Higgs Boson Limits and Discovery: Some Lessons Learned from the Tevatron (and LEP)

Tom Junk
Fermilab

TeV4LHC Forum, Cosener's House
May 7-8, 2009

- Tools in use at the Tevatron for Exclusion and Discovery
- Examples
- Practical suggestions for combinations
- Do's and Don'ts -- things to be careful about

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  - TLimit in ROOT is out of date -- no fits for nuisance parameters, no shape errors or bin-by-bin errors

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

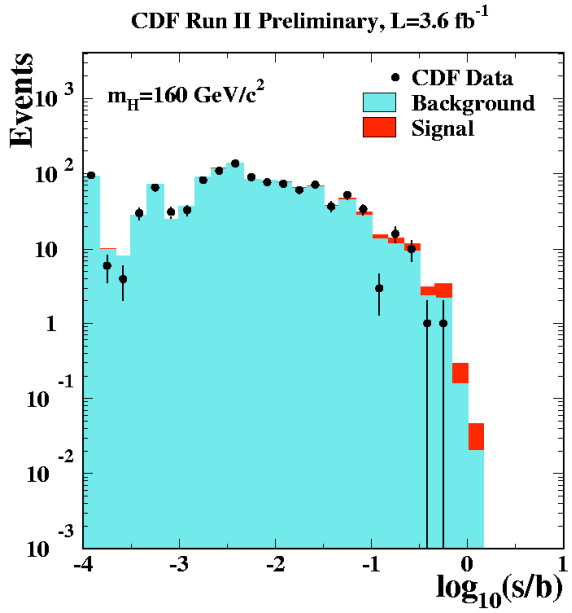
Measure *r*: $$0.68 = \int_{r_{low}}^{r_{high}} L'(data \mid r)\pi(r)dr$$

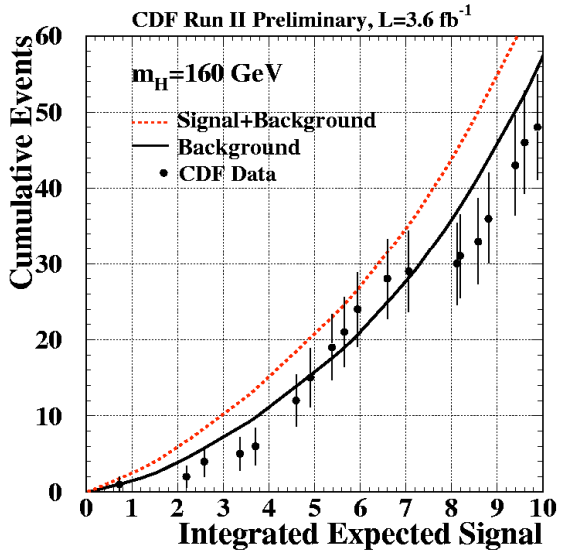
Typically $\pi(r)$ is constant
Other options possible.
Sensitivity to priors a concern.

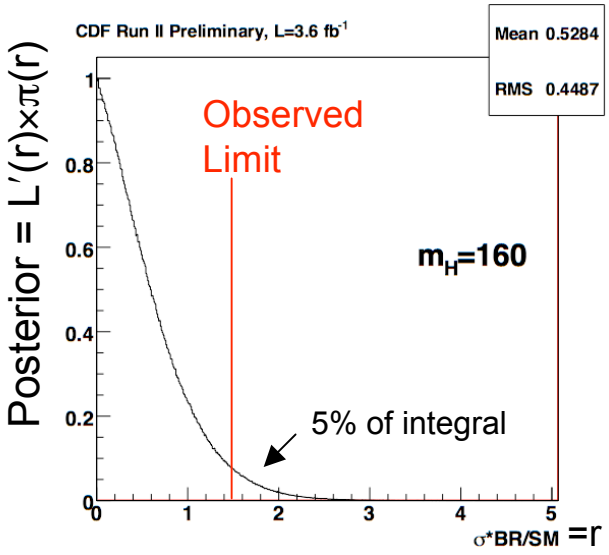
$$r = r_{\max}{}^{+(r_{high}-r_{\max})}_{-(r_{\max}-r_{low})}$$

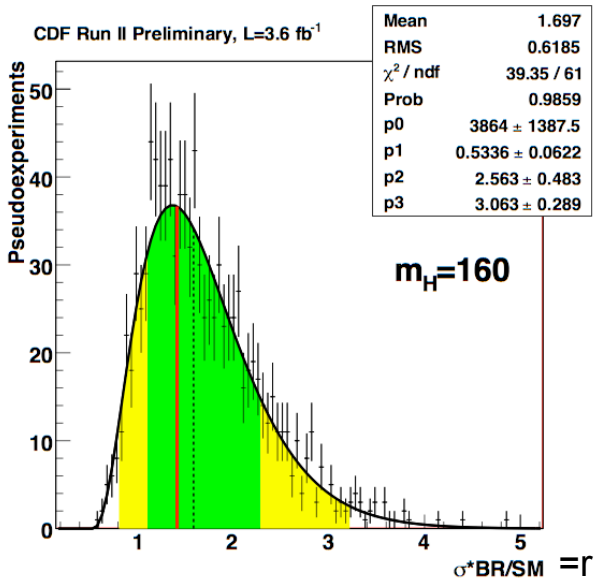
Usually: shortest interval containing 68% of the posterior (other choices possible)

Bayesian Example: CDF Higgs Search at m_H=160 GeV



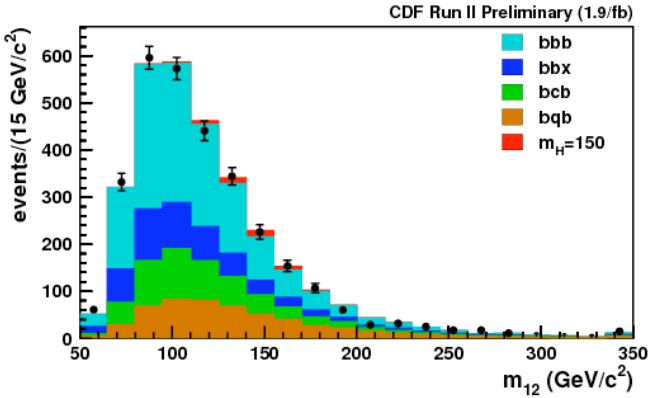


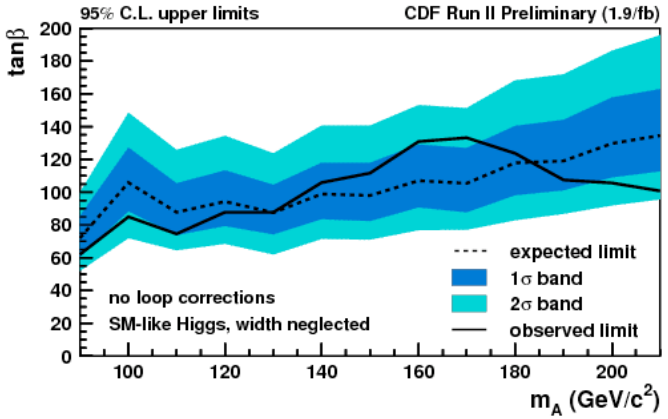


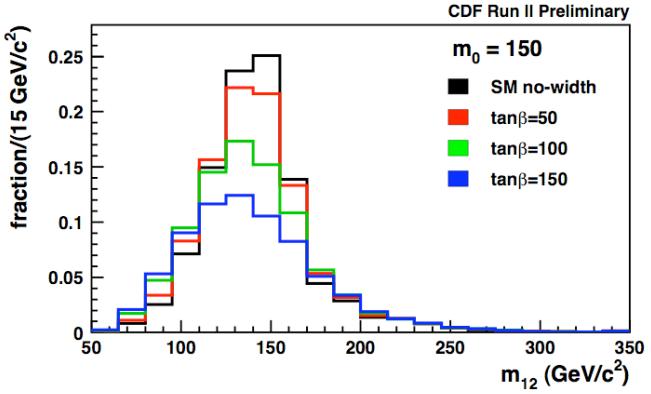


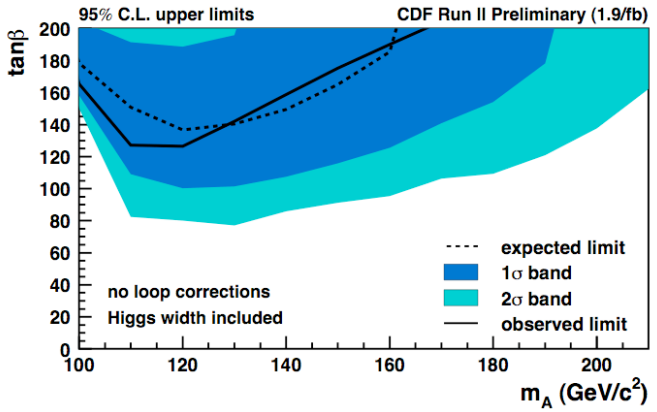
An Example Where Usual Bayesian Software Doesn't Work

- Typical Bayesian code assumes fixed background, signal shapes (with systematics) -- scale signal with a scale factor and set the limit on the scale factor
- But what if the kinematics of the signal depend on the cross section? Example -- MSSM Higgs boson decay width scales with $\tan^2\beta$, as does the production cross section.
- Solution -- do a 2D scan and a two-hypothesis test at each m_A,$\tan\beta$ point



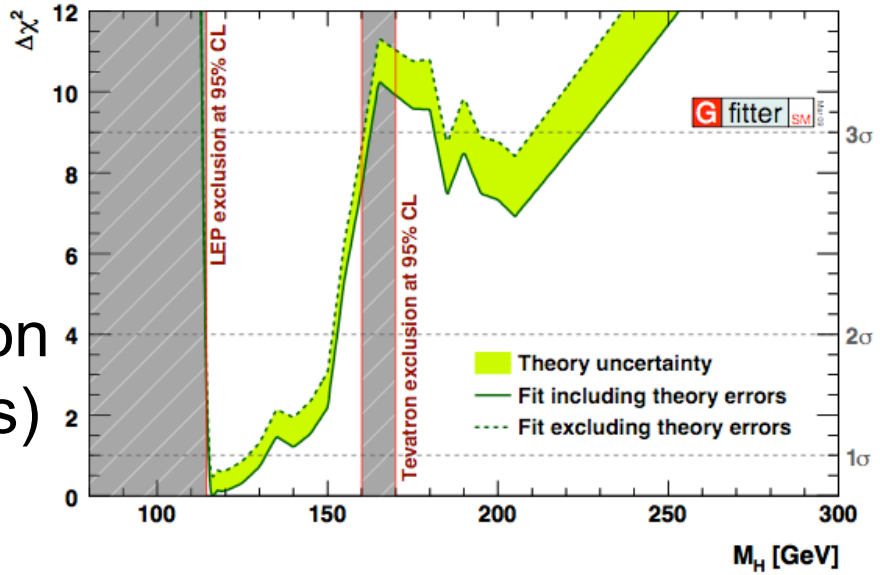




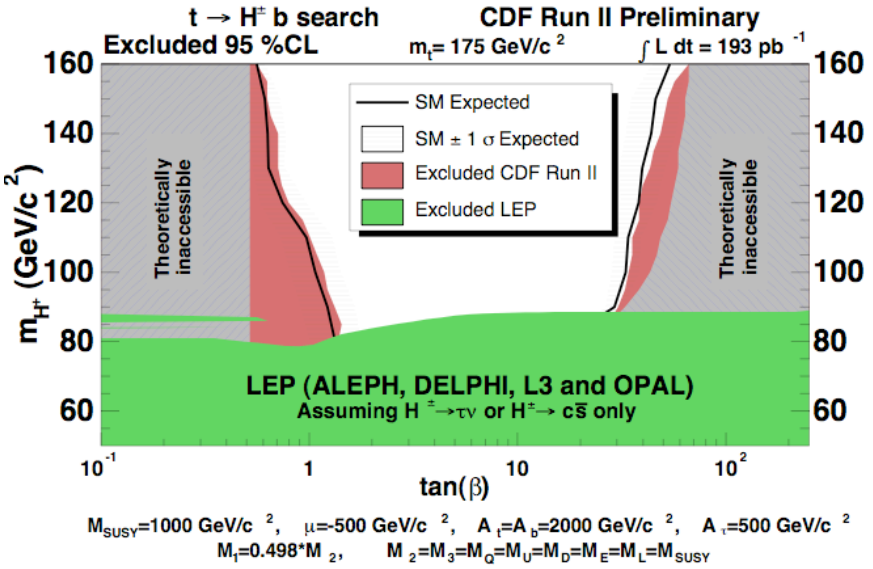


Priors in Non-Cross-Section Parameters

Example: take a flat prior in m_H; can we discover the Higgs boson by process of elimination? (assumes exactly one Higgs boson exists, and other SM assumptions)



Example: Flat prior in log(tanβ) -- even with no sensitivity, can set non-trivial limits..



Bayesian Discovery?

Bayes Factor

$$B = L'(data \mid r_{\max}) / L'(data \mid r = 0)$$

Similar definition to the profile likelihood ratio, but instead of maximizing L, it is averaged overr nuisance parameters in the numerator and denominator.

Similar criteria for evidence, discovery as profile likelihood.

Physicists would like to check the false discovery rate,
and then we're back to p-values.

But -- odd behavior of *B* compared with p-value for even a simple case

J. Heinrich, CDF 9678
http://newton.hep.upenn.edu/~heinrich/bfexample.pdf

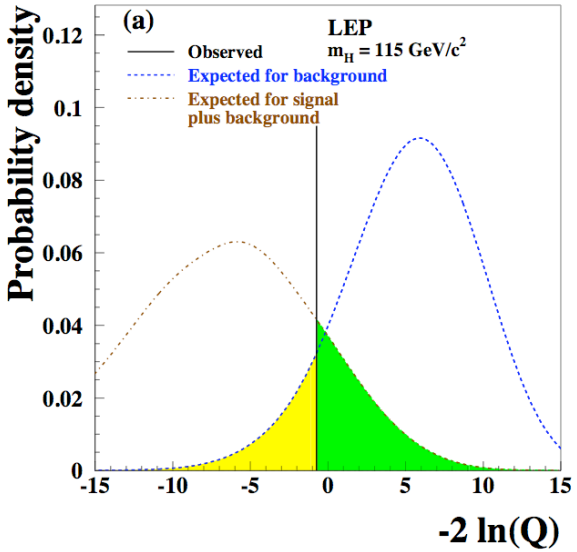
Mini-Review: CL_s Limits

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

Mini-Review: CL_s Limits

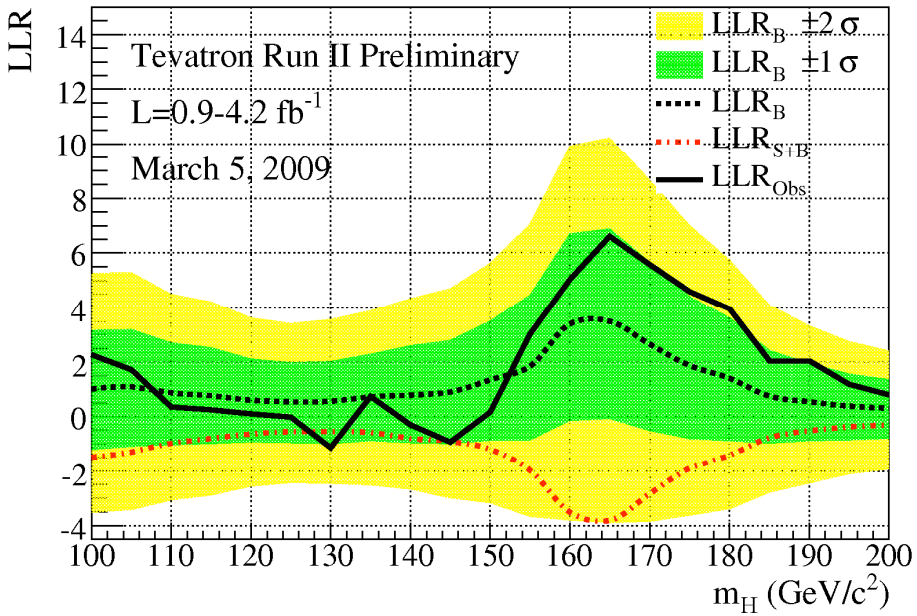


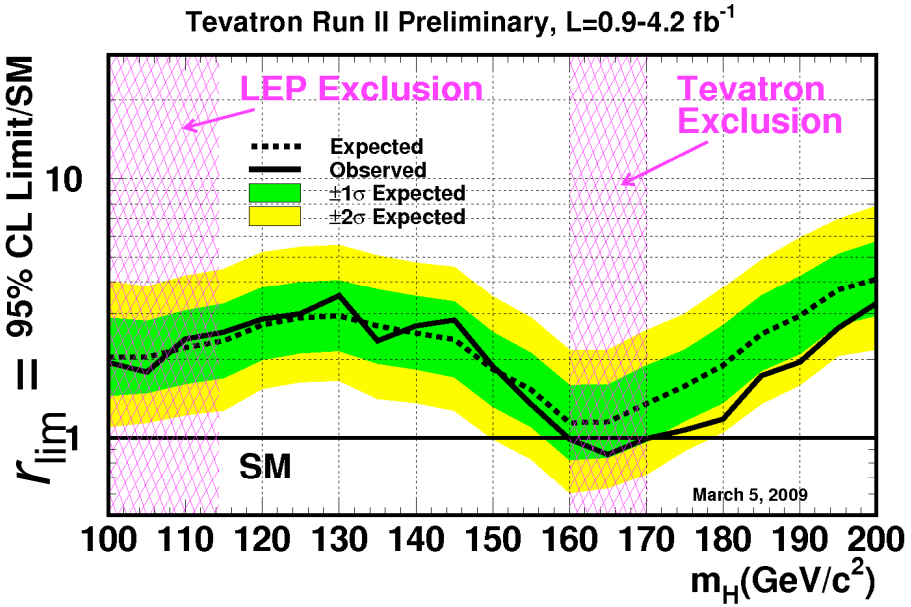
p-values:
Yellow area = $1\text{-}CL_b = 1\text{-}P(-2\ln Q > -2\ln Q_{obs} | \text{ b only})$
Green area = $CL_{s+b} = P(-2\ln Q > -2\ln Q_{obs} | \text{ s+b})$

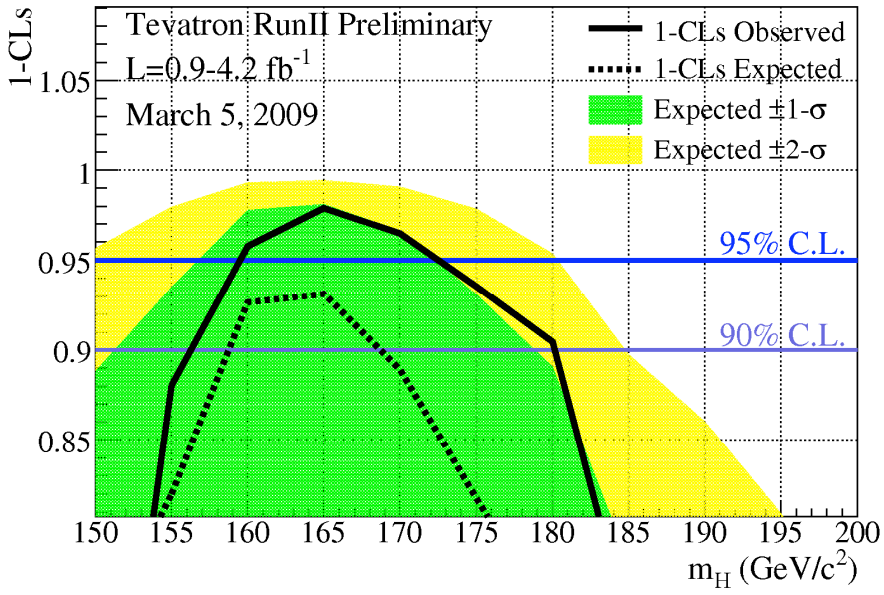
$CL_s \equiv CL_{s+b}/CL_b \geq CL_{s+b}$
Exclude if $CL_s < 0.05$
Vary r until $CL_s = 0.05$ to get r_{lim}

- 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 Higgs Combination Cross-Checked Two Ways







Very similar results --

- Comparable exclusion regions
- Same pattern of excess/deficit relative to expectation

n.b. Using CL_{s+b} limits instead of CL_s or Bayesian limits would extend the bottom of the yellow band to zero in the above plot, and the observed limit would fluctuate accordingly. We'd have to explain the 5% of m_H values we randomly excluded without sufficient sensitivity.

Practical Suggestions 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:

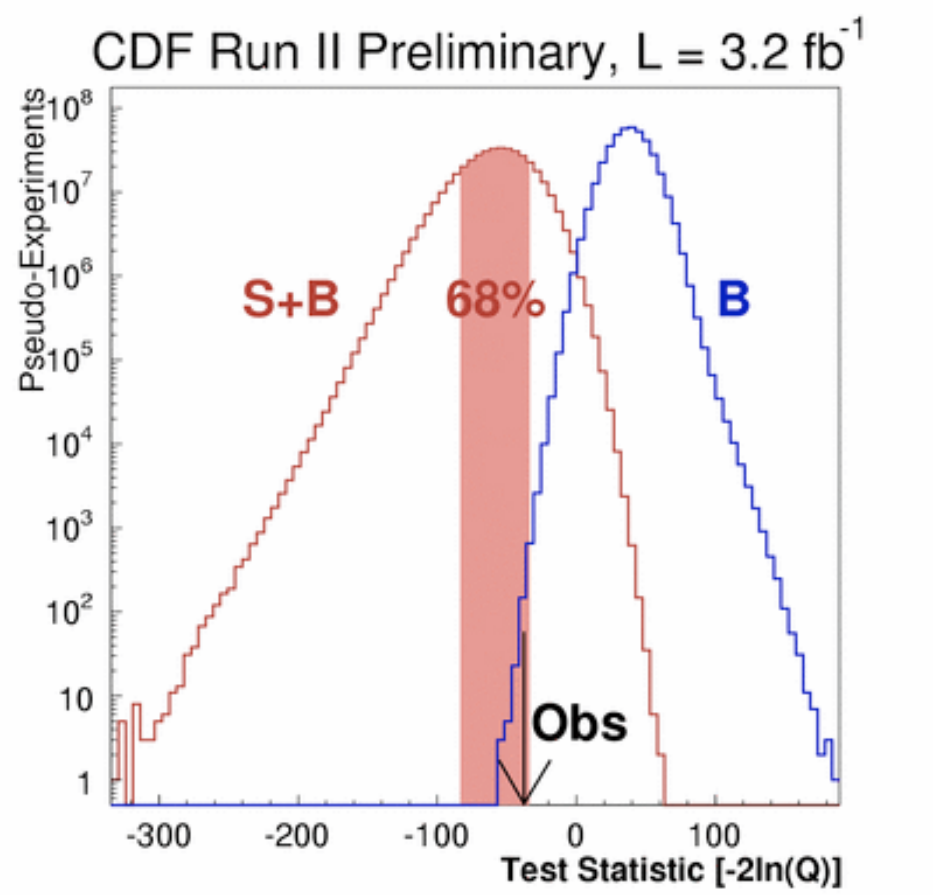
- Lots of work/CPU!
- Have to share preliminary histos (your competitors may find your mistakes!)

Practical Suggestions 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!

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.

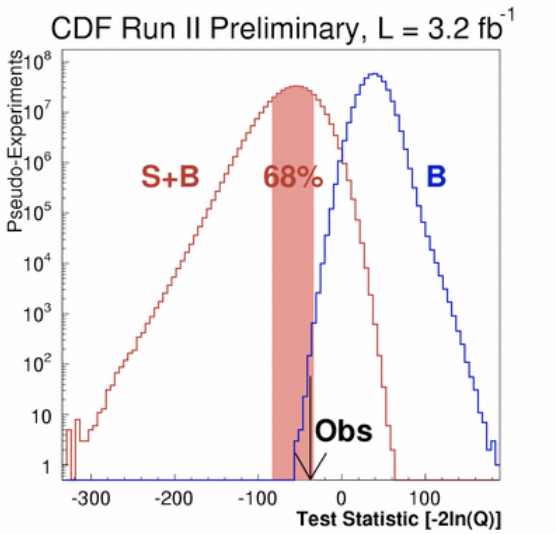
5σ: p-value of 2.77×10^{-7} or less.

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.

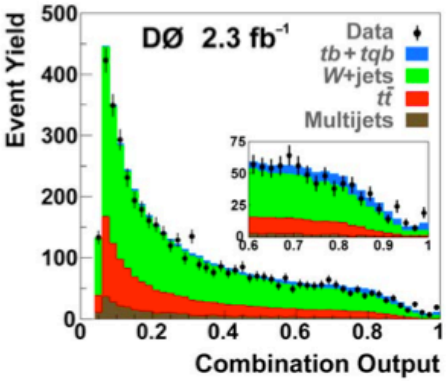
Another Avenue towards using Bayesian Techniques for Discovery

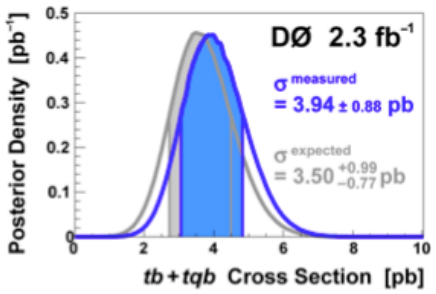
- D0 measures the single top cross section with a Bayesian technique
- The measured cross section is used as a test statistic for the p-value for significance. Pseudoexperiments fluctuate systematics.

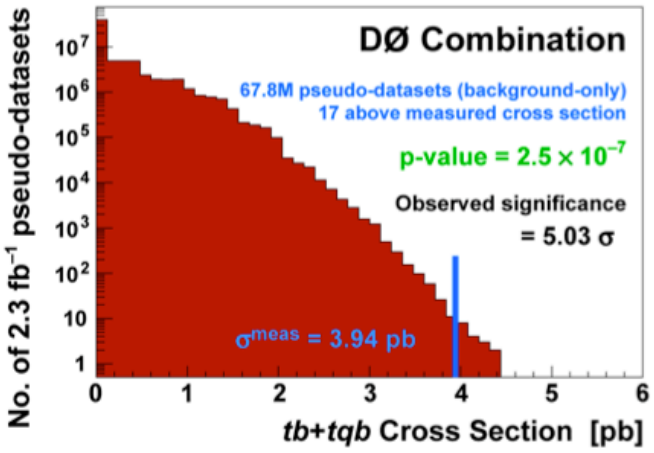
Combined Results

$$\sigma(p\bar{p} \to tb + X, tqb + X) = 3.94 \pm 0.88 \text{ pb}$$

(m_t=170GeV)







$$p - value = 2.5 \times 10^{-7}$$

Measured Significance = 5.03σ

35

C. Gerber,
D0 single Top
Fermilab seminar
March 10, 2009

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?

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?

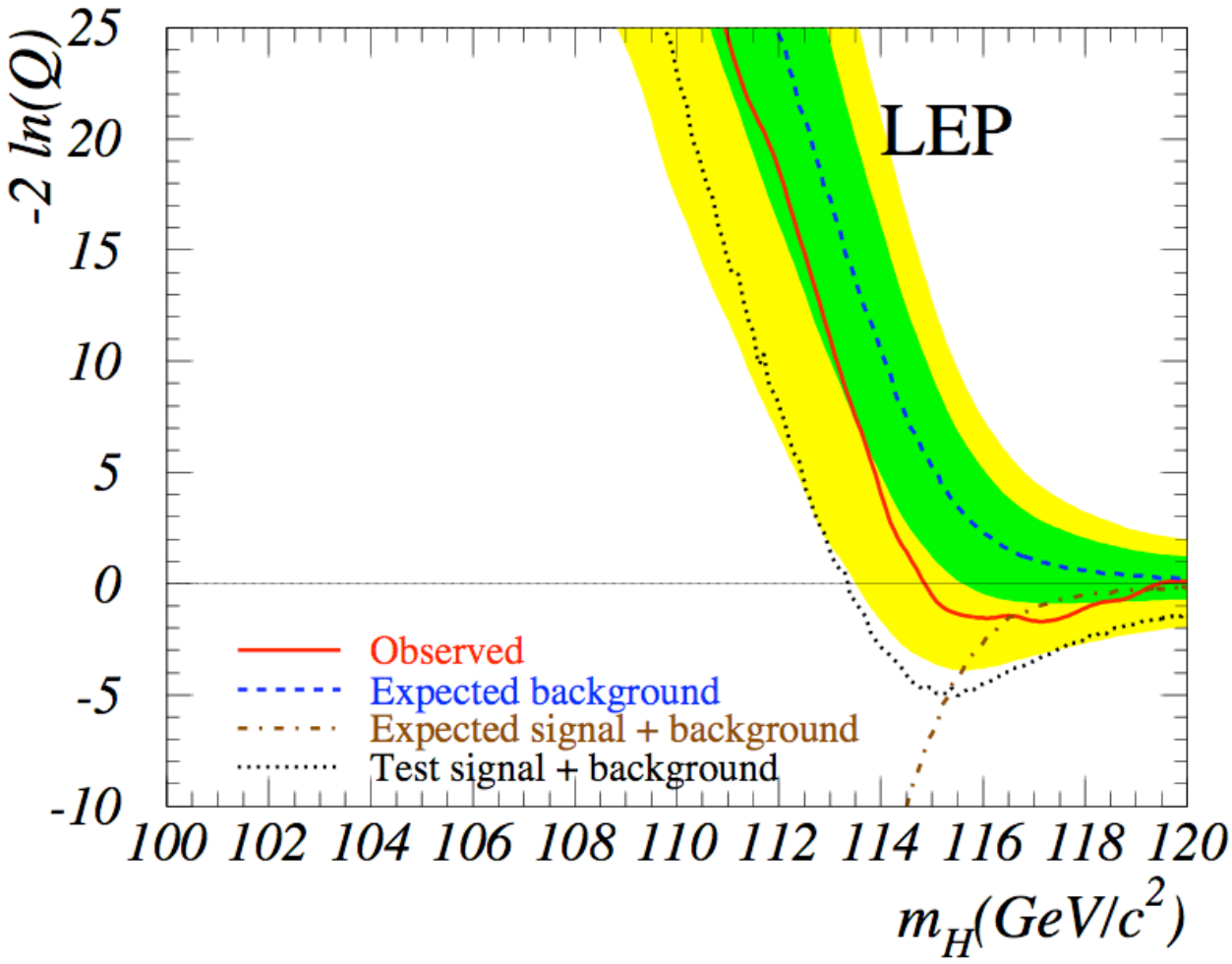
- Channels have different resolutions (or is resolution even the right word for a multivariate discriminant?
- Channels vary their weight in the combination as cross sections and branching ratios change with m_H

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.

Next to impossible unless somehow analyzers supply how each pseudo-dataset looks at each test mass.

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:

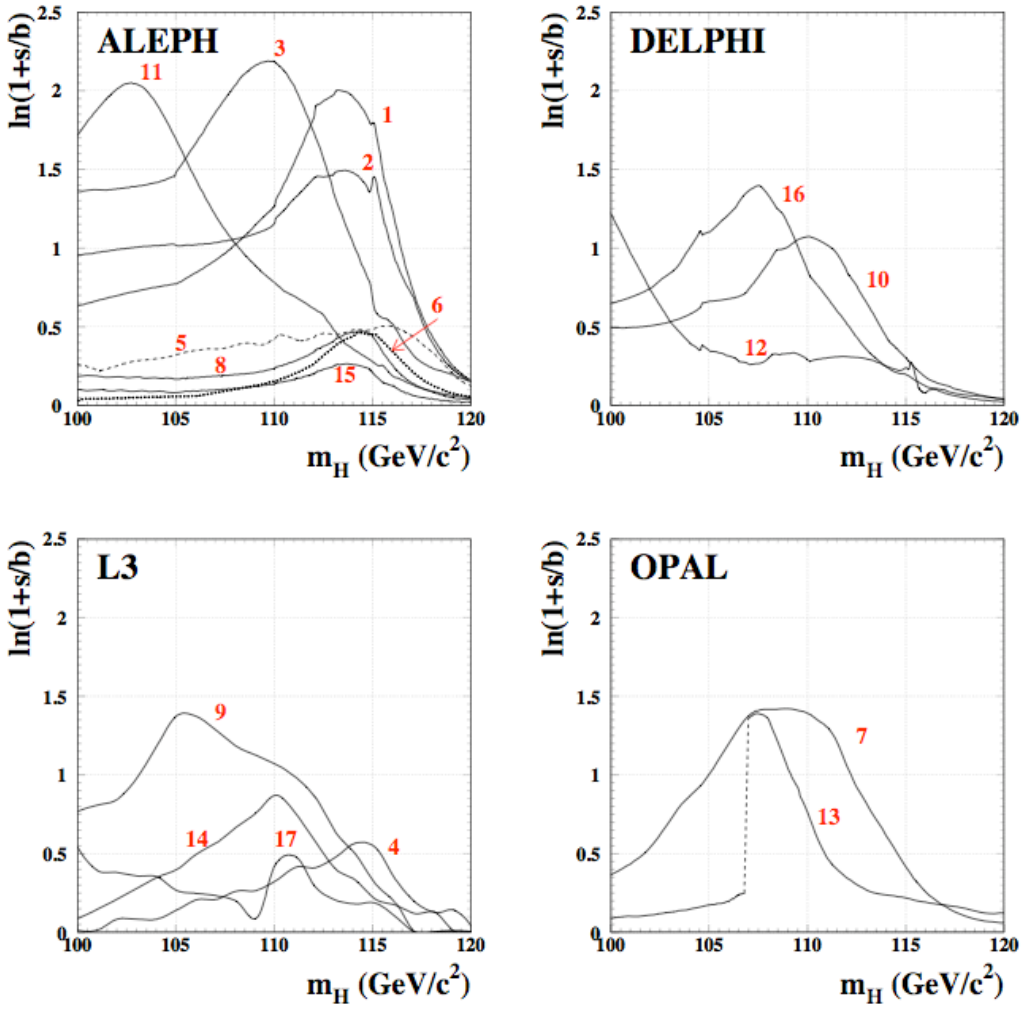


Individual Candidates Can Make a Big Difference

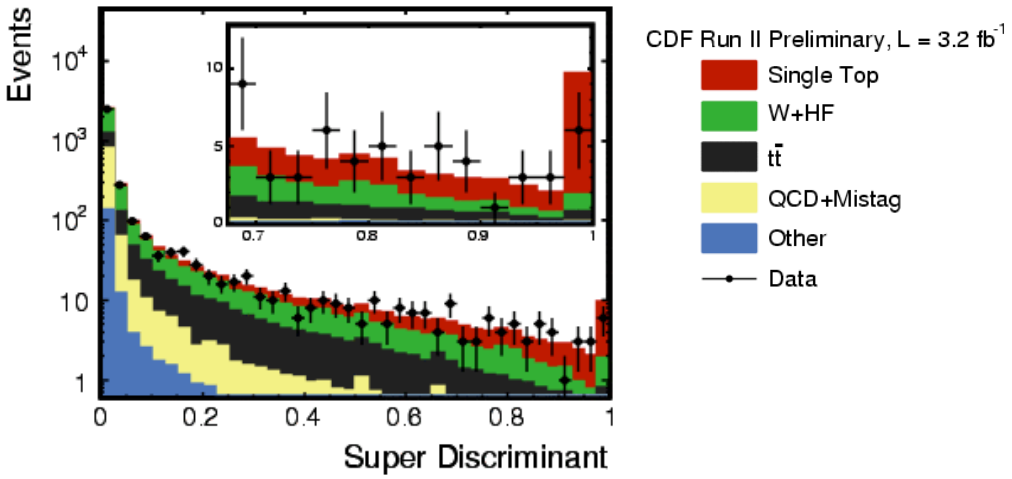
At LEP -- can follow individual candidates' interpretations as functions of test mass

if s/b is high enough near each one.

Fine mass grid -- smooth interpolation of predictions -- some analysis switchovers at different m_H for optimization purposes

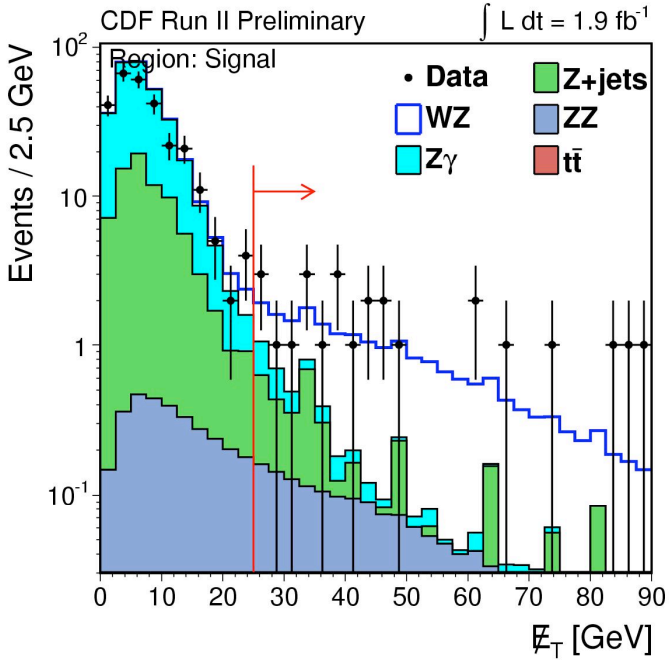


Even at a Hadron Collider, High s/b is possible!



Example -- CDF single top observation

Low s/b, high rate bins in the same histogram

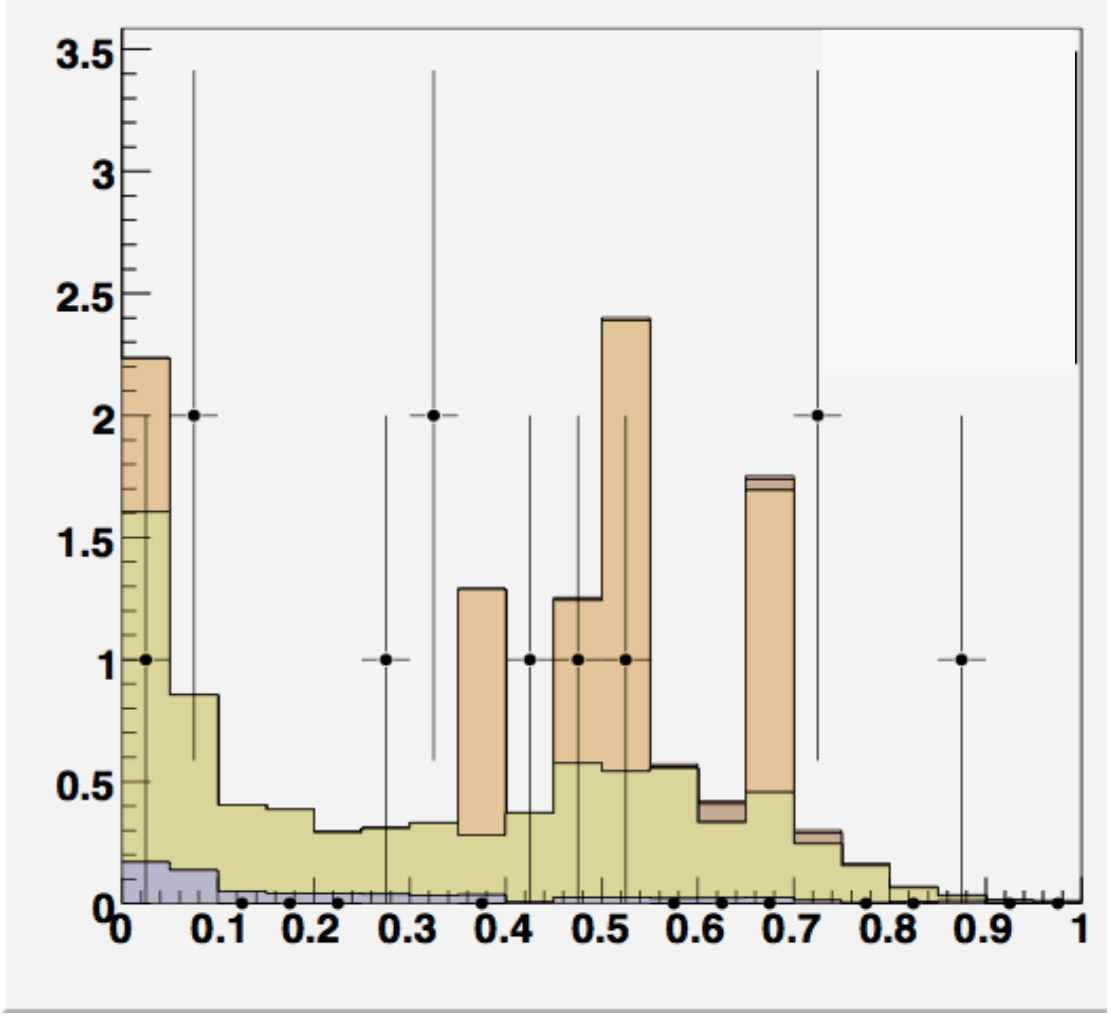


Example -- CDF's trilepton WZ measurement

Another issue -- each bin contains predictions using weighted MC -- and events have a broad spectrum of weights.

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.

Overcovers for discovery, undercovers for limits?

Measurement and Discovery are Very Different

Buzzwords:

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

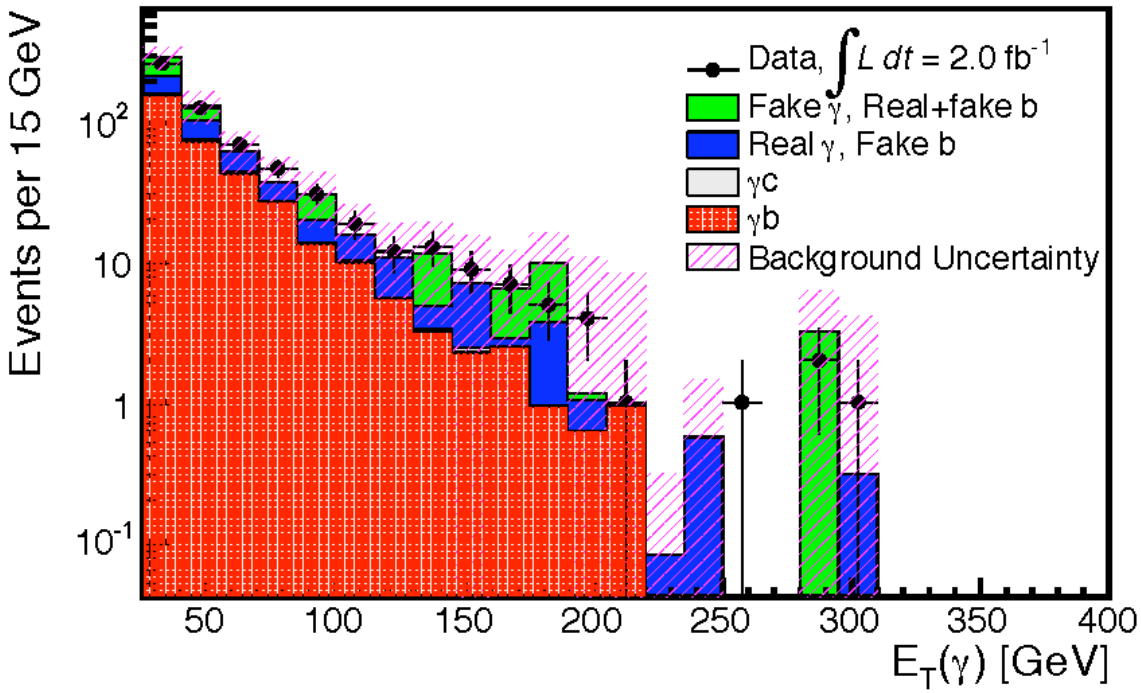
You can have a discovery and a poor measurement!
Example: Expected $b=2\times10^{-7}$ events, expected signal=1 event, observe 1 event, no systematics.

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

MC Statistics and "Broken" Bins



NDOF=?

- Limit calculators cannot tell if the background expectation is really zero or just a downward MC fluctuation.
- Real background estimations are sums of predictions with very different weights in each MC event (or data event)
- Rebinning or just collecting the last few bins together often helps.

- Advice: Make your own visible underflow and overflow bins (do not rely on ROOT's underflow/overflow bins -- they are usually not plotted. Limit calculators should ignore ROOT's u/o bins).

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.