# Statistical Tools in Collider Experiments

## Multivariate analysis in high energy physics

Pauli Lectures - 06/02/2012

**Nicolas Chanon - ETH Zürich**

ETH Institute for Particle Physics

# Main goals of these lessons

- Have an understanding of what are **multivariate analyses**

- How they are used in **high energy physics**

- Answer to the questions : what is a **neural network** ? a **boosted decision tree** ? what are the multivariate methods currently used in HEP ?

- Become familiar with problems related with **training and application** of multivariate methods

- Be aware of the **systematic uncertainties** related to multivariate techniques

- Be able to understand the results of **new physics searches at Tevatron or LHC** in the form where they are presented usually, and how they were produced

# Introductory comments

- In these lectures, examples will be mainly taken from Higgs boson searches at LHC

- Will focus on multivariate methods commonly used in the high energy physics community

- Theory will be addressed as a tool for practical usage

# Exercises

- Proposed exercises will follow the progress of the lecture

- Problem inspired by Higgs searches in H->2photons channel at LHC

- **Goal** : be able to estimate the sensitivity of a search for a small peak over a huge background, using multivariate methods

- **3 exercises** :
    - Setting up Root and TMVA environment, TMVA basics
    - Using a MVA method inside the analysis
    - Estimation of analysis sensitivity

# Outline

1. Introduction
2. Multivariate methods
3. Optimization of MVA methods
4. Application of MVA methods in HEP
5. Understanding Tevatron and LHC results

# Lecture 1. Introduction

# Content of this lecture

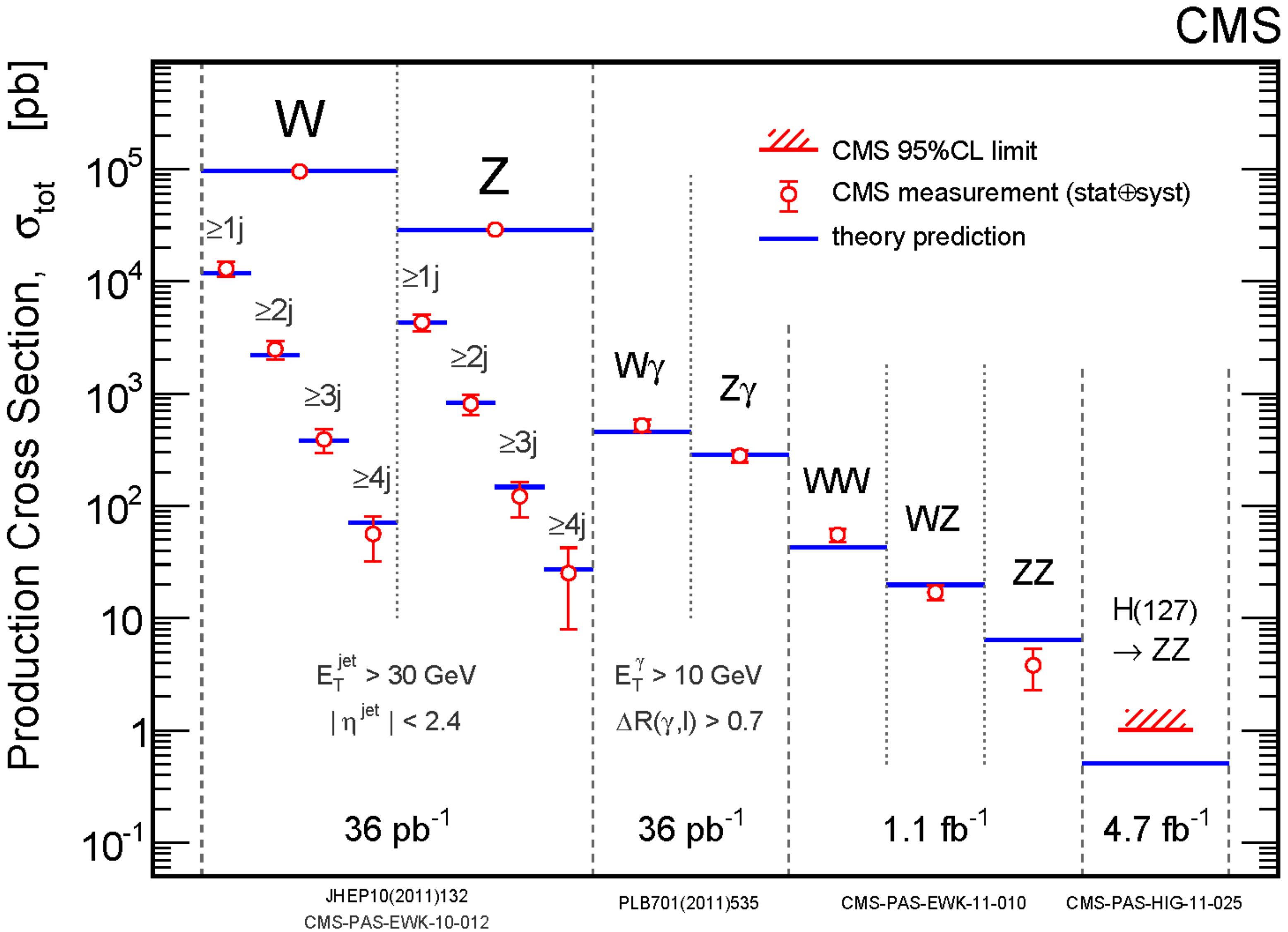
- **Introduction**
  - Experimental problems in high energy physics
  - The problem : how to distinguish signal from background ?

- **Multivariate analyses examples in HEP**
  - At the Tevatron
  - At the LHC

- **Presentation of commonly used multivariate methods**

# Searching for rare signals

**Higgs and new physics cross-sections are small...**



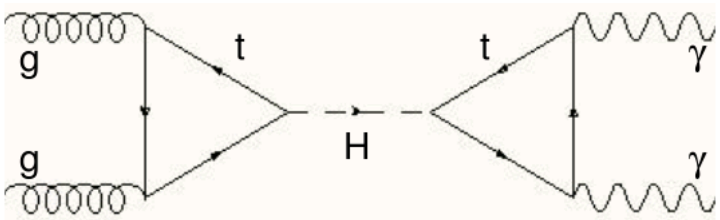

Examples of background to H→ZZ searches

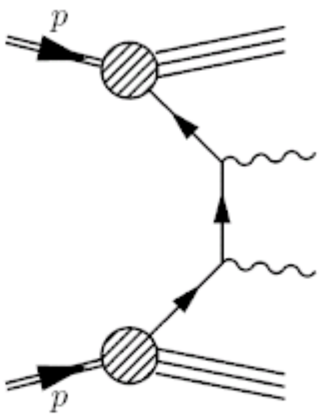
5 orders of magnitude

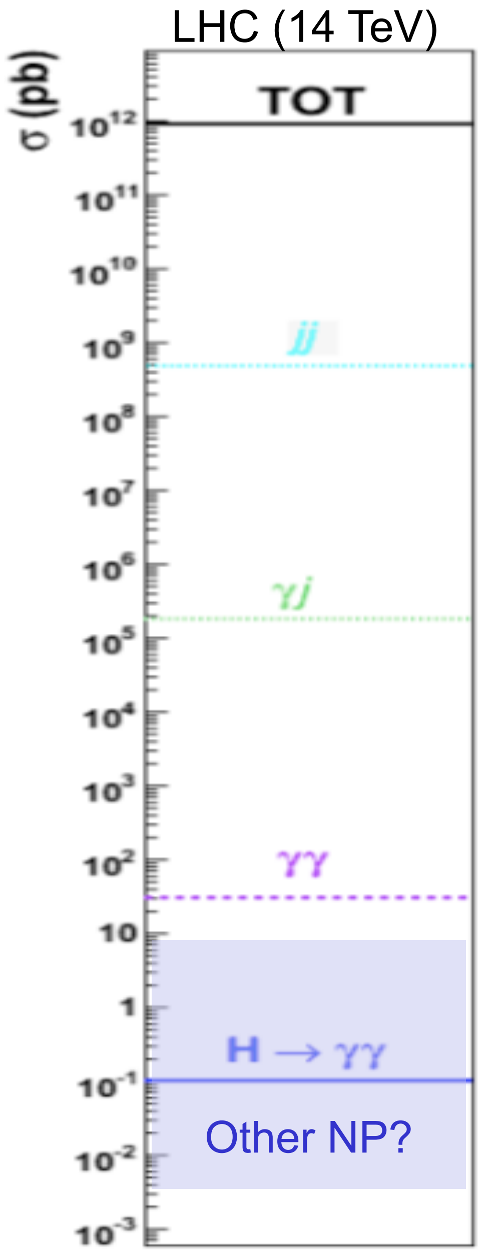
# Over huge backgrounds

**To achieve a discovery, huge background reduction rate needed**

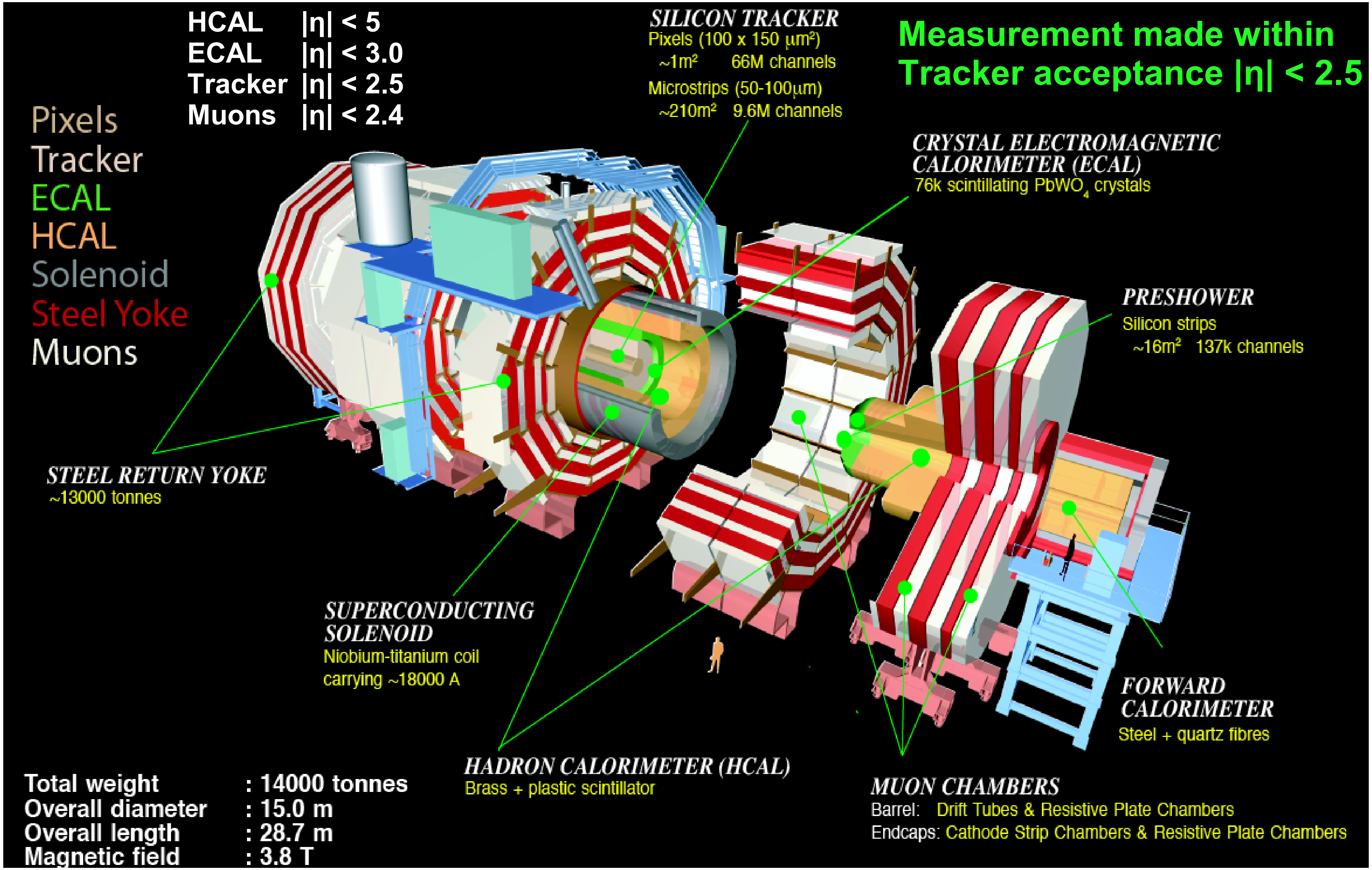
- Example of **H→γγ** : typically 9 orders of magnitude under the QCD jets background
- **Reducible background** : jet-jet, photon-jet
    - Jets can be mis-identified as photons
    => can be suppressed by tight photon identification criteria
- **Irreducible background :** photon-photon
    - Non-resonant diphoton continuum
    => Can be discriminated using kinematic properties



# With a given detector (here, CMS)

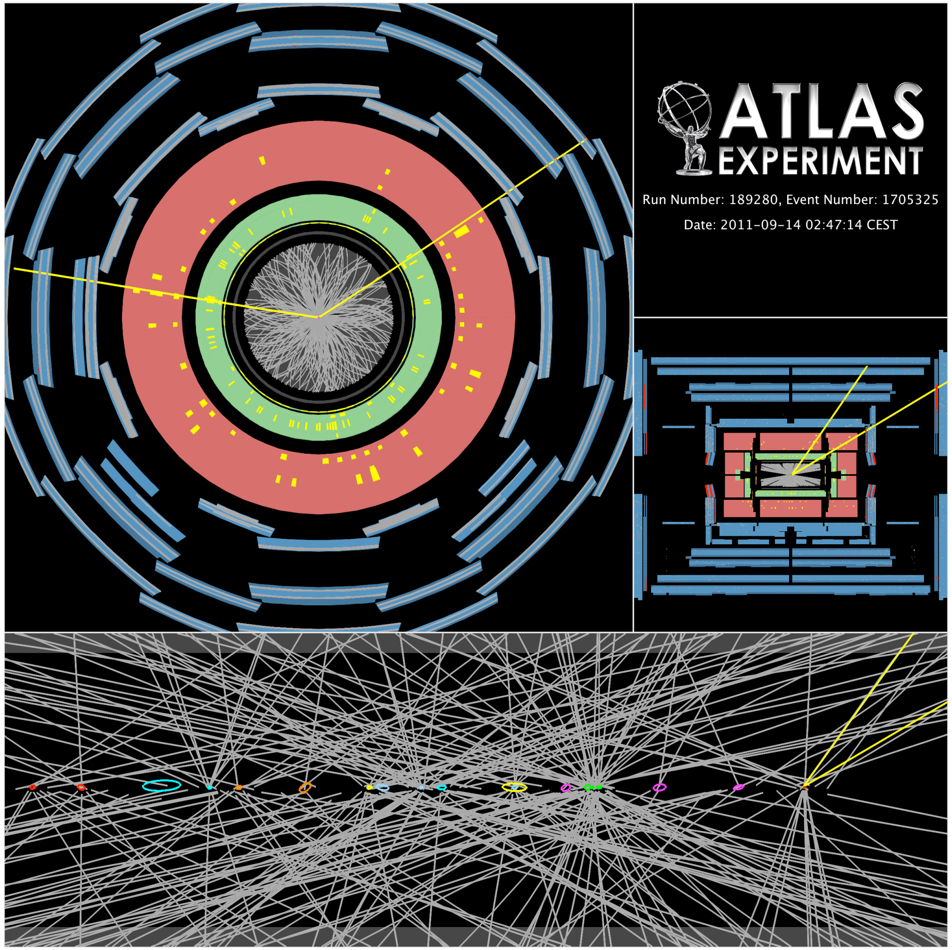


# Experimental issues

**Experimental challenges :**

- Detector calibration
- Identification of the tracks / energy deposits in the sub-detectors
- Particle reconstruction
- Particle identification
- Finding the vertex of hard interaction among all pile-up vertices
- Discriminate the signal process against all other background processes
- ...

- **Multivariate methods can help for that**




Collision with 20 pile-up events recorded with the ATLAS detector

# Multivariate analysis : Definitions

**MultiVariate Analysis :**

- Set of statistical analysis methods that simultaneously analyze multiple measurements (variables) on the object studied
- Variables can be dependent or correlated in various ways

**Classification / regression :**

- **Classification** : discriminant analysis to separate classes of events, given already known results on a training sample
- **Regression** : analysis which provides an output variable taken into account the correlations of the input variables

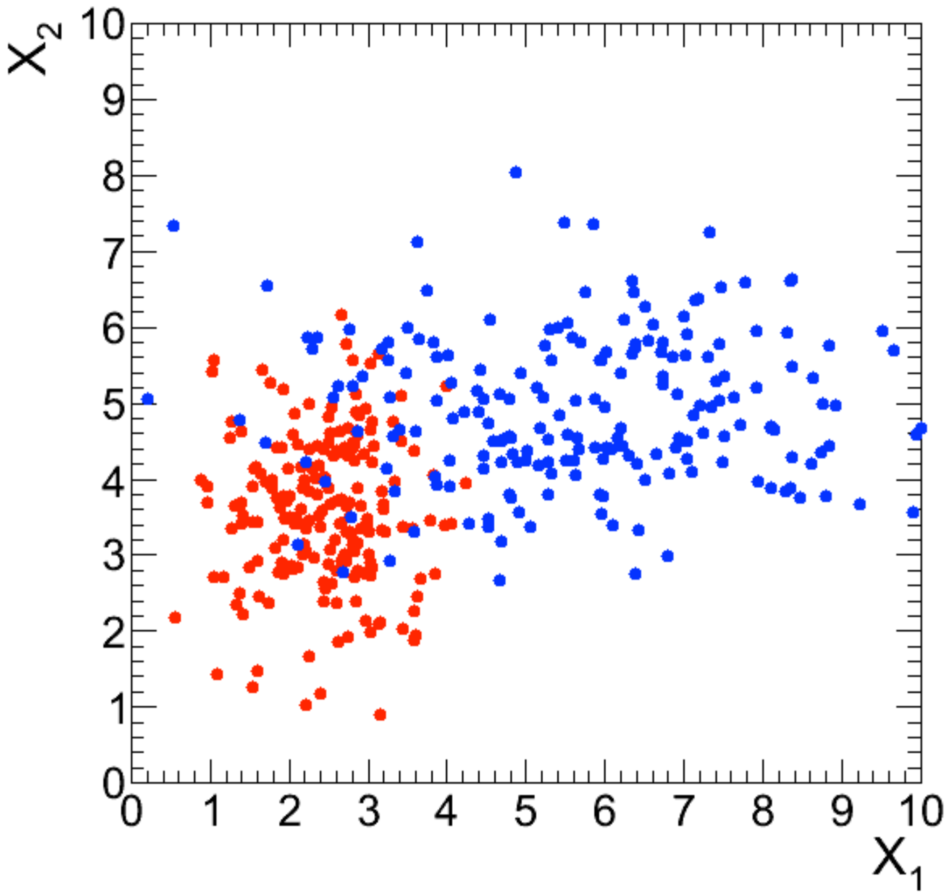
**Statistical learning :**

- **Supervised learning :** the multivariate method is trained over a sample were the result is known (e.g. Monte-Carlo simulation of signal and background)
- **Unsupervised learning :** no prior knowledge is required. The algorithm will cluster events in an optimal way

# Event classification

- Focus here on **supervised learning for classification**.
- Use case in particle physics : **signal/background discrimination**

- Assume we have two populations (signal and background) and two variables

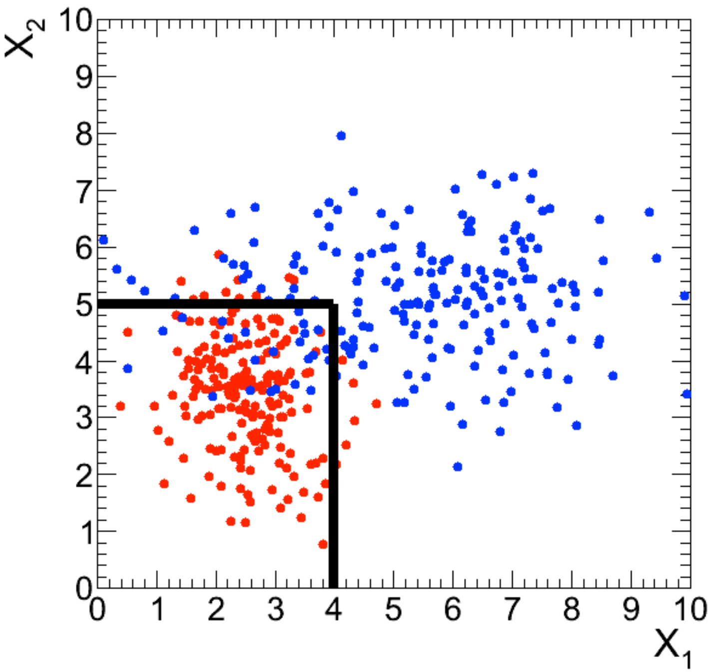



- How to decorrelate, what decision boundary (on X1 and X2) to choose, to decide if an event is signal or background ?

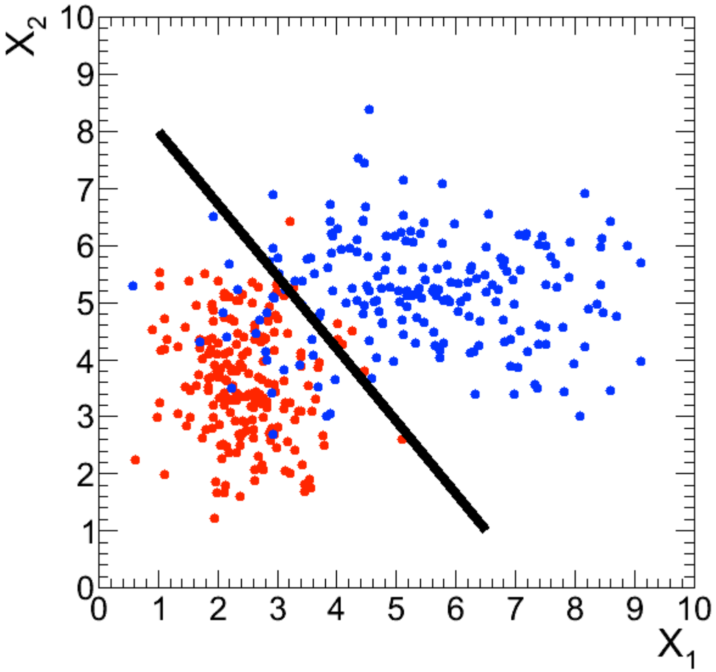
# Event classification

- **Possible solutions :** rectangular cuts, Fisher, non-linear contour

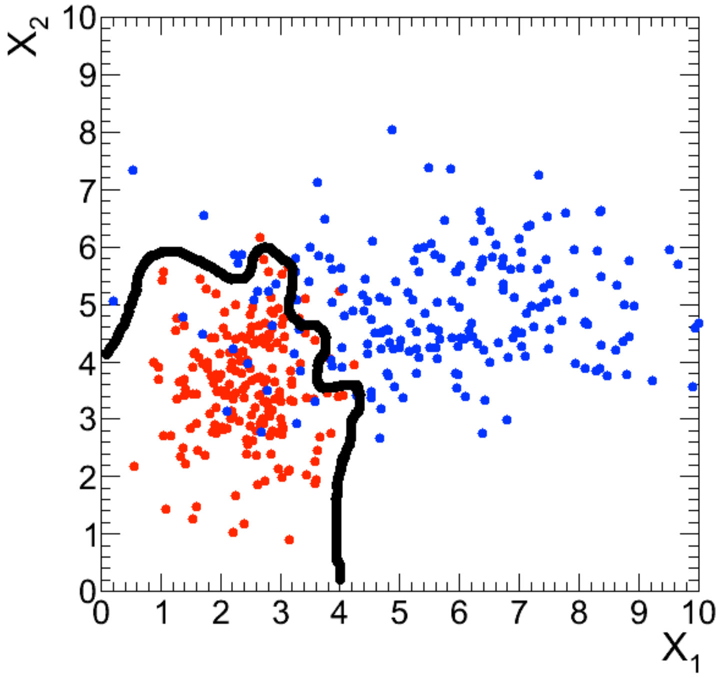
**Rectangular cuts**




**Linear (Fisher)**




**Non-linear**



# Multivariate analyses in HEP

- **Signal/background discrimination :**
  - **Object reconstruction :** discriminate against instrumental background (electronic noise...)
  - **Object identification :** e.g. electron, bottom quark identification, to improve the rejection other objects resembling (e.g. jets)
  - **Discriminating physics process against physics backgrounds**. Many examples, e.g. single top against W+jets, H->WW against WW background...

- **Improving the energy measurement**, via regression. Allows to narrow the reconstructed mass peak, improve the resolution.

- **Estimate the sensitivity of the analysis :**
  - **Sensitivity to signal exclusion or discoveries :** Likelihood of the data to be consistent with background only or signal+background hypothesis
  - **Combination** of many channels

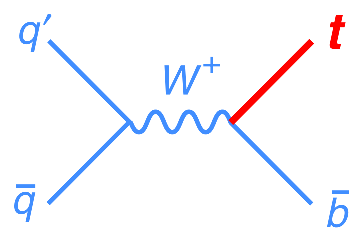
  => exclusion limits or discoveries

# MVA examples in HEP : Tevatron

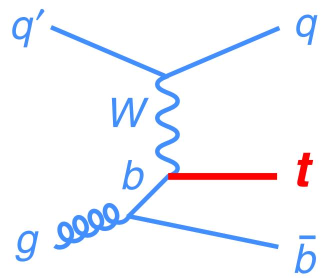
## Single top discovery

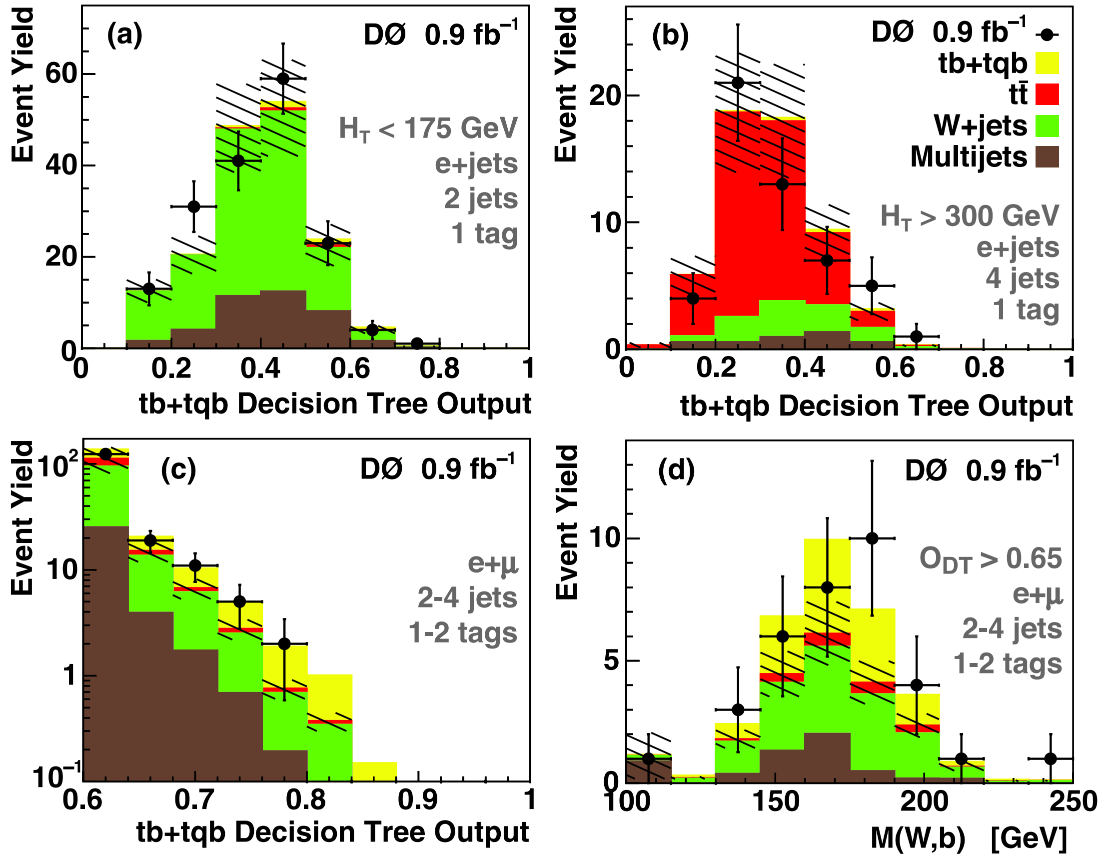
PhysRevLett.98.181802

(a)



(b)






- When published, very controversial

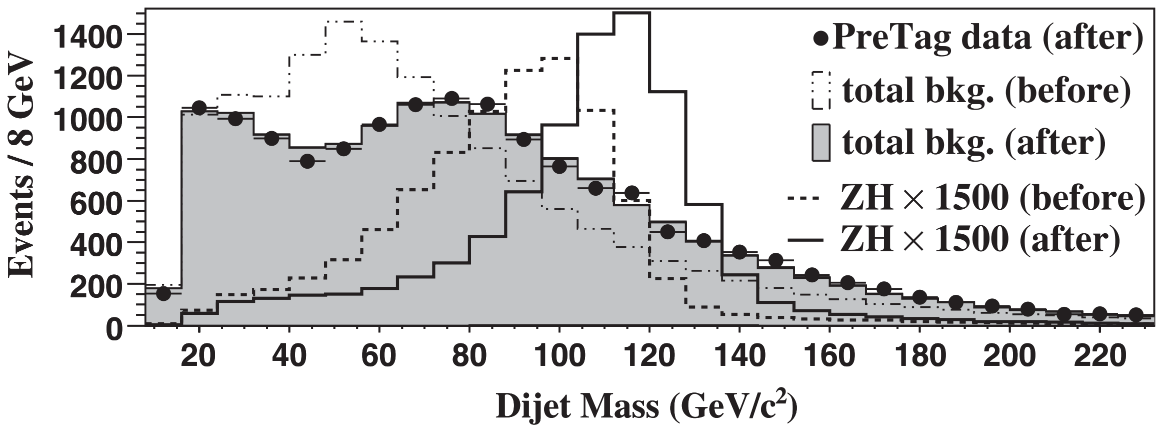
- 36 boosted decision trees used to discriminate signal from background

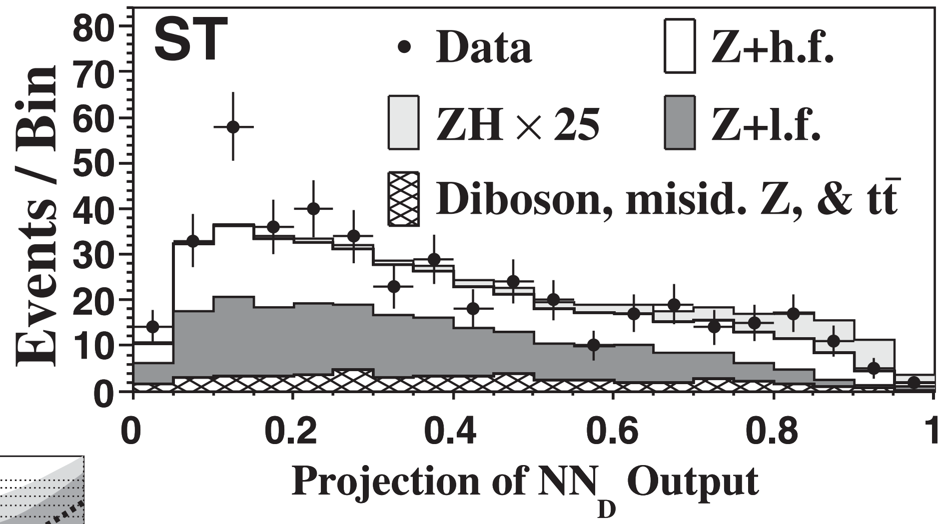
- First measurement of the single top cross-section, today well established

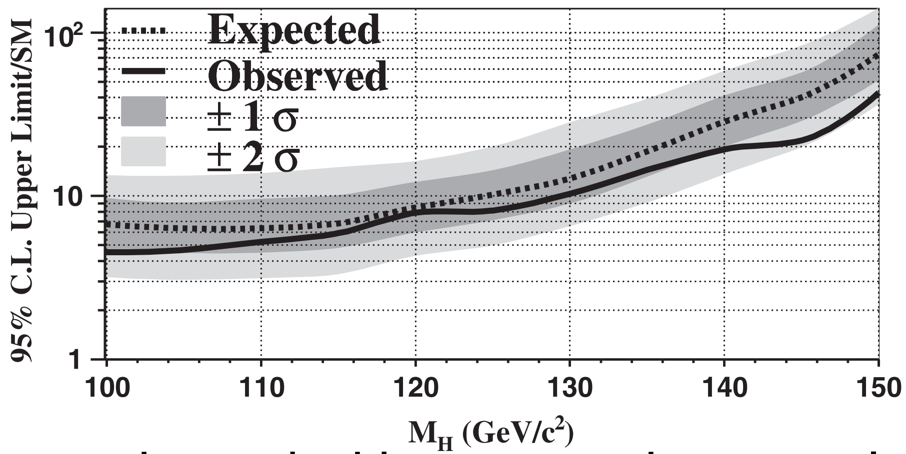
# MVA examples in HEP : Tevatron

## ZH→llbb searches at CDF

PRL 105, 251802 (2010)



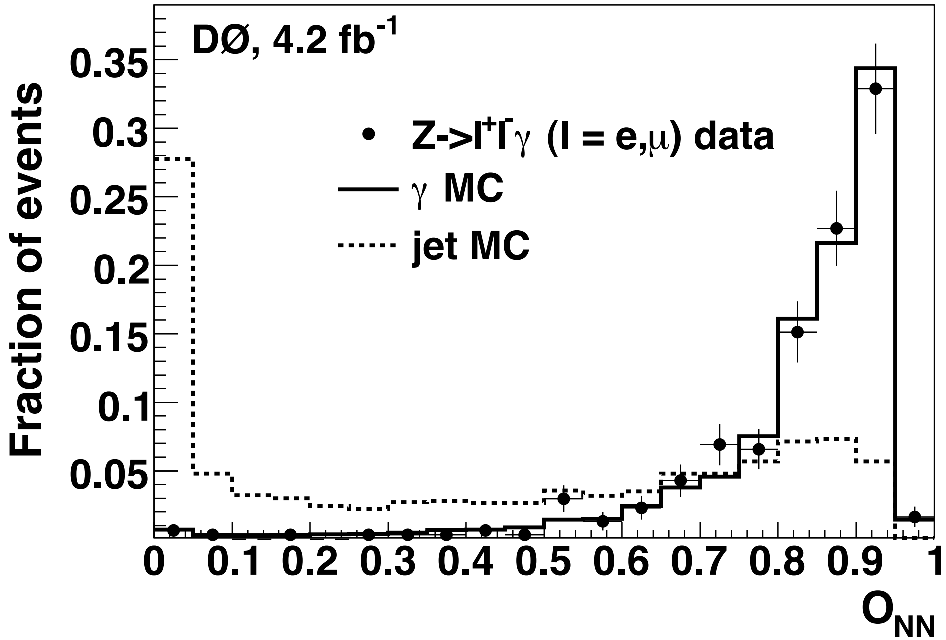



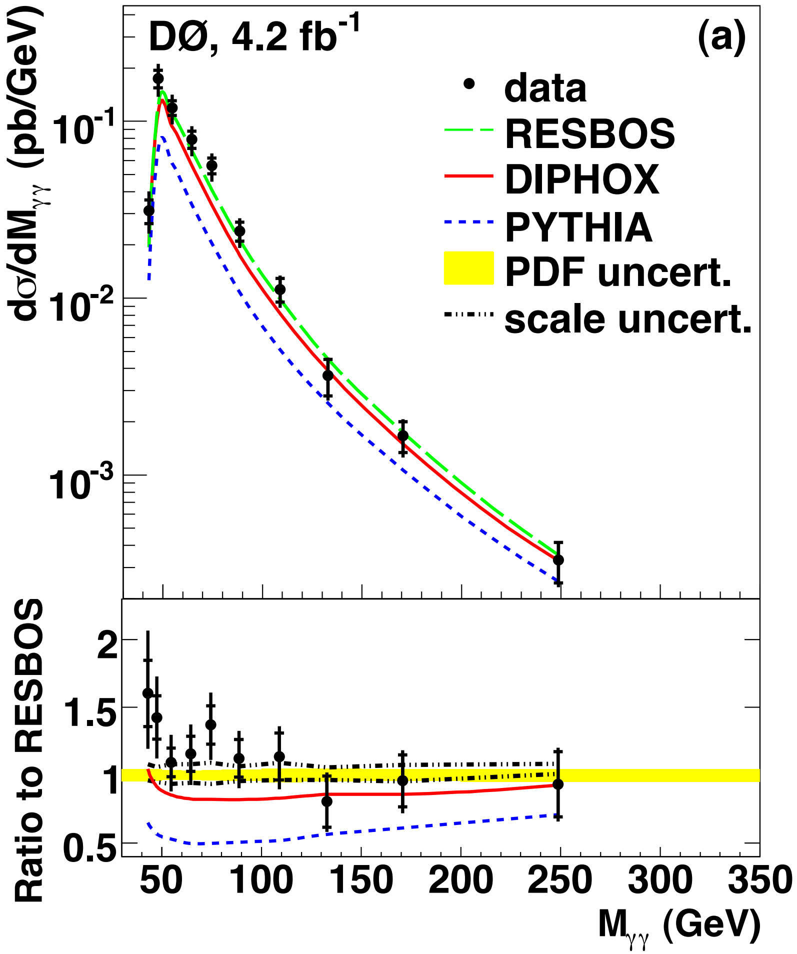

- b-jet energy estimated with a regression neural network, to improve dijet mass resolution
- b-tagging with neural networks, used to compute the final limits

# MVA examples in HEP : Tevatron

## Photon identification at D0 and applications

arxiv:1002.4917v3



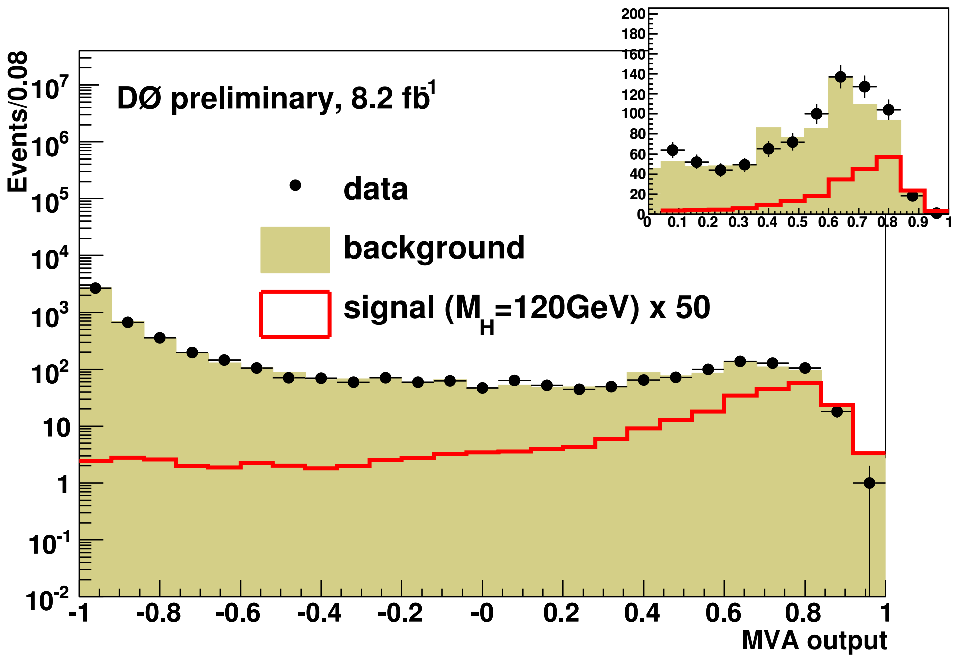




- Neural network for Photon Id based on calorimeter energy deposit and track variables in an isolation cone around the photon
- Used to identify and measure the diphoton+X cross-section

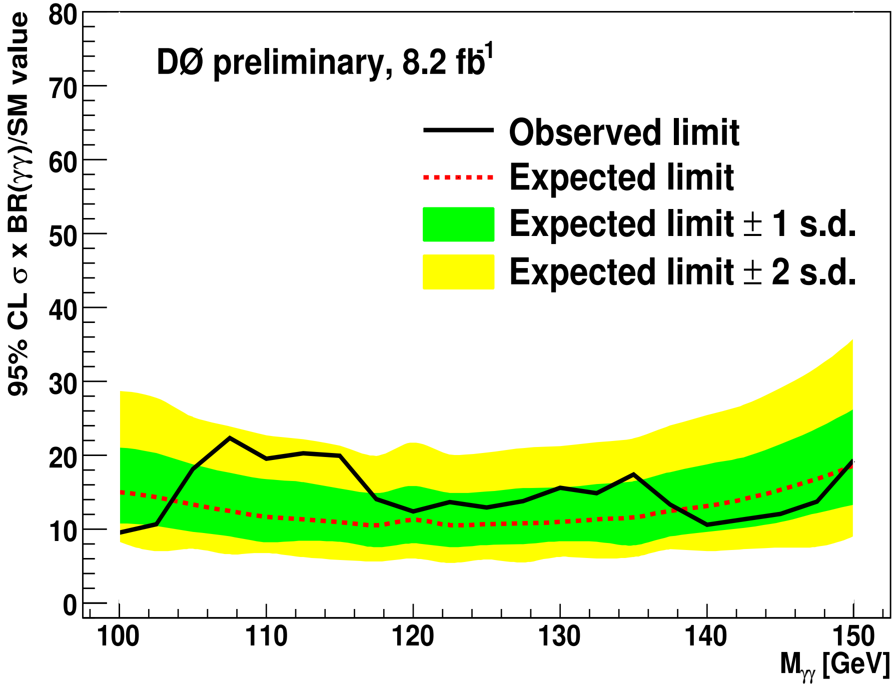
# MVA examples in HEP : Tevatron

## H→γγ searches at D0

DØ Note 6177-CONF




(c) $M_H = 120$ GeV




- Identify photons with the neural network (reduces fake photons processes)
- Boosted decision tree with kinematic variables to improve the sensitivity against the diphoton continuum (+30%)
- The BDT includes the invariant mass of the diphoton system as input

# MVA examples in HEP : LHC

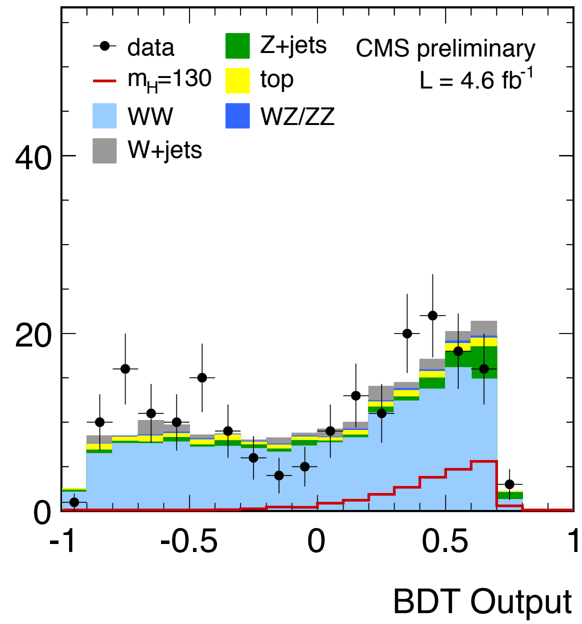
## H→WW→llvv searches in CMS

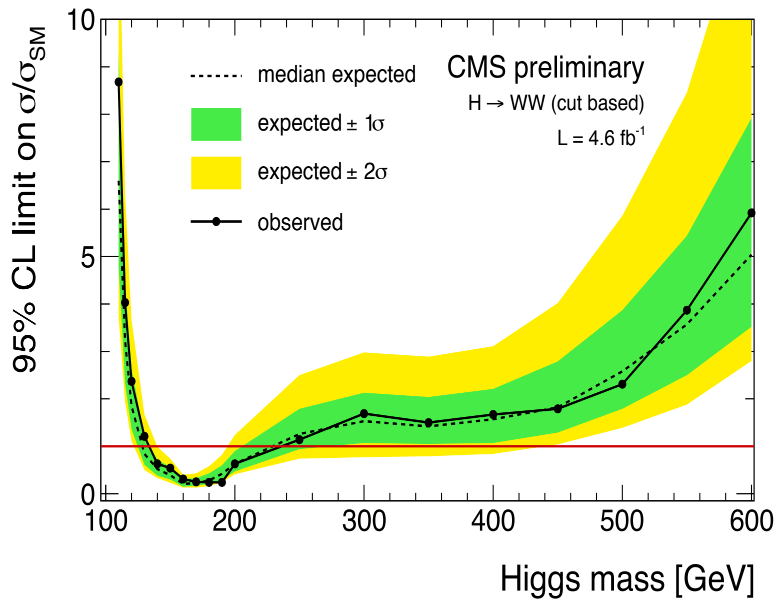
CMS-PAS-HIG-11-024

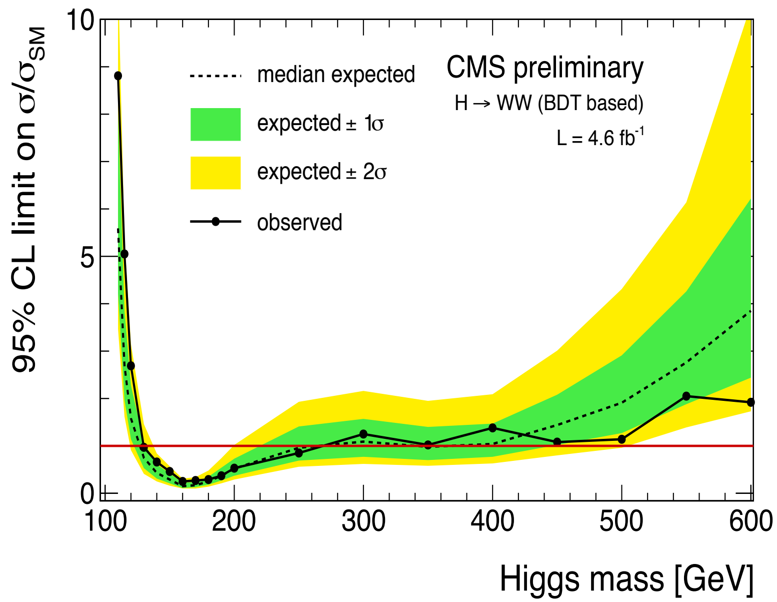
- 3 channels : 0-jet, 1-jet, 2-jet

- Electron identification with a multivariate technique : 50% more background rejection for the same signal efficiency
- Boosted decision tree in 0-jet and 1-jet channels : kinematic variables

- Limits improved by using BDT



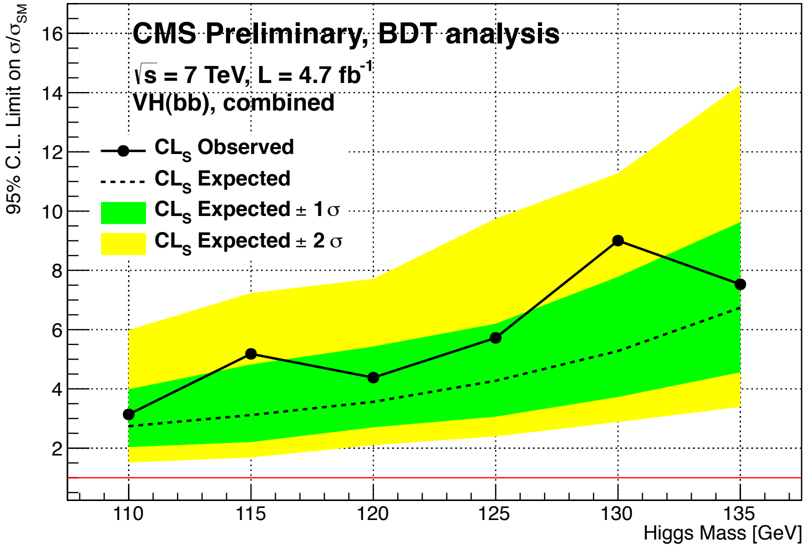



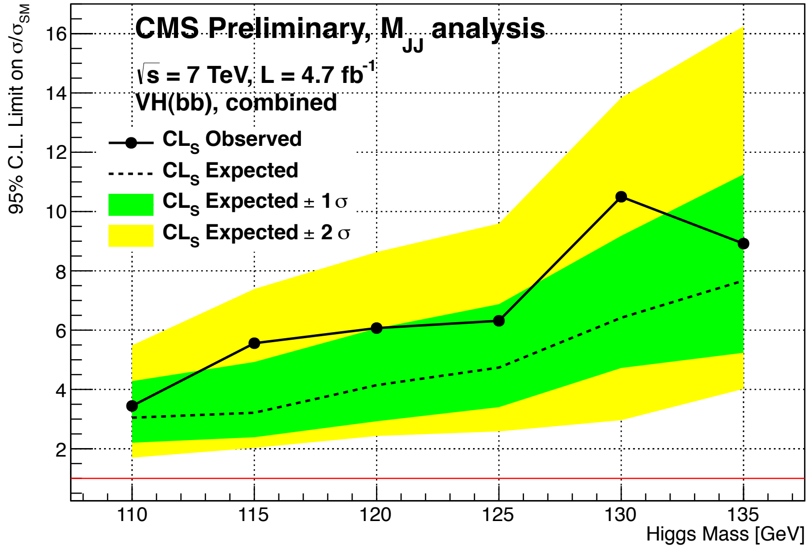



# MVA examples in HEP : LHC

## H→bb searches in CMS

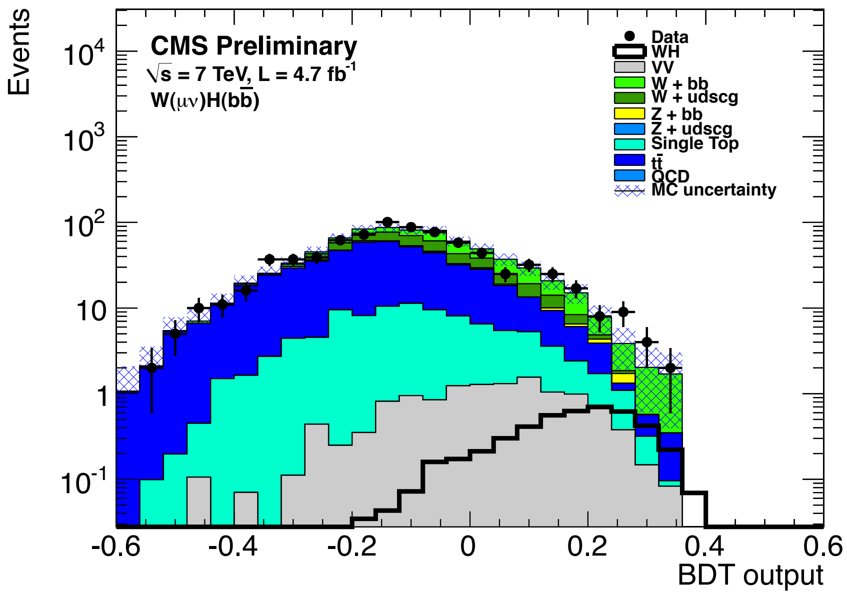
CMS-PAS-HIG-11-031




- Searches for VH, H→bb
- 5 channels : W→eν,μν, Z→ee,μμ, Z→νν

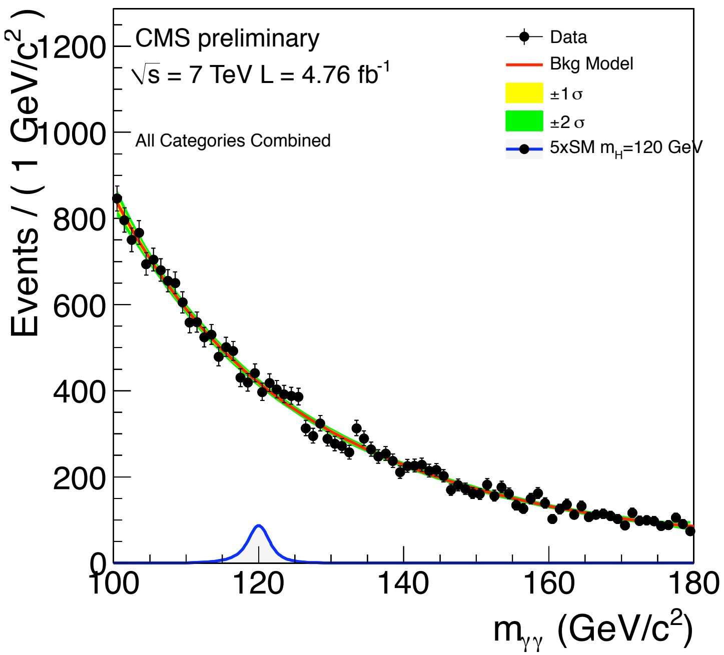
- B-tagging selection on a likelihood discriminant (track impact parameter + secondary vertices information)
- Boosted decision trees for the kinematics

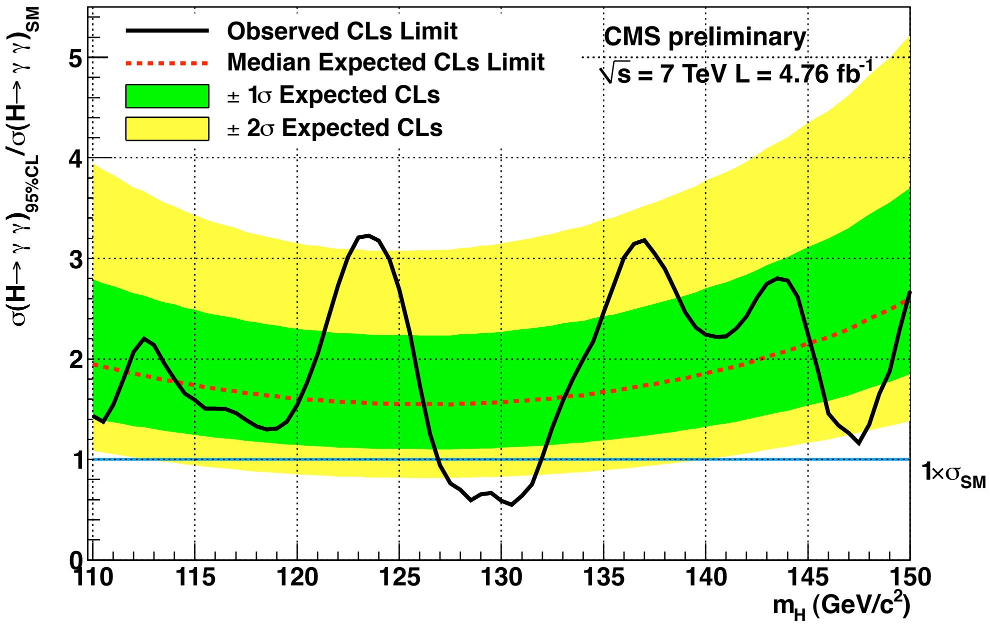


# MVA examples in HEP : LHC

## H→γγ searches in CMS

CMS-PAS-HIG-11-030



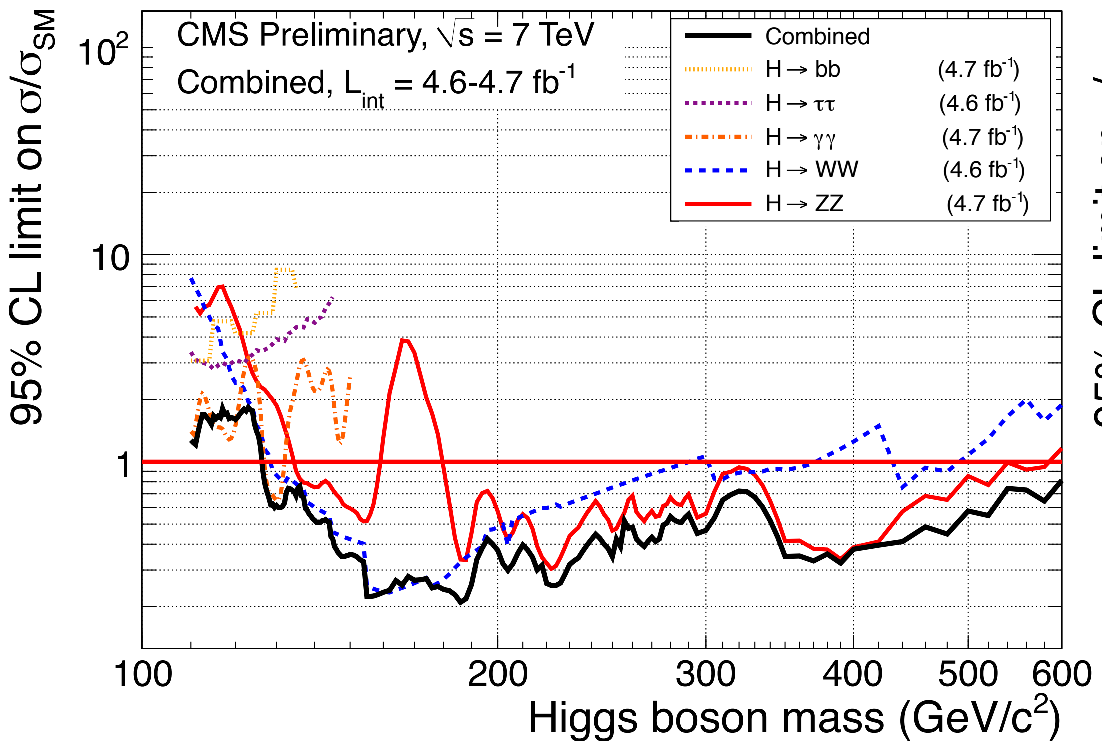



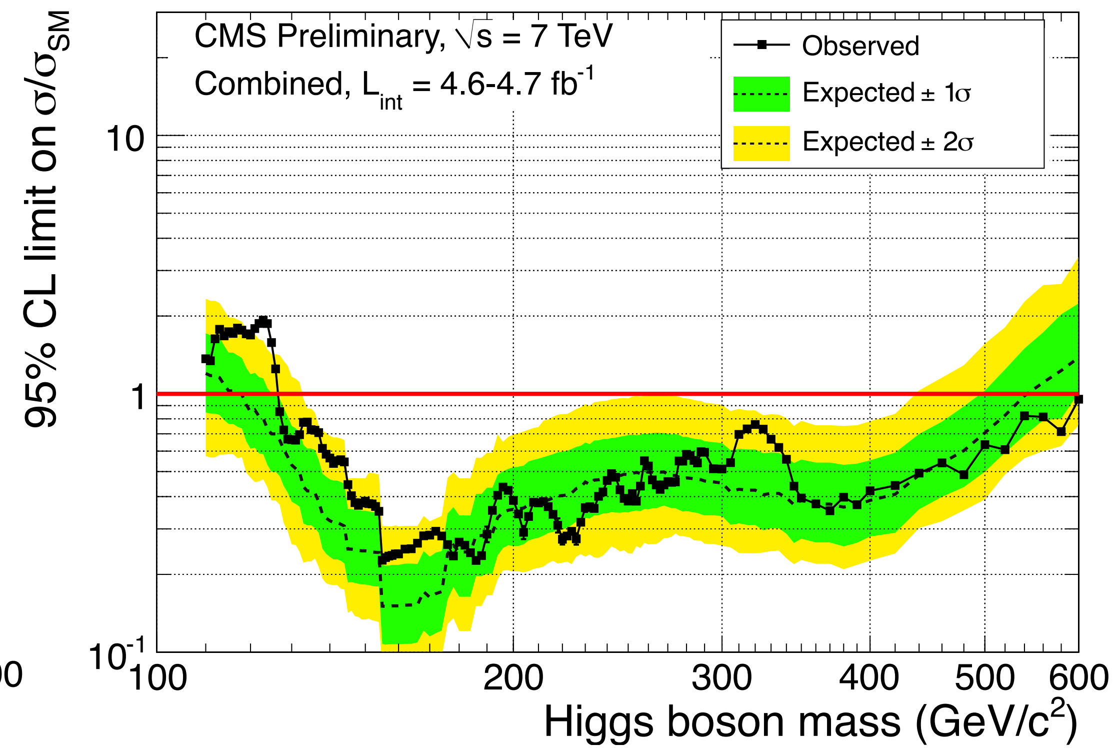

- Hard interaction vertex identified with a BDT using diphoton kinematics and track variables
- Photon energy estimated with a BDT regression from geometry and energy deposit variables (10% improvement on the limit)

# MVA examples in HEP : LHC

## Combination of all channels in CMS

CMS-PAS-HIG-11-032




- Combination can be seen as a grand multivariate analysis
- Limits are set with CLs method
- Exclusion at 95% confidence level : 127-600 GeV

# Plenty of multivariate methods...

**Example of MVA methods :**
- Rectangular cut optimization
- Fisher
- Likelihood
- Neural network
- Decision tree
- Support Vector Machine
- ...

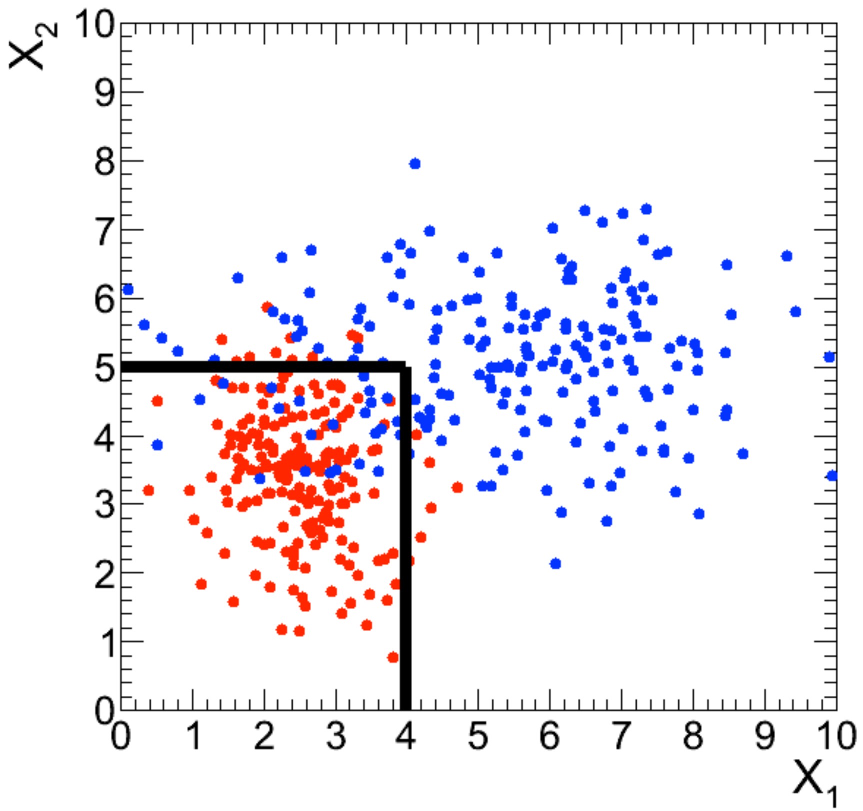
**Characteristics :**
- Level of complexity and transparency
- Performance in term of background rejection
- Way of dealing with non-linear correlations
- Speed of training
- Robustness while increasing the number of input variables
- Robustness against overtraining

# Rectangular cuts

- Simplest multivariate method, very intuitive
- All HEP analyses are using rectangular cuts, not always completely optimized

**Rectangular cuts optimization :**
- Grid search, Monte-Carlo sampling
- Genetic algorithm
- Simulated annealing

**Characteristics :**
- Difficult to discriminate signal from background if non-linear correlations
- Optimization difficult to handle with high number of variables




Define the signal region :

a1 < x1 < a2,

b1 < x2 < b2

...

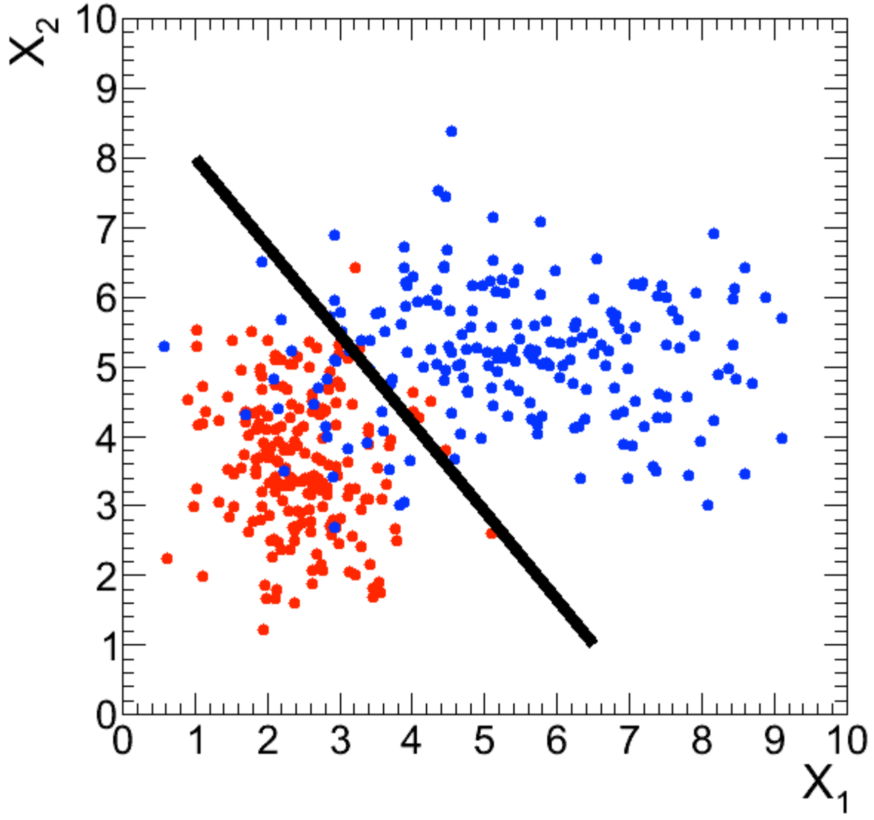
# Fisher discriminant

**Fisher method :**

- Cut on a linear combination of the input variables

  $y < a.x1 + b.x2$
- This corresponds to an hyper-plan in the variable phase-space
- Very efficient if linear correlations

- Again, difficult to handle non-linear correlations
- More easily trained than rectangular cuts



# Likelihood estimator

- The likelihood ratio is defined by : $$y_{\mathcal{L}}(i) = \frac{\mathcal{L}_S(i)}{\mathcal{L}_S(i) + \mathcal{L}_B(i)}$$

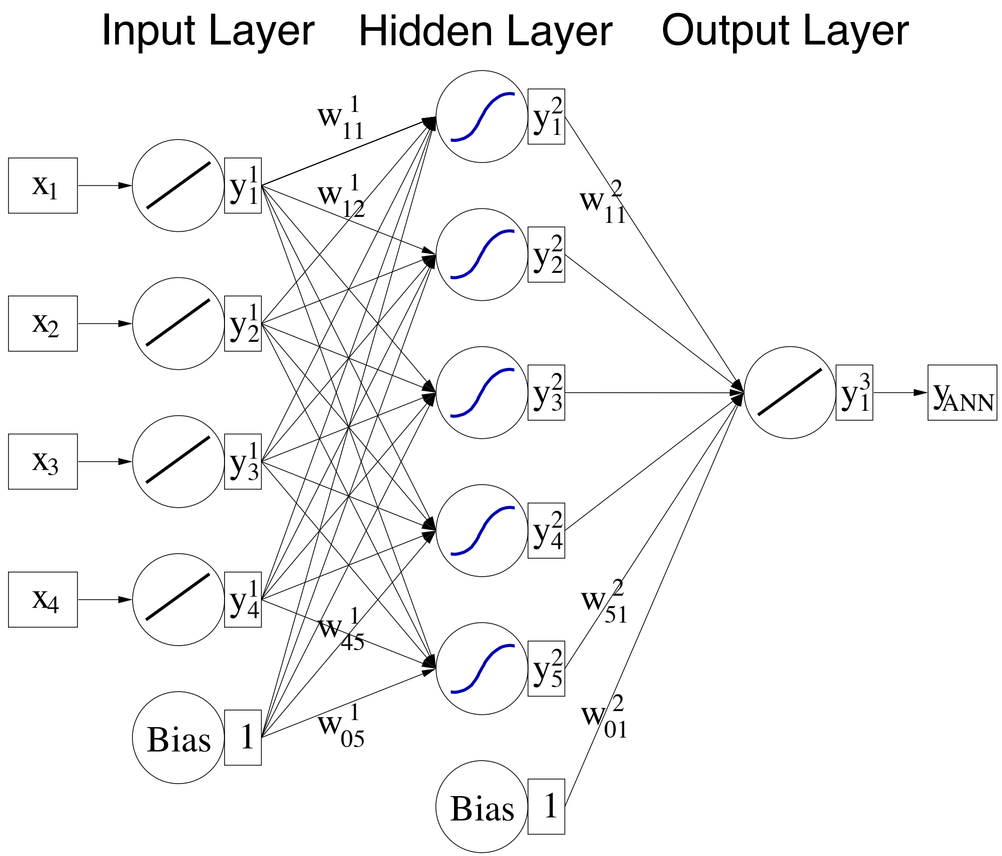
$$\mathcal{L}_{S(B)}(i) = \prod_{k=1}^{n_{\mathrm{var}}} p_{S(B),k}(x_k(i))$$

is the product of the probability function for each variables.

- Optimal when no correlation between the variables
- This likelihood method does not take into account the correlations and is therefore sub-optimal in presence of correlations
- Refinements exist to take into account the correlations

# Neural network

- Most commonly used : the **multi-layer perceptron**
- Composed of neurons taking as input a linear combination of the previous neuron outputs
- Activation function (usually tanh) transforms the linear combination
- Weights for each neurons are found during the training phase by minimizing the error on the neural network output

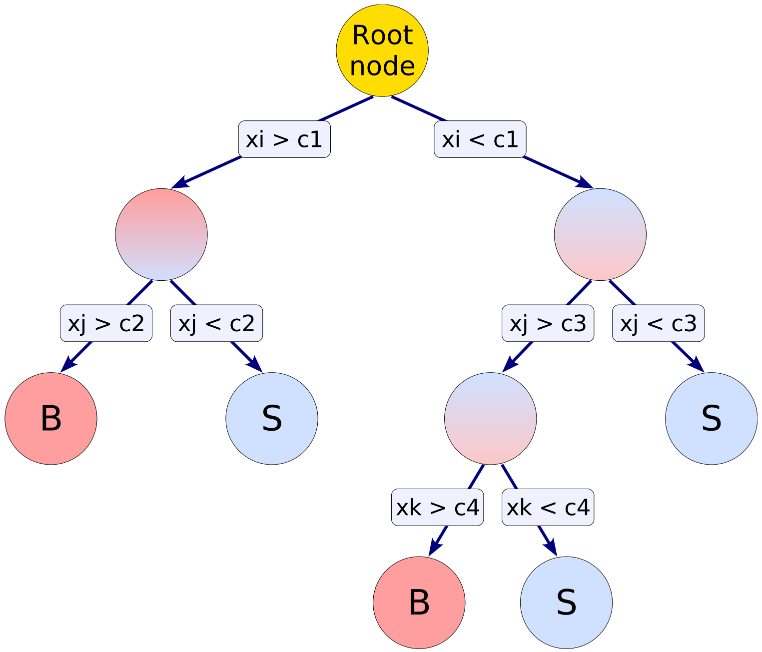



- Neural networks are universal approximators : takes advantage of correlations

- Quite stable against overtraining and against increasing number of variables

# Decision tree

- A **decision tree** is a binary tree : a sequence of cuts paving the phase-space of the input variables
- Repeated yes/no decisions on each variables are taken for an event until a stop criterion is fulfilled
- Trained to maximize the purity of signal nodes (or the impurity of background nodes)

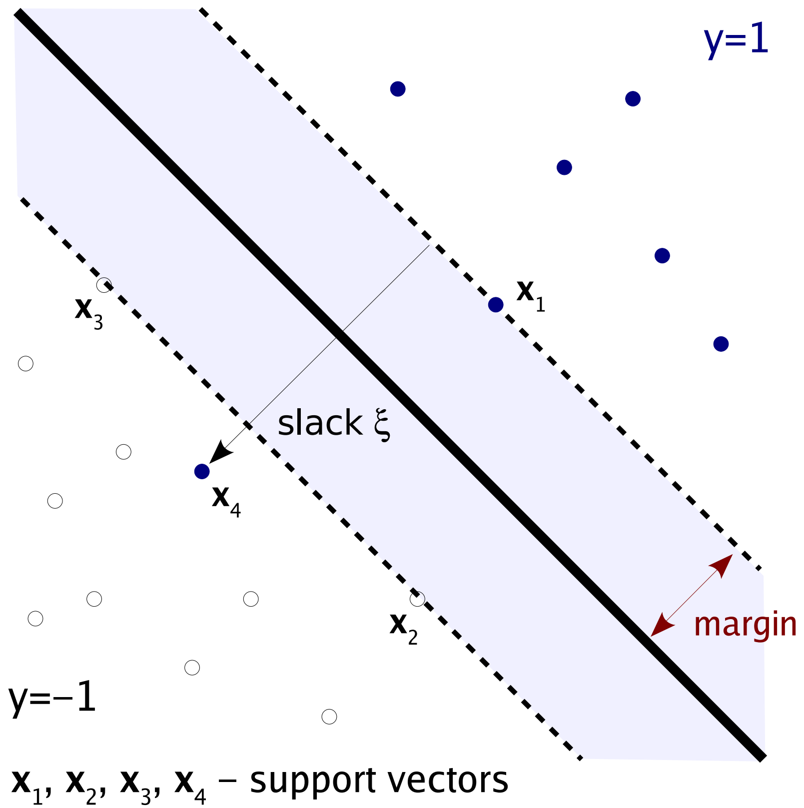



- Decision trees are **extremely sensitive to the training samples**, therefore to overtraining

- To stabilize their performance, one uses different techniques :
    - **- Boosting**
    - - Bagging
    - - Random forests

# Support Vector Machine

- **Idea** : build a hyperplane that separate signal and background vectors (events) using only a subset of all training vectors (support vectors)
- Position of the hyperplane found by maximizing the margin between it and the support vectors
- Higher dimensions spaces are used by non-linear transformation, using kernel functions such as the gaussian basis




- SVM can be competitive with NN and BDT but is often less discriminant : often data are non-separable, therefore sensitive to all the SVM parameters

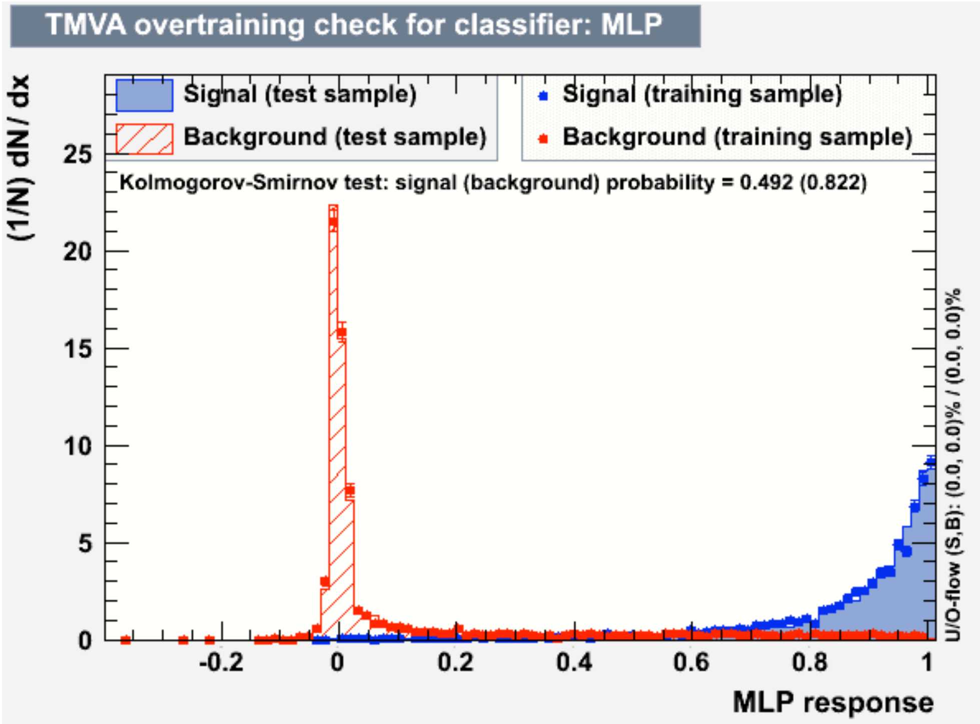
- In some cases this method performs very well

# Training and application

**Training / test samples**

- For all multivariate methods, two samples are used :
    - Training sample
    - Test sample
- This is mandatory to check that the training has converged to a solution which does not depend on the statistical fluctuations of the training sample
- Generally speaking, MVA should be applied (or tested) in events where the response is not known

- Training is time-consuming, especially while increasing the number of variables (and depending on the method)
- Application is usually faster : it uses a set of weights used in the MVA output computation



# Which method to choose ?

From TMVA manual

| Criteria | | MVA Method: Cuts | Likelihood | PDE-RS / k-NN | PDE-Foam | H-Matrix | Fisher / LD | MLP | BDT | RuleFit | SVM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Performance | No or linear correlations | ⋆ | ⋆⋆ | ⋆ | ⋆ | ⋆ | ⋆⋆ | ⋆⋆ | ⋆ | ⋆⋆ | ⋆ |
| | Nonlinear correlations | ○ | ○ | ⋆⋆ | ⋆⋆ | ○ | ○ | ⋆⋆ | ⋆⋆ | ⋆⋆ | ⋆⋆ |
| Speed | Training | ○ | ⋆⋆ | ⋆⋆ | ⋆⋆ | ⋆⋆ | ⋆⋆ | ⋆ | ○ | ⋆ | ○ |
| | Response | ⋆⋆ | ⋆⋆ | ○ | ⋆ | ⋆⋆ | ⋆⋆ | ⋆⋆ | ⋆ | ⋆⋆ | ⋆ |
| Robustness | Overtraining | ⋆⋆ | ⋆ | ⋆ | ⋆ | ⋆⋆ | ⋆⋆ | ⋆ | ○ | ⋆ | ⋆⋆ |
| | Weak variables | ⋆⋆ | ⋆ | ○ | ○ | ⋆⋆ | ⋆⋆ | ⋆ | ⋆⋆ | ⋆ | ⋆ |
| Curse of dimensionality | | ○ | ⋆⋆ | ○ | ○ | ⋆⋆ | ⋆⋆ | ⋆ | ⋆ | ⋆ | |
| Transparency | | ⋆⋆ | ⋆⋆ | ⋆ | ⋆ | ⋆⋆ | ⋆⋆ | ○ | ○ | ○ | ○ |