Science & Technology Facilities Council
Rutherford Appleton Laboratory

# Hands-on Statistics

Tim Adye
Rutherford Appleton Laboratory

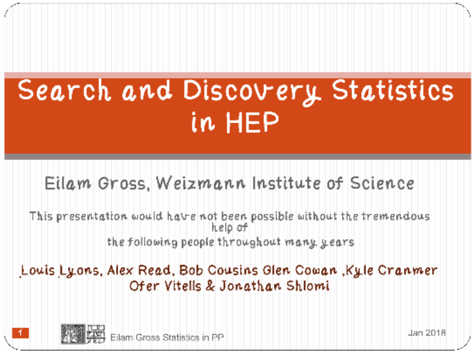
PPD Advanced Graduate Lectures
11th May 2020

# Introduction

- This is not a statistics lecture!
  - Instead, I hope to give some hints on how to use some of the statistical methods you learned from the real expert lecturers.
    - I apologise in advance for sloppy language, unintroduced terms, and complete lack of proofs

- For a thorough introduction, I recommend:

1. CERN Academic Training Lecture series, which has had 3–4 hour lectures by different HEP statistics experts every couple of years. I have found particularly helpful:
   1. Eilam Gross in 2018
   2. Glen Cowan in 2012
   3. Kyle Cranmer in 2011



Search and Discovery Statistics in HEP

Eilam Gross, Weizmann Institute of Science

This presentation would have not been possible without the tremendous help of the following people throughout many years

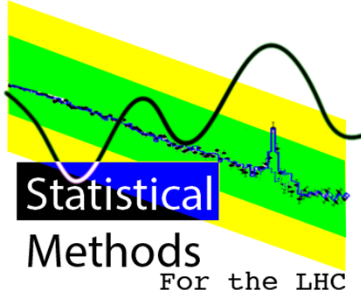
Louis Lyons, Alex Read, Bob Cousins Glen Cowan ,Kyle Cranmer Ofer Vitells & Jonathan Shlomi

1

Eilam Gross Statistics in PP

Jan 2018

2. "Statistics Methods for the LHC" – online documentation from ATLAS, with RooFit / RooStats / RooUnfold code examples.



# Introduction

- There are alternative concepts and methods that we could use, but I will only discuss techniques most common in Particle Physics today, notably:
  1. Frequentist statistics
     - Bayesian used in most other fields
  2. profile likelihood ratio developed for the LHC
     - You are probably already familiar with the other common method, least squares ($\chi^2$) fit
  3. CLs limits
     - not widely accepted outside our field

- Due to lack of time or personal experience, I will not discuss
  1. combination of results (BLUE etc)
     - I will mention Likelihood combination
  2. goodness-of-fit ($\chi^2$, KS tests, etc)
  3. ... or any techniques used on event data, before the final statistical interpretation
     - multivariate discrimination, machine learning, sPlots, etc.

# Lecture plan

- Building a model
  1. PDF ⊗ data → Likelihood
  2. Asimov dataset

- Some typical types of statistical analysis
  - Testing a model, with an example from LHC Run1 Higgs measurements, demonstrates all three stages:
    3. Measurement
    4. Discovery
    5. Exclusion
  - ~~Presenting results without a model:~~
    ~~6. Unfolding~~ (but see RooUnfold, very old slides)

- Summary

# Model building

# PDF, dataset, and likelihood

- All the statistical tests we will be considering are based on the likelihood

$$L(\boldsymbol{\alpha}) = \prod_c \prod_i P_c(x_i|\boldsymbol{\alpha}) \cdot \prod_j C_j(\theta_j|\alpha_j)$$

1. $L(\boldsymbol{\alpha})$ is a function of various parameters ($\boldsymbol{\alpha}$), some of which we wish to determine
2. $P_c(x_i|\boldsymbol{\alpha})$ is the probability density function (PDF) for channel $c$, evaluated for each member of the dataset, $x_i$
   - The use (or not) of the parameters, $\boldsymbol{\alpha}$, in the different channels determines how they are constrained by the data
   - eg. for binned data in histogram $h$, with bins, $i$, $P_h(n_i|\boldsymbol{\alpha}) = \mathrm{Poisson}(n_i|\nu_i(\boldsymbol{\alpha}))$
3. $C_j(\theta_j|\alpha_j)$ are additional PDFs that do not depend on the data
   - eg. constraint terms for systematic uncertainties, $C_j(\theta_j|\alpha_j) = \mathrm{Gaussian}(\mu_j|\alpha_j, \sigma_j)$

- Bear in mind:
  - PDFs ($P_c(x)$ and $C_j(\theta)$) must be normalised to 1, or a constant independent of $\boldsymbol{\alpha}$
    - The likelihood, on the other hand, is not normalised
  - The absolute value of the likelihood ($L(\boldsymbol{\alpha})$) is irrelevant, only changes WRT $\boldsymbol{\alpha}$
  - It is usually used as $-\ln L$, or more commonly, $-2\ln L$
    - maximum likelihood is at minimum of $-2\ln L$
    - in the Asymptotic limit, $-2\ln L$ is distributed like a $\chi^2$

$$-2\ln L(\boldsymbol{\alpha}) = \sum_c \sum_i -2\ln P_c(x_i|\boldsymbol{\alpha}) + \sum_j -2\ln C_j(\theta_j|\alpha_j)$$

# RooFit

- RooFit is a tool for creating models
  - RooAbsPdf: base class for PDFs. Will often be constructed from many PDF types.
    - eg. RooGaussian, RooProdPdf, RooSimultaneous
    - these are functions of each other,
      and of RooRealVar parameters that can be mapped to fit parameters
    - can be constructed directly from C++ or Python, or via a "factory" from a specification
      - eg. SUM::model (f*RooGaussian::g(x,m[0],1), RooChebychev::c(x,{a0[0.1],a1[0.2],a2[-0.3]}))
  - RooAbsData: abstract dataset type. Can hold binned and/or unbinned data
  - RooStats::ModelConfig: holds configuration information for a single model
    - PDF, POIs, NPs, observables, etc
  - RooWorkspace: container for PDFs, datasets, and ModelConfigs
    - This can be saved to a workspace.root file to allow separate statistical analysis
    - everything needed should be stored here, allowing sharing, combining, archiving

- RooFit also provides fitting and basic statistical analysis tools
  - RooNLLVar: $-\ln L$ constructed from PDF and dataset
  - RooMinimizer: uses Minuit to minimise RooNLLVar for specified parameters
- RooStats provides higher-level statistical analysis tools
  - eg. ProfileLikelihoodTestStat, AsymptoticCalculator, FrequentistCalculator, HypoTestInverter

# HistFactory and pyhf

- HistFactory is a tool for creating models of binned data with systematics

$$L(\boldsymbol{\alpha}) = \prod_c \prod_i \mathrm{Poisson}(n_i | \nu_i(\boldsymbol{\alpha})) \cdot \prod_j C_j(\theta_j | \alpha_j)$$

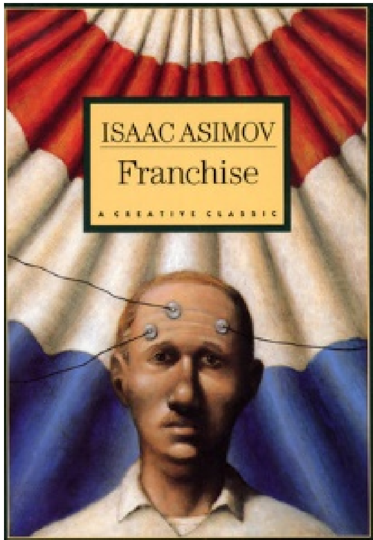
  - Multiple disjoint channels, multiple samples contributing to each with additional (possibly shared) systematics
  - Many analyses can use HistFactory instead of calling RooFit directly.
  - Model specified with XML, which refers to histograms in hist.root files

- pyhf is a reimplementation of HistFactory in pure-Python
  - no dependence on ROOT or RooFit
  - XML+histograms specification replaced by JSON
  - full conversion of models from HistFactory and back
  - reproduces HistFactory results, but much faster
    - tested on ATLAS Run-2 sbottom search, which ran >20 times faster

# Asimov dataset

- An Asimov dataset [1] is generated for a particular set of model parameters such that the maximum likelihood best-fit value of all those parameters are equal to their generated values.
  - ie. maximising $L_A(\mu, \boldsymbol{\theta} | \mu_0, \boldsymbol{\theta}_0)$ will yield $\hat{\mu} = \mu_0, \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$
  - When used in a statistical test, it will return the result expected from that model configuration
    - eg. $p_0$ calculated using Asimov dataset generated with $\mu = 0$ will return the p-value expected from no signal
- Asimov datasets are built as binned datasets, in which the event count in each bin is set to the expected event yield for the chosen model parameters.
  - For unbinned models, a binned distribution is generated with chosen binning fine enough to reproduce all significant features of the model.
  - Note this means the Asimov dataset can look different from data or toy datasets: fractional bin contents or unbinned→binned
- For RooFit models:

```
dataset = RooStats::AsymptoticCalculator::GenerateAsimovData (pdf, observables);
```



[1] Named for SF author, Isaac Asimov, whose 1955 short story, *Franchise*, envisaged the 2008 US Presidential Election decided by one voter representative of the entire electorate. [arXiv:1007.1727]

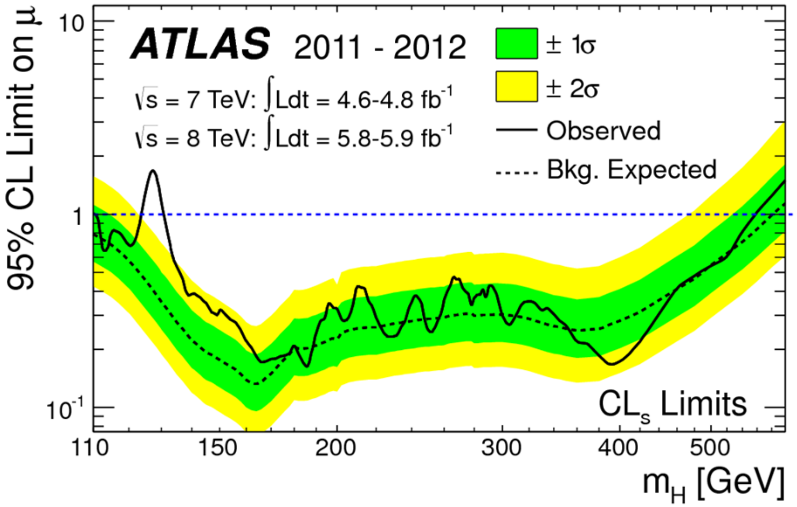
As an Asimov fan of old, this name makes me very happy.

# Statistical tests

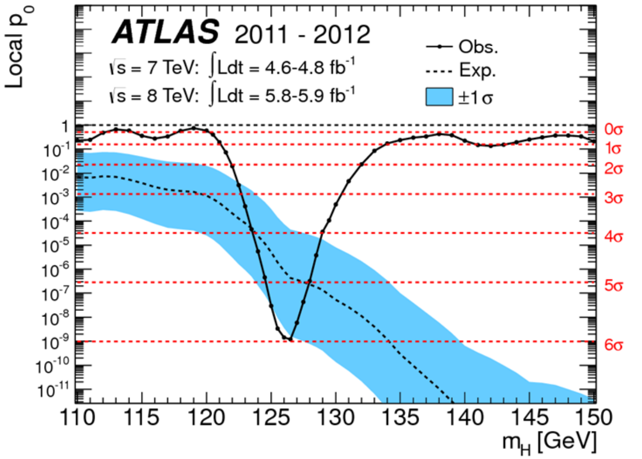
# The three ages of discovery

PLB 716 (2012) 1

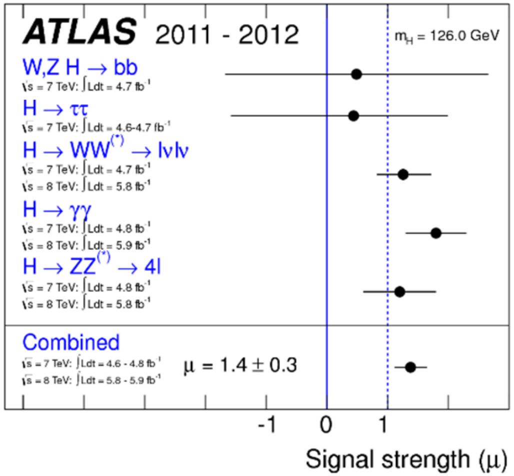
1. **Exclusion**: CLs

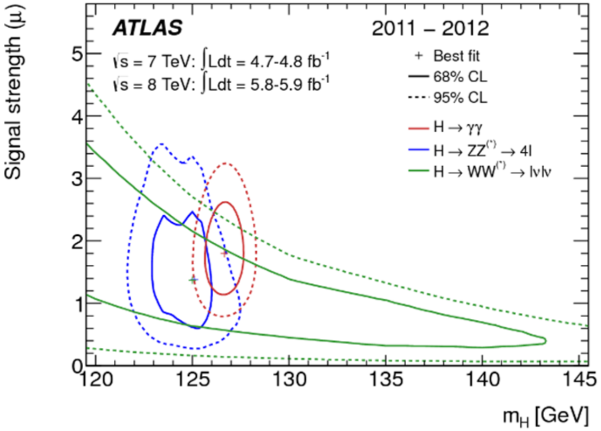


2. **Discovery**: $p_0$



3. **Measurement**: $\hat{\mu} \pm \sigma$
   or more generally,
   confidence intervals





# Measurement

# Measurement: profile likelihood ratio

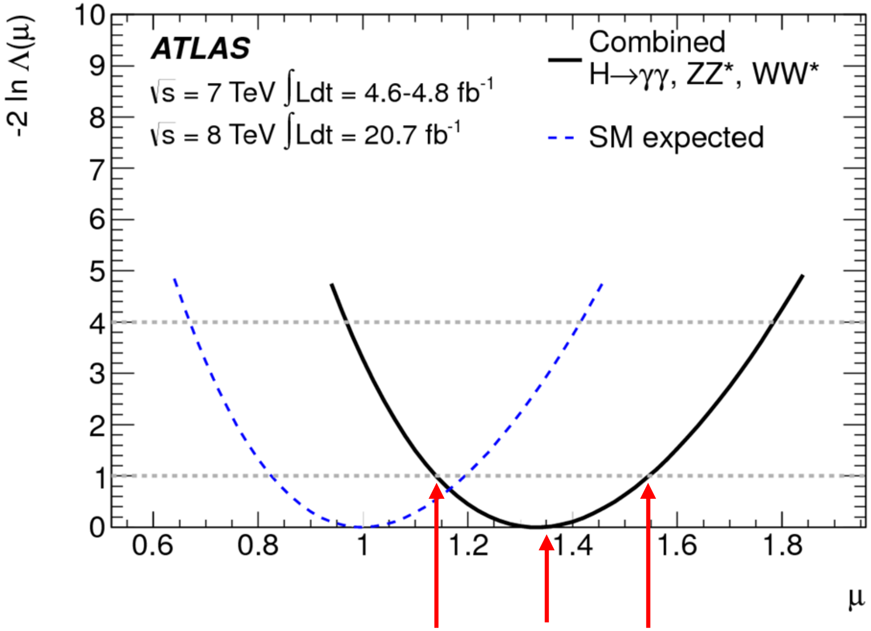
PLB 726 (2013) 88

- The likelihood is a function of our parameters of interest (POI), here $\mu$, and various nuisance parameters (NP), $\boldsymbol{\theta}$: $L(\mu,\boldsymbol{\theta})$.
  - Note that the $\boldsymbol{\theta}$ are often dependent on $\mu$.
- We form the profile likelihood ratio as: $\Lambda(\mu) = \dfrac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})}$

maximise $L(\mu,\boldsymbol{\theta}(\mu))$
for all $\boldsymbol{\theta}(\mu)$
with specified $\mu$

maximise $L(\mu,\boldsymbol{\theta})$
for all $\mu,\boldsymbol{\theta}$

  - $\Lambda(\mu)$ can be evaluated with two fits:
    1. $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are the "best fit" (maximum likelihood estimate, MLE) values of $\mu$ and $\boldsymbol{\theta}$
    2. $\hat{\hat{\boldsymbol{\theta}}}(\mu)$ are the "conditional best fit" values for all the NPs at a given, specified, $\mu$.
- Plot $-2\ln\Lambda(\mu)$ against $\mu$
  - Minimum is at $-2\ln\Lambda(\hat{\mu}) = 0$ (by definition)
  - In the asymptotic limit (large N),
    - this will be distributed like a $\chi^2_1$ distribution
      - or $\chi^2_n$ for $n$ POIs
    - so 68% confidence interval is the range where $-2\ln\Lambda(\mu) < 1$

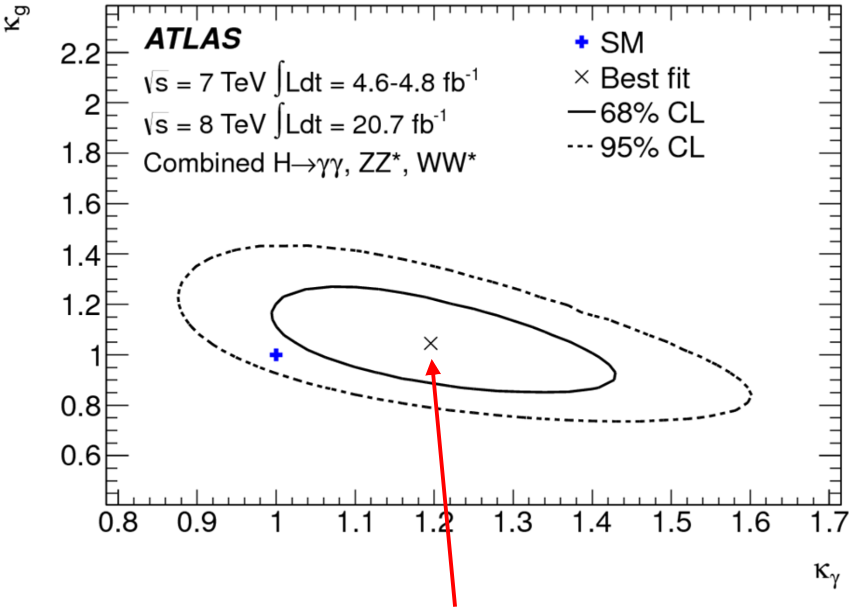


$\mu = 1.33^{+0.21}_{-0.18}$ $-1\sigma$ $\hat{\mu}$ $+1\sigma$

# Measurement: 2D contours

PLB 726 (2013) 88

- For multiple POIs
  - calculate $-2\ln\Lambda(\boldsymbol{\mu})$ for all points on a grid and
  - draw contours for regions $-2\ln\Lambda(\boldsymbol{\mu}) < D^{-1}(\chi^2_n)$,
    - where $D^{-1}(\chi^2_n)$ is the inverse of the cumulative $\chi^2_n$ distribution, for $n$ POIs. [1]
    - 2D contours:
      - $D^{-1}(\chi^2_2(68\%))$ = 2.30
      - $D^{-1}(\chi^2_2(95\%))$ = 6.18



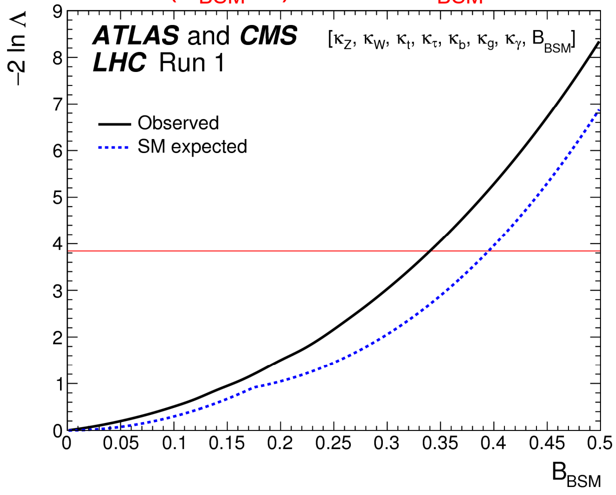
$-2\ln\Lambda(\hat{\kappa}_\gamma, \hat{\kappa}_g) = 0$

[1] $D^{-1}(\chi^2_n(p))$ = ROOT::Math::chisquared_quantile (p, n)

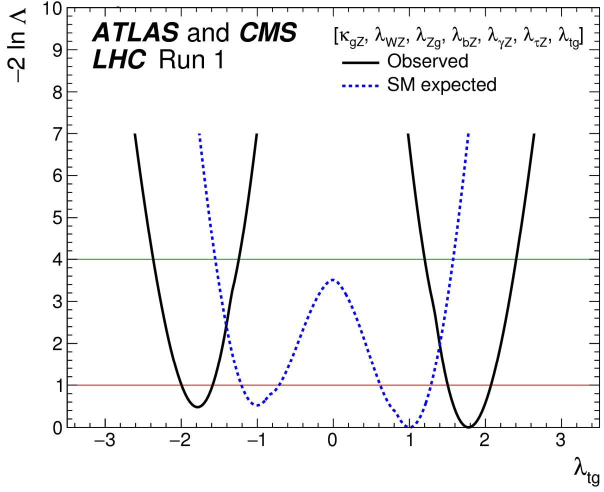
# Measurement: variations on a theme

JHEP 08 (2016) 045

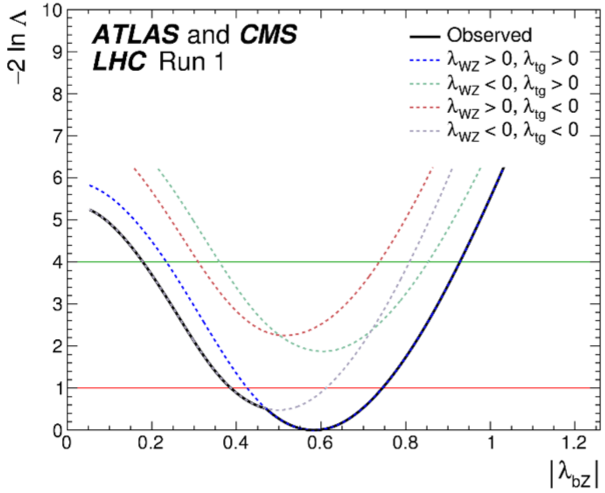
95% confidence interval
with ($B_{BSM}$≥0) bound: $B_{BSM} < 0.34$



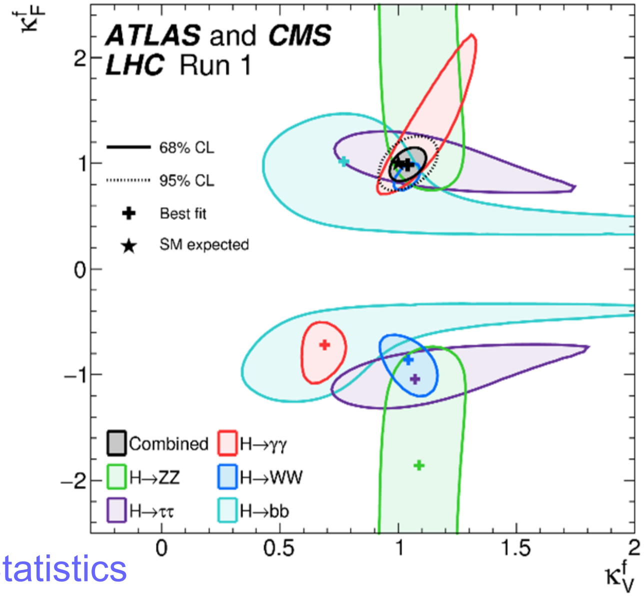
disjoint confidence interval:
$\lambda_{tg} = [-2.00, -1.59] \cup [1.50, 2.07]$



kink due to different sign combinations
of profiled NPs: $|\lambda_{bZ}| = 0.58^{+0.16}_{-0.20}$

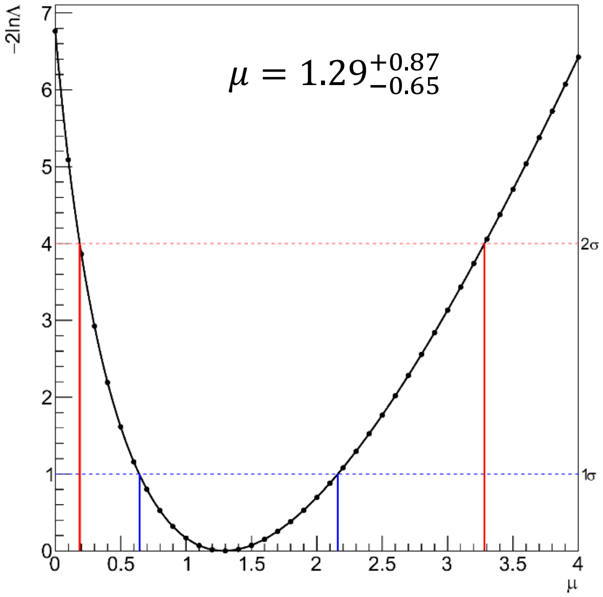


multiple contours for different
channels and their combination



# Measurement: scanning the likelihood curve

- To calculate a single PLR, require two fits:
  - $$-2\ln\Lambda(\mu) = -2\ln\frac{L\left(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu)\right)}{L(\hat{\mu},\hat{\boldsymbol{\theta}})}$$ $$= -2\ln L(\mu,\,\hat{\hat{\boldsymbol{\theta}}}(\mu)) - 2\ln L\left(\hat{\mu},\,\hat{\boldsymbol{\theta}}\right)$$
  - The second term is independent of $\mu$, so only needs to be evaluated once
  - … or not at all, if the minimum can be determined from the curve
    - removes ambiguity from the offset calculated in two ways (unconditional vs conditional fits)
    - should be ~quadratic near minimum, so can use a quadratic interpolation of lowest 3 points



- Can (approximately) cross-check the result with the unconditional fit for $-2\ln L\left(\hat{\mu},\,\hat{\boldsymbol{\theta}}\right)$:
  - $\hat{\mu}$ should agree within the precision of the fit and of the interpolation
  - inverse Hessian at the minimum is the local covariance matrix, so $\sigma_0^2 = H^{-1}(\mu,\mu)$
    - Minuit will calculate (symmetric) errors from the Hessian
      - run with strategy=2, or call Hesse() explicitly.
    - Minuit's Minos() is similar to the curve scan, but without user control or diagnostic plot
  - Example comparison: $\mu = 1.29^{+0.87}_{-0.65}$ (curve) with $\mu = 1.29 \pm 0.73$ (Hessian)

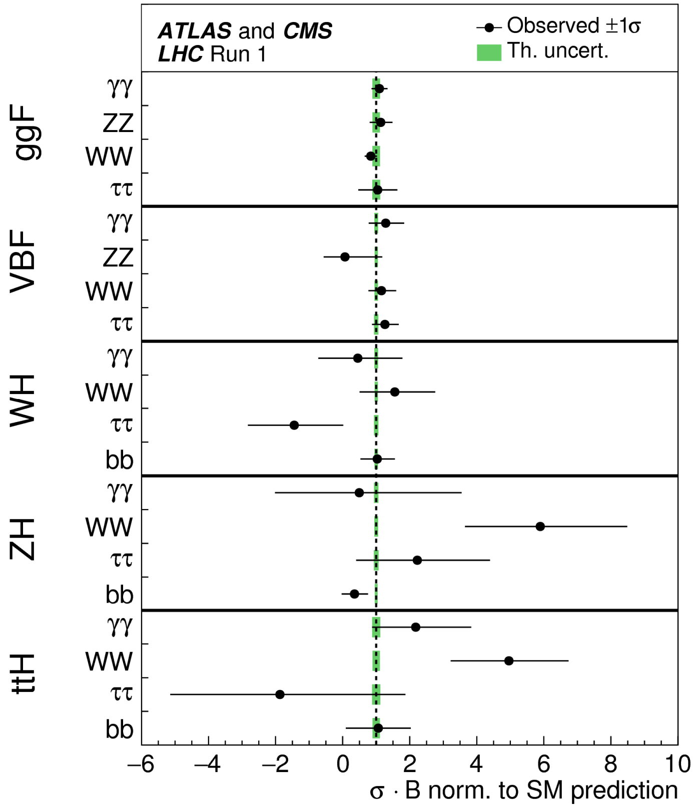
# Measurement: technical issues

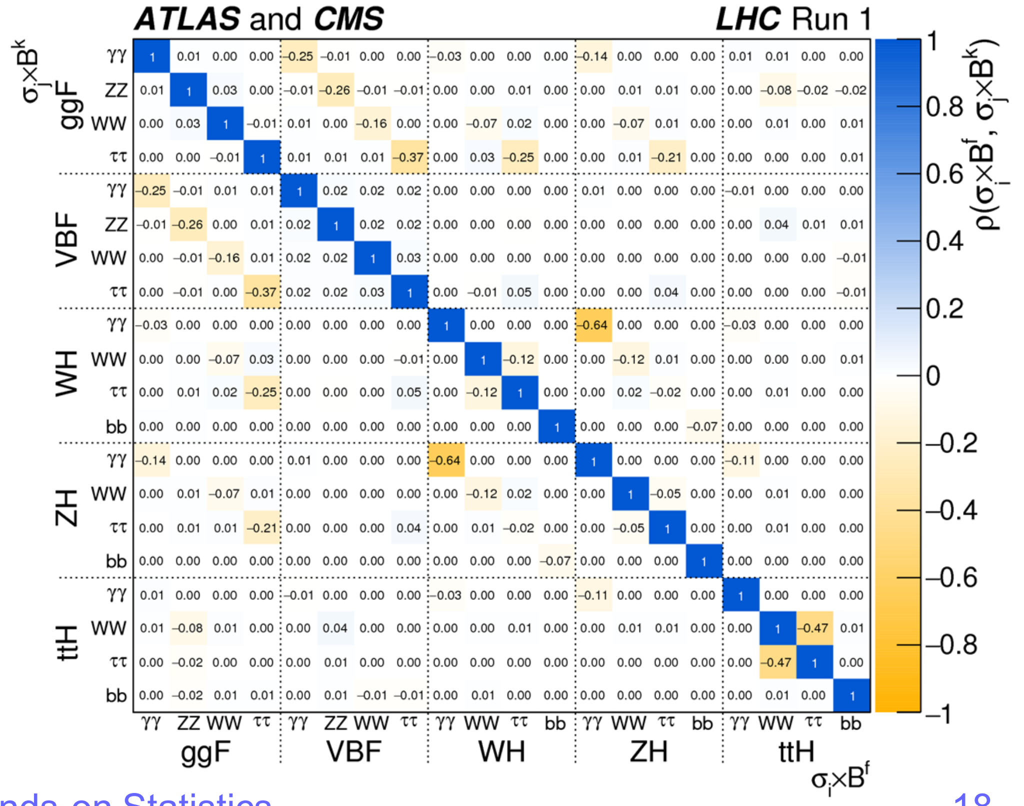
- Sometimes significant CPU requirements
  - Time = (likelihood evaluation time) * (number of evaluations to fit) * (number of fits)
  - Mitigations:
    - Simplify likelihood (faster likelihood evaluation)
    - Reduce or combine number of NPs (simplifies likelihood and fewer fit cycles)
    - Use fewer points in scan and interpolate (quadratic or spline)
      - 2D interpolation is more cumbersome
        - ROOT's TGraph2D can do linear interpolation of contours (use GetContourList() to extract)
    - Run different points in parallel, eg. in batch or on the Grid.
- Fit problems
  1. Fit failures reported by Minuit (or other minimiser)
  2. Bumpy curve, kinks, or bad points
  - Possible causes:
    - Numerical precision in likelihood evaluation
    - Undefined component (eg. –ve log for some parameter values) in likelihood evaluation
    - Minuit tolerance settings
    - NPs hitting their parameter limit
    - Some POIs or NPs don't budge from initial position
    - Minuit can't "tunnel" from secondary minimum

# Measurement: correlation matrix

JHEP 08 (2016) 045

- For ≥ 3 POIs, it is not often practical to show contours
  - requires scanning a large number of points
  - results not easy to visualise
- Another option is to provide the correlation matrix at the best-fit point for all POIs
  - calculate using inverse Hessian $\rho(\mu_1,\mu_2) = H^{-1}(\mu_1,\mu_2)\,/\,\left(H^{-1}(\mu_1,\mu_1)H^{-1}(\mu_2,\mu_2)\right)^{1/2}$
  - but beware that the correlations at the best-fit can be quite different elsewhere



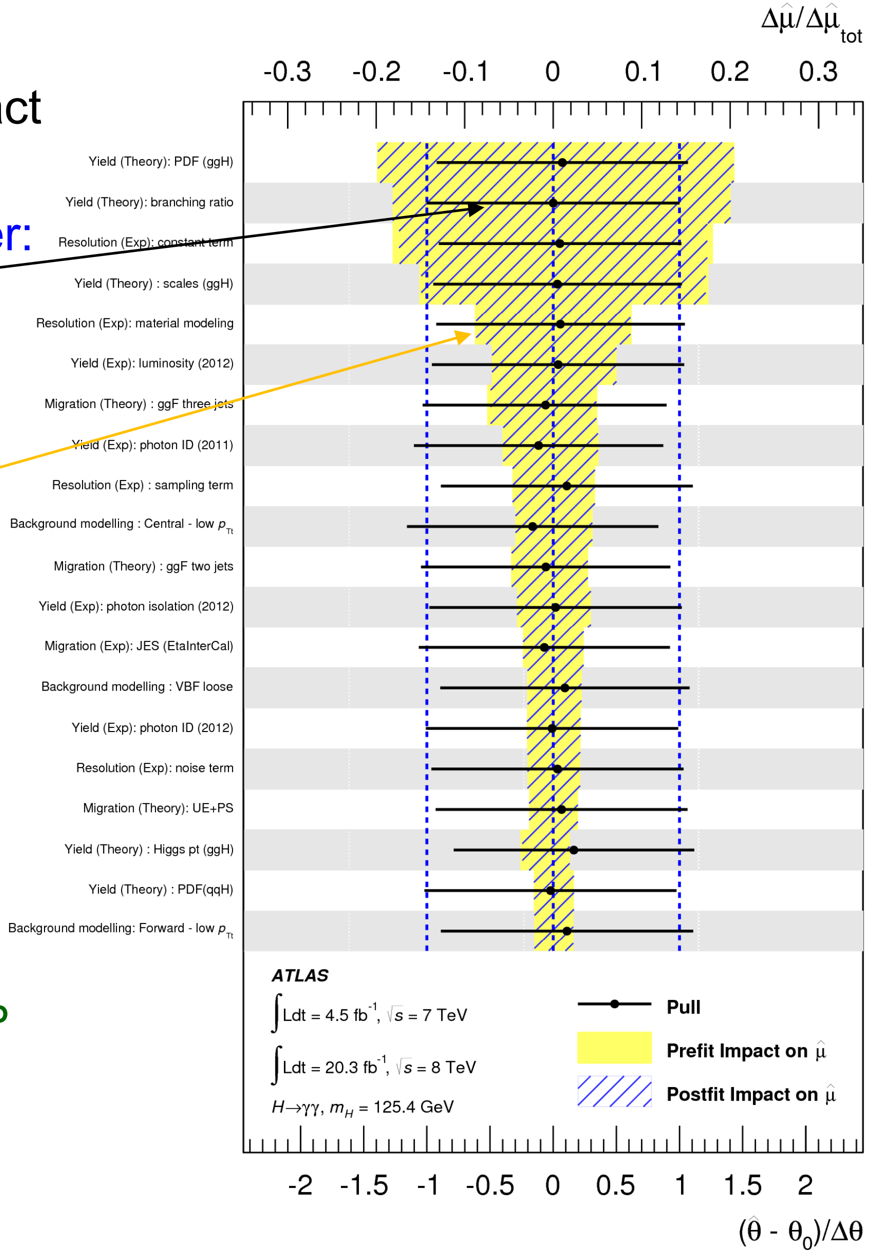


| | | ggF | | | | VBF | | | | WH | | | | ZH | | | | ttH | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\gamma\gamma$ | ZZ | WW | $\tau\tau$ | $\gamma\gamma$ | ZZ | WW | $\tau\tau$ | $\gamma\gamma$ | WW | $\tau\tau$ | bb | $\gamma\gamma$ | WW | $\tau\tau$ | bb | $\gamma\gamma$ | WW | $\tau\tau$ | bb |
| ggF | $\gamma\gamma$ | 1 | 0.01 | 0.00 | 0.00 | −0.25 | −0.01 | 0.00 | 0.00 | −0.03 | 0.00 | 0.00 | 0.00 | −0.14 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 |
| | ZZ | 0.01 | 1 | 0.03 | 0.00 | −0.01 | −0.26 | −0.01 | −0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | −0.08 | −0.02 | −0.02 |
| | WW | 0.00 | 0.03 | 1 | −0.01 | 0.01 | 0.00 | −0.16 | 0.00 | 0.00 | −0.07 | 0.02 | 0.00 | 0.00 | −0.07 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 |
| | $\tau\tau$ | 0.00 | 0.00 | −0.01 | 1 | 0.01 | 0.01 | 0.01 | −0.37 | 0.00 | 0.03 | −0.25 | 0.00 | 0.00 | 0.01 | −0.21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
| VBF | $\gamma\gamma$ | −0.25 | −0.01 | 0.01 | 0.01 | 1 | 0.02 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 |
| | ZZ | −0.01 | −0.26 | 0.00 | 0.01 | 0.02 | 1 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.01 | 0.01 |
| | WW | 0.00 | −0.01 | −0.16 | 0.01 | 0.02 | 0.02 | 1 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | −0.01 |
| | $\tau\tau$ | 0.00 | −0.01 | 0.00 | −0.37 | 0.02 | 0.02 | 0.03 | 1 | 0.00 | −0.01 | 0.05 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | −0.01 |
| WH | $\gamma\gamma$ | −0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1 | 0.00 | 0.00 | 0.00 | −0.64 | 0.00 | 0.00 | 0.00 | −0.03 | 0.00 | 0.00 | 0.00 |
| | WW | 0.00 | 0.00 | −0.07 | 0.03 | 0.00 | 0.00 | 0.00 | −0.01 | 0.00 | 1 | −0.12 | 0.00 | 0.00 | −0.12 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
| | $\tau\tau$ | 0.00 | 0.01 | 0.02 | −0.25 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | −0.12 | 1 | 0.00 | 0.00 | 0.02 | −0.02 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
| | bb | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1 | 0.00 | 0.00 | 0.00 | −0.07 | 0.00 | 0.00 | 0.00 | 0.00 |
| ZH | $\gamma\gamma$ | −0.14 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | −0.64 | 0.00 | 0.00 | 0.00 | 1 | 0.00 | 0.00 | 0.00 | −0.11 | 0.00 | 0.00 | 0.00 |
| | WW | 0.00 | 0.01 | −0.07 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | −0.12 | 0.02 | 0.00 | 0.00 | 1 | −0.05 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
| | $\tau\tau$ | 0.00 | 0.01 | 0.01 | −0.21 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.01 | −0.02 | 0.00 | 0.00 | −0.05 | 1 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
| | bb | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | −0.07 | 0.00 | 0.00 | 0.00 | 1 | 0.00 | 0.00 | 0.00 | 0.00 |
| ttH | $\gamma\gamma$ | 0.01 | 0.00 | 0.00 | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 | −0.03 | 0.00 | 0.00 | 0.00 | −0.11 | 0.00 | 0.00 | 0.00 | 1 | 0.00 | 0.00 | 0.00 |
| | WW | 0.01 | −0.08 | 0.01 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 1 | −0.47 | 0.01 |
| | $\tau\tau$ | 0.00 | −0.02 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | −0.47 | 1 | 0.00 |
| | bb | 0.00 | −0.02 | 0.01 | 0.01 | 0.00 | 0.01 | −0.01 | −0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 1 |

# Validation of model

PRD 90 (2014) 112015

- The NPs' effect on a model can be tested by determining by their post-fit pulls and impact on the POI
  - Often (perhaps confusingly) displayed together:
  1. NP best-fit value and error
     - relative to nominal, $(\hat{\theta}-\theta_0)/\Delta\theta$, here indicated by blue dotted lines at $0 \pm 1$.
     - refers to scale at the bottom
  2. Impact of NP's error on POI
     - $\pm\Delta\hat{\mu} = \hat{\mu}(\hat{\theta} \pm \sigma_\theta) - \hat{\mu}$
       - important to check relative sign of impact if correlating NPs in a combined workspace
     - can use pre-fit (nominal) and/or post-fit NP errors
     - refers to scale at the top, here relative to the total error, $\Delta\hat{\mu}/\Delta\hat{\mu}_{\text{tot}}$
     - Size of impact indicates importance of each NP



# Discovery

# Discovery significance

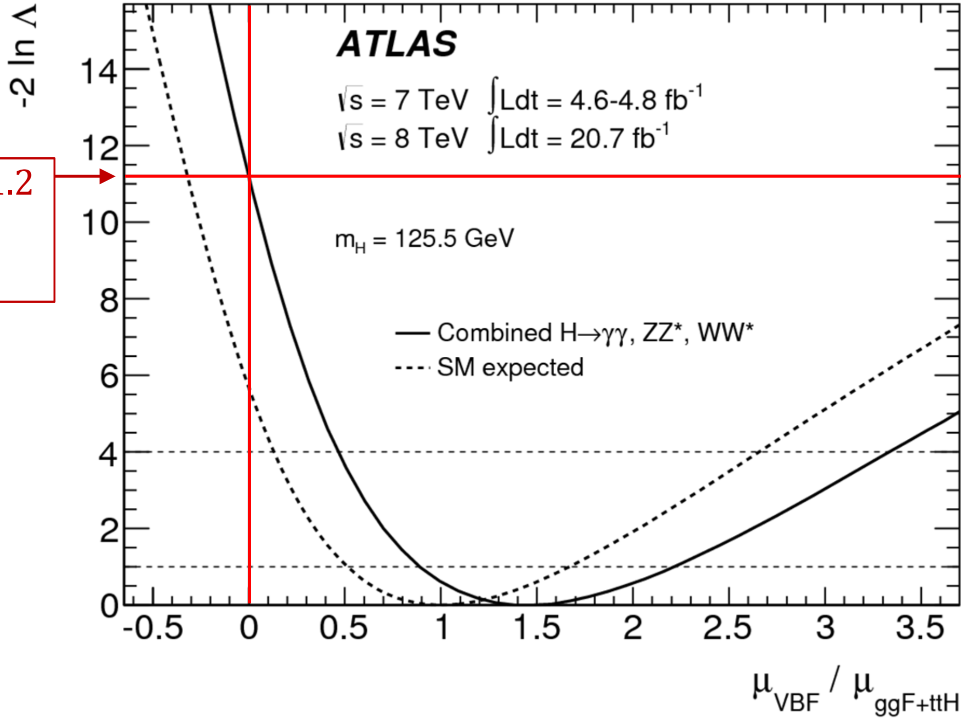
PLB 726 (2013) 88

- In the asymptotic limit (large N), the PLR, $\Lambda(\mu) = \frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})}$, gives the compatibility between $\mu$ and $\hat{\mu}$ hypotheses.
  - Where $\mu$ is a ratio relative to the SM (eg. $\mu = \sigma/\sigma_{\mathrm{SM}}$), we can test
    1. Compatibility with background-only hypothesis: $Z_0 = \sqrt{-2\ln\Lambda(\mu=0)}$
    2. Compatibility with SM (1 POI): $Z_{\mathrm{SM}} = \sqrt{-2\ln\Lambda(\mu=1)}$
    3. Compatibility with SM ($n$ POIs): $Z_{\mathrm{SM}} = D^{-1}(\chi_n^2(-2\ln\Lambda(\boldsymbol{\mu})))$
  - $Z_\mu$ is the significance ($N\sigma$), which (assuming $\chi_1^2$ for 1 POI) has equivalent p-value, $p_\mu = s\,\Phi(-Z_\mu)$, where
    - $s = 1$ for single-sided test like $p_0$ [1]
    - $s = 2$ for double-sided test like $p_{\mathrm{SM}}$
    - $\Phi(Z)$ is the Gaussian CDF [2]

$-2\ln\Lambda = 11.2$

$Z_0 = 3.3\sigma$

$p_0 = 0.04\%$



- $p_0$ interpreted as the significance of a signal, relative to a background-only hypothesis

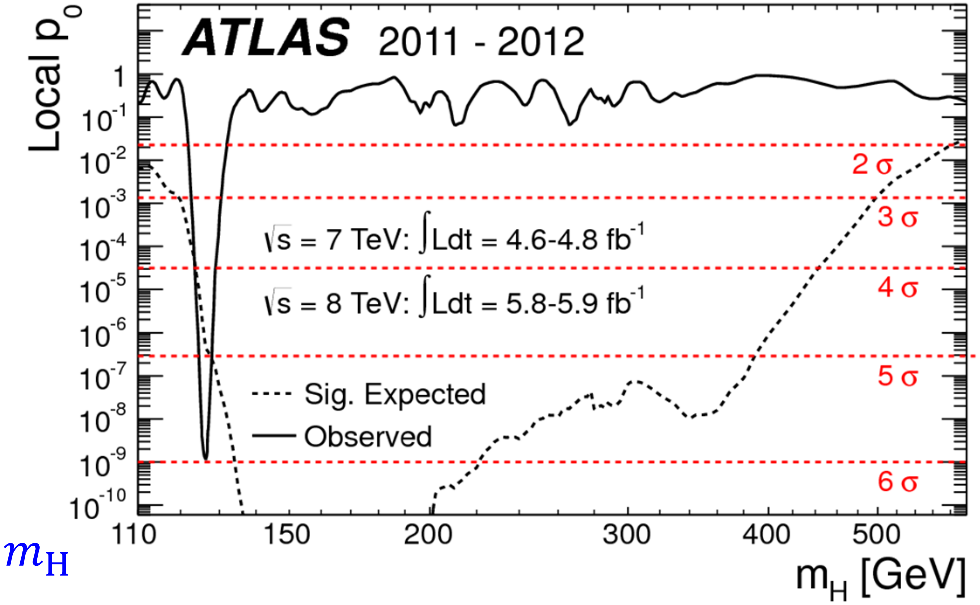
[1] 1-sided p-value is capped at $p_0 < 0.5$. Can uncap by using $-Z_0$ for $\hat{\mu} < 0$

[2] $\Phi(Z)$ = ROOT::Math::gaussian_cdf(Z)
$\Phi^{-1}(p)$ = ROOT::Math::gaussian_quantile(p,1.0)

# Discovery: $p_0$ vs $m_\mathrm{H}$

PLB 716 (2012) 1

- Each mass hypothesis ($m_H$) has its own likelihood function, $L_{m_H}(\mu, \boldsymbol{\theta})$, eg.
  1. $m_H$ hypothesis in kinematic fits
  2. $\mu = \sigma/\sigma_\mathrm{SM}(m_H)$ so need $m_H$-specific SM production XS and decay BR [LHC-H-XS-WG]
  3. each combined likelihood includes accessible decay modes at specified $m_\mathrm{H}$



- $p_0$ vs $m_H$ plot is the result of ~independent fits to each $L_{m_H}$ [1]
  - The largest local significance is $6.0\sigma$ ($p_0{\sim}10^{-9}$) at $m_H = 126.5$ GeV
    - the result of many (part-correlated) searches across the full $110 \le m_H < 600$ GeV range
  - correct for the “look-elsewhere effect” using Gross-Vitells formula [arXiv:1005.1891]:
    - $p_\mathrm{global} = p_\mathrm{local} + \langle N(c_0)\rangle e^{-(c-c_0)/2} \qquad = 10^{-9} + 9 \cdot e^{-6.0^2/2} = 1.4 \cdot 10^{-7} \;\to 5.1\sigma$

- Still using asymptotic approximation, which we may not be confident in for new signal

  → test with toys

[1] except in $m_H$ measurement, use single likelihood $L(m_H, \boldsymbol{\theta})$

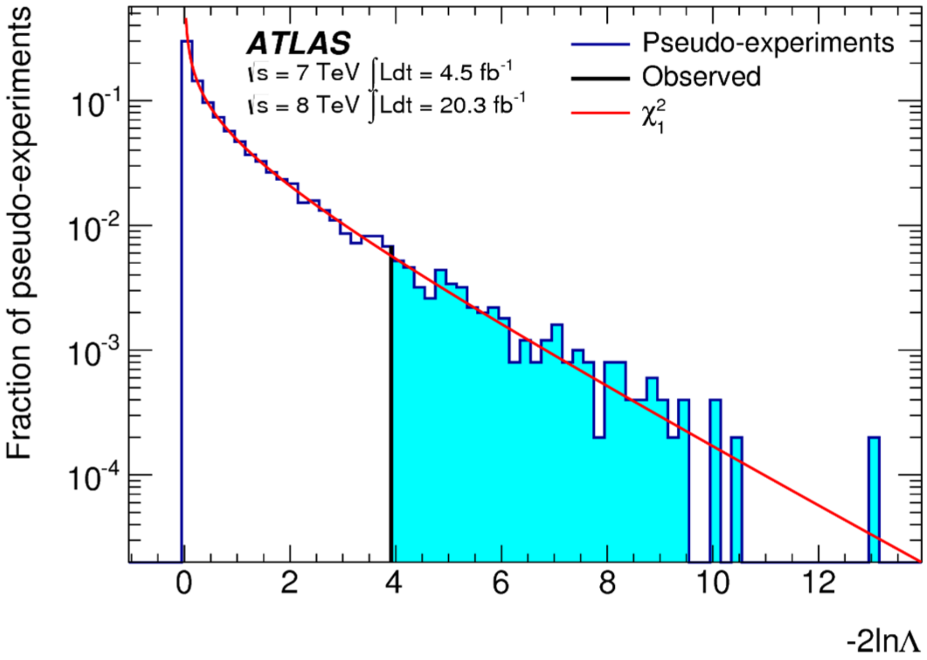
# Discovery with toys

PRD 90 (2014) 052004

- Toy MC (AKA "Monte Carlo pseudo-experiments") can be generated directly from the components of the likelihood function
  1. For each toy, generate
     1. toy dataset (pdf.generate(obs)), with $\mu, \boldsymbol{\theta}$ determined from expectation or fit to data
     2. set of global observables (pdf.generate(globObs))
        - simulates variation of "NP truth"
  2. Calculate a test statistic, $t_\mu = -2\ln\Lambda(\mu)$, requiring:
     1. conditional fit, under hypothesis being tested, eg. $\mu = 0$, background-only for $p_0$
     2. unconditional fit for best-fit $\mu$ for this toy
  - RooStats::FrequentistCalculator() can be used to run this procedure for many toys

Example distribution of $t_0$

(here for a 2-sided test of compatibility of two signals, not 1-sided signal significance)



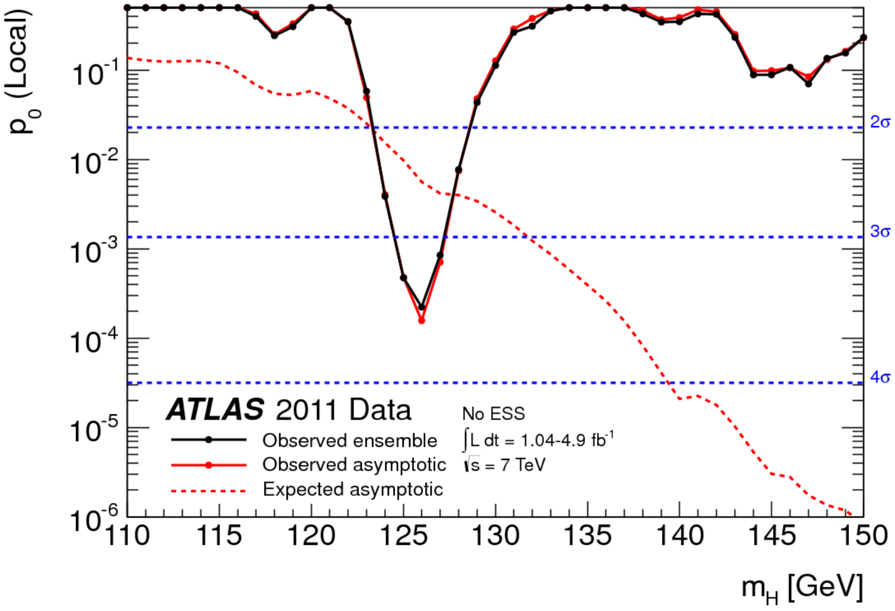
- The observed p-value is just the fraction of toys with test statistic larger than the observed:
  - $p_0 = N_{\text{toys}}\left(t_0 > t_{0,\text{obs}}\right) / N_{\text{toys}}$

# Discovery with toys

PLB 716 (2012) 1
PLB 710 (2012) 49

- For the 2012 ATLAS Higgs discovery
  - the $6.0\sigma$ local significance was reduced to $5.9\sigma$ by including the effect of energy-scale systematics
  - ESS could only be measured using toys at $m_H = 126.5$ GeV
    - limited by CPU time available (used extrapolation from 300k toys)

- The cross-check with toys is more clearly seen with a previous sample
  - lower significance → smaller number of toys required



# Exclusion

# Exclusion: CLs

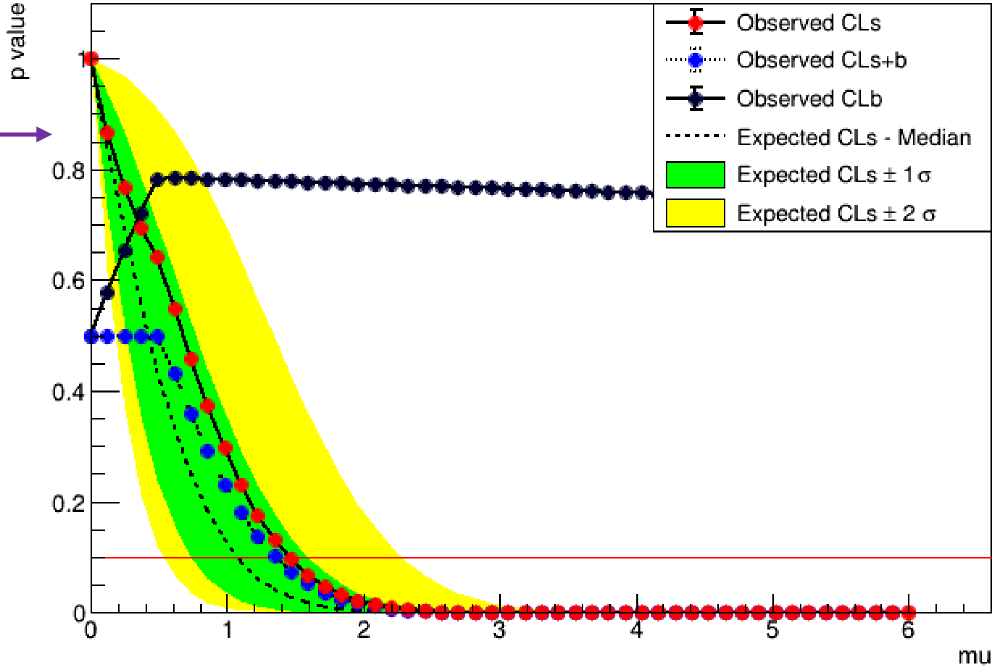
- CLs: $p'_\mu = p_\mu/(1 - p_0)$
  - CLs divides the tested p-value ($\mathrm{CL}_{s+b}$) by the background-exclusion p-value ($\mathrm{CL}_b$)
    - normally has little effect, but it is useful to inhibit a fluctuation spuriously excluding a hypothesis to which we have little sensitivity
  - $p_\mu$ and $p_0$ can be calculated as described previously for toys
- For a 95% CL limit, reject a particular $\mu$ (s+b) hypothesis if $p'_\mu \leq 0.05$.
  - to obtain a limit, find $\mu_{\mathrm{up}}$, the $\mu$ value for which $p'_\mu = 0.05$

- For toys, this means generating/fitting toys for various $\mu$ and interpolating $\mu_{\mathrm{up}}$
- Asymptotic limit obtained using the procedure from Asimov Paper [arXiv:1007.1727]
  - $q_\mu \quad = -2\ln\Lambda(\mu)$ PLR for observed data
  - $q_{\mu,A} = -2\ln\Lambda_A(\mu|0)$ PLR for background-only Asimov dataset
  - $p'_\mu = (1 - \Phi(\sqrt{q_\mu}\,))$ / $\Phi(\sqrt{q_{\mu,A}} - \sqrt{q_\mu})$
    - scan $\mu$ to find $\mu_{\mathrm{up}}$ for which $p'_\mu = 0.05$.
  - For the median expected limit, $\mu_{\mathrm{up}} = 1.96\,\sigma(\mu_{\mathrm{up}})$ $[\Phi^{-1}(1 - 0.05/2) = 1.96]$
    - where $\sigma(\mu_{\mathrm{up}}) = \mu_{\mathrm{up}}/\sqrt{q_{\mu_{\mathrm{up}},A}}$, so again requires a numerical determination of $\mu_{\mathrm{up}}$
    - The expected bands, median$\pm N\sigma$, $\mu_{\mathrm{up}+N} = \left(\Phi^{-1}\big(1 - 0.05\Phi(N)\big) + N\right)\cdot\sigma(\mu_{\mathrm{up}+N})$

# CLs procedure

- For RooFit models, see:
  1. RooStats StandardHypoTestInvDemo.C tutorial, or
  2. ATLAS CLs tutorial
- In summary, create an asymptotic or toy calculator:
  1. RooStats::AsymptoticCalculator  calc (data, bModel, sbModel);    // or
  2. RooStats::FrequentistCalculator  calc (data, bModel, sbModel);
- and pass that to the hypothesis test inverter:

```
RooStats::HypoTestInverter  hypo (calc);
result = hypo.GetInterval();
RooStats::HypoTestInverterPlot (,,result);
```

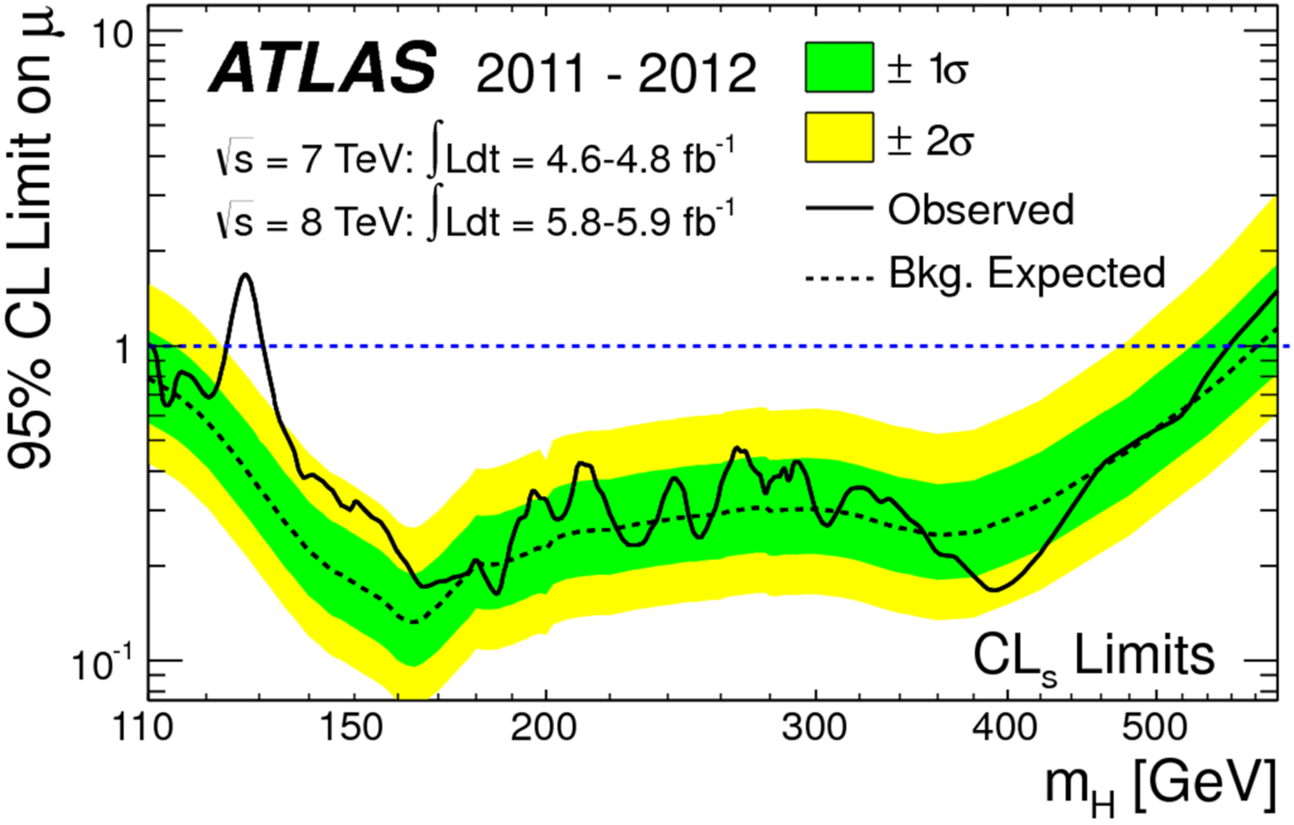


- For HistFactory-style models, pyhf has built-in tools to calculate CLs

# Exclusion limit vs $m_H$

PLB 716 (2012) 1

- In Higgs search, plot $\mu_{\text{up}}$ vs $m_H$
  - different likelihood for each $m_H$, as before



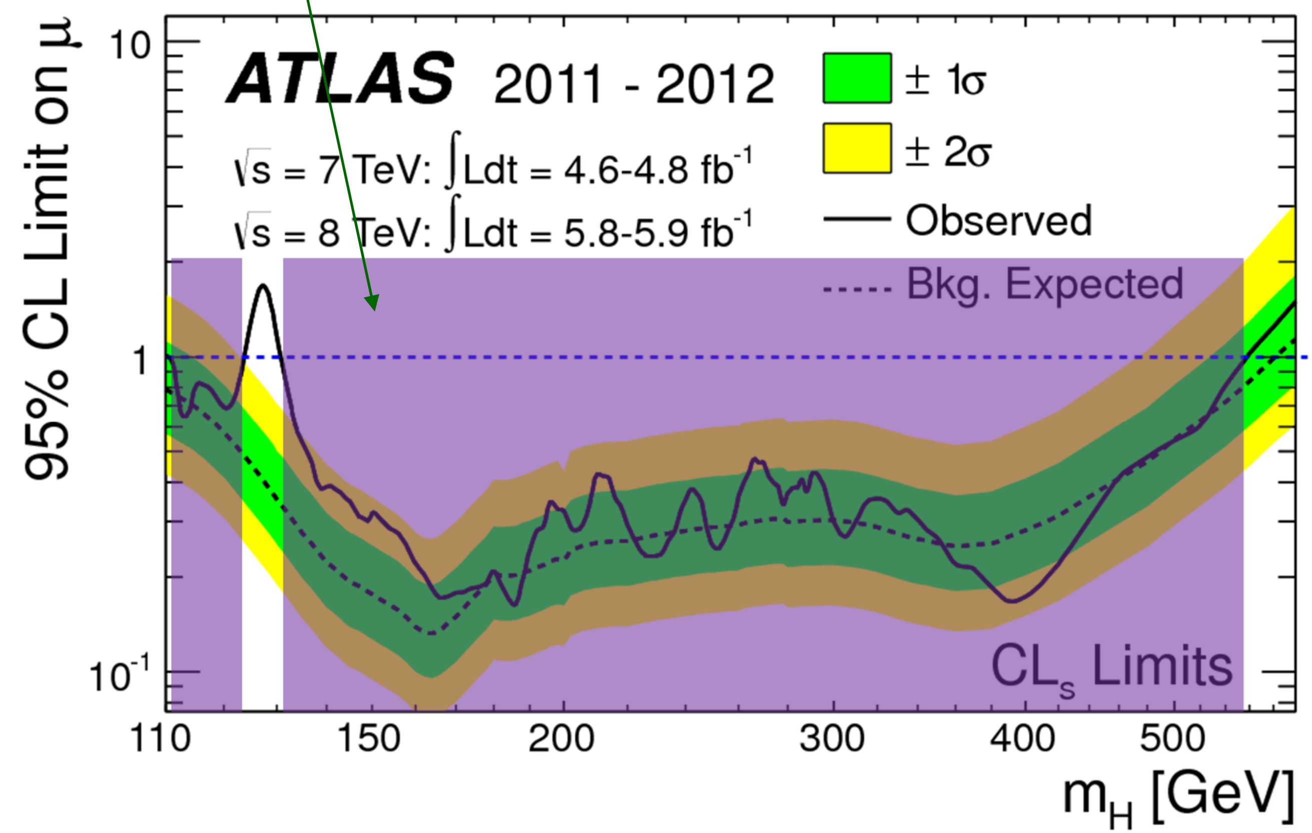
# Exclusion limit vs $m_H$

PLB 716 (2012) 1

- In Higgs search, plot $\mu_{\text{up}}$ vs $m_H$
  - different likelihood for each $m_H$, as before

**Observed**

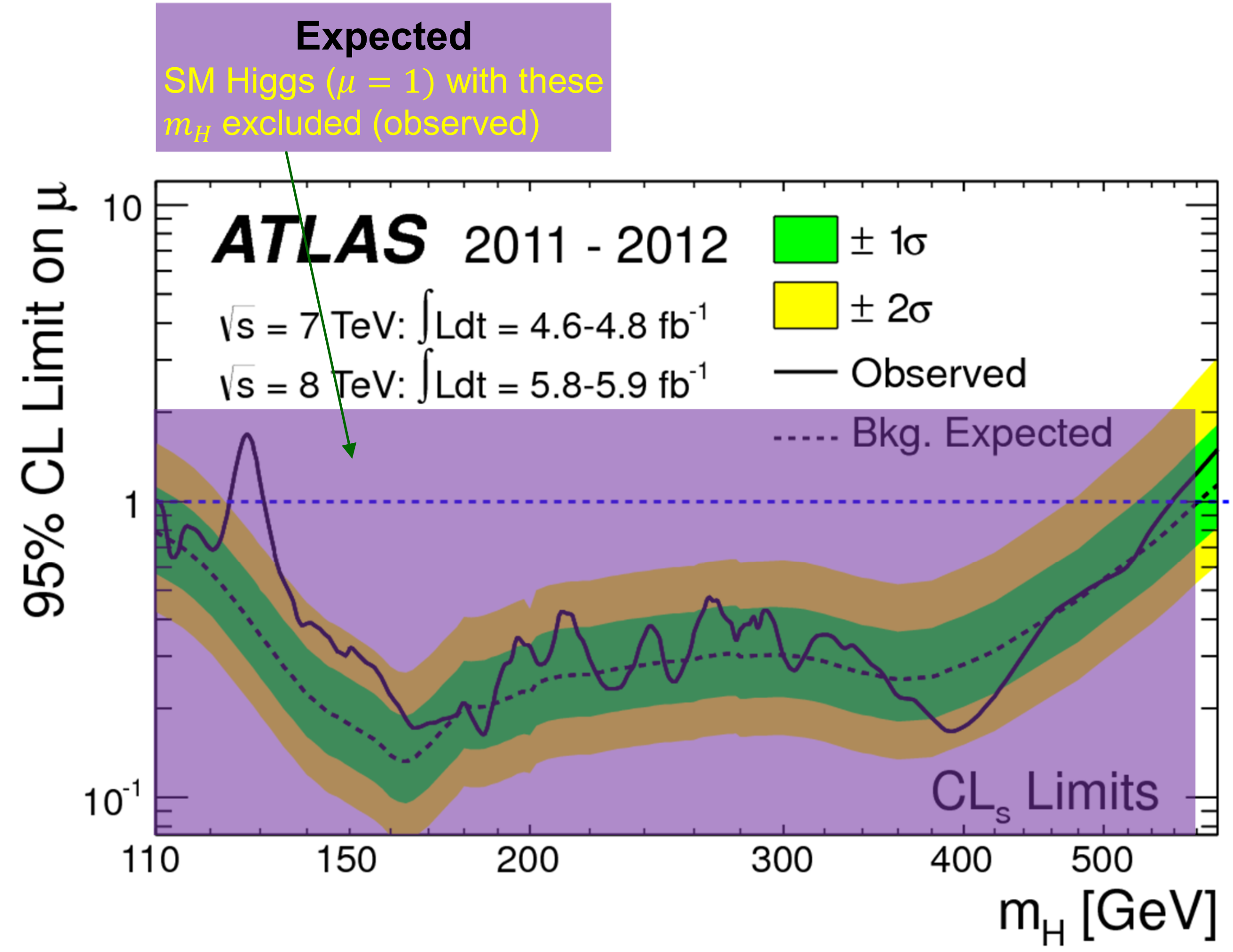
SM Higgs ($\mu = 1$) with these $m_H$ excluded (observed)



# Exclusion limit vs $m_H$

PLB 716 (2012) 1

- In Higgs search, plot $\mu_{\text{up}}$ vs $m_H$
  - different likelihood for each $m_H$, as before



**Expected**
SM Higgs ($\mu = 1$) with these
$m_H$ excluded (observed)

# Summary

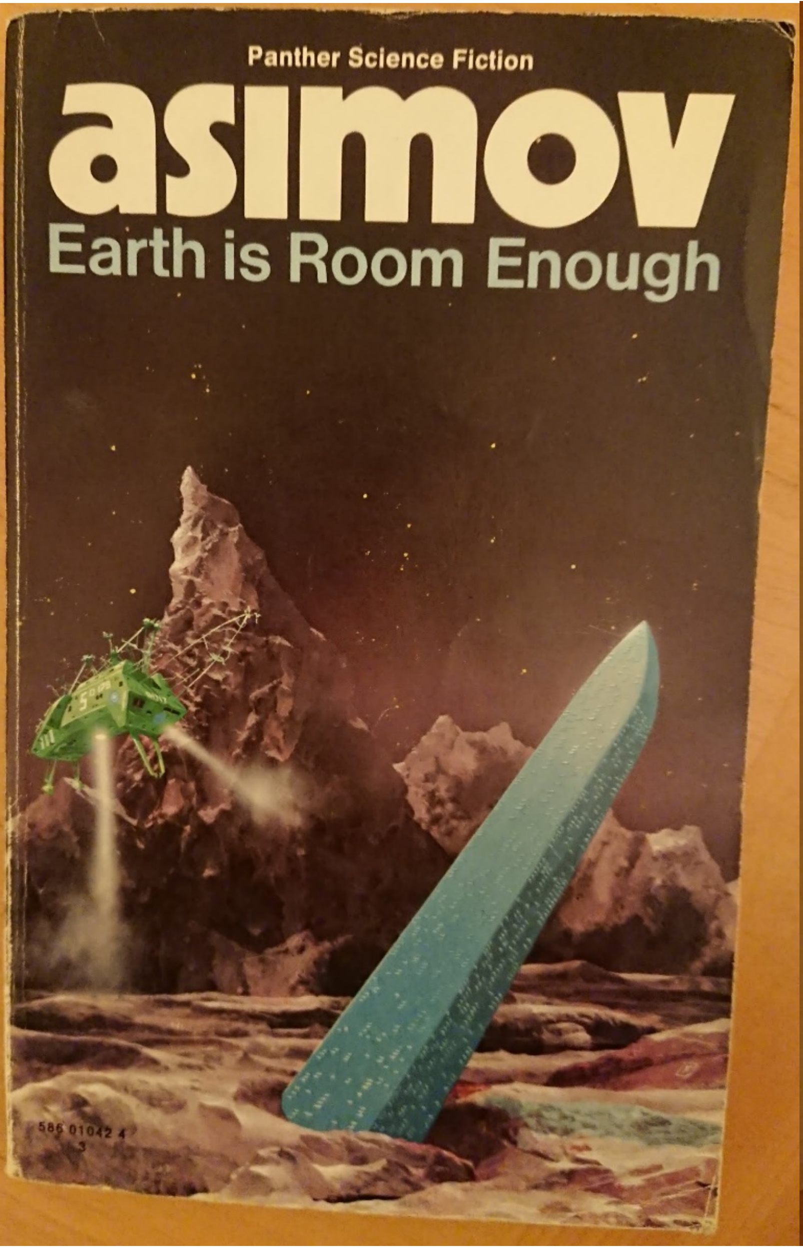
# Summary

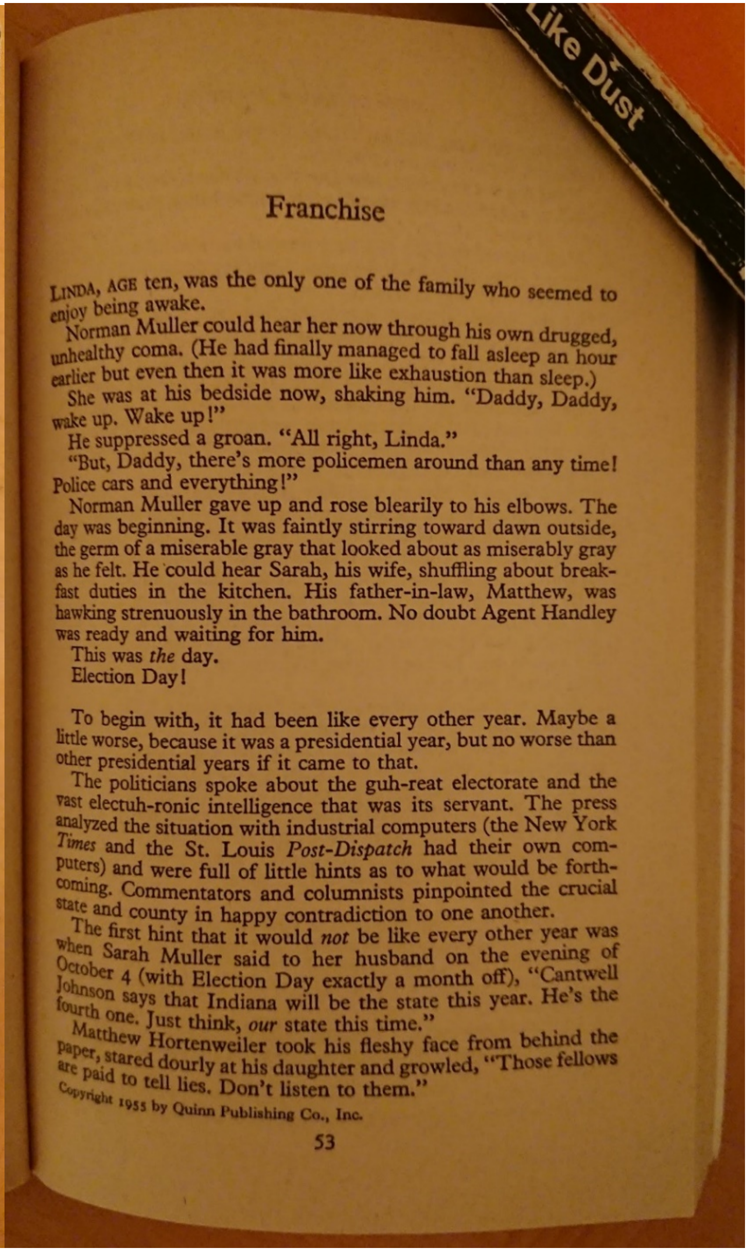
- Build models with
  - RooFit (C++, Python, or factory)
  - HistFactory (XML)
  - pyhf (JSON)
- Keep model and data in RooFit workspace files
  - Asimov dataset

- Statistical tests
  - Measurement, scanning profile likelihood ratio
    - RooFit
  - Discovery with profile likelihood ratio, asymptotic or toys
    - RooFit, RooStats
  - Exclusion with CLs, asymptotic or toys
    - RooFit, RooStats, or pyhf

# Backup

# Isaac Asimov – Franchise





## Franchise

LINDA, AGE ten, was the only one of the family who seemed to enjoy being awake.

Norman Muller could hear her now through his own drugged, unhealthy coma. (He had finally managed to fall asleep an hour earlier but even then it was more like exhaustion than sleep.)

She was at his bedside now, shaking him. "Daddy, Daddy, wake up. Wake up!"

He suppressed a groan. "All right, Linda."

"But, Daddy, there's more policemen around than any time! Police cars and everything!"

Norman Muller gave up and rose blearily to his elbows. The day was beginning. It was faintly stirring toward dawn outside, the germ of a miserable gray that looked about as miserably gray as he felt. He could hear Sarah, his wife, shuffling about breakfast duties in the kitchen. His father-in-law, Matthew, was hawking strenuously in the bathroom. No doubt Agent Handley was ready and waiting for him.

This was *the* day.

Election Day!

To begin with, it had been like every other year. Maybe a little worse, because it was a presidential year, but no worse than other presidential years if it came to that.

The politicians spoke about the guh-reat electorate and the vast electuh-ronic intelligence that was its servant. The press analyzed the situation with industrial computers (the New York *Times* and the St. Louis *Post-Dispatch* had their own computers) and were full of little hints as to what would be forthcoming. Commentators and columnists pinpointed the crucial state and county in happy contradiction to one another.

The first hint that it would *not* be like every other year was when Sarah Muller said to her husband on the evening of October 4 (with Election Day exactly a month off), "Cantwell Johnson says that Indiana will be the state this year. He's the fourth one. Just think, *our* state this time."

Matthew Hortenweiler took his fleshy face from behind the paper, stared dourly at his daughter and growled, "Those fellows are paid to tell lies. Don't listen to them."

53

The Asimov dataset [arXiv:1007.1727] is named for SF author, Isaac Asimov, whose 1955 short story, *Franchise*, envisaged the 2008 US Presidential Election decided by one voter representative of the entire electorate.

This is my copy of the story, in a collection.