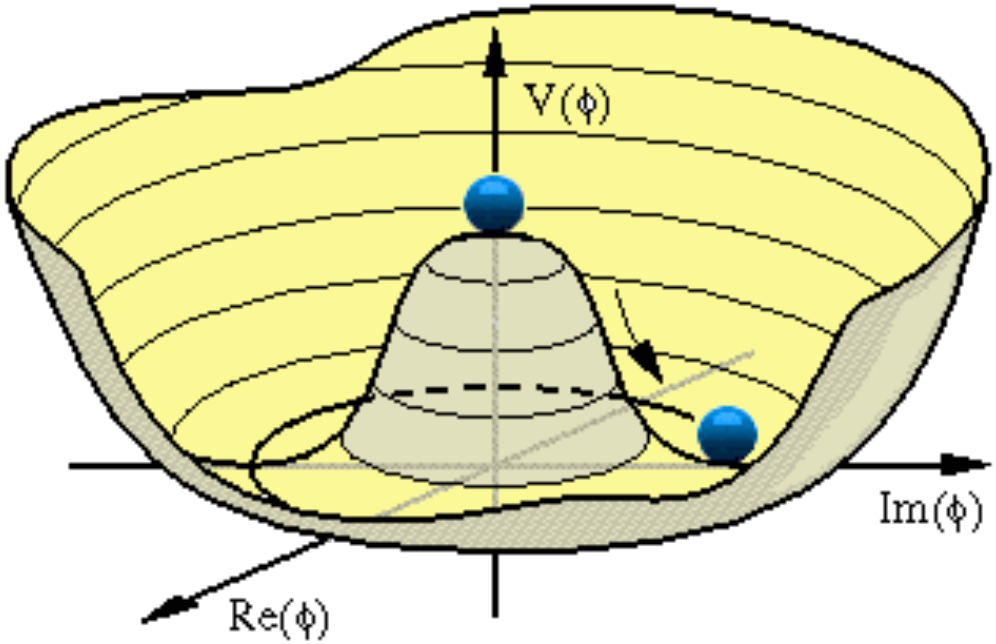
Data Analysis and Statistical Methods in Experimental Particle Physics



Thomas R. Junk
Fermilab

TRISEP 2014
Sudbury, Ontario

With public material from the ATLAS and CMS experiments at CERN, and the CDF and D0 experiments at the Tevatron

Lecture Outline

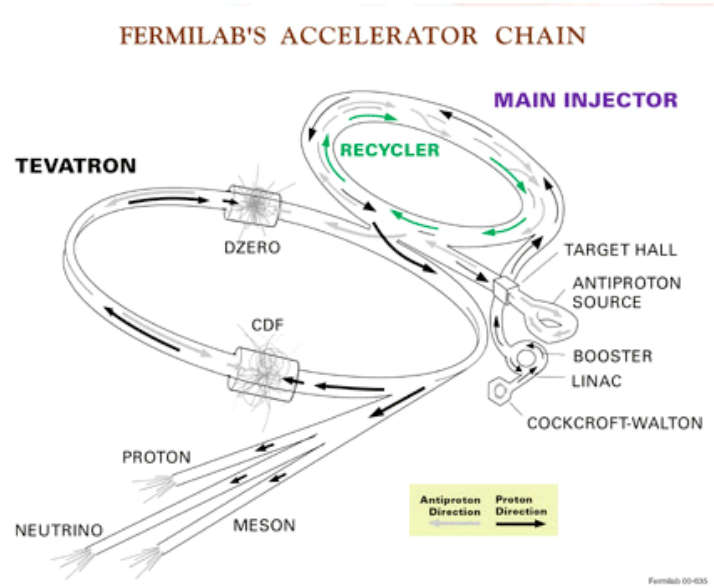
Lecture 1: Introduction, Gaussian Approximations

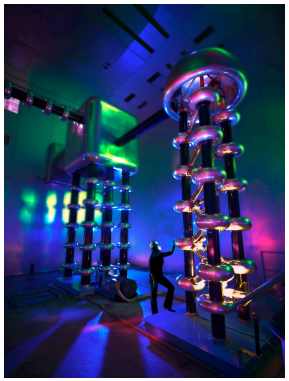
Lecture 2: Data Analysis and Systematic Uncertainties

Lecture 3: Frequentist and Bayesian Interpretation

Lecture 4: Higgs Boson Searches, Discovery, and Measurements

The Tevatron and Associated Accelerators





Cockroft-Walton

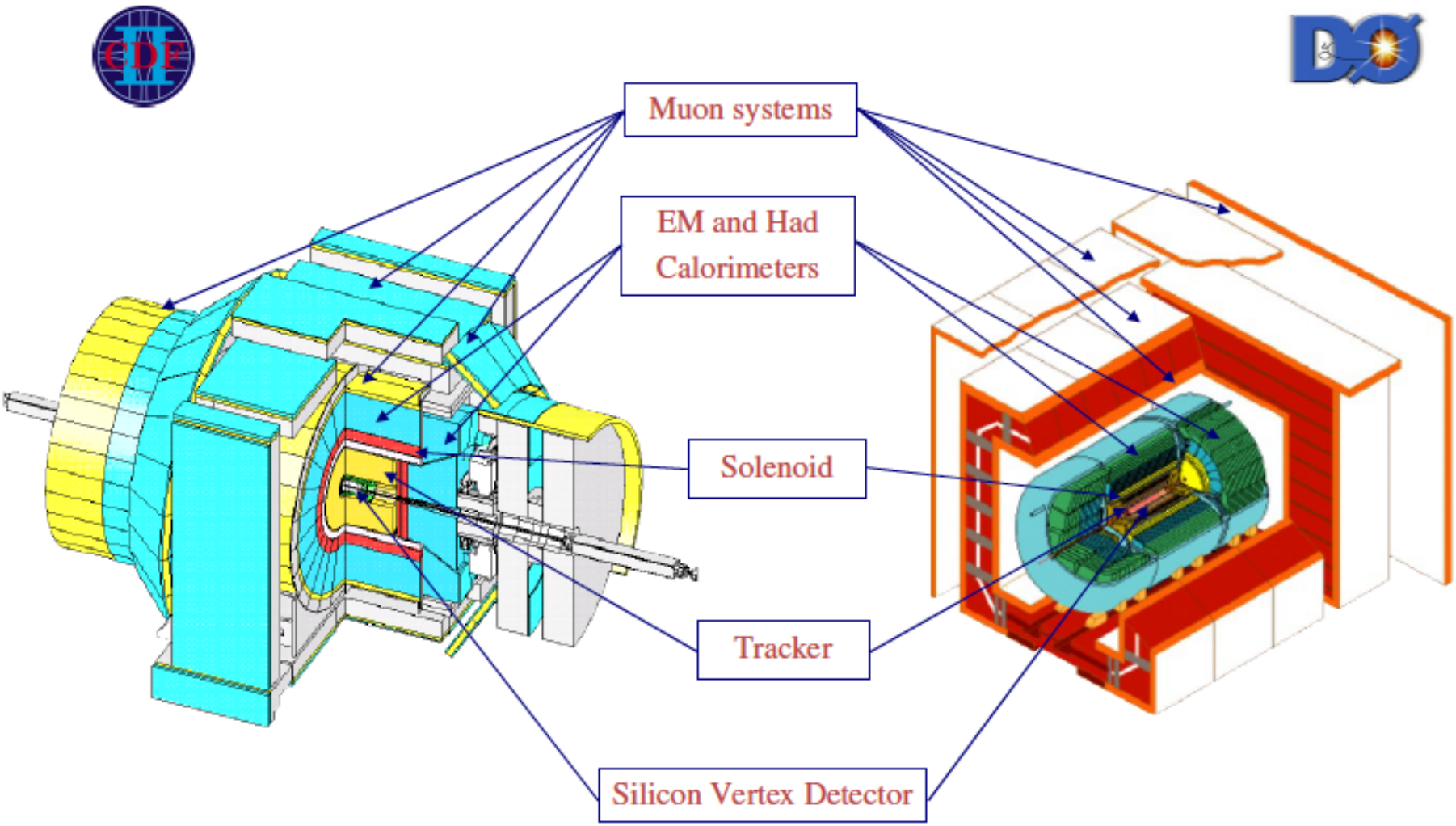
Linac

Booster

pbar debuncher and accumulator

ppbar collisions at 1.96 TeV
Luminosity up to 400E30 $cm^{-2}s^{-1}$
A very good week: ~80 pb^{-1}

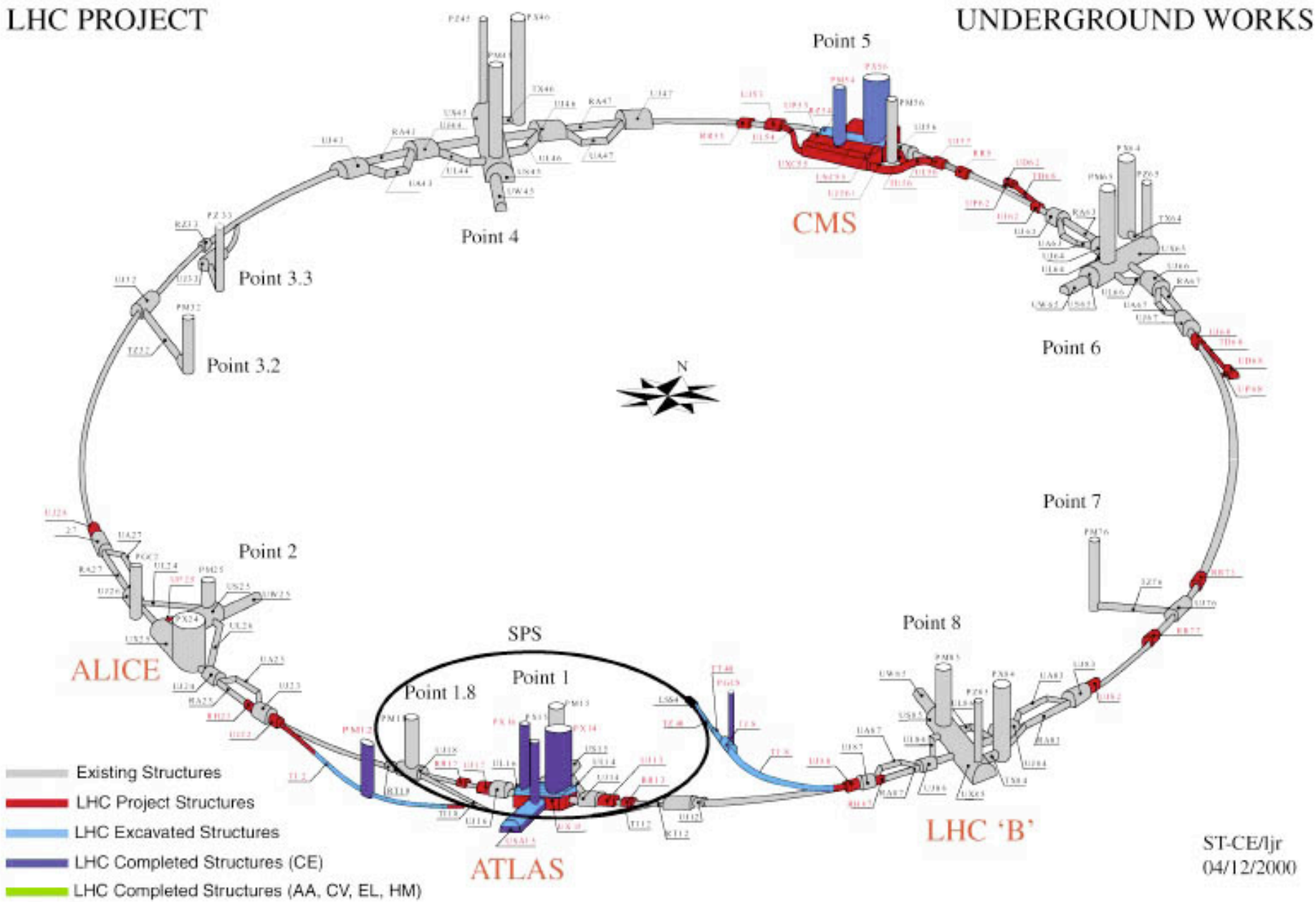
The CDF and D0 Detectors



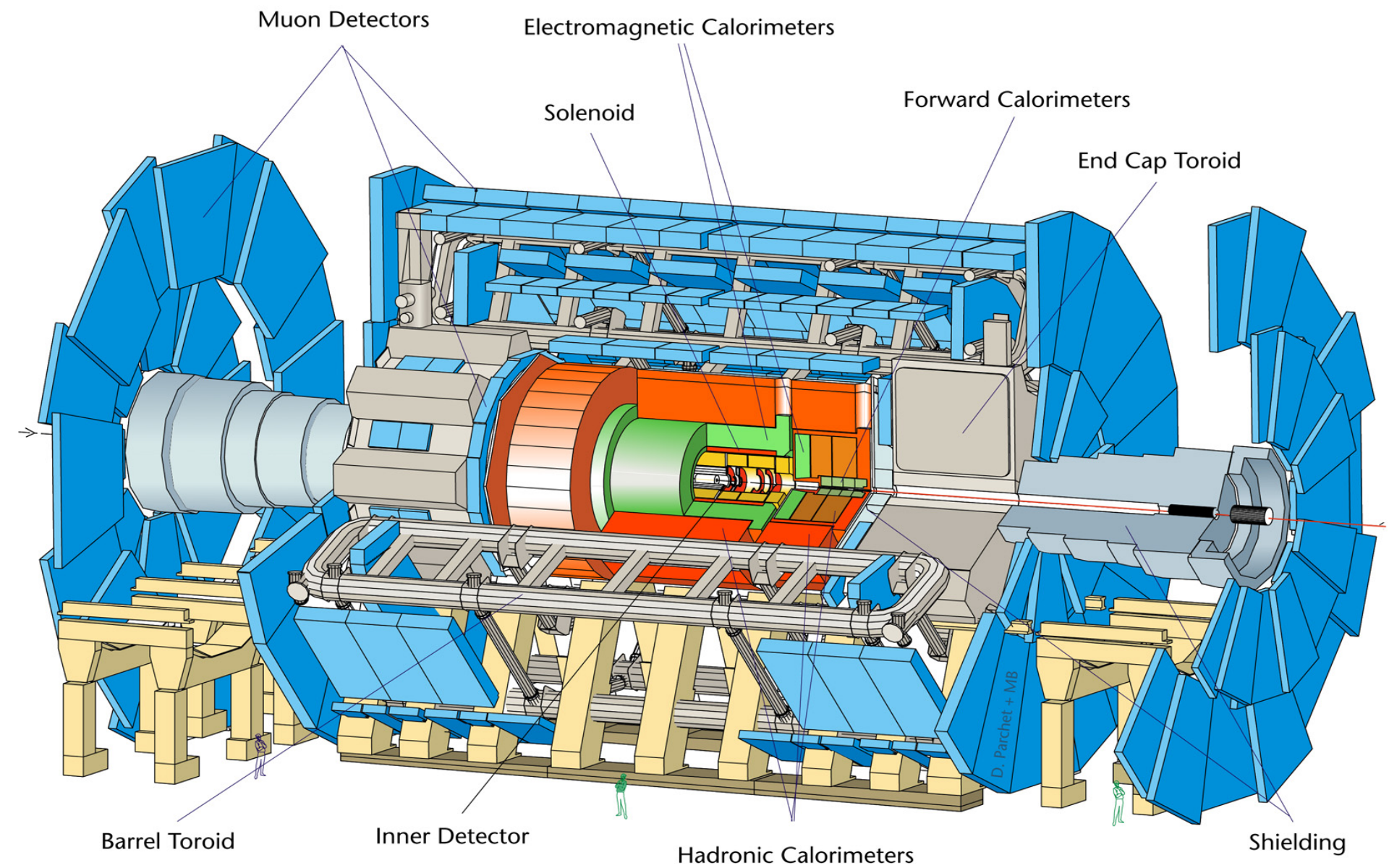
The Large Hadron Collider at CERN from the air

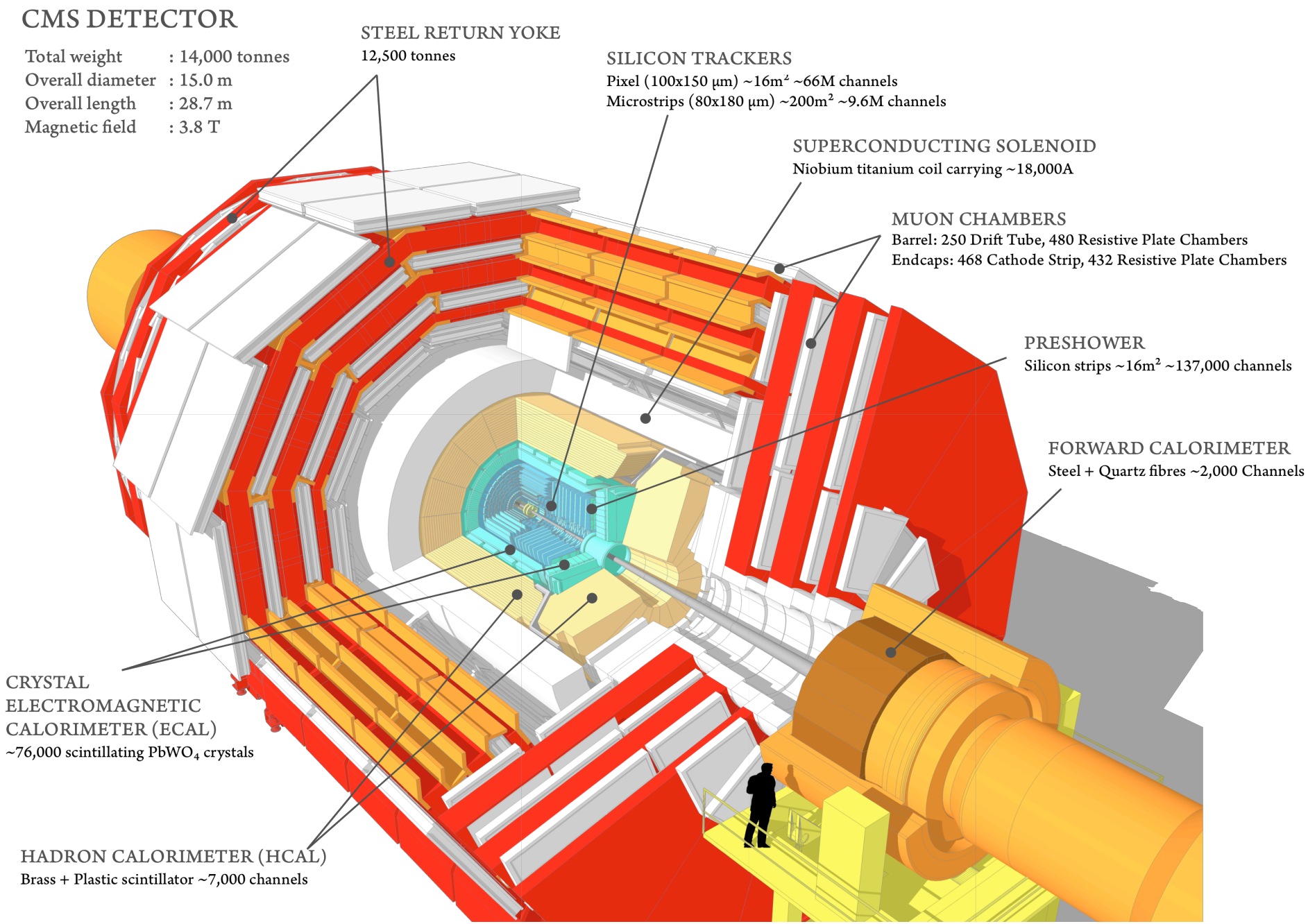
Re-uses the 27 km LEP tunnel

Strong superconducting magnets for pp collisions at 7, 8, and 13-14 TeV

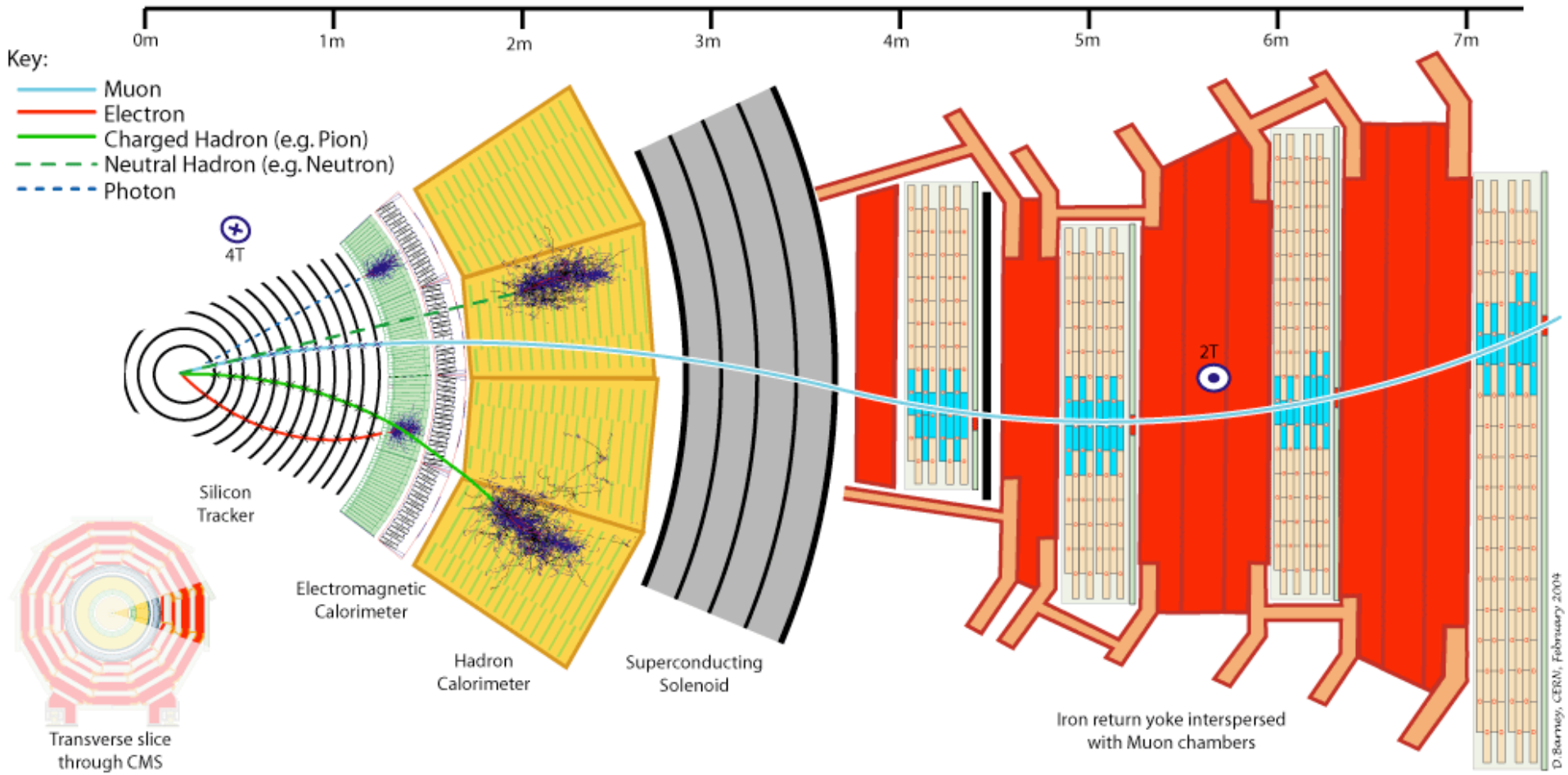

LHC PROJECT
UNDERGROUND WORKS
Point 5
CMS
Point 4
Point 3.3
Point 3.2
Point 6
Point 7
Point 2
ALICE
SPS
Point 1
Point 1.8
Point 8
LHC 'B'
ATLAS
N
Existing Structures
LHC Project Structures
LHC Excavated Structures
LHC Completed Structures (CE)
LHC Completed Structures (AA, CV, EL, HM)
ST-CE/ljr
04/12/2000

The ATLAS Detector at the LHC



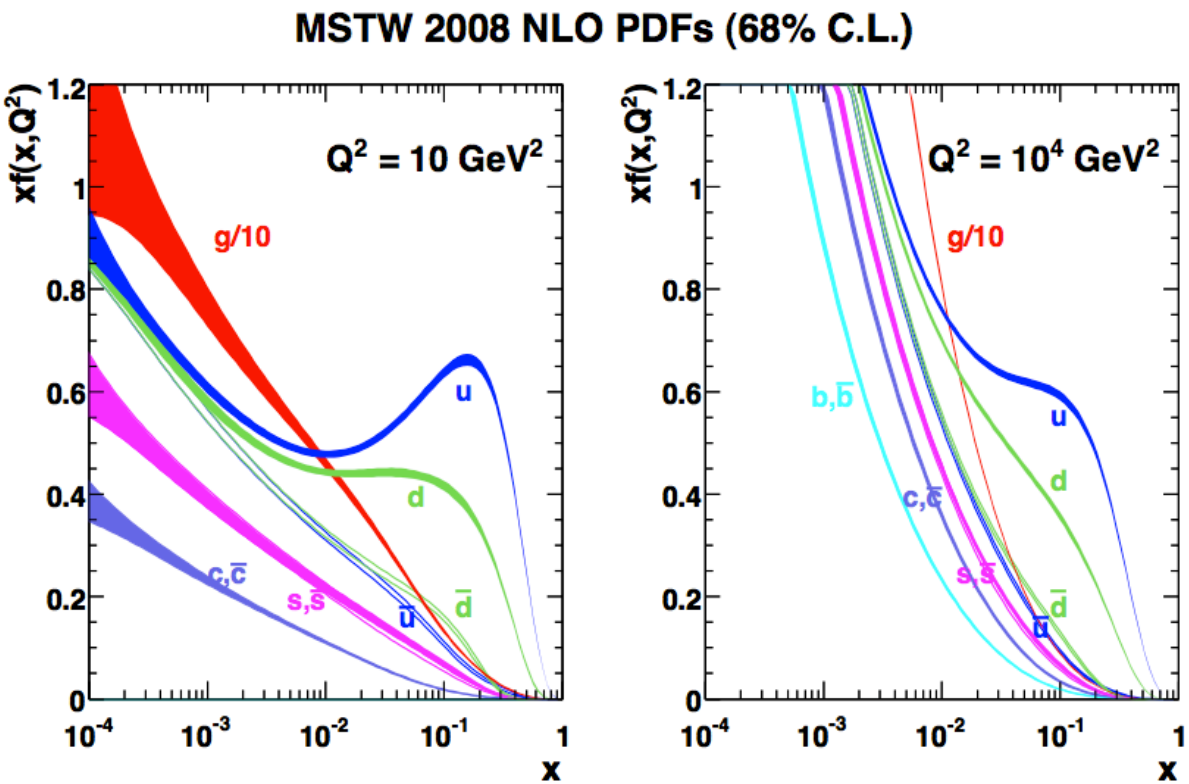

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO4 crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

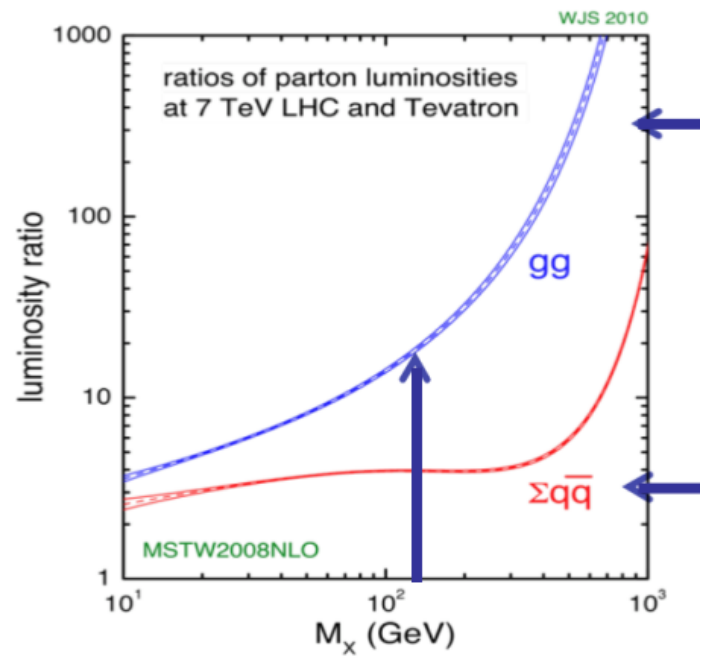
A Slice Through CMS



Parton Distribution Functions

Two effects: LHC has larger beam energy, thus larger Q^2, and also it's *pp* and Tevatron is $p\bar{p}$

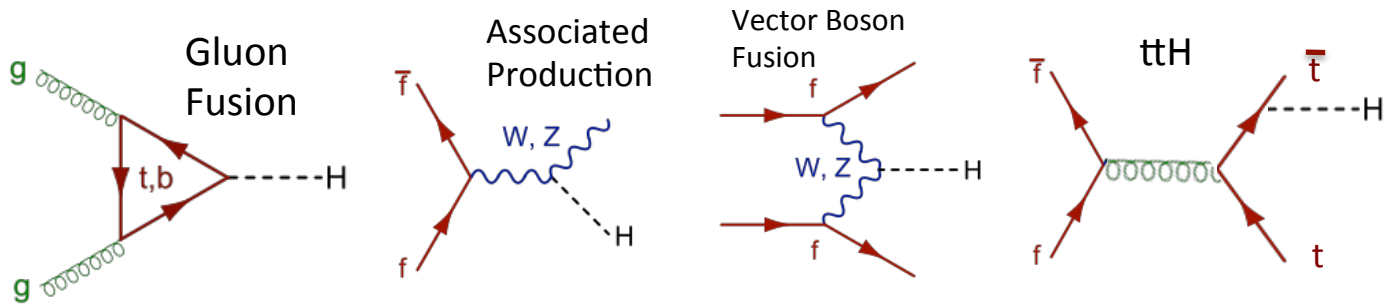


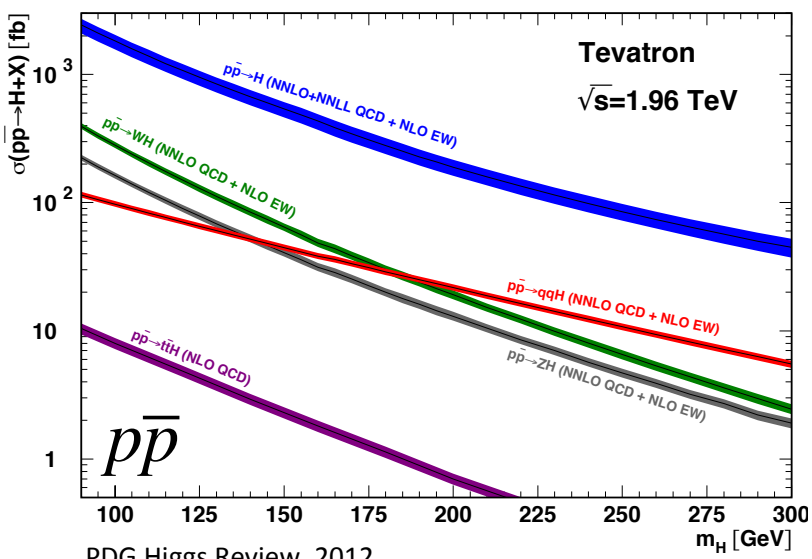


Tevatron Energies

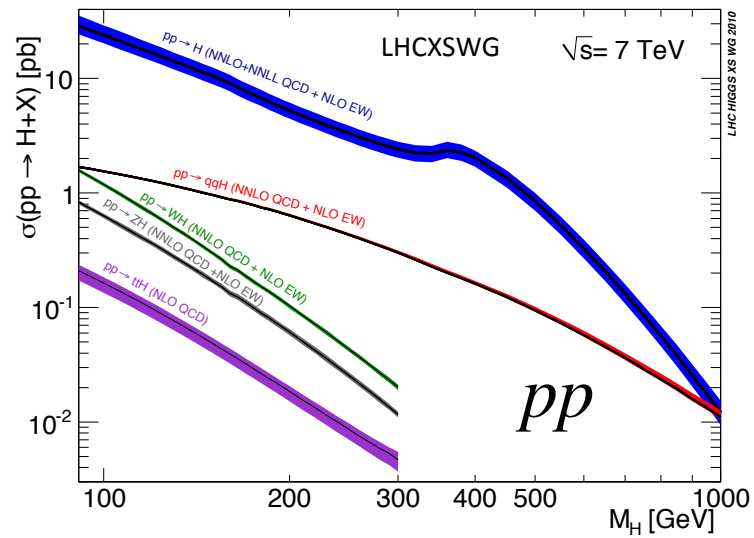
LHC Energies

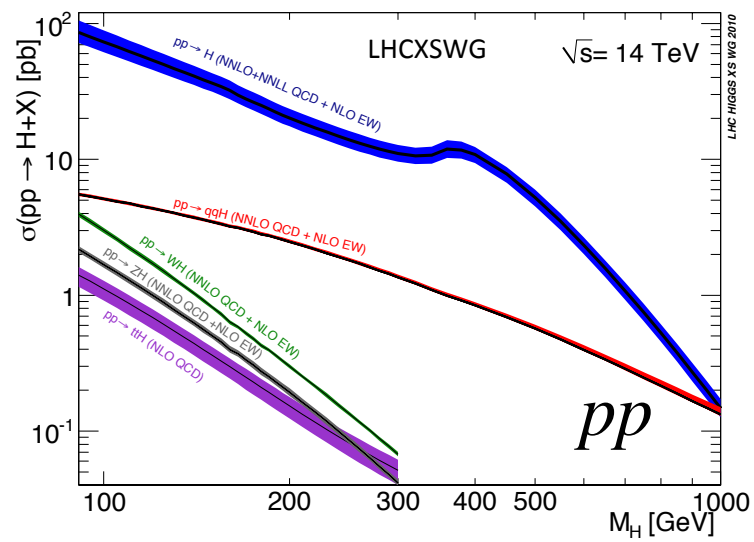
SM Higgs Boson Production Processes and Cross Sections



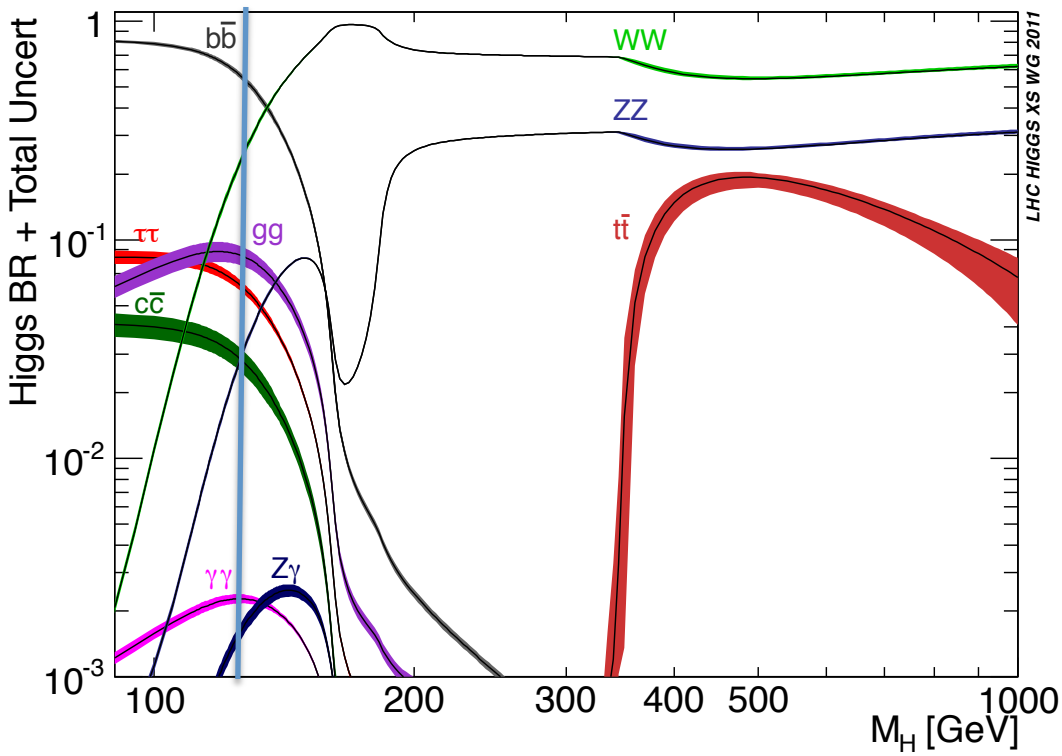


PDG Higgs Review, 2012





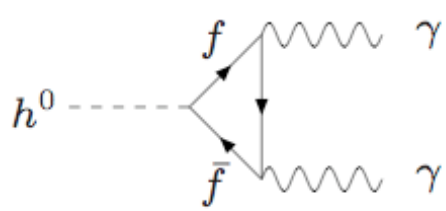
SM Higgs Boson Decay Branching Ratios

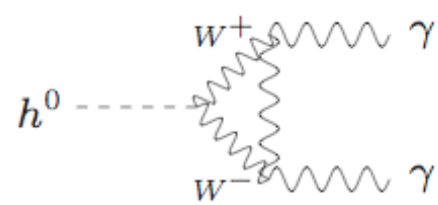


At m_H=125 GeV,

Br(H→bb) ≈ 58%
Br(H→ττ) ≈ 6.4%
Br(H->WW) ≈ 22%
Br(H→ZZ) ≈ 2.7%
Br(H->γγ) ≈ 0.23%

The main contributions to Γ(H->γγ):
a top loop and a W loop. Since one is a fermion and the other a boson, they interfere destructively.
The W loop is dominant.
New heavy particles could contribute! Adding to or subtracting from Γ(H->γγ)

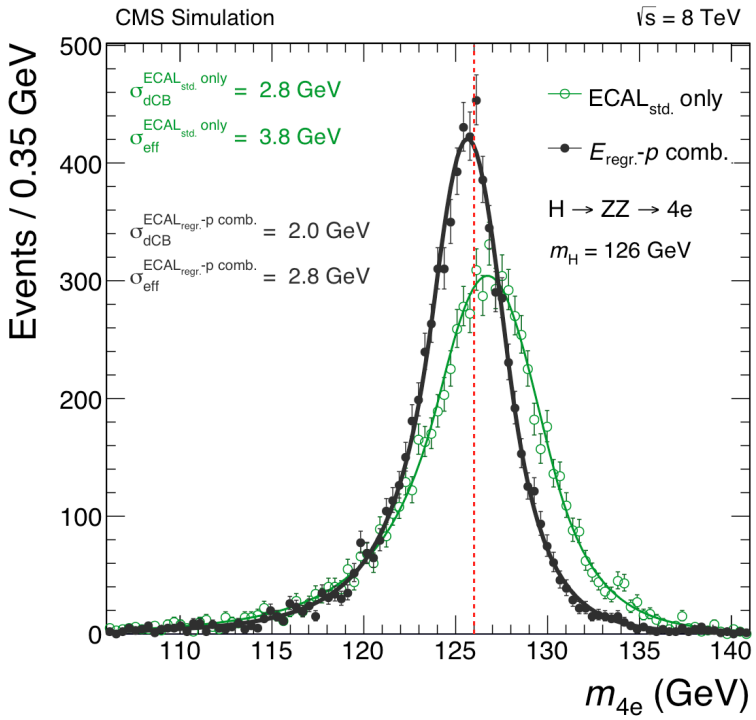


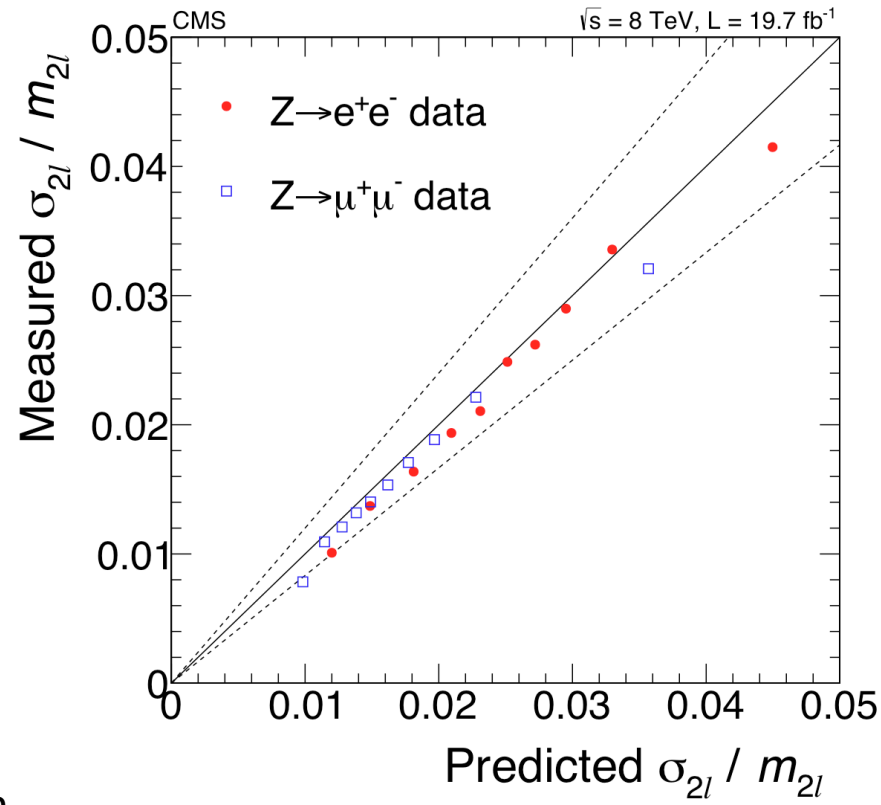


CMS Search for H→ZZ→4 leptons

CMS, arXiv:1312:5353
Phys. Rev. D 89, 092007

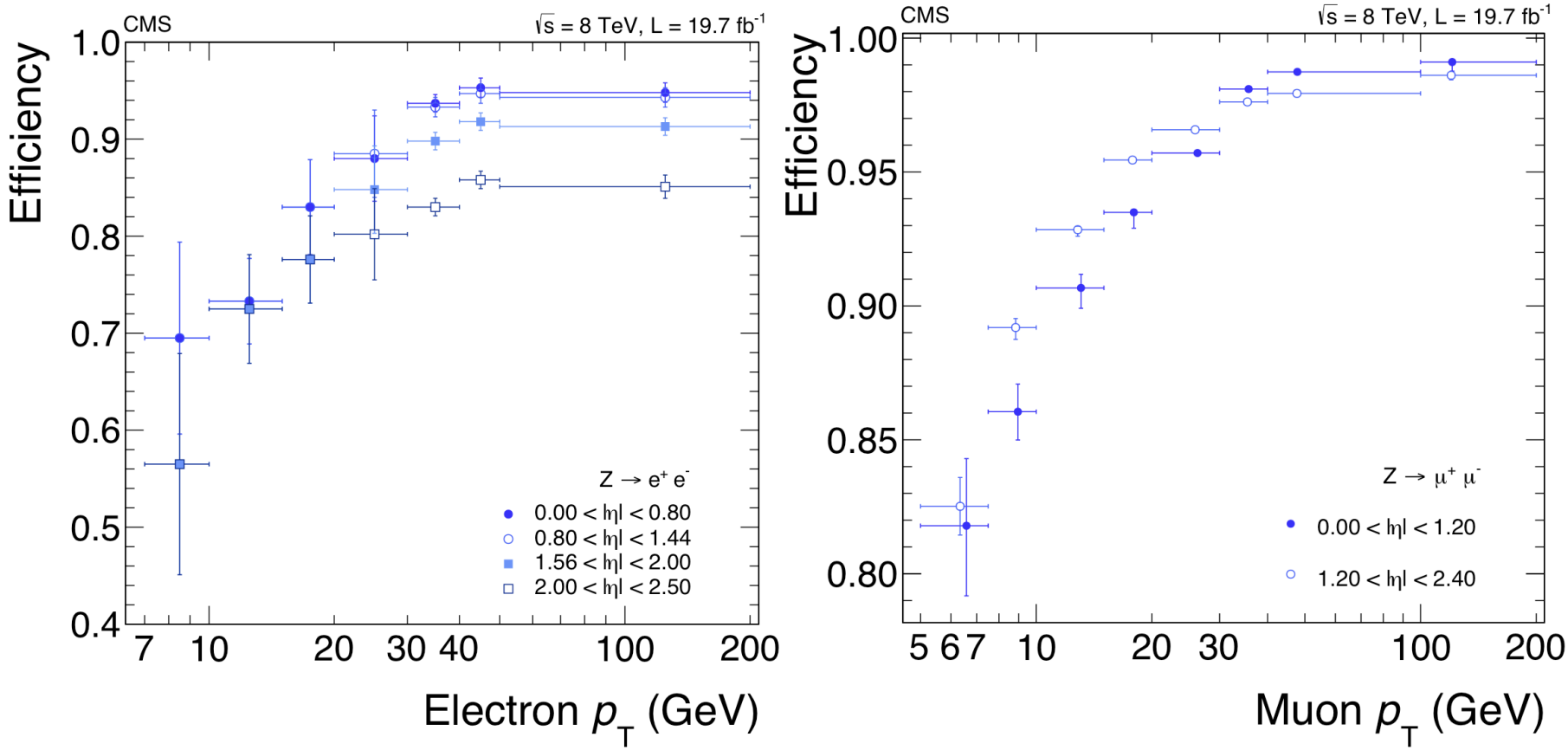
Preliminaries – Reconstructed mass resolution





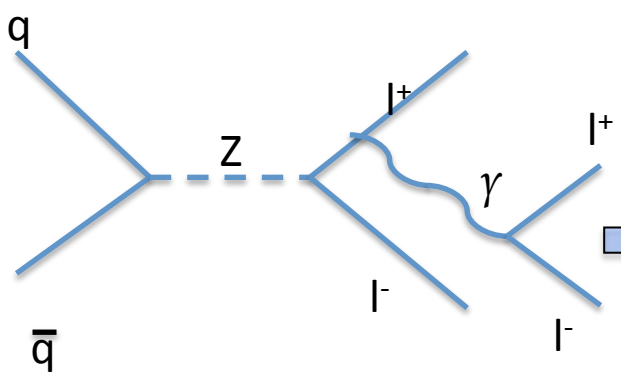
MC Higgs boson signal resolution in the 4e channel. Improvements to resolution from categorizing electrons and using NN's on the calorimeter clusters

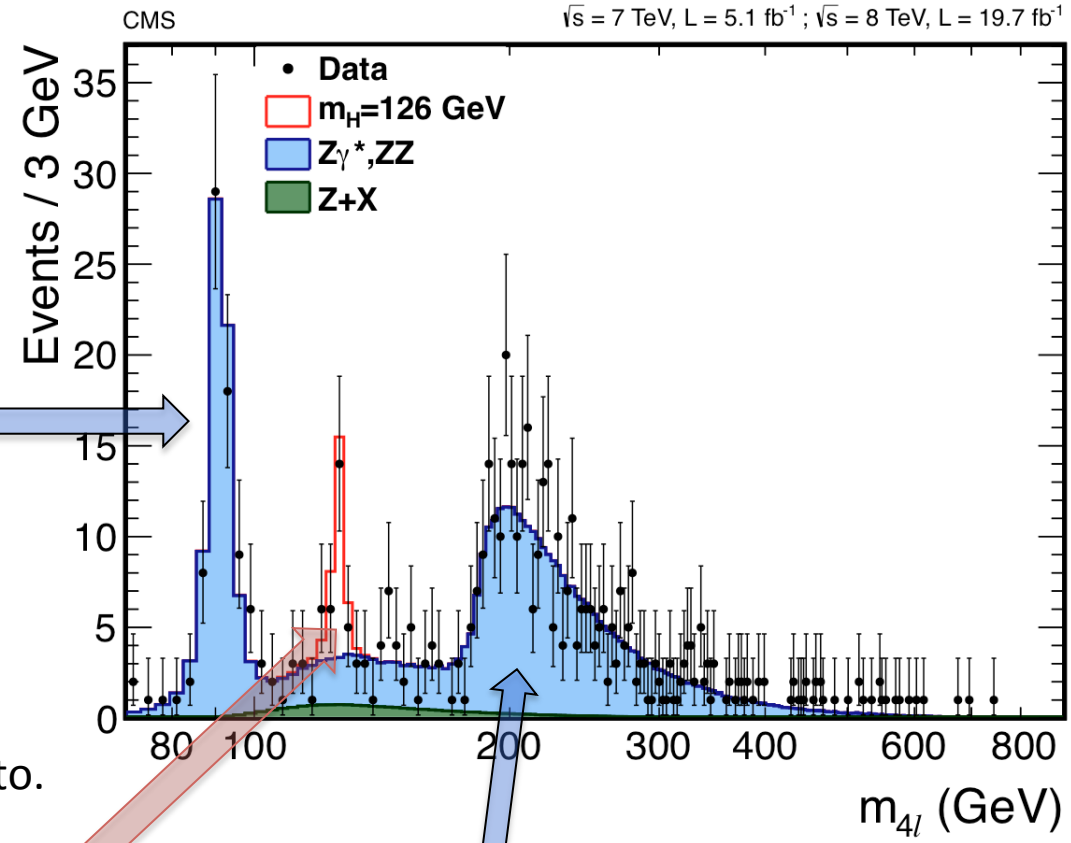
Lepton Selection Efficiency Measurements using CMS Data – Tag and Probe



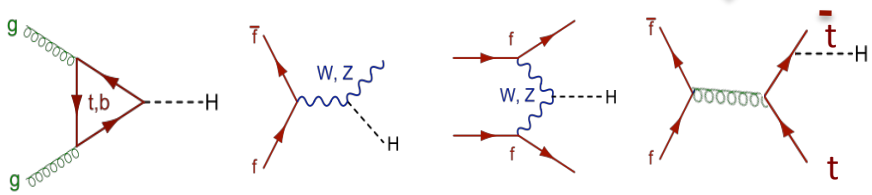
The M(4L) Distribution

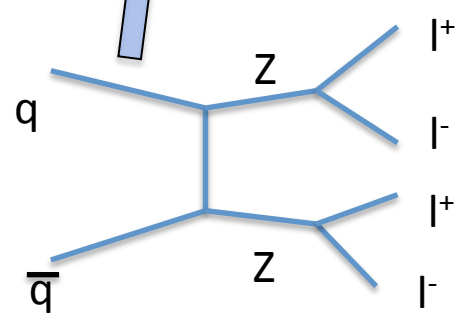
Summed plot: eeee, eeμμ, μμμμ





I'm not too sure what the error bars on the data points correspond to.



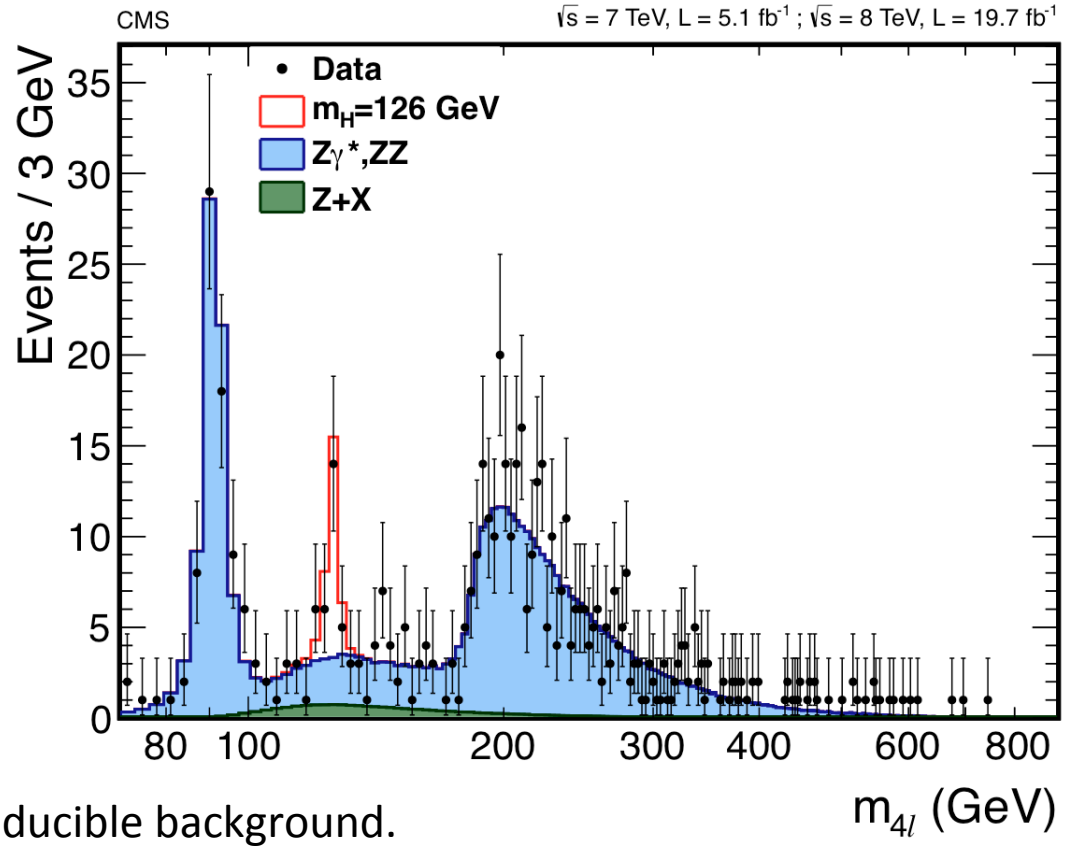


Background Estimation

Complicated sideband fits –

Drell-Yan + FSR conversion
well known from theory
(and LEP measurements!)
Known differentially at NNLO
(at least!)

Diboson production – also
has precise theory predictions.
But I'd not fit them with a single
nuisance parameter to the data.
MCFM used to predict dominant irreducible background.



Fake backgrounds: Z+X, estimated using same-sign data and Z+jets data with loose lepton ID.

Statistical Procedures Recommended by the LHC (2011)

ATLAS Collaboration and CMS Collaboration, *Procedure for the LHC Higgs boson search combination in Summer 2011*, ATLAS-PHYS-PUB-2011-011, CMS NOTE-2011/005 (2011).

Recommends:

- Profile likelihood used for maximum-likelhood fits. Nuisance parameters are fit using external constraints and the data likelihood function.

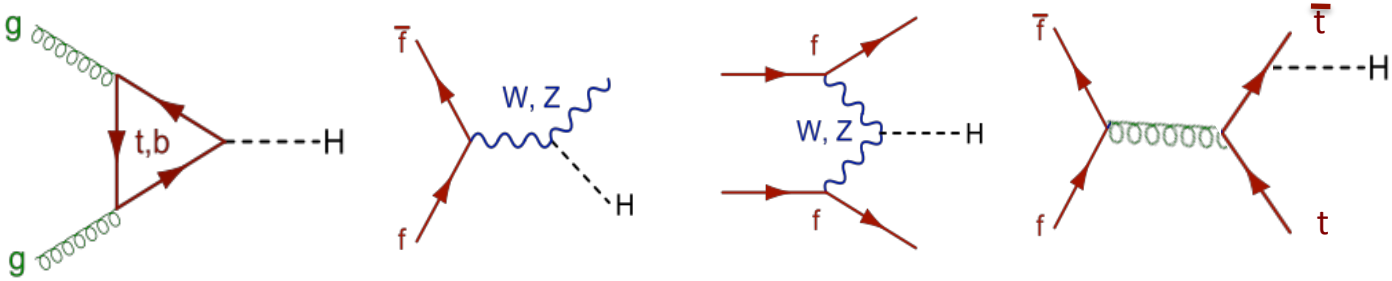
- Profile likelihood ratio test statistic used to sort experimental outcomes for computing p-values.

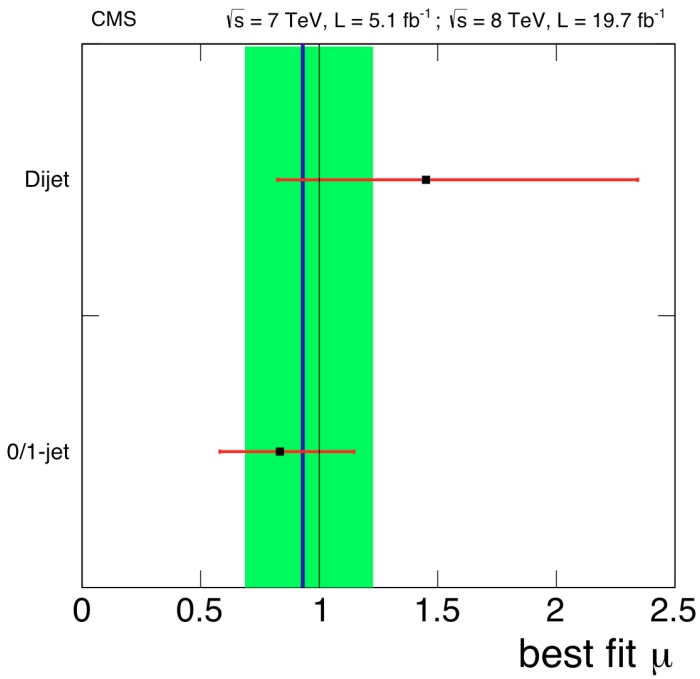
- Limits: CL_s using profile-likelihood ratio as the test statistic

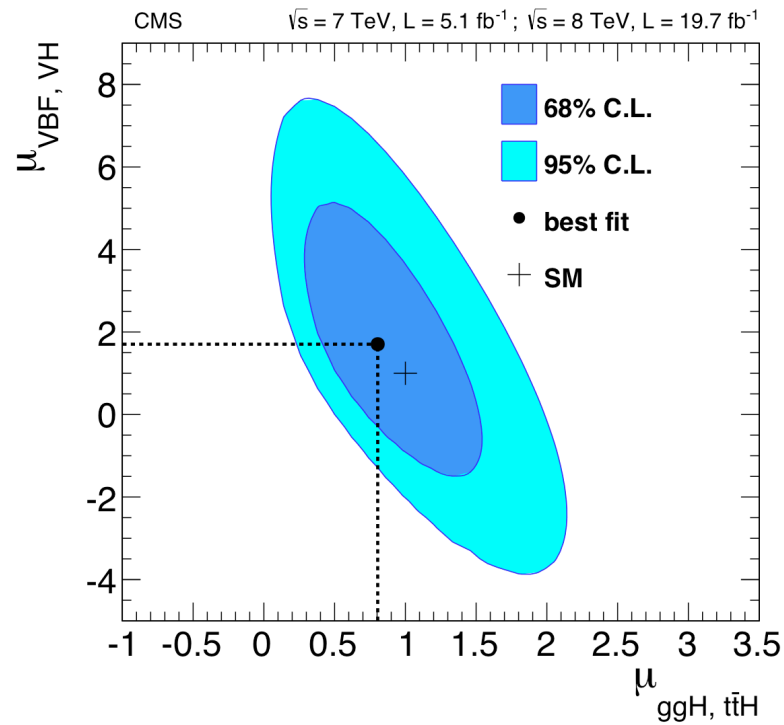
- Systematic uncertainties:
  Best-fit values (include systematic priors in these fit) are then fixed in order to generate pseudoexperiments. (fitting without fluctuating)

Profile Likelihood method

Divide and Conquer has a Side Benefit – More Physics





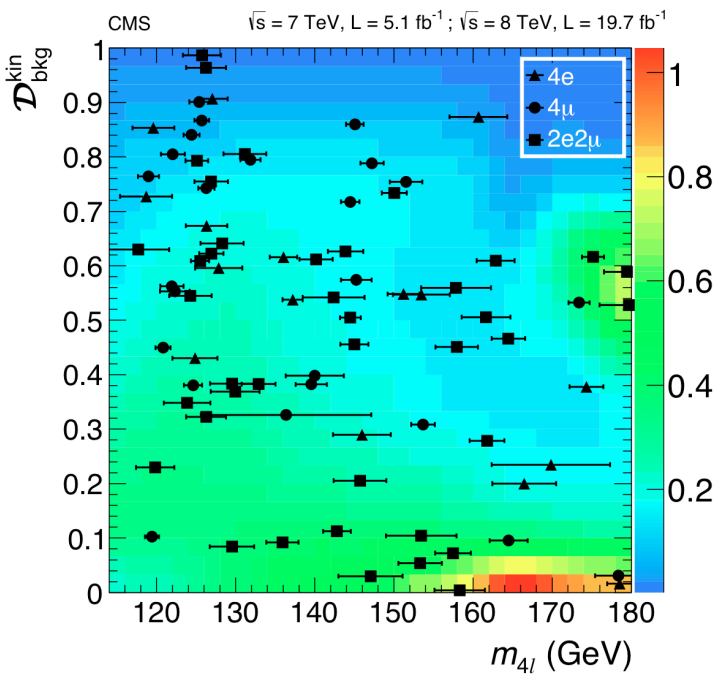


Profile Likelihood used – fit for parameters of interest and nuisance parameters

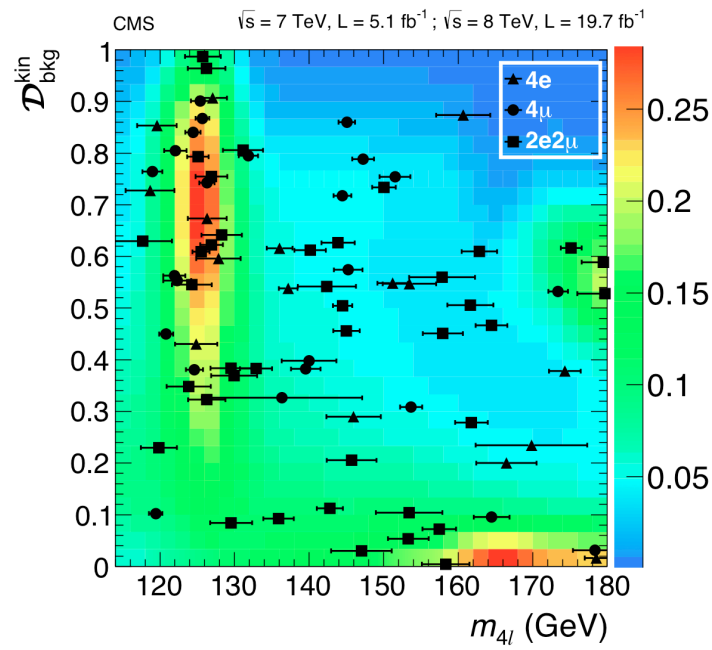
Divide and Conquer Again – but This Time Reduce Dimensionality with a Likelihood-based Discriminant using non-mass information

$$\mathcal{D}_{\rm bkg}^{\rm kin} = \frac{\mathcal{P}_{0^+}^{\rm kin}}{\mathcal{P}_{0^+}^{\rm kin} + \mathcal{P}_{\rm bkg}^{\rm kin}} = \left[1 + \frac{\mathcal{P}_{\rm bkg}^{\rm kin}(m_{Z_1}, m_{Z_2}, \vec{\Omega}|m_{4\ell})}{\mathcal{P}_{0^+}^{\rm kin}(m_{Z_1}, m_{Z_2}, \vec{\Omega}|m_{4\ell})}\right]^{-1}$$

Angular information and masses of the two ll pairs, designed to be independent of m_{4l}

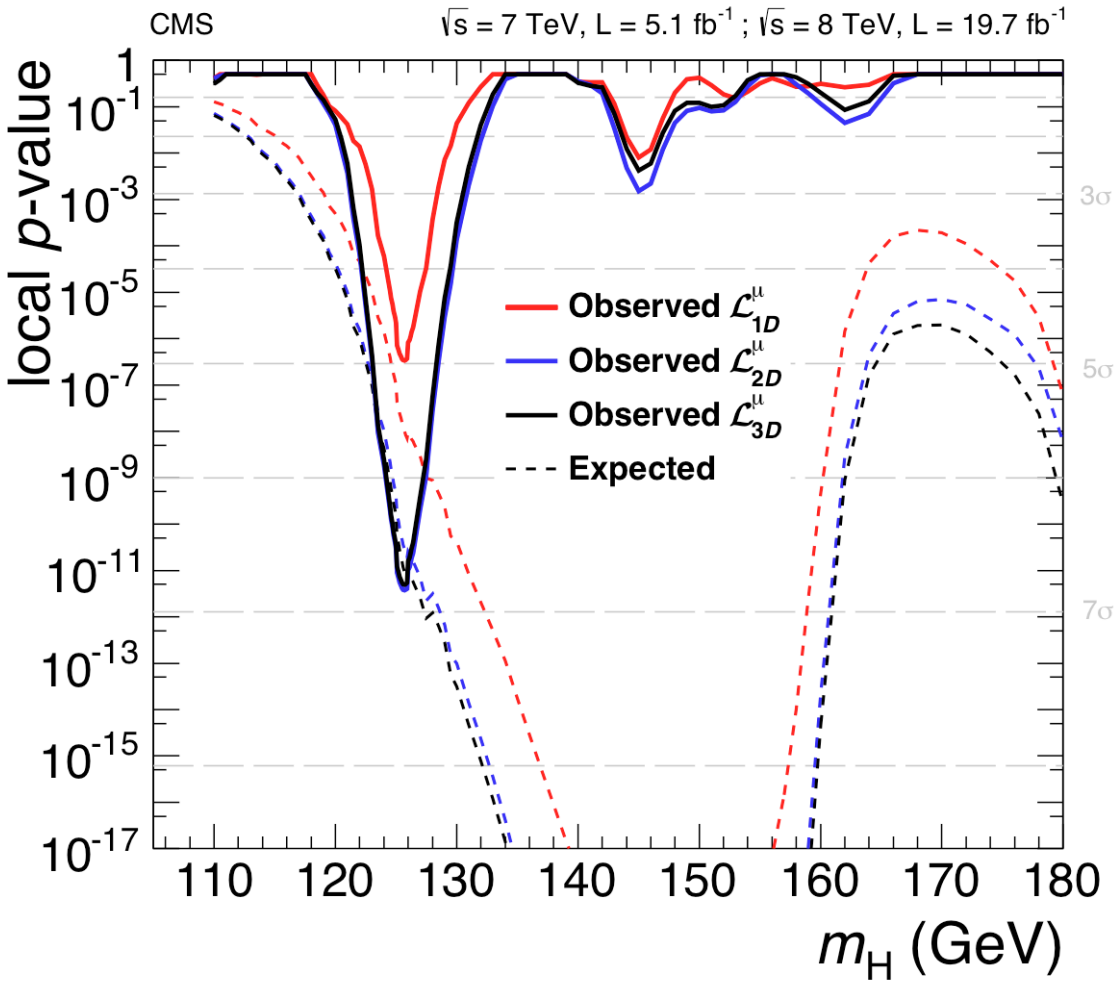


Data and Background-Only prediction



Data and Signal+Background prediction

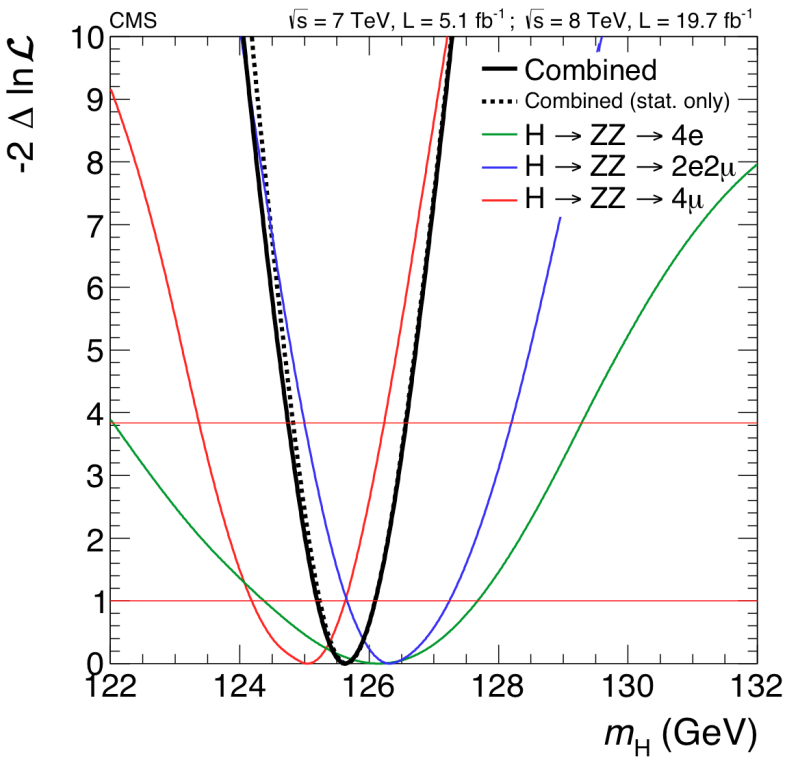
Observed and Expected p-values, CMS ZZ4l

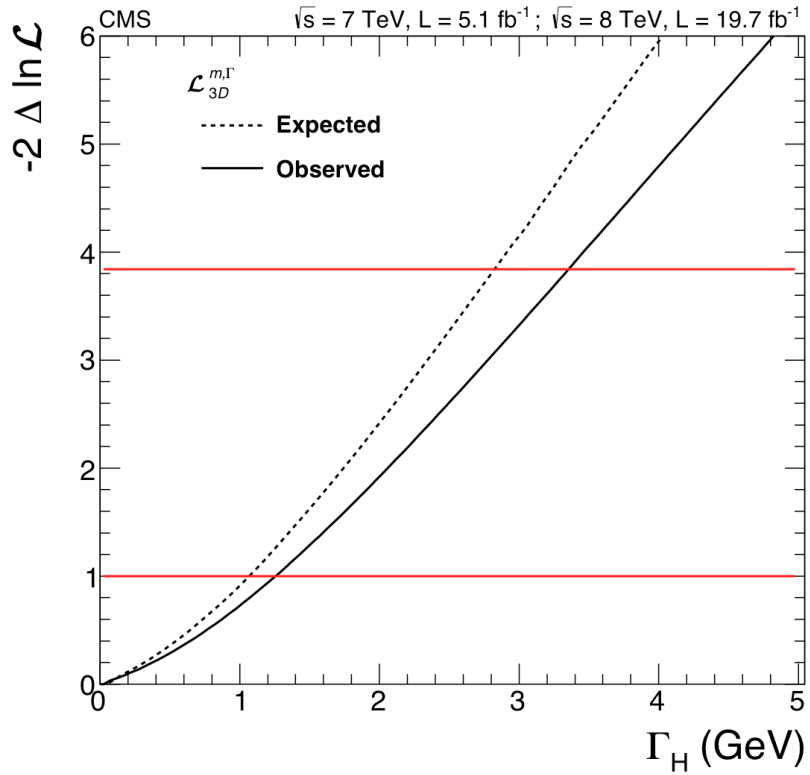


1D fit – just fit the m_{4l} Distribution
2D fit – m_{4l} vs. D_{kin}^{bkg} Distribution

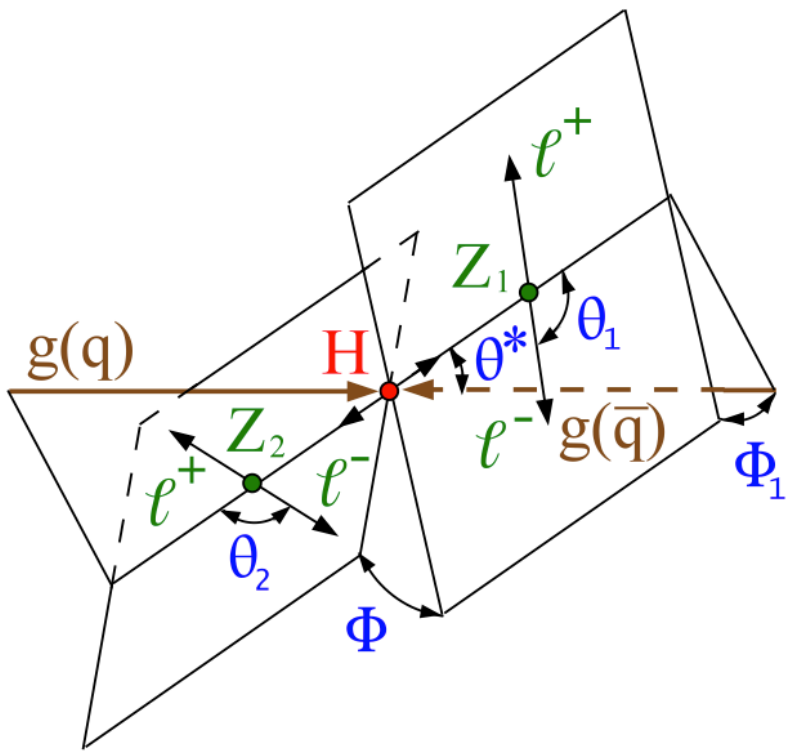
3D fit: Include $p_T(4l)$ as an extra variable for 0, 1-jet events, and a jet-based likelihood discriminant for 2-jet events.

Mass and Width

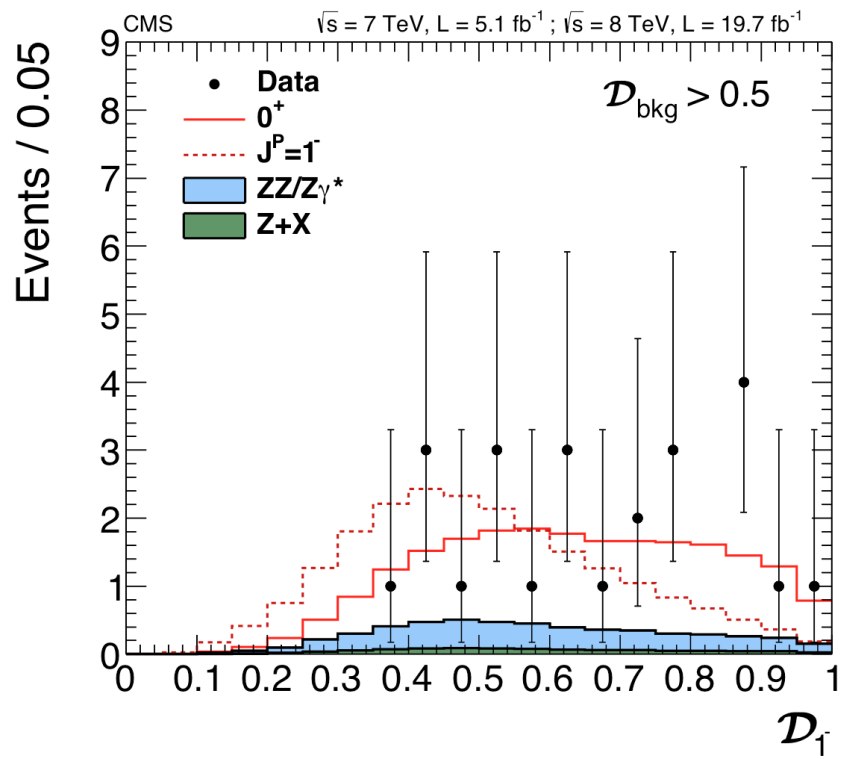




Spin and Parity: CMS ZZ4l

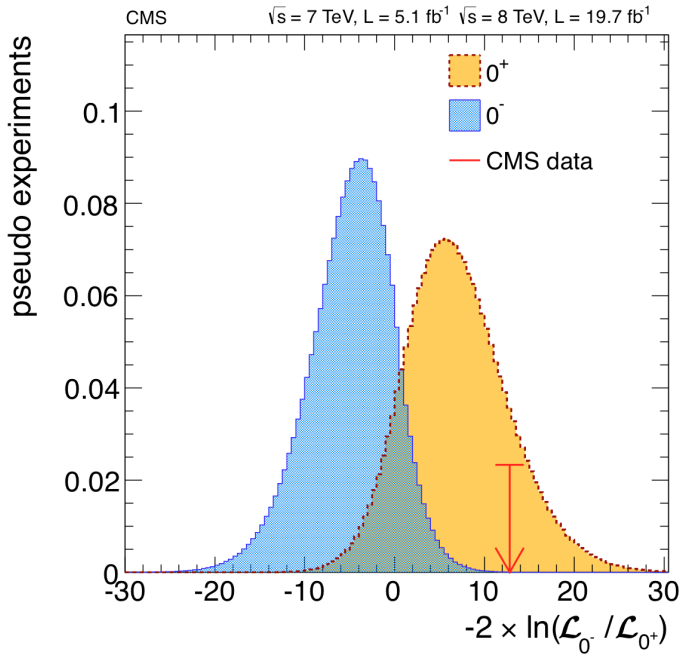


Example: Vector Higgs Discriminant



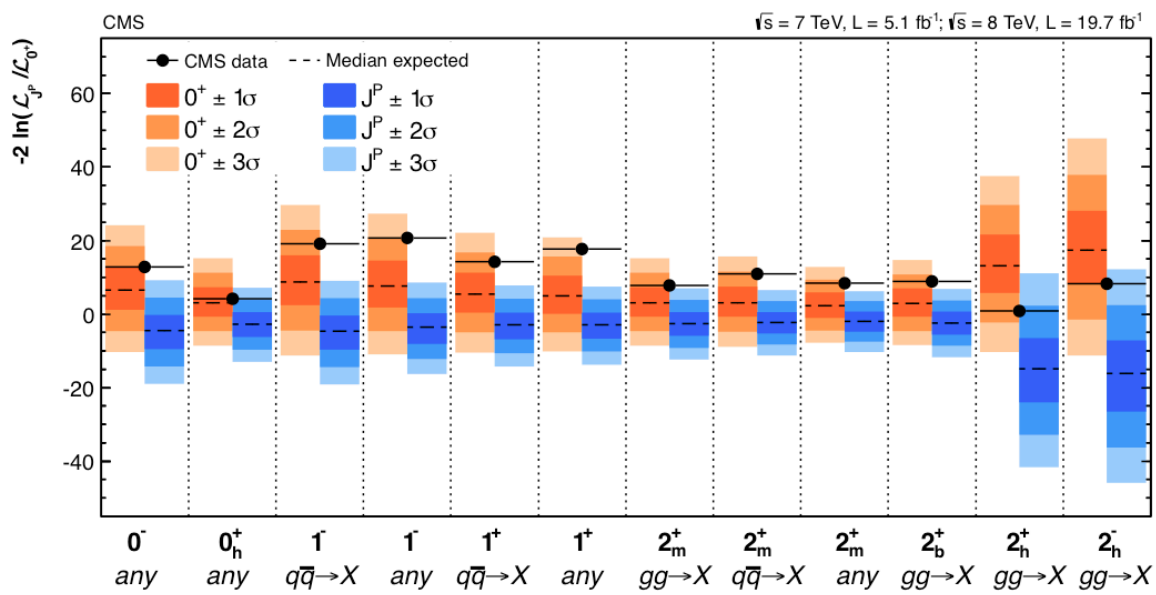
Fairly pure sample of signal!

Spin and Parity p-values – Run SM and Exotic Pseudoexperiments and use LLR

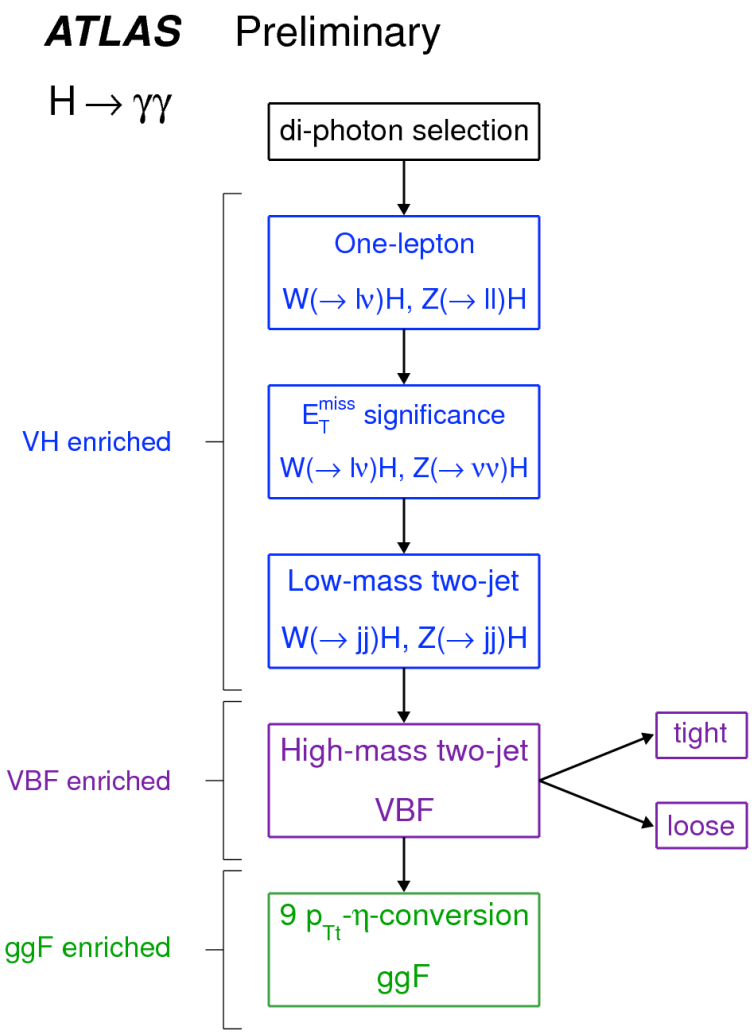


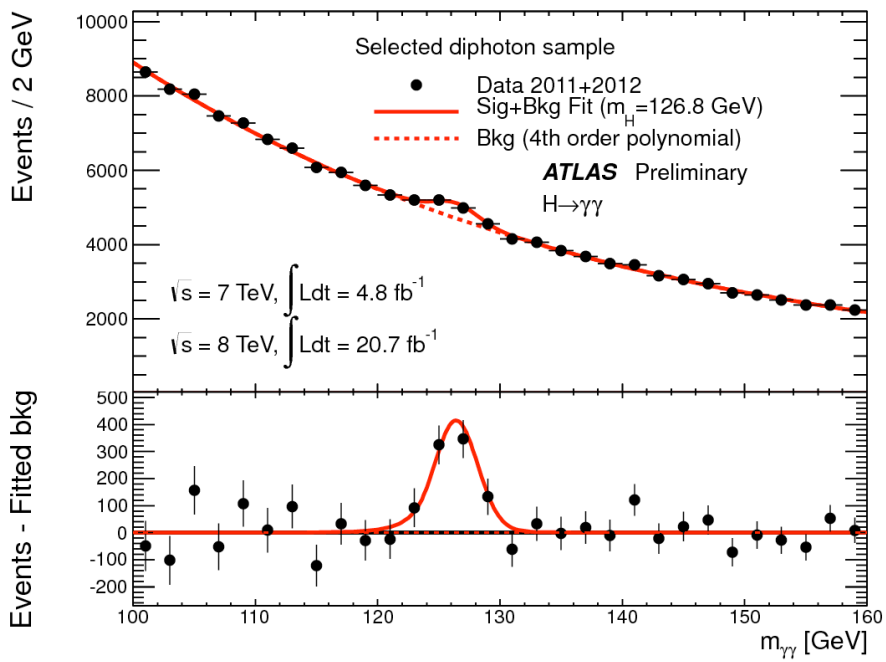
Example: Pseudoscalar hypothesis

Repeat for several different models



ATLAS-CONF-2013-012: H→γγ



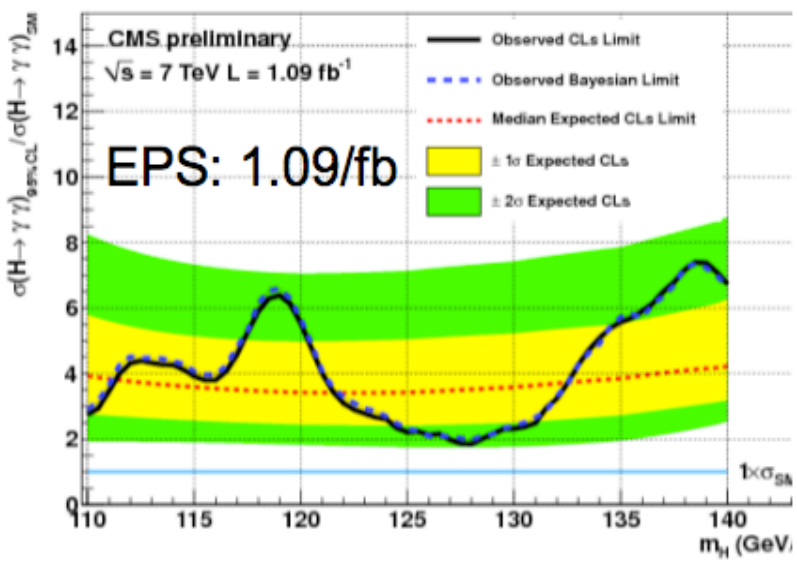


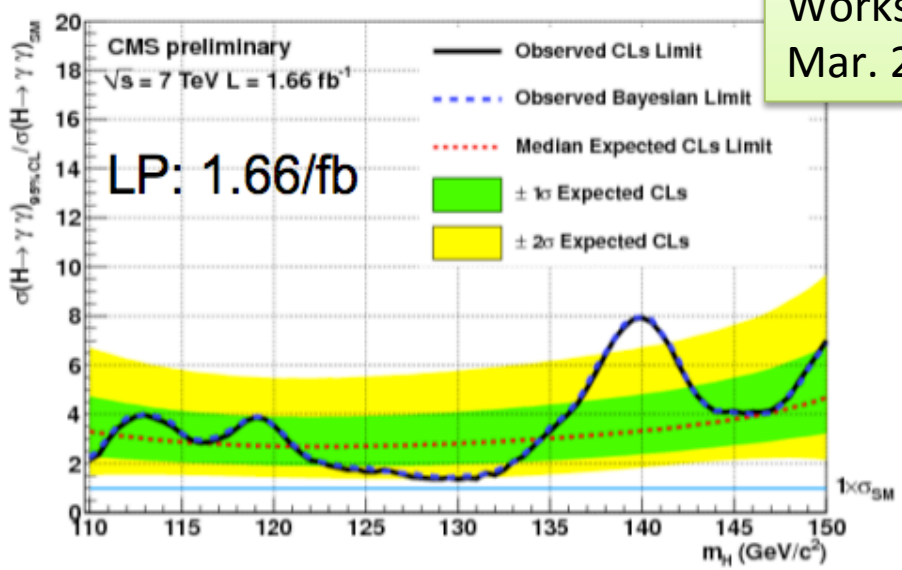
It wasn't always so clear...

Show ATLAS animation

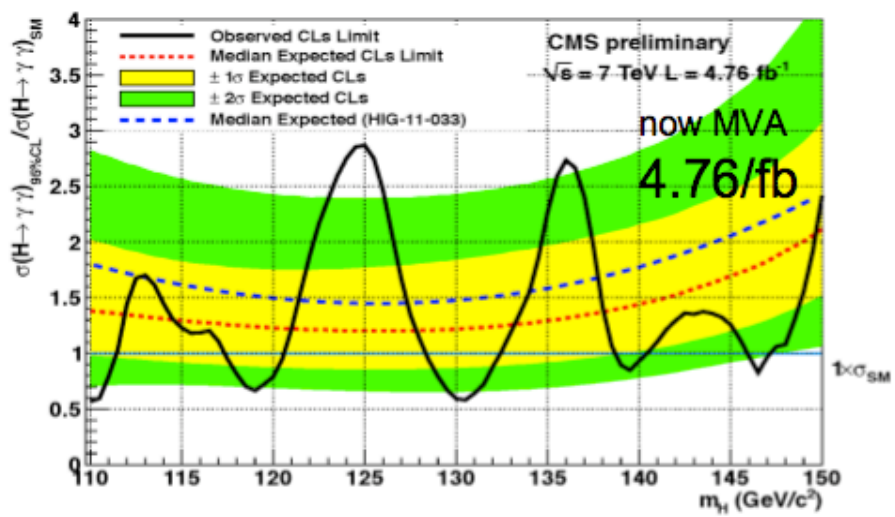
C. Paus, Implications Workshop, Mar. 27, 2012

CMS History: $H \rightarrow \gamma\gamma$

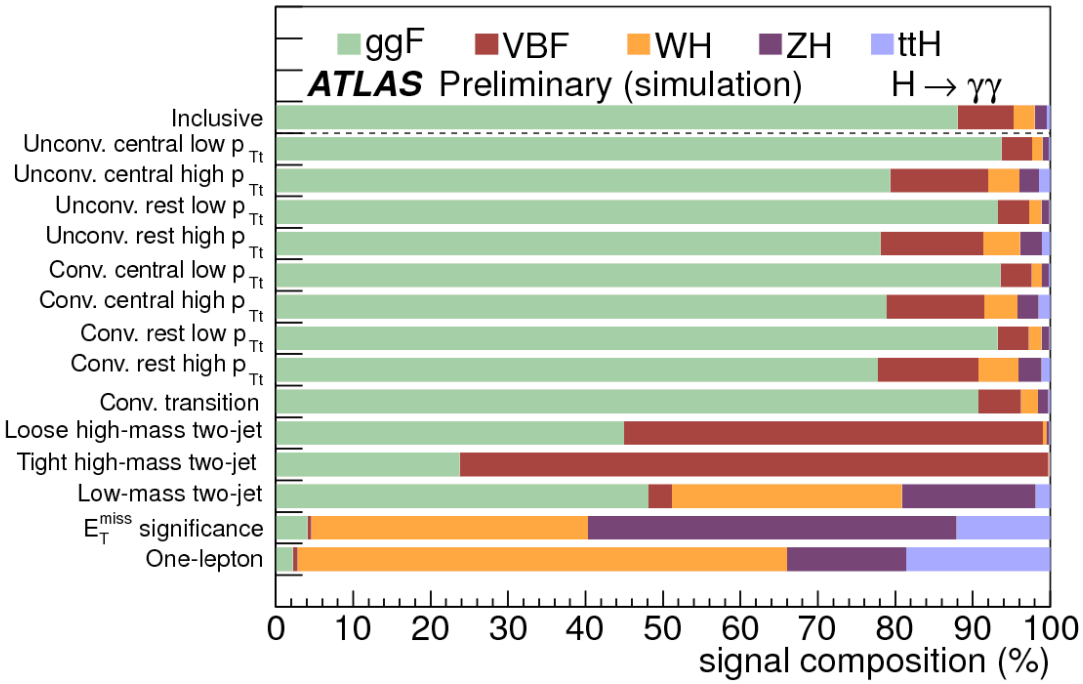


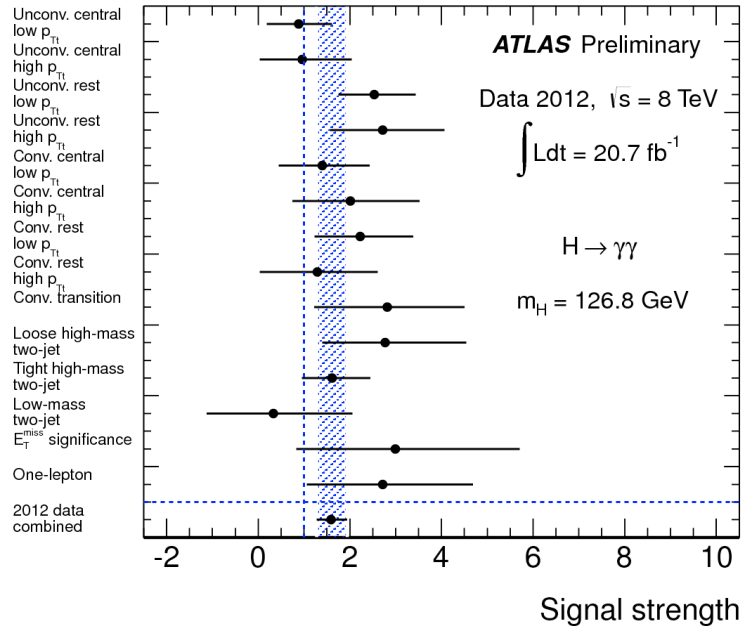


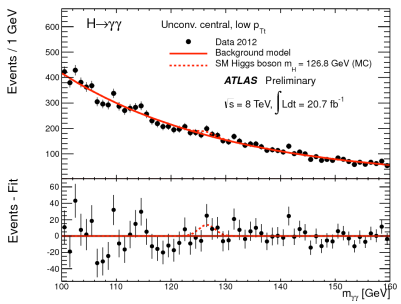
- EPS (1.09/fb) LP (1.66/fb) Dec 19 (4.76/fb)
- 'peaks' come and go
- we are getting into interesting territory, and peaks can also stay

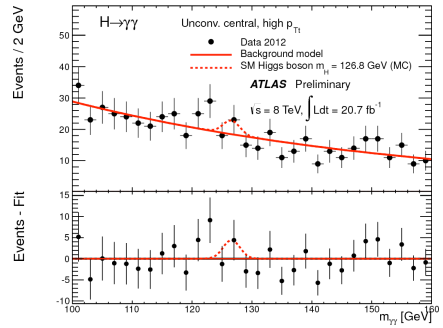


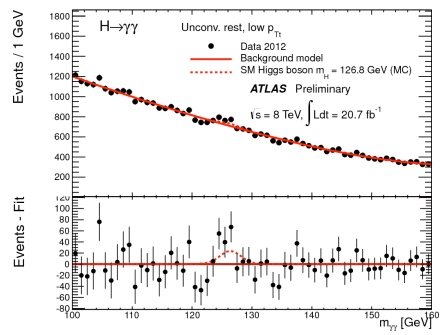
Divide and Conquer Again – Several categories of photons, p_T, and jets/no jets

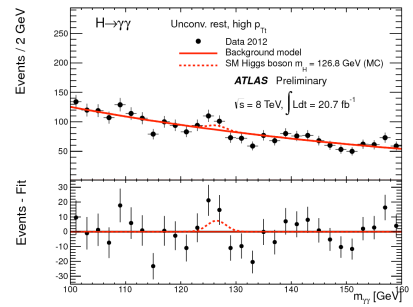






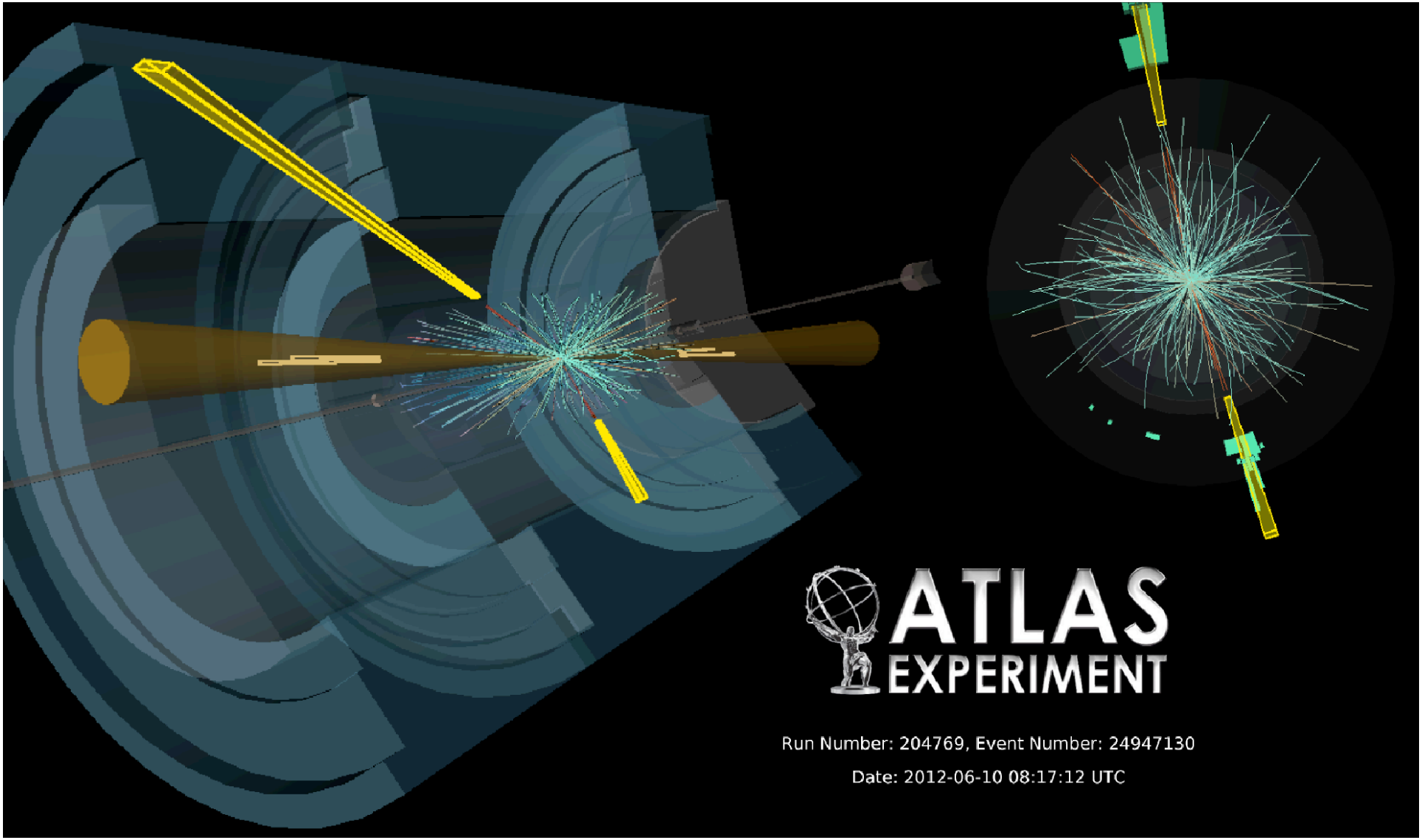




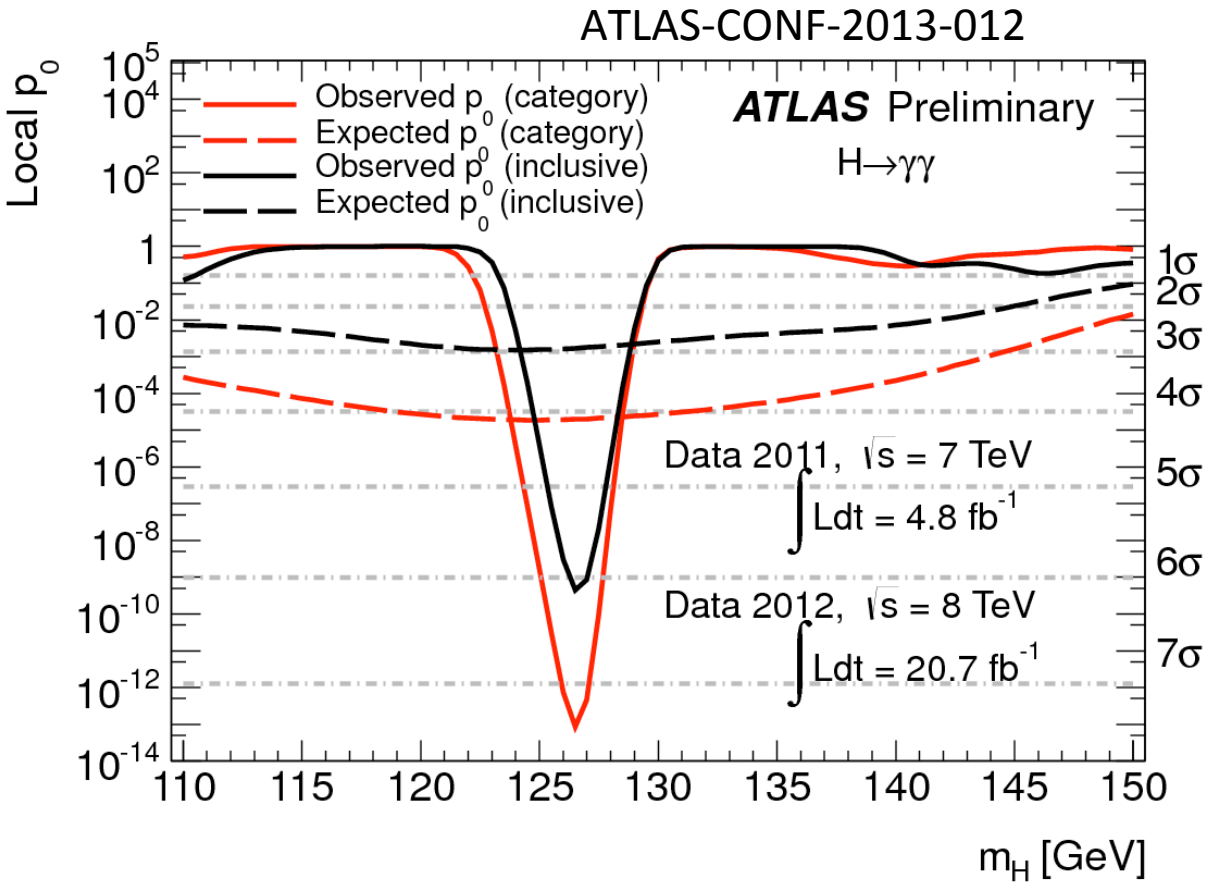


And many more. Backgrounds fit to smooth functions (polynomials, exponentials of polynomials)

A VBF Candidate. H→$\gamma\gamma$+2 jets event

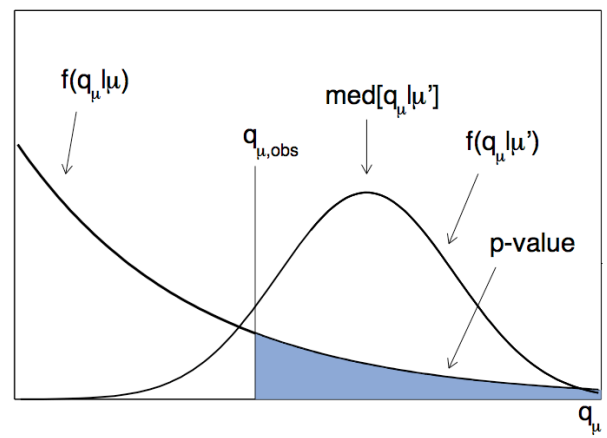


ATLAS H→γγ p-values – adding it all together and combining channels

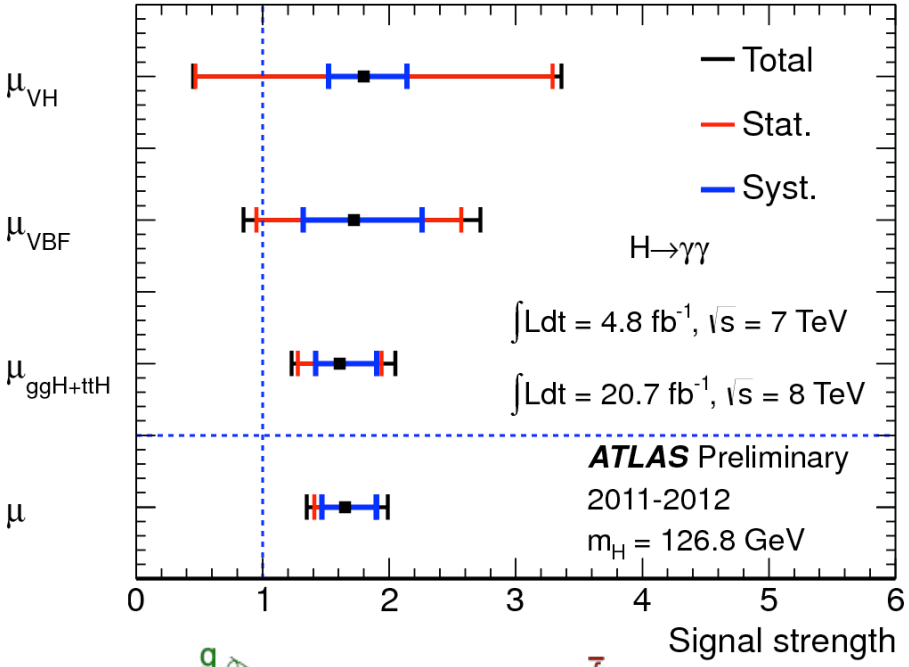


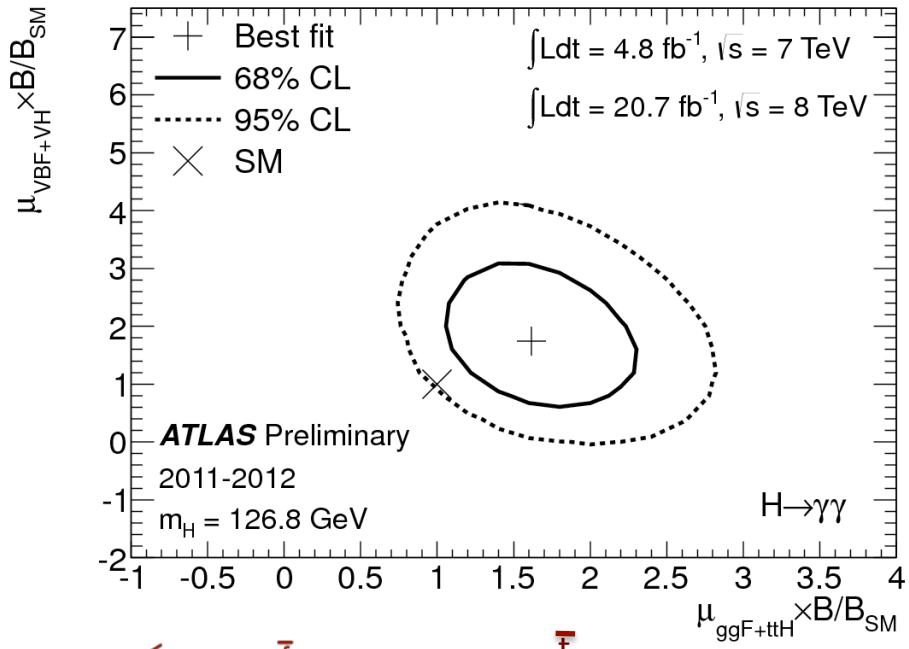
$$q_\mu \equiv 2\ln\left(\frac{L(\text{data}|\hat{\mu},\hat{\nu})}{L(\text{data}\,|\,\mu,\hat{\hat{\nu}})}\right)$$

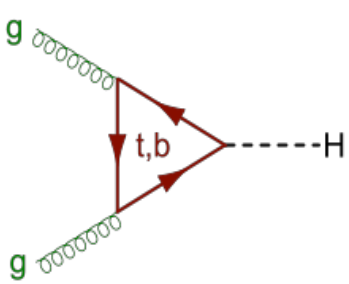
Channels combined assuming SM ratios

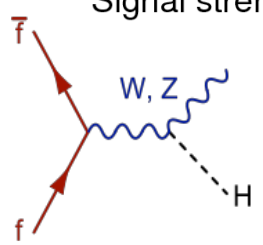


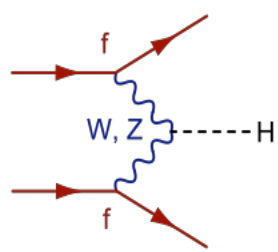
ATLAS H→γγ Results by Physics Process

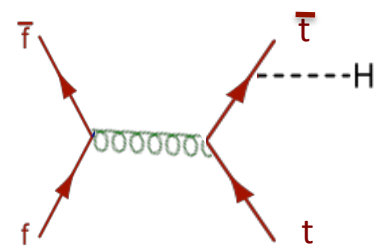


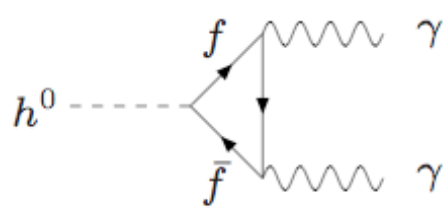


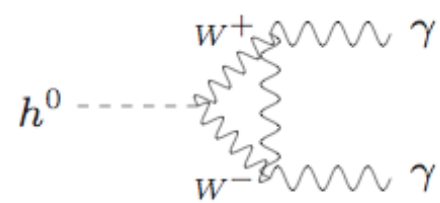






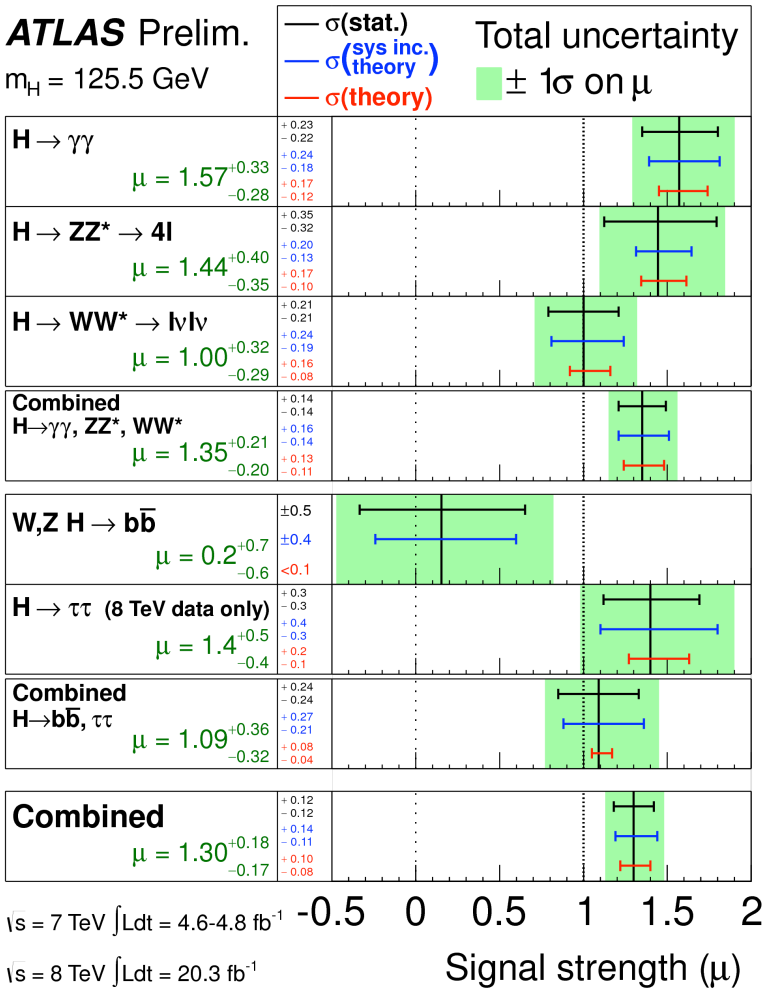


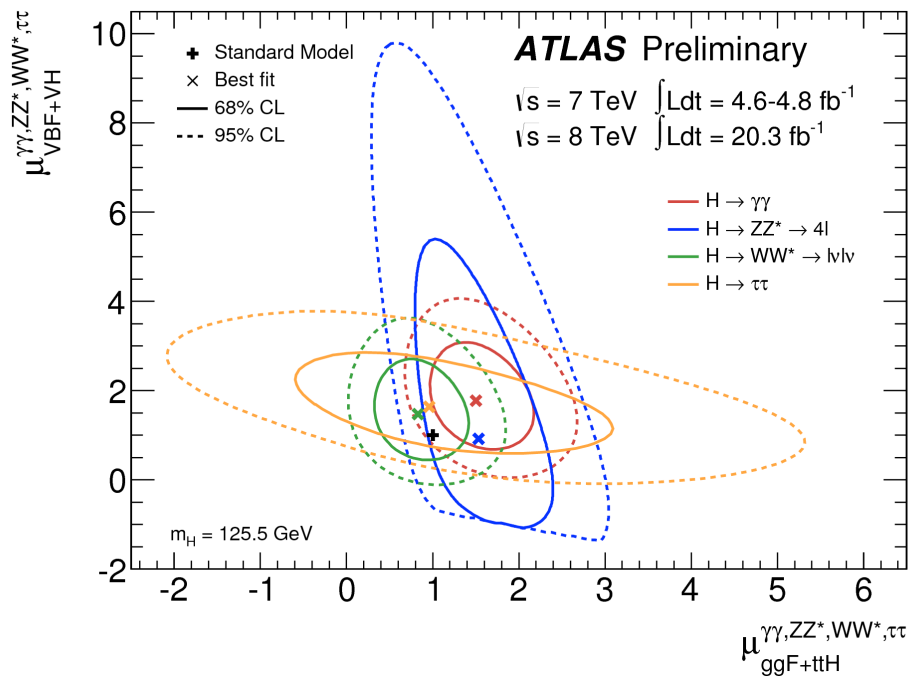




Coupling Fits From the Separate Channel Results

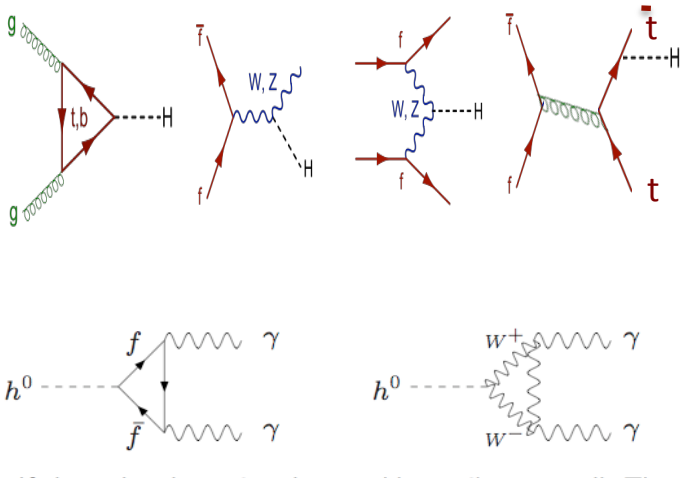
ATLAS-CONF-2014-009





Signal strengths μ should be positive definite. Profile likelihood fits allow for negative values.

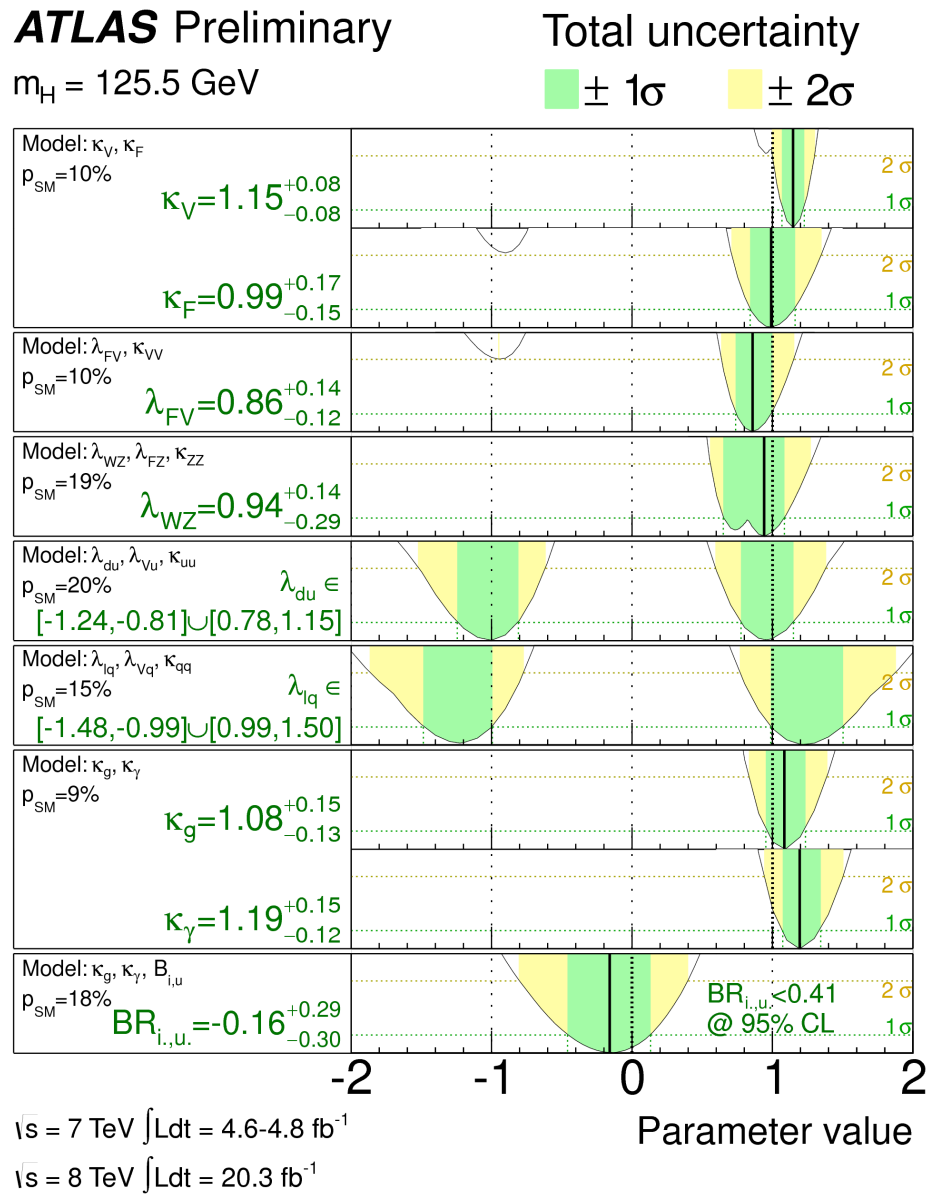
Coupling Fits: ATLAS-Conf-2014-009



Kappa parameters scale couplings at the matrix-element level.
A. David et al., arXiv:1209.0040

Cross sections depend quadratically on the kappas.

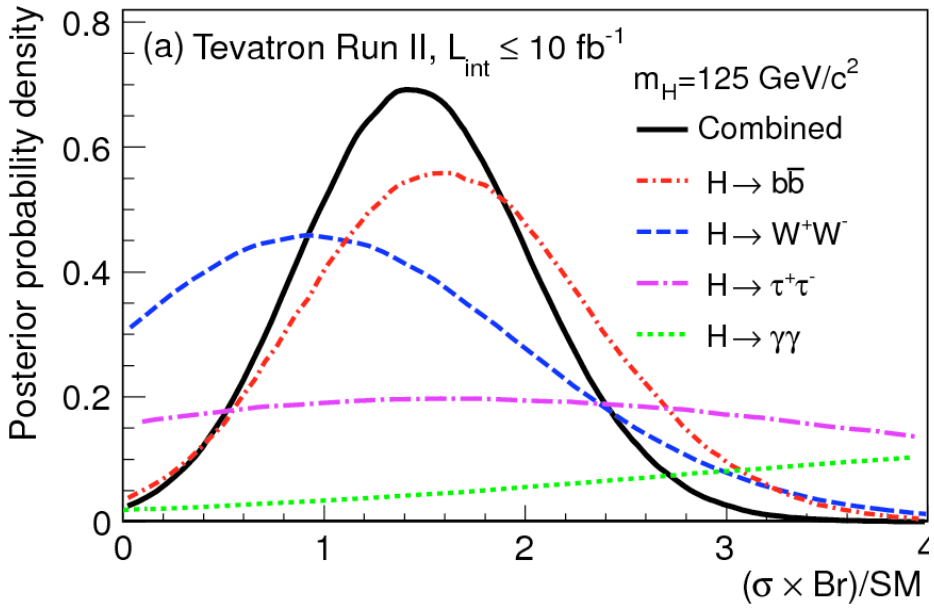
Get bimodal likelihoods and posteriors

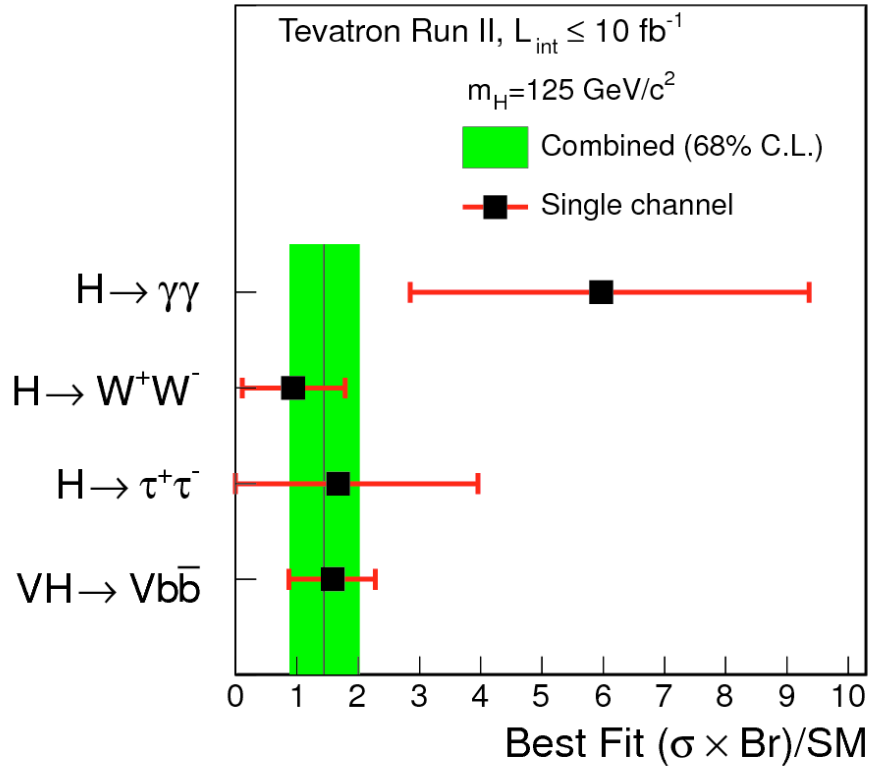


Tevatron Cross-Section Fits by Channel

Bayesian posteriors:

$$L'(data\,|\,r) = \int L(data\,|\,r,\nu)\pi(\nu)\,d\nu$$





Tevatron local significance at m_H=125 GeV: 3σ in combination using LLR as a test statistic. T. Aaltonen et al., CDF and D0 Collaborations, Phys. Rev. D **88**, 052014 (2013).

Higgs Boson Couplings to Fermions and Bosons

From the Review
of Particle Physics
Higgs Boson Section:
Carena, Grojean,
Kado, and Sharma

Tevatron: Bayesian Posteriors
LHC: Profile likelihood contours

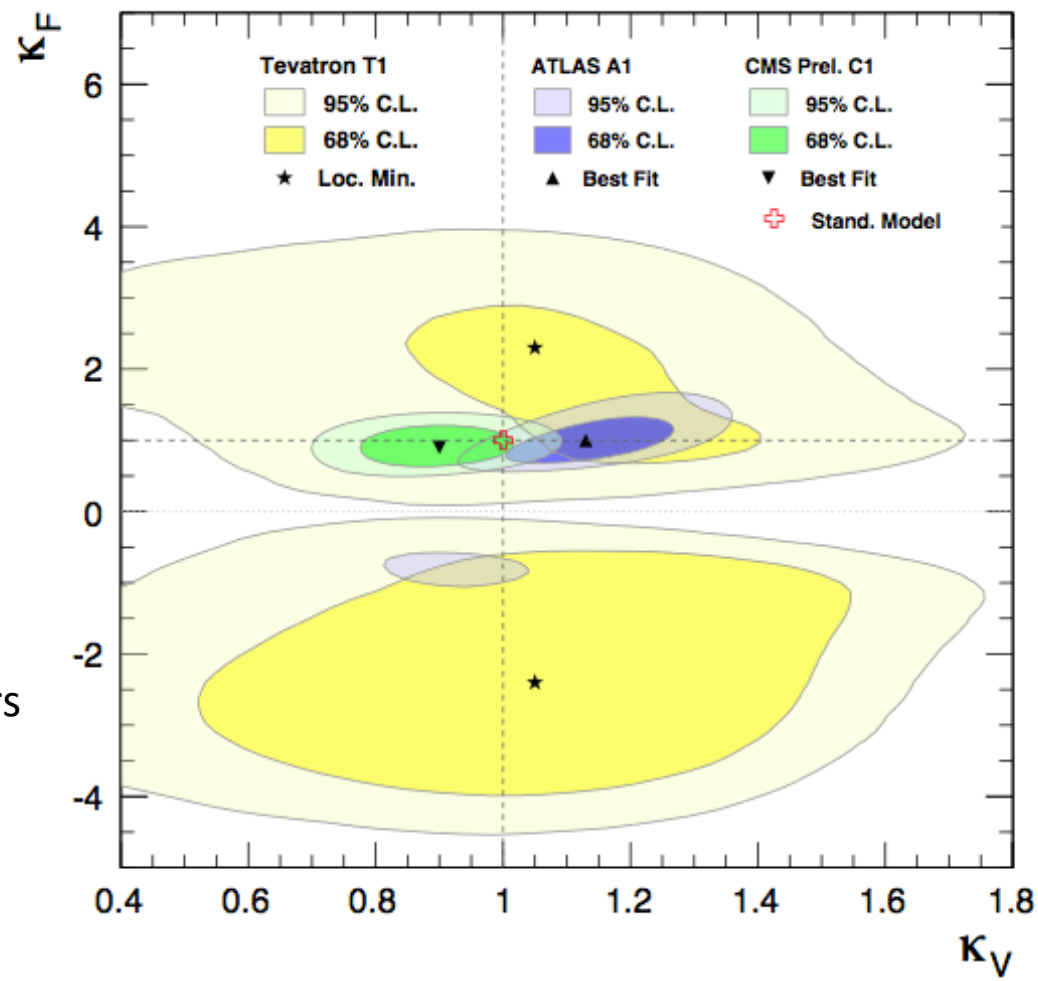
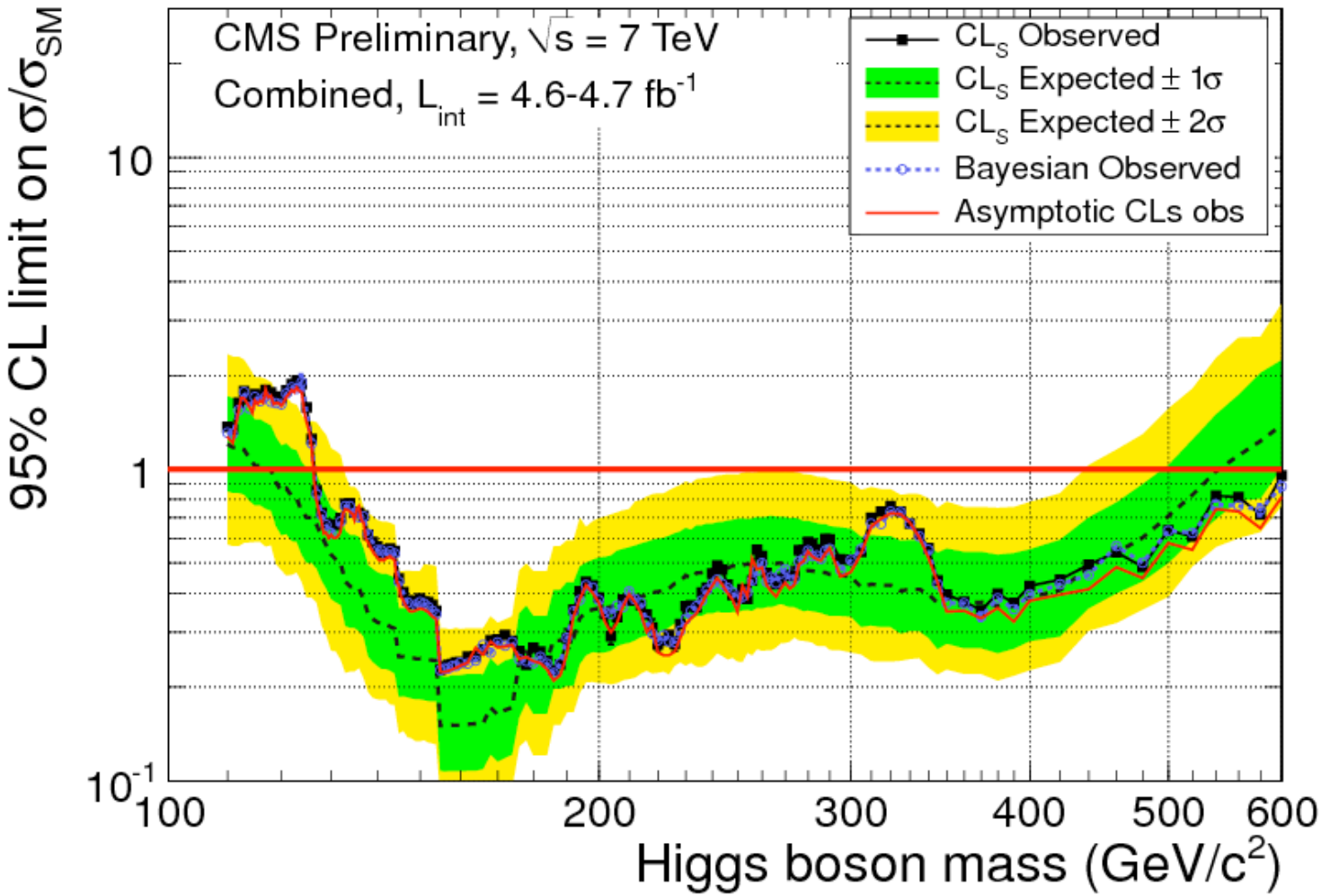


Figure 18: Likelihood contours of the global fit in the (κ_F, κ_V) plane for the ATLAS A1 [119], the CMS C1 [120] and the D0 and CDF combined T1 [108] results.

Bayesian and CL_s Limits Tend to be Close Together

CMS-HIG-11-032



CERN-EP-98-046

Different observed limits from CL_s using different test statistics.

Method A: LLR (but no fits)

Method B: Bayeian posterior integral

Method C: Fractional event counting

But the sensitivities are hard to tell apart

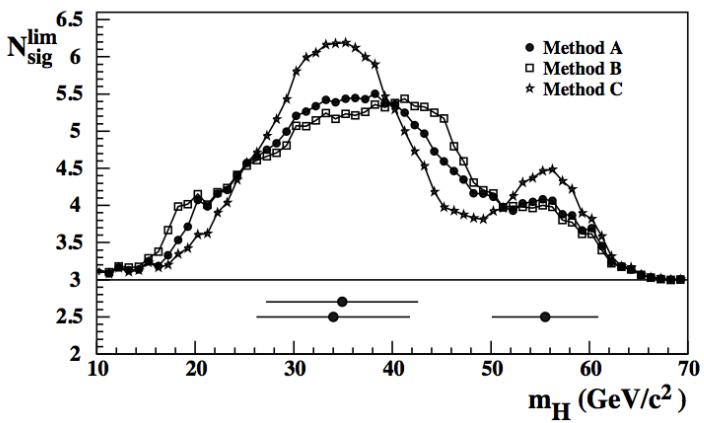


Figure 1: *Test (I); the 95% confidence level "experimental" upper bound for the number of signal events,* N_{sig}^{lim}*, as a function of* m_H*, obtained by Methods A, B and C. The smallest value, three, obtained far from the candidate events, is indicated by the horizontal line. The three dots with error bars show the masses and assumed resolutions of the candidate events.*

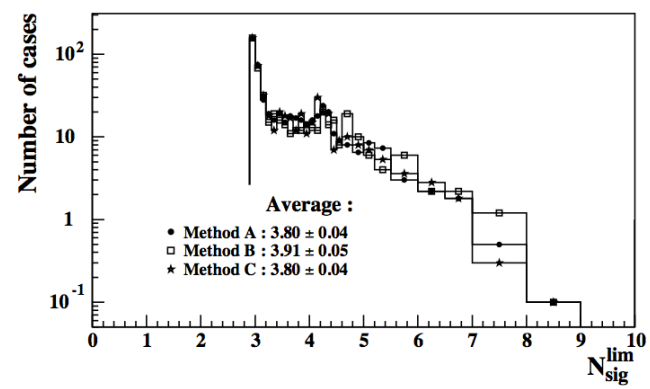
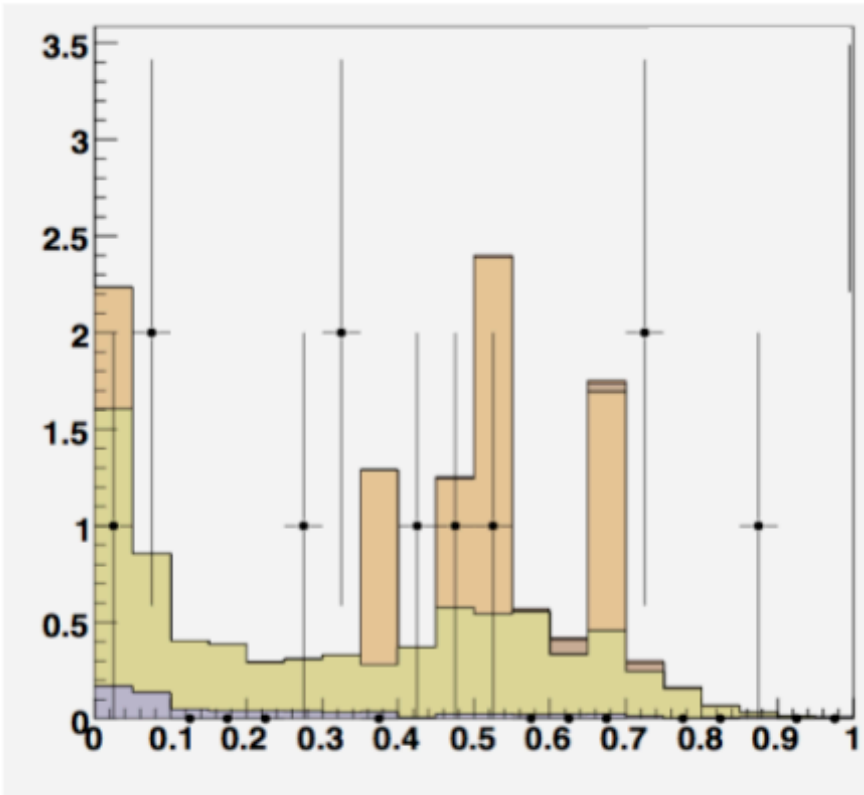


Figure 2: *Test (I); the distribution of the 95% confidence level "experimental" upper limit for the number of signal events,* N_{sig}^{lim}*, assuming* m_H*=40* GeV/c^2*, in 600 gedanken experiments with background only, as obtained by Methods A, B and C. The average values of the limits are also given in the figure.*

A Pitfall -- Not Enough MC (data) To Make Adequate Predictions

An Extreme Example (names removed)



Cousins, Tucker and Linnemann tell us prior predictive p-values undercover with 0±0 events are predicted in a control sample.

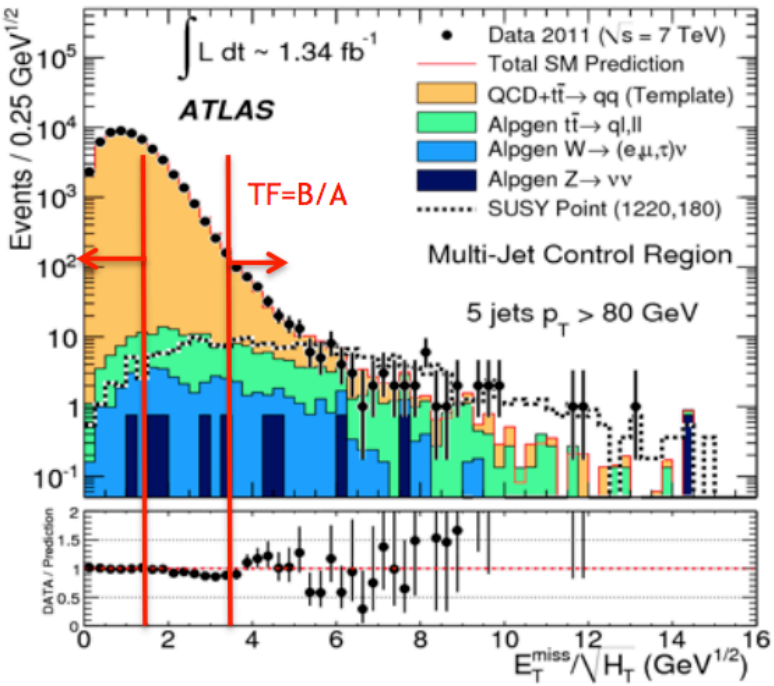
CTL Propose a flat prior in true rate, use joint LF in control and signal samples. Problem is, the mean expected event rate in the control sample is n_{obs}+1 in control sample. Fine binning → bias in background prediction.

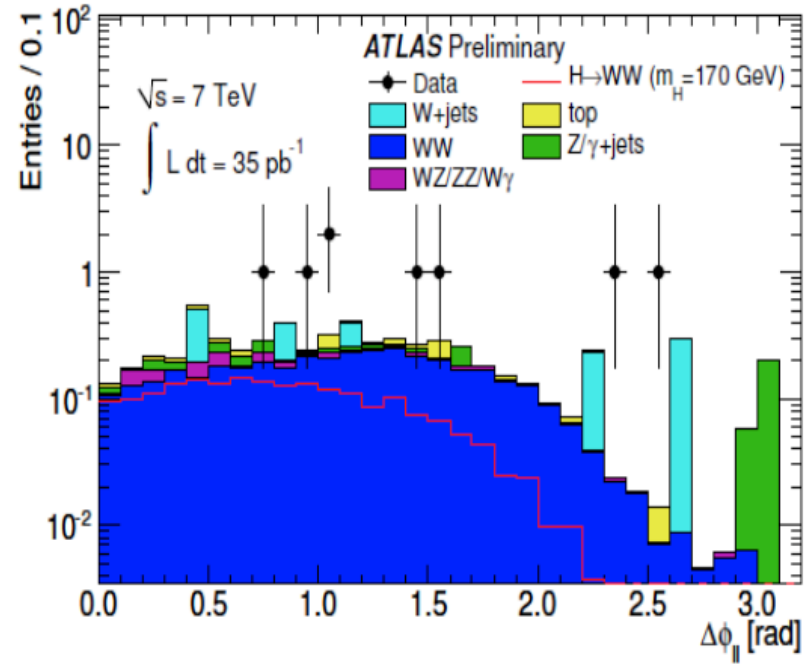
Questions: What's the shape we are trying to estimate? What is the uncertainty on that shape?

Overcovers for discovery, undercovers for limits?

Some Very Early Plots from ATLAS

Suffer from limited sample sizes in control samples and Monte Carlo
Nearly all experiments are guilty of this, especially in the early days!





Data points' error bars are not sqrt(n). What are they? I don't know. How about the uncertainty on the prediction?

The left plot has adequate binning in the "uninteresting" region. Falls apart on the right-hand side, where the signal is expected.
Suggestions: More MC, Wider bins, transformation of the variable (e.g., take the logarithm).
Not sure what to do with the right-hand plot except get more modeling events.

Optimizing Histogram Binning

Two competing effects:

1) Separation of events into classes with different s/b improves the sensitivity of a search or a measurement. Adding events in categories with low s/b to events in categories with higher s/b dilutes information and reduces sensitivity.

→ Pushes towards more bins

2) Insufficient Monte Carlo can cause some bins to be empty, or nearly so. This only has to be true for one high-weight contribution.

Need reliable predictions of signals and backgrounds in each bin

→ Pushes towards fewer bins

Note: It doesn't matter that there are bins with zero data events – there's always a Poisson probability for observing zero.

The problem is inadequate prediction. Zero background expectation and nonzero signal expectation is a discovery!

Overbinning = Overlearning

A Common pitfall – Choosing selection criteria after seeing the data.
"Drawing small boxes around individual data events"

The same thing can happen with Monte Carlo Predictions –

Limiting case – each event in signal and background MC gets its own bin.
→Fake Perfect separation of signal and background!.

Statistical tools shouldn't give a different answer if bins are shuffled/sorted.

Try sorting by s/b. And collect bins with similar s/b together. Can get arbitrarily good performance from an analysis just by overbinning it.

Note: Empty data bins are okay – just empty prediction is a problem. It is our job however to properly assign s/b to data events that we did get (and all possible ones).

A Comment on low s and low b

Bins with tiny s and tiny b can have large s/b (Louis Lyons: large s/sqrt(b) is suspicious)

Naturally occurring in HEP and others seeking discovery:

1) Each beam crossing has very small s and b but has the same s/b as neighboring beam crossings. Can make a histogram of the search for new physics separately for each beam crossing. Same s and b predictions, just scaled down very small.

   Adding is the same as a more elaborate combination if the histograms were accumulated under identical conditions (all rates, shapes, and systematics are the same)

2) Surveillance video catching a criminal – each frame has a small s, b, but still worthwhile to collect each frame (and analyze them separately)

The 2011 CERN Unfolding/ Deconvolution Workshop

http://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=107747

And look at the talks for Thursday, January 20, 2011 at the bottom of the page

Available Software, Tools, Documentation

CDF Statistics Committee
http://www-cdf.fnal.gov/physics/statistics/statistics_home.html
Useful for documentation. Provides advice for common, thorny questions

BaBar Statistics Working Group
http://www.slac.stanford.edu/BFROOT/www/Statistics/

ROOSTATS
https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome
A very complete toolset. I haven't used it (but I should have). It's in common use at the LHC

MCLIMIT
http://www-cdf.fnal.gov/~trj/mclimit/production/mclimit.html
Used on CDF, some use on D0 and LHC. Limits, cross sections, p-values, both Frequentist and Bayesian tools

PHYSTAT.ORG
http://www.phystat.org
Maintained by Jim Linnemann. We toolsmiths really should keep it up to date...

Available Software, Tools, Documentation

PDG Probability and Statistics Reviews (ed. Glen Cowan)

http://pdg.lbl.gov/2012/reviews/rpp2012-rev-probability.pdf
http://pdg.lbl.gov/2012/reviews/rpp2012-rev-statistics.pdf

If these links get out of date, just search pdg.lbl.gov for the mathematical reviews
Excellent brief reference, but maybe a little too brief to learn the material.

Good Reads:

Frederick James, "Statistical Methods in Experimental Physics", 2nd edition, World Scientific, 2006

Louis Lyons, "Statistics for Nuclear and Particle Physicists" Cambridge U. Press, 1989

Glen Cowan, "Statistical Data Analysis" Oxford Science Publishing, 1998

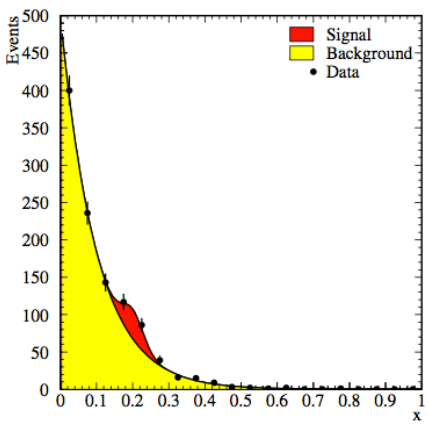
Roger Barlow, "Statistics, A guide to the Use of Statistical Methods in the Physical Sciences", (Manchester Physics Series) 2008.

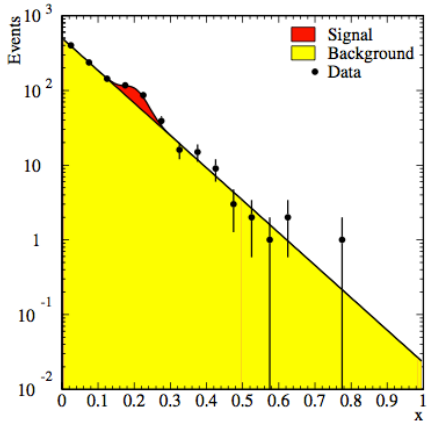
Bob Cousins, "Why Isn't Every Physicist a Bayesian" Am. J. Phys **63**, 398 (1995).

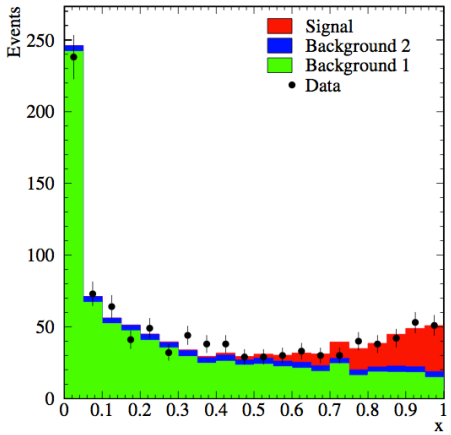
A Rather Complete and Lengthy Homework Exercise – Banff Challenge 2

Look at the bottom of this page:

http://www-cdf.fnal.gov/~trj/







Two problems:

1) Find a bump on a steeply falling background with low stats on one end

2) Perform a hypothesis test with an arbitrary discriminant distribution and two background components

Simulated datasets drawn from hidden true values supplied to challengees. 2 months given to solve problems. Answers and some approaches given on that web site.

Extras

Statistical Uncertainties on Systematic Uncertainties?

Answer: No. But some systematic uncertainties are difficult to evaluate properly.

See Roger Barlow's "Systematic errors: Facts and Fictions",
arXiv: hep-ex/0207026

The idea: If a systematic uncertainty is estimated by comparing two data samples or two MC samples, or data vs. MC, then if one or both of them have a limited size, then the magnitude of the systematic can be poorly constrained.

Ideally, work harder (run more MC) to get a better prediction of the expected signal and background, under all assumptions of systematic variation.

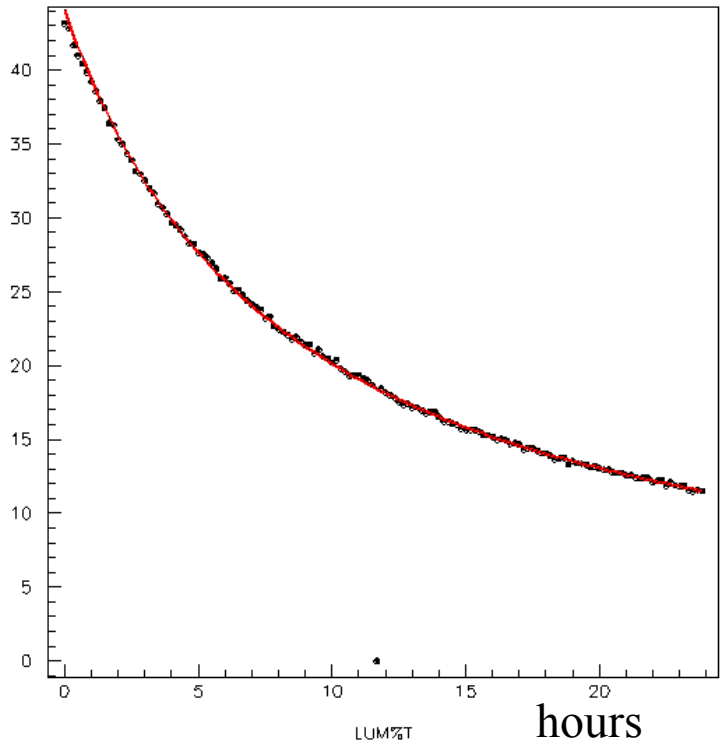
Monte Carlo Statistical Uncertainty is a Systematic Uncertainty
but don't double-count it for each separate MC variation of each nuisance parameter. Easy to do by comparing central vs. varied MC samples.

Statistically weak tests should be handed as cross checks. If they are consistent, consider the test to have passed, but do not add systematic uncertainty.
If they fail, however, and a discrepancy between two MC's or data and MC cannot be understood and fixed, then a systematic uncertainty is called for.

Outliers

- Sometimes they're obvious, often they are not.
- Best to make sure that the uncertainties on all points honestly include all known effects. Understand them!

Lum E30



L. Ristori,
Instantaneous
Luminosity at CDF vs. time
(a Tevatron store in 2005)

CERN-EP-98-046

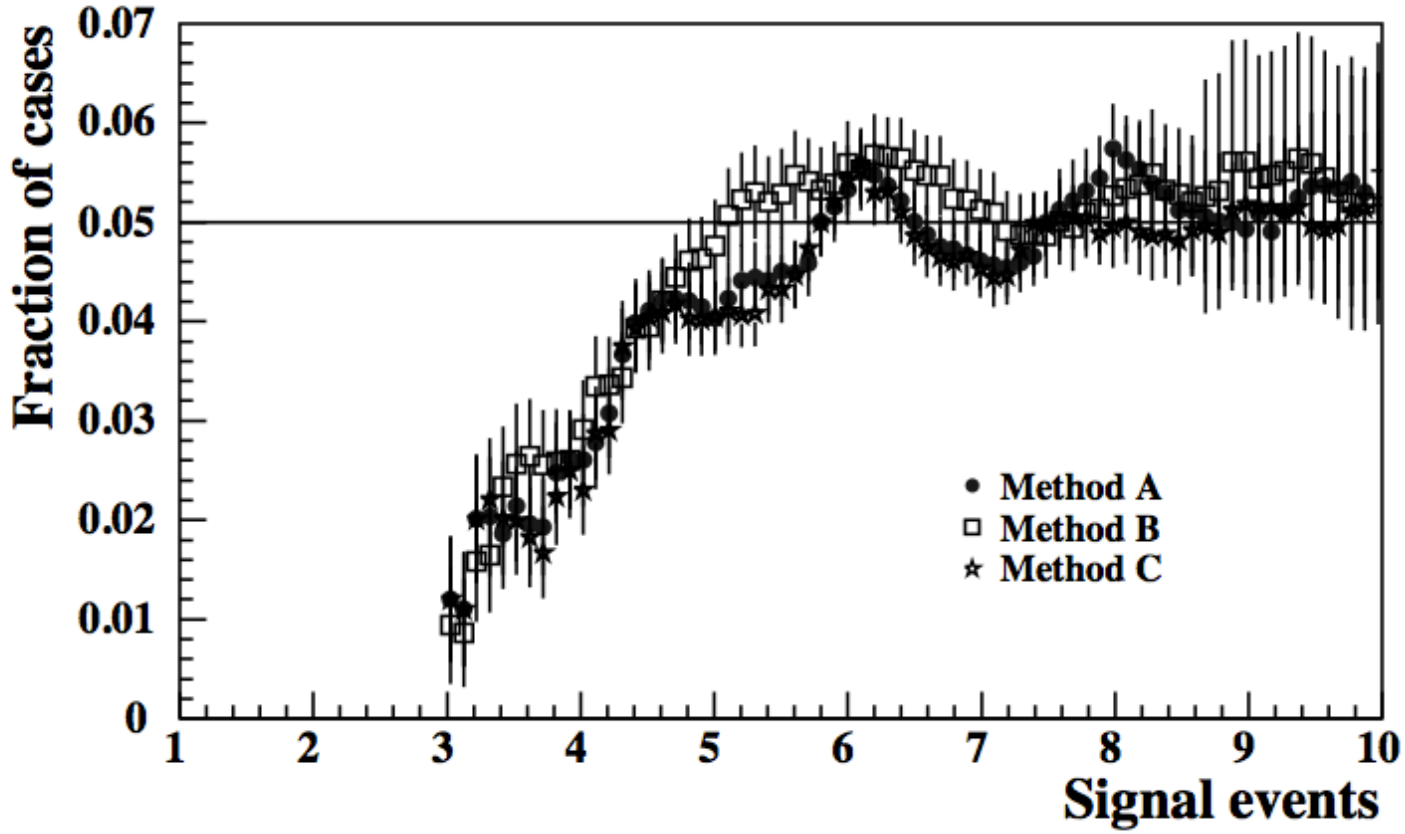
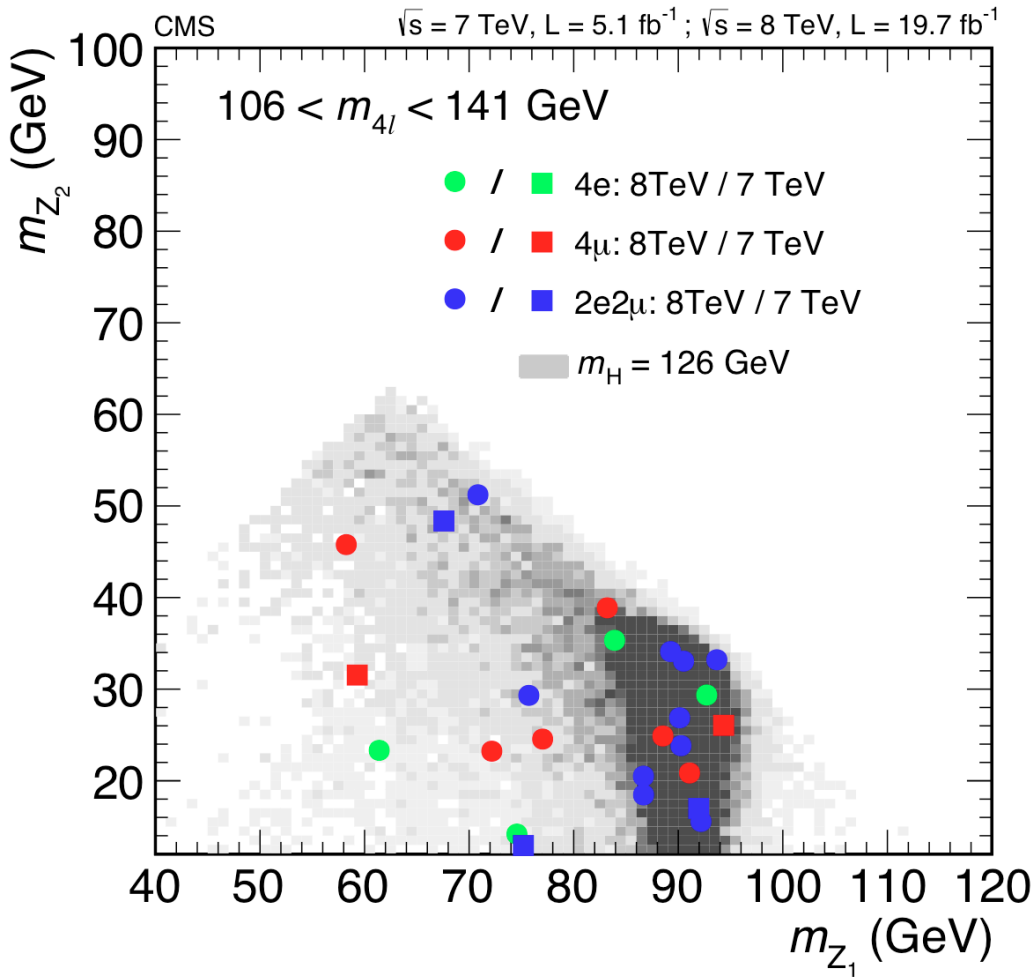
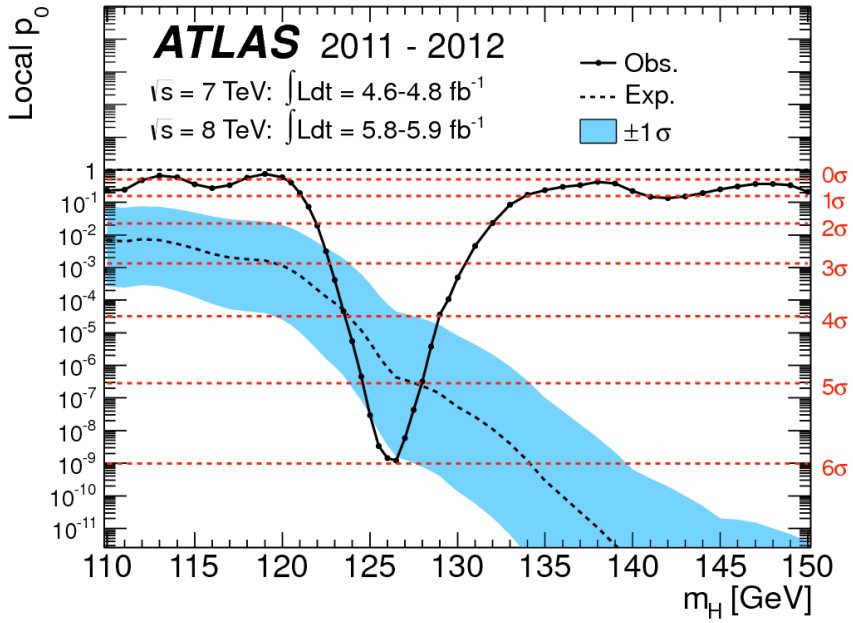


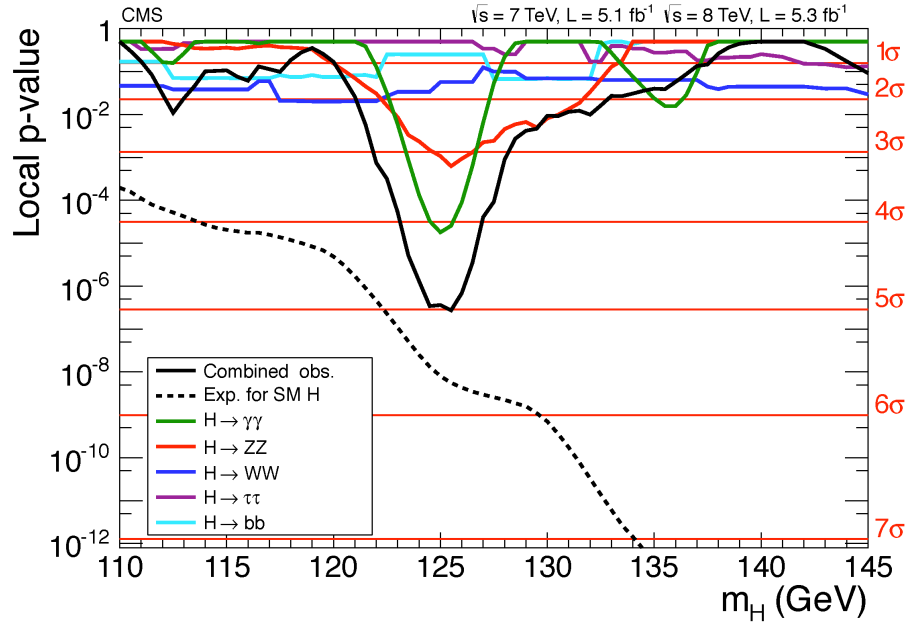
Figure 3: *Test (II); the fraction of gedanken experiments with signal+background in which an existing Higgs boson with 77* GeV/c^2 *mass is actually excluded using Methods A, B and C, as a function of the signal event rate.*


CMS
√s = 7 TeV, L = 5.1 fb-1 ; √s = 8 TeV, L = 19.7 fb-1
106 < m4l < 141 GeV
4e: 8TeV / 7 TeV
4μ: 8TeV / 7 TeV
2e2μ: 8TeV / 7 TeV
mH = 126 GeV
mZ2 (GeV)
mZ1 (GeV)
100
90
80
70
60
50
40
30
20
40
50
60
70
80
90
100
110
120

ATLAS and CMS's combinations of search channels – Discovery publications



ATLAS Combined H→γγ, H→ZZ, H→WW
Search Observation PLB 716, 1 (2012)



CMS Combined H→γγ, H→ZZ, H→WW, H→ττ, H→bb
Search Observation PLB 716, 30 (2012)

"Wiggles" in the expected p-value at m_H an artifact of using the best-fit nuisance parameters (in this case, signal acceptance) from the data when generating pseudo-experiments A true signal pushes up the best-fit acceptance (or at least doesn't push it down).

Analyses optimized only on MC and not data however.

Constraining the Couplings of the Higgs Boson to Fermions and Gauge Bosons

We follow the procedures and notation of the LHC Higgs Cross Section WG
A. David et al., arXiv:1209.0040

The model: SM-like, but

Hff couplings are scaled together by κ_f

HWW coupling is scaled by κ_W
HZZ coupling is scaled by κ_Z

For some studies, we scale the HWW and HZZ couplings by $\kappa_W=\kappa_Z=\kappa_V$

Standard Model is recovered if $\kappa_f = \kappa_W = \kappa_Z = 1$

Constraining Couplings

Step 1: Scale cross sections for each process according to couplings

$$\sigma(gg \to H) = \sigma_{SM}(gg \to H)(0.95\kappa_f^2 + 0.05\kappa_f\kappa_V)$$

$$\sigma(WH) = \sigma_{SM}(WH)\kappa_V^2$$

$$\sigma(ZH) = \sigma_{SM}(ZH)\kappa_V^2$$

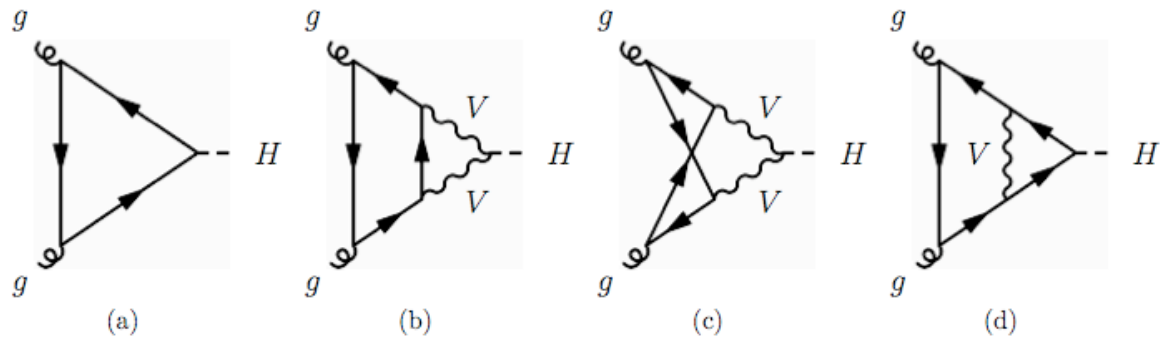
LO relations, but mostly true at higher order (most QCD affects the colored initial-state particles. There is gg→WH at higher order, however.)

$$\sigma(VBF) = \sigma_{SM}(VBF)\kappa_V^2$$

-- Pretty much by definition! Unless NNLO VBF includes the EW ggH piece. From Bolzoni, Moch, Maltoni, and Zaro's papers, it seems as if the EW ggH piece is not in the NNLO VBF calculation.

A. David *et al.*, LHCHXSWG-2012-001. arXiv:1209.0040

Two-Loop Electroweak Contributions to the ggH Coupling

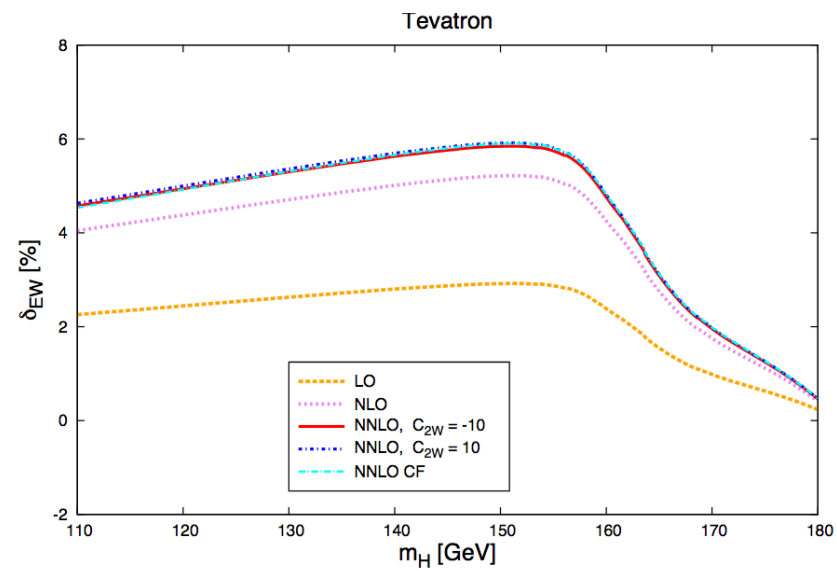


Aglietti, Bonciani, Degrassi and Vicini, arXiv:hep-ph/0610033

Anastasiou, Boughezal, Petriello, JHEP 0904 003 (2009)

Actis, Passarino, Sturm, Uccirati PLB 670, 12 (2008)

Grazzini and de Florian PLB 674, 291 (2009)



- Approximately a 5% contribution to the gg→H production cross section at m_H=125 GeV/c^2. Main contribution is from interference with the LO process.
- Contribution not included in VBF calculation (HAWK: Denner, Dittmaier, Mück)

Constraining Couplings

Step 2: Recompute all Higgs boson decay branching ratios from scaled partial widths

$$\Gamma(H \to gg) = \Gamma_{SM}(H \to gg)(0.95\kappa_f^2 + 0.05\kappa_f\kappa_V)$$

$$\Gamma(H \to W^+W^-) = \Gamma_{SM}(H \to W^+W^-)\kappa_V^2$$

$$\Gamma(H \to b\bar{b}) = \Gamma_{SM}(H \to b\bar{b})\kappa_f^2$$

$$\Gamma(H \to \tau^+\tau^-) = \Gamma_{SM}(H \to \tau^+\tau^-)\kappa_f^2$$

$$\Gamma(H \to c\bar{c}) = \Gamma_{SM}(H \to c\bar{c})\kappa_f^2$$

$$\Gamma(H \to ZZ) = \Gamma_{SM}(H \to ZZ)\kappa_V^2$$

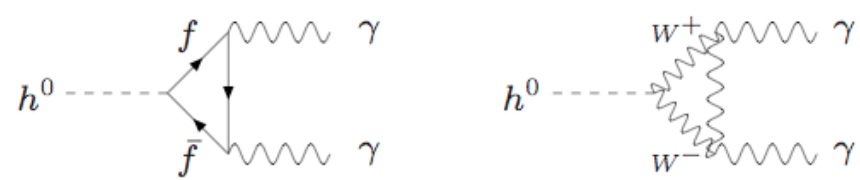
$$\Gamma(H \to \gamma\gamma) = \Gamma_{SM}(H \to \gamma\gamma)\left|\alpha\kappa_V + \beta\kappa_f\right|^2$$

Other modes, like $H \to \mu^+\mu^-$ and $H \to Z\gamma$ have very small widths

$$Br(H \to X\bar{X}) = \frac{\Gamma(H \to X\bar{X})}{\sum_i \Gamma_i}$$

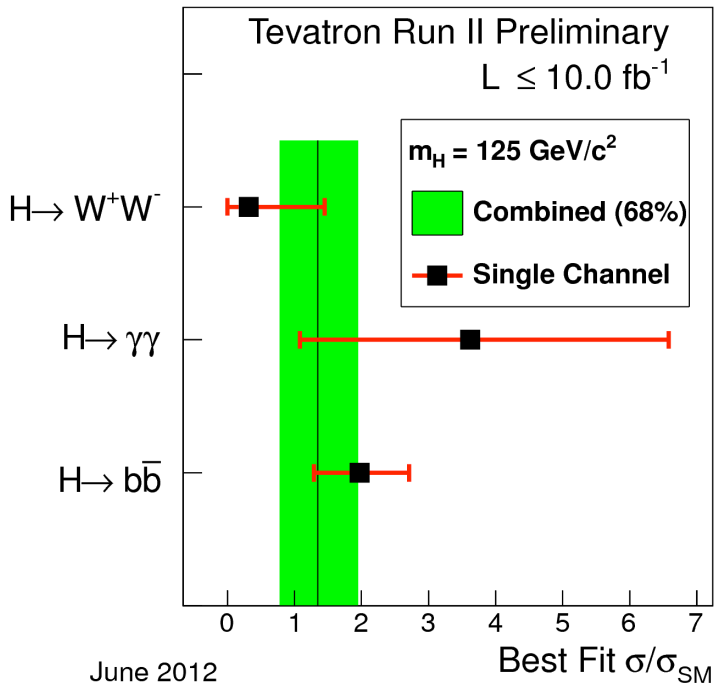
α and β come from Spira et al., arXiv:hep-ph/9504378

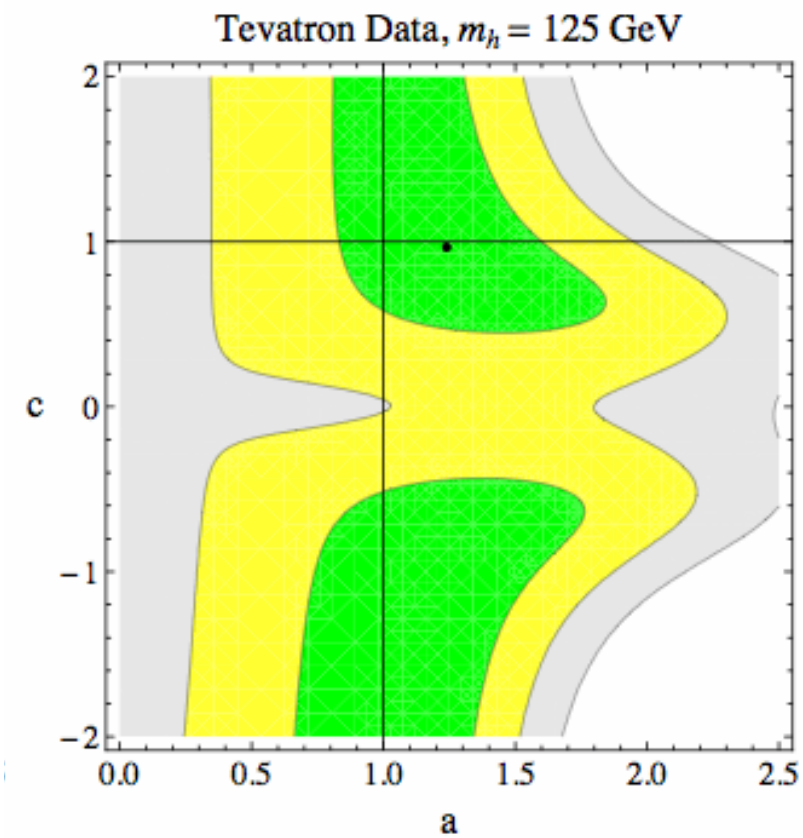
α=1.28 β = -0.21



Some work from theorists

Espinosa, Grojean, Mühlleitner, and Trott, "First Glimpse at Higgs' Face", arXiv:1207.1717v2 (updated Aug. 21 with post-ICHEP data) JHEP 1212, 045 (2012).





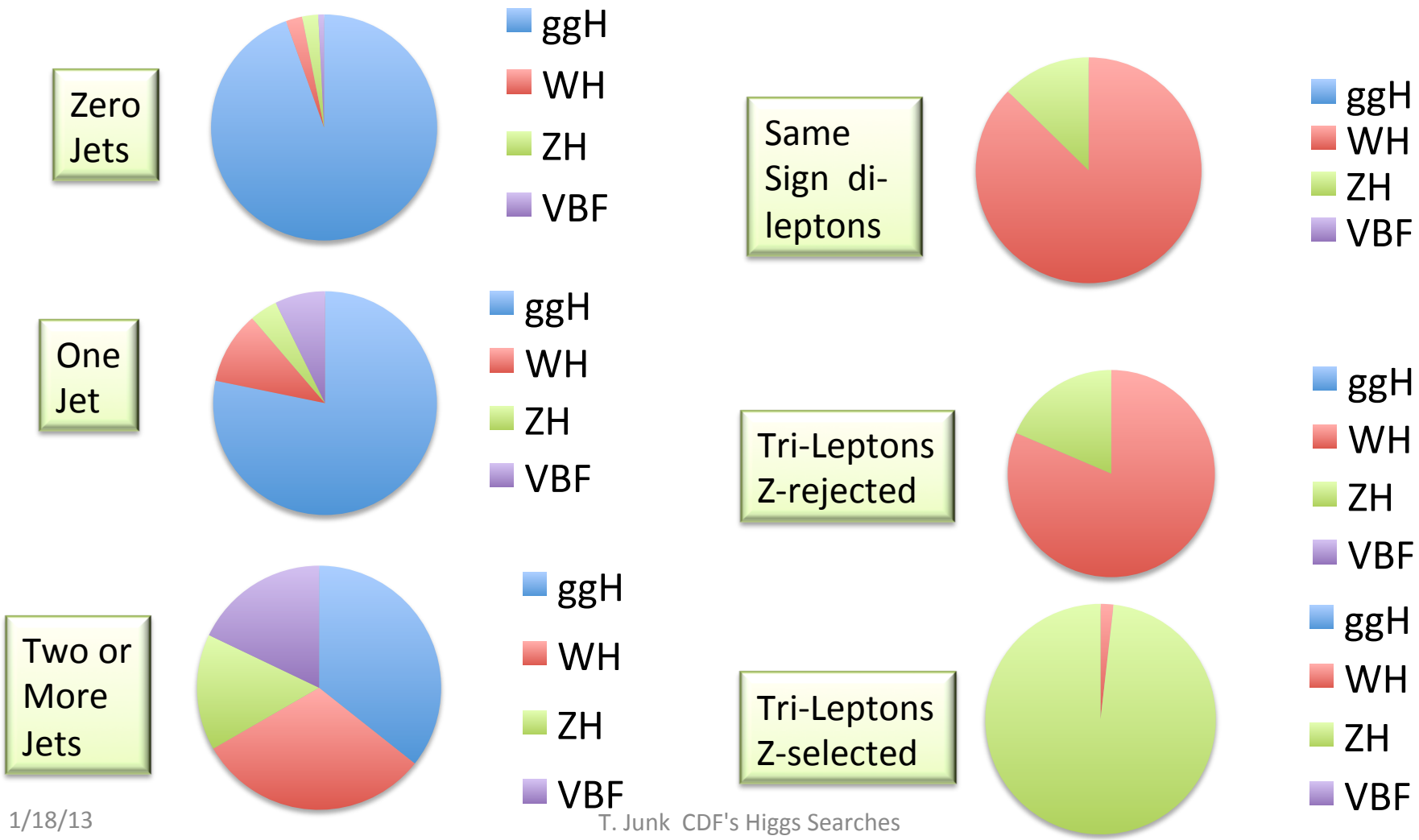
$a=\kappa_V$, $c=\kappa_F$

More Complete Treatment of Signal Scalings

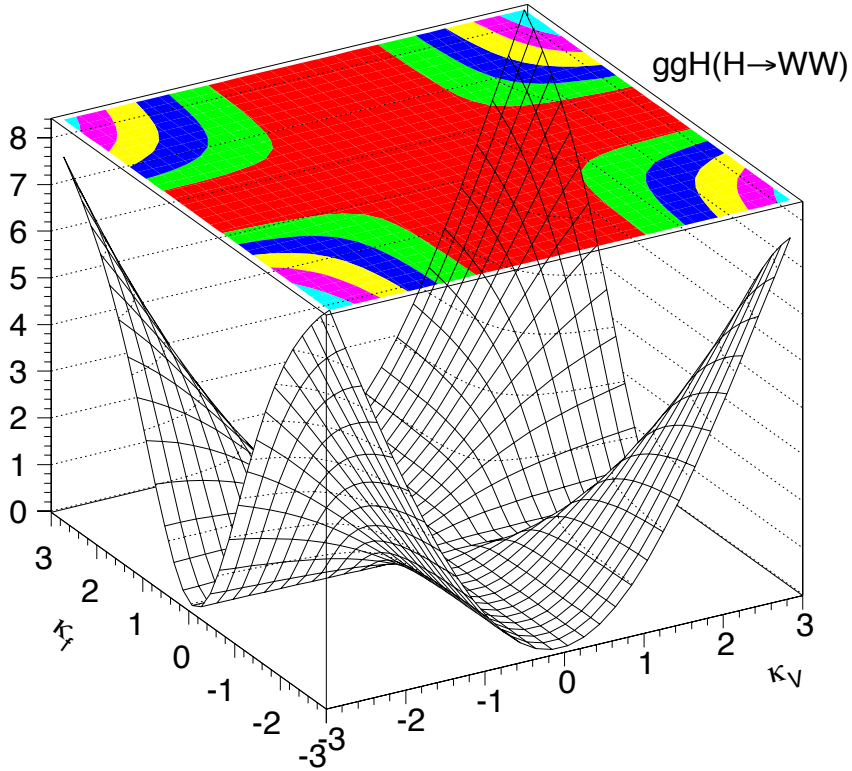
Example: HWW Search. All Signals have H→WW, but different mixtures of production mechanisms in each search channel.

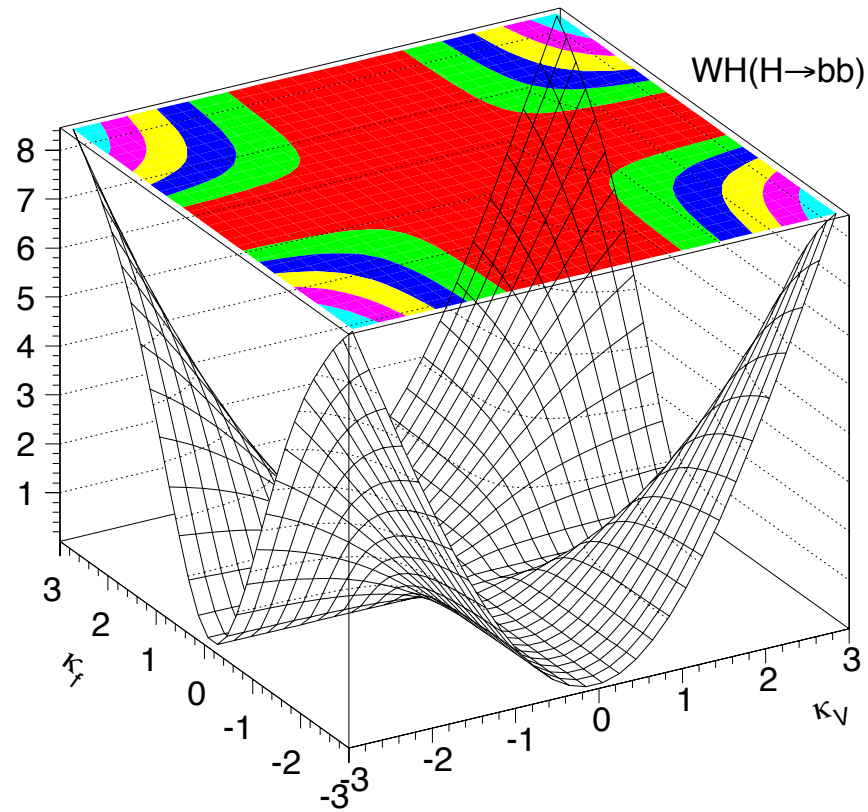
Each signal contribution should be scaled by the appropriate factor

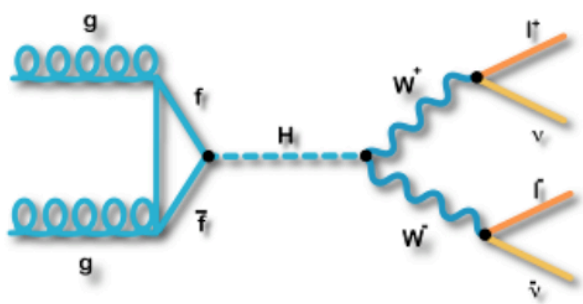
Pie charts show relative contributions.

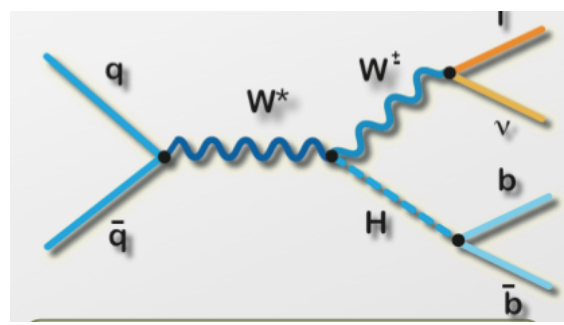


Signal Rate Enhancement Factors for gg→H→WW and WH->lvbb

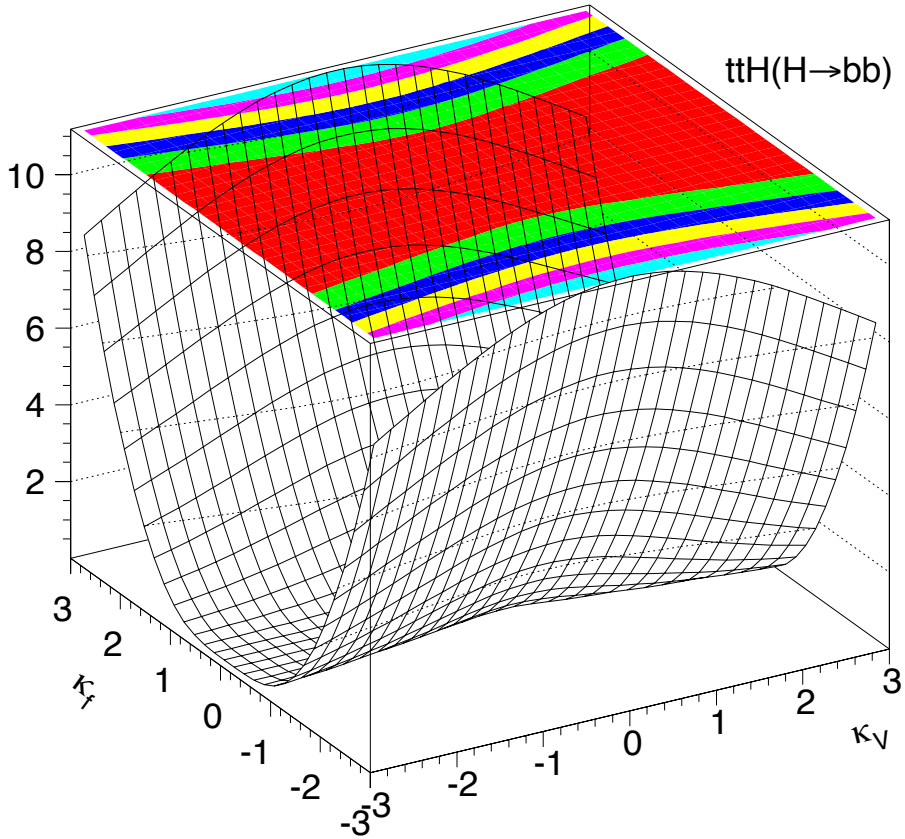


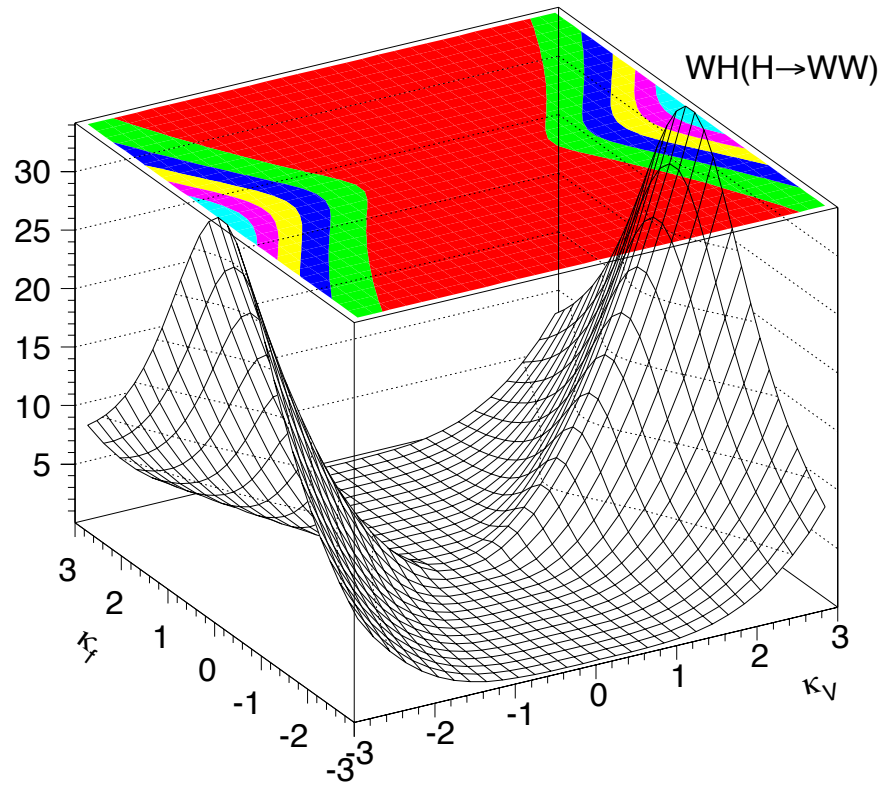


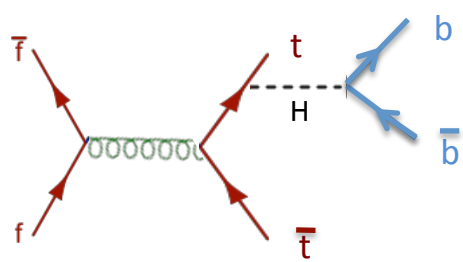


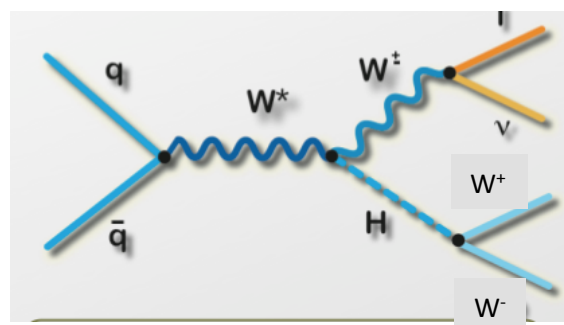


Signal Rate Enhancement Factors for ttH→ttbb and WH->WWW

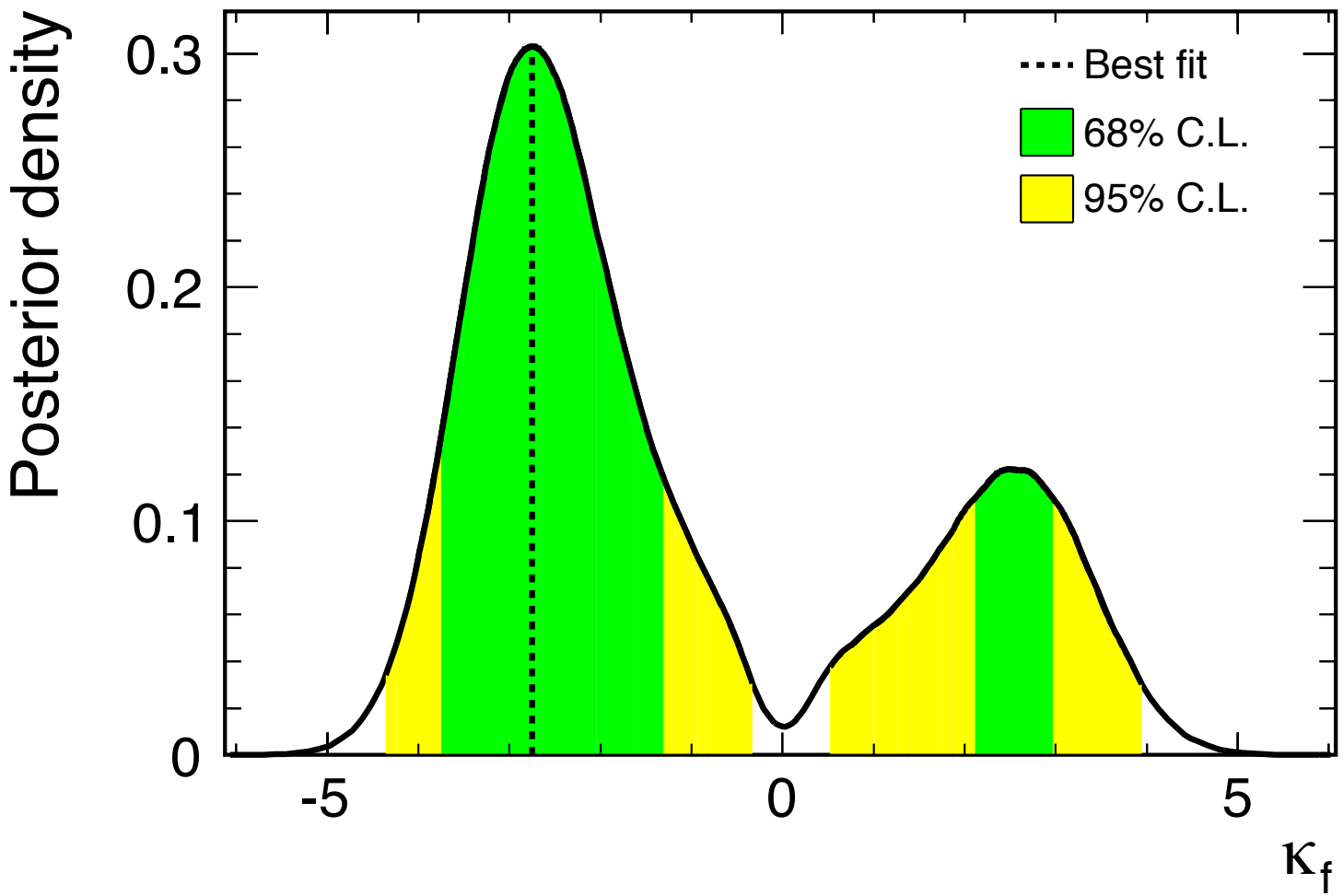








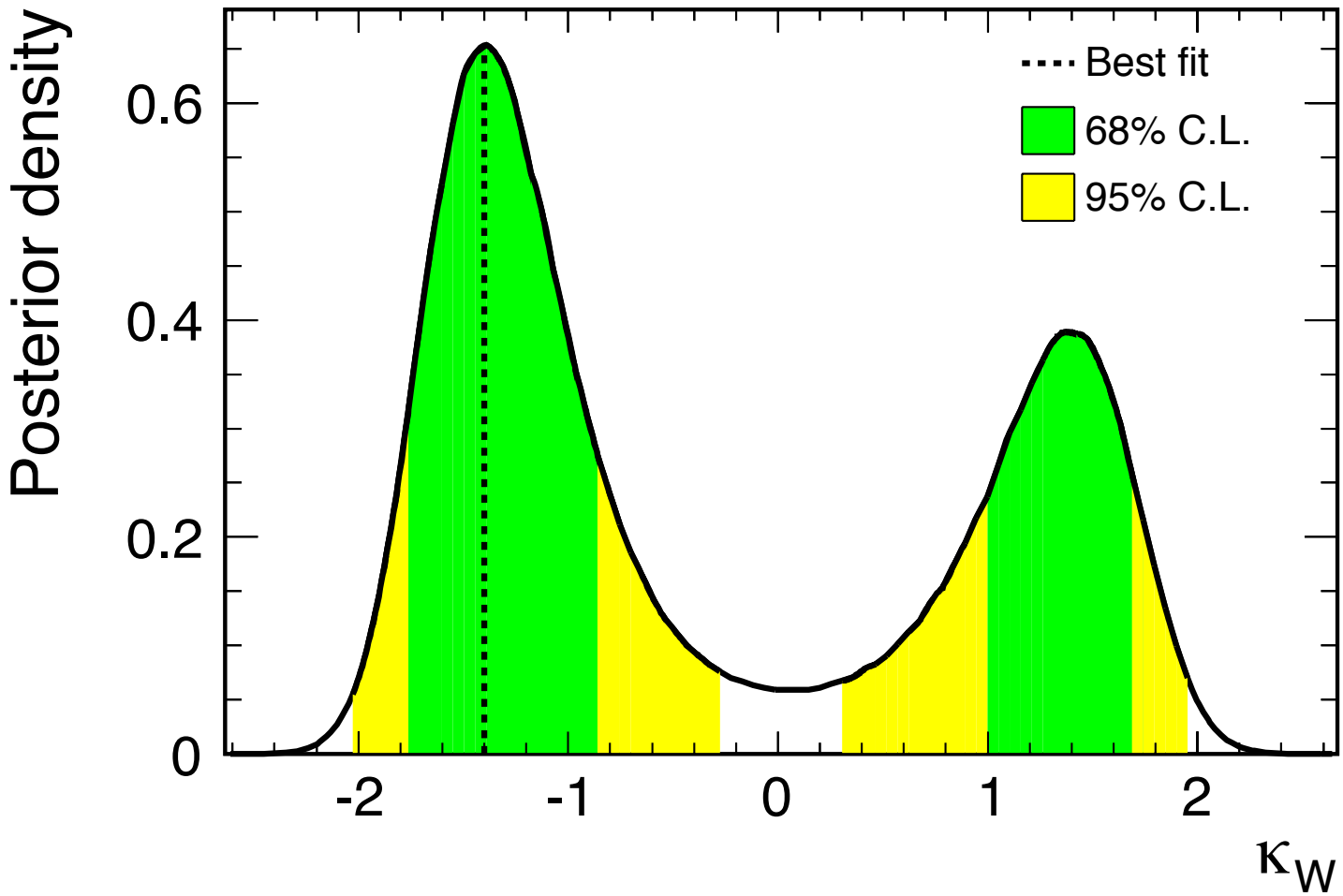
Posterior Constraint on the Hff Coupling Factor κ_f



- Uniform prior assumed
- $\kappa_W = \kappa_Z = 1$ assumed

Excess in the $H \rightarrow \gamma\gamma$ searches drives the asymmetry from positive and negative coupling scale factors

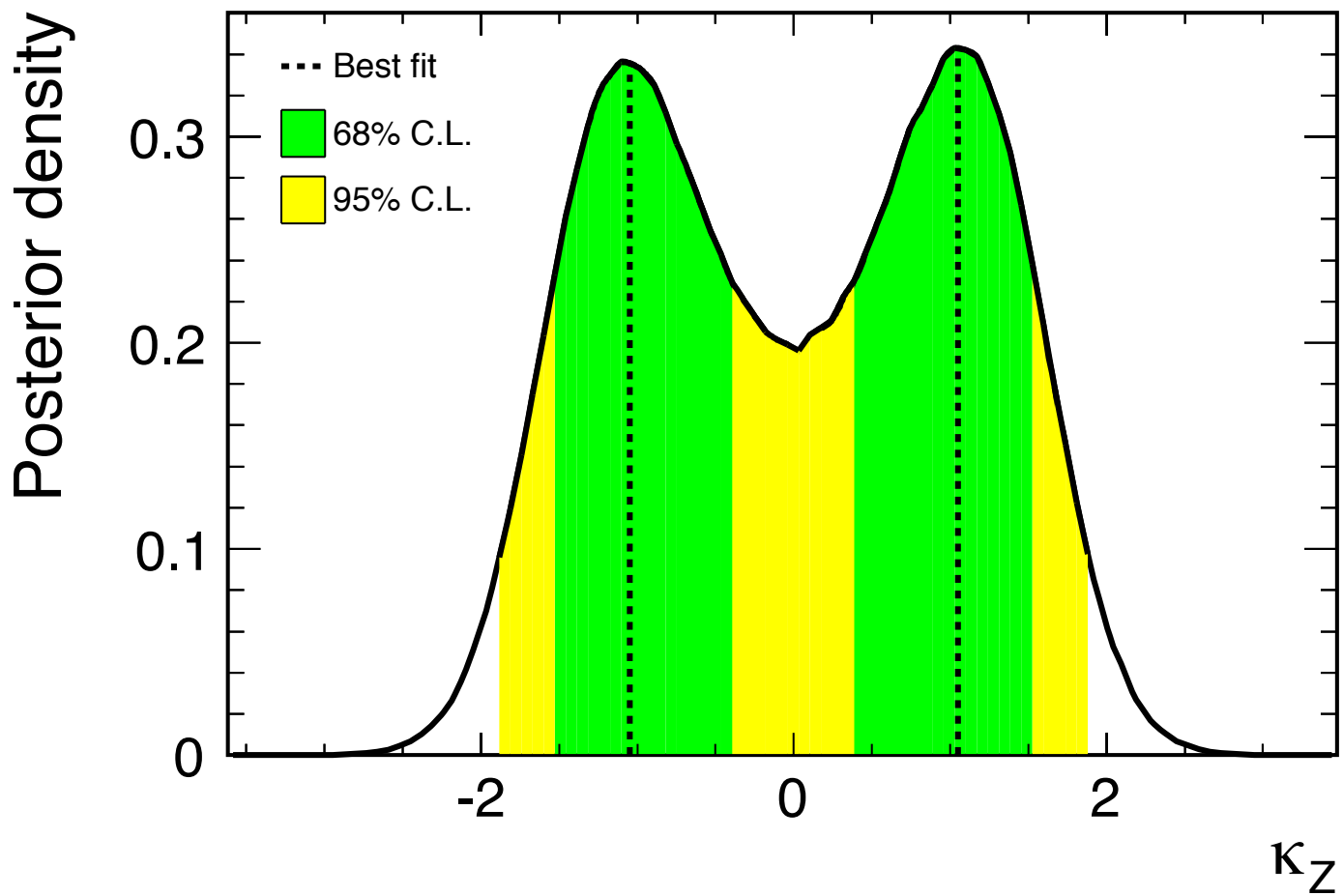
Posterior Constraint on the HWW Coupling Factor κ_W



- Uniform prior assumed
- $\kappa_f = \kappa_Z = 1$ assumed

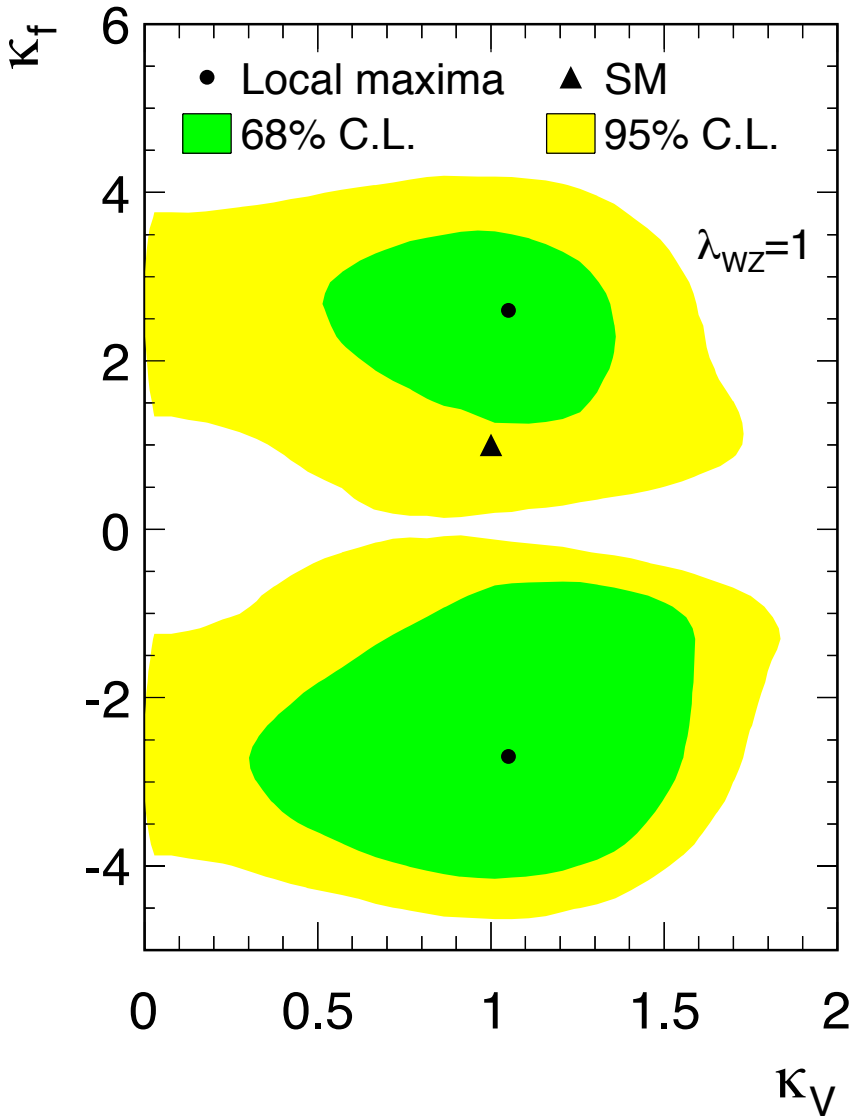
Excess in the H→γγ searches drives the asymmetry from positive and negative coupling scale factors

Posterior Constraint on the HWW Coupling Factor κ_Z



- Uniform prior assumed
- $\kappa_f = \kappa_W = 1$ assumed

Two-Dimensional Constraints: Bosons vs. Fermions

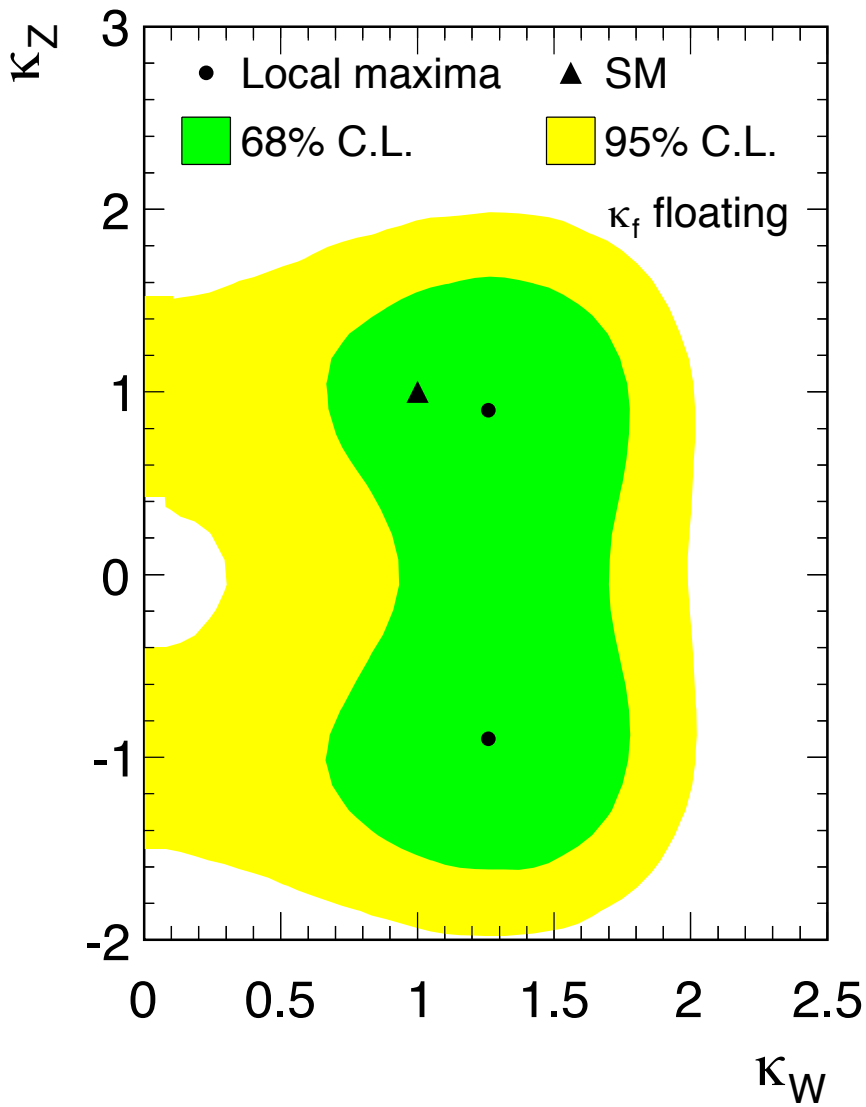


Uniform prior assumed

Large κ_V with small κ_f constrained by trileptons, same-sign dileptons

Large κ_f and small κ_V constrained by $t\bar{t}H \rightarrow t\bar{t}bb$

Two-Dimensional Constraints: W vs. Z Couplings



Uniform prior assumed

κ_f integrated over ("marginalized")

Less constraint on HZZ than on HWW

All couplings consistent with SM predictions