Modern Jet Algorithms and their performance

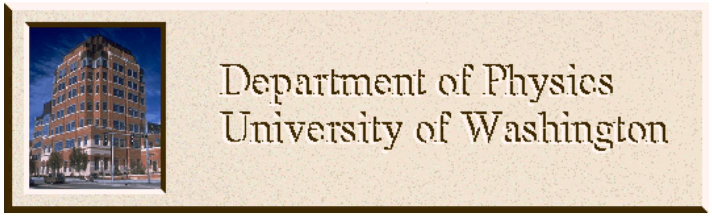
Steve Ellis

Big Picture:

For the next decade the focus of particle physics phenomenology will be on the LHC. The LHC will be both very exciting and very challenging -

- addressing a wealth of essential scientific questions
- with new (wonderfully precise) detectors
- operating at high energy *and* high luminosity (eventually)
- most of the data will be about hadrons (jets).

Theory and Experiment must work together to make the most of the data.



Department of Physics
University of Washington

US ATLAS Hadronic Final State Forum
8/23/10

Why am I here?

- I’ve been in jet physics since the beginning – so can provide some historical perspective on jets – A Jets Primer!

- I’m a theorist – so focus on theory/data interplay!
  What is reliable and what is not –

- We have learned a lot in the run-up to the LHC, and have many new tools (FASTJet, SPARTYJet, …)

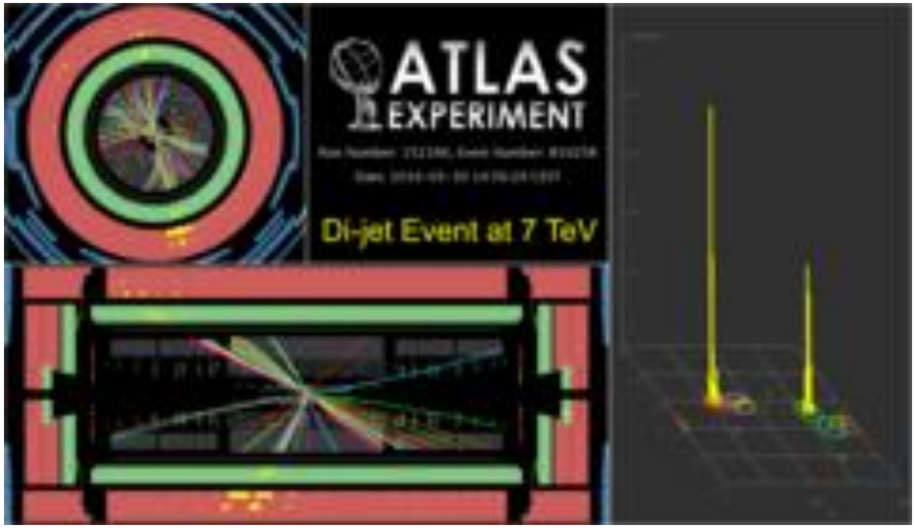
For a recent summary see “Jetography” by Gavin Salam, 0906.1833 (but I recall some history differently!)

Outline

- Why jets?
- Old and New lessons for Cone and Recombination (kT) jets
- When does theory work?
- Intro to understanding jet masses & substructure – see also Michael Spannowsky & Thomas Gadfort

What is a jet at the LHC?

- The answer is “qualitatively” simple to see -
- JET = collimated spray of hadronic “stuff” in the detector



So what is the big deal?

Why Jets & Algorithms?

- Focus on large energy exchange processes
(to make interesting stuff)
⇒ resolve the partons within beam protons!

- Partons (q's and g's) are colored and radiate when isolated in phase space ⇒ (~collinear) showers of colored partons
⇒ "jets" at parton level

- Long distance dof are color-less hadrons
⇒ jets of hadrons in the detector,
but cannot (strictly) arise from single parton

⇒ ID "jets" with algorithm (a set of rules) applied to both hadrons and partons

Jet Physics: The Basis of QCD Collider Phenomenology

Long distance physics = complicated (all orders showering of colored objects, nonperturbative hadronization = organization into color singlets)

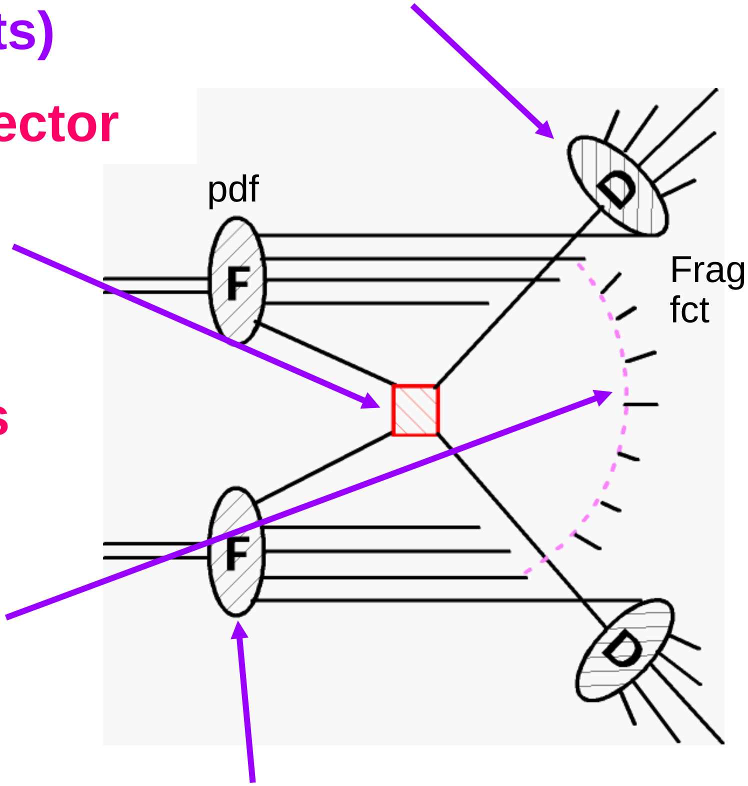
Exp’ters Measure this in the detector

Short distance physics = simple (perturbative)

Theorists Want to talk about this

Correlated by Underlying Event (UE) color correlations + PU

Stuck with this, small?



More long distance physics, but measured in pdfs

Defining Jets – No Unique/Correct Answer

- Map the observed (hadronic) final states onto the (short-distance) partons by summing up all the approximately collinear stuff (shower), ideally on an event-by-event basis.

- Need rules for summing ⇒ jet algorithm
  Start with list of partons/particles/towers
  End with list of jets (and stuff not in jets)

E.g.,

- Cone Algorithms, based on geometry – "non-local" sum over core of shower

  Simple, "well" suited to hadron colliders with Underlying Events (UE)

- Recombination (or kT) Algorithm, based on "local" pair-wise merging of local objects to "undo" shower

  Tends to "vacuum up" soft particles, "well" suited to e+e- colliders

Generally - IDEAL ALGORITHM

- *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 (all orders in PertThy)*
- *Detector Independence: there should be no dependence on cell type, numbers, or size*
- *Level Independence: The algorithms should behave equally at the parton, particle, and detector levels.*
- *Uniformity: everyone uses the same algorithms (theory and experiment, different experiments)*

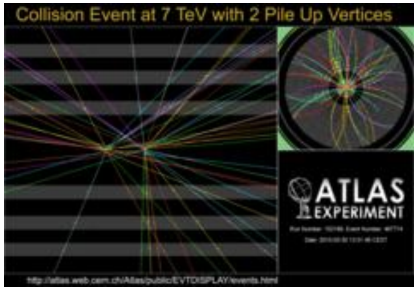
Historically never entirely true!

Looking for the hidden truth – a jet performs!!

But it is never that simple!!!

Jet issues can arise from -

- Systematics of specific algorithm
- Higher - order perturbative contributions (not just single parton)
- Showering – sum of all orders (leading-log, soft-collinear) emissions - smears energy distribution – “splash-out”
- Hadronization – nonperturbative re-organization into color singlet hadrons (confinement) – “splash-out”
- “Uncorrelated” contributions of rest of collision (UE) – “splash-in”
- Uncorrelated contributions of overlapping collisions (PU) “splash-in”



"Lessons" about the Jet Systematics – Test Case

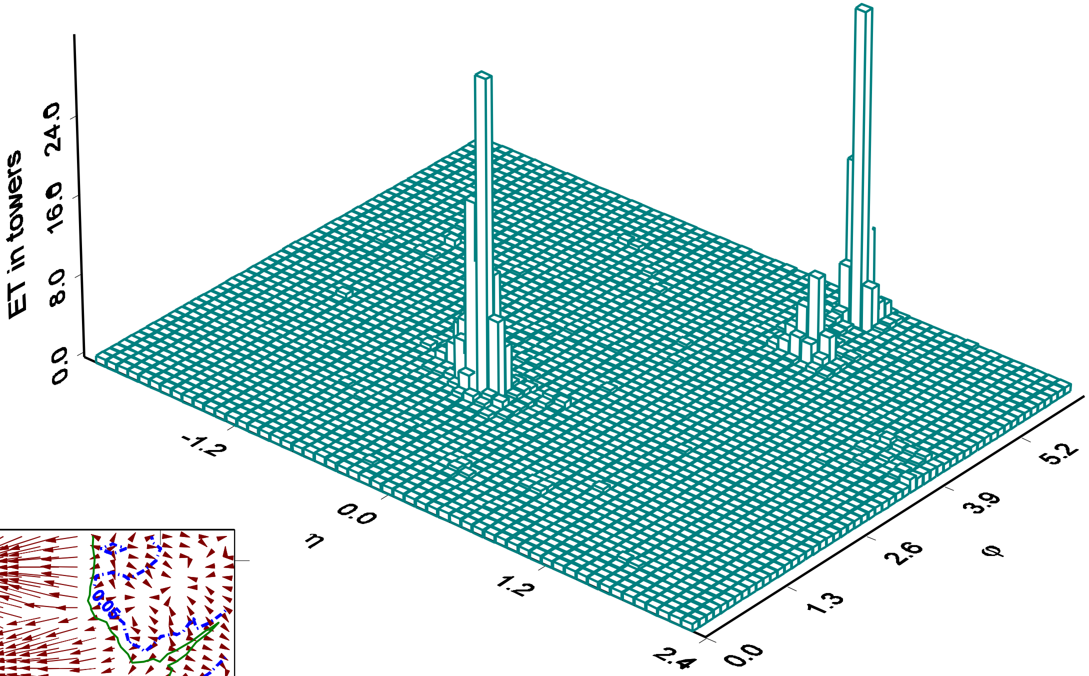
Cone Algorithm – focus on the core of jet (1990 Snowmass)

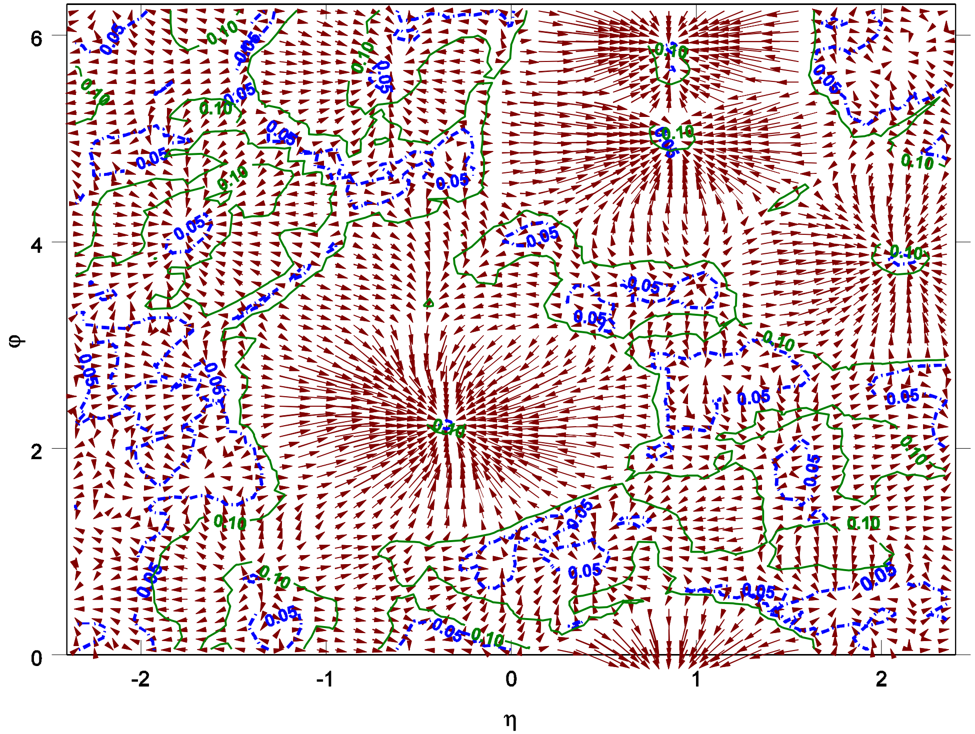
- Jet = "stable cone" ⇒ 4-vector of cone contents || cone direction
- Well studied – several issues

- **Cone Algorithm – particles, calorimeter towers, partons in cone of size R, defined in angular space, *e.g.*, (*y*,*φ*),**
- **CONE center -** $\left(y^C, \varphi^C\right)$
- **CONE i** $\in$ ***C iff*** $\Delta R^i \equiv \sqrt{\left(y^i - y^C\right)^2 + \left(\varphi^i - \varphi^C\right)^2} \leq R$
- **Cone Contents ⇒ 4-vector** $P_\mu^C = \sum_{i \subset C} p_\mu^i$
- **4-vector direction** $\bar{y}^C = 0.5 \ln\left[\frac{P_0^C + P_z^C}{P_0^C - P_z^C}\right] \; ; \; \bar{\varphi}^C = \arctan\left[\frac{P_y^C}{P_x^C}\right]$
- **Jet = stable cone** $\left(\bar{y}^C, \bar{\varphi}^C\right) = \left(y^C, \varphi^C\right)$

Find by iteration, *i.e.*, put next trial cone at $\left(\bar{y}^C, \bar{\varphi}^C\right)$

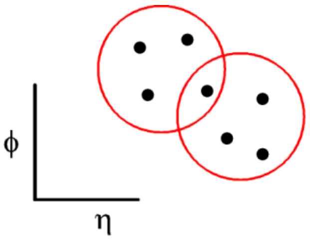
Iteration ⇒ “flow” to the “hot” spots





Iteration ⇒ meant to be insensitive to initial point/initial clustering

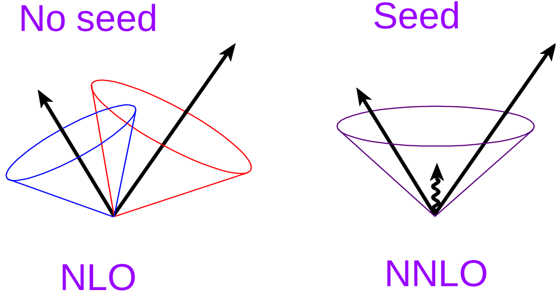
Cone Lessons: (Life is not so simple)



1) Stable Cones can and do Overlap: need rules for merging and splitting, split/merge algorithm has new parameter f_{merge} but not the same for D0 and CDF

2) <u>Seeds</u> – experiments only look for jets around active regions (save computer time)

⇒ problem for theory, IR sensitive (Unsafe?) at NNLO



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

This is a BIG deal philosophically – but not a big deal numerically (in data)

⇒ Use SEEDLESS version (SISCone) at the LHC

Seeds and Sensibility – An Important Lesson for Theory – Experiment interaction

- 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)

⇒ DON'T use seeds in perturbation theory, correct for them in data analysis, or better, USE SEEDLESS Algorithm (SISCone)!!

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

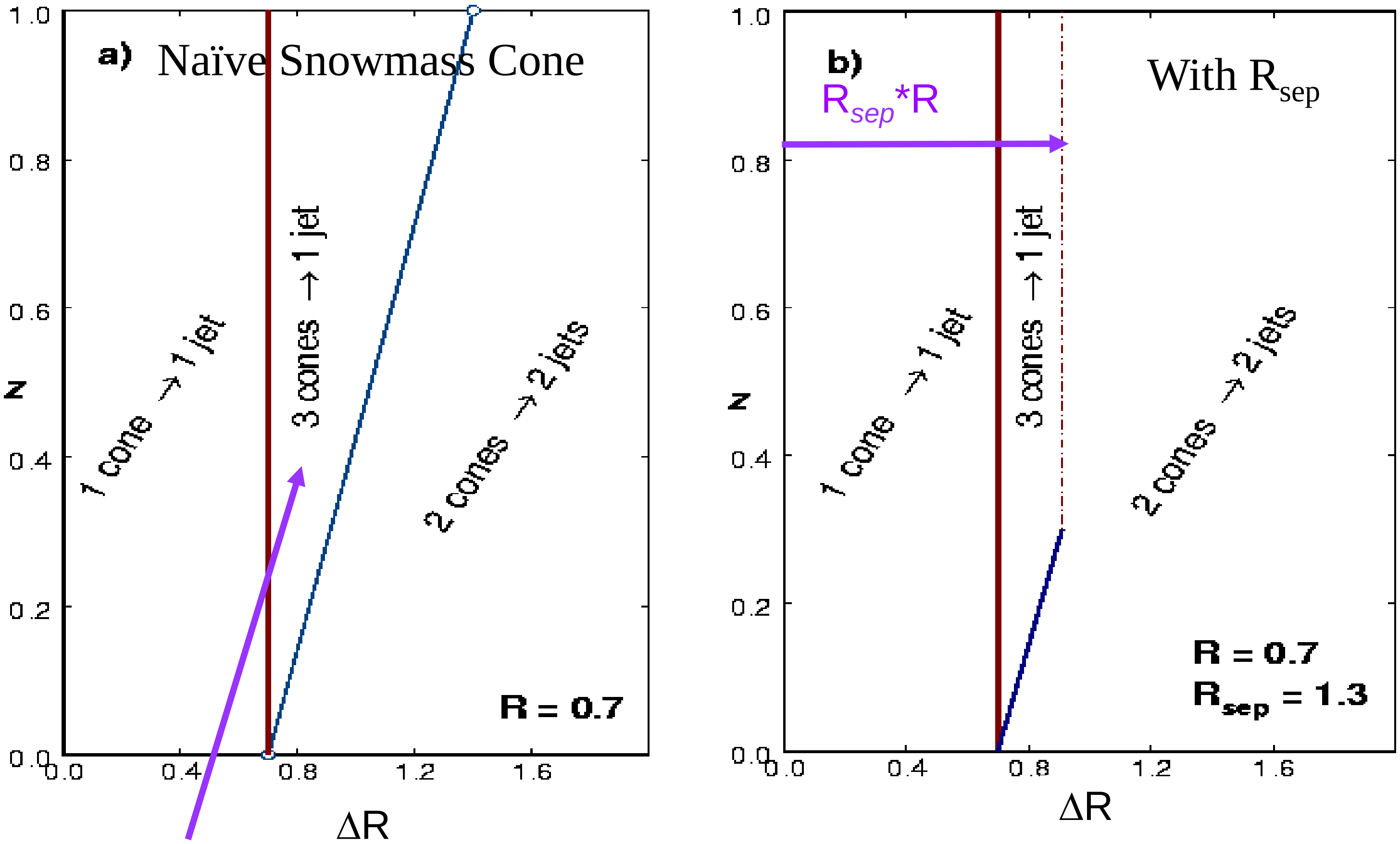
In the data seeds vs seedless is a few % correction (e.g., lower the Seed p_T threshold) and this is small compared to other corrections

Remember this lesson at the LHC!!

2 Parton (NLO) Phase Space

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

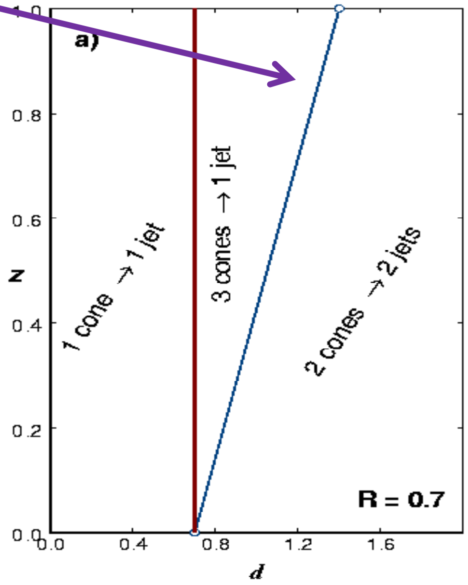
  Simulate the missed middle cones with R_{sep} (not very helpful at higher orders)



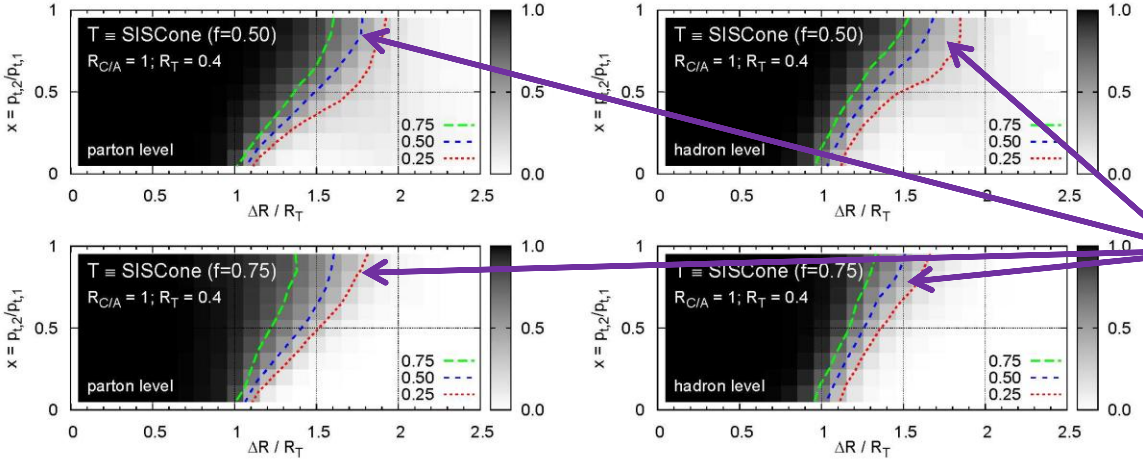
~10% of cross section here

Cone Lesson - (Even if Seedless)

3) Splash-out from smearing of energetic parton at edge of cone – can be quantitatively relevant



- Study by G. Salam* – find 2 (sub)jets with a different algorithm and see when SISCone merges them



(Shower only)

(Shower + Had-n)

Seldom merge in the corner, another reason to cut corner (R_{sep})

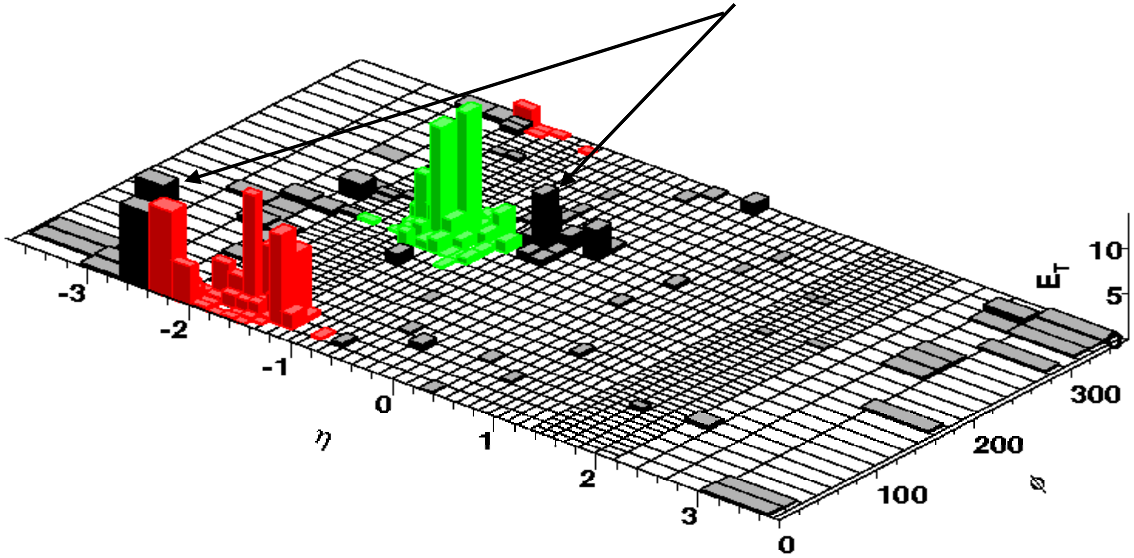
*Similar to earlier CDF study

(Related) Cone Lesson: (Even if Seedless)

4) Dark Towers - Energy in secondary showers may not be clustered in any jet

- Expected stable cone *not* stable due to smearing from showering/hadronization (compared to PertThy)

- Under-estimate E_T (~ 5% effect for jet cross section)



Cone Fixes -

1. All experiments use the same split/merge parameters and 0.8 > f_{merge} > 0.5 to avoid over-merging, over-splitting (jet size stable *vs* jet pT or PU) - Not true at the Tevatron…

2. NOTE: “progressive-removal” seeded cones - find cone jets one at a time starting with *largest* pT seed and REMOVE jet constituents from further analysis. This is NOT collinear safe!

3. Use Midpoint Cone (with seeds) in Run II – only fix to NNLO

4. Use seedless cone algorithm (e.g., SIScone)!!!!

5. No “good” solution yet to Dark towers except to look for 2nd pass jets after removing the 1st pass jets from the analysis (employed in SISCone).

Conclude ⇒ Cone algorithm well understood, but several issues!

Recombination – focus on undoing the shower pairwise

Merge partons, particles or towers pairwise based on "closeness" defined by minimum value of

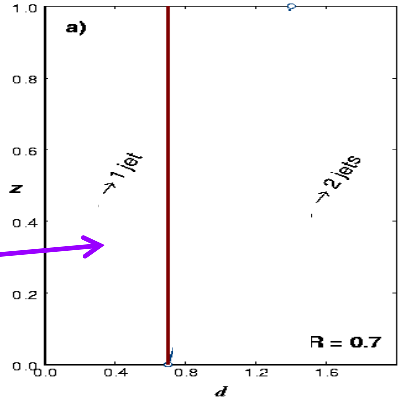
$$k_{T,(ij)}^2 \equiv \mathrm{Min}\left[\left(p_{T,i}^2\right)^{\alpha}, \left(p_{T,j}^2\right)^{\alpha}\right] \frac{\left(y_i - y_j\right)^2 + \left(\phi_i - \phi_j\right)^2}{D^2}, k_{T,i}^2 = \left(p_{T,i}^2\right)^{\alpha}$$

If $k_{T,(ij)}^2$ is the minimum, merge pair (add 4-vectors and redo list;

If $k_{T,i}^2$ is the minimum → *i* is a jet!

(no more merging for *i*), 1 parameter *D*

(NLO, equals cone for D = R, R_{sep} = 1)



α = *1,* ordinary k_T, recombine soft stuff first (undo kT ordered shower)

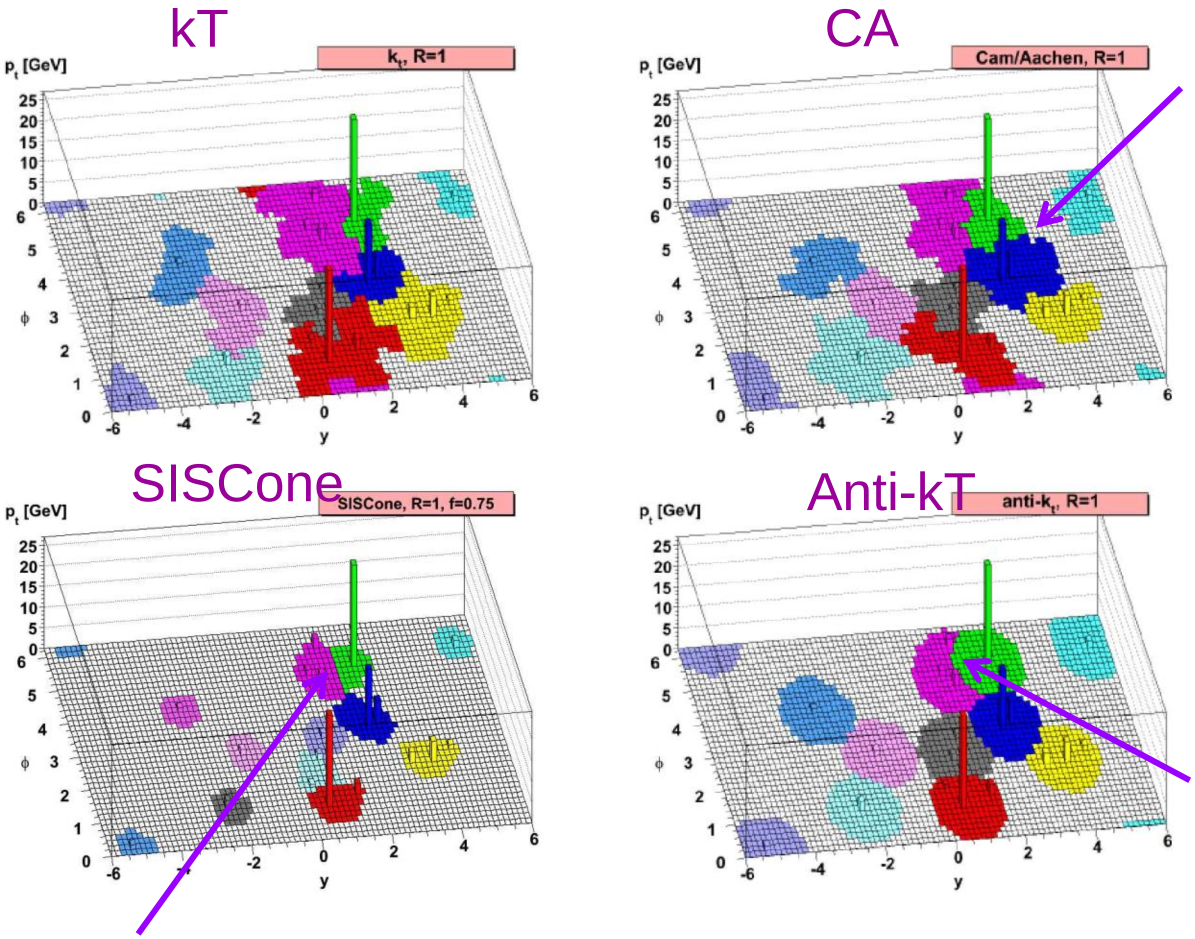
α = *0, Cambridge/Aachen* (CA), controlled by angles only (undo angle ordered shower)

α = *-1, Anti-*k_T, just recombine stuff around hard guys – cone-like with seeds ***THE NEW GUY!!*** (not matched to shower structure)

Recombination Lessons:

- 👍 Jet identification is unique – no merge/split stage
- 👍 “Everything (interesting) in a jet”, no Dark Towers
- 👎 Resulting jets are more amorphous for $\alpha \geq 0$, energy calibration more difficult (subtraction for UE + PU?)
- 👍 But for $\alpha < 0$, Anti-kT (Carriari, Salam & Soyez), jet area seems stable and geometrically regular * - the “real” cone algorithm (but large pT jets take a bite out of small pT one)

Jet Areas – from Salam & Carriari, Salam & Soyez



kT

CA

SISCone

Anti-kT

Amorphous edges

Anti-kT very regular leading jets,

But bites others

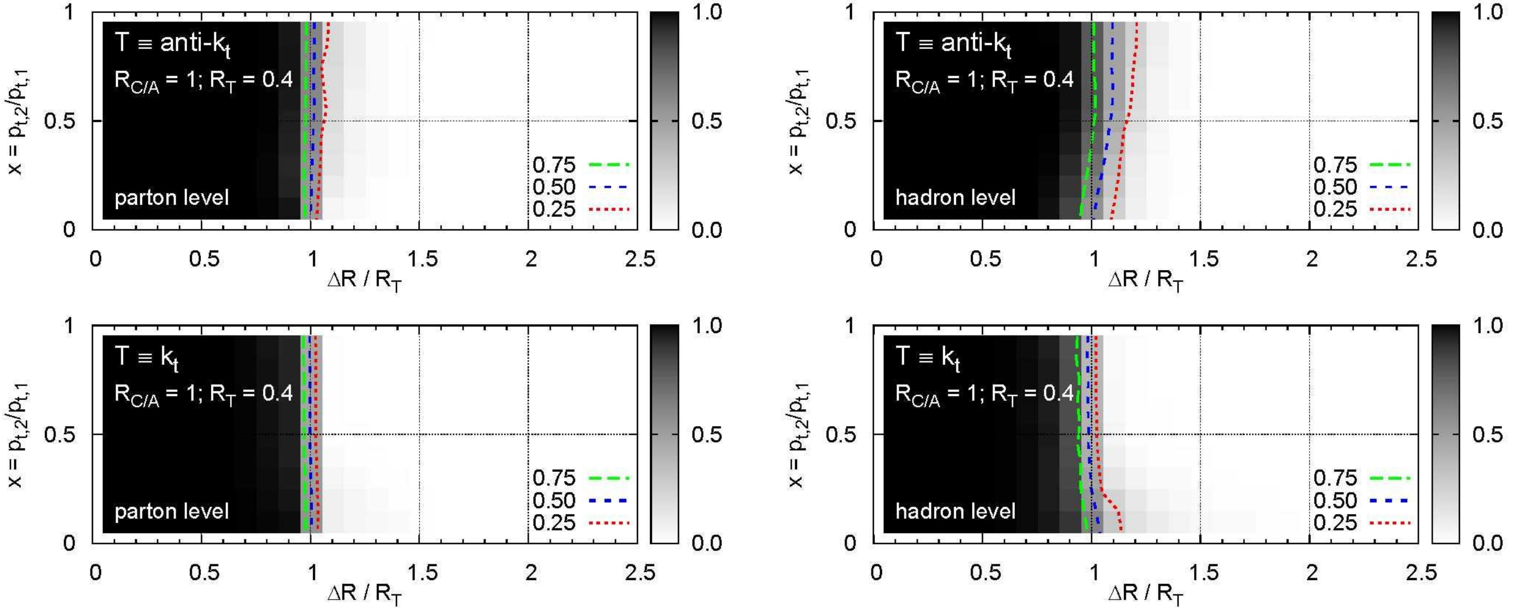
S/M effect

Recombination Lessons:

- 👍 Jet identification is unique – no merge/split stage
- 👍 "Everything (interesting) in a jet", no Dark Towers
- 👎 Resulting jets are more amorphous for $\alpha \geq 0$, energy calibration more difficult (subtraction for UE + PU?)
- 👍 But for $\alpha < 0$, Anti-kT (Carriari, Salam & Soyez), jet area seems stable and geometrically regular * - the "real" cone algorithm (but large pT jets take a bite out of small pT one)
- 👍 Energetic partons seldom near edge of jets – merging matches PT theory, showering less of a issue

Recombination & showering (Salam)

Find energetic subjets (~last merging) with CA, see what anti-kT and kT find – do they still merge insider R – YES



(Shower only)

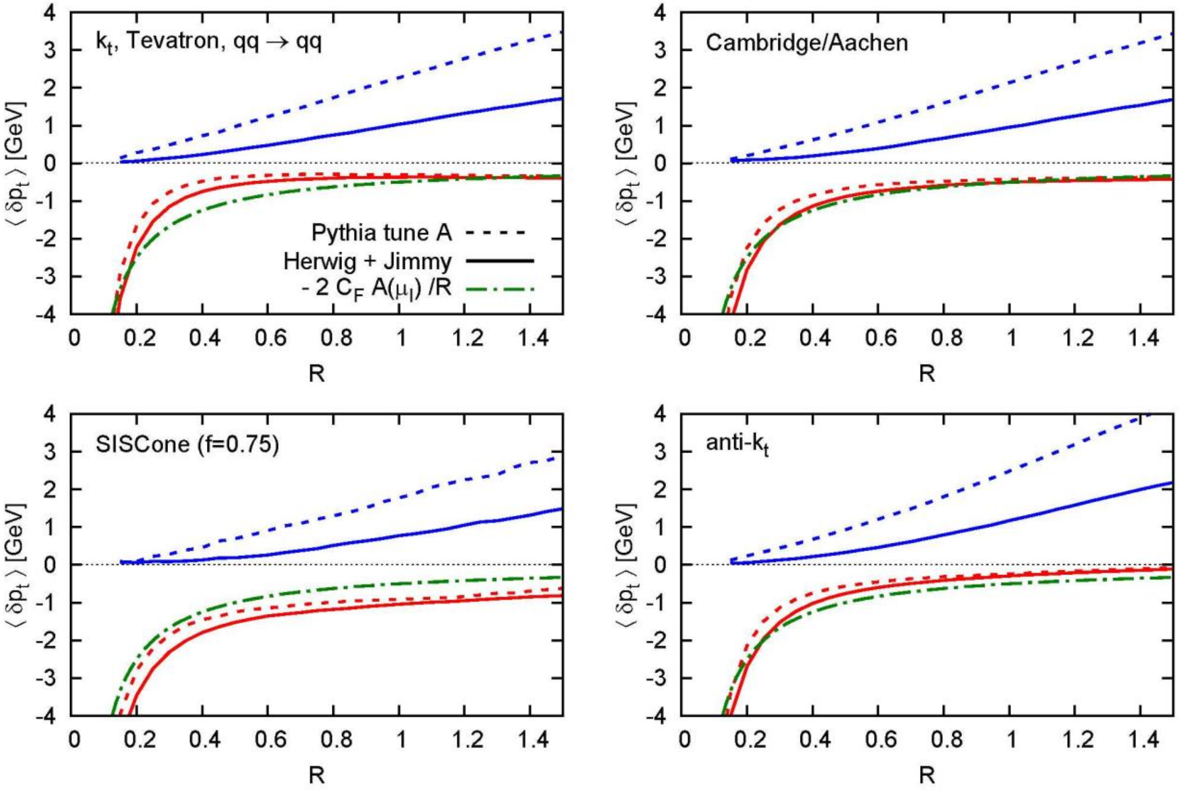
(Shower + had’n,
more smearing)

Recombination Lessons:

- 👍 Jet identification is unique – no merge/split stage
- 👍 “Everything” in a jet, no Dark Towers
- 👎 Resulting jets are more amorphous for $\alpha \geq 0$, energy calibration more difficult (subtraction for UE + PU?)
- 👍 But for $\alpha < 0$, Anti-kT (Carriari, Salam & Soyez), jet area seems stable and geometrically regular * - the “real” cone algorithm (but large pT jets take a bite out of small pT one)
- 👍 Energetic partons seldom near edge of jets – merging matches PT theory, showering less of a issue
- 👎 Analysis can be very computer intensive (time grows like N^3, recalculate list after each merge)
- 👍 New version FASTJet (Salam & Soyez) goes like N^2 or N ln N ($\alpha \geq 0$), plus scheme for finding areas (and UE correction)

Hadronization (splash-out) & UE (Splash-out) issues:

- Measure and Correct jet properties – MC & analytic Tevatron - pJ/$\Delta R^2 \sim$ 0.5 GeV for UE,



UE (MC's don't agree)

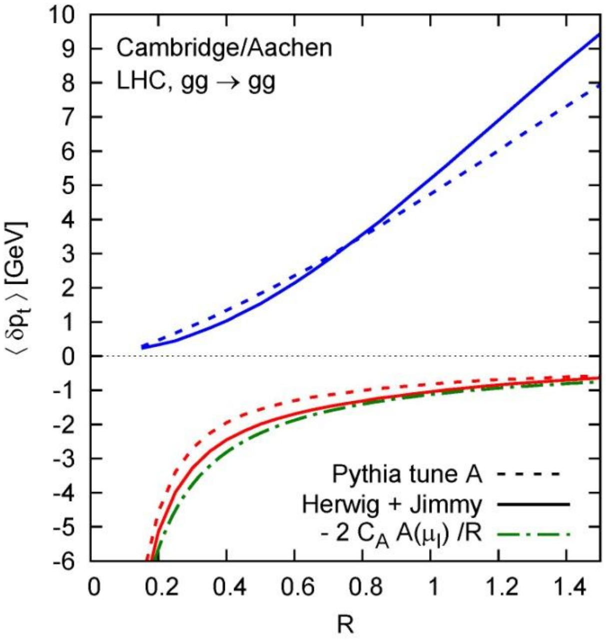
hadronization

Similar for different algorithms and approx. cancel for R ~ 0.7

Studies by Dasgupta, Magnea and Salam

At the LHC, expect - $pJ/\Delta R^2 \sim 1.5$ GeV

- Here gluons!,
  better UE agreement
  Still approx. cancel for
  R ~ 0.6



- If “groom” jets (pruning, trimming, filtering), may lose cancellation!

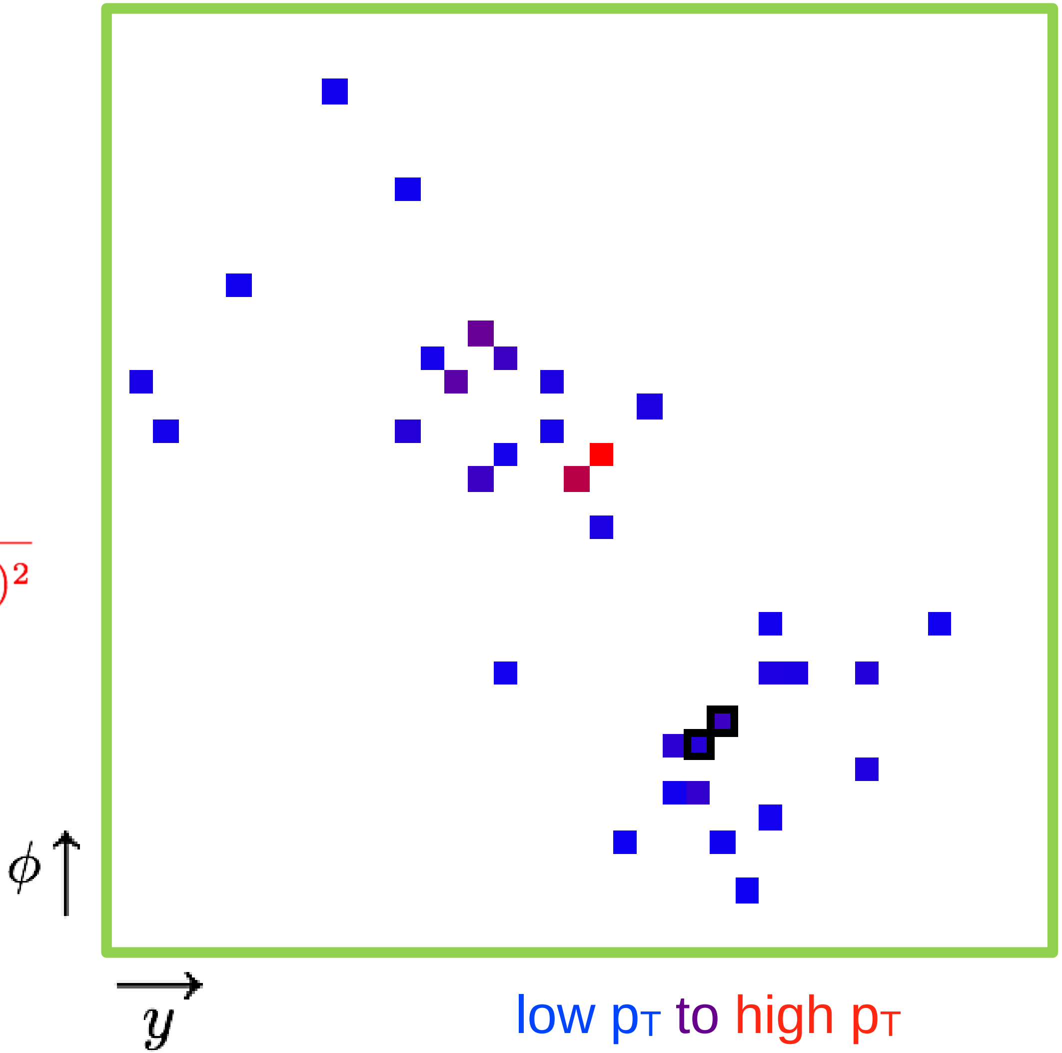
Using Recombination Algorithms at the LHC – Here CA algorithm in action – “natural” substructure at each merging!

Think of starting with calorimeter cells, recombine “closest” pair at each step leading to larger p_T

For CA close in quantity

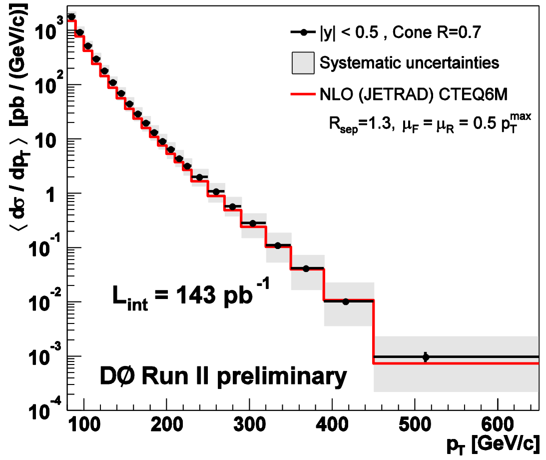
$$\Delta R_{ij} = \sqrt{(y_i - y_j)^2 + (\phi_i - \phi_j)^2}$$

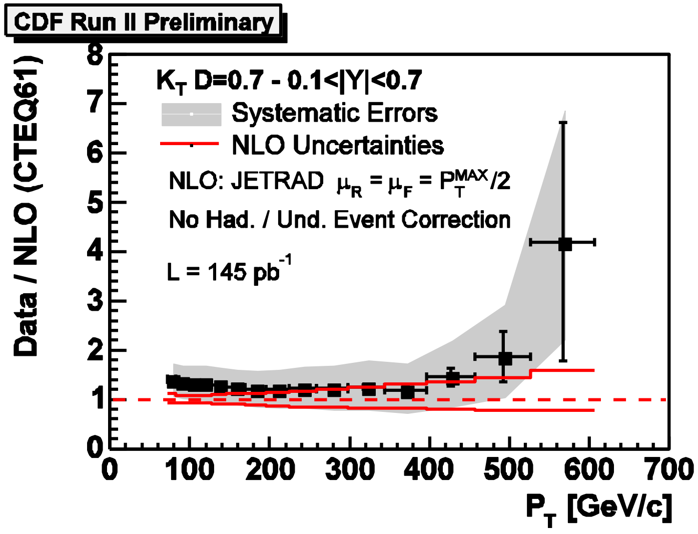
(0.05 x 0.05) Cells
with E > 1 GeV

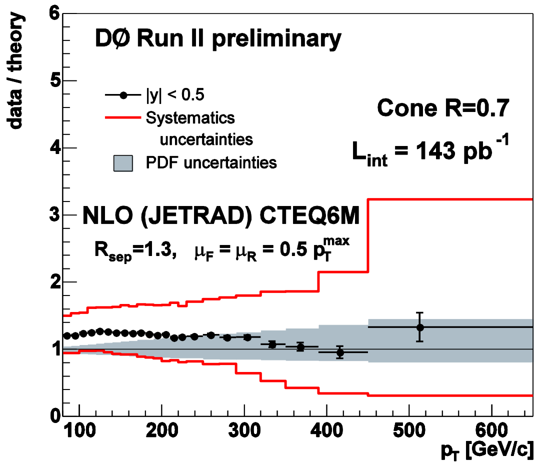


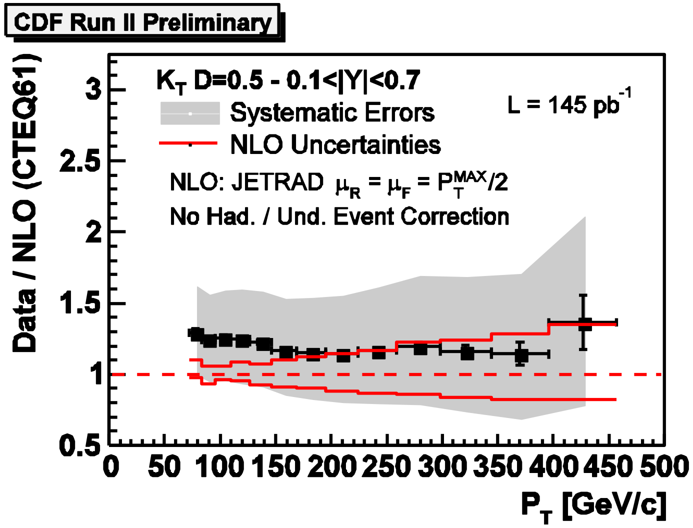
low p_T to high p_T

Tevatron



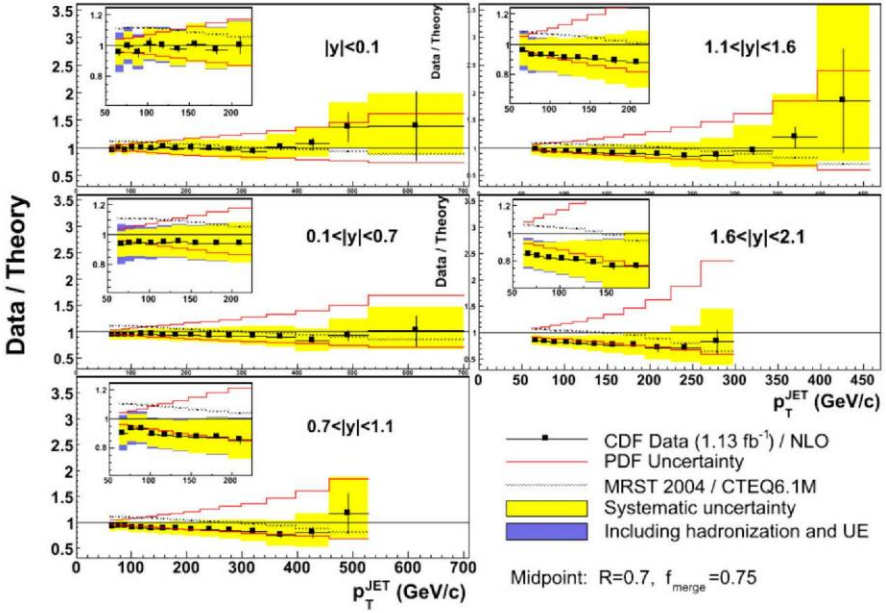






Both algorithms work and yield similar results, ~10%

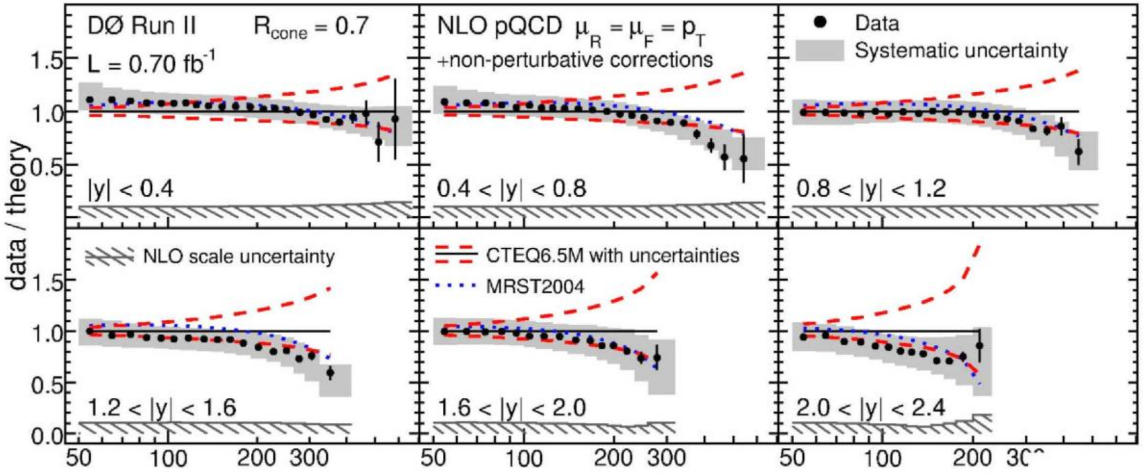
Latest Cone Data – from Tevatron



PDF uncertainty tends to dominate, ~ 10%, especially at large p_T

Note: CDF uses $\mu = p_T/2$

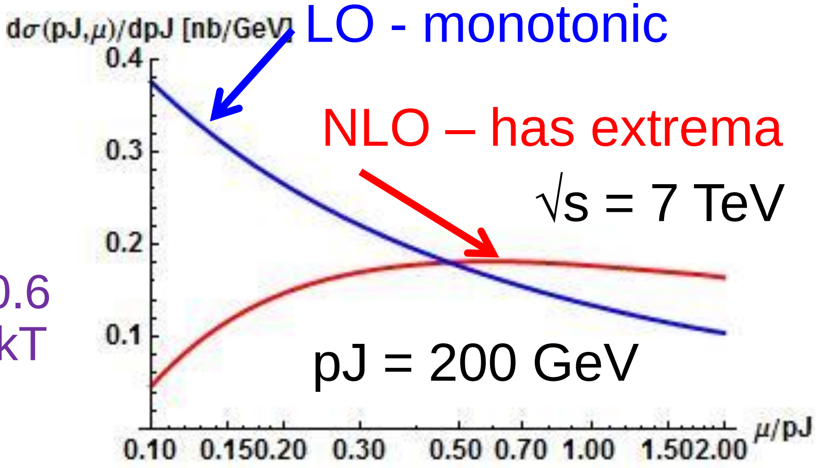
Note: D0 uses $\mu = p_T$



Aside on scale choices in Jet Perturbation Theory - How large a pT is large enough?

- The History is choose $\mu_{UV} = \mu_{CO} = pJ/2$ based on
  NLO ~ μ “independent” there
  NLO ~ LO there
  The Pert Thy tells us the scale!

D = 0.6
Anti-kT



LO - monotonic

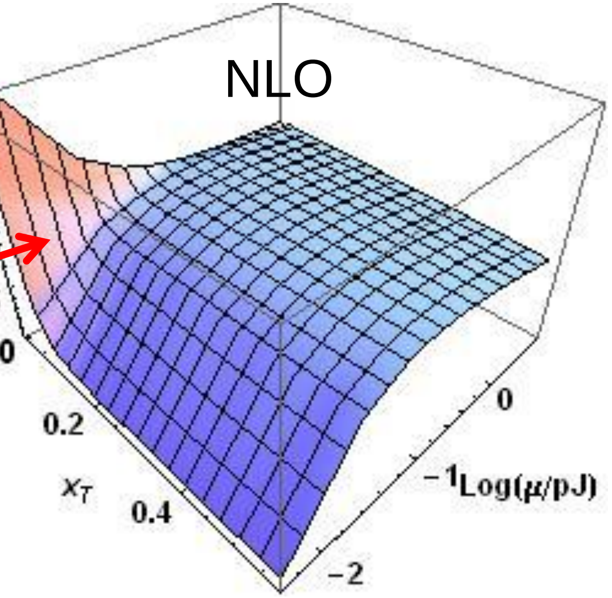
NLO – has extrema

$\sqrt{s}$ = 7 TeV

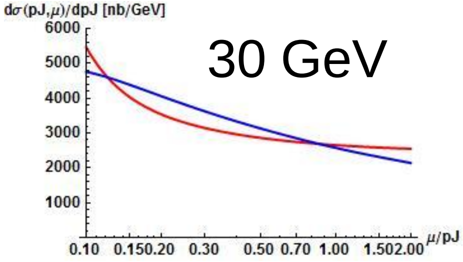
pJ = 200 GeV

But not true for $x_T = 2p_T/\sqrt{s} < 0.03$, large ln(xT)

NLO becomes monotonic again at small pJ
pJ < 100 GeV at 7 TeV
NLO not reliable

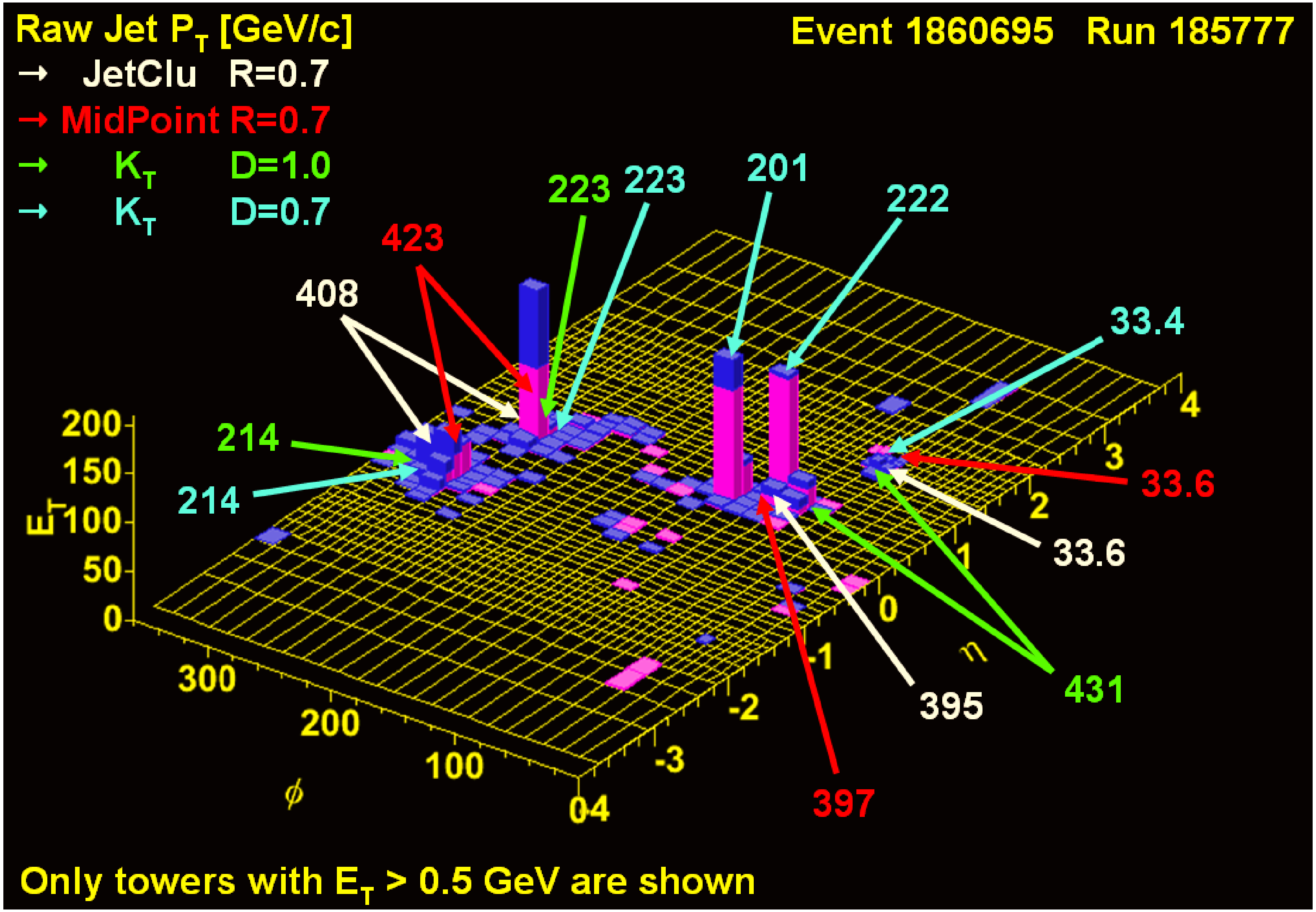


NLO



30 GeV

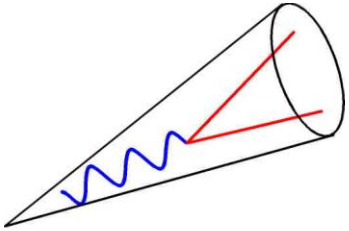
Final Lesson: Different algorithms ⇒ slightly different jets (same CDF event)



EM, **Hadronic**

Goals at LHC Different ⇒ Different Role for Jets!

- Find Physics Beyond the Standard Model (BMS), need < 10% precision!
- BSM Event structure likely different from QCD, more jets? Different structure within jets? Must be able to reconstruct masses from multi-jets & also from *single jets*
- Want to select events/jets by non-QCD-ness



- Highly boosted SM and non-SM particles – W, Z, top, Higgs, SUSY ⇒ *single* jet instead of 2 or 3 jets, focus on masses and substructure of jets
- Much recent progress, but lots of work still to be done – see next talks!!

Recall Jets History at Hadron Colliders

- JETS I – Cone style jets applied to data at the SpbarpS, and Run I at the Tevatron to map final state hadrons onto LO (or NLO) hard scattering, initially 1 jet ⇔1 parton (test QCD)

Little attention paid to masses of jets or the internal structure, except for energy distribution within a jet

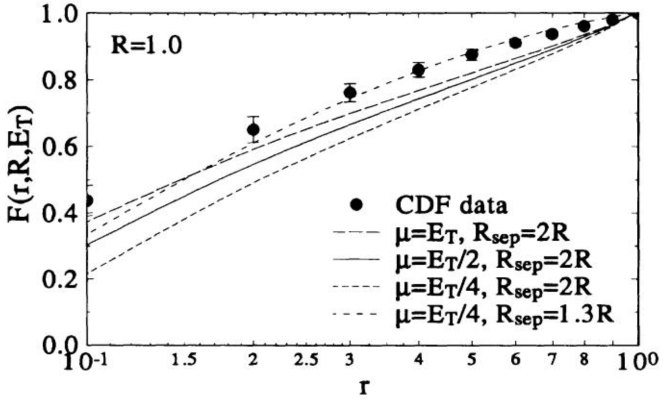


FIG. 2. $F(r,R,E_T)$ vs r for $R=1.0$, $\sqrt{s}=1800$ GeV, $E_T=100$ GeV, and $0.1<|\eta|<0.7$ with $\mu=E_T/4$, $E_T/2$, E_T compared to data from CDF [7]; the dot-dashed curve is explained in the text.

- JETS II – Run II & LHC, starting to look at structure of jets: masses and internal structure – a jet renaissance

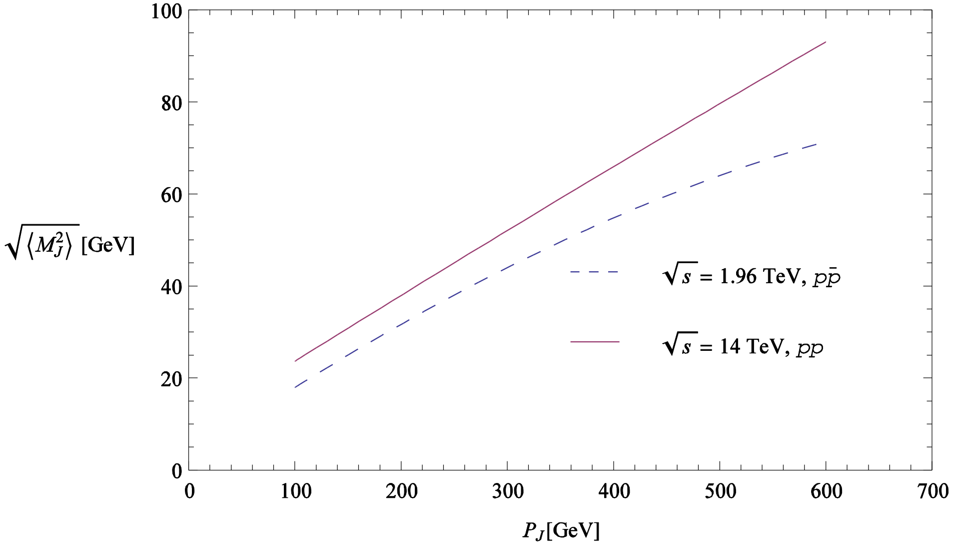
Jet Masses in pQCD:

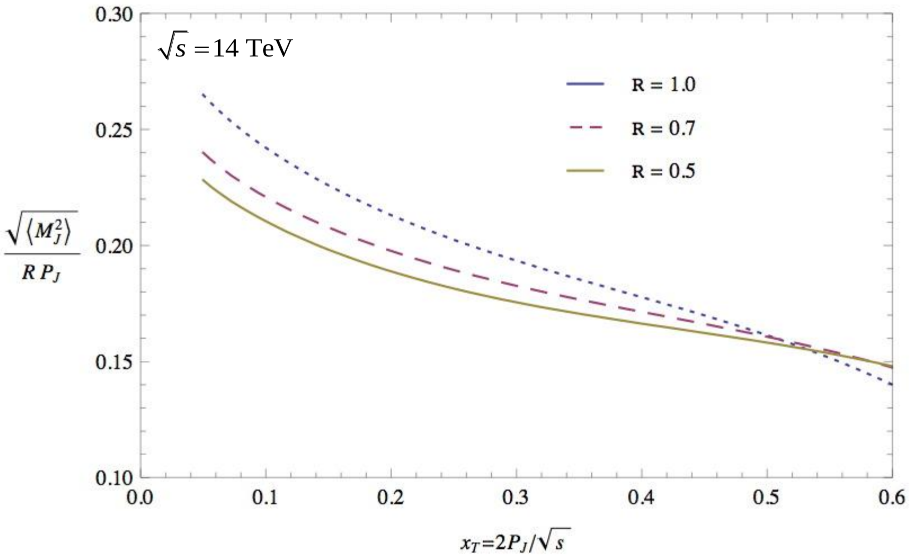
Dimensions

- In NLO PertThy $\sqrt{p_{J,\mu}p_J^{\mu}} \Rightarrow \sqrt{\left\langle M^2 \right\rangle_{NLO}} = f\left(\frac{p_J}{\sqrt{s}}\right)\sqrt{\alpha_s\left(p_J\right)}\,p_J R$

Phase space from dpfs, *f* ~ 1

Jet Size, *R*, *D* ~ $\Delta\theta$, determined by jet algorithm

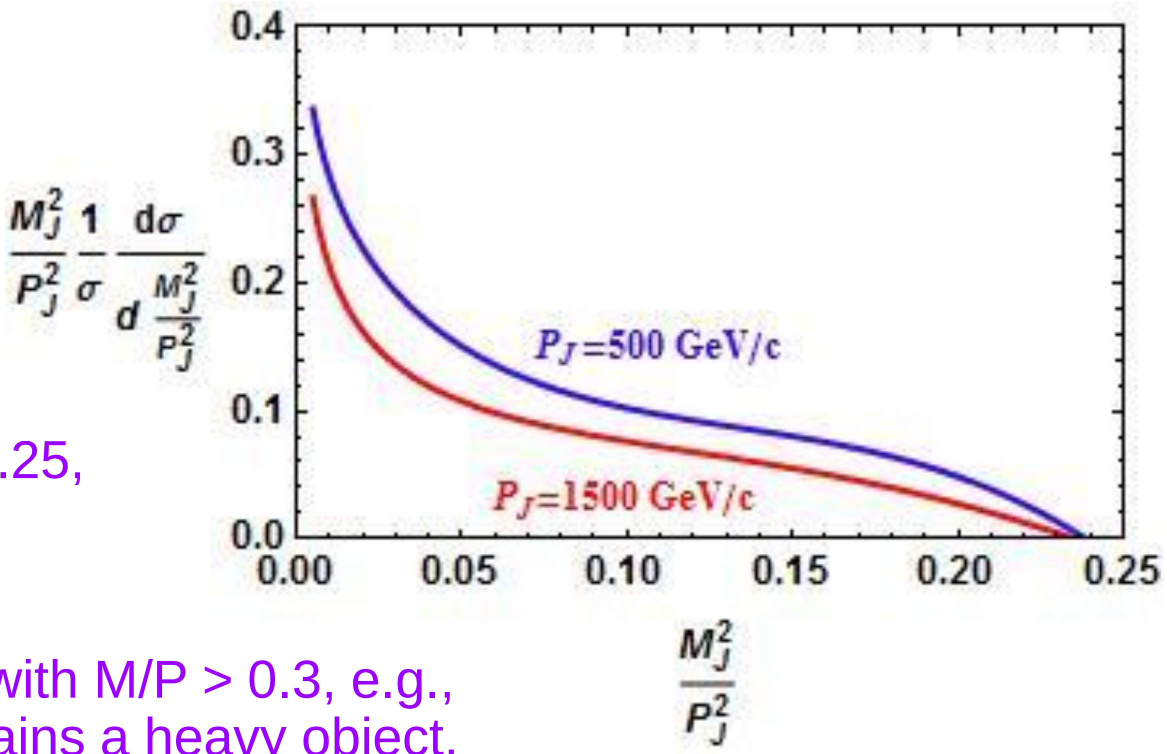




Useful QCD "Rule-of-Thumb" $\Rightarrow \sqrt{\left\langle M^2 \right\rangle_{NLO}} \sim 0.2\, p_J R$

Mass for fixed P_J at NLO

For Cone, R = 0.7
or kT, D = 0.7

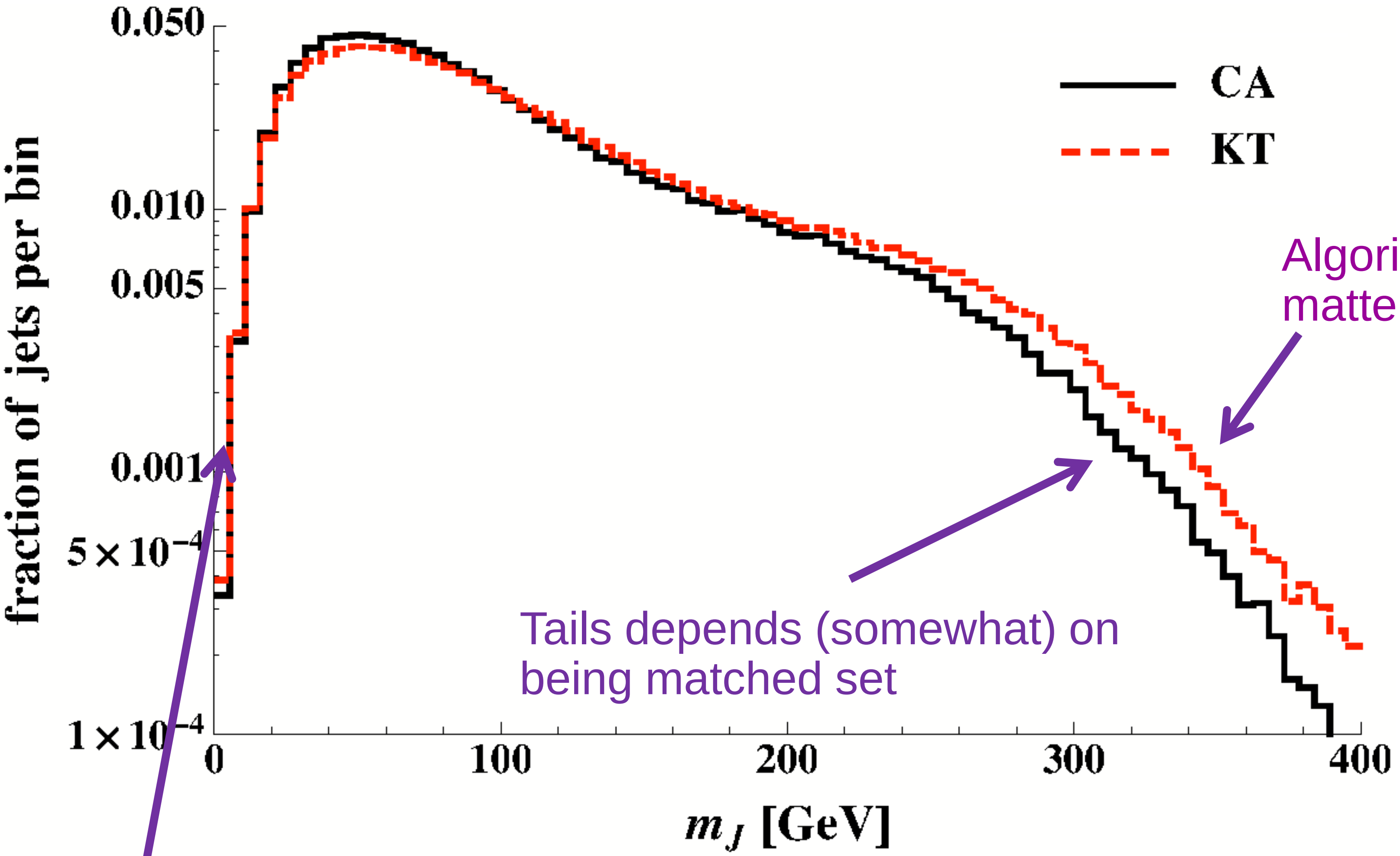


Peaked at low mass,
cuts off for $(M/P)^2 > 0.25$,
$M/P > 0.5$

$\Rightarrow$ Selecting on jets with $M/P > 0.3$, e.g.,
because the jet contains a heavy object,
already suppresses the QCD background;

Want heavy particle boosted enough to be in a jet (use large-ish R,D ~1),
but not so much to be QCD like ($\sim 2 < \gamma < 5$)

Jet Mass in PYTHIA (showered & matched set)
D = 1, 500 GeV/c < pT < 700 GeV/c



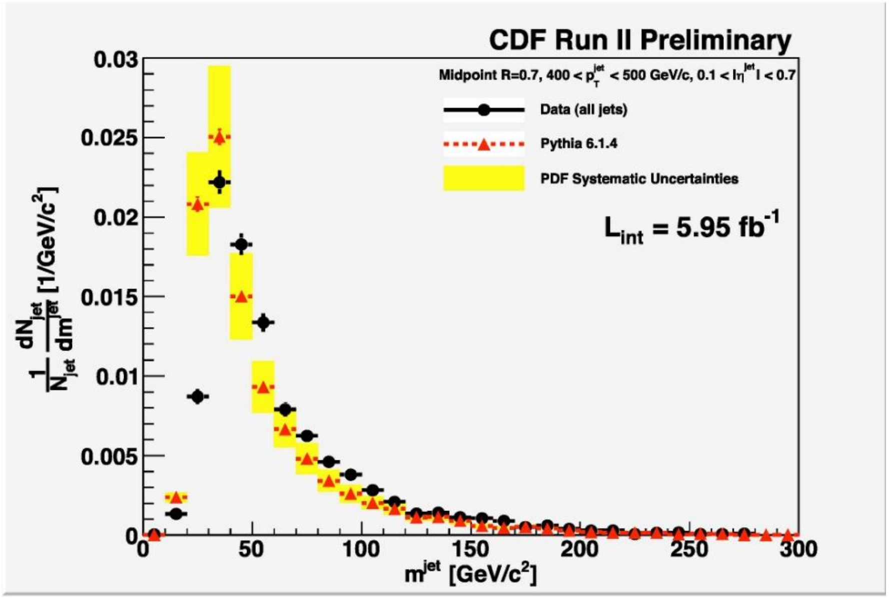
Algorithm matters

Tails depends (somewhat) on being matched set

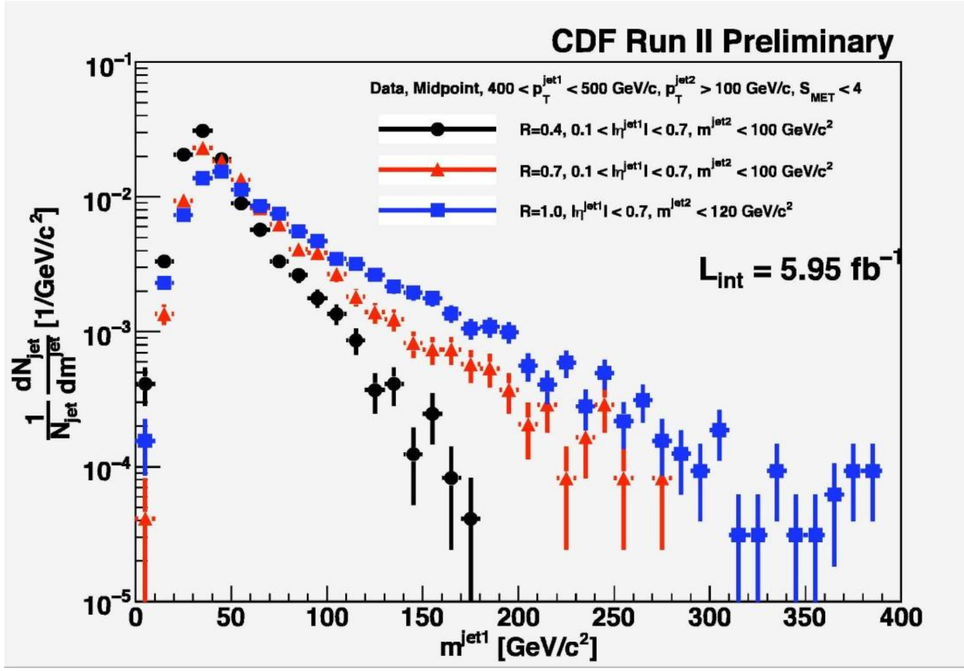
Turns over

Jet Mass – CDF Data

(CDF/PUB/JET/PUBLIC/10199 7/19/10)



At least qualitatively the expected shape – masses slightly larger than MC – need the true hard emissions (as in matched sets)

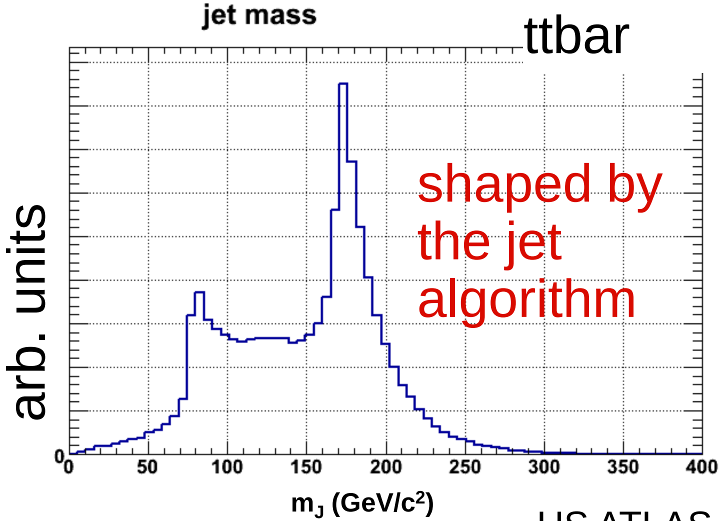


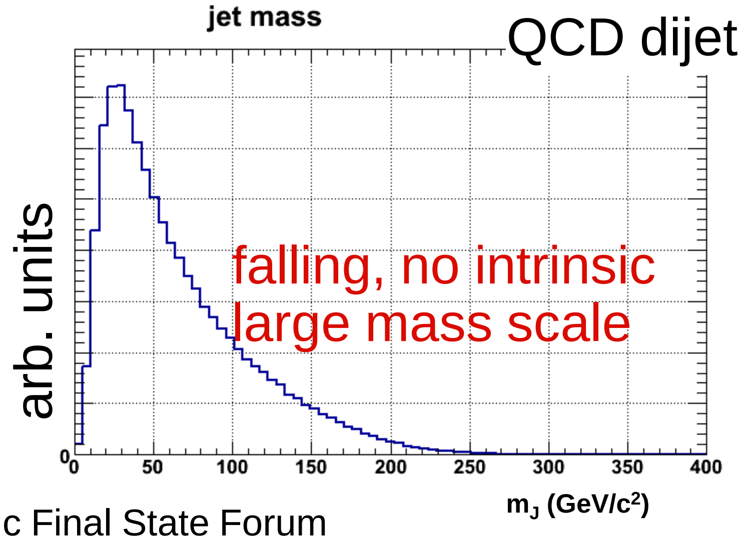
Large mass tail grows, as expected, with jet size parameter in the algorithm - You find what you look for!

Heavy Particles Searches with Single Jets - Issues

- 👎 QCD multijet production rate >> production rate for heavy particles
- 👍 In the jet mass spectrum, production of non-QCD jets may appear as local excesses (bumps!) but must be enhanced using analyses
- 👍 Use jet substructure as defined by **recombination algorithms** ($\alpha \geq 0$, not anti-kT) to refine jets
- 👎 Algorithm will systematically shape distributions
- Example - top quark as surrogate new particle.

$\boldsymbol{\sigma}_{\text{ttbar}} \approx 10^{-3}\boldsymbol{\sigma}_{\text{jj}}$





Jet Substructure at the LHC – (at least 2 approaches)

- Jet Tagging – select for specific substructure characteristic of search target, e.g., top quark
- Jet Grooming – “cleanup” jet to make any inherent mass scale more apparent (bump in mass distribution) – also reduces impact of UE, PU and algorithm details

Summary/Conclusions:

- It will take time to understand the SM at the LHC, but we understand jets much better now than we did at the beginning of Run I
- It is essential to test and validate a variety of jet algorithms – the familiar ones like cones, whose issues we need to re-confirm, and the less familiar ones like Anti-kT, whose issues we need to uncover – different algorithms find (slightly) different jets and will likely have different uses
- It is essential that the different Collaborations document the algorithms they use – and try to use the same ones some of the time
- It is essential to study and understand the role of the Underlying Event and Pile-Up (splash-in) and Showering and Hadronization (splash-out) in jets at the LHC

Summary/Conclusions:

- In comparing to perturbative QCD results, it is important to let the calculation define the appropriate scale. When logs are large, it will be important to sum them. Soft Collinear Effective Theory (SCET) techniques may be useful. I think they will be!

- It is essential to study and understand the properties of jets – masses and substructure – validate by IDing top jets, W/Z jets in LHC data

  Recombination algorithms ($\alpha \geq 0$) provide natural substructure

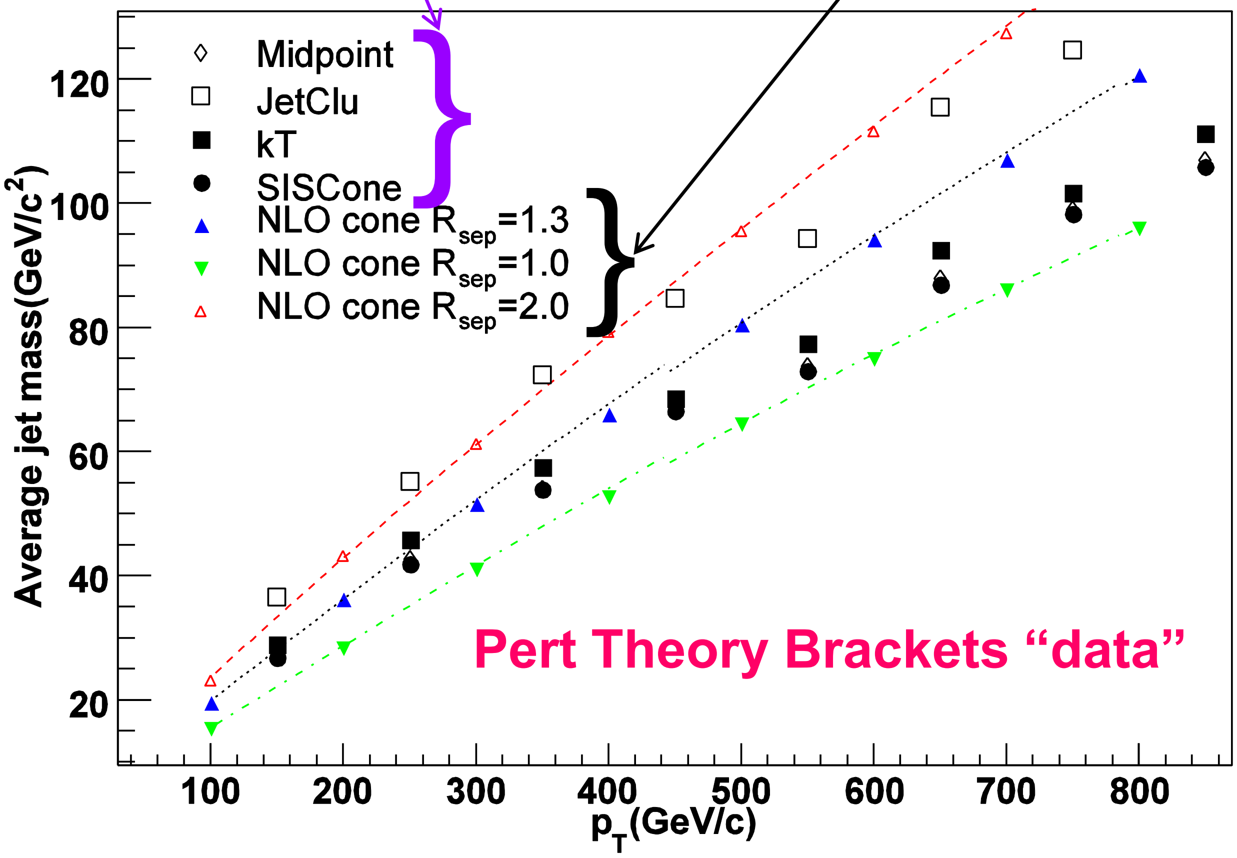
  $\Rightarrow$ single jets and their substructure will likely play a role in the search for BSM physics, along with heavy flavor tags, correlations with other jets (pair production), MET, *etc*.

Extra Detail Slides

Compare to (simulated) LHC data: (R_{sep} scales R)

Various algorithms applied to simulated LHC data
(diamond, square, circle)

NLO Cone Theory, various R_{sep} values (lines, triangles)



Pert Theory Brackets "data"

R_{sep} = 2, Snowmass

R_{sep} =1.3 EKS

R_{sep} = 2, kT