

Boosted decision trees in practice

Yann Coadou

CPPM Marseille

esipap...

European School of Instrumentation
in Particle & Astroparticle Physics

Archamps, 30 January 2014



- 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**

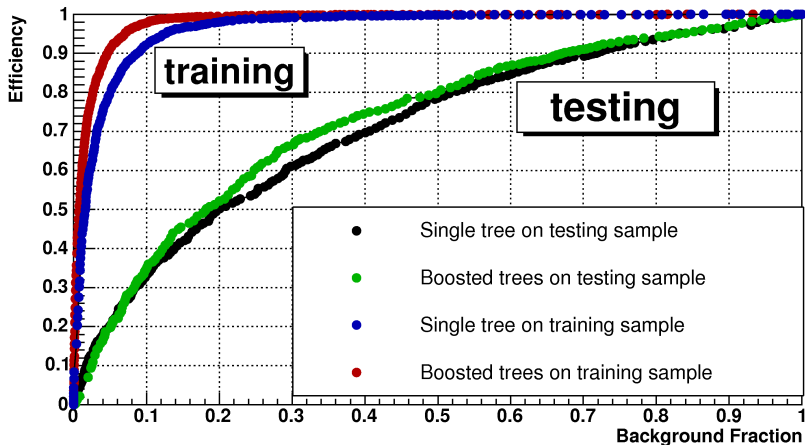


!!! VERY IMPORTANT !!!

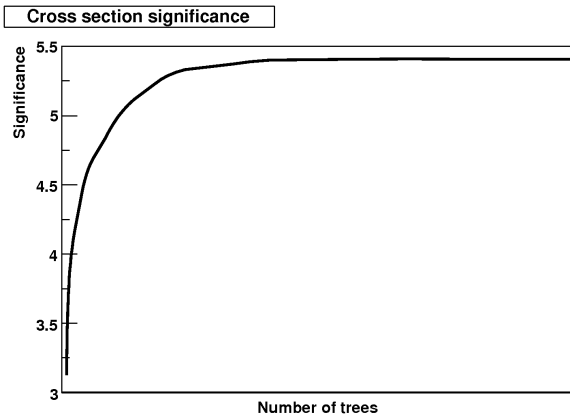
**Understand your inputs well
before you start playing with multivariate techniques**



Efficiency vs. background fraction

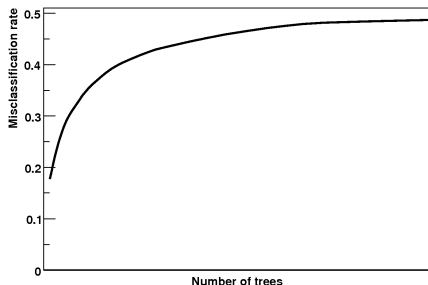


- Clear overtraining, but still better performance after boosting

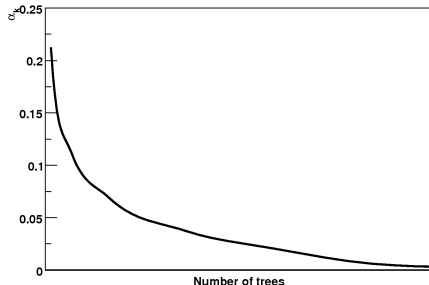


- More relevant than testing error
- Reaches plateau
- Afterwards, boosting does not hurt (just wasted CPU)
- Applicable to any other figure of merit of interest for your use case

Misclassification rate for each tree



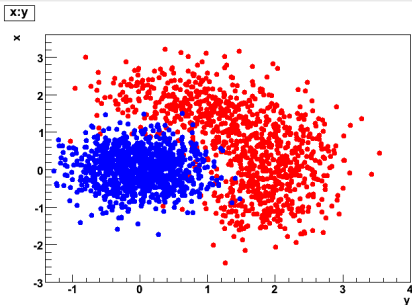
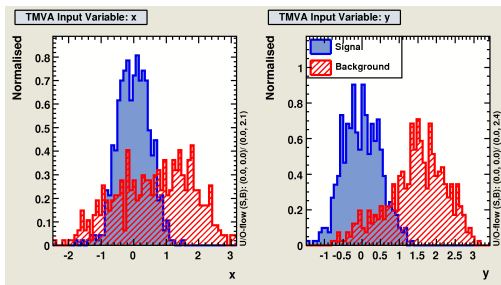
Tree weight α_k

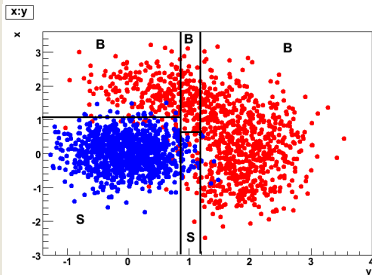
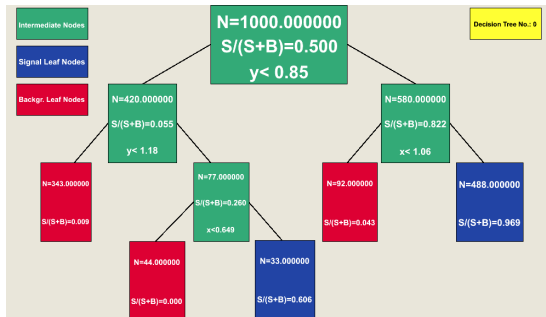


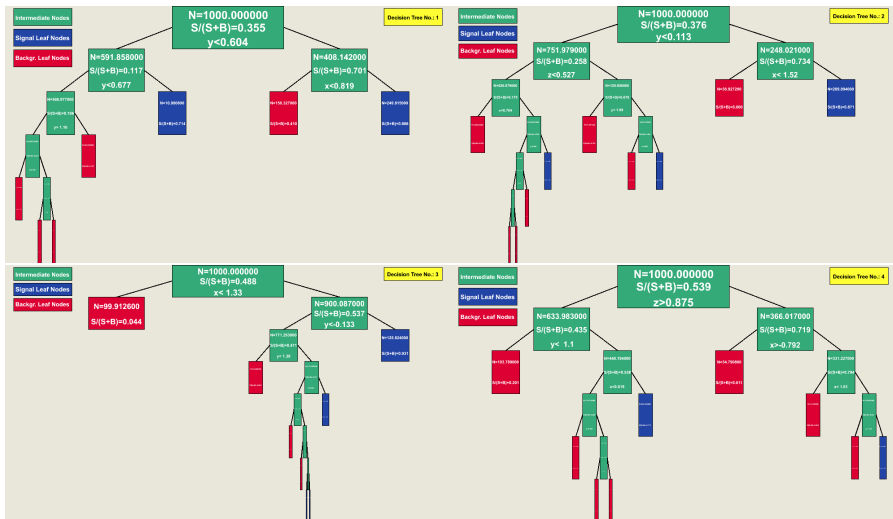
- First tree is best, others are minor corrections
- Specialised trees do not perform well on most events \Rightarrow decreasing tree weight and increasing misclassification rate
- Last tree is not better evolution of first tree, but rather a pretty bad DT that only does a good job on few cases that 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**

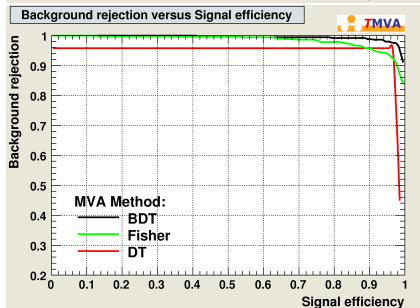
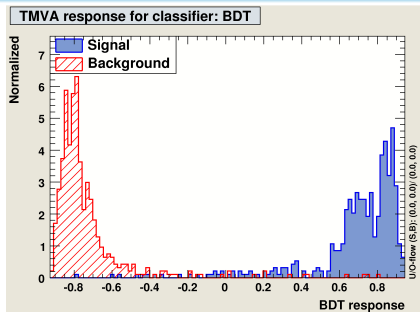
- Using TMVA and some code modified from G. Cowan's CERN academic lectures (June 2008)





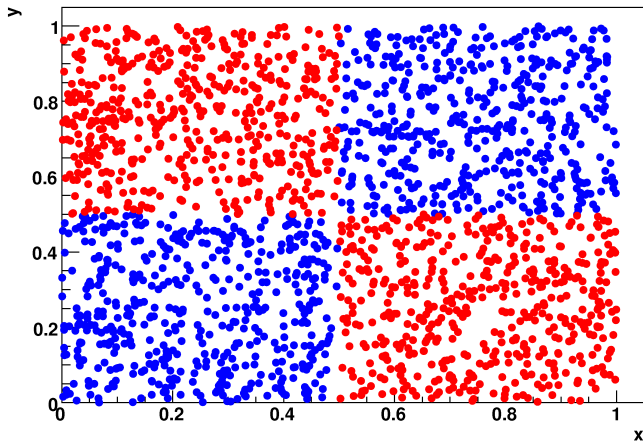


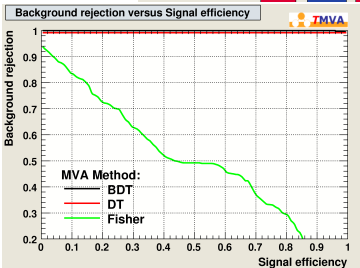
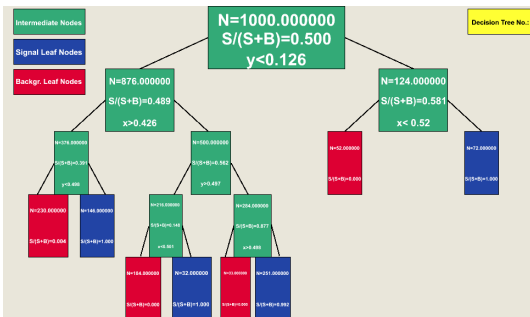
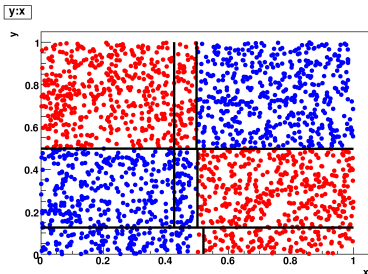
- Specialised trees



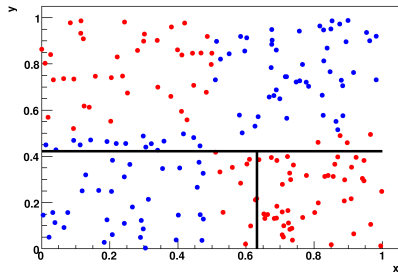
Concrete example: XOR

y:x





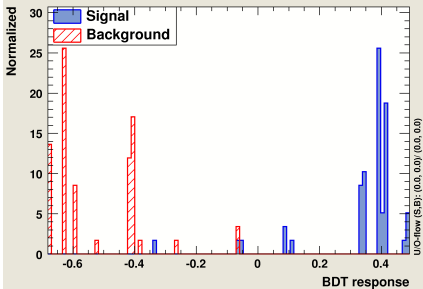
y:x



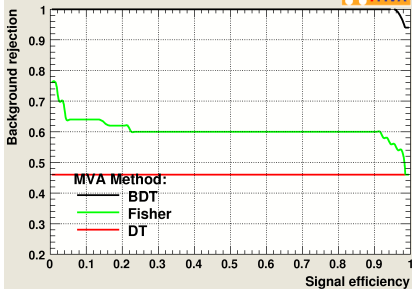
Small statistics

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

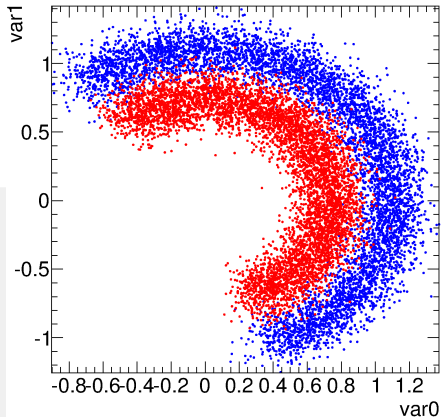
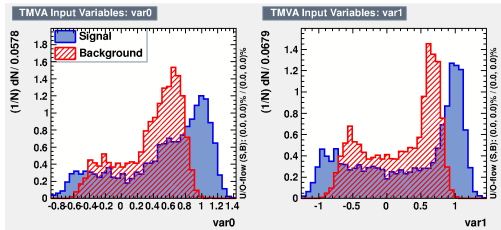
TMVA response for classifier: BDT



Background rejection versus Signal efficiency



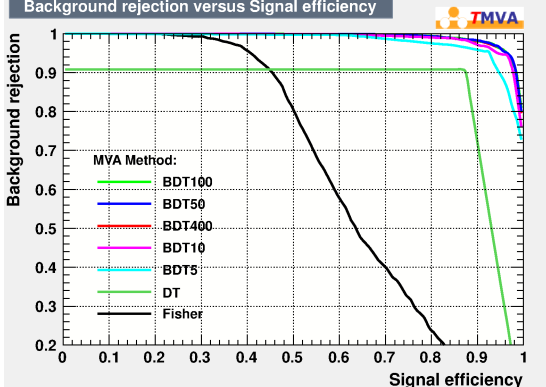
- Using TMVA and create_circ macro from `$ROOTSYS/tmva/test/createData.C` to generate dataset



Boosting longer

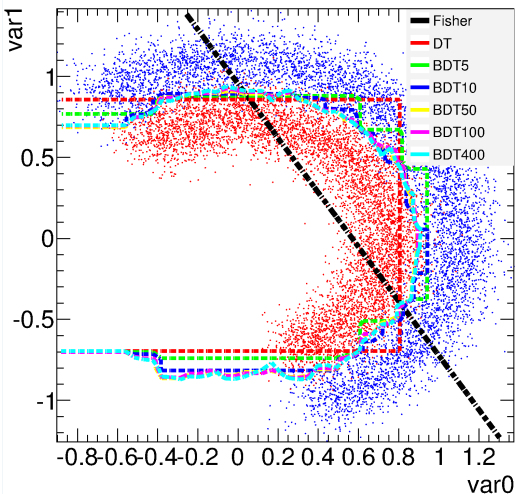
- Compare performance of Fisher discriminant, single DT and BDT with more and more trees (5 to 400)
- All other parameters at TMVA default (would be 400 trees)

Background rejection versus Signal efficiency



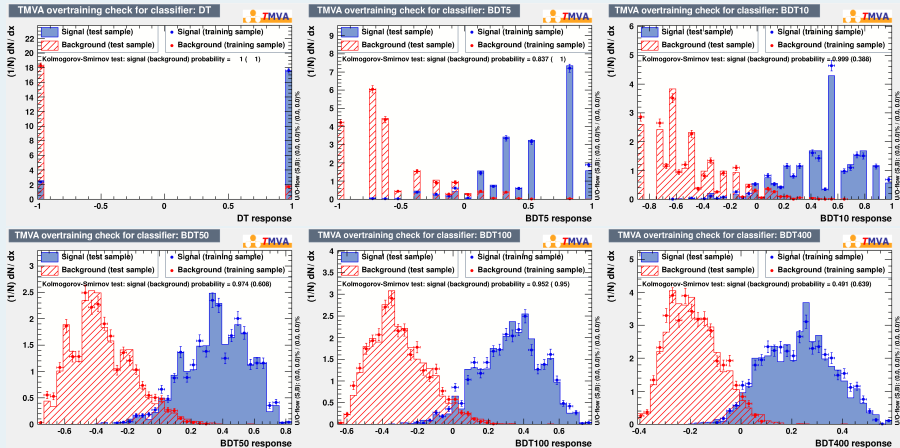
- Fisher bad (expected)
- Single (small) DT: not so good
- More trees \Rightarrow improve performance until saturation

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 \Rightarrow improve performance (less step-like, closer to optimal separation) until saturation
- Largest BDTs: wiggle a little around the contour \Rightarrow picked up features of training sample, that is, overtraining

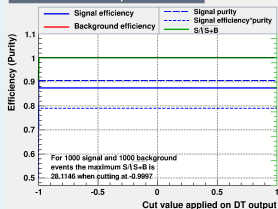
Training/testing output



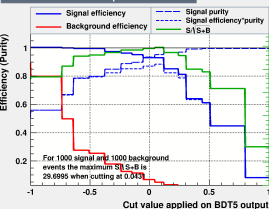
- Better shape with more trees: quasi-continuous
- Overtraining because of disagreement between training and testing?
Let's see

Performance in optimal significance

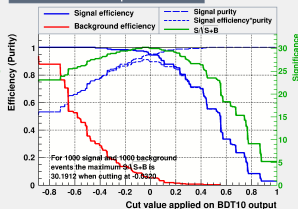
Cut efficiencies and optimal cut value



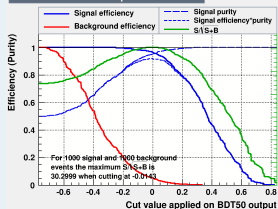
Cut efficiencies and optimal cut value



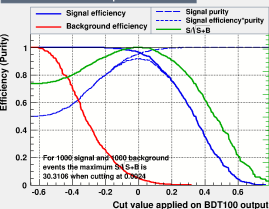
Cut efficiencies and optimal cut value



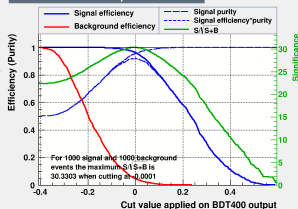
Cut efficiencies and optimal cut value



Cut efficiencies and optimal cut value



Cut efficiencies and optimal cut value

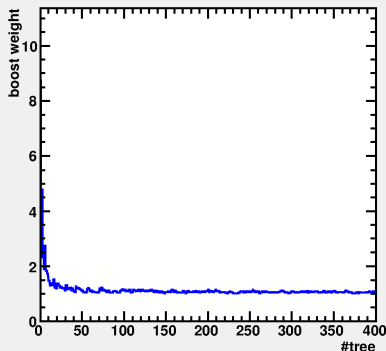


- Best significance actually obtained with last BDT, 400 trees!
- But to be fair, equivalent performance with 10 trees already
- Less “stepped” output desirable? \Rightarrow maybe 50 is reasonable

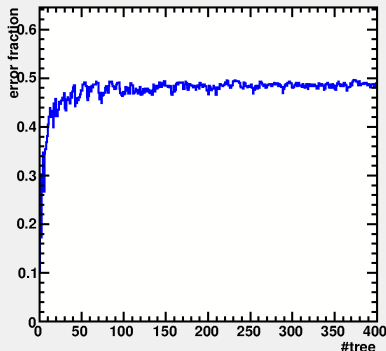
Control plots

- 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

Boost weights vs tree



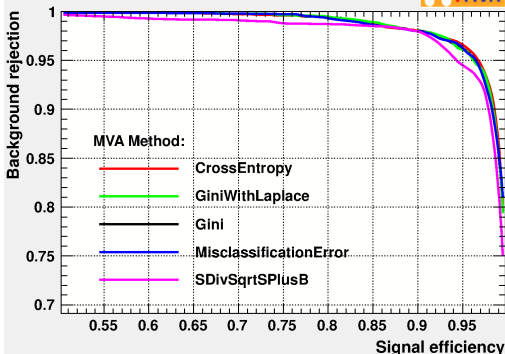
error fraction vs tree number



Separation criterion for node splitting

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

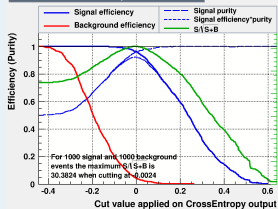
Background rejection versus Signal efficiency



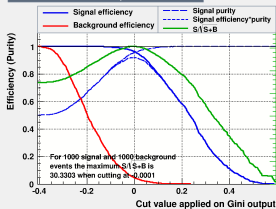
- Very similar performance (even zooming on corner)
- Small degradation (in this particular case) for $\frac{s}{\sqrt{s+b}}$: only criterion that doesn't respect good properties of impurity measure (see yesterday: maximal for equal mix of signal and bkg, symmetric in p_{sig} and p_{bkg} , minimal for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

Performance in optimal significance

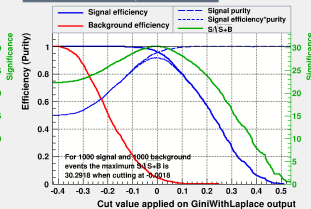
Cut efficiencies and optimal cut value



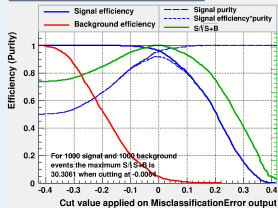
Cut efficiencies and optimal cut value



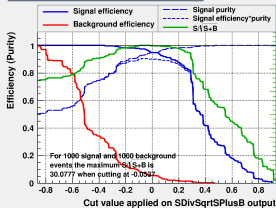
Cut efficiencies and optimal cut value



Cut efficiencies and optimal cut value



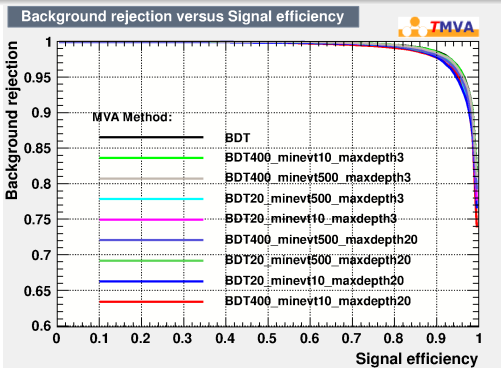
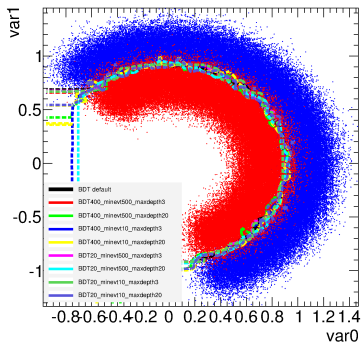
Cut efficiencies and optimal cut value



- 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



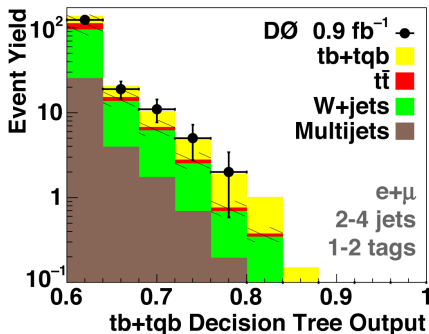
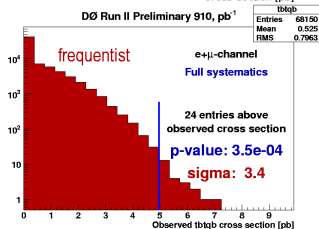
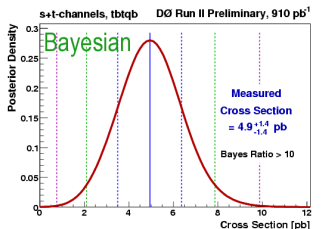
- Overall: very comparable performance. Depends on use case.

- 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

$\sigma_{s+t} = 4.9 \pm 1.4 \text{ pb}$
 $p\text{-value} = 0.035\% (3.4\sigma)$
 SM compatibility: 11% (1.3σ)



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

Object Kinematics

$p_T(\text{jet1})$
 $p_T(\text{jet2})$
 $p_T(\text{jet3})$
 $p_T(\text{jet4})$
 $p_T(\text{best1})$
 $p_T(\text{notbest1})$
 $p_T(\text{notbest2})$
 $p_T(\text{tag1})$
 $p_T(\text{untag1})$
 $p_T(\text{untag2})$

Angular Correlations

$\Delta R(\text{jet1}, \text{jet2})$
 $\cos(\text{best1}, \text{lepton})_{\text{besttop}}$
 $\cos(\text{best1}, \text{notbest1})_{\text{besttop}}$
 $\cos(\text{tag1}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{tag1}, \text{lepton})_{\text{btaggedtop}}$
 $\cos(\text{jet1}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{jet1}, \text{lepton})_{\text{btaggedtop}}$
 $\cos(\text{jet2}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{jet2}, \text{lepton})_{\text{btaggedtop}}$
 $\cos(\text{lepton}, Q(\text{lepton}) \times z)_{\text{besttop}}$
 $\cos(\text{lepton}_{\text{besttop}}, \text{besttop}_{\text{CMframe}})$
 $\cos(\text{lepton}_{\text{btaggedtop}}, \text{btaggedtop}_{\text{CMframe}})$
 $\cos(\text{notbest}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{notbest}, \text{lepton})_{\text{besttop}}$
 $\cos(\text{untag1}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{untag1}, \text{lepton})_{\text{btaggedtop}}$

Event Kinematics

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

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

Object Kinematics

$p_T(\text{jet1})$
 $p_T(\text{jet2})$
 $p_T(\text{jet3})$
 $p_T(\text{jet4})$
 $p_T(\text{best1})$
 $p_T(\text{notbest1})$
 $p_T(\text{notbest2})$
 $p_T(\text{tag1})$
 $p_T(\text{untag1})$
 $p_T(\text{untag2})$

Angular Correlations

$\Delta R(\text{jet1}, \text{jet2})$
 $\cos(\text{best1}, \text{lepton})_{\text{besttop}}$
 $\cos(\text{best1}, \text{notbest1})_{\text{besttop}}$
 $\cos(\text{tag1}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{tag1}, \text{lepton})_{\text{btaggedtop}}$
 $\cos(\text{jet1}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{jet1}, \text{lepton})_{\text{btaggedtop}}$
 $\cos(\text{jet2}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{jet2}, \text{lepton})_{\text{btaggedtop}}$
 $\cos(\text{lepton}, Q(\text{lepton}) \times z)_{\text{besttop}}$
 $\cos(\text{lepton}_{\text{besttop}}, \text{besttop}_{\text{CMframe}})$
 $\cos(\text{lepton}_{\text{btaggedtop}}, \text{btaggedtop}_{\text{CMframe}})$
 $\cos(\text{notbest}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{notbest}, \text{lepton})_{\text{besttop}}$
 $\cos(\text{untag1}, \text{alljets})_{\text{alljets}}$
 $\cos(\text{untag1}, \text{lepton})_{\text{btaggedtop}}$

Event Kinematics

Aplanarity(alljets, W)
 $M(W, \text{best1})$ (“best” top mass)
 $M(W, \text{tag1})$ (“ b -tagged” top mass)
 $H_T(\text{alljets})$
 $H_T(\text{alljets} - \text{best1})$
 $H_T(\text{alljets} - \text{tag1})$
 $H_T(\text{alljets}, W)$
 $H_T(\text{jet1}, \text{jet2})$
 $H_T(\text{jet1}, \text{jet2}, W)$
 $M(\text{alljets})$
 $M(\text{alljets} - \text{best1})$
 $M(\text{alljets} - \text{tag1})$
 $M(\text{jet1}, \text{jet2})$
 $M(\text{jet1}, \text{jet2}, W)$
 $M_T(\text{jet1}, \text{jet2})$
 $M_T(W)$
 Missing E_T
 $p_T(\text{alljets} - \text{best1})$
 $p_T(\text{alljets} - \text{tag1})$
 $p_T(\text{jet1}, \text{jet2})$
 $Q(\text{lepton}) \times \eta(\text{untag1})$
 \sqrt{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: $H_T(\text{alljets}, W)$. But detector not perfect \Rightarrow capture the essence from several variations usually helps “dumb” MVA

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

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)

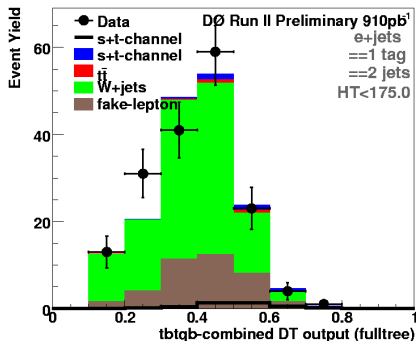
Ensemble generation

- 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)

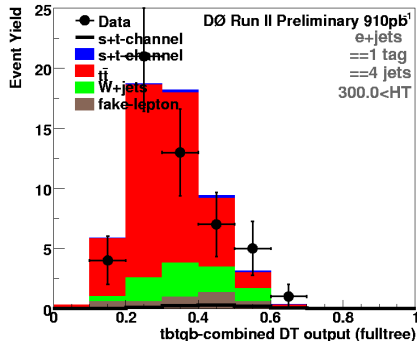
Achieved linear response to varying input cross sections and negligible bias

- Validate method on data in no-signal region

- **“W+jets”**: = 2 jets,
 $H_T(\text{lepton}, \cancel{E}_T, \text{alljets}) < 175 \text{ GeV}$

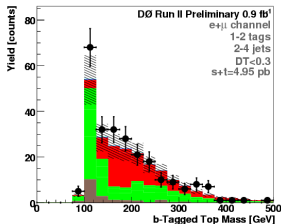


- **“ttbar”**: = 4 jets,
 $H_T(\text{lepton}, \cancel{E}_T, \text{alljets}) > 300 \text{ GeV}$

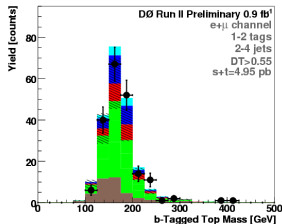


- Good agreement

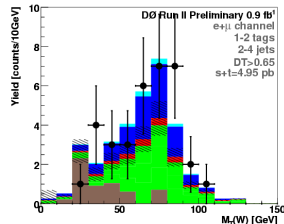
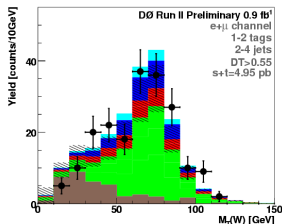
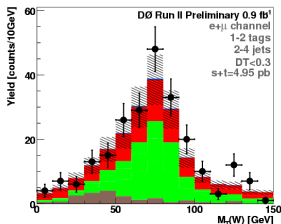
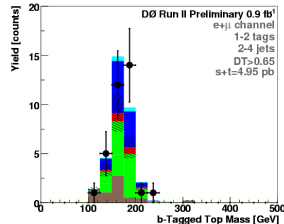
$DT < 0.3$



$DT > 0.55$

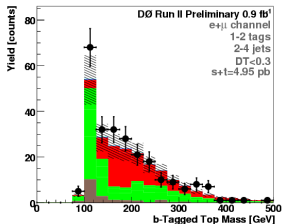


$DT > 0.65$

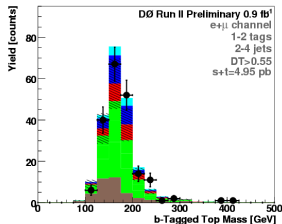


- High BDT region = shows masses of real t and $W \Rightarrow$ expected
- Low BDT region = background-like \Rightarrow expected

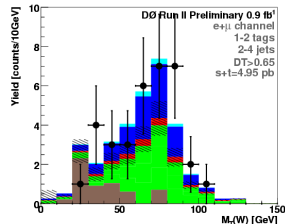
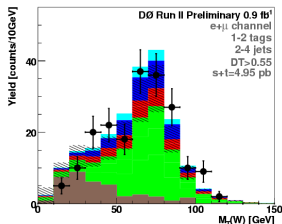
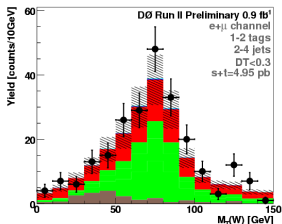
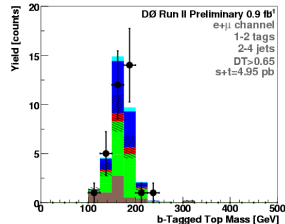
$DT < 0.3$



$DT > 0.55$

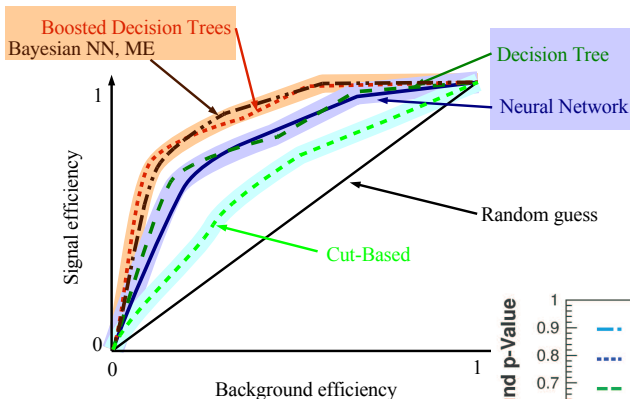


$DT > 0.65$



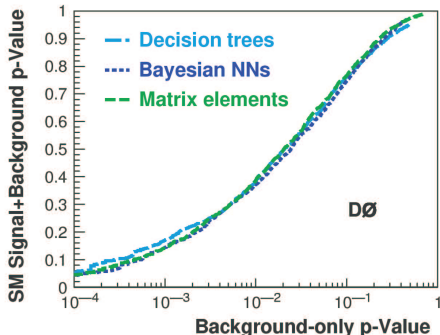
- High BDT region = shows masses of real t and $W \Rightarrow$ expected
- Low BDT region = background-like \Rightarrow expected
- Above doesn't tell analysis is ok, but not seeing this could be a sign of a problem

Comparison for D0 single top evidence



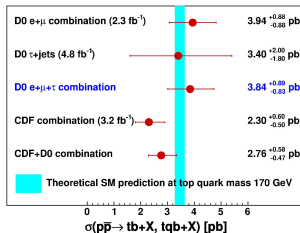
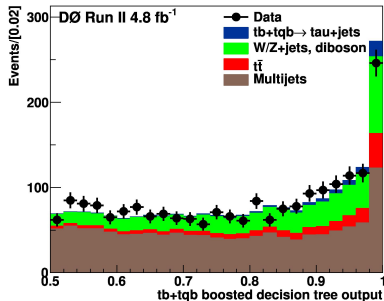
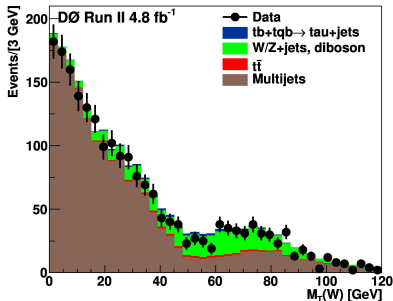
- Cannot know *a priori* which method will work best
- \Rightarrow Need to experiment with different techniques

Power curve



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

- Tau ID BDT and single top search BDT

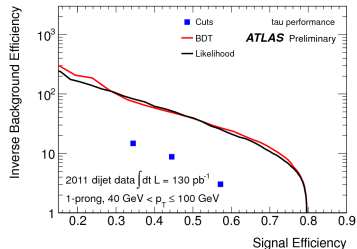
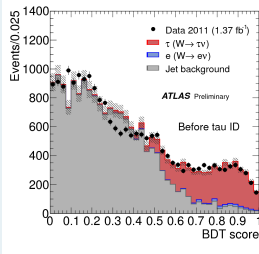


- 4% sensitivity gain over e + μ analysis

▶ PLB690:5,2010

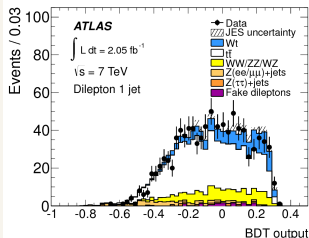
ATLAS tau identification

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



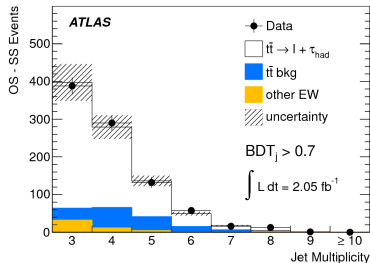
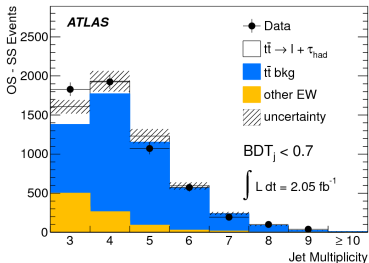
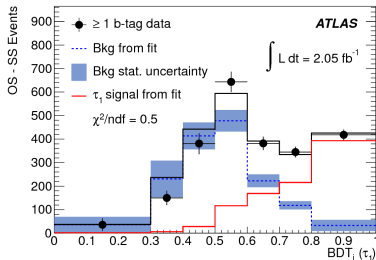
ATLAS Wt production evidence

- [Phys.Lett. B716 \(2012\) 142-159](#)
- BDT output used in final fit to measure cross section
- Constraints on systematics from profiling



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

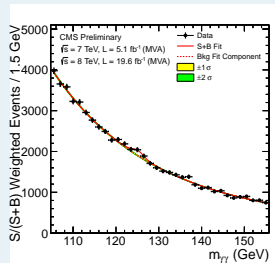
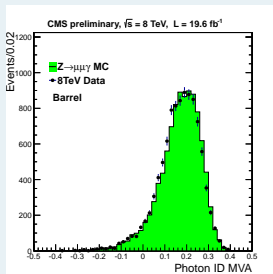
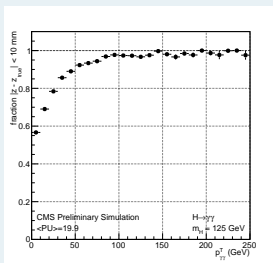
- BDT for tau ID: one to reject electrons, one against jets
- Fit BDT output to get tau contribution in data



CMS $H \rightarrow \gamma\gamma$ result

► CMS-PAS-HIG-13-001 Hard to use more BDT in an analysis:

- vertex selected with BDT
- 2nd vertex BDT to estimate probability to be within 1cm of interaction point
- photon ID with BDT
- photon energy corrected with BDT regression
- event-by-event energy uncertainty from another BDT
- several BDT to extract signal in different categories



- 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), e.g. in ATLAS now
- LHCb: $B_{(s)}^0 \rightarrow \mu\mu$ search, selection of $B_s^0 \rightarrow J/\psi\phi$ for ϕ_s measurement
- More and more underway

Dedicated code








- 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/>







Dedicated code

- 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://statpatrec.sourceforge.net/>
 - **TMVA (Toolkit for MultiVariate Analysis)**
Integrated in ROOT, complete manual <http://tmva.sourceforge.net>

-  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)

-  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)