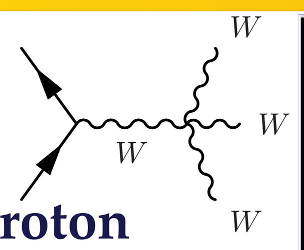
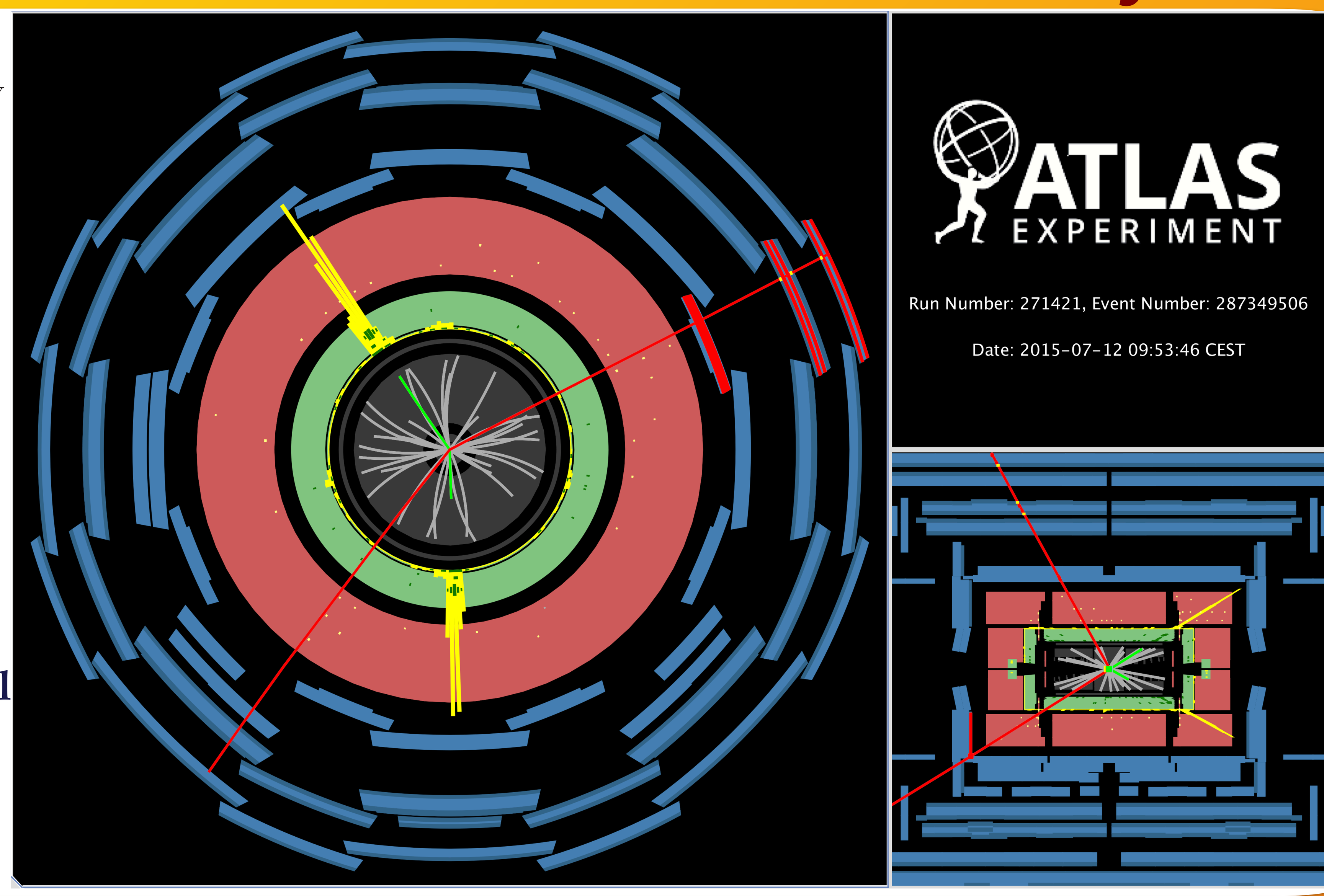
Signal Region Identification with Applications to the ATLAS W±W±W∓ Analysis

Introduction



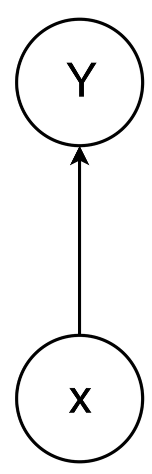
ATLAS is general purpose detector designed to study **proton-proton collisions** at the CERN Large Hadron Collider (LHC) in Switzerland.
Collisions happen every 25 ns: every collision is an example of dataset.
All the examples are **independent** from the other.
In each collision, a certain physical process happens with a **random probability**: the low probability ones are more interesting.
Layers of varying detector technology are used to measure the properties (**charge, mass, momentum, type**) of particles produced in each collision.
These properties are the features X. An example is shown on the right.

This study specifically targets applying machine learning methods in **signal and background separation:**
The signal is a certain rare physical process (W±W±W∓) that could have similar recorded features with the background processes.



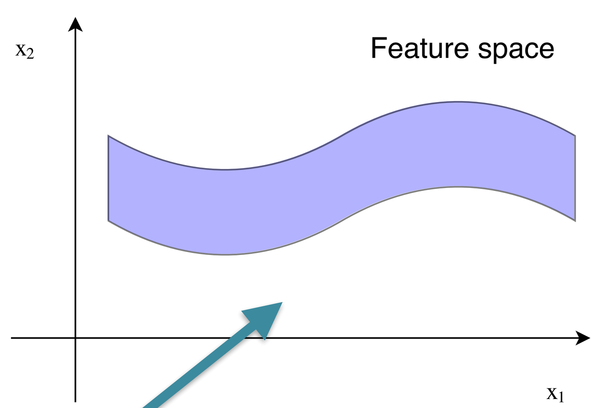
The probabilistic model

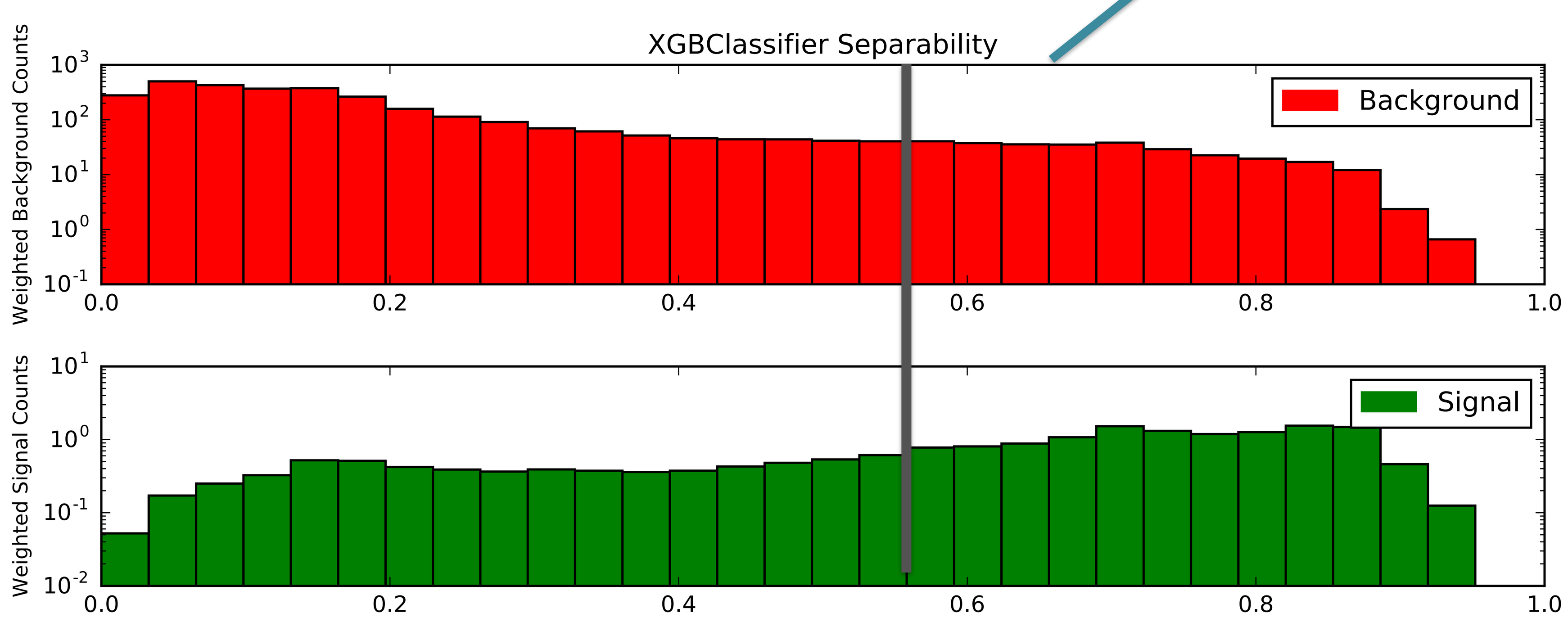
The graphical model is a **simple logistic regression**:
given the features X: physical quantities measured by the detector: energy, angular distribution,
the model predicts the label distribution Y: signal or background.



Three different architectures are tested:
- simple logistic regression (one layer NN)
- deep neural network (three-layer NN)
- decision trees (XGBoost)

Output of the classifier is used for distinguish signal and background in the feature space.
Example from XGBoost is shown below.



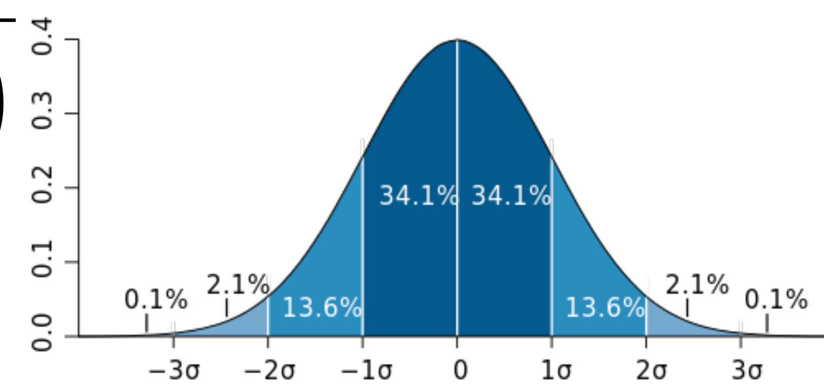


The physical metric: significance

A well established metric in physics is the **significance** of a signal with respect to the background.
Given a region in the feature space, the number of signal events (S) and background events (B) are counted; thus Z (significance):

$$Z = \sqrt{2\left(\left(s+b\right)\log\left(1+\frac{s}{b}\right)-s\right)}$$

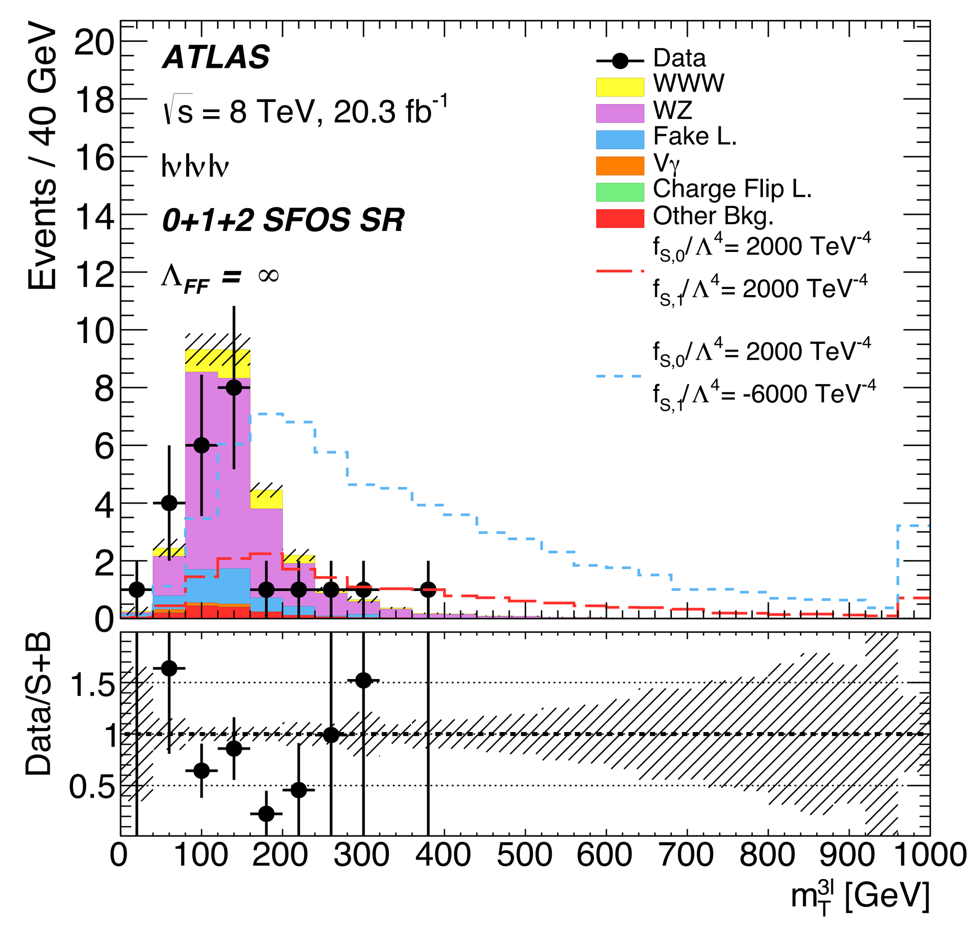
This is further interpreted in **standard deviations** of gaussian distribution:



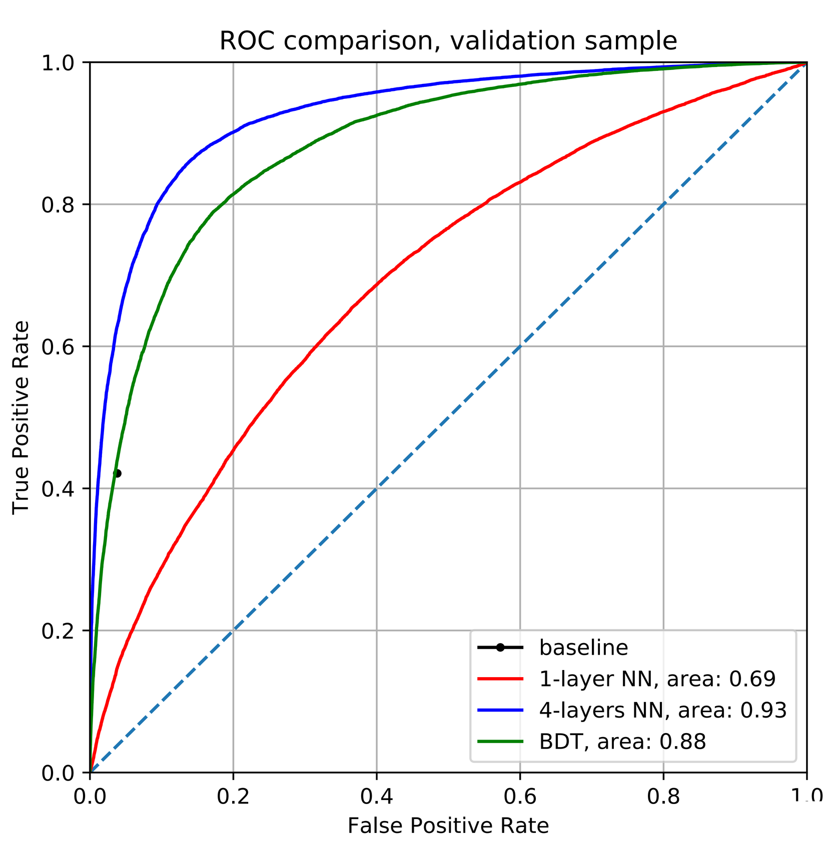
Z = 3 means probability (signal ~ stat. fluctuations of background) < 0.1%

The **baseline** is the result which was previously published by the ATLAS experiment. It was obtained with no machine learning, but simple, **rectangular cuts**.

On the right is the data-prediction comparison on the previous analysis using baseline method.



Results: signal & background separation

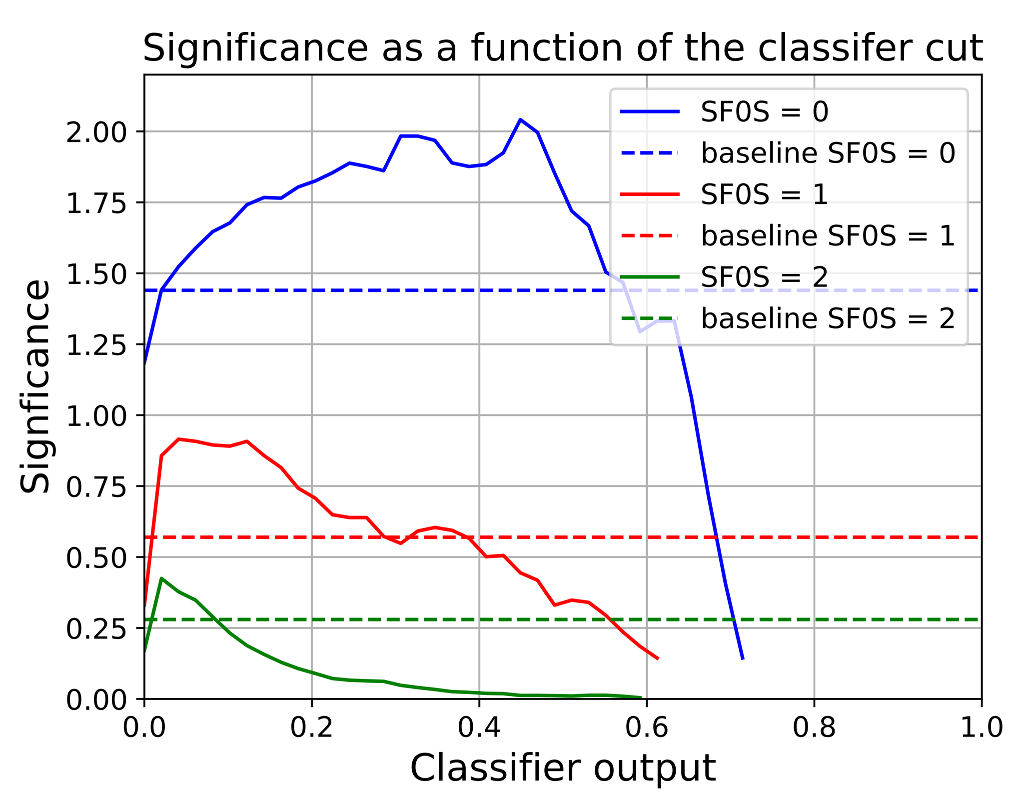


Evaluation of the various architectures through ROC curves are shown on the right.
Notably, the baseline is performing extremely well.

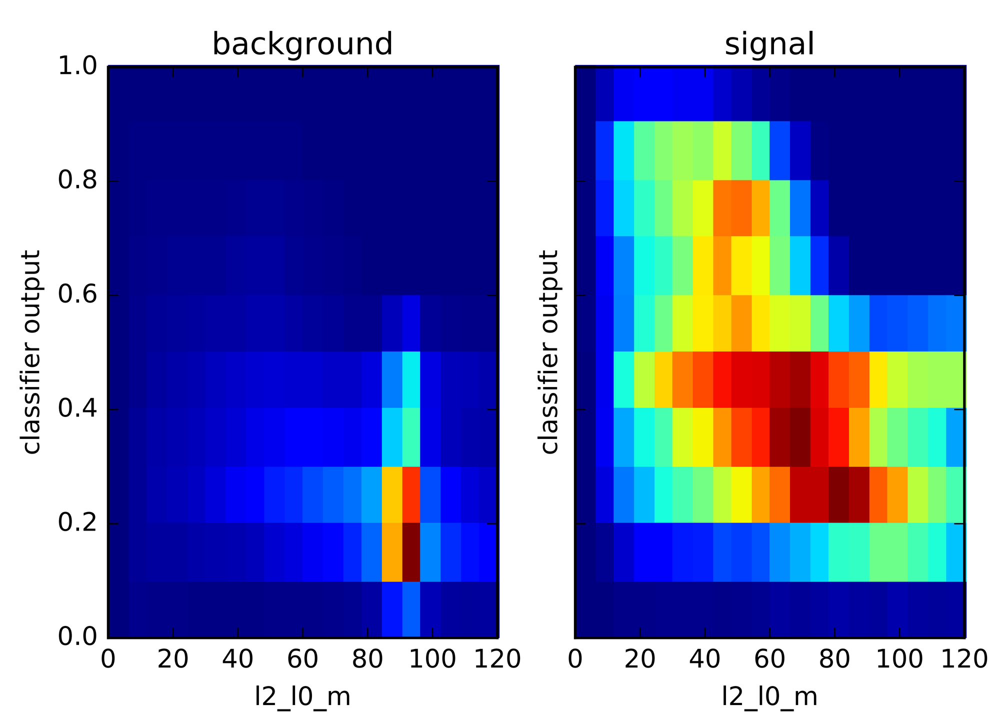
Both deep neural networks and boosted decision trees perform **better** than the baseline, as expected.

The output of the deep neural network classifier is then used to evaluate and optimize the significance.

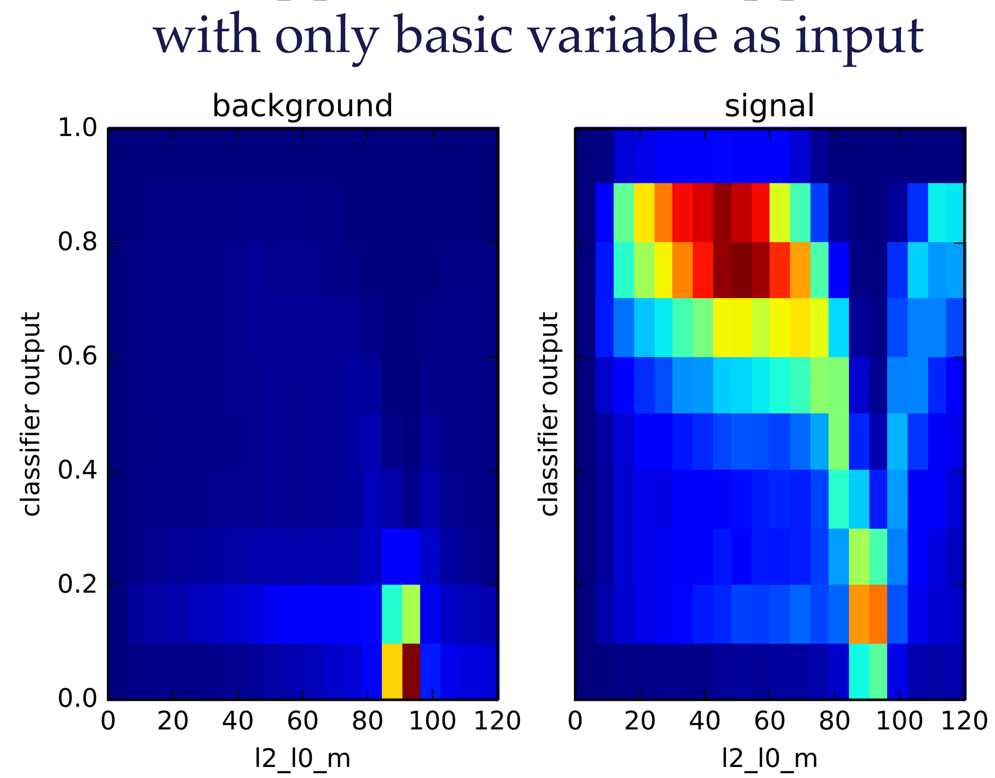
Using the DNN output, the significance with respect to the baseline **increases**.



Interpretations of outputs



with only basic variable as input



with high level variable input

Is machine **learning physics**?

This variable is the mass of the invariant sum of two 4-vectors. The non-linear computation involves **special relativity**.

By comparing the distribution in the model output: in the top row, although without explicit direct input, the model exploits the difference in **correlations** between signal and background, and gives a significant distinguish power in this variable.

In other words, the model **learns the distribution** as if it gets the direct input in the bottom row.

Nicolo Foppiani, Jonah Philion, Baojia Tong, CS281 2017