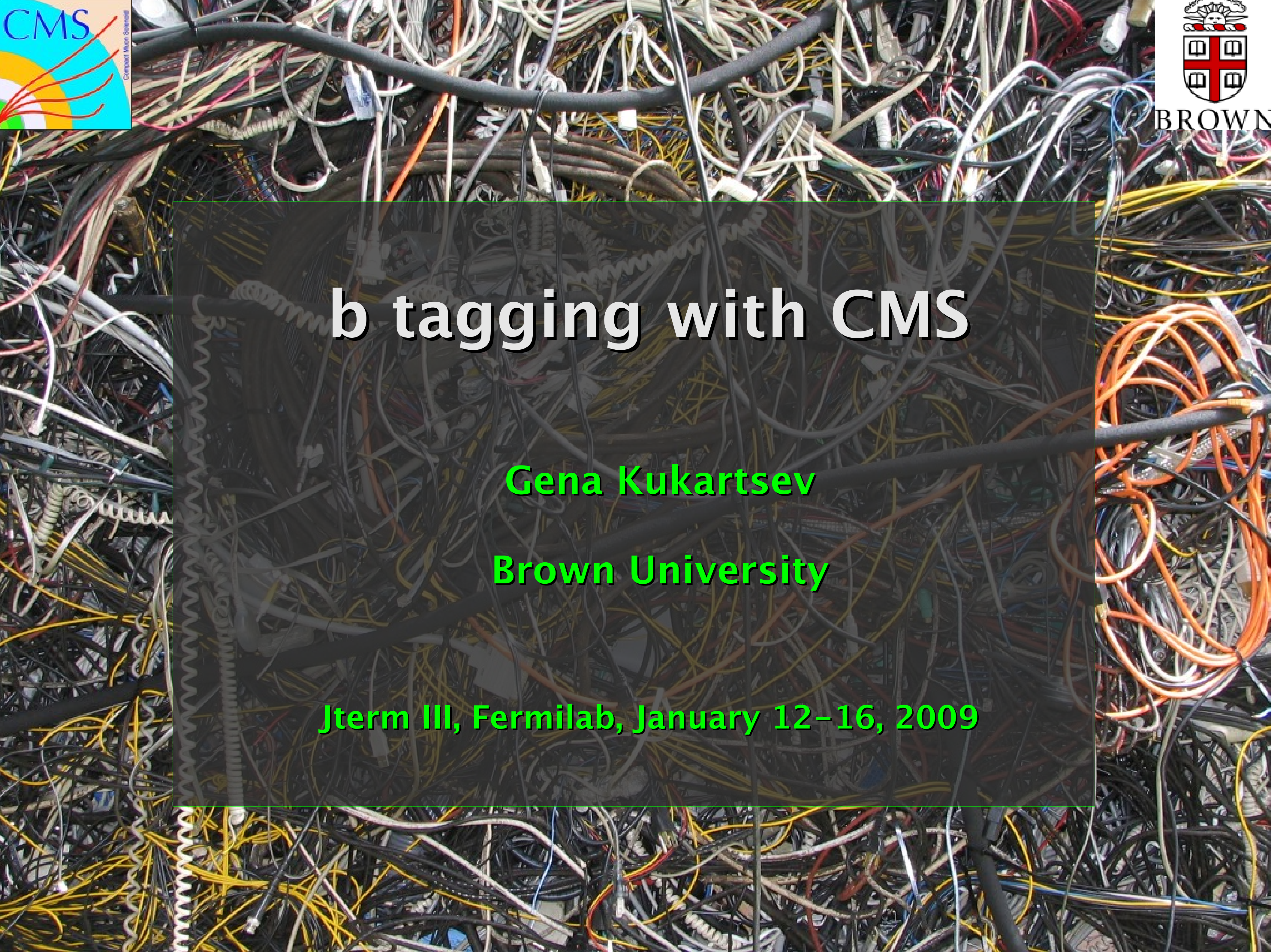


b tagging with CMS

Gena Kukartsev

Brown University

Jterm III, Fermilab, January 12-16, 2009

Outlook

- Many different algorithms implemented in CMSSW:
  - Universal / Specialized
  - Robust (for early data) / Complex
  - Different purity ranges
- “Simple” taggers: require no calibrations
  - The discriminant is the presence of a reconstructed object plus a single observable
  - Suitable for startup (allowing for misalignment) and HLT
- “Complex” taggers: require calibrations or training – not good for startup
  - Combine many observables with an MVA tool (Likelihood, NN)
  - Require calibrations (on real data)
- Negative tagging
  - Expect light jets to have symmetric distribution
  - Impact parameter, flight distance
  - Can be used to model non-b jets background for calibration on data

How we tell b jets

- Principle: exploit (weak) decay and production properties of b-hadrons:
  - Lifetime: τ ~1.5 ps ($c\tau$ ~450μm) → for p = 20 GeV/c, decay length ~ 1.8 mm
    - secondary (tertiary) decay vertex; displaced tracks with large IP
  - High mass (~5.2 GeV) and decay multiplicity (~5 charged tracks)
  - Decay kinematics (e.g. rapidities)
  - Hard b-quark fragmentation function:
    - decay products with larger p_T relative to b hadron flight direction
  - Semi-leptonic decays (per lepton flavour: BR ~ 11%, ~20% incl. cascade)
- Problem:
  - charmed hadrons: long lifetime (D+: 312 μm, D0: 123 μm) and mass (~1.87GeV)
  - strange hadrons (K^0_S): remove track pairs with mass ~ $m(K^0_S)$
  - conversions ($\gamma \rightarrow ee$): remove vertices too far away, tracks with IP too large

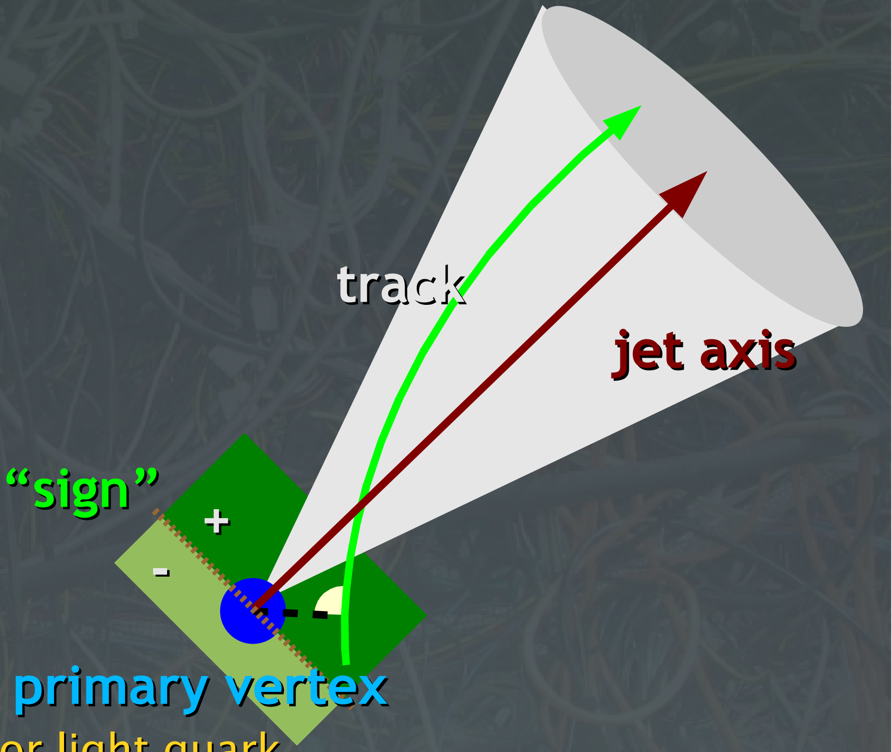
Physics case for b tagging

- New physics states like to couple to t and b quarks
- Higgs boson may decay to b and anti-b quarks most of the time
- t quarks decay into b quarks
  - There will be many t quarks produced at LHC
  - Cross section of the ttbar production is high, ~800pb
  - Will create an important background to new physics searches
  - t quark is interesting on its own: not as well understood as we sometimes think

b–tagging algorithms

- Impact parameter
  - Track counting: TrackCountingHighEff, TrackCountingHighPur
  - Jet probability: JetProbability, JetBProbability
  - ImpactParameterMVA
- Secondary vertex
  - SimpleSecondaryVertex
  - CombinedSecondaryVertex, CombinedSecondaryVertexMVA
- Soft lepton
  - (simple) SoftMuonByPt, SoftMuonByIP3d
  - SoftMuon, SoftMuonNoIP
  - SoftElectron

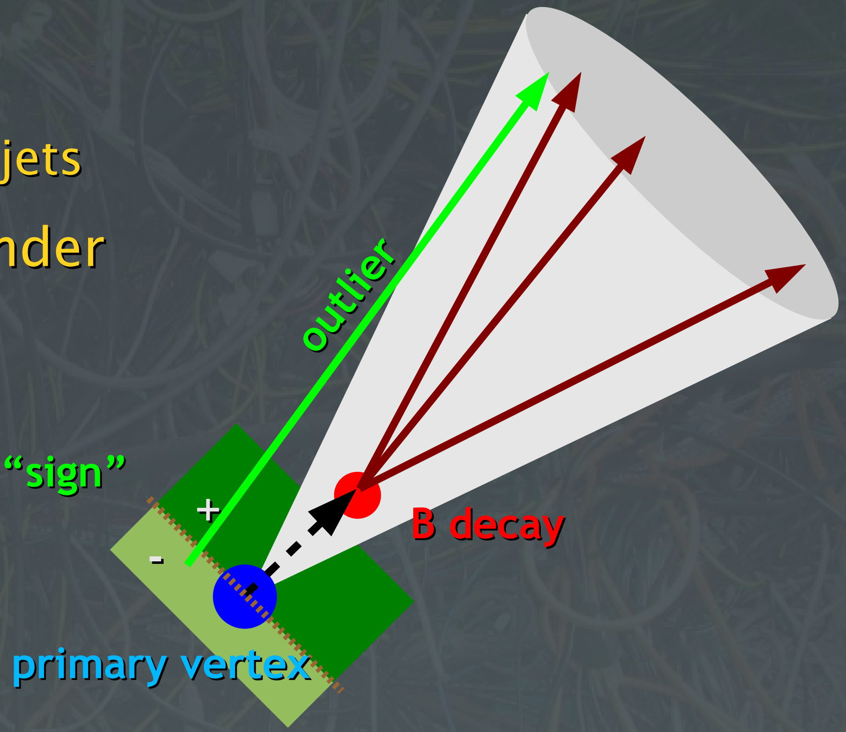
Impact parameter taggers



- Using Impact parameter significance
- Simple taggers:
  - Track Counting, high efficiency
    - SIP3d of 2nd track
  - Track Counting, high purity
    - SIP3d of 3rd track
- Complex taggers:
  - Track Probability
    - Likelihood that each track comes from a b or light quark
    - Combine the likelihood of all tracks into a jet likelihood
  - Track B Probability
    - Only use first 4 tracks
    - Re-weight "b" likelihood
  - Impact Parameter MVA
    - Combined all track IPs via MVA

Secondary vertex tagger

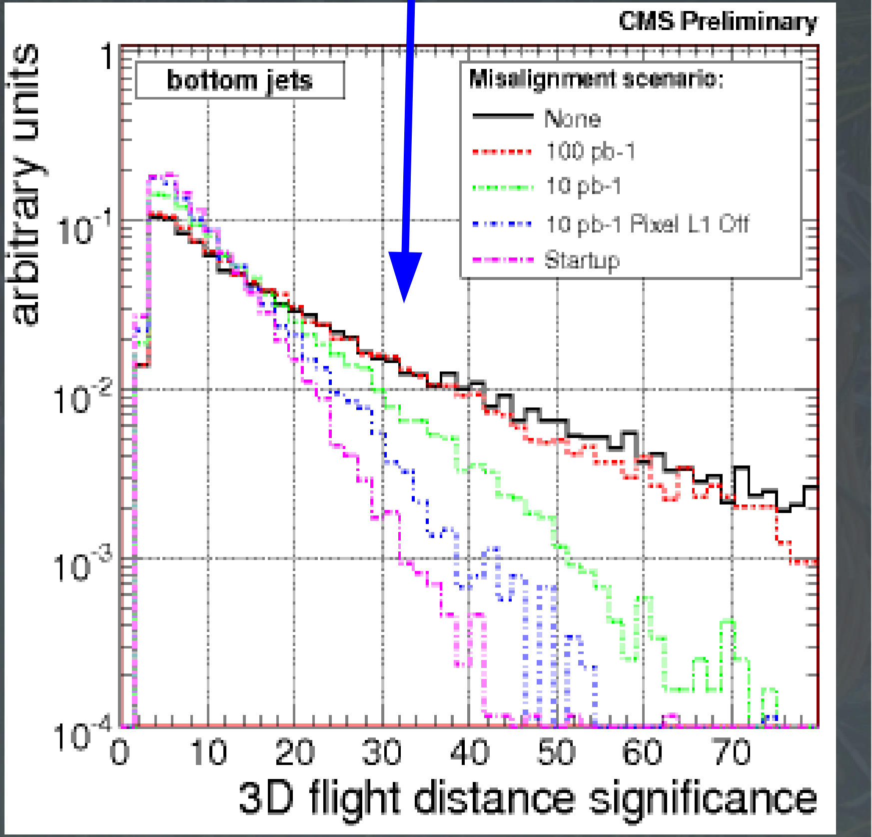
- Inclusive vertex reconstruction in jets
  - Long lifetime of *b*–hadrons
  - Secondary vertices can be expected in the jets
- Vertex reconstruction: Adaptive Vertex Finder
  - Build a Vertex
  - Downweight incompatible tracks
  - Repeat until all tracks are used
- Reject primary vertex
- Mass constraint
  - Reject M(vertex) > 6.5 GeV
  - Reject KS
- Simple tagger: Simple Secondary Vertex

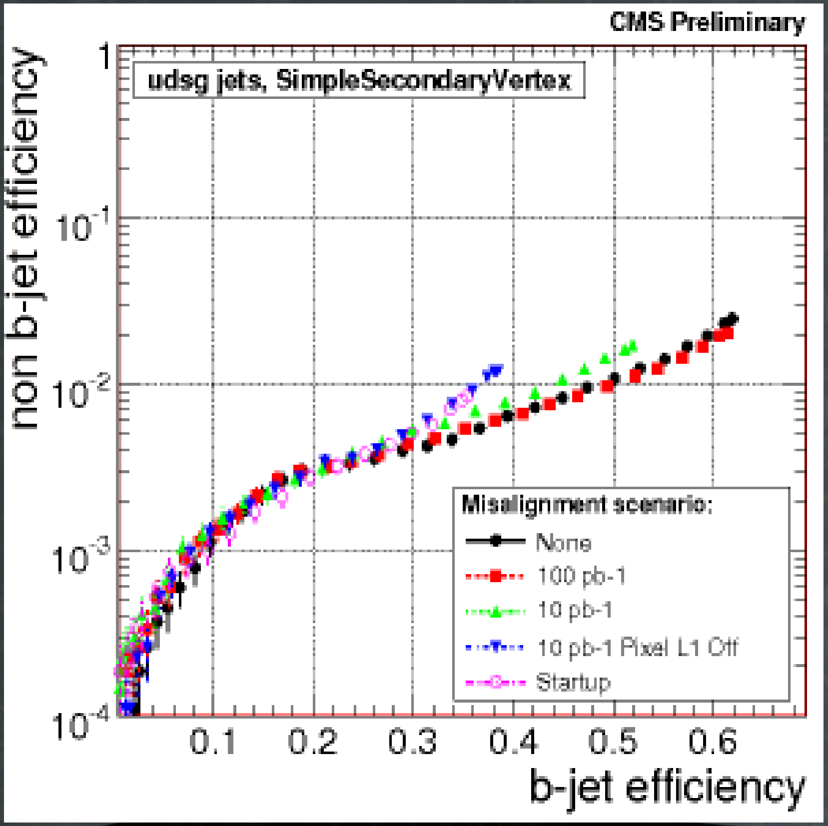


Simple Secondary Vertex tagger

- Simple Secondary Vertex tagger: signed 3D flight distance significance
  - Least sensitive to tracker misalignment
  - smaller systematics
  - limited b-jet tagging efficiency !

→ will only reach ~70% with ideal detector





Impact of Misalignment on the CMS b-Tagging
C. Saout, A. Scheurer, F.-P. Schilling, A. Schmidt
BTV-07-003, CMS AN 2007/47 (approved)

Combined secondary vertex

- Define *Vertex Categories* depending on result of vertex search:

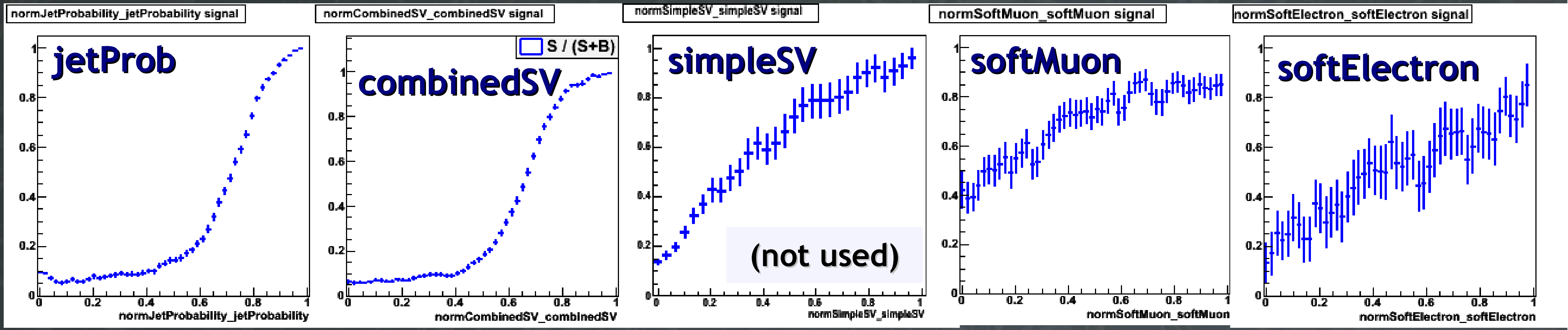
  1) "RecoVertex": at least one accepted SV candidate found

  2) "PseudoVertex" : built from tracks incompatible with the primary vertex (IP sign. > 2), if at least two such tracks are present

  3) "NoVertex": the rest

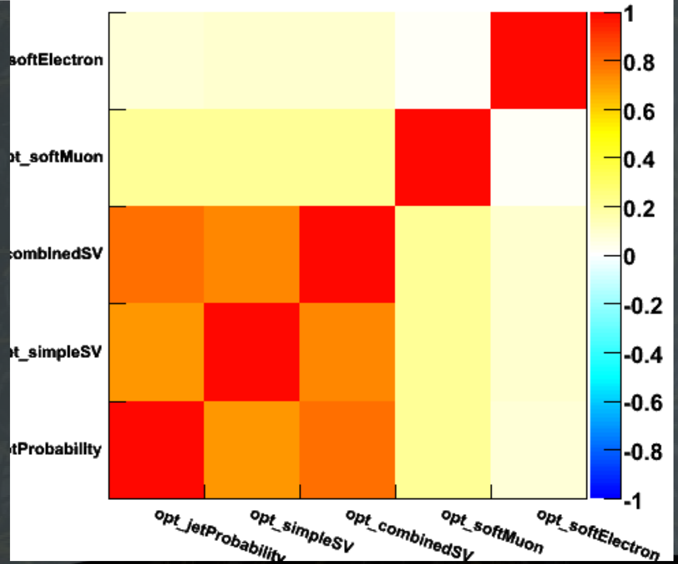
- Input variables
  - All track information (SIP3d of all tracks, contribution to inv. mass ...)
  - SV informations (flight distance, SV invariant mass, …)
- Complex Taggers:
  - Combined Secondary Vertex
    - For each variable, likelihood to come from a b, c or light quark
    - Ratio of combined likelihoods as b-tag discriminator
    - Depends on priors for charm and light content, and PDFs for inputs
  - Combined Secondary Vertex MVA
    - Uses a Neural Network (via the MVA framework) instead of likelihoods
  - Training has an intrinsic dependence on PDFs of input variables

MVA combined tagger

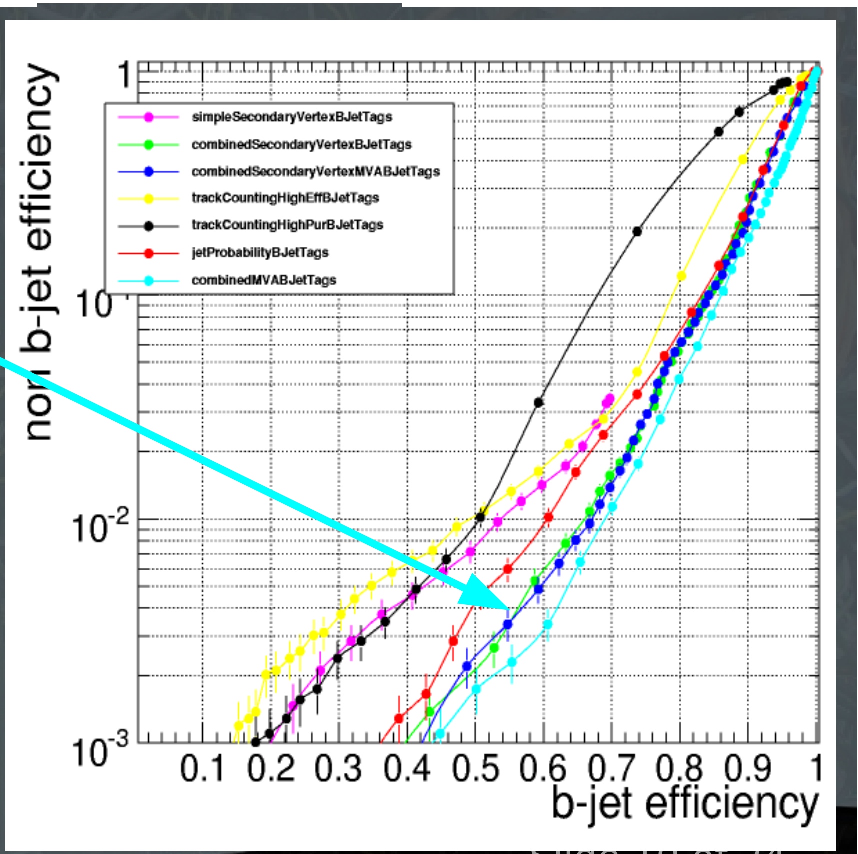


Combine discriminators of lifetime and soft lepton information (C.Saout)

correlation matrix

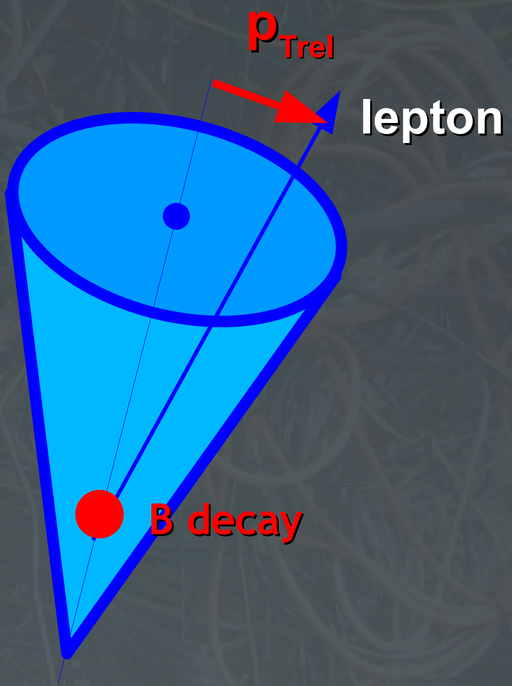


CombinedMVA



Soft lepton tagger

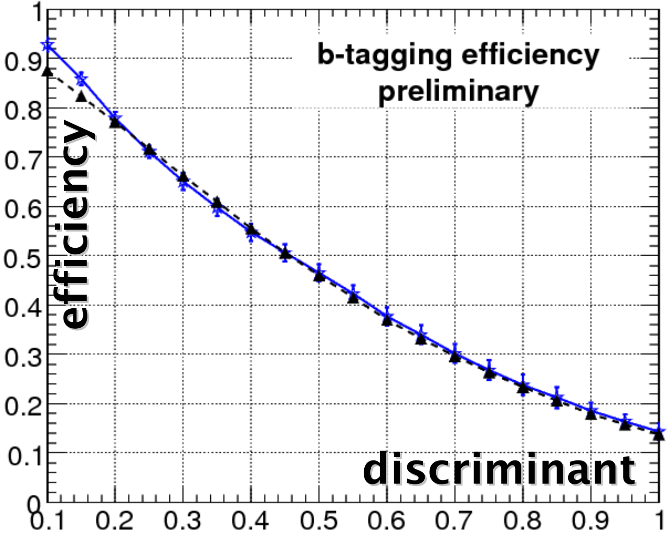
- Identify b-jets through electron or muon from semi-leptonic decay of b
  - inclusive direct & cascade decay branching ratio to each lepton family ~20%
- Simple taggers:
  - Soft Muon by SIP3d
    - requires a muon with SIP3d > cut
  - Soft Muon by pTrel
    - requires a muon with pTrel > cut
- Complex taggers:
  - Combine kinematic variables of leptons with NN
    - pTrel, SIP3d, DeltaR, rapidity...
  - Soft Muon, Soft Electron
  - Soft Muon NoIP
    - All but SIP3d
  - Calibrations are outdated: work in progress !

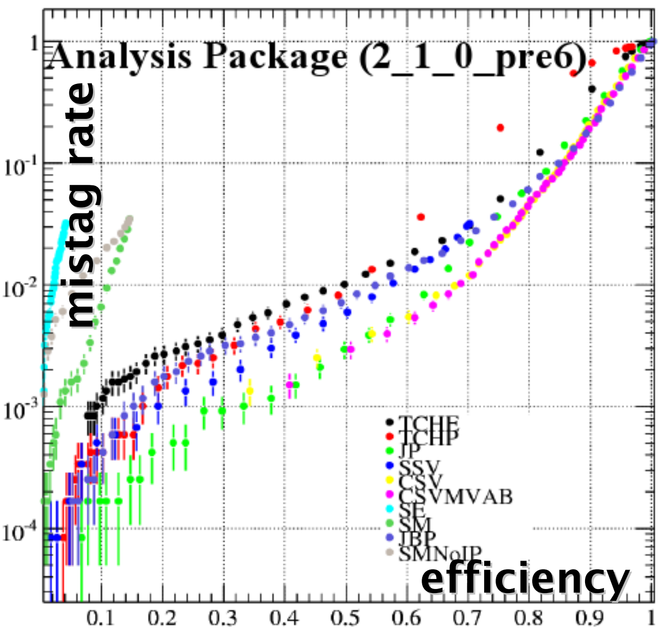


b tagging efficiency

- We can never know for sure whether the jet is a b jet or not
- The output of a tagger is a "discriminant: usually, a floating point number, usually from 0 to 1, a measure of "b-flavorness" of a jet
- b-tagging group will at first provide performance estimates for a limited number of "working points":
    - Loose (e.g. discriminant > 0.5),
    - medium, tight b tags etc...
    - Working points are tuned in terms of "mistag rate" – probability of tagging a non-b quark jet (e.g. 10% – loose, 1% – medium, 0.1% – tight)

Unavoidable compromise:
tag most b jets and do a lot of mistakes (high efficiency)
or
tag precious few b jets yet be almost sure in each of them (high purity)

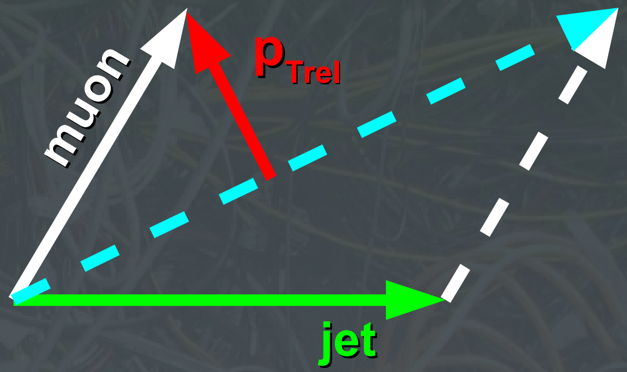




Performance measurement

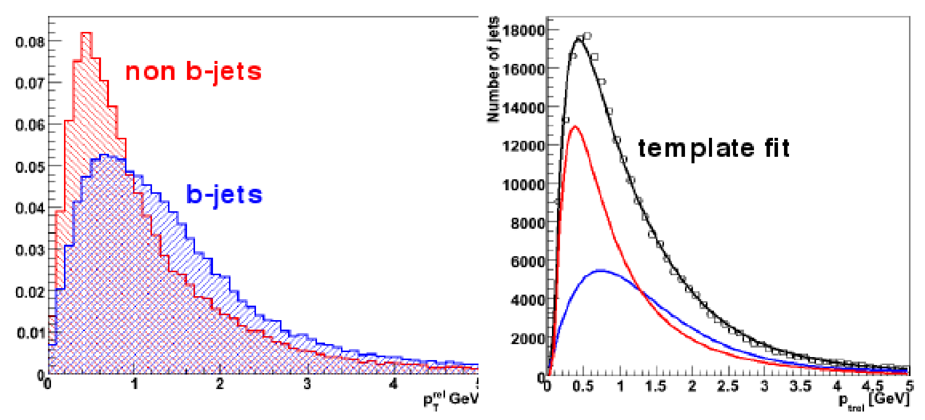
- Effort put on performance measurement from data:
  - Tagging efficiencies and mistag rates from the data
  - Understanding / Calibration of input objects
  - Comparison of Monte Carlo and data
- We want to be able to measure as fast as possible, as reliably as possible, efficiencies and mistag rates from data
  - Possibly, have as many cross check methods as possible
- Efficiency measurement (wiki):
  - "System 8" and p_{Trel} method (CMS AN-2007/046 – approved)
  - From *tt* events (CMS AN-2007/30)
- Measurement of mistagging rate (wiki):
  - Measurement of light quark tagging efficiency using multi-jet events and negative impact parameter tags (CMS AN-2007/048 – approved)
- CDF methods being implemented (efficiency & mistag, link)
- New ideas being proposed and pursued (since we have time... link)

pTrel



- Consider a jet with a muon in it
- P_{Trel} is the muon transverse momentum relative to the jet axis
- P_{Trel} is likely to be larger in a b jet than in any other jet

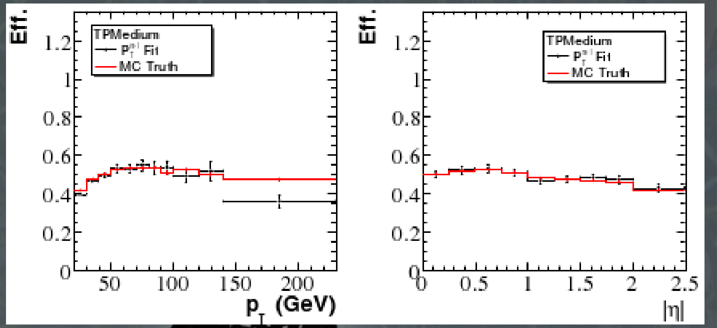
$$p_{Trel} = \frac{\vec{p}_\mu \times (\vec{p}_\mu + \vec{p}_{jet})}{\vec{p}_\mu + \vec{p}_{jet}}$$



- We get the shapes of the P_{Trel} distribution for b jets and non-b jets from Monte Carlo
- Fit the P_{Trel} distribution of a jet sample with the templates and find the amount of b jets in the sample

pTrel method

- Consider a dataset
  - two reconstructed jets where one jet has a muon
- Subset, where the muon jet is tagged
- Use p_{Trel} template fit to find the amount of
  - b jets with muons (**n_b**) – from the main set
  - Tagged b jets with muons (**n_b^{tag}**) – from the subset
- b-tagging efficiency will be $$\epsilon_b = \frac{n_b^{tag}}{n_b}$$
- This is measured as a function of jet pseudorapidity and transverse momentum
- The main systematical uncertainty comes from the dependence on the Monte Carlo for the p_{Trel} templates (currently ~15%)



p_{Trel} **pseudorapidity**

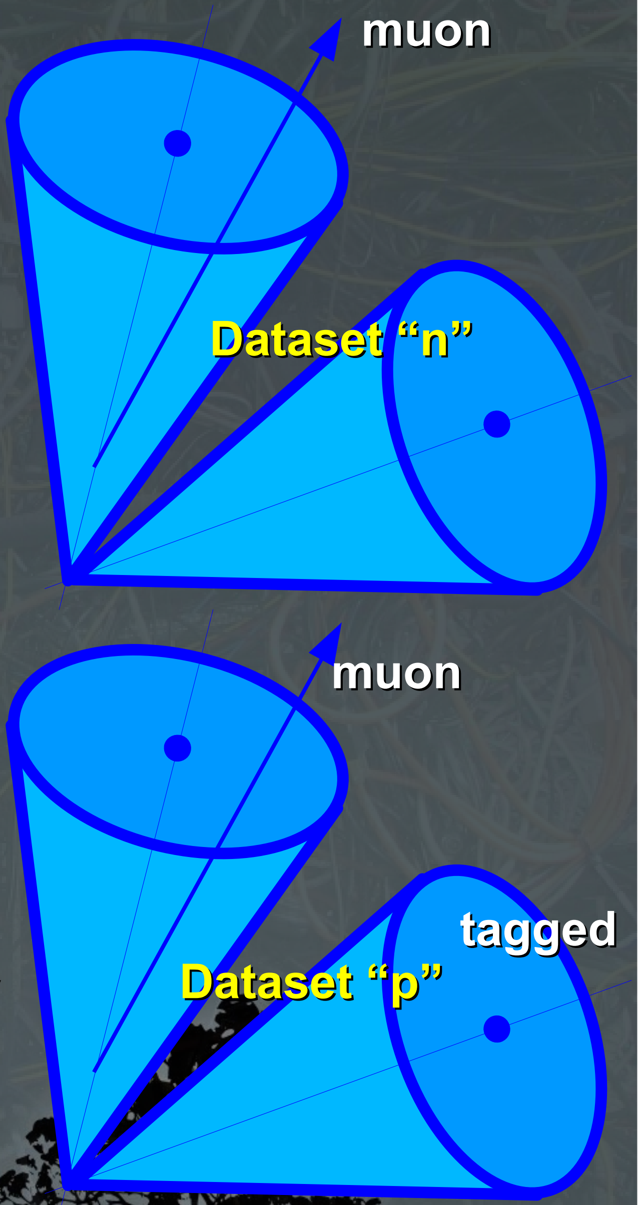
CMS Analysis Note 2007/046

Counting method

- A variation of the p_{Trel} method
- Consider two datasets
  - Two reconstructed jets where one jet has a muon **(n events)**
  - Same, and the “away” jet is tagged **(p events)**
- Use p_{Trel} template fit to find $\mathbf{n_b, n_b^{tag}, p_b, p_b^{tag}}$
- If one assumes that the “away” jets in the (n) sample are mostly non-b jets then the b-tagging efficiency is

$$\epsilon_b = \frac{p_b^{tag}}{p_b} = \frac{p_b^{tag}}{p - p_{cl}} = \frac{p_b^{tag}}{p - n_{cl} \cdot \epsilon_{cl}} = \frac{p_b^{tag}}{p - (n - n_b) \cdot \epsilon_{cl}}$$

- This assumption is questiionable, not supported by the existing Monte Carlo
- **There is no clear advantage over pure p_{Trel} method as of today**



CMS Analysis Note 2007/046

”System 8”

- Minimal dependence on MC
    - Only correlation factors between taggers
- Two independent taggers
- Two different data sets
    - As in the counting method
- Two jet categories: b and non-b (cl)
- 8 equations, 8 unknowns
- Solve numerically and get b-tagging efficiency as a function of jet pseudorapidity and transverse momentum

$$n = n_b + n_{cl}$$

$$p = p_b + p_{cl}$$

$$n^{tag} = \epsilon_b^{tag} n_b + \epsilon_{cl}^{tag} n_{cl}$$

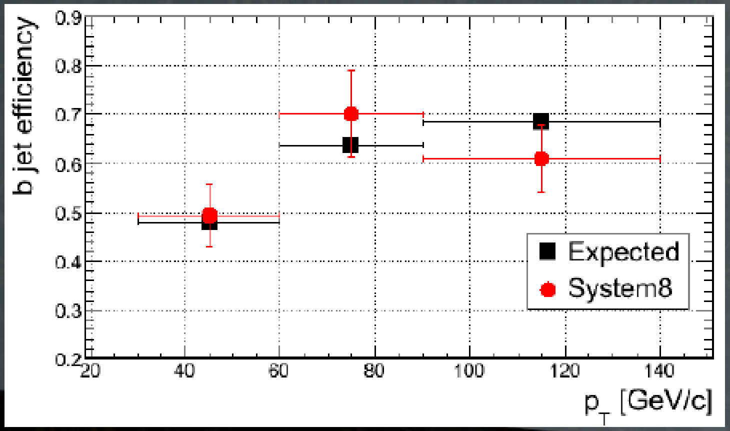
$$p^{tag} = \beta \epsilon_b^{tag} p_b + \alpha \epsilon_{cl}^{tag} p_{cl}$$

$$n^{\mu} = \epsilon_b^{\mu} n_b + \epsilon_{cl}^{\mu} n_{cl}$$

$$p^{\mu} = \delta \epsilon_b^{\mu} p_b + \gamma \epsilon_{cl}^{\mu} p_{cl}$$

$$n^{tag,\mu} = \kappa_b \epsilon_b^{tag} \epsilon_b^{\mu} n_b + \kappa_{cl} \epsilon_{cl}^{tag} \epsilon_{cl}^{\mu} n_{cl}$$

$$p^{tag,\mu} = \kappa_b \beta \delta \epsilon_b^{tag} \epsilon_b^{\mu} p_b + \kappa_{cl} \alpha \gamma \epsilon_{cl}^{tag} \epsilon_{cl}^{\mu} p_{cl}$$



Left-hand side - observables

Correlation factors (greek) – from MC

All the rest is solved for

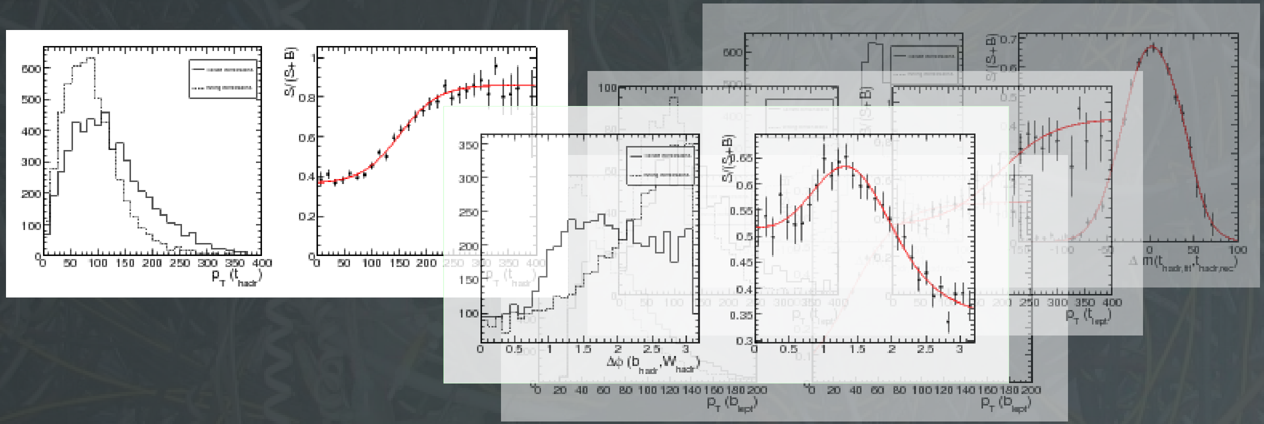
CMS Note 2008/081

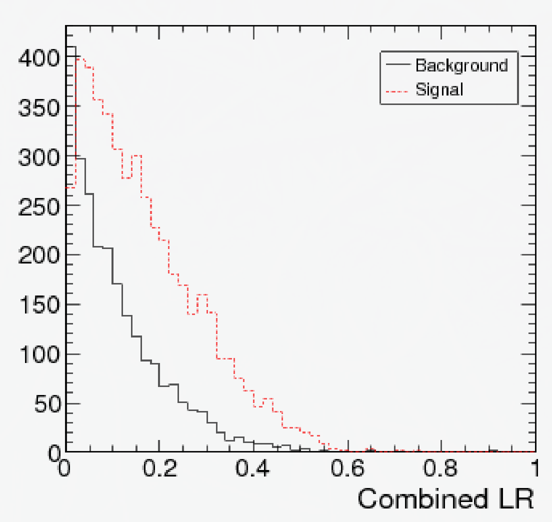
The likelihood ratio method

- The method is identical with both semileptonic and fully leptonic events
- Use enriched b jet sample
  - Build a multivariate classifier for jets
  - Use the classifier to select b jets

$$t\bar{t} \rightarrow (W^+ b)(W^- \bar{b}) \rightarrow (l^+ \nu b)(\bar{c} s \bar{b})$$

$$t\bar{t} \rightarrow (W^+ b)(W^- \bar{b}) \rightarrow (l^+ \nu b)(l^- \bar{\nu} \bar{b})$$





Extract the efficiency:

$$\epsilon_b = \frac{1}{x_b}\left[x_{tag} - \epsilon_0 (1 - x_b)\right]$$

fraction of b jets

fraction of tagged jets

mistag rate

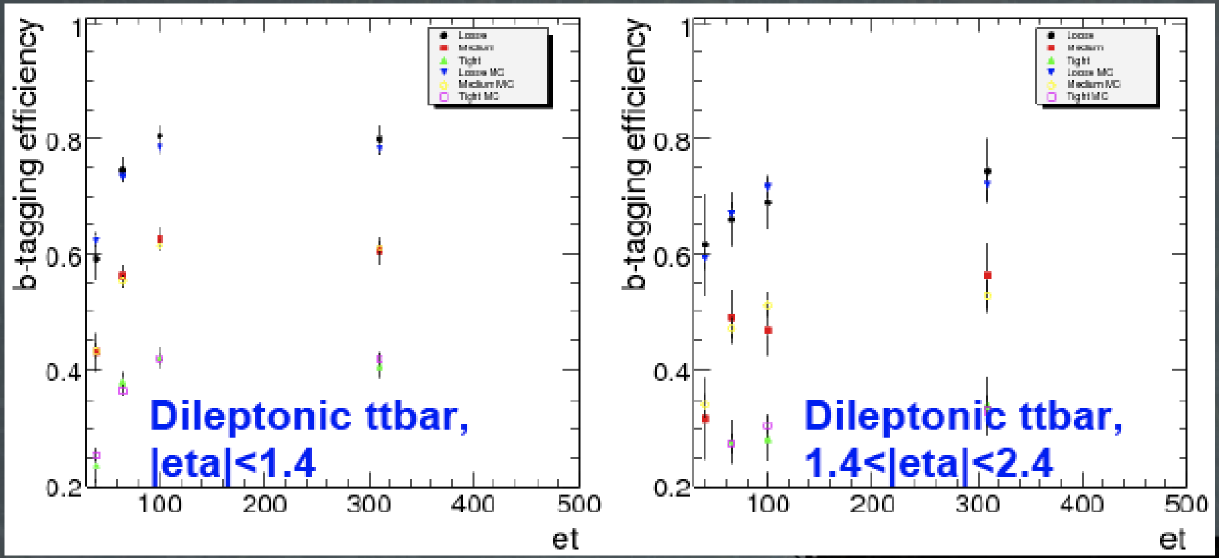
Joris Maes, Thomas Speer
inspired by
Steven Lowette

CMS AN-2007/30

Likelihood ratio method

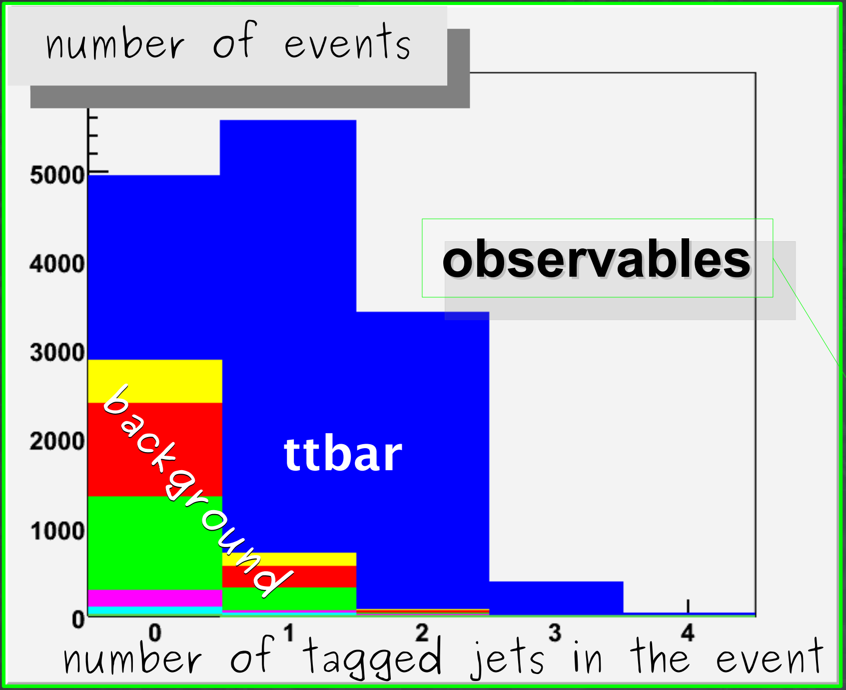
- b tagging efficiency is estimated inclusively for 10/pb scenario. Beyond that – as a function of jet pseudorapidity and transverse momentum
- Challenges
  - Combinatorics of the event candidates (semileptonic)
  - Low statistics (dilepton)
- Main sources of systematical uncertainty
  - Likelihood ratio cut bias
  - Initial and final state radiation

CMS AN-2007/30



Tag consistency method

With semileptonic ttbar events: $t\bar{t} \to (W^+ b)(W^- \bar{b}) \to (l^+ \nu b)(\bar{c} s \bar{b})$



Extract b-tagging efficiency by counting the events with a given amount of tags, using the known jet flavor content of the data

..... background

jet flavor (MC)

F_0_0_4 = 2.23843
F_0_0_5 = 0.803966
F_0_0_6 = 0.230982
F_0_1_1 = 0.522502
F_0_1_2 = 0.602384
...

ML fit enforces consistency between expected and observed

b-tagging efficiency

optional: mistag rates, ttbar cross section

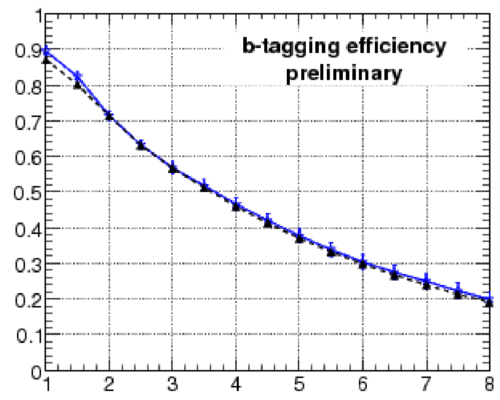
Gena Kukartsev, Meenakshi Narain

CMS AN-2007/30

Tag consistency method

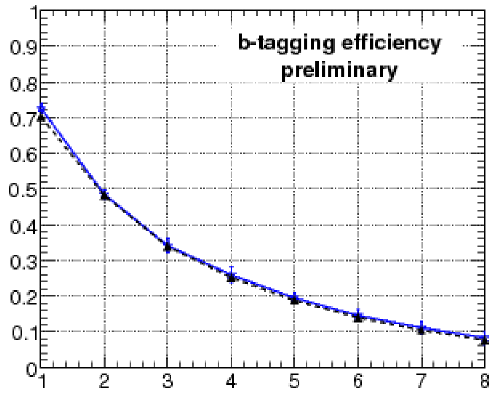
Working point	Tagger	discriminant	b tagging efficiency	“official”
Loose	track counting	>2.0, high efficiency	0.726 ± 0.014(stat) ± 0.001(JES)	0.71
	jet probability	>0.26	0.715 ± 0.012 ± (<<0.001)	0.75
Medium	track counting	>4.6, high efficiency	0.538 ± 0.013 ± 0.001	0.53
	jet probability	>0.50	0.547 ± 0.013 ± (<<0.001)	0.54
Tight	track counting	>4.7, high purity	0.299 ± 0.013 ± 0.001	0.3
	jet probability	>0.76	0.334 ± 0.014 ± (<<0.001)	0.33

Track Counting, high efficiency



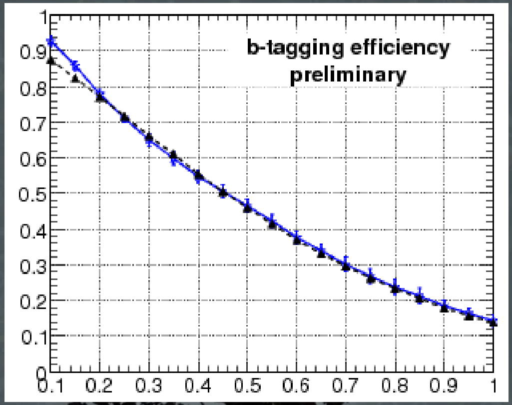
discriminant

Track Counting, high purity



discriminant

Jet probability



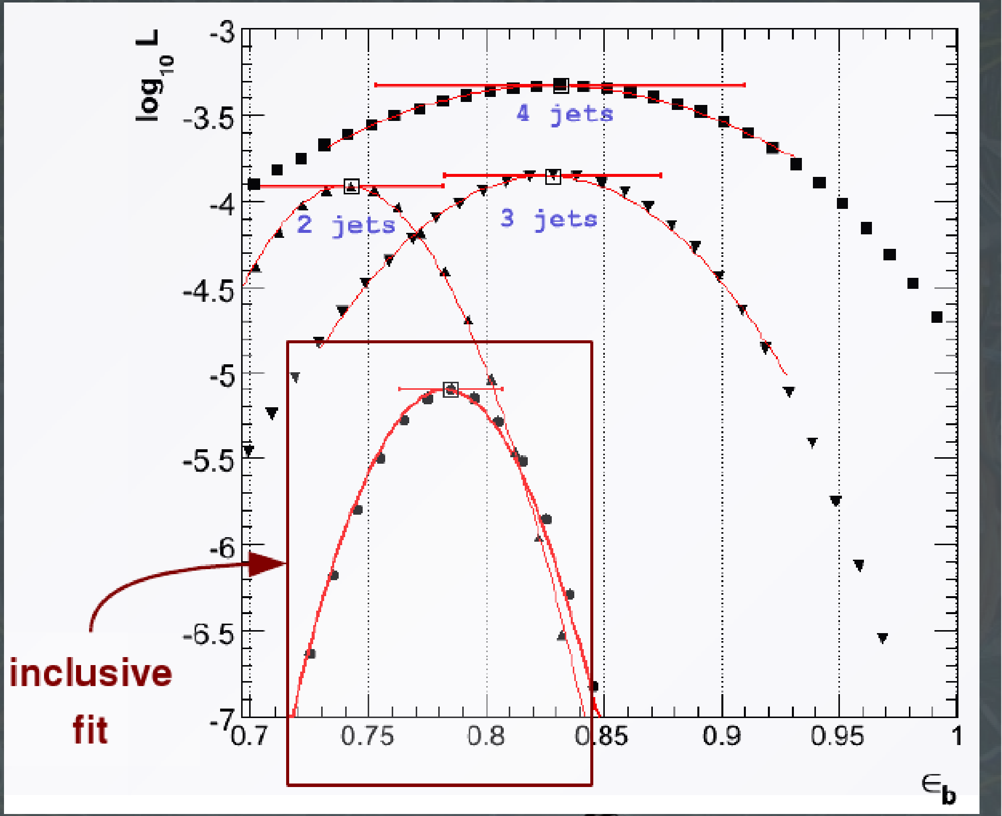
discriminant

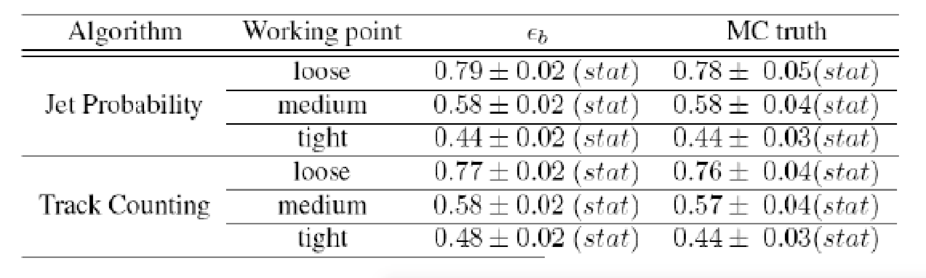
Tag consistency (dilepton)

With fully leptonic ttbar events:

$t\bar{t} \rightarrow (W^+ b)(W^- \bar{b}) \rightarrow (l^+ \nu b)(l^- \bar{\nu} \bar{b})$

- Using low jet multiplicity events
- Use lepton-jet correlation for jet-lepton assignement
- Data-driven method to measure jet purity and derive background levels
- Systematical uncertainties are to be studied





Algorithm	Working point	ϵ_b	MC truth
Jet Probability	loose	$0.79 \pm 0.02\ (stat)$	$0.78 \pm 0.05 (stat)$
	medium	$0.58 \pm 0.02\ (stat)$	$0.58 \pm 0.04 (stat)$
	tight	$0.44 \pm 0.02\ (stat)$	$0.44 \pm 0.03 (stat)$
Track Counting	loose	$0.77 \pm 0.02\ (stat)$	$0.76 \pm 0.04 (stat)$
	medium	$0.58 \pm 0.02\ (stat)$	$0.57 \pm 0.04 (stat)$
	tight	$0.48 \pm 0.02\ (stat)$	$0.44 \pm 0.03 (stat)$

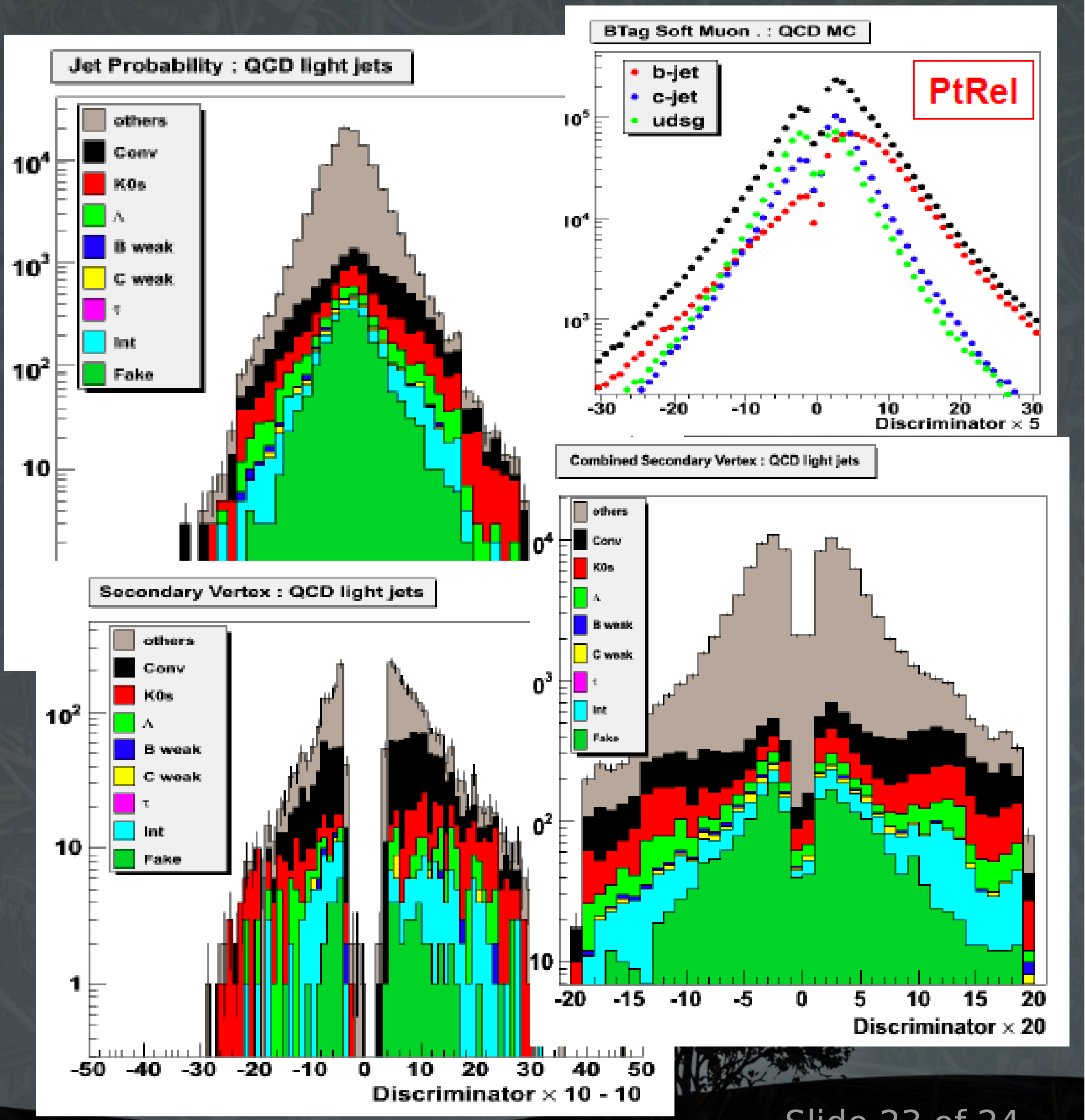
Pedro Ferreira da Silva

Negative tags

- Sample with “small” contamination of *b/c* jets: multi-jet or γ+jets samples
- Negative tagger:
  - Simple taggers, reverse the input (IPS, lifetime)
  - Complex taggers, to be defined more carefully
- Check that it is indeed symmetric on light jets
- (Correct and) use it to extract positive light mistag

$$\varepsilon_{tag}^{Data}(udsg) = R_l \times \varepsilon_{neg}^{Data}(all)$$

$$R_l = \frac{\varepsilon_{tag}^{MCqcd}(udsg)}{\varepsilon_{neg}^{MCqcd}(all)}$$

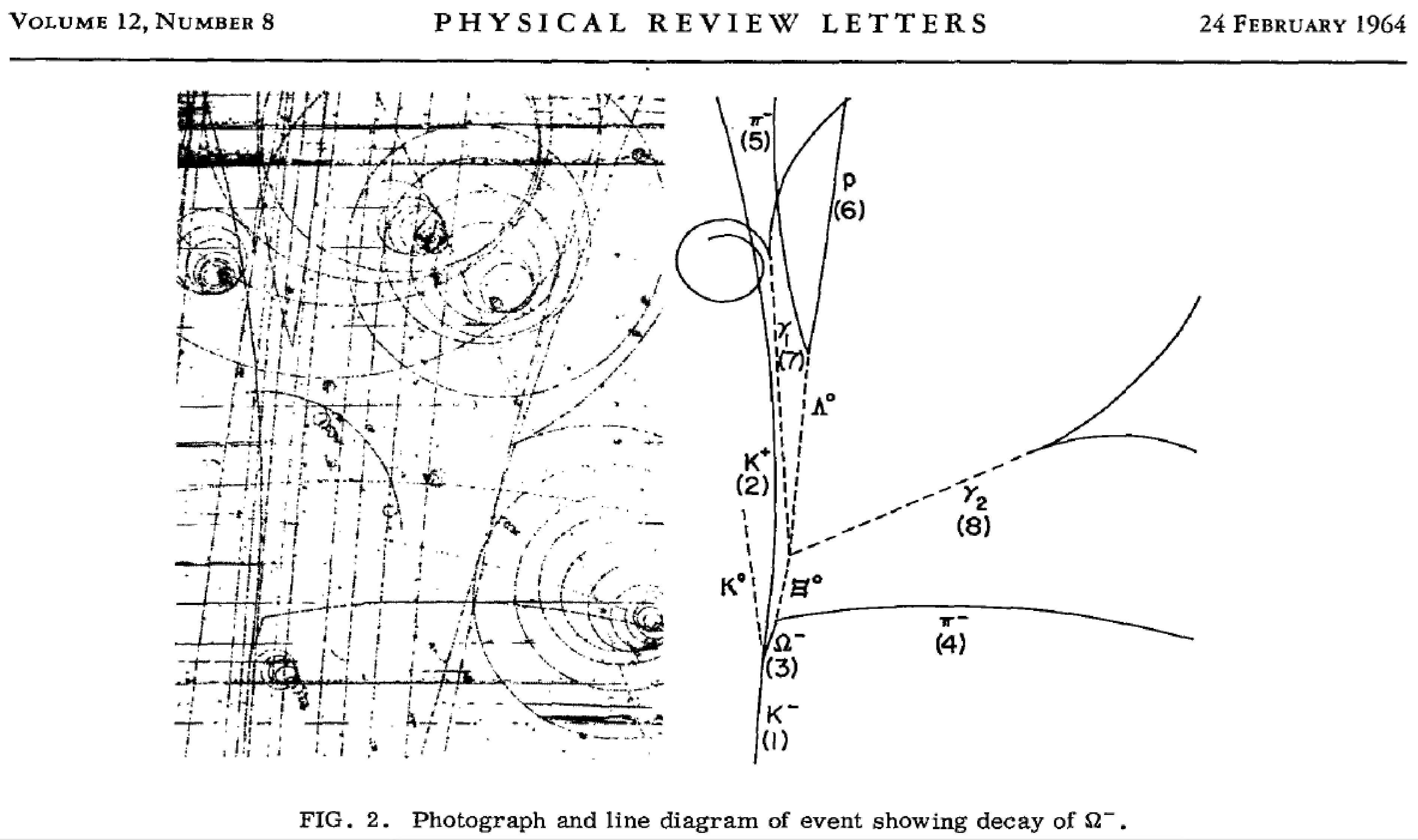


Summary

- Several b taggers are implemented
- Documentation is available
  - Software guide: SWGuideBTagging
  - Workbook: WorkBookBTagging
  - Tutorial: slides (pdf)
- Efficiency measurements
  - Several methods are available
  - Many improvements since CSA07
  - Summer08/Fall08 studies are under way
  - New uncertainty estimates should be available soon

Backup

Omega-minus



VOLUME 12, NUMBER 8 PHYSICAL REVIEW LETTERS 24 FEBRUARY 1964

FIG. 2. Photograph and line diagram of event showing decay of Ω^-.

Tag consistency

More tagged categories:

$$L=-2\log\prod_{i=0}^{4} Poisson(N_i, <N_i>)$$

Include background in the fit:

$$<N_n> = <N_n^{ttbar}> + <N_n^{background}> =$$

$$L\cdot\sigma_{t\bar{t}}\cdot\epsilon_{t\bar{t}}\cdot\left(\sum_{i,j,k} F_{ijk}^{ttbar}\sum_{i'+j'+k'=n}^{i'\le i,\, j'\le j,\, k'\le k}(\ldots)+\frac{\sigma_{background}}{\sigma_{t\bar{t}}}\cdot\frac{\epsilon_{background}}{\epsilon_{t\bar{t}}}\cdot\sum_{i,j,k} F_{ijk}^{background}\sum_{i'+j'+k'=n}^{i'\le i,\, j'\le j,\, k'\le k}(\ldots)\right)$$

Tag consistency method

$$L = -\log\left(Poisson(N_1, <N_1>) \times Poisson(N_2, <N_2>) \times Poisson(N_3, <N_3>)\right)$$

The expected number of tagged jets:

$$<N_n> = \sum_{i,j,k} F_{ijk} \sum_{i'+j'+k'=n}^{i'\le i,\, j'\le j,\, k'\le k} C_i^{i'} \epsilon_b^{i'} (1-\epsilon_b)^{(i-i')} C_j^{j'} \epsilon_c^{j'} (1-\epsilon_c)^{(j-j')} C_k^{k'} \epsilon_b^{k'} (1-\epsilon_b)^{(k-k')}$$

Each F_{ijk} stands for a particular event jet content

Here is an example how F_{113} contributes to $<N_1>$

$$\underline{<N_1>} = F_{113} \times \left(\underline{1\cdot\epsilon_b(1-\epsilon_c)(1-\epsilon_l)^3} + \underline{1\cdot(1-\epsilon_b)\epsilon_c(1-\epsilon_l)^3} + \underline{3\cdot(1-\epsilon_b)(1-\epsilon_c)\epsilon_l(1-\epsilon_l)^2}\right)$$

In other words, the one tagged jet may be the b jet, the c jet or any of the three light jets

Repeat that for all F_{ijk} for $<N_1>$, $<N_2>$ and $<N_3>$, signal and background.

Likelihood ratio method

The fraction of tagged jets in the jet sample:

$$x_{tag}=\epsilon_b x_b+\epsilon_c x_c+\epsilon_l x_l$$

So, the tagging efficiency:

$$\epsilon_b=\frac{1}{x_b}\left[x_{tag}-\epsilon_c x_c-\epsilon_l x_l\right]$$

We define $x_0=x_c+x_l$. Considering $x_b+x_c+x_l=1$, the probability to tag a non-b jet is

$$\epsilon_0 x_0=\epsilon_c x_c+\epsilon_l x_l$$

All this indeed leads us to

$$\epsilon_b=\frac{1}{x_b}\left[x_{tag}-\epsilon_0(1-x_b)\right]$$