Data Analysis and Statistical Methods in Experimental Particle Physics

Thomas R. Junk
Fermilab

University of D0 Lecture Series. Nov. 8, 15, Dec. 6, 15

Lecture 2: Data Analysis Issues and Systematic Uncertainties

- Example Analyses using approximate Gaussian statistics:
  - Large Data Set Cross Section Measurement
  - A prominent mass peak on a smooth background
  - TGC analysis at LEP2 with multiple peaks in the likelihood

- Multivariate analyses
  - Neural Networks
  - Boosted Decision Trees
  - Matrix Elements

Measuring a Cross Section

Number of observed events: counted

Background: Measured from data / calculated from theory

$$\sigma^{meas} = \frac{N_{obs} - N_{BG}}{L \cdot \varepsilon}$$

$$"L" = \int L\,dt$$

Measured Cross section σ

Efficiency: optimized by experimentalist

Integrated Luminosity: Determined by accelerator, trigger prescale, …

Many thanks to B. Heinemann for the slides

Uncertainty on the Measured Cross section

- You will want to minimize the uncertainty:

$$\frac{\delta\sigma}{\sigma} = \sqrt{\frac{\delta N_{obs}^2 + \delta N_{BG}^2}{(N_{obs} - N_{BG})^2} + \left(\frac{\delta\mathcal{L}}{\mathcal{L}}\right)^2 + \left(\frac{\delta\epsilon}{\epsilon}\right)^2}$$

"Fractional Uncertainties Add in Quadrature"

- Thus you need:
  - N_{obs}-N_{BG} small (I.e. N_{signal} large)
    - Optimize selection for large acceptance and small background
  - Uncertainties on efficiency and background small
    - Hard work you have to do
  - Uncertainty on luminosity small
    - Usually not directly in your power

Slide from B. Heinemann, 2008

Luminosity Measurements and Uncertainties

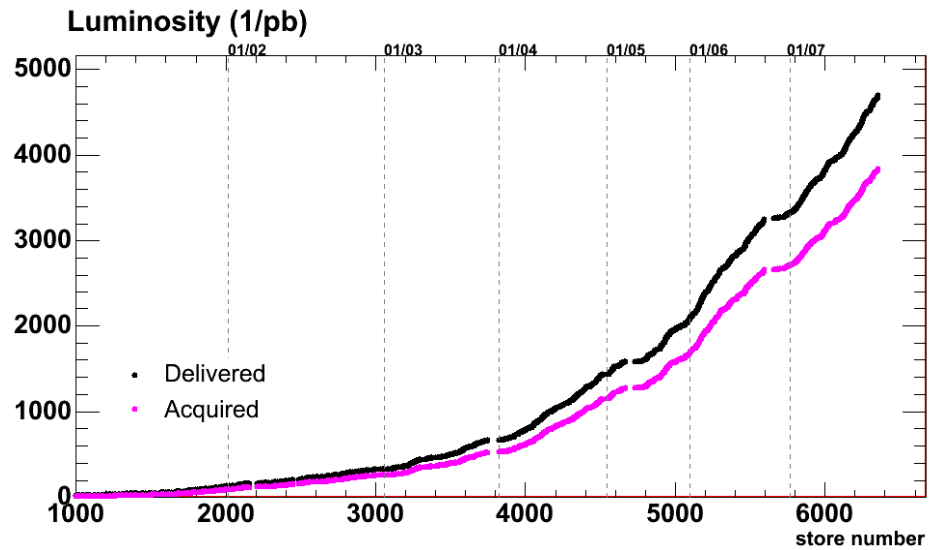
- Many different ways to measure it:
  - Beam optics
    - LHC startup: precision ~20-30%
    - Ultimately: precision ~5%
  - Relate number of interactions to total cross section
    - absolute precision ~4-6%, relative precision much better
  - Elastic scattering:
    - LHC: abslute precision ~3%
  - Physics processes:
    - W/Z: precision ~2-3% ?
- Need to measure it as function of time:
  - $L(t) = L_0 e^{-t/\tau}$ with $\tau \approx 14$h at LHC and L_0 = initial luminosity

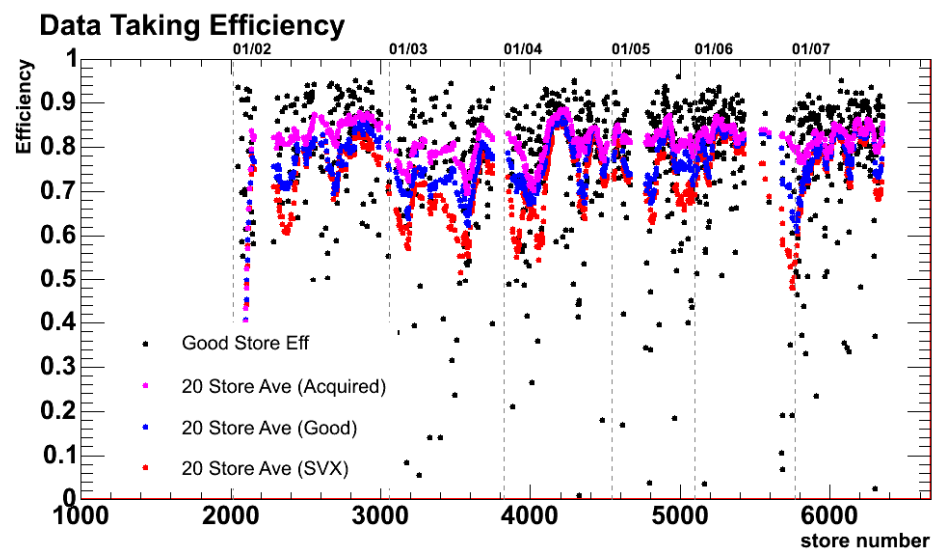
Luminosity Estimates are a "Shared Resource" – One example of a calibration shared by many groups

Slide from B. Heinemann, 2008

Your Luminosity

- Your data analysis luminosity is not equal to LHC/Tevatron luminosity!
- Because:
  - The detector is not 100% efficiency at taking data
  - Not all parts of the detector are always operational/on
  - Your trigger may have been off / prescaled at times
  - Some of your jobs crashed and you could not run over all events
- All needs to be taken into account
  - Severe bookkeeping headache





Slide from B. Heinemann, 2008

A Problem with that Uncertainty Formula

$$\frac{\delta\sigma}{\sigma} = \sqrt{\frac{\delta N_{obs}^2 + \delta N_{BG}^2}{(N_{obs} - N_{BG})^2} + \left(\frac{\delta\mathcal{L}}{\mathcal{L}}\right)^2 + \left(\frac{\delta\epsilon}{\epsilon}\right)^2}$$

$$\sigma^{meas} = \frac{N_{obs} - N_{BG}}{L \cdot \varepsilon}$$

Both the integrated luminosity in the denominator and the N_{BG} in the numerator depend on the luminosity estimate, because some backgrounds are estimated using

Theory cross section x Integrated Luminosity x branching ratios x cut acceptance.

Other backgrounds may be estimated using data-based techniques (more on this later)

→ Missing a correlation!

Handling Correlations the Easy Way

$$\sigma^{meas} = \frac{N_{obs} - N_{BG}}{L \cdot \varepsilon}$$

1) Identify ***independent*** sources of systematic uncertainty. Usually they have names and are listed in tables of systematic uncertainties. These are called ***nuisance parameters***

Luminosity estimate depends on:

- Inelastic pp (or ppbar) cross section
- Luminosity monitor acceptance

or, if using a data-based luminosity extraction

- Inclusive W or Z cross section theory prediction, and
- Lepton identification systematic uncertainty

Note – you cannot measure the inclusive Z cross section using the second method.

continued:

Handling Correlations the Easy Way

$$\sigma^{meas} = \frac{N_{obs} - N_{BG}}{L \cdot \varepsilon}$$

2) Evaluate the impact of each nuisance parameter on your answer, holding the others fixed:

$$\frac{d\sigma^{meas}}{d\nu_i}$$

were ν_i is the ith nuisance parameter.

Tip – you can often collect nuisance parameters together if they all affect the result in the same way. "Integrated Luminosity" is a perfectly good nuisance parameter most of the time, as predictions depend on it.

But sometimes you can't. Suppose pp→Z is one of the background sources, and you are using the measured Z rate to constrain the luminosity in the data.

These can even be non-overlapping data. Z→ee constrains the lumi, while Z→hadrons is a background, for example.

Then the inclusive Z cross section assumed becomes the nuisance parameter (and its impact partially cancels!)

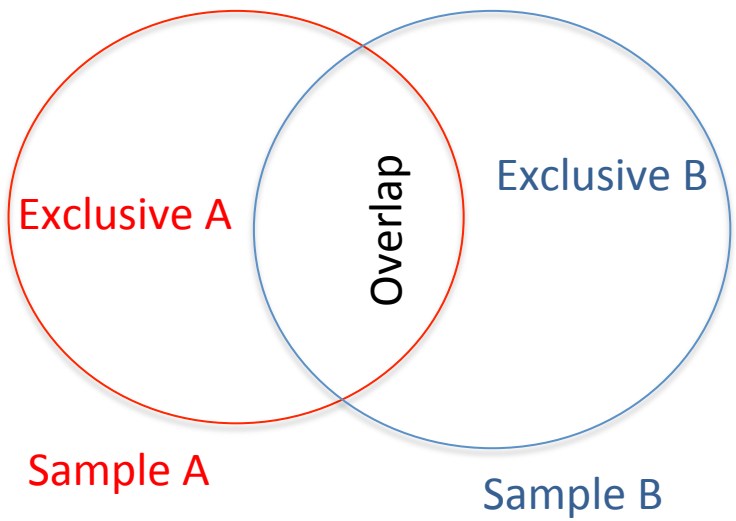
Handling Correlations the Easy Way

Tip: Sometimes uncertainties are correlated in a nontrivial way:

Example: Two sources of background are estimated using data control samples, but these control samples share some but not all of their data events.

Suggestion: Always seek an uncorrelated parameterization. We know that those control samples are partially correlated due to the overlaps. We estimate the correlation by knowing the fractions in the exclusive and inclusive samples.

You can always break down partially correlated uncertainties into pieces – a fully correlated piece and uncorrelated pieces.



In this case, Exclusive A, Exclusive B, and Overlap may be the nuisance parameters for evaluating the stat. uncertainty from these control samples

Handling Correlations the Easy Way

Putting it all together:
Once you have an uncorrelated basis, just add the uncertainties in quadrature.

$$\delta\sigma^{meas} = \sqrt{\sum_i \left(\frac{d\sigma^{meas}}{d\nu_i} \delta\nu_i \right)^2}$$

A nuisance parameter is any value you assumed in order to do your analysis which you do not know the exact value of (usually all of them).

Much of the work is devoted to identifying a proper set of nuisance parameters, and constraining their possible values, preferably with data.

We must frequently ask theorists for help!

Example of Data-Driven Background Estimates

Seek Z→ee events, but there are misreconstructed W(→ev)+jets events where Missing E_T is small and a jet fakes an electron

Typical cuts: Require small Missing E_T, opposite-sign electrons, electron isolation, centrality, and P_T (usually > 20 GeV), and m_{ee} close to m_Z

Standard technique: Count same-sign events passing all the other requirements.

Assumption: Jets faking electrons do so with random charge assignments: Can just use the count of same-sign events as the W+jets background.

A hole in the assumption: The charge of the W is anticorrelated with the charge of the leading particles in the accompanying jet.

Measure the hole: Using a sample purified in W+jets (high missing E_T), measure the charge correlation between the W→e and "fakeable objects" in the accompanying jets.

Acceptance / Efficiency

- Actually rather complex:
  - Many ingredients enter here
  - You need to know:

$$\varepsilon_{total} = \frac{\text{Expected Number of Events used in Analysis}}{\text{Expected Number of Events Produced}}$$

- Ingredients:
  - Trigger efficiency
  - Identification efficiency
  - Kinematic acceptance
  - Cut efficiencies

Slide from B. Heinemann, 2008

Trigger Efficiency

Triggers typically select events with

- Isolated leptons with p_T> a threshold (20 GeV typical at the Tevatron)
- Missing Transverse Energy (various thresholds, depending on other objects in the event
- Jets – total energy, reconstructed jet counts above E_T thresholds

Triggers are difficult to model in Monte Carlo

- Rely on partially reconstructed information – whatever an FPGA can compute in a few microseconds
- Triggers sometimes get updated
- Trigger hardware sometimes fails and gets repaired/upgraded

Most reliable way to estimate trigger efficeincy – with datasets collected on overlapping triggers.

- Selects some fraction of events also selected by desired trigger.

Check what fraction of events passed the target trigger that "should" have.

Example: Using the isolated lepton trigger to check the MET + jets trigger

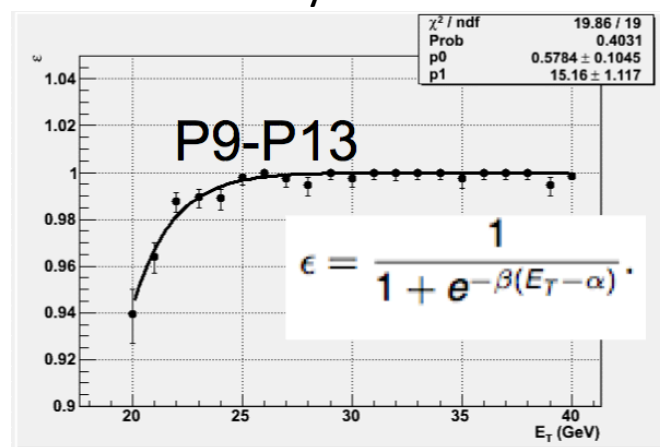
Trigger and ID Efficiency for e's and μ's

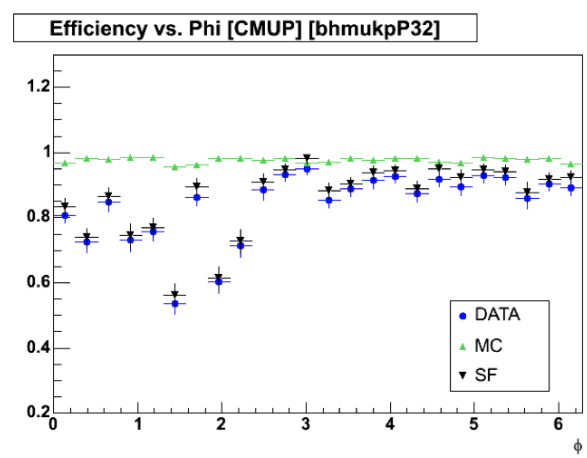
- Can be measured using Z's with tag & probe method
  - Statistically limited

$$\varepsilon_{trig} = \frac{N_{trig}}{N_{ID}}$$

- Can also use trigger with more loose cuts to check trigger with tight cuts to map out
  - Energy dependence
    - turn-on curve decides on where you put the cut
  - Angular dependence
    - Map out uninstrumented / inefficient parts of the detectors, e.g. dead chambers
  - Run dependence
    - Temporarily masked channels (e.g. due to noise)

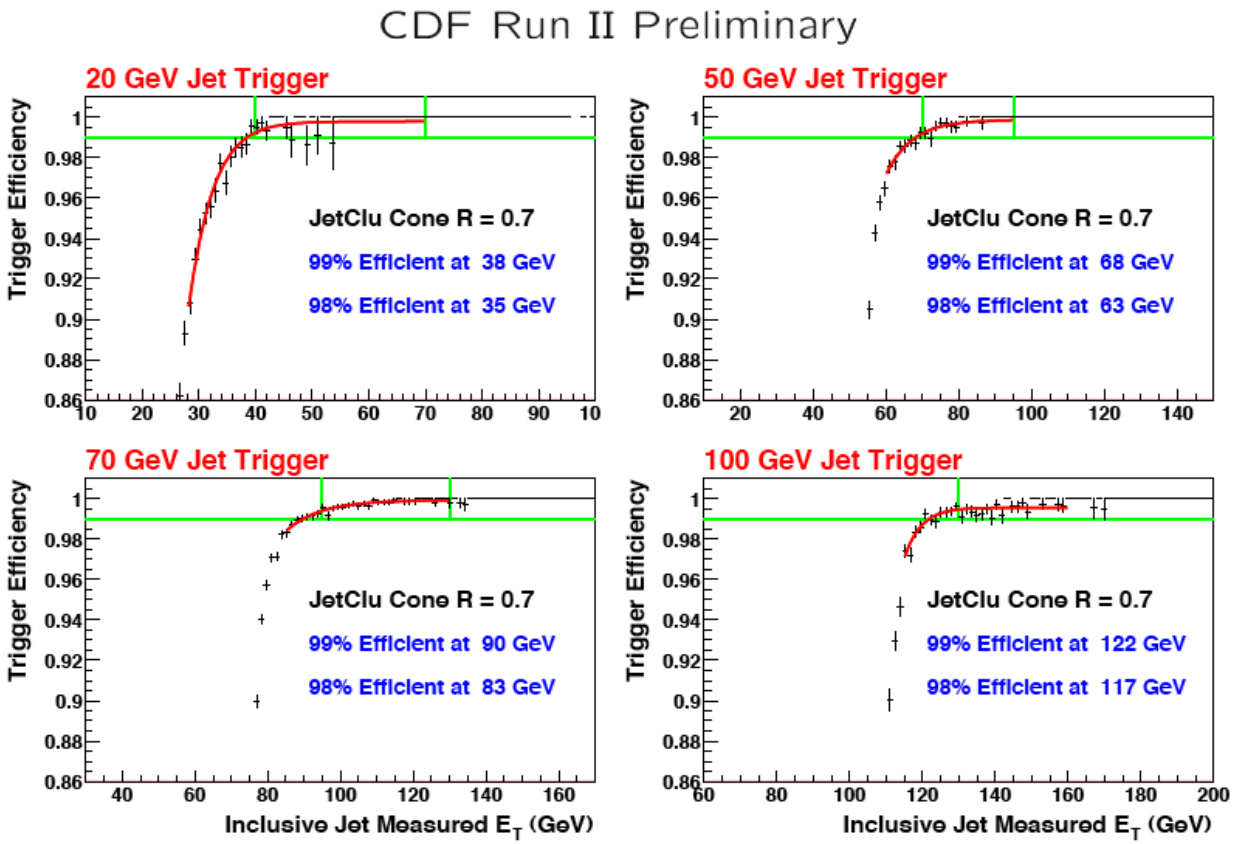
CDF Level 2 Calorimeter efficiency





Slide from B. Heinemann, 2008

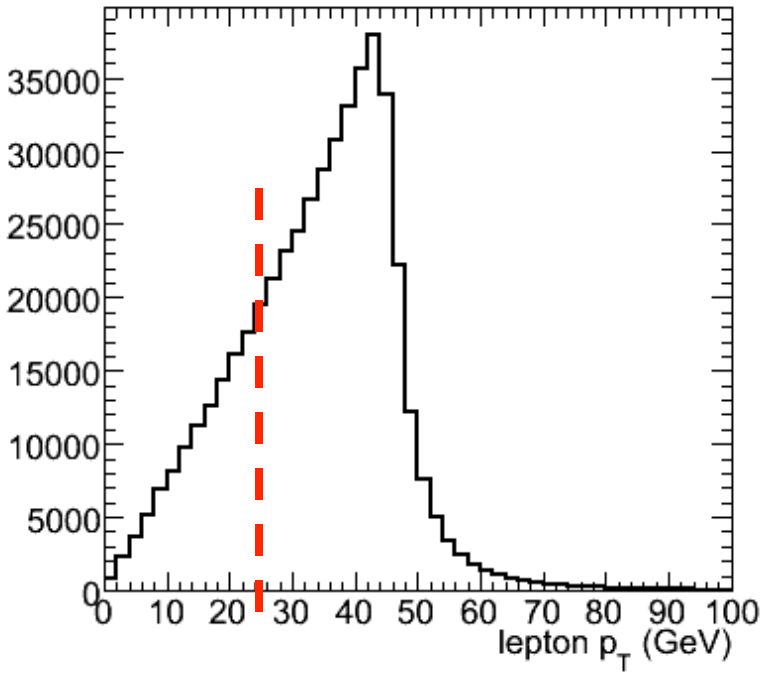
Jet Trigger Efficiencies

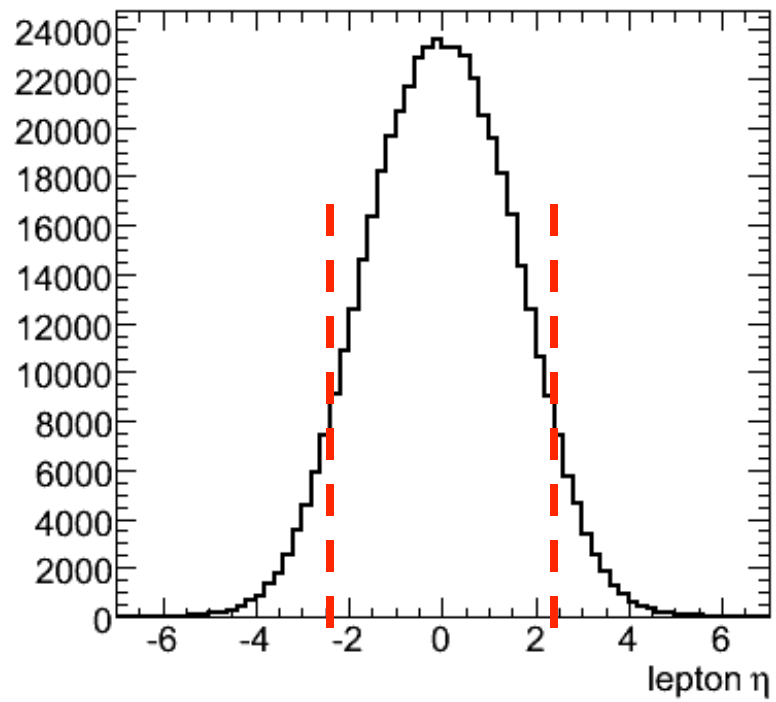


- Bootstrapping method:
  - E.g. use MinBias to measure Jet-20, use Jet-20 to measure Jet-50 efficiency … etc.
- Rule of thumb: choose analysis cut where ε>90-95%
  - Difficult to understand the exact turnon

Slide from B. Heinemann, 2008

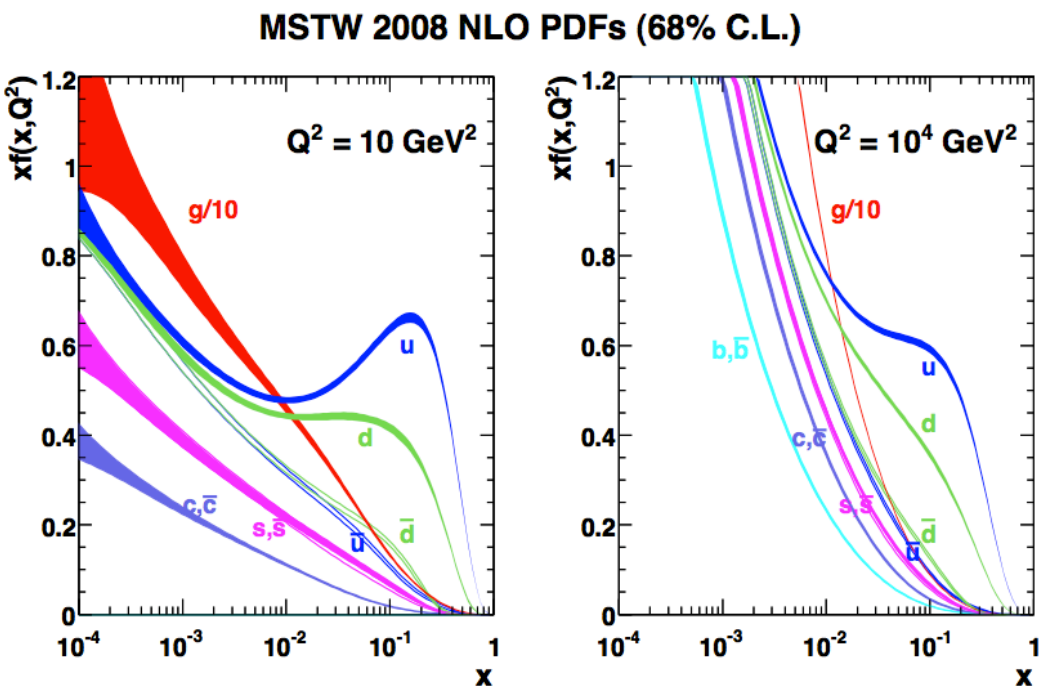
Acceptance of Kinematic Cuts: Z's

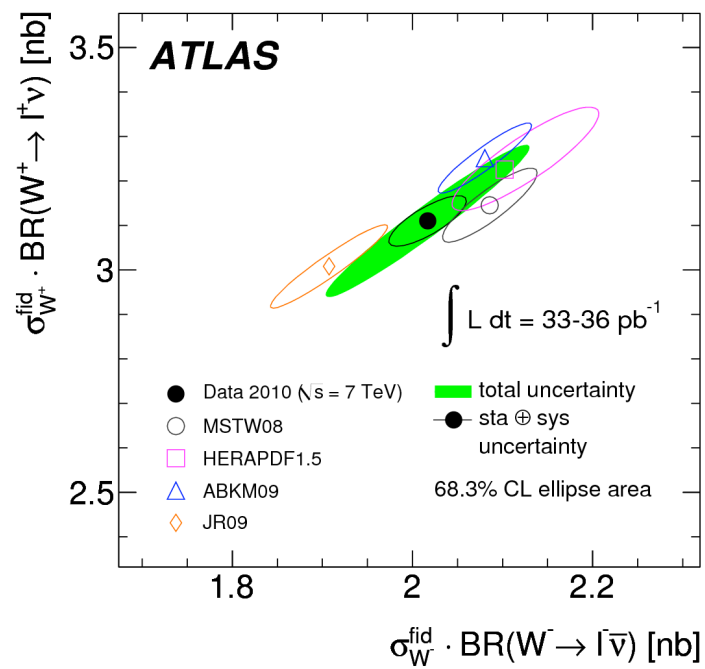




- Some events are kinematically outside your measurement range
- E.g. at Tevatron: 63% of the events fail either p_T or η cut
  - Need to understand how certain these 63% are
  - Best to make acceptance as large as possible
    - Results in smaller uncertainties on extrapolation

Parton Distribution Functions





Affect analysis in two ways:

1) Changes the cross section prediction (not a problem for the signal, that's what we're measuring! But an issue for backgrounds).
2) Changes the differential distributions – mostly via p_z and η

Measurements of standard-candle processes such as pp→W and Z constrain PDF's

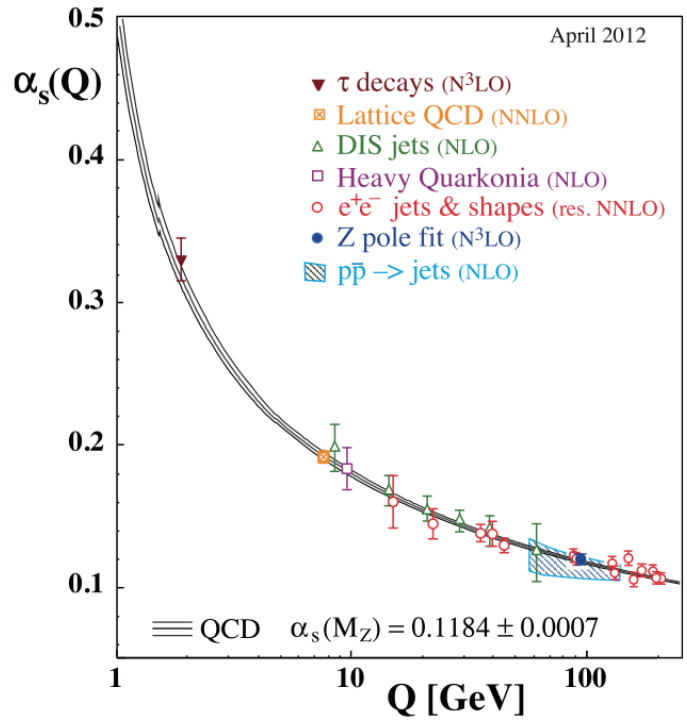
Factorization and Renormalization Scales

Fixed-order matrix-element calculations in Monte Carlo generators (Pythia, MadEvent, Alpgen, etc) are missing higher-order corrections.

Parton showers cover some of this but not all.

Changing the scale at which α_s is computed changes the predicted cross sections as well as differential shapes, giving more weight to some events than others.

Visible in p_T spectra in gg→H, ppbar→Z, etc.



Would like to constrain as many properties of background processes with data control samples as possible.

Not possible for signals that haven't been observed yet!

Jet Energy Scale (JES)

Simulation of jet energies is fraught with possible errors

- Incomplete material description
- Incorrect nuclear cross sections
- Incomplete modeling of hadronic showers
- Incorrect modeling of quark and gluon fragmentation and hadronization

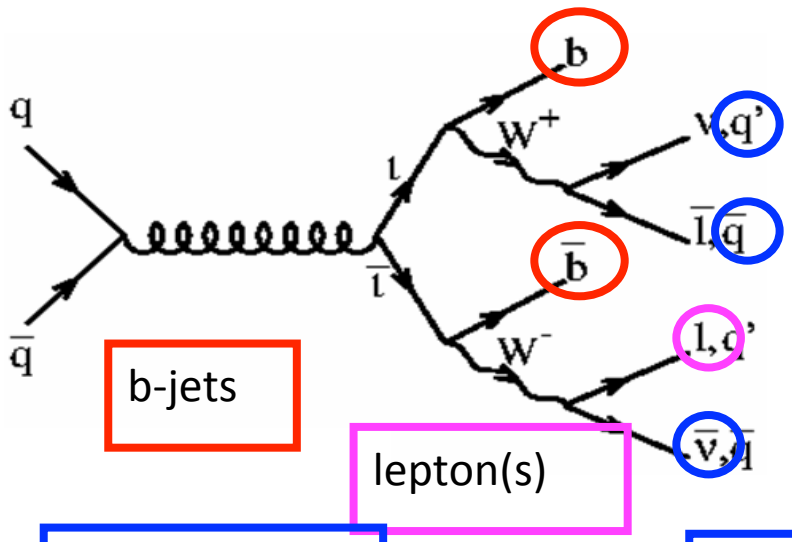
JES is an important ingredient in estimating selection efficiencies if jets are required or vetoed. Typically a jet E_T threshold must be passed in order for a jet to get identified as such.

Measurements:

- Test-beam calorimetry determinations
- In-situ calibrations
  - Photon-jet balancing
  - Z-jet balancing
  - W→jets in ttbar events

All require extrapolations and assumptions. For example, photons+jets and Z+jets have different quark/gluon content in the jets. W→jets has (almost) no b-quark content, but we may be interested in calibrating the jet energy scale for b's

Measuring JES in situ for a top quark mass measurement

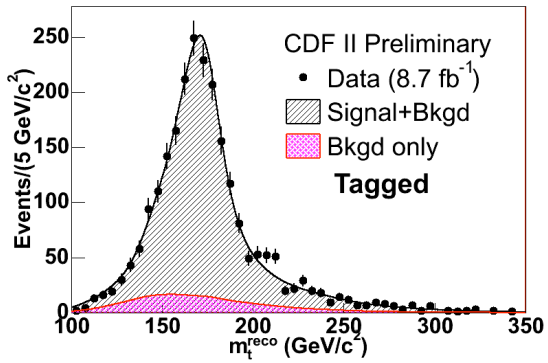


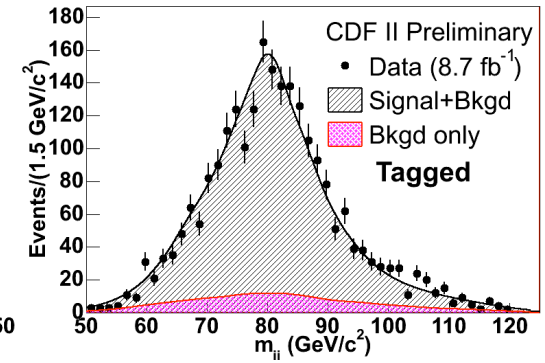
b-jets

lepton(s)

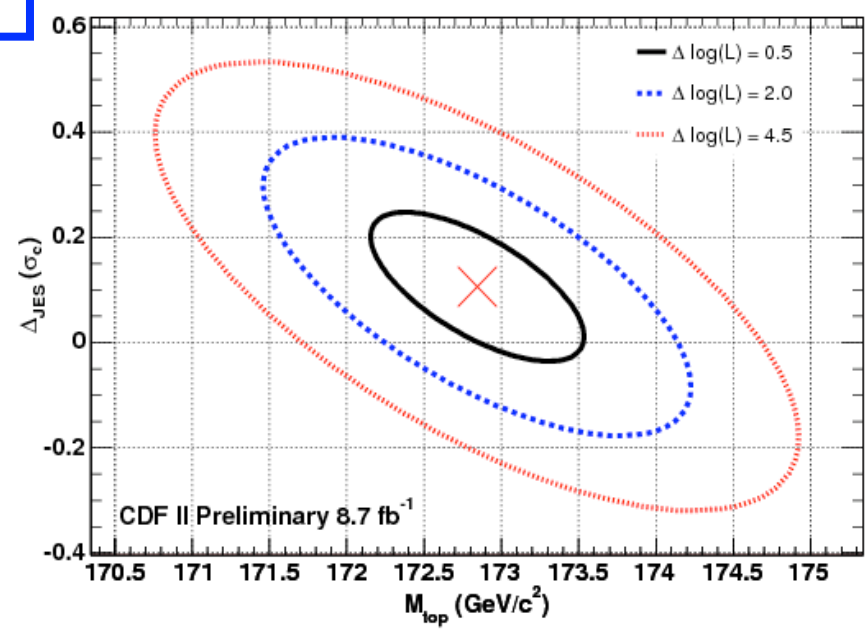
missing ET

more jets





Note – ΔlogL = 0.5 is used here as we want a 1D measurement of m_t; JES is just a nuisance parameter.



Systematic Uncertainties vs. Cross Checks

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

A typical cross-check of an analysis:
Change selection cuts, rerun analysis, see if you get a different answer.

Question: How different does it have to be before we get unhappy?

- There's a statistical component: we expect *some* change: tightening cuts removes some events, loosening them adds new ones, but most will be shared. What to expect?

- There may be a genuine systematic effect, but it only samples events near the cut being varied.
  - These events may not be that important anyway
  - Does not test events far away from the cut which may be more important
  - If your "best" (i.e. highest s/b) events are next to cuts, then there may be an analysis optimization issue lurking in there.

Systematic Uncertainties vs. Cross Checks

Varying cuts: assuming all events contribute the same amount to the answer, the width of the expected difference (Gaussian approx) is:

$$\sigma_{x_1 - x_2} = \sqrt{\sigma_2^2 - \sigma_1^2}$$

Easy for computing p-values – how many sigma we are different from zero is an estimate of how significant the discrepancy is.

If we see a 1 or 2-sigma effect? Count it as a systematic uncertainty in the result? Roger and I say no: It's a robustness check, not an indication that there's a problem.

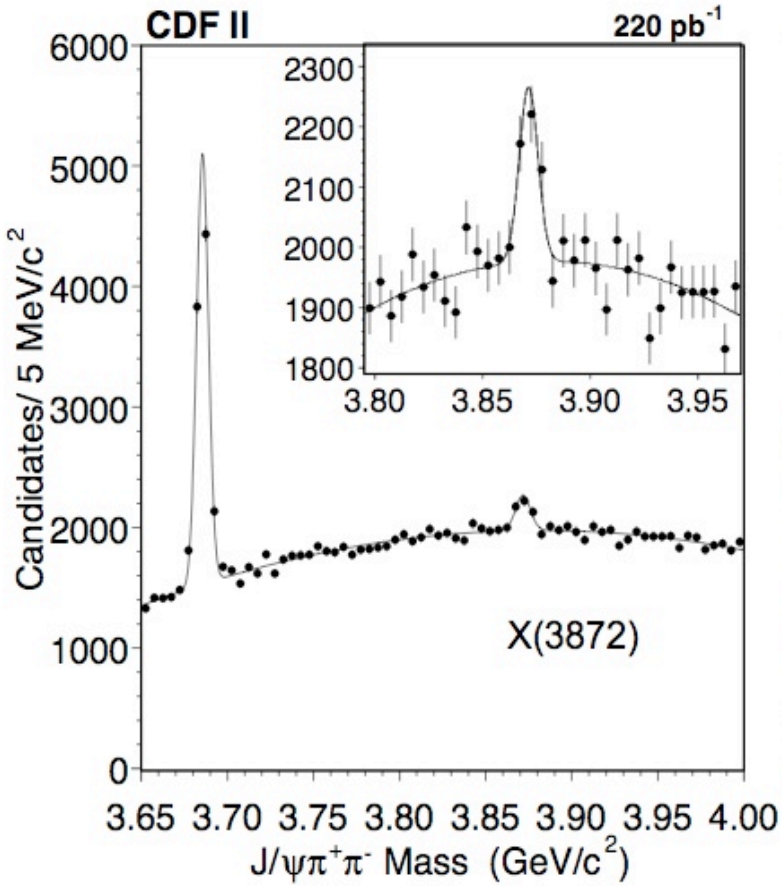
If the robustness check fails, try to identify what the assumption is in the model that's wrong. Model parameters are almost always uncertain: what knobs are there you can turn that can fix the problem?

Taking it as a systematic uncertainty penalizes diligence.

Also statistically weak cross-checks would penalize the total uncertainty. Some cross checks just are not that strong.

"On-Off" Example

Select events with J/ψ(→ll) $\pi^+\pi^-$ candidates. Lots of nonresonant background whichis poorly understood *a priori*, but there's a *lot* of it.



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

Typical strategy:
Fit the background outside of the signal peak, and interpolate the background under the signal to subtract it off.

The ratio of events in the sidebands to the background prediction under the signal is called **τ**

"Weak" Sideband Constraints

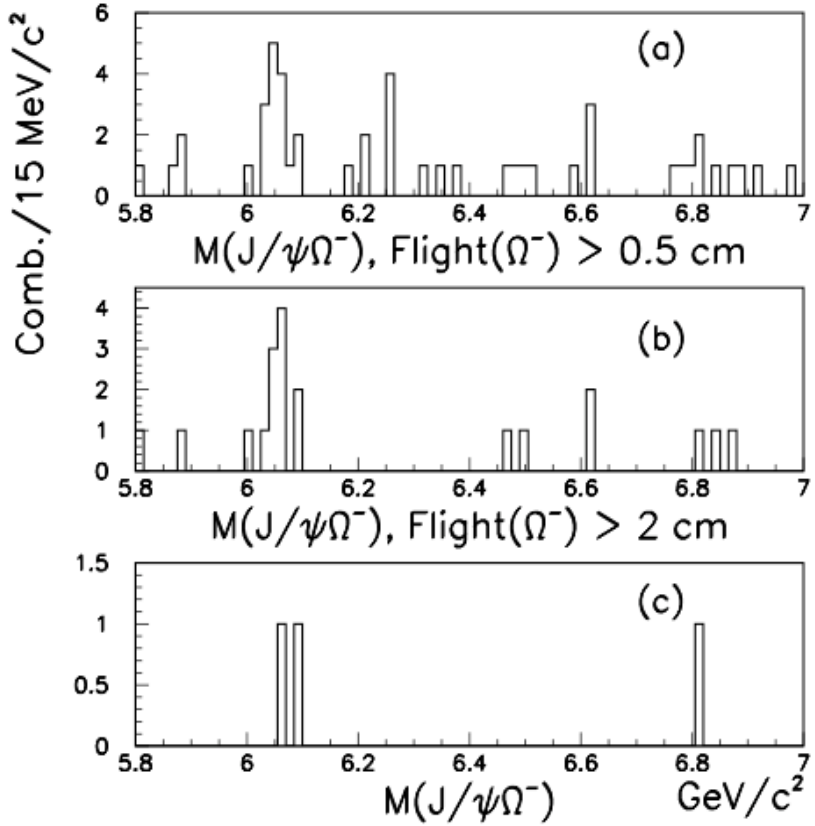


FIG. 8: (a,b) The invariant mass distribution of $J/\psi\,\Omega^-$ combinations for candidates where the transverse flight requirement of the Ω^- is greater than 0.5 cm and 2.0 cm. (c) The invariant mass distribution of $J/\psi\,\Omega^-$ combinations for candidates with at least one SVXII measurement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

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.

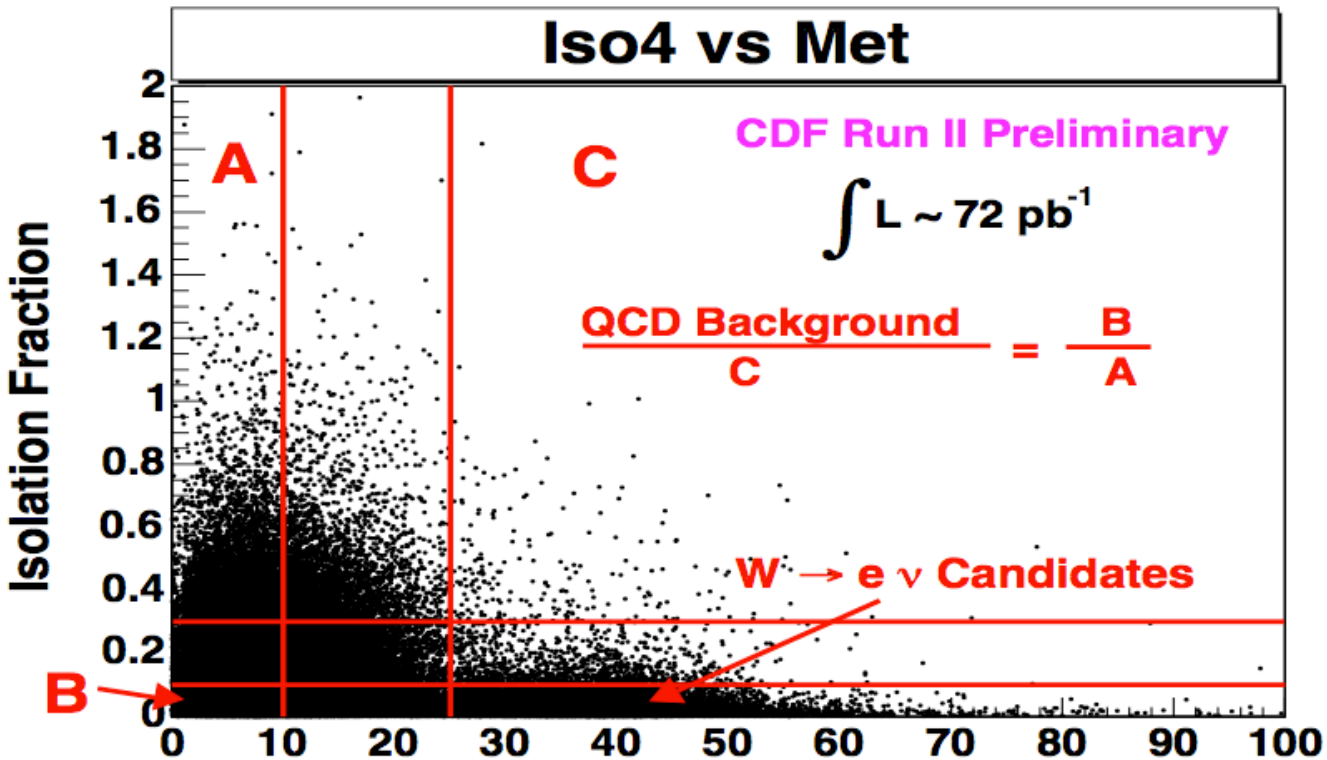
Classical example: Ray Davis's Solar Neutrino Deficit observation. Comparing data (neutrino interactions on a Chlorine detector at the Homestake mine) with a model (John Bahcall's Standard Solar Model). Calibrations of detection system were exquisite. But it lacked a standard candle.

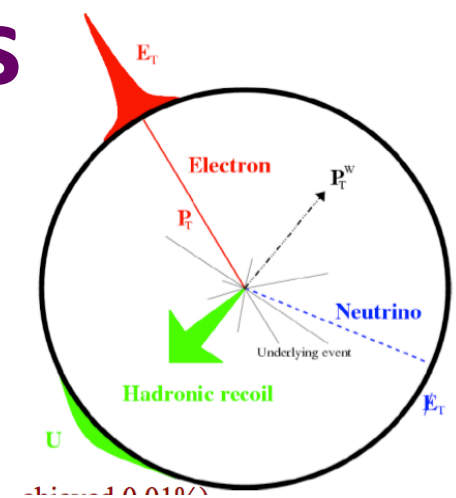
How to incorporate systematic uncertainties? Fewer options left.

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

"ABCD" Methods

CDF's W Cross Section Measurement





Isolation fraction=

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.

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

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

"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

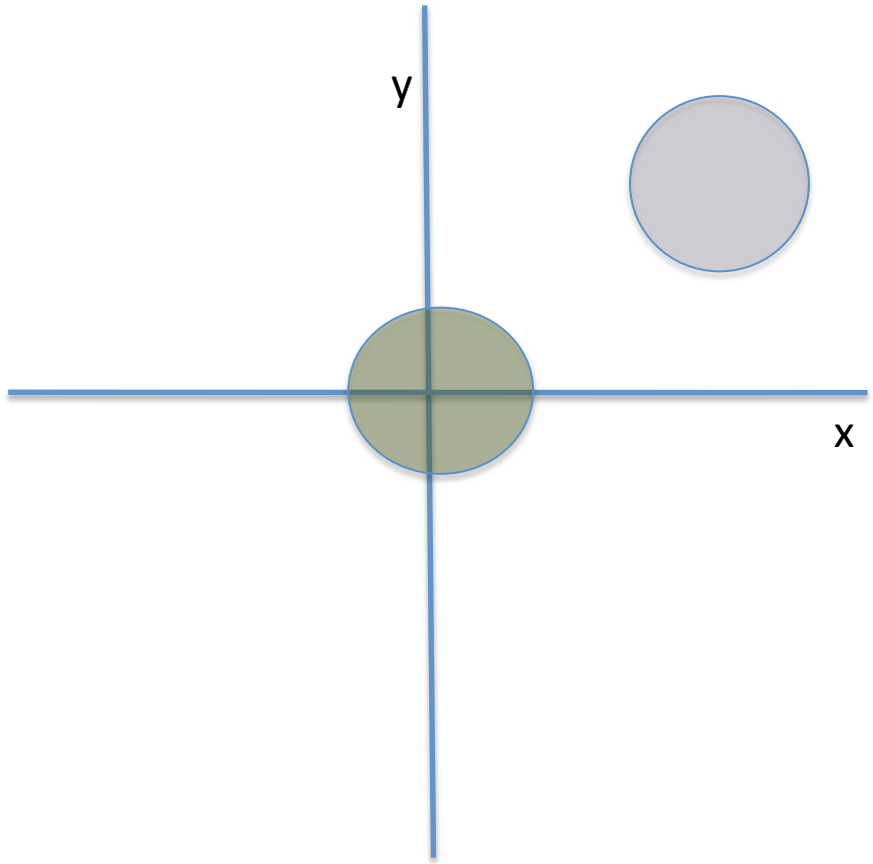
Disadvantages

- 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

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.

Underlying parameters may not scale the observation linearly

$$\sigma^{meas} = \frac{N_{obs} - N_{BG}}{L \cdot \varepsilon}$$

This assumes a signal is adding incoherently (QM sense) to a background.

But: Rates are proportional to matrix elements ***squared***.

Coupling parameters come in quadratically at least!

Sometimes signals and backgrounds intefere with each other quantum mechanically!

Example: gg→H→WW interferes with gg→WW. Campbell, Ellis, and Williams, JHEP 1110, 005 (2011)

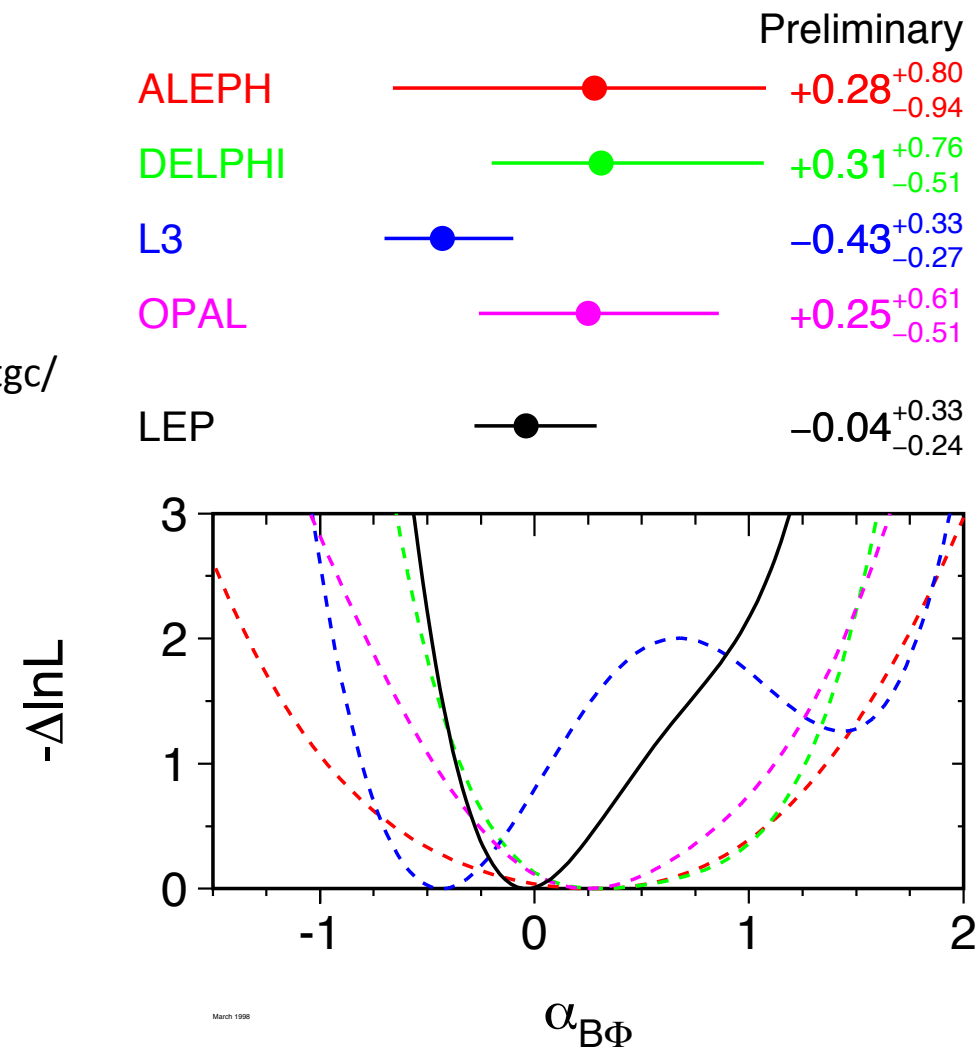
Another example: Seeking charged Higgs bosons in top quark decay. By changing the branching ratios, the effect of the presence of new physics can reduce the expected data counts. Negative signal? Or just less "background"? More on this later.

Example of a Multimodal Likelihood Function

LEP2 Triple Gauge Coupling
Constraints from 1998

http://lepewwg.web.cern.ch/LEPEWWG/lepww/tgc/

With more data, ambiguities were resolved, so I had to go back a ways to find a good example.



Multivariate Analyses

These are an important tool for optimizing sensitivity

- Reduce expected uncertainties on measurements
- Raise chances of discovering particles that are truly there
- Improve the ability to exclude particles that are truly absent

BUT:

- There are many ways to make a mistake with them: More work!
  - Optimizing them
    - Best input variables
    - Best choice of MVA
  - Validating them
    - Validate modeling of inputs *and* outputs
    - Check for overtraining
  - Propagate systematic uncertainties through them
    - Rates
    - Shapes
    - Bin-by-bin

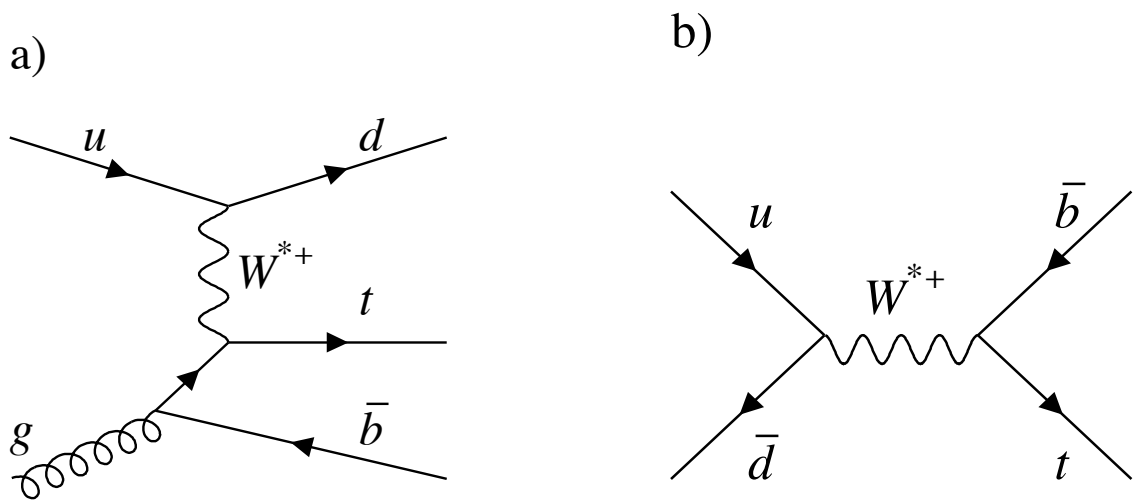
When MVA's provide the most benefit

If there are several reconstructed quantities per event that are useful for separating signal from background or measuring properties of signal.

If there's just one such variable, there can be no additional gain.

MVA's ***reduce dimensionality*** – start with many reconstructed quantities and reduce them down to one.

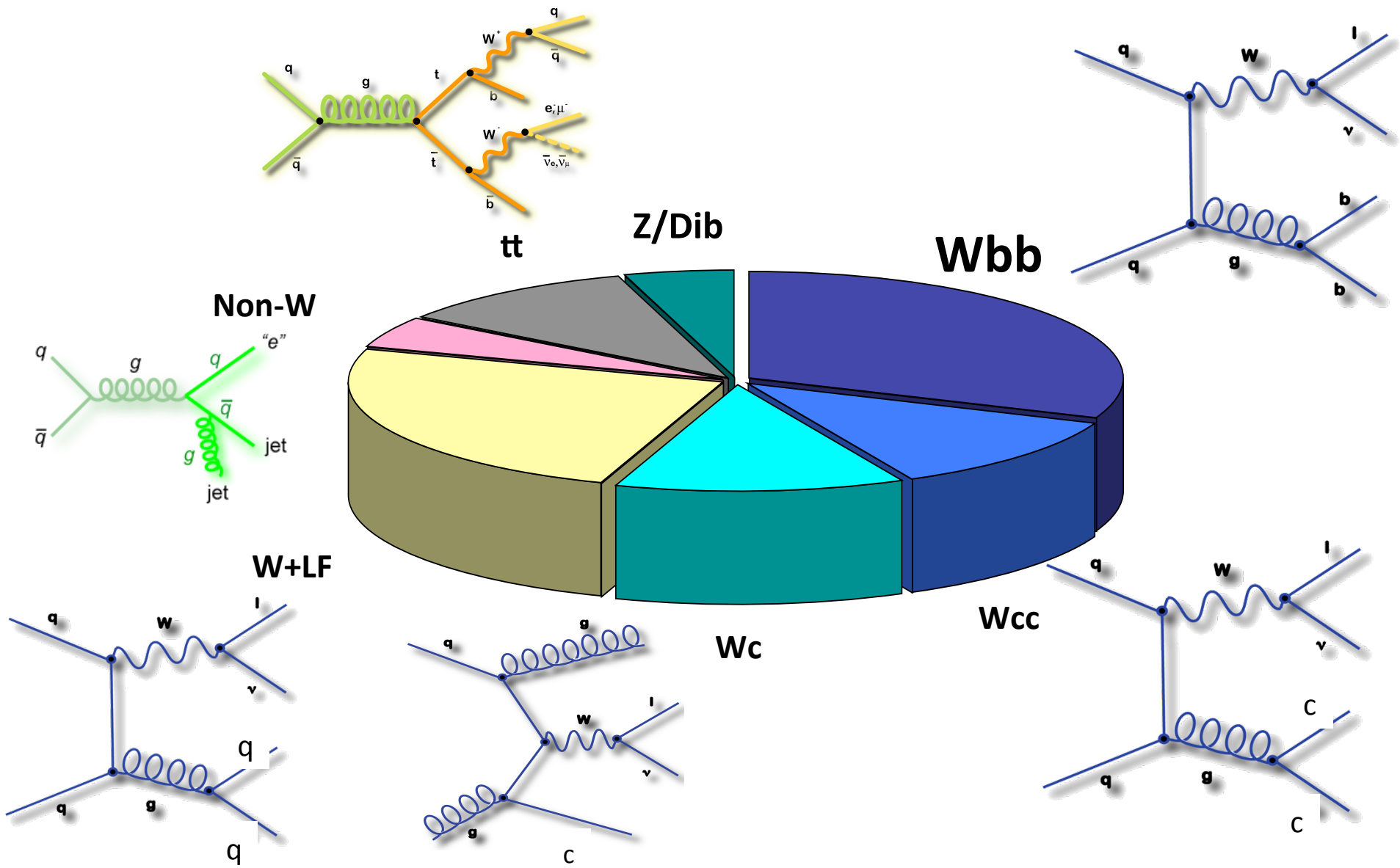
My favorite example – single top at the Tevatron



We know all about the top quark – mass spin, couplings.

s/b is small ~1:15, and uncertainty on background is about 30%. Need some way to purify signal and background

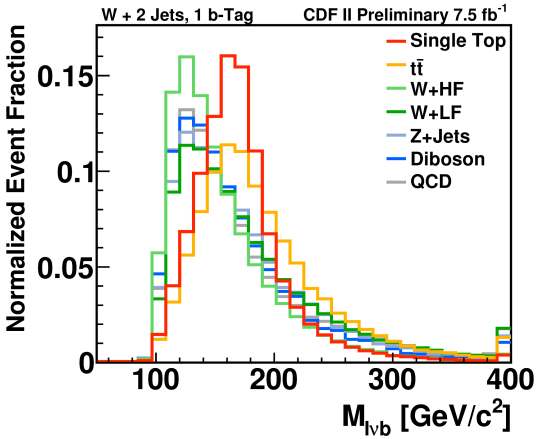
Backgrounds to Single Top Production

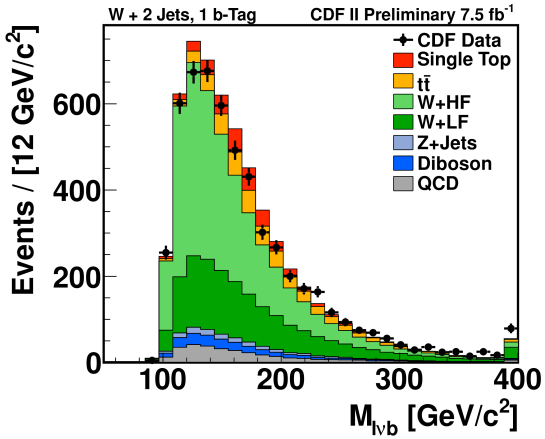


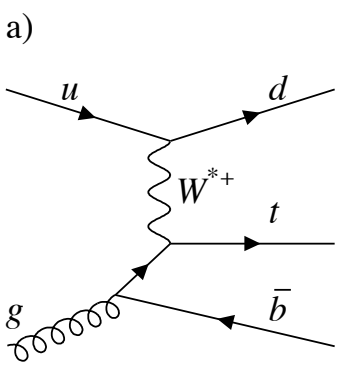
Single Top at the Tevatron MVA Example

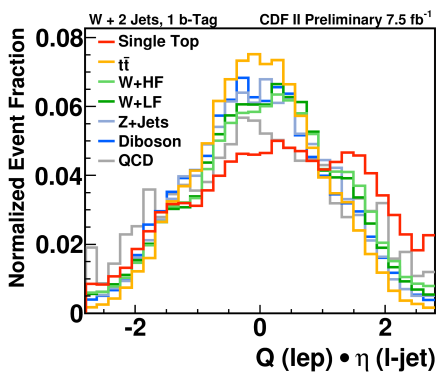
Select events with W→lν, two or three jets, one or more b-tags

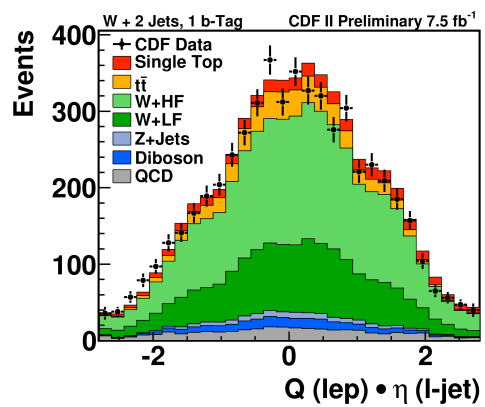
Not an easy bump-on-a-background search – the bump is too wide!
(poor mass resolution due to missing neutrinos)





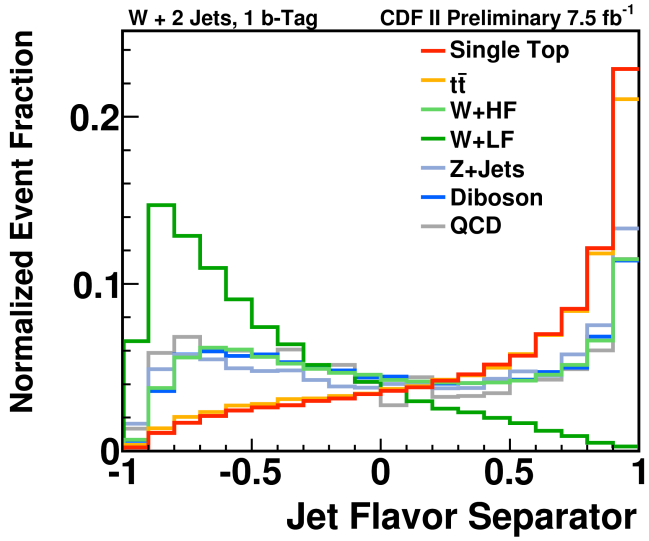


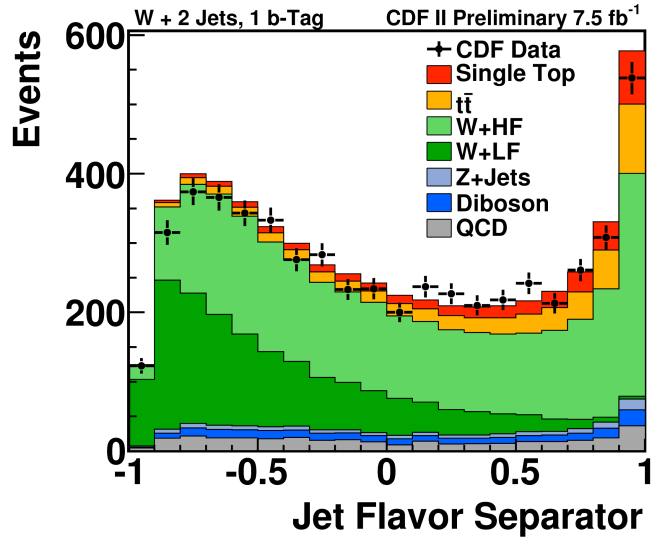




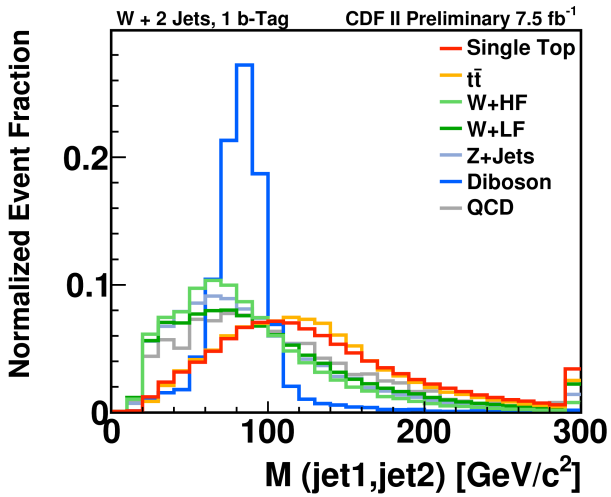
Another clever variable (suggested by C.P. Yuan)
Qxη

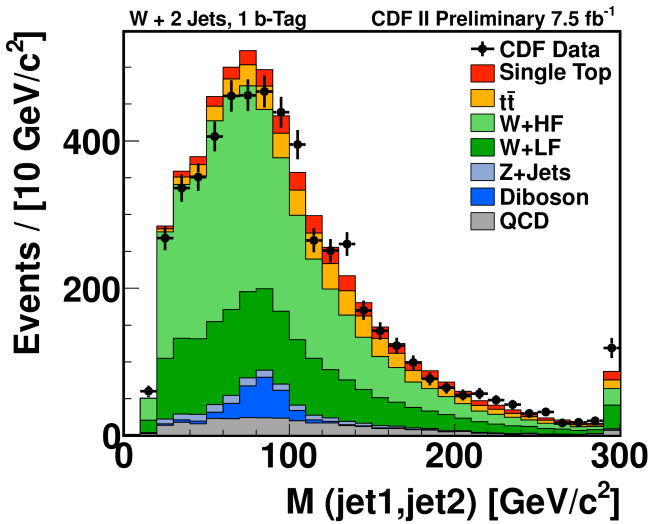
Single Top at the Tevatron MVA Example





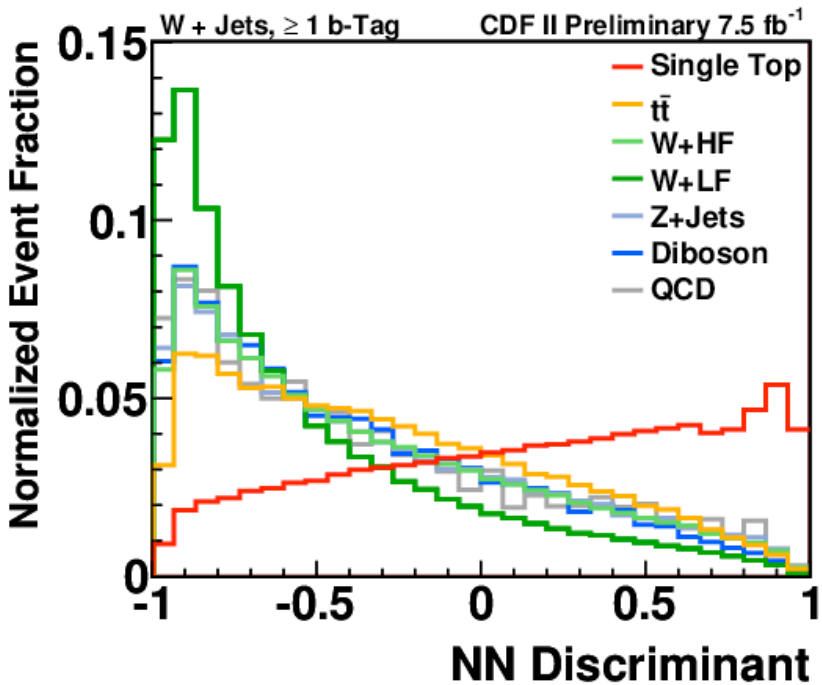
B-tag flavor separator – W+1-tag events are full of mistagged light-flavor and charm. This helps separate them

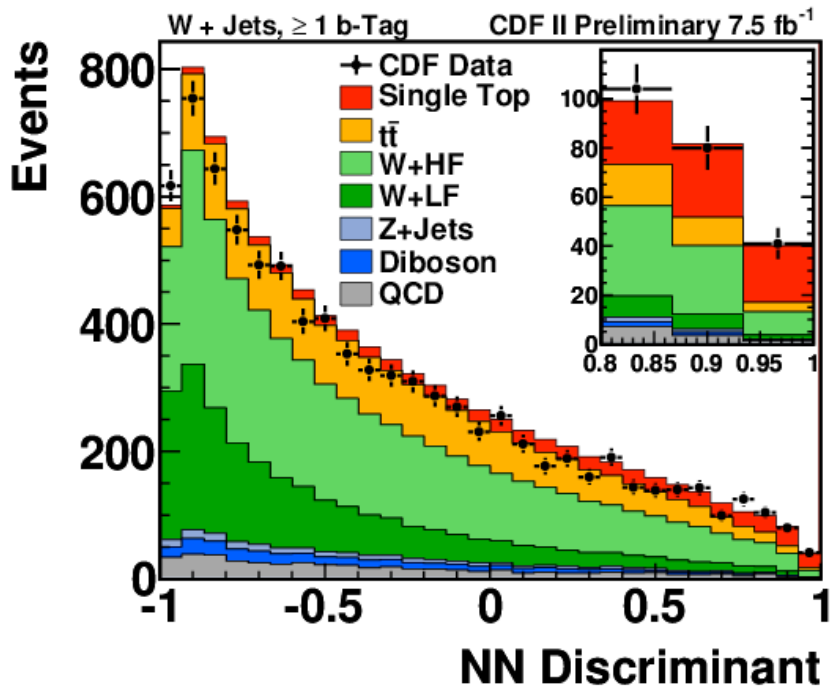




m_{jj} Surprised us a bit since it is not characteristic of single top. But it is for the background! (gluons are massless)

CDF's 7.5 fb^{-1} Single Top MVA output

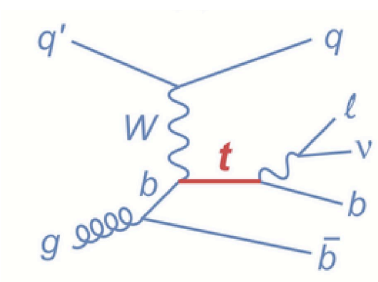


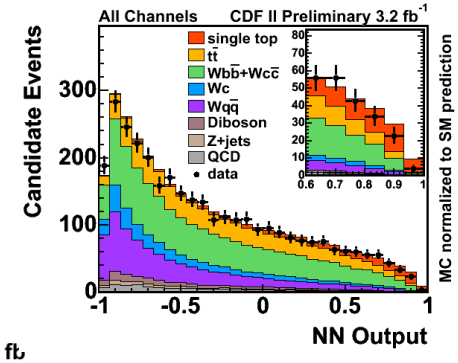


You can cut and count using the MVA output and use the statistical methods we discussed, or do something more sophisticated.

Checking Input Distributions to an MVA

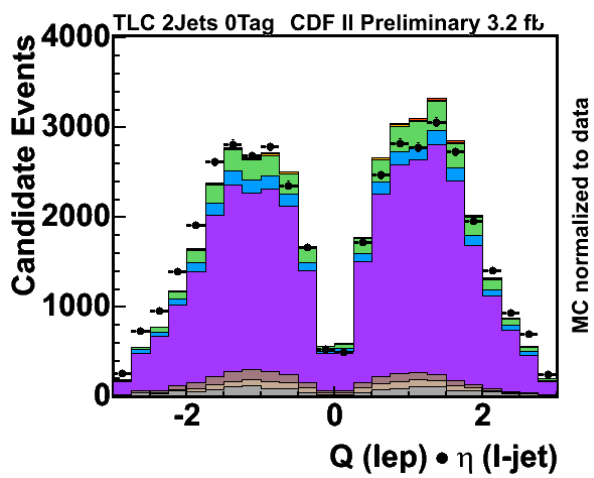
- Relax selection requirements – show modeling in an inclusive sample (example – no b-tag required for the check, but require it in the signal sample)
- Check the distributions in sidebands (require zero b-tags)
- Check the distribution in the signal sample for all selected events
- Check the distribution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

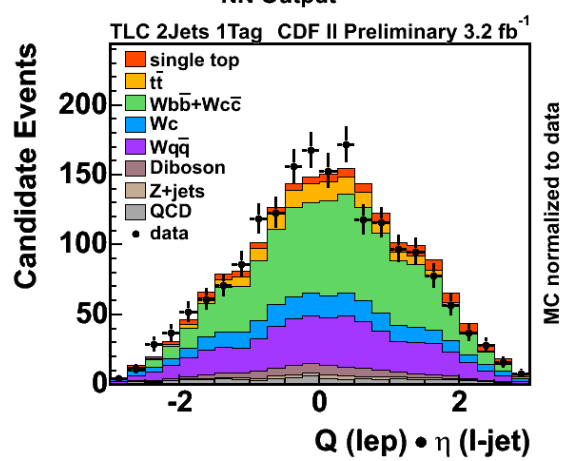


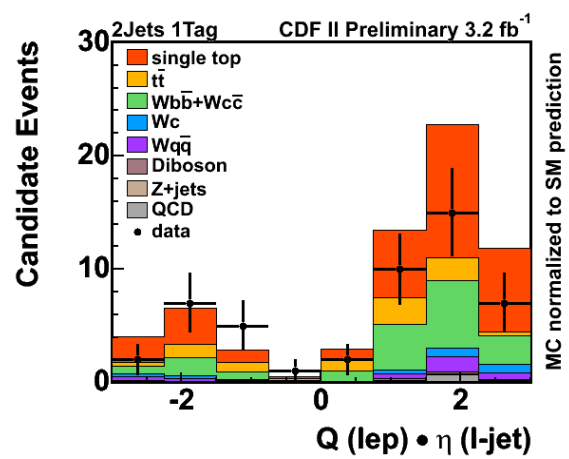


Example: $Q_{lepton} * \eta_{untagged\ jet}$ in CDF's single top analysis. Good separation power for t-channel signal.

Phys.Rev.D82:112005 (2010)







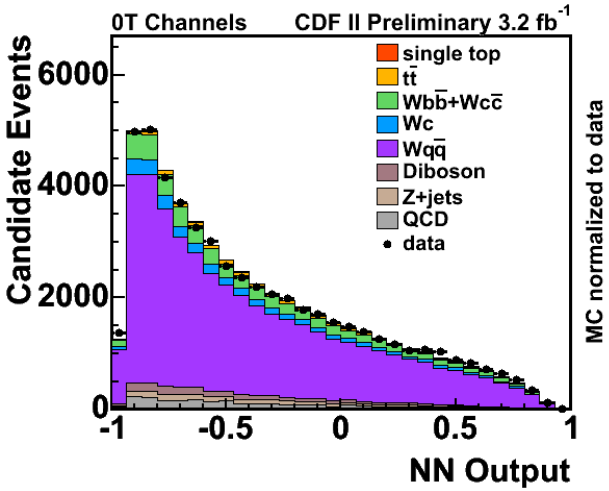
highest |η| jet as a well-chosen proxy

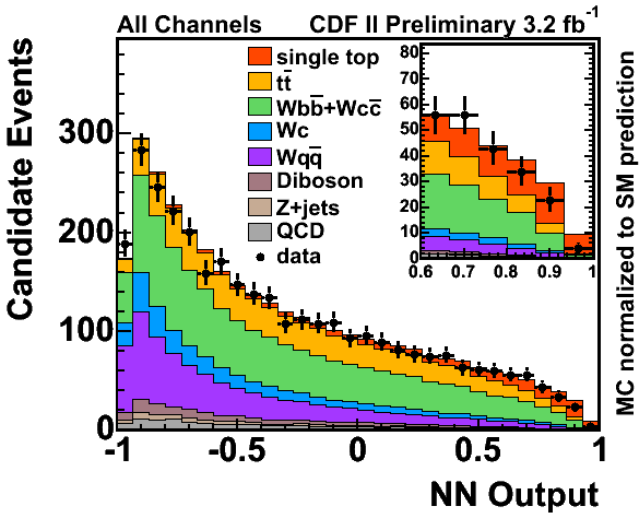
Checking MVA Output Distributions

- 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



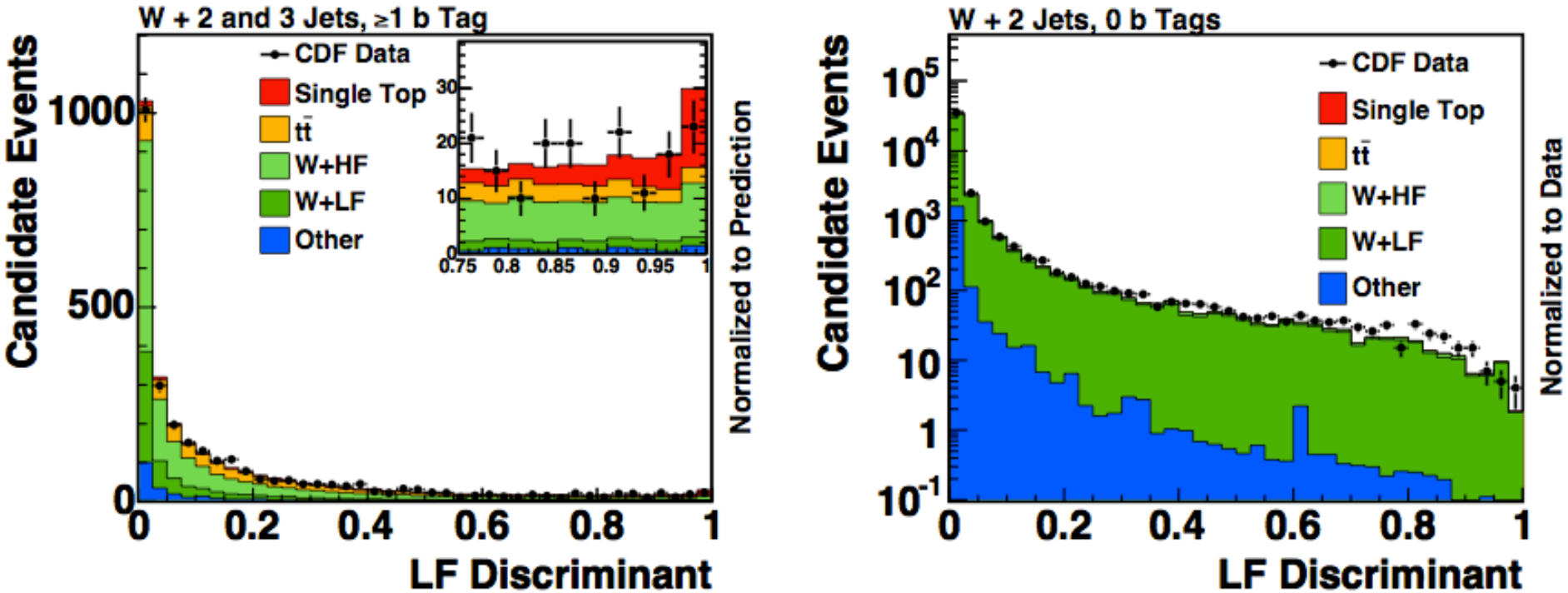


Phys.Rev.D82:112005 (2010)

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)

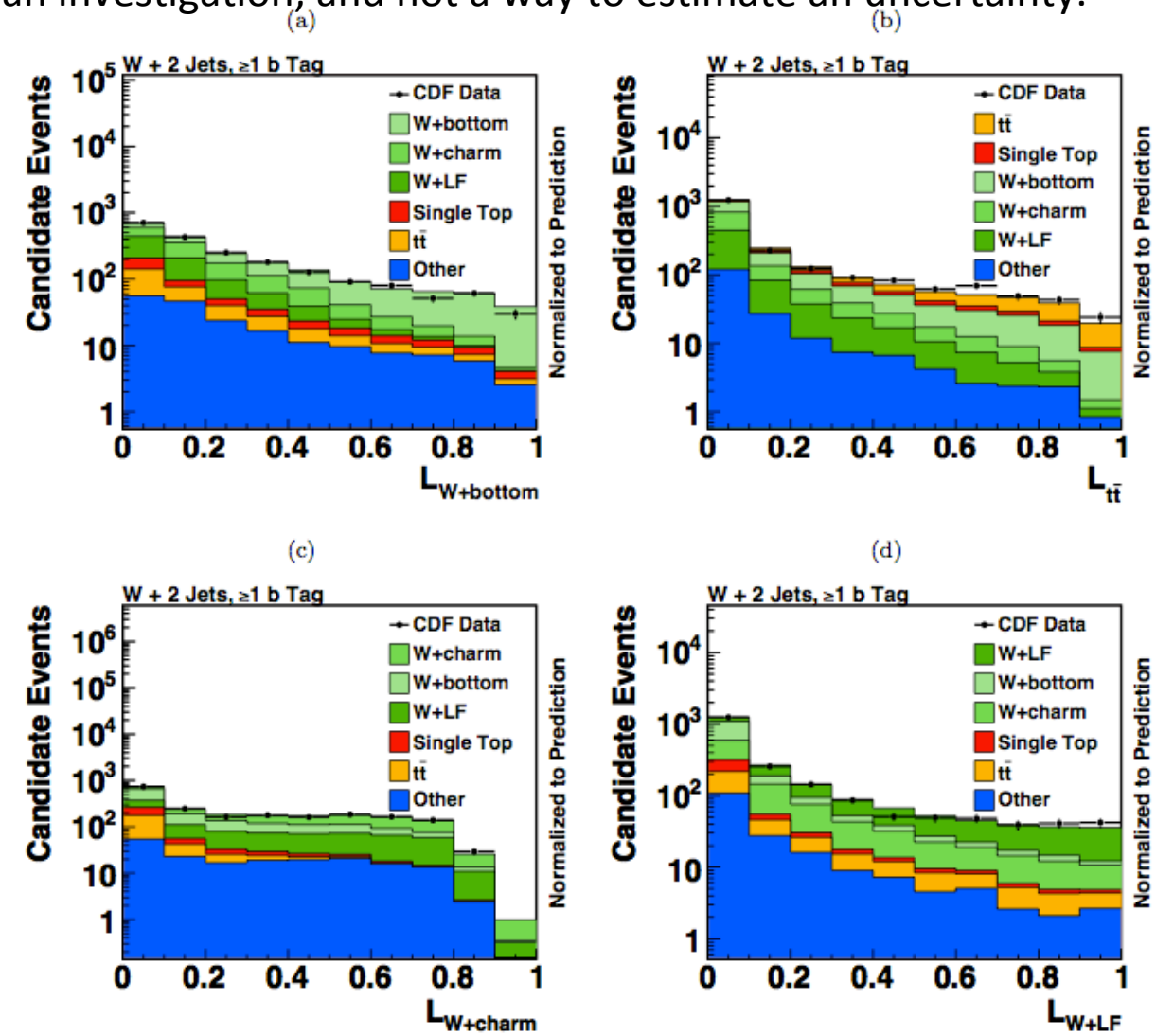


Strategy: Assess a shape systematic covering the difference between data and MC – extrapolate the uncertainty from the control sample to the signal sample.

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).

CDF's 7.5 fb^{-1} Single Top MVA Systematic Uncertainties

Nuisance Parameters
Listed by Name

Rate and Shape Uncertainties

Source of Uncertainty	Rate	Shape	Processes affected
Jet energy scale	0–8%	X	all
Initial and final state radiation	0–6%	X	single top, $t\bar{t}$
Parton distribution functions	0–1%	X	single top, $t\bar{t}$
Acceptance and efficiency scale	1–7%		single top, $t\bar{t}$, diboson, Z/γ^*+jets
Luminosity	6%		single top, $t\bar{t}$, diboson, Z/γ^*+jets
Jet flavor separator		X	all
Mistag model		X	W+light
Non-W model		X	Non-W
Factorization and renormalizatio		X	$Wb\bar{b}$
Jet η and ΔR distribution		X	W+light
Non-W normalization	40%		Non-W
$Wb\bar{b}$ and $Wc\bar{c}$ norm	30%		$Wb\bar{b}$, $Wc\bar{c}$
Wc normalization	30%		Wc
Mistag normalization	10–20%		W+light
$t\bar{t}$ normalization	8%		$t\bar{t}$
Monte Carlo generator	3–7%		single top, $t\bar{t}$
Single top normalization	7%		single top
Top mass	2-12%	X	single top, $t\bar{t}$

$*$ X indicates the sources of uncertainty from shape variation
$*$ Sources listed below double line are used only in $|V_{tb}|$ measurement

Example MVA Methods

Coded up in TMVA – comes with recent versions of ROOT

- Feed-Forward Neural Networks (multi-layer perceptrons)
  Abbreviations: NN, ANN, MLP
- Boosted Decision Trees
- Matrix Elements

All are just functions of the reconstructed event observables.

We could devise our own functions if it suited our needs and we were smart enough.

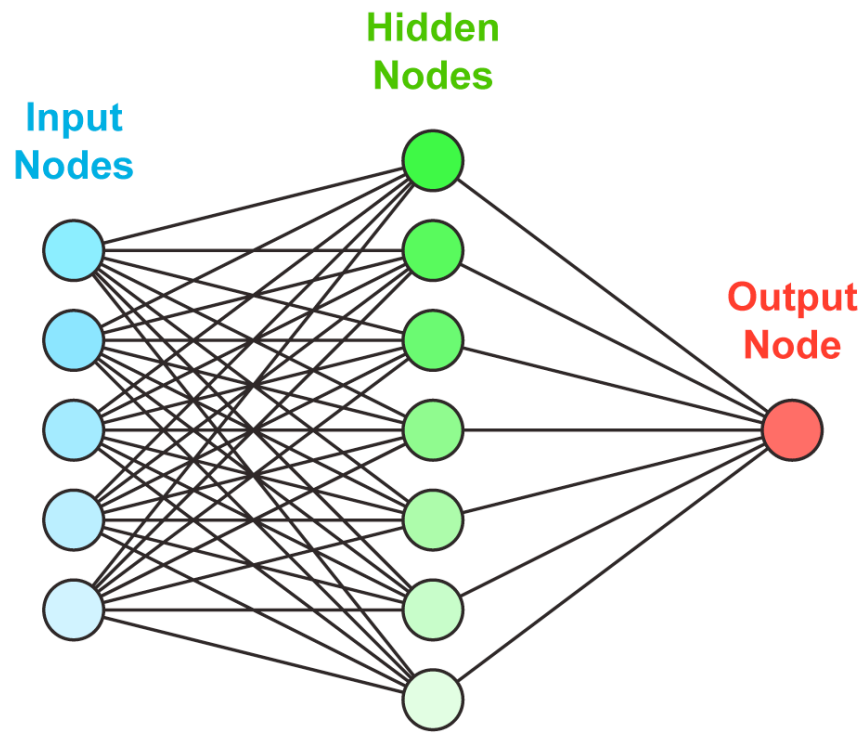
These are machine derived, so we call it ***machine learning.***

See, for example, P. Bhat, **Ann.Rev.Nucl.Part.Sci. 61 (2011) 281-309**

A Neural Network

Inputs to node *i* have weights w_i. Outputs are sigmoid functions of the weighted inputs.

$$o_i = S\left(\sum_i w_{ij} v_j\right)$$

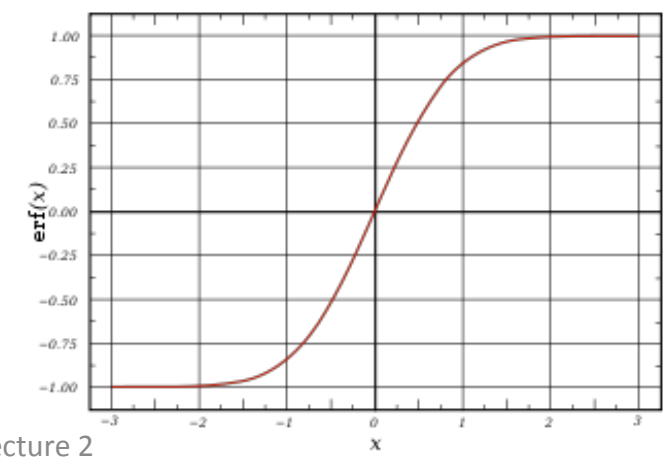


S is any of these:

S(x) = $\mathrm{Tan}^{-1}(x)$,
$x/\mathrm{sqrt}(1+x^2)$
$1/(1+\exp(-x))$
$\tanh(x)$

Or any other s-shaped function

Main features: Nonlinearity, monotonicity



Training a Neural Network

The weights w_{ij} are arbitrary. We may choose them, as well as the structure of the network, to optimize our analysis.

We would like to classify events as signal (output = 1) or background (output = 0).

Ad-hoc figure of merit: Minimize the sum of squares of errors made by the network:

$$E = \sum_{events} \left(O_{desired} - O_{obtained} \right)^2$$

Why this function?

Well, it's easy to differentiate with respect to the weights for each event.

Back-propagation training: Loop over training events (some signal, some background) and adjust the weights each time according to how the adjustment will improve *E*.

Weighted events are okay with most MVA training programs. But it's worth checking to see how they respond to negative-weight events!

Adjustable parameters: "learning rate" – how big the steps in w_{ij} are scaled by the derivative. How many events to use to train, how many spins through the training sample to use ("epochs")

Training a Neural Network

Critique of standard Neural Networks:

- No one really cares about $E = \sum_{events} \left(O_{desired} - O_{obtained} \right)^2$

  We care about the best expected uncertainty
  on cross section or property measurements
  Best expected limits if a particle is not there
  Best expected chances of discovery if a particle is there

- Addition of non-useful variables (random noise) can hurt overall performance
- Inputs can have very broad ranges of behavior
  discrete, large ranges, small ranges, mixtures ..

  (can be mitigated by clever preprocessing)
- Advantages – can make use of correlations between input variables by forming nearly arbitrary functions of them.

- Experience with them shows that it is usually better to
  - Give it the best variables already as inputs
  - Pre-select the data into samples so the NN has less work to do
    (fewer sources of backgrond that are imporant0

Training a Neural Network

Often the question arises: How big should the training samples be?

NN training figure of merit usually results in NN output = purity of the bin, normalized to training sample sizes.

Change the signal training fraction – change the purity of the total training sample.

But: Any invertible function of a discriminant has the same discriminating power as the original discriminant.

-- Corresponds to a rebinning of the output.

-- So no real need for variable-size bins, as long as you can transform the variable.

Desire – separate events in high s/b bins from those in lower s/b bins
-- adding bin contents with low s/b to higher s/b ones dilutes the sensitivity
Extreme limit – put everything in one bin. Not very sensitive! We're better off classifying events by categories than collecting them all together.

Overtraining

If a training sample is small, and the NN has many nodes and weights, it is possible for the NN to "learn" the properties of individual events in the training sample and get them classified correctly all the time.

This may not be representative of any other sample (like the data).

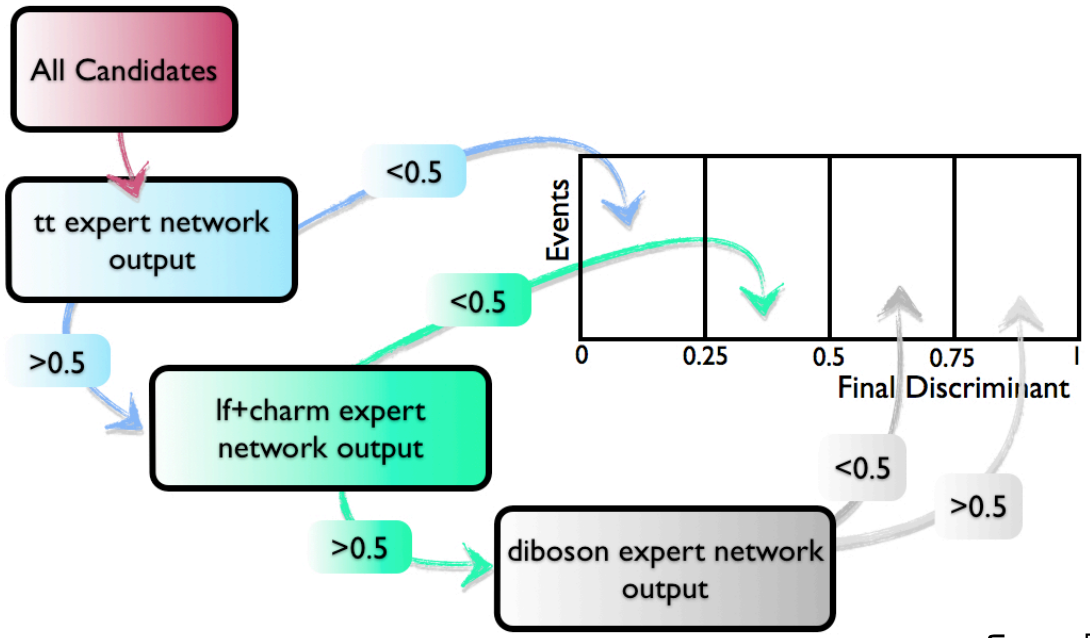
The network may not perform as well as it thinks it is performing if only the training sample is used to judge.

Ensure that overtraining does not affect correctness:
Use different events to train a NN and to test it.

Even if it's overtrained, then the independent evaluation of its performance is not systematically biased by this effect.

The NN may not be fully optimal, however.

Example of Giving NN's Some Help – Cascading NN Stages



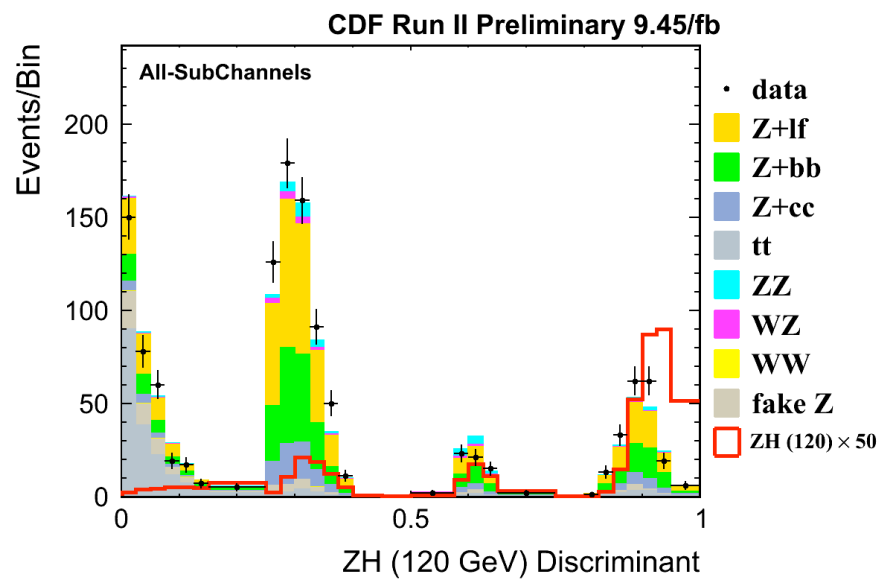
CDF's

ZH→llbb search

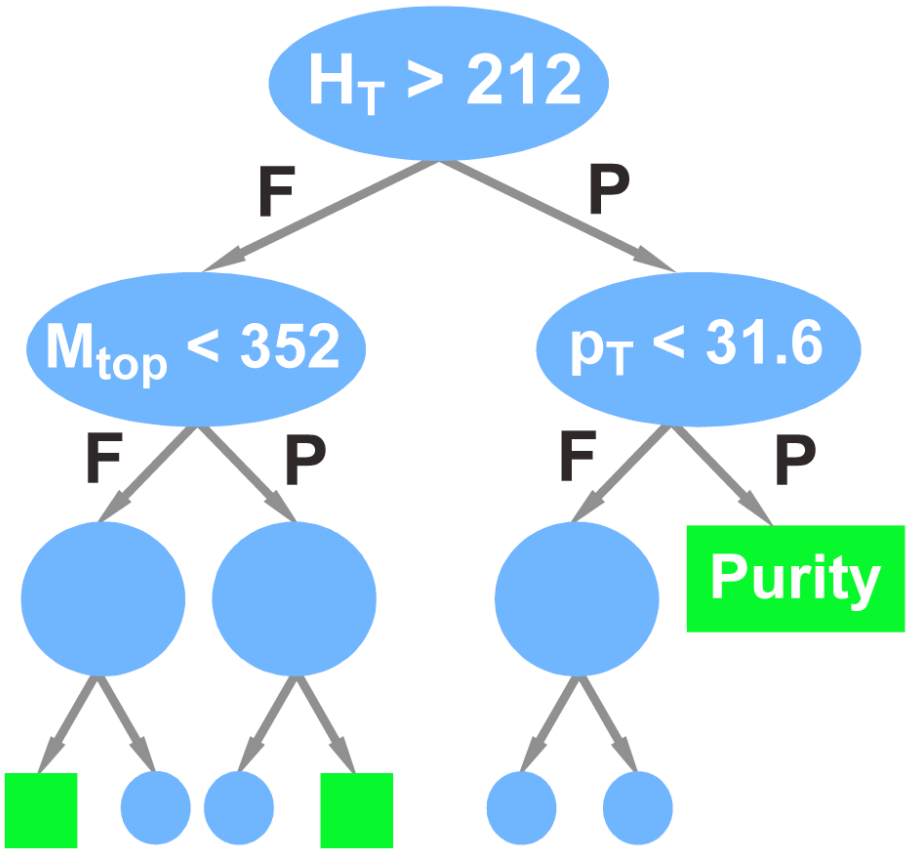
Further help:

Event selection is lljj, with m_{ll} near M_Z. One or two b-tags, with loose or tight b-tagging requirements.

Split sample up into b-tag categories:
Tight-Tight
Tight-Loose
Single Tight
Loose-Loose



(Boosted) Decision Trees



Original work by J. Friedman in the 1980's

Look through the list of input variables; Try sliding a cut along each one and find the cut on a variable that maximizes the purity difference on both sides of the cut.

"Gini index" – p(1-p), where p=purity zero for perfect separation.

Iterate the search for the best cut on the best variable for each subset of events thus divided. Stop when you run out of enough MC to predict the contents of a sample.

- Advantages over NN's: not as sensitive to the addition of "noise" variables – they just never get cut on
- The Gini index is also just a proxy for what we really care about.

Boosting Decision Trees

Decision tree training rather sensitive to random fluctuations.

Two cuts which are almost as good can get re-ordered in the training process based on the presence of a small number of training MC events.

The first cut has a profound impact on the behavior of the rest of the tree

Would like to retrain many trees and average the behavior – knock off the sharp edges.

Retrain after reweighting events that have been misclassified: Boost their weights so that further retrainings have a better shot at classifying them properly.

Sort bins by purity and average the resulting discriminants.

Matrix-Element Discriminants

- Calculate probability density of an event resulting from a given process

Phase space factor:
Integrate over unknown or poorly measured quantities

Parton distribution functions

$$P(p_l^\mu, p_{j1}^\mu, p_{j2}^\mu) = \frac{1}{\sigma} \int d\rho_{j1} d\rho_{j2} dp_\nu^z \sum_{comb} \phi_4 \, | M(p_i^\mu) |^2 \frac{f(q_1) f(q_2)}{| q_1 || q_2 |} W_{jet}(E_{jet}, E_{part})$$

Inputs:
lepton and jet 4-vectors – no other information needed!

Matrix element:
Different for each process. Leading order, obtained from MadGraph

Transfer functions:
Account for detector effects in measurement of jet energy

- The input variables are the same for all matrix elements – adding a new matrix element requires more calculation but does not use any different information from the data

Matrix-Element Discriminants

In principle, nothing performs better than these.

If processes cannot be separated because they contribute to the final state in the same way, this is all there is.

BUT:

- Four-vectors are imperfectly measured. Transfer functions are also imperfect.

- Only the modeling needs systematics; construction of the discriminant does not incur additional systematics, so even if the discriminant is imperfect or naive, it's okay – just an optimization question.

- Matrix elements are usually leading-order only.

- Particles are sometimes not reconstructed at all, even when they should be

- Some processes do not have well defined matrix elements – like data-derived fakes.

- Non-kinematic information is important, too, such as b-tags (help reduce combinatorics)

- Not clear whether integrating over all possibilities or just picking the best one is the most optimal for the purposes we set out for (more on this later).

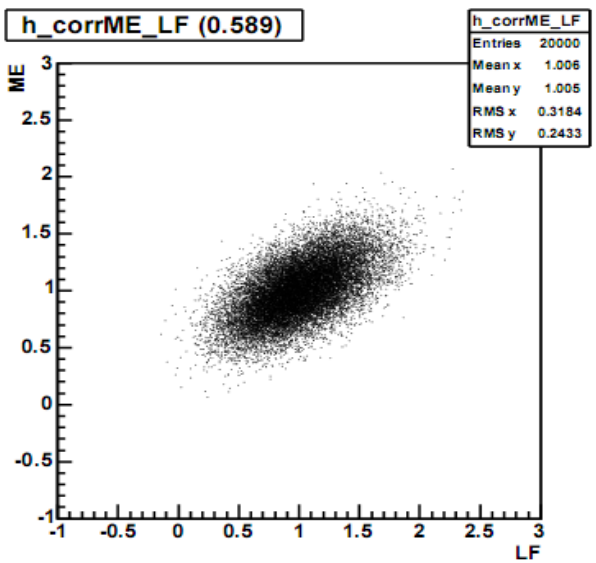
Several Analyses on the Same Data

- 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!

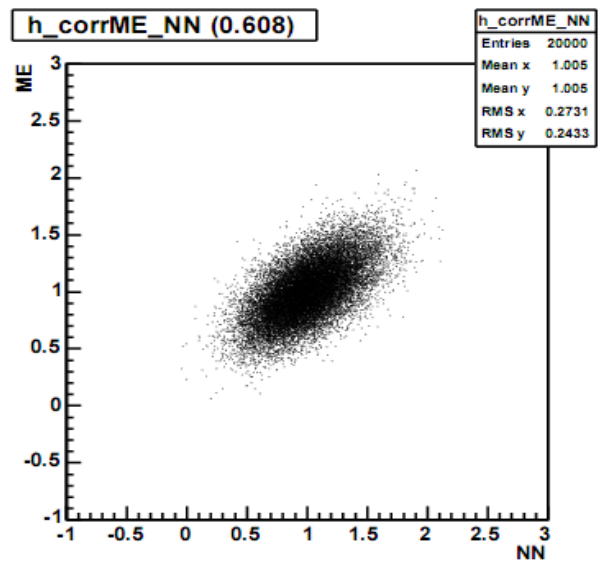
- What to do?
  - Pick one and publish it – criterion: best sensitivity. Median expected limit, median expected discovery sensitivity, median expected measurement uncertainty. How to pick it if the result is 2D? Need a 1D figure of merit.
  - Can check consistency with pseudoexperiments. A p-value using 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

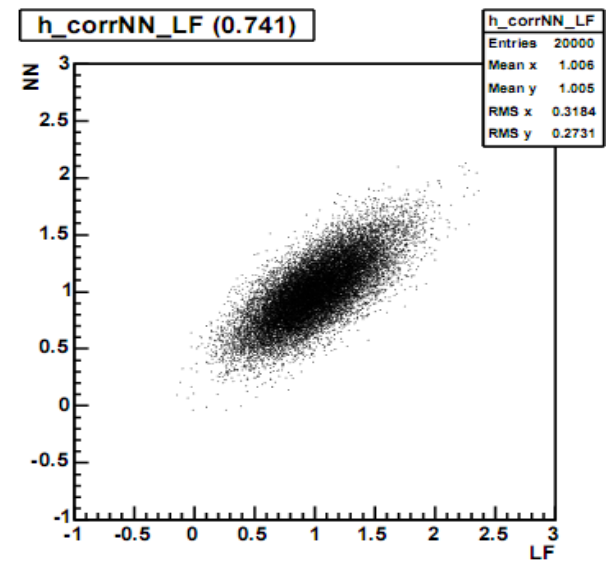
LF-ME 58.9%



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..)