Informal session on tools and techniques

Luis Roberto Flores Castillo

University of Wisconsin-Madison
September 7, 2012

XV Mexican School of Particles and Fields
September 6-15, 2012

Benemérita Universidad Autónoma de Puebla
Puebla, Mexico

Foreword

- The goal of this session is to provide some relatively unstructured time for questions and answers on
  - The Higgs search in ATLAS
  - The techniques used in the search
  - Some topics not mentioned yesterday
- If you want to ask in Spanish, please go ahead
- I suggest a list of topics and give a few comments on each, but feel free to propose others

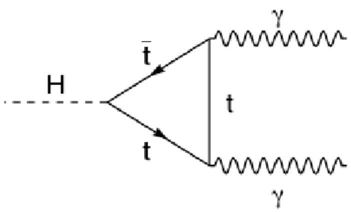
Suggested topics

- Full simulation, fast simulation, toy MC
- Control regions and cross checks
- “Blind” analyses
- Combination of search channels
- Multivariate methods
  - Likelihood ratios
  - Artificial Neural Networks
  - Boosted decision trees
- Organization of a big collaboration

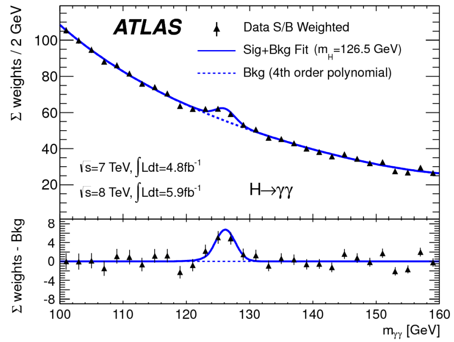
5-slide reminder (and a bit more)

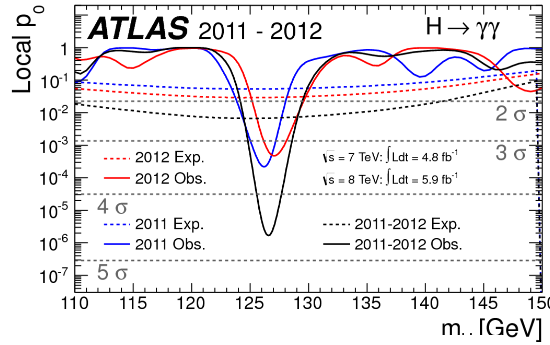
Three most sensitive channels

Strength at m_H=126 GeV

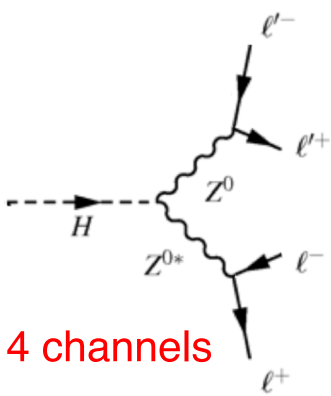


10 categories

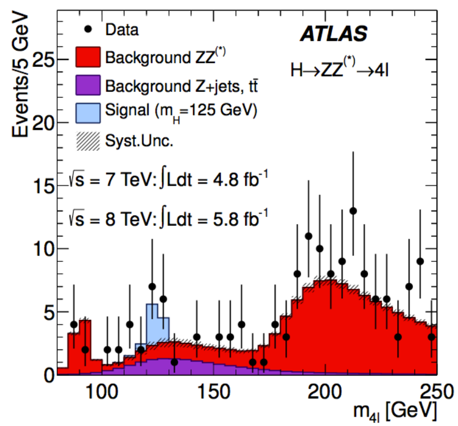


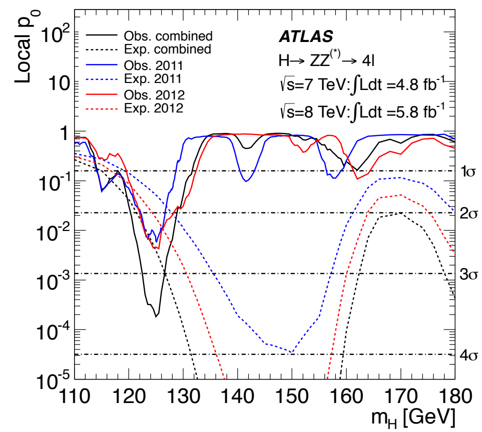


1.8 ± 0.5

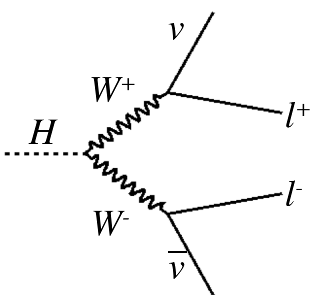


4 channels



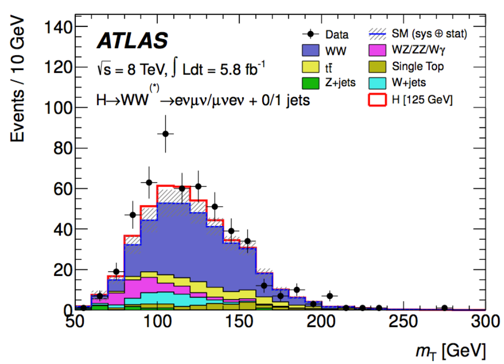


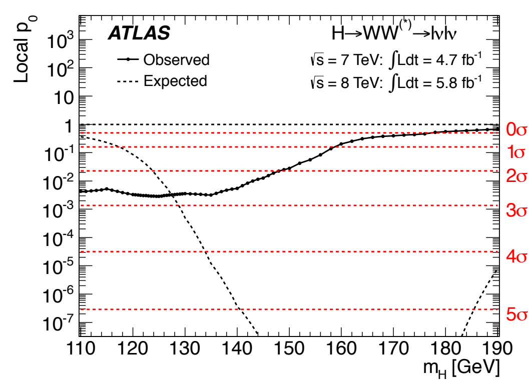
1.2 ± 0.6



3 channels

(based on jet multiplicity)





1.4 ± 0.5

Combination summary

Higgs Boson Decay	Subsequent Decay	Sub-Channels	m_H Range [GeV]	$\int L\,dt$ [fb^{-1}]	Ref.
		2011 $\sqrt{s}=7$ TeV			
$H \to ZZ^{(*)}$	4ℓ	$\{4e, 2e2\mu, 2\mu 2e, 4\mu\}$	110–600	4.8	[87]
	$\ell\ell\nu\bar{\nu}$	$\{ee, \mu\mu\} \otimes$ {low, high pile-up}	200–280–600	4.7	[125]
	$\ell\ell q\bar{q}$	{b-tagged, untagged}	200–300–600	4.7	[126]
$H \to \gamma\gamma$	–	10 categories $\{p_{\mathrm{Tt}} \otimes \eta_\gamma \otimes \text{conversion}\} \oplus \{\text{2-jet}\}$	110–150	4.8	[127]
$H \to WW^{(*)}$	$\ell\nu\ell\nu$	$\{ee, e\mu/\mu e, \mu\mu\} \otimes$ {0-jet, 1-jet, 2-jet} $\otimes$ {low, high pile-up}	110–200–300–600	4.7	[106]
	$\ell\nu qq'$	$\{e, \mu\} \otimes$ {0-jet, 1-jet, 2-jet}	300–600	4.7	[128]
$H \to \tau\tau$	$\tau_{\mathrm{lep}}\tau_{\mathrm{lep}}$	$\{e\mu\} \otimes \{\text{0-jet}\} \oplus \{\ell\ell\} \otimes \{\text{1-jet, 2-jet}, VH\}$	110–150	4.7	[129]
	$\tau_{\mathrm{lep}}\tau_{\mathrm{had}}$	$\{e, \mu\} \otimes \{\text{0-jet}\} \otimes \{E_{\mathrm{T}}^{\mathrm{miss}} < 20\text{ GeV}, E_{\mathrm{T}}^{\mathrm{miss}} \geq 20\text{ GeV}\}$ $\oplus \{e, \mu\} \otimes \{\text{1-jet}\} \oplus \{\ell\} \otimes \{\text{2-jet}\}$	110–150	4.7	
	$\tau_{\mathrm{had}}\tau_{\mathrm{had}}$	{1-jet}	110–150	4.7	
$VH \to Vbb$	$Z \to \nu\nu$	$E_{\mathrm{T}}^{\mathrm{miss}} \in \{120-160, 160-200, \geq 200\text{ GeV}\}$	110–130	4.6	[130]
	$W \to \ell\nu$	$p_{\mathrm{T}}^{W} \in \{< 50, 50-100, 100-200, \geq 200\text{ GeV}\}$	110–130	4.7	
	$Z \to \ell\ell$	$p_{\mathrm{T}}^{Z} \in \{< 50, 50-100, 100-200, \geq 200\text{ GeV}\}$	110–130	4.7	
		2012 $\sqrt{s}=8$ TeV			
$H \to ZZ^{(*)}$	4ℓ	$\{4e, 2e2\mu, 2\mu 2e, 4\mu\}$	110–600	5.8	[87]
$H \to \gamma\gamma$	–	10 categories $\{p_{\mathrm{Tt}} \otimes \eta_\gamma \otimes \text{conversion}\} \oplus \{\text{2-jet}\}$	110–150	5.9	[127]
$H \to WW^{(*)}$	$e\nu\mu\nu$	$\{e\mu, \mu e\} \otimes$ {0-jet, 1-jet, 2-jet}	110–200	5.8	[131]

- 4+4+2+10+18+6+4+7+1+3+4+4+4+10+6 = 87 !!!
- Common parameters: m_H, μ, lumi uncertainty,

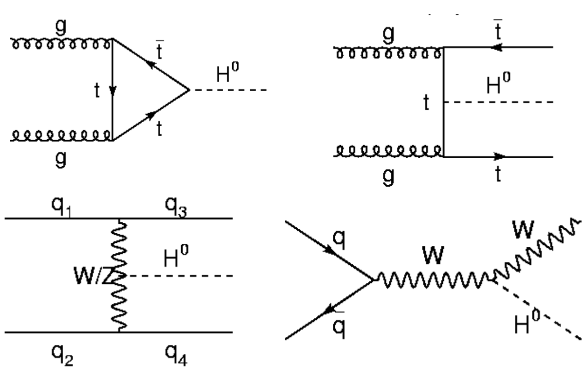
Correlated systematic uncertainties

- Integrated luminosity (3.9% for 2011, 3.6% for 2012)
- Electron and photon trigger and identification efficiencies
- Electron and photon energy scales: five parameters (calibration method, presampler ES in B and EC, material)
- Muon reconstruction, separate for ID and MS
- Jet energy scale and missing transverse energy (dependent on p_T, η, jet flavor, specific treatment for b-jets)
- Sources affecting 7 & 8 TeV data fully correlated
- Uncertainties on background estimates based on control samples considered uncorrelated between 7 and 8 TeV

Correlated systematic uncertainties

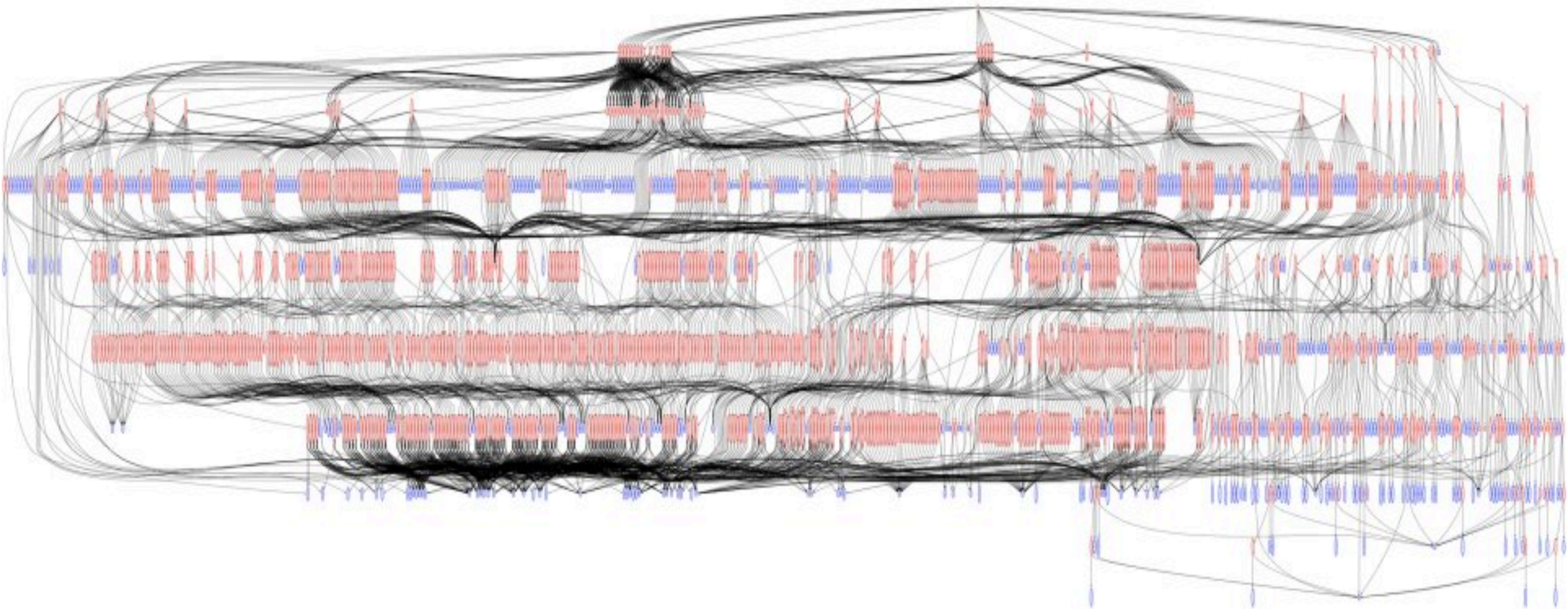
Theory uncertainties: mostly correlated for signal predictions

- QCD scale uncertainties for m_H=125 GeV :
  - ~8% for ggF
  - 1% VBF and WH/ZH
  - +4%, -9% for ttH
- Uncertainties on predicted branching ratios ~ 5%
- Parton Distribution Functions:
  - 8% for predominantly gluon-initiated ggF and ttH
  - 4% for predominantly quark-initiated VBF and WH/ZH



- Higgs production w/additional jets in $\gamma\gamma$, $l\nu l\nu$, $\tau\tau$ reduced to 25%
- Additional unc. on signal normalization: ±150%×$(m_H/\text{TeV})^3$ (4% for m_H = 300 GeV)

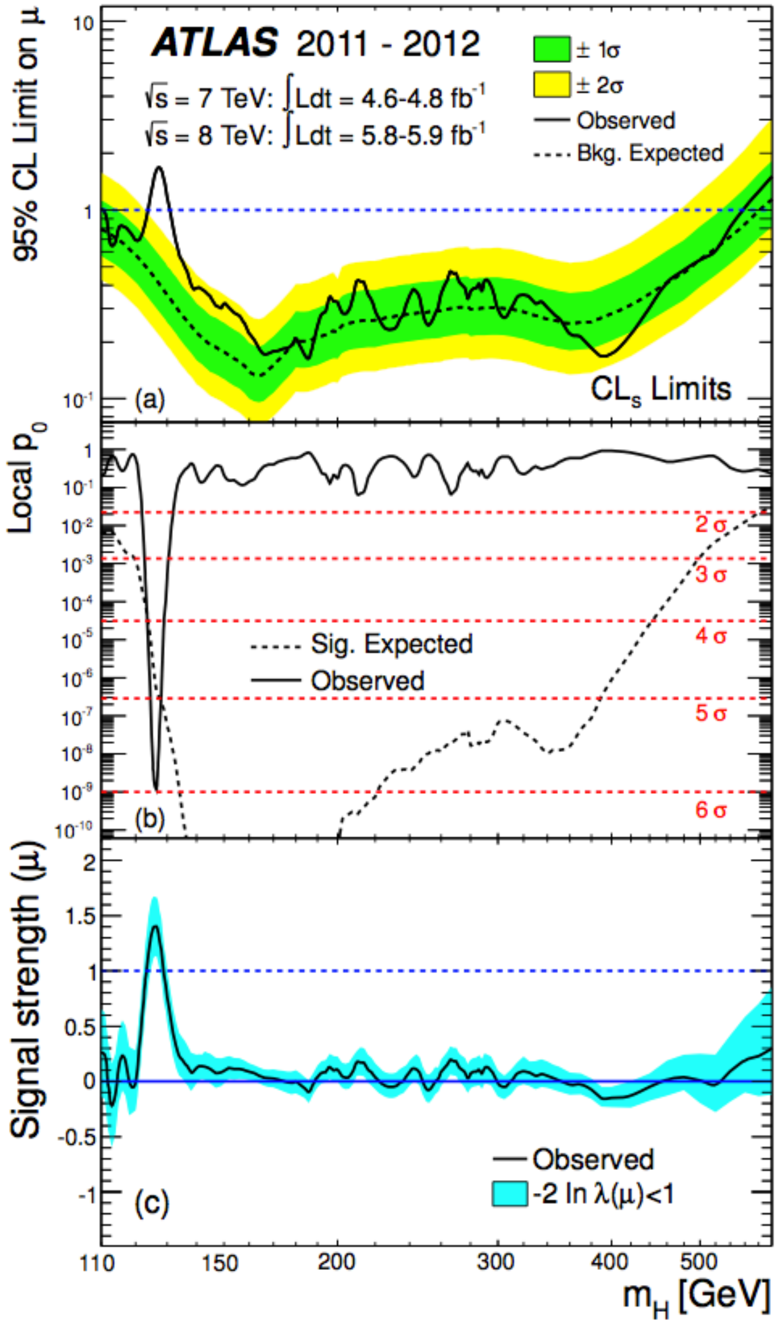
ATLAS combination model



- Each channel has its own data streams, triggers, control regions, main backgrounds, systematic uncertainties, …
- Each team develops its own code for the analysis
- All are put it into a common file format to allow the combination
- Non-negligible amount of work on just *naming conventions!*

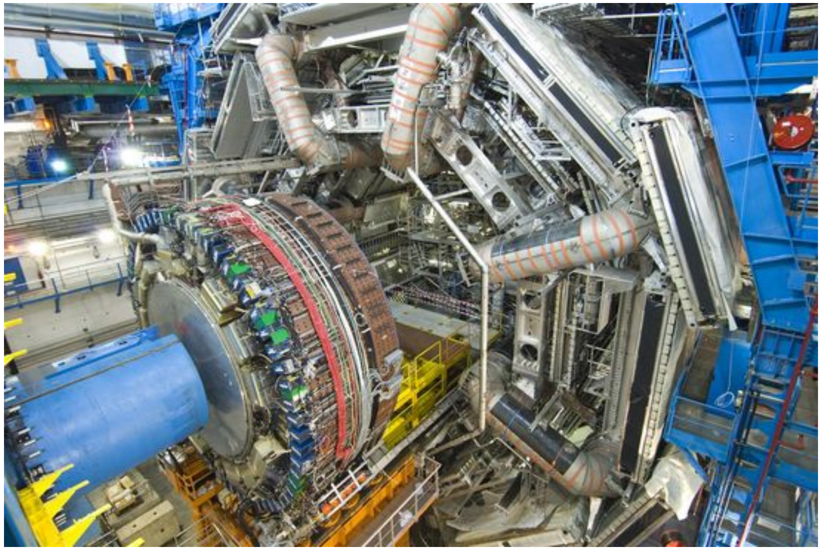
Three views of the combination

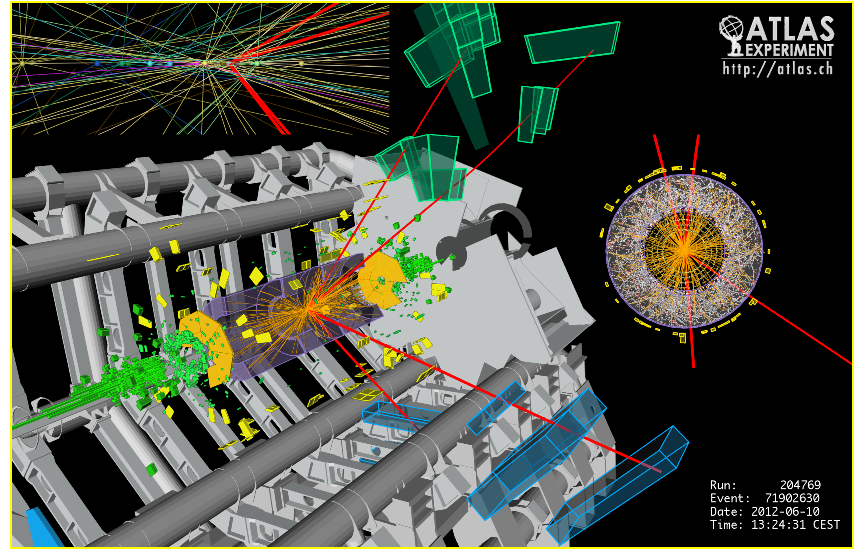
- As a limit: fluctuations around expected … except ~125 GeV
- Probability that the excess comes from background only: below 2σ everywhere … except ~125 GeV
- Signal strength (SM=1): compatible with 0 … except ~125 GeV

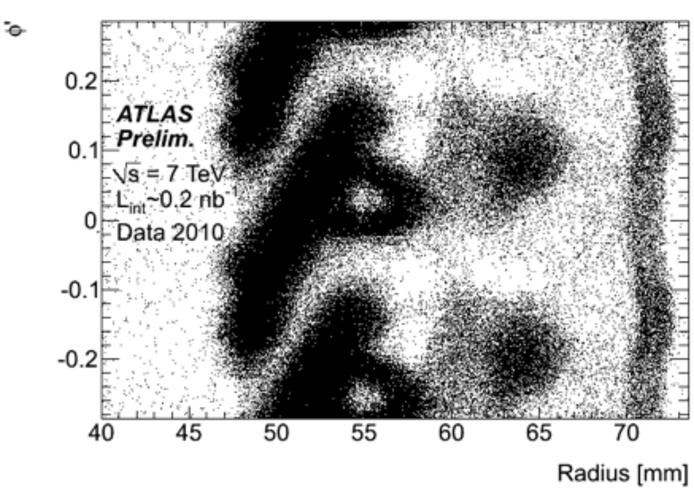


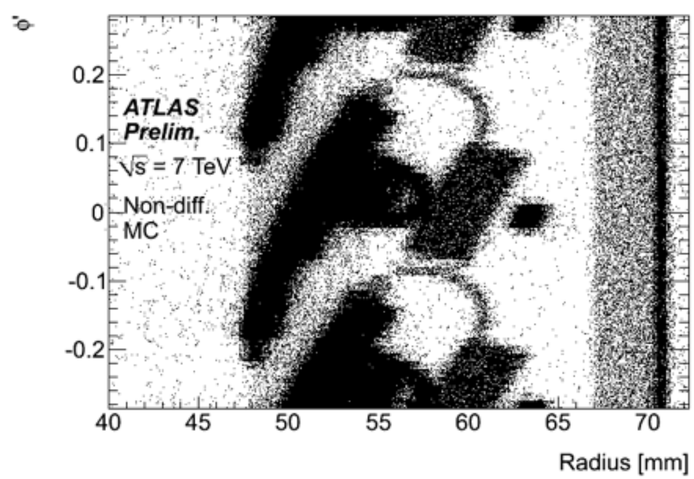
Full vs fast simulation, toy MC

Full detector simulation

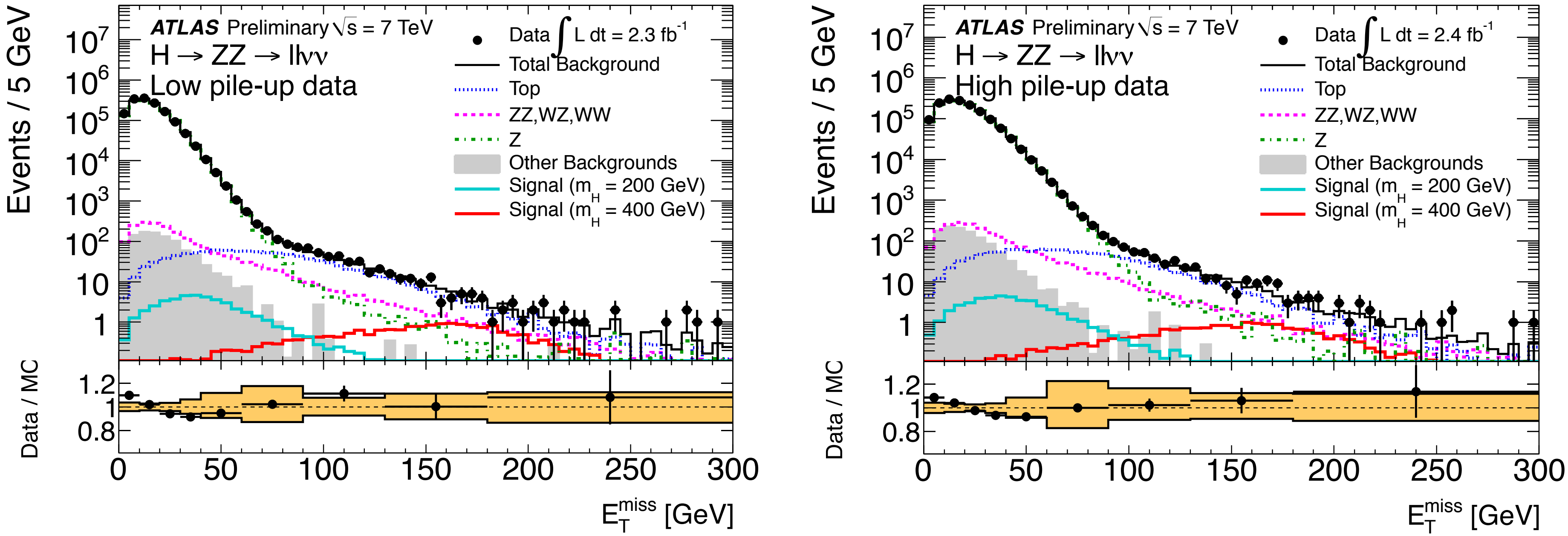








Full detector simulation

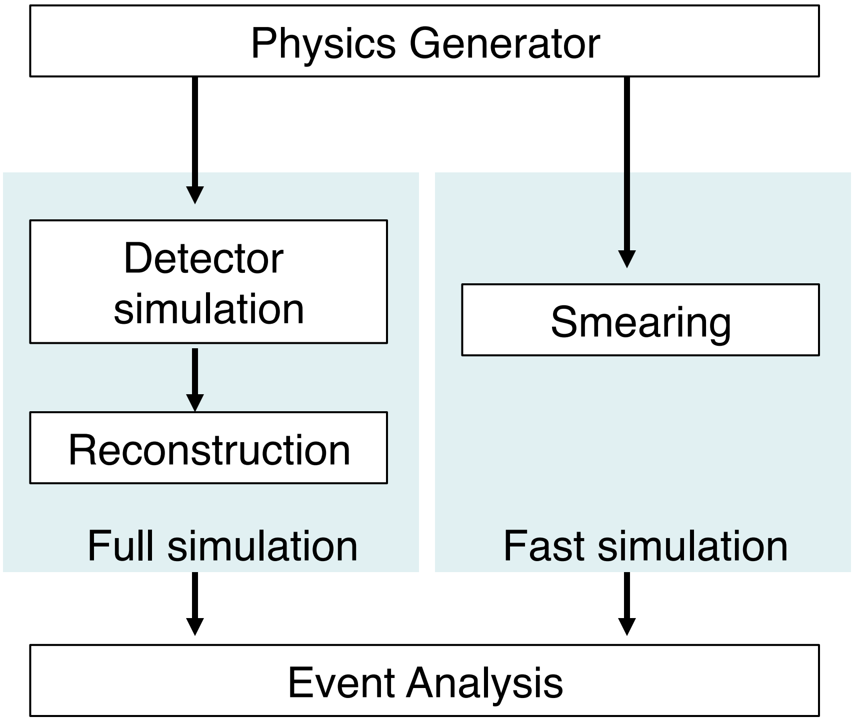


- Missing transverse momentum distribution for events with exactly two oppositely charged electrons or muons with $|m_{ll}-m_Z|<15$ GeV

Full vs Fast vs toy MC

Full simulation:

- detailed simulation of
  - particles' passage through detector material
  - Magnetic fields
  - Particle trajectories
  - Hits left
  - Triggers
  - …
- Reconstruction algorithms: same as applied in data



Fast simulation:

- Apply resolution functions as measured in data or full simulation.

Toy MC:

- Use only final distributions; e.g., to test fit procedures.

Control regions and cross checks

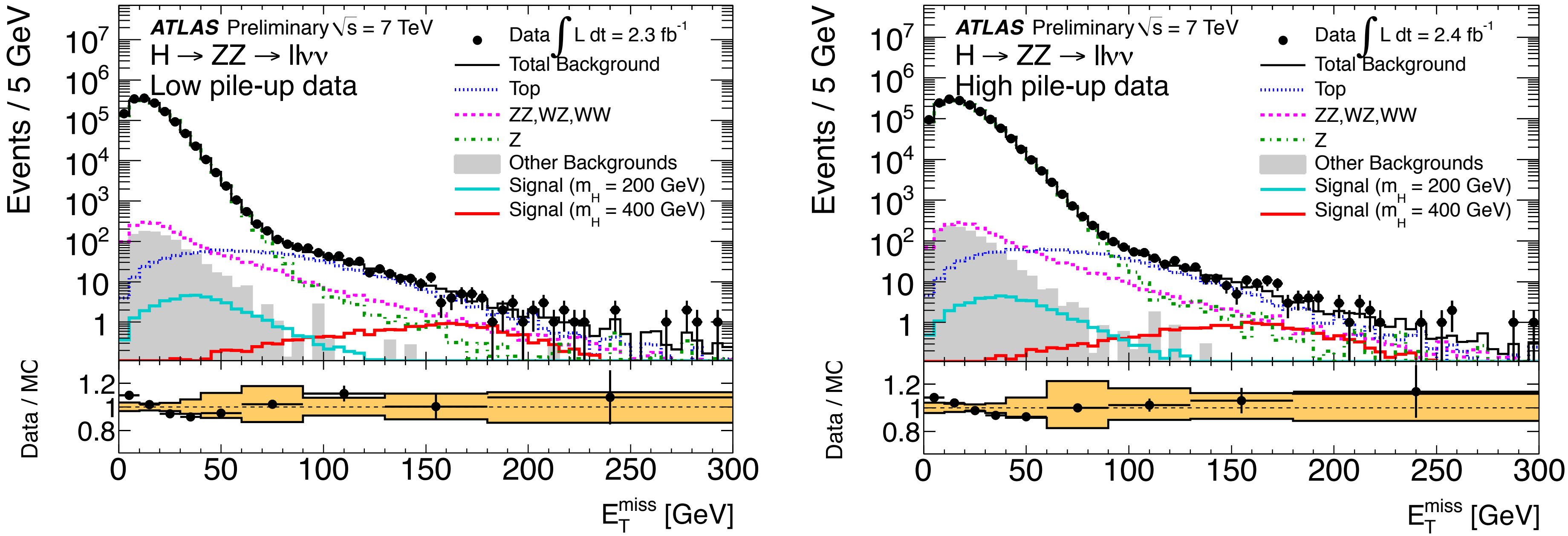
Control regions

- In a search (or measurement), we are interested in the set of events that satisfy some specific cuts (the “signal region”)
  – Example: in the $H \rightarrow 4l$ search: Events with 4 leptons, all isolated and with small impact parameters.

- A “control region” refers to a sample of events obtained by varying the main selection cuts; examples:
  – Only 2 leptons are isolated; no requirement on the rest
  – 2 leptons are required to *fail* the impact parameter cut
  – Two electrons satisfy only less strict criteria than usual (to be called an “electron”, a set of cells in the calorimeter should pass a large number of cuts)

- Objectives:
  – Check MC/data agreement in a larger sample
  – Estimate the number of events from processes other than signal

Control regions and cross checks



- Missing transverse momentum distribution for events with exactly two oppositely charged electrons or muons with $|m_{ll}-m_Z|<15$ GeV

Summary of background estimations in H→4l

8 TeV

Method	Estimated number of events
4μ	
m_{12} fit: $Z+$jets contribution	$0.51\pm 0.13 \pm 0.16^{\dagger}$
m_{12} fit: $t\bar{t}$ contribution	$0.044\pm0.015\pm0.015^{\dagger}$
$t\bar{t}$ from $e^{\pm}\mu^{\mp}+\mu^{\pm}\mu^{\mp}$	$0.058\pm0.015\pm0.019$
$2e2\mu$	
m_{12} fit: $Z+$jets contribution	$0.41\pm 0.10 \pm 0.13^{\dagger}$
m_{12} fit: $t\bar{t}$ contribution	$0.040\pm0.013\pm0.013^{\dagger}$
$t\bar{t}$ from $e^{\pm}\mu^{\mp}+\mu^{\pm}\mu^{\mp}$	$0.051\pm0.013\pm0.017$
$2\mu 2e$	
$\ell\ell+e^{\pm}e^{\mp}$	$4.9\pm 0.8 \pm 0.7^{\dagger}$
$\ell\ell+e^{\pm}e^{\pm}$	$4.1\pm 0.6 \pm 0.8$
$3\ell+\ell$ (same-sign)	$3.5\pm 0.5 \pm 0.5$
$4e$	
$\ell\ell+e^{\pm}e^{\mp}$	$3.9\pm 0.7 \pm 0.8^{\dagger}$
$\ell\ell+e^{\pm}e^{\pm}$	$3.1\pm 0.5 \pm 0.6$
$3\ell+\ell$ (same-sign)	$3.0\pm 0.4 \pm 0.4$

7 TeV

Method	Estimated number of events
4μ	
m_{12} fit: $Z+$jets contribution	$0.25\pm 0.10 \pm 0.08^{\dagger}$
m_{12} fit: $t\bar{t}$ contribution	$0.022\pm0.010\pm0.011^{\dagger}$
$t\bar{t}$ from $e^{\pm}\mu^{\mp}+\mu^{\pm}\mu^{\mp}$	$0.025\pm0.009\pm0.014$
$2e2\mu$	
m_{12} fit: $Z+$jets contribution	$0.20\pm 0.08 \pm 0.06^{\dagger}$
m_{12} fit: $t\bar{t}$ contribution	$0.020\pm0.009\pm0.011^{\dagger}$
$t\bar{t}$ from $e^{\pm}\mu^{\mp}+\mu^{\pm}\mu^{\mp}$	$0.024\pm0.009\pm0.014$
$2\mu 2e$	
$\ell\ell+e^{\pm}e^{\mp}$	$2.6\pm 0.4 \pm 0.4^{\dagger}$
$\ell\ell+e^{\pm}e^{\pm}$	$3.7\pm 0.9 \pm 0.6$
$3\ell+\ell$ (same-sign)	$2.0\pm 0.5 \pm 0.3$
$4e$	
$\ell\ell+e^{\pm}e^{\mp}$	$3.1\pm 0.6 \pm 0.5^{\dagger}$
$\ell\ell+e^{\pm}e^{\pm}$	$3.2\pm 0.6 \pm 0.5$
$3\ell+\ell$ (same-sign)	$2.2\pm 0.5 \pm 0.3$

More than one method per channel, compatible results

Uncertainties 20%-70% depending on background and data sample

Summary of background estimations in H→4l

8 TeV

Method	Estimated number of events
4μ	
m_{12} fit: $Z+$jets contribution	$0.51\pm 0.13\pm 0.16^{\dagger}$
m_{12} fit: $t\bar{t}$ contribution	$0.044\pm 0.015\pm 0.015^{\dagger}$
$t\bar{t}$ from $e^{\pm}\mu^{\mp}+\mu^{\pm}\mu^{\mp}$	$0.058\pm 0.015\pm 0.019$
$2e2\mu$	
m_{12} fit: $Z+$jets contribution	$0.41\pm 0.10\pm 0.13^{\dagger}$
m_{12} fit: $t\bar{t}$ contribution	$0.040\pm 0.013\pm 0.013^{\dagger}$
$t\bar{t}$ from $e^{\pm}\mu^{\mp}+\mu^{\pm}\mu^{\mp}$	$0.051\pm 0.013\pm 0.017$
$2\mu 2e$	
$\ell\ell+e^{\pm}e^{\mp}$	$4.9\pm 0.8\pm 0.7^{\dagger}$
$\ell\ell+e^{\pm}e^{\pm}$	$4.1\pm 0.6\pm 0.8$
$3\ell+\ell$ (same-sign)	$3.5\pm 0.5\pm 0.5$
$4e$	
$\ell\ell+e^{\pm}e^{\mp}$	$3.9\pm 0.7\pm 0.8^{\dagger}$
$\ell\ell+e^{\pm}e^{\pm}$	$3.1\pm 0.5\pm 0.6$
$3\ell+\ell$ (same-sign)	$3.0\pm 0.4\pm 0.4$

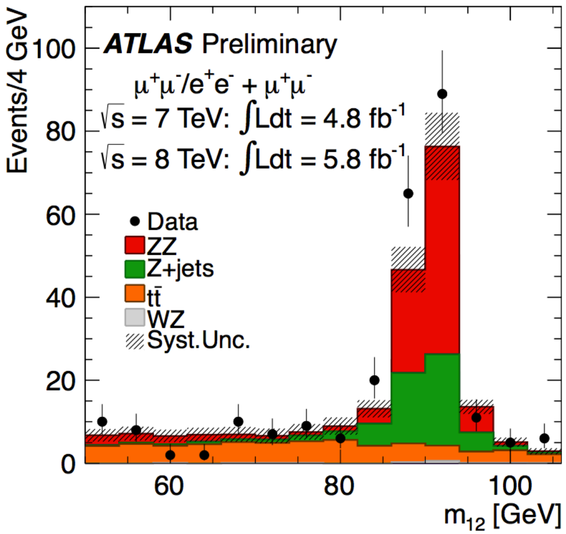
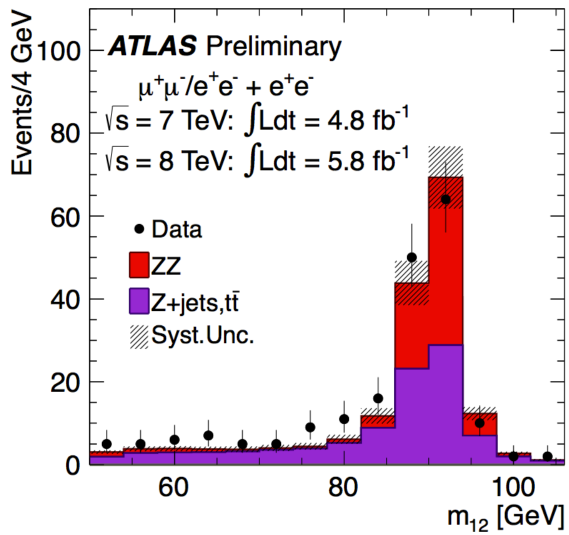
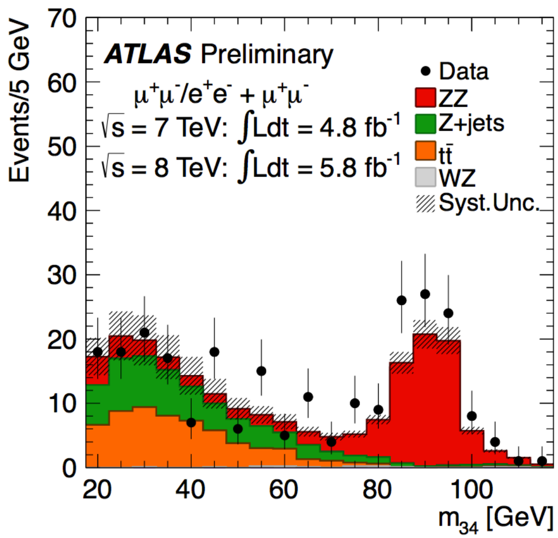
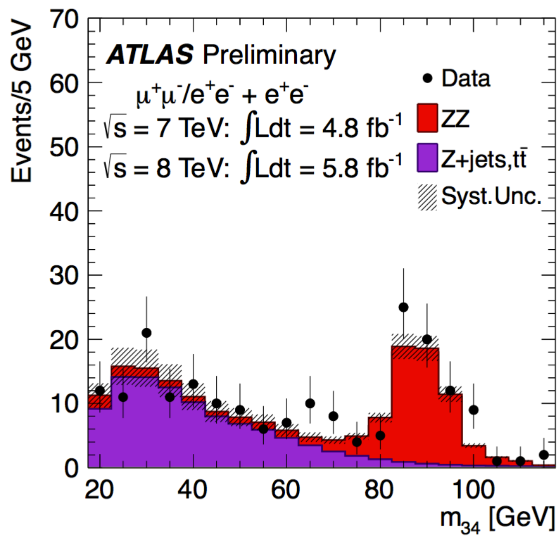
7 TeV

Method	Estimated number of events
4μ	
m_{12} fit: $Z+$jets contribution	$0.25\pm 0.10\pm 0.08^{\dagger}$
m_{12} fit: $t\bar{t}$ contribution	$0.022\pm 0.010\pm 0.011^{\dagger}$
$t\bar{t}$ from $e^{\pm}\mu^{\mp}+\mu^{\pm}\mu^{\mp}$	$0.025\pm 0.009\pm 0.014$
$2e2\mu$	
m_{12} fit: $Z+$jets contribution	$0.20\pm 0.08\pm 0.06^{\dagger}$
m_{12} fit: $t\bar{t}$ contribution	$0.020\pm 0.009\pm 0.011^{\dagger}$
$t\bar{t}$ from $e^{\pm}\mu^{\mp}+\mu^{\pm}\mu^{\mp}$	$0.024\pm 0.009\pm 0.014$
$2\mu 2e$	
$\ell\ell+e^{\pm}e^{\mp}$	$2.6\pm 0.4\pm 0.4^{\dagger}$
$\ell\ell+e^{\pm}e^{\pm}$	$3.7\pm 0.9\pm 0.6$
$3\ell+\ell$ (same-sign)	$2.0\pm 0.5\pm 0.3$
$4e$	
$\ell\ell+e^{\pm}e^{\mp}$	$3.1\pm 0.6\pm 0.5^{\dagger}$
$\ell\ell+e^{\pm}e^{\pm}$	$3.2\pm 0.6\pm 0.5$
$3\ell+\ell$ (same-sign)	$2.2\pm 0.5\pm 0.3$

More than one method per channel, compatible results

Uncertainties 20%-70% depending on background and data sample

Control Regions

	Z+mm	Z+ee
On-shell Z		
Subleading Z		

- Isolation and impact parameter cuts not applied to sub-leading di-lepton
- Normalized to data-driven estimates
- Good data/MC agreement in shape and normalization

“Blinding”

Blind analyses

- The main idea of a “blind” analysis is to avoid looking at the signal region before the analysis procedure is fixed

- The reason: avoid biases.

- Example:

  A modified cut would add three 4-lepton candidates to the peak; should you use it?
  Answer: you should not be asking that question!!

  – Analysis decisions should be made based on EXPECTED sensitivity (i.e., without looking at your data)

  – Otherwise, the significance becomes meaningless

  [it is like testing a new medicine and counting only patients where it worked]

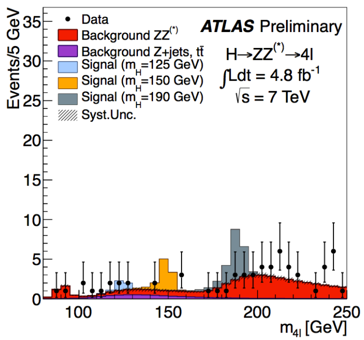
Blind analyses

- It can be done in several ways:
  - Removing a mass range from all plots
  - Adding a “distortion” to the data
  - Use only control regions until procedure is settled
  - Use only “old” data until the procedure is settled

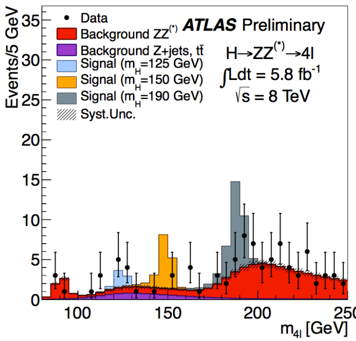
- Not always clear-cut since, as time goes on, pile-up increases, there are software upgrades, running conditions change, reconstruction quantities need to be studied by looking at the most recent data

Combination of channels

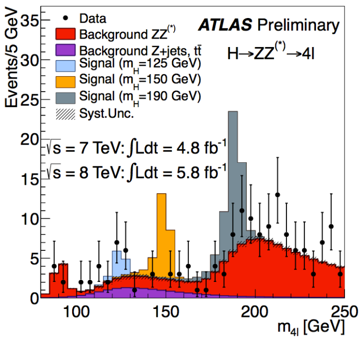
Combination of channels



+



=



?

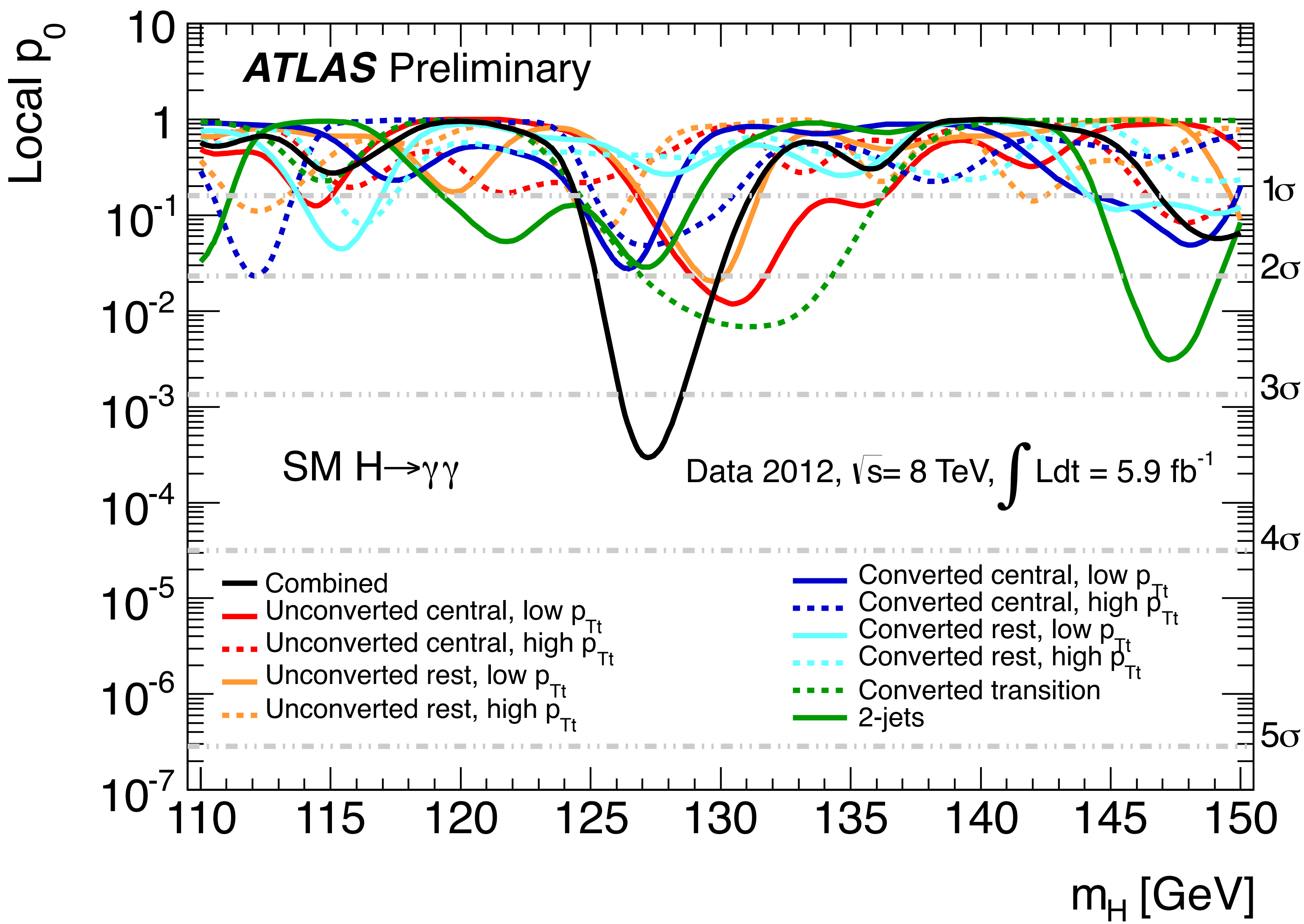
- More than adding histograms
- The reason:
  - imagine a high S/B, low stats, buried into a low S/B, high stats search. Adding them together basically throws away the significance of the one with few events.
  - Instead, treating each separately and adding their likelihoods would have a better significance than the best of them.

Ten categories in H→γγ

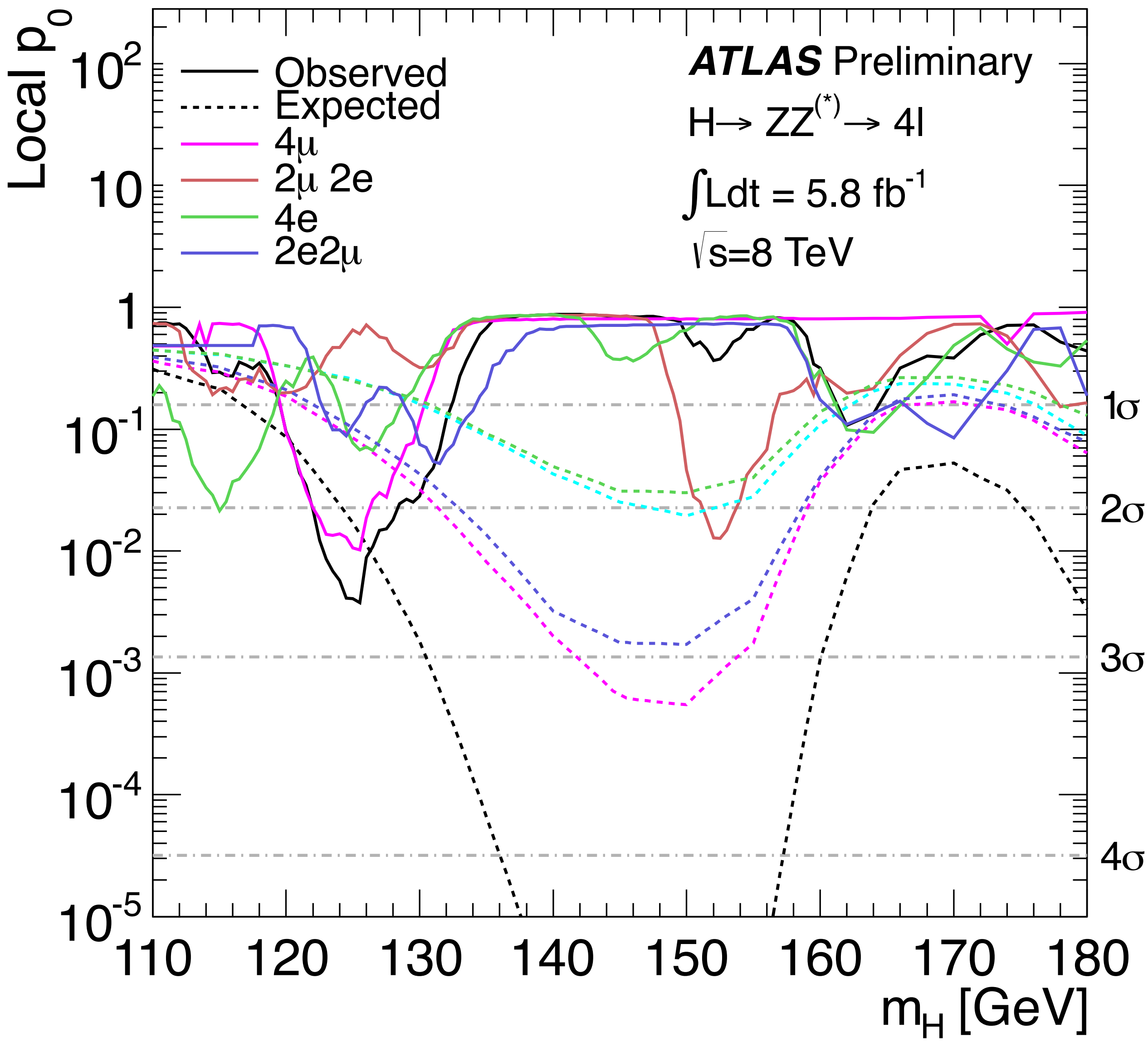
$\sqrt{s}$	7 TeV		8 TeV		
$\sigma \times B(H \rightarrow \gamma\gamma)$ [fb]		39		50	FWHM
Category	N_D	N_S	N_D	N_S	[GeV]
Unconv. central, low p_{Tt}	2054	10.5	2945	14.2	3.4
Unconv. central, high p_{Tt}	97	1.5	173	2.5	3.2
Unconv. rest, low p_{Tt}	7129	21.6	12136	30.9	3.7
Unconv. rest, high p_{Tt}	444	2.8	785	5.2	3.6
Conv. central, low p_{Tt}	1493	6.7	2015	8.9	3.9
Conv. central, high p_{Tt}	77	1.0	113	1.6	3.5
Conv. rest, low p_{Tt}	8313	21.1	11099	26.9	4.5
Conv. rest, high p_{Tt}	501	2.7	706	4.5	3.9
Conv. transition	3591	9.5	5140	12.8	6.1
2-jet	89	2.2	139	3.0	3.7
All categories (inclusive)	23788	79.6	35251	110.5	3.9

- Highest (2-jet) and lowest (conv. rest, low-p_{Tt}) sensitivities

Ten categories in $H \rightarrow \gamma\gamma$



Four channels in $H \to ZZ \to 4l$



Multivariate methods

Multivariate methods

Several MVA methods available

- Likelihood ratio
- Neural networks
- Boosted decision trees

They generally improve the sensitivity.

How much depends on how optimum the original analysis was.

Attention should be paid to

- Evaluation of systematic uncertainties
- Having enough MC to train the MVA

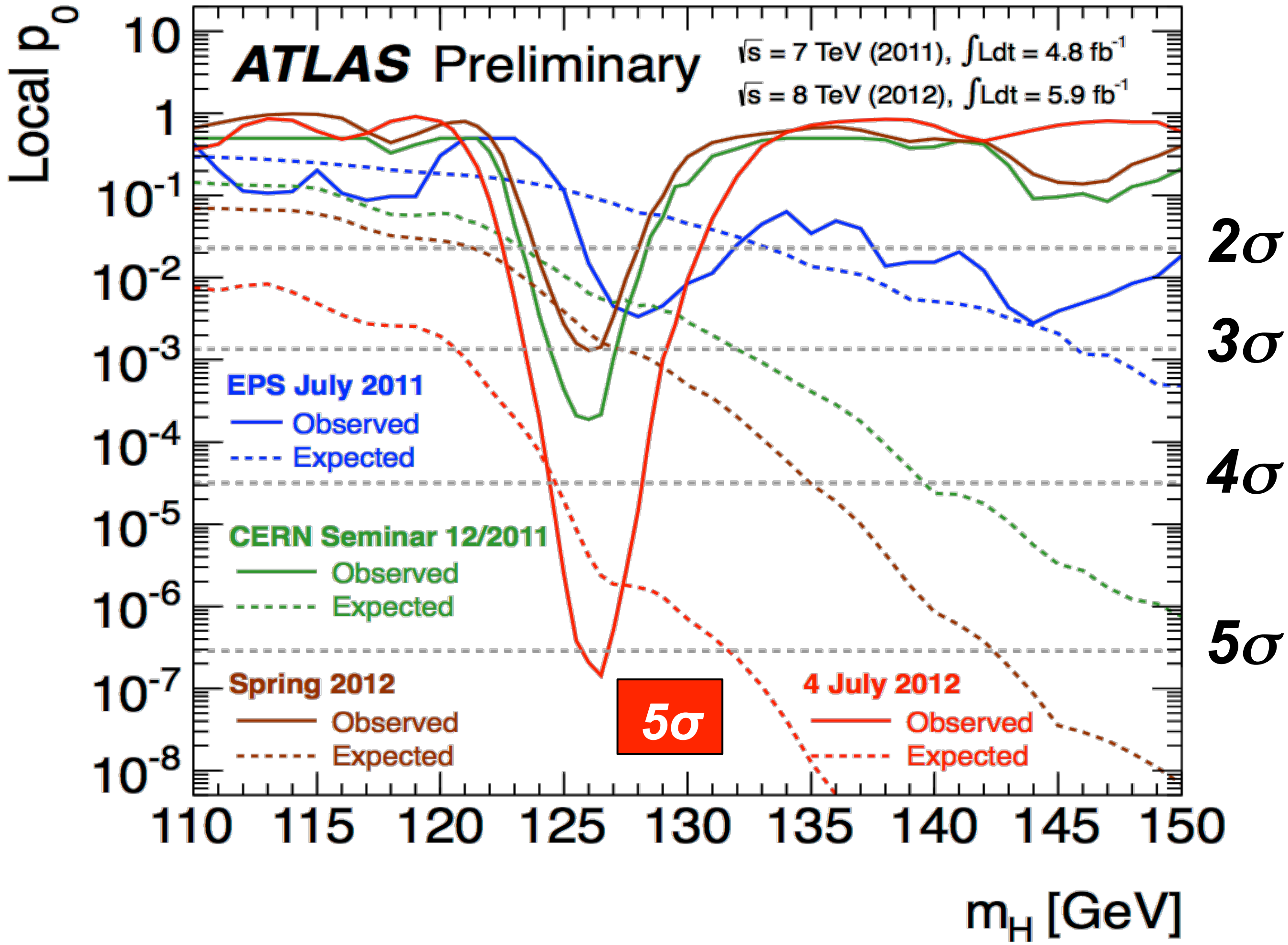
Organization of a big collaboration

ATLAS structure

- 3000 people, 1000 graduate students
- Many vital things to work on:
  - Detectors, trigger, data preparation, software, computing
  - Physics analysis is only the last step in the chain
- Groups:
  - Combined performance, simulation, statistics, detector systems, luminosity, data taking, upgrade
  - Physics: SM, B, Top, SUSY, Higgs, Exotics, Heavy Ions, MC
- Within Higgs group:
  - 7 groups
  - Each covers several analyses
- In one analysis:
  - Analysis group, editors, editorial board, PubCom, signoff.

Evolution of the excess over time

Higgs combination Evolution over time



EPS July 2011

Council Dec 2011

Spring 2012

ICHEP 2012
(July 4, 2012)

Backup

Background estimation *ll+ee*

[nominal]

Z+XX control samples

X: **E**lectrons from heavy flavor,
Electrons from photon **C**onversions,
jets misidentified as electrons (“**F**akes”)

The idea

- Loosen requirements on the two subleading electrons
- Classify each of the two as (E)lectron, (C)onversion or (F)ake
  Nine types of events (EE, EC, EF, CE, CC, CF, FE, FC, FF) [p_T-ordered]
- Using MC-based efficiencies, determine how many of each type is expected in the signal region

- Classification as **E**lectron, **C**onversion or **F**ake based on
  - Transition radiation hits,
  - Number of hits in the innermost pixel layer (the *b*-layer),
  - Fraction of energy deposited in first layer of the EM calorimeter,
  - Lateral containment along φ in the 2nd layer of the EM calorimeter

Background estimation

ll+ee

[nominal]

Z+XX control samples

- Events on each class (based on reconstruction quantities) are a mixture of *true* ee, ec, ef, …

	ee	ec	ef	ce	…
EE					
EC					
EF					
…					

- Composition fractions from MC are used to obtain the expected true composition of each class
  - Limited Z+XX MC; efficiencies obtained from Z+X MC
  - Reweighted to Z+XX p_T spectrum
  - Verified good agreement w/data after isolation, IP and all cuts.
- Final estimate:
  expected true composition * efficiency (true class → signal region)
  $\Sigma_j \Sigma_i$ (true type i)*(efficiency of true i to be reco'd as j in the signal region)
- Low event numbers; toy MC used to obtain central value and uncertainty

Background estimation $ll+ee$

[nominal]

Data/MC comparison

	$4e$		$2\mu 2e$	
	Data	MC	Data	MC
EE	32	22.7±4.8	31	24.9±5.0
EC	6	6.0±2.5	2	1.9±1.4
EF	18	19.0±4.4	26	15.3±3.9
CE	4	8.8±3.0	6	5.1±2.3
CC	1	5.3±2.3	6	4.2±2.0
CF	12	8.8±3.0	15	15.3±3.9
FE	16	5.7±2.4	12	8.4±2.9
FC	6	6.5±2.6	7	4.3±2.1
FF	12	17.4±4.2	16	33.6±5.8
Total	107	100±10	121	113±11

(8 TeV data)

- **Opposite-sign subleading electrons**
- **Estimate based on same-sign subleading electrons also obtained as cross check**