Interpretability and Validation of Machine Learning Models

Andrew Larkoski
Reed College

Snowmass CompF3 Workshop, August 10, 2020

Learning from a Machine

Andrew Larkoski
Reed College

Snowmass CompF3 Workshop, August 10, 2020

Caveats and Biases

Interested in Machine Learning only insofar as ***I*** learn more physics

Physicists aren’t computer scientists

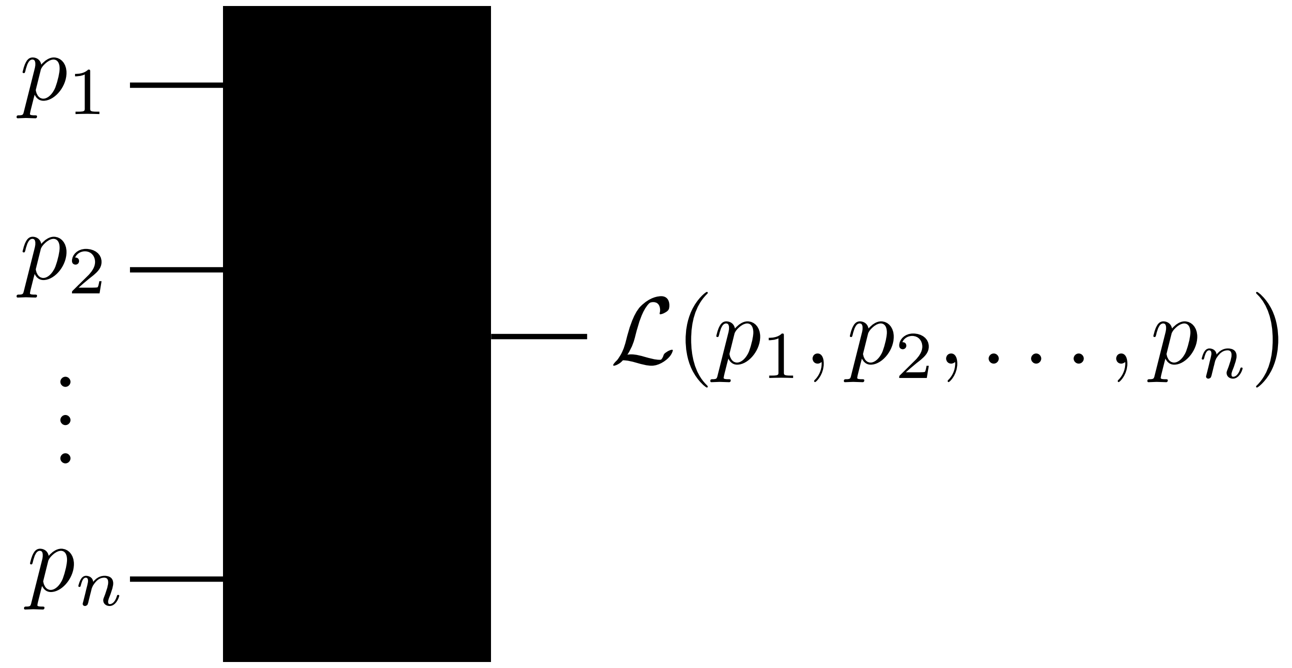
Amazing opportunity: working with a machine *requires* precise understanding of data and problem at hand

When one is restricted one is most creative: can just thinking like a machine lead to new insights?

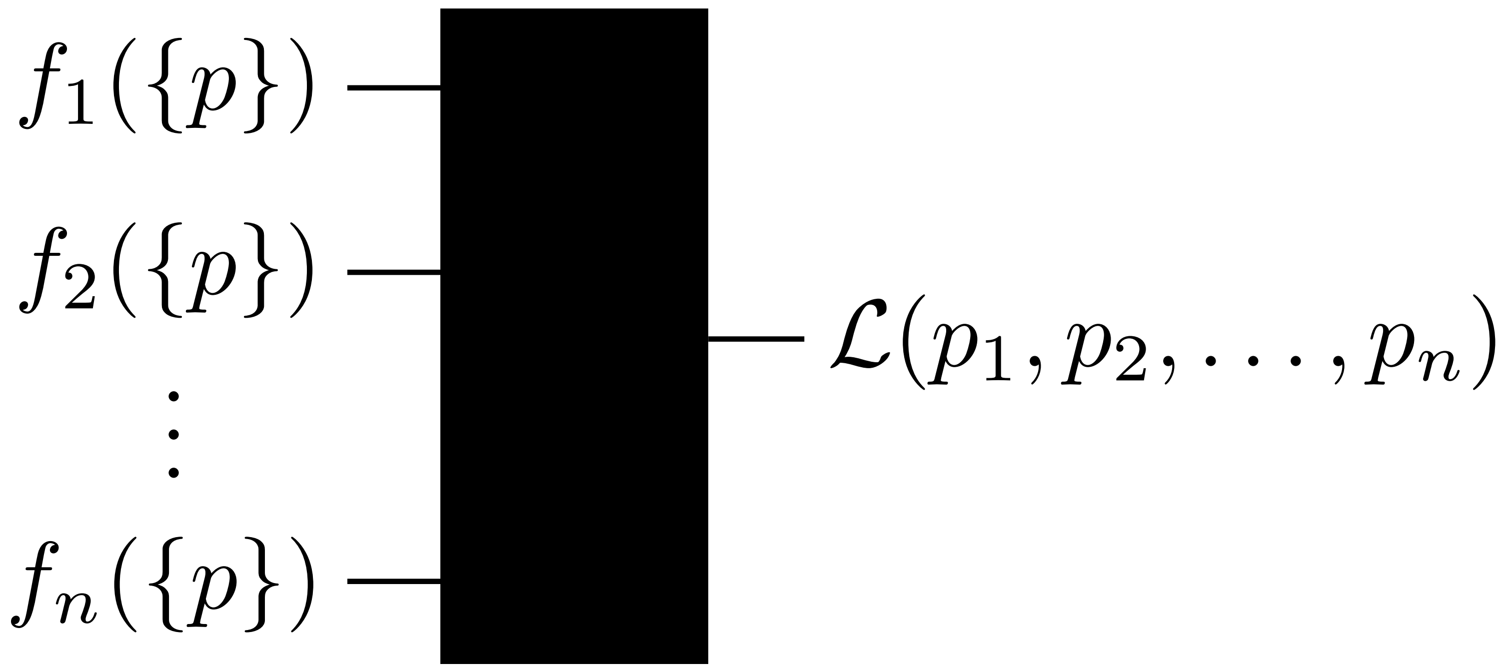
Thinking Like a Machine: Binary Discrimination

Universal Approximation Theorem

Cybenko 1989, …



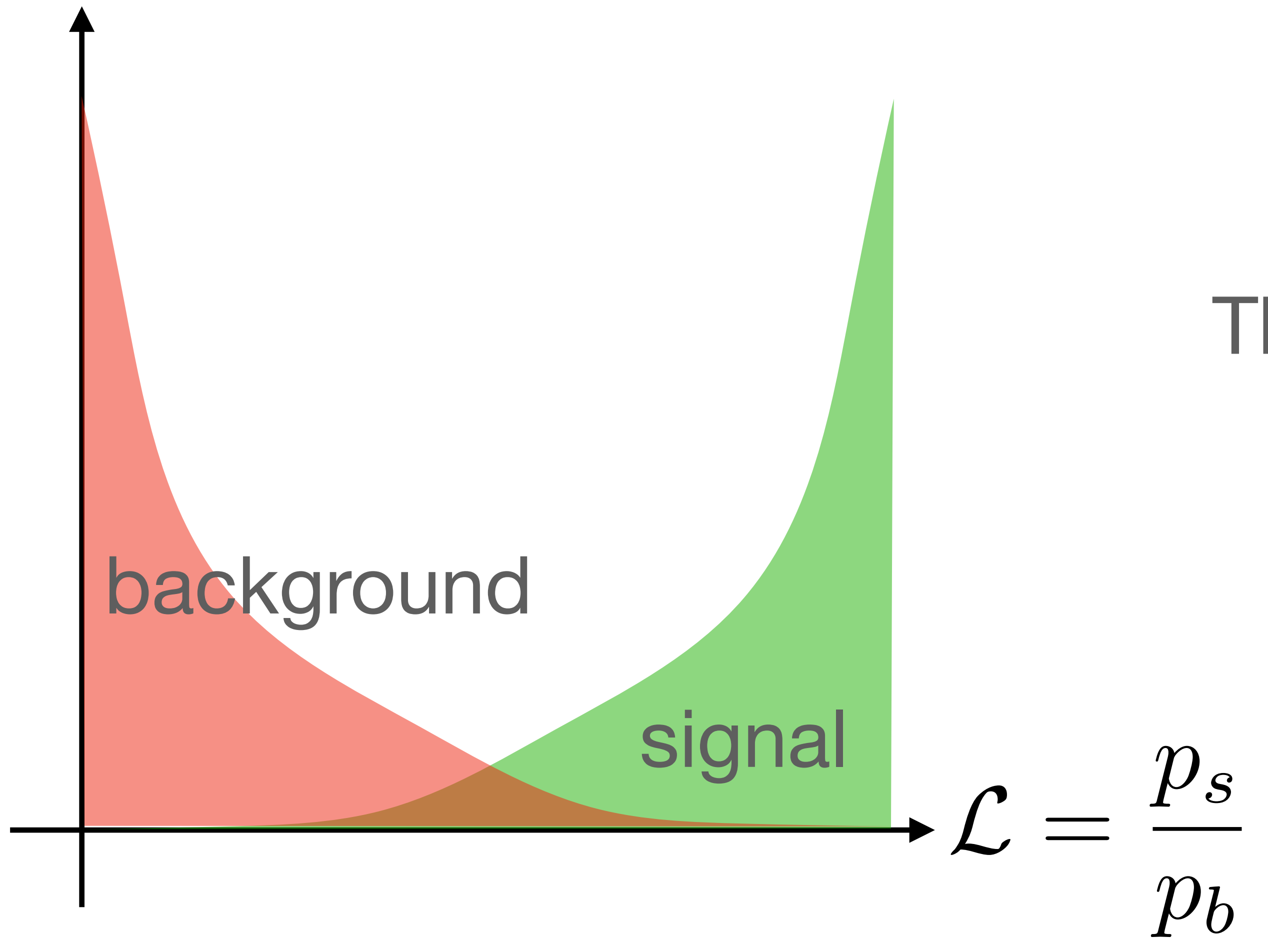
A "good" machine can output **any** function of the inputs



Work with inputs that we as human physicists understand the best

Neyman-Pearson Lemma

Neyman, Pearson 1933



The optimal discriminant is monotonic in the likelihood ratio

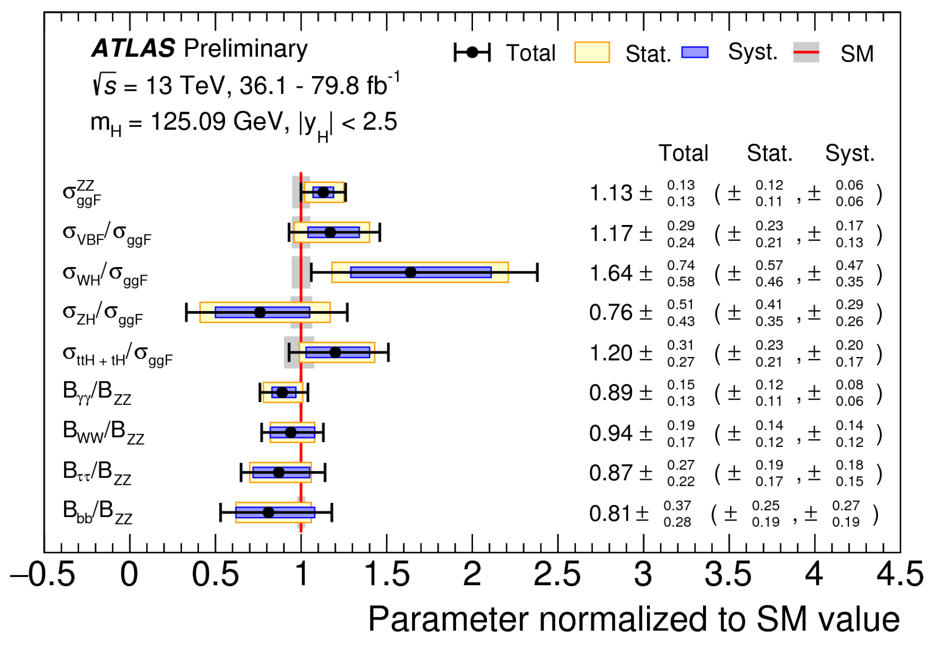
No longer an "art" to constructing observables

Don't let the machine do anything you couldn't possibly understand

Can learn a lot from reformulation of the problem

Canonical Example: Quark vs. Gluon Jets

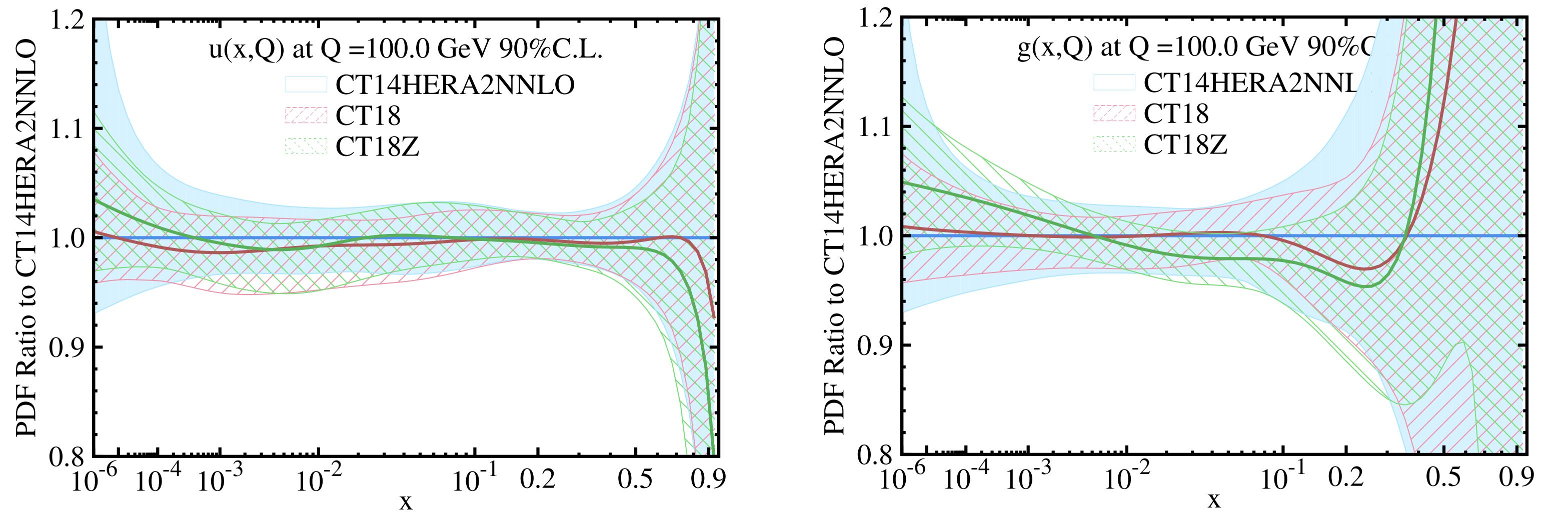
Importance in LHC Program



ATLAS-CONF-2018-031

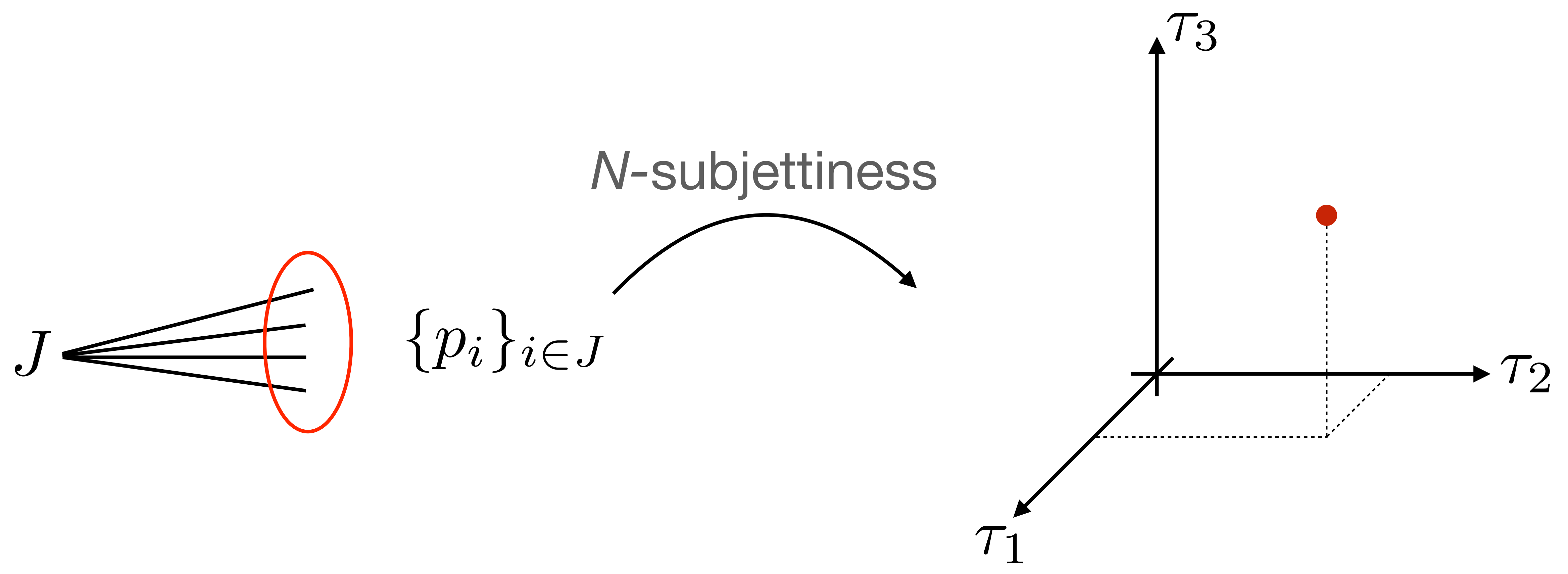
Higgs Physics
$H \to bb$ and $H \to gg$
are ~70% of total width

Gluon PDFs
Large uncertainties at large x



arXiv:1908.11394

Good Input Variables



$$\tau_N^{(\beta)} = \frac{1}{p_{TJ}} \sum_{i\in J} p_{Ti} \min\left\{R_{1i}^{\beta}, R_{2i}^{\beta}, \ldots, R_{Ni}^{\beta}\right\}$$

arXiv:1108.2701,1011.2268

In general, infrared and collinear safe observables enable theory understanding

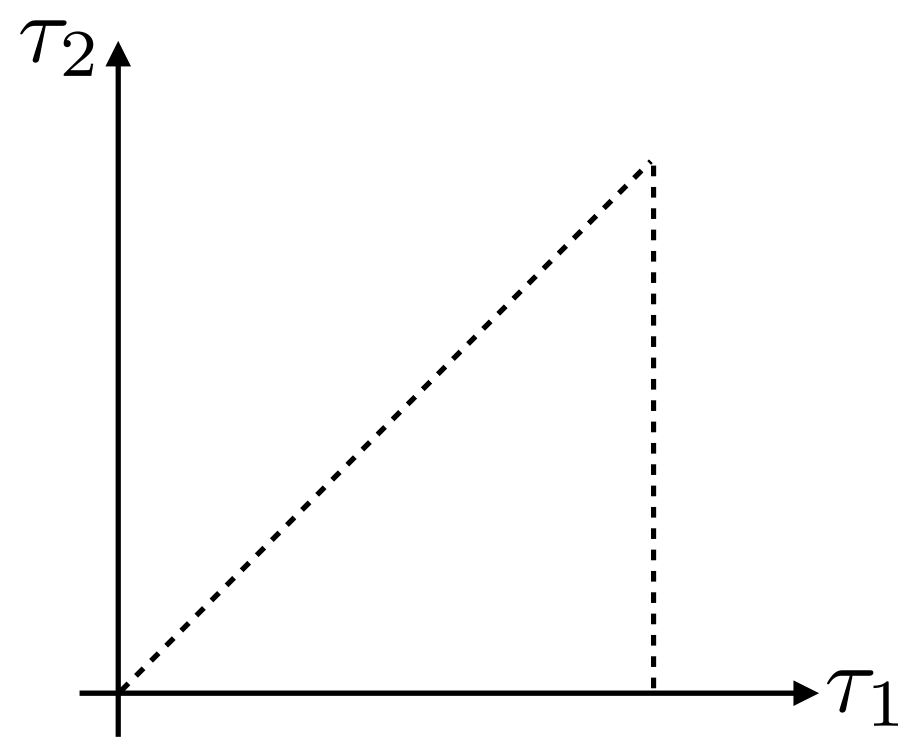
Measure a collection of *N*-subjettiness observables on jets

M-body phase space is 3*M*-4 dimensional

arXiv:1704.08249

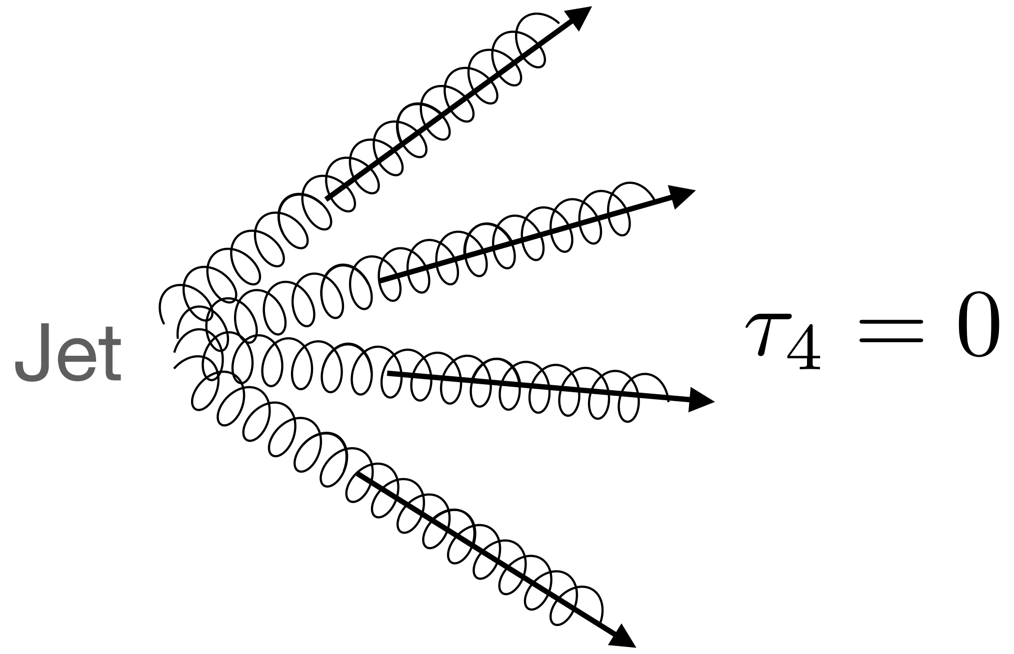
Canonical Example: Quark vs. Gluon Jets

Simplified Phase Space



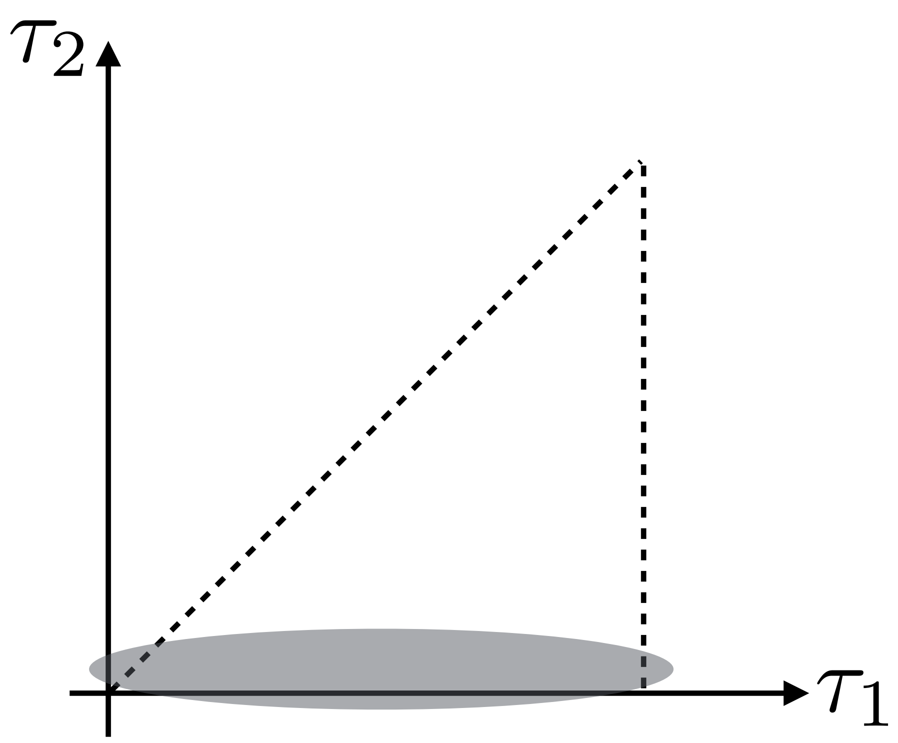
$\tau_2 < \tau_1$ as particles are always closer to one of two axes than a single axis

$\tau_2 \to 0$ limit is degenerate limit by IRC safety

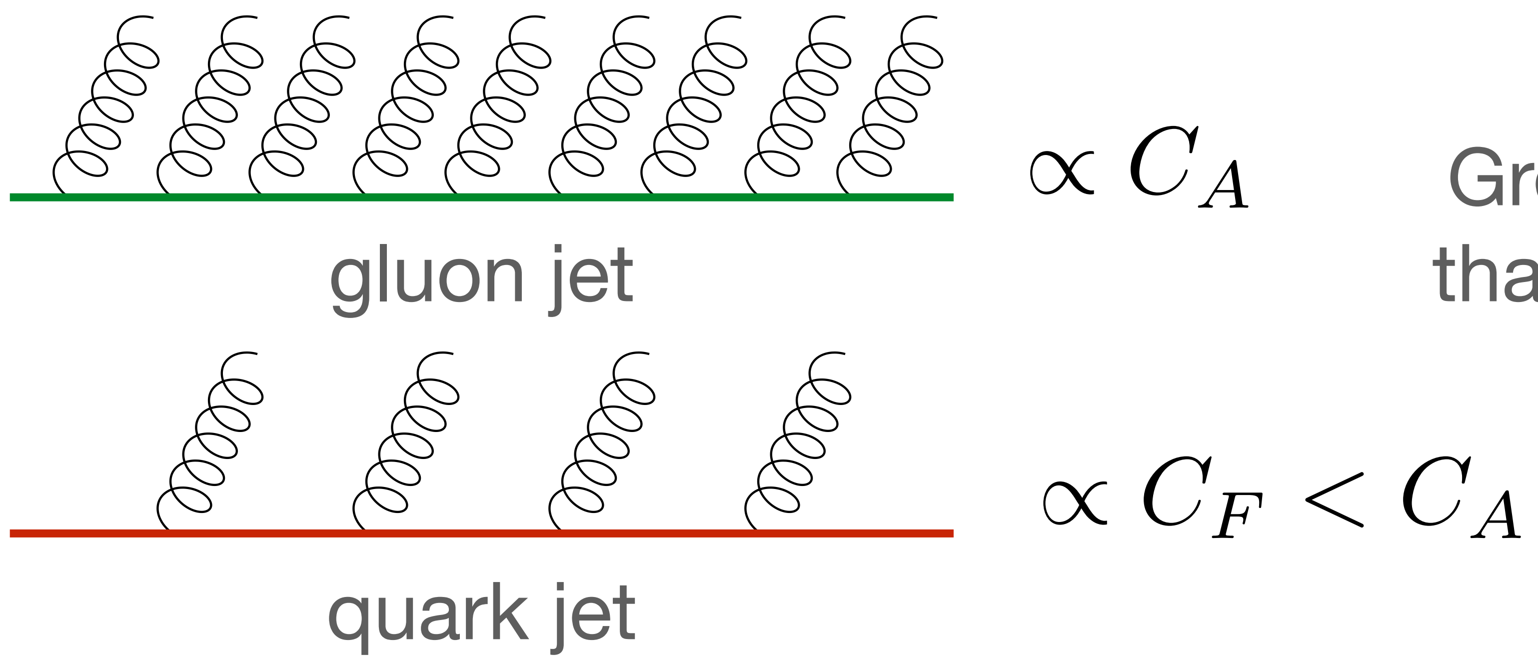


In general, $\tau_N \to 0$ limit means that *N* or fewer particles are resolved

Sudakov Suppression in IRC limit



Particle production as Poisson process exponentially suppresses singular region



Greater suppression for gluons than quarks; controlled by color Casimirs

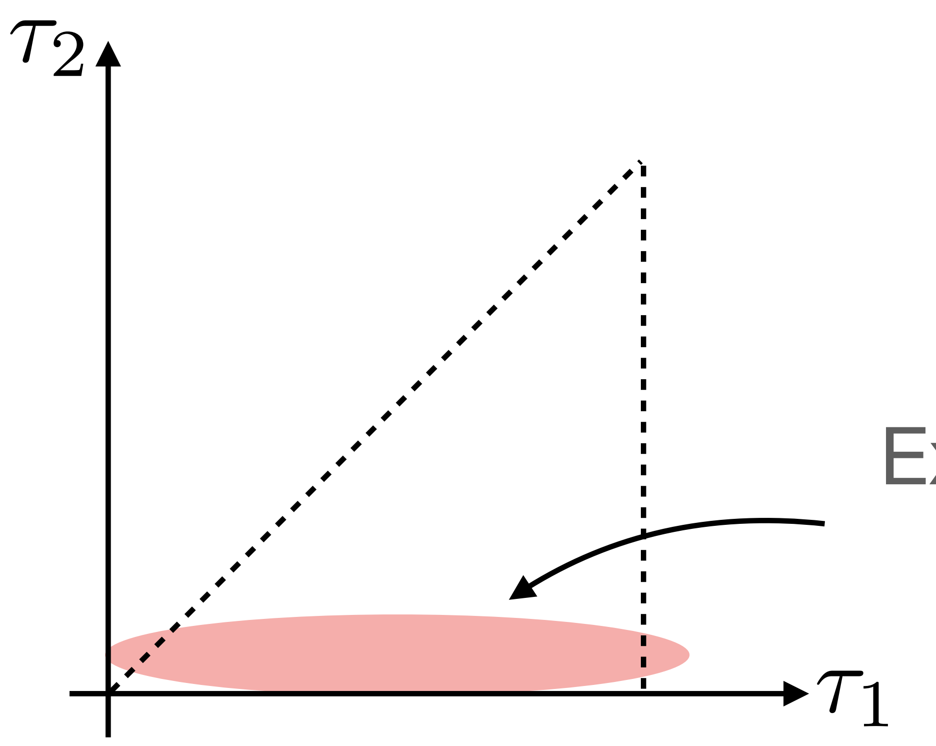
By picking a “nice” form of input data, we learn general features of the distribution

Likelihood *cannot* be an arbitrary function of inputs

arXiv:1906.01639, 2007.04319

Canonical Example: Quark vs. Gluon Jets

Likelihood Ratio



$$\mathcal{L} = \frac{p_g(\{\tau_N\})}{p_q(\{\tau_N\})}$$

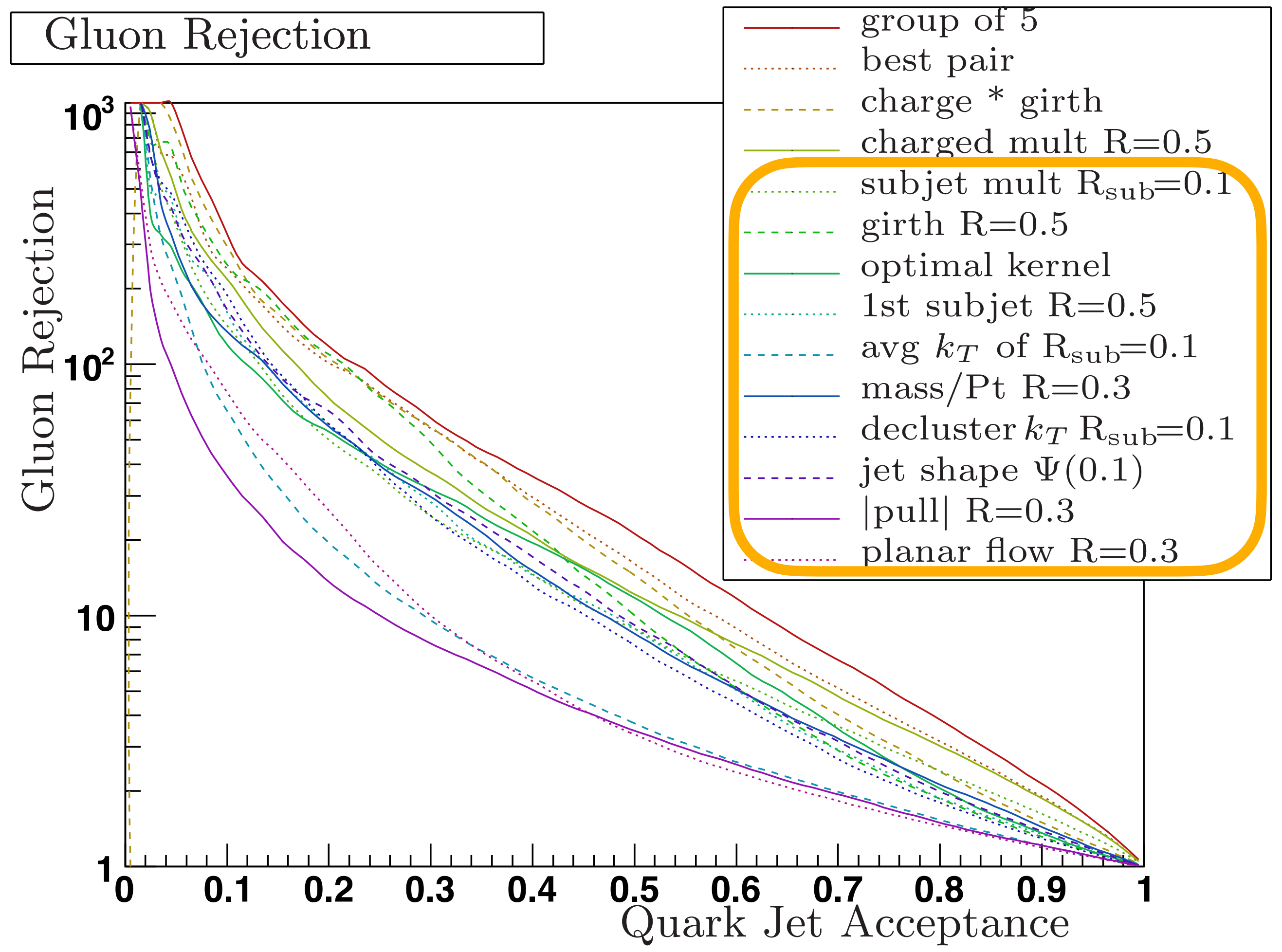
Exponentially more likely to be quark than gluon here

Entire divergent region is mapped to unique value $\mathscr{L} = 0$

Likelihood ratio for quark versus gluon discrimination is IRC safe!

Consequences

Universal approximation theorem implies that likelihood is IRC safe with **any** collection of inputs



arXiv:1106.3076

Solves long-known observation: IRC safe observables are known to be good discriminants

Closing the feedback loop: give the machine inputs to simplify its task

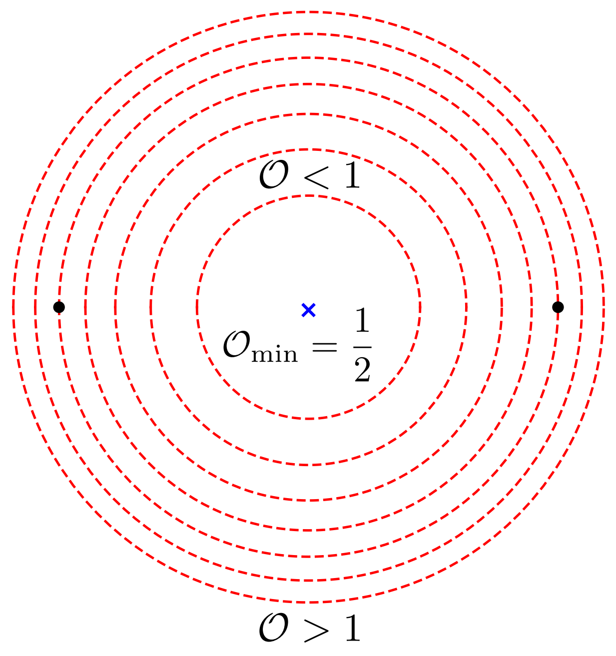
Use expert knowledge to get a real machine closer to ideal

arXiv:1906.01639, 2007.04319

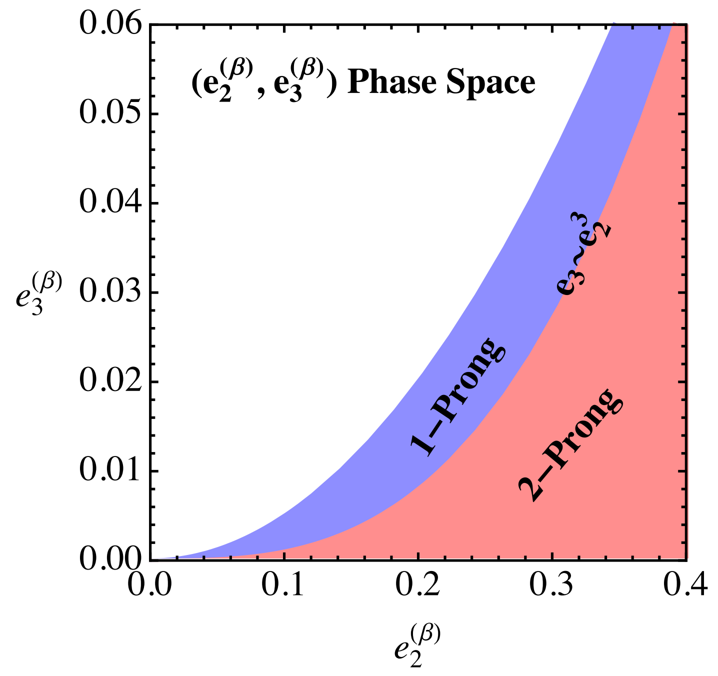
Beyond Binary Discrimination

How else can we think like a machine?

Other binary discrimination problems

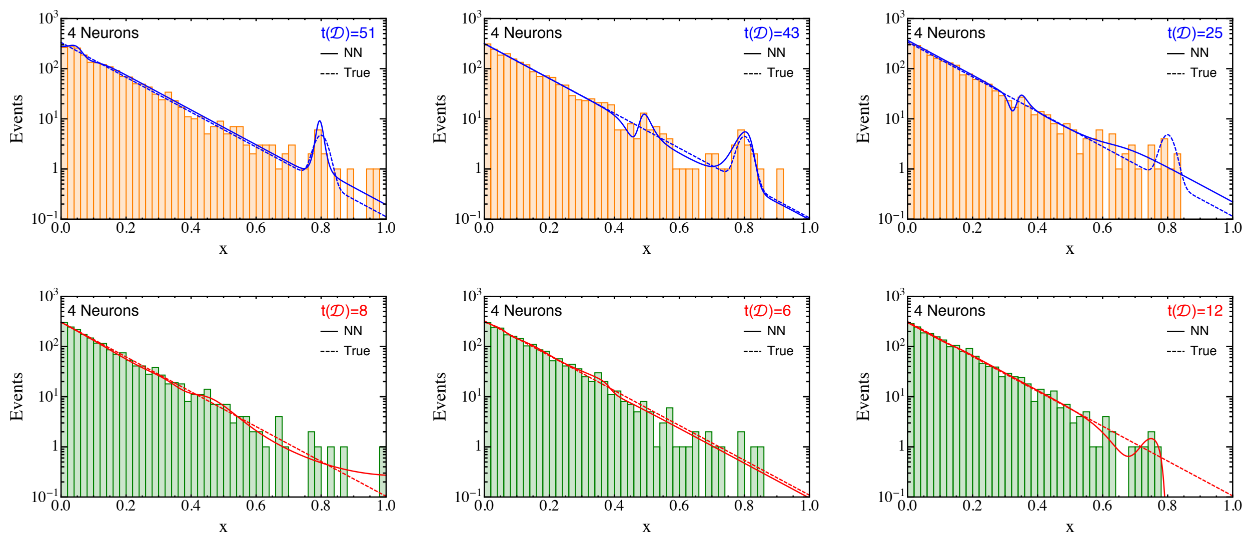


arXiv:2006.14680



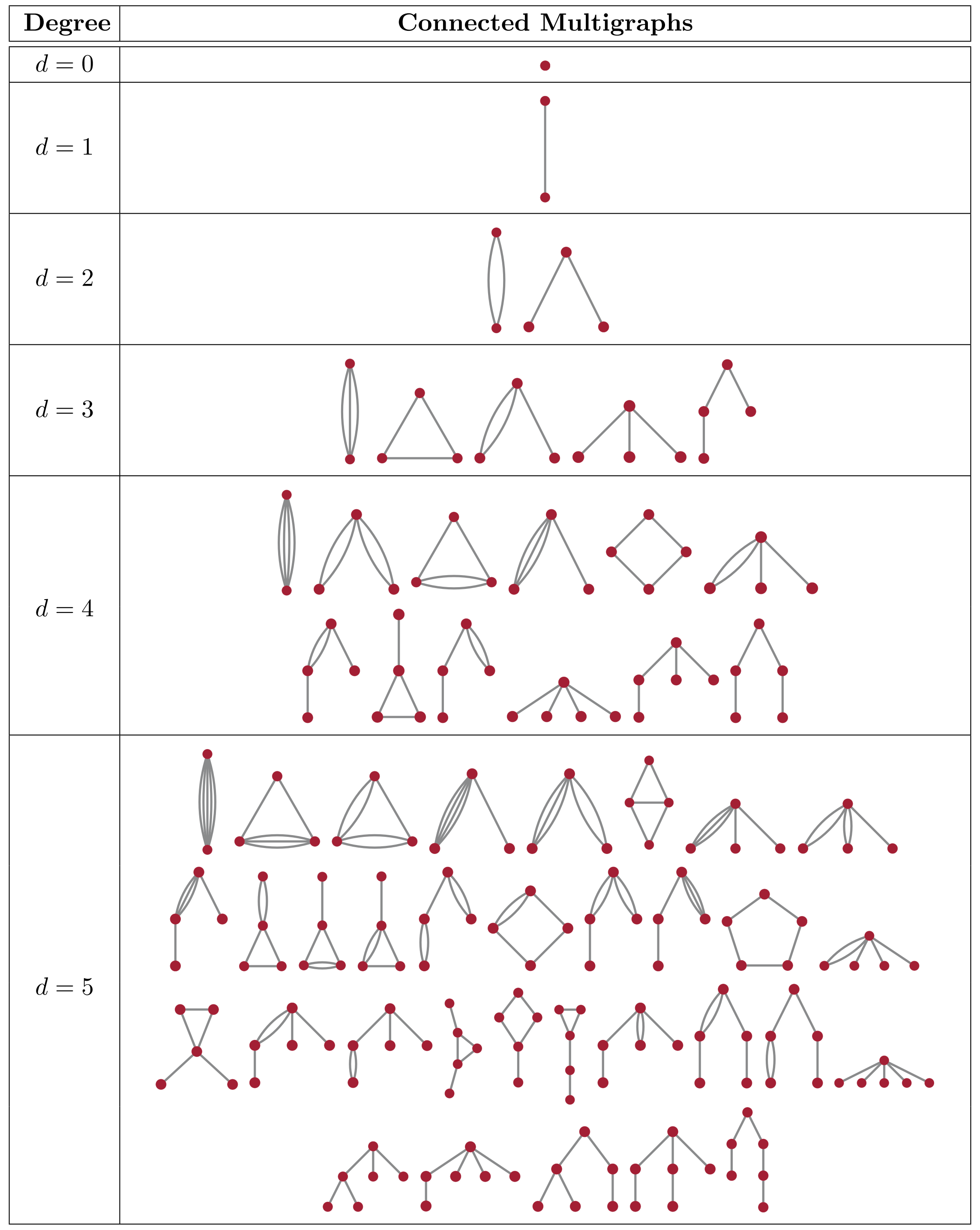
arXiv:1409.6298

Regression



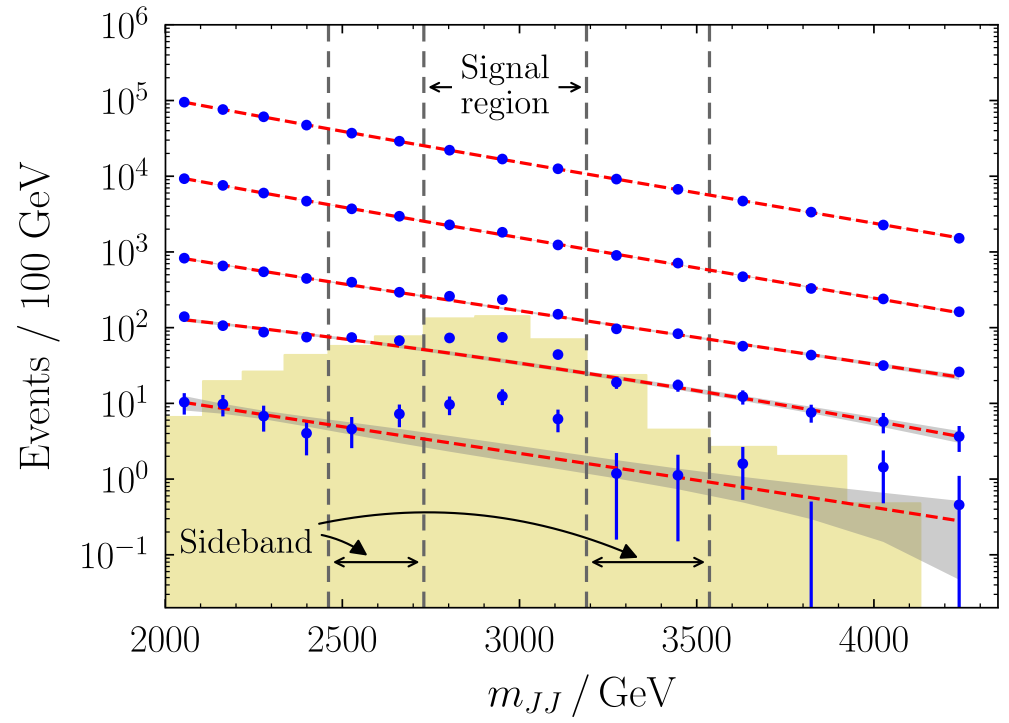
arXiv:1806.02350

(Over)complete function bases

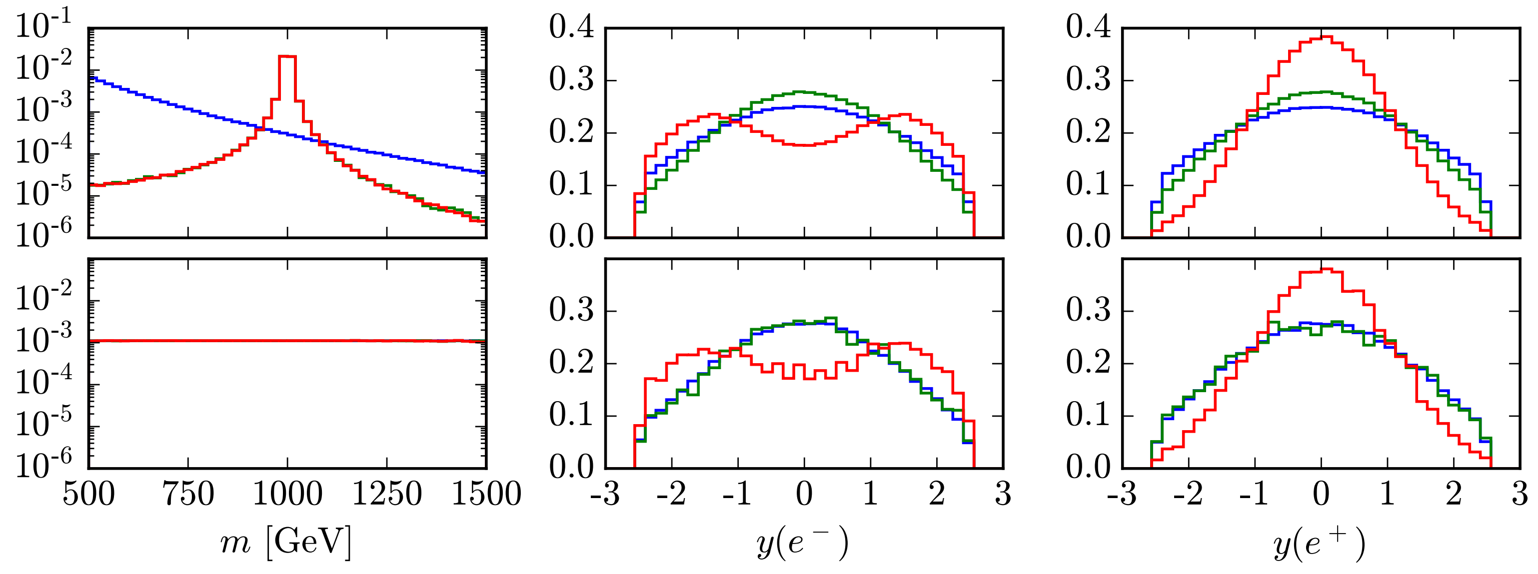


arXiv:1712.07124

Multi-label classification/ anomaly detection



arXiv:1805.02664

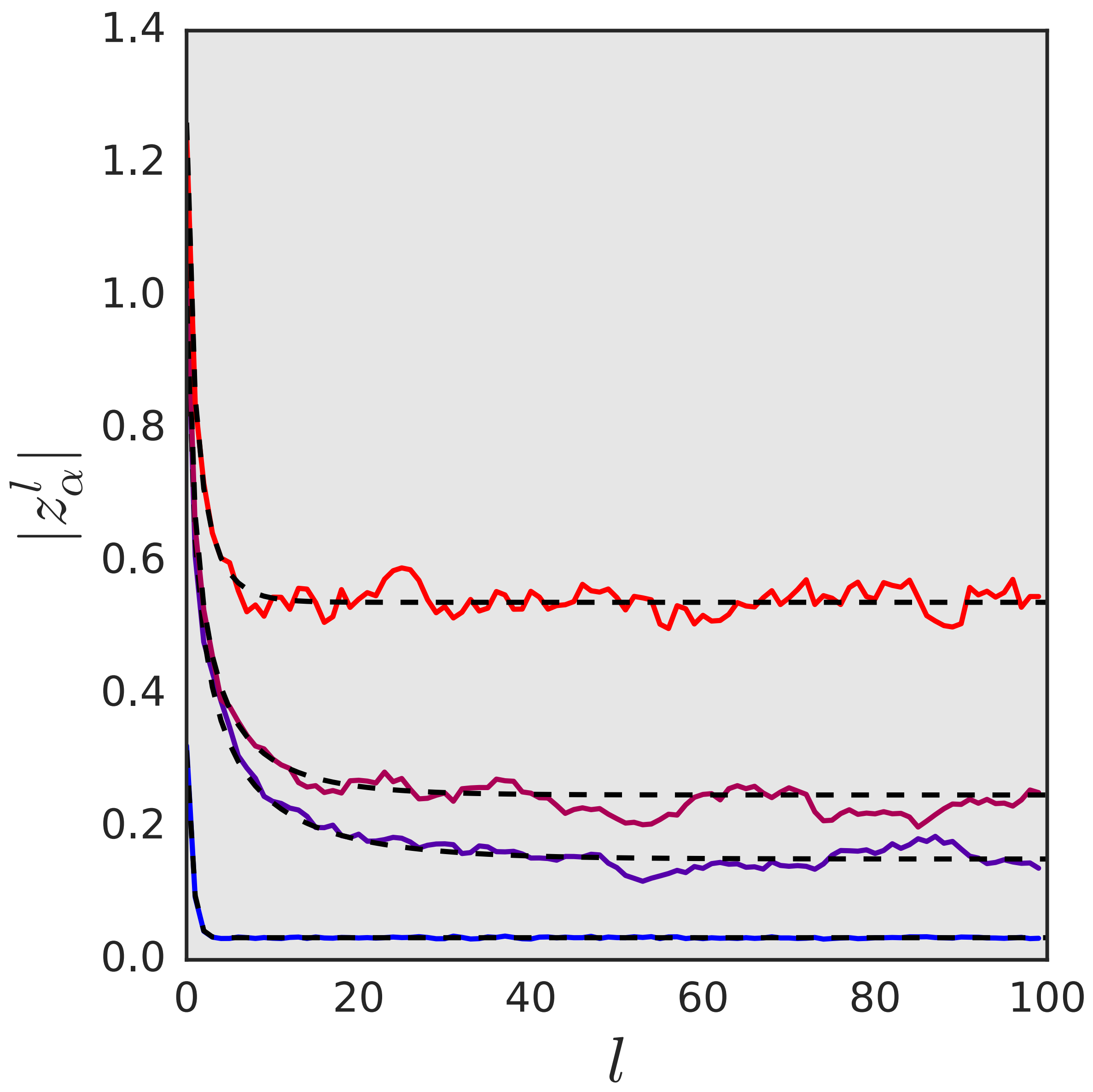


arXiv:1709.10106

Are we trusting the machine to identify physics too much or not enough?

Physicists Learning about a Neural Network

Treating a Neural Network as a Statistical System

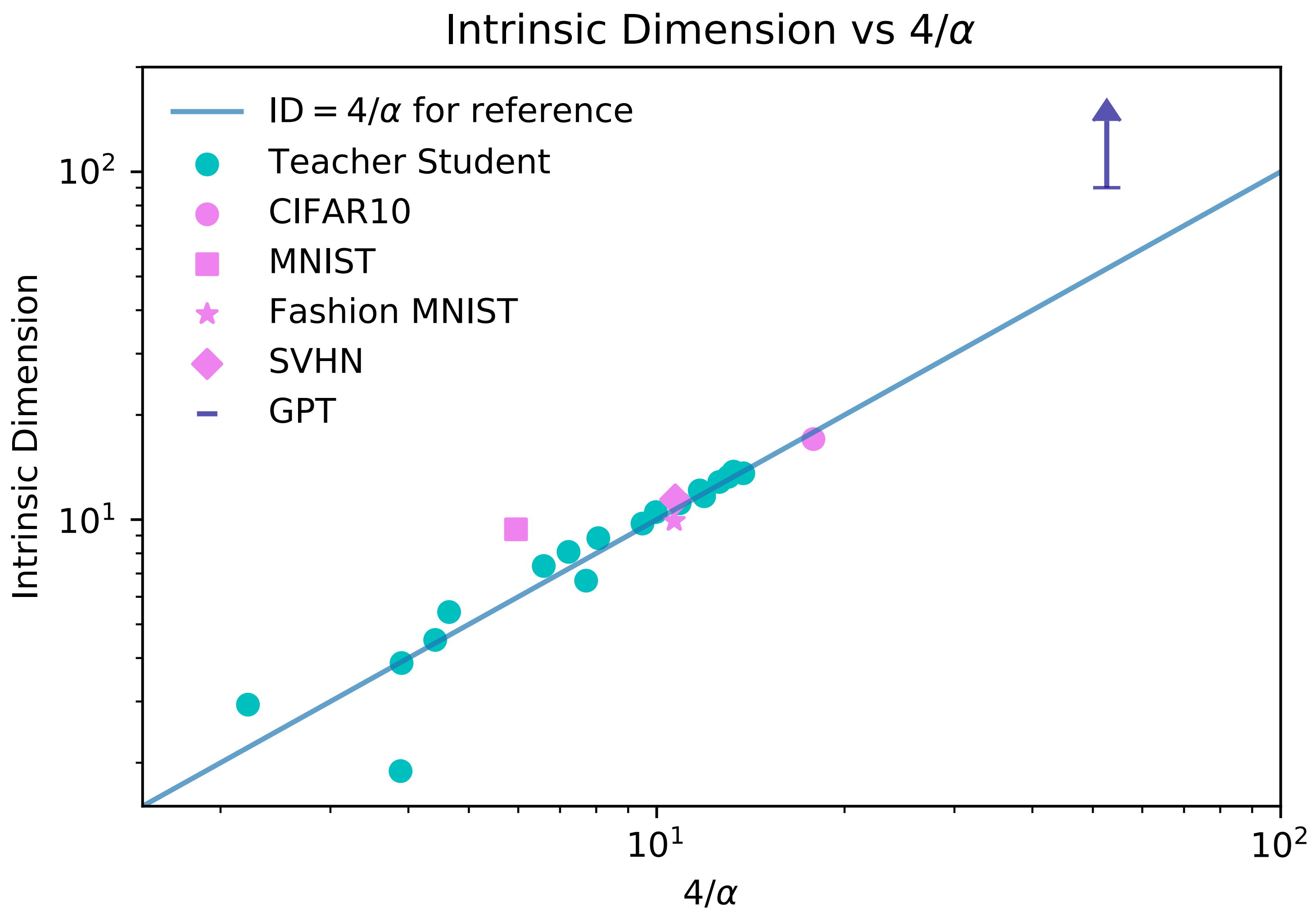


arXiv:1710.06570

Excellent agreement between mean-field theory and NN instantiation

Results from Ph.D. theorists who now work at Google

Identifying Rules for Scaling Laws in Neural Networks



arXiv:2004.10802

Critical exponents are a manifestation of universality

Results from an active physics professor and grad student

Neither example is on hep-ex, hep-ph, or hep-th!

Is this Physics, or CS and Statistics?

Where else can a machine actively teach us physics?