Machine Learning approach for CMS H$\rightarrow \tau\tau$ analysis

- Master thesis -

E. Kreuzgruber

HEPHY Students' Day

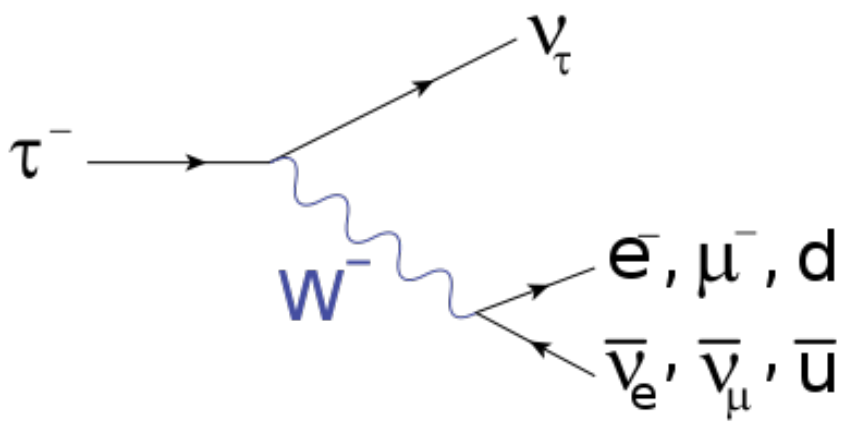
05/09/18

Outline

- Motivation behind H$\rightarrow \tau\tau$ analysis
- Machine Learning in general
- Ingredients for good training

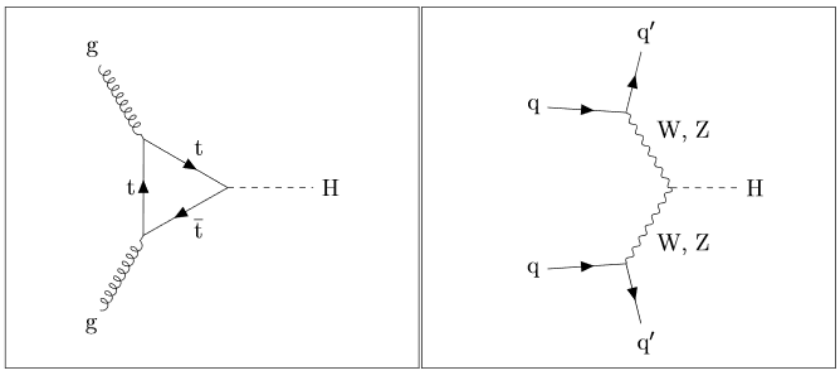
H$\rightarrow \tau\tau$ decay

- Fermion masses via Yukawa couplings
- High branching ratio in $\tau\tau$ channel
- decay modes of τ - hadronically (τ_h), leptonically (e,μ)
- we look at $e\tau$, $\mu\tau$, $\tau\tau$



Higgs production and decay

- main Higgs production modes at LHC:
  - Gluon-gluon fusion (ggH)
  - Vector boson fusion (VBF)



(a) ggF (b) VBF

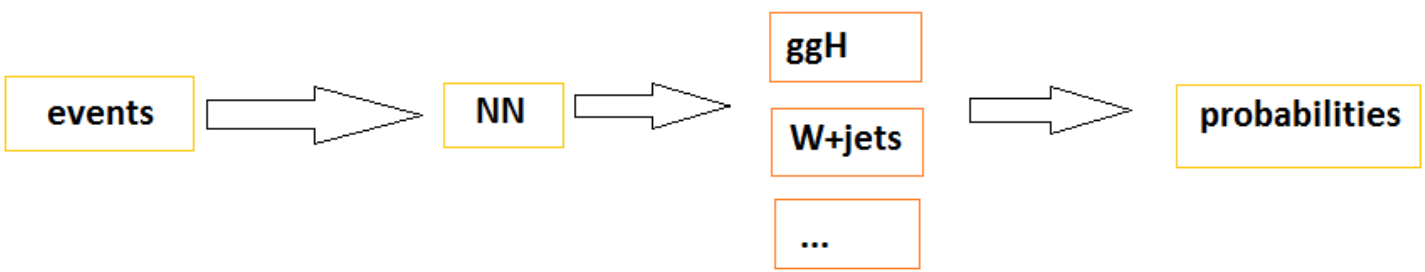
- main background: $Z \to \tau\tau$
- W+jets, multijets/QCD; or decay with lepton or jet faking τ

Machine Learning
General idea

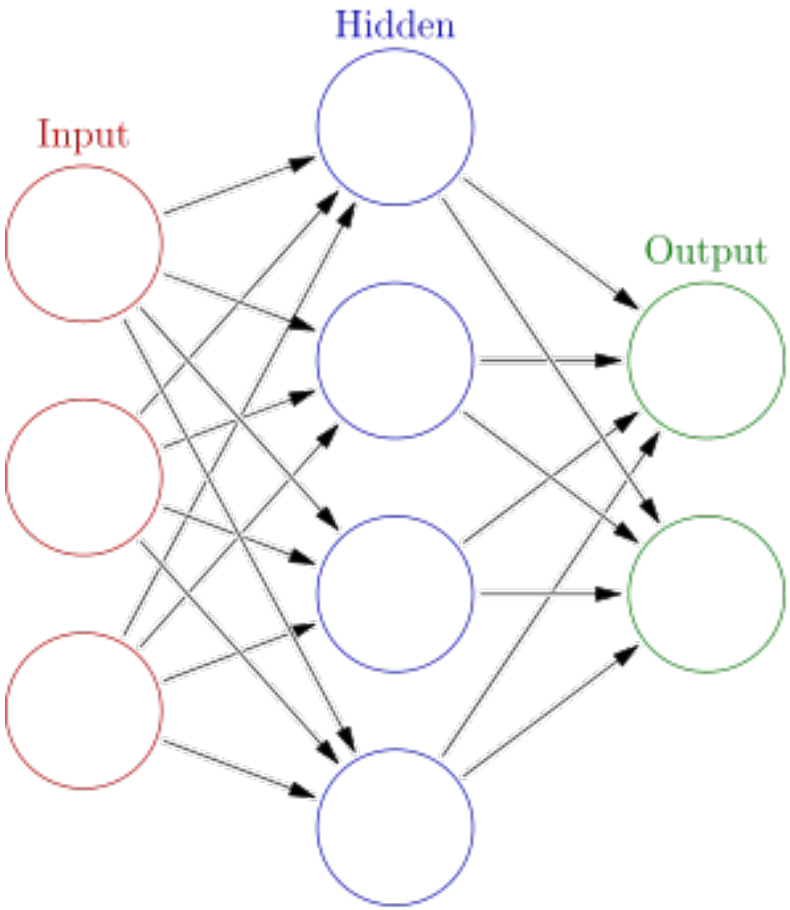
- Take MC simulated data
- split into two datasets (even/odd)
- train on one, test on the other
- switch & repeat
- good performance $\rightarrow$ apply on data

Machine Learning
Algorithms

- Boosted decision trees
  - XGBoost
  - $\rightarrow$ Markus
- Multiclass neural networks
  - Keras
  - $\rightarrow$ KIT



Neural networks



- neurons connected via weights
- take weighted input as pre-activation
- apply activation function: f.ex. lin, relu, tanh
- calculate output to next layer

Keras

customize structure of model:

- Architecture - 2 layers, 200 neurons each
- Activation function - tanh
- Optimizer and loss function

Problem: not always enough statistics for every class

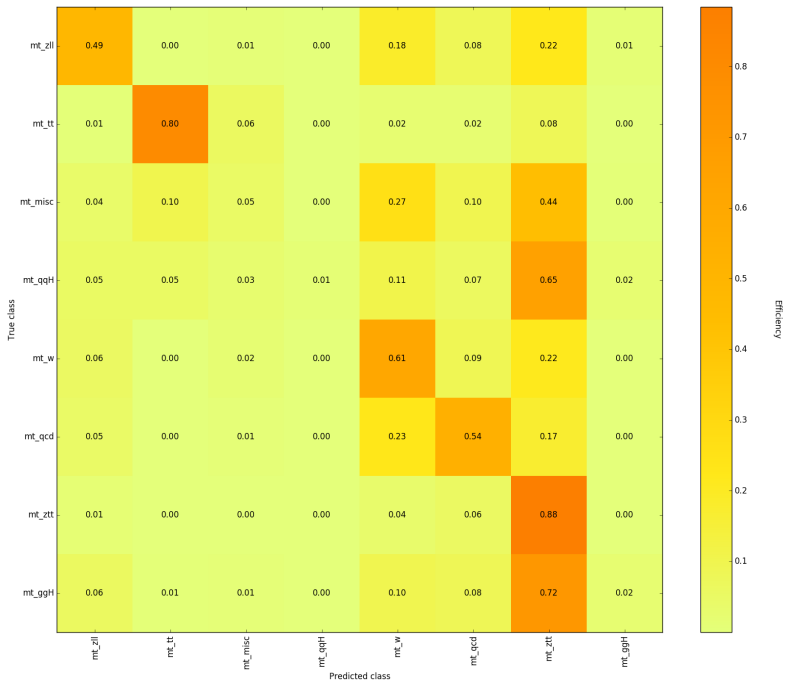
Large background samples/small signal samples $\rightarrow$ adjust weights for NN accordingly

Output: probability vector $\rightarrow$ associate class with highest probability

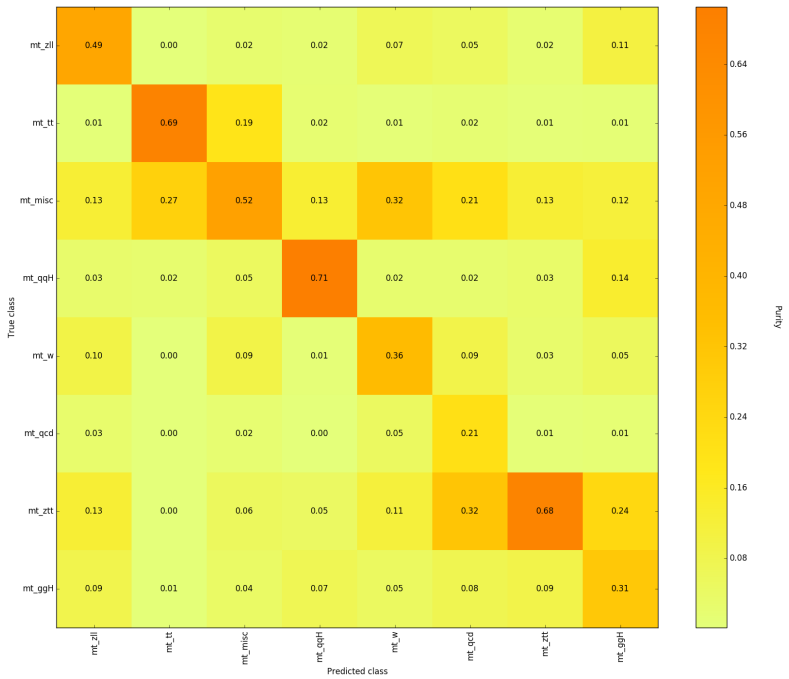
Confusion matrices

Check quality of training/testing

How well does predicted class match the actual class?



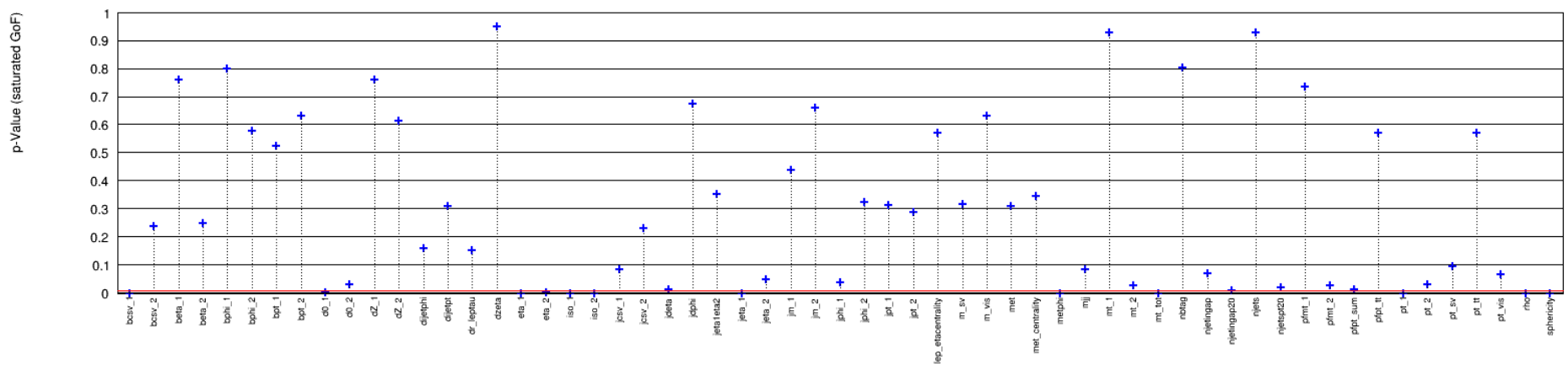
efficiency representation



purity representation

Input variable selection

- find best variables for training
- use saturated goodness of fit test
- 1D (marginal distributions) and 2D gof



Data cards

Put all information we have into data cards to use for training:

- QCD estimation (SS)
- Systematic uncertainties
- t-shapes
- Jet energy corrections
- categories (0-jet, boosted, vbf)

Thank you for your attention!