Introduction to particle physics: experimental part

- Expected results and toys
  - Pseudo-experimens and Asimov datasets
  - Dealing with non-asymptotic situations
- Profiling
- Look-Elsewhere Effect
- Bayesian method
- Presentation of results

Slides extracted from N. Berger lectures at CERN Summer School 2019

Prof. dr hab. Elżbieta Richter-Wąs

Generating Pseudo-data

Model describes the distribution of the observable: **P(data; parameters)**

⇒ Possible outcomes of the experiment, for given parameter values

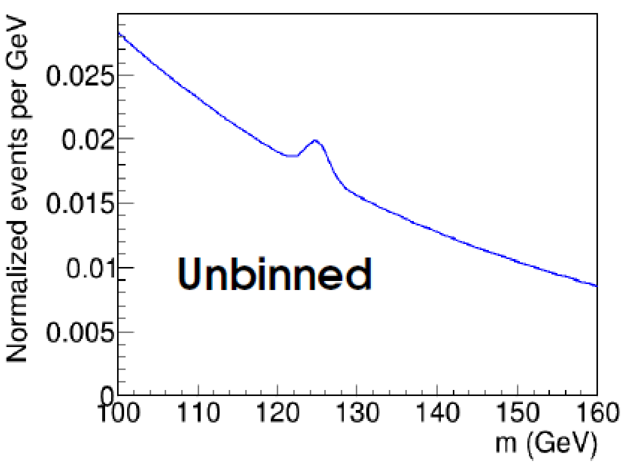
Can draw random events according to PDF : **generate *pseudo-data***

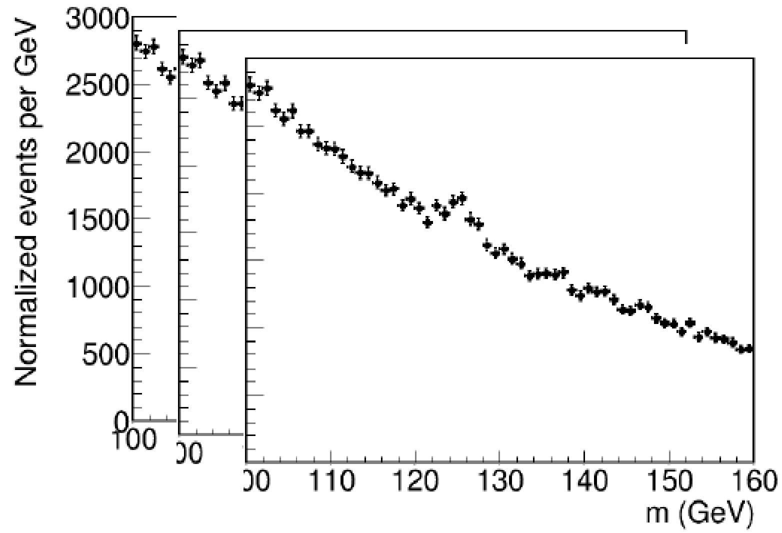
$$P(\lambda=5)$$

2, 5, 3, 7, 4, 9, ….

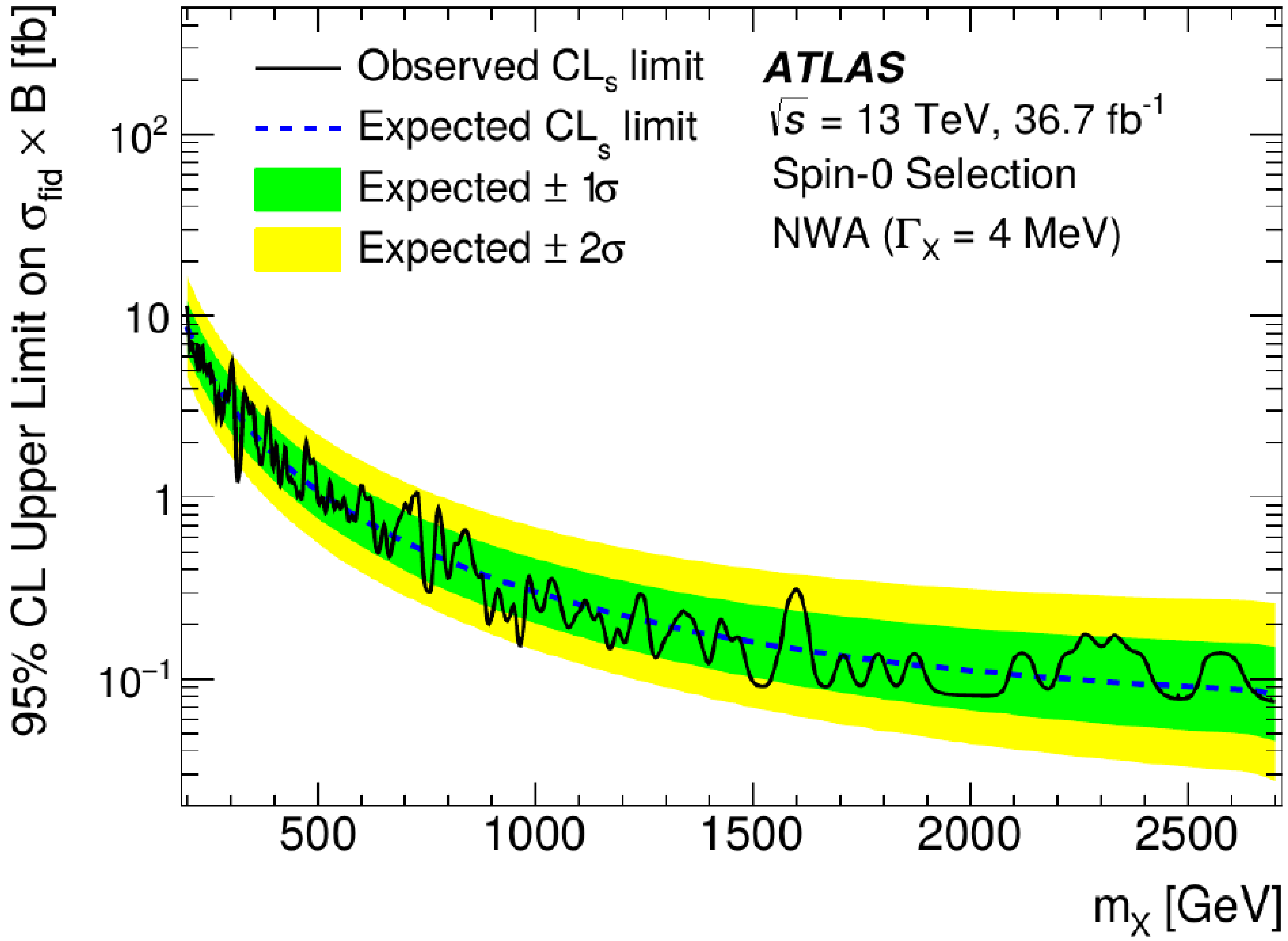
Each entry = separate "experiment"

Generate





Expected results



Expected limits: Toys

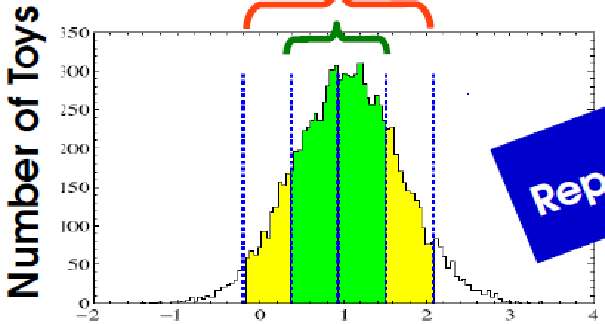
***Expected* results**: median outcome under a given hypothesis
→ usually B-only for searches, but other choices possible.

Two main ways to compute:

→ Pseudo-experiments (*toys*):

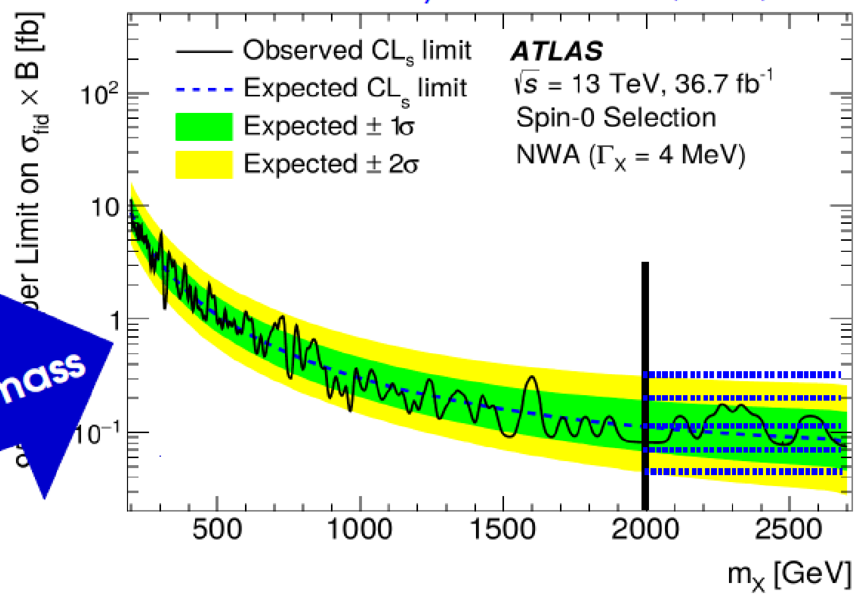
- Generate a pseudo-dataset in B-only hypothesis
- Compute limit
- Repeat and histogram the results
- Central value = median, bands based on quantiles

68% of toys **95% of toys**



Eur.Phys.J.C71:1554,2011 **Computed limit**

Phys. Lett. B 775 (2017) 105



Expected limits: Asimov Datasets

***Expected* results**: median outcome under a given hypothesis
→ usually B-only by convention, but other choices possible.

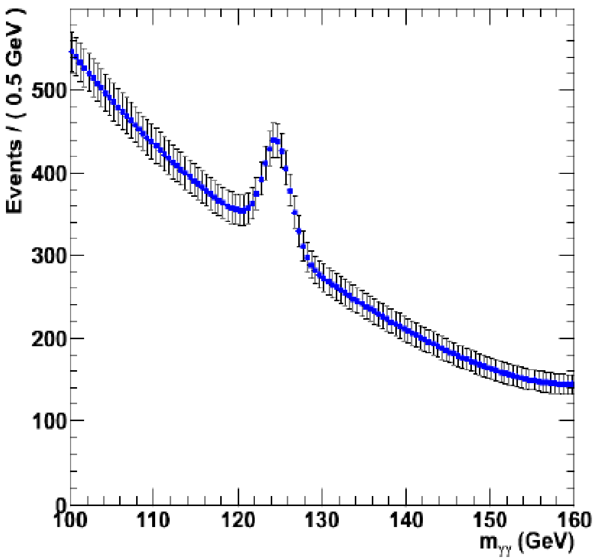
Two main ways to compute:

Strictly speaking, Asimov dataset if
$\hat{X} = X_0$ for all parameters X,
where X_0 is the generation value

→ Asimov Datasets

- Generate a "perfect dataset" – *e.g.* for binned data, set bin contents carefully, no fluctuations.
- Gives the median result immediately:
  median(toy results) ↔ result(median dataset)
- Get bands from asymptotic formulas:
  Band width

$$\sigma^2_{S_0,A} = \frac{S_0^2}{q_{S_0}(\mathrm{Asimov})}$$



⊕ **Much faster (1 "toy")**

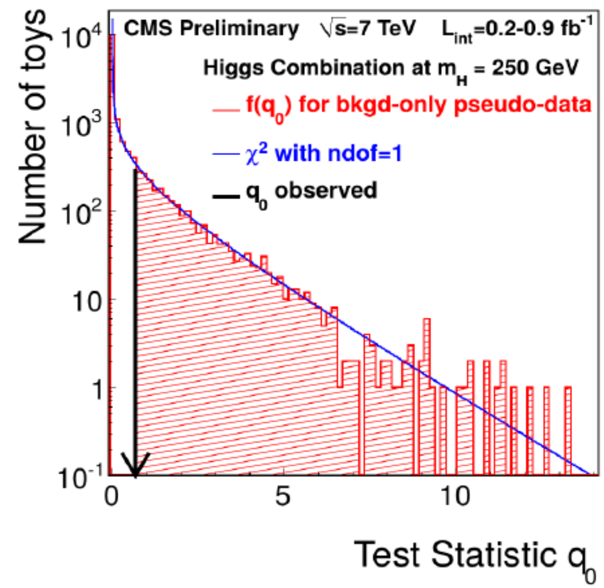
⊖ **Relies on Gaussian approximation**

9

Beyond Asymptotics: Toys

CMS-PAS-HIG-11-022

Asymptotics usually work well, but break down in some cases – e.g. **small event counts.**



Solution: generate ***pseudo data*** (**toys**) using the PDF, under the tested hypothesis

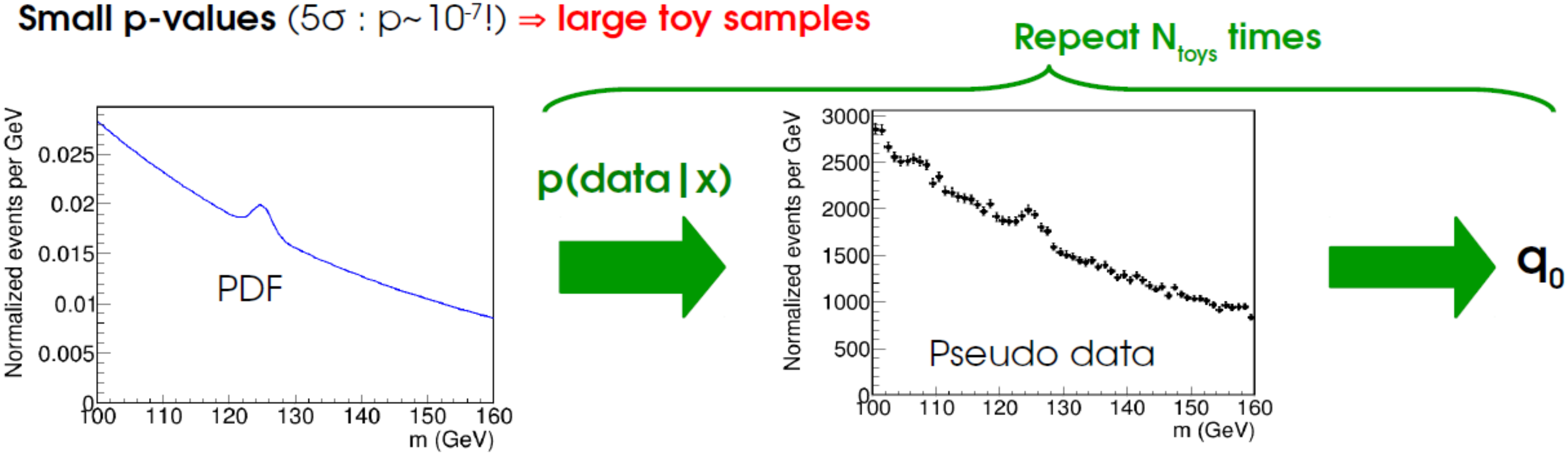
→ Also randomize the observable (θ^{obs}) of each auxiliary experiment: $G(\theta^{obs};\theta,\sigma_{syst})$

→ Samples the true distribution of the PLR

⇒ Integrate above observed PLR to get the p-value

→ Precision limited by number of generated toys,

Small p-values (5σ : p~10^{-7}!) ⇒ **large toy samples**

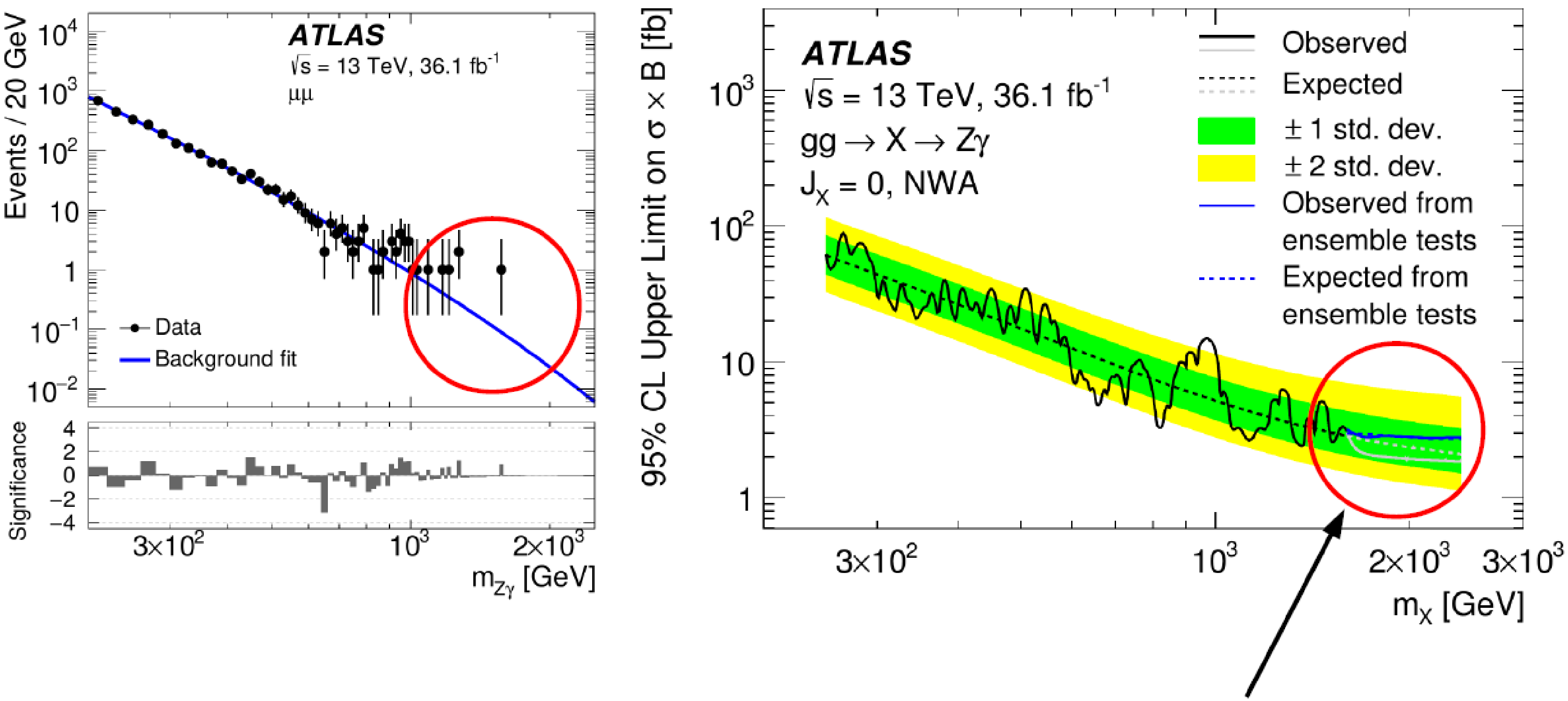


Toys: Example

ATLAS X→Zγ Search: covers 200 GeV < m_X < 2.5 TeV

JHEP 10 (2017) 112

→ for m_X > 1.6 TeV, low event counts ⇒ derive results from toys



Asymptotic results (in gray) give optimistic result compared to toys (**in blue**)

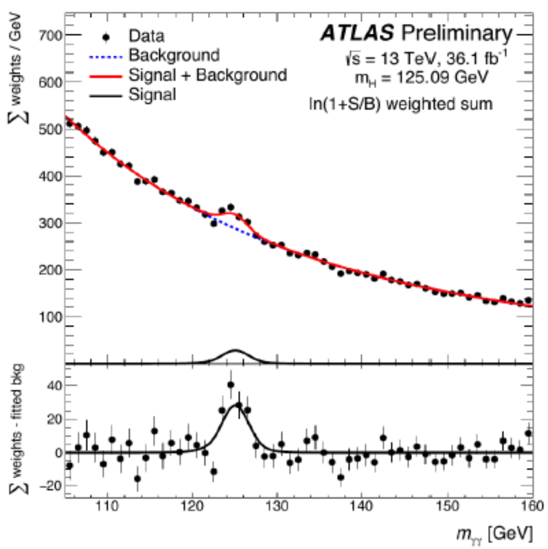
Remarks

Short answer: The high-signal, low-background experiments have been done already (although a surprise would be welcome...)

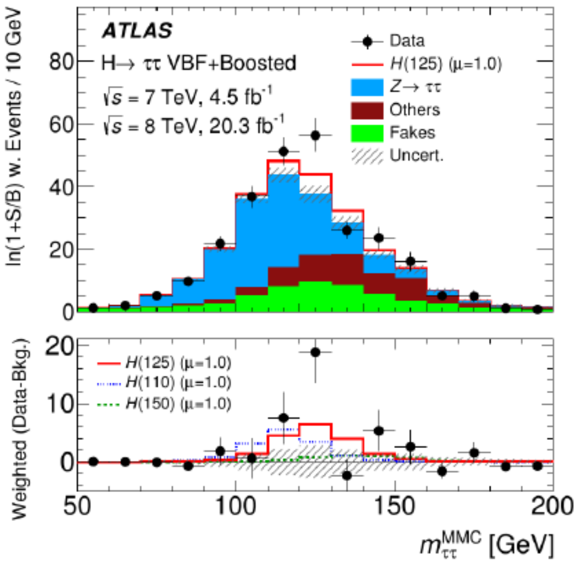
e.g. at LHC:

- **High background levels**, need precise modeling
- **Large systematics,** need to be described accurately
- **Small signals**: need optimal use of available information :
  - **Shape analyses** instead of counting
  - **Categories** to isolated signal-enriched regions

ATLAS-CONF-2017-045



JHEP 12 (2017) 024



Profiling

How to deal with nuisance parameters in likelihood ratios ?

→ **Let the data choose ⇒ use the best-fit values (*Profiling*)**

⇒ **Profile Likelihood Ratio** (PLR)

$$t_{S_0} = -2 \log \frac{L(S=S_0, \hat{\hat{\theta}}(S_0))}{L(\hat{S}, \hat{\theta})}$$

$\hat{\hat{\theta}}(S_0)$ best-fit value for $S=S_0$ (conditional MLE)

$\hat{\theta}$ overall best-fit value (unconditional MLE)

Wilks' Theorem: same properties as plain likelihood ratio

$$f(t_{S_0} \mid S=S_0) = f_{\chi^2(n_{dof}=1)}(t_{S_0})$$

also with NPs present

→ Profiling "builds in" the effect of the NPs

⇒ **Can use t_{S0} to compute limits, significance, etc. in the same way as before**

Systematics implementation

Prototype: NP measured in a separate ***auxiliary* experiment**
e.g. luminosity measurement

→ Build the combined likelihood of the main+auxiliary measurements

$$L(\mu,\theta;\mathbf{data}) = L_{\mathbf{main}}(\mu,\theta;\mathbf{main\ data})\ L_{\mathbf{aux}}(\theta;\mathbf{aux.\ data})$$

Independent measurements: ⇒ just a product

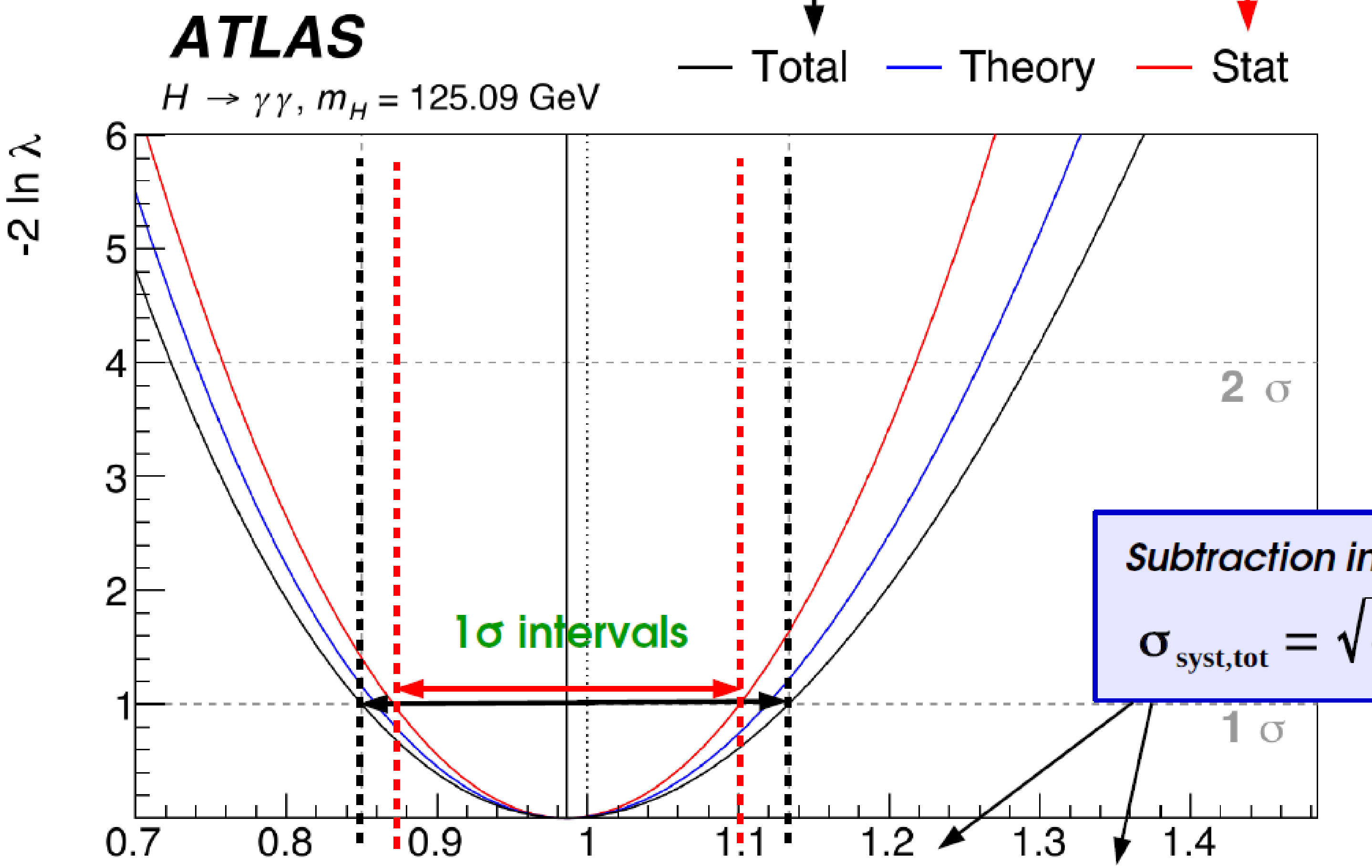
Gaussian form often used by default: $L_{\mathbf{aux}}(\theta;\mathbf{aux.\ data}) = G(\theta^{\mathbf{obs}};\theta,\sigma_{\mathbf{syst}})$

→ Often no clear setup for auxiliary measurements
e.g. theory uncertainties on missing HO terms from scale variations
→ **Implemented in the same way nevertheless** ("pseudo-measurement")

Uncertainty decomposition

All systematics NPs excluded : statistical uncertainty only

All systematics NPs included: stat+syst uncertaintes



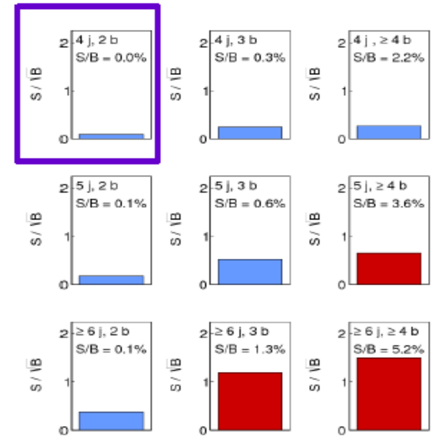
Subtraction in quadrature

$$\sigma_{\text{syst,tot}} = \sqrt{\sigma^2_{\text{total}} - \sigma^2_{\text{stat}}}$$

$$\mu = 0.99 \pm 0.12\,(\text{stat}) \pm 0.06\,(\text{syst}) \pm 0.06\,(\text{theo})$$

Profiling example

$t\bar{t}H \to b\bar{b}$

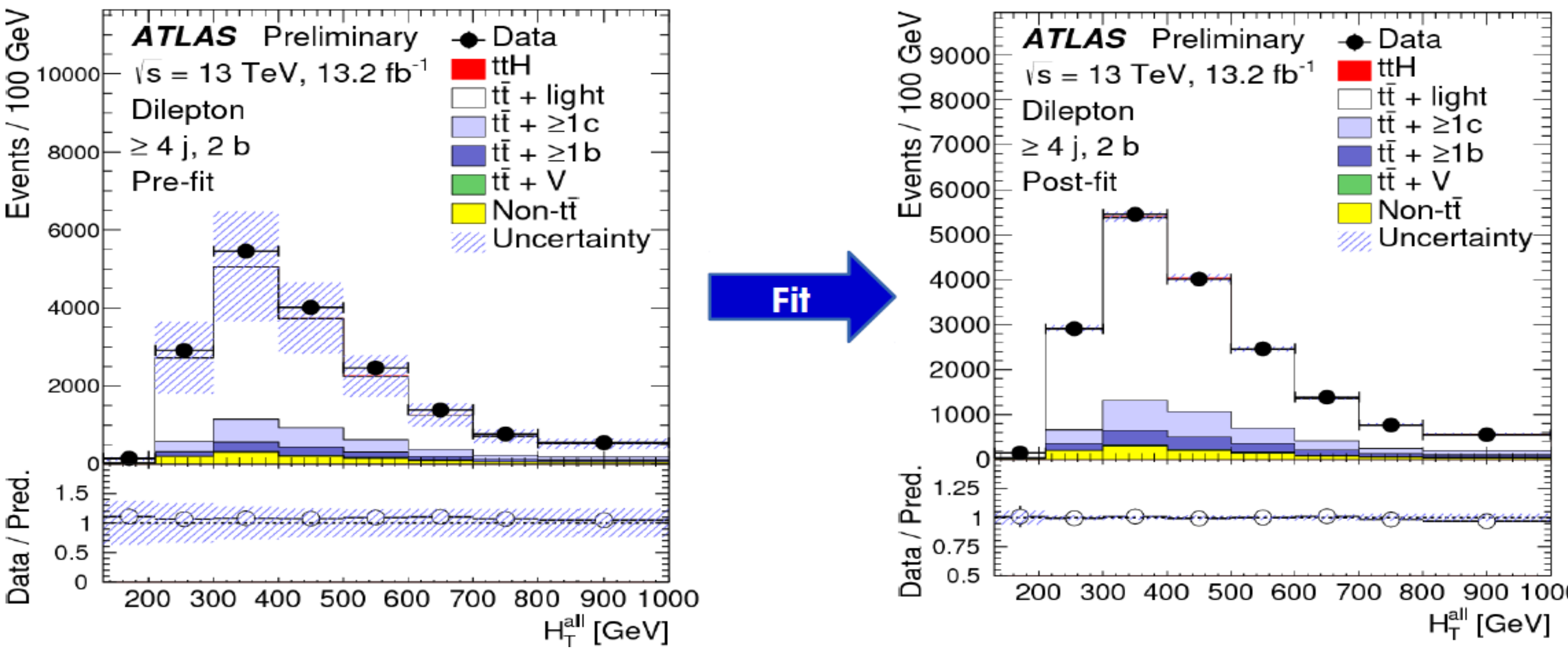


Analysis uses low-S/B categories to constrain backgrounds.

→ **Reduction in large uncertainties on $t\bar{t}$ bkg**

→ **Propagates to the high-S/B categories** through the statistical modeling

⇒ **Care needed in the propagation** (e.g. different kinematic regimes)



ATLAS-CONF-2016-08

Profiling Takeaways

When testing a hypothesis, use the best-fit values of the nuisance parameters: ***Profile Likelihood Ratio***.

$$\frac{L(\mu=\mu_0, \hat{\hat{\theta}}_{\mu_0})}{L(\hat{\mu}, \hat{\theta})}$$

Allows to include systematics as uncertainties on nuisance parameters.

Profiling systematics includes their effect into the total uncertainty.

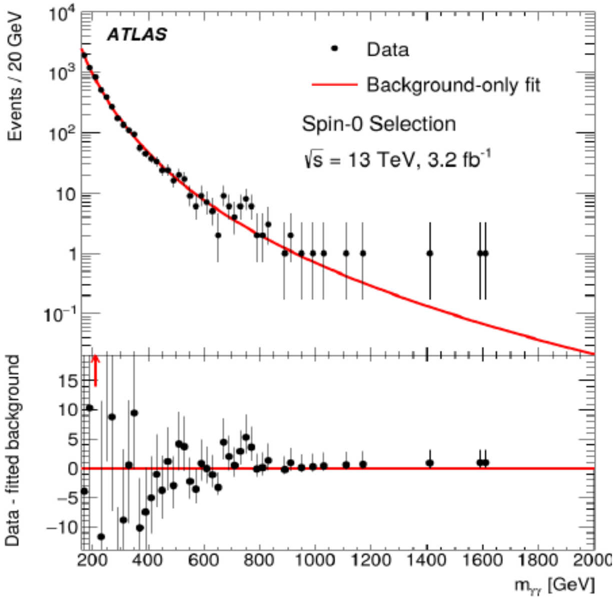
Gaussian:

$$\sigma_{total} = \sqrt{\sigma_{stat}^2 + \sigma_{syst}^2}$$

Guaranteed to work well as long as everything is Gaussian, but typically also robust against non-Gaussian behavior.

Profiling can have unintended effects – need to carefully check behavior

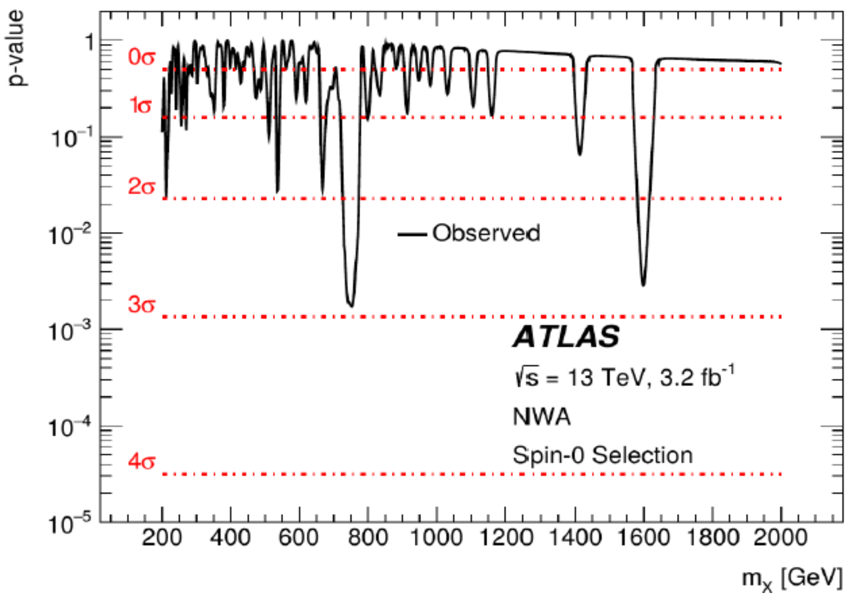
Look-Elsewhere effect



Sometimes, unknown parameters in signal model e.g. p-values as a function of m_X

⇒ Effectively: **multiple, simultaneous searches**

→ If e.g. small resolution and large scan range, **many independent experiments**



→ More likely to find an excess **anywhere in the range**, rather than in a **predefined** location

⇒ ***Look-elsewhere effect*** (LEE)

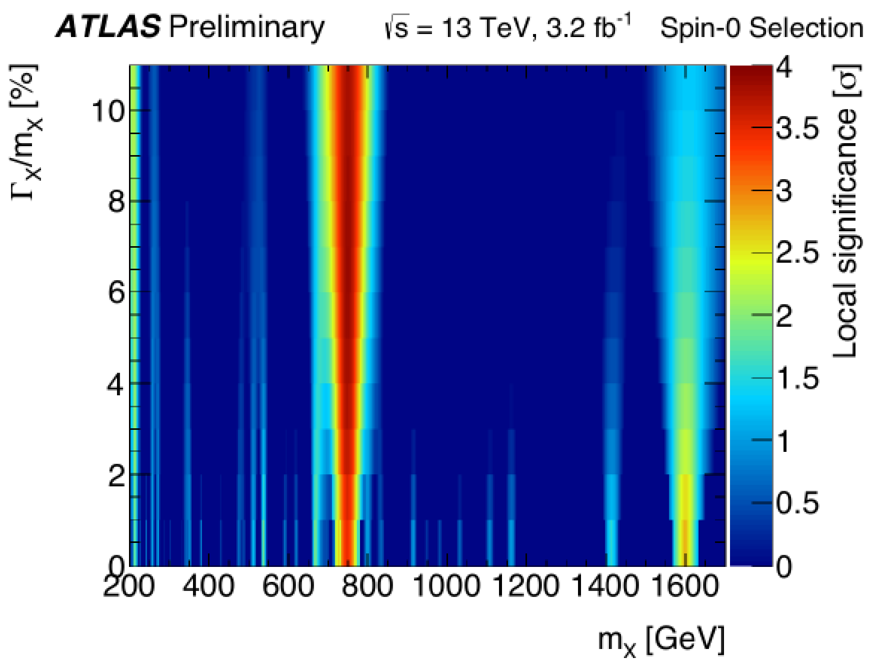
Global Significance

Probability for a fluctuation ***anywhere*** in the range → **Global** p-value.
at a given location → **Local** p-value

For searches over a parameter range, **the global p-value is the relevant one**

→ Accounts for the actual search procedure: look for an excess anywhere in the scanned range

→ Depends on the scanned parameter ranges



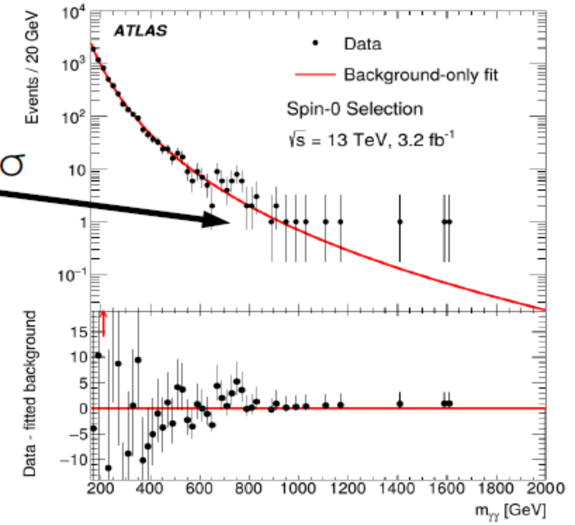
e.g. $X \rightarrow \gamma\gamma$:

- $200 < m_X < 2000$ GeV
- $0 < \Gamma_X < 10\%\ m_X$.

→ p_{local} is what comes out of the usual formulas

How to compute p_{global} (or N_{trials}) ?

Global Significance from Toys



Local 3.9σ

Principle: repeat the analysis in toy data:

→ generate pseudo-dataset

→ perform the search, scanning over parameters as in the data

→ report the largest significance found

→ repeat many times

⇒ The frequency at which a given Z_0 is found ***is*** the global p-value

e.g. **X→γγ Search**: **Z_{local} = 3.9σ** (⇒ p_{local} ~ 5 10^{-5}),

→ However we are scanning 200 < m_X < 2000 GeV and 0 < Γ_X < 10% m_X !

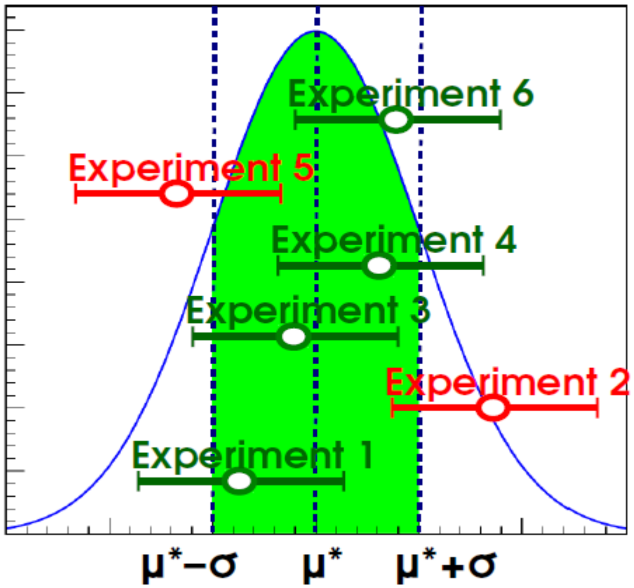
→ Toys : find such an excess **2%** of the time somewhere in the range

⇒ p_{global} ~ 2 10^{-2}, **Z_{global} = 2.1σ** Less exciting, and better indication of true Z!

⊕ **Exact treatment**

⊖ **CPU-intensive** especially for large Z (need ~O(100)/p_{global} toys)

Frequentist vs. Bayesian



All methods described so far are **frequentist**

- Measurement outcomes are random
- Parameters value are **fixed but unknown**

Must be careful about meaning:

→ **"5σ Higgs discovery"**

- → if there is really no Higgs, such fluctuations are observed in only one in 3 million experiments : **P(data | no Higgs) is small**

This is not the crucial question! What we would really like to know is
What is the probability that the excess we see is a fluctuation

→ we want **P(no Higgs |data)** – but all we have is **P(data | no Higgs)**
However **P(no Higgs |data)** is not well-defined in the frequentist framework

Frequentist vs. Bayesian

Can use **Bayes' theorem** to address this:

$$P(\text{no Higgs}|\text{data}) = \frac{P(\text{data}|\text{no Higgs})}{P(\text{data})} P(\text{no Higgs})$$

same as in the frequentist formalism (=likelihood)

Prior Probability

irrelevant normalization factor

Can compute P(no Higgs | data), **if we provide P(no Higgs)**

→ An hypothesis ("no Higgs") is now considered something random

- Is the presence of the Higgs in a experiment randomly chosen ?
- In fact, different definition of p: ***degree of belief***, not from frequencies.
- P(no Higgs) ***Prior degree of belief*** – critical ingredient in the computation

Compared to frequentist PLR:

⊕ **answers the "right" question**

⊖ **answer depends on the prior**

⊕ **In practice, frequentist and Bayesian methods usually give similar results**

"Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentist use impeccable logic to deal with an issue that is of no interest to anyone." - Louis Lyons

Bayesian methods

Probability distribution (= likelihood) :

→ Same as frequentist case, but treat systematics by **integrating over priors**, instead of profiling:

→ Integrate out θ to get P(μ) : $$P(\mu) = \int P(\mu, \theta) C(\theta)\, d\theta$$

→ Use probability distribution P(μ) directly for limits & intervals

e.g. define 68% CL ("Credibility Level") interval (A, B) by: $$\int_A^B P(\mu)\, d\mu = 68\,\%$$

⊖ No simple way to test for discovery

⊖ Integration over NPs can be CPU-intensive (but can use MCMC methods)

Priors : most analyses use flat priors in the analysis variable(s)

⇒ **Parameterization-dependent**: if flat in σ×B , them not flat in couplings….

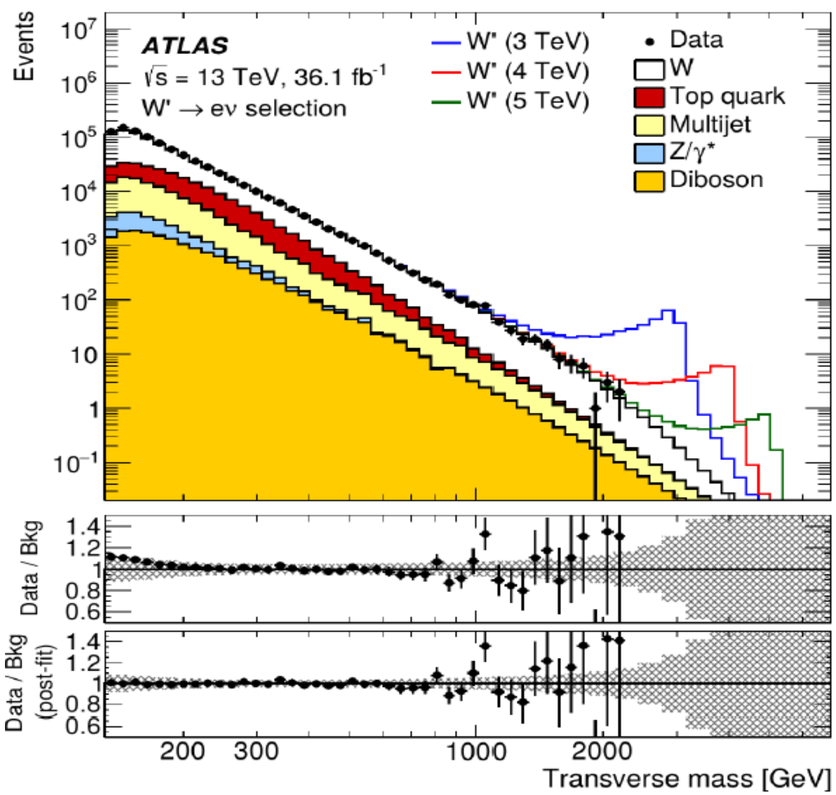
→ Can use the Jeffreys' or reference priors, but difficult in practice

Bayesian methods

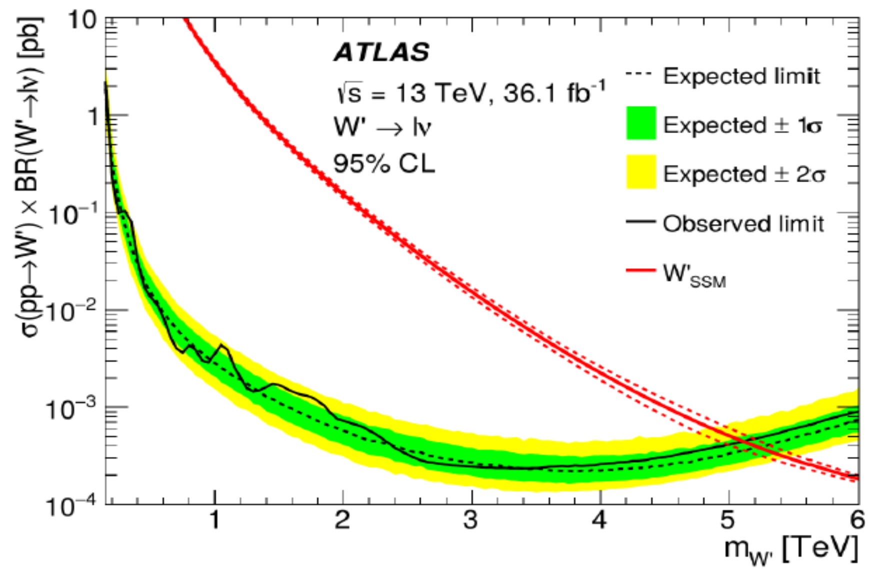
Example: W'→lν Search

arXiv:1706.04786

- **POI**: **W' σ×B** → use **flat prior over** [0, +∞[.
- **NPs**: syst on **signal ε** (6 NPs), **bkg** (6), **lumi** (1) → integrate over Gaussian priors



Trigger	Multijet background
Lepton reconstruction and identification	Top extrapolation
Lepton momentum scale and resolution	Diboson extrapolation
E_T^{miss} resolution and scale	PDF choice for DY
Jet energy resolution	PDF variation for DY
Pile-up	EW corrections for DY
	Luminosity



48

Why 5σ?

One-sided discovery: **5σ ⇔ p_0 = 3 10^{-7} ⇔ 1 chance in 3.5M**

→ Overly conservative ?

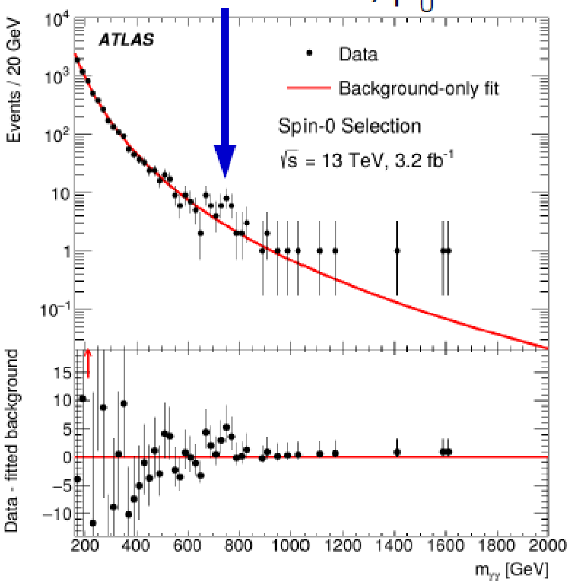
→ Do we even control such small probabilities ?

Reasons for sticking with 5σ (from Louis Lyons):

- **LEE** : searches typically cover multiple independent regions
  ⇒ Global p-value is the relevant one
  N_{trials} ~ 1000 : local 5σ ⇔ O(10^{-4}) more reasonable
- **Mismodeled systematics**: factor 2 error in syst-dominated analysis ⇒ factor 2 error on Z…
- **History**: 3σ and 4σ excesses do occur regularly, for the reasons above
- **"Subconscious Bayes Factor"** : p-value should be at least as small as the subjective p(S):

Local 3.9σ, p_0 = 5E-5

Global 2.1σ, p_0 = 2E-2



Extraordinary claims require extraodinary evidence

⇒ Stay with 5σ…

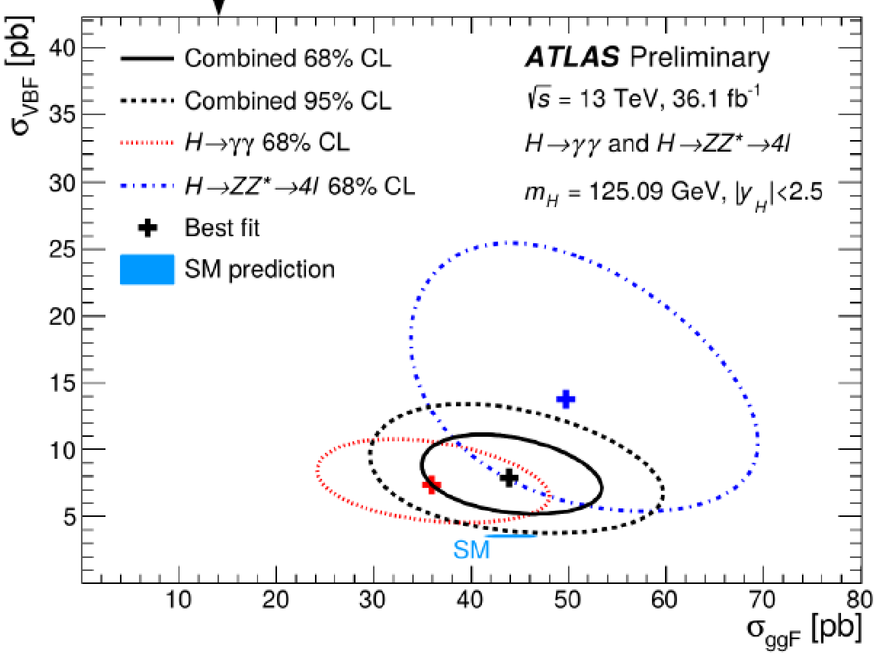
Reparametrisation

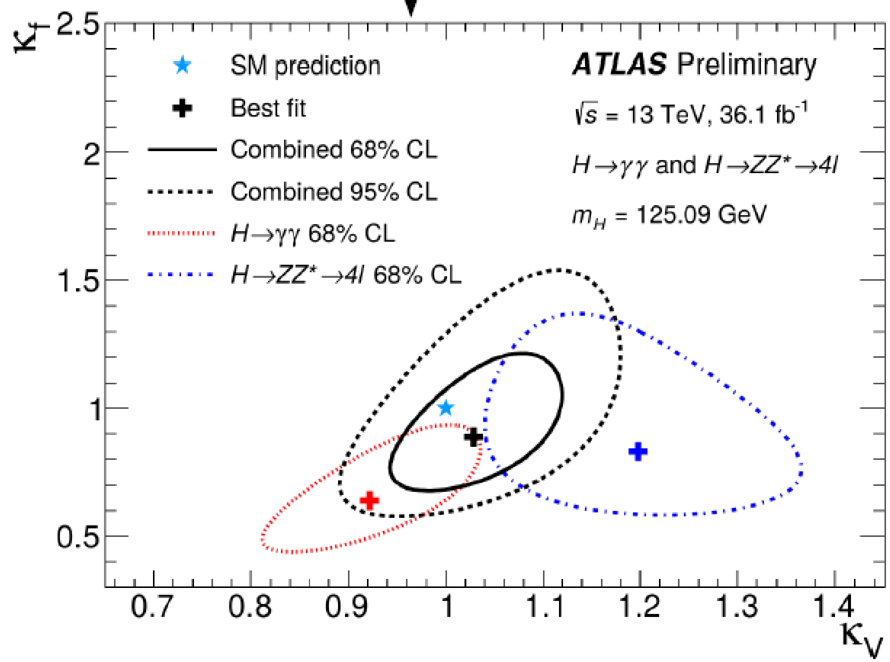
Start with basic measurement in terms of e.g. **σ×B**

→ How to measure derived quantities (couplings, parameters in some theory model, etc.) ? **→ just reparameterize the likelihood:**

e.g. Higgs couplings: σ_{ggF}, σ_{VBF} sensitive to Higgs coupling modifiers κ_V, κ_F.

$$L(\sigma_{ggF}, \sigma_{VBF}) \xrightarrow[\sigma_{VBF} \to \sigma_{VBF}(\kappa_V, \kappa_F)]{\sigma_{ggF} \to \sigma_{ggF}(\kappa_V, \kappa_F)} L(\sigma_{ggF}(\kappa_V, \kappa_F), \sigma_{VBF}(\kappa_V, \kappa_F)) \equiv L'(\kappa_V, \kappa_F)$$

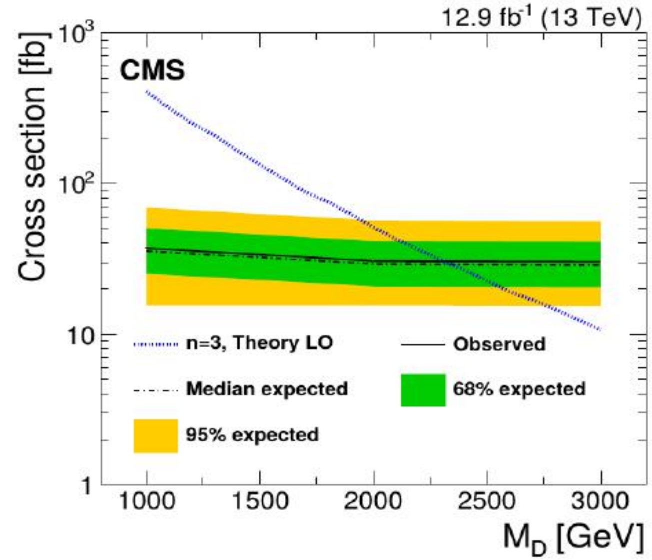


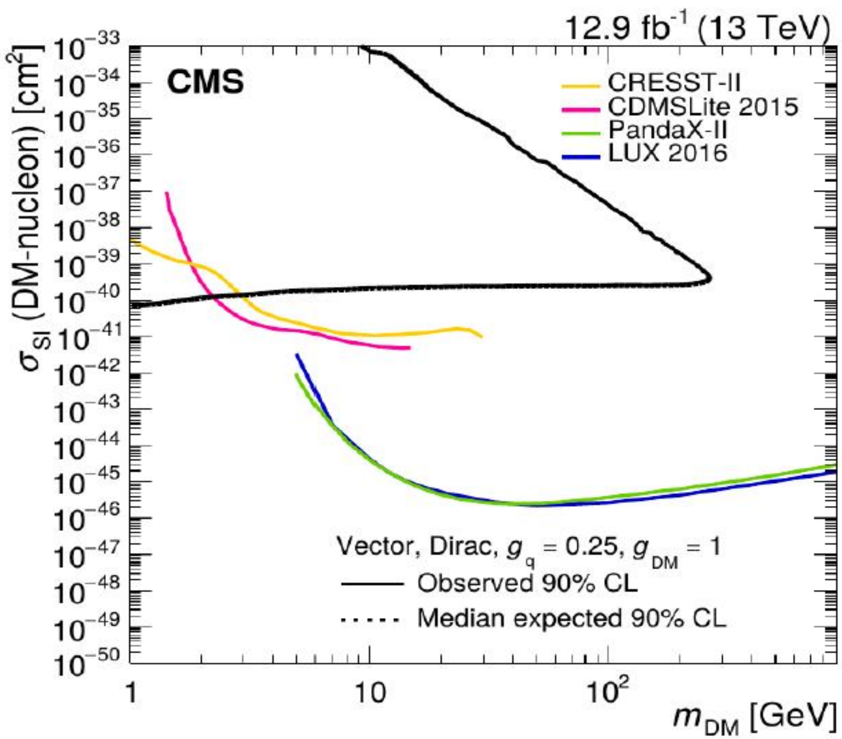


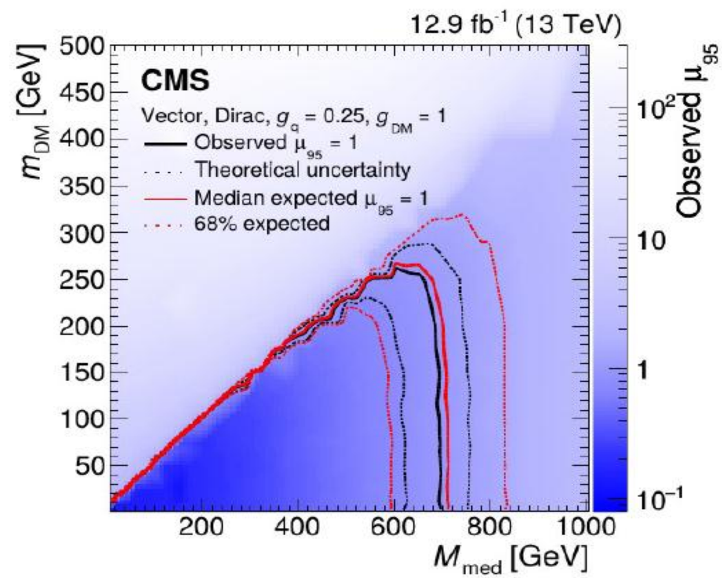
Reparametrisation: Limits

Reparameterization: Limits

CMS Run 2 Monophoton Search: measured **N_s** in a counting experiment reparameterized according to various DM models





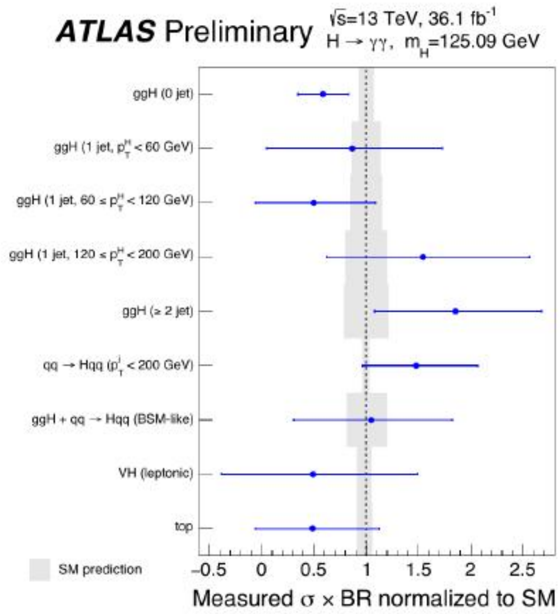


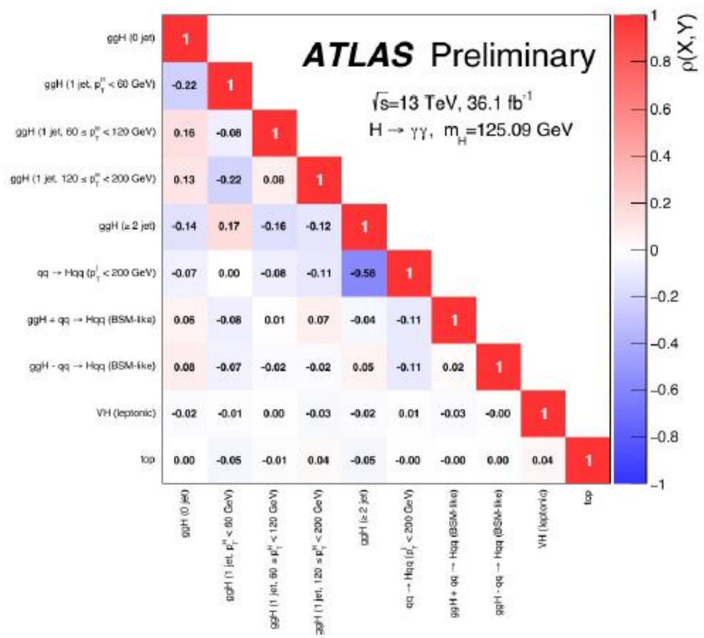
Presentation of results

→ Cannot test every model : need to make enough information public so that others (theorists) are able to do it independently

⇒ **Gaussian case**: sufficient to provide measurements + covariance matrix

→ For example using the HEPData repository.





Non-Gaussian case: no simple method

Conclusions

- Significant evolution in the statistical methods used in HEP
- Variety of methods, adapted to various situations and target results
- Allow to
  - model the statistical process with high precision in difficult situations (large systematics, small signals)
  - make optimal use of available information
- Implemented in standard RooFit/RooStat toolkits within the ROOT framework, as well as other tools (BAT)
- Still many open questions and areas that could use improvement
  → e.g. how to present results with all available information