introduction

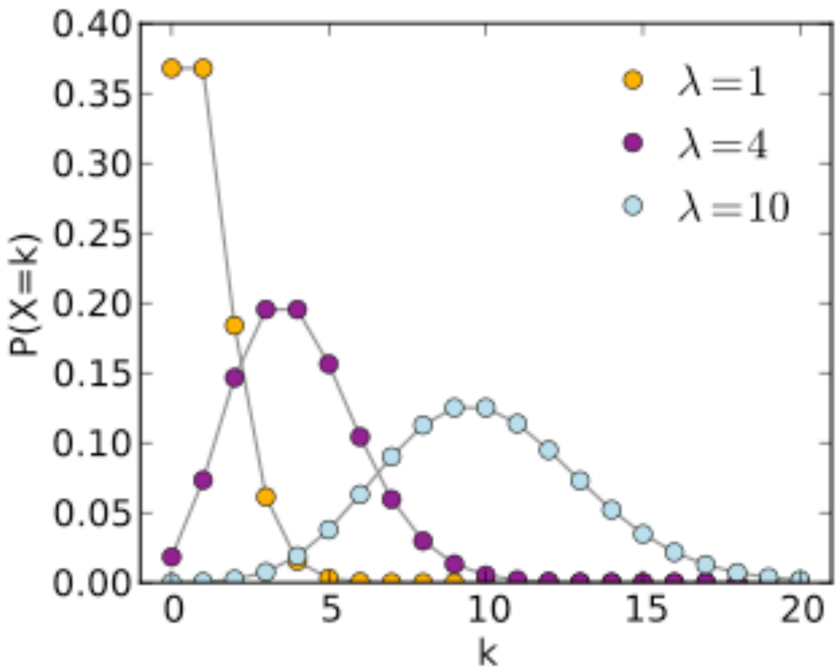
- my interpretation of “data analysis techniques” is here “doing a data analysis”
- follow the steps from the beginning (data taking) to the end (the result)
  - the luminosity
  - the trigger, from the point of view of the analysis
  - the reconstruction and detector response
  - the simulation
  - differential cross-section measurement: a di-jet correction
  - searches: the H > WW > lvlv
  - multivariate techniques

thanks to the following people, for interesting discussions, for liberally “borrowing” slides, or both: D. Benedetti, C. Bernet, T. Camporesi, G. Cowan, K. Cranmer, K. Ellis, S. Gennai , A. Ghezzi, A. Hoecker, R. Van Kooten, M. Nguyen, M. Paganoni, M. Pelliccioni, E. Rizvi, R. Rossin ...

the pile-up

the pile-up

- At LHC, the interaction rate is higher than the bunch crossing rate
- Within a bunch crossing in LHC, more interactions happen
- An event of interesting physics will be **recorded together with other events overlapped**, that are proton-proton interactions with low physics interest
- they are equivalent to a non-interesting event (**minimum bias**)



- given an average number of interactions, the number of PU events per bunch-crossing is expected to have roughly a poissonian distribution

measure the pile-up

- multiply the luminosity (per bunch) by the minimum bias cross-section (71.3 mb) gets the expected rate per bunch:

$$\mathrm{Rate}_{\mathrm{pileup}_{\mathrm{xing,ls}}} = \mathscr{L}_{\mathrm{xing,ls}} \cdot \sigma_{\mathrm{minimum\ bias}}$$

- divide by the revolution frequency of a bunch to get the number of PU events:

$$\mathscr{N}_{\mathrm{pileup}_{\mathrm{xing,ls}}} = \frac{\mathscr{L}_{\mathrm{xing,ls}} \cdot \sigma_{\mathrm{minimum\ bias}}}{\mathrm{cirulation\ rate}}$$

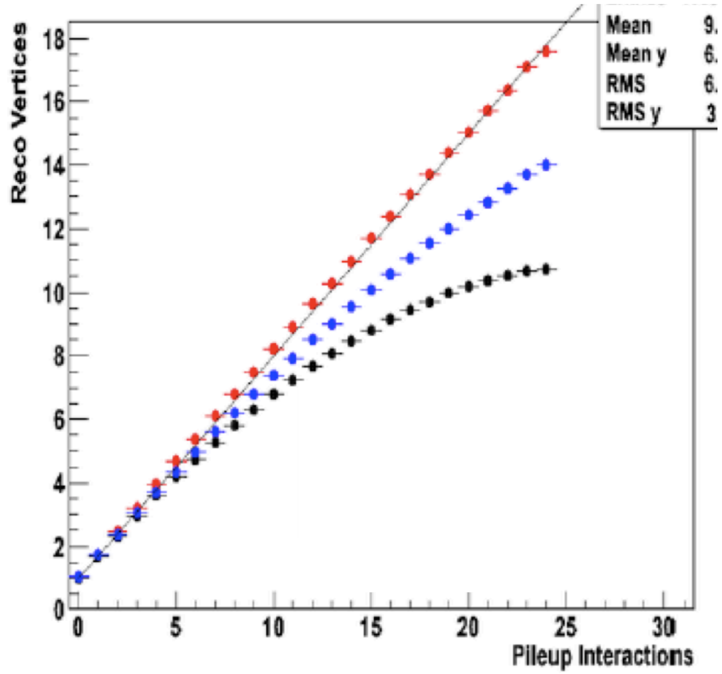
- calculate average distributions over longer periods, weighting by the luminosities

effects of pile-up

- fill in the detector with deposits:
  - **jet reconstruction** algorithms incorporate pile-up deposits
  - **lepton isolation** cones are filled in with pile-up deposits
  - **new jets** might appear in the event
  - more hits in the **tracker** appear
  - the **trigger** is affected
  - **MET** resolution worsens
  - ....

how to deal with it

- apply strict r**equirements on the vertexing of tracks** - need a precise vertex reconstruction algorithm
- measure the **pile-up density** event by event, and use it to subtract from the jets energy a pile-up term (FastJet)
- do the same with isolation cones

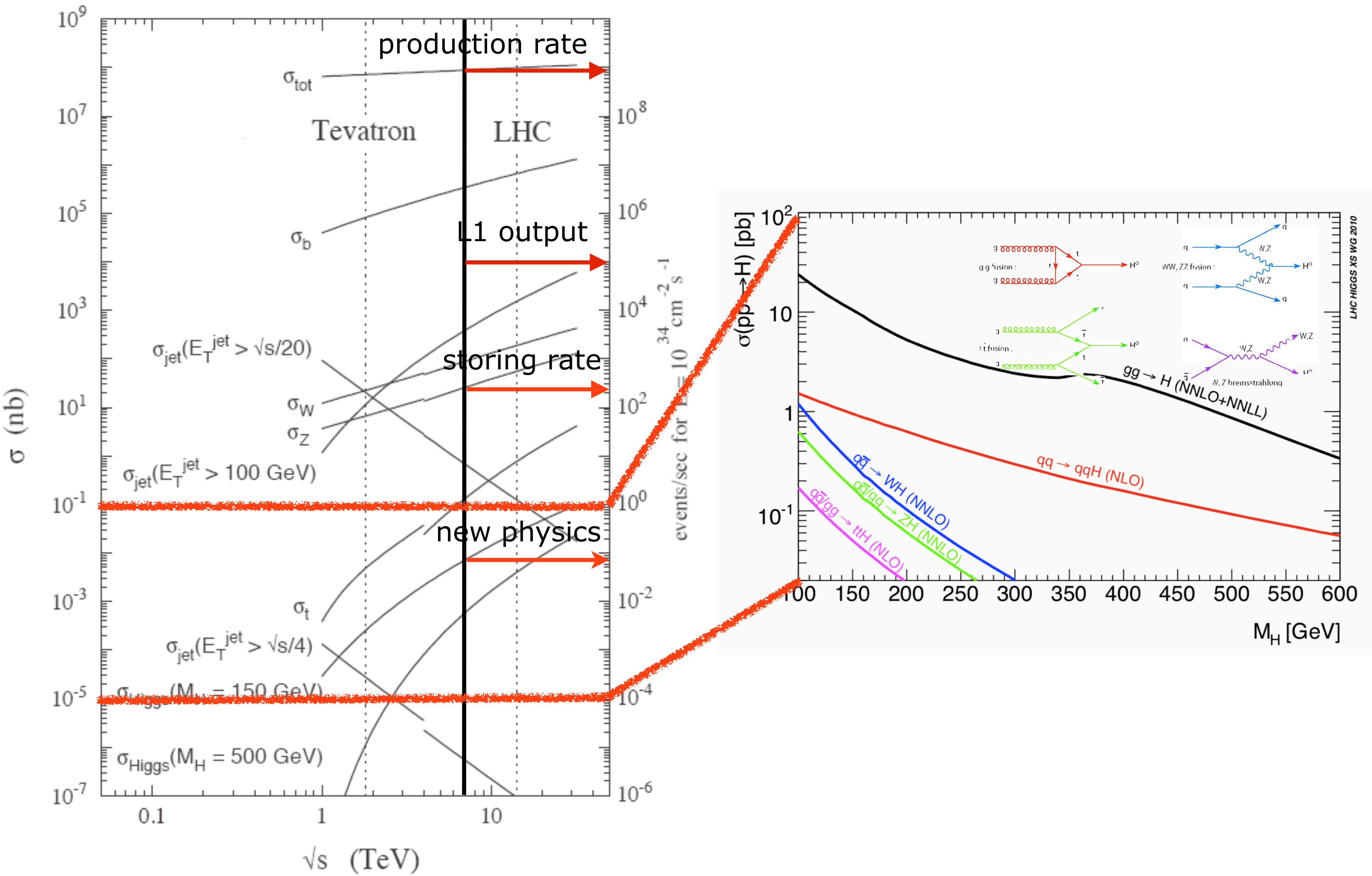


- subtract in the isolation cone the contribution of tracks that do not aim at the same vertex of the lepton
- reconstruct the MET only with particles that aim at a given vertex

M. Cacciari, G. Salam and G. Soyez, **FastJet** http://www.lpthe.jussieu.fr/~salam/fastjet/

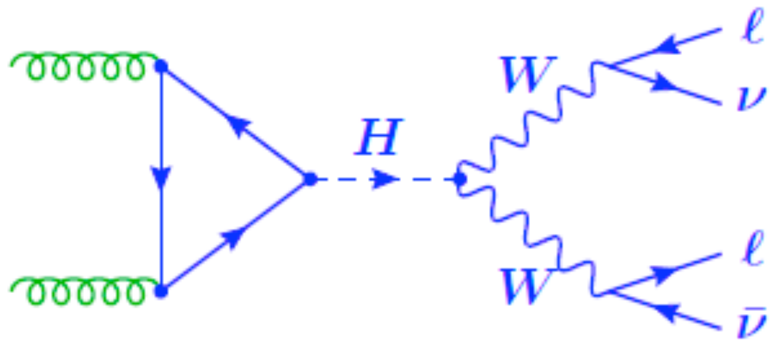
H > WW > lvlv

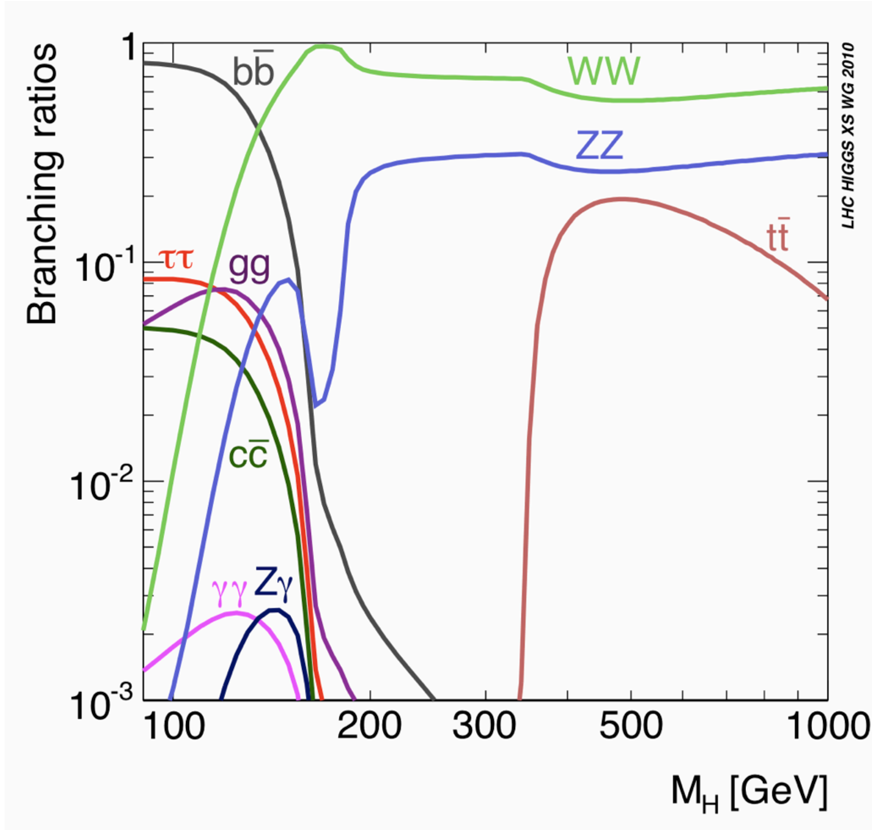
one plot for the Higgs boson



H > WW > lνlν

$$gg \to H \to WW \to \ell\bar{\nu}\bar{\ell}\nu$$

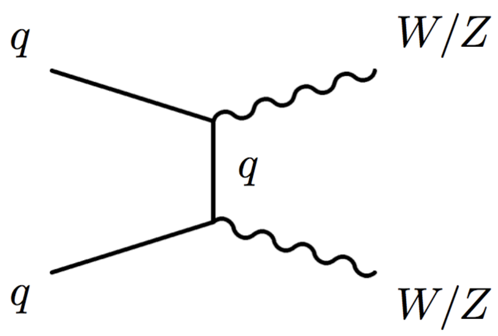


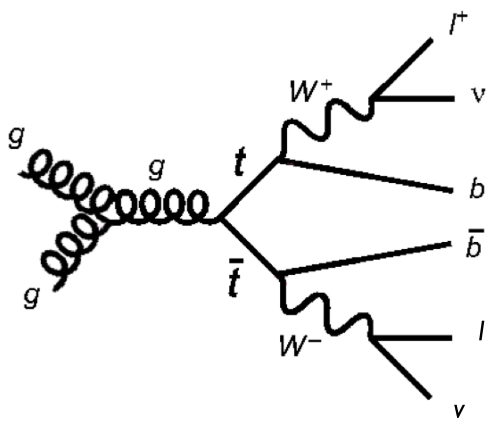


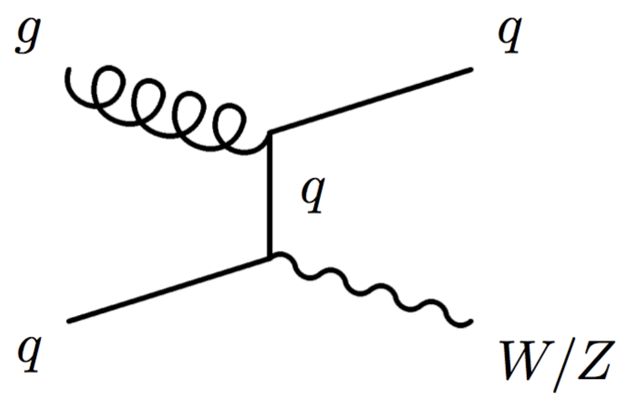
- **pros**
  - main production channel over a large mass range
  - main decay channel for intermediate and high masses
- **cons**
  - no invariant mass reconstruction is possible, since two neutrinos escape the detection

the backgrounds

- two identified leptons + missing energy in the final state







σ = 4.5 pb	σ = 15 pb	W: σ = 31 10^3 pb Z: σ = 3.5 10^3 pb
irreducible: same final state of the signal, exploit different kinematics of the production	there are two additional b-jets in the detector, due to the top decay, veto on jets (or on b-jets)	jets in the detector can give a lepton-like signature (non prompt leptons, or fake leptons form track+calo deposit): very high cross section

the triggers and first steps

muons are easily identified in the detector

electrons need to be separated from jets already at trigger level: higher thresholds and ID requirements

cross-triggers are not symmetric, to maximize the efficiency while keeping the rate low

doubleMu7
Ele17_CaloIdL_CaloIsoVL_Ele8_CaloIdL_CaloIsoVL
Mu8_Ele17_CaloIdL
Mu17_Ele8_CaloIdL

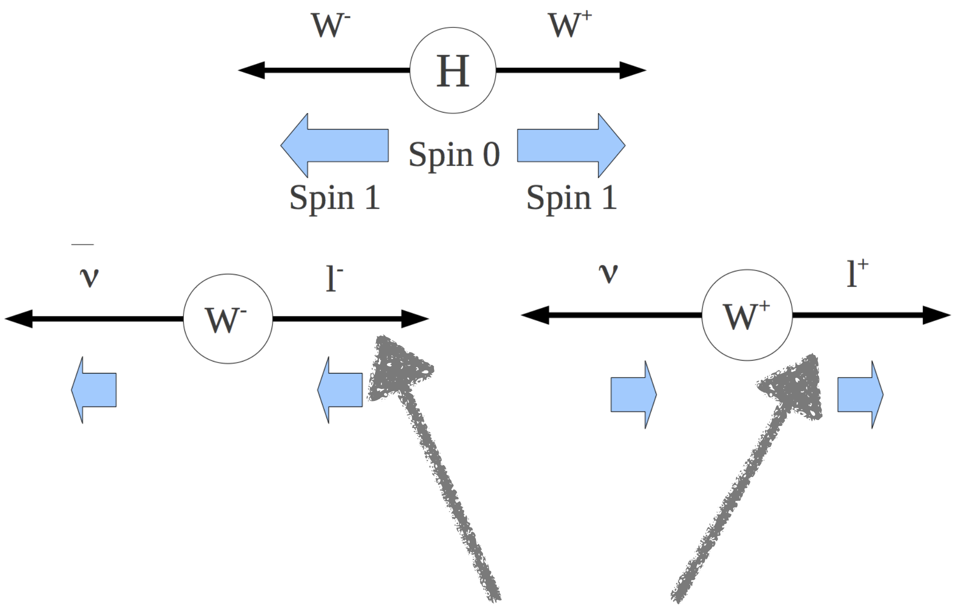
efficiencies calculated with the tag&probe on each leg separately

the analysis starts on samples selected by those triggers (if more than one trigger selects the events, any double countings have to be eliminated)

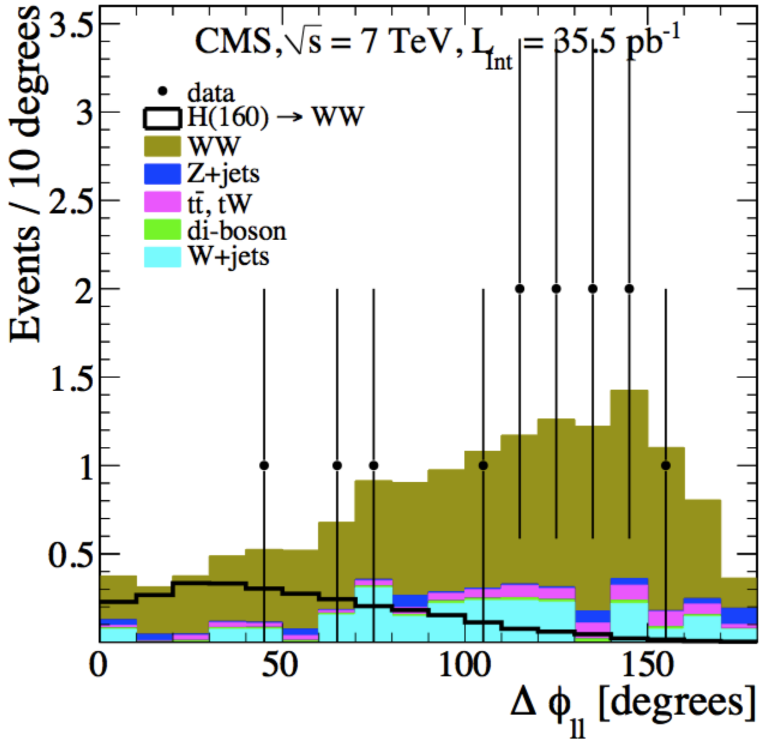
the analysis

- no invariant mass reconstruction --> **counting experiment**:
- isolate a phase space region where the signal-to-background ratio is maximized
- count the number of events
- compare to standard model expectations

example of a discriminating variable;



the two leptons are
close to each other



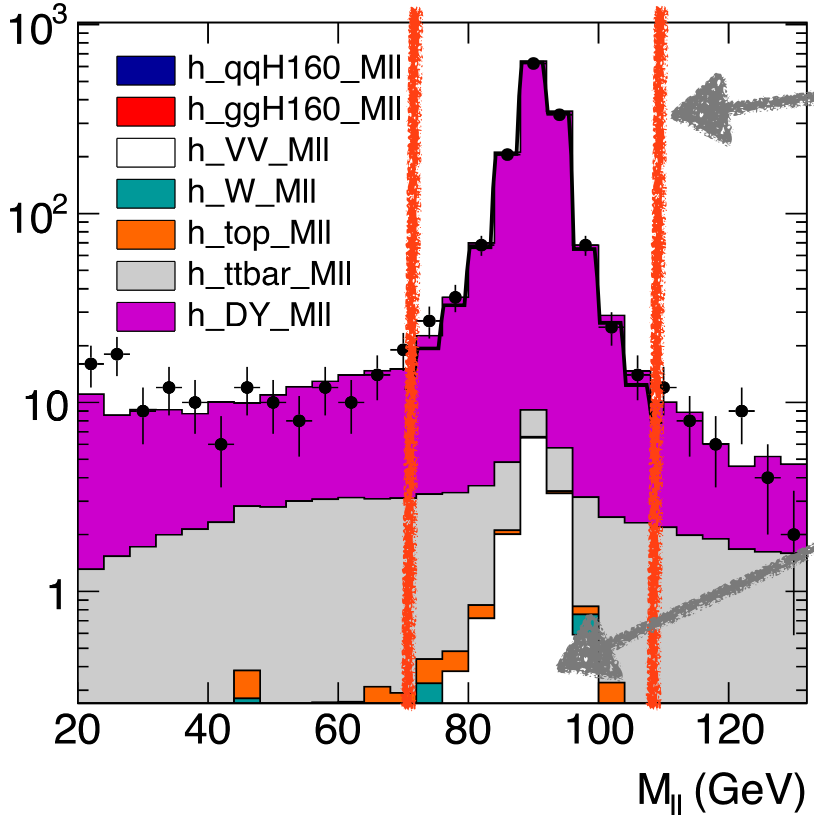
fight the backgrounds

$$\sigma = \frac{N_{obs} - N_{bkg}}{\varepsilon \cdot \int \mathcal{L} dt}$$

- evaluate (and subtract) backgrounds in the signal phase space region
- the simulation is reliable as much as the description of the theoretical model **and** the description of the detector
- determine the amount of **backgrounds from data** when these assumptions fail (the systematic uncertainty is expected to be big)
- the more the simulation is trusted, the less is compulsory to rely on data

get the cross-section

- assume a good knowledge of the background from simulation (efficiency wrt various selections)
- the absolute cross-section is the only missing information
- fit (or count) the simulation to data, and get the cross-section value



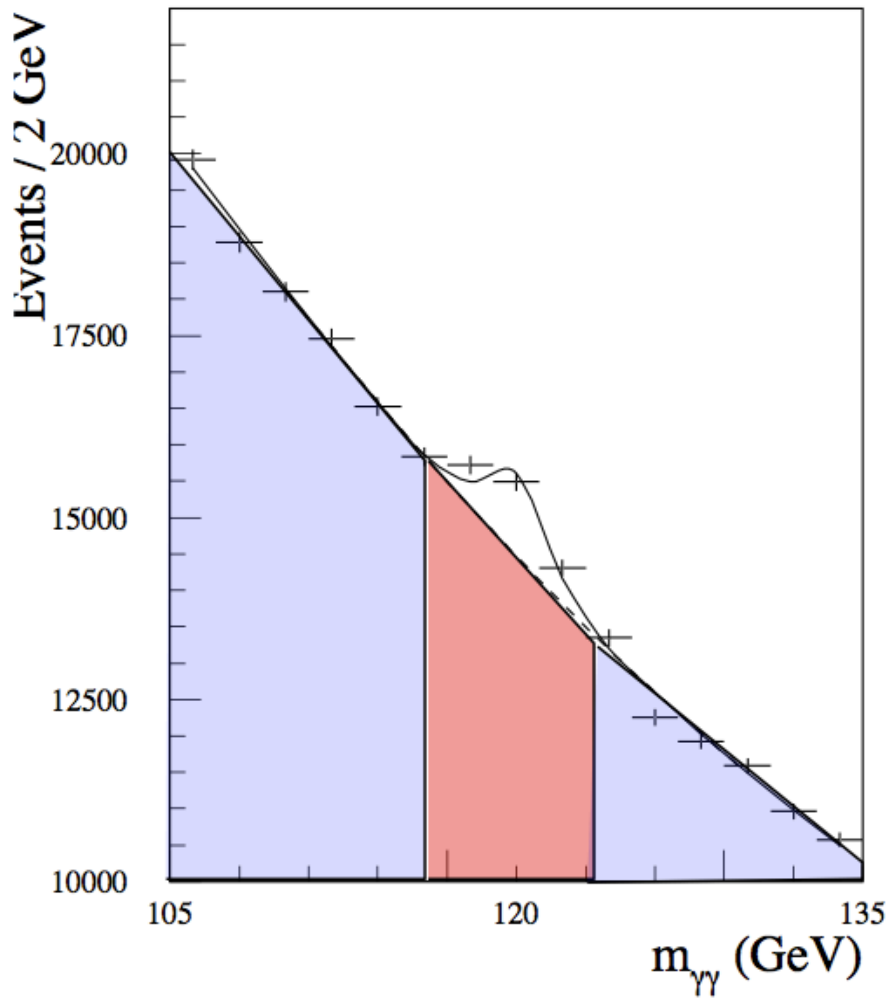
the Z contribution is dominant under the Z resonance peak

the uncertainty is due to the statistics available and the contamination of other samples

evaluate the **impact on the analysis**: probably does not need the same precision as a cross-section measurement

side-bands

- when the background is **expected to behave smoothly**, for example in case of random combinations
- assume a simple shape, and **extrapolate the background under the signal peak** from the sides
- **fit the distribution** with signal shape (a resonance) and background (exponential, linear)



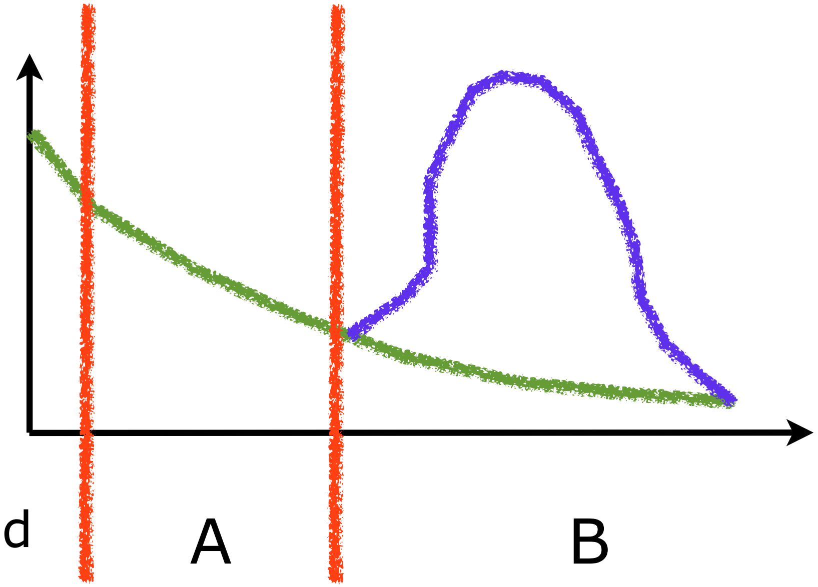
control region

- assume the **knowledge of the background shape**, at a certain level of the analysis
- fit the shape to data in a **signal-free region**, where that background is dominant and extrapolate it to the signal region
- in case of low statistics, **count the number of events** and extrapolate

$$N^{bkg-A}_{\text{inferred}} = N^{bkg-B}_{DATA} \left(\frac{N^{bkg-A}_{MC}}{N^{bkg-B}_{MC}} \right)$$

uncertainty due to the data statistics and other systematic sources

uncertainty due to: background model, control region contamination, propagation of other systematic sources, MC stats



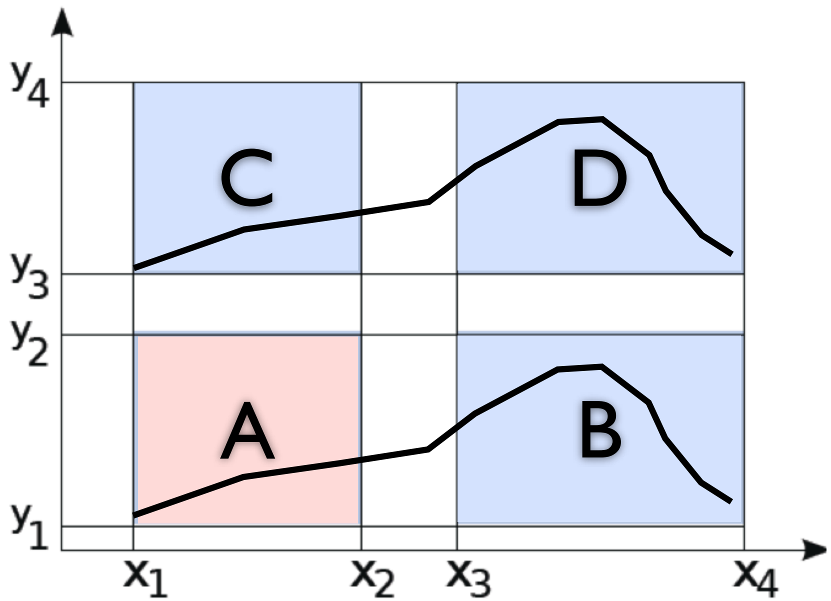
ABCD method

- measure also the **shape from data**, to perform the extrapolation from a control region to a signal region
- assume the bkg pdf to be factorized: $f^{bkg}(x,y) = f_x^{bkg}(x) \cdot f_y^{bkg}(y)$
- the **correlation check** done with simulation is a less stringent requirement than the good description of the shape
- in case of low statistics, **count the number** of events and extrapolate

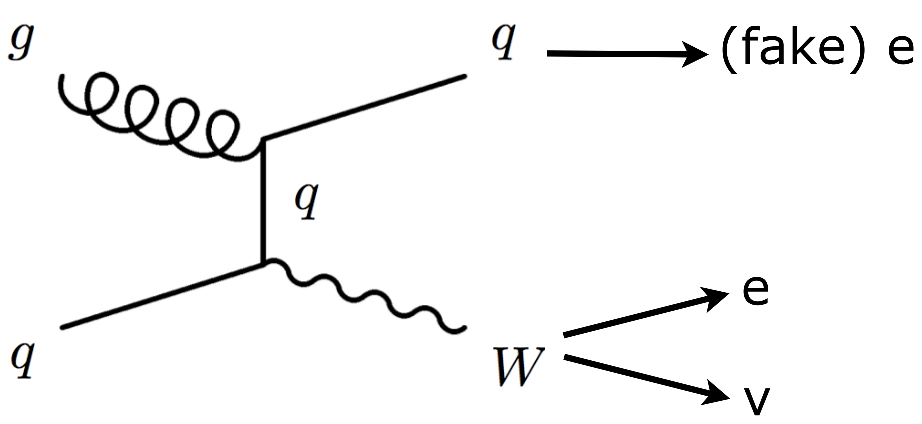
$$N_A^{bkg} = N_B^{bkg} \cdot \frac{N_C^{bkg}}{N_D^{bkg}}$$

uncertainty due to the data statistics and other systematic sources

uncertainty due to: control region contamination, propagation of other systematic sources



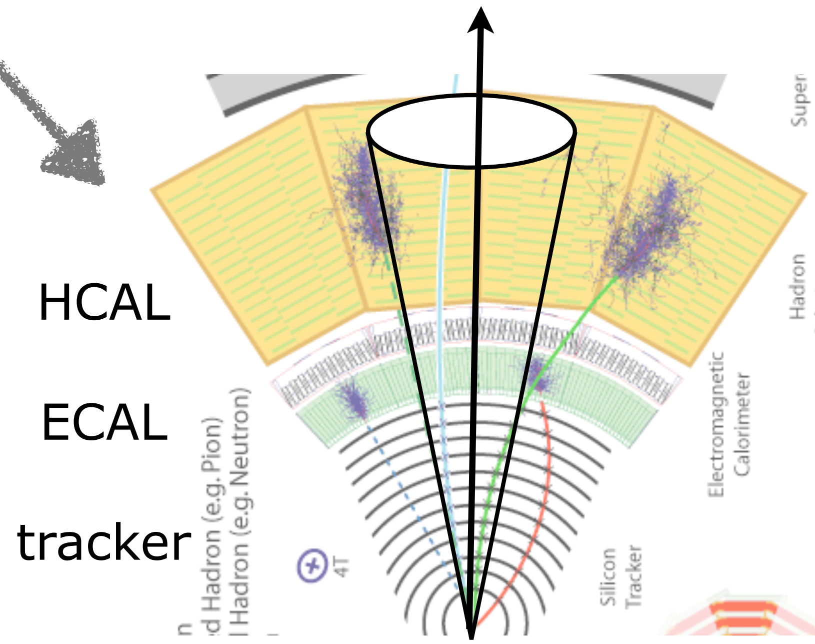
W+jets background



W: σ = 31 10^3 pb
Higgs: σ ~ 1 pb

- lepton identification and isolation are meant to reduce the probability for a jet to mimic a prompt lepton (fake rate)

- goodness of fit in the tracker
- track pointing to the primary vertex
- electrons shower shape variables
- goodness of fit for muons
- geometrical matching between different sub-detectors responses



in the simulation, the contamination critically depends on the detector description

fake rate

- **measure from data** the fake rate and use it to evaluate the background contamination
- sample with **no prompt leptons**: QCD dijet (di-jet trigger is therefore necessary for the Higgs search!)

$$r = \frac{\text{identified, isolated (fake) lepton}}{\text{fakeable object}}$$

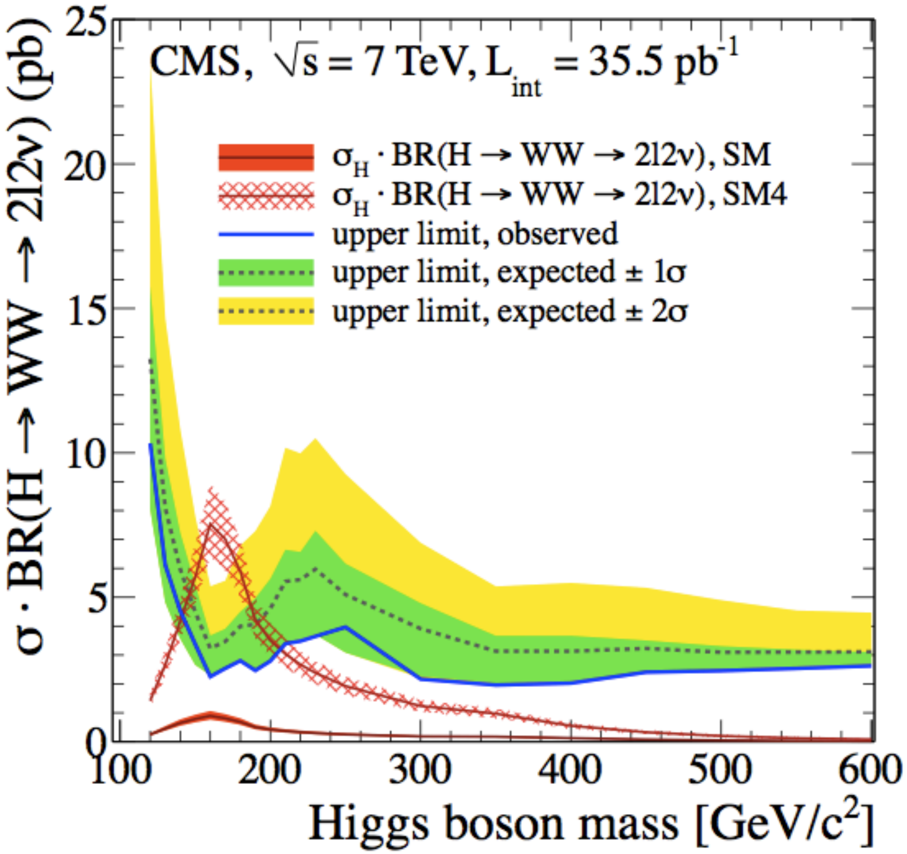
survives the selections of the analysis

definition depends on the object (e,μ) and the statistics available, and is part of the systematics

- select a (almost pure) **W+jets sample** by loosening the ID on one lepton (single lepton triggers necessary for Higgs search!)
- multiply by *r* to get the **number of events in the signal region**
- hypothesis: **the fake rate is the same**

the result

- with the number of measured events, and the estimated backgrounds, one draws the conclusion

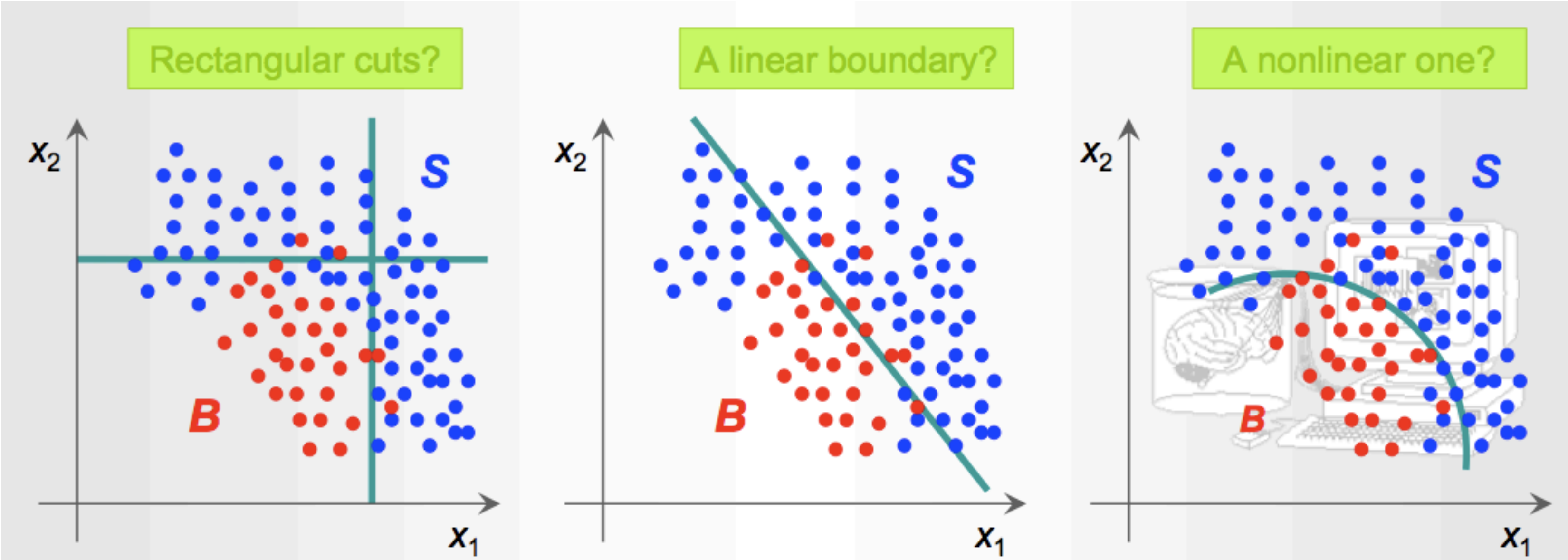


- for each Higgs mass, the selections choice has been optimized on a multi-dimensional rectangular grid

- is it the best choice?

multi variate techniques

multi variate techniques



- rectangular selections do not fully exploit the **topology of the events**
- build **more sophisticated discriminants** to separate signal from backgrounds
- need a **good knowledge** of both signal and background
- need high Monte Carlo **statistics**

Toolkit for Multivariate Data Analysis with ROOT, http://tmva.sourceforge.net/

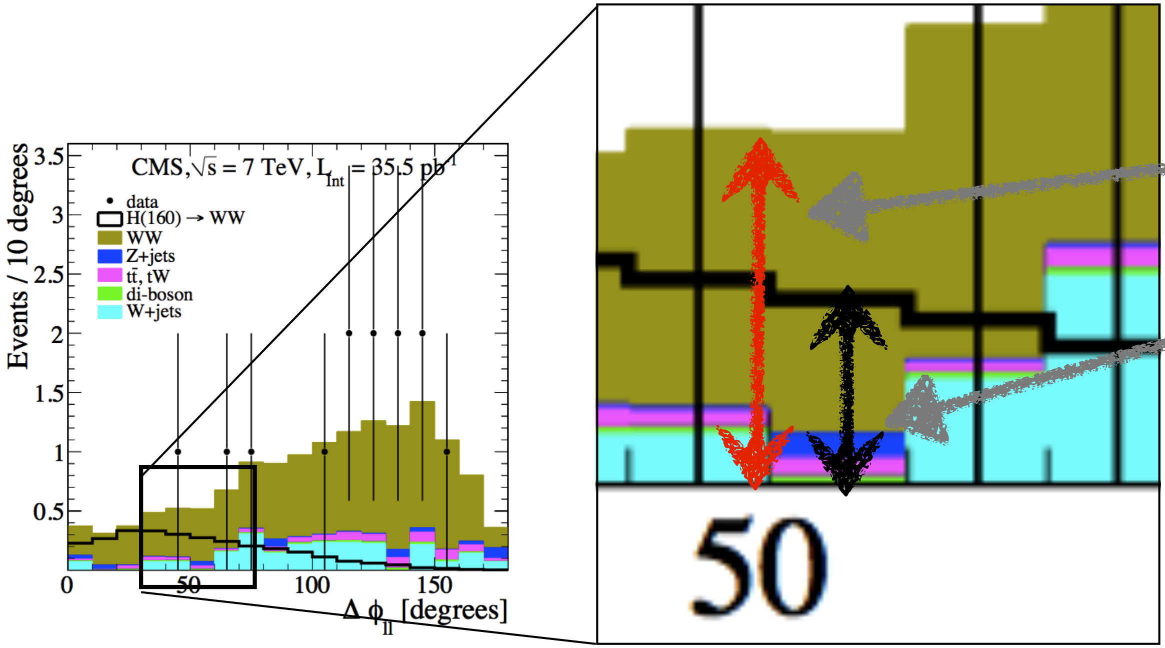
H. Voss, **Multivariate Data Analysis and Machine Learning in High Energy Physics**

likelihood discriminant

search for a classification of the events, that maps the set of the analysis variables into a single one

$$y_i = f(\vec{x}_i) : \mathbb{R}^n \rightarrow \mathbb{R}$$

$$R_L(i) = \frac{L_S(i)}{L_B(i) + L_S(i)}$$



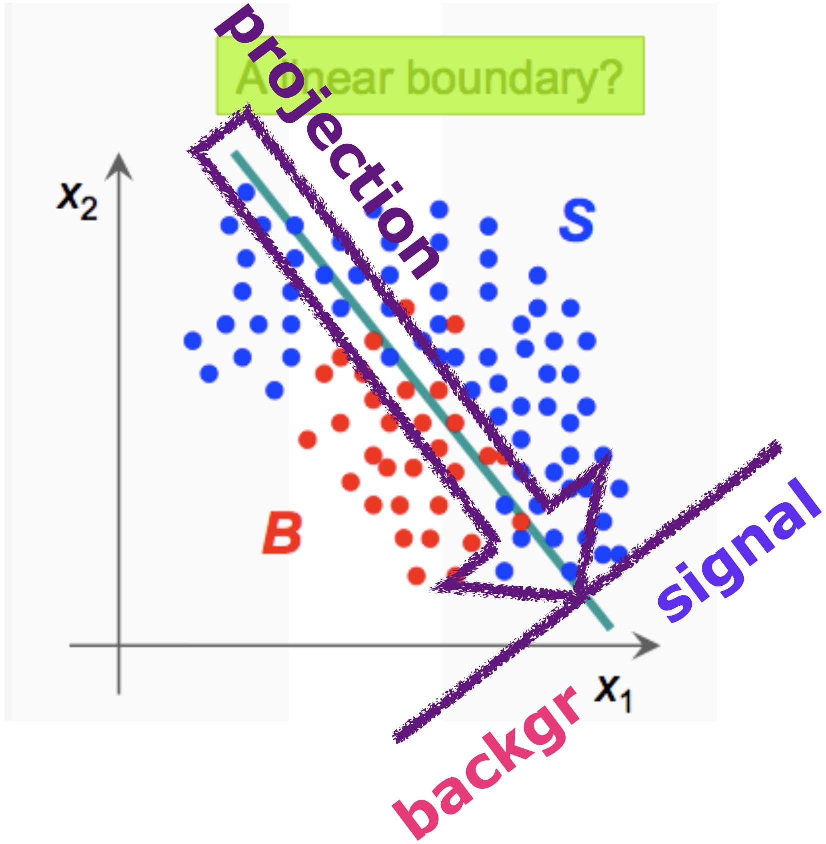
estimates for each event the confidence of being signal-like

For more, uncorrelated, variables: "easily" built
For linearly correlated variables, first decorrelate them

$$L_S(i) = f_S(\vec{x}_i) = \prod_{j \epsilon (\text{vars})} f_{S,j}(x_{ij})$$

fisher discriminant

- **project high-dimensional dataset onto a line** and perform classification in this one-dimensional space
- **optimization**: maximize the distance between means, while minimizing the variance within each class
- very effective with **linear correlations**



build the linear combination:

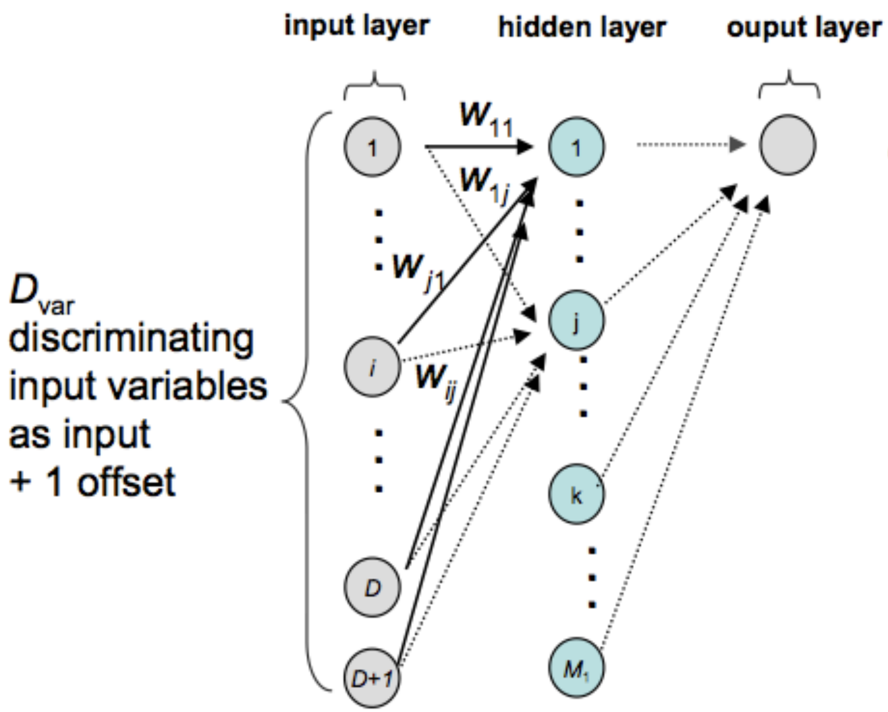
$$y(\vec{x}, \vec{w}) = w_0 + \sum x_i \cdot w_i$$

find the best weight by minimizing the criterion:

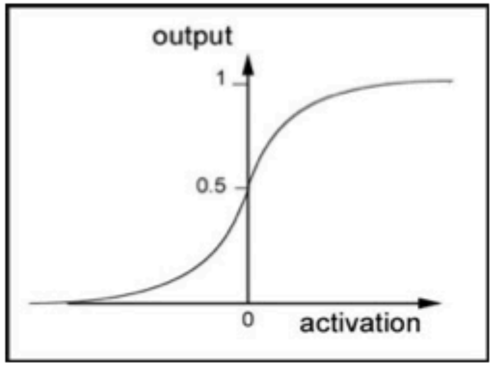
$$J(\vec{w}) = \frac{\left(\langle y_S \rangle - \langle y_B \rangle\right)^2}{\sigma^2_{y_S} + \sigma^2_{y_B}}$$

neural networks

- to cope with non-linear correlations, try a more sophisticated combination of the inputs



output: $$y(x)=\sum_{i}^{M} w_{0i} A\left(w_{i0} + \sum_{j=1}^{D} w_{ij} \cdot x_j \right)$$



"Activation" function e.g. sigmoid:

$$A(x)=\left(1+e^{-x}\right)^{-1}$$

or tanh

input variables, none of them is a smoking gun

factors of the "base" in which the non linear y is decomposed

the non-linear base element

need to find the weights, i.e. to train the neural network

on the training

loss function: how many times I make a mistake in the classification

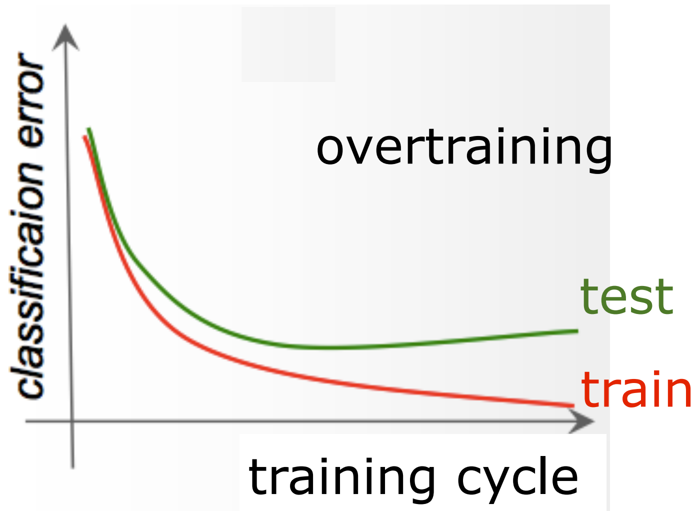
$$L(w) = \sum_{i}^{\text{events}} (\underbrace{y(x_i)}_{\text{predicted event type}} - \underbrace{y(C)}_{\text{true event type}})^2$$

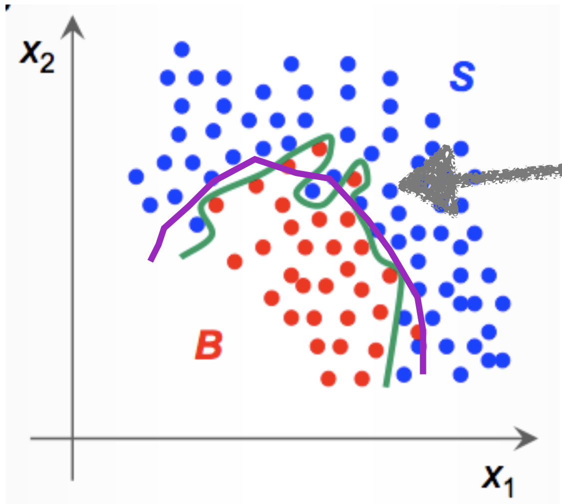
values of y:
1 = signal
0 = background

minimize the loss function:
- start from random weights
- change them according to the L gradient
- loop several times on the training samples

$$w^{n+1} = w^n + \eta \cdot \vec{\nabla}_w L \quad (w)$$

divide the simulated sample into **training** and **testing**,
continue the training until the performances on the training stabilize,
stop before the ones on the testing worsen

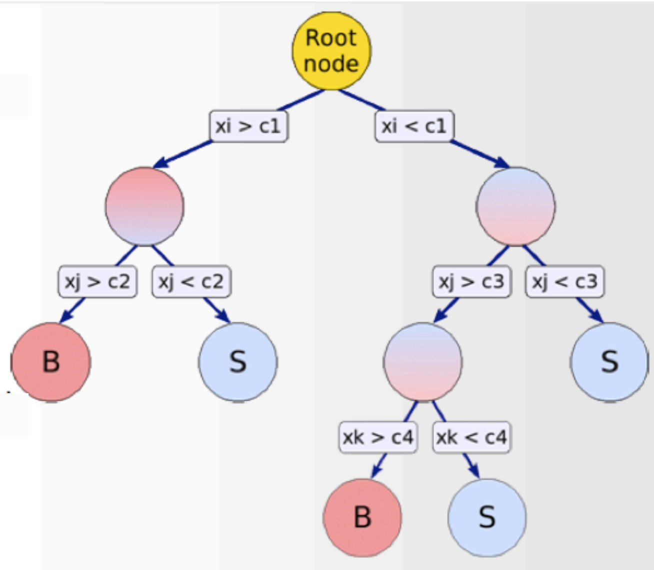


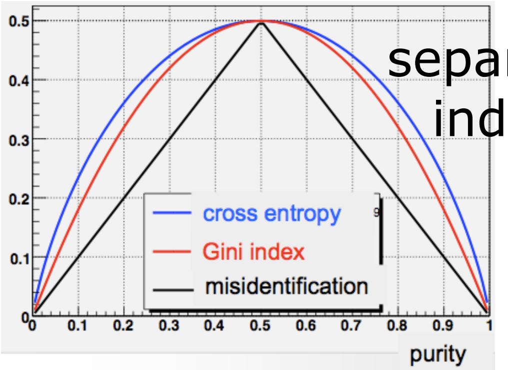


in overtraining, the NN is adapting to statistical fluctuations of the training sample

boosted decision trees

- rank the variables in terms of **discriminating power**
- apply **subsequent selections** in each of the variables
- **stop** when:
  - minimal #events per node
  - maximum number of nodes
  - maximum depth specified
  - a split doesn't give a minimum separation gain





separation indices

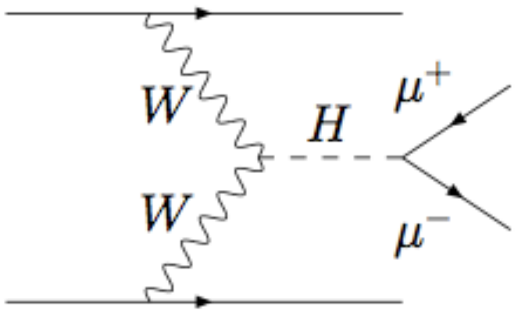
- in each final node (leaf) return S/B discrimination (discrete or continuous)
- independent of monotonous variable transformations
- immune against outliers
- weak variables are ignored
- very sensitive to statistical fluctuations in training data

boosting: combine a whole forest of Decision Trees, derived from the same sample, e.g. using different event weights.

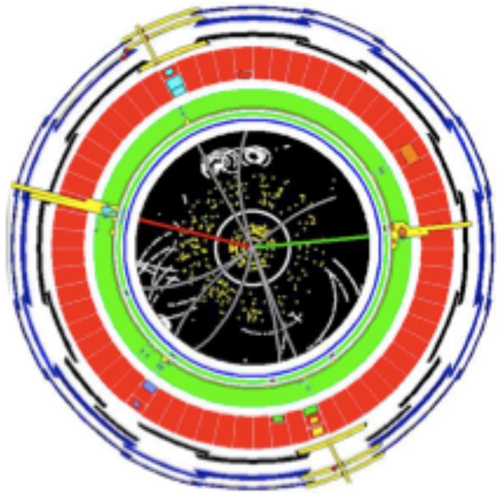
matrix elements

- the MVA techniques is **describe the final state topology** with a parametrization, built on the simulation
- matrix elements are this description, at the level of the physical process

need to include the effects due to the detector for each physics object considered



$\oplus$



$$P(\mathbf{x}|M_t) = \frac{1}{N}\int d\Phi \; |\mathcal{M}_{t\bar{t}}(p;M_t)|^2 \prod_{jets} f(p_i, j_i) f_{PDF}(q_1) f_{PDF}(q_2)$$

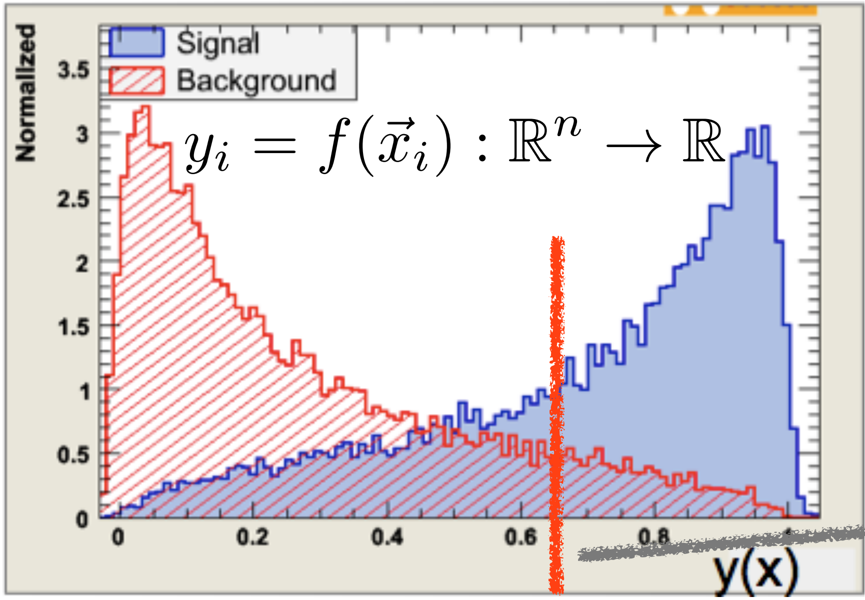
Phase-space Integral

Matrix Element

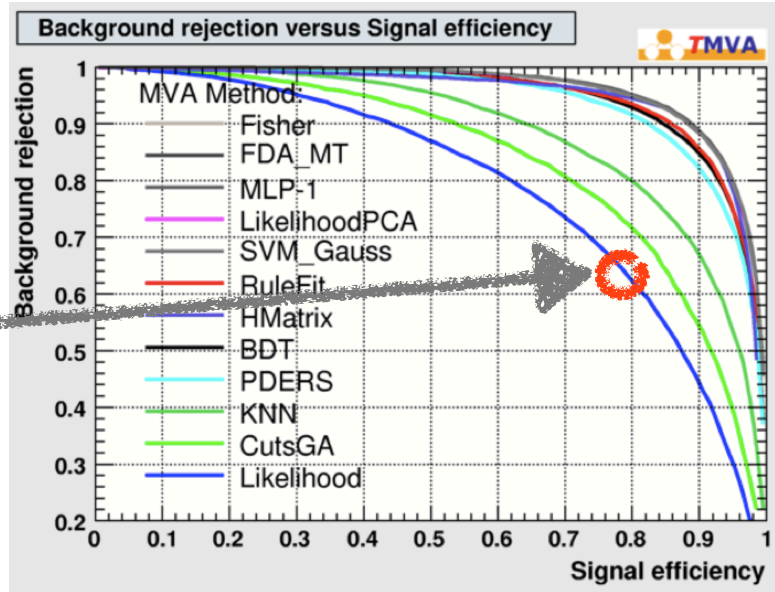
Transfer Functions

calculate the probability for each background and **build a likelihood ratio**

selecting the events



for each discriminant, now make the choice: what is signal, what is background?



Receiver Operating Characteristics (ROC) Curve: how efficiency versus purity

- choose the working point by maximizing a figure of merit:

$$\frac{S}{\sqrt{B}}$$

search: sensitivity over background fluctuations

$$\frac{S}{\sqrt{S+B}}$$

known signal: sensitivity over fluctuations of the total sample

$$\frac{S}{\sqrt{B+(\Delta B)^2}}$$

search: sensitivity over background fluctuations plus systematics

summary table

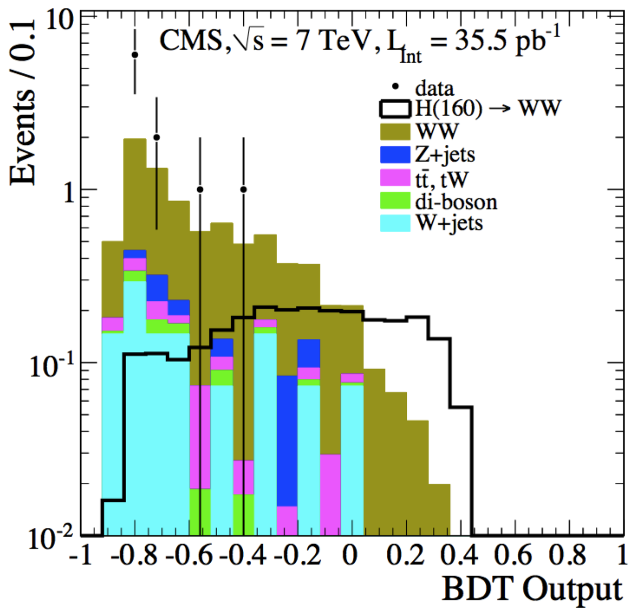
- useful table for the choice of the method to be used, among the ones provided by TMVA

Criteria		Classifiers								
		Cuts	Likeli-hood	PDERS / k-NN	H-Matrix	Fisher	MLP	BDT	RuleFit	SVM
Perfor-mance	no / linear correlations	😐	🙂	🙂	😐	🙂	🙂	😐	🙂	🙂
	nonlinear correlations	😐	🙁	🙂	🙁	🙁	🙂	🙂	😐	🙂
Speed	Training	🙁	🙂	🙂	🙂	🙂	😐	🙁	😐	🙁
	Response	🙂	🙂	🙁/😐	🙂	🙂	🙂	😐	😐	😐
Robust-ness	Overtraining	🙂	😐	😐	🙂	🙂	🙁	🙁	😐	😐
	Weak input variables	🙂	🙂	🙁	🙂	🙂	😐	😐	😐	😐
Curse of dimensionality		🙁	🙂	🙁	🙂	🙂	😐	🙂	😐	😐
Transparency		🙂	🙂	😐	🙂	🙂	🙁	🙁	🙁	🙁

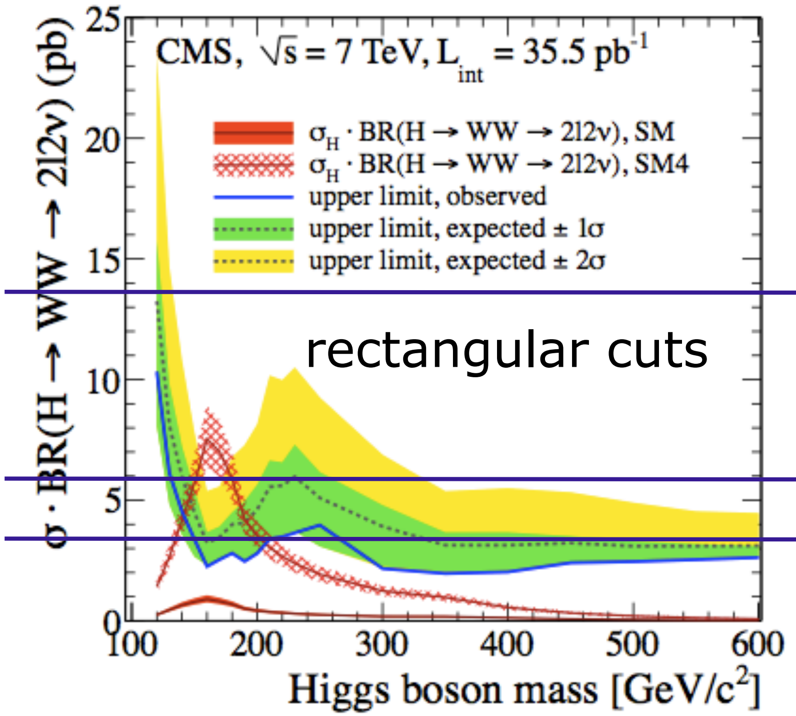
what about systematics?

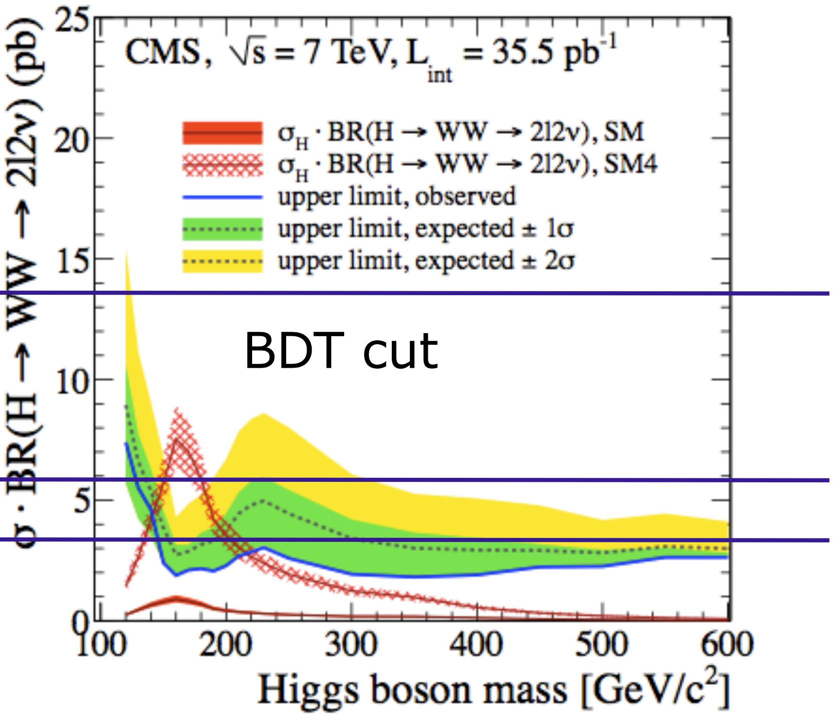
- **in terms of training**, a systematic effect yields a sub-optimal discriminant
- **in terms of results**, a systematic in the model reflects in the the efficiency and purity estimates, and in the event counts
- **compare data to MC** for the *y(x)* variable
- train the discriminator on **different models**
- try and understand effects by training on **reduced sets of variables**

the H > WW > lvlv case



boosted decision tree output, for the SM Higgs hypothesis of 160 GeV mass





in conclusion

- the luminosity
- the trigger, from the point of view of the analysis
- the reconstruction and detector response
- the simulation
- differential cross-section measurement: a di-jet correction
- searches: the H > WW > lvlv
- multivariate techniques

- data are arriving... when the going gets tough, the toughs get going, ... and have fun!