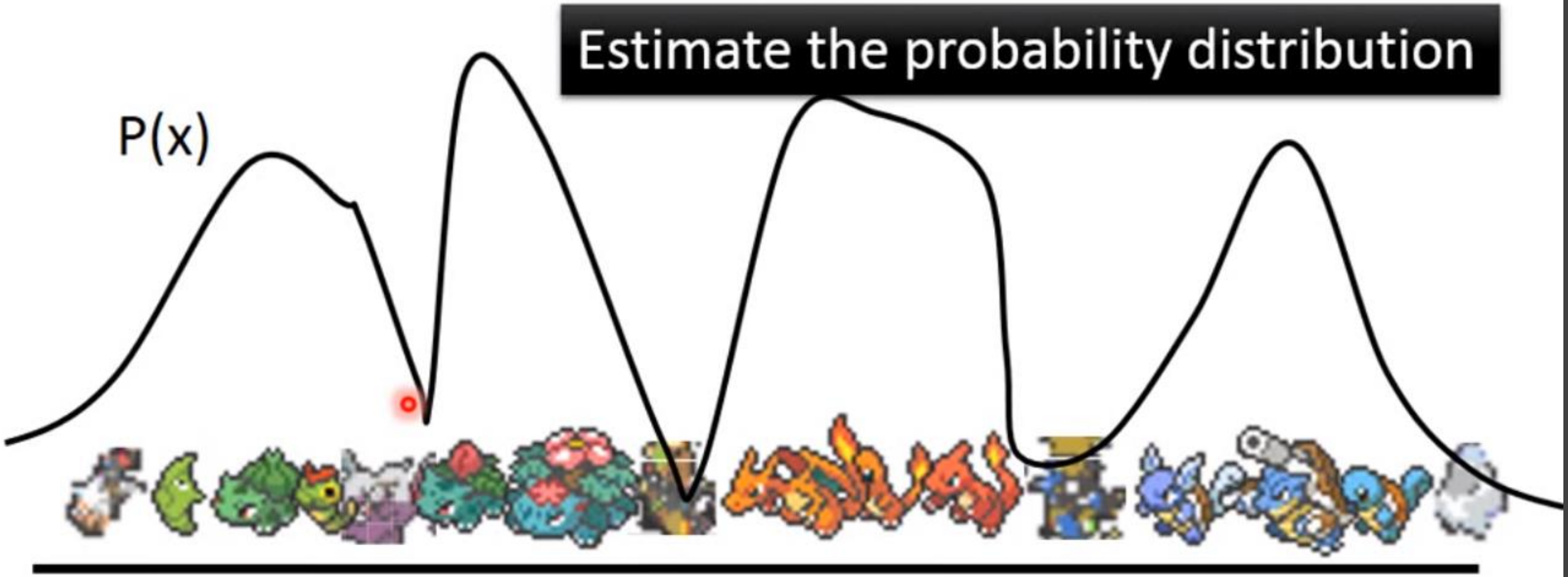
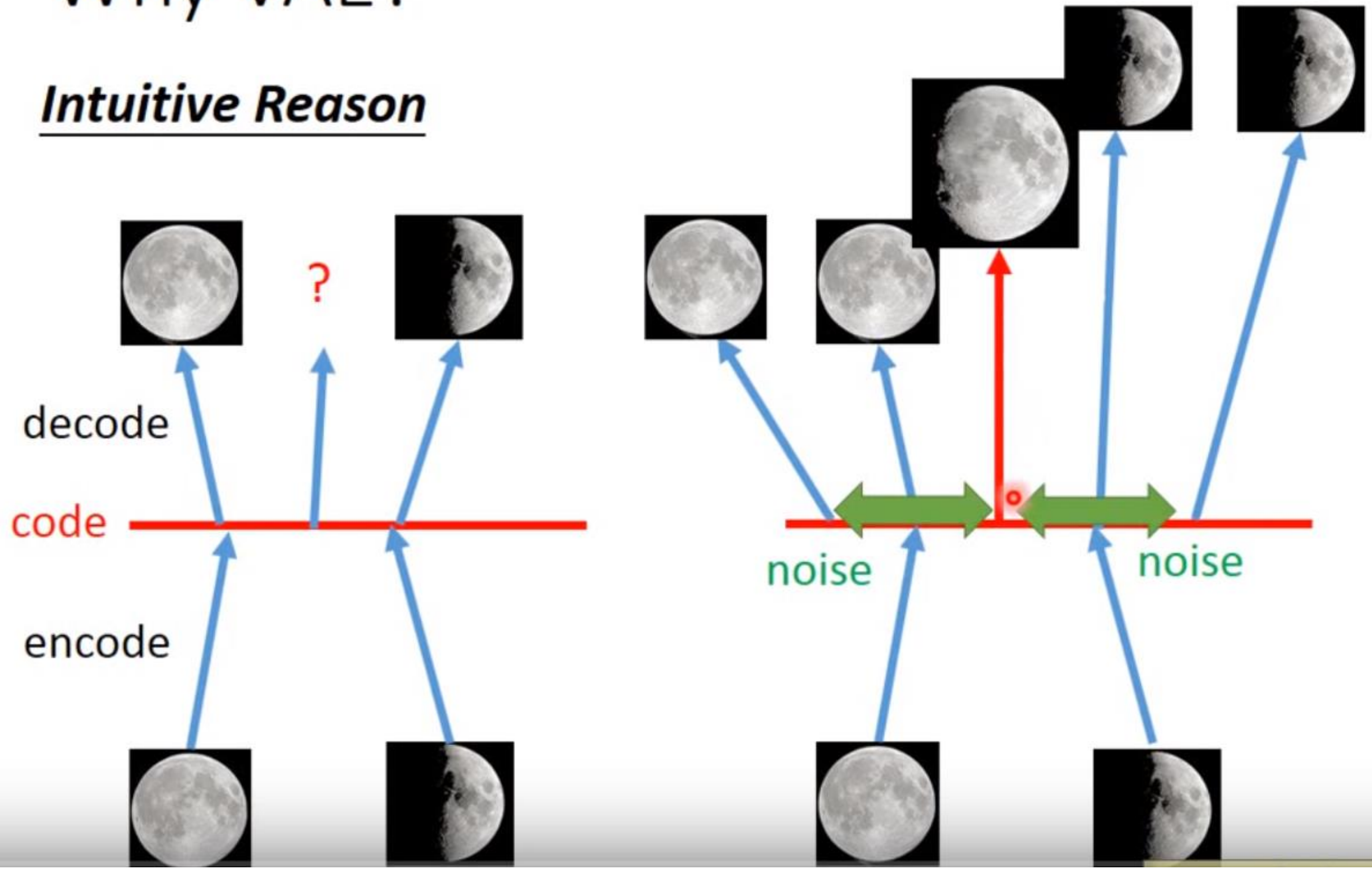
Why VAE?

- Back to what we want to do



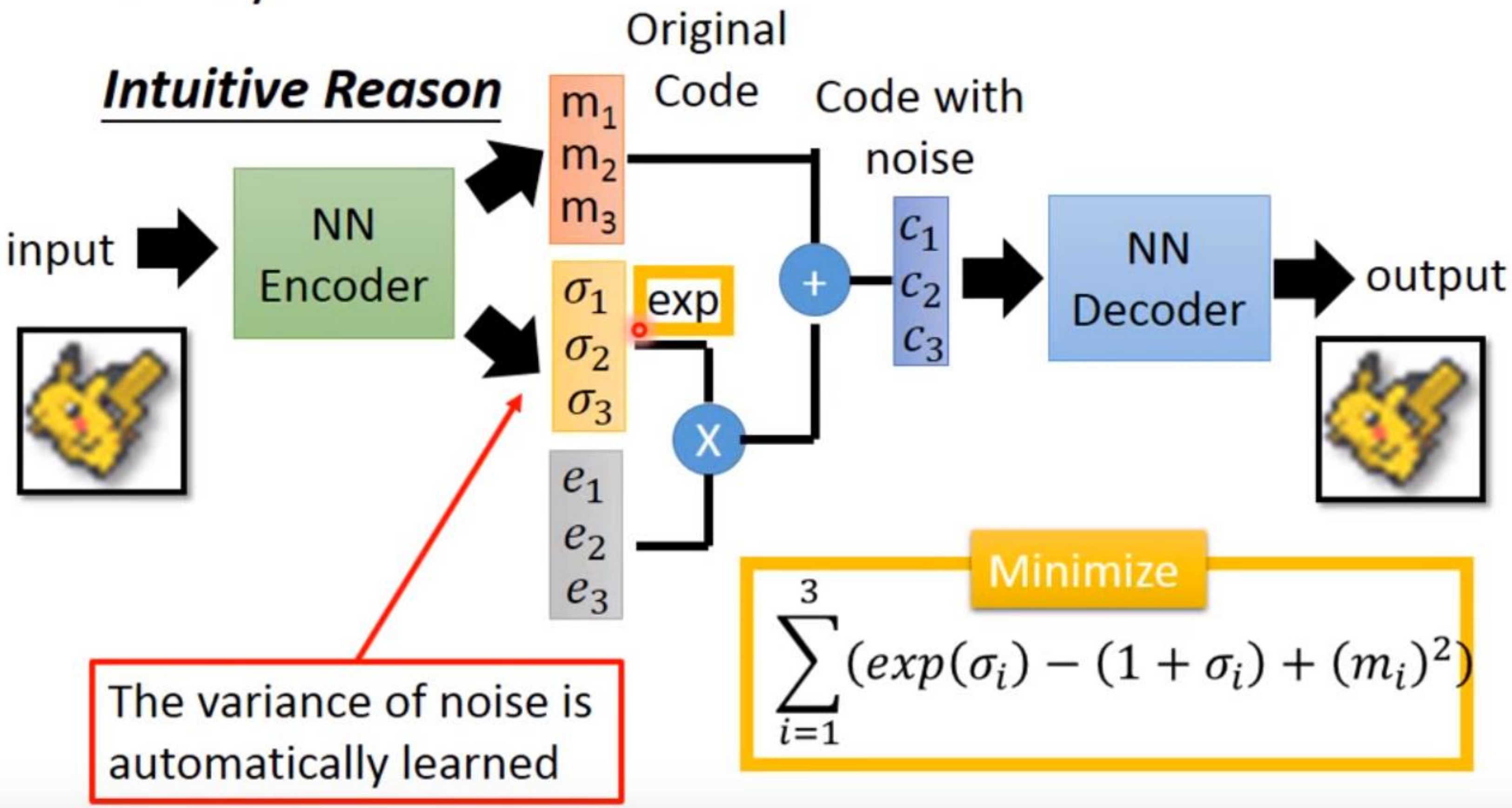
Why VAE?

Intuitive Reason



Why VAE?

What will happen if we only minimize reconstruction error?

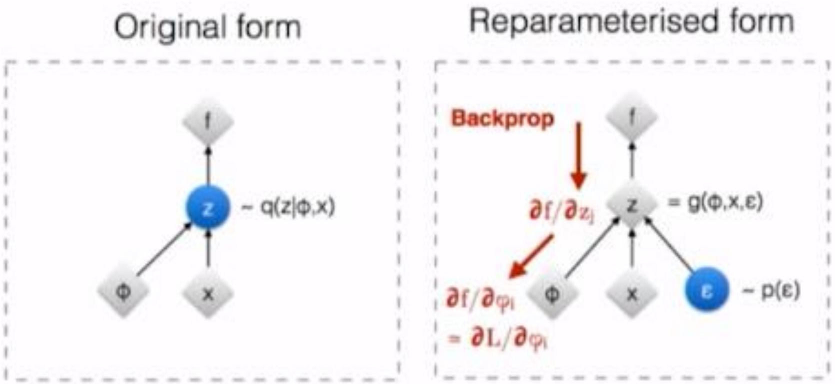


Re-Parameterization Trick - (The great thing about VAE)

Backpropagation not possible through random sampling!

$$z^{(i,l)} = \mu^{(i)} + \sigma^{(i)} \odot \varepsilon_i$$

$$\varepsilon_i \sim N(0,1)$$



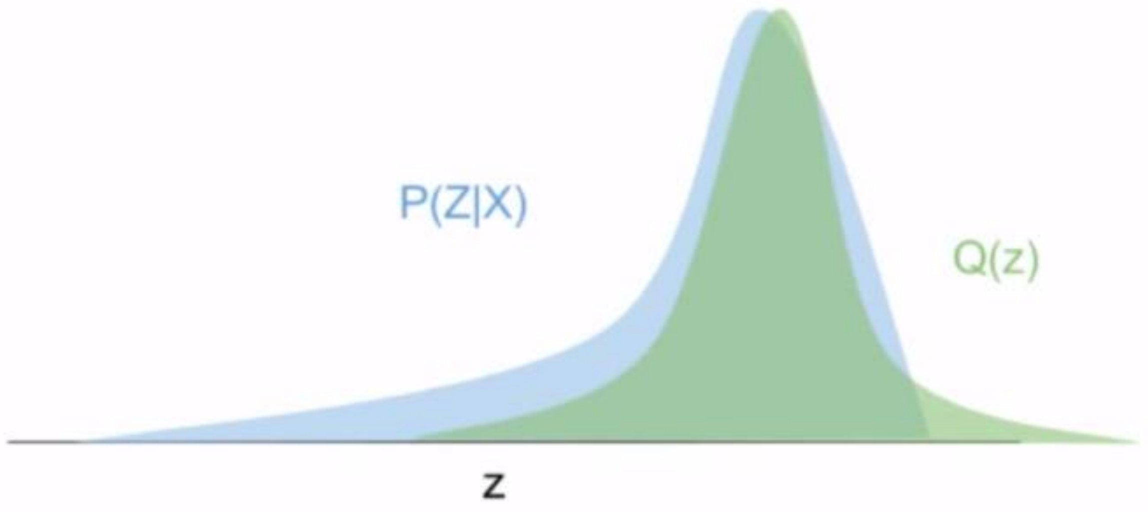
: Deterministic node

: Random node

[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

[Sampling Generative Networks https://arxiv.org/abs/1609.04468]

KL-Divergence - Background



Gaussian Mixture Model

$$P(x) = \sum_m P(m)P(x|m)$$

How to sample?

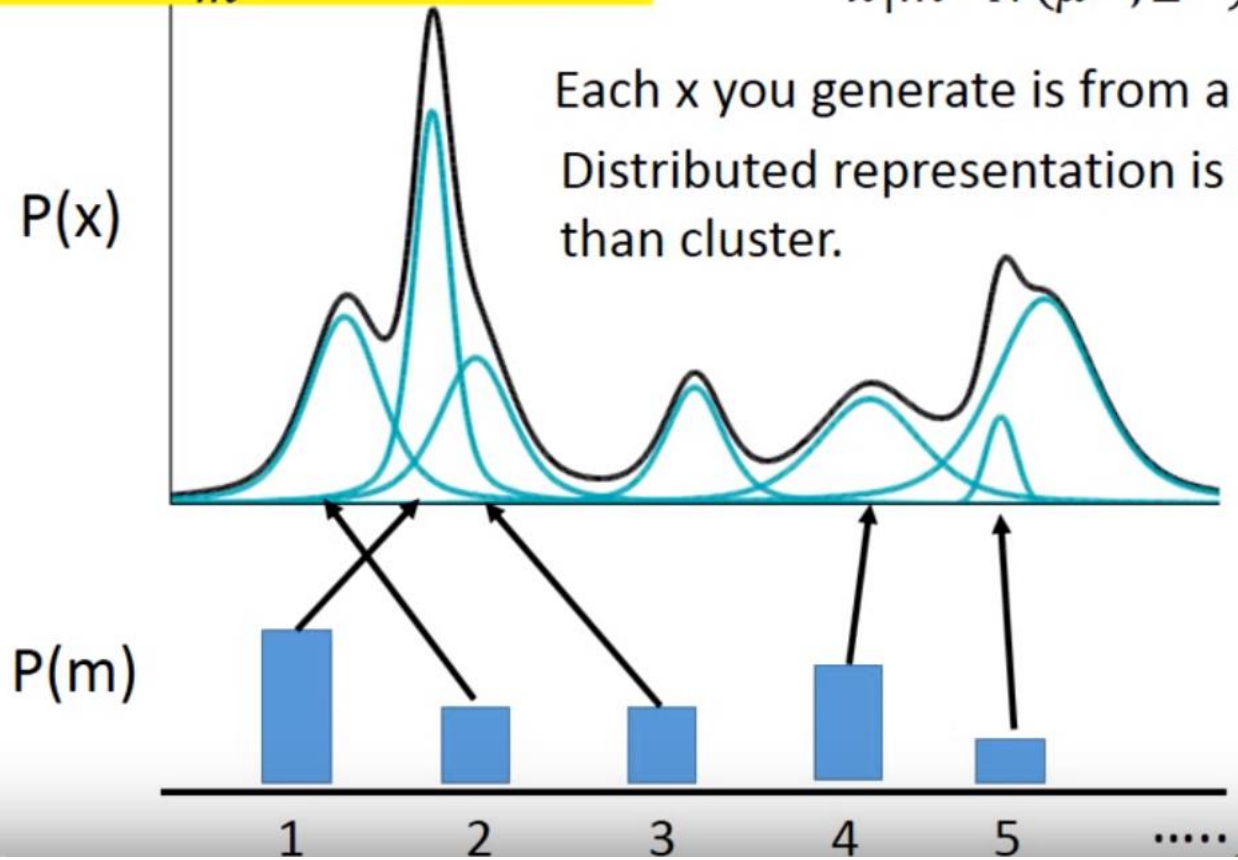
$m \sim P(m)$ (multinomial)

m is an integer

$x|m \sim N(\mu^m, \Sigma^m)$

Each x you generate is from a mixture

Distributed representation is better than cluster.



So VAE is version of Distributed representation of GMM

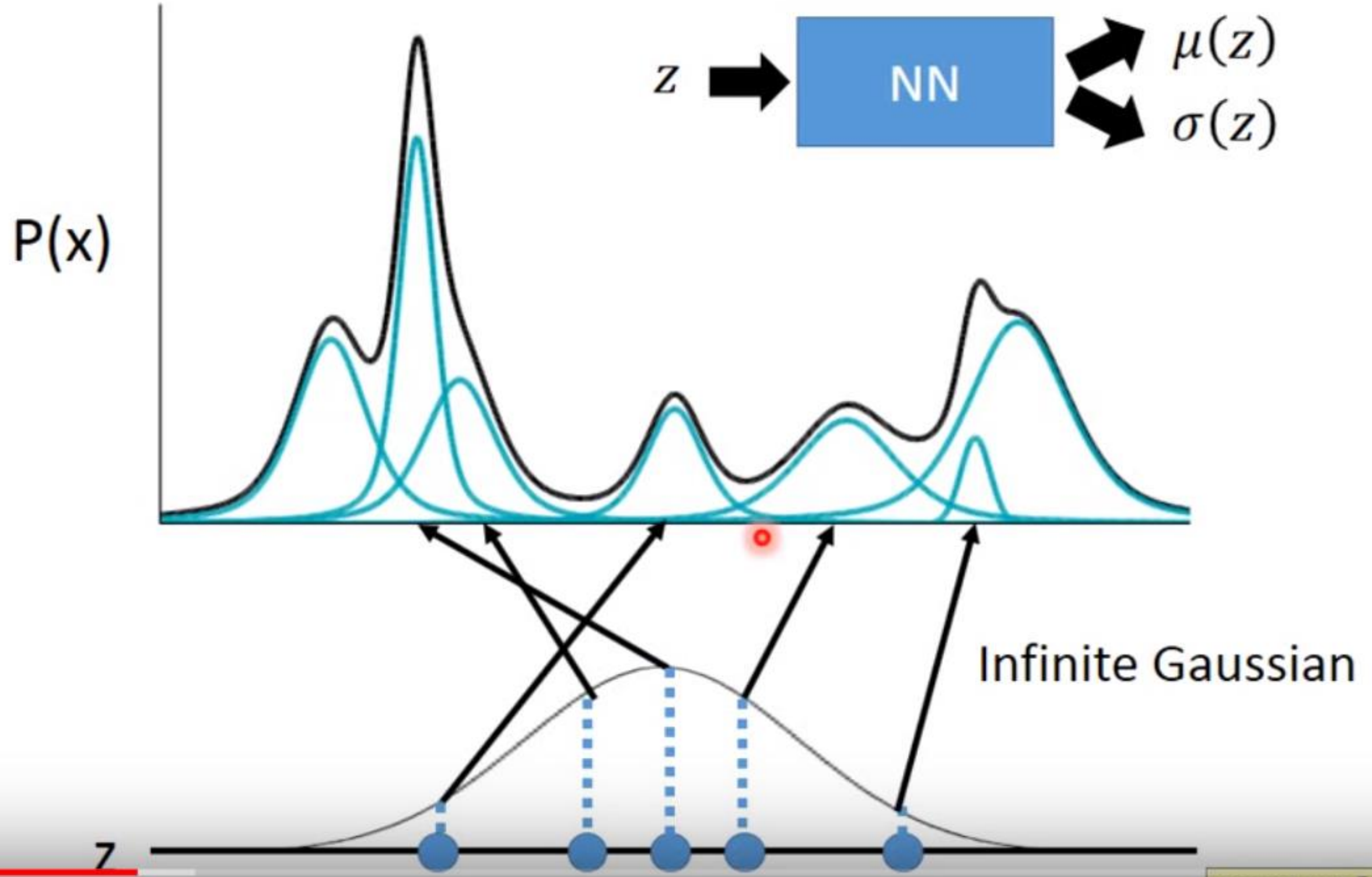
VAE

$z \sim N(0, I)$

$x|z \sim N\left(\mu(z), \sigma(z)\right)$

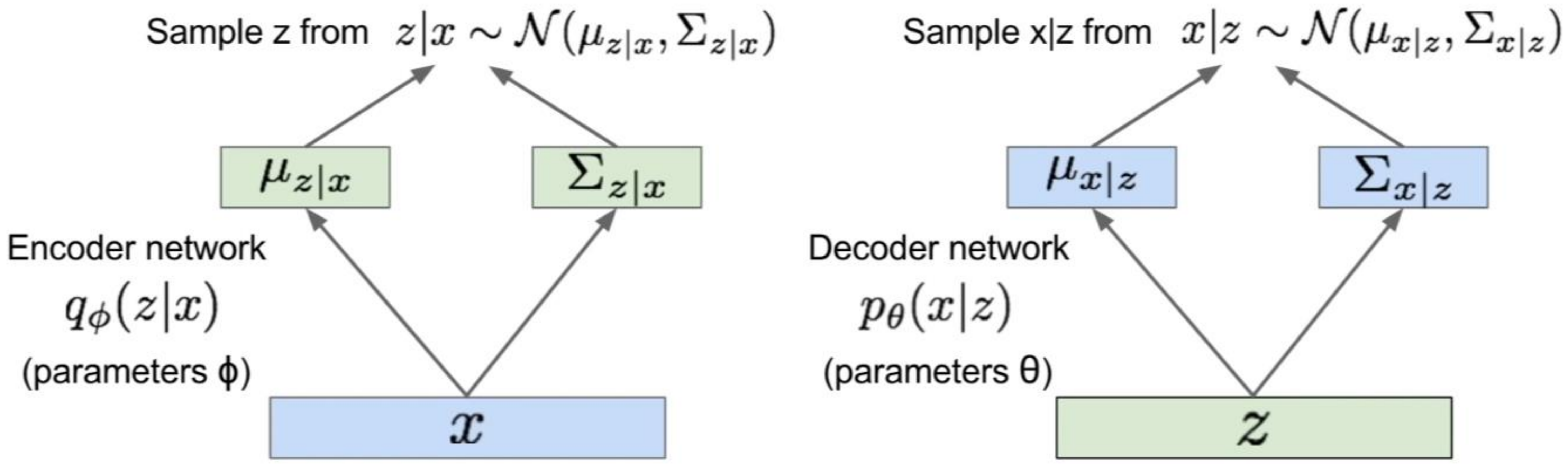
z is a vector from normal distribution

Each dimension of z represents an attribute



Variational Autoencoders

Since we're modeling probabilistic generation of data, encoder and decoder networks are probabilistic



Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

Variational Autoencoders

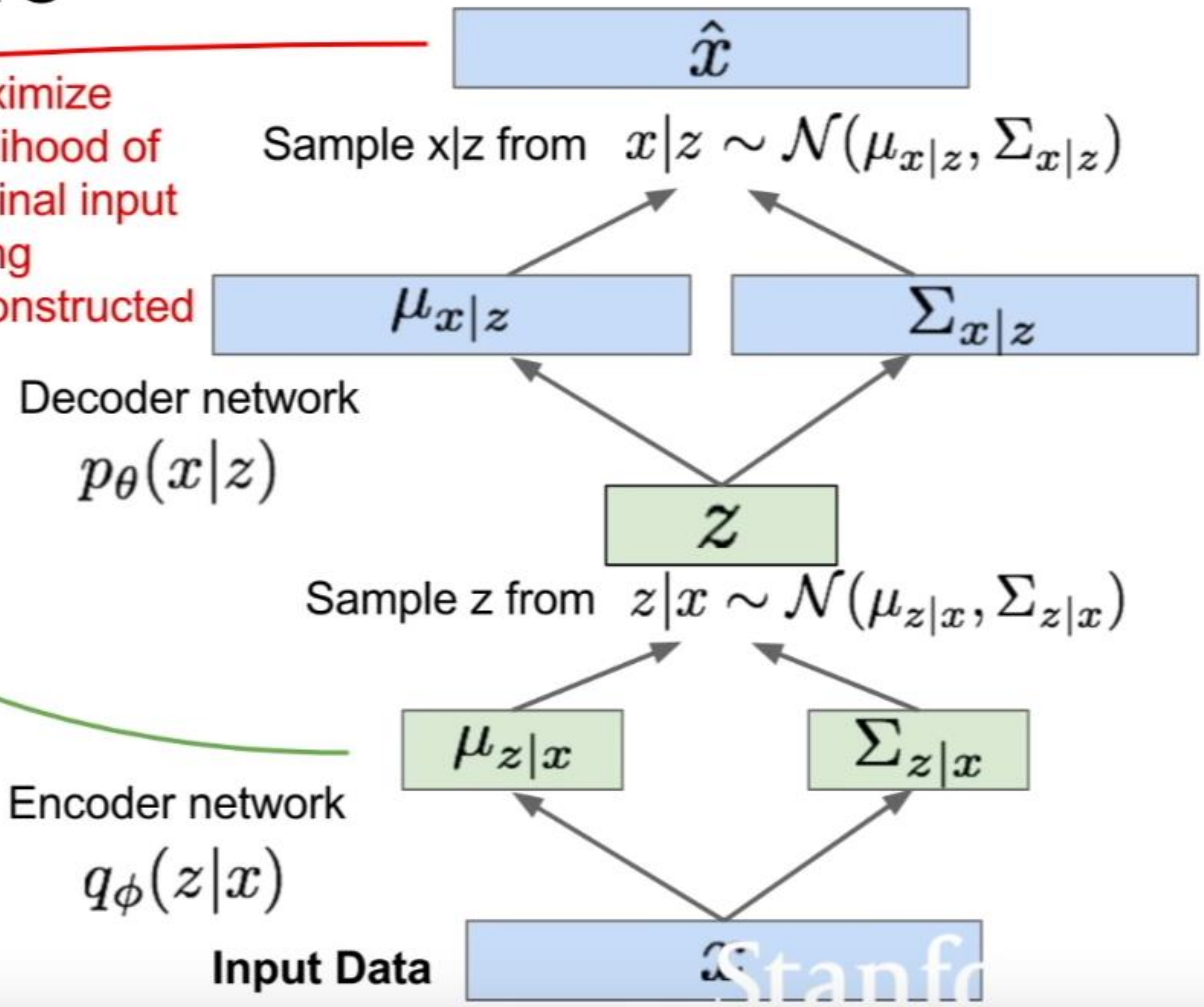
Putting it all together: maximizing the likelihood lower bound

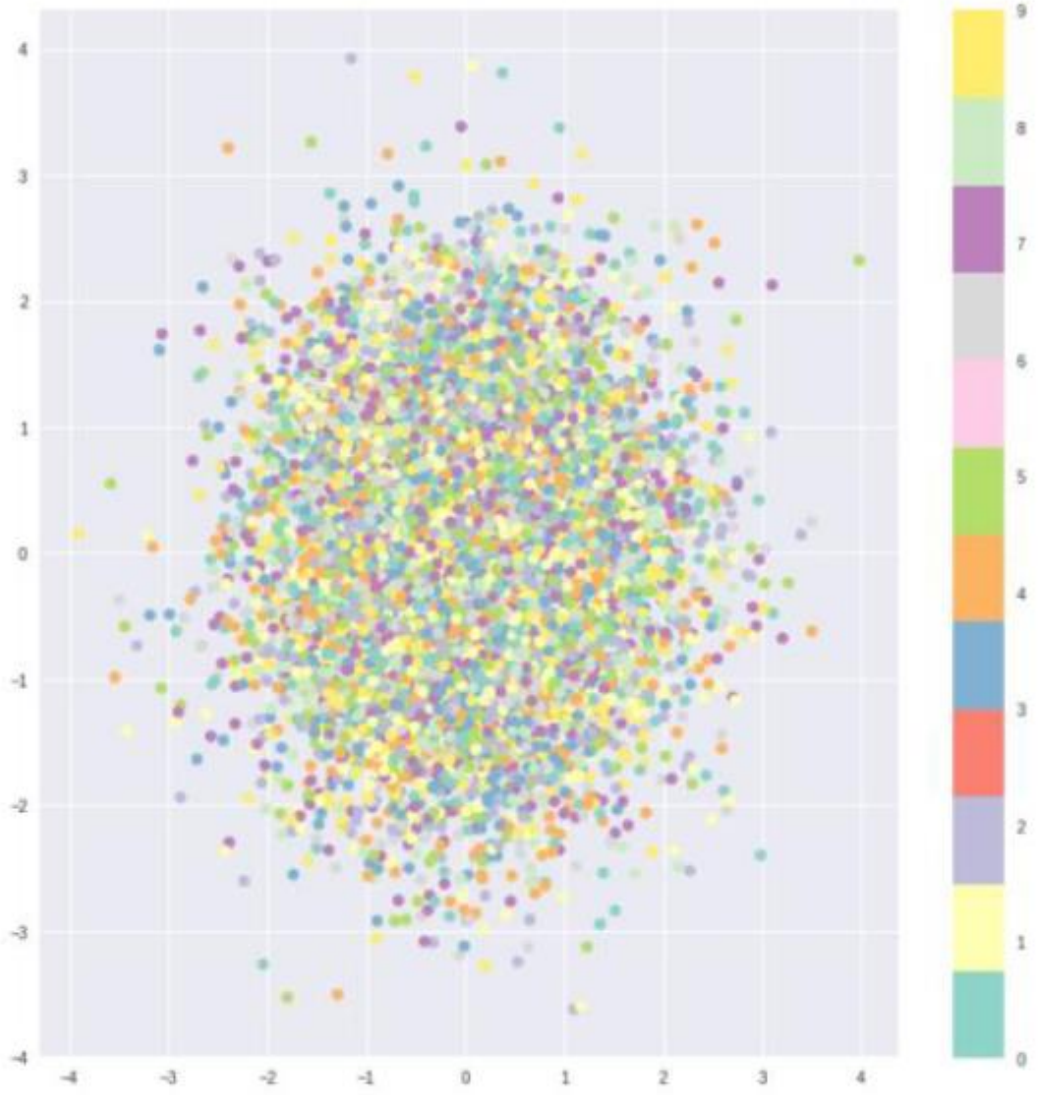
$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

Maximize likelihood of original input being reconstructed

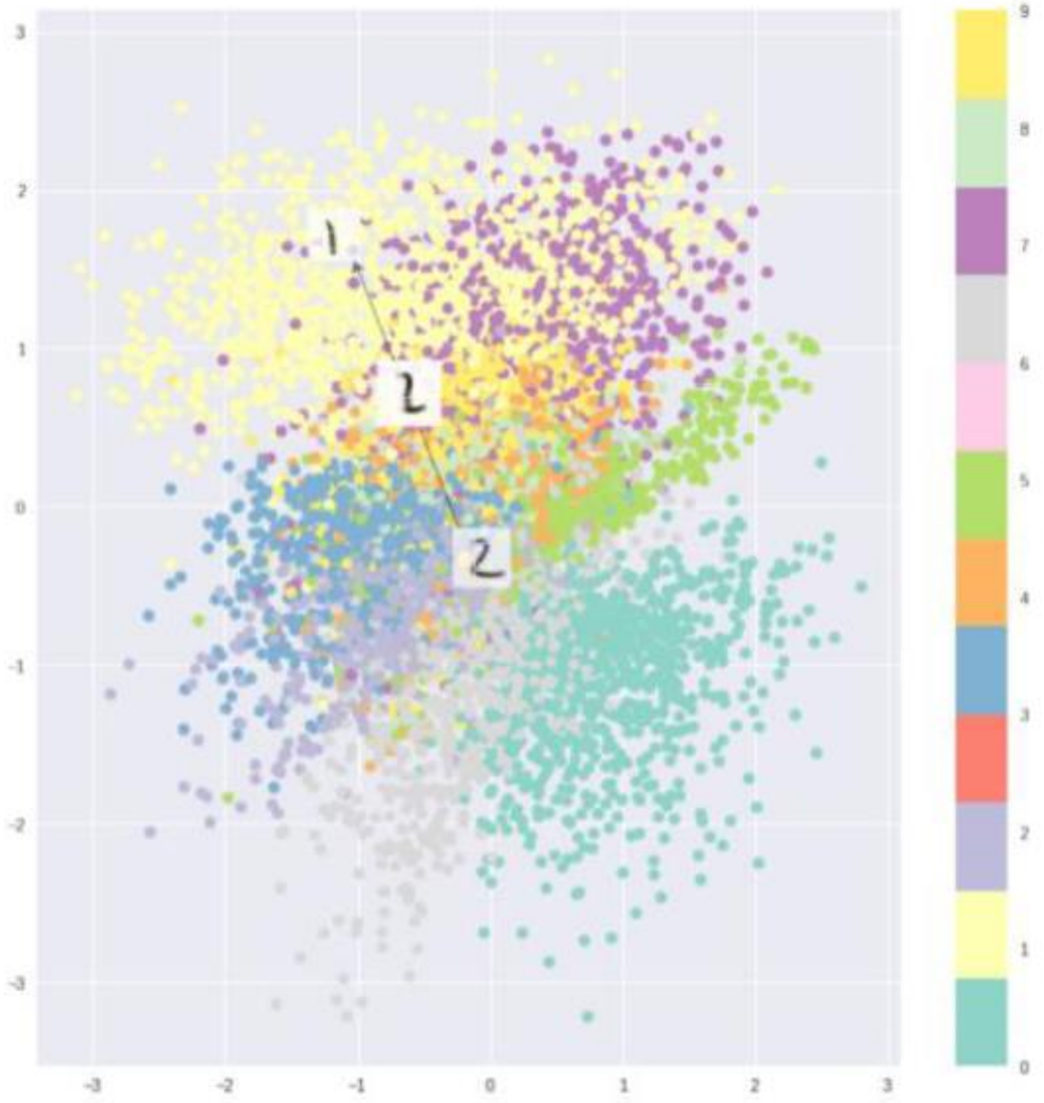
Make approximate posterior distribution close to prior

For every minibatch of input data: compute this forward pass, and then backprop!

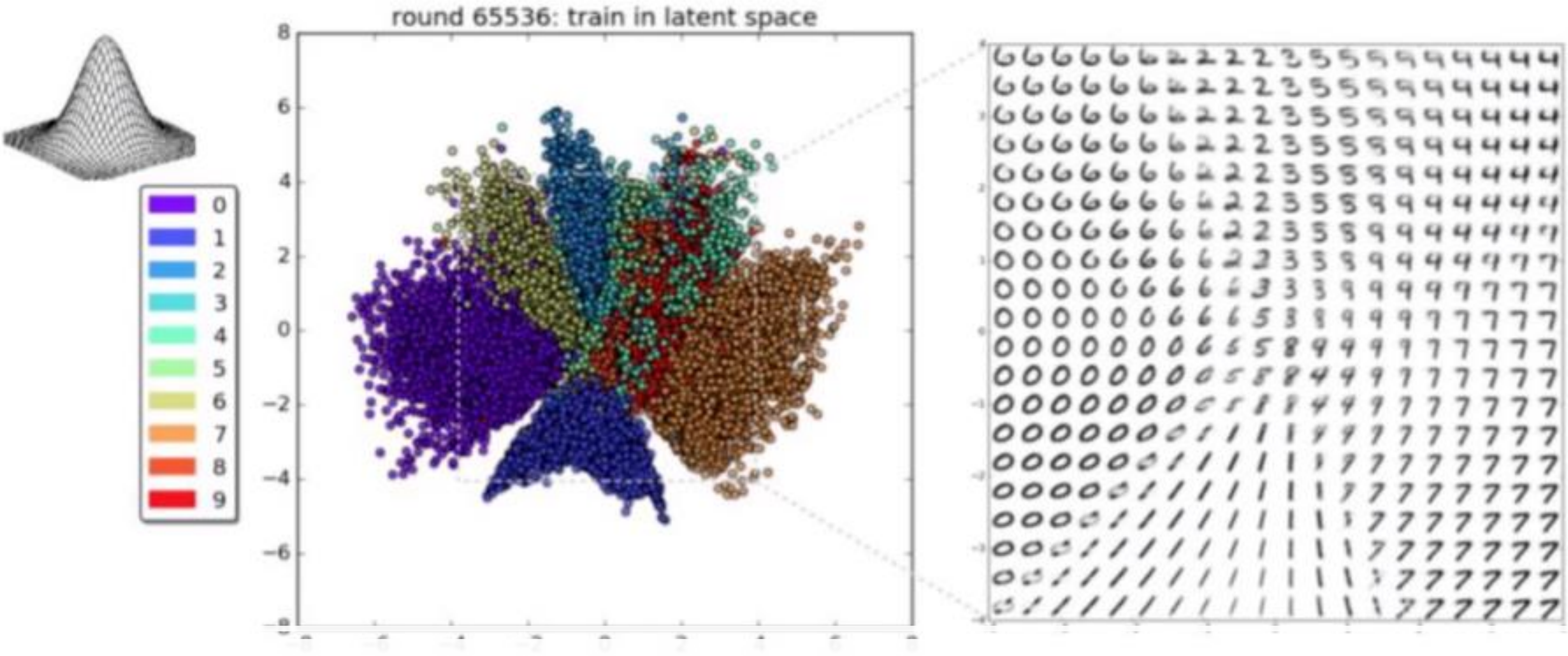




Optimizing using pure KL divergence loss



Optimizing using both reconstruction loss and KL divergence loss



Cross entropy Sampling