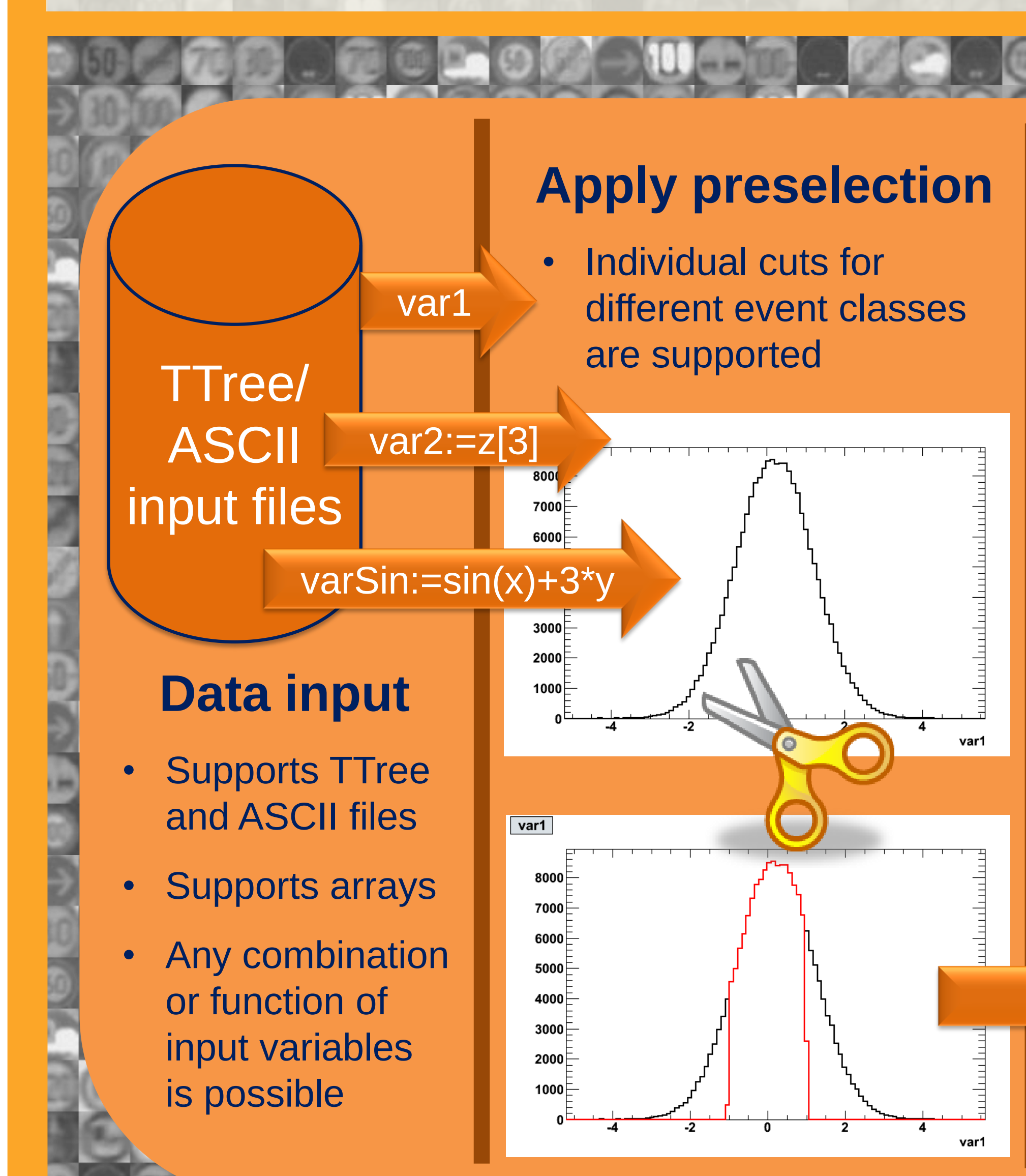
TMVA 4 – Toolkit for Multivariate Data Analysis in ROOT

TMVA core developer team: A. Höcker, P. Speckmayer, J. Stelzer, J. Therhaag, E. v. Törne, H. Voss

TMVA provides a large set of sophisticated multivariate analysis techniques for both classification and regression tasks in HEP. All methods are embedded in a powerful yet user-friendly framework capable of handling the preprocessing of the input data as well as the evaluation and comparison of the MVA algorithms. TMVA is fully integrated in the popular ROOT data analysis framework.

preprocessing



Apply preselection

- Individual cuts for different event classes are supported

Data input

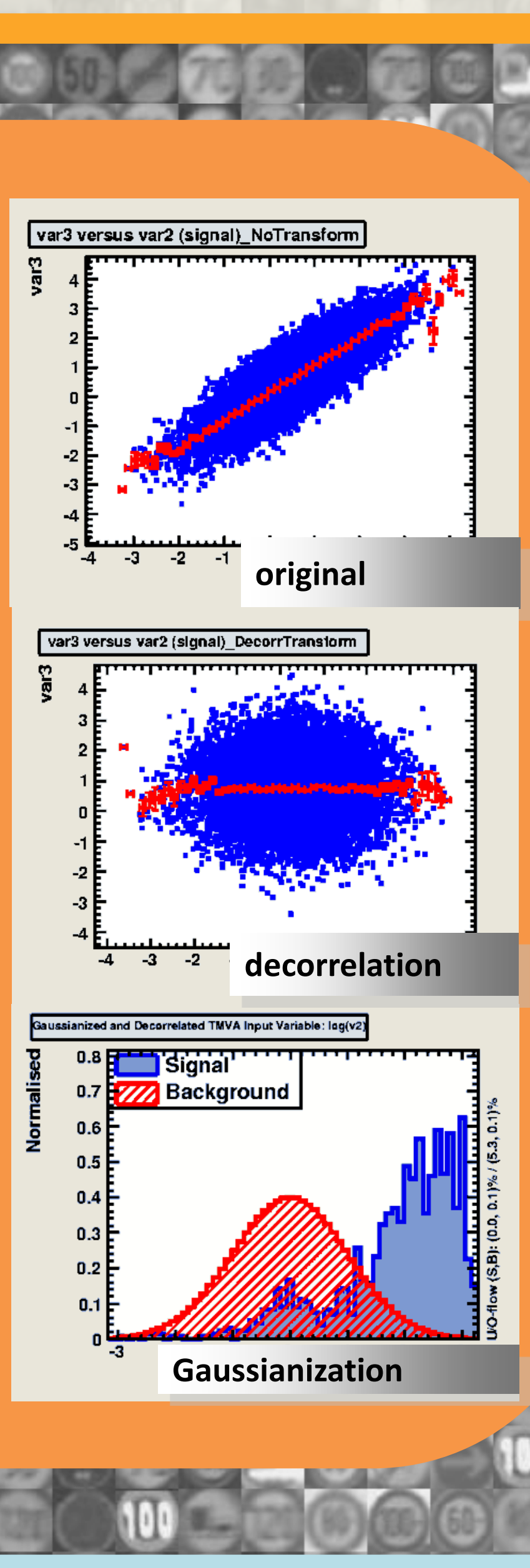
- Supports TTree and ASCII files
- Supports arrays
- Any combination or function of input variables is possible

Use event weights

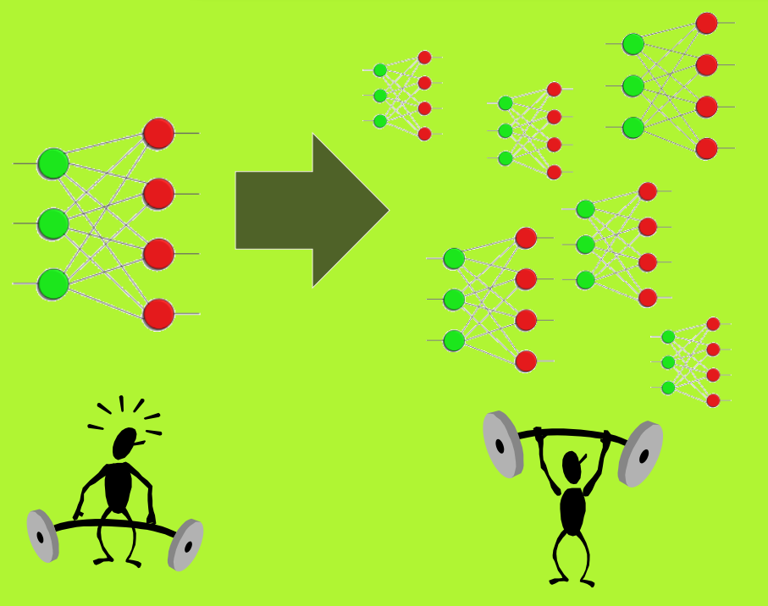
- Supports event-by-event weights, weights for individual files/trees and weights for different classes

Transformations

- Supports individual transf. for each method
- Transformations can be chained
- NEW: Transformation of variable subsets
- TMVA knows:
  - Normalisation
  - Decorrelation
  - Principal component analysis
  - Gaussianisation



meta-methods

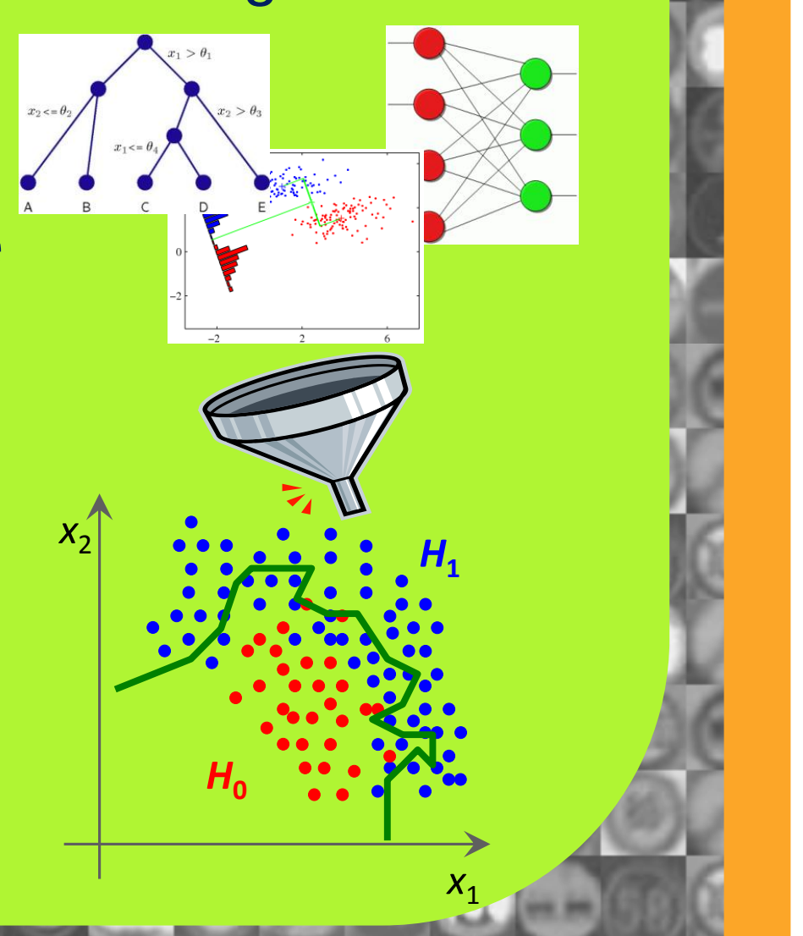


Generalized boosting

- TMVA4 can not only boost decision trees, but any MVA method available
- Ensemble of “weak learners” often outperforms complicated algorithms

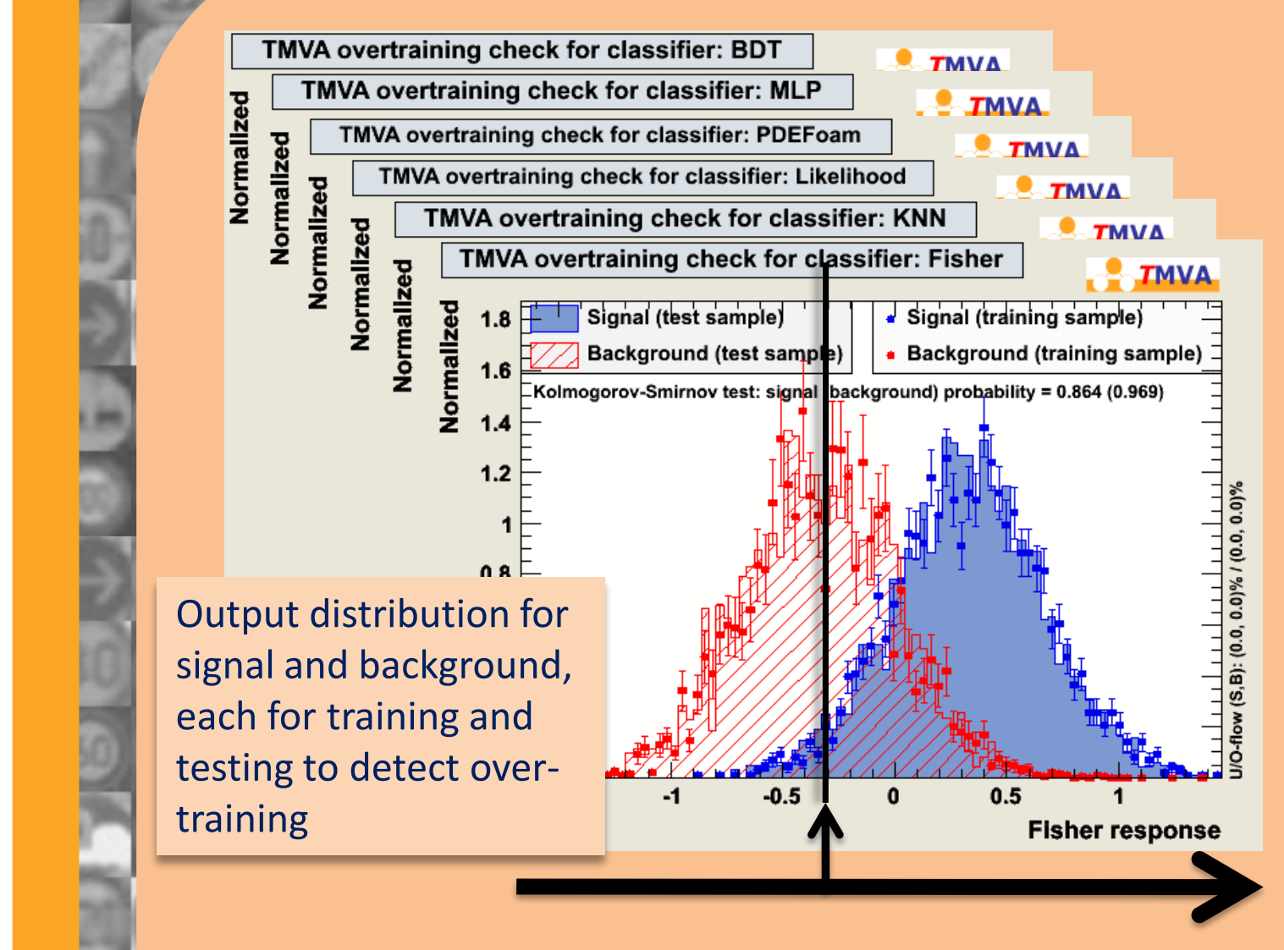
Classifier combination

- TMVA4 can use different methods in different parts of the input phase-space, taking into account characteristic features of the underlying data
- Combine all methods to obtain a powerful meta-method which is optimally adjusted to the problem



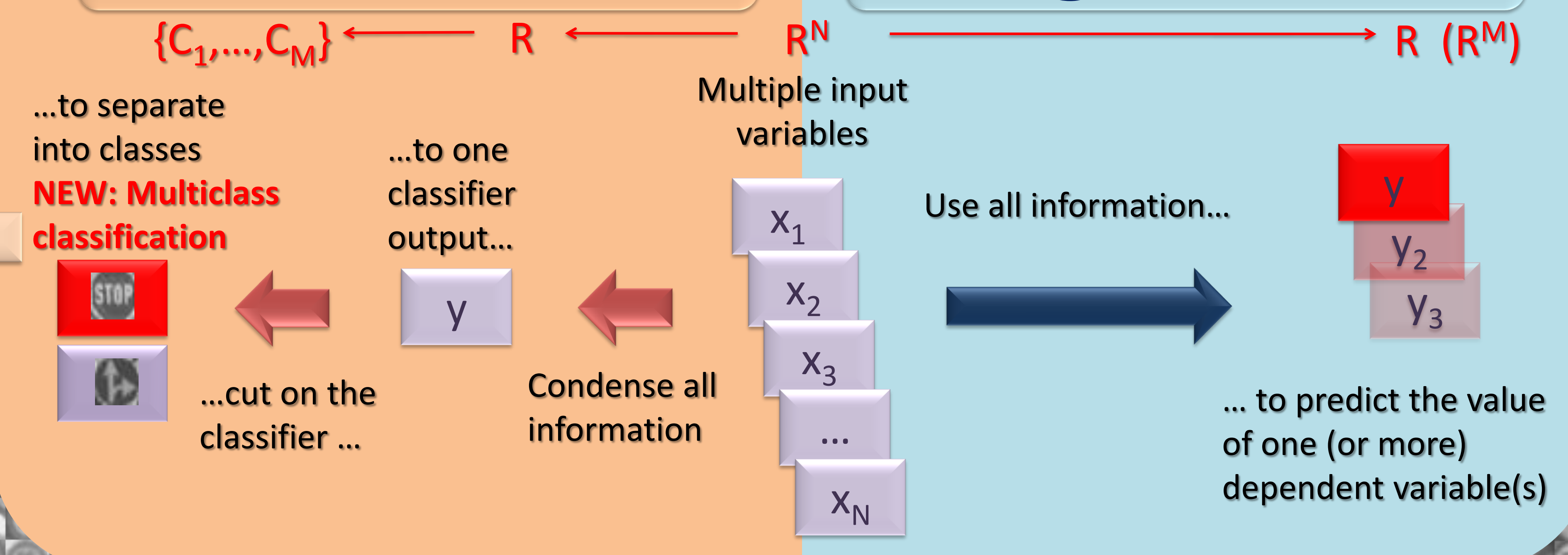
classification

regression



Output distribution for signal and background, each for training and testing to detect over-training

Scan over classifier output variable values creates set of (ε_{sig},1-ε_{bkgd}) points -> ROC curves



…to separate into classes

NEW: Multiclass classification

…to one classifier output…

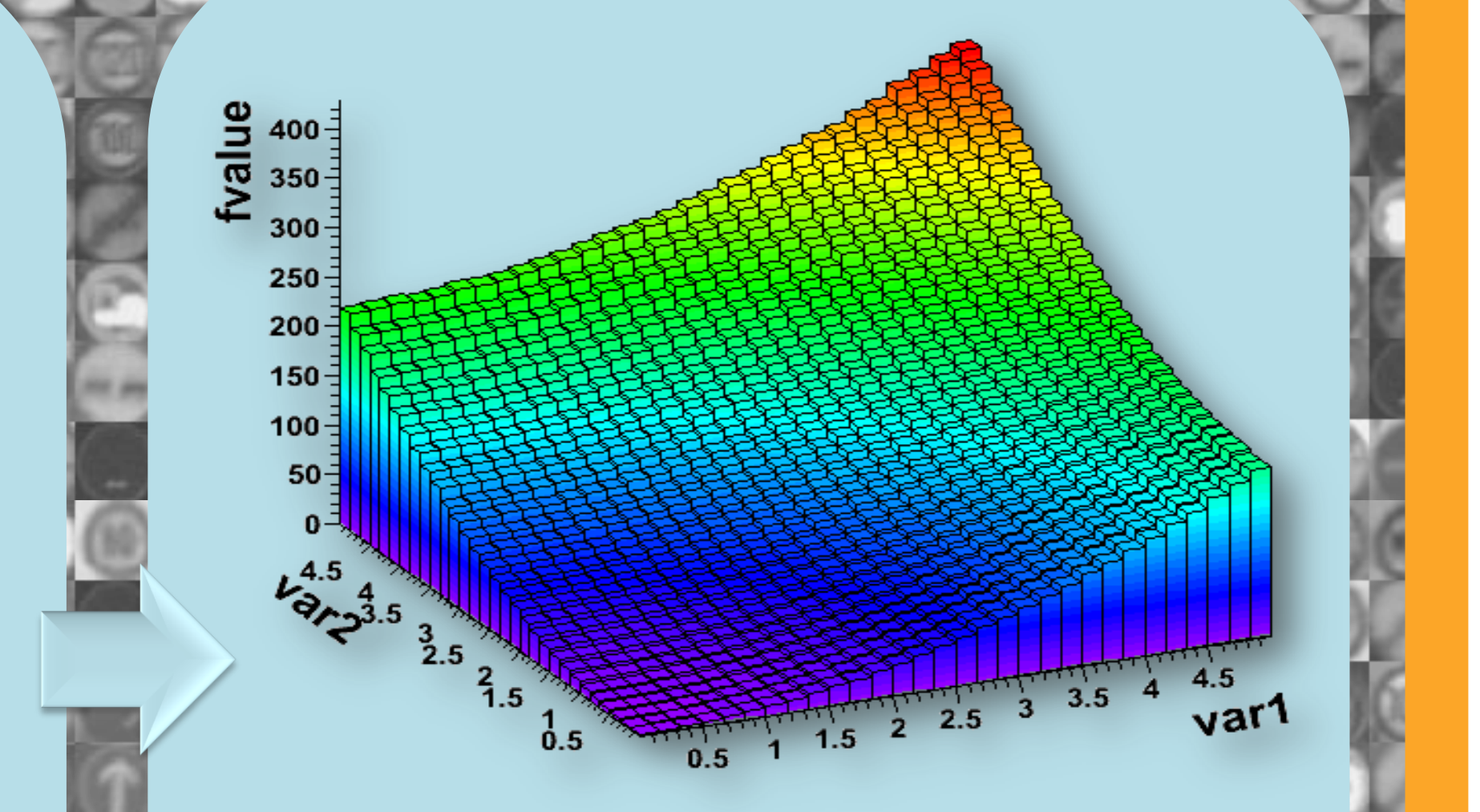
…cut on the classifier …

Condense all information

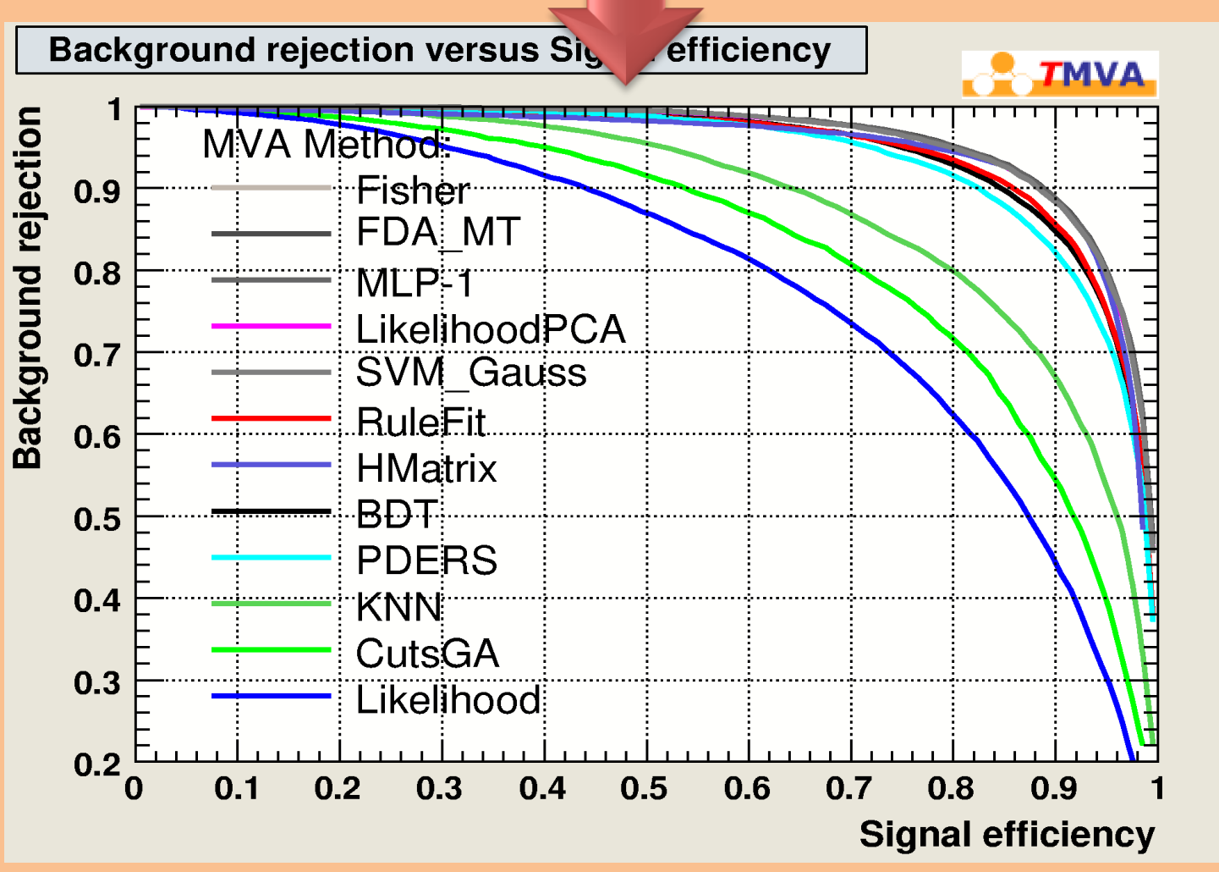
Multiple input variables

Use all information…

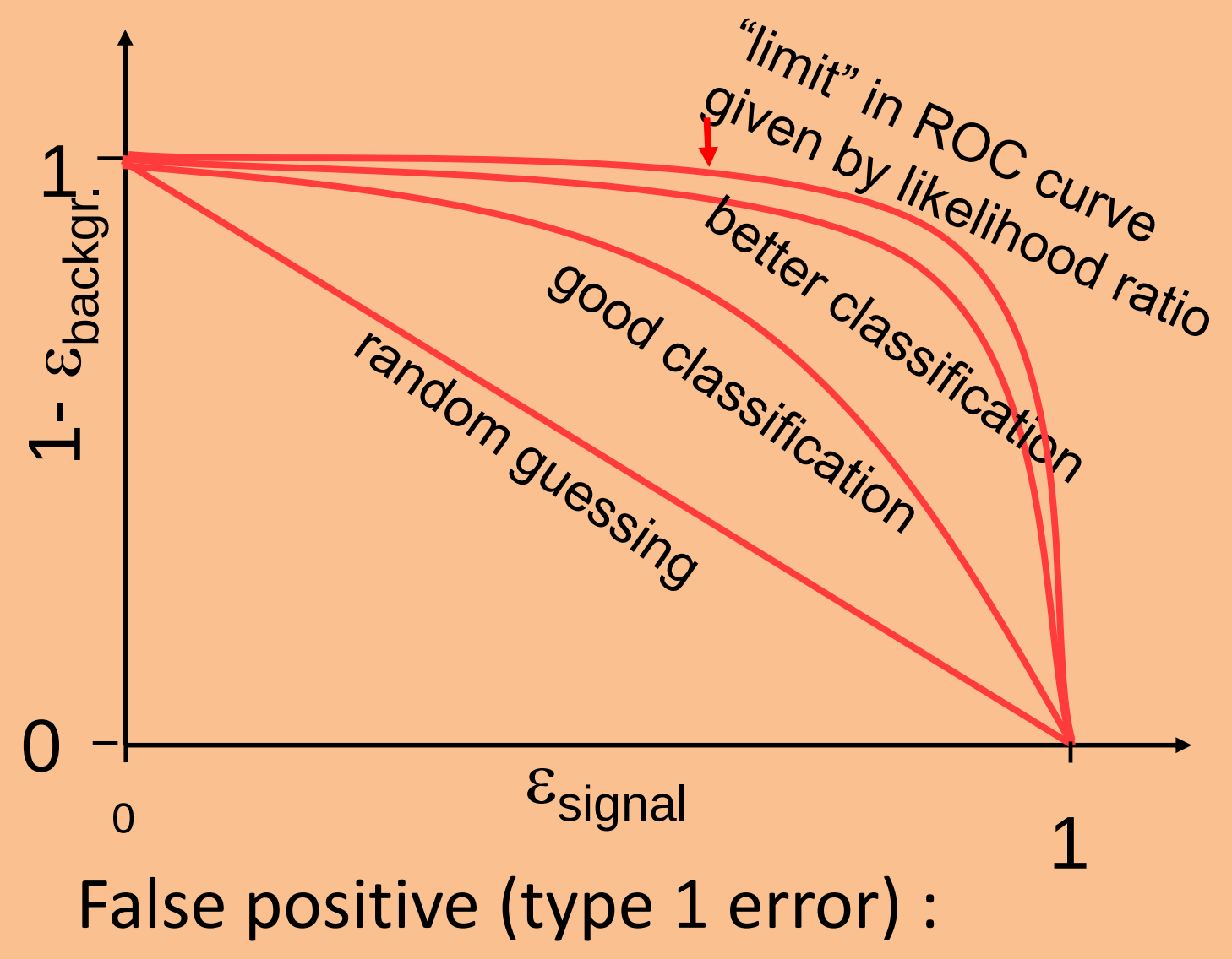
… to predict the value of one (or more) dependent variable(s)



Example: Estimation of target as a function of two variables



ROC curve describes performance of a binary classifier by plotting the false positive vs. the true positive fraction



False positive (type 1 error) :

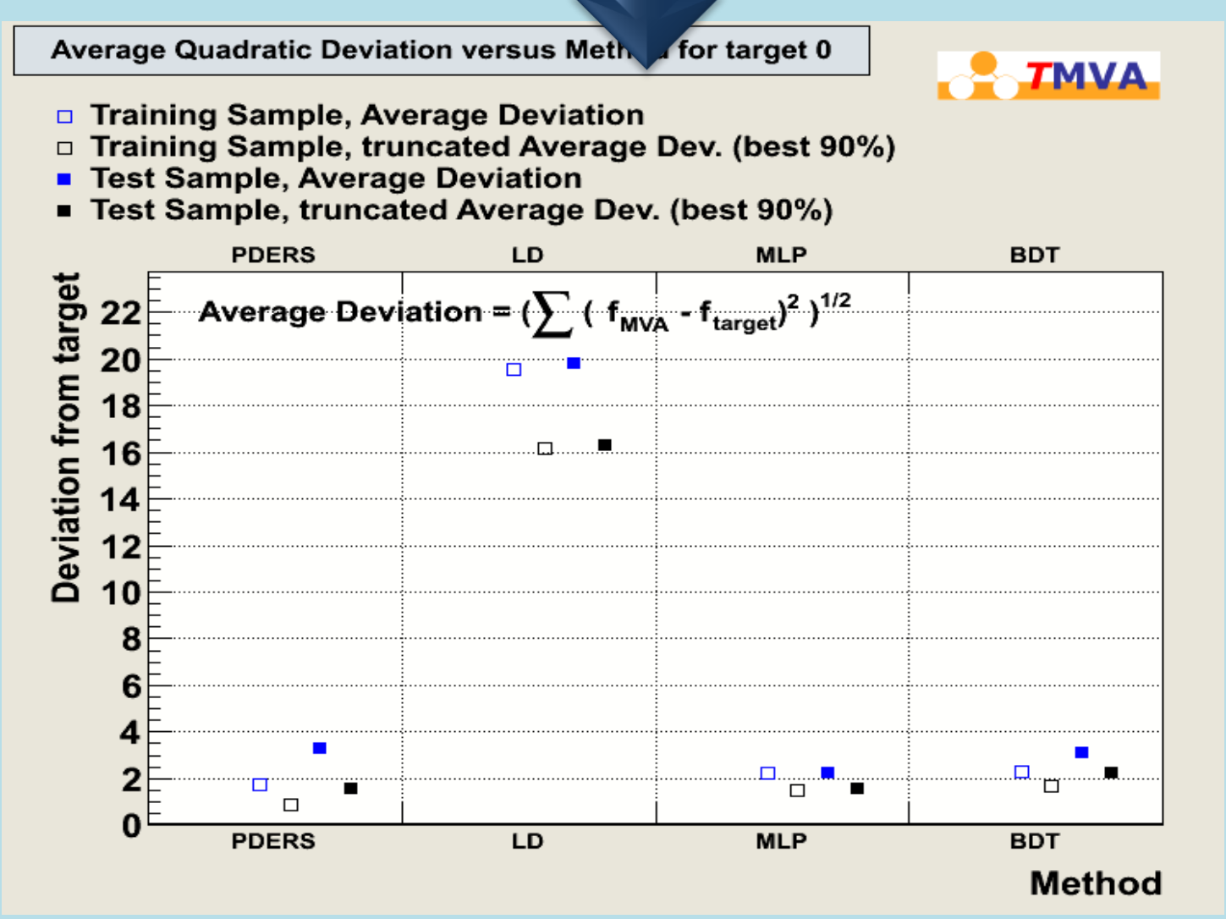
classify background event as signal

-> loss of purity

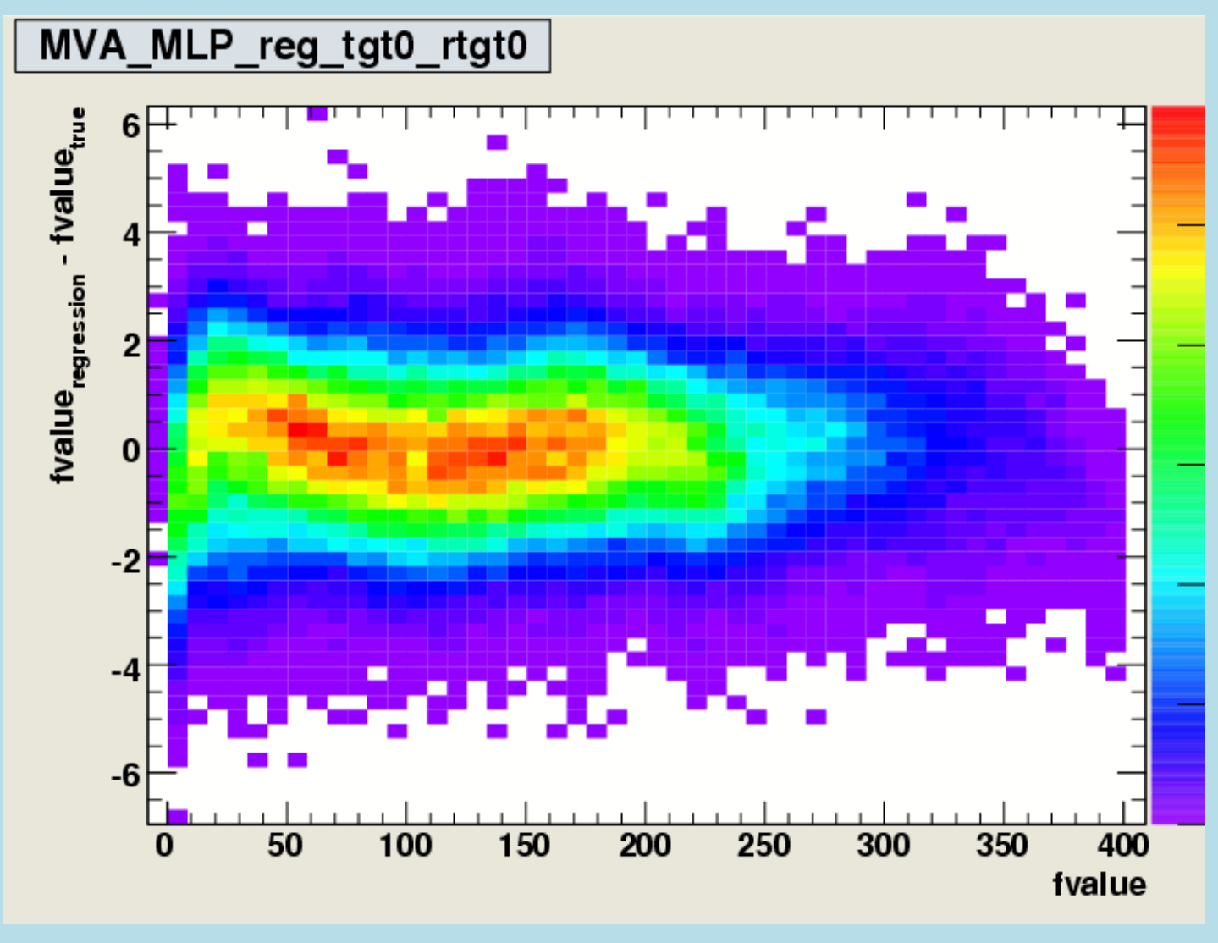
False negative (type 2 error) :

fail to identify a signal event as such

-> loss of efficiency



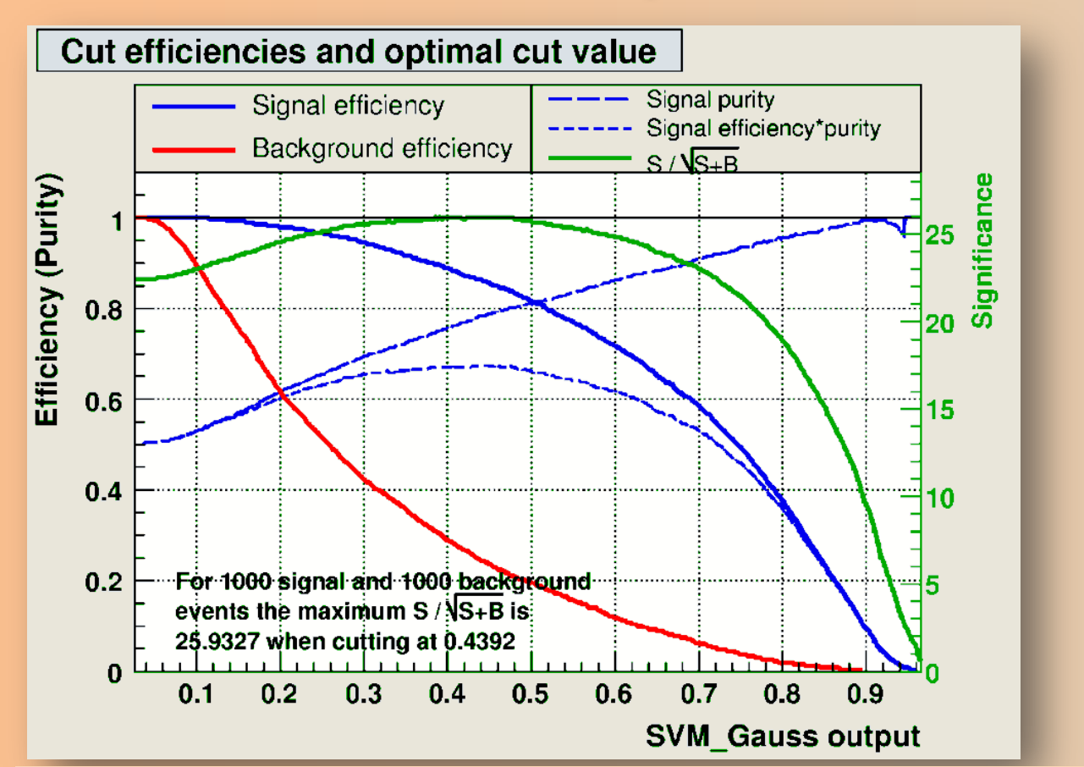
Show average quadratic deviation of true and estimated value for both training a and testing



Show estimated value minus true value as a function of the true value

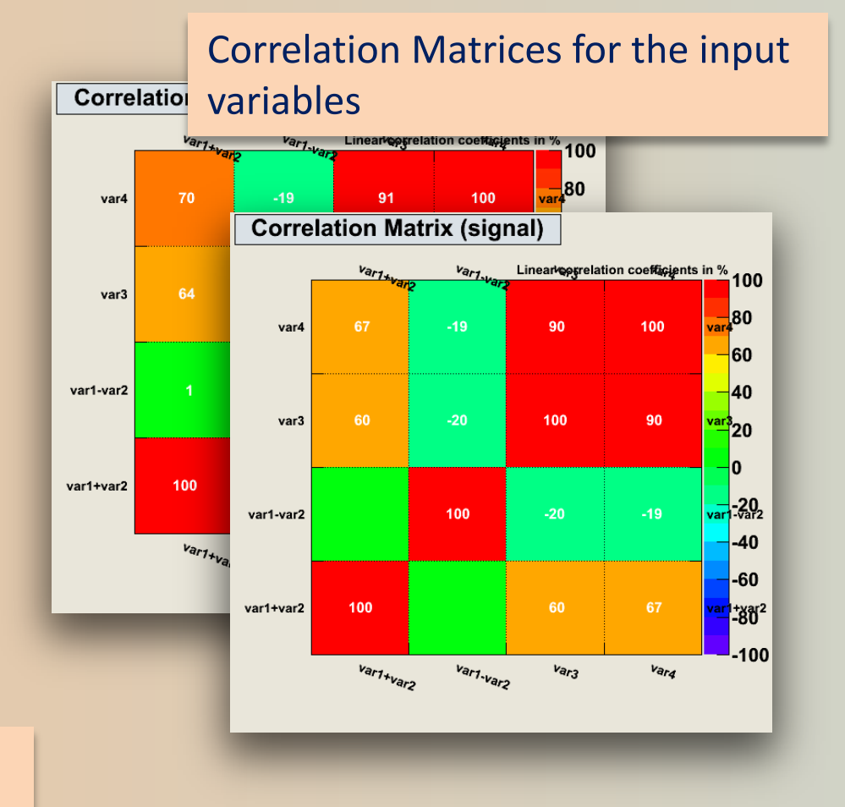
evaluation & assessment

TMVA provides many evaluation macros to produce plots and numbers which help the user to decide on the best classifier and settings for an analysis

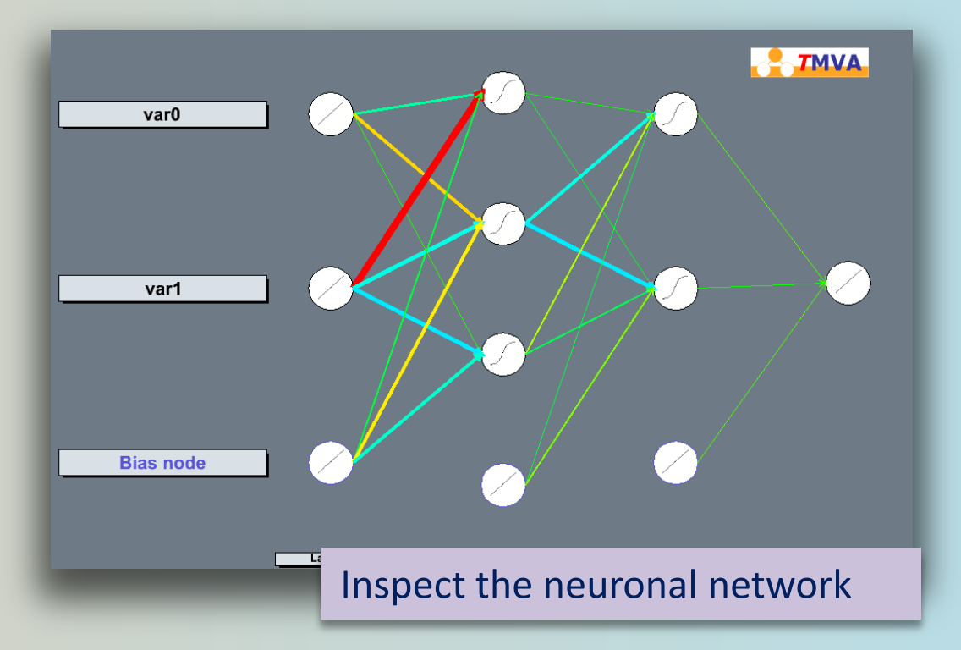


Working Point: Find optimal cut on a classifier output (=optimal point on ROC curve) depending on the problem:

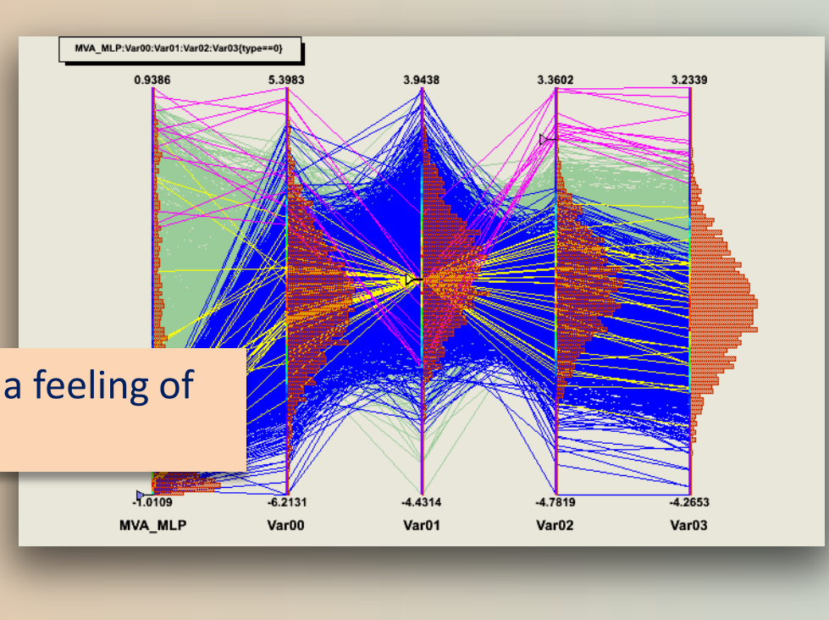
- Cross section measurement: maximum of S/√(S+B)
- Signal Search: maximum of S/√(B)
- Precision measurement: high purity
- Trigger selection: high efficiency



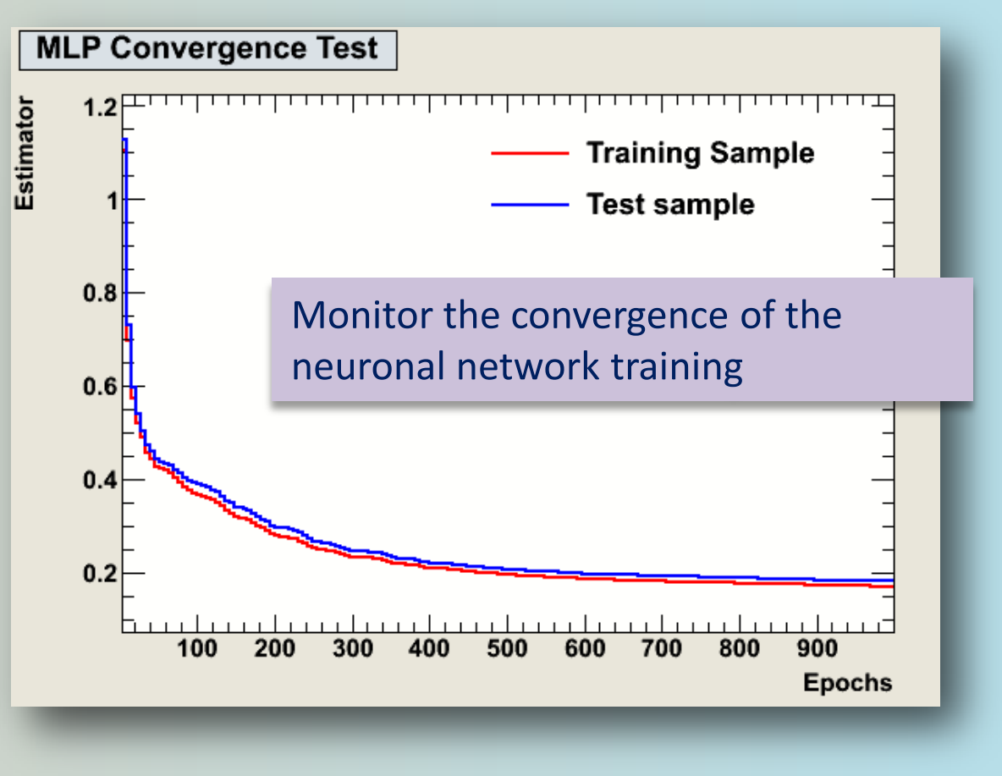
Correlation Matrices for the input variables



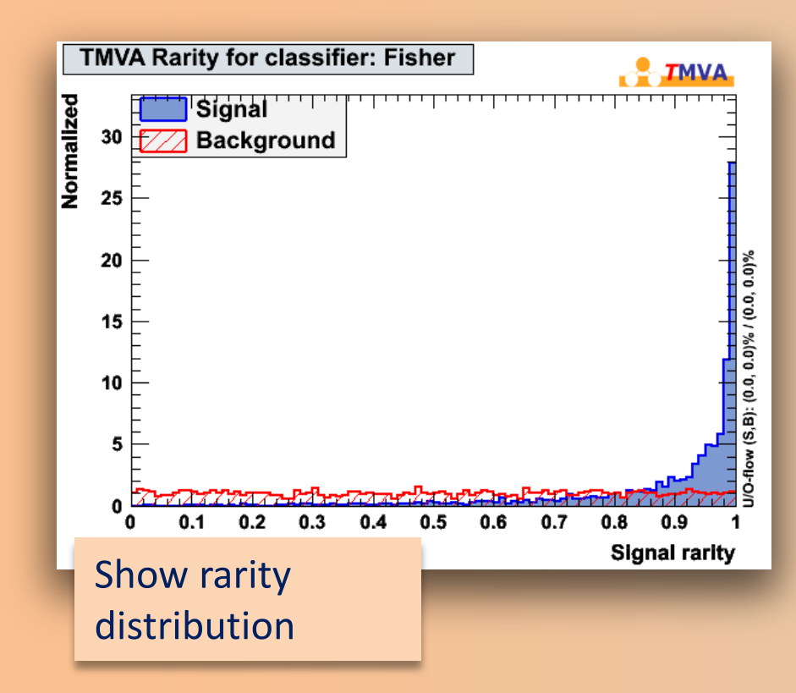
Inspect the neuronal network



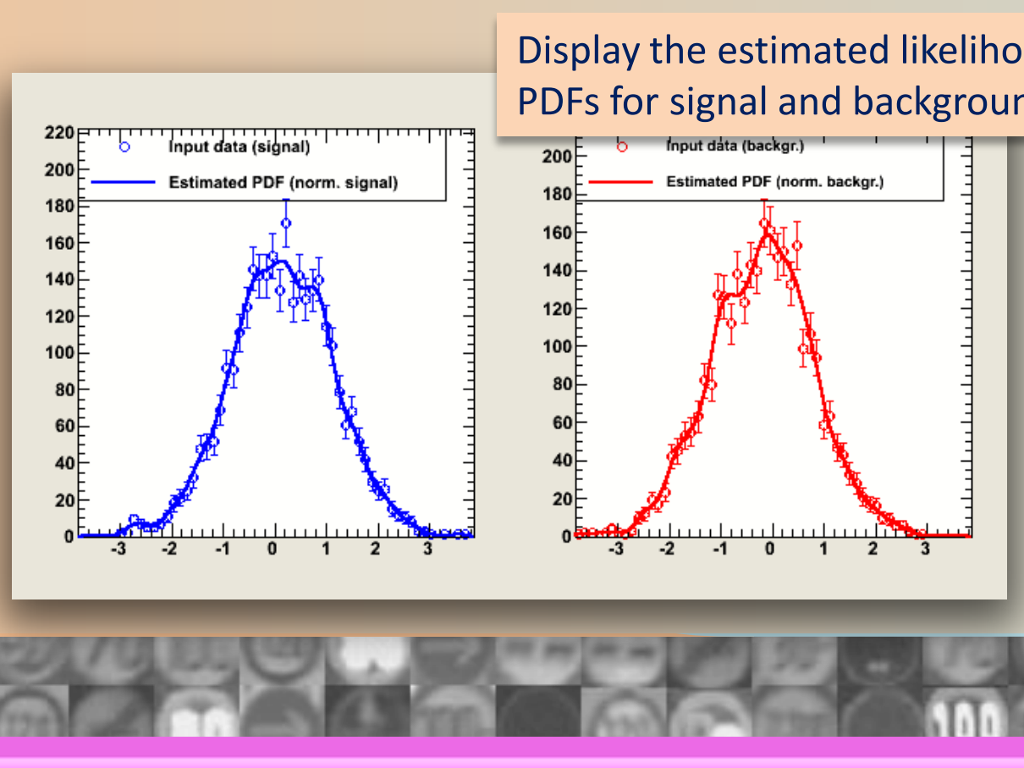
Parallel coordinates (give a feeling of the variable correlations)



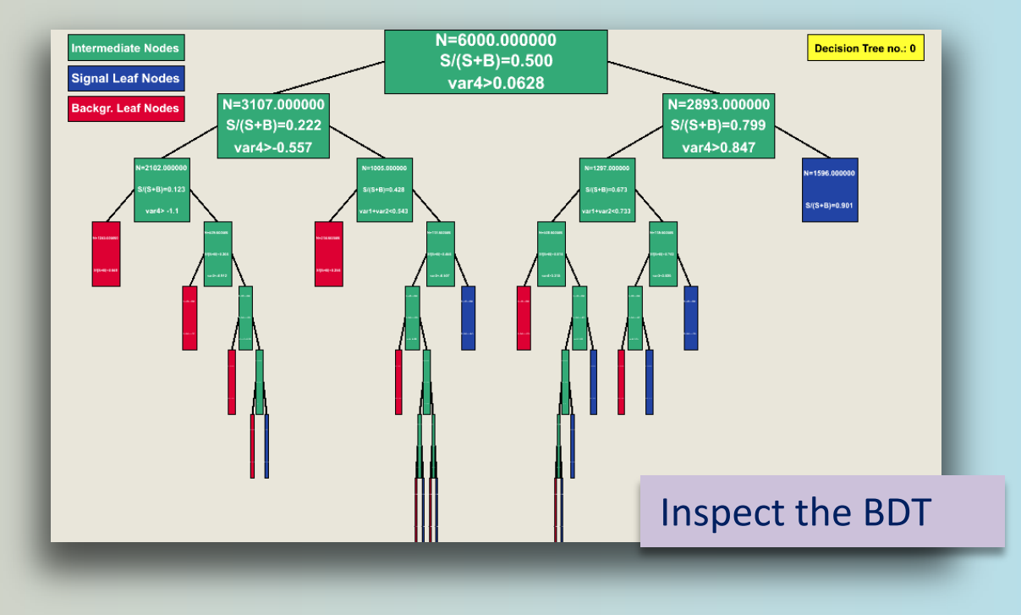
Monitor the convergence of the neuronal network training



Show rarity distribution



Display the estimated likelihood PDFs for signal and background



Inspect the BDT

summary & new developments

- Many MVA methods implemented
- One common platform/interface for all MVA methods
- Wide range of data pre-processing capabilities
- Common input and analysis framework (ROOT scripts)
- Train and test all methods on same data sample and evaluate consistently

Criteria		TMVA classifier overview								
		Cuts	Likelihood	PDERS/ k-NN	H-Matrix	Fisher	MLP	BDT	RuleFit	SVM
Performance	no / linear correlations	😐	😊	😊	😐	😊	😊	😐	😊	😊
	nonlinear correlations	😐	☹	😊	☹	☹	😊	😊	😐	😊
Speed	Training	☹	😊	😊	😊	😊	😐	☹	😐	☹
	Response	😊	😊	☹ 😐	😊	😊	😊	😐	😐	😐
Robustness	Overtraining	😊	😐	😐	😊	😊	☹	☹	😐	😐
	Weak input variables	😊	😊	☹	😊	😊	😐	😐	😐	😐

- **Automatic tuning of MVA methods to assist the user and optimize performance**
- Cross validation to make optimal use of the available input data
- **Multiclass option for all methods**
- Flexible variable transformations
- Extended set of example scripts to familiarize the user with the features and options of TMVA