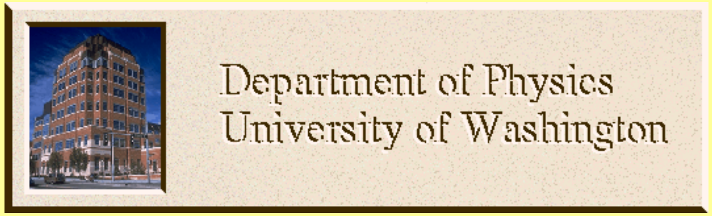
Implementing Jet Algorithms: A Practical Jet Primer

Stephen D. Ellis
University of Washington



West Coast LHC Theory Network
UC Davis December 2006

Outline:

- Jet Jargon
- Big Picture Jet Goals for LHC
- Cone Details & Lessons from the Tevatron
- kT – the hope for the future?
- Jets & BSM issues (at the LHC)
- Summary

See TeV4LHC QCD Report hep-ph/610012

Jet Jargon:

- IR safety – Dave : Thy - cancel singularities, Exp – lower sensitivity to soft stuff

- Cone algorithm – Dave; "stable" cones & "fixed" geometry
- Split/merge issue – Overlapping cones – Dave
- Seeds – IR sensitivity – Dave : fix in data, NOT apply to theory
- Rsep – match NLO Pert Thy to experiment (does NOT break cone)
- JETCLU (Run I CDF) & Ratcheting
- MidPoint Cone Algorithm – A Fix for Run II : Always look for stable cone between 2 stable cone
- Dark Towers – Dave's "Walls": Energetic towers not in any stable cone
- Search Cone Algorithm – a CDF NOT fix in Run II

- kT algorithm – pairwise reconstruction, softest first – Dave
- Underlying Event (UE) and the kT algorithm
- Pile-up – collisions overlapping in time

The Goal at the LHC is a 1% (Precision) Description of Strong Interaction Physics (where Tevatron Run I is ~ 10%)

To this end we want to *precisely* map

- physics at 1 meter, *i.e*., what we can measure in the detector, *e.g*., $E(y,\phi)$

On To

- physics ⑥ 1 fermi, *i.e*., what we can calculate with small numbers of partons, leptons and gauge bosons as functions of E, y, ϕ

We "understand" what happens at the level of short distance partons and leptons, *i.e*., perturbation theory is simple, can reconstruct masses, *etc*.

Thus

We want to map the observed (hadronic) final states onto a representation that mimics the kinematics of the (short-distance) partons; ideally on a event-by-event basis.

But

We know that the (short-distance) partons shower (perturbatively) and hadronize (nonperturbatively), *i.e*., spread out as they evolve from short to long distances, and there *must be* color correlations.

"<u>SOLUTION</u>": associate "nearby" hadrons or partons into JETS via ALGORITHMS, *i.e.*, rules that can be applied to data and theory

- <u>Cone Algorithms</u>, *e.g.*, Snowmass, based on "fixed" geometry (well suited to hadron colliders with UEs)

- <u>k_T Algorithm,</u> based on pairwise merging, nearest, lowest p_T first (familiar at e^+e^- colliders), tends to "vacuum up" soft particles

👍 **Render PertThy IR & Collinear Safe**

👎 **But mapping of hadrons to partons can *never* be 1 to 1, event-by-event!   Colored states ≠ singlet states!**

Goals of IDEAL ALGORITHM (Motherhood)

- *Fully Specified: including defining in detail any preclustering, merging, and splitting issues*
- *Theoretically Well Behaved: the algorithm should be infrared and collinear safe (and insensitive) with no ad hoc clustering parameters (e.g., R_{SEP})*
- *Detector Independence: there should be no dependence on cell type, numbers, or size*
- *Order Independence: The algorithms should behave equally at the parton, particle, and detector levels.*
- *Uniformity: everyone uses the same algorithms*

Defining a Jet with Algorithm-

- Start with a list of particles (4-vectors) and/or calorimeter towers (energies and angles)
- End with lists of particles/towers, one list for each jet
- And a list of particles/towers not in any jet – the spectators – remnants of the initial hadrons not involved in the short distance physics (but there must be some correlations and ambiguity)

<u>Fundamental Issue</u> – Compare Experiments to each other & to Theory

Warning:

<u>We should all use the same algorithm!!</u>
(as closely as humanly possible), *i.e*. both ATLAS & CMS (and theorists).

This is NOT the case at the Tevatron, even in Run II!!

And should NOT be the case if experiments use seeds, etc. – CORRECT for these in data analysis (already correct for detector effects, hadronization)

Observations:

- Iterative Cone Algorithm

  Has detailed issues (merge/split, seeds, dark towers), which only became clear with serious study (and this is a *good* thing)

  And now we know (most of) the issues and can correct for them

- The k_T Algorithm

  May have detailed issues ("vacuum" effect, UE and pile-up sensitivity,..), but much less mature experience at hadron colliders

  We need to find out with the same sort of serious study (history says issues *will* arise)

Run I - *Snowmass Cone Algorithm*

- **Cone Algorithm – particles, calorimeter towers, partons in cone of size R, defined in angular space, *e.g.*, (η,φ)**

- **CONE center - (η^C,φ^C)**

- **CONE i $\in$ *C iff*** $$\sqrt{\left(\eta^i-\eta^C\right)^2+\left(\varphi^i-\varphi^C\right)^2}\le R$$

- **Energy** $$E_T^C=\sum_{i\subset C}E_T^i$$

- **Centroid** $$\overline{\eta}^C=\sum_{i\subset C}E_T^i*\eta^i/E_T^C\ ;\ \overline{\varphi}^C=\sum_{i\subset C}E_T^i*\varphi^i/E_T^C$$

- **Jet is defined by "stable" cone**

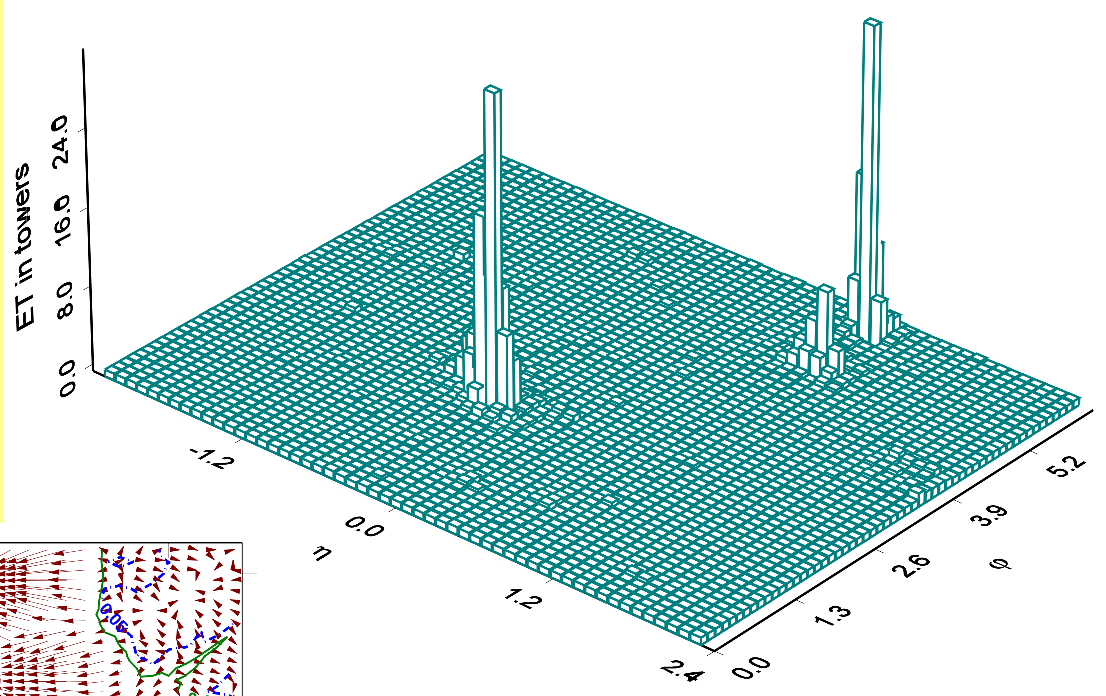
$$\eta^J = \eta^C = \overline{\eta}^C \;\; ; \;\; \varphi^J = \varphi^C = \overline{\varphi}^C \;\; ; \;\; \vec{F}^C = 0$$

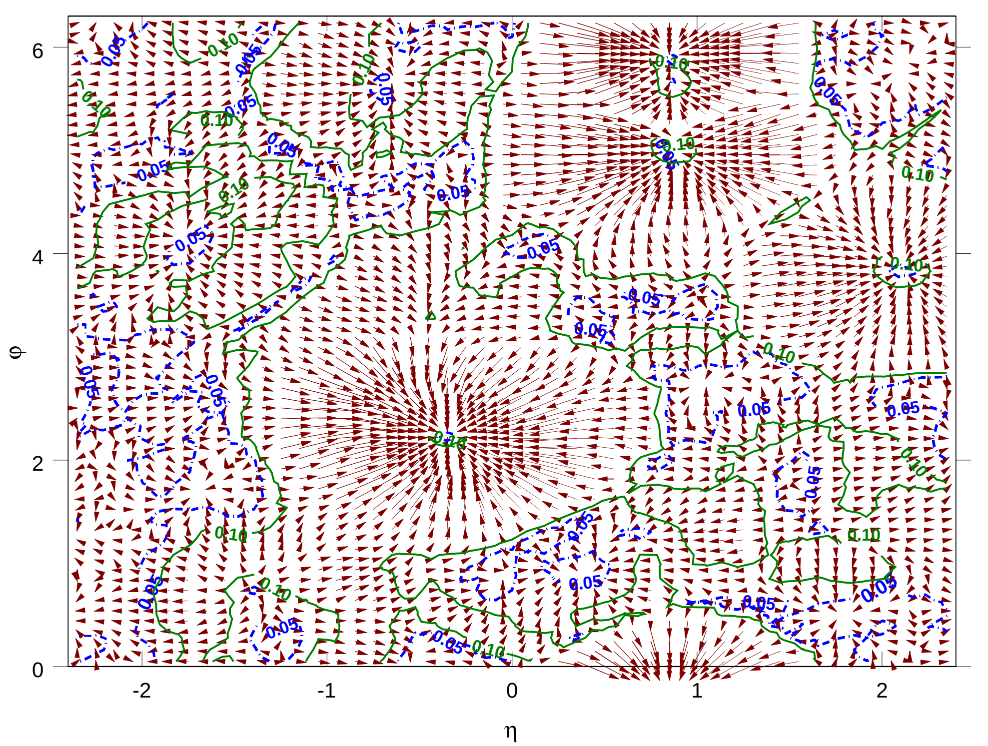
- **Stable cones found by iteration: start with cone anywhere (and, in principle, *everywhere*), calculate the centroid of this cone, put new cone at centroid, iterate until cone stops "flowing", *i.e.*, stable ⇒ Proto-jets (prior to split/merge)**

- ***"Flow vector"*** $\vec{F}^C = \left(\overline{\eta}^C - \eta^C, \overline{\varphi}^C - \varphi^C\right)$

⇒ *unique, discrete jets event-by-event* (at least in principle)

Example Lego & Flow

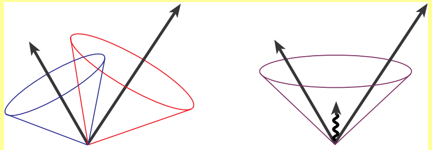




Run I Issues (Life gets more complex):

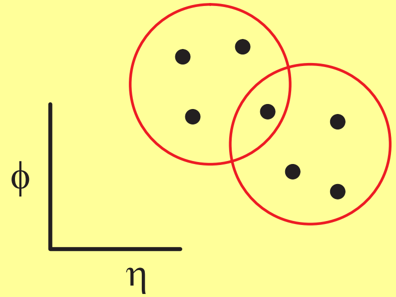
Cone: Seeds – only look for jets under brightest street lights, *i.e.*, near very active regions

⇒ problem for theory, IR sensitive at NNLO



Stable Cones found by iteration (E_T weighted centroid = geometric center) can Overlap,

⇒ require Splitting/Merging scheme merge if share energy fraction > *fmerge* parameter



⇒ Different in different experiments

⇒ Don't find "possible" central jet between two well separated proto-jets (partons)

Cones: Seeds and Sensibility -

- Tension between desire

  To Limit analysis time (for experiments) with seeds

  To Use identical algorithms in data and perturbation theory

- Seeds are intrinsically IR sensitive (MidPoint Fix only for NNLO, not NNNLO)

$\Rightarrow$ DON'T use seeds in perturbation theory, correct for them in data analysis

In the theory they are a big deal – IR UNsafety (Yikes)!!!!!!

In the data seeds vs seedless is a few % correction (e.g., lower the Seed p_T threshold) and this is small compared to other corrections – [Run I jets results are meaningful!!]

To understand these issues consider Snowmass "Potential"

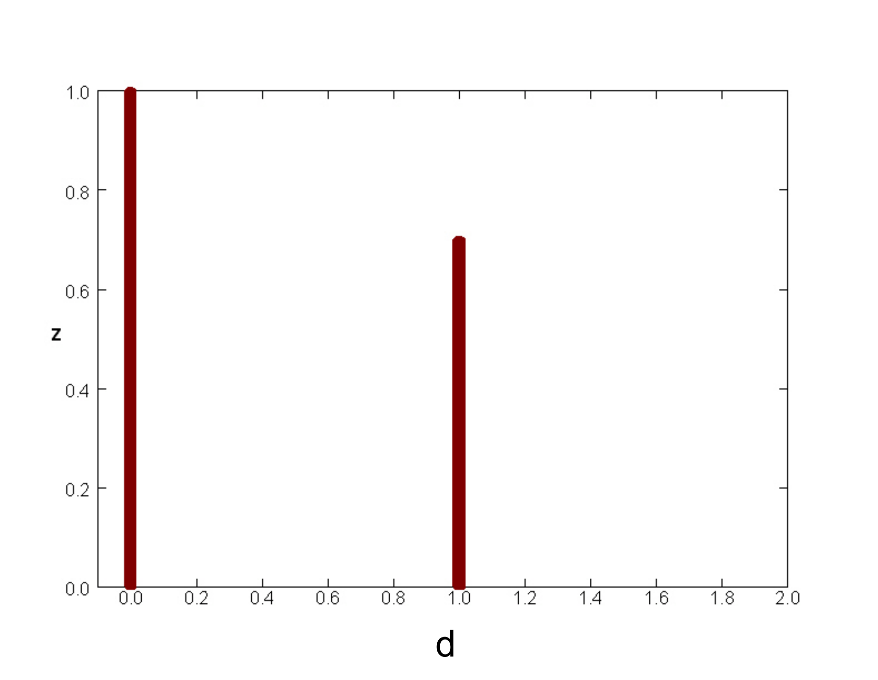
- In terms of 2-D vector $\vec{r} = (\eta, \varphi)$ or (y, φ) define a "potential"

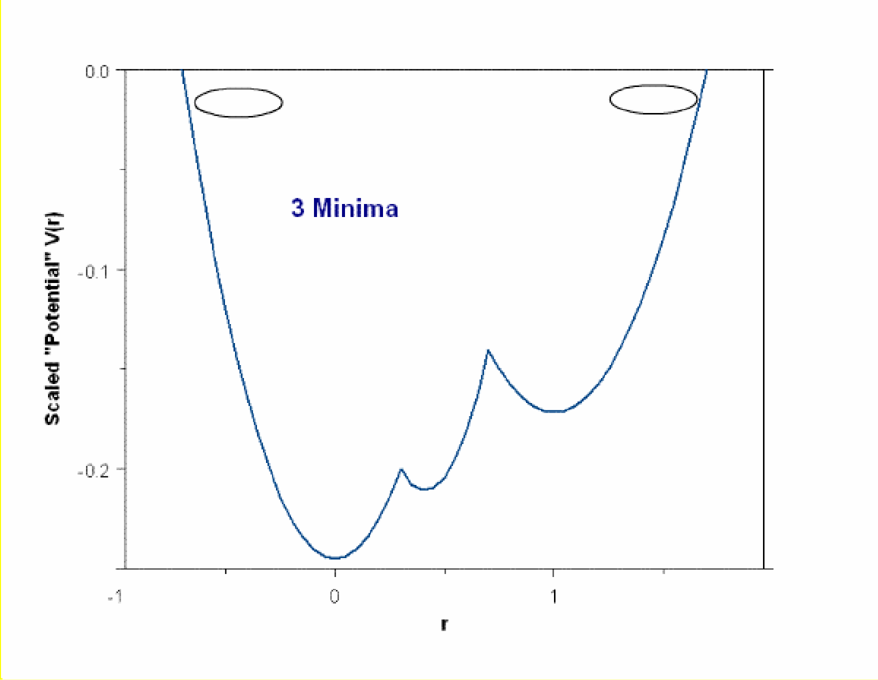
$$V(\vec{r}) \equiv -\frac{1}{2} \sum_i E_T^i \left(R^2 - \left(\vec{r}^{\,i} - \vec{r} \right)^2 \right) \Theta \left(R^2 - \left(\vec{r}^{\,i} - \vec{r} \right)^2 \right)$$

- Extrema are the positions of the stable cones; gradient is "force" that pushes trial cone to the stable cone, *i.e.,* the flow vector

$$\vec{F}(\vec{r}) = -\vec{\nabla} V(\vec{r}) = \sum_i E_T^i \left(\vec{r}^{\,i} - \vec{r} \right) \Theta \left(R^2 - \left(\vec{r}^{\,i} - \vec{r} \right)^2 \right)$$

(THE) Simple Theory Model - 2 partons (separated by d < 2R): yield potential with 3 minima – trial cones will migrate to minima from seeds near original partons ⇒ miss central minimum





$z = p_{\min} / p_{\max}$, d = separation

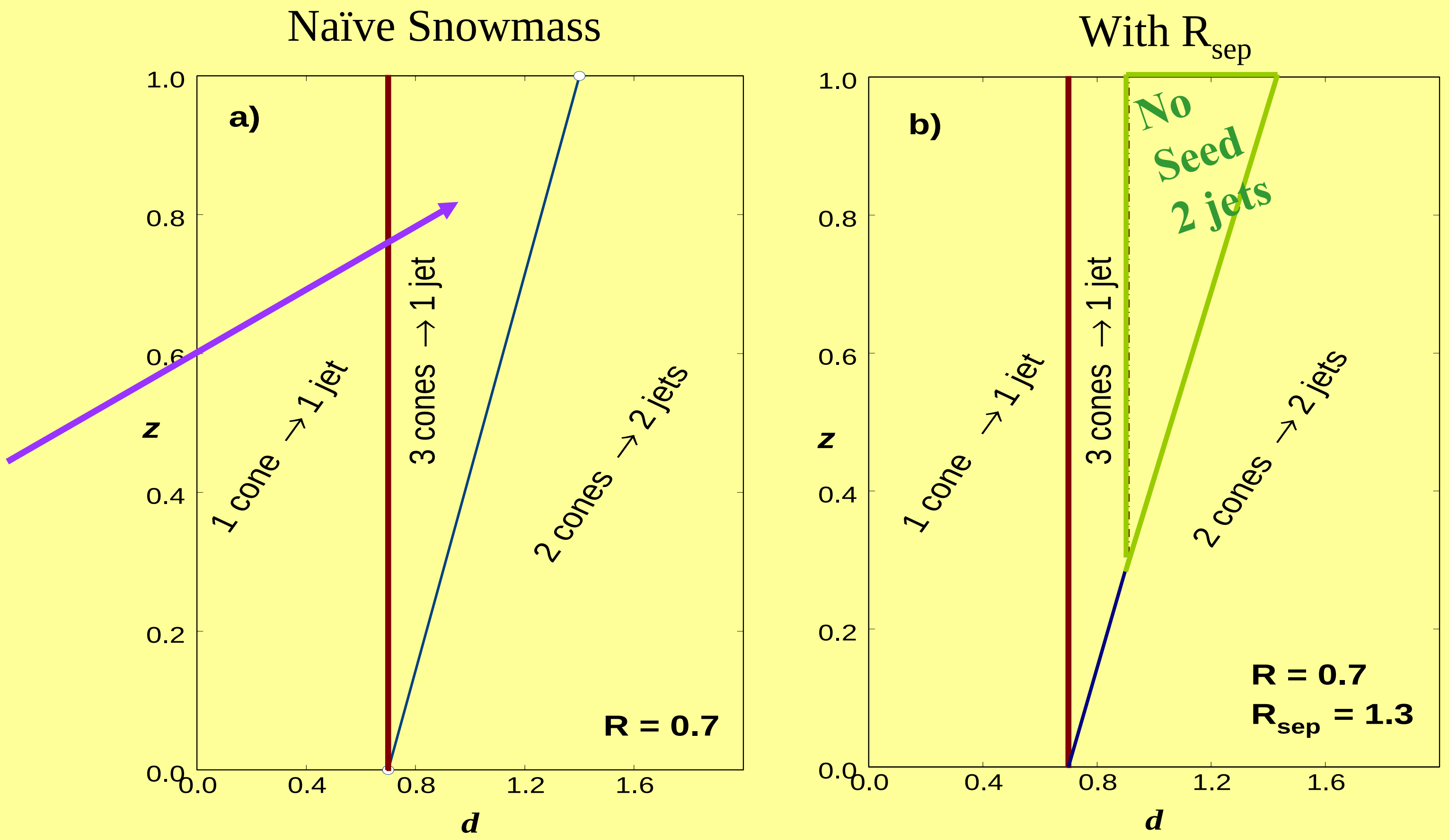
Smearing of order R

Numerical issue:

- Seeds can mean missed configurations with 2 partons in 1 Jet, NLO Perturbation Theory – d = parton separation, $z = p_2/p_1$,,

  Simulate the missed middle cones with R_{sep}

< 10% of cross section here



Run I Cone Issues (Life gets more complex):

3) Kinematic variables:
$E_{T,Snow} \neq E_{T,CDF} \neq E_{T,4D} = p_T$ (5 % differences)

Different in different experiments and in theory

4) Other details –
- Energy Cut on towers kept in analysis (*e.g.*, to avoid noise)
- (Pre)Clustering to find seeds (and distribute "negative energy")
- Energy Cut on precluster towers
- Energy cut on clusters
- Energy cut on seeds kept

5) Starting with seeds find stable cones by iteration, but in JETCLU (CDF), "once in a seed cone, always in a cone", the "ratchet" effect

Detailed Differences mean Differences in:

- UE contributions
- Calorimeter info *vs* tracking info
- Non-perturbative hadronization (& showering) compared to PertThy
- (Potential) Impact of Higher orders in perturbation theory
- Mass reconstruction

To address these issues, the Run II Study group Recommended

<u>Both</u> experiments use

- (legacy) Midpoint Algorithm – always look for stable cone at midpoint between found cones
- Seedless Algorithm
- k_T Algorithms

- Use identical versions except for issues required by physical differences (in preclustering??)

- Use (4-vector) E-scheme variables for jet ID and recombination

A NEW issue for Iterative Cone Algorithms – DARK TOWERS (Dave's Walls)

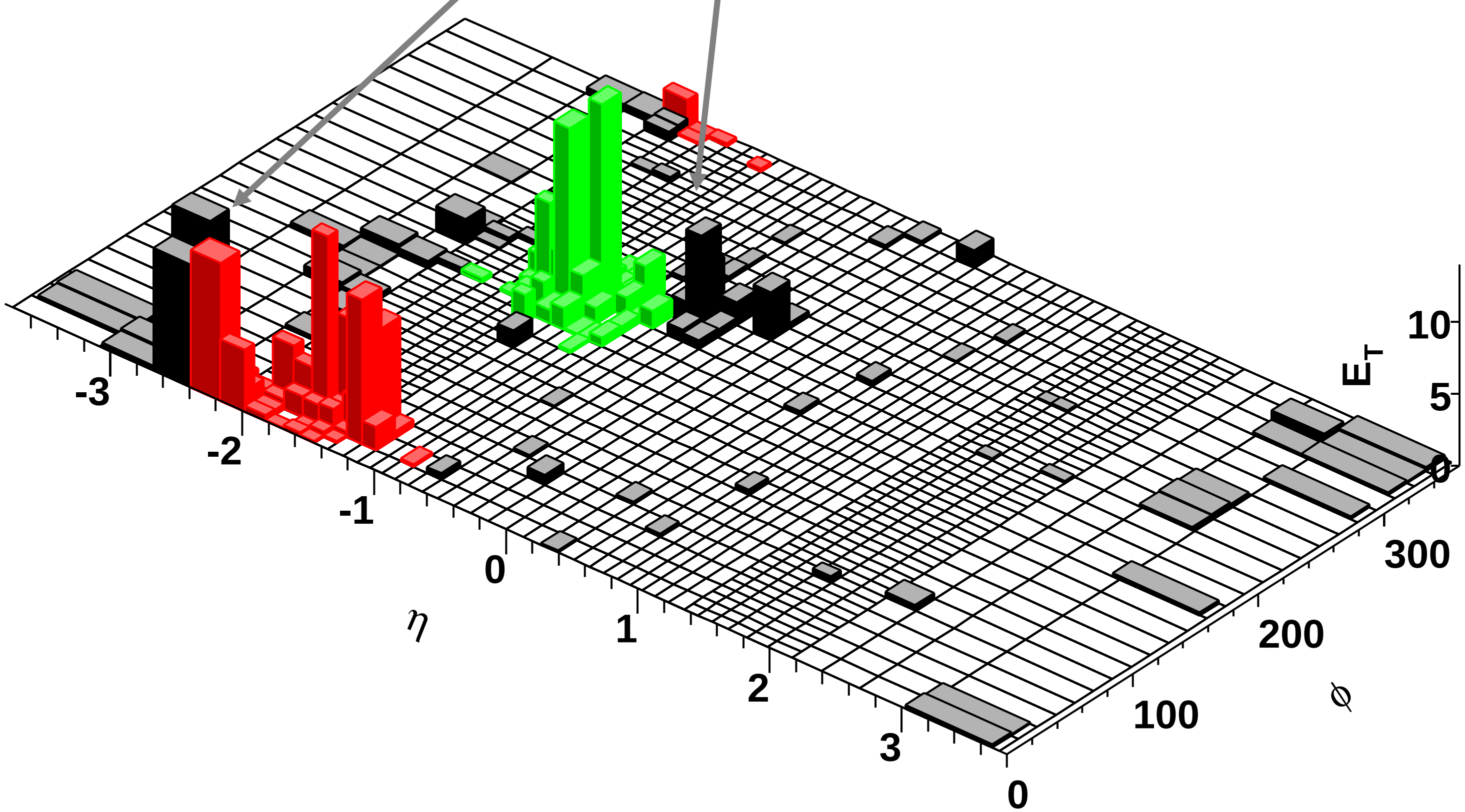
- Compare jets found by JETCLU (with ratcheting) to those found by MidPoint and Seedless Algorithms

- "Missed Energy" – when energy is smeared by showering/hadronization do not always find stable cones expected from perturbation theory

  ⇒ 2 partons in 1 cone solutions
  ⇒ or even second cone

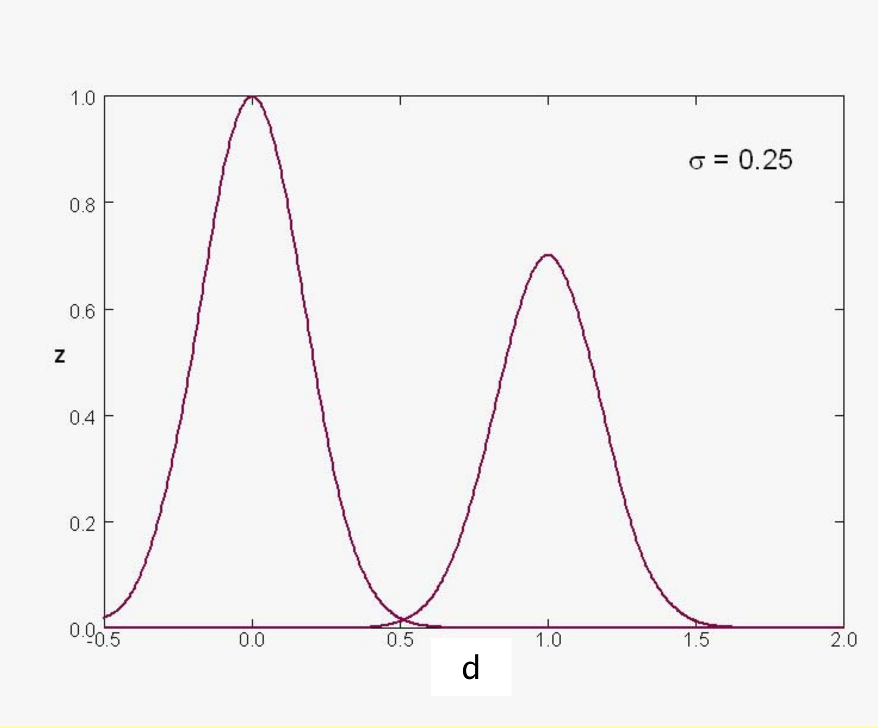
  Under-estimate E_T – new kind of Splashout

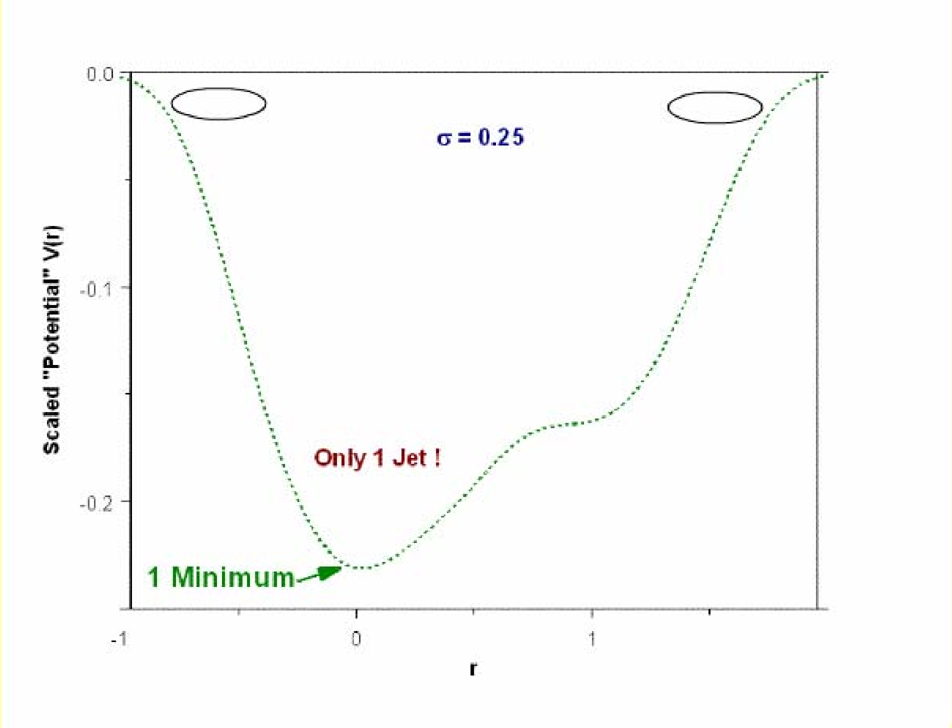
Missed or Dark Towers (not in any stable cone) – How can that happen? (Dave's issue with "walls")



Results from M. Tönnesmann

Why Dark towers? Include smearing (~ showering & hadronization) in simple picture, find only 1 stable cone (no midpoint stable cone & dark towers)



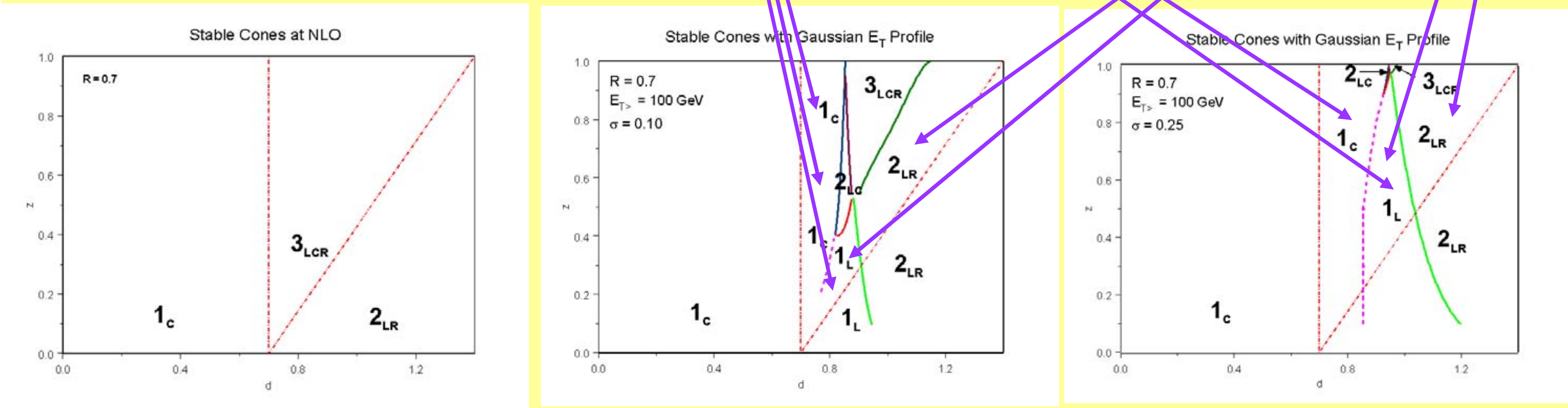


Compare with smearing: MidPoint will still miss 2-in-1 Jets (R_{sep} < 2)

Dark towers (no R stable cone)

Missing MidPoint (no C stable cone)

σ = 0 σ = 0.1 σ = 0.25



Proposed Fix with smaller radius *Search Cone* – Used by CDF

- Over compensates with (too) many found stable cones, so use larger f_merge (f_CDF > f_D0)

- (Re)Introduces IR-sensitivity through soft stable search cones (R' < R) that, when expanded to R, can envelop and merge nearby pairs of energetic partons, which themselves do not correspond to a stable cone (R)

- NOT A COMPLETE SOLUTION!!

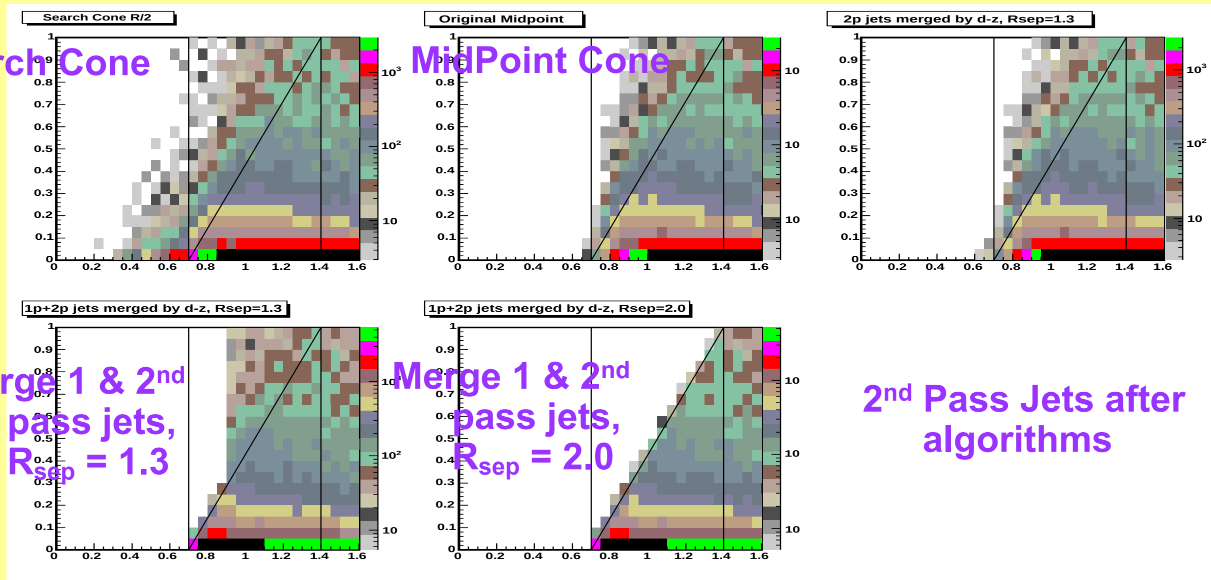
Better(?) - Consider a Dark Tower Correction based on Comparison to pQCD

- Take multiple passes at data

  1st pass jets = found by Cone Algorithm
  2nd pass jets = missed by Cone Algorithm (but found if remove 1st pass jet)

- Merge if in correct region of (d, z) plane (?)

  ⇒ Correct to data!



The k_T Algorithm

- Merge partons, particles or towers pair-wise based on "closeness" defined by minimum value of

$$d_{ij}^2 \equiv \mathrm{Min}\left[p_{T,i}^2, p_{T,j}^2 \right] \frac{\left(y_i - y_j\right)^2 + \left(\phi_i - \phi_j\right)^2}{D^2}, d_i^2 = p_{T,i}^2$$

  If d_{ij}^2 is the minimum, merge pair and redo list;
  If d_i^2 is the minimum -> i is a jet! (no more merging for i),

  1 parameter D (?), at NLO R = 0.7, R_{sep} = 1.3 ⇔ D = 0.83

- Jet identification is unique – no merge/split stage

- Resulting jets are more amorphous, energy calibration difficult (subtraction for UE?), and analysis can be very computer intensive (time grows like N^3, recalculate list after each merge)
  But new version (Cacciari & Salam) goes like N ln N (only recalculate nearest neighbors)

In the future: (comments, not criticisms)

- When we look carefully will we find problems and add details ? History says yes! (See below)

- The (official?) k_T webpage has 5 parameters to specify the implementation, resolution variable, combination scheme, etc.

- Recall the Cambridge k_T (e^+e^-) algorithm that added angular ordering to get rid of “junk jets” (resolution variable 🕒 ordering variable) and “soft-freezing” to reduce mis-clustering

Jet Algorithm Summary:

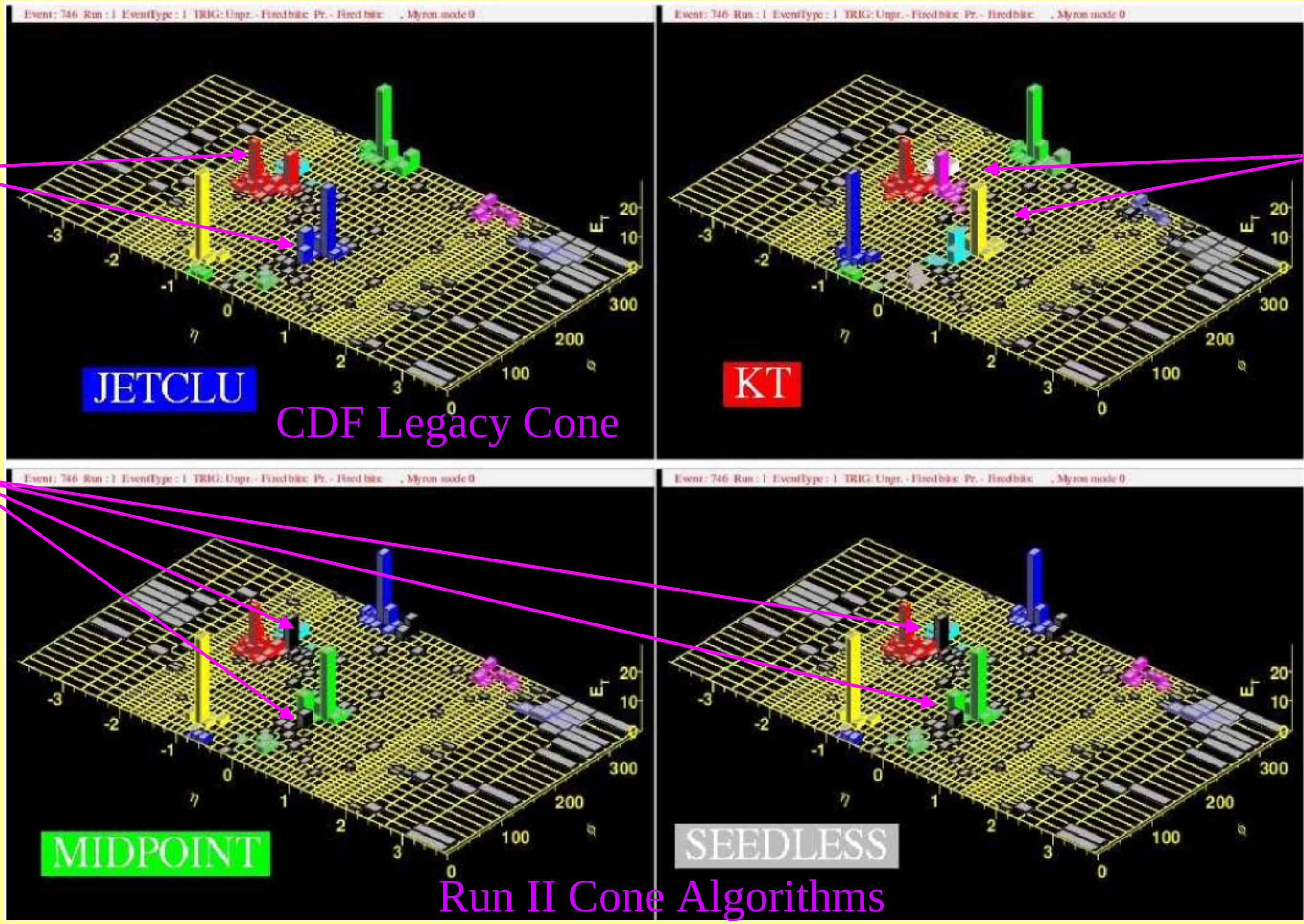
- Seeds & pQCD are a bad mix (not IRS). **It is better to correct for seeds during the analysis of the data (a small correction) and compare to theory w/o seeds (so no IRS issue) !!**
- Dark towers are a real 5 - 10% effect, but the search cone fix aggravates the IRS issue – better to recognize as a correction during the analysis of the data (or the theory), along with corrections for detector, UE, hadronization, seeds, and missing 2-in-1 configurations
- Compare corrected experimental numbers to pQCD without seeds and R_{sep} = 2
- Need serious phenomenology study of the k_T algorithm

Same Event – slightly different jets

Merged jets

UN Merged jets

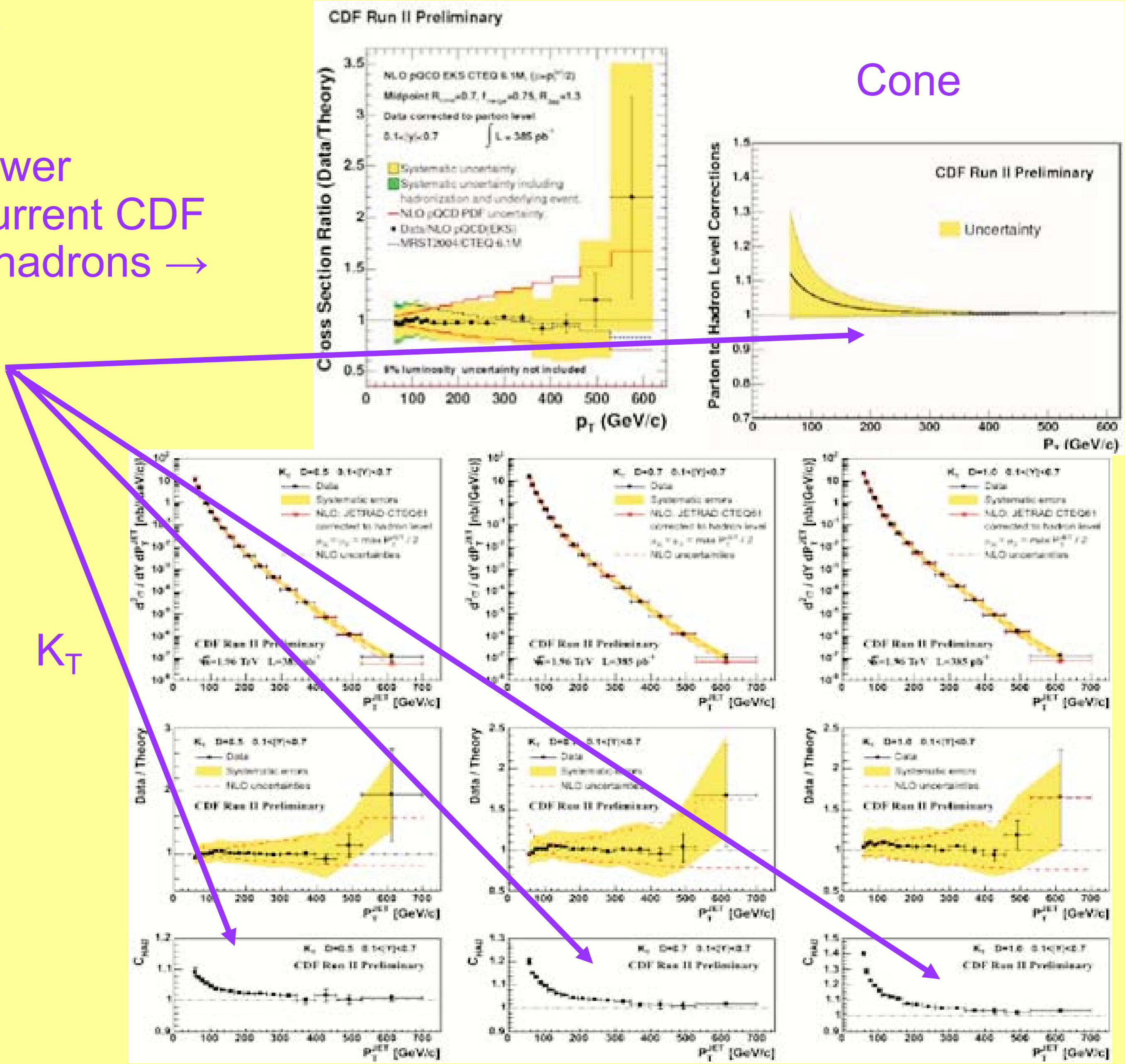
Dark towers

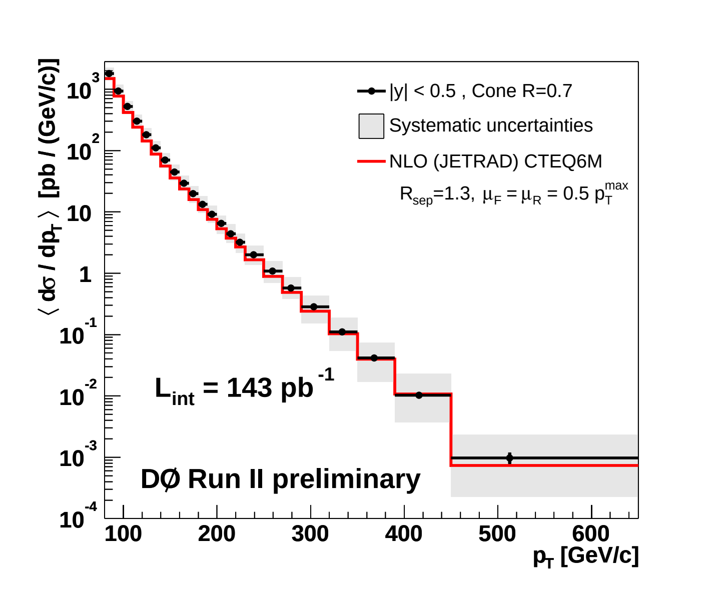


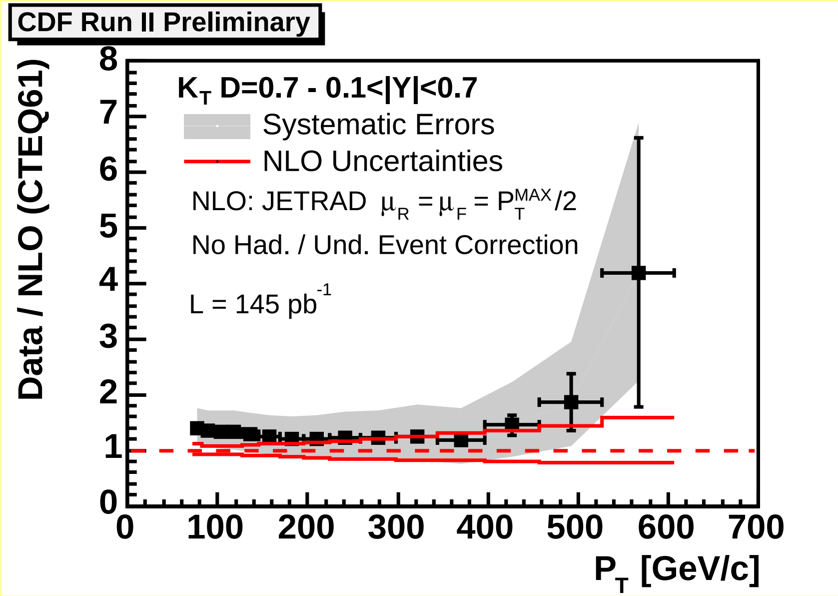
Corrections

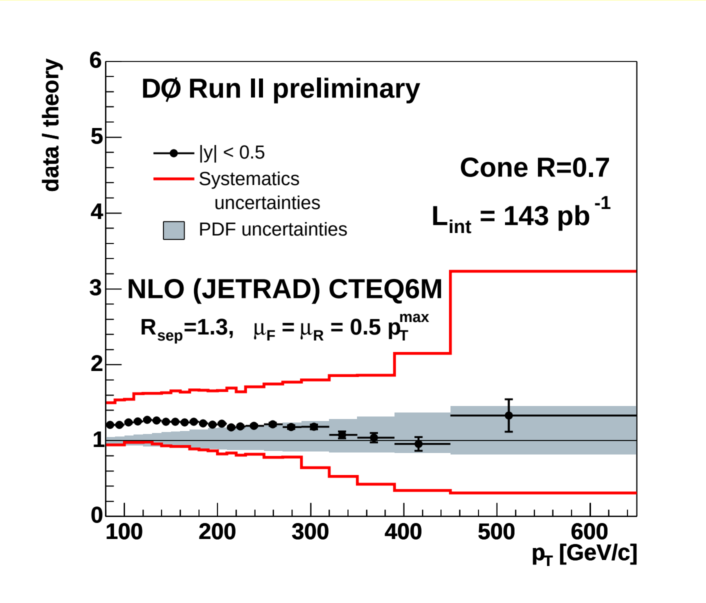
Seed and Dark Tower corrections ≤ current CDF corrections for hadrons → partons

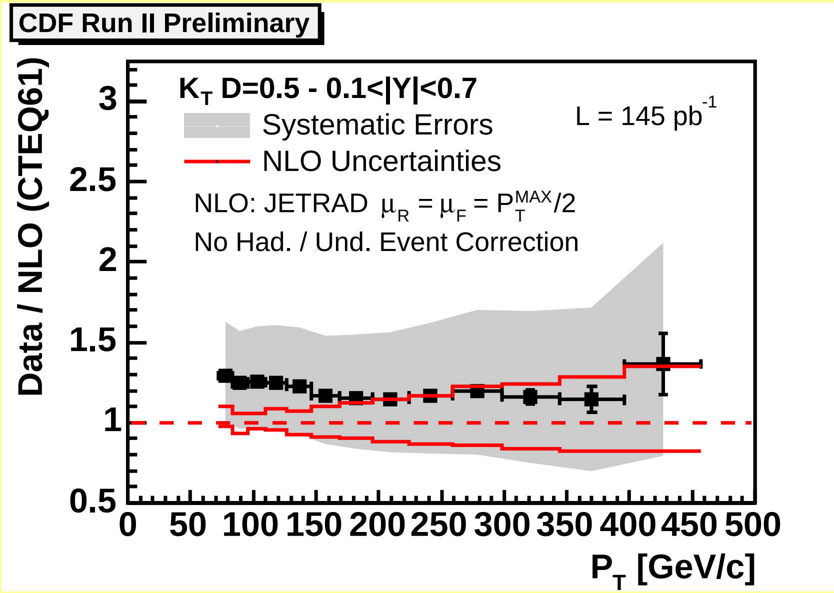
K_T




|y| < 0.5 , Cone R=0.7
Systematic uncertainties
NLO (JETRAD) CTEQ6M
R_sep=1.3, μ_F = μ_R = 0.5 p_T^max
L_int = 143 pb^-1
DØ Run II preliminary
⟨ dσ / dp_T ⟩ [pb / (GeV/c)]
p_T [GeV/c]


CDF Run II Preliminary
K_T D=0.7 - 0.1<|Y|<0.7
Systematic Errors
NLO Uncertainties
NLO: JETRAD μ_R = μ_F = P_T^MAX/2
No Had. / Und. Event Correction
L = 145 pb^-1
Data / NLO (CTEQ61)
P_T [GeV/c]


DØ Run II preliminary
|y| < 0.5
Systematics uncertainties
PDF uncertainties
Cone R=0.7
L_int = 143 pb^-1
NLO (JETRAD) CTEQ6M
R_sep=1.3, μ_F = μ_R = 0.5 p_T^max
data / theory
p_T [GeV/c]


CDF Run II Preliminary
K_T D=0.5 - 0.1<|Y|<0.7
Systematic Errors
NLO Uncertainties
L = 145 pb^-1
NLO: JETRAD μ_R = μ_F = P_T^MAX/2
No Had. / Und. Event Correction
Data / NLO (CTEQ61)
P_T [GeV/c]

Goals at LHC Different $\Rightarrow$ Different Figure of Merit for Jet algorithm?

- Find Physics Beyond the Standard Model
- Event structure likely different from QCD, more jets? Overlap? Different structure within jets?
- Want to select on non-QCD-ness
- Highly boosted SM particles – W, Z, top $\Rightarrow$ single jet instead of 2 or 3 jets, focus on substructure in jets

LHC and BSM Goals

- Many questions, but some answers from LHC Olympics $\Rightarrow$ learn about phenomenological challenges of LHC (a pedagogical tool)

  Study “Black Boxes” (BB) of simulated events containing unknown BSM signal that has been processed by realistic detector simulation (PGS), *i.e.*, events are lists of (sometimes mis-IDed and mis-measured) objects (leptons, photons, jets & MET)

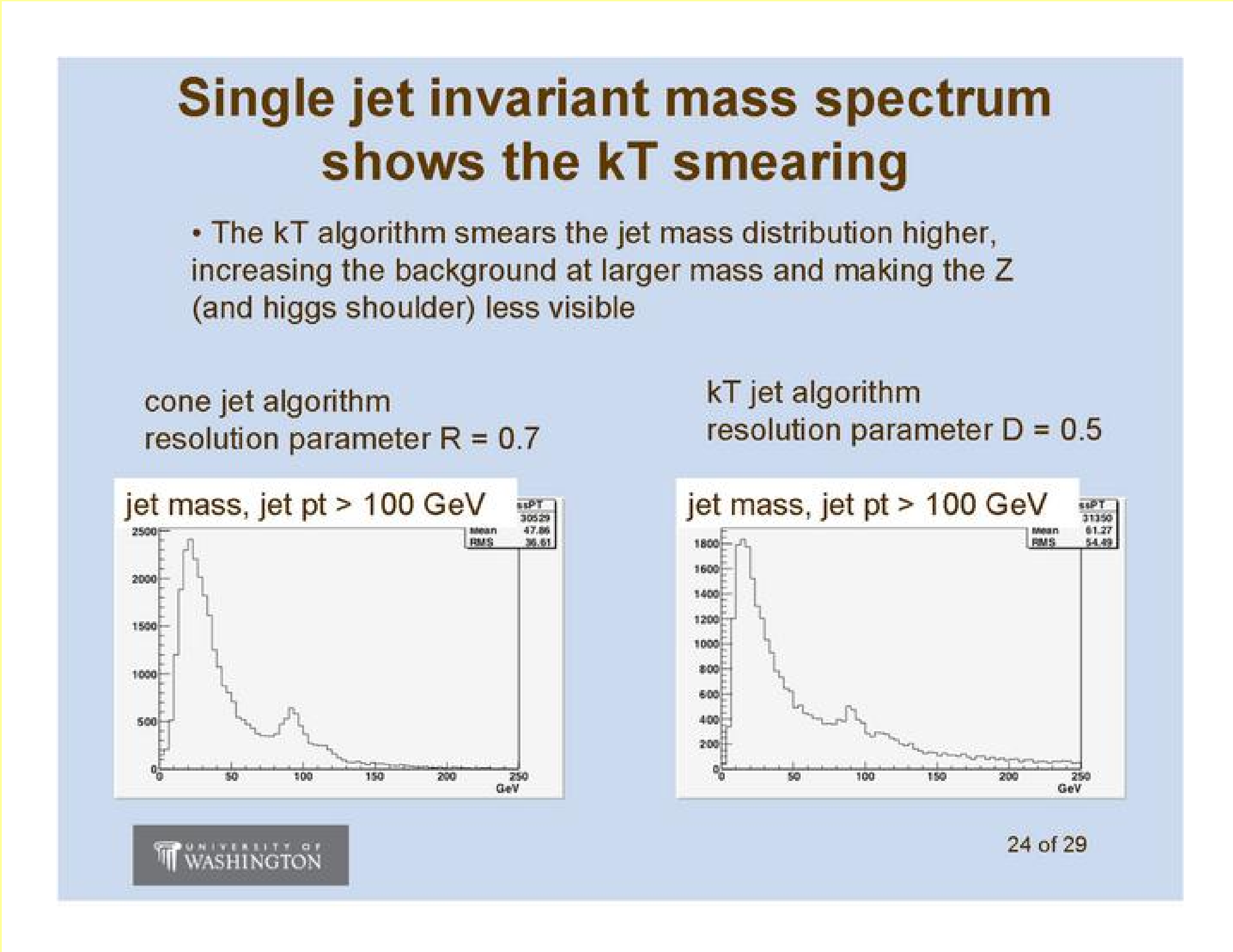
  Try to ID the new physics – difficult even when no real SM background

  Jets play central role and PGS 3.0 used cone jets, while PGS 4.0 uses kT jets - compare

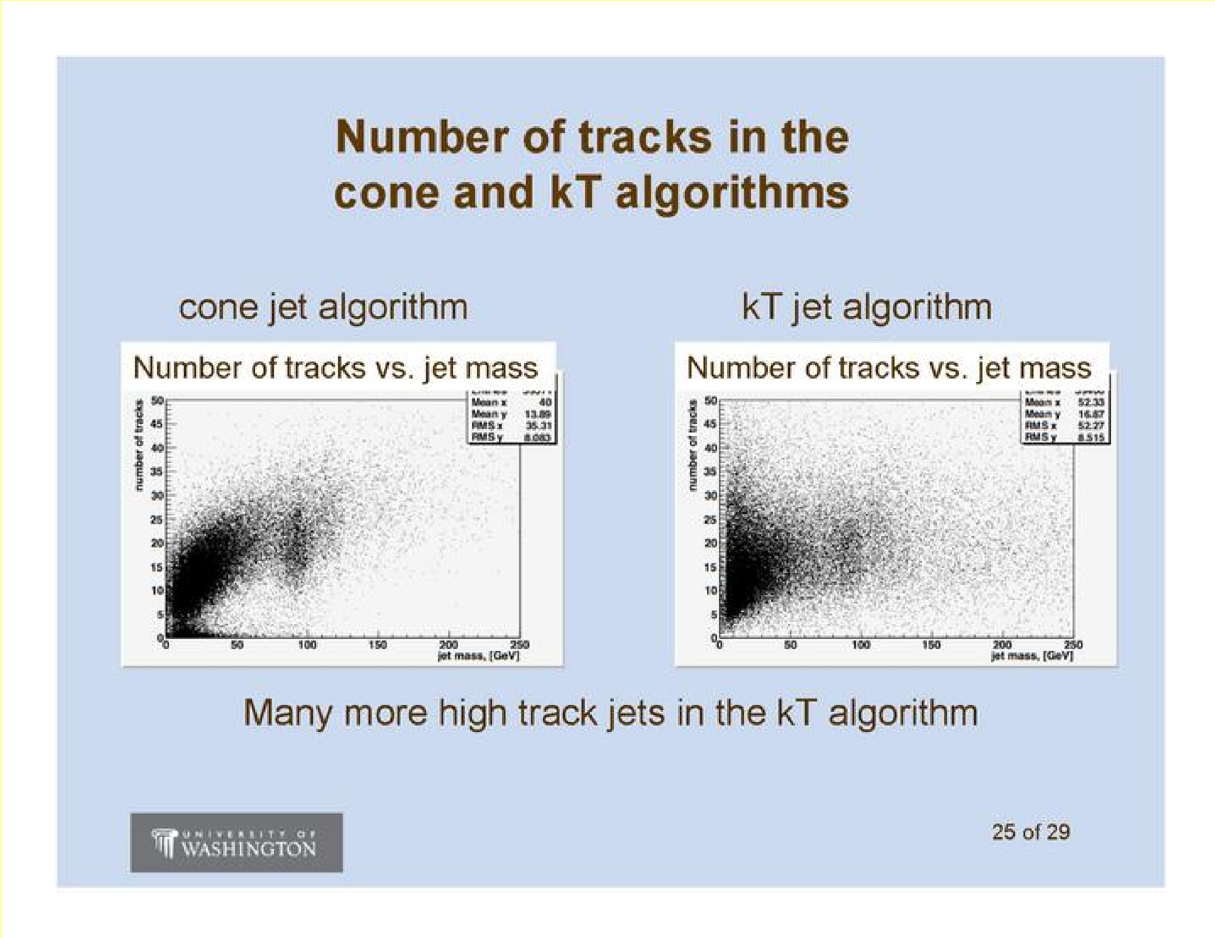
Interesting comparison in context of LHC Olympics – new physics at few TeV scale means highly boosted particles decay into 1, instead of 2 (or more jets)

From Jon Walsh at KITP

UW BB with 2 kinds of jets



Larger fluctuations in jet properties (# of charged tracks) with kT algorithm



LHC environment -

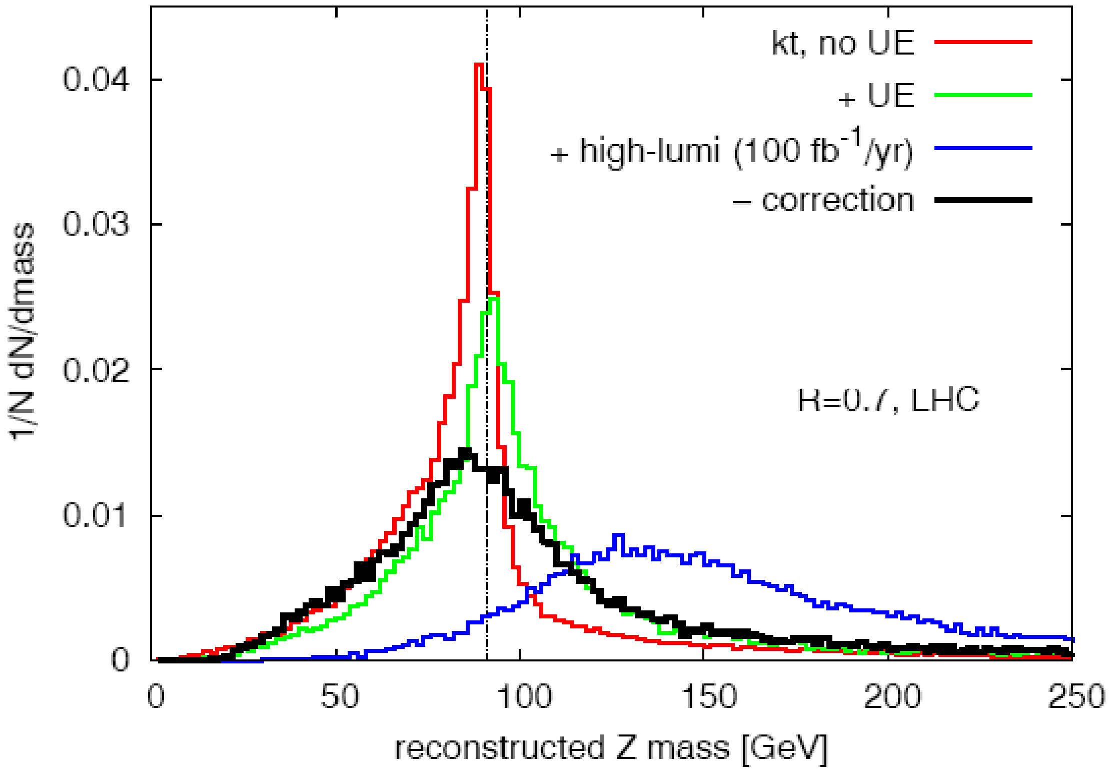
May be much “noisier” at the LHC

- Enhanced UE ?

- Pile-up at large Luminosity – multiple events in each time bucket (most min-bias)

Studies from Matteo Cacciari & Gavin Salam

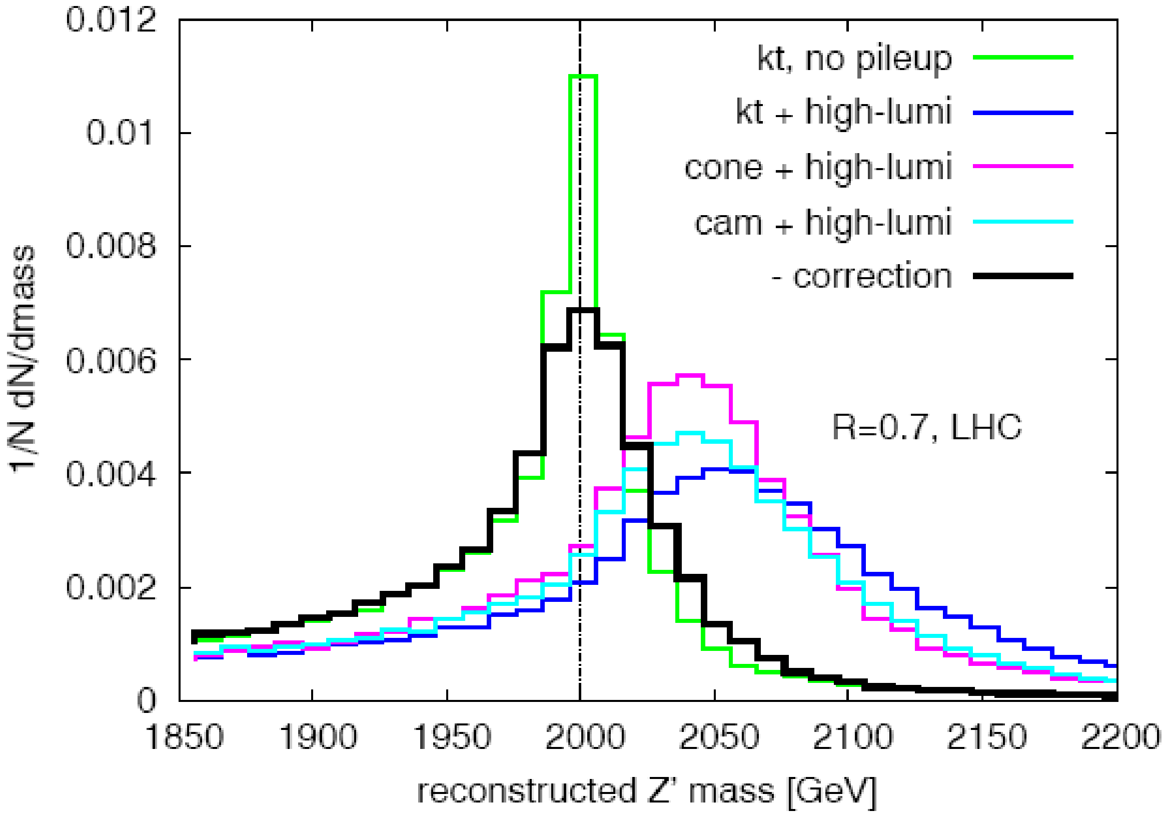
Try reconstructing M_Z from $Z \to 2$ jets, with subtraction of UE/MB



Some loss in resolution, but good value for the Z mass

Talk at MC@LHC 7/2006

Z' reconstruction – can fix with detailed jet-by-jet analysis! Need to verify can do this in real detector, *i.e.*, measure jet area



Uncorrected cone better than k_t.

Cam is intermediate ($\langle A_{cam}\rangle \simeq \langle A_{cone}\rangle$, but fluctuations larger)

Corrected Cam (and k_t) is best.

If New Physics $\Rightarrow$ New Jet Structure

- E.g., Produce particles in separate sector of theory, The Hidden Valley of Strassler

  [hep-ph/0607160, hep-ph/0605193, hep-ph/0604261]

  Decay back into SM particles with

  More jets

  Enhanced heavy flavor

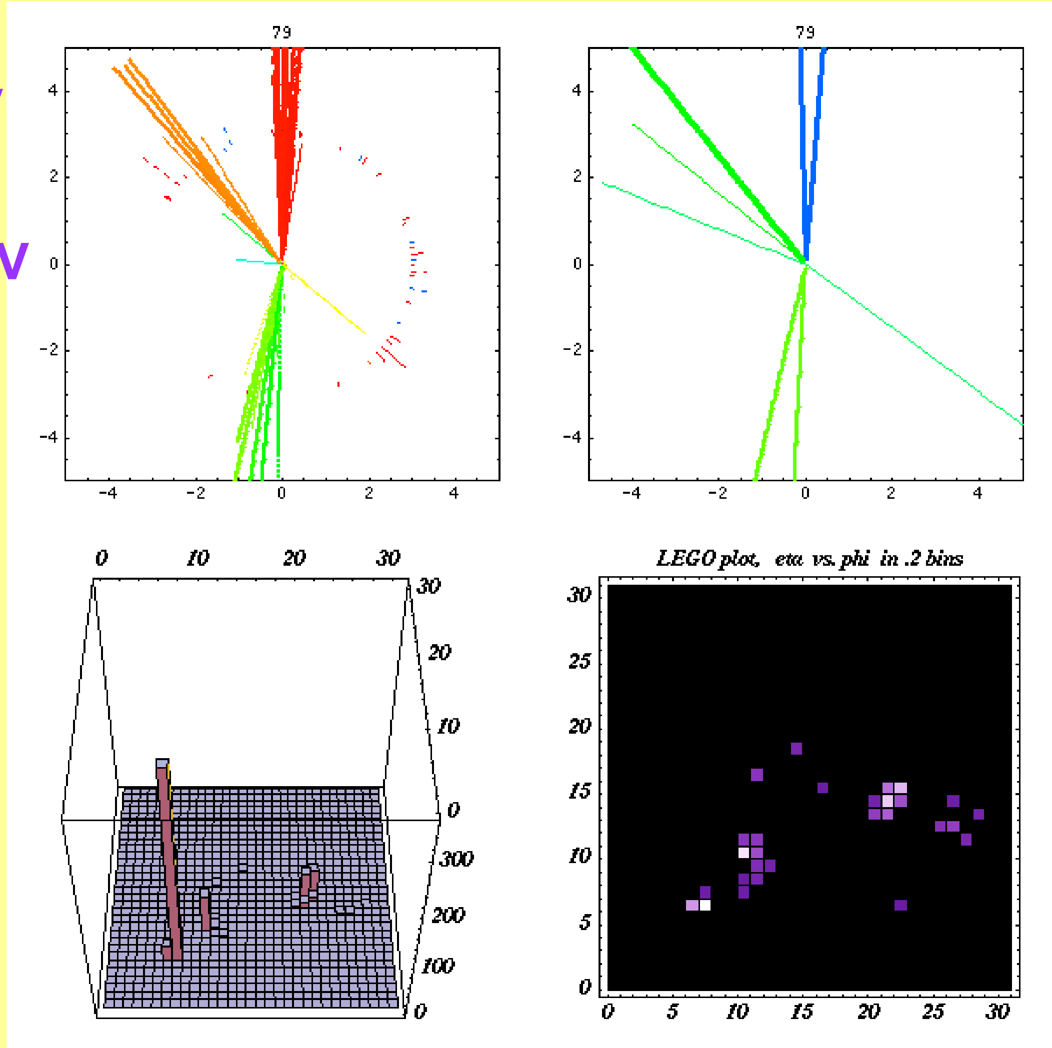
  Displaced vertices (if long lifetimes)

Simulated (Strassler) Events – many b's & jets

M_Z' = 3 TeV

M_Vpion = 30 GeV

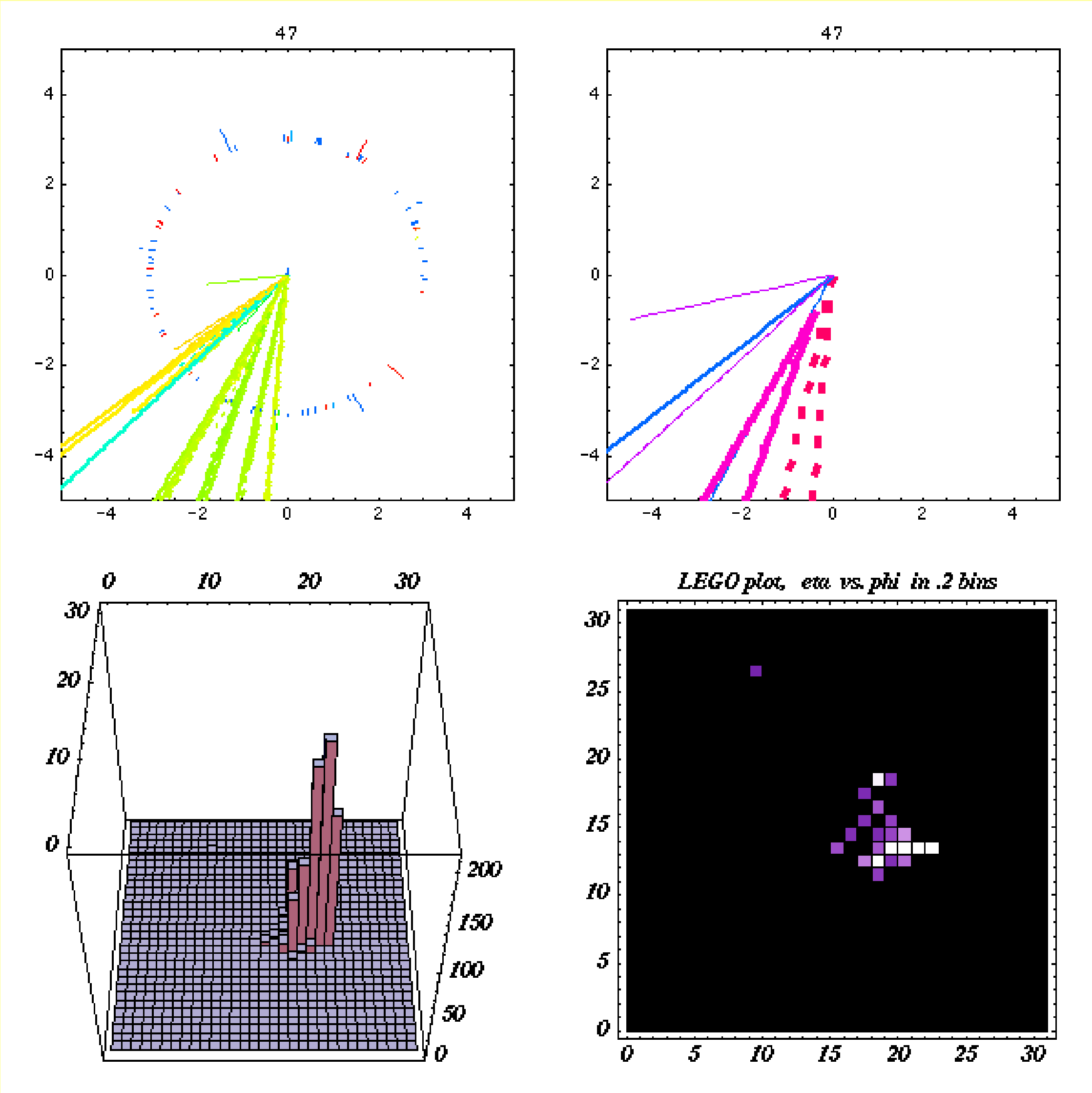
3 jets



8 b's

Hidden Valley is 2-flavor QCD-like

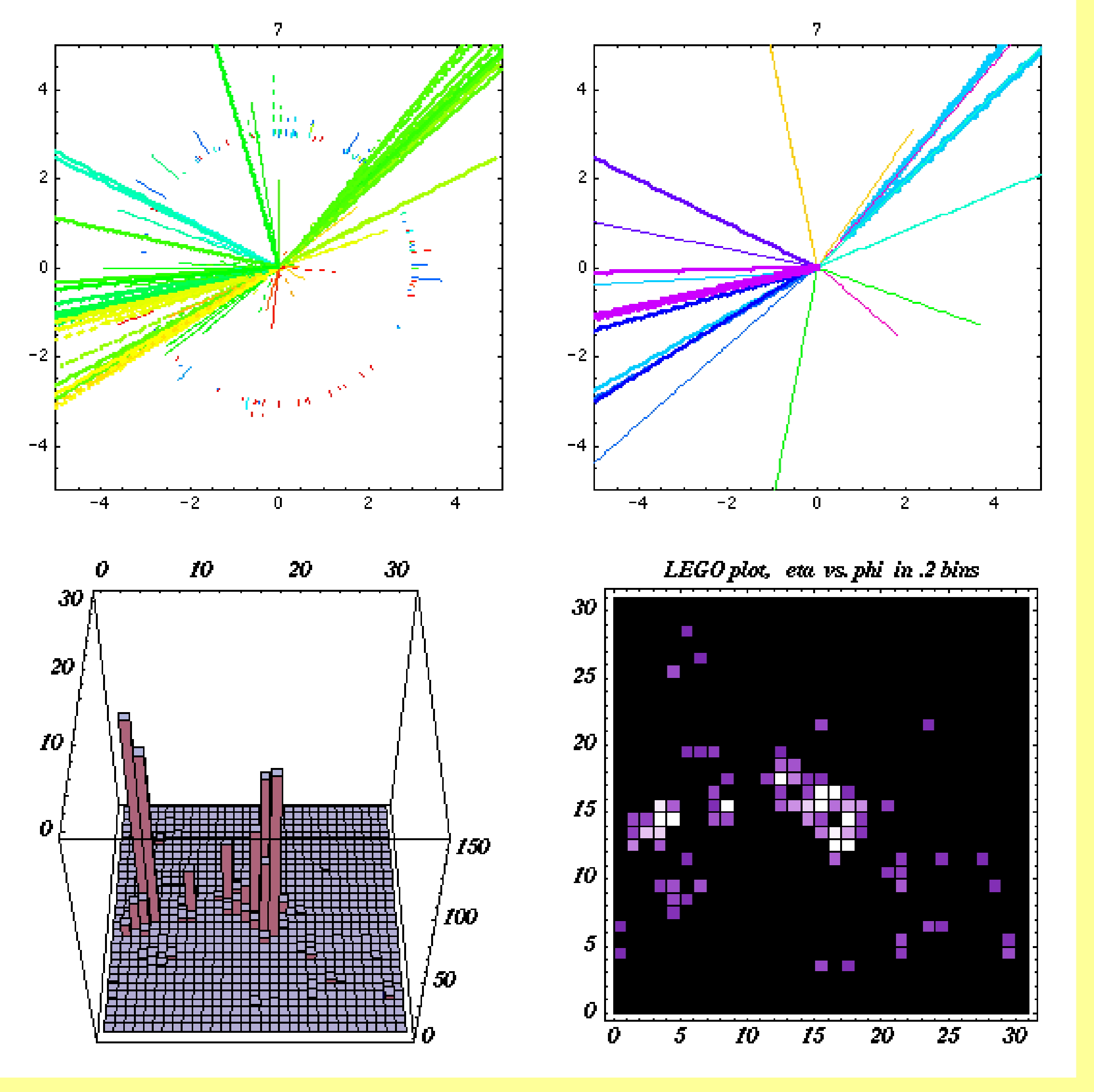
Some with taus & Missing ET



6 b's & 2 taus

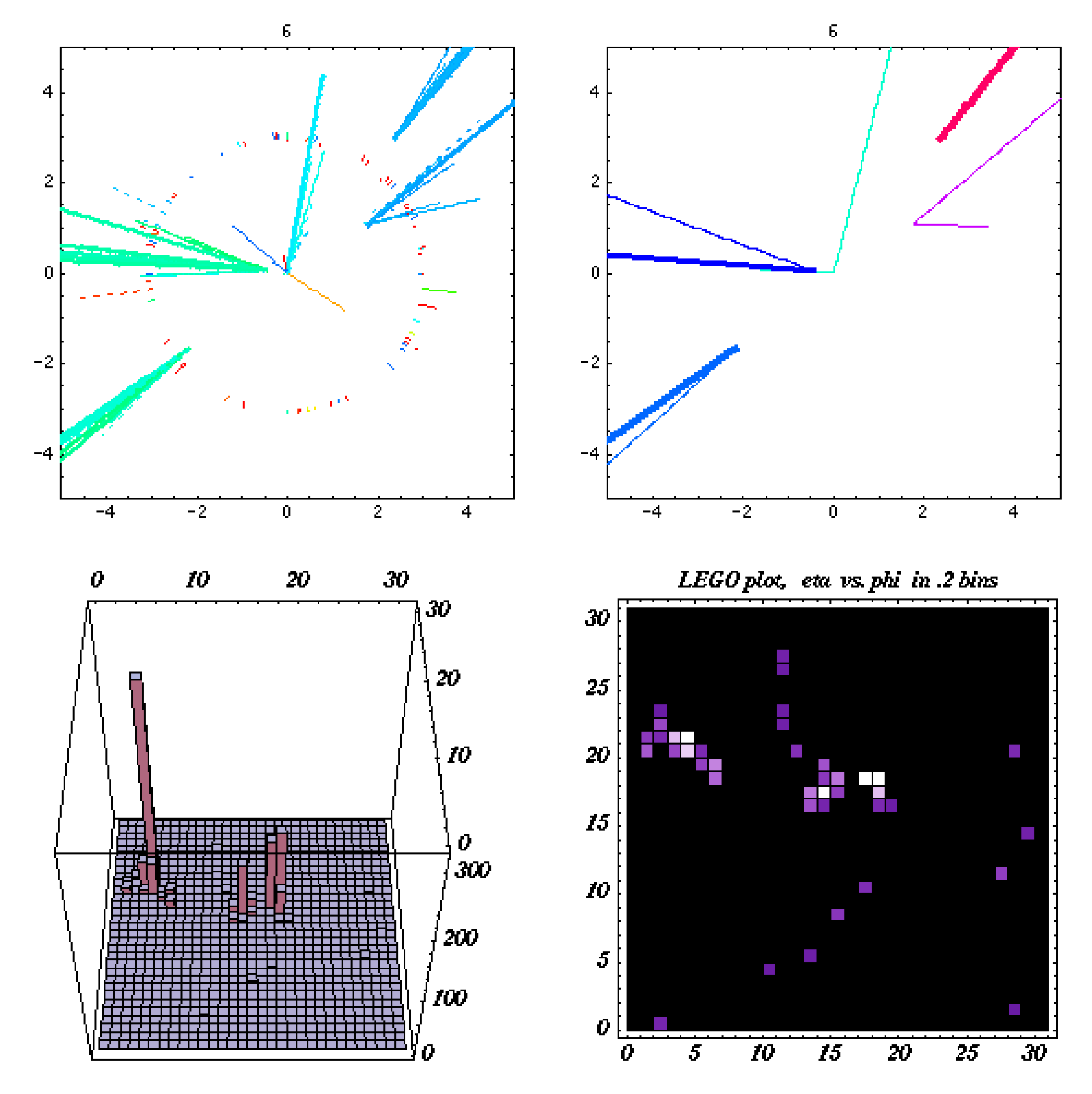
1 jet?

More b's & messy jets



20 b's

Displaced Jet Vertices



Summary -

- Iterative Cone jets have many issues, but they are the devils we know and can (largely) correct for.
- kT jets do not exhibit these devils, but may have their own, especially in the noisy LHC world. Can we learn to correct for them?
- Can we tell SM jets from BSM jets? Is the sub-jet structure the answer?
- Do we need a different analysis tool?

Extra Detail Slides

Dictionary of Hadron Collider Terminology

EVENT

HADRON-HADRON COLLISION

Primary (Hard) Parton-Parton Scattering

Initial-State Radiation (ISR) = Spacelike Showers associated with Hard Scattering

Underlying Event

Multiple Parton-Parton Interactions: Additional parton-parton collisions (in principle with showers etc) in the same hadron-hadron collision.

= Multiple Perturbative Interactions (MPI)

= Spectator Interactions

Beam Remnants: Left over hadron remnants from the incoming beams. Colored and hence correlated with the rest of the event →

Fragmentation

Perturbative:

Final-State Radiation (FSR)

= Timelike Showers

= Jet Broadening and Hard Final-State Bremsstrahlung

Non-perturbative:

String / Cluster Hadronization

(Color Reconnections?)

PILE-UP: Additional hadron-hadron collisions recorded as part of the same event.

From Peter Skands

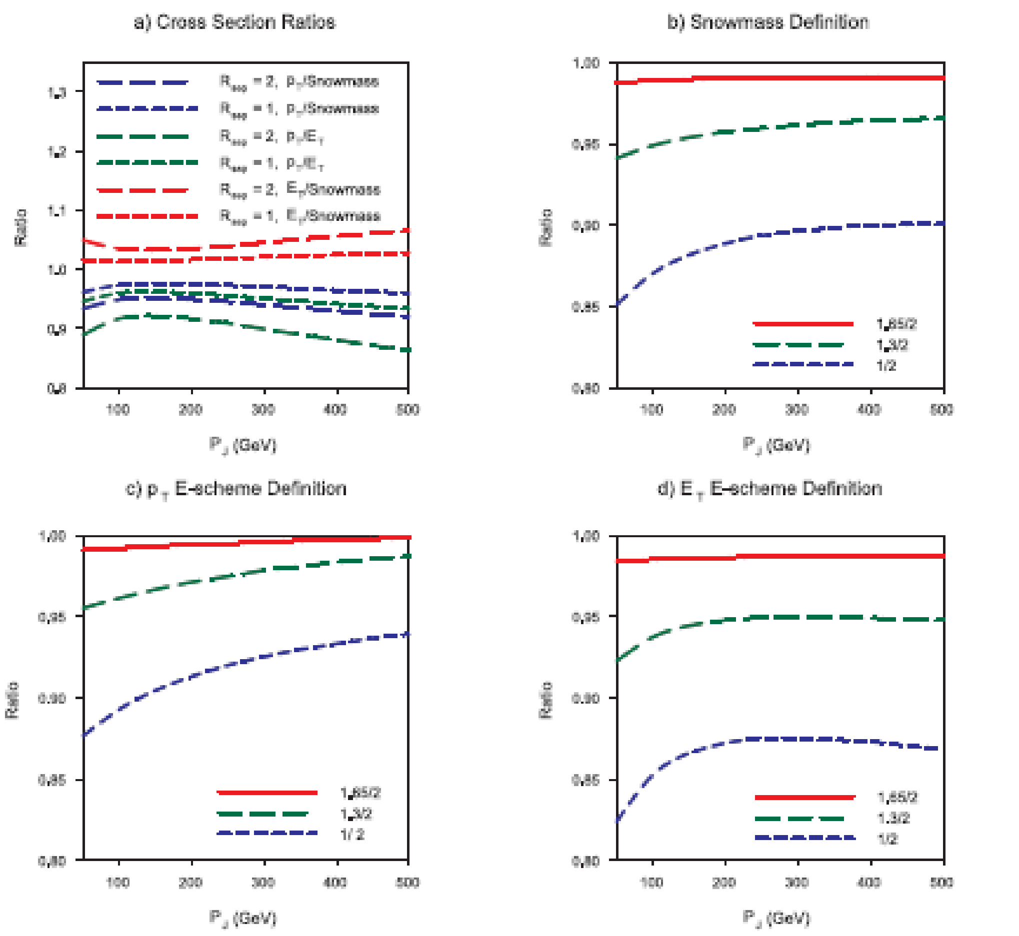


Fig. 2.0.6: Ratios of the NLO inclusive cone jet cross section versus the jet momentum for 3 definitions of the kinematics for various values of R_{sep}. The Snowmass definition uses p_T weighting and a jet momentum defined by the scalar sum $P_J = \sum_k p_{T,k}$. The E-scheme algorithms use 4-vector kinematics (as recommended for Run II) and either $P_J = p_T = |\vec{p}_T|$ (c) or $P_J = E_T = E \sin\theta$ (d). The parts of the figure illustrate a) ratios of different choices of P_J versus P_J for $R_{sep} = 2$ and $R_{sep} = 1$; b) ratio to the default value $R_{sep} = 2$ for $R_{sep} = 1.65$, $R_{sep} = 1.3$ and $R_{sep} = 1$ using the Snowmass definitions for the kinematics and for P_J; c) the same as b) except using 4-vector kinematics and $P_J = p_T$; d) the same as c) but with $P_J = E_T$.