# RooStats: a tool for statistical data analysis

Grégory Schott

Karlsruhe Institute of Technology (KIT)

heptools Higgs mini-workshop
Torino - November 24th 2009

# Outline

- What is RooStats? It's a collaborative project between ATLAS, CMS and ROOT to provide a consolidated set of statistical tools
  - TWiki: https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome
- This is not a talk about statistics
  - See: http://indico.cern.ch/conferenceDisplay.py?confId=73545
- Neither about the software implementation
  - See: http://indico.cern.ch/conferenceDisplay.py?confId=72320
- Description and motivation of the project

# Motivations

- Statistical interpretation of data in an analysis
  - useful to have a common, well tested package
- Combination of analyzes within an experiment
- Possible combination of Atlas and CMS analyzes
- Be able to compare statistical methods
- Generalize and cleanup statistical tools in ROOT
- Want to agree on statistical conventions
  - avoid *apples-to-oranges* comparisons

# Features

- Rely on RooFit: provides a developed & flexible basis
- Extension to complex problems
  - Work on arbitrary data and model and can handle many observable, parameter of interest and nuisance parameters
- Combine at analysis level
  - Retain full information for treating correlations
- All statistical methods start from description of likelihood function (or PDF)

# Likelihood analysis

- Simple likelihood: $L_i(n_i|r,s_i,b_i) = \frac{e^{-rs_i-b_i}}{n_i!}(rs_i+b_i)^{n_i}$
  - Can be extended to binned likelihood
- Multiple channels: $L(r) = \prod_i L_i(n_i|r,s_i,b_i)$
- With observables; extended, unbinned likelihood:

$$L(\vec{x}|r,s,b,\vec{\theta}_s,\vec{\theta}_b) = \frac{e^{-rs-b}}{n!}(rs+b)^n \prod_{j=1}^{n}(rsf_s(\vec{x}_j|\vec{\theta}_s)+bf_b(\vec{x}_j|\vec{\theta}_b))$$

  - $f_s$, $f_b$ signal and background distribution from MC or control samples

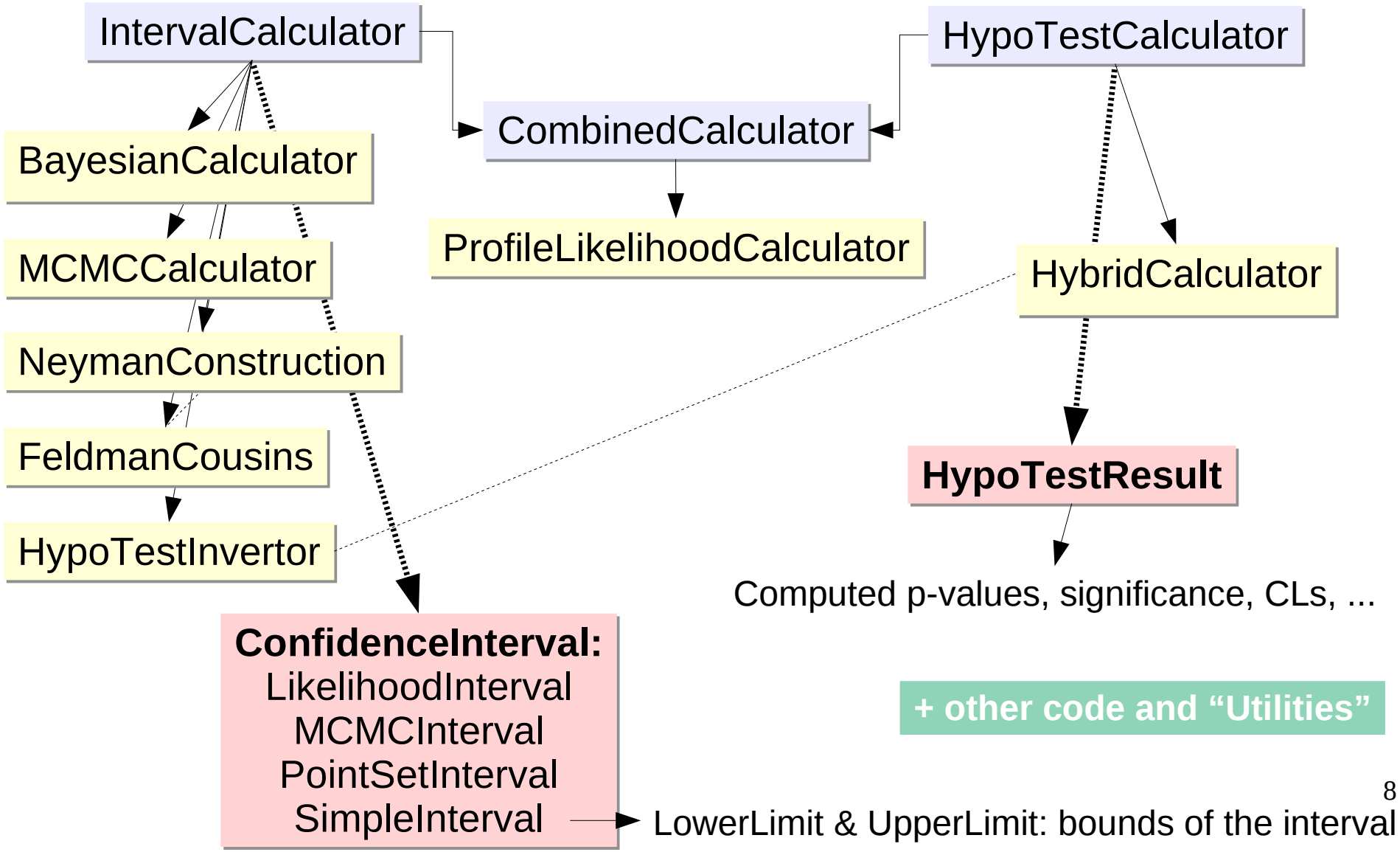
## *(taken from R. Cousins)*

- Once the statistical problem is described, various methods can be easily applied and compared
  - Bayesian, Frequentist, Likelihood ratio, “CLs”, ...
- The community can ask the result be shown with one or another method and to study sampling properties
  - If methods agree → important check of robustness
  - If methods disagree → we learn something
    - The results are answers to different questions
    - Bayesian methods can have poor frequentist properties
    - Frequentist methods can badly violate likelihood principle

# Workspace

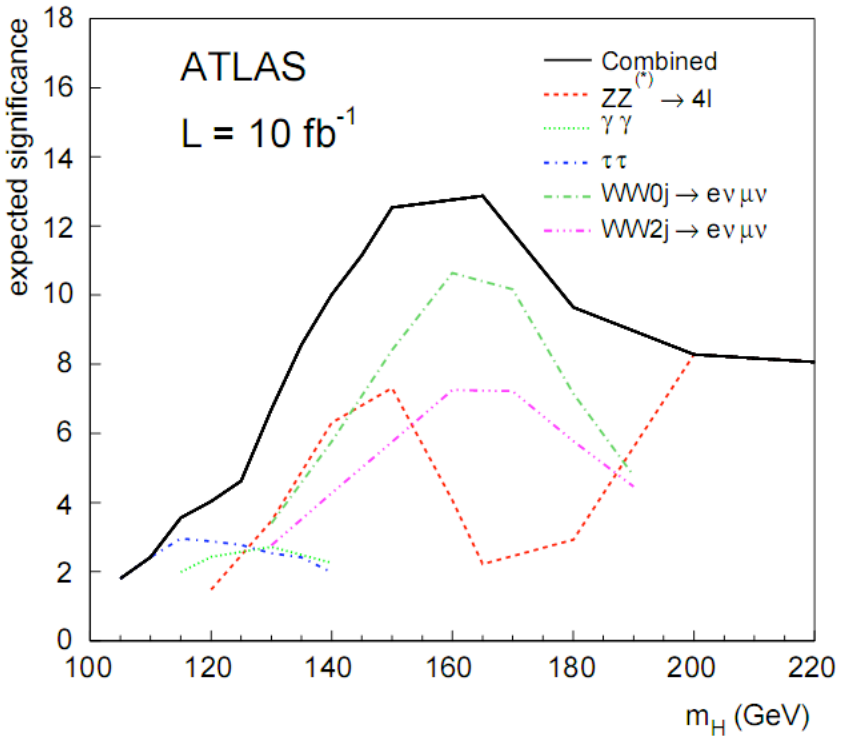
- Developed workspace class to facilitate combinations
- Workspaces contains any RooFit object, in particular:
  - Data (binned or unbinned)
  - PDF model
  - Uncertainty / shape of nuisance parameters
- Utilities to correlate objects or ease the description of the PDF model (Factory)
- Can be saved to file, easily shared and used in combination with Workspaces of other analyzes
- Allow to eventually distribute data and model in an electronic form once analysis has been published

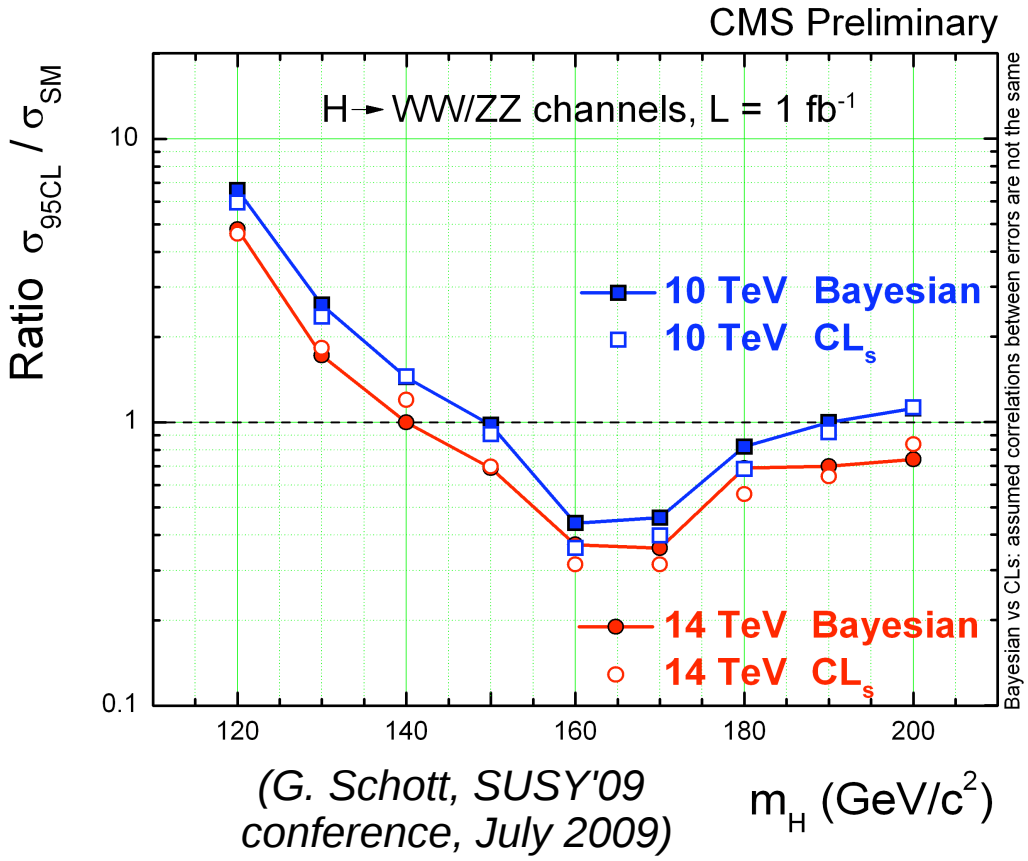
# Overview of classes in RooStats



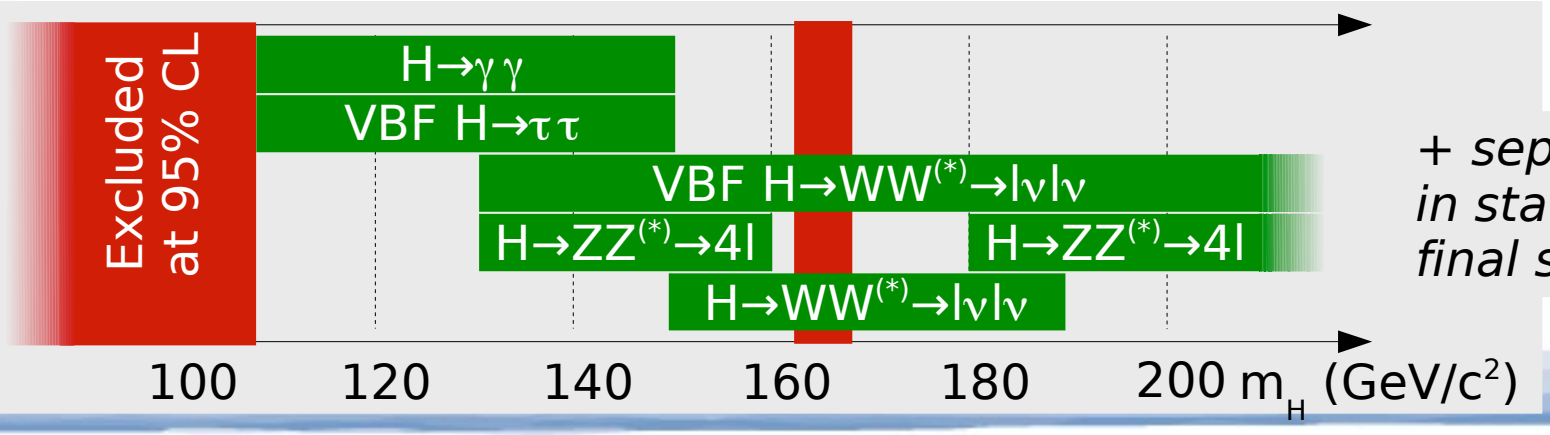
# Some Atlas/CMS Higgs projections

Median expected exclusion
*(CSC report [arXiv:0901.0512](arXiv:0901.0512))*




*(G. Schott, SUSY'09 conference, July 2009)*




*+ separation in stat. indep. final states*

# Summary

- Code in CMS and ATLAS combined and improved to form the RooStats project
- RooStats available from ROOT since December 2008 (new release next week 5.25.04)
  - Common implementation of methods
  - Speak common language for combination
  - Flexible enough to accommodate all cases
- Some improvements needed:
  - Consolidation / speed / documentation / testing
  - Open project, new contributors are welcome

# Conclusion

- Issues with combination:
  - Need common software → RooStats
  - Need convergence on details of statistical methods → ATLAS-CMS statistics forum
  - Inputs (MC generators, previous measurements)
  - Decompose various effect; determine correlations
  - Best knowledge on systematic uncertainties
    - determine and model them

  → Need discussion and agreement to converge