MCnet-LPCC Summer School
July 2012, CERN

Jets

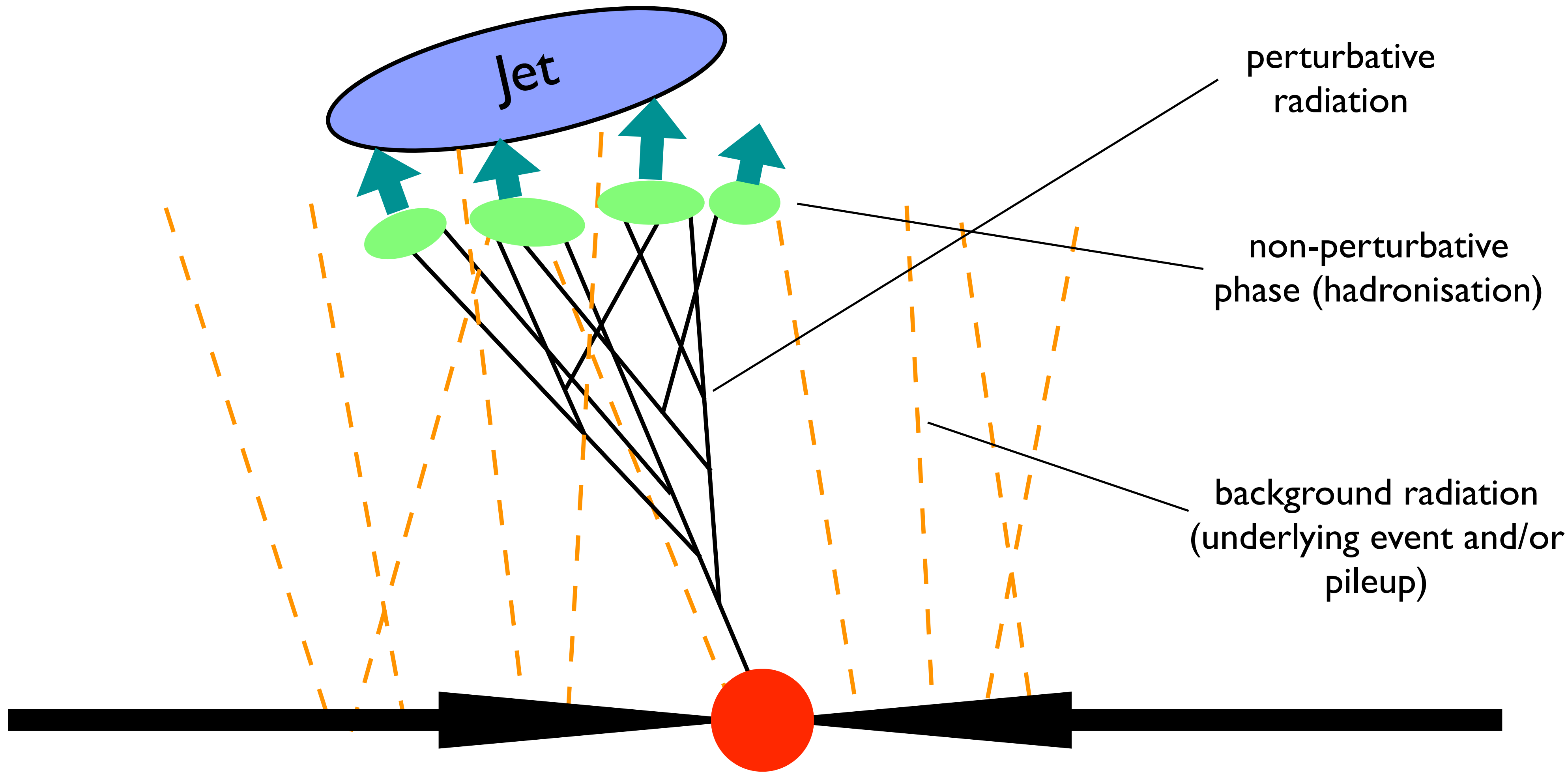
Lecture 2

Matteo Cacciari
LPTHE Paris

Outline

- Lecture 2: jet properties and substructure
  - *Corrections to a jet transverse momentum*
    - *jet areas*
    - *background estimation and subtraction*
  - *Jet substructure techniques*
    - *taggers and groomers for boosted objects*

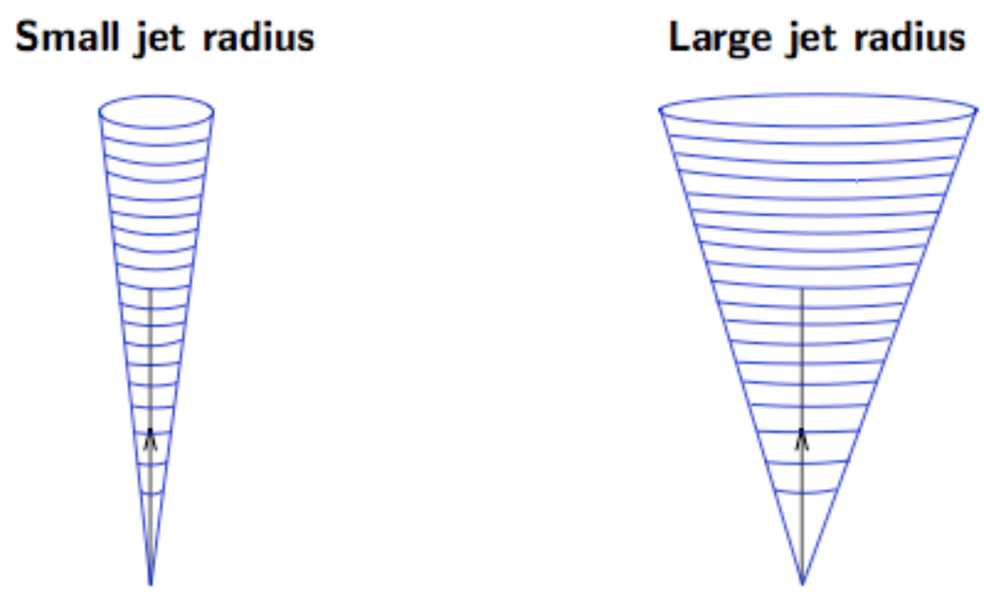
What contributes to a jet's transverse momentum?



Contributions to a jet p_t

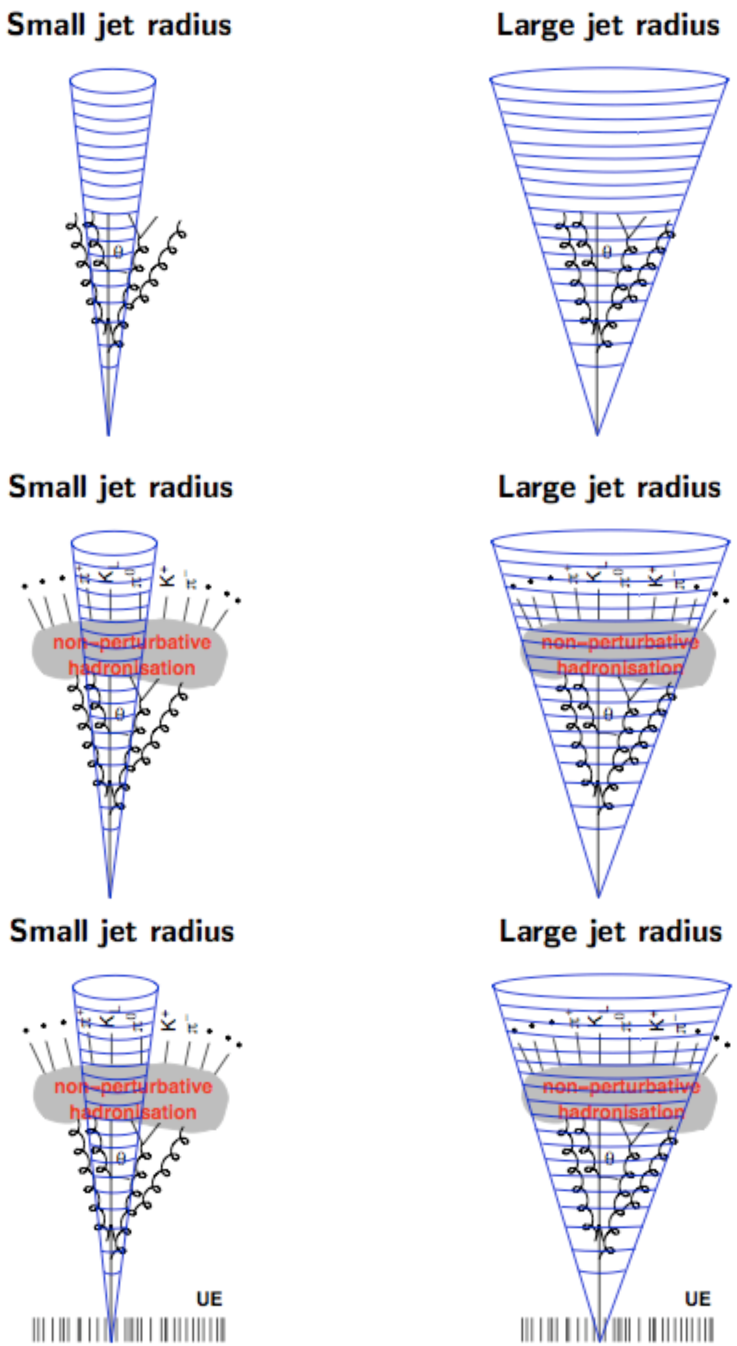
- Ideally, we'd like a jet to contain all radiation emitted by a parton, i.e. reproduce exactly the parton. This is impossible for a number of reasons:
  - First and foremost, *a parton is not a physical observable, while a jet is*. There cannot be exact equivalence between them
  - A *jet has finite extent*, and part of the radiation (perturbative or non-perturbative) emitted by the parton may end up beyond the jet's boundaries, leading to a decrease of the jet momentum (with respect to the parton's)
  - A jet does not fragment in a vacuum. *Background radiation* (underlying event and/or pileup) can affect its momentum in at least two ways:
    - Some background radiation will be clustered with the jet, increasing its momentum
    - Some hard particles (from the parton fragmentation process) may not cluster with the rest of the jet because of the disturbing presence of other nearby particles (they will form other jets with them). Alternatively, hard particles that in the vacuum would not have clustered with the jet will

Effects of jet 'radius'



Irrelevant for a single-particle jet

Effects of jet ‘radius’



perturbative radiation:
large radius better (lose less)

non-perturbative hadronisation:
large radius better (lose less)

underlying event:
large radius worse (capture more)

R-dependent effects

Perturbative radiation: $$\Delta p_t \simeq \frac{\alpha_s(C_F, C_A)}{\pi} p_t \ln R$$

Hadronisation: $$\Delta p_t \simeq -\frac{(C_F, C_A)}{R} \times 0.4 \text{ GeV}$$

Underlying Event: $$\Delta p_t \simeq \frac{R^2}{2} \times (2.5 \text{ — } 15 \text{ GeV})$$

Tevatron LHC

(small-R limit results)

Analytical estimates: Dasgupta, Magnea, Salam, arXiv:0712.3014

Jets ‘reach’

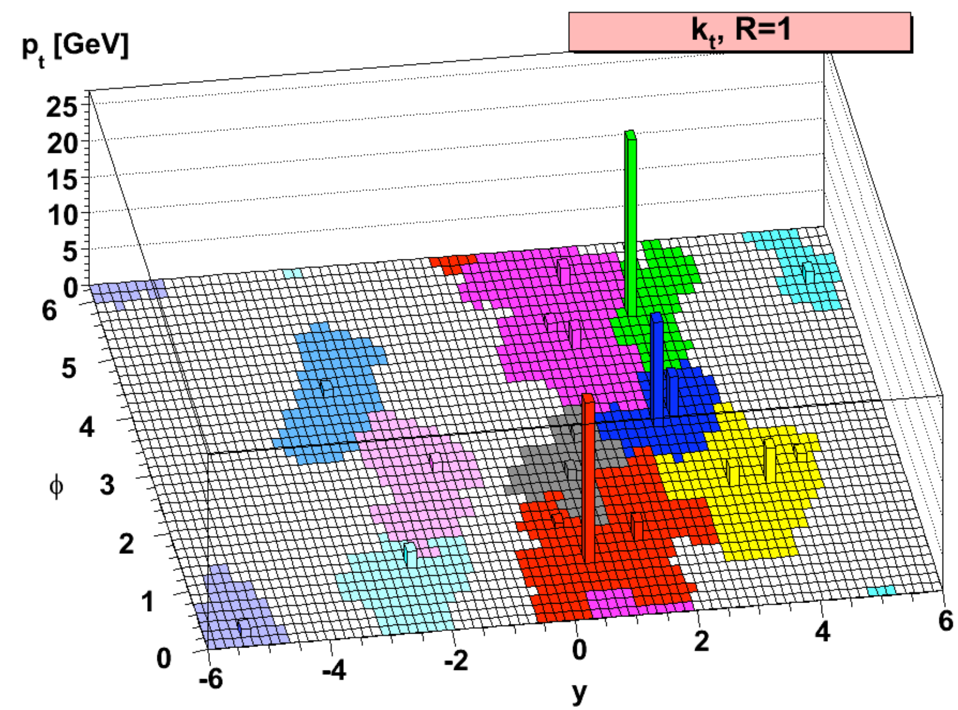
Algorithmically, a jet is simply a collection of particles

For a number of reasons, it is however useful to consider its **spatial extent**, i.e. given the position of its axis, up to where does it collect particles? What is its shape?

These details are important for a number of corrections of various origin: perturbative, non-perturbative (hadronisation), detector related, etc

Note that the intuitive picture of a jet being a cone (of radius R) is **wrong.** This is what k_t jets can look like:

(more later about what this plot really means)



From jet 'reach' to jet areas

MC, Salam, Soyez, 0802.1188

Not one, but three **definitions** of a jet's size:

- **Passive** area

  Place a single soft particle in the event, measure the extent of the region where it gets clustered within a given jet

  Reach of jet for **pointlike** radiation

- **Active** area

  Fill the events with many soft particles, cluster them together with the hard ones, see how many get clustered within a given jet

  Reach of jet for **diffuse** radiation

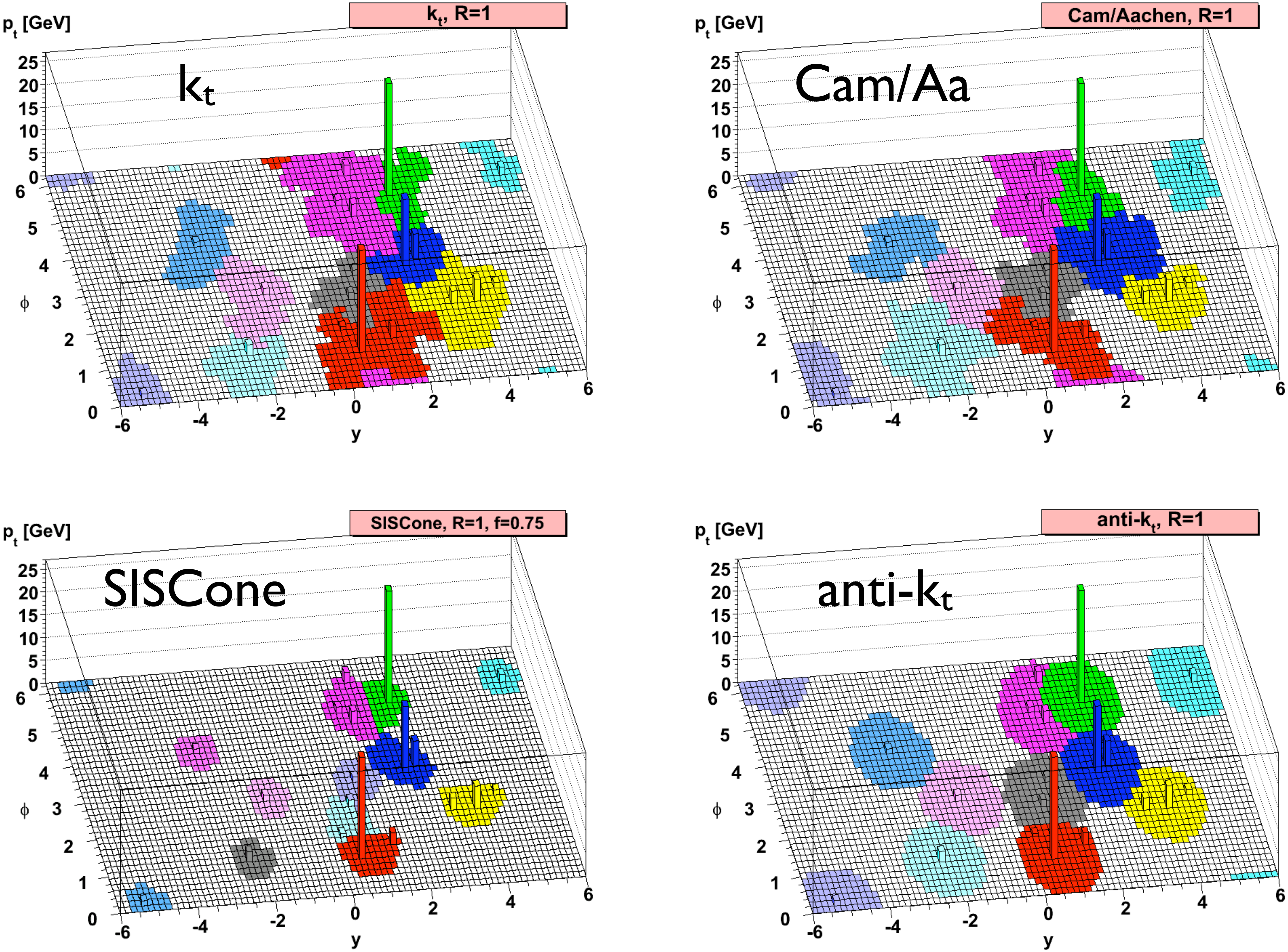
- **Voronoi** area

  Sum of areas of intersections of Voronoi cells of jet constituents with circle of radius R centred on each constituent

  Coincides with passive area for k_t algorithm

(In the large number of particles limit all areas converge to the same value)

```
 /// specify where to place the ghosts, their area, how many times
 /// to repeat the clustering
 double rapmax = 2.5;
 int nrepeat = 1;
 double ghost_area = 0.01;
 GhostedAreaSpec gas(rapmax, nrepeat, ghost_area);

 /// use this configuration to define the area
 /// A sensible default for gas is provided
 AreaDefinition area_def(active_area, gas);

/// construct cluster sequence with areas
  ClusterSequenceArea(input_particles, jet_def, area_def);
  ....
  jets[0].area()
```



kt, R=1
kt
Cam/Aachen, R=1
Cam/Aa
SISCone, R=1, f=0.75
SISCone
anti-kt, R=1
anti-kt
pt [GeV]
φ
y

Jet areas: the single hard particle case

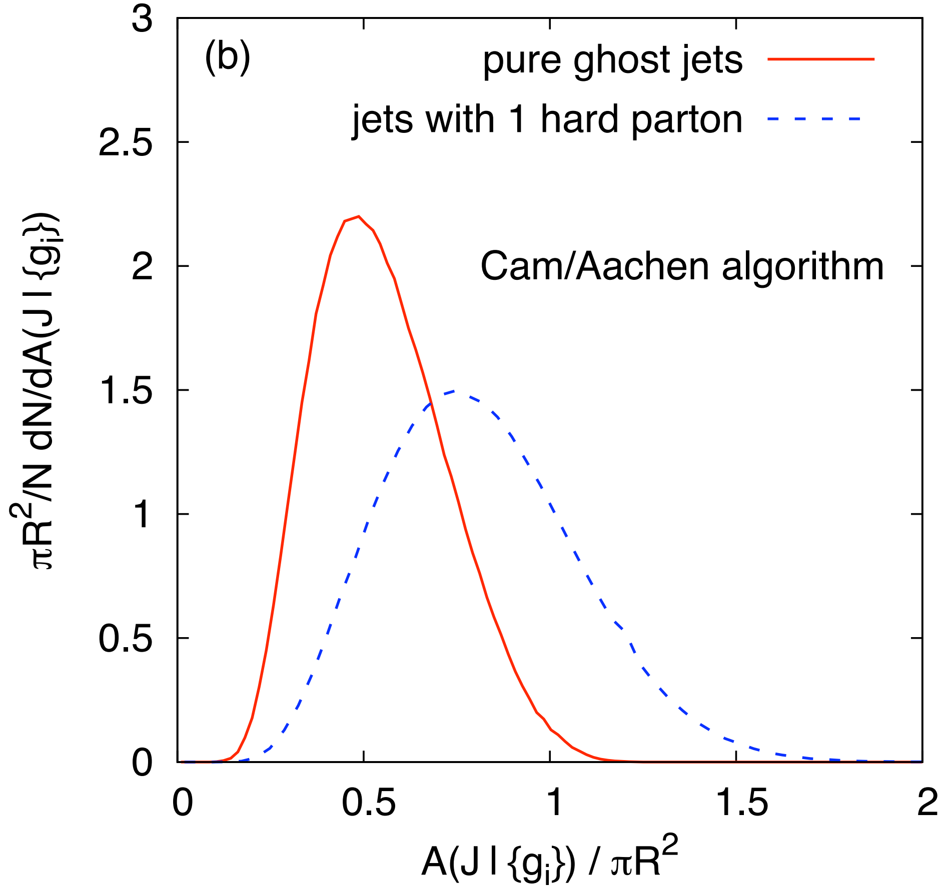
It is worth noting that, for a jet made of a **single hard particle**, while **passive** areas are indeed **πR^2, active** areas are **not**

Active areas	k_t	Cam/Aa	SISCone	anti-k_t
$\langle A \rangle / \pi R^2$	0.81	0.81	1/4	1

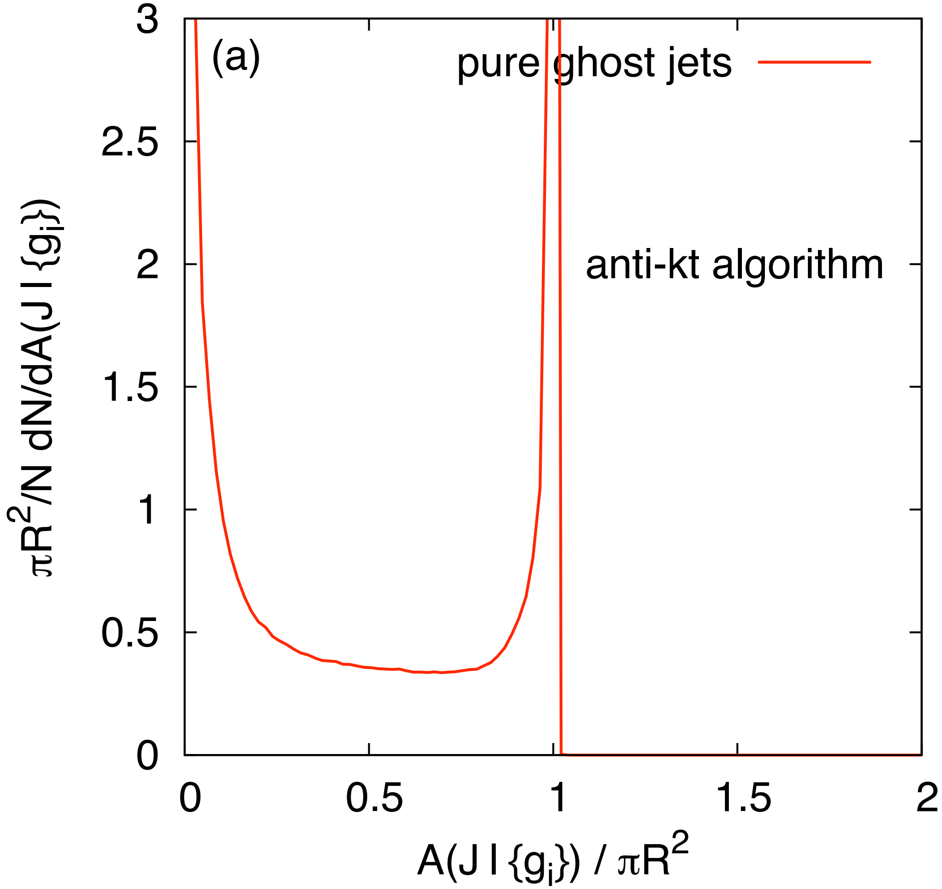
Only anti-k_t has the behaviour one would naively expect, i.e. area = πR^2

Active area distributions

k_t and Cam/AA



anti-k_t



For a roughly uniformly soft background, anti-k_t gives many small jets and many large ones

(you can't fill a plane with circles!)

Jet area: summary

- Jets CAN have an area, but one must define it
- The jet (active) area expresses the susceptibility of a jet to contamination from a uniform background
- Different jet algorithms can have very different area properties:
  - Jet areas in many algorithms can fluctuate significantly from a jet to another. Isolated hard jets in anti-k_t are one exception
  - Jet areas can depend on a jet's p_t, driven by a (calculable) anomalous dimension that is specific to each jet algorithm. Anti-k_t jets are again an exception, in that the anomalous dimension is zero

Effect of background

How are the hard jets modified by the background?

Susceptibility
(how much bkgd gets picked up)

Jet areas

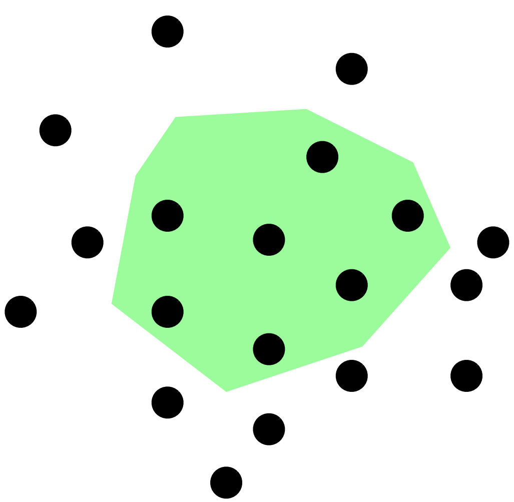
Resiliency
(how much the original jet changes)

Backreaction

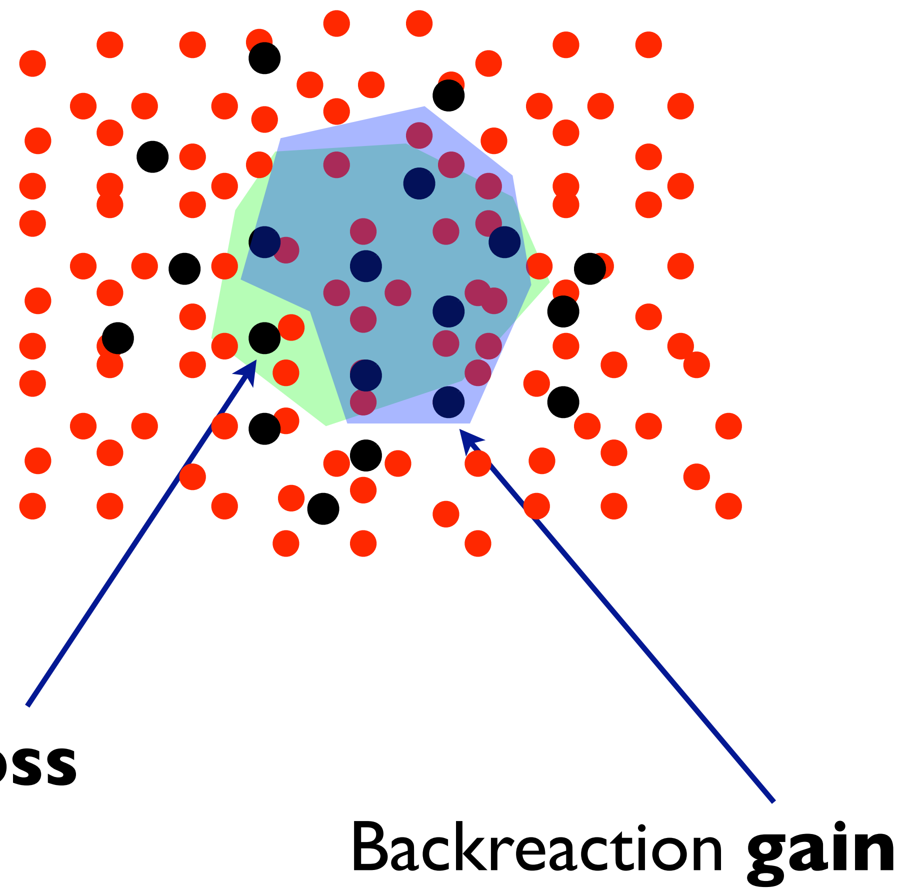
Resiliency: backreaction

"How (much) a jet changes when immersed in a background"

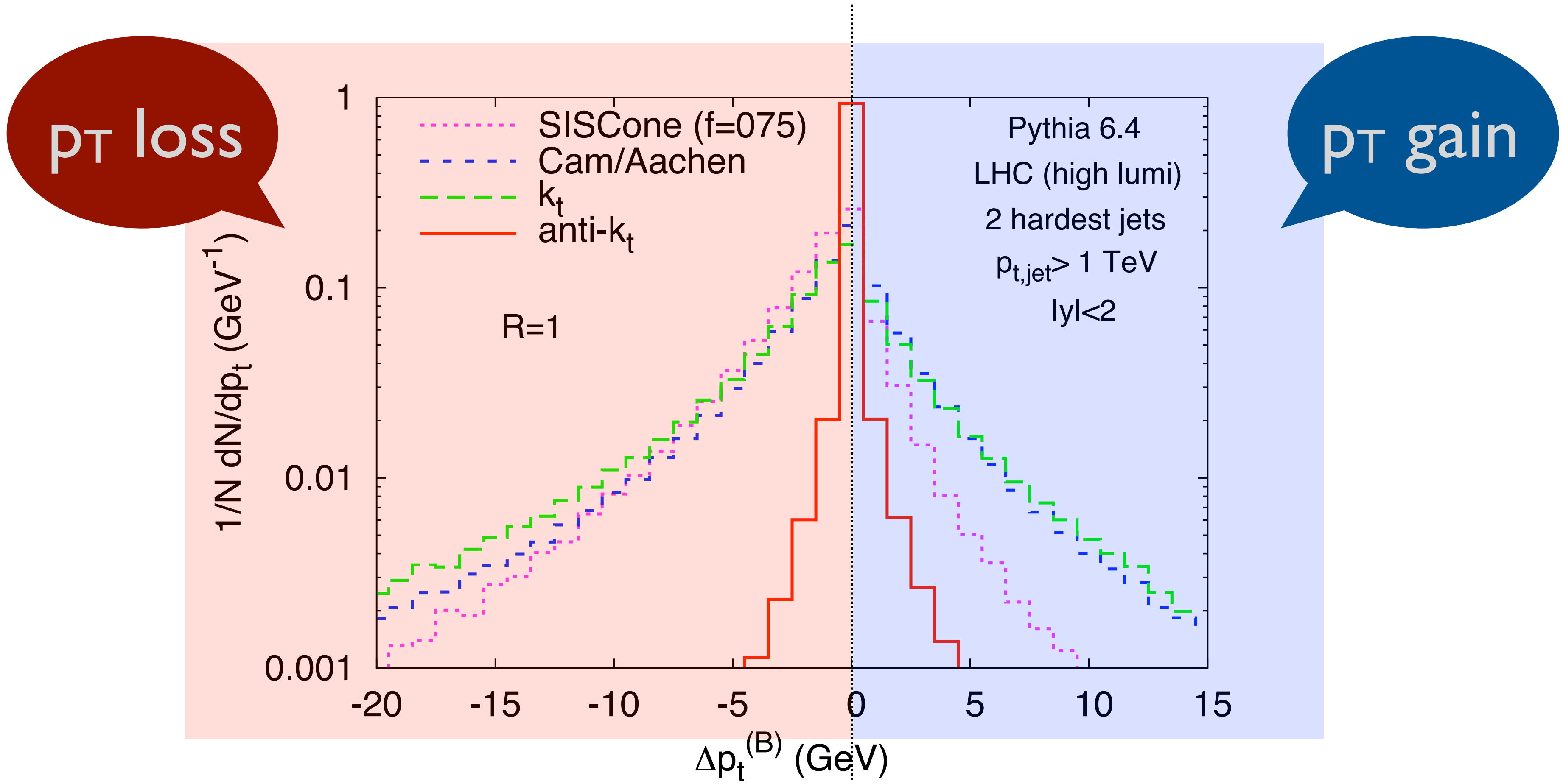
Without background



With background



Resiliency: backreaction



Anti-k_t jets are much more resilient to changes from background immersion

(NB. Backreaction is a minimal issue in pp background and at large pt.
Can be much more important in Heavy Ion collisions)

Hard jets and background

Modifications of the hard jet

$$\Delta p_t = \rho A \pm (\sigma\sqrt{A} + \sigma_\rho A + \rho\sqrt{\langle A^2\rangle - \langle A\rangle^2}) + \Delta p_t^{BR}$$

Background momentum density (per unit area)

background

'susceptibility'

back-reaction

'resiliency'

Background subtraction

Once ρ has been measured, it can be used to correct the transverse momentum of the hard jets:

$$p_T^{\text{hard jet, corrected}} = p_T^{\text{hard jet, raw}} - \rho \times \text{Area}_{\text{hard jet}}$$

MC, Salam, 0707.1378

If ρ is measured on an event-by-event basis, and each jet subtracted individually, this procedure will remove many fluctuations and generally improve the resolution of, say, a mass peak

$$\Delta p_t = \rho A \pm (\sigma\sqrt{A} + \sigma_\rho A + \rho\sqrt{\langle A^2\rangle - \langle A\rangle^2}) + \Delta p_t^{BR}$$

Irreducible fluctuations: uncertainty of the subtraction

Background estimation and subtraction

```
// constructor for a background estimator
JetMedianBackgroundEstimator bge(Selector sel,
                                 JetDefinition jet_def,
                                 AreaDefinition area_def);

// an alternative background estimator
//GridMedianBackgroundEstimator bge(Selector sel, grid_step);

bge.set_particles(input_particles);
....
double rho = bge.rho(jet);   // extract rho estimation
```

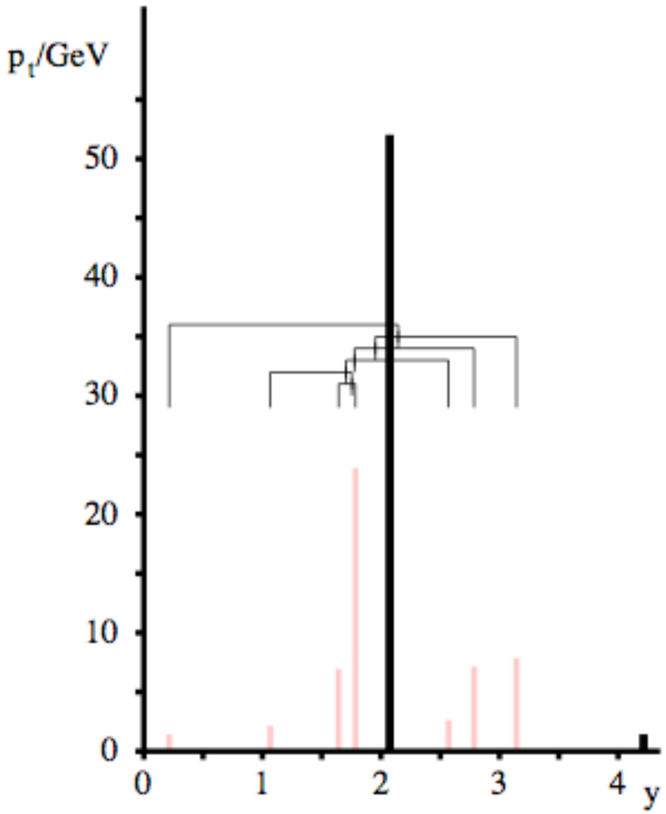
```
// define a subtractor
Subtractor sub(&bge);

// apply it to a jet (or a vector of jets)
PseudoJet subtracted_jet = sub(jet);
```

The IRC safe algorithms

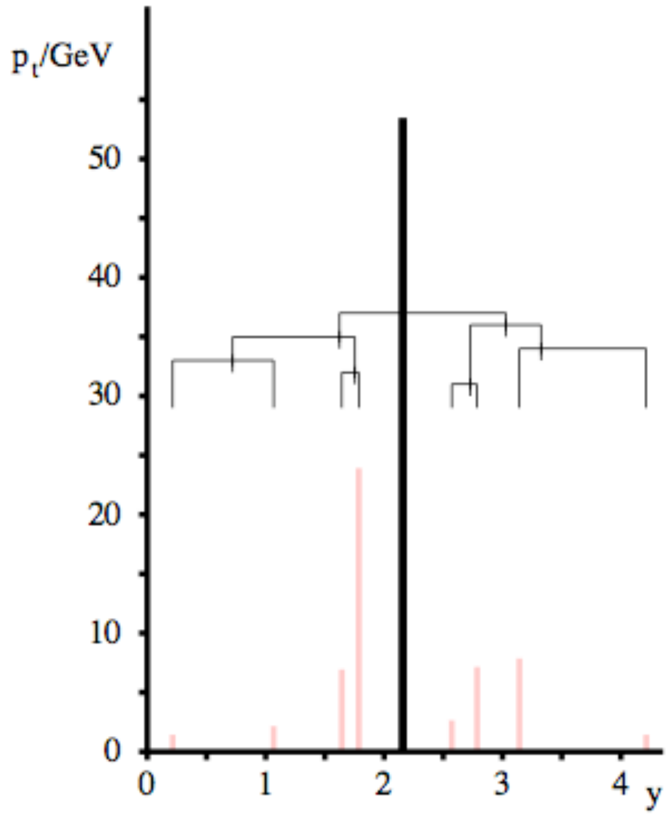
	Speed	Regularity	UE contamination	Backreaction	Hierarchical substructure
k_t	☺ ☺ ☺	☂	☂☂	☁☁	☺ ☺
Cambridge /Aachen	☺ ☺ ☺	☂	☂	☁☁	☺ ☺ ☺
anti-k_t	☺ ☺ ☺	☺ ☺	☁/ ☺	☺ ☺	✘
SISCone	☺	☁	☺ ☺	☁	✘

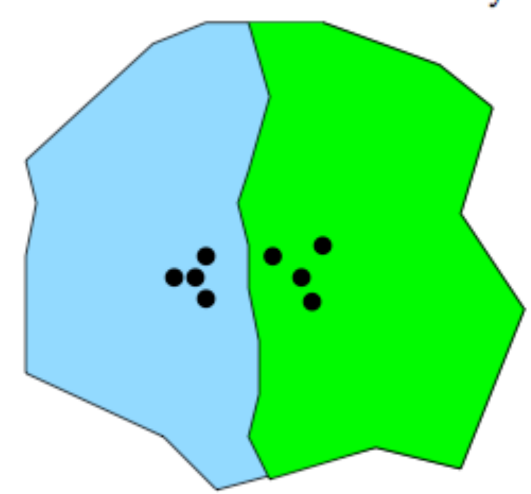
Hierarchical substructure

anti-k_t algorithm

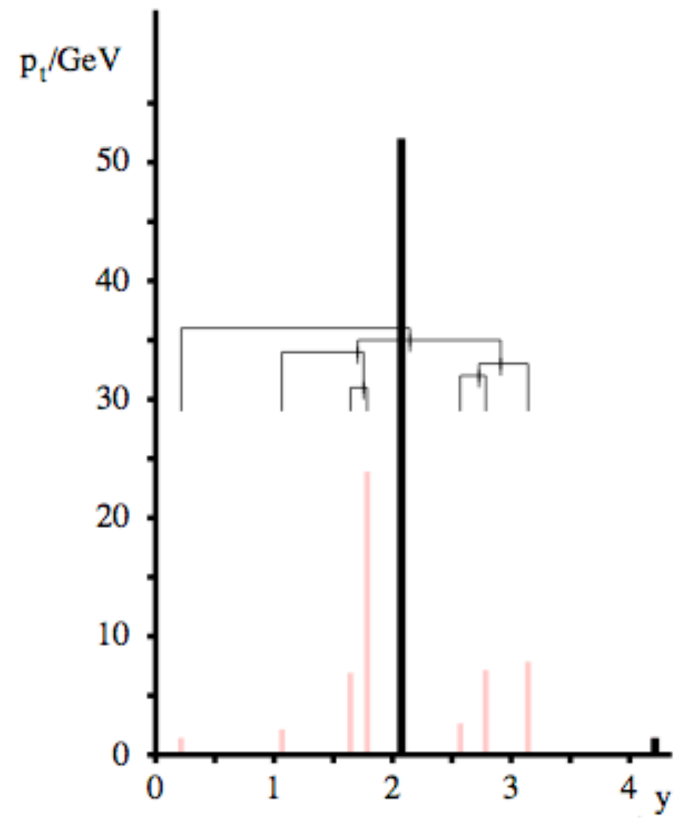


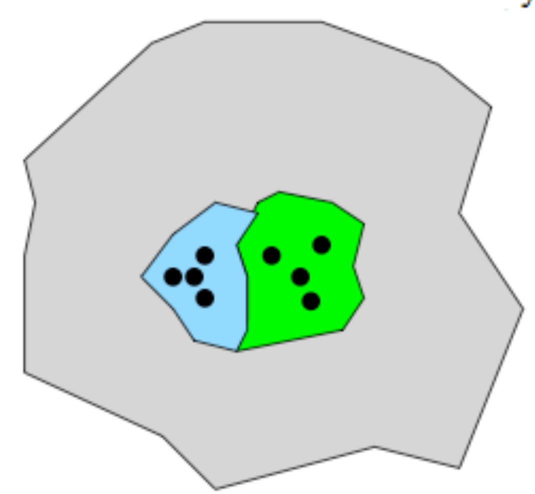
k_t algorithm





Cambridge/Aachen

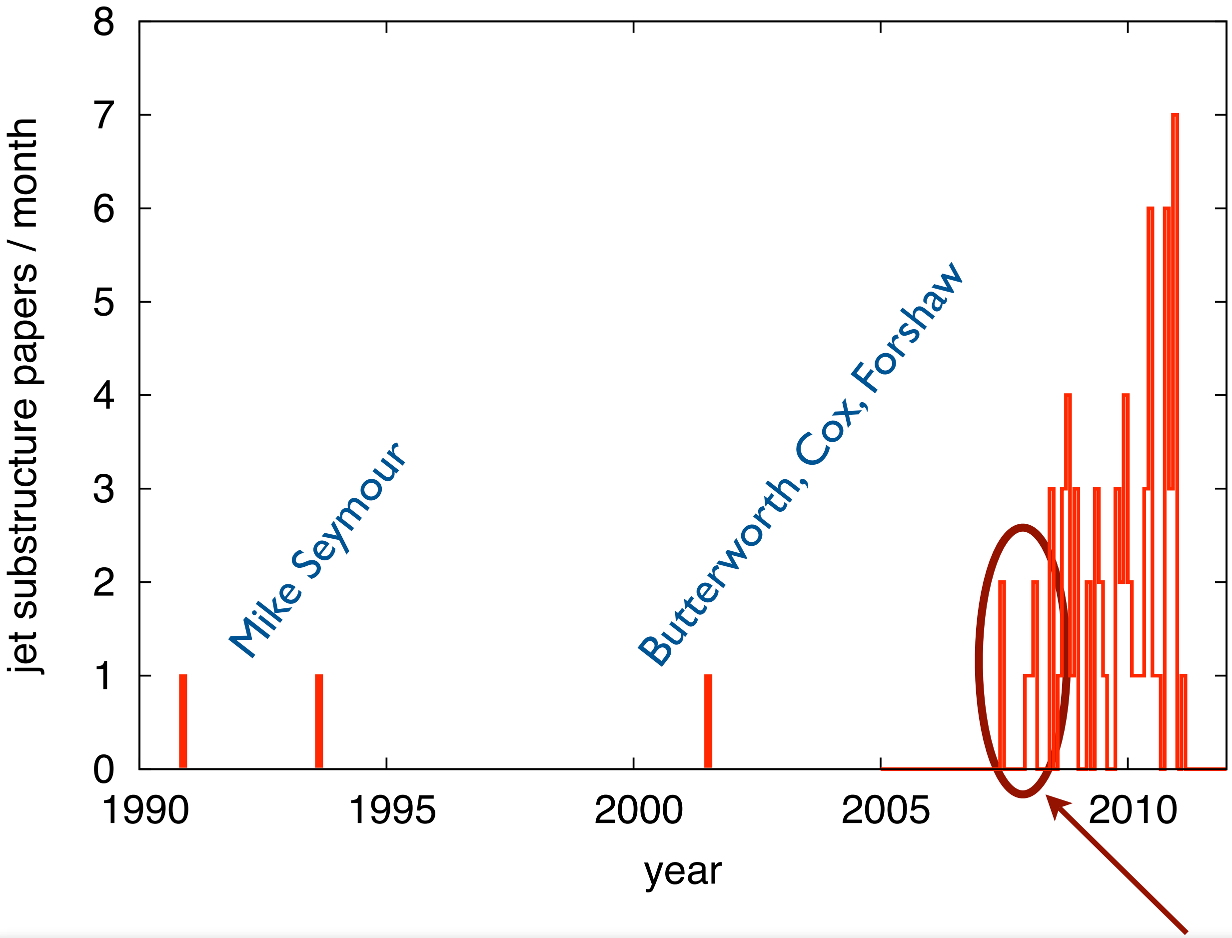




Slide by
Gavin Salam

‘Jet substructure’ papers in SPIRES

Number of papers containing the words ‘jet substructure’ and ‘LHC’



More than 70 papers after 2008.
(the histogram dates from early 2011)

Pioneered by M. Seymour in the early ‘90s, rebooted by BDRS paper

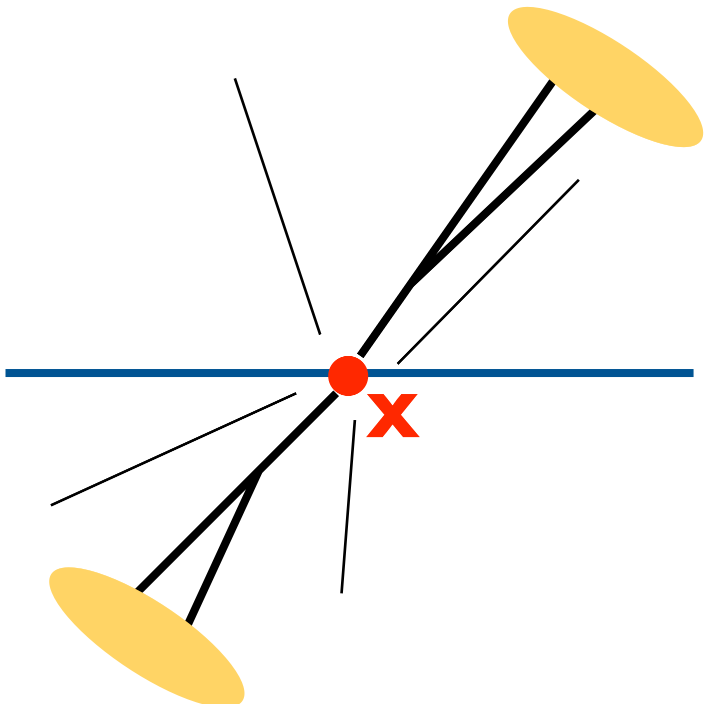
15. **Jet substructure as a new Higgs search channel at the LHC.**
Jonathan M. Butterworth, Adam R. Davison (University Coll. London), Mathieu Rubin, Gavin P. Salam (Paris, LPTHE).
Published in **Phys.Rev.Lett. 100 (2008) 242001**
e-Print: **arXiv:0802.2470 [hep-ph]**

Jet substructure

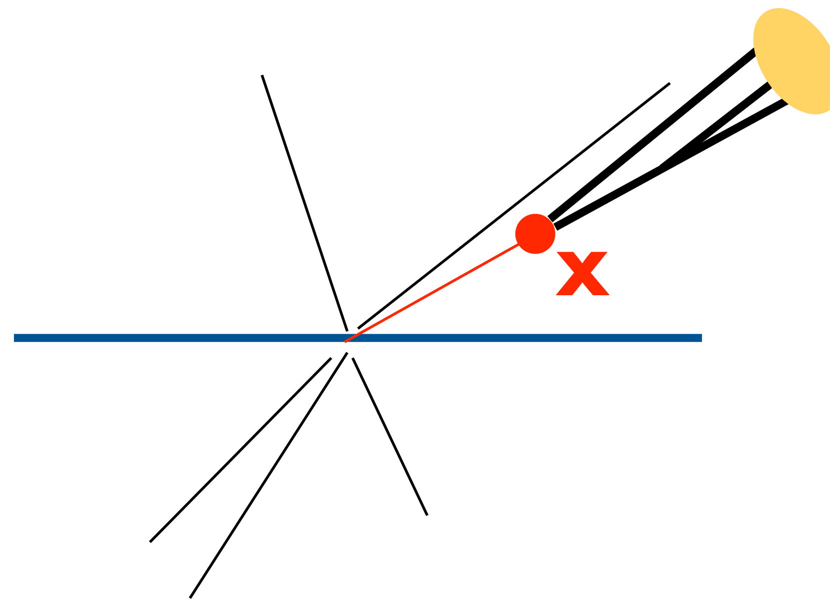
- The substructure of a jet (i.e. the ability to further resolve smaller components) can be exploited to
  - *tag a particular structure inside the jet, i.e. a massive particle*
    - *Examples: Higgs (2-prongs decay), top (3-prongs decay)*
  - *remove background contamination from the jet or its components*
    - *Examples: filtering, trimming, pruning*

In the following I'll be mainly illustrating the BDRS tagger/filter as a pedagogical example, and also list other approaches

Why boosted objects



Heavy particle X at **rest**



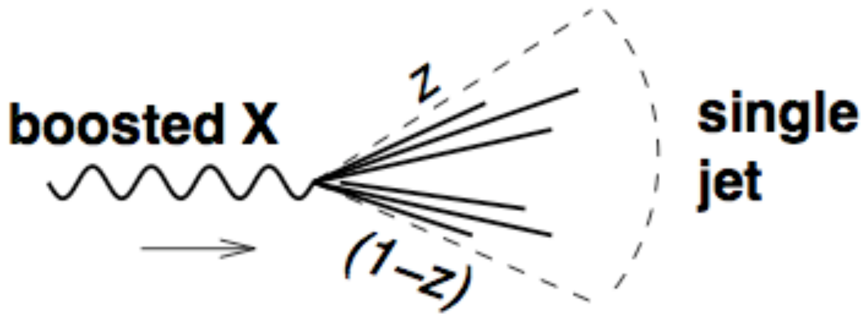
Boosted heavy particle X

Easy to resolve jets and calculate invariant mass, but signal very likely swamped by background (eg H→bb v. tt →WbWb)

Cross section very much reduced, but acceptance better and some backgrounds smaller/ reducible

Why substructure

Scales: m ~ 100 GeV, p_t ~ 500 GeV



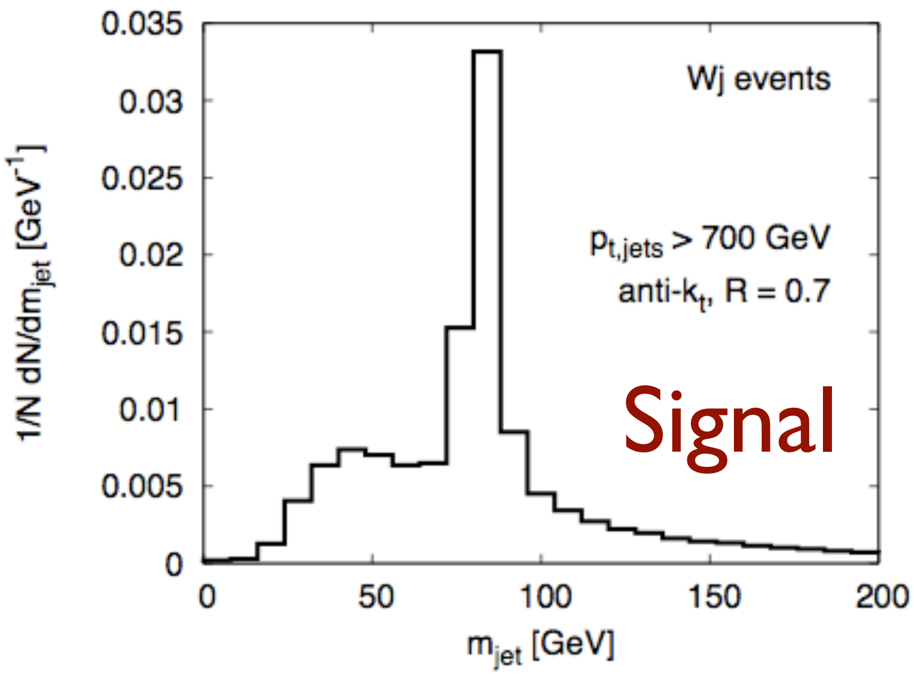
$$R \gtrsim \frac{m}{p_t} \frac{1}{\sqrt{z(1-z)}}$$

- need small R (< 2m/p_t ~ 0.4) to resolve two prongs
- need large R (>~ 3m/p_t ~ 0.6) to cluster into a single jet

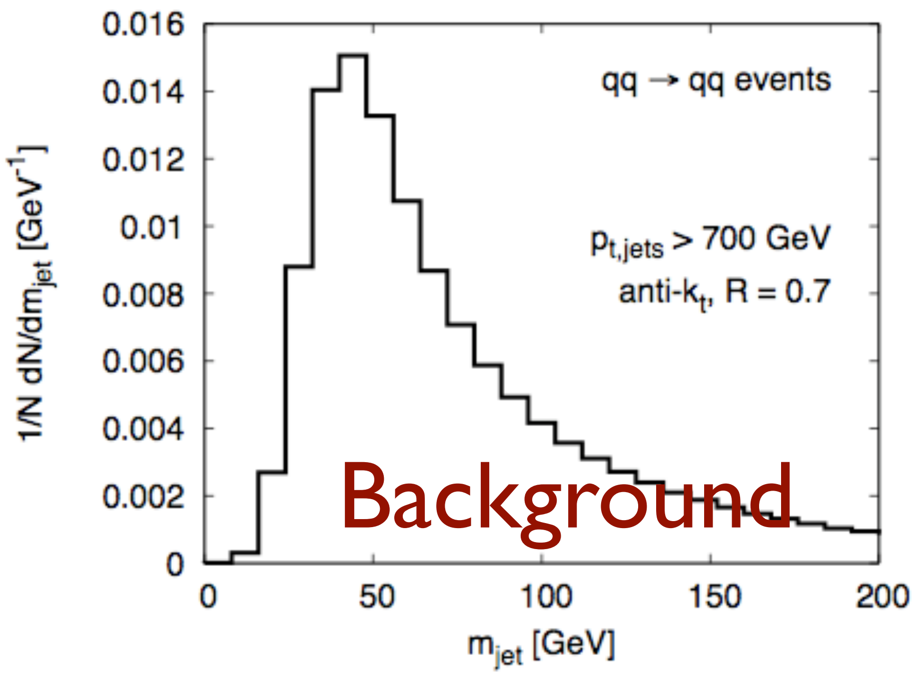
Possible strategies

- Use large R, get a single jet : background large
- Use small R, resolve the jets : what is the right scale?
- Let an algorithm find the ‘right’ substructure

G. Salam



A heavy object decaying into a single jet naturally gives it a mass...

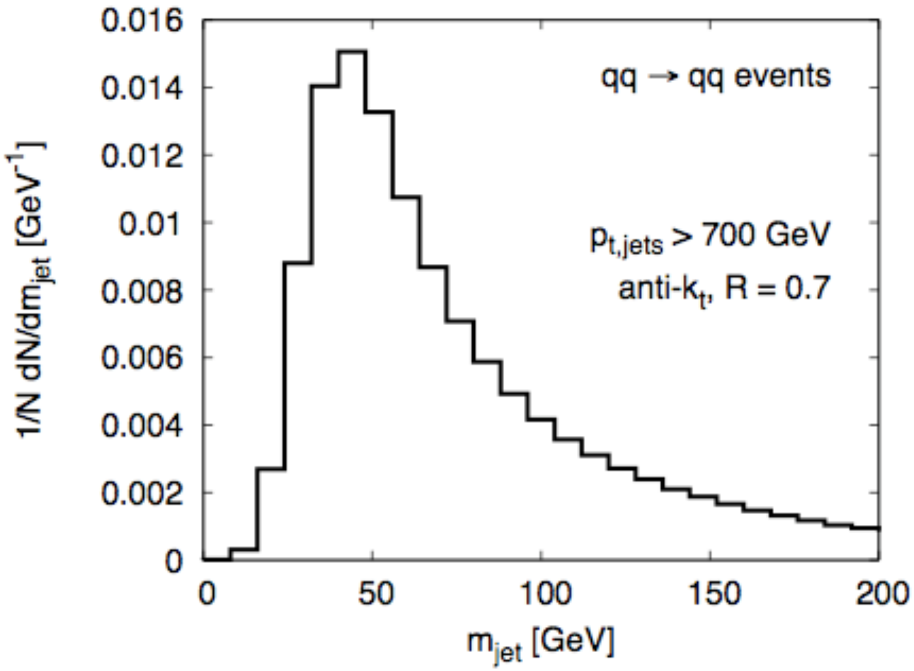


... but pure QCD jets can be massive too:

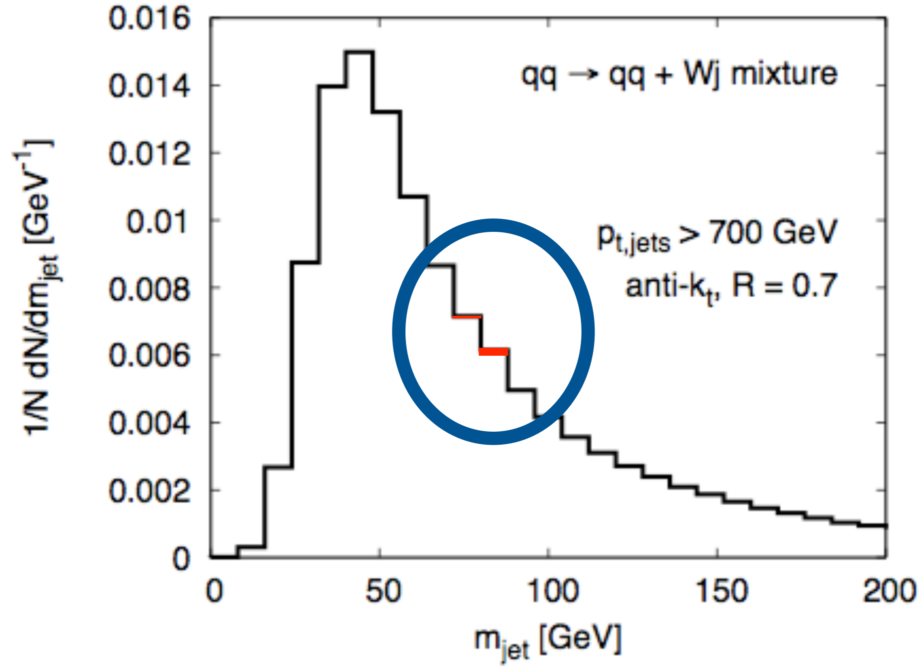
$$\frac{dN}{d\ln m} \sim \alpha_s \ln \frac{p_t R}{m} \times \text{Sudakov}$$

Mass of a single jet

Summing 'signal' and 'background' (with appropriate cross sections) shows how much the background dominates



Background only



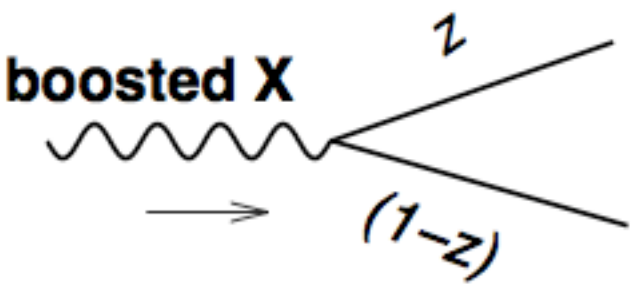
Signal + background

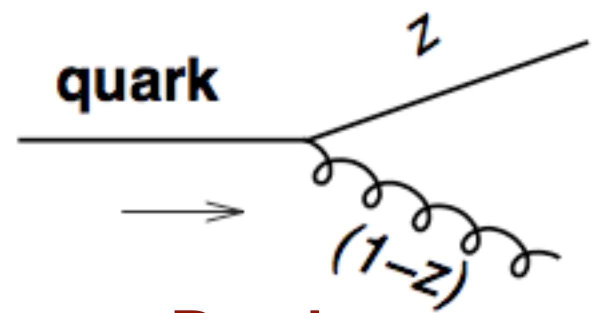
Practically identical

This means that one can't rely on the invariant mass only.
An appropriate strategy must be found to reduce the background and enhance the signal

QCD v. heavy decay

A possible approach for reducing the QCD background is to identify the two prongs of the heavy particle decay, and put a cut on their momentum fraction





($p_{tj} = zp_t$)

($p_{ti} = (1\text{-}z)zp_t$)

Signal:

$$P(z) \sim \text{constant}$$

Will split mainly **symmetrically**

Background:

$$P(z) \sim \frac{1+z^2}{1-z} \qquad P(z) \sim \frac{1+(1-z)^2}{z}$$

Will split mainly **asymmetrically**

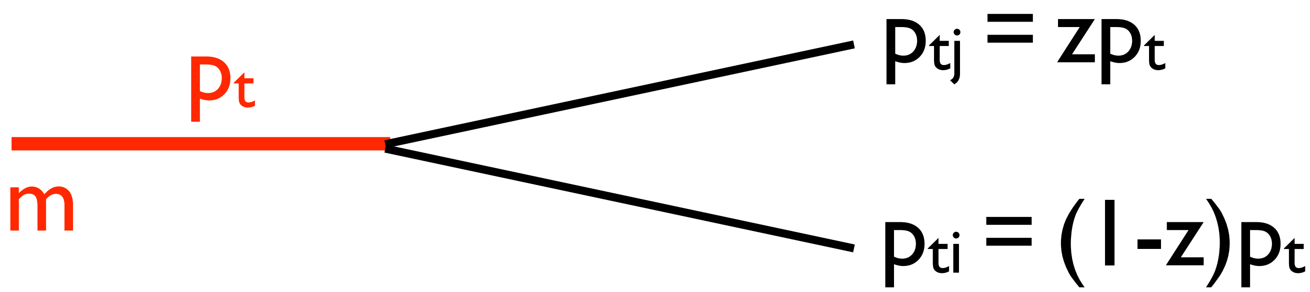
Potential tagger: asymmetric splitting

Possibly implemented via a cut on

$$y = min(p_{ti}^2, p_{tj}^2)\frac{\Delta R_{ij}^2}{m^2} \simeq \frac{min(p_{ti}, p_{tj})}{max(p_{ti}, p_{tj})}$$

Splittings and distances

Quasi-collinear splitting ($p_{tj} < p_{ti}$)



Invariant mass:

$$m^2 \simeq p_{ti} p_{tj} \Delta R_{ij}^2 = (1-z) z p_t^2 \Delta R_{ij}^2$$

k_t distance:

$$d_{ij} \overset{(p_{tj} < p_{ti})}{=} z^2 p_t^2 \Delta R_{ij}^2 \simeq \frac{z}{1-z} m^2$$

For a given mass, the background will have smaller distance d_{ij} than the signal, therefore it will tend to **cluster earlier** in the k_t algorithm

Potential tagger: last clustering in k_t algorithm

This is where the hierarchy of the k_t algorithm becomes relevant.
QCD radiation is clustered first, and only at the end the symmetric, large-angle splittings due to decays are reclustered

Alternative algorithms

- Suppose that for some reasons (which will become clearer later) one does not with to use the k_t algorithm
  - *One must then find a way to determine what the **relevant splitting** (i.e. the one due to the decay, not to QCD radiation) is.*

A possible approach is to use a Mass-Drop requirement:
the clustering is **progressively undone**, and a splitting is the relevant one if both subjects are much less massive than their combination

A generic substructure approach to tagging will

- Cluster initially with a large R, so as to collect all the decay products of a boosted heavy particle into a single jet
- Decluster this jet into subjets, using some conditions to decide when to stop the declustering (i.e. find the 'relevant splitting'), possibly including kinematical cuts to reduce the QCD background.
  - *The stopping criterion automatically finds the 'right size' for the distance between the two prongs of the heavy particle decay*
- Optionally add a final 'cleaning' procedure to remove as much as possible spurious soft/background radiation

Generig tagging/grooming

Fat-jet finding
Usually anti-kt, R ≈ 1

↓ large p_t, large mass fat-jet, signal or background

Tagging step

↓ signal jet candidate, still background-contaminated

grooming step

↓ final candidate, potentially with little background contamination

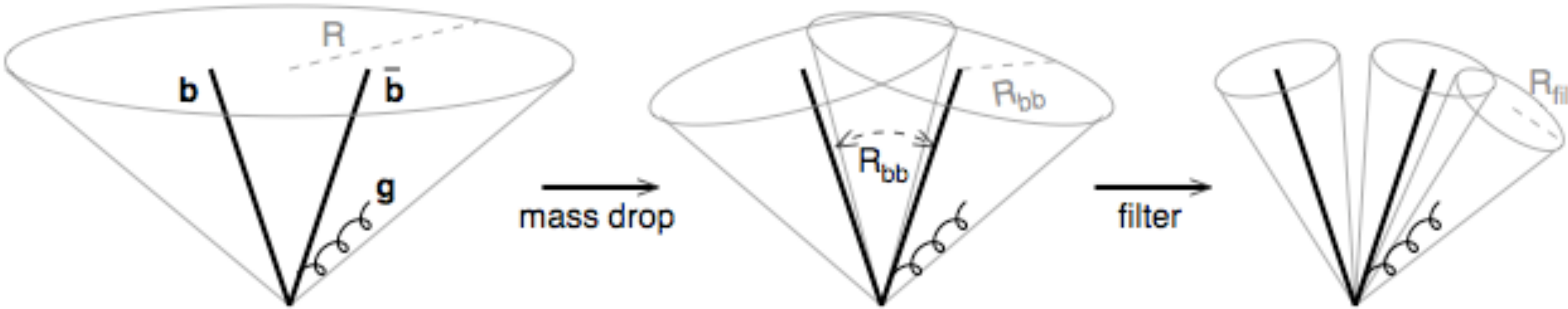
Note that in some taggers (i.e. pruning) the tagging and grooming steps are not explicitly factorised

Also, some tools may actually not follow rigidly this scheme

The BDRS tagger

These ideas led to the first 'modern' implementation of a boosted tagger

15. **Jet substructure as a new Higgs search channel at the LHC.**
Jonathan M. Butterworth, Adam R. Davison (University Coll. London), Mathieu Rubin, Gavin P. Salam (Paris, LPTHE).
Published in **Phys.Rev.Lett. 100 (2008) 242001**
e-Print: **arXiv:0802.2470 [hep-ph]**



It's a two-prongs tagger for boosted Higgs, which

- Uses the **Cambridge/Aachen** algorithm (see why in the next slide)
- Employs a **Mass-Drop** condition (as well as an asymmetry cut) to find the relevant splitting (i.e. 'tag' the heavy particle)
- Includes a post-processing step, using '**filtering**' (introduced in the same paper) to clean as much as possible the resulting jets of UE contamination

Why C/A and not k_t

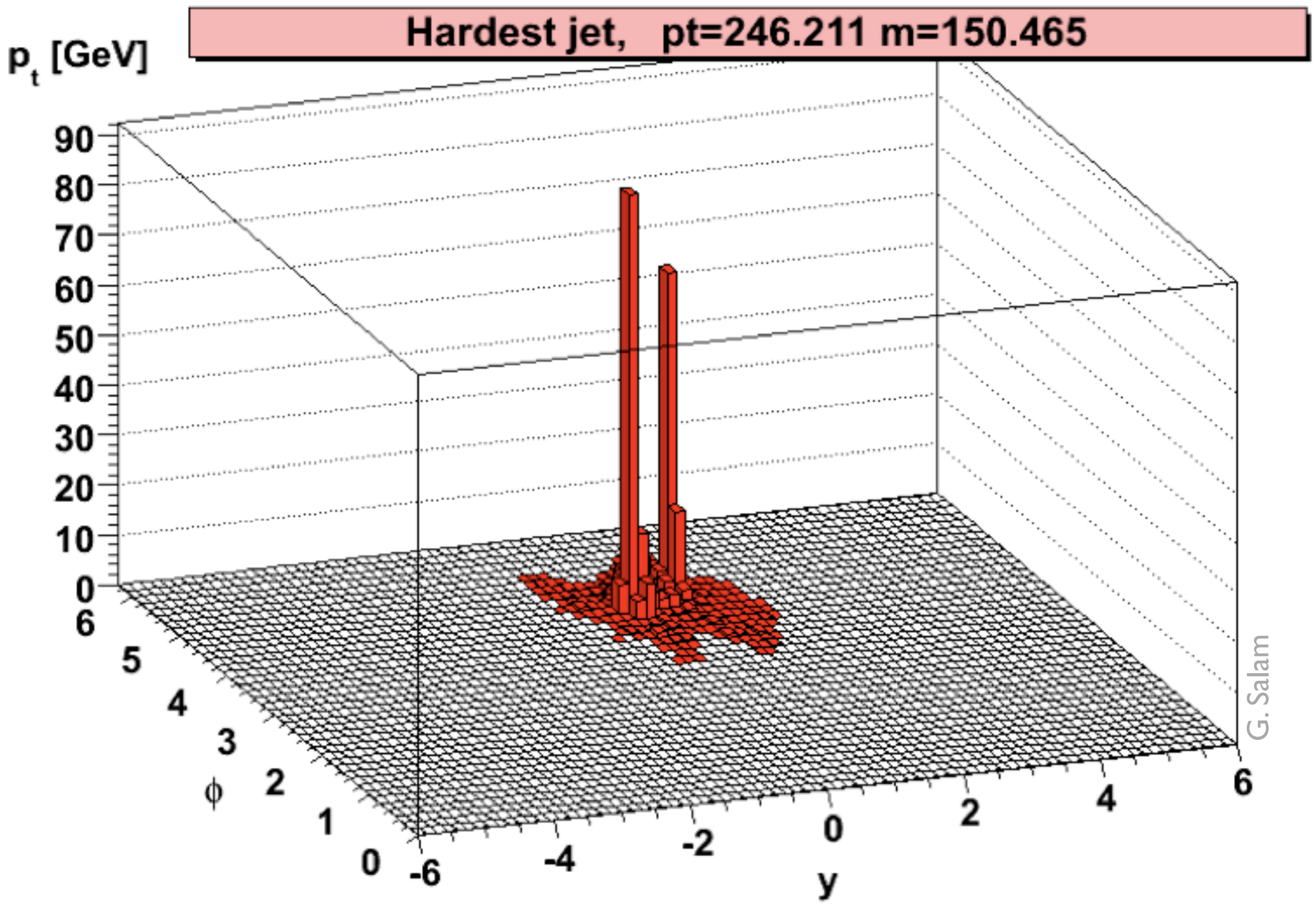
While k_t gives a 'natural relevant splitting' at its first declustering, there are a number of reasons why Cambridge/Aachen has been preferred

- k_t's 'relevant subjets' tend to include more soft radiation than needed, eventually leading to poor resolution (large areas and fluctuations)
- The angle-based clustering distance of Cambridge/Aachen ensures that at the relevant splitting the radii of the jets of the two prongs are similar to the distance between the two prongs themselves. This ensures that, because of angular ordering, these jets contain essentially all the radiations emitted by the decay products of the heavy particles (b quarks, in the case of BDRS)
- Cambridge/Aachen allows one to obtain naturally clustering sequences for any R with a single run, which is useful in the filtering step

Boosted Higgs tagger

$pp \rightarrow ZH \rightarrow \nu\bar{\nu}b\bar{b}$

Butterworth, Davison, Rubin, Salam, 2008



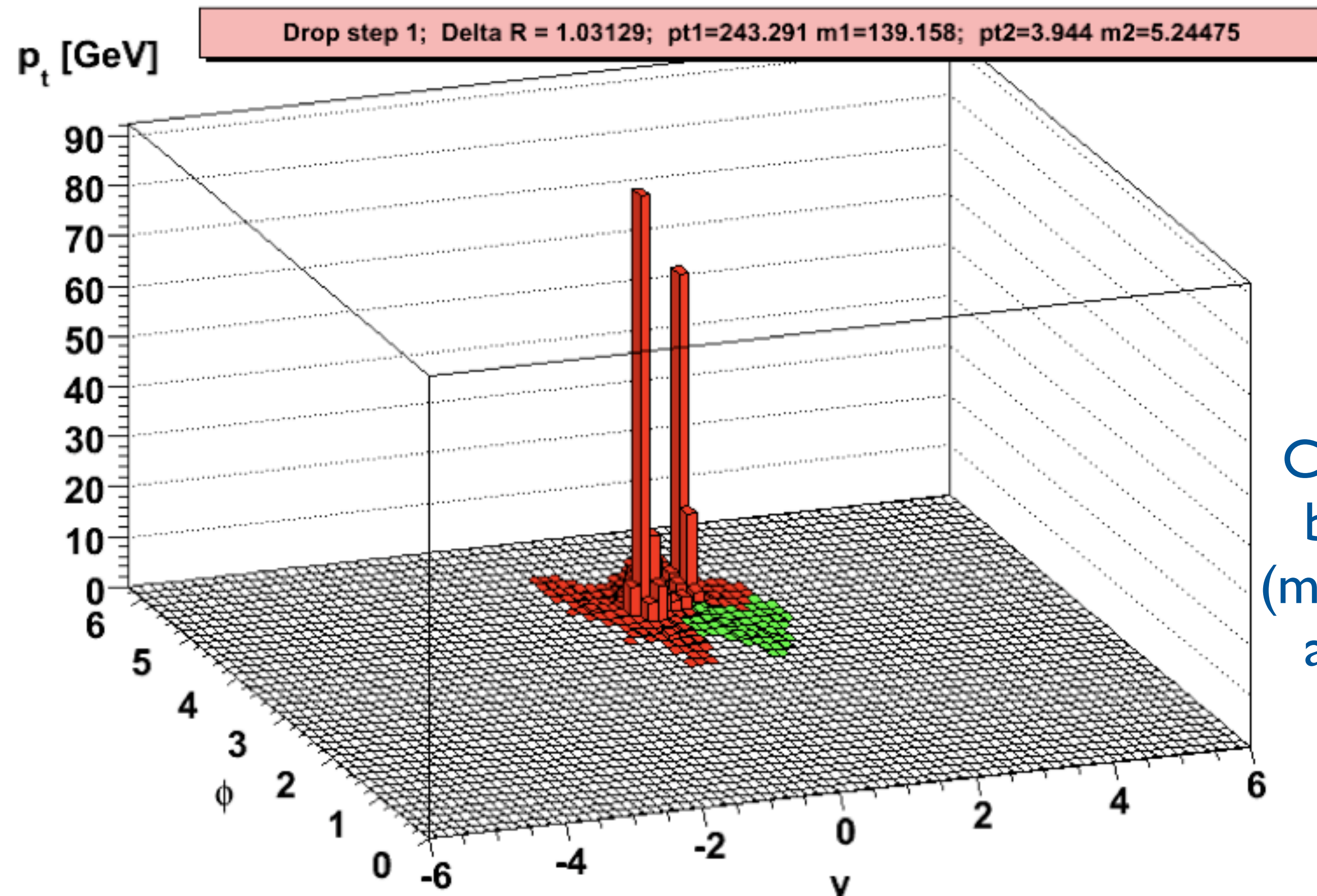
Start with the hardest jet

Use C/A with large R=1.2

m_j = 150 GeV

Boosted Higgs tagger

$pp \to ZH \to \nu\nu bb$



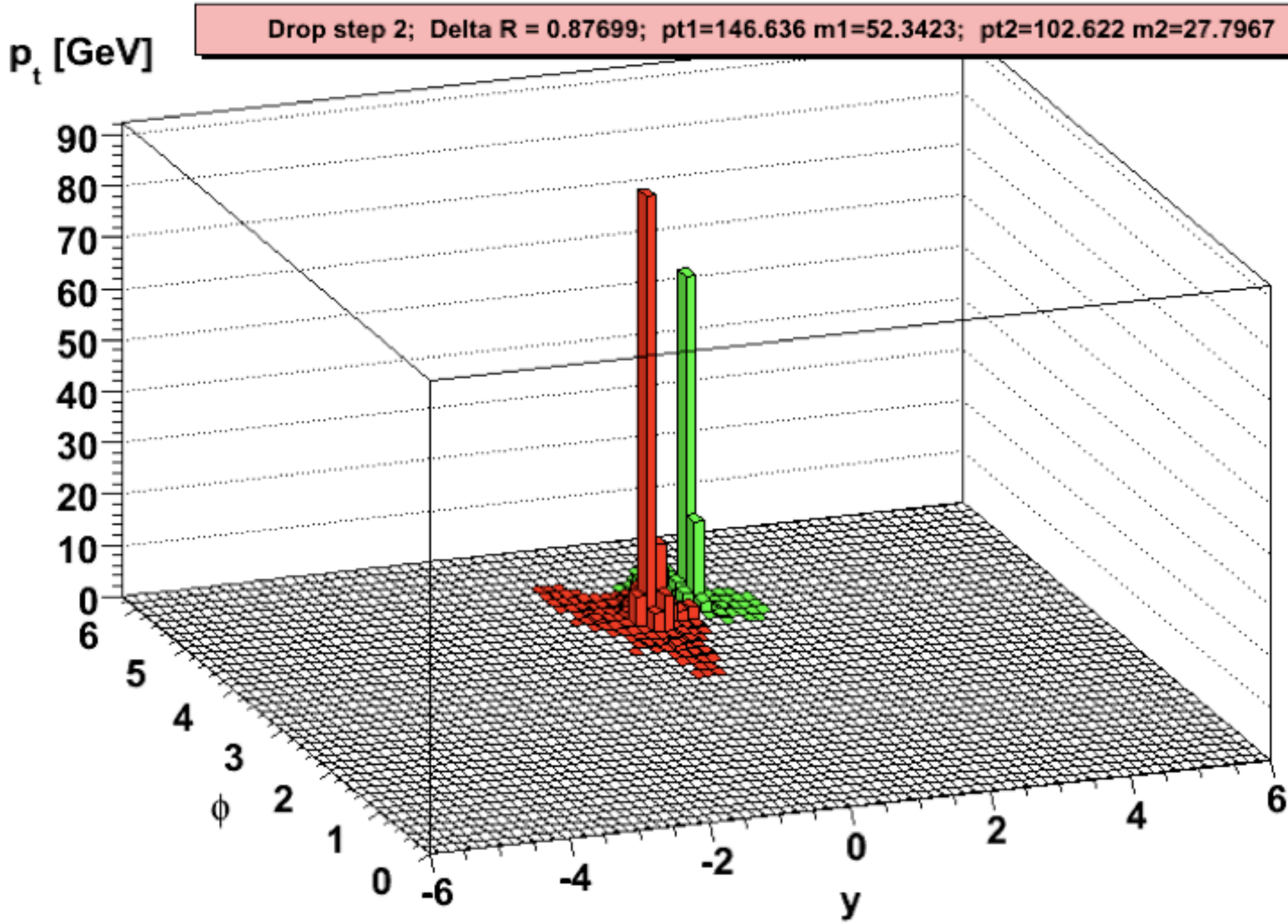
Undo last step of clustering

Check how the mass splits between the two subjets (m_1 = 139 GeV, m_2 = 5 GeV) and how asymmetric the splitting is

If $\frac{max(m_1, m_2)}{m_j} > \mu$ or $\frac{min(p_{t1}^2, p_{t2}^2)}{m_j^2} \Delta R_{12}^2 < y_{cut}$ repeat

Boosted Higgs tagger

$pp \to ZH \to \nu\nu bb$



m_1 = 52 GeV, m_2 = 28 GeV

Stop when a large mass drop is observed

(and recombine these two jets)

[NB. Parameters used μ = 0.67 and y_{cut} = 0.09]

Jet substructure as filter

The ***jet substructure*** can be exploited to help **removing contamination** from a soft background

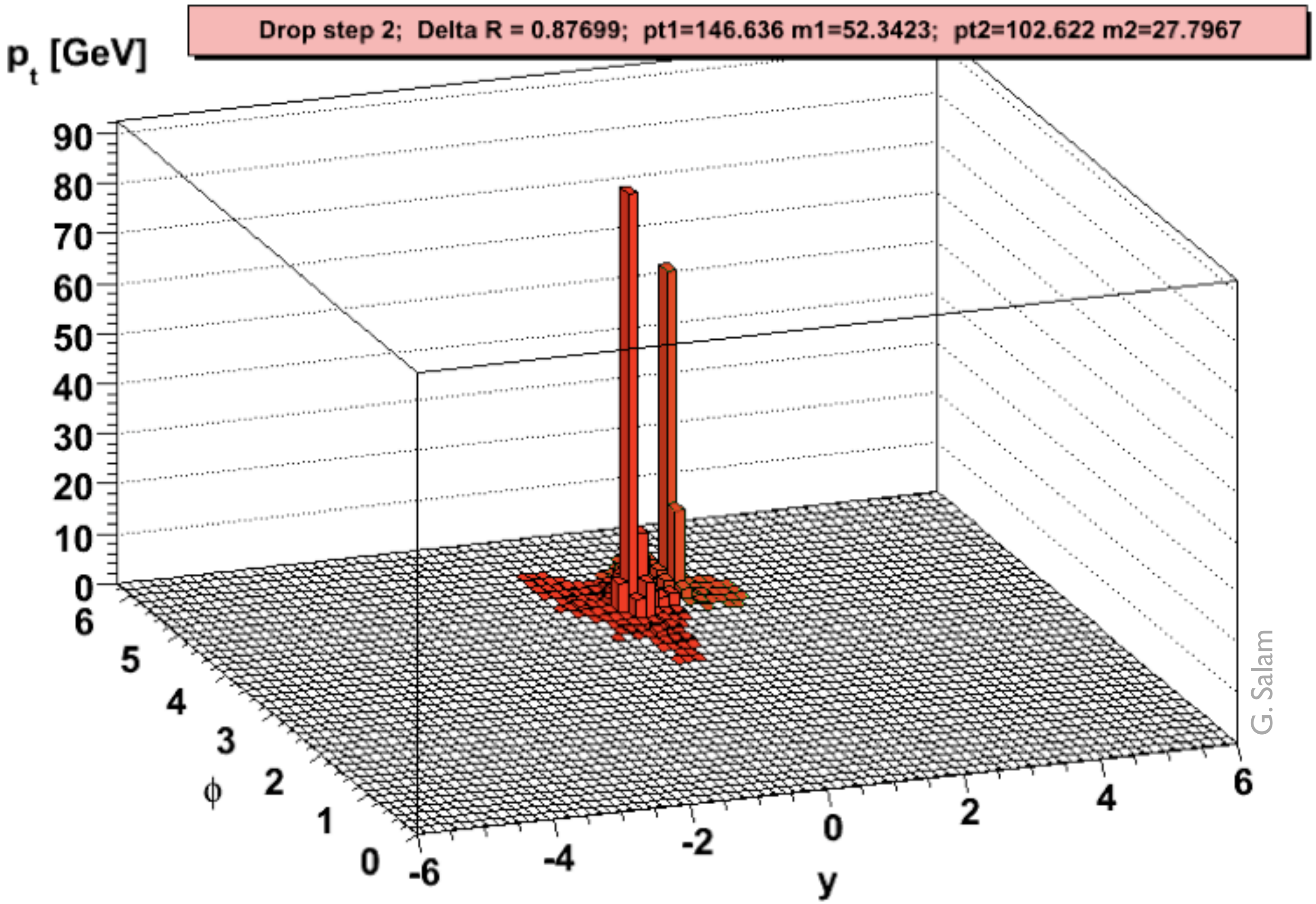
- Jet 'filtering' — Butterworth, Davison, Rubin, Salam, 2008
  Break jet into subjets at distance scale R_{filt}, retain n_{filt} hardest subjets
- Jet 'trimming' — Krohn, Thaler, Wang, 2009
  Break jet into subjets at distance scale R_{trim}, retain subjets with $p_{t,subjet} > \varepsilon_{trim}\, p_{t,jet}$
- Jet 'pruning' — S. Ellis, Vermilion, Walsh, 2009
  While building up the jet, discard softer subjets when $\Delta R > R_{prune}$ and $\min(p_{t1},p_{t2}) < \varepsilon_{prune}\,(p_{t1}+p_{t2})$

Aim: limit sensitivity to background while retaining bulk of perturbative radiation

Filtering, trimming and pruning are in the end effectively quite similar.
These and similar tools are collectively called **groomers**

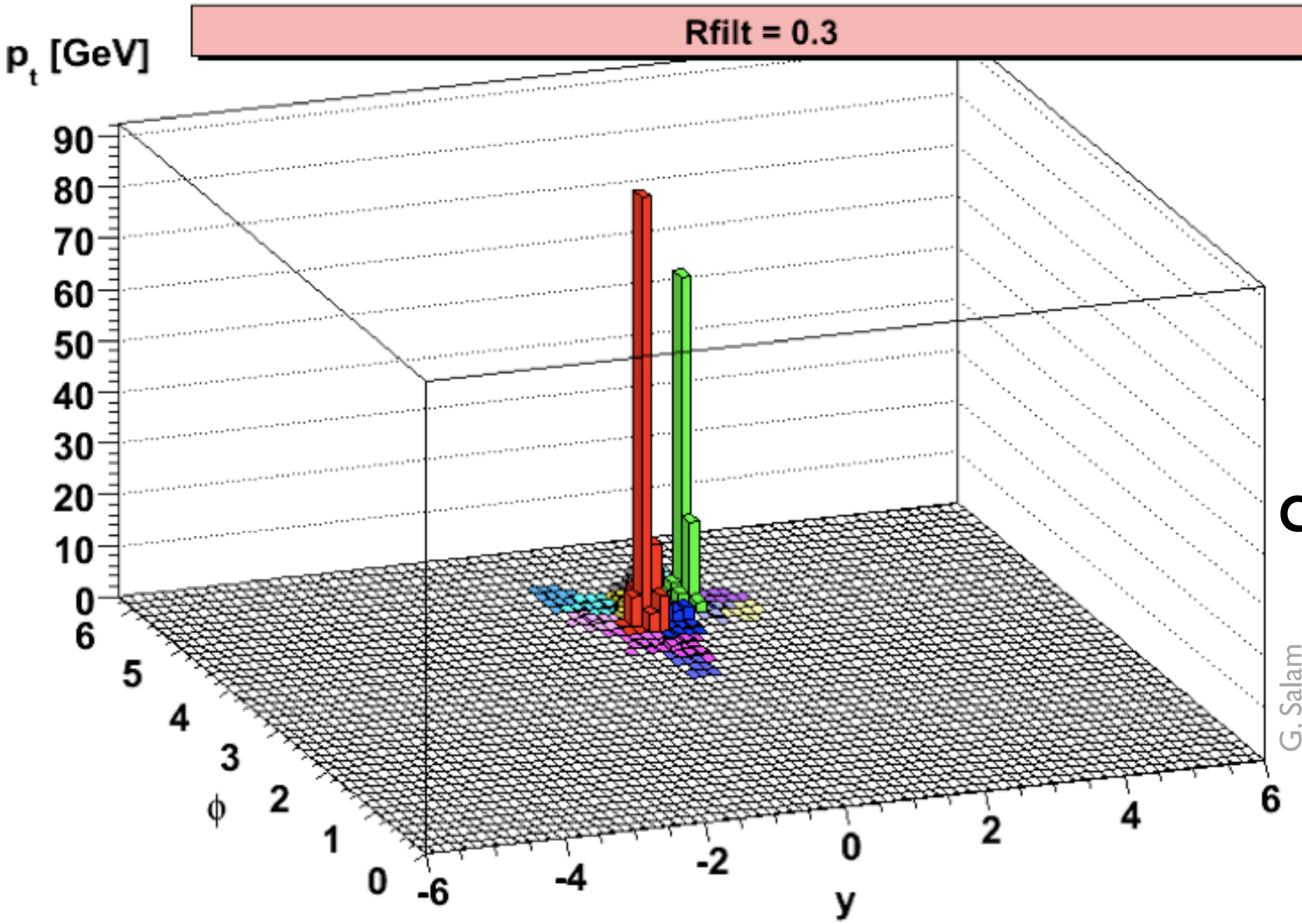
Filtering in action

Butterworth, Davison, Rubin, Salam, arXiv:0802.2470



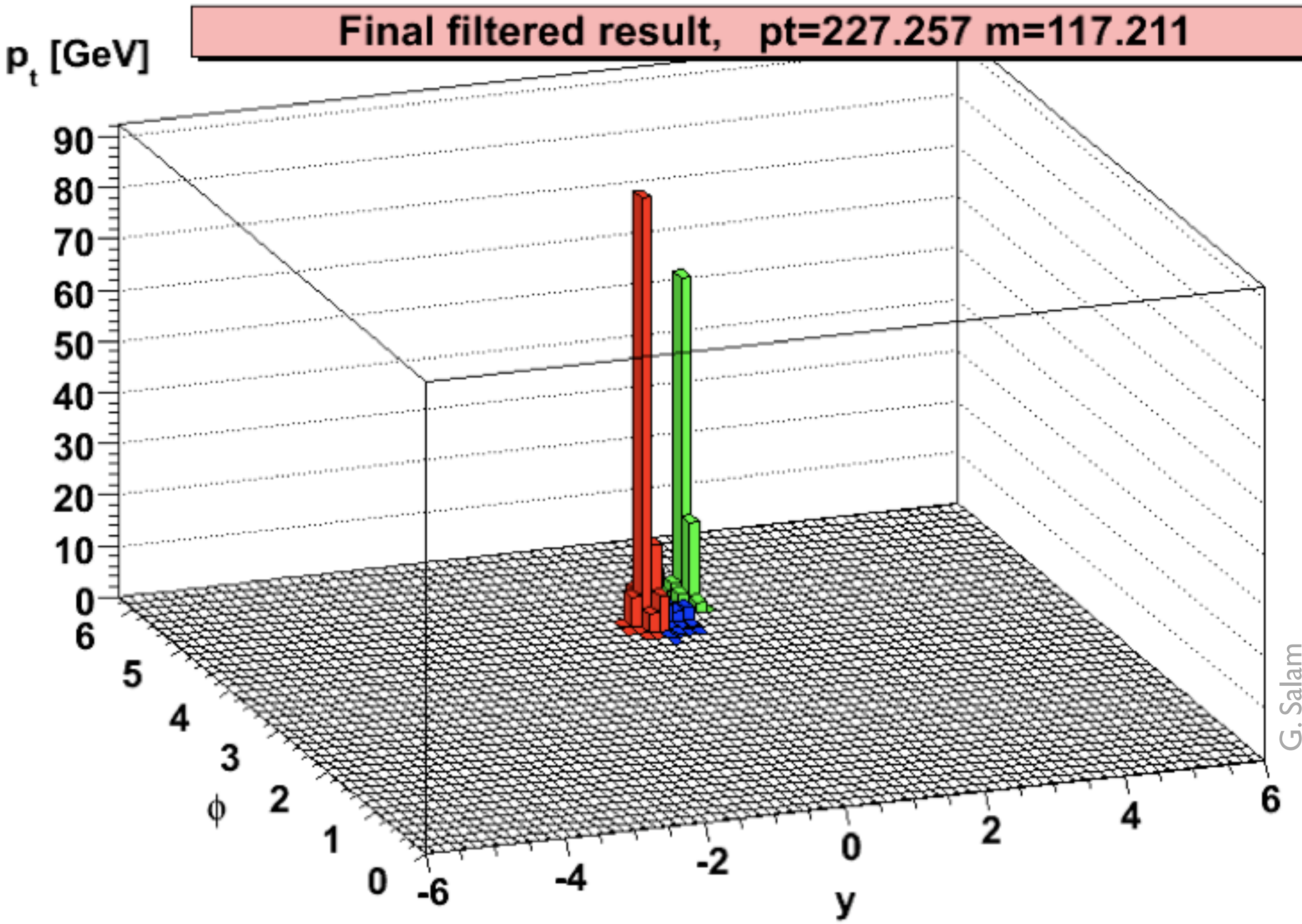
Start with a jet

Filtering in action



Recluster the contituents with R_{filt}

Filtering in action



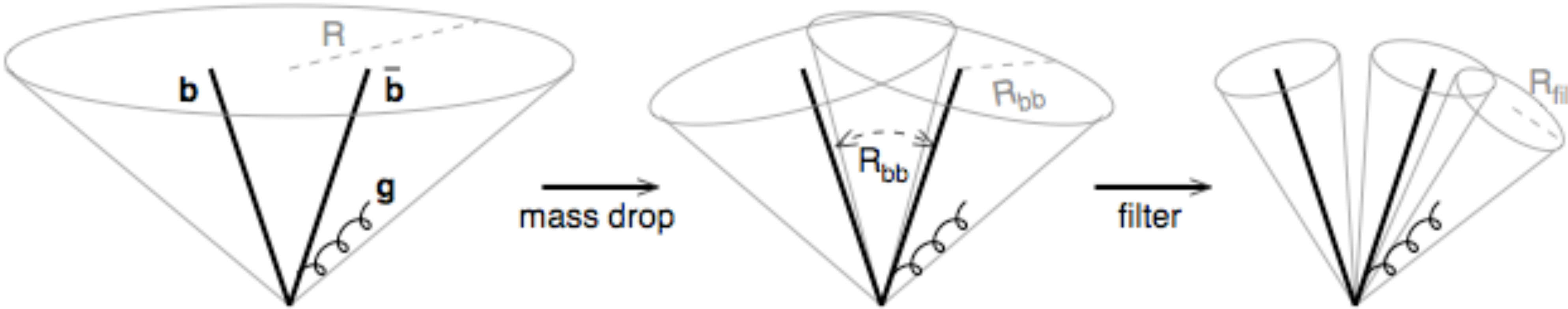
Only keep the n_{filt} hardest jets

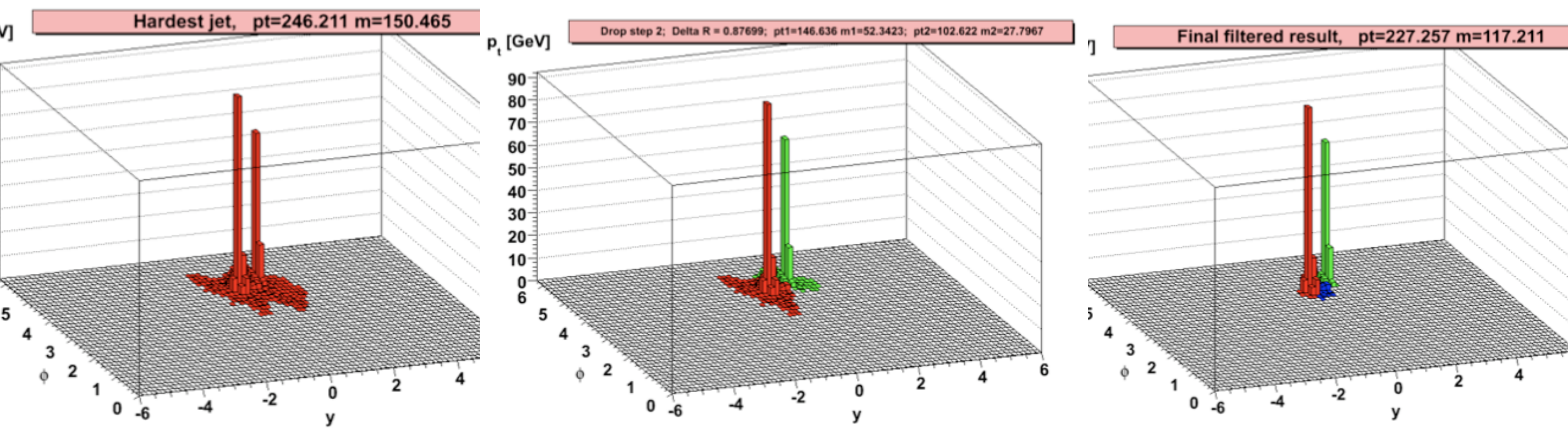
The low-momentum stuff surrounding the hard particles has been removed

Boosted Higgs analysis

$pp \to ZH \to \nu\bar{\nu}b\bar{b}$

Butterworth, Davison, Rubin, Salam, 2008





Cluster with a large R

Undo the clustering into subjets, until a large mass drop is observed

Re-cluster with smaller R, and keep only 3 hardest jets

BDRS in FastJet

```
#include “fastjet/tools/MassDropTagger.hh”
#include “fastjet/tools/Filter.hh”

JetDefinition jet_def(cambridge_algorithm, 1.2);
ClusterSequence cs(input_particles, jet_def);

// define the tagger and use it
MassDropTagger md_tagger(0.667, 0.09);
PseudoJet tagged = md_tagger(jets[0]);

// define the filter and use it
Filter filter(0.3,SelectorNHardest(3));
Pseudojet higgs = filter(tagged);     // this is the Higgs!!
```

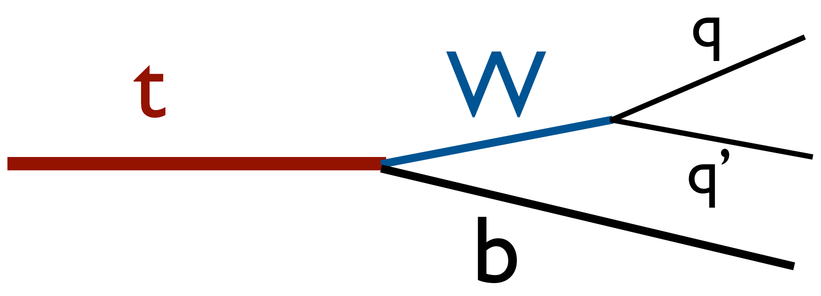
The real analysis is slightly more refined (b-tagging, dynamical filter radius, etc) but the main features are already present here

Pruning in FastJet

```
#include "fastjet/tools/Pruner.hh"

JetDefinition jet_def(cambridge_algorithm, 1.2);
ClusterSequence cs(input_particles, jet_def);

// define the pruner and use it
double zcut = 0.1;
double rcut_factor = 0.5;

Pruner pruner(cambridge_algorithm, zcut, rcut_factor);

PseudoJet tagged = pruner(jets[0]);
```

Top tagging

In order to tag a (boosted) top one must now identify 3-prongs structures originating from the top decay



Simplest approach: iterate declustering of k_t jets, beyond a first 'relevant splitting'

Early examples (2008):

- *ATLAS top tagger: put cuts on jet mass and d_{ij} scale*
- *Thaler-Wang: decluster to exactly 2 or 3 jets, put cuts on jet mass*

Many more top taggers after these

Example: top taggers

[1 jet $\gtrsim$ 2 partons]
└[An example]

Tagging boosted top-quarks

Many papers on top tagging in '08–'11: jet mass + something extra.

Questions

- What efficiency for tagging top?
- What rate of fake tags for normal jets?

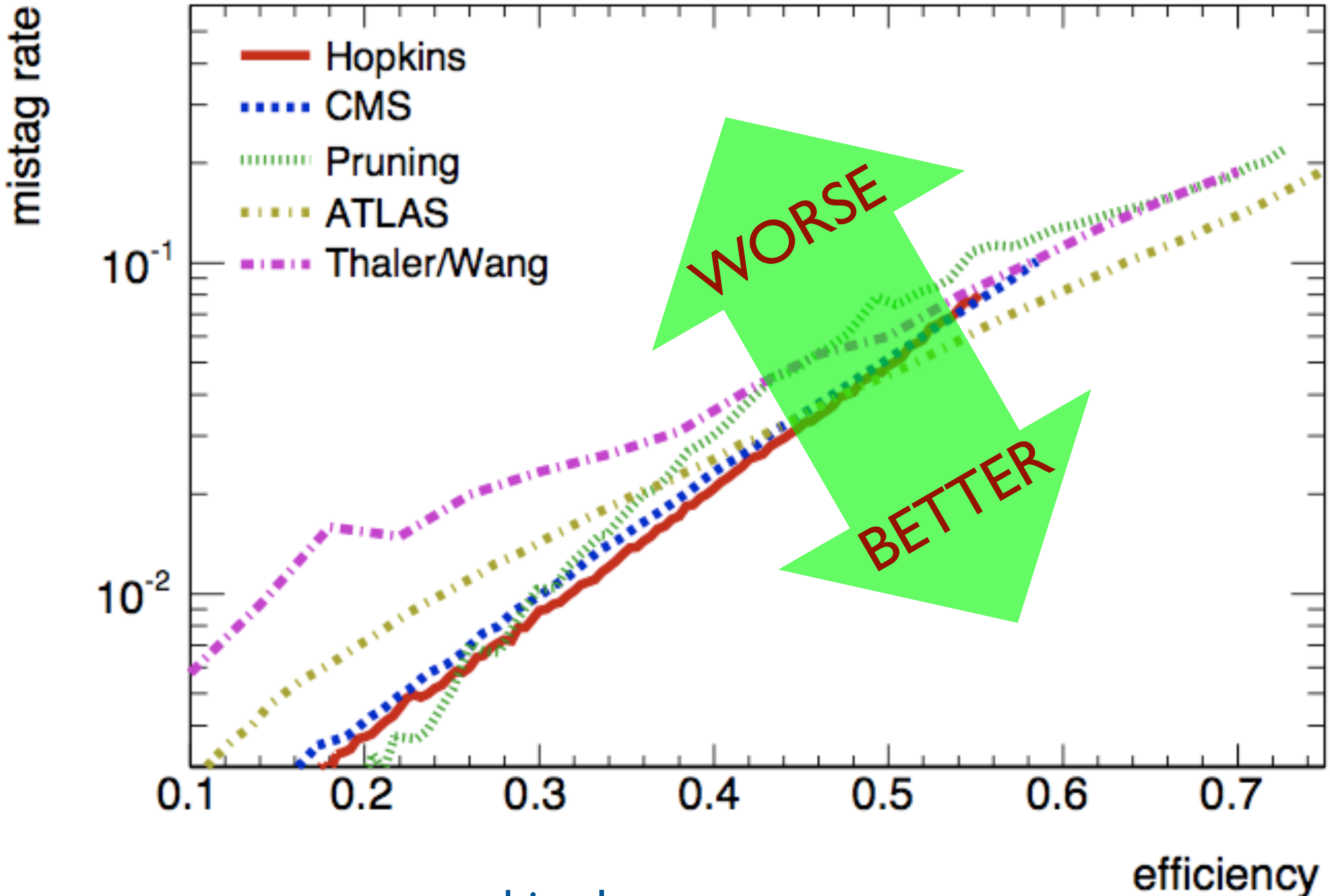
Rough results for top quark with $p_t \sim 1$ TeV

	"Extra"	eff.	fake
[from T&W]	just jet mass	50%	10%
Brooijmans '08	3,4 k_t subjets, d_{cut}	45%	5%
Thaler & Wang '08	2,3 k_t subjets, z_{cut} + various	40%	5%
Kaplan et al. '08	3,4 C/A subjets, $z_{cut} + \theta_h$	40%	1%
Ellis et al. '09	C/A pruning	10%	0.05%
ATLAS '09	3,4 k_t subjets, d_{cut} MC likelihood	90%	15%
Chekanov & P. '10	Jet shapes	60%	10%
Almeida et al. '08–'10	Template + shapes	13%	0.02%
Thaler & v Tilburg '10	Subjettiness	40%	2%
Plehn et al. '09–'10	C/A MD, θ_h/Dalitz [busy evs, $p_t \sim 300$]	35%	2%

Jets lecture 3 (Gavin Salam) CERN Academic Training March/April 2011 26 / 29

Comparison of top taggers

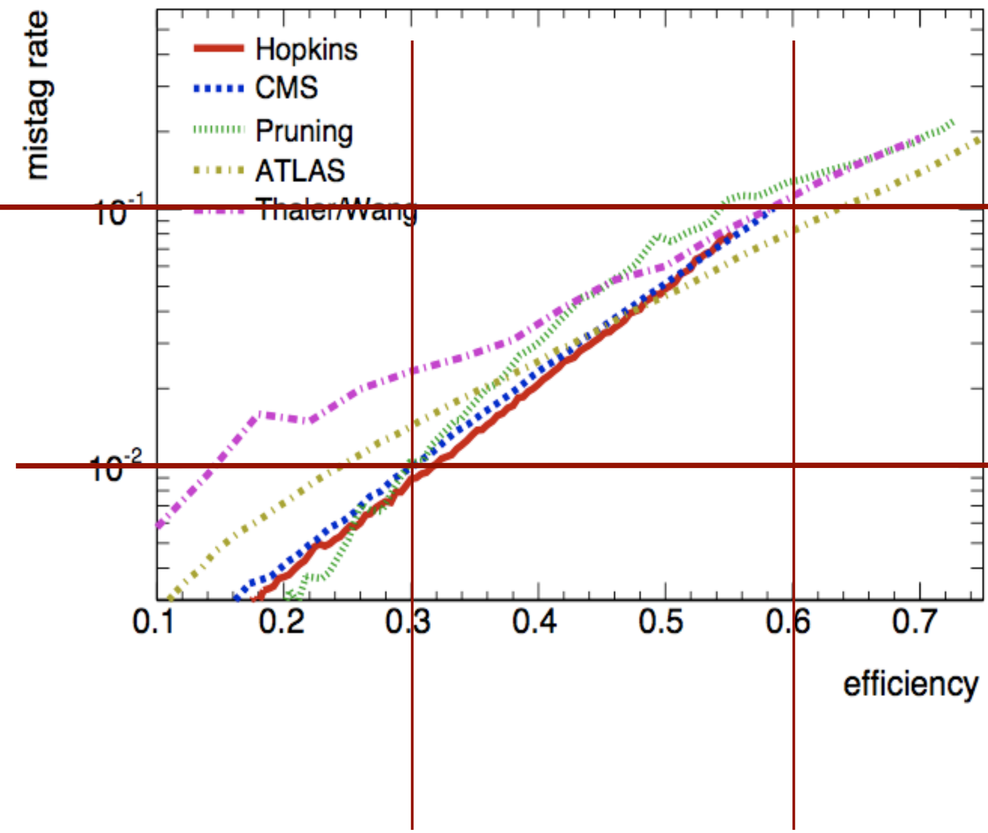
Boost 2010 proceedings, arXiv:1012.5412



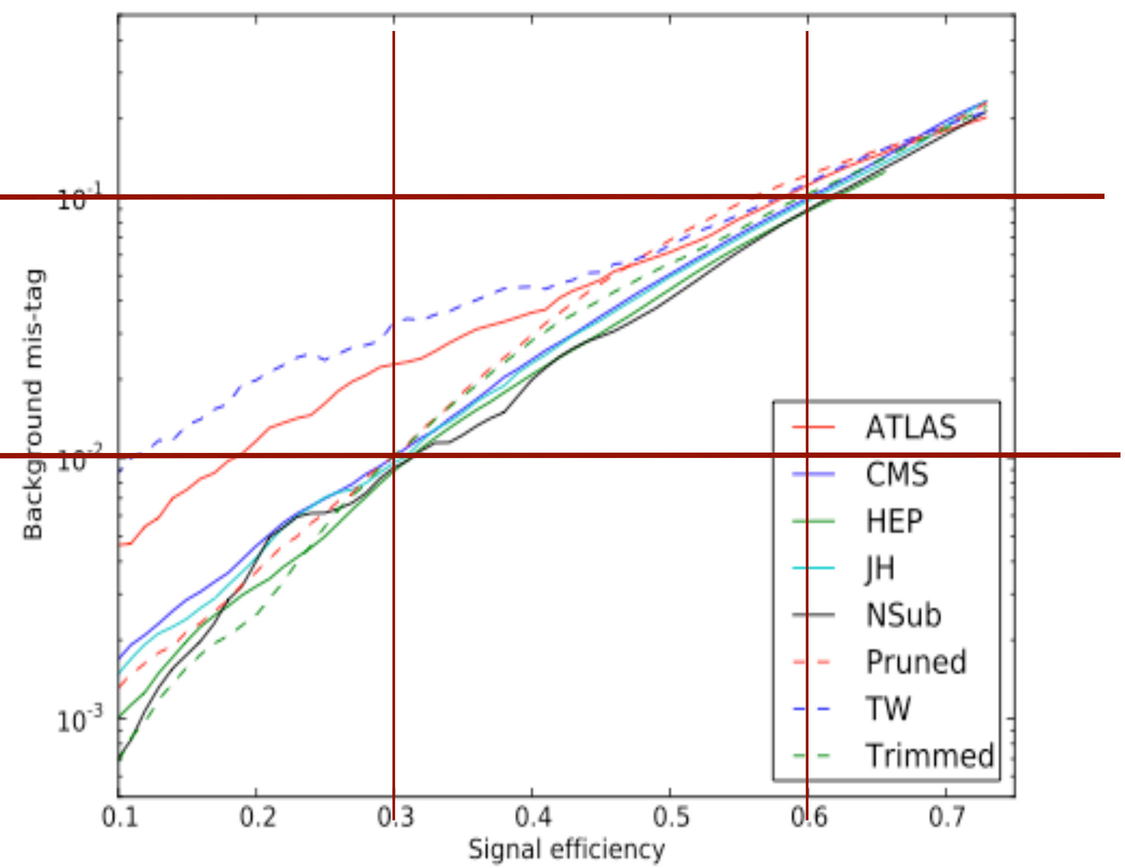
Even more curves now on this plot

Comparison of top taggers

Boost **2010** proceedings, arXiv:1012.5412



Boost **2011** proceedings, arXiv:1201.0008



Law of diminishing returns: improvement has become very hard

Boost 2012 is under way right now in Valencia.

Concluding remarks

- Proper (IRC-safe) definition of jet algorithms and efficient implementations have allowed for the study and exploitation of a number of jet properties
  - jet areas
    - *background determination and subtraction*
  - jet substructure
    - *taggers*
    - *groomers*

- Many new physics search strategies based on jet substructure are being explored and commissioned right now at the LHC
  - *As soon as more data (and more boosted particles) are available, we should see the first results from them*

- While these tools seem mature and are being refined, one can hope that new radical ideas will again start a 'revolution' and deeply change the field

Backup

Areas as a dynamical jet property

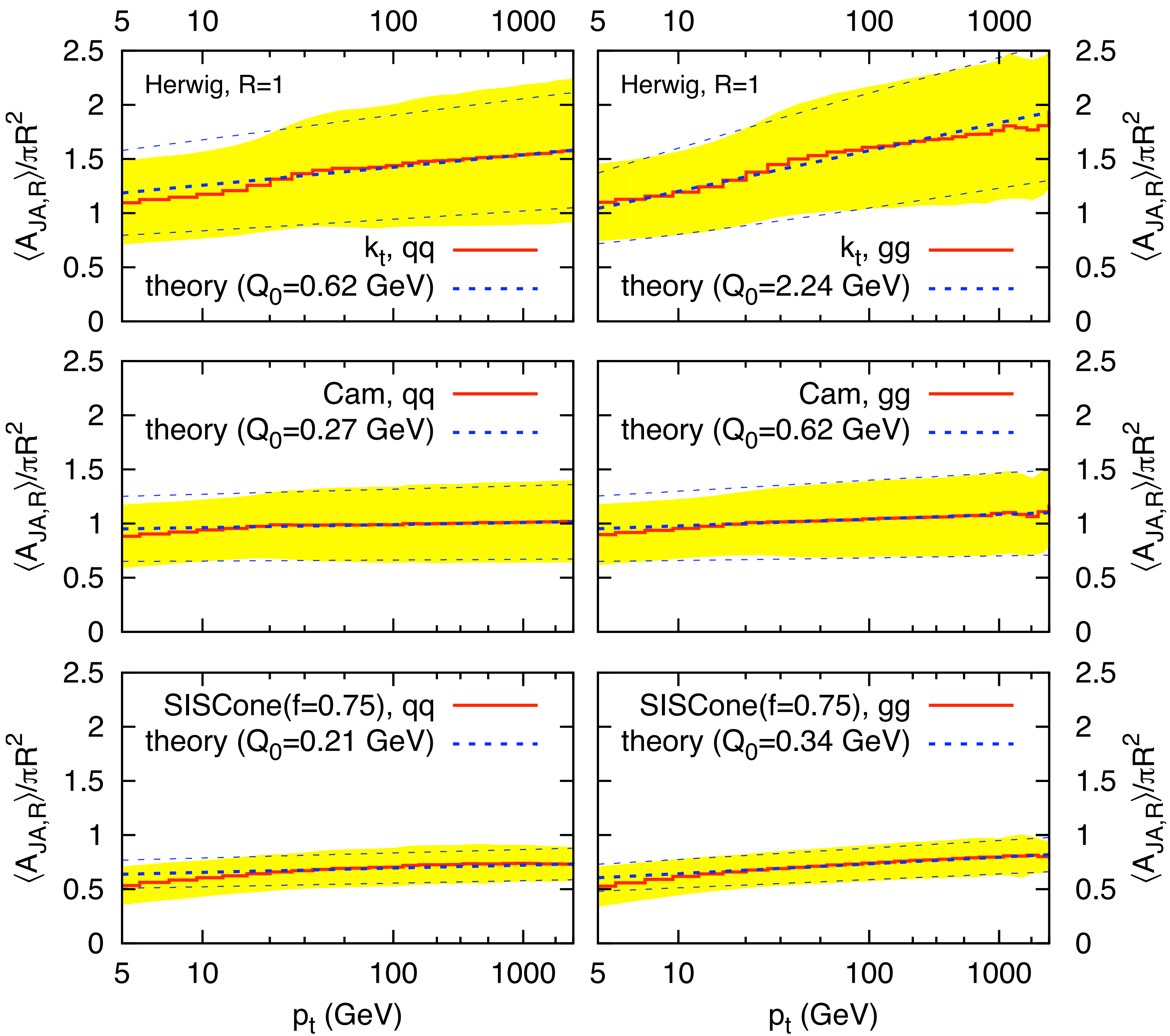
The average area of a jet can change with its p_t:

$$\langle \Delta A \rangle = D \, \frac{C_1}{\pi b_0} \ln \frac{\alpha_s(Q_0)}{\alpha_s(R p_{t1})}$$

	k_t	Cam/Aa	SISCone	anti-k_t
D	0.52	0.08	0.12	0

Again, only anti-k_t has a typical area that does **not** increase with p_t

Jet areas scaling violations



Averages and dispersions evolution from Monte Carlo simulations (dijet events at LHC) in good agreement with simple LL calculations

Area scaling violations are a legitimate observable.

(Though they might not be the best place where to measure α_s ....)

Jet areas scaling violations

MC, Salam, Soyez, arXiv:0802.1189

Check anti-k_t behaviour: scaling violations indeed absent, as predicted