PURDUE
UNIVERSITY
CALUMET

b-tagging at CMS

Pratima Jindal
Purdue University Calumet
(on behalf of the LPC b-Id group)

JTerm IV
August 3rd 2009

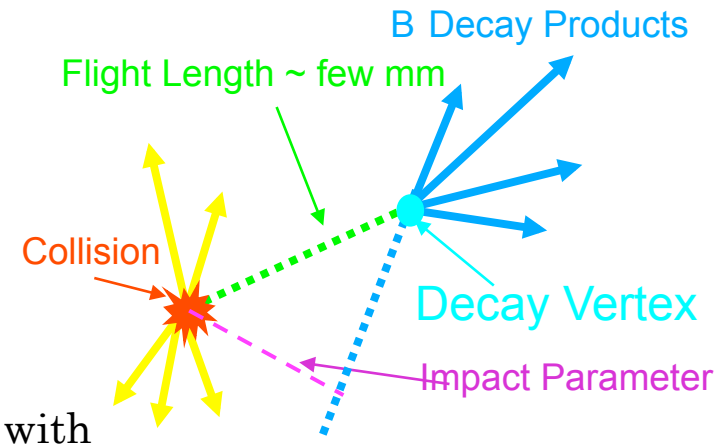
Outline

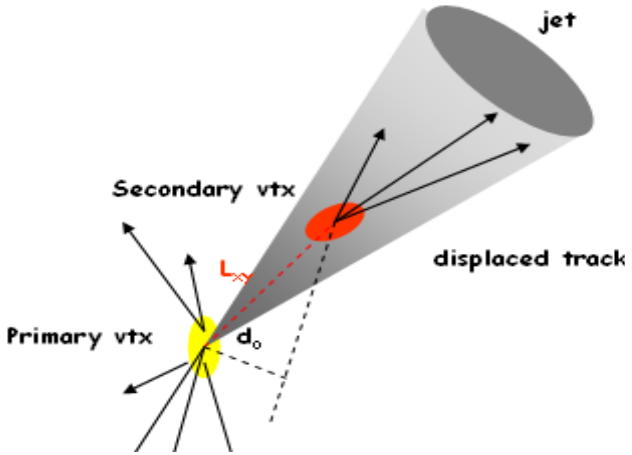
PURDUE UNIVERSITY CALUMET

- Identifying *b*-jets
- *b*-tagging Algorithms
  - Impact Parameter based
  - Secondary Vertex based
  - Soft Lepton based
- Misalignment Scenarios
- Performance Measurements Methods
  - pTrel Method
  - System8 Method
  - Mistag rate measurement using negative tags
  - Top quark based: Likelihood ratio method
- Scale Factors and Tag Rates
- Conclusion

Identifying *b*-Jets

- The CMS collaboration has implemented several algorithms to discriminate jets coming from the hadronization of *b* quarks from those from lighter quarks
- Exploit the properties of *b* hadrons to distinguish *b*-jets from light (*u,d,s,g*) jets:
  - large lifetime ~1.5 ps (large decay length: 20 GeV *B*-hadron decays after ~2 mm)
    - search for tracks or vertexes displaced w.r.t. primary vertex
  - large mass ~5 GeV
    - search for leptons, from semileptonic B decays, with large transverse momentum w.r.t. jet axis
- Tagged *b*-jets will be used to identify top quarks and in searches of the Higgs boson and other non Standard Model searches.
- A reliable estimate of the performance of these algorithms is therefore crucial, and methods to estimate efficiencies and mistag rates directly from data are needed.
- CMS has prepared several strategies to extract efficiencies and rejection rates from data, which should work even on the first data (10 pb^{-1}).





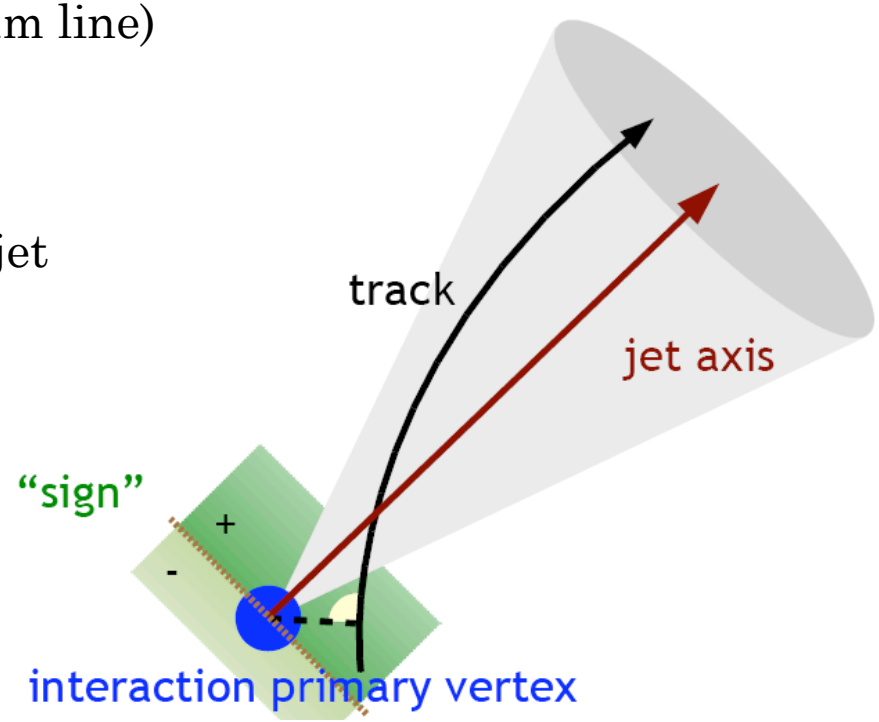
b-tagging Algorithms

- Impact Parameter based:
  - Track counting High Eff/Pur (2)
  - Jet[B]Probability (2)
- Secondary Vertex based
  - Simple Secondary Vertex
  - Combined Secondary Vertex
- Leptons:
  - Soft mu by IP 3d
  - Soft mu by Ptrel
  - Soft electron
- Combined
  - Combined MVA

All the performance plots shown in the following slides are for Summer08 QCDPt80 samples

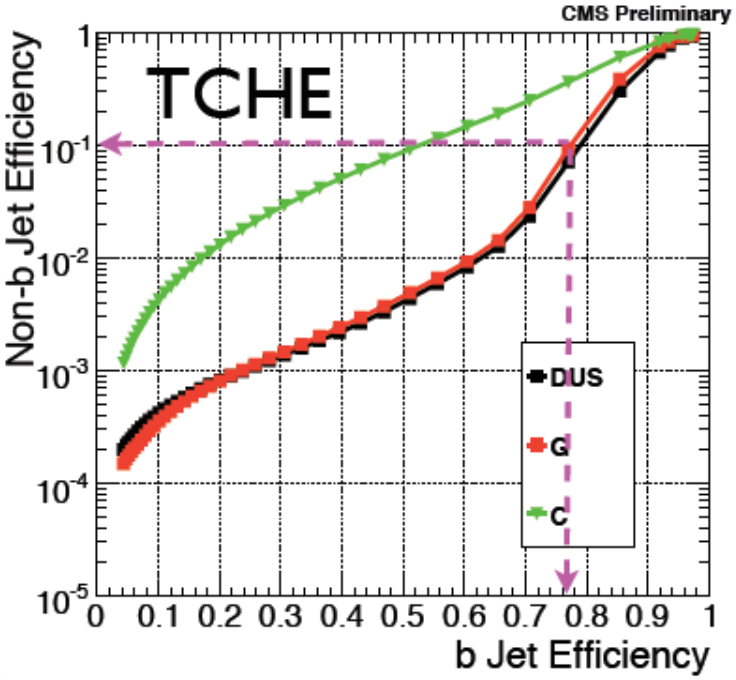
Impact Parameter

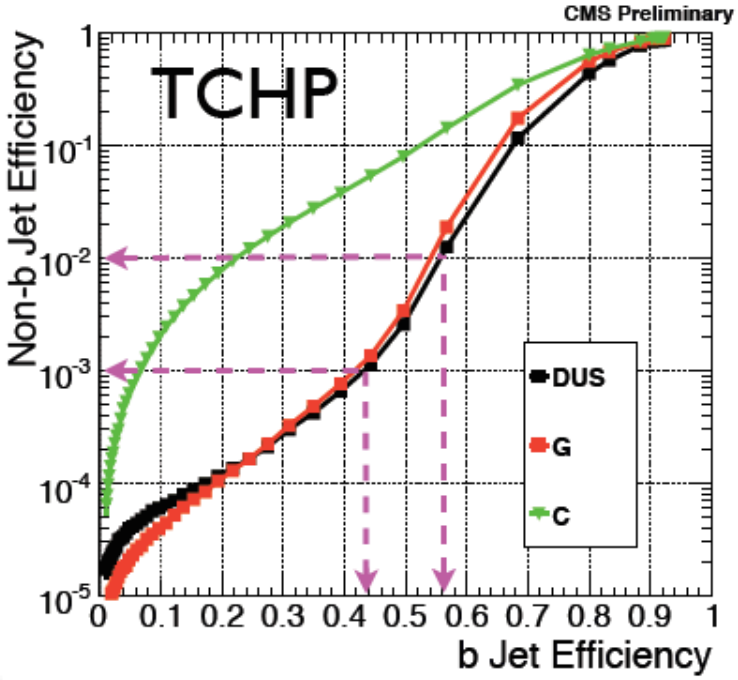
- Large B hadron lifetime → large impact parameter, d_0, of B decay products
- Search for tracks displaced w.r.t. primary vertex
- Use either 2D (transverse plane to beam line) or 3D impact parameter
- Use "sign" of the impact parameter:
  - positive if track intersection with jet axis is along jet direction
- Impact parameter significance, d_0/sd_0 is used as discriminant between signal (true b-jets) and background (fake b-jets)



Track Counting

- Identify a jet as a "b-jet" if there are at least N tracks each with a 3D significance of the impact parameter exceeding S
  - if one is interested in a high efficiency for b-jets, the second track can be used
  - for higher purity selections the third track is a better choice

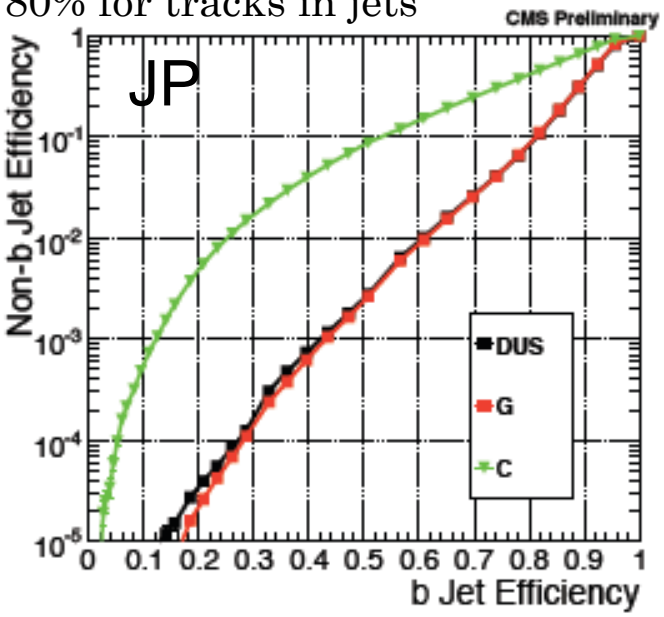


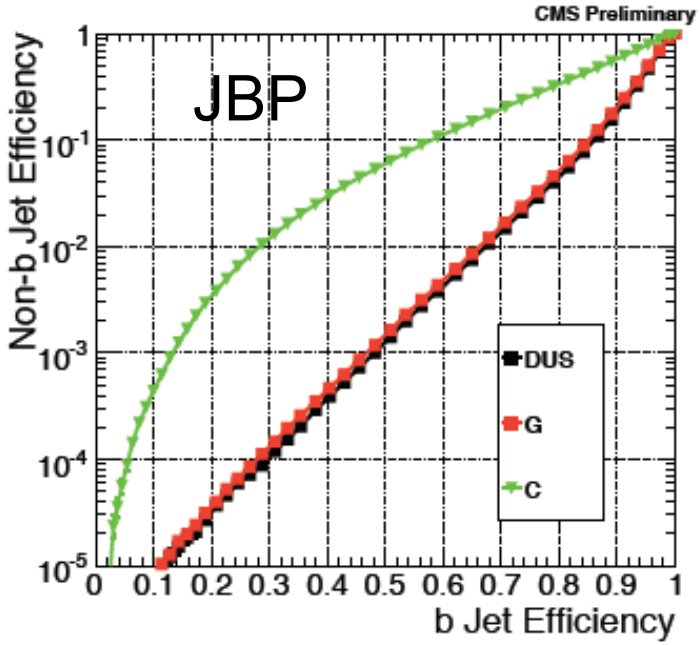


Mistag rate versus efficiency for the "track counting high efficieny" (left) and "track counting high purity" (right) taggers

Jet [B] Probability

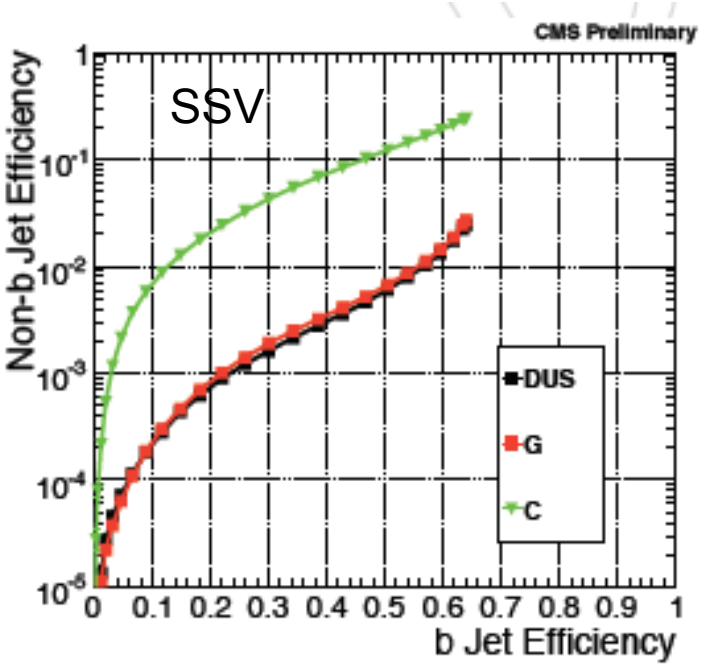
- Jet Probability
  - Determine the probability that each track in jet comes from primary vertex
  - Estimate a combined probability that all tracks in jet come from PV
  - Use combined probability as discriminant
- Jet B Probability
  - Estimate how likely it is that the four most displaced tracks are compatible with the primary vertex
  - The selection 4 comes from the fact that the average charged track multiplicity in weak B hadron decays is 5, and average track reconstruction efficiency is around 80% for tracks in jets

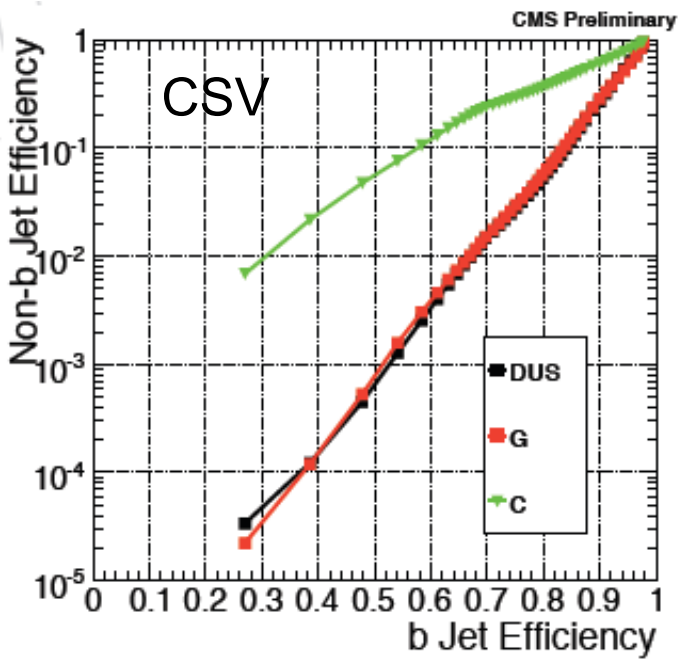




Secondary Vertex

- Simple Secondary Vertex
  - Based on reconstruction of atleast one "Secondary Vertex"
  - The significance of the 3D flight distance is used as a discriminator
- Combined Secondary Vertex
  - Involves the use of secondary vertices, together with other lifetime information, like the IP significance, vertex mass, number of tracks at the vertex, decay lengths …

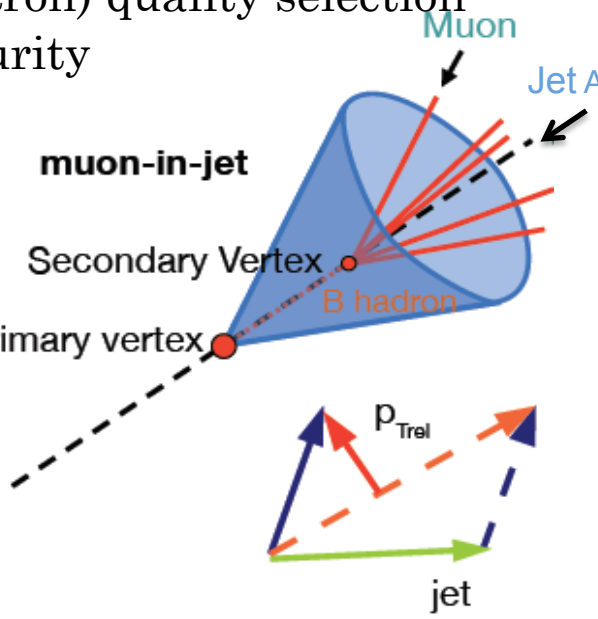




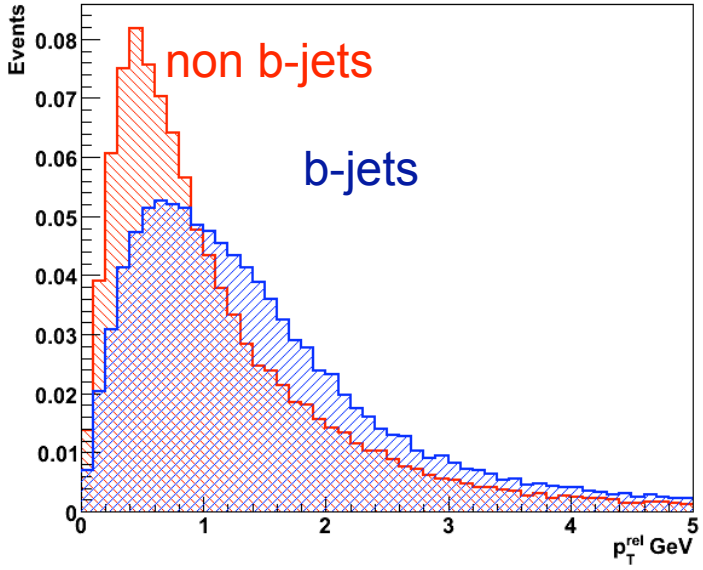
Soft Leptons

- Exploit large semileptonic branching fraction of B decays (Br(b → lepton) ≈ 10%) along with the large b-hadron mass (~5 GeV)
- Leptons from b decays are characterized by:
  - large impact parameter w.r.t. PV
  - large transverse momentum w.r.t. jet axis (p_{Trel})
  - large angular distance w.r.t. jet axis
- Lepton (muon or electron) quality selection improves b-tagging purity

$$p_{Trel} = \frac{p^{\mu} \times p^{\mu+jet}}{|p^{\mu+jet}|}$$



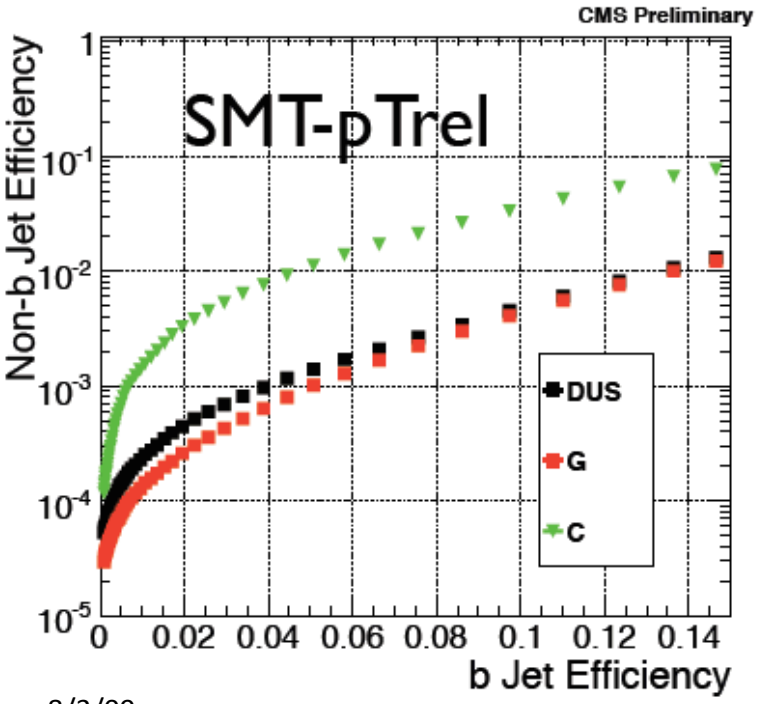
Allows to distinguish *b*-jets from *c*- and light-jets

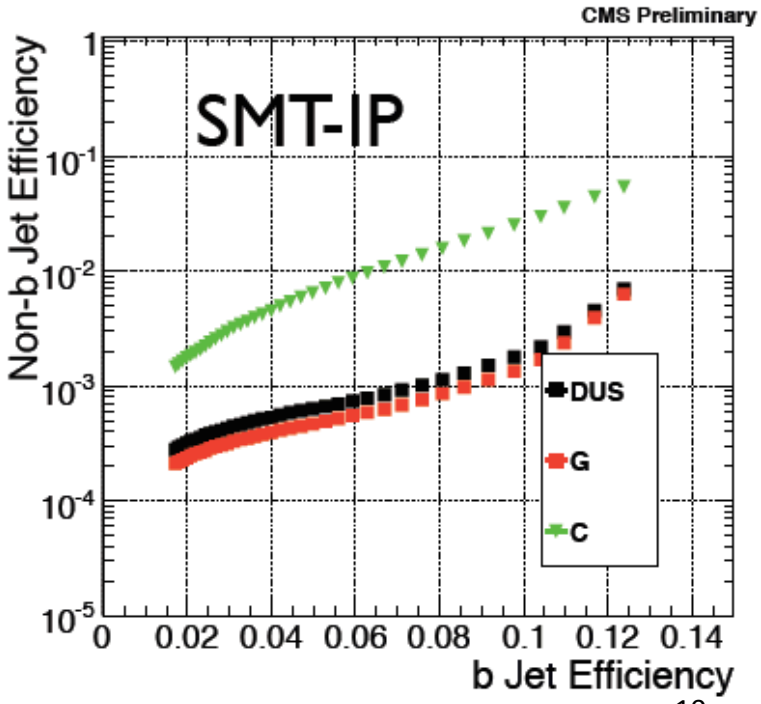


For muon-jets, p_{Trel} is defined as the p_T of the muon relative to the direction of the muon+jet axis

Muon *b*-tagging

- Two algorithms which, in addition to searching for a global muon within the jet
  - Use the $\mathbf{p_{Trel}}$ to the jet axis
  - Use the **IP significance** (only when positive) of the muon w.r.t. to the jet
- Tagger is very robust against tracker misalignment but suffers from a much lower efficiency





Operating Points for Summer/Fall '08 Samples

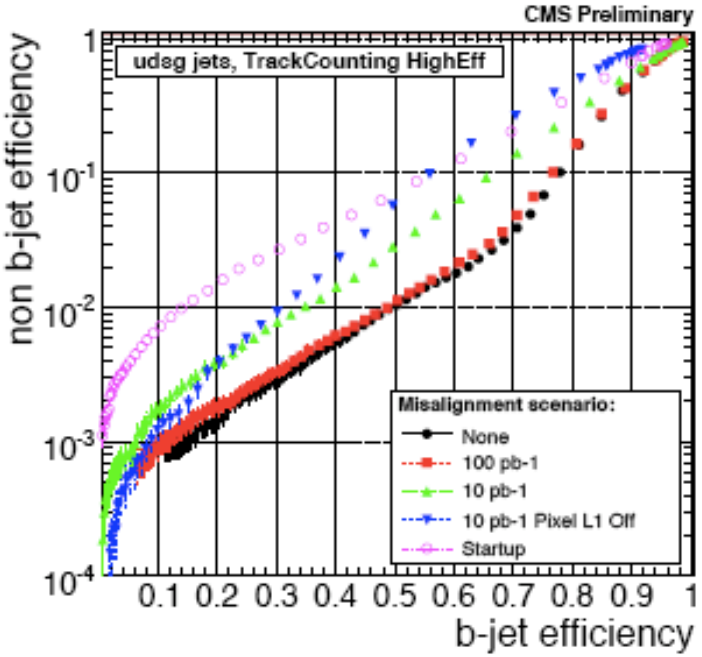
- Three Operating points are defined that select an average fraction of approximately 10%, 1% or 0.1% of *udsg* tagged jets

Tagger	Point	Discriminator	light mistag	b-efficiency
trackCountingHighEff	Loose	2.03	0.1	0.82
	Medium	4.38	0.01	0.65
	Tight	14.2	0.001	0.24
trackCountingHighPur	Loose	1.47	0.1	0.69
	Medium	2.36	0.01	0.6
	Tight	5.36	0.001	0.38
jetProbability	Loose	0.241	0.1	0.85
	Medium	0.49	0.01	0.65
	Tight	0.795	0.001	0.38
jetBProbability	Loose	1.1	0.1	0.85
	Medium	1.37	0.01	0.8
	Tight	1.39	0.001	0.79
simpleSecondaryVertex	Loose	1.25	0.1	0.72
	Medium	2.05	0.01	0.63
	Tight	4.07	0.001	0.23
combinedSecondaryVertex	Loose	0.387	0.1	0.87
	Medium	0.838	0.01	0.71
	Tight	0.94	0.001	0.52

Sample: Summer08 QCD dijet pt-hat 80-120 GeV
NB: Use calibrated IC5 jets.
Cuts: jet p_T > 30 GeV and jet eta < 2.4

Misalignment Scenarios

- Performance of these taggers has been studied under several misaligned scenarios (BTV-07-003)
- For the beginning of data taking, lifetime based b-tagging algorithms can be used. For a light flavor mistagging rate similar to the ideal detector case, the b-jet efficiency is reduced by 10-30%, depending on the degree of misalignment
- Simple Secondary Vertex tagger has been proved to be robust and has a sufficient performance at the same time



b-jet efficiency versus non b-jet efficiency for the various misalignment scenarios for the TrackCounting (high efficiency) algorithm for light flavor jets

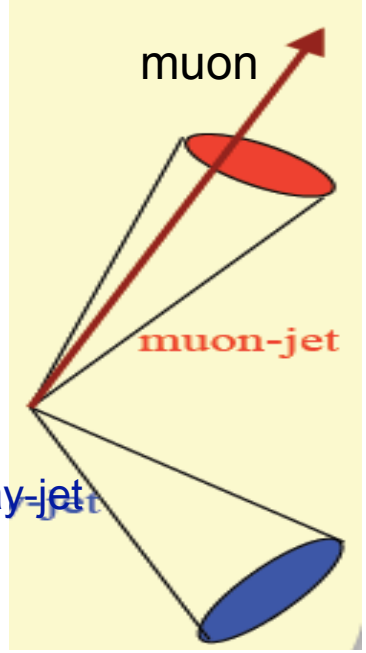
The study is based on a sample of 20K inclusive *ttbar* events produced with PYTHIA6

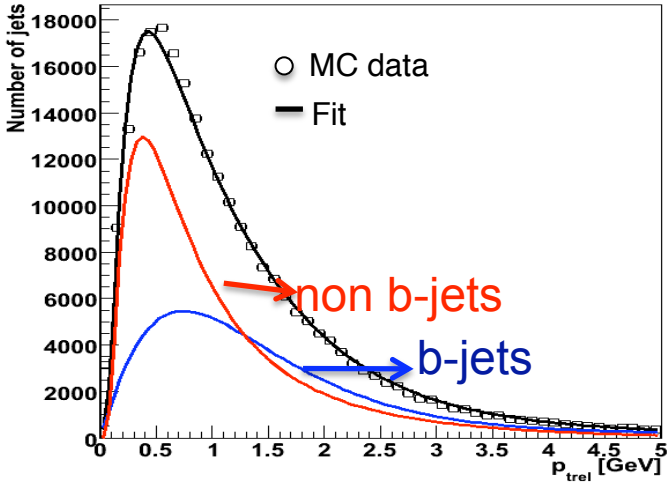
PURDUE
UNIVERSITY
CALUMET

Performance Measurement Methods

p_{Trel} Method

- The method is based on data samples that have atleast two reconstructed jets and a non-isolated muon close to one of the jets
- Determine *b*-content of sample by fitting the p_{Trel} distribution of the muon-jet to a linear combination of p_{Trel} templates for *b* and *non-b* jets before and after tagging the muon-jet with a tagger



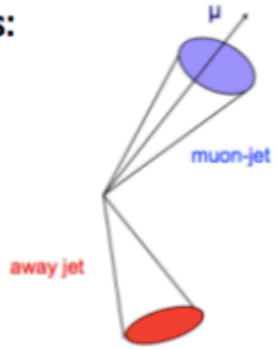


- The *b*-tagging efficiency is calculated as the ratio between the number of *b*-jets after and before tagging as determined by the p_{Trel} fits

System8 Method

BTV-07-001

- Based on two independent b-taggers and two samples:
  - muon-in-jet with away-jet
  - muon-in-jet with tagged-away-jet
- Two categories: b-jets and non b-jets (udsg)
- Need MC to estimate correlation parameters



$$n = n_b + n_{cl}$$
$$p = p_b + p_{cl}$$

muon-in-jet + away-jet
muon-in-jet + tagged-away-jet

$$n^{\text{tag}} = \varepsilon_b^{\text{tag}} n_b + \varepsilon_{cl}^{\text{tag}} n_{cl}$$
$$p^{\text{tag}} = \beta\, \varepsilon_b^{\text{tag}} p_b + \alpha\, \varepsilon_{cl}^{\text{tag}} p_{cl}$$

apply "probe" tagger

$$n^{\text{pTrel}} = \varepsilon_b^{\text{pTrel}} n_b + \varepsilon_{cl}^{\text{pTrel}} n_{cl}$$
$$p^{\text{pTrel}} = \delta\, \varepsilon_b^{\text{pTrel}} p_b + \gamma\, \varepsilon_{cl}^{\text{pTrel}} p_{cl}$$

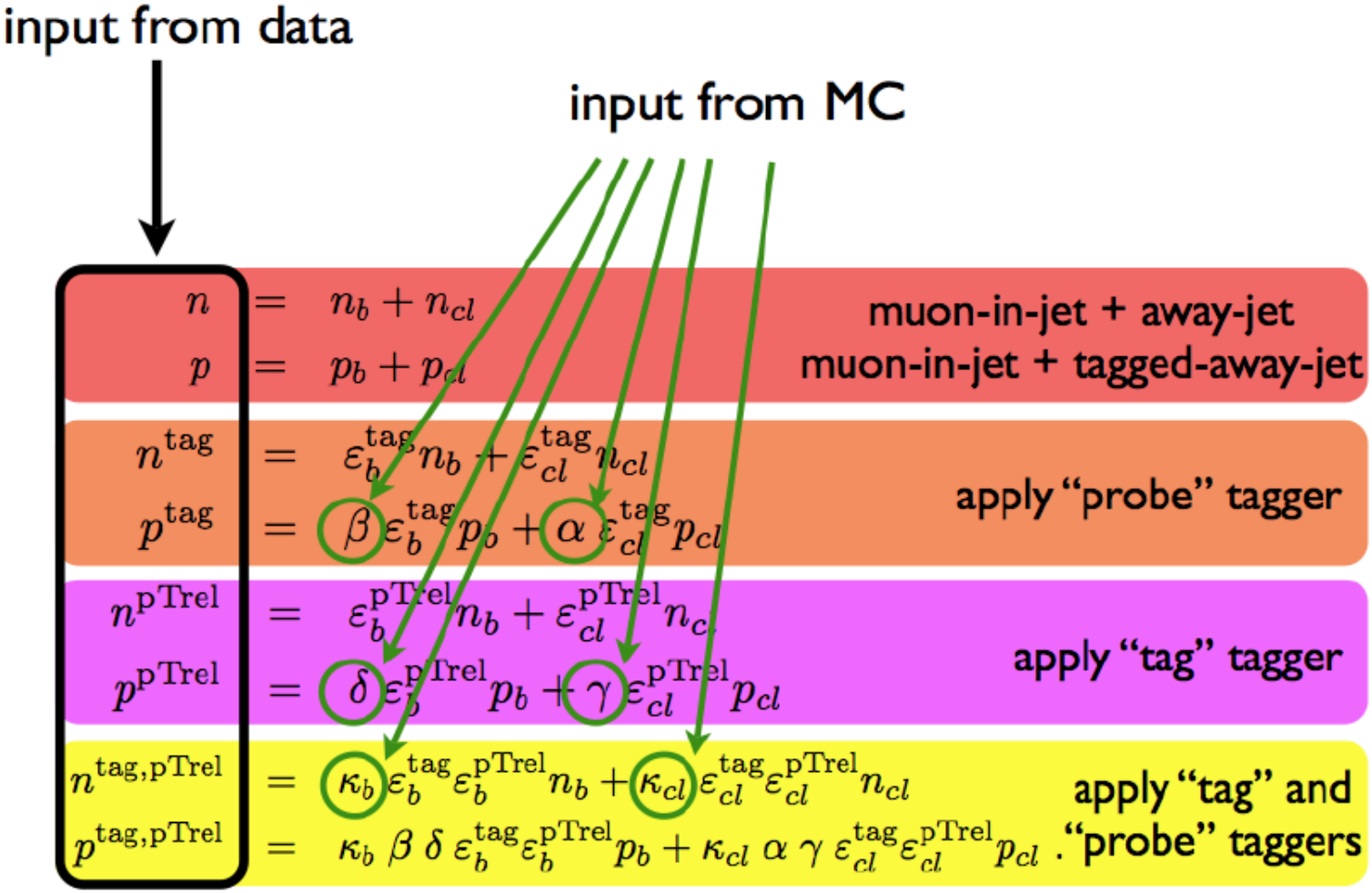
apply "tag" tagger

$$n^{\text{tag,pTrel}} = \kappa_b\, \varepsilon_b^{\text{tag}} \varepsilon_b^{\text{pTrel}} n_b + \kappa_{cl}\, \varepsilon_{cl}^{\text{tag}} \varepsilon_{cl}^{\text{pTrel}} n_{cl}$$
$$p^{\text{tag,pTrel}} = \kappa_b\, \beta\, \delta\, \varepsilon_b^{\text{tag}} \varepsilon_b^{\text{pTrel}} p_b + \kappa_{cl}\, \alpha\, \gamma\, \varepsilon_{cl}^{\text{tag}} \varepsilon_{cl}^{\text{pTrel}} p_{cl}$$

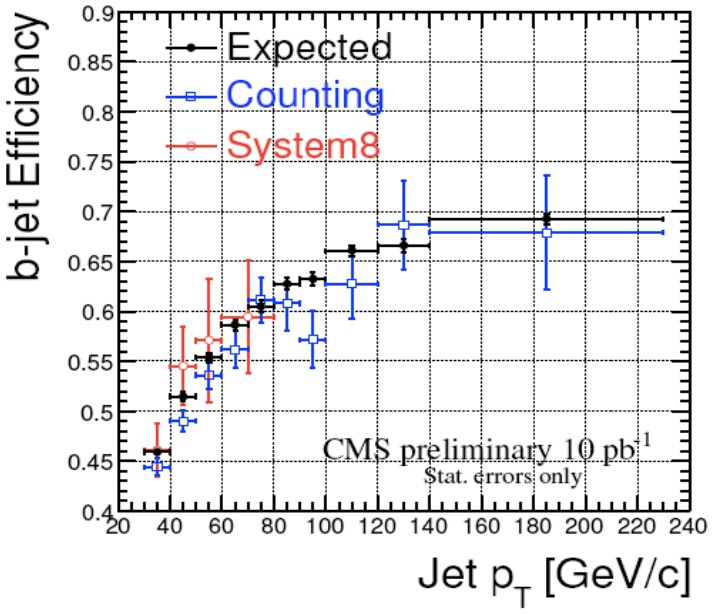
apply "tag" and "probe" taggers

System8 Method

output: efficiencies and fractions



Results from System8 and p_{Trel}

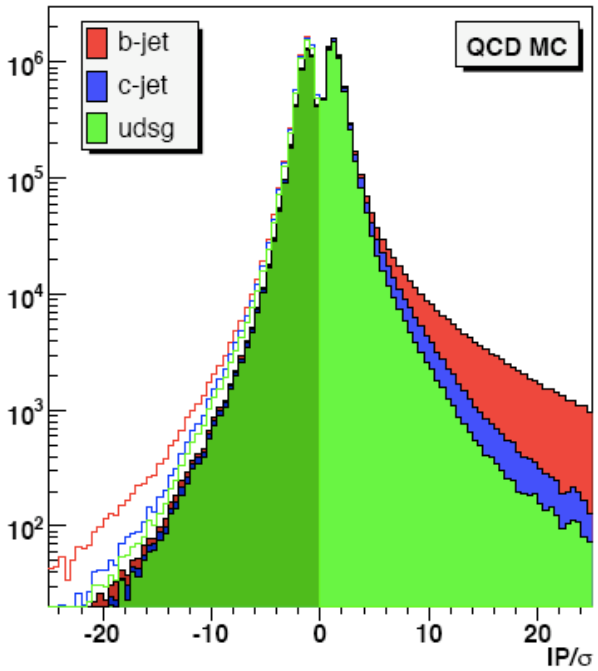


- Measured *b*-efficiency from System8 and p_{Trel} method as a function of the jet p_T (requiring p_T > 30 GeV), compared to the MC true (expected) efficiency

- The expected uncertainty on the *b*-tagging efficiency measured from muon+jet data for different luminosity scenarios is evaluated to be 15% for 10 pb^{-1}, 10% for 100 pb^{-1} and 5-6% for 1 fb^{-1}
- The p_{Trel} method is dominated by systematic uncertainty and the System8 by statistical

Mistag rate measurement using negative tags

- The method extracts rejection rates from light quarks looking at tracks with negative impact parameter, and using these distributions to model the mistag rate due to detector effects like resolutions and badly reconstructed tracks
- Uses the lifetime tagger algorithms



- Distribution of negative and positive discriminators should be approximately symmetric for *uds* and gluon jets

- However light *udsg*-jets are also affected by displaced processes such as long lived particles, hadronic interaction with the material and displacements originating from fake and badly reconstructed tracks

Mistag rate measurement using negative tags

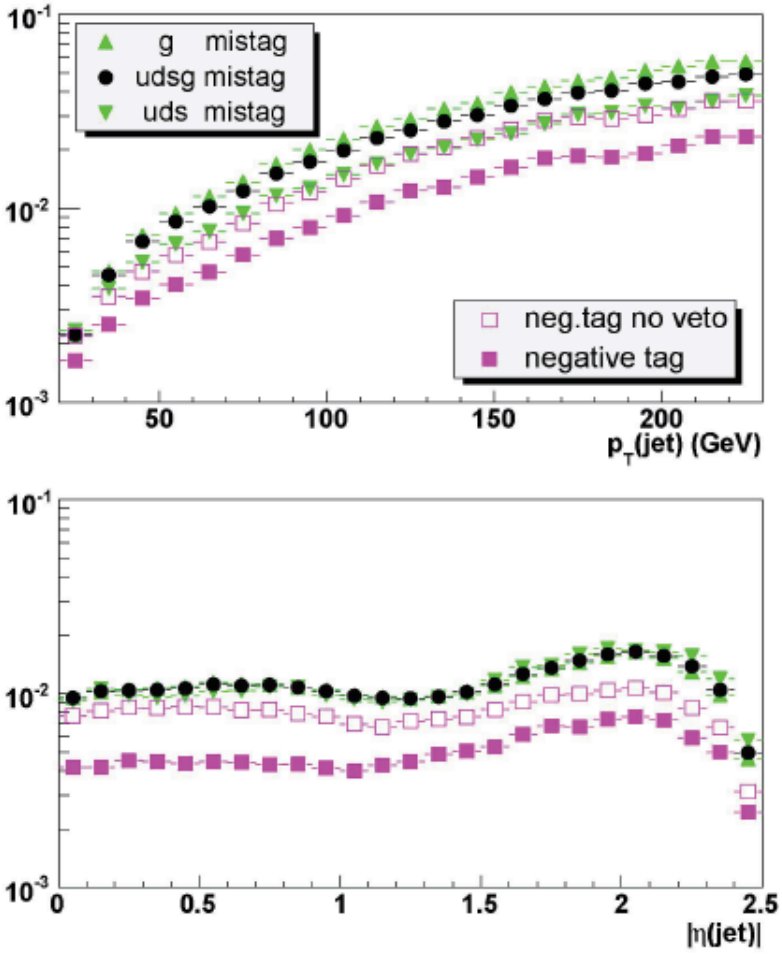
The mistag efficiency due to *udsg*-jets is evaluated as:

$$\epsilon_{data}^{mistag} = \epsilon_{data}^{-} \cdot R_{light}$$

ϵ_{data}^{-} is the negative tag rate in multi-jet data

$$R_{light} = \epsilon_{MC}^{mistag} / \epsilon_{MC}^{-}$$

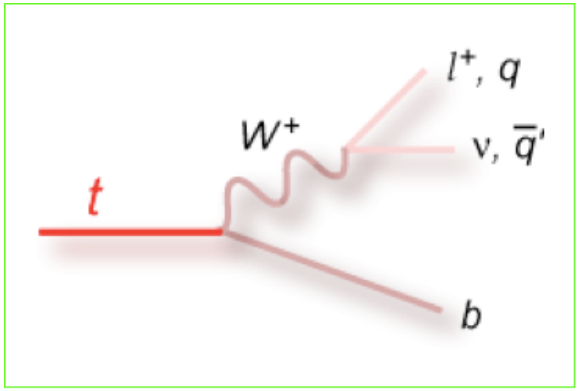
- R_{light} is the ratio between the mistag efficiency of *udsg*-jets and the negative tag rate for all jets in simulation
- The evaluation of the mistag efficiency is sensitive to the fractions of c and b quarks in the negative tag jet sample. The *c* and *b* fractions can be significantly reduced by applying a positive tag veto: the negative tag jet is rejected if it has any track with IP/σ_{IP} >4



Mistag efficiency and negative tag rate

Top Quark based: Likelihood Ratio Method

- Several methods that take advantage of the large number of top events produced in the LHC
- These methods will be very useful in the long term
- Measure the *b*-jet performance from *ttbar* events by isolating a jet sample enriched in *b*-jet content using a likelihood ratio



$$\mathcal{L} = \prod_i \frac{f_i(x_i)}{1 - f_i(x_i)}$$

- Build a multivariate classifier for jets
- Use the classifier to select *b* jets

Extract the efficiency: $\epsilon_b = \frac{1}{x_b}\left[x_{tag} - \epsilon_{cl}(1 - x_b)\right]$

fraction of b jets ← x_b

x_{tag} → fraction of tagged jets

ϵ_{cl} → mistag rate

B-tagging group plans to provide

PURDUE UNIVERSITY CALUMET

- Inclusive b, c and light tagging efficiencies in the form of multidimensional parameterizations denoted "Tag-Rate" (TR).
- TRs are derived from one-dimensional functions (E_T, η ,…) assuming that they can be factorized.
- TRs can be implemented as
  - Histograms (no biases)
  - Functions (more tolerant to low statistics)
- Ratio of data to MC efficiencies gives the scale factor (SF).
- This will be provided in principle only for one jet collection recommended by JetMET (SisCone5)

Tag rates and scale factor definitions

- Scale factors

$$SF_b(E_T,\eta) = \frac{\varepsilon^{data}_{b\to\mu}(E_T,\eta)}{\varepsilon^{mc}_{b\to\mu}(E_T,\eta)}$$

$$SF_c = SF_b$$

$$SF_l(E_T,\eta) = \frac{\varepsilon^{data}_{-}(E_T,\eta)}{\varepsilon^{mc}_{-}(E_T,\eta)}$$

- Tag rates

$$TR_b(E_T,\eta) = \frac{\varepsilon^{mc}_{b}(E_T,\eta)}{\varepsilon^{mc}_{b\to\mu}(E_T,\eta)} \times \varepsilon^{data}_{b\to\mu}(E_T,\eta) = SF_b(E_T,\eta) \times \varepsilon^{mc}_{b}(E_T,\eta)$$

$$TR_c(E_T,\eta) = SF_b(E_T,\eta) \times \varepsilon^{mc}_{c}(E_T,\eta)$$

$$TR_l(E_T,\eta) = \frac{\varepsilon^{mc}_{l}(E_T,\eta)}{\varepsilon^{mc}_{-}(E_T,\eta)} \times \varepsilon^{data}_{-}(E_T,\eta) = SF_l(E_T,\eta) \times \varepsilon^{mc}_{l}(E_T,\eta)$$

- In data, simply apply the tagger for each working point
- When analyzing MC, two options are available:
  - Actually tag the jets (direct tagging) and use SF to the MC efficiency to correct to data (*used at CDF*)
  - Do not tag and apply TR to get a probability to have a jet tagged (*employed by D0*)

Event Tagging Probability

- Needs to be derived by weighting each reconstructed jet in the event by a TR according to its flavor f, its E_T and its η .
- The probability to have at least one tag in a given event is the complement of the probability that none of the jets is tagged:

$$P_{\text{event}}^{\text{tag}}(0 \text{ tag}) = \prod_{i=1}^{N_{jets}} [1 - TR_{f_i}(E_{T_i}, \eta_i)]$$

$$P_{\text{event}}^{\text{tag}}(\geq 1 \text{ tag}) = 1 - P_{\text{event}}^{\text{tag}}(0 \text{ tag})$$

- The probability to have exactly one tag and two or more is:

$$P_{\text{event}}^{\text{tag}}(1 \text{ tag}) = \sum_{i=1}^{N_{jets}} TR_{f_i}(E_{T_i}, \eta_i) \prod_{j \neq i}^{N_{jets}} [1 - TR_{f_j}(E_{T_j}, \eta_j)]$$

$$P_{\text{event}}^{\text{tag}}(\geq 2 \text{ tag}) = P_{\text{event}}^{\text{tag}}(\geq 1 \text{ tag}) - P_{\text{event}}^{\text{tag}}(1 \text{ tag})$$

- The expected number of tagged events for a particular process is given by:

$$N_{\text{event}}^{\text{tag}} = N_{\text{event}}^{\text{pretag}} \bar{P}_{\text{event}}^{\text{tag}}$$

Average of the perevent Tagging probability over a sample of events for the process under consideration

Conclusions

PURDUE UNIVERSITY CALUMET

- Various *b*-tagging algorithms exploit characteristics of B hadrons
  - lifetime, mass, semileptonic decays
- SimpleSecondaryVertex and TrackCounting proved to be robust against tracker misalignment
  - Even with the cosmic allignment we are in good shape for simple *b*-tagging algorithms
- Use simple and more robust *b*-taggers at the beginning of data taking
  - track counting, displaced vertex, lepton tagging
- Use multivariate *b*-tagging techniques later as data is better understood to take full advantage of available information
- Several methods are present to measure *b*-tagging efficiency and mistag rate from collider data.
- All methods rely to some extent on the Monte Carlo information
- The robustness of each method depends on their sensitivity to the amount of data and the way simulated information is used by them
- It is important to develop several strategies to take advantage of their complimentary features