Practical Issues in Statistical Interpretation of Tevatron Data

Thomas R. Junk
Fermilab

Progress on Statistical Issues in Searches Conference
SLAC National Accelerator Laboratory
June 5, 2012

- Multiple Parameters of Interest
- Handling Nuisance Parameters
- Overbinning, Smoothing, and Distributions that Ought Not to be Smoothed
- Model validation with Multivariate Analyses
- ABCD Methods
- Look-Elsewhere

Fermilab from the Air

Tevatron ring radius=1 km
Commissioned in 1983

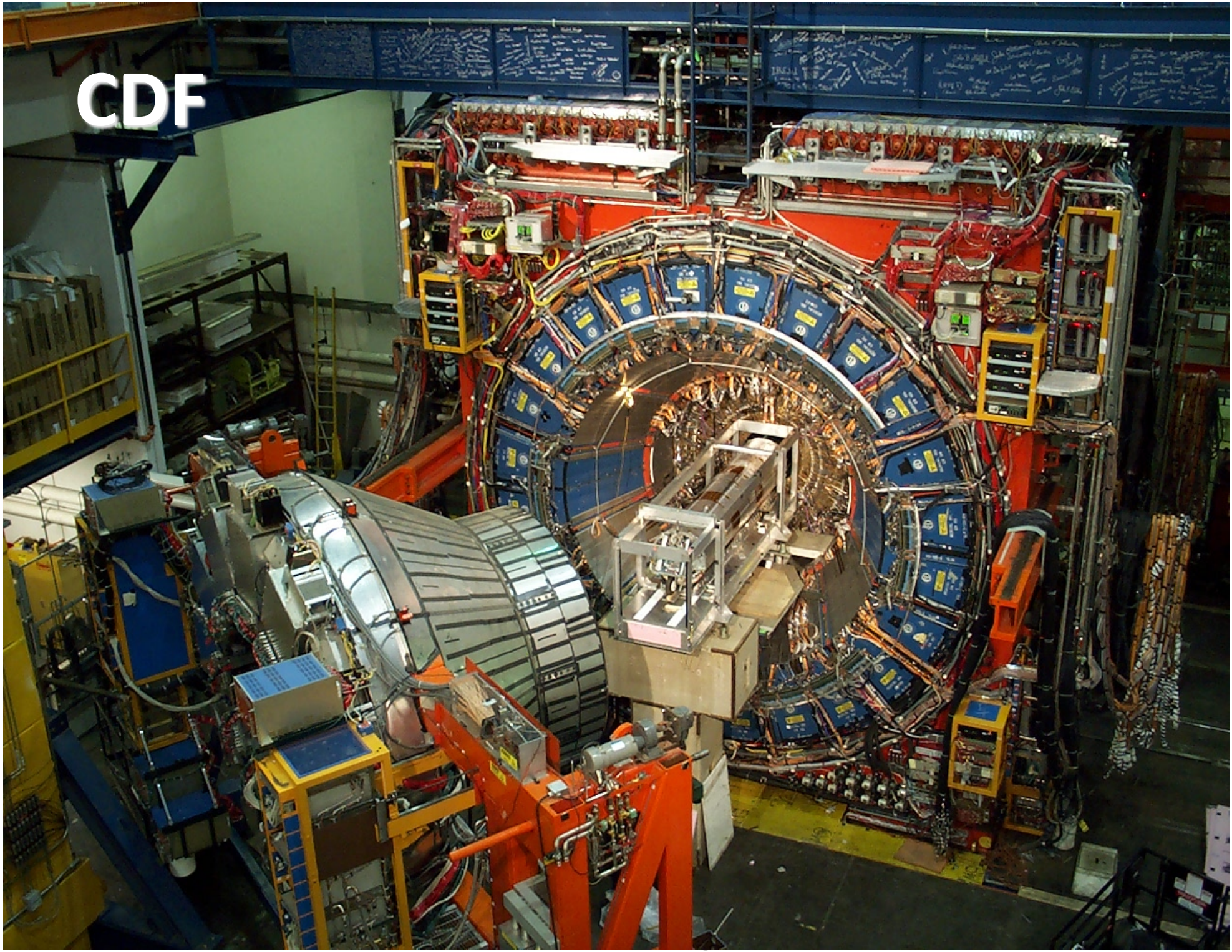
Protons-antiproton collisions for Run I and Run II

Run II: $\sqrt{s_{p\bar{p}}} = 1.96\ \mathrm{TeV}$

Main Injector commissioned for Run II

Recycler used as another antiproton accumulator

Run II ended Sep. 30, 2011
10 fb^{-1} of analyzable data/ experiment.
500+ papers/experiment and counting!


CDF

Two (or more) Parameters of Interest

For quoting Gaussian uncertainties on **single** parameters. Ellipse is a contour of 2ΔlnL=1

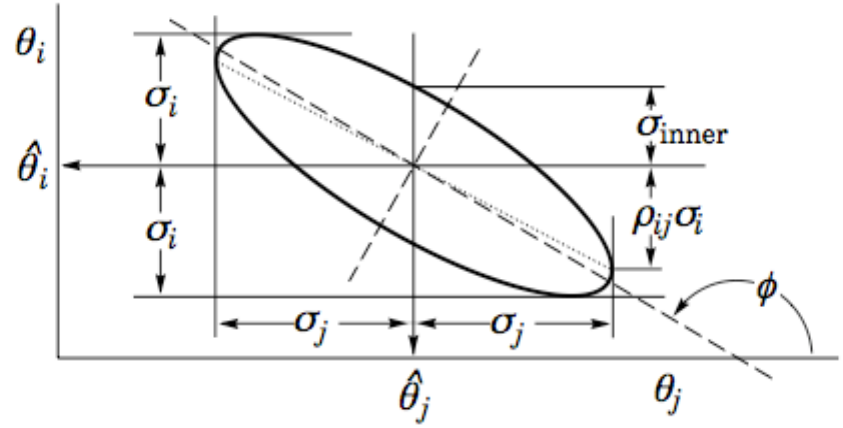


Figure 33.5: Standard error ellipse for the estimators $\widehat{\theta}_i$ and $\widehat{\theta}_j$. In this case the correlation is negative.

For displaying joint estimation of several parameters

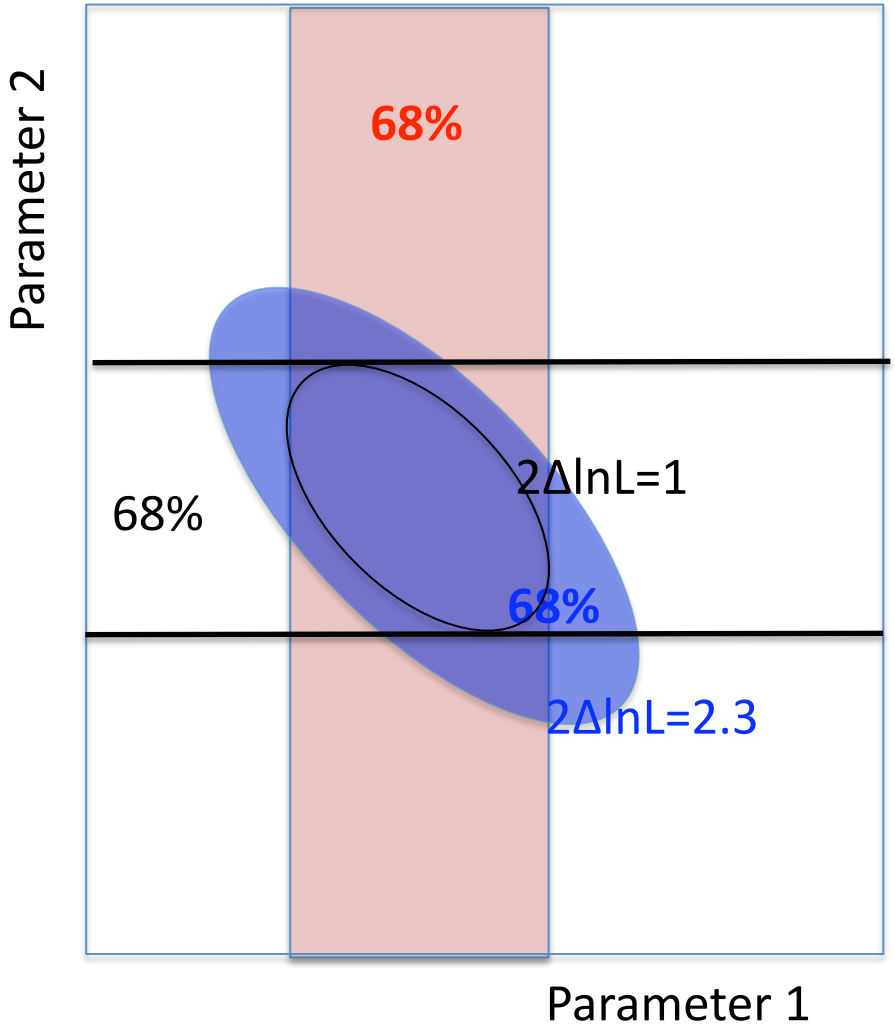
Table 33.2: $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

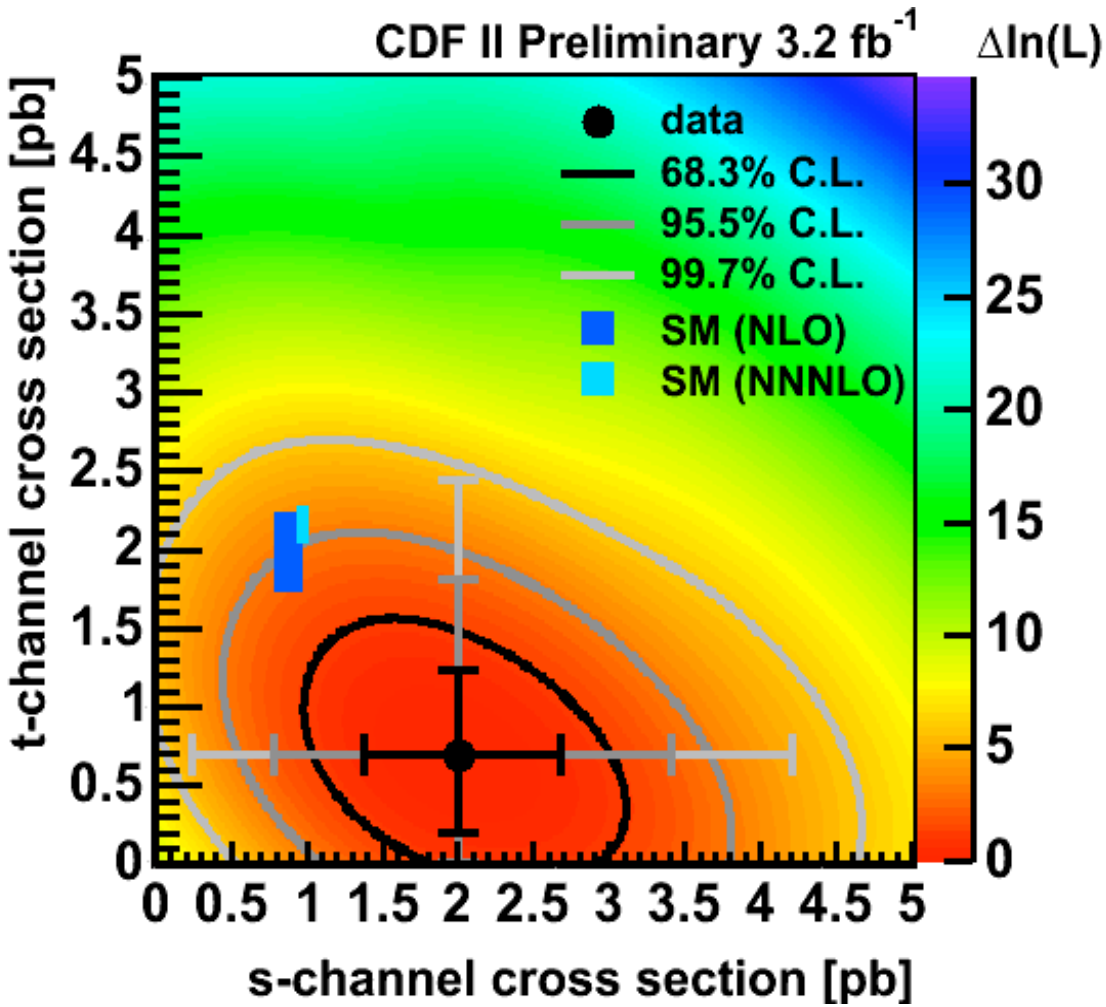
$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

From the 2011 PDG Statistics Review

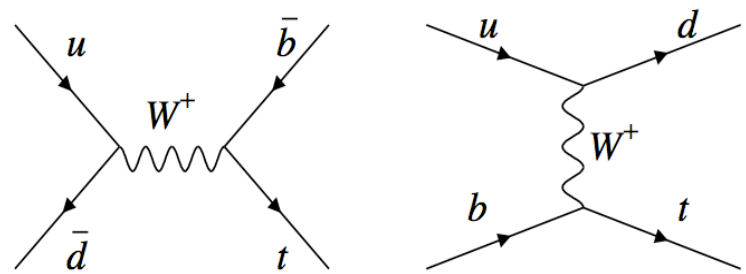
http://pdg.lbl.gov/2011/reviews/rpp2011-rev-statistics.pdf

1D or 2D Presentation

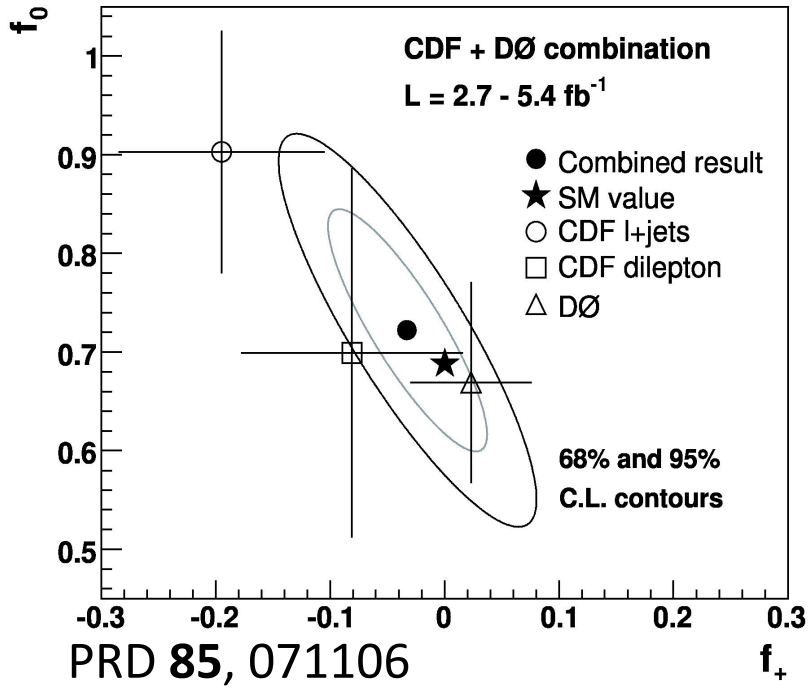




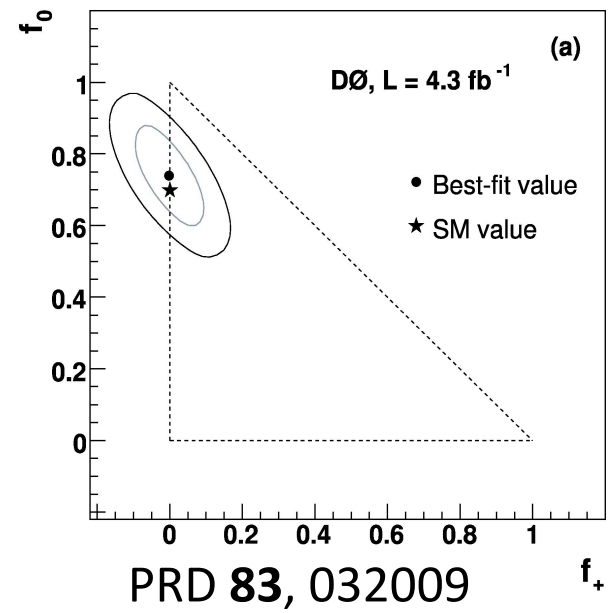
I prefer when showing a 2D plot, showing the contours which cover in 2D. The 2ΔlnL=1 contour only covers for the 1D parameters, one at a time.



A Variety of ways to show 2D Fit results

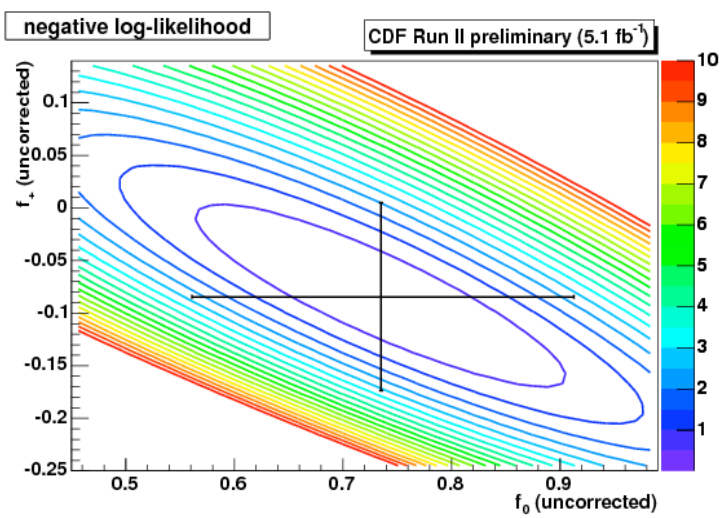


PRD **85**, 071106

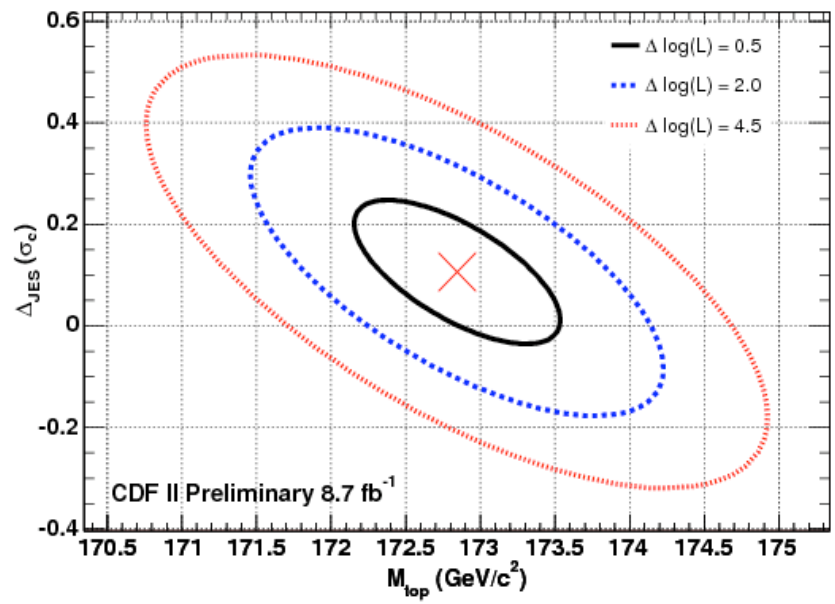


PRD **83**, 032009

68% and 95% contours



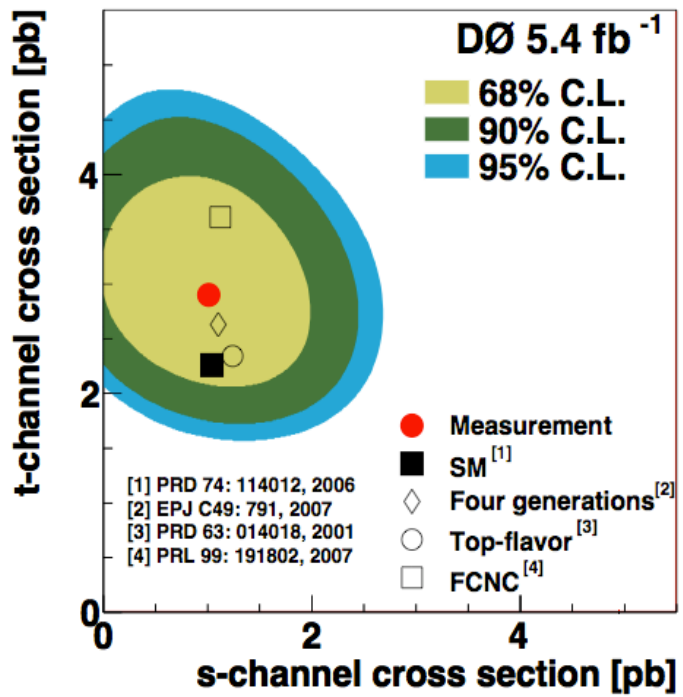
http://www-cdf.fnal.gov/physics/new/top/2011/WhelDil/index.html



Hypothesis Testing with Two Parameters of Interest?

See for example, D0's evidence for t-channel single top production:
Phys.Lett. B **705** (2011) 313-319

"... using the log-likelihood approach
... we compute for the first time the significance of the *tqb* cross section independently of any assumption on the production rate of *tb*."

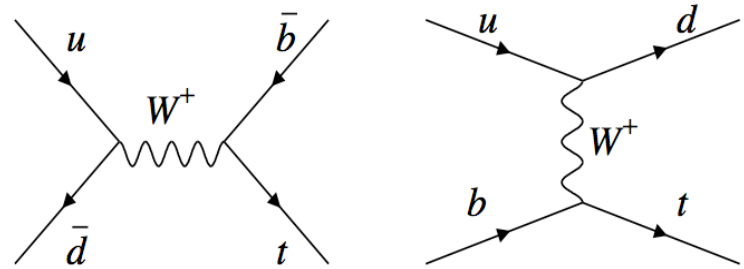


In this specific case the correlation is small, and this claim isn't so bad. But to calculate a p-value $p(\lambda \geq \lambda_{obs} | \text{signal}=0)$, one needs a sample space of pseudoexperiments, and thus an assumption of the s-channel rate.

Hypothesis Testing with Two Parameters of Interest?

Another issue: A variation on the LEE theme.

s-channel and t-channel single top events are differ in their kinematic distributions



For observation and measurement of the total single top production rate σ_{s+t}, we'd train our MVA's on the Standard Model mixture of both.

For separate measurement of σ_s and σ_t, we have choices of training strategies. Single-tag events – t-channel; Double-tag events – s-channel.

Suppose we want to do a hypothesis test on just the t-channel? Re-optimize all MVA's with t-channel as the signal and s-channel as background. Re-do this for s-channel.

Ideally we'd pick the most sensitive MVA (highest expected significance) for the test we want to do, but there is a temptation to pick the one with the highest measured significance (we tell analyzers to pick the most sensitive).

If you have an excess of data events that could be s- or t-channel (can't tell), this strategy may end up giving an observation of both (or neither), when we're really only sure there's at least one process present.

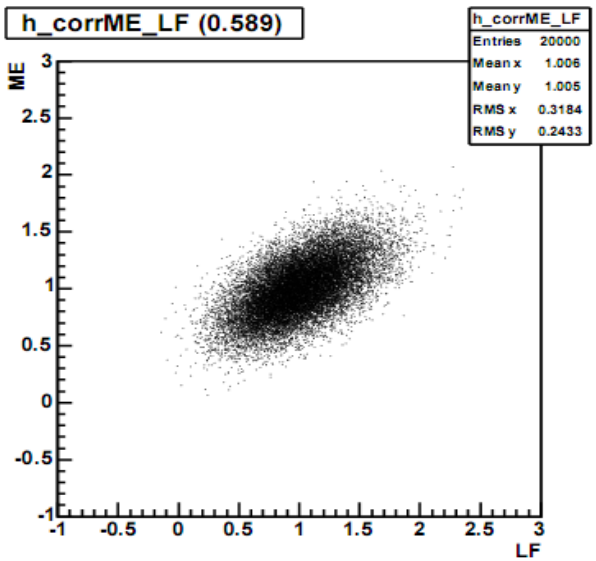
Several Analyses on the Same Data

- Different groups are interested in the same search/measurement using the same data.
- May have slightly different selection requirements (Jet energies, lepton types, missing Et, etc).
- Usually have different choices of MVA or even training strategies for the same MVA
- Always will give different results!

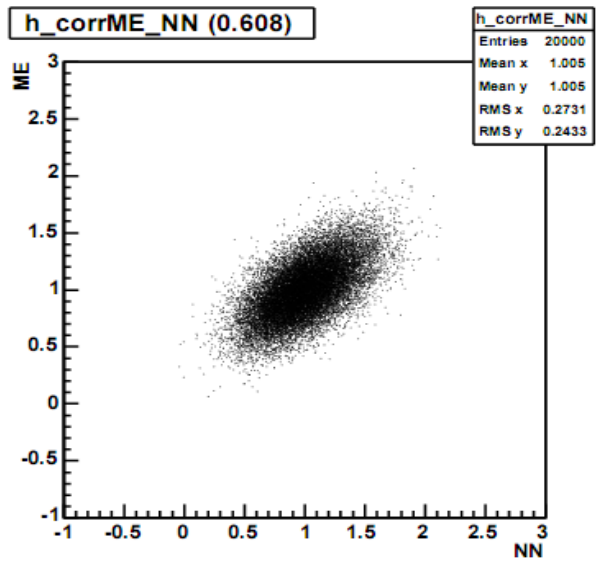
- What to do?
  - Pick one and publish it – criterion: best sensitivity. Median expected limit, median expected p-value, median expected measurement uncertainty. How to pick it if the result is 2D? Need a 1D figure of merit.
  - Can check consistency with pseudoexperiments. A p-value using Δ(measurement) as a test statistic. What's the chance of running two analyses on the same data and getting a result as discrepant as what we got?
  - Combine MVA's into a super-MVA
    - Keeps everyone happy and involved
    - Usually helps sensitivity
    - Requires coordination and alignment of each event in data and MC
    - Easiest when overlap in data samples is 100%. Otherwise have to break sample up into shared and non-shared subsets and analyze them separately
- What not to do: Pick the one with the "best" observed result. (LEE!)

An Example of Running Three Analyses on the Same Events in Monte Carlo Repetitions

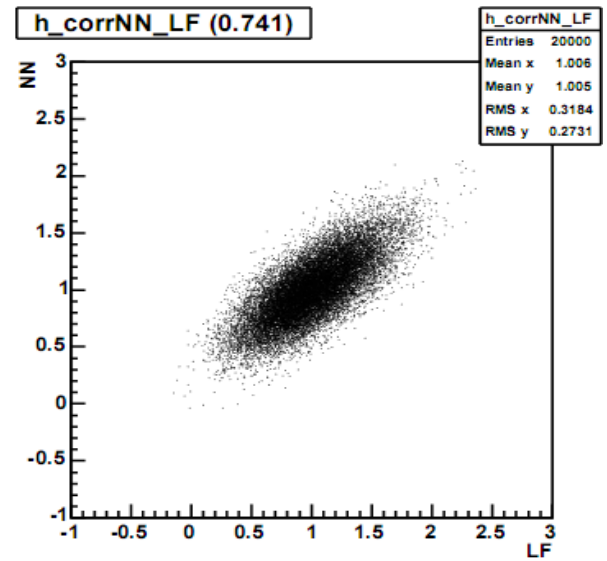
LF-ME 58.9%



ME-NN 60.8%



LF-NN 74.1%



Different questions can be asked: What's the distribution of the maximum difference between the measurements any two teams? What's the quadrature sum of the pairwise differences? Condition on the sum? (Probably not..)

Systematic Uncertainty Handling

For a very thorough review, see Luc Demortier's note:
"P Values: What They Are and How to Use Them "

http://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf

Plausible options:

1) Prior-Predictive method
2) Supremum method
3) Confidence-Level method
4) Plug-In p-values
5) Define all nuisance parameters to be parameters of interest
6) Define only the important nuisance parameters to be parameters of interest

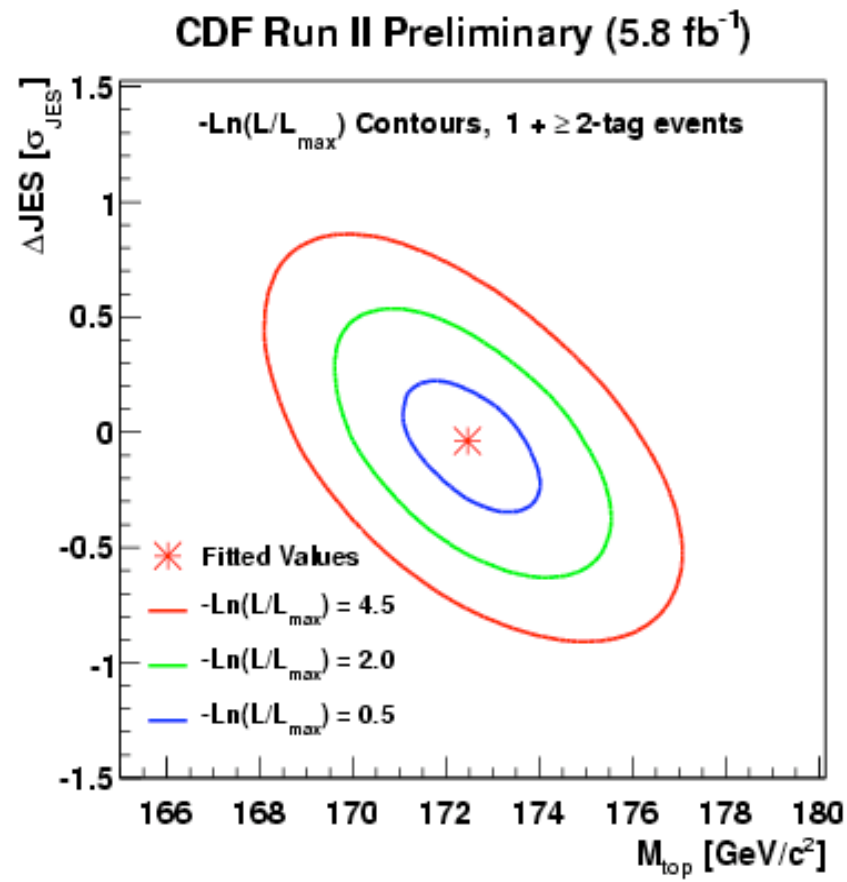
The prior-predictive method is a mixture of Bayesian and Frequentist reasoning

The supremum method is very conservative and I argue not fully non-Bayesian. It also produces mixed results – can have an outcome which is an excess over background when setting limits and a deficit when computing a p-value.

Treat Nuisance Parameters As Parameters of Interest!

- One person's nuisance parameter is another's parameter of interest.

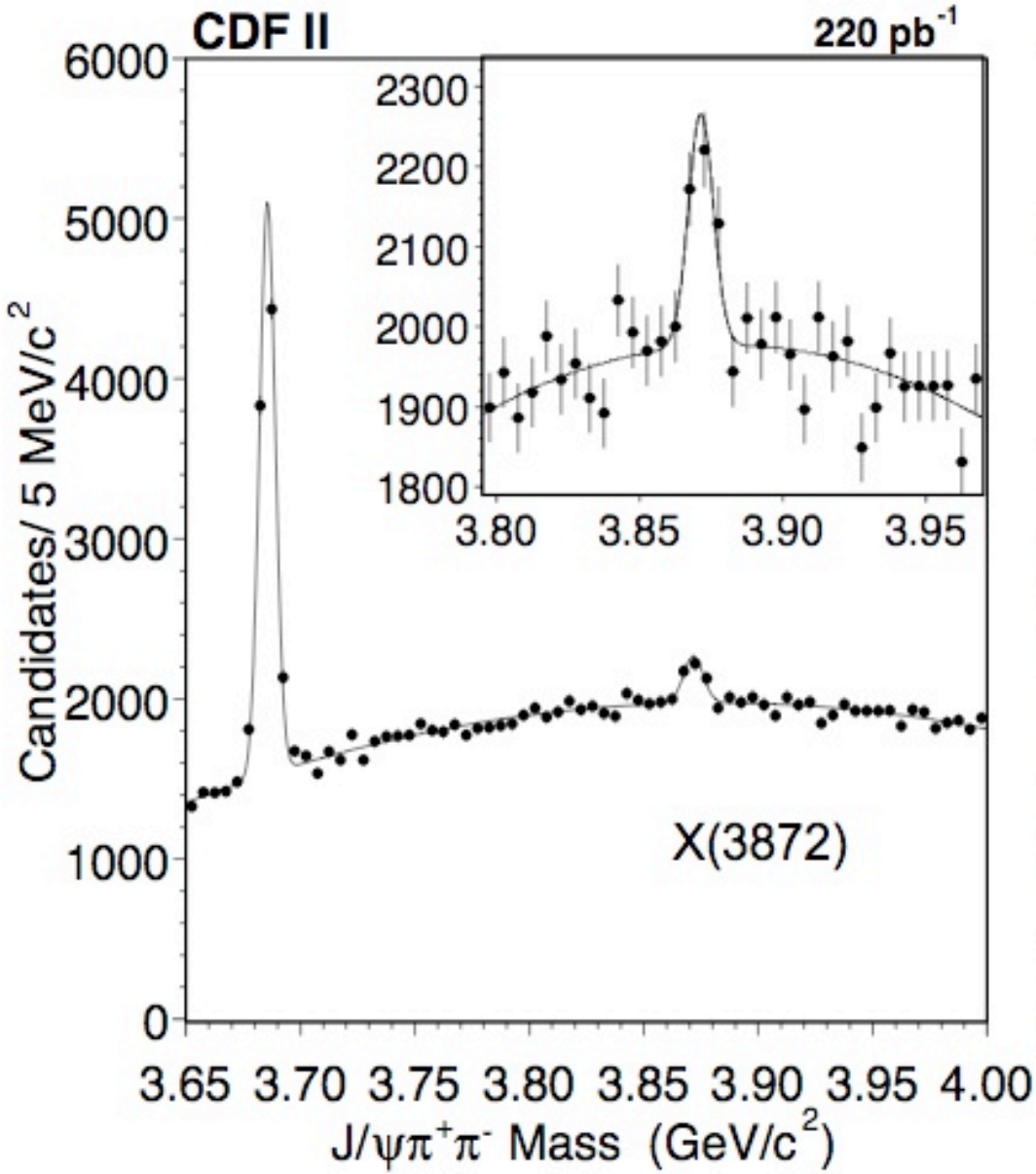
- Really only good if you have one dominant source of systematic uncertainty, and you want to show your joint measurement of the nuisance parameter and the parameter of interest.



Difficult to apply to cases with many nuisance parameters.

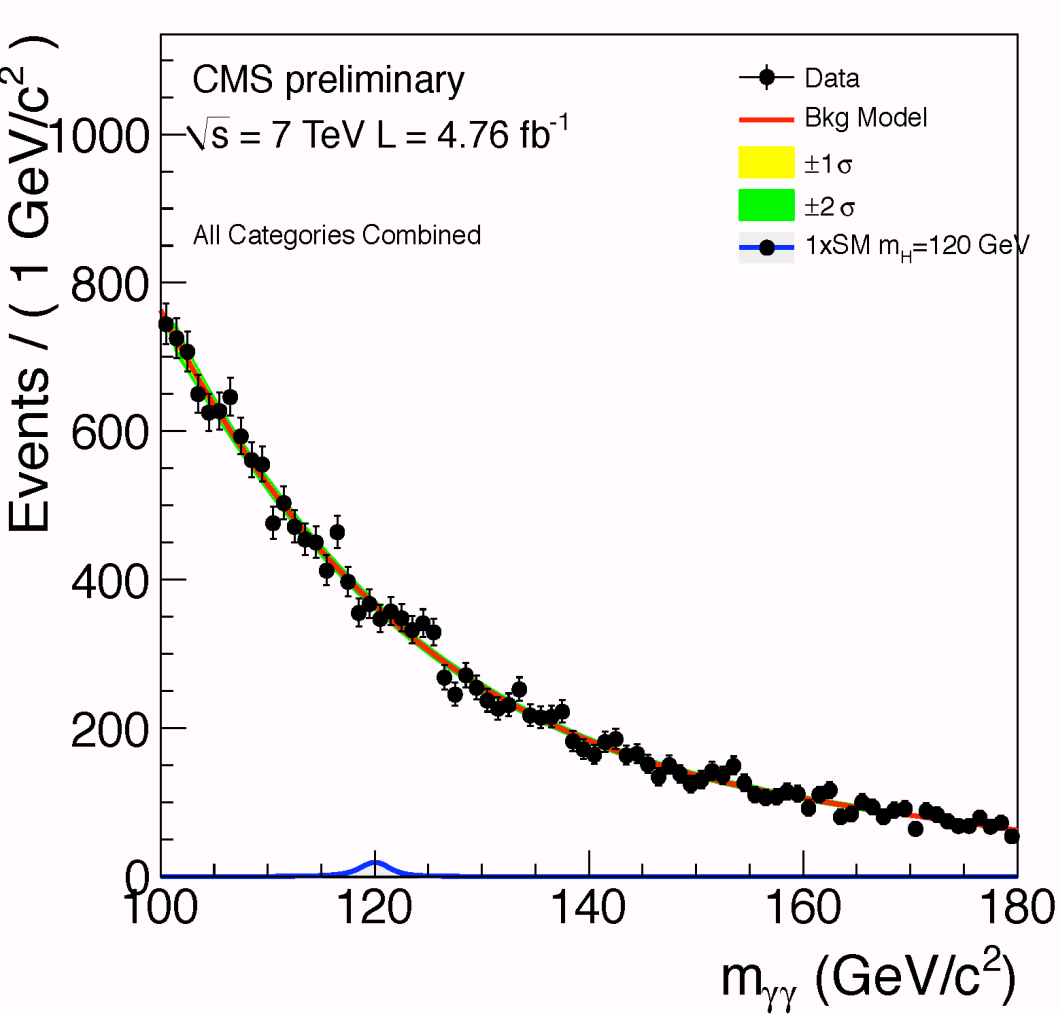
Example: top quark mass (parameter of interest), vs. CDF's jet energy scale in all-hadronic ttbar events. Doesn't follow my suggestion! But one parameter (JES) is not a parameter of "interest" If it were, we'd use ΔLnL=1.15 instead of 0.5

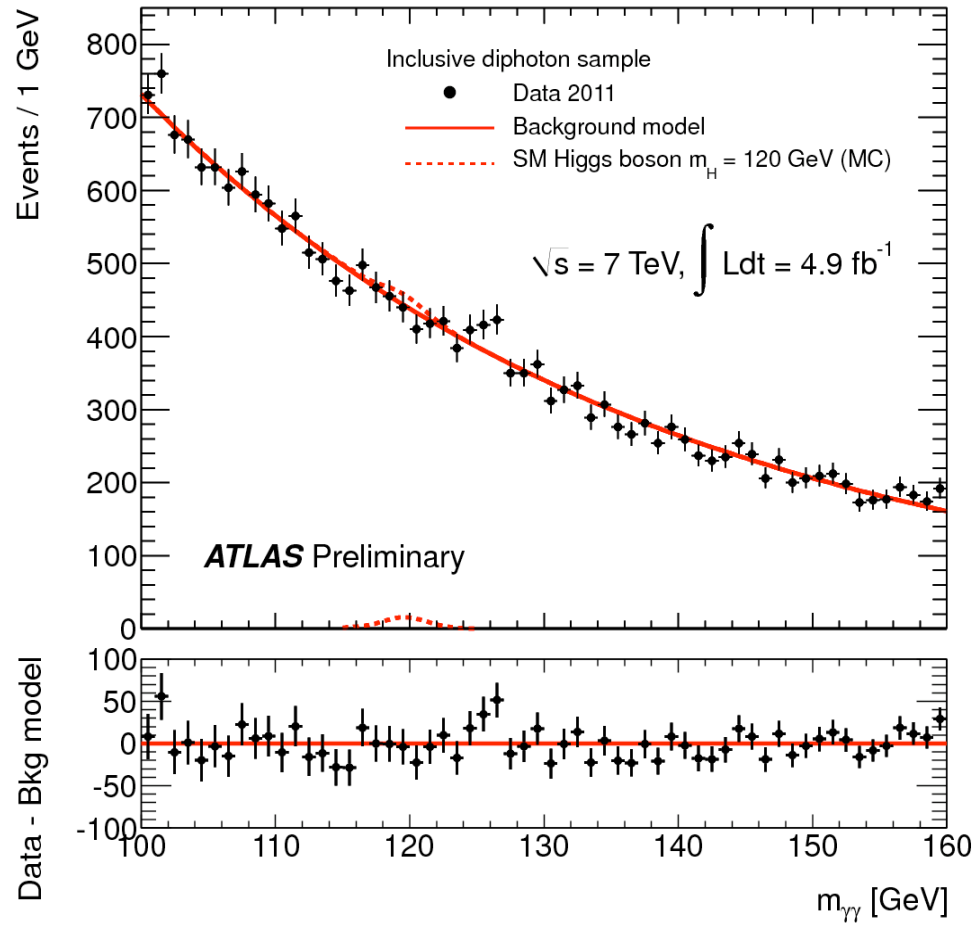
“Strong” Sideband Constraints



Guess a shape that fits the backgrounds, and fit it with a signal.

Another Strong Sideband Constraint Example





"Weak" Sideband Constraints

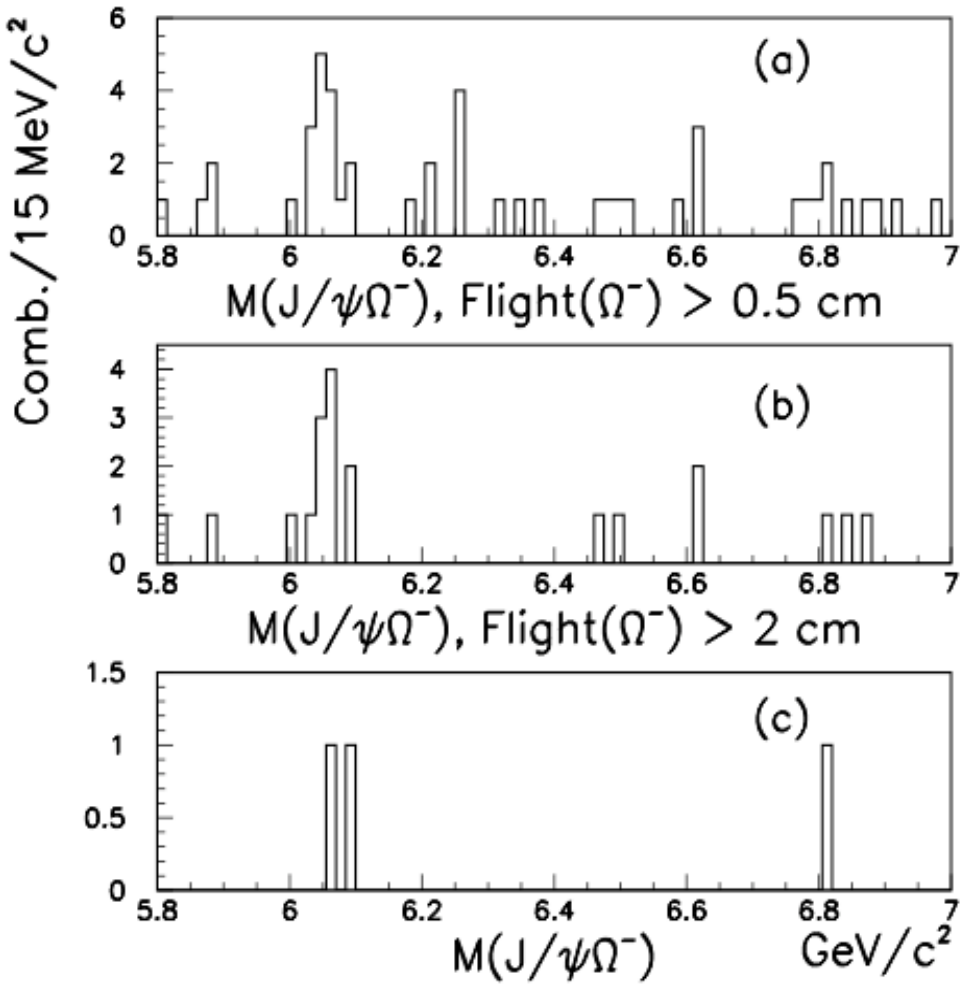


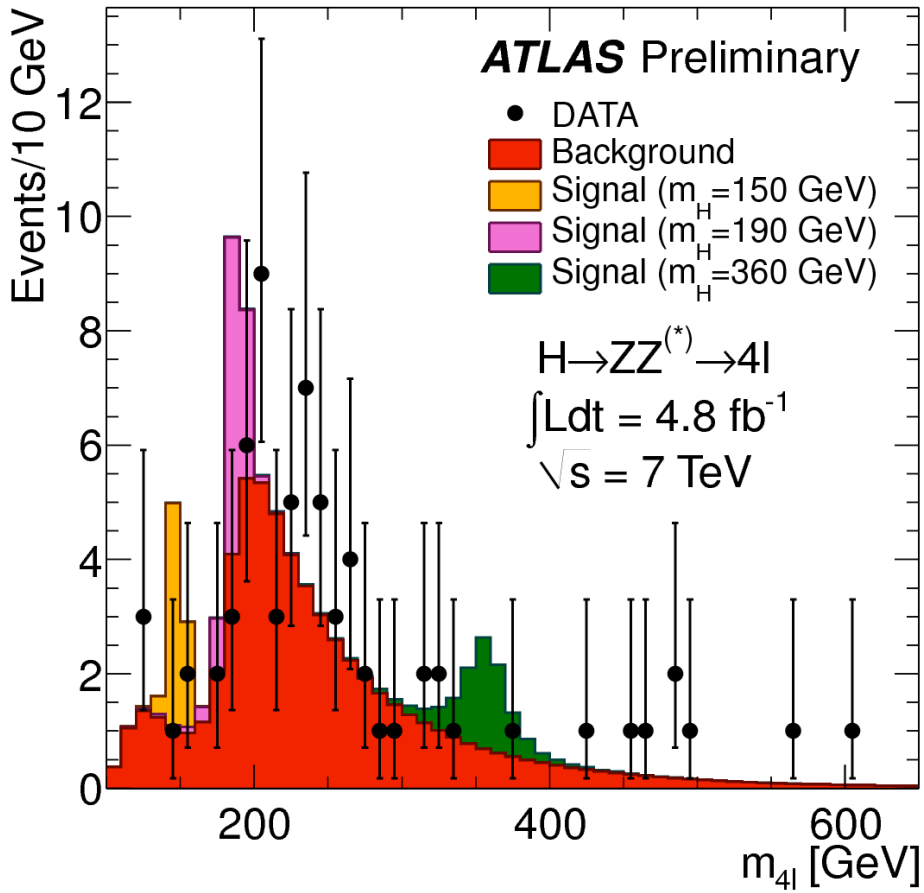
FIG. 8: (a,b) The invariant mass distribution of $J/\psi\,\Omega^-$ combinations for candidates where the transverse flight requirement of the Ω^- is greater than 0.5 cm and 2.0 cm. (c) The invariant mass distribution of $J/\psi\,\Omega^-$ combinations for candidates with at least one SVXII measurement on the Ω^- track. All other selection requirements are as in Fig. 5(c).

CDF's Ω_b observation paper:

Phys.Rev. D80 (2009) 072003

A Mixture of Theory and Data is Needed for a more Complicated situation

H→ZZ→Four Leptons



Main Backgrounds:

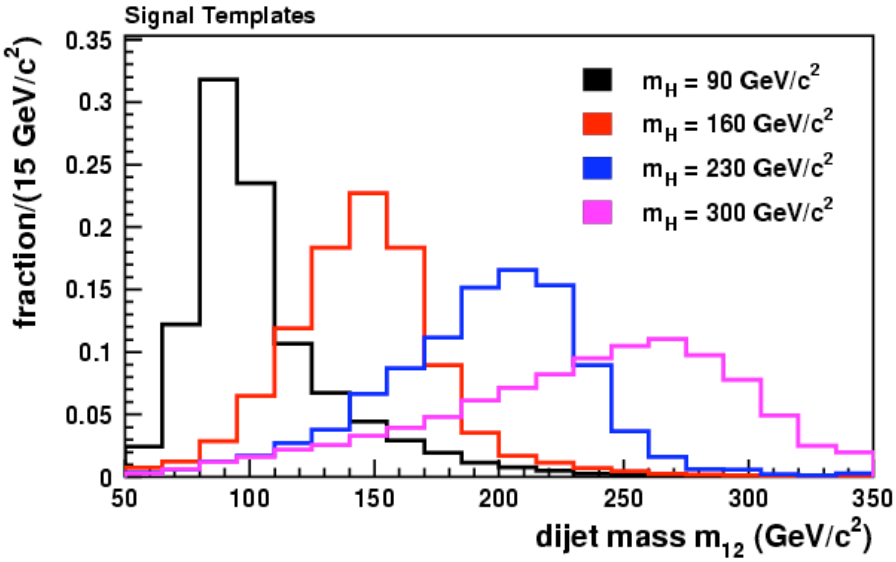
pp → ZZ (MC)*theory
pp → Z+jets, ZW+jet(s), ... Data
pp → ttbar

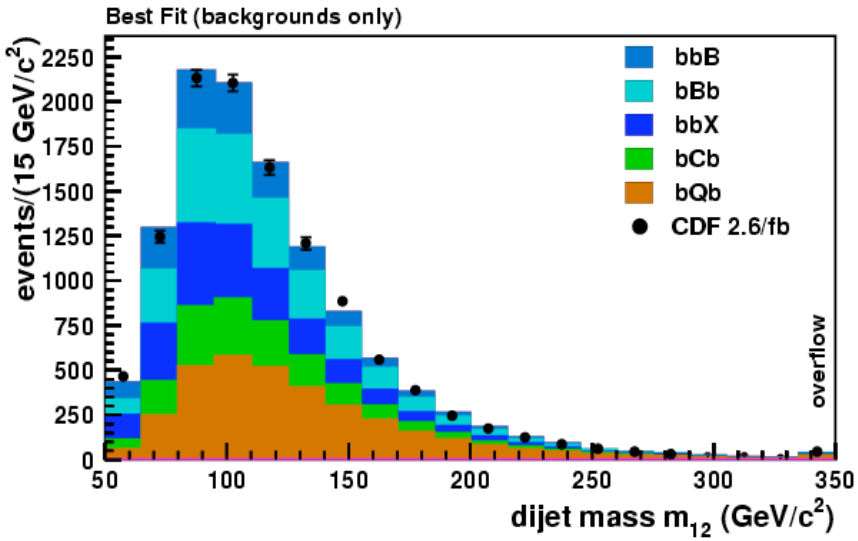
Low-mass 4L side: off-shell Z's, "radiative tail", and other backgrounds

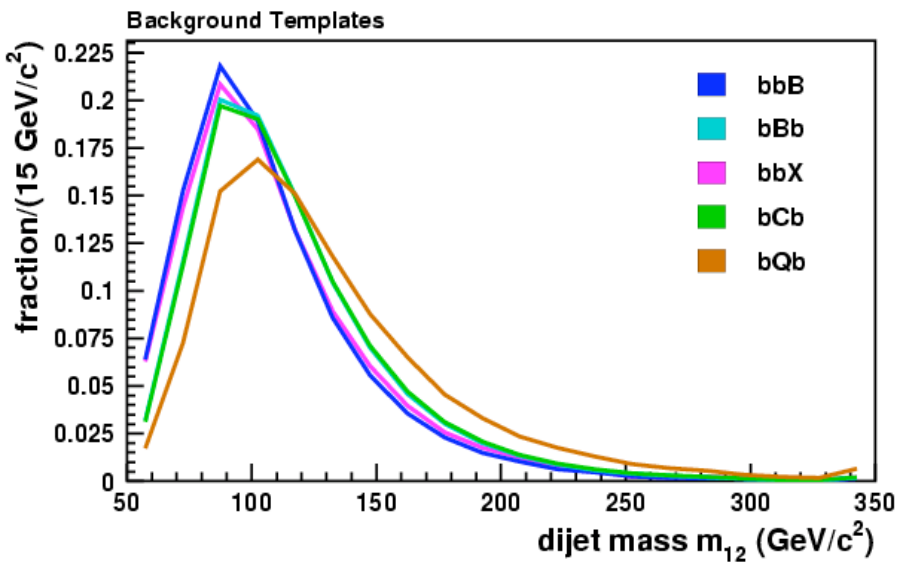
Dependent on theoretical predictions of the shape of the dominant ZZ background.

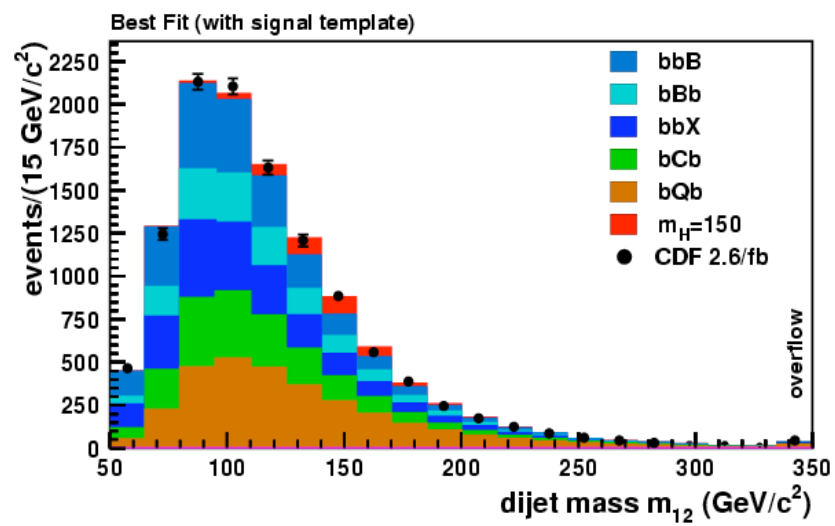
With more data, replace "bad" systematiccs with "good" ones (theory replaced with data). But in the early stages, the "bad" systematic uncertainties are smaller!

Another Weak Sideband Constraint Example that Looks Like a Strong Sideband Constraint



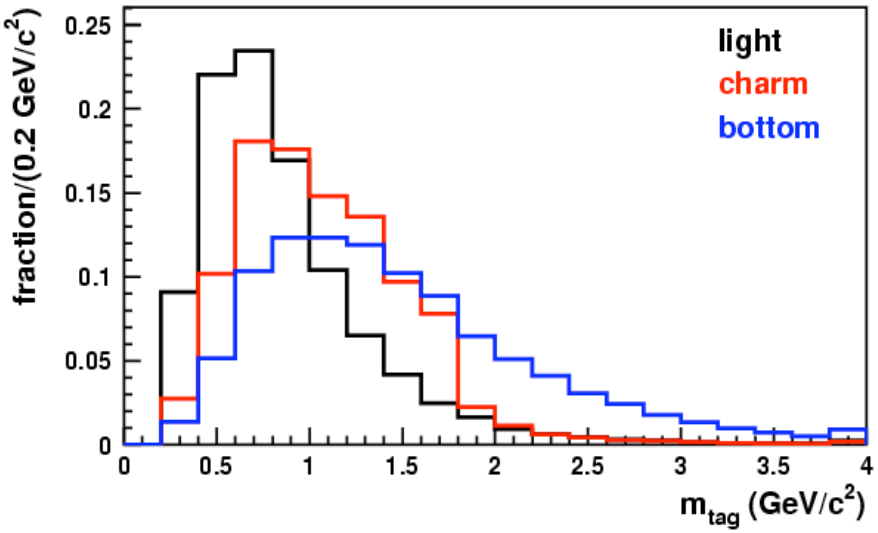


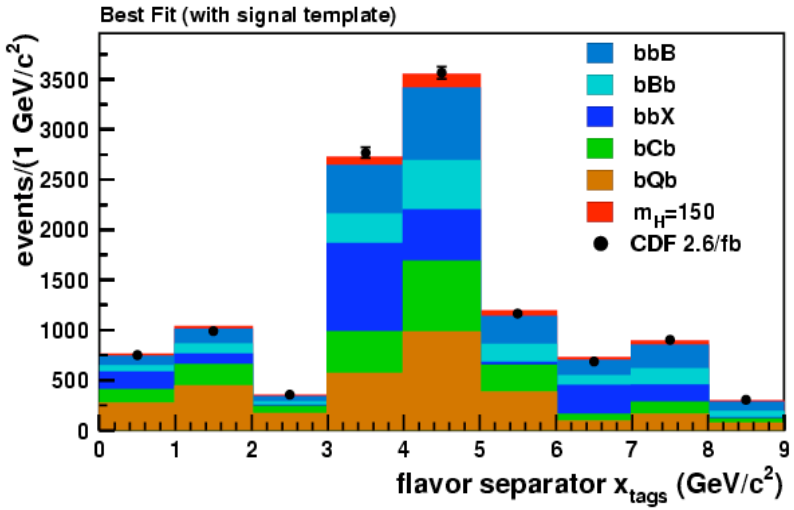


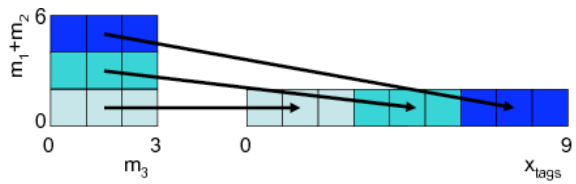


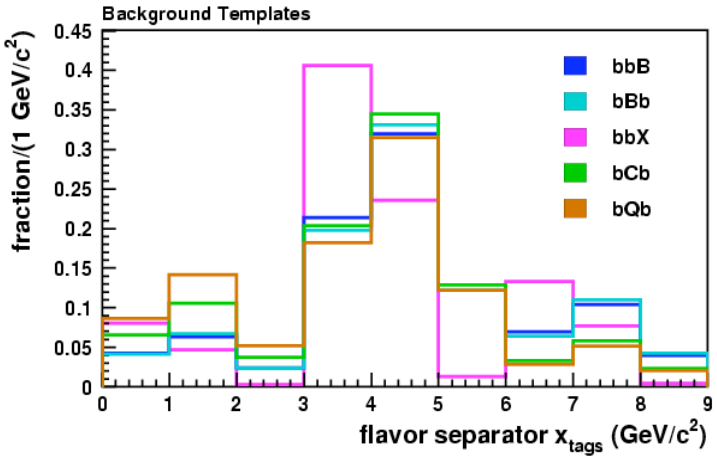
Phys.Rev. D85 (2012) 032005
e-Print: arXiv:1106.4782 [hep-ex]

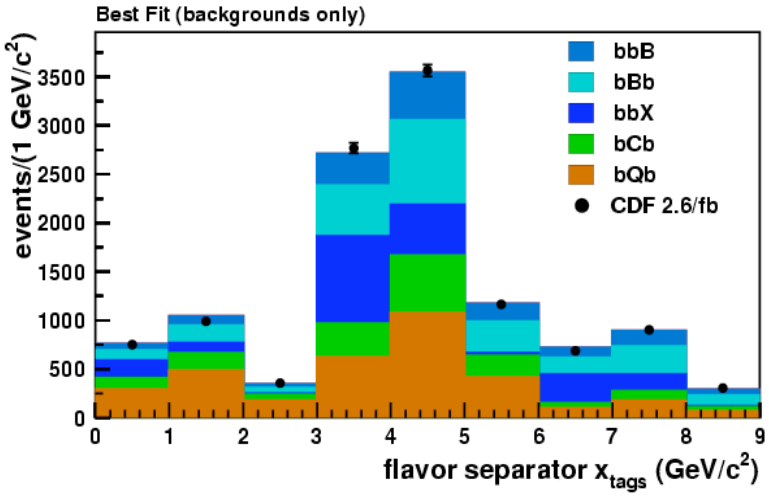
Breaking the Flavor Degeneracy with a Tag Variable











No Sideband Constraints?

Example: Counting experiment, only have a priori predictions of expected signal and background

All test statistics are equivalent to the event count – they serve to order outcomes as more signal-like and less signal-like. More events == more signal-like.

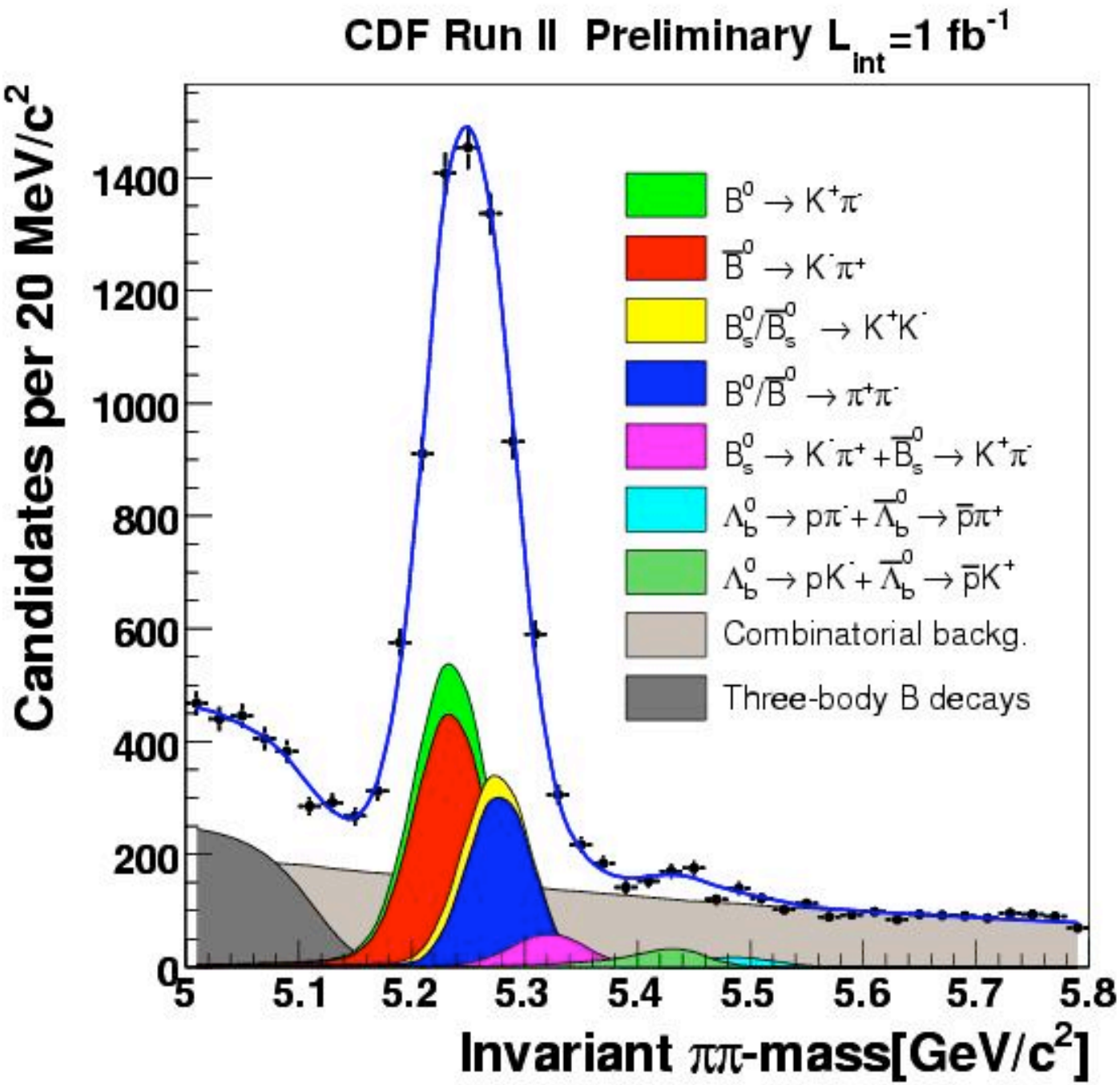
Classical example: Ray Davis's Solar Neutrino Deficit observation. Comparing data (neutrino interactions on a Chlorine detector at the Homestake mine) with a model (John Bahcall's Standard Solar Model). Calibrations of detection system were exquisite. But it lacked a standard candle.

How to incorporate systematic uncertainties? Fewer options left.

Another example: Before you run the experiment, you have to estimate the sensitivity. No sideband constraints yet (except from other experiments).

Prior predictive method then is equivalent to the profile method using the control samples to estimate nuisance parameters. And it's more general in cases that the signal contamination of the sidebands is important.

Several "on's", Several "off's"



More than one “off” sample

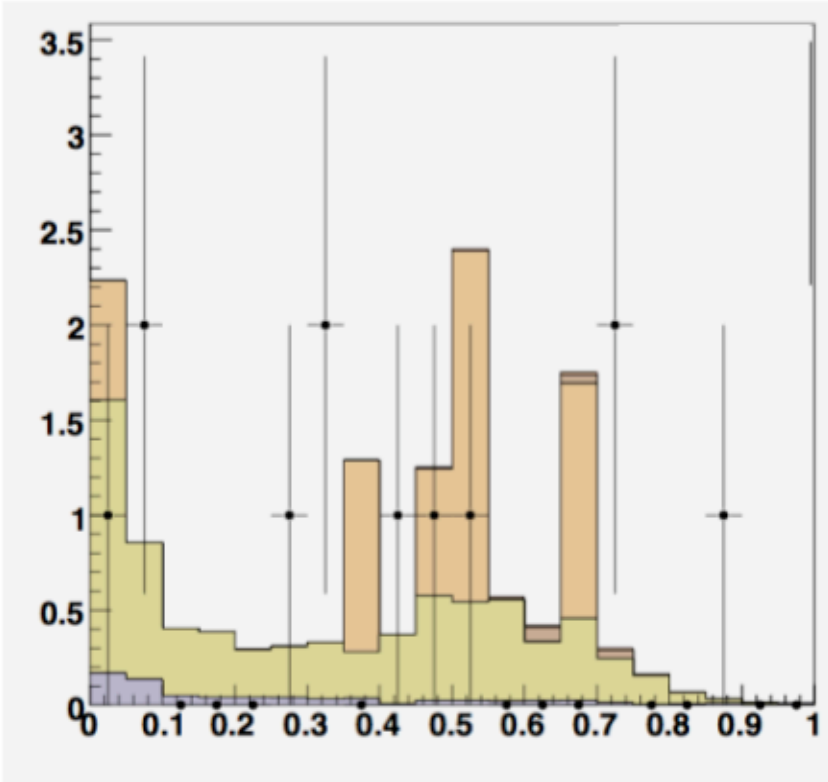
Conflicting estimates of background – what to do?

Very typical example: Pythia vs. Herwig (here the “off” samples are Monte Carlo).
“Take the difference as a systematic”
“Take the average and half the difference as a systematic”
Try to learn something about which one is more reliable.

But you can invert more than one cut and have “conflicting” off samples in the data too. Really the extrapolation factors or the sample composition estimates are what’s wrong, not the actual data.

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?

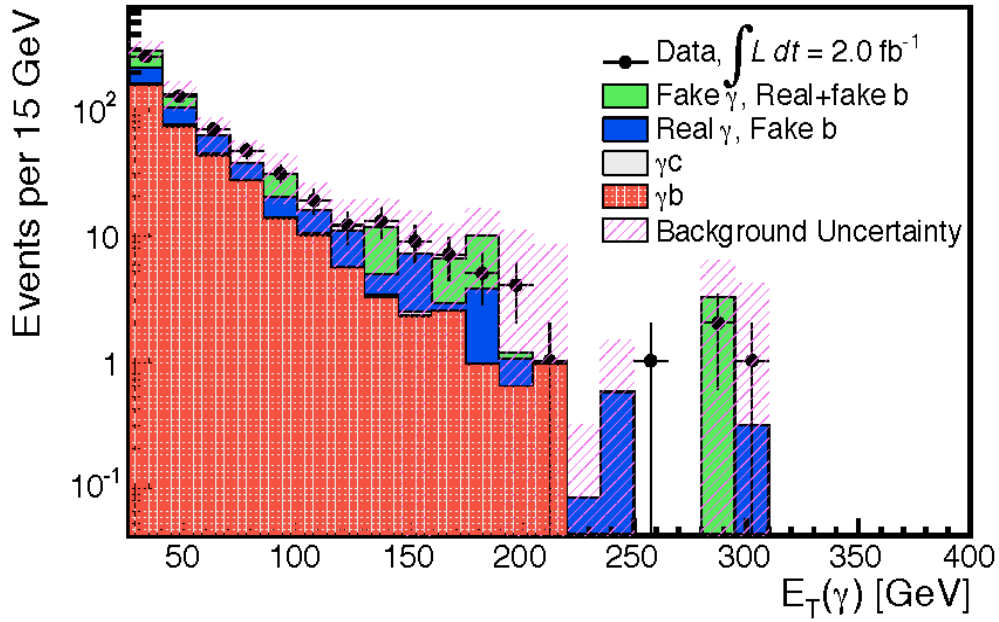
See Glen Cowan's talk yesterday

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

Lesson learned: Try to do a better job with the predictions! Statistical methods won't save us.

MC Statistics and "Broken" Bins

Overbinning is like overtraining a NN. s, b, and d can all be in different bins. A histogram can be partially overbinned, like this one here:

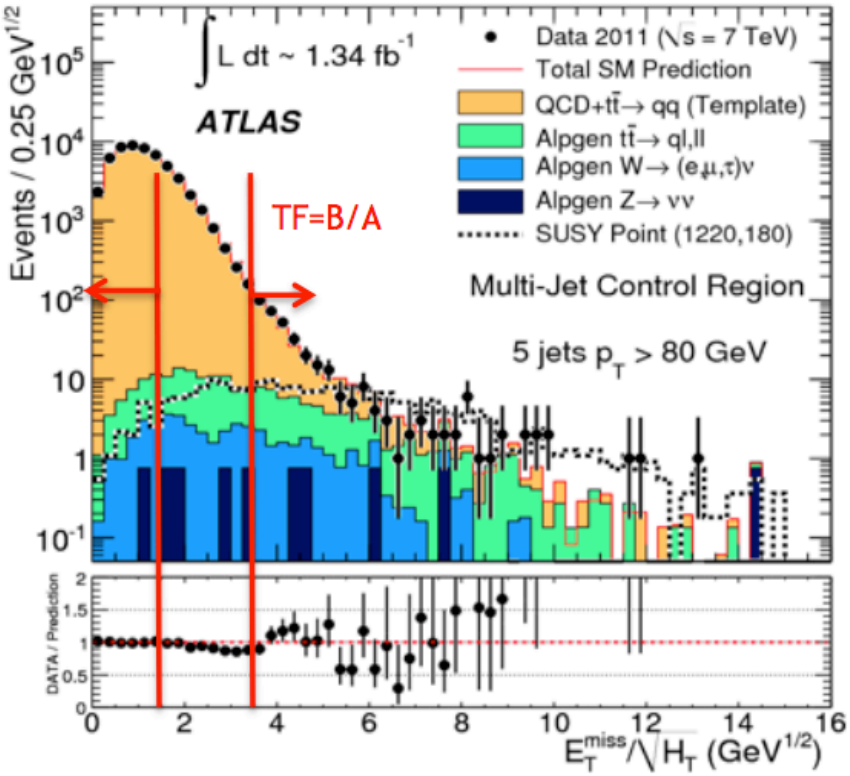


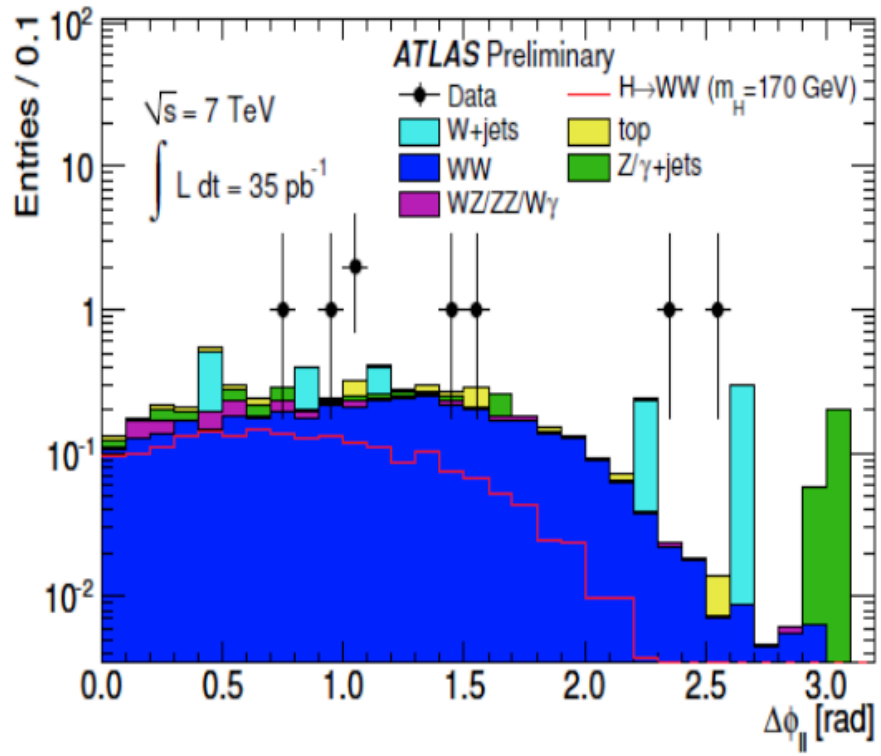
NDOF=?

- Limit calculators/discovery tools 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.
- Problem compounded by requiring shape uncertainties to be evaluated! Alternate shape MC samples are often even more thinly populated than the nominal samples. Validation of adequate preparation of results is necessary? (but what are the criteria?)

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!





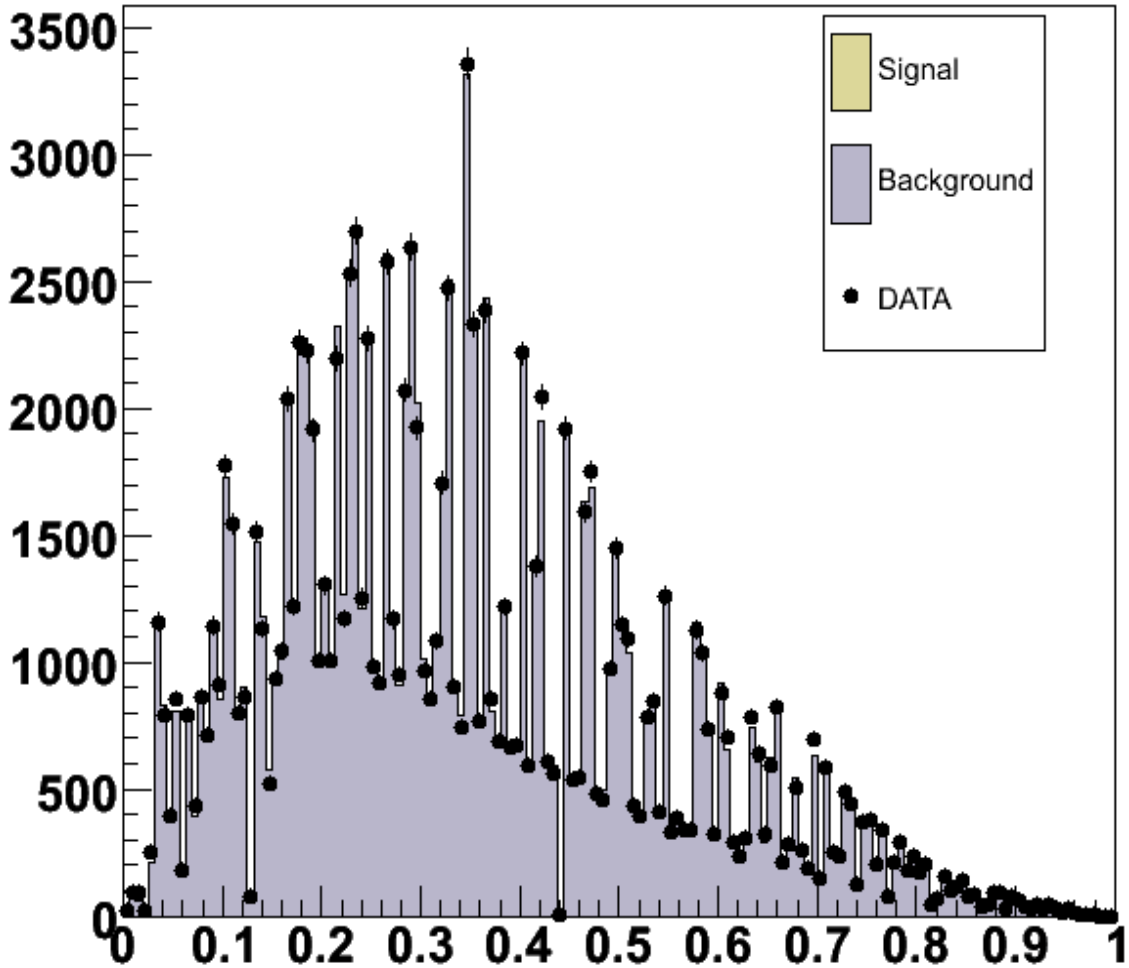
Data points' error bars are not sqrt(n). What are they? I don't know. How about the uncertainty on the prediction?

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.

This Histogram is Probably Okay

The binning is a little odd, though. You can get this kind of distribution from a decision tree or a likelihood MVA. (forest of delta functions)

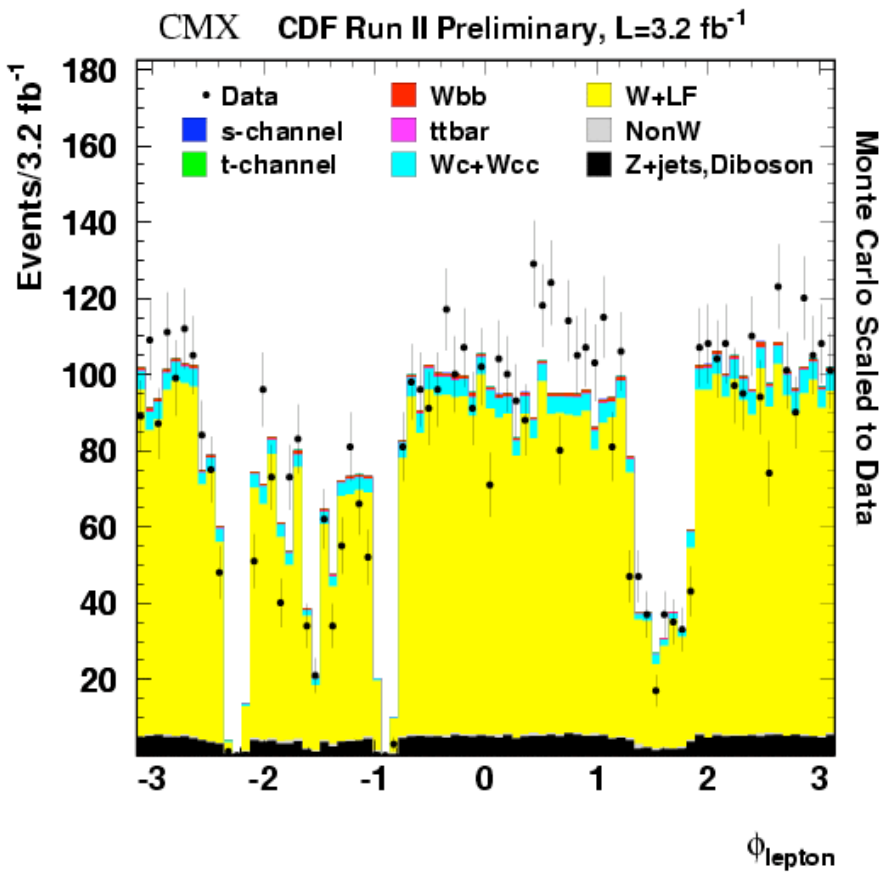
Watch out though! Smoothing and some kinds of interpolations (E.G. horizontal morphing a la Alex Read) are inappropriate for this distribution.

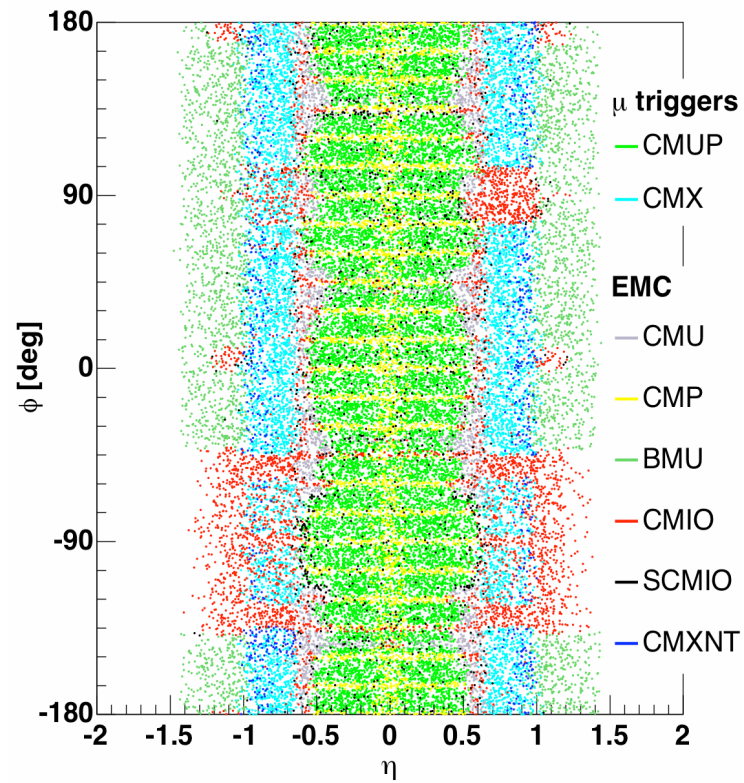


MVA mark

Sometimes distributions like these have natural causes: Lepton ϕ distributions for detectors with many cracks, for example.

A More Common Example – muon coverage at high angles.





No smoothing/extrapoation allowed here!

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.

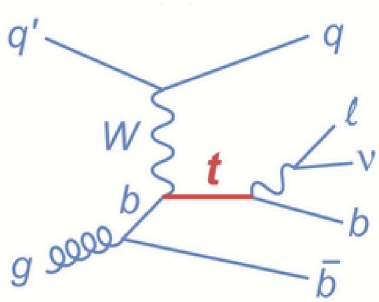
Note: Empty data bins are okay – just empty prediction is a problem. It is our job however to properly assign s/b to data events that we did get (and all possible ones).

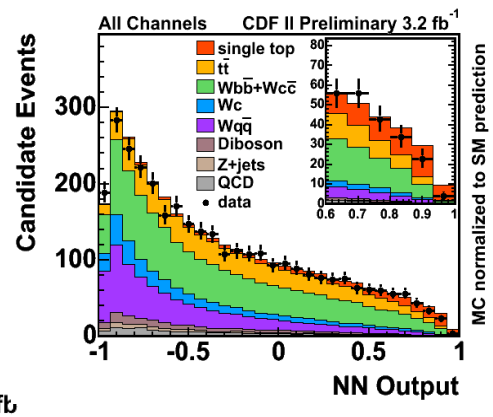
Model Validation

- Not normally a statistics issue, but something HEP experimentalists spend most of their time worrying about.
- Systematic Uncertainties on predictions are usually constrained by data predictions.
- Often discrepancies between data and prediction are the basis for estimating systematic uncertainty

Checking Input Distributions to an MVA

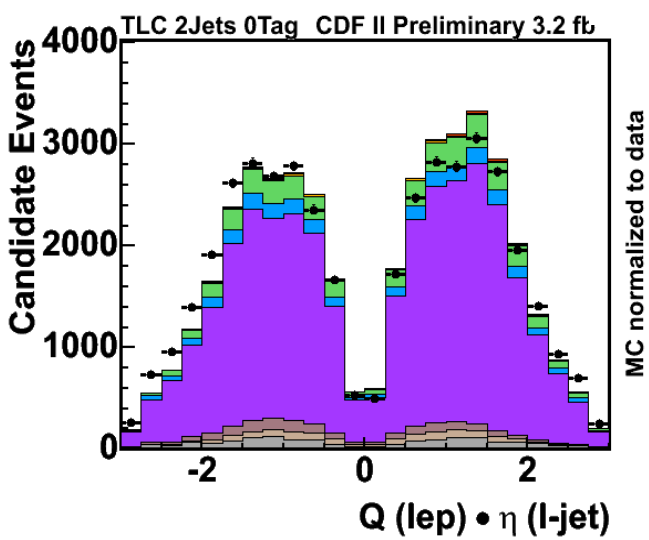
- Relax selection requirements – show modeling in an inclusive sample (example – no b-tag required for the check, but require it in the signal sample)
- Check the distributions in sidebands (require zero b-tags)
- Check the distribution in the signal sample for all selected events
- Check the distribution after a high-score cut on the MVA

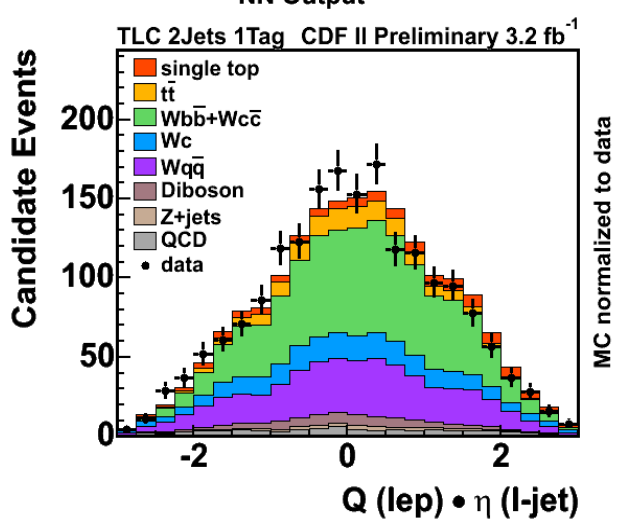


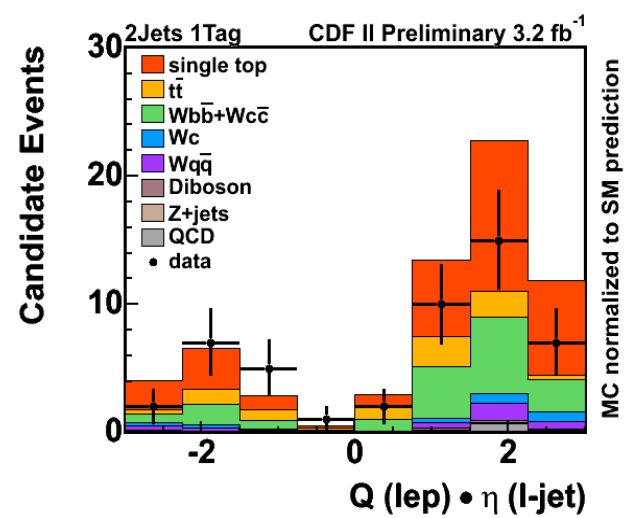


Example: $Q_{lepton} * \eta_{untagged\ jet}$ in CDF's single top analysis. Good separation power for t-channel signal.

Phys.Rev.D82:112005 (2010)







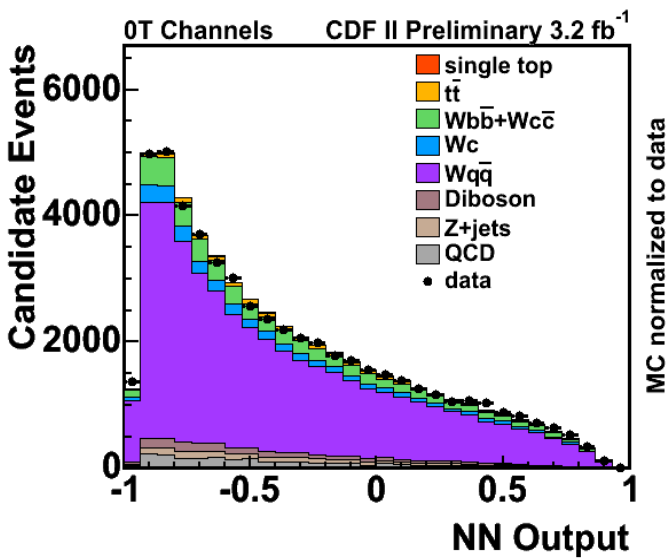
highest |η| jet as a well-chosen proxy

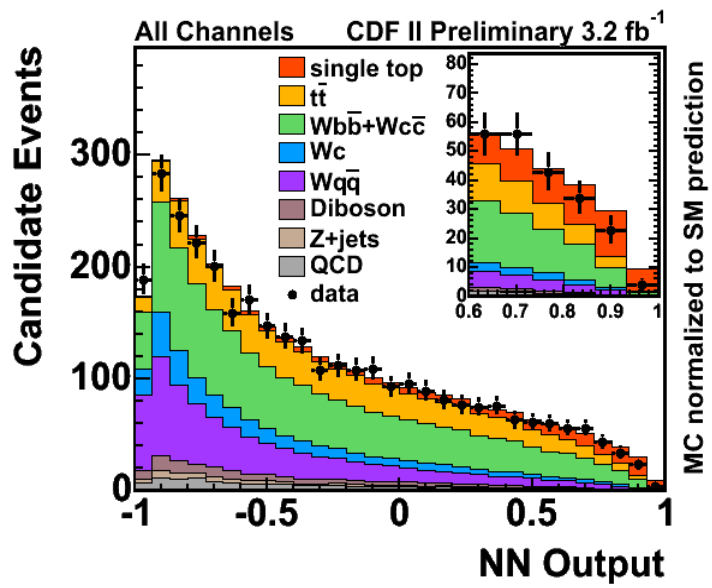
Checking MVA Output Distributions

- Calculate the same MVA function for events in sideband (control) regions
- For variables that are not defined outside of the signal regions, put in proxies. (sometimes just a zero for the input variable works well if the quantity really isn't defined at all – pick a typical value, not one way off on the edge of its distribution)
- Be sure to use the same MVA function as for analyzing the signal data.

Example: CDF NN single-top
NN validated using events with
zero b-tag

signal region



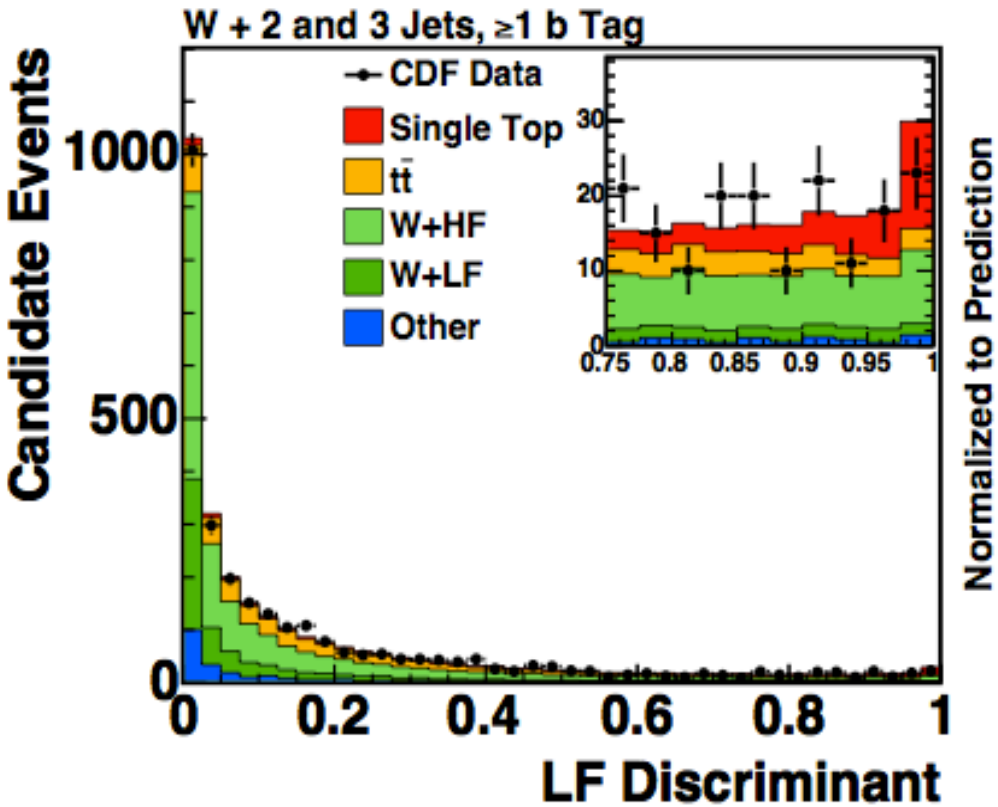


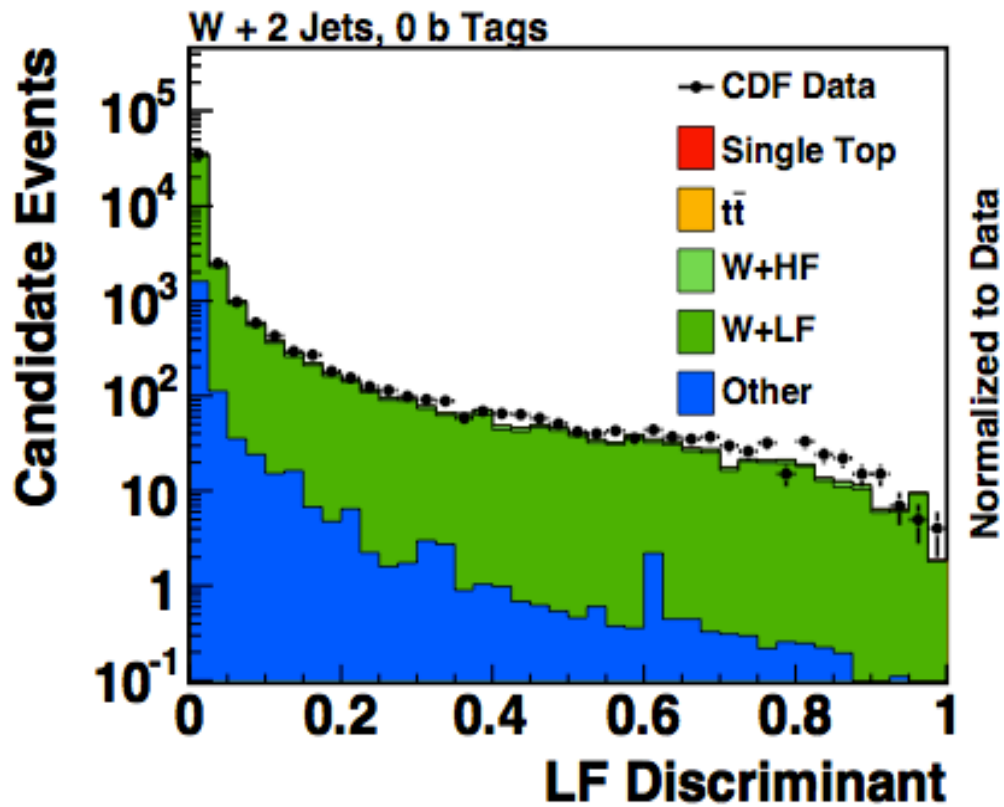
Phys.Rev.D82:112005 (2010)

A Comparison in a Control Sample that is Less than Perfect

CDF's single top Likelihood Function discriminant checked in untagged events

(a) Phys.Rev.D82:112005 (2010) (b)



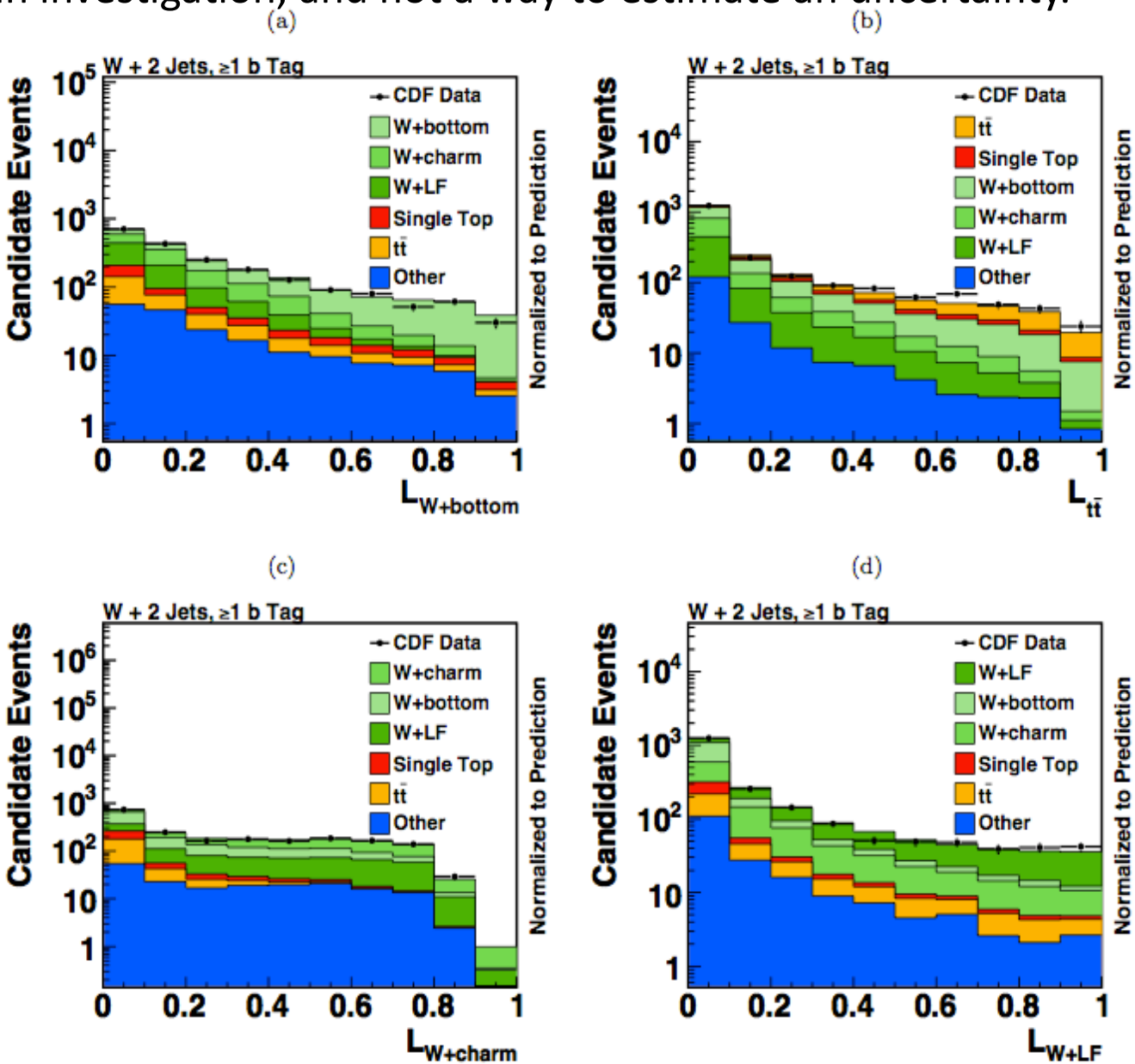


Strategy: Assess a shape systematic covering the difference between data and MC – extrapolate the uncertainty from the control sample to the signal sample.

If the comparison is okay within statistical precision, do not asses an additional uncertainty (even/especially if the precision is weak). Barlow, hep-ex/0207026 (2002).

Another Validation Possibility – Train Discriminants to Separate Each Background

Same input variables as signal LF. LF has the property that the sum of these plus the signal LF is 1.0 for each event. Gives confidence. If the check fails, it's a starting point for an investigation, and not a way to estimate an uncertainty.



Phys.Rev.D82:112005 (2010)

Model Validation with MVA's

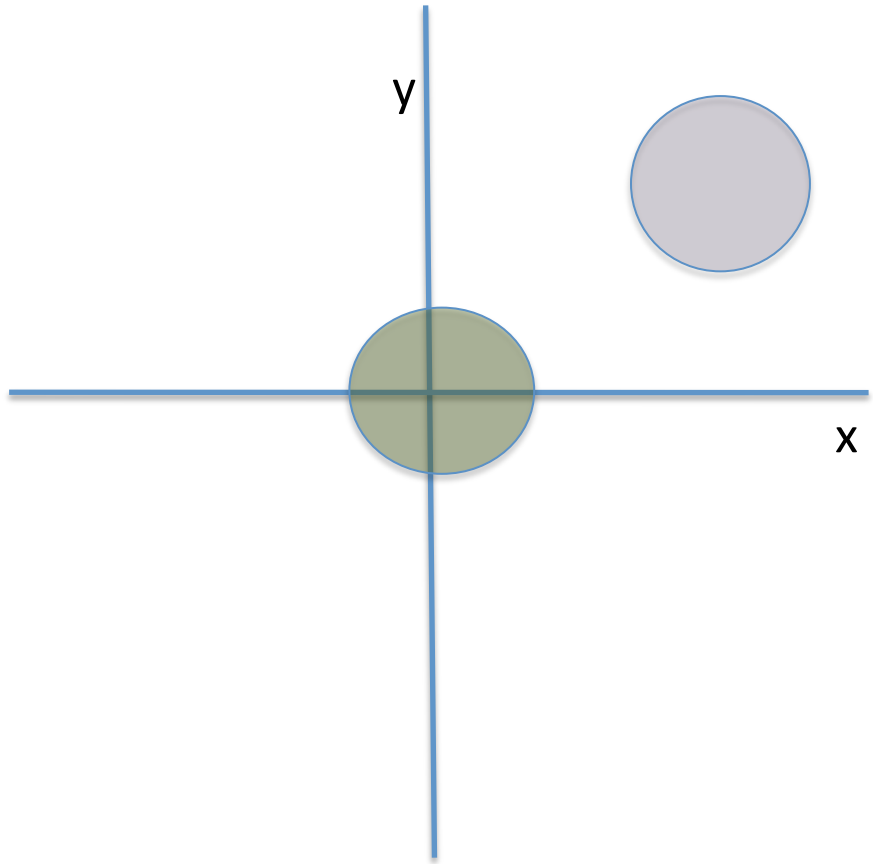
- Even though input distributions can look well modeled, the MVA output could still be mismodeled.
  Possible cause – correlations between one or more variables could be mismodeled
- Checks in subsets of events can also be incomplete.
  A sum of distributions whose shapes are well reproduced by the theory can still be mismodeled if the relative normalizations of the components is mismodeled.

- Can check the correlations between variables pairwise between data and prediction
- Difficult to do if some of the prediction is a one-dimensional extrapolation from control regions (e.g., ABCD methods).

- My favorite: Check the MVA output distribution in bins of the input variables! We care more about the MVA output modeling than the input variable modeling anyway.
- Make sure to use the same normalization scheme as for the entire distribution – do not rescale to each bin's contents.

Ideally, we'd try to find a control sample depleted in signal that has exactly the same kind of background as the signal region (usually this is unavailable).

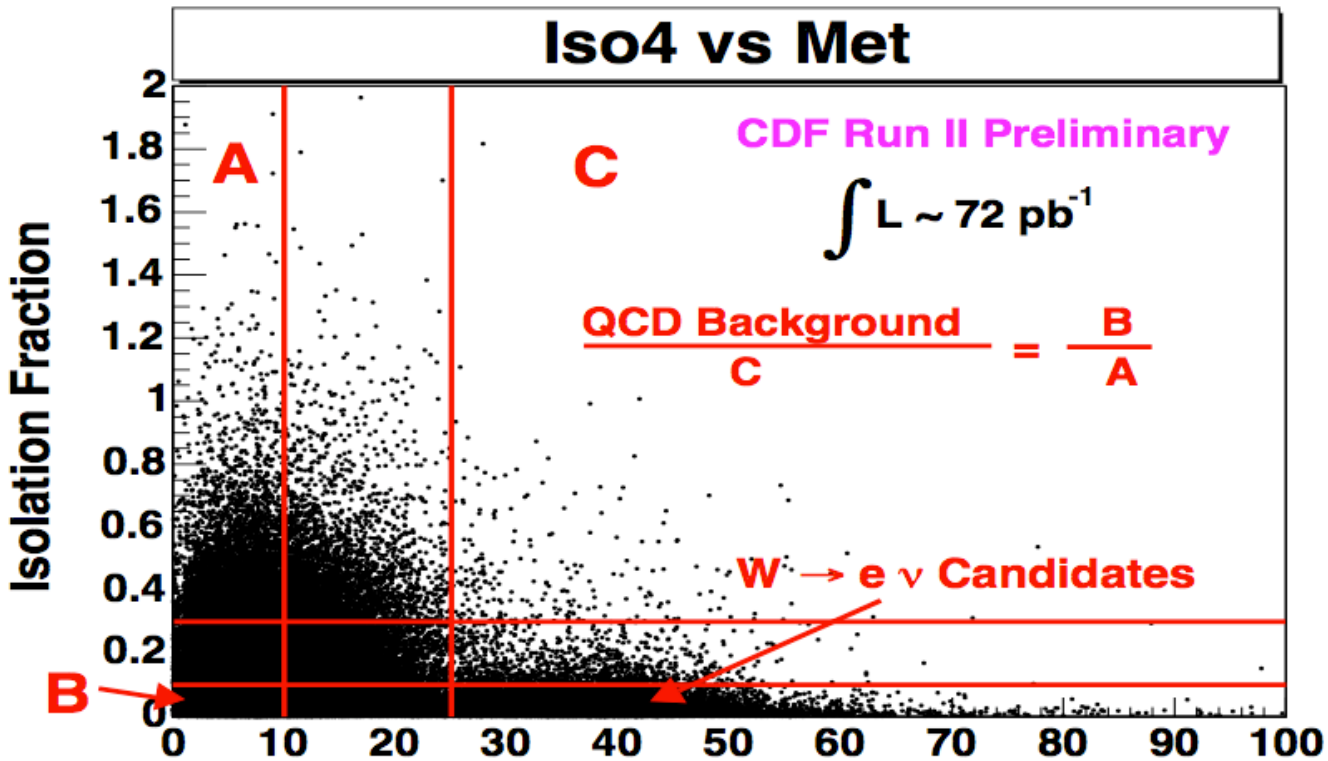
The Sum of Uncorrelated 2D Distributions may be Correlated

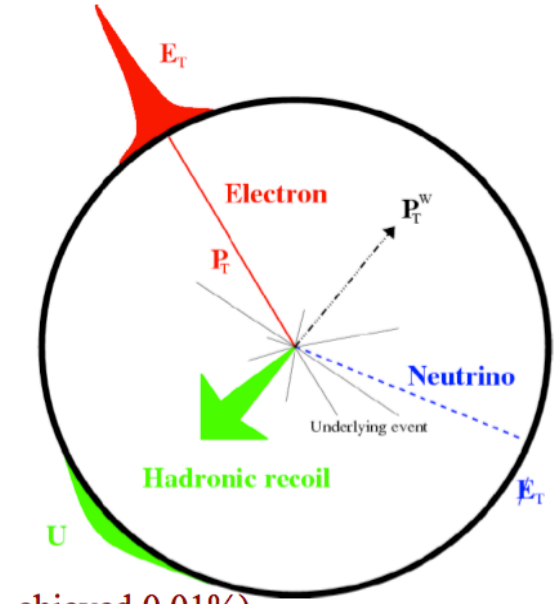


Knowledge of one variable helps identify which sample the event came from and thus helps predict the other variable's value even if the individual samples have no covariance.

"ABCD" Methods

CDF's W Cross Section Measurement





Energy in a cone of radius 0.4 around lepton candidate not including the lepton candidate / Energy of lepton candidate

Want QCD contribution to the "D" region where signal is selected.

ABCD methods are really just on-off methods where τ is measured using data samples

Assumes: MET and ISO are uncorrelated sample by sample
Signal contribution to A,B, and C are small and subtractable

"ABCD" Methods

Advantages

- Purely data based, good if you don't trust the simulation
- Model assumptions are injected by hand and not in a complicated Monte Carlo program (mostly)
- Model assumptions are intuitive

Disadvantages

- The lack of correlation between MET and ISO assumption may be false. e.g., semileptonic B decays produce unisolated leptons and MET from the neutrinos.
- Even a two-component background can be correlated when the contributions aren't by themselves.
- Another way of saying that extrapolations are to be checked/assigned sufficient uncertainty
- Works best when there are many events in regions A,B, and C. Otherwise all the problems of low stats in the "Off" sample in the On/Off problem reappear here. Large numbers of events → Gaussian approximation to uncertainty in background in D
- Requires subtraction of signal from data in regions A, B, and C → introduces model dependence
- Worse, the signal subtraction from the sidebands depends on the signal rate being measured/tested.

→ A small effect if s/b in the sidebands is small

→ You can iterate the measurement and it will converge quickly

Examples of ABCD Methods

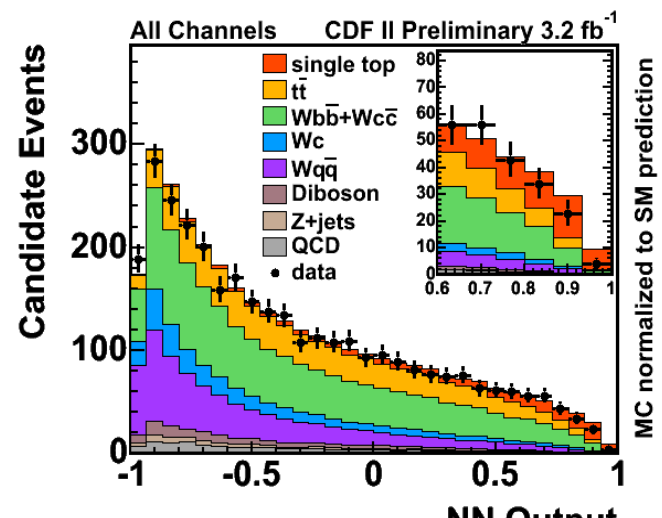
- Sideband calibration of background under a peak. ("what if the background peaks also where the signal peaks?")
- **The on-off problem with τ=A/C**. Very frequently samples A and C are in MC simulations, where we can be sure not to contaminate the background estimations wtih signal. Example: Using the MC to estimate acceptance for a cut for background, to be scaled with a data control sample.

But we pay the price of unknown MC mismodeling.

Uncorrelated variable assumption == assumption that τ is the same in the data and the MC. (check modeling of shape of distribution in the MC)

Equivalent of previous problem: Even if the background shapes are well modeled by the MC, if there are multiple background processes which contribute, they can have different fractional contributions, distorting the total shapes.

- Fitting an MVA shape to the data. Low-score MC = A, High-Score MC = C
  Low-score data = B, High-score Data=D.



An Approximate LEE Correction for Peak Hunting

See E. Gross and O. Vitells, **Eur.Phys.J. C70 (2010) 525-530**.

Approximate formula applies to bump hunts on a smooth background.

Requires a few fully simulated pseudoexperiments with complete p-value calculations over the region of interest. Count up-crossings of a threshold. Extrapolates to higher thresholds assuming large-sample behavior. Specifically, that the LR test statistic has a chisquared distribution.

An interesting feature – specific to bump hunts but may be more general:

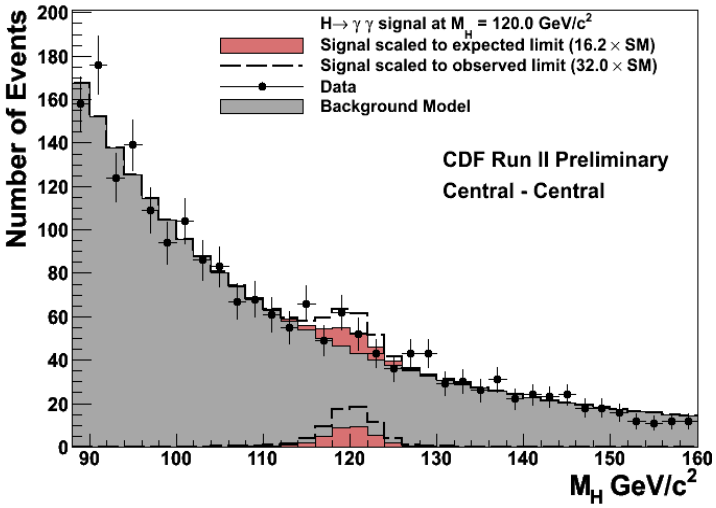
As the expected significance goes up, so does the LEE correction

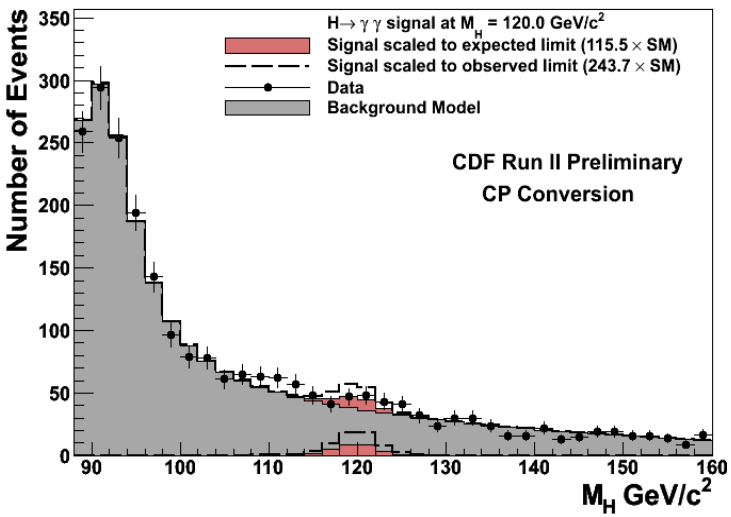
This makes lots of sense: LEE depends on the number of separate models that can be tested. As we collect more data, we can measure the position of the peak more precisely.

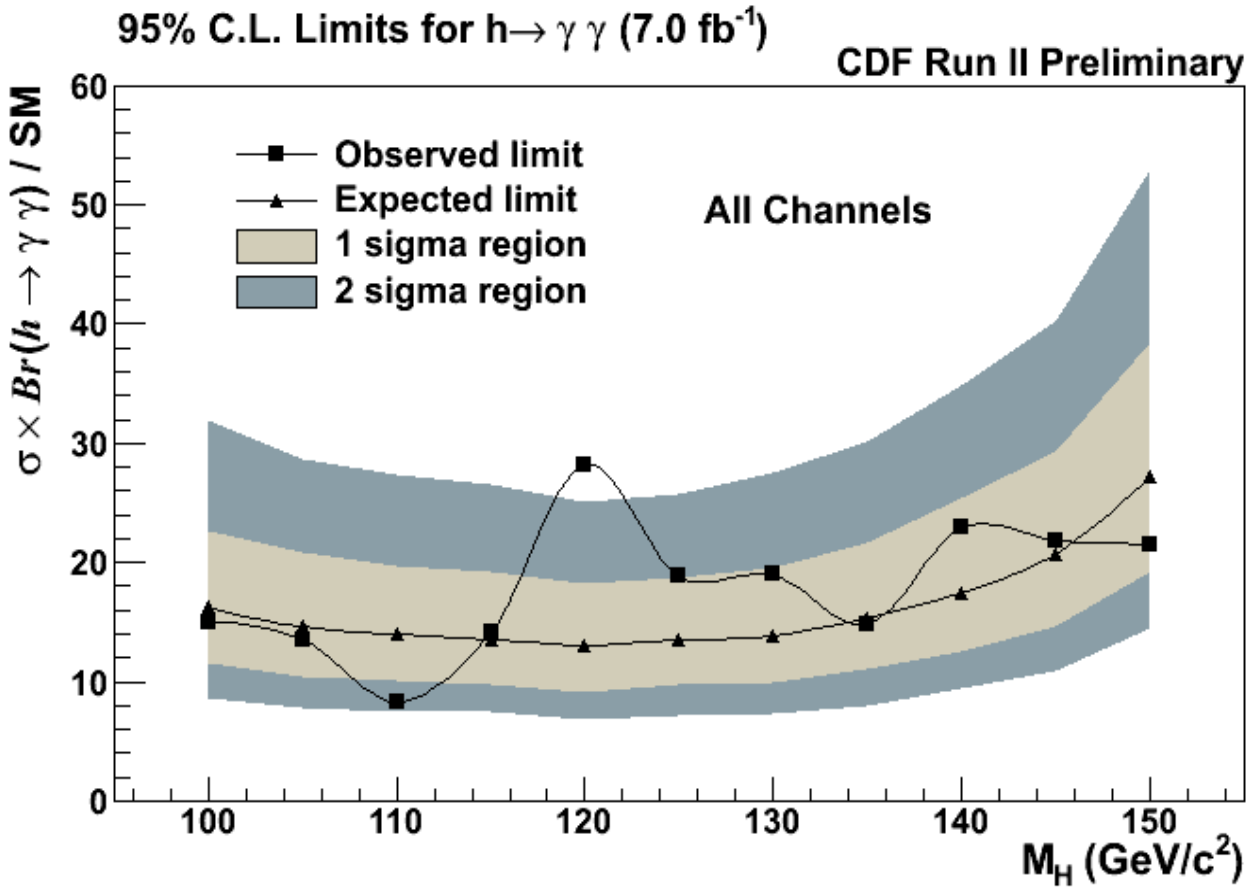
So we can tell more peaks apart from each other, even with the same reconstruction resolution.

But: Combine a poor resolution low s/b search with a high resolution high s/b but very tiny s and very tiny b search – may not get the right answer.

CDF's 2011 H→γγ Search







+2 other channels with smaller excesses

Insufficient sensitivity to a SM Higgs boson. Rate ruled out by other searches (gg→H→WW for example). So we know the bump is a stat fluctuation.

CDF+D0
Higgs Search
Channels
Combined

>300 nuisance
parameters

arXiv:1203.3774

TABLE I: Luminosity, explored mass range and references for the different processes and final states ($\ell = e$ or μ) for the CDF analyses. The generic labels "2×", "3×", and "4×" refer to separations based on lepton categories.

Channel	Luminosity (fb^{-1})	m_H range (GeV/c^2)	Reference
$WH \to \ell\nu b\bar{b}$ 2-jet channels 4×(TT,TL,Tx,LL,Lx)	9.45	100-150	[17]
$WH \to \ell\nu b\bar{b}$ 3-jet channels 3×(TT,TL)	9.45	100-150	[17]
$ZH \to \nu\bar{\nu}b\bar{b}$ (SS,SJ,1S)	9.45	100-150	[18]
$ZH \to \ell^+\ell^- b\bar{b}$ 2-jet channels 2×(TT,TL,Tx,LL)	9.45	100-150	[19]
$ZH \to \ell^+\ell^- b\bar{b}$ 3-jet channels 2×(TT,TL,Tx,LL)	9.45	100-150	[19]
$H \to W^+W^-$ 2×(0 jets,1 jet)+(2 or more jets)+(low-$m_{\ell\ell}$)	9.7	110-200	[20]
$H \to W^+W^-$ (e-τ_{had})+(μ-τ_{had})	9.7	130-200	[21]
$WH \to WW^+W^-$ (same-sign leptons)+(tri-leptons)	9.7	110-200	[20]
$WH \to WW^+W^-$ tri-leptons with 1 τ_{had}	9.7	130-200	[21]
$ZH \to ZW^+W^-$ (tri-leptons with 1 jet)+(tri-leptons with 2 or more jets)	9.7	110-200	[20]
$H \to ZZ$ four leptons	9.7	120-200	[22]
$H + X \to \tau^+\tau^-$ (1 jet)+(2 jets)	8.3	100-150	[23]
$WH \to \ell\nu\tau^+\tau^-/ZH \to \ell^+\ell^-\tau^+\tau^-$ ℓ-τ_{had}-τ_{had}	6.2	100-150	[24]
$WH \to \ell\nu\tau^+\tau^-/ZH \to \ell^+\ell^-\tau^+\tau^-$ (ℓ-ℓ-τ_{had})+(e-μ-τ_{had})	6.2	100-125	[24]
$WH \to \ell\nu\tau^+\tau^-/ZH \to \ell^+\ell^-\tau^+\tau^-$ ℓ-ℓ-ℓ	6.2	100-105	[24]
$ZH \to \ell^+\ell^-\tau^+\tau^-$ four leptons including τ_{had} candidates	6.2	100-115	[24]
$WH + ZH \to jjb\bar{b}$ (SS,SJ)	9.45	100-150	[25]
$H \to \gamma\gamma$ (CC,CP,CC-Conv,PC-Conv)	10.0	100-150	[26]
$t\bar{t}H \to WWb\bar{b}b\bar{b}$ (lepton) (4jet,5jet,≥6jet)×(SSS,SSJ,SJJ,SS,SJ)	9.45	100-150	[27]
$t\bar{t}H \to WWb\bar{b}b\bar{b}$ (no lepton) (low met,high met)×(2 tags,3 or more tags)	5.7	100-150	[28]

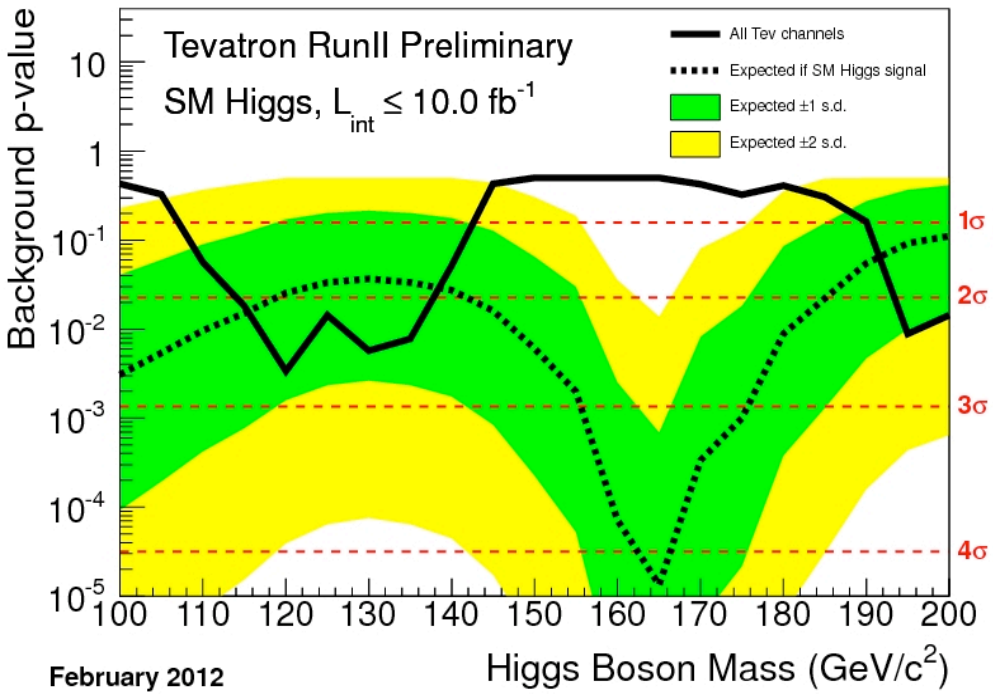
TABLE II: Luminosity, explored mass range and references for the different processes and final states ($\ell = e, \mu$) for the D0 analyses.

Channel	Luminosity (fb^{-1})	m_H range (GeV/c^2)	Reference
$WH \to \ell\nu b\bar{b}$ (TST,LDT,TDT)×(2,3 jet)	9.7	100-150	[29]
$ZH \to \nu\bar{\nu}b\bar{b}$ (MS,TS)	9.5	100-150	[30]
$ZH \to \ell^+\ell^- b\bar{b}$ (TST,TLDT)×(ee,$\mu\mu$,ee_{ICR},$\mu\mu_{trk}$)	9.7	100-150	[31]
$H+X \to \ell^\pm\tau^\mp_{had}jj$	4.3-6.2	105-200	[32]
$VH \to e^\pm\mu^\pm + X$	9.7	115-200	[33]
$H \to W^+W^- \to \ell^\pm\nu\ell^\mp\nu$ (0,1,2+ jet)	8.6-9.7	115-200	[34]
$H \to W^+W^- \to \mu\nu\tau_{had}\nu$	7.3	115-200	[32]
$H \to W^+W^- \to \ell\bar{\nu}jj$	5.4	130-200	[35]
$VH \to \ell\ell\ell + X$	9.7	100-200	[36]
$VH \to \tau\tau\mu + X$	7.0	115-200	[37]
$H \to \gamma\gamma$	9.7	100-150	[38]

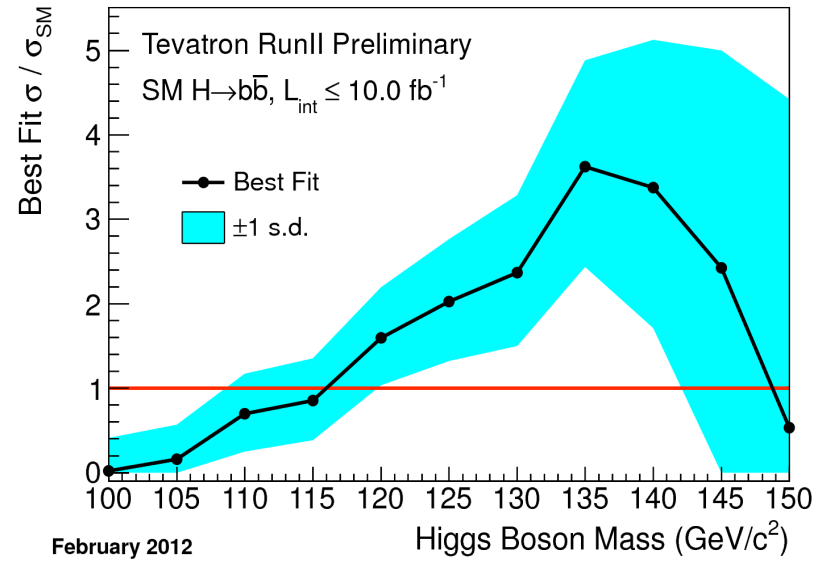
Tevatron Higgs Search LEE

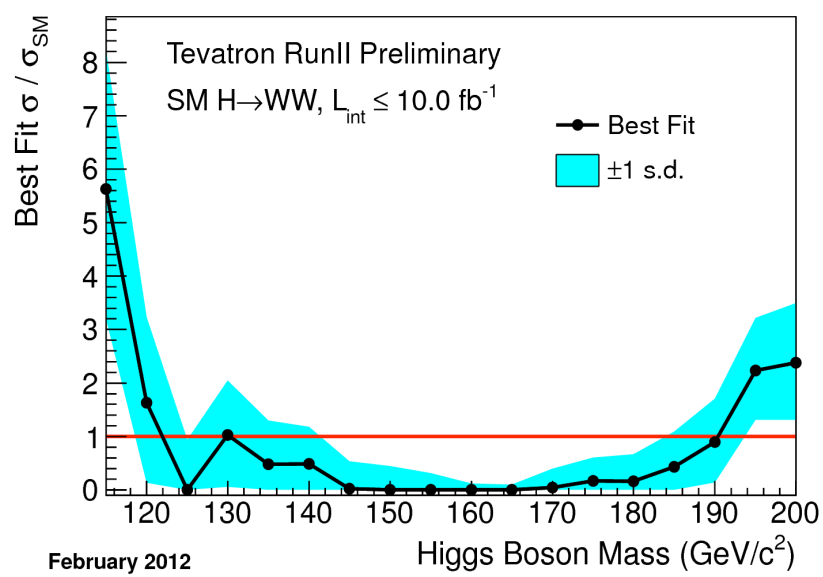
Local p-value vs Higgs boson mass

Bands show expectation assuming a signal is present (at each m_H separately)



Cross Section times Branching Ratio Fits vs. m_H

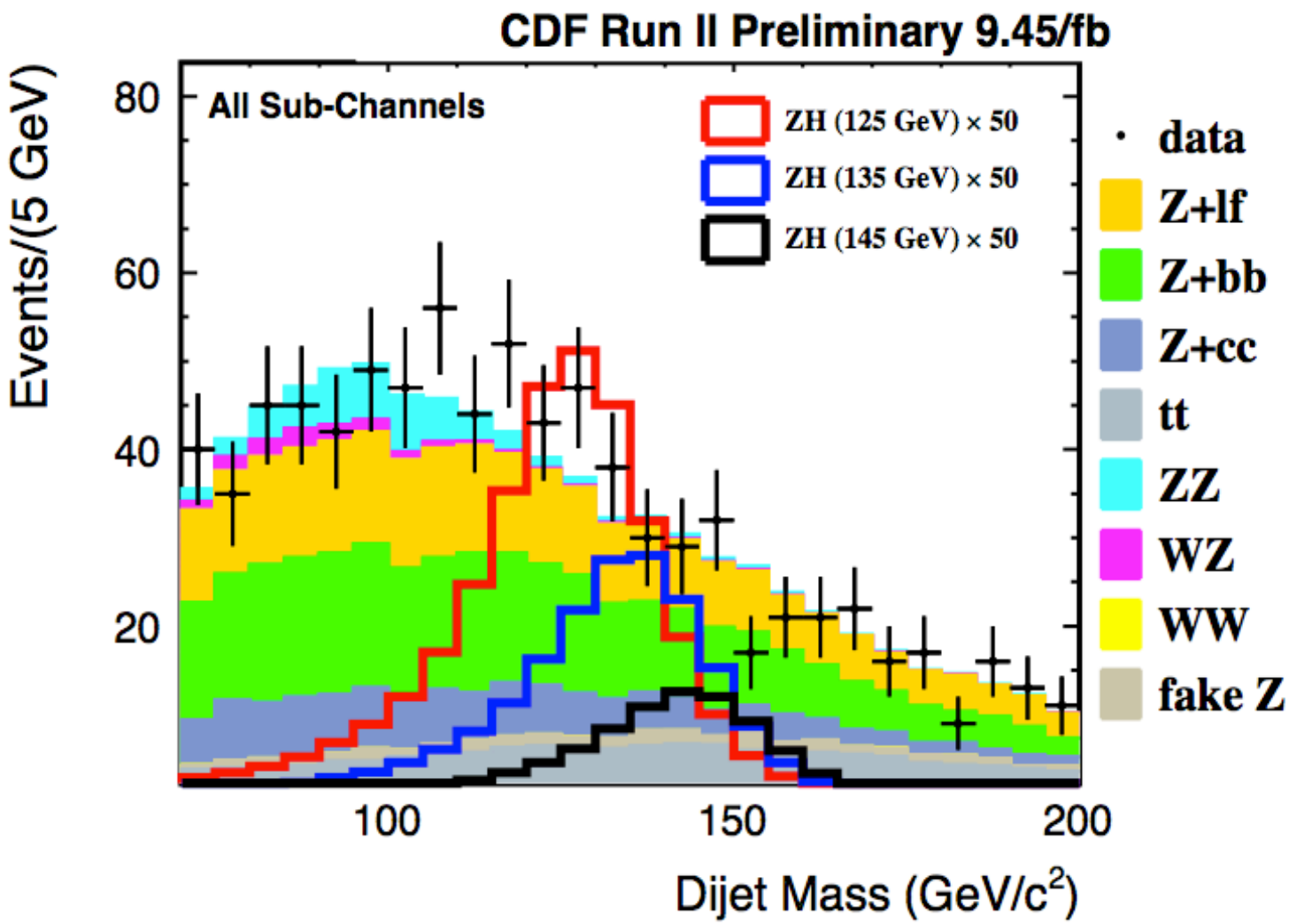




A complication: Most search channels train their MVA's separately at each m_H to optimize sensitivity

Signal, Background, and Data in the ZH→llbb Search

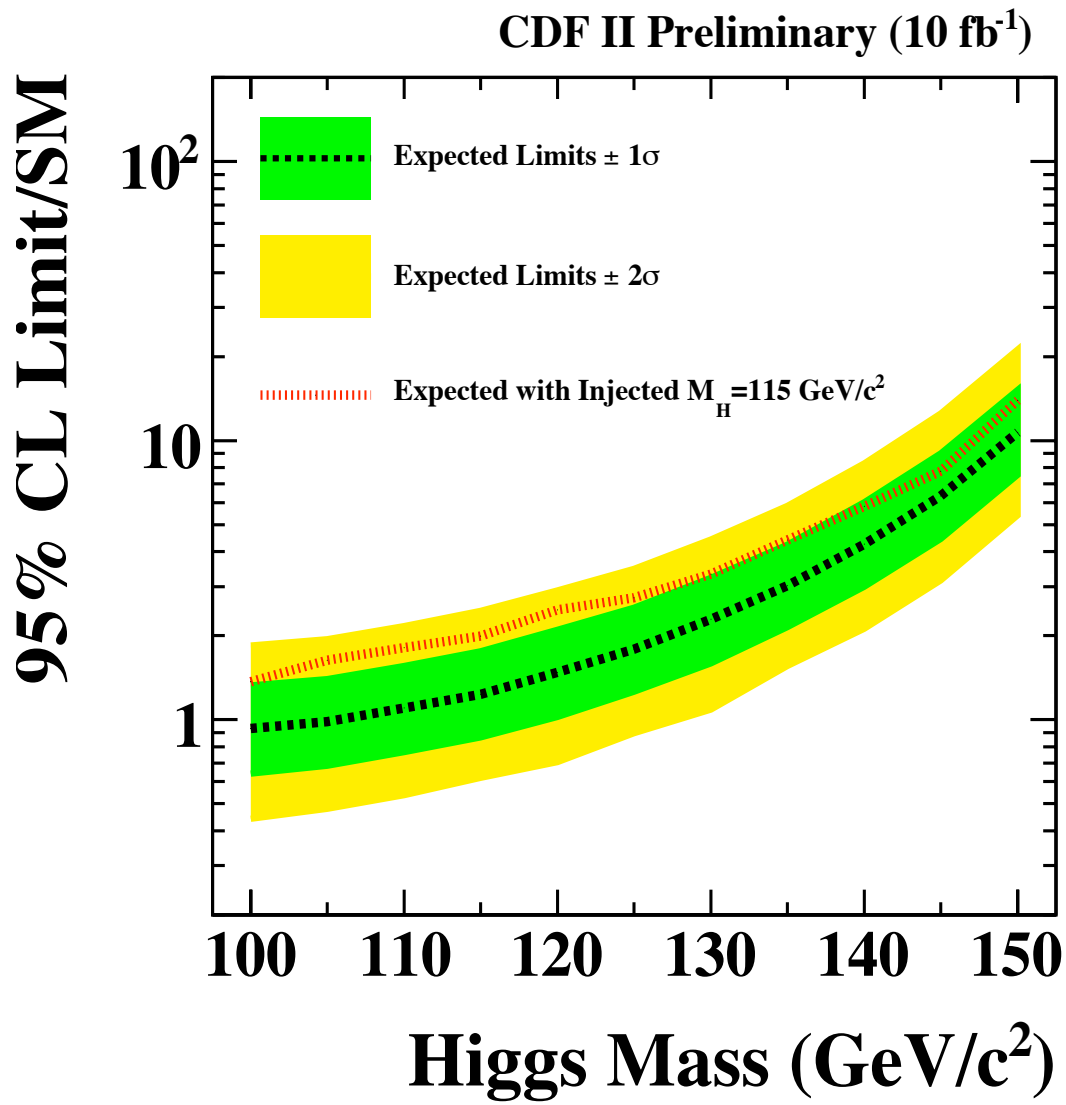
Reconstructed m_{jj} distribution



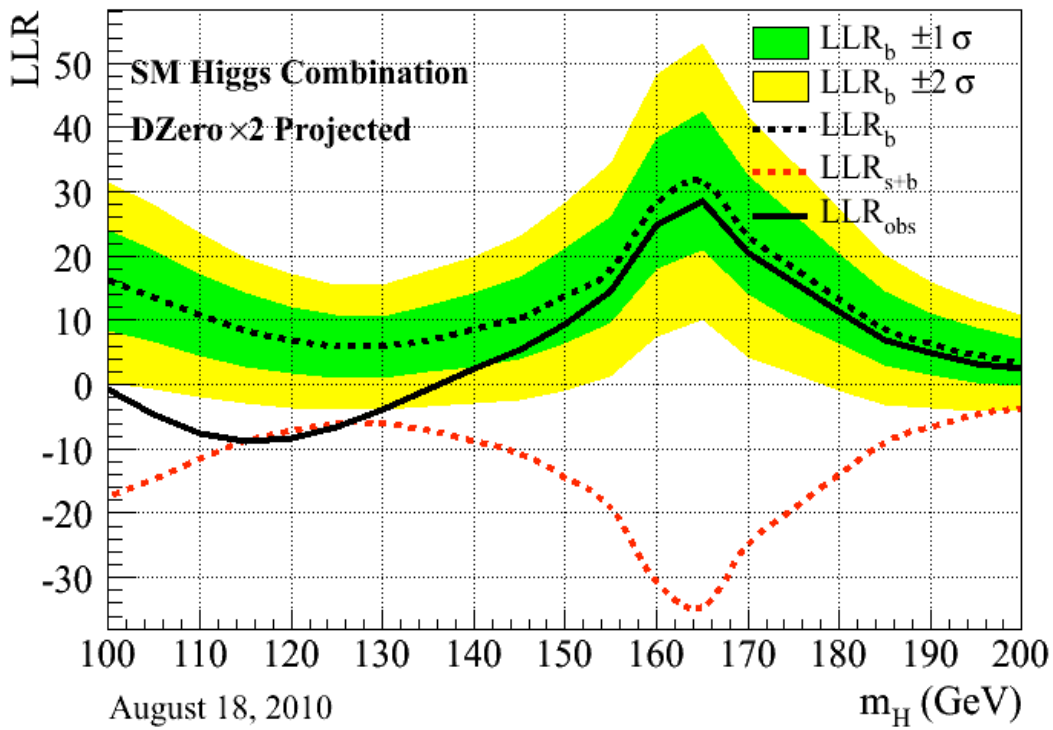
You can tell m_H is on the high side of the range only by what's missing and not what's there!

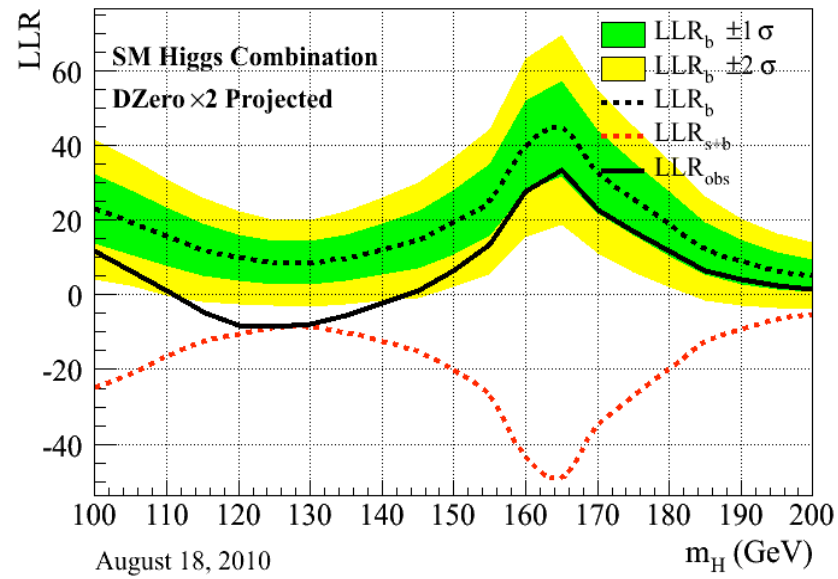
CDF's WH channel expectation (x3 luminosity to simulate the presence of other channels: llbb, METbb)

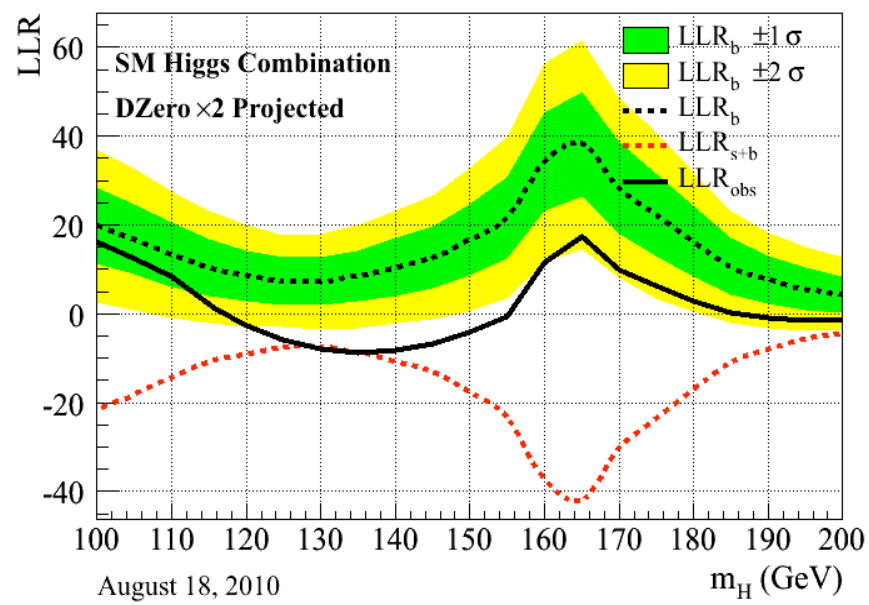
With a 115 GeV signal injected



Stirring it All Together – D0's LLR Test







Assuming observed and expected +3 sigma excess, and median outcome. Resolution from $-2\Delta LLR = \Delta\chi^2=1$

Resolution at 115 GeV: ±5 GeV
Resolution at 135 GeV: ~±10 GeV

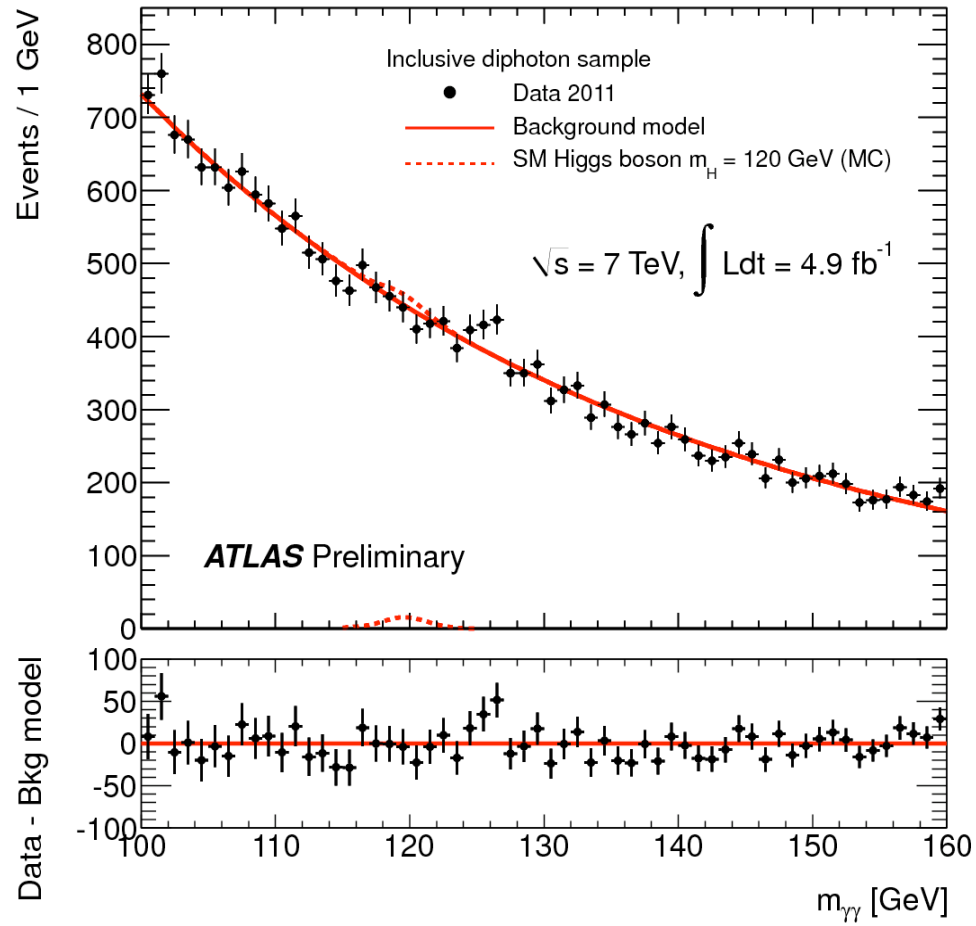
An interesting Bias Bill Murray Showed at The Next Stretch of the Higgs Magnificent Mile Conference

Seek a bump on a smooth background
Example: LHC (or Tevatron) H→γγ search.

Allow m_H to float and pick the m_H that maximizes the fitted cross section.

The fitted cross section will be biased upwards and the position resolution of "lucky" outcomes will be worse than unlucky ones even if a signal is truly present.

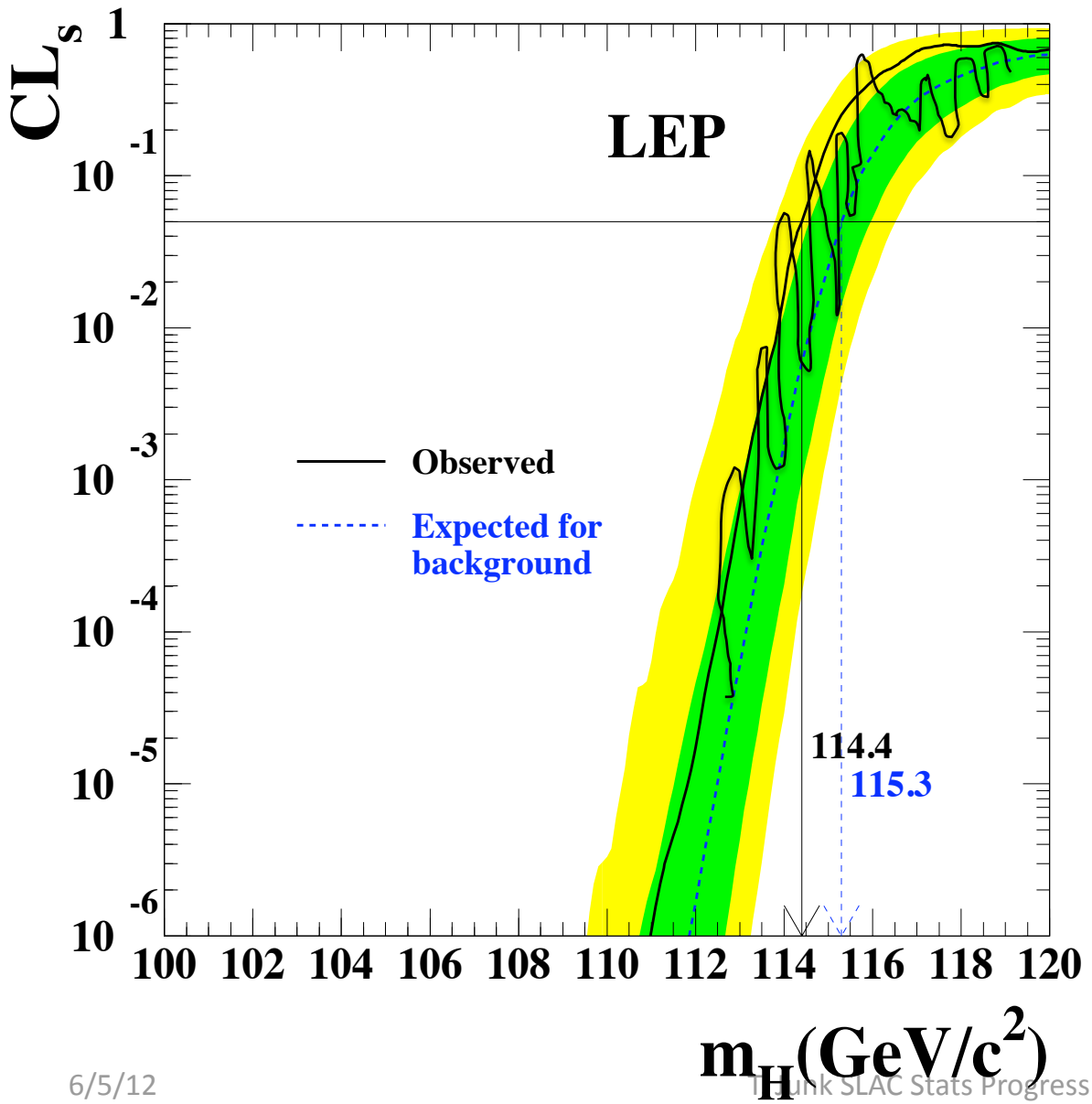
Why? A true bump can coalesce with a fluctuation either to the left or to the right of the bump (two chances to fluctuate upwards).



Effect can be substantial! Calibrate with simulated experimental outcomes (FC).

https://twindico.hep.anl.gov/indico/conferenceOtherViews.py?view=standard&confId=856

LEE for Limits?



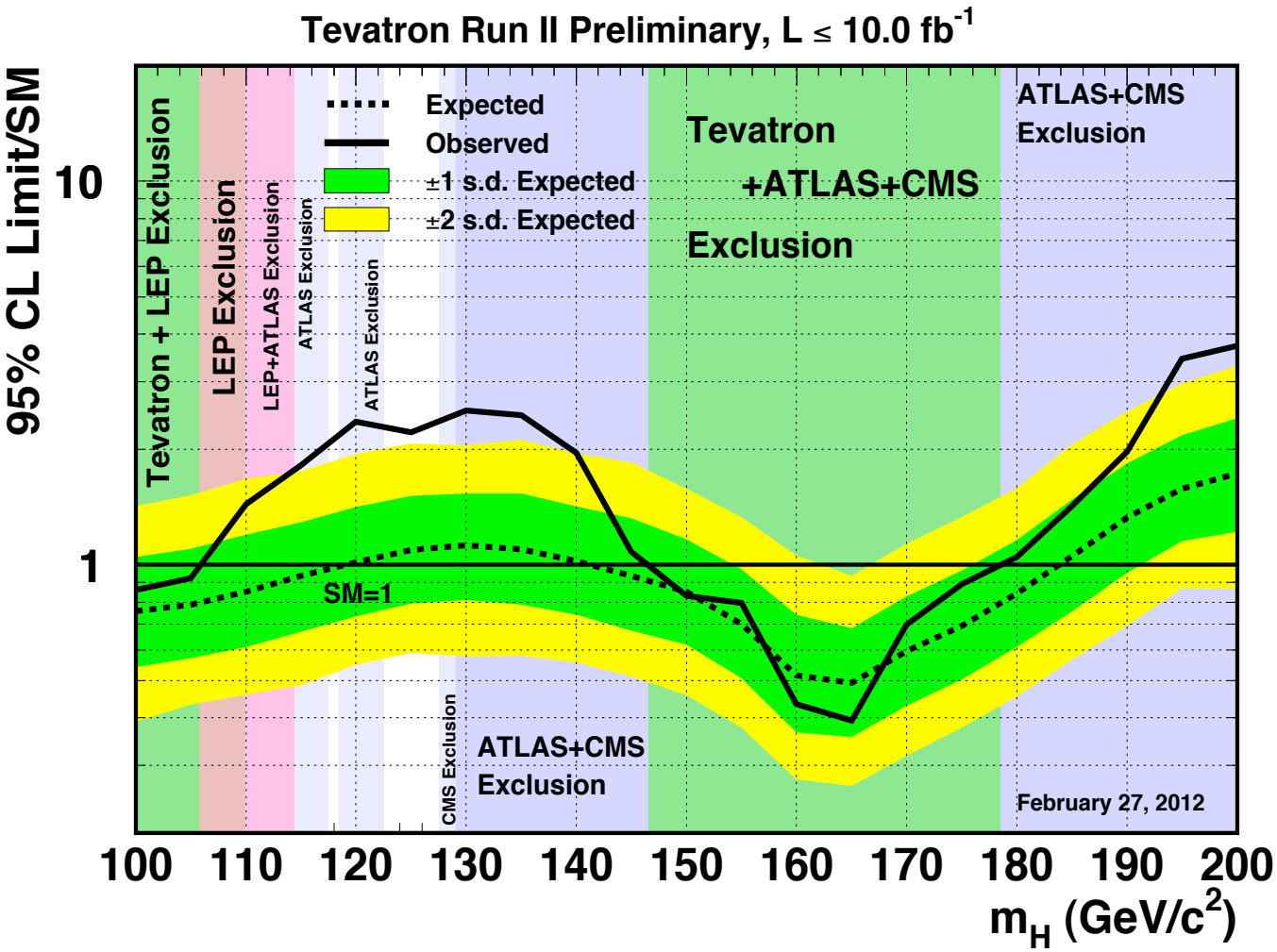
No, but there is the opposite effect. We take the most conservative mass exclusion. If the CLs curve crosses several times we quote the smallest (LEP).

Hard to say what the median expected limit is – the place where the median CLs crosses the line is higher than the median lowest limit.

LHC and Tevatron experiments quote multiple disjoint m_H limits.

No LEE: justification – each test at each m_H is an independent search with its own error rate, assuming a particle is truly there at each mass one at a time.

LEE for Limits?



Multiple experiments searching for the same particle.

Multiple chances to falsely exclude a particle that's actually there.

Very easy to take the union of excluded regions, but this does not have 95% coverage.

The best thing to do is to combine for a single interpretation.

But the limits are of secondary importance here.

Where is "Elsewhere?"

A collider collaboration is typically very large; >1000 Ph.D. students. ATLAS+CMS is another factor of two. (Four LEP collaborations, Two Tevatron collaborations).

Many ongoing analyses for new physics. The chance of seeing a bump somewhere is large. What is the LEE?

Do we have to correct our previously published p-values for a larger LEE when we add new analyses to our portfolio?

How about the physicist who goes to the library and hand-picks all the largest excesses? What is LEE then?

"Consensus" at the Banff 2010 Statistics Workshop: LEE should correct only for those models that are tested within a single published analysis. Usually one paper covers one analysis, but review papers summarizing many analyses do not have to put in additional correction factors.

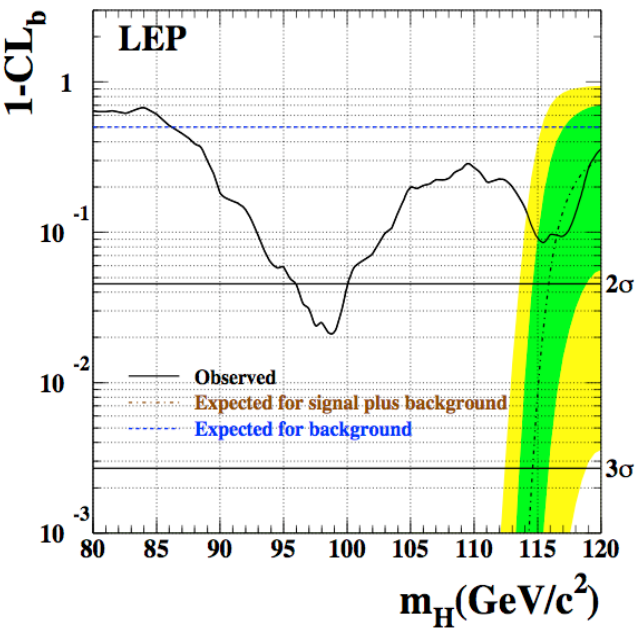
For the Winter 2012 Higgs search analyses, we had several LEE's computed, depending on the mass range defined to be elsewhere.

Caveat lector.

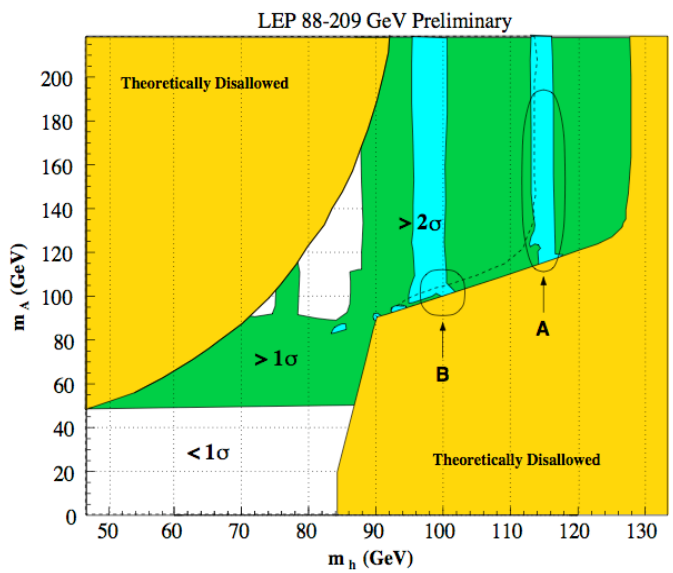
Where is "Elsewhere?"

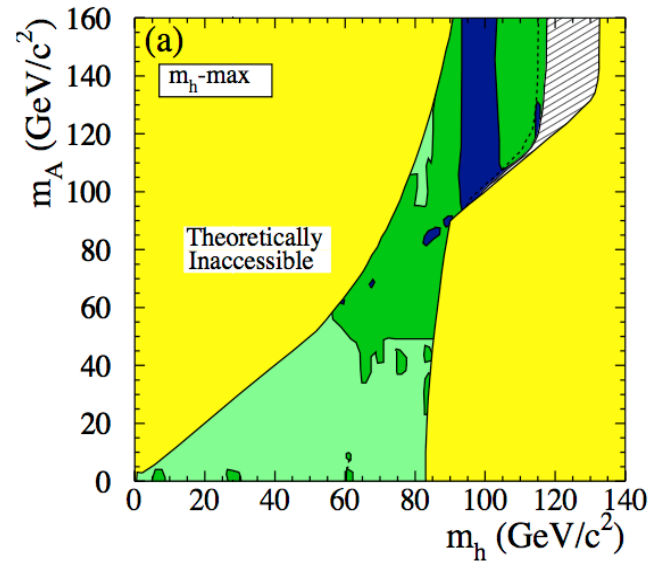
LEE is often hard enough to evaluate. Right way to do it – compute p-value of p-values simulate experiment assuming zero signal many times and for each simulated outcome find the model with the smallest p-value.
Multidimensional models are harder, and LEE is worse.

Kane, Wang, Nelson, Wang, Phys. Rev. D **71**, 035006 (2005)



ALEPH, DELPHI, L3, OPAL, and the LHWG
Phys.Lett. B565 (2003) 61-75





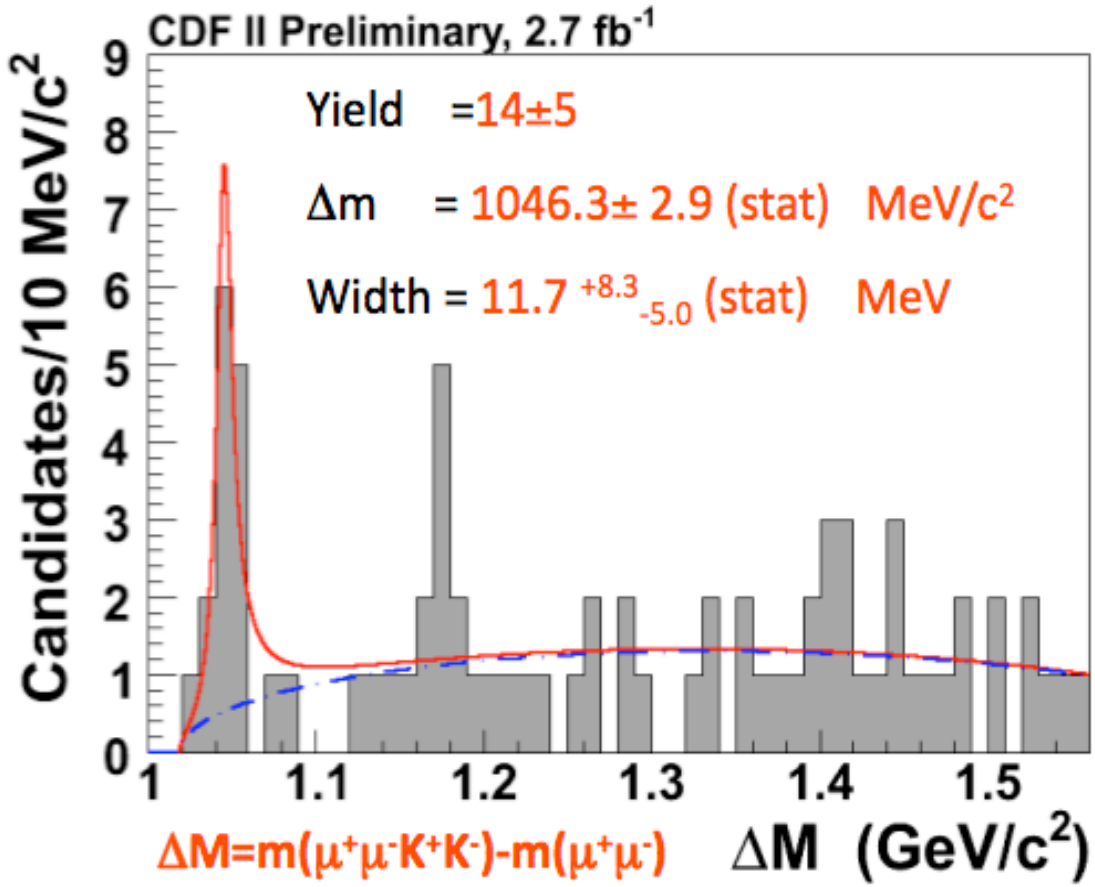
ALELPH, DELPHI, L3, OPAL, and the LHWG
Eur.Phys.J. C47 (2006) 547-587

Two excesses seen; proposed models explain both with two Higgs bosons. Combined local significance is greater, but LEE now is much larger (and unevaluated). Published plot grays out region beyond experimental sensitivity.

Search for structures in $J/\psi\phi$ mass--Data

- We model the Signal (S) and Background (B) as:

S: S-wave relativistic Breit-Wigner B: Three-body decay Phase Space



Convoluted with resolution (1.7 MeV)

Slide from K. Yi, Fermilab Joint Experimental/Theoretical Physics Seminar, March 17, 2009

How many bumps do you see?

$\sqrt{(-2\log(L_{max}/L_0))}$=5.3, need Toy MC to determine significance for low statistics

What if we don't have a signal model, and we're just on a hunting expedition? What's LEE now?

Choosing a Region of Interest

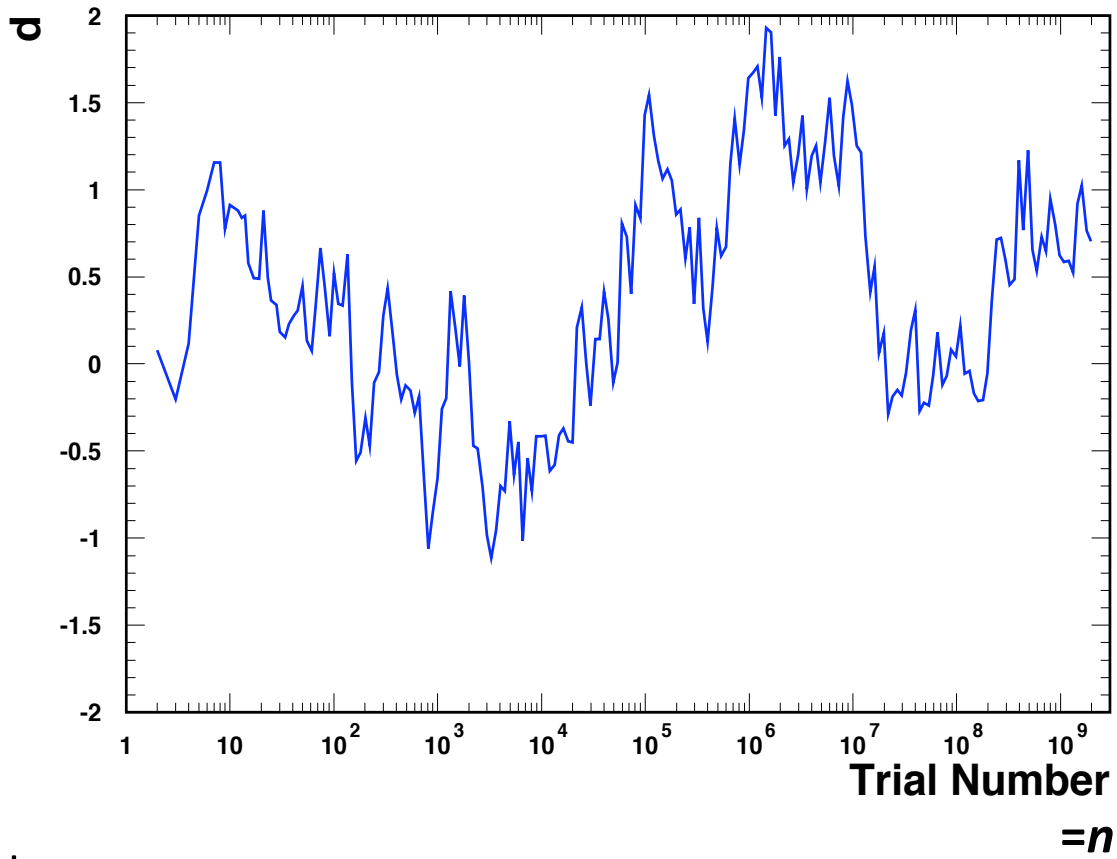
- I do not have a foolproof prescription for this, just some thoughts.
- Analyses are designed to optimize sensitivity, but LEE dilutes sensitivity. There is a penalty for looking for many independently testable models. Can we optimize this?
- But you should always do a search anyway! If you expect to be able to test a model, you should.
- Testing previously excluded models? We do this anyway, just in case some new physics shows up in a way that evaded the previous test.
- There is no such thing as a model-independent search. Merely building the LHC or the Tevatron means we had something in mind. And the SM (or just our implementation of it) is wrong, but possibly not in a way that is both interesting and testable.

Look ElseWHEN

Running averages converge on correct answer, but the deviations in units of the expected uncertainty have a random walk in the logarithm of the number of trials

$$d_n = \frac{\sum_{k=1}^{n} r_k / n}{1/\sqrt{n}}$$

The r_k are IID numbers drawn from a unit Gaussian.



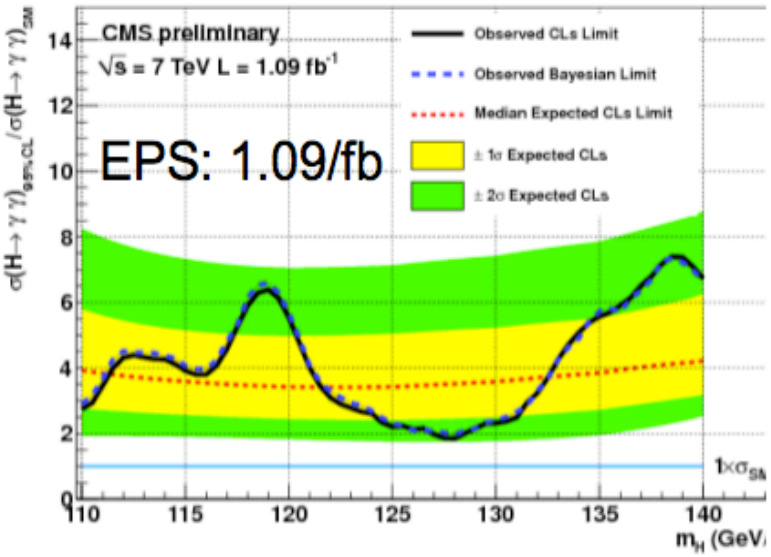
It's possible to cherry-pick a dataset with a maximum deviation. "Sampling to a foregone conclusion"

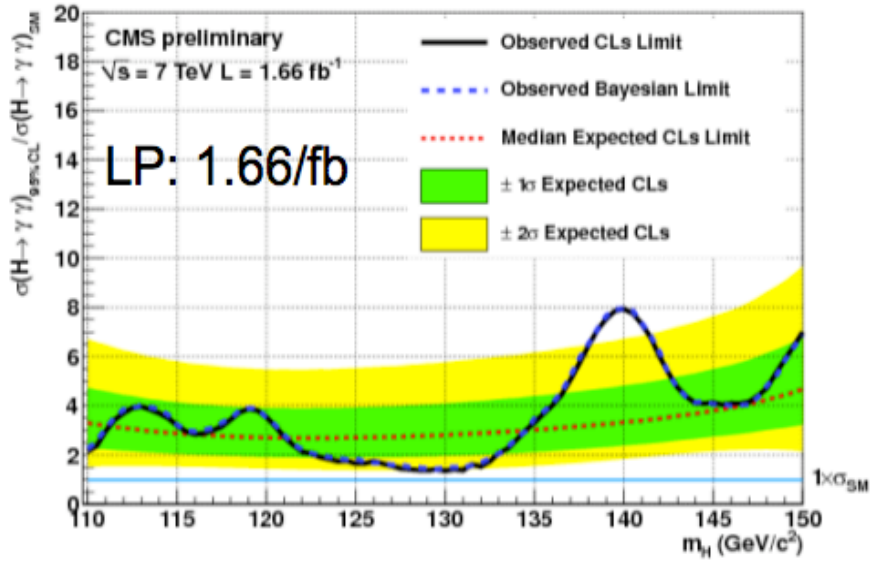
Stopping Rule: In HEP, we (almost always!) take data until our money is gone. We produce results for the major conferences along the way. Some will coincidentally stop when the fluctuations are biggest. We take the most recent/largest data sample result and ignore
(or should!) results performed on smaller data sets. p-values still distributed uniformly from 0 to 1. A recipe for generating "effects that go away"

Look ElseWHEN

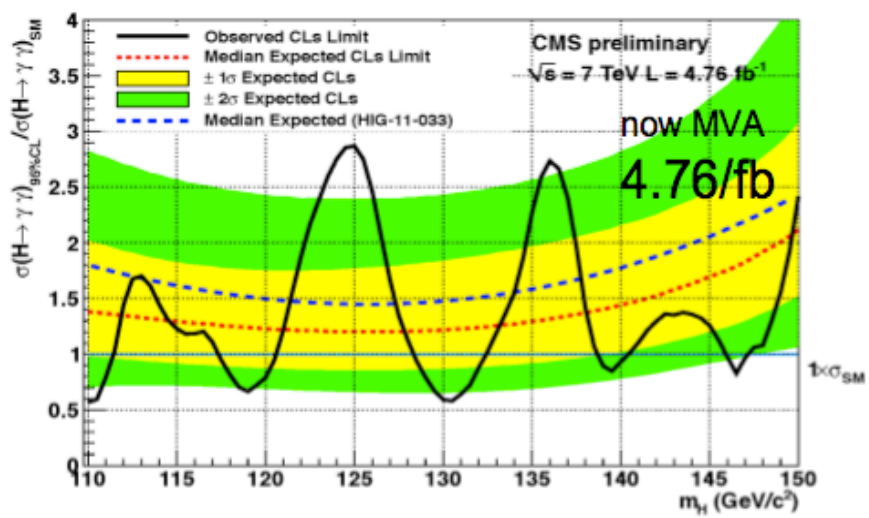
CMS History: $H \rightarrow \gamma\gamma$

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



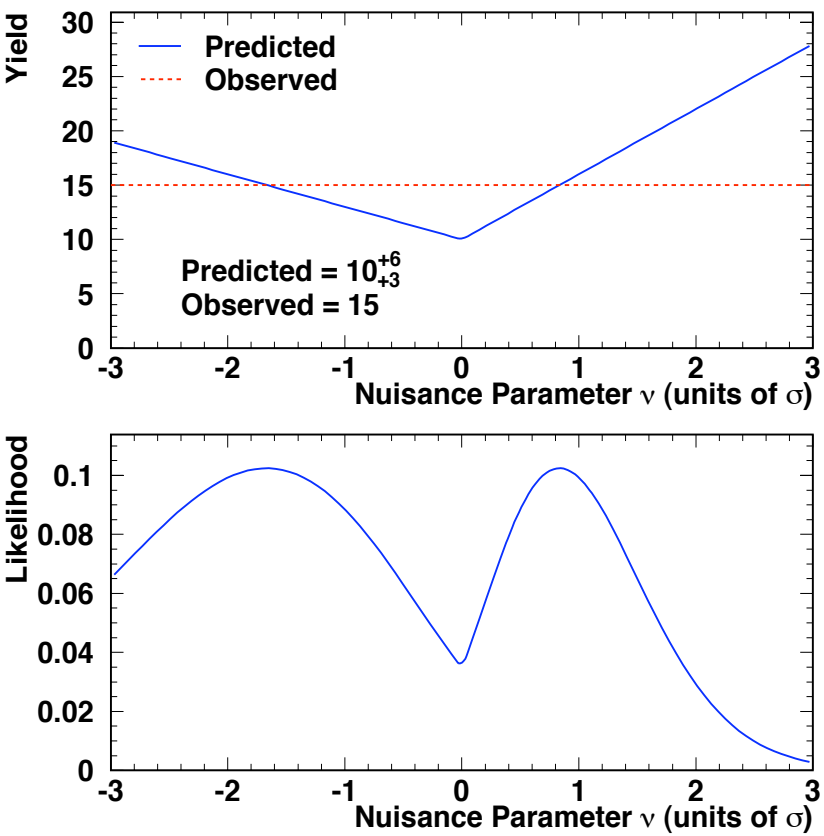


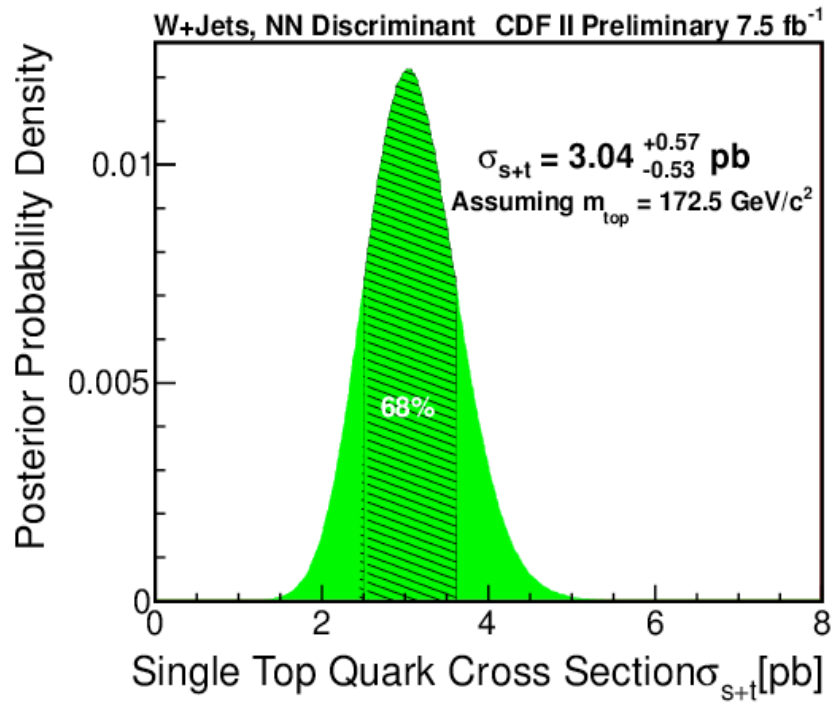
- EPS (1.09/fb) LP (1.66/fb) Dec 19 (4.76/fb)
- 'peaks' come and go
- we are getting into interesting territory, and peaks can also stay



https://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=162621

Parameter Estimation – Marginalize or Profile?





If Pred = 10^{-6}_{-3}, and obs=15, then the likelihood would have one maximum, but it would have a corner. MINUIT may quote inappropriate uncertainties as the second derivative isn't well defined.

The corner can be smoothed out – See
R. Barlow, http://arxiv.org/abs/physics/0406120,
http://arxiv.org/abs/physics/0401042
http://arxiv.org/abs/physics/0306138

But I know of no way
to get rid of the double-peak
(nor should there be a way --
it can be a real effect. See the LEP2 TGC measurements

Analysis Optimization in Isolation or in Combination?

Typical situation:

A measurement has a statistical and a systematic uncertainty, where the statistical uncertainty includes "good" systematics that are constrained by the data, and the "bad" ones never get better constrained no matter how much data are collected.

We sometimes have a choice of how to analyze marked Poisson data.

1) aggressive reconstruction making assumptions about particle distributions – more statistical power per event at the cost of introducing systematic uncertainty

2) more model-independent analysis with fewer assumptions – less statistical power per event but better control over systematics.

→ Combination with other measurements (from other data runs or other collaborations) is like collecting more data. Method 1 hits the systematic limit and loses weight in the combination even though it may be the most powerful method by itself.

More general: With little data, we are more dependent on our assumptions, with more data we can relax the assumptions and constrain our models.

Recommendation: For combinations, optimize for the large luminosity case.

Correlations among Uncertainties – When is it Conservative, when not?

- Within a channel – contributions that add together: including correlations usually weakens the sensitivity (always: sensitivity is expected)
- Between channels – accounting for correlations is not conservative
  One channel's observed data becomes another "off" sample for another's.
  Have to trust all the τ factors, and even offsets from central predictions in order to put in these correlations.

- Overestimating the impacts of systematic uncertainty on a prediction is not conservative if a correlation is taken into account. Can result in underestimated systematic error on a combined result.

Example (systematic uncertainty 1 is 100% correlated, syst uncertainty 2 is 100% correlated)

Measurement 1: $m_1 = 5 \pm 1$ (syst1) ± 1 (syst2) Combine with BLUE: $m_{best}=2m_1-m_2$
Measurement 2: $m_2 = 5 \pm 1$ (syst1) ± 2 (syst2) → $m_{best}=5 \pm 1$ (syst1) ± 0 (syst2)

Here accounting for correlation and an overestimated systematic uncertainty results in an aggressive result.

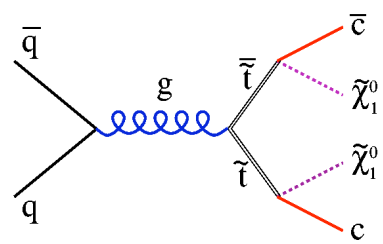
Extra Slides

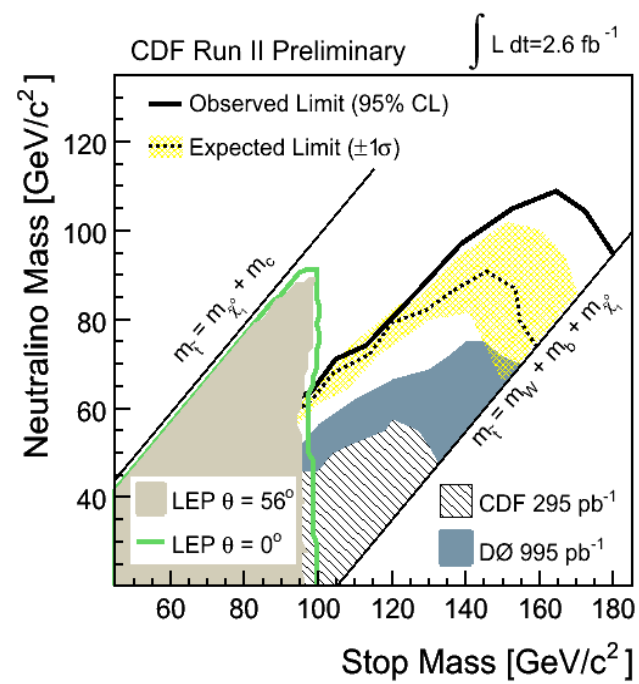
Where is "Elsewhere?"

- Most searches for new physics have a "region of interest"
  - Definition is a choice of the analyzer/collaboration
  - Often bounded below by previous searches, bounded above by kinematic reach of the accelerator/detector
  - Limits the amount of work involved in preparing an analysis. Sometimes a 2D search involves lots of training of MVA's and checking sidebands and validation of inputs and outputs

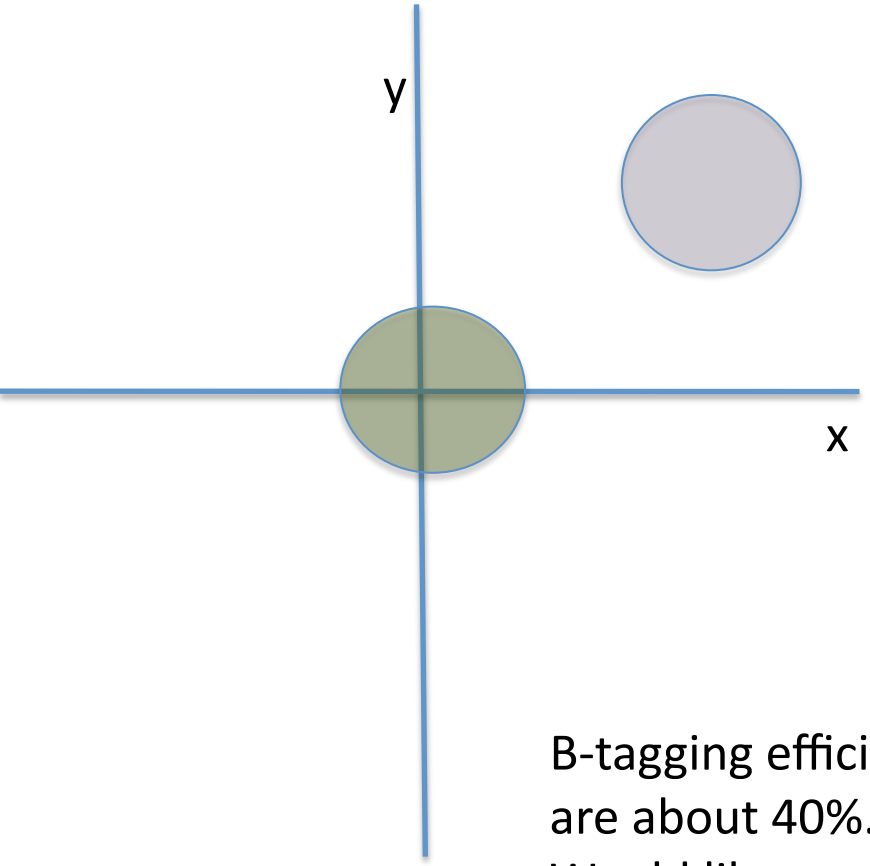
Example: A search for pair-produced stop quarks which decay to c+Neutralino

If $M_{stop}>m_W+m_b+m_{neutralino}$ then another analysis takes over.



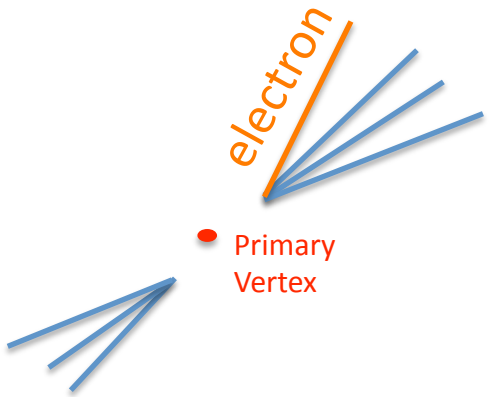


An Example: Double-Tag Methods



Dijet events at LEP1/SLD

Z→u,ubar
d,dbar
s,sbar
b,bbar
leptons
neutrinos



A double-vertex-B-tagged event with a semileptonic decay

B-tagging efficiencies (efficiency of finding the displaced vertex) are about 40%. We do not trust MC modeling of the b-tag efficiency. Would like to measure the B-tag efficiency and the Br(Z→b,bbar) branching fraction together in the same data. Count events with 0, 1, and 2 vertex tags. Enough information to solve for the Br and the efficiency.

x=b-tag of jet 1, y=b-tag of jet 2. Assume uncorrelated probabilities for tagging the jets. But the flavor of the jets is correlated! It is this flavor correlation that allows us to extract Br and Tag eff.

On-Off Measurements – Averaged or Combined

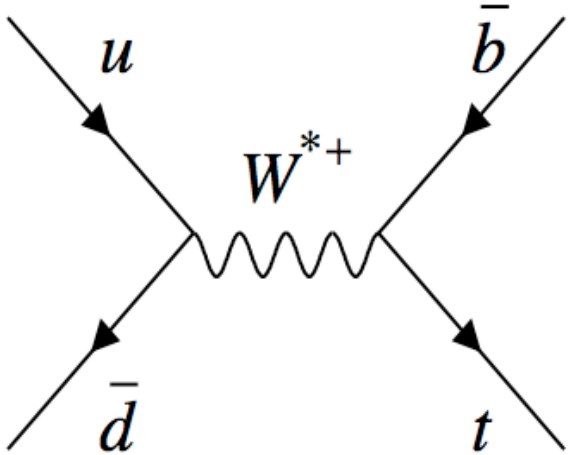
- One global on-off measurement vs. breaking the data into subsamples
- Assume the "off" data are collected along with the "on" data (control sample on the other side of a cut for example)

- Global on-off measurement allows each data subsample's off measurements to help measure each other data subsample's on sample's backgrounds. Assumption which may be false: you are allowed to do this. If the detector or accelerator changed part way through the run, then you may need to break the samples up.

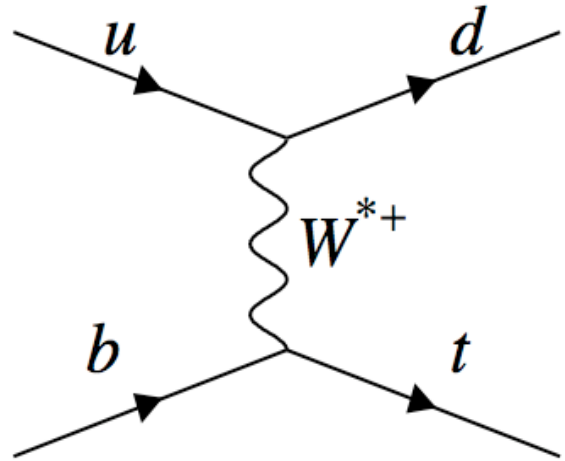
- Breaking them apart allows only each subsample's off measurements to calibrate the background in the corresponding on samples.

- Same for ABCD methods – averaging subsamples

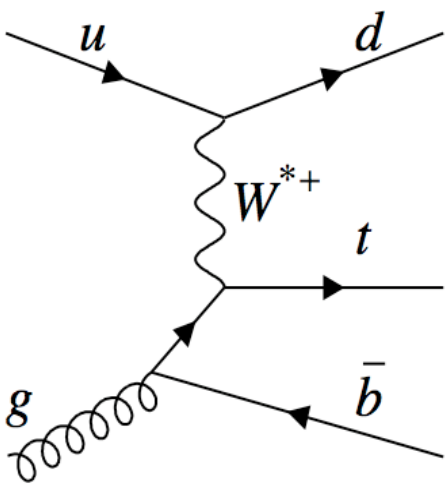
Single Top Production Mechanisms

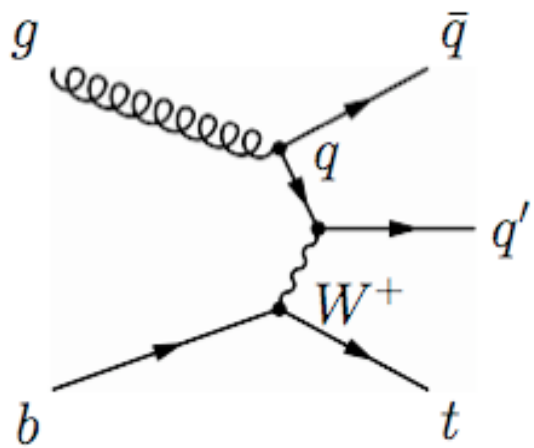


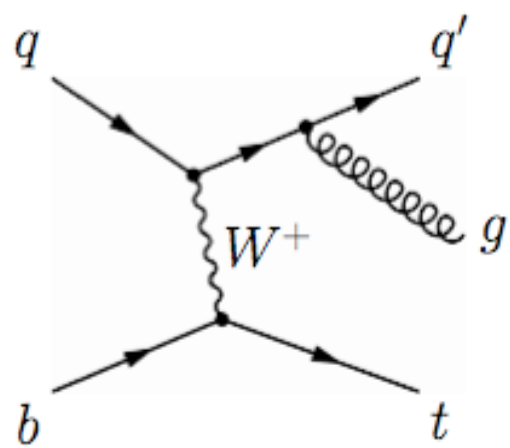
"*s*-Channel"



"*t*-Channel"



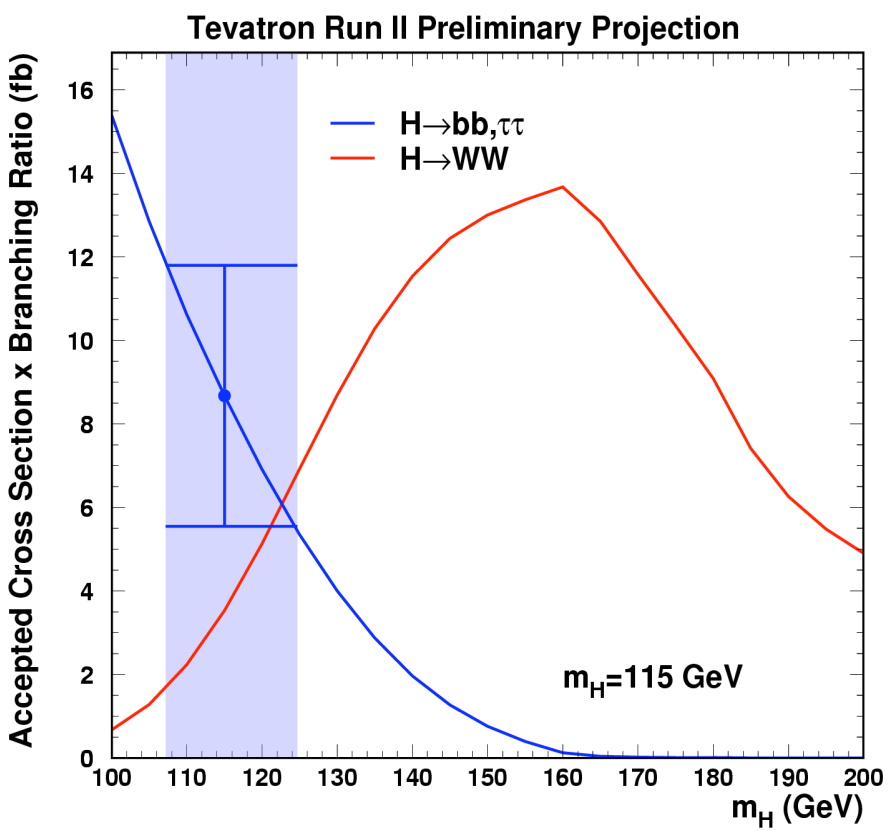


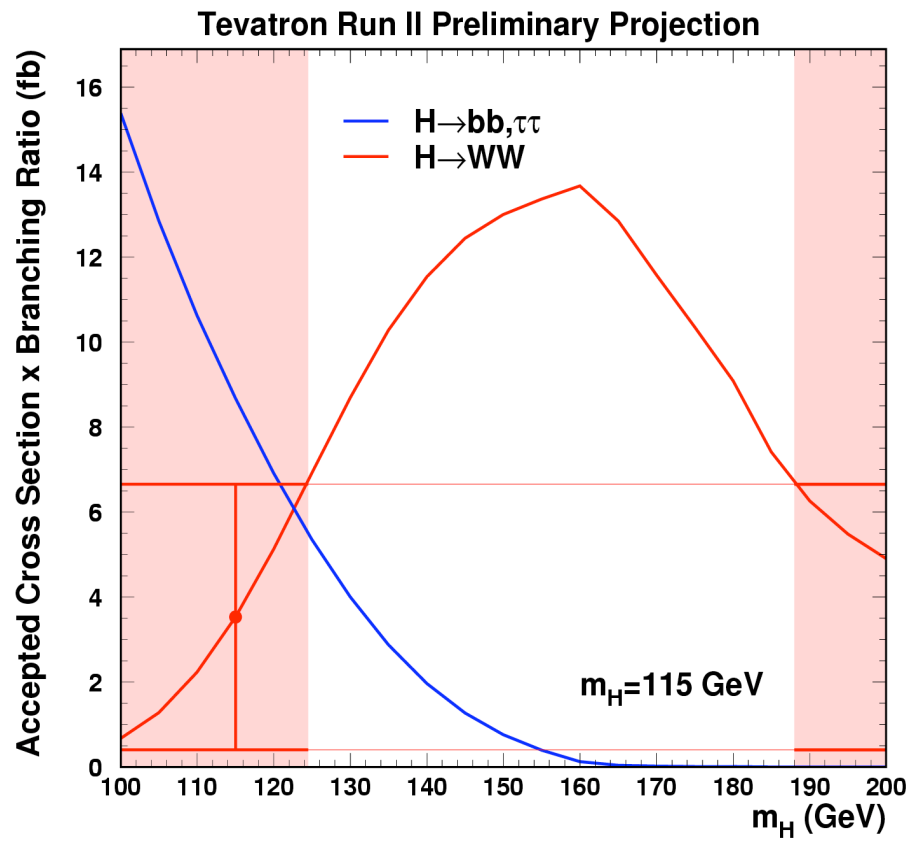


"NLO Contributions to *t*-Channel Production"

Leveraging our Rate Measurements to Measure the Higgs Boson Mass

Assuming SM cross sections and branching fractions, measured rates are strong functions of m_H. Example at m_H=115 GeV, assuming +3 sigma excess, and a median outcome in both the bb,ττ channels and the WW channels:





Tau channels can contribute here, even with less precise m_{rec} than the bb channels

An Extreme Example from Georgios Choudalakis

Ten MC events, used to estimate a background b, but with different weights.

τ_1=0.1
τ_2=0.2
τ_3=0.3
τ_4=0.4
τ_5=0.5
τ_6=0.6
τ_7=0.7
τ_8=0.8
τ_9=0.9
τ_{10}=1.0

The sum is 5.5 = b

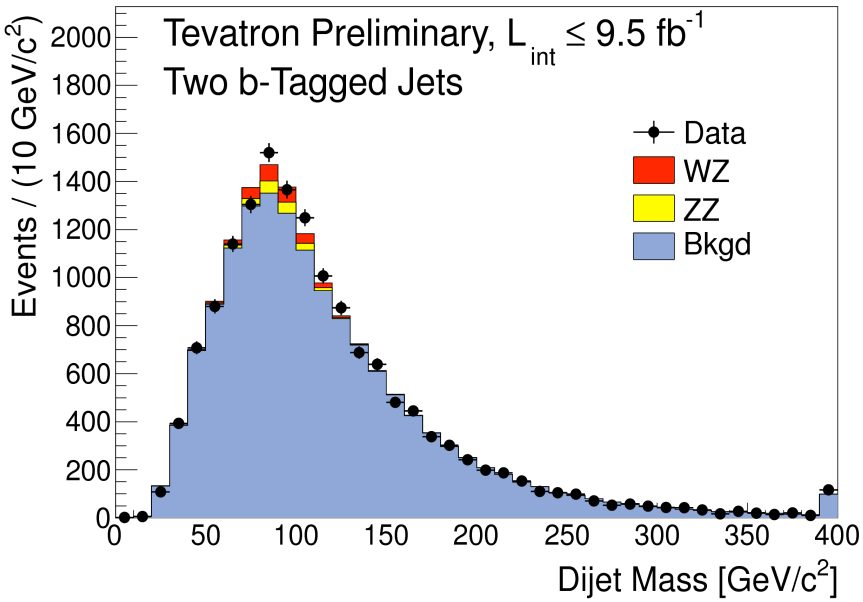
But what to use for the prior on b?

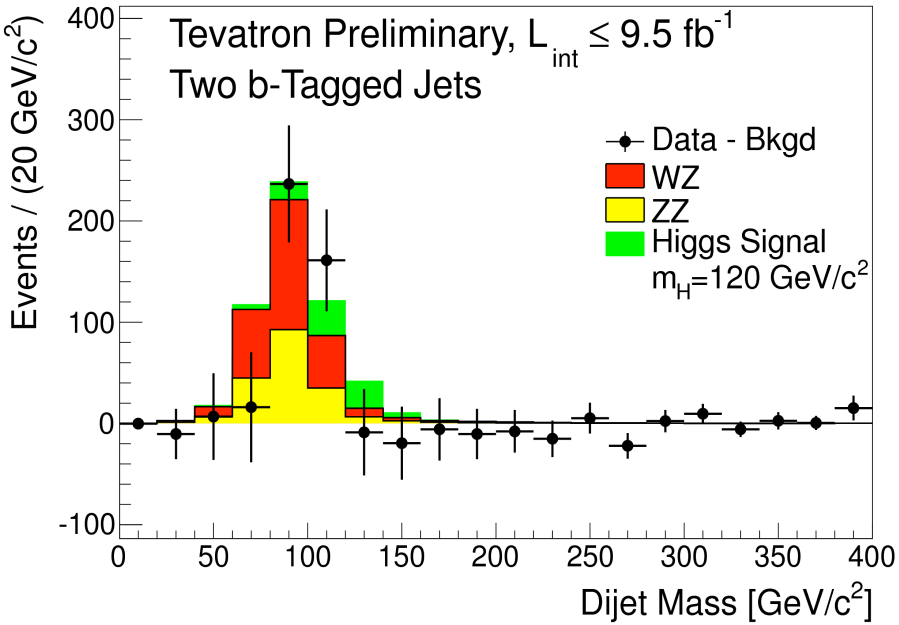
Are there any possible (and possibly large) weights which are not represented here? Could we have gotten a MC event with weight=100?

Very little information about the distribution of the weights is present here.

Need acceptance as a function of weight.

General limit/discovery tools – do we need a histogram of weights for each bin of each signal and background contribution? What if this is insufficient anyway (as it is in this case).


Tevatron Preliminary, $L_{int} \leq 9.5\ fb^{-1}$
Two b-Tagged Jets
Data
WZ
ZZ
Bkgd
Events / (10 GeV/c^2)
Dijet Mass [GeV/c^2]


Tevatron Preliminary, $L_{int} \leq 9.5\ fb^{-1}$
Two b-Tagged Jets
Data - Bkgd
WZ
ZZ
Higgs Signal
m_H=120 GeV/c^2
Events / (20 GeV/c^2)
Dijet Mass [GeV/c^2]

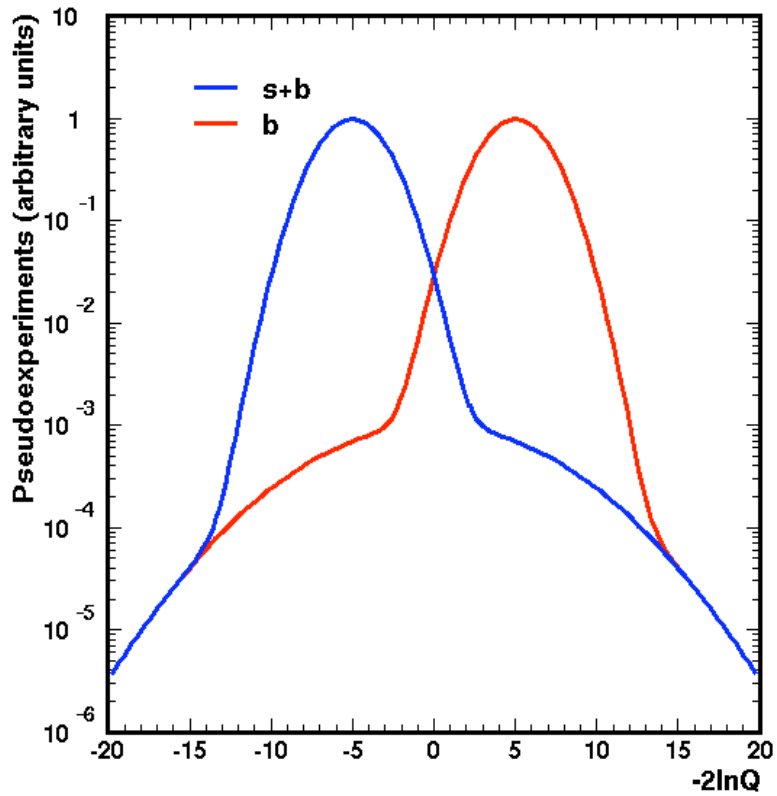
Interesting Behavior of CL_s

CL_s may not be a monotonic function of -2lnQ

Tails in the -2lnQ distribution shared in the s+b and b-only hypothesis (fit failures)

Distributions are sums of two Gaussians each. The wide Gaussian is centered on zero.

Practical reason this could happen – every thousandth experimental outcome, the fit program "fails" and gives a random answer.

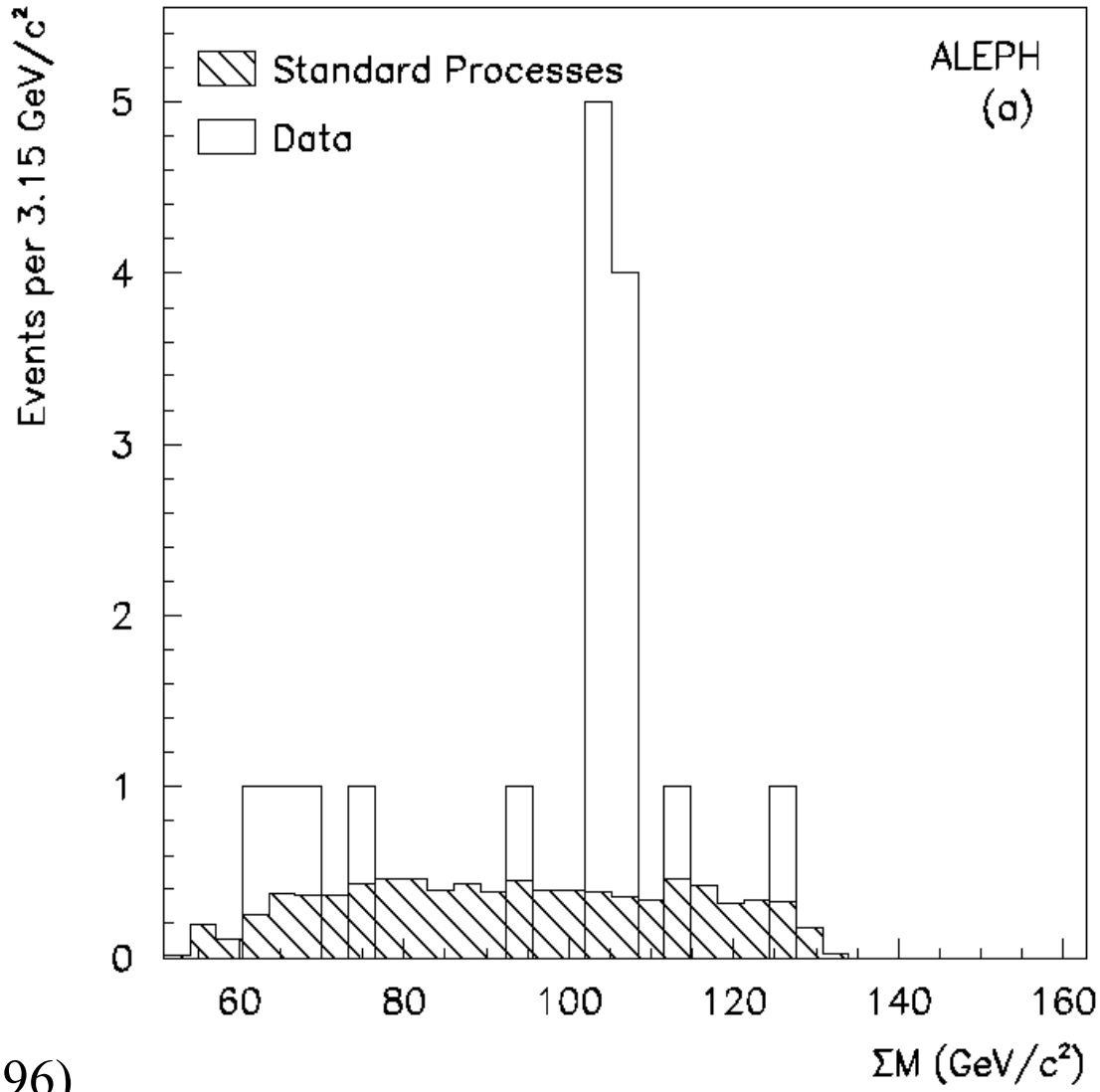


CL_s=1 for
-2lnQ < -15 or
-2lnQ > +15

Not really a pathology of the method, but rather a reflection that the test statistic isn't always doing its job of separating s+b-like outcomes from b-like outcomes in some fraction of the cases.

A Bump that Got Away

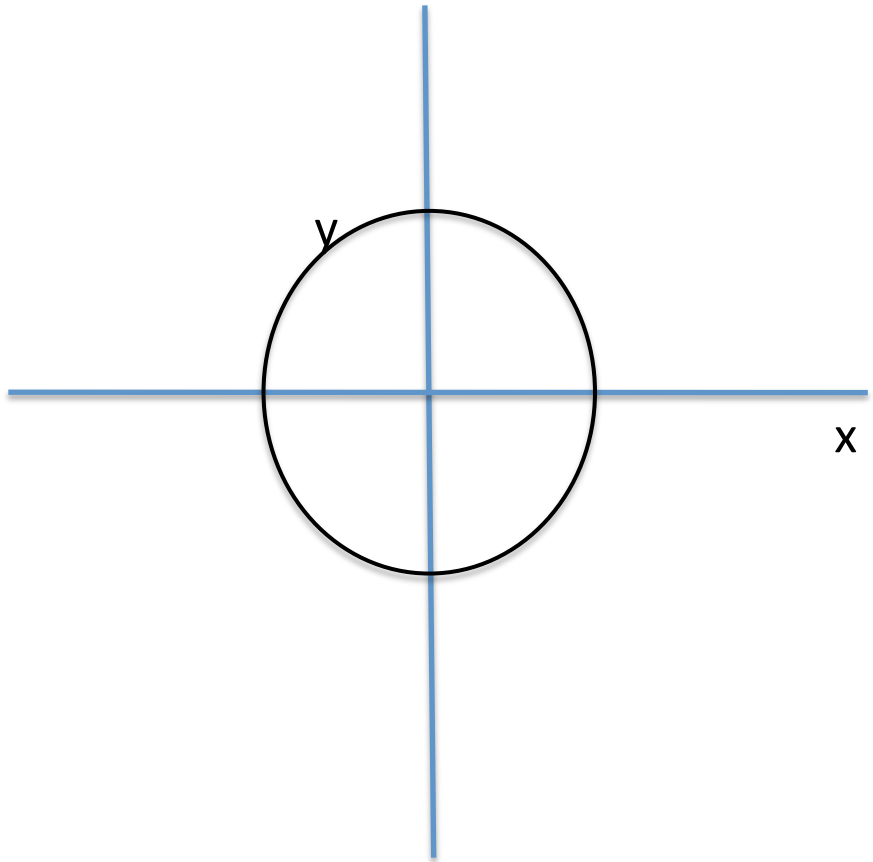
"the width of the bins is designed to correspond to twice the expected resolution ... and their origin is deliberately chosen to maximize the number of events found in any two consecutive bins"



ALEPH Collaboration, Z. Phys. C71, 179 (1996)

Dijet mass sum in $e^+e^- \rightarrow jjjj$

A Sample with Zero Covariance is Not Necessarily Uncorrelated



Example – perimeter of a circle. Knowledge of x provides knowledge of y up to a 2-fold ambiguity. But the covariance of the sample vanishes!

Something to watch out for with Principal Components Analysis – does not remove correlation, only covariance.

Cases to be Careful about Applying the LEE Approximation

Not all searches are bump hunts on a smooth background

– Multivariate Analyses are usually trained up
at each mass separately, and there is not a single distribution we can look elsewhere in.

Statistical effects only. If there's a systematic effect in the background modeling,
a "signal" may grow in significance with additional data in a way that's not described here.

The mismodeling may be concentrated in a small portion of the histogram (this is not
a LEE effect but a more difficult question).

Background parameterization may grow in sophistication as data are collected.

Not all LR test statistic distributions are modeled well by chisquared distributions.

Combine a large-data-sample bump hunt with a high s/b, low-background (say, b=1e-5)
search and the distribution of the LR is a convolution of chisquared and Poisson.