Boosted decision trees: single top quark production evidence at DØ

Yann Coadou

CERN

Workshop on Top Physics: from the TeVatron to the LHC
Grenoble, 20 October 2007

Wish I were here...

Them

"We don't want to work more!!!"

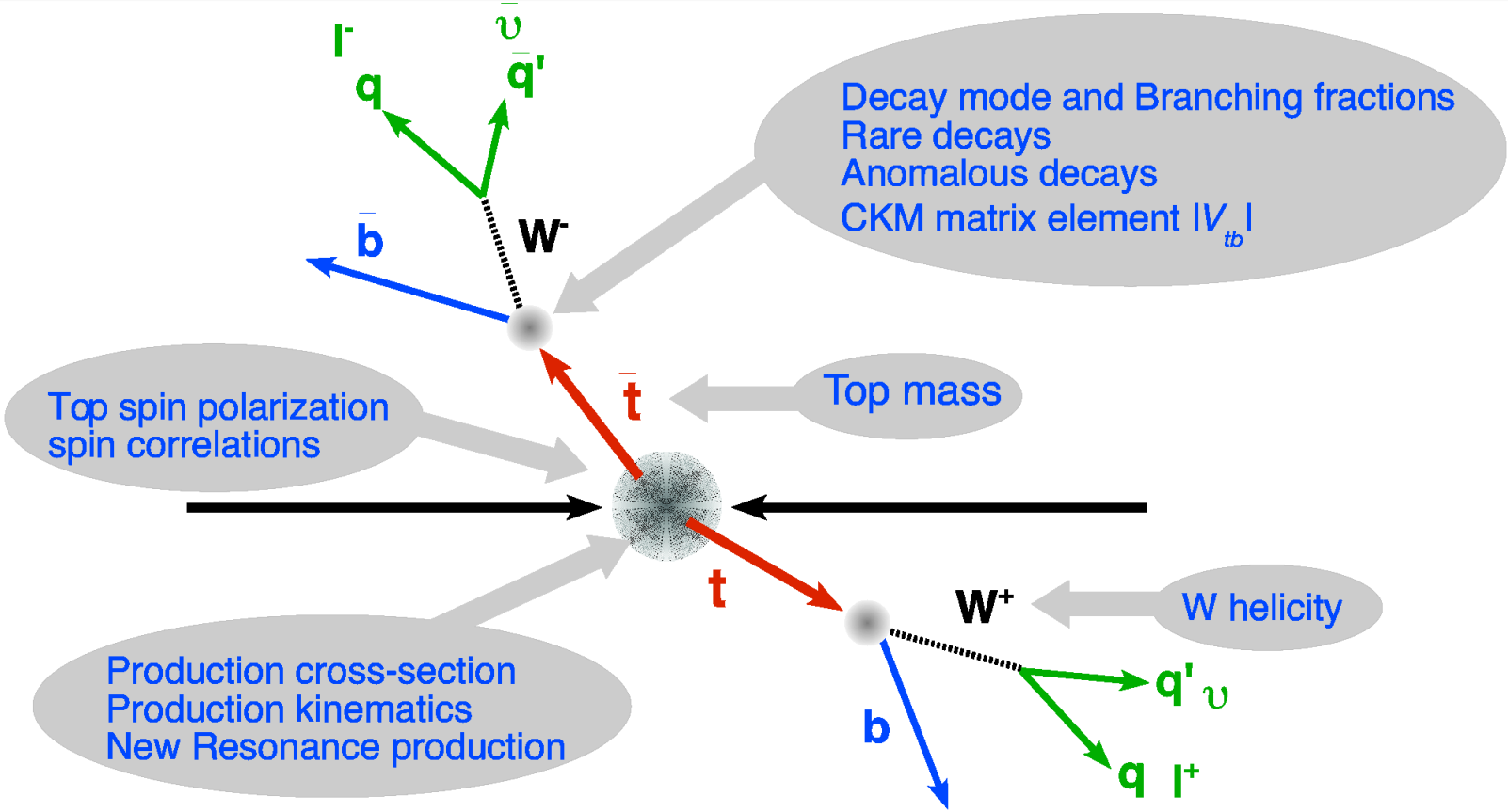
Poor Me

"I want to go to Grenoble to work on Saturday!"

Thank you Harrison!

Top quark physics

- top quark discovered in 1995 by CDF and DØ at the Tevatron
- Heaviest of all fermions
- Couples strongly to Higgs boson
- So far only observed in pairs, only at the Tevatron

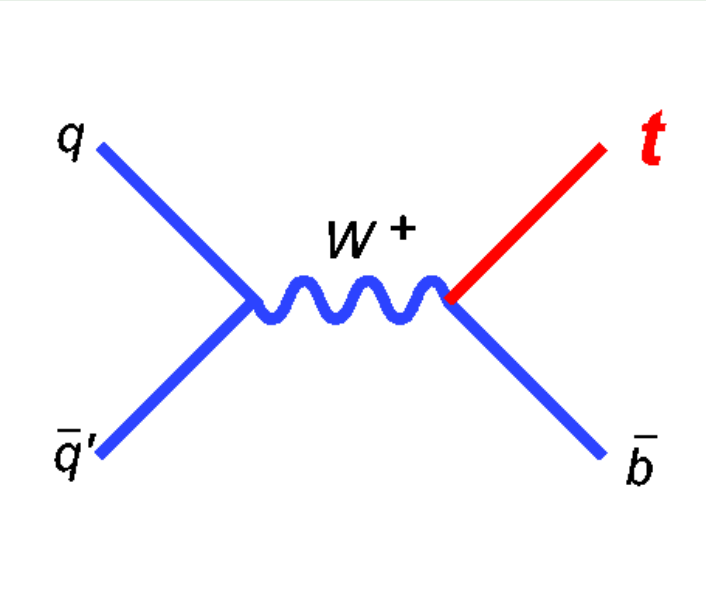


©B.Vachon

Single top quark production

- Never observed before: electroweak production

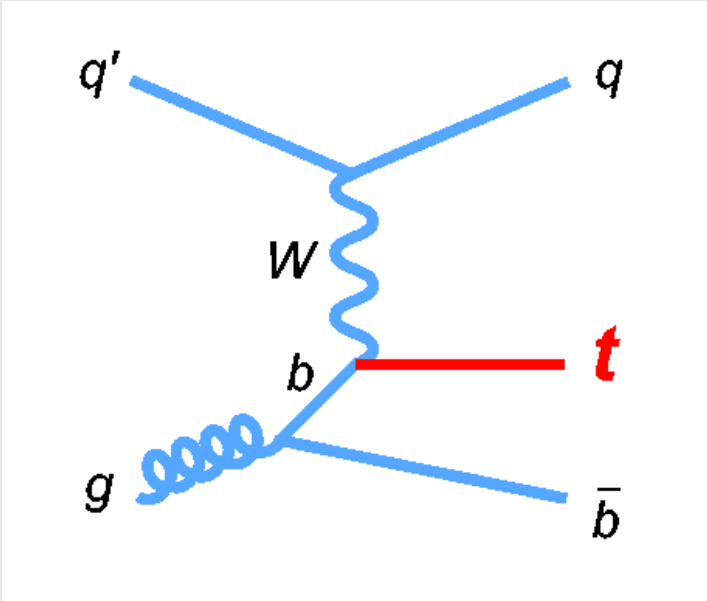
s-channel (tb)



- $\sigma_{NLO} = 0.88 \pm 0.11$ pb (*)
- previous limits (95% C.L.):

Run II DØ: < 5.0 pb (370 pb^{-1})
Run II CDF: < 3.1 pb (700 pb^{-1})

t-channel (tqb)



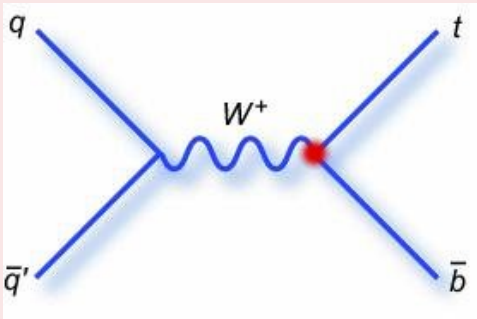
- $\sigma_{NLO} = 1.98 \pm 0.25$ pb(*)
- previous limits (95% C.L.):

Run II DØ: < 4.4 pb (370 pb^{-1})
Run II CDF: < 3.2 pb (700 pb^{-1})

(*) $m_t = 175$ GeV, Phys.Rev. D70 (2004) 114012

Why do we care? — $|V_{tb}|$, new physics

- Has never been observed before!
- It should happen in SM
- First measurement of $|V_{tb}|$



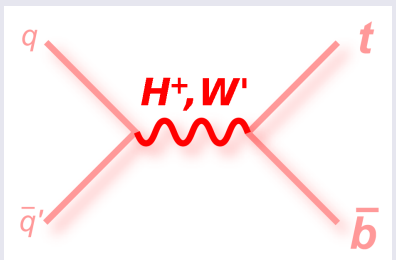
Direct access to $|V_{tb}|$

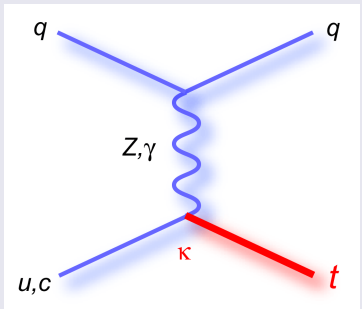
$$V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}$$

- In SM, from constraints on V_{td} and V_{ts}: $|V_{tb}| = 0.9991^{+0.000034}_{-0.000004}$
- New physics, e.g. 4^{th} generation: $0.07 < |V_{tb}| < 0.9993$

New physics

- s and t channel cross sections differently sensitive
- s-channel: charged resonances (heavy W' boson, charged Higgs boson, charged top pion, etc.)
- t-channel: new interactions (FCNC, 4^{th} generation, etc.)



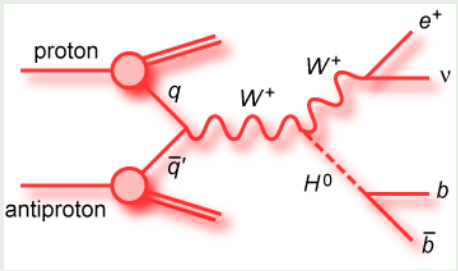


Why do we care? — Spin, Higgs, analysis techniques

Top quark spin

- Large mass ⇒ top quark decays before it can hadronize (no top jets)
- First chance to study a bare quark!
- Top polarization reflected in angular distributions of decay products
- SM predicts high degree of left-handed tops ⇒ possible sign of new physics, or help pin down what new physics

Higgs searches



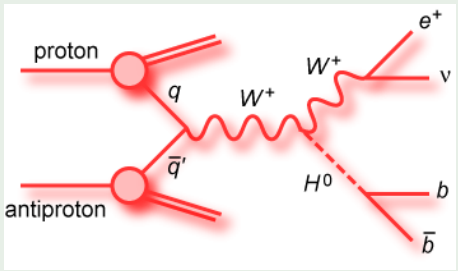
- Important background to *WH* associated Higgs production
- As soon as we discover it, somebody will try to get rid of it....

Why do we care? — Spin, Higgs, analysis techniques

Top quark spin

- Large mass ⇒ top quark decays before it can hadronize (no top jets)
- First chance to study a bare quark!
- Top polarization reflected in angular distributions of decay products
- SM predicts high degree of left-handed tops ⇒ possible sign of new physics, or help pin down what new physics

Higgs searches



- Important background to *WH* associated Higgs production
- As soon as we discover it, somebody will try to get rid of it....

Advanced analysis techniques

- Test of techniques to extract small signal out of large background
- If tools don't work for single top, forget about the Higgs and other small signals
- If tools don't work at Tevatron, not much hope for LHC

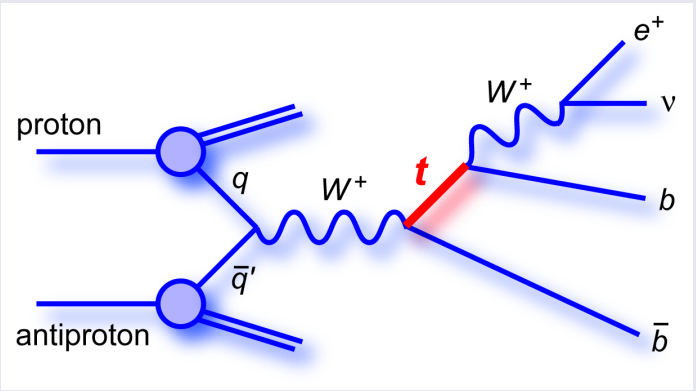
Before we go on...

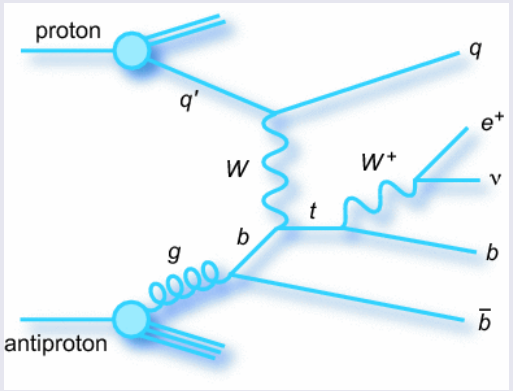
!!! VERY IMPORTANT !!!

Understand your inputs well
before you start playing with multivariate techniques

Event selection

Signal: CompHEP+Pythia





Signature

- isolated lepton
- $\not\!\!E_T$
- jets
- at least 1 b-jet

Backgrounds

W+jets

- Most difficult background
- Alpgen+Pythia (MLM matching between matrix elements and parton shower)
- Heavy flavour fraction and normalization from data

$t\bar{t}$ ($m_t = 175$ GeV)

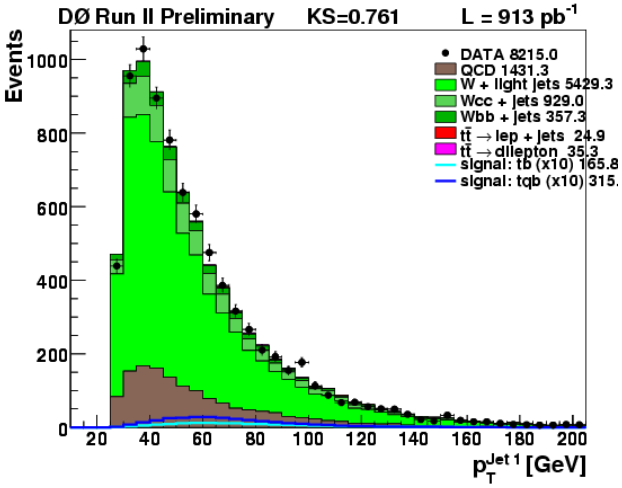
- Alpgen+Pythia (MLM)
- Normalized to $\sigma_{NNLO} = 6.8$ pb

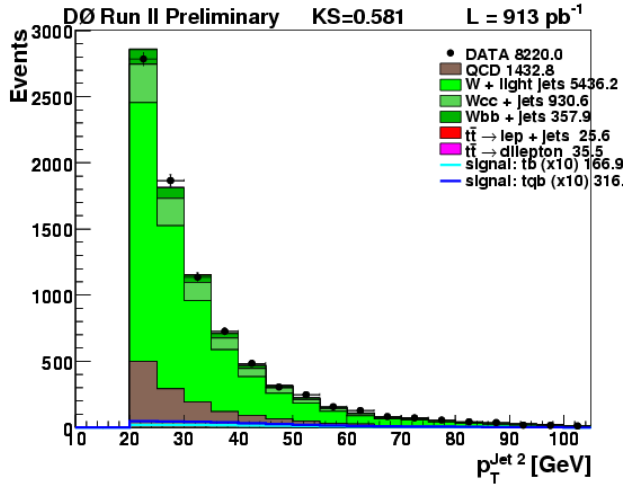
Multijet events

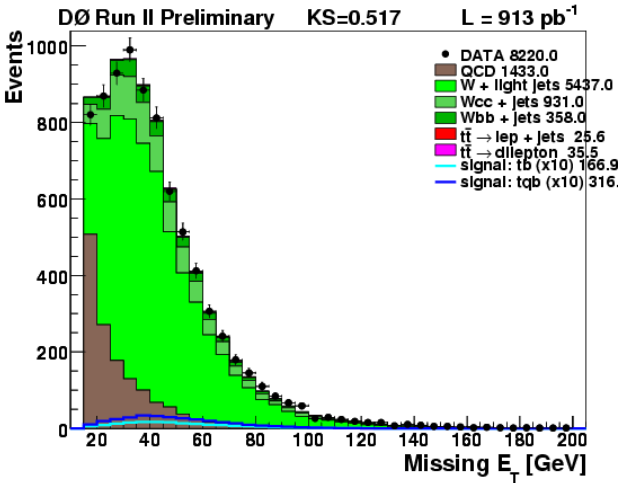
- misidentified lepton, from data

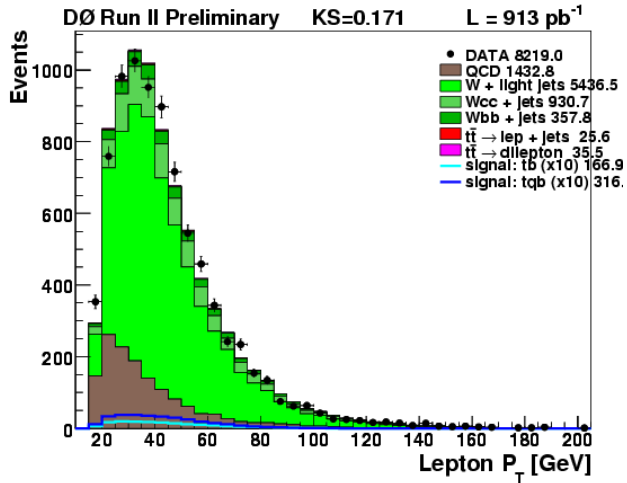
Event selection — Agreement before tagging

- Normalize W+jets and multijet to data before tagging
- Checked 90 variables, 4 jet multiplicities, electron + muon
- Good description of data

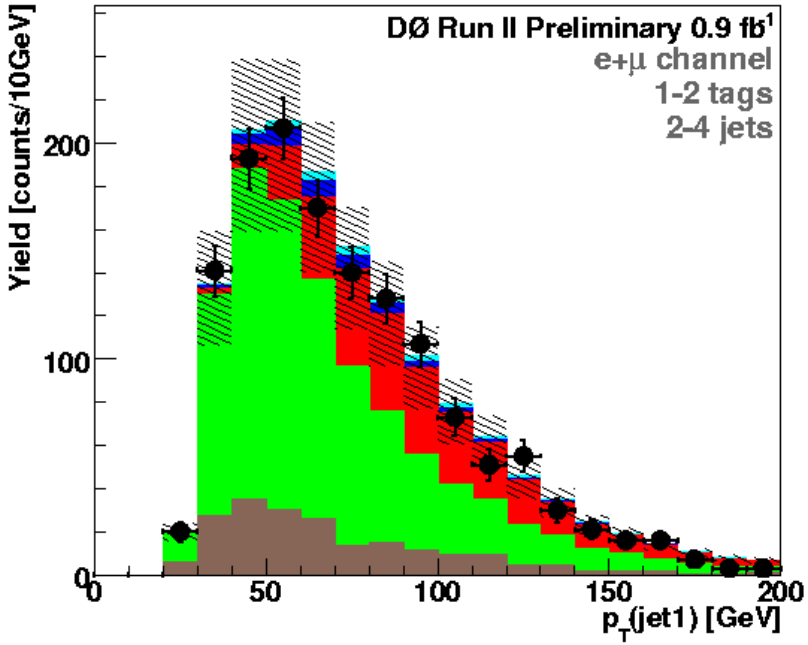






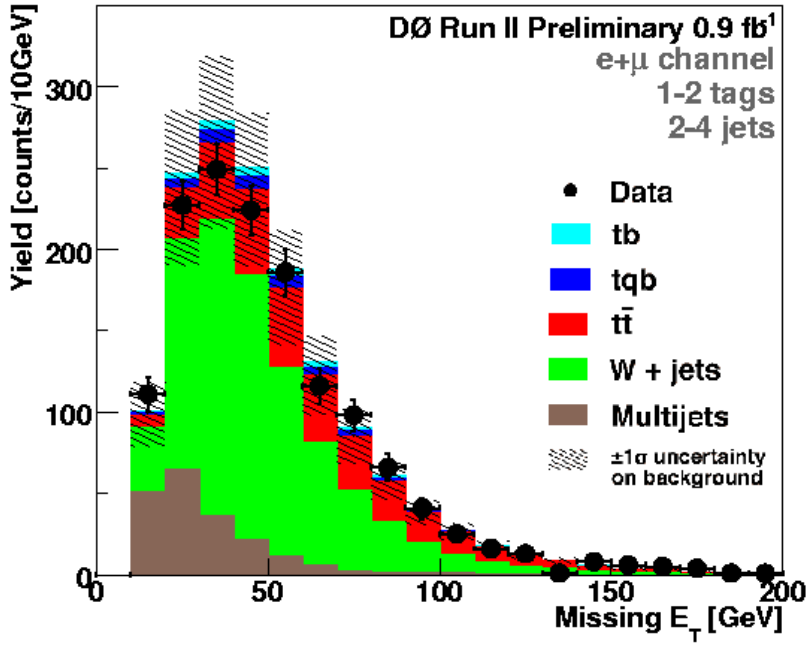


Agreement after tagging



- It took 8 months to get there...
- The signal is smaller than the background uncertainty
- Need to get smarter

Totals	2 Jets	3 Jets	4 Jets
Data	697	455	246
Total Background	685	460	253
Signal	36	20	6



Multivariate analysis techniques

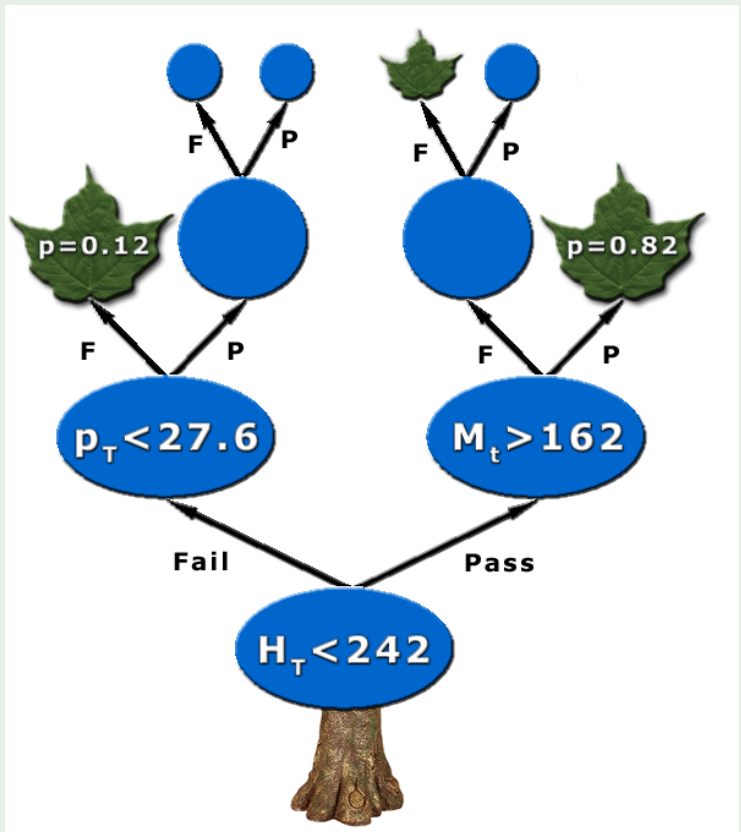
- Bayesian neural networks
- Matrix element
- Boosted decision trees

- All three analyses have similar sensitivity and give compatible measurements
- Details about decision trees only today

Decision trees

- Machine-learning technique, widely used in social sciences
- Idea: recover events that fail criteria in cut-based analysis

- Start with all events = first node
  - sort all events by each variable
  - for each variable, find splitting value with best separation between two children (mostly signal in one, mostly background in the other)
  - select variable and splitting value with best separation, produce two branches with corresponding events ((F)ailed and (P)assed cut)
- Repeat recursively on each node
- Splitting stops: terminal node = leaf



- DT output = leaf purity, close to 1 (0) for signal (bkg)

Ref: Breiman *et al*, "Classification and Regression Trees", Wadsworth (1984)

Tree construction parameters

Normalization of signal and background before training

- same total weight for signal and background events

Selection of splits

- list of questions ($variable_i > cut_i$?)
- goodness of split (next slide)

Decision to stop splitting (declare a node terminal)

- minimum leaf size (100 events)
- insufficient improvement from splitting

Assignment of terminal node to a class

- signal leaf if purity > 0.5
- background otherwise

Splitting a node

Impurity $i(t)$

- maximum for equal mix of signal and background
- symmetric in p_{signal} and $p_{background}$
- minimal for node with either signal only or background only
- strictly concave $\Rightarrow$ reward purer nodes

- Decrease of impurity for split s of node t into children t_L and t_R (goodness of split):
  $\Delta i(s,t) = i(t) - p_L \cdot i(t_L) - p_R \cdot i(t_R)$
- Aim: find split s^* such that:

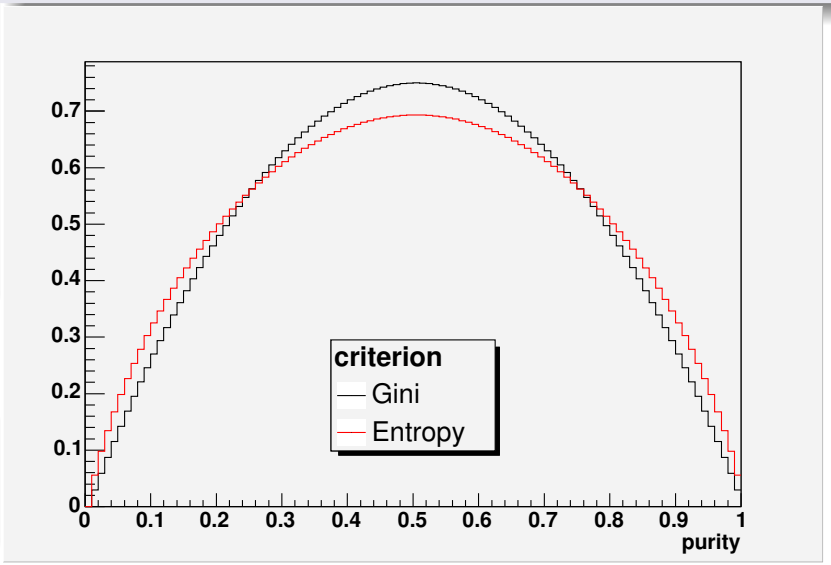
$$\Delta i(s^*,t) = \max_{s \in \{\text{splits}\}} \Delta i(s,t)$$

- Maximizing $\Delta i(s,t) \equiv$ minimizing overall tree impurity

Examples

$Gini = 1 - \sum_{i=s,b} p_i^2 = \frac{2sb}{(s+b)^2}$

$entropy = -\sum_{i=s,b} p_i \log p_i$



Decision Trees - 49 input 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$_{\mathrm{besttop}}$,besttop$_{\mathrm{CMframe}}$)
- cos(lepton$_{\mathrm{btaggedtop}}$,btaggedtop$_{\mathrm{CMframe}}$)
- cos(notbest,alljets)$_{\mathrm{alljets}}$
- cos(notbest,lepton)$_{\mathrm{besttop}}$
- cos(untag1,alljets)$_{\mathrm{alljets}}$
- cos(untag1,lepton)$_{\mathrm{btaggedtop}}$

Event Kinematics

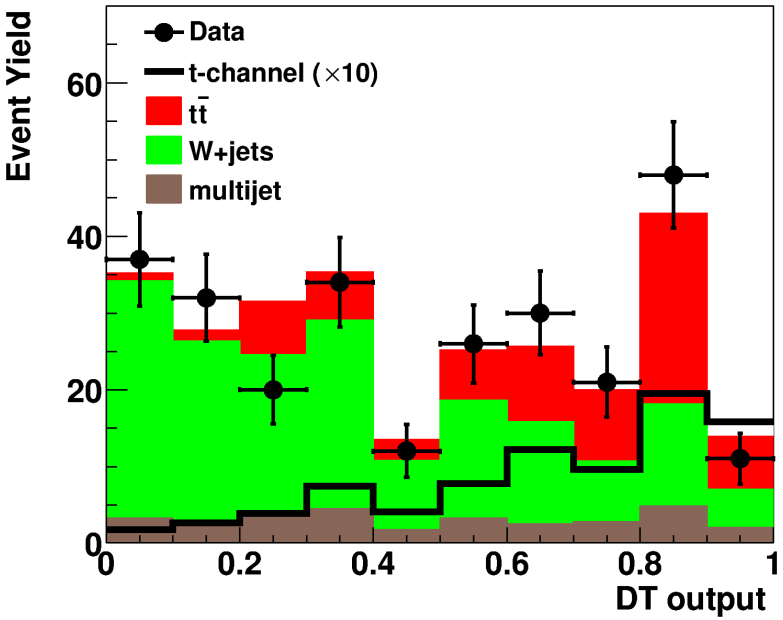
- Aplanarity(alljets, W)
- M(W,best1) ("best" top mass)
- M(W,tag1) ("b-tagged" top mass)
- H_T(alljets)
- H_T(alljets$-$best1)
- H_T(alljets$-$tag1)
- H_T(alljets, W)
- H_T(jet1,jet2)
- H_T(jet1,jet2, W)
- M(alljets)
- M(alljets$-$best1)
- M(alljets$-$tag1)
- M(jet1,jet2)
- M(jet1,jet2, W)
- M_T(jet1,jet2)
- 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, lost some sensitivity
- Same list used for all channels

Decision tree output

Measure and apply

- Take trained tree and run on independent pseudo-data sample, determine purities
- Apply to data
- Should see enhanced separation (signal right, background left)
- Could cut on output and measure, or use whole distribution to measure



Limitations

- Instability of tree structure
- Piecewise nature of output

Advantages

- DT has human readable structure (no black box)
- Training is fast
- Deals with discrete variables
- No need to transform inputs
- Resistant to irrelevant variables

Boosting a decision tree

Boosting

- Recent technique to improve performance of a weak classifier
- Recently used on decision trees by GLAST and MiniBooNE
- Basic principle on DT:
  - train a tree T_k
  - $T_{k+1} = \text{modify}(T_k)$

Ref: Freund and Schapire, "Experiments with a new boosting algorithm", in *Machine Learning: Proceedings of the Thirteenth International Conference*, pp 148-156 (1996)

AdaBoost algorithm

- Adaptive boosting
- Check which events are misclassified by T_k
- Derive tree weight α_k:
  $err_k = \frac{\sum_{i=1}^{N} w_i \times \text{isMisclassified}_k(i)}{\sum_{i=1}^{N} w_i}$
  $\alpha_k = \beta \times \ln((1 - err_k)/err_k)$
- Increase weight of misclassified events by e^{α_k}
- Train again to build T_{k+1}
- Boosted result of event i:
  $T(i) = \sum_{k=1}^{N_{\text{tree}}} \alpha_k T_k(i)$

- Averaging $\Rightarrow$ dilutes piecewise nature of DT
- Usually improves performance

Analysis validation

Ensemble testing

- Test the whole machinery with many sets of pseudo-data
- Like running DØ experiment 1000s of times
- Generated ensembles with different signal contents (no signal, SM, other cross sections, higher luminosity)

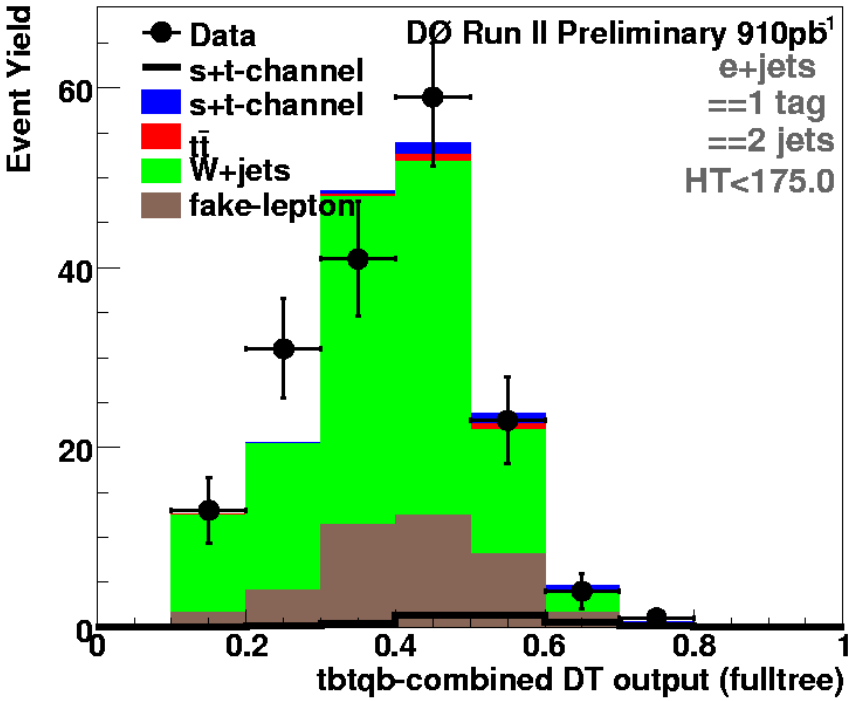
Ensemble generation

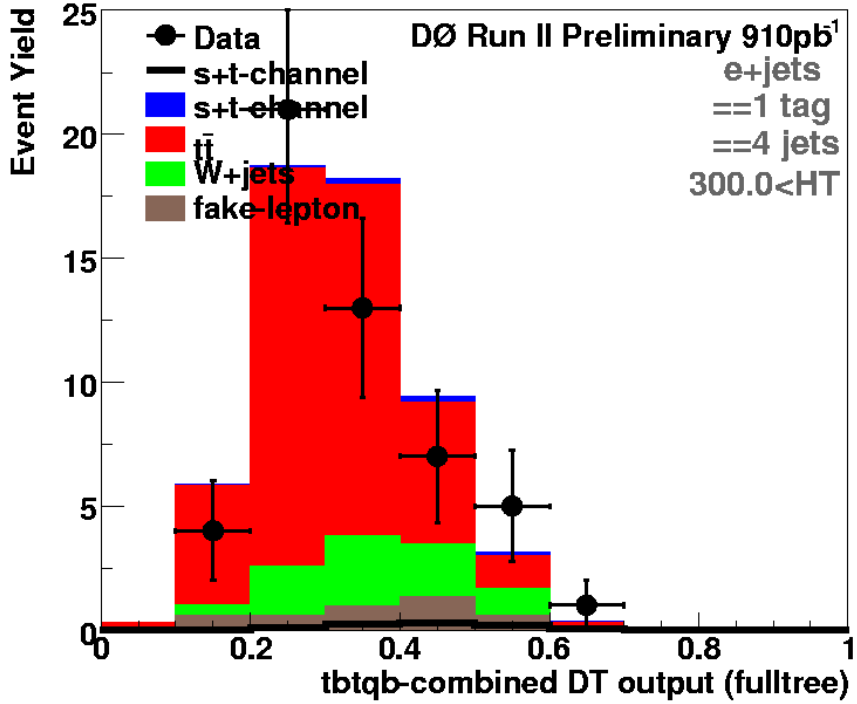
- Pool of weighted signal + background events
- Fluctuate relative and total yields in proportion to systematic errors, reproducing correlations
- Randomly sample from a Poisson distribution about the total yield to simulate statistical fluctuations
- Generate pseudo-data set, pass through full analysis chain (including systematic uncertainties)

All analyses achieved linear response to varying input cross sections and negligible bias

Cross-check samples

- Validate methods on data in no-signal region
- **"W+jets": =2jets, H_T(lepton,$\not\!\!E_T$,alljets) $<$ 175 GeV**
- **"ttbar": =4jets, H_T(lepton,$\not\!\!E_T$,alljets) $>$ 300 GeV**

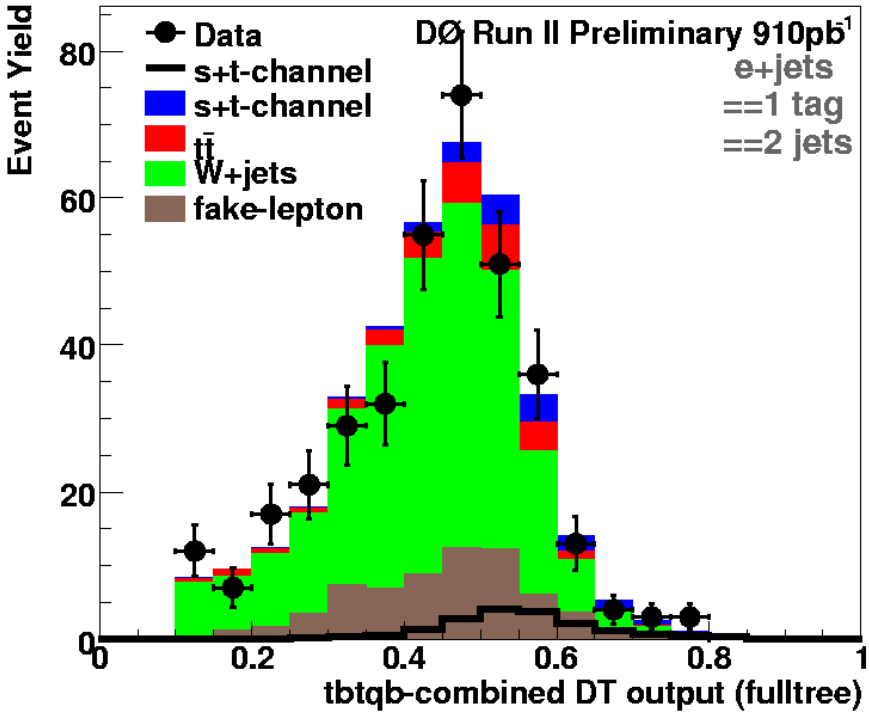




- Good agreement of model with data

Decision trees on data

- 36 different decision trees
  - 3 signals (s,t,$s+t$)
  - 2 leptons (e,μ)
  - 3 jet multiplicities (2,3,4 jets)
  - 2 b-tag multiplicities (1,2 tags)
- For each signal train against the sum of backgrounds



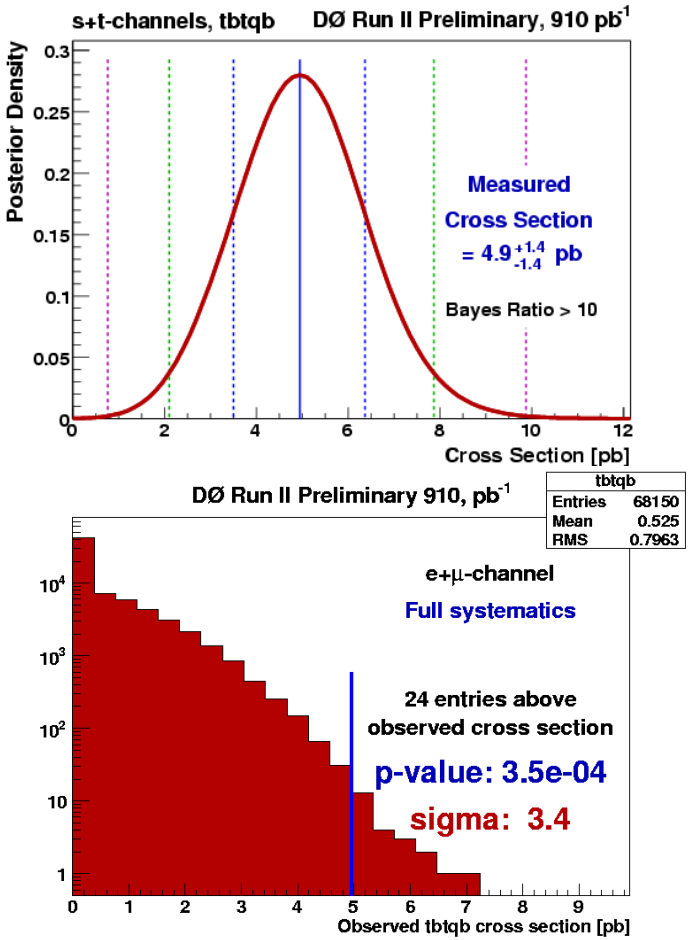
Results

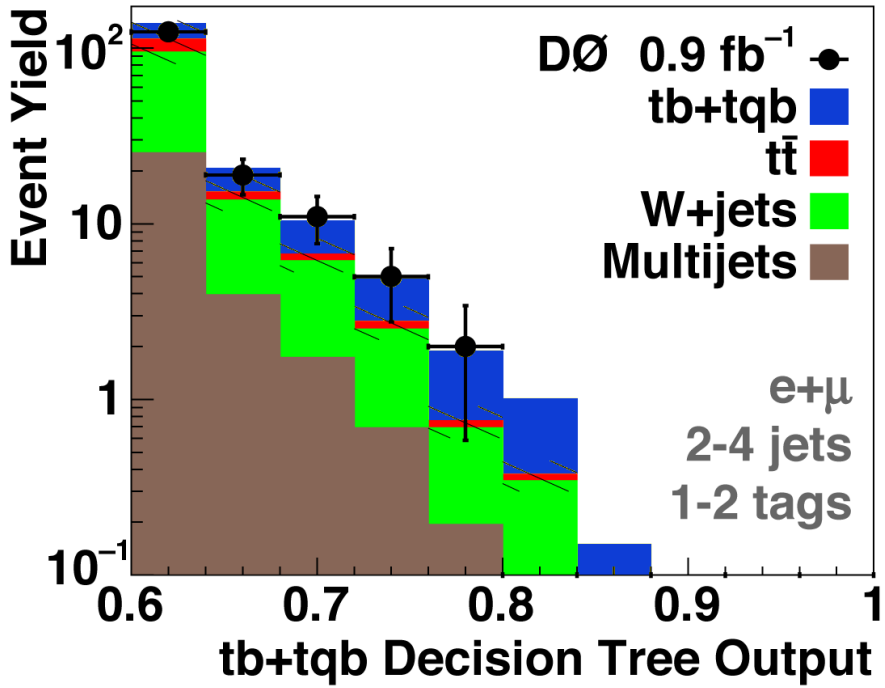
$\sigma_{s+t} = 4.9 \pm 1.4$ pb

p-value $= 0.035\%$ (3.4σ)

SM compatibility: 11% (1.3σ)

Evidence for single top production!

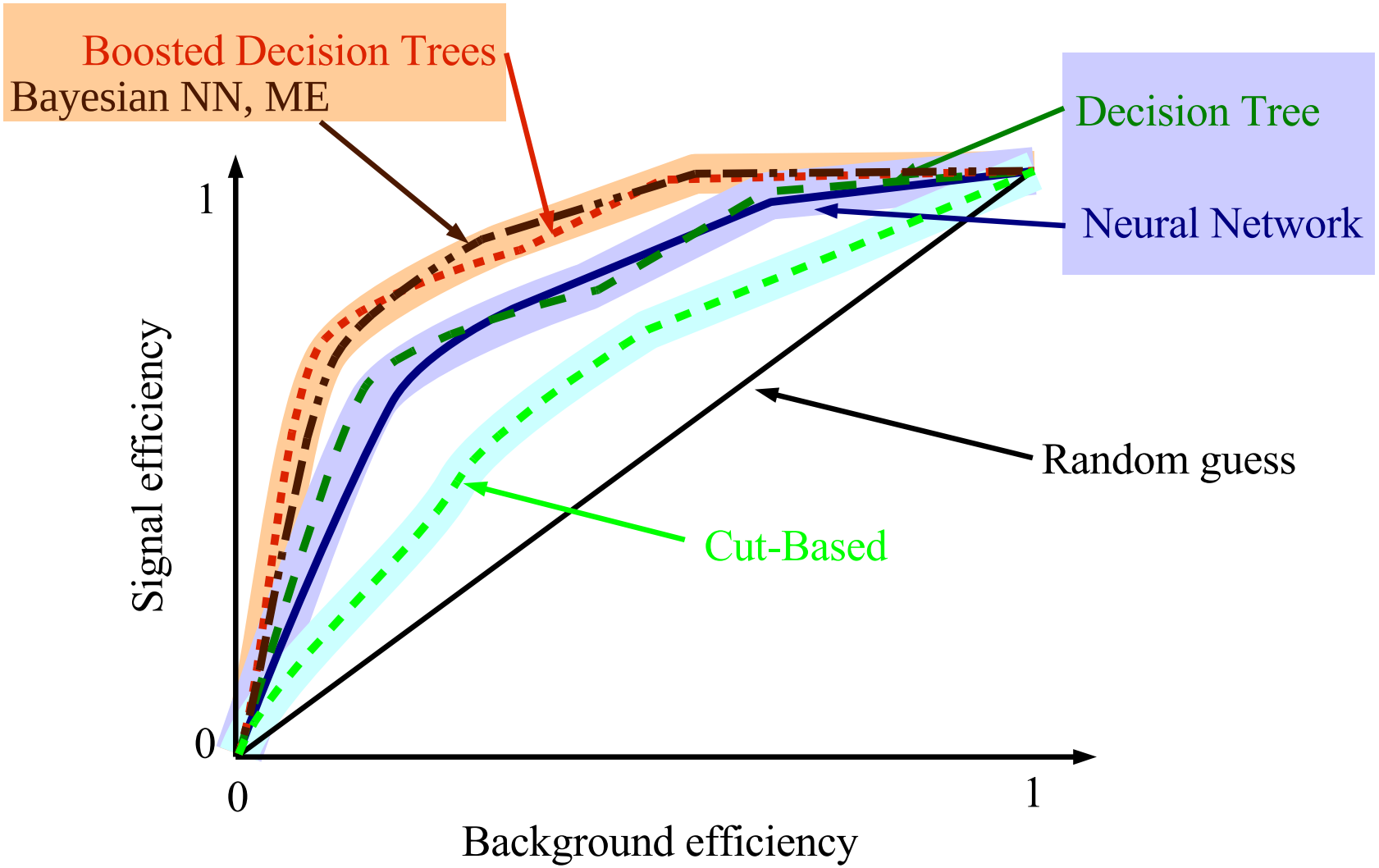




$\sigma_s = 1.0 \pm 0.9$ pb

$\sigma_t = 4.2^{+1.8}_{-1.4}$ pb

Comparison for DØ single top analysis



- Cannot know *a priori* which method will work best
- $\Rightarrow$ Need to experiment with different techniques

Conclusion

First evidence for single top quark production (DØ decision trees)

$$\sigma(p\bar{p} \rightarrow tb + X, tqb + X) = 4.9 \pm 1.4 \text{ pb}$$

3.4σ significance

First direct measurement of $|V_{tb}|$ (DØ decision trees)

$$|V_{tb} f_1^L| = 1.3 \pm 0.2$$

assuming $f_1^L = 1$: $0.68 < |V_{tb}| \leq 1$ @ 95% CL

(Always assuming $V_{td}^2 + V_{ts}^2 \ll V_{tb}^2$ and pure $V-A$ and CP-conserving Wtb interaction)

Published in Phys. Rev. Lett. 98, 181802 (2007) (hep-ex/0612052)

New preliminary combination of DT, ME and BNN

$$\sigma(p\bar{p} \rightarrow tb + X, tqb + X) = 4.7 \pm 1.3 \text{ pb}$$

3.6σ significance

- A lot more data already at hand

Single top prospects — Tevatron and LHC

Tevatron

- By 2009 we should have observed single top production and measured its cross section to 15-20%
- $|V_{tb}|$ is then known to $\sim$10%

LHC

- Much larger production rates:
  $$\sigma_s^{t/\bar{t}} = 6.6/4.1 \text{ pb } (\pm 10\%)$$
  $$\sigma_t^{t/\bar{t}} = 156/91 \text{ pb } (\pm 5\%)$$
  $$\sigma_{tW}^{t/\bar{t}} = 34/34 \text{ pb } (\pm 10\%)$$
- Try to observe all three channels (s-channel challenging)
- $|V_{tb}|$ measured to percent level
- Large samples $\Rightarrow$ study properties