What have we learned from the Tevatron for the LHC

Yuri Gershtein

with shamelessly stolen slides from too many people to list here

Outline

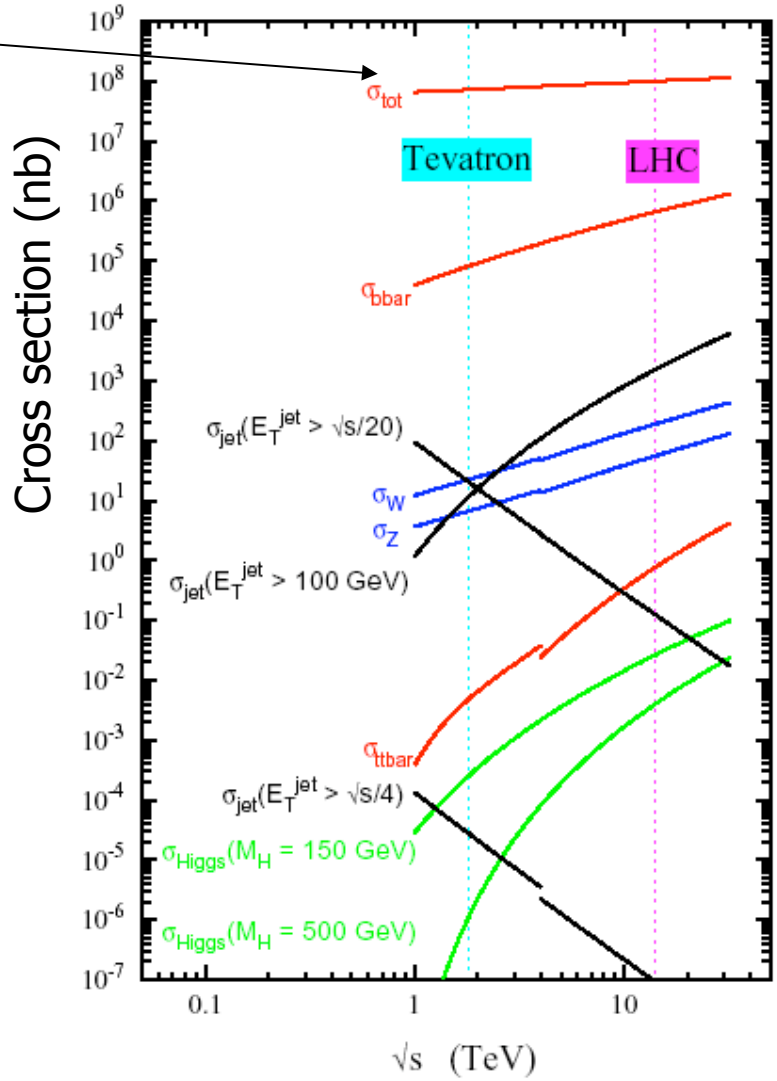
- Tevatron: brief history
- physics at hadron colliders
  - particle detection, trigger, reconstruction, …
- QCD lessons
  - pdf's, NLO, NNLO, …
- Flavor lessons:
  - b-physics at hadron collider is possible
- Precision measurements
  - W mass
- Top quark physics
- Advanced analysis: multivariate methods
  - top and higgs
- Will not talk about the new phenomena searches
  - techniques are the same, and no discoveries have been made so far

Summary of lessons so far

- LHC will be a 20-30 year program. Be patient!
  - although the delays affected careers of young people and are generally quite frustrating
- Hadron colliders are very messy (but the way to get to the energy frontier)
  - underlying event
  - large occupancies
  - huge total cross-sections – pile-up
  - trigger shapes everything
- Yet, it is possible to do precision measurements!

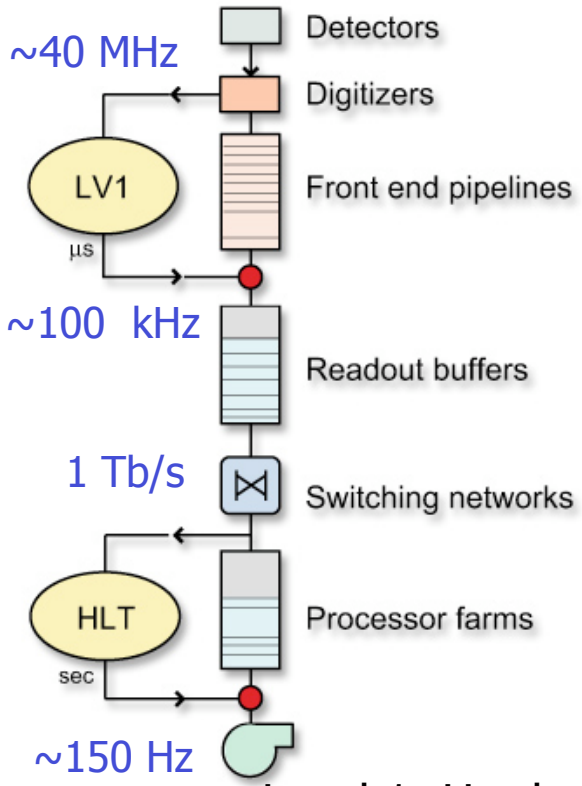
Trigger: Selecting the interesting events (I)

- Our starting point is here
  - At the LHC the rate for all collisions is 40MHz!
  - Although ideal, it's impossible to keep all the events
- Need to decide a priori which are the "interesting" events to keep/filter
- Need to be selective
  - enhance rare processes
  - reduce common ones
- If we make bad/unwise choices we will throw away the new physics!
  - If you don't trigger on it, it's gone forever!
- Theory plays a role in guiding these choices
  - Important to have good communication between theorists and experimentalists for coming up with new triggers
- Physics priorities of collaboration is another consideration…

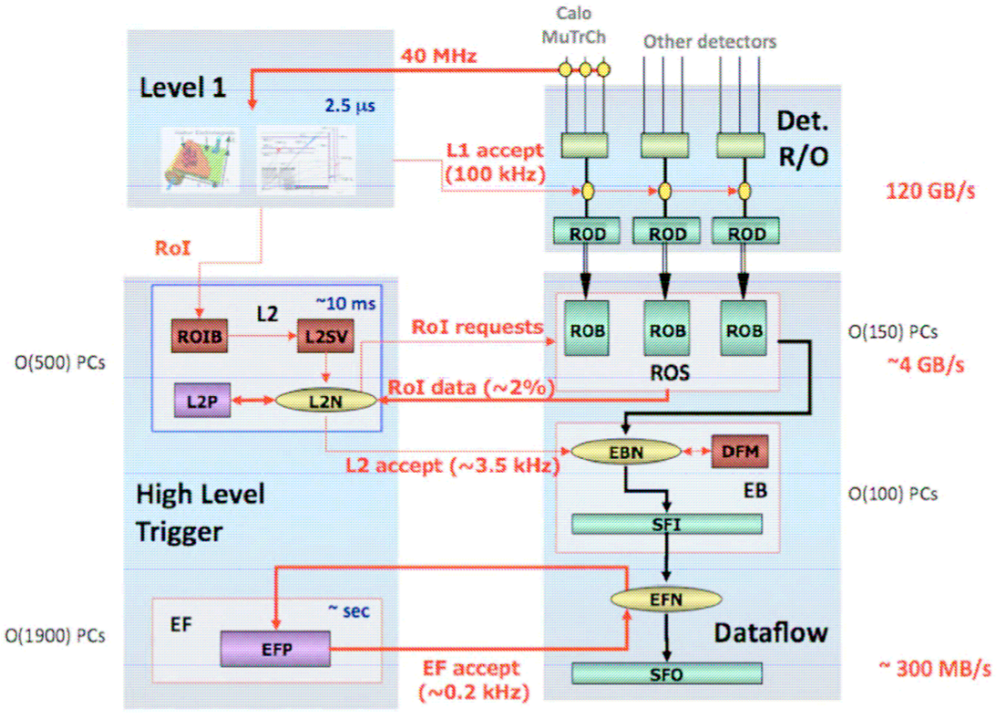


CMS and ATLAS Triggers

CMS



ATLAS



Level 1: Hardware based (electronics)
Level 2: Software based
The decision to keep ~1/200,000 events happens every second.
No room for mistakes!

Tevatron Trigger menu

- Level 1:
  - crude reconstruction of tracks, calorimeter clusters, muon tracks
  - some spatial matching between sub-systems
- Level 2:
  - silicon IP information
  - refinement of selection (i.e. topological cuts)
- Level 3:
  - full detector information, basically a simplified reconstruction results are available
- Typical triggers

  jet, multijet, acoplanar jets, jets + MET

  single electron/photon/muon/tau

  two objects (ee, e+mu, etc)

  Rate for "low pT" physics is high, and one needs to be inventive to keep rejection high

Trigger: limitations

- Physics at hadron colliders is in many respects similar to looking only under a lamp post.

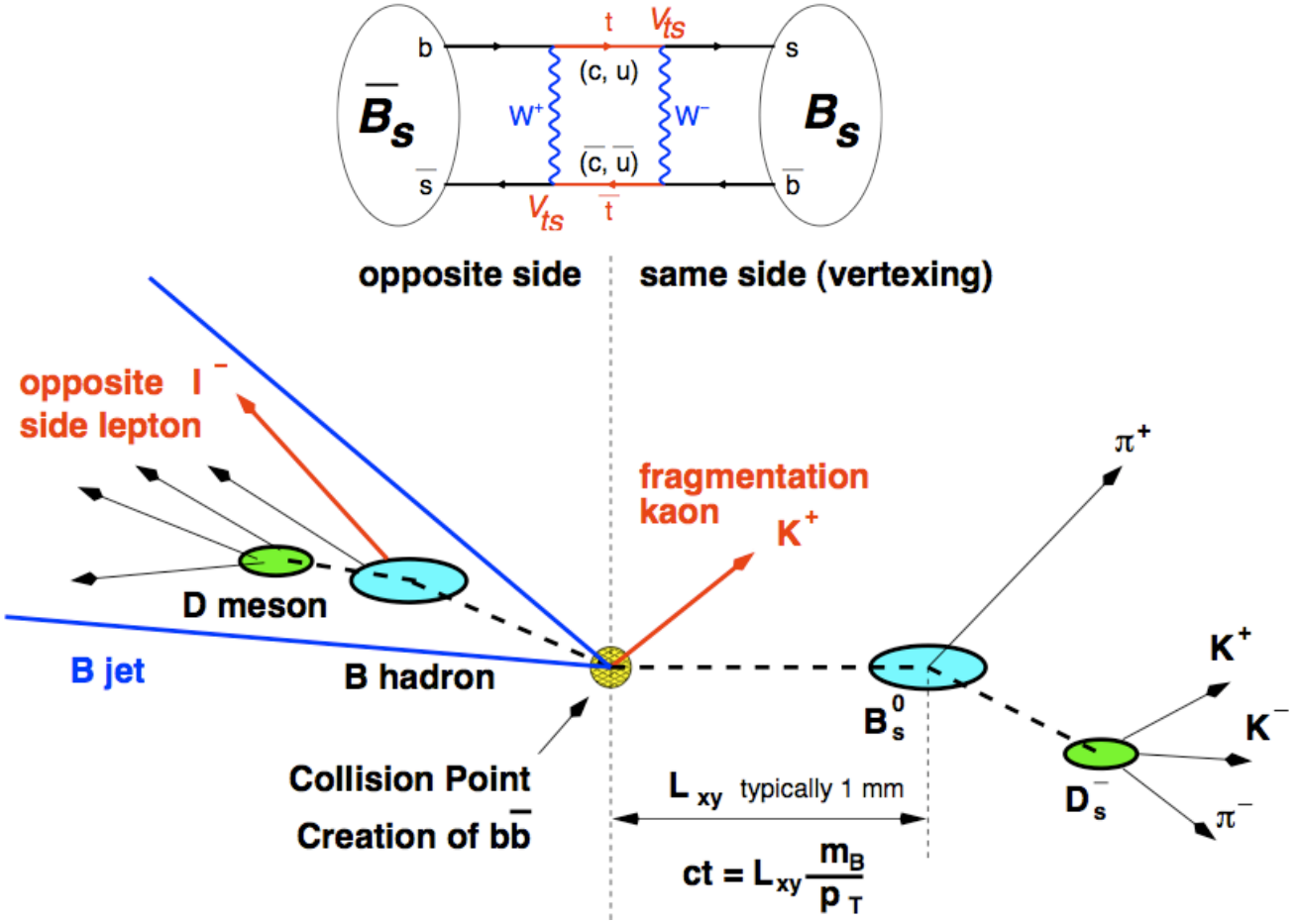
- Since the backgrounds are high, any non-standard signature will fail the trigger unless a new specific trigger is designed
  - i.e. long-lived particles
- Some signatures call for very low pT thresholds
  - some new phenomena
  - flavor physics

Tevatron Triggers for flavor physics

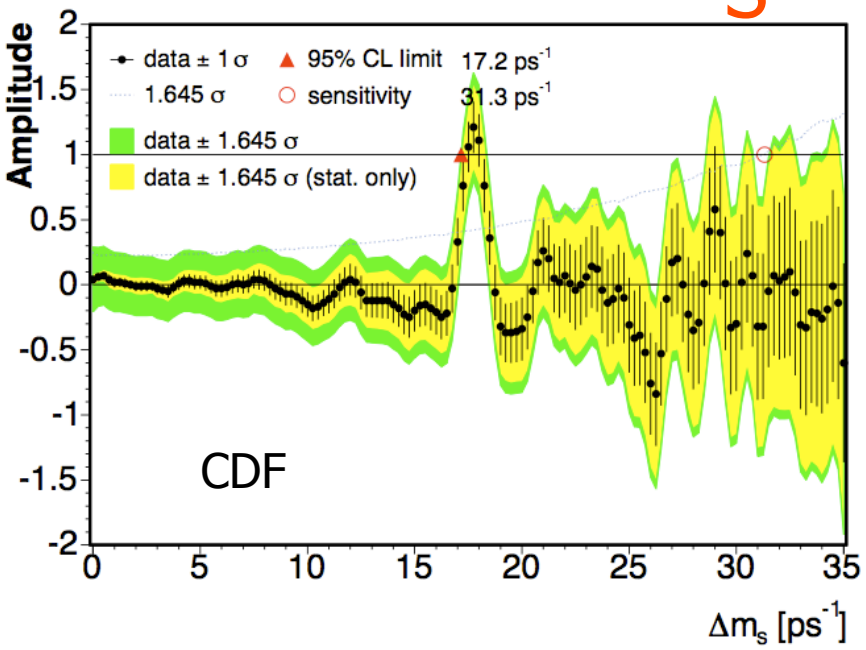
- Level 1: limited options
  - one or two low p_T muons (trigger threshold is the key)
  - two tracks (CDF only)
- Level 2:
  - silicon IP information
- Level 3:
  - particle combinations, mass windows, etc…

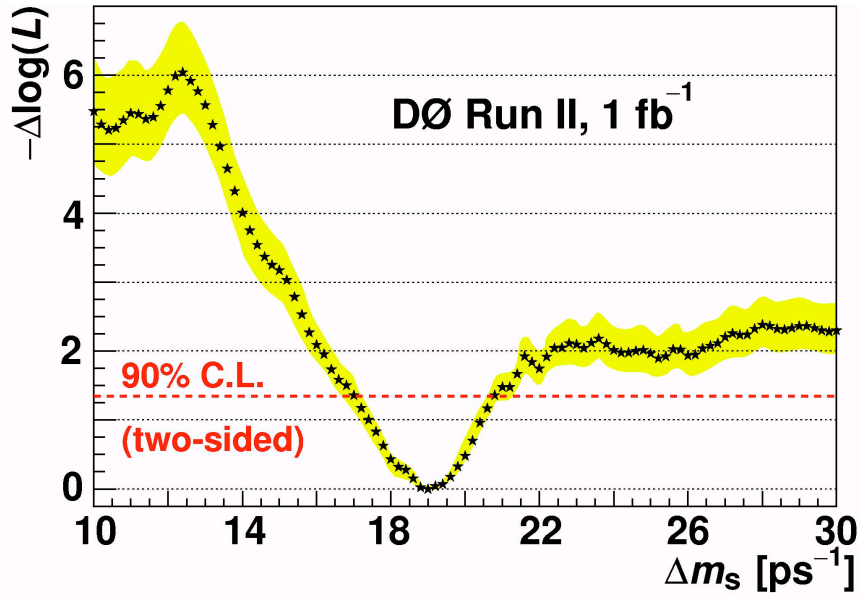
B_S (B_d) mixing

- Gives access to V_{ts} (V_{td}) and sensitive to new physics
- Interference between $B_{d,s} \to \overline{B}_{d,s} \to X_{CP}$ and $B_{d,s} \to X_{CP}$ provides a window to CP violation



B_S mixing

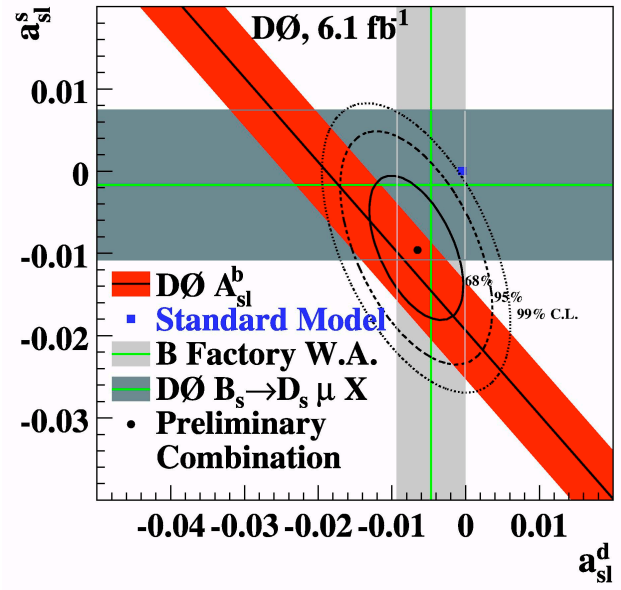




- Recent DØ result – deviation from SM prediction of dilepton charge assymetry: more $\mu^+\mu^+$ pairs are produced compared to $\mu^-\mu^-$

$$A_{\rm sl}^b = -0.00957 \pm 0.00251\ ({\rm stat}) \pm 0.00146\ ({\rm syst})$$

$$A_{\rm sl}^b(SM) = (-2.3^{+0.5}_{-0.6}) \times 10^{-4}$$

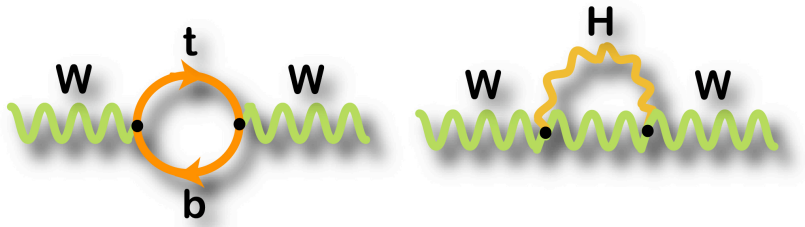


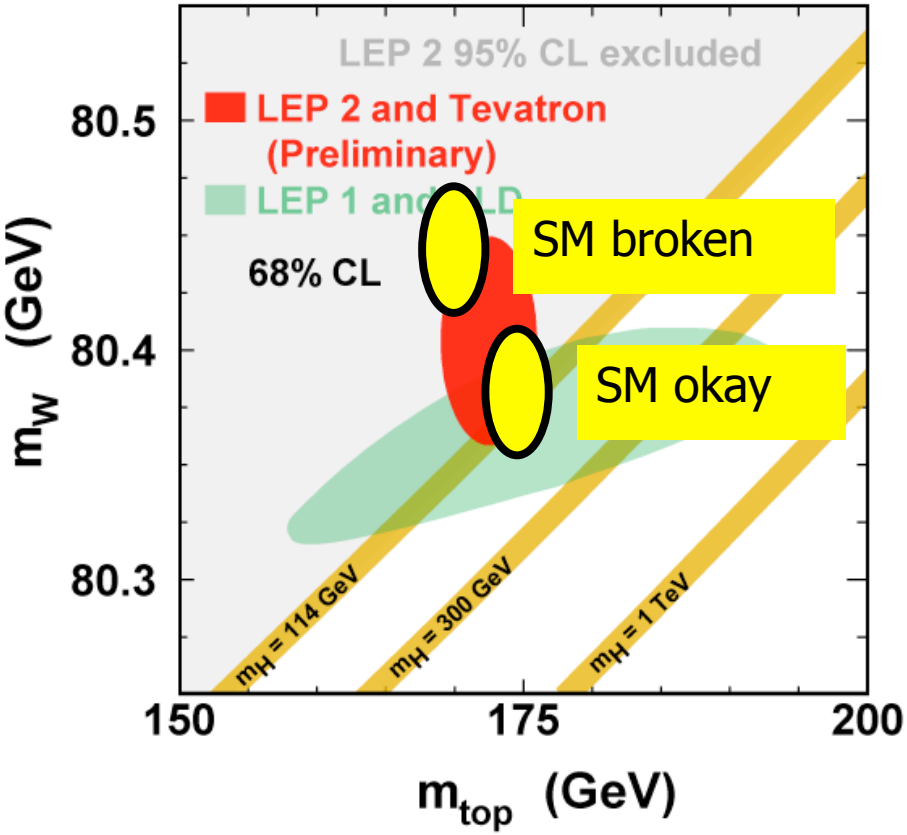
Precision Measurement of Electroweak Sector of the Standard Model

- **W boson mass**
- **Top quark mass**
- **Implications for the Higgs boson**

The W boson, the top quark and the Higgs boson

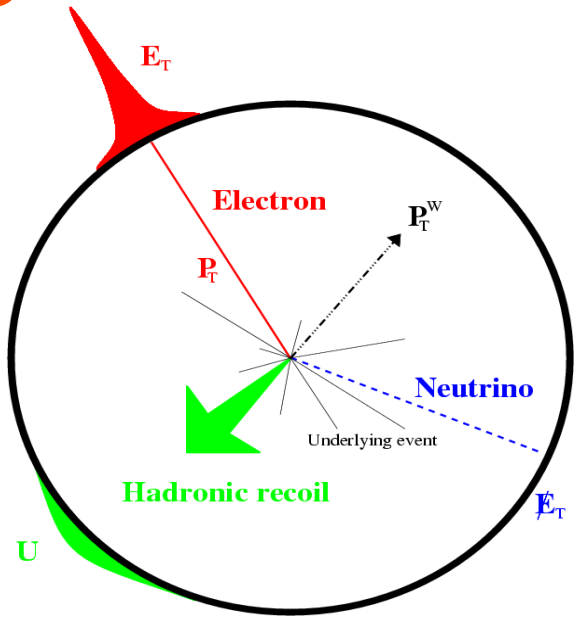
- Top quark is the heaviest known fundamental particle
  - Today: m_{top}=173.1±1.3 GeV
  - Run 1: m_{top}=178±4.3 GeV/c^2
  - Is this large mass telling us something about electroweak symmetry breaking?
    - Top Yukawa coupling:
    - <H>/(√2 mtop) = 1.005 ± 0.008
- Masses related through radiative corrections:
  - $m_W \sim M_{top}^2$
  - $m_W \sim \ln(m_H)$
- If there are new particles the relation might change:
  - Precision measurement of top quark and W boson mass can reveal new physics





W Boson mass

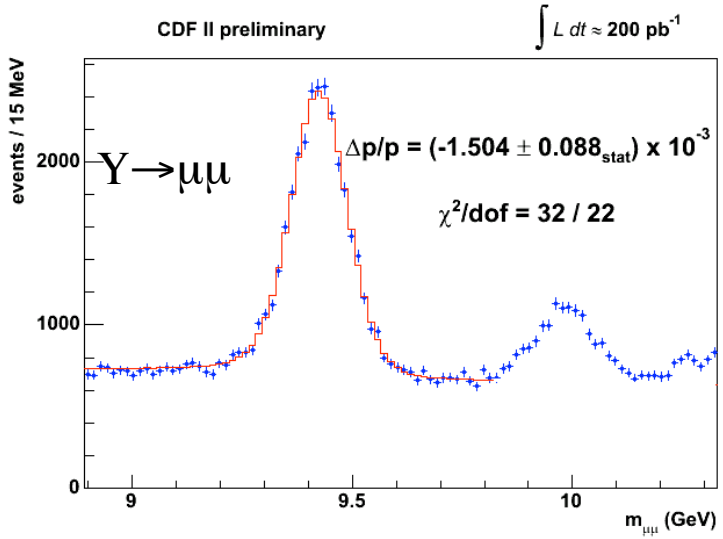
- Real precision measurement:
  - LEP: M_W=80.367±0.033 GeV/c^2
  - Precision: 0.04%
    - => Very challenging!
- Main measurement ingredients:
  - Lepton p_T
  - Hadronic recoil parallel to lepton: $u_{||}$
  - Missing ET
- Z→ll superb calibration sample:
  - but statistically limited:
    - About a factor 10 less Z's than W's
    - Most systematic uncertainties are related to the size of Z sample
      - Will scale with $1/\sqrt{N_Z}$ (=$1/\sqrt{L}$)

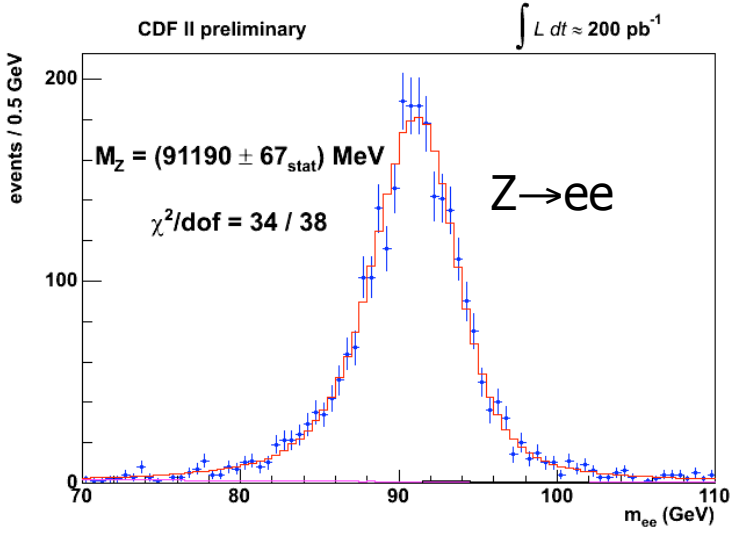


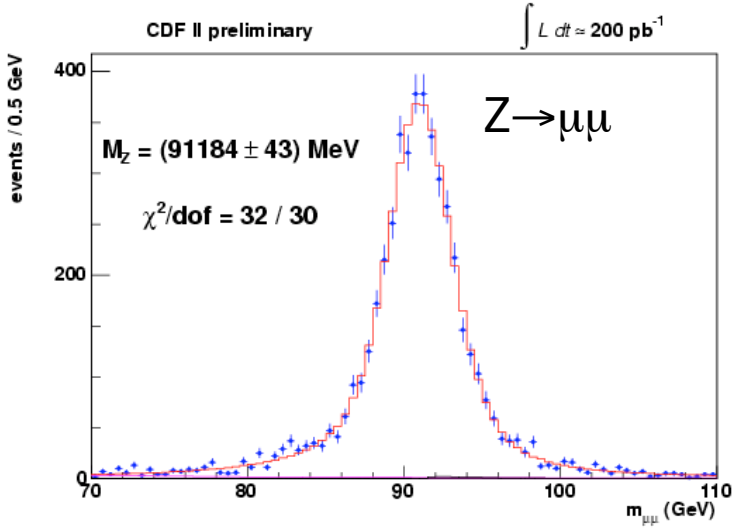
$$m_T = \sqrt{2p_T^l \not{p}_T\,(1-\cos\Delta\phi)},$$
$$\not{p}_T \approx |p_T + u_{||}|$$

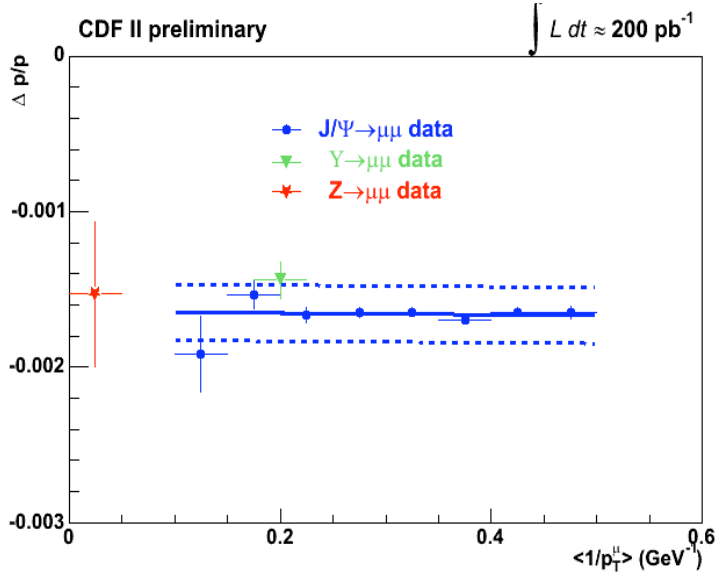
$$m_T \approx 2p_T\sqrt{1+u_{||}/p_T} \approx 2p_T + u_{||}$$

Lepton Momentum Scale









- **Systematic uncertainty on momentum scale: 0.04%**

Systematic Uncertainties

m_T Fit Uncertainties			
Source	$W \to \mu\nu$	$W \to e\nu$	Correlation
Tracker Momentum Scale	17	17	100%
Calorimeter Energy Scale	0	25	0%
Lepton Resolution	3	9	0%
Lepton Efficiency	1	3	0%
Lepton Tower Removal	5	8	100%
Recoil Scale	9	9	100%
Recoil Resolution	7	7	100%
Backgrounds	9	8	0%
PDFs	11	11	100%
W Boson p_T	3	3	100%
Photon Radiation	12	11	100%
Statistical	54	48	0%
Total	60	62	-

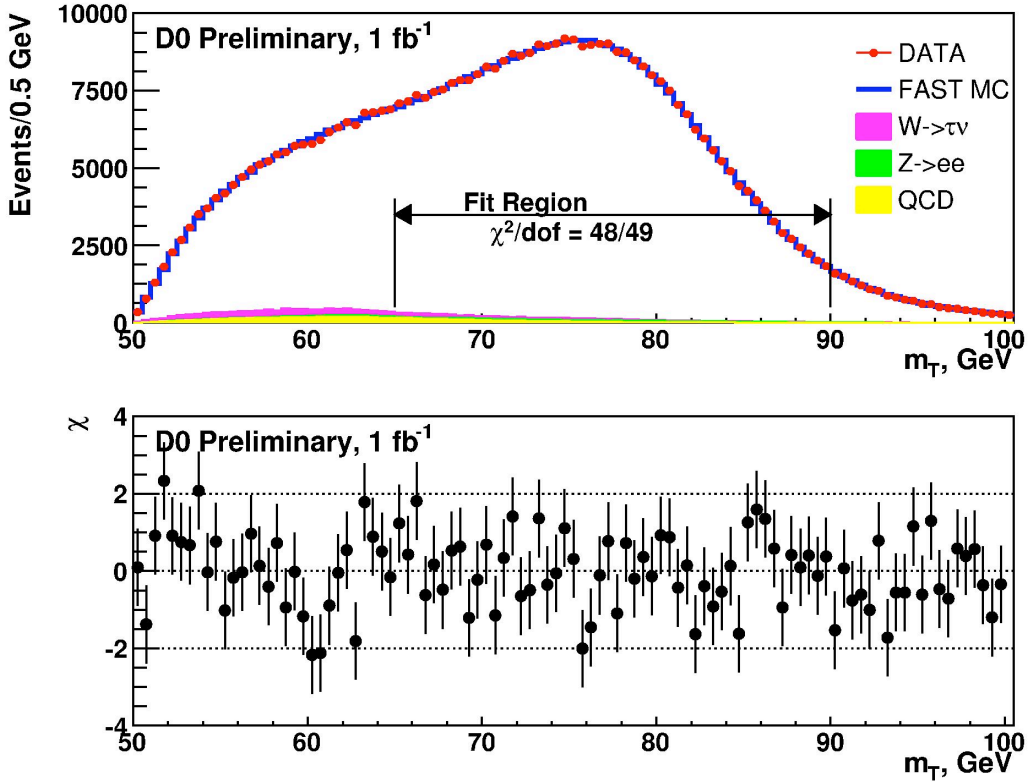
TABLE IX: Uncertainties in units of MeV on the transverse mass fit for m_W in the $W \to \mu\nu$ and $W \to e\nu$ samples.

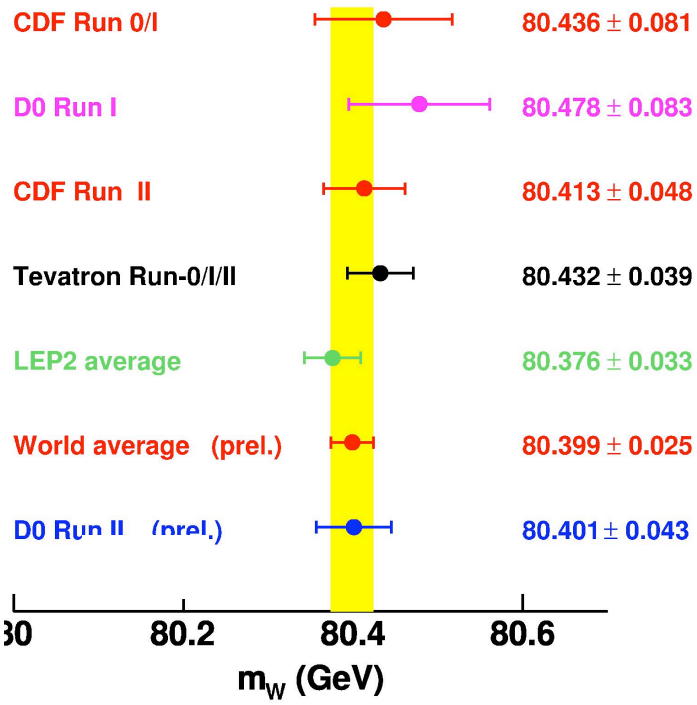
Limited by data statistics

Limited by data and theoretical understanding

- Overall uncertainty 60 MeV for both analyses
  - Careful treatment of correlations between them
- Dominated by stat. error (50 MeV) vs syst. (33 MeV)

W Boson Mass





New world average:

M_W=80399 ± 23 MeV

Ultimate precision:

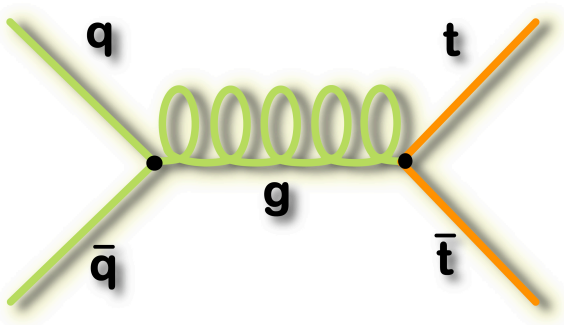
Tevatron: 15-20 MeV

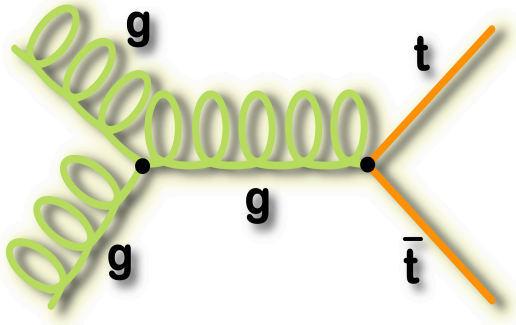
LHC: unclear (5 MeV?)

Top Quark Production and Decay

- At Tevatron, mainly produced in pairs via the strong interaction

85%





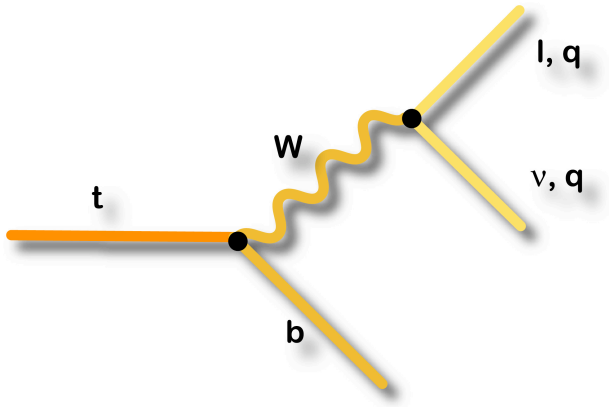
15%

- Decay via the electroweak interactions $Br(t \rightarrow Wb) \sim 100\%$
  Final state is characterized by the decay of the W boson

Dilepton

Lepton+Jets

All-Jets



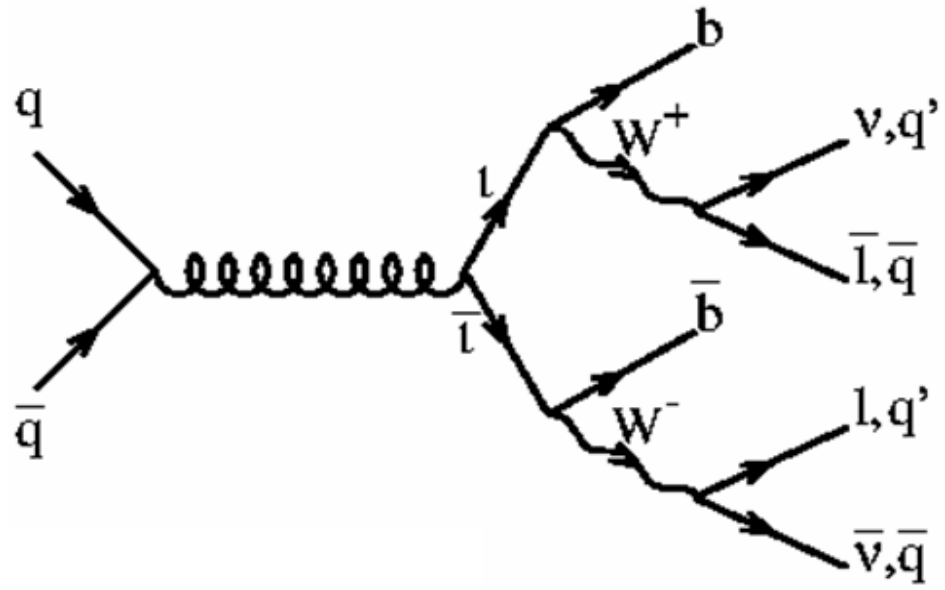
Different sensitivity and challenges in each channel

How to identify the top quark

SM: $t\bar{t}$ pair production, Br(t→bW)=100% , Br(W→lν)=1/9=11%

dilepton	***(4/81)***	**2 leptons + 2 jets + missing E_T**
l+jets	***(24/81)***	**1 lepton + 4 jets + missing E_T**
fully hadronic	***(36/81)***	**6 jets**

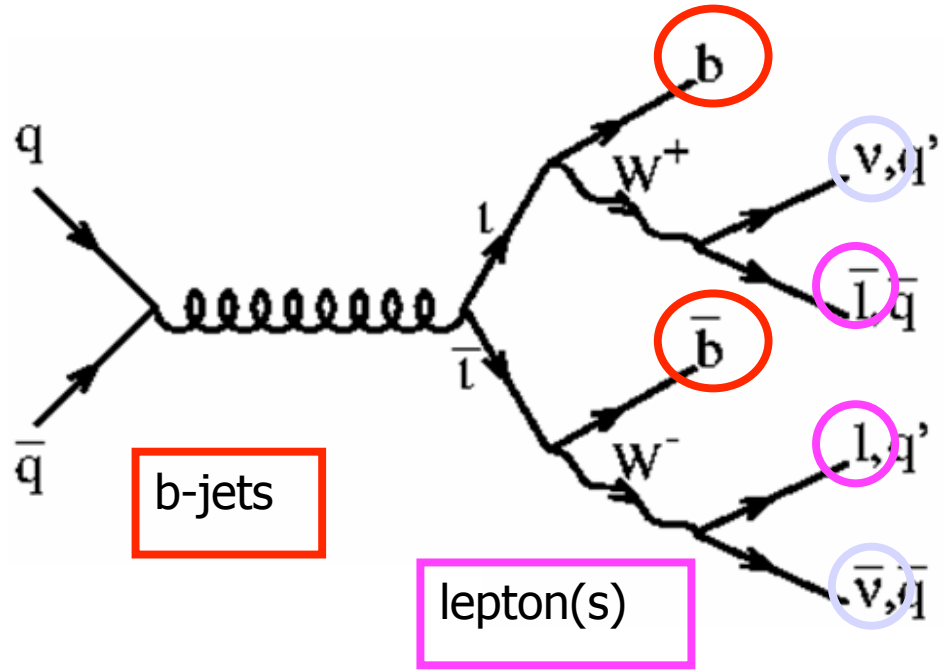
(here: l=e,μ)



How to identify the top quark

SM: $t\bar{t}$ pair production, Br(t→bW)=100% , Br(W->lν)=1/9=11%

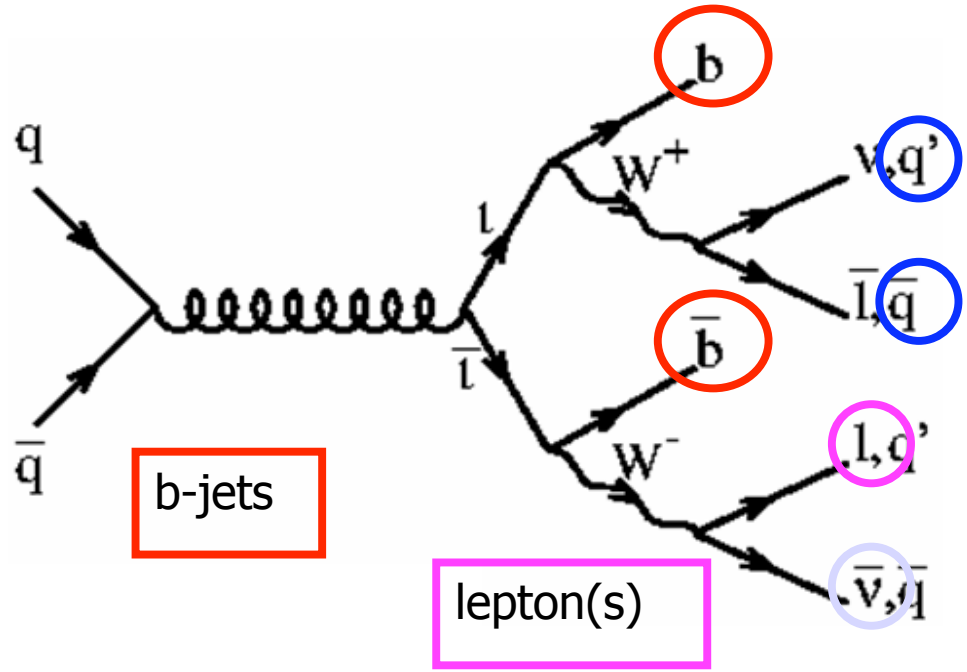
dilepton	***(4/81)***	**2 leptons + 2 jets + missing E_T**
lepton+jets	***(24/81)***	**1 lepton + 4 jets + missing E_T**
fully hadronic	***(36/81)***	**6 jets**



How to identify the top quark

SM: $t\bar{t}$ pair production, Br(t→bW)=100% , Br(W->lν)=1/9=11%

dilepton	***(4/81)***	**2 leptons + 2 jets + missing E_T**
lepton+jets	***(24/81)***	**1 lepton + 4 jets + missing E_T**
fully hadronic	***(36/81)***	**6 jets**



b-jets

lepton(s)

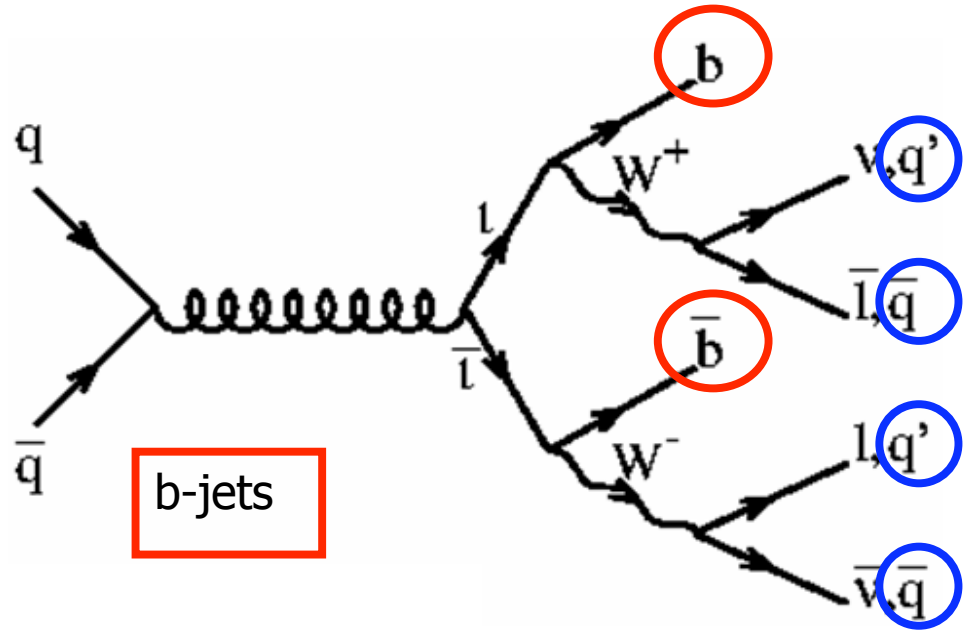
missing ET

more jets

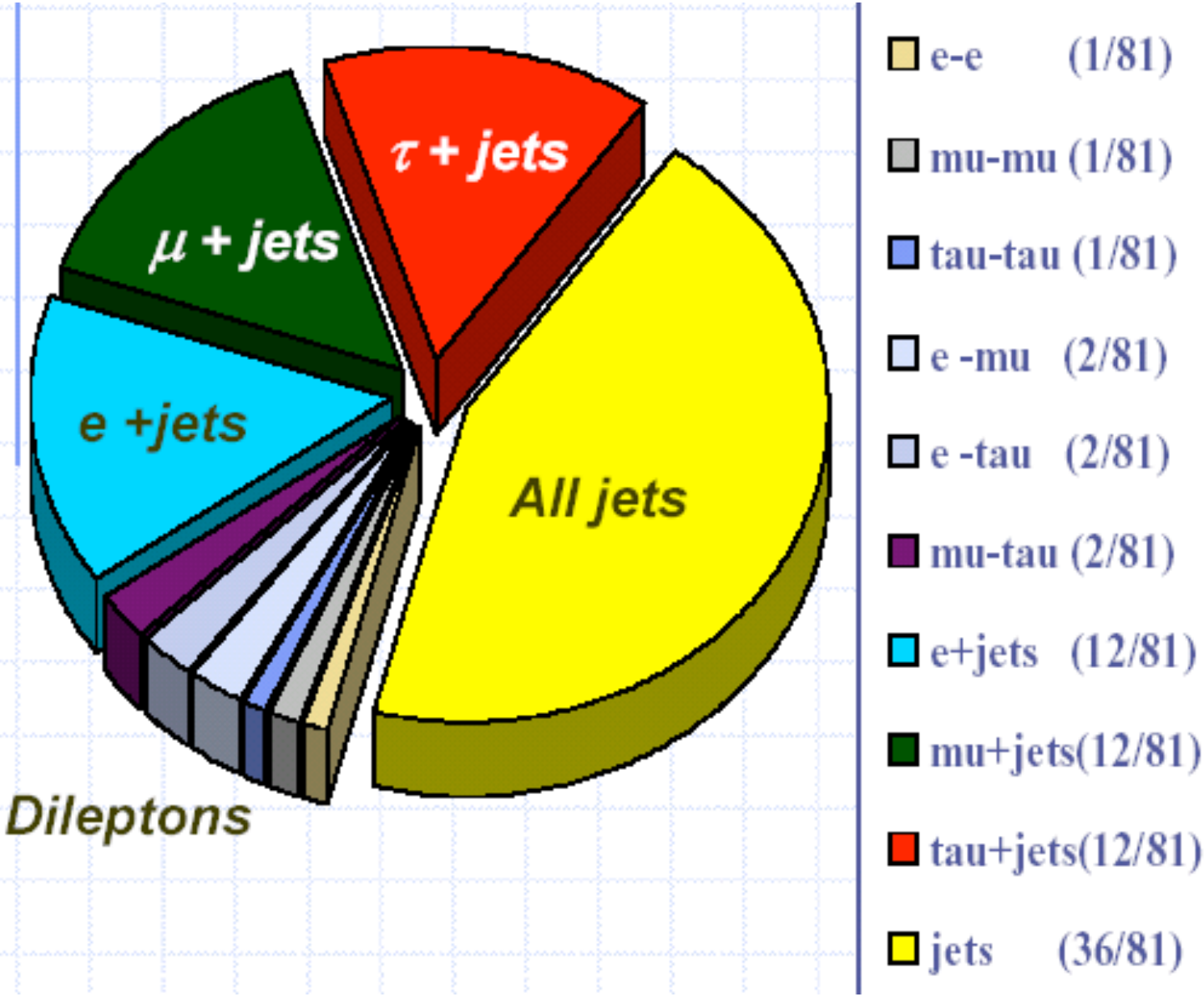
How to identify the top quark

SM: $t\bar{t}$ pair production, Br(t→bW)=100% , Br(W->lν)=1/9=11%

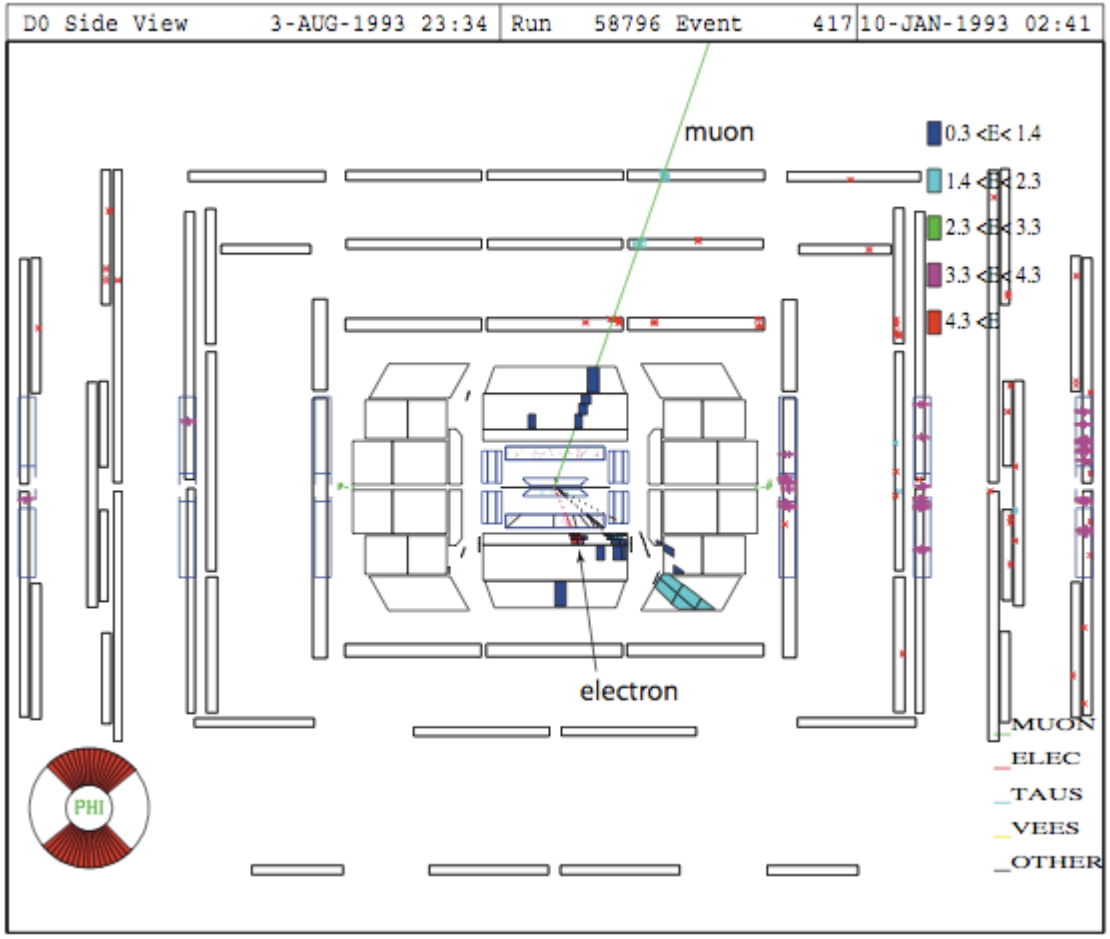
dilepton	***(4/81)***	**2 leptons + 2 jets + missing E_T**
lepton+jets	***(24/81)***	**1 lepton + 4 jets + missing E_T**
fully hadronic	***(36/81)***	**6 jets**



Top Event Categories



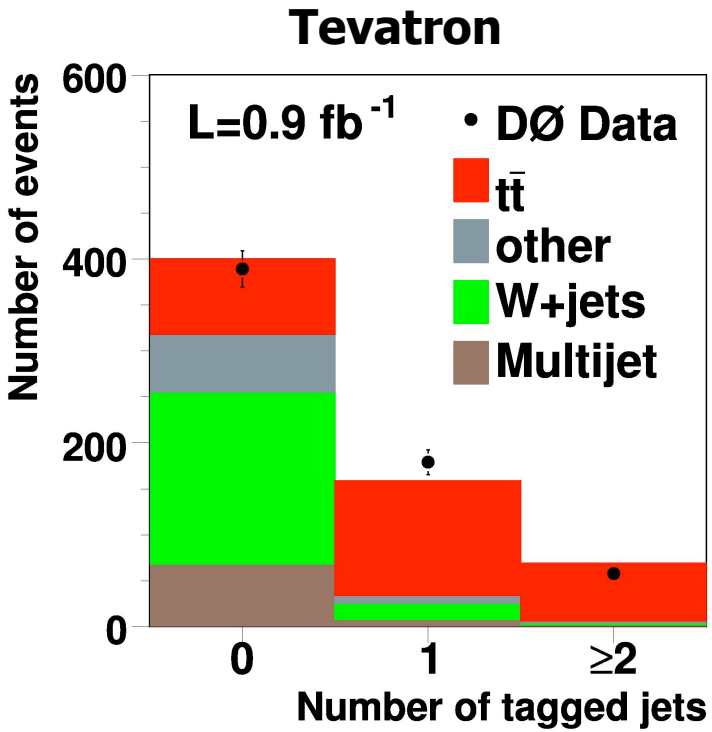
Power of one event

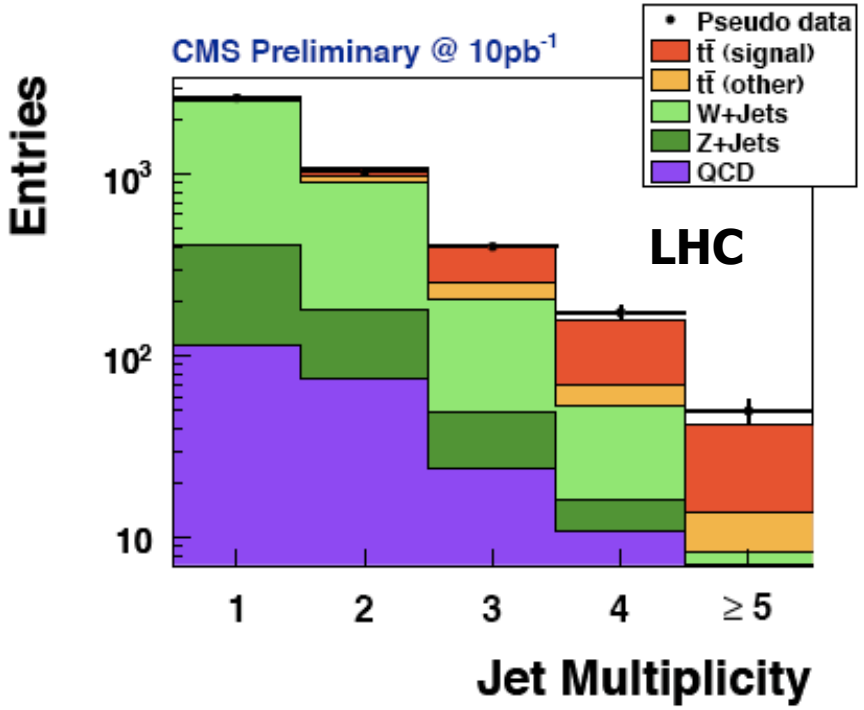


- found in 1993 at DØ
  - electron
  - muon
  - 3 jets
  - Missing ET
- survived all optimized and re-optimized cuts of all Run I analyses
- top mass was "measured" using this one event to be

  163 ± 36 GeV

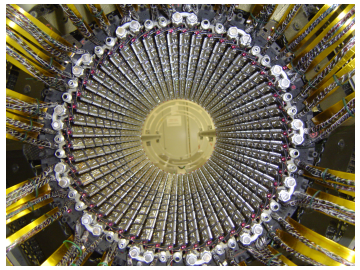
Finding the Top at Tevatron and LHC



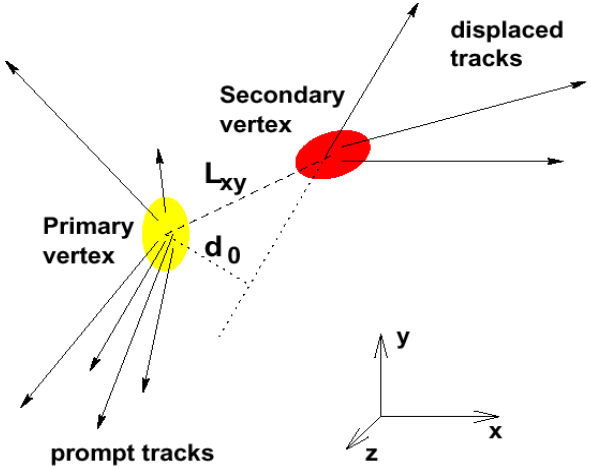


- Tevatron:
  - Top is overwhelmed by backgrounds:
  - Even for 4 jets S/B is only about 0.8
  - Use b-jets or topological analysis to purify sample
- LHC
  - Signal clear even without b-tagging: S/B is about 1.5-2

Finding the b-jets

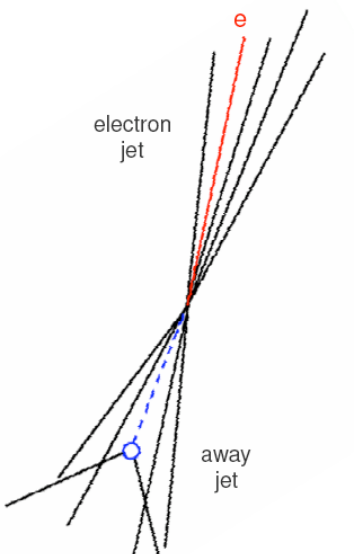


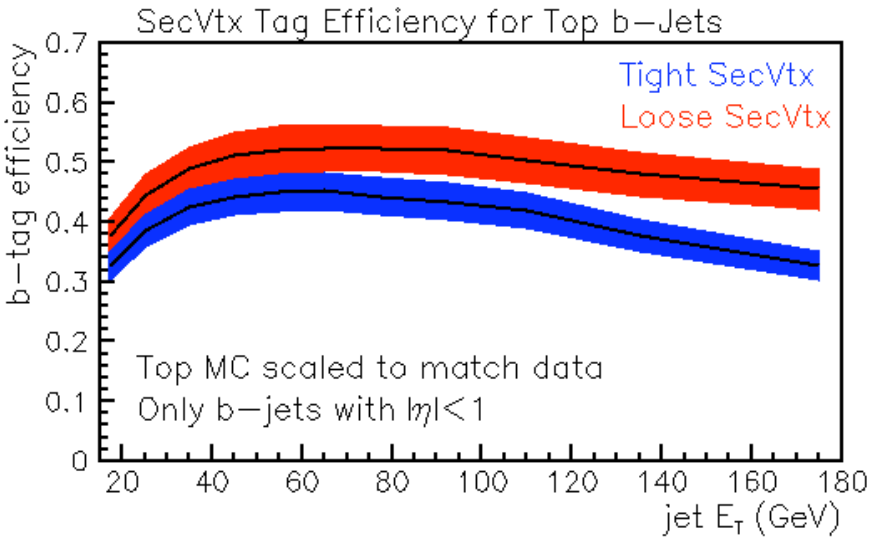
- Exploit large lifetime of the b-hadron
  - B-hadron flies before it decays: $d=c\tau$
    - Lifetime τ =1.5 ps^{-1}
    - $d=c\tau$ = 460 μm
    - Can be resolved with silicon detector resolution
- Soft lepton tag
  - i.e. muon from B decay with large d_0
- "Secondary Vertex":
  - reconstruct primary vertex:
    - resolution ~ 30 μm
  - Search tracks inconsistent with primary vertex (large d_0):
    - Candidates for secondary vertex
    - See whether three or two of those intersect at one point
  - Require displacement of secondary from primary vertex
    - Form L_{xy}: transverse decay distance projected onto jet axis:
      - $L_{xy}>0$: b-tag along the jet direction => real b-tag or mistag
      - $L_{xy}<0$: b-tag opposite to jet direction => mistag!
    - Significance: e.g. $\delta L_{xy} / L_{xy} > 7$ (i.e. 7σ significant displacement)
- More sophisticated techniques exist

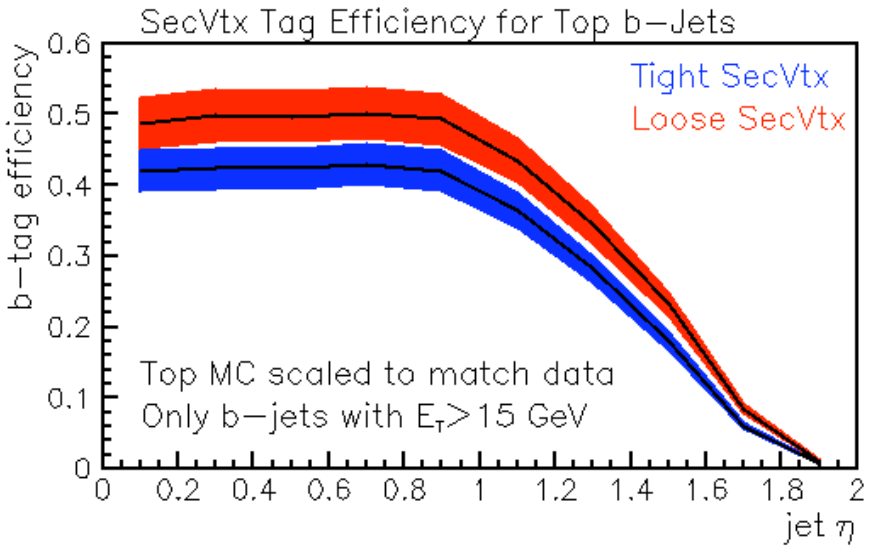


Characterise the B-tagger: Efficiency

- Efficiency of tagging a true b-jet
  - Use Data sample enriched in b-jets
  - Select jets with electron or muons
    - From semi-leptonic b-decay
  - Measure efficiency in data and MC



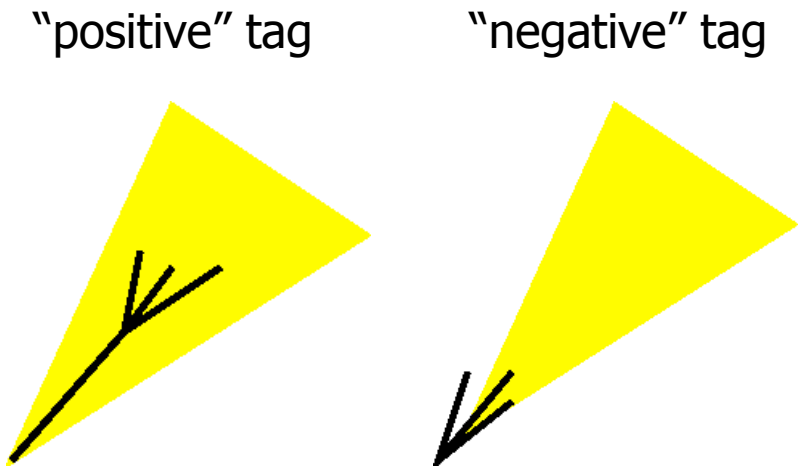


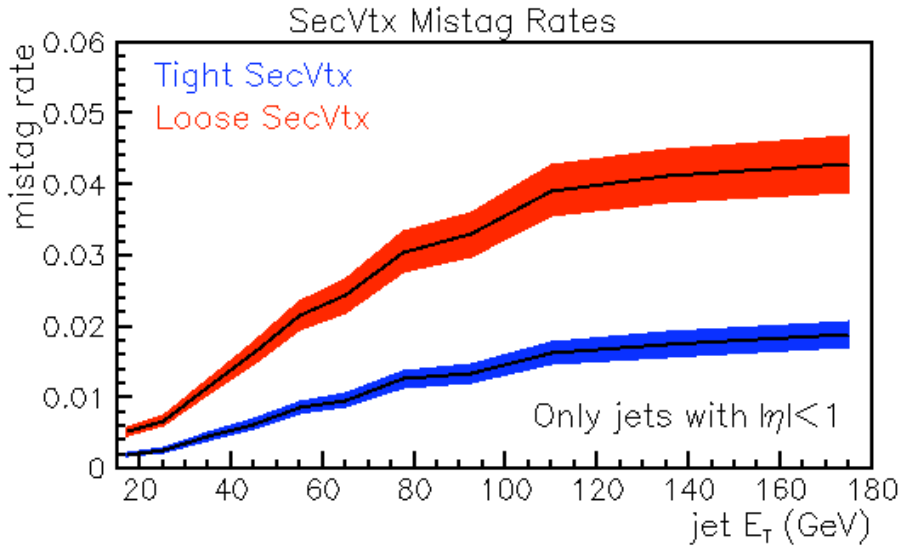


Achieve efficiency of about 40-50% at Tevatron

Characterise the B-tagger: Mistag rate

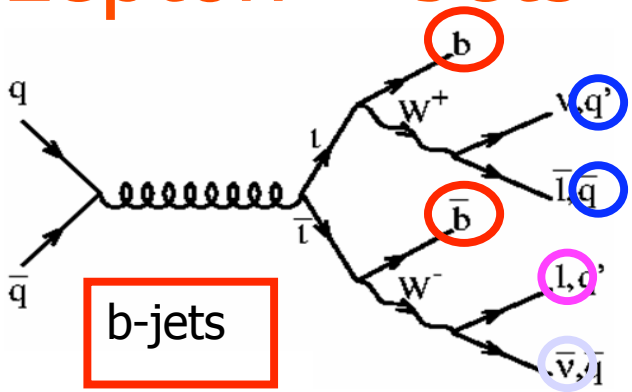
- Mistag Rate measurement:
  - Probability of light quarks to be misidentified
  - Use "negative" tags: $L_{xy}<0$
    - Can only arise due to misreconstruction
  - Mistag rate for E_T=50 GeV:
    - Tight: 0.5% (ε=43%)
    - Loose: 2% (ε=50%)
  - Depending on physics analyses:
    - Choose "tight" or "loose" tagging algorithm

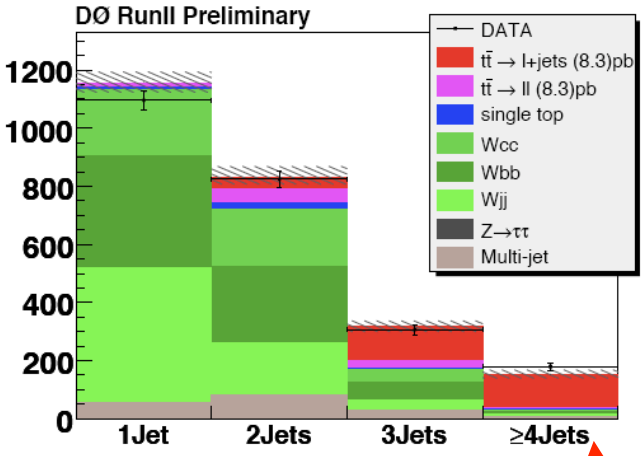


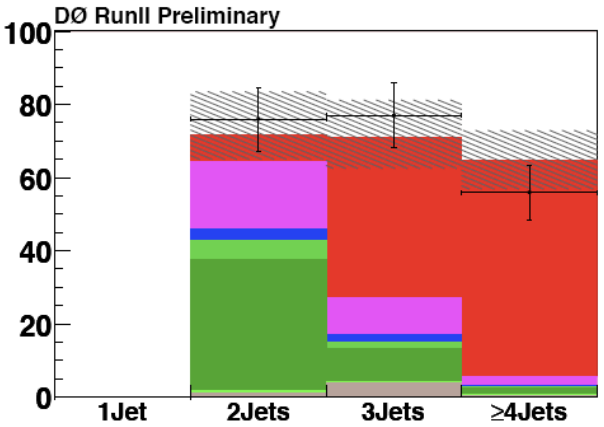


The Top Signal: Lepton + Jets

- Select:
  - 1 electron or muon
  - Large missing E_T
  - 1 or 2 b-tagged jets







double-tagged events, nearly no background

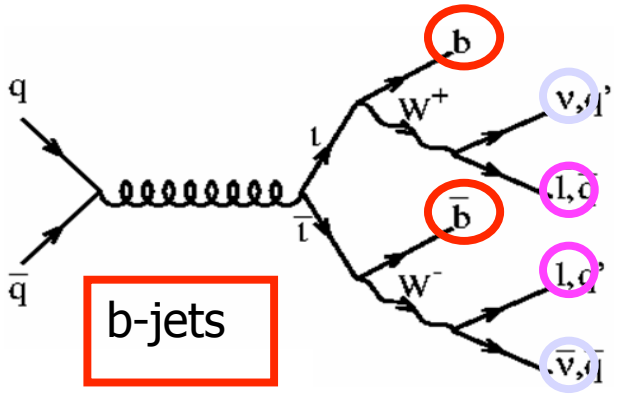
Check backgrounds

Top Signal

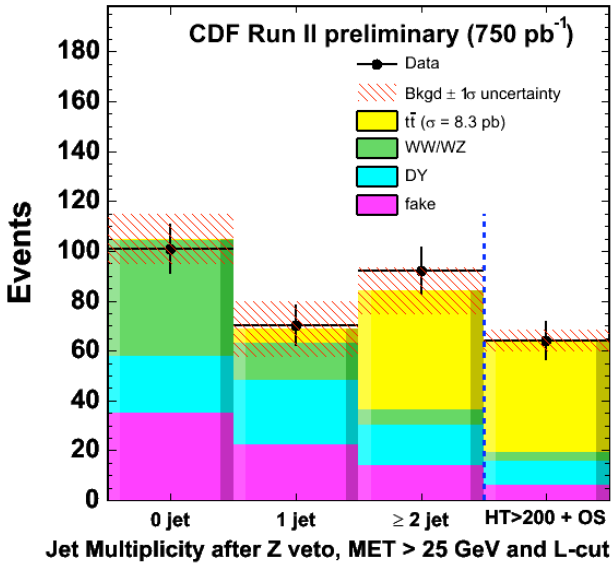
$\sigma(t\bar{t}) = 8.3^{+0.6}_{-0.5}(\text{stat}) \pm 1.1\ (\text{syst})$ pb

The Top Signal: Dilepton

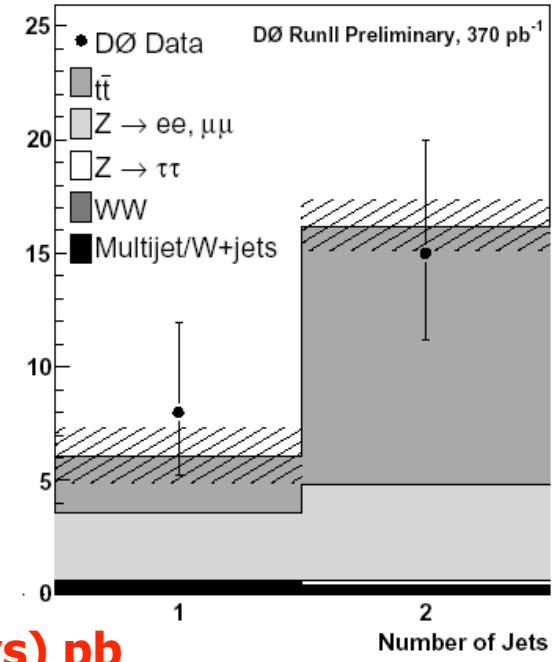
- Select:
  - 2 leptons: ee, eμ, μμ
  - Large missing E_T
  - 2 jets (with or w/o b-tag)



w/o b-tag



with b-tag



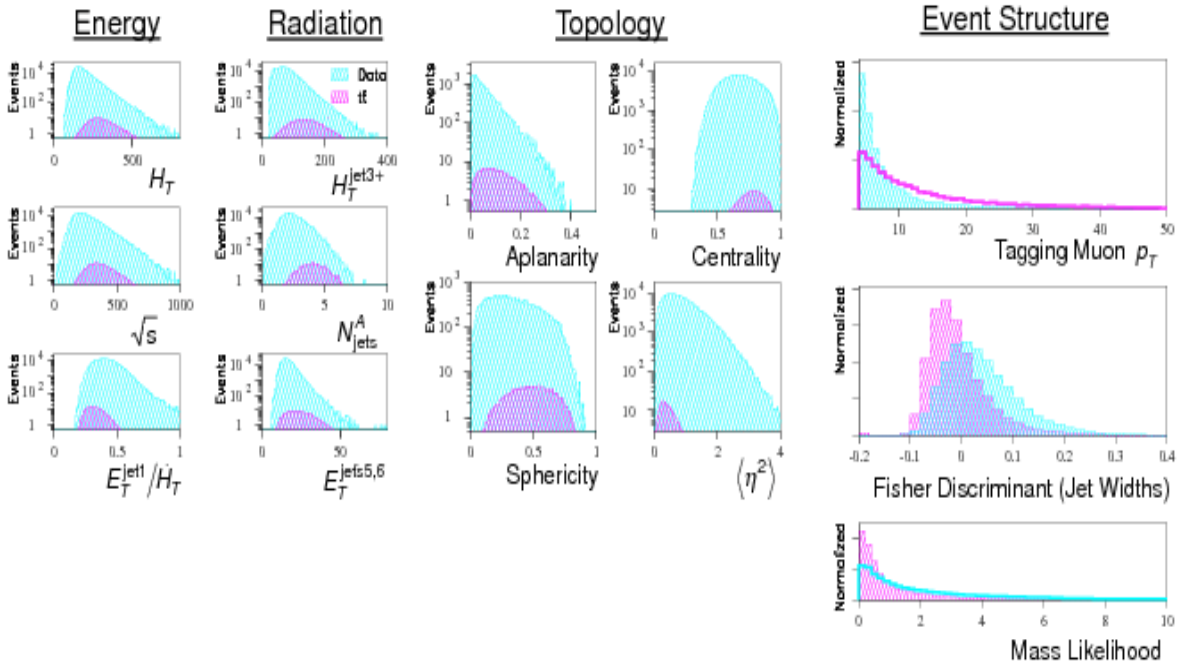
σ = 6.2 ± 0.9 (stat) ± 0.9 (sys) pb

Top to six jets

- The hardest channel
  - no leptons or MET in the final state
  - the main background is QCD
- hard even when requiring two tight b-tags, paying $0.4^2=0.16$ in branching (CDF)
- Was also observed in Run I by DØ, without magnetic field or silicon tracker
  - was made possible by the use of Neural Networks, one of the first analyses from major HEP experiment to employ them
  - exploit subtle differences in event kinematics and jet shape (top jets are quark, QCD multijets are gluon)

Top → All jets

- used 18 variables!



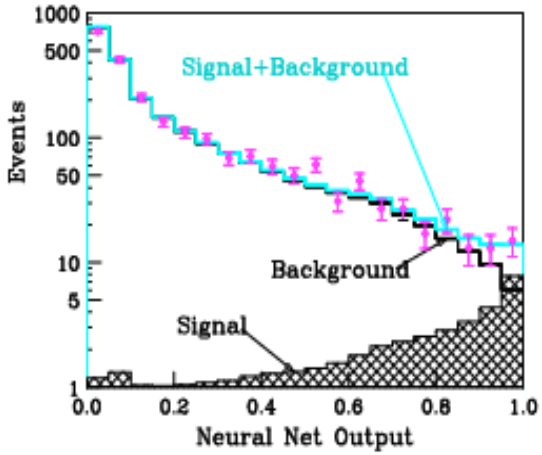


Figure 2: Neural network variables for the DØ $t\bar{t} \rightarrow$ alljets analysis. The first 10 variables are used in one network, and the output from that network is used together with the last three variables in a second network.

Multivariate Analyses

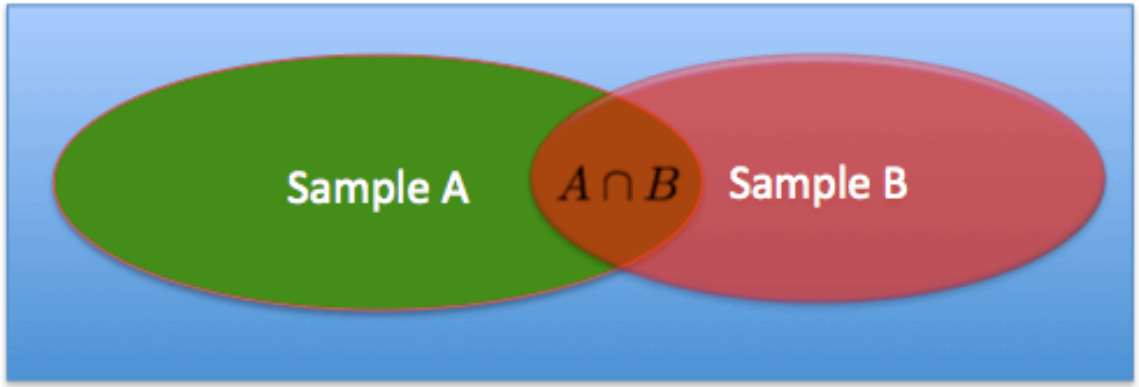
- Advantages and disadvantages
- Neural Networks
- Decision Trees
- Matrix Element methods

What Multivariate Analyses Are

- We start with a data sample of interesting events: U
  - Each event can be described in terms of n dimensions (or n *discriminating variables*) of interest.
- This sample contains more than one class of events: A, B, ...
- Lets just consider the case of two classes (simple to generalize to more)
- So: $A \subset U, \text{ and } B \subset U$

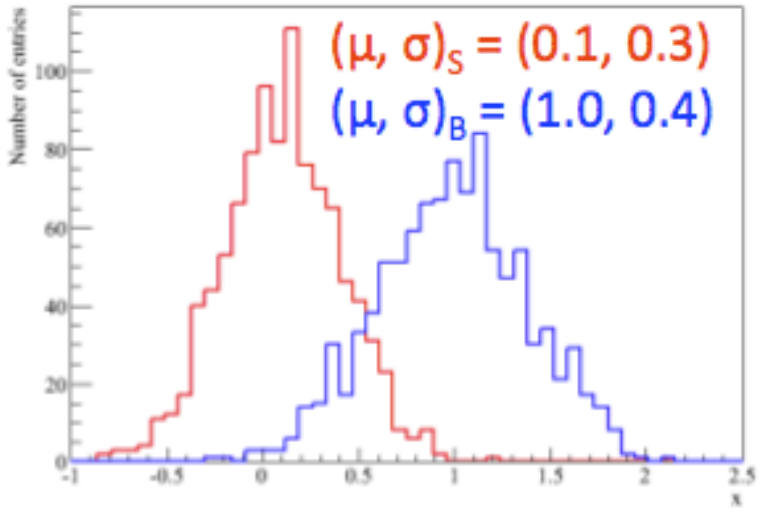
$$A \cap B \neq \emptyset$$

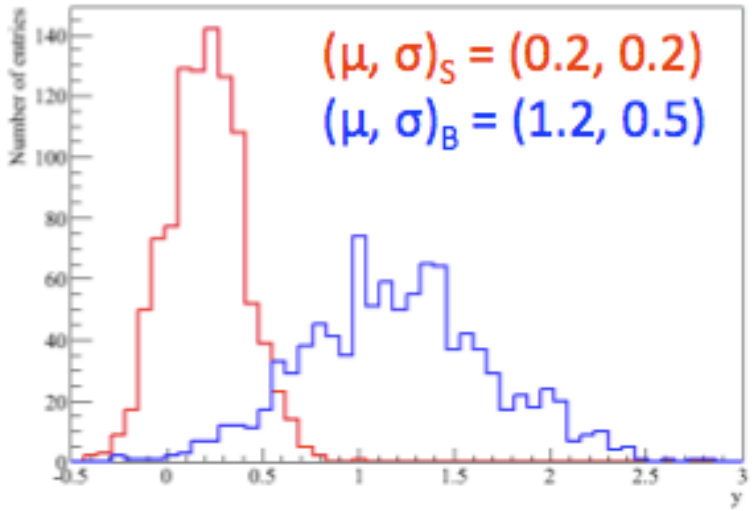
Note: If the intersection of A and B is null, then the problem is not interesting, and we can easily separate the two classes of interest with a set of cuts.



Q) How can we optimally separate classes A and B in n dimensions?

Cuts on Variables





- Very intuitive and visual
- Example:
  - two variables X and Y
  - both show separation between "signal" and "background"
  - cuts on both will improve purity
  - cut optimization may be a little complicated if variables are correlated, but it's an easily solvable problem:
    random grid search
- Systematic error is relatively easy to estimate

What Multivariate Analyses Are

- Consider the event e_i:
  - $e_i = e_i(\underline{x}) = e_i(x_1, x_2, x_3, ..., x_n)$, which is the i^{th} event of a dataset U.
  - How do we determine the A*ness* or B*ness* of a given event $e_i = e_i(\underline{x})$?

•We need some way to assign a probability to the hypothesis that event e_i is of class A.

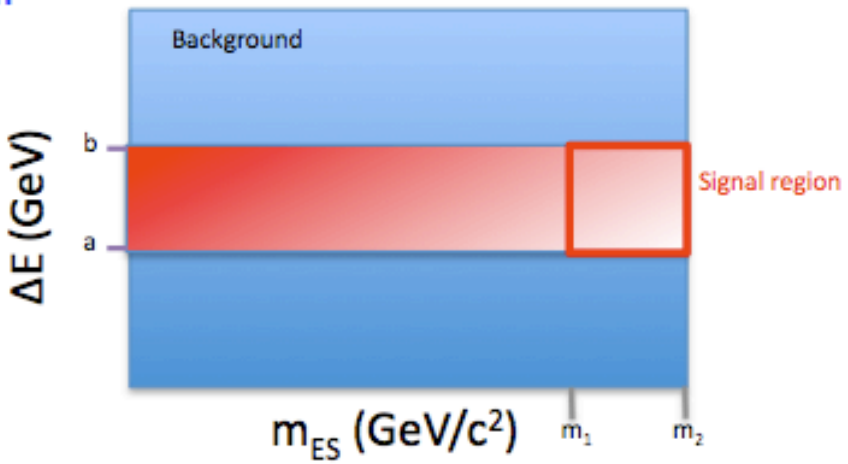
$$P(e_i \in A) \leq 1$$

•The complement is the probability that e_i is in the class B (as we are only considering two classes).

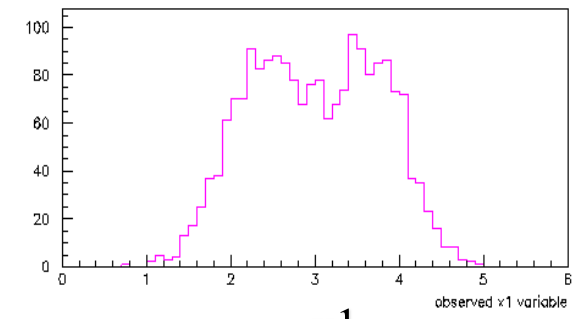
$$\overline{P(e_i \in A)} = P(e_i \in B) \leq 1$$

•Most of the time we can't tell for certain if an event e_i is of class A or class B.

Think of the familiar example of a signal region, where we know that the signal purity will be higher in this region, than outside it.

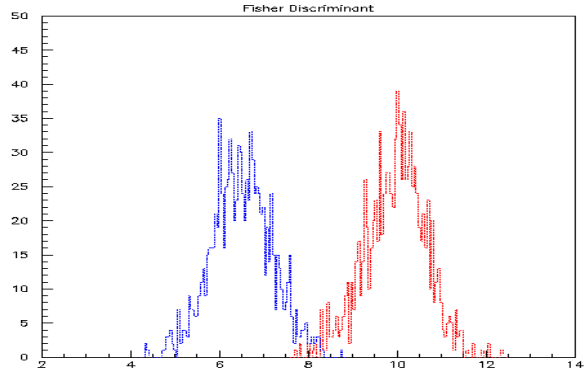


What Multivariate Analyses Are

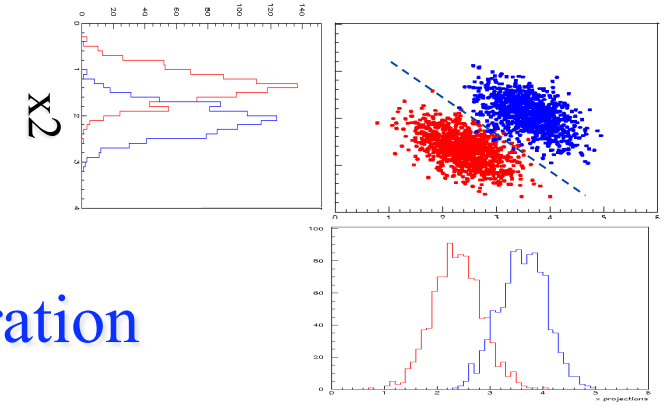


x1

x2



D(x1,x2)=2.014x1 + 1.592x2



x2

x1

Want to achieve maximum separation

$U = A \cup B$ → algorithm → Output

Linear Algorithms

- The example above can be solved by a simple linear algorithm, like a Fisher's Discriminant

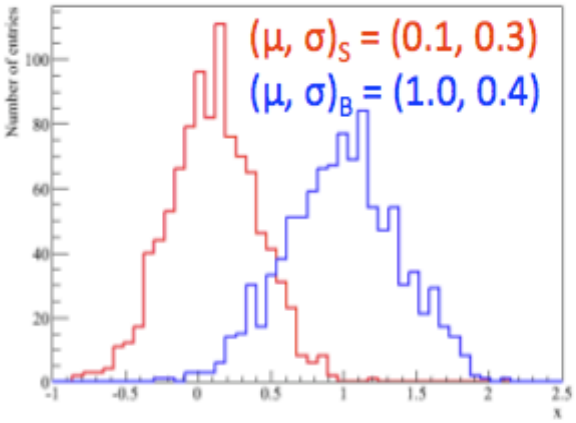
$$\Im_i = \sum_{j=1}^{j=N_{\text{var}}} \alpha_j x_{ij} = W \cdot \vec{x}_i$$

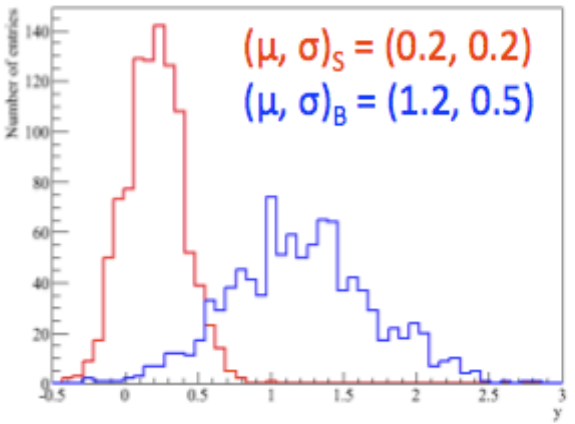
- Finding matrix W is fairly straightforward exercise
  - need to maximize the difference between mean values of $\Im_i$ for signal and background while minimizing their RMS'es

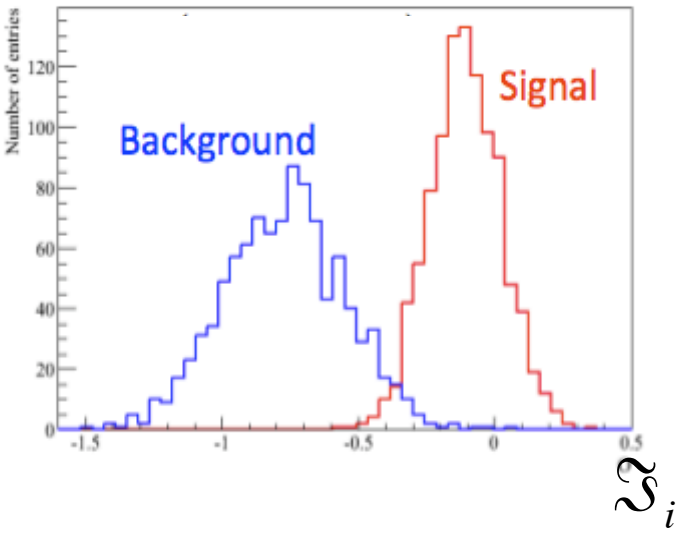
- what this amounts to is finding the optimal coordinate system in the N_{var}-dimensional space by linear transformations

Fisher's Discriminant

- Let's see an example

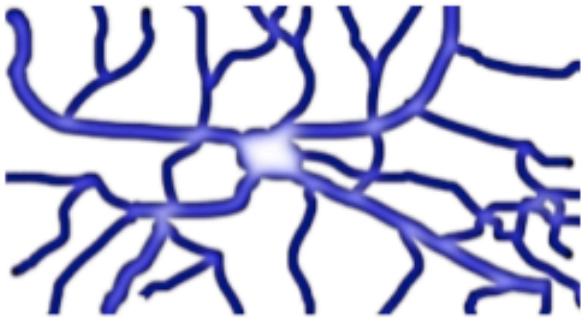


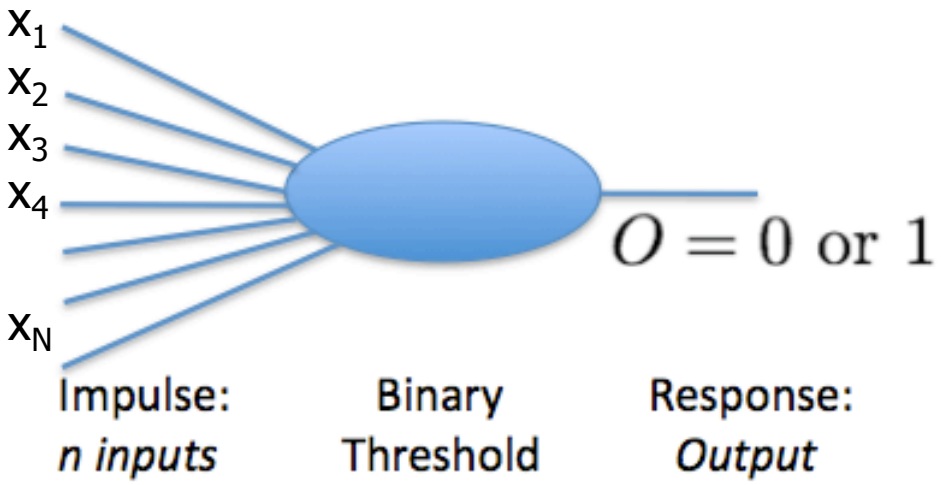




Neural Networks

- These are non-linear algorithms:
  - Called Artificial Neural Networks: ANN or just Neural Networks NN.
  - The fundamental building block of a NN is the perceptron (algorithmic analogy of a neuron).





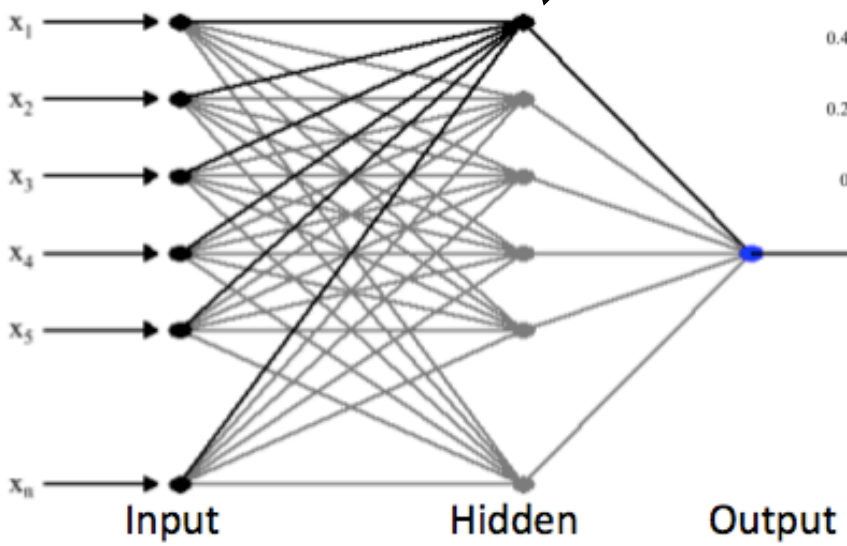
$$y = \vec{w} \cdot \vec{x} + b$$

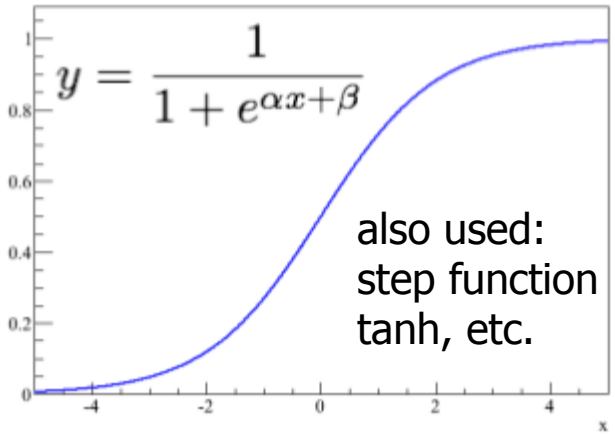
if *y>0* then *O=1*, else *O=0*

Neural Networks (Multi-Layer Perceptron)

- n inputs
- 1 hidden layer of n nodes
- 1 output

$$y = f(\vec{w} \cdot \vec{x} + b)$$

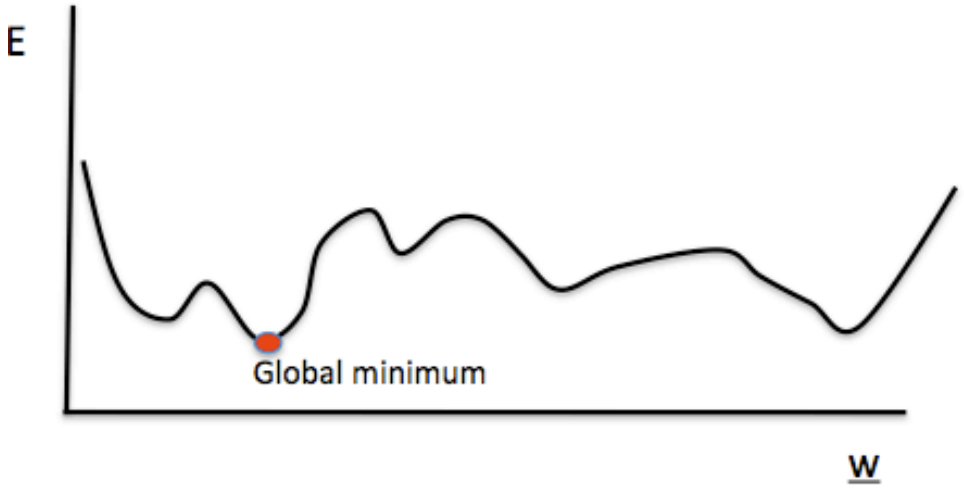




- Decide on the activation function to use for each node/layer.
- Determine the weights used to evaluate y_i for each node.
- Check that we have not over-trained our network

Training an MLP

- This is a multi-parameter problem.
- There are many minima, and we want to converge on the global minimum, not on a local one.



There are many nodes, hence many many weight parameters to determine when training an MLP.

This is a complicated problem akin to a multi-dimensional ML fit with many free parameters.

- Determining the global minimum can be non-trivial.

but, there are now fairly advanced tools to do that, like TMVA

Training an MLP

- In order to train the MLP we need two samples of data:
  - Sample A, which is a data-set containing M entries of class A events.
  - Sample B, which is a data-set containing M entries of class B events.

You don't have to use equal numbers of events for both classes, however not doing so will affect the convergence of your network. You are advised to keep to using equal number of events in samples A and B.

- How do we know when training has finished?
  - Just compare the error against some anticipated threshold?
  - Just compare the error gradient against some anticipated threshold?
  - Compare the error obtained against a validation sample.

Training an MLP: How much data do we need?

- As a rule of thumb, the number of events scales with the complexity of the network as follows:

$$
\begin{aligned}
M &= \text{sample size} \\
W &= \text{number of weight parameters} \\
N &= \text{number of nodes} \\
\epsilon &= \text{error threshold}
\end{aligned}
$$

If there is a single hidden layer, to avoid failing to train a net properly you want to make sure that the training sample size M satisfies:

$$M > O\left(\frac{W}{\epsilon}\right)$$

If your sample doesn't satisfy this then you run a high risk of misclassification of events.

If the network is more complicated then you should try to ensure that:

$$M > O\left[\frac{W}{\epsilon}\log(N/\epsilon)\right]$$

Baum & Haussler, Neural Comp. 1 151-160 (1989)

Training an MLP: How much data do we need?

- Example:
  - an MLP with 1 hidden layer of 10 nodes,10 inputs and 1 output node (so... W=(10+1)×10), and the misclassification error level you want to achieve is 0.1:

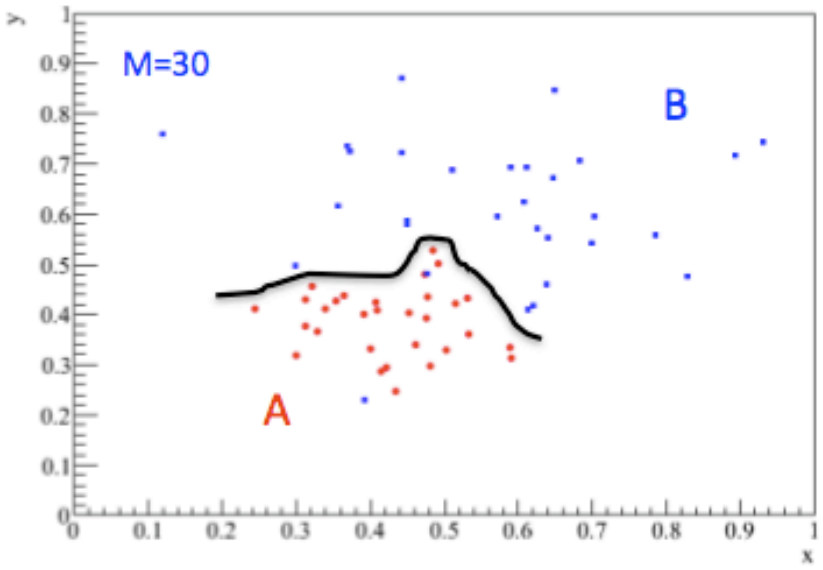
$$M > O\left(\frac{W}{\epsilon}\right)$$

  - You want more than 1100 training events to have a reasonable chance of obtaining an optimal separation of signal and background.

  - This doesn't mean that you get a properly trained net – you have to do some more checks to ensure that!

Training an MLP: Validation

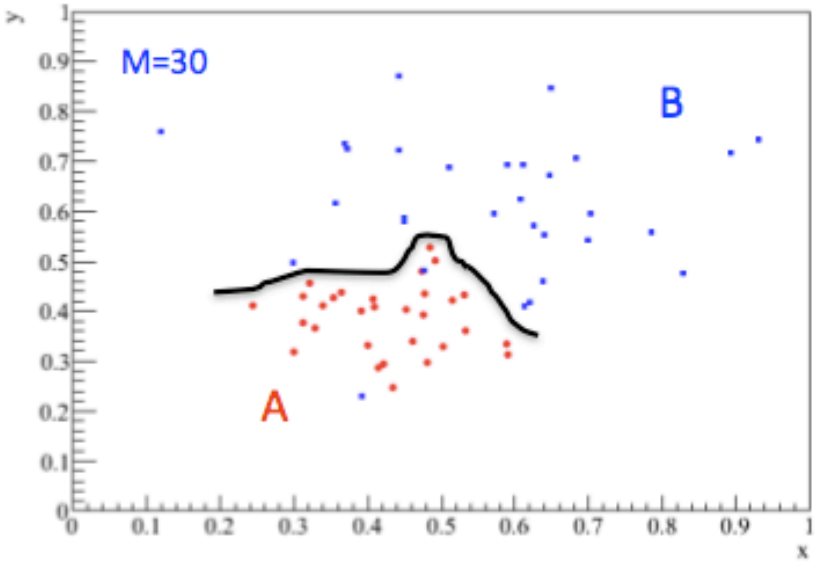
- Overtraining occurs when you have obtained weights that are tailored to your specific sample of A and B events, rather than being a true representation of the optimal discrimination between the classes.

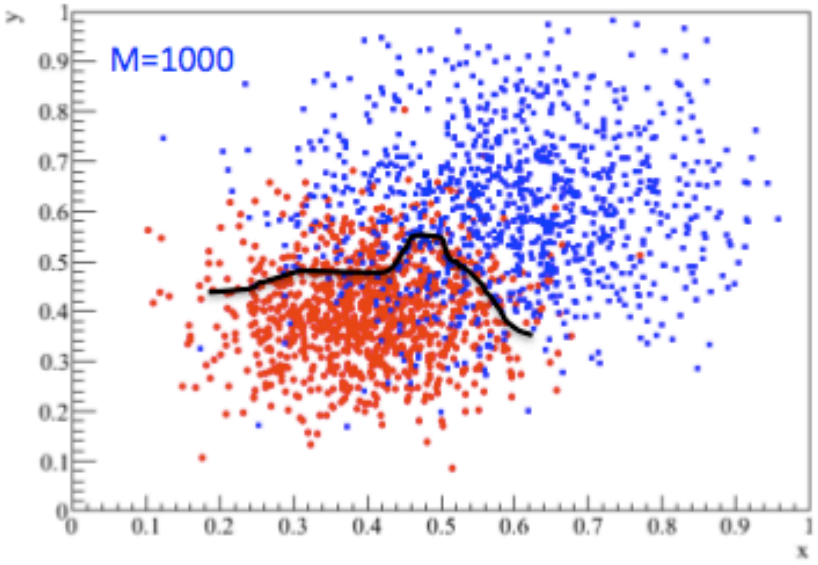


Is the line a reasonable boundary to use as a cut between A and B?

Training an MLP: Validation

- Overtraining occurs when you have obtained weights that are tailored to your specific sample of A and B events, rather than being a true representation of the optimal discrimination between the classes.



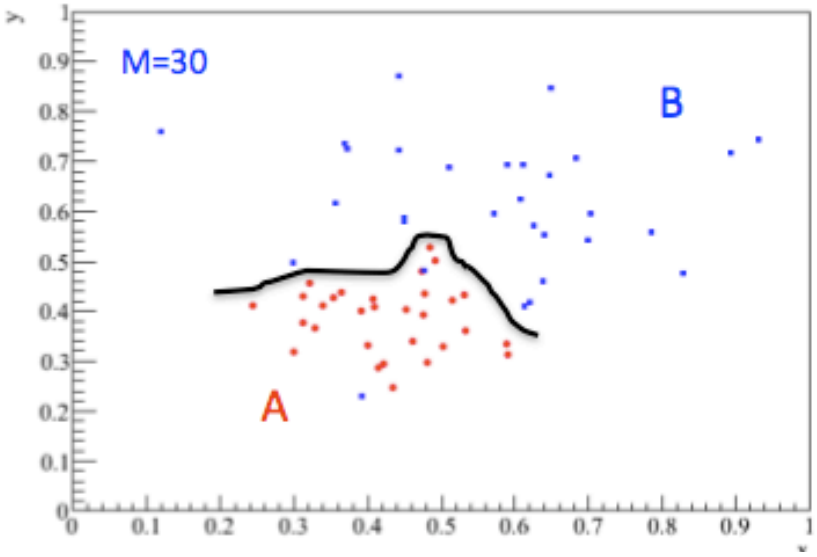


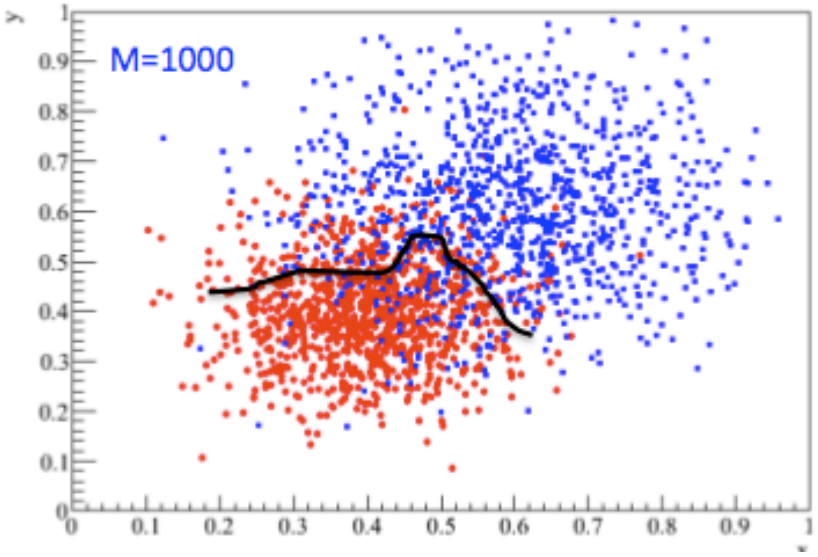
The boundary has been tailored to the initial sample of statistics and (in this case) is not the best choice of boundary for a separate sample.

This is an illustrates the need to have sufficient data to train. It highlights the issue of statistical fluctuations in data. **Don't tune on features of a specific data set!**

Training an MLP: Validation

- Overtraining occurs when you have obtained weights that are tailored to your specific sample of A and B events, rather than being a true representation of the optimal discrimination between the classes.





A solution is to use a statistically independent sample to check the result of the training, and to stop training only when the the training and reference samples give the same performance (within tolerance).

Training an MLP: Validation

- It is important that we have sufficient data to use in training:
  - Makes sure that the result is sensible.
  - Means that we can use an equal amount of data as a reference to compare against.
  - This can be a tough constraint as we often resort to MLPs when we want to extract every last bit of information from the data, and usually don't have events to spare!

- Similarly make sure that you don't over-train your MLP. How do you know if you are converging on a general feature of the data, or just a specific feature of your dataset?
  - Use a validation sample!

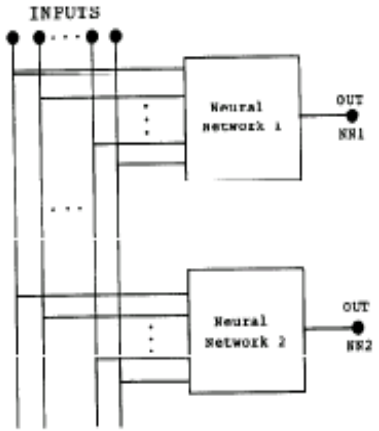
- The temptation is to use all data to train.
  - Don't do it as you can't guarantee the result is sensible!

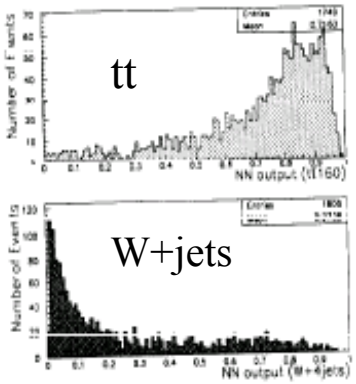
The Top Quark

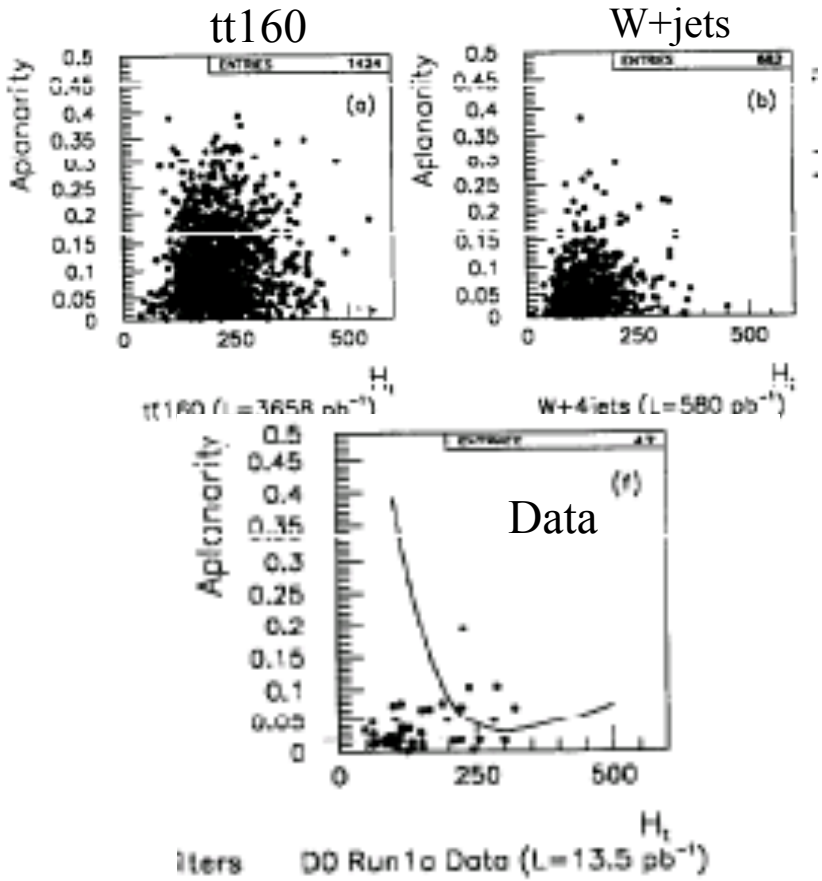
DØ DPF94

Analysis with 2 and 5 variables

NN Analysis tt → e+jets channel







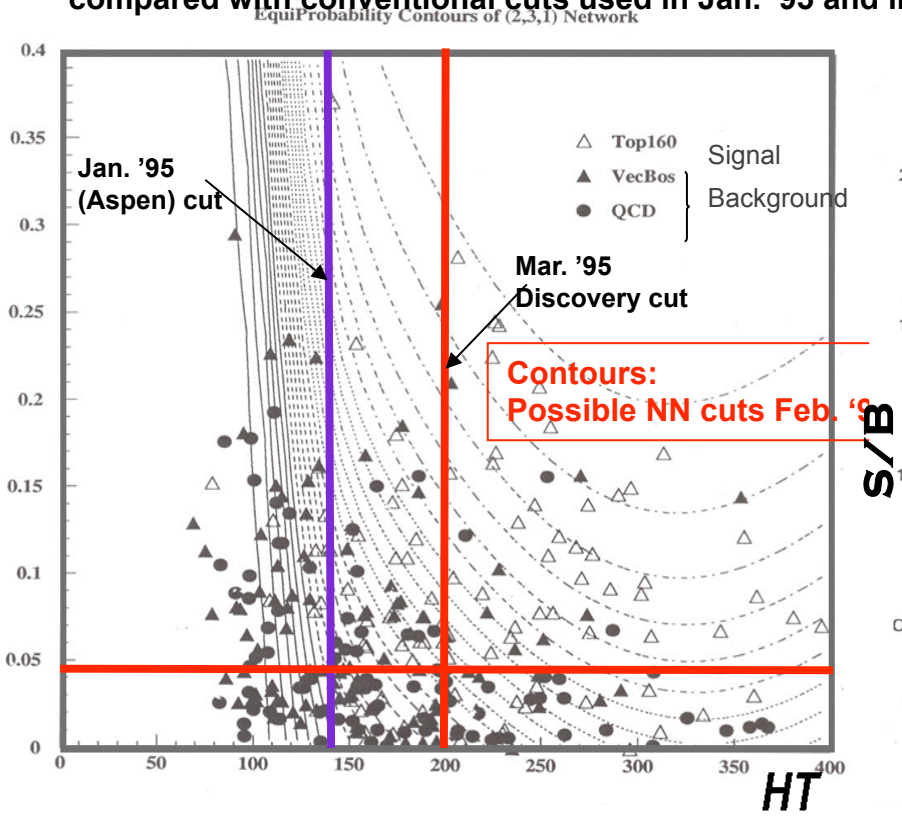
- If DØ had pursued the NN analysis in other channels, the evidence and/or discovery may have come sooner!

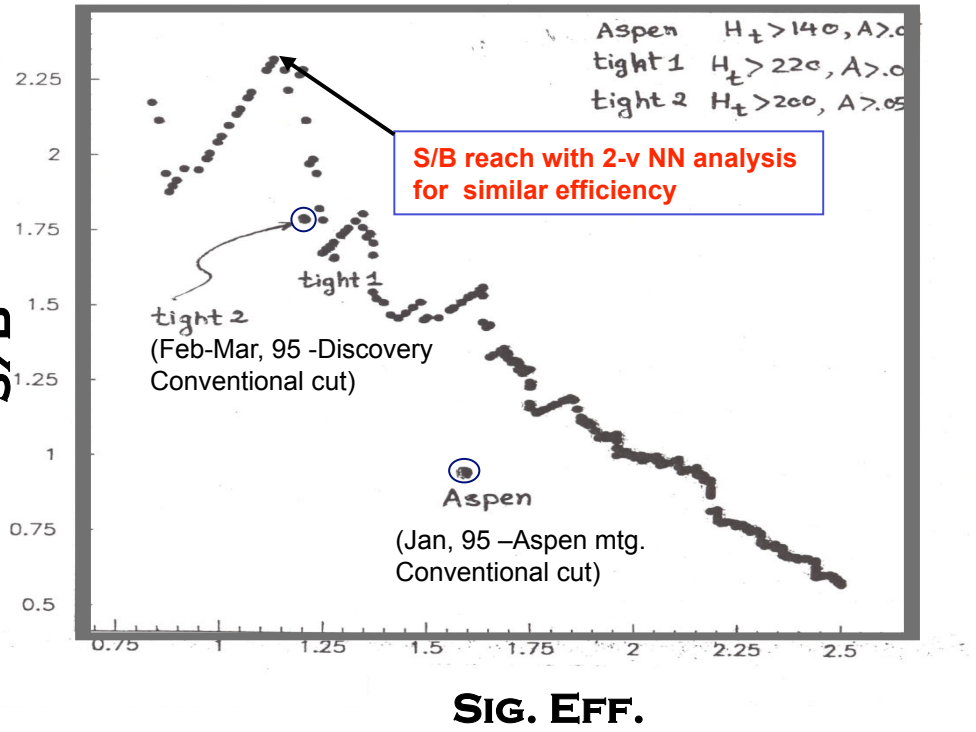
Cut Optimization
Feb. '95

P. Bhat, H.Prosper, E. Amidi
D0 Top Marathon, Feb. '95

Aplanarity & HT variables
Letpon+jets channels

Neural Network Equi-probability Contour cuts from 2-variable analysis compared with conventional cuts used in Jan. '95 and in Observation paper



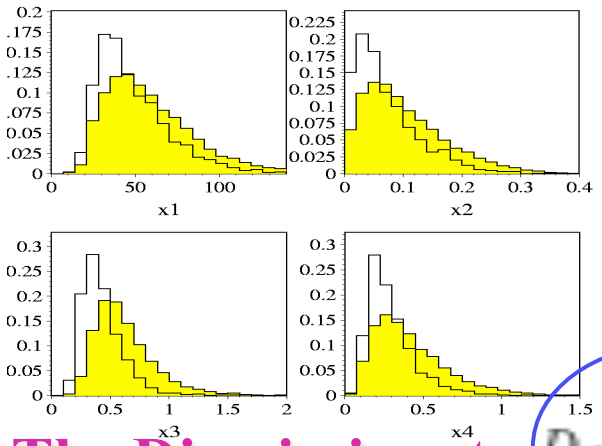


Measurement of the Top Quark Mass

1996

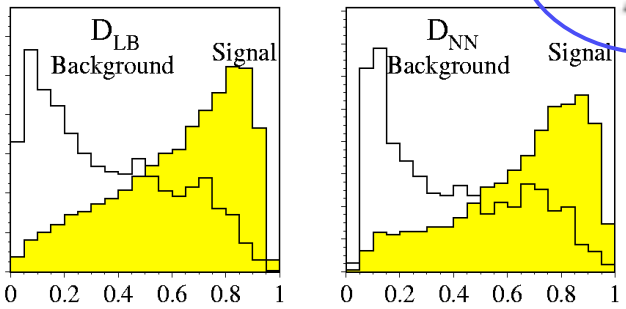
First significant physics result using multivariate methods

Discriminant variables



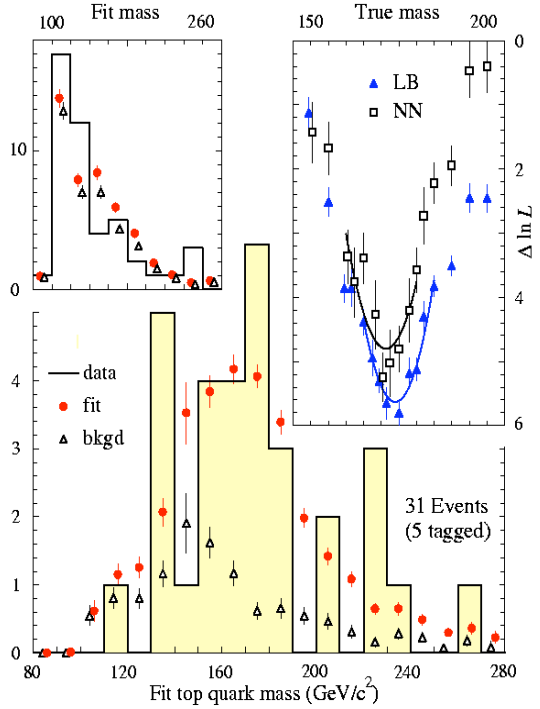
The Discriminants

$$D = \frac{P_s}{P_s + P_b}$$



LB: Low-bias maximum likelihood
NN: Neural Networks

DØ Lepton+jets



$m_t = 173.3 \pm 5.6(\text{stat.}) \pm 6.2\ (\text{syst.})\ \text{GeV}/c^2$

Fit performed in 2-D: $(D_{LB/NN}, m_{fit})$

Statistical error for the same data sample reduced from 11.7 GeV to 5.6 GeV!

Decision Trees

A comparatively new method of analysis

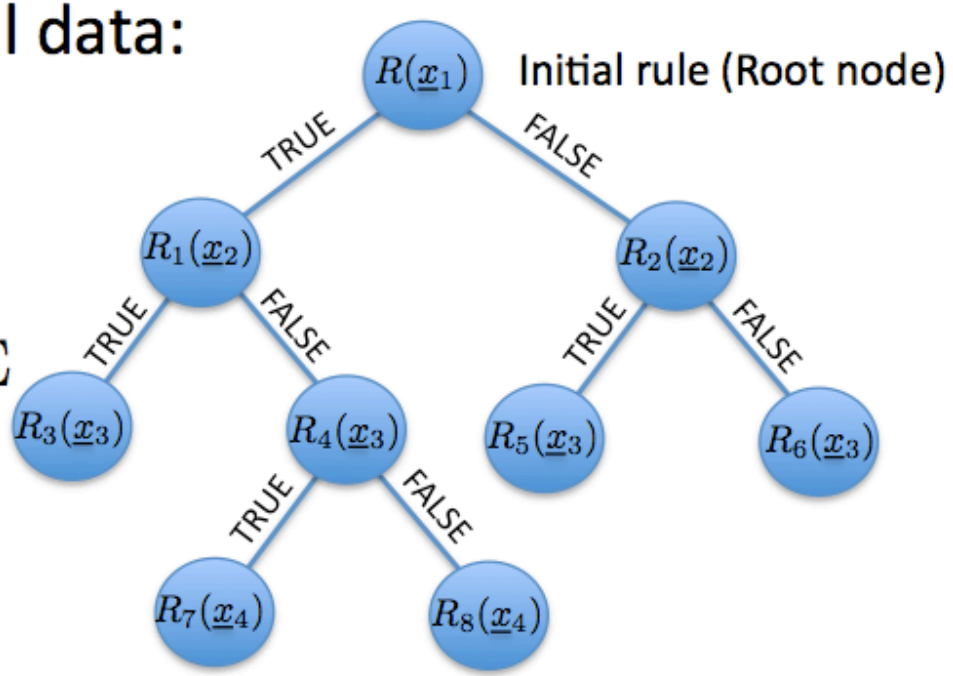
- seems more visual (perception, mostly)
- Trees can be
  - Binary
  - Boosted
  - Bagged
- Trees in a Forest

Decision Trees

- Apply the initial rule to all data:
  - Divide into two classes (with a binary output)

$$R(\underline{x}_1) = \underline{x} > \underline{x}_i \text{ TRUE}$$
$$= \underline{x} < \underline{x}_i \text{ FALSE}$$

  - Each successive layer divides the data further into Signal (class A)/ Background (class B) classifications.

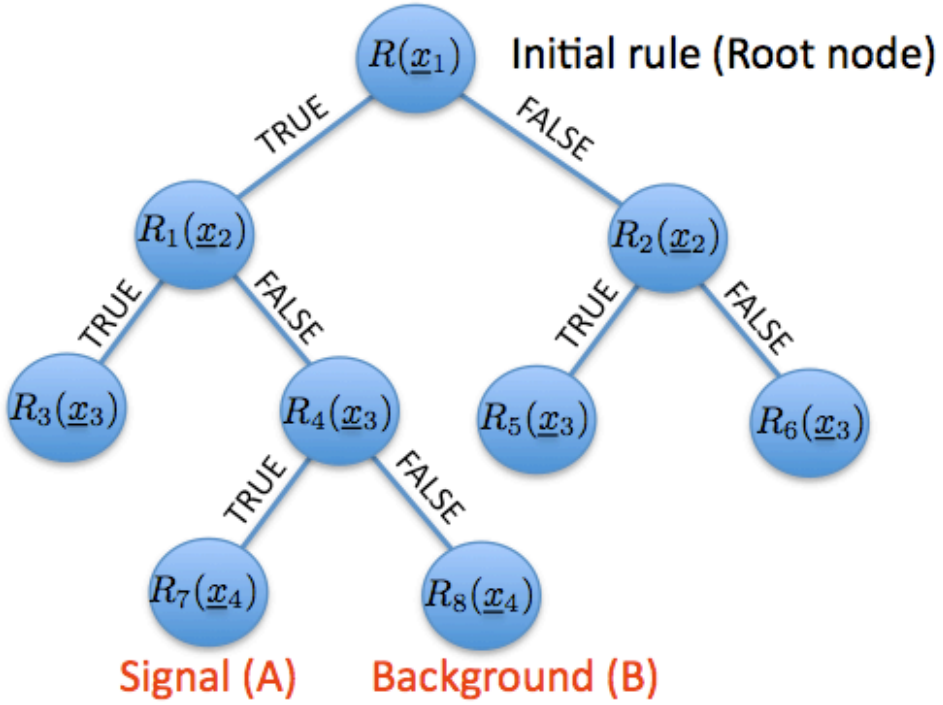


- The classification for a set of cut values will have a classification error.
  - So just as with a NN one can vary the cut values x_i in order to train the tree.

Decision Trees

- What is happening?

- Each nodes uses the sub-set of discriminating variables that give the best separation between classes.

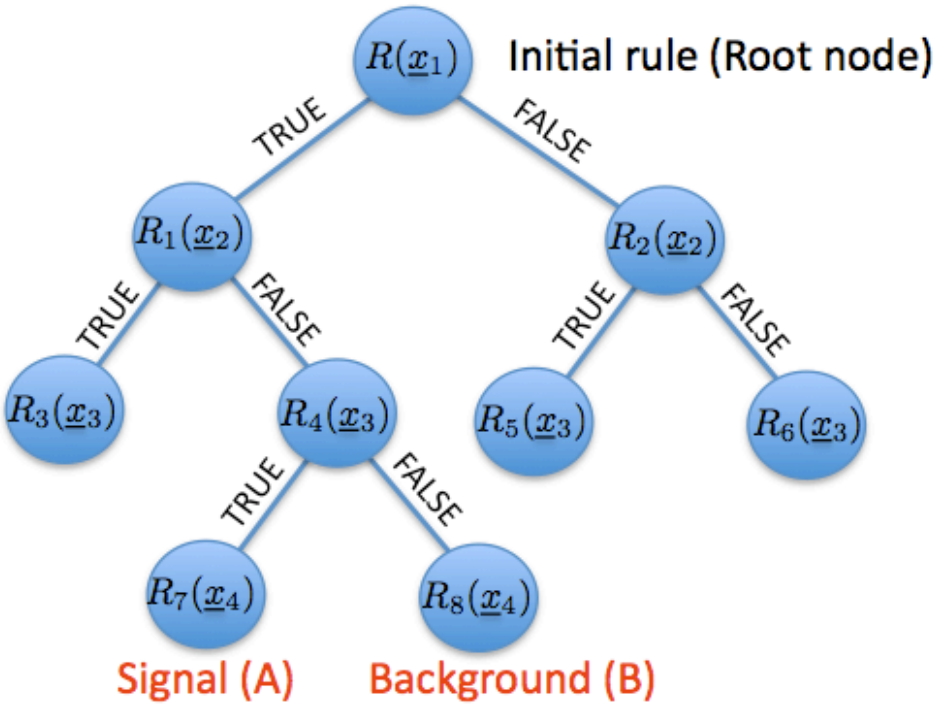


- Some variables may be used by more than one node.

- Other variables may never be used.

Decision Trees

- What is happening?
- The bottom of a tree just looks like a sub-sample of events subjected to a cut based analysis.



- There are many bottom levels to the tree:
  - ... so there are many signal / background regions defined by the algorithm.

Decision Trees

- Binary Decision Tree has the following pros/cons:

- Pros:
  - Easy to understand and interpret.
  - More flexibility in the algorithm when trying to separate classes of events.
    - Able to obtain better separation between classes of events than a simple cut-based approach.

- Cons:
  - Instability with respect to statistical fluctuations in the training sample.

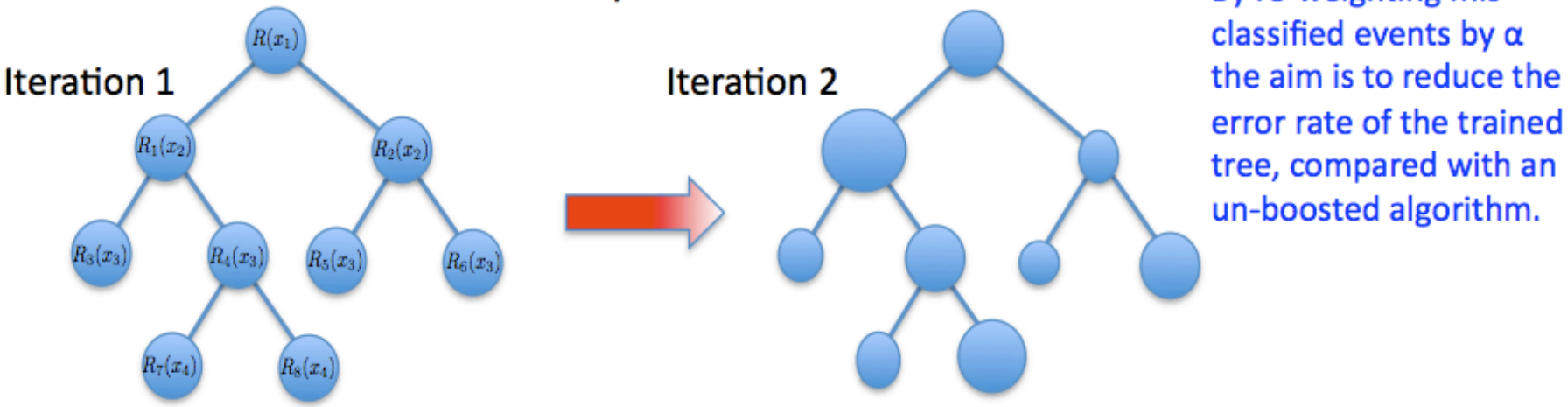
- It is possible to improve upon the binary decision tree algorithm to try and overcome the instability or susceptibility of overtraining.

Boosted Decision Trees

- At each stage in training there may be some mis-classification of events (error rate).
  - Assign a greater event weight α to mis-classified events in the next training iteration.

$$\alpha = \frac{1-\epsilon}{\epsilon} \qquad \epsilon = \text{error rate}$$

  - Re-weight whole sample so that the sum of weights remains the same, then iterate.



By re-weighting mis-classified events by α the aim is to reduce the error rate of the trained tree, compared with an un-boosted algorithm.

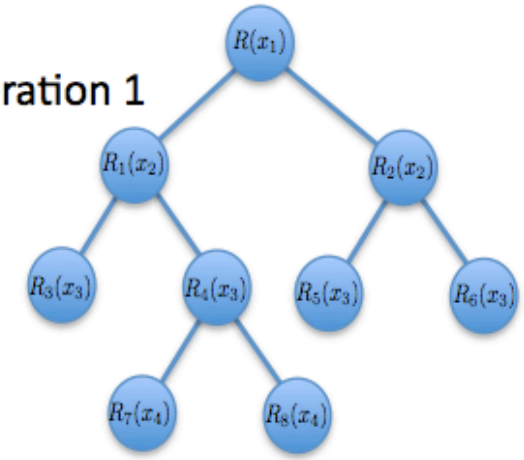
Boosted Decision Trees

- At each stage in training there may be some mis-classification of events (error rate).
  - Assign a greater event weight α to mis-classified events in the next training iteration.

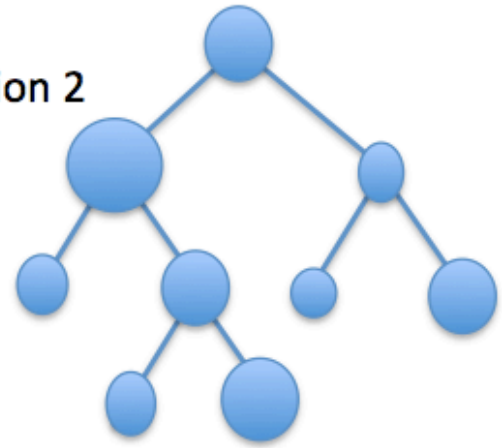
$$\alpha = \frac{1-\epsilon}{\epsilon} \qquad \epsilon = \text{error rate}$$

  - Re-weight whole sample so that the sum of weights remains the same, then iterate.

Iteration 1



Iteration 2



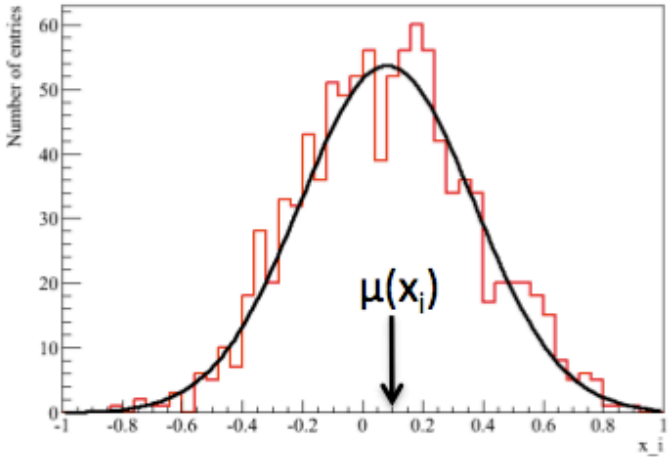
The resulting Boosted Decision Tree tends to be more stable than a normal Decision Tree.

Bagged Decision Trees

- Aim: To improve the stability of a Decision Tree algorithm.
- Solution: Sample the training data used to determine the solution.
- Take the average solution of a number of re-sampled solutions.
  - This re-sampling removes the problem of fine tuning on statistical fluctuations.

Like choosing the mean value of a cut at each level.

Again, the results tend to be more stable than just using a decision tree.



Forests

- A given decision tree may not be stable, so instead we can grow a forest.
  - For a forest, the classification of an event e_i in sample A or B is determined as the dominant result of all of the tree classifications for that event.

    e.g. In a forest of 100 trees, if there are 80 classifications of type A, and 20 of type B, the event e_i is considered to be of type A.

  - Trees in a forest use a common training sample, and are typically boosted.

Matrix Element Analysis (ME)

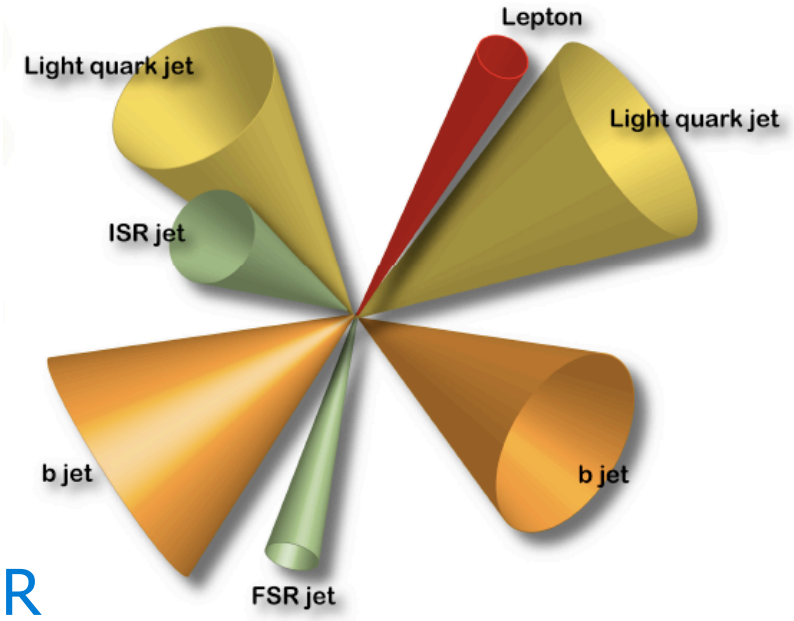
Tries to address the problem of the choice of variables

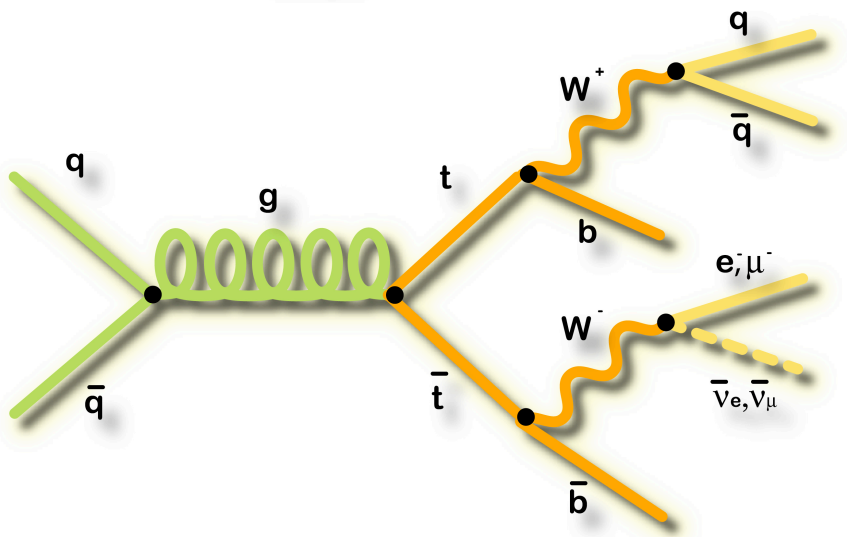
- choice may be a problem – number of variables can grow very large, so one needs huge training sample, increased sensitivity to noise, etc.
  - with modern training methods it is not as big a problem as it used to be
- take theoretical matrix element for the signal and try to map observed variables to the theoretical ones
  - before one gets into gory details seems that it guarantees the best possible set of variables
  - plus, no training is required – no false minima!

was first used by DØ to measure top mass

Top Mass Measurement: $t\bar{t} \rightarrow (bl\nu)(bqq)$

- 4 jets, 1 lepton and missing E_T
  - Which jet belongs to what?
  - Combinatorics!
- B-tagging helps:
  - 2 b-tags =>2 combinations
  - 1 b-tag => 6 combinations
  - 0 b-tags =>12 combinations
- More combinatorics from ISR/FSR
- Two Strategies:
  - Template method:
    - Uses "best" combination
    - Chi2 fit requires $m(t)=m(\bar{t})$
  - Matrix Element method:
    - Uses all combinations
    - Assign probability depending on kinematic consistency with top





First ME application: top mass

If we could access all parton level quantities in the events (the four momentum for all final and initial state particles), then we would simply evaluate the differential cross section as a function of the mass of the top quark for these partons. This way we would be using our best knowledge of the physics involved.

Since we do not have the partonlevel information for data, we use the differential cross section and integrate over everything we do not know.

$$P_{t\bar{t}}(x) = \frac{1}{\sigma_{tot}} \int d\sigma(y) dq_1 dq_2 f(q_1) f(q_2) W(x,y)$$

y is *parton* kinematic variables
x is *measured* kinematic variables
W(x,y) is a *transfer function*

Transfer Function for e+jets

W(x,y) probability of measuring **x** when **y** was produced (***x*** jet variables, ***y*** parton variables):

$$W(x,y) = \delta^3(p_e^y - p_e^x)\prod_{j=1}^{4} W_{jet}(E_j^y, E_j^x)\prod_{i=1}^{4}\delta^2(\Omega_i^y - \Omega_i^x)$$

where

E^y	energy of the produced quarks
E^x	measured and corrected jet energy
p^y_e	produced electron momenta
p^x_e	measured electron momenta
$\Omega^y_j\ \Omega^x_j$	produced and measured jet angles

Energy of electrons is considered well measured, an extra integral is done for events with muons. Due to the excellent granularity of the DØ calorimeter, angles are also considered as well measured. A sum of two Gaussians is used for the jet transfer function (W_{jet}), parameters extracted from MC simulation.

Event Probability

- in the first analysis, 5 jet events were discarded
- use all combinations, including two solutions for neutrino p_z

$$P_{t\bar{t}} = \frac{1}{\sigma_{tot}} \int d\rho_1 dm_1^2 dM_1^2 dm_2^2 dM_2^2 \sum_{comb,\nu} |M|^2 \frac{f(q_1) f(q_2)}{|q_1||q_2|} \phi_6 W_{jet}(x,y)$$

2(in) + 18(final) = 20 degrees of freedom

3(e) + 8(Ω1..Ω4) + 3(P_{in}=P_{final})+1(E_{in}=E_{final}) = 15 constraints

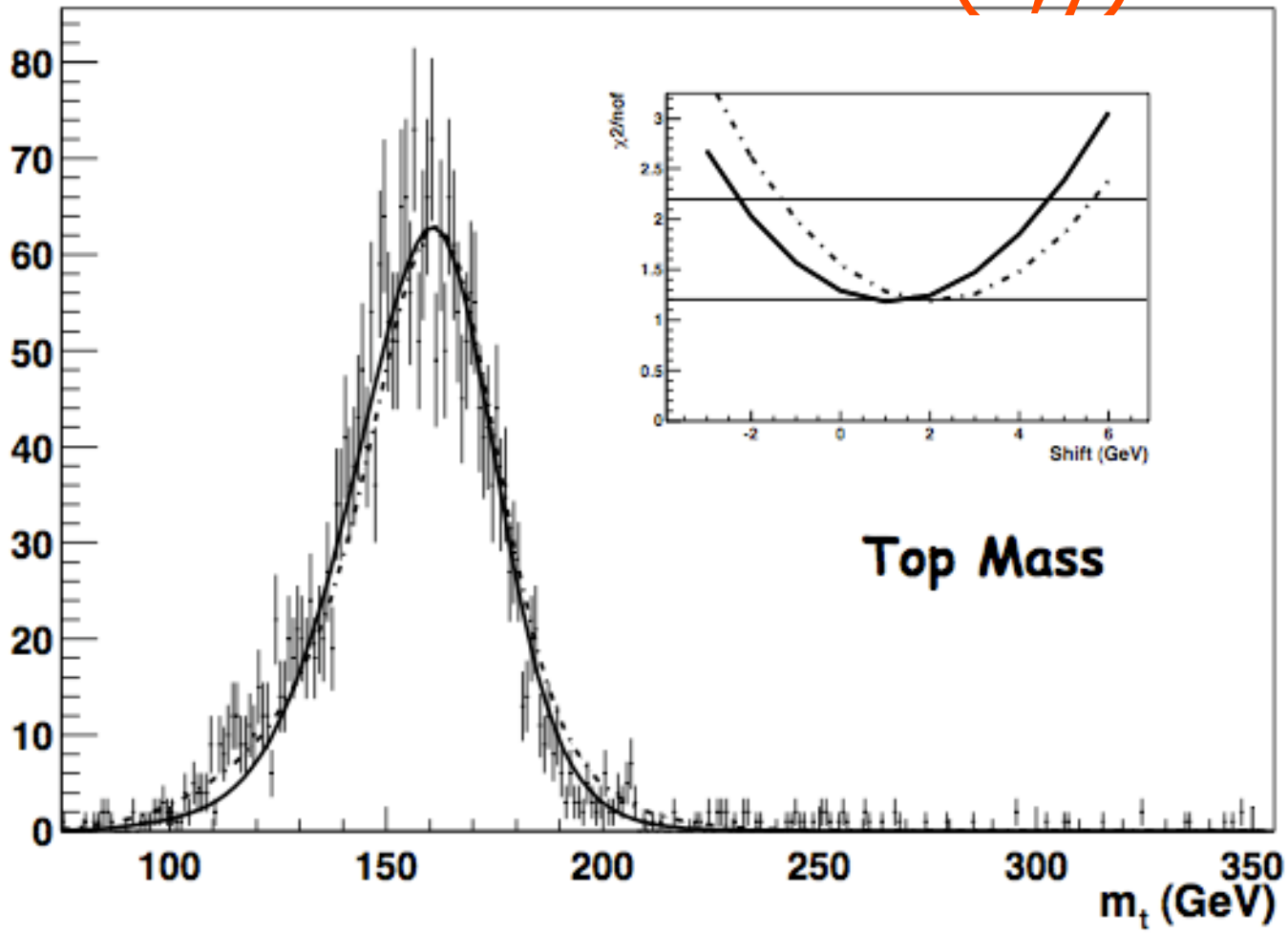
20 – 15 = 5 integrals

Sum over 24 combinations of jets, all values of the neutrino momentum are considered. Because it is L.O., we use only 4-jet events.

ρ_1	momentum of one of the jets	m_1, m_2	top mass in the event
M_1, M_2	W mass in the event	$f(q_1), f(q_2)$	parton distribution functions (CTEQ4) for qq incident chann.
q_1, q_2	initial parton momenta	ϕ_6	six particle phase space
$W(x,y)$	probability of measuring x when y was produced in the collision		

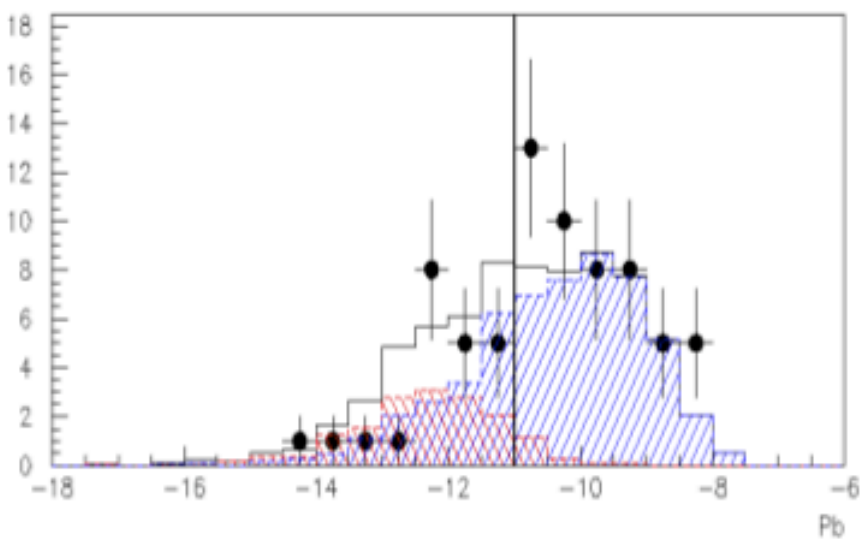
- make similar event probabilities for the background
- discriminator is $P_{sig}/(P_{sig}+P_{bkg})$

Transfer Function: full simulation vs. direct calculation with W(x,y)

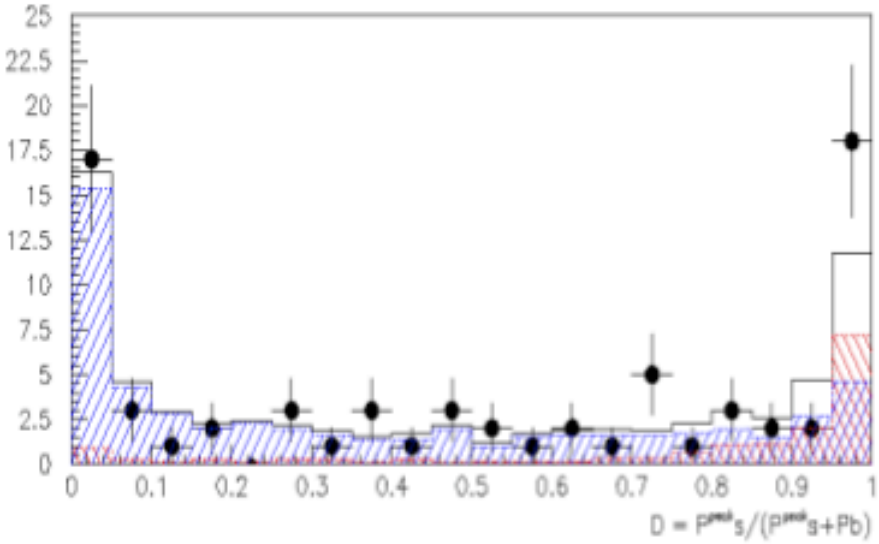


Histogram: HERWIG events after **full DØ reconstruction**, using the standard criteria
Solid Line: Calculated by using the **transfer function on partons**
Dashed: Same as solid, but with a variant transfer function

Run I data



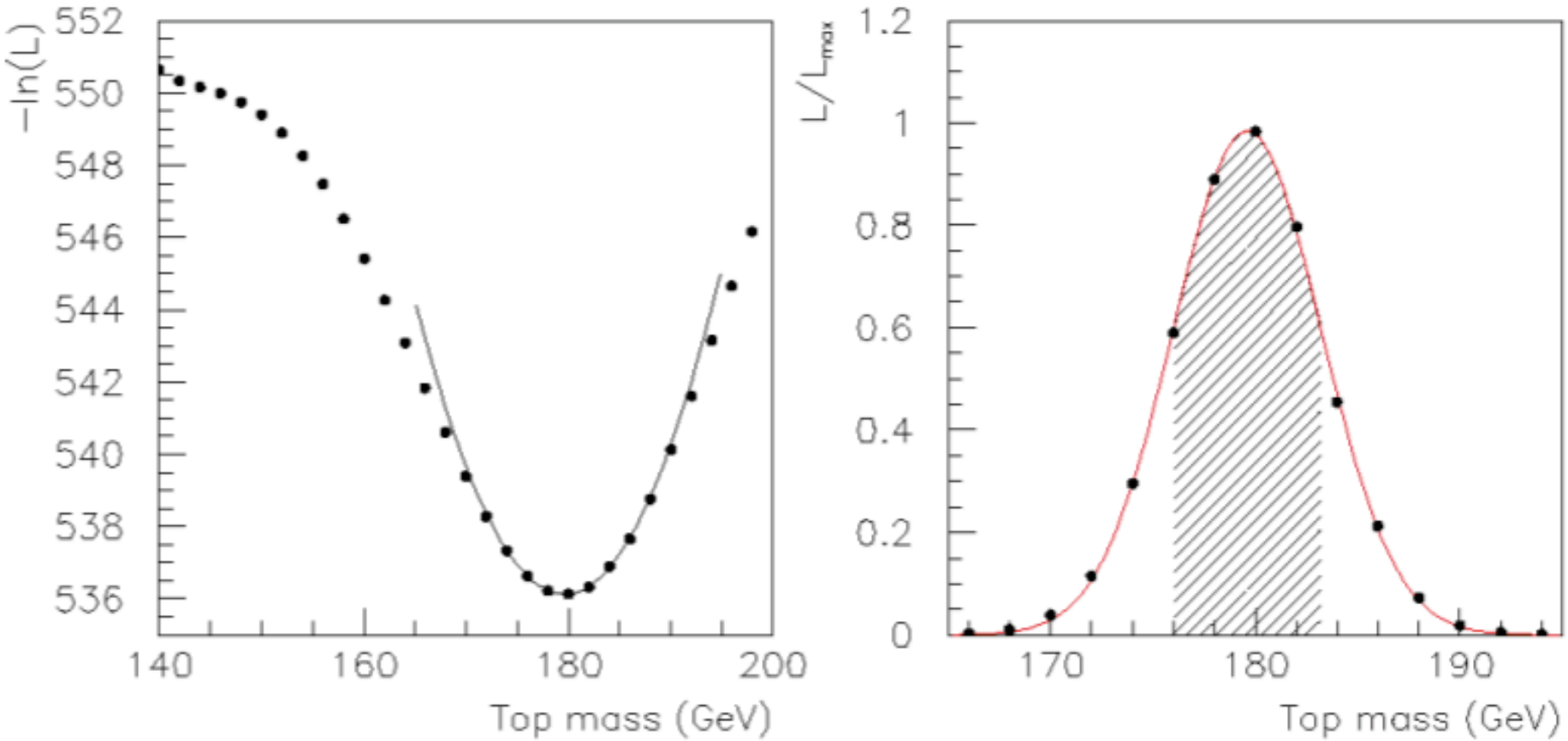
Background probability



Discriminator

Comparison of (16 signal + 55 background) MC and data sample before the background probability selection.

Run I data



$M_t = 180.1 \pm 3.6$ GeV $\pm$ SYST - preliminary

This new technique improves the statistical error on M_t from 5.6 GeV [PRD 58 52001, (1998)] to 3.6 GeV. This is equivalent to a factor of 2.4 in the number of events. 22 events pass our cuts, from fit: (12 s + 10 b)

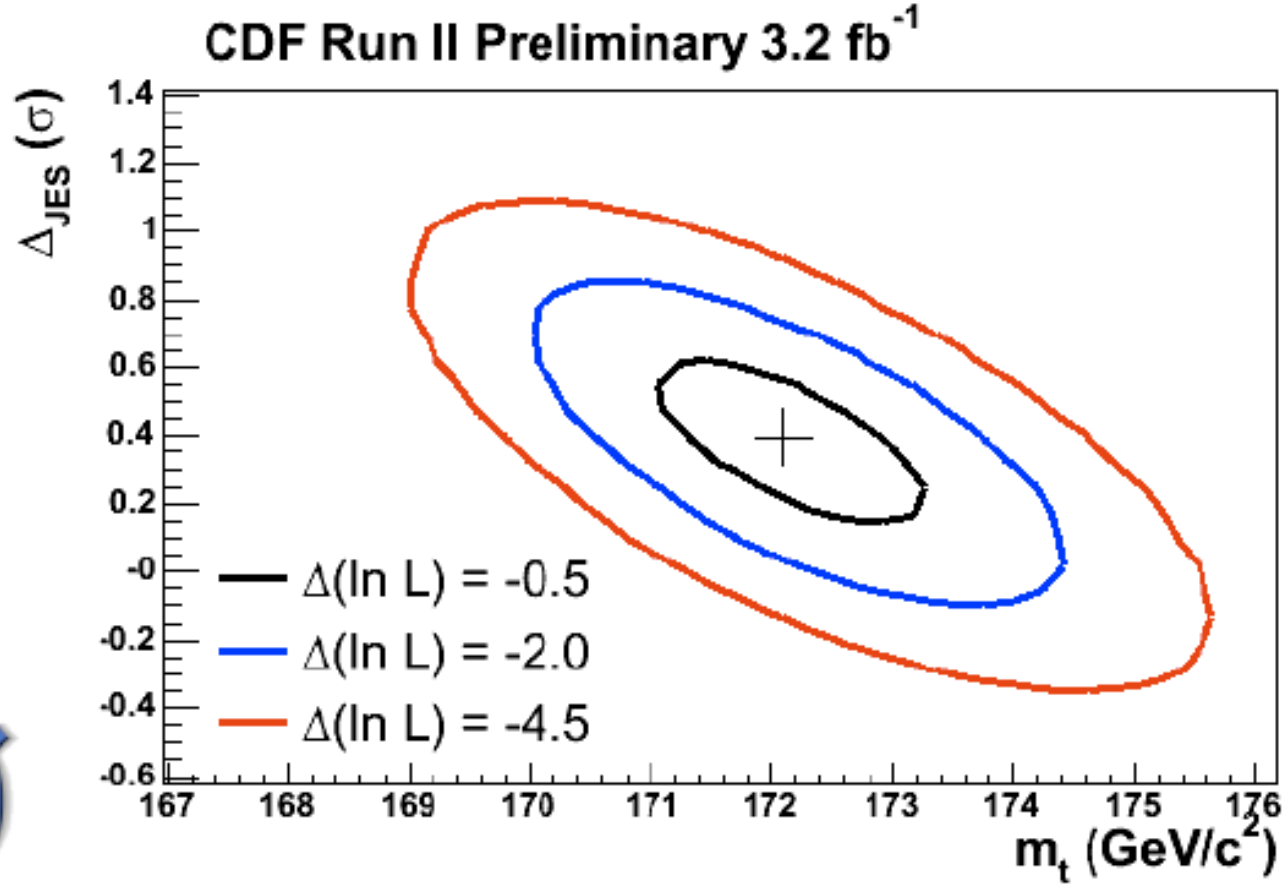
Top mass in Run II

- ME Run I measurement:

  M_t=180.1 ± 3.6 (stat) ±4.0 (syst) GeV

- In Run II - much larger statistics
  - measurements become limited by systematics
  - the largest one is jet energy scale (JES)
- Another idea from DØ: instead of varying JES in top mass likelihood to get a systematic error on the mass, find minimum of likelihood that is a function of BOTH top mass and JES

Example Results on m_{top}



m_{top} =
173.7 ± 0.8 (stat) ± 1.6 (syst) GeV

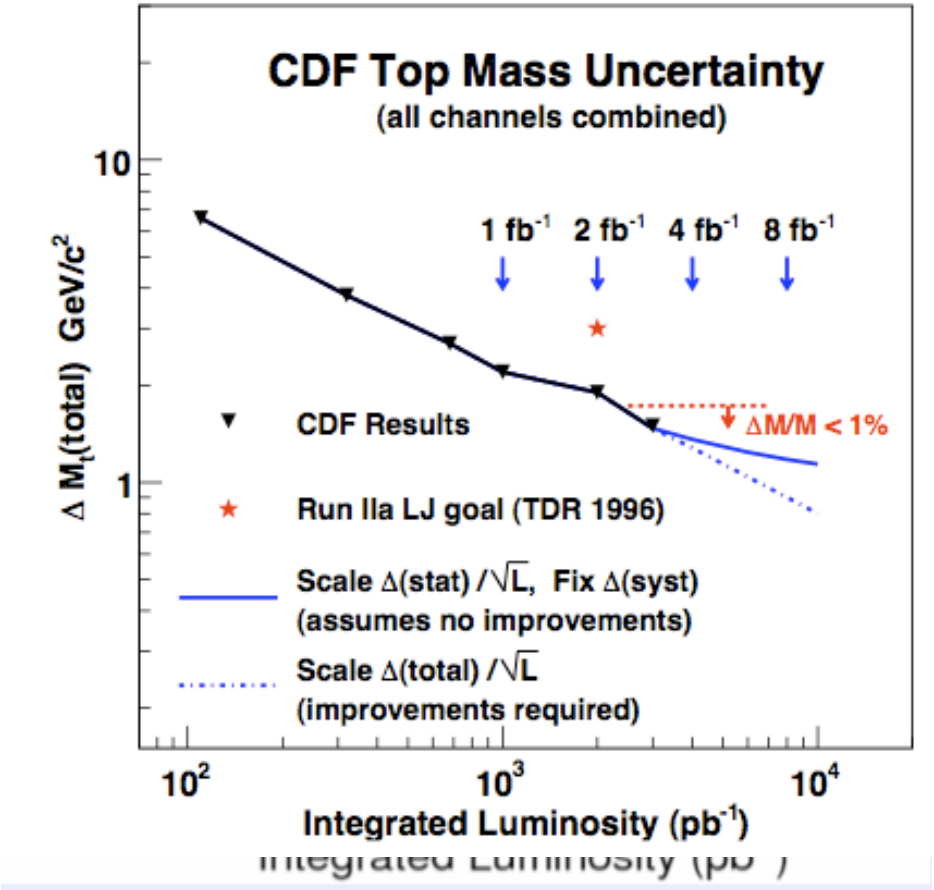
3.6 fb^{-1} ±1.0%

m_{top} =
172.1 ± 0.9 (stat) ± 1.3 (syst) GeV

3.2 fb^{-1} ±0.9%

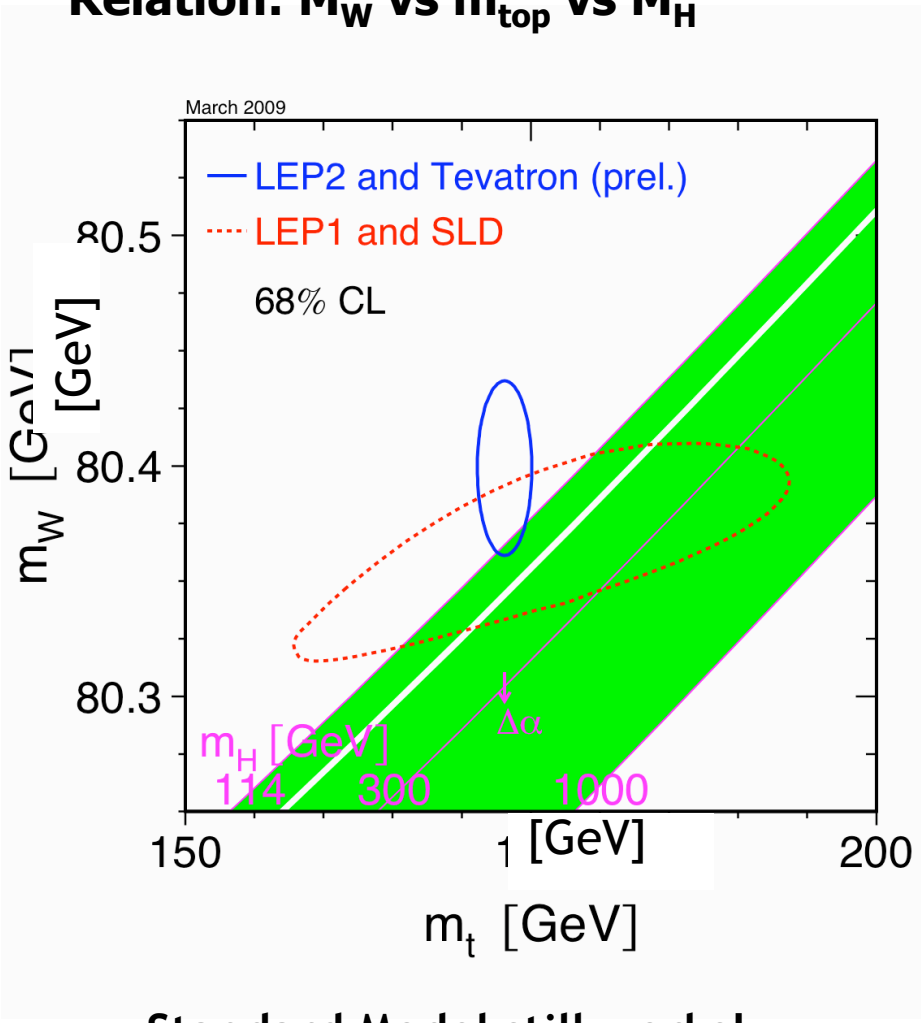
Future of M_{top}

- Once experimentalists have data, there is no limit to our ingenuity!!



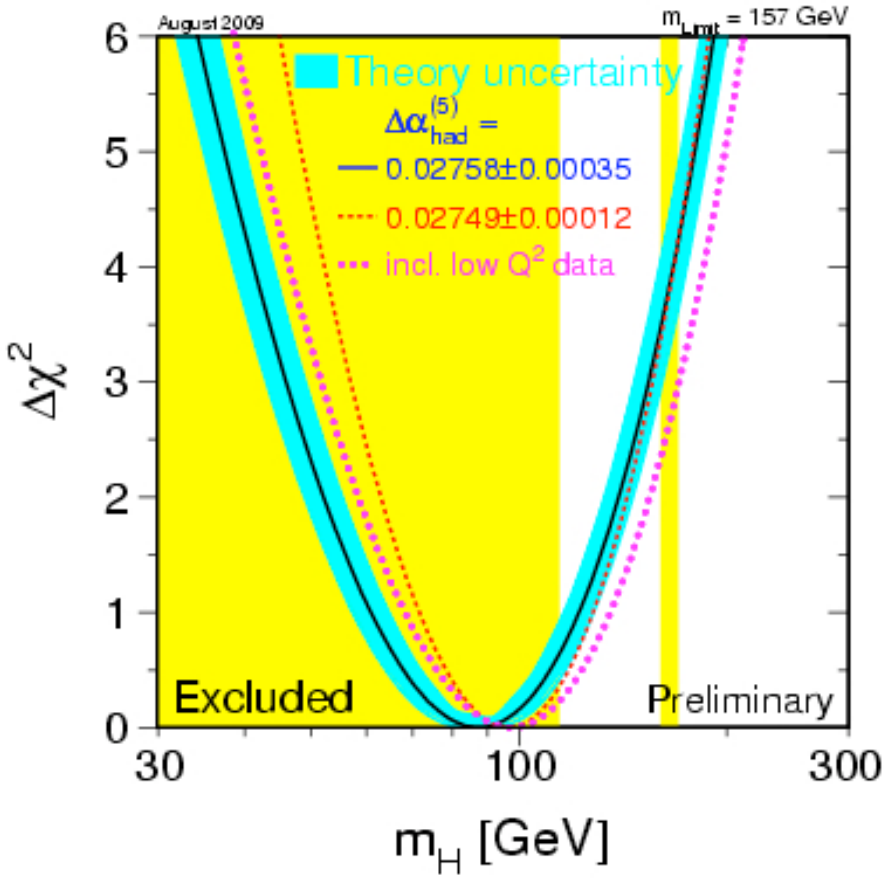
Implications for the Higgs Boson

Relation: M_W vs m_{top} vs M_H



Standard Model still works!

$m_H = 87^{+35}_{-26}$ GeV

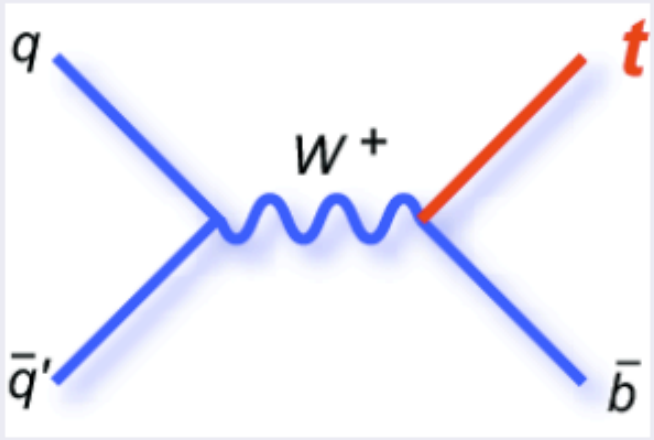


Indirect constraints:
m_H<163 GeV @95%CL

Single top

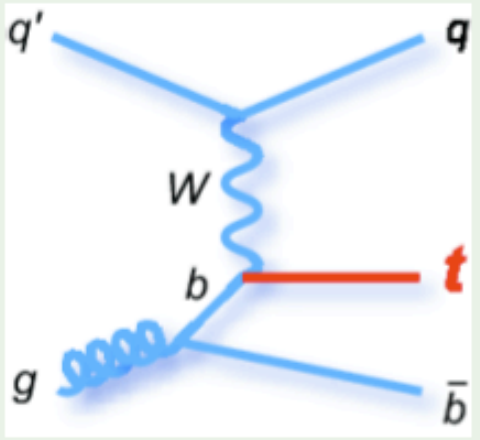
- A.k.a. electroweak production of top quark
- A great way to test if top is actually THE top quark

s-channel (tb)



- $\sigma_{NLO} = 0.88 \pm 0.11$ pb(*)

t-channel (tqb)



- $\sigma_{NLO} = 1.98 \pm 0.25$ pb (*)

- Final state – W + 2 b-jets (+ sometimes q)
  - same as low mass Higgs
- **A field day for multivariate analyses**

Decision Trees

DT Choices

- 1/3 of MC for training
- Adaboost $\beta = 0.2$
- Boosting cycles $= 20$
- Signal leaf if purity > 0.5
- Minimum leaf size $= 100$ events
- Same total weight to signal and background to start

Analysis Strategy

- Train 36 separate trees: $(s,t,s+t)$ x (e,μ) x (2,3,4 jets) x (1,2 tags)
- For each signal train against the sum of backgrounds

Decision Tree for the First Evidence Analysis: 49 variables!

Object Kinematics

- p_T(jet1)
- p_T(jet2)
- p_T(jet3)
- p_T(jet4)
- p_T(best1)
- p_T(notbest1)
- p_T(notbest2)
- p_T(tag1)
- p_T(untag1)
- p_T(untag2)

Angular Correlations

- ΔR(jet1,jet2)
- cos(best1,lepton)$_{\mathrm{besttop}}$
- cos(best1,notbest1)$_{\mathrm{besttop}}$
- cos(tag1,alljets)$_{\mathrm{alljets}}$
- cos(tag1,lepton)$_{\mathrm{btaggedtop}}$
- cos(jet1,alljets)$_{\mathrm{alljets}}$
- cos(jet1,lepton)$_{\mathrm{btaggedtop}}$
- cos(jet2,alljets)$_{\mathrm{alljets}}$
- cos(jet2,lepton)$_{\mathrm{btaggedtop}}$
- cos(lepton,Q(lepton)$\times z$)$_{\mathrm{besttop}}$
- cos(lepton,besttopframe)$_{\mathrm{besttopCMframe}}$
- cos(lepton,btaggedtopframe)$_{\mathrm{btaggedtopCMframe}}$
- cos(notbest,alljets)$_{\mathrm{alljets}}$
- cos(notbest,lepton)$_{\mathrm{besttop}}$
- cos(untag1,alljets)$_{\mathrm{alljets}}$
- cos(untag1,lepton)$_{\mathrm{btaggedtop}}$

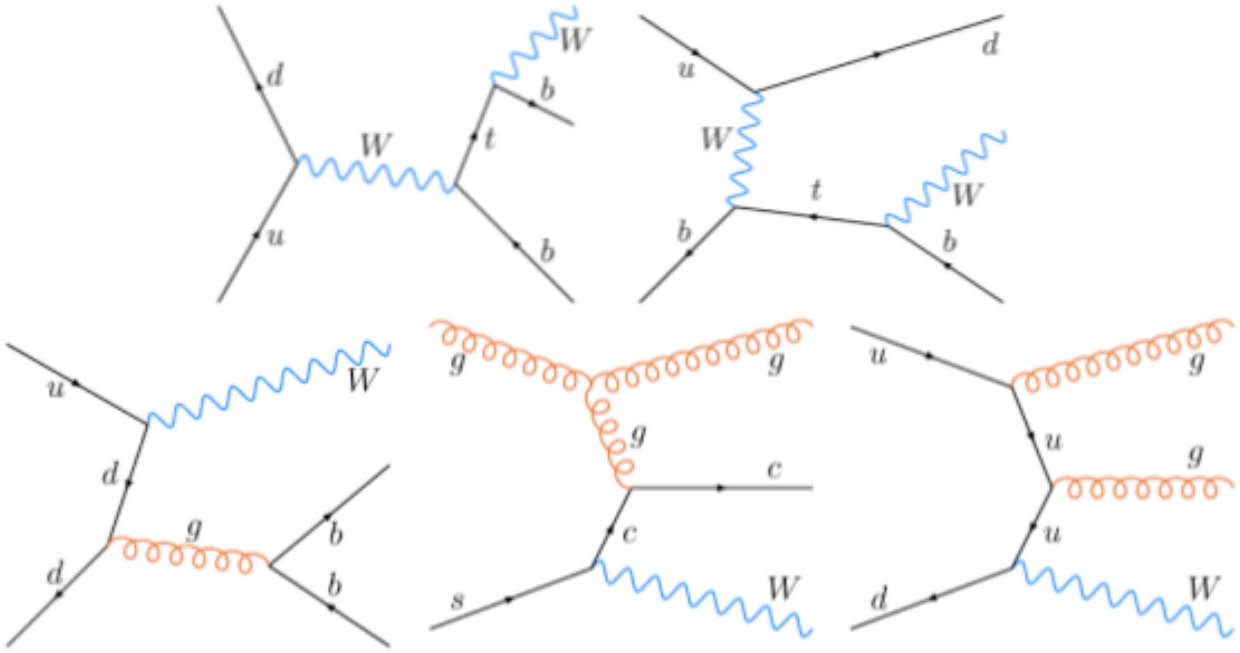
Event Kinematics

- Aplanarity(alljets,W)
- M(W,best1) ("best" top mass)
- M(W,tag1) ("b-tagged" top mass)
- H_T(alljets)
- H_T(alljets$-$best1)
- H_T(alljets$-$tag1)
- H_T(alljets,W)
- H_T(jet1,jet2)
- H_T(jet1,jet2,W)
- M(alljets)
- M(alljets$-$best1)
- M(alljets$-$tag1)
- M(jet1,jet2)
- M(jet1,jet2,W)
- M_T(jet1,jet2)
- M_T(W)
- Missing E_T
- p_T(alljets$-$best1)
- p_T(alljets$-$tag1)
- p_T(jet1,jet2)
- Q(lepton)$\times\eta$(untag1)
- $\sqrt{\hat{s}}$
- Sphericity(alljets,W)

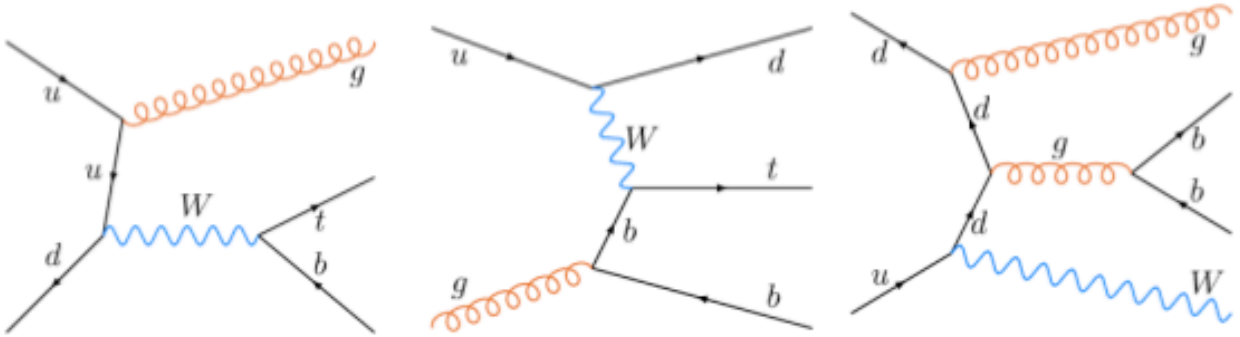
- Adding variables does not degrade performance
- Tested shorter lists, lose some sensitivity
- Same list used for all channels

Matrix Element (Elements!)

2-jets:



3-jets:



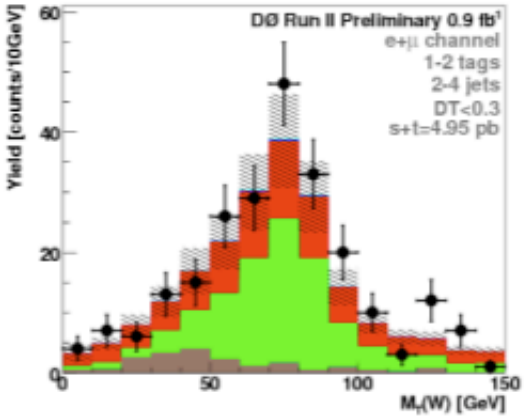
Neural Network

- used modification of MLP similar in spirit to bagging a forest of decision trees

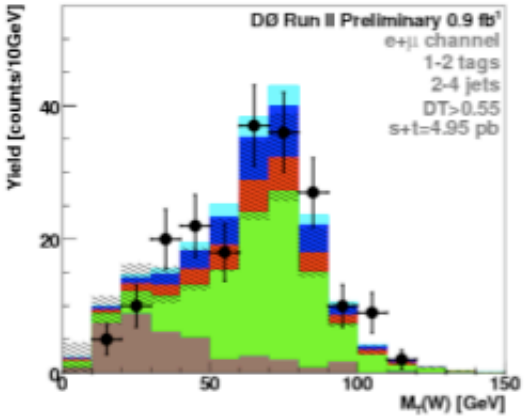
- A different sort of neural network:
  - Instead of choosing one set of weights, find posterior probability density over all possible weights
  - Averaging over many networks weighted by the probability of each network given the training data
  - Less prone to overtraining
  - For details see: http://www.cs.toronto.edu/radford/fbm.software.html
- Use 24 variables (subset of DT variables)

First Evidence

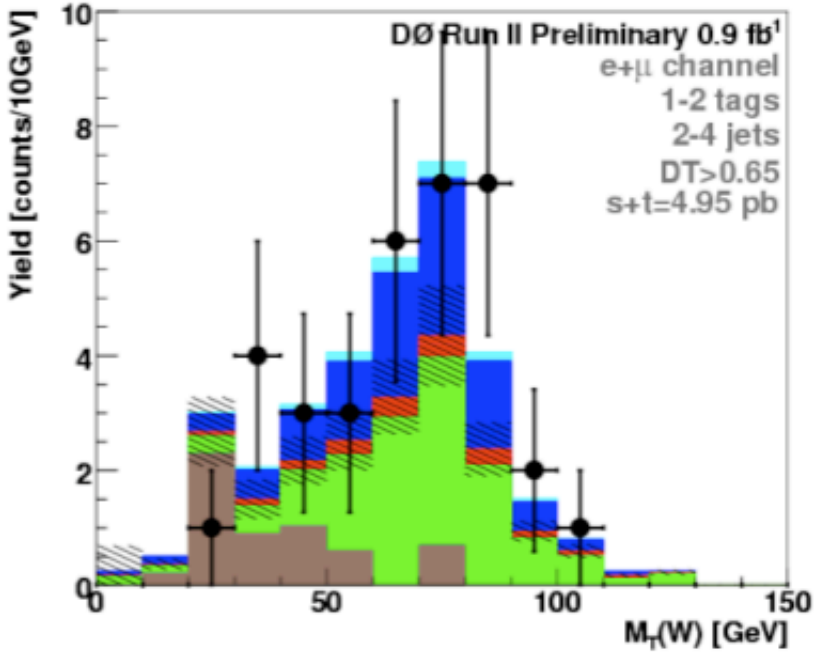
$DT < 0.3$



$DT > 0.55$

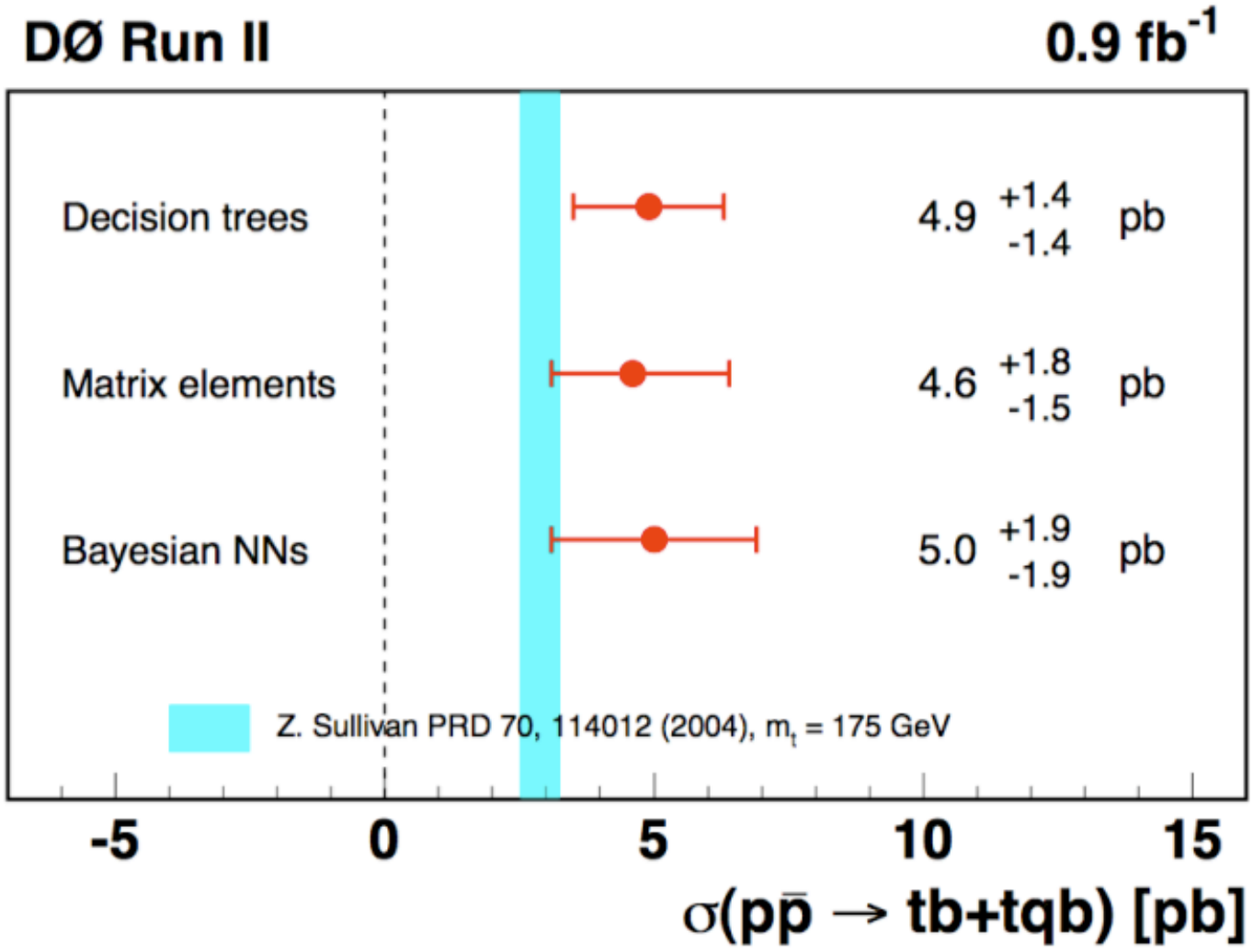


$DT > 0.65$



- Excess in high DT output region.

First Evidence



more blood can be squeezed from this stone!

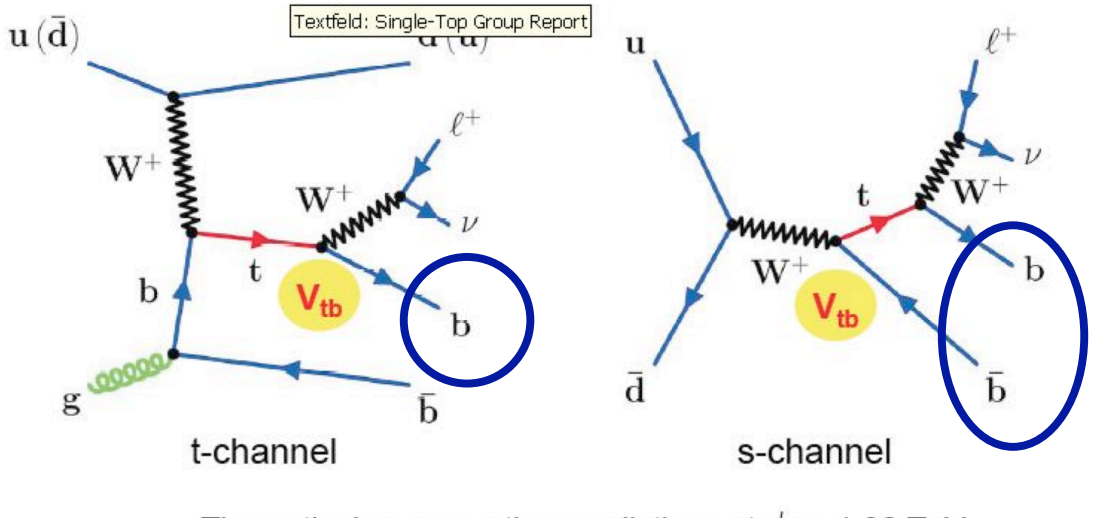
Technique	Electron	Muon
DT vs ME	52%	58%
DT vs BNN	56%	48%
ME vs BNN	46%	52%

Also measured the cross section in 400 members of the SM ensemble with all three techniques and calculated the linear correlation between each pair:

	DT	ME	BNN
DT	100%	39%	57%
ME		100%	29%
BNN			100%

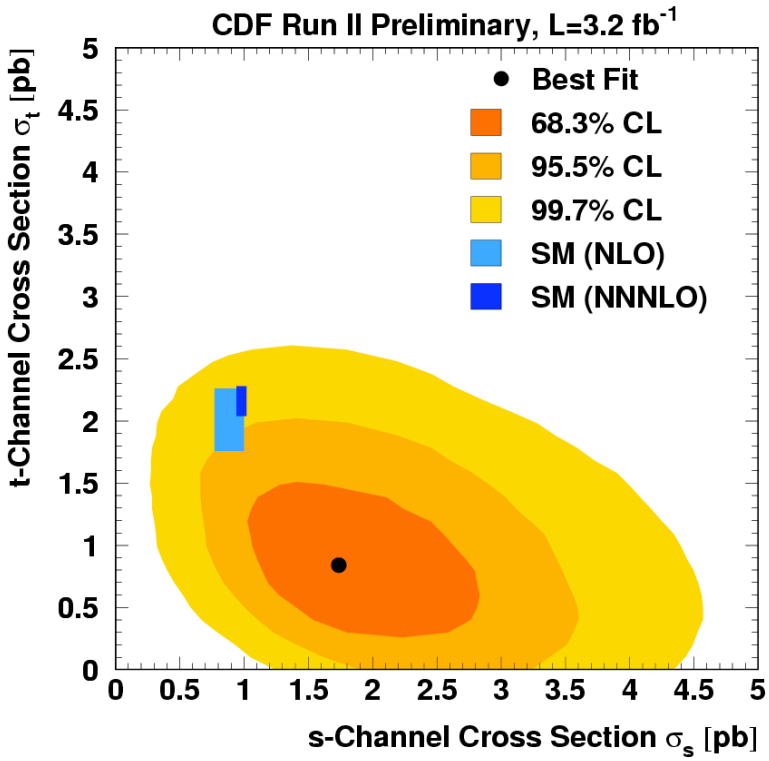
Single Top: now at CDF

- Single top observed. 3.2fb^{-1} $\sigma_{ST} = 2.3^{+0.6}_{-0.5}$ pb, $5.9\,\sigma$ significance
- Separately measure s and t channel production.
  - Measurement driven by statistics of single and double tag events



$\sigma_t = 0.8 \pm 0.4$ pb

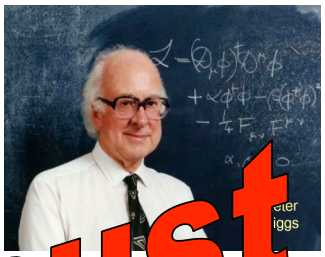
$\sigma_s = 1.8 \pm {}^{0.7}_{-0.5}$ pb



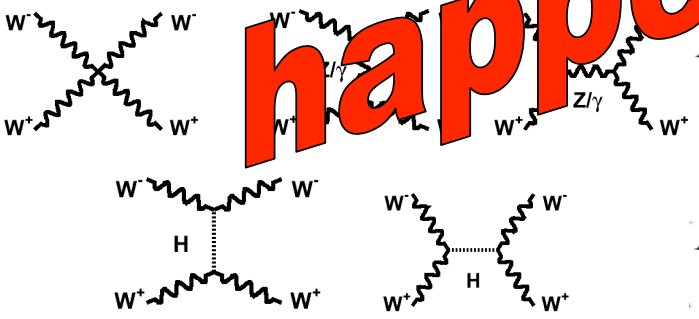
The Higgs Boson

- Electroweak Symmetry breaking caused by scalar Higgs field
- vacuum expectation value of the Higgs field $\langle\Phi\rangle = 246\ \text{GeV}/c^2$
  - gives mass to the W and Z gauge bosons,
    - $M_W \propto g_W\langle\Phi\rangle$
  - fermions gain a mass by Yukawa interactions with the Higgs field,
    - $m_f \propto g_f\langle\Phi\rangle$
  - Higgs boson couplings are proportional to mass
- Higgs boson prevents unitarity violation of WW cross section
  - $\sigma(pp \rightarrow WW) > \sigma(pp \rightarrow \text{anything})$
    - => illegal!
    - At $\sqrt{s}$=1.4 TeV!

Peter Higgs



Something new must happen at the LHC!



$$A \approx g^2 \frac{E^2}{M_W^2}$$

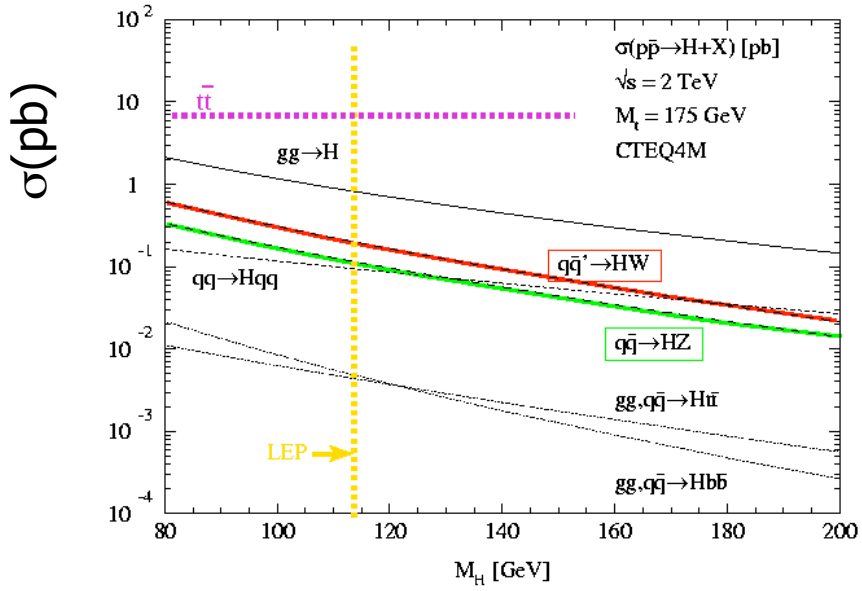
$$A \approx -g^2 \frac{E^2}{M_W^2}$$

Terms which grow with energy cancel for E >> M_H

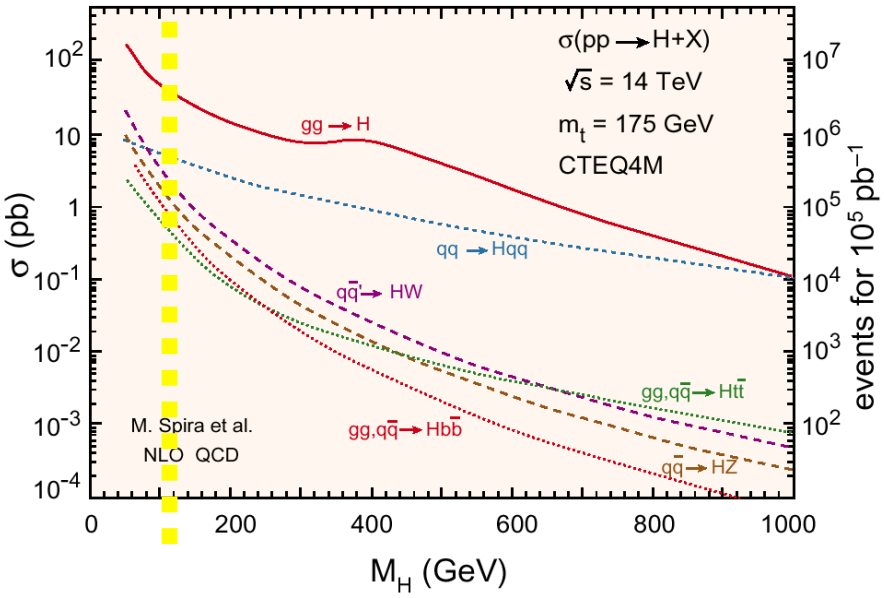
This cancellation requires $M_H < 800$ GeV

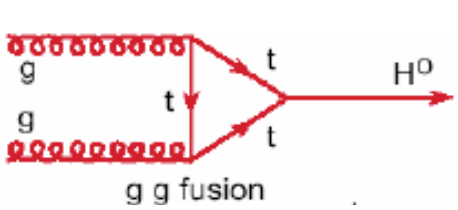
Higgs Production: Tevatron and LHC

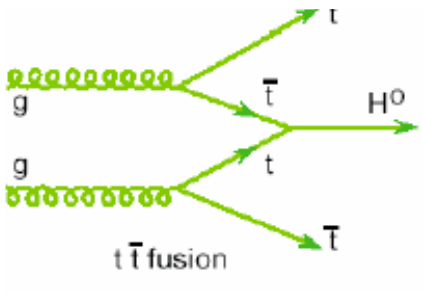
Tevatron

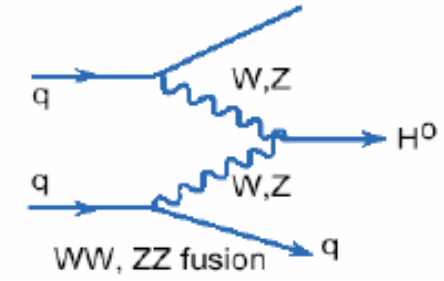


LHC





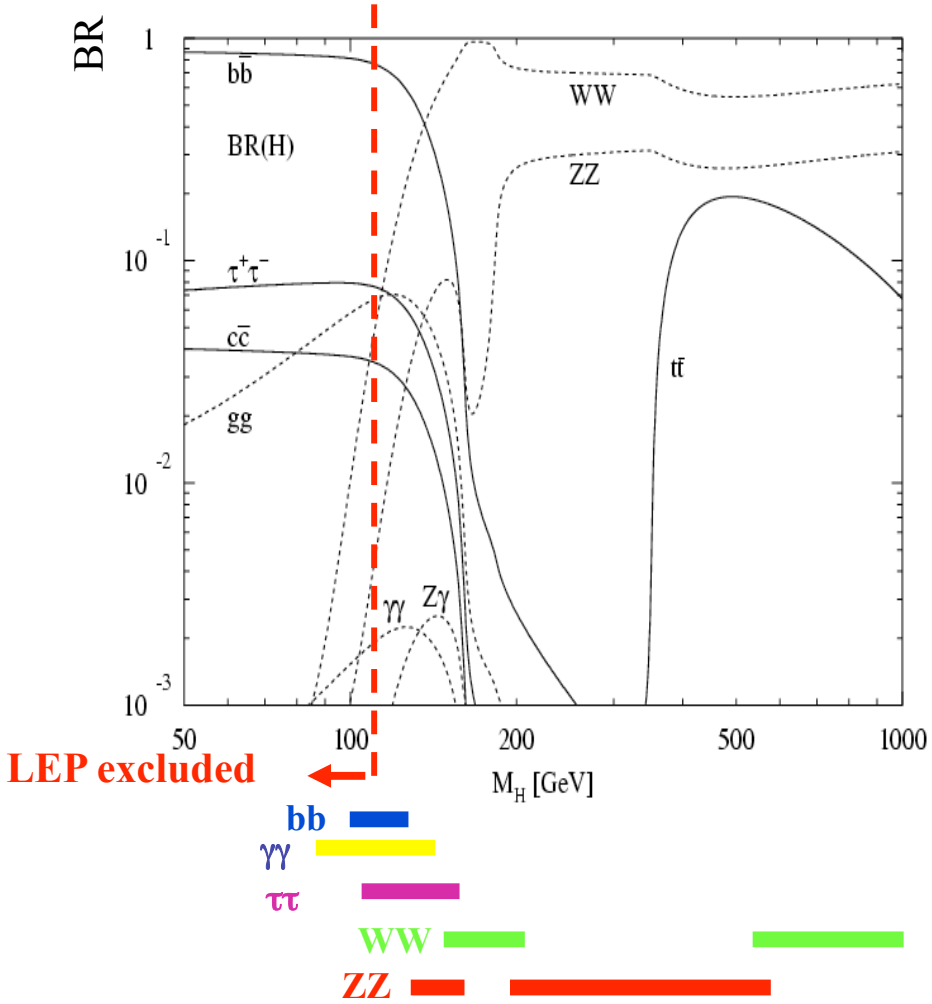




dominant: gg→ H, subdominant: HW, HZ, Hqq

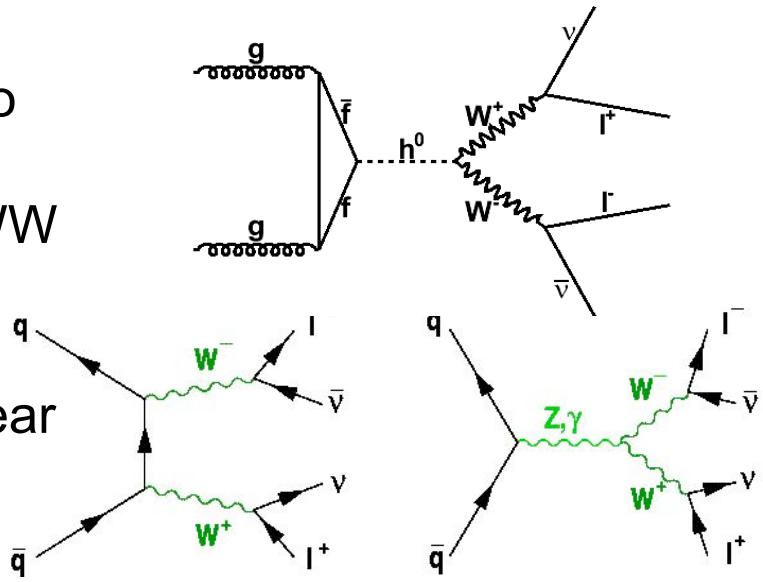
Higgs Boson Decay

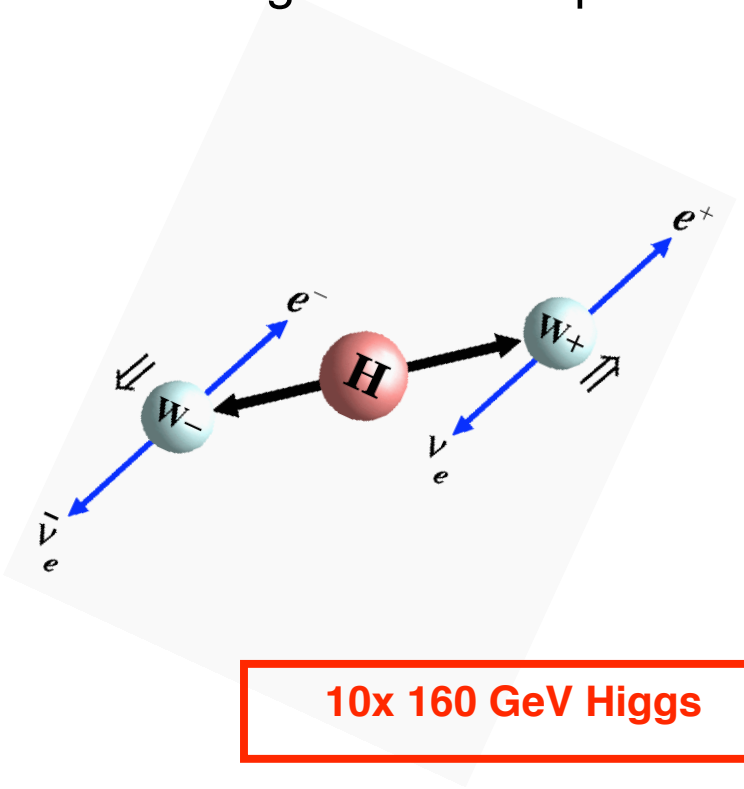
- Depends on Mass
- M_H<130 GeV/c^2:
  - $b\bar{b}$ dominant
  - WW and $\tau\tau$ subdominant
  - $\gamma\gamma$ small but useful
- M_H>130 GeV/c^2:
  - WW dominant
  - ZZ cleanest

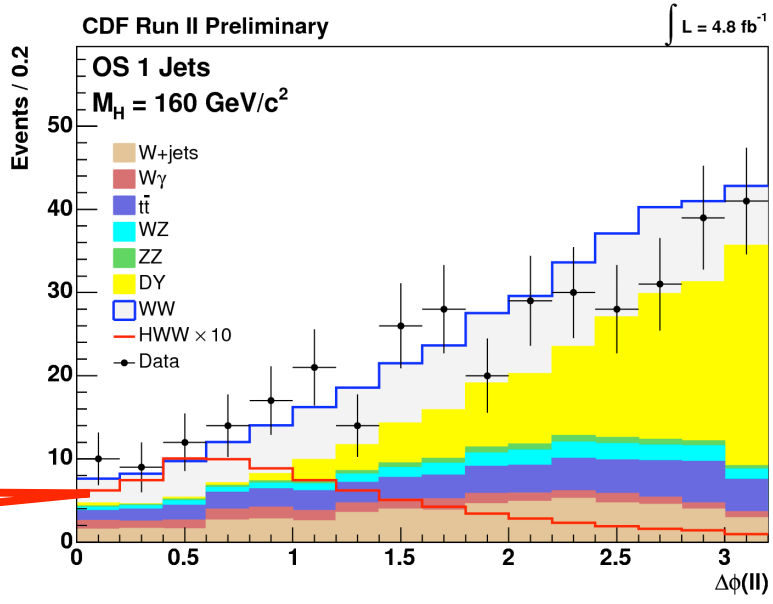


High Mass: $H \rightarrow WW^{(*)} \rightarrow l^+l^-\nu\nu$

- Higgs mass reconstruction impossible due to two neutrinos in final state
- Make use of spin correlations to suppress WW background:
  - Higgs is scalar: spin=0
  - leptons in $H \rightarrow WW^{(*)} \rightarrow l^+l^-\nu\nu$ are collinear
- Main background: WW production





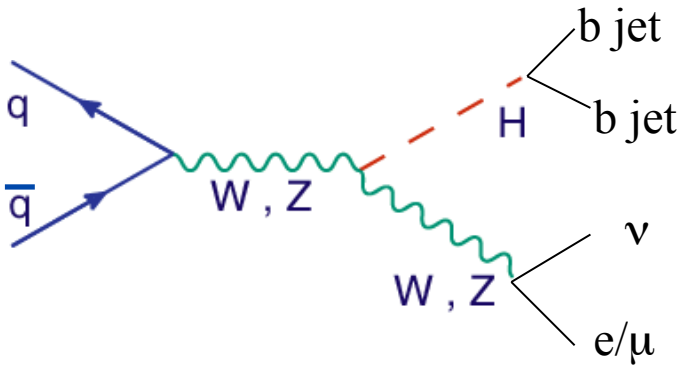


10x 160 GeV Higgs

Low Mass Higgs: m_H<140 GeV

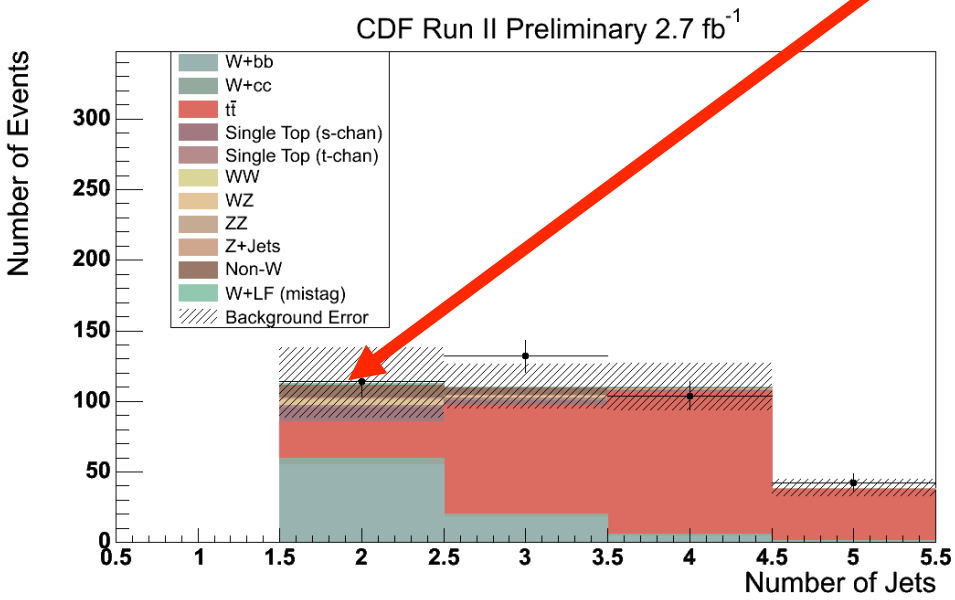
- Tevatron:
  - WH($\rightarrow$bb), ZH($\rightarrow$bb)
- LHC:
  - H($\rightarrow\gamma\gamma$), qqH($\rightarrow\tau\tau$/WW*)
  - may be other modes with very high L

WH→lνbb



- **WH selection:**
  - **1 or 2 tagged b-jets**
  - **electron or muon with p_T > 20 GeV**
  - **E_T^{miss} > 20 GeV**

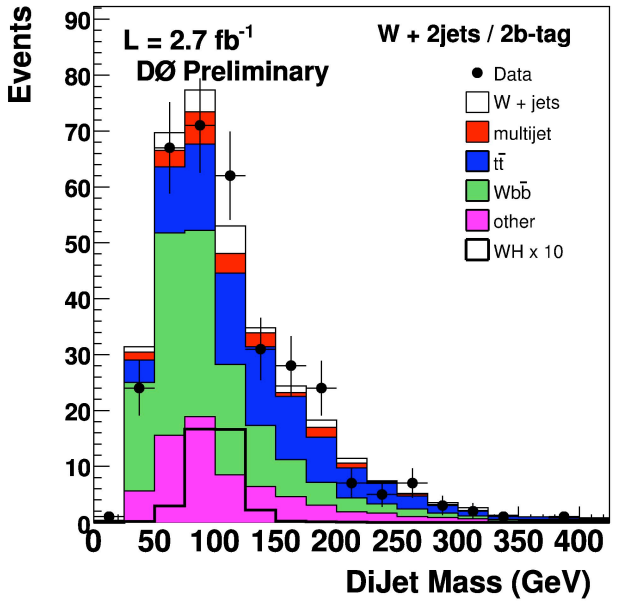
Looking for 2 jets

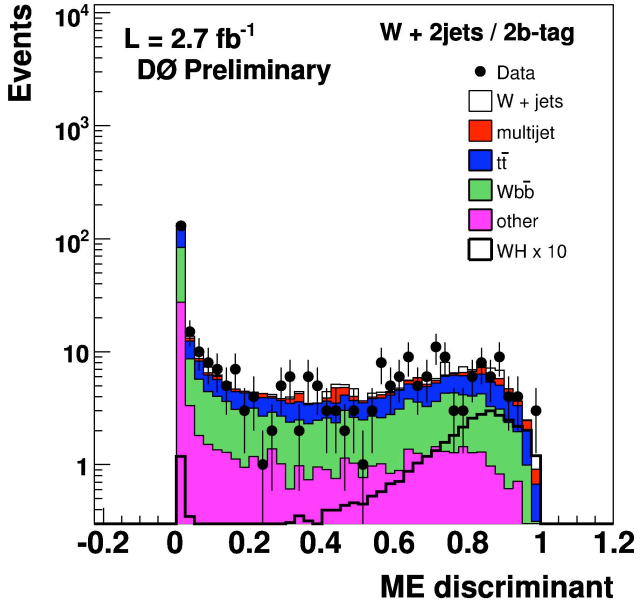


Expected Numbers of Events for 2 b-tags:
WH signal: 1.6
Background: 110±25

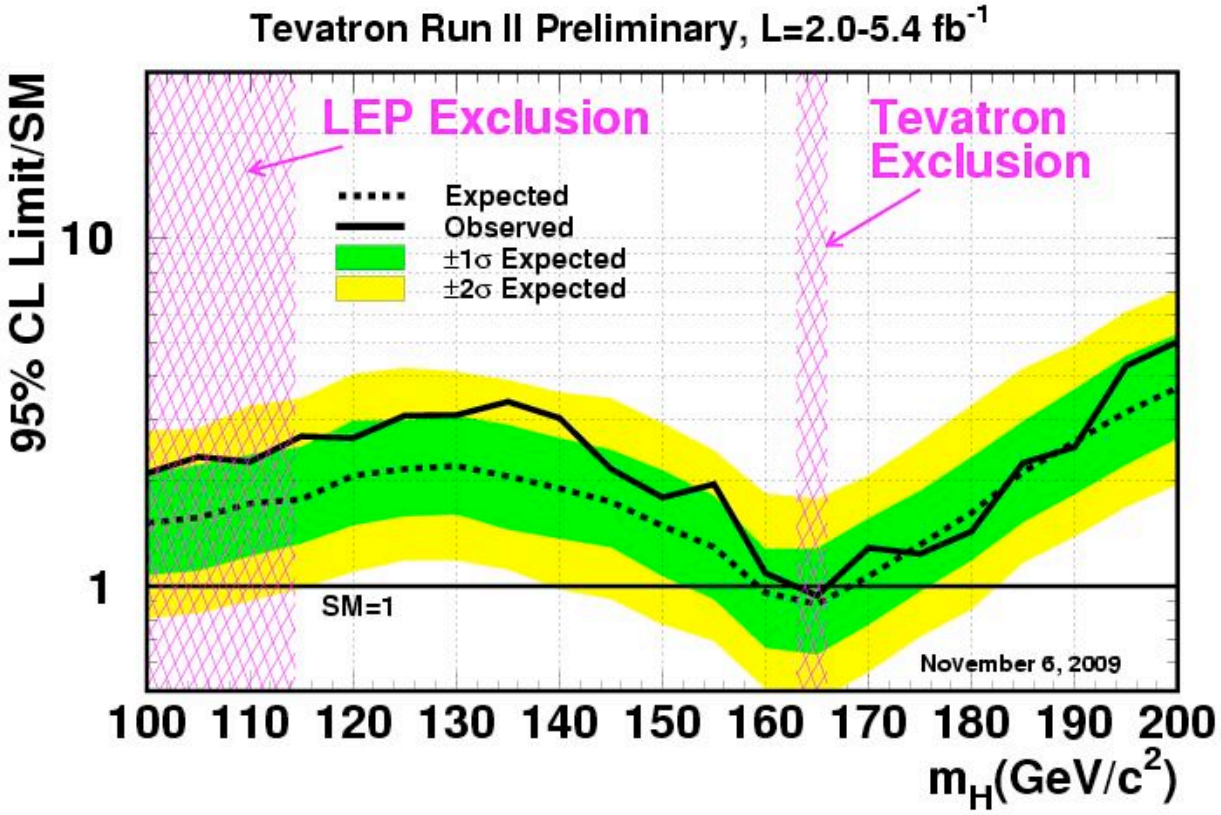
WH Dijet Mass distributions

- Use discriminant to separate signal from backgrounds:
  - Invariant mass of the two b-jets
    - Signal peaks at m(bb)=m_H
    - Background has smooth distribution
  - More complex:
    - Neural network or other advanced techniques
- Backgrounds still much larger than the signal:
  - Further experimental improvements and luminosity required
  - E.g. b-tagging efficiency (40->60%), *NN/ME selection*, higher lepton acceptance
- Similar analyses for ZH





Tevatron Combined Status



- Combine CDF and DØ analyses from all channels at low and high mass
  - Exclude m_H=163-166 GeV/c^2 at 95% C.L.
  - m_H=120 GeV/c^2: limit/SM=2.8

Future of the Tevatron

- Running confirmed till 2011
  - chance to exclude large fraction of Higgs masses if Higgs is not there
  - very slim chance of evidence
- Running until 2014 is being considered
- My personal view
  - LHC schedule is not important – probably only Tevatron can measure h→bb if higgs mass is low
  - There is also value in beam asymmetry – makes measurements like top charge asymmetry (which is currently ~2σ anomalous) possible
  - Unfortunately physics is not the only consideration…