Analysis Walkthrough - Higgs

ATLAS Induction day

Ruggero Turra

INFN Milano

21 Oct 2019

Higgs sector in the Standard Model

$$L_H = (D_\mu \phi)^\dagger (D_\mu \phi) - V(\phi)$$

Spontaneous symmetry breaking gives mass to the vector bosons:

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2 \qquad \phi = \begin{pmatrix} 0 \\ v + H \end{pmatrix} \qquad v = \sqrt{\frac{-\mu^2}{2\lambda}}$$

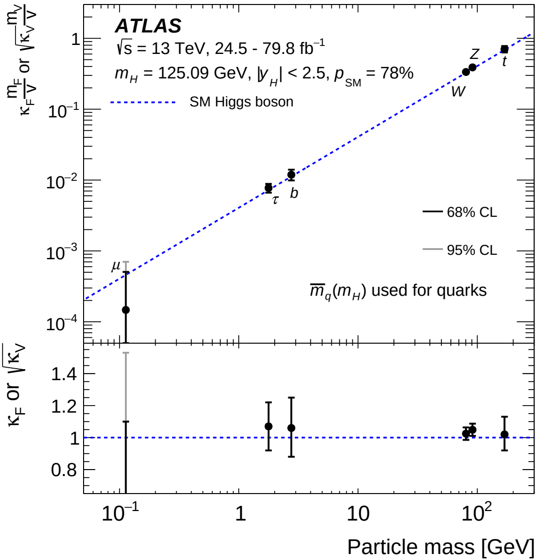
$$(D_\mu \phi)^\dagger (D_\mu \phi) = \frac{1}{2}(\partial_\mu H)(\partial^\mu H) + (v + H)^2 \left[\frac{g^2}{4} W_\mu^\dagger W^\mu + \frac{g^2}{8 \cos^2 \theta_W} Z_\mu Z^\mu \right]$$

So we have obtained mass term: $m_W = m_Z \cos\theta_W = \frac{1}{2} g v$ and $m_H = \sqrt{2\lambda v^2}$
and coupling terms proportional to m_W^2/v and m_Z^2/v or λ for the H self coupling

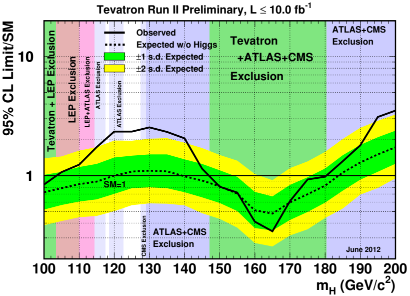
Fermion mass terms are introduced from Yukawa interactions with the Higgs field:

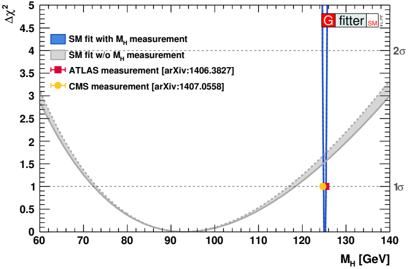
$$L_Y = -\frac{1}{2}(v + H)\lambda_f \bar{f} f$$

where $m_f = \lambda_f \frac{v}{\sqrt{2}}$ and coupling proportional to m_f/v



Works pretty well! Data in agreement with SM (at least within the error bars)

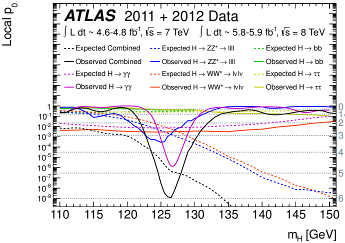




Theoretical, indirect and direct contraints on the Higgs boson mass.

Very small mass region not excluded.

Higgs discovery



Physics Letters B 716 (2012) 1–29

Contents lists available at SciVerse ScienceDirect

Physics Letters B

www.elsevier.com/locate/physletb

ELSEVIER

Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC ☆

ATLAS Collaboration *

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

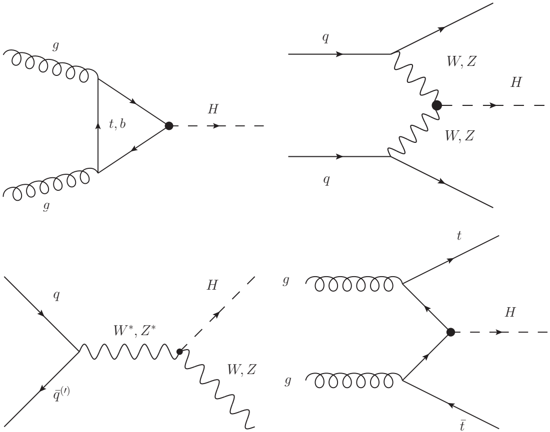
A R T I C L E I N F O

Article history:
Received 31 July 2012
Received in revised form 8 August 2012
Accepted 11 August 2012
Available online 14 August 2012
Editor: W.-D. Schlatter

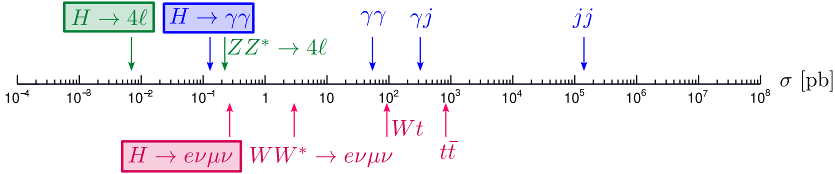
A B S T R A C T

A search for the Standard Model Higgs boson in proton–proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\mathrm{stat}) \pm 0.4\,(\mathrm{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

Higgs at LHC



g
t, b
H
g
q
W, Z
H
W, Z
q
q
H
W*, Z*
W, Z
q̄(′)
t
g
H
g
t̄



$H \to 4\ell$
$H \to \gamma\gamma$
$ZZ^* \to 4\ell$
$\gamma\gamma$
γj
jj
10^{-4} 10^{-3} 10^{-2} 10^{-1} 1 10 10^{2} 10^{3} 10^{4} 10^{5} 10^{6} 10^{7} 10^{8}
σ [pb]
Wt
$t\bar{t}$
$H \to e\nu\mu\nu$
$WW^* \to e\nu\mu\nu$

Higgs after 2012

If you read carefully the Higgs discovery paper you will read:

> “*discovery of a new particle [. . .] consistent with the SM Higgs boson hypothes [. . .] more data are needed to assess its nature in detail*”

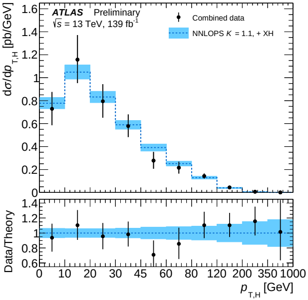
The Higgs group focus on:

- Measure the properties of the Higgs boson trying to constrain the SM (indirect search of new physics) searching for (small) deviations: precision measurements
  - Usually this means to combine many analyses focusing on different decay/production modes
- Direct search for new physics (e.g. new excess with signature similar to the one of the SM Higgs Boson)

Precision physics: measurement and interpretations

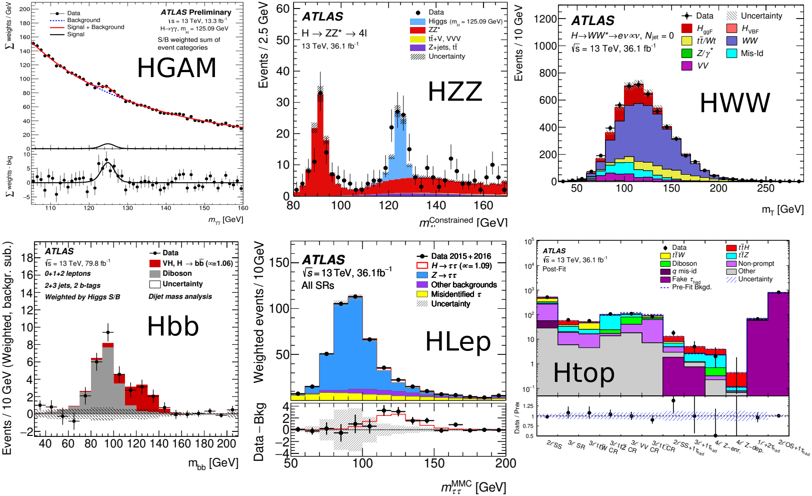
Measurements:

- Cross sections
  - fiducial (e.g. $|y_H| < 2.5$)
  - per production mode (e.g. ggF 0 jets, ...)
  - differential (e.g. $p_T^H, N_j, \ldots$)
- Mass



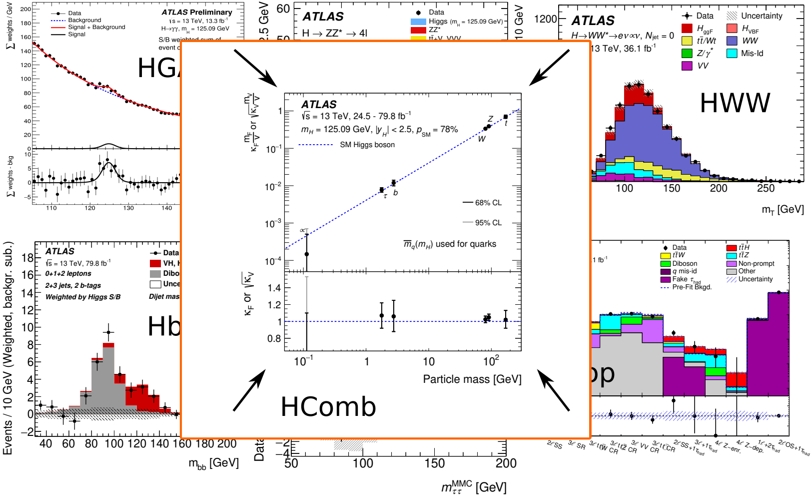
The second part is interpretations:

- How the measured $\sigma \times Br$ are compatible with the SM?
- From several $\sigma_i \times Br_f$ (different production and decay modes) infer the coupling between Higgs and SM particles
- Constrain Wilson coefficients in Effective Filed Theory
- Probe specific BSM models: two-Higgs-doublet model (2HDM), spin $\neq 0$, CP-odd fermion interactions, ...
- Width (SM: 4 MeV): indirectly



Suggestion

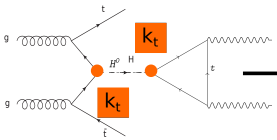
Choose one group connected to the performance study you are doing

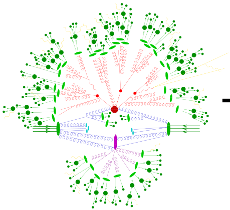


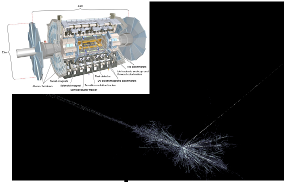
Suggestion

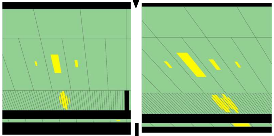
Choose one group connected to the performance study you are doing

How analyses are done

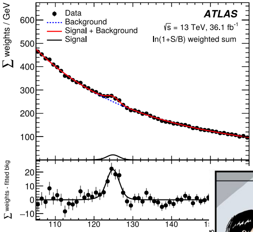




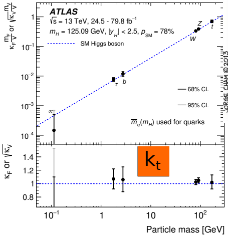




```
************************************************************************
* HGamPhoto * HGamPhoto * HGamPhoto * HGamPhoto * HGamPhoto * HGamPhoto *
************************************************************************
* 29.888656 * 0.9771484 * 603.58868 * 0.9633792 * -0.000668 * -0.001239 *
* 169.17051 * 0.9525480 *         0 * 0.9821917 * 0.0184388 * 0.0036611 *
*           *           *           *           *           *           *
* 0.0904846 * 0.9897341 *         0 * 0.9662012 * -0.000481 * -0.000493 *
* 12.654571 * 0.9385359 *         0 * 0.9735243 * 0.0025533 * 0.0021982 *
* 89.138832 * 0.9304037 * 591.24932 * 0.9727476 * 0.0049735 * 0.0023095 *
* 17.837196 * 0.9876792 *         0 * 0.9716276 * -0.000455 * -0.001094 *
* 23.108921 * 0.9700428 * 576.01599 * 0.9815456 * 0.0232527 * 0.0248024 *
* 28.042976 * 0.9733156 *         0 * 0.9607071 * -0.001524 * -0.001623 *
* 40.976474 * 0.9586699 * 123.75949 * 0.9570868 * 0.0025046 * -0.002294 *
* 77.887596 * 0.9788494 * 86.849739 * 0.9480844 * 0.0076855 * -0.001612 *
* 47.115467 * 0.9218847 *         0 * 0.9654291 * 0.0104289 * 0.0103499 *
* 63.076416 * 0.6504288 *         0 * 0.9387179 * 0.0159983 * 0.0151415 *
* 65.951156 * 0.9655177 * 637.56341 * 0.9576712 * -0.000718 * -0.000567 *
* 147.03241 *  0.974675 * 215.25540 * 0.9597544 * -0.004000 * 0.0012847 *
* 32.016052 * 0.9842315 *         0 * 0.9623845 * 0.0001832 * 0.0089165 *
* 47.263755 * 0.9828590 * 639.28521 * 0.9466156 * -0.013425 * 0.0042643 *
```



interpretation



select
categorize
modelize the data

Before starting

Before starting you should be able to answer to

- What is the impact of your result? Why is it interesting?
- What is your main background?
- How will you estimate the background? How precise you must be? Can you estimate it from data control regions? Or do you need prediction? How accurate are they?
- What are your observables? How will you model them?
- How much luminosity you need to make your result interesting?
- Do you expect to be dominated by experimental or theoretical systematics?
- Do you have theoretical prediction to compare? How are they accurate?

Expression of interest meeting

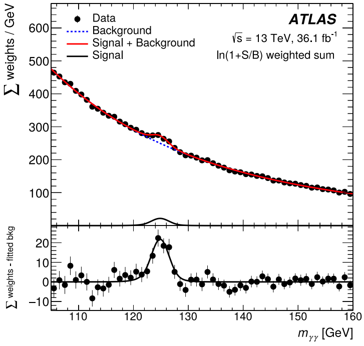
You don't need to finalize your analysis to be able to answer: back of the envelope calculation should be enough

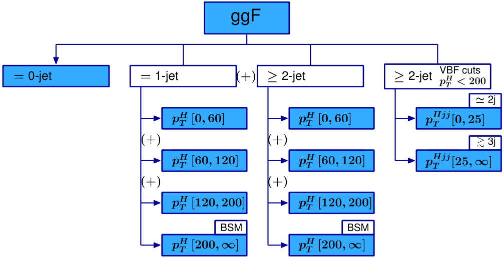
And: what is your timeline? Targeting a conference? Paper/CONF note? Who will work on it?

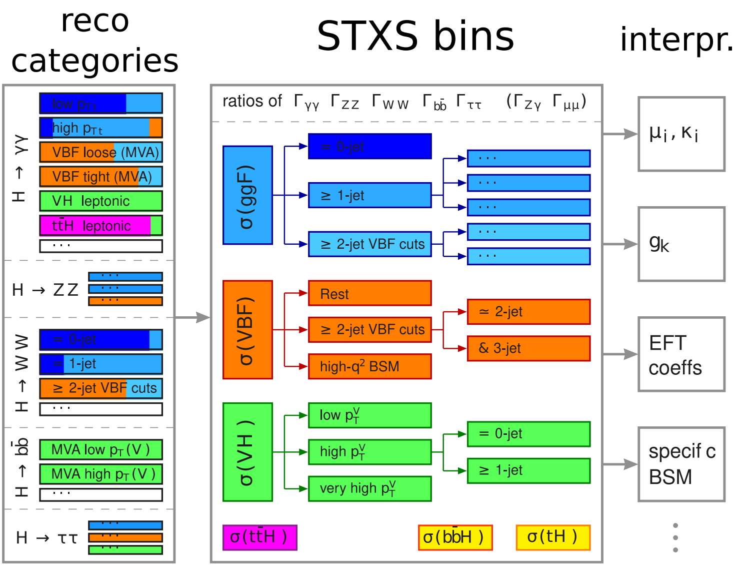
When you have these answers organize the expression of interest meeting within your subgroup

Analysis example: $H \to \gamma\gamma$

- Decay fully reconstructed and good resolution: peak in $m_{\gamma\gamma}$
- Backgroud under the peak can be estimated fitting the sidebands:
  - How to do the fit? How to evaluate the systematic on the background model?
- Large statistics: split events in categories to measure separately Simplified Cross Sections (STXS) cross sections (ggF with 0 jet, . . .)
- STXS bins are too many, due to statistics we merge some
- For most of the categories the main background is $qq/gg \to \gamma\gamma$ and fakes (γj)






reco categories
STXS bins
interpr.
H → γγ
low p_{Tt}
high p_{Tt}
VBF loose (MVA)
VBF tight (MVA)
VH leptonic
$t\bar{t}H$ leptonic
. . .
H → ZZ
H → WW
= 0-jet
= 1-jet
≥ 2-jet VBF cuts
H → $b\bar{b}$
MVA low $p_T(V)$
MVA high $p_T(V)$
H → ττ
ratios of $\Gamma_{\gamma\gamma}$ Γ_{ZZ} Γ_{WW} $\Gamma_{b\bar{b}}$ $\Gamma_{\tau\tau}$ $(\Gamma_{Z\gamma}$ $\Gamma_{\mu\mu})$
σ(ggF)
= 0-jet
≥ 1-jet
≥ 2-jet VBF cuts
σ(VBF)
Rest
≥ 2-jet VBF cuts
high-q^2 BSM
≃ 2-jet
& 3-jet
σ(VH)
low p_T^V
high p_T^V
very high p_T^V
= 0-jet
≥ 1-jet
σ($t\bar{t}H$)
σ($b\bar{b}H$)
σ(tH)
μ_i, κ_i
g_k
EFT coeffs
specif c BSM

Monte Carlo samples

Analyses are designed using MC simulated samples (or control regions)

- For precision physics many signals (Higgs signals) and background samples are reused for several analyses
- Higgs signal samples are quite standard. The main complication is to choose the setup for theoretical systematics (ask Physics Modeling Group – PMG)
- If possible background is data-driven: this doesn't mean you don't need MC

Having all the samples with all the information you need, understading exactly what is the meaning of each variable is one of the most difficult part of the analysis.

Usually each (sub)group has some experts taking care of samples production

Selection and categorization

We select only events passing some criteria, to suppress background

- Quality selection of objects (photons, electrons, . . .) are quite standard (test several working points)

Check the performance of the selections, give feedback to performance groups (e/gamma, . . .)

- Kinematic selection are driven by your process ($H \to \gamma\gamma$), and limited by the detector acceptance and by the trigger
- Usually events are categorized in good/bad events, or to measure multiple quantities at the same time (ggF/VBF, . . . productions)
  - Simple cuts, machine learning, . . .
  - More agressive optimization are usually more model-dependent

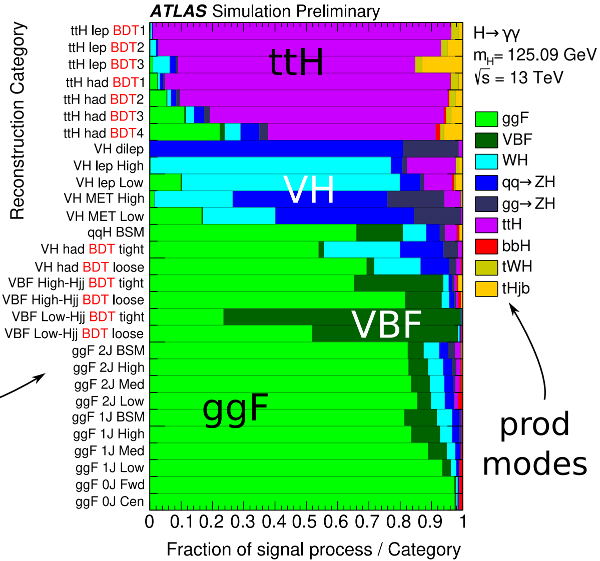
Optimization

- We want to measure different production modes (ggF+0J, ... VBF with 2J, ... ttH)
- We need to separate these kind of events in categories
- The categorization is a function of reconstructed quantities
  - Choose your metric
  - Choose the input quantities
  - Choose the algorithm: NN / BDT / ...
  - Sometimes a couple of cuts are enough

Suggestion

This is the main part of the analysis where people are very competitive.

Share your work, ideas, mistakes, ...: work as a team.



29 reco-categories inspired by STXS

prod modes

Machine learning

- Machine learning is very popular and there are a lot of very powerful tools

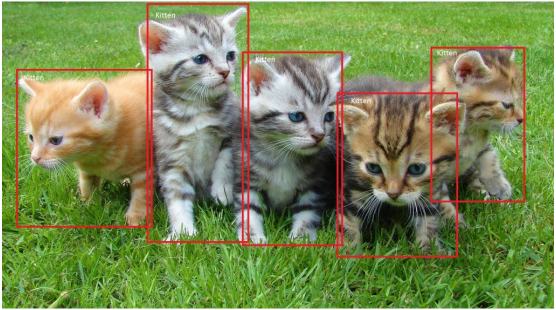
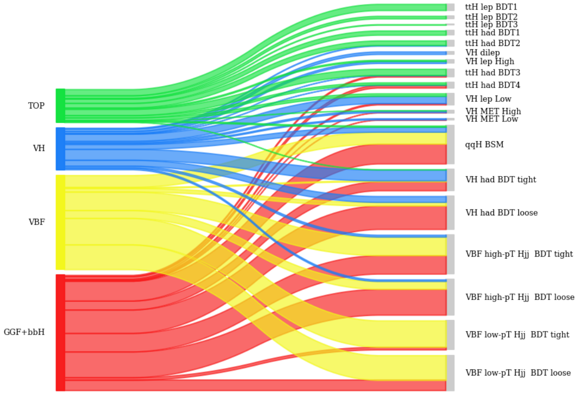


Figure: Real photo of real cats made by milions of pixels

In HEP we mostly train on MC: be sure your algorithm is not learning a quantity not well modeled
Convincing people that you can well model (e.g. estimate efficiencies with small systematic) a black box taking as input everything you can measure is hard. Are you sure the systematic variations you are doing are enough for such a complicated algoritm?

You have defined your selection and categorization. You have to model your observables as a function of the parameters you want to measure (e.g. Higgs production cross sections)

- What is the $m_{\gamma\gamma}$ shape of signal and background?
- If N VBF events are generated, how many do you expect to select in each category $\rightarrow$ efficiency estimation (from MC)

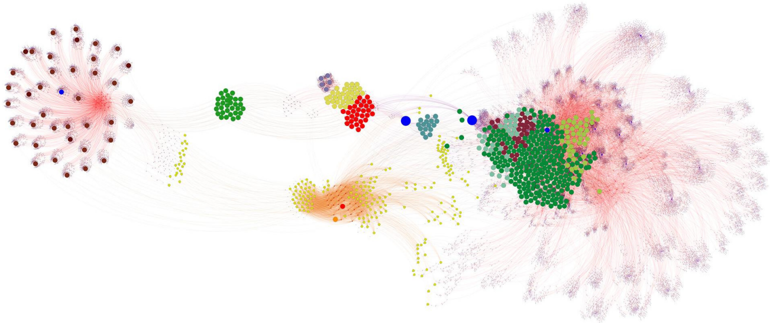


Modeling of the background

- The background shape is a sort of quick decreasing function. Since the signal is very small, we need to model the background very precisely otherwise we are introducing a bias in the signal estimation.
- Some analyses prefer to use the shape from the MC (usually the normalization is from data): you need a very good generator and a lot of events
- In $H \to \gamma\gamma$ since the function seems to be quite simple we are assuming it is an analytical function (exponential, ...)
- To check how good our assumption works we try on a background-only MC: in this case if we do a s+b fit we should find $s = 0$. The deviation (spurious signal) is assumed to be our systematic
- This is a limitation of the analysis: the MC has not infinite statistics: with this procedure it is quite likely you overestimate the systematic
- Spurious signal is one of the main limitation of the analysis. Solutions:
  - Brute force: generate more MC events (using some tricks)
  - Inject more information in your MC: assume it is smooth. Smooth with some algorithm to remove high-frequency fluctuations.
  - ...

The statistical analysis

Once you have all the informations (expected cross sections, efficiencies, shape modeling, systematics) lets put everything into a statistical model (RooFit workspace) to make a likelihood



- It's complicated, but this is just the graph of the expression $P(\text{observable}|\text{cross sections})$ to make the likelihood
- Complexity is mosty given by the large number of systematics

The plan

- Discuss frequently the progress inside your analysis meeting (weekly)
- Keep your subgroup (HGam) updated, discuss common/main problems (e.g. samples, background modeling, change in timeline) with them
- Nominate internal editors and document your progress in the supporting note
- When analysis is mature ask for an Editorial Board and ask for a group approval
- Nominate editors of the paper and start writing it, discuss with EB
- Before unblinding (look at the data) you need the approval of your group (Higgs unblinding approval)
- If everything is ok, go for ATLAS circulation of your draft paper

Suggestion

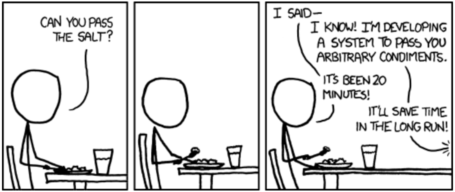
This is a lot of meetings. Make them productive:

- present in a clear way (e.g. never show a plot without axis labels)
- keep your audience interested

Random suggestions for the analysis

You will spend a lot of time coding...

- Code depending on the contex: are you writing a new reconstruction algorithm or are you just filling histogram with different systematics? How long your code will live?
- At analysis level usually you are dominated by the time you need to write the code and understand if it is correct: not by the running time
- You also will spend a lot of time undertanding the code you wrote two months ago (or remembering in which folder you put it): use gitlab, write README, write in-code documentation. This is help you and others.
- At analysis level you have more freedom on the tools you can use: ROOT+RDataFrame / uproot+pandas / jupyter notebook / ...



HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE?
(ACROSS FIVE YEARS)

HOW MUCH TIME YOU SHAVE OFF	HOW OFTEN YOU DO THE TASK					
	50/DAY	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
1 MINUTE	8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
5 MINUTES	9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
30 MINUTES		6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
1 HOUR		10 MONTHS	2 MONTHS	10 DAYS	2 DAYS	5 HOURS
6 HOURS				2 MONTHS	2 WEEKS	1 DAY
1 DAY					8 WEEKS	5 DAYS