Karlsruhe Institute of Technology

Last steps of the analysis

HEP analysis ecosystems workshop, Amsterdam | 22 May 2017

Andrew Gilbert

Outline

- **How are the final results of an analysis produced?**
  - Examples of different types of results
  - Software tools
  - Model building
  - Example: CMS datacards

- **How are results from different analyses combined?**
  - Strategies - combine measurements vs. combine data
  - Example: Top quark mass combination
  - Example: LHC Higgs combination

- **Reinterpretation of results after publication**
  - Sharing the event data
  - Sharing final results
  - Sharing analysis selection
  - Sharing likelihood information

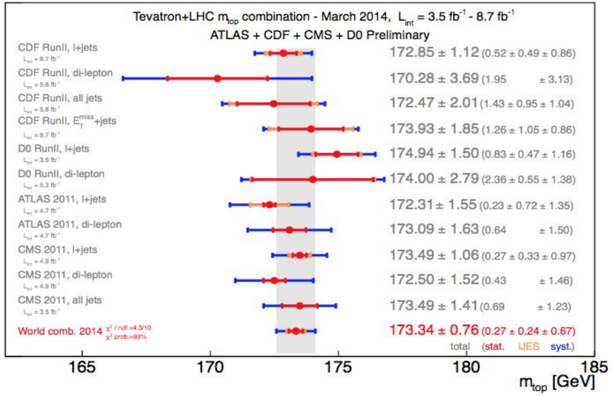
- **Disclaimer**: I come from a CMS Higgs physics background - examples in this talk are biased!

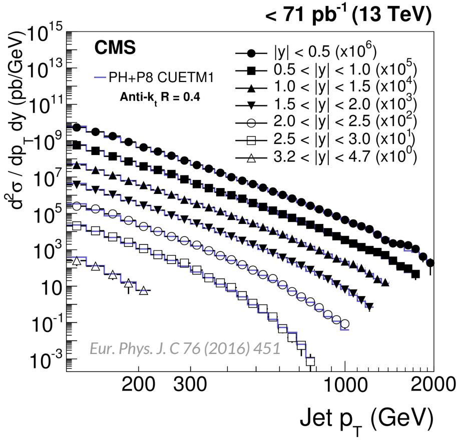
Different kinds of results

- **Measurements** of known processes: masses, cross sections, ratios, asymmetries

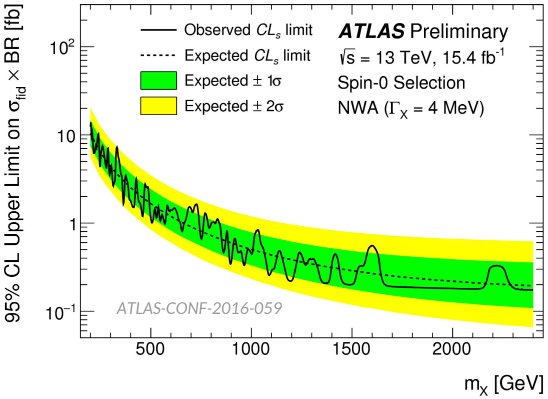
- **Searches** for hypothesised processes
  - Set **exclusion limits** if process is not observed
  - Quantify level of deviation from the standard model expectation, e.g. **p-value**

- **Many common software tools and methods used to produce these**

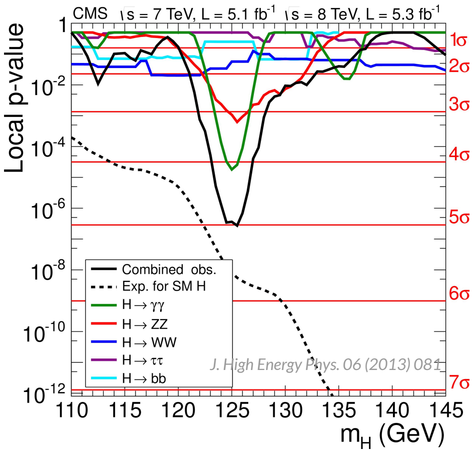




Eur. Phys. J. C 76 (2016) 451



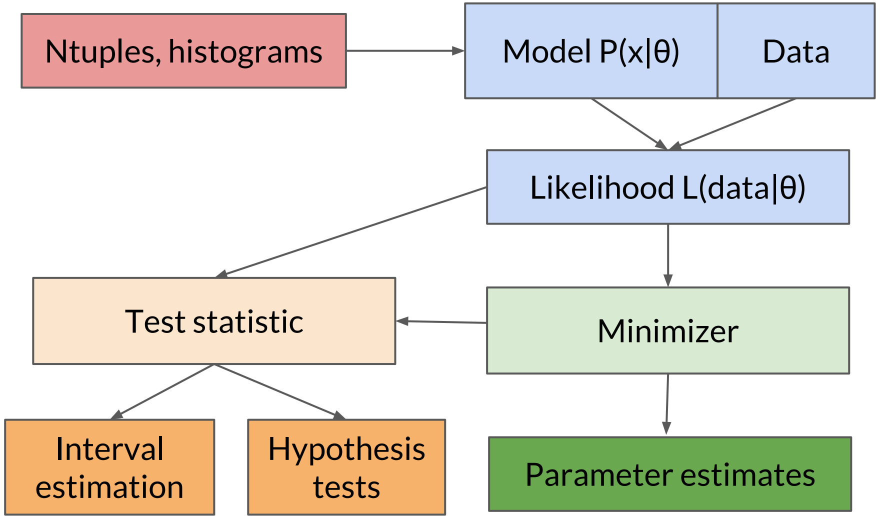
ATLAS-CONF-2016-059

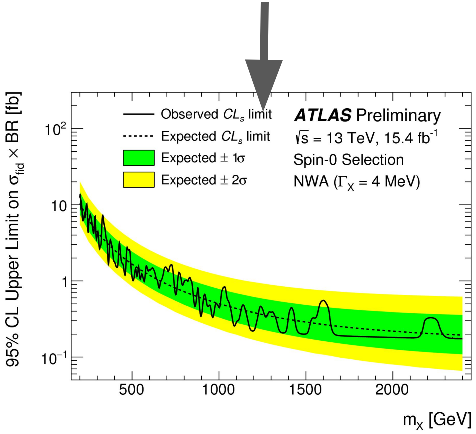


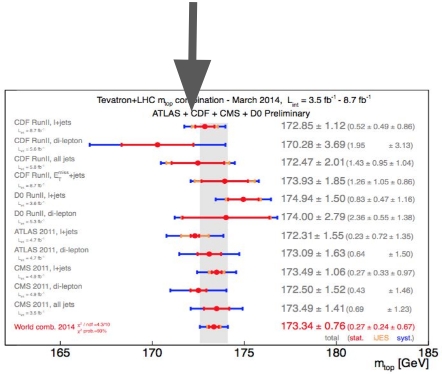
J. High Energy Phys. 06 (2013) 081

Software Tools

- Numerical results typically extracted via a fit of a model to the data
- Model expressed as a probability density function P(x|θ)
  - x = set of observed quantities
  - θ = parameters that alter the model prediction
- Model may be constructed by Monte Carlo simulation, analytic functions
- Define a likelihood function for the observed data given the model
- Use numerical minimization of the log-likelihood to estimate parameter values
- Test-statistics to distinguish between hypotheses and calculate intervals

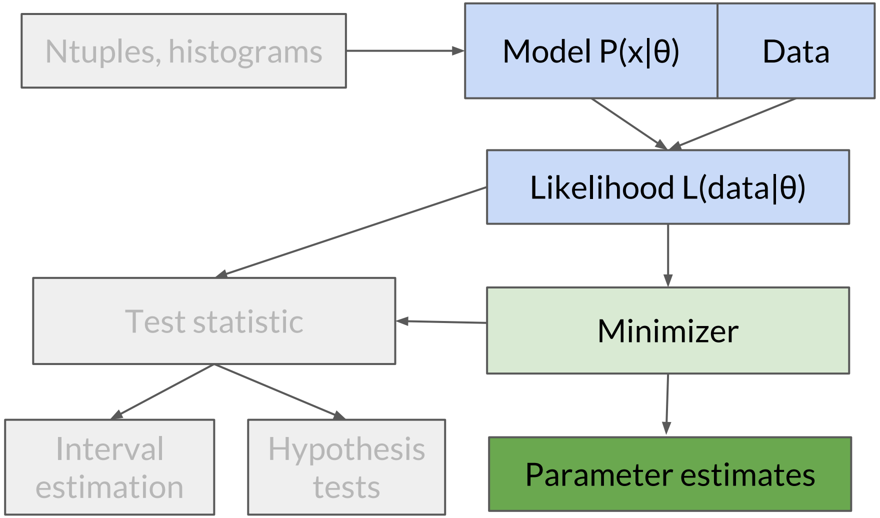






Software Tools

- The RooFit framework is used extensively to define the model, variables, data and likelihood function
- Each represented by C++ objects
- Many commonly-used PDFs included, can be used as building blocks for more complicated models. Also straightforward for users to write entirely new PDF classes
- RooFit can normalize PDFs, generate toy MC data and make plots for arbitrary models
- Interfaces with ROOT::Math::Minimizer for minimisation via Minuit (most common), simplex and other routines
- Provides the RooWorkspace container for persisting all model information in a ROOT file



Ntuples, histograms

Model P(x|θ)

Data

Likelihood L(data|θ)

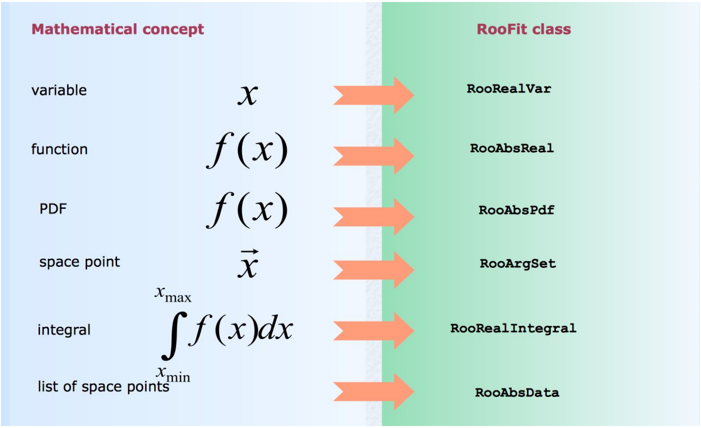
Test statistic

Minimizer

Interval estimation

Hypothesis tests

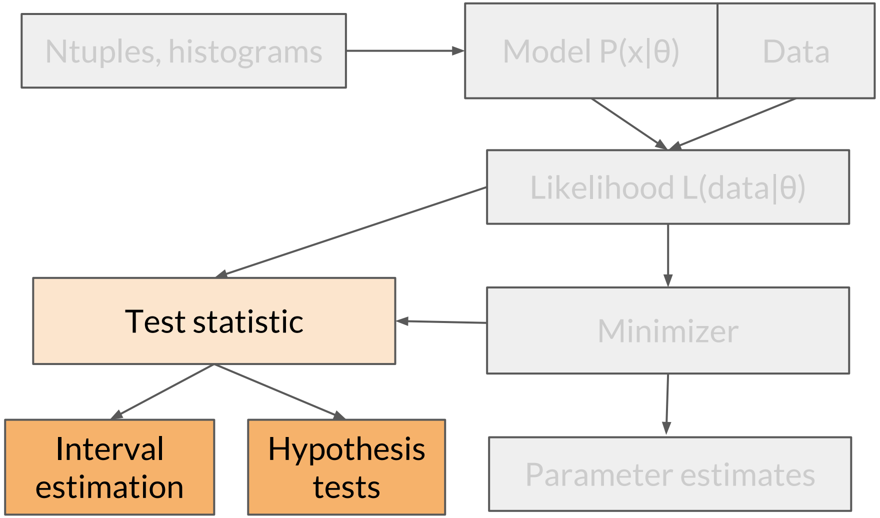
Parameter estimates



Mathematical concept		RooFit class
variable	x	RooRealVar
function	$f(x)$	RooAbsReal
PDF	$f(x)$	RooAbsPdf
space point	$\vec{x}$	RooArgSet
integral	$\int_{x_{\min}}^{x_{\max}} f(x)dx$	RooRealIntegral
list of space points		RooAbsData

Software Tools

- The RooStats framework is built on top of ROOT and RooFit to provide a range of statistical methods that can be applied to arbitrary RooFit models
- Implements commonly used interval calculators: e.g. Profile likelihood, Bayesian with support Markov-chain integration, Feldman-Cousins
- Also provides classes for hypothesis testing, e.g:
  - **Frequentist**, with built-in toy dataset evaluation to build the test-statistic distributions
  - **Asymptotic** - widely used by the LHC experiments as it avoids the computing-intensive step of generating and fitting toy datasets


Ntuples, histograms

Model P(x|θ)

Data

Likelihood L(data|θ)

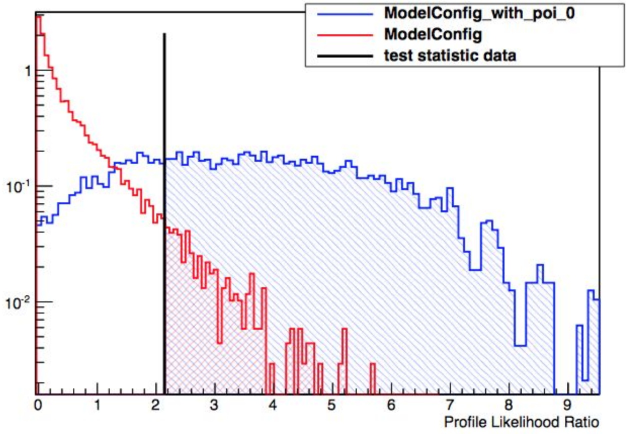
Test statistic

Minimizer

Interval estimation

Hypothesis tests

Parameter estimates



RooWorkspace

- Container class for RooFit objects that preserves links between variables and functions
- Can store all data, PDFs, uncertainties that are defined for an analysis
- Provides a convenient “factory” language for quickly defining new objects

```
RooWorkspace w;
w.factory(“Gaussian::g(x[-10,10],mean[-10,10],sigma[3])”);
```

- Allows a separation between producing the model and running statistical methods
  - Can be saved to a ROOT file
  - Possible to edit and merge workspaces for combinations

```
RooWorkspace w;
w.import(myPdf);
w.import(myData);
w.writeToTfile(“workspace.root”);
```

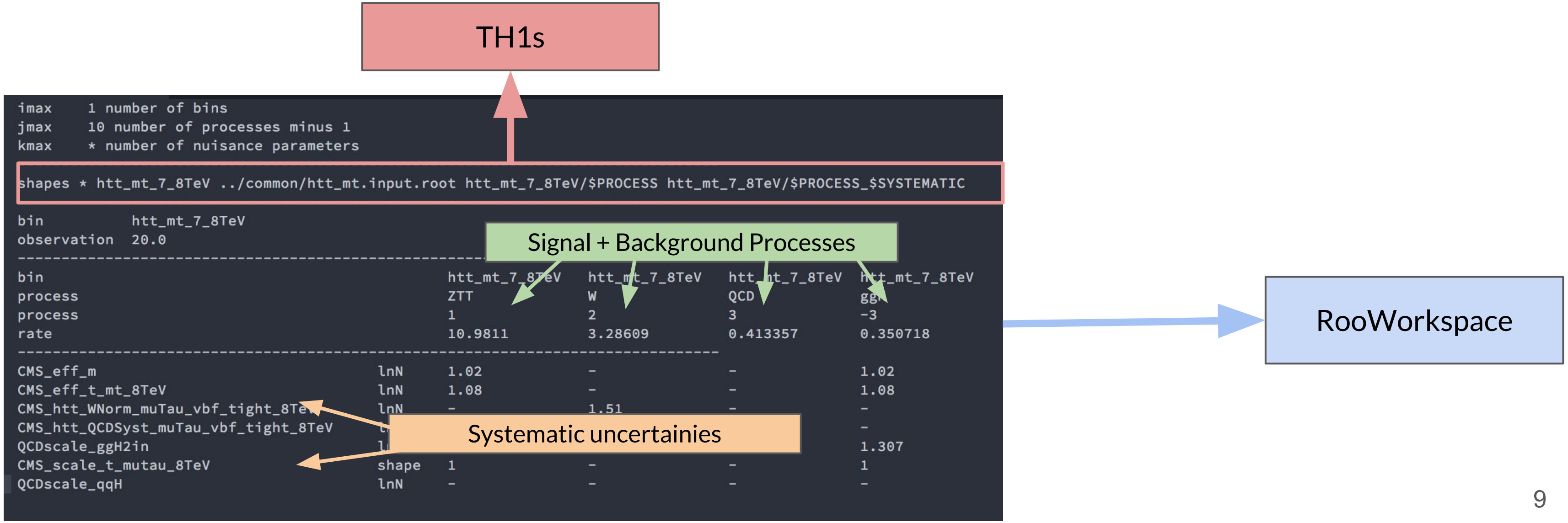
Model Building

- A typical analysis today can contain O(10) channels, each with O(10) processes, and O(100) systematic uncertainties
- Uncertainties may change both the normalisation and shape of the expected distributions
- Assembling the RooFit model “by-hand” in each case laborious and repetitive
- ⇒ **Use higher-level tools to automate the construction of the model**

- **RooStats** includes the HistFactory tool - configuration of template-based models in C++ or XML
- Experiments also build frameworks to automate model construction and perform common statistical tasks:
  - HistFitter: originally developed for ATLAS supersymmetry searches. Built on top of RooFit, RooStats and HistFactory. Provides complete framework for model construction, fitting and hypothesis testing and presentation of results.
  - **Combine**: Used extensively in CMS. Originally developed within the Higgs group but now used widely for SM, top, SUSY and exotic searches. Provides datacard format for specifying models, python classes for applying signal parameterisation, simple interface for running RooStats methods and additional fit diagnostics.

Example: the CMS datacard format

- Users write plain-text datacards describing: channels, data, contributing processes, systematic uncertainties
- Cards can represent self-contained counting experiments or refer to pre-existing TH1s or RooFit PDFs for building shape analyses
- Datacards can easily be combined before processing to make the RooWorkspace



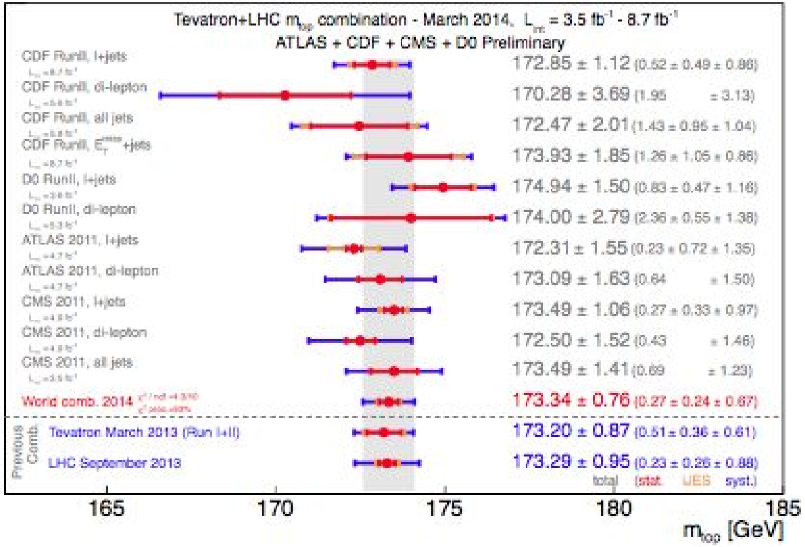
```
imax    1 number of bins
jmax    10 number of processes minus 1
kmax    * number of nuisance parameters

shapes * htt_mt_7_8TeV ../common/htt_mt.input.root htt_mt_7_8TeV/$PROCESS htt_mt_7_8TeV/$PROCESS_$SYSTEMATIC

bin          htt_mt_7_8TeV
observation  20.0
-----------------------------------------------
bin                                              htt_mt_7_8TeV   htt_mt_7_8TeV   htt_mt_7_8TeV  htt_mt_7_8TeV
process                                          ZTT             W               QCD            gg
process                                          1               2               3              -3
rate                                             10.9811         3.28609         0.413357       0.350718
--------------------------------------------------------------------------------
CMS_eff_m                               lnN      1.02            -               -              1.02
CMS_eff_t_mt_8TeV                       lnN      1.08            -               -              1.08
CMS_htt_WNorm_muTau_vbf_tight_8Te       lnN      -               1.51            -              -
CMS_htt_QCDSyst_muTau_vbf_tight_8TeV                                                            -
QCDscale_ggH2in                                                                                 1.307
CMS_scale_t_mutau_8TeV                  shape    1               -               -              1
QCDscale_qqH                            lnN      -               -               -              -
```

Combination methodology

- Often multiple analyses/experiments measuring or searching for the same signals
- Best sensitivity or smallest uncertainty achieved via a combined measurement
- Higgs discovery with 5σ significance only possible in July 2012 because searches in different decay channels were readily combinable
- **Ideal approach**: the most rigorous method for combining two analyses is to combine the individual likelihoods:
  - Given $L(n_A \mid \theta_A)$ and $L(n_B \mid \theta_B)$ construct $L(n_A, n_B \mid \theta_A \cup \theta_B)$
  - Where there may be some common parameters between sets A and B
- In practice not always possible:
  - Different software used to encode likelihood
  - Requires common signal parameterization and consistent treatment of common systematic uncertainties
- Combinations at the likelihood-level within experiments are commonplace
  - At the LHC greatly facilitated by widespread use of RooFit and workspaces
- When likelihood combination not feasible, can combine measurements directly

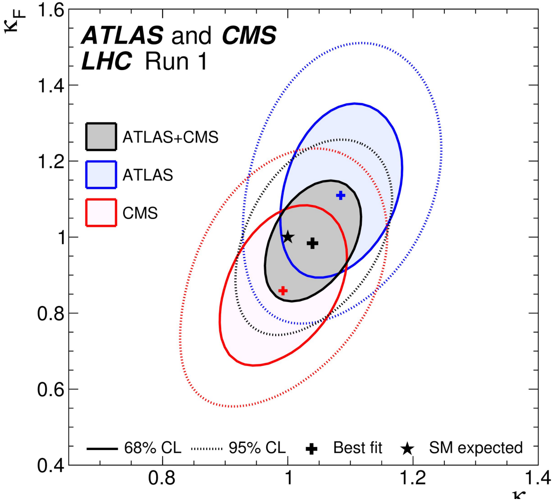
Combination of measurements

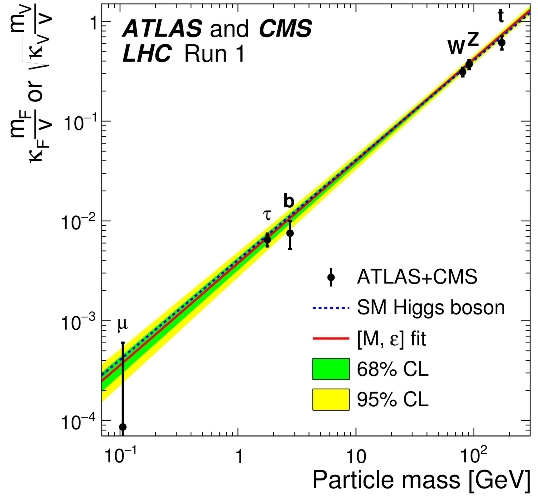
- **Example**: LHC+Tevatron combined top mass measurement
- Using the BLUE (Best Linear Unbiased Estimate) framework
- Used to combine a number of estimates for a singl observable
- Determines coefficients for a linear of combinatio of input measurements by minimising total uncertainties on the combined result
- Assumes all uncertainties are described by Gaussian PDFs
- Takes statistical and systematic uncertainties into account as well as correlations in the latter between the two measurements



ATLAS+CMS Higgs Combination

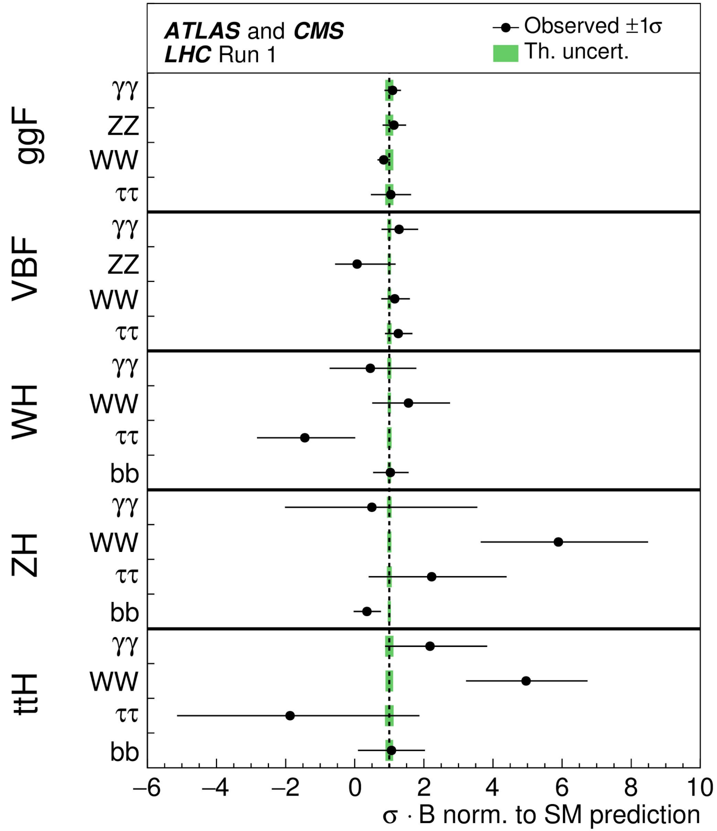
- Combined analysis of Run 1 data to extract couplings and signal strengths
- Results produced by combining RooFit workspaces
- Made possible by early agreement in 2011 by the LHC Combination Group for common treatment of systematic uncertainties
- **Combined workspace facts and figures:**
    - 62k data points
    - 12k function objects
    - 4300 nuisance parameters (many related to finite MC statistics)
    - Minuit successfully able to minimize the combined likelihood function: ~ 1-2 hours per fit
- Full set of results required O(30k) individual fits
    - Achieved fast turnaround by running on the grid - ideal since CPU-dominated task with minimal I/O - can run at any site





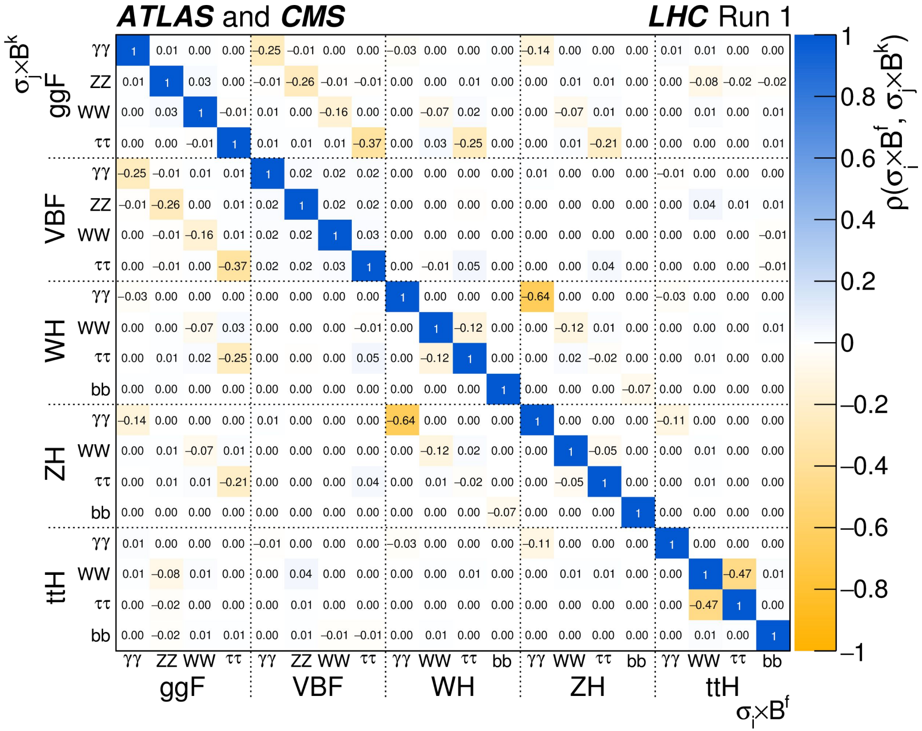
ATLAS+CMS Higgs Coupling Combination

- Results presented for ~ 20 different signal parameterisations
- But impossible to cover every interesting model that exists now, let alone what might be devised in the future
- How can results be reinterpreted in the future?
- Provide results in the most general model possible: signal strength for each combination of Higgs production and decay mode



ATLAS+CMS Higgs Coupling Combination

- But the measurements + uncertainties are not enough!
- Not possible to distinguish between all possible signal processes processes with current dataset and analysis selections ⇒ correlations are important
- Also publish the correlation matrix
- In principle results with more constrained parametrisations could be reproduced from this
- In practice there are limitations - only an approximation of 21-dimensional likelihood function



		ggF				VBF				WH				ZH				ttH			
		γγ	ZZ	WW	ττ	γγ	ZZ	WW	ττ	γγ	WW	ττ	bb	γγ	WW	ττ	bb	γγ	WW	ττ	bb
ggF	γγ	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
	ττ	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	γγ	–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
	ττ	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	γγ	–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
	ττ	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	γγ	–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
	ττ	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	γγ	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
	ττ	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

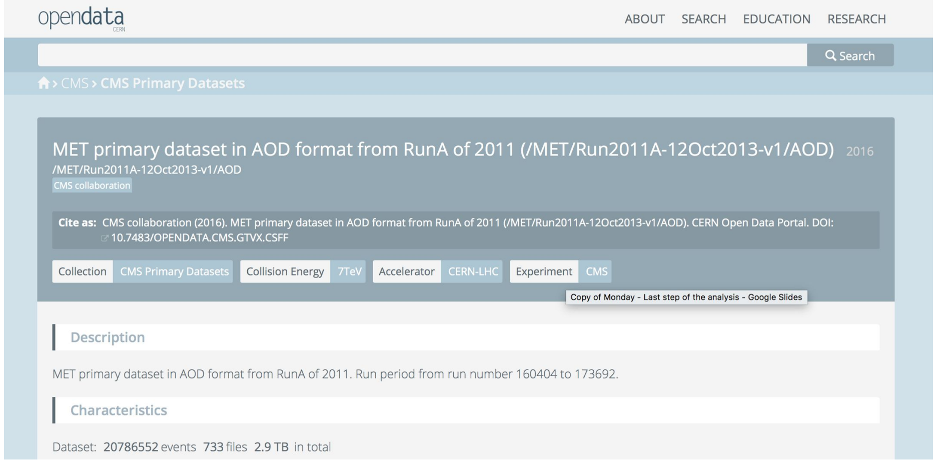
Analysis reinterpretation

- Different motivations / needs for sharing data
- Typical case: phenomenologists want to test if new model is excluded by an analysis
- Will only discuss a few approaches here, many more are used and under development
- See LHC forum on BSM results interpretation for a larger set of software tools in use

- Main approaches:
    - Release the event data itself, allows for entirely new analysis outside the collaborations
    - Publish measurements and limits in such a way that they can be directly reinterpreted, i.e. exclude process **X** with a cross section above **Y pb**. Phenomenologists need only calculate cross section of **X** in their favourite model.
    - Publish simplified information about the likelihood, such that other signal expectations could be inserted
    - Rarely done: publish the full likelihood model (i.e. the RooWorkspace)

CERN OpenData

- Initiative to release reconstructed LHC collision data & MC simulation for public analysis
- ATLAS, LHCb, CMS and ALICE have all released example data primarily for education purposes
- CMS has also released ~ 300 TB of √s = 7 TeV data + VM containing software needed to analyse it



MET primary dataset in AOD format from RunA of 2011 (/MET/Run2011A-12Oct2013-v1/AOD) 2016

/MET/Run2011A-12Oct2013-v1/AOD

CMS collaboration

Cite as: CMS collaboration (2016). MET primary dataset in AOD format from RunA of 2011 (/MET/Run2011A-12Oct2013-v1/AOD). CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.GTVX.CSFF

Collection: CMS Primary Datasets | Collision Energy: 7TeV | Accelerator: CERN-LHC | Experiment: CMS

Copy of Monday - Last step of the analysis - Google Slides

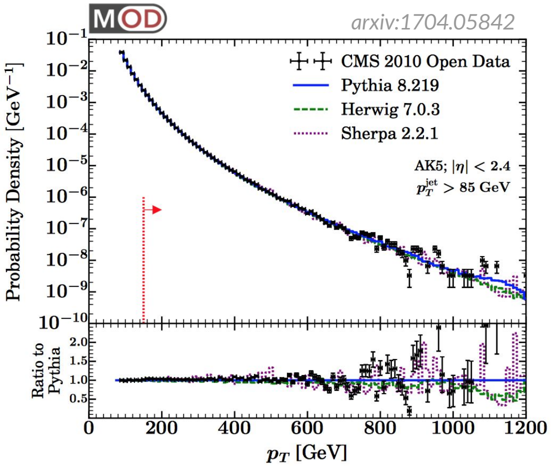
Description

MET primary dataset in AOD format from RunA of 2011. Run period from run number 160404 to 173692.

Characteristics

Dataset: 20786552 events 733 files 2.9 TB in total

- A first analysis using the OpenData has already appeared

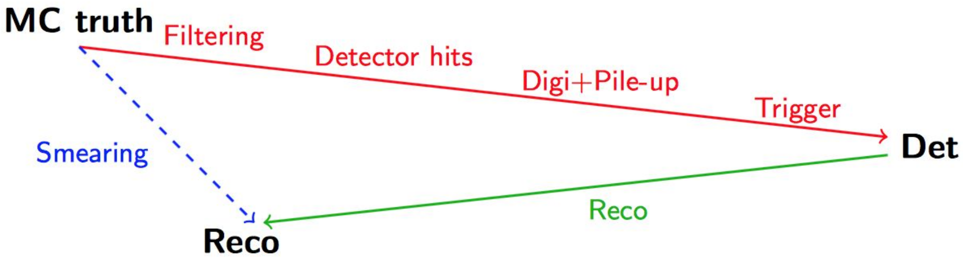


arxiv:1704.05842

- Open publication of the event-level data offers greatest potential for new interpretations and searches - but also the steepest learning curve

Rivet

- Framework for analysing and validating events produced by MC generators
- Analysts provide “rivet routines” defining selections and observables that can be applied to events in the HepMC format
- Until recently mainly used by SM measurements as required unfolded distributions of observables for comparison - unfolding not typically used in BSM searches
- Recently possible to add smearing functions of MC truth information to approximate detector response & reconstruction:



MC truth

Filtering

Detector hits

Digi+Pile-up

Trigger

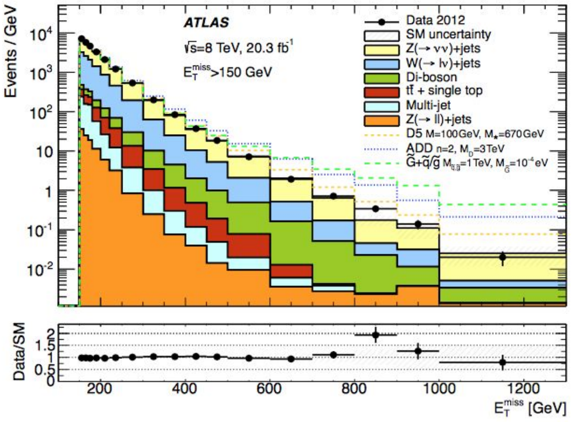
Det

Smearing

Reco

Reco

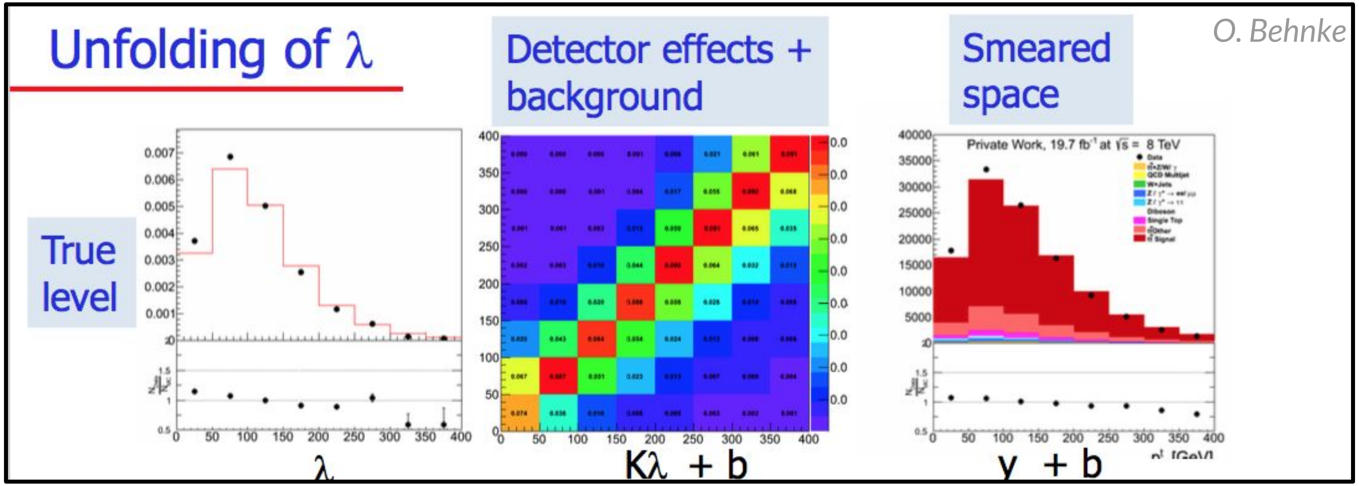
- Increasing numbers of BSM routines published by ATLAS and CMS, e.g. ATLAS monojet search



Eur. Phys. J. C (2015) 75

Unfolding

- General problem of un-smearing a distribution from the measured to the truth level

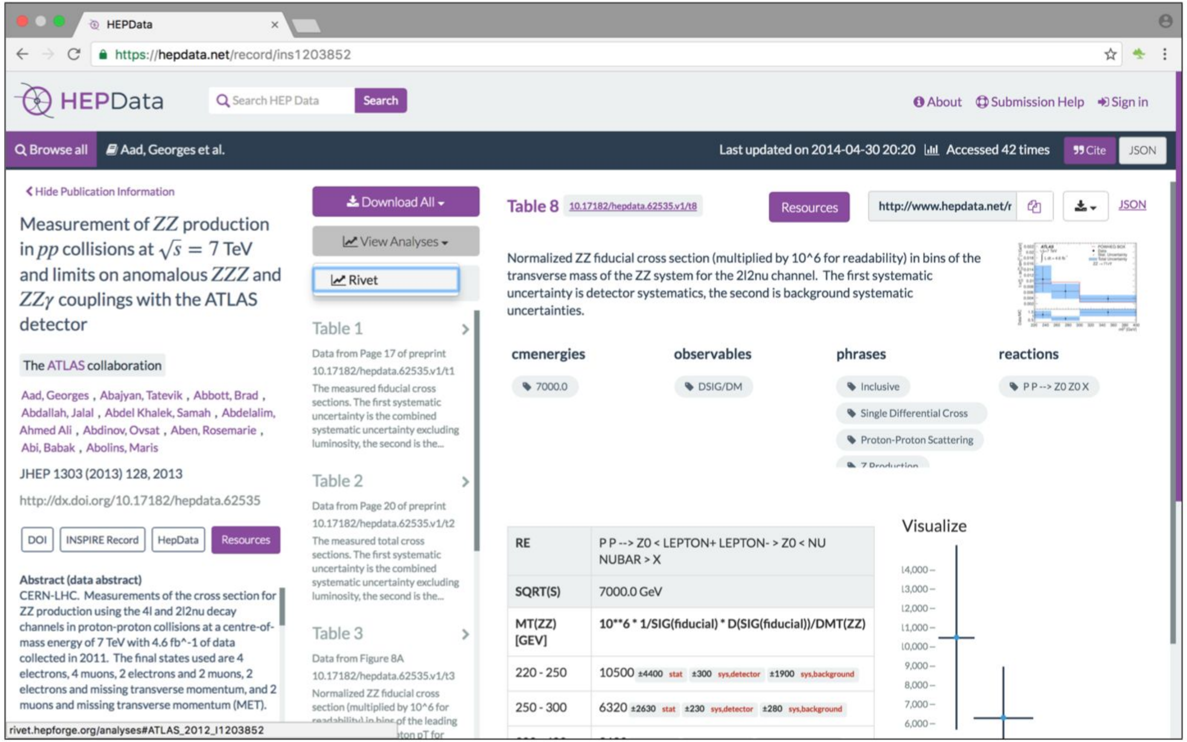


- Two packages in common use: TUnfold and RooUnfold

- Both support propagation of statistical uncertainties through the unfolding procedure, as well as common regularisation methods: e.g. D'Agostini iteration, singular value decomposition, Tikhonov

HEPData

- Online archive of the raw data points and values contained in HEP publication figures
- Avoids the need for extracting numbers from the image files and facilitates the the reinterpretation of results
- Data can visualised online as well as exported in different format e.g. CSV, ROOT, YAML



Measurement of ZZ production in pp collisions at $\sqrt{s} = 7$ TeV and limits on anomalous ZZZ and $ZZ\gamma$ couplings with the ATLAS detector

The ATLAS collaboration

Aad, Georges , Abajyan, Tatevik , Abbott, Brad , Abdallah, Jalal , Abdel Khalek, Samah , Abdelalim, Ahmed Ali , Abdinov, Ovsat , Aben, Rosemarie , Abi, Babak , Abolins, Maris

JHEP 1303 (2013) 128, 2013

http://dx.doi.org/10.17182/hepdata.62535

Abstract (data abstract)
CERN-LHC. Measurements of the cross section for ZZ production using the 4l and 2l2nu decay channels in proton-proton collisions at a centre-of-mass energy of 7 TeV with 4.6 fb^-1 of data collected in 2011. The final states used are 4 electrons, 4 muons, 2 electrons and 2 muons, 2 electrons and missing transverse momentum, and 2 muons and missing transverse momentum (MET).

Table 1
Data from Page 17 of preprint
10.17182/hepdata.62535.v1/t1
The measured fiducial cross sections. The first systematic uncertainty is the combined systematic uncertainty excluding luminosity, the second is the...

Table 2
Data from Page 20 of preprint
10.17182/hepdata.62535.v1/t2
The measured total cross sections. The first systematic uncertainty is the combined systematic uncertainty excluding luminosity, the second is the...

Table 3
Data from Figure 8A
10.17182/hepdata.62535.v1/t3
Normalized ZZ fiducial cross section (multiplied by 10^6 for readability) in bins of the leading

Table 8 10.17182/hepdata.62535.v1/t8

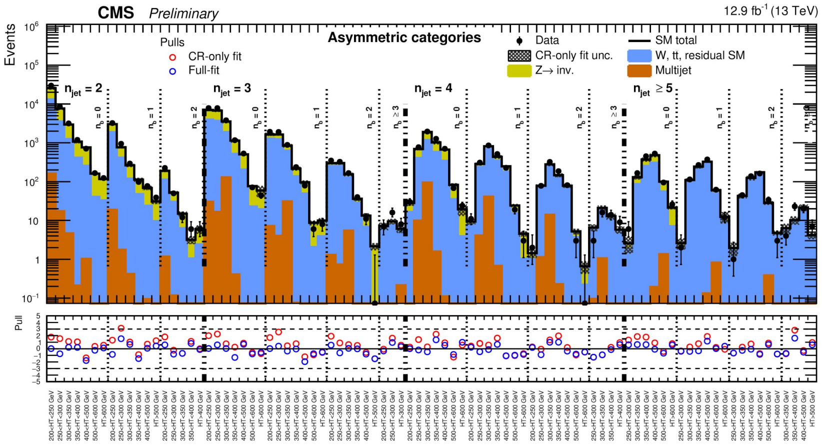
Normalized ZZ fiducial cross section (multiplied by 10^6 for readability) in bins of the transverse mass of the ZZ system for the 2l2nu channel. The first systematic uncertainty is detector systematics, the second is background systematic uncertainties.

cmenergies	observables	phrases	reactions
7000.0	DSIG/DM	Inclusive Single Differential Cross Proton-Proton Scattering Z Production	P P --> Z0 Z0 X

RE	P P --> Z0 < LEPTON+ LEPTON- > Z0 < NU NUBAR > X
SQRT(S)	7000.0 GeV
MT(ZZ) [GEV]	10**6 * 1/SIG(fiducial) * D(SIG(fiducial))/DMT(ZZ)
220 - 250	10500 ±4400 stat ±300 sys,detector ±1900 sys,background
250 - 300	6320 ±2630 stat ±230 sys,detector ±280 sys,background

Simplified Likelihood approach

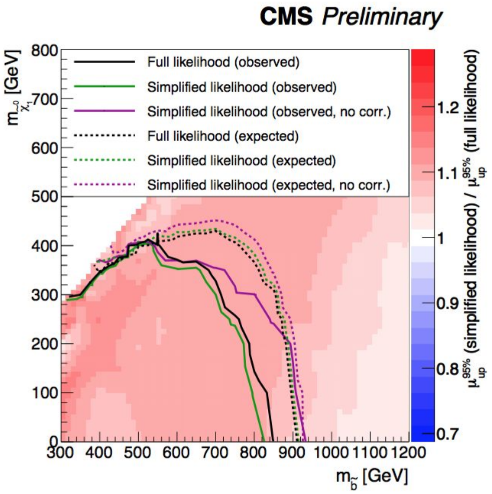
- Defining model-independent limits not always possible
- BSM physics searches typically use large number of bins/categories, exploiting the shape information of multiple variables
- Instead of full likelihood, provide an approximate one using the covariance matrix for background yields of each bin
  - Encodes statistical and systematic uncertainties of the full likelihood and correlations between bins



CMS-PAS-SUS-16-016

Example: CMS Hadronic SUSY search

SL yields accurate approximation of full model-dependent limits



CMS-NOTE-2017/001

Summary

- **For statistical analysis and combinations within and between experiments:**
  - RooFit + RooStats recommended for the flexibility in defining models, ease of likelihood-level combination and persistence via RooWorkspace
- **Identify features of high-level frameworks that could be shared more widely**
- **O(1000) parameter fits are becoming increasingly common**
  - Is there scope for performance gains here? In the NLL evaluation and/or MINUIT algorithm
- No one-size-fits-all solution for the **reinterpretation of results**, but facilities like Rivet and HEPdata should be used where possible
  - Release of likelihoods (simplified or full) could benefit from a common software framework for running fits and creating signal parameterisations

Backup