The Future is Open

Adventures with Public Collider Data

Jesse Thaler

Fermilab Colloquium — September 30, 2020

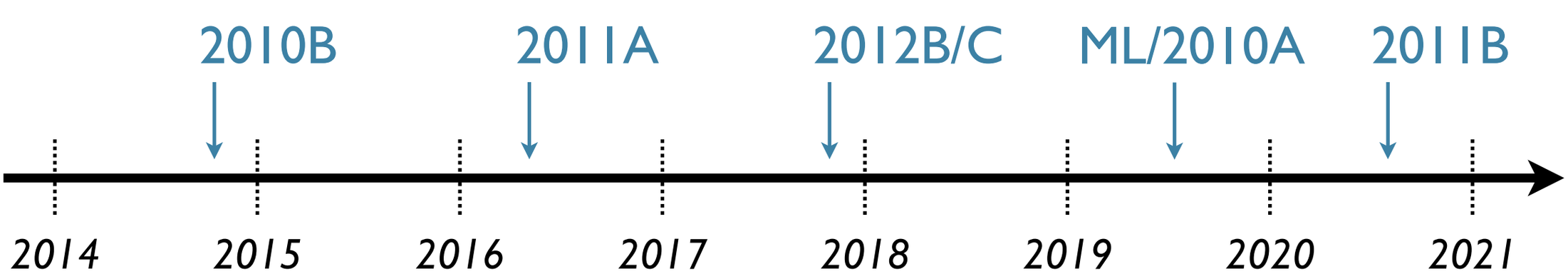
Wearing my New Hat…

The NSF AI Institute for
Artificial Intelligence and Fundamental Interactions

Postdoctoral fellowship opportunity (Oct 20 deadline):
http://iaifi.org/fellows.html

The CMS experiment is pioneering the release of research-grade public collider data

[First Release, 2014; Second Release, 2016; Third Release, 2017; Fourth Release, 2019; Fifth Release, 2020]

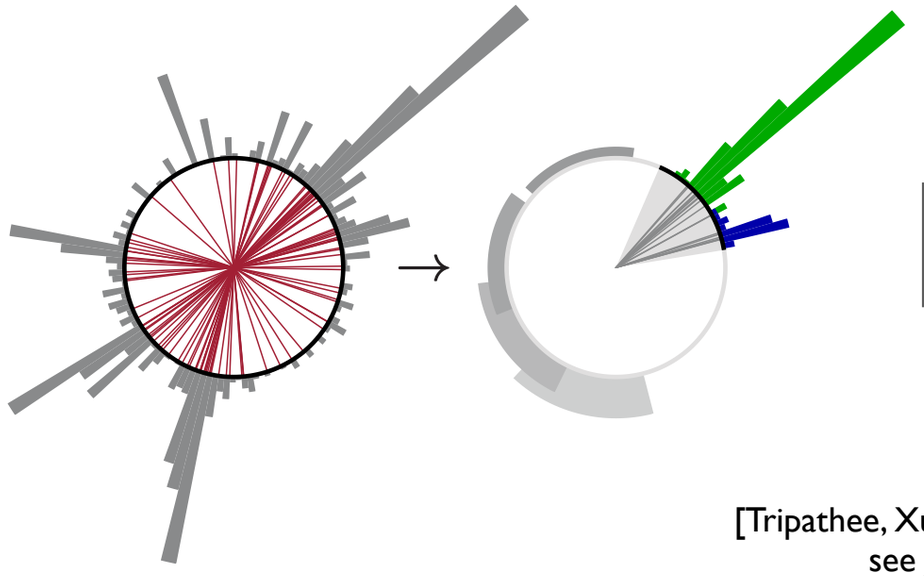


External (and internal) users are taking advantage of this unique scientific resource

PRL **119,** 132003 (2017) PHYSICAL REVIEW LETTERS week ending 29 SEPTEMBER 2017

Exposing the QCD Splitting Function with CMS Open Data

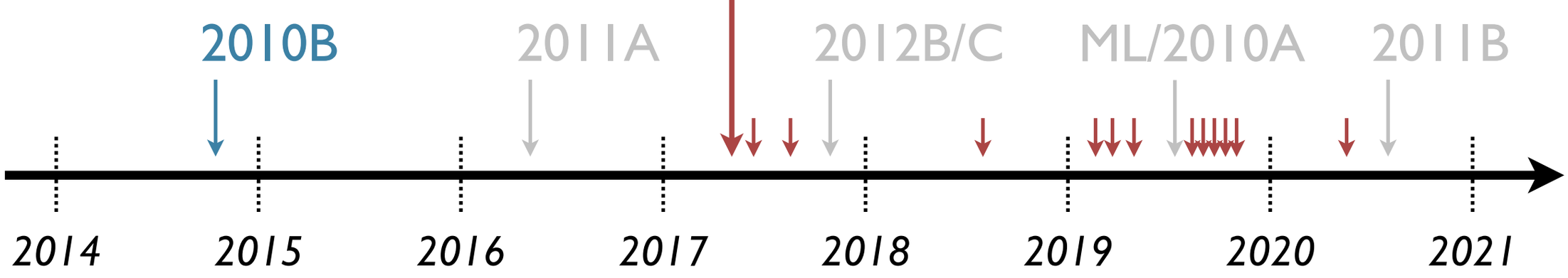
Andrew Larkoski,[1,*] Simone Marzani,[2,†] Jesse Thaler,[3,‡] Aashish Tripathee,[3,§] and Wei Xue[3,‖]



MOD

MIT Open Data

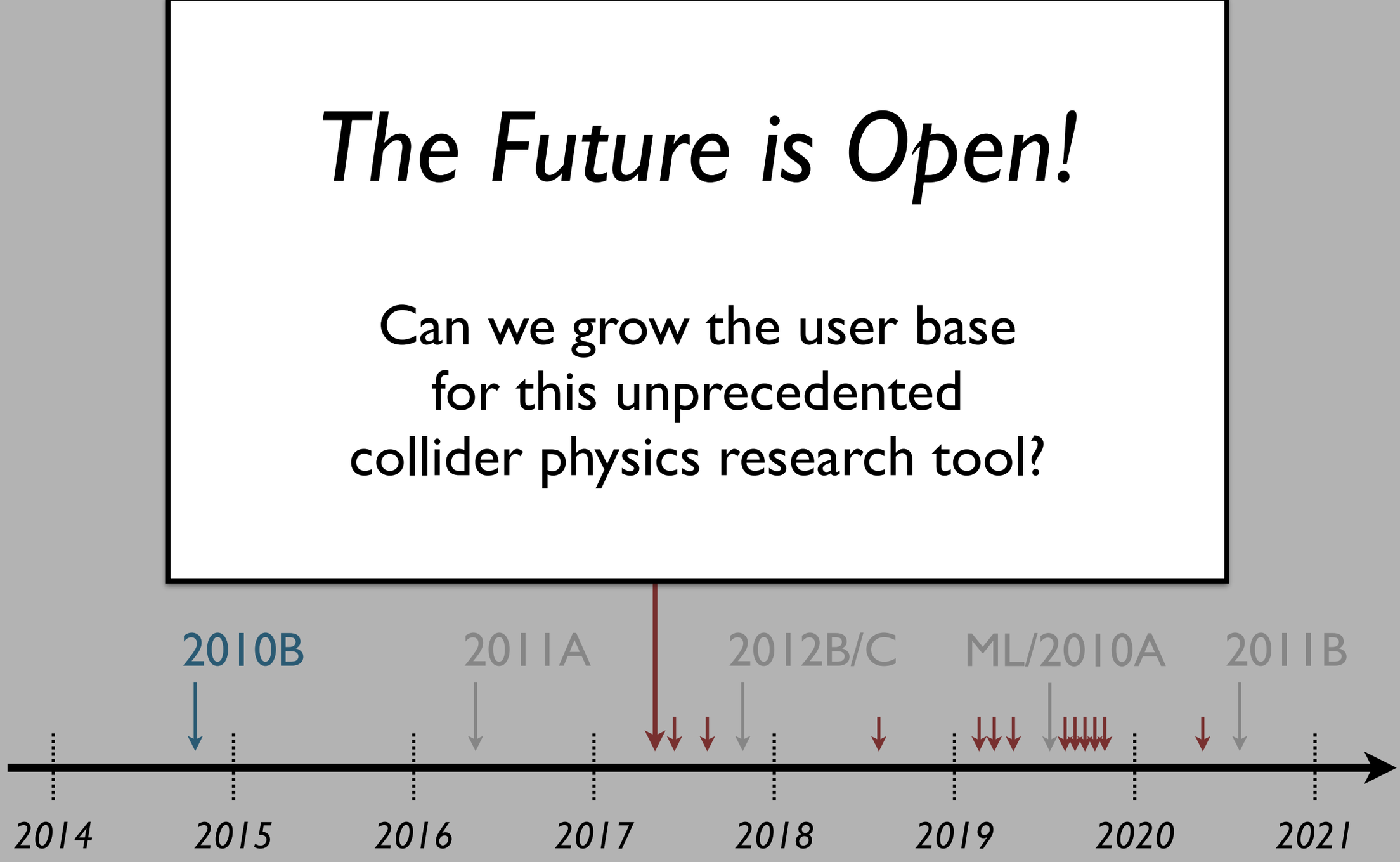
[Tripathee, Xue, Larkoski, Marzani, JDT, PRL 2017; see JDT, Fermilab Wine & Cheese 2016]



External (and internal) users are taking advantage of this unique scientific resource

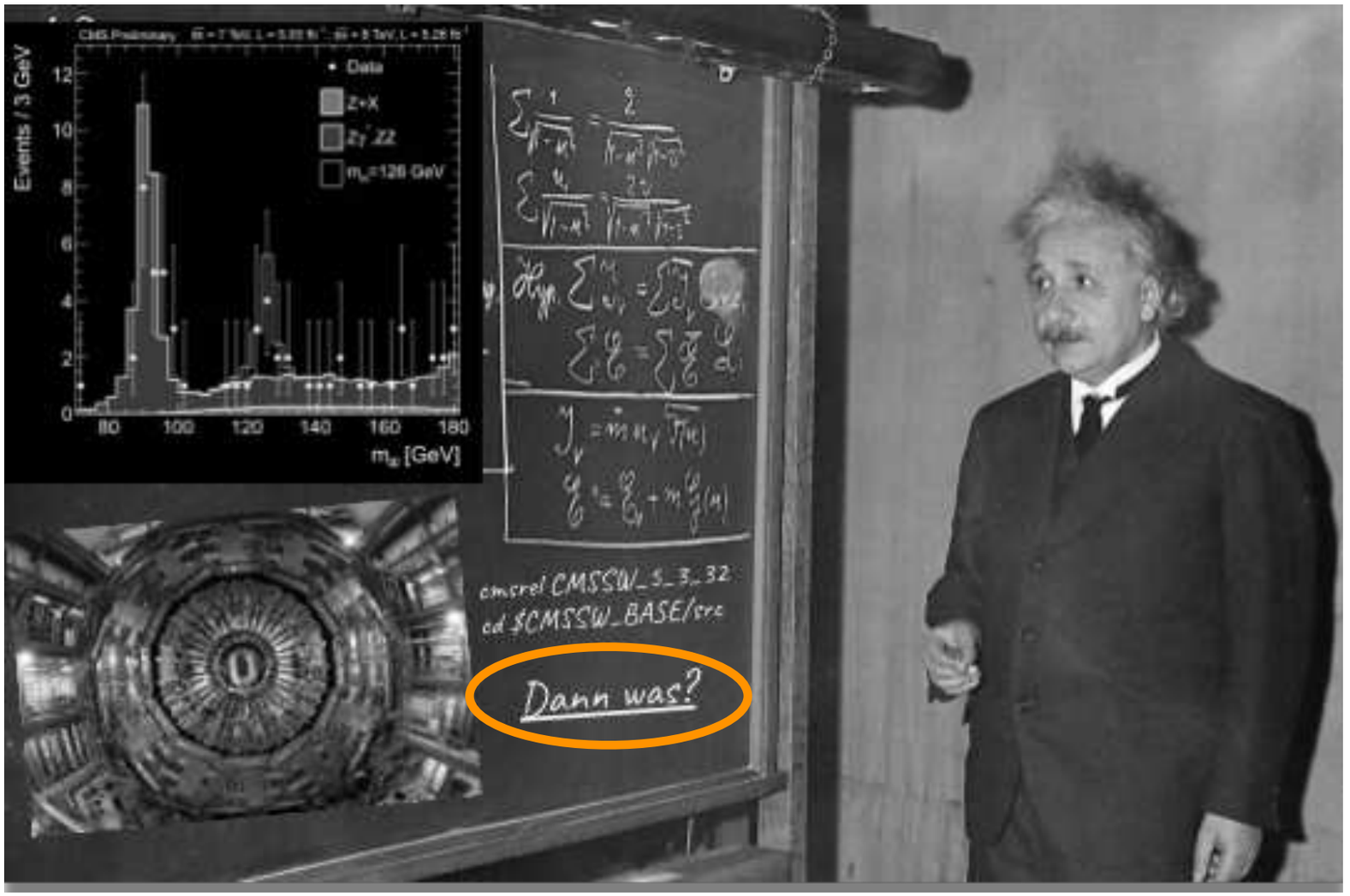
The Future is Open!

Can we grow the user base
for this unprecedented
collider physics research tool?



CMS Open Data Workshop for Theorists

Remote Fermilab LPC Workshop, September 30 – October 2, 2020



Organizing Committee:
- Matthew Bellis (Siena College)
- Edgar Carrera (U. San Francisco de Quito)
- Kati Lassila-Perini (U. of Helsinki)
- Jesse Thaler (MIT)

Local Organizing Committee:
- Gabriele Benelli (Brown U.)
- Christian Herwig (Fermilab)
- Julie Hogan (Bethel U. and Brown U.)
- Clemens Lange (CERN)
- Andrew Melo (Vanderbilt U.)
- Nada Mohamed (Siena College)
- Stephen Mrenna (Fermilab)
- Kevin Pedro (Fermilab)
- Emanuele Usai (Brown U.)
- David Yu (Brown U.)

LPC Events Committee:
- Gabriele Benelli (Brown U., Co-Chair)
- Kevin Pedro (Fermilab, Co-Chair)

LPC Coordinators:
- Cecilia Gerber (UIC)
- Sergo Jindariani (Fermilab)

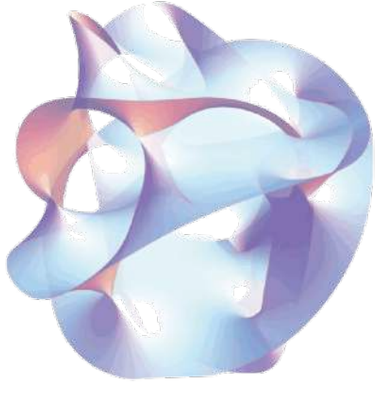
Goal: Enable theorists to use real experimental collider data in their research

[CMS Open Data Workshop, Sept-Oct 2020]

Cartoon of Collider Physics Community

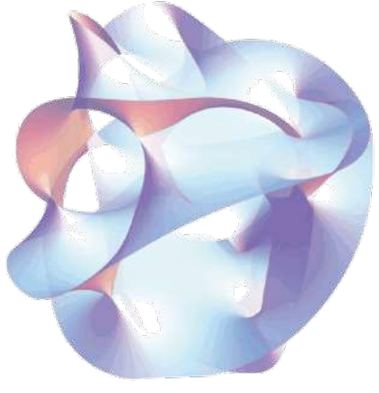
"Theorists"

"Experimentalists"



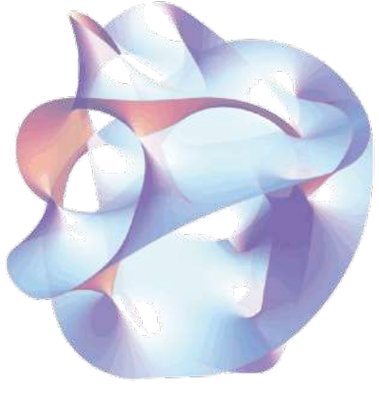
Cartoon of Collider Physics Community

"Theorists" *"Phenomenologists"* *"Experimentalists"*

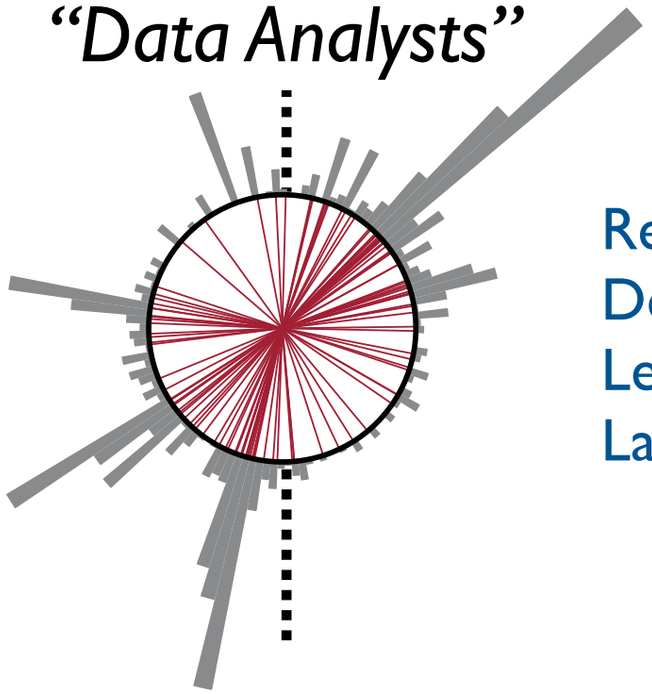


Cartoon of Collider Physics Community

"Theorists" *"Phenomenologists"* *"Experimentalists"*



"Data Analysts"

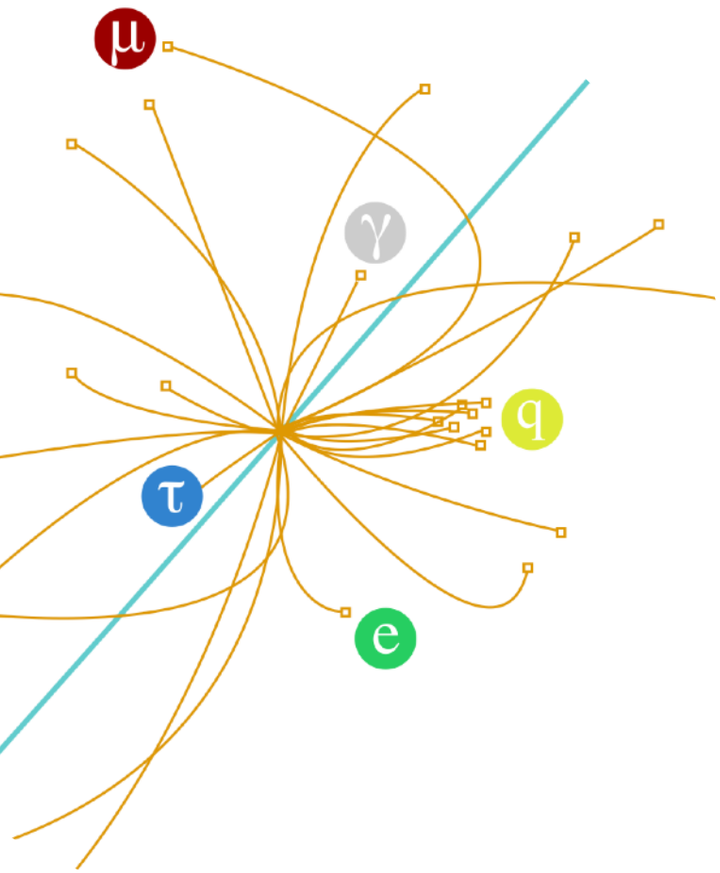


Simulated datasets
Calculational uncertainties
Proof-of-concept studies
Small collaborations

Real datasets
Detector uncertainties
Legacy analyses
Large collaborations

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

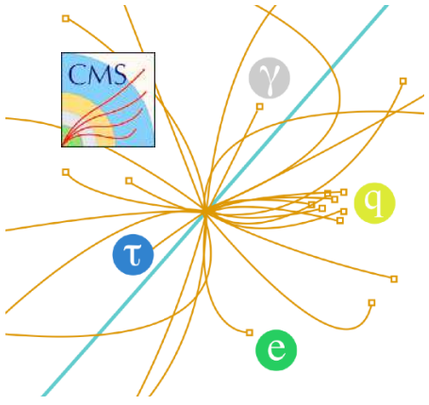
Facilitating dialogue between theory and experiment

Researching physics in and beyond the standard model

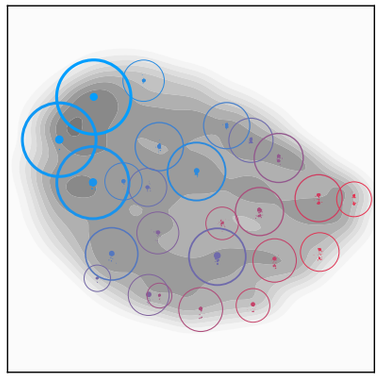
These are only possible with sustained investment in public data initiatives

[JDT, CMS Blog 2017]

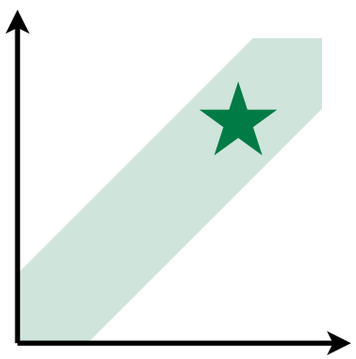
Outline



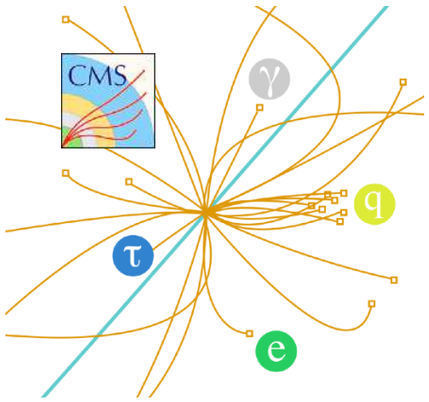
Introducing the CMS Open Data



Adventures with Public Collider Data



Looking to the Past, Present, and Future



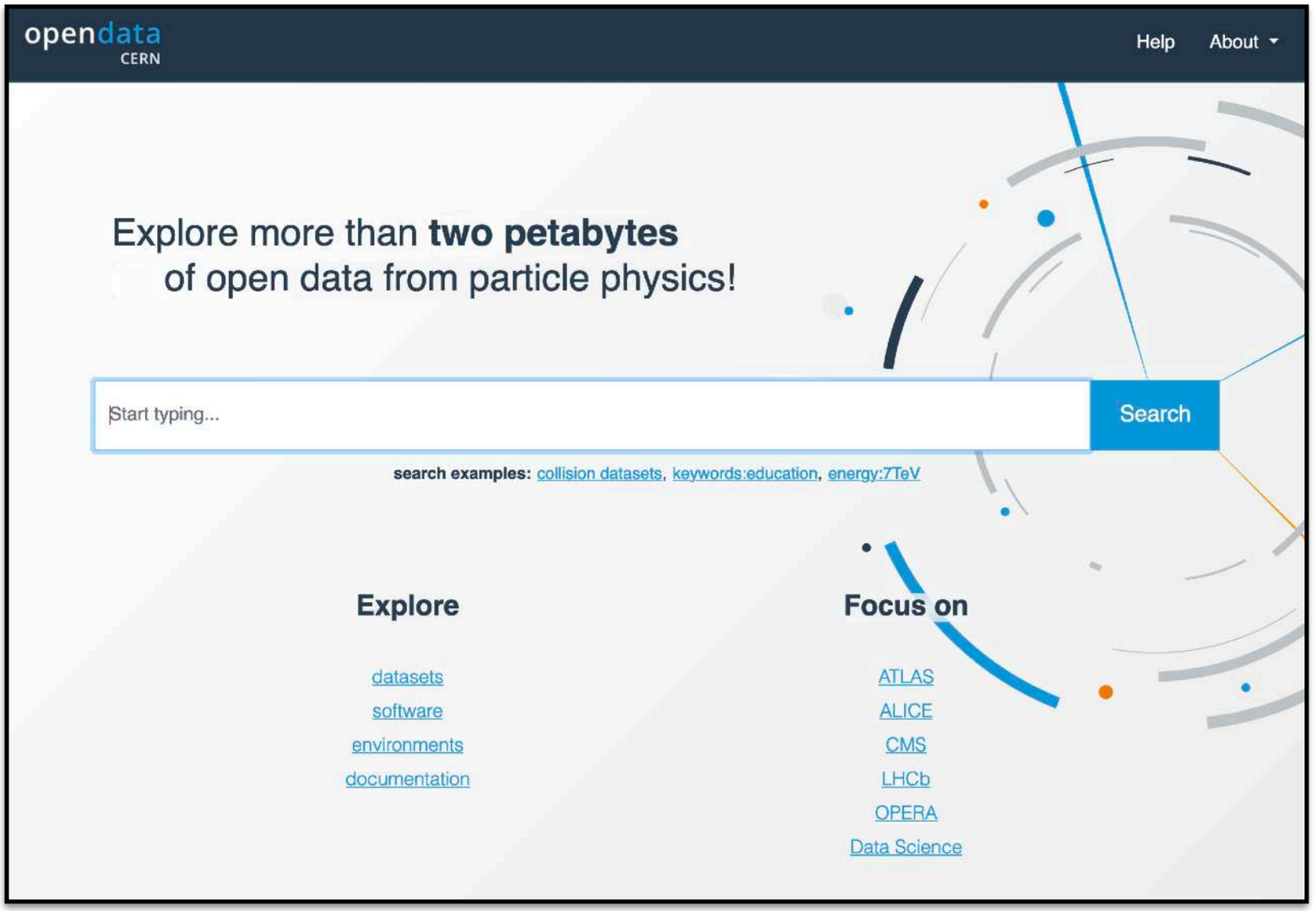
Introducing the CMS Open Data

Adventures with Public Collider Data

Looking to the Past, Present, and Future

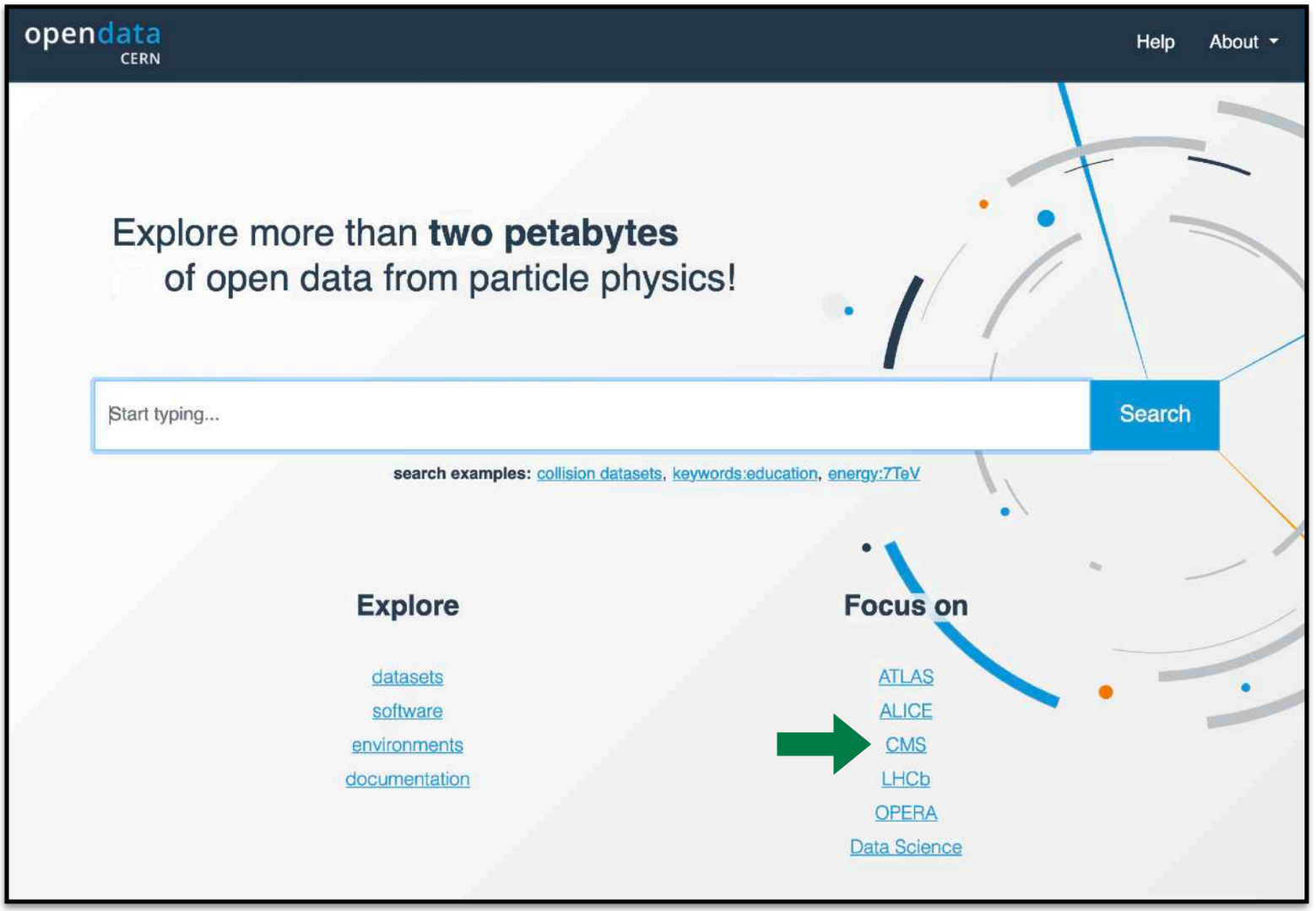
What is CMS Open Data?

http://opendata.cern.ch/



What is CMS Open Data?

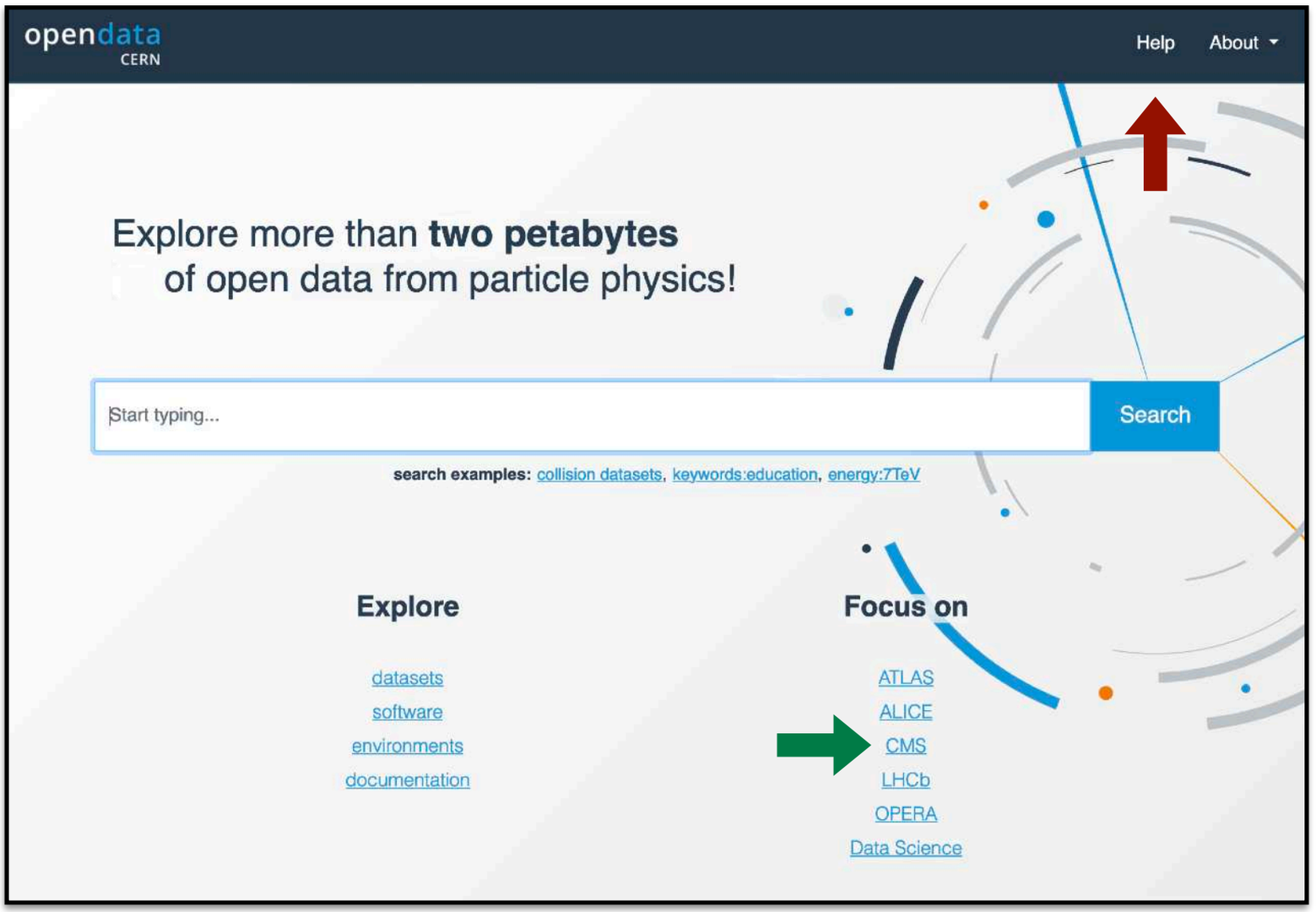
http://opendata.cern.ch/



Same data format (AOD) and software framework (CMSSW) used by CMS

What is CMS Open Data?

http://opendata.cern.ch/



Same data format (AOD) and software framework (CMSSW) used by CMS

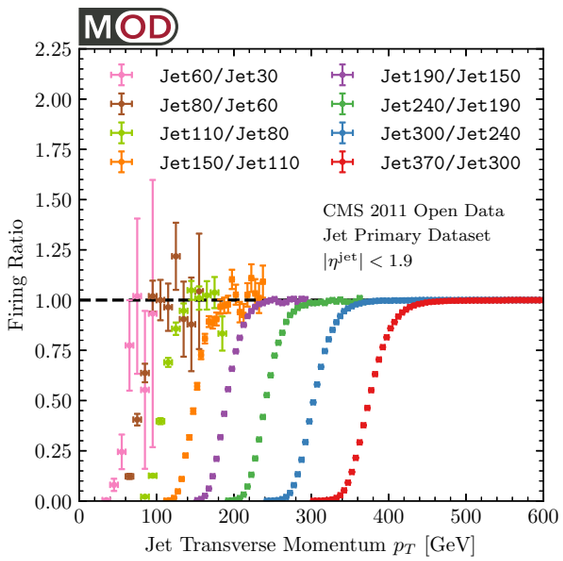
Same steep learning curve (hence this week's LPC workshop)

What is CMS Open Data?

http://opendata.cern.ch/

Same experimental challenges!

Accounting for triggers



[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020]

"Ich habe das [Strahlenergie Korrekturen] mit Stumpf und Stil verlernt." ~ Emmy Noether

Acceptances, efficiencies, uncertainties

	Central Value	Uncertainty
$\mathcal{L}$	$2.11\,\text{fb}^{-1}$	2.2%
A_Z	0.392	2.4%
$\sqrt{\epsilon_{\text{tr}}^Z}$ (*i.e.* per muon)	0.924	2.4%
$\sqrt{\epsilon_{\text{iso}}^Z}$ (*i.e.* per muon)	0.966	1.5%
Background	—	1.0%
Combined ($\mathcal{L}\, A_Z\, \epsilon_{\text{tr}}^Z\, \epsilon_{\text{iso}}^Z$)	$0.659\,\text{fb}^{-1}$	5.3%

$$I_{\text{comb}} = \frac{\left(p_T^{\text{track}} + E_T^{\text{ECAL}} + E_T^{\text{HCAL}}\right)_{R<0.3}}{p_T^{\mu}}$$

[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019]

New opportunities for investigation (and discovery)!

In Backup

"Researching physics in and beyond the Standard Model"

All 19 papers (thus far) using CMS Open Data

Standard Model Analyses

[Tripathee, Xue, Larkoski, Marzani, JDT, PRL 2017, PRD 2017]
[Apyan, Cuozzo, Klute, Saito, Schott, Sintayehu, JINST 2020]

BSM Searches

[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019]
[Lester, Schott, JHEP 2019]

Machine Learning Studies

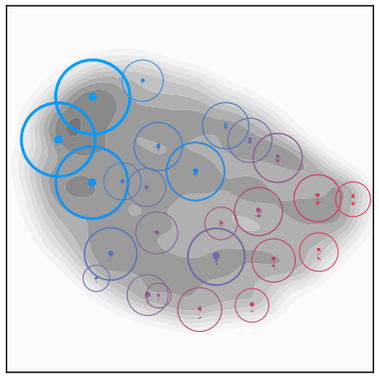
[Fernández Madrazo, Heredia Cacha, Lloret Iglesias, Marco de Lucas, EPJWoC 2019]
[Andrews, Paulini, Gleyzer, Poczos, CSBS 2020]
[Andrews, Alison, An, Bryant, Burkle, Gleyzer, Narain, Paulini, Poczos, Usai, NIM 2020]
[Moreno, Nguyen, Vlimant, Cerri, Newman, Periwal, Spiropulu, Duarte, Pierini, PRD 2020]
[Knapp, Dissertori, Cerri, Nguyen, Vlimant, Pierini, arXiv 2020]

And More!

[Pata, Spiropulu, arXiv 2019]
[Paktinat Mehdiabadi, Fahim, JPG 2019]
[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020]

Please contact me if I missed your CMS Open Data study!

Introducing the CMS Open Data



Adventures with Public Collider Data

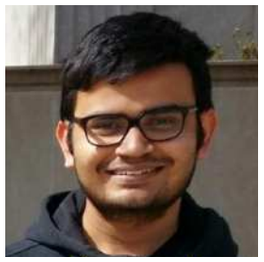
Looking to the Past, Present, and Future

MIT Open Data Publications

"Educating future scientists"

QCD Splitting Function:

[Tripathee, Xue, Larkoski, Marzani, JDT, PRL 2017, PRD 2017]


Aashish Tripathee

Wei Xue

Andrew Larkoski

Simone Marzani

Dimuon Resonance Hunt:

[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019]

Cari Cesarotti

Matt Strassler

Yotam Soreq

Wei Xue

Event Space Geometry:

[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020]

Radha Mastandrea

Preksha Naik

Patrick Komiske

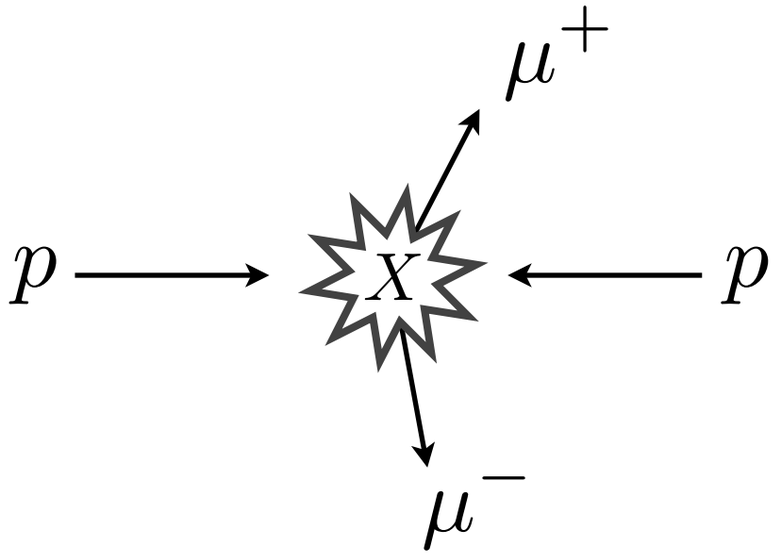
Eric Metodiev

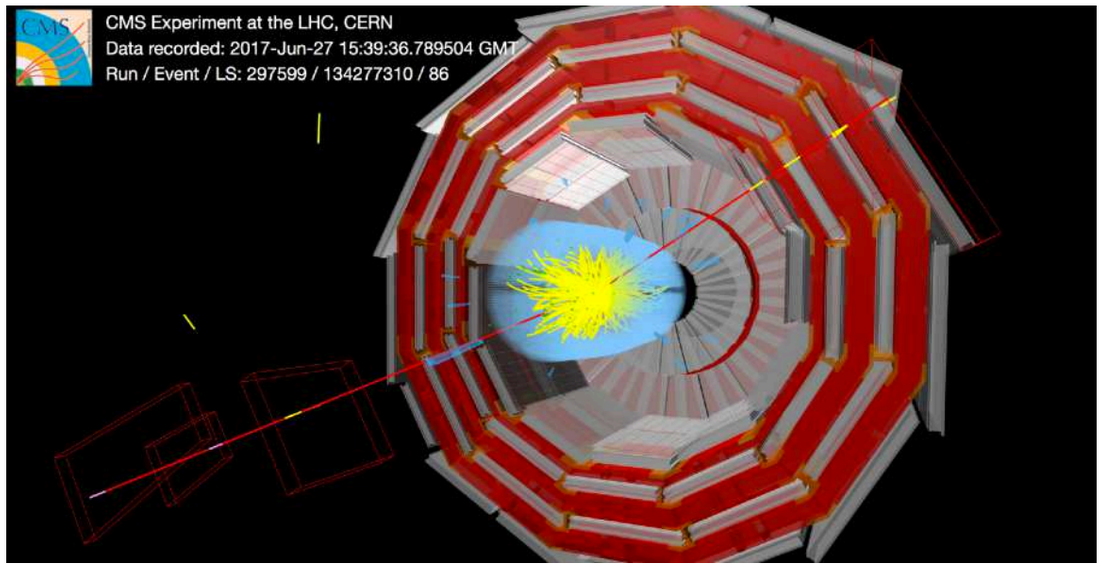
Summer interns: Alexis Romero, Ed Hirst, Max Henderson, Joe Zhou

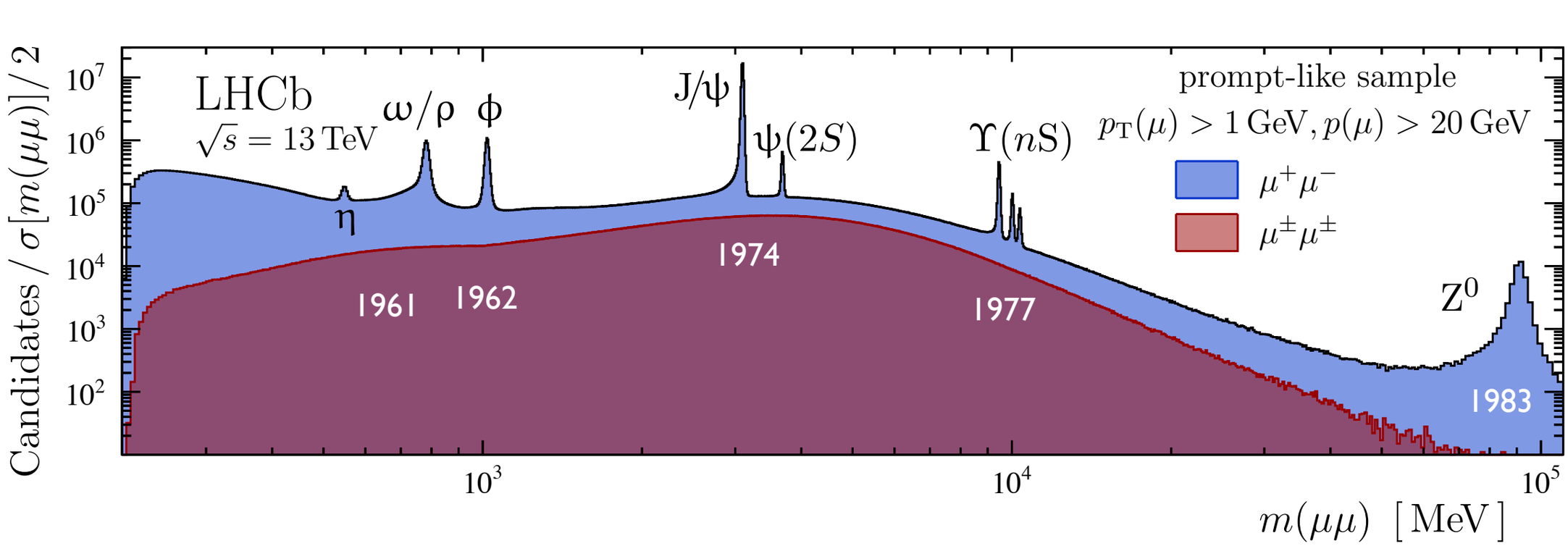
Ongoing work: Serhii Kryhin, Ian Moult, …

Dimuon Resonance Hunt

New Physics in Dimuons?



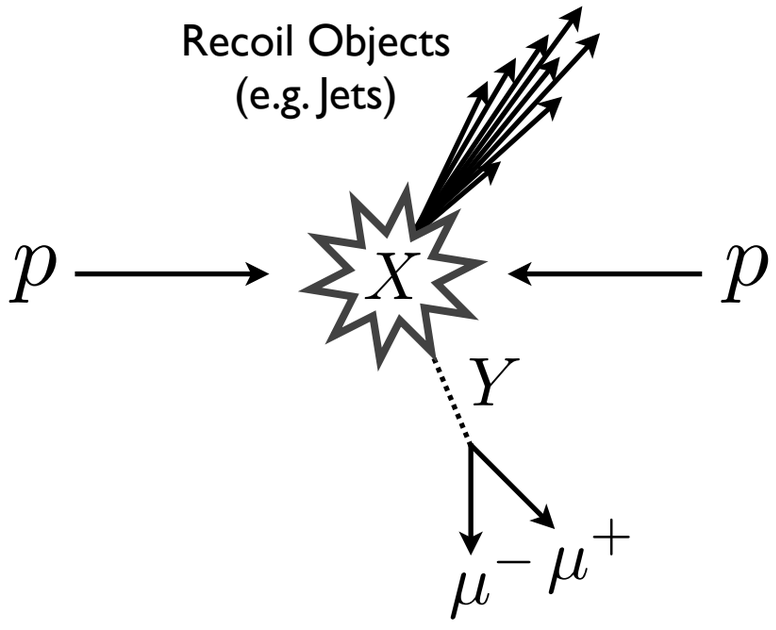




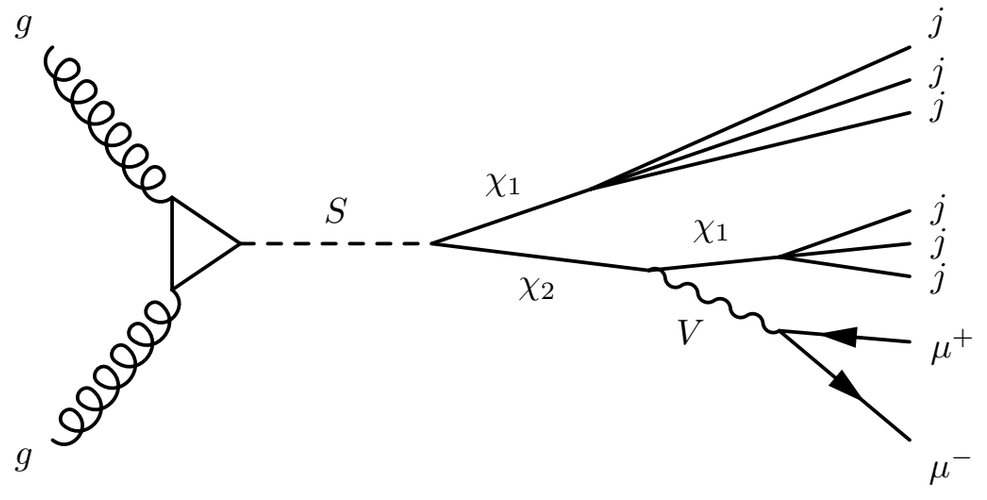
A dimuon resonance for the 2020s? (In addition to H → μ⁺μ⁻)

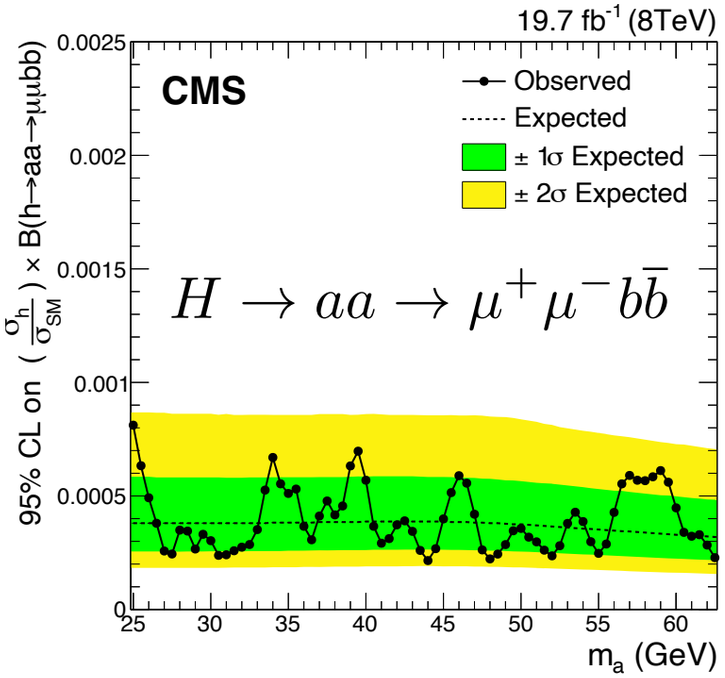
[LHC, PRL 2018]

Giving Dimuons a (Lorentz) Boost



E.g. Portal to a Hidden Sector



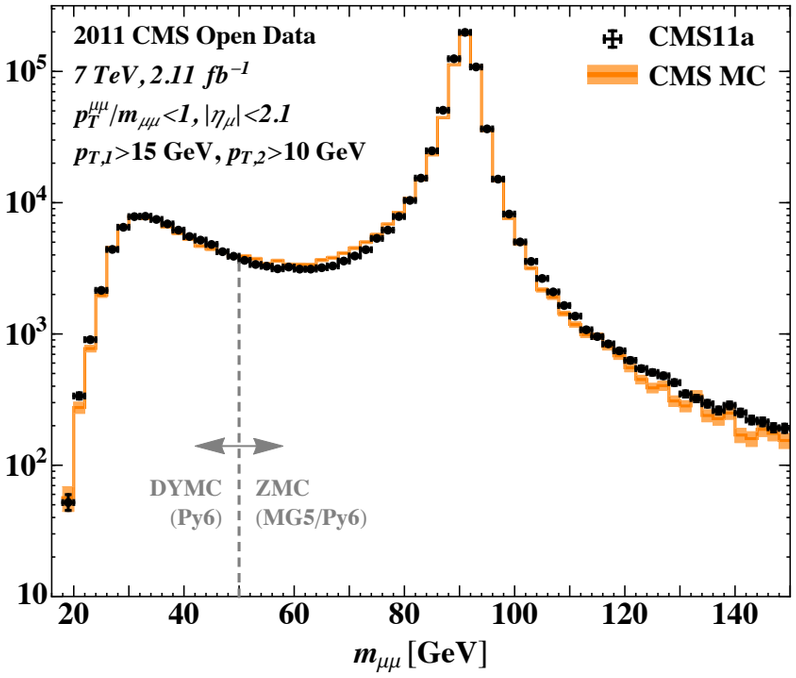


CMS has searches for boosted $\mu^+\mu^-$ plus *specific* recoil objects

No search (to our knowledge) for *generic* boosted $\mu^+\mu^-$ signature

[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019;
motivated by Strassler, Zurek, PLB 2007; limit from CMS, JHEP 2017]

On the Road to Discovery?



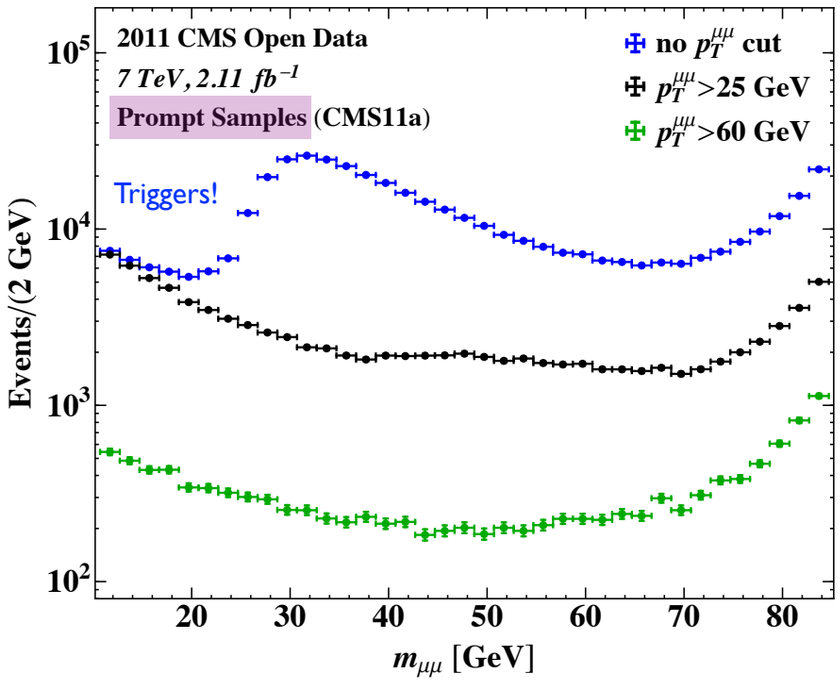
(Re)measuring the Z Boson

Our Result: σ = 974 ± 52 pb

Systematics limited by luminosity, acceptance, isolation efficiency, and background uncertainties

CMS 2011: σ = 986 ± 31 pb

NNLO: σ = 970 ± 30 pb



Reaching into the Unknown

CMS tight muon selection*

Prompt (or isolated) muon requirement

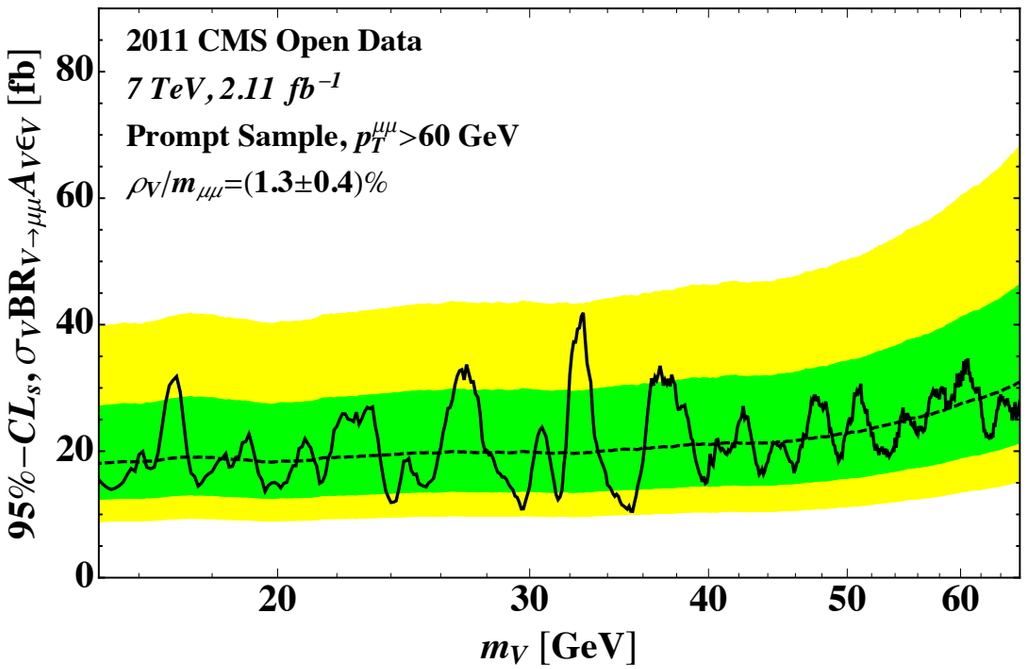
Dimuon transverse momentum cut

[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019; cf. CMS, JHEP 2011; CMS, JHEP 2013; Gavin, Li, Petriello, Quackenbush, CPC 2011]

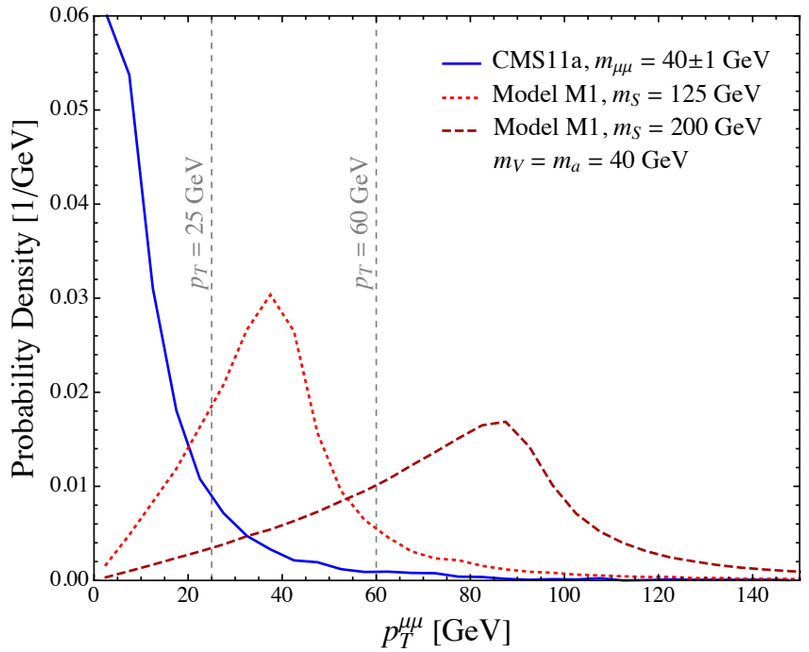
Pushing to the Limit

"Facilitating dialogue between theory and experiment"

We didn't discovery anything…



… but we could have!



Derived new limits, which could be an order of magnitude stronger using LHC Run 2 data!

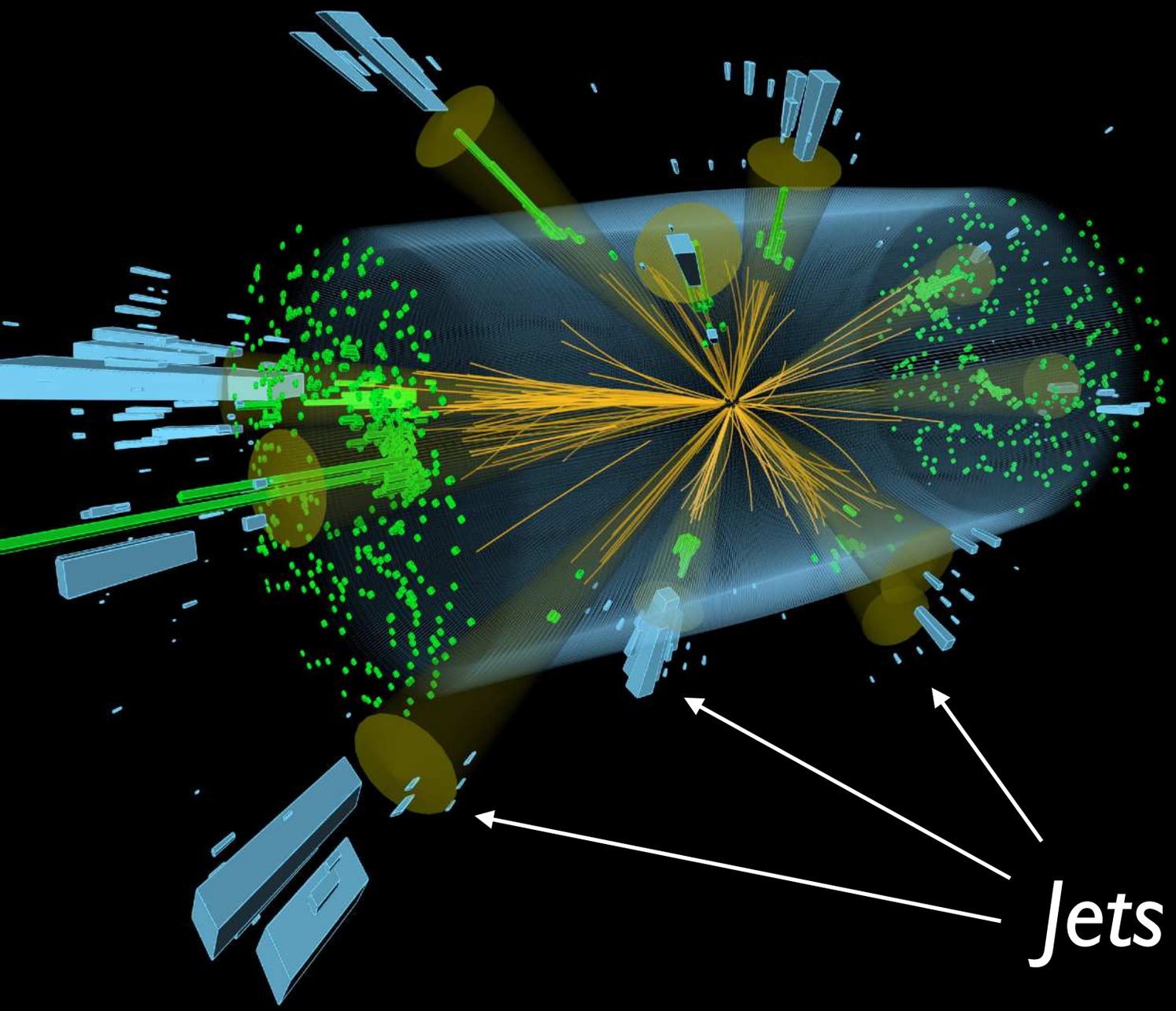
"Although searches using current open data are unlikely to uncover BSM phenomena on their own, they can help demonstrate the value of certain search strategies and justify the application of those strategies by the experimental collaborations on much larger data sets."

[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019]

Preliminary: A New View on QCD

One Collider Event

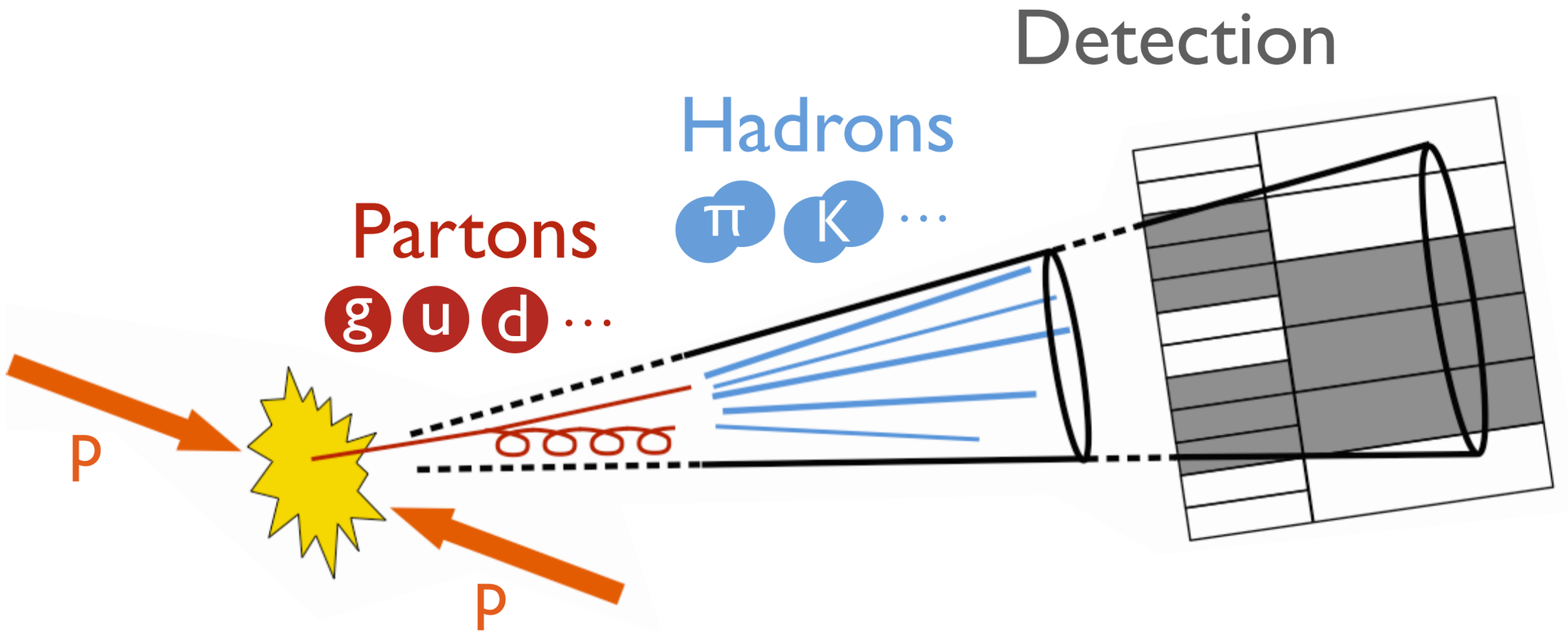
Collection of points in momentum space



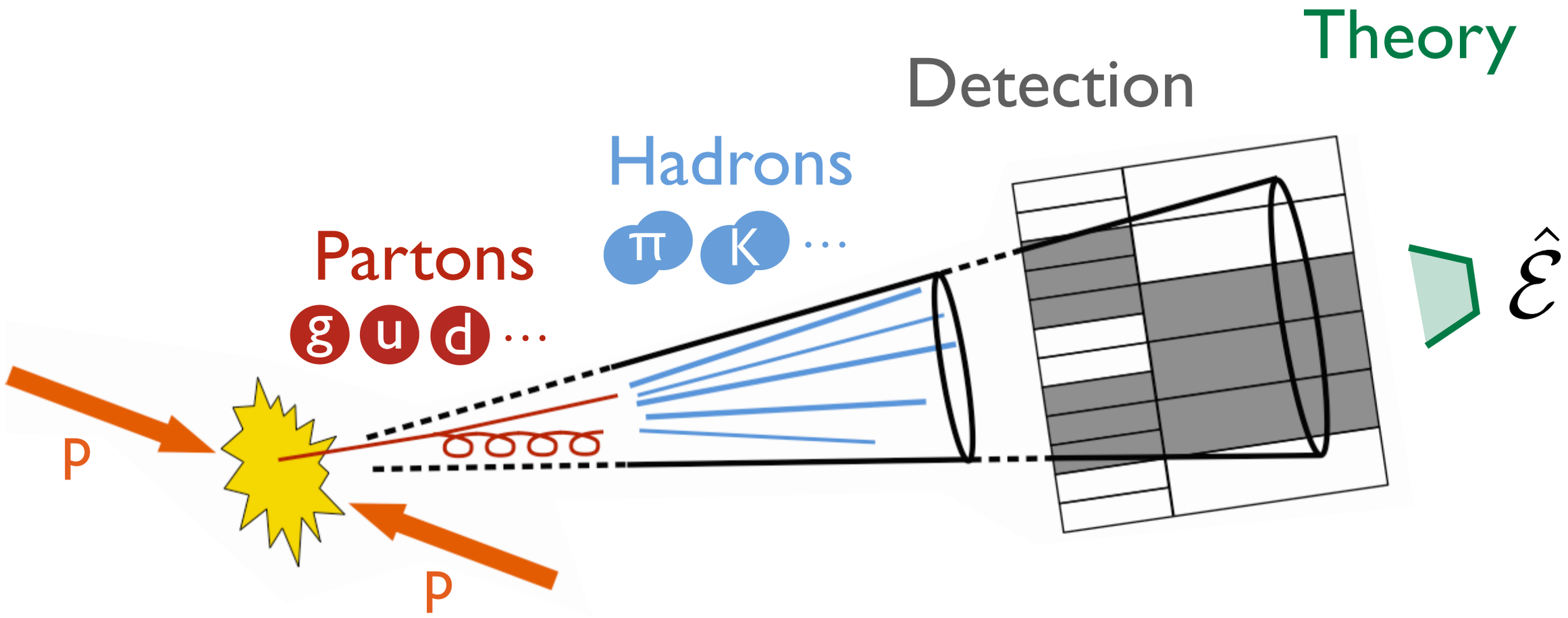
T E H M

γ	photon	elementary
$e^{\pm}$	electron	elementary
$\mu^{\pm}$	muon	elementary
$\pi^{\pm}$	pion	composite
$K^{\pm}$	kaon	composite
K^0_L	K-long	composite
$p/\overline{p}$	proton	composite
$n/\overline{n}$	neutron	composite

Dynamics of Jet Formation



Dynamics of Jet Formation



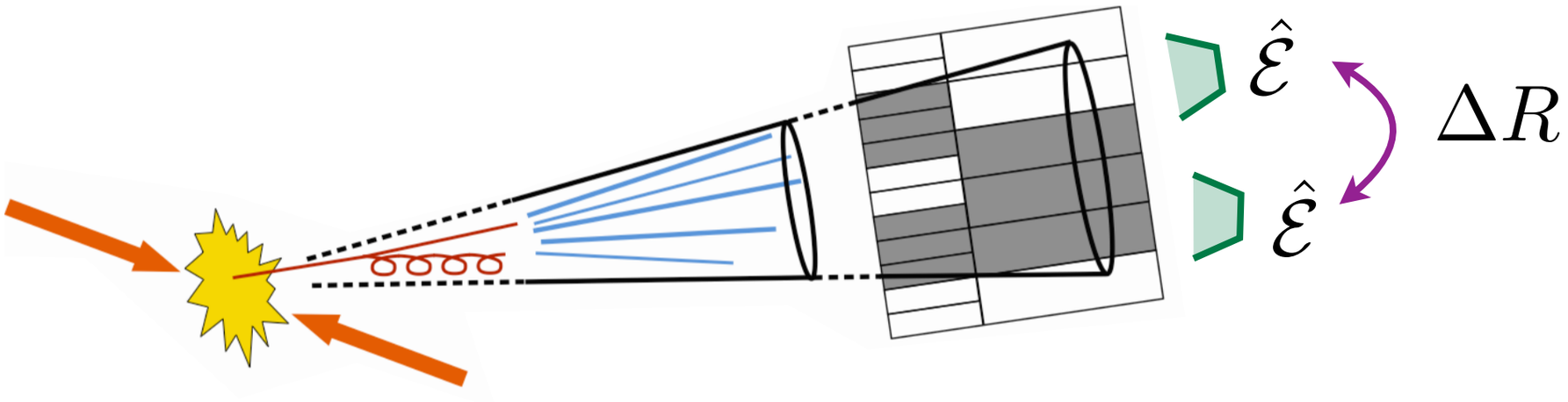
Energy Flow:

Robust to hadronization and detector effects

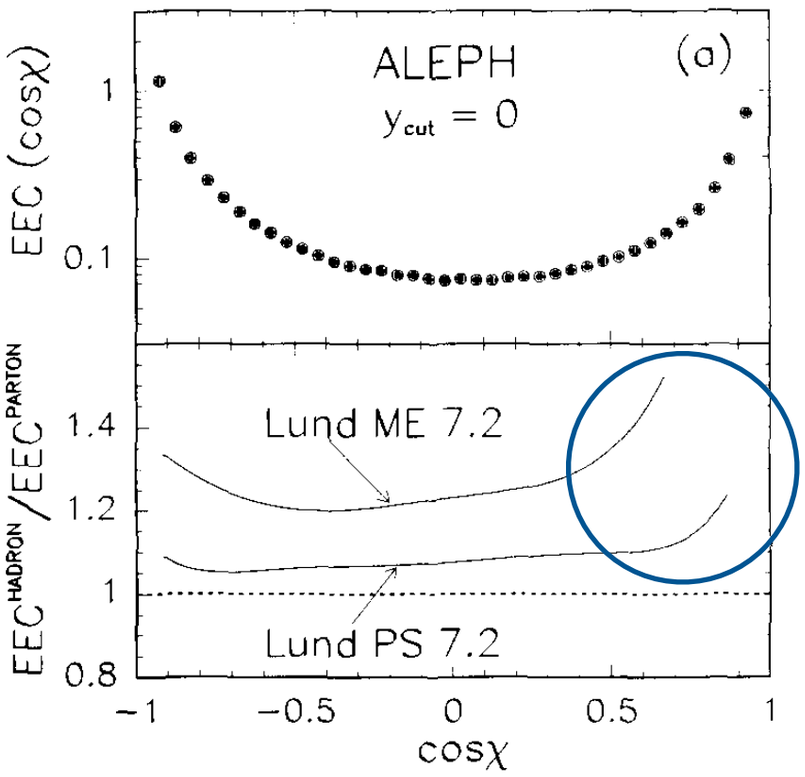
$$\hat{\mathcal{E}} \simeq \lim_{t\to\infty} \hat{n}_i T^{0i}(t, vt\hat{n})$$

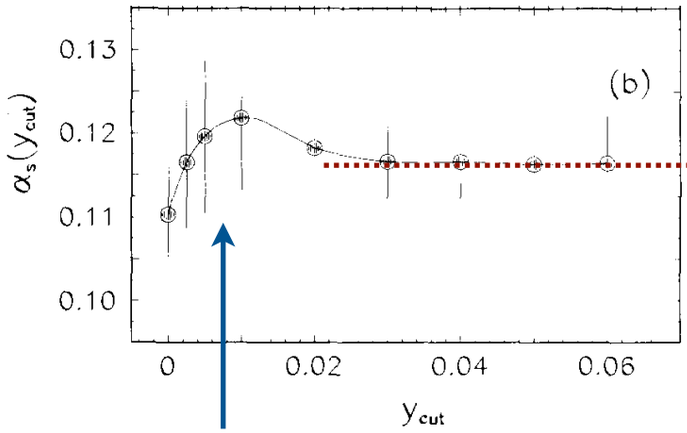
[see e.g. Sveshnikov, Tkachov, PLB 1996; Hofman, Maldacena, JHEP 2008; Mateu, Stewart, JDT, PRD 2013; Belitsky, Hohenegger, Korchemsky, Sokatchev, Zhiboedov, PRL 2014; Chen, Moult, Zhang, Zhu, PRD 2020]

Energy-Energy Correlators



A long history in probing collinear dynamics of QCD





Extracting the strong coupling constant

Theoretical challenges with small angle (collinear) limit

[Basham, Brown, Ellis, Love, PRL 1978; ALEPH, PLB 1991; see Chen, Moult, Zhang, Zhu, PRD 2020]

“Try it on the open data!”

Jesse Thaler July 24, 2019 at 2:01 PM

EE(E)C?

To: Ian Moult

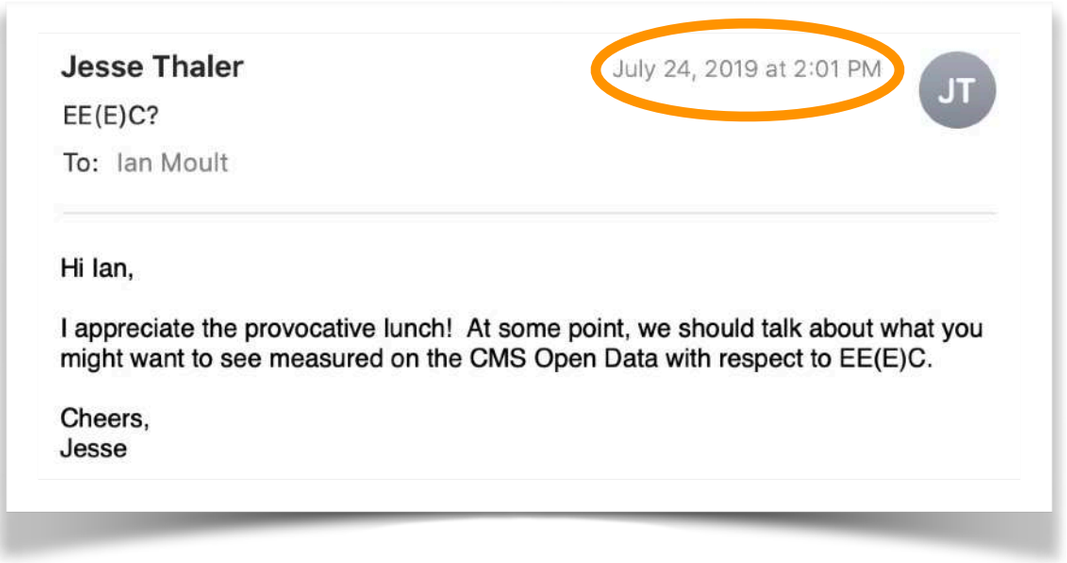
Hi Ian,

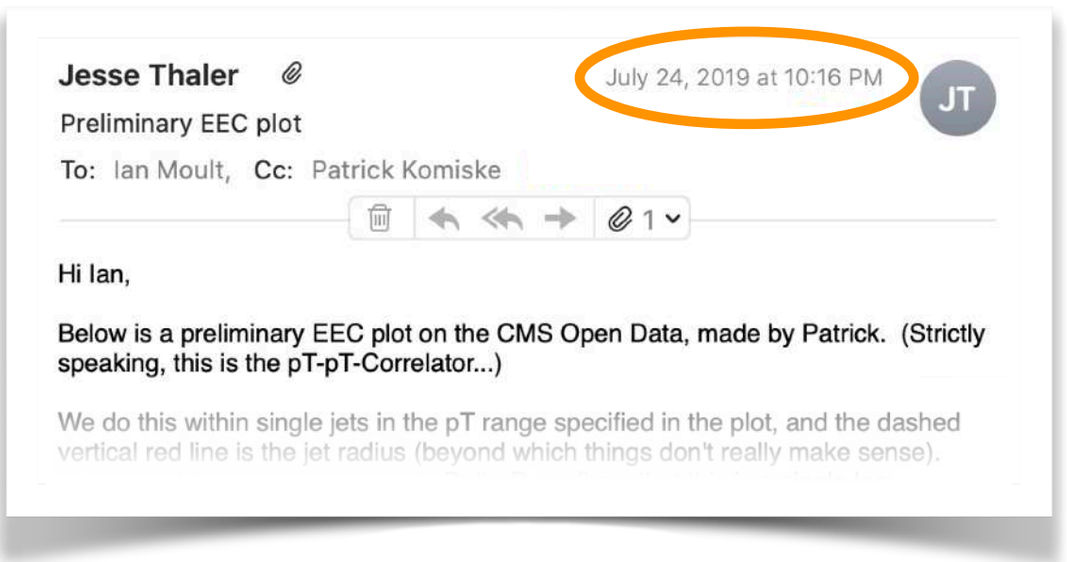
I appreciate the provocative lunch! At some point, we should talk about what you might want to see measured on the CMS Open Data with respect to EE(E)C.

Cheers,
Jesse

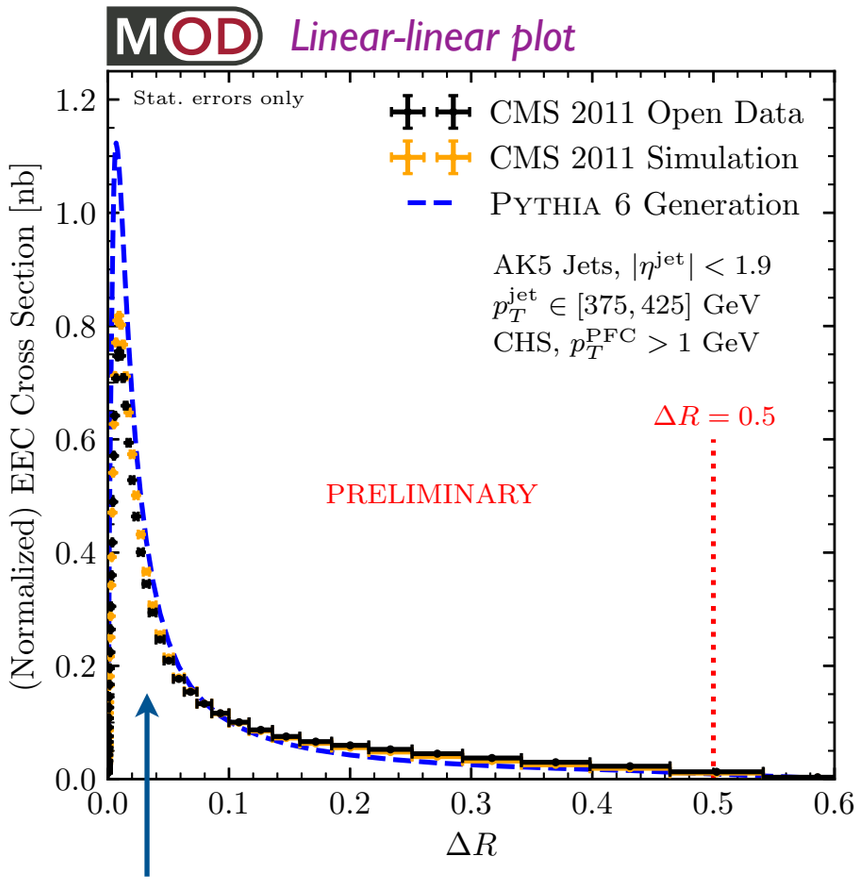
[Komiske, Moult, JDT, et al., in progress; see talk by Moult, BOOST 2019]

“Try it on the open data!”





First Jet EEC Plot from the LHC (!)

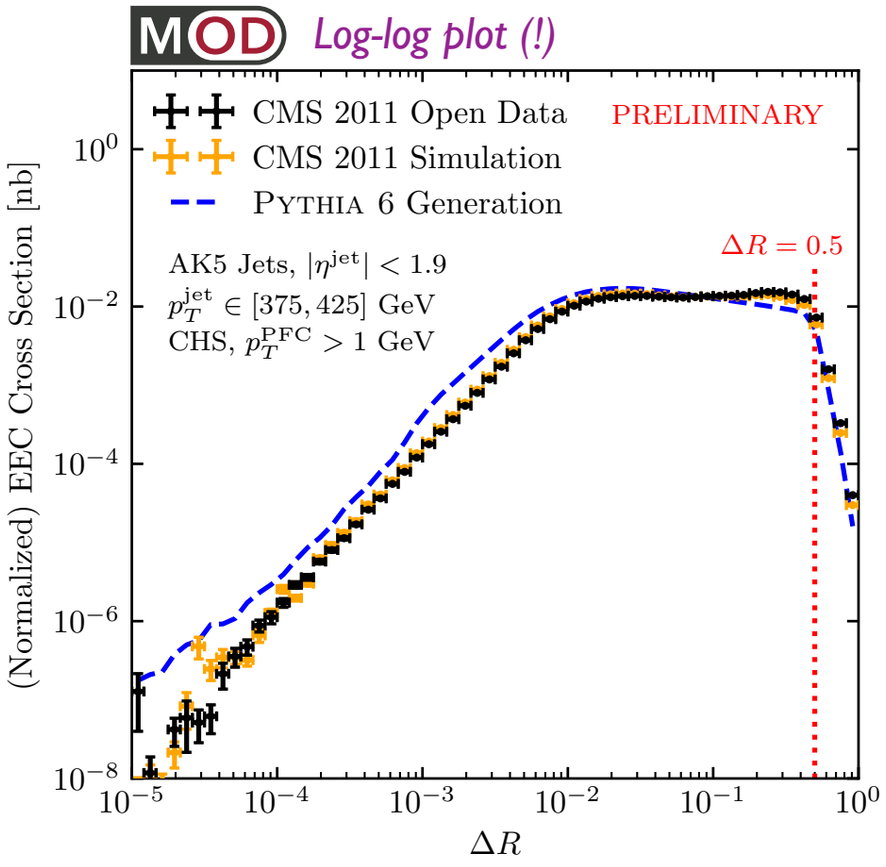


Are we learning something about small angle limit of QCD?

[Komiske, Moult, JDT, et al., in progress; see talk by Moult, BOOST 2019]

QCD Phase Transition in Jets?

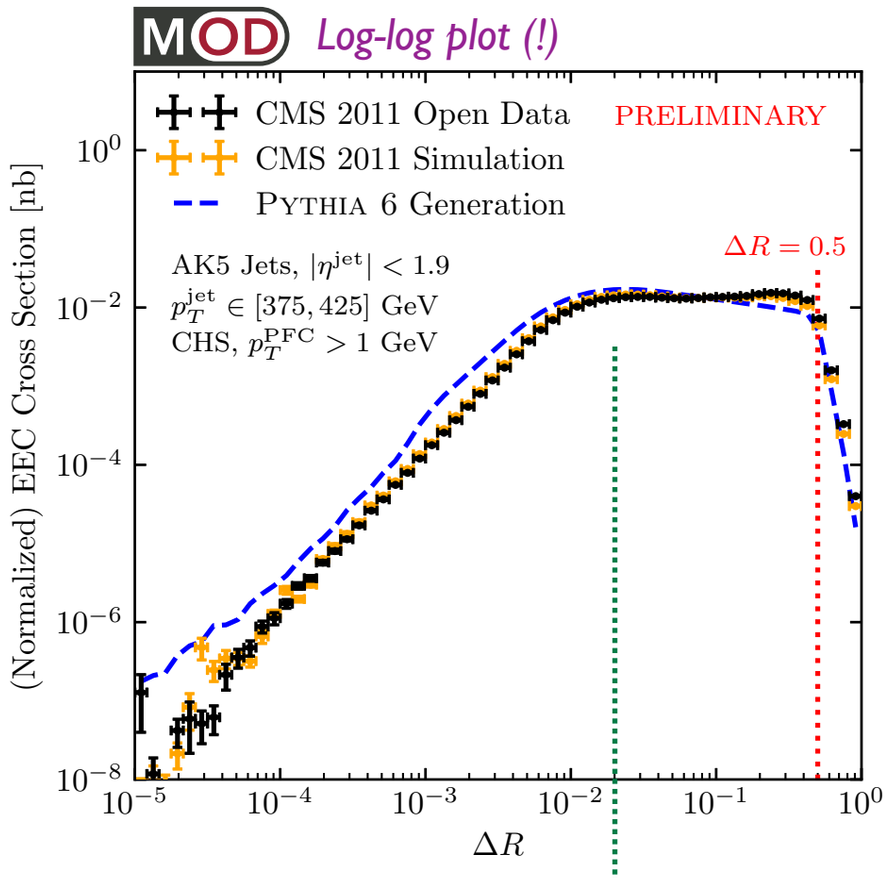
"Enabling exploratory/ proof-of-principle studies"

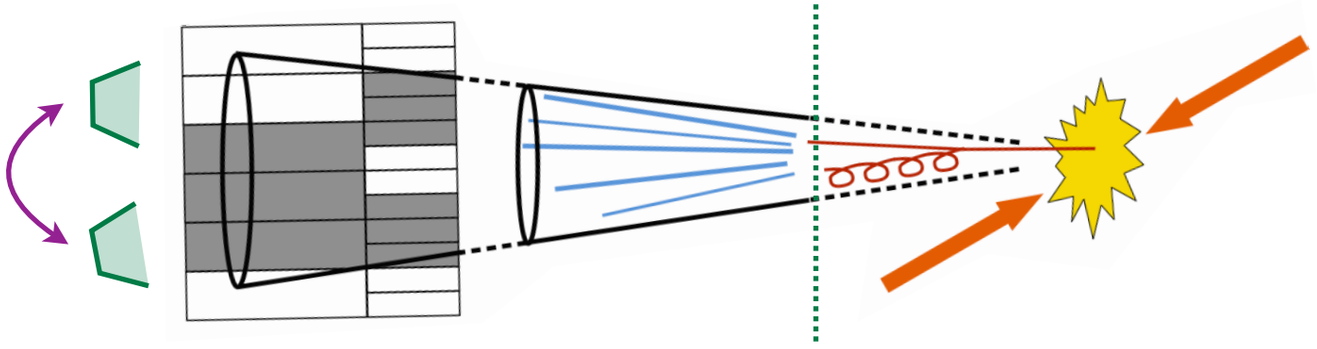


[Komiske, Moult, JDT, et al., in progress; see talk by Moult, BOOST 2020]

QCD Phase Transition in Jets?

"Enabling exploratory/ proof-of-principle studies"





Leveraging complementary perspectives

Theory: EEC probes different physics from standard jet observables

Data Analysis: Scaling behavior transition only apparent in log-log space

Experiment: Impressive angular resolution from CMS tracking system

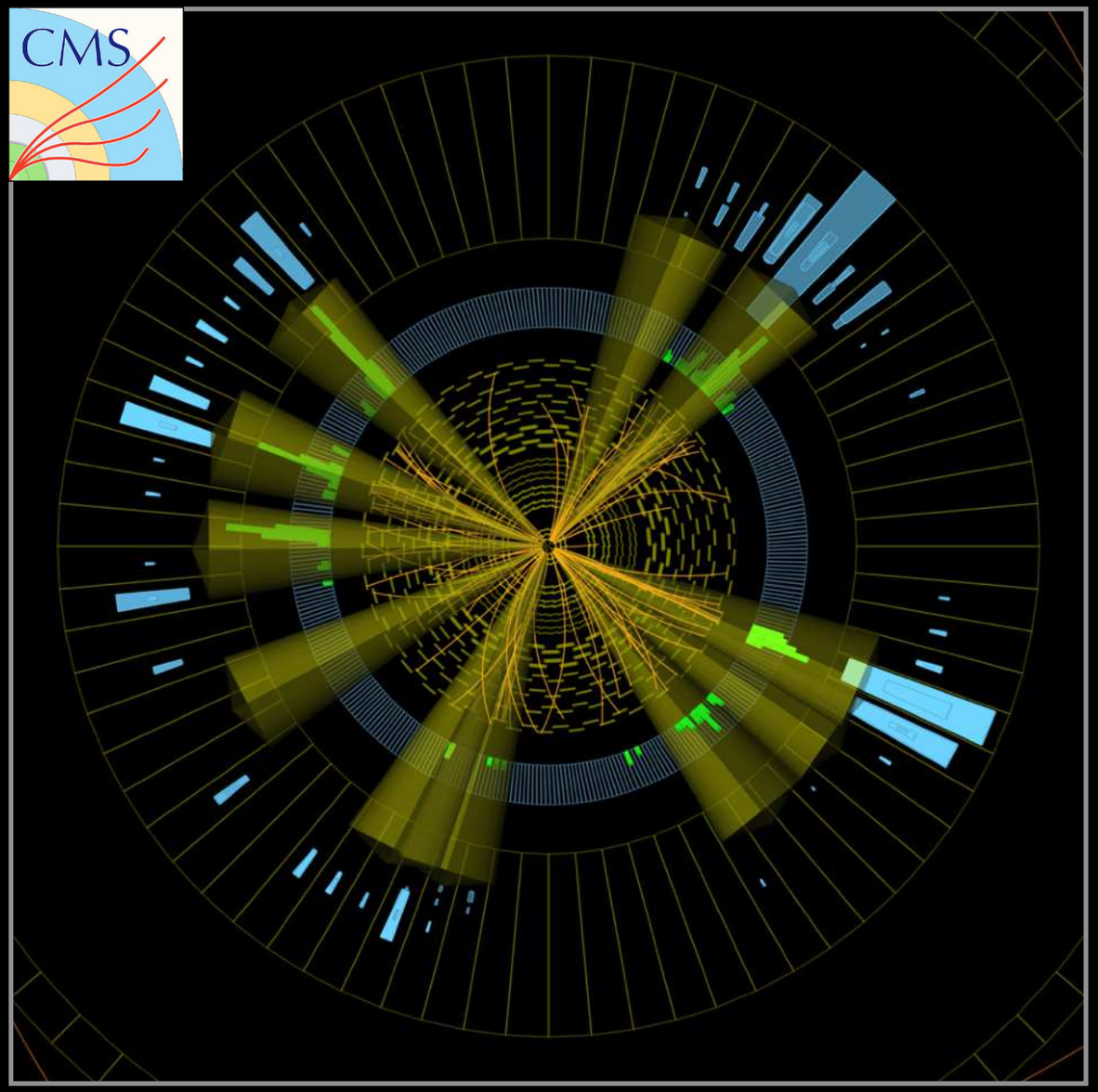
"I doubt that anyone can envision the result before seeing the open data plot."
— Not Emmy Noether

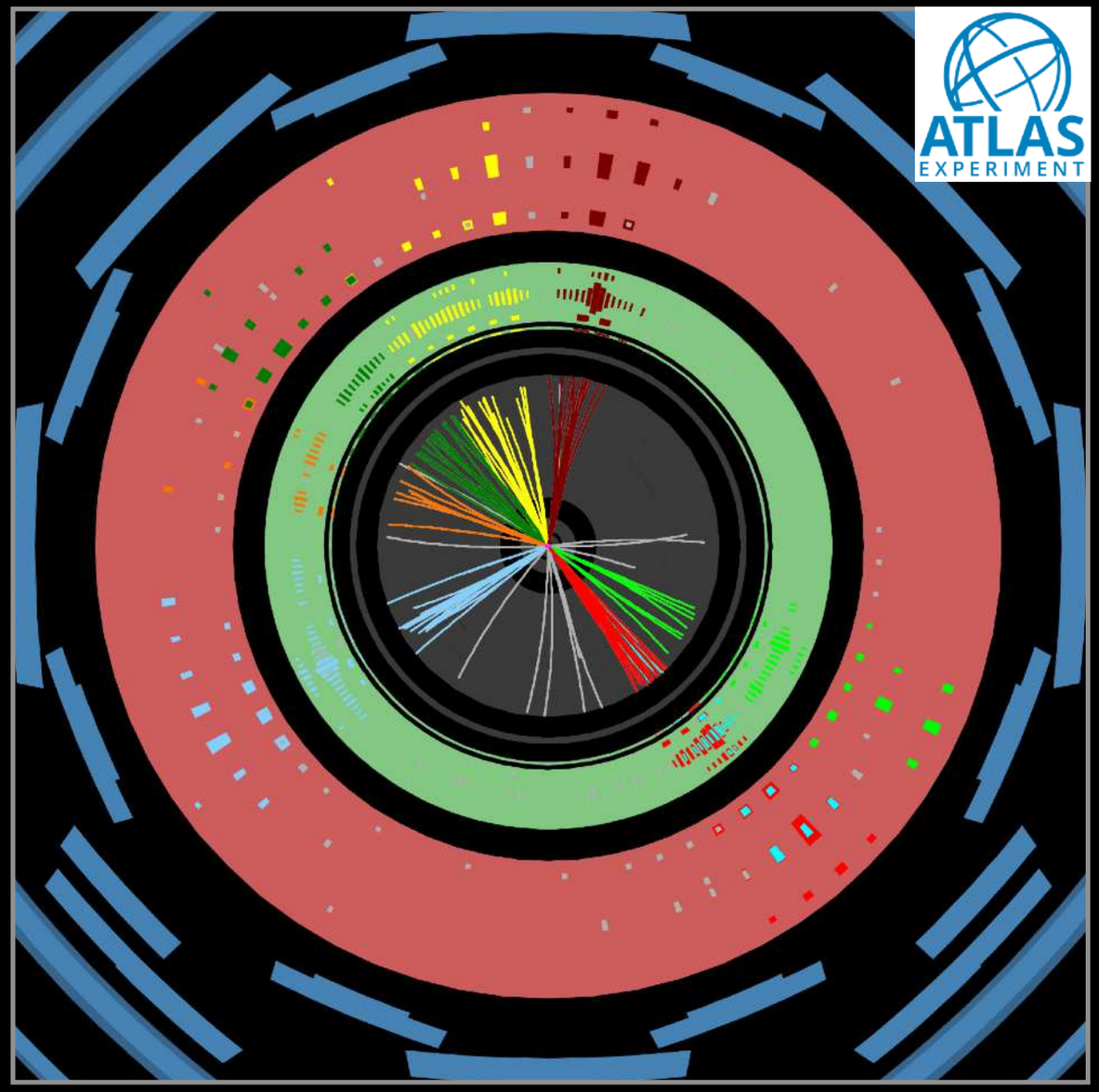
[Komiske, Moult, JDT, et al., in progress; see talk by Moult, BOOST 2020]

Event Space Geometry

Two Collider Events

Two collections of points in momentum space

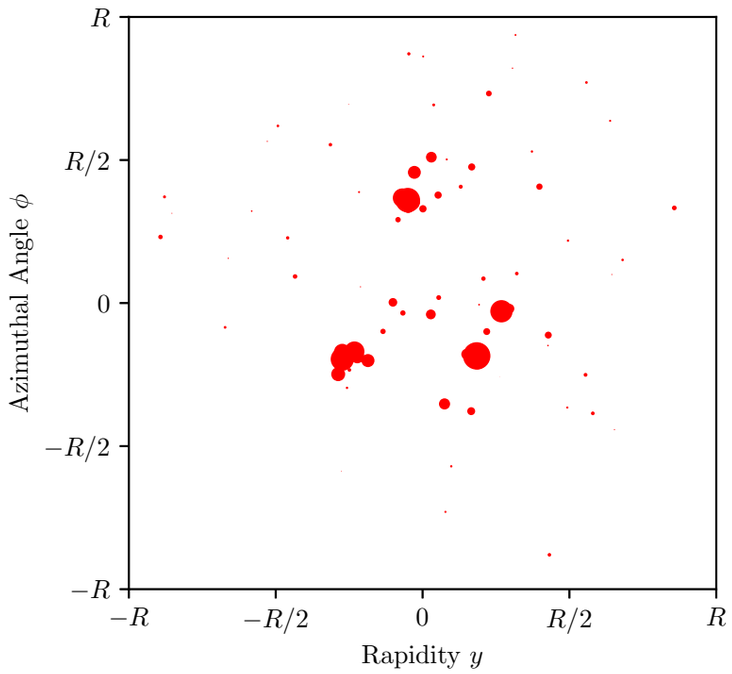


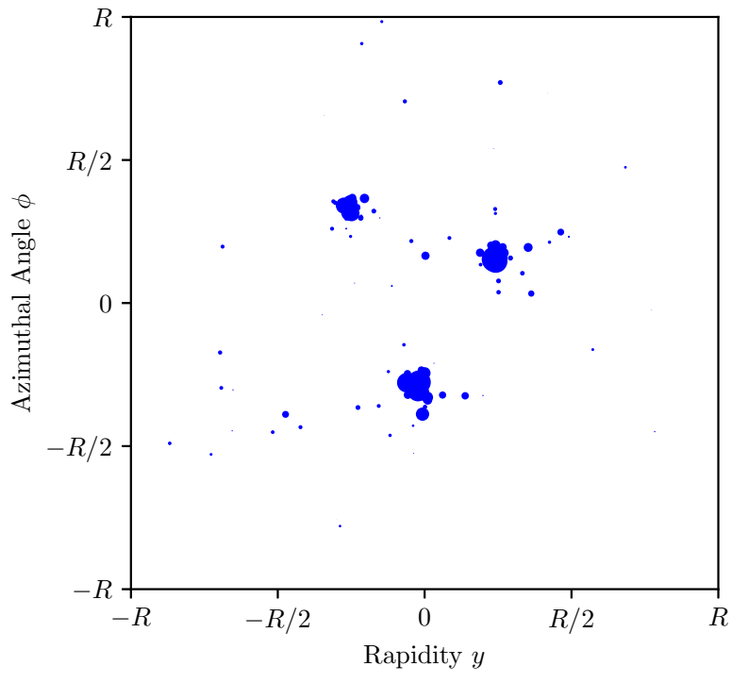


How "close" are these? (8.5 km?)

Similarity of Two Jets?

$$\mathcal{E}(\hat{n}) = \sum_i E_i \, \delta(\hat{n} - \hat{n}_i)$$

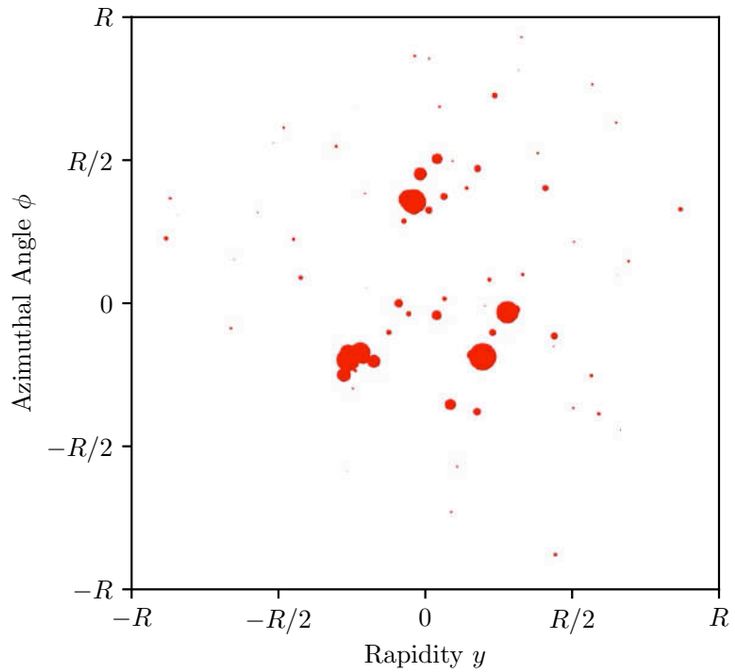




[Komiske, Metodiev, JDT, PRL 2019; code at Komiske, Metodiev, JDT, energyflow.network]

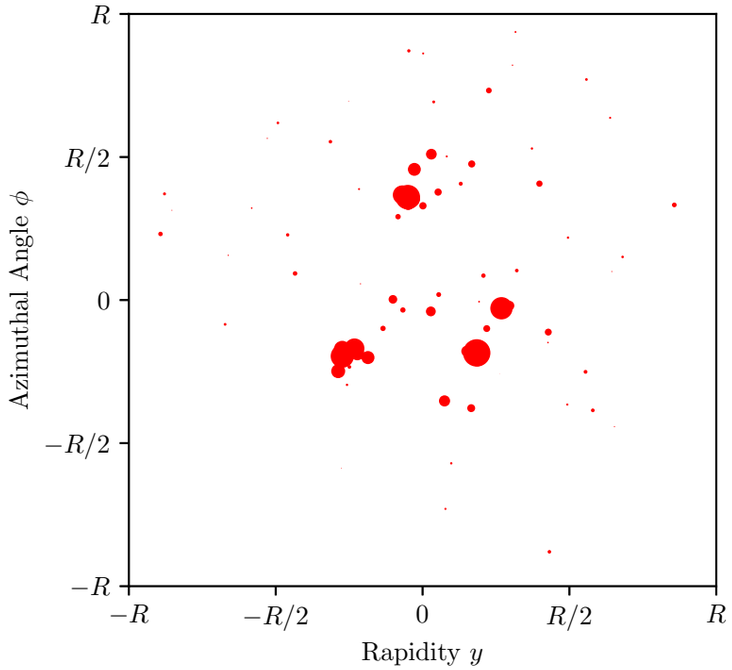
Similarity of Two Jets?

$$\mathcal{E}(\hat{n}) = \sum_i E_i \, \delta(\hat{n} - \hat{n}_i)$$

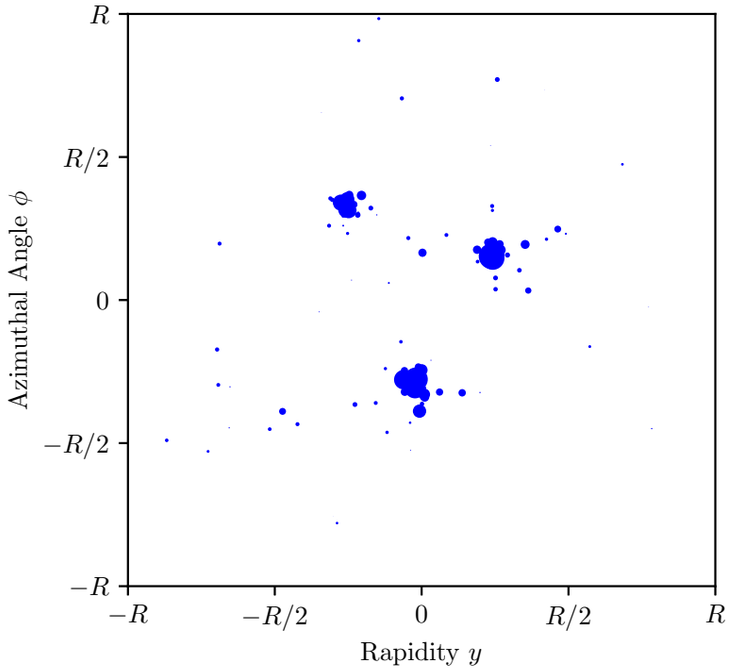


Optimal Transport:

Earth Mover's Distance
a.k.a. 1-Wasserstein metric



125.4 GeV

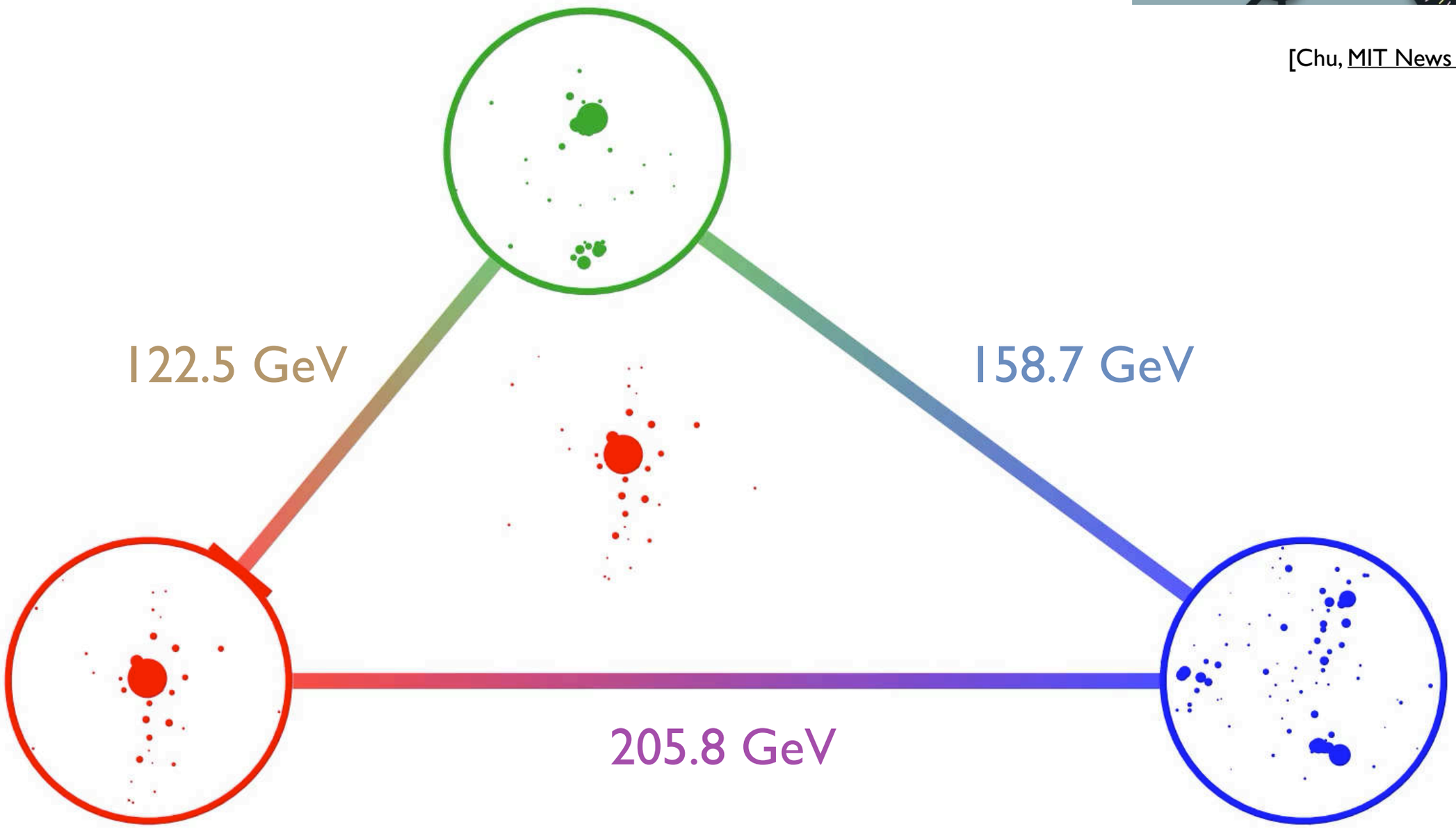


[Komiske, Metodiev, JDT, PRL 2019; code at Komiske, Metodiev, JDT, energyflow.network]

Triangulating the Space of Jets

Three jets from the CMS Open Data

[Chu, MIT News July 2019]

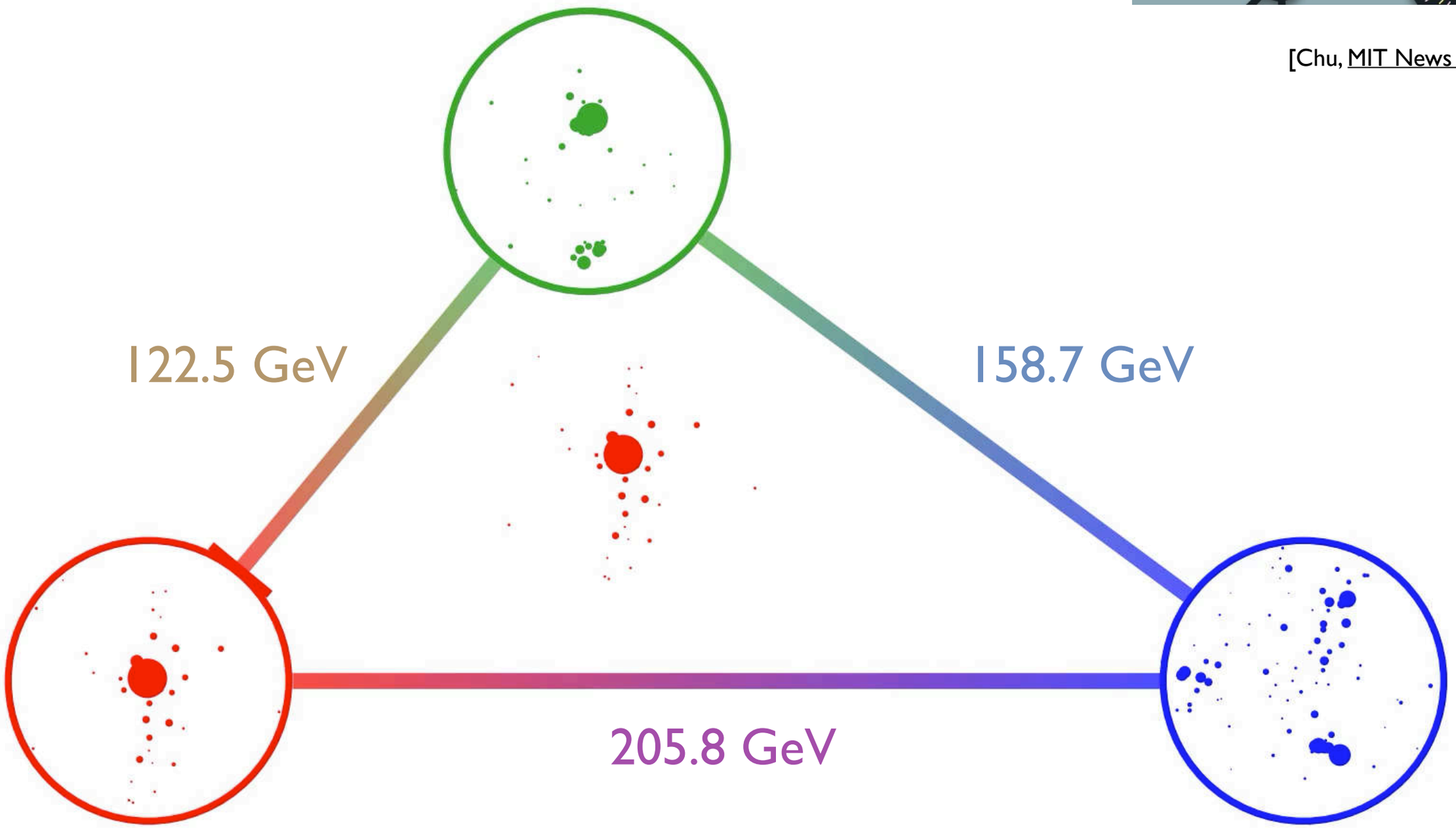


[Komiske, Metodiev, JDT, PRL 2019; code at Komiske, Metodiev, JDT, energyflow.network]

Triangulating the Space of Jets

Three jets from the CMS Open Data

[Chu, MIT News July 2019]



[Komiske, Metodiev, JDT, PRL 2019; code at Komiske, Metodiev, JDT, energyflow.network]

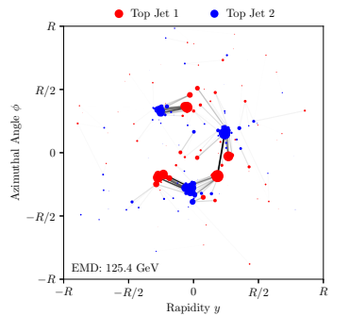
What can you do with a metric?

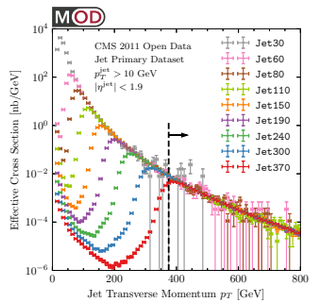
"Facilitating curiosity-driven research"

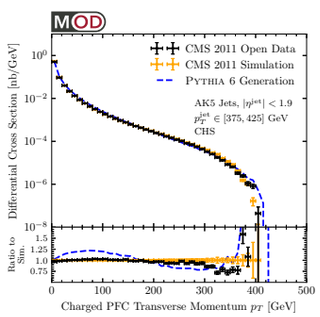
And 1,690,984 jets from the CMS Open Data?

$$\mathrm{EMD}(\mathcal{E}, \mathcal{E}') = \min_{\{f\}} \sum_i \sum_j f_{ij} \frac{\theta_{ij}}{R} + \left| \sum_i E_i - \sum_j E'_j \right|$$

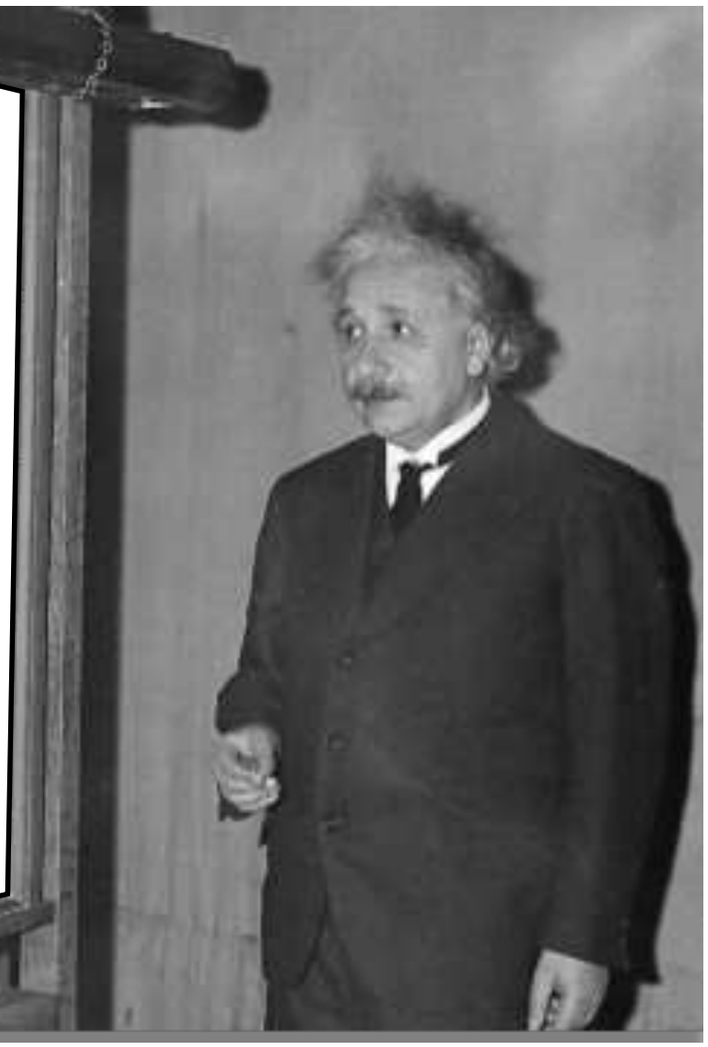
$$0 \leq \mathrm{EMD}(\mathcal{E}, \mathcal{E}') \leq \mathrm{EMD}(\mathcal{E}, \mathcal{E}'') + \mathrm{EMD}(\mathcal{E}', \mathcal{E}'')$$







Dann was?



Use our preprocessed jet samples!

DOI 10.5281/zenodo.3340205 - CMS 2011A Jets, pT > 375 GeV

[https://energyflow.network/docs/datasets/]

Dimensionality of Space of Jets

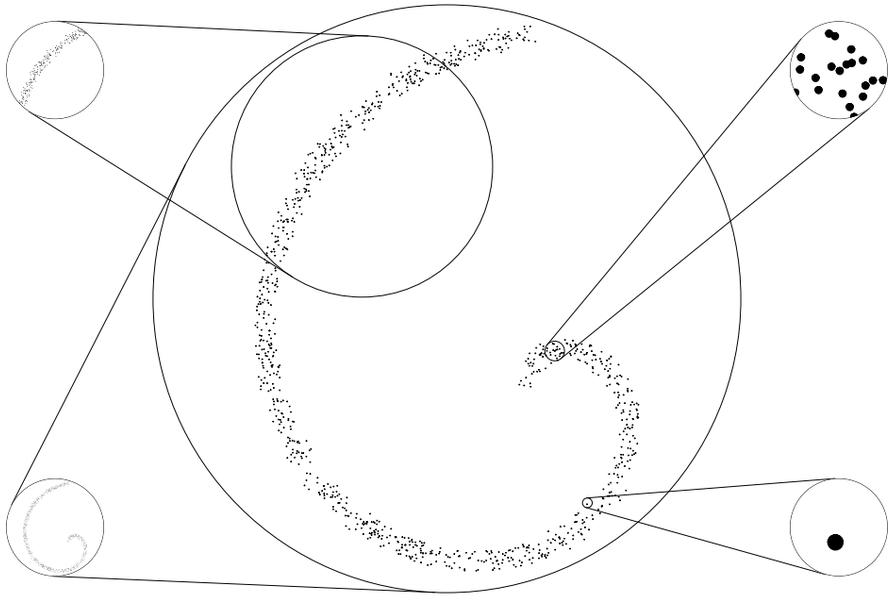
$$N_{\text{neighbors}}(r) \sim r^{\dim}$$

$$\Rightarrow \quad \dim(r) \sim r \frac{\partial}{\partial r} \ln N_{\text{neighbors}}(r)$$

[Grassberger, Procaccia, PRL 1983; Kégl, NIPS 2002]

dim≃1

dim≃2



dim≃2

(eventually 0)

dim≃0

Dimensionality of Space of Jets

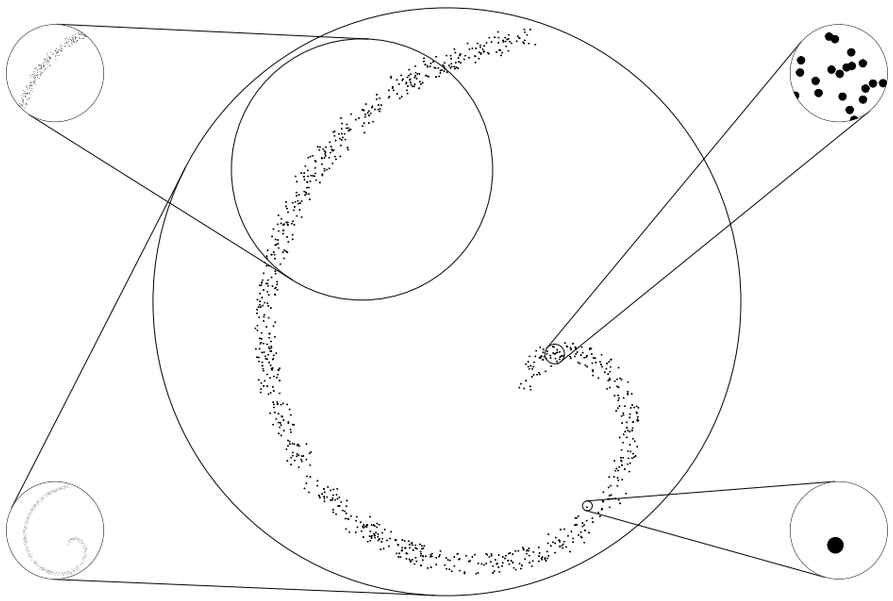
$$N_{\text{neighbors}}(r) \sim r^{\dim}$$

$$\Rightarrow \quad \dim(r) \sim r \frac{\partial}{\partial r} \ln N_{\text{neighbors}}(r)$$

[Grassberger, Procaccia, PRL 1983; Kégl, NIPS 2002]

dim≃1

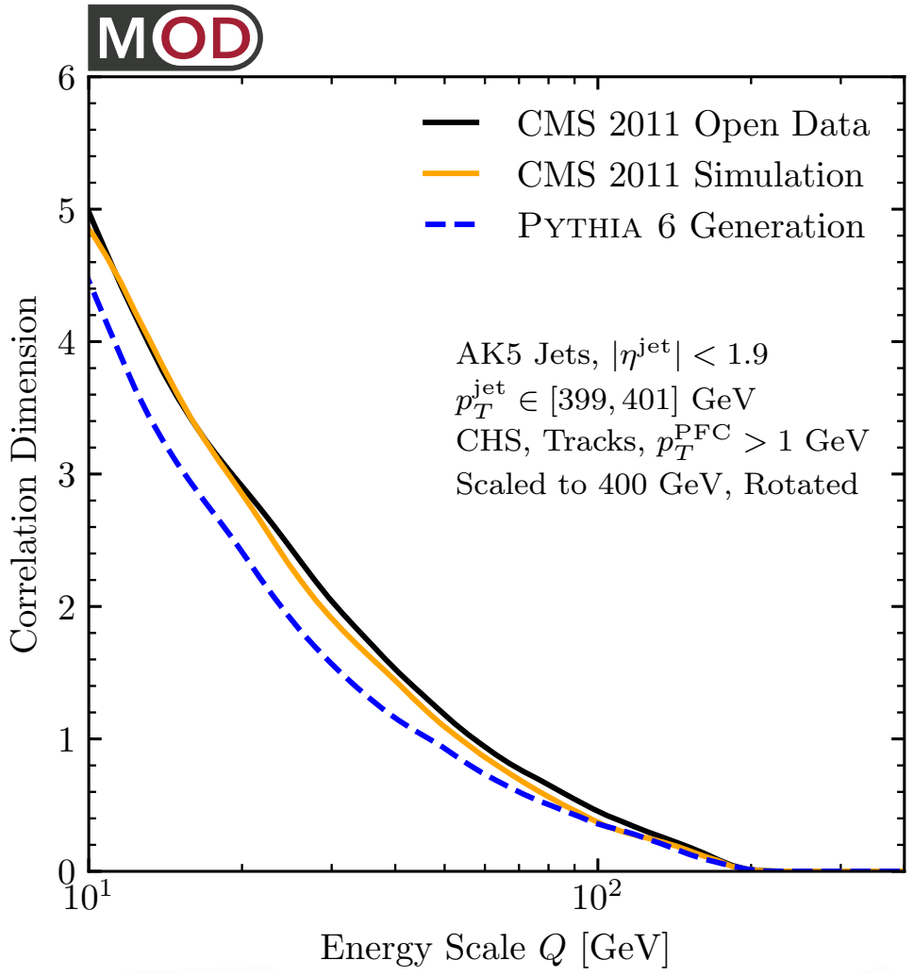
dim≃2

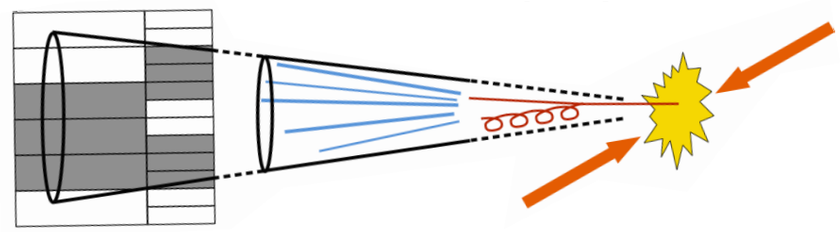


dim≃2

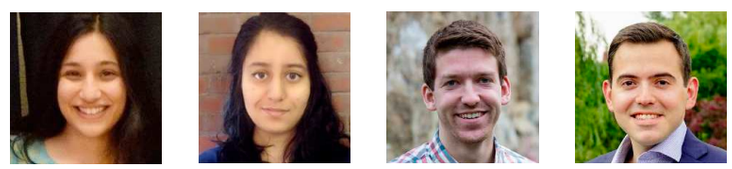
(eventually 0)

dim≃0

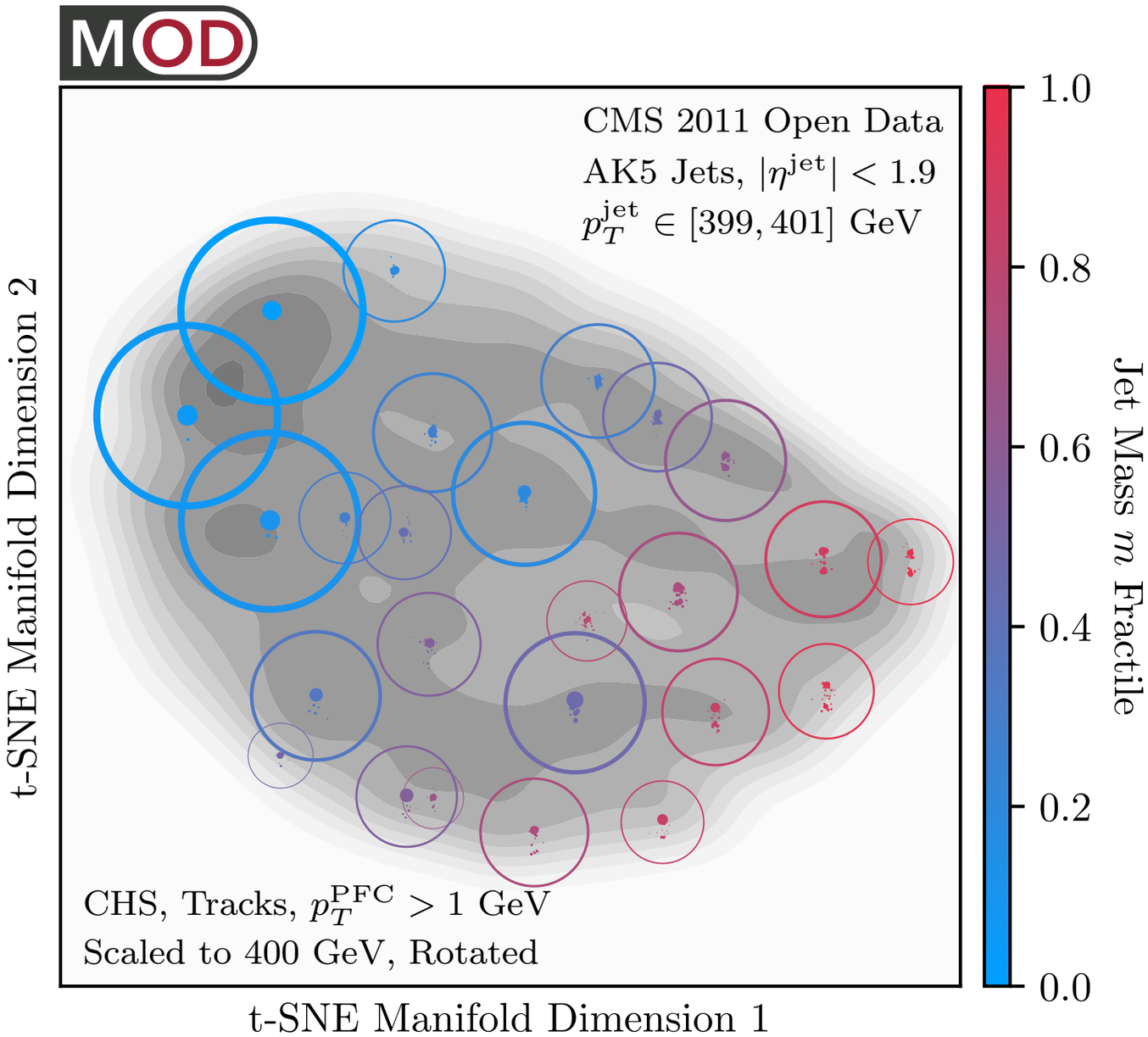


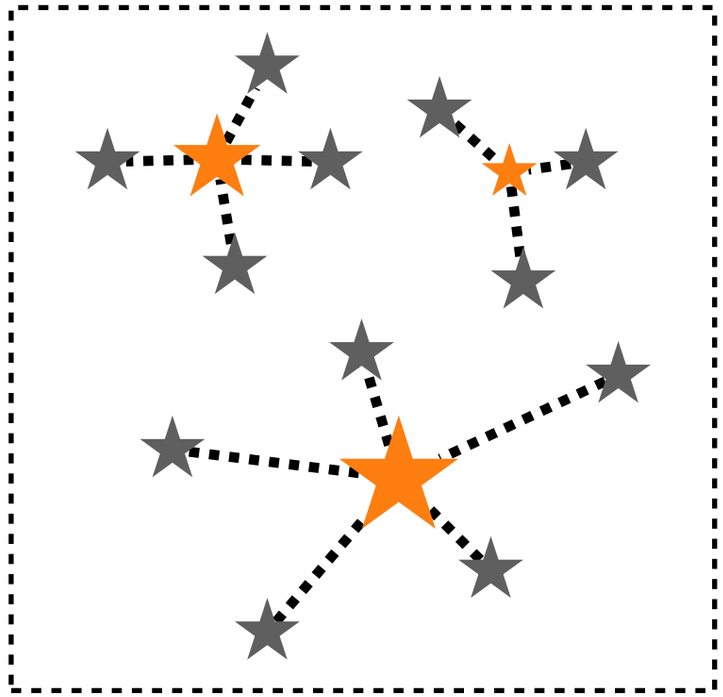


[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020]



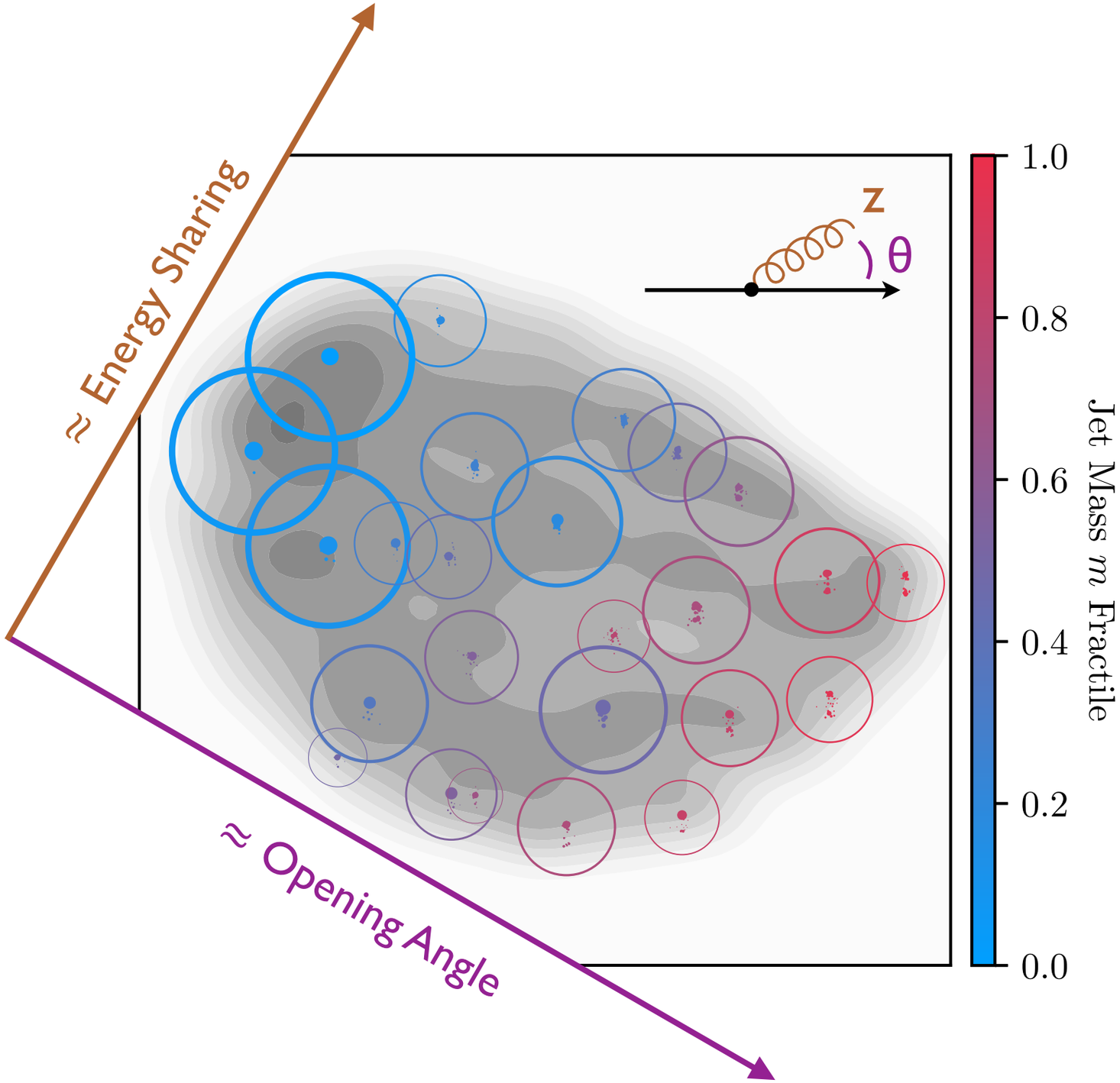
Most Representative Jets





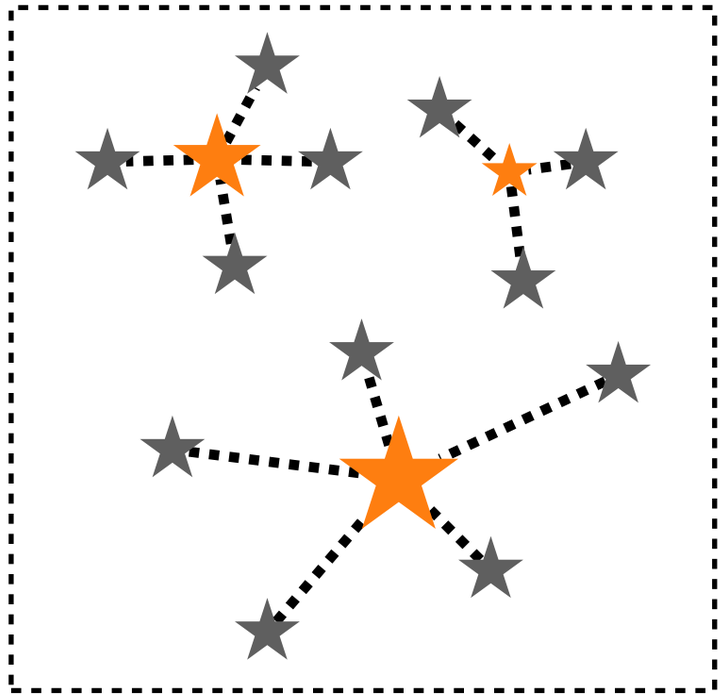
[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020; using van der Maaten, Hinton, JMLR 2008]

Most Representative Jets



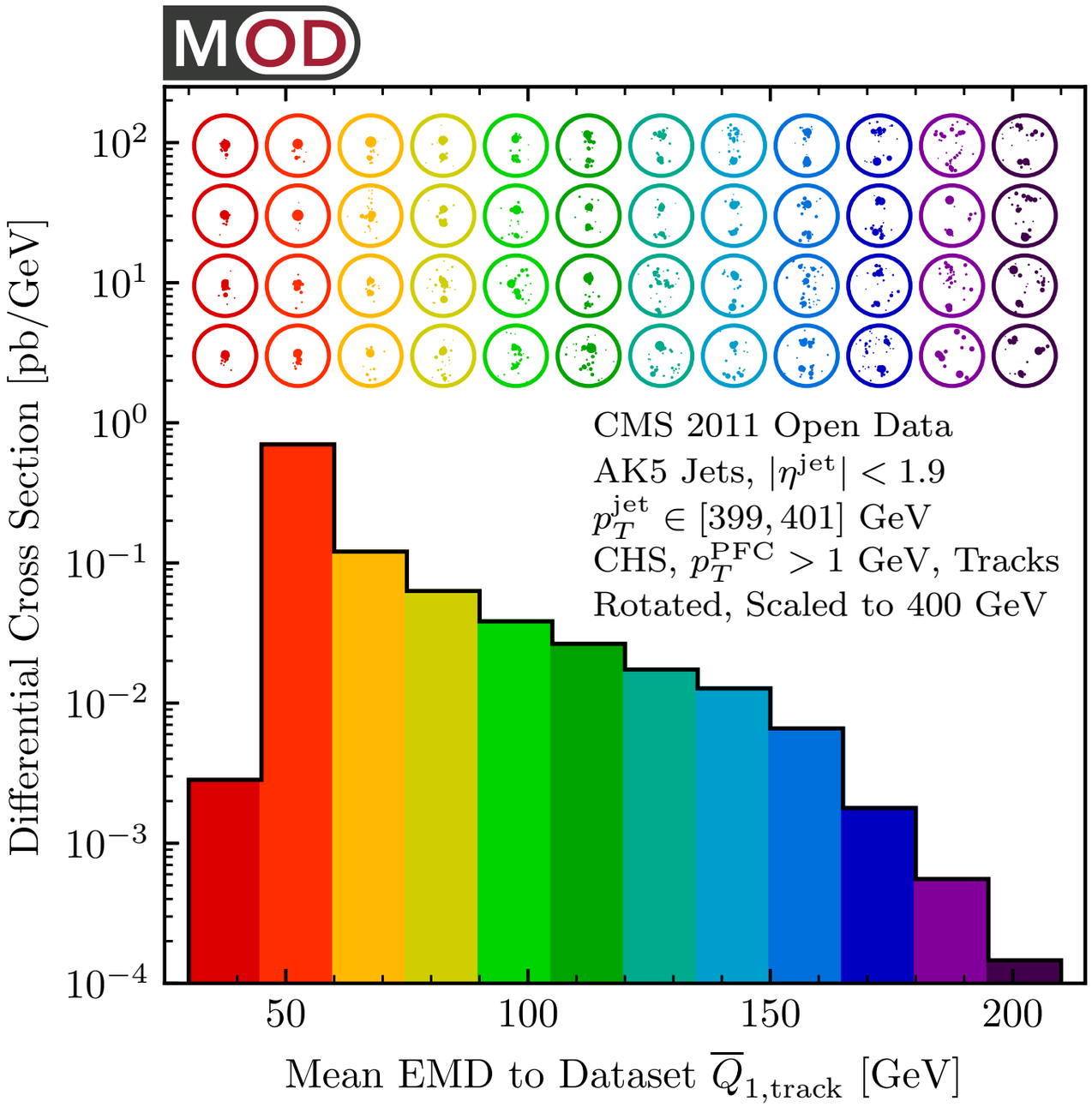
k-medoids

Arranged via t-SNE



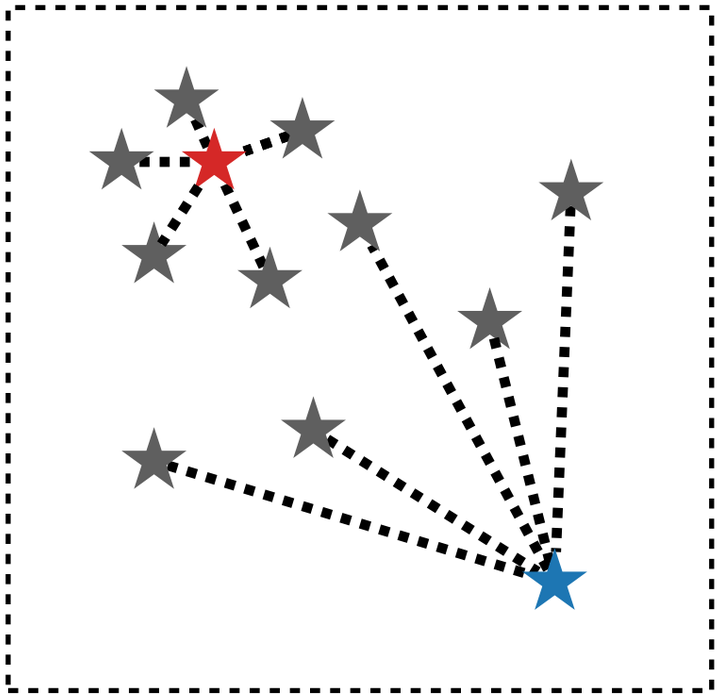
[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020; using van der Maaten, Hinton, JMLR 2008]

Least Representative Jets



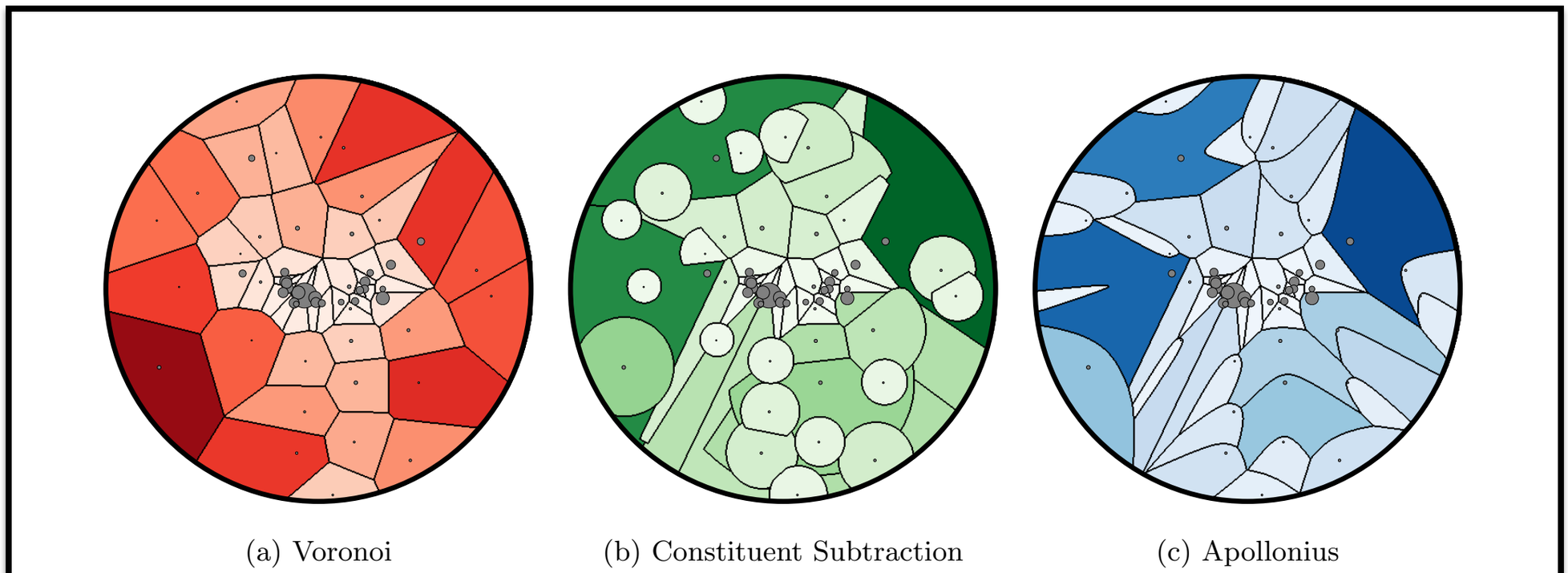
New Physics?

Or tails of QCD?



[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020]

When your theory paper just needs a jet…



(a) Voronoi (b) Constituent Subtraction (c) Apollonius

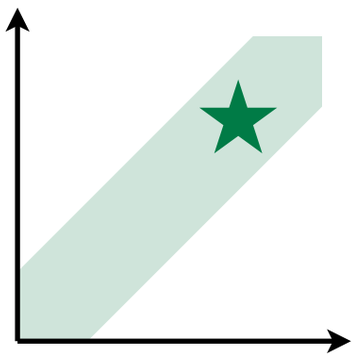
Figure 9: Three pileup subtraction procedures shown for a jet from the 2011A CMS Jet Primary Dataset. The jet constituents are shown as gray disks at their locations in the rapidity-azimuth plane with sizes proportional to their transverse momenta. The boundary is at a distance $R = 0.5$ from the jet axis at the center. The color intensity of each region

Use our preprocessed jet samples! DOI 10.5281/zenodo.3340205 - CMS 2011A Jets, pT > 375 GeV

[Komiske, Metodiev, JDT, JHEP 2020;
using https://energyflow.network/docs/datasets/]

Introducing the CMS Open Data

Adventures with Public Collider Data



Looking to the Past, Present, and Future

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.

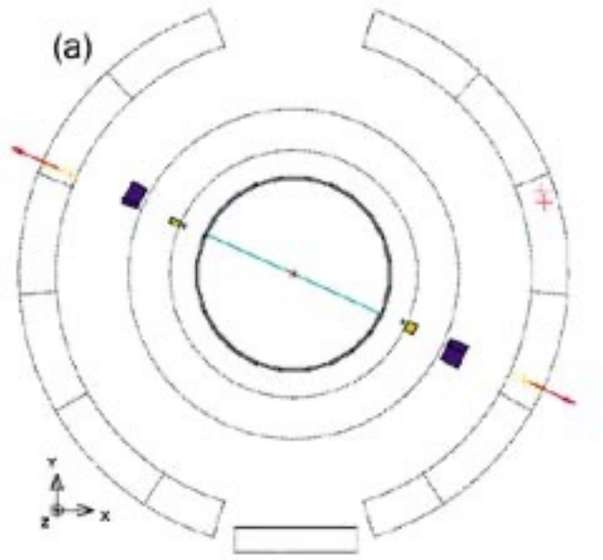
[Tripathee, Xue, Larkoski, Marzani, JDT, PRD 2017]

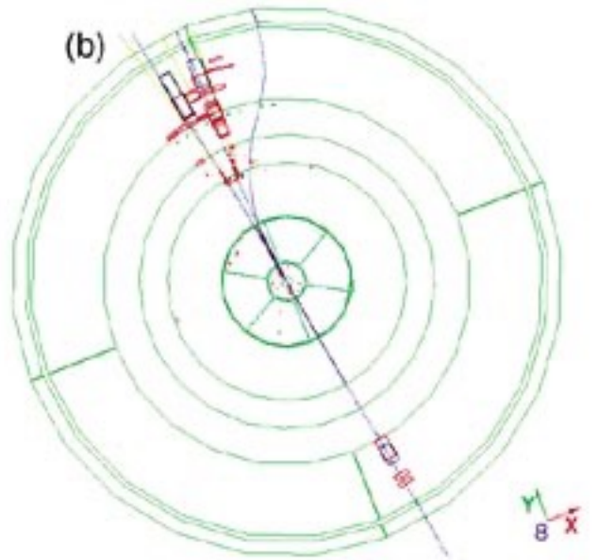
Looking to the Past…

"Stress-testing archival data strategies"

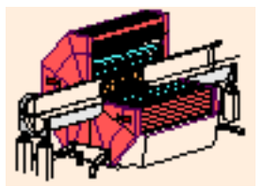
Large Electron Positron Collider (LEP)

OPAL

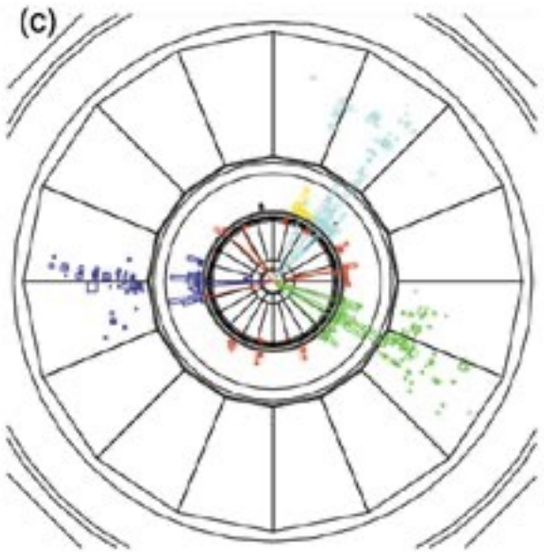


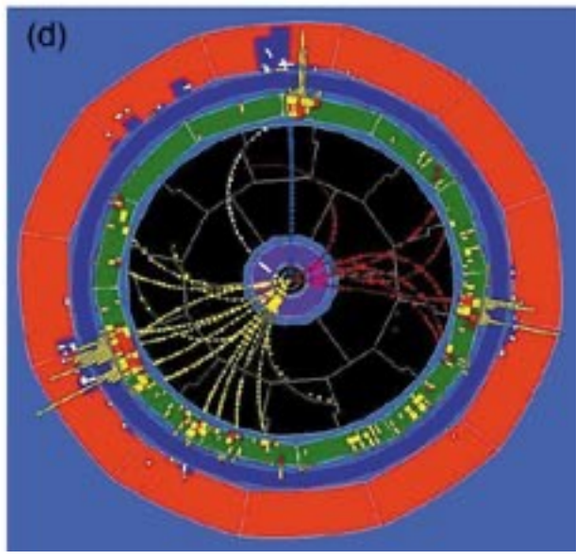


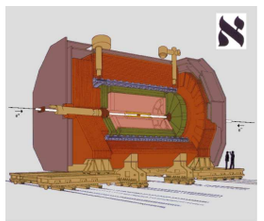
DELPHI



L3

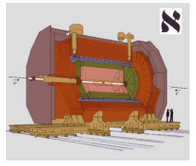






ALEPH

ALEPH Hunt for Exotic Higgs Decay



JOSHUA BATSON SCIENCE 01.23.15 6:45 AM

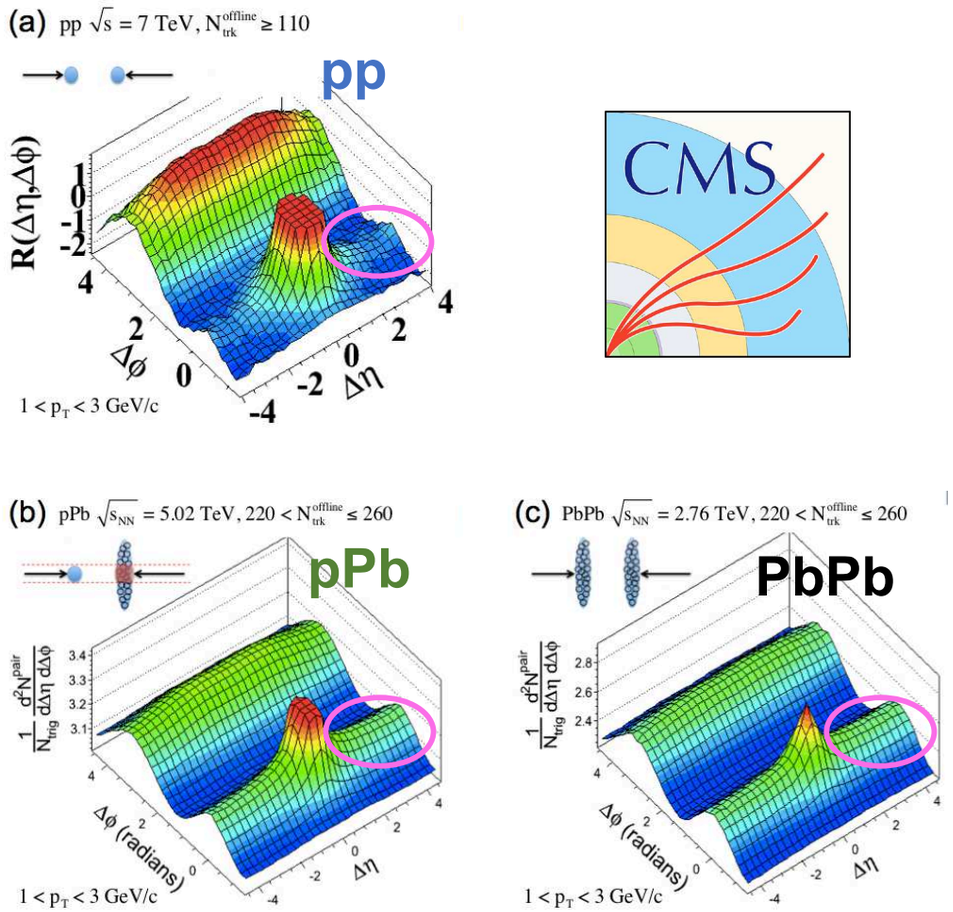
HOW THREE GUYS WITH $10K AND DECADES-OLD DATA ALMOST FOUND THE HIGGS BOSON FIRST

"Kyle Cranmer clicks for the Higgs! Also pictured: Itay Yavin (standing), James Beacham (sitting), and Veuve Clicquot (boxed)"

[ALEPH, JHEP 2010; Batson, Wired 2015; in backup see Kile, von Wimmersperg-Toeller, arXiv 2017, JHEP 2018, arXiv 2017]

ALEPH Confronts the CMS Ridge

2010 pp surprise!



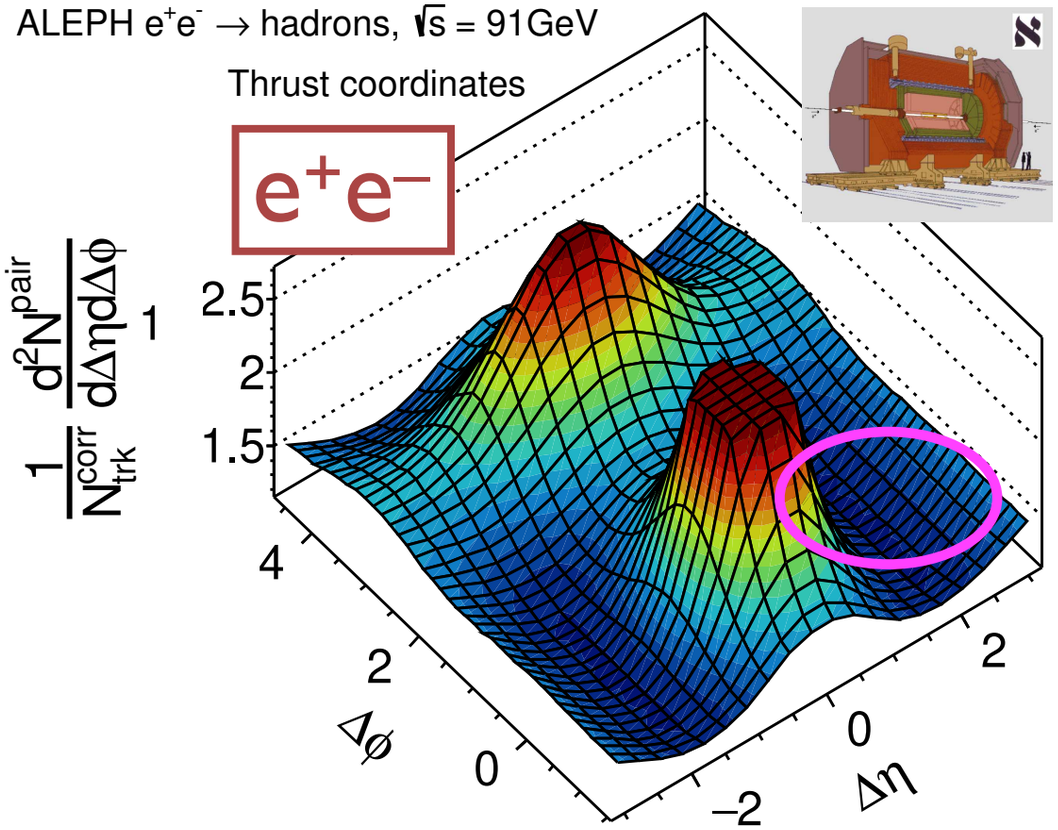
[CMS, JHEP 2010]

ALEPH Confronts the CMS Ridge

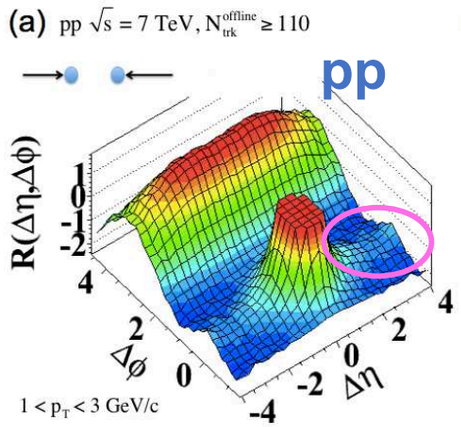
1990–95 e^+e^- data

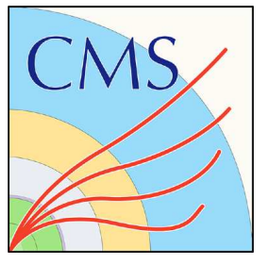
⇓ ⟵ 2010 pp surprise!

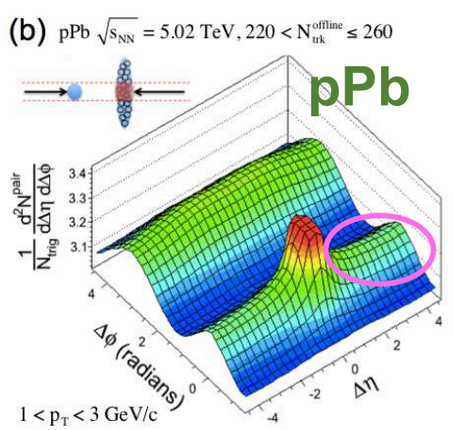
2019 e^+e^- analysis

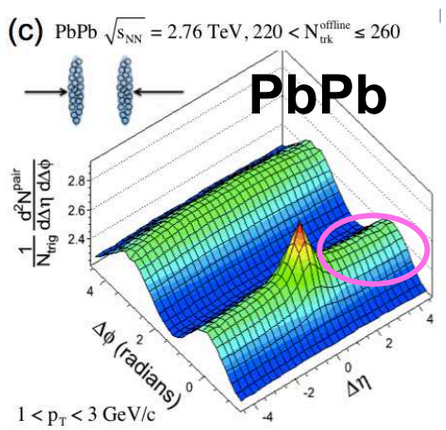


vs.









[Badea, Baty, Chang, Innocenti, Maggi, McGinn, Peters, Sheng, JDT, Lee, PRL 2019]

[CMS, JHEP 2010]

Looking to the Present…

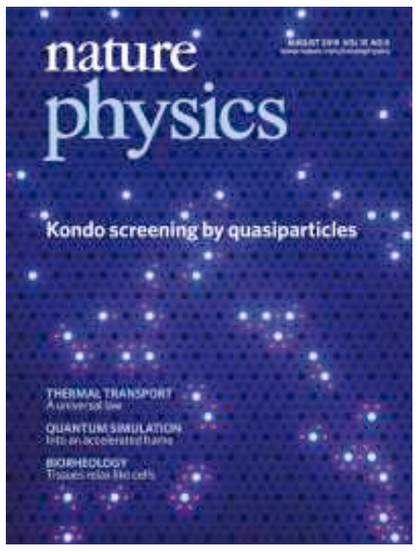
"I support open data in principle, but I'm worried about the adverse effects this might have on the ability of large experimental collaborations to function effectively."

[Strassler, JDT, Nature Physics 2019; see public data case studies in astrophysics: Planck, Fermi, LIGO/Virgo]

Looking to the Present…

"I support open data in principle, but I'm worried about the adverse effects this might have on the ability of large experimental collaborations to function effectively."

As am I!



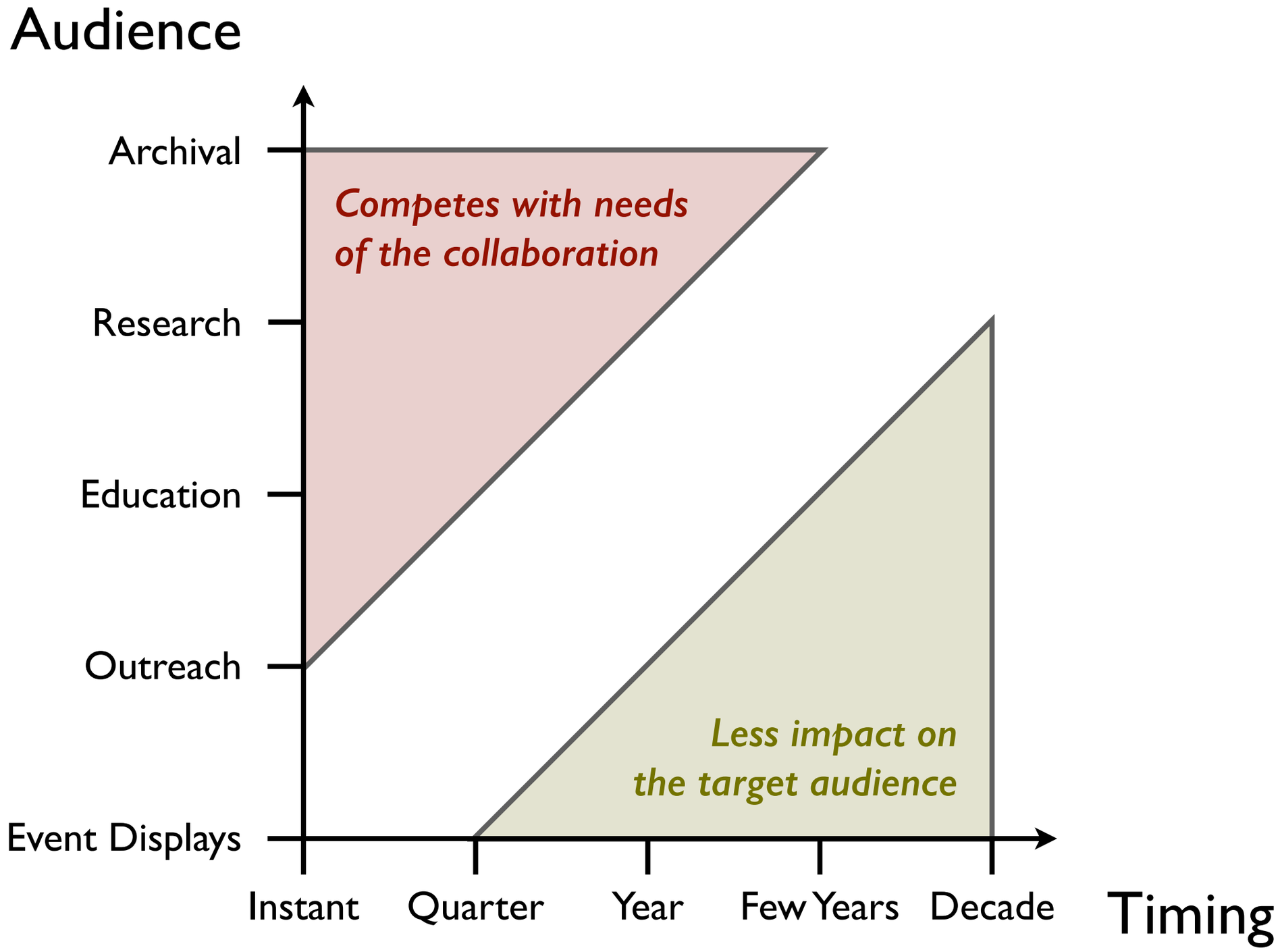
COLLECTION | 01 AUGUST 2019

Viewpoints on data and code sharing

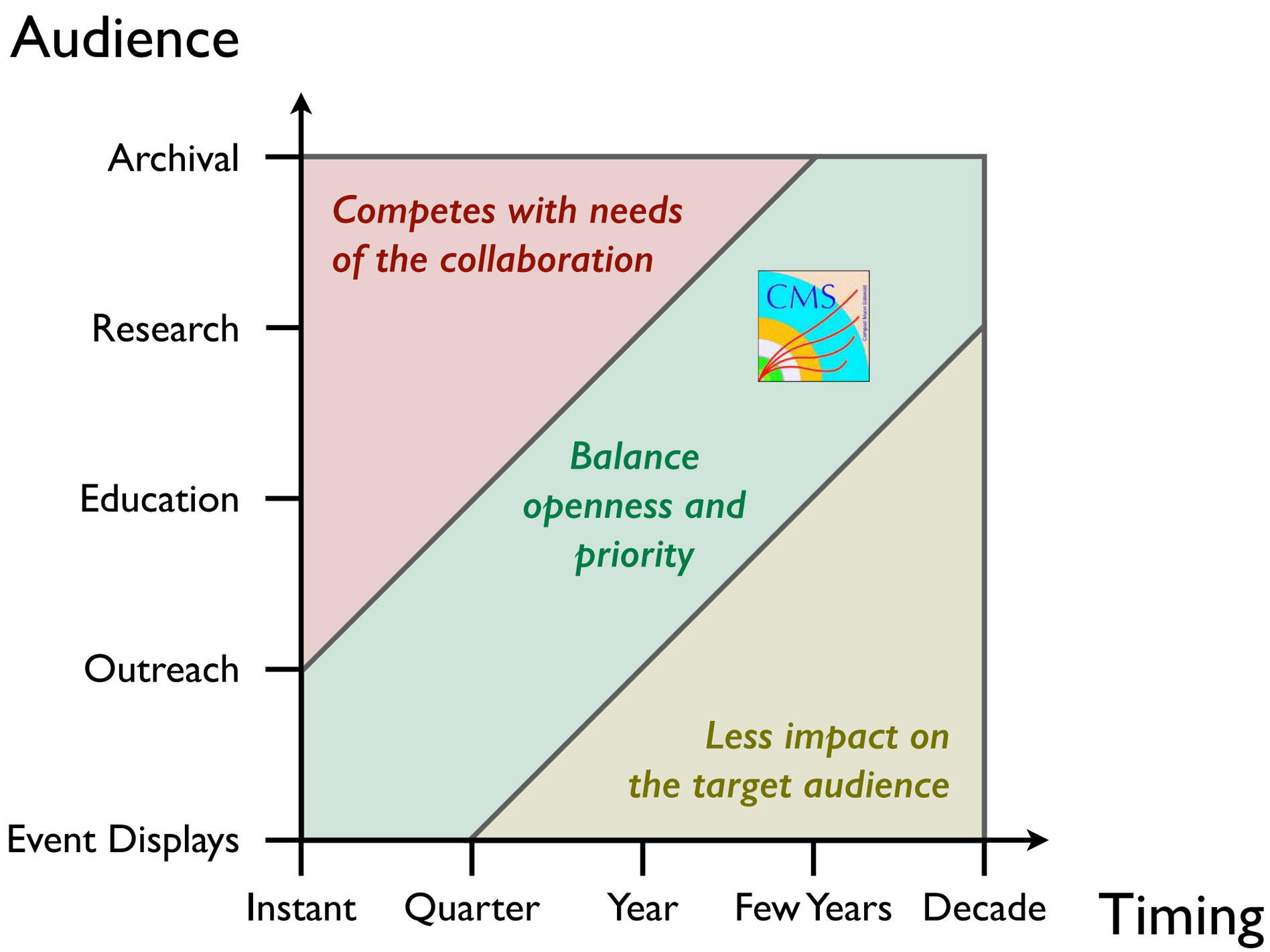
"Only those who spent years building the experiments have earned quick access."

[Strassler, JDT, Nature Physics 2019; see public data case studies in astrophysics: Planck, Fermi, LIGO/Virgo]

What do we mean by "Public Data"?



What do we mean by "Public Data"?



Looking to the Future…

SnowMass2021

Complementarily of pp and e^+e^- colliders:

HiLumi HL-LHC PROJECT ⇔ ILC 250 · CEPC · FCC ee · CLIC 380

Pushing to higher energies: ILC-1000, SPPC, FCC-hh, CLIC-3000, …

Humanity may never again build a 14 TeV proton-proton collider!

(or a 2 TeV proton-antiproton collider,
or an 800-mile-baseline neutrino experiment, or …)

Can we nevertheless capitalize on future scientific ingenuity?

[2021 APS DPF Community Planning Process]

Data preservation (and outside analyses) require significant resources:

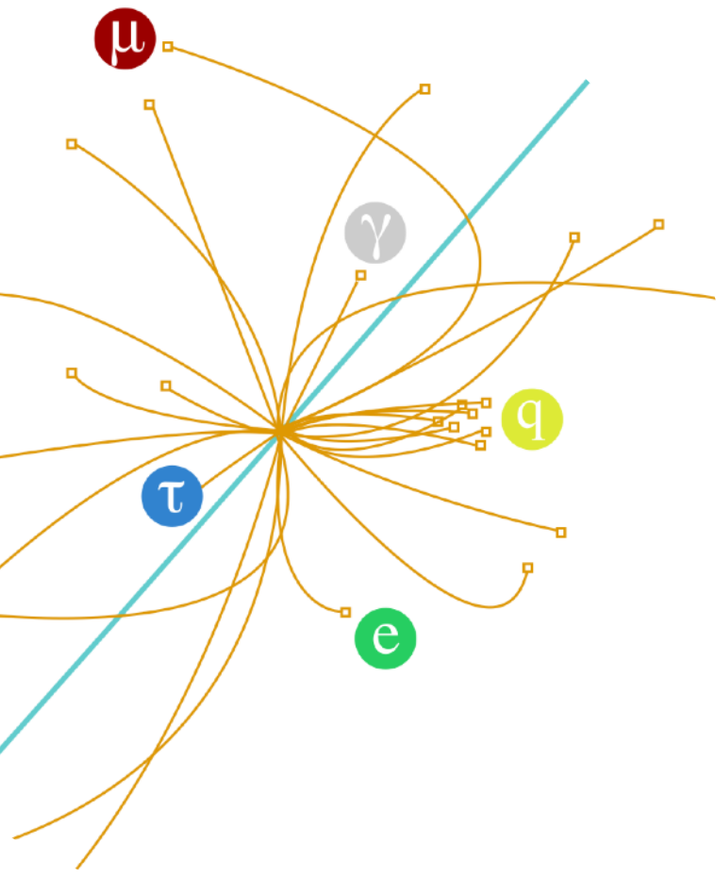
People, time, ideas, and money

Multiple archival models needed to capitalize on future advances in theory, experiment and data analysis:

[Chen, et al., Nature Physics 2018]

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

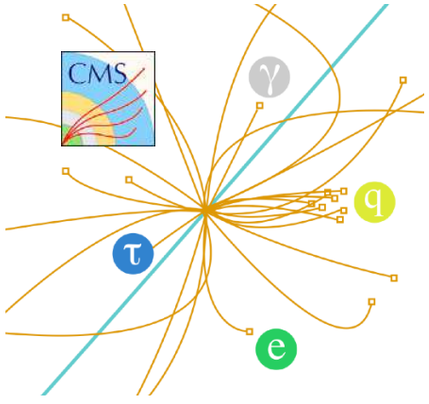
Facilitating dialogue between theory and experiment

Researching physics in and beyond the standard model

These benefits are worth sustained investment in public data initiatives

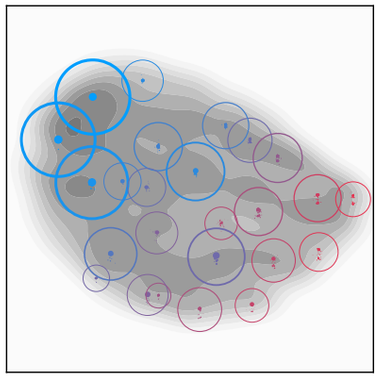
[JDT, CMS Blog 2017]

Summary



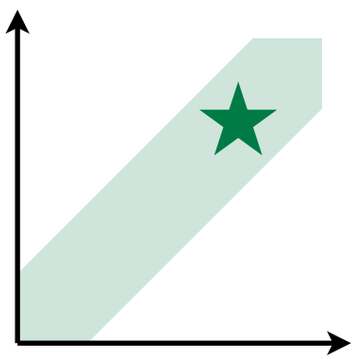
Introducing the CMS Open Data

Unique collider data set with enormous scientific value



Adventures with Public Collider Data

Exploratory studies in and beyond the Standard Model



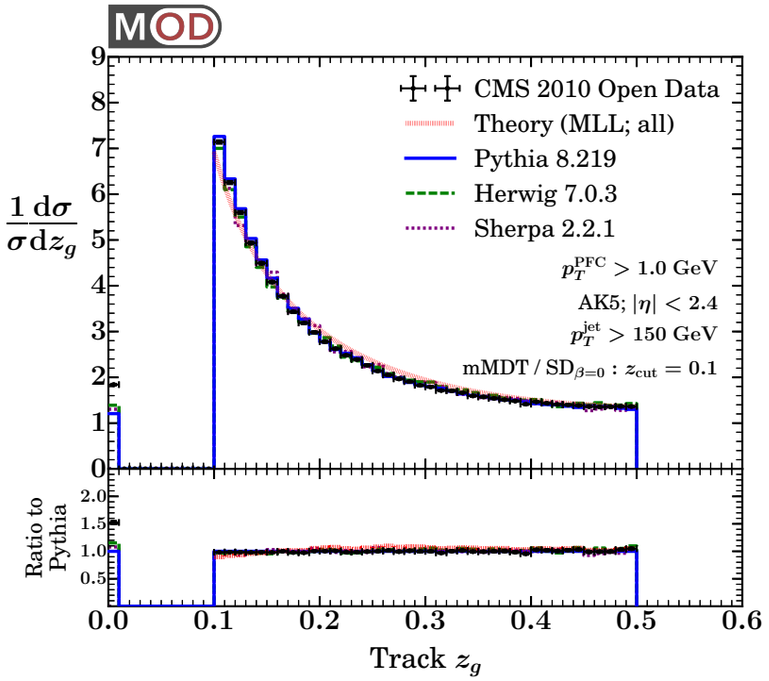
Looking to the Past, Present, and Future

Sustained investment from outreach to research to archives

Backup Slides

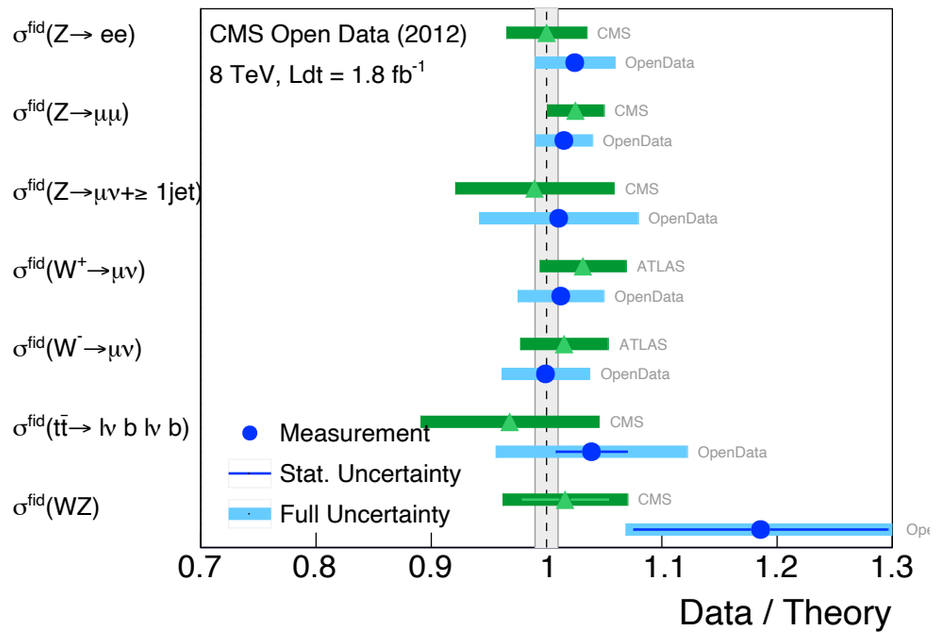
Standard Model Analyses

Jet Substructure Studies



[Tripathee, Xue, Larkoski, Marzani, JDT, PRL 2017, PRD 2017; based on Larkoski, Marzani, JDT, PRD 2015]

Electroweak Benchmarks

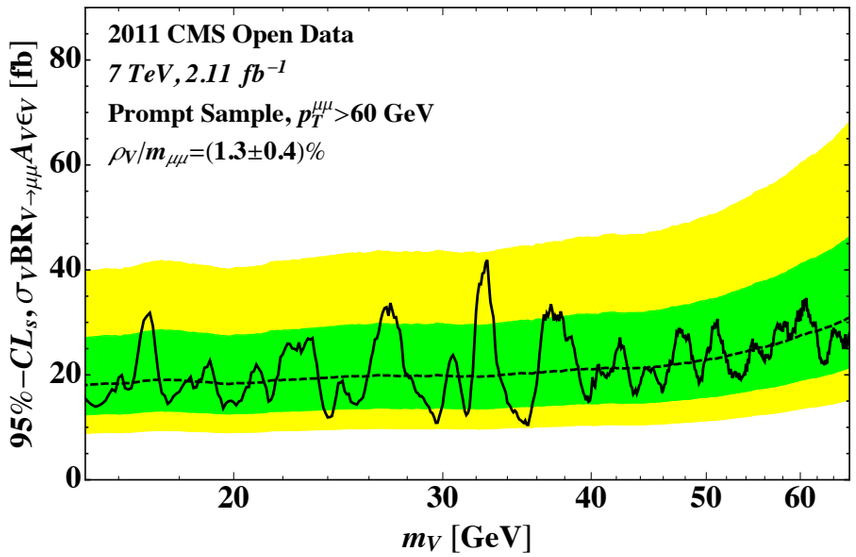


[Apyan, Cuozzo, Klute, Saito, Schott, Sintayehu, JINST 2020]

2010 data ⇒ 2014 release ⇒ 2015 idea ⇒ 2017 analysis

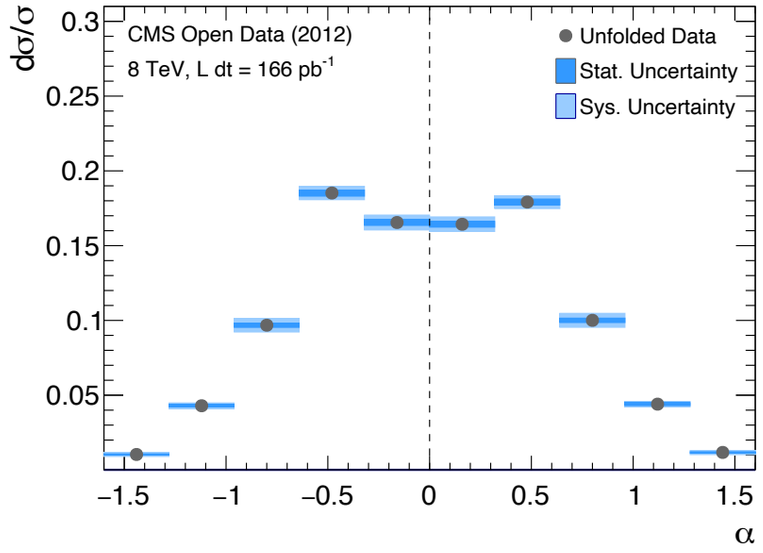
BSM Searches

Dimuon Resonance Hunt



[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019]

Non-Standard Parity Violation

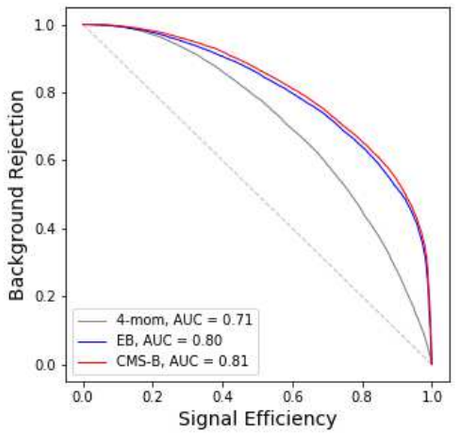


[Lester, Schott, JHEP 2019]

"It is hard to imagine any reason why every possible attempt should not be made to test and re-test the fundamental symmetries of nature every time a door opens onto a new energy range."

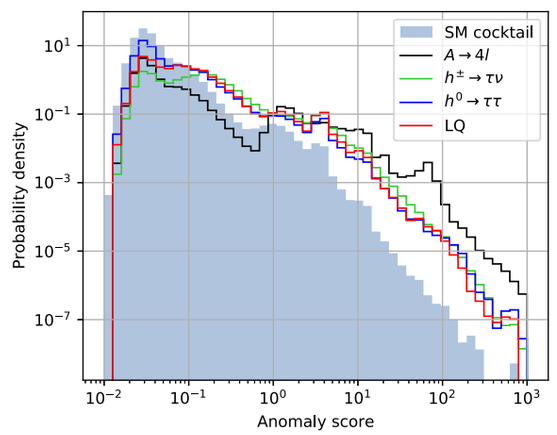
Machine Learning Studies

End-to-End Classification



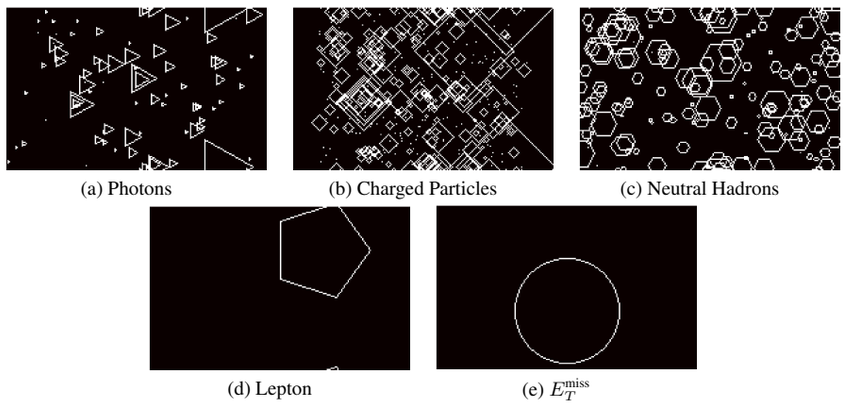
[Andrews, Paulini, Gleyzer, Poczos, CSBS 2020; see also Andrews, Alison, An, Bryant, Burkle, Gleyzer, Narain, Paulini, Poczos, Usai, NIM 2020]

Anomaly Detection



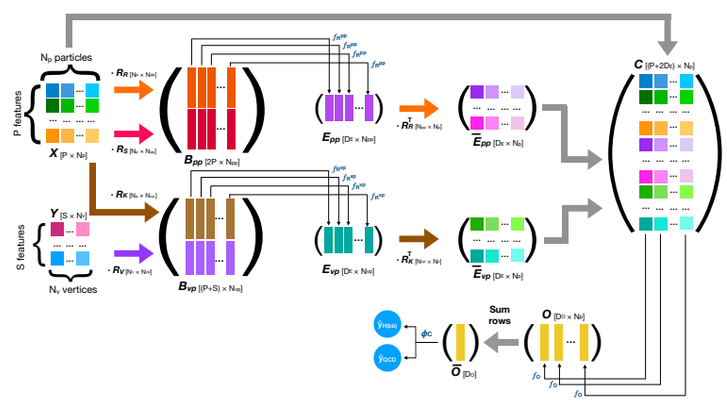
[Knapp, Dissertori, Cerri, Nguyen, Vlimant, Pierini, arXiv 2020]

Computer Vision Techniques



[Fernández Madrazo, Heredia Cacha, Lloret Iglesias, Marco de Lucas, EPJWoC 2019; figure from Nguyen, Weitekamp, Anderson, Castello, Cerri, Pierini, Spiropulu, Vlimant, CSBC 2019]

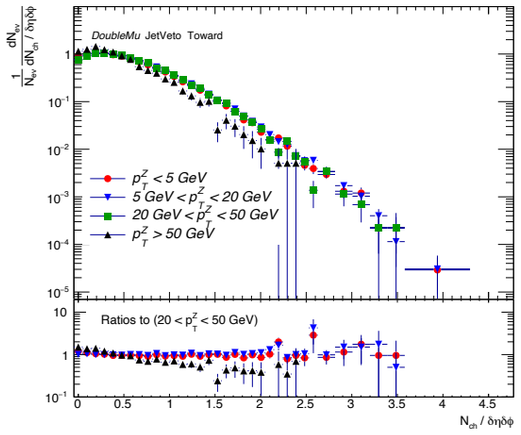
Interaction Networks



[Moreno, Nguyen, Vlimant, Cerri, Newman, Periwal, Spiropulu, Duarte, Pierini, PRD 2020]

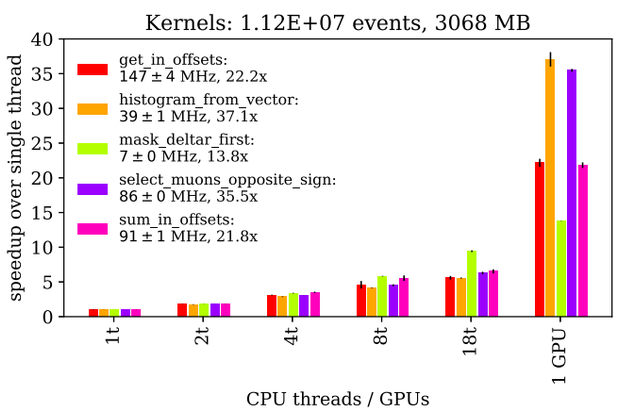
And More!

Underlying Event Studies



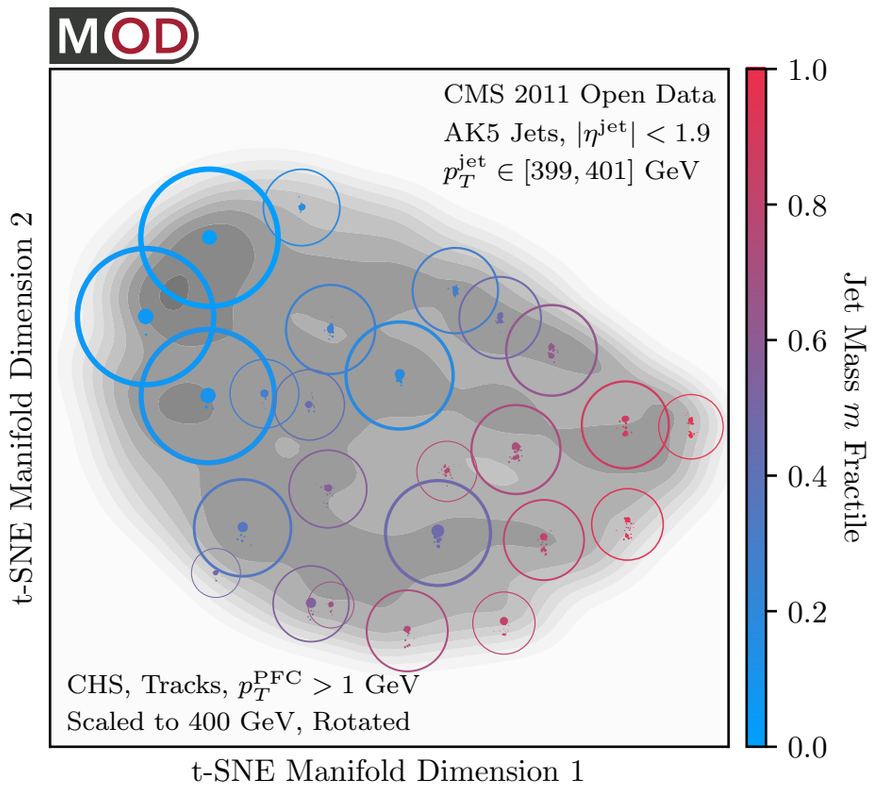
[Paktinat Mehdiabadi, Fahim, JPG 2019]

GPU Acceleration



[Pata, Spiropulu, arXiv 2019]

Event Space Geometry



[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020]

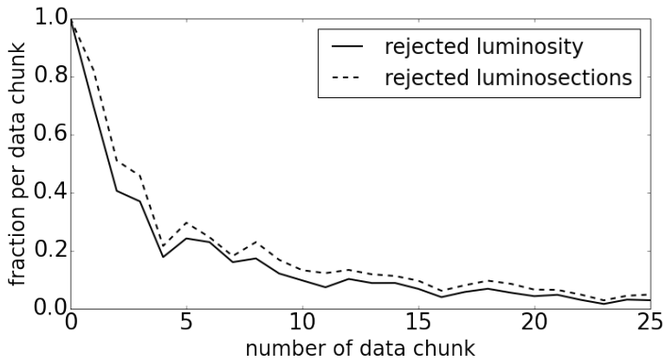
Please contact me if I missed your CMS Open Data study!

And Even More!

Addendum

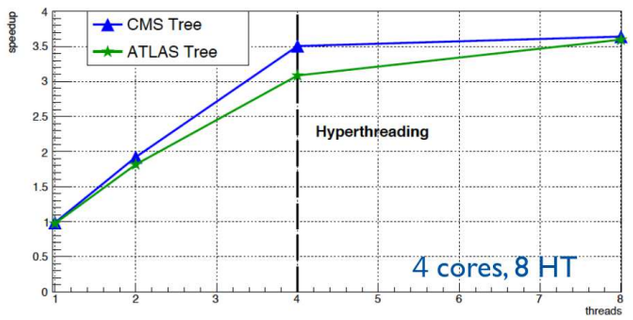
Thanks to Achim Geiser for the bibliography!

Data Quality



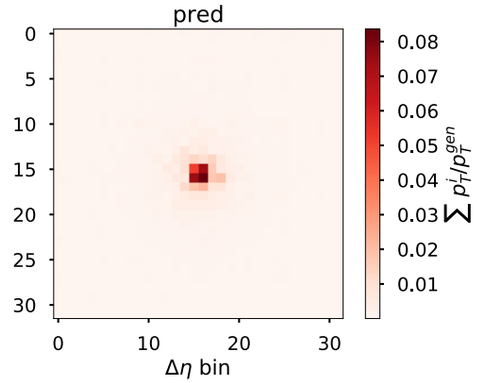
[Borisyak, Ratnikov, Derkach, Ustyuzhanin, JPCS 2017]

Parallel Processing



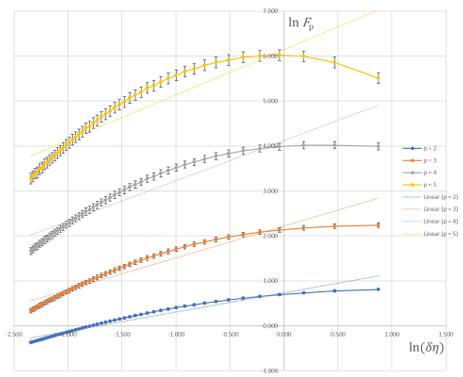
[Piparo, Tejedor, Guiraud, Ganis, Mato, Moneta, Valls Pla, Canal, JPCS 2017]

Generative Models



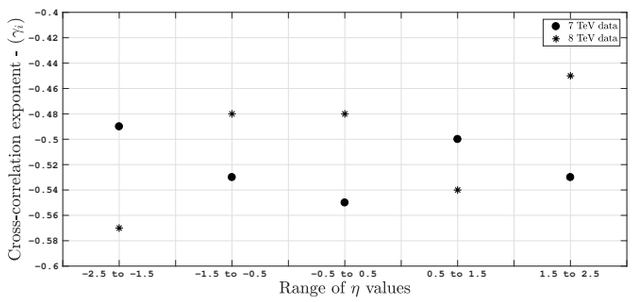
[Musella, Pandolfi, CSBS 2018]

Rapidity Fluctuations



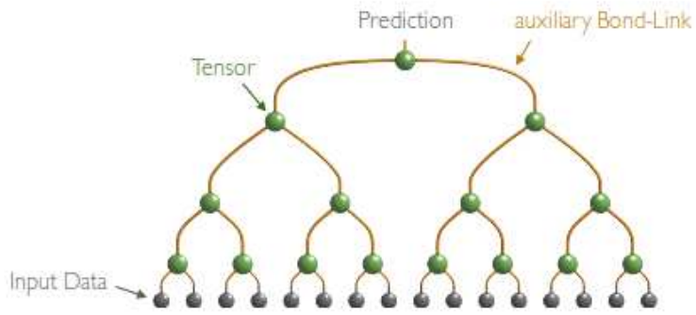
[Ong, Yuen, Ang, Chan, Oh, EPJWoC 2019]

Symmetry Scalings



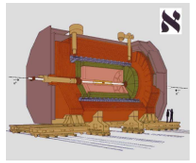
[Bhaduri, Bhaduri, Ghosh, AHEP 2020]

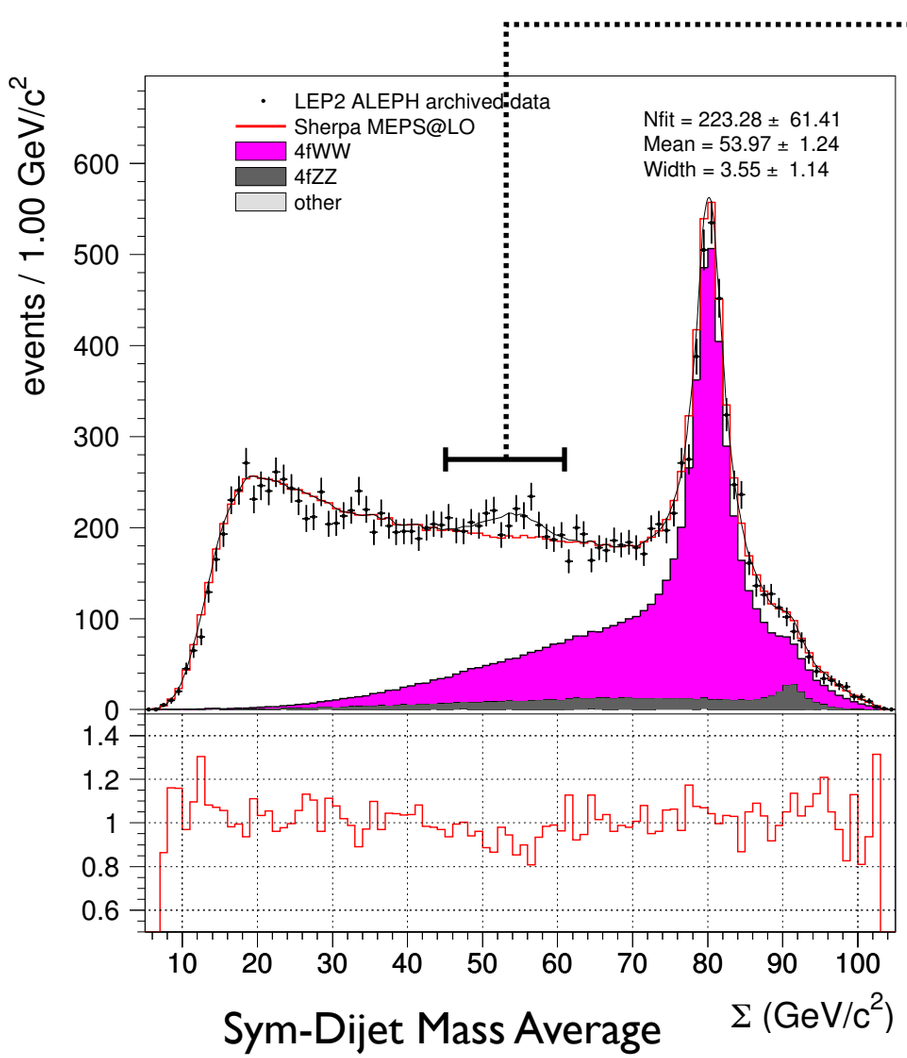
Tensor Networks (LHCb)



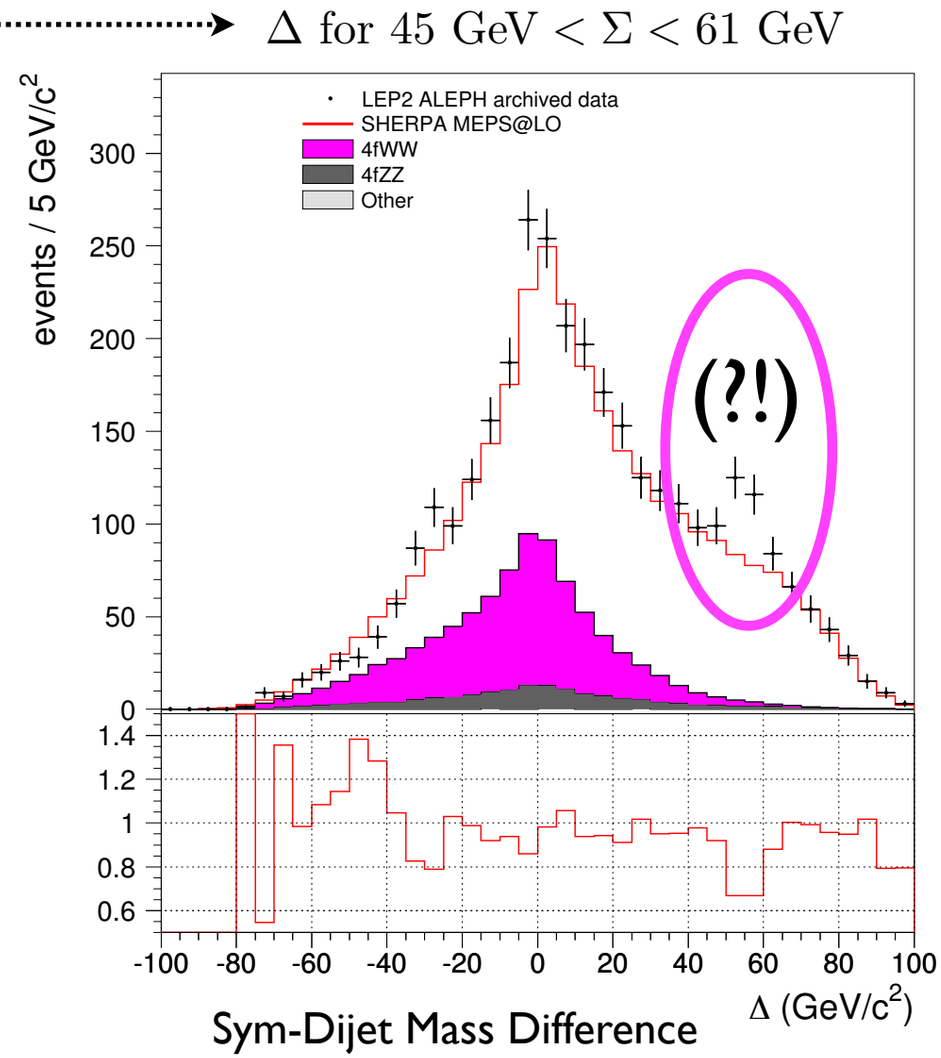
[Trenti, Sestini, Gianelle, Zuliani, Felser, Lucchesi, Montangero, arXiv 2020]

ALEPH Puzzle in Quad-Jet Kinematics





Δ for 45 GeV $< \Sigma <$ 61 GeV



"Whether the excesses described here ultimately are explained by QCD or physics beyond the Standard Model, our results demonstrate the lasting utility of the archived LEP data."

[Kile, von Wimmersperg-Toeller, arXiv 2017, JHEP 2018, arXiv 2017]