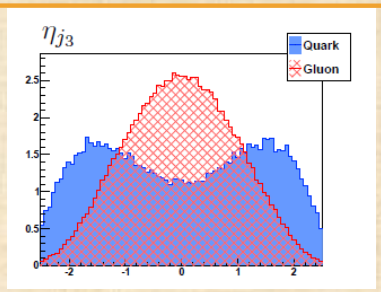
Multivariate Overview

Matthew Schwartz

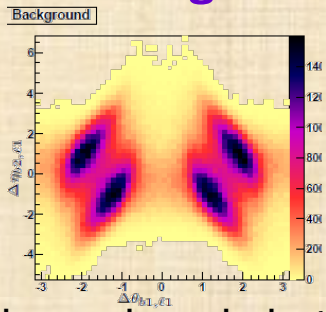
Harvard University

March 24, Boost 2011

WHY USE A MULTIVARIATE APPROACH?



•We can think about and visualize **single variables**



• Two variables are harder

• Nobody who thinks in 11 dimensions is in this room!

• There things that **computers are just better** at.

• Multivariate approach lets you figure out how well you could **possibly do**

EFFICIENCY

Save you the trouble or looking for good variables (project killer)

FRAMING

See if simple variables can do as well (establishes the goal)

POWER

Sometimes they are really necessary (**e.g. ZH**)

MULTIVARIATE (MVA) BASICS

Lots of methods (all in the TMVA package for root)

- Boosted Decision Trees
- Artificial Neural Networks
- Fischer Discriminants
- Rectangular cut optimization
- Projective Likelihood Estimator
- H-matrix discriminant
- Predictive learning/Rule ensemble
- Support Vector Machines
- K-nearest neighbor
- …

Useful in many areas of science, such as artificial intelligence

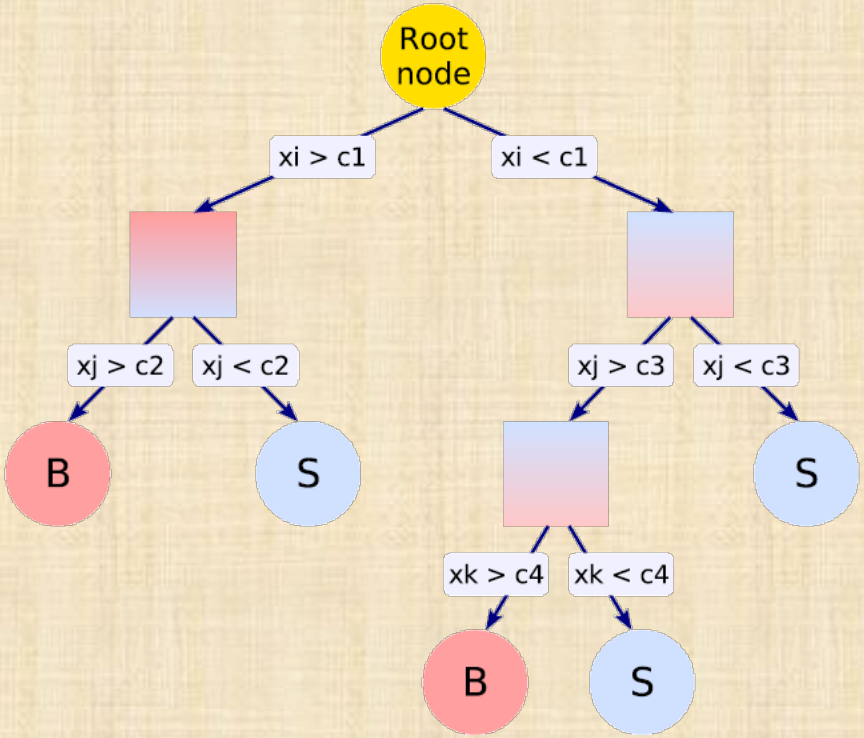
For particle physics, **Boosted Decision Trees** work best

Easy to understand

Train fast

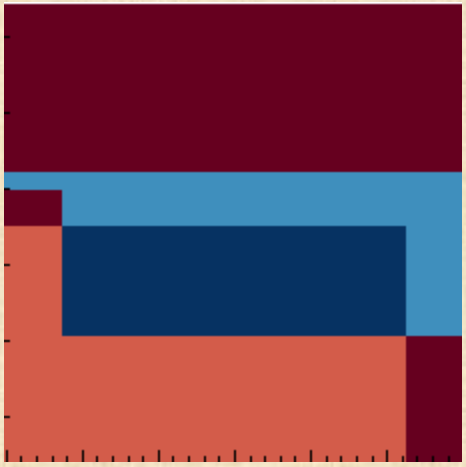
Excellent efficiencies

BOOSTED DECISION TREES



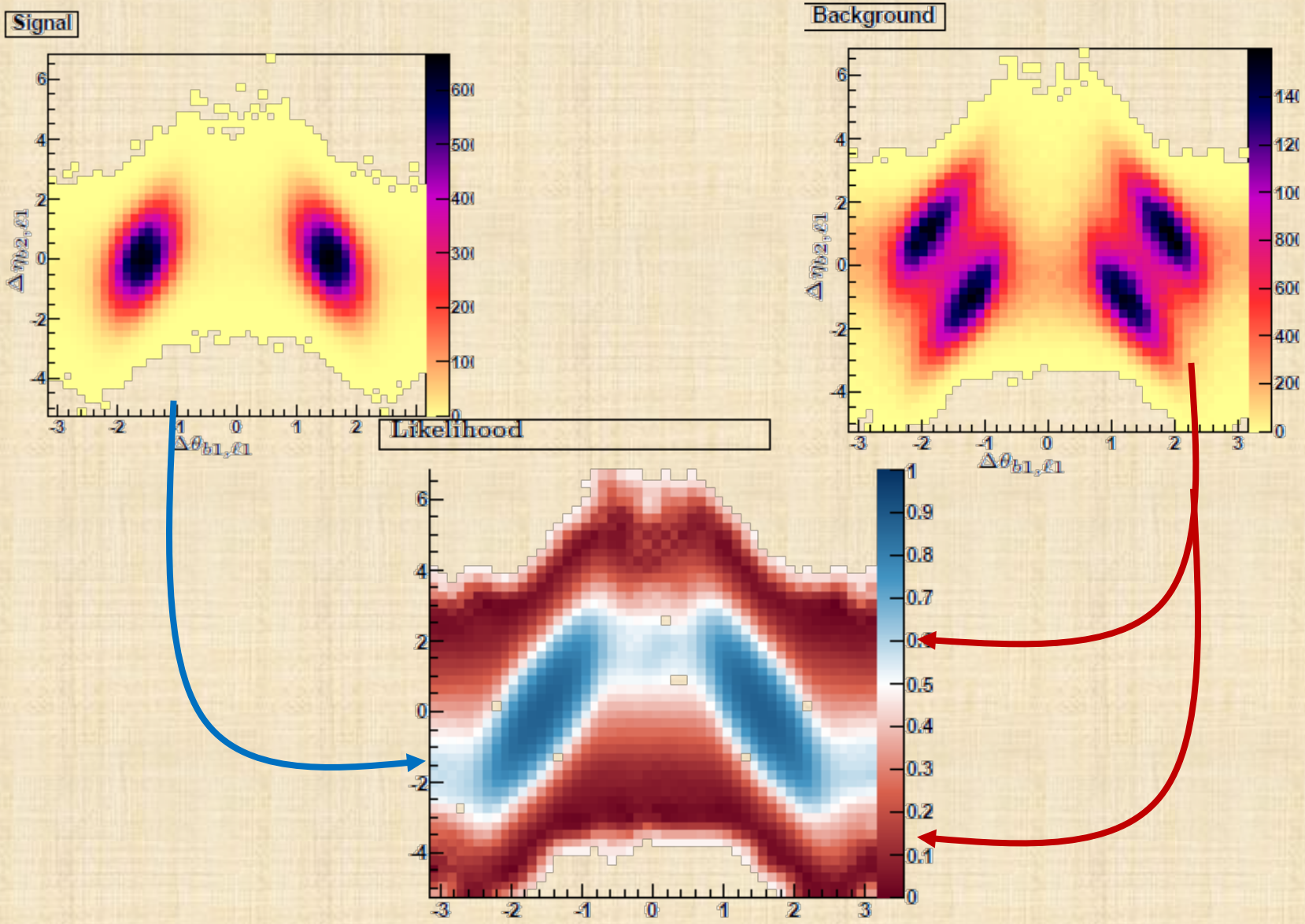
One decision tree

Two decision trees

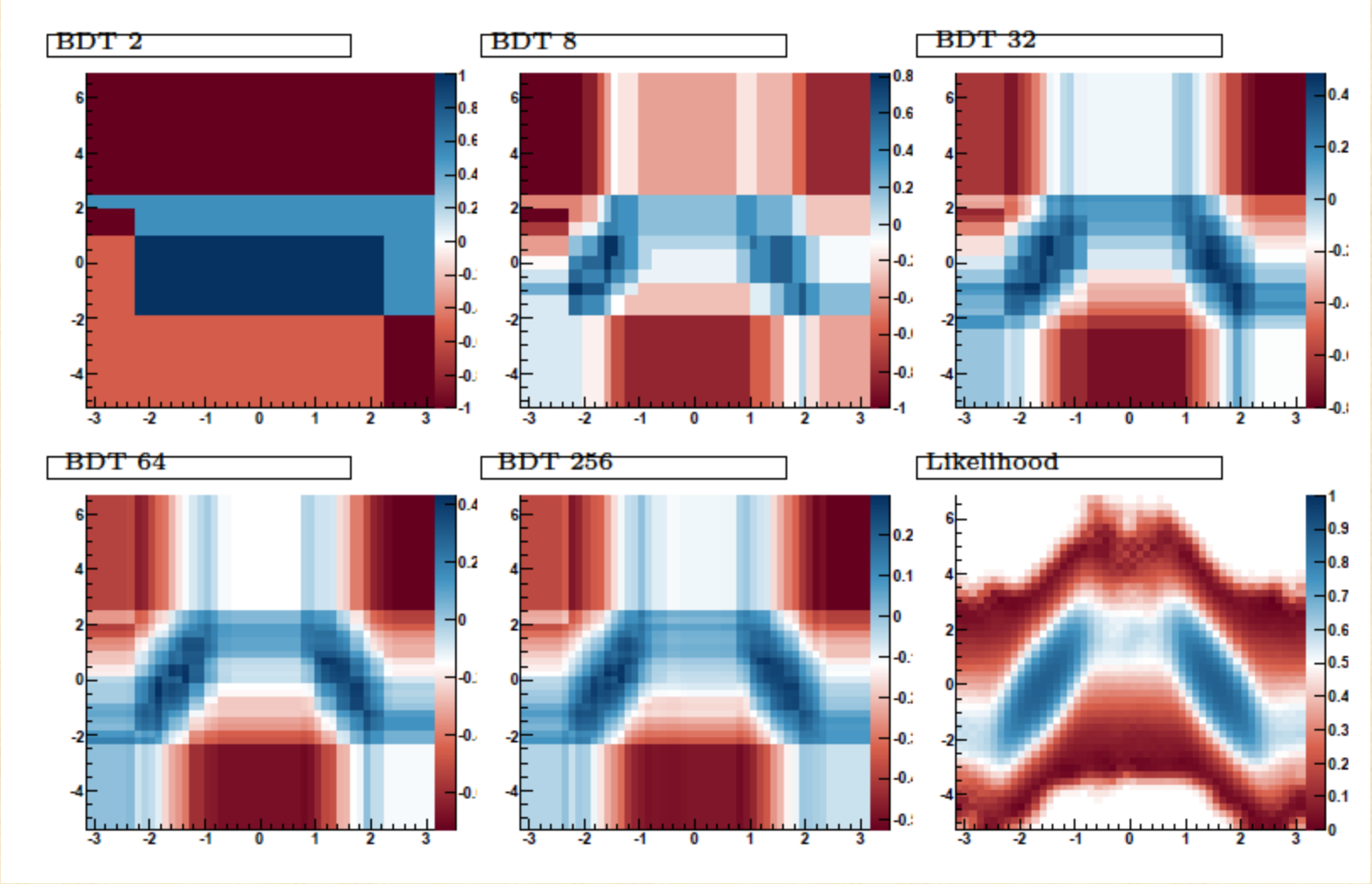


- **Boosting** : train successive trees on misclassified events by enhancing their importance

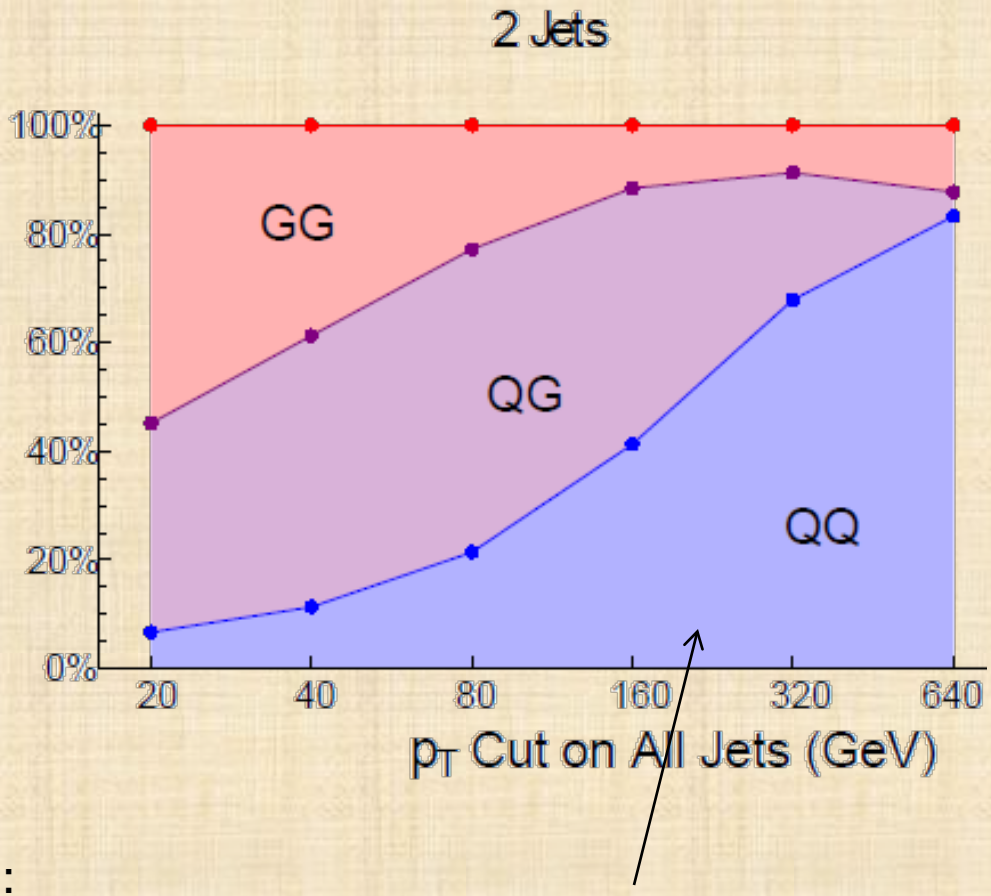
EXACT SOLUTION: LIKELIHOOD



MULTIDIMENSIONS: APPROXIMATE



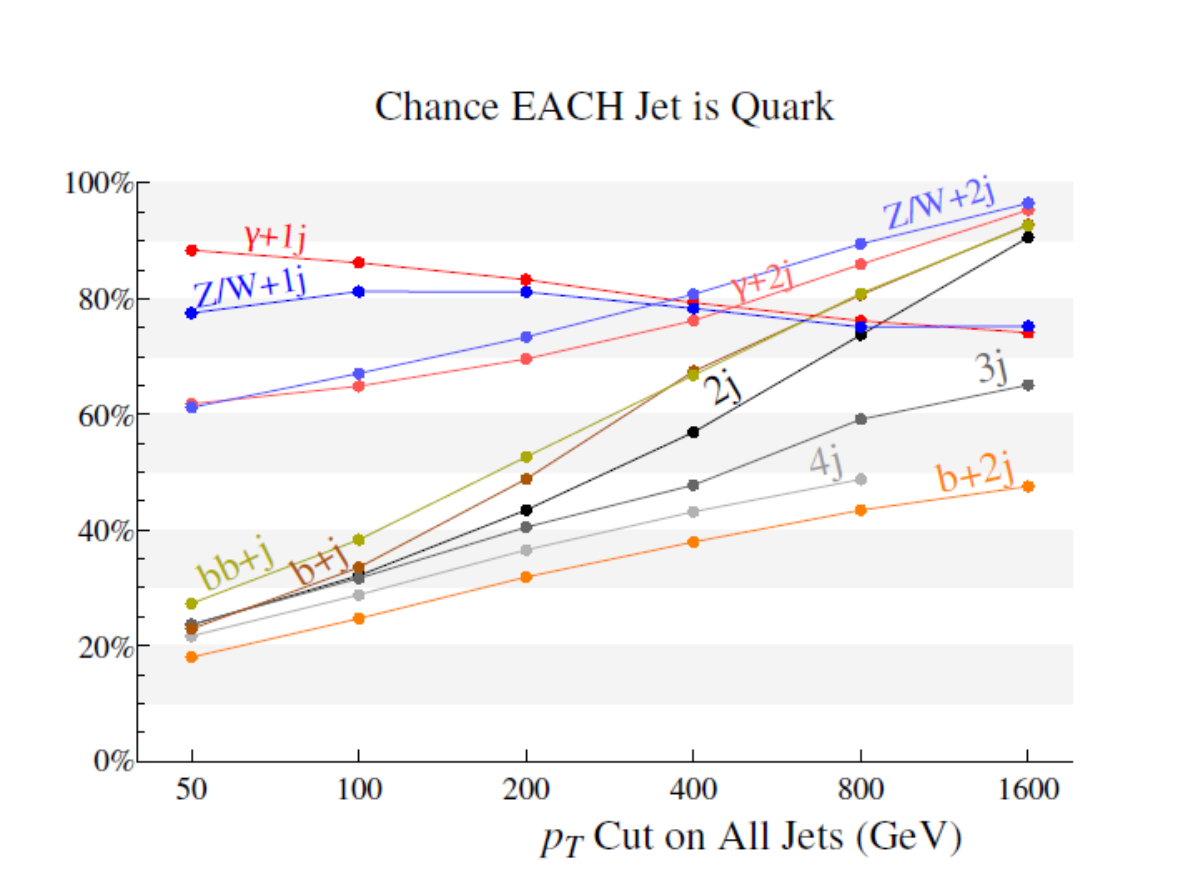
EXAMPLE: WHERE ARE THE QUARK JETS?



(Aside for Pekka:

Maybe one reason jet masses more correlated at high pT)

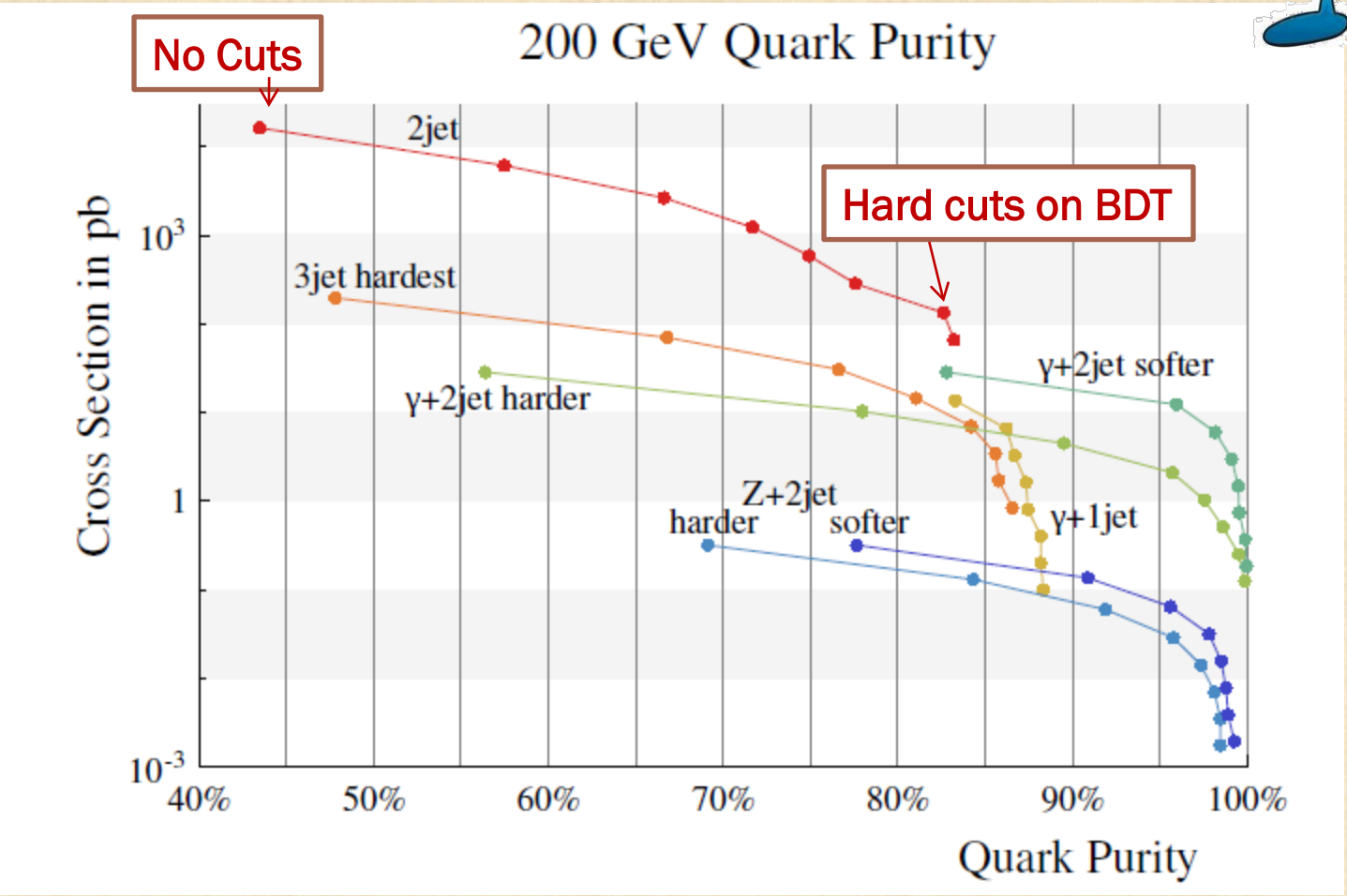
LOOK AT ALL SAMPLES



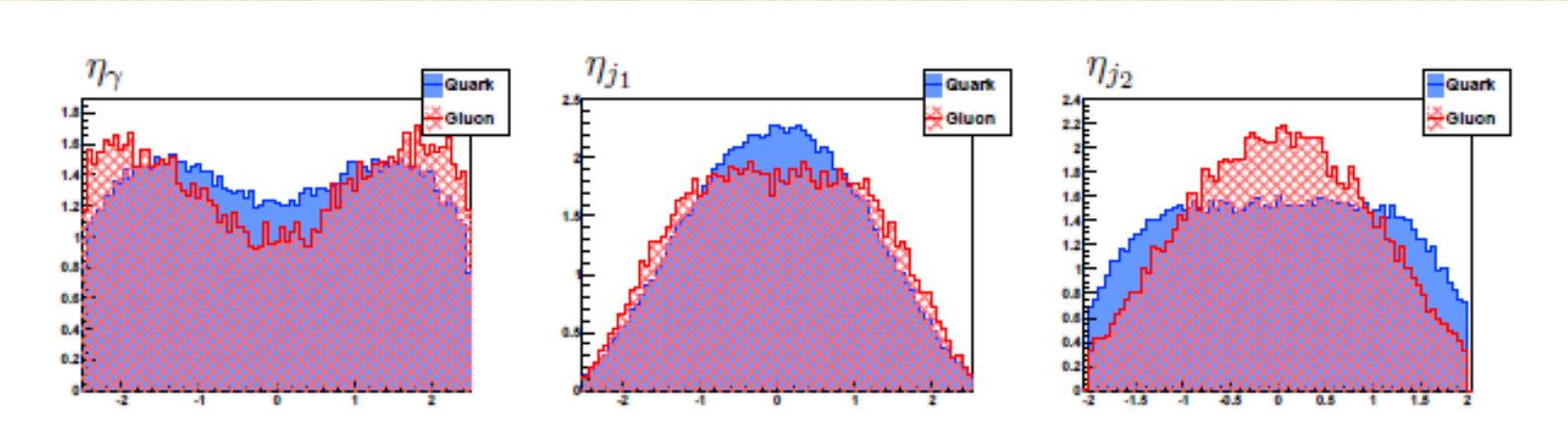
- What about cross sections?
- Can cuts purify the samples?

THROW THEM INTO THE BDT

Optimize efficiency using BDT classifier with parton momenta as inputs (6 or 9 inputs)

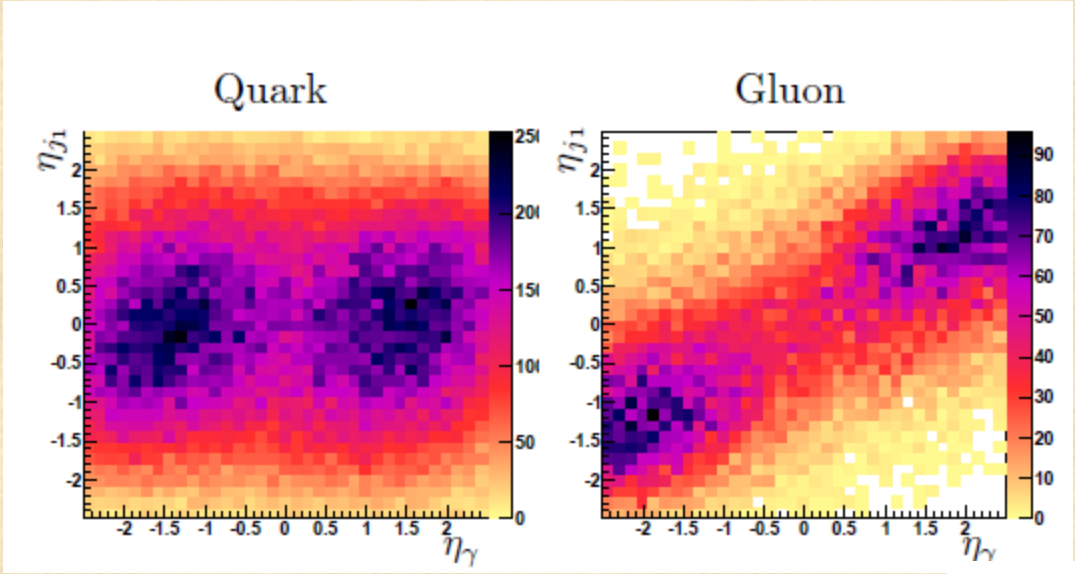


NOW LOOK AT THE γ + 2 JETS SAMPLE

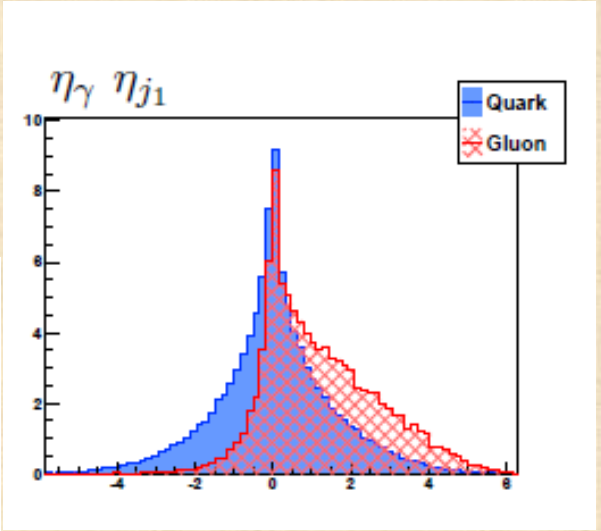


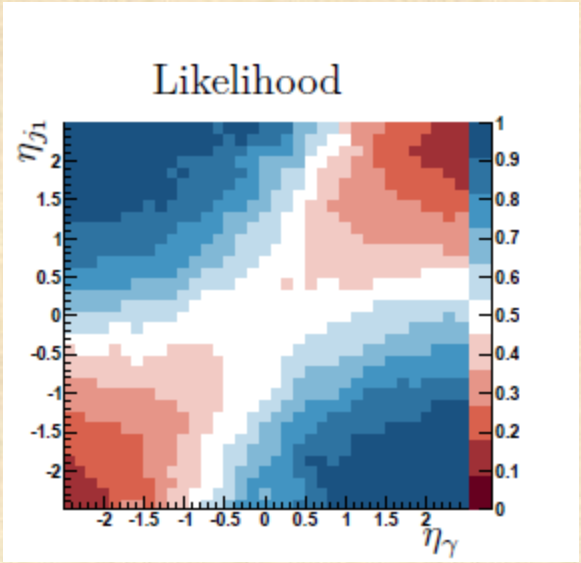
- Look at the best discriminants, ranked by cuts
- The rapidity of the photon and the rapidity of the second hardest jet look good
- But cutting on just η_γ or just η_{J2} does not help much

LOOK AT CORRELATIONS

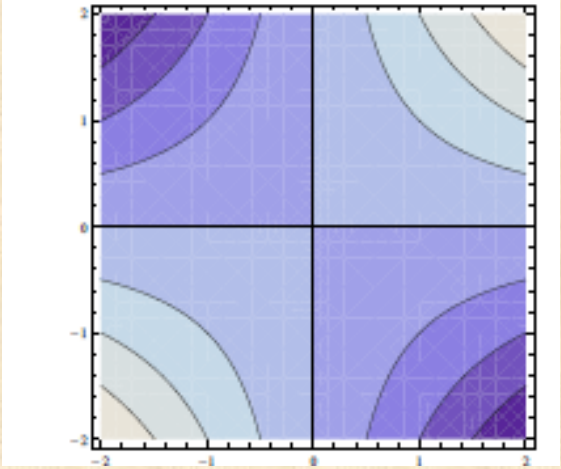


Distribution of $\eta_\gamma \eta_{j1}$



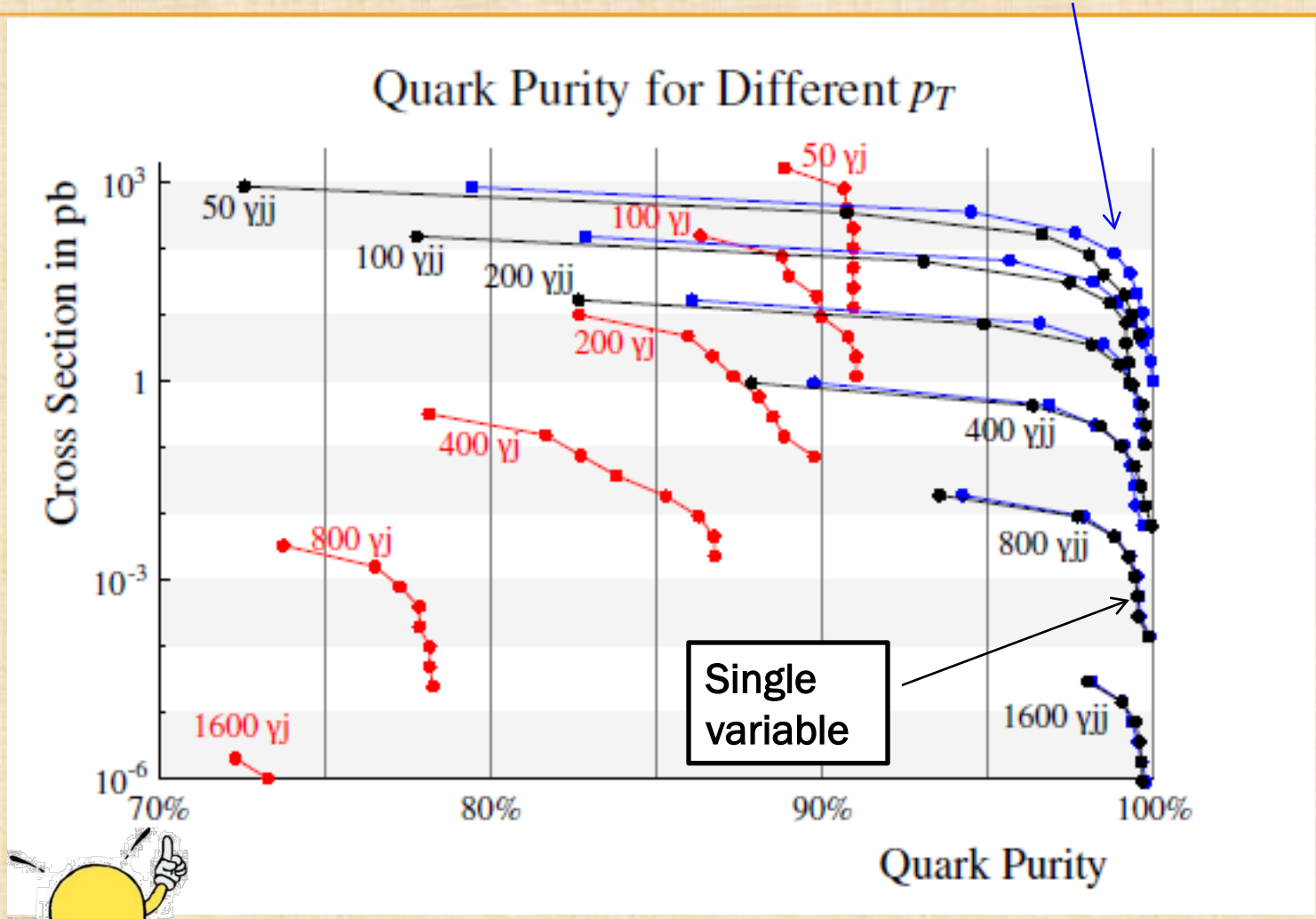


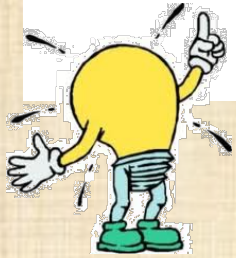
Contours of $\eta_\gamma \eta_{j1}$



BEST SINGLE VARIABLE

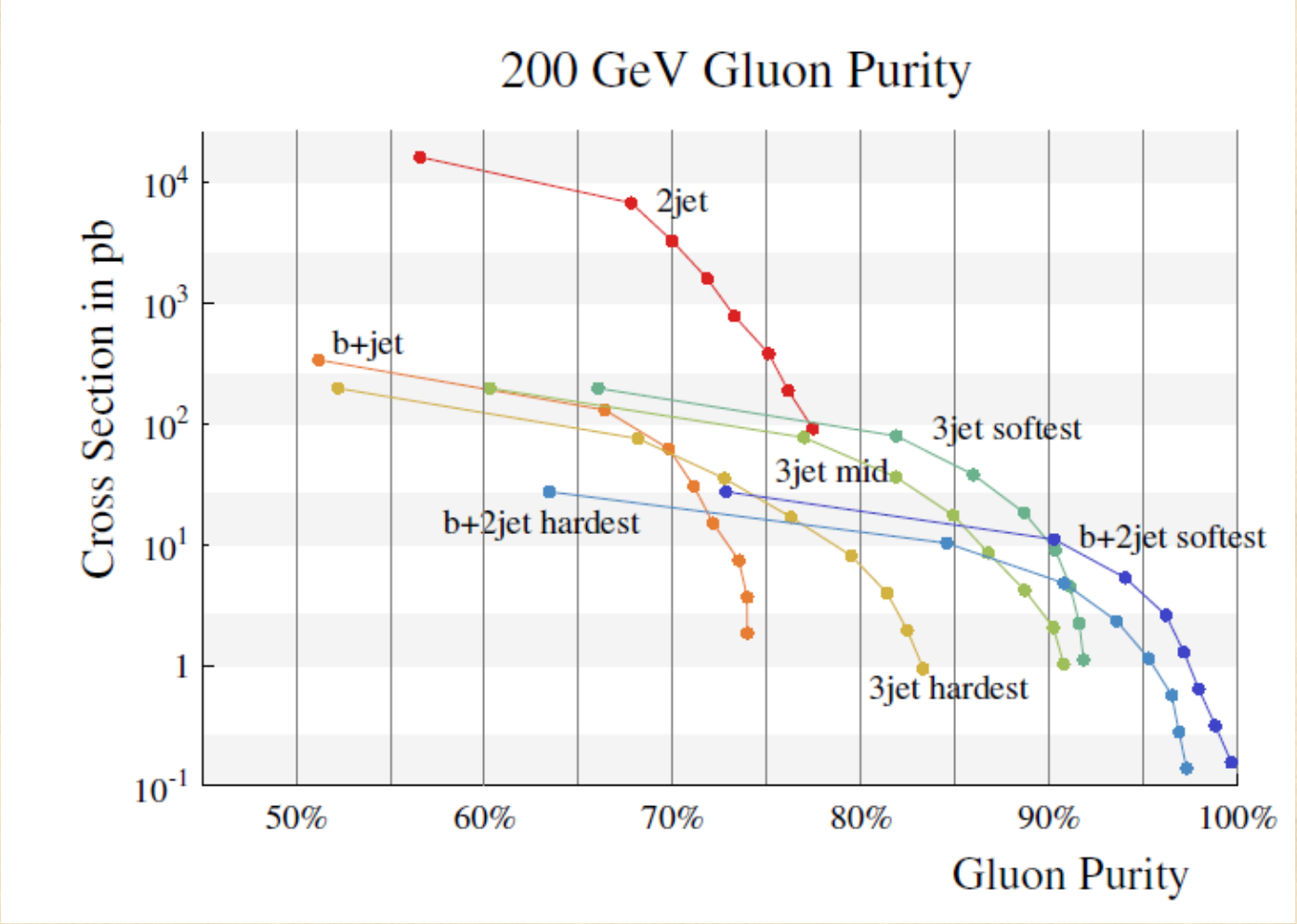
BDT results





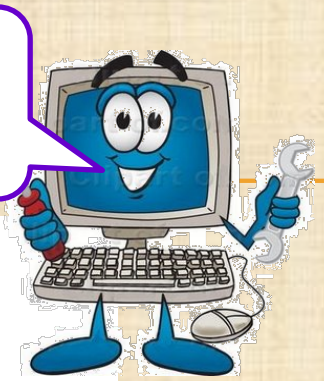
BDTs led us to the variable,
but with the variable we **don't need BDTs**

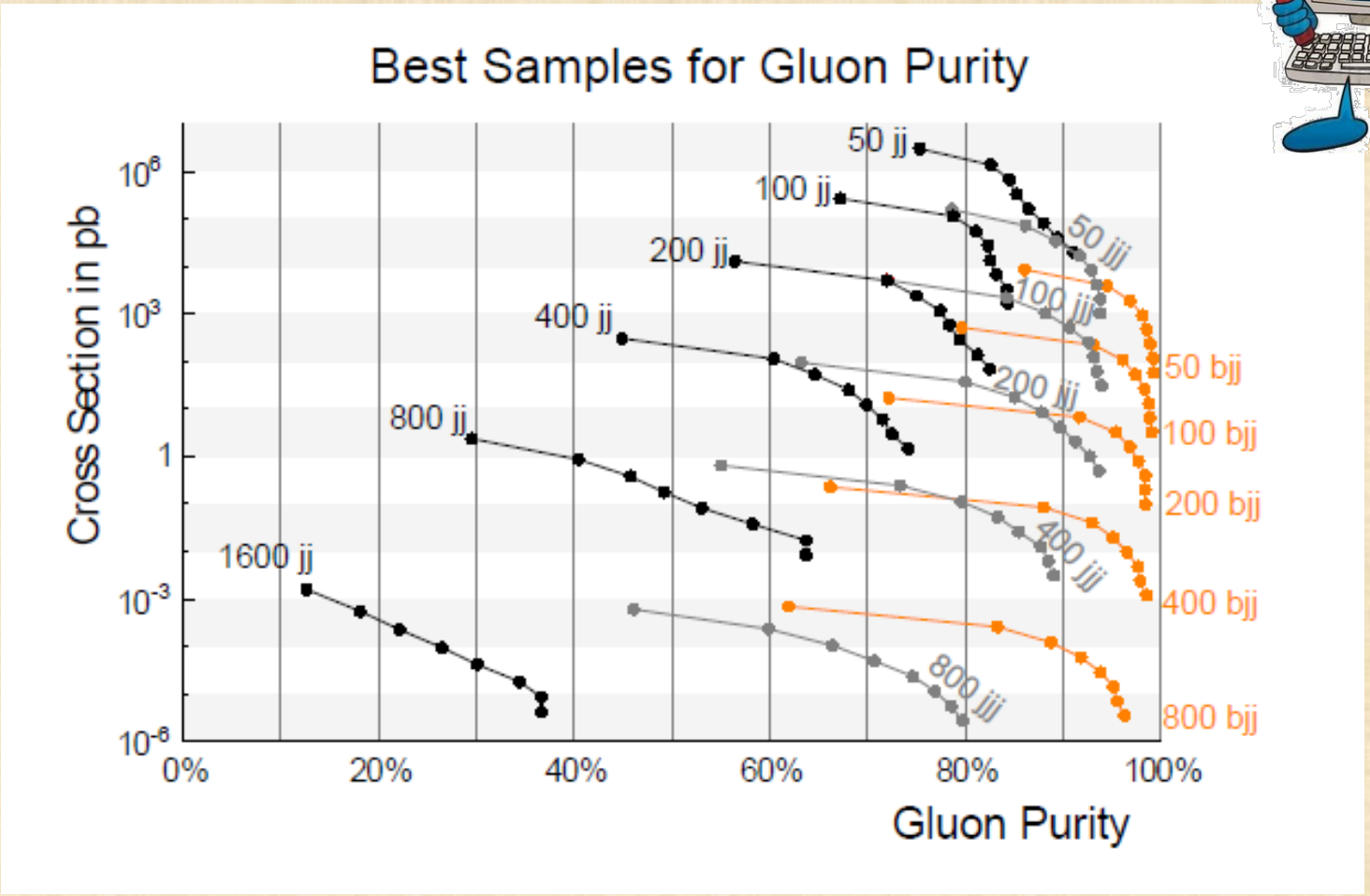
WHAT ABOUT PURE GLUONS?



b+2 jets or trijets look promising

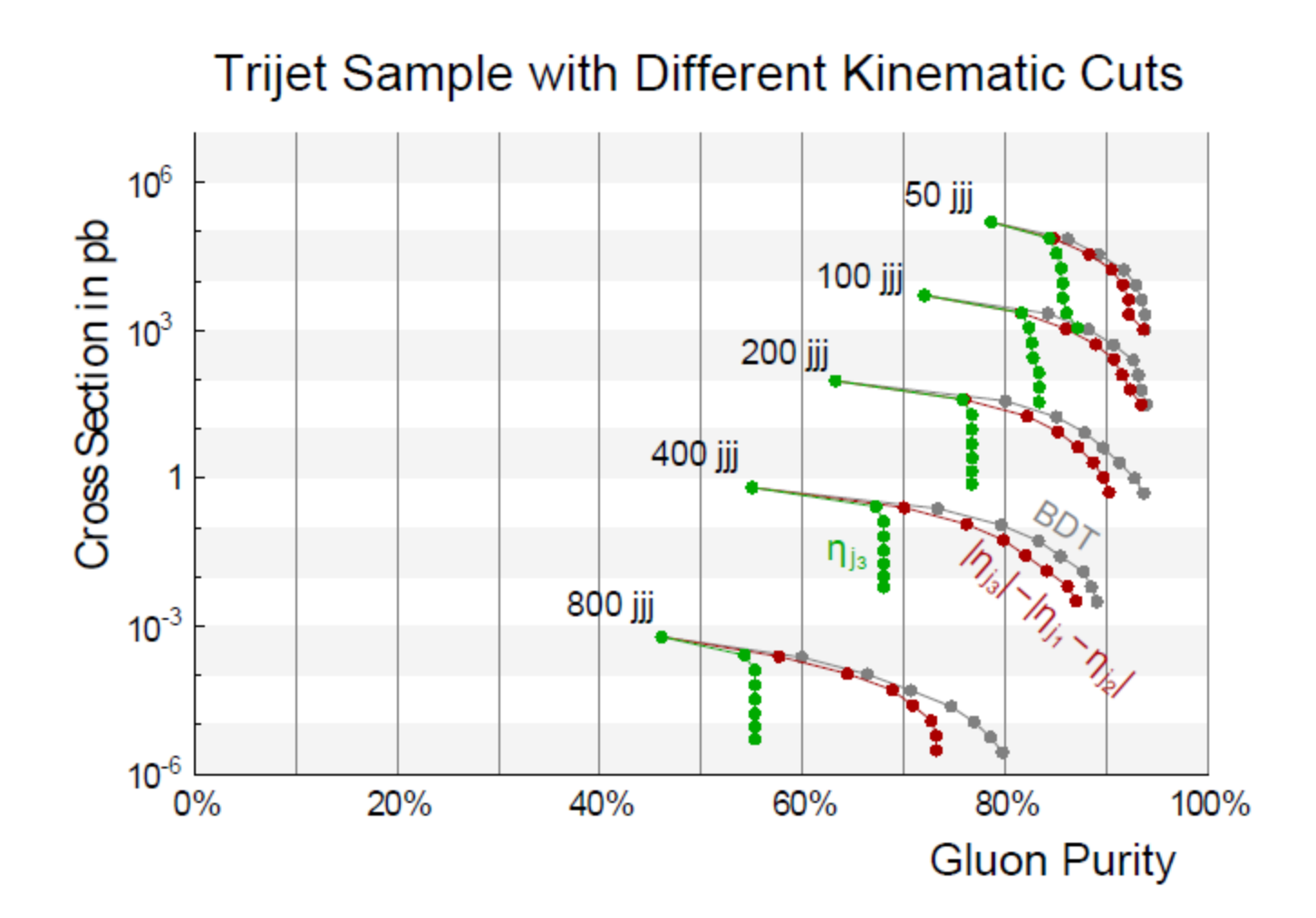
THROW IT AT THE BDT





- Now try to find a single variable that works as well…

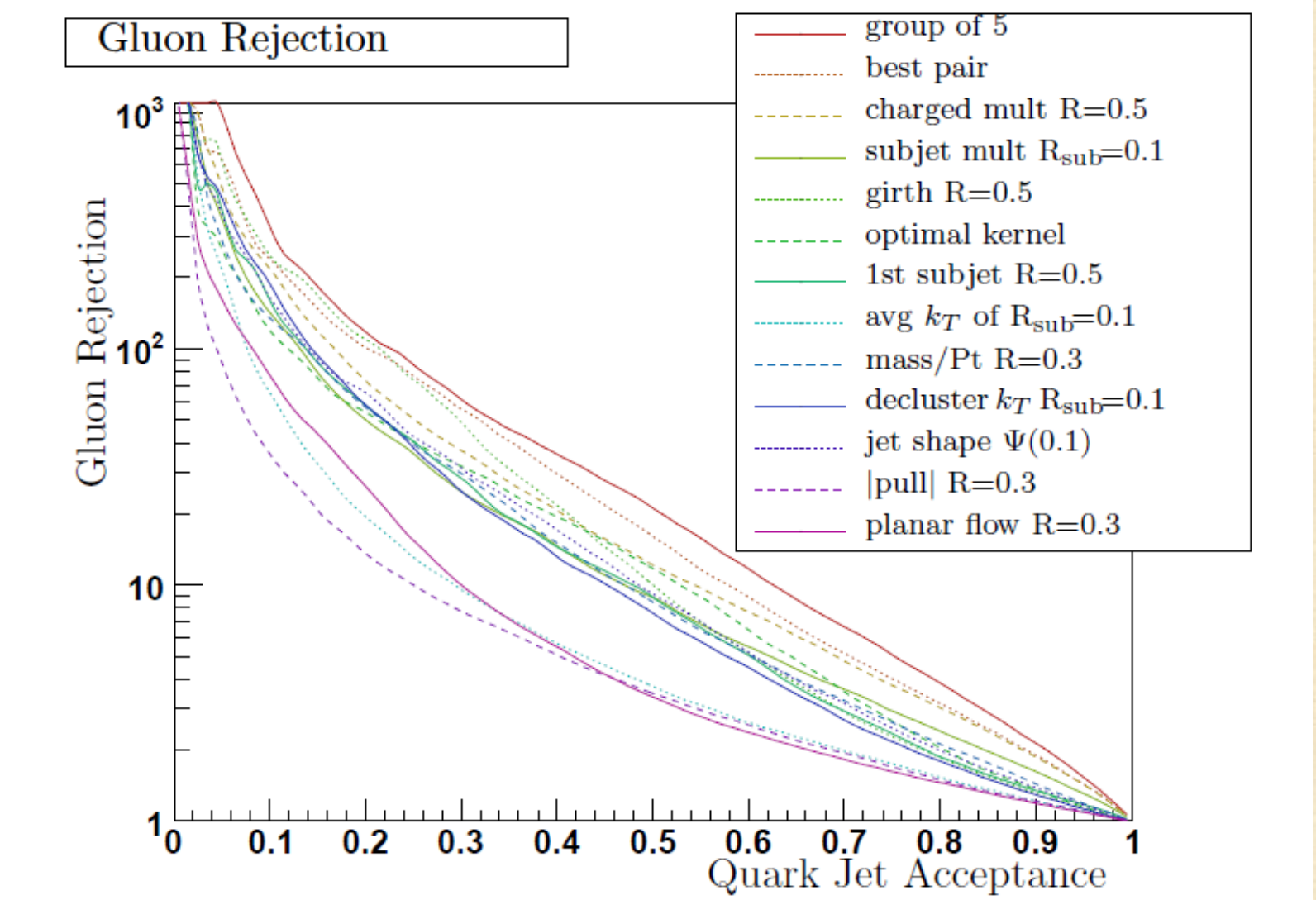
FINDING PURE GLUON JETS



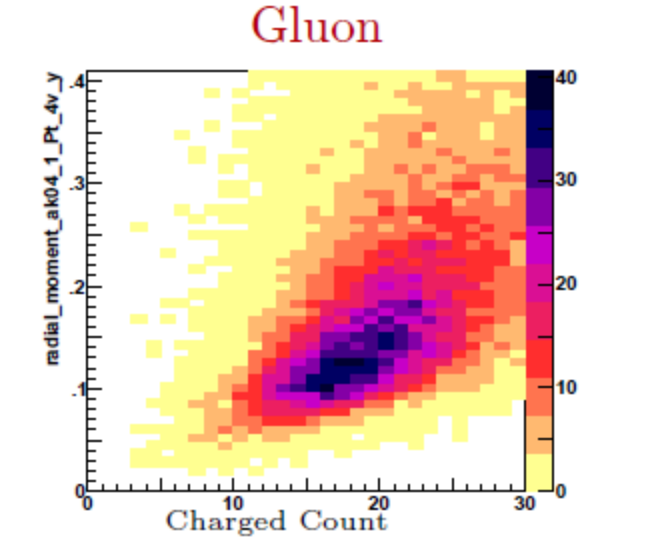
SUMMARY OF FINDING QUARKS/GLUONS

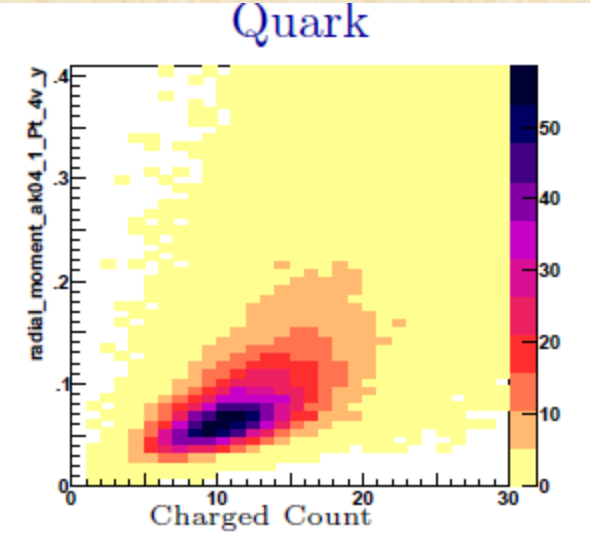
- For quarks, look at gamma + jet
  - cut on $\eta_\gamma \eta_{j1} + \Delta R_{\gamma j_2}$
- For gluons
  - Look at b+2 jets
  - look at trijets
    - Cut on $|\eta_{j3}| - |\eta_{j1} - \eta_{j2}|$

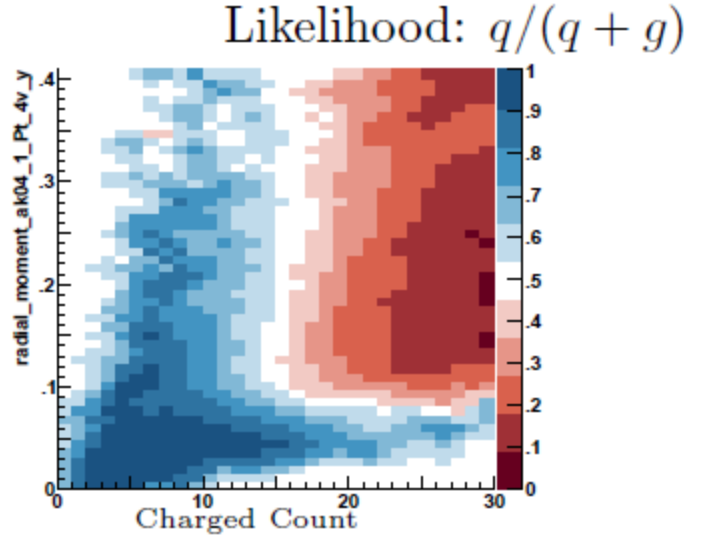
HOW TO TELL QUARKS FROM GLUONS



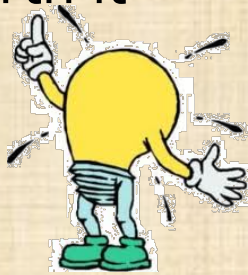
BEST PAIR WORKS PRETTY GOOD







- Can get 50% quarks and 4% gluons
- Need one "count" variable
  and one "moment" variable
- Beyond that, probably not worth it

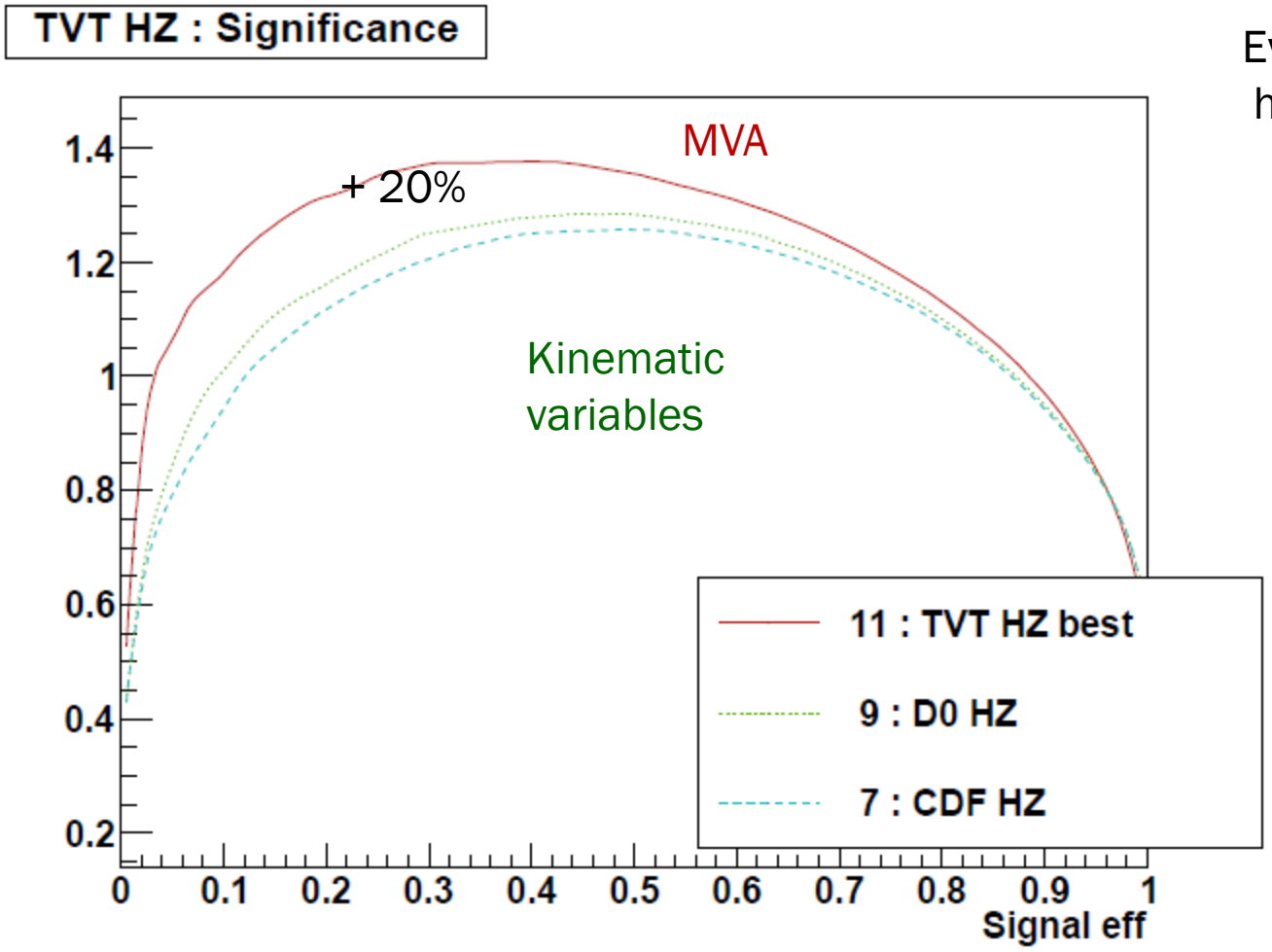


Sometimes one variable is good enough

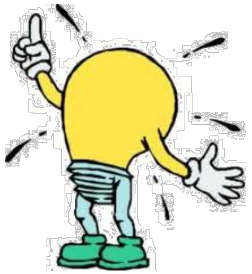
but

sometimes many variables are needed

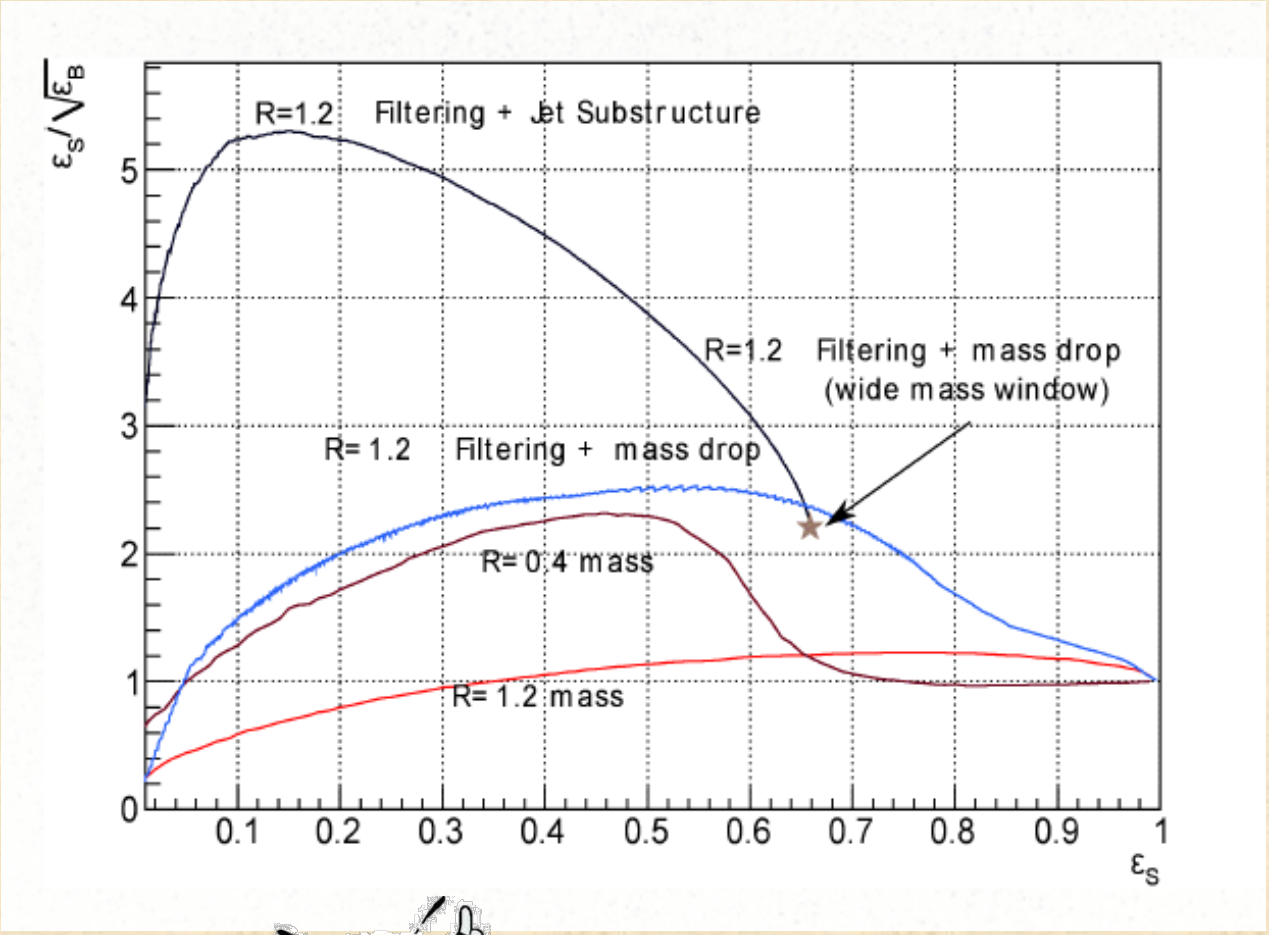
MVA POWERFUL FOR HIGGS SEARCH



Every little bit helps

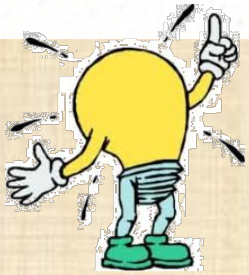


OPTIMIZE W TAGGING



<u>7 MC insensitive variables</u>

1. Jet mass with R=0.5
2. Jet mass with R=0.4
3. Filtered Jet mass
4. Mass of hardest subjet, from filtering
5. Mass of second hardest subjet, from filtering
6. Ratio of the pT's of the two hardest subjets
7. Planar Flow



This is the **ultimate goal** of W-tagging

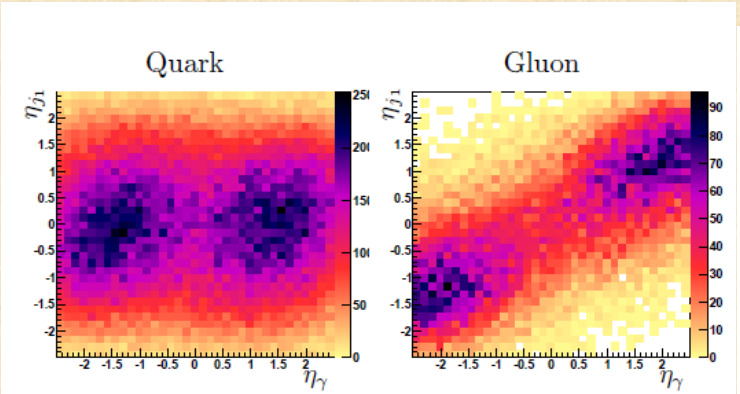
Challenge: can fewer variables do as well?

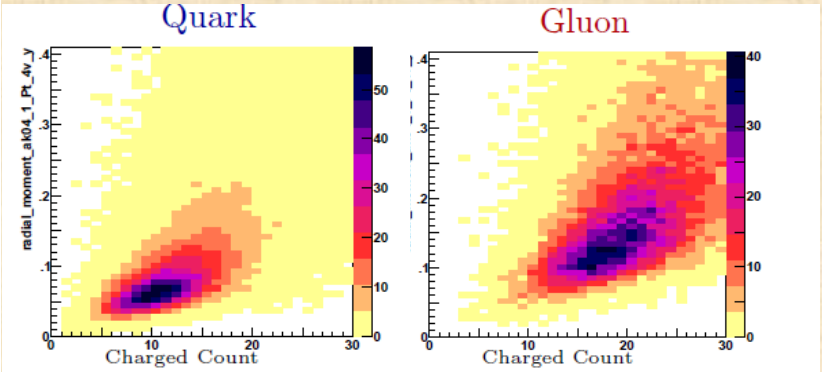
GENERAL OBSERVATIONS

- **For matrix-element level (fixed # dof)**
  - BDT much **easier to implement** than matrix element method
  - MVA kinematic discriminants can be **similar to cutting on one** smart variable (e.g. in finding pure samples of Quark and Gluon jets)

- **For showered samples (W-tagging, HZ)**
  - **Subtle correlations** make **thinking difficult**
  - Some discriminants (like pull) **don't work on their own** but work well as the 5th or 6th variable added to a BDT
  - Sometimes **few variables are enough** (e.g. Quark vs Glue)
  - Sometimes **many variables are needed** (e.g. W-tagging)

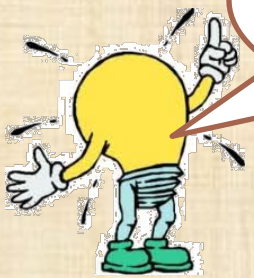
- **We desperately need data on 2d correlations!!**





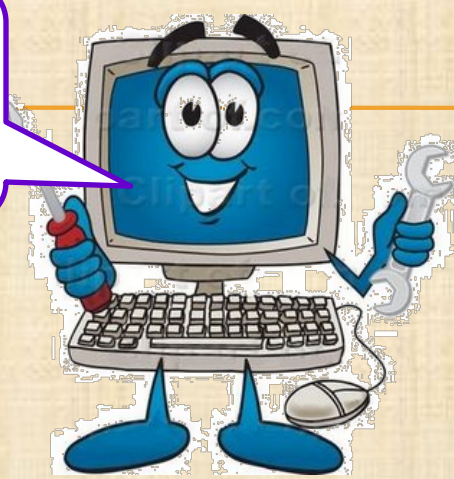
Most useful with **pure sampes** (quark, glue, hadronic-W, hadronic-top)

CONCLUSIONS



Boosted Decision Trees are here to **help** you

Let me do the work for you!



Multivariate analysis quickly tells you how well you could **possibly do**

FRAMING

See if simple variables can do as well

POWER

Sometimes MVA is really necessary

EFFICIENCY

Saves you the trouble or looking for good variables

- Need to have **correlations in real data**
  - If 2D distributions agree with the Monte Carlo, the BDTs can be trusted
  - If they don't agree, we have a great opportunity to improve the MCs