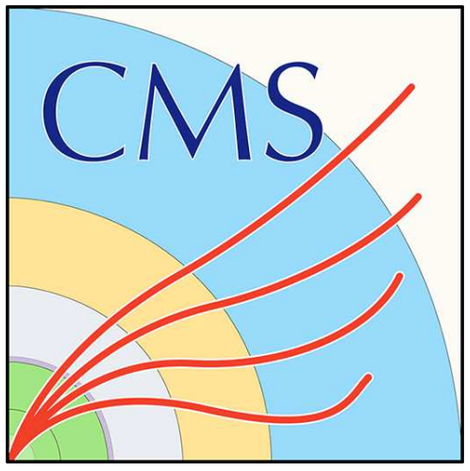
# The Future is Open

# Jet Substructure with CMS Public Data

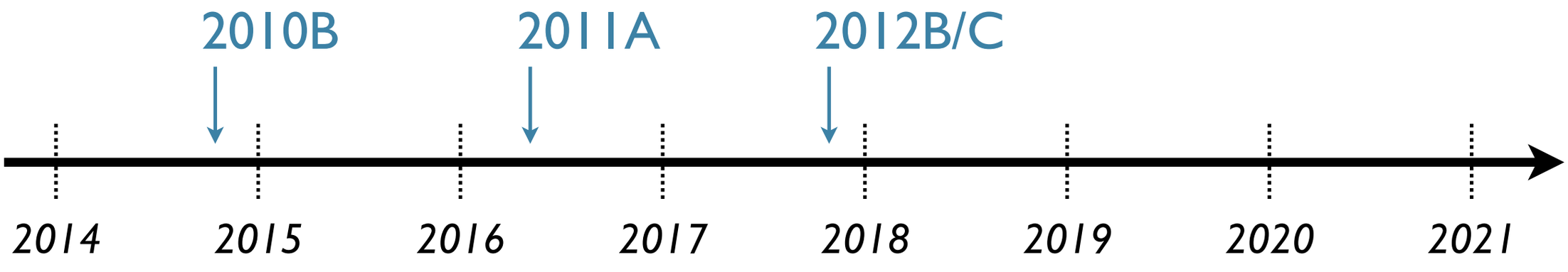
Jesse Thaler

CMS Week, CERN — June 25, 2018

# CMS and CERN are pioneering the release of research-grade public collider data

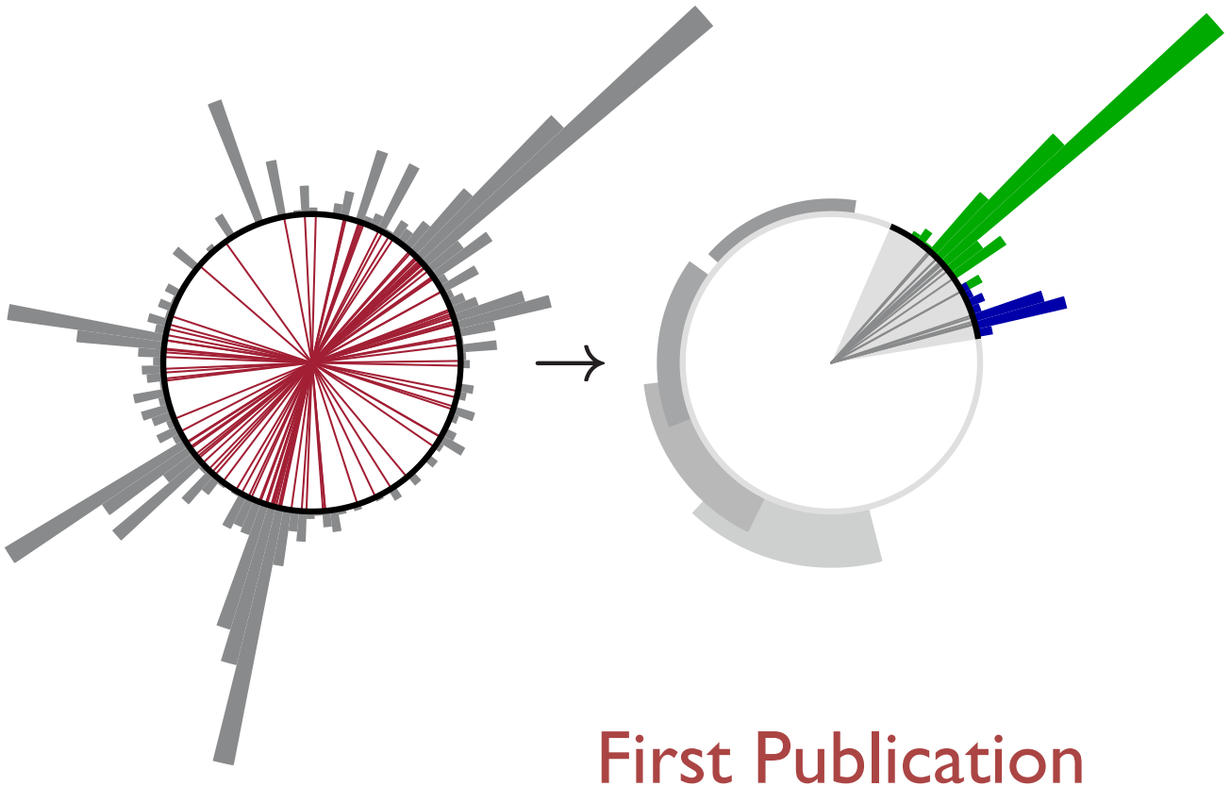


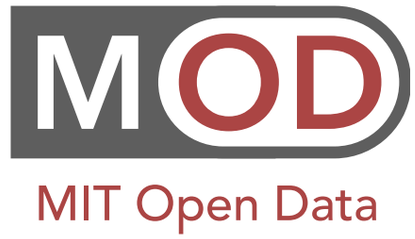


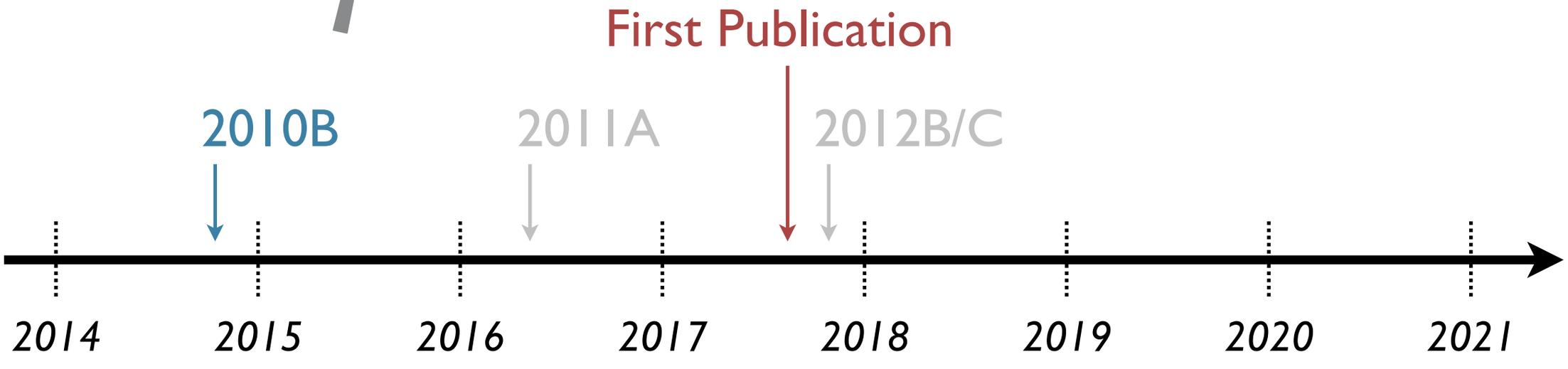


# Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,[1,*] Simone Marzani,[2,†] Jesse Thaler,[3,‡] Aashish Tripathee,[3,§] and Wei Xue[3,‖]





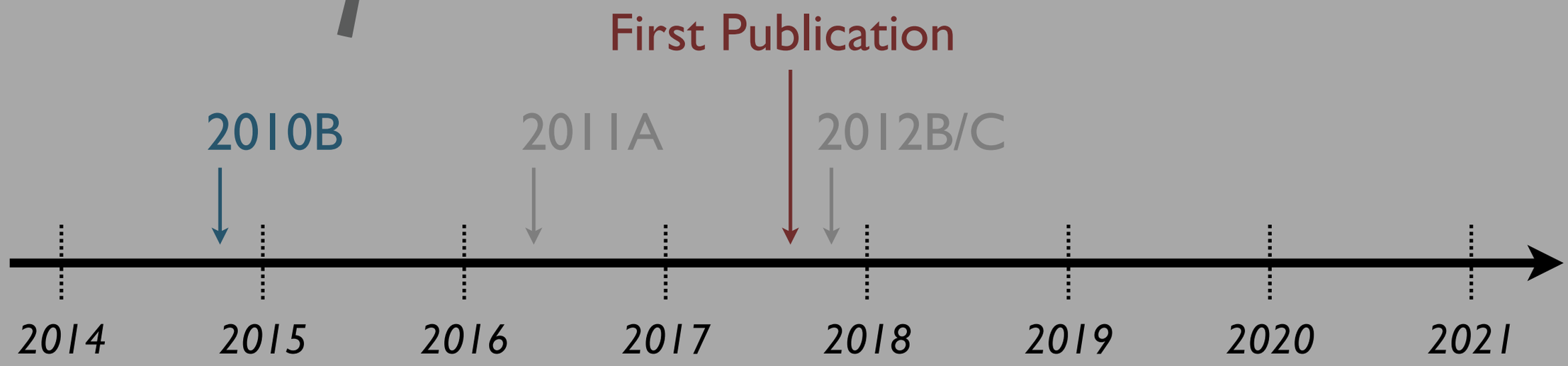
# Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,[1,*] Simone Marzani,[2,†] Jesse Thaler,[3,‡] Aashish Tripathee,[3,§] and Wei Xue[3,‖]

*A Milestone for Public Collider Data*

*A Milestone for Jet Physics*

*An Opportunity/Challenge for our Community*

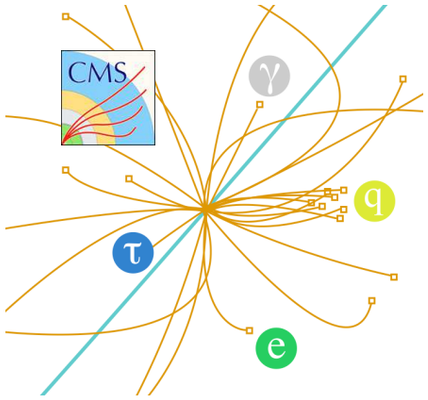



*Viability of public collider data depends on interest/enthusiasm of particle physics community*

Goals of this talk:

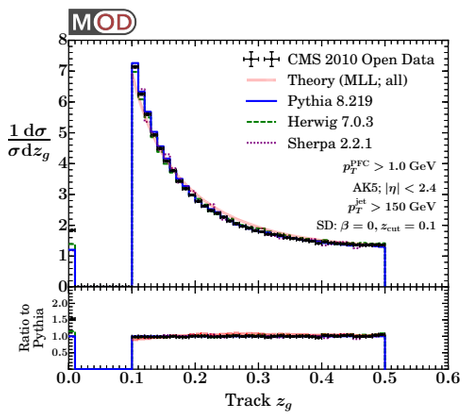
Highlight the opportunities
Expose the challenges
Inspire you to help

# Outline



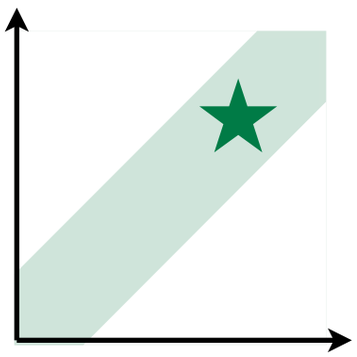

## Using the CMS Open Data

*Recent progress processing the 2011 dataset*



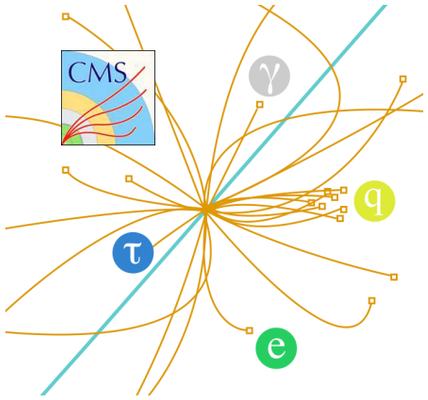

## Jet Substructure and QCD Splittings

*Highlights from our 2010 publications*



## The Future of Public Collider Data

*Back to the future with ALEPH data*



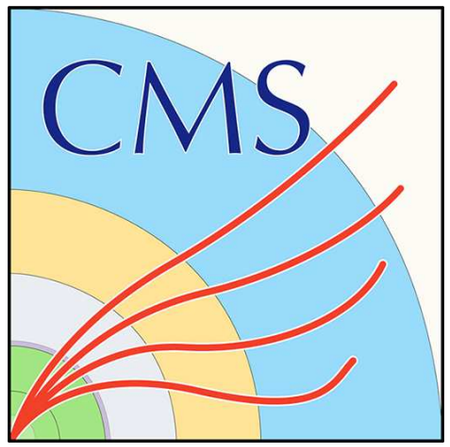

# Using the CMS Open Data




## Jet Substructure and QCD Splittings



## The Future of Public Collider Data




opendata
CERN

Kati Lassila-Perini,
Achim Geiser, ...

opendata.cern.ch/research/CMS

*November 2014:*

Run 2010B
7 TeV, 32 $pb^{-1}$

*>20 TB, no MC*

*(First publication: QCD)*

*April 2016:*

Run 2011A
7 TeV, 2.5 $fb^{-1}$

*>100 TB, with MC*

*(In the pipeline: BSM, ML)*

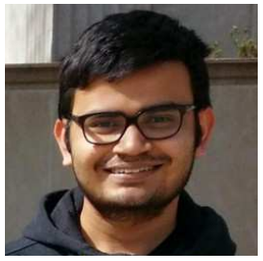
*December 2017:*

Run 2012B/C
8 TeV, 11.6 $fb^{-1}$

*>1 PB, with MC*

# The MIT Open Data Team

## 2010 QCD:



**Aashish Tripathee**

Wei Xue

Andrew Larkoski

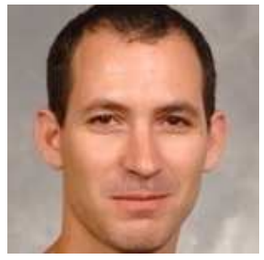
Simone Marzani

Summer intern: Alexis Romero

CMS advice: Sal Rappoccio

## 2011 BSM:

Matt Strassler



Yotam Soreq

Wei Xue

Cari Cesarotti

Raffaele D'Agnolo

## 2011 ML:

**Radha Mastandrea**

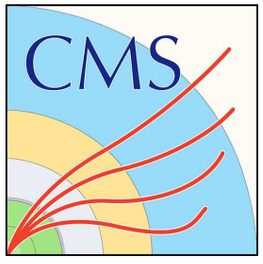
**Preksha Naik**

Patrick Komiske

…

*Today:*
*very preliminary 2011 results*
*(only 16%, still debugging)*

# Key Challenge: Initial Data Processing




*CernVM + CMSSW 5.3.32 (for 2011)*

AOD Format (CMS Root)

RAW → RECO → "Analysis Object Data"

*Access via XRootD, write custom EDAnalyzer*

For novices, very steep learning curve for using Root and understanding overall data structure (esp. triggers)

Jet Primary Dataset: 4.7 TB for 3 million AOD events

Daunting amount of information!

# Our Strategy: Simplified Analysis Framework

*MODProducer + MODAnalyzer + FastJet 3.3.1*

MOD

MOD Format (ASCII)

For 2010: Cross-check with flat Root n-tuples

*Access via External Hard Drive*

Text files as educational tool and debugging strategy
Access only essential information, sacrifice flexibility

Jet Primary Dataset: ~500 GB for 3 million MOD events

New for 2011: Separate processing for triggers/luminosity

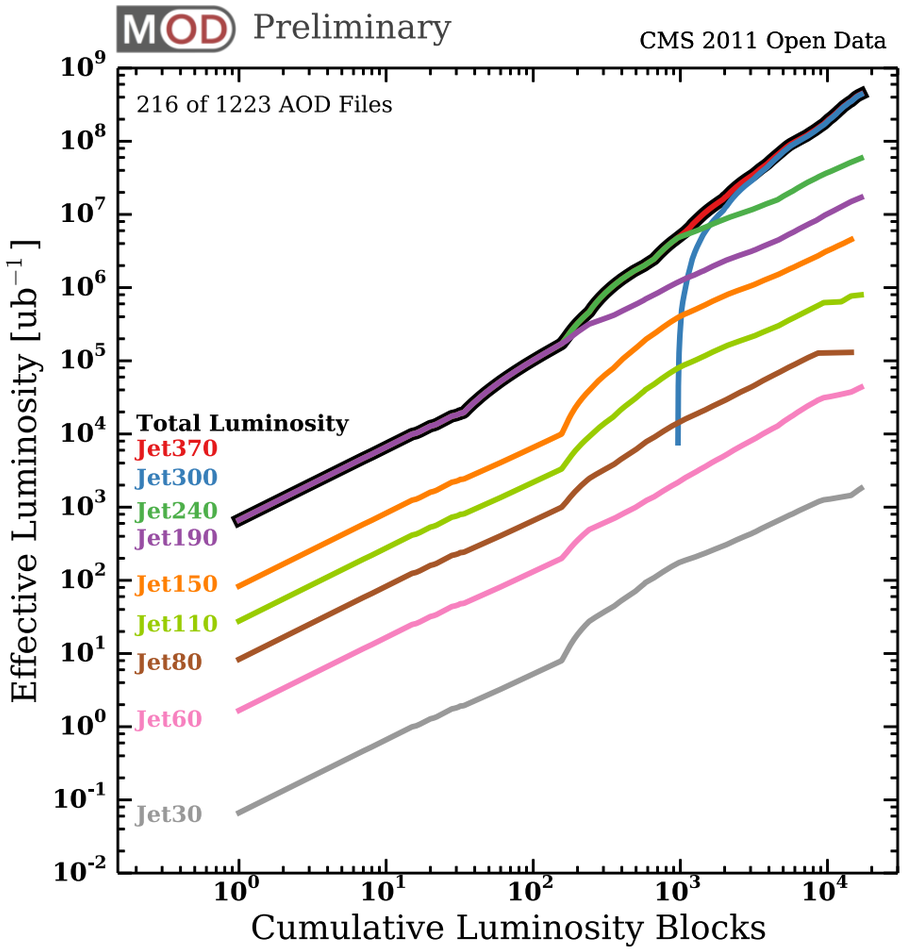
# Example MOD Metadata (Simplified)

```
BeginFile Version 6 CMS_2011A Data Jet

# File                Filename  TotalEvents  ValidEvents  IntLumiDel  IntLumiRec
  File    A850-02163E008D77         30284        29057    895036.1    884409.2

#Block    RunNum    LumiBlock       Events       Valid?  IntLumiDel  IntLumiRec
 Block    160578          366           53           1        28.6        27.5
 Block    160578          367           58           1        28.6        27.5
 Block    160578          368           38           1        28.6        27.5
 Block    160578          369           40           1        28.6        27.5
 Block    160578          370           57           1        28.6        27.5
 Block    160578          371           46           1        28.6        27.5
 Block    160578          372           37           1        28.6        27.5
   ...

# Trig                    Name    Present   Valid   Fired  EffLumiDel  EffLumiRec  AvePrescale
  Trig          HLT_Jet240_v2      16530       9       1     27228.1     27026.7          3.0
  Trig  HLT_DiJetAve140U_v4      13754       5       1     11705.3     11582.4          1.0
  Trig          HLT_Jet240_v1      13754       5       1     18387.4     18205.5          1.0
  Trig          HLT_Jet150_v2      16530       9       2      2722.8      2702.6         30.0
  Trig          HLT_Jet190_v2      16530       9       2      8168.4      8108.0         10.0
  Trig          HLT_Jet190_v1      13754       5       2      6939.9      6872.4          3.0
  Trig   HLT_DiJetAve15U_v4      13754       5       1         1.0         1.0       7500.0
  Trig   HLT_DiJetAve110_v2      11772       3       1       358.3       356.0         75.0
   ...

EndFile
```

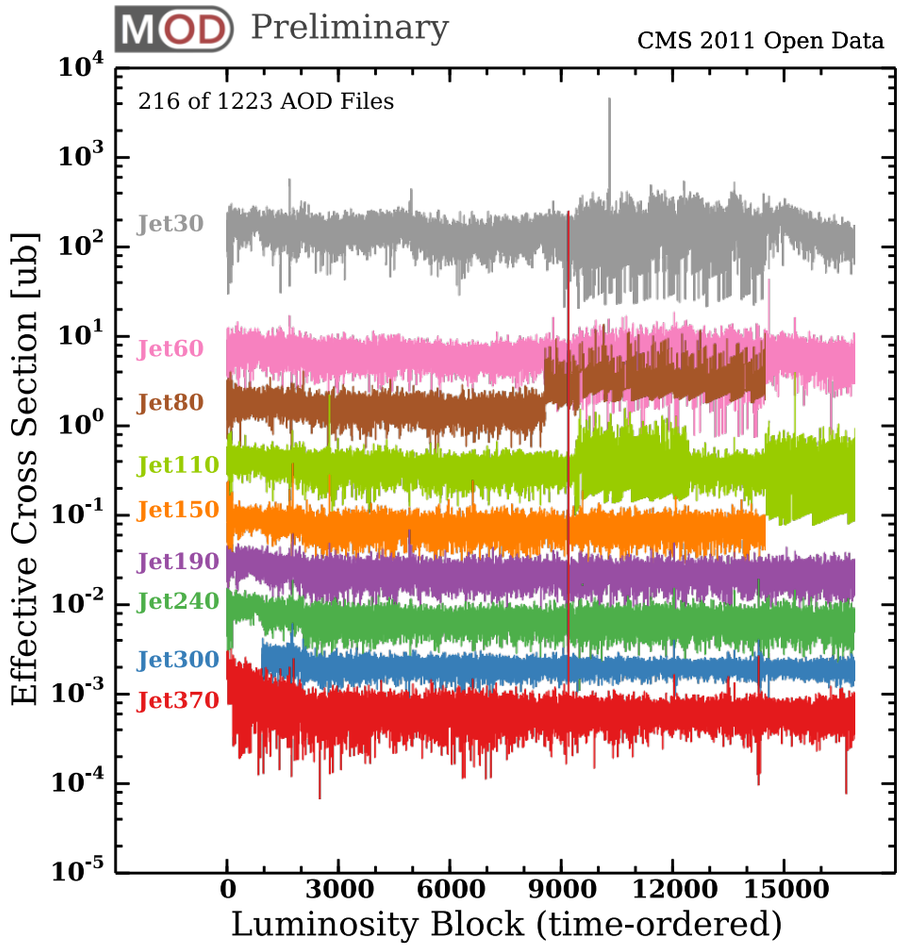
*New for 2011: Effective luminosity information per trigger*

# Example MOD Metadata (Simplified)

## Effective Luminosity per Trigger




## Testing Trigger Consistency




*New for 2011: Effective luminosity information per trigger*

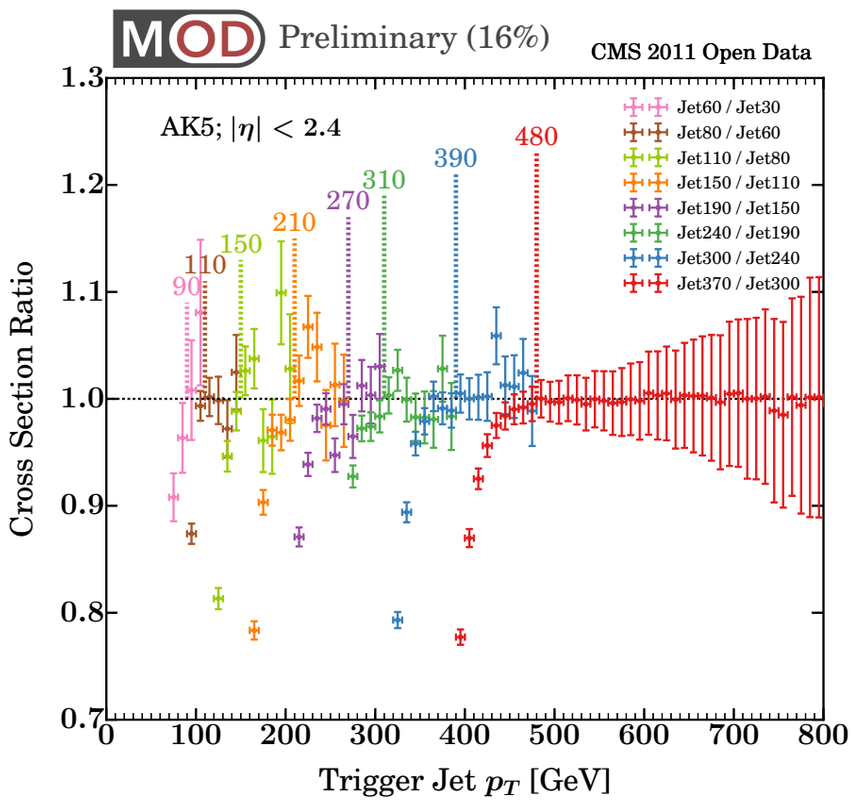
# Example MOD Event (Simplified)

```
BeginEvent Version 6 CMS_2011A Data Jet

# /Run2011A/Jet/MOD/12Oct2013-v1/20000/000D4260-D23E-E311-A850-02163E008D77.mod

#Cond      RunNum     EventNum    LumiBlock     NPV     Timestamp   msOffset
 Cond      160578     38142433          366       4    1300254008      84656

#Trig                        Name      Prescale_1   Prescale_2      Fired?
 Trig          HLT_DiJetAve30U_v4               1           15           0
 Trig          HLT_DiJetAve50U_v4               1            3           1
 Trig               HLT_Jet110_v1               1            1           1
 ...

# AK5        px        py        pz    energy     jec    area   no_of_const   neu_had_frac     ...
  AK5    -48.53     91.23    922.46    928.25    1.15    0.77             3           0.17     ...
  AK5     27.14    -27.95   -176.24    180.60    1.11    0.71            14           0.11     ...
  AK5      6.87    -27.39   -127.71    130.89    1.13    0.59            10           0.14     ...
  ...

# PFC        px        py        pz    energy  pdgId
  PFC      3.05     -2.27    -18.08     18.48    211
  PFC      3.51     -3.48    -21.66     22.22    211
  PFC      2.83     -3.01    -20.00     20.42   -211
  PFC      2.89     -2.37    -18.40     18.77    211
  PFC      1.21     -1.31     -7.58      7.79   -211
  PFC      1.62     -2.72    -12.17     12.58   -211
  PFC      7.15     -7.56    -46.86     48.01     22
 ...

EndEvent
```

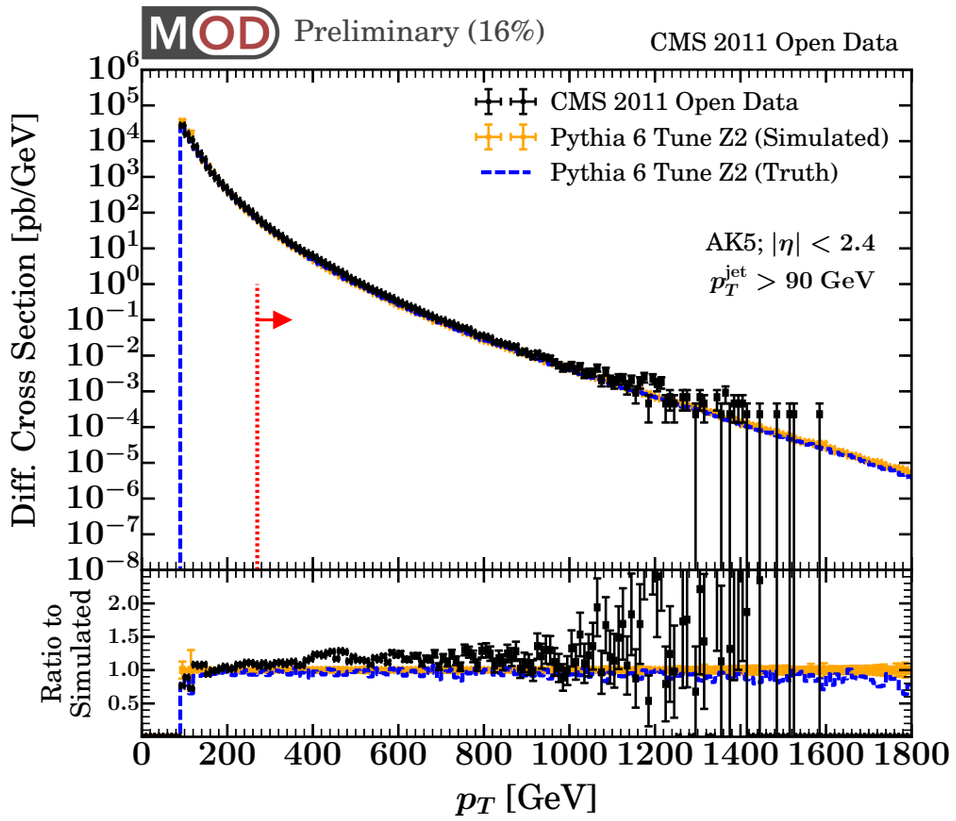
*Crucial: JEC factors, jet quality criteria, and particle flow candidates*

# Example MOD Event (Simplified)

## Trigger Turn-on Behavior




## Hardest Jet pT

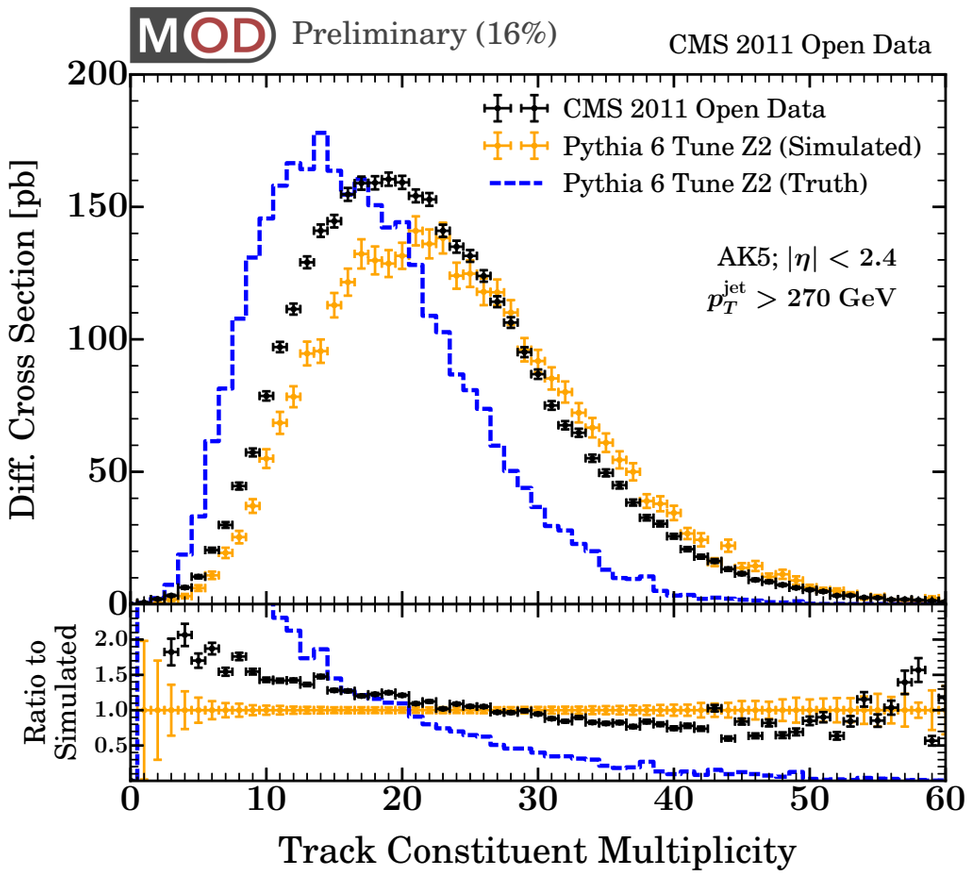



*New for 2011:*
Detector Simulated Data (!)

*Crucial: JEC factors, jet quality criteria, and particle flow candidates*

# Basic Jet Properties

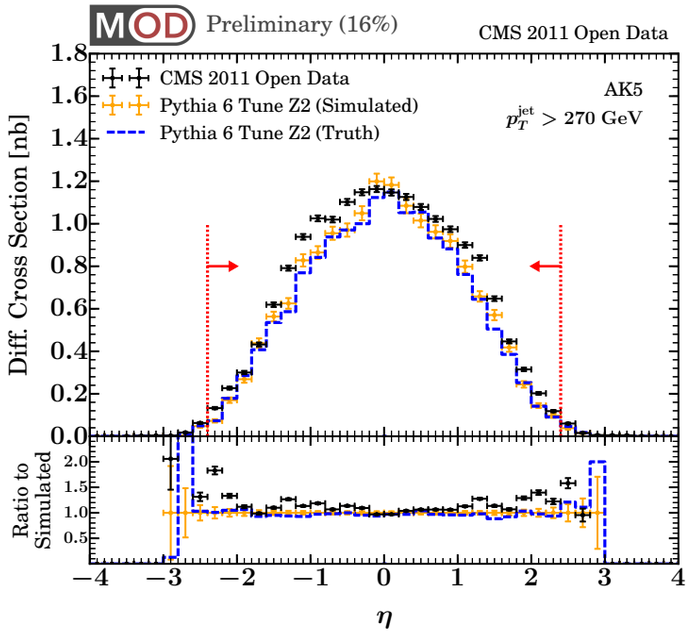
## Track Multiplicity



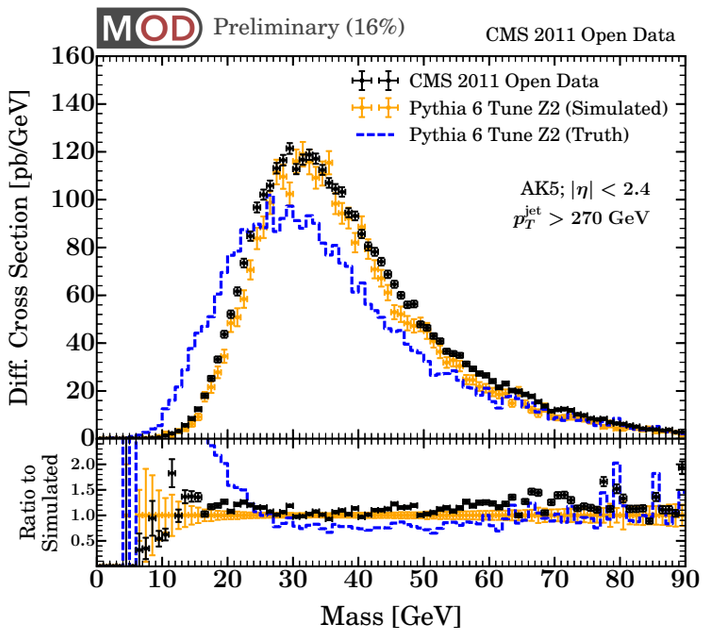

Expected mismatch with truth from strange hadrons ($c\tau_0 \in [10,1000]$ mm)

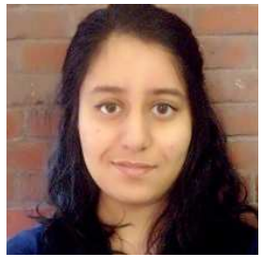
Mismatch with simulated under study

## Pseudorapidity




## Jet Mass (without JMC)





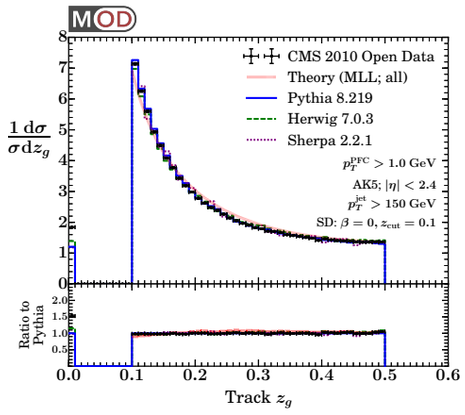
## Preliminary 2011 Processing:

## 2 students, ~6 months

(First collider physics project,
previous experience in Python/C++,
using template from 2010 analysis)

*Opportunity to streamline, since steps common to almost every collider data project*

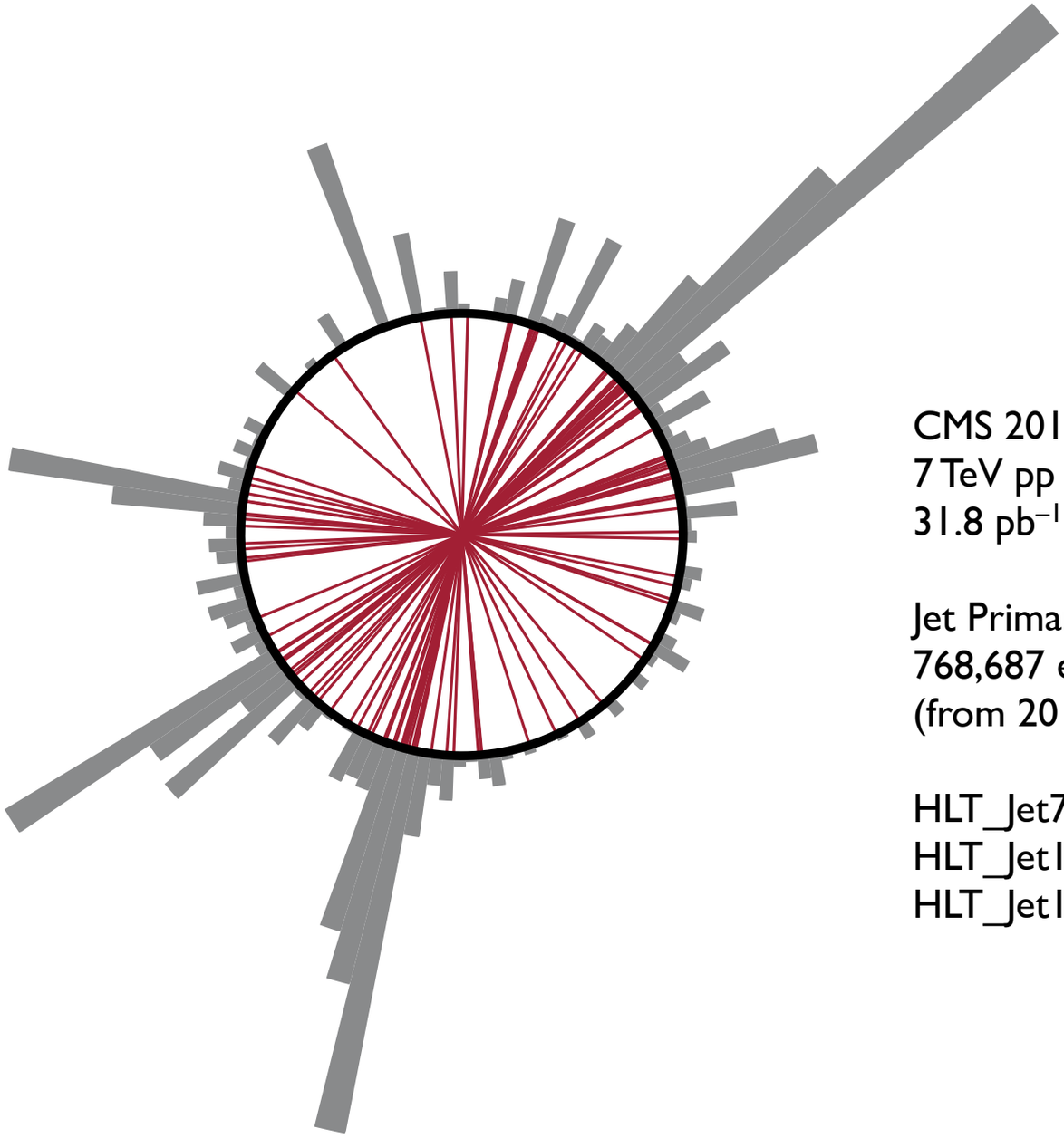
Using the CMS Open Data




# Jet Substructure and QCD Splittings



The Future of Public Collider Data

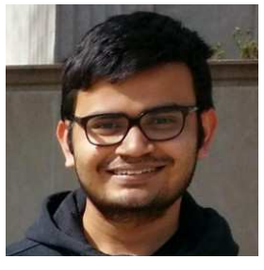


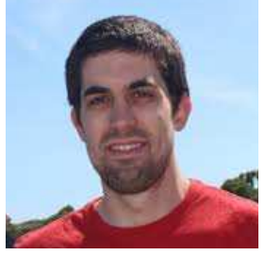
CMS 2010 Run B
7 TeV pp
31.8 $pb^{-1}$

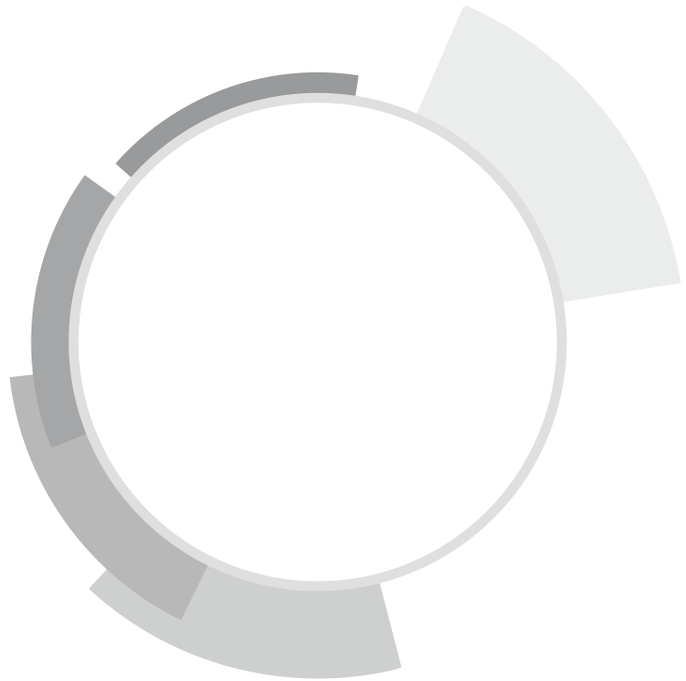
Jet Primary Dataset
768,687 events
(from 20 million)

HLT_Jet70U
HLT_Jet100U
HLT_Jet140U

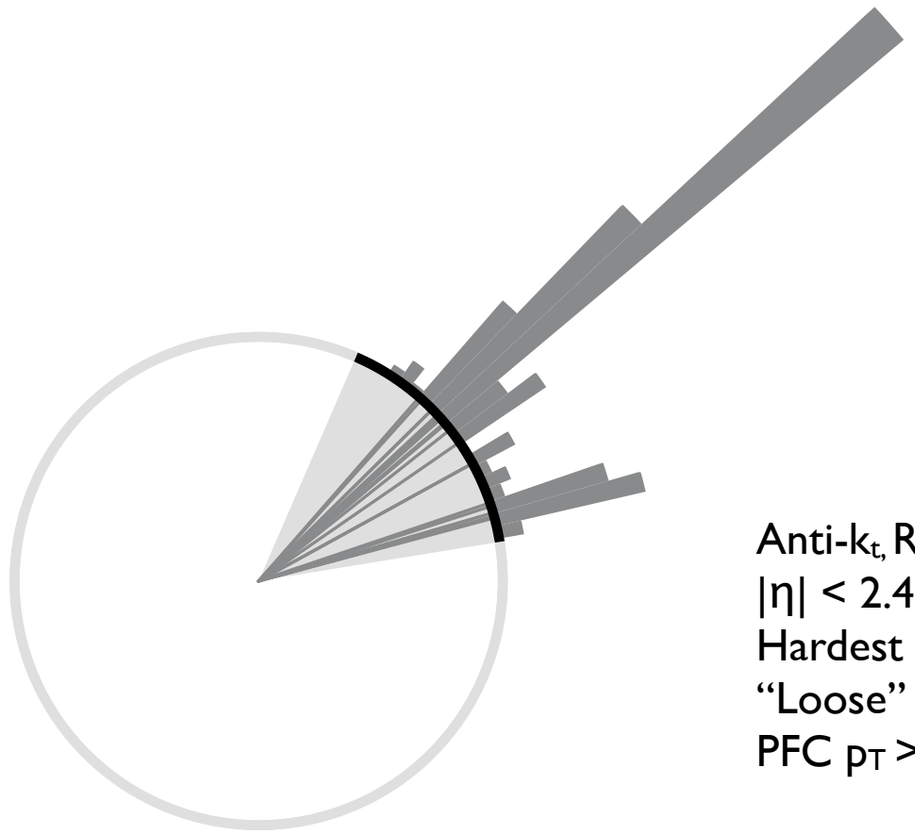
MOD





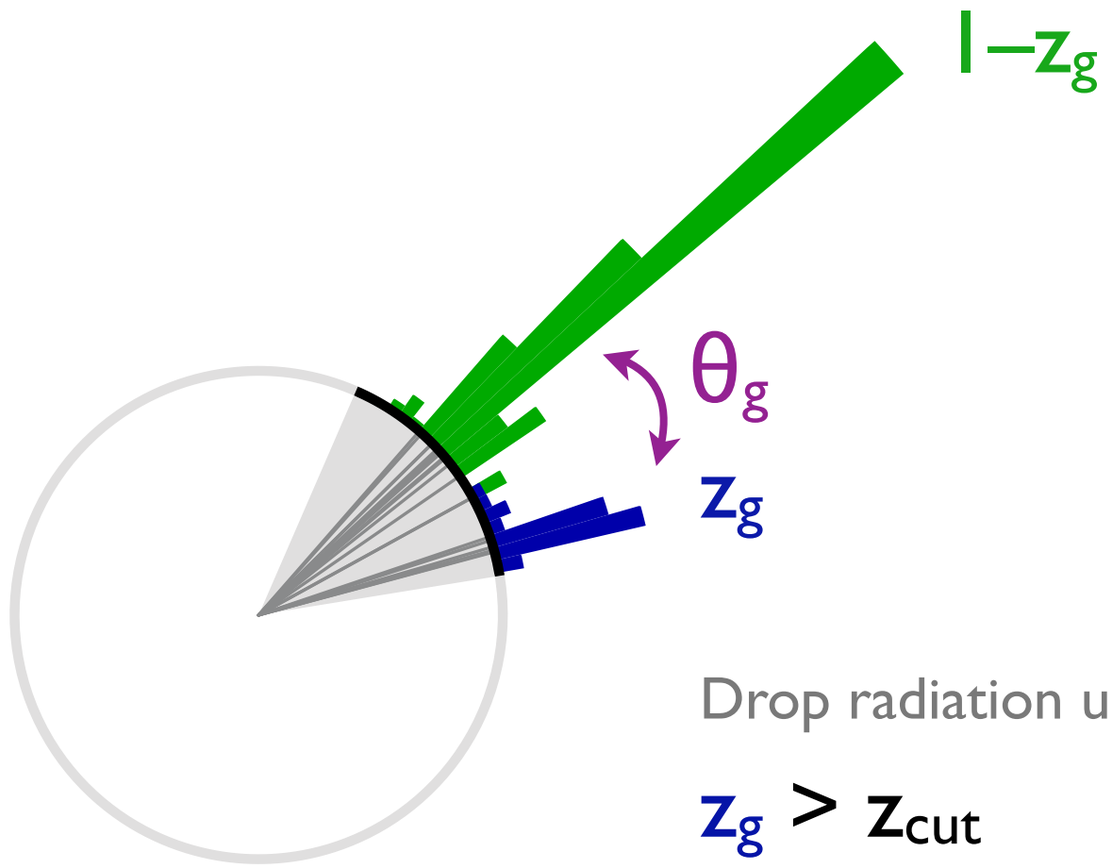



Anti-$k_t$, R = 0.5
$|\eta| < 2.4$
$p_T > 20$ GeV
"Loose" JQC



Anti-$k_t$, R = 0.5
$|\eta| < 2.4$
Hardest $p_T$ > 150 GeV
“Loose” JQC
PFC $p_T$ > 1 GeV

# Jet Grooming: mMDT/Soft Drop $\beta = 0$

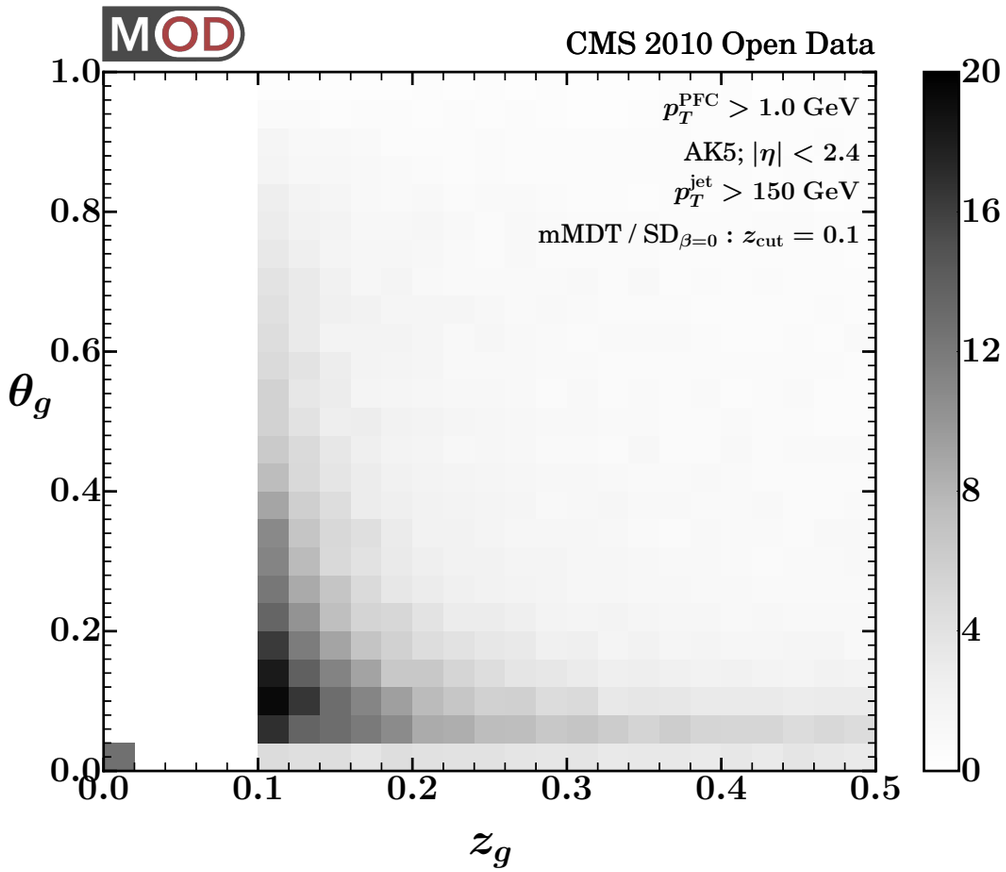


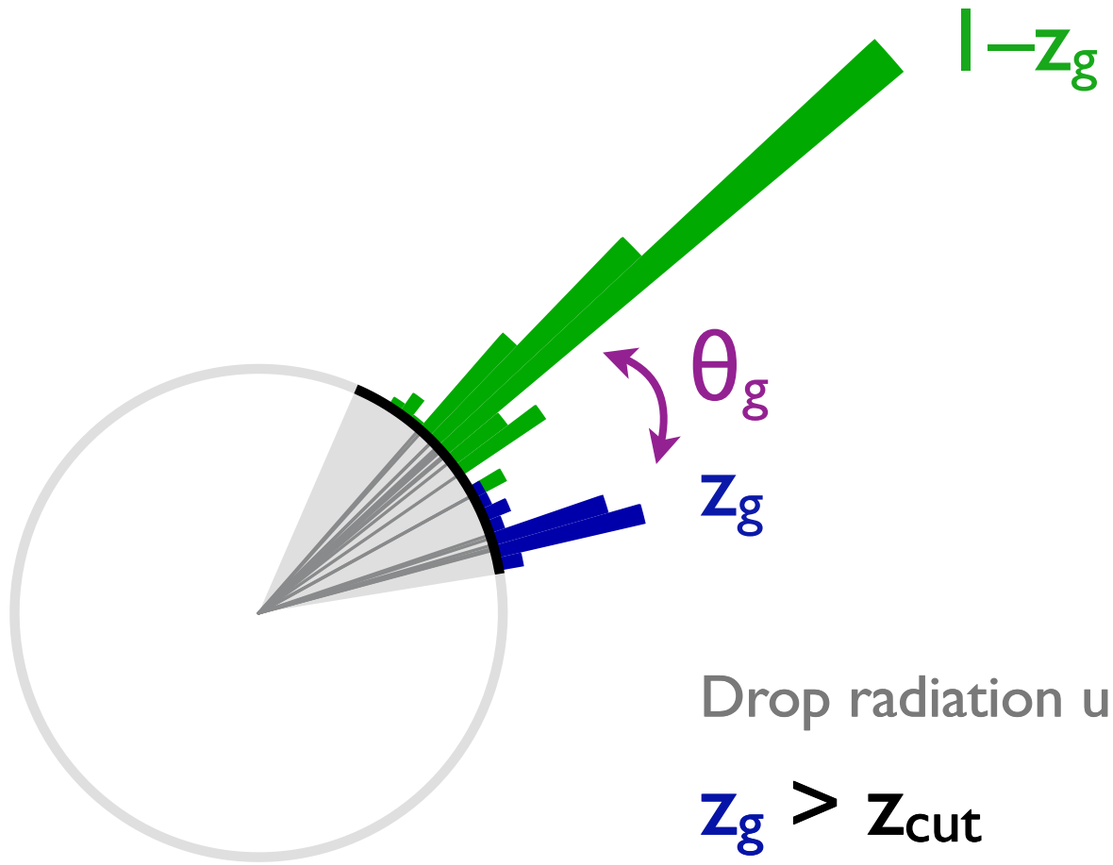

Drop radiation until:

$z_g > z_{cut}$

[Larkoski, Marzani, Soyez, JDT, 1402.2657; Dasgupta, Fregoso, Marzani, Salam, 1307.0007;
see also Butterworth, Davison, Rubin, Salam, 0802.2470]

# Soft/Collinear Behavior



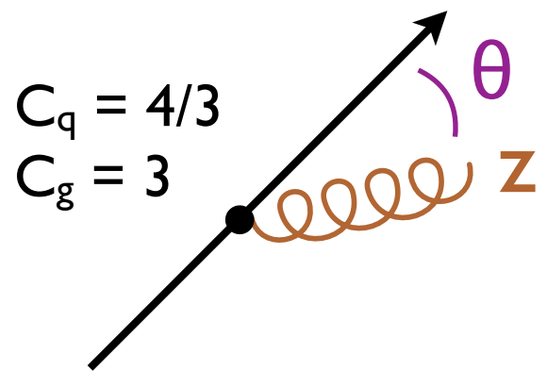




Drop radiation until:

$z_g > z_{\mathrm{cut}}$

$$\mathrm{d}P_{i\to ig} \simeq \frac{2\alpha_s}{\pi} C_i \frac{\mathrm{d}\theta}{\theta} \frac{\mathrm{d}z}{z}$$

Collinear Soft




$$z \approx z_g \quad (!)$$

[Larkoski, Marzani, JDT, 1502.01719; see also Larkoski, JDT, 1307.1699]

# Soft/Collinear Behavior

*Perfect application of CMS Open Data*

2010 data ⇒ 2014 release ⇒ 2015 idea ⇒ 2017 analysis

Benefits from low trigger thresholds and low pileup

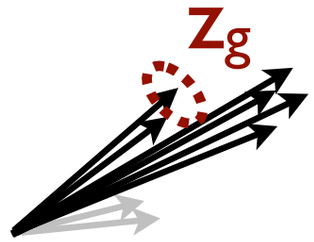
$z_g$

$\mathrm{d}P_i$

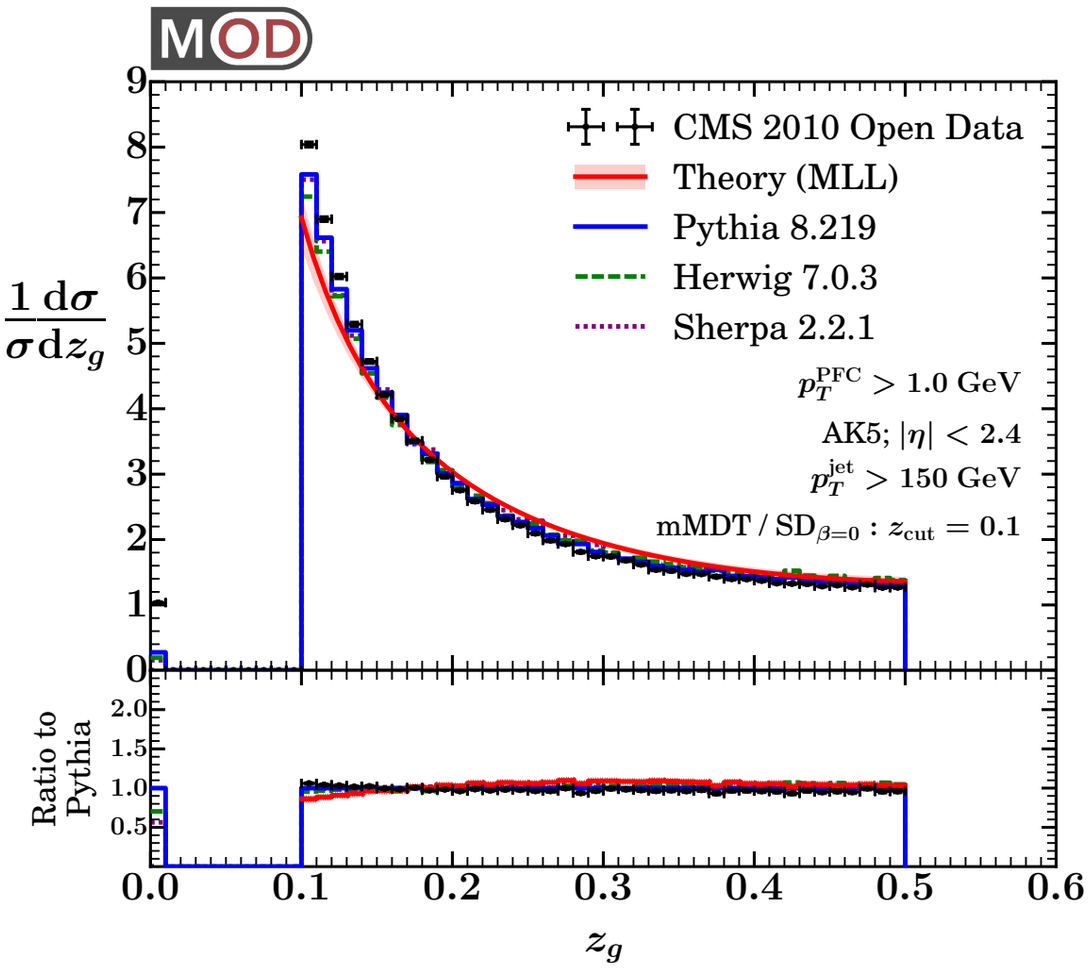

$z_g$ (!)

[Larkoski, Marzani, JDT, 1502.01719; see also Larkoski, JDT, 1307.1699]

# Exposing the QCD Splitting Function

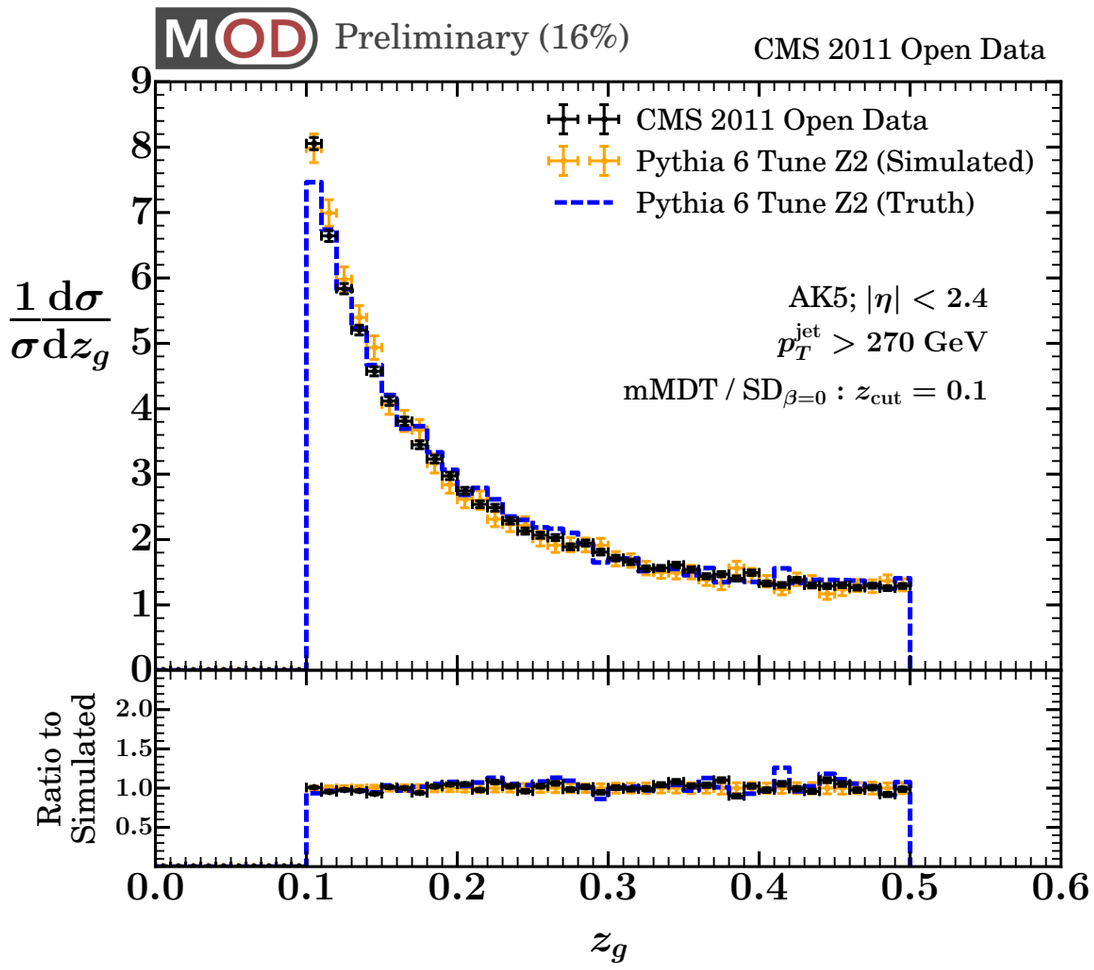



## 2010 Analysis




[Larkoski, Marzani, JDT, Tripathee, Xue, 1704.05066]

## Initial 2011 Results

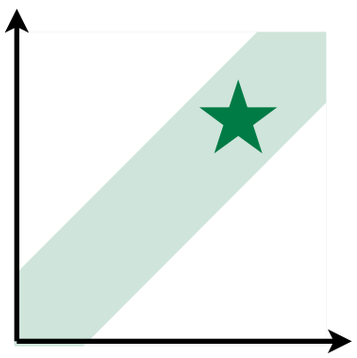



Clear impact from detector
(N.B.: no PFC $p_T$ cut imposed)

# Using the CMS Open Data




## Jet Substructure and QCD Splittings



# The Future of Public Collider Data

# Jet substructure studies with CMS open data

Aashish Tripathee,[1,*] Wei Xue,[1,†] Andrew Larkoski,[2,‡] Simone Marzani,[3,§] and Jesse Thaler[1,‖]

## V. ADVICE TO THE COMMUNITY

### A. Challenges

### B. Recommendations

## VI. CONCLUSION

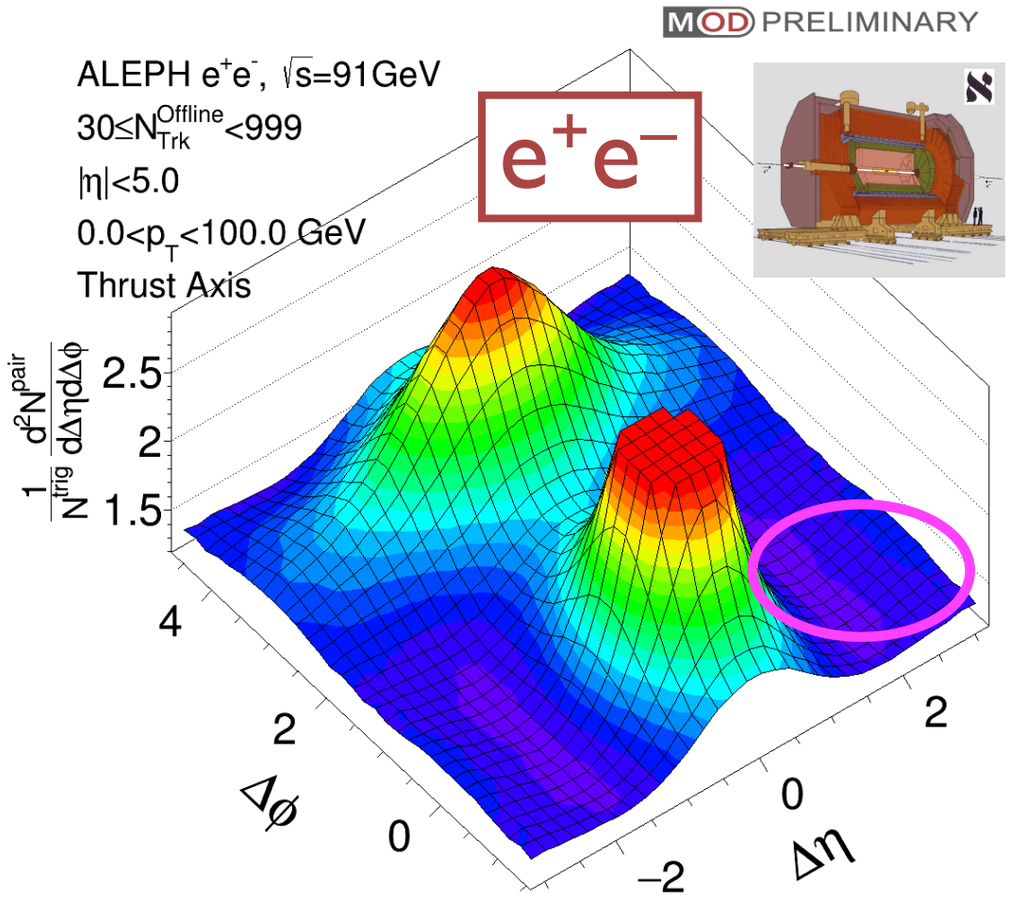
As the LHC explores the frontiers of scientific knowledge, its primary legacy will be the measurements and discoveries made by the LHC detector collaborations. But there is another potential legacy from the LHC that could be just as important: granting future generations of physicists access to unique high-quality data sets from proton-proton collisions at 7, 8, 13, and 14 TeV.

# E.g. ALEPH Confronts the CMS Ridge

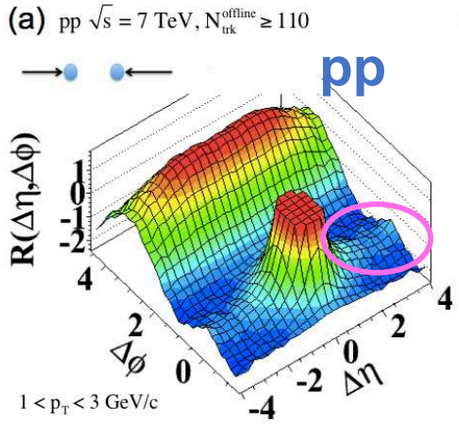
1990–95 $e^+e^-$ data

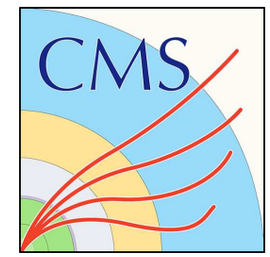
⇓ ⟵ 2010 pp surprise!

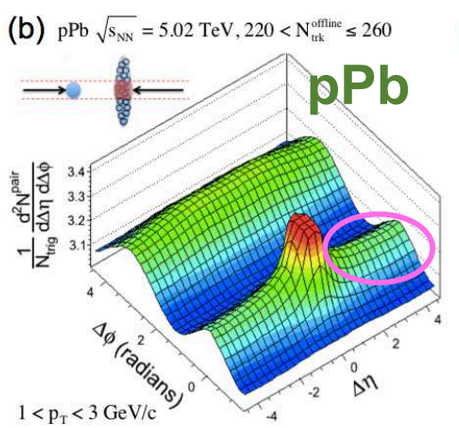
2018 $e^+e^-$ analysis

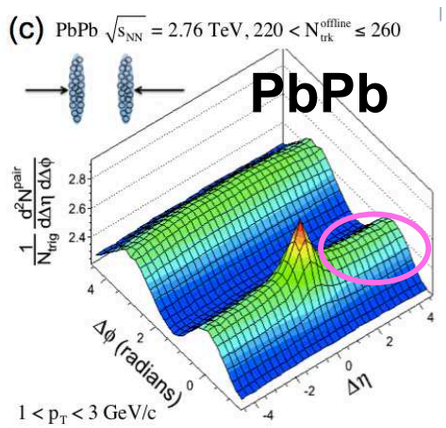



vs.



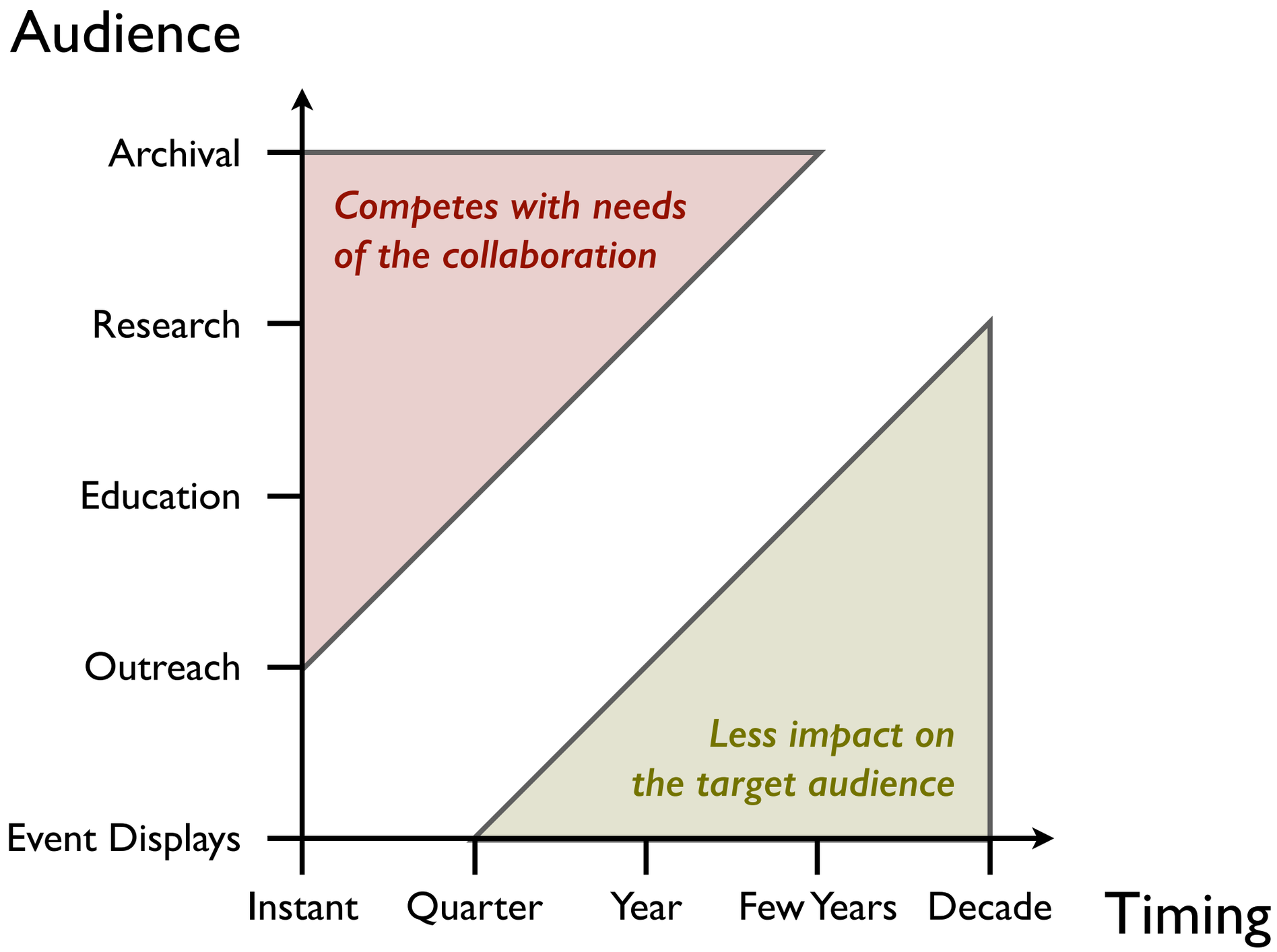



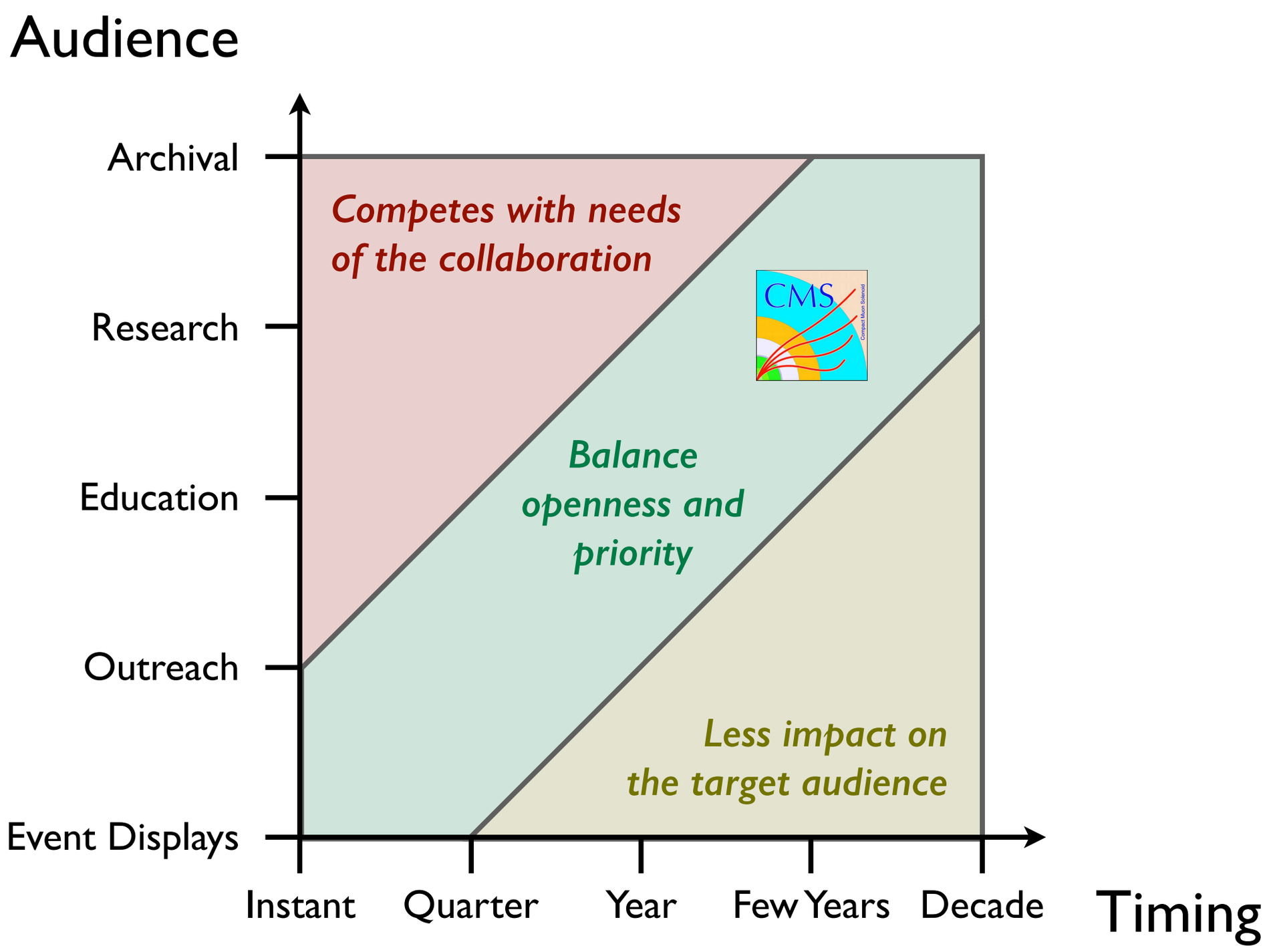




[Badea, Baty, Chang, Innocenti, **Yen-Jie Lee**, Maggi, McGinn, Peters, Sheng, JDT, appeared at **Quark Matter 2018**]

# Different Options for "Public Data"



# Different Options for "Public Data"



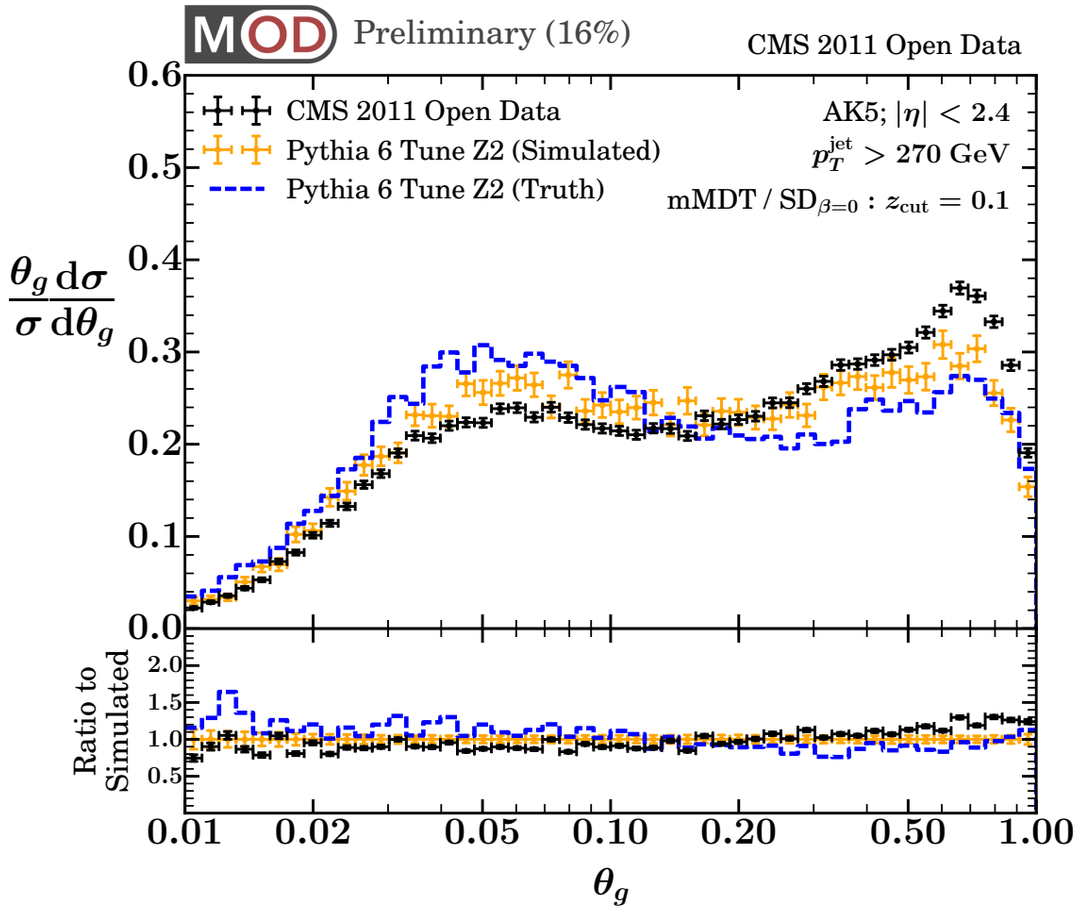
*Data preservation (and outside analyses)*
*require significant resources:*

People, time, ideas, and money

To help justify these resources,
let me address three common concerns
about public collider data raised by our work

# Balance between Sophistication and Exploration

*"There is no way you can do an external analysis with the same degree of sophistication as within the collaboration"*

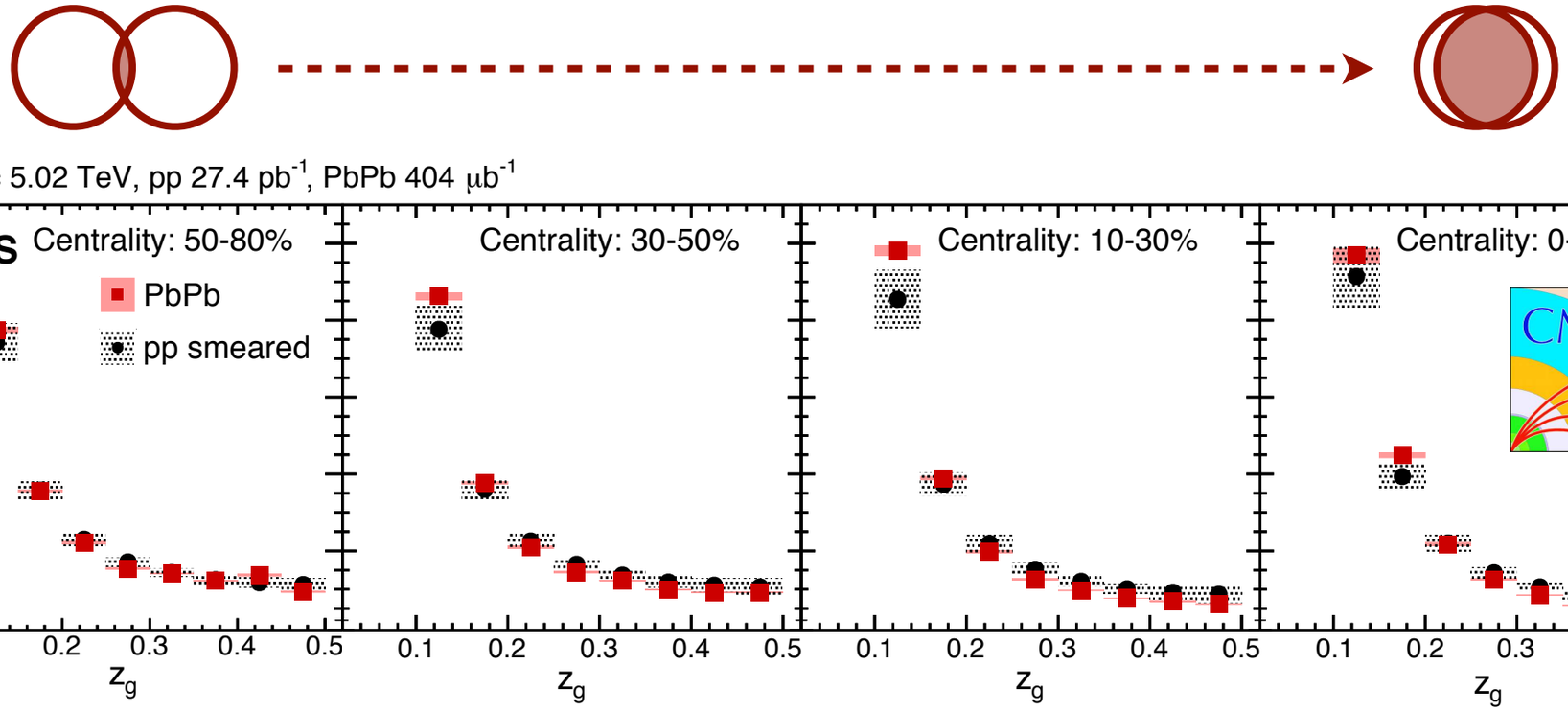



Agreed (mostly)

But with unexpected theoretical/experimental issues at play, value in exploratory studies

# Synergy between Internal and External Efforts

*"This work competes with ongoing collaboration analyses without the scrutiny of internal review"*

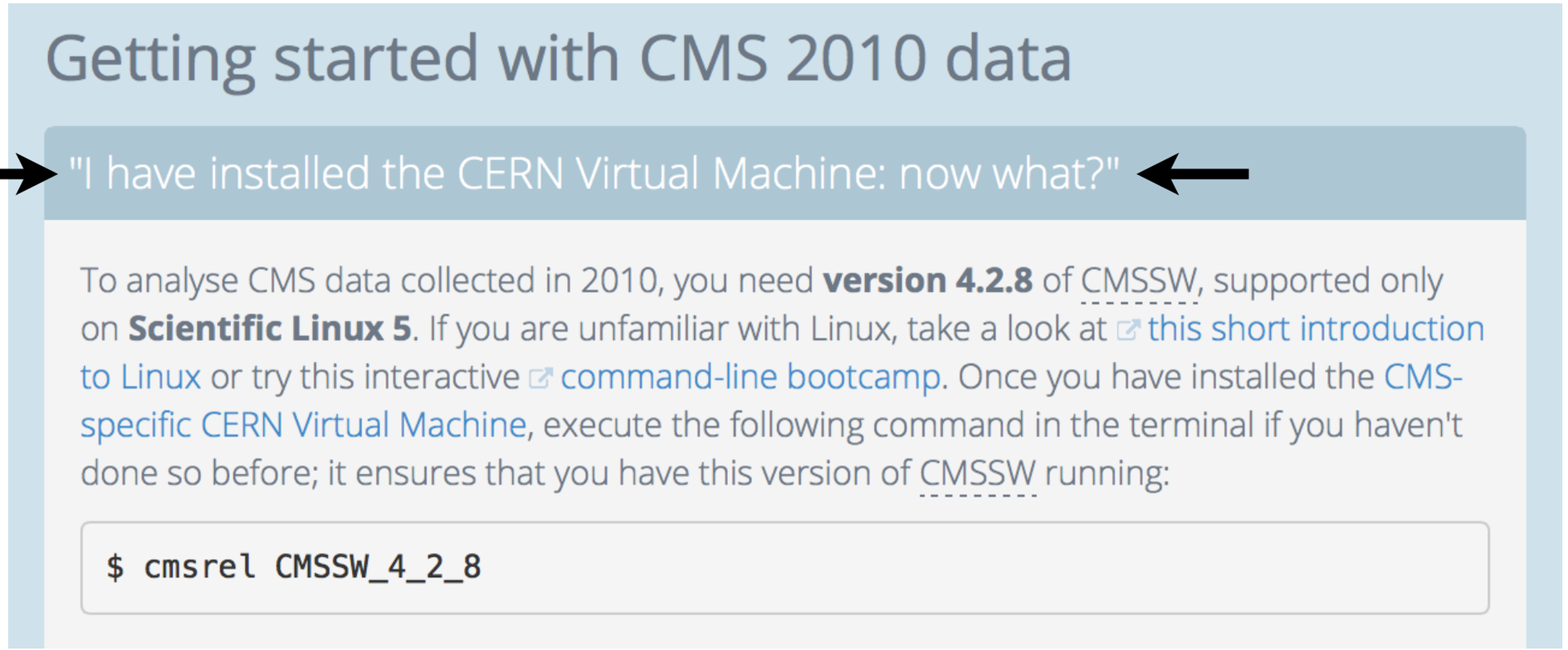



Agreed, but fine line between "compete" and "complement"
Important to have robust peer review (e.g. dual referees)

[CMS, 1708.09429 ⇒ **Phys. Rev. Lett. 120:142302**]

# Value of Open-Ended Investigations

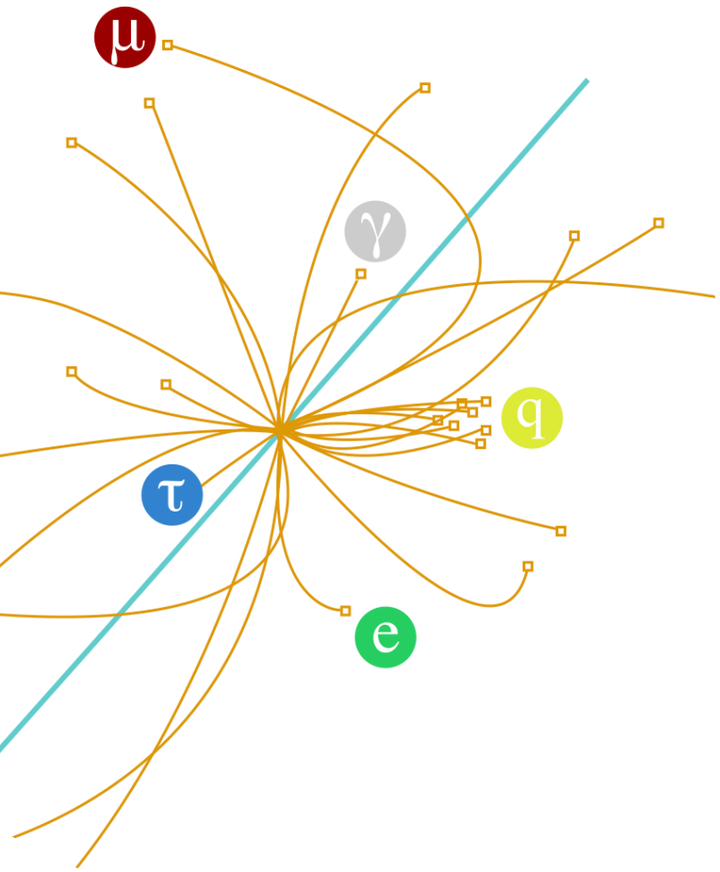
*"If you really wanted to do this jet substructure measurement, you should have joined CMS as an associate member"*




Agreed, but what I really wanted to do is figure out the answer to this question (curiosity-driven research)

# My View

*The CMS Open Data is a fantastic resource, with many exciting applications*




Educating future scientists

Stress-testing archival data strategies

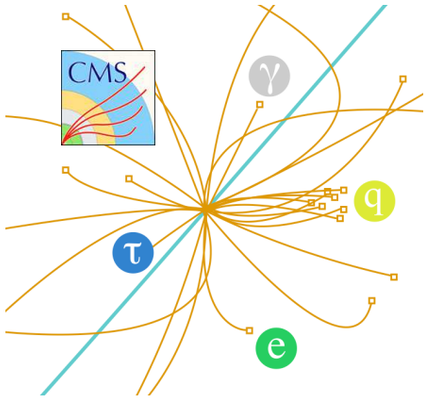
Enabling exploratory/proof-of-principle studies

Facilitating dialogue between theory and experiment

Researching physics in and beyond the standard model

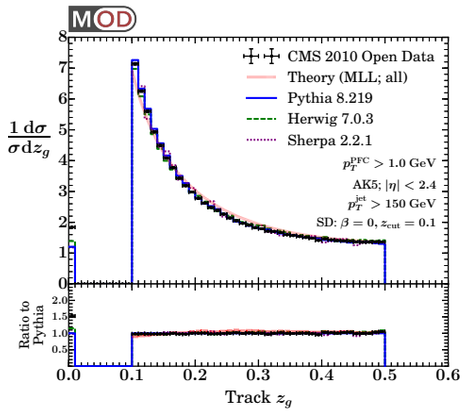
*These are only possible with sustained investment in public data initiatives*

# Summary



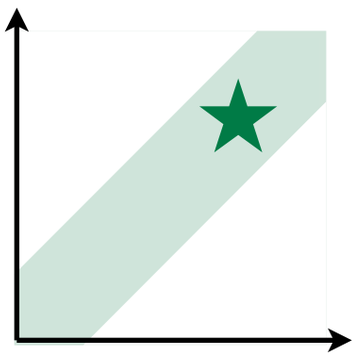

## Using the CMS Open Data

*Unique collider data set, ideal for exploratory studies*




## Jet Substructure and QCD Splittings

*Exposing the universal singularity structure of gauge theories*



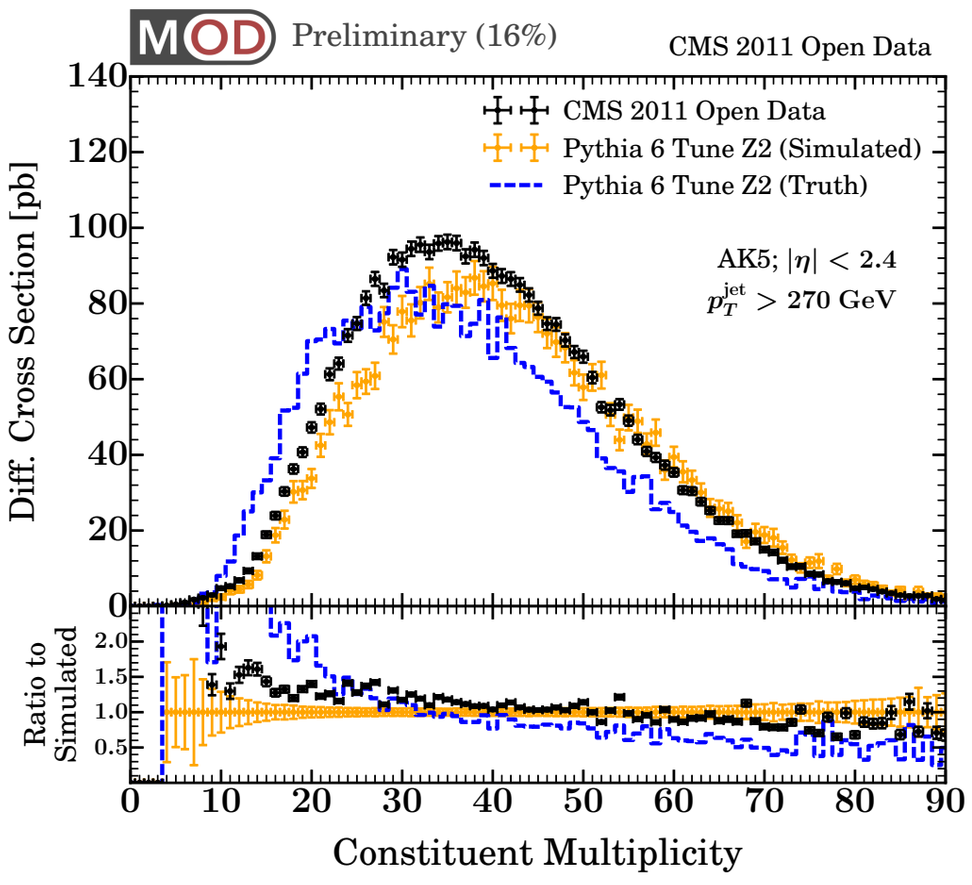
## The Future of Public Collider Data

*Sustained investment from outreach to research to archives*

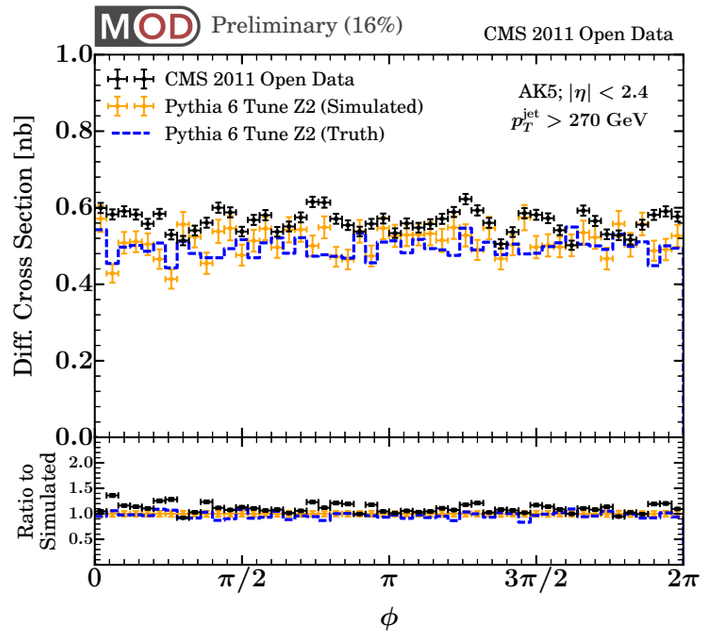
# *Backup Slides*

# Additional 2011 Plots

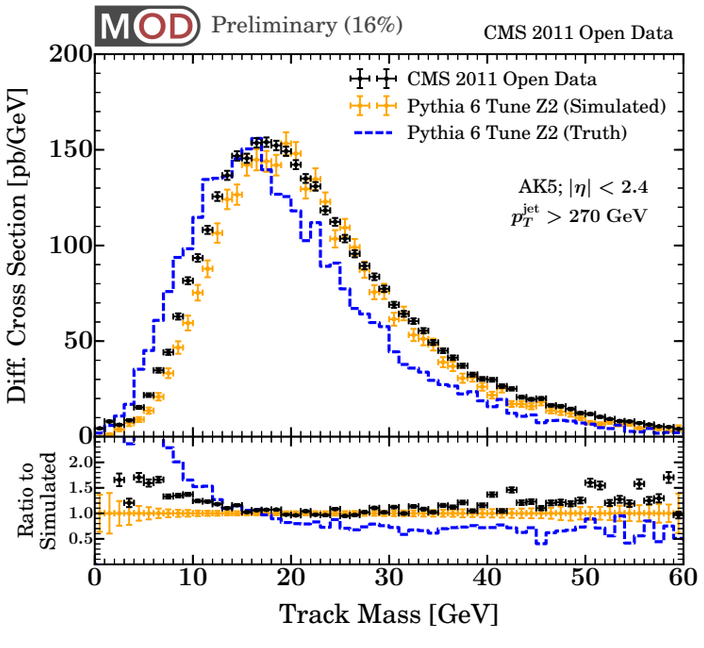
## Constituent Multiplicity



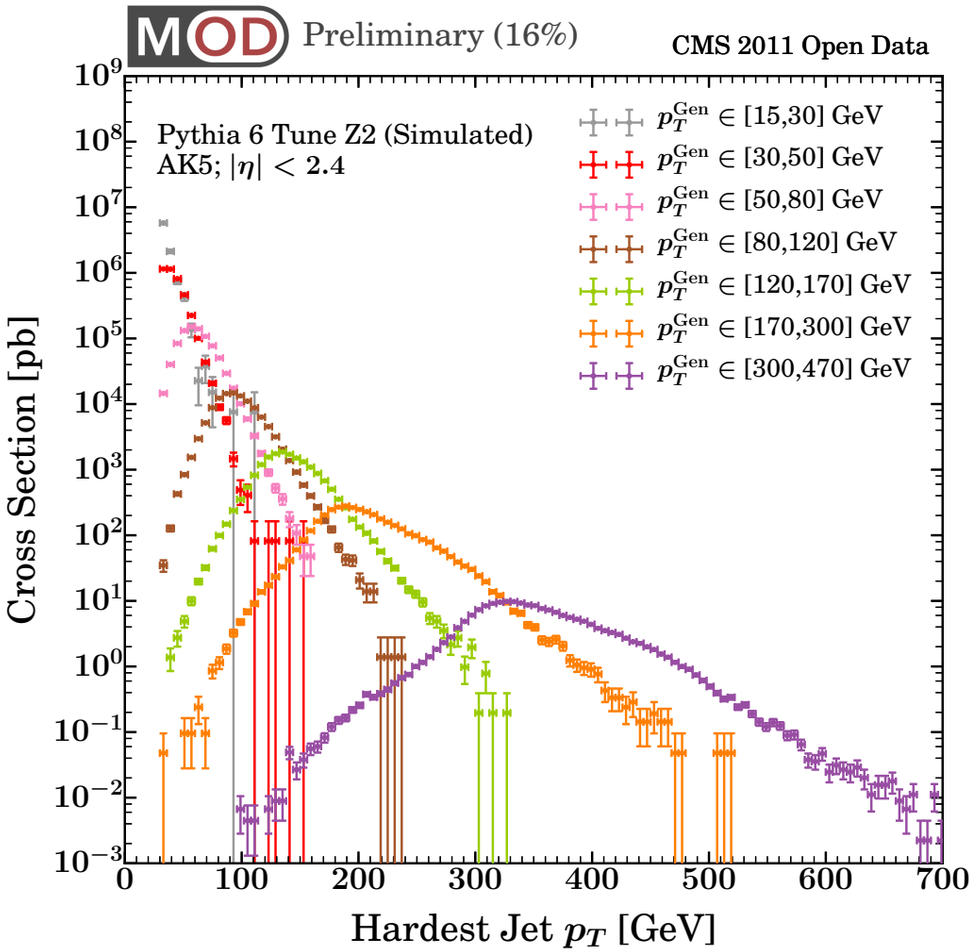

## Azimuth

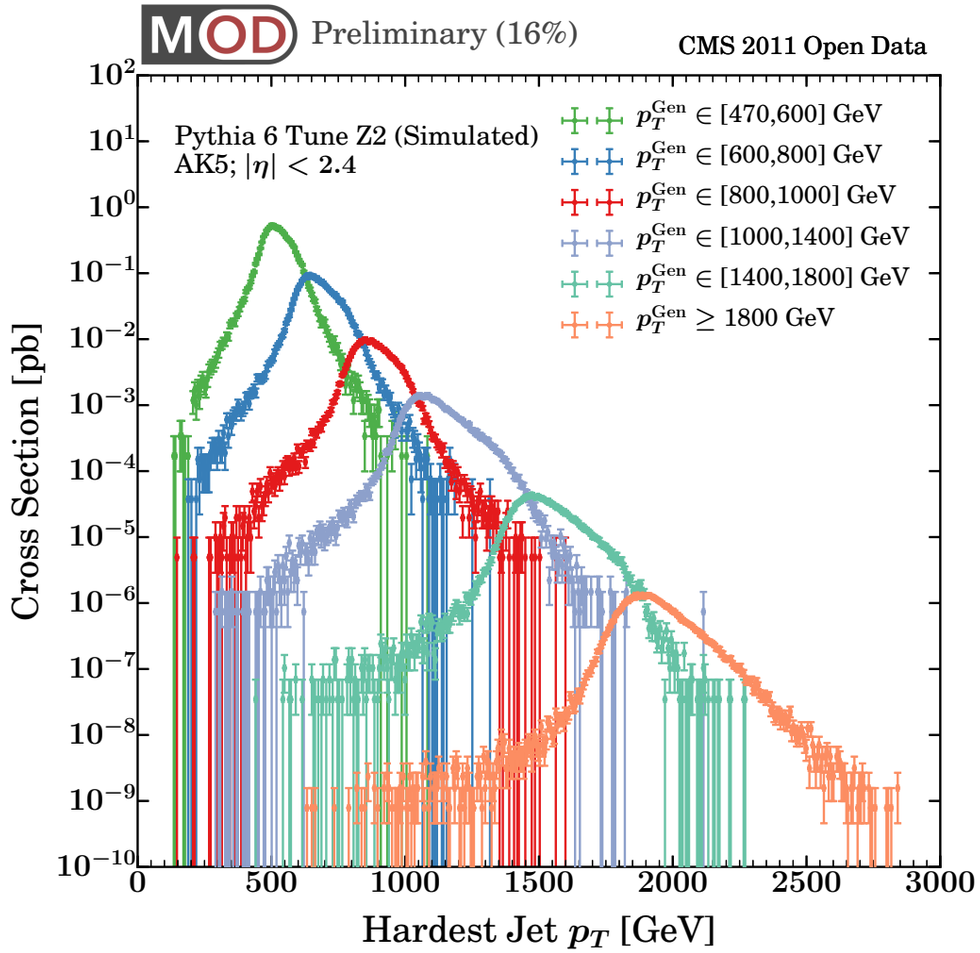



## Track Mass

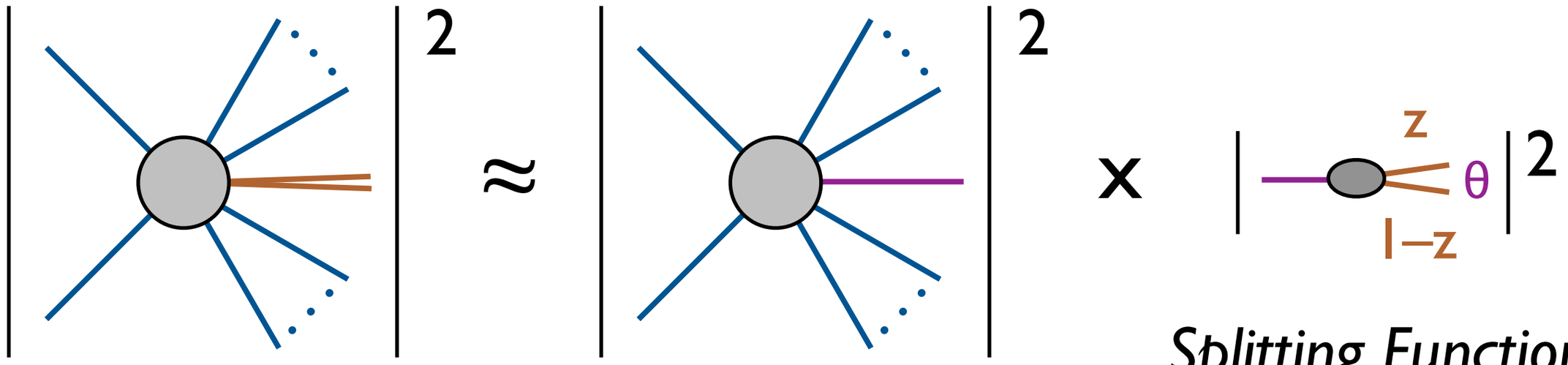


# Sewing Together Simulated Data Samples



# Textbook QCD: Universal Collinear Limit

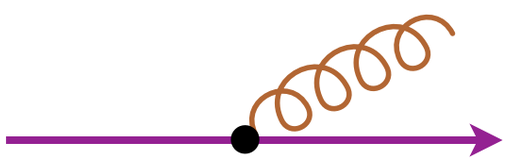



2→n

2→n–1

*Splitting Function*

1→2



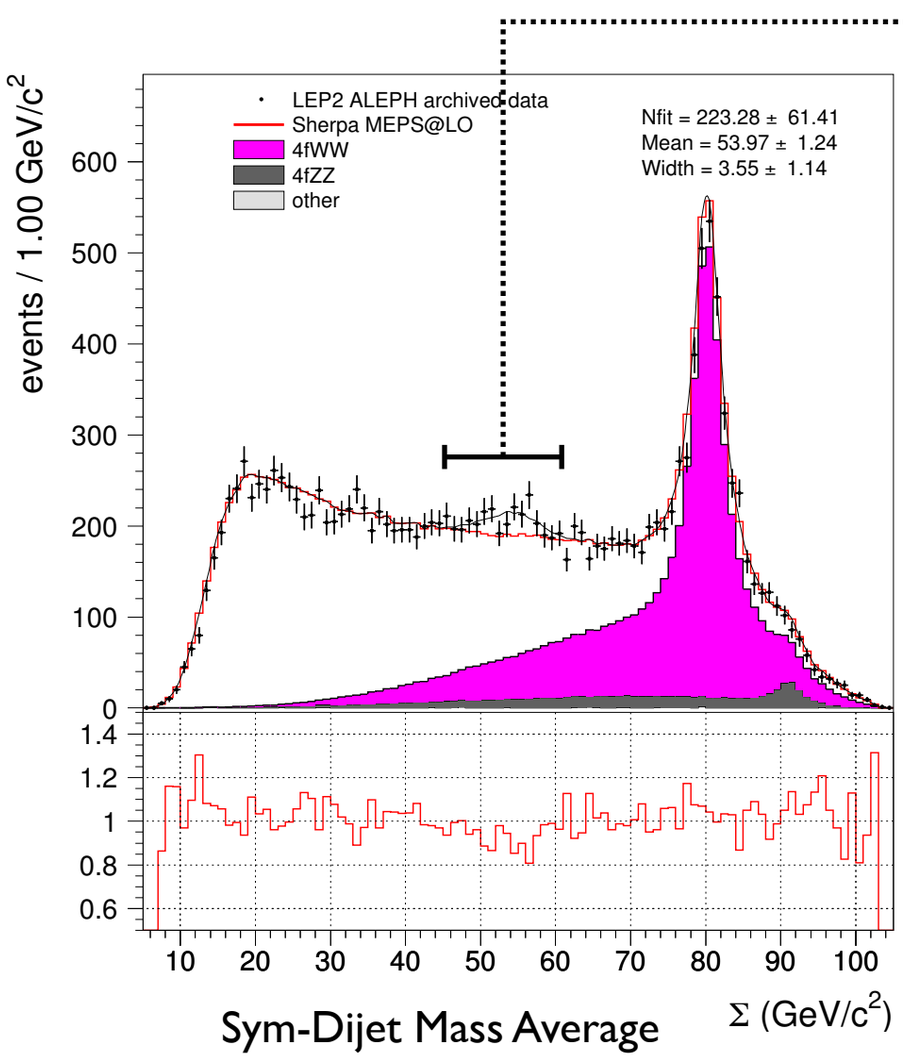
$C_q = 4/3$
$C_g = 3$

$$\mathrm{d}P_{i\to ig} \simeq \frac{2\alpha_s}{\pi} C_i \frac{\mathrm{d}\theta}{\theta} \frac{\mathrm{d}z}{z}$$

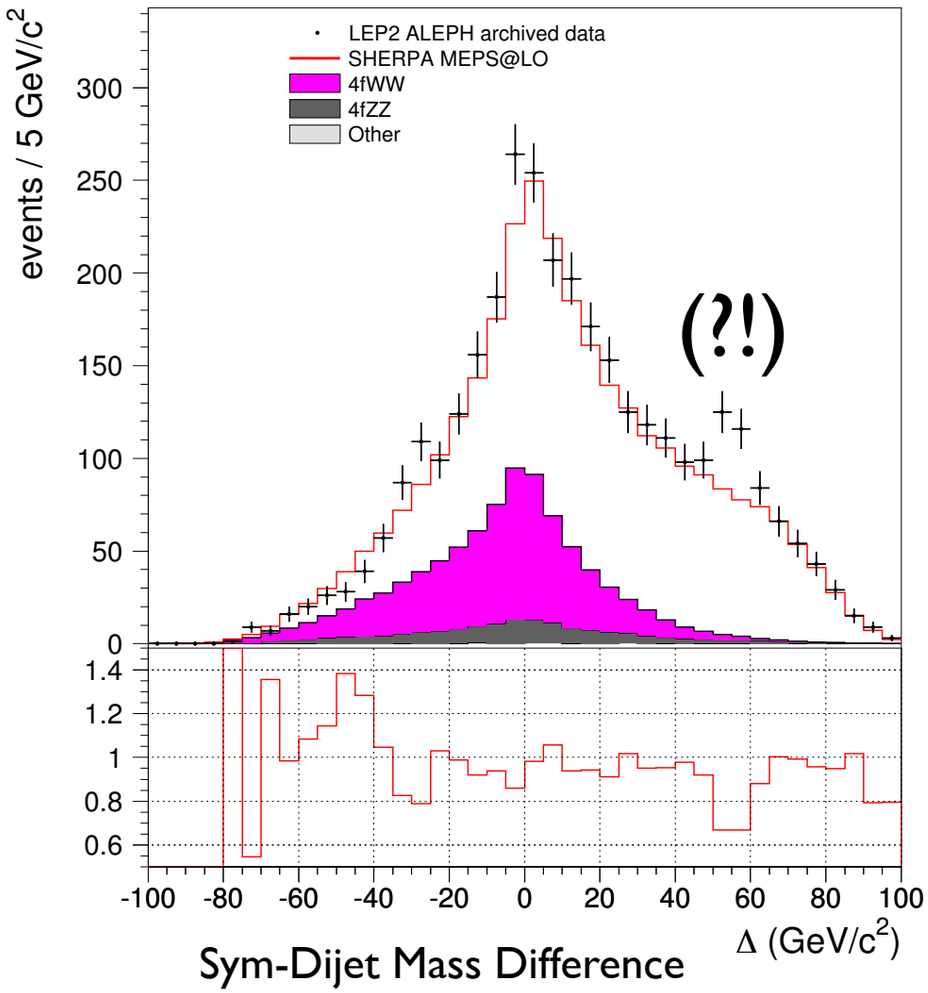
Collinear singularity

Soft singularity

# A Quad-Jet Puzzle in Archival ALEPH Data




$\Delta$ for 45 GeV $< \Sigma <$ 61 GeV

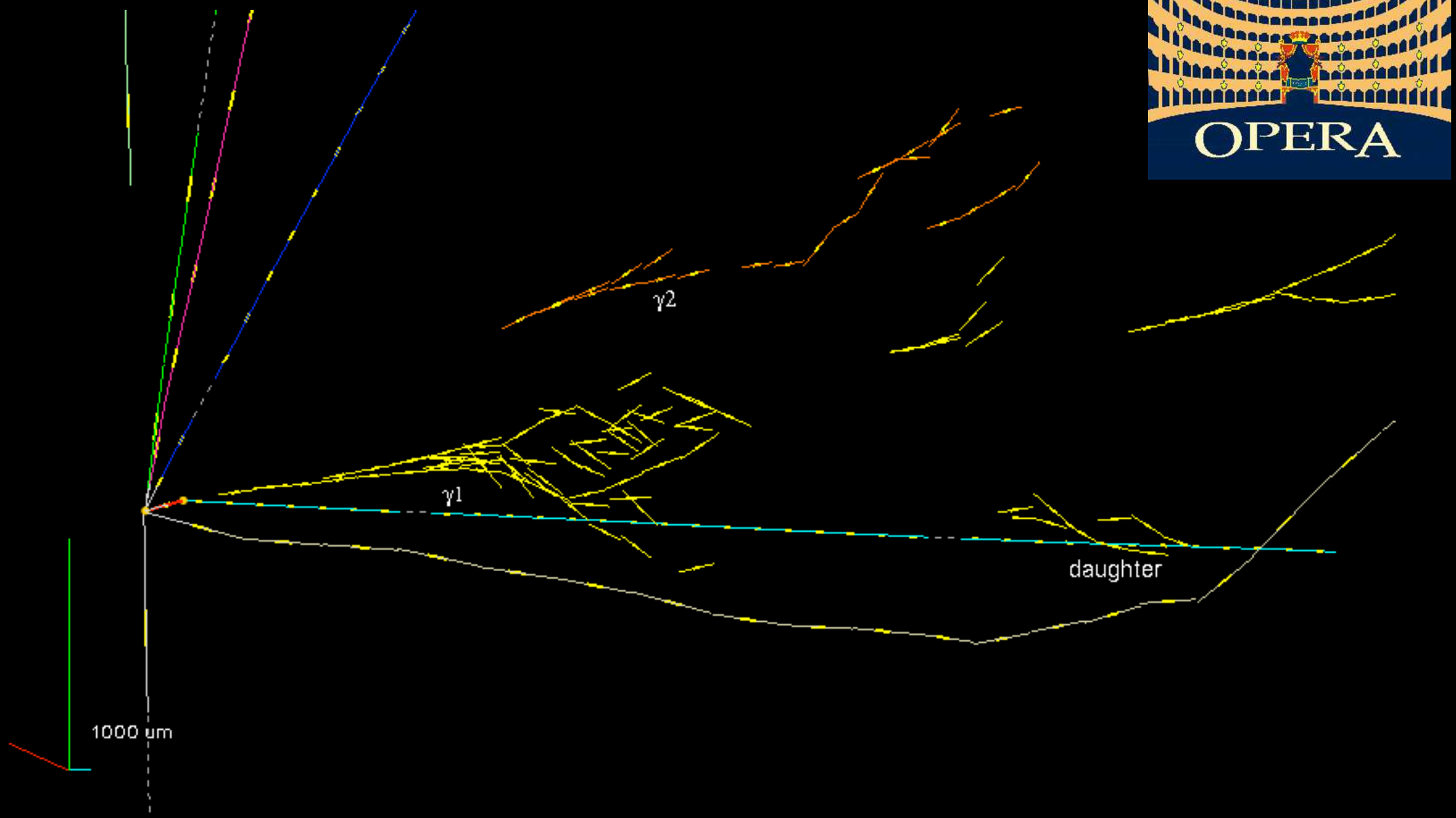



[Kile, von Wimmersperg-Toeller, 1706.02242, 1706.02255, 1706.02269]

# *Example of stress-testing archival data strategies*

```
----- Begin Fatal Exception 07-Jun-2018 14:34:22 EDT-----------------------
An exception of category 'FileOpenError' occurred while
   [0] Constructing the EventProcessor
   [1] Constructing input source of type PoolSource
   [2] Calling RootInputFileSequence::initFile()
   [3] Calling StorageFactory::open()
   [4] Calling XrdFile::open()
Exception Message:
Input file root://eospublic.cern.ch//eos/opendata/cms/Run2011A/Jet/AOD/12Oct2013-v1/20001/76C7AE0D-1E3F-E311-9EB3-00304
8D2BD66.root was not found, could not be opened, or is corrupted.
   Additional Info:
      [a] XrdClient::Open(name='root://eospublic.cern.ch//eos/opendata/cms/Run2011A/Jet/AOD/12Oct2013-v1/20001/76C7AE0D
-1E3F-E311-9EB3-003048D2BD66.root', flags=0x10, permissions=0666) => error '' (errno=10000)
      [b] Current server connection: root://eospublic-srv-m1.cern.ch:1094//eos/opendata/cms/Run2011A/Jet/AOD/12Oct2013-
v1/20001/76C7AE0D-1E3F-E311-9EB3-003048D2BD66.root
----- End Fatal Exception -------------------------------------------------
```

# *New: Public neutrino data!*




*Derived data from 2008-2012 ⇒ Released May 2018*