Data ex Machina
Machine Learning with Jets in CMS Open Data

Machine Learning for Jet Physics 2020

Eric M. Metodiev
Center for Theoretical Physics
Massachusetts Institute of Technology

Patrick Komiske
Radha Mastandrea
Preksha Naik
Jesse Thaler
CMS Open Data
A new public dataset for jet studies

Unsupervised Learning
A metric for collider events

Supervised Learning
Training directly on collider data
CMS Open Data
A new public dataset for jet studies

Unsupervised Learning
A metric for collider events

Supervised Learning
Training directly on collider data
opendata.cern.ch

Explore more than two petabytes of open data from particle physics!

Search examples: collision datasets, keywords:education, energy:7TeV

Explore
- datasets
- software
- environments
- documentation

Focus on
- ATLAS
- ALICE
- CMS
- LHCb
- OPERA
- Data Science

Get started
CMS 2011A Jet Primary Dataset (+ Simulation)

2.3 fb$^{-1}$ of 7 TeV proton-proton collision data. [link]

~2 million $R = 0.5$ jets with $p_T > 375$ GeV, $|\eta| < 1.9$
Our processed jet dataset is public!

[Komiske, Mastandrea, EMM, Naik, Thaler, 1908.08542]

~2 million $R = 0.5$ anti-$kT$ jets recorded by CMS

$p_T > 375$ GeV, $|\eta| < 1.9$

Jets as lists of particle flow candidates:

$[p_T, y, \phi, ID, \text{vertex}]$

Plus additional information:

Jet energy correction factors

Monte Carlo samples

Detector simulation

Zenodo record

Binder demo to download and read the dataset
CMS Open Data
A new public dataset for jet studies

Unsupervised Learning
A metric for collider events

Supervised Learning
Training directly on collider data
CMS Open Data
A new public dataset for jet studies

Unsupervised Learning
A metric for collider events

Supervised Learning
Training directly on collider data
When are two jets similar?

Many unsupervised methods rely on a **distance matrix**. Need a physically-sensible **metric** between jets!

These two jets “look” similar, but have different numbers of particles, flavors, and locations.

---

**How do we quantify this?**

“Space of Jets”
The Energy Mover’s Distance

“Energy” Mover’s Distance: the minimum “work” (energy x angle) to rearrange one jet (pile of energy) into another

- Energy Mover’s Distance: the minimum “work” (energy x angle) to rearrange one jet (pile of energy) into another

![Diagram illustrating the Energy Mover’s Distance](image)

See Today’s Talk on Optimal Transport

[Komiske, EMM, Thaler, 1902.02346]
Earth Mover’s Distance: [Peleg, Werman, Rom] [Rubner, Tomasi, Guibas]
Six Decades of Collider Techniques

1960

Taming infinities

1962-1964

Infrared Safety

[Kinoshita, JMP 1962]

[Lee, Nauenberg, PR 1964]

1977

Event Shapes

1977

Thrust, Sphericity

[Farhi, PRL 1977]

[Georgi, Machacek, PRL 1977]

1993

Jet Algorithms

$k_T$ jet clustering

[Ellis, Soper, PRD 1993]

[Catani, Dokshitzer, Seymour, Webber, NPB 1993]

1997-1998

Jet Substructure

C/A jet clustering

[Wobisch, Wengerl, 1998]

[Dokshitzer, Leder, Moretti, Webber, JHEP 1997]

2010-2015

N-(subjettiness, XCone

[Stewart, Tackmann, Waalewijn, PRL 2010]

[Thaler, Van Tilburg, JHEP 2011]

[Stewart, Tackmann, Thaler, Vermilion, Wilkason, JHEP 2015]

2014-2019

Constituent Subtraction

[Berta, Spousta, Miller, Leitner, JHEP 2014]

[Berta, Masetti, Miller, Spousta, JHEP 2019]

2020

Pileup

And many more!

[Stewart, Tackmann, Waalewijn, PRL 2010]

[Thaler, Van Tilburg, JHEP 2011]

[Stewart, Tackmann, Thaler, Vermilion, Wilkason, JHEP 2015]
Six Decades of Collider Techniques as Optimal Transport!

[Komiske, EMM, Thaler, to appear]

Smooth function of energy distribution are finite in QFT

Event shapes as distances to the 2-particle manifold

Jet Algorithms

\[ \mathcal{E} = \text{argmin} \ EMD(\mathcal{E}, \mathcal{E}') \]

Jet Substructure

\[ t(\mathcal{E}) = \min_{|\mathcal{E}'|=2} EMD(\mathcal{E}, \mathcal{E}') \]

Jets are N-particle event approximations

Subtract a pileup as a uniform distribution

EMD(\mathcal{E}, \mathcal{E}') < \delta \rightarrow |\mathcal{O}(\mathcal{E}) - \mathcal{O}(\mathcal{E}')| < \epsilon

Taming infinities

Event Shapes

Jet Algorithms

Jet Substructure

Pileup

1960

1962-1964

1977

1993

1997-1998

2010-2015

2014-2019

2020

Infrared Safety

Thrust, Sphericity

\[ k_T \] jet clustering

C/A jet clustering

N-(sub)jettiness, XCones

Constituent Subtraction

And many more!

[Farhi, PRL 1977]

[Georgi, Machacek, PRL 1977]

[Flohr, Soper, PRD 1993]

[Catani, Dokshitzer, Seymour, Webber, NPB 1993]

[Wobisch, Wengler, 1998]

[Dokshitzer, Leder, Moretti, Webber, JHEP 1997]

[Stewart, Tackmann, Waalewijn, PRL 2010]

[Thaler, Van Tilburg, JHEP 2011]

[Stewart, Tackmann, Thaler, Vermilion, Wilkason, JHEP 2015]
Visualizing the Manifold

What does the space of jets look like?

[t-SNE embedding]

[van der Maaten, Hinton, JMLR 2008]
Visualizing the Manifold

What does the space of jets look like?

[t-SNE embedding: 25-medoid jets shown]

[van der Maaten, Hinton, JMLR 2008]
[Komiske, Mastandrea, EMM, Naik, Thaler, 1908.08542]
Visualizing the Manifold

What does the space of jets look like?

[van der Maaten, Hinton, JMLR 2008]
[Komiske, Mastandrea, EMM, Naik, Thaler, 1908.08542]
Correlation Dimension

Conceptual Idea

\[ N_{\text{neighbors}}(r) \propto r^{\text{dim}} \]

\[ \text{dim}\approx 1 \quad \text{dim}\approx 2 \quad \text{dim}\approx 0 \]

(\text{eventually } 0)

\[ \text{dim}(Q) = Q \frac{\partial}{\partial Q} \ln \sum_{i=1}^{N} \sum_{j=1}^{N} \Theta[\text{EMD}(\varepsilon_i, \varepsilon_j) < Q] \]

Experimental Data

Dimension blows up at low energies.

Theoretical Calculation

See extra slides for calculation sketch.

See Jack’s Talk for more on dimensionality

[Grassberger, Procaccia, PRL 1983]  [Kegl, NeurIPS 2002]

[Komiske, Mastandrea, EMM, Naik, Thaler, I 908.08542]
CMS Open Data
A new public dataset for jet studies

Unsupervised Learning
A metric for collider events

Supervised Learning
Training directly on collider data
CMS Open Data
A new public dataset for jet studies

Unsupervised Learning
A metric for collider events

Supervised Learning
Training directly on collider data
Training on pure samples: **Cat** jets vs. **Dog** jets

Cat Jets

Dog Jets

Classifier

1 vs. 0
Training on mixed samples: **Cat jets vs. Dog jets**

This defines an equivalent classifier to the pure case!
Training on pure samples: Quark jets vs. Gluon jets

Quark Jets

Gluon Jets

Classifier

1

0

vs.
Training on mixed samples: Quark jets vs. Gluon jets

Classifier

[EMM, B. Nachman, J. Thaler, 1708.02949]
[P.T. Komiske, EMM, B. Nachman, M.D. Schwartz, 1801.10158]
[L. Dery, B. Nachman, F. Rubbo, A. Schwartzman, 1702.00414]
[T. Cohen, M. Freytsis, B. Ostdiek, 1706.09451]
To reduce sample dependence, we train an EFN on tracks with $p_T^{\text{PFC}} > 1$ GeV and remove pileup.

Or high-dimensional unfolding? See Patrick’s Talk
What is the model learning?

$$\text{EFN} = F \left( \sum_{i=1}^{M} p_{Ti} \Phi(y_i, \phi_i) \right)$$

Learn these

See Patrick's Talk at ML4Jets 2018
CMS Open Data
A new public dataset for jet studies

Unsupervised Learning
A metric for collider events

Supervised Learning
Training directly on collider data
Welcome to EnergyFlow

EnergyFlow is a Python package containing a suite of particle physics tools:

- **Energy Flow Polynomials**: EFPs are a collection of jet substructure observables which form a complete linear basis of IRC-safe observables. EnergyFlow provides tools to compute EFPs on
The End
Thank you!
Extra Slides
Most Representative Jets

Jet Mass: $m = (\sum_{i=1}^{M} p_i^\mu)^2$

Measures how “wide” the jet is.

Jet Mass Histogram

Komiske, Mastandrea, EMM, Naik, Thaler, 1908.08542
When are two jets similar?

The energy flow (distribution of energy) is the information that is robust to: fragmentation, hadronization, detector effects, …

Energy flow ⇔ Infrared and Collinear (IRC) Safe information

[F.V. Tkachov, 9601308]
[N.A. Sveshnikov, F.V. Tkachov, 9512370]
[P.S. Cherzor, N.A. Sveshnikov, 9710349]
When are two jets similar?

Treat jets as distributions of energy: \[ E(\hat{n}) = \sum_{i=1}^{M} E_i \delta(\hat{n} - \hat{n}_i) \]

Ignoring particle flavor, charge…
The Energy Mover’s Distance

Review: The Earth Mover’s Distance

**Earth Mover’s Distance**: the minimum “work” (stuff x distance) to rearrange one pile of dirt into another

Metric on the space of (normalized) distributions: *symmetric, non-negative, triangle inequality*

Distributions are close in EMD ⇔ their expectation values are close.

Also known as the 1-Wasserstein metric.
The Energy Mover’s Distance

Energy Mover’s Distance: the minimum “work” \((\text{energy} \times \text{angle})\) to rearrange one event (pile of energy) into another

\[
\text{EMD}(\mathcal{E}, \mathcal{E}') = \min_{\{f\}} \left( \sum_{i=1}^{M} \sum_{j=1}^{M'} f_{ij} \frac{\theta_{ij}}{R} + \sum_{i=1}^{M} E_i - \sum_{j=1}^{M'} E'_j \right)
\]

Difference in radiation pattern

Difference in total energy

[Komiske, EMM, Thaler, 1902.02346]
The Energy Mover’s Distance

From Earth to Energy

**Energy Mover’s Distance**: the minimum “work” \((\text{energy} \times \text{angle})\) to rearrange one event (pile of energy) into another

\[ \text{EMD(\mathcal{E}, \mathcal{E}')} + \text{EMD(\mathcal{E}', \mathcal{E}'')} \geq \text{EMD(\mathcal{E}, \mathcal{E}'')} \]

**EMD** has dimensions of energy

**True metric as long as**

\[ R \geq \frac{1}{2} \theta_{\text{max}} \]

\( R \) is the jet radius, for conical jets

**Solvable via Optimal Transport problem.**

~1ms to compute EMD for two jets with 100 particles.
The Energy Mover’s Distance
From Earth to Energy

**Energy Mover’s Distance**: the minimum “work” \((\text{energy} \times \text{angle})\) to rearrange one event (pile of energy) into another

[Komiske, EMM, Thaler, 1902.02346]

\[ \text{EMD}(\varepsilon, \varepsilon') + \text{EMD}(\varepsilon', \varepsilon'') \geq \text{EMD}(\varepsilon, \varepsilon'') \]

https://energyflow.network
Energy Moving and IRC Safety

Events close in EMD are close in any infrared and collinear safe observable!

Additive IRC-safe observables:

\[ \text{EMD}(\mathcal{E}, \mathcal{E}') \geq \frac{1}{RL} |\mathcal{O}(\mathcal{E}) - \mathcal{O}(\mathcal{E}')| \]

“Lipschitz constant” of \( \Phi \)
i.e. bound on its derivative

\[ \mathcal{O}(\mathcal{E}) = \sum_{i=1}^{M} E_i \Phi(\hat{n}_i) \]
e.g. \( \beta \geq 1 \) jet angularities:

[Berger, Kucs, Sterman, 0303051]
[Larkoski, Thaler, Waalewijn, 1408.3122]

\[ |\lambda(\beta)(\mathcal{E}) - \lambda(\beta)(\mathcal{E}')| \leq \beta \text{ EMD}(\mathcal{E}, \mathcal{E}') \]
Old Observables in a New Language

**$N$-subjettiness** is the EMD between the event and the closest $N$-particle event.

$$
\tau_N^{(\beta)}(\mathcal{E}) = \min_{\text{axes}} \sum_{i=1}^{M} E_i \min \{ \theta_{1,i}^{\beta}, \theta_{2,i}^{\beta}, \ldots, \theta_{N,i}^{\beta} \}
$$

$$
\tau_N(\mathcal{E}) = \min_{|\mathcal{E}'|=N} \text{EMD}(\mathcal{E}, \mathcal{E}').
$$

$\beta \geq 1$ is $p$-Wasserstein distance with $p = \beta$.

**Thrust** is the EMD between the event and two back-to-back particles.

$$
t(\mathcal{E}) = E - \max_n \sum_i |\vec{p}_i \cdot \hat{n}| \quad \rightarrow \quad t(\mathcal{E}) = \min_{|\mathcal{E}'|=2} \text{EMD}(\mathcal{E}, \mathcal{E}')
$$

with $\theta_{ij} = \hat{n}_i \cdot \hat{n}_j$, $\hat{n} = \vec{p}/E$. 

Geometry in the space of events
Quantifying Pileup and Detector Effects with EMD

**Gen./Sim.** EMD universally quantifies pileup and detector effects.

- **Gen./Sim. EMD: 44.4 GeV**
  - Translated Azimuthal Angle $\phi$
  - Translated Rapidity $y$
  - CMS 2011 Simulation
  - AK5 Jets
  - All PFCs
  - Scaled to 400 GeV
  - EMD: 44.4 GeV

- **Gen./Sim. EMD: 33.7 GeV**
  - Translated Azimuthal Angle $\phi$
  - Translated Rapidity $y$
  - CMS 2011 Simulation
  - AK5 Jets, CHS
  - All PFCs
  - Scaled to 400 GeV
  - EMD: 33.7 GeV

- **Gen./Sim. EMD: 6.7 GeV**
  - Translated Azimuthal Angle $\phi$
  - Translated Rapidity $y$
  - CMS 2013 Simulation
  - AK5 Jets, CHS
  - Tracks, $p_T^{PFC} > 1$ GeV
  - Scaled to 400 GeV
  - EMD: 6.7 GeV

+ charged hadron subtraction

+ Tracks only, $p_T^{PFC} > 1$ GeV cut

See extra slides for histograms. Can also quantify hadronization effects this way.
Exploring the Space of Jets: Visualizing the Manifold

Visualize the space of events with t-Distributed Stochastic Neighbor Embedding (t-SNE).

Finds an embedding into a low-dimensional manifold that respects distances.

What does the space of jets look like?
Exploring the Space of Jets: Visualizing the Manifold

What does the space of jets look like?

Quantify (and calculate!) dimension of the space of jets. See extra slides.
Exploring the Space of Jets: Correlation Dimension

Intuition:

\[ N_{\text{neighboring}}(r) \propto r^{\dim} \]

\[ \dim(r) = r \frac{\partial}{\partial r} \ln N_{\text{neighbors}}(r) \]

Correlation dimension:

\[ \dim(Q) = Q \frac{\partial}{\partial Q} \ln \sum_{i=1}^{N} \sum_{j=1}^{N} \theta[\text{EMD}(\varepsilon_i, \varepsilon_j) < Q] \]

Energy scale \( Q \) dependence

Count neighbors in ball of radius \( Q \)
Exploring the Space of Jets: Correlation Dimension

At LL: \( \text{dim}_{q/g}(Q) = -\frac{8\alpha_s C_{q/g}}{\pi} \ln \frac{Q}{p_T/2} \)

\[ C_q = C_F = \frac{4}{3} \]
\[ C_g = C_A = 3 \]

+ 1-loop running of \( \alpha_s \)

**EMD: Intrinsic Dimension**
PYTHIA 8.235, \( \sqrt{s} = 14 \text{ TeV} \)
\( R = 1.0, p_T \approx 500 \text{ GeV} \)

**MC**
- Blue: Quark Jets
- Red: Gluon Jets
- Black: Hadrons
- Dashed: Partons
- Dotted: Theory, LL

Dimension blows up at low energies.

Jets are “more than fractal”
CMS Open Data

Many exciting physics applications with the CMS Open Data already.

Exposing the QCD splitting function

[Tripathee, Xue, Larkoski, Marzani, Thaler, 1704.05842]
[Larkoski, Marzani, Thaler, Tripathee, Xue, 1704.05066]

Looking for parity violation in jets

[Lester, Schott, 1904.11195]

Searching for dimuon resonances

[Cesarotti, Soreq, Strassler, Thaler, Xue, 1902.04222]

Analyzing collision data with deep learning techniques

[Madrazo, Cacha, Iglesias, de Lucas, 1708.07034]
[Andrews, Paulini, Gleyzer, Poczos, 1807.11916]
[Andrews, et al., 1902.08276]
Quantifying Pileup and Detector Effects with EMD

Gen./Sim. EMD universally quantifies pileup mitigation and detector effects.
Jet Substructure Observables

Study jet substructure at truth and detector level.

\[ m^2 = \left( \sum_{i \in \text{Jet}} p_i^\mu \right)^2 \]

\[ M = \sum_{i \in \text{Jet}} 1 \]

\[ p_T^D = \frac{\sum_{i \in \text{Jet}} p_{T,i}^2}{\left( \sum_{i \in \text{Jet}} p_{T,i} \right)^2} \]

Similar to: [Larkoski, Marzani, Thaler, Tripathee, Xue, 1704.05066]
Exploring the Space of Jets: Correlation Dimension

QCD jets are simplest.

W jets are more complicated.

Top jets are most complex.

“Decays” have \textasciitilde constant dimension.

**Data ex Machina: ML with Jets in CMS Open Data**

Eric M. Metodiev, MIT
Exploring the Space of Jets: Correlation Dimension

QCD jets are simplest.

W jets are more complicated.

Top jets are most complex.

“Decays” have \( \sim \)constant dimension.

Fragmentation becomes more complex at lower energy scales.

**EMD: Intrinsic Dimension**

*PYTHIA 8.235, \( \sqrt{s} = 14 \) TeV

\( R = 1.0, p_T \in [500, 550] \) GeV

MC
QCD jets are simplest.

W jets are more complicated.

Top jets are most complex.

“Decays” have ~constant dimension.

Fragmentation becomes more complex at lower energy scales.

Hadronization becomes relevant at scales around 20 GeV.
Exploring the Space of Jets: Correlation Dimension

Can we understand this analytically?

Pileup & Detector Effects

AK5 Jets, $|\eta^{\text{jet}}| < 1.9$

$\frac{p_T^{\text{jet}}}{\text{GeV}} \in [375, 425]$ Scaled to 400 GeV

PRELIMINARY

Dimension blows up at low energies.

Jets are “more than fractal”
Jet Kinematic Distributions


(CMS 2011 Open Data
CMS 2011 Simulation
PYTHIA 6 Generation

AK5 Jets, $|\eta^{jet}| < 1.9$
$p_T^{jet} \geq 375$ GeV

(CMS 2011 Open Data
CMS 2011 Simulation
PYTHIA 6 Generation

AK5 Jets
$p_T^{jet} \geq 375$ GeV
Quantifying event modifications: Hadronization

\[ \lambda(\beta=1) = \sum_{i=1}^{M} E_i \theta_i \]

\[ \lambda(\beta=1) = 111.1 \text{GeV} \]

\[ \lambda(\beta=1) = 111.6 \text{GeV} \]

\[ \mathcal{E} = \mathcal{E}_{\text{partons}} \]

\[ \mathcal{E}' = \mathcal{E}_{\text{hadrons}} \]

\[ |\lambda(\beta=1)(\mathcal{E}) - \lambda(\beta=1)(\mathcal{E}')| \leq \text{EMD}(\mathcal{E}, \mathcal{E}') \]
Exploring the Space of Events: Jet Classification

Classify $W$ jets vs. QCD jets

Look at a jet’s nearest neighbors (kNN) to predict its class.

Optimal IRC-safe classifier with enough data.

Nearing performance of ML.
Exploring the Space of Events

Use EMD as a measure of event similarity

Unsupervised clustering algorithms can be used to cluster events

Jets are clusters of particles

???? are clusters of jets

VP Tree: $O(\log(N))$ neighbor query time

Much more to explore.
Exploring the Space of Events: $W$ jets

$W$ jets are 2-pronged:

$z$: Energy Sharing of Prongs
$\theta$: Angle between Prongs
$\varphi$: Azimuthal orientation

Constrained by $W$ mass:

$$z(1 - z)\theta^2 = \frac{p_{\mu j}^2}{p_T^2} = \frac{m_W^2}{p_T^2}$$

Hence we expect a **two-dimensional** space of $W$ jets.

After $\varphi$ rotation: **one-dimensional**
Exploring the Space of Jets: Correlation Dimension

Sketch of leading log (one emission) calculation:

$$\dim_{q/g}(Q) = Q \frac{\partial}{\partial Q} \ln \sum_{i=1}^{N} \sum_{j=1}^{N} \Theta[\text{EMD}(\mathcal{E}_i, \mathcal{E}_j) < Q]$$

$$= Q \frac{\partial}{\partial Q} \ln \Pr[\text{EMD} < Q]$$

$$= Q \frac{\partial}{\partial Q} \ln \Pr[\lambda^{(\beta=1)} < Q; C_{q/g} \rightarrow 2 \ C_{q/g}]$$

$$= Q \frac{\partial}{\partial Q} \ln \exp \left( -\frac{4\alpha_s C_{q/g}}{\pi} \ln^2 \frac{Q}{p_T/2} \right)$$

$$= -\frac{8\alpha_s C_{q/g}}{\pi} \ln \frac{Q}{p_T/2}$$

+ 1-loop running of $\alpha_s$