Multi-Class Classification Methodology and Application in HEP

Stefan Wunsch
stefan.wunsch@cern.ch

January 18, 2017

Multi-Class Classification Methods and Methodology

TMVA Multi-Class Methods

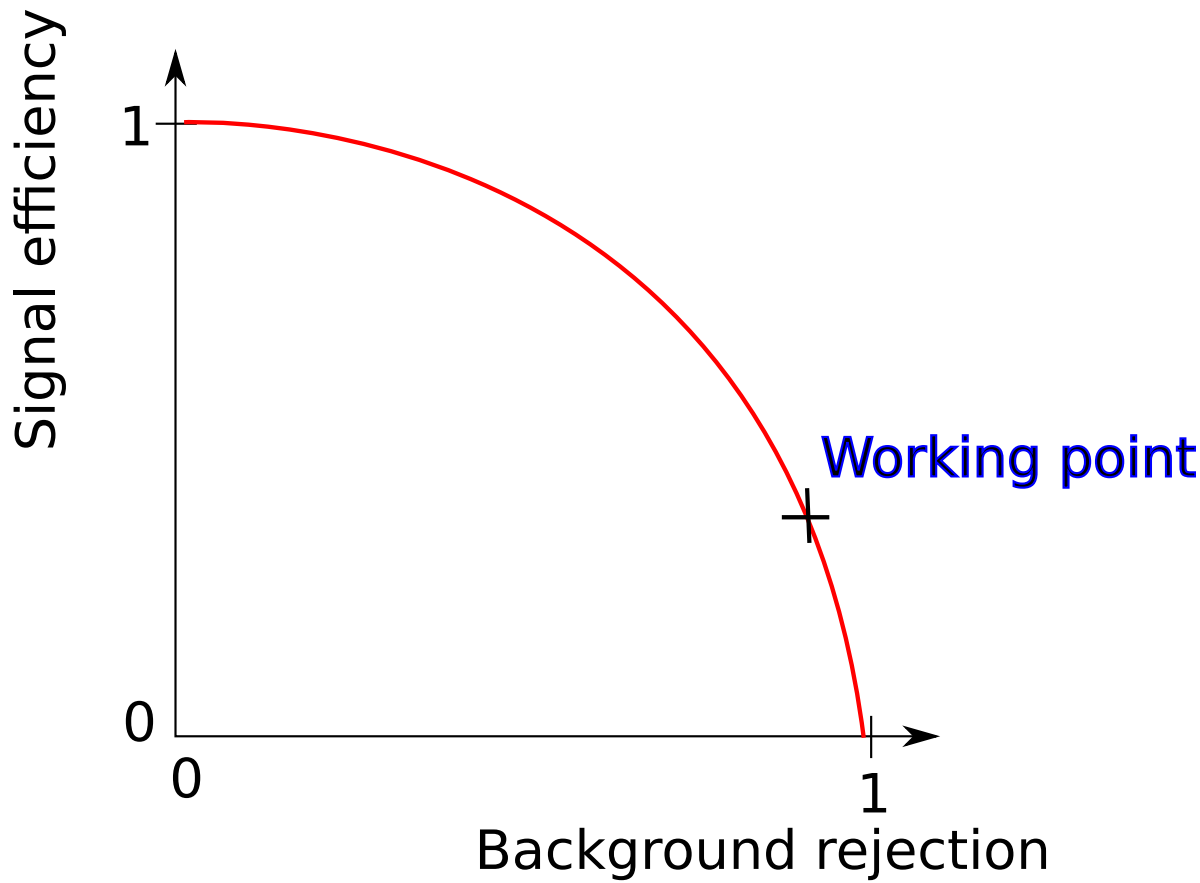
- **Status October 2013** (TMVA Users Guide)

MVA method	Classi-fication	Regress-ion	Multi-class/target classification	Multi-class/target regression
Cut optimisation	•	○	○	○
Likelihood	•	○	○	○
PDE-RS	•	•	○	○
PDE-Foam	•	•	•	•
k-NN	•	•	○	•
H-Matrix	•	○	○	○
Fisher	•	○	○	○
LD	•	•	○	○
FDA	•	•	○	○
MLP	•	•	○	•
TMlpANN$^{(\star)}$	•	○	○	○
CFMlpANN	•	○	○	○
SVM	•	○	○	○
BDT	•	•	○	○
RuleFit	•	○	○	○

TMVA Multi-Class Methods (2)

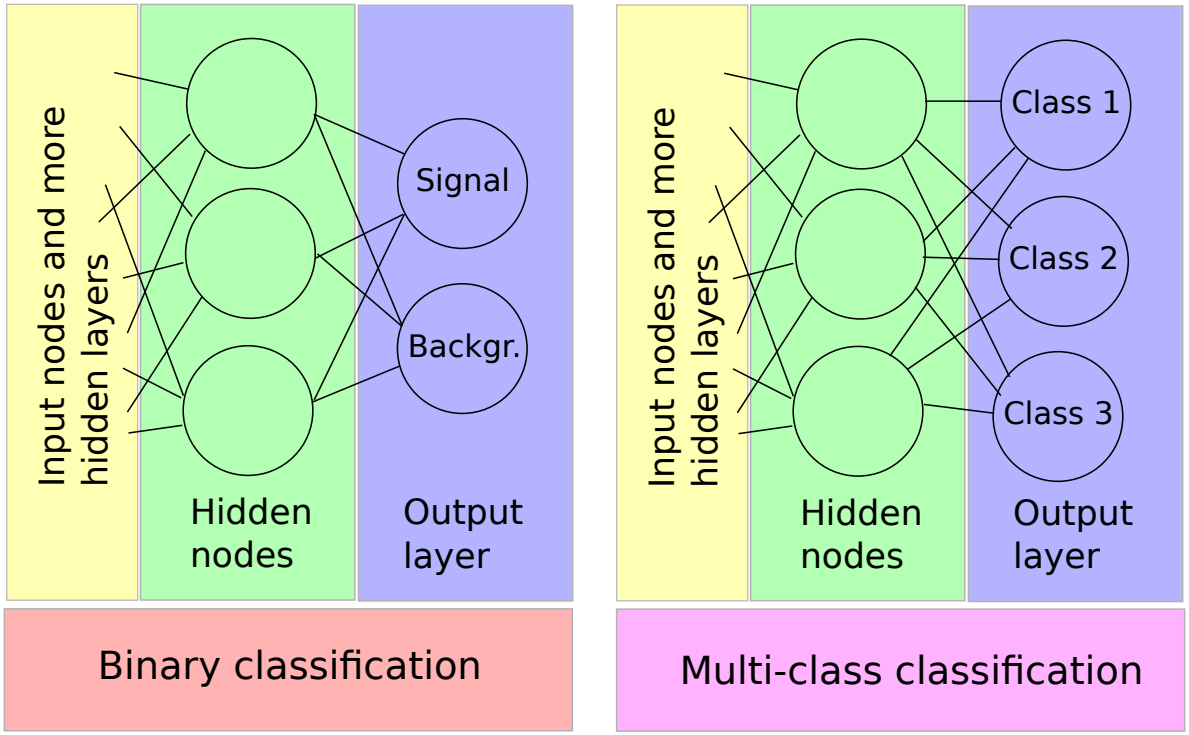
- **Status January 2017** (ROOT master branch)
- **TMVA methods capable of multi-class classification**:
  - BDT
  - Probability density estimator (PDE-Foam)
  - Function discriminant analysis (FDA)
  - DNN (TMVA implementation)
  - DNN (PyMVA Keras interface for Theano and Tensorflow)
  - Random Forest (PyMVA scikit-learn interface)
  - Gradient Tree Boosting (PyMVA scikit-learn interface)
  - AdaBoost Classifier (PyMVA scikit-learn interface)
- Satisfying list of methods
- **Weak point:** Missing multi-class LDA/Fisher method
  $\rightarrow$ Excellent performance reference for any other method
  - Unfortunately scikit-learn LDA/QDA implementations do not support event weights

Binary Classification vs Multi-Class Classification



- **Main difference:** Losing the ability to apply working points
  - Excessively used in physics!

Binary Classification vs Multi-Class Classification (2)



- Neural network example
- Softmax activation $\rightarrow$ Outputs represent probabilities
  $\rightarrow$ Output is normed to 1
- N output nodes $\rightarrow$ $N-1$ degrees of freedom

Binary Classification vs Multi-Class Classification (3)

- **Solution:** Applying modified event weights and select always output node with highest probability
  - Moving the working point selection to the training step
  - Tell the optimizer your preferences beforehand
- **Details:**
  - **High signal efficiency:** Multiply event weights of target class with large value

    It is important to find as many events as possible of this class, no matter how many false classifications we do.
  - **High background rejection:** Multiply event weights of target class with small value

    It is more important to get the other classes right than classifying events of this class correctly.

Binary Classification vs Multi-Class Classification (4)

- **Previous slide:** Categorical cross-entropy loss function

- **Outlook:** Is it possible to define a loss function that incorporates the working point and can be optimized on it?

TMVA: Genetic Fitter Cuts

- **Special case:** Separate signal and background (only two classes) with multi-class MVA method
- **Task:** Find best working point on all output nodes, which maximizes signal efficiency and background rejection
- Implemented in `TMVA::Factory::EvaluateAllMethods`

```
Calculating best set of cuts for class Signal
FitterBase : Optimisation, please be patient ...
           : Elapsed time: 3.4 sec
           :   cutValue[Signal] = 0.454085;
           :   cutValue[Background_0] = -0.521825;
           :   cutValue[Background_1] = -0.829396;
           :   cutValue[Background_2] = -0.46266;
```

Evaluation Metrics

- **Simple metric:** Global accuracy
  - How many events are correctly classified?
- **More sophisticated:** Migration matrix
- **Migration matrix:** Fill event weights in matrix representation
- Example: $H \to \tau\tau$ event classification

```
                 CLASSIFIED CLASS
                 Ztt      TT      W+Jets Zll     ggH     qqH
TRUE  Ztt        [2.2924 0.0005 0.0983 0.0145 0.0031 0.0001]
CLASS TT         [0.0621 0.0407 0.0956 0.0038 0.0015 0.0001]
      W+Jets     [0.2133 0.0009 0.3933 0.0124 0.0006 0.0000]
      Zll        [0.1004 0.0002 0.0778 0.1077 0.0014 0.0000]
      ggH        [0.0107 0.0000 0.0022 0.0009 0.0006 0.0000]
      qqH        [0.0009 0.0000 0.0003 0.0000 0.0002 0.0001]
```

Migration Matrix: Efficiency Representation

- Normalize rows respective to principal axis
- Alternative: Normalize on sum of event weights in each row

		CLASSIFIED CLASS					
		Ztt	TT	W+Jets	Zll	ggH	qqH
TRUE	Ztt	[1.0000	0.0002	0.0429	0.0063	0.0014	0.0001]
CLASS	TT	[1.5269	1.0000	2.3511	0.0928	0.0360	0.0033]
	W+Jets	[0.5425	0.0023	1.0000	0.0316	0.0015	0.0000]
	Zll	[0.9319	0.0016	0.7226	1.0000	0.0128	0.0001]
	ggH	[18.437	0.0281	3.8743	1.4683	1.0000	0.0652]
	qqH	[7.2164	0.0356	2.2743	0.3801	1.5109	1.0000]

Migration Matrix: Purity Representation

- Normalize columns respective to principal axis
- Alternative: Normalize on sum of event weights in each column

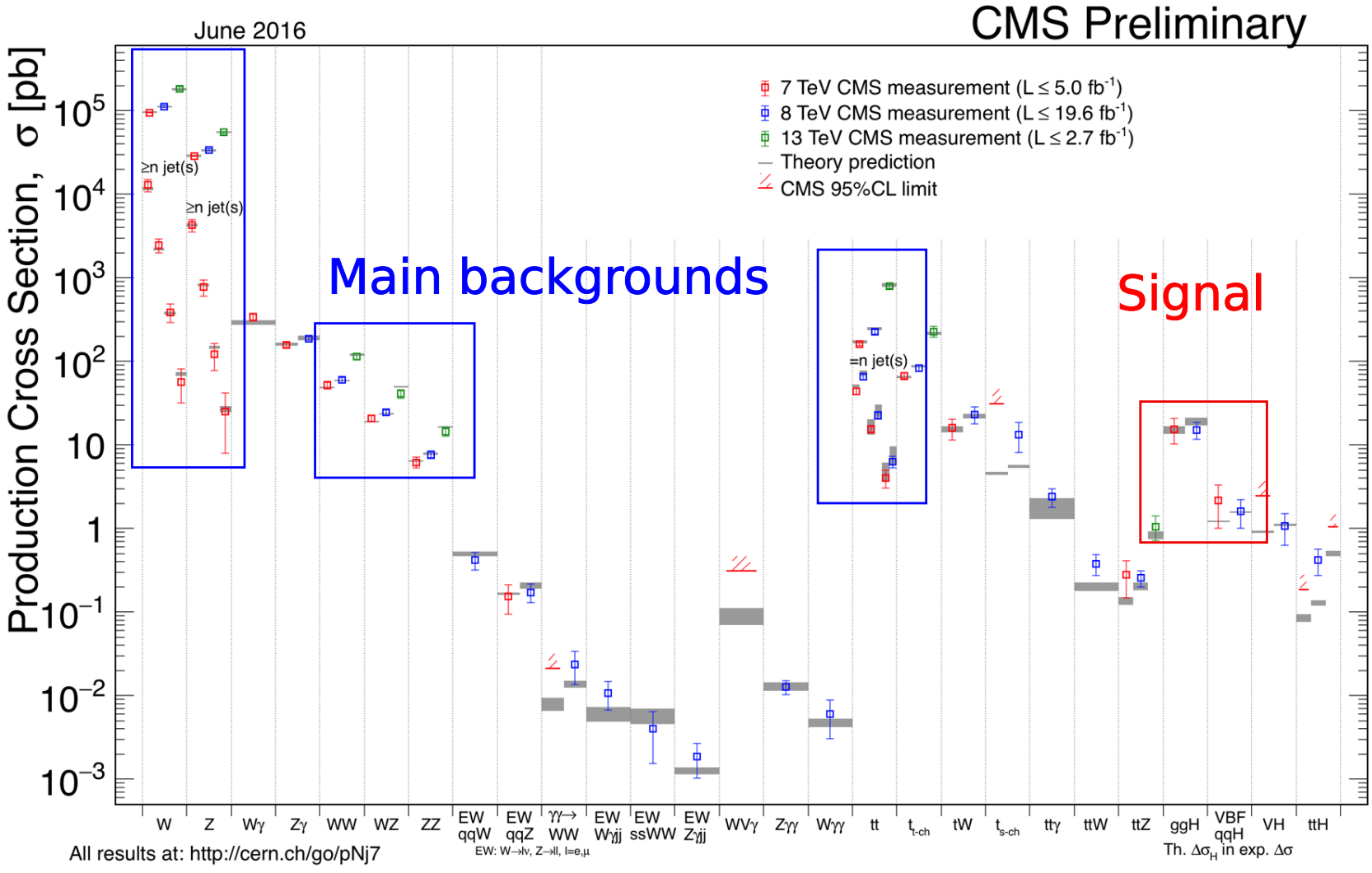
		CLASSIFIED CLASS					
		Ztt	TT	W+Jets	Zll	ggH	qqH
TRUE	Ztt	[1.0000	0.0116	0.2499	0.1343	5.4040	1.0402]
CLASS	TT	[0.0271	1.0000	0.2431	0.0350	2.5213	1.0994]
	W+jets	[0.0931	0.0221	1.0000	0.1154	0.9865	0.0000]
	Zll	[0.0438	0.0043	0.1979	1.0000	2.3722	0.1154]
	ggH	[0.0047	0.0004	0.0057	0.0079	1.0000	0.3067]
	qqH	[0.0004	0.0001	0.0007	0.0004	0.3210	1.0000]

HEP Application

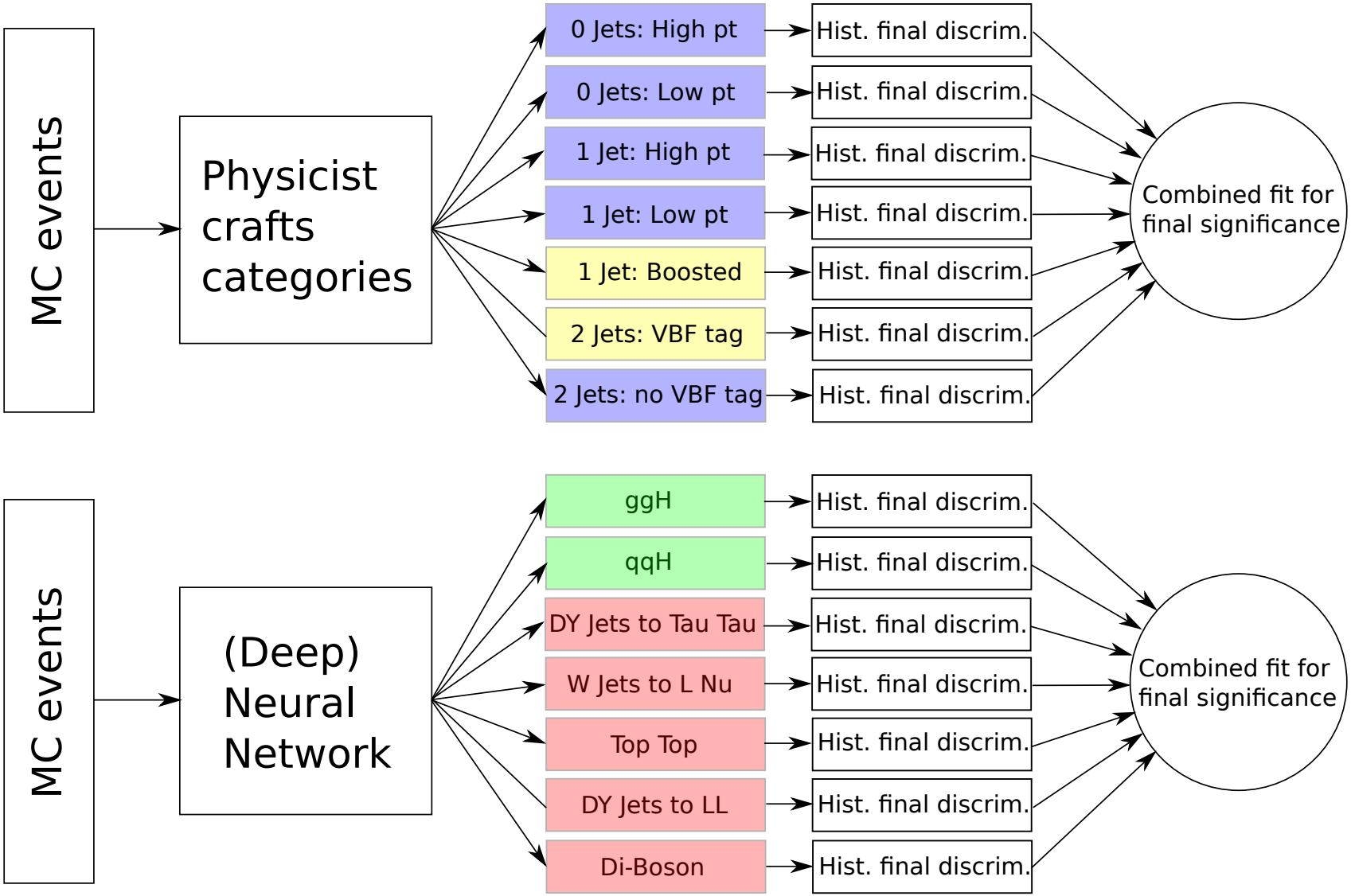
HEP Application: Event Classification

- **Application:** Event classification
  - Which process produced an event in data?
- **Analysis:** Standard model $H \to \tau\tau$
  - Discovery analysis
  - Trying to push signal significance further
  - Complicated because of tiny signal, neutrinos and jets in final state
- **Approach:** Classifying events in production processes
  - Signal classes: Gluon-fusion Higgs production and VBF Higgs production
  - Background classes: DY Jets, W+Jets, $t\bar{t}$, QCD, ...
  - Training on Monte Carlo
  - **Natural approach compared to conventional cut based analysis**

Production Processes

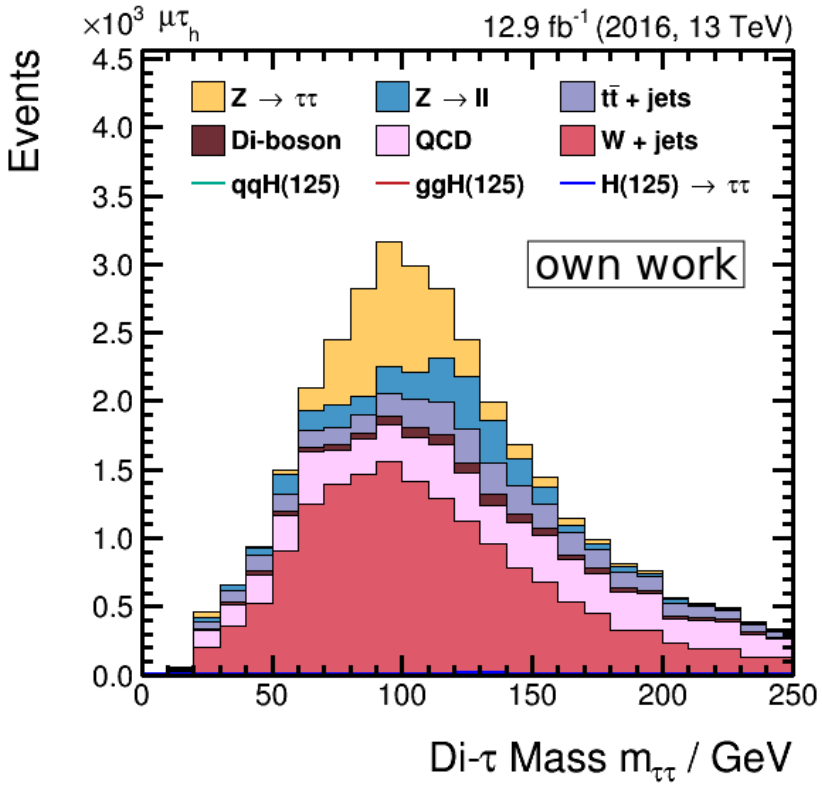


Standard Model $H \to \tau\tau$ MVA Event Classification

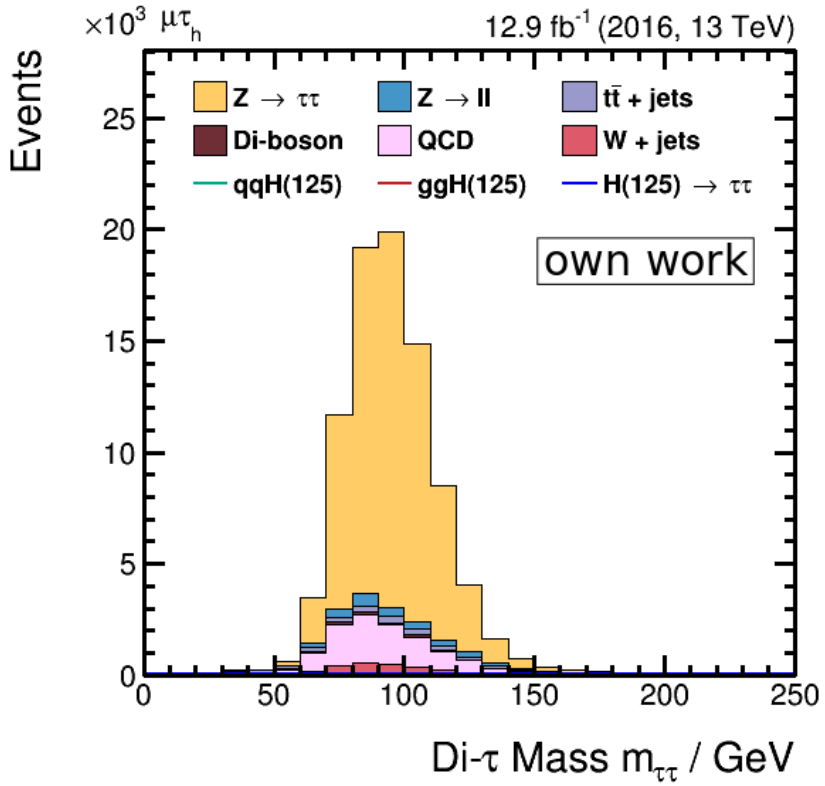


Examples: Background Classes

- Channel: Muon + hadronic Tau
- Plots contain only Monte Carlo events
- **MVA classification in production processes**



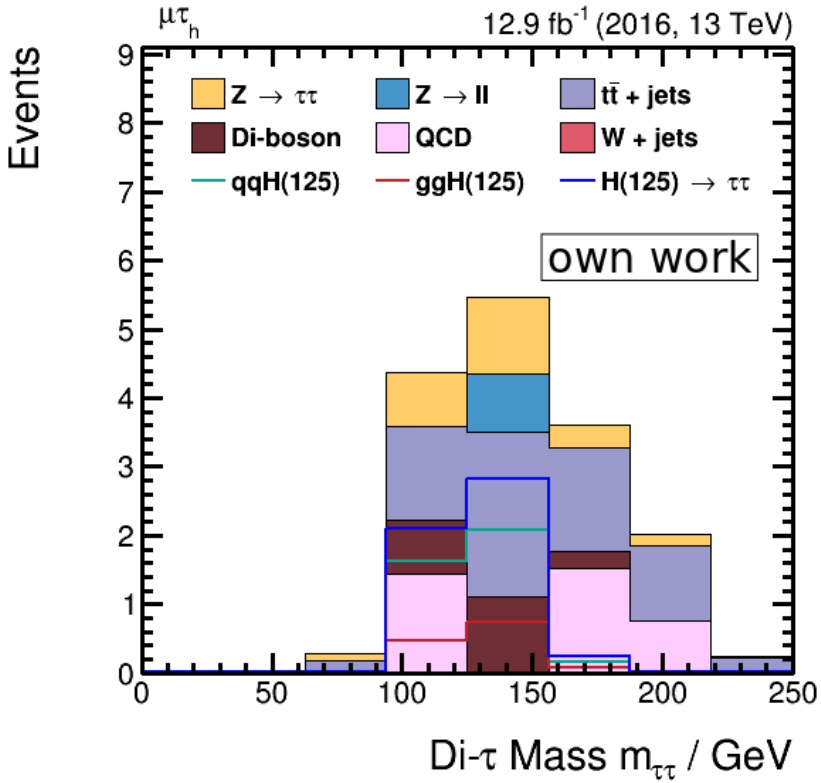
W+Jets



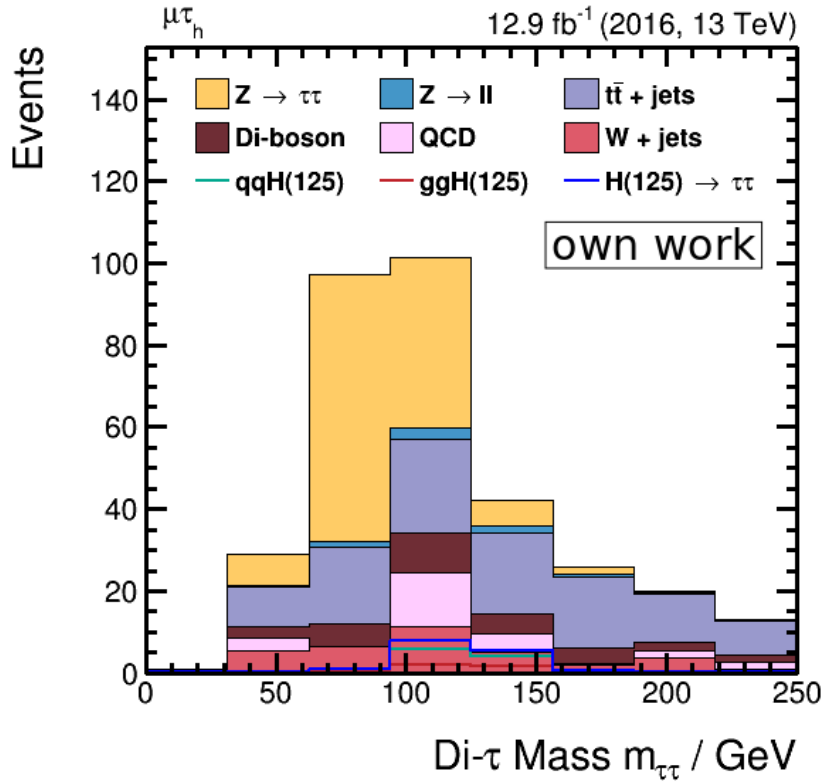
Z to 2 Taus

Examples: Signal Classes

- **Left:** MVA classification
- **Right:** Conventional cut based signal class
- **MVA:** Very pure signal class
  $\rightarrow$ Hard to achieve with cut based categorization



Higgs from VBF



2 Jets with VBF tag

Recap: Working Point

- Shown plots are tuned on **high purity**
  $\rightarrow$ Assign **event weights as small as possible** to signal classes

- Discovery analysis is mainly driven by pure classes (significant bins)

- **Highly versatile** event classification technique

- **Problem:** Too small event weights prune classes (output nodes) completely

Outlook

- MVA multi-class event classification looks promising
- **But:** Nothing worth without uncertainties $\rightarrow$ Working on it!

- Technical problems mostly solved
- Improved loss functions?