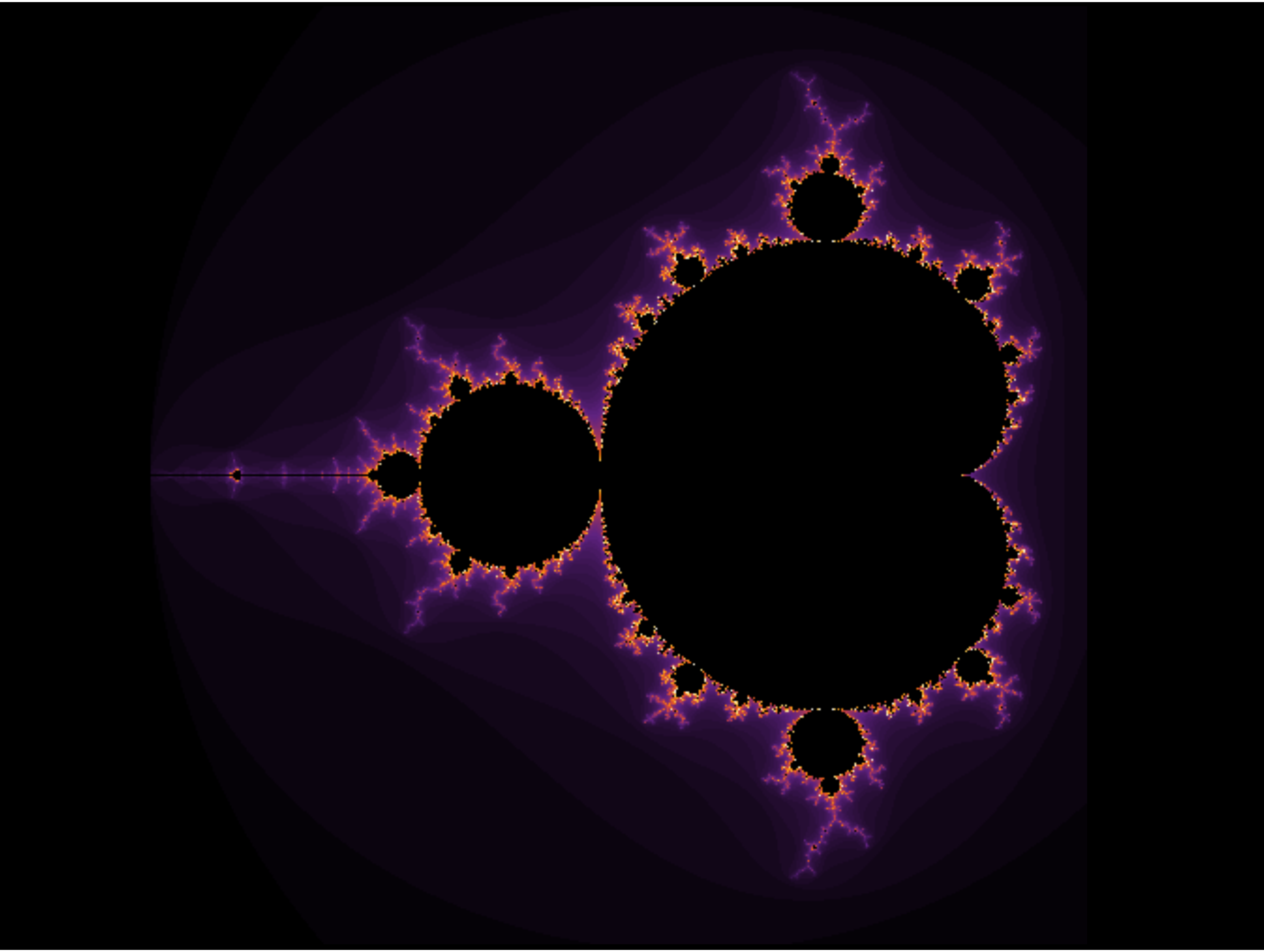
How Much Information is in a Jet?

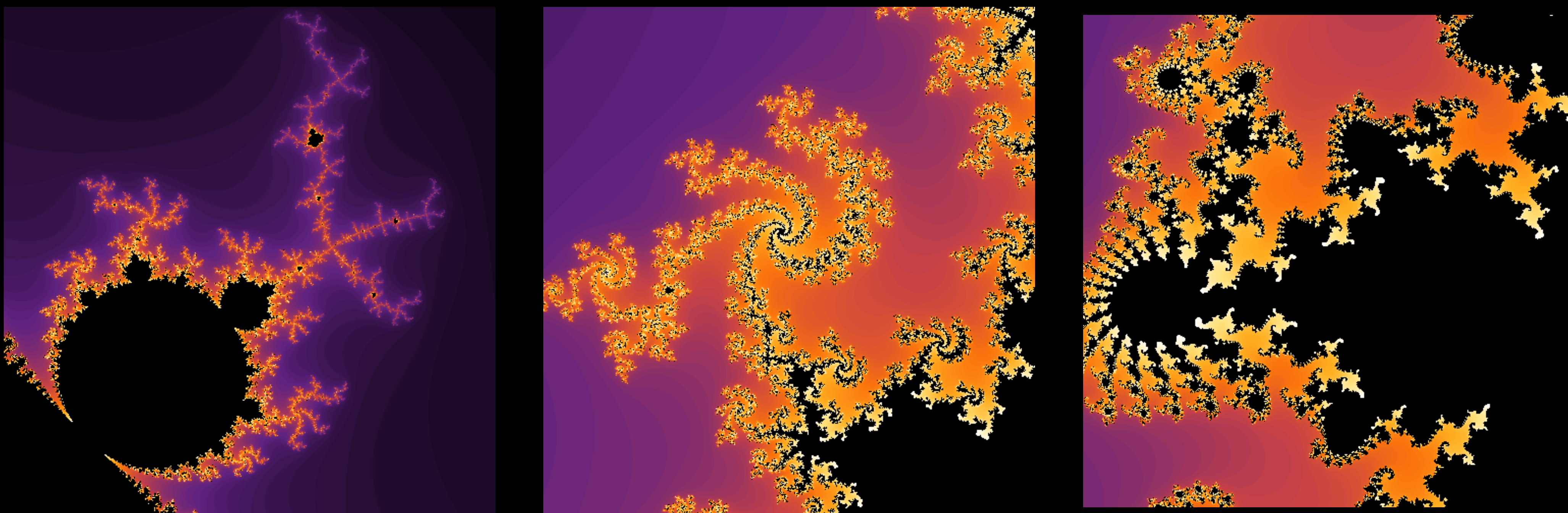
Andrew Larkoski
Reed College

with Kaustuv Datta,

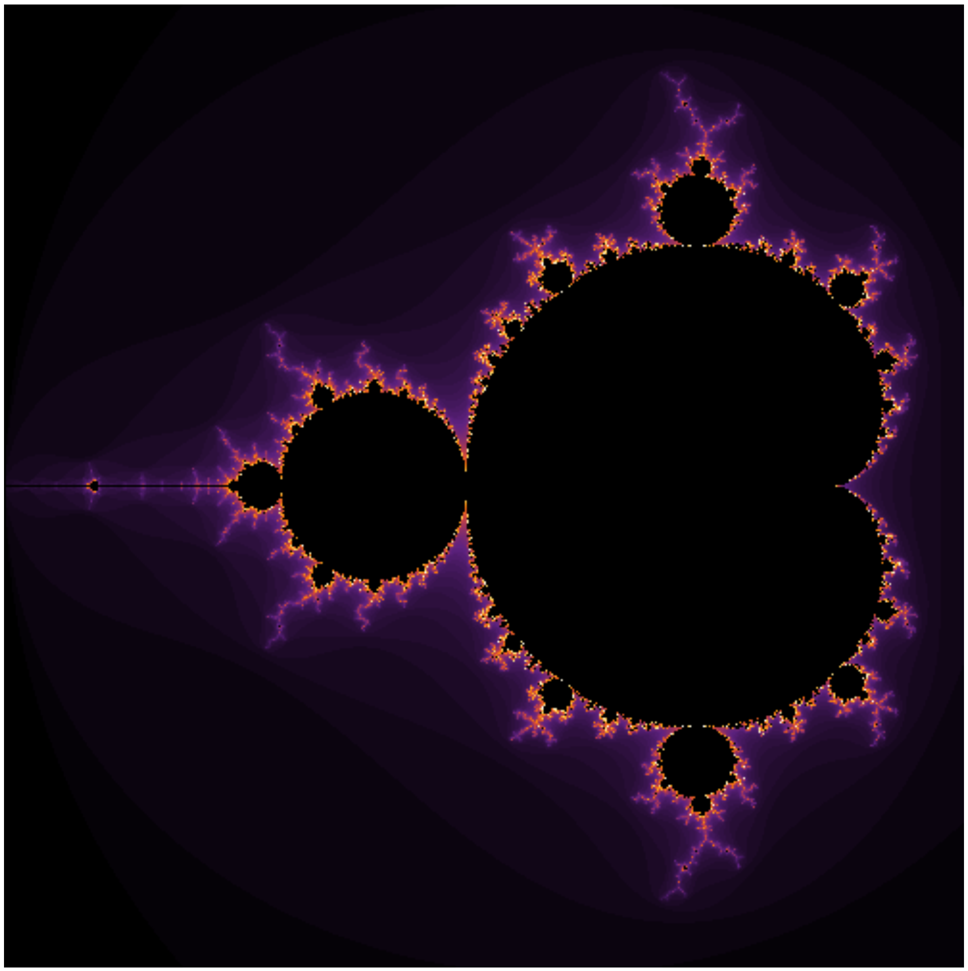
JHEP **1706**, 073 (2017) [arXiv:1704.08249]

CMS SMP-J, January 23, 2018





Complexity and Information



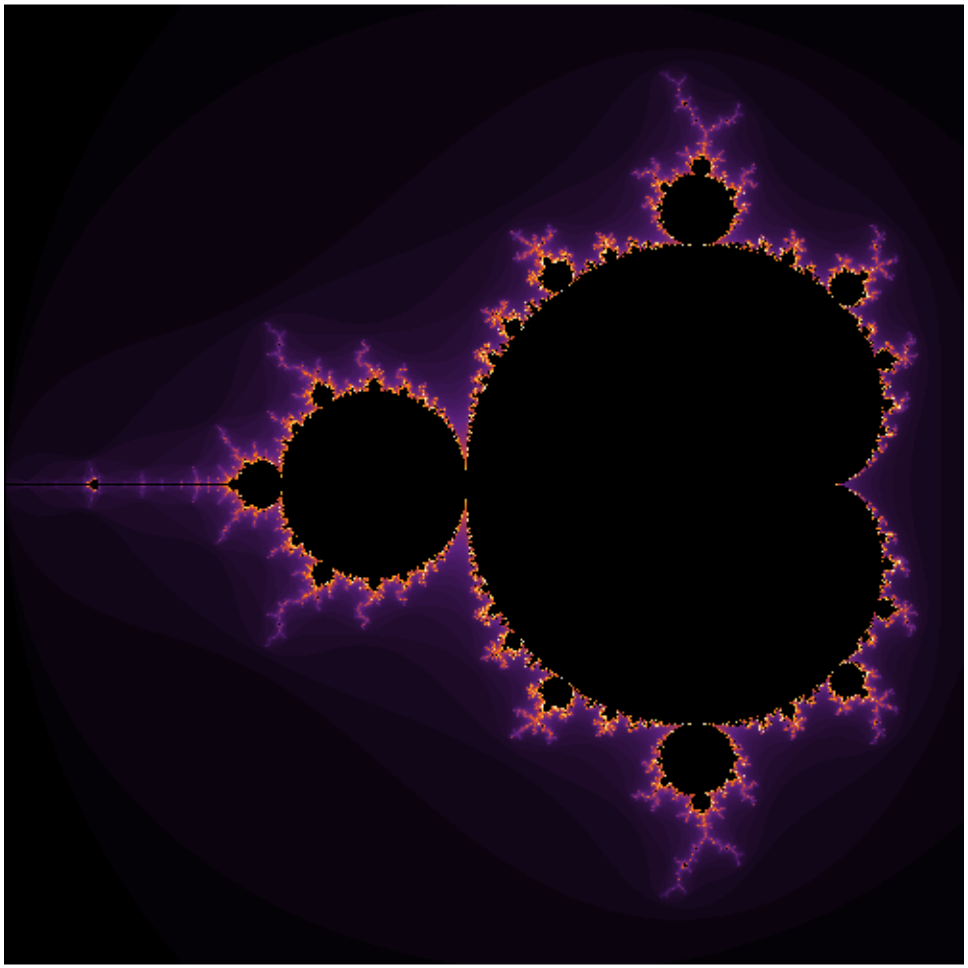
This image:

500 x 500 = 250,000 pixels

8-bit color in each pixel

Total information in bits
≈ 2 Mbits

Complexity and Information



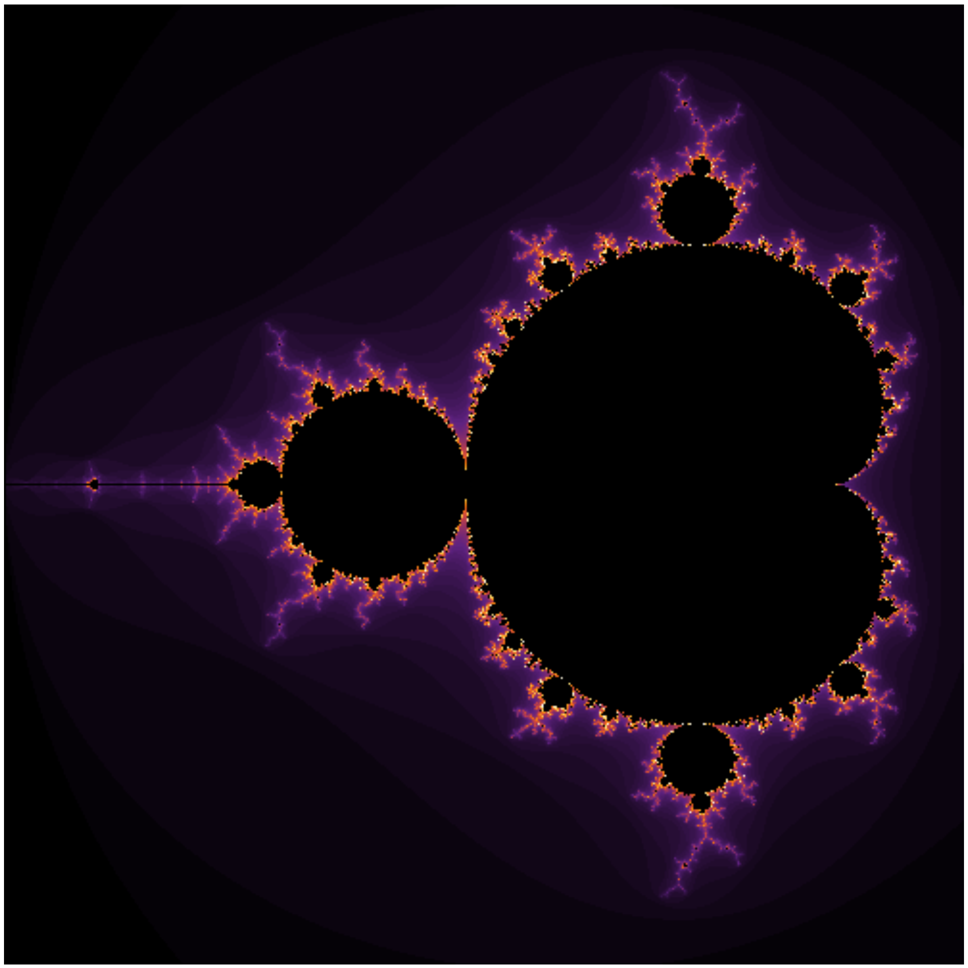
Mandelbrot set:
Defined by recursively applying

$$f(z) = z^2 + c$$

Complexity does not mean explosion of information content

Fractals can have *apparent* arbitrary complexity from simple rules

Complexity and Information

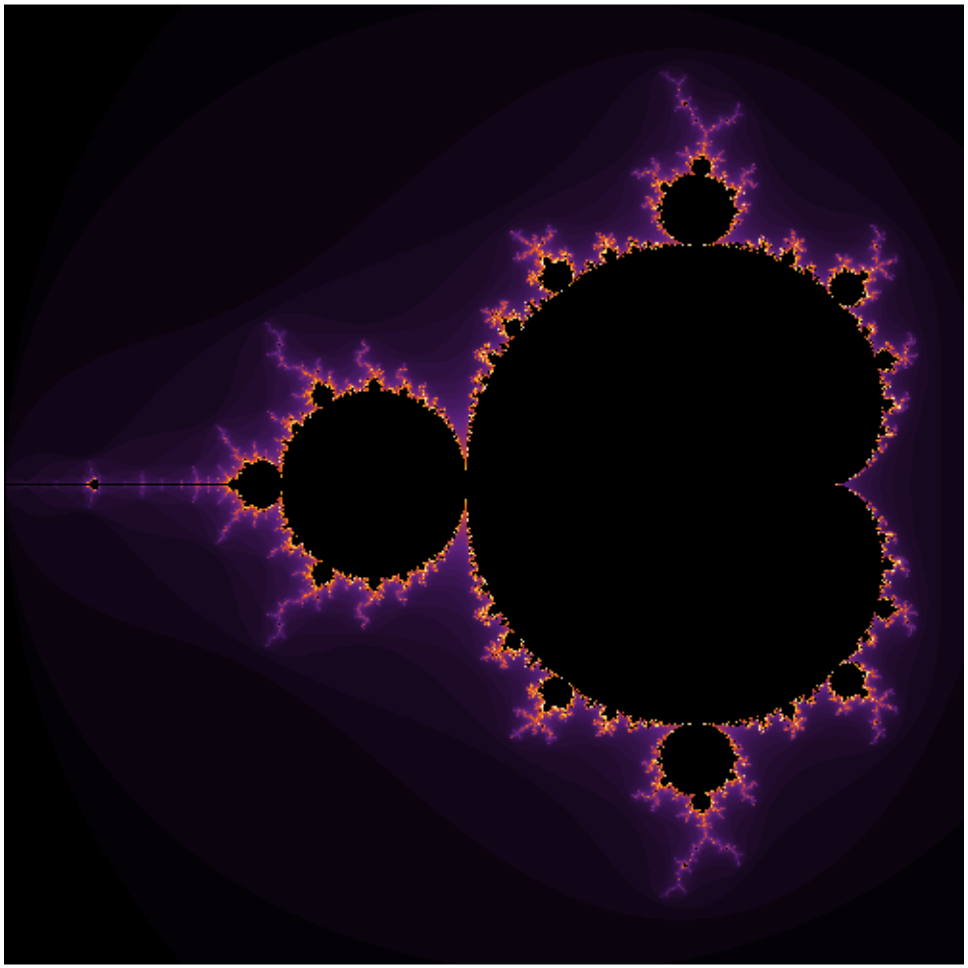


Kolmogorov Complexity:
The information of the simplest computer program that can construct the object

Example pseudo-program:

For each pixel c_i
  For $n < n_{max}$, do
    $z_0 = 0$;
    $z_{n+1} = z_n^2 + c_i$;
  Color pixel c_i from z_{nmax}

Complexity and Information



Number of bits in image:
~2 Mbits

Number of bits in program
(Kolmogorov complexity):
~100s of bits

Takeaway:
Just because something looks complex, doesn't mean it is

Caveats

I am a theoretical physicist

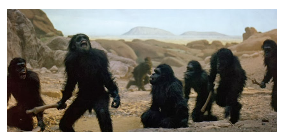
I don't know much about machine learning
(nor do I want to know much)

Motto: "What I cannot understand, I should not create."
~Feynman^{-1}

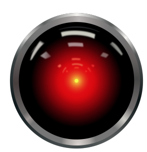
Machine Learning on Jets

My nightmare as a physicist:

Hundreds of input variables → → Mysterious output



Any organizing principle?
Can the input be simplified?
Is there any hope for a **human** to understand the output?

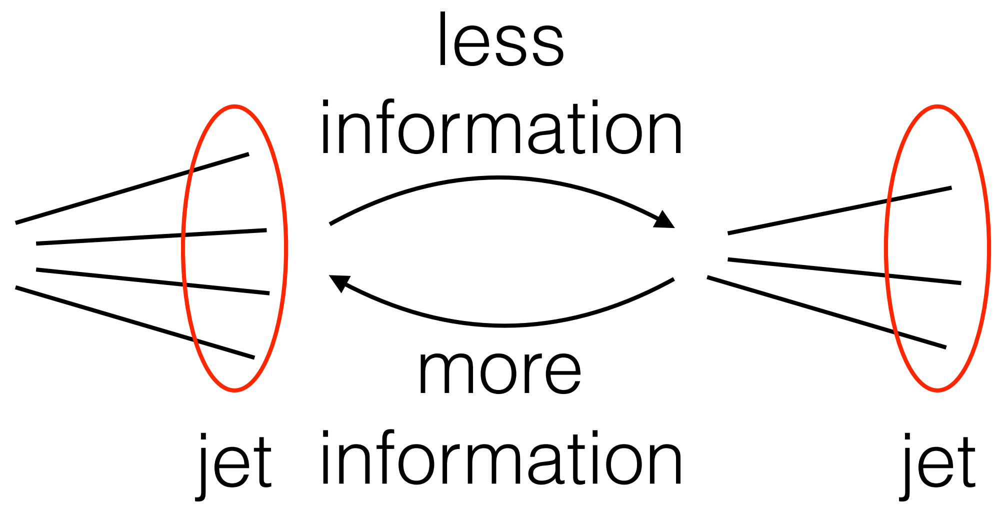


Human Learning on Jets

To make progress, use the guiding principles:

Systematic Improvability

Including more or less information in jet description is well-defined



Direct Calculability (technical)

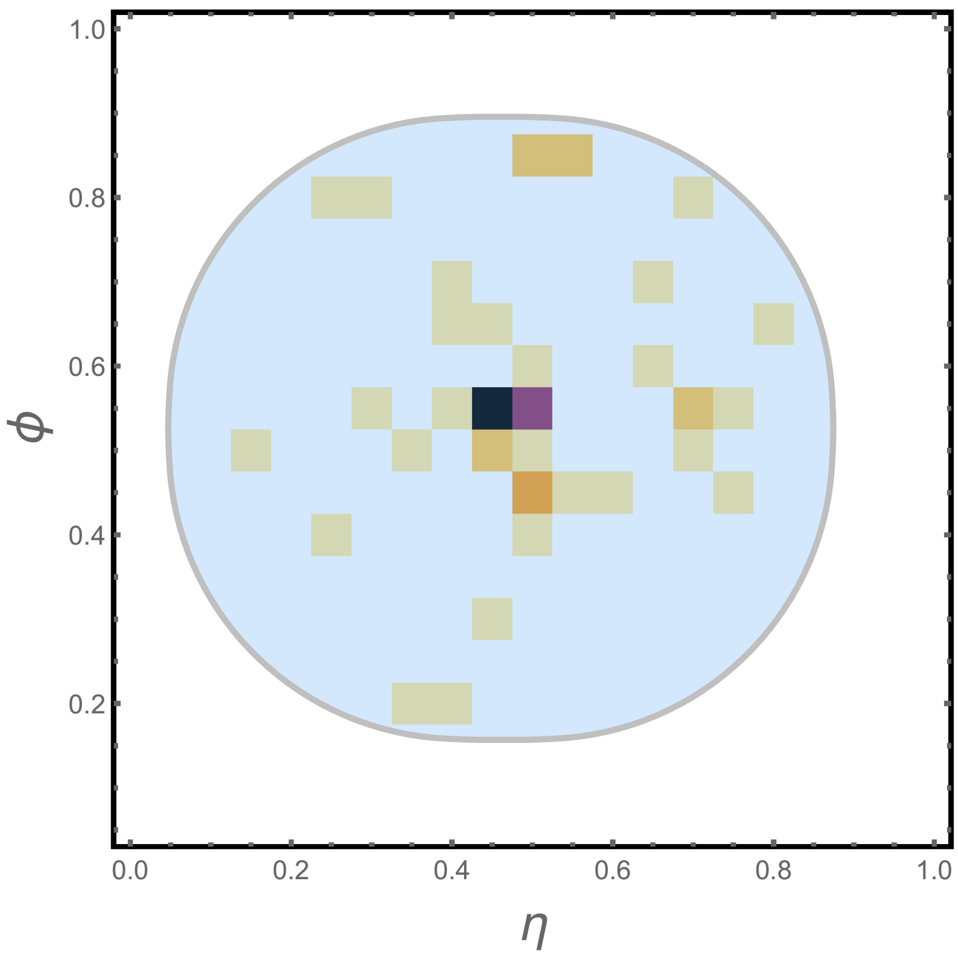
$$\tau_N^{(\alpha)} = \frac{1}{p_{TJ}} \sum_{i \in \text{jet}} p_{Ti} \min\left\{\Delta R_{i1}^{\alpha}, \ldots, \Delta R_{iN}^{\alpha}\right\}$$

Sensitive to radiation off of *N* axes in the jet

“Infrared and collinear safe”

Human Learning on Jets

Systematically resolve more structure in the jet



Full Jet

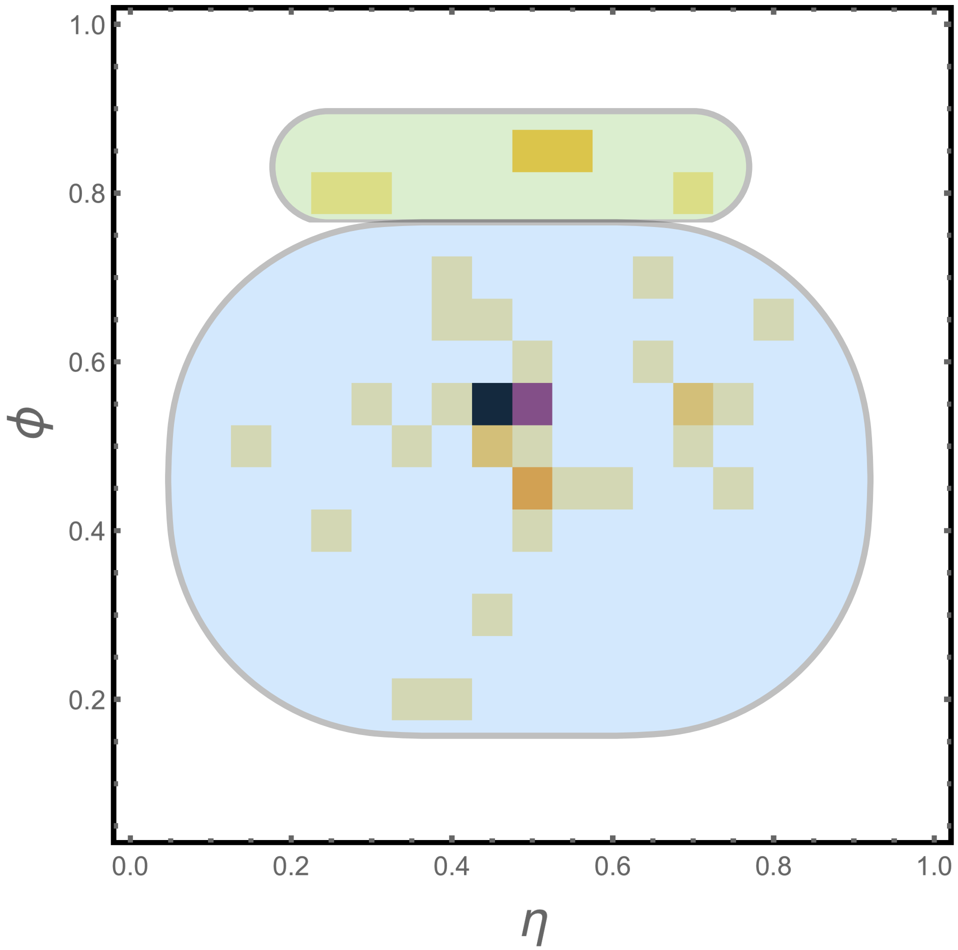
Net p_T, η, ϕ selected for

1 useful quantity:
jet invariant mass

Restrict m_J in a range about the mass of interest

Human Learning on Jets

Systematically resolve more structure in the jet



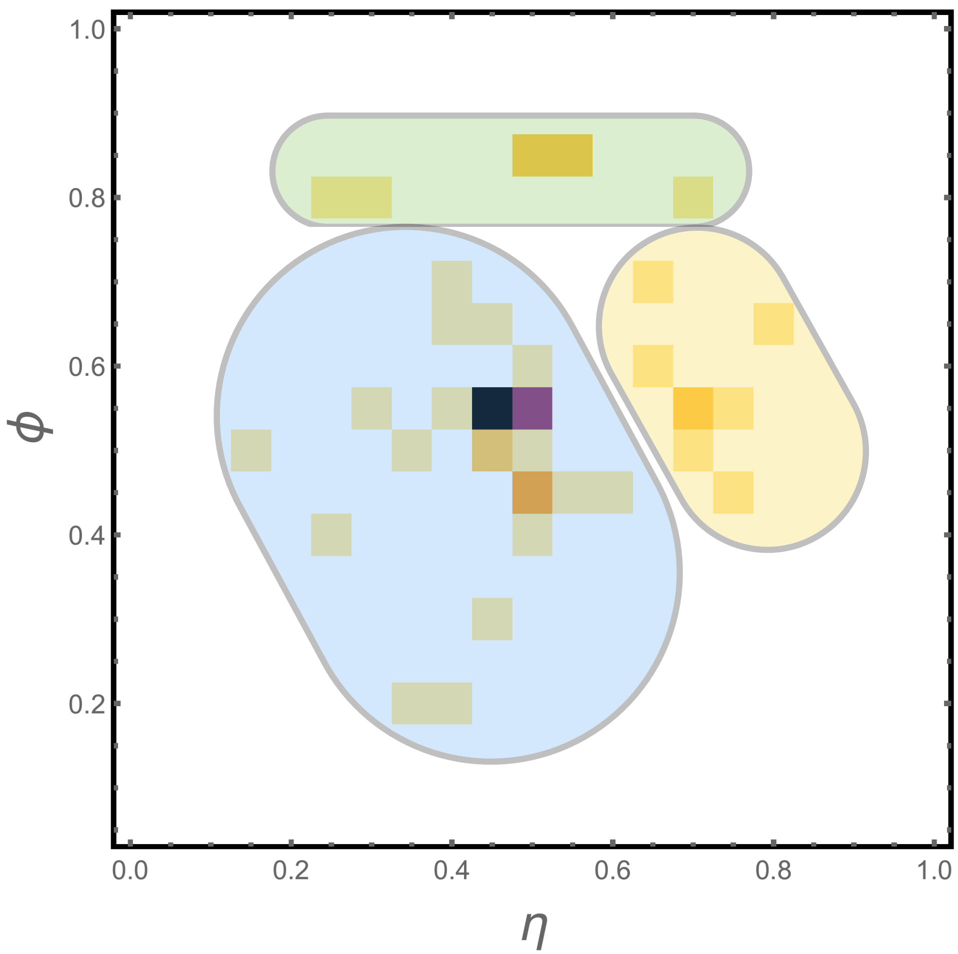
Two Subjets

Net p_T, η, ϕ, m_J selected for

2 useful quantities:
relative p_T fraction
relative angle

Human Learning on Jets

Systematically resolve more structure in the jet



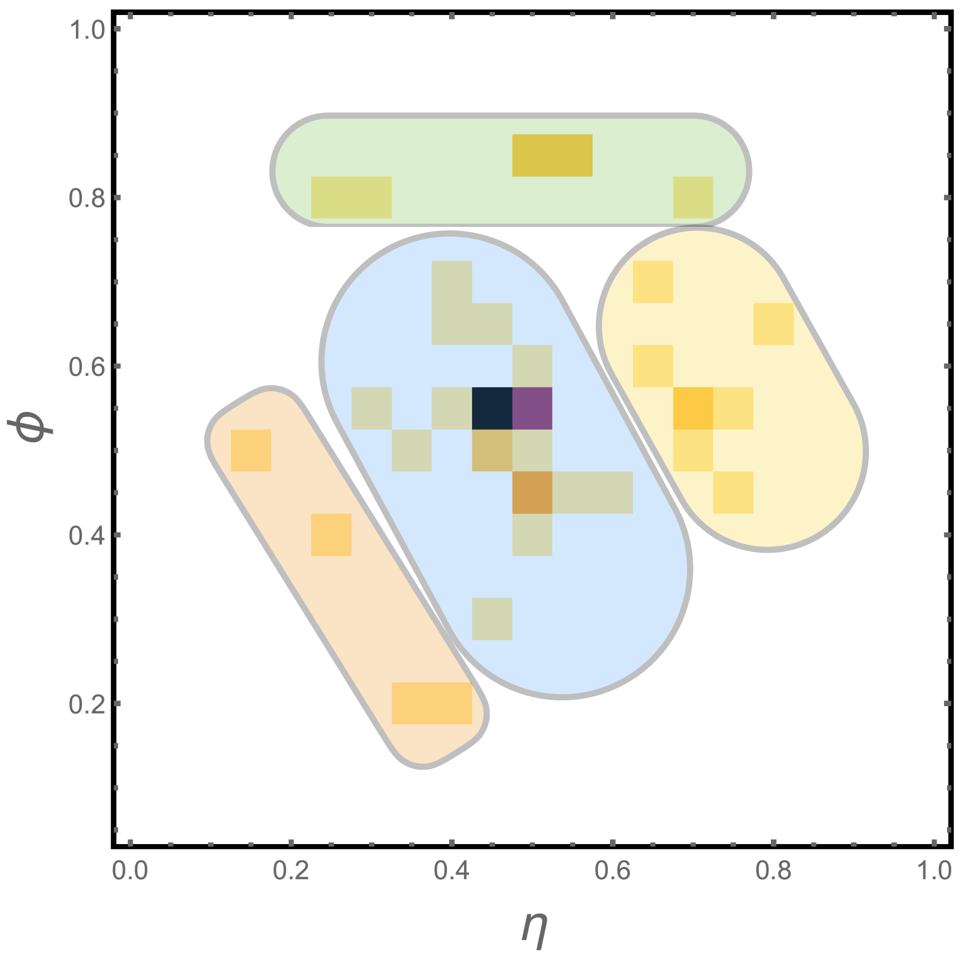
Three Subjets

Net p_T, η, ϕ, m_J selected for

5 useful quantities:
2 relative p_T fractions
3 relative angles

Human Learning on Jets

Systematically resolve more structure in the jet



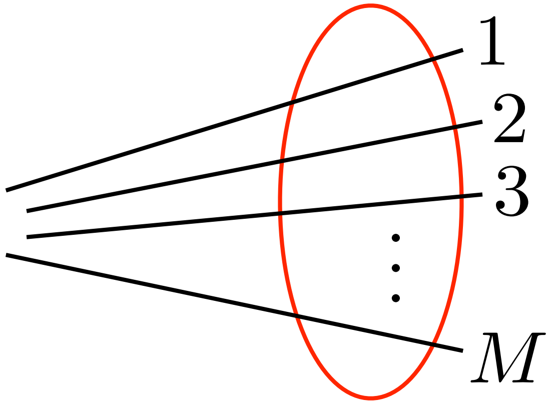
Four Subjets

Net p_T, η, ϕ, m_J selected for

8 useful quantities:
3 relative p_T fractions
5 relative angles

Can continue to resolve arbitrary structure

Human Learning on Jets



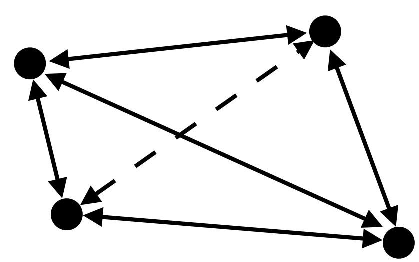
jet

Measure observables to resolve M-body phase space

$$\sigma \sim \int \prod_{i=1}^{M} \left[\frac{d^4 p_i}{(2\pi)^4} 2\pi \delta(p_i^2 - m_i^2) \right] \delta^{(4)} \left(Q - \sum_{i=1}^{M} p_i \right) |\mathcal{M}|^2$$

3M - 4 dimensional phase space

In general:
M - 1 relative p_T fractions
2M - 3 relative angles



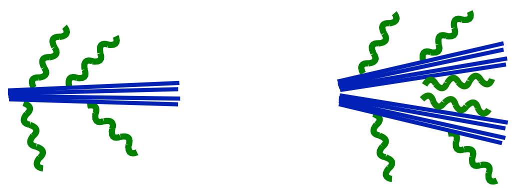
4 particle example

M-body Phase Space Machine Learning

Measure observables sensitive to
2-, 3-, 4-, 5-, and 6-body phase space + jet mass

Analyzed with a deep neural network on GPUs

Calculated ROC curves for QCD vs. Z boson



If information is finite, should see saturation

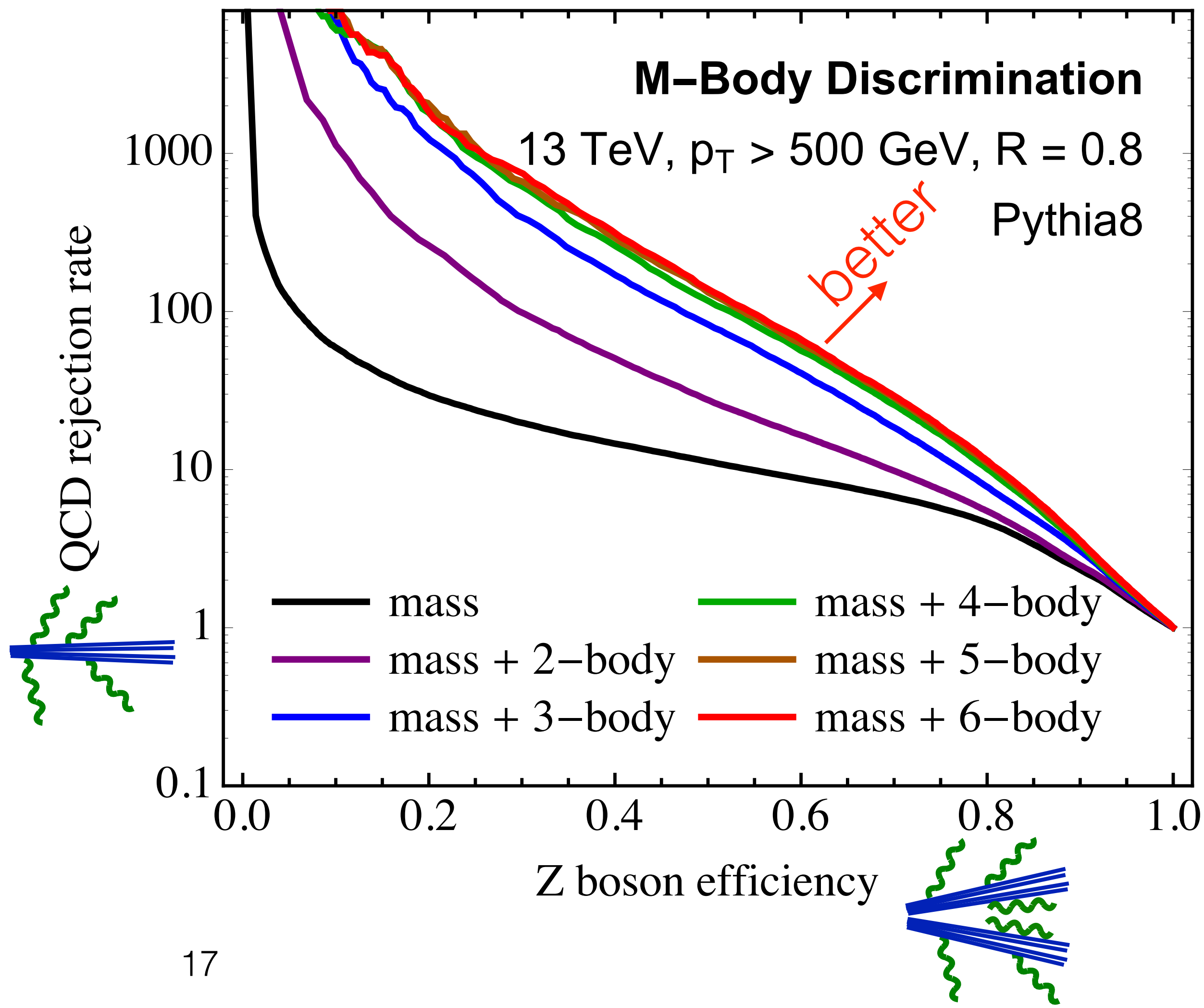
M-body Phase Space Machine Learning

Measure observables sensitive to
2-, 3-, 4-, 5-, and 6-body phase space + jet mass

Results:

Saturation observed at 4-body phase space!

4-body phase space = 8 dimensional



Why does this approach work?

Apparently there's very little information useful for discrimination

Why?

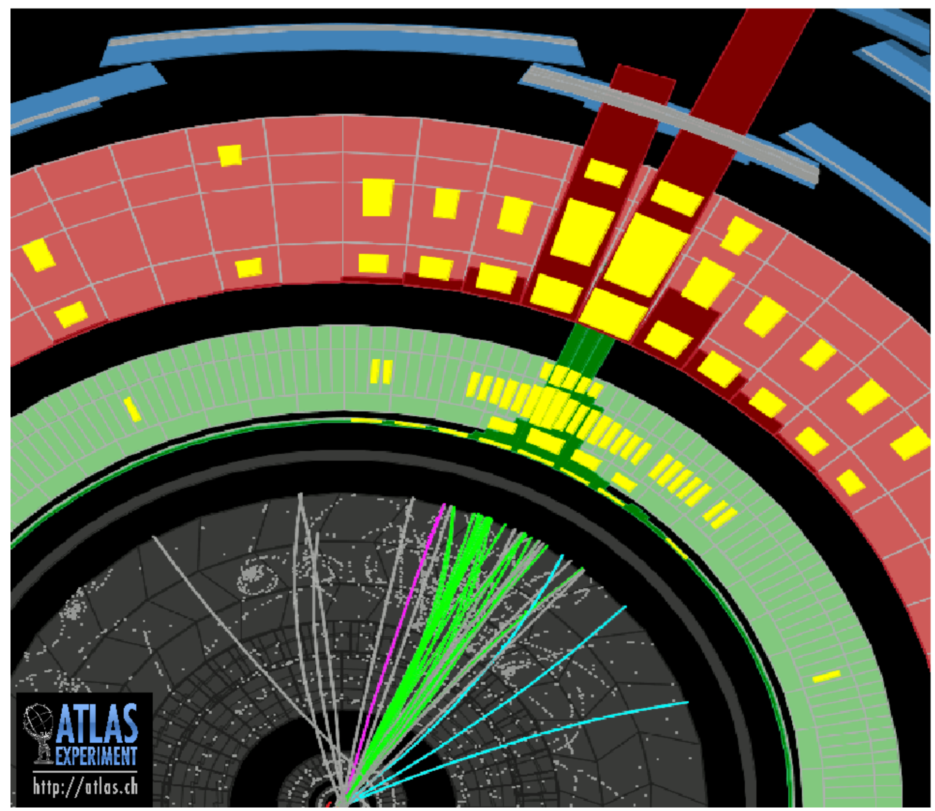
This jet has 30 particles

Information to define all particles:

$3 \times 32 \times 30 \approx 3000$ bits

(p_T,η,ϕ) — 9 digits — particles

ATLAS 2011



Why does this approach work?

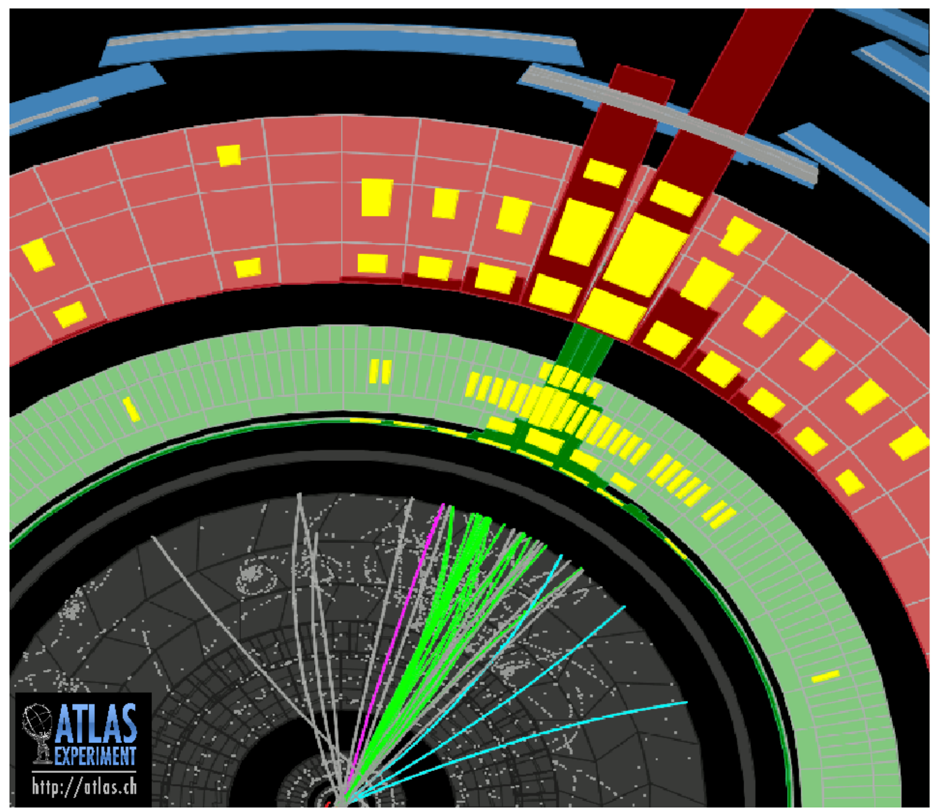
Essentially all particle production in QCD is governed by the surprisingly simple DGLAP equation:

$$Q^2 \frac{df_i(x,Q^2)}{dQ^2} = \int_x^1 \frac{dz}{z} \frac{\alpha_s}{2\pi} P_{ij\leftarrow k}\left(\frac{x}{z}\right) f_k(z,Q^2)$$

Recursive just like Mandelbrot set

Corresponding Kolmogorov complexity will be small

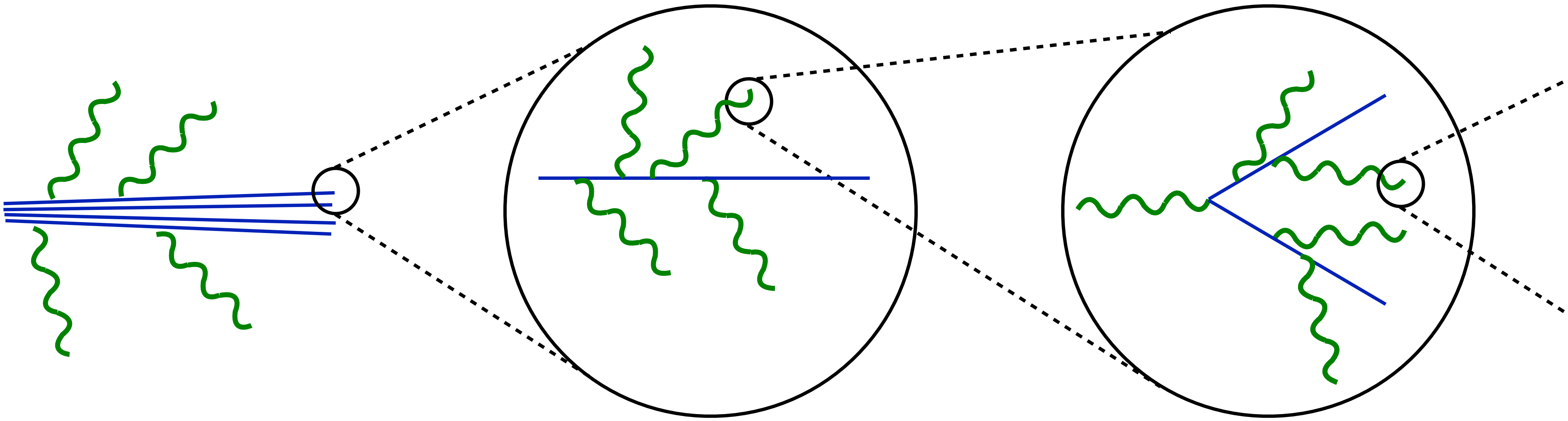
ATLAS 2011



Why does this approach work?

Essentially all particle production in QCD is governed by the surprisingly simple DGLAP equation:

$$Q^2 \frac{df_i(x,Q^2)}{dQ^2} = \int_x^1 \frac{dz}{z}\frac{\alpha_s}{2\pi} P_{ij\leftarrow k}\left(\frac{x}{z}\right) f_k(z,Q^2)$$



Seemingly-complex, fractal-like substructure of a jet

Conclusions

There isn't that much information in a jet:
particle production is recursive

Need to use techniques that exploit this feature

Resolving 4 subjets is sufficient to saturate possible QCD vs. Z boson discrimination