Prospects for Neural Network Applications in the LHC Data Analysis

Kyle Cranmer
University of Wisconsin-Madison

June 3, 2003
Cracow School of Theoretical Physics

Outline:

- Introduction
- Potential Applications
- The Impact of Neural Networks to the Search for a Light Higgs Boson
- Conclusions

A Special Thank You to

The Neural Network Session is sponsored by Center of Excellence of the European Comission

Kolmogorov's Superposition Theorem

THEOREM 1 (KOLMOGOROV'S SUPERPOSITION THEOREM)
For each $d \geq 2$ there exist continuous functions $\phi_q : [0,1] \to \mathbb{R}$, $q = 0, \ldots, 2d$ and constants $\lambda_p \in \mathbb{R}$, $p = 1, \ldots, d$ such that the following holds true: for each continuous function $F : [0,1]^d \to \mathbb{R}$ there exists a continuous function $g : [0,1] \to \mathbb{R}$ such that

$$F(x_1, \ldots, x_d) = \sum_{q=0}^{2d} g\left(\sum_{p=1}^{d} \lambda_p \phi_q(x_p)\right).$$

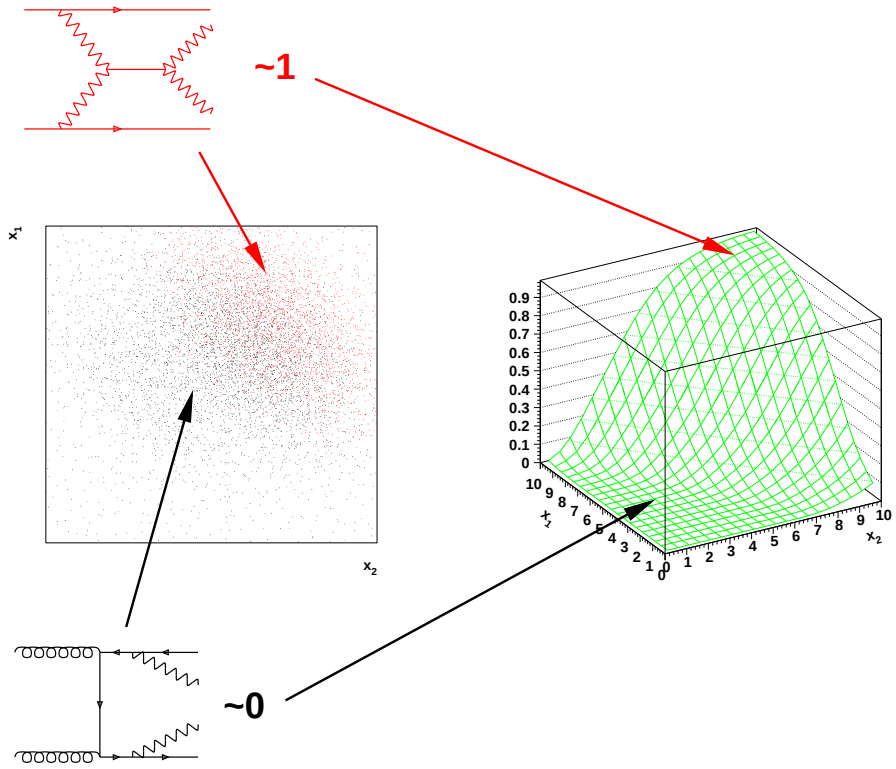
Note, ϕ_q and λ_p are independent of the represented function F.

Kolmogorov's paper, published in 1957, did not refer to neural networks directly; instead, it was in response to Hilbert's 13^{th} problem posed in 1900 to the International Congress of Mathematicians in Paris.

Exactly 30 years later Hecht-Nielsen noticed the application to the theory of neural networks: each continuous function $F : [0,1]^n \to \mathbb{R}$ can be implemented by a feed-forward neural network with continuous activation functions $t : [0,1] \to \mathbb{R}$.

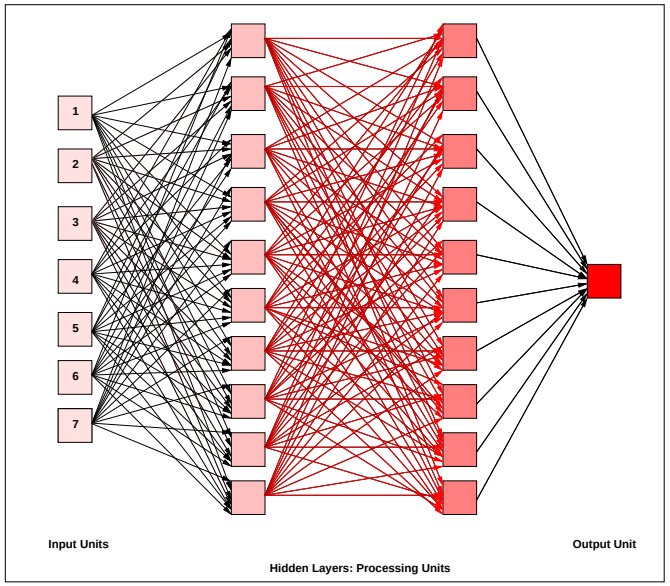
From Theory we can predict distributions of x for signal & background

If we associate signal with 1 & background with 0, we define a function on x



Kolmogorov's Theorem tells us a Neural Network can *represent* this function

What do Neural Networks look like?



The Network is made of Nodes N_j which process inputs I_i

The node is composed of a linear operation and a non-linear transfer function σ

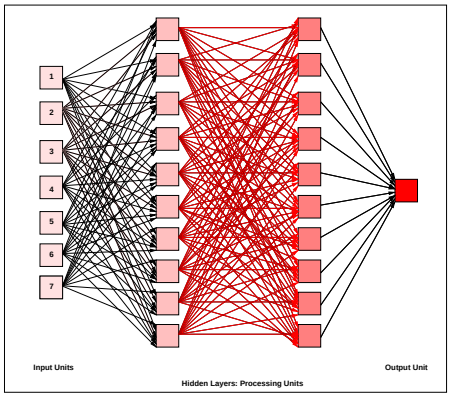
$$N_j = \sigma\left(\sum_i W_{ij} I_i + \delta_j\right) \tag{1}$$

Examples of $\sigma(x)$:

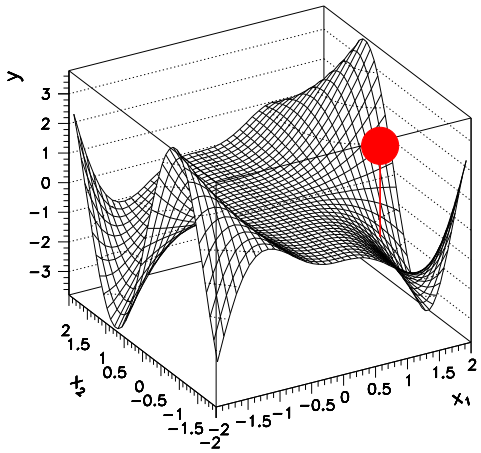
- $\arctan(x)$
- $1/(1+\exp(-x))$
- The Heaviside function

Neural Networks gained prominance when the PDP Group introduced the *backpropogation learning algorithm*.

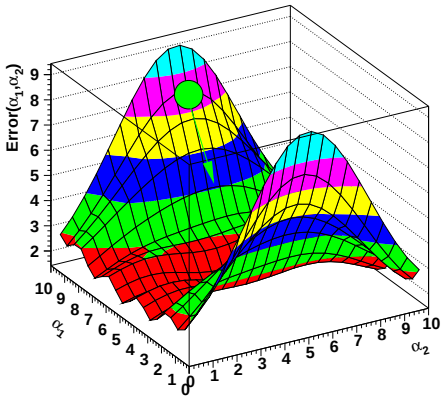
Schematic of Backpropagation



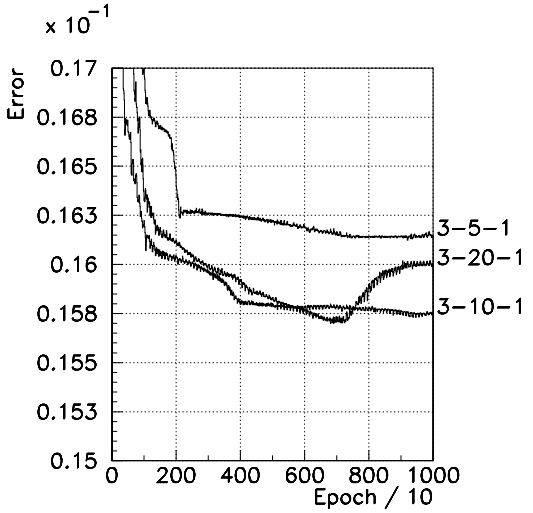
1) Start with some parameters α which correspond to some $f_\alpha(x)$



2) Calculate the Error $|y - f_\alpha(x)|$



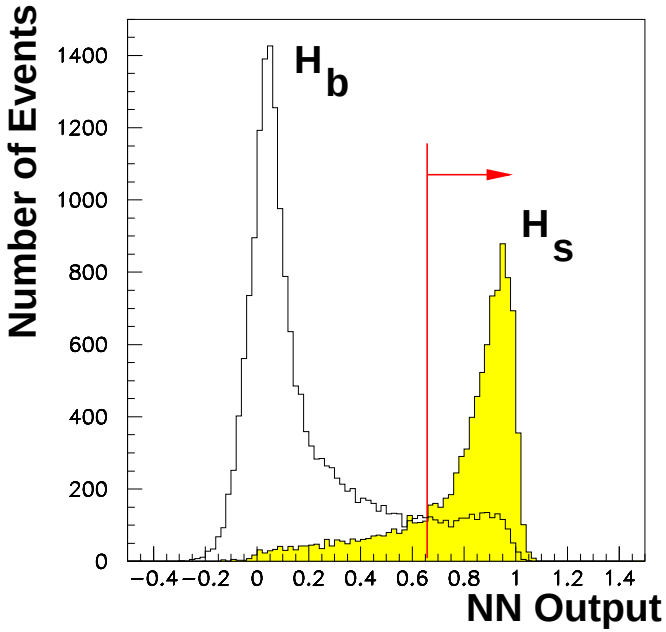
3) Apply the Chain rule to find $\Delta\alpha$ of steepest decent



4) Repeat Many Times until Error is minimized

Once the neural network has been trained, we *test* with an independent sample.

Typically, histograms of the `NN Output` are made which show the discrimination between signal and background.

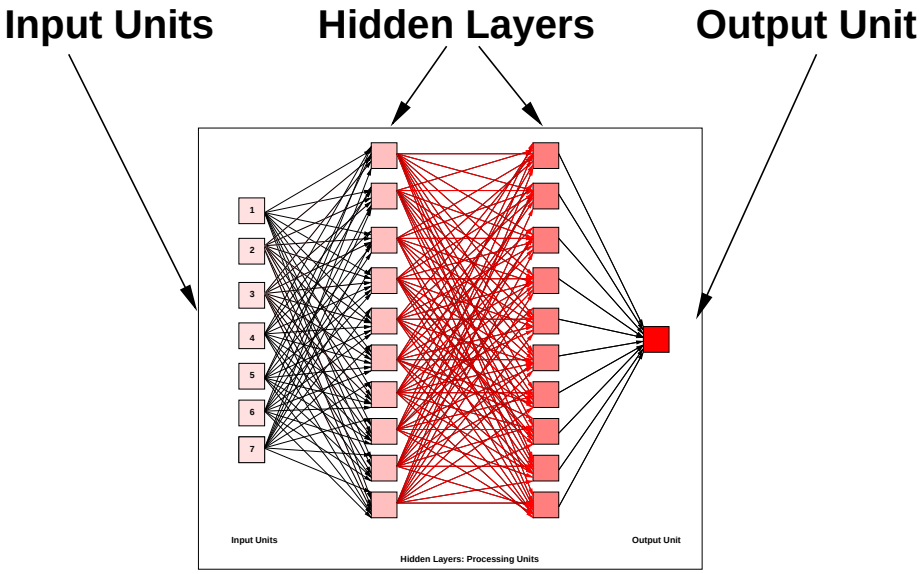


Candidate events are chosen by requiring that `NN Output` is greater than some NN_{cut}.

The choice of the NN_{cut} can be chosen to optimize:
- S/B - The Signal-to-Background Ratio
- $S/\sqrt{B}$ - The Gaussian Significance
- $P(S+B;B)$ - The Poisson Significance

It is also possible to use the shape in a statistical calculation based on the likelihood ratio. In that case you weight each event x_i by $\log(1+H_s(x_i)/H_b(x_i))$

The architecture of a Neural Network has impact on both its generalization properties and the training time.



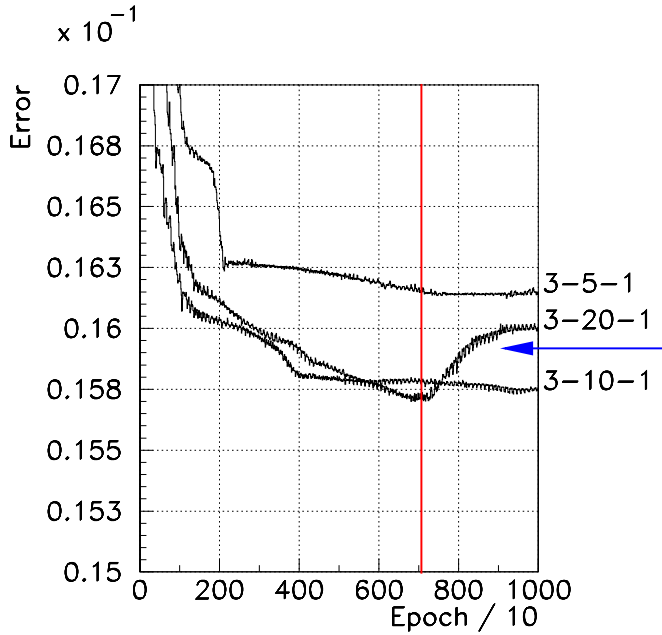
7−10−10−1

Kolmogorov's Theorem suggests
at least twice as many hidden nodes as input variables.

The choice of Architecture is largely a matter of trial-and-error.

In physics a specific phase-space point may correspond to signal *AND* background.

Given a finite training sample, the network might just learn the samples
a phenomenon called *overtraining*.

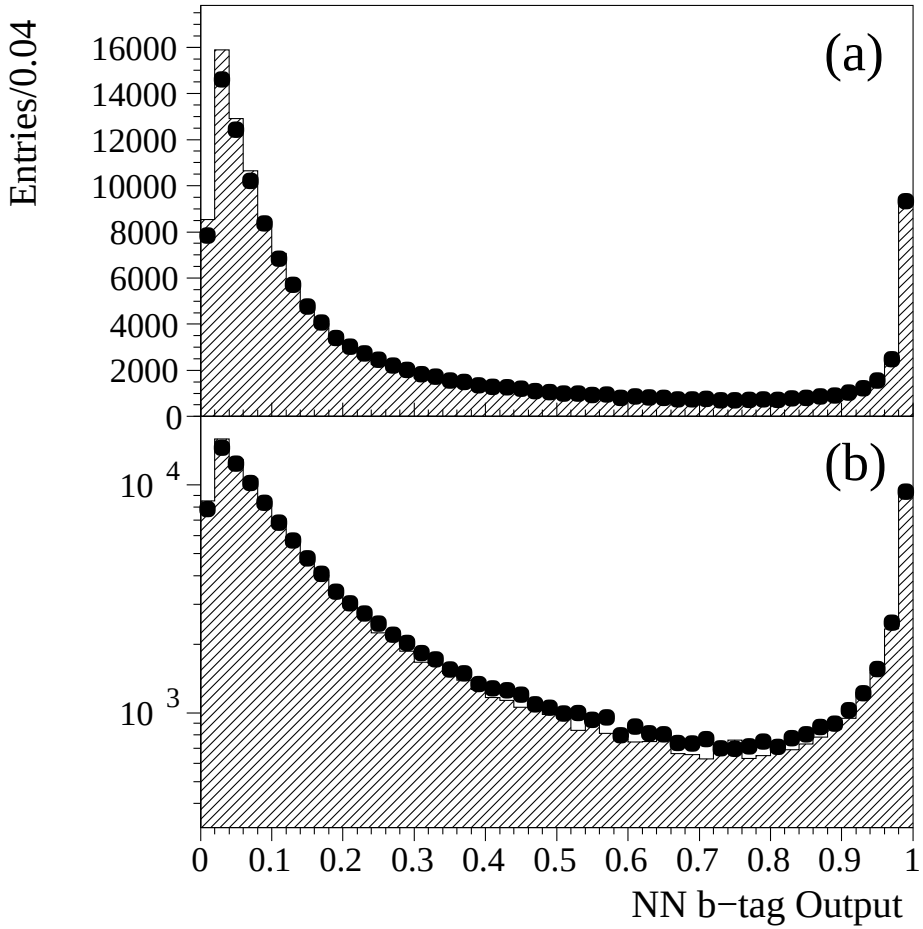


This Network was overtrained

Training Phase

Testing Phase

Once you have real data, you should validate the Monte Carlo you used to train the network



Example: This is a b-tagging Neural Network from ALEPH
Data & Monte Carlo agree very well

Applications of Neural Networks to the LHC

Neural Networks can be applied all along the analysis chain

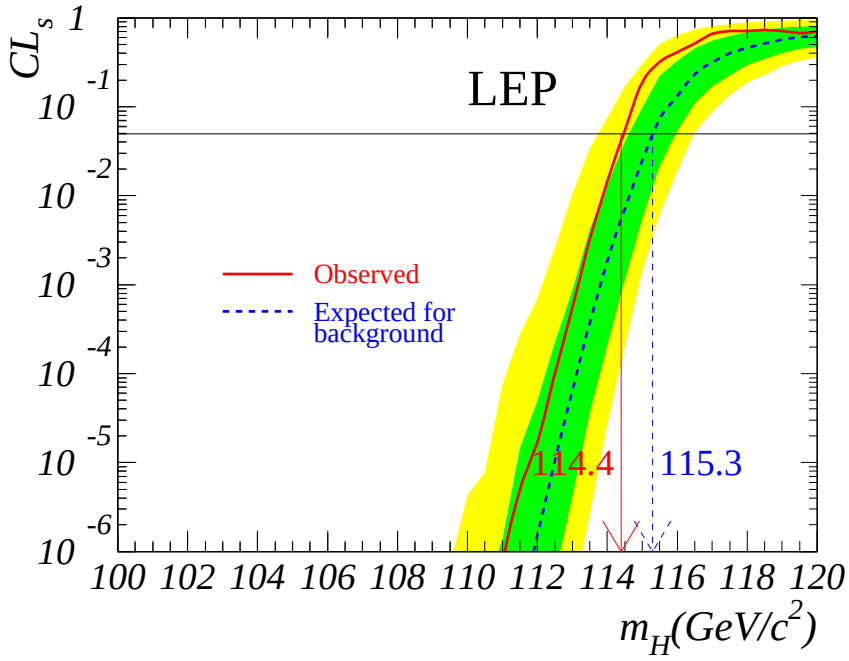
Triggering: Hardware NN's have been used for triggering

Track Fitting: Some are studying the use of NN's to fit tracks

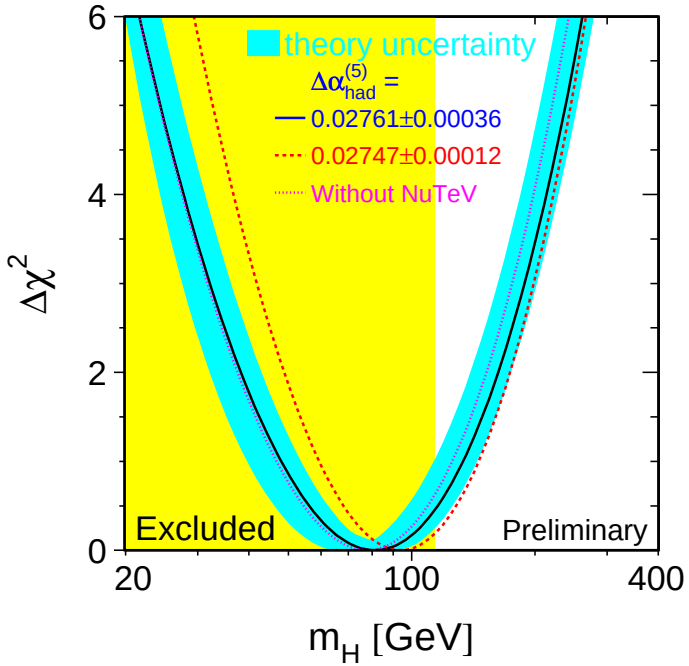
Flavor Tagging: A multivariate problem well-suited for NN's
Used at LEP, Tevatron, and BaBar

New Particle Searches: Searches for Higgs & SUSY will use NN's

Motivation for a Light Higgs



LEP direct search limit places
$M_H > 114.4$ GeV at 95% Confidence

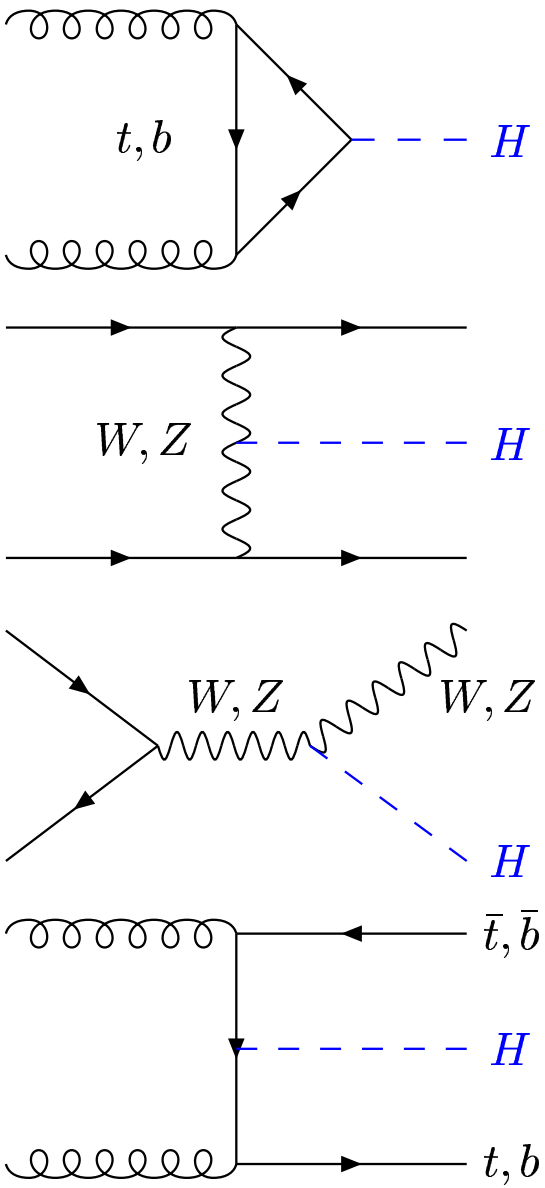


LEP Electroweak Fits limit
$M_H < 211$ GeV at 95% Confidence

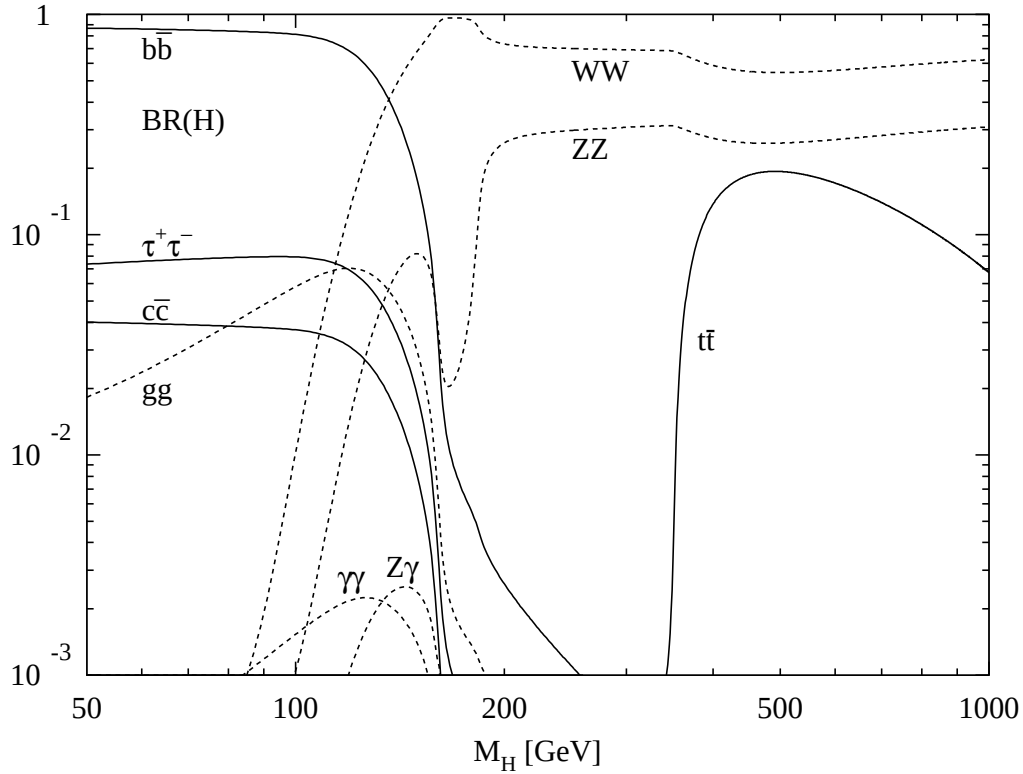
The MSSM predicts lightest Higgs to have $M_h < 135$ GeV

The low mass region is very exciting and very challenging!

Production and Decay of the Standard Model Higgs

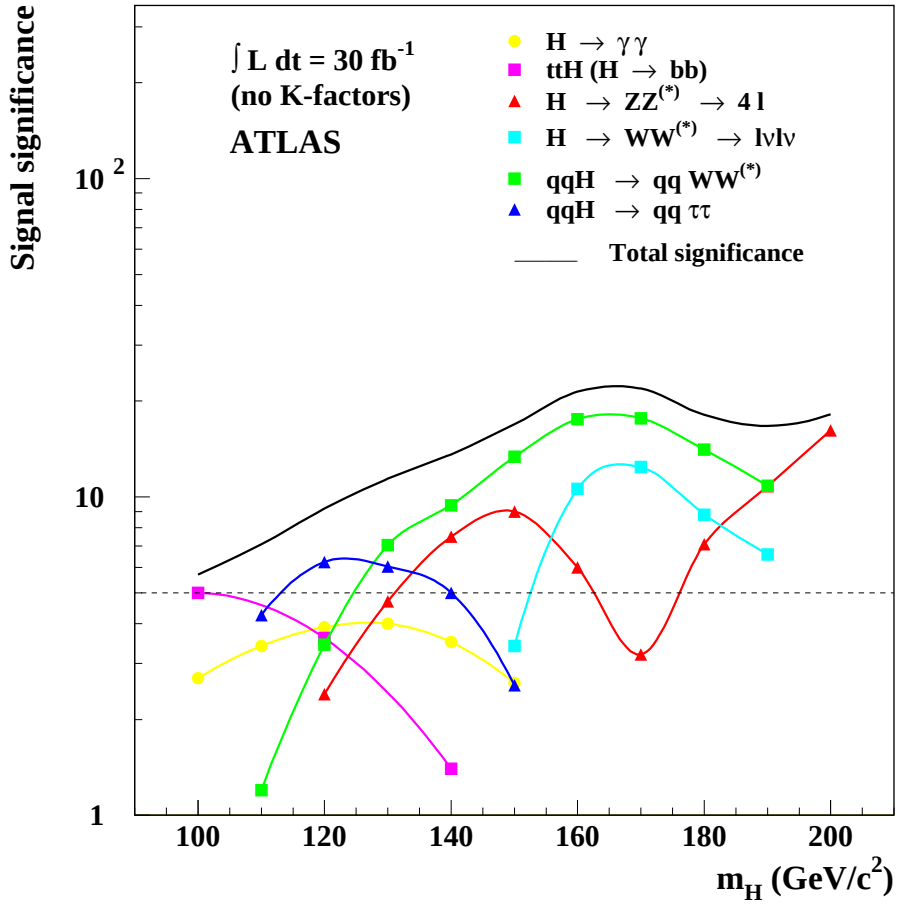


M. Spira Fortsch. Phys. 46 (1998)



- Gluon-Gluon Fusion dominant production process.
- Vector Boson Fusion (VBF) $\approx 20\%$ of gg at 120 GeV

- $H \to b\bar{b}$ dominant at low mass, but hard to trigger
- Forward Tagging Jets of VBF help S/B



VBF channels significantly improve the low mass range

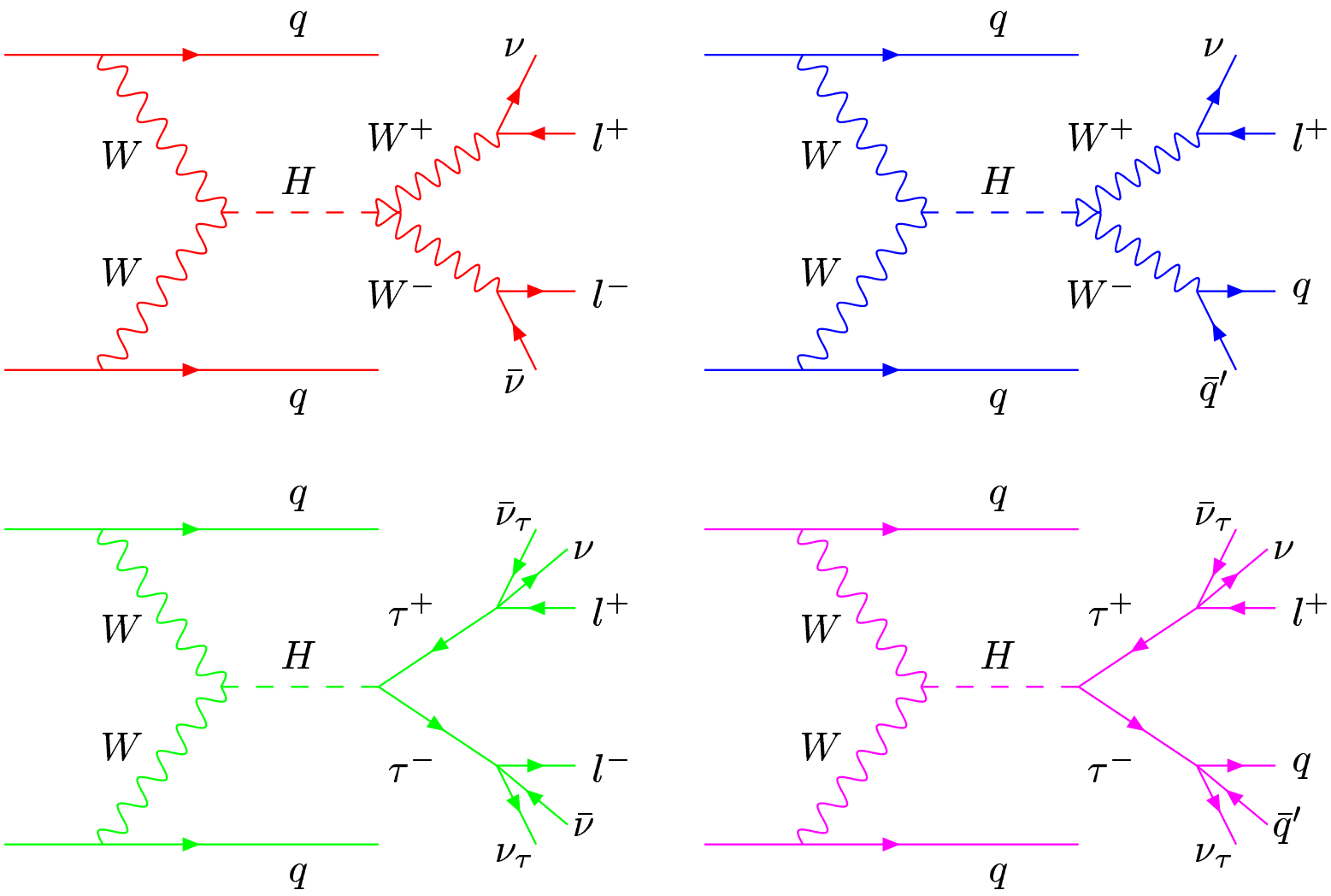
- $qqH \to qqWW^*$
- $qqH \to qq\tau\tau$

Several production and decay modes are available, so we can measure properties

5σ significance can be reached over the full mass range with 30 fb^{-1}

What is the impact of Neural Networks on the VBF Channels?

The $qqH(H \to WW)$ and $qqH(H \to \tau\tau)$ Final States



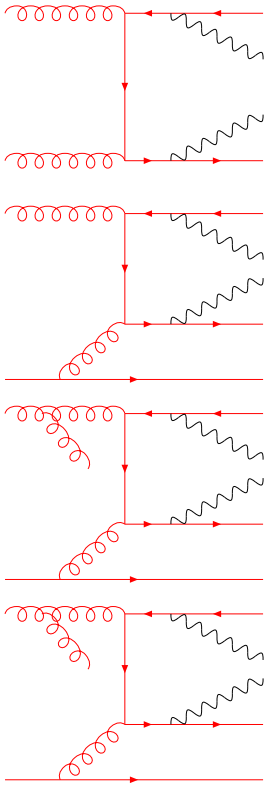
A brief overview of the main backgrounds for each VBF channel:

X = Dominant

x = sub-dominant

	$t\bar{t}$	WW+jets	W+jets	Z+jets
$H \to WW \to l\nu l\nu$	X	X		x
$H \to WW \to l\nu jj$	x	x	X	
$H \to \tau\tau \to l\nu\nu l\nu\nu$	x	x		X
$H \to \tau\tau \to l\nu\nu j\nu$	x			X

The $t\bar{t}$ Background

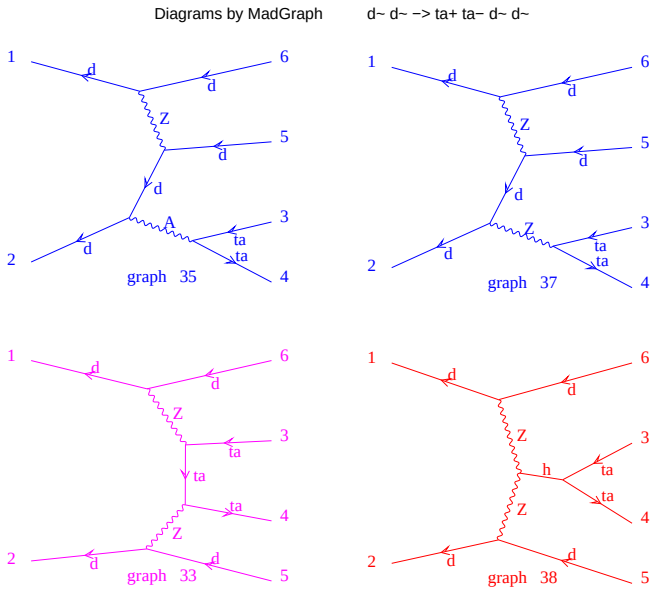
Process	Example Diagram
$t\bar{t}$	
$t\bar{t}j$	
$t\bar{t}jj$	
$t\bar{t}(jj)$ w/ FWE	

$t\bar{t}$ is a major background and appears in each of the VBF channels.

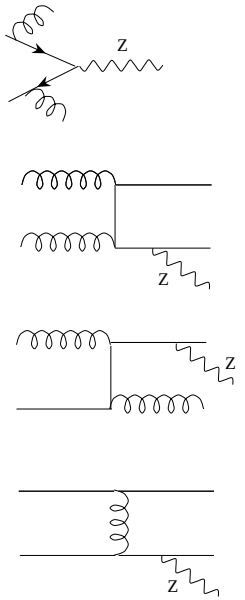
About 80% of the time one of the tagging jets comes from ISR and the other from a b-jet (from the top decay)

Divergences in $t\bar{t}j$ & $t\bar{t}jj$ are a problem, difficult to model well

EW Zjj



QCD Zjj

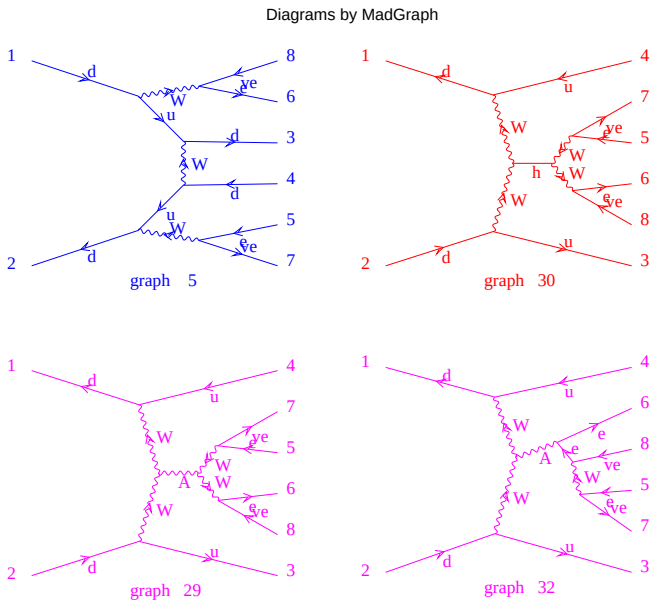


Identical color structure as signal

Larger cross-section than EW Zjj

$Z \to \tau\tau$ Irreducable background for $H \to \tau\tau$

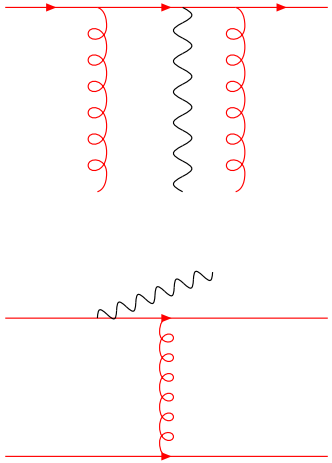
EW WWjj



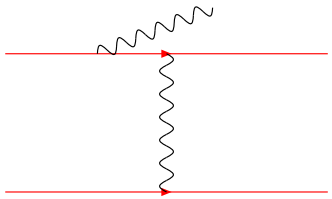
Identical color structure

Very small cross section

QCD Wjj



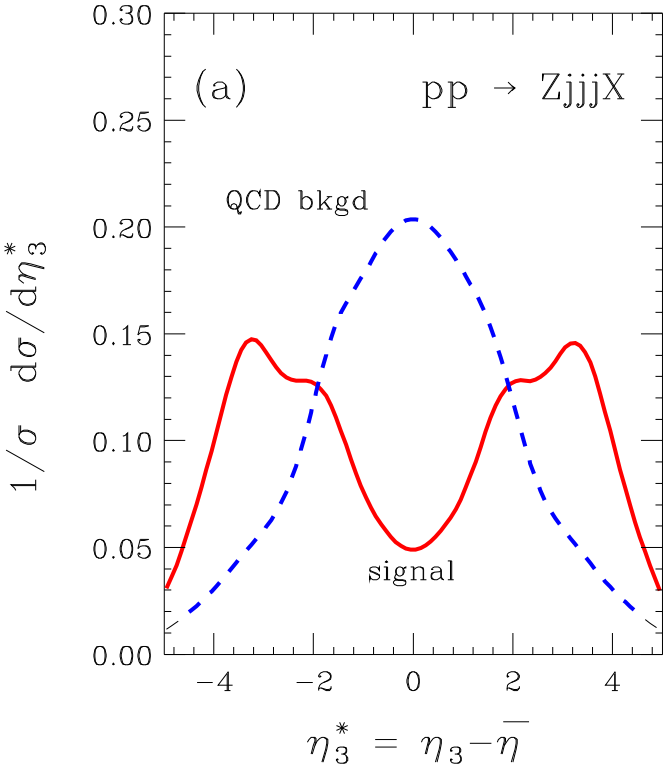
EW Wjj



The differing color structures between the VBF Signal and Backgrounds motivates a Central Jet Veto

$$\eta^* = \eta_3 - \frac{\eta_1 + \eta_2}{2} \qquad (2)$$

From the exact $Z + 3$ jet M.E. Calculation, we expect EW processes to be depleted near $\eta^* = 0$.

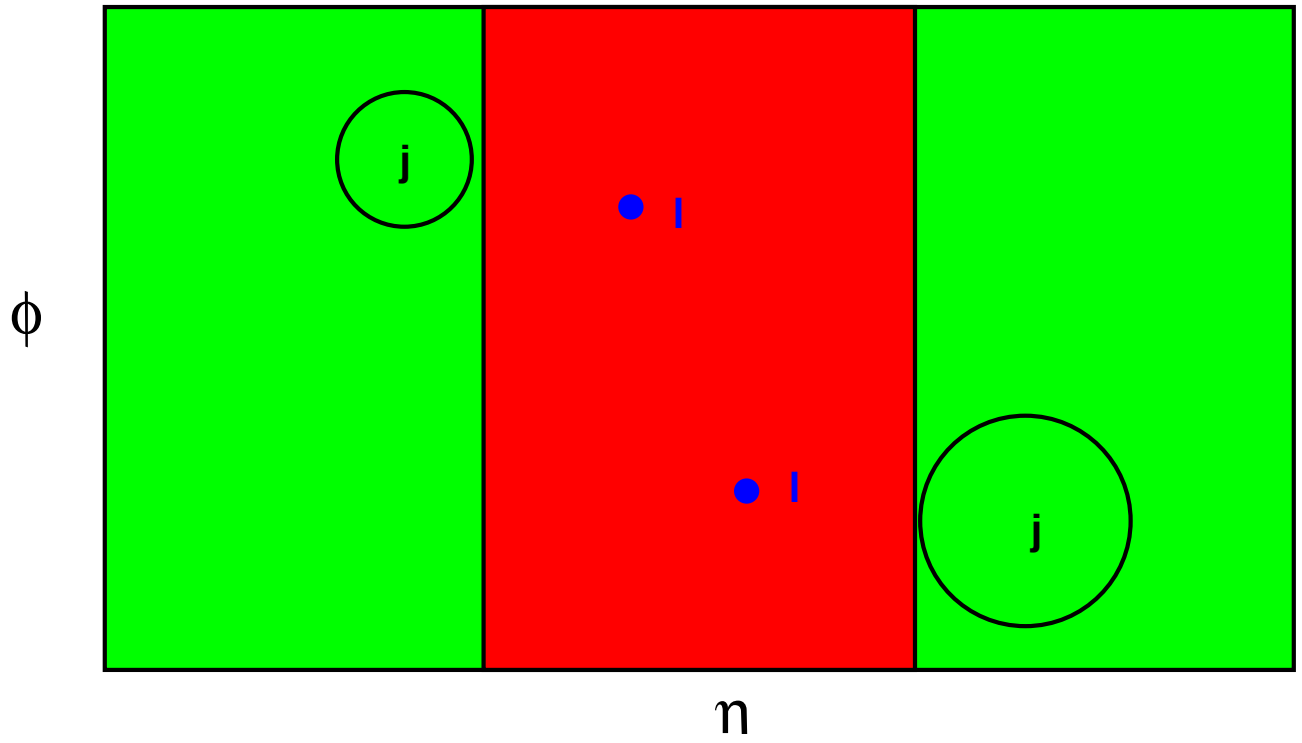


Rainwater, Szalapski, & Zeppenfeld hep-ph/9605444

Look for two forward *tagging jets*

Look for the Higgs decay products between the tagging jets

Veto the event if hard jets between the tagging jets



Vector Boson Fusion is a complicated final state

Expect correlations between tagging jets and the Higgs 4-momenta

Because the Higgs is a scalar, we expect the distribution of the "Higgs decay products" to be different for background

Exploiting these correlations is the job of a multivariate analysis

Choosing Variables

Intuitively we want as much information as possible in our variables x_i

But we need exponentially more training samples as the dimensionality d grows "The Curse of Dimensionality" viz. $N \propto e^d$

This tradeoff means we want the variables to have little redundancy

Boos, Dudko, & Ohl (EPJ C 11) have suggested that angular and singular variables from Matrix Elements as 'optimal' variables.

Including detector effects, these variables may not be optimal even in theory. Practicalities of the Multivariate Algorithm may in fact dominate.

As an experimentalist the choice of input variables is a practical matter guided by several considerations.

- Is the variable trustworthy?
  Is the Monte Carlo simulation prone to theoretical uncertainties?
  Are the relevant aspects of the detector simulation well-modeled?

- Does the variable discriminate between signal and background?

- Is the variable strongly correlated with other variables already included?
  If so, are the correlations well modeled?

Choosing Variables (cont'd)

As a first step, we see how much the current cut analysis is improved by applying Neural Networks

Thus, we restrict ourselves to kinematic variables which were used or can be derived from the variables used in the cut analysis.

- $\Delta\eta_{ll}$ - the pseudo-rapidity difference between the two leptons,
- $\Delta\phi_{ll}$ - the azimuthal angle difference between the two leptons,
- M_{ll} - the invariant mass of the two leptons,
- $\Delta\eta_{jj}$ - the pseudo-rapidity difference between the two tagging jets,
- $\Delta\phi_{jj}$ - the azimuthal angle difference between the two tagging jets,
- M_{jj} - the invariant mass of the two tagging jets,
- M_T - the transverse mass.

where $M_T = \sqrt{(E_T^{ll} + E_T^{\nu\nu})^2 - (\vec{P}_T^{ll} + \not{p}_T)^2}$

There are a pleathora of learning algorithms available.
I will leave it to the other speakers to describe them.

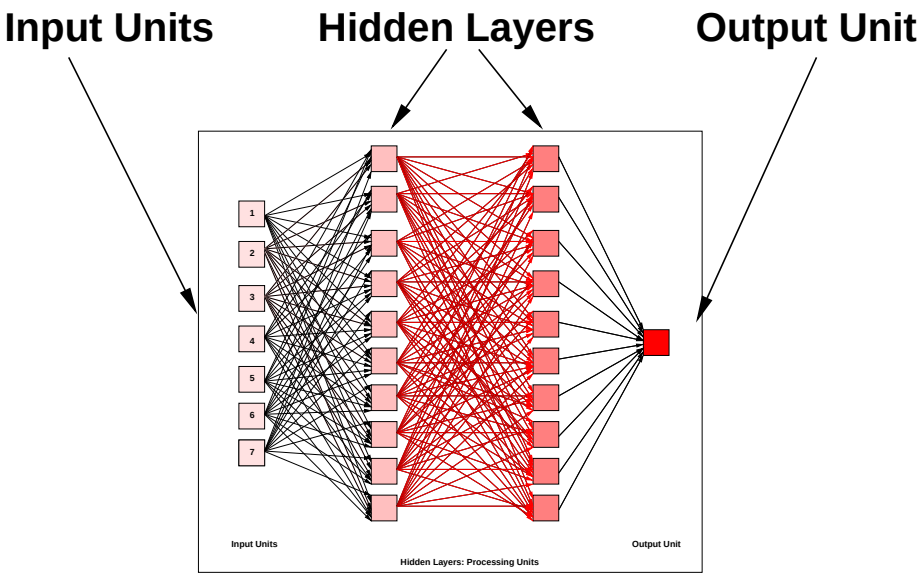
We performed two analyses:

1) With `SNNS` we used backpropagation with momentum
   A learning parameter $\eta = 0.01$ and momentum term $\mu = 0.01$ were used

2) With `MLPfit` we used Broyden - Fletcher - Goldfarb - Shanno method with a reset frequency of 400 and a τ value for the line search of 1.2.

The two methods agreed within the expected variability of different training runs.

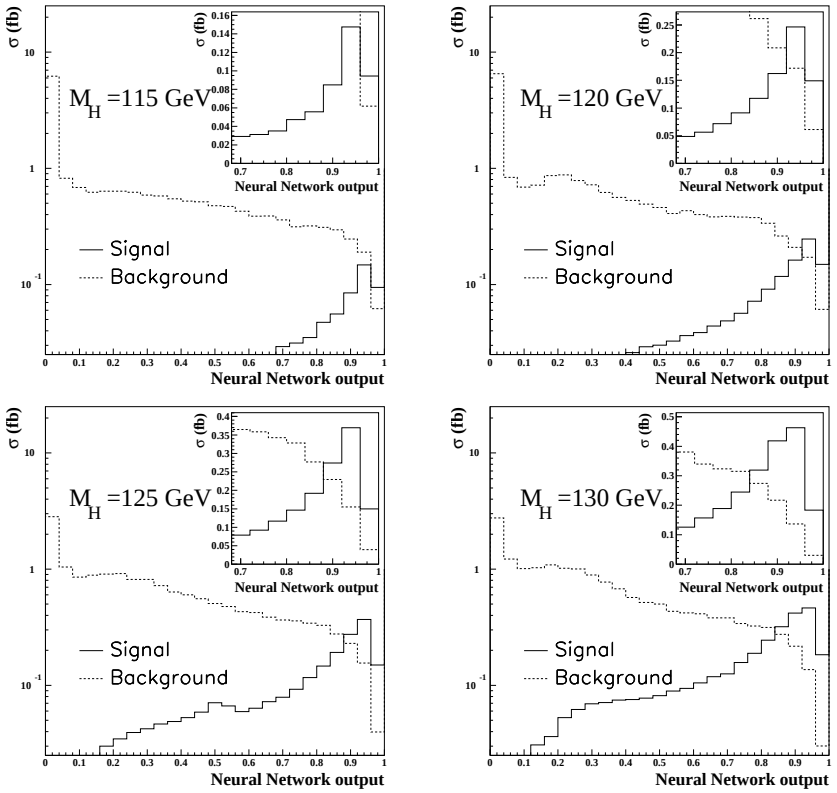
We tried using a variety of Networks



7−10−10−1

After some experimentation, we used both
7-10-10-1 and 7-10-3-1 architectures

The neural network output for signal (solid line) and background (dashed line) for $M_H = 115 - 130$ GeV with $H \to W^+W^- \to e^\pm \mu^\mp \not{p}_T$.

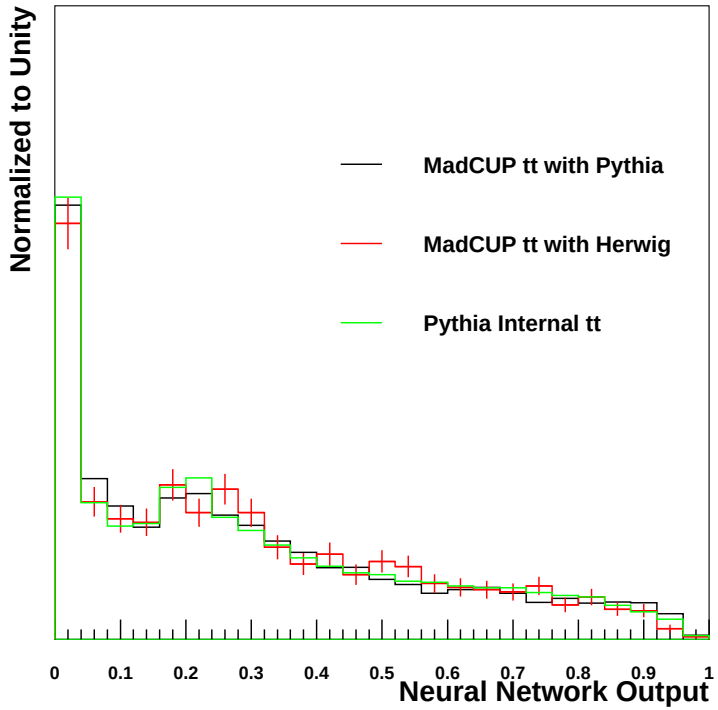


Notice that the signal is concentrated near 1 and the background is concentrated near 0.

Stability of the Neural Network

We do not yet have data to validate the neural network with...
... so we tried varying the Parton Shower model in the dominant $t\bar{t}$ background

By using the same Matrix Element with Pythia & Herwig,
the Parton Shower Systematic can be isolated.



We used the MadCUP Matrix Elements for $t\bar{t}$ and interfaced them with Pythia & Herwig via the Les Houches interface.

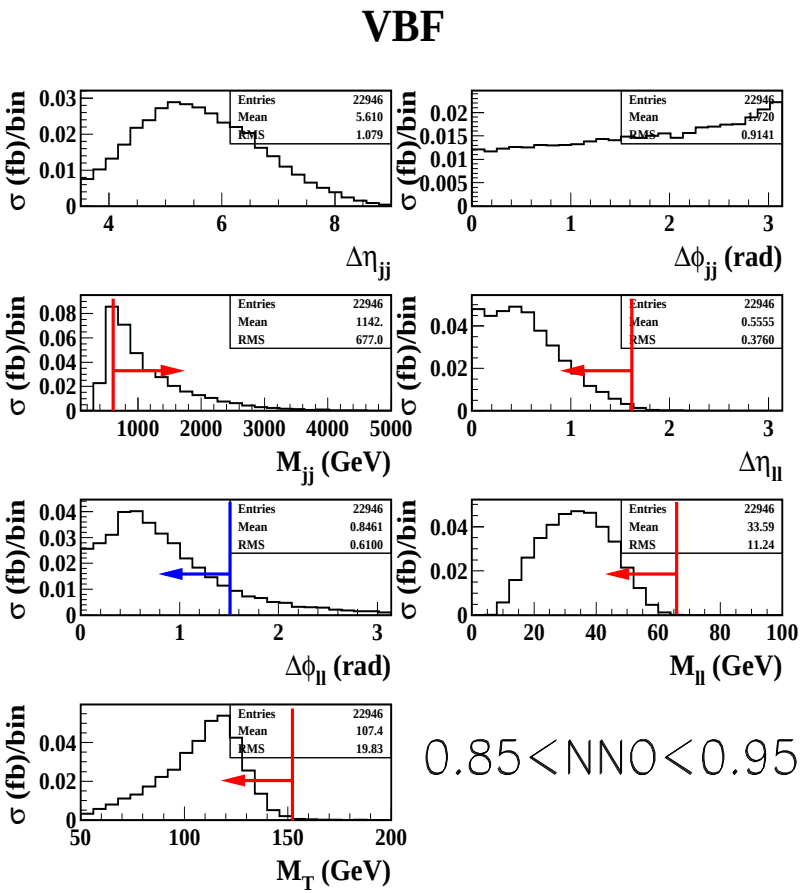
For Herwig, use 4.5 Million events

For Pythia, use 27.3 Million events

X-sec $(W \to \{e\nu, \mu\nu\}) = 20.8$ pb

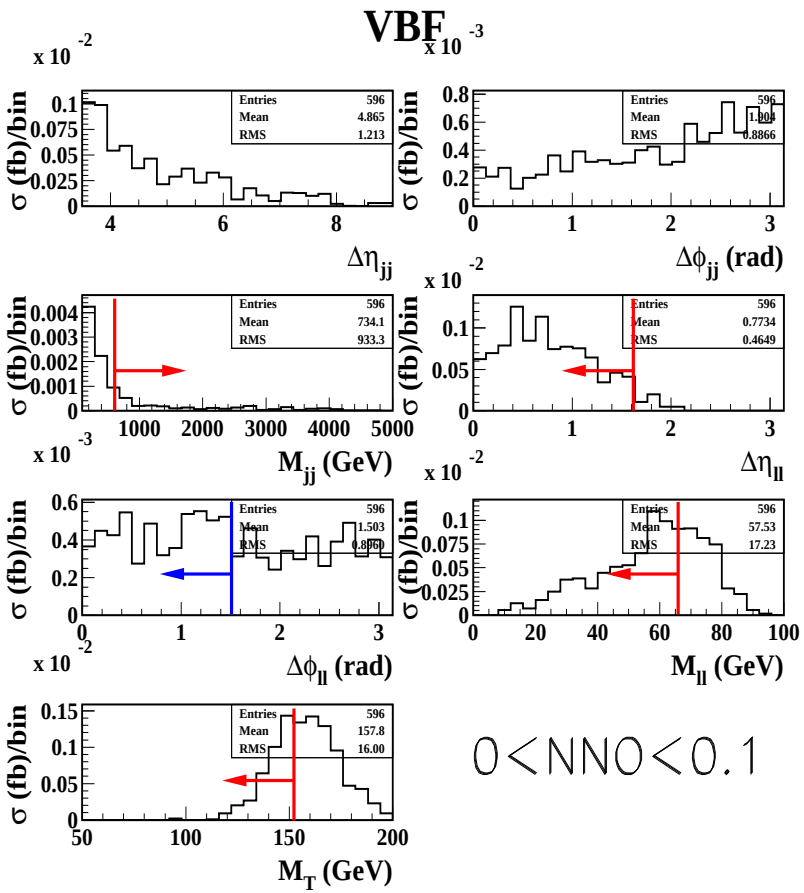
The Signal-like & Background-like Regions

Signal−like Region



0.85<NNO<0.95

Background−like Region

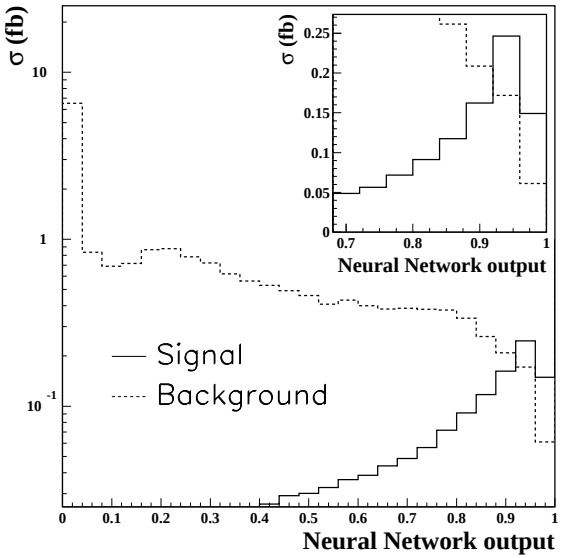


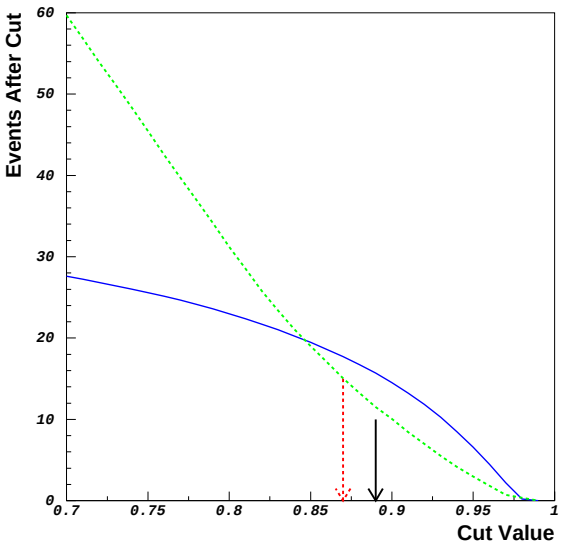
0<NNO<0.1

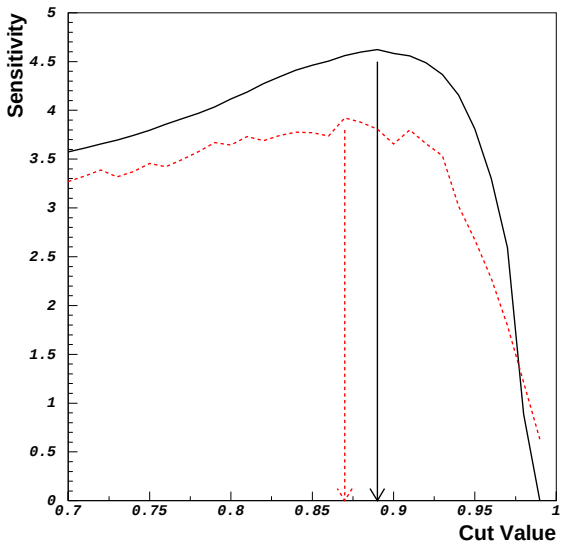
Features similar to the Cut Analysis

Features different than the Cut Analysis

Finding the Optimal cut on the Neural Network Output





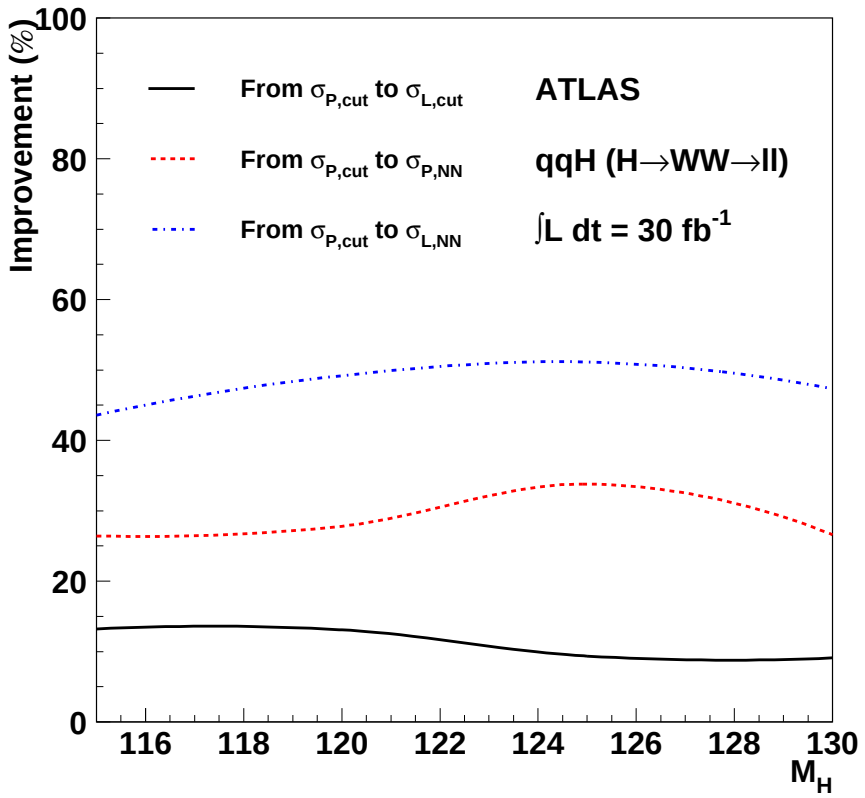


Varry the cut on the Neural Network output

Calculate Number of signal events s and background events b that survive cut

Maximize some sensitivity *e.g.* $s/\sqrt{b}$ or $P(s+b;b)$

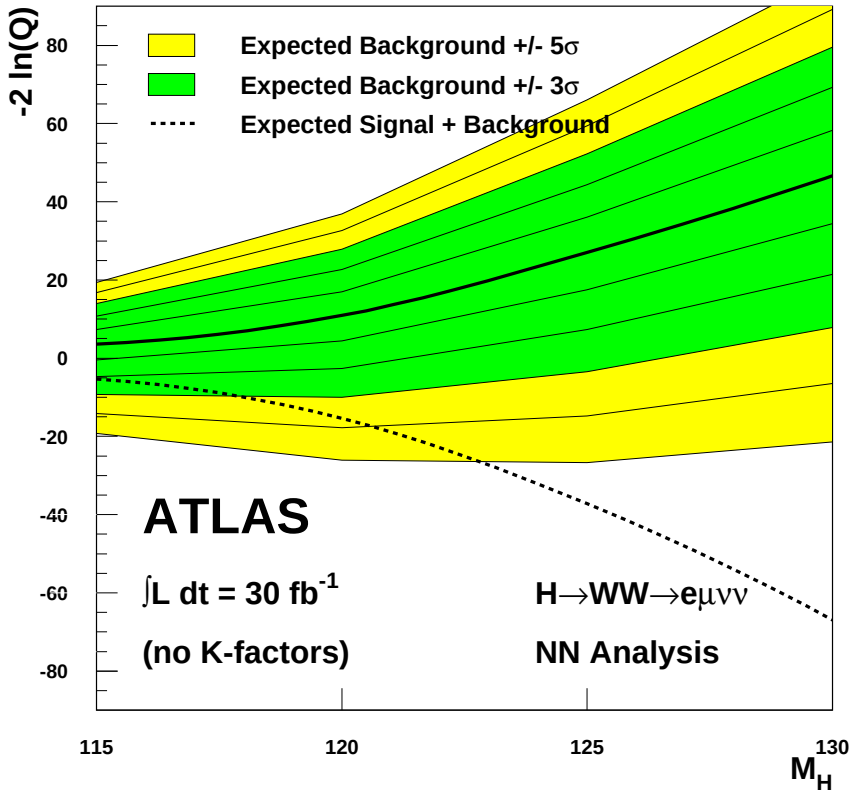
Result: For the $H \to WW \to ll\nu\nu$ analysis



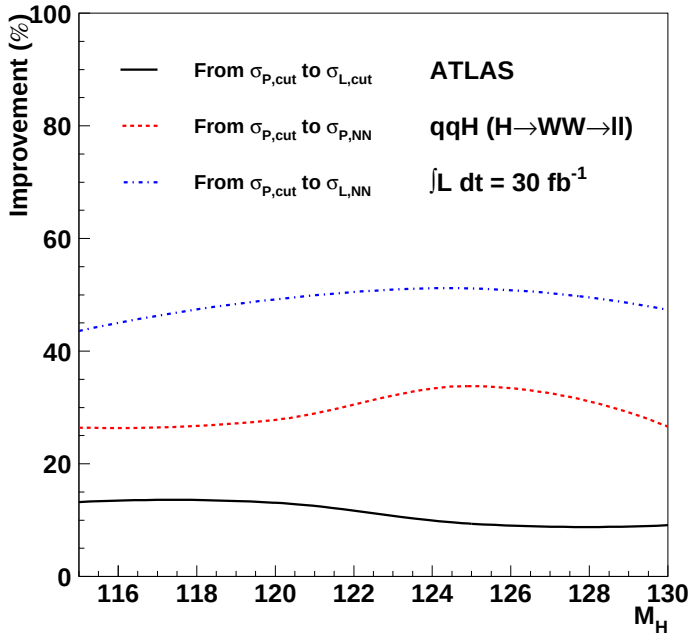
The Poisson Significance increased $\approx 30\%$ with Neural Networks

It is also possible to use the shape of the Neural Network
in a likelihood-ratio calculation

In that case each event x_i is weighted by $Q = 1 + H_s(x_i)/H_b(x_i)$



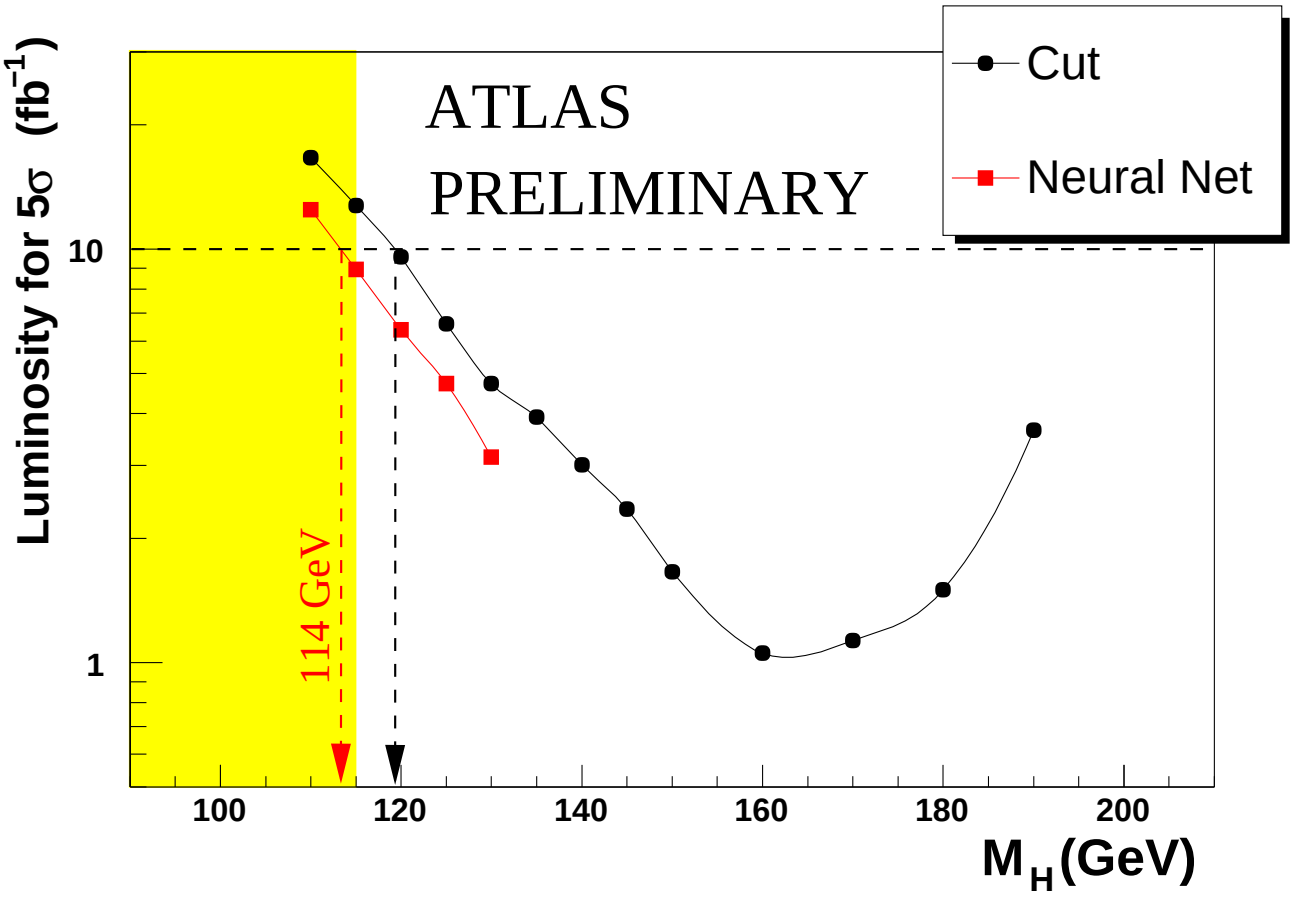
Impact on Significance



Using statistics based on the Likelihood Ratio signficance increased $10 - 15\%$ for the cut analysis.

Combining the Neural Network with Likelihood Ratio increased the signficance $40 - 50\%$ relative to the cut analysis with Poissonian statistics.

A similar analysis was done for $H \to \tau\tau$



Preliminary work suggests that Neural Networks cause a significant impact to the amount of luminosity required to reach the 5σ discovery threshold.

Neural Networks have great potential in the complex analyses encountered at the LHC

Expect Neural Networks to be used for
Flavor Tagging & New Particle Searches

Preliminary work for Higgs produced via Vector Boson Fusion
show 40-50% improvment in significance

Still A lot of work to be done!