Representation Learning of Collider Events

Jack Collins

SLAC NATIONAL ACCELERATOR LABORATORY

BERKELEY LAB

How Much Information is in a Jet?

Kaustuv Datta and Andrew Larkoski

Physics Department, Reed College, Portland, OR 97202, USA

Menu

(Absolutely no substitutions)

Aperetif
How much information is in a jet?

Appetizer
Autoencoder Introduction

Fish Course
The Metric Space of Collider Events

Main Course
The Variational Autoencoder:
a pedagogical introduction

Cheese Selection
Application to top jets

Dessert
Mystery Special

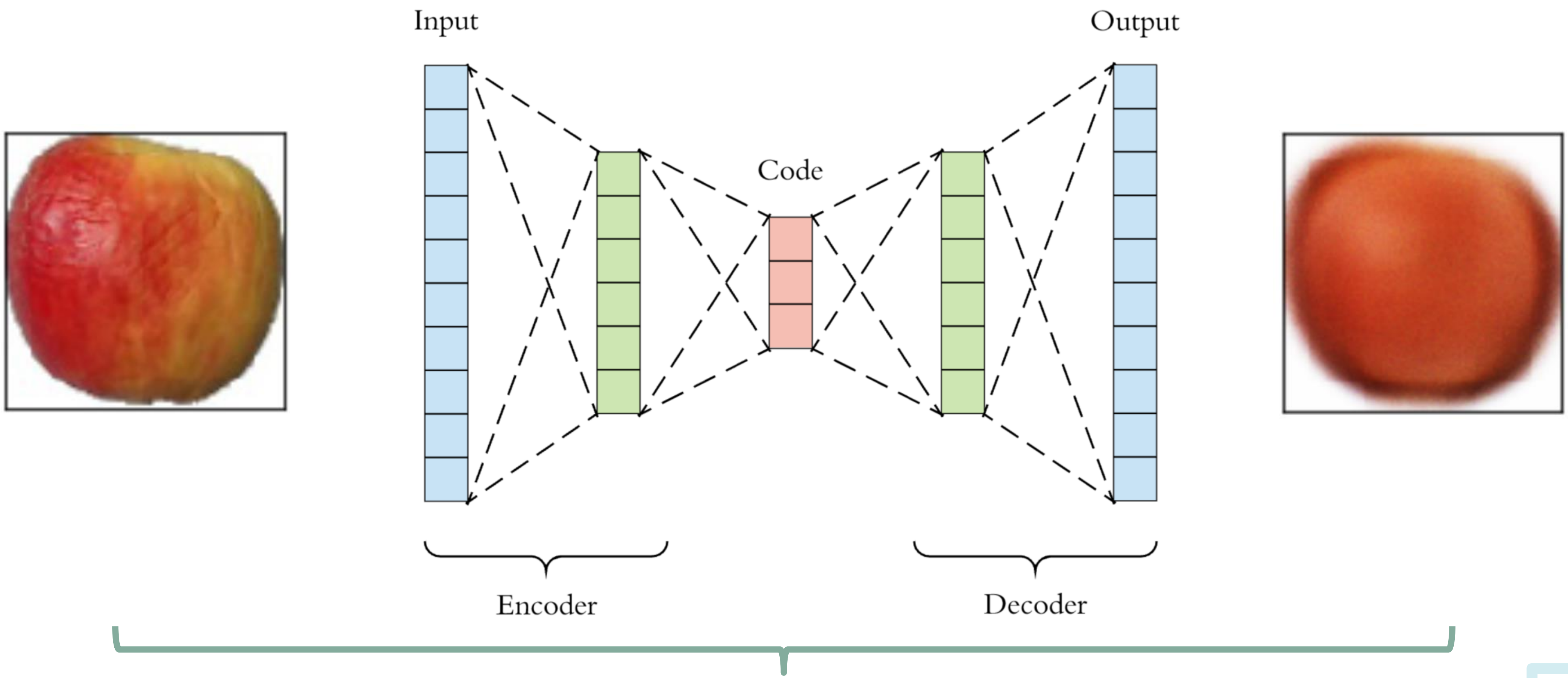
Palate Cleanser
Conclusions

Digestif
A sketchy derivation

Appetizer

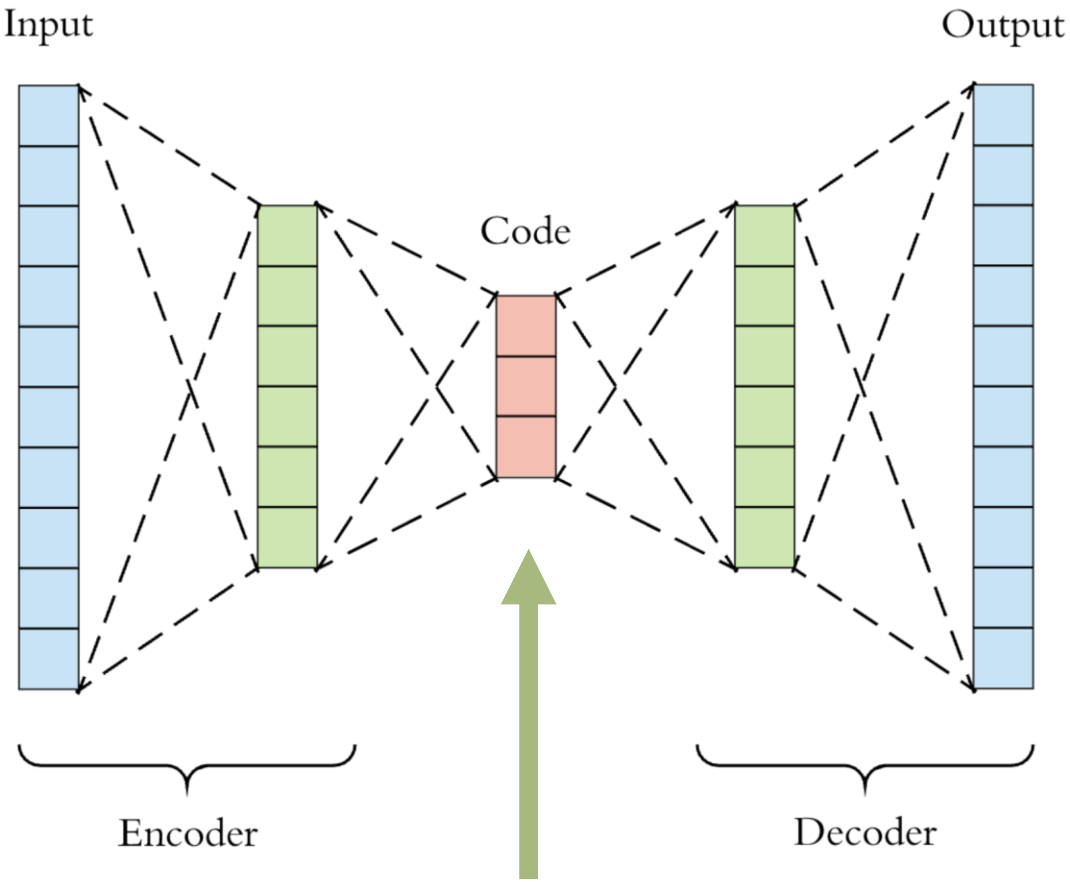
Autoencoder introduction

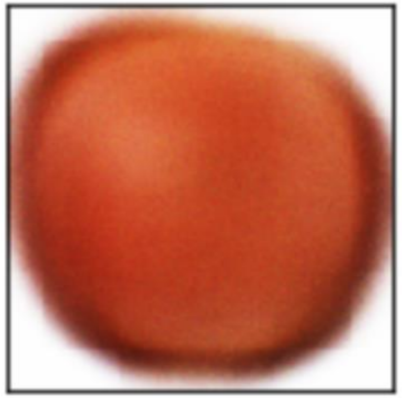
The Plain Autoencoder



Loss = |Output – Input| *(what is this for jets?)*

The Plain Autoencoder





Latent space =?= Learnt representation

Fish Course

The Metric Space of Collider Events

Some of the next few slides are taken directly from a talk by Jesse Thaler at SLAC in 2019, http://www.jthaler.net/talks/jthaler_2019_04_SLAC_EMD.pdf

The Space of Collider Events

Jesse Thaler

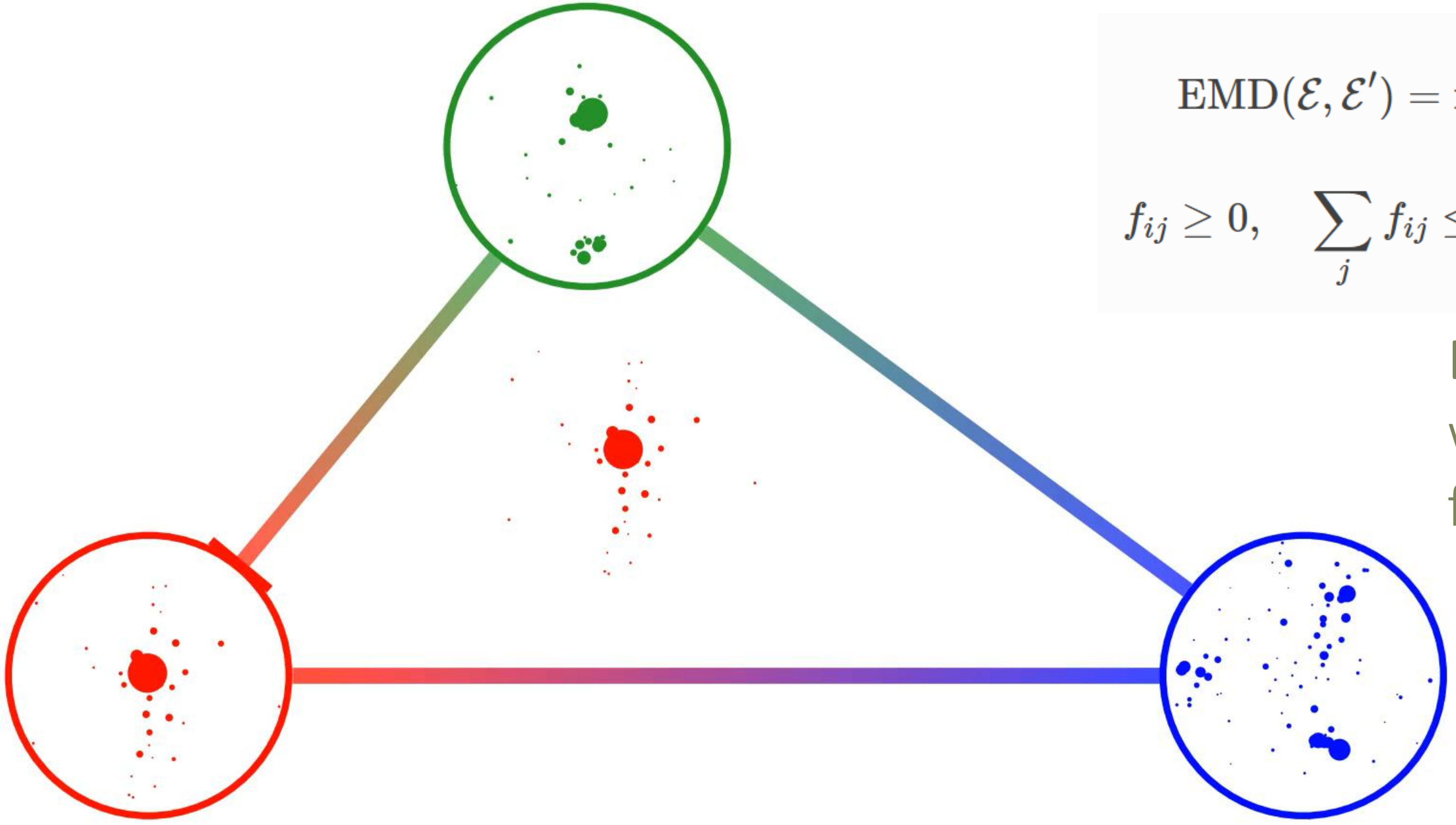
MIT

with Patrick Komiske & Eric Metodiev, 1902.02346

EPP Theory Seminar, SLAC — April 24, 2019

Earth Movers Distance

*Cost to transform one jet into another = Energy * distance*



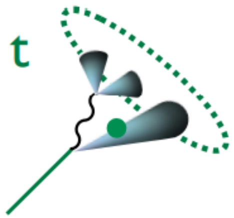
$$\mathrm{EMD}(\mathcal{E},\mathcal{E}') = \min_{\{f_{ij}\}} \sum_{ij} f_{ij} \frac{\theta_{ij}}{R} + \left| \sum_i E_i - \sum_j E'_j \right| ,$$

$$f_{ij} \geq 0, \quad \sum_j f_{ij} \leq E_i, \quad \sum_i f_{ij} \leq E'_j, \quad \sum_{ij} f_{ij} = E_{\min},$$

Defines a metric space in which jets or collider events form a geometric manifold.

Taken from https://energyflow.network/docs/emd/, Eric Metediov, Patrick Komiske III, Jesse Thaler

Visualizing Top Quark Evolution



500 GeV

Top Quark

Decay
EMD: 161.1 GeV

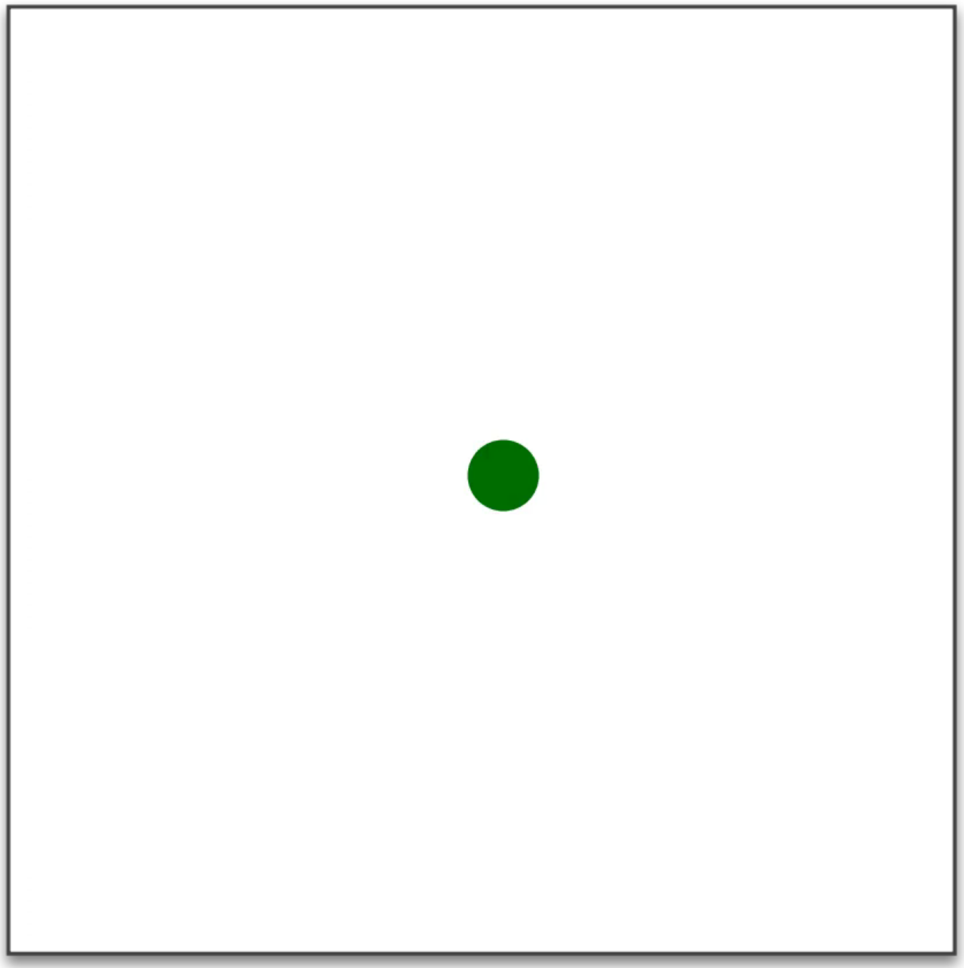
Three Quarks

Showering
EMD: 47.1 GeV

Partons

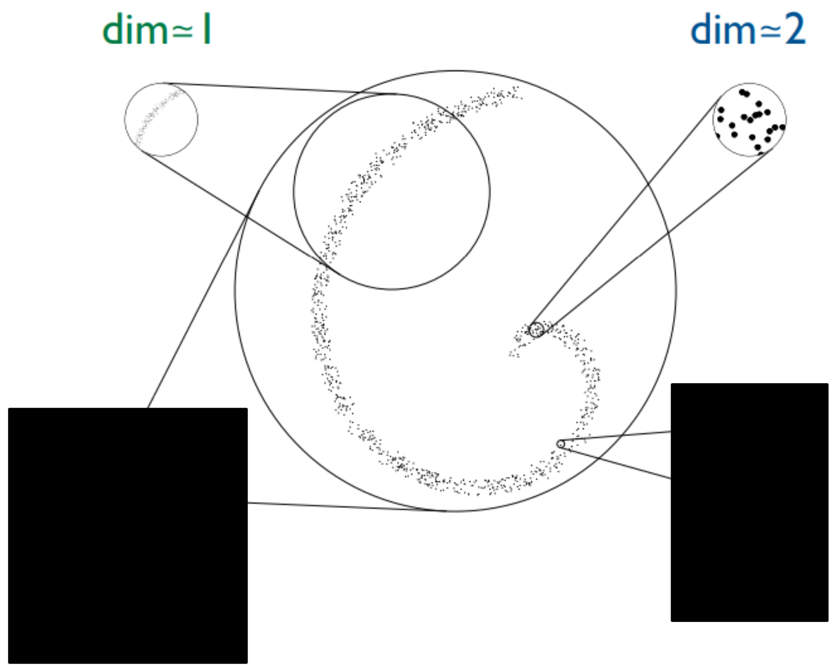
Hadronization
EMD: 27.0 GeV

Hadrons



Quantifying Dimensionality

Correlation Dimension: $\dim(Q) = Q \frac{\partial}{\partial Q} \ln \sum_i \sum_j \Theta\big(\text{EMD}(\mathcal{E}_i, \mathcal{E}_j) < Q\big)$



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

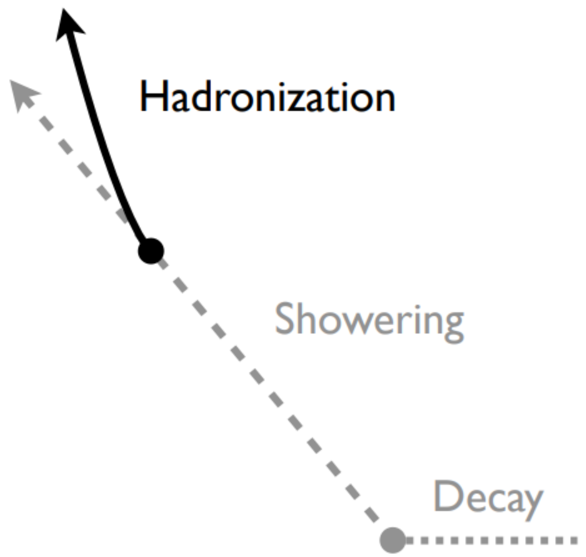
$$\Downarrow$$

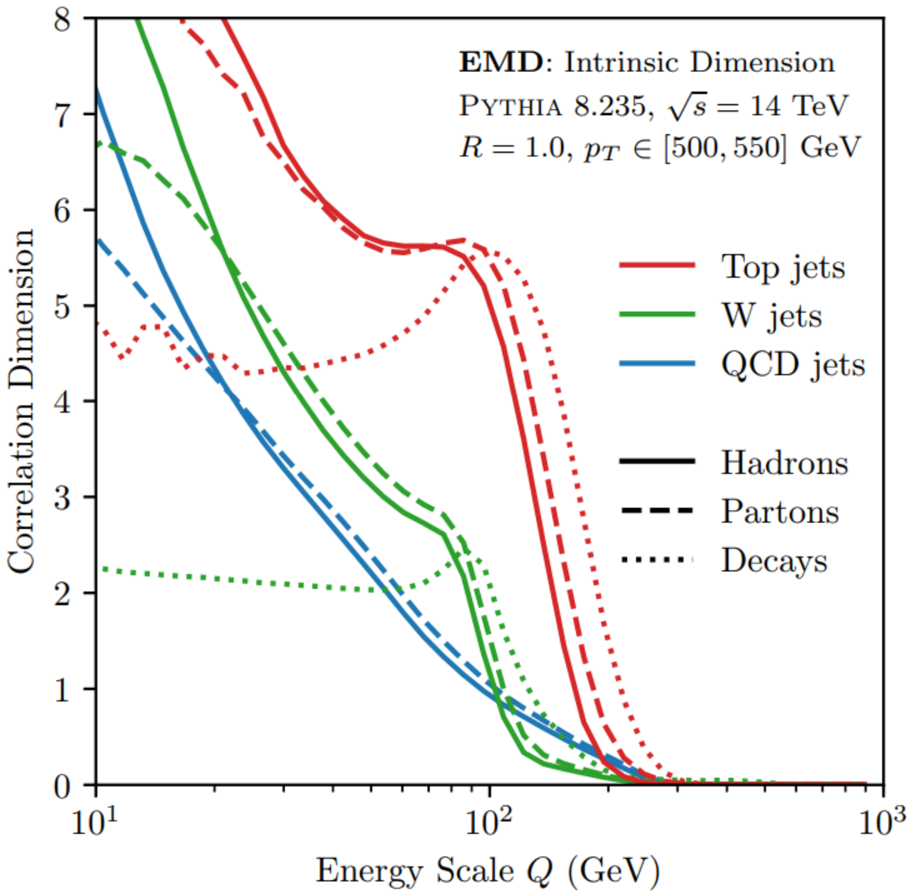
$$\dim(r) \sim r \frac{\partial}{\partial r} \ln N_{\text{neighbors}}(r)$$

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

Hadron-Level Dimension

$$\dim(Q) = Q \frac{\partial}{\partial Q} \ln \sum_i \sum_j \Theta(\text{EMD}(\mathcal{E}_i, \mathcal{E}_j) < Q)$$





Increasing complexity: multi-body phase space
perturbative emissions
non-perturbative dynamics

[Komiske, Metodiev, JDT, 1902.02346]

Preliminary Calculation

Leading Log:

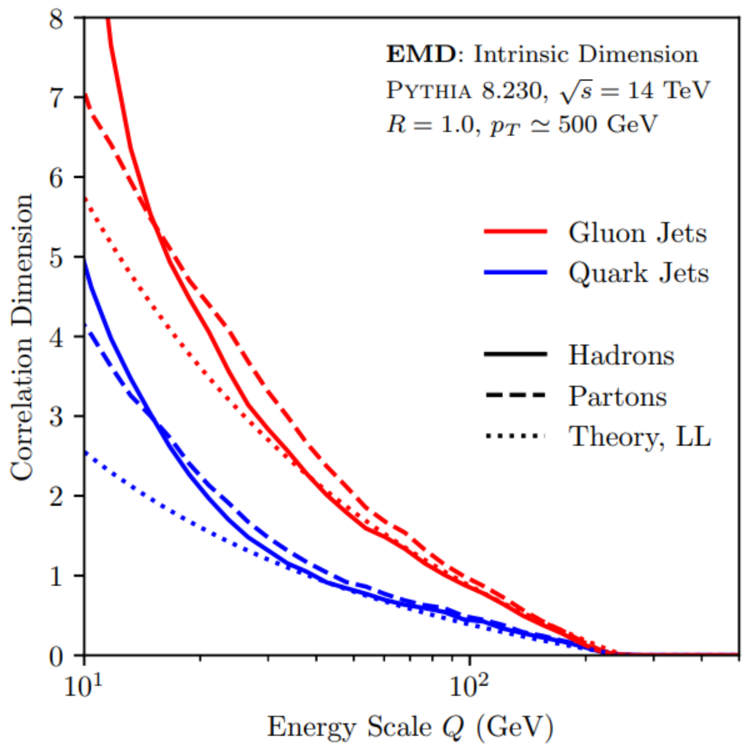
(single log, since dim has derivative)

$$\dim_i(Q) \simeq -\frac{8\alpha_s}{\pi} C_i \ln \frac{Q}{p_T}$$

Color Factor

$C_A = 3$

$C_F = 4/3$



Point of EMD

EMD is a distance metric that encodes physical information.

Different physical processes that are relevant for jet formation are associated with different EMD scales.

E.g. top pT ~ 500 GeV, top mass ~ 150 GeV, QCD ~ 1 – 150 GeV.

The geometry of the jet manifold under EMD reflects physical processes. **Unlike jet image differences.**

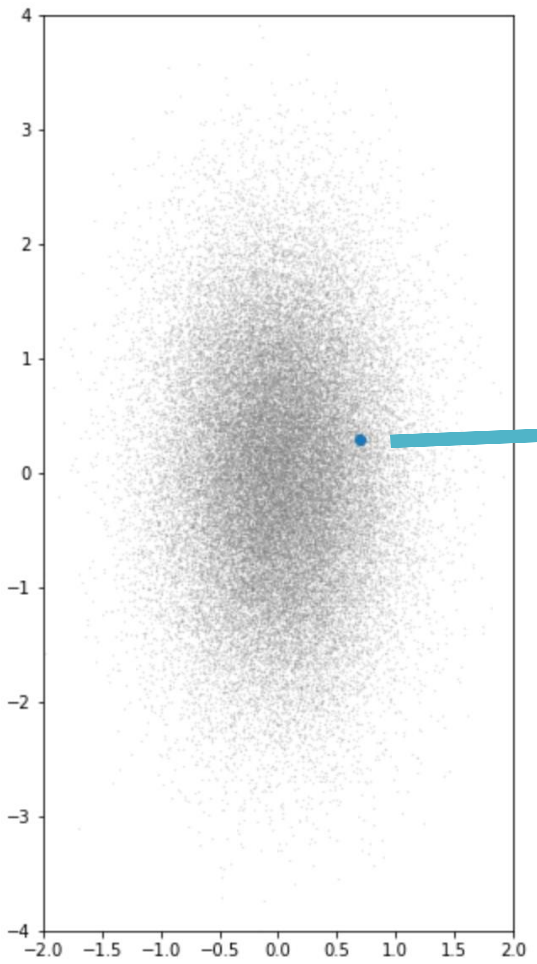
→ Can learn physics from physics < -- > geometry map?

Main Course

The Variational Autoencoder

The Plain Autoencoder

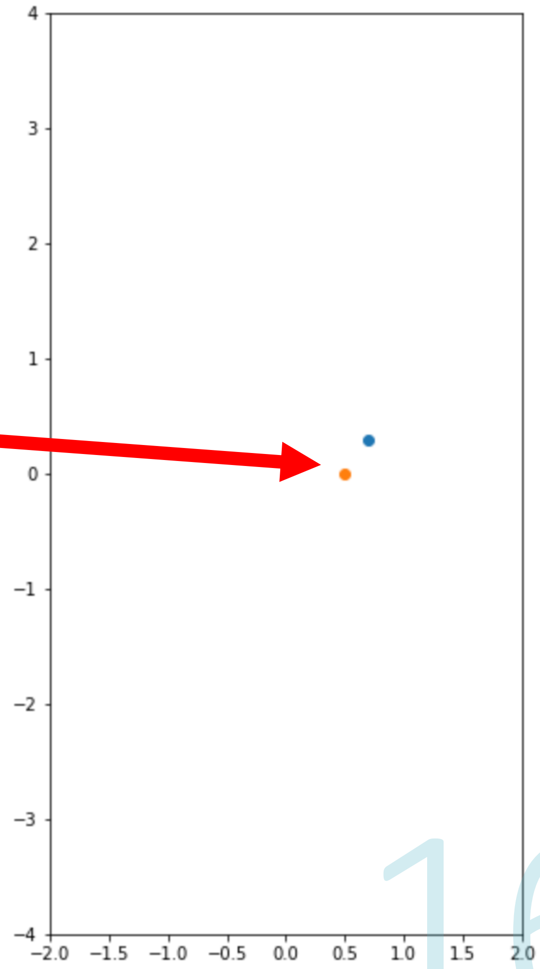
Garbage



AE

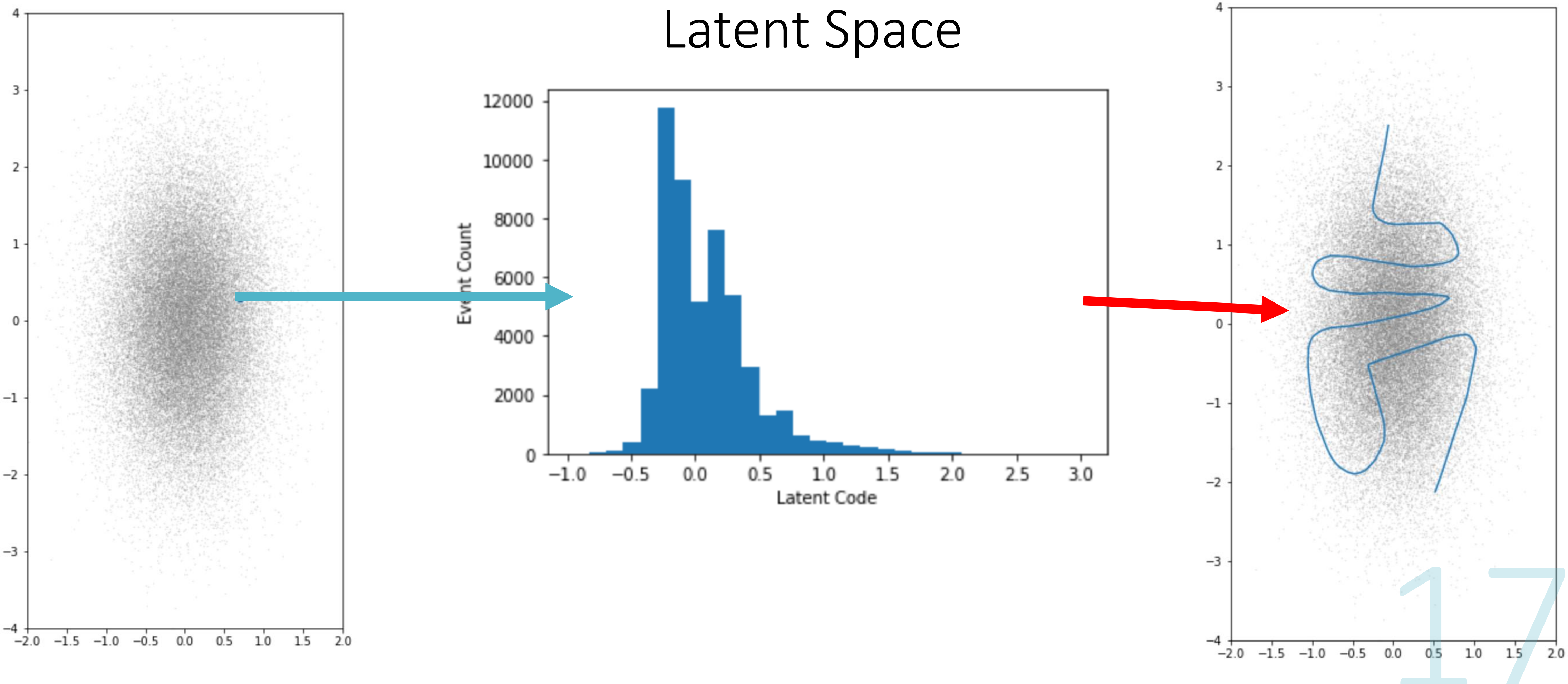
(1D latent space)

Rec. Loss = $|\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2$



The Plain Autoencoder

Garbage

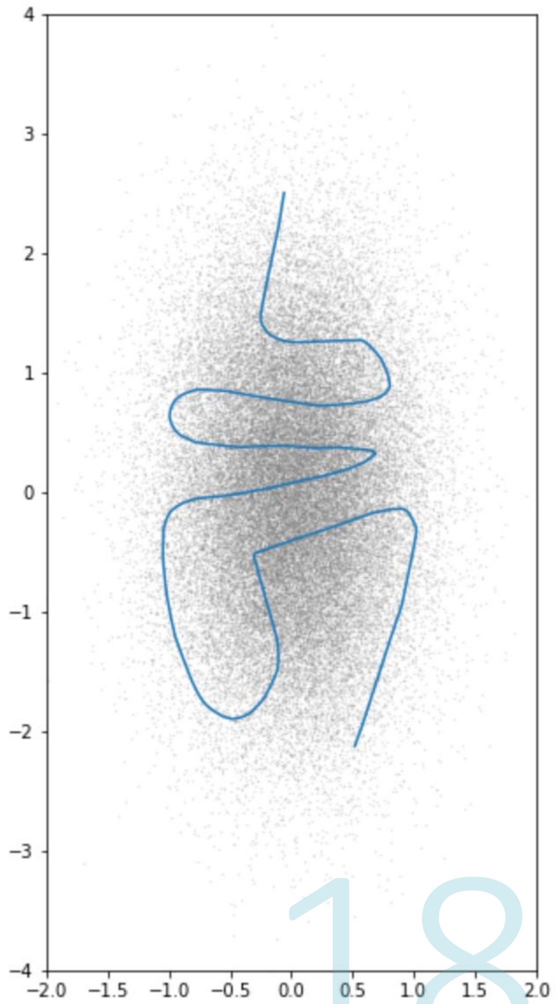


The Plain Autoencoder

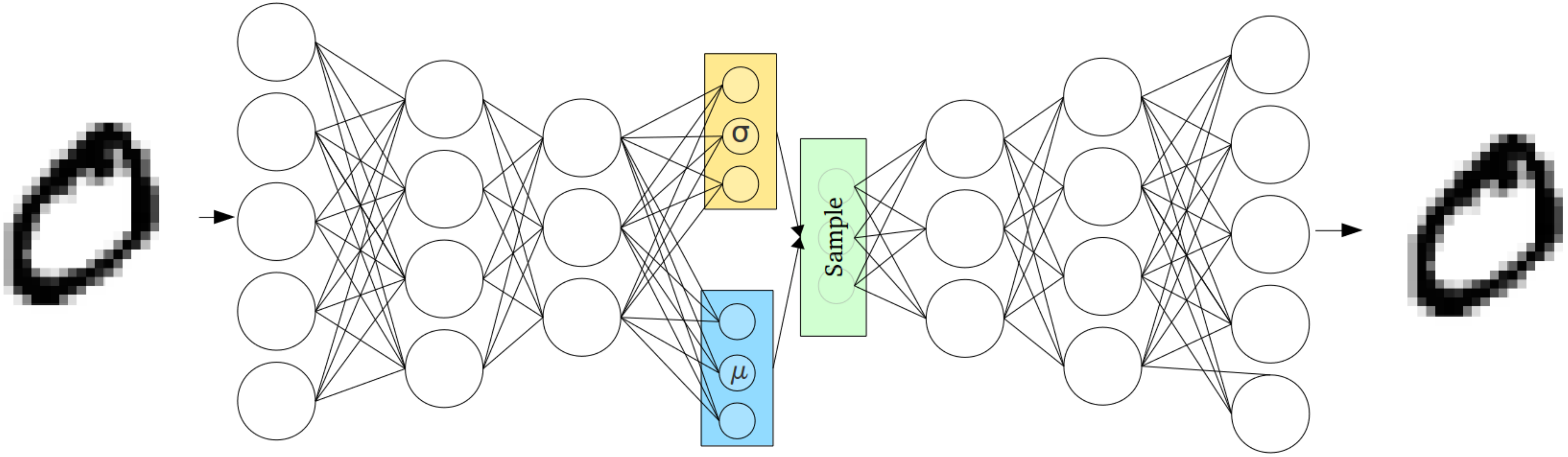
Garbage

1. The AE learns some dense packing of the data space

2. The latent representation is highly coupled with the expressiveness of the network architecture of the encoder and decoder



The Variational Autoencoder



$$\text{Loss} = \underbrace{|\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2/\beta^2}_{\text{Reconstruction error}} - \underbrace{\sum_i \tfrac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)}_{KL(q(z|x)||p(z)) \sim \text{"Information cost"}}$$

Reconstruction error

KL(q(z|x)||p(z)) ~ "Information cost"

The Variational Autoencoder

Information and the loss function

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2/\beta^2 - \sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2 - \beta^2 \sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

The Variational Autoencoder

Information and the loss function

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2/\beta^2 - \sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

1) $\boldsymbol{\beta}$ is dimensionful!

The same dimension as the distance metric,
e.g. GeV.

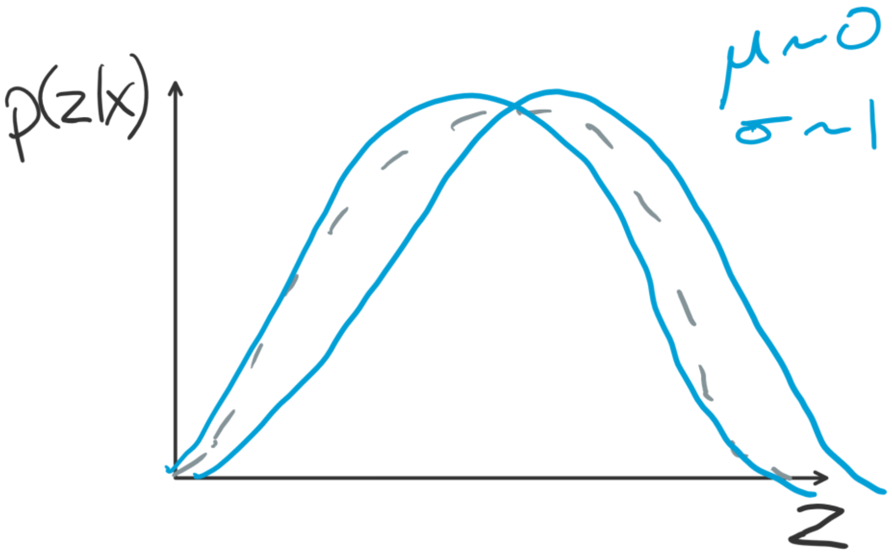
$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2 - \beta^2 \sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

The Variational Autoencoder

Information and the loss function

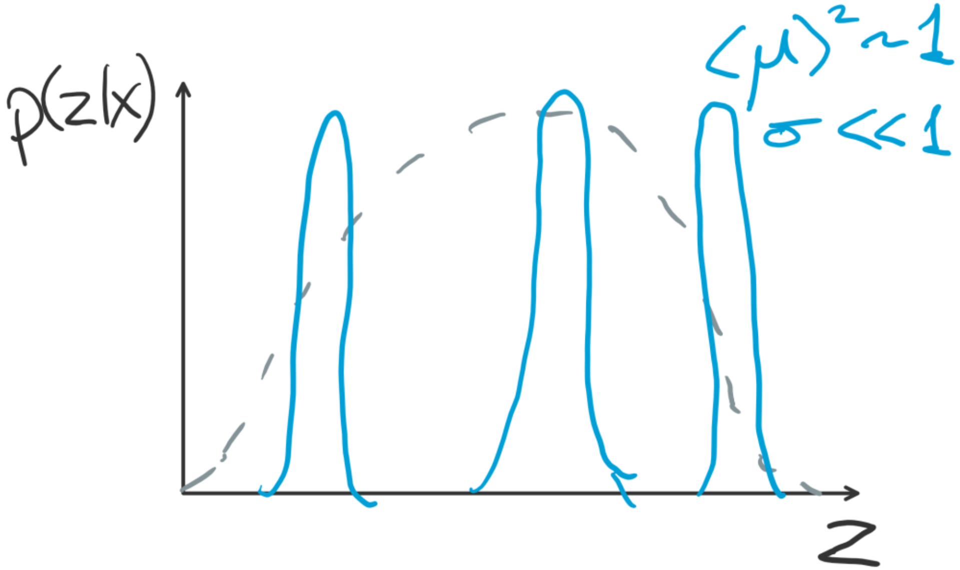
$\boldsymbol{\beta} \rightarrow \infty$

No info encoded in latent space



$\boldsymbol{\beta} \ll$ ***Lengthscale***

Info encoded in latent space



$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2 - \beta^2 \sum_i \frac{1}{2}\left(1 + \log \sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

The Variational Autoencoder

Information and the loss function

$\boldsymbol{\beta} \rightarrow \infty$

No info encoded in latent space

$\boldsymbol{\beta} \ll \textit{Lengthscale}$

Info encoded in latent space

2) $\boldsymbol{\beta}$ is the cost for encoding information

The encoder will only encode information about the input to the extent that its usefulness for reconstruction is sufficient to justify the cost.

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2 - \beta^2 \sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

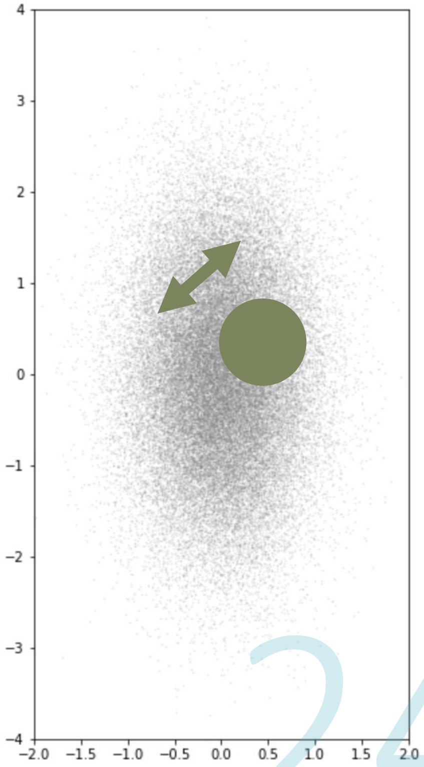
The Variational Autoencoder

Information and the loss function

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2/\beta^2 - \sum_i \frac{1}{2}\left(1 + \log \sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

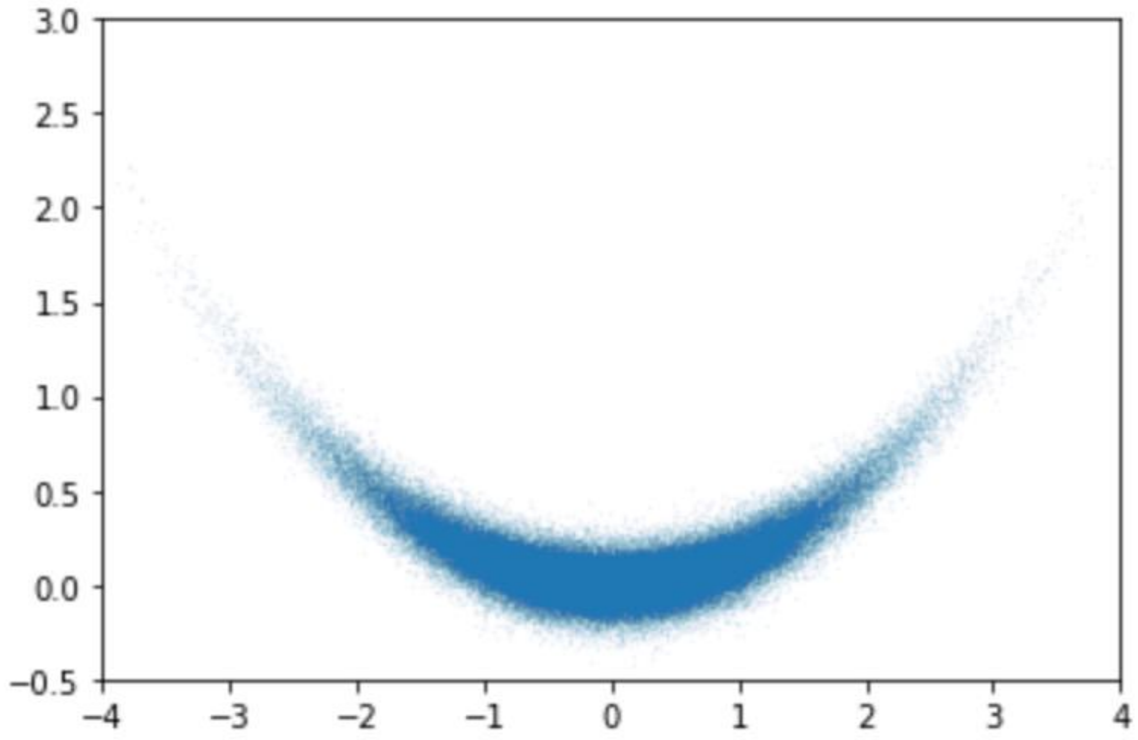
3) $\boldsymbol{\beta}$ is the distance resolution in reconstruction space

The stochasticity of the latent sampling will smear the reconstruction at scale $\sim \beta$



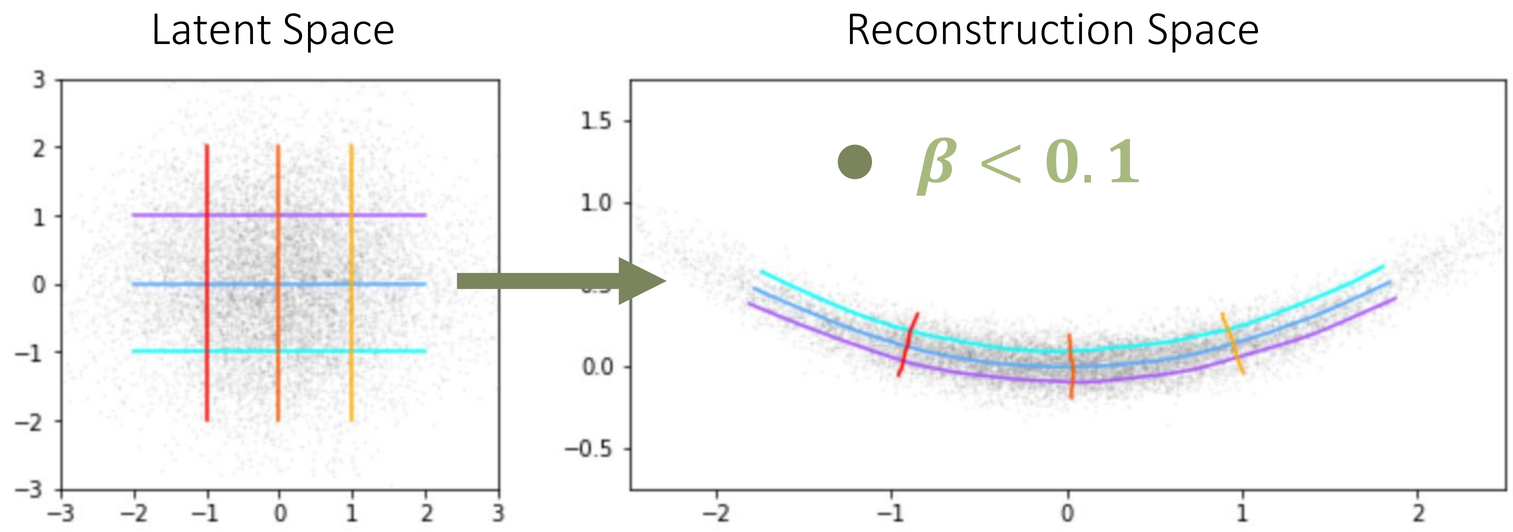
The Variational Autoencoder

Bananas



The Variational Autoencoder

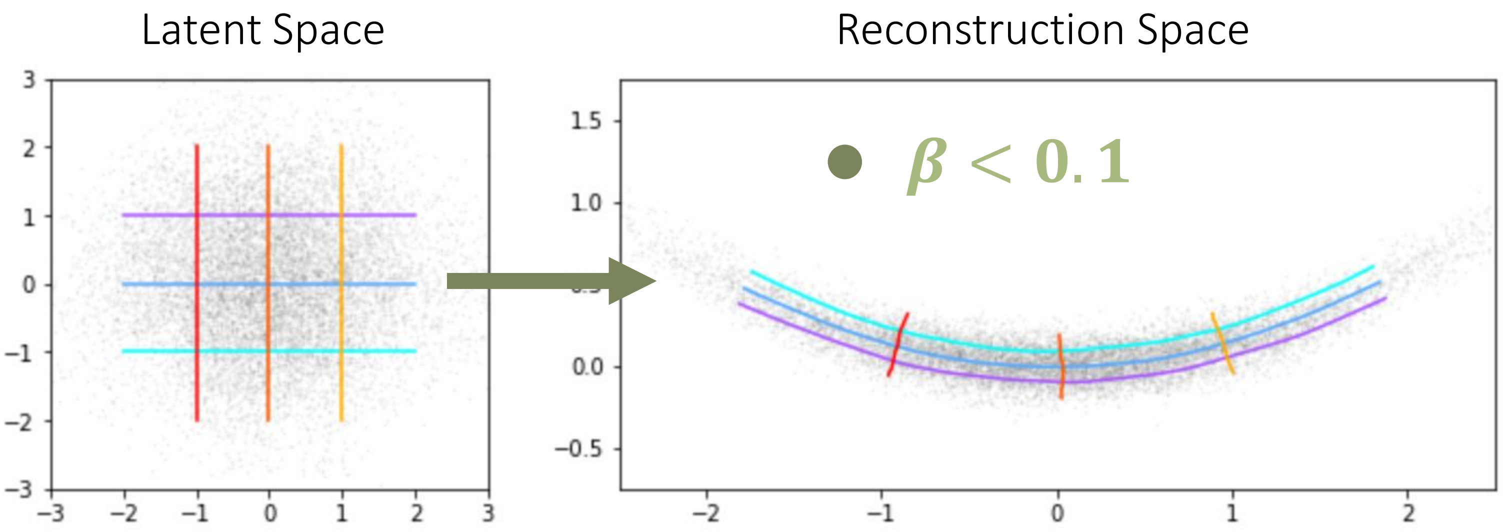
Bananas



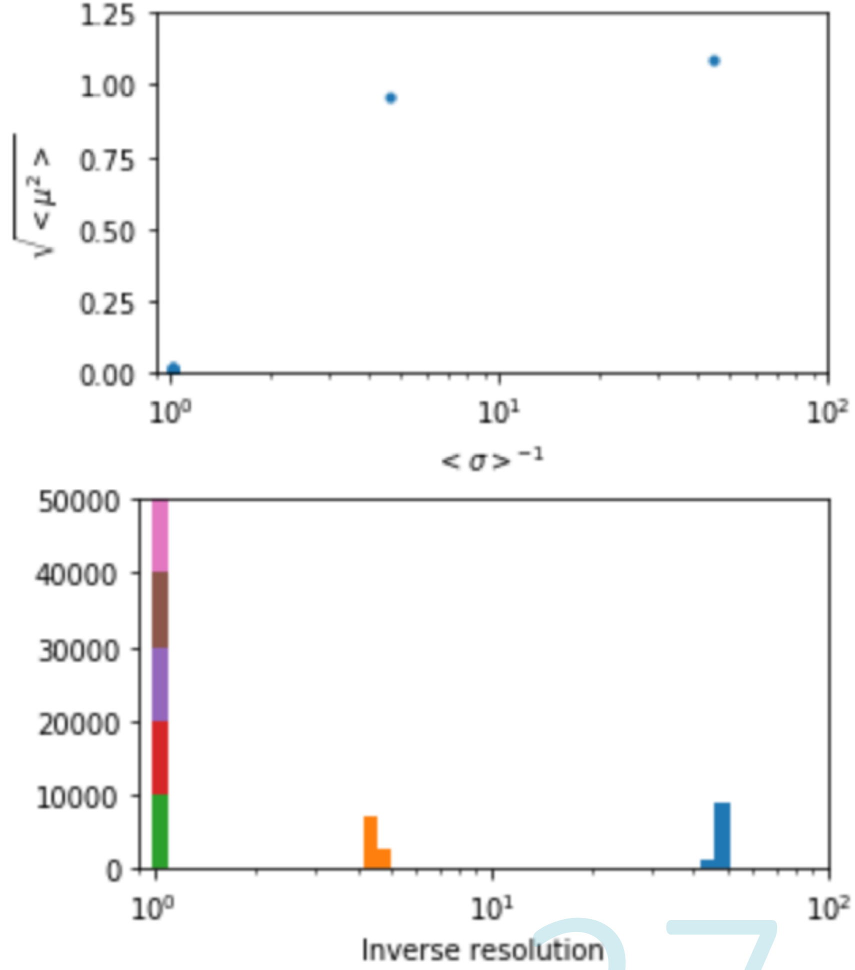
- $\beta < 0.1$

The Variational Autoencoder

Bananas



- $\beta < 0.1$

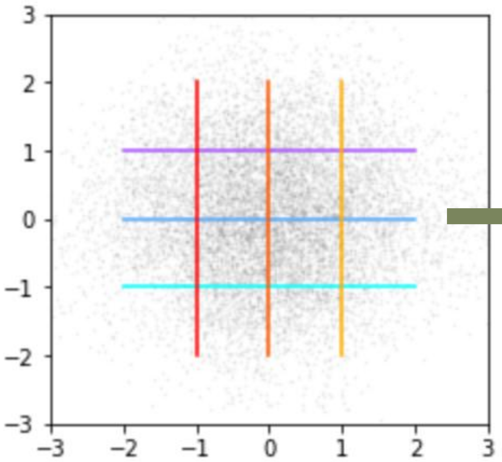


The Variational Autoencoder

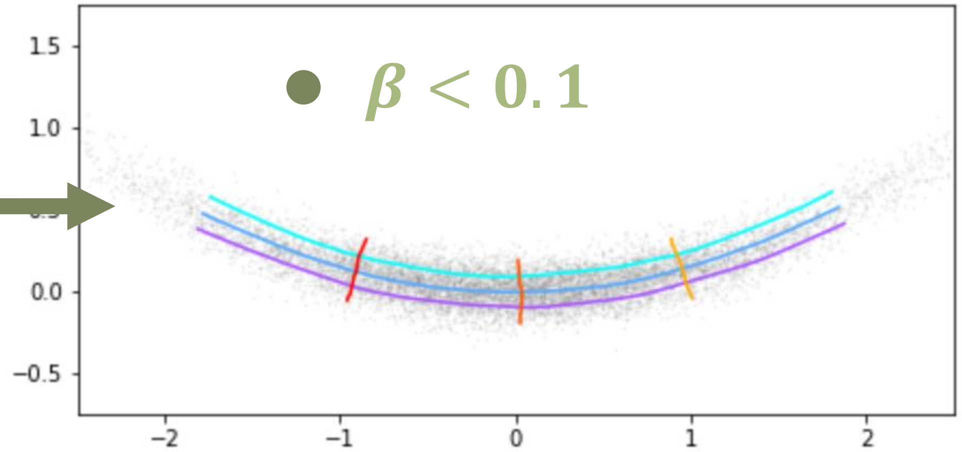
Bananas

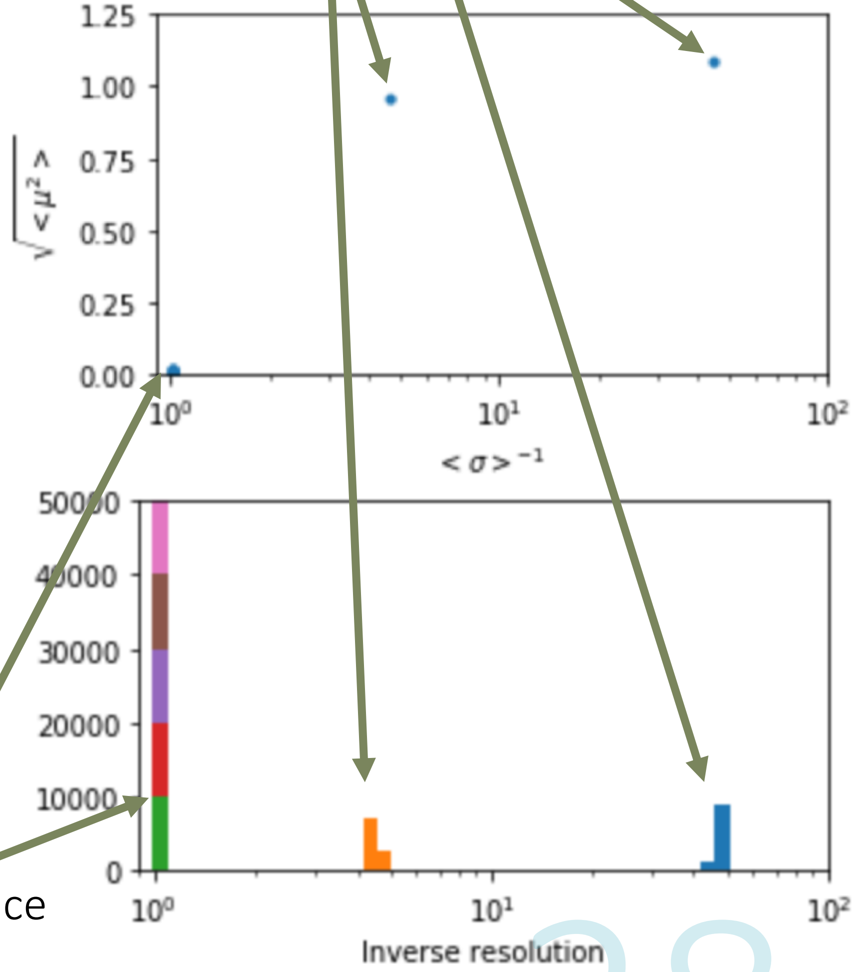
2 latent space dimensions learn something

Latent Space



Reconstruction Space





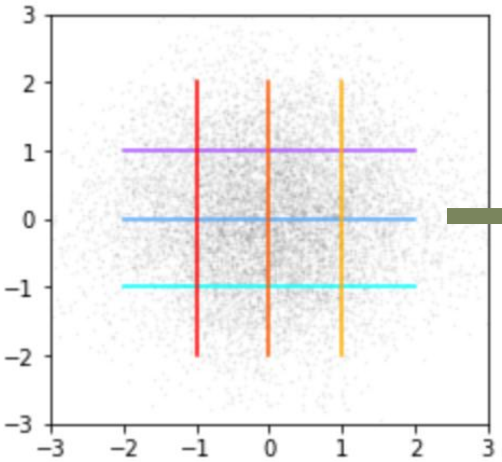
The VAE is doing non-linear PCA

8 Excess latent space dimensions learn nothing

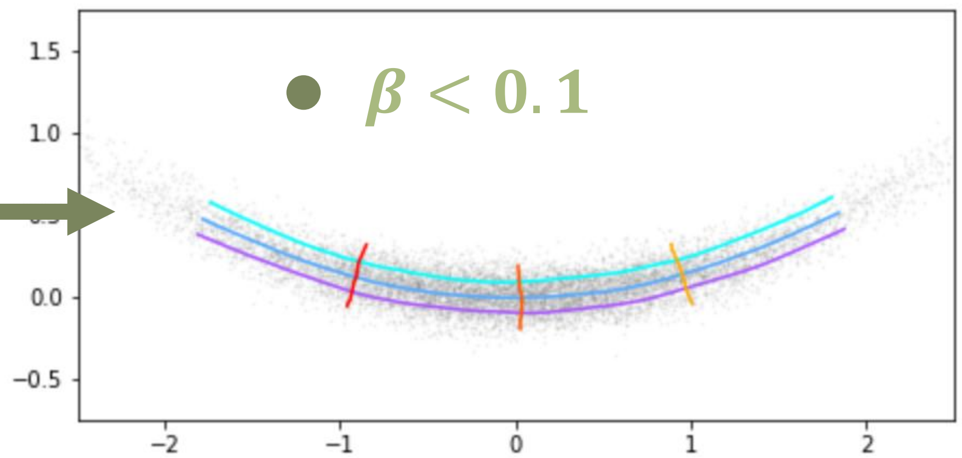
The Variational Autoencoder

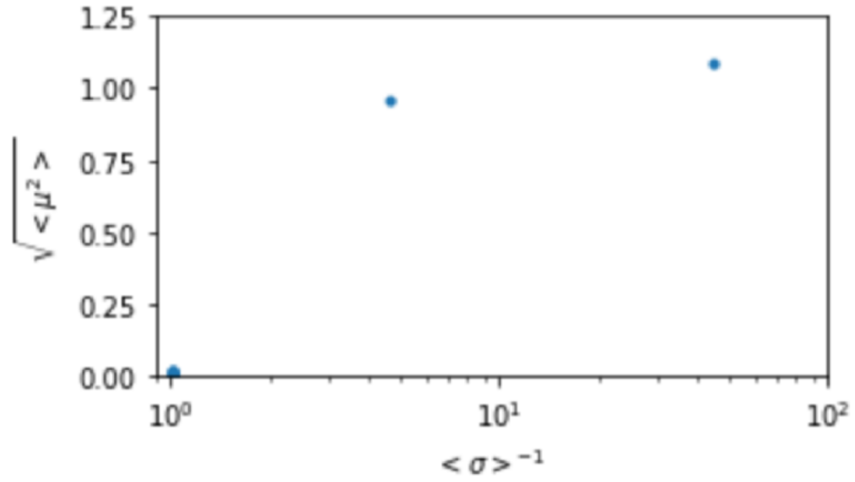
Bananas

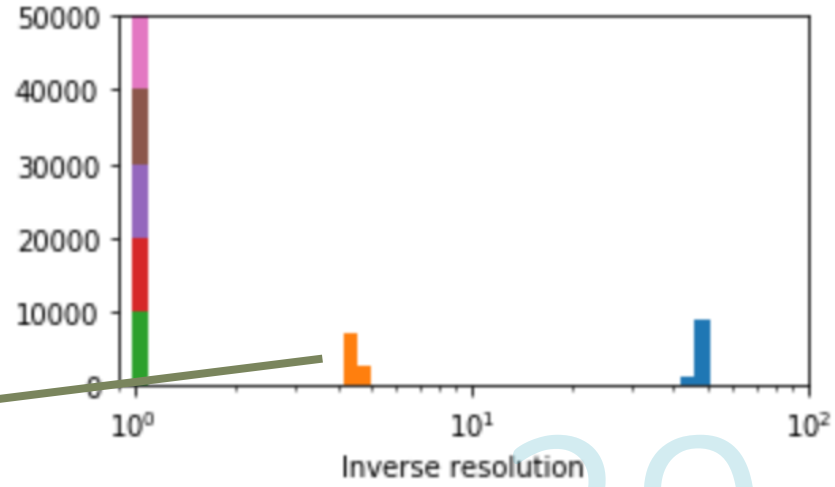
Latent Space



Reconstruction Space







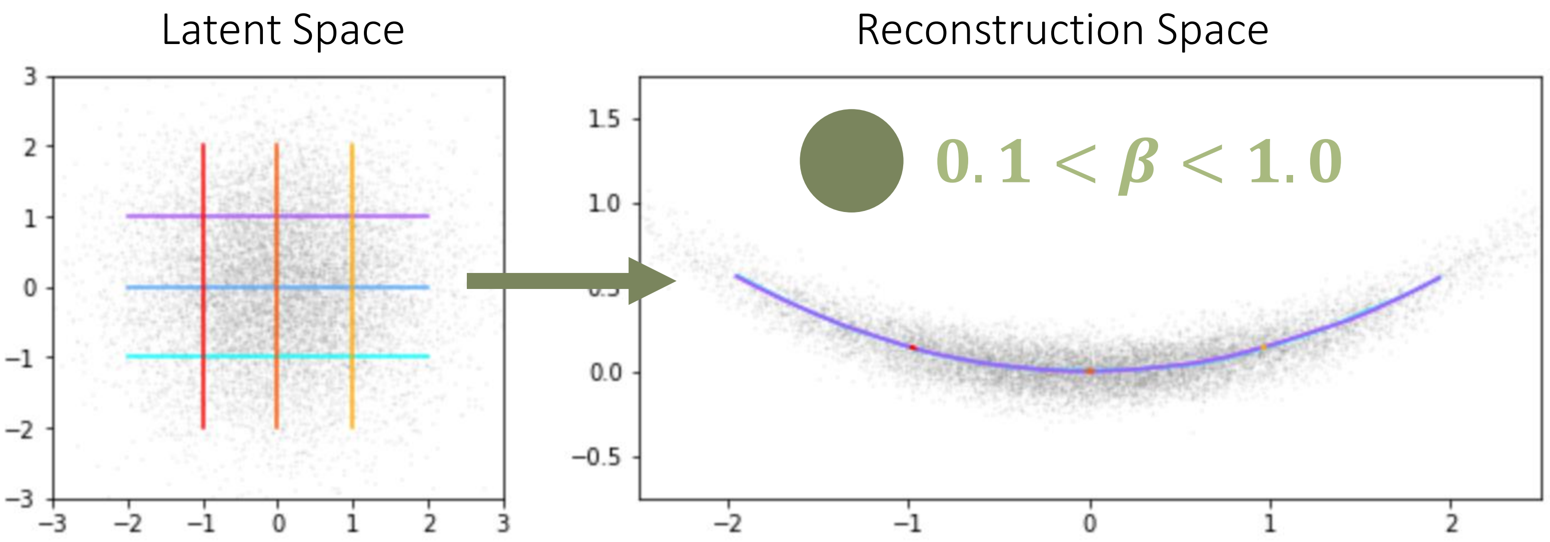
The VAE is doing non-linear PCA

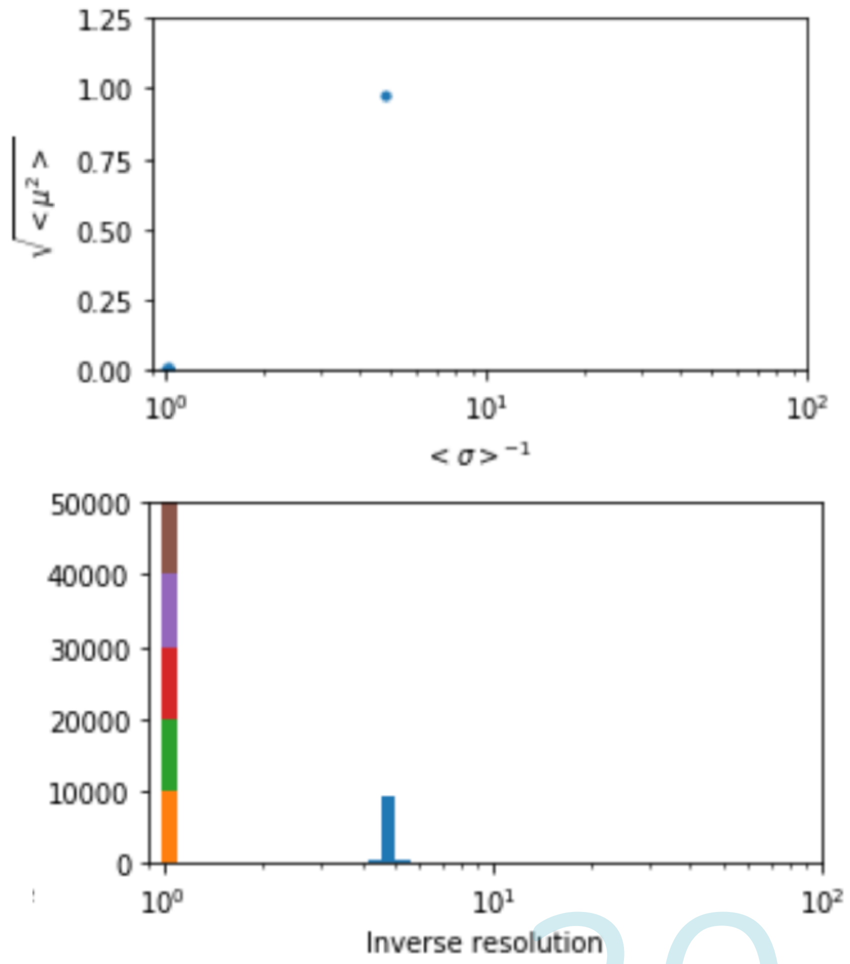
Size = β/σ

Spectrscopy

The Variational Autoencoder

Bananas

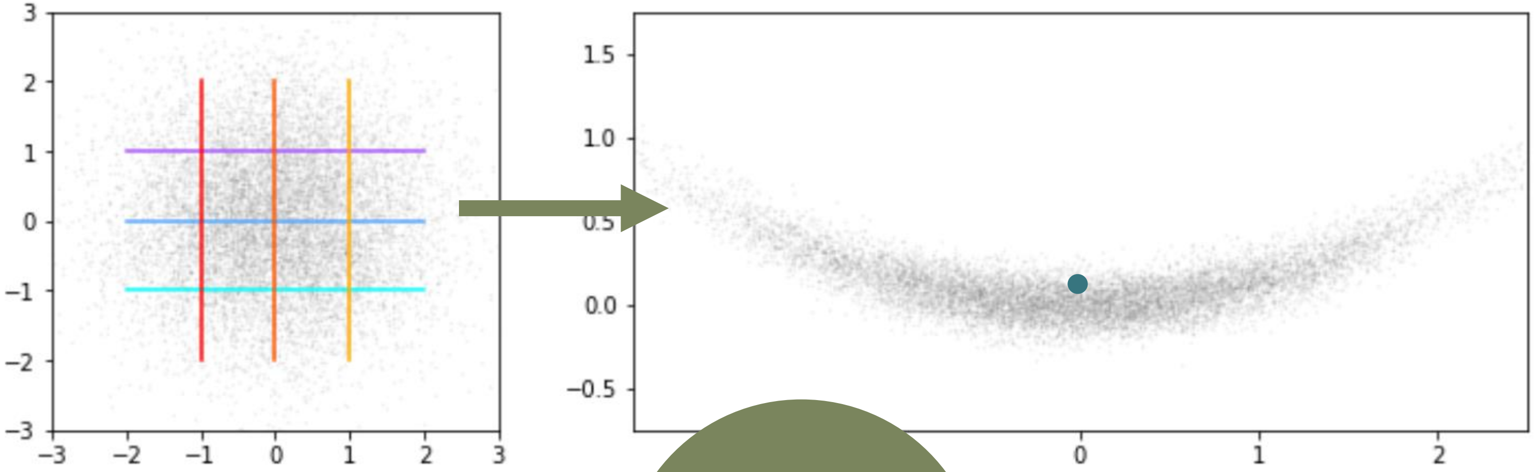


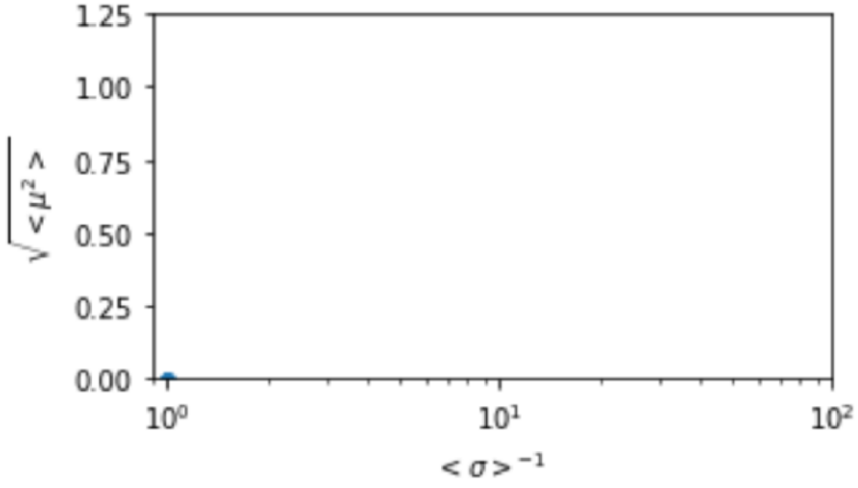


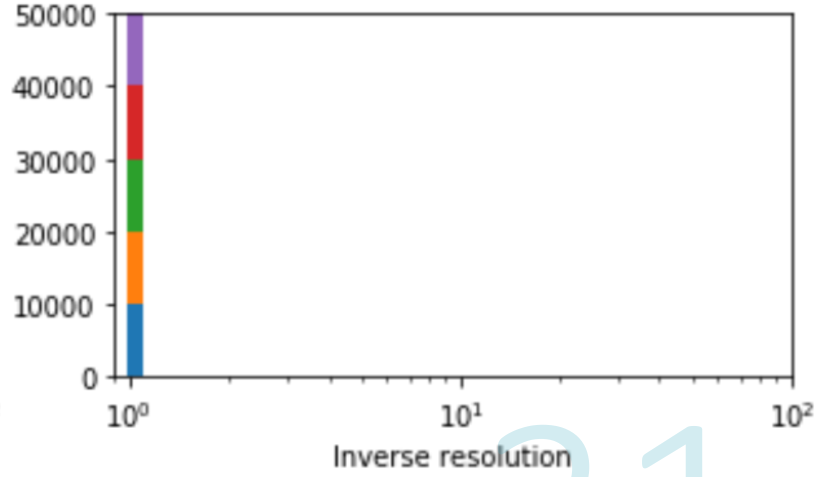
The Variational Autoencoder

Bananas

Latent Space　　Reconstruction Space







$\beta \gg 1$

The Variational Autoencoder

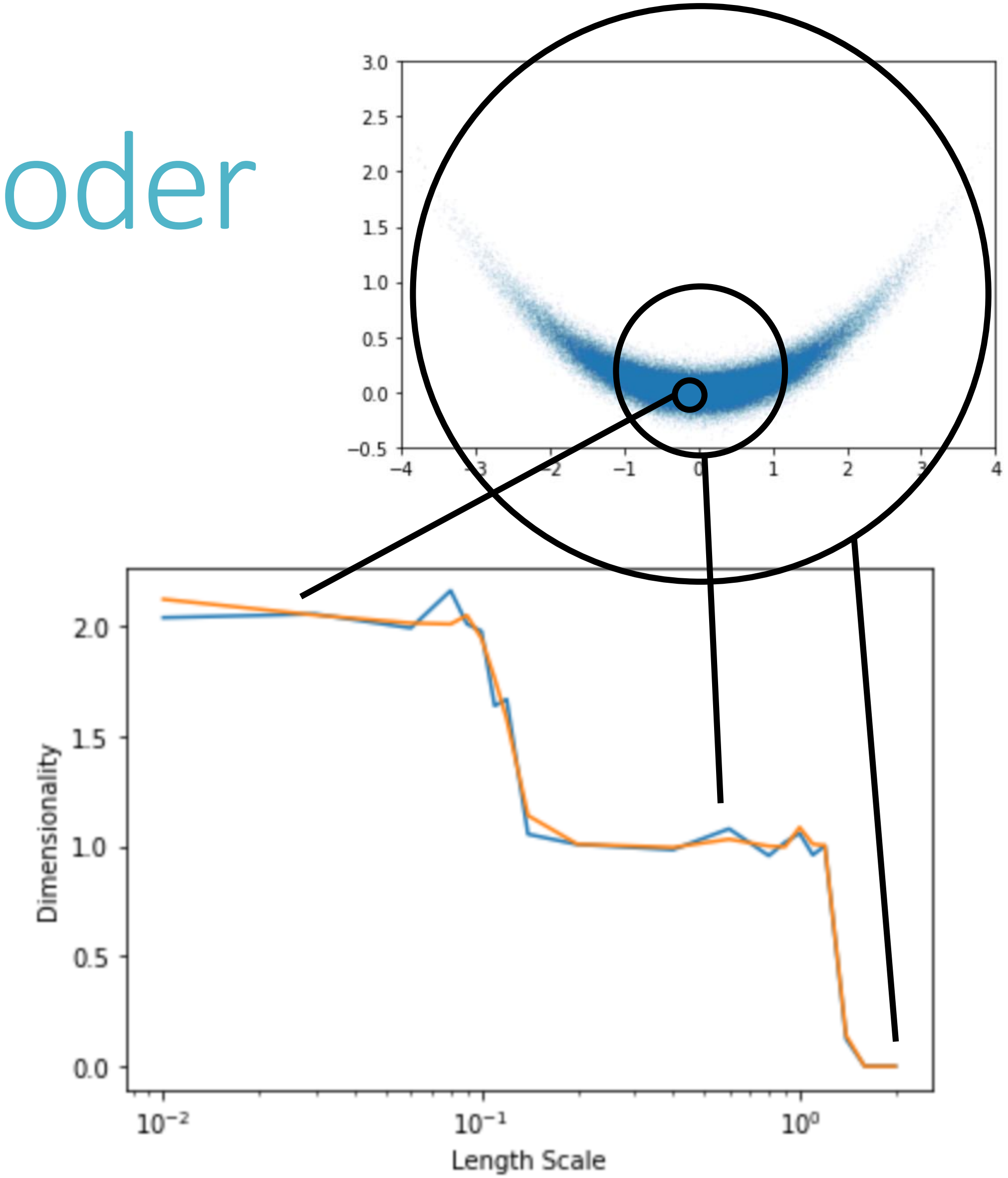
Dimensionality

$$D_1 \equiv 2 \frac{d\langle|\Delta \boldsymbol{x}|^2\rangle}{d\,\beta^2}$$

Variation of resolution with scale (think $\langle r^2 \rangle = D\,\sigma^2$ for D-dimensional Gaussian).

$$D_2 \equiv \frac{d\,KL}{d \log \beta}$$

Variation of information with scale.



I am still trying to work out formally the meaning of these expressions, but they have an air of truthiness about them and empirically give sensible results.

The Variational Autoencoder

What is the point

Dimensionality Analysis

$$D_1 \equiv 2\frac{d\langle|\Delta \boldsymbol{x}|^2\rangle}{d\,\beta^2}$$

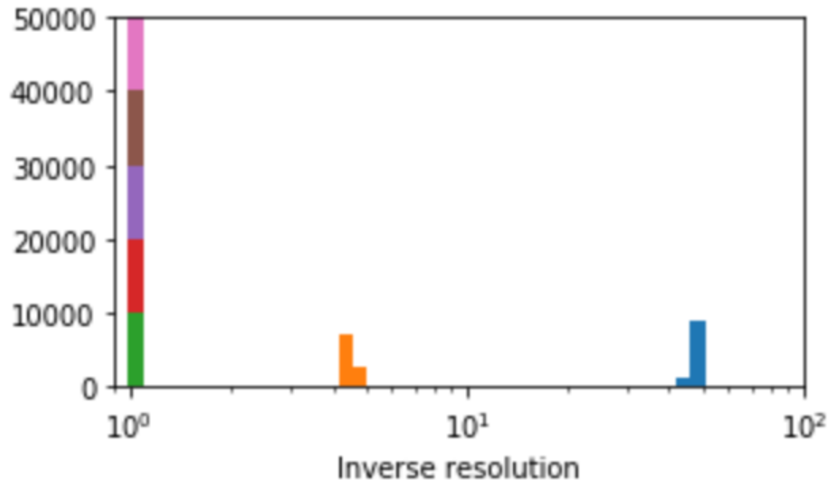
$$D_2 \equiv \frac{d\,KL}{d\log\beta}$$

The VAE learns a scale-dependent representation of the true data geometry.

Properties of the data geometry can be inferred directly from the statistics of the latent encoding.

Properties of the ***physics*** *can be learnt from the latent space?*

Spectral Analysis



The Variational Autoencoder

What is new?

Dimensionality Analysis

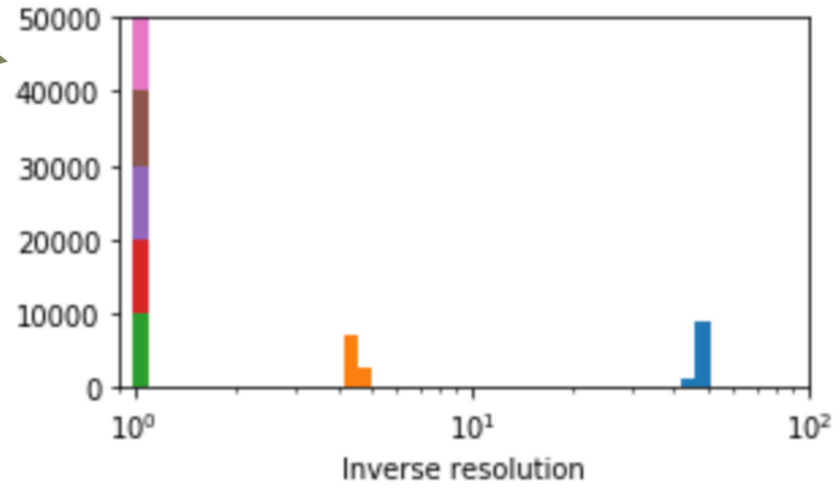
$$D_1 \equiv 2\frac{d\langle|\Delta\boldsymbol{x}|^2\rangle}{d\,\beta^2}$$

$$D_2 \equiv \frac{d\;KL}{d\log\beta}$$

Are these new?

I have never seen them before.

Spectral Analysis

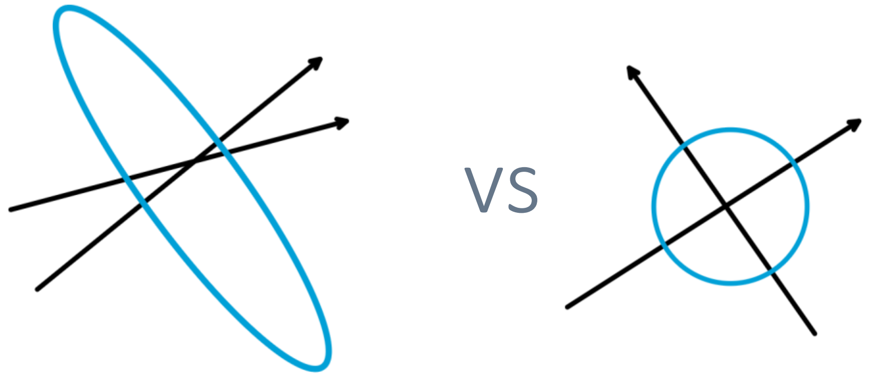


This maps directly to lengthscales in the data manifold space

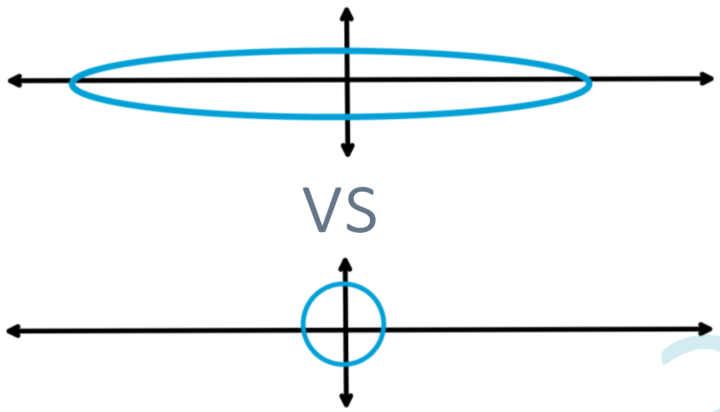
The Variational Autoencoder

Orthogonalization and Organization is Information-Efficient

Orthogonalization:



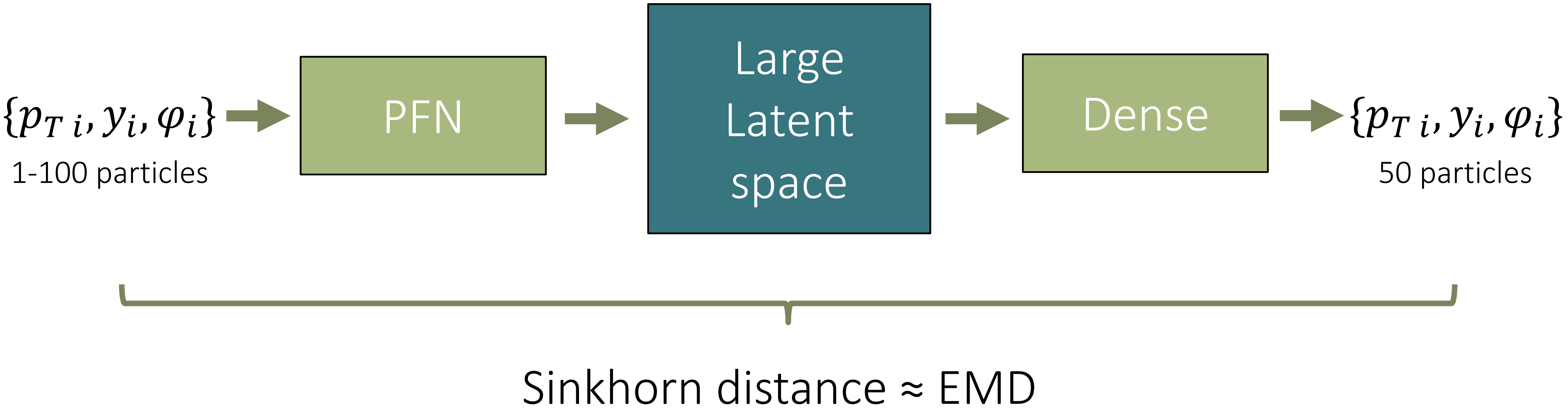
Organization:



Cheese Course

Application to Top Jets

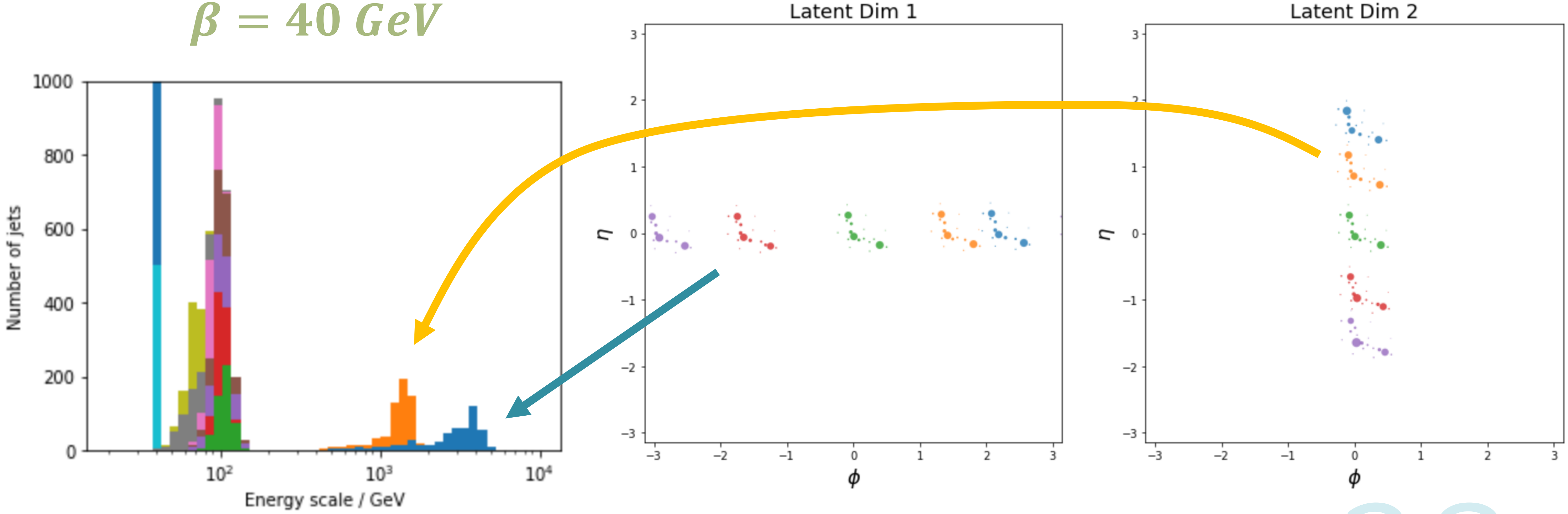
Jet VAE


{pT i, yi, φi}
1-100 particles
PFN
Large Latent space
Dense
{pT i, yi, φi}
50 particles
Sinkhorn distance ≈ EMD

Exploring the Learnt Representation

Top Jets

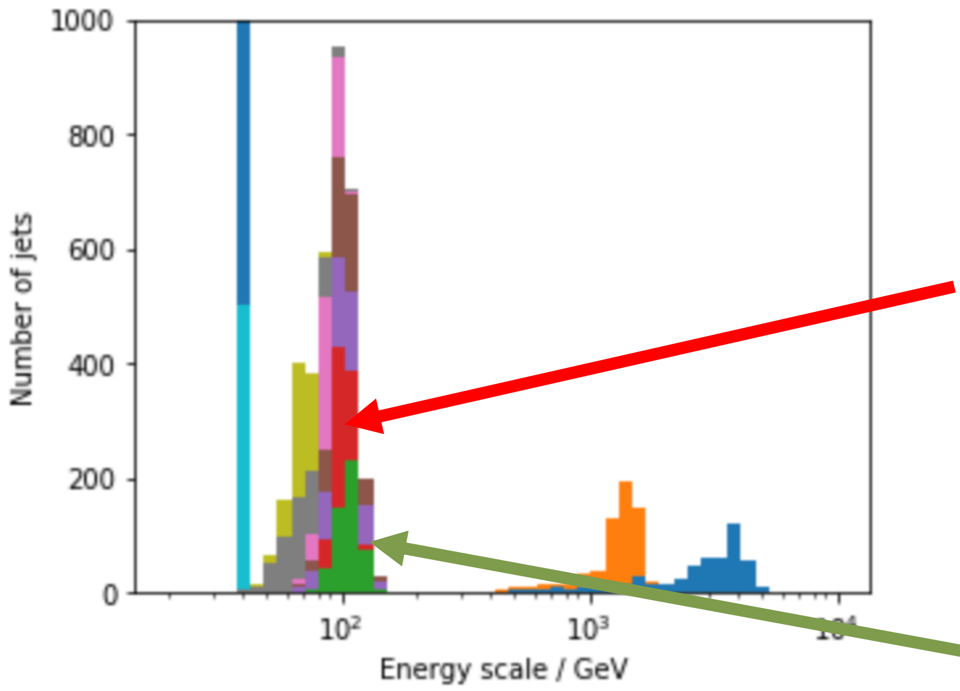
$\beta = 40\ GeV$

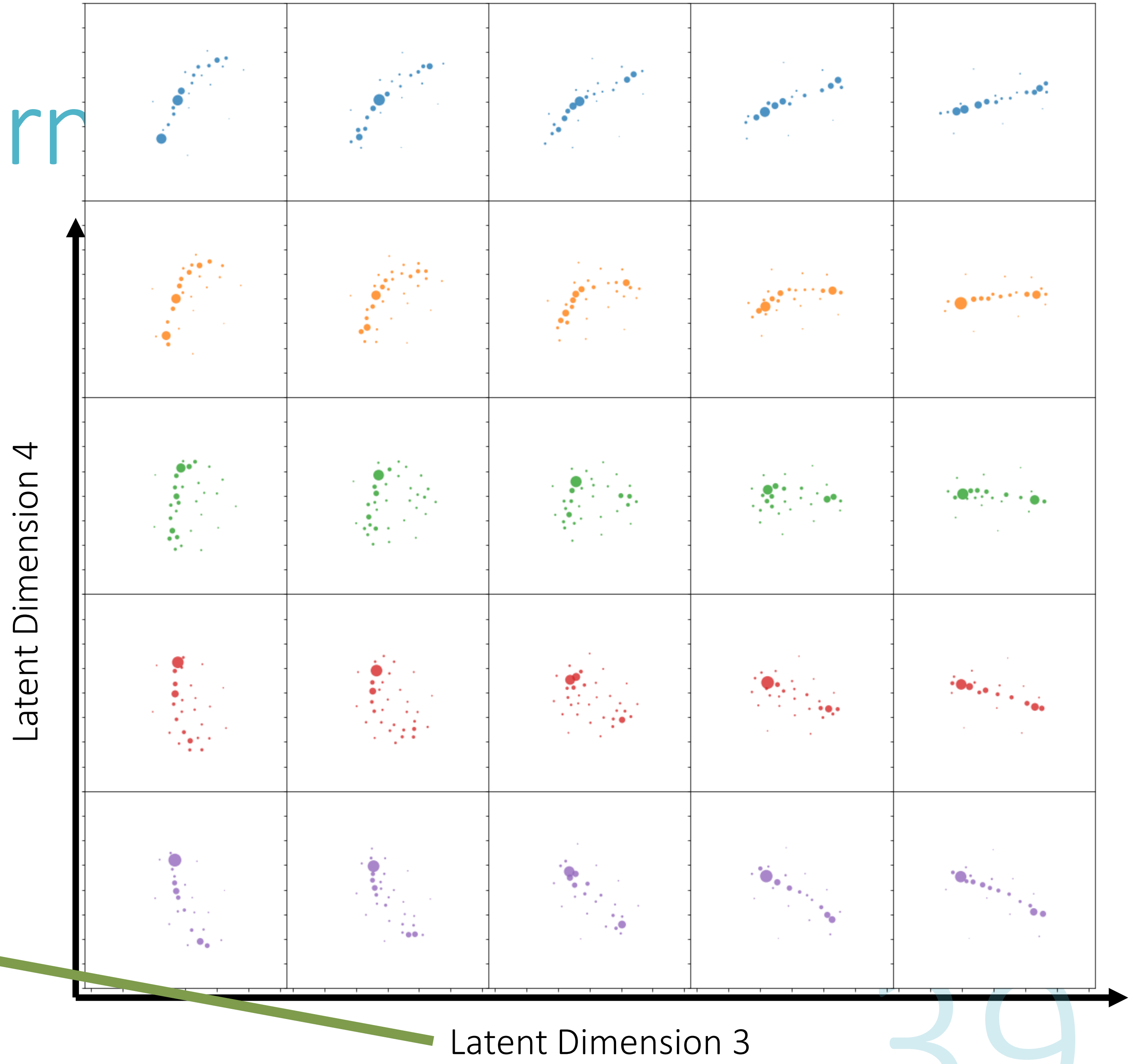


Exploring the Learn

Top Jets

$$\beta = 40\ GeV$$

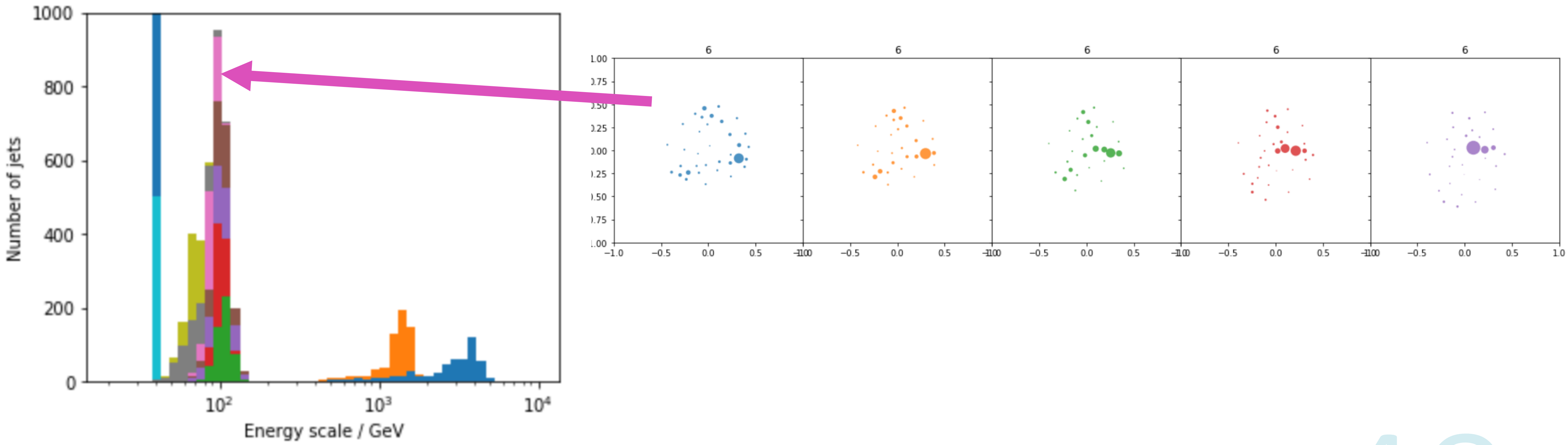




Exploring the Learnt Representation

Top Jets

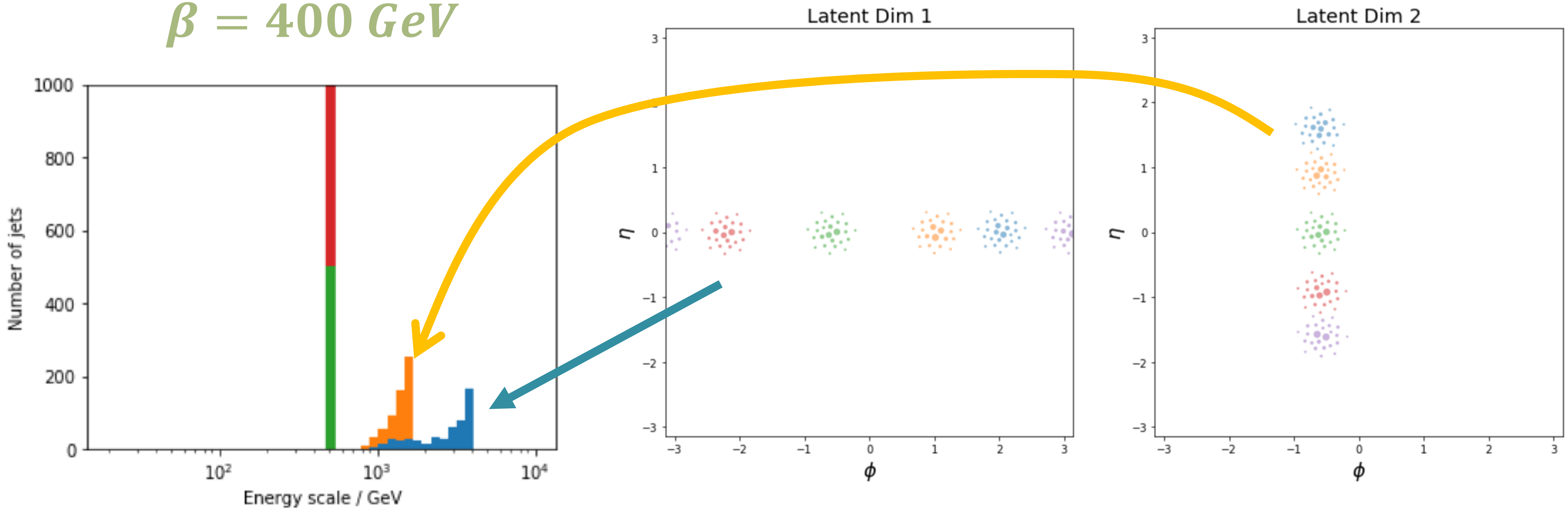
$\boldsymbol{\beta = 40\ GeV}$



Exploring the Learnt Representation

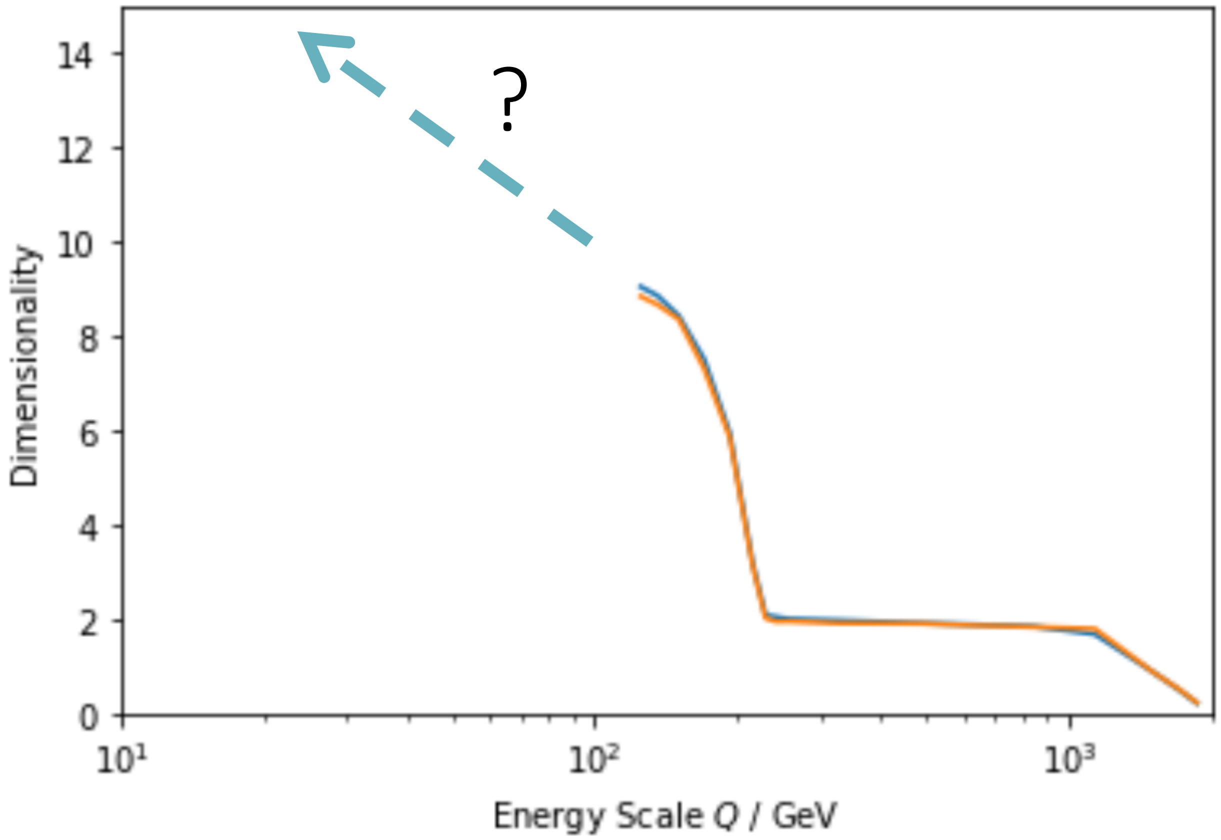
Top Jets

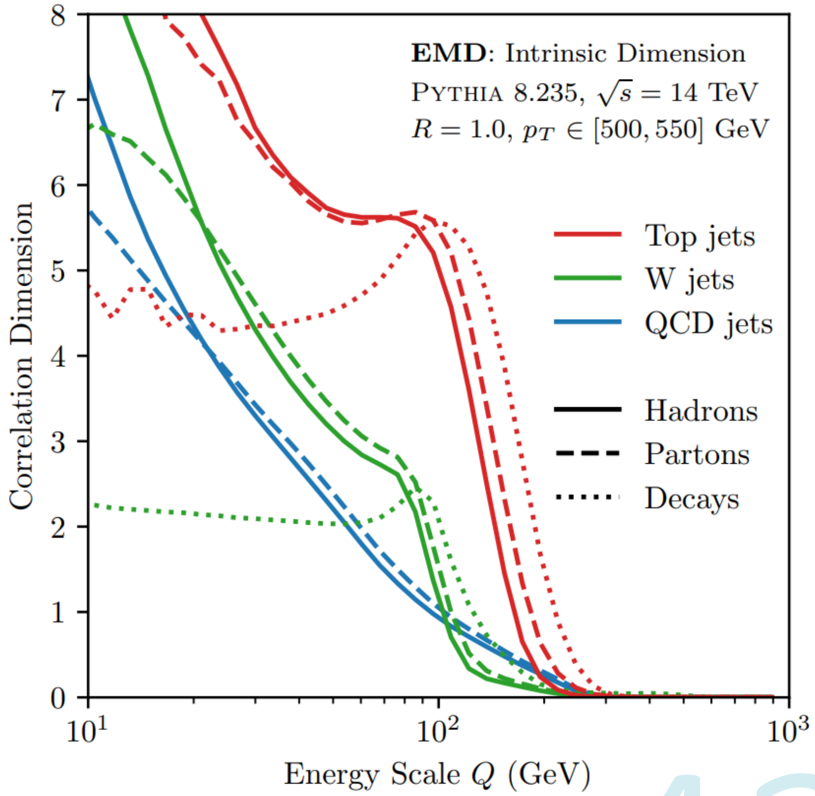
$\beta = 400\ GeV$



Exploring the Learnt Representation

Dimensionality





What is the point?

"Can we learn something new from dimensionality and geometry? Maybe something in the nonperturbative regime?"

Anonymous Professor A

"Once you have understood the geometry of the data manifold you have understood everything about the problem"

Anonymous Professor B

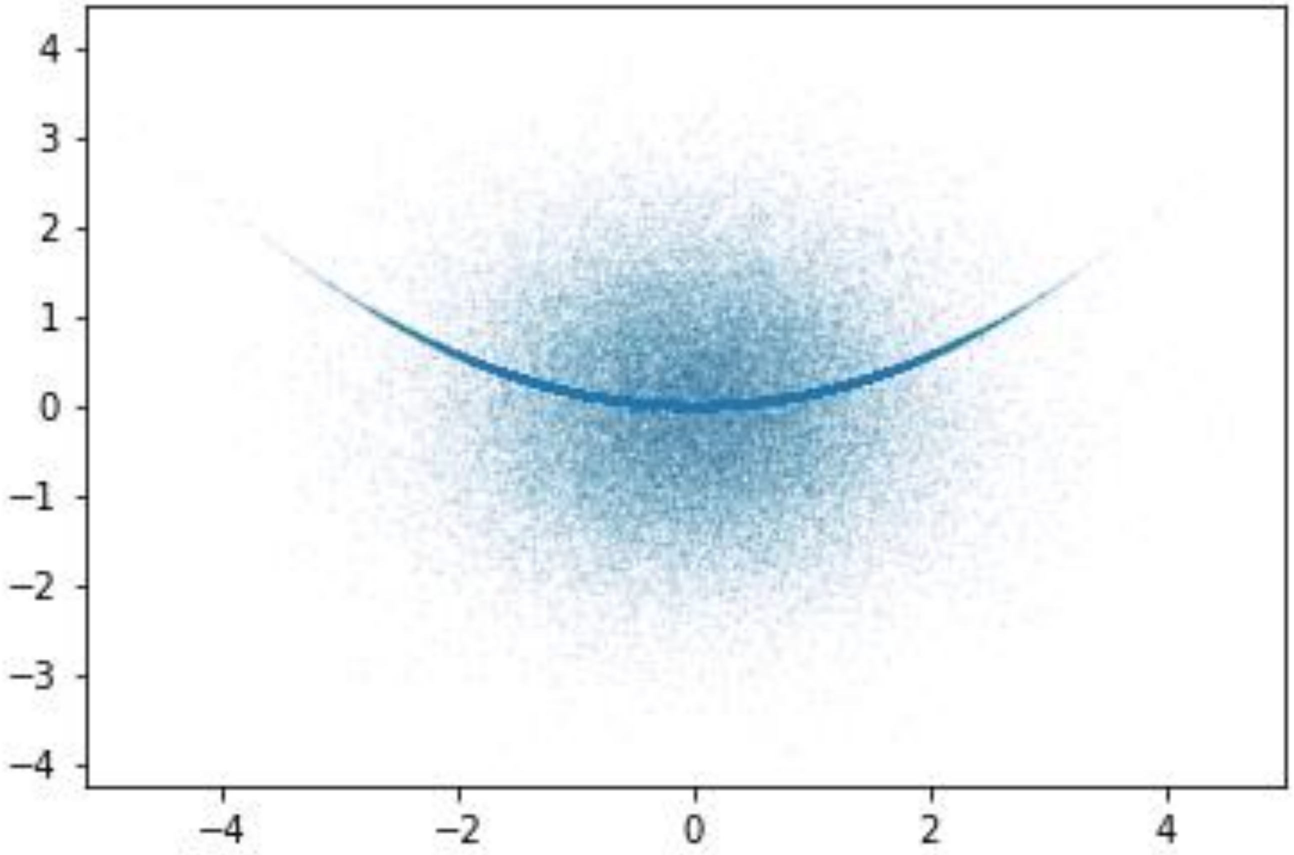
"Ehhhh, I don't know, probably not."

These are not exact quotes, just based on recollection, please don't take them too seriously!

Dessert

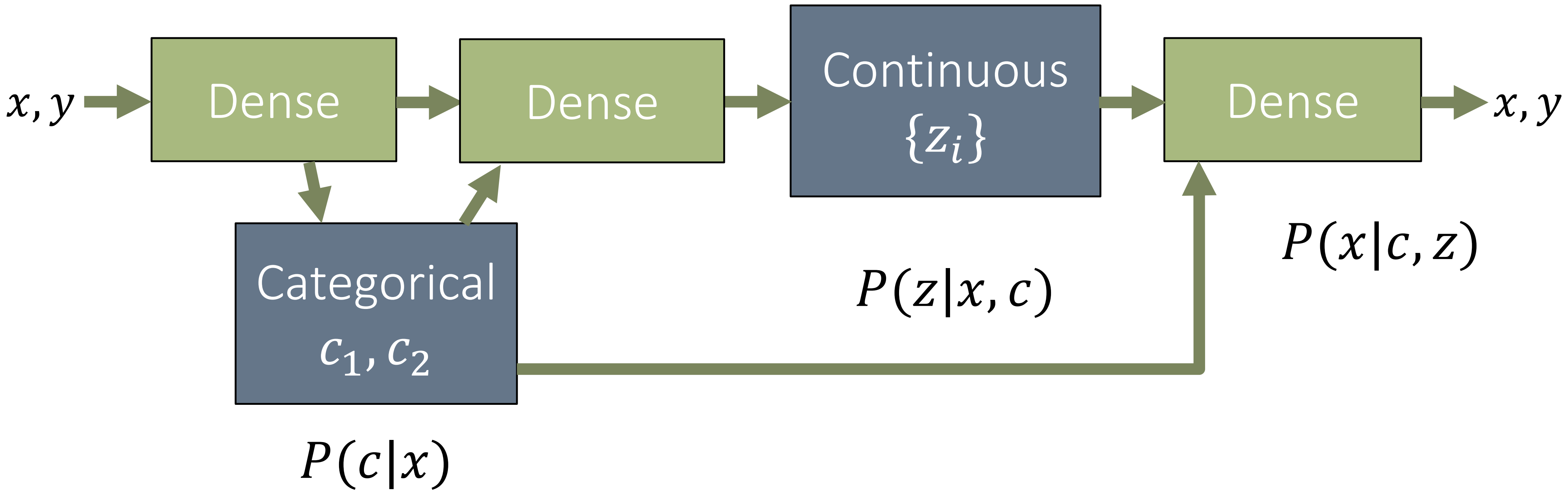
Unsupervised Classification

A Mixed Sample



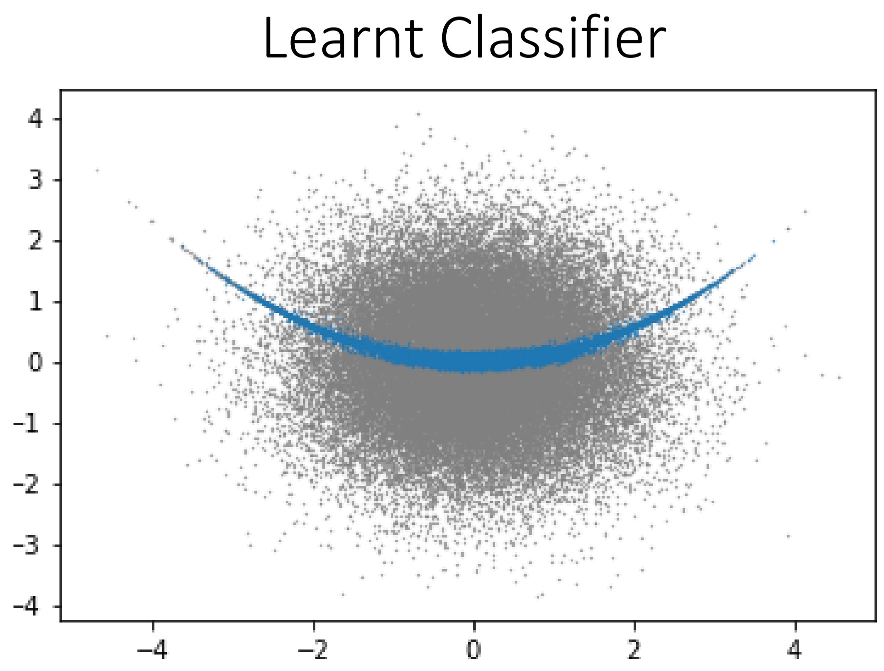
A Mixed Sample

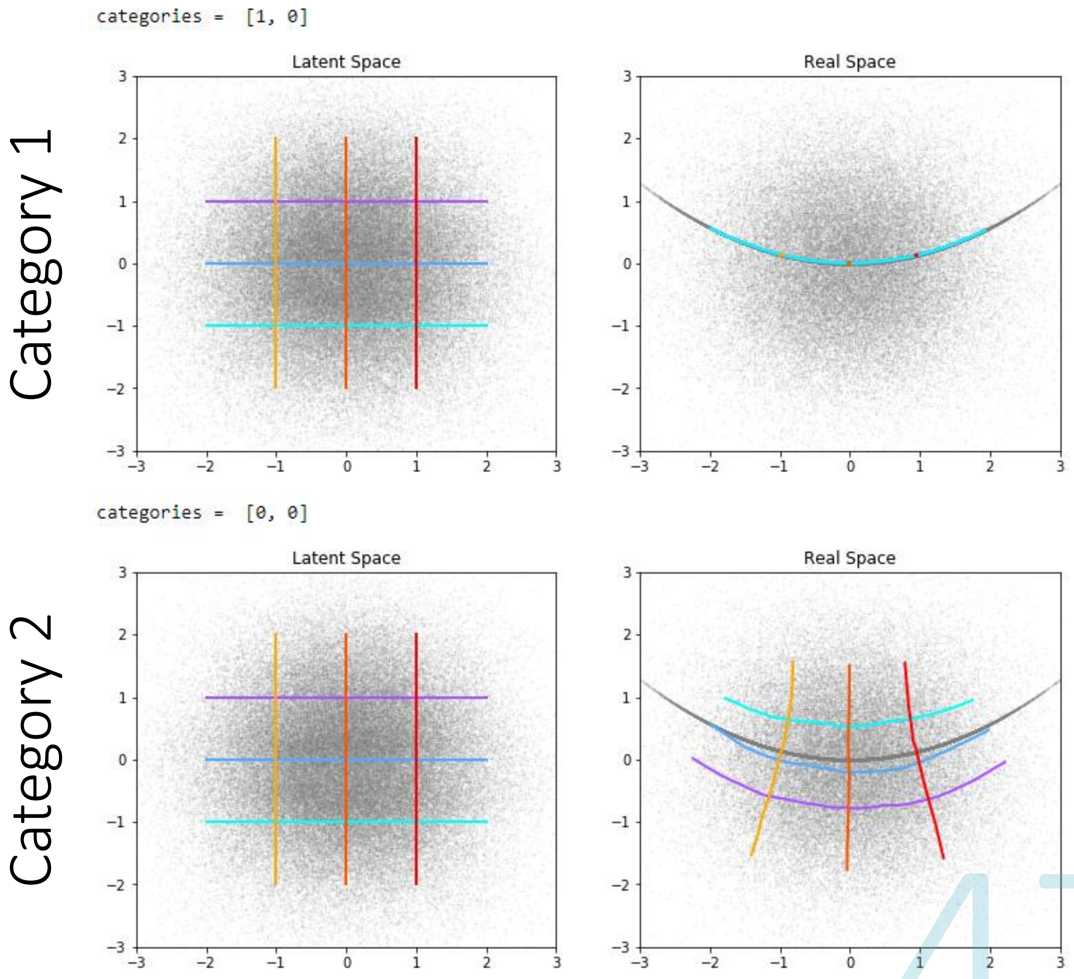
VAE structure



A Mixed Sample

VAE structure

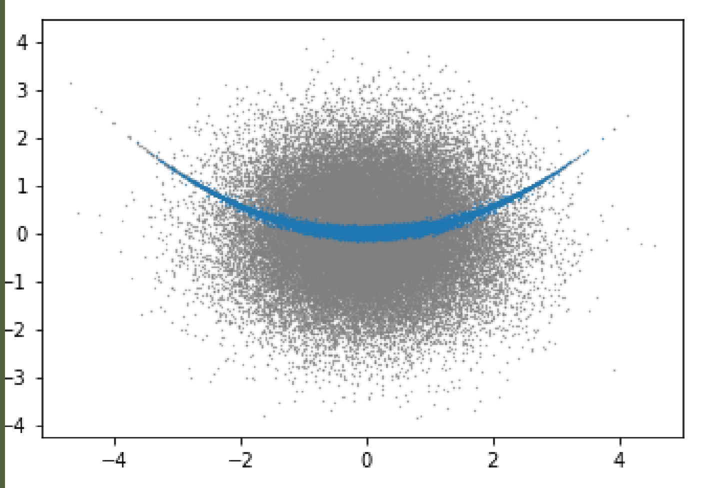


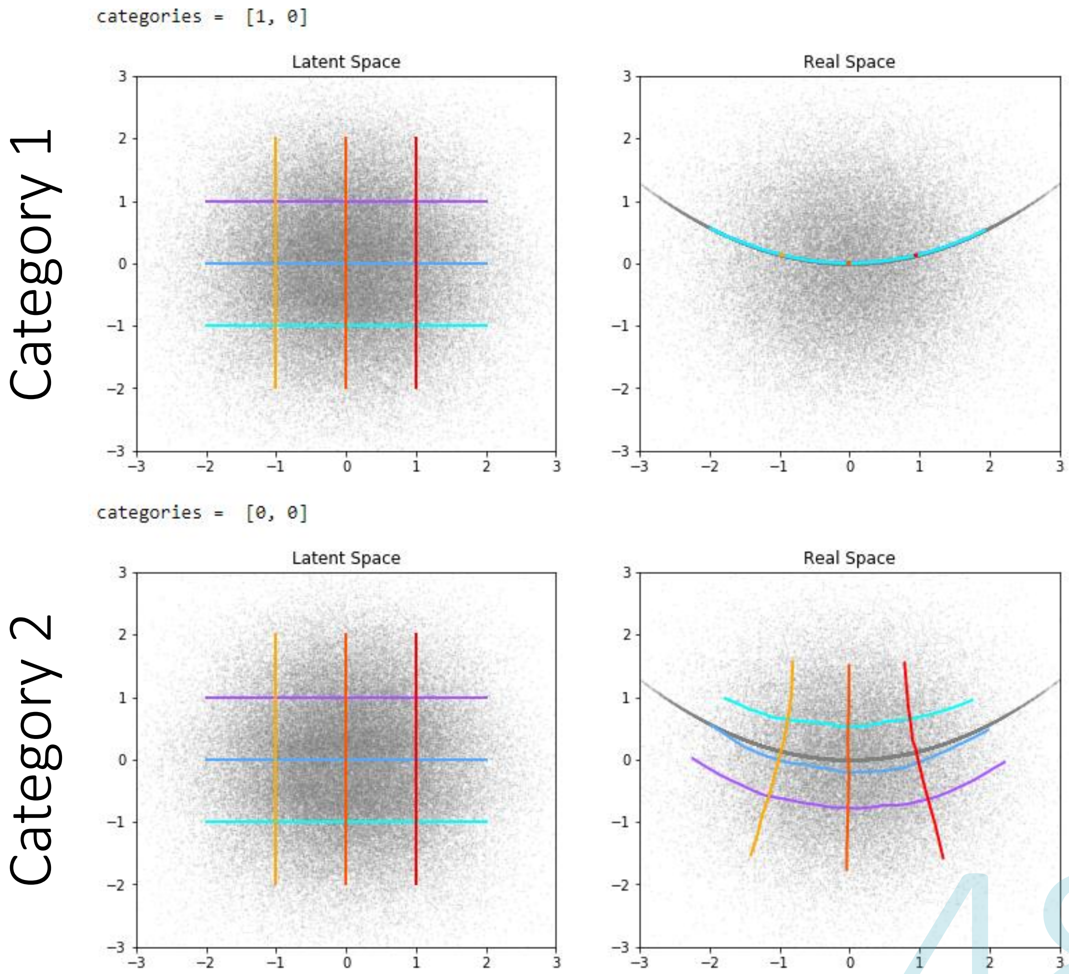


A Mixed Sample

VAE structure

The VAE has spontaneously learnt to classify banana vs not banana.

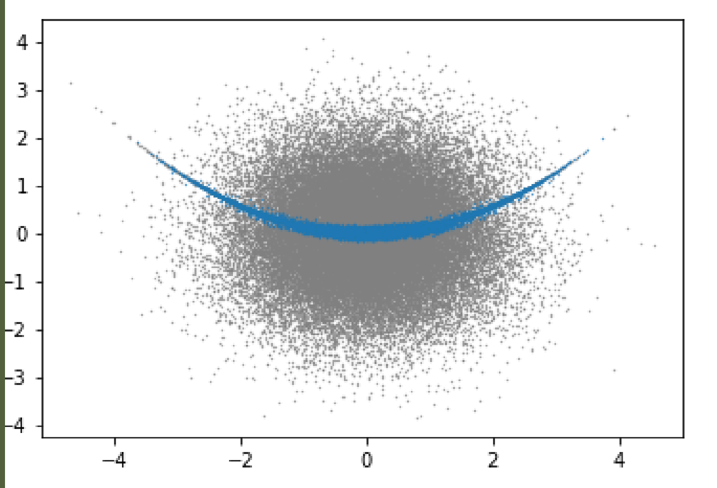




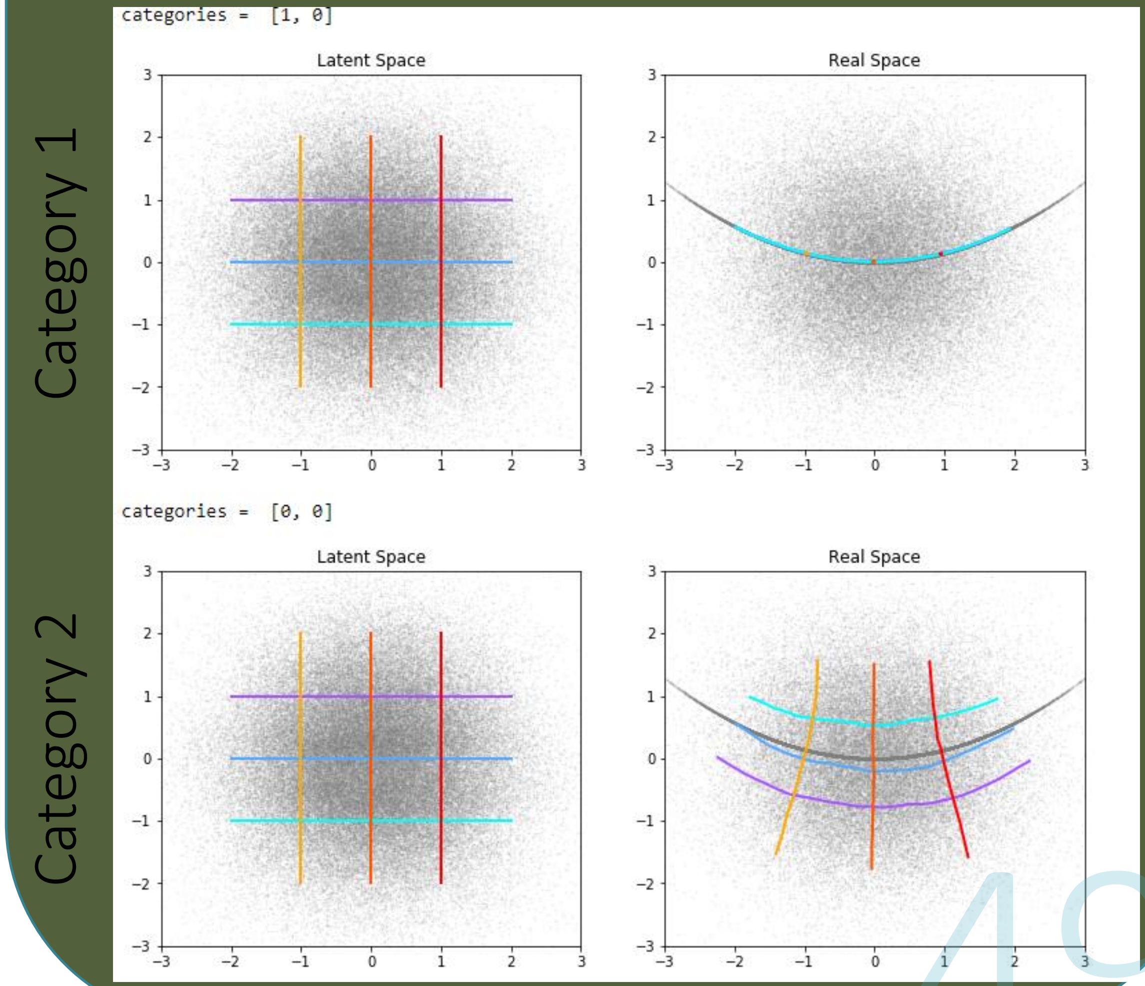
A Mixed Sample

VAE structure

The VAE has spontaneously learnt to classify banana vs not banana.



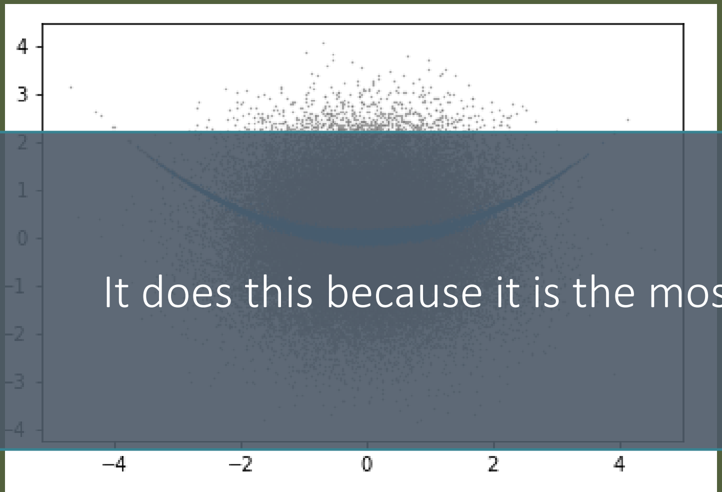
The VAE has learnt to map banana class to a 1D manifold, and not-banana class to a 2D manifold.



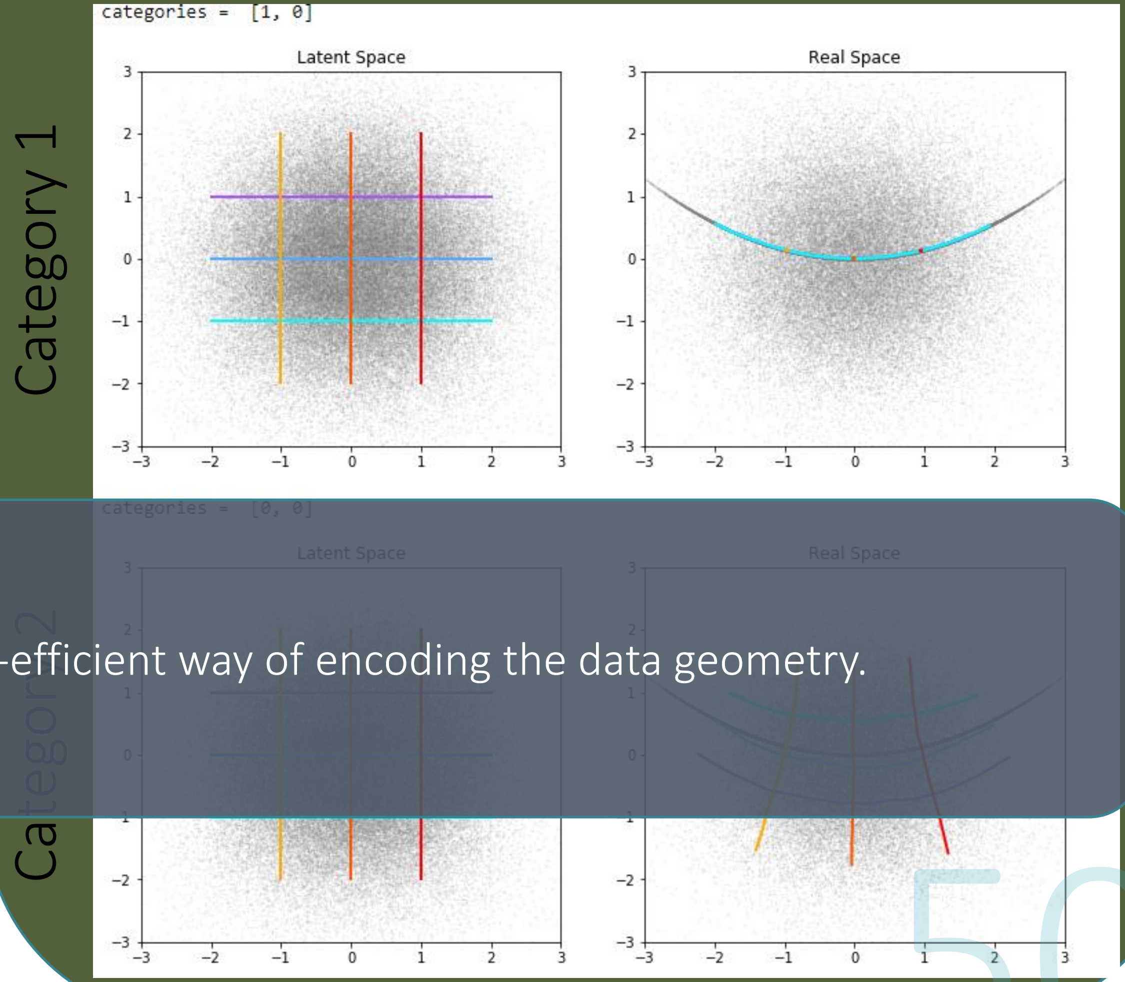
A Mixed Sample

VAE structure

The VAE has spontaneously learnt to classify banana vs not banana.



The VAE has learnt to map banana class to a 1D manifold, and not-banana class to a 2D manifold.



It does this because it is the most information-efficient way of encoding the data geometry.

A Note on Topology

The regular Gaussian VAE is trying to learn a mapping from the real data manifold M to the latent space R^N, because that is the structure imposed on the latent space.

The real data manifold might not be topologically equivalent to R^N. E.g. the φ coordinate of the jet is on S^1. In this case the plain VAE learns to cut the circle at an arbitrary position, which is not ideal. If I give it a latent space in $R^N \times (S^1)^M$, it should optimally learn to put periodic coordinates on S^1's… What about S^k?

A mixed sample is a superposition of manifolds $M_1 \times M_2 \times$…. This can be modelled using a categorical variable before the continuous ones.

My philosophy: give the VAE as many options for latent category and topology as I can think of and practically implement, and then attempt to learn the structure of the dataset by studying how it chooses to use them. Take latent dim -> infinity.

Is this new? I have never seen other people take this approach before, but I'm ignorant of the literature.

Palate Cleanser
Conclusions

The VAE is trying to learn a simple representation of the ***geometry*** of the data manifold on which it is trained.

The latent space statistics can be studied to learn about the learnt geometry.

If the geometry of the data manifold reflects the underlying physics responsible for generating it, then maybe the geometry can teach us about the physics.

Special thanks to

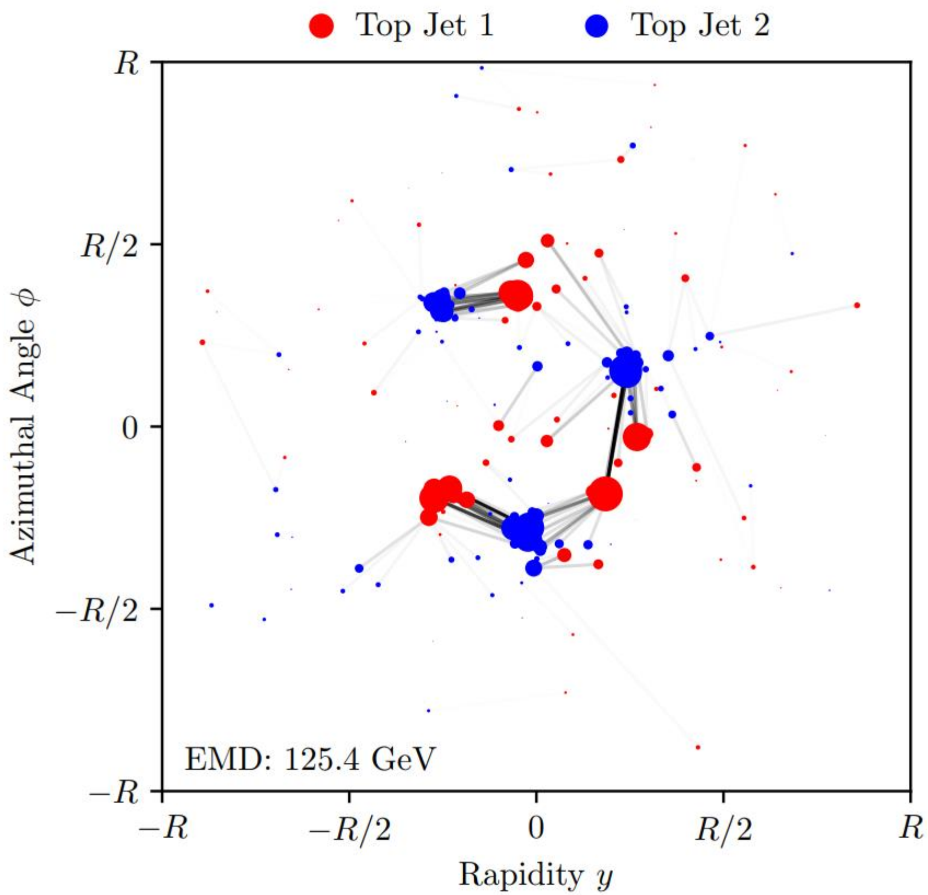
Digestif

Analytic results for Dimensionality

(blackboard)

Reconstruction Error

Sinkhorn Distance ≈ EMD



arXiv:1902.02346

Sinkhorn's algorithm; start with $\Delta R_{ij}, p_{Ti}, p_{Tj}$ then:

$K_{ij} = \exp(\Delta R_{ij}/\tau)$
$u_i = \mathbf{1}_i$
$v_i = \mathbf{1}_j$

Repeat N times:

$u_i = p_{Ti}/(K.v)_i$
$v_i = p_{Tj}/\left(K^T.u\right)_j$

Return $T_{ij} = u_i K_{ij} v_j$

The Variational Autoencoder

Doesn't suffer from curse of dimensionality

Toy data generated from:

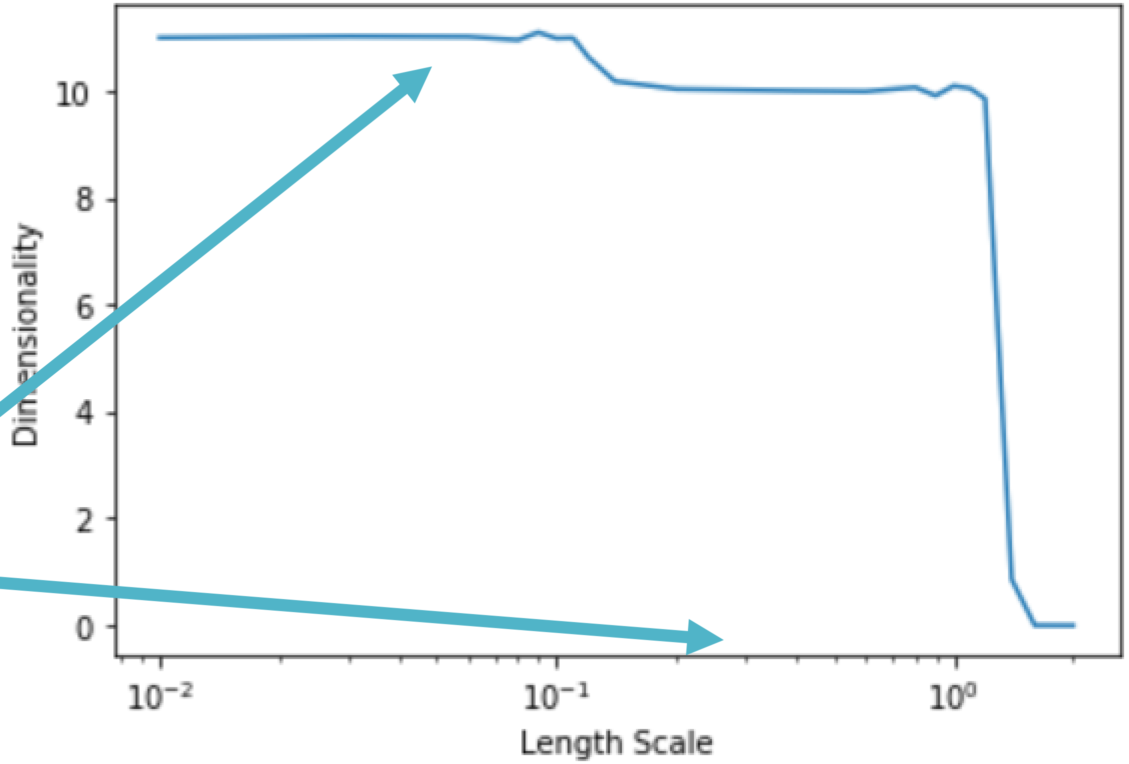
$$P(\vec{x}) = [\prod_{i=1}^{10} N_i(\mu = 0, \sigma = 1)]\ N_{11}\ (\mu = 0, \sigma = 0.1)$$

With $N_{tot} = 5 * 10^5$ points

Typical distance to neighbour $\sim N_{tot}^{-1/10} \sim 0.3$

Correlation dimension runs into sparsity limit before the small dimension is even discovered!

The VAE finds the small dimension.



Future Directions

1. What is the point?

2. Alternative latent priors?

3. Alternative metrics?

The Variational Autoencoder

ML Engineer:

"A VAE is a fancy AE with regulated stochastic latent space sampling"

Bayesian statistician:

*"A VAE is a probability model trained to extremize the **E**vidence **L**ower **BO**und on the posterior distribution p(x)"*

The Variational Autoencoder:

Dimensionality

$$\langle|\Delta \boldsymbol{x}|^2\rangle = \sum\langle|\Delta x_i|^2\rangle = D\rho^2 + \sum_{i>D} S_i^2$$

$$D = \frac{d\langle|\Delta \boldsymbol{x}|^2\rangle}{d\rho^2}$$

Setting $\frac{dL}{d\sigma} = 0$ implies:

1. $\rho = \beta$
2. $D = \frac{d\,KL}{d\log\beta}$