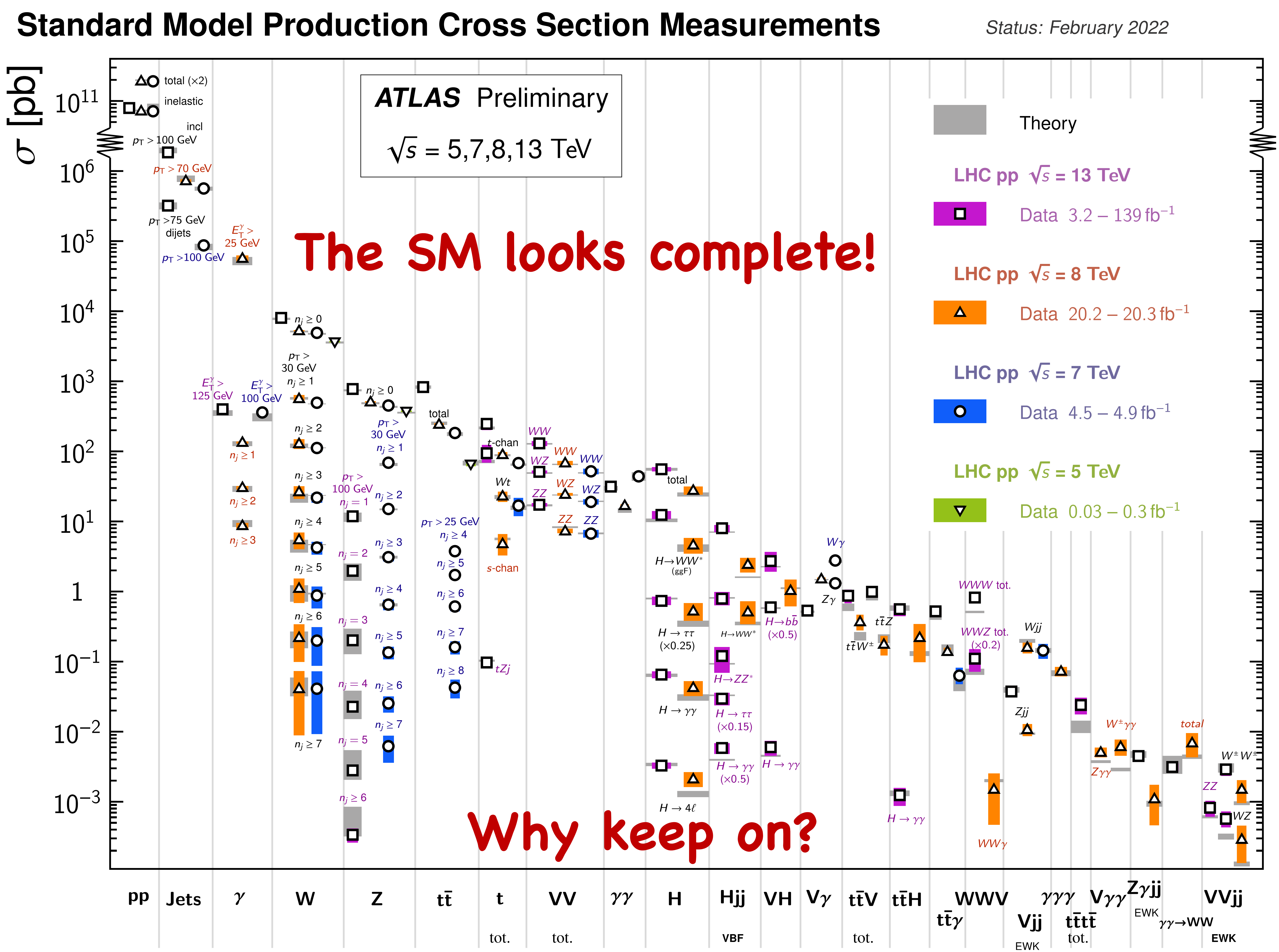
Search for New Physics with Machine Learning

concepts, applications and recent progress

Roman Pasechnik
Lund University

HAMLET-PHYSICS 2024

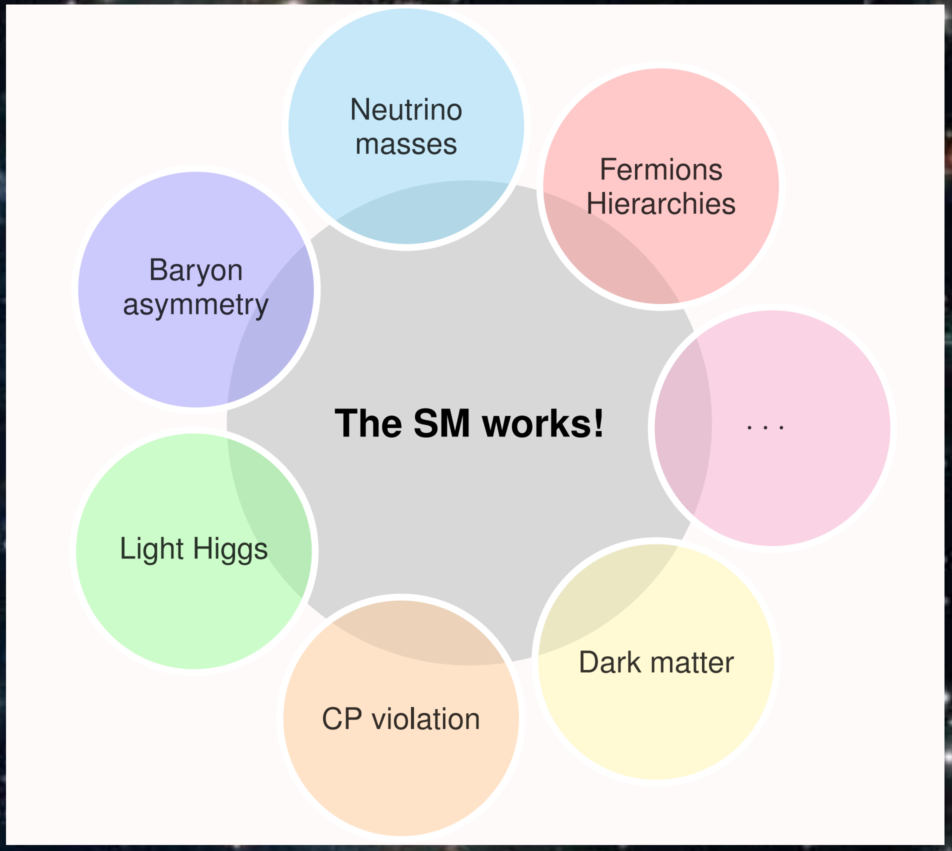
Impressive performance of the Standard Model



The SM looks complete!

Why keep on?

Open questions



Unexplained phenomena:

- Dark Matter
- Matter-antimatter asymmetry
- Dark Energy

Unsatisfactory structure of the SM:

- Hierarchy problems (Higgs, flavour)
- Naturalness
- Quantum Gravity etc

Plethora of Beyond the SM theories to test against data

NASA, ESA, AND THE HUBBLE HERITAGE TEAM (STSCI/AURA)

Goals of Particle Physics

- **Explore physics at the highest energy scale (TeV scale at the LHC)**

  - search of new Higgs scalars (= Higgs "partners")
  - search for low-energy traces of supersymmetry (SUSY)
  - investigate various scenarios of physics beyond the SM

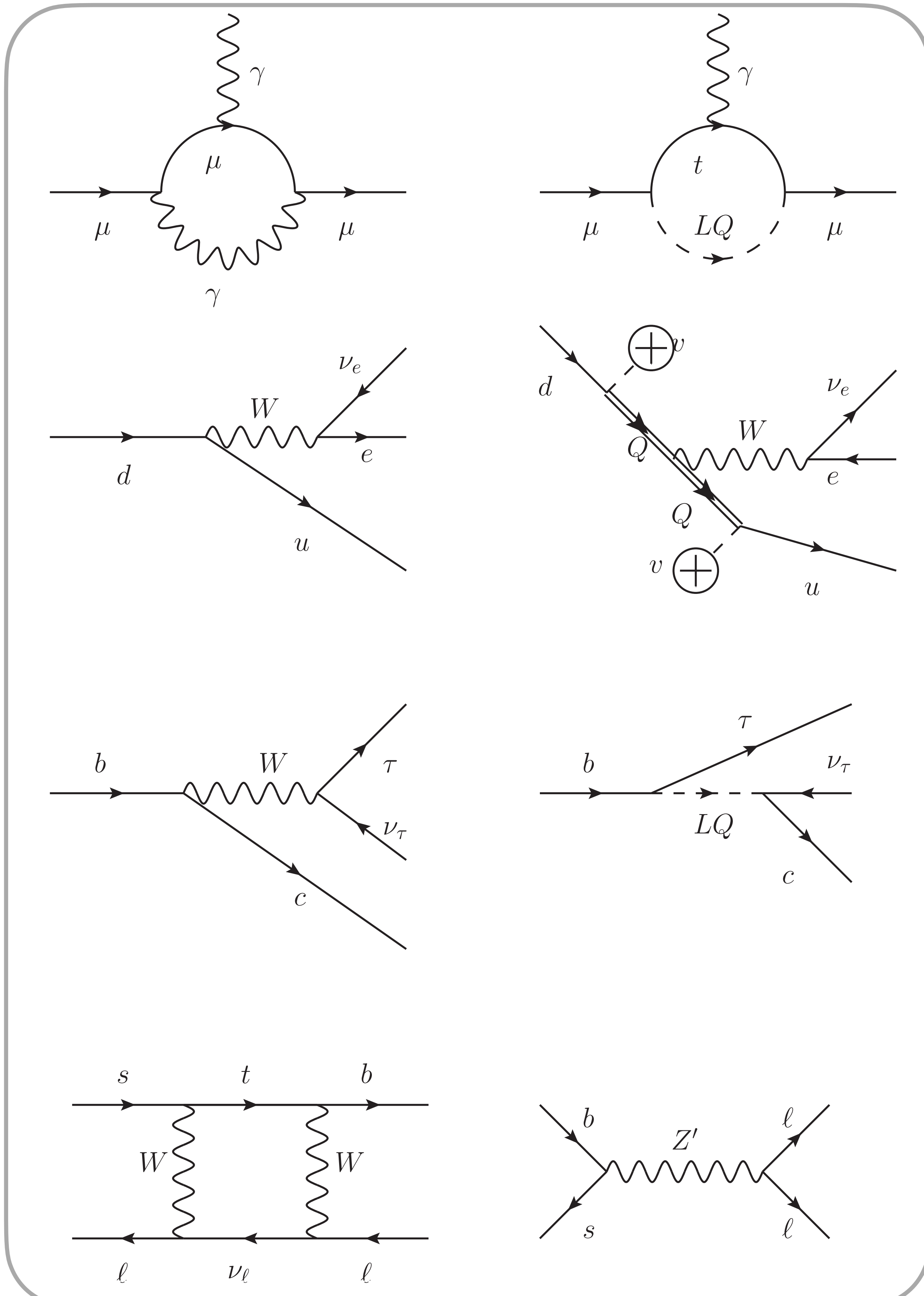
- **Precision measurements of SM processes:**

  - test nontrivial predictions of the SM, including very rare processes
  - search for deviations from the SM sensitive to new physics at high scales
  - improve precision of the SM parameter measurements
  - study the Higgs boson and the EWSB sector
  - study QCD dynamics and parton content of the proton
  - QCD/MC tools development
  - exploit the SM measurements as "Standard Candles" to tune and test detector performance

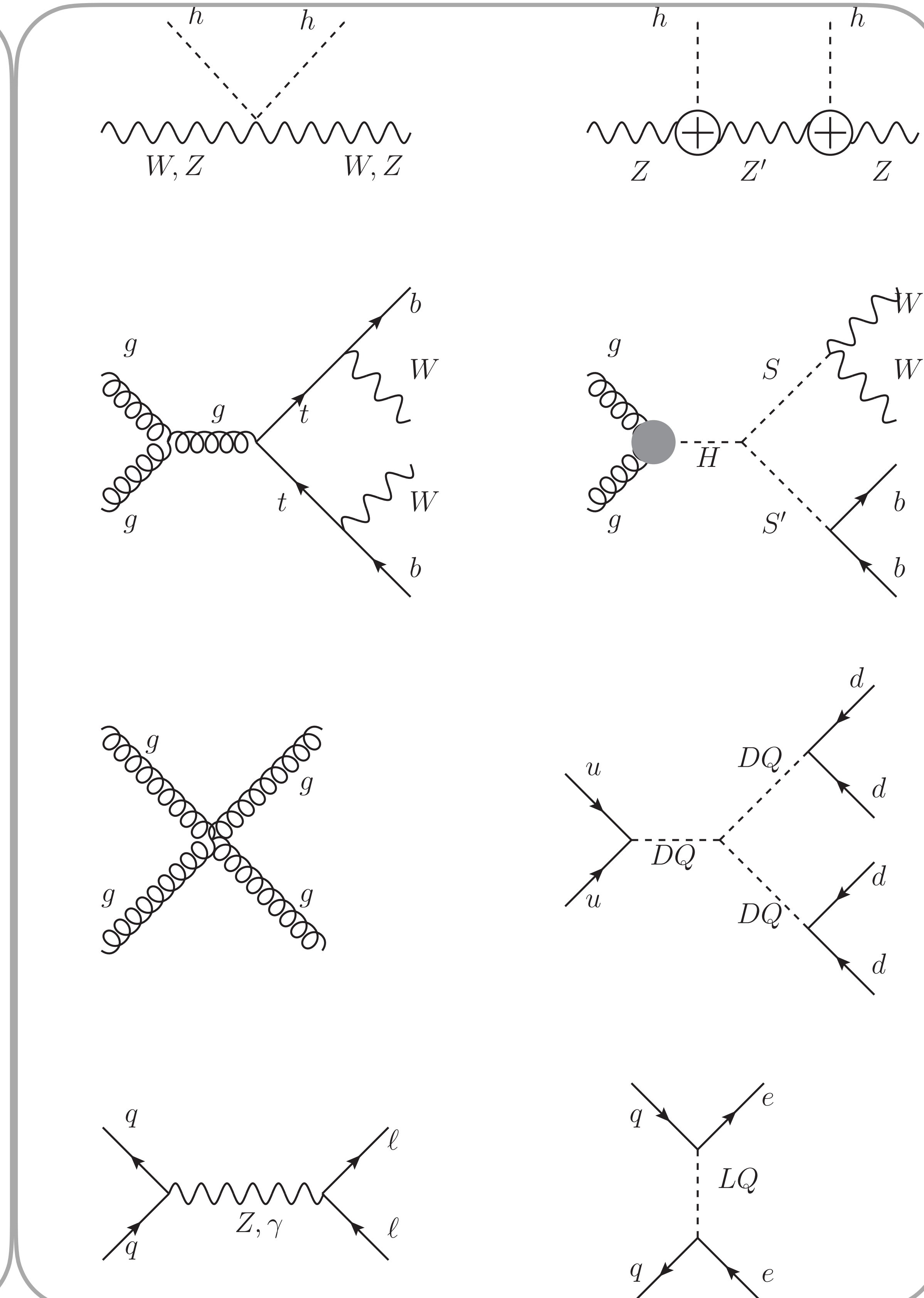
Emerging anomalies in collider data

Crivellin, Mellado, Nature Reviews Physics 6, 294 (2024)

Standard Model | **New Physics**



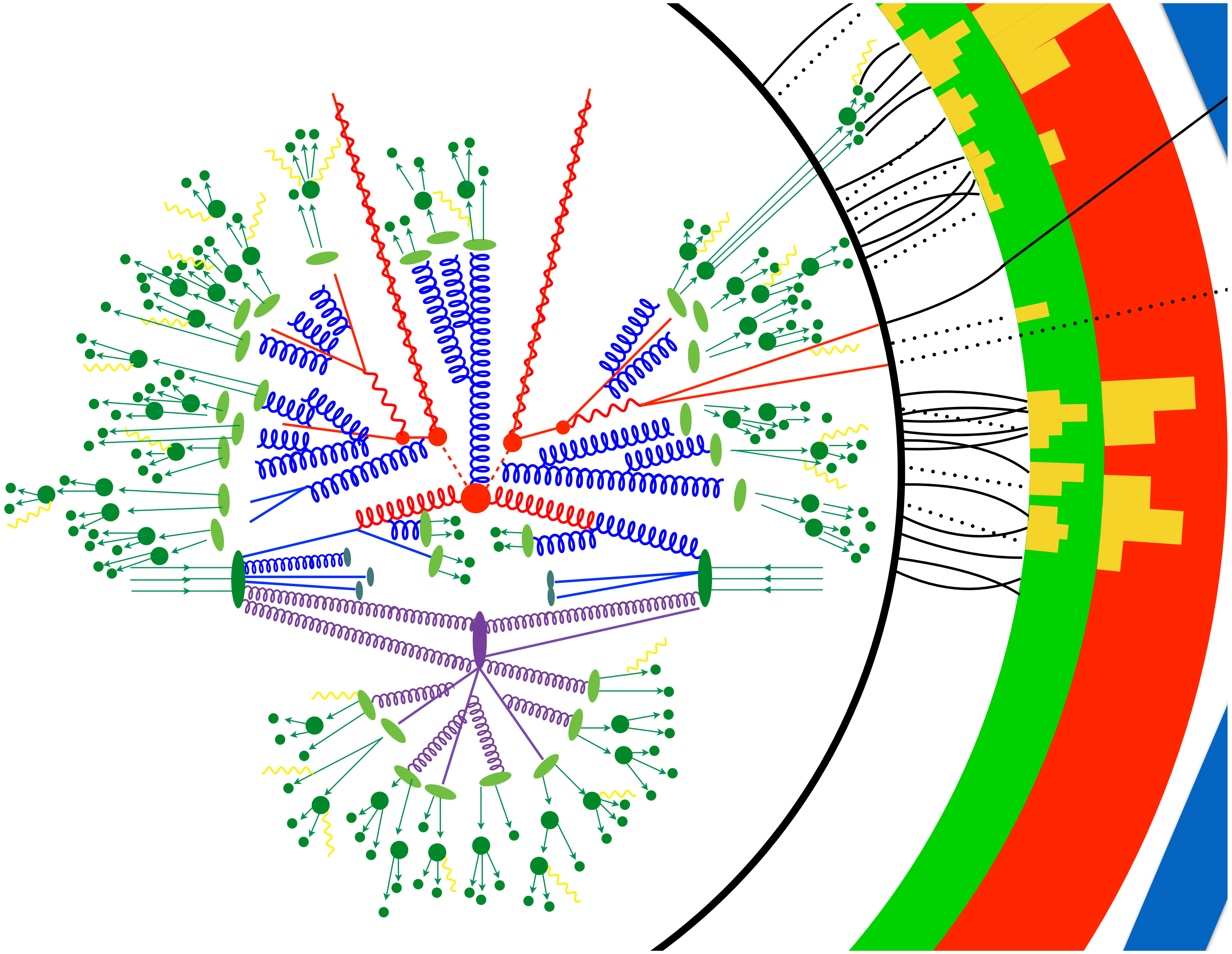
Standard Model | **New Physics**



Why ML is relevant for Particle Physics

Karagiorgi, Kasieczka, Kravitz, Nachman, Shih, arXiv:2112.03769

Microscopic physics (particle interactions, scattering…) $\Rightarrow$ Large-distance phenomena (hadronisation, propagation, jets…) $\Rightarrow$ Detector response



Particle physics experiments produce enormous datasets of high complexity!

ML is a powerful new tool that enables us to get more physics out of these big datasets

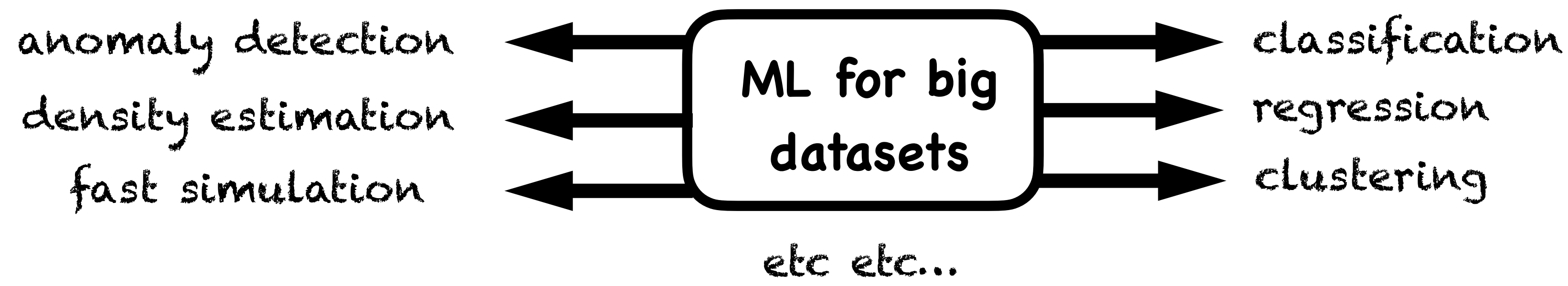
We live in a very special point of history similar to invention of a telescope

ML would enable us to see features we could not see before — “telescope” for Big Data!

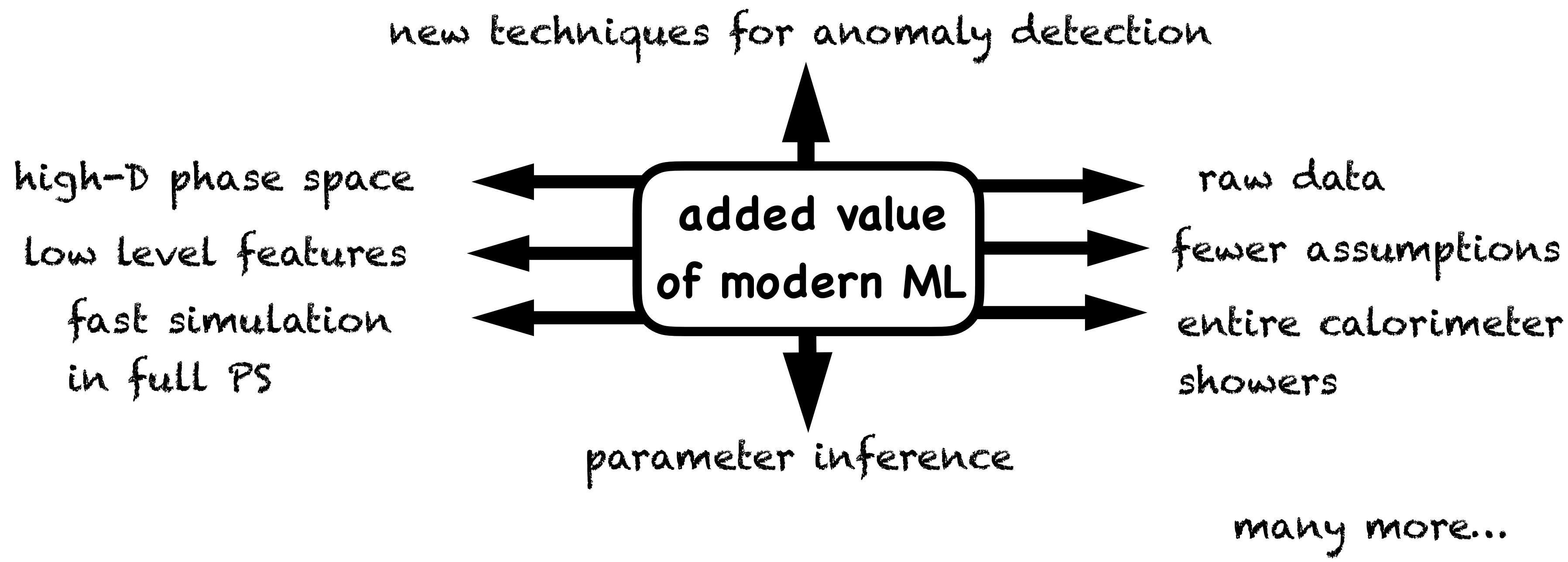
What ML can do for particle physicists

Mehta et al, Phys. Rept. 810, 1–124 (2019)

What we could do before, but ML does better job ⇐



What we could not do before, but ML makes it possible ⇐

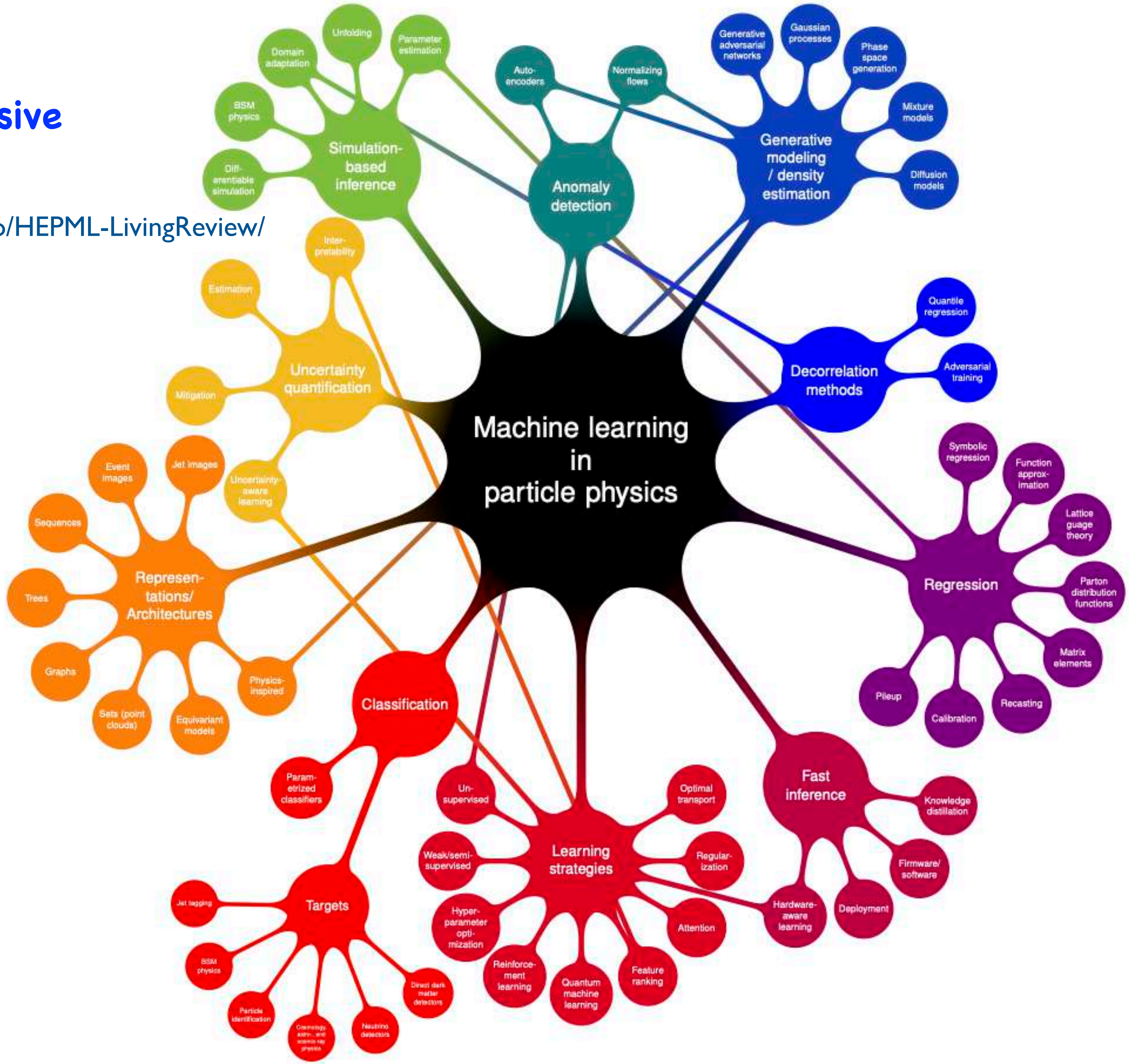


Biggest advantages of ML

- ✦ greatly enhanced sensitivity/precision [x10-100]
- ✦ accelerated simulation [e.g. fast simulation]
- ✦ accelerated/efficient extraction of physics [inference]
- ✦ ML is cross-disciplinary [same methods can be used in many fields]
- ✦ ML is not only for experimentalists - theorists use "simulated data" a lot!

For comprehensive review, see

https://iml-wg.github.io/HEPML-LivingReview/

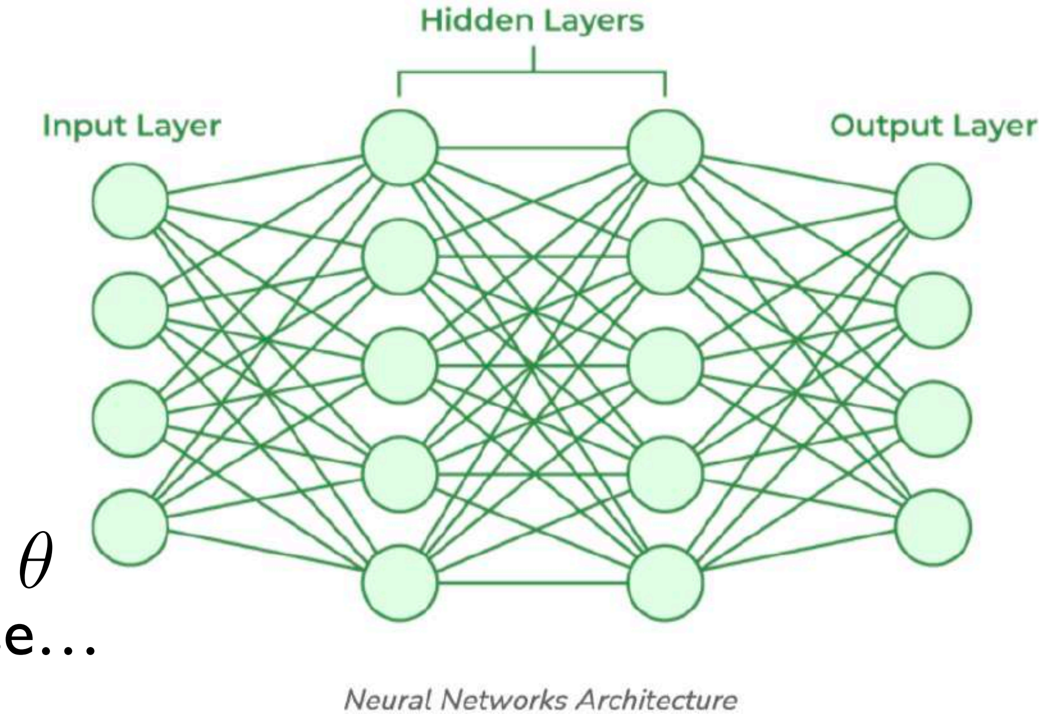


https://github.com/jmduarte/Nomological_Net_ML_Particle_Physics?tab=readme-ov-file

Learning from data

ML = sophisticated curve fitting!

Example: feed-forward NN (MLP)



- the data instances $\vec{x}_i \subset R^d \qquad i = 1 \dots N$ are considered to be drawn i.i.d. from some data distribution $p_{\rm data}(x)$

- often we are interested to "learn" a function $f(x, \theta)$ from the data for some parameters (e.g. weights, biases) θ
  $\Rightarrow$ Neural Network, or made up of NNs, or smth else…

- the learning process (e.g. NN training) is optimisation of an objective (loss function)

$$L(\theta) = \sum_{\rm data} \mathcal{L}[f(x,\theta)]$$

- supervised learning [regression or classification] — we want $f(x,\theta)$ to take a specific form e.g. binary classification $f(\vec{x}_i, \theta) = y_i$ for truth labels $y_i = 1 \text{ or } 0$ (QCD jets vs top jets)

- the objective is to get as close to the truth labels as possible e.g. minimise the mean squared error (MSE) loss

$$\mathcal{L} = (f(\vec{x}_i, \theta) - y_i)^2$$

- Truth labels often come from simulation, or when categorisation of the data is obvious e.g. hand-labelled data (cat vs dog, natural images etc)

Less than supervised

Data often come from experiments (e.g. LHC) without truth labels (non human interpretable) while simulations are not perfect -> What to do?

- ✦ in Particle Physics the data is very complex and not clearly separable into categories
  $\Rightarrow$ large overlap in their distributions

- ✦ Unsupervised learning — no labels available at all — data-driven/simulation-free approach!
  $\Rightarrow$ uncovering hidden patterns, structures or relationships in the data
  $\Rightarrow$ Examples: clustering similar data points, reducing dimensionality for visualisation

- ✦ "Noisy" labels: weakly supervised learning — a powerful tool in New Physics searches:
  $\Rightarrow$ data-derived labels that correlate with a given category [e.g. signal S vs background B] but may not be EXACTLY in that category
  $\Rightarrow$ Splitting the data into "signal region" [S-enriched] + "control region" [B-enriched], a generative model is applied for anomaly detection

- ✦ A mix of labelled and unlabelled data: semi-supervised learning
  $\Rightarrow$ simulation+data to mitigate the simulation effects, or when parts cannot be labeled

- ✦ Data-driven methods to learn the objective: self-supervised learning — by using symmetries or deleting parts of the data and trying to fill that in [relevant in Large Language Models]
  $\Rightarrow$ useful for learning embeddings (e.g. using data-derived labels on jets related or not related by rotation, one learns a jet representation encoding the symmetry)

Generative ML

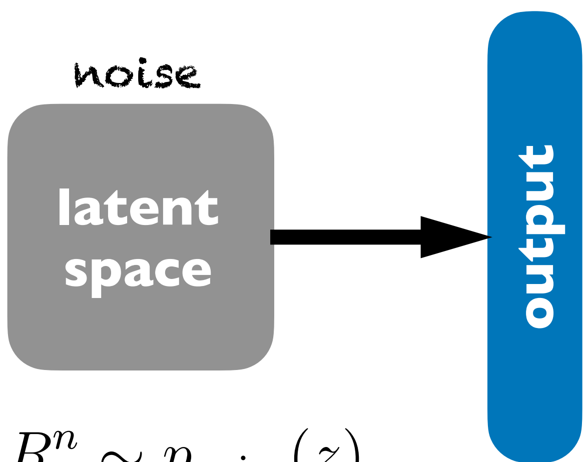
Can we generate more samples that follow the same distribution as the data?

✦ We want to learn the data probability distribution $p_{\rm data}(x)$ [density estimation] and then sample from it — often, a very difficult task!
$\Rightarrow$ we can learn to sample from $p_{\rm data}(x)$ without actually learning this function

✦ Generative modelling — to learn $p_{\rm model}(x)$ as close as possible to $p_{\rm data}(x)$ and then sample to generate (dream up) synthetic data capturing underlying patterns of the original dataset

data

noise



synthetic data sampled from $p_{\rm model}(x)$

learnt explicitly or implicitly

$z \subset R^n \sim p_{\rm noise}(z)$

typically, a simple distribution like $N(0,1)^n$

Methods of generative ML

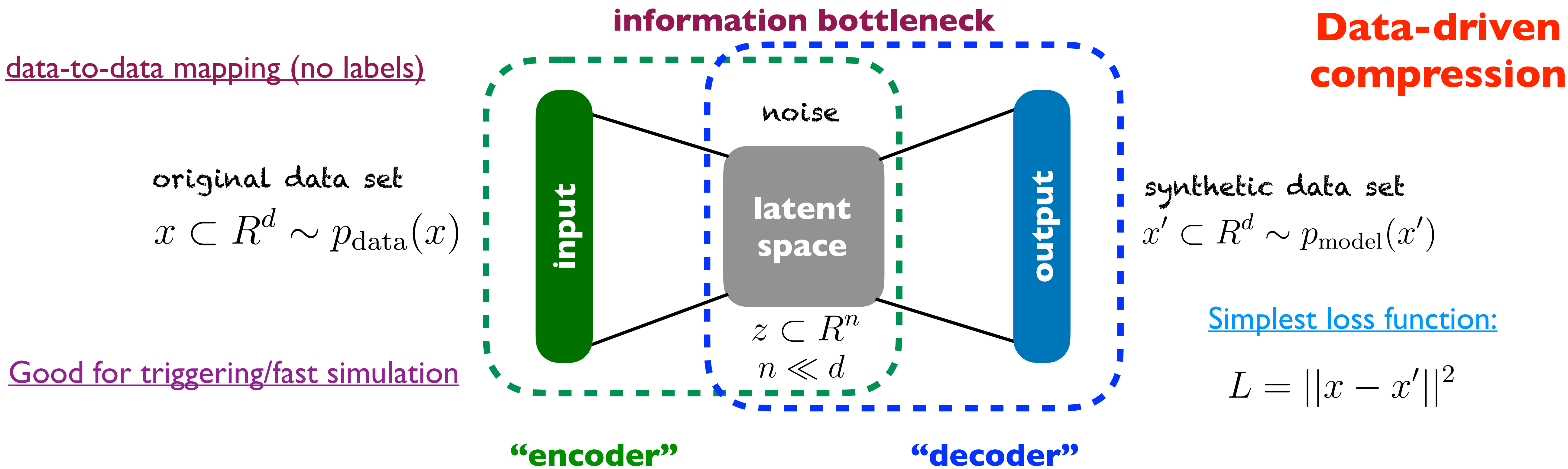
$\Rightarrow$ Generative Adversarial Networks (GANs)
$\Rightarrow$ Variational Autoencoders (VAE)
$\Rightarrow$ Normalising Flows
$\Rightarrow$ Diffusion Models

Common applications in Particle Physics:

✦ fast simulation e.g. "surrogate" modelling (training on few samples to generate more), phase space sampling (integrations)
✦ anomaly detection — outliers (clean up data), group anomalies (bumps from NP)
✦ simulation based inference e.g. EFT fits

Autoencoders

✦ AE maps data back to itself through reduced latent space trying to figure out a latent representation of the data that captures its essential features



✦ AE trained on "normal" events as an "anomaly detector" [only outliers, not overdensity]
$\Rightarrow$ an outlier causes the loss to have a "cluster" at large values
$\Rightarrow$ anomalies may be separated in the latent space — compression enhances clustering
BUT no guarantee that AE learns sensible latent space [no probabilistic interpretation]

✦ Simulation-based AE for New Physics search

Farina, Nakai, Shih, PRD 101, 075021 (2020)
Heimel et al, SciPost Phys. 6, 030 (2019)

$\Rightarrow$ take simulated jet images as data: QCD jets as B and NP jets as S (anomaly)
$\Rightarrow$ if S is rare, AE separates the anomaly well (in the tail of the loss)

✦ AE as a "complexity detector" Finke et al, JHEP 06, 161 (2021)
$\Rightarrow$ train on QCD jets, finds top jets, BUT train on top jets, does not find QCD jets

Variational Autoencoders

Can we enforce the latent space to have a suitable probabilistic interpretation?

- VAE as a latent variable model: $z \sim p(z)$ (the "prior") while $x \sim p_\theta(x|z)$ we get a set $\{x_i, z_i\}$ drawn from $p(x, z)$ — by integrating out z we get data distribution $p_\theta(x)$

- To determine the conditional probability $p_\theta(x|z)$ — learn it by maximising the maximum likelihood estimation (MLE) w.r.t NN parameters θ

"decoder"

$$\text{MLE} = \sum_{x \sim p_{\text{data}}(x)} \log p_\theta(x)$$ with Bayesian evidence $$p_\theta(x) = \int p_\theta(x|z)p(z)dz$$ **hard!**

"encoder"

- Variational "posterior" $r_\psi(z|x)$ — still samples z-space but differently depending on x

- Utilising MC sampling of the integral from the posterior and applying Jensen inequality:

Kullback-Leibler divergence

$$\log p_\theta(x) \geq \sum_{z \sim r_\psi(z|x)} \log \frac{p_\theta(x|z)p(z)}{r_\psi(z|x)} = \log p_\theta(x) - \text{KL}(r_\psi(z|x)||p_\theta(z|x))$$ "evidence lower bound" (ELBO)

true posterior for a given θ

- Taking normal distribution $p_\theta(x|z) = N(\mu_\theta(z), \beta)$

$$\text{ELBO} = E_{r_\psi(z|x)}\left[-\frac{||x - \mu_\theta(z)||^2}{2\beta^2}\right] - \text{KL}(r_\psi(z|x)||p(z))$$

VAE as a "regularised" vanilla AE with a "smoothing" KL-term (posterior tends to the prior)

Maximised!

reconstruction error of vanilla AE

Normalising Flows

Invertible map between the data and latent spaces

Papamakarios at al,
J. ML Res., 22(57) 1, 2021

✦ We can achieve both — get the latent space z and do density estimation (DE)

$$z = f(x, \theta) \longleftrightarrow x = f^{-1}(z, \theta)$$

$$p(z) \longleftrightarrow p_\theta(x) = p(z)\left|\det \frac{df}{dz}\right|$$

Optimisation problem:
to fit parameters θ
to the data

✦ We directly optimise the negative-log likelihood [tends to perform better than VAE]

$$L = -\sum_x \log p_\theta(x)$$

best model that generalises best to the unseen test set

✦ Trade off: more ambitious/challenging, bad scaling with dimensions;
$\Rightarrow$ Jacobian of $d \times d$ matrix takes $O(d^3)$ operations repeated many times

✦ A small family of invertible functions with upper-triangular matrices: $O(d)$ operations!
$\Rightarrow$ "autoregressive transformation" $z_1 = f_1(x_1)\,, \quad z_2 = f_2(x_1, x_2)\,, \quad \ldots \quad z_n = f_2(x_1 \ldots x_n)$

✦ To gain on expressivity [= having enough parameters to learn the transform], one can chain multiple such transforms, permute between them —
$\Rightarrow$ Masked Autoregressive Flow (MAF): $z_i = \alpha_i(x_1, ..x_{i-1}, \theta) z_i + \mu_i(x_1..x_{i-1}, \theta)$ where the coefficients can be thought as outputs of one big NN (slow sampling)
$\Rightarrow$ Inverse Autoregressive Flow (IAF): inverse algorithm (slow training)

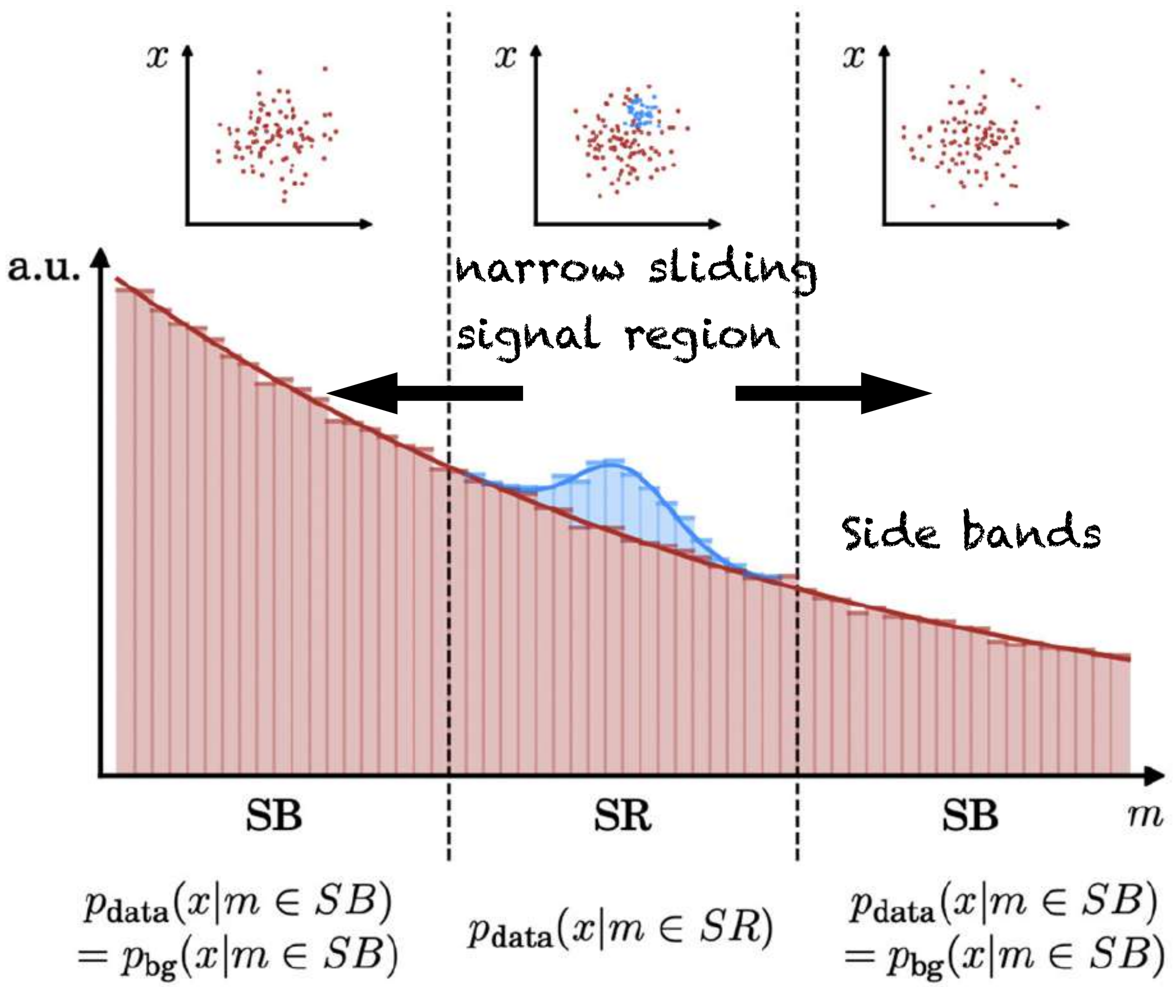
✦ "Density distillation": first train the MAF then distill it into an IAF [fit IAF to MAF]

Resonant anomaly detection: SIC

Golling et al, EPJC 84, 241(2024)

How we use ML in searches for new phenomena in Particle Physics?

✦ Assume S is localised (resonant) in some feature (typically, invariant mass) and B is smooth



✦ Inclusive bump hunt — standard technique for new particle searches at colliders:
- $\Rightarrow$ split the distribution into SR and SBs
- $\Rightarrow$ smooth interpolation provides fully data-driven B in the SR
- $\Rightarrow$ discovery significance via Poisson $\sigma = \dfrac{S}{\sqrt{B}}$

✦ How ML can enhance the bump?
- $\Rightarrow$ multivariate bump hunt: looking for correlated excesses in other features x (e.g. jet substructure, missing energy etc)

✦ Anomaly Score $R(x)$ — large for S, and small for B — must be uncorrelated with the mass

✦ Significance Improvement Characteristic (SIC) — how much the significance is improved by a cut on $R(x)$ based upon how many S,B events survived:

$$\text{SIC} = \frac{\epsilon_S}{\sqrt{\epsilon_B}}$$ for cut efficiencies $\epsilon_{S,B}$

We want a ML technique that produces large SIC in a data-driven way

Proof of concept: inject a small signal to simulated B, see how it shows up in SIC

CWoLa Hunting

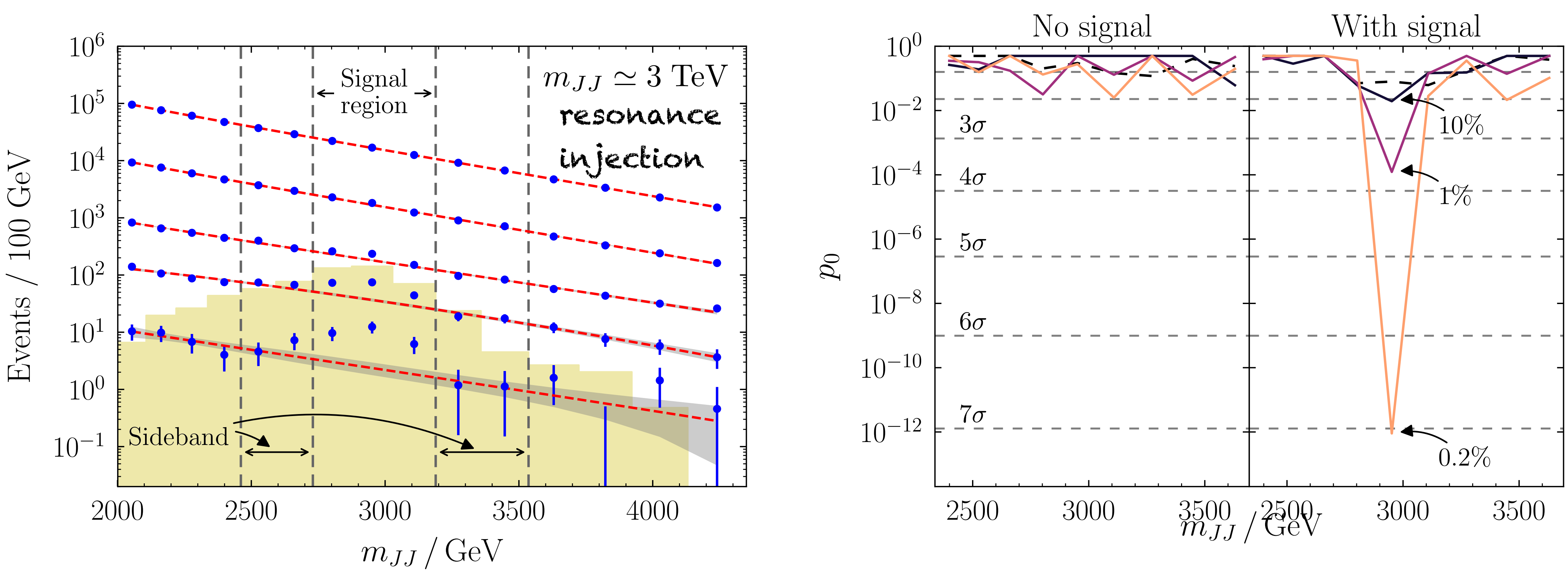
How can we evaluate anomaly score and improve discovery significance?

Collins, Howe, Nachman, PRD 99, 014038 (2019)
ATLAS search in real data: 2005.02983

- For any S, the S-to-B likelihood ratio $p_{\rm S}(x)/p_{\rm B}(x)$ is an optimal S/B classifier!
  $\Longrightarrow$ how do we learn approximations to $p_{\rm S,B}(x)$ using ML?

- "Classification W/o Labels" (CWoLa) — train a classifier to learn SB vs SR to approximate the ideal anomaly score $R_{\rm ideal} = p_{\rm data}(x)/p_{\rm B}(x)$ fully correlated with the ratio $p_{\rm S}(x)/p_{\rm B}(x)$
  $\Longrightarrow$ if features x and the mass are uncorrelated in B, train a binary classifier to learn $R(x)$ using noisy labels SR (label 1) and SB (label 0) — weak supervision!
  $\Longrightarrow$ output: probability that a given x comes from SR

Bayes' theorem: $$p({\rm SR}|x) = \frac{p_{\rm SR}(x)}{p_{\rm SR}(x)+p_{\rm SB}(x)} = \frac{R(x)}{1+R(x)} \longrightarrow R(x)$$



Anomaly detectors with correlation

What happens if features are correlated with mass?

- CWoLa stops working — becomes just a B-to-B classifier!

- "Anomaly detection with density estimation" (ANODE) Nachman, Shih, PRD 101, 075042 (2020)
  $\Longrightarrow$ no classifier, train conditional DEs (Normalising Flows) to learn SR and SB densities conditioned on mass

- "Classifying Anomalies thorough outer DE" (CATHODE) Hallin et al, PRD 106, 055006 (2022)
  combine best of CWoLa and ANODE: learn SB density in ANODE, interpolate into SR, train a classifier on SR data in CWoLa — great performance!

- CURTAINs method Raine, Klein, Sengupta, Golling, Front. Big Data 6, 899345 (2023)
  instead of Normalising Flows, do invertible NN that learns to map SB to SB using optimal transport loss

- LaCATHODE method Hallin et al, PRD 107, 114012 (2023)
  using CATHODE in latent space

- Methods using simulation with re-weighting:
  SALAD Andreassen, Nachman, Shih, PRD 101, 095004 (2020)
  FETA Golling, Klein, Mastandrea, Nachman, PRD 107, 096025 (2023)
  SA-CWola Benkendorfer, Pottier, Nachman, PRD 104, 035003 (2021)

Credit: David Shih

Anomaly detectors: performance

Golling et al, EPJC 84, 241(2024)

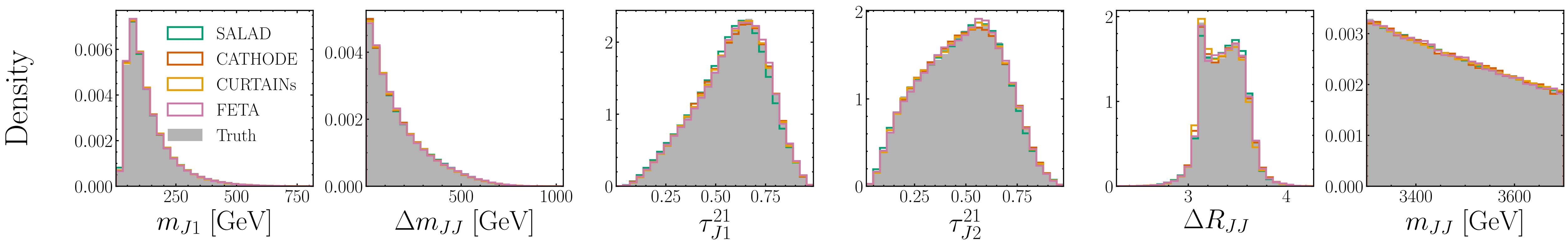
✦ ML model building: LHC Olympics R&D data set - fully labeled

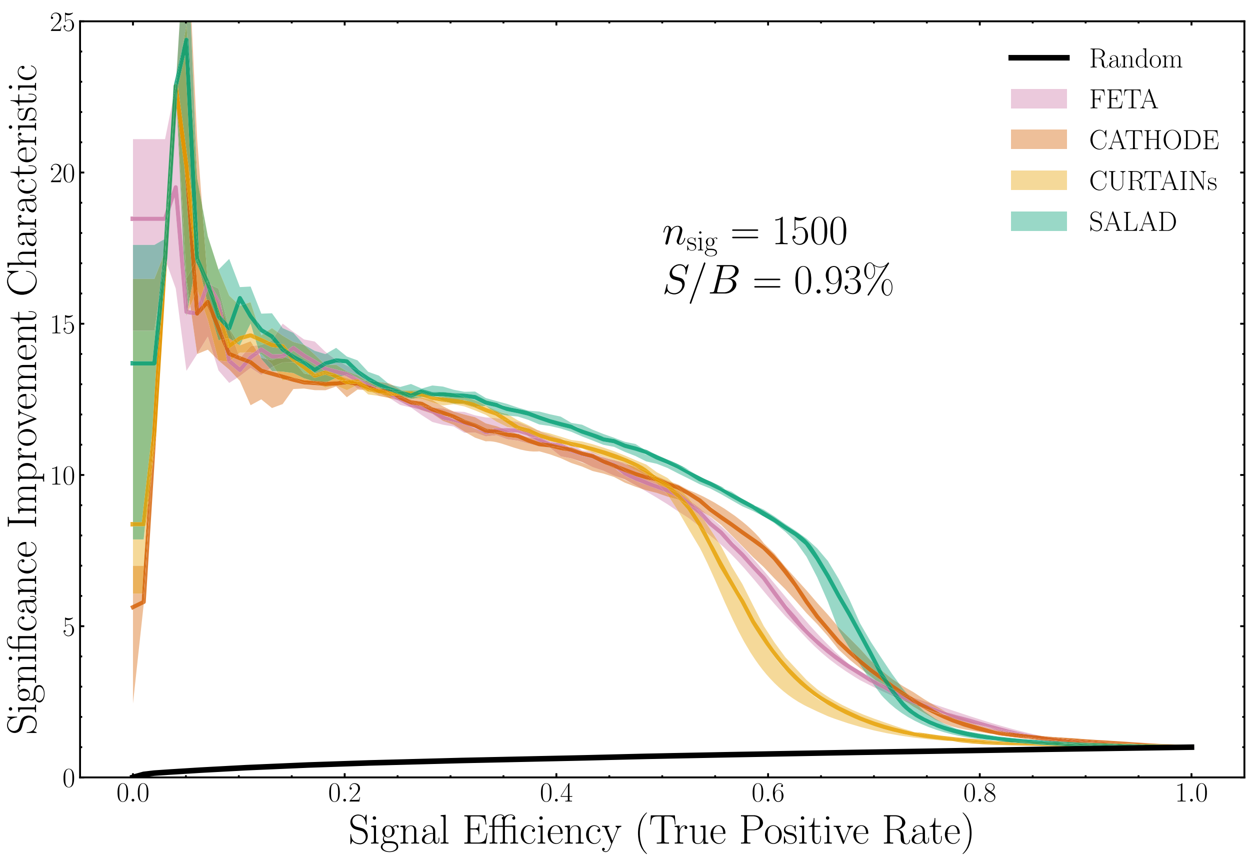
Pythia + Delphes: 13 TeV pp, leading jet $p_T > 1.2$ TeV

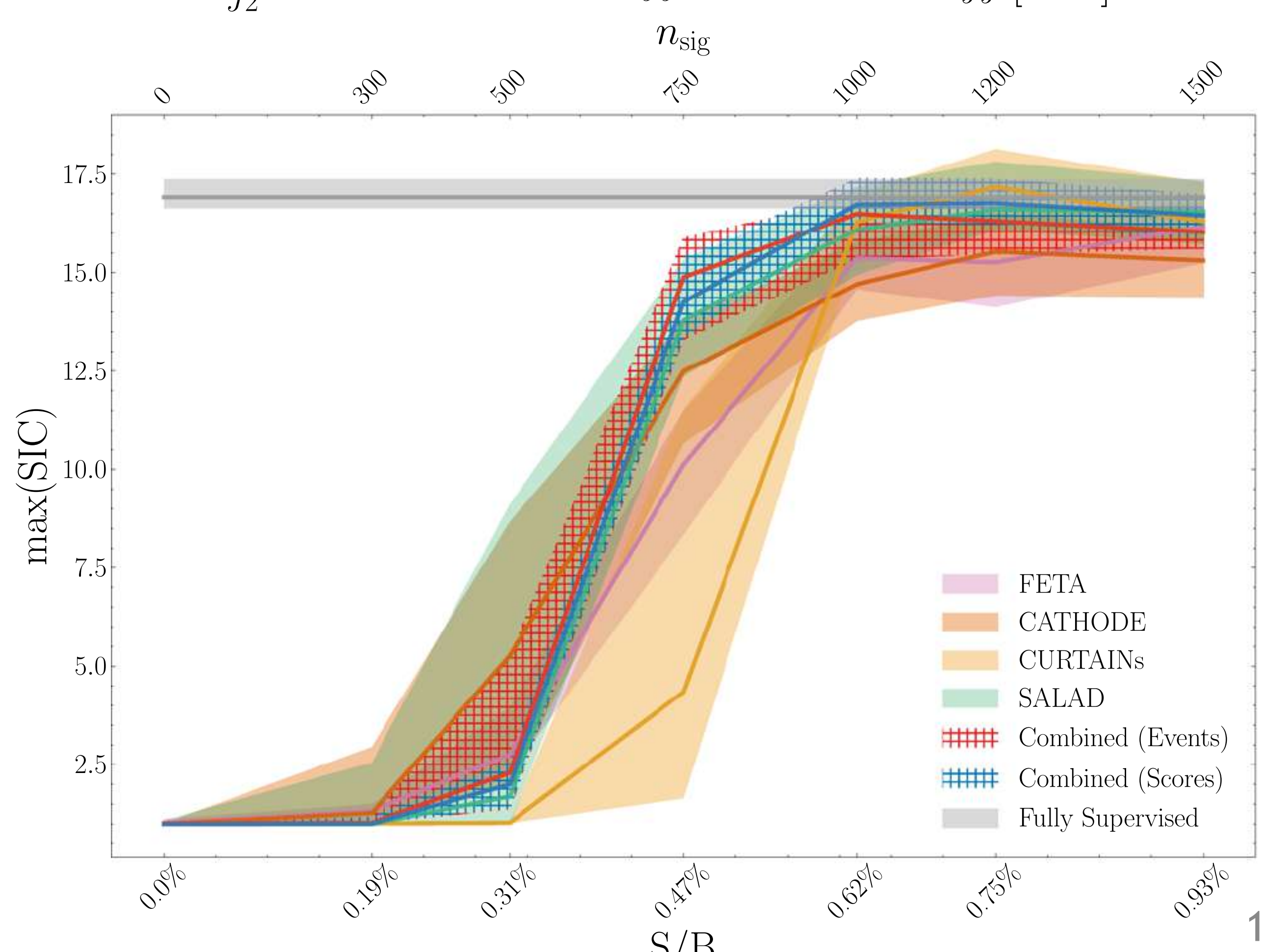
SM background: QCD di-jets

R&D signal: $pp \to Z' \to X(\to jj) + Y(\to jj)$ $\quad m_{Z'} = 3.5$ TeV $\quad m_Y = 100$ GeV $\quad m_X = 500$ GeV

Synthetic data in 5D feature space + di-jet mass (SM B only)



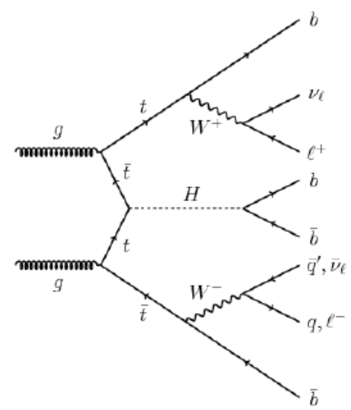


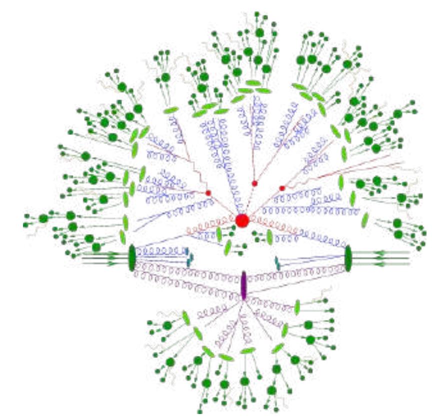


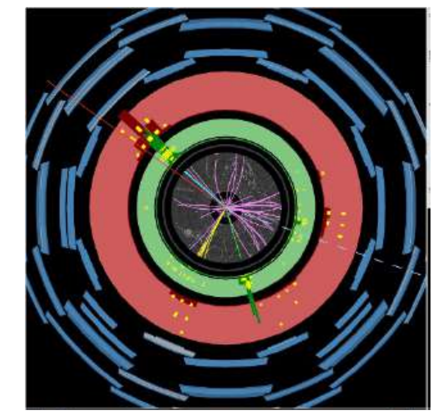
Simulation based inference

Data generation

Given theory parameters - what detectors observe?





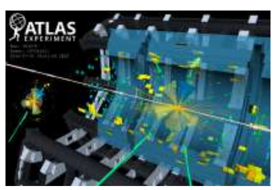


O(20) Fundamental physics parameters θ

O(10) particles

O(100) particles

O(10^8) detector elements

$p($ 

$|$  $)$?

100M detector elements

SM / BSM parameters

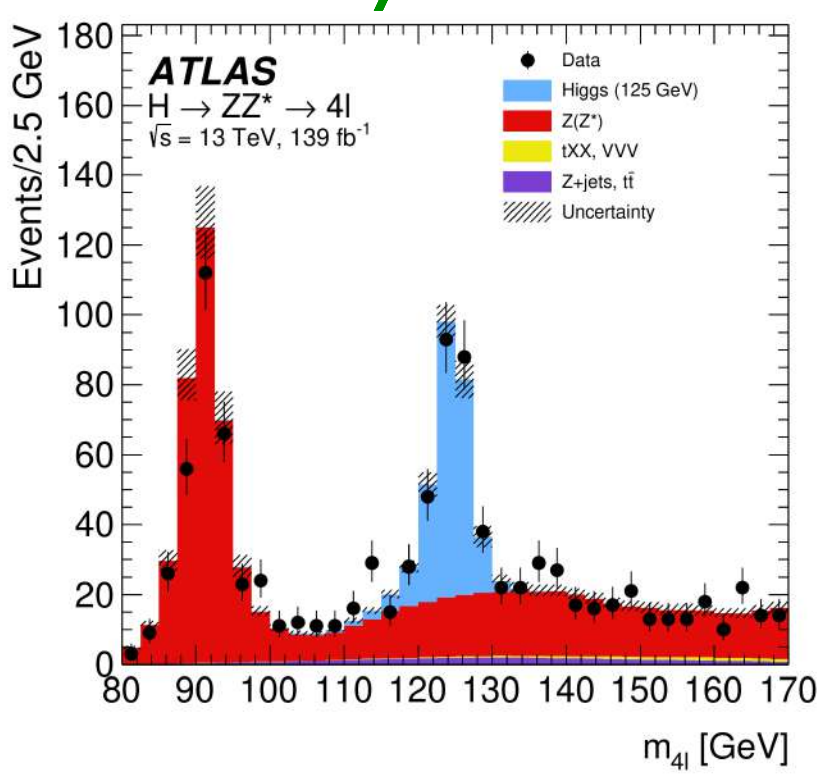
$$p(x|\theta) = \int dz\, p(x|z_h)p(z_h|z_p)p(z_p|\theta)$$

unobserved random processes: mathematically intractable!

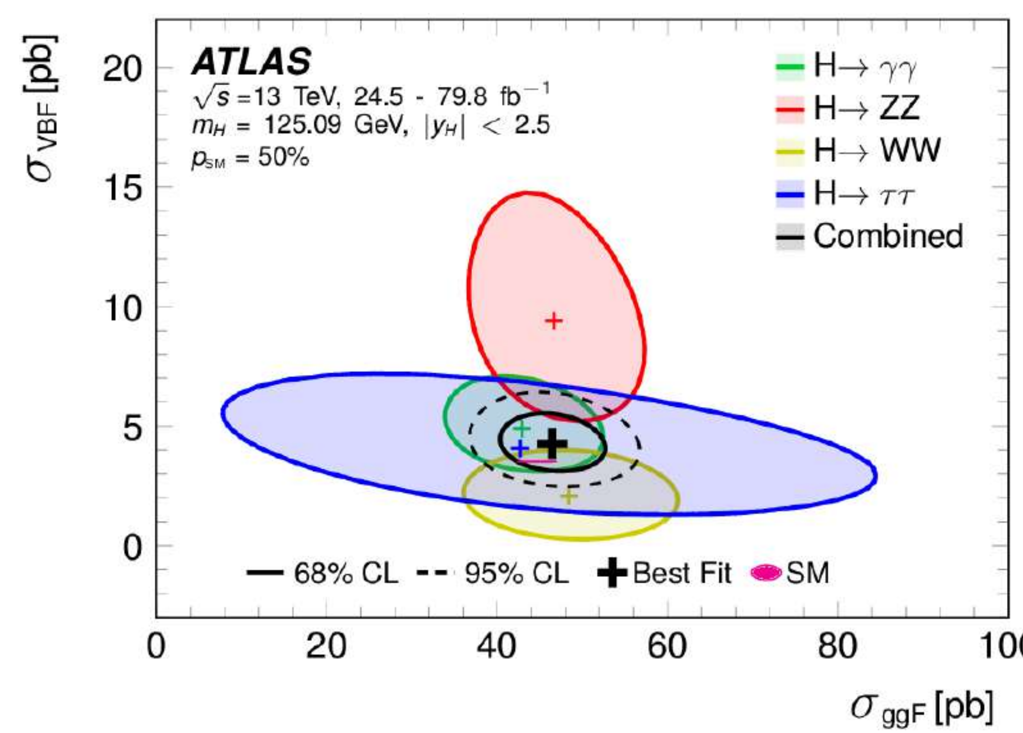
Density estimation

Likelihood for inference

in 1D we do this all the time!



$\Rightarrow$



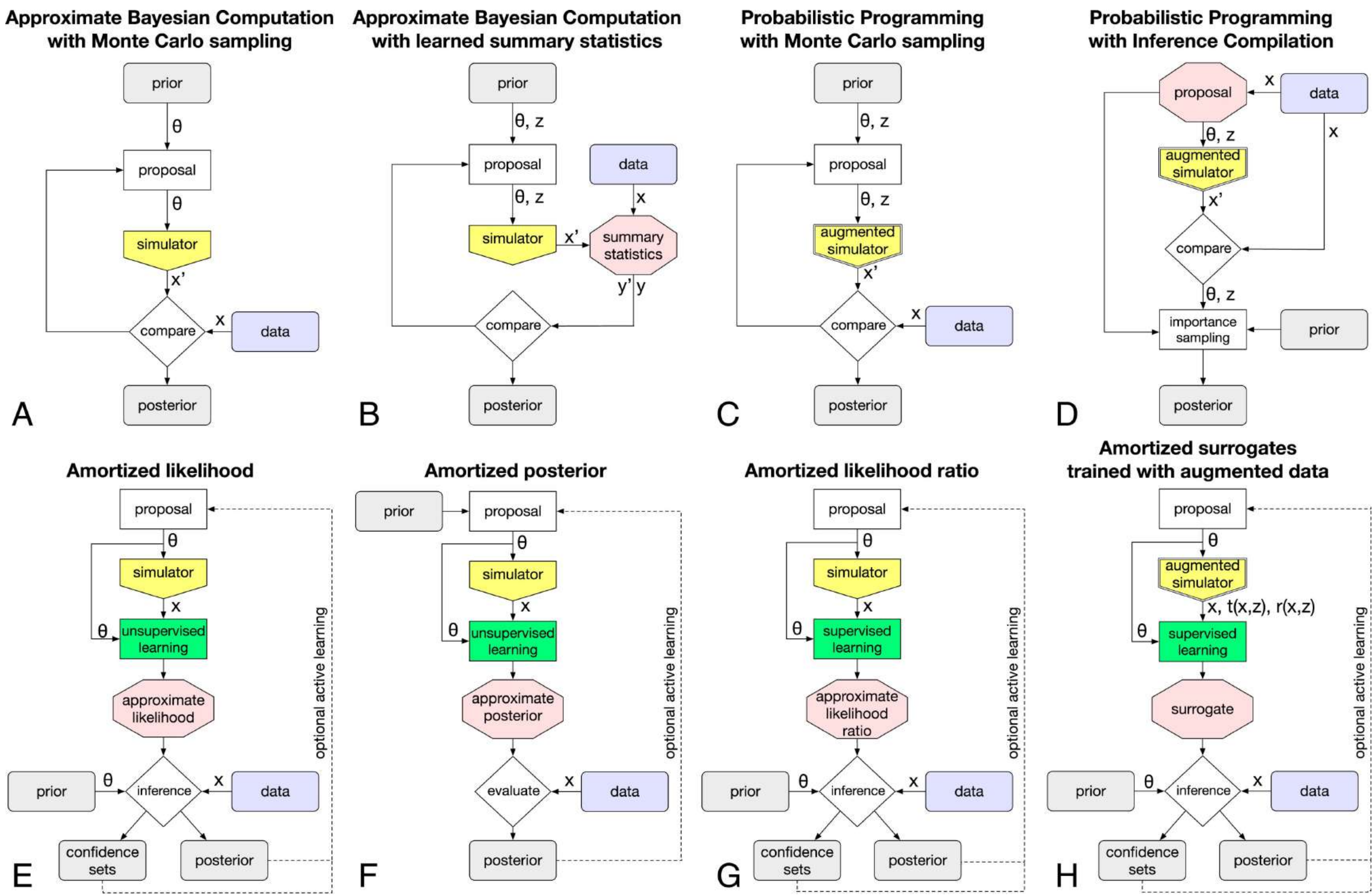
This is extremely challenging in large d's!

Image credit: Michael Kagan

Simulation based inference with ML

Summary of different approaches:

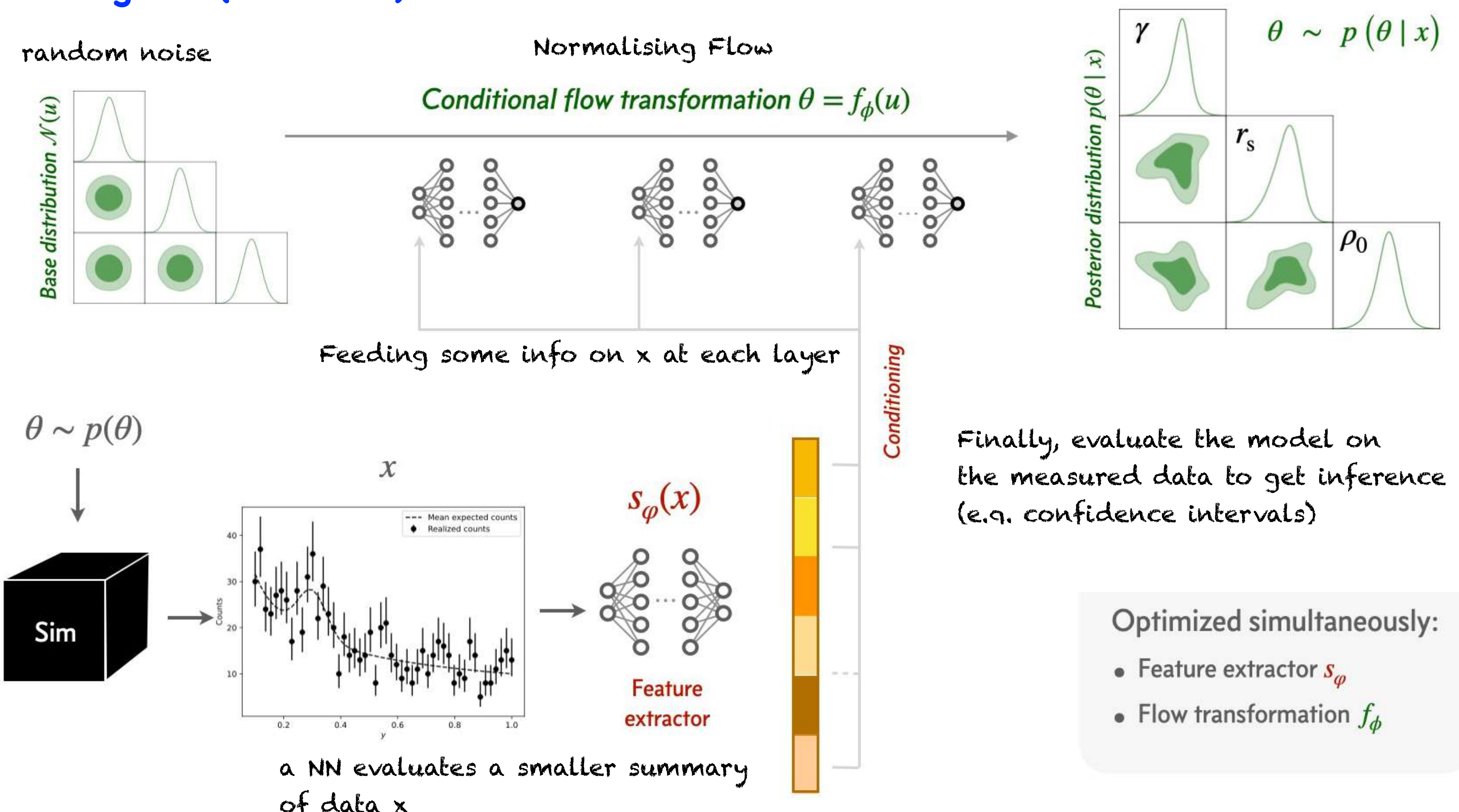
Cranmer, Brehmer, Louppe, PNAS 117, 30055 (2020)



Use simulator to train NN to approximate likelihood ratio (using classifier), likelihood or posterior (using Normalising Flow), then use NN to do parameter inference on observed data

Conditional neural posterior DE

Getting posterior distribution of model parameters for a given (simulated) data



Physics-informed NN architectures may offer new ways to gain on efficiency of training and precision

Image credit: S. Mishra-Sharma

Summary

- ✦ ML is a new exciting research field that bridges Data Science, theoretical and experimental physics together
- ✦ ML is revolutionising research in Particle Physics and other data-intensive frontier fields offering a huge improvement in precision and computational efficiency, and new applications
- ✦ ML enables to hunt for subtle features in large and complex datasets and potentially to infer the best fundamental physics model combining vast amounts of data from different measurements and theoretical constraints
- ✦ While New Physics remains elusive, ML offers new opportunities for future discoveries