Analysis update

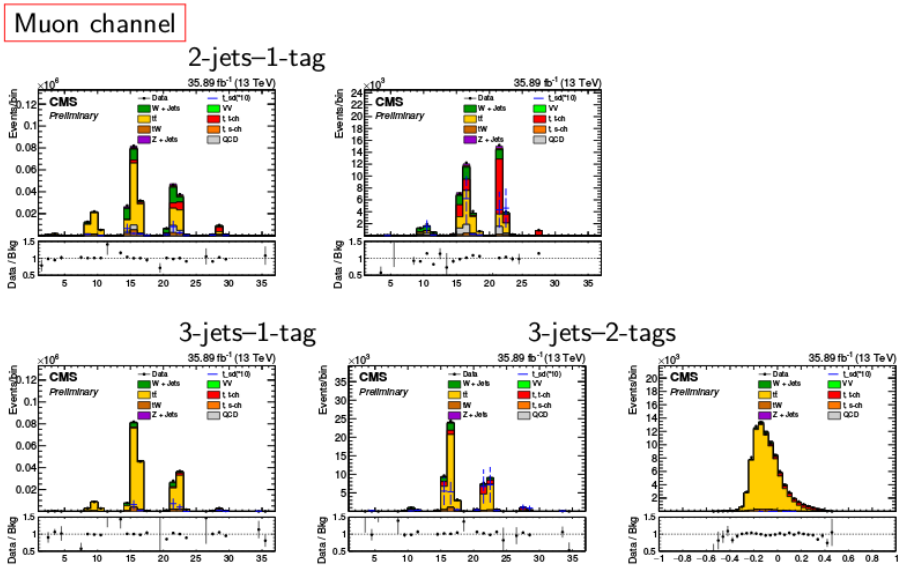
Lukas Layer

Overview over the first time

- Bureaucracy + 2 workshops CERN and Padua
- Learning the naples framework and combine ($\rightarrow$ reading a lot of code)
- Learn the analysis done by Agostino: Measurement of the CKM matrix elements $|V_{tb}|^2, |V_{ts}|^2, |V_{td}|^2$ in single top events
- Now work on: Training of the BDTs with alternative methods + Production of datacards with combine harvester
- Still a lot of things to understand...

Analysis overview (Agostino)

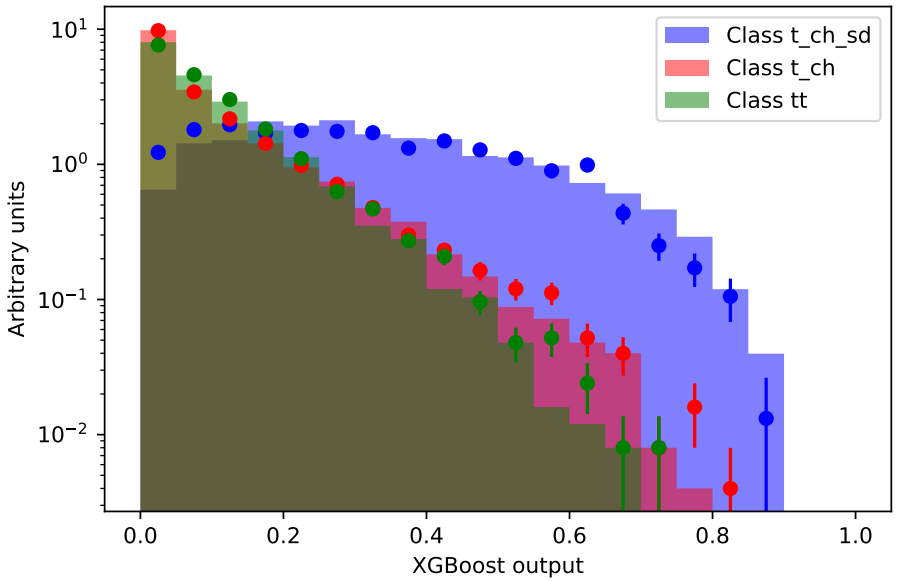
- Fit to standard single top t-channel to extract $|V_{tb}|^2$ and $|V_{ts}|^2 + |V_{td}|^2$
- central and signal region in 2j1t, 3j1t, one region in 3j2t
- MVA analysis in the different regions for electron and muon

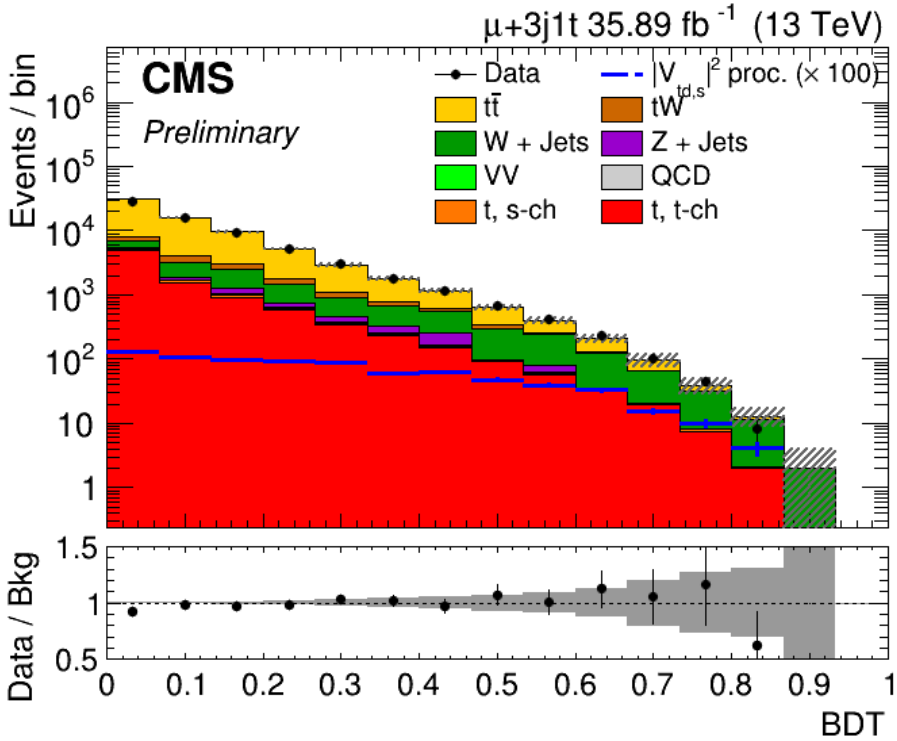


→ **challenges: many nuisances, little discrimination, statistical limitations**

Training of BDTs: fast checks on SWAN

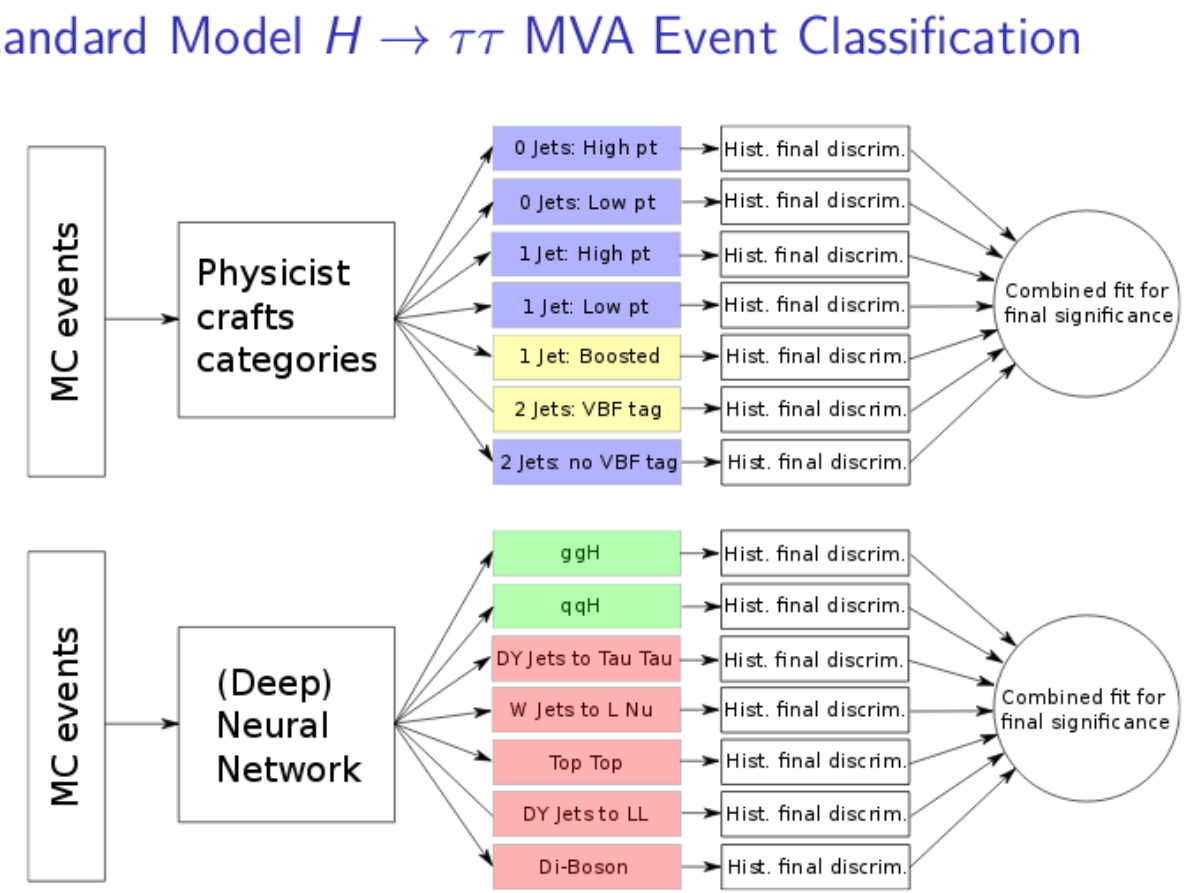
- Implementation on CERN SWAN (Service for Web based ANalysis)
- Possible use of (multiclass) training with TMVA, Keras, XGBoost, h20... with jupyter notebooks, pandas etc.
- Fast plot of stacks
- First scripts written but need to be validated





Why multiclass classification

- Multiple nodes with probability for each class
- $\rightarrow$ possible definition of categories
- Natural way of controlling the training of multiple sources, in particular importance!



MVA variable validation

- ML exploits deep correlations
- 1D check of variables not enough - also correlations important
- Best way of doing this checks?

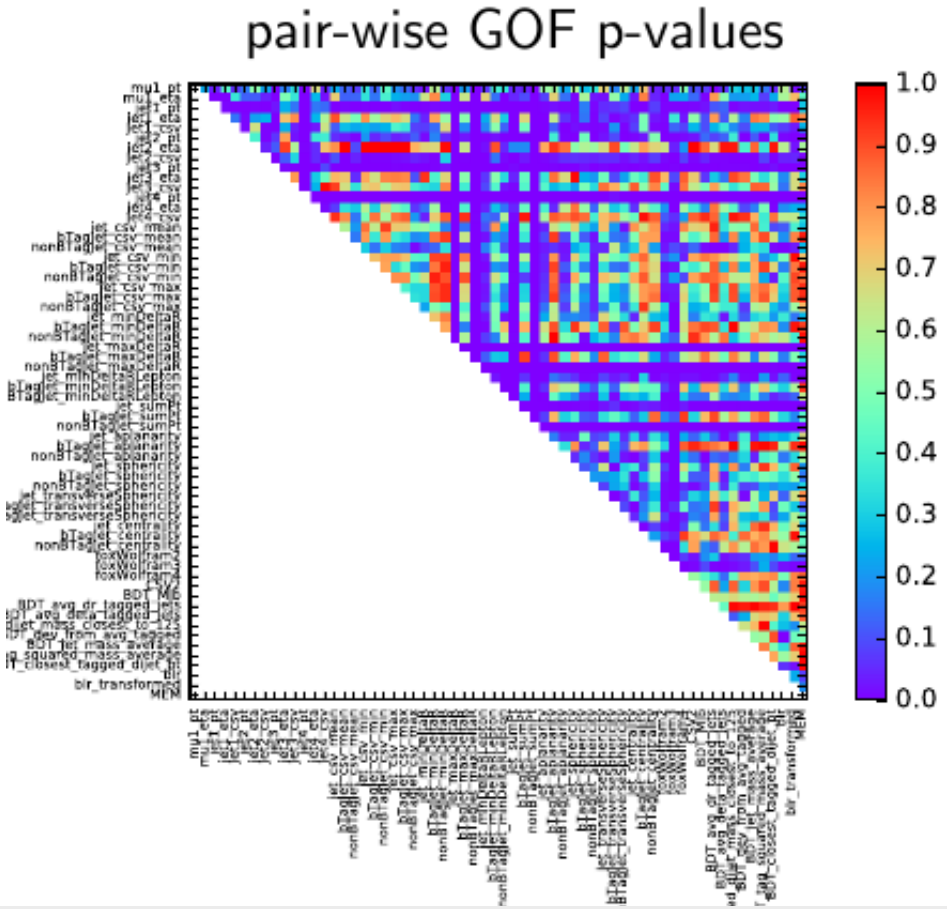
- KS test for data-MC 2D histogram (CMS ttH)

  https://indico.cern.ch/event/668017/contributions/2947012/

- Nearest-Neighbors Two-Sample Test https://indico.cern.ch/event/752376/contributions/3164224/attachments/1732946/2802184/DeSimone.pdf

- Domain adversarial NN

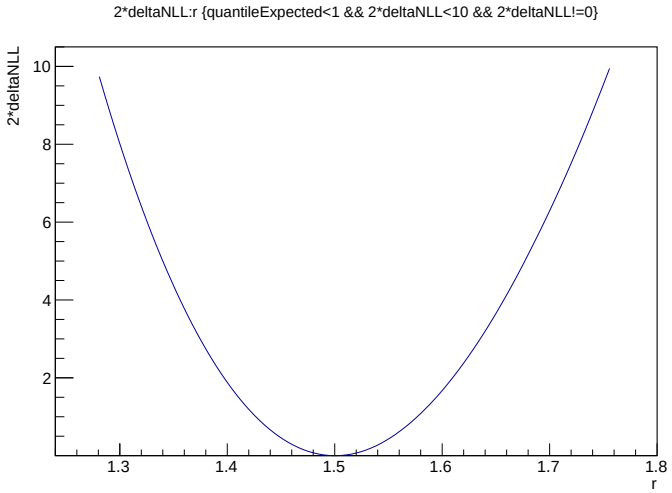
  https://indico.cern.ch/event/668017/contributions/2947022/attachments/1630499/2598803/IML_talk_JMMA.pdf



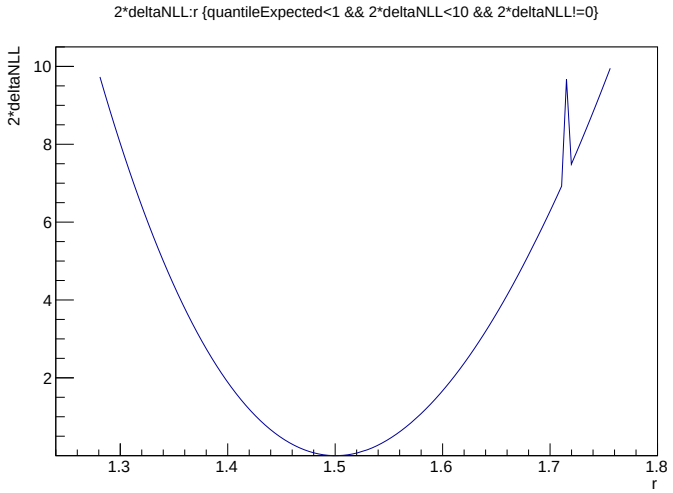
Combine Harvester

- implement datacard production with combine harvester
- get independent from local framework
- crosscheck with naples framework with fit of single top t-channel without MC statistics

r = 1.50026 +/- 0.0753699



r = 1.50045 +/- 0.075378



- values numerically not the same, but very similar
- combine harvester makes life easier

Different treatment of MC statistics

Naples framework

- include standard single top t-channel uncertainty if relative fraction of events > 20
- include sample uncertainty if $e_i/n_i^{tot} > 0.04$

Combine harvester

- bin contents x_i, bin errors e_i with e_i/x_i larger than the bin-by-bin addition threshold
- Rank bins from lowest to highest fraction error: e_i^2/e_{TOT}^2
- Start removing bin errors from the lowest ranked bins as long as the total sum of squared error removed (r_{sub}) is less than the merge threshold
- Scale up the remaining errors by a factor $\sqrt{(}1/(1 - r_{sub})$, which ensure the total bin error from all processes remains the same

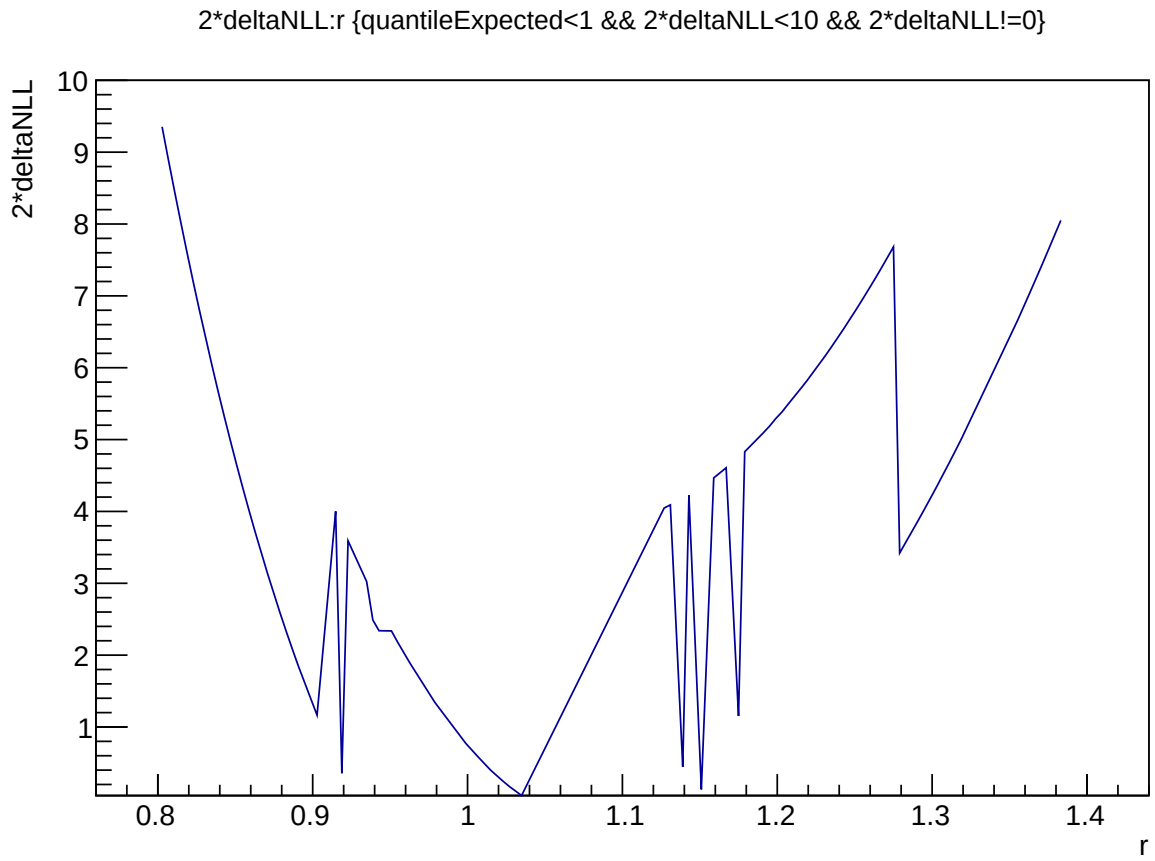
Different treatment of MC statistics continued

Combine

- Barlow-Beeston-lite approach: attempt to assign a single nuisance parameter to scale the sum of the process yields in each bin, constrained by the total uncertainty instead of requiring separate parameters, one per process
- effective number of unweighted events: $n_{tot}^{eff} = n_{tot}^2 / e_{tot}^2$
- If $n_{tot}^{eff} <= n_{threshold}$ separate uncertainties will be created for each process
- If $n_{tot}^{eff} > n_{threshold}$: single Gaussian-constrained Barlow-Beeston-lite parameter is created that will scale the total yield in the bin
- ML of each nuisance parameter has a simple analytic form $\rightarrow$ for models with large numbers of bins this can reduce the fit time and increase the fit stability

Comparison of MC treatment

- still ongoing $\rightarrow$ several small technical problems
- will be shown in next group meeting
- in general: multiple minima on likelihood scan



Conclusions

Combine harvester

- Production of datacard with harvester implemented
- Harvester implemented on a high level
- Comparison of different MC statistics treatment ongoing

Multiclass

- First scripts written and usable on SWAN
- Validation ongoing
- MVA variable validation would be helpful
- Little statistics for single top channed sd