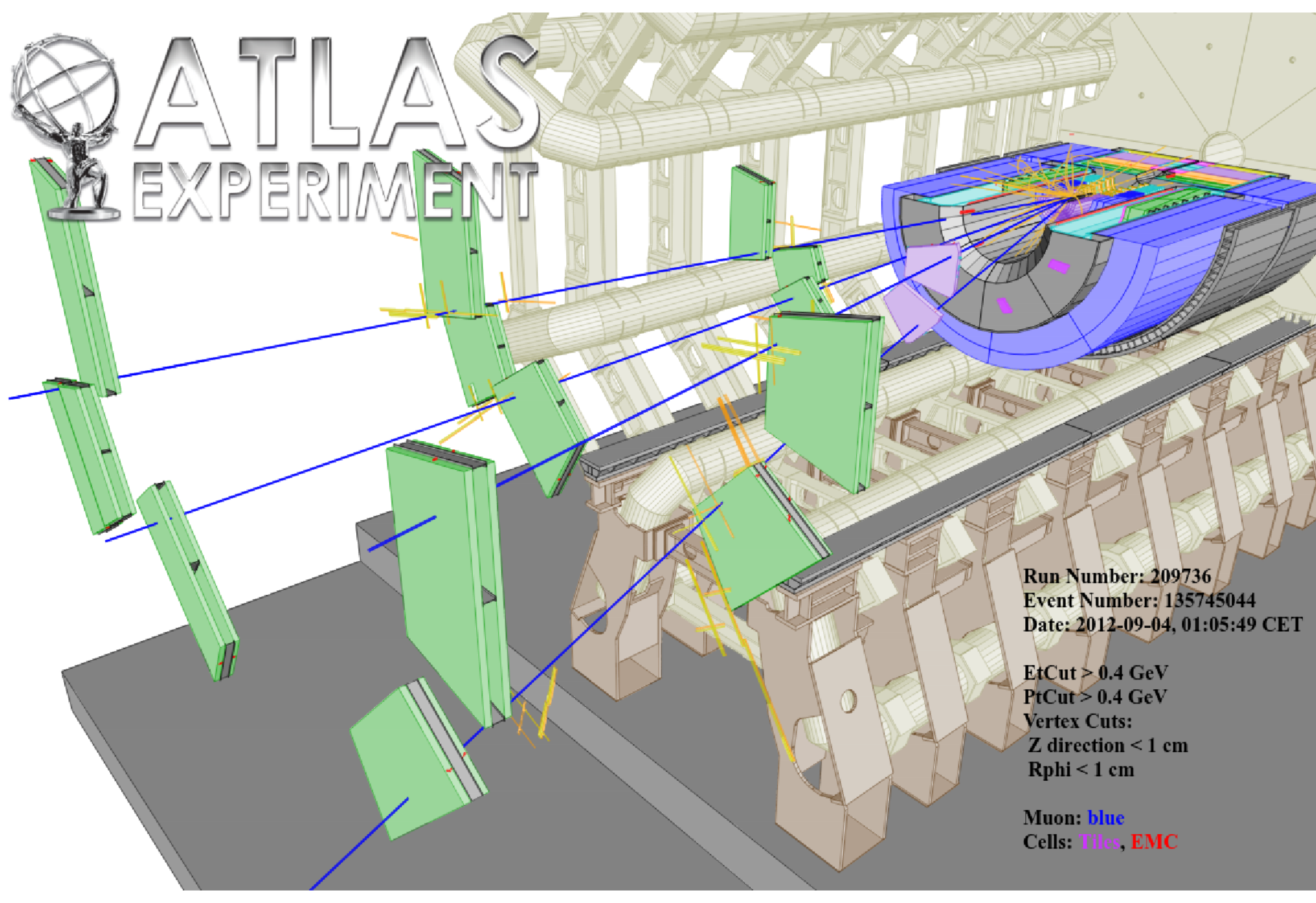
NEW YORK UNIVERSITY
Center For Data Science

ALFRED P. SLOAN FOUNDATION

Gordon and Betty MOORE FOUNDATION

The collaborative statistical modeling tools used to discover the Higgs boson

Sven Kreiss, Kyle Cranmer
Center for Cosmology and Particle Physics, NYU

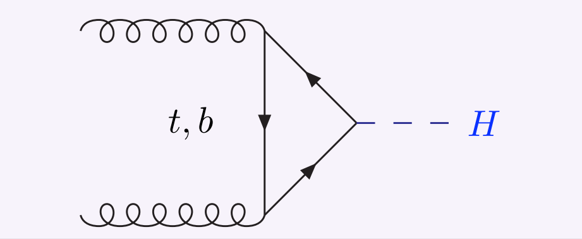
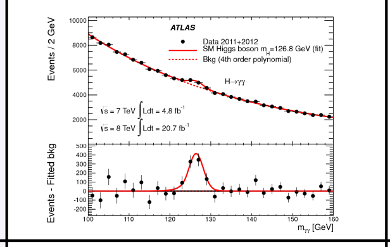
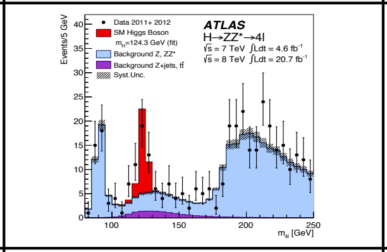
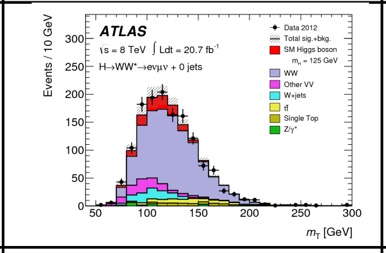
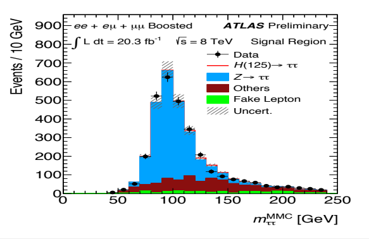
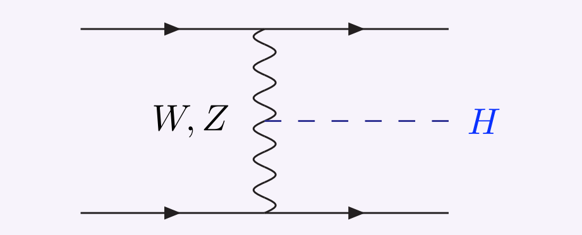
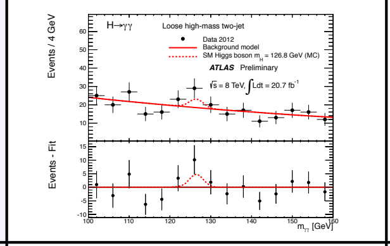
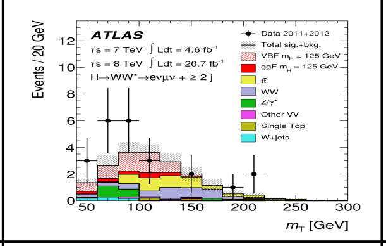
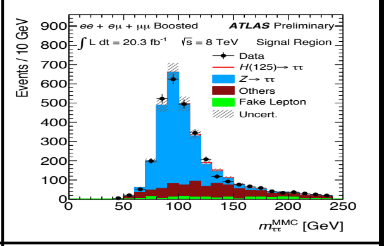
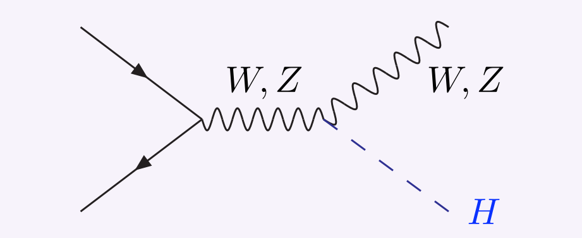
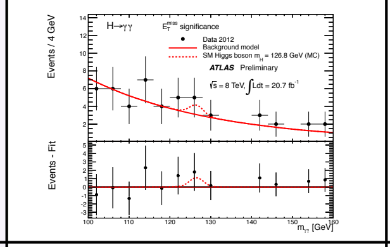
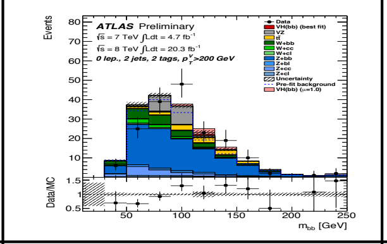
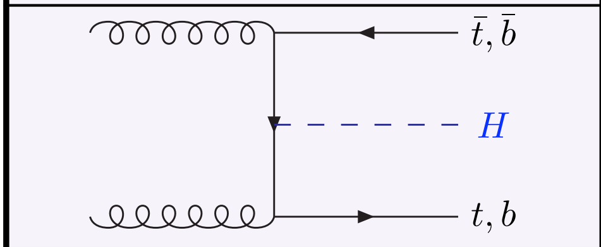
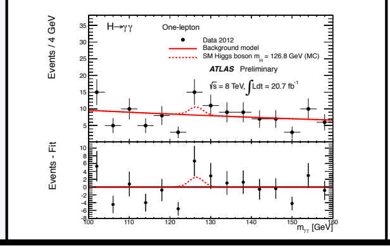


The Data Challenge

- raw data rate > 1 TB / sec
- 10^{15} collisions with 2 MB / collision
- can only save 1 in 10^5 collisions
- still, experiments produce 15 PB / year

Needle in a Haystack

- ~4 billion collisions needed to produce 1 Higgs boson
- classification algorithms designed for each possible production & decay
- ~1 trillion collisions needed to produce a Higgs that passes selection
- depending on production & decay mode, selected events may be:
  - high signal-to-background with sharp feature as in H→ZZ
  - low signal-to-background with sharp feature as in H→γγ
  - intermediate signal-to-background with broad feature as in H→WW
- our theory predicts the expectations in each production & decay

Production\Decay	γγ	ZZ	WW	bb	ττ
					
					
					
					

From Big Data to Big Models

- teams of 20-200 scientists address each of the production & decay modes
- they use theory and a detailed detector simulation to model the data
- modeling approaches range from parametric to non-parametric techniques
- systematic uncertainties of our complicated detectors must be modeled, too

Finally, we combine the statistical models for these different categories of events into a coherent and comprehensive statistical model (see figure).

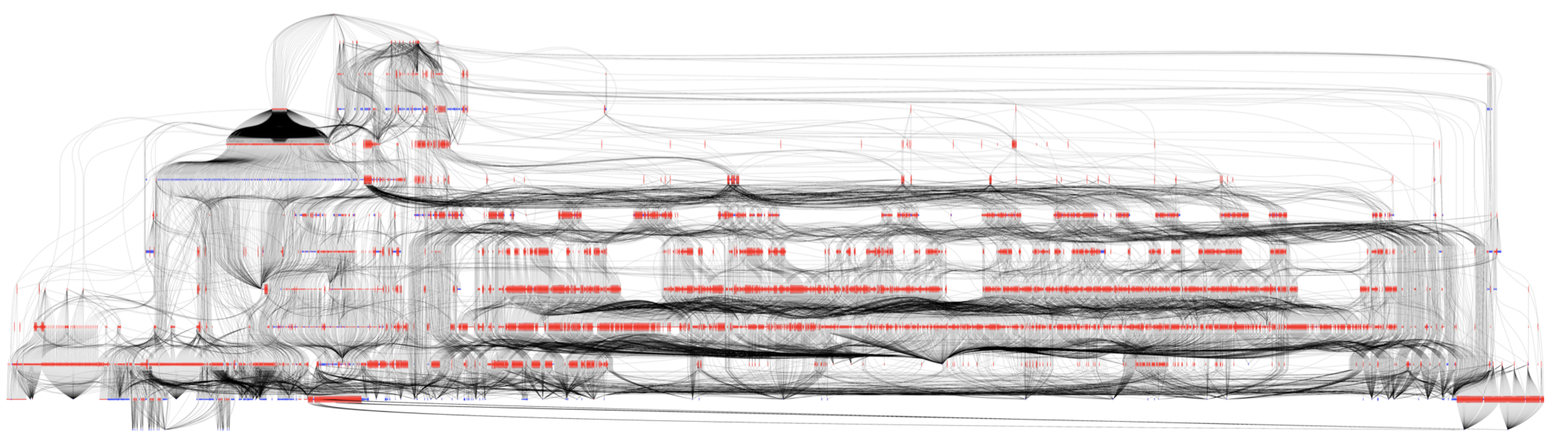
- because we use the same detector for each of the individual searches, we must take special care of correlated systematic uncertainties

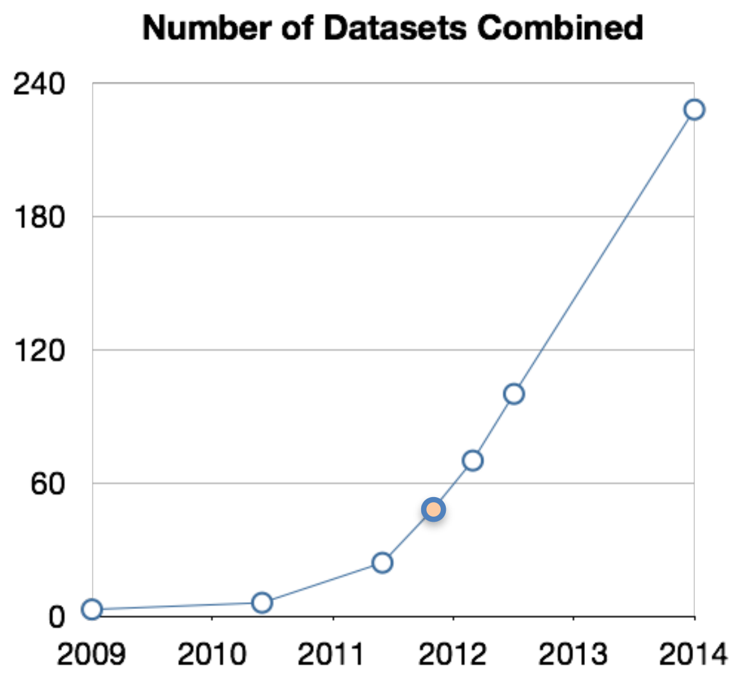
$$p(x_{ce}, a_p\,|\boldsymbol{\alpha}) = \prod_{c\in\text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha})\right] \cdot \prod_{p\in\text{systematics}} f_p(a_p|\alpha_p)$$

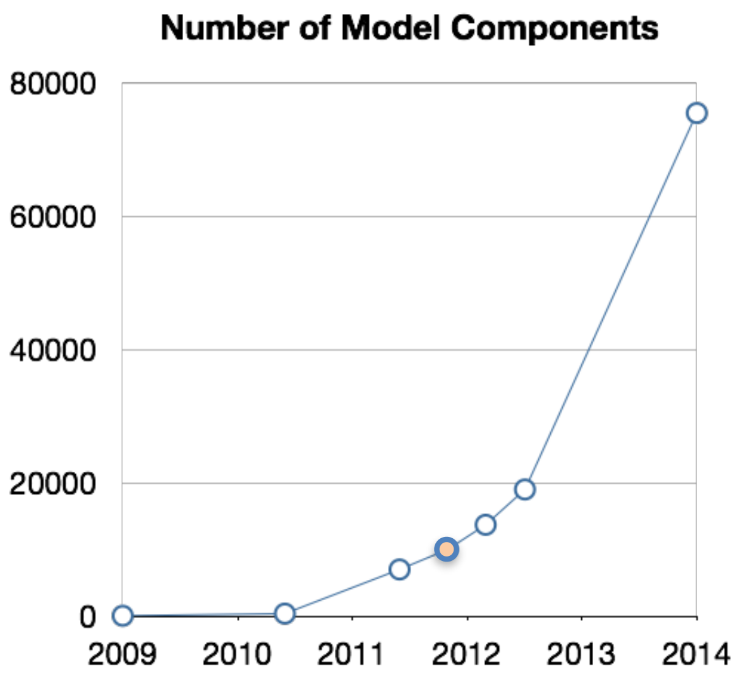
Since introducing the RooFit/RooStats collaborative statistical modeling tools, the growth of model complexity has grown roughly exponentially in several metrics.

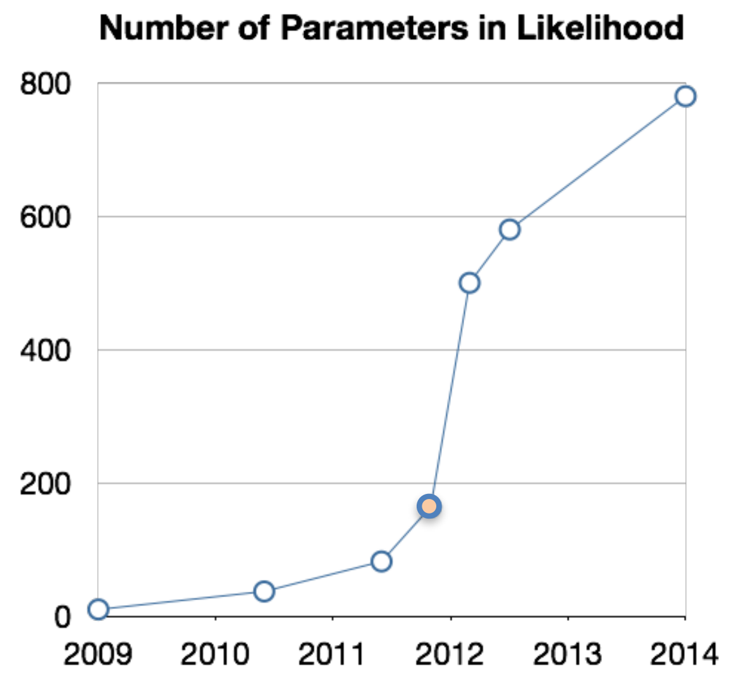
- model complexity leading to new statistical and computational challenges

Visualization of Combined Statistical Model (Nov. 2011)




Number of Datasets Combined

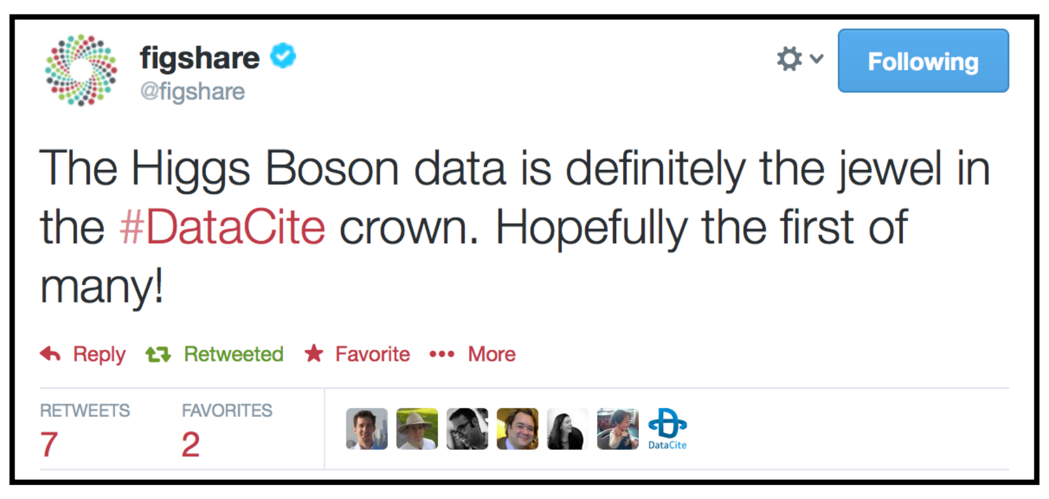

Number of Model Components


Number of Parameters in Likelihood

Digitally Published Statistical Models

This technology allows for unprecedented ability to publish complex statistical models; enabling reproducibility and a broader range of reuse by the community.

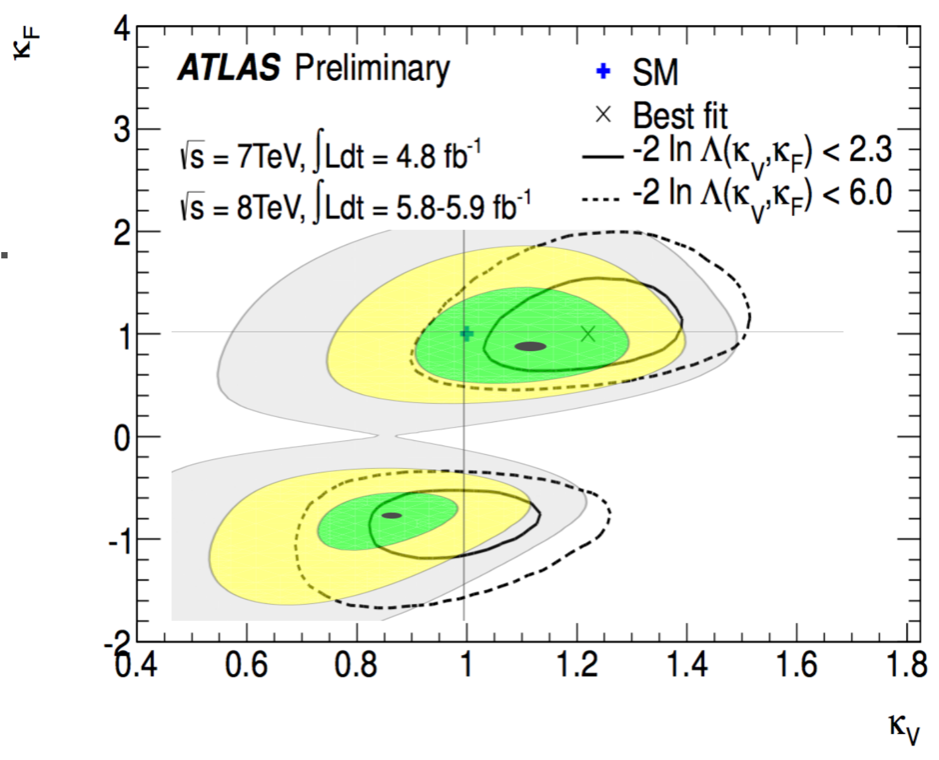
Simplified forms of the ATLAS Higgs likelihoods have been published, assigned DOIs, and are now being used and cited by others.


figshare
@figshare

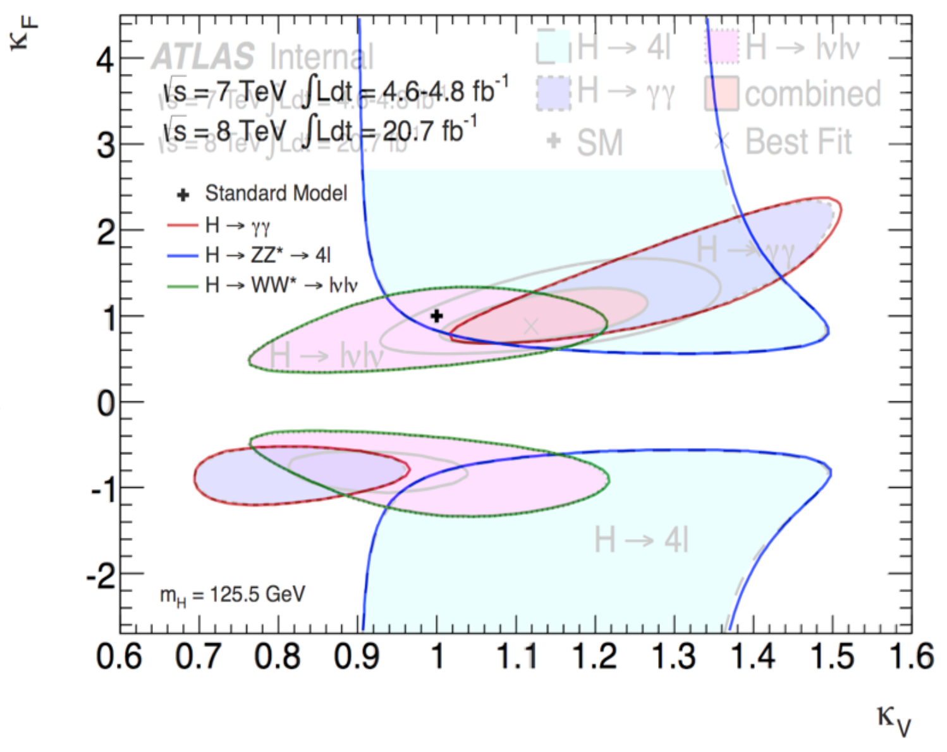
The Higgs Boson data is definitely the jewel in the #DataCite crown. Hopefully the first of many!

These published likelihoods have vastly improved the ability for theorists outside of the collaborations to reproduce our high-level inference on the properties of this particle and enhanced the scientific discourse around profound discovery.

Before



After


ATLAS Internal
$\sqrt{s}$ = 7 TeV $\int$Ldt = 4.6-4.8 fb^{-1}
$\sqrt{s}$ = 8 TeV $\int$Ldt = 20.7 fb^{-1}
H → 4l
H → lνlν
H → γγ
combined
SM
Best Fit
Standard Model
H → γγ
H → ZZ* → 4l
H → WW* → lνlν
H → γγ
H → lνlν
H → 4l
m_H = 125.5 GeV
κ_F
κ_V