Machine Learning at LHC

Jernej F. Kamenik

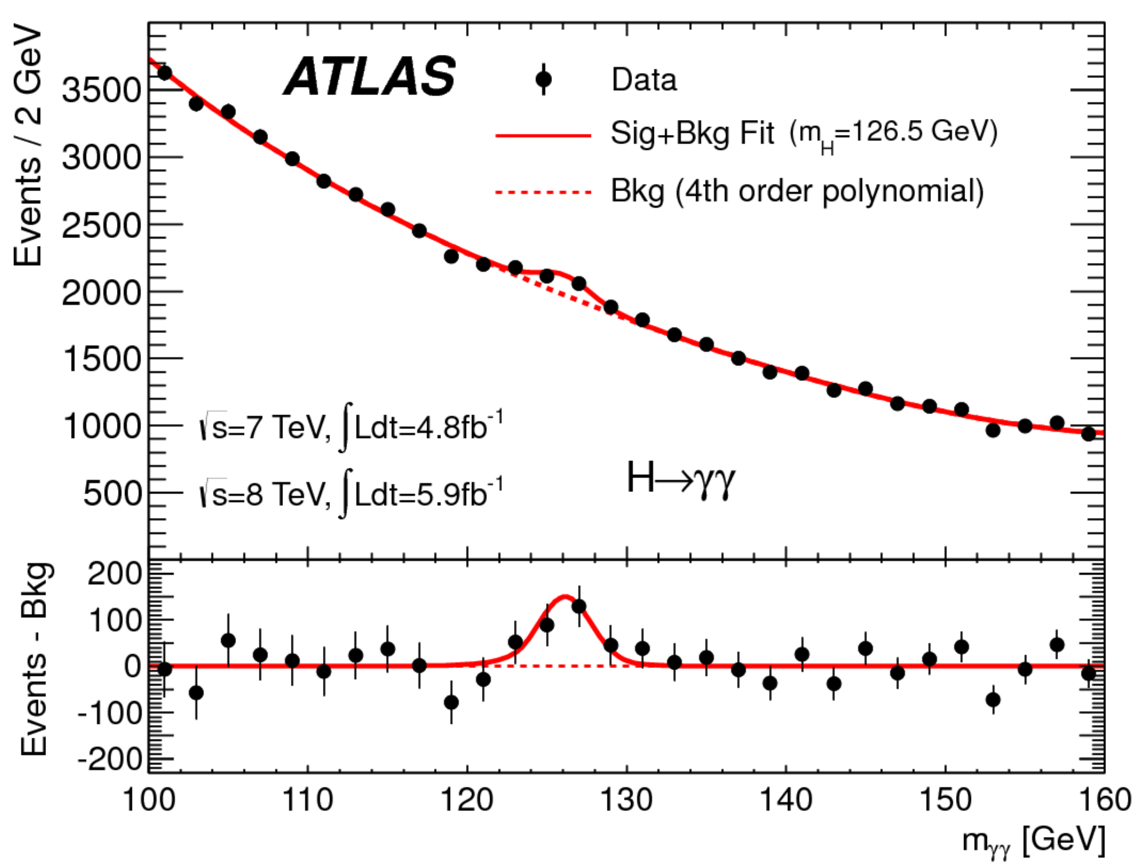
Institut
"Jožef Stefan"
Ljubljana, Slovenija

Univerza v Ljubljani
Fakulteta za matematiko in fiziko

Split
6/10/2022

What’s next for particle physics?

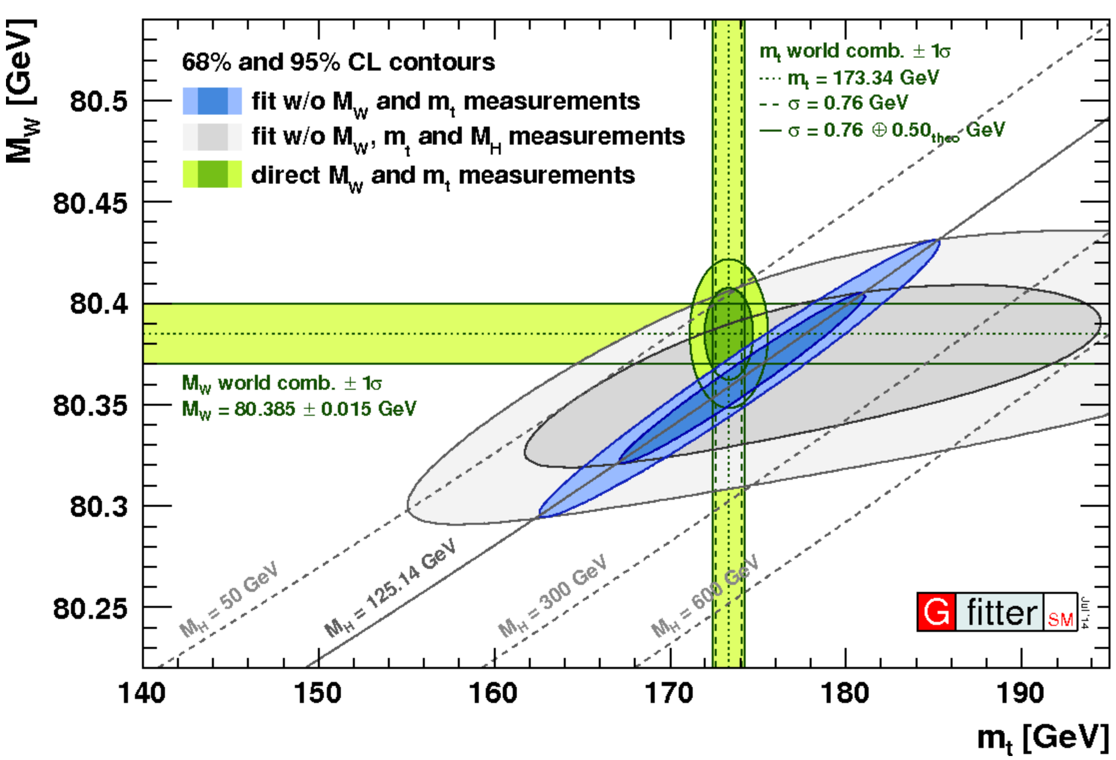
In 2012 last d.o.f predicted within SM - Higgs boson was discovered at LHC.

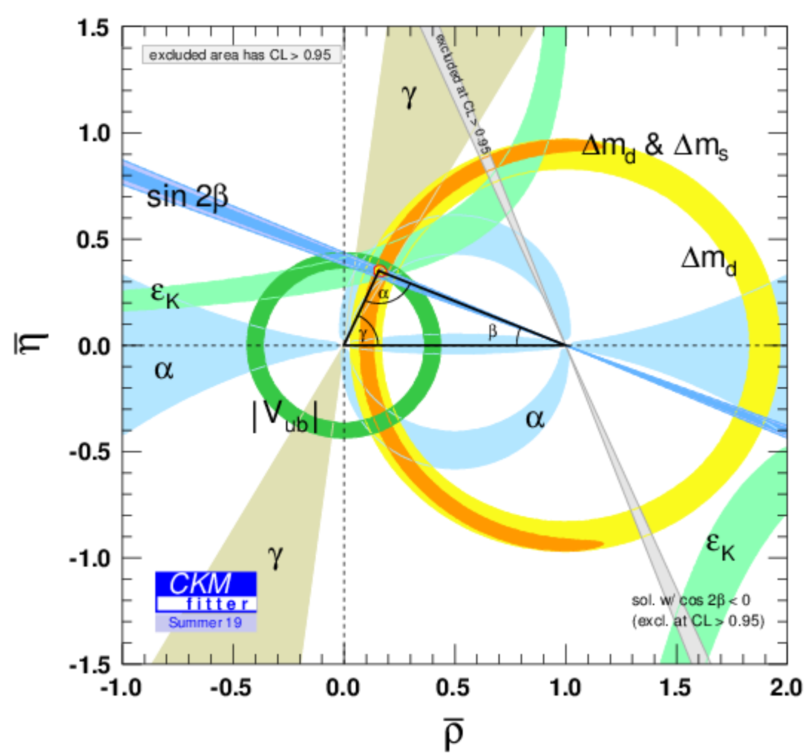


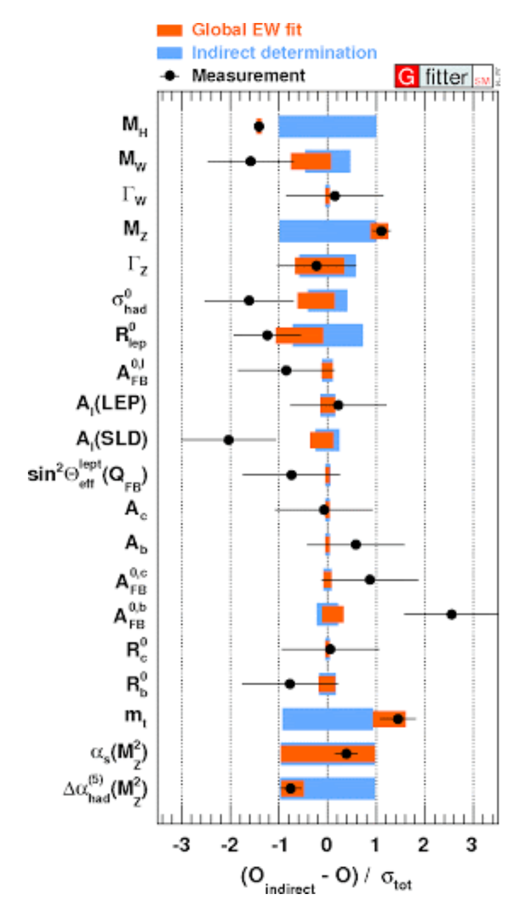
What’s next for particle physics?

SM immensely successful

- ✔ predictive up to very large energy scales $\Lambda \gg M_{\mathrm{Planck}} \simeq 10^{16}\,\mathrm{TeV}$
- ✔ all key predictions confirmed
- ✔ in excellent agreement with precision measurements in particle physics experiments over past 40+ years







What's next for particle physics?

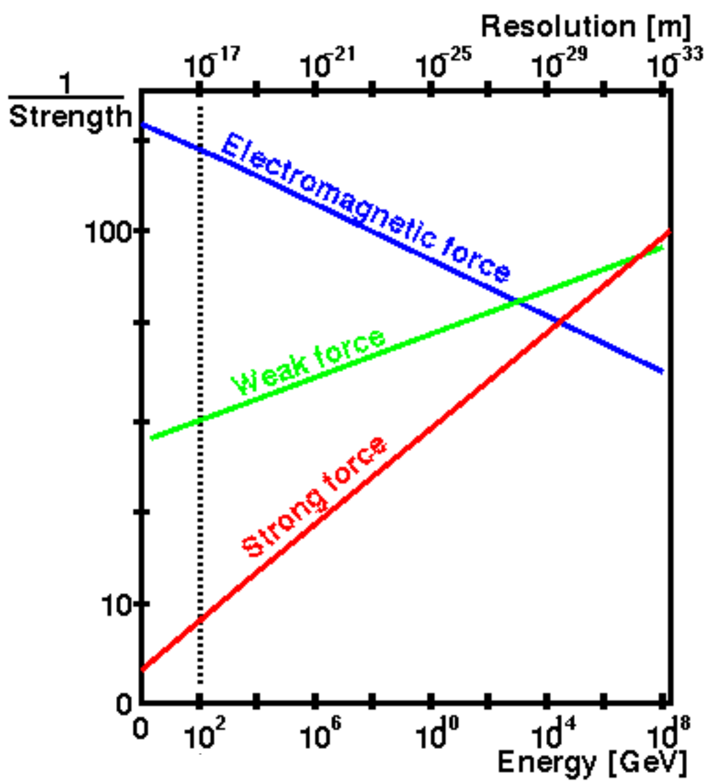


image source: J. Bondi

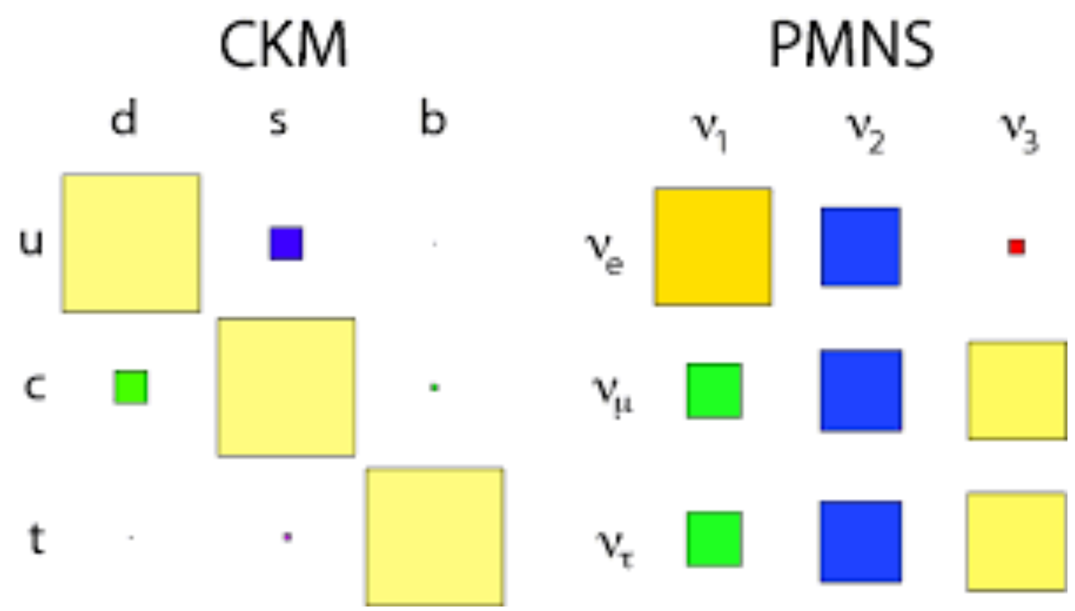
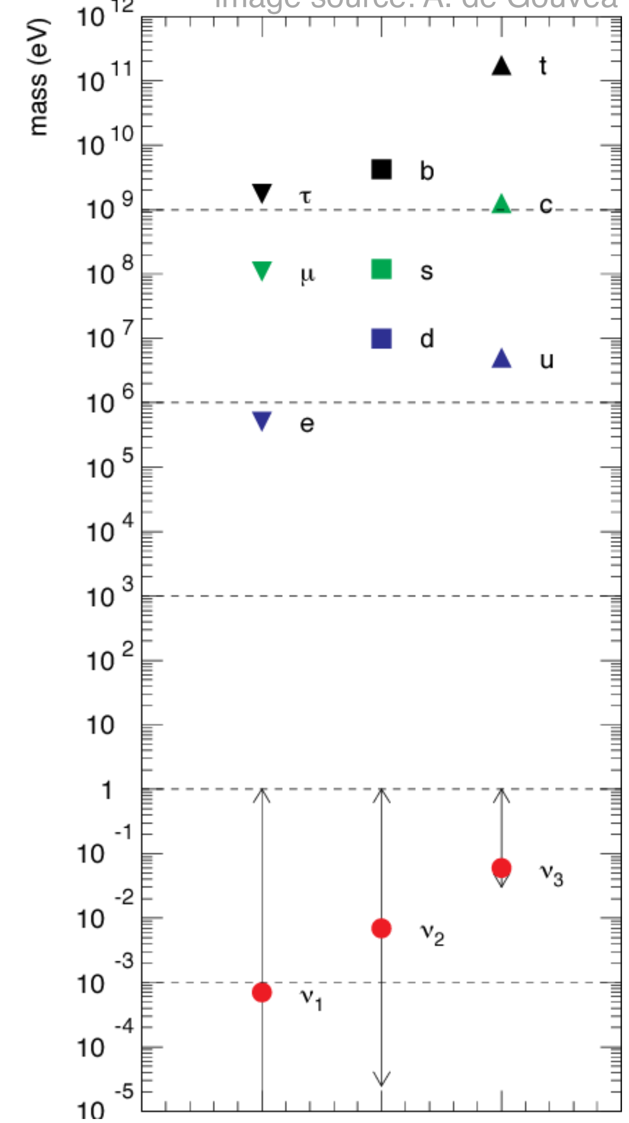


image source: J. A. Romeu

image source: A. de Gouvea



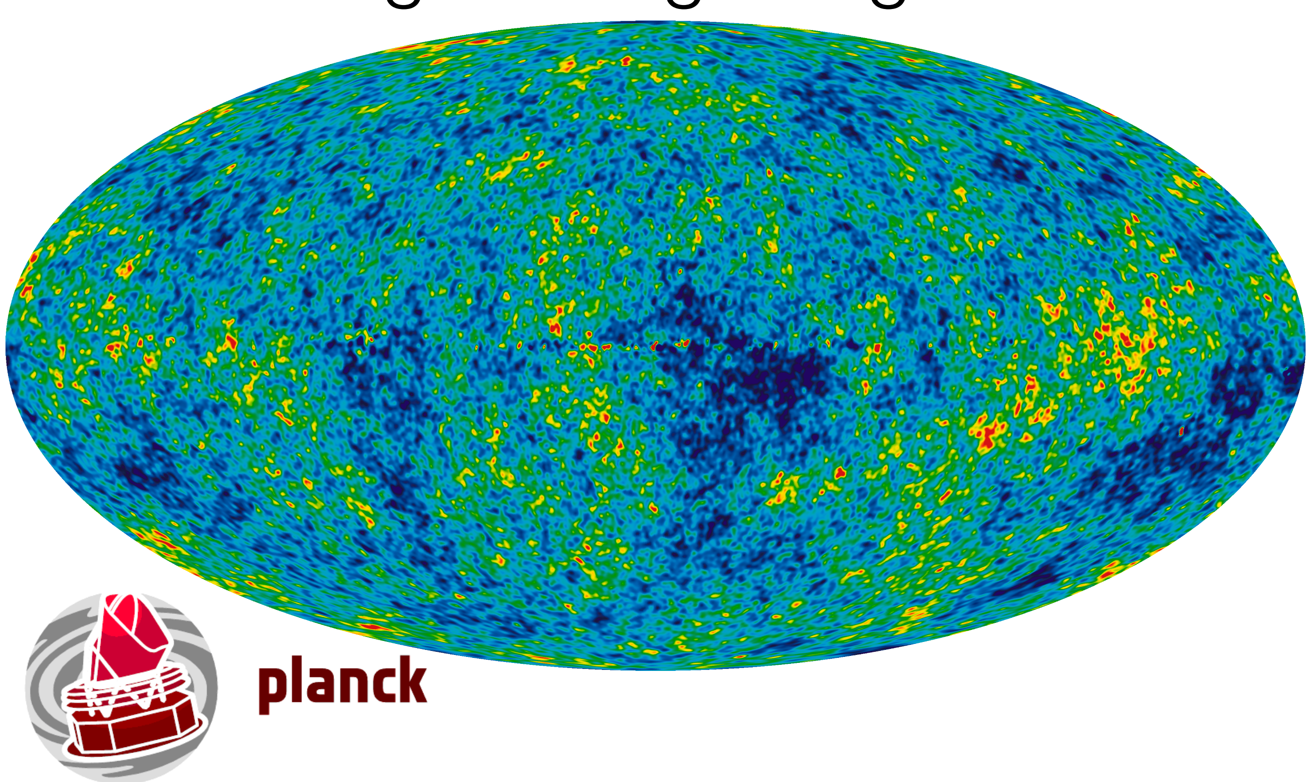
Several outstanding theoretical puzzles

- ✘ unification of fundamental forces (with gravity)
- ✘ observed patterns of parameters describing particle flavors
- ✘ observed hierarchies of mass scales in nature

What's next for particle physics?

Particle physics vs. Cosmology

- Origin of most of gravitational mass and energy in Universe
- Origin of matter-antimatter asymmetry in Universe
- Origin of Big Bang



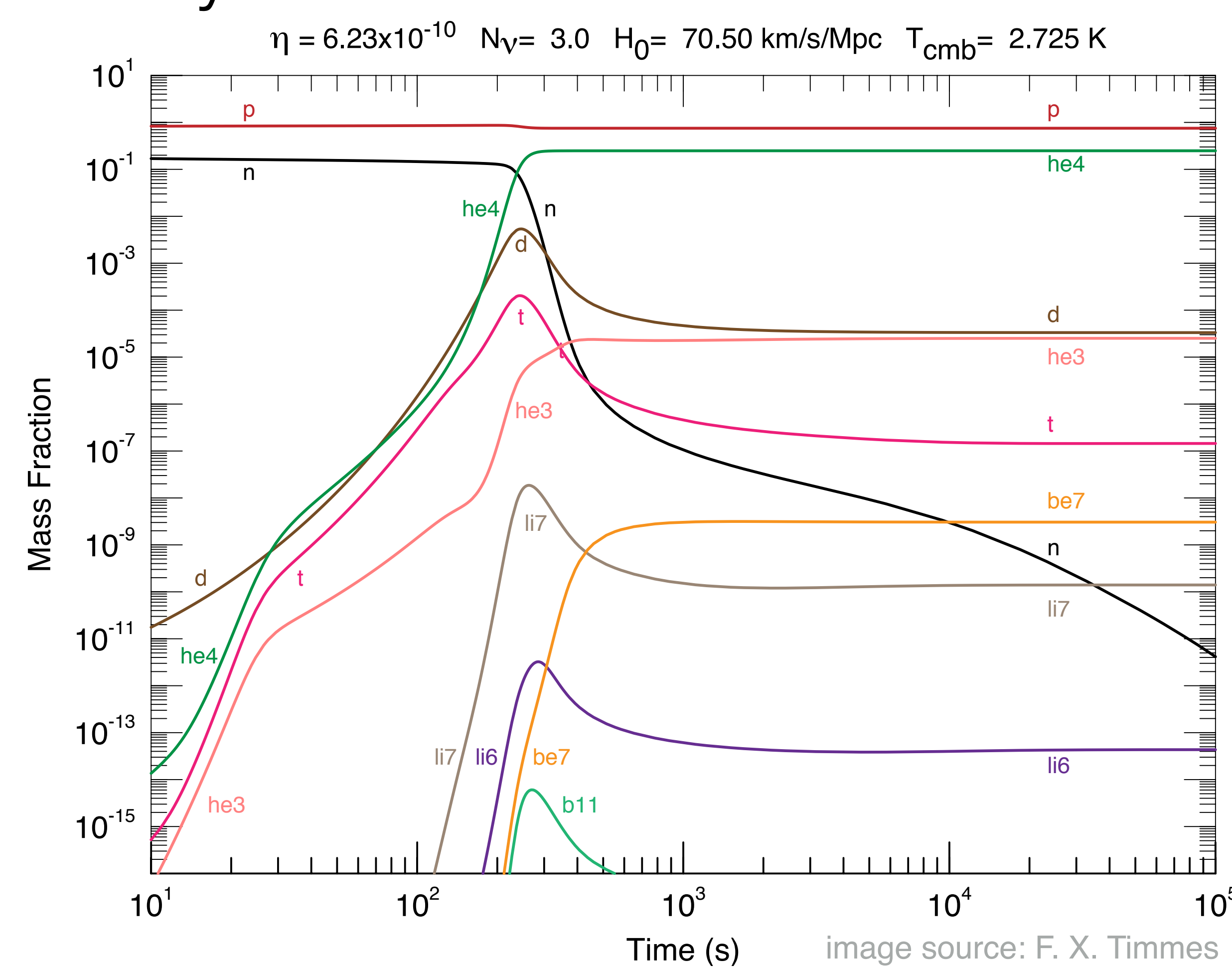


image source: F. X. Timmes

Searching for unknown physics BSM

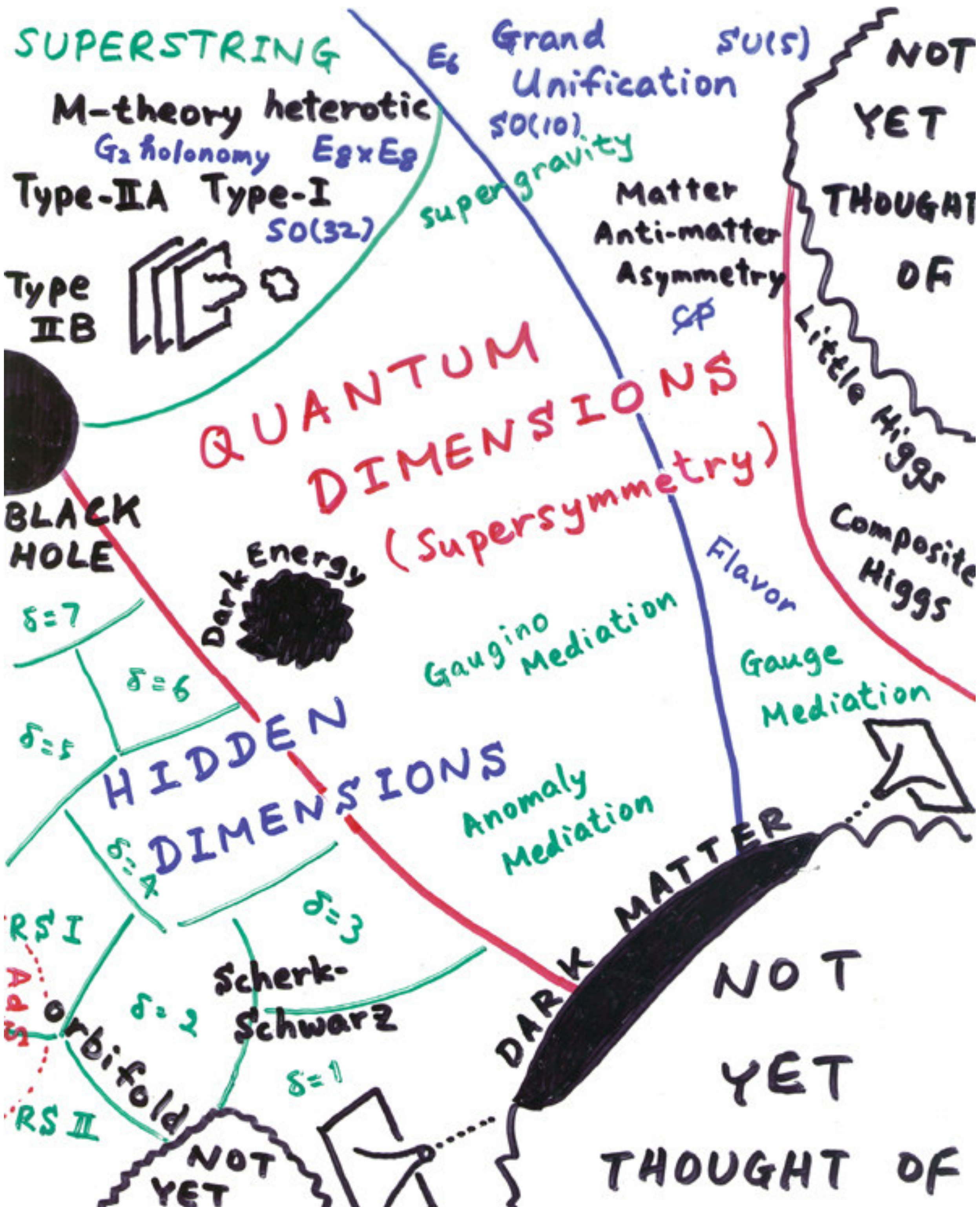


image source: H. Murayama

Theory

Multitude of theoretical proposals addressing SM shortcomings

Few unambiguous predictions experimentally accessible with current technology

Most prospective directions possibly not yet conceived

Searching for unknown physics BSM

Experiment

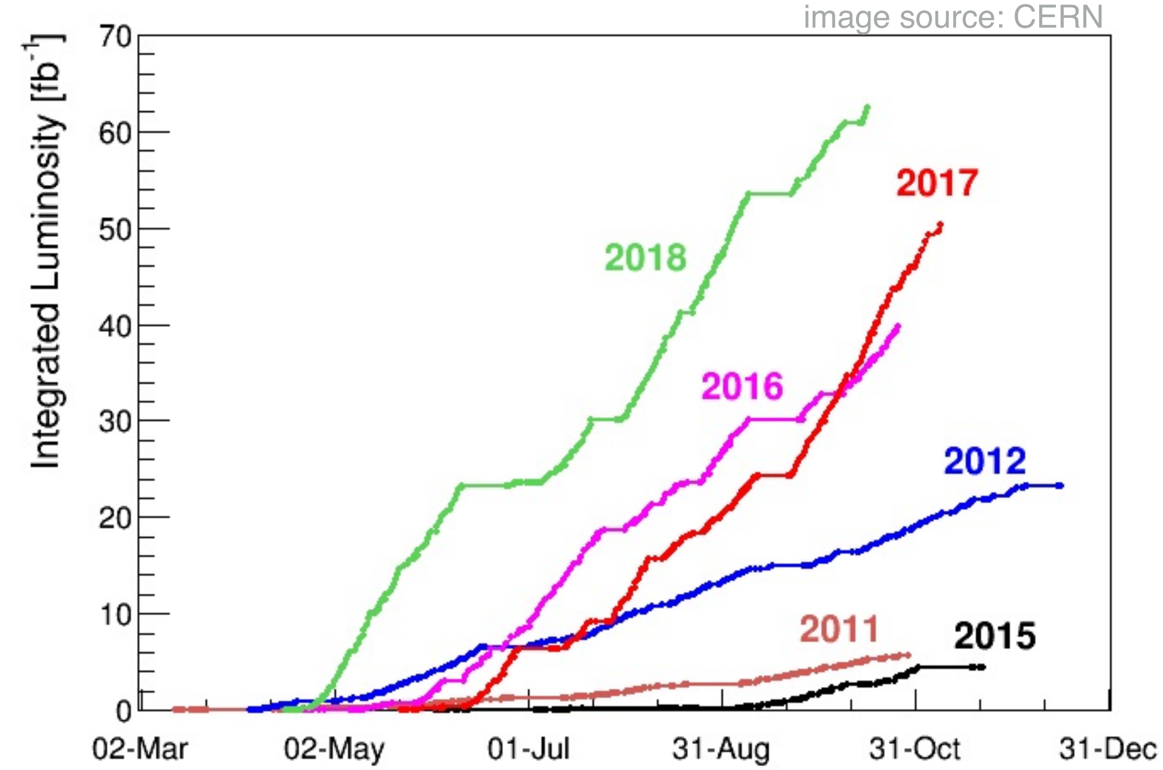
LHC produced abundance of data (experiments so far recorded ~200PB of most interesting events)

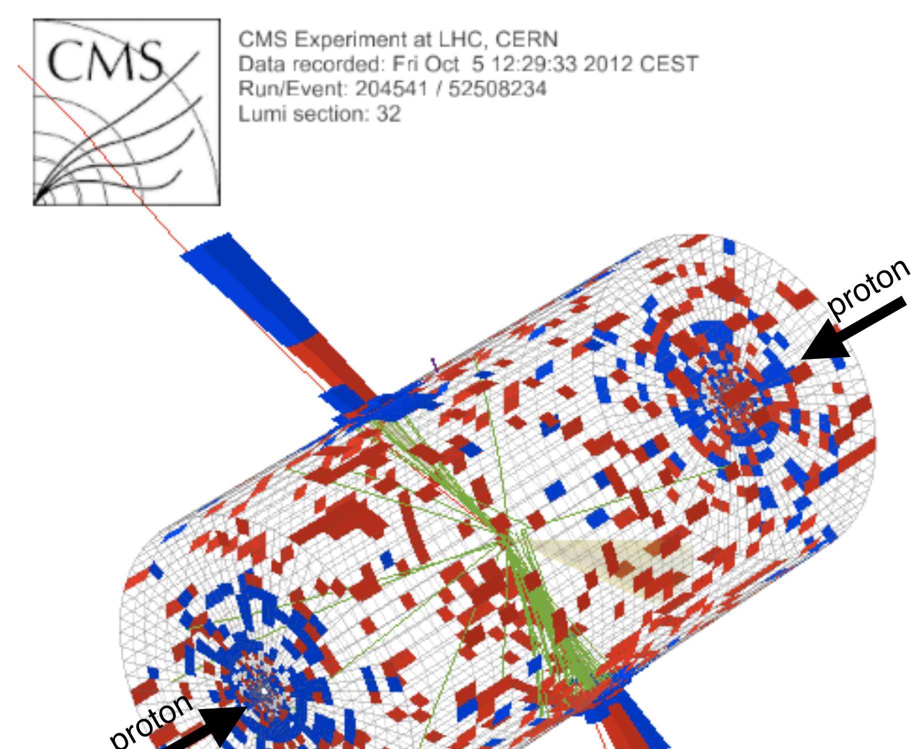
Expected to increase by order of magnitude in next decade

Challenge to discern interesting events from mundane backgrounds

Which events are interesting?

image source: CERN





Searching for unknown physics BSM

Discovering the Higgs boson was like searching for a needle in a haystack…

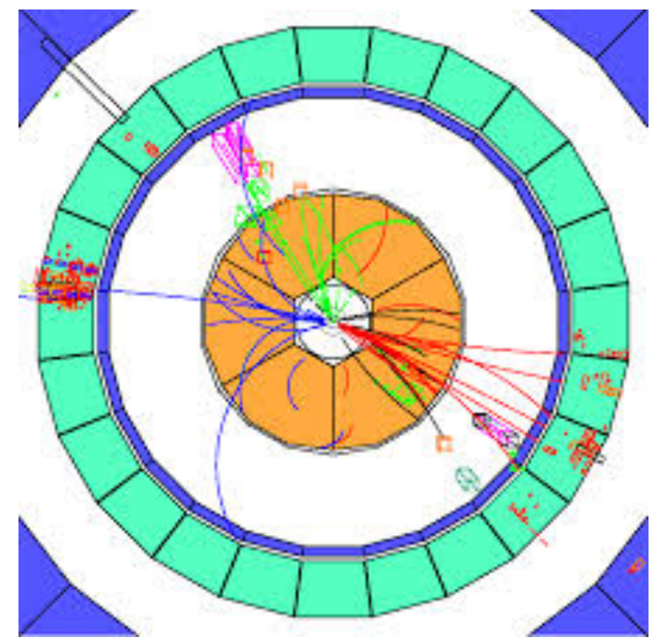
image source: J. Hill

…at least we knew how the needle looked like.

Machine Learning in Particle Physics

ML (subset of AI) uses statistical learning algorithms to build models based on data

ML has been used in HEP, particularly in experimental applications since late 20th century



Track finding with neural networks
Carsten Peterson
Nuclear Instruments and Methods in Physics Research Section A (1989) 279 (3): 537-5459.

The Use of Neural Networks in High-Energy Physics
Bruce Denby
Neural Computation (1993) 5 (4): 505–549.

In past 10-15 years, exploration of ‘deep learning’ approaches in HEP closely follows advances in algorithms & available computing power

Machine Learning in (Particle) Physics

ML plays increasingly important role in HEP including in

⇒ Data reconstruction and analysis
(particle flow, object reconstruction & classification)

see e.g. Pata et al., 2101.08578
Kasieczka et al., 1902.09914
Brehmer et al., 1907.10621
…

⇒ First-principles theory calculations & detector simulations
(MC event generation, Lattice simulations,…)

see e.g. Boyda et al., 2202.05838
Butter et al., 2203.07460
Albergo et al., 2101.08176
…

⇒ Detector and Accelerator design and operation
(Differentiable detectors, ML triggers, Defect detection)

see e.g. Dorigo, 2203.13818
Govorkova et al.,2108.03986
Akchurin et al., 2203.08969

⇒ Anomaly Detection for BSM physics searches
(rest of this talk…)

Anomaly detection in (Particle) Physics

⇒ Common goal: to classify among different (un)known physical processes ⟺ to learn/approximate likelihoods

Crucial to understand physics learned by the machine

see e.g. Faucett, Thaler & Whiteson, 2010.11998

⇒ Helps to understand systematics & validate assumptions (i.e. MC, control region dependence)

Challenge of uncovering and characterizing possible unexpected signals in (LHC) data.

see e.g. Kasieczka et al., 2101.08320

⇒ Need to identify signal regions, construct null-hypothesis tests, mitigate potentially large look elsewhere effects…

{

Supervised ML
a.k.a. Universal Function Approximation

Input: representation of model $p(x)$, finite number of examples $\{x_i\}$ sampled/computed/generated from p

Output: mapping $f(x \to z)$ minimizing a loss function $\mathcal{L}(f, \{x_i\})$

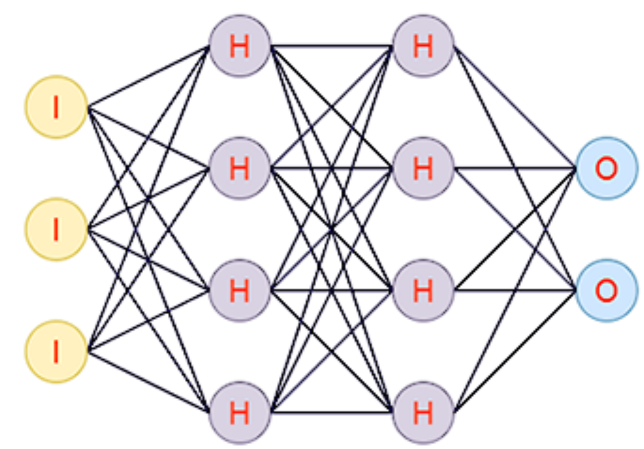
see e.g. B. Nachman, 1909.03081

Common example: model of two distributions $p_s(x)\,,\;\; s = 0\,,\,1$
with loss function $\mathcal{L} = -[s_i \log(z_i) + (1 - s_i)\log(1 - z_i)]$

(cross-entropy)

$\Rightarrow f$ will approximate $f(x) \sim \dfrac{p_0(x)/p_1(x)}{1 + p_0(x)/p_1(x)}$ (likelihood ratio)

Typical implementation in terms of (deep) neural networks



Supervised ML
a.k.a. Universal Function Approximation

Example: distinguishing boosted massive resonances from QCD jets

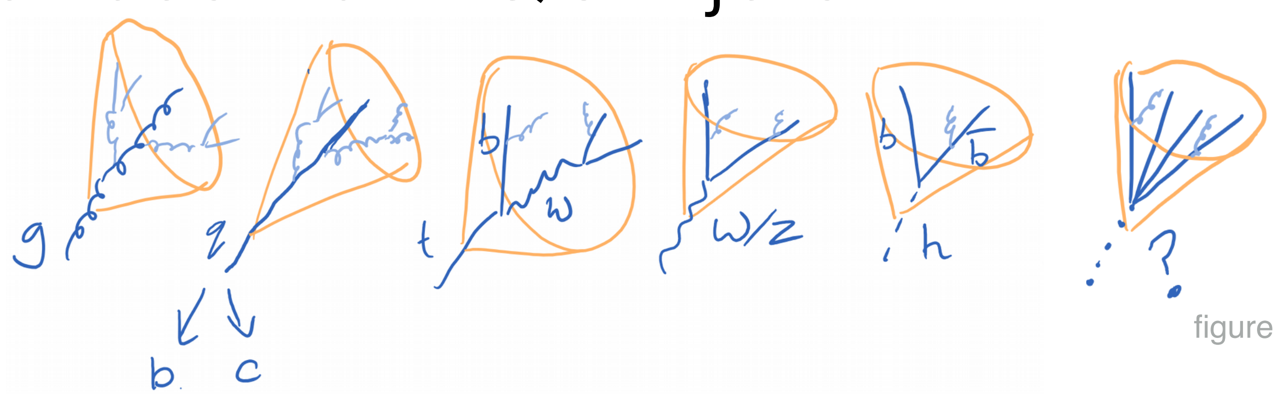
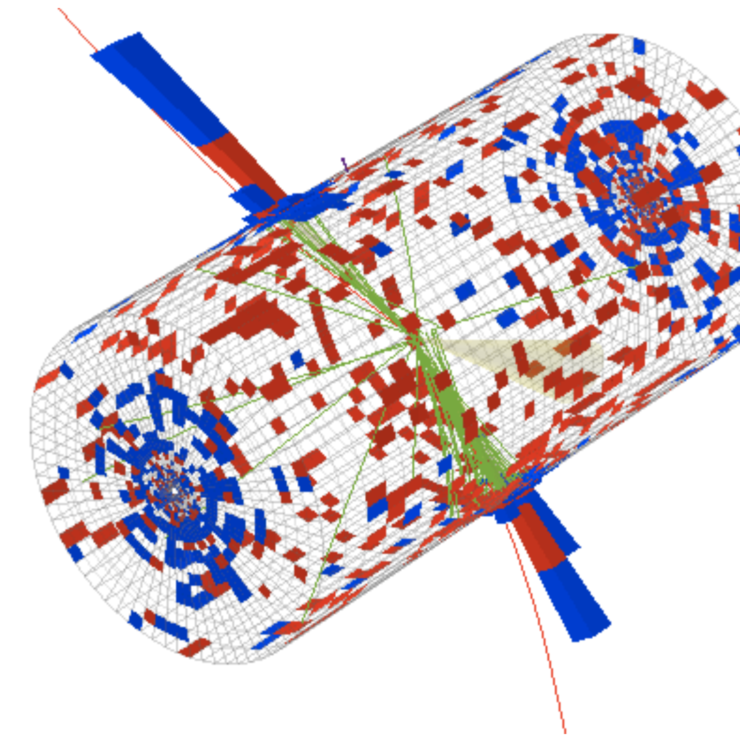


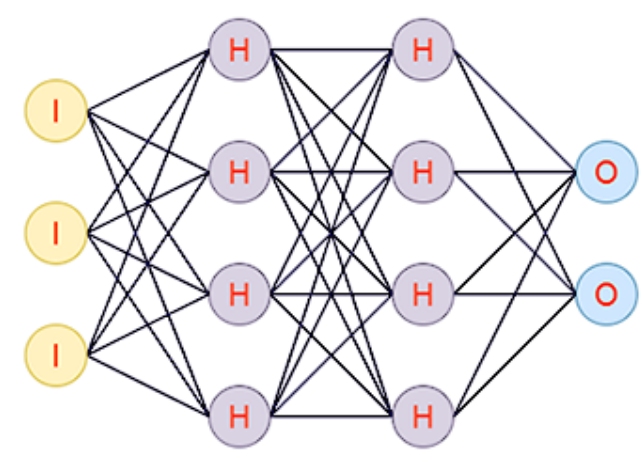
figure by J. Collins



Input (x)*:* detector readings

(particle tracks, calorimeter clusters)

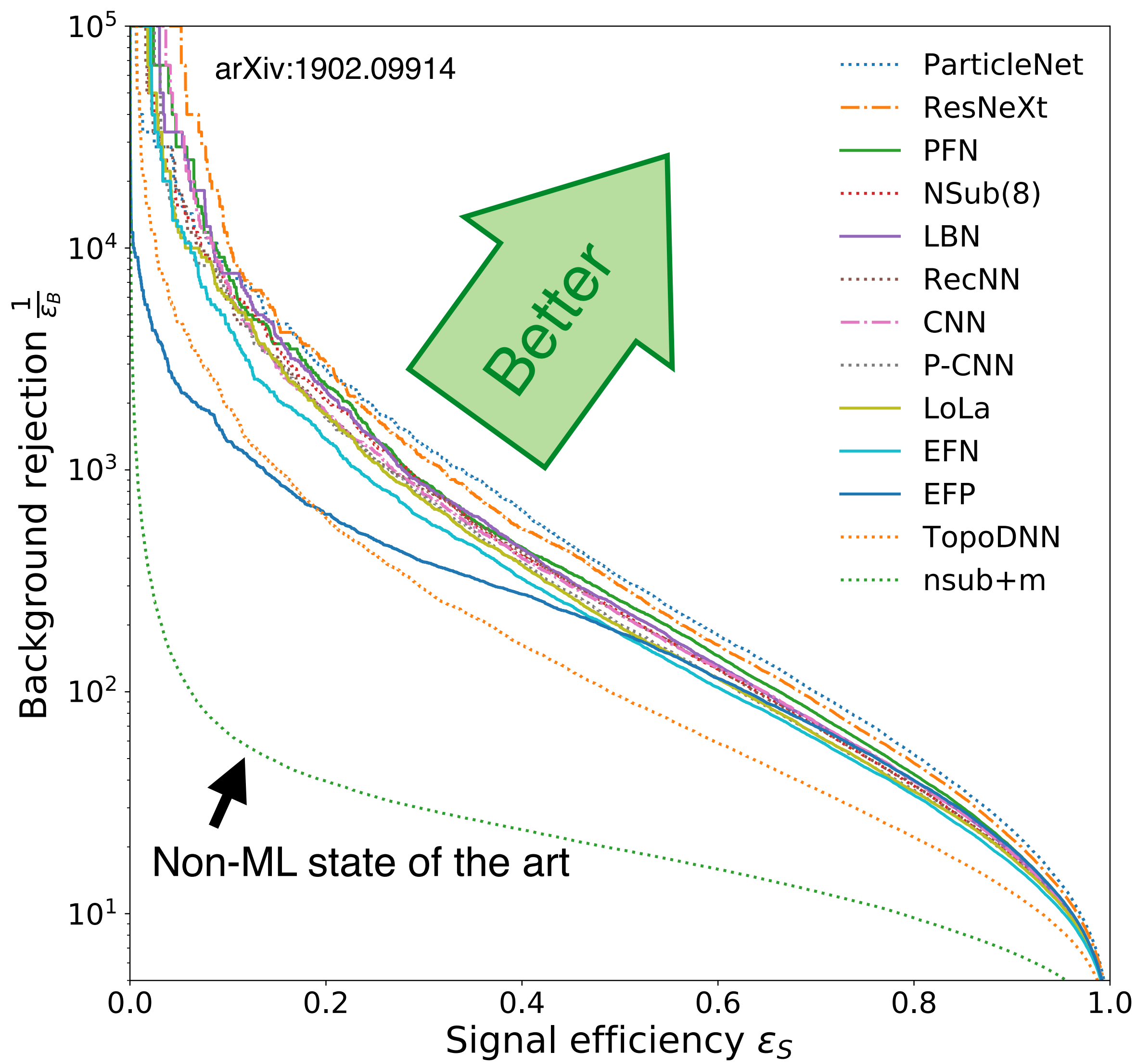
Method: DeepNN trained on artificial (MC) or pre-tagged (labelled) samples $\{x_i\}$

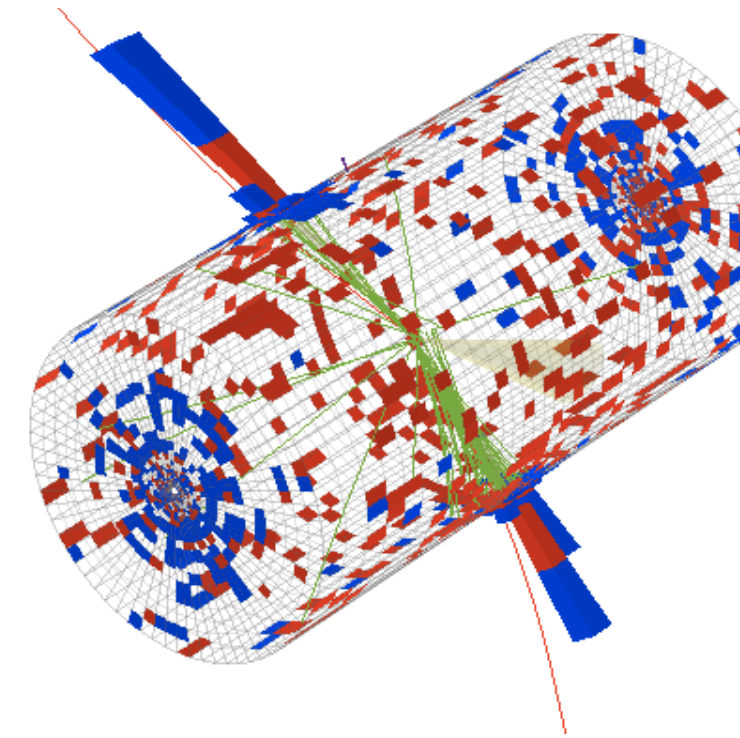


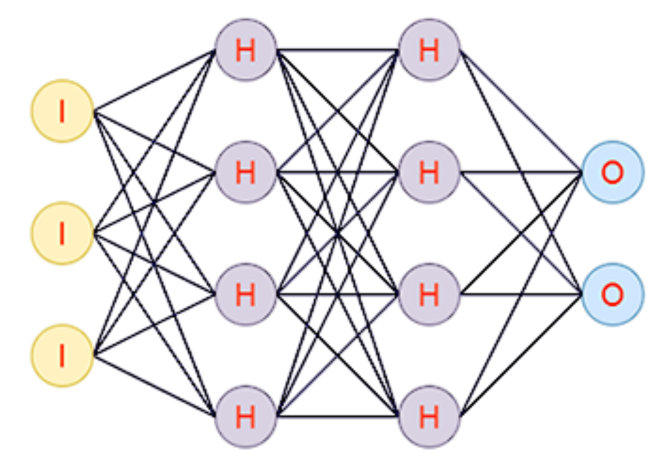
Output: parametric classifier (f) with some (ROC) performance curve

Supervised ML
a.k.a. Universal Function Approximation

Example: QCD vs. t-quark (top-tagging)







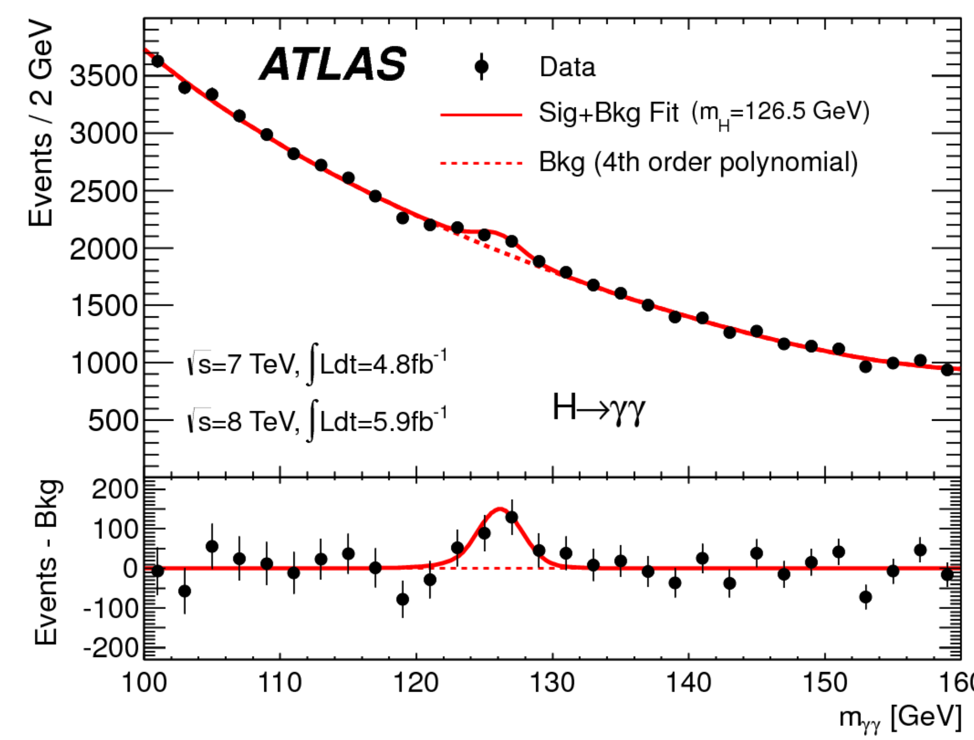
Output: parametric classifier (f) with some (ROC) performance curve

}

Challenges of ML Applications to BSM

Signal model $p_S(x)$ may not be known, unknown number of unlabelled examples $\{x_i\}$ sampled from p_S *may* be present *somewhere* in data.

Background model $p_B(x)$ (SM) known imperfectly, cannot be relied upon to subtract from data.



However, often reasonable to assume signal is quasi-localized (unevenly distributed) in given dataset (e.g. resonant).

⇒ "Weakly- & Unsupervised" ML

Weakly-Supervised ML example

Metodiev, Nachman & Thaler, 1708.02949

see also Nachman & Shih, 2001.04990

...

Classification from mixed samples: pure (signal, background) samples not available in real data

$$p_{M_1}(\vec{x}) = f_1\, p_S(\vec{x}) + (1 - f_1)\, p_B(\vec{x}),$$
$$p_{M_2}(\vec{x}) = f_2\, p_S(\vec{x}) + (1 - f_2)\, p_B(\vec{x}),$$

1.) Assume f_1, f_2 known (e.g. from MC), then simply

$$h^{M_1/M_2}_{\text{optimal}}(\vec{x}) = p_{M_1}(\vec{x})/p_{M_2}(\vec{x})$$

2.) Assume only $f_1 > f_2$ then use monotonicity of

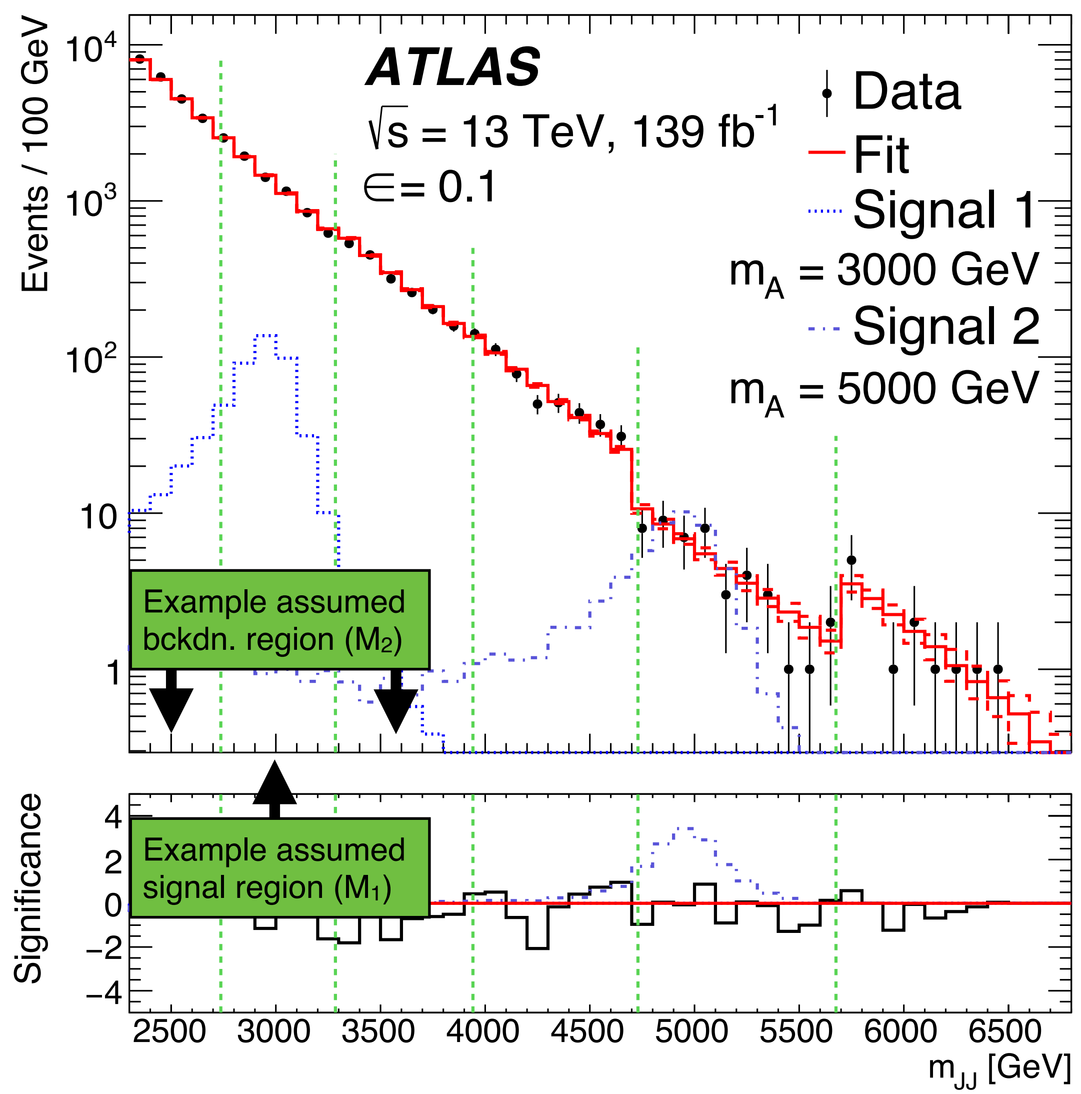
$$\frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1 - f_1)\, p_B}{f_2\, p_S + (1 - f_2)\, p_B}$$

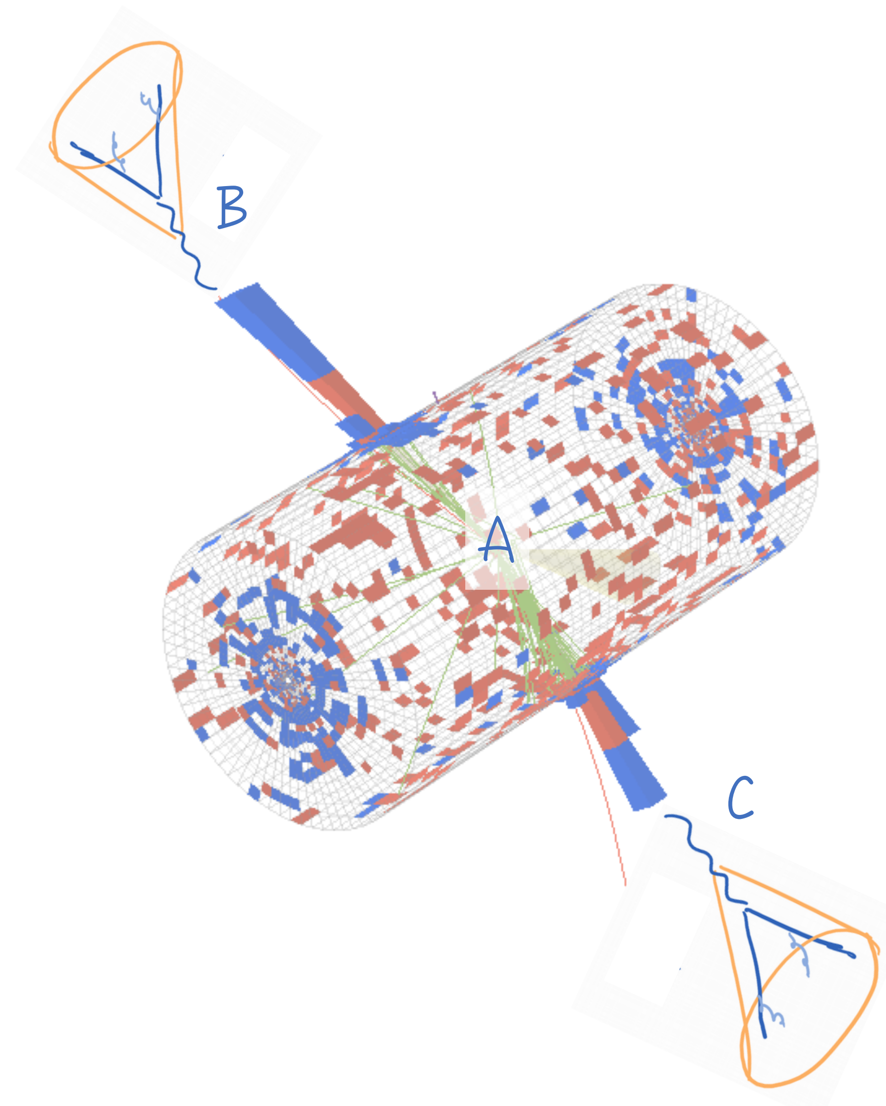
Train (NN) classifier only to distinguish two mixed samples!

Weakly-Supervised ML example

ATLAS, 2005.02983

Example: Model agnostic BSM search in di-jet events



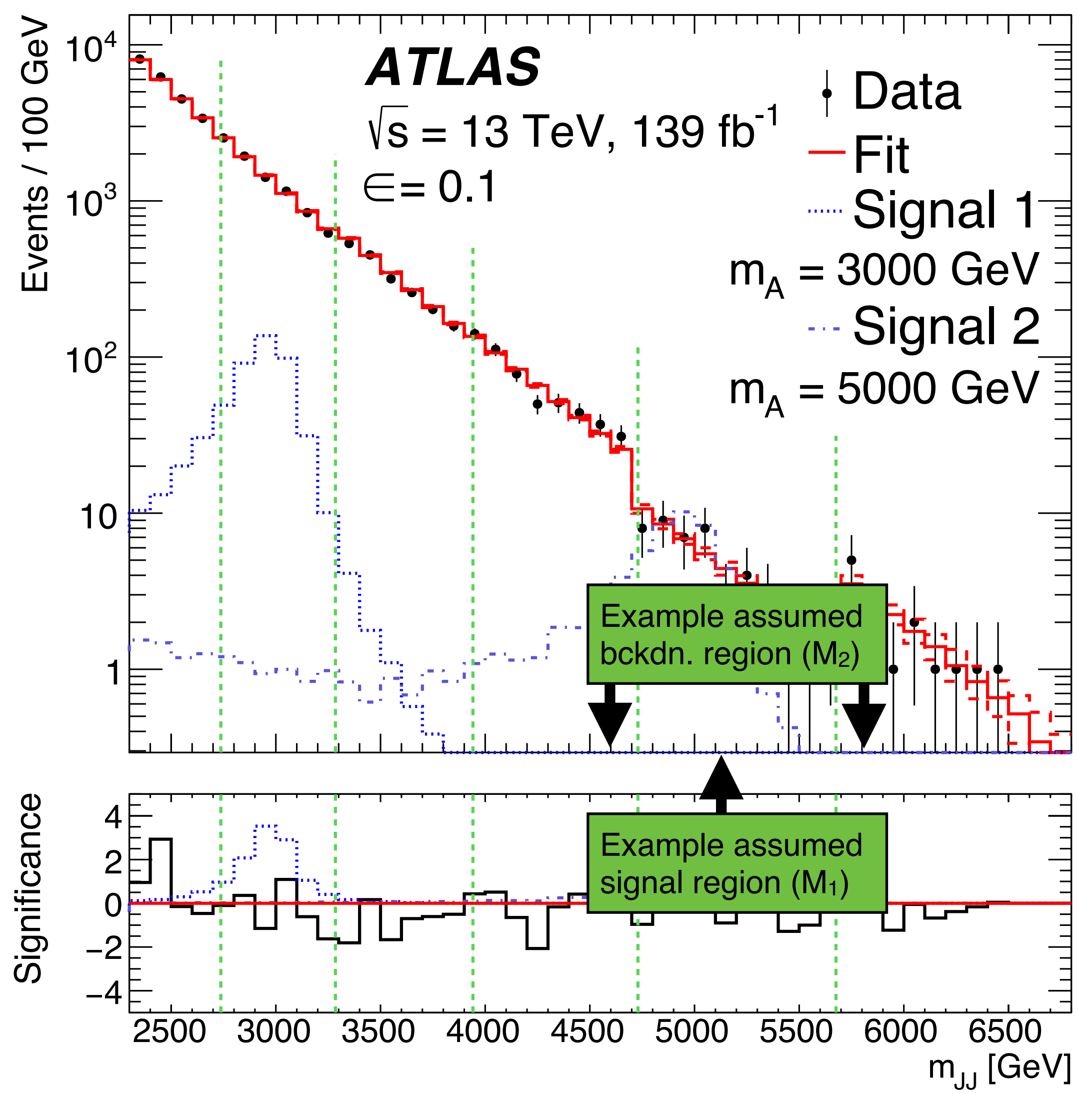


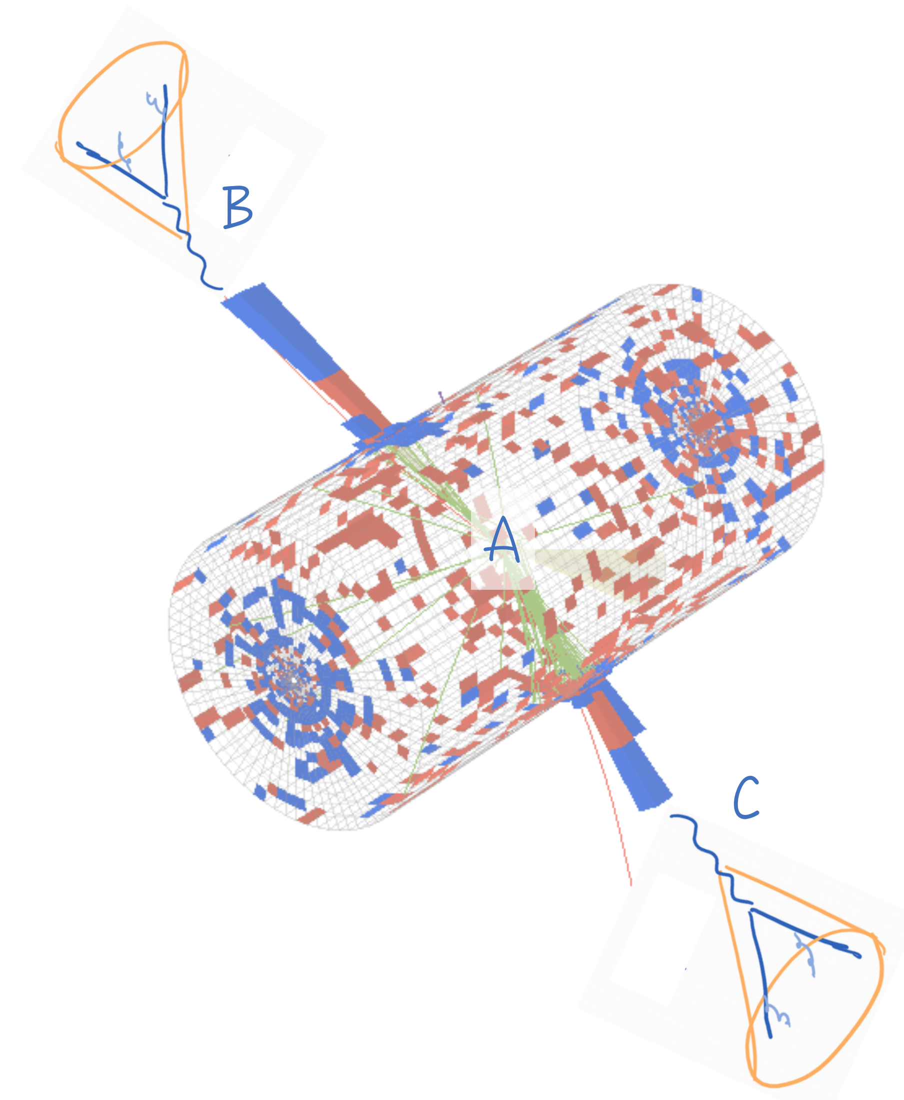
Train (NN) classifier only to distinguish two mixed samples!

Weakly-Supervised ML example

ATLAS, 2005.02983

Example: Model agnostic BSM search in di-jet events



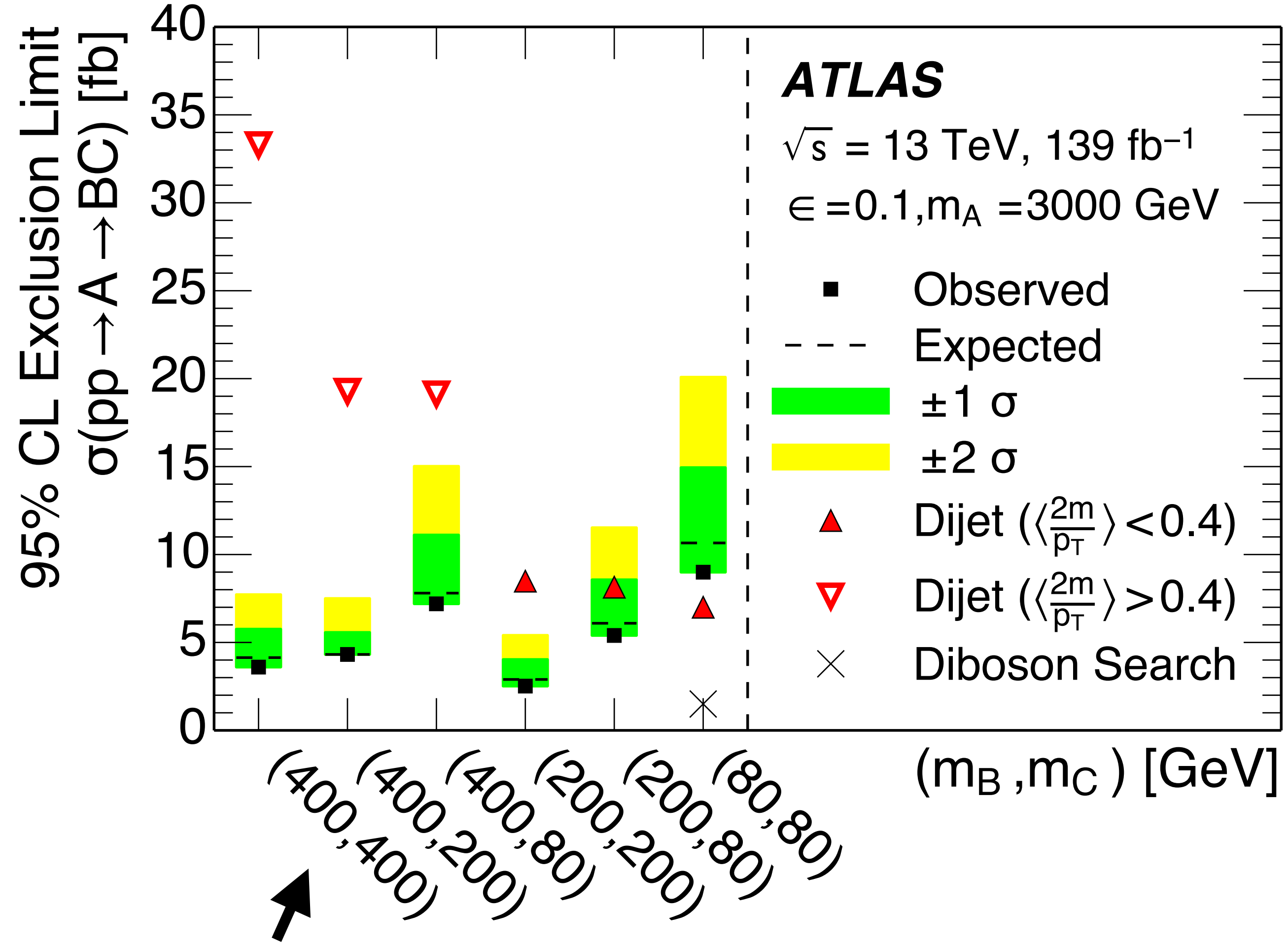


Train (NN) classifier only to distinguish two mixed samples!

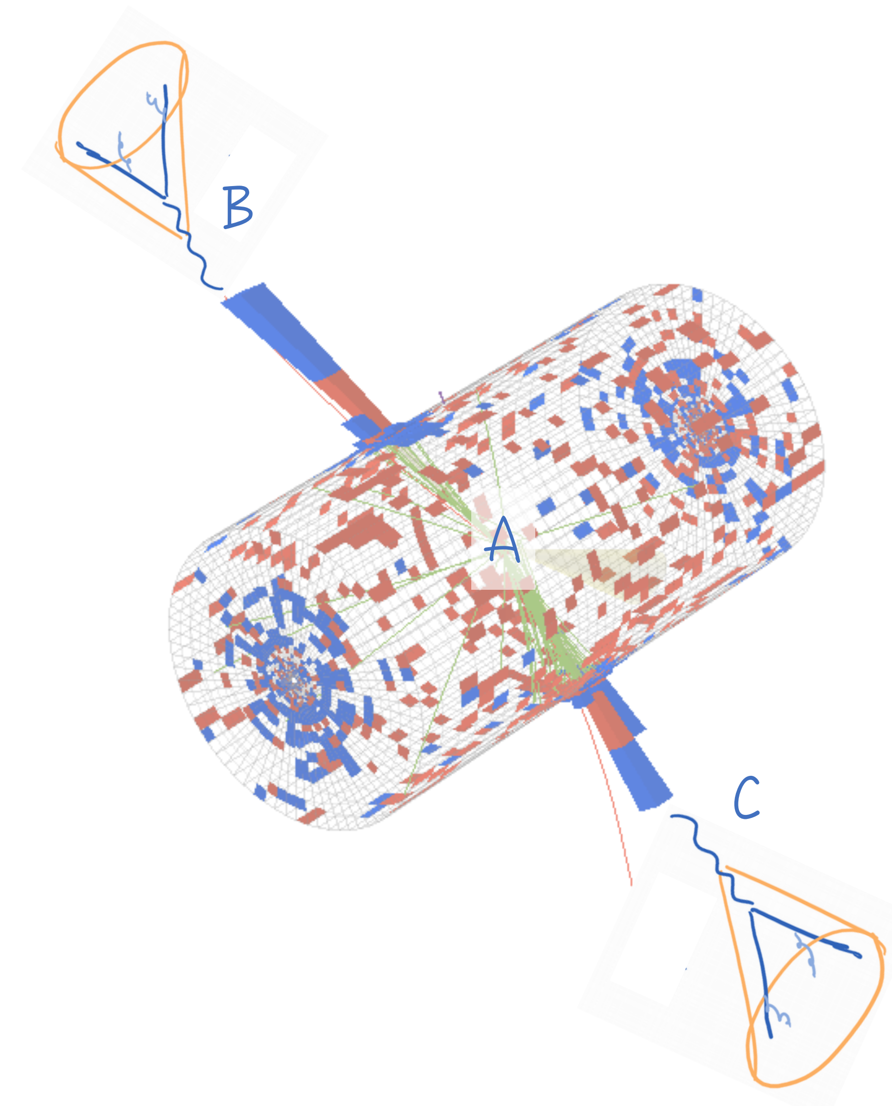
Weakly-Supervised ML example

ATLAS, 2005.02983

Example: Model agnostic BSM search in di-jet events



Example BSM models



Train (NN) classifier only to distinguish two mixed samples!

Challenges of (weakly supervised) ML for BSM

Scanning over possible signal-rich regions (M_1) can accumulate large trails factors (look-elsewhere effect).

see e.g. Bayer, Seljak & Robnik, 2108.06333

Crucial de-correlation of features $\{x_i\}$ and scanning variable (e.g. di-jet invariant mass M_{jj})

see e.g. Benkendorfer, Le Pottier & Nachman, 2009.02205

Assuming weakly supervised classifier s uncovers localized excess in data…

What is the physics contained in $s(x)$?

see e.g. Bortolato et al., 2103.06595
Dillon et al., 1904.04200, 2005.12319

How is it sensitive to biases & systematics of $\{x_i\}$?

see e.g. Nachman, 1909.03081
Gosh & Nachman, 2109.08159

How to quantify the significance of (lack of) detection ?

J. F. K. & Szewc, 2210.02226

Null-hypothesis test for Anomaly detection

J. F. K. & Szewc, 2210.02226

In essence classification from mixed samples

$$p_{M_1}(\vec{x}) = f_1\, p_S(\vec{x}) + (1-f_1)\, p_B(\vec{x}),$$
$$p_{M_2}(\vec{x}) = f_2\, p_S(\vec{x}) + (1-f_2)\, p_B(\vec{x}),$$

defines a mixture model assuming conditional independence: $p(s(\vec{x}), y|\pi) = (1-\pi)\, p(s(\vec{x})|B)p(y|B) + \pi\, p(s(\vec{x})|S)p(y|S)$

Classifier output = anomaly score ~*S*/*B* scanning variable defining $M_{1,2}$

⇒ Null-hypothesis test = statistical independence of s & y

$$p(s(\vec{x}), y|\pi = 0) \qquad I(s,y|z) = \int ds\, dy\; p(s,y|z) \log \frac{p(s,y|z)}{p(s|z)p(y|z)} = 0$$

⇒ Clear statement on the presence of signal (*p*-value)

⇒ No fixed anomaly score cuts or *B* extrapolations!

Conclusions

Machine learning methods have long history of applications in particle physics

Today part of standard toolbox of HEP data analysis

Example: Higgs boson discovery and characterization

Open-ended nature of BSM physics searches presents novel challenges ⇒ opportunities for ML

see e.g. Shanahan et al., 2209.07559
Carleo et al., 1903.10563

Recent exciting progress in weakly- and un-supervised ML approaches to BSM physics
(only one direction covered today)

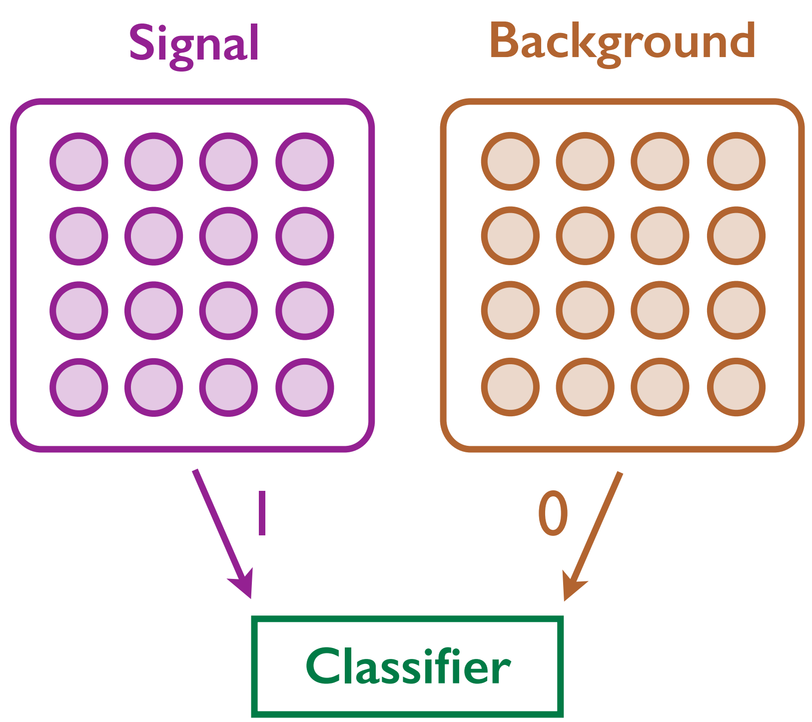

Extras

Jet classification: basics

x list of observables useful for distinguishing S from B

$p_S(x)$ and $p_B(x)$ - probability distributions of x for S and B

classifier $h(x)$ close to 1 for S and close to 0 for B - to be learned by minimizing loss function (e.g mean-square)



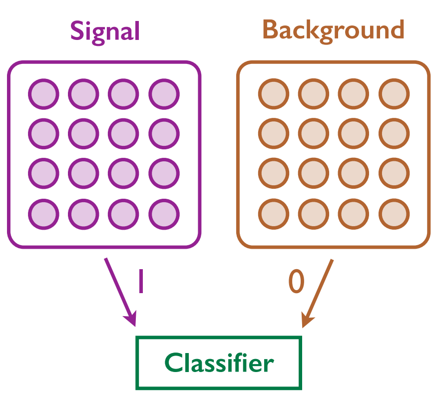
Jet classification: basics

x list of observables useful for distinguishing S from B

$p_S(x)$ and $p_B(x)$ - probability distributions of x for S and B

classifier $h(x)$ close to 1 for S and close to 0 for B - to be learned by minimizing loss function (e.g mean-square)

receiver operating characteristic (ROC) curve



$$\epsilon_S = \int \mathrm{d}\vec{x}\, p_S(\vec{x})\, \Theta(h(\vec{x}) - c)$$
$$\epsilon_B = \int \mathrm{d}\vec{x}\, p_B(\vec{x})\, \Theta(h(\vec{x}) - c)$$

Neyman-Pearson lemma: $h_{\text{optimal}}(\vec{x}) = p_S(\vec{x})/p_B(\vec{x})$ (likelihood ratio)

If x - low dimensional, can use histograms directly, otherwise use supervised ML (BDTs, NNs, …)

Jet classification: basics

x list of c

$p_S(x)$ anc

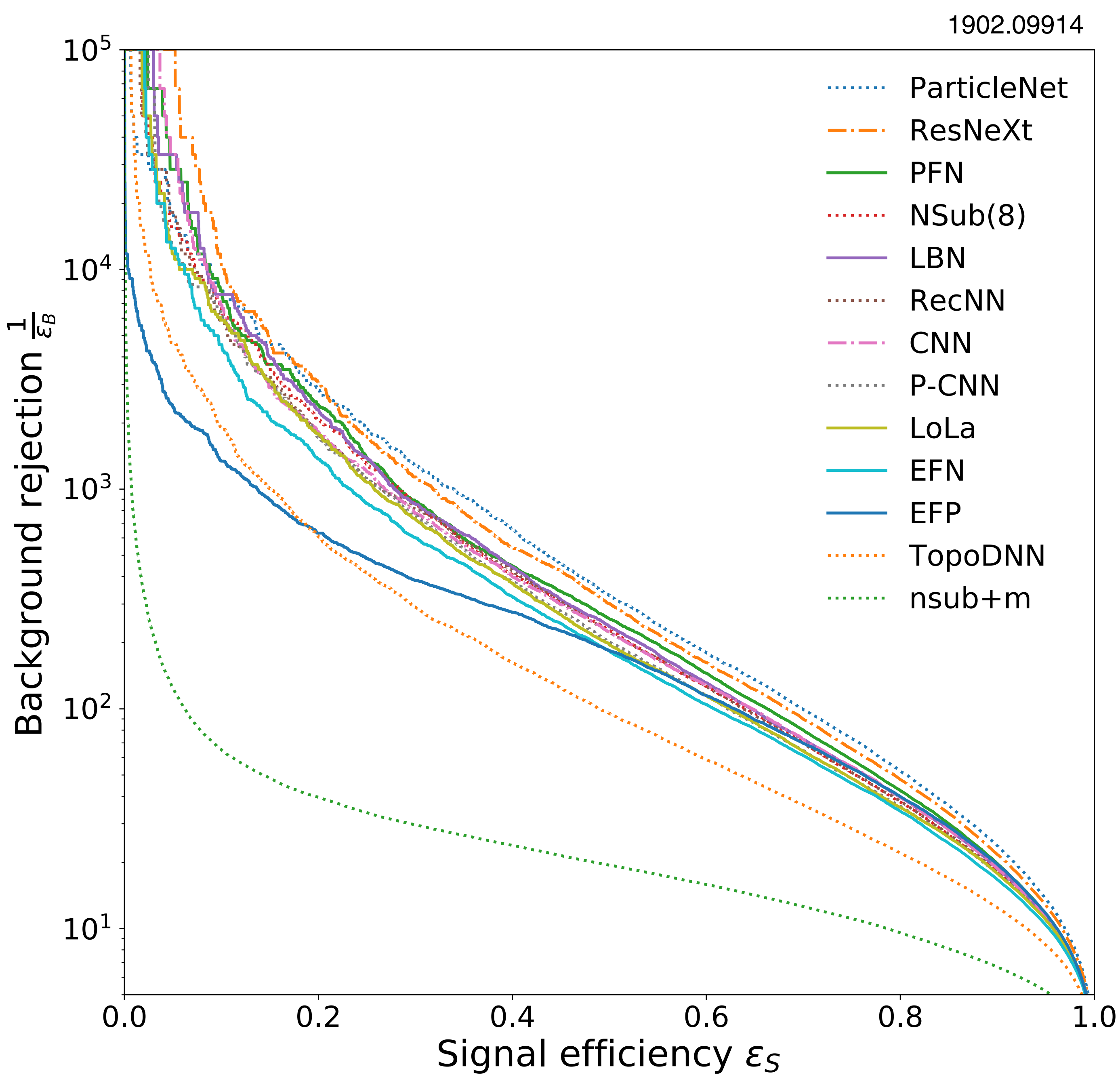
classifier
learned b

receiver

Neyman-

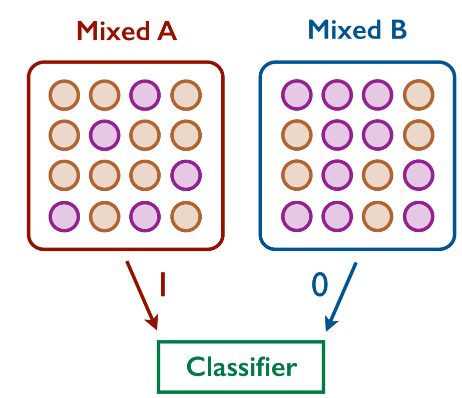
If x - low
otherwis

Example: QCD j vs. t classification (top-tagging)



Jet classification: mixed samples

Classification from mixed samples: pure samples not available in real data



$$p_{M_1}(\vec{x}) = f_1\, p_S(\vec{x}) + (1-f_1)\, p_B(\vec{x}),$$
$$p_{M_2}(\vec{x}) = f_2\, p_S(\vec{x}) + (1-f_2)\, p_B(\vec{x}),$$

1.) Assume f_1, f_2 known (e.g. from MC), then simply

$$h^{M_1/M_2}_{\text{optimal}}(\vec{x}) = p_{M_1}(\vec{x})/p_{M_2}(\vec{x})$$

2.) Assume only $f_1 > f_2$ then use monotonicity of

$$\frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1-f_1)\, p_B}{f_2\, p_S + (1-f_2)\, p_B}$$

(Classification Without Labels)

Metodiev, Nachman & Thaler, 1708.02949