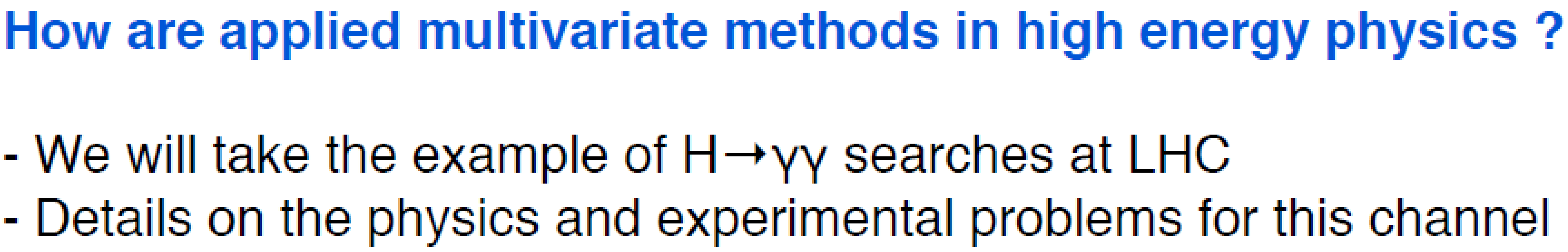
Machine Learning and Multivariate Techniques in HEP data Analyses



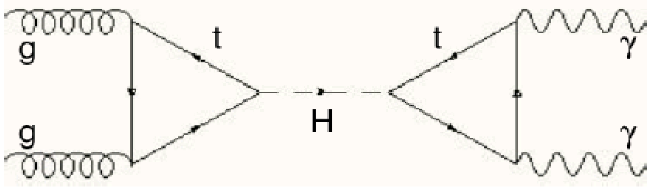
How are applied multivariate methods in high energy physics ?

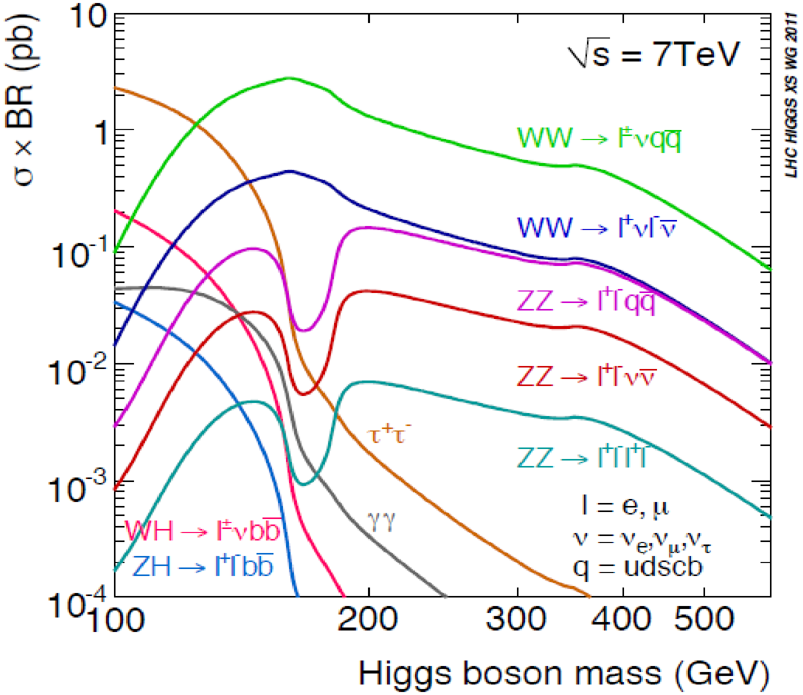
- We will take the example of H→γγ searches at LHC
- Details on the physics and experimental problems for this channel

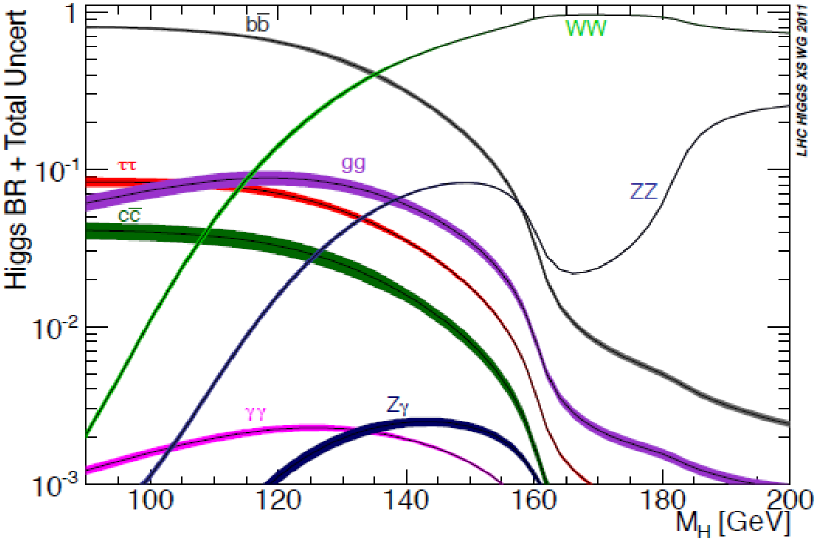
Extracted from lectures by N. Chanon, ETH Zurich.

H->γγ at LHC: signal

- H→γγ produced mainly via gluon fusion
- Branching ratio ~0.2%

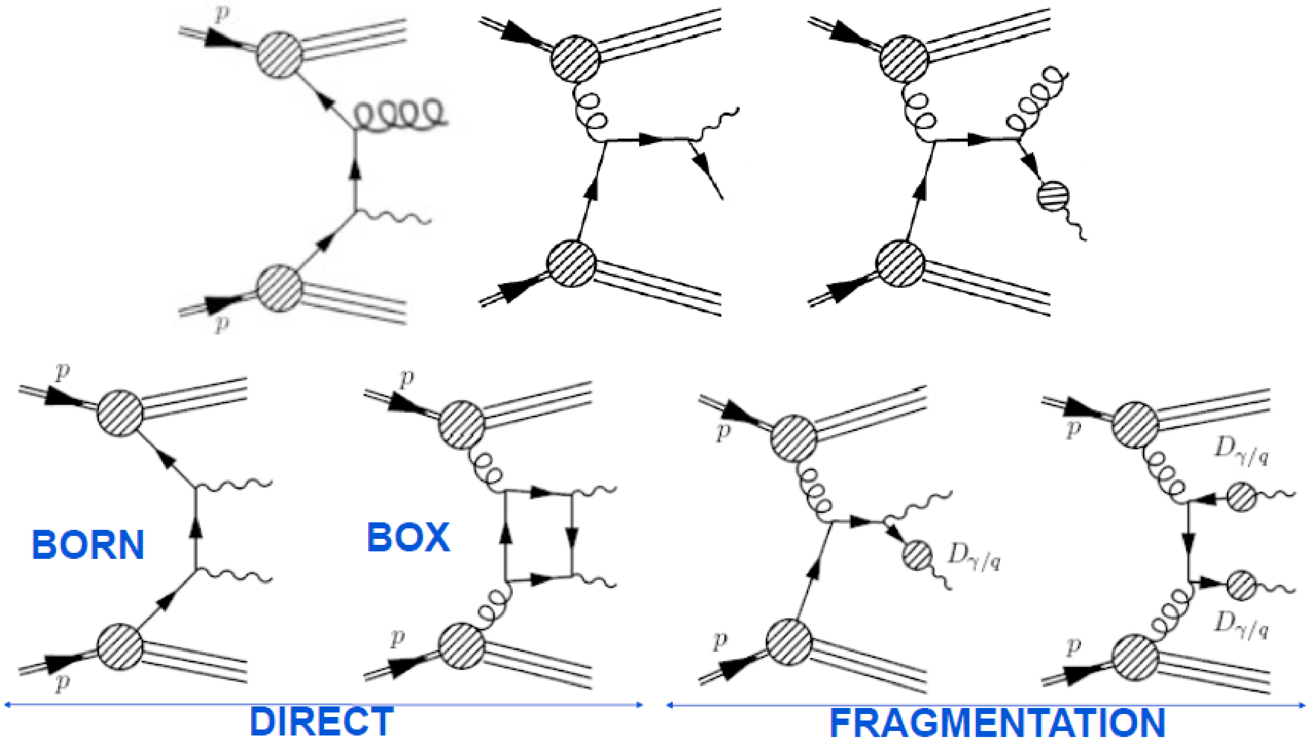


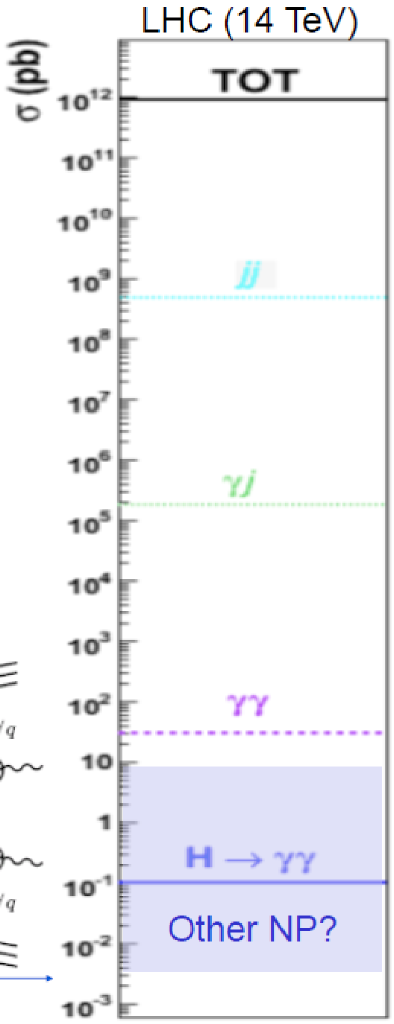




H->γγ at LHC: background

- Huge background over 9 order of magnitude in cross-section
- Dijet, gamma-jet processes with jets faking photons
- Diphoton continuum



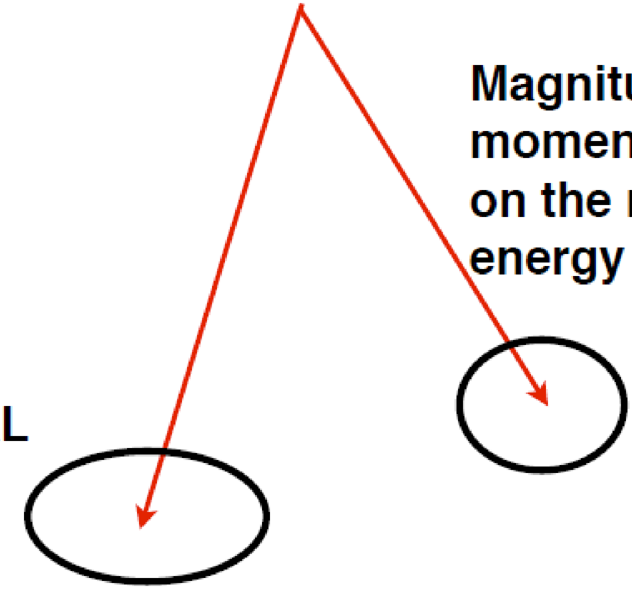


H->γγ at LHC: issues

- This channel suffers from small branching ratio and huge background.
- But it has the best sensitivity at low mass
- Reason : CMS and ATLAS have very good resolution on the γγ invariant mass

Main issues for H→γγ :

- Vertex identification
- Energy resolution
- Photon identification
- Kinematics
- Analysis sensitivity



Hard interaction vertex

Magnitude of the 4-momentum depends on the reconstructed energy

Position in the ECAL

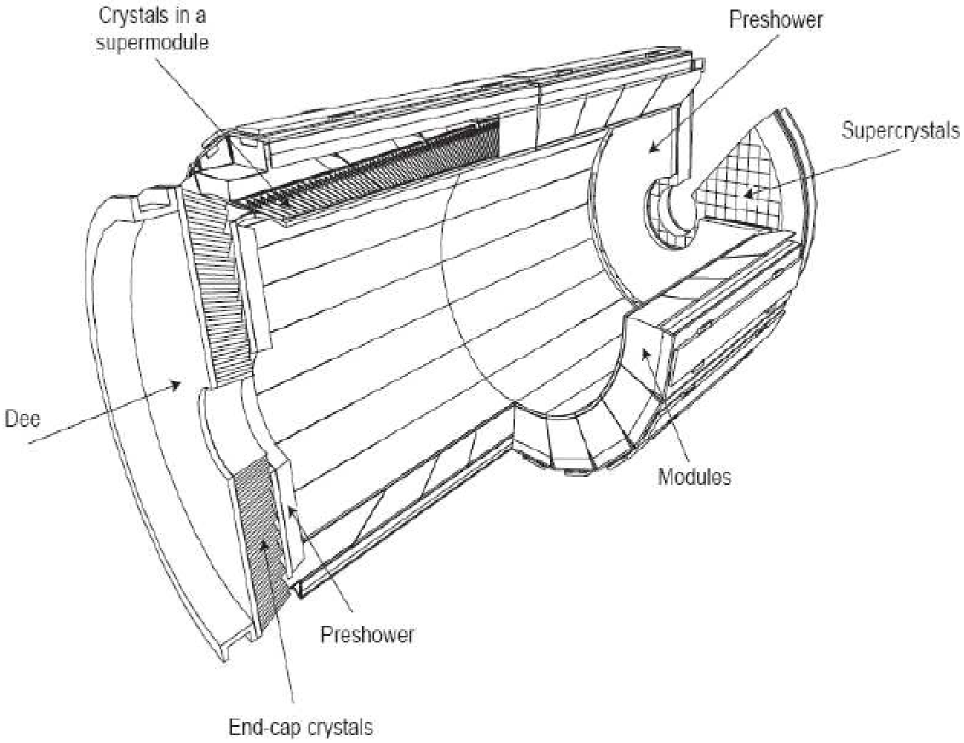
Shape of the energy deposit

CMS electromagnetic calorimeter

The **ECAL** is made of scintillating crystals of PbWO4 :

- **Barrel** : 36 “supermodules” with 1700 crystals each (coverage IηI<1.48)
- **Endcaps** : 268 “supercrystals” with 25 crystals each (coverage 1.48<IηI<3.0)

Furthermore, a **preshower** made of silicon strip sensors is located in front of the endcaps (1.65<IηI<2.6)



Energy resolution (measured in electron test beam) :

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E(GeV)}} \oplus \frac{b}{E(GeV)} \oplus c$$

a = 2.8% stochastic term

b = 12% noise term

c = 0.3% constant tern

CMS electromagnetic calorimeter

The **ECAL** is made of scintillating crystals of PbWO4 :

- **Barrel** : 36 “supermodules” with 1700 crystals each (coverage Iηl<1.48)
- **Endcaps** : 268 “supercrystals” with 25 crystals each (coverage 1.48<lηl<3.0)

Furthermore, a **preshower** made of silicon strip sensors is located in front of the endcaps (1.65<lηl<2.6)

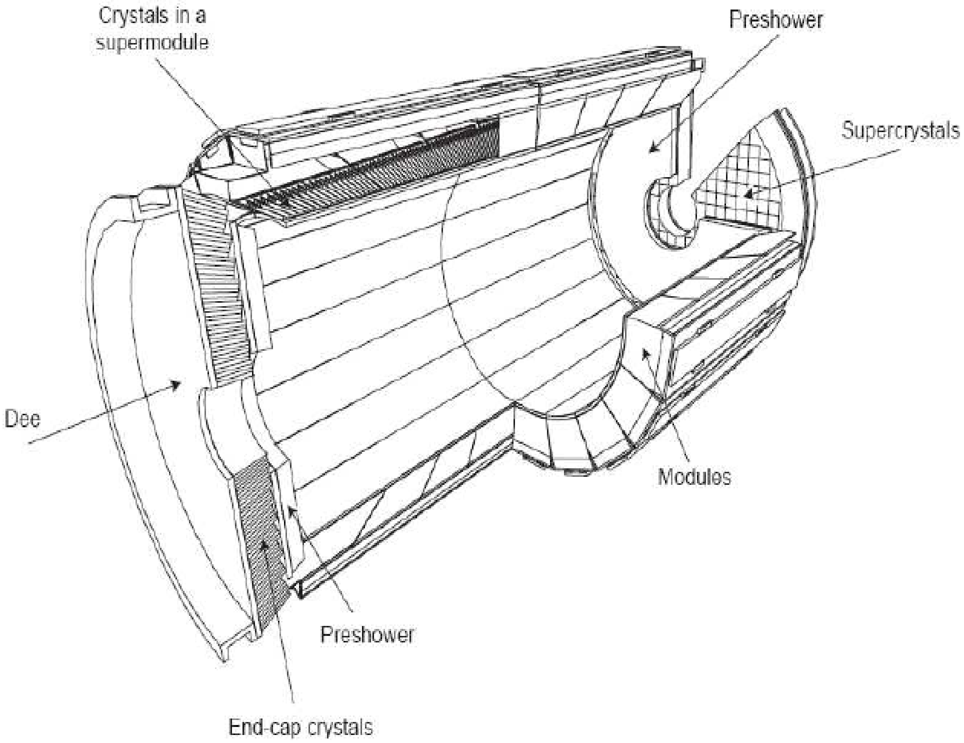
Energy resolution (measured in electron test beam) :

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E(GeV)}} \oplus \frac{b}{E(GeV)} \oplus c$$

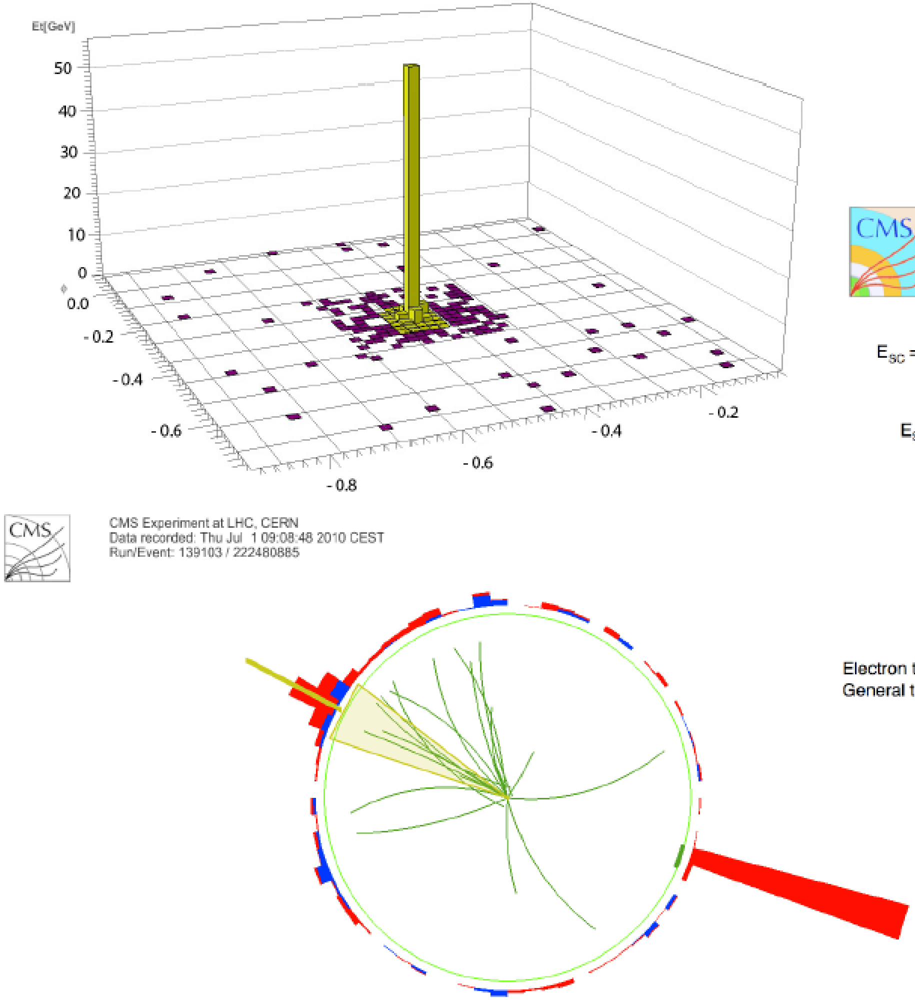
a = 2.8% stochastic term

b = 12% noise term

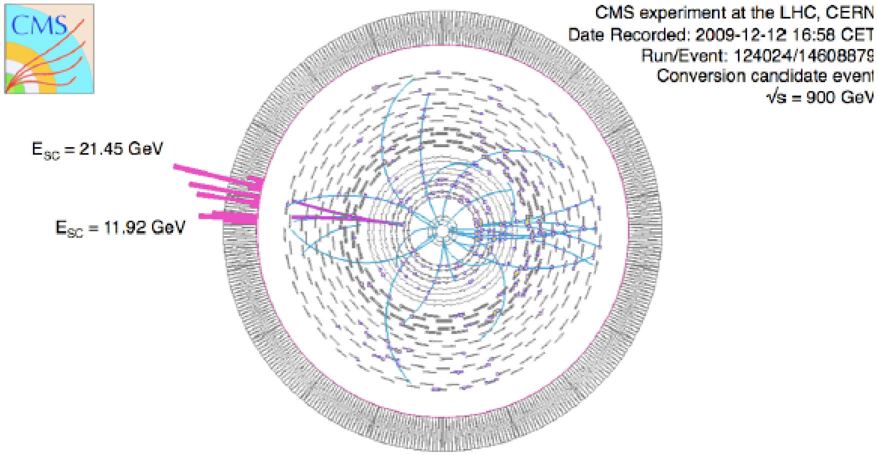
c = 0.3% constant tern



Photon event display



CMS Experiment at LHC, CERN
Data recorded: Thu Jul 1 09:08:48 2010 CEST
Run/Event: 139103 / 222480885



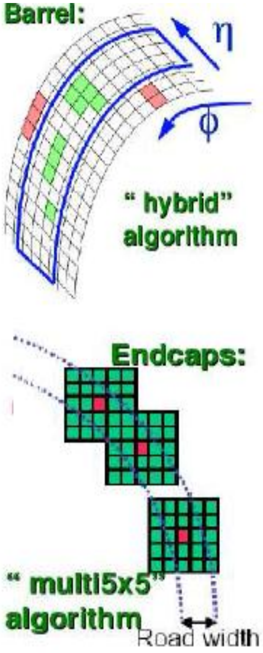
CMS experiment at the LHC, CERN
Date Recorded: 2009-12-12 16:58 CET
Run/Event: 124024/14608879
Conversion candidate event
√s = 900 GeV

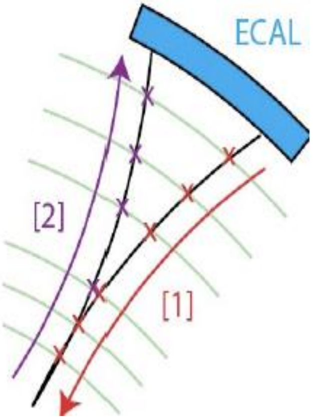
Electron tracks are shown in purple, and their superclusters in pink in the ECAL.
General tracks are in blue and tracker clusters (silicon strips) are shown by small squares.

Photon reconstruction

Photons are reconstructed with energy deposits in **ECAL** crystals

- **Barrel** : take advantage of the 3.8 T magnetic field which bends the charged particles trajectory (in case of a photon conversion)
- **Endcap** : merge contiguous 5 × 5-crystal matrices around the most energetic crystals





Converted photons :

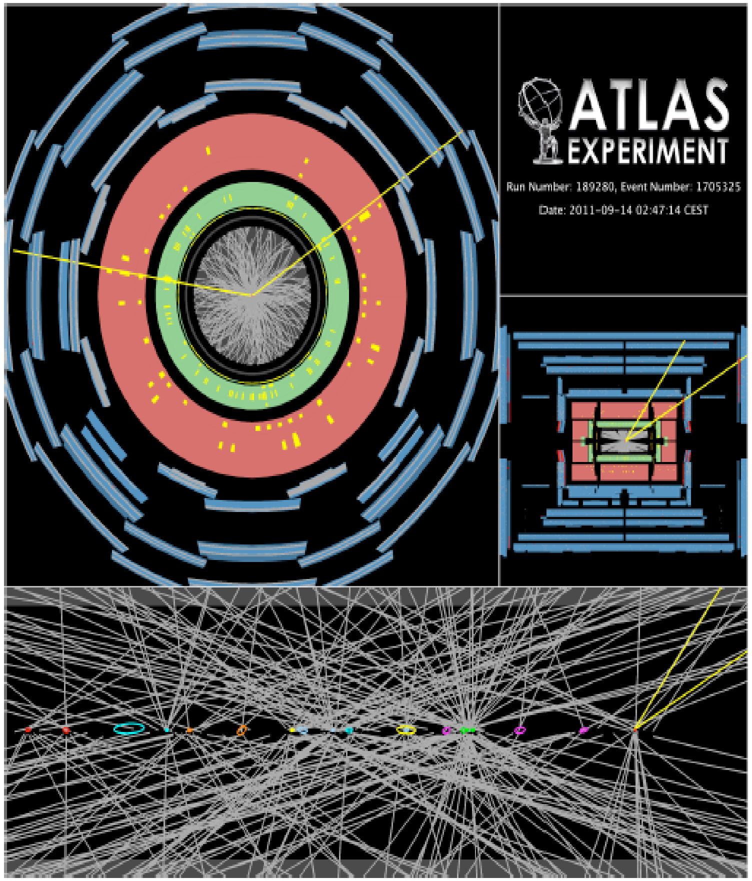
- Start from **energy deposits in ECAL**
- **Track finding** proceeds inward and outwards, taking into account electron energy loss by bremsstrahlung
- Select the e^+/e^- pair with the best vertex fit χ^2

H->γγ at LHC: vertexing

- Up to ~20 pile-up events per bunch crossing in 2011
- How to identify the hard interaction vertex ?
- Usual vertexing algorithm uses reconstructed tracks. Choose the vertex having the **highest sum pt squared**.

For H→γγ we have additional information :

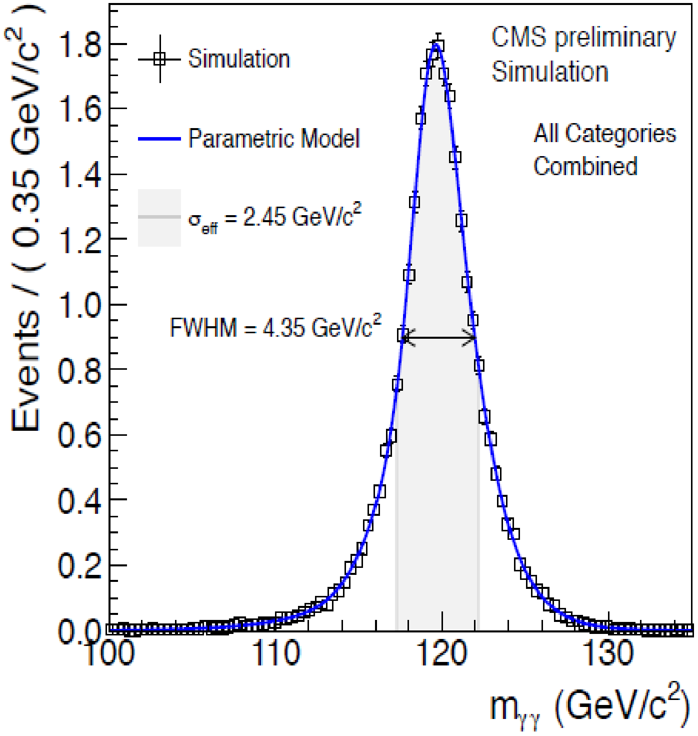
- ATLAS : calorimeter pointing (photon conversion tracks pointing)
- CMS : multivariate method using tracks + diphoton kinematics, combined with conversion information


ATLAS EXPERIMENT
Run Number: 189280, Event Number: 1705325
Date: 2011-09-14 02:47:14 CEST

H->γγ at LHC: energy resolution

- Higgs natural width is zero from an experimental point of view in the γγ channel
- So the **experimental width** is driven by **how well the photon energy is reconstructed** (once measured the position in the ECAL and the vertex found)

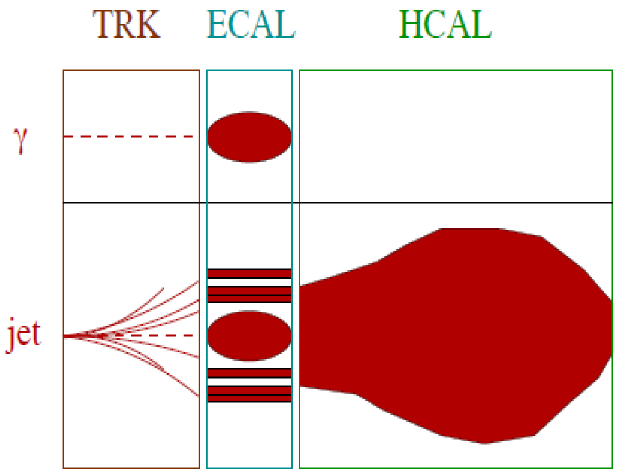
- CMS : PbWO4 crystals calorimeter, subject to **loss of transparency**
- Clustering of the energy deposited is affected by the **tracker material** in front of the ECAL
- Corrections to get back the reconstructed energy to the energy at the vertex might not be optimal
- CMS : energy regression

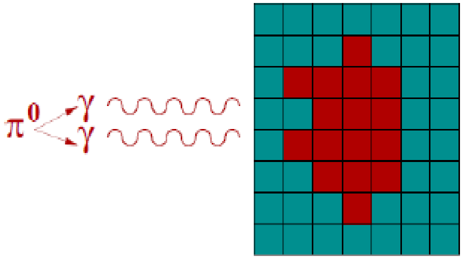


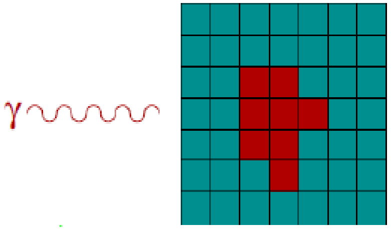
H->$\gamma\gamma$ at LHC: photon identification

Why jets can fake photons ?

- Isolated boosted pi0 decaying to 2 photons can be reconstructed in one single supercluster





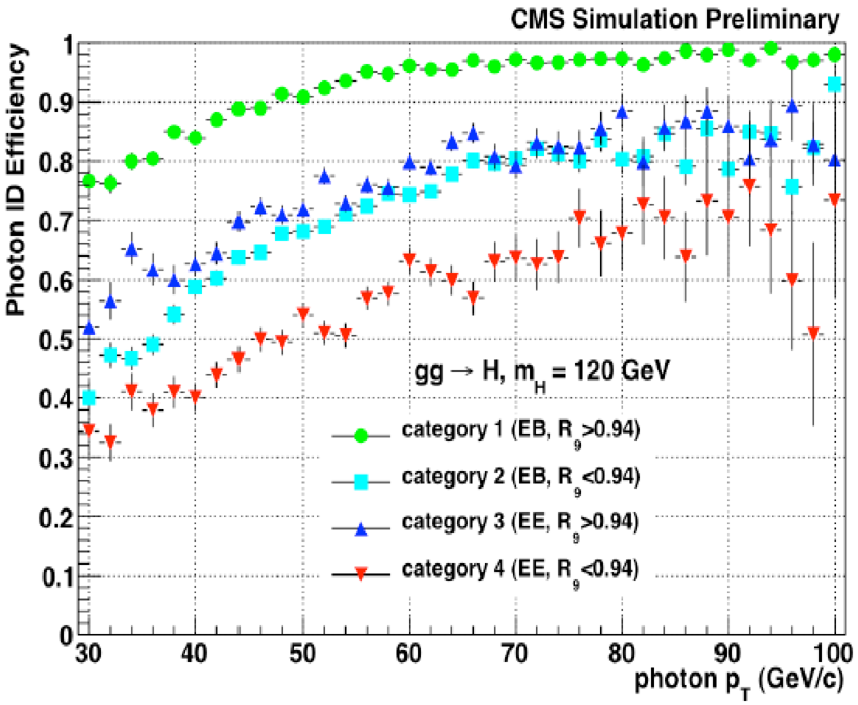


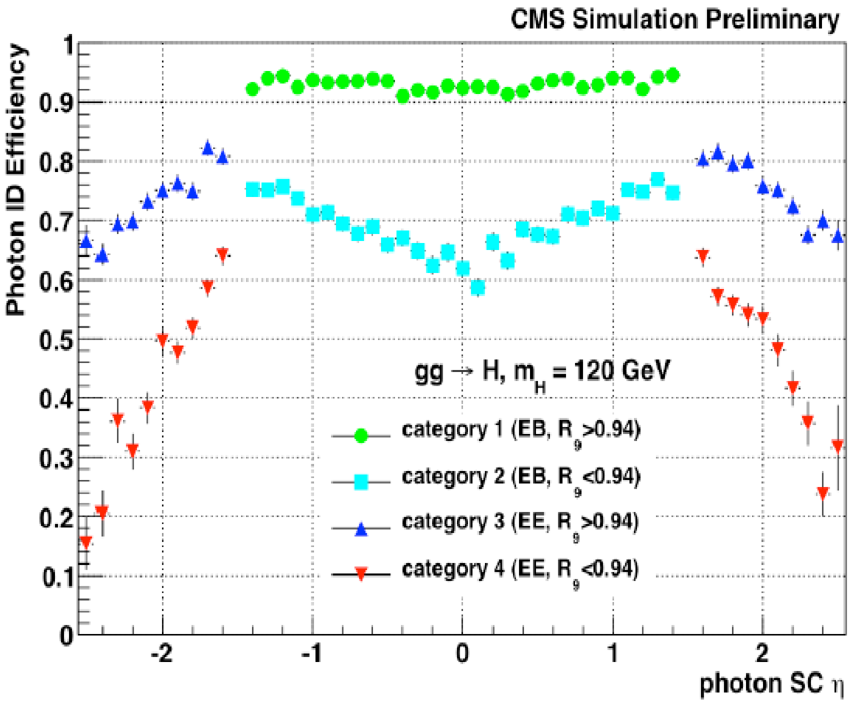
Photon identification :

- **Electron rejection** : the energy deposit should not be matched to hits in the pixel detector
- The **transverse shape** of the energy deposits in ECAL should be compatible with a single photon shower
- **Isolation** : in a cone **$\Delta R<0.4$** around the photon, use $\sum E_T$ of energy deposits in **ECAL**, **HCAL** and $\sum p_T$ of the charged particles measured in the **tracker**

H->γγ at LHC: photon identification

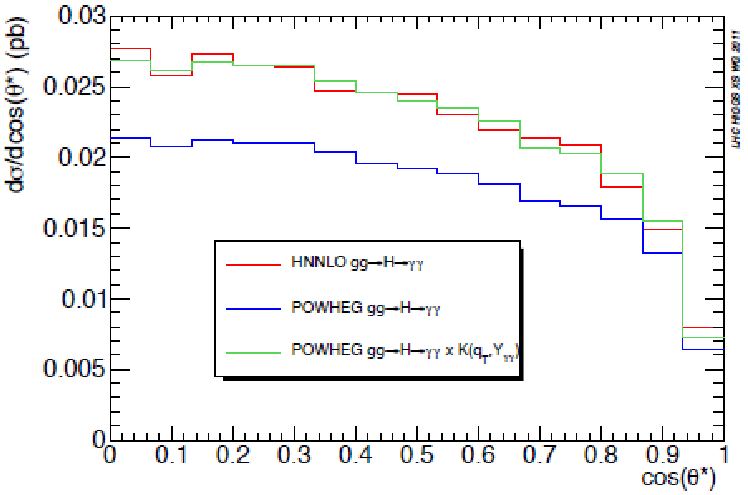
- In CMS photon identification is achieved using cuts on :
- **3 cluster shape variables** : H/E, transverse shape of the electromagnetic deposit, R9 = E3x3/Esupercluster
- **3 Isolation variables** : ECAL+HCAL+tracker in 0.3, 0.4 cones according to the wrong and right vertex hypothesis, Tracker isolation alone

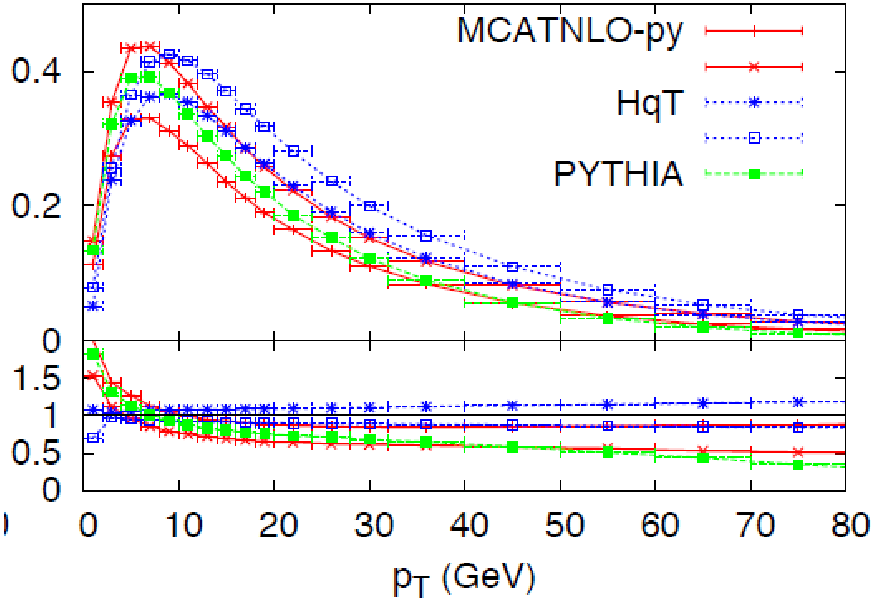


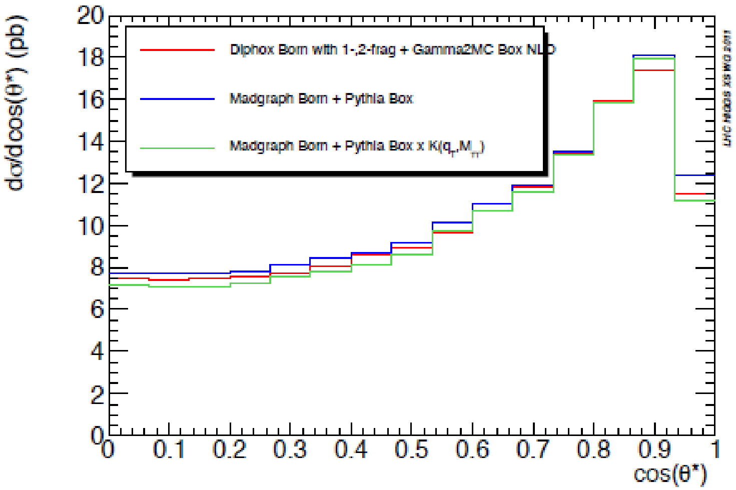


H->γγ at LHC: kinematics

- **Photon pT threshold :** usually asymmetric, pT>40,30 or 25 GeV
- **cos(θ*) :** can be discriminant in some kinematical regime
- **Diphoton pT** as discriminant variable : a myth for the gluon fusion

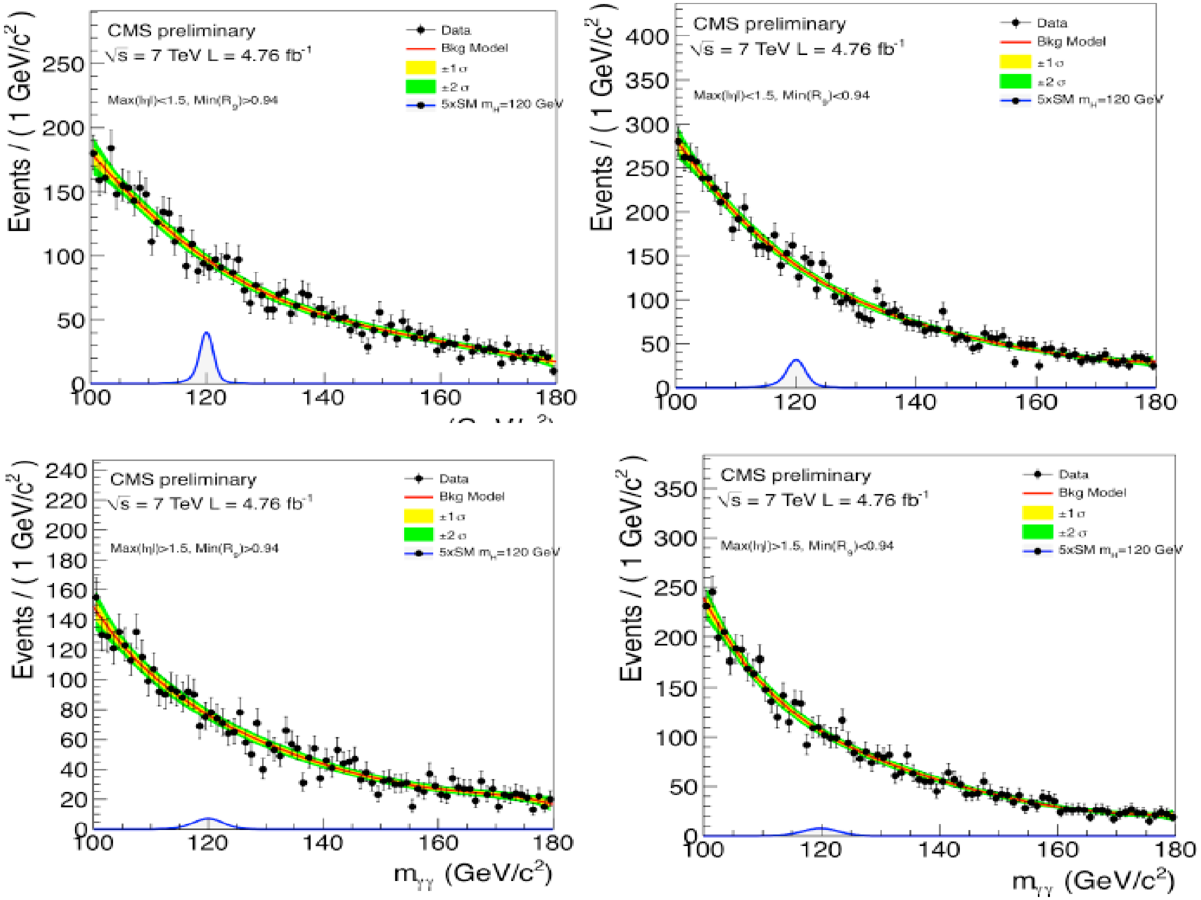






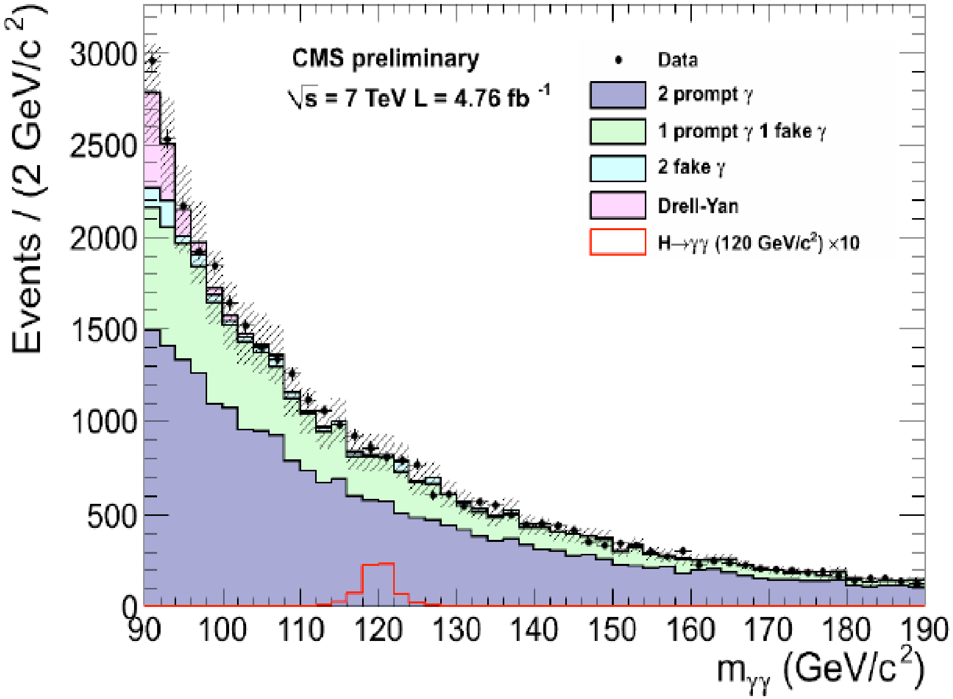
H->γγ at LHC: diphoton categories

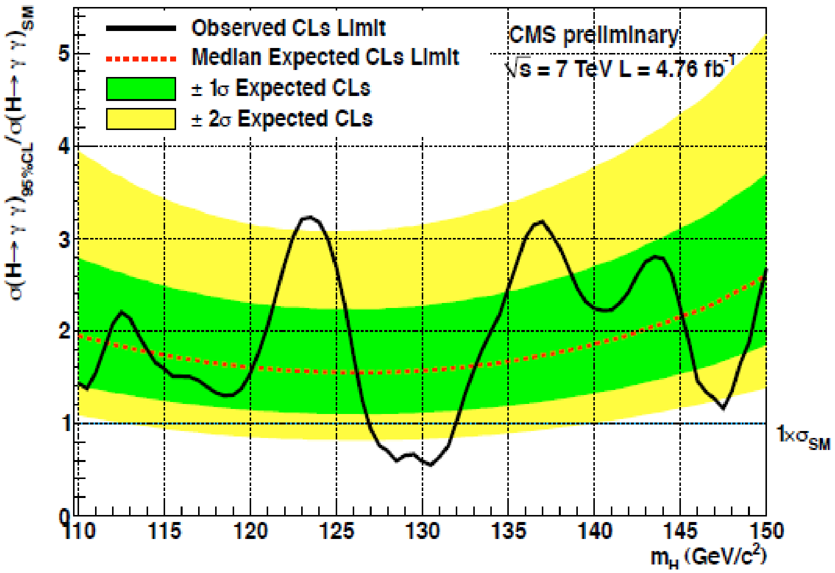
- CMS : 4 eta-r9 categories to improve mass resolution and s/b ratio
- ATLAS : 9 categories (eta / conversion / pt thrust)



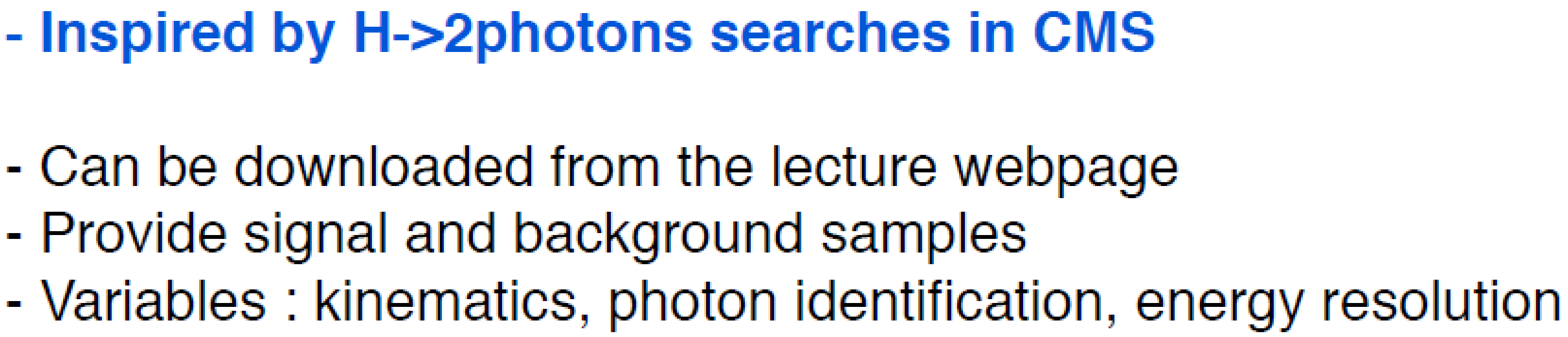
H->γγ at LHC: analysis sensitivity

- Fit of the diphoton invariant mass distribution in data (how to choose the fit function ?)
- MC is not used to derive the sensitivity
- Unbinned CLs method





H->$\gamma\gamma$ at LHC: exercises



- Inspired by H->2photons searches in CMS

- Can be downloaded from the lecture webpage
- Provide signal and background samples
- Variables : kinematics, photon identification, energy resolution

https://people.phys.ethz.ch/~pheno/Lectures2012_StatisticalTools/mva_exercise/

H->γγ at LHC: exercises

Installing ROOT

- Simplest option is probably to download the binaries (just unpack it)
- Do not forget to source bin/thisroot.sh

Download the exercises on the webpage

- Pdf with instructions and questions
- The samples in the ROOT format

Having a look to the samples :

- root -l Sample.root

Running TMVA

- Go the the directory tmva
- Classifier training can be launched using TMVAClassification.C
- Once the classifiers trained, one can investigate with TMVAGui.C
- One can also have a look to the training output : TMVA.root

Exercise: samples

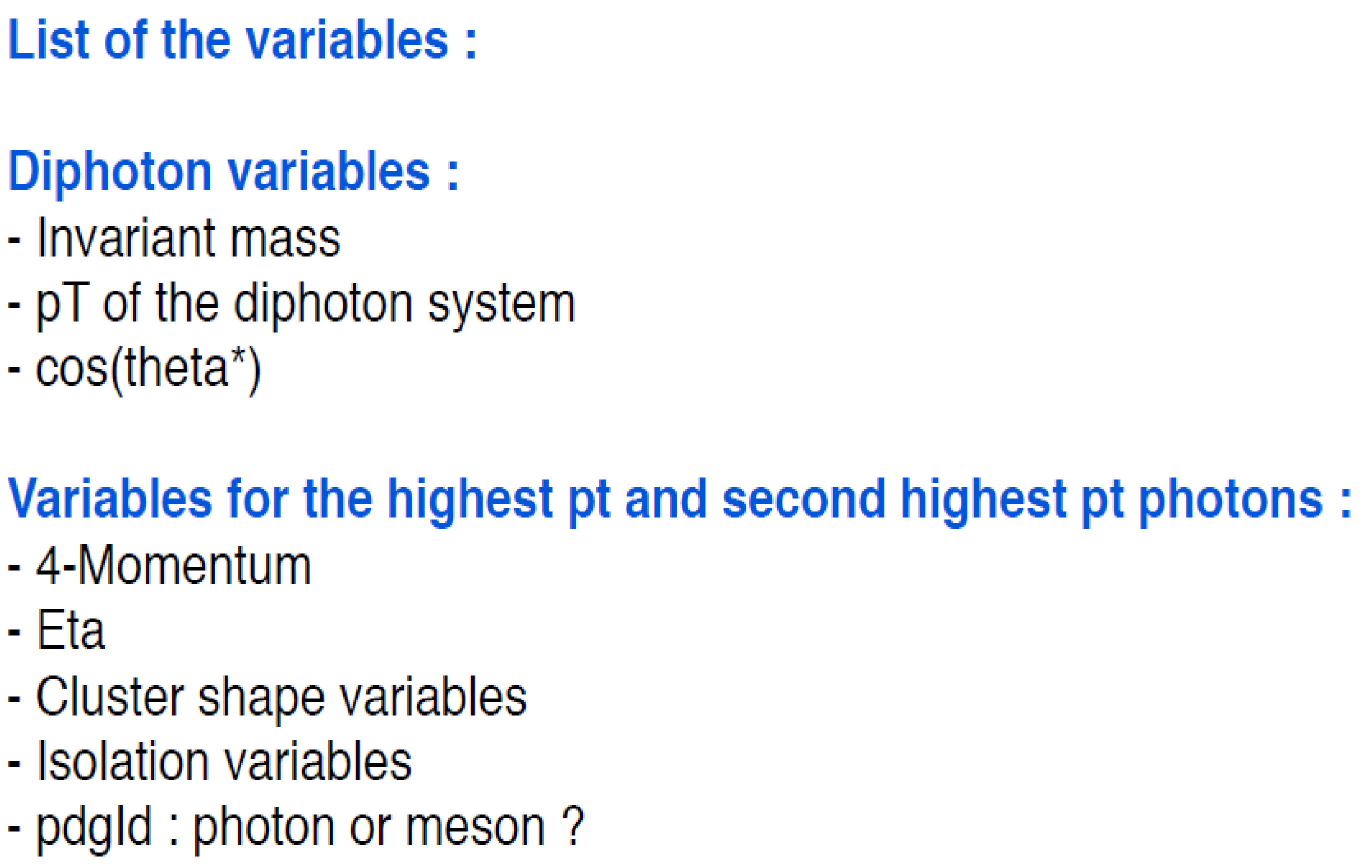
Samples provided were generated using Pythia :

- **gg→H→γγ mH=120 GeV** (100kevt generated) - forget other production mechanisms
- **γγ Born** (1Mevt)
- **γγ Box** (1Mevt)
- **γ+Jet** (20Mevt - lack of statistics)
- Dijet background was not generated (1000x more events would have been needed due to the small jet→γ misidentification rate)

Experiment simulation

- Events have been passed into a (home-made) program which emulates the experiment
- Energy smearing due to finite detector effects
- Energy deposits variables
- Important correlations taken into account

Exercise: variables



List of the variables :

Diphoton variables :

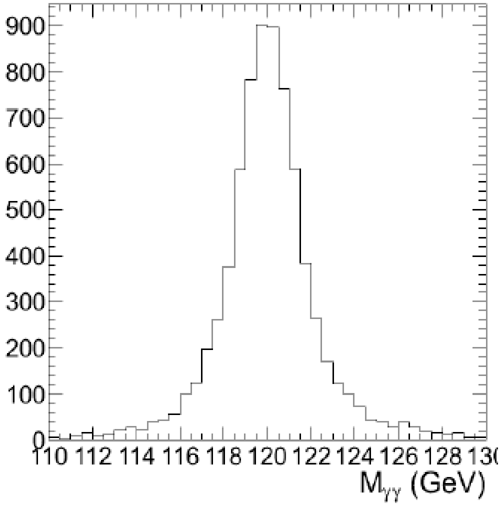
- Invariant mass
- pT of the diphoton system
- cos(theta*)

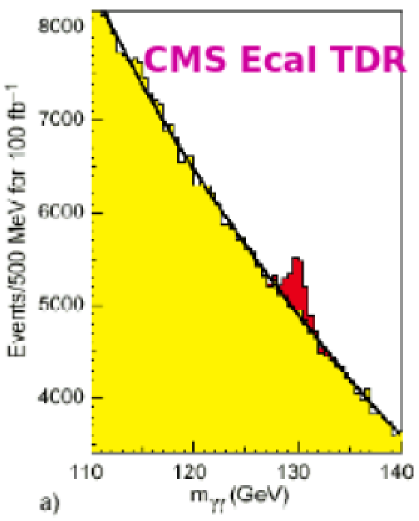
Variables for the highest pt and second highest pt photons :

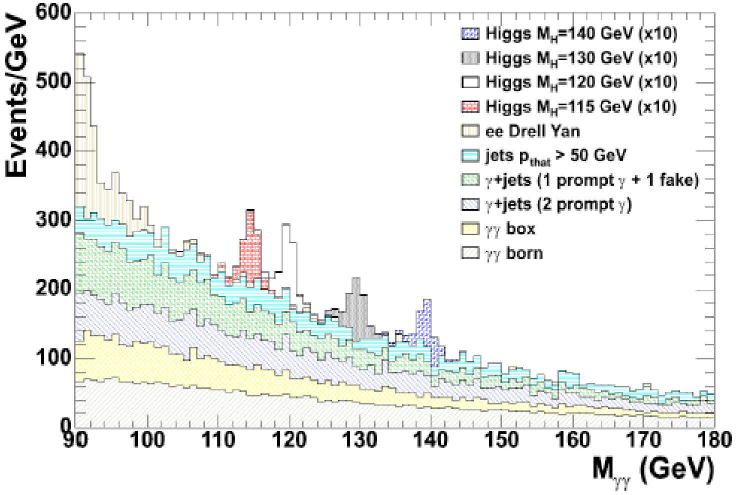
- 4-Momentum
- Eta
- Cluster shape variables
- Isolation variables
- pdgId : photon or meson ?

Exercise: invariant mass

- In the exercise, the diphoton mass resolution is different from the one we get in reality, but the order of magnitude is the right one
- Look for a sharp peak in a steeply falling background
- After photon identification, the jet-jet and gamma+jet background is much reduced

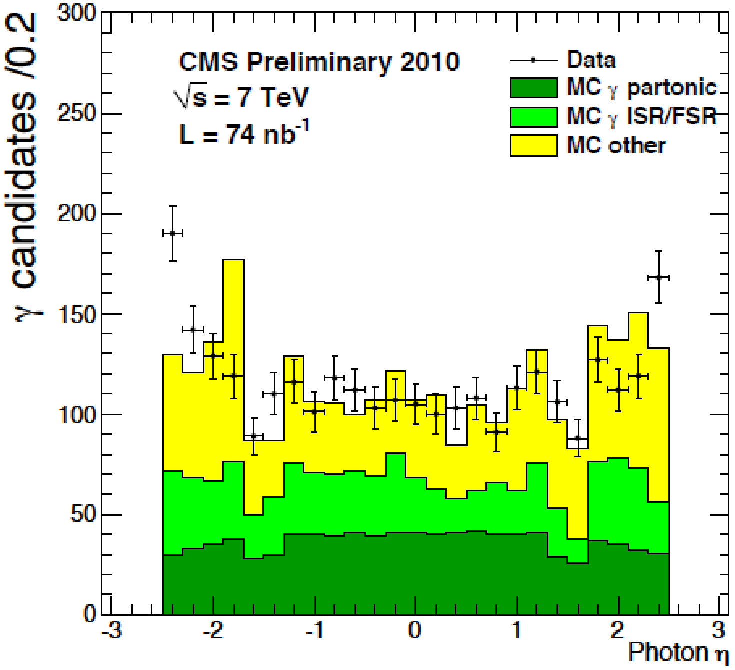






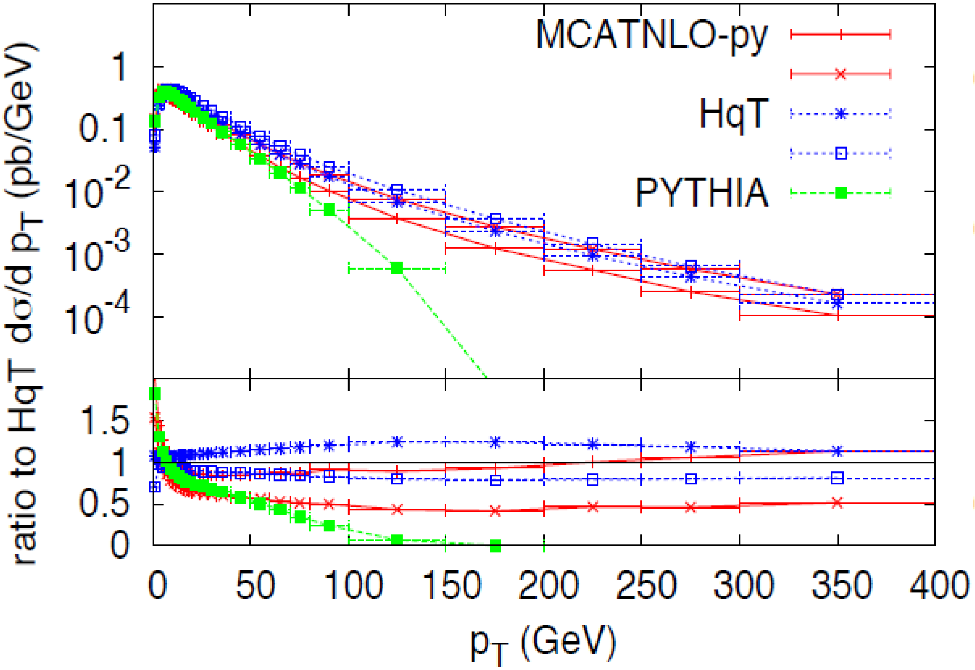
Exercise: photon kinematics

- In the exercise, the photon Pt is smeared
- The reconstruction efficiency (η-dependent) is not taken into account. This gives more photons in the barrel-endcap transition region than expected experimentally



Exercise: photon kinematics

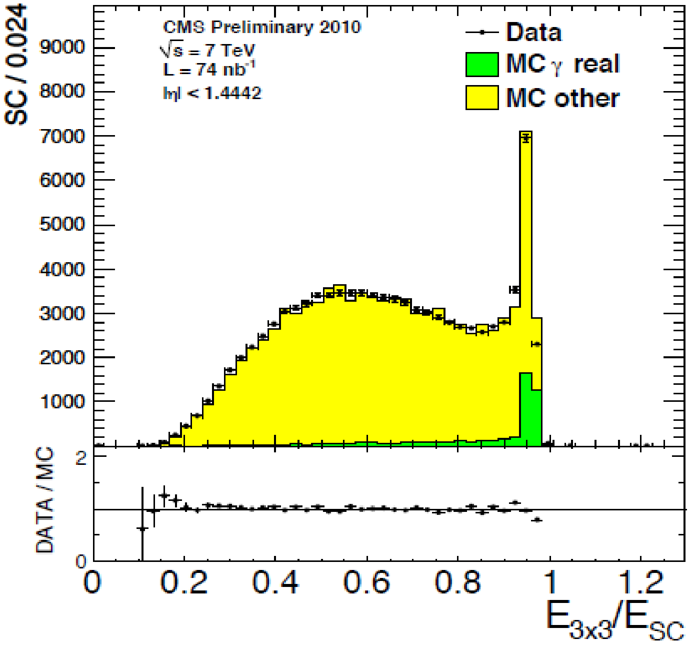
- The diphoton transverse momentum is only LO+LL from Pythia here
- Can be used for the purpose of demonstration, but the discriminating power is much reduced in reality
- cos(theta*) can also be used, but it is difficult to make it very discriminant with the trigger thresholds actually used

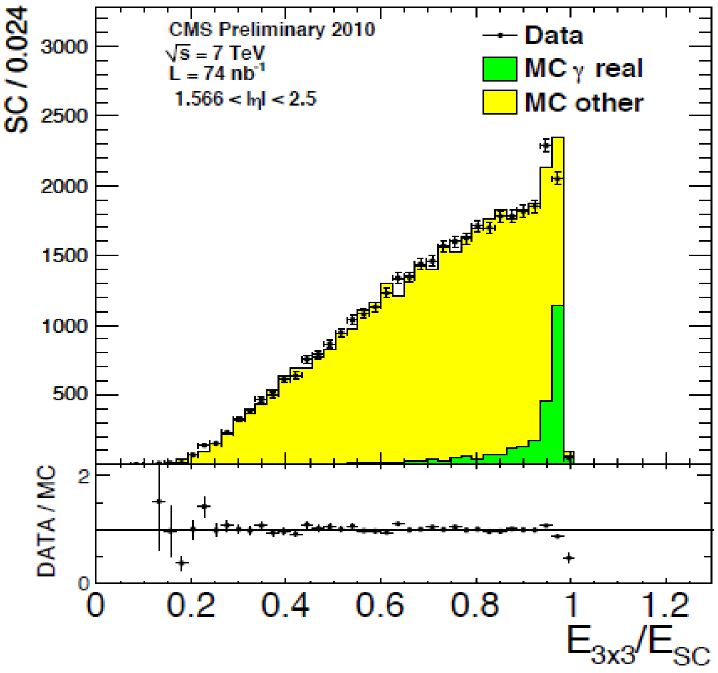


Exercise: R9 cluste shape variable

R9 = E3x3/Esupercluster

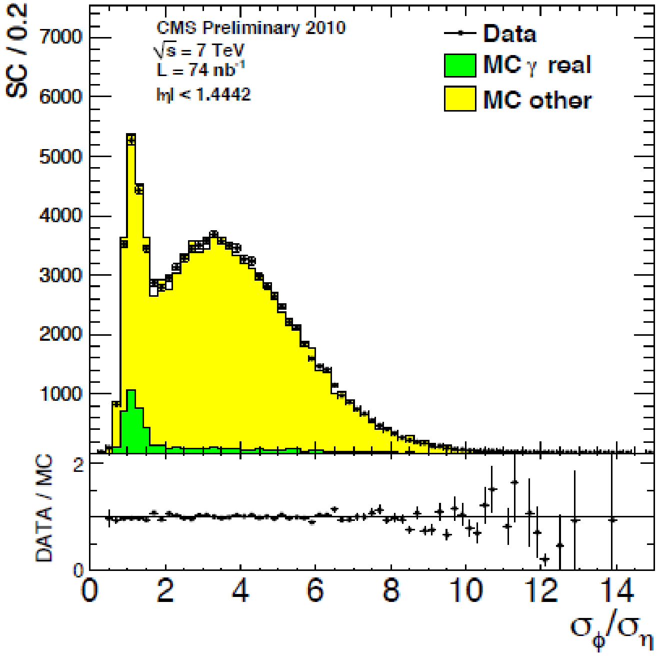
- High R9 : unconverted photon, very good energy resolution
- Low R9 : converted photon, poor energy resolution
- π0 also located at low R9
- R9 is very η-dependent

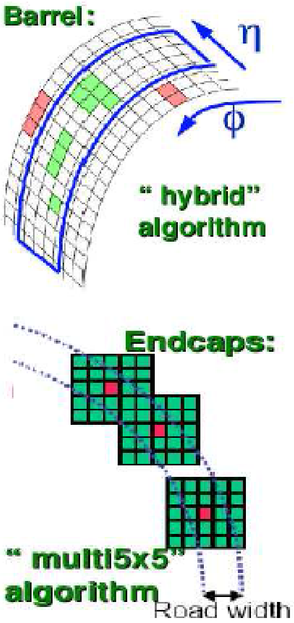


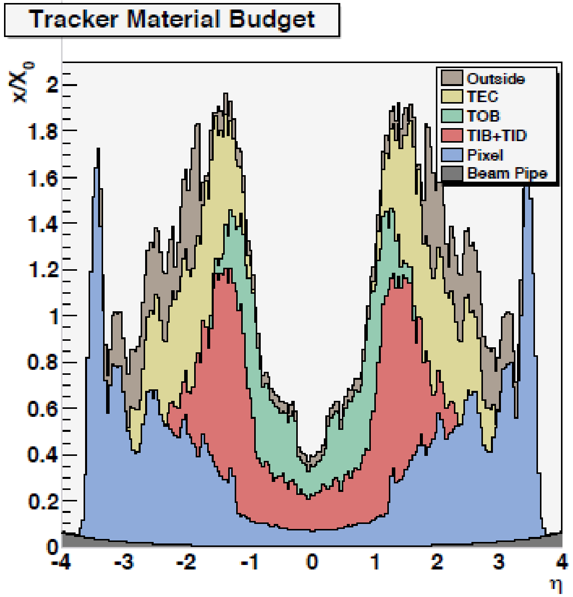


Exercise: brem cluter shape variable

- A photon energy deposit is broader in ϕ than in η, due to the magnetic field which bent the conversion trajectory around the z axis
- The η-width is broader for a π0 than a photon
- The clustering algorithm is affected by the material in front of the ECAL : strongly η-dependant
- The photon energy resolution is strongly dependent on σϕ/ση







Exercise: isolation energy

Isolation energy is defined in a ΔR cone of 0.3 or 0.4 around the photon

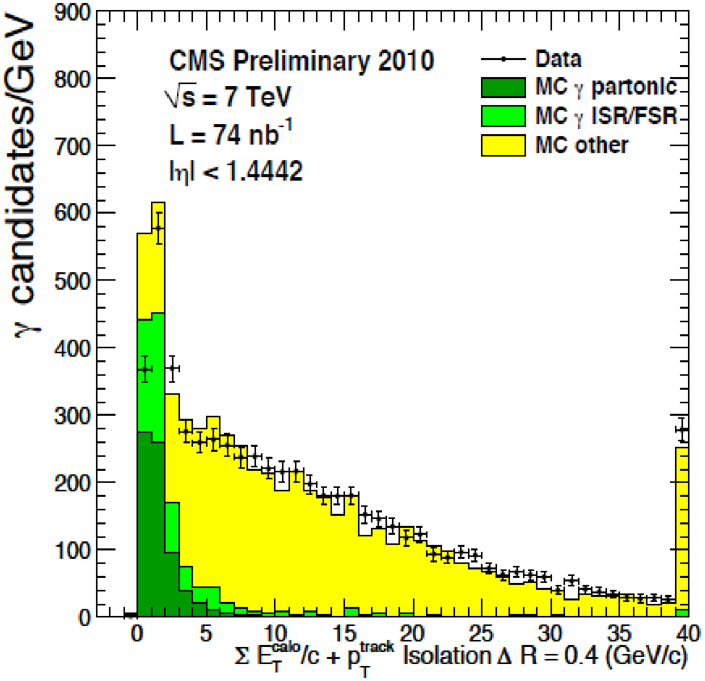
- Tracker isolation : Sum pT of the tracks reconstructed inside the cone
- ECAL, HCAL isolation : Sum Et of the energy deposits inside the cone

Isolation energy is coming from :

- Underlying event
- Pile-up
- QCD/QED radiation

- Prompt photons are isolated
- Neutral mesons within jets are less isolated

- In the exercise, 0.3 and 0.4 cones are used



Exercise: possible multi-variate methods

To improve the H→γγ analysis sensitivity, one can use several multi-variate methods :

Vertexing MVA

- Used in CMS results since Summer 2011
- In the exercises, no pile-up. The vertex is assumed to be correctly reconstructed.

Energy regression

- Used in CMS results since Dec 13
- Can be tried with the samples provided in the exercises

Photon identification with MVA

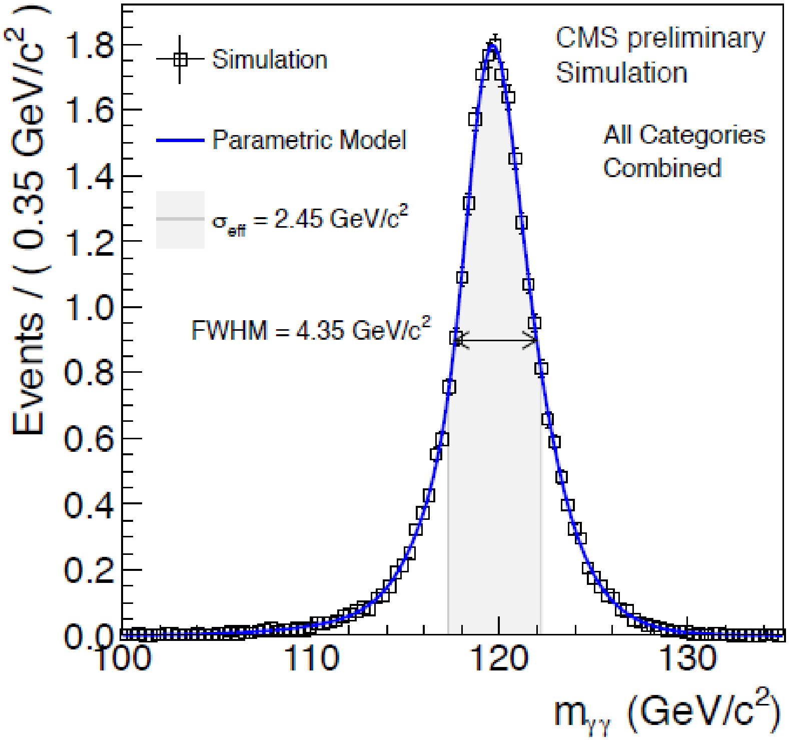
- Photon identification performed with rectangular cuts for the moment
- Can be tried in the exercises

Kinematics MVA

- Only the invariant mass is used for the moment - no MVA
- Can be tried in the exercises

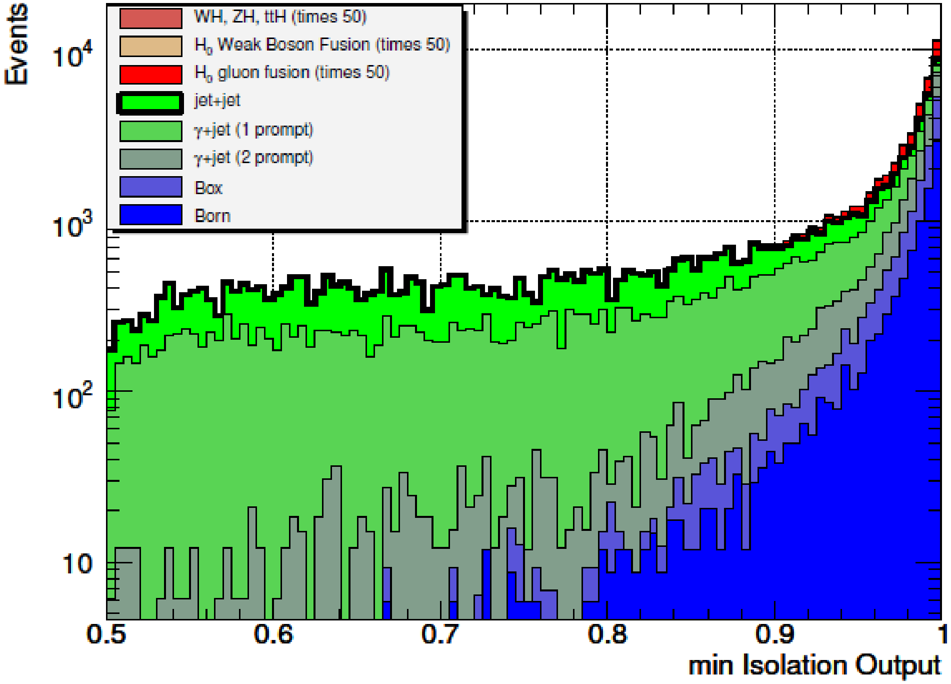
Exercise: energy regression

- Perform a regression from the reconstructed energy to the generated energy, using many geometrical variables and cluster shape variables
- This improved a lot the invariant mass resolution



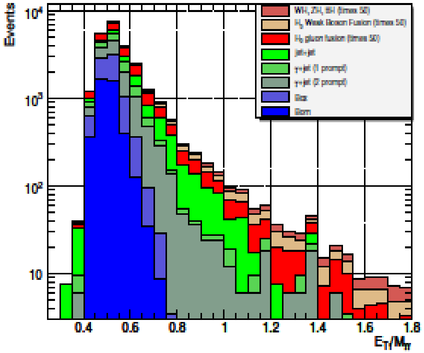
Exercise: photon identification

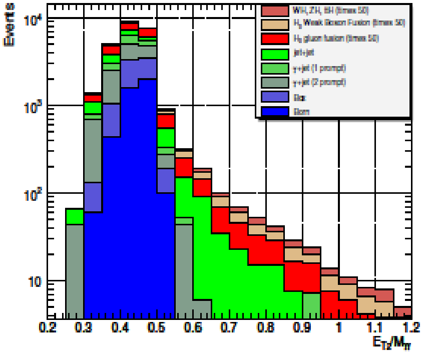
- In CMS Physics TDR vol. II, a photon identification NN was used :
- Uses ECAL, HCAL, Tracker isolation
- And R9 cluster shape variable

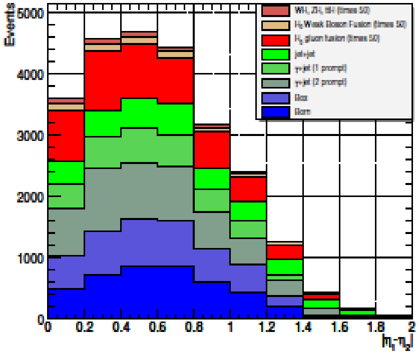


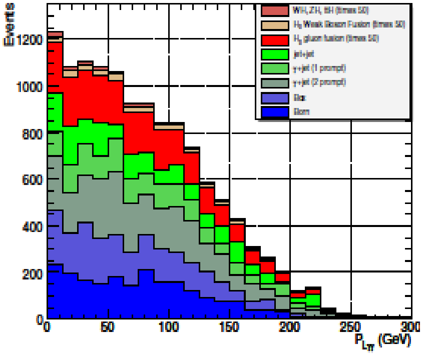
Exercise: kinematics MVA

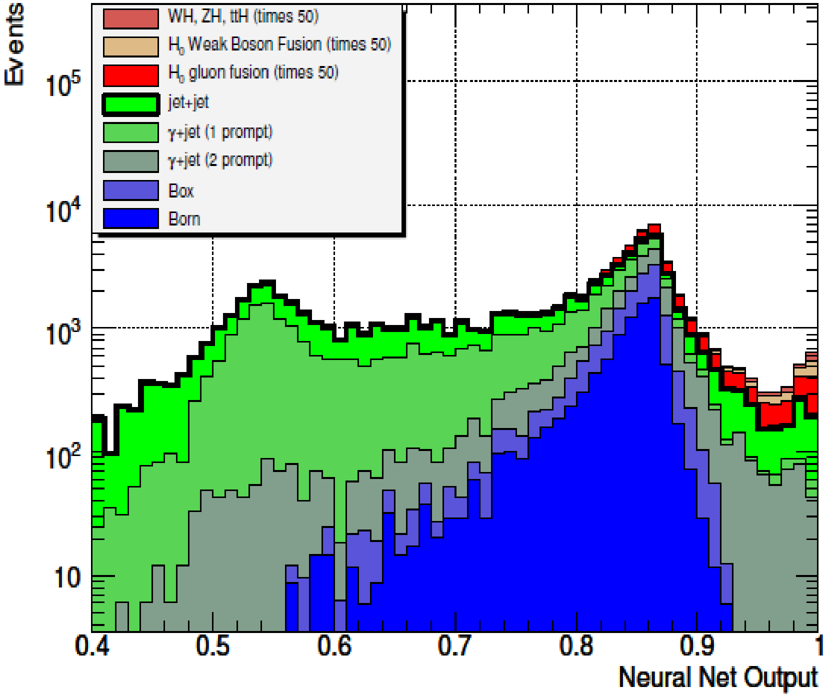
- In **CMS Physics TDR vol. II, a global NN** was used :
- ET/M of the two photons, pseudo-rapidity difference, pT of the diphoton system
- The two outputs of the NNisol


Events
10^4
10^3
10^2
$E_{T1}/M_{\gamma\gamma}$
WH, ZH, ttH (times 50)
H_0 Weak Boson Fusion (times 50)
H_0 gluon fusion (times 50)
jet+jet
γ+jet (1 prompt)
γ+jet (2 prompt)
Box
Born


Events
10^4
10^3
10^2
$E_{T2}/M_{\gamma\gamma}$
WH, ZH, ttH (times 50)
H_0 Weak Boson Fusion (times 50)
H_0 gluon fusion (times 50)
jet+jet
γ+jet (1 prompt)
γ+jet (2 prompt)
Box
Born


Events
5000
4000
3000
2000
1000
0
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2
$|\eta_1-\eta_2|$
WH, ZH, ttH (times 50)
H_0 Weak Boson Fusion (times 50)
H_0 gluon fusion (times 50)
jet+jet
γ+jet (1 prompt)
γ+jet (2 prompt)
Box
Born


Events
$P_{L\gamma\gamma}$ (GeV)
WH, ZH, ttH (times 50)
H_0 Weak Boson Fusion (times 50)
H_0 gluon fusion (times 50)
jet+jet
γ+jet (1 prompt)
γ+jet (2 prompt)
Box
Born


Events
10^5
10^4
10^3
10^2
Neural Net Output
WH, ZH, ttH (times 50)
H_0 Weak Boson Fusion (times 50)
H_0 gluon fusion (times 50)
jet+jet
γ+jet (1 prompt)
γ+jet (2 prompt)
Box
Born