Searching for new physics with autoencoders

ML4Jets
November 16, 2018

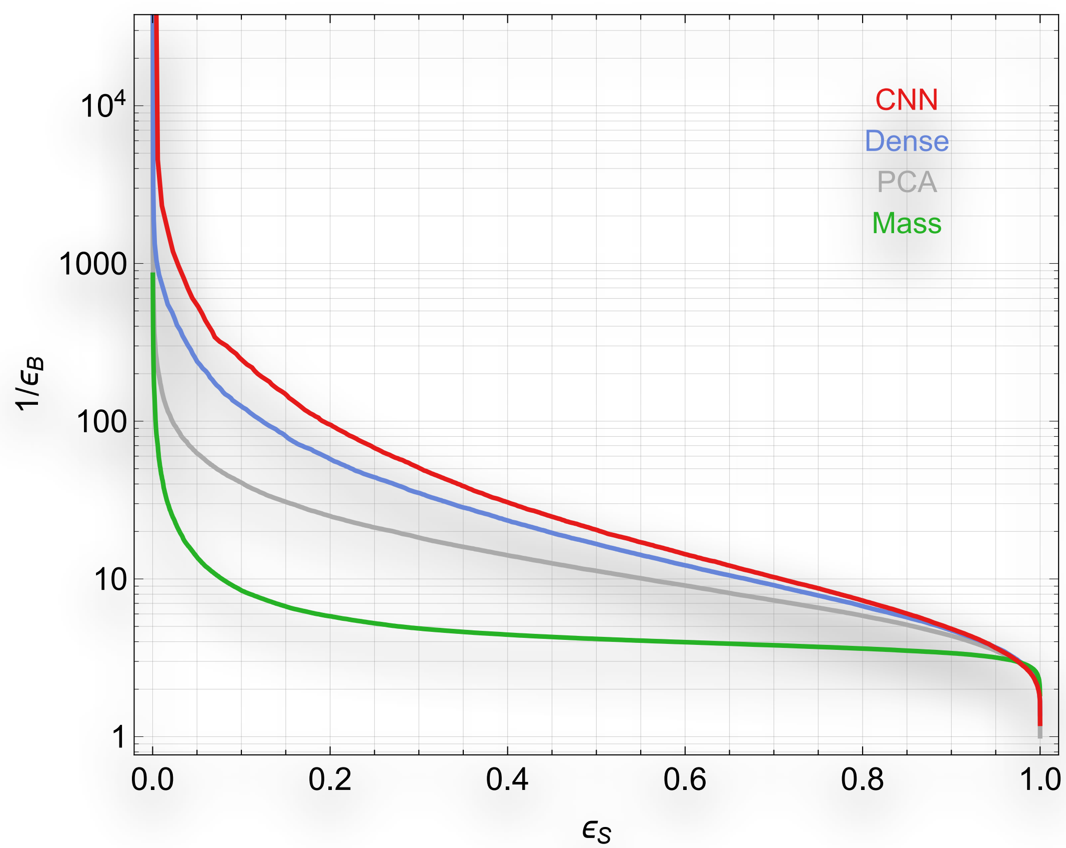
Marco Farina
Stony Brook University

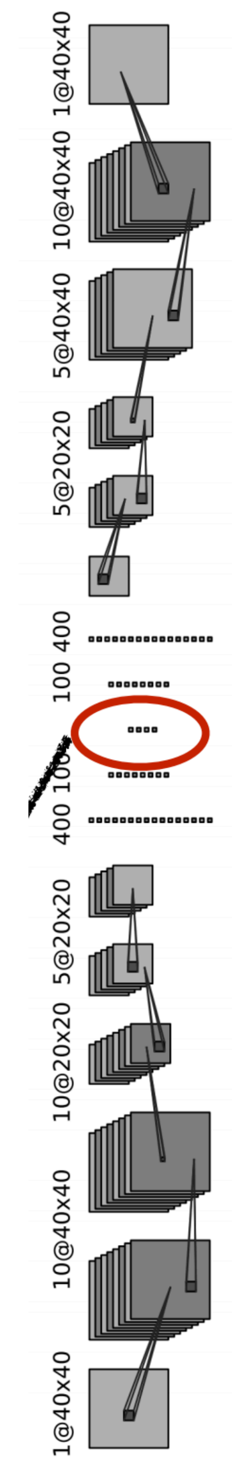
Based on Farina, Nakai, Shih ‘18

Thanks Gregor!

Does it work?

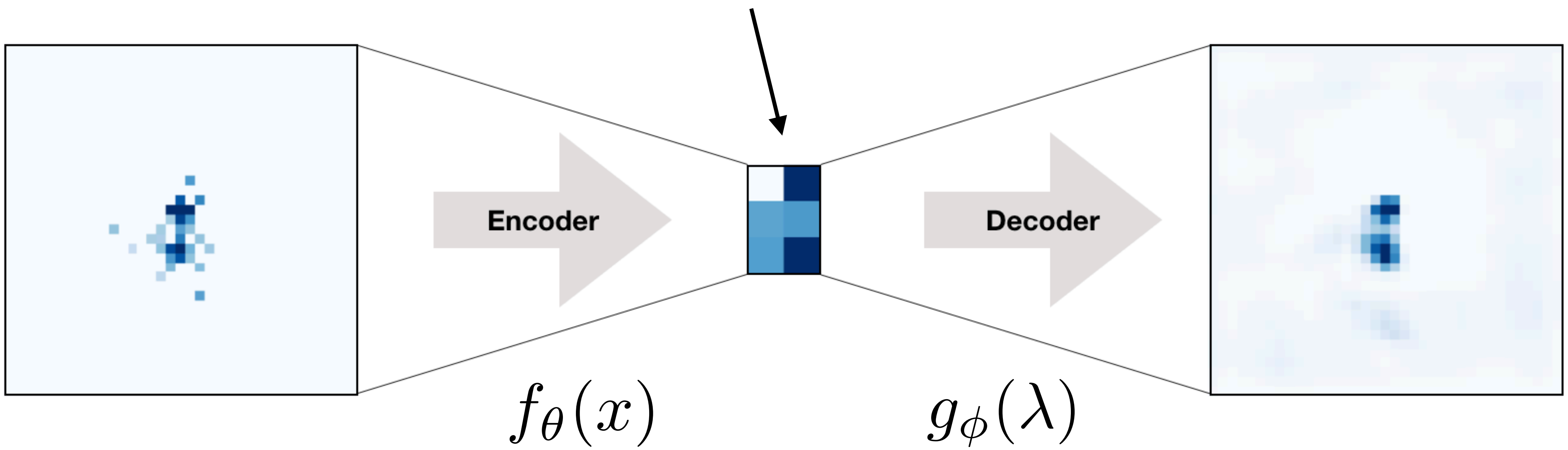
- Train on QCD only
- Test on top vs QCD
- Cut on loss function as discriminator
  - Large loss → autoencoding failure → anomaly





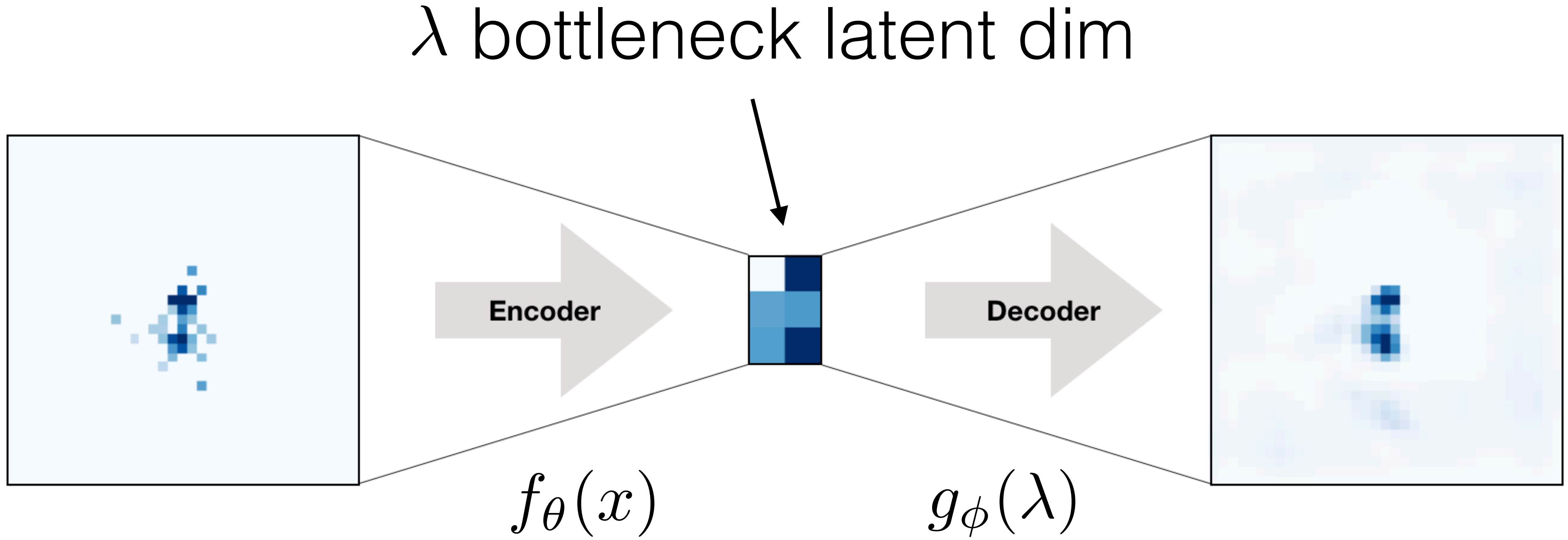
Autoencoders

λ bottleneck latent dim



$$\ell(\theta, \phi) = |x - g_\phi(f_\theta(x))|^2$$

Autoencoders

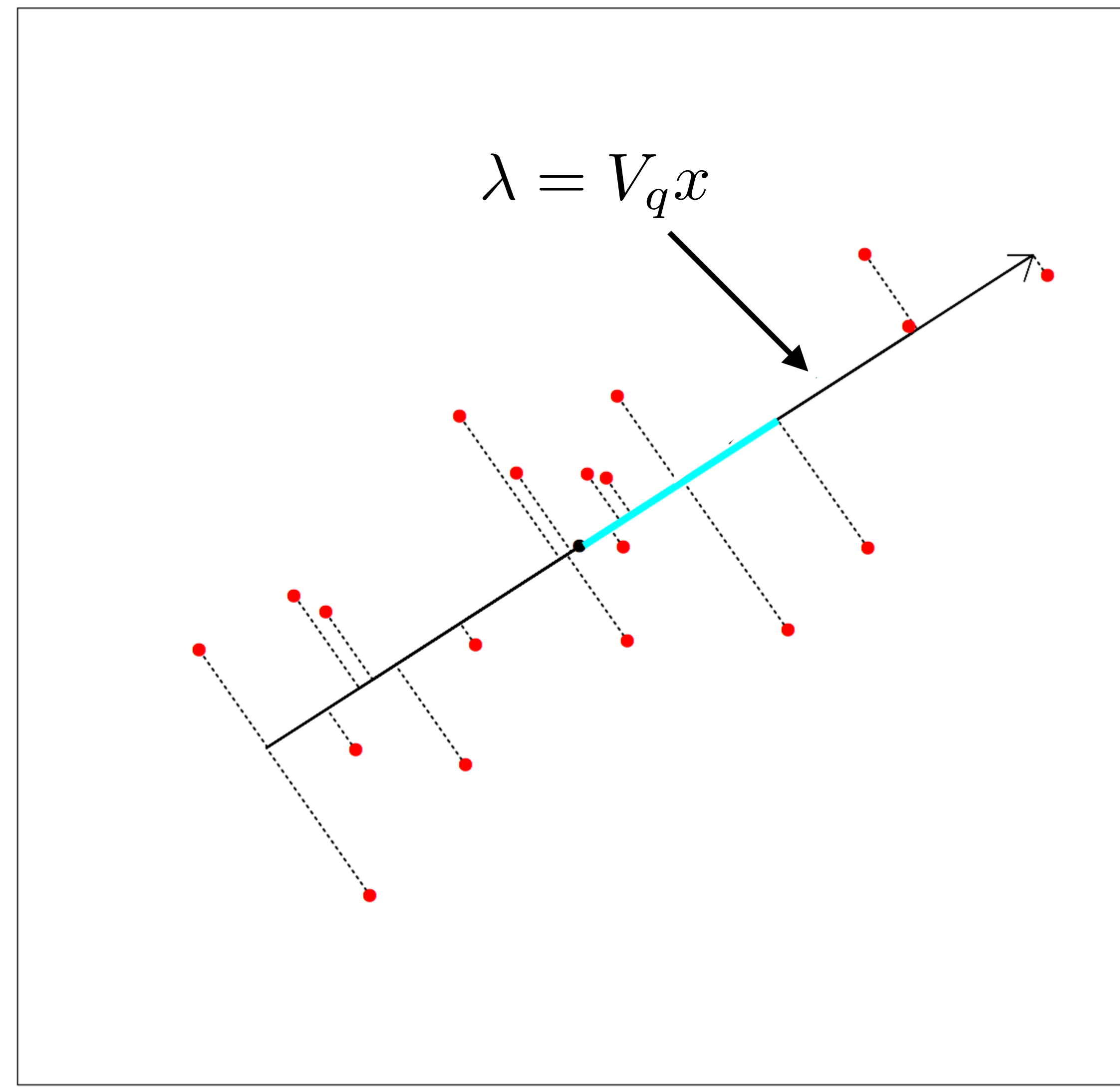


- PCA
- DNN
- CNN

PCA

Principal component analysis: linear transformation
Lower dimensional representation of data $q < n$

$$\ell(V_q) = |x - V_q^T V_q x|^2$$



Find best V_q

Maximum variance ⟷ Minimum error

Data

We generated:

- QCD jets as train/background
- tops and 400 GeV gluinos (with 3j RPV decay) as signal/anomaly

Transformed sample in images

Preprocessing as in Macaluso, Shih '18

	CMS
Jet sample	13 TeV $p_T \in (800, 900)$ GeV, $\lvert\eta\rvert < 1$ PYTHIA 8 and DELPHES particle-flow match: $\Delta R(t, j) < 0.6$ merge: $\Delta R(t, q) < 0.6$ 1.2M + 1.2M
Image	37×37 $\Delta\eta = \Delta\phi = 3.2$

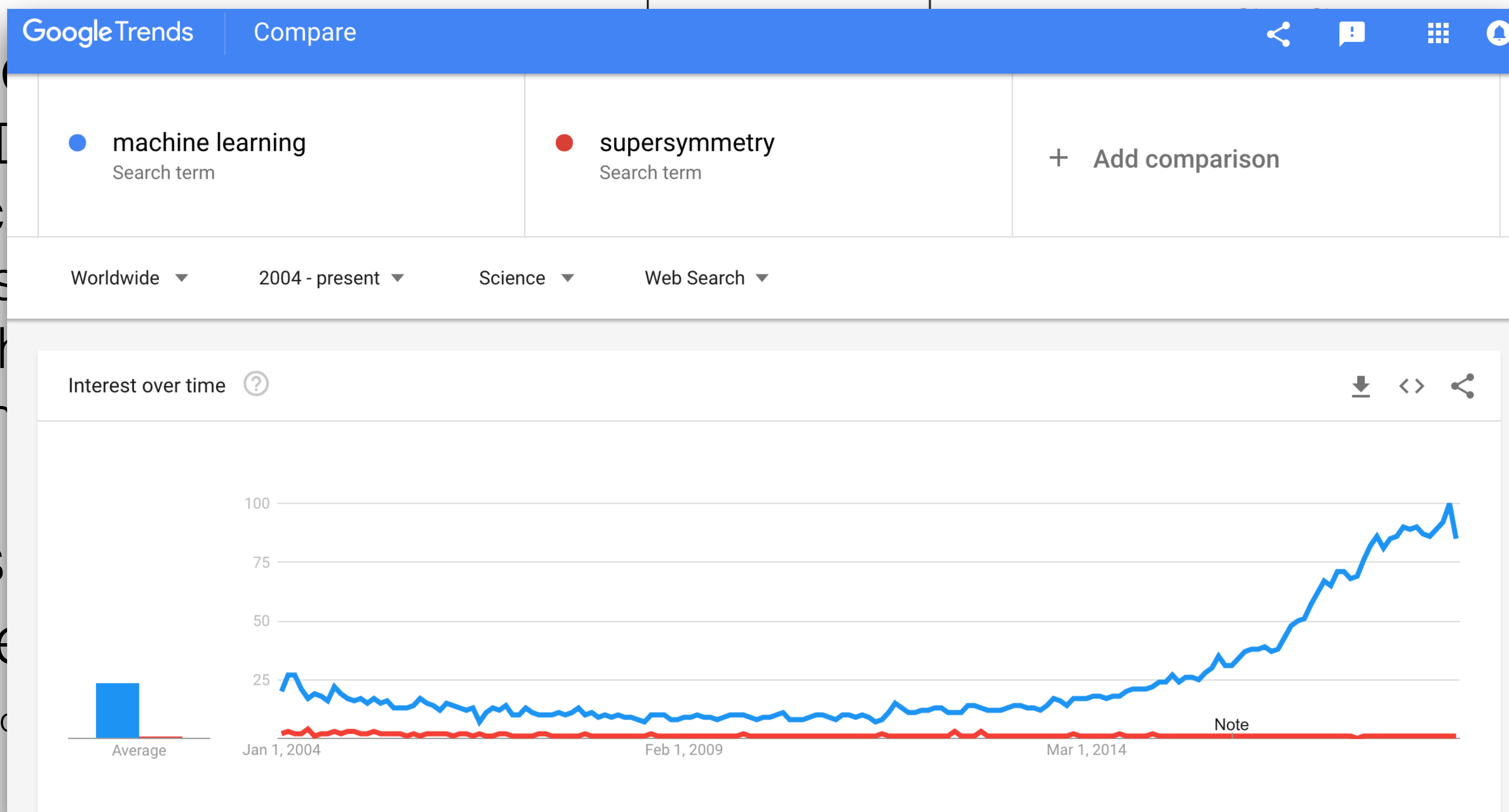
Data

We g

- QCD
  bac
- tops
  (with
  sign

Trans
image

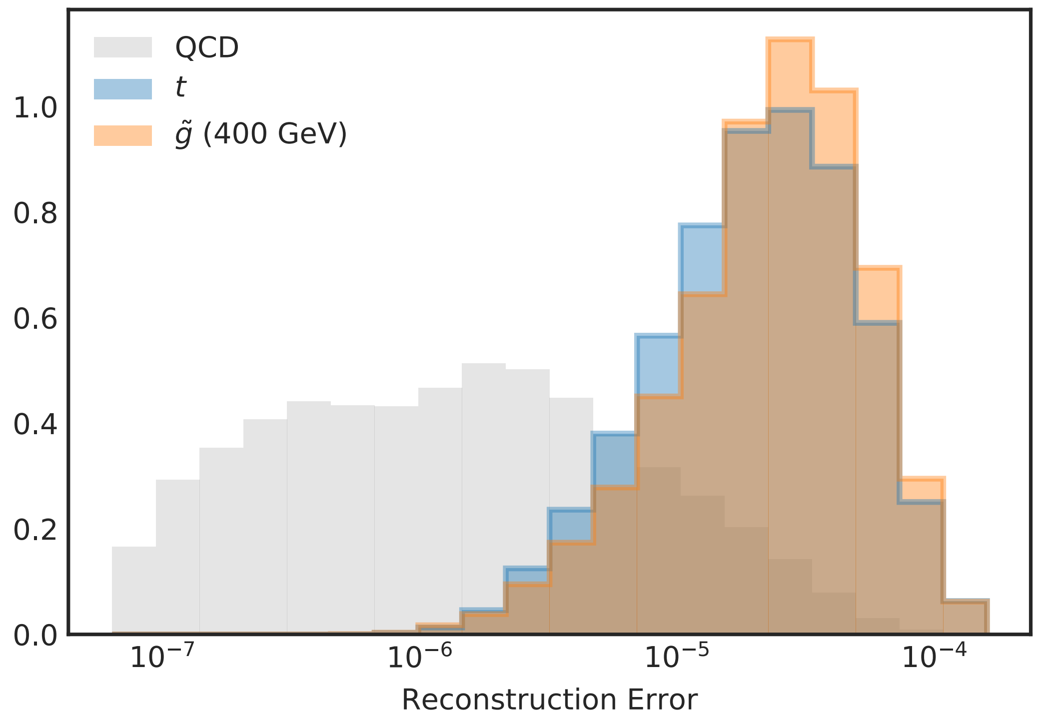
Prepro

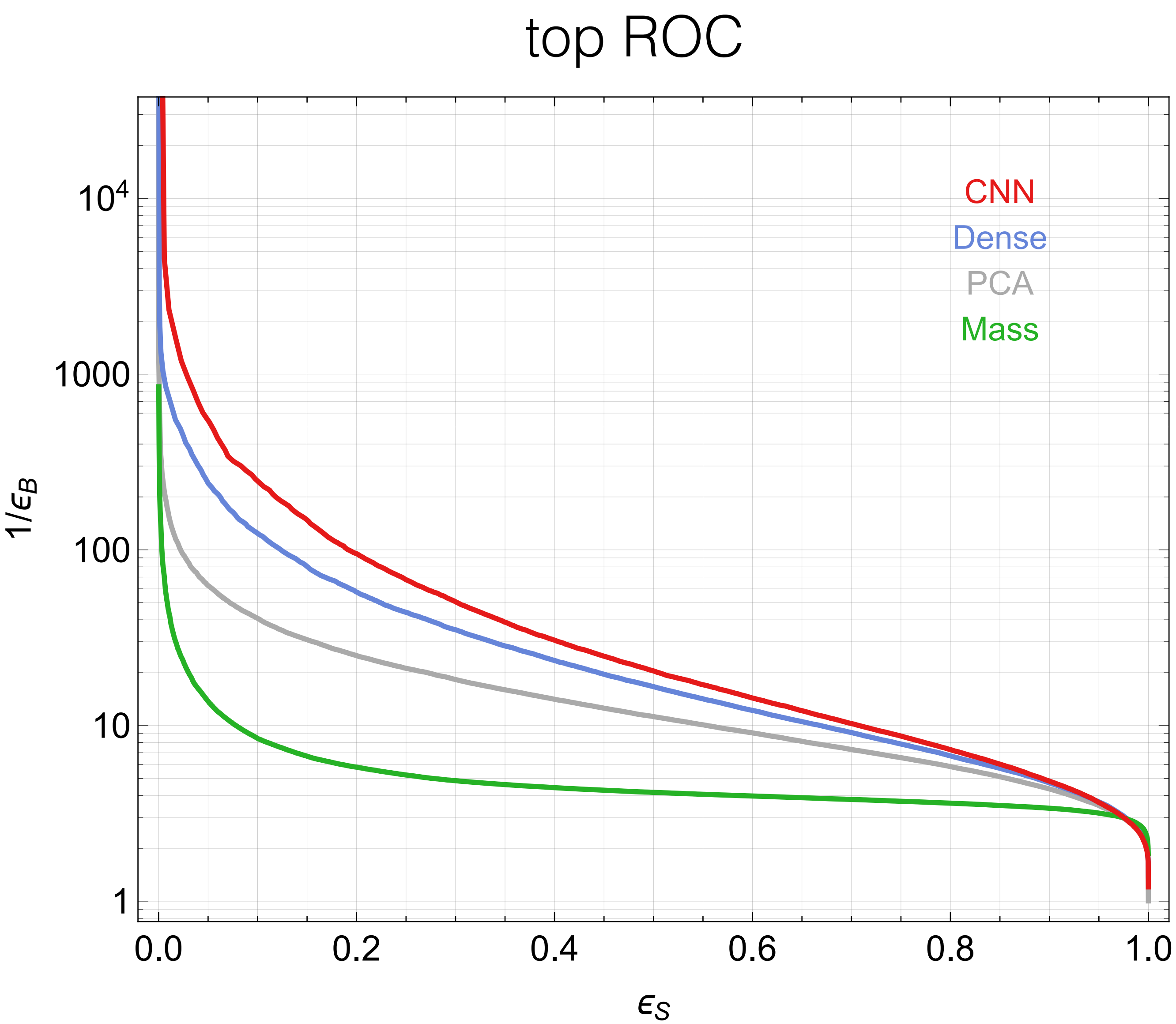


Anomalous jets detection

After training on QCD jets…

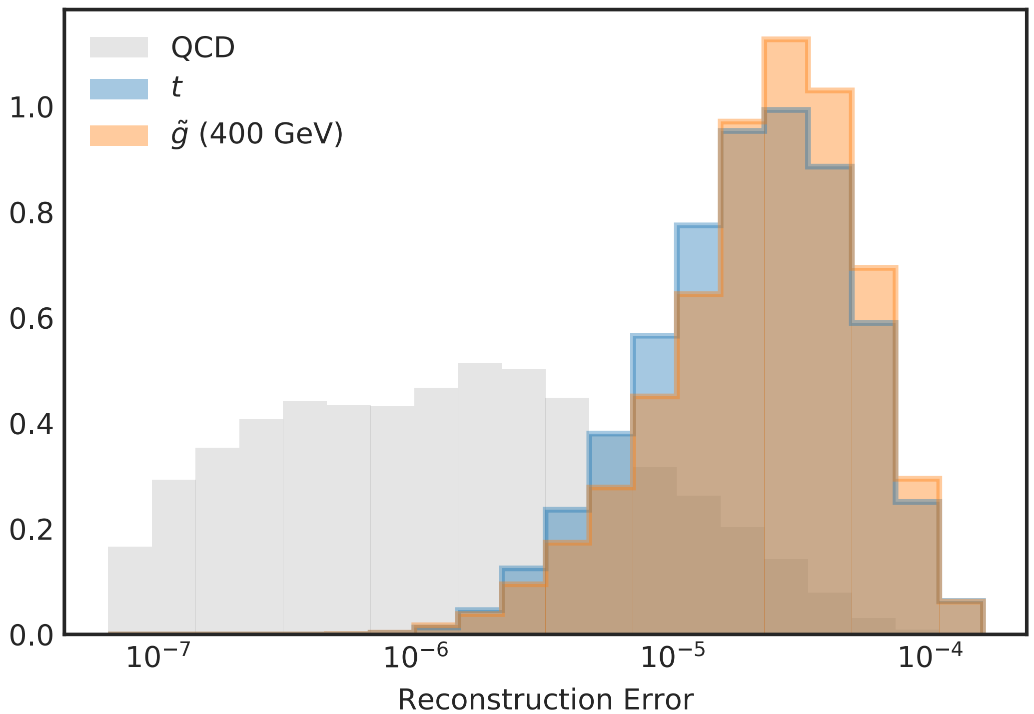
Anomalous jets detection

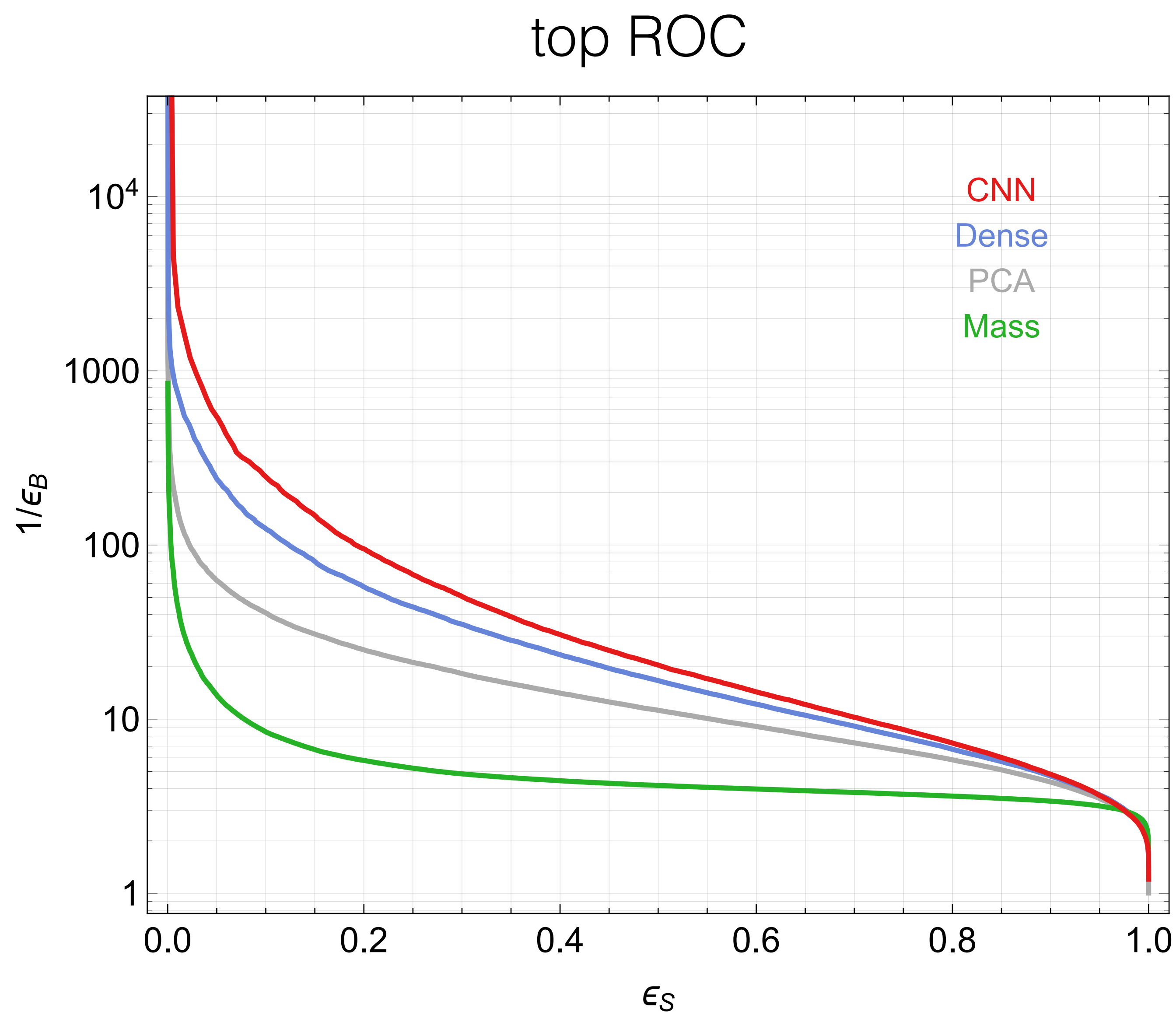




Tag anomaly using cut on reconstruction error

Anomalous jets detection



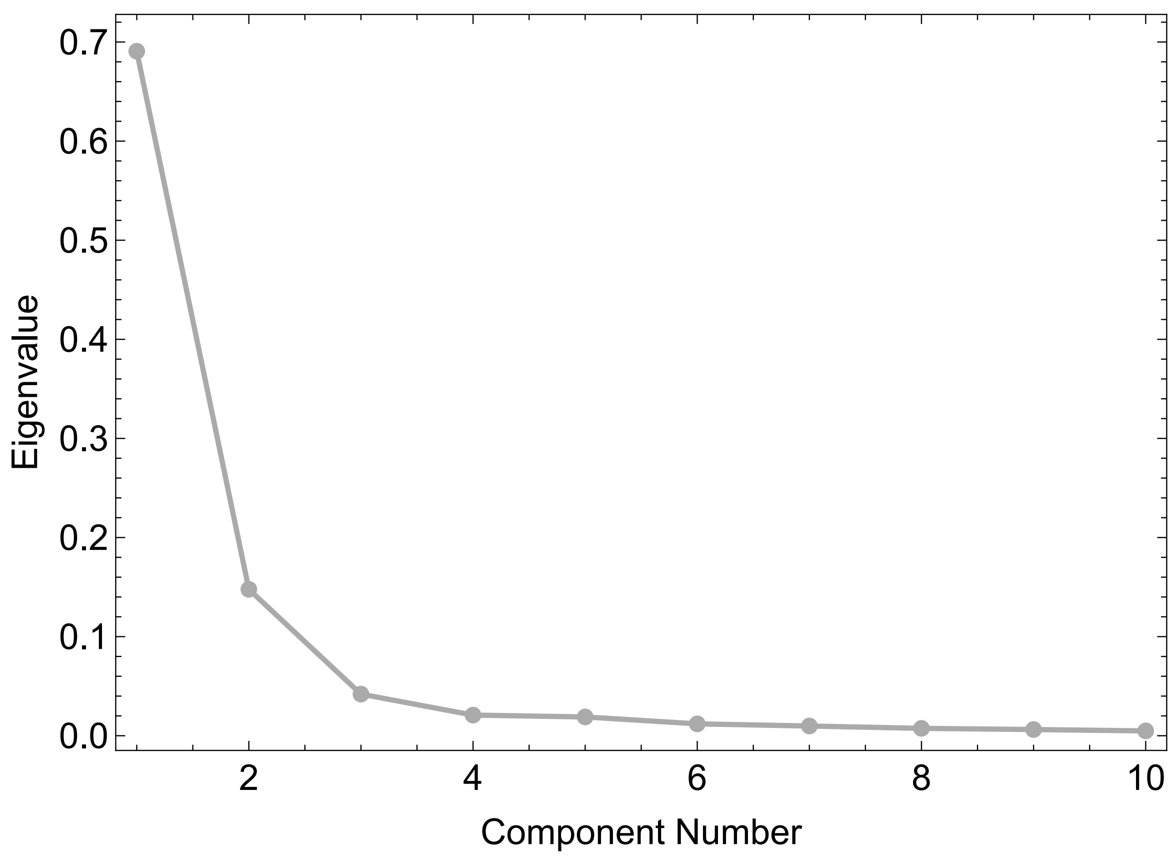


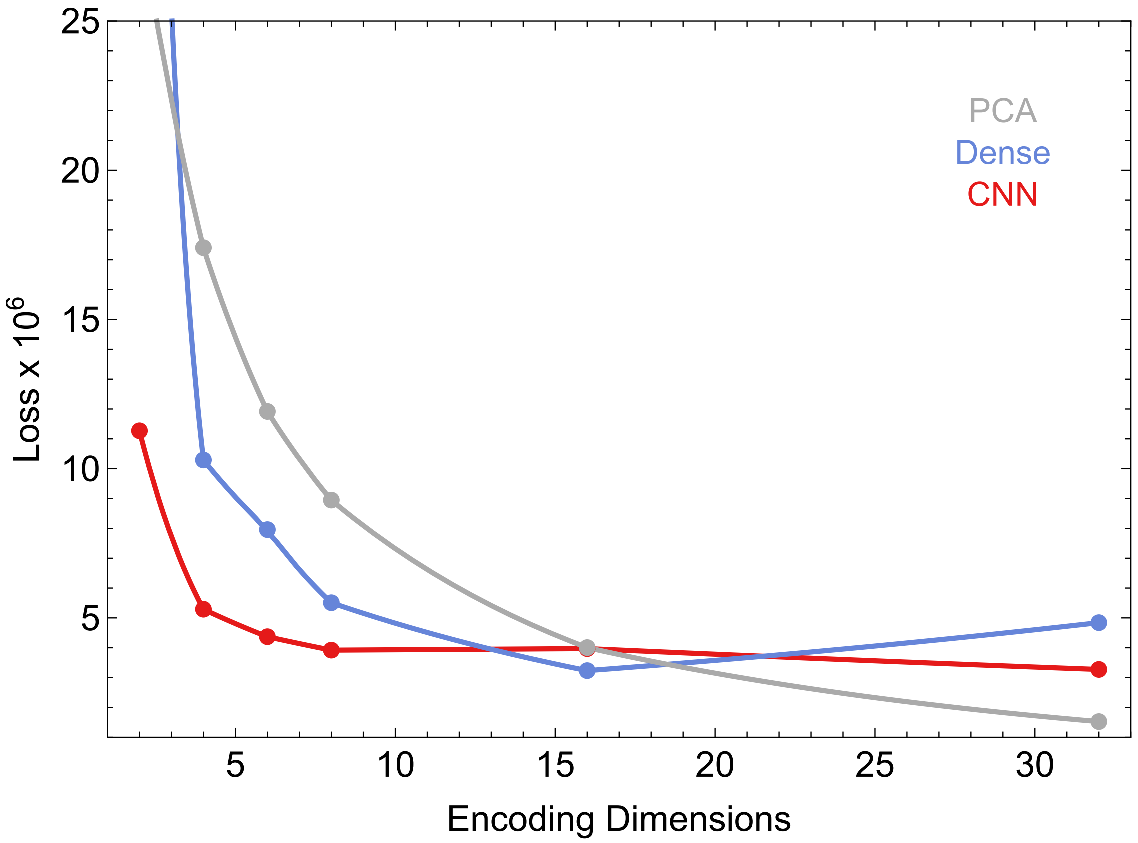
$$E_{10} \sim 0.7$$

$$E_{100} \sim 0.15$$

Encoding dim

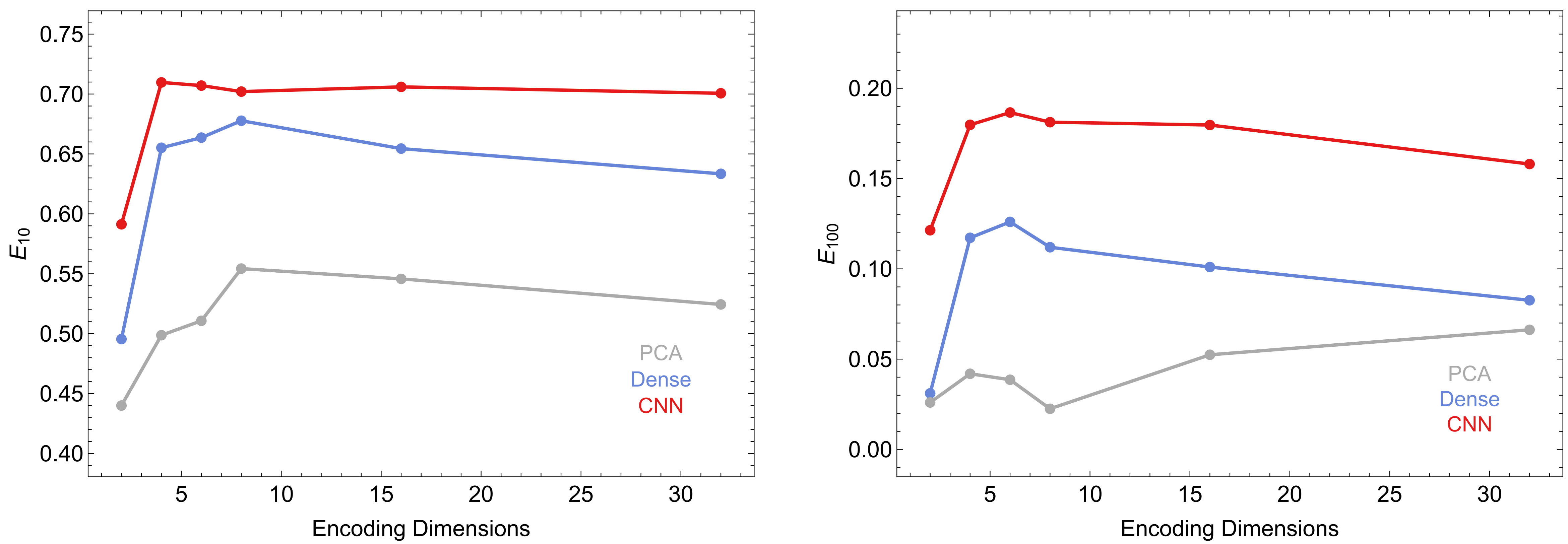
Expressivity vs Triviality





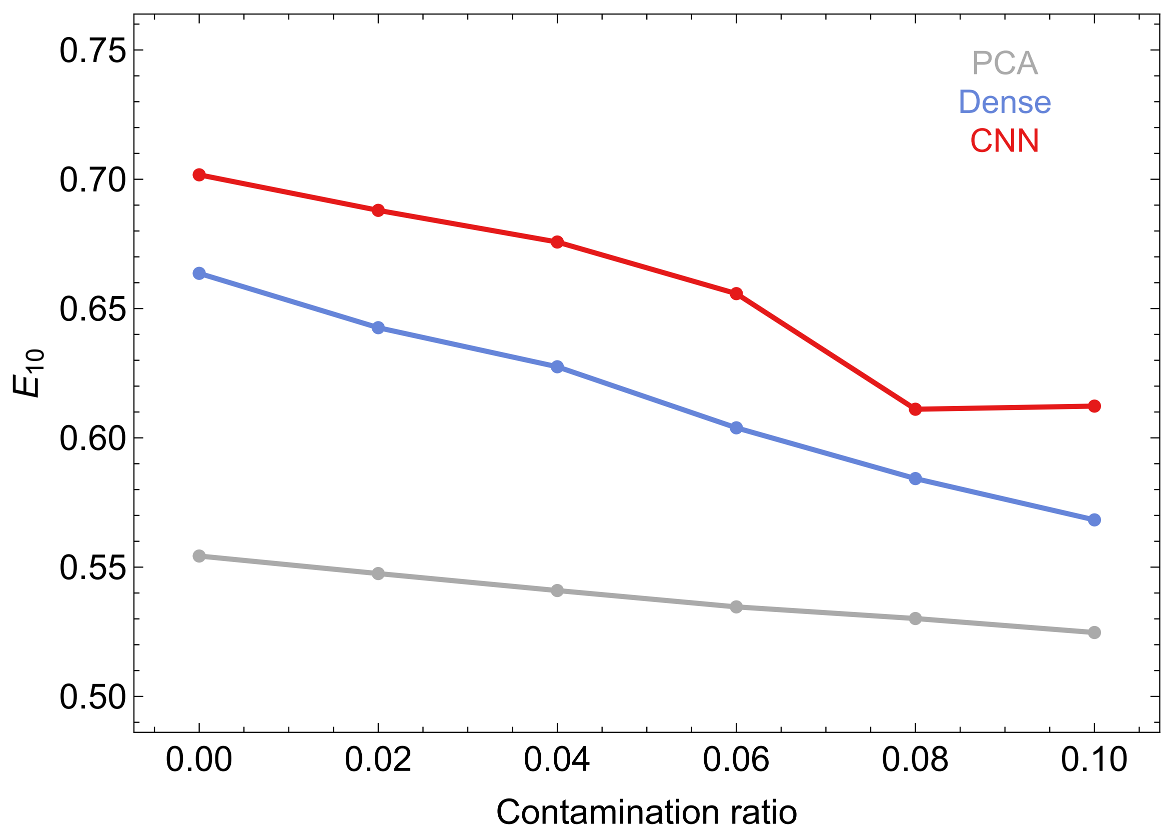
Encoding dim

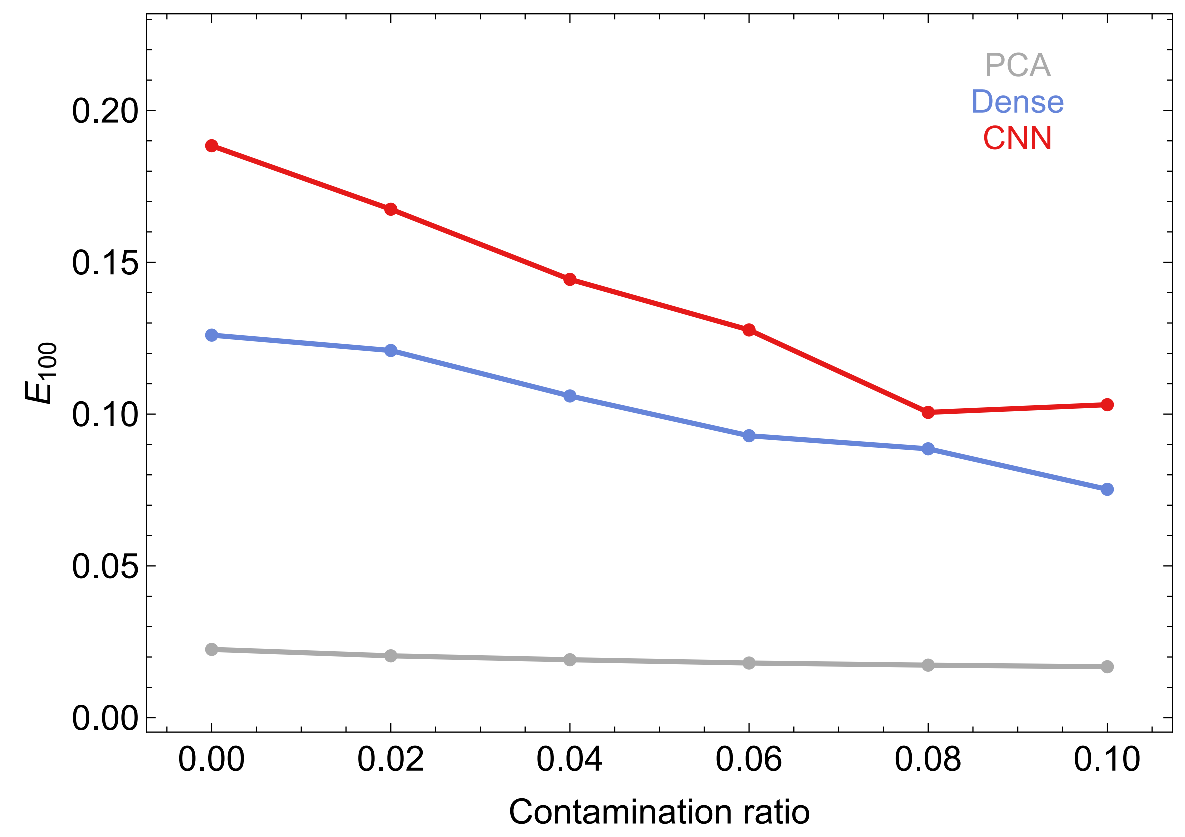
Expressivity vs Triviality



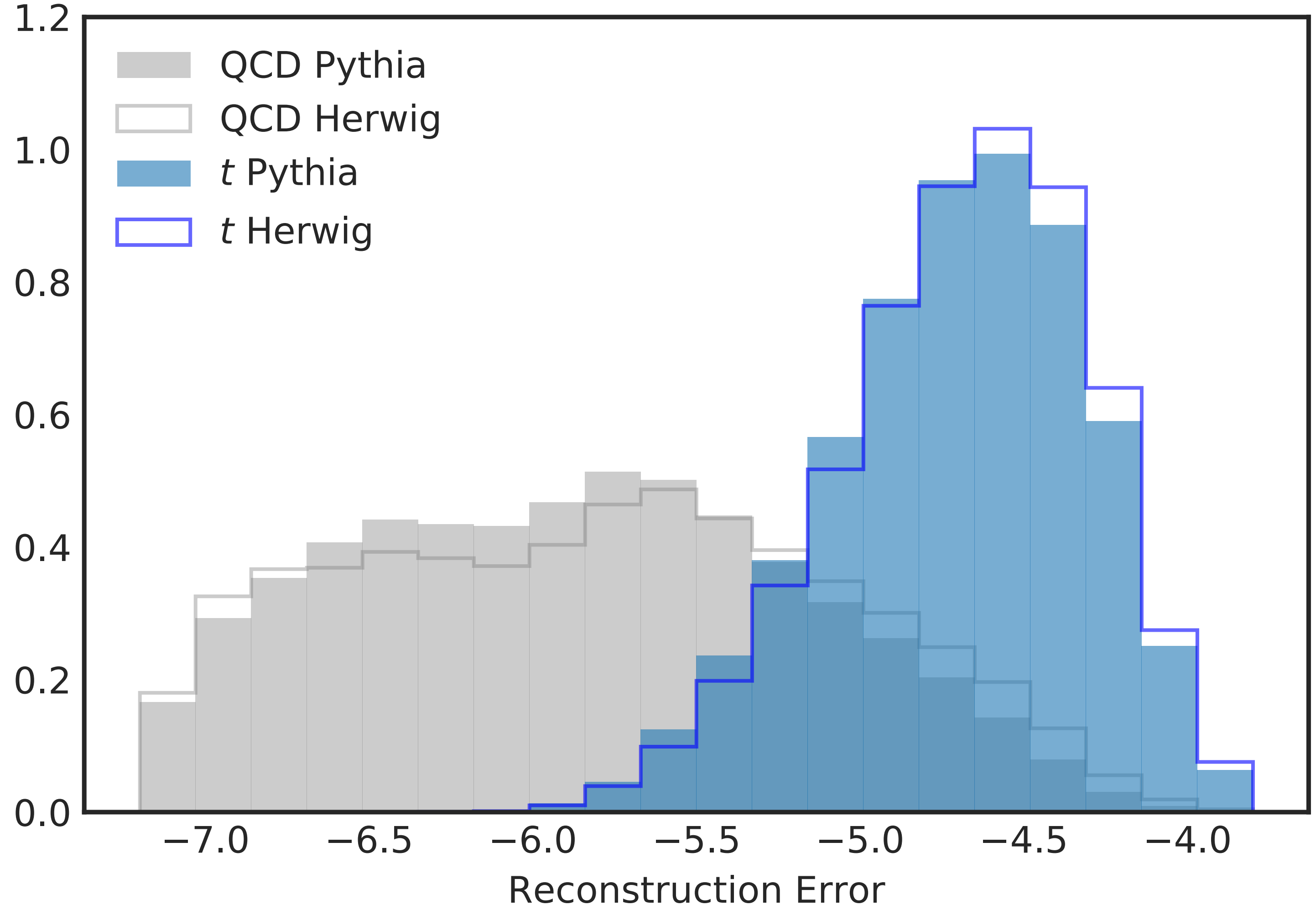
Train set contamination

Cheating so far: weakly supervised vs unsupervised





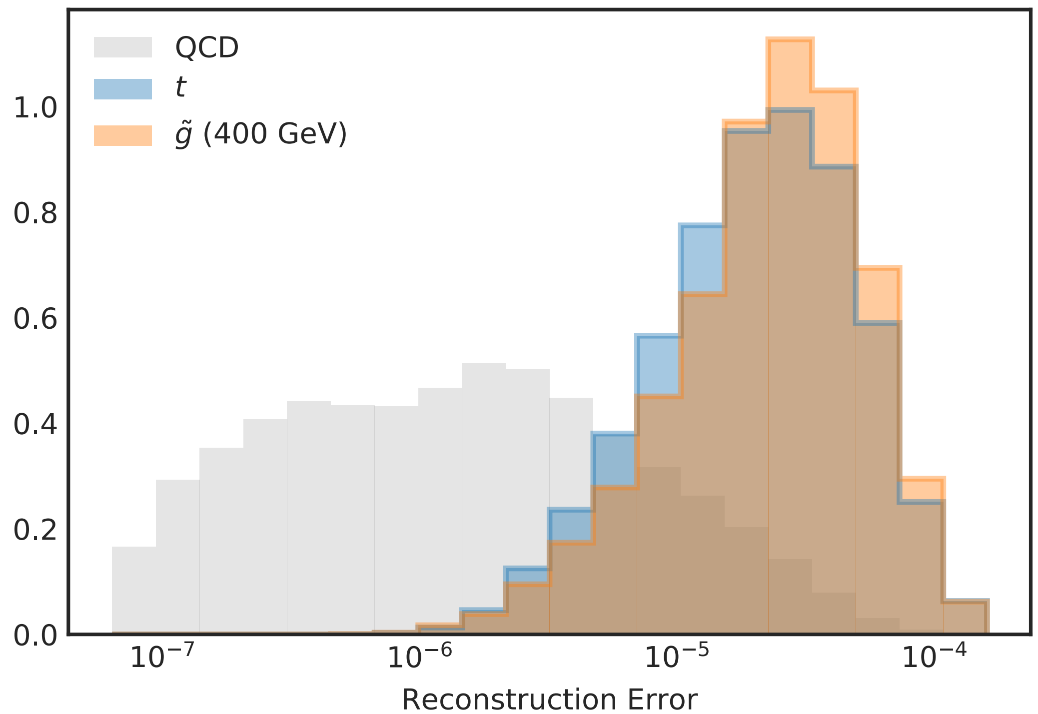
Robustness?



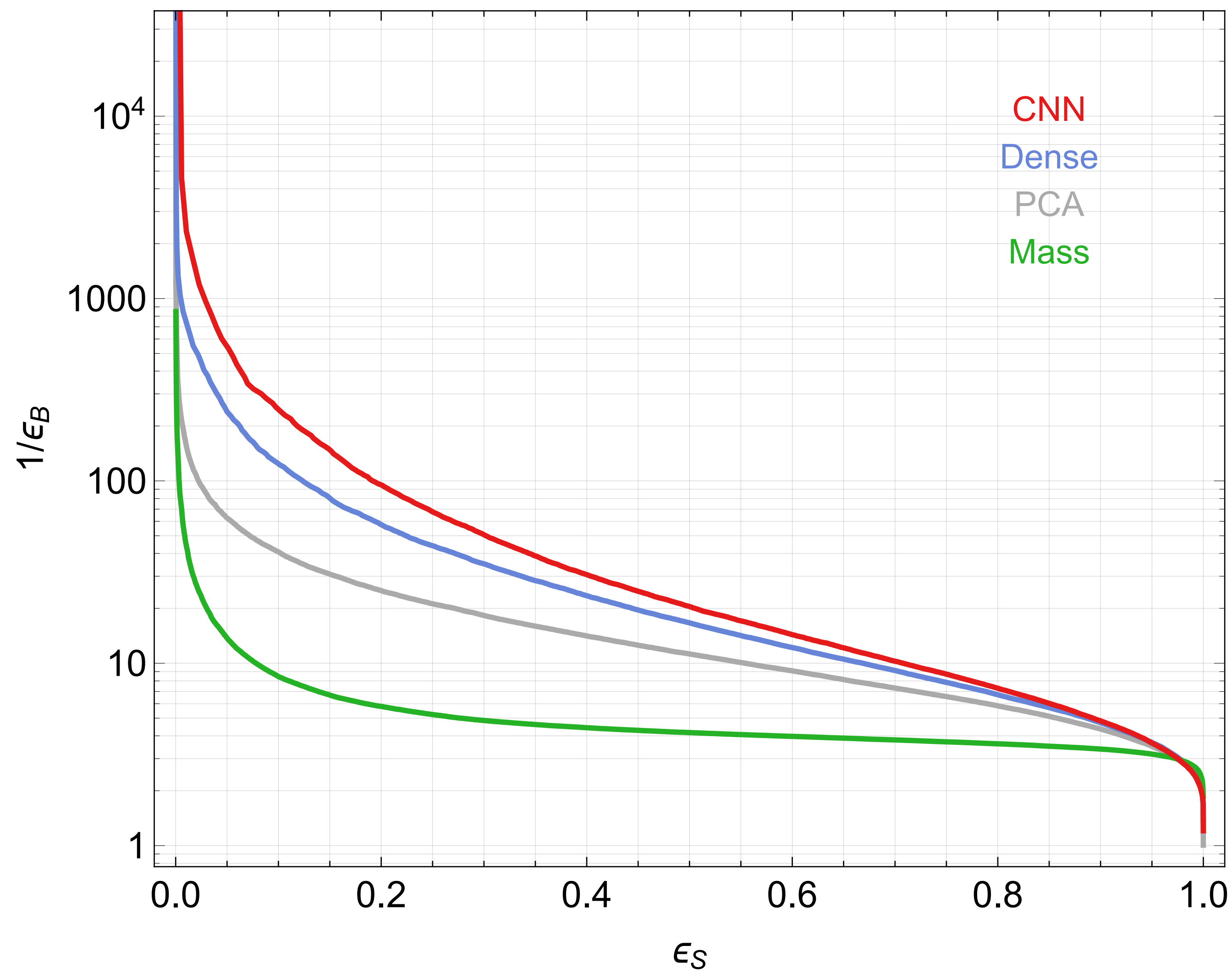
Different generators, as a proxy for training on simulation vs real data

Very qualitative…

Anomalous jets detection



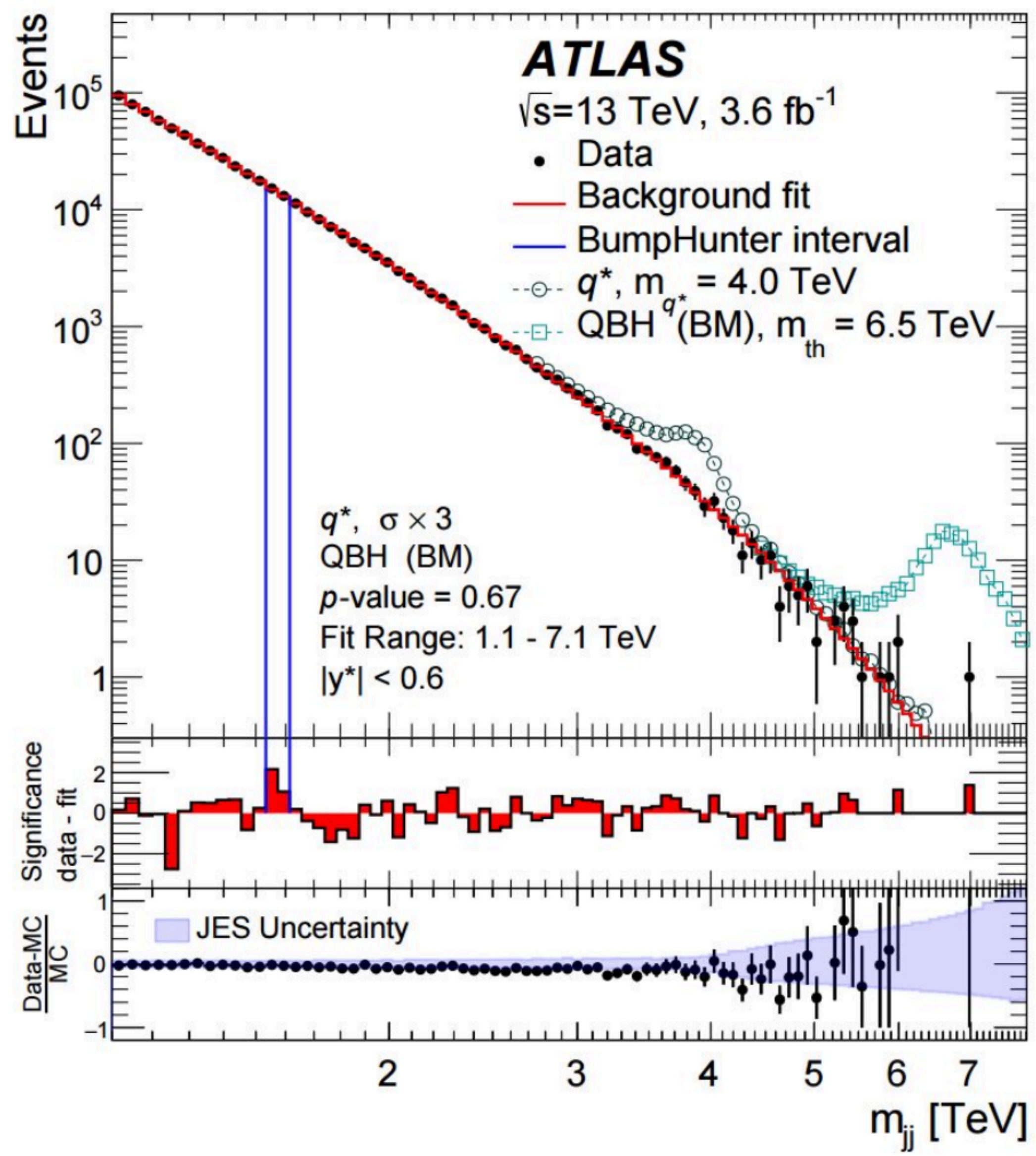
	t	$\tilde{g}$
PCA	0.51 / 0.04	0.98 / 0.36
Dense	0.66 / 0.13	0.90 / 0.39
CNN	0.70 / 0.19	0.77 / 0.23



Build test statistic given expected number of events passing the cut or…

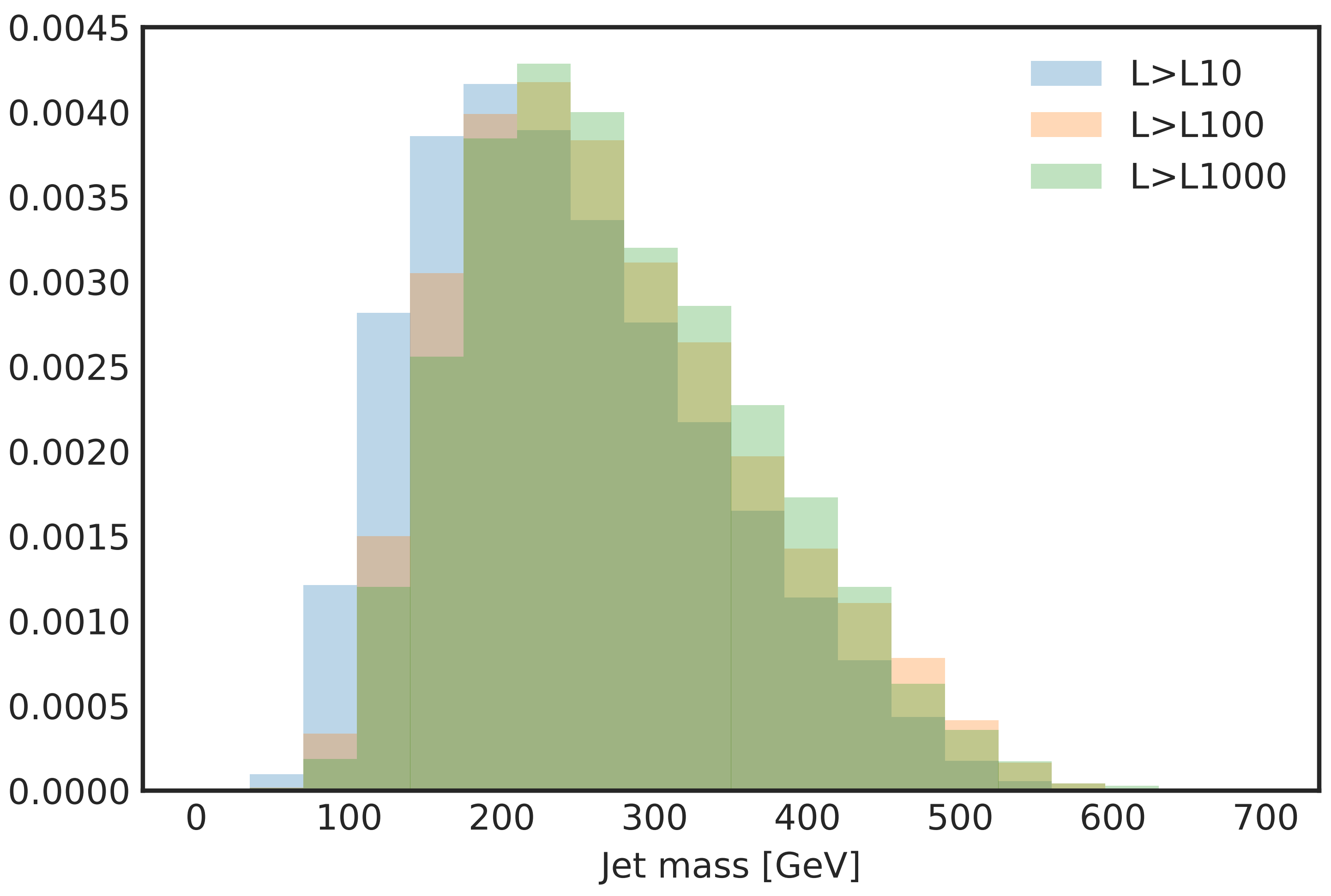
Enhanced bump hunt?

Use AE to reduce background in conventional searches



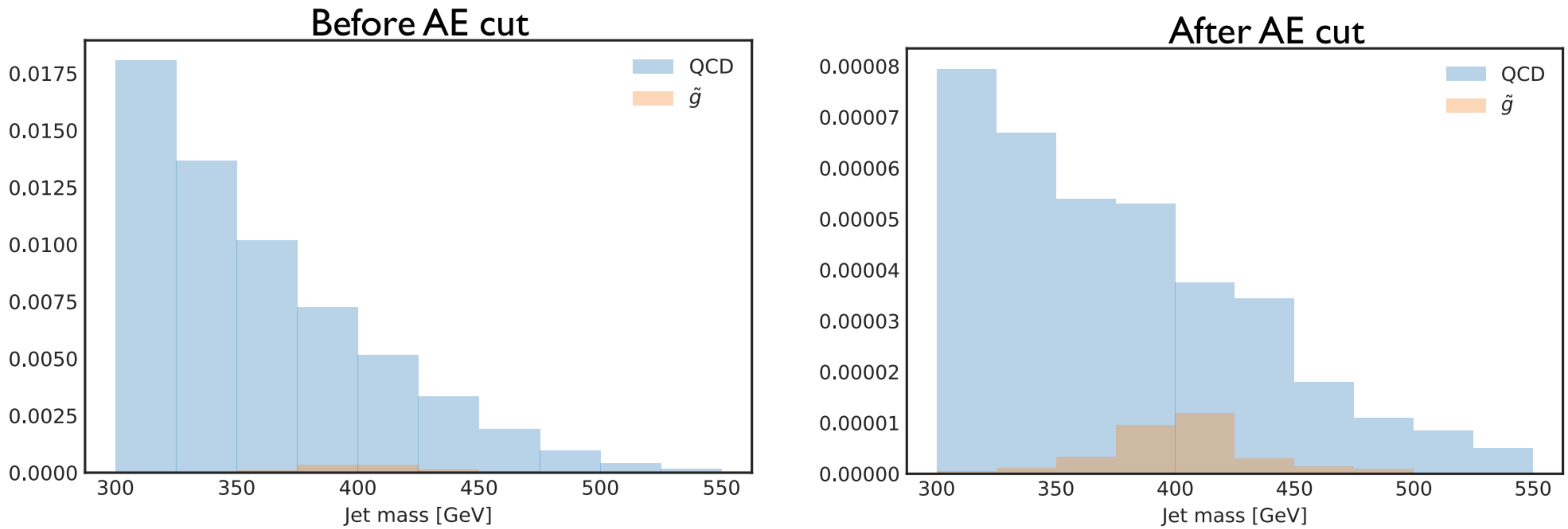
Crucial to avoid distortions of the spectrum

Enhanced bump hunt?



Mass distribution stable against reconstruction error cuts

Enhanced bump hunt?

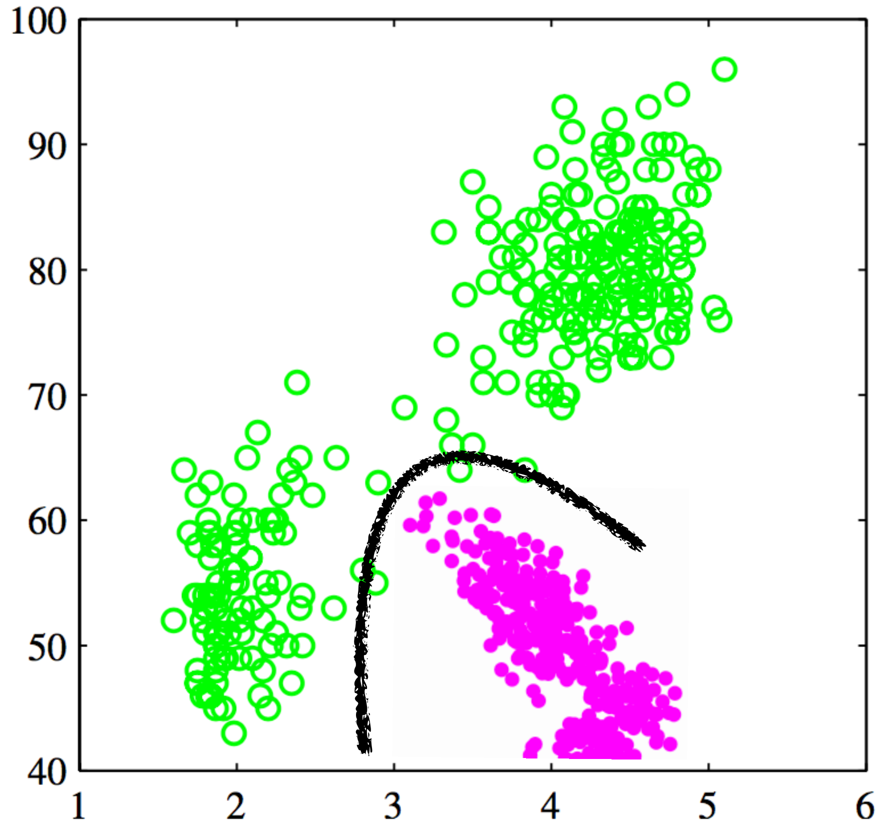


Use AE to deplete data from bkg events.
Followed by bump hunt? Other techniques?

Very different tasks

Training

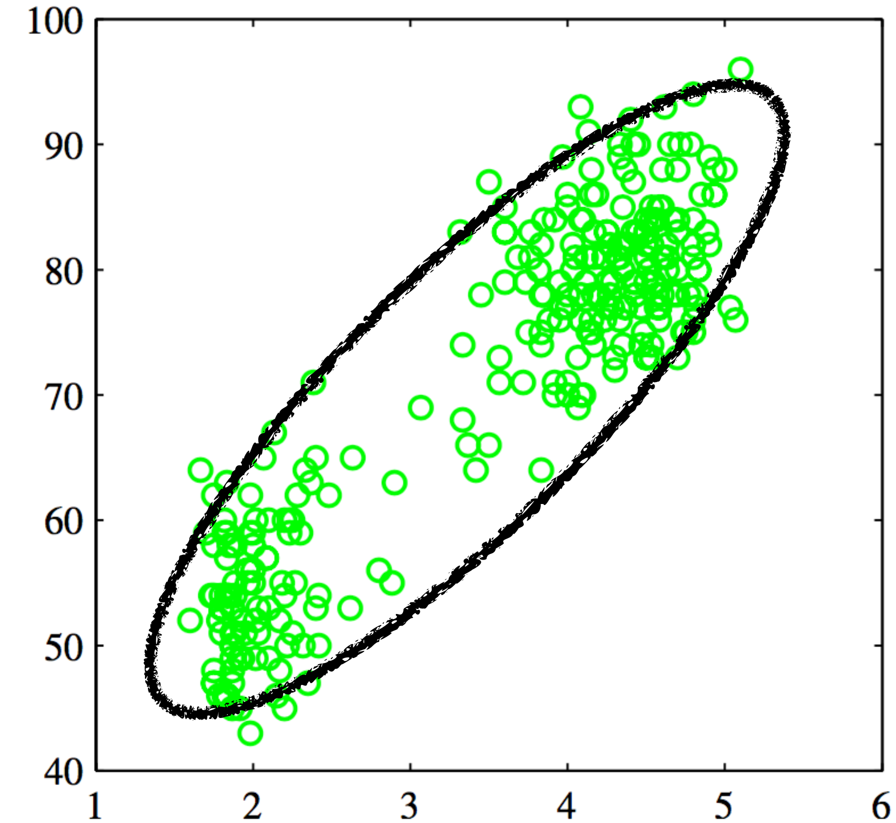
Supervised



Optimized decision boundary

vs

Unsupervised

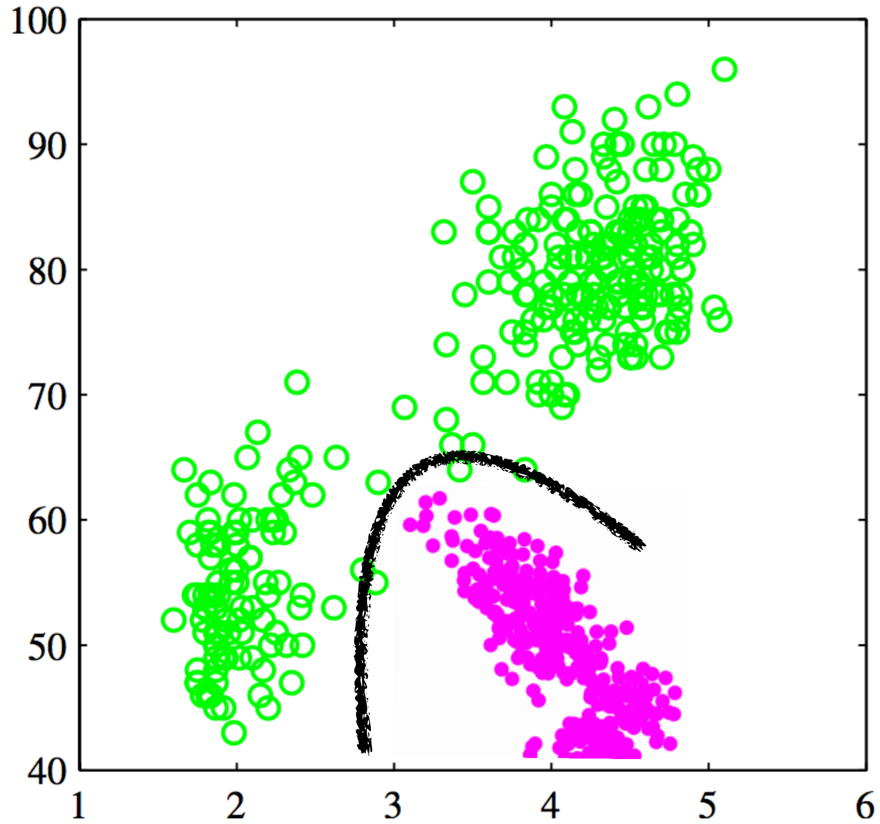


Optimized modeling of input data

Very different tasks

Inference

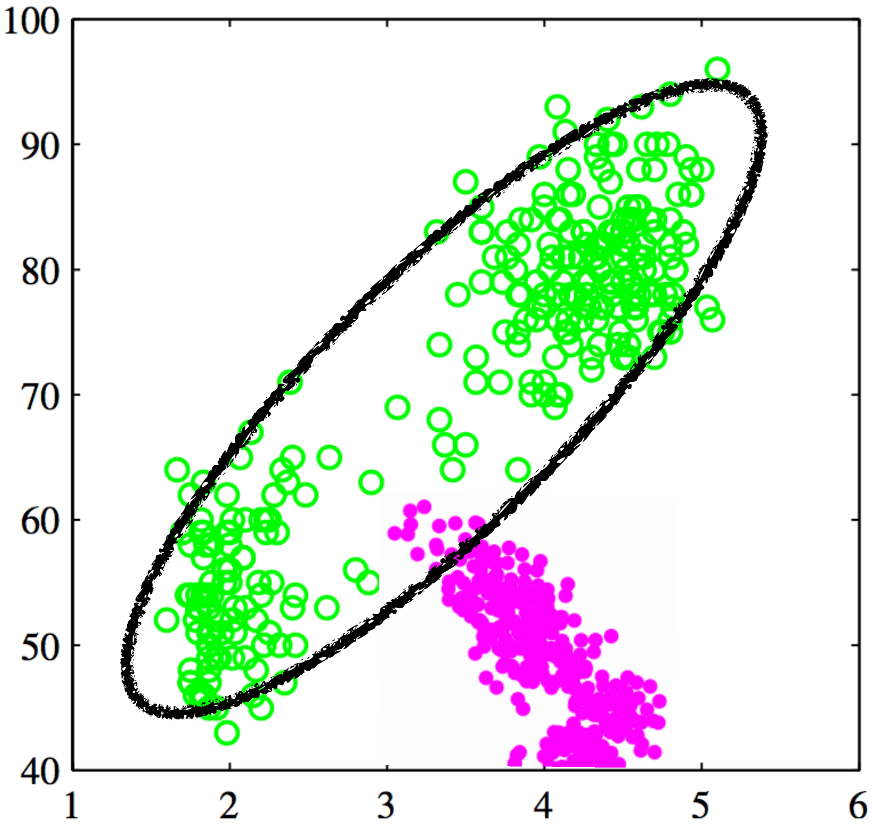
Supervised



Optimized decision boundary

vs

Unsupervised

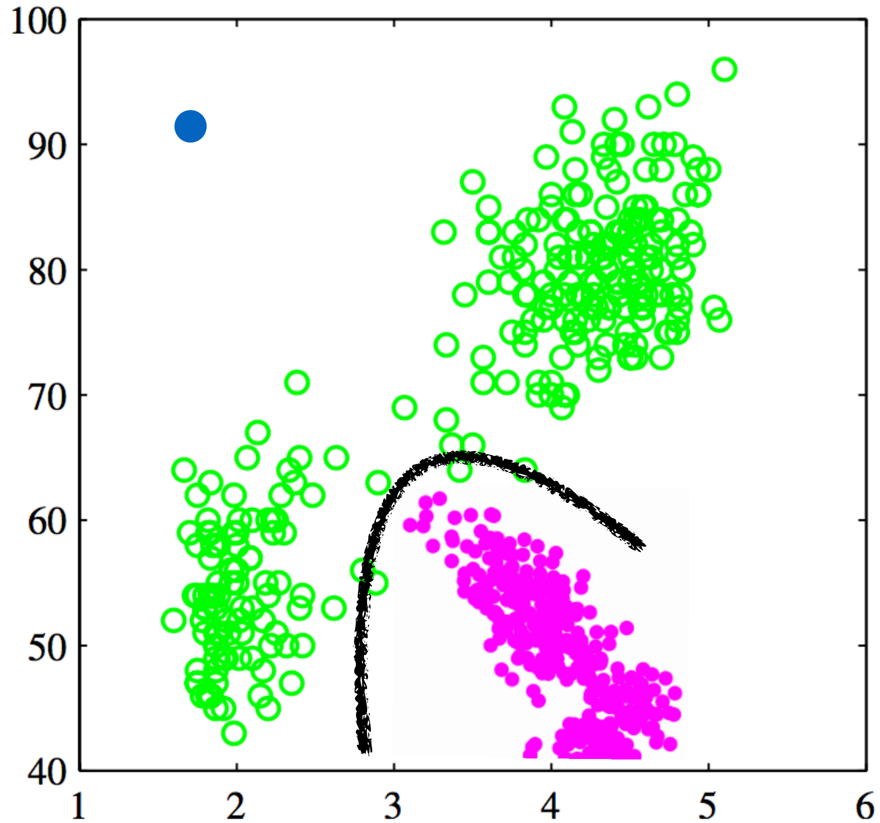


Optimized modeling of input data

Very different tasks

Inference

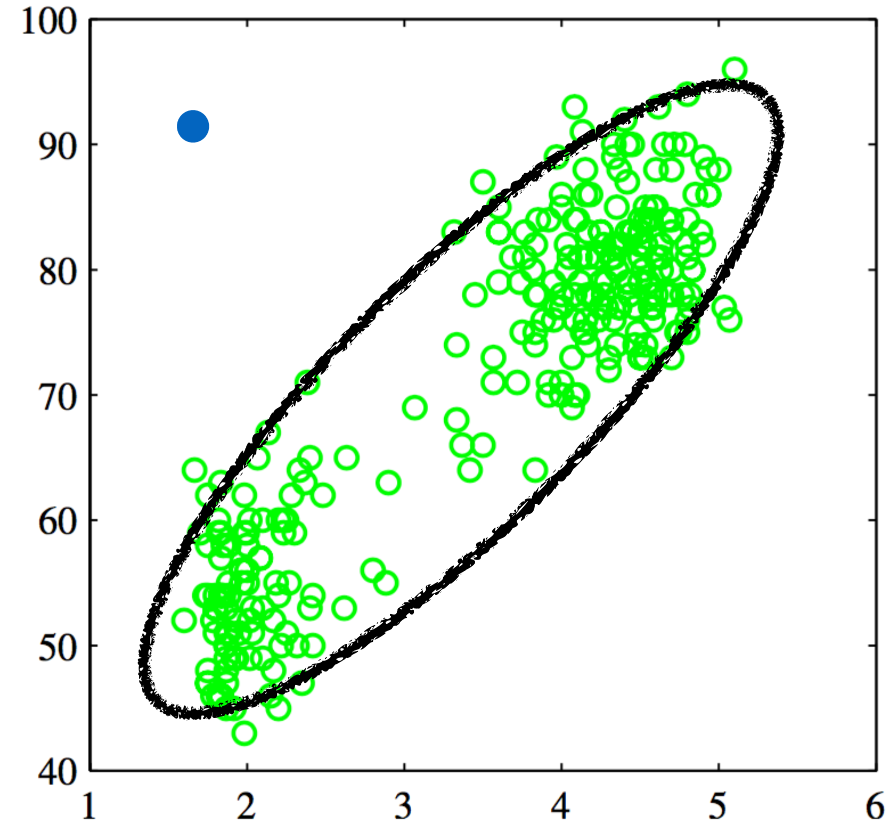
Supervised



Optimized decision boundary

vs

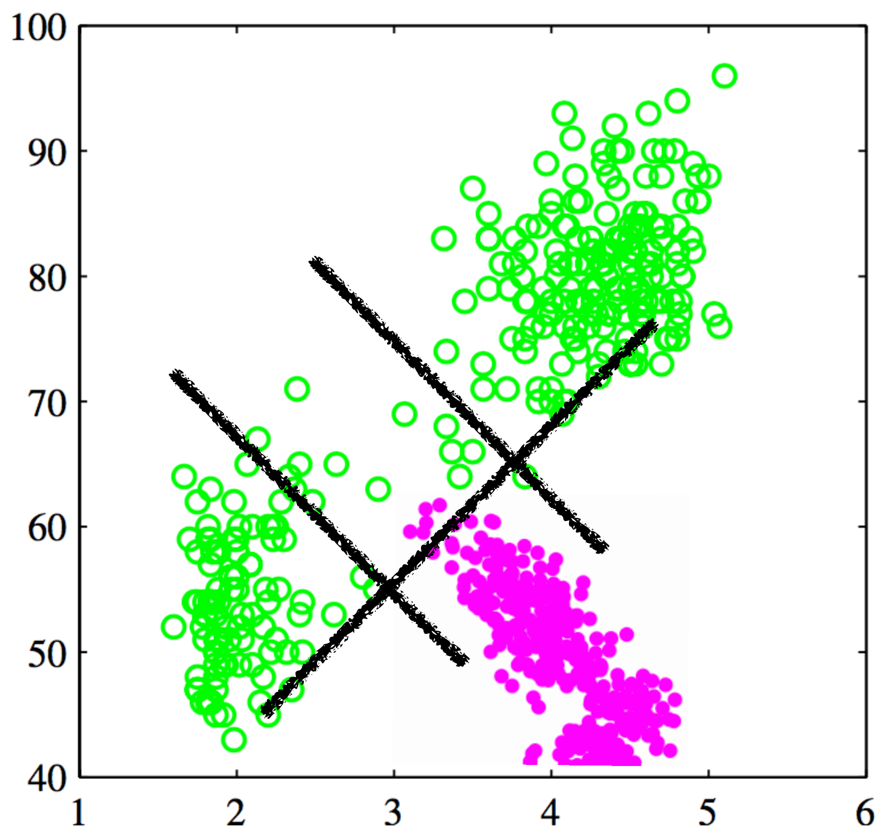
Unsupervised



Optimized modeling of input data

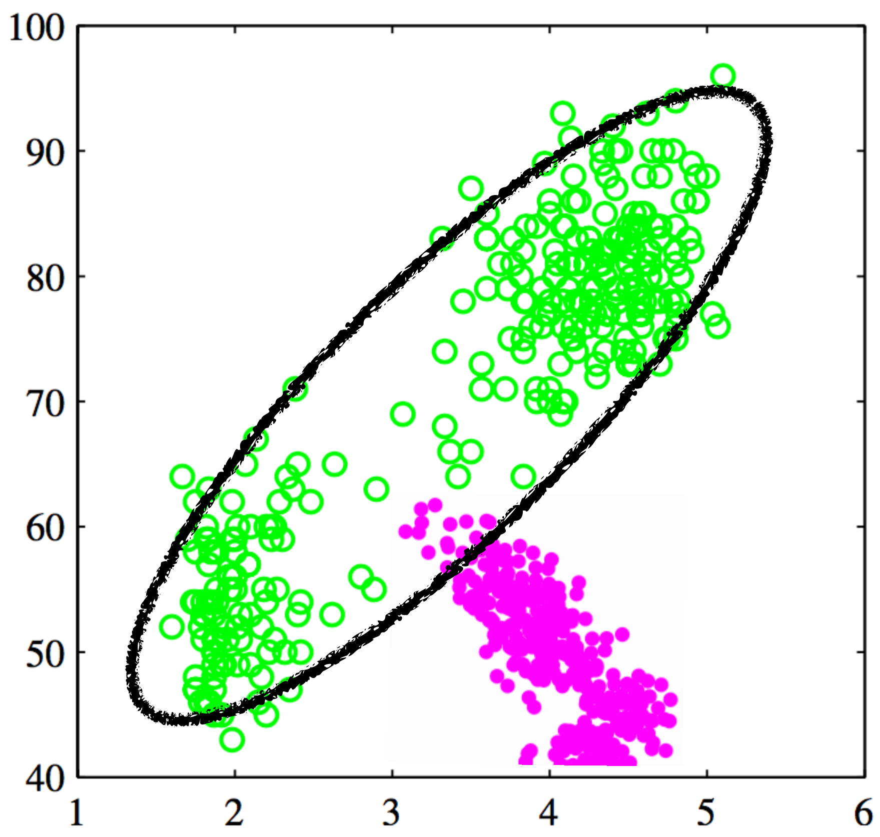
Very different tasks

CWoLa hunting



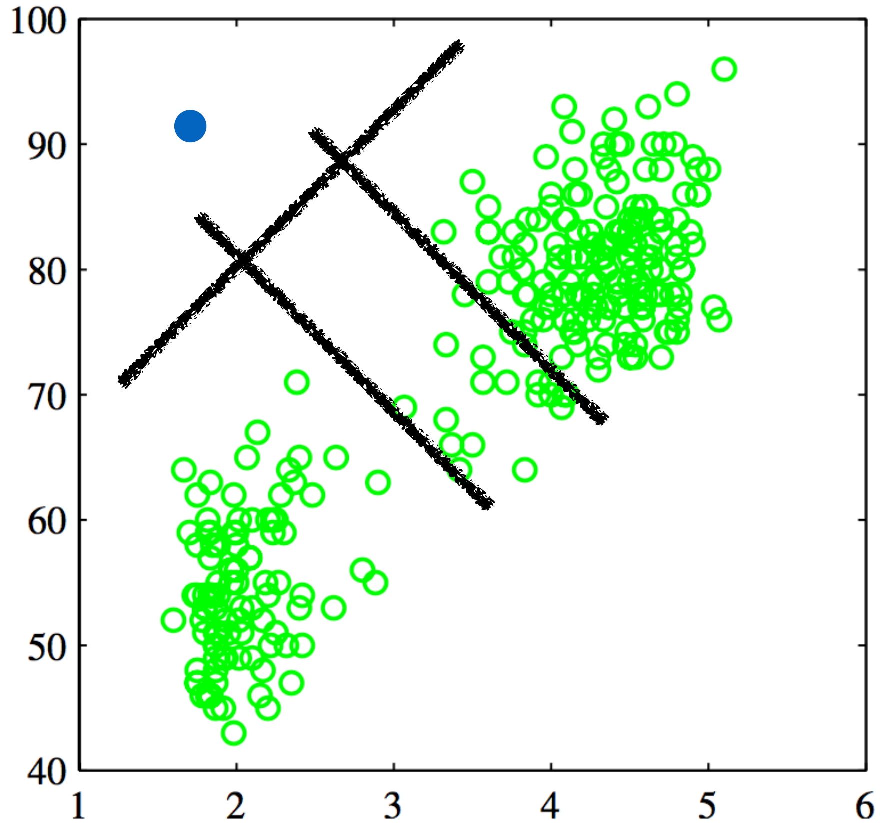
vs

Autoencoders



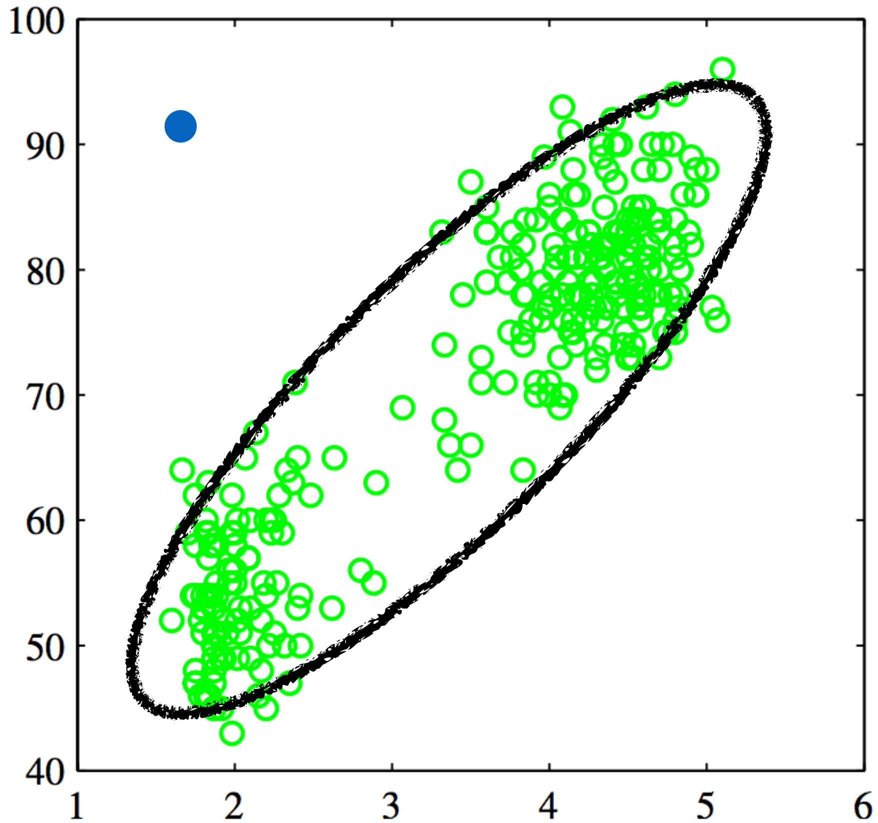
Very different tasks

CWoLa hunting



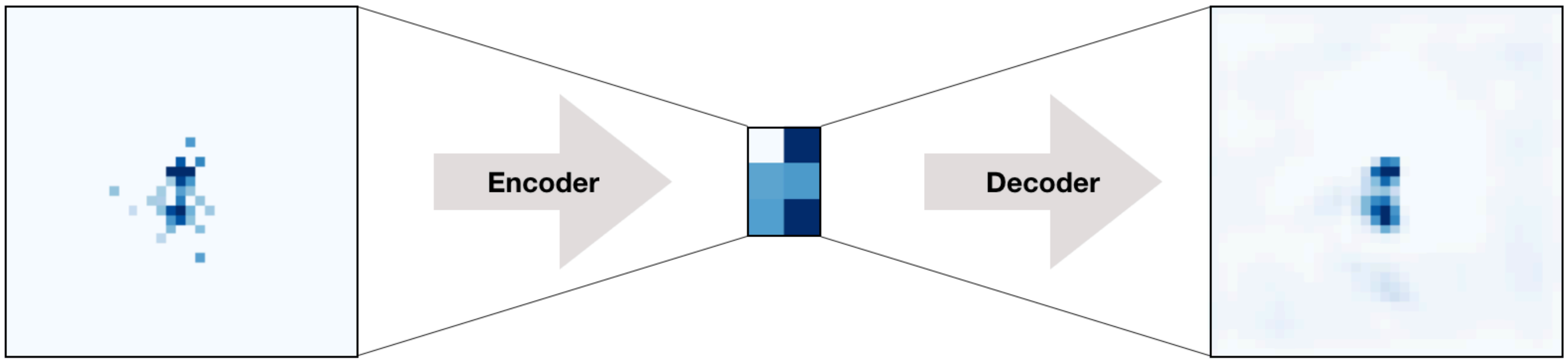
vs

Autoencoders



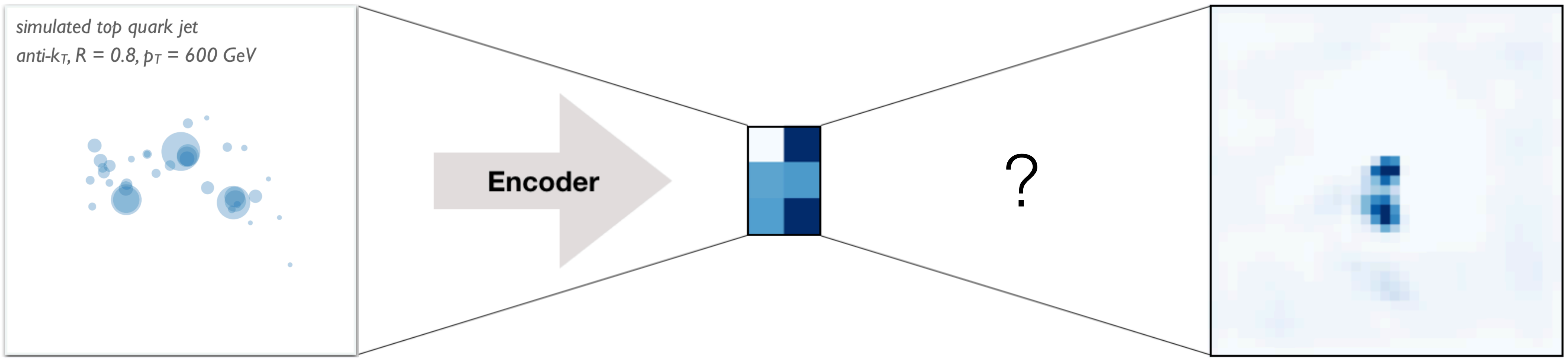
Which one is better? When? Can they be combined?

Other jet representations?



- 4-vectors matrix (see Gregor's talk)
- sequence (RNN/seq2seq)

Other jet representations?



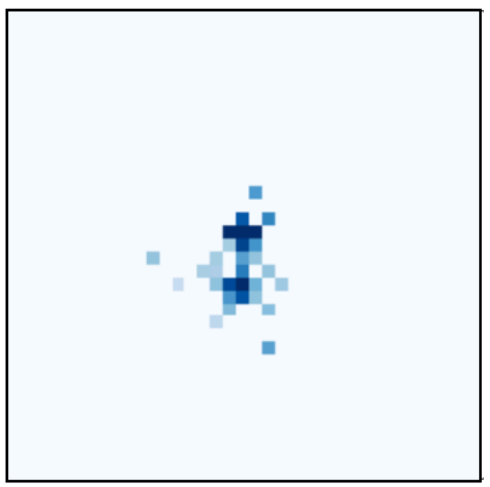
Huilin's talk

- 4-vectors matrix (see Gregor's talk)
- sequence (RNN/seq2seq)
- binary tree (?)
- point cloud
- …

Is it really useful/relevant?

Choice of loss?

What's a good "distance" between two jets?
E.g.

1) 

$$\ell = |p_T^{(\mathrm{pred})} - p_T^{(\mathrm{true})}|^2 \longrightarrow \ell = log\left(\frac{p_T^{(\mathrm{pred})} + \bar{p}}{p_T^{(\mathrm{true})} + \bar{p}}\right)^2$$

Kominske, Metodiev, Nachman, Schwartz '17

2)

$$\tilde{k}_j = \begin{pmatrix} \tilde{k}_{0,j} \\ \tilde{k}_{1,j} \\ \tilde{k}_{2,j} \\ \tilde{k}_{3,j} \end{pmatrix} \xrightarrow{\mathrm{LoLa}} \begin{pmatrix} \tilde{k}_{0,j} \\ \tilde{k}_{1,j} \\ \tilde{k}_{2,j} \\ \tilde{k}_{3,j} \\ \sqrt{\tilde{k}_j^2} \end{pmatrix}$$

$$L_{\mathrm{auto}} = \sum_{j=1}^{40} \sum_{i=0}^{3} \left(\tilde{k}_{i,j}^{\mathrm{in}} - \tilde{k}_{i,j}^{\mathrm{auto}}\right)^2$$

Jet (binary) classification

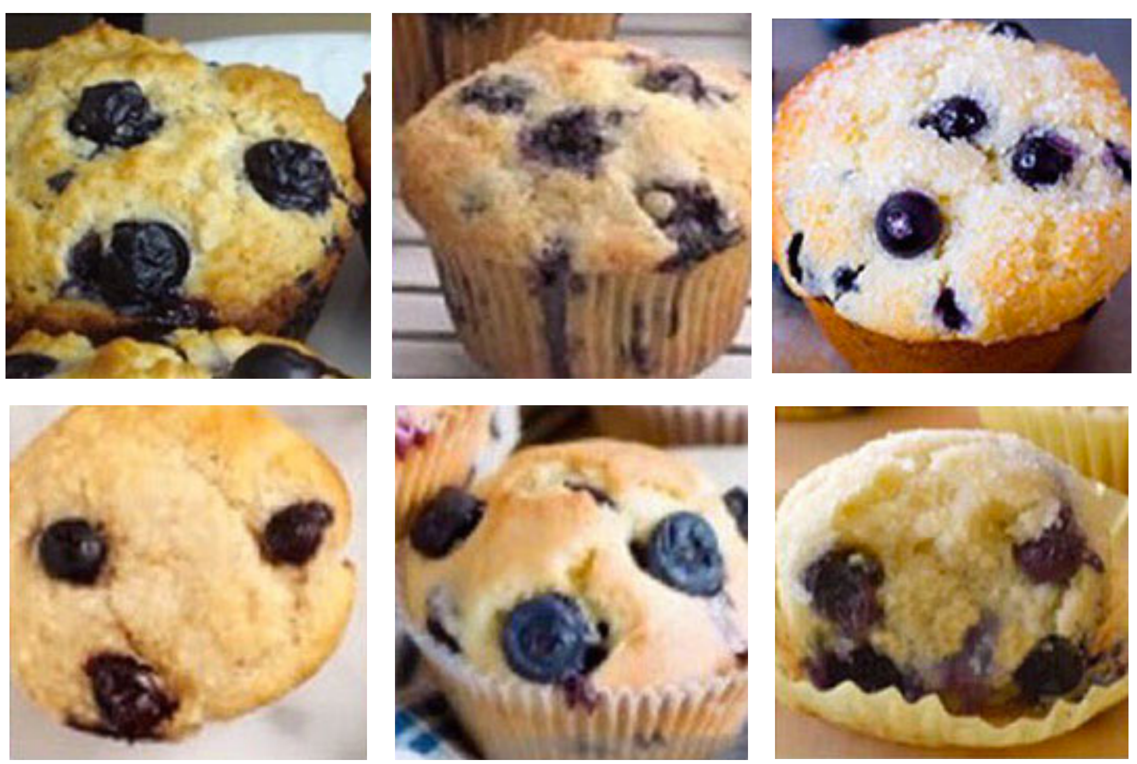
Muffin or chihuahua?

$$p(\mathrm{dog}|x) = f(x)$$

Jet anomaly detection

Given data from $p(x)$. Can I tell if $\hat{x}$ comes from the same distribution?

Train set

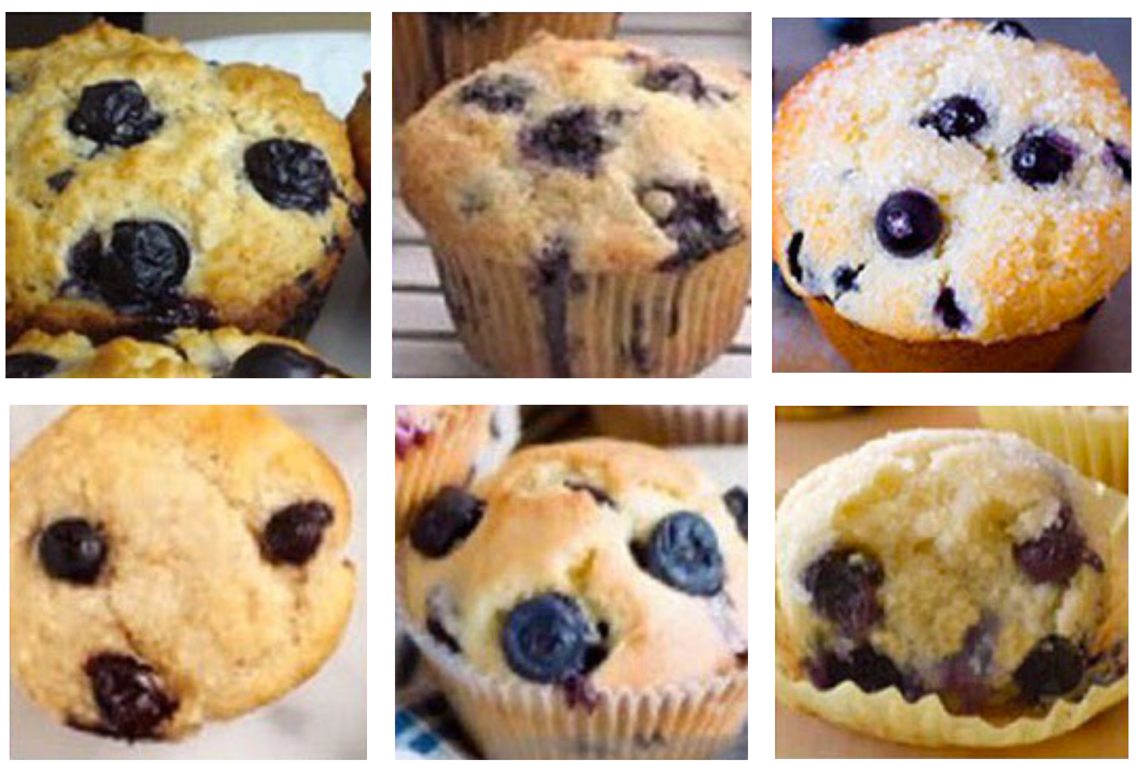


$$p(\mathrm{muffin}|\hat{x}) > \epsilon$$

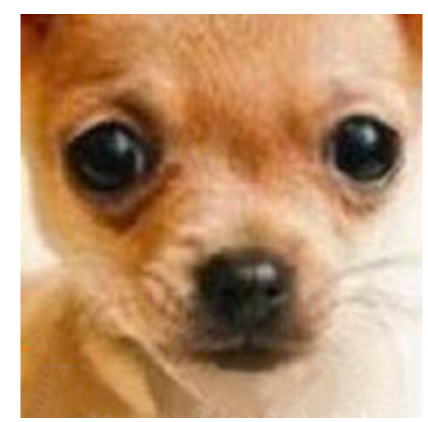
Jet anomaly detection

Given data from $p(x)$. Can I tell if $\hat{x}$ comes from the same distribution?

Train set



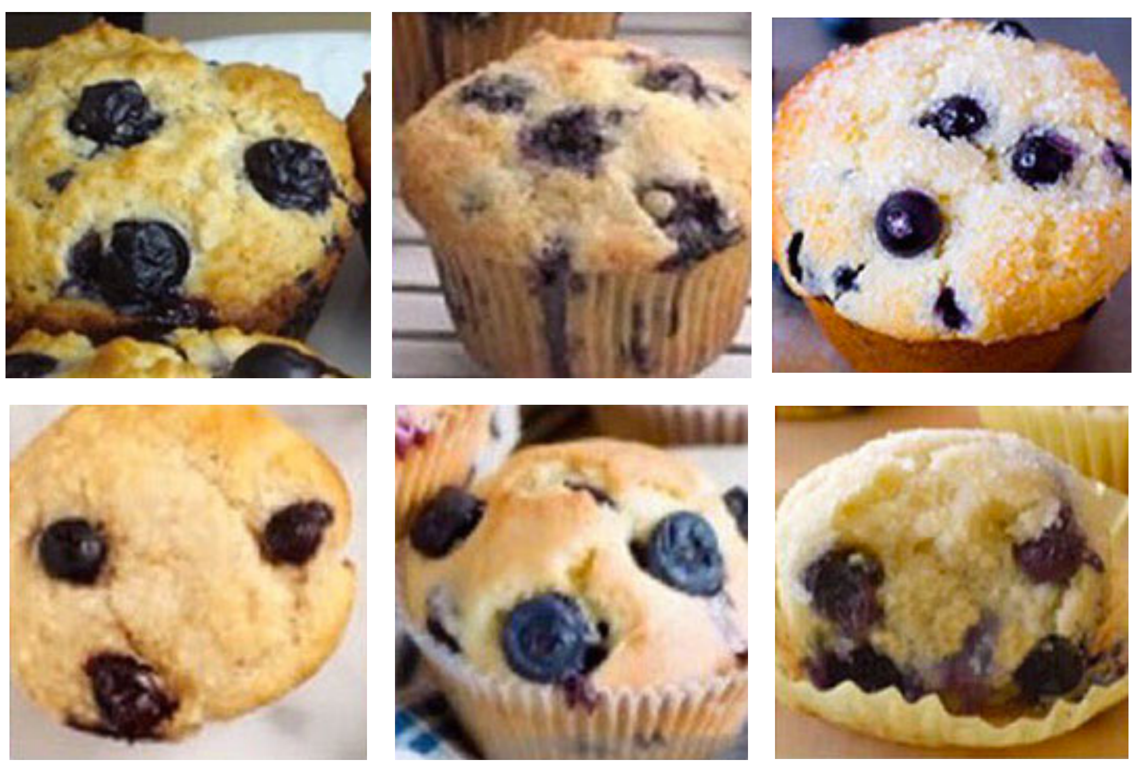
$$p(\mathrm{muffin}|\hat{x}) > \epsilon$$



Anomaly detection

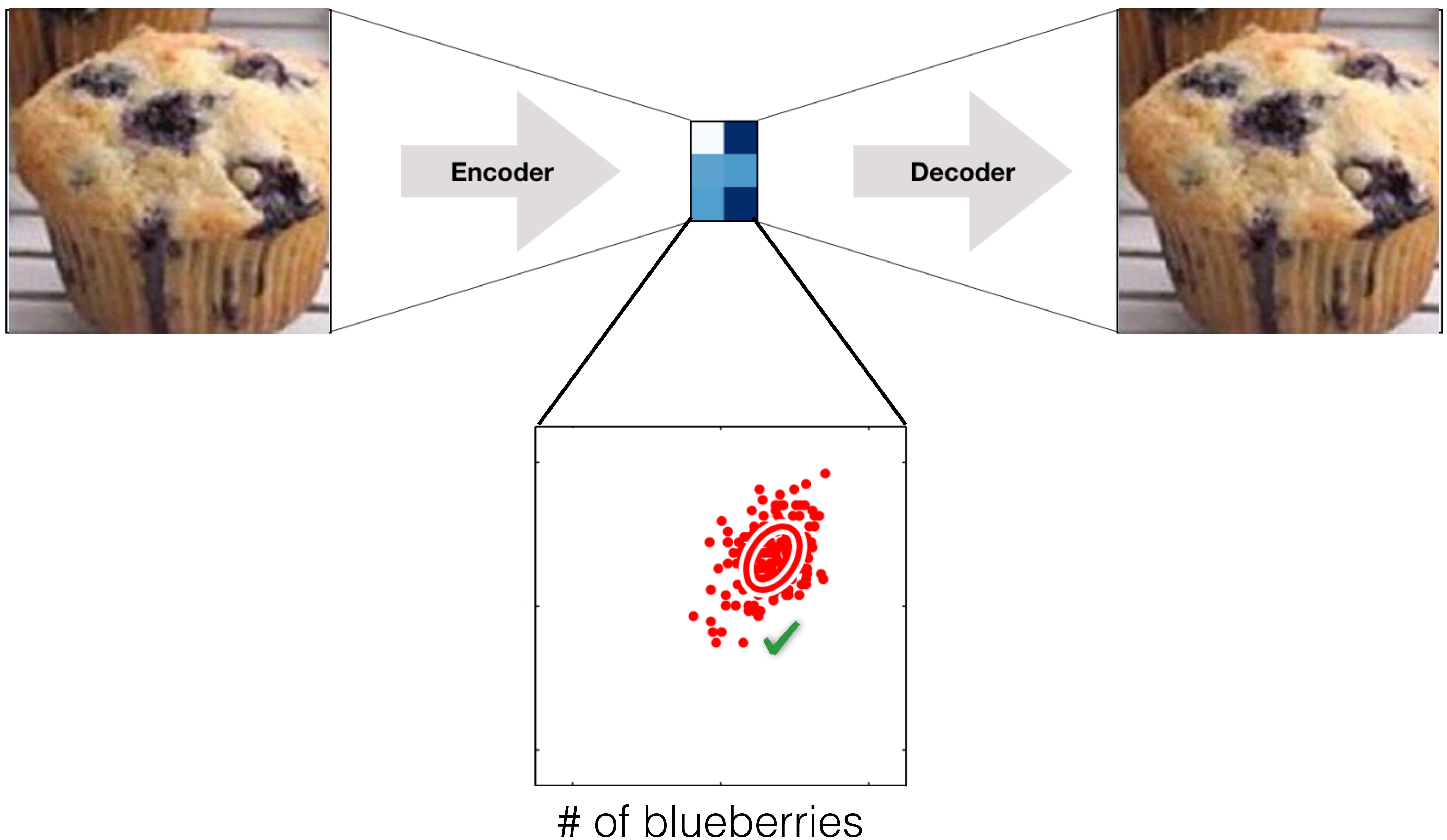
Given data from $p(x)$. Can I tell if $\hat{x}$ comes from the same distribution?

Train set

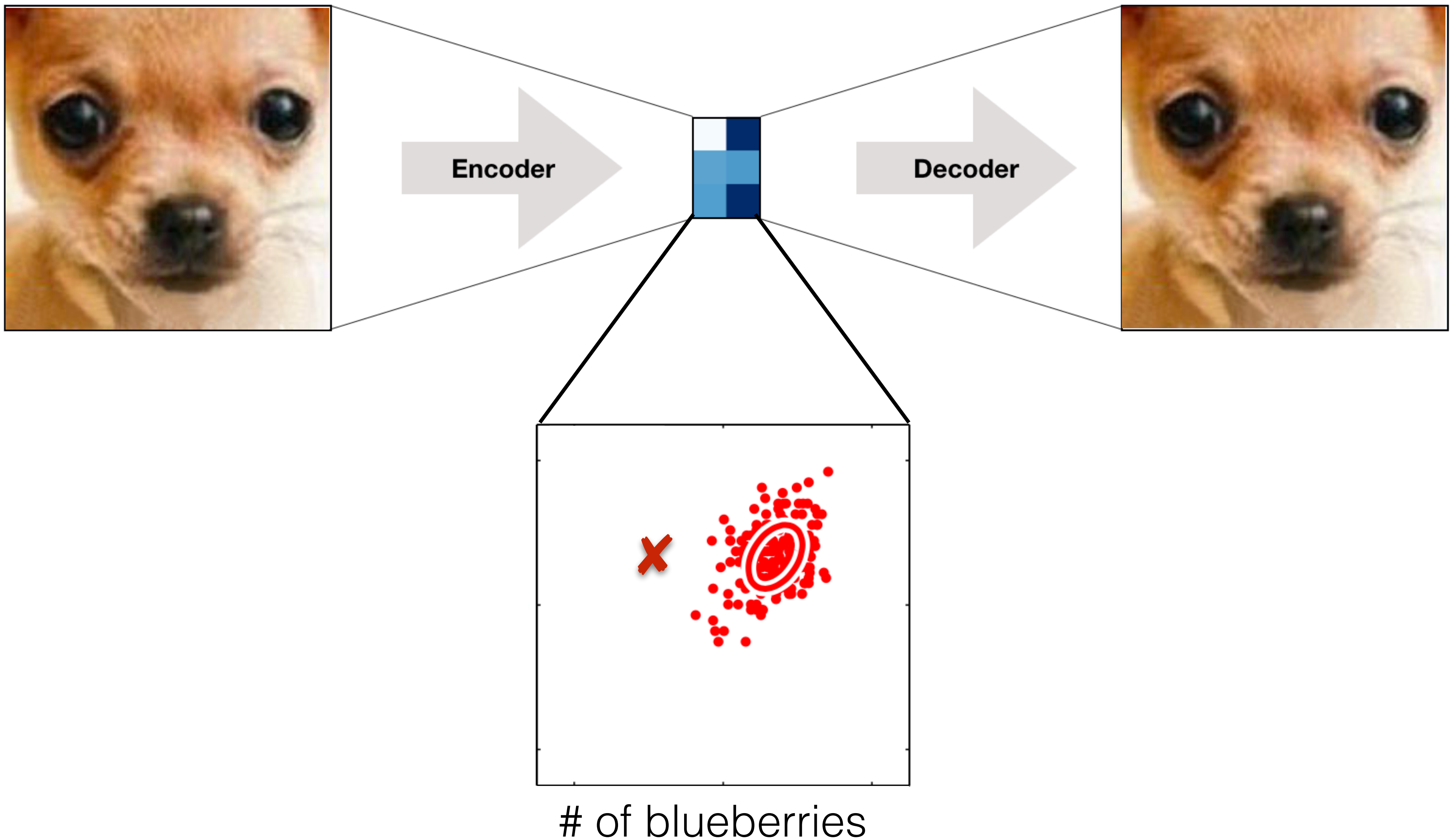


$$p(\text{muffin}|\hat{x}) > \epsilon$$

Future directions?

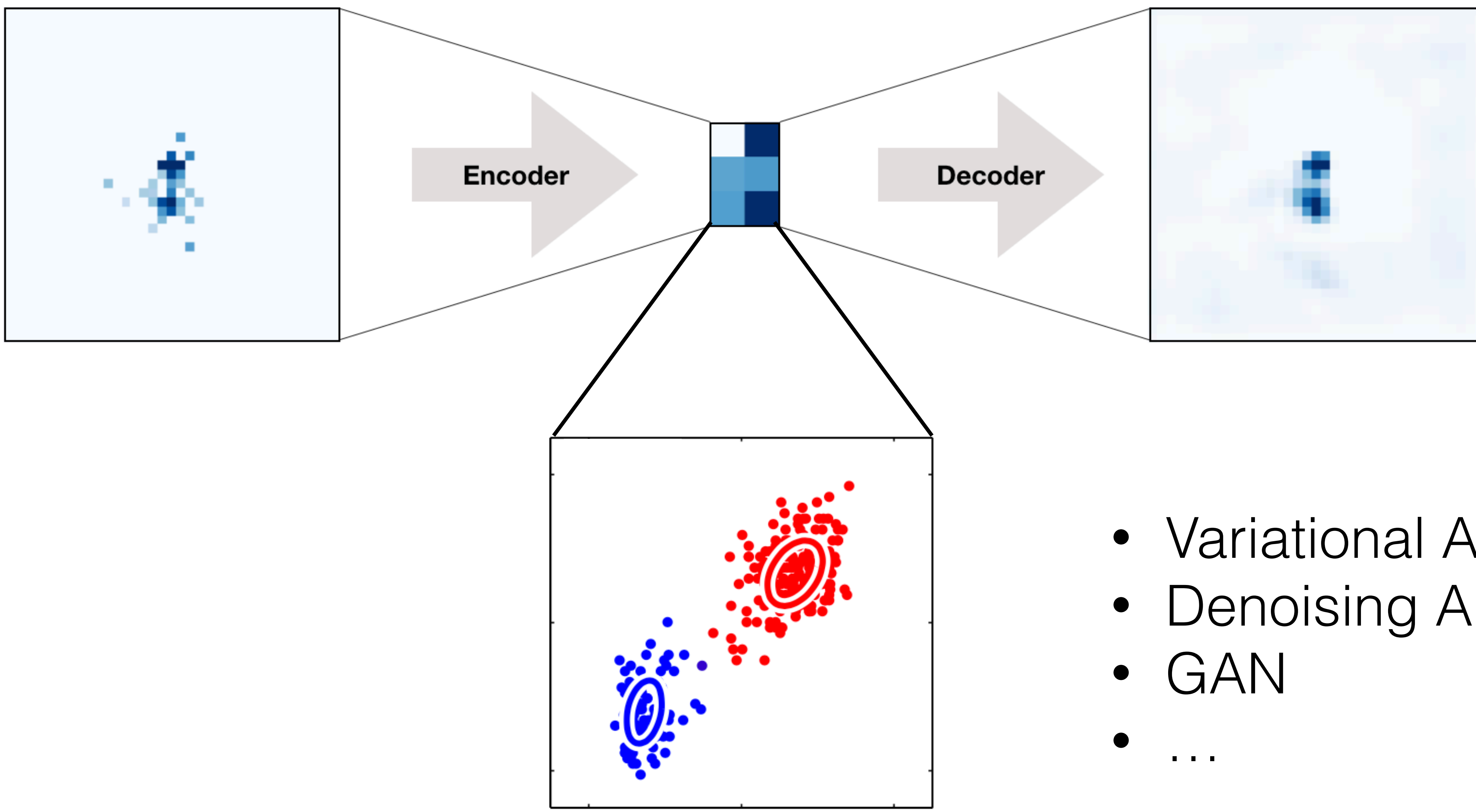


Future directions?



Future directions?

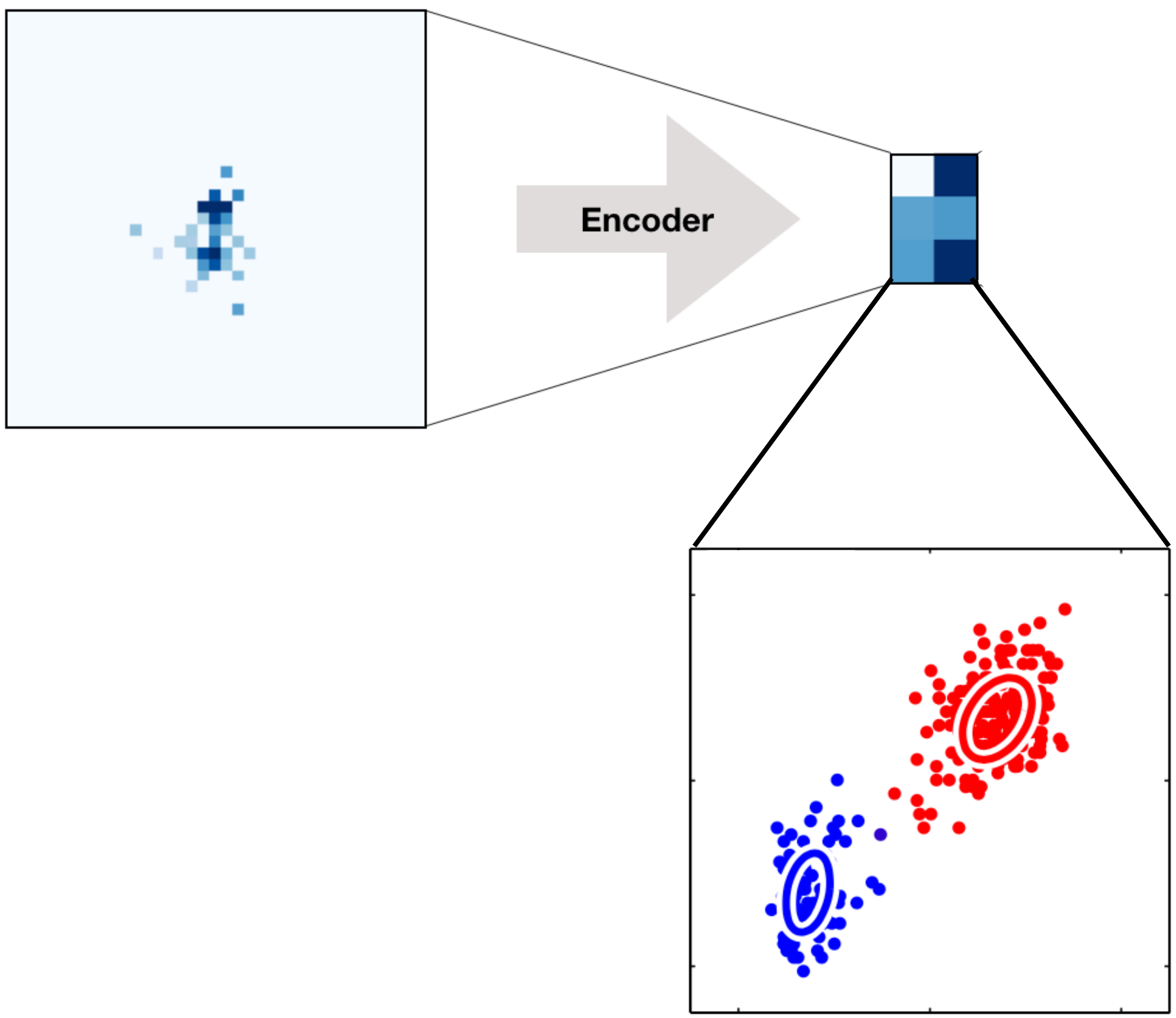
For now we have abandoned all probabilistic interpretation and discarded latent space information



- Variational AE
- Denoising AE
- GAN
- …

Future directions?

For now we have abandoned all probabilistic interpretation and discarded latent space information



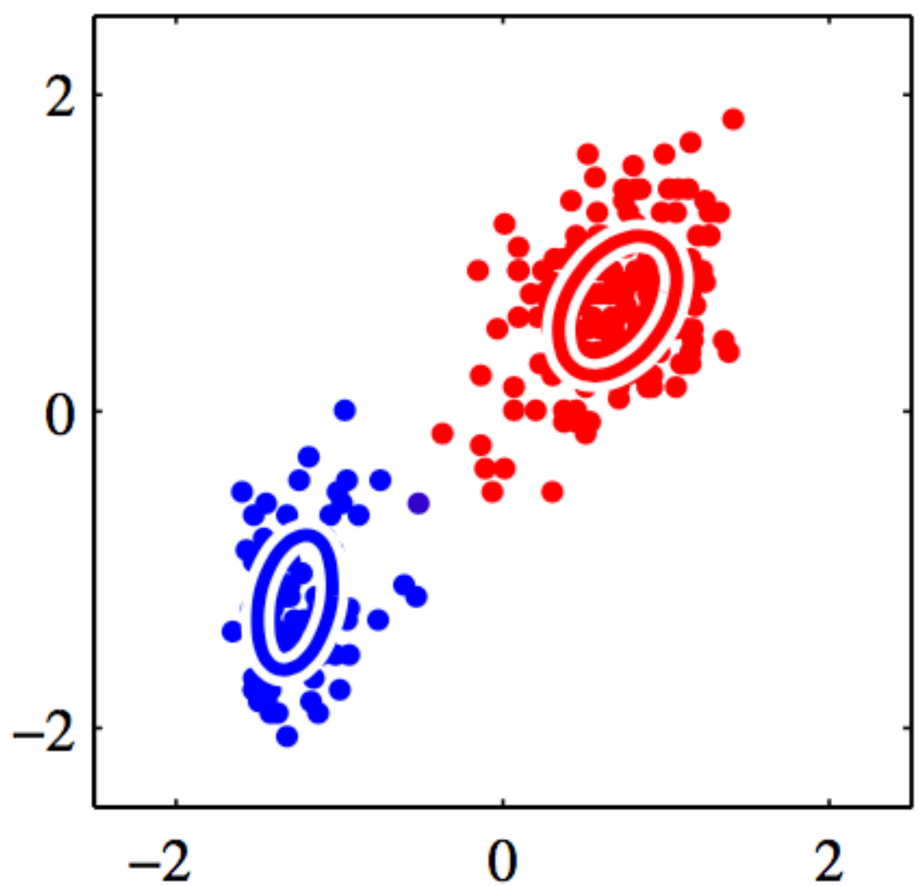
Conclusion

Autoencoders can be powerful tools for anomaly searches in jet physics

First baby steps in unsupervised territory

One example

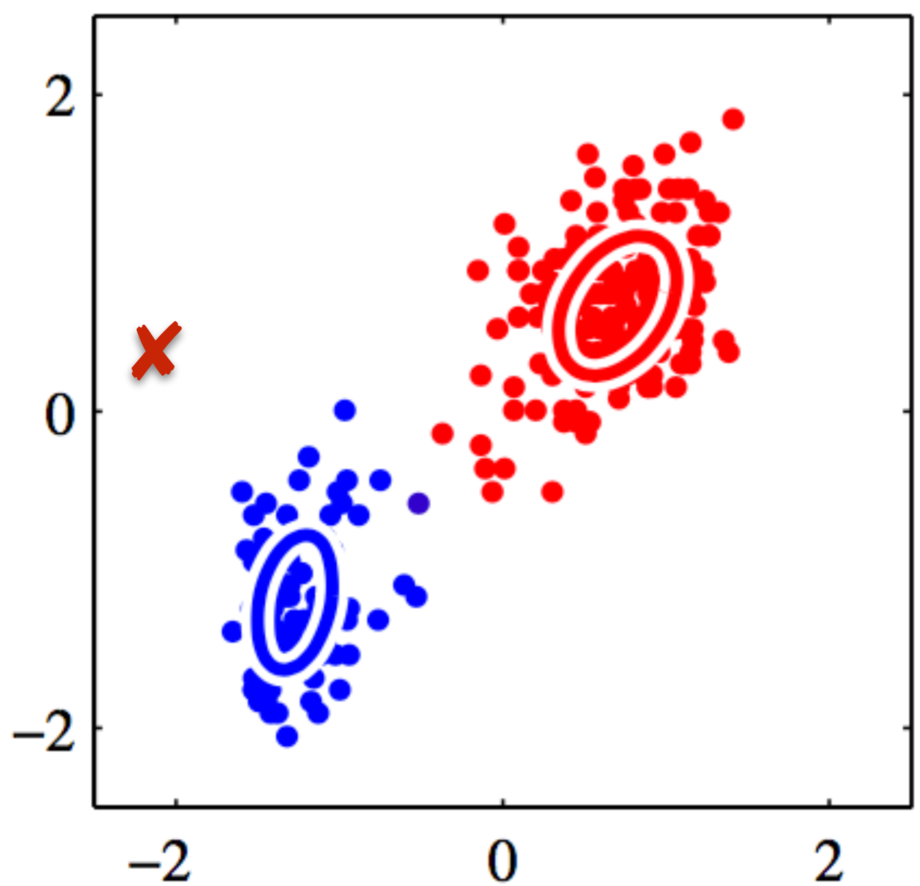
Model distribution in some way, e.g. as a mixture (of gaussians). Maximum likelihood parameter estimation.



$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

One example

Model distribution in some way, e.g. as a mixture (of gaussians). Maximum likelihood parameter estimation.

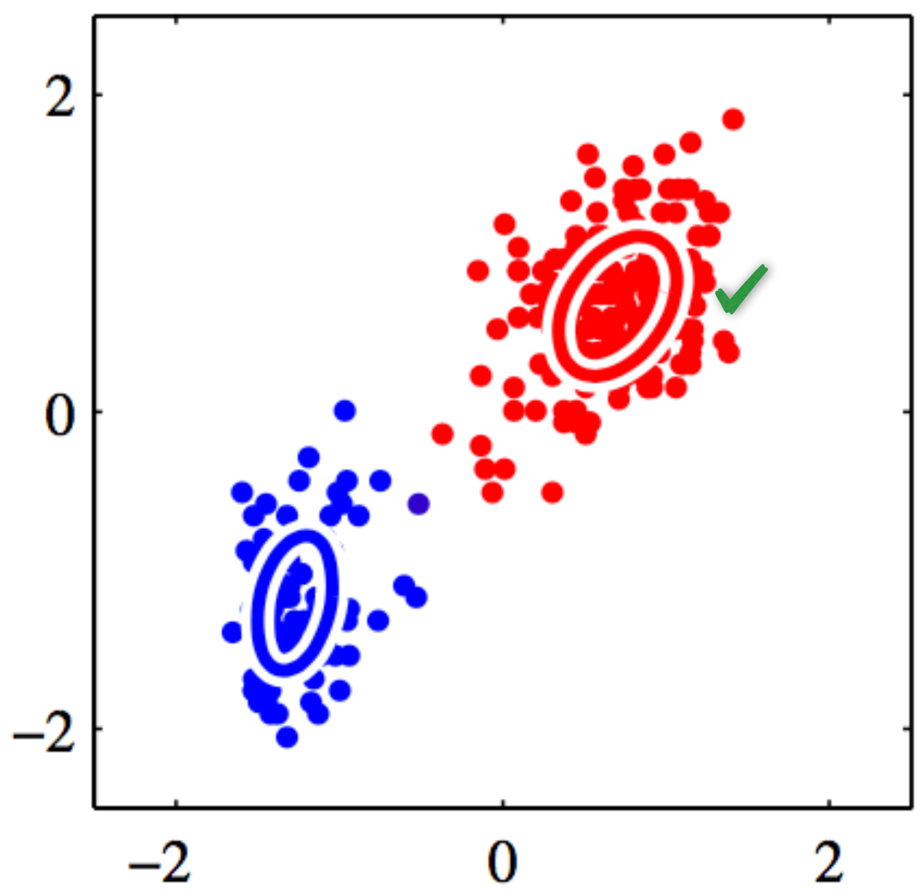


$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

$$p(\hat{x}) > \epsilon$$

One example

Model distribution in some way, e.g. as a mixture (of gaussians). Maximum likelihood parameter estimation.

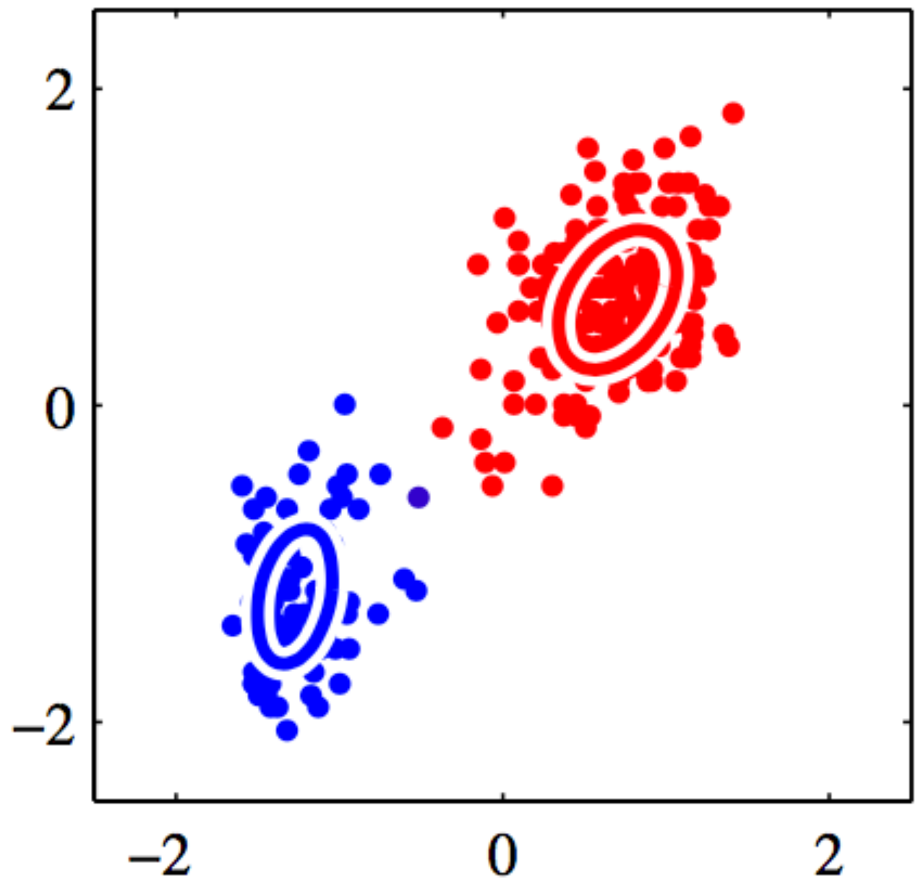


$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

$$p(\hat{x}) > \epsilon$$

One example

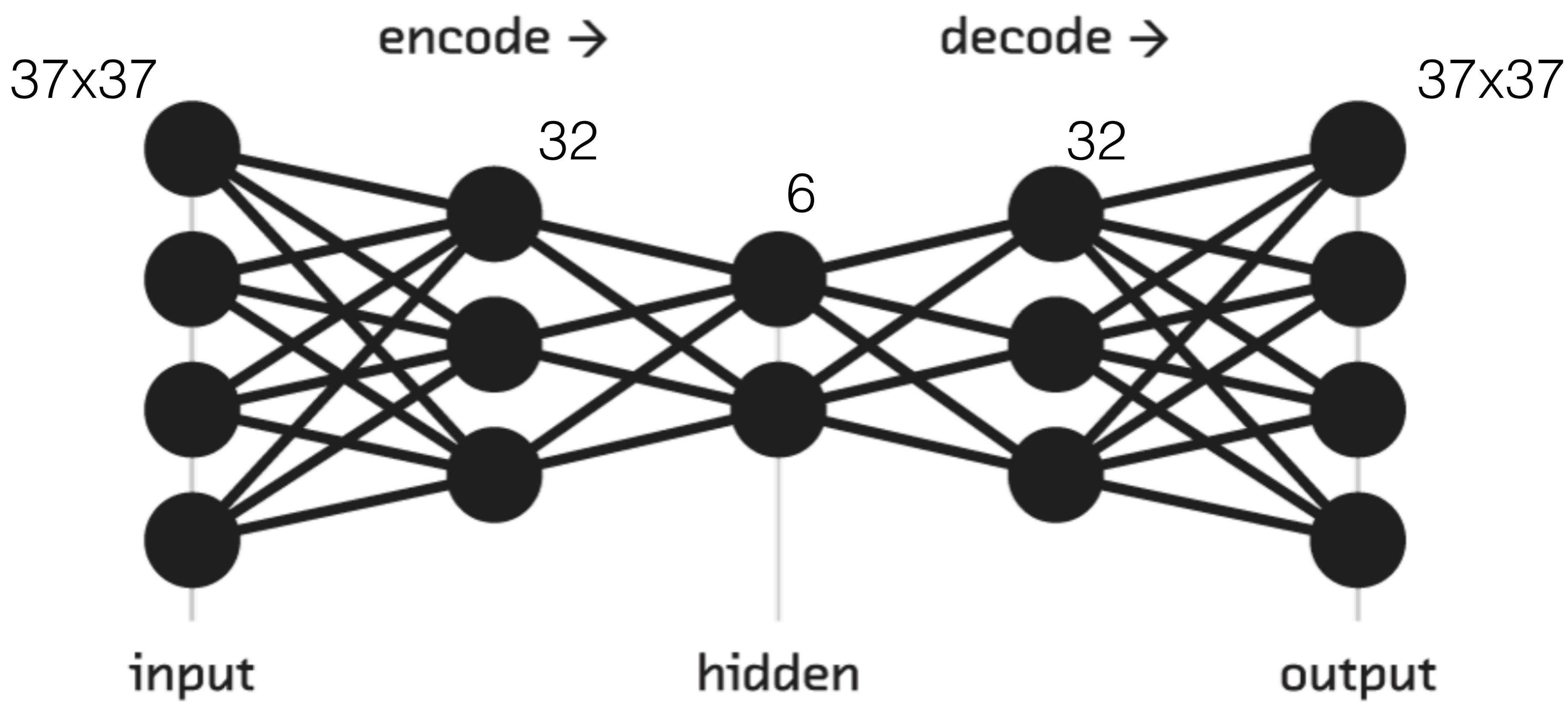
Model distribution in some way, e.g. as a mixture (of gaussians). Maximum likelihood parameter estimation.



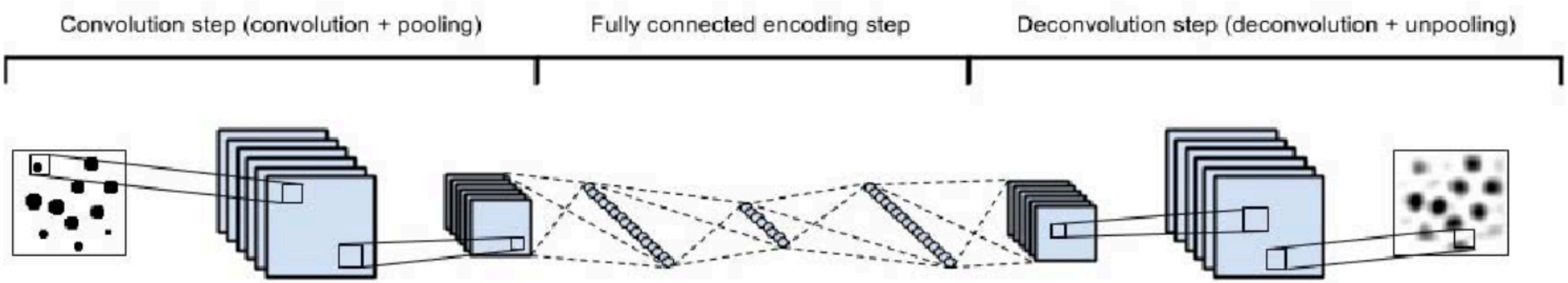
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Not feasible for large dimensionality data

DNN



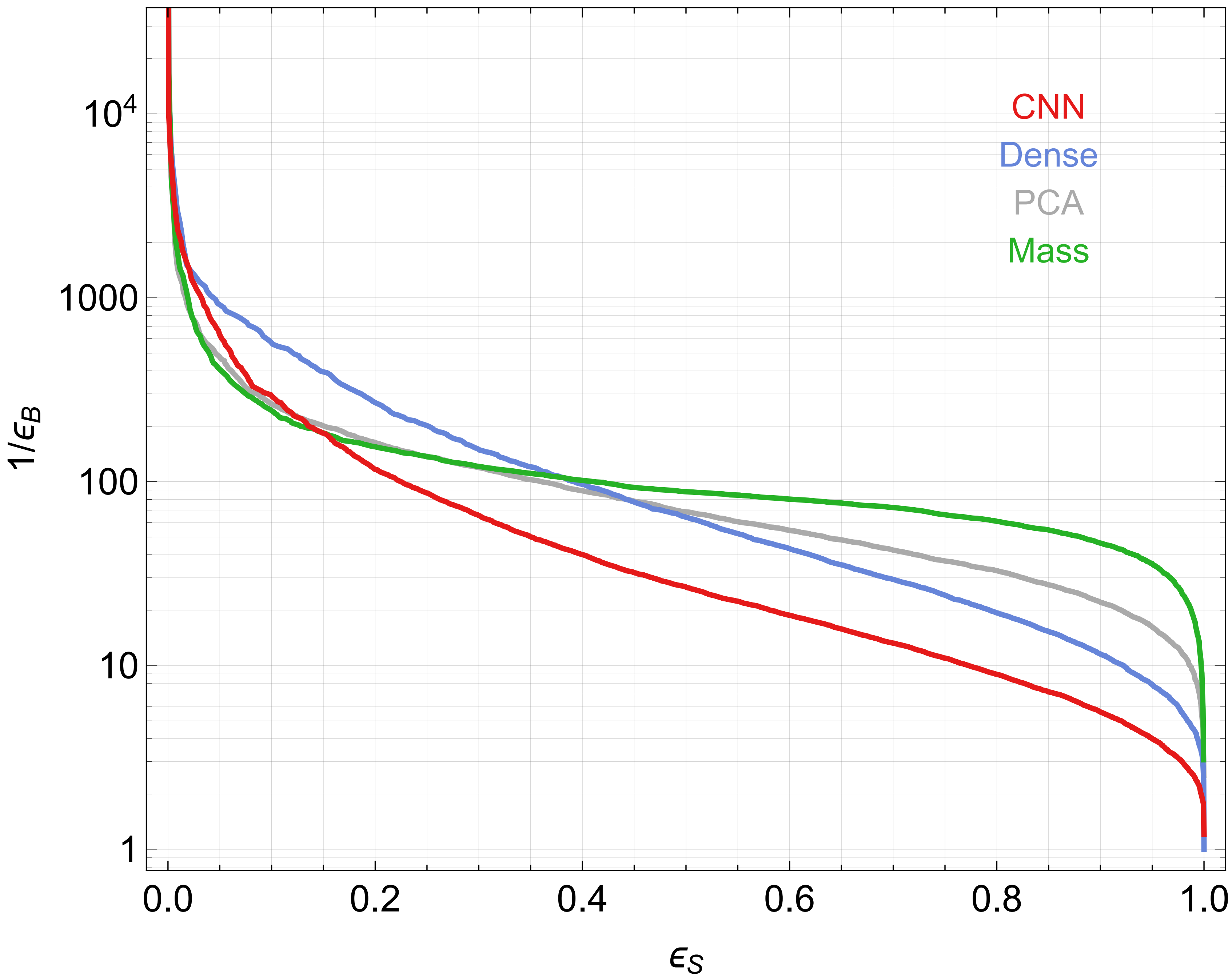
CNN



Ke, Lin, Huang '17

```
layer=Conv2D(128, kernel_size=(3, 3),
             activation='relu',padding='same')(layer)
layer=MaxPooling2D(pool_size=(2, 2),padding='same')(layer)
layer=Conv2D(128, kernel_size=(3, 3),
             activation='relu',padding='same')(layer)
layer=MaxPooling2D(pool_size=(2, 2),padding='same')(layer)
layer=Conv2D(128, kernel_size=(3, 3),
             activation='relu',padding='same')(layer)
layer=Flatten()(layer)
layer=Dense(32, activation='relu')(layer)
layer=Dense(6)(layer)
encoded=layer
```

Gluino ROC



CNN
Dense
PCA
Mass
10^4
1000
100
10
1
$1/\epsilon_B$
0.0
0.2
0.4
0.6
0.8
1.0
ϵ_S