Boosted decision trees in practice

Yann Coadou

CPPM Marseille

School of Statistics SOS2012, Autrans

1 June 2012

Outline

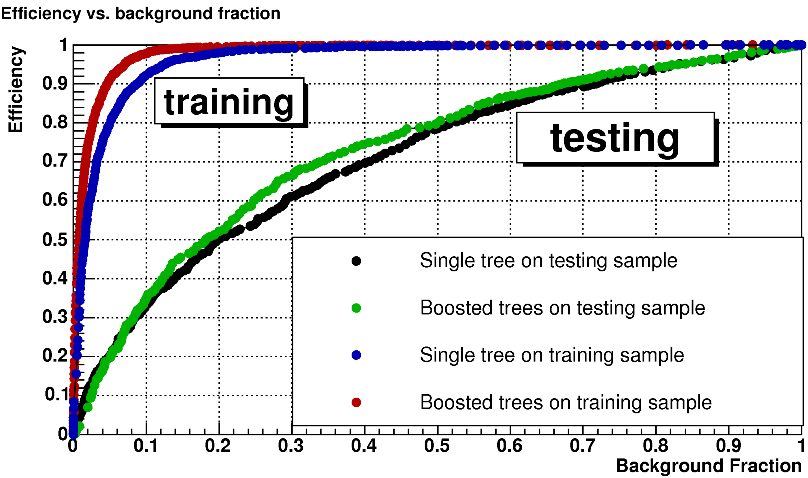
1. **BDT performance**
   - Overtraining?
   - Clues to boosting performance
2. **Concrete examples**
   - First is best
   - XOR problem
   - Circular correlation
   - Many small trees or fewer large trees?
3. **BDTs in real life (. . . or at least in real physics cases. . .)**
   - Single top search at D0
   - More applications in HEP
4. **Software and references**

Before we go on...

!!! VERY IMPORTANT !!!

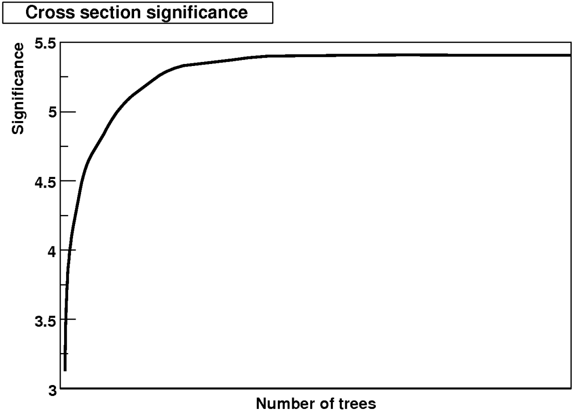
Understand your inputs well
before you start playing with multivariate techniques

Training and generalisation error



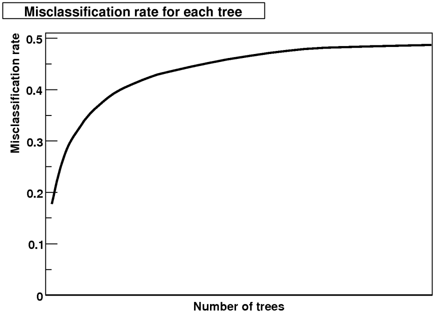
- Clear overtraining, but still better performance after boosting

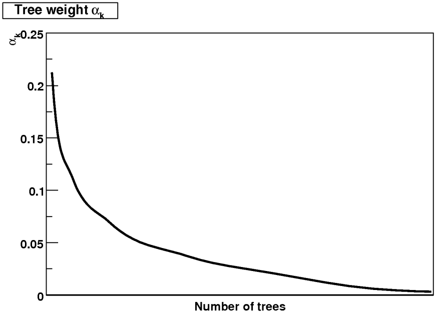
Cross section significance



- More relevant than testing error
- Reaches plateau
- Afterwards, boosting does not hurt (just wasted CPU)
- Applicable to any other figure of merit of interest for your use case

Clues to boosting performance



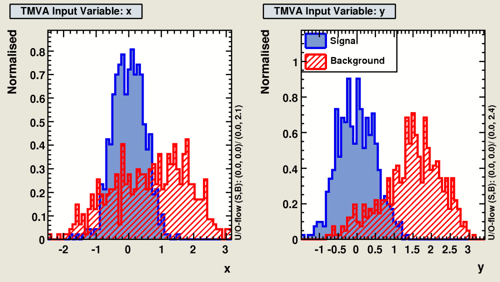


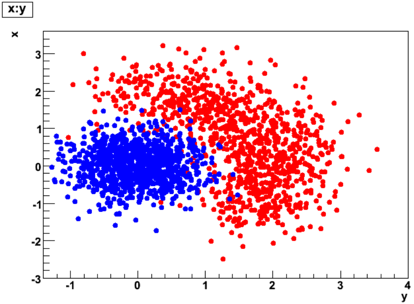
- First tree is best, others are minor corrections
- Specialised trees do not perform well on most events ⇒ decreasing tree weight and increasing misclassification rate
- Last tree is not better evolution of first tree, but rather a pretty bad DT that only does a good job on few cases that the other trees couldn't get right

1. BDT performance
   - Overtraining?
   - Clues to boosting performance
2. Concrete examples
   - First is best
   - XOR problem
   - Circular correlation
   - Many small trees or fewer large trees?
3. BDTs in real life (. . . or at least in real physics cases. . .)
   - Single top search at D0
   - More applications in HEP
4. Software and references

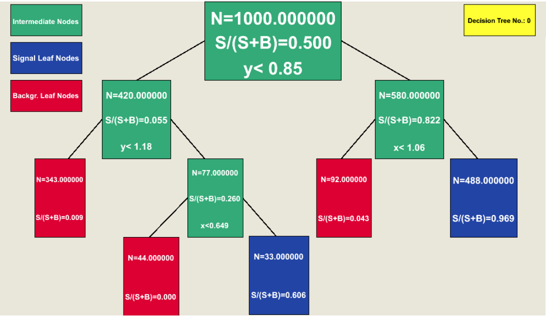
Concrete example

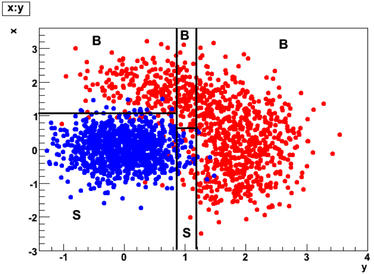
- Using TMVA and some code modified from G. Cowan's CERN academic lectures (June 2008)



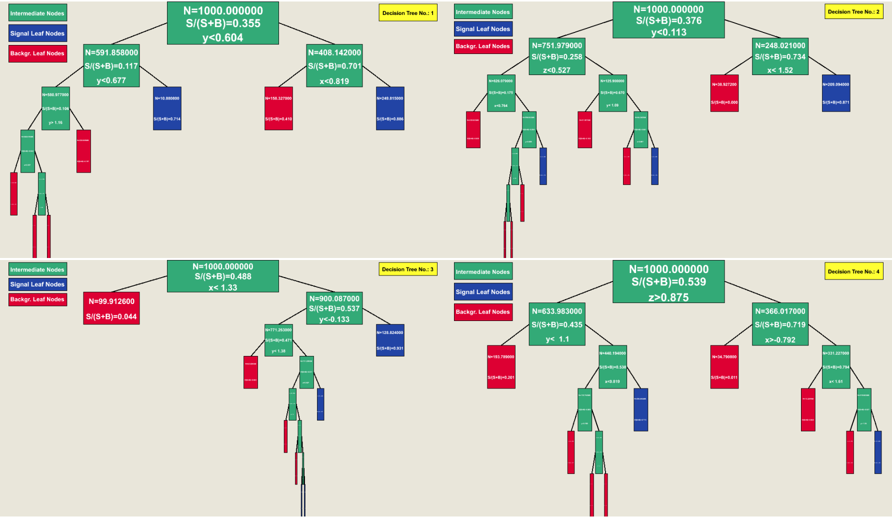


Concrete example



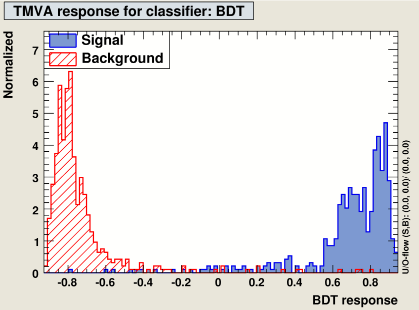


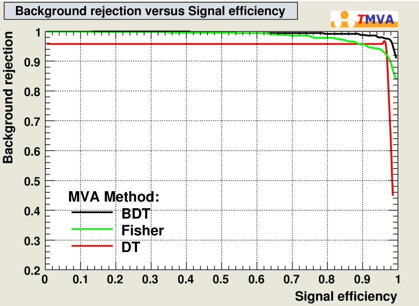
Concrete example



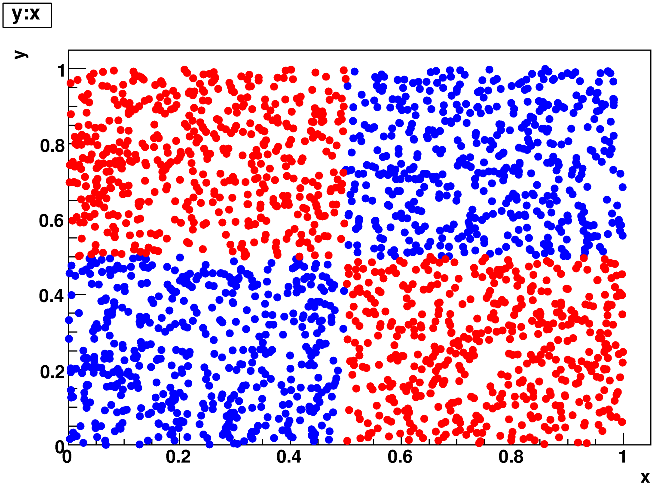
- Specialised trees

Concrete example

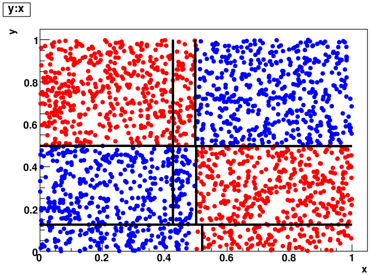


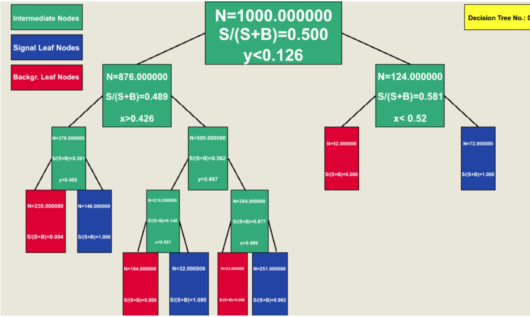


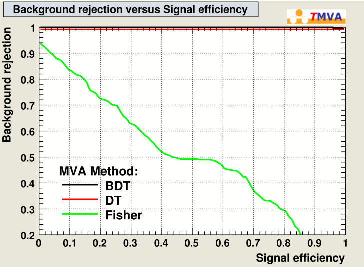
Concrete example: XOR



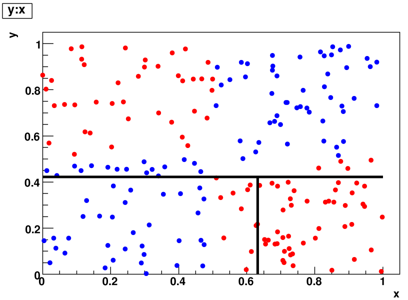
Concrete example: XOR





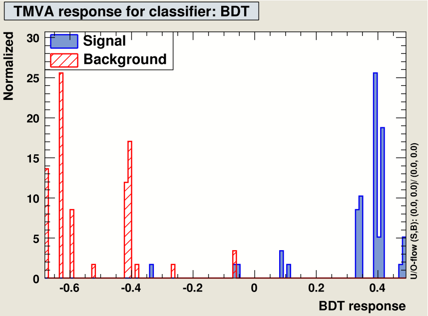


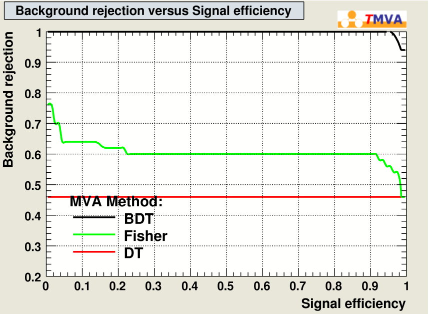
Concrete example: XOR with 100 events



Small statistics

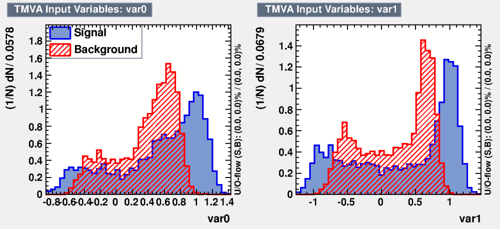
- Single tree or Fischer discriminant not so good
- BDT very good: high performance discriminant from combination of weak classifiers

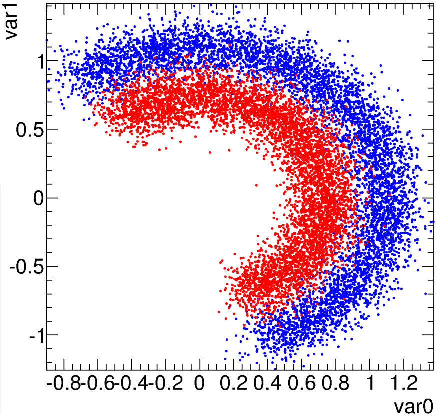




Circular correlation

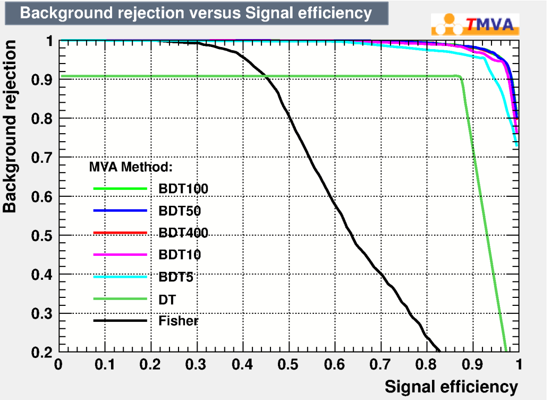
- Using TMVA and `create_circ` macro from `$ROOTSYS/tmva/test/createData.C` to generate dataset





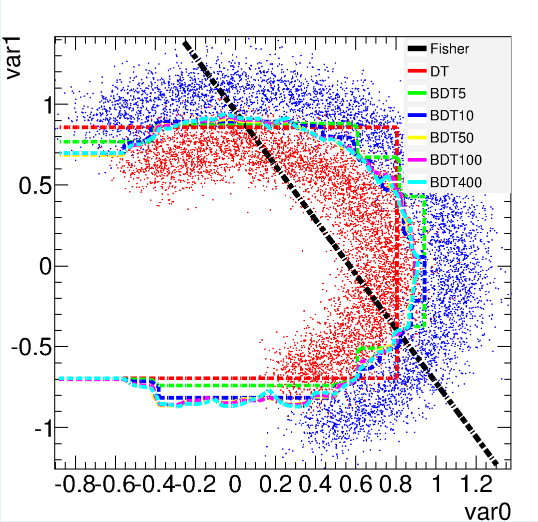
Boosting longer

- Compare performance of Fisher discriminant, single DT and BDT with more and more trees (5 to 400)
- All other parameters at TMVA default (would be 400 trees)



- Fisher bad (expected)
- Single (small) DT: not so good
- More trees ⇒ improve performance until saturation

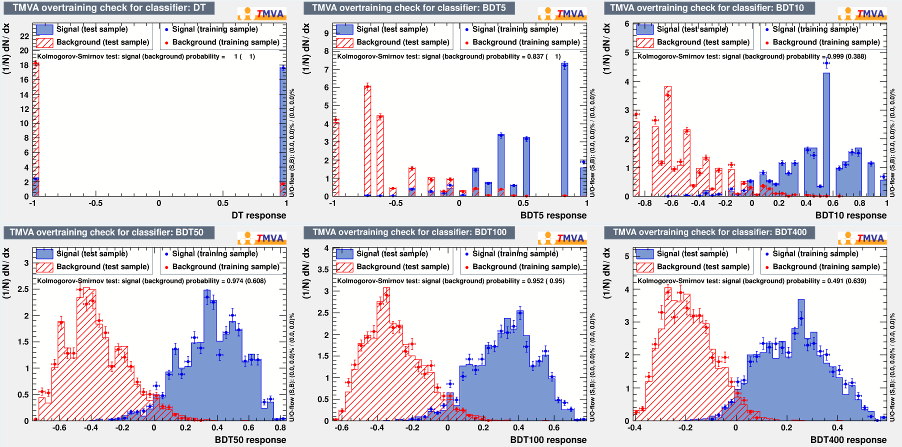
Decision contours



- Fisher bad (expected)
- Note: max tree depth = 3
- Single (small) DT: not so good. Note: a larger tree would solve this problem
- More trees ⇒ improve performance (less step-like, closer to optimal separation) until saturation
- Largest BDTs: wiggle a little around the contour ⇒ picked up features of training sample, that is, overtraining

Circular correlation

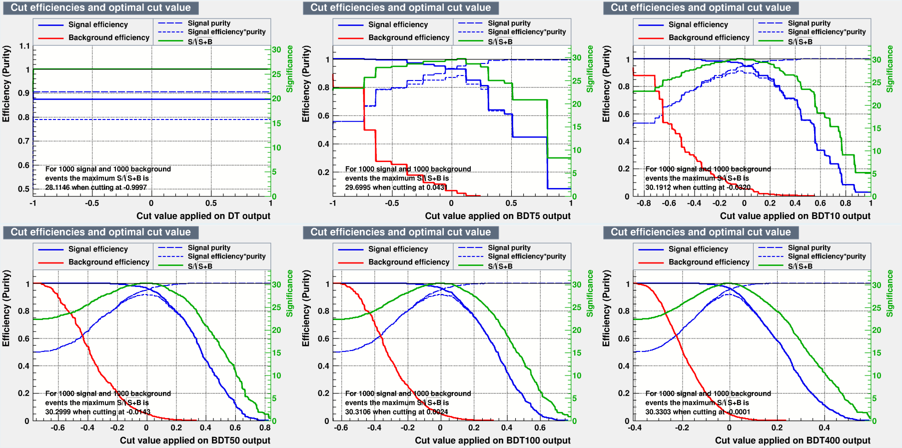
Training/testing output



- Better shape with more trees: quasi-continuous
- Overtraining because of disagreement between training and testing? Let's see

Circular correlation

Performance in optimal significance

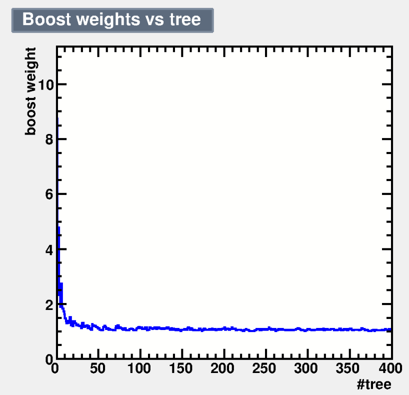


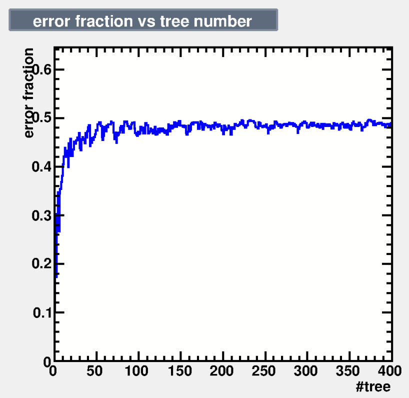
- Best significance actually obtained with last BDT, 400 trees!
- But to be fair, equivalent performance with 10 trees already
- Less "stepped" output desirable? ⇒ maybe 50 is reasonable

Circular correlation

Control plots

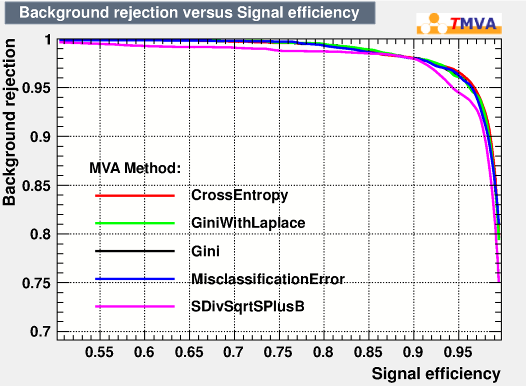
- Boosting weight decreases fast and stabilises
- First trees have small error fractions, then increases towards 0.5 (random guess)
- $\Rightarrow$ confirms that best trees are first ones, others are small corrections





Separation criterion for node splitting

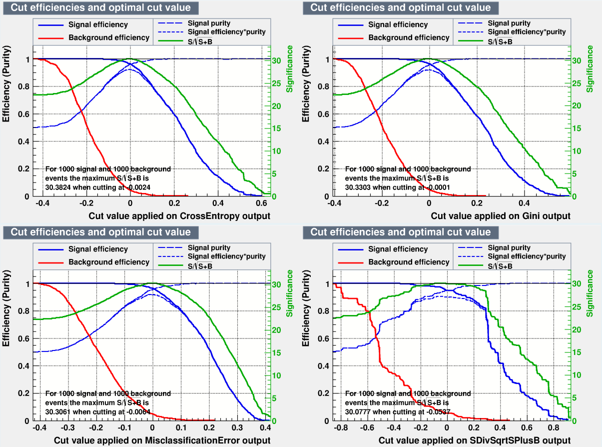
- Compare performance of Gini, entropy, misclassification error, $\frac{s}{\sqrt{s+b}}$
- All other parameters at TMVA default



- Very similar performance (even zooming on corner)
- Small degradation (in this particular case) for $\frac{s}{\sqrt{s+b}}$: only criterion that doesn't respect good properties of impurity measure (see yesterday: maximal for equal mix of signal and bkg, symmetric in p_{sig} and p_{bkg}, minimal for node with either signal only or bkg only, strictly concave)

Circular correlation

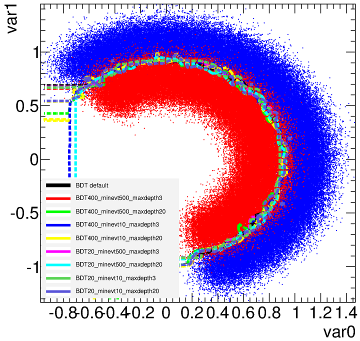
Performance in optimal significance

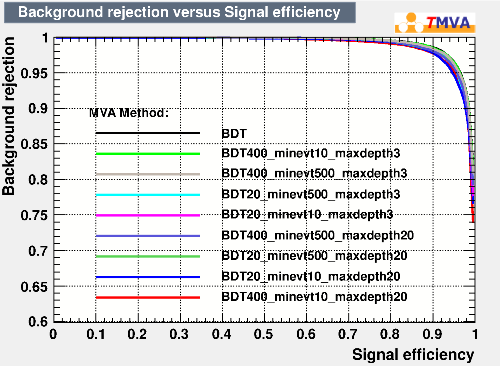


- Confirms previous page: very similar performance, worse for BDT optimised with significance!

Many small trees or fewer large trees?

- Using same `create_circ` macro but generating larger dataset to avoid stats limitations
- 20 or 400 trees; minimum leaf size: 10 or 500 events
- Maximum depth (max number of cuts to reach leaf): 3 or 20





- Overall: very comparable performance. Depends on use case.

1. BDT performance
   - Overtraining?
   - Clues to boosting performance
2. Concrete examples
   - First is best
   - XOR problem
   - Circular correlation
   - Many small trees or fewer large trees?
3. **BDTs in real life (...or at least in real physics cases...)**
   - Single top search at D0
   - More applications in HEP
4. Software and references

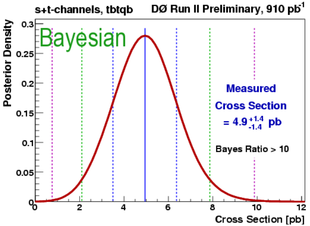
Single top production evidence at D0 (2006)

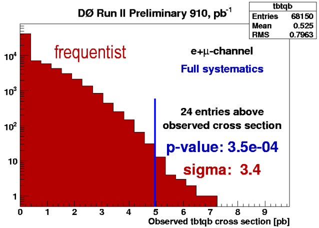
- Three multivariate techniques: BDT, Matrix Elements, BNN
- Most sensitive: BDT

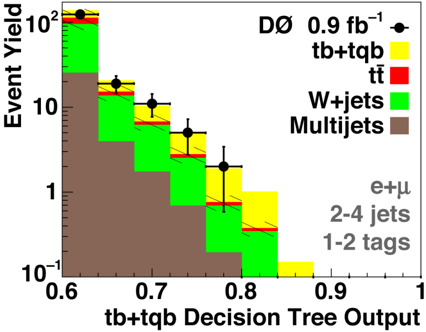
$\sigma_{s+t} = 4.9 \pm 1.4$ pb

p-value $= 0.035\%$ (3.4σ)

SM compatibility: 11% (1.3σ)







$$\sigma_s = 1.0 \pm 0.9 \text{ pb}$$
$$\sigma_t = 4.2^{+1.8}_{-1.4} \text{ pb}$$

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 did not degrade performance
- Tested shorter lists, lost some sensitivity
- Same list used for all channels

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 did not degrade performance
- Tested shorter lists, lost some sensitivity
- Same list used for all channels
- Best theoretical variable: H_T(alljets, W). But detector not perfect ⇒ capture the essence from several variations usually helps "dumb" MVA

Decision tree parameters

BDT choices

- 1/3 of MC for training
- AdaBoost parameter $\beta = 0.2$
- 20 boosting cycles
- Signal leaf if purity > 0.5
- Minimum leaf size $=$ 100 events
- Same total weight to signal and background to start
- Goodness of split - Gini factor

Analysis strategy

- Train 36 separate 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

Analysis validation

Ensemble testing

- Test the whole machinery with many sets of pseudo-data
- Like running D0 experiment 1000s of times
- Generated ensembles with different signal contents (no signal, SM, other cross sections, higher luminosity)

Ensemble generation

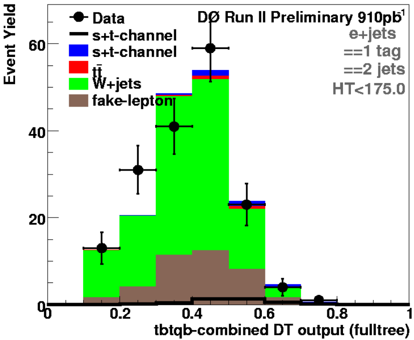
- Pool of weighted signal + background events
- Fluctuate relative and total yields in proportion to systematic errors, reproducing correlations
- Randomly sample from a Poisson distribution about the total yield to simulate statistical fluctuations
- Generate pseudo-data set, pass through full analysis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

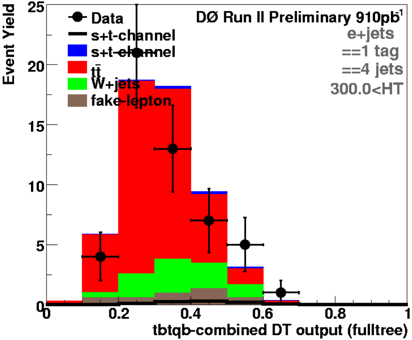
Achieved linear response to varying input cross sections and negligible bias

- Validate method on data in no-signal region

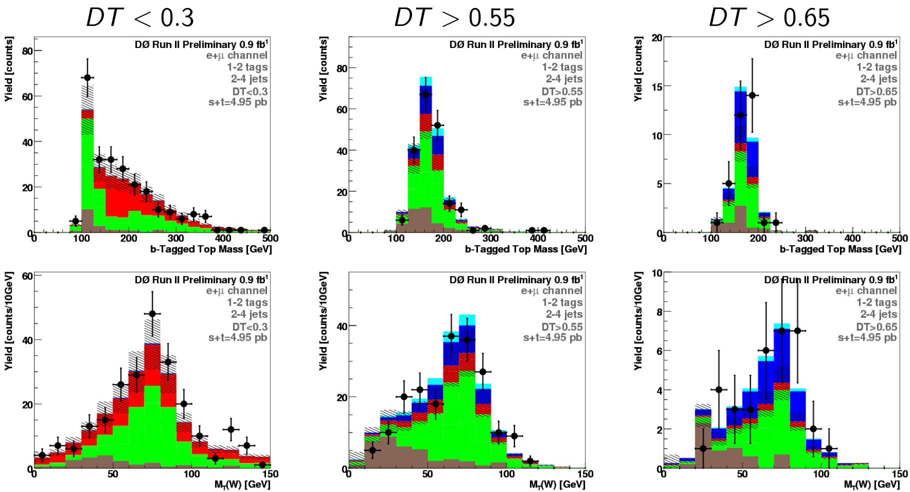
- **"W+jets"**: $= 2$ jets, $H_T(\text{lepton}, \not{E}_T, \text{alljets}) < 175$ GeV

- **"ttbar"**: $= 4$ jets, $H_T(\text{lepton}, \not{E}_T, \text{alljets}) > 300$ GeV





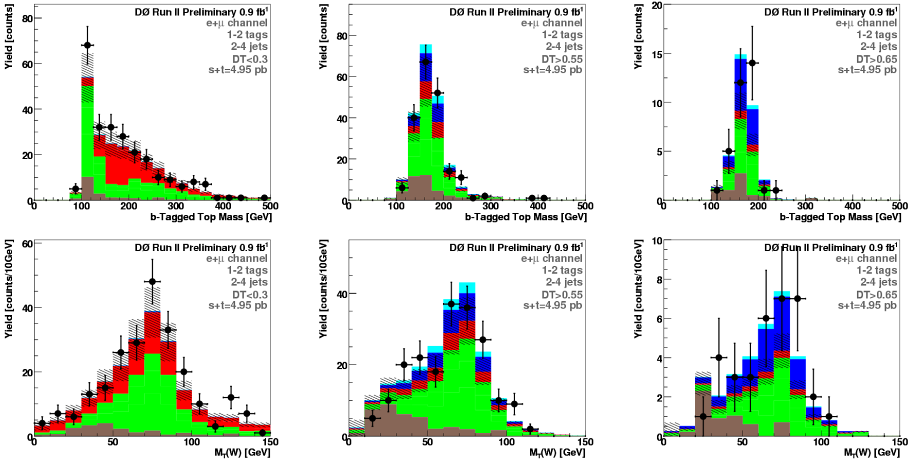
- Good agreement



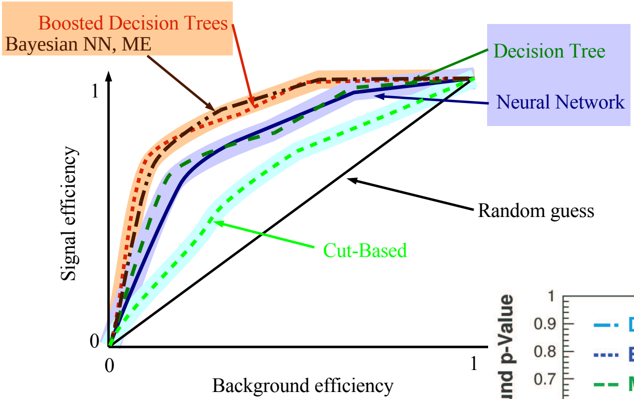
- High BDT region = shows masses of real t and W $\Rightarrow$ expected
- Low BDT region = background-like $\Rightarrow$ expected

Boosted decision tree event characteristics

$DT < 0.3$ $DT > 0.55$ $DT > 0.65$

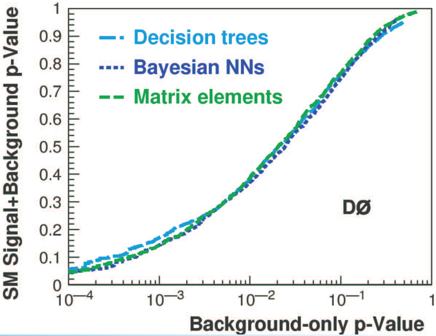


- High BDT region = shows masses of real t and W $\Rightarrow$ expected
- Low BDT region = background-like $\Rightarrow$ expected
- Above doesn't tell analysis is ok, but not seeing this could be a sign of a problem



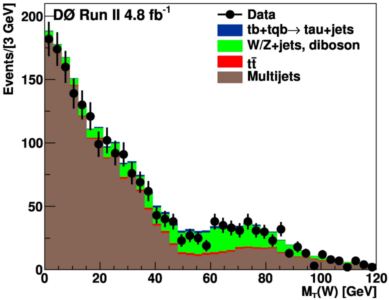
- Cannot know *a priori* which method will work best
- ⇒ Need to experiment with different techniques

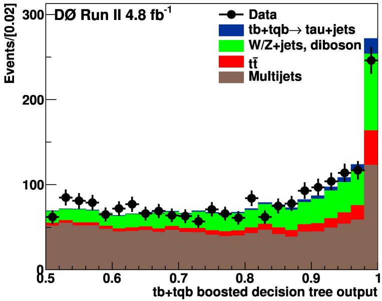
Power curve

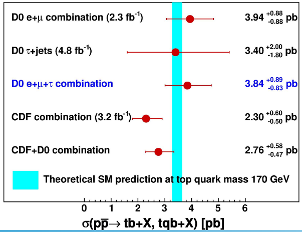


Search for single top in tau+jets at D0 (2010)

- Tau ID BDT and single top search BDT



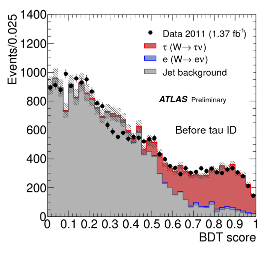


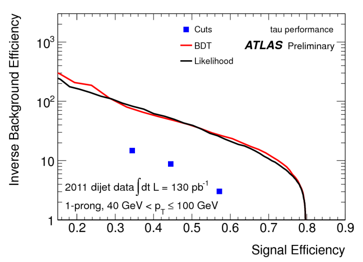


- 4% sensitivity gain over $e + \mu$ analysis
- PLB690:5,2010

ATLAS tau identification

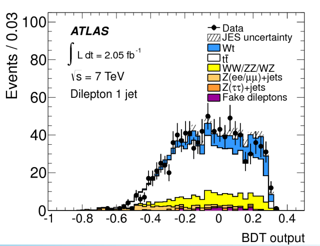
- Now used both offline and online
- Systematics: propagate various detector/theory effects to BDT output and measure variation





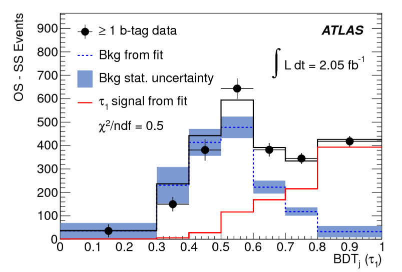
Hot from the press: ATLAS Wt production evidence (since Monday)

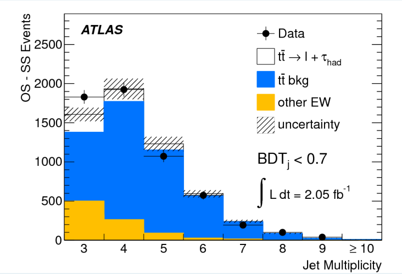
- arXiv:1205.5764
- BDT output used in final fit to measure cross section
- Constraints on systematics from profiling

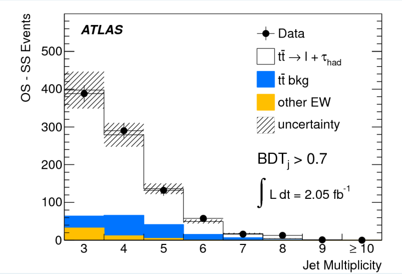


ATLAS $t\bar{t} \to e/\mu + \tau$+jets production cross section

- BDT for tau ID: one to reject electrons, one against jets
- Fit BDT output to get tau contribution in data







Boosted decision trees in HEP studies

- MiniBooNE (e.g. physics/0408124 NIM A543:577-584, physics/0508045 NIM A555:370-385, hep-ex/0704.1500)
- D0 single top evidence (PRL98:181802,2007, PRD78:012005,2008)
- D0 and CDF single top quark observation (PRL103:092001,2009, PRL103:092002,2009)
- D0 tau ID and single top search (PLB690:5,2010)
- Fermi gamma ray space telescope (same code as D0)
- BaBar (hep-ex/0607112)
- ATLAS/CMS: Many other analyses
- *b*-tagging for LHC (physics/0702041)
- LHCb: $B^0_{(s)} \to \mu\mu$ search, selection of $B^0_s \to J/\psi\phi$ for ϕ_s measurement
- More and more underway

Boosted decision tree software

- Historical: CART, ID3, C4.5
- D0 analysis: C++ custom-made code. Can use entropy/Gini, boosting/bagging/random forests
- MiniBoone code at http://www-mhp.physics.lsa.umich.edu/~roe/

Much better approach

- Go for a fully integrated solution
  - use different multivariate techniques easily
  - spend your time on understanding your data and model
- Examples:
  - Weka. Written in Java, open source, very good published manual. Not written for HEP but very complete
    http://www.cs.waikato.ac.nz/ml/weka/
  - StatPatternRecognition
    http://www.hep.caltech.edu/~narsky/spr.html
  - **TMVA (Toolkit for MultiVariate Analysis). Now integrated in ROOT, complete manual** **http://tmva.sourceforge.net**

References I

- L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and Regression Trees*, Wadsworth, Stamford, 1984
- J.R. Quinlan, "Induction of decision trees", *Machine Learning*, 1(1):81–106, 1986
- J.R. Quinlan, "Simplifying decision trees", *International Journal of Man-Machine Studies*, 27(3):221–234, 1987
- R.E. Schapire, "The strength of weak learnability", *Machine Learning*, 5(2):197–227,1990
- Y. Freund, "Boosting a weak learning algorithm by majority", *Information and computation*. 121(2):256–285, 1995
- Y. Freund and R.E. Schapire, "Experiments with a New Boosting Algorithm" in *Machine Learning: Proceedings of the Thirteenth International Conference*, edited by L. Saitta (Morgan Kaufmann, San Fransisco, 1996) p. 148
- Y. Freund and R.E. Schapire, "A short introduction to boosting" *Journal of Japanese Society for Artificial Intelligence, 14(5):771-780 (1999)*

References II

- Y. Freund and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, 55(1):119–139, 1997
- J.H. Friedman, T. Hastie and R. Tibshirani, "Additive logistic regression: a statistical view of boosting", *The Annals of Statistics*, 28(2), 377–386, 2000
- L. Breiman, "Bagging Predictors", *Machine Learning*, 24 (2), 123–140, 1996
- L. Breiman, "Random forests", *Machine Learning*, 45 (1), 5–32, 2001
- B.P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor, Nucl. Instrum. Methods Phys. Res., Sect.A 543, 577 (2005); H.-J. Yang, B.P. Roe, and J. Zhu, Nucl. Instrum.Methods Phys. Res., Sect. A 555, 370 (2005)
- V. M. Abazov *et al.* [D0 Collaboration], "Evidence for production of single top quarks,", Phys. Rev. D**78**, 012005 (2008)