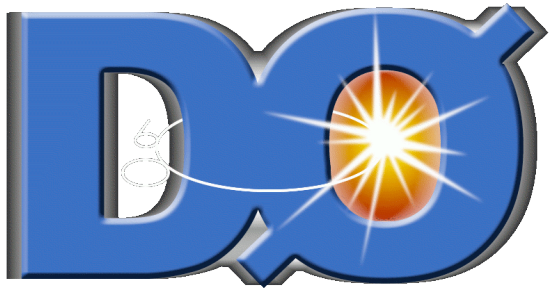
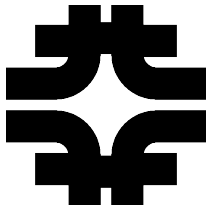
Statistical Techniques for Combining Tevatron Higgs Boson Searches

Tom Junk
Fermilab





- Exchanging Experimental Results
- Bayesian Limits
- CL_s Limits
- Discovery Techniques
- Tevatron Combined Results

http://tevnphwg.fnal.gov
http://www-cdf.fnal.gov/physics/new/hdg/hdg.html
http://www-d0.fnal.gov/Run2Physics/WWW/results/higgs.htm

Exchanging Experimental Results for Combination

At LEP and at the Tevatron, we exchanged histograms of observed and predicted events.

Many advantages:

- Crosscheck analyzers' work:
  Signal and background checksum
  Limit/discovery recalculations
  check for "broken" bins
  - s>0 when b=0
  - any observation or prediction <0
- Can make control plots
- Can try a great variety of statistical treatments
  -- Profile Likelihood, Bayesian, CL_s and compare each one
- Can make expected limits/LLR distributions without approximations
- Can draw the $\pm 1\sigma$, $\pm 2\sigma$ bands on expected limits with MC
- Can point to excesses and deficits to explain why limits and p-values are as they are
- Can accommodate new cross sections and branching ratios by scaling
- Can pick and choose signals if more than one expected (e.g., do 4th gen analysis with H→WW without WH, ZH and VBF)
- Pre-binned histograms mean combiners don't have to choose binning, reducing mistakes, inconsistencies
- possibly less work for the analyzer

Disadvantages:

- Lots of work/CPU!
- Have to share preliminary histos (your competitors may find your mistakes!)

Exchanging Experimental Results for Combination

Systematic uncertainties itemized by named source

- Asymmetric Rate errors on each predicted component

- Shape errors supplied as alternate shape histograms
  Bin-by-bin ratios are inconvenient -- example m_{jj} histogram where the variation is "horizontal" and not vertical.
  Need shape interpolators/extrapolators to use them.
  Typically $\pm 1\sigma$ shape variations are explored one source at a time by analyzers. Analyzers will ask combiners to extrapolate out to arbitrary $\pm n\sigma$ shapes (!)
  -- practical difficulty: How to estimate 5σ systematics?

- Bin-by-bin independent uncertainties (MC statistics)

- Names used to categorize correlations in a way easy to understand and check

- Give names to exchanged template histograms please!

Cross Section and Branching Ratio Alignment

TABLE I: The production cross sections and decay branching fractions for the SM Higgs boson assumed for the combination.

m_H (GeV/c^2)	$\sigma_{gg\to H}$ (fb)	σ_{WH} (fb)	σ_{ZH} (fb)	σ_{VBF} (fb)	$\sigma_{t\bar{t}H}$ (fb)	$B(H\to b\bar{b})$ (%)	$B(H\to c\bar{c})$ (%)	$B(H\to\tau^+\tau^-)$ (%)	$B(H\to W^+W^-)$ (%)	$B(H\to ZZ)$ (%)	$B(H\to\gamma\gamma)$ (%)
100	1861	291.9	169.8	99.5	8.000	80.33	3.542	7.920	1.052	0.1071	0.1505
105	1618	248.4	145.9	93.3	7.062	78.57	3.463	7.821	2.307	0.2035	0.1689
110	1413	212.0	125.7	87.1	6.233	75.90	3.343	7.622	4.585	0.4160	0.1870
115	1240	181.9	108.9	79.07	5.502	71.95	3.169	7.288	8.268	0.8298	0.2029
120	1093	156.4	94.4	71.65	4.857	66.49	2.927	6.789	13.64	1.527	0.2148
125	967	135.1	82.3	67.37	4.279	59.48	2.617	6.120	20.78	2.549	0.2204
130	858	116.9	71.9	62.5	3.769	51.18	2.252	5.305	29.43	3.858	0.2182
135	764	101.5	63.0	57.65	3.320	42.15	1.854	4.400	39.10	5.319	0.2077
140	682	88.3	55.3	52.59	2.925	33.04	1.453	3.472	49.16	6.715	0.1897
145	611	77.0	48.7	49.15	2.593	24.45	1.075	2.585	59.15	7.771	0.1653
150	548	67.3	42.9	45.67	2.298	16.71	0.7345	1.778	68.91	8.143	0.1357
155	492	58.9	37.9	42.19	2.037	9.88	0.4341	1.057	78.92	7.297	0.09997
160	439	50.8	33.1	38.59	1.806	3.74	0.1646	0.403	90.48	4.185	0.05365
165	389	44.6	30.0	36.09	1.607	1.29	0.05667	0.140	95.91	2.216	0.02330
170	349	40.2	26.6	33.58	1.430	0.854	0.03753	0.093	96.39	2.351	0.01598
175	314	35.6	23.7	31.11	1.272	0.663	0.02910	0.073	95.81	3.204	0.01236
180	283	31.4	21.1	28.57	1.132	0.535	0.02349	0.059	93.25	5.937	0.01024
185	255	28.2	18.9	26.81	1.004	0.415	0.01823	0.046	84.50	14.86	0.008128
190	231	25.1	17.0	24.88	0.890	0.340	0.01490	0.038	78.70	20.77	0.006774
195	210	22.4	15.3	23	0.789	0.292	0.01281	0.033	75.88	23.66	0.005919
200	192	20.0	13.7	21.19	0.700	0.257	0.01128	0.029	74.26	25.33	0.005285

All channels must use consistent predictions

Tevatron Correlated Systematic Errors I

Total Systematic error count: 129 (not counting bin-by-bin errors)

Note: correlation in errors on backgrounds between experiments helps sensitivity! One experiment is another experiment's control sample.

Luminosity: 3.8% Correlated CDF and DØ σ_{inel}(ppbar)
4.4% detector-specific

Diboson Cross Sections:
Defined for
$75 < m_H < 105$ GeV

$$\sigma_{W^+W^-} = 11.34^{+0.56}_{-0.49}\ (\text{scale})\ ^{+0.35}_{-0.28}(\text{PDF})\text{pb}$$

$$\sigma_{W^\pm Z} = 3.22^{+0.20}_{-0.17}\ (\text{scale})\ ^{+0.11}_{-0.08}\ (\text{PDF})\ \text{pb}$$

$$\sigma_{ZZ} = 1.20^{+0.05}_{-0.04}\ (\text{scale})\ ^{+0.04}_{-0.03}\ (\text{PDF})\ \text{pb}$$

ttbar Cross Section: Moch and Uwer, evaulated at
m_t=173 ± 1.2 GeV is (using MSTW2008 PDFs)

$$\sigma_{t\bar{t}} = 7.04^{+0.24}_{-0.36}\ (\text{scale}) \pm 0.14(\text{PDF}) \pm 0.30(\text{mass})$$

Tevatron Correlated Systematic Errors II

Signal Cross Section uncertainties (using MSTW PDFs)

WH, ZH: ± 5%

gg→H: ±17.5% (weighted scale over jet samples)
± 12.5% (weighted PDF). Pt spectra reweighted to NNLO+NNLL predictions

Errors taken from Anastasiou, Dissertori, Stockli, and Webber

VBF: ± 10%

Theory errors applied to SM interpretations, but taken off for cross-section times branching ratio limits.

CDF-DØ Uncorrelated errors:
- K-factors (data driven)
- trigger efficiency
- b-tag efficiency and mistags
- jet energy scale
- lepton ID, fakes and conversions
- MET modeling

Correlated *within* CDF and D0 where appropriate

Steps Required for Combination

- Histograms and named rate ***and shape*** errors exchanged
- Check stacked histograms and systematic tables
  with analysis documentation total counts:
    data, signal, background
    - look for bins with b=0 and have data events (bad!)
- Repeat individual channel limits -- compare against
  approved results.
- Assess correlations on systematics

CDF and D0 teams each do three combinations, using Bayesian and CL_s techniques.

CDF

D0

Tevatron

Consistency at the better than 10% level required for all combinations at all test masses. Quote Bayesian limits (historical)

Mini-Review: Bayesian Limits

$$L(r,\theta) = \prod_{channels} \prod_{bins} P_{Poiss}(data \mid r,\theta)$$

Where r is an overall signal scale factor, and θ represents all nuisance parameters.

$$P_{Poiss}(data \mid r,\theta) = \frac{(rs_i(\theta) + b_i(\theta))^{n_i} e^{-(rs_i(\theta)+b_i(\theta))}}{n_i!}$$

where n_i is observed in each bin i, s_i is the predicted signal for a fiducial model (SM), and b_i is the predicted background. Dependence of s_i and b_i on θ includes rate, shape, and bin-by-bin independent uncertainties.

Mini-Review: Bayesian Limits

Including uncertainties on nuisance parameters θ

$$L'(data \mid r) = \int L(data \mid r,\theta)\pi(\theta)d\theta$$

where $\pi(\theta)$ encodes our prior belief in the values of the uncertain parameters. Usually Gaussian centered on the best estimate and with a width given by the systematic. The integral is high-dimensional. Markov Chain MC integration is quite useful!

Useful for a variety of results:

Limits:
$$0.95 = \int_0^{r_{\lim}} L'(data \mid r)\pi(r)dr$$

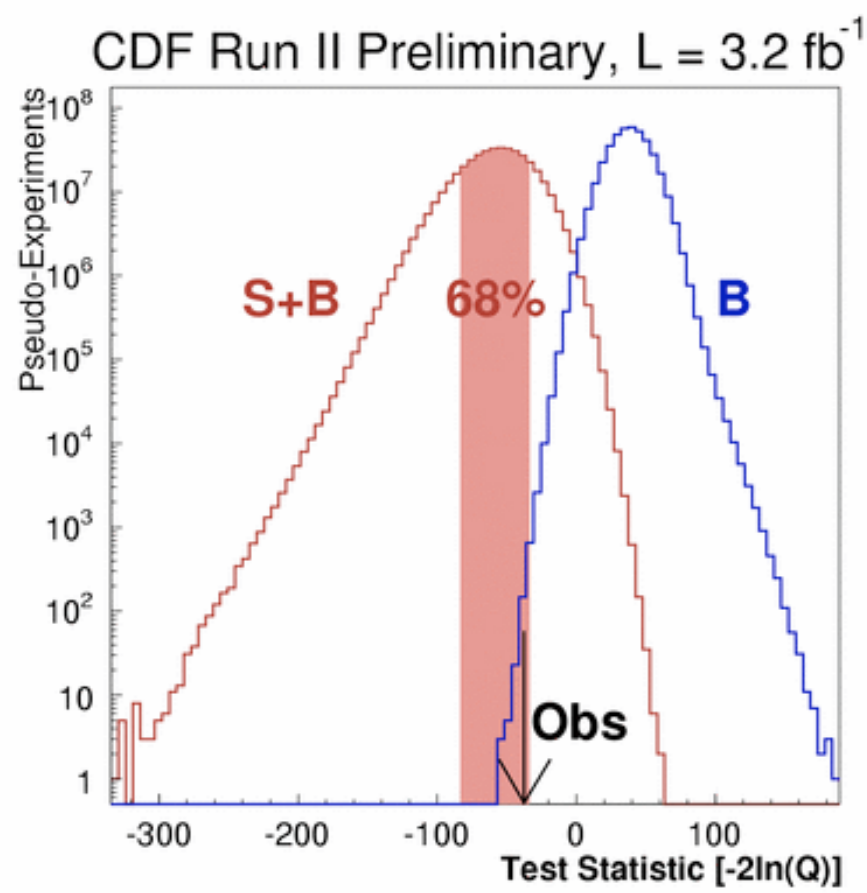
Typically $\pi(r)$ is constant
Other options possible.
Sensitivity to priors a concern.

Measure *r*:
$$0.68 = \int_{r_{low}}^{r_{high}} L'(data \mid r)\pi(r)dr$$

$$r = r_{\max}{}^{+(r_{high}-r_{\max})}_{-(r_{\max}-r_{low})}$$

Usually: shortest interval containing 68% of the posterior (other choices possible)

Discovery with p-Values



Example: CDF single top.

$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

100 M s+b and b-only pseudoexperiments, each with fluctuated nuisance parameters, and fit twice.

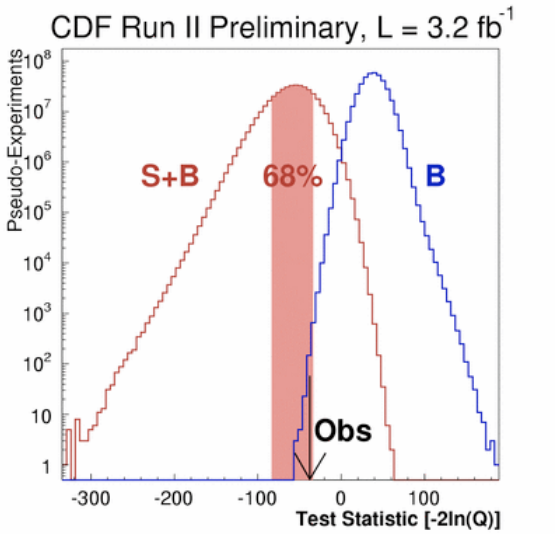
5σ: p-value of 2.77×10^{-7} or less.

3σ: p-value of 1.35×10^{-3} or less

2σ: p-value of 2.28% or less

Buzzword: "Prior Predictive ensemble"

Fitting and Fluctuating



$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

- Monte Carlo pseudoexperiments are used to get p-values.
- Test statistic -2lnQ is not uncertain for the data.
- Distribution from which -2lnQ is drawn is uncertain!

- Nuisance parameter fits in numerator and denominator of -2lnQ **do not incorporate systematics into the result**.
  Example -- 1-bin search; all test statistics are equivalent to the event count, fit or no fit.
- Instead, we fluctuate the probabilities of getting each outcome since those are what we do not know. Each pseudoexperiment gets random values of nuisance parameters.
- Can also try values of nuisance parameters that maximize the p-value, but that's very conservative (called the supremum p-value, still needs choices of parameter ranges).
- Why fit at all? It's an optimization. Fitting reduces sensitivity to the uncertain true values and the fluctuated values. For stability and speed, you can choose to fit a subset of nuisance parameters (the ones that are constrained by the data). Or do constrained or unconstrained fits, it's your choice.
- If not using pseudoexperiments but using Wilk's theorem, then the fits are important for correctness, not just optimality.

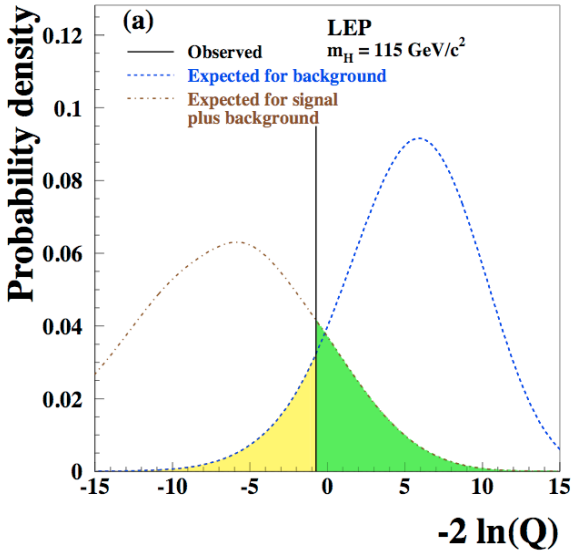
Mini-Review: CL_s Limits

- Based on p-values using the log likelihood ratio as the test statistic. Neyman-Pearson lemma says LLR is the uniformly most powerful test statistic, although the Neyman-Pearson one fits for the parameter of interest, not just the nuisance parameters, making the null hypothesis a subset of the test hypothesis

$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

Glen Cowan (really Pearson's) LLR also fits for s (actually $r \times s$) in the numerator, while $r = 0$ in the denominator

Mini-Review: CL_s Limits



p-values:

Yellow area = $1\text{-}CL_b = 1\text{-}P(-2\ln Q > -2\ln Q_{obs} \mid \text{b only})$

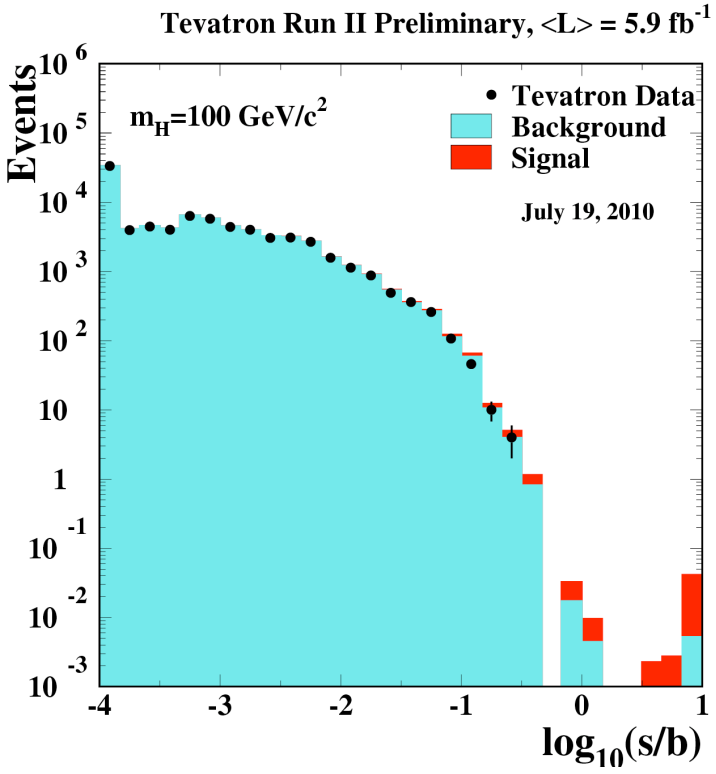
Green area = $CL_{s+b} = P(-2\ln Q > -2\ln Q_{obs} \mid \text{s+b})$

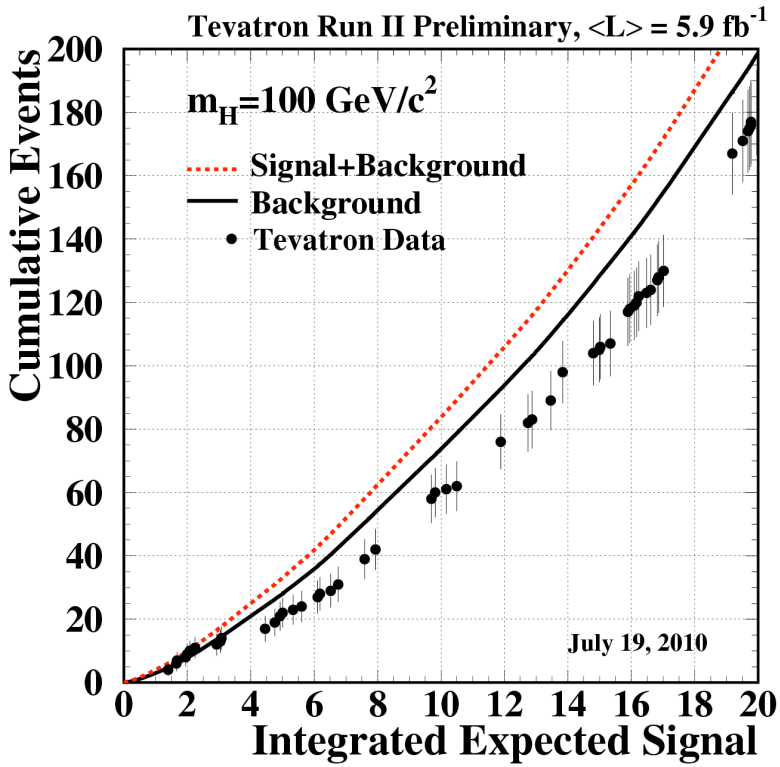
$CL_s \equiv CL_{s+b}/CL_b \geq CL_{s+b}$

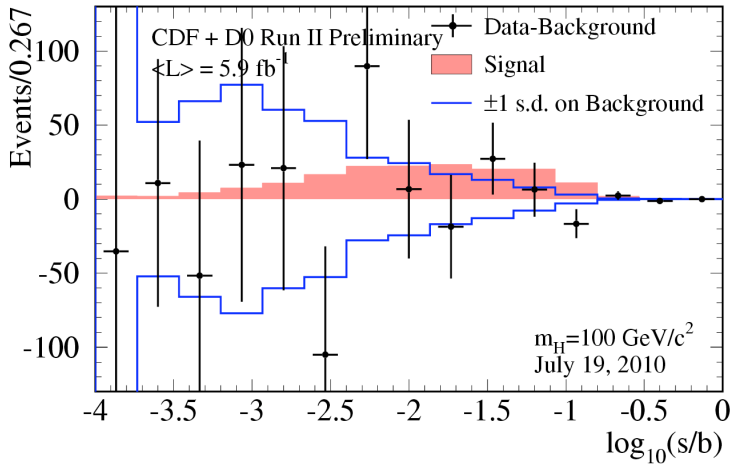
Exclude if $CL_s < 0.05$

Vary r until $CL_s = 0.05$ to get r_{lim}

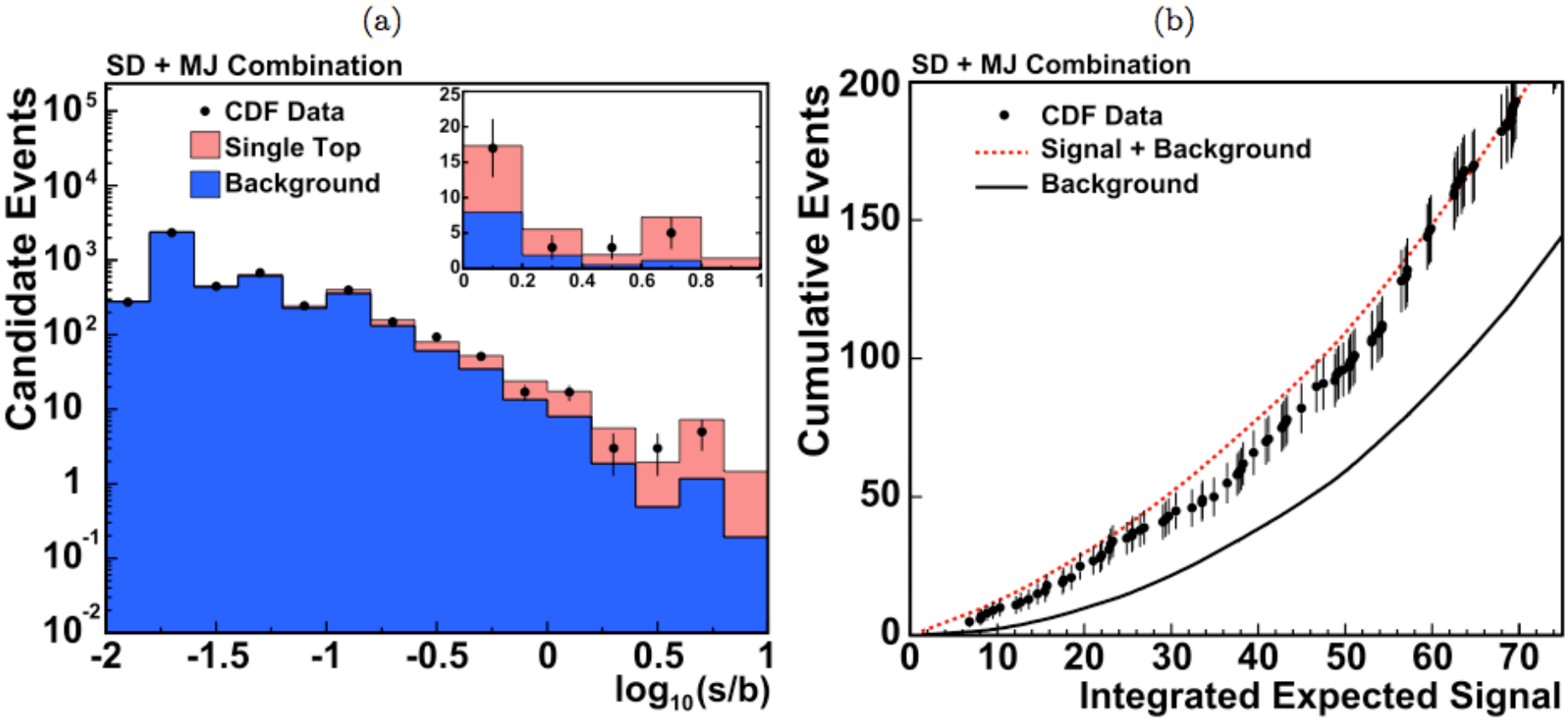
- Advantages:
  - Exclusion and Discovery p-values are consistent.
    Example -- a 2σ upward fluctuation of the data with respect to the background prediciton appears both in the limit and the p-value as such
  - Does not exclude where there is no sensitivity
    (big enough search region with small enough resolution and you get a 5% dusting of random exclusions with CL_{s+b})


Tevatron Run II Preliminary, <L> = 5.9 fb-1
Events
mH=100 GeV/c2
Tevatron Data
Background
Signal
July 19, 2010
log10(s/b)

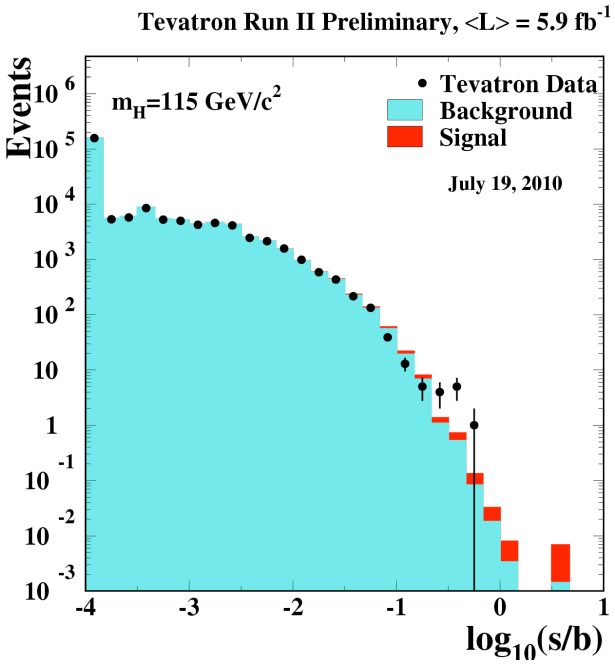

Tevatron Run II Preliminary, <L> = 5.9 fb-1
Cumulative Events
mH=100 GeV/c2
Signal+Background
Background
Tevatron Data
July 19, 2010
Integrated Expected Signal

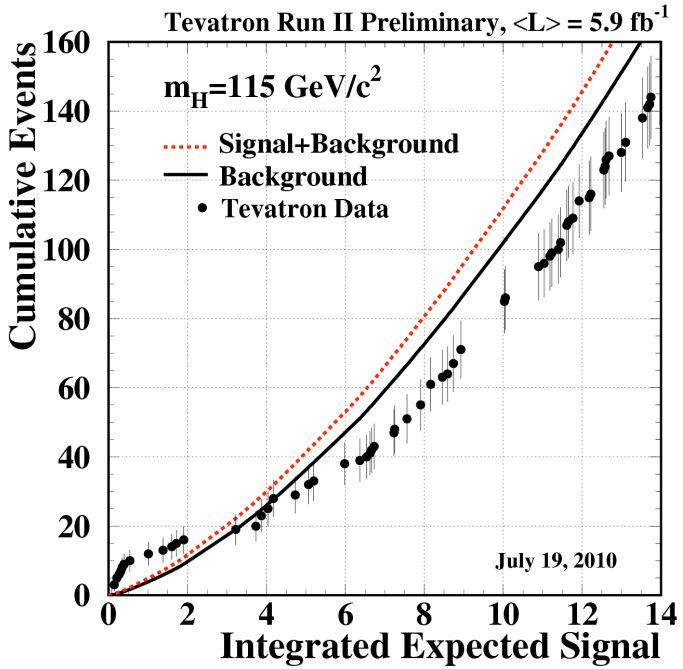

CDF + D0 Run II Preliminary
<L> = 5.9 fb-1
Data-Background
Signal
±1 s.d. on Background
Events/0.267
mH=100 GeV/c2
July 19, 2010
log10(s/b)

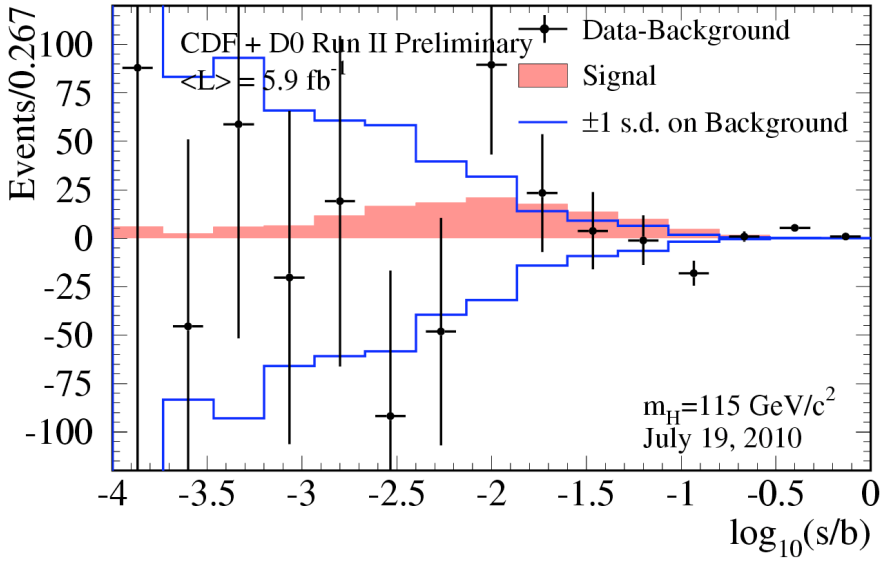
What These Look Like for a 5.0σ Observation



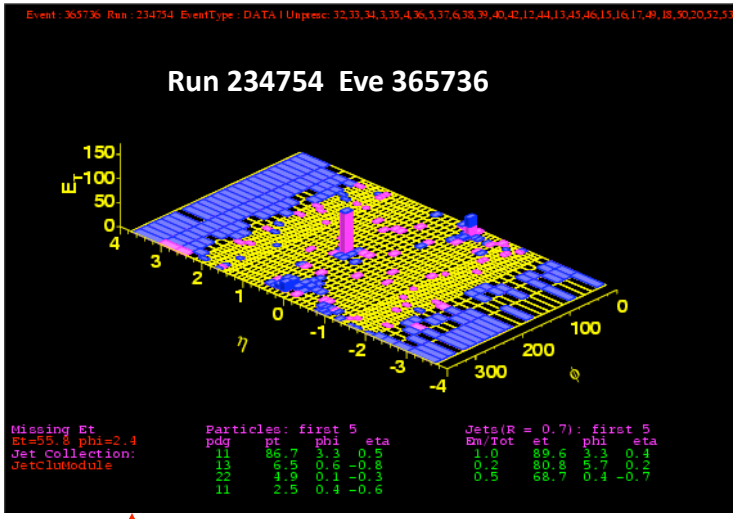
CDF Single Top, 3.2 fb^{-1}

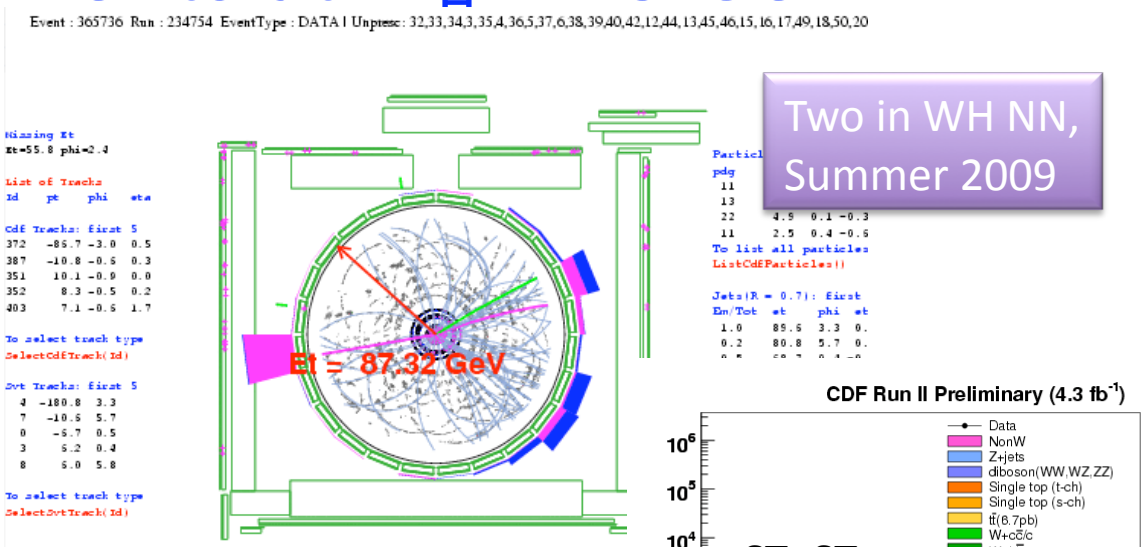

Tevatron Run II Preliminary, <L> = 5.9 fb-1
Events
mH=115 GeV/c2
Tevatron Data
Background
Signal
July 19, 2010
log10(s/b)


Tevatron Run II Preliminary, <L> = 5.9 fb-1
Cumulative Events
mH=115 GeV/c2
Signal+Background
Background
Tevatron Data
July 19, 2010
Integrated Expected Signal

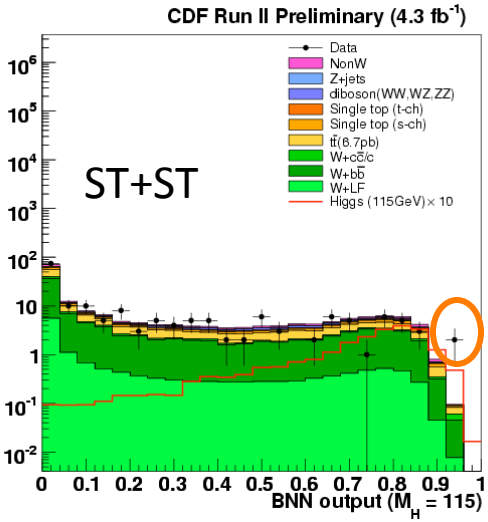

Events/0.267
CDF + D0 Run II Preliminary
<L> = 5.9 fb-1
Data-Background
Signal
±1 s.d. on Background
mH=115 GeV/c2
July 19, 2010
log10(s/b)

The Top Five Events at m_H=115 GeV



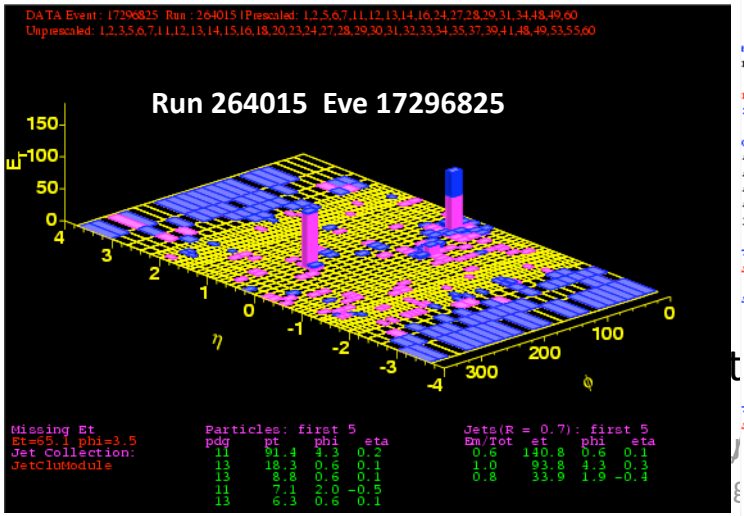


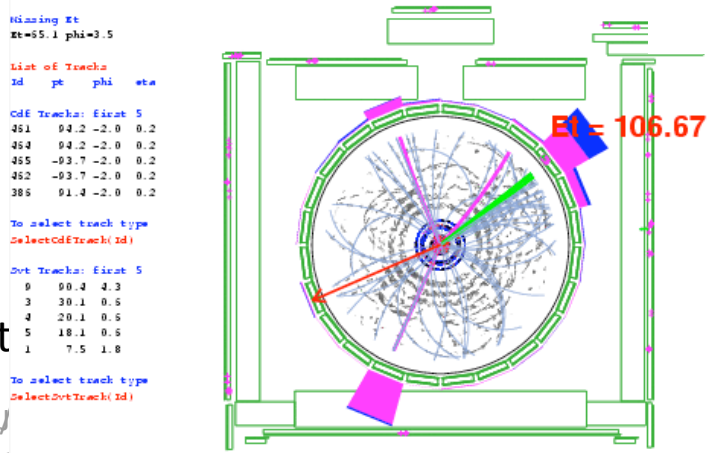
Two in WH NN, Summer 2009



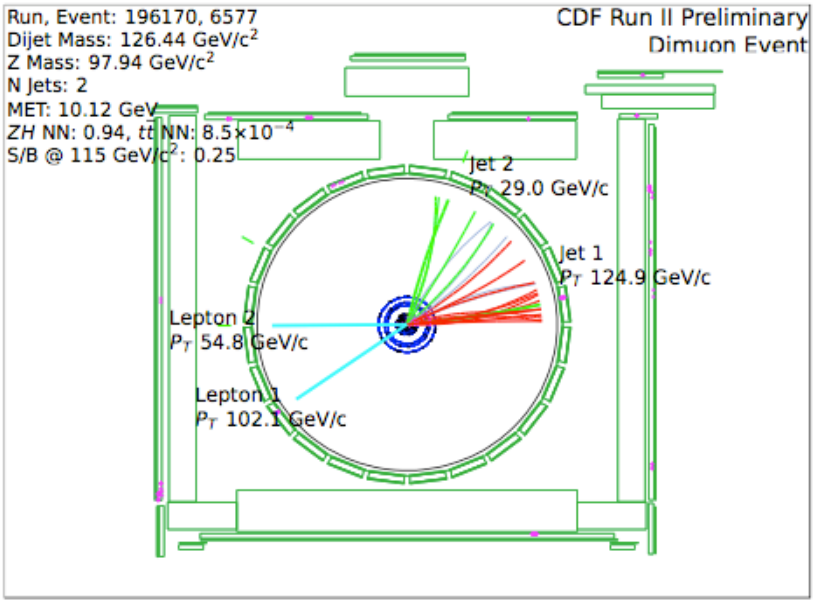
This was found as the most Higgs like event at 1.9/fb analysis (event display was blessed).

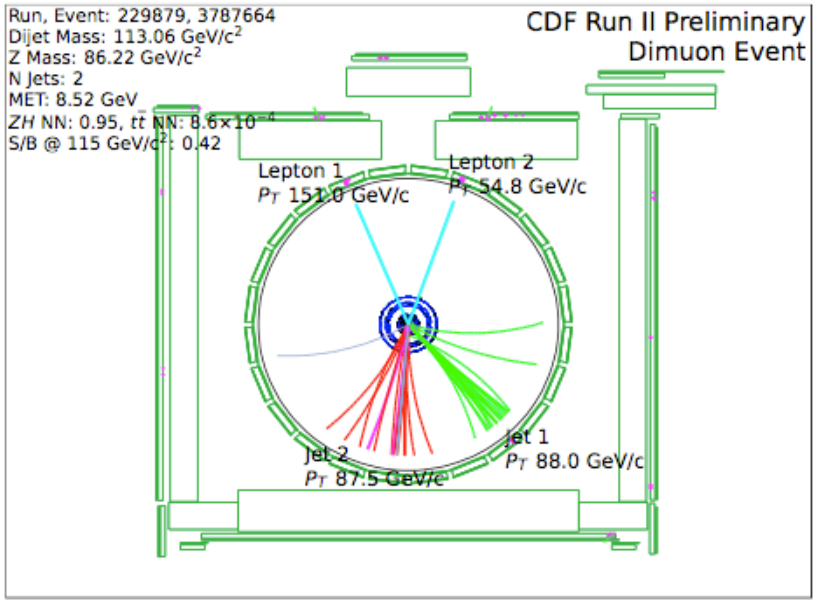
Our BNN re-find this event!!

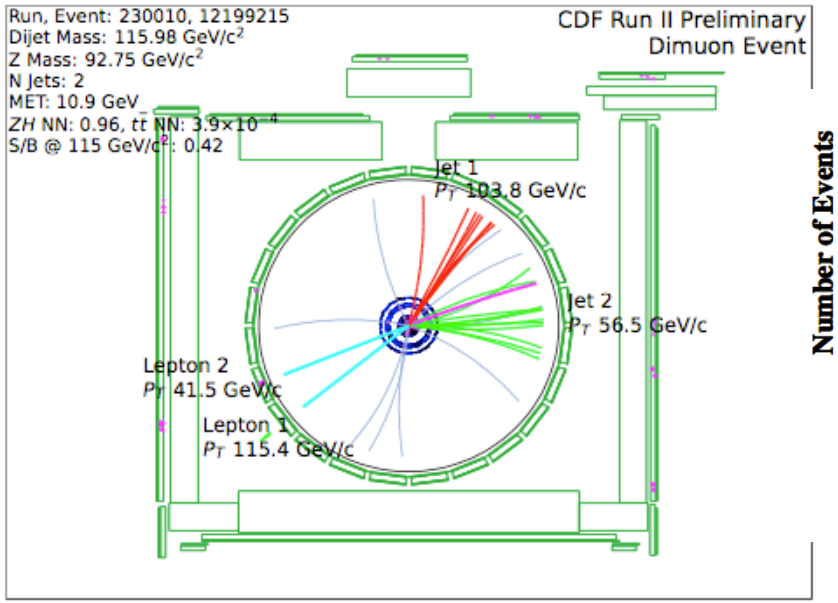


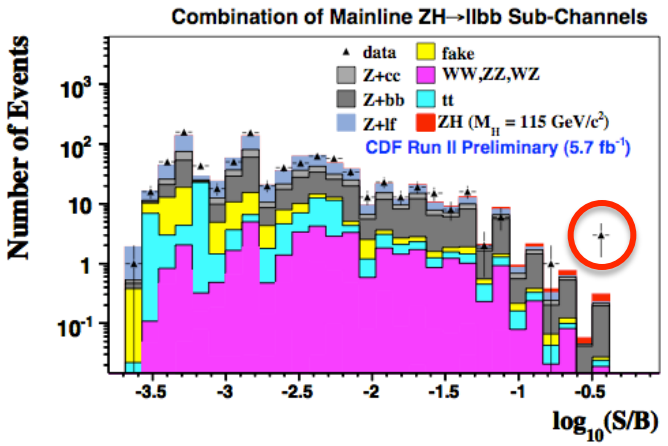


Three High s/b Candidates in the llbb Channels

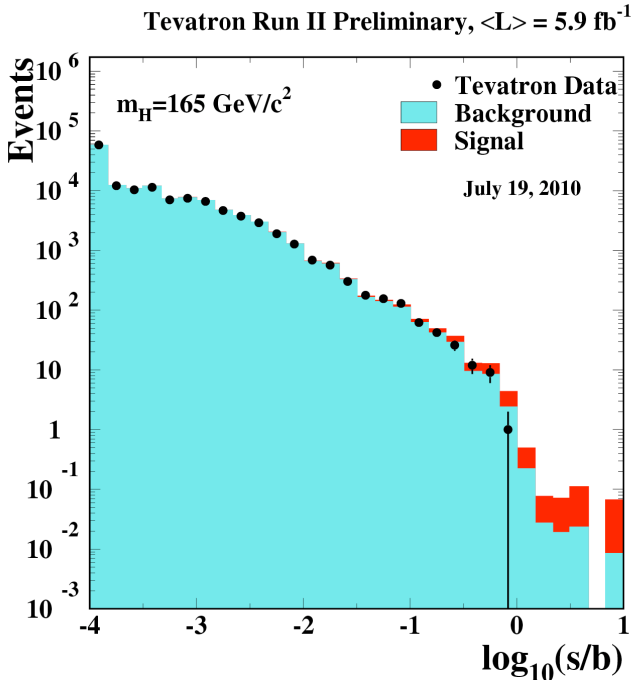


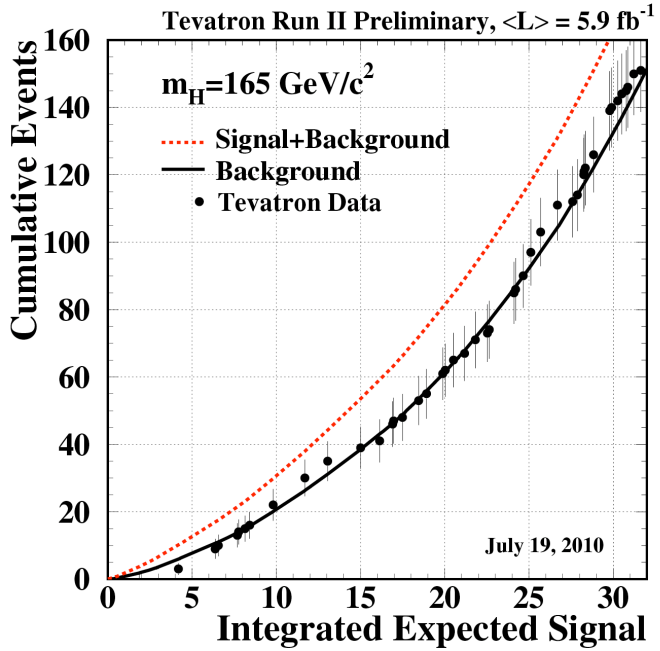


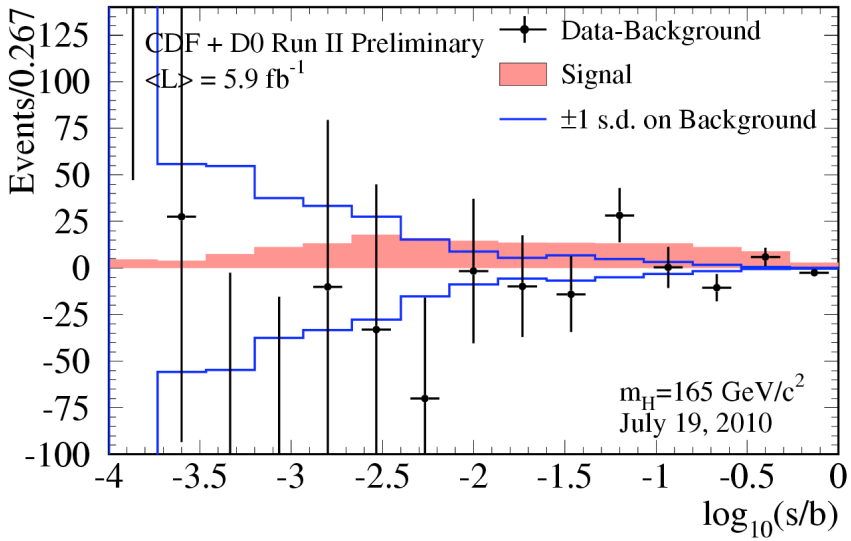




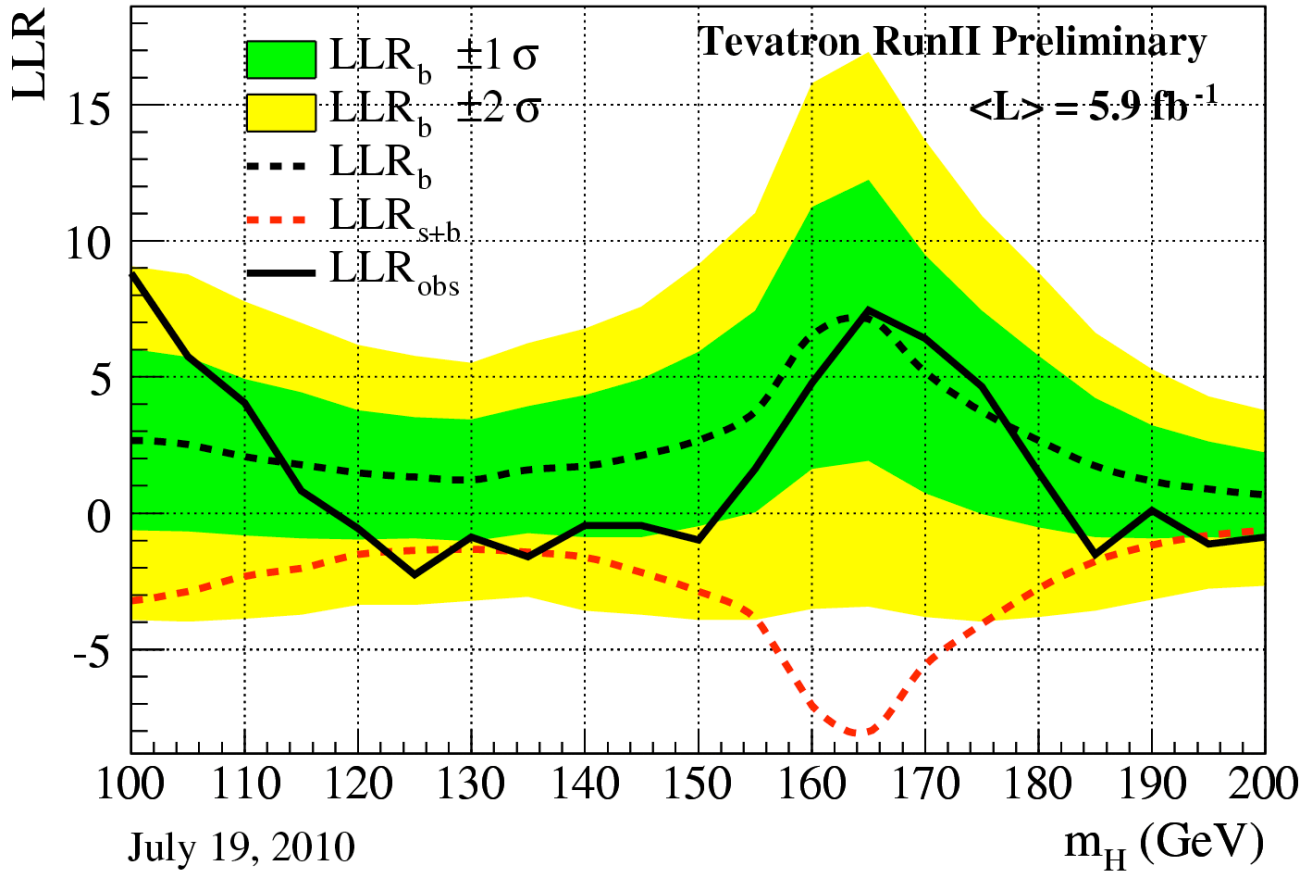
s/b of each of these events is 0.3 to 0.4


Tevatron Run II Preliminary, <L> = 5.9 fb-1
mH=165 GeV/c2
Tevatron Data
Background
Signal
July 19, 2010
Events
log10(s/b)

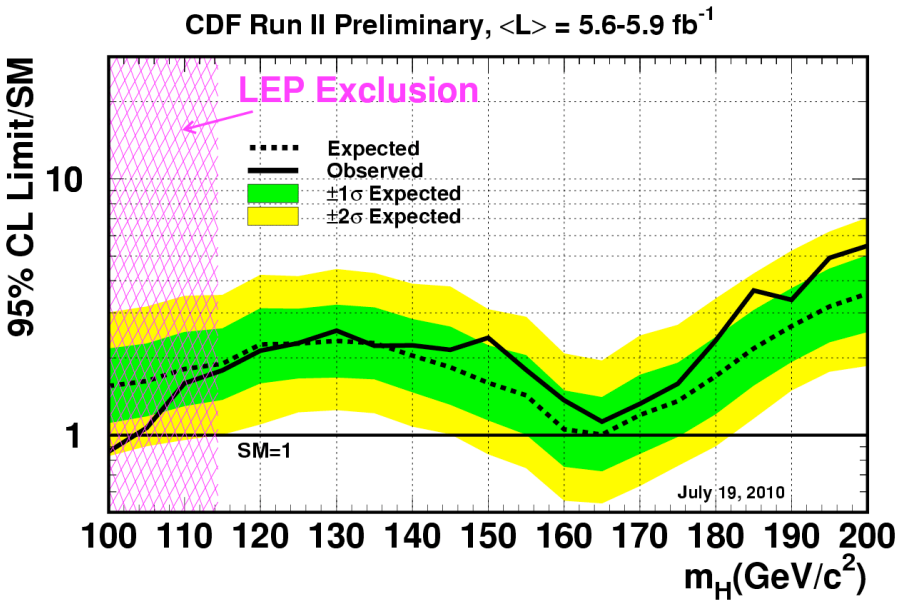

Tevatron Run II Preliminary, <L> = 5.9 fb-1
mH=165 GeV/c2
Signal+Background
Background
Tevatron Data
July 19, 2010
Cumulative Events
Integrated Expected Signal

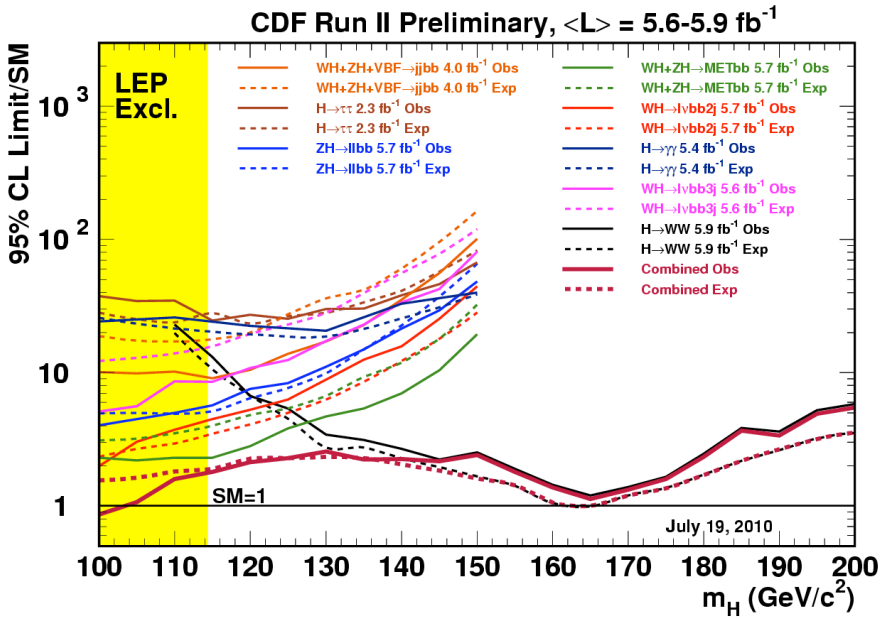

CDF + D0 Run II Preliminary
<L> = 5.9 fb-1
Data-Background
Signal
±1 s.d. on Background
mH=165 GeV/c2
July 19, 2010
Events/0.267
log10(s/b)

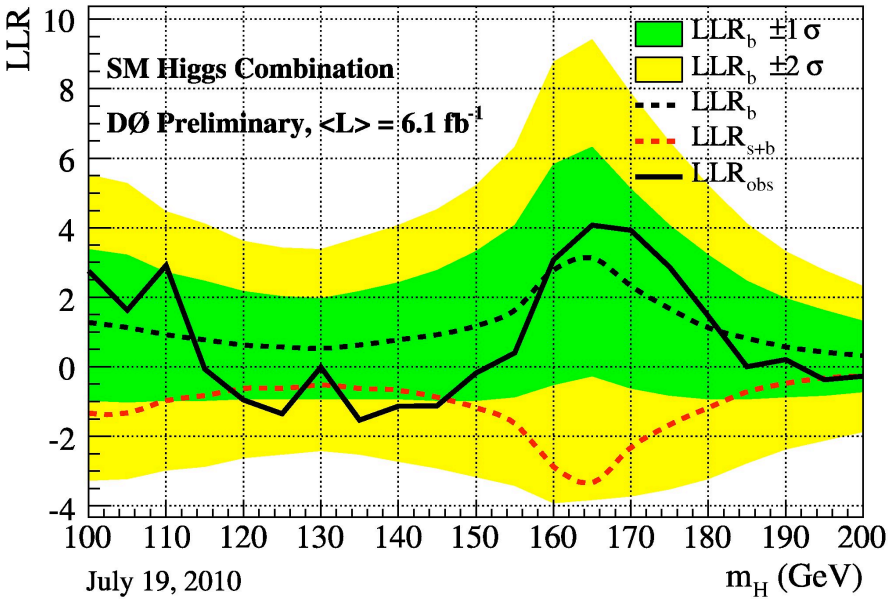
Looking for a Signal

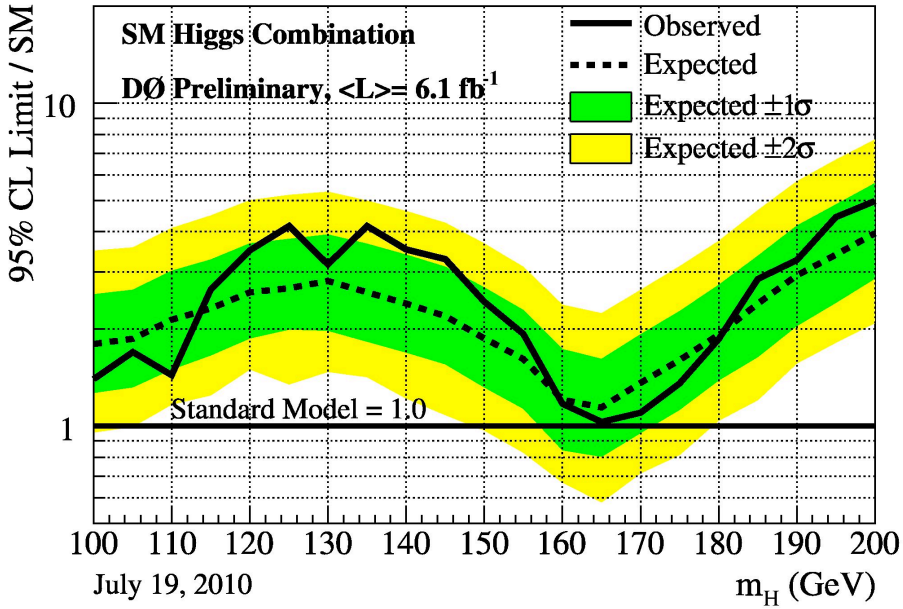


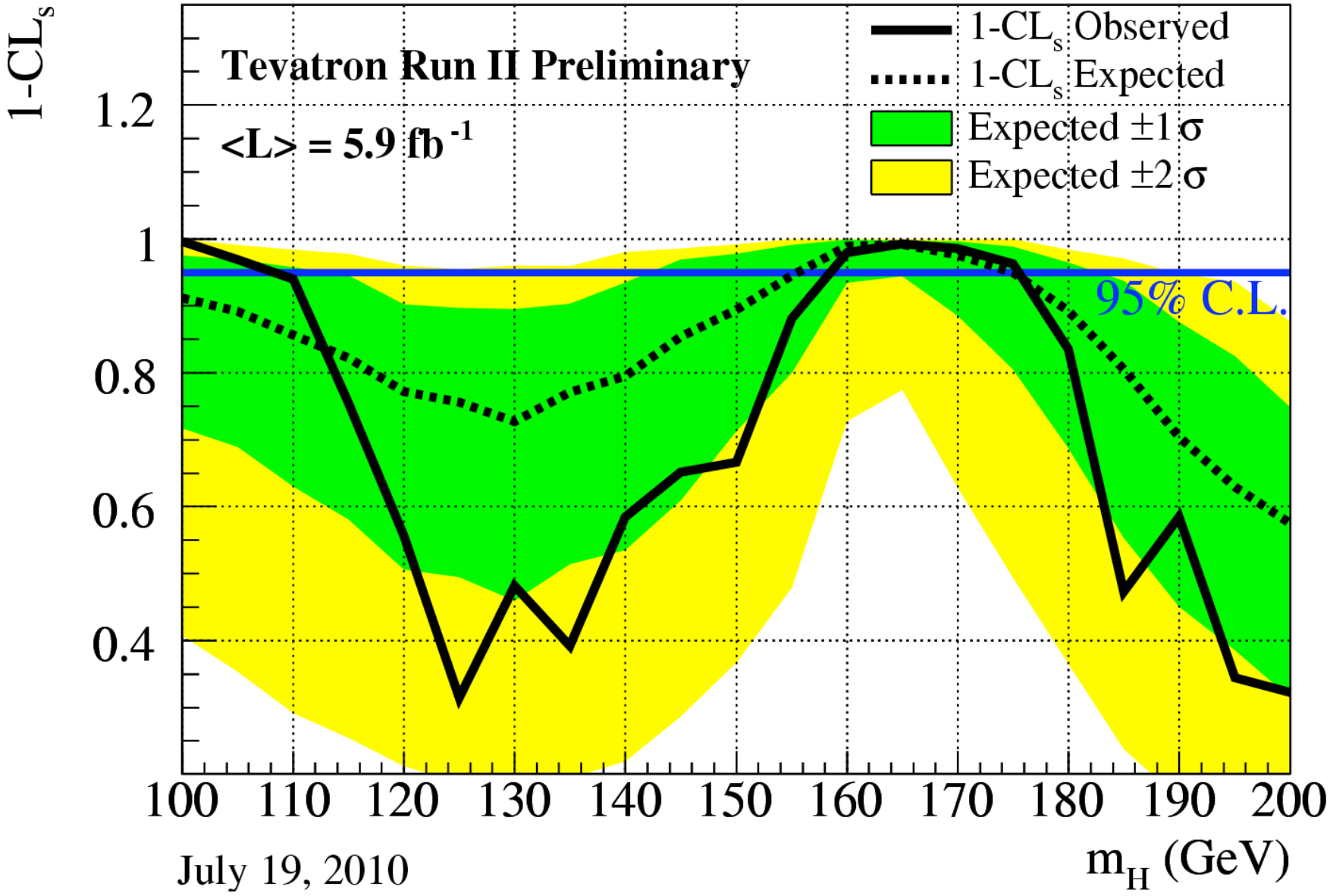
$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$


CDF Run II Preliminary, <L> = 5.6-5.9 fb-1
95% CL Limit/SM
LEP Exclusion
Expected
Observed
±1σ Expected
±2σ Expected
SM=1
July 19, 2010
mH(GeV/c2)


CDF Run II Preliminary, <L> = 5.6-5.9 fb-1
95% CL Limit/SM
LEP Excl.
WH+ZH+VBF→jjbb 4.0 fb-1 Obs
WH+ZH+VBF→jjbb 4.0 fb-1 Exp
H→ττ 2.3 fb-1 Obs
H→ττ 2.3 fb-1 Exp
ZH→llbb 5.7 fb-1 Obs
ZH→llbb 5.7 fb-1 Exp
WH+ZH→METbb 5.7 fb-1 Obs
WH+ZH→METbb 5.7 fb-1 Exp
WH→lνbb2j 5.7 fb-1 Obs
WH→lνbb2j 5.7 fb-1 Exp
H→γγ 5.4 fb-1 Obs
H→γγ 5.4 fb-1 Exp
WH→lνbb3j 5.6 fb-1 Obs
WH→lνbb3j 5.6 fb-1 Exp
H→WW 5.9 fb-1 Obs
H→WW 5.9 fb-1 Exp
Combined Obs
Combined Exp
SM=1
July 19, 2010
mH (GeV/c2)

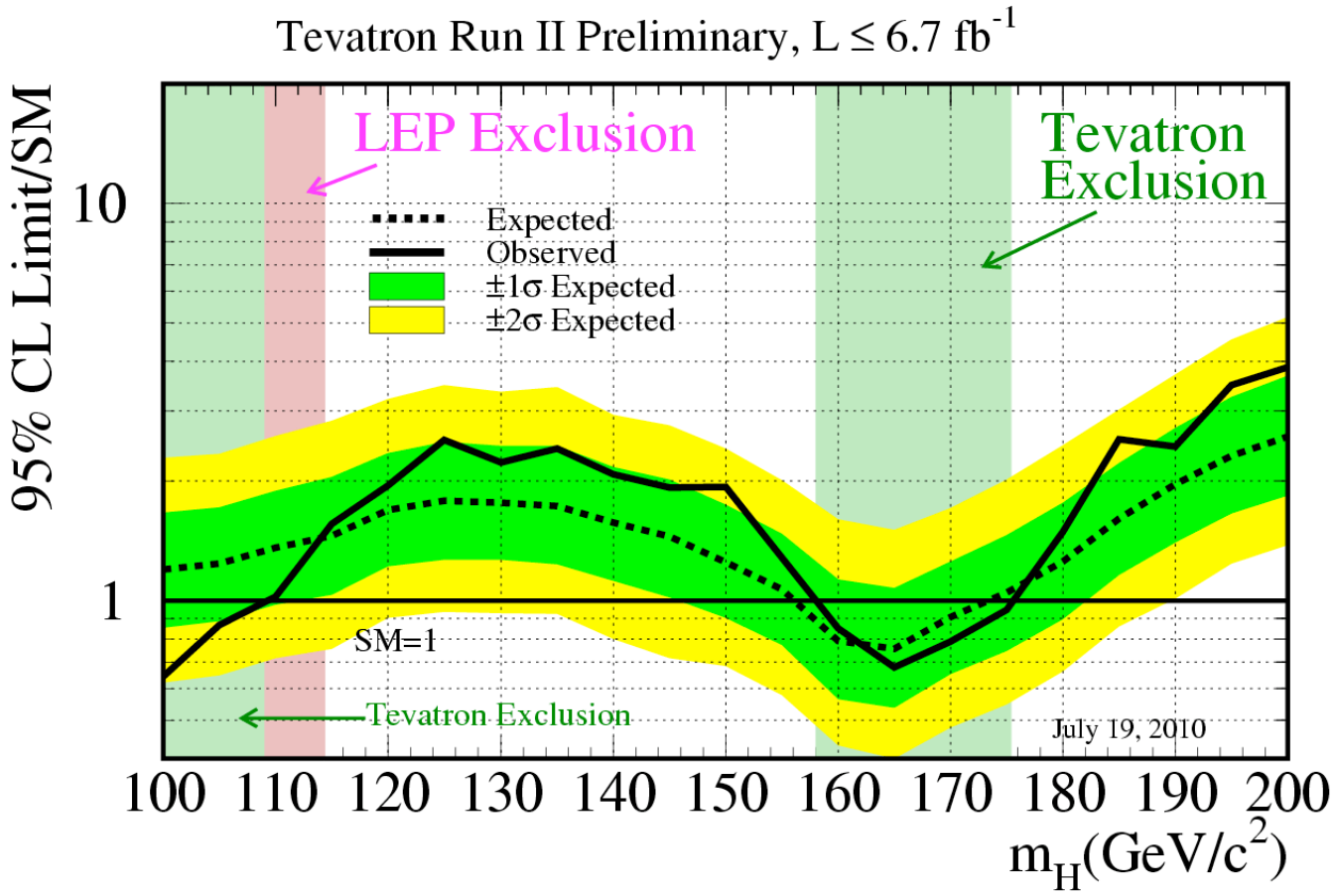

LLR
SM Higgs Combination
DØ Preliminary, <L> = 6.1 fb-1
LLRb ±1 σ
LLRb ±2 σ
LLRb
LLRs+b
LLRobs
July 19, 2010
mH (GeV)


95% CL Limit / SM
SM Higgs Combination
DØ Preliminary, <L>= 6.1 fb-1
Observed
Expected
Expected ±1σ
Expected ±2σ
Standard Model = 1.0
July 19, 2010
mH (GeV)


Tevatron Run II Preliminary
<L> = 5.9 fb -1
1-CLs Observed
1-CLs Expected
Expected ±1 σ
Expected ±2 σ
95% C.L.
1-CLs
1.2
1
0.8
0.6
0.4
100 110 120 130 140 150 160 170 180 190 200
mH (GeV)
July 19, 2010

Tevatron Observed and Expected Limits

Bayesian



Excluded regions:

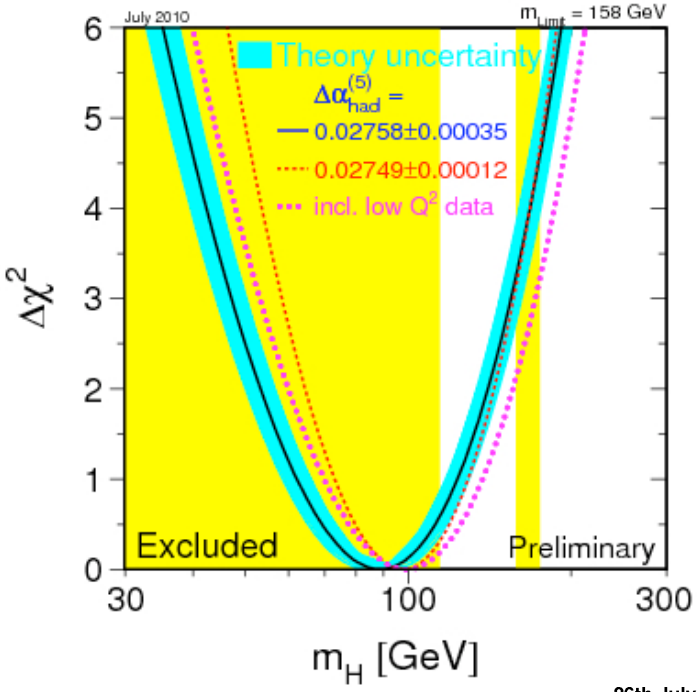
$158 < m_H < 175$ GeV

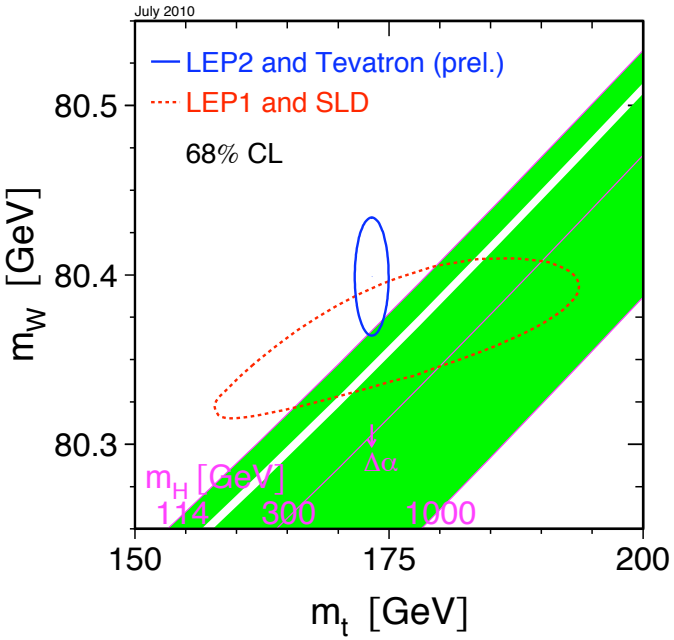
$100 < m_H < 109$ GeV

Expected Exclusion
(if no signal is present):

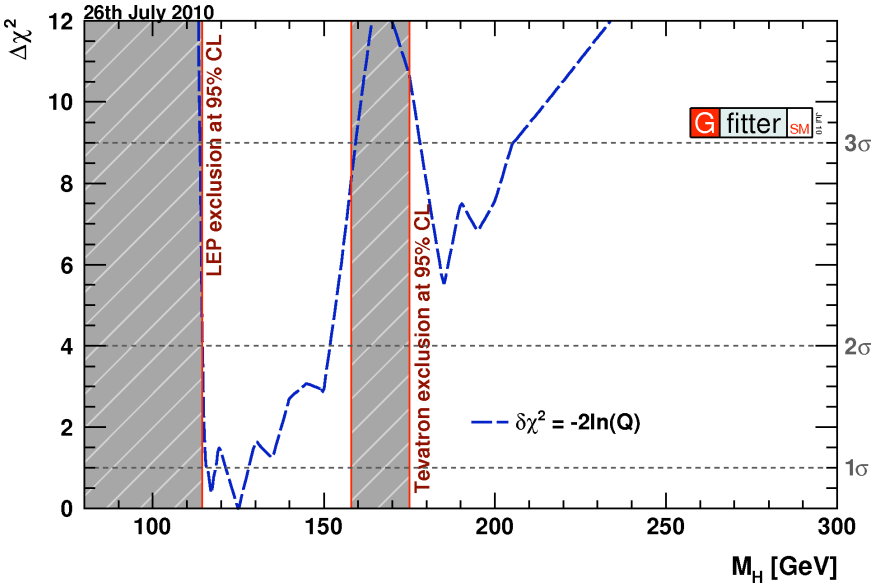
$156 < m_H < 173$ GeV

We have a large impact on the Global Fits

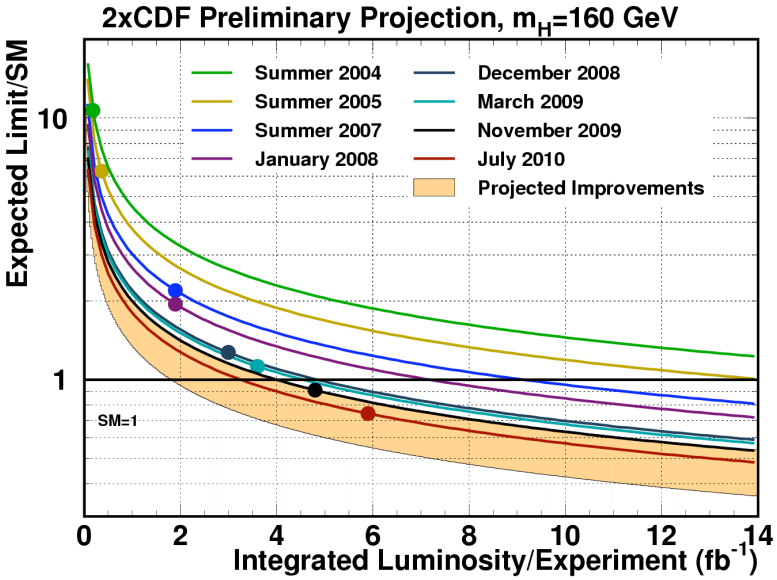


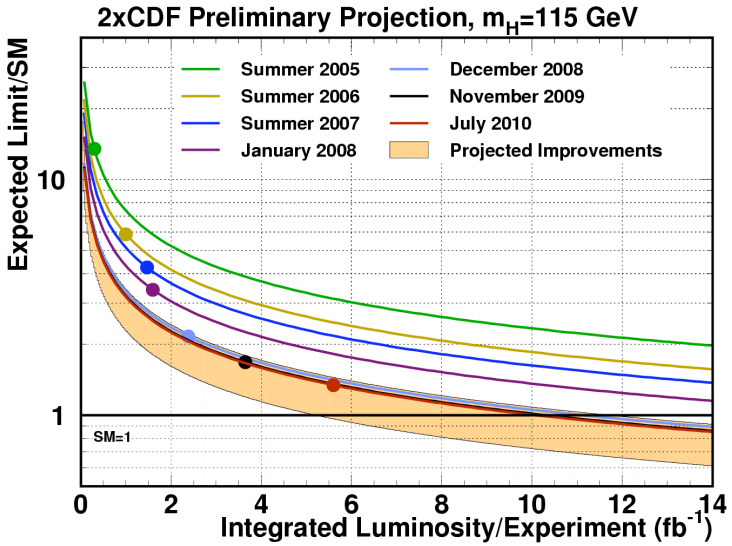


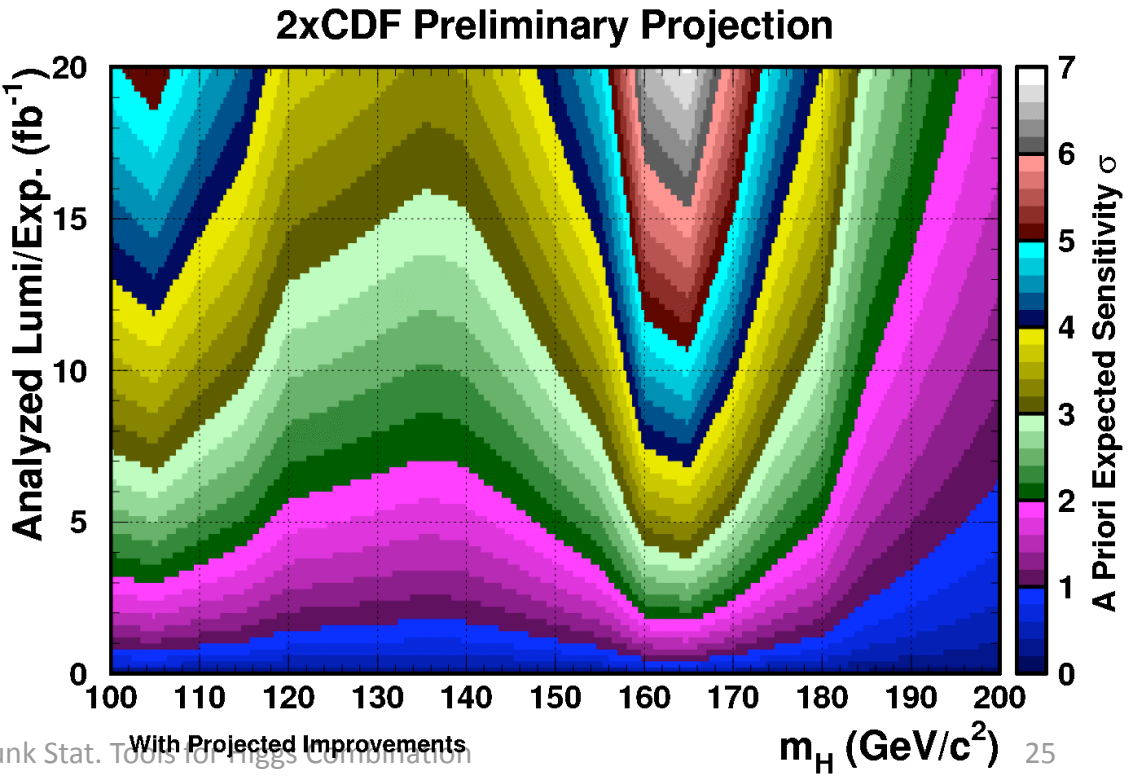
Gfitter Collaboration:



Projections for Future Performance

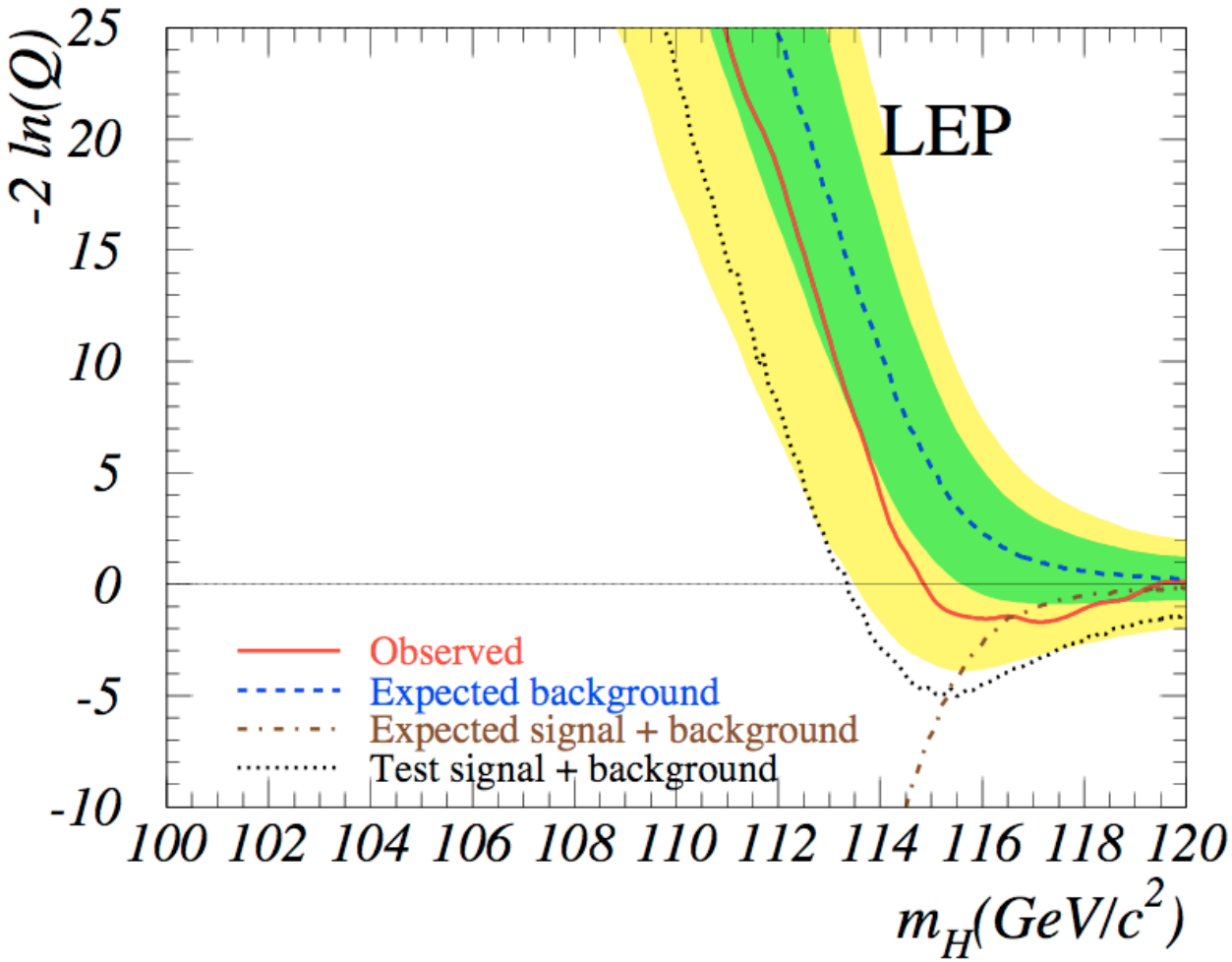






Look-Aside Histograms

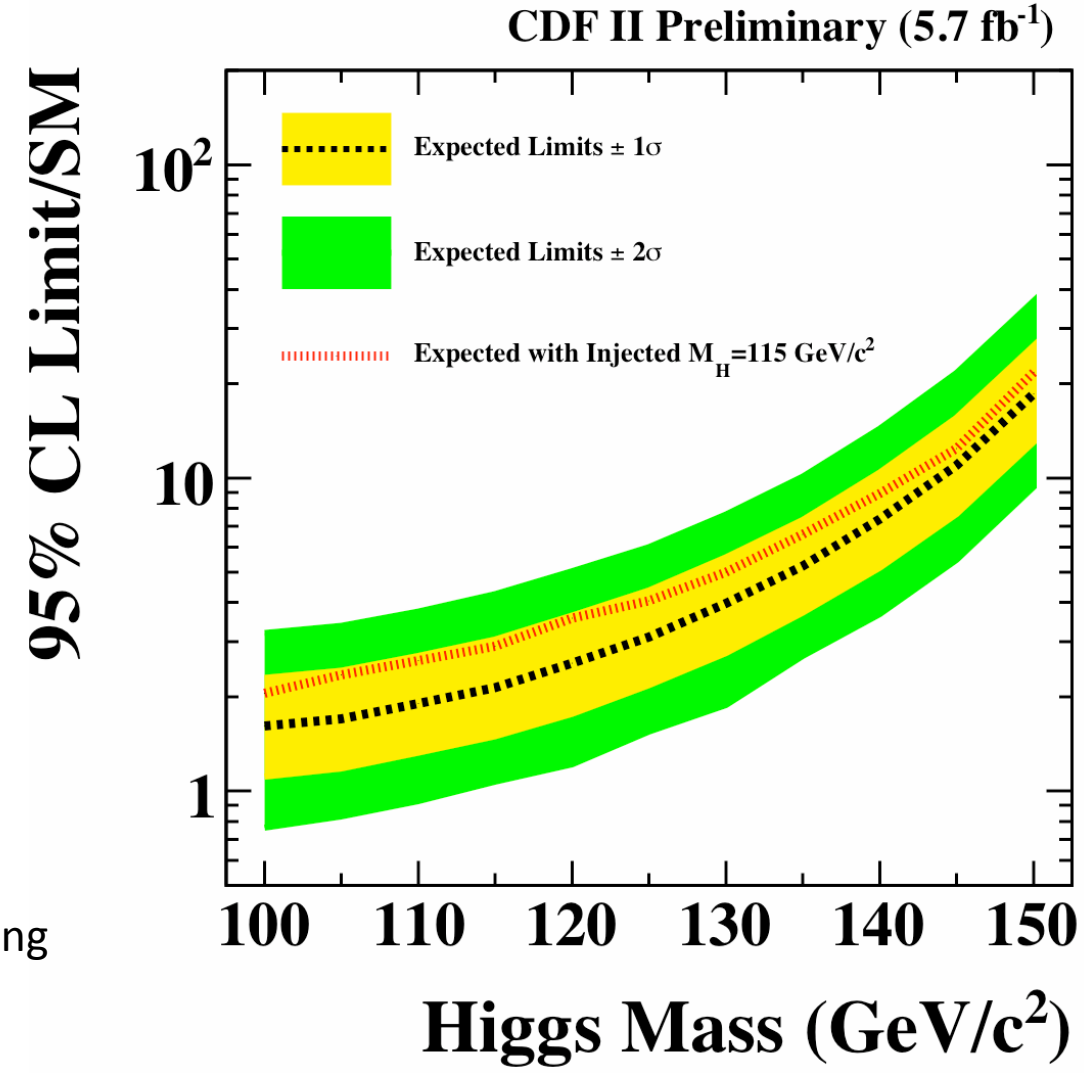
What does a signal with $m_H=m_1$ look like when seeking $m_H=m_2$?
So far, not done at Tevatron. Not needed to study the trials factor, but needed to make this plot:



26

Studies of Injecting a Signal at m_H=115 GeV

- lvbb, METbb, and llbb channels included
- Inject SM*1.0 signal at m_H=115 GeV on top of SM backgrounds, and generate pseudoexperiments with that.
- Analyze 115 signal+background pseudoexperiments at other test masses – 100 GeV to 150 GeV
- Find the median expected limit assuming signal is there (compute it just as you would without the signal) and compare with the distribution of limits assuming the signal is completely absent.

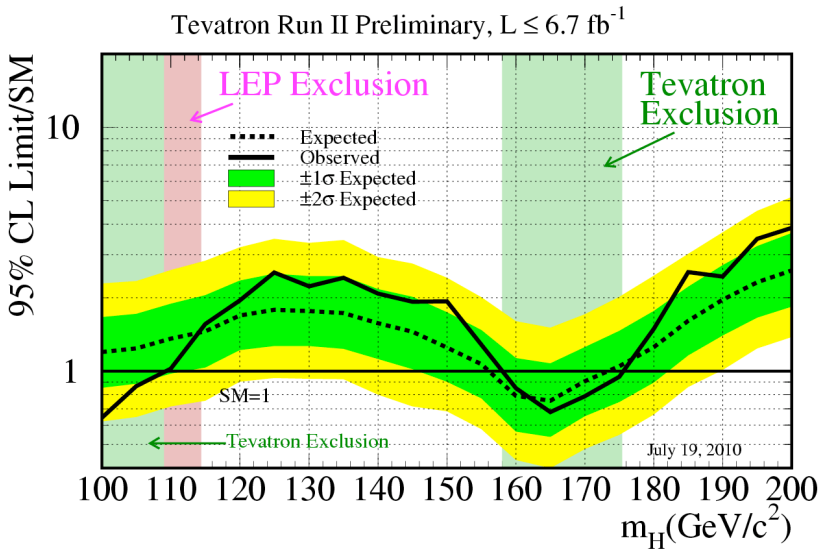


Summary

The Tevatron
is doing very well!

We will run at least
through 2011

We are asking to
run another 3 years.

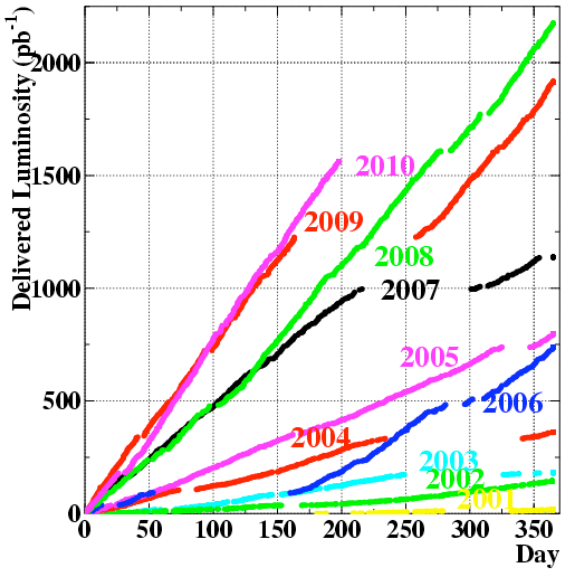


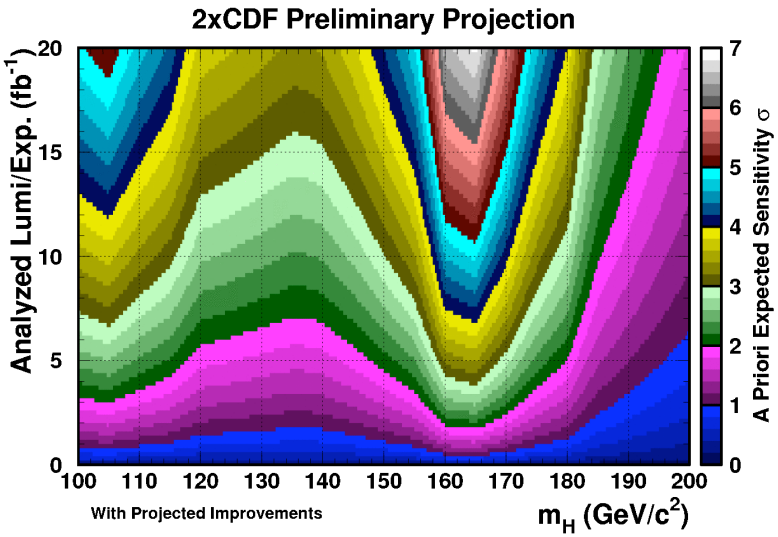
Excluded regions:

$158 < m_H < 175$ GeV
$100 < m_H < 109$ GeV

Expected Exclusion
(if no signal is present):

$156 < m_H < 173$ GeV





Backup Material

Commonly Used Tools for Setting Limits and Discovering New Processes in use at the Tevatron

- Bayesian limits -- common at CDF
  - genlimit code by Joel Heinrich, added to mclimit code by Tom Junk
  - Implements posterior integrated over systematic uncertainties with a flat prior in cross section in 1D
  - Method described in PDG statistics reveiw
  - Extra feature -- “correlated prior”

- CL_s limits -- common at D0, but used at CDF as well.
  - Collie code by Wade Fisher in use at D0
  - Method described in PDG statistics review
  - mclimit was originally designed to do CL_s and still does.
  - TLimit in ROOT is out of date -- no fits for nuisance parameters, no shape errors or bin-by-bin errors

Measurement and Discovery are Very Different

Buzzwords:

- Measurement = “Point Estimation”
- Discovery = “Hypothesis Testing”

You can have a discovery and a poor measurement!
Example: Expected $b=2\times10^{-7}$ events, expected signal=1 event, observe 1 event, no systematics.

p-value ~2×10^{-7} is a discovery! (hard to explain that event with just the background model). But have ±100% uncertainty on the measured cross section!

In a one-bin search, all test statistics are equivalent. But add in a second bin, and the measured cross section becomes a poorer test statistic than the ratio of profile likelihoods.

In all practicality, discriminant distributions have a wide spectrum of s/b, even in the same histogram. But some good bins with b<1 event

Sociological Issues

- Discovery is conventionally 5σ. In a Gaussian asymptotic case, that would correspond to a ±20% measurement.

- Less precise measurements are called "measurements" all the time

- We are used to measuring undiscovered particles and processes. In the case of a background-dominated search, it can take years to climb up the sensitivity curve and get an observation, while evidence, measurements, etc. proceed.

- Referees can be confused.

The Trials Factor

- Also called the "Look Elsewhere Effect"
- Bump-hunters are familiar with it.

What is the probability of an upward fluctuation as big as the one I saw *anywhere* in my histogram?

-- Lots of bins → Lots of chances at a false discovery
-- Approximation: Multiply smallest p-value by the number of "independent" models sought (not histogram bins!).
Bump hunters: roughly (histogram width)/(mass resolution)
Criticisms:
Adjusted p-value can now exceed unity!
What if histogram bins are empty?
What if we seek things that have been ruled out already?

It's not bins, but the number of independent hypotheses being tested that matters!

Low mass resolution at the Tevatron – not very many independent excesses possible.

The Trials Factor

More seriously, what to do if the p-value comes from a big combination of many channels each optimized at each m_H sought?

- Channels have different resolutions (or is resolution even the right word for a multivariate discriminant?
- Channels vary their weight in the combination as cross sections and branching ratios change with m_H

Proper treatment -- want a p-value of p-values!
(use the p-value as a test statistic)
Run pseudoexperiments and analyze each one at each m_H studied. Look for the distribution of smallest p-values.

Difficult but possible.