# Data-Driven Background Estimates at the LHC

Stéphanie Beauceron
GDR Terascale
4th November 2010, Brussel

On Behalf of ATLAS and CMS Collaboration

Vrije Universiteit Brussel

iihe BRUXELLES BRUSSEL

# Outline

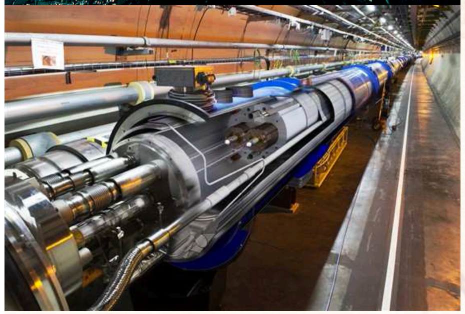
LHC is starting the searches

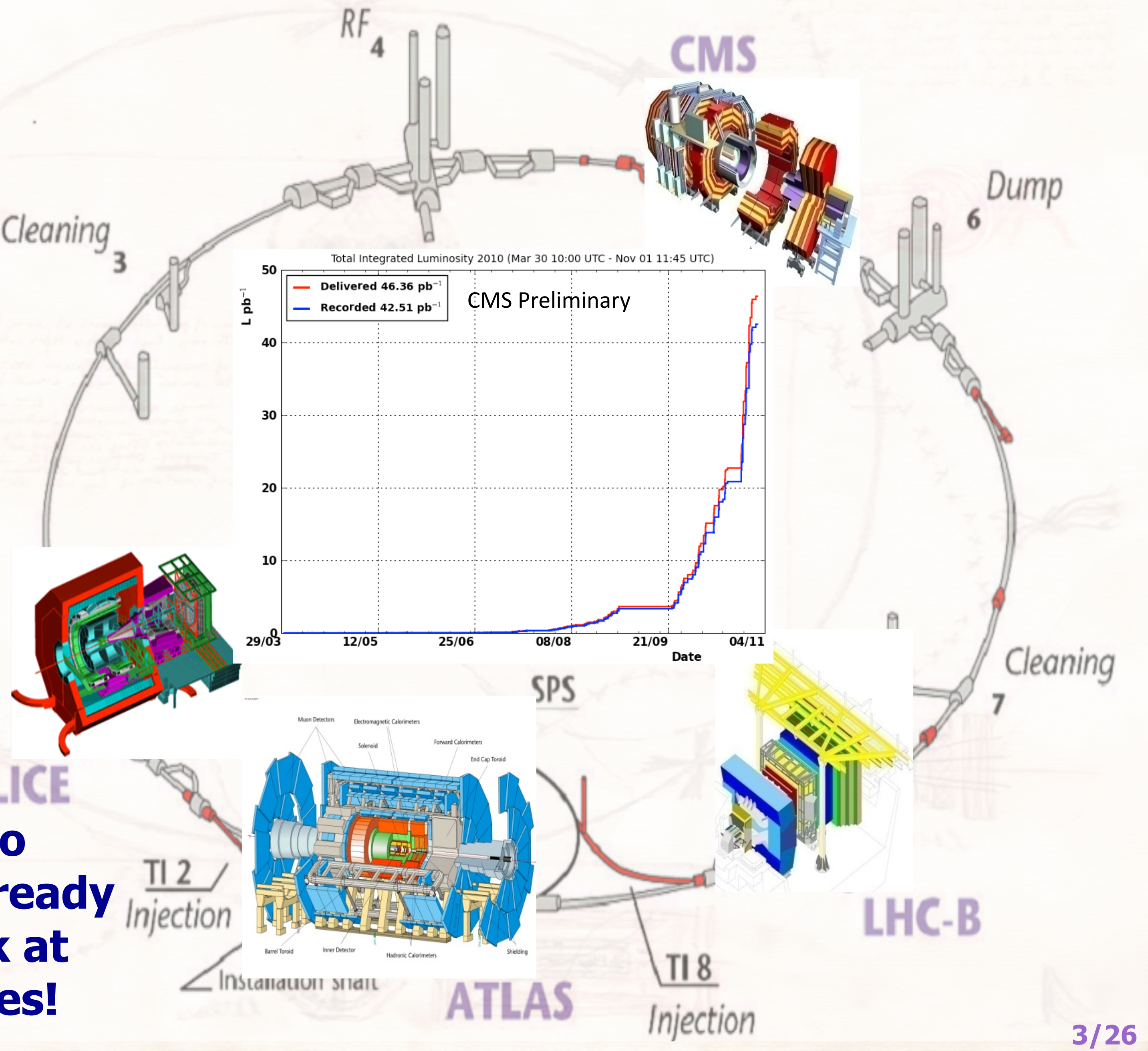
Main methods on the market:

- Fit
- Scaling
- Templates
- Replacement Method
- Matrix Methods

Conclusion

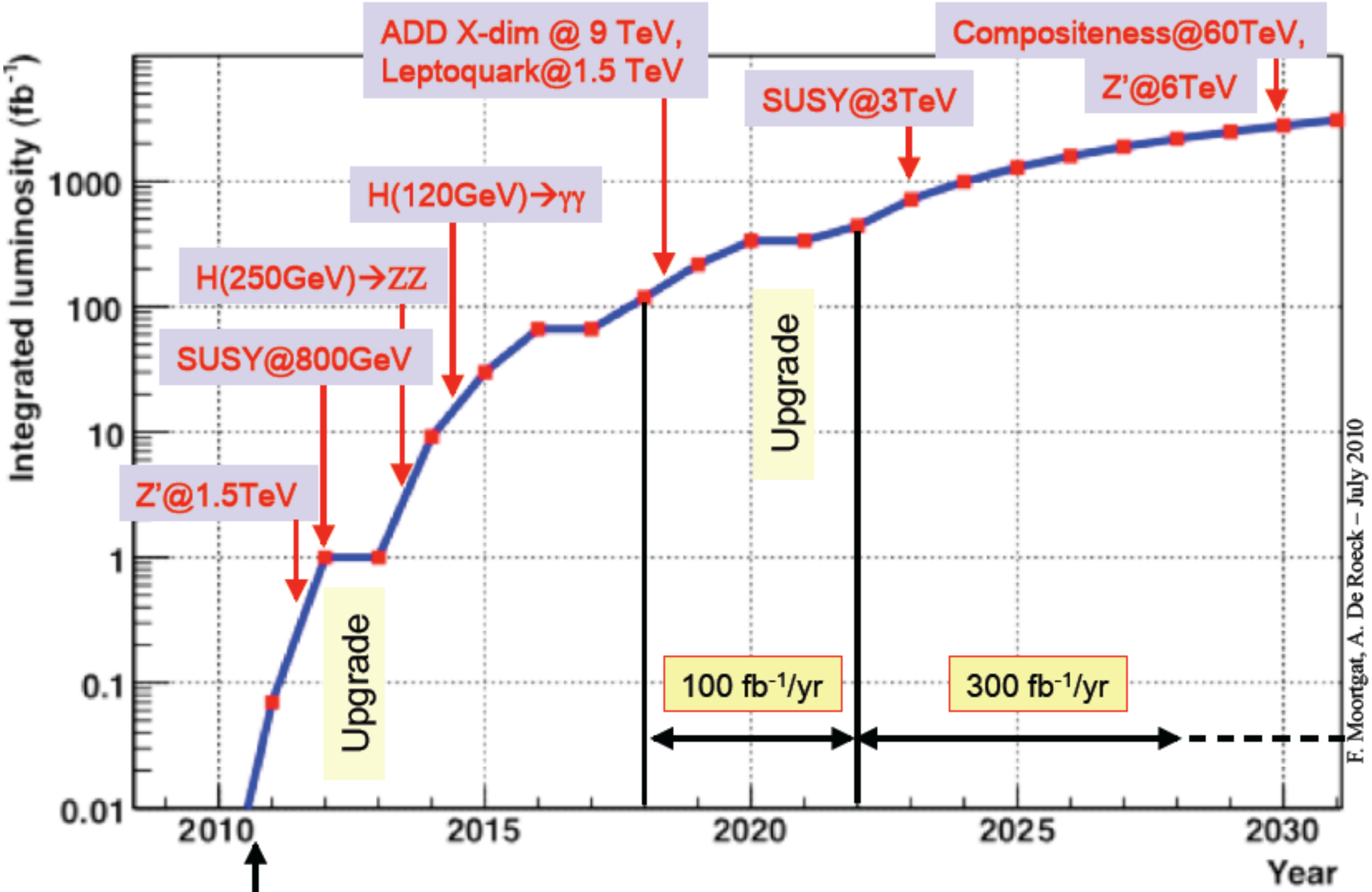
# LHC Opens Window to Searches






**More than**
**40 pb-1 delivered to**
**ATLAS and CMS, already**
**a lot of data to look at**
**and to start searches!**

# Road Map to Discoveries

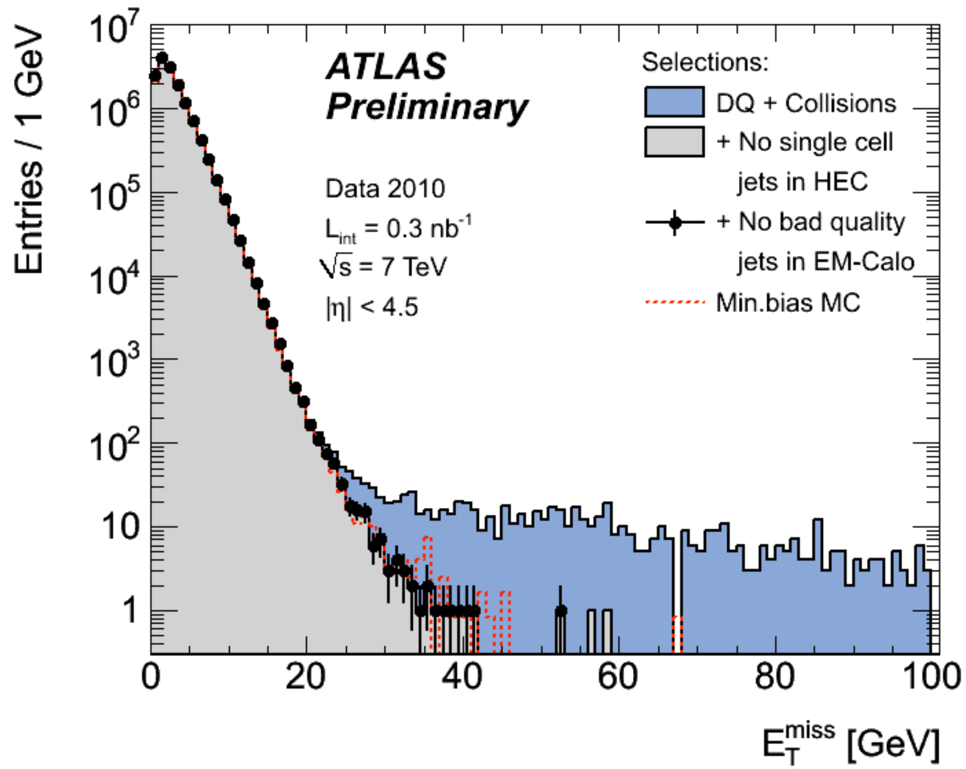



First physics run in 2010-2011: 1 $fb^{-1}$ @ 7 TeV

# Commissioning Detectors: Understanding the Variables…

**From the first collisions day, a lot of results have been appearing very quickly**

**→ Understanding and commissioning of the detector is in well advanced stage**
**→ Mandatory before exploring new territories…**




**Standard Model signals are becoming background of searches, need to have a proper evaluation of their contamination in signal area (too large to number of events to be simulated).**

# Which Methods to Use…

**Depending of the signal studied, different kind of background:**

- **Resonance like signal:**

→ **Propagation Fit and subtract background from the fit**

→ **Factorization cuts**

- **Looking in tail of distributions (on top of previous):**

→ **Templates**

→ **Replacement Method**

→ **Various Matrix Method**

→ **Various techniques can be used for cross check, some time mandatory to do them in sequence**

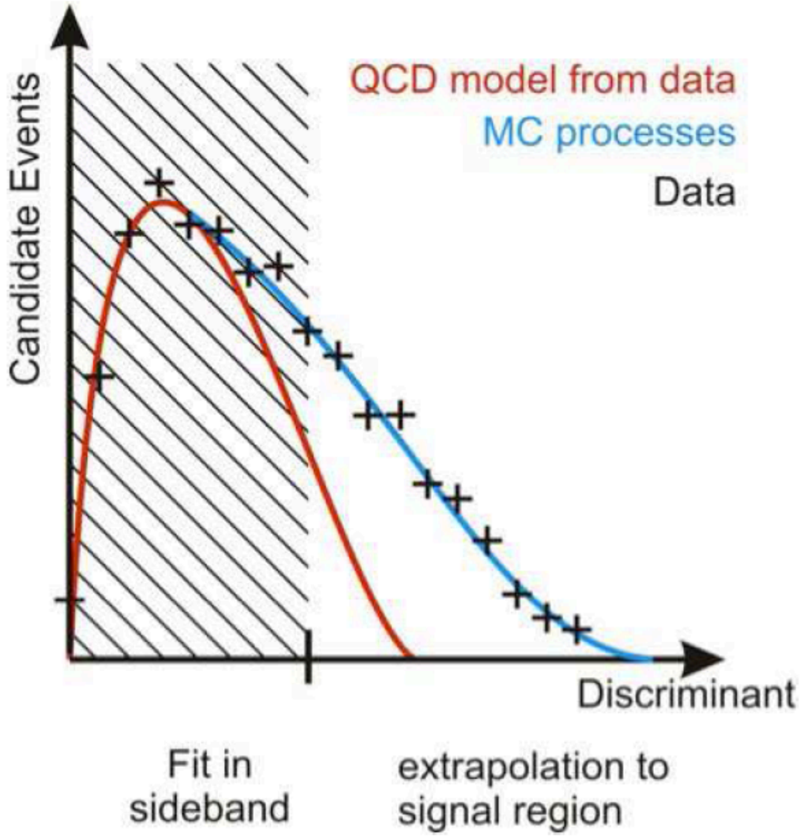
# Fit Propagation

**Find a control region in phase space where SM background dominates.**

**Use measurements in this region to infer SM background in signal region.**

**Should ensure the fit function is valid in the signal area.**

**Ex: Searches with isolated leptons to determine contamination from non isolated leptons.**

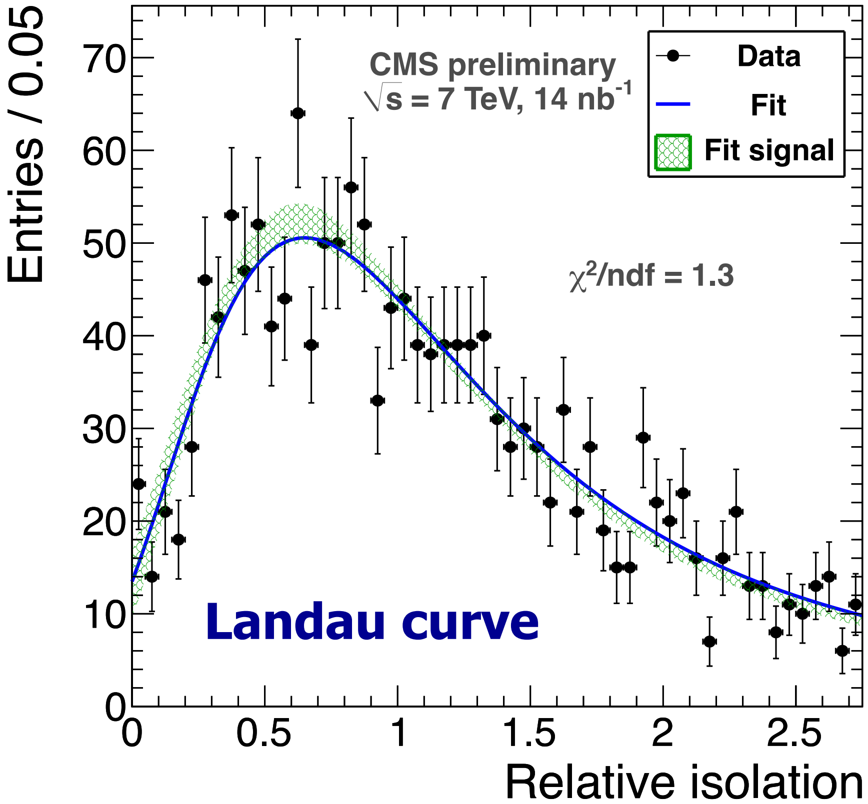


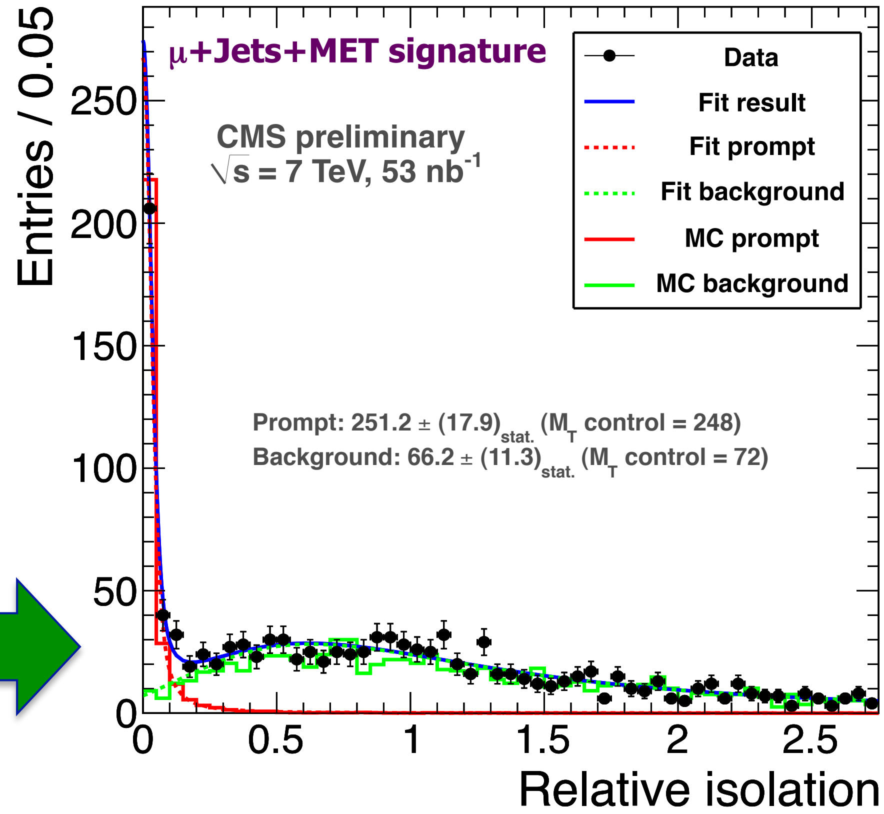

**Variation: Fit of multiple contributions**

# $\mu$+Jets+ME$_T$ Signature

**Looking at samples after full selection except isolation.**
**Determine the shape of the function to fit in a background like sample.**

CMS PAS SUS-10-001




**Fit of signal can also be done using simulation.**
**➔ Good agreement between fit estimation, data and simulation control**

# e+Jets+$ME_T$ Signature

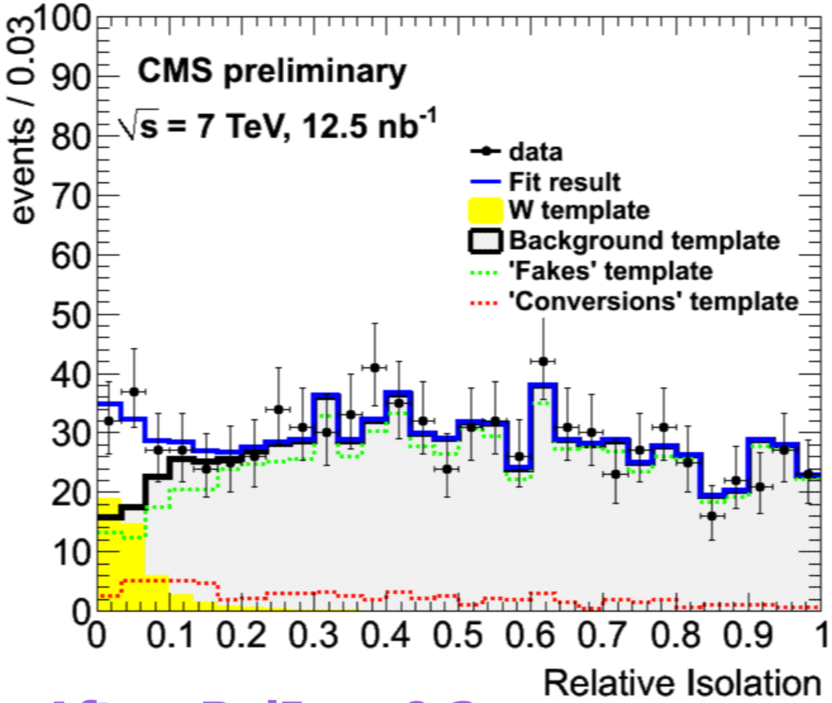
**Two kinds of background:**

- **heavy-flavor decays and jets mis-identified as electron**
- **electrons due to photon conversion**

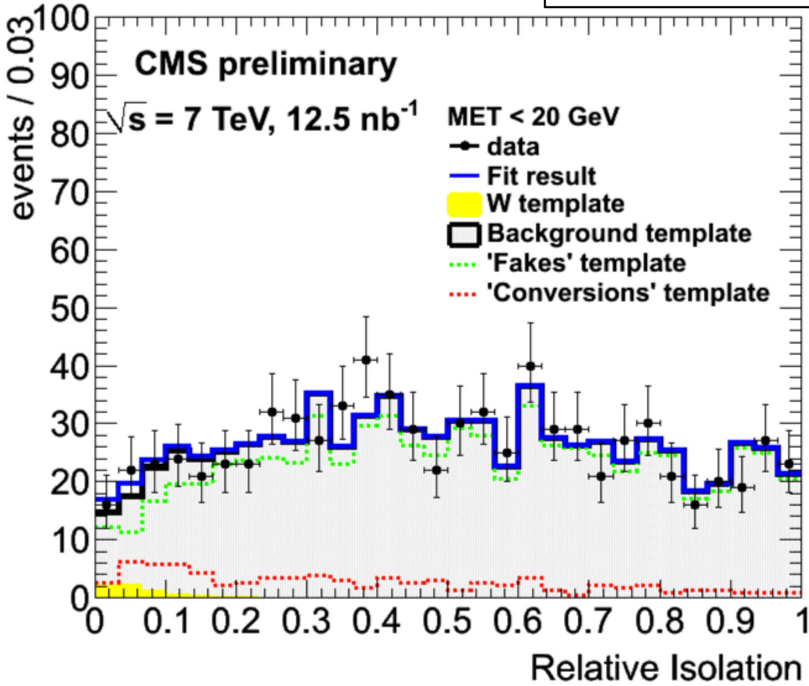
**Select control samples dominated by each of above sources by inverting selection cuts**

**Perform fit using Relative Isolation (RelIso = $p_T(e)/\Sigma E_{T\ R<0.3}$) distributions for each background.**

CMS PAS SUS-10-001




**After: RelIso<0.3**
**Predicted : 224 ± 13**
**Observed : 263**




**After (RelIso<0.3)**
**Predicted : 215 ± 13**
**Observed : 215**

# Factorization Cuts/Scaling

**Determine all efficiencies of the cuts selection and weight a background like sample by all efficiencies.**

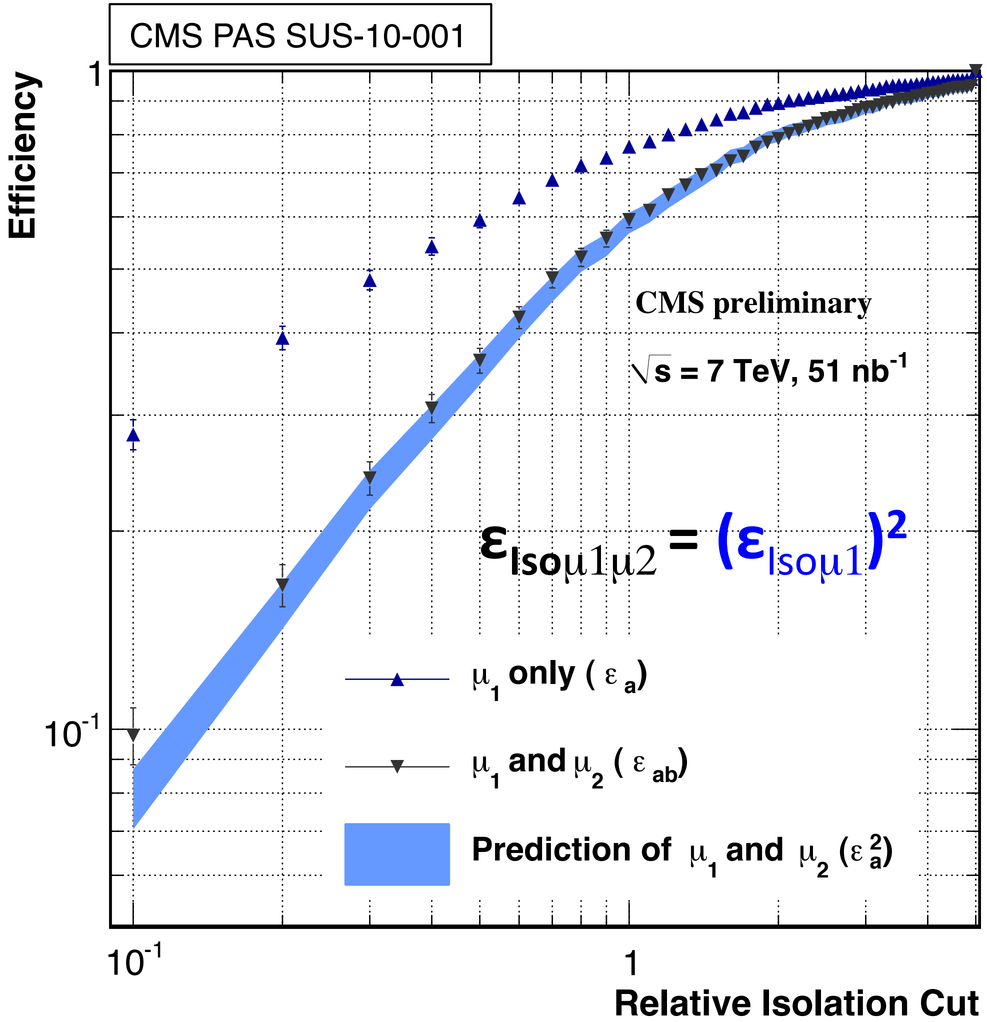
**Mainly to ensure that a given SM background can be neglected in the final selection, or using higher statistics sample:**

- **Berends-Giele scaling method:** $\mathrm{W}^{\geq 4\mathrm{jets}} = \mathrm{W}^{2\mathrm{jets}} \cdot \sum_{i=2}^{\infty} (\mathrm{Z}^{2\mathrm{jets}}/\mathrm{Z}^{1\mathrm{jet}})^i$
- **Scaling distribution according to resolution etc**

**Need to control the correlation between cuts and/or ensure that selection do not bais scaling.**

# Same Sign di-Muons Searches

**Selection cuts are uncorrelated**
**➔ selection efficiency for each cut measured in control samples**




**Di-Muons samples before isolation (dominated by multijet events)**

**Isolation of $\mu_1$ = $\varepsilon_{Iso\mu1}$**

**Isolation of $\mu_2$ = $\varepsilon_{Iso\mu2}$**

$\varepsilon_{AllCuts} = \varepsilon_{Iso\mu1} \cdot \varepsilon_{Iso\mu2}$

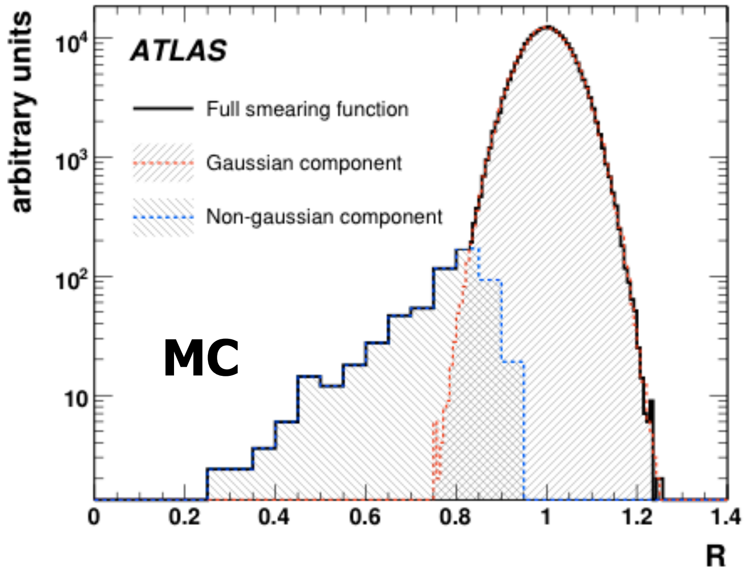
**Good agreement between prediction and observed ➔ multijet background can be scaled down by $(\varepsilon_{Iso\mu1})^2$**

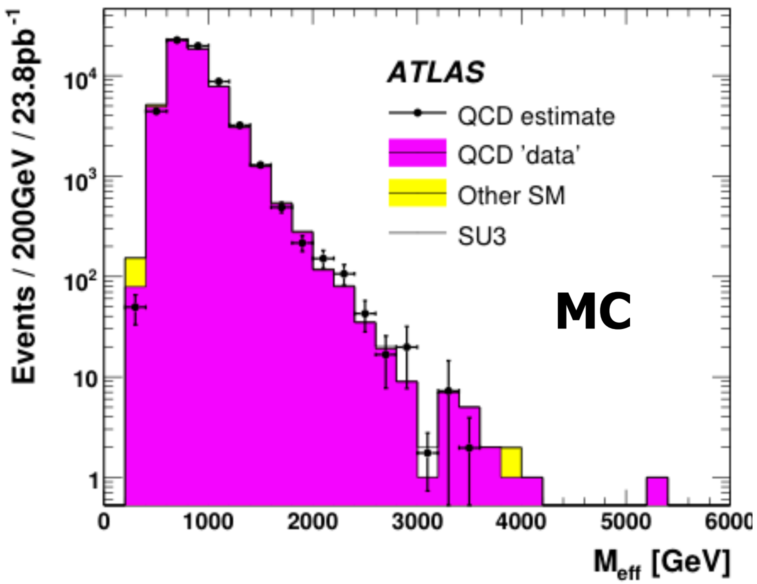
# Smearing

arXiv:0901.0512 (2009)

**Modify Monte Carlo samples to mimic the data:**
**Mostly used for QCD events to introduce Jet Resolution and its effect on missing ET.**

- **Derive Gaussian part of smearing function from γ + jet control sample**
- **Derive non-Gaussian part from Mercedes events (⅄), requiring that the MET is co-linear with one of the jets**
- **Combine smearing functions, normalising with di-jet sample**
- **Apply smearing function to low MET events to predict the tail in the high MET signal region.**

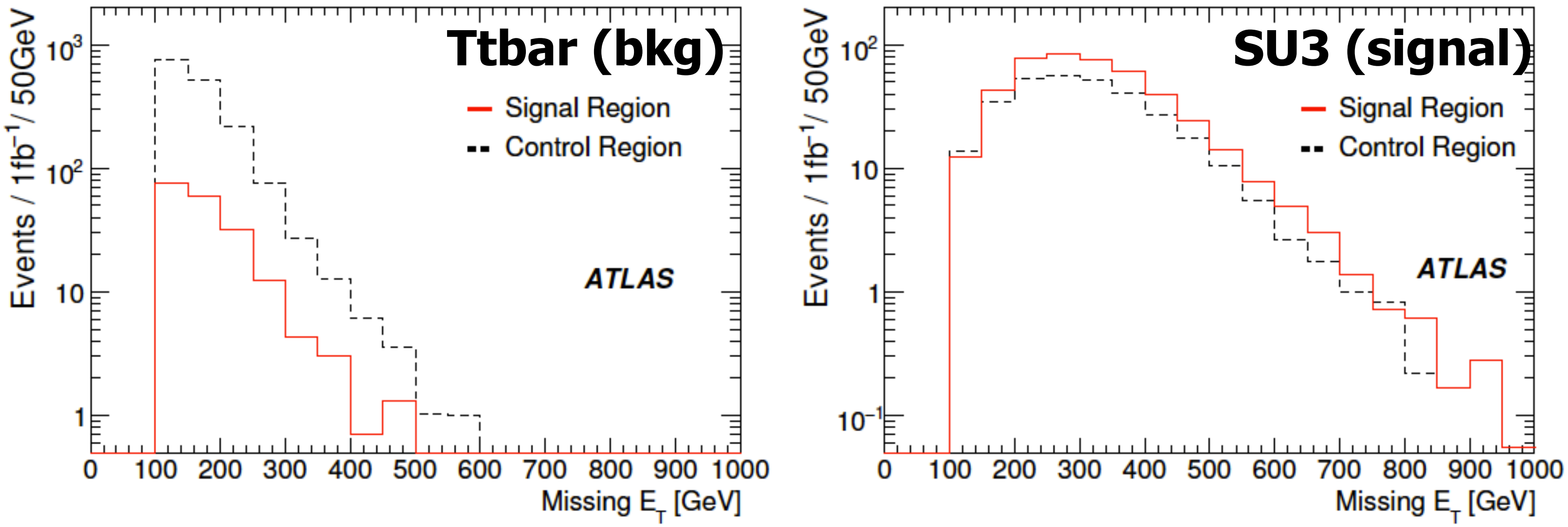


# Templates

**Define a signal-depleted control sample**
**Determine the shape of background in this region**
**Propagate the shape of the background in a signal like region.**
**Need to understand the variables shape in control region to port it in signal region.**

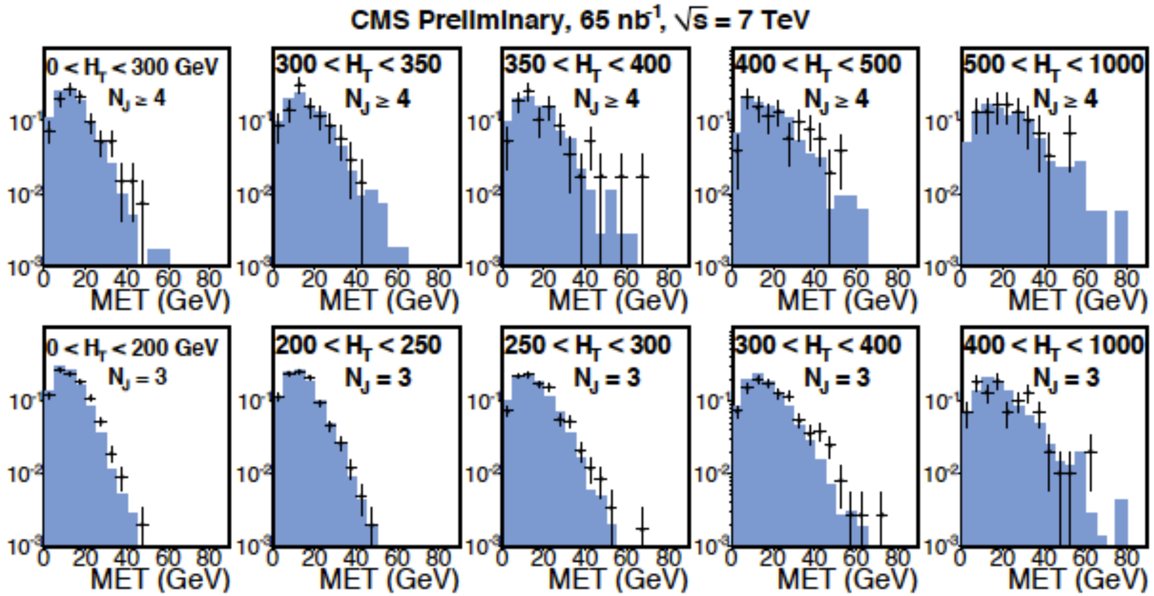
CERN-OPEN-2008-20



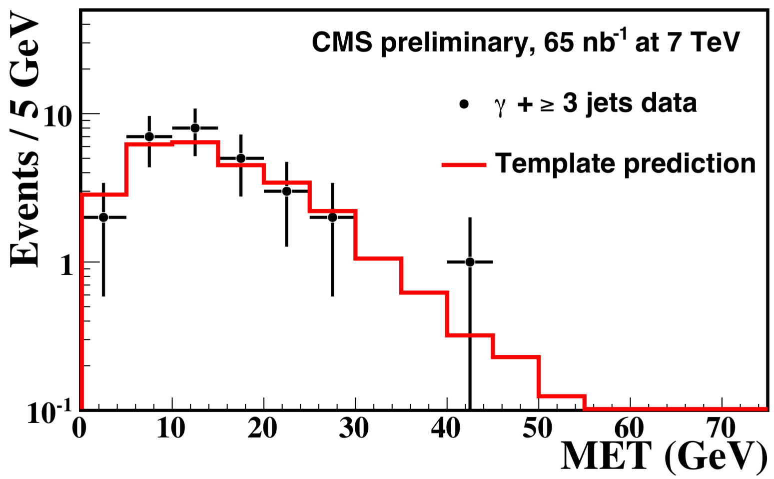
# Lepton+jets+MET Signatures

CMS PAS SUS-10-001

- **MET background from real MET (e.g. in W/Z) and MET due to mis-measurements**
- **Use MET templates from multi-jet events to predict MET for g +jets events**




MET templates from multi-jet events




**Good agreement between predicted and observed distributions:**
**for MET > 15 GeV**
**predicted = 12.5**
**observed = 11**

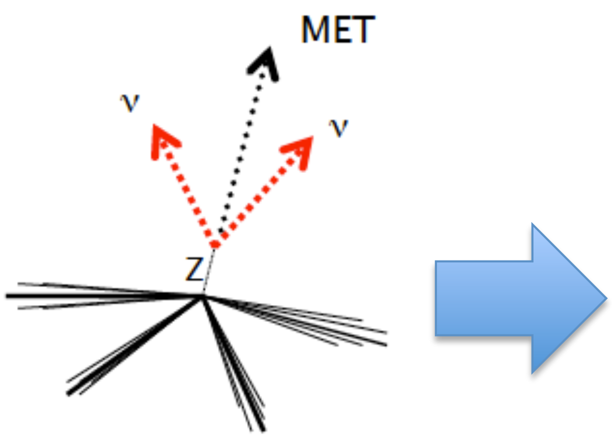
# Replacement Method

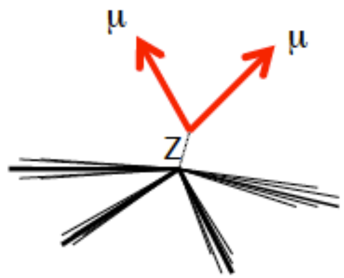
**Use a none standard model process identified from data and "modify" it in order to simulate another standard model process.**

**Example:**

**Large missing $E_T$ searches + jets:**

**Z +jets → νν + jets ➔ irreducible background**

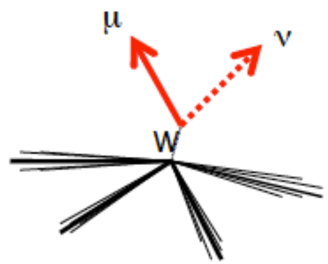



**Z → ll + jets**

Strength: very clean

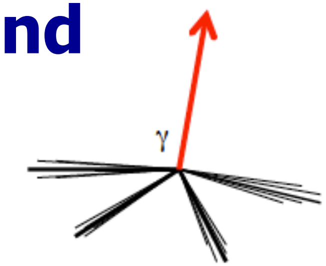
Weakness: low statistics




**W → lν + jets**

Strength: larger statistics

Weakness: background from SM and SUSY




**γ + jets**

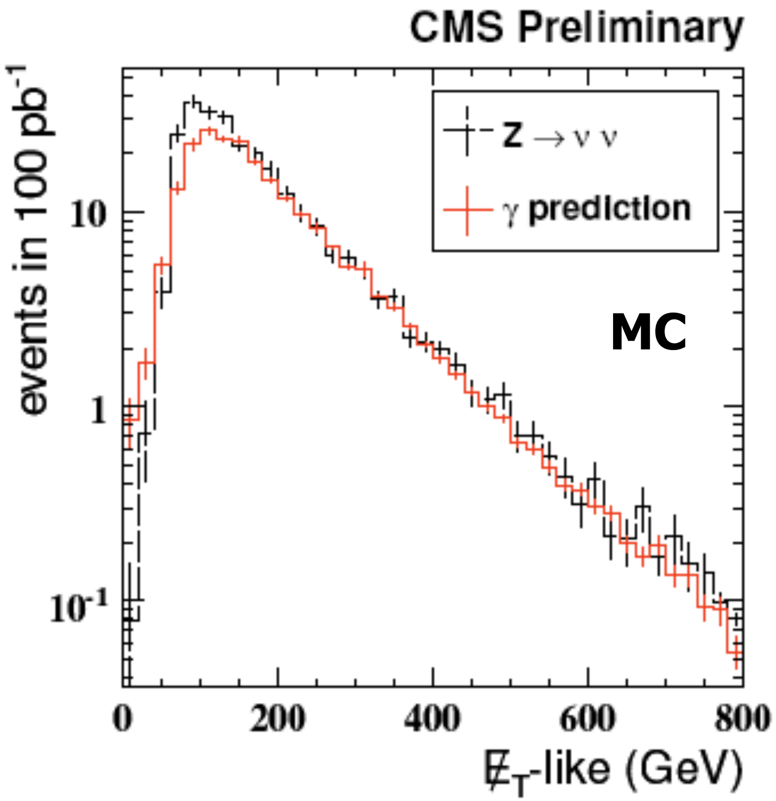
Strength: large statistics and clean at high $E_T$

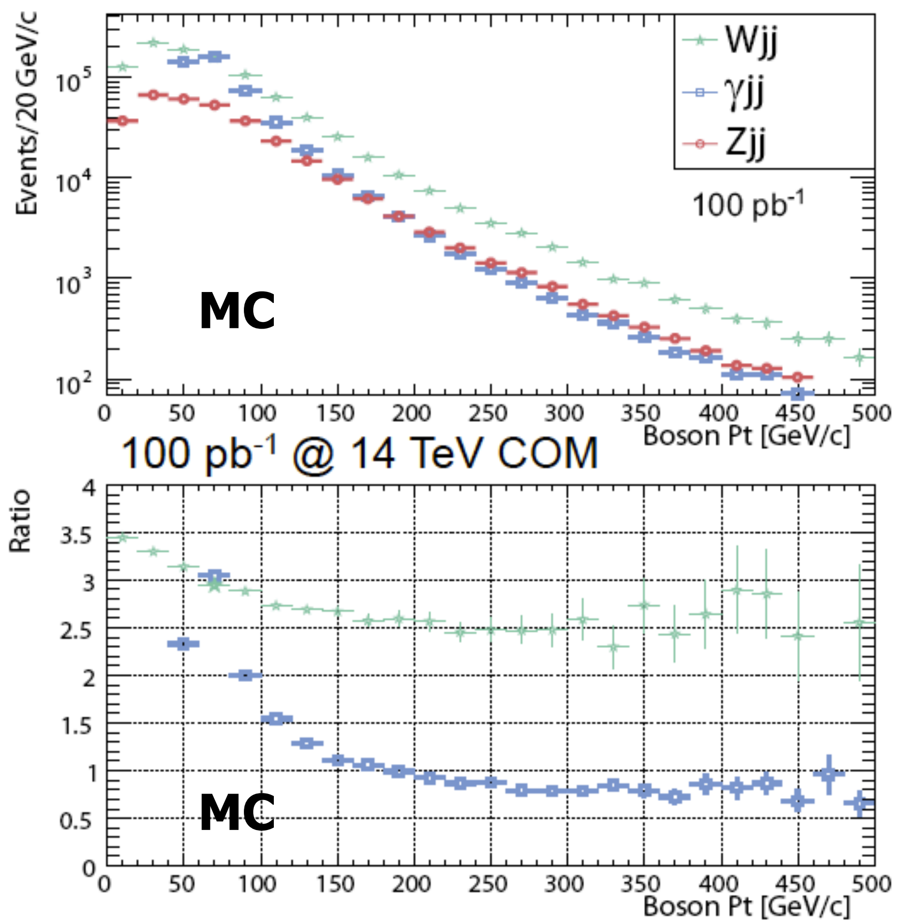
Weakness: background at low $E_T$, theoretical errors

# Z+jets→νν +jets

CMS-PAS-SUS-08-002

**Select γ + ≥3 jets with E(γ)>150 GeV**
**Remove photon from the event**
**Recalculate MET**
**Normalise with σ(Z+jets)/σ(γ+jets) from MC or measurements**




**Good agreement between prediction and estimation.**

# Matrix Method "à la DØ"

## (or Tight/Loose Ratio)

**An initial sample containing $N_{loose}$ events**
**→Applying an additional cut to reach a second sample containing $N_{tight}$ events which is a subset of the initial sample**
**Each sample contains a given number of signal ($N_{real}$) like and background ($N_{fake}$) like. Fraction are changing as follow:**

Cut

$$\begin{aligned} N^{\text{loose}} &= N^{\text{loose}}_{\text{real}} + N^{\text{loose}}_{\text{fake}} \\ N^{\text{tight}} &= \epsilon_{\text{real}} N^{\text{loose}}_{\text{real}} + \epsilon_{\text{fake}} N^{\text{loose}}_{\text{fake}} \end{aligned}$$

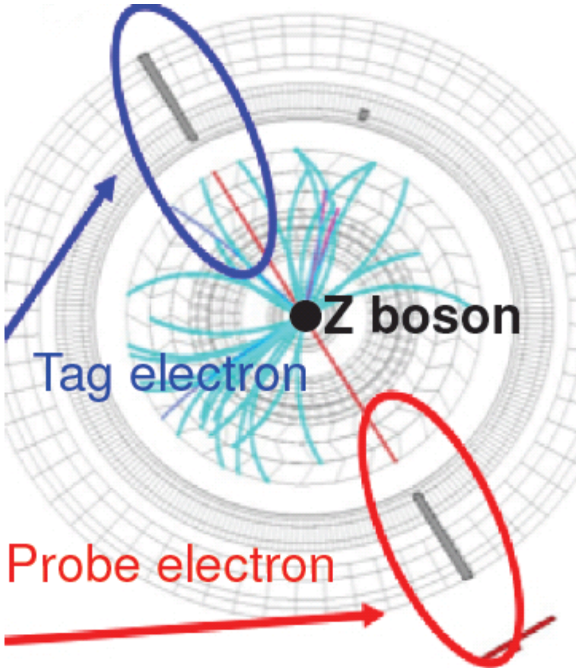
**Challenge: calculating $\varepsilon_{real}$ and $\varepsilon_{fake}$**

**Mainly used to determine multi jets background in analysis selecting on leptons.**

# Determining the parameters

**When using leptons, use Tag and Probe to compute $\varepsilon_{tight}$:**

- **require a the $l^+l^-$ pair to be within a $m_Z$ window**
- **high lepton purity can be reached with tight ID cuts on the "tag" and the $m_Z$ window**

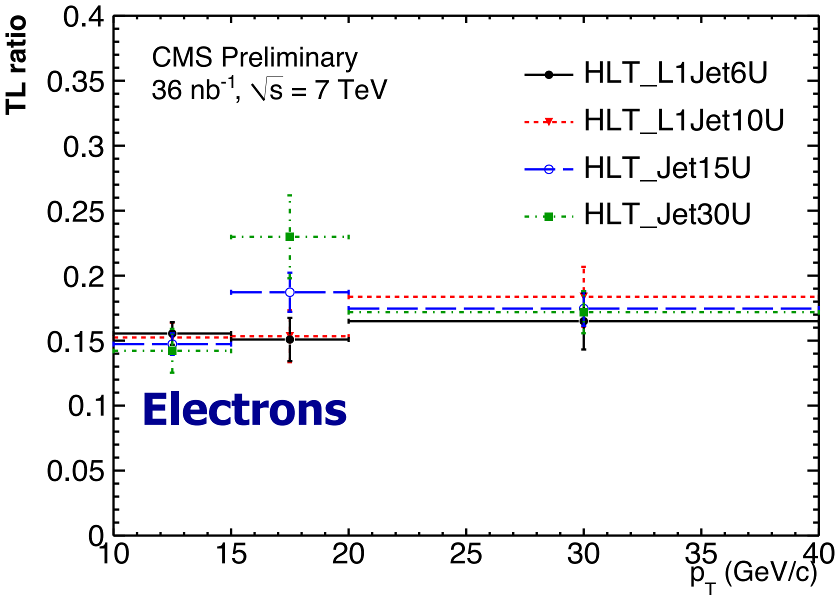


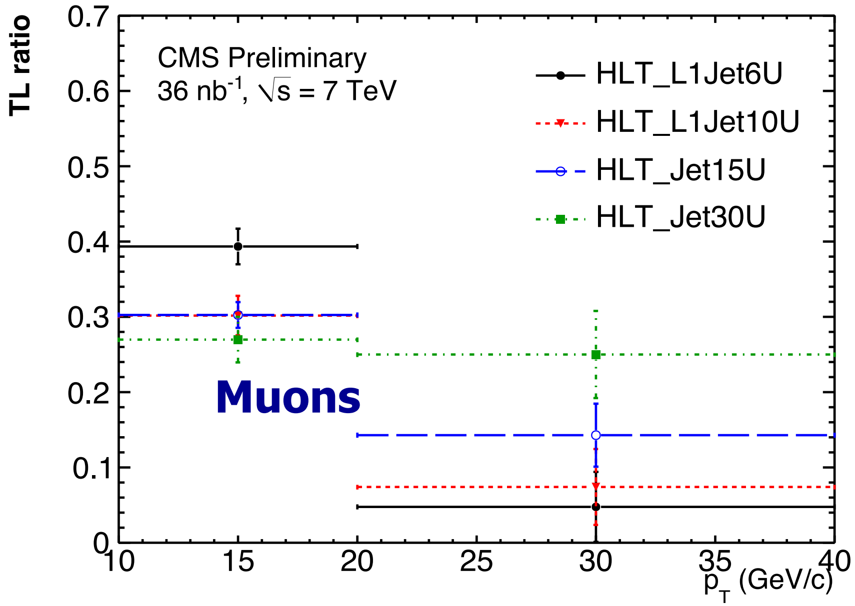

**For $\varepsilon_{fake}$, look for background dominated samples (jets dominated samples, lepton-jets back to back or W+jets with W in the other lepton flavor)**

# Same Sign Searches

**Use a jets dominated control sample (loose lepton-id & isolation) to measure $\varepsilon_{fake}$ (= "TL ratio" ) as function of kinematics variables**

**Tight-to-Loose-Ratios using different jet-triggered samples**




| Channel | Predicted | Observed |
|---|---|---|
| $ee$ | $0.43^{+0.18}_{-0.14}$ | 0 |
| $e\mu$ | $0.14^{+0.18}_{-0.09}$ | 1 |
| $\mu\mu$ | $0.22^{+0.51}_{-0.18}$ | 0 |

**Consistency in predicted & observed number of events.**

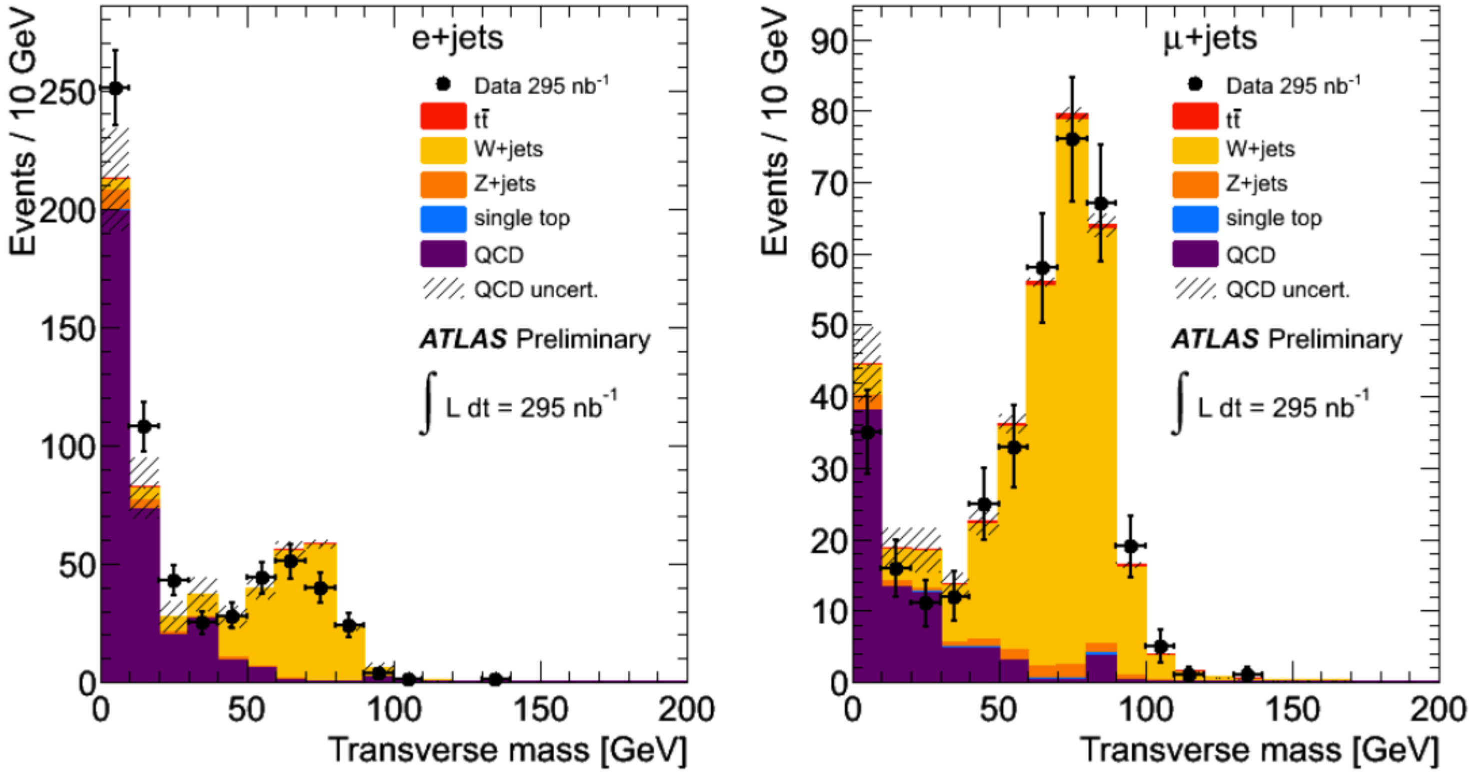
CMS PAS SUS-10-001

# Top Rediscovery

**In sample for $\varepsilon_{fake}$, contamination of signal can appear. Equation of $N_{loose}$ and $N_{tight}$ can be rewritten and by iteration, bias on $\varepsilon_{fake}$ can be removed.**

ATLAS-CONF-2010-087

**QCD is estimated by this method in each of the bin of the distribution for semi-leptonic top searches.**




**Fair agreement between data and the sum of MC samples and multijets estimation.**

# Extension to Di-Leptons

**The system of equation can be written for Di-Lepton final states searches:**

$$\begin{bmatrix} N_{TT} \\ N_{TL} \\ N_{LT} \\ N_{LL} \end{bmatrix} = \begin{bmatrix} rr & rf & fr & ff \\ r(1-r) & r(1-f) & f(1-r) & f(1-f) \\ (1-r)r & (1-r)f & (1-f)r & (1-f)f \\ (1-r)(1-r) & (1-r)(1-f) & (1-f)(1-r) & (1-f)(1-f) \end{bmatrix} \begin{bmatrix} N_{RR} \\ N_{RF} \\ N_{FR} \\ N_{FF} \end{bmatrix}$$

**With:**

**$f=\varepsilon_{fake}$** **$N_{TT}$ = Number of events in Tight-Tight**

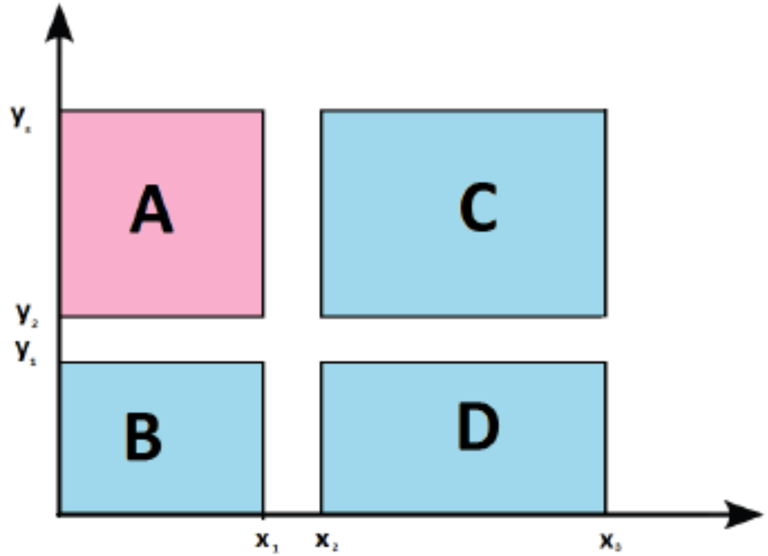
**$r=\varepsilon_{real}$** **$N_{LL}$ = Number of events in Loose-Loose**

**By solving the equation, each sample composition ($N_{RR}$ = Number of events containing two real leptons) can be found.**

# Matrix Method "à la CDF"

## (ABCD method/$M_T$/Tiles)

**Simplified version of the matrix method "à la DØ".**
**Splitting a 2D phase space by 2 criteria to obtain a signal like area and background like area:**

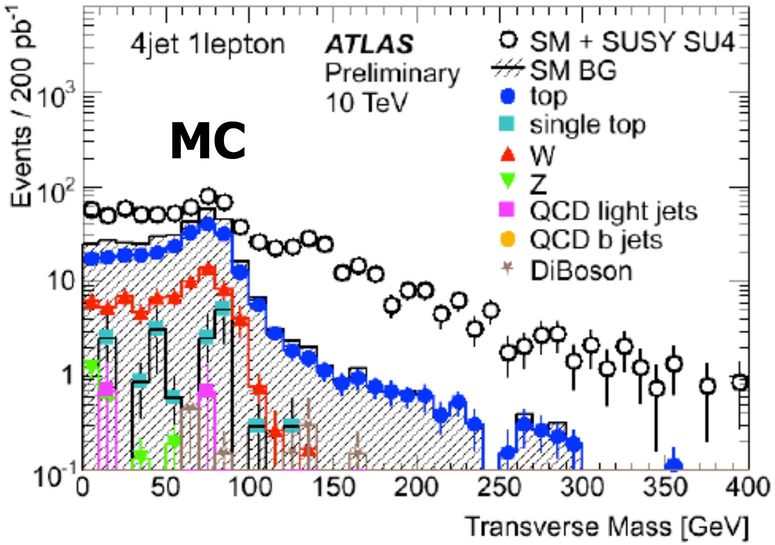


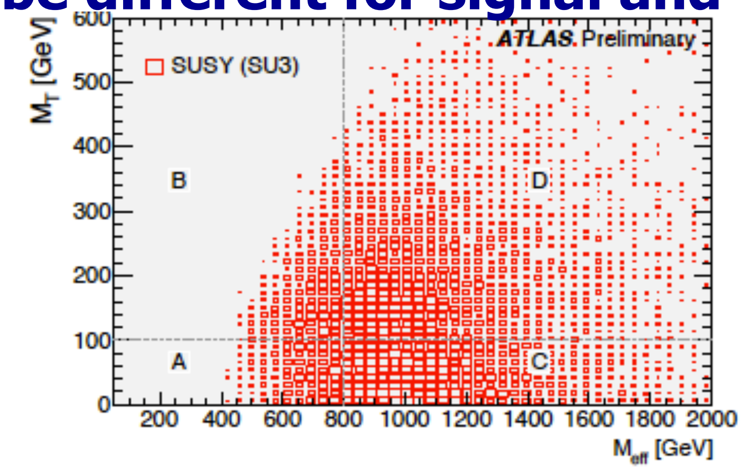

**Hypothesis:**
**- Neglecting signal contribution in regions B and D**
**- X variables has no effect on studied background**
**- Assuming that variables x and y are uncorrelated**
**➔ Number of background events in signal region A can be evaluated as $N_A = N_B \times N_C / N_D$.**
**Main issue: find uncorrelated variables**

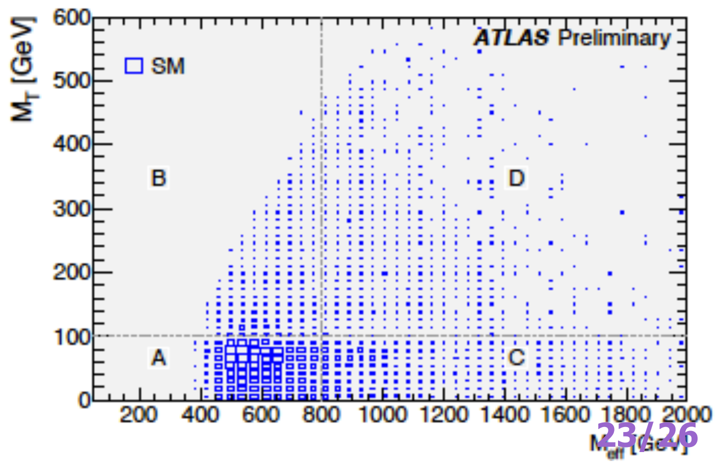
# Tiles Method

ATL-PHYS-PUB-2009-077

**Variation of Matrix Method "à la CDF":**
**Use $M_T$ and $M_{eff}$($=\Sigma\ E_T$ of ALL objects) as the two variables**
**($M_T$ > 100 GeV, W decay is background).**
**Each quadrant is named tiles.**
**Hypothesis:**
**-Relative inclusive fractions of SM background events in each tile are predicted by MC simulation.**
**- Discriminating variables are mutually independent for signal events.**
**- In presence of signal, the distributions of events among the tiles need to be different for signal and background.**



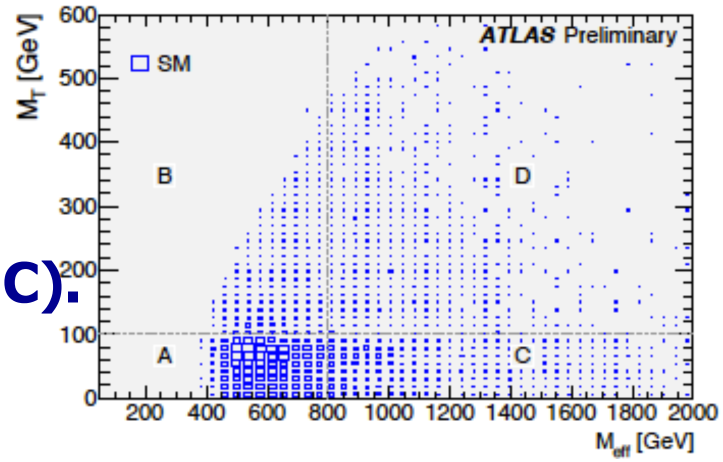



# 2x2 Tiles Method

**In each of the tile:**

$$\bar{N}_A = f_A^{SM}\bar{N}^{SM} + f_A^S\bar{N}^S, \qquad \bar{N}_B = f_B^{SM}\bar{N}^{SM} + f_B^S\bar{N}^S$$

$$\bar{N}_C = f_C^{SM}\bar{N}^{SM} + f_C^S\bar{N}^S, \qquad \bar{N}_D = f_D^{SM}\bar{N}^{SM} + f_D^S\bar{N}^S$$




**Where the f represents respectively the fraction of SM/Signal in a given tile (from MC). Requiring further that the signal variables be independent:**

$$f_A^S = (1 - f_{M_{eff}}^S)(1 - f_{M_T}^S), \qquad f_B^S = (1 - f_{M_{eff}}^S) f_{M_T}^S,$$

$$f_C^S = f_{M_{eff}}^S(1 - f_{M_T}^S), \qquad f_D^S = f_{M_{eff}}^S f_{M_T}^S,$$

**➔ System can be solved:**

$$N^{SM} = \frac{1}{2(f_A f_D - f_B f_C)} \Big\{ f_D N_A - f_C N_B - f_B N_C + f_A N_D - \Big[ \Big( -(f_C N_B) - f_D (N_A + 2N_B) + f_B N_C + f_A N_D + 2 f_B N_D \Big)^2 - 4(f_D N_B - f_B N_D)\Big( (f_C + f_D)(N_A + N_B) - (f_A + f_B)(N_C + N_D) \Big) \Big]^{1/2} \Big\}$$

**➔ And signal:** $N^S = N_A + N_B + N_C + N_D - N^{SM}$

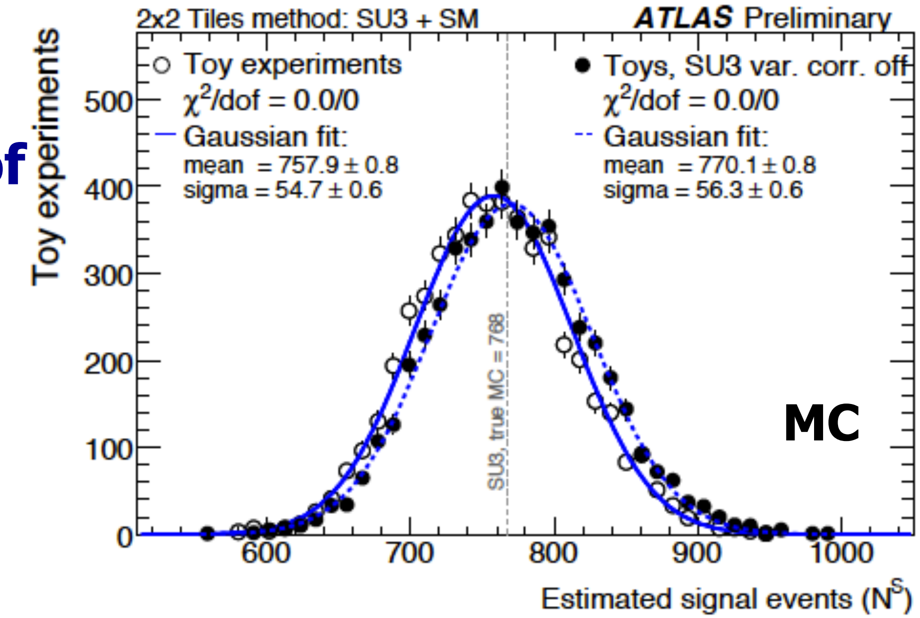
# NxN Tiles Method

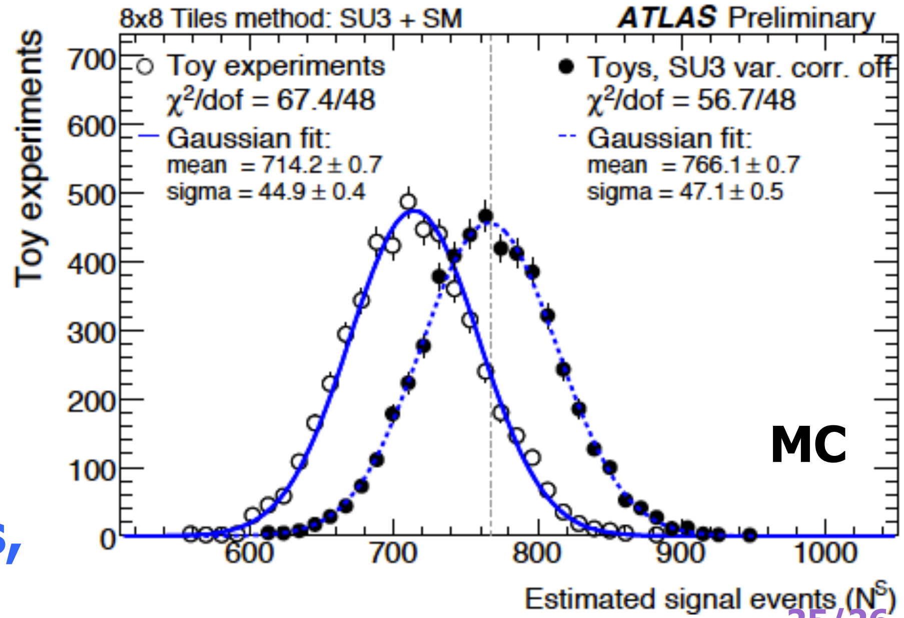
ATL-PHYS-PUB-2009-077

**Split the phase space in N tiles, $N^2$ equations can be written.**
**Ignoring signal correlation in each of the tiles, the problems is over constraint**
**➔ Define extended negative log-likelihood:** $$-\ln\mathscr{L} = \sum_{i,j=1}^{n}\left(\bar{N}_{ij} - N_{ij}\ln\bar{N}_{ij}\right)$$

**Minimizing $-\ln\mathscr{L}$ ⇔ solving an unbinned maximum-likelihood (ML) fit, where the background and signal probability density functions (PDF) are one two-dimensional and two one-dimensional binned histograms.**

**➔ Improve information content of the fit (more precise determination**
**➔ Probes the signal shape in 2D**
**➔ But signal correlation in each tiles, induce a bias…**



# Conclusion

- **LHC is delivering a huge chuck of data that experiments are currently using for commissioning and looking for new physics.**
- **A large variety of method to estimate SM process from data have been looked at over MC to understand the bias and are currently exercised on data.**
- **The variety of methods allows cross check and combination of them to reduce systematic/bias.**

**➔ Moriond results will integrate all this and perhaps we will see some signal above the SM background…**

# Charge Asymmetry

**In case of dilepton searches, use the symmetry in the charge of multijet background to determine it.**

**Same sign searches:**

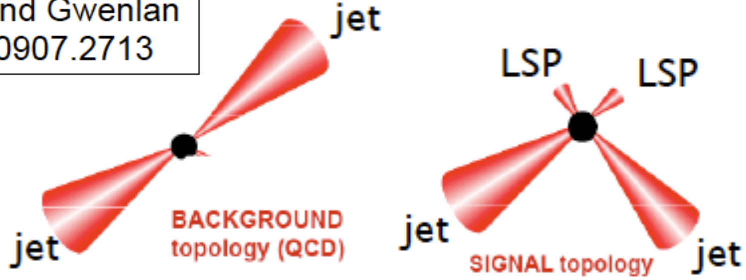
- **Very low Standard Model background rate**
- **Backgrounds from charge mis-identified**

**Opposite sign searches:**

- **Use opposite-sign, opposite-flavor sample to subtract SM background**

# New Variables: All Hadronic Searches

Barr and Gwenlan
arXiv:0907.2713




$$\alpha_T = \frac{E_{T\,j2}}{M_{T\,j1j2}} = \frac{\sqrt{E_{T\,j2}/E_{T\,j1}}}{\sqrt{2(1-\cos\Delta\varphi)}}$$

**A new variable combining angular and energy measurements ($\alpha_T$)**
**No dependence on MET ➔ robust**
**Originally proposed for di-jet events ➔ generalised up to 6 jets**
**Perfectly balanced events have $\alpha_T$=0.5**
**Mis-measurement of either jet leads to lower values**
**Studies the variation of the variable as function of others**

PRL101:221803 (2008) & CMS-PAS-SUS-09-001