Jet Superstructure

and Multivariate Studies

Matthew Schwartz

Harvard University

Based on work

with J. Gallicchio, **PRL**, 105:022001,2010 (superstructure)
with K. Black, J. Gallicchio, J. Huth, M. Kagan and B. Tweedie arXiv:1010.3698 (multivariate higgs search)
with Z. Han and Y. Cui, arXiv:1012.2077 (multivariate W-tagging)

LHC Physics Day, CERN
February 4, 2011

INTRODUCTION

A lot of recent work on **jet substructure** and some on jet **superstructure**

- Masses, angularities, filtering/trimming/pruning, subjetiness, planar flow, …
- Interesting theory questions
    - What is optimal?
    - Can we trust monte carlos?
    - Can we compute them more accurately in QCD?
- Variables are useful, but highly correlated
    - e.g. jet mass and jet p_T are closely related

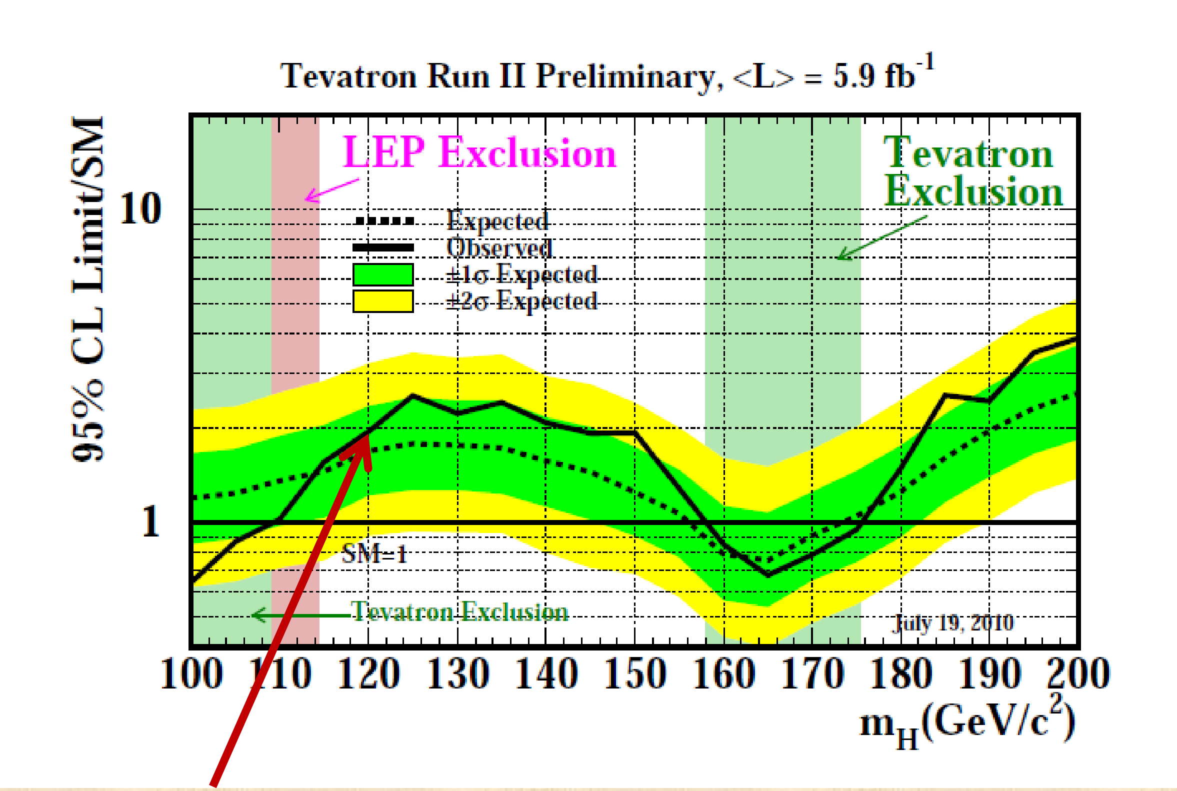
Why **do experimentalists** use multivariate methods: Neural Networks (NN), Boosted Decision Trees (BDT), etc, but **theorists do not**?

- Experimentalists want to see things **early** -- every little bit helps
- To theorists, the difference between 10 fb^{-1} and 100 fb^{-1} is 0
- NNs and BDTs are complicated – theorists are scared of **black boxes**

To properly **appreciate** jets, we must get used to studying variables *and* their correlations

HOW DO WE FIND A LIGHT HIGGS?

Tevatron



- Need a factor of 2 improvement in significance for m_H=120
- Double statistics gives √2
- Where will the other √2 come from?

LHC

- Important search channel is

  pp ⟶ W/Z + H

  H ⟶ bb

- Abandoned by ATLAS and CMS

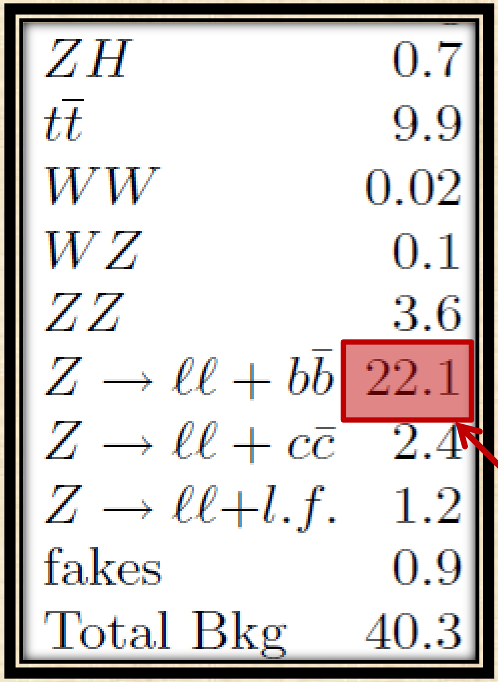
  too much background

- Recently high P_T W/Z + H revived,
  - Requires P_T > 200
  - Lose 95% of signal

How good can we do in W/Z + (H ⟶ bb)?

FOCUS ON $pp \to HZ \to b\bar{b}l^+l^-$

CDF note 10235 (summer 2010)



ZH	0.7
$t\bar{t}$	9.9
WW	0.02
WZ	0.1
ZZ	3.6
$Z \to \ell\ell + b\bar{b}$	22.1
$Z \to \ell\ell + c\bar{c}$	2.4
$Z \to \ell\ell + l.f.$	1.2
fakes	0.9
Total Bkg	40.3

Dominant background is the irreducible one

CDF employs multivariate approach

Inputs to the neural net are

- Missing transverse energy
- Dijet mass
- tt matrix element output
- ZH matrix element output
- Sum of leading jet Pt's
- number of jets

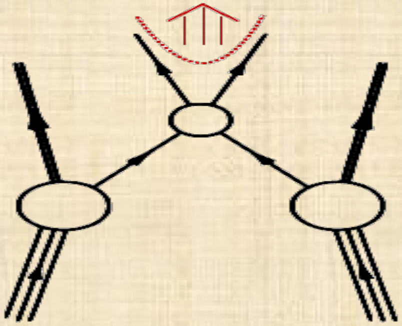
Parton-level kinematics

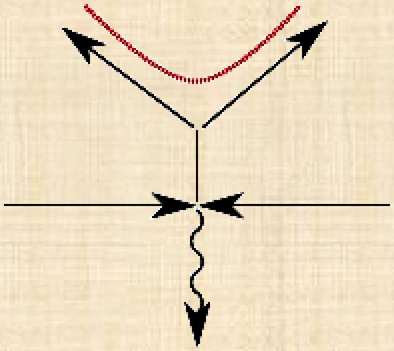
Questions:

- Are there smarter more comprehensive inputs?
- Can we trust the multivariate approach?

ONE THING THEY IGNORE: COLOR

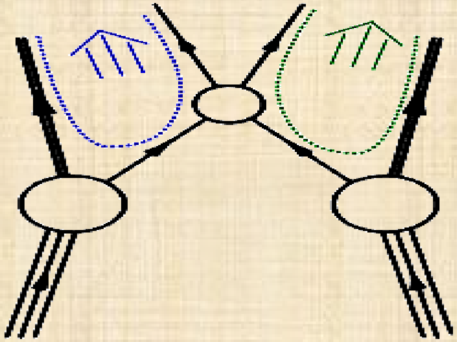
Signal

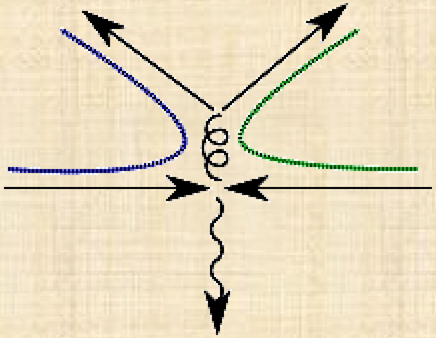




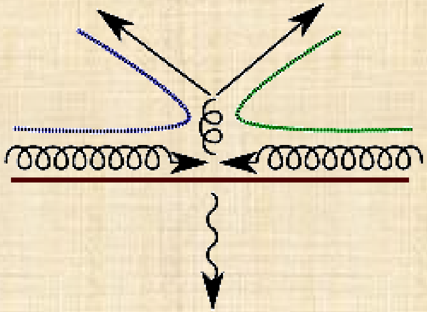
$H \to b\bar{b}$

Background





$q\bar{q} \to Zb\bar{b}$

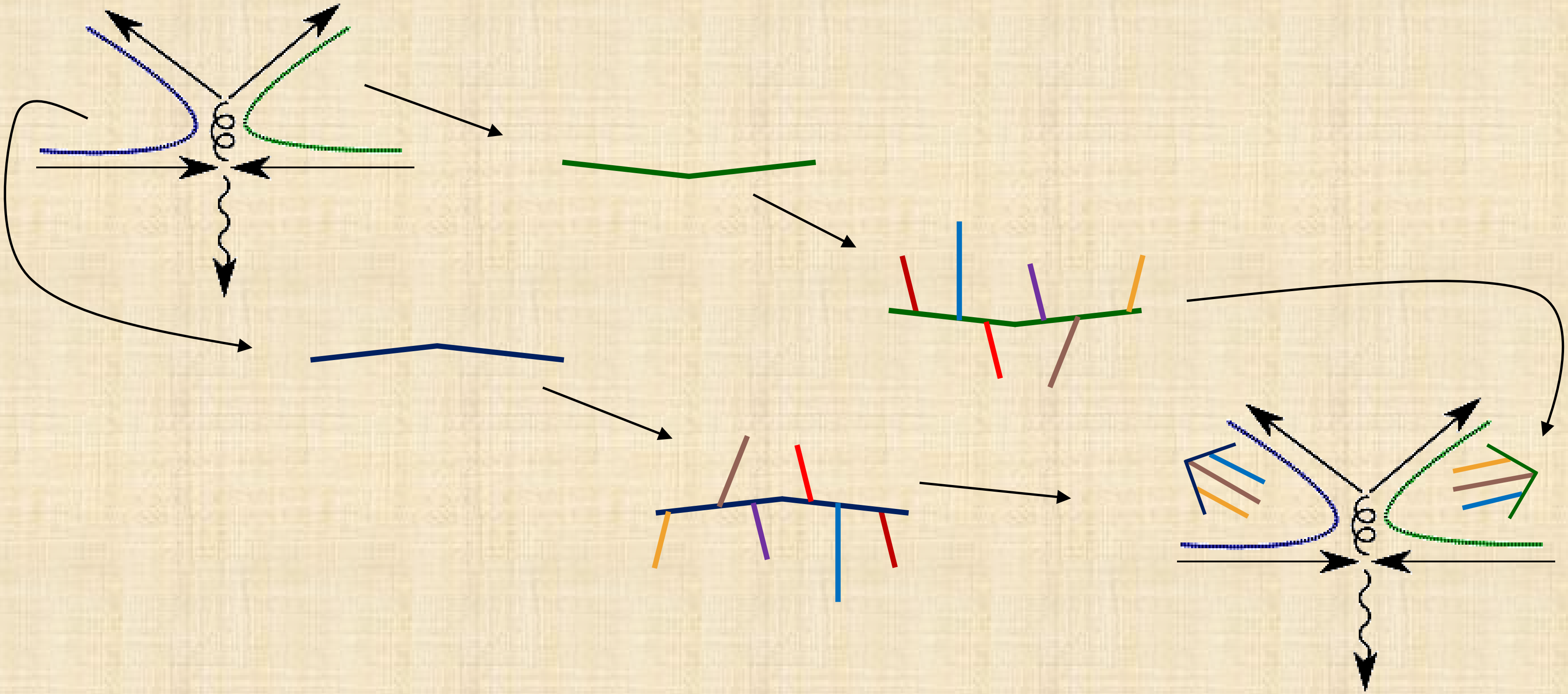


$gg \to Zb\bar{b}$

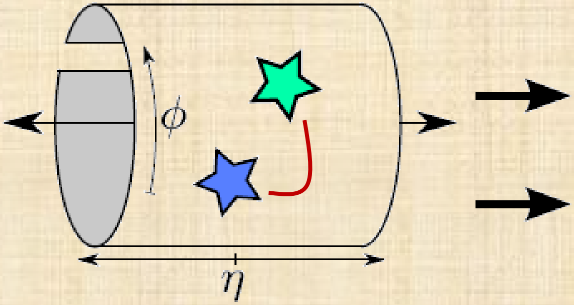
HOW DO THEY SHOW UP?

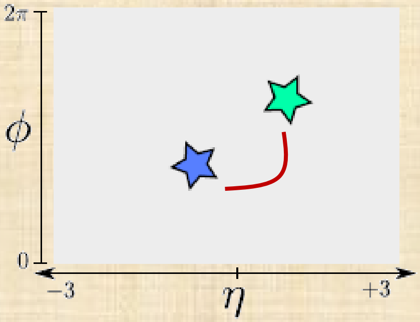
Monte Carlo simulation

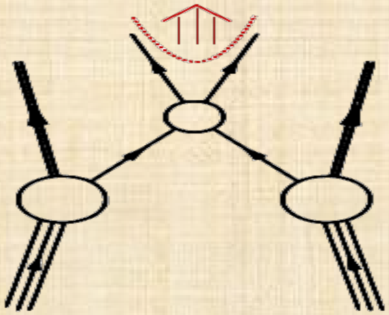
- Color coherence (angular ordering, e.g. Herwig)
- Color string showers in its rest frame (pt ordering, e.g. Pythia)
    - Boost → string showers in string-momentum direction



HOW DO THEY SHOW UP?





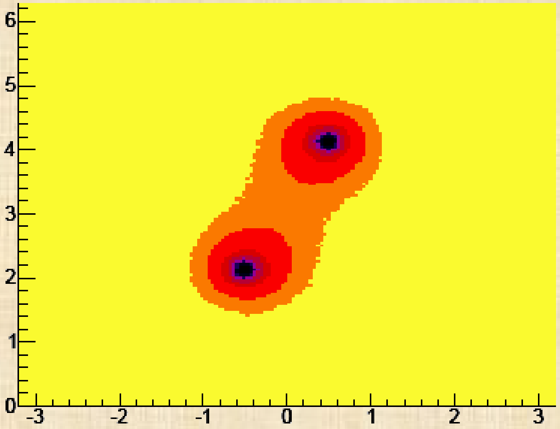


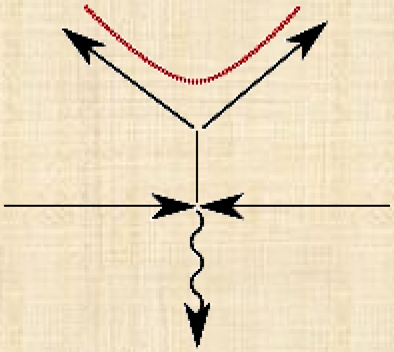
Shower same event *millions* of times

Higgs:

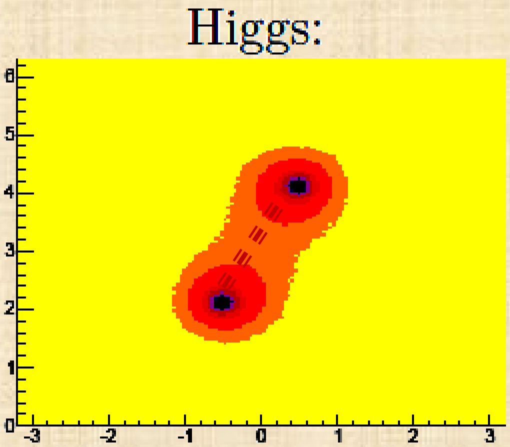
$\Delta\eta_{b\bar{b}} = 1$
$\Delta\phi_{b\bar{b}} = 2$

Add up E_T in each cell:

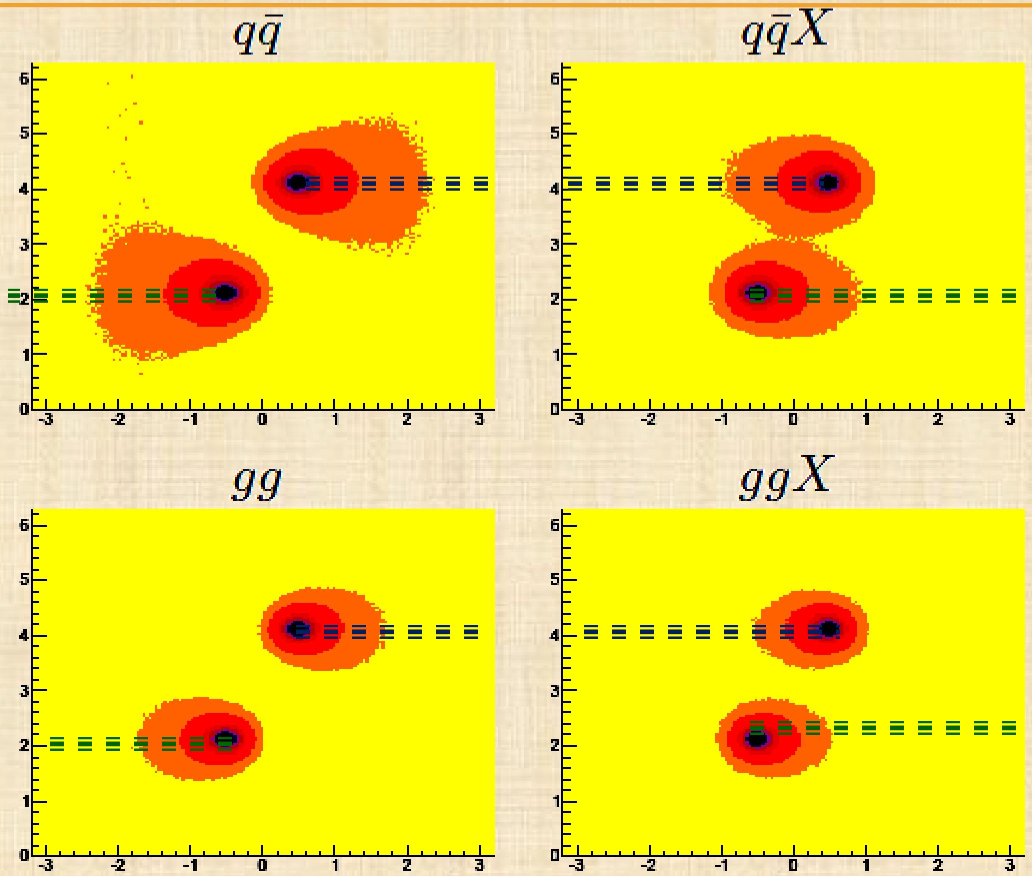




SIGNAL VS BACKGROUND



Higgs:

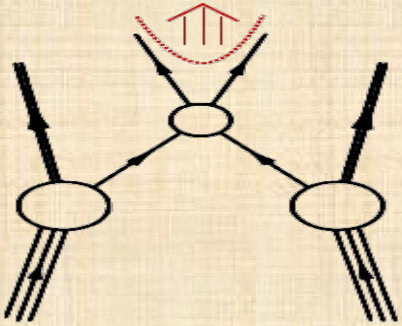


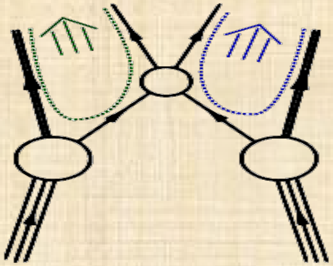
$q\bar{q}$

$q\bar{q}X$

gg

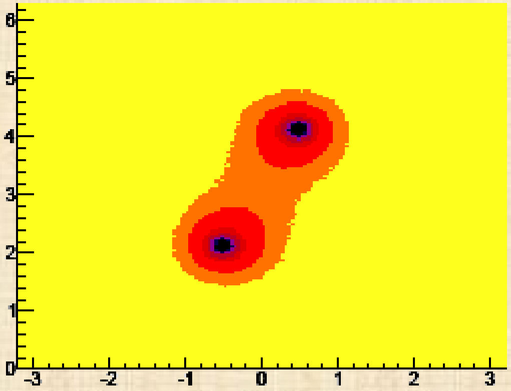
ggX



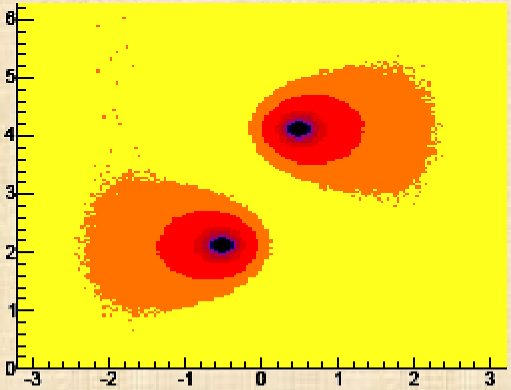


HOW CAN WE USE IT?

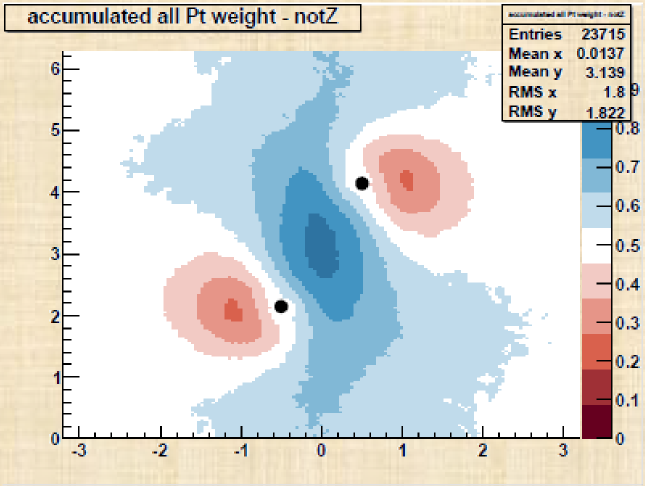
Higgs:



$q\bar{q}$

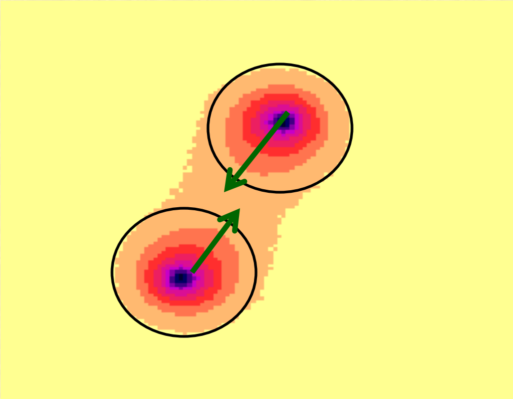


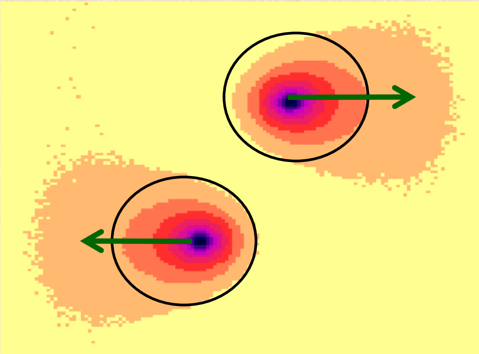
Baysean **probability** that each bit of radiation is **signal**



- Most useful radiation is R = 0.5 – 1.5 away
- Pattern depends strongly on **kinematics**
- Can we find a *simpler* or more *universal* discriminant?

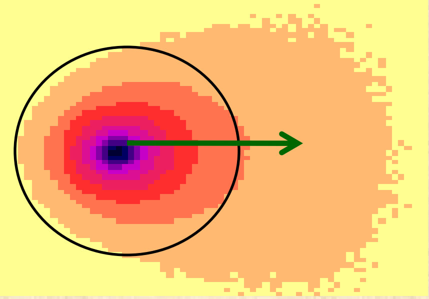
PULL

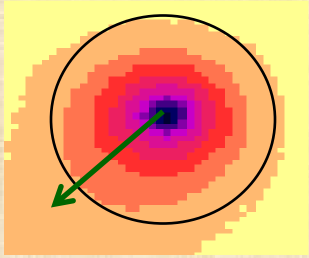




- Find **jets** (e.g. anti-k_T)
- Construct **pull vector** (~ dipole moment) on radiation in **jet**

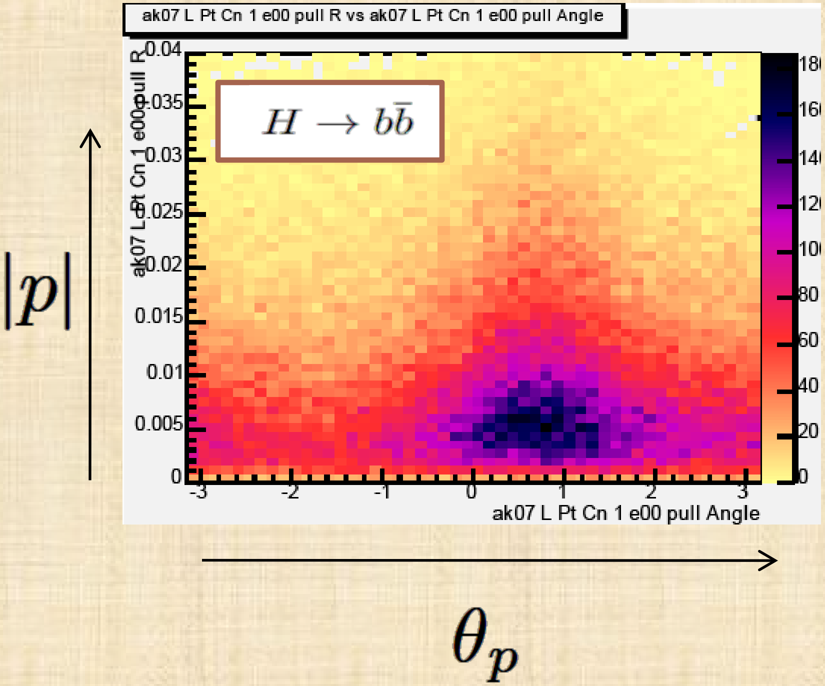
$$\vec{p} = \sum_i \frac{E_T^i \, |r_i|}{E_T^{jet}} \, \vec{r}_i$$

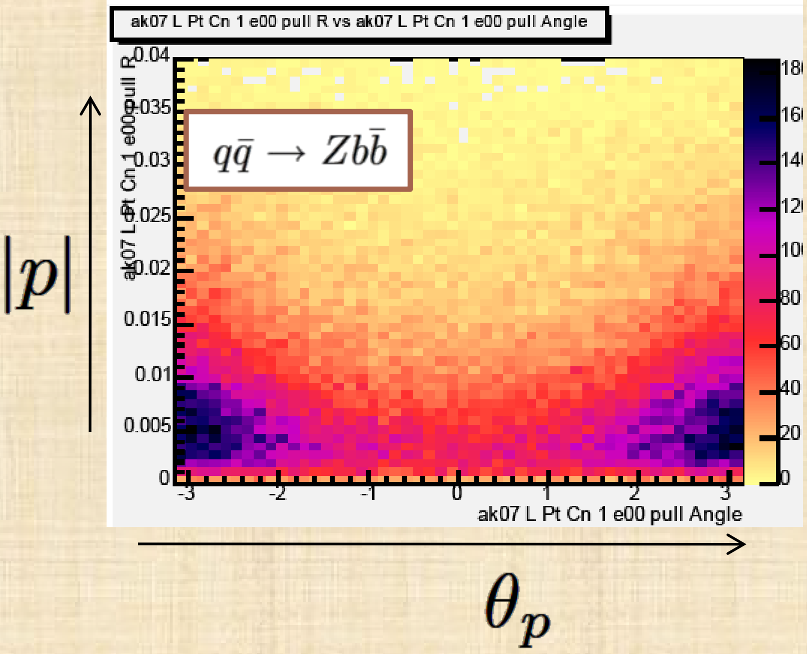




PULL VECTOR IN RADIAL COORDS

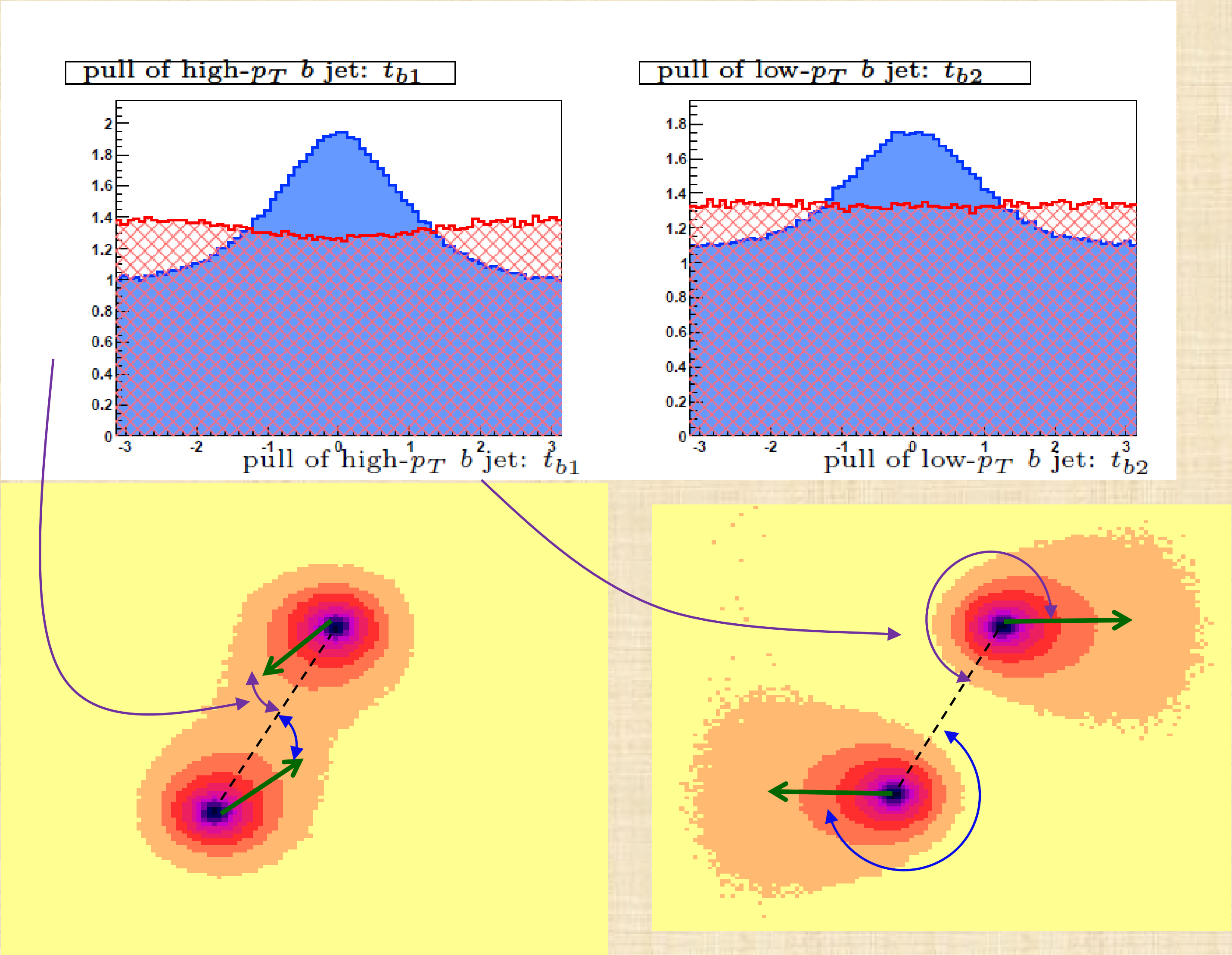
$$\vec{p} = \sum_i \frac{E_T^i \, |r_i|}{E_T^{jet}} \vec{r}_i$$



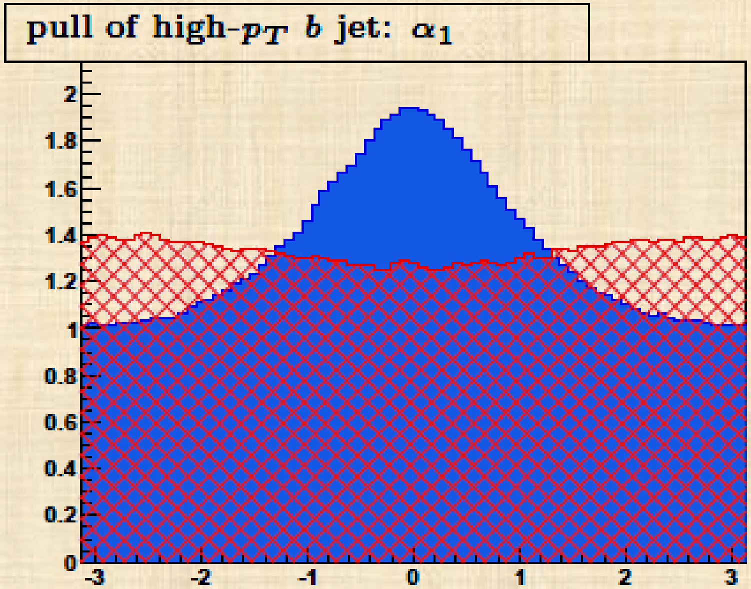


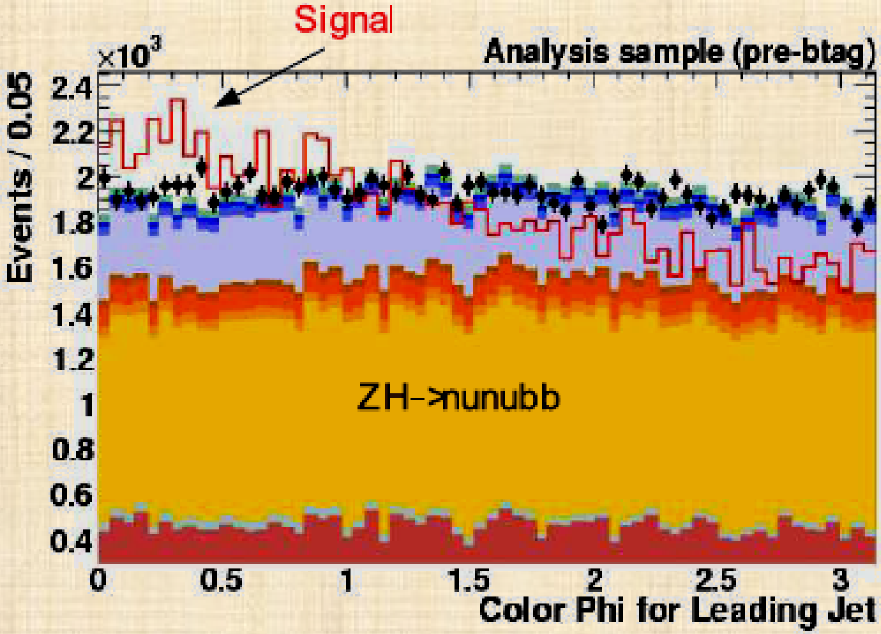
- Angle much more important than length
- Look at relative pull angles

SIGNAL VS BACKGROUND



PULL HAS BEEN MEASURED!



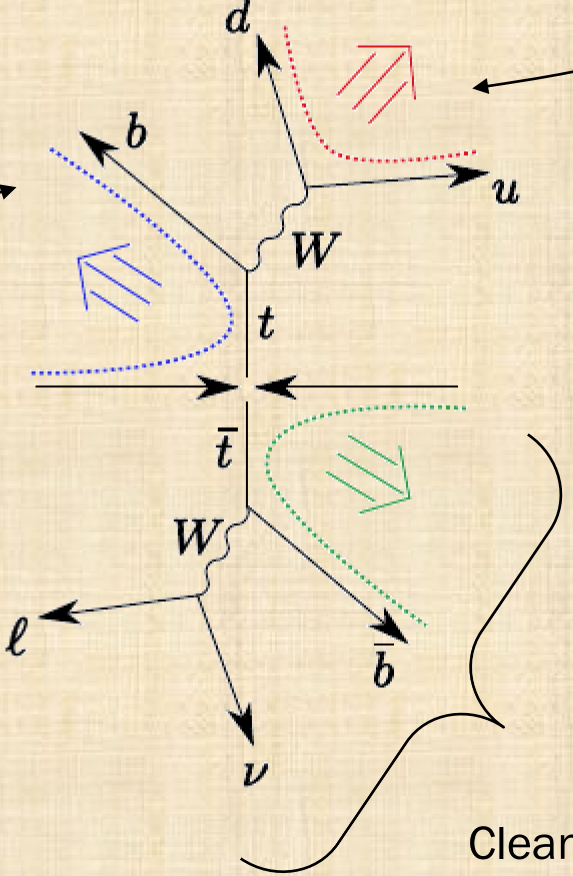


DØ

Note 6087-CONF Aug 2010, Andy Haas: $ZH \to b\bar{b}\nu\bar{\nu}$

(consistent with flat background)

CAN WE VALIDATE? YES! ON TTBAR

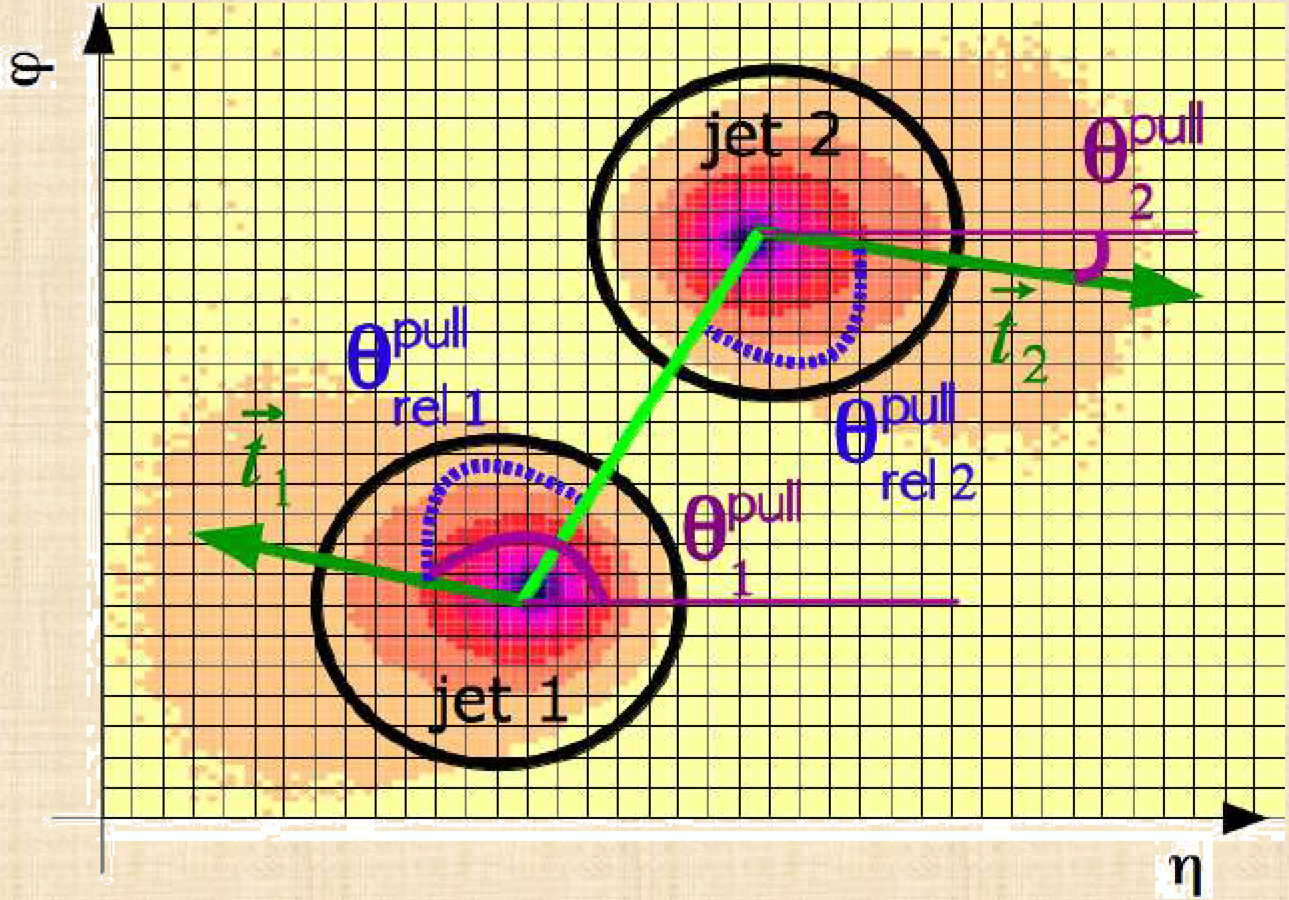


b-tag

Measure pull

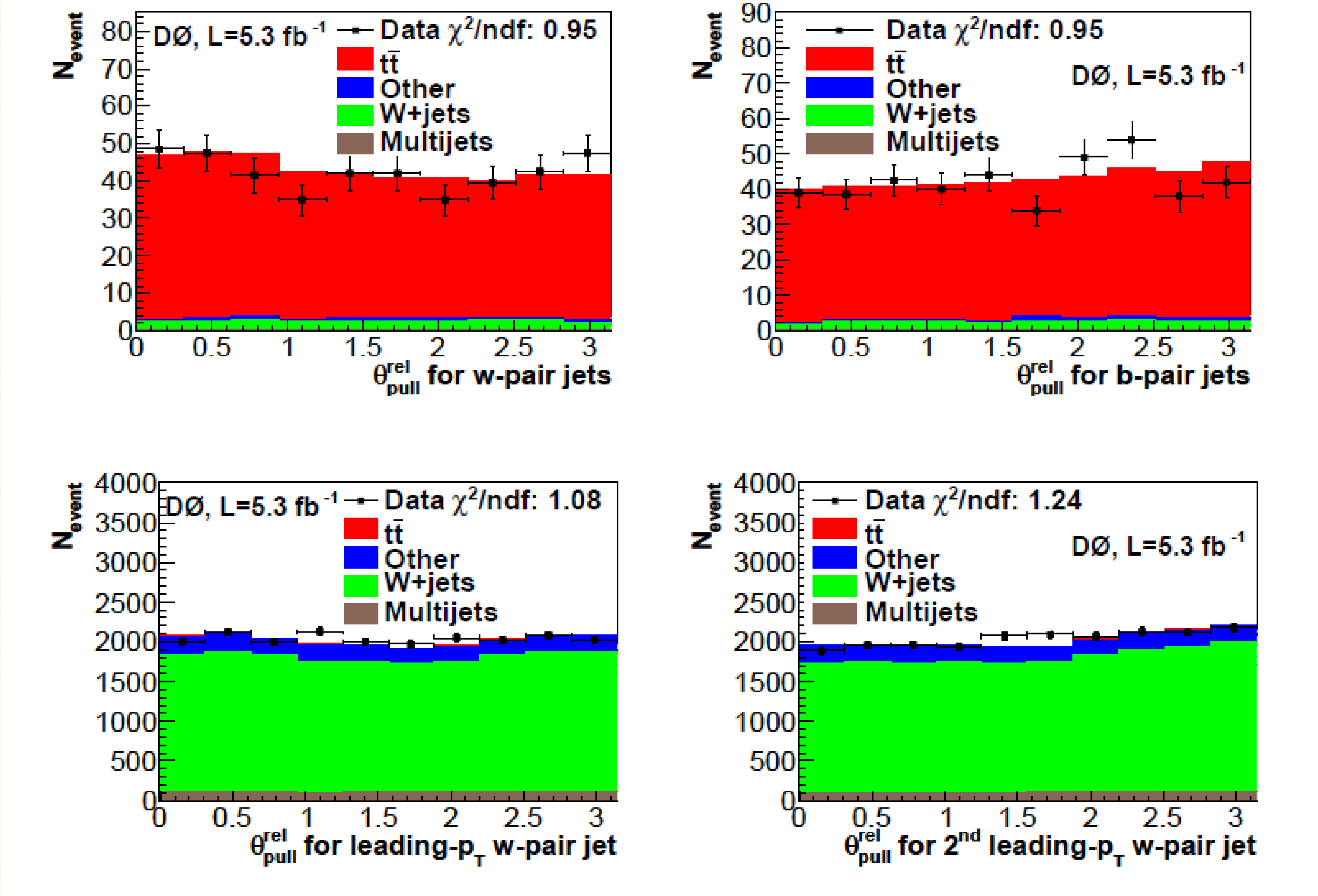
Clean top tag on leptonic side

MANY PULL ANGLES

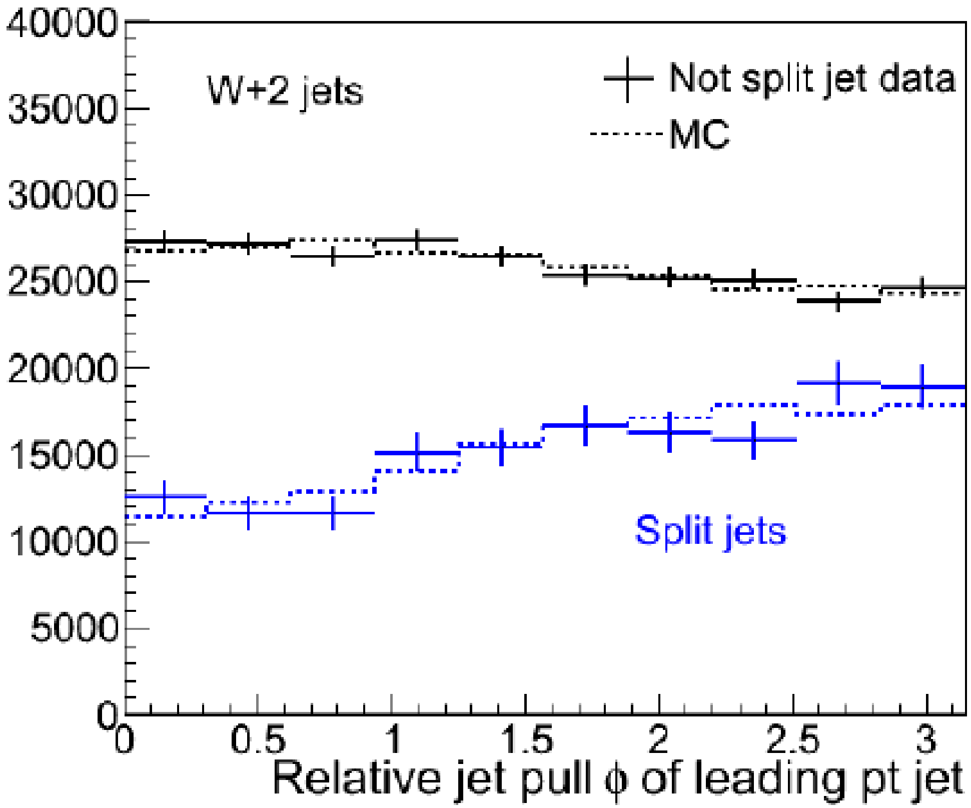


MEASURED BY DZERO

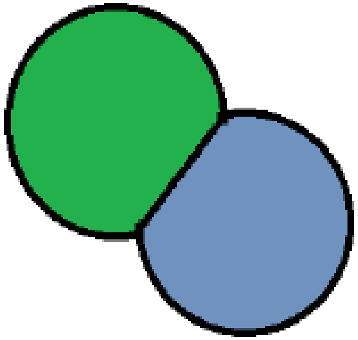
Andy Haas and Yvonne Peters, hep-ex:1101.0648



NON-FLAT PULL

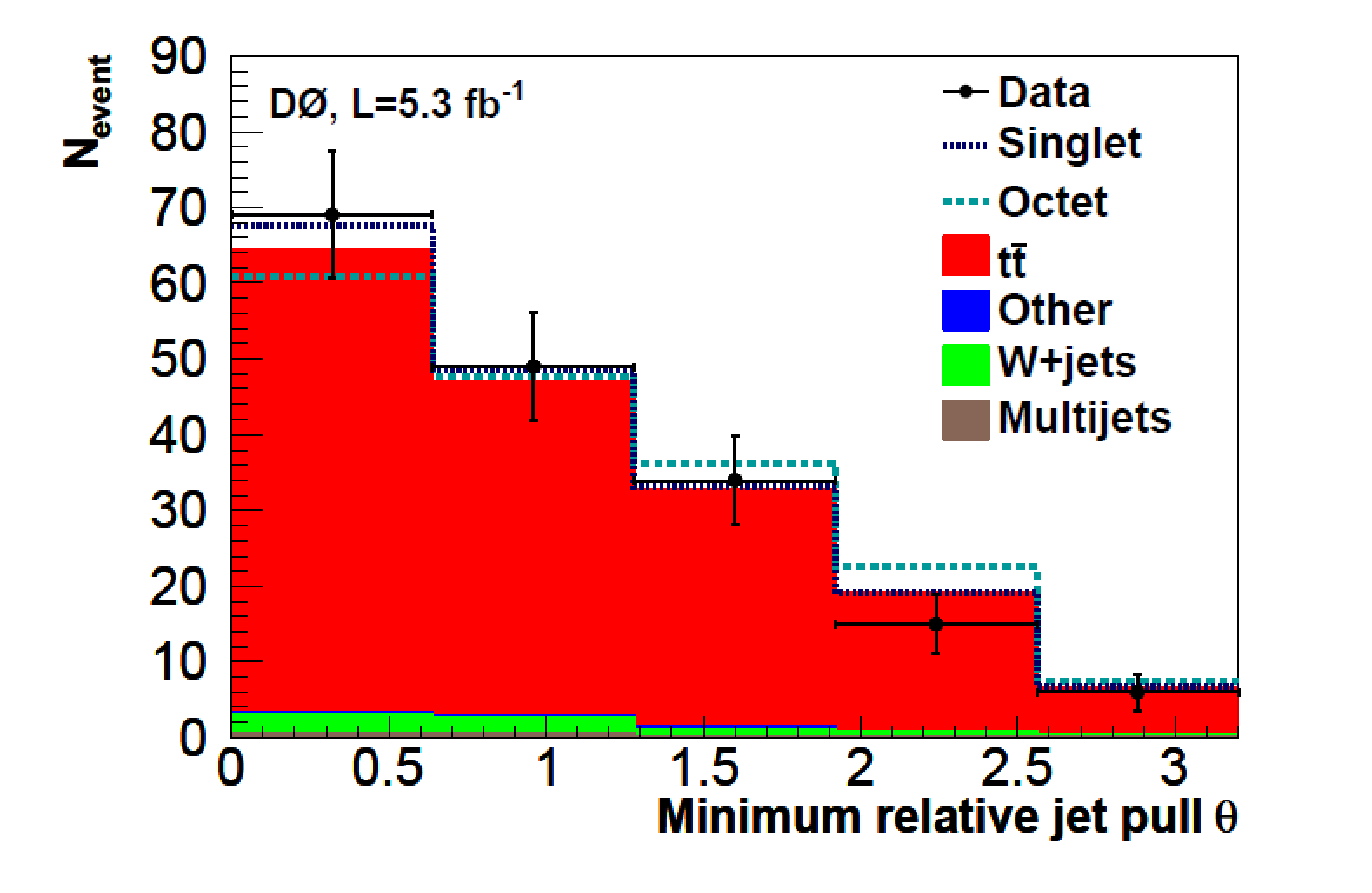


Noise/pileup area smaller towards other jet!

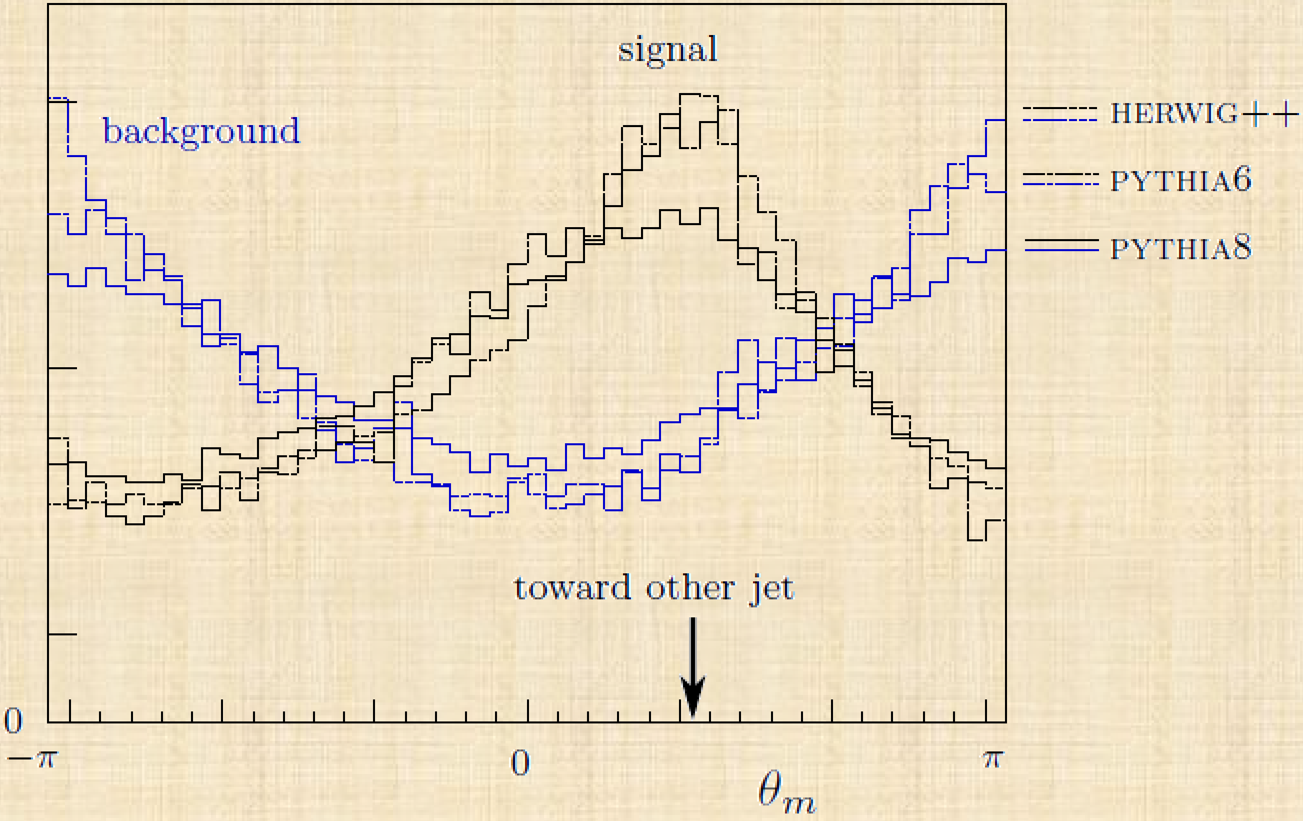


Cells are assigned to the *nearest jet*

RULED OUT COLOR OCTET W



PYTHIA VS HERWIG



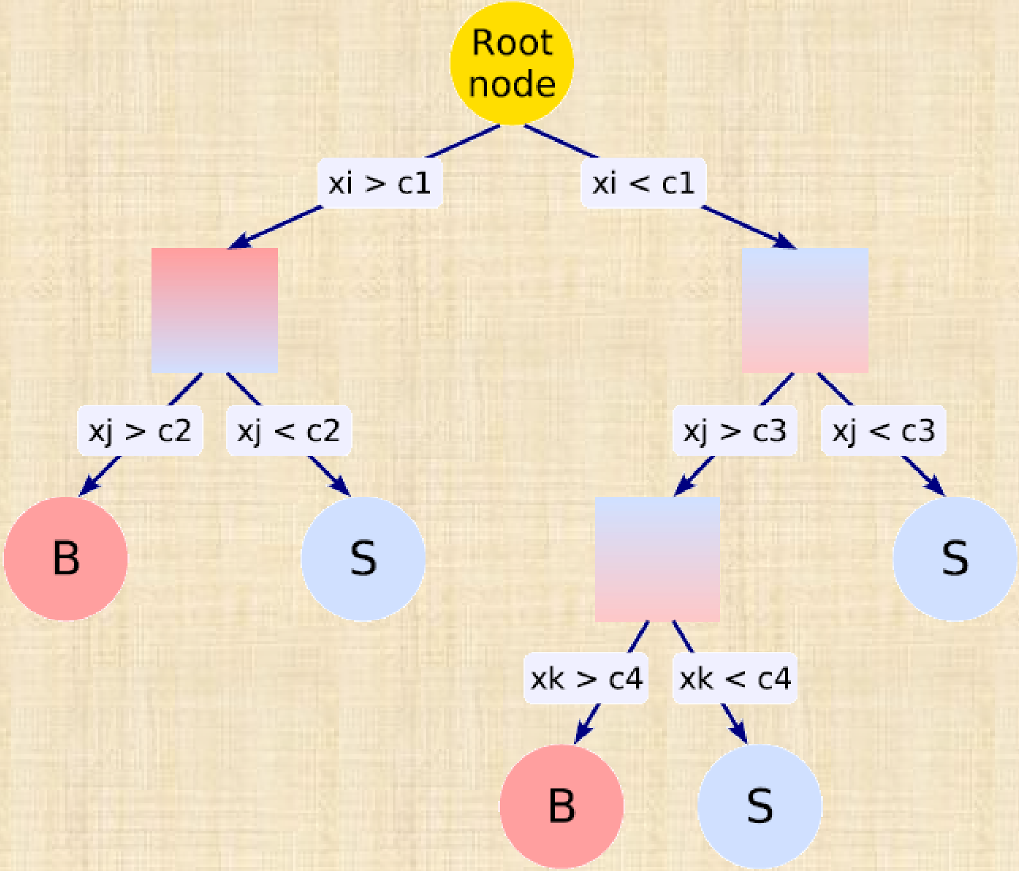
Seems robust.

Can we calculate pull??? Good theory question…

HOW DOES PULL HELP

- According to Dzero it gives around **5% improvement** in pp -> ZH -> νν bb

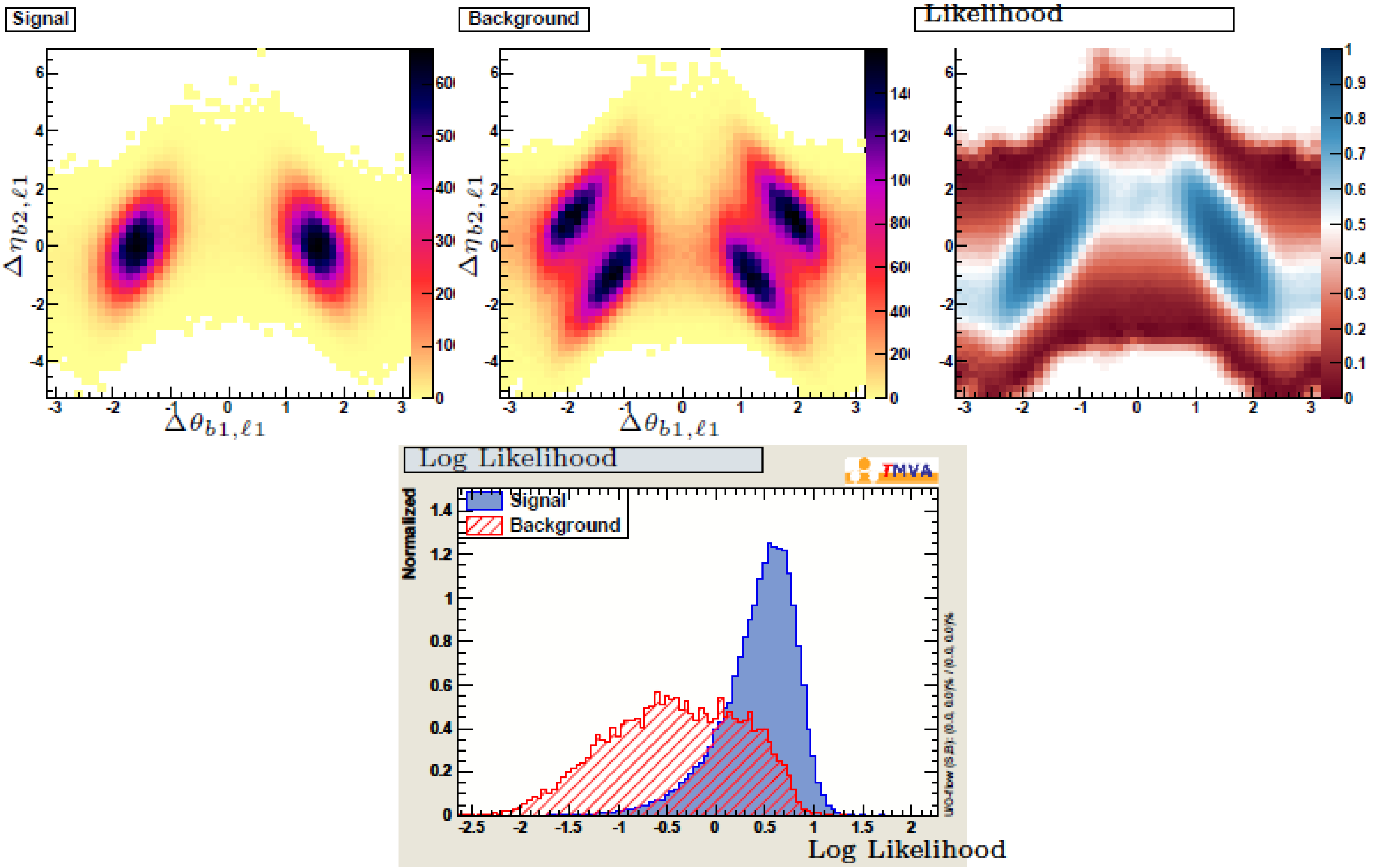
- How does that work? **<u>Boosted Decision Trees</u>**



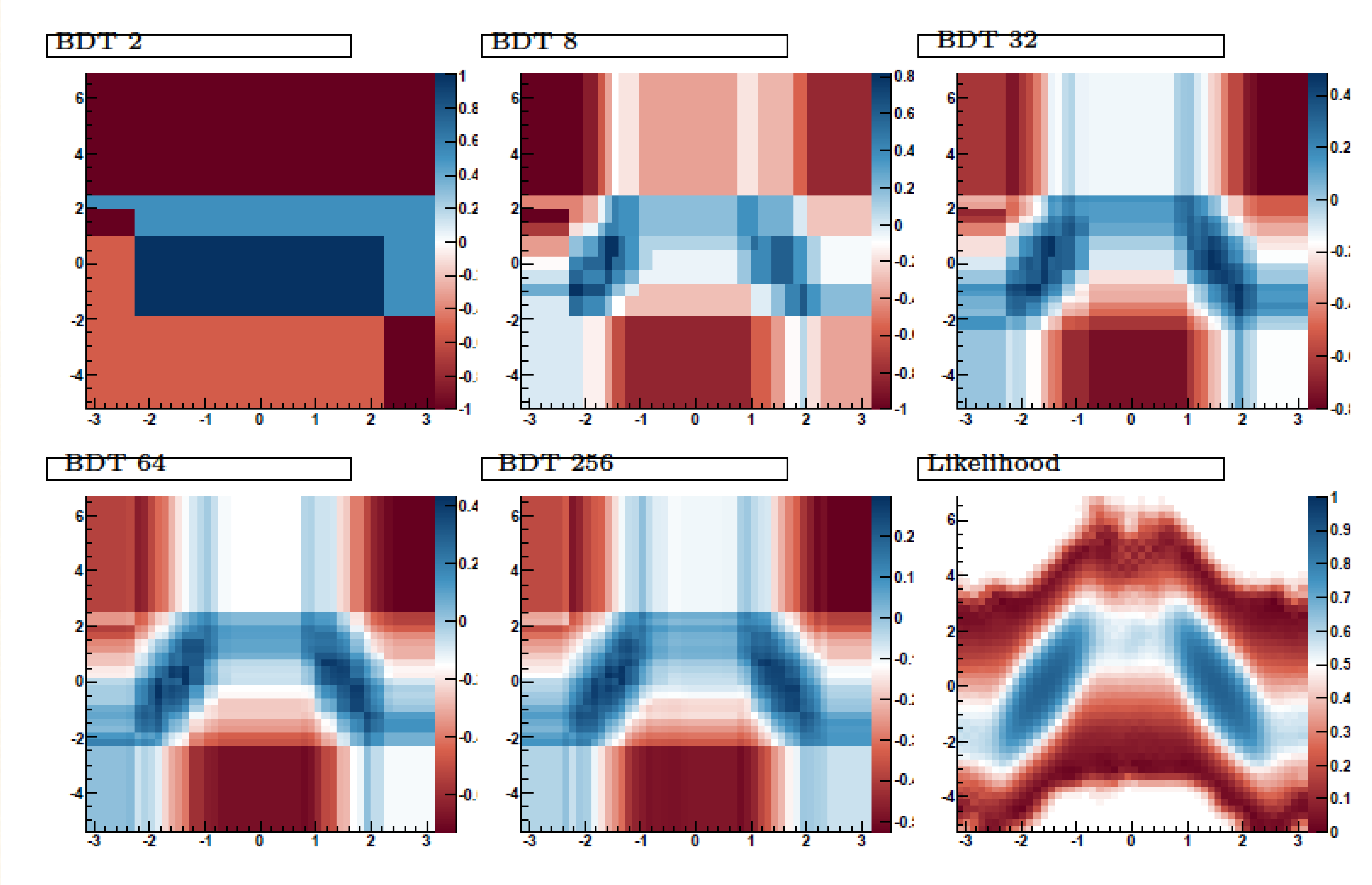
- Train multiple trees and have them vote

- Approximates exact solution.

EXACT SOLUTION: LIKELIHOOD

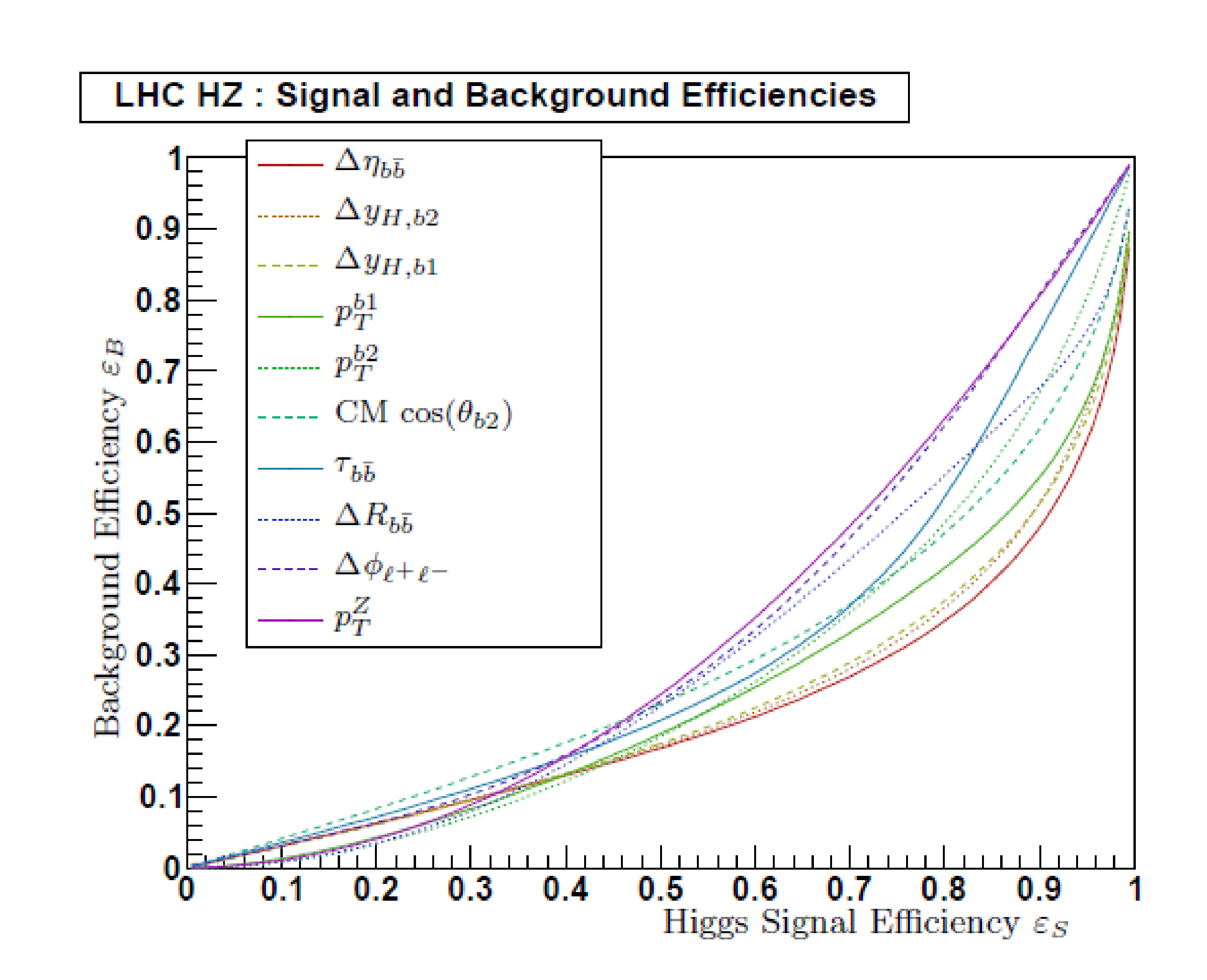


MULTIDIMENSIONS: APPROXIMATE



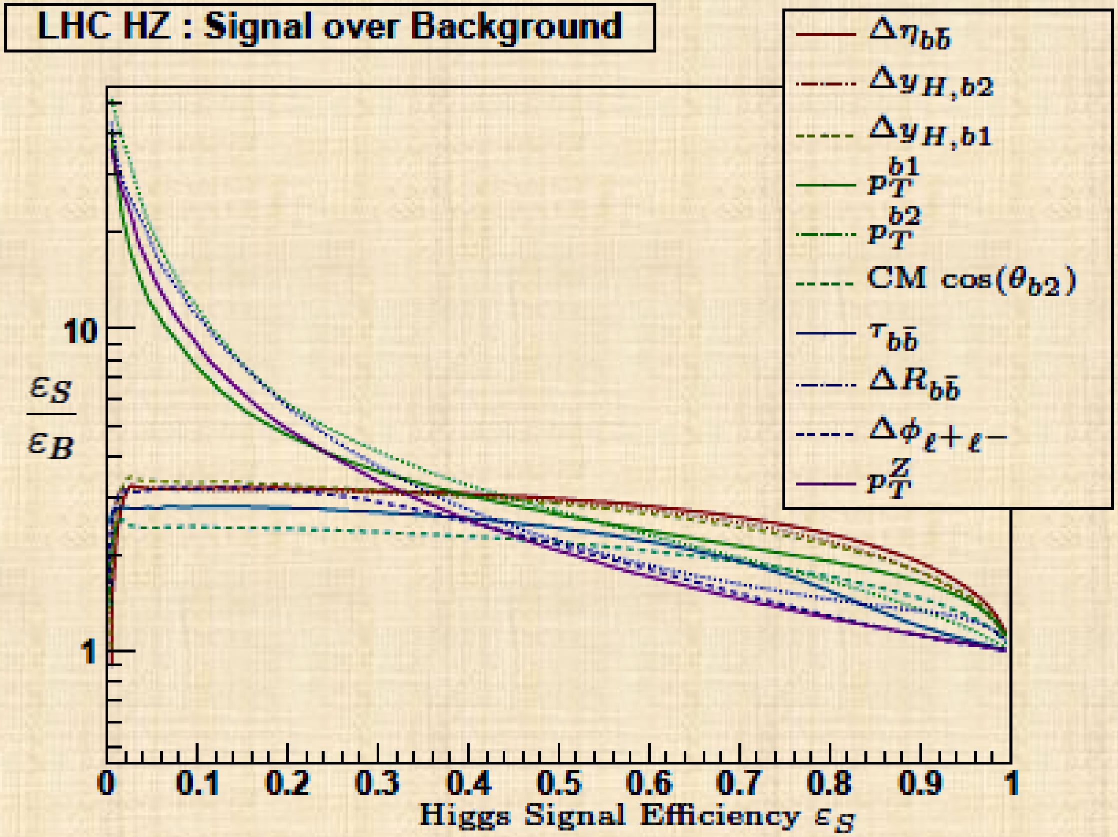
VISUALIZE IMPROVEMENT

Receiver Operator Characteristic (ROC)



TAKE RATIOS

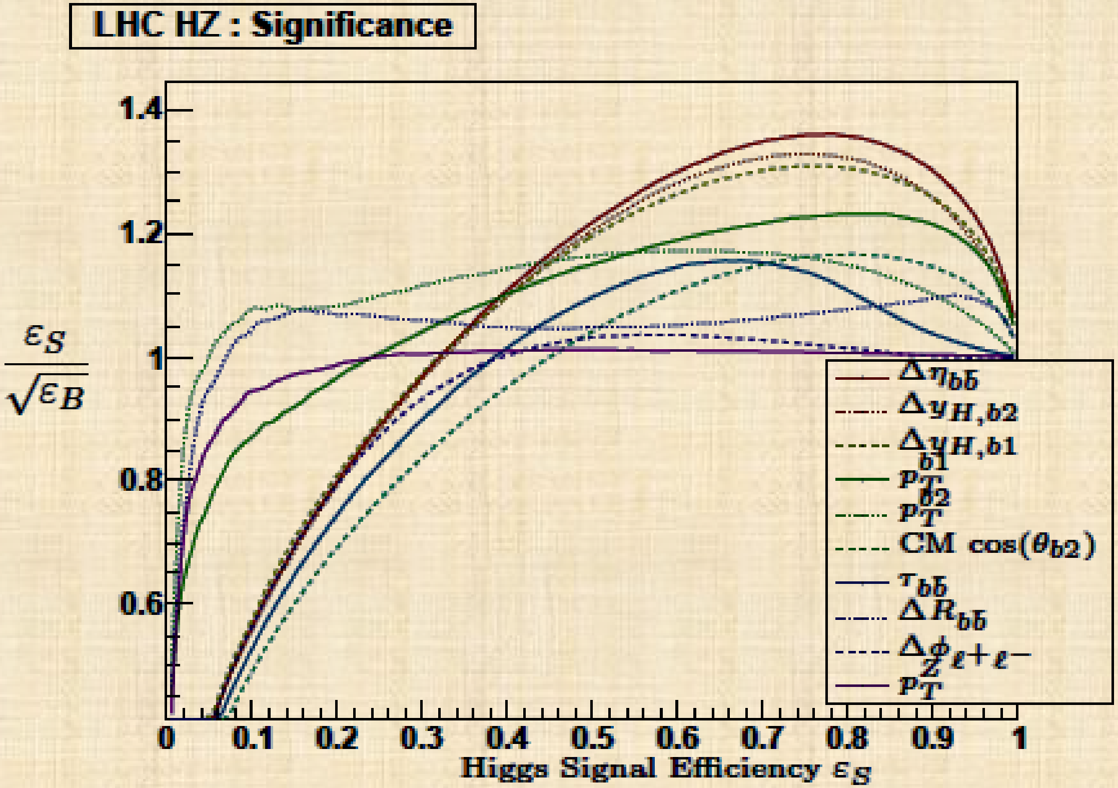
$$\frac{S}{B} \quad \xrightarrow{\text{cut}} \quad \frac{\varepsilon_S S}{\varepsilon_S B} = \left(\frac{\varepsilon_S}{\varepsilon_B}\right) \frac{S}{B}\,.$$



- •Hard cuts make arbitrary large improvement in S/B
- •S/B important, but misleading to optimize or compare variables

SIC CURVES

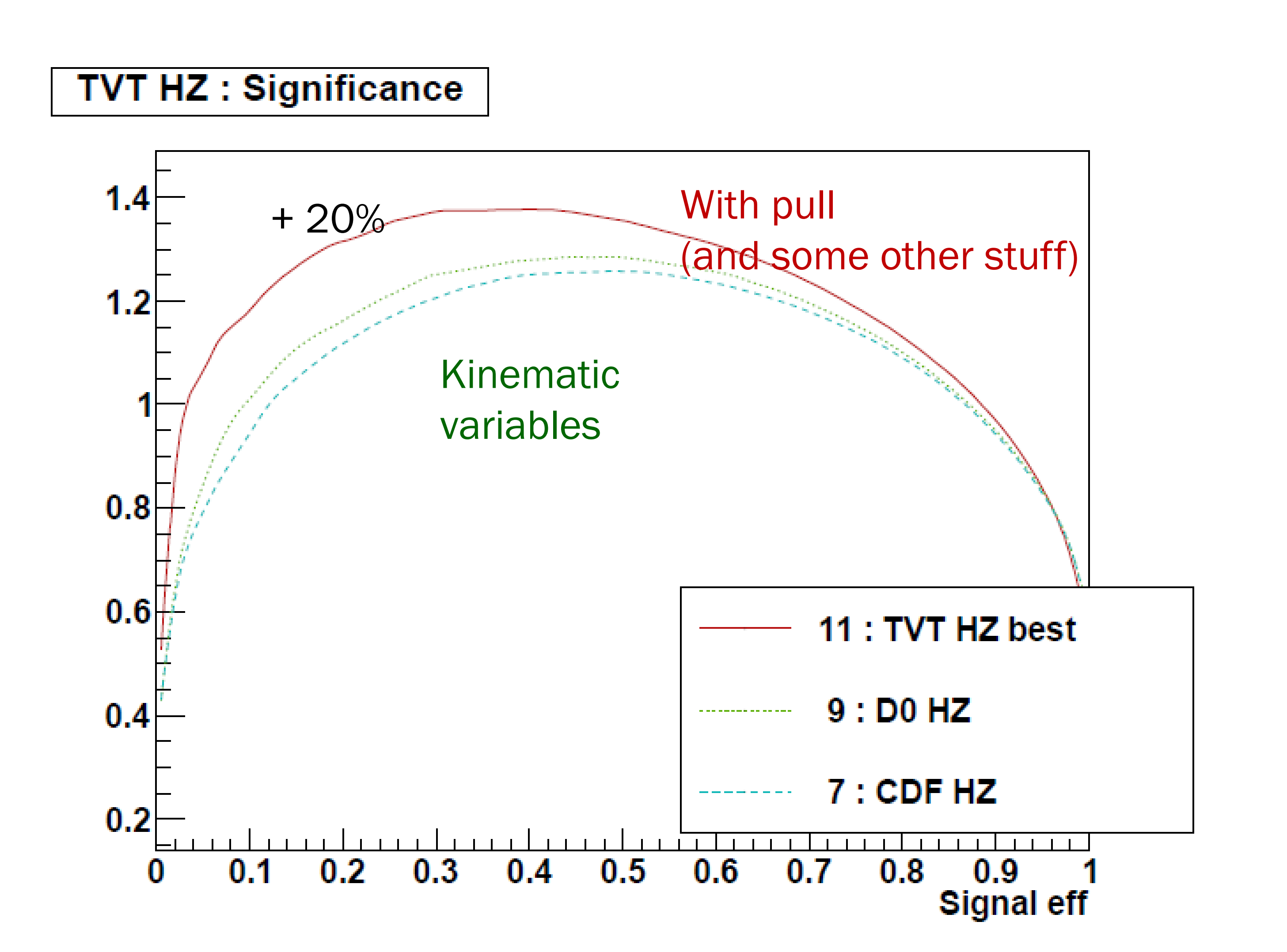
$$\sigma \equiv \frac{S}{\sqrt{B}} \quad \xrightarrow{\text{cut}} \quad \frac{\varepsilon_S S}{\sqrt{\varepsilon_B B}} = \left(\frac{\varepsilon_S}{\sqrt{\varepsilon_B}}\right)\sigma$$



$$\text{SIC} \equiv \frac{\varepsilon_S}{\sqrt{\varepsilon_B}}$$

- Nice **visualization**
- Has maximum at **interesting** place
- Well defined way to **compare variables**

ADDING PULL HELPS

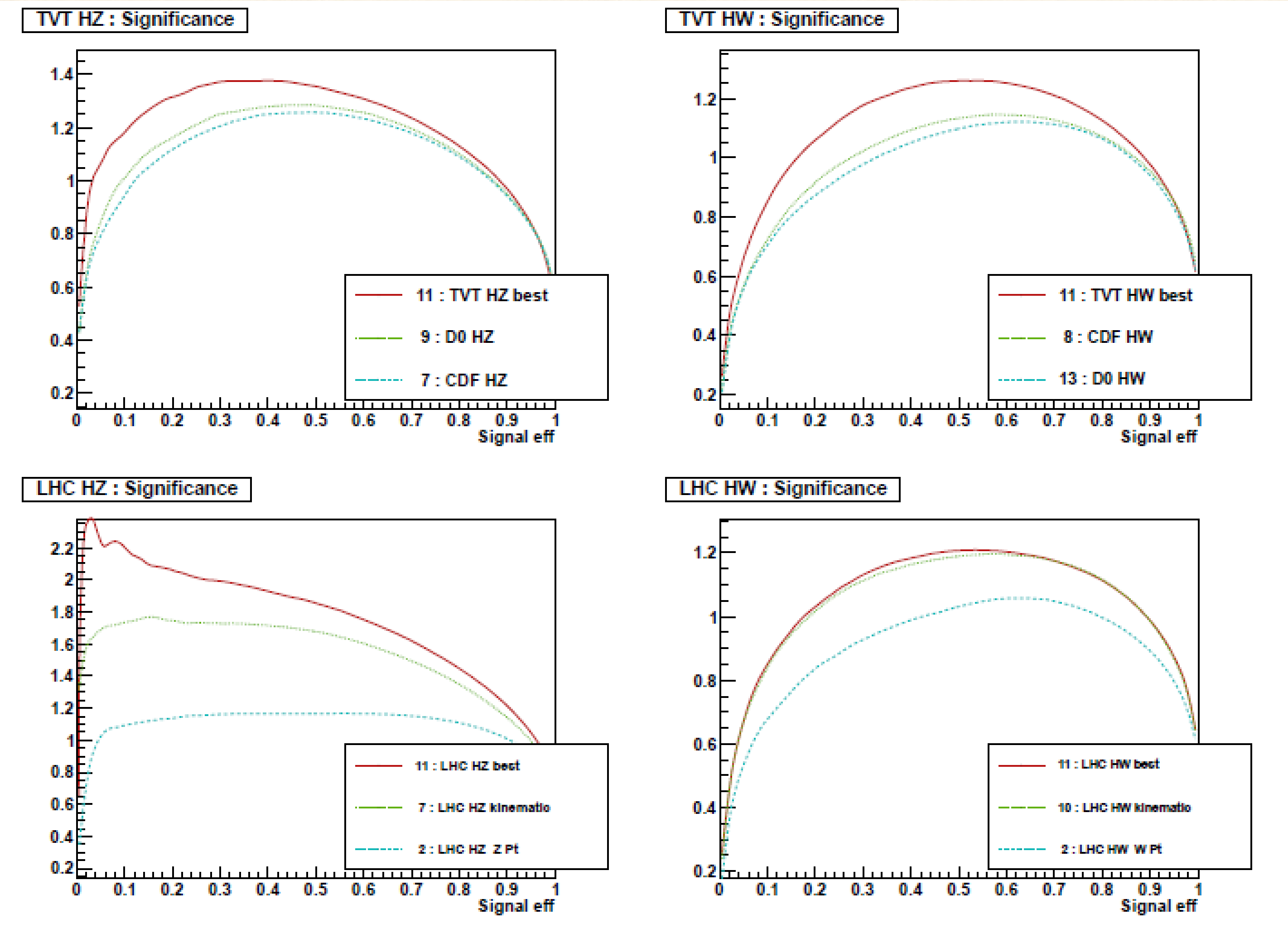


+ 20%

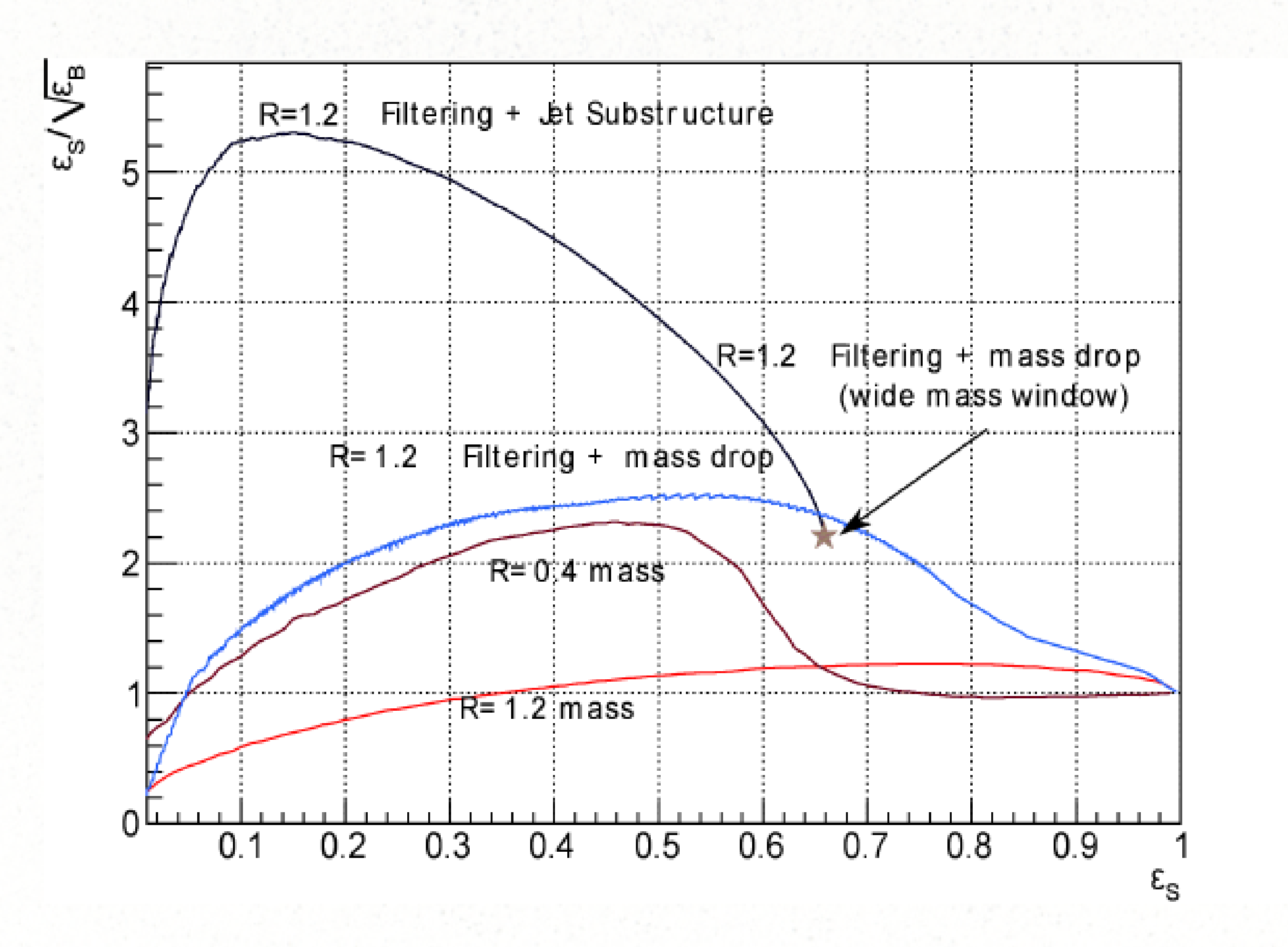
With pull
(and some other stuff)

Kinematic
variables

ALSO AT THE LHC



OPTIMIZE W TAGGING



Huge improvement in significance with multivariate approach

CONCLUSIONS

New machine needs new tricks

- Jet substructure extremely useful
  - Jet mass, Jet shapes, R-cores, Splitting scales
  - Filtering, Trimming, Pruning

- Complimentary information in superstructure
  - Sensitive to other jets – global information
  - Measures color flow

- Correlations are subtle
  - Multivariate techniques are essential
  - Boosted Decision Trees work well if used carefully
  - Proper visualization makes comparisons much easier
    - e.g. SIC curves

$$\mathrm{SIC} \equiv \frac{\varepsilon_S}{\sqrt{\varepsilon_B}}$$