Lecture 18: Variational Autoencoders

CS109B Data Science 2

Pavlos Protopapas, Mark Glickman, and Chris Tanner

Outline

Motivation for Variational Autoencoders (VAE)

Mechanics of VAE

Separability of VAE

Training and the math behind everything

Outline

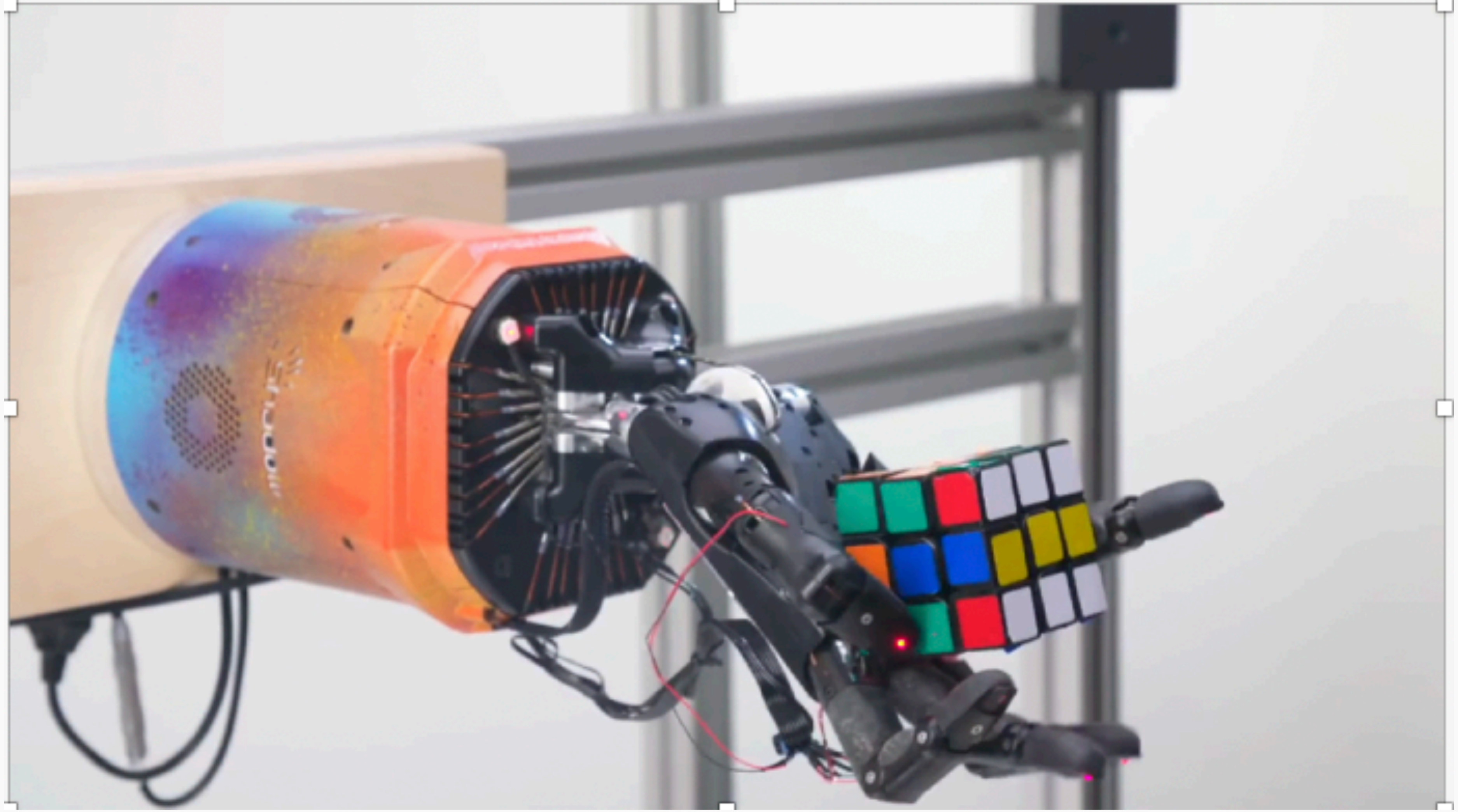
Motivation for Variational Autoencoders (VAE)

Mechanics of VAE

Separability of VAE

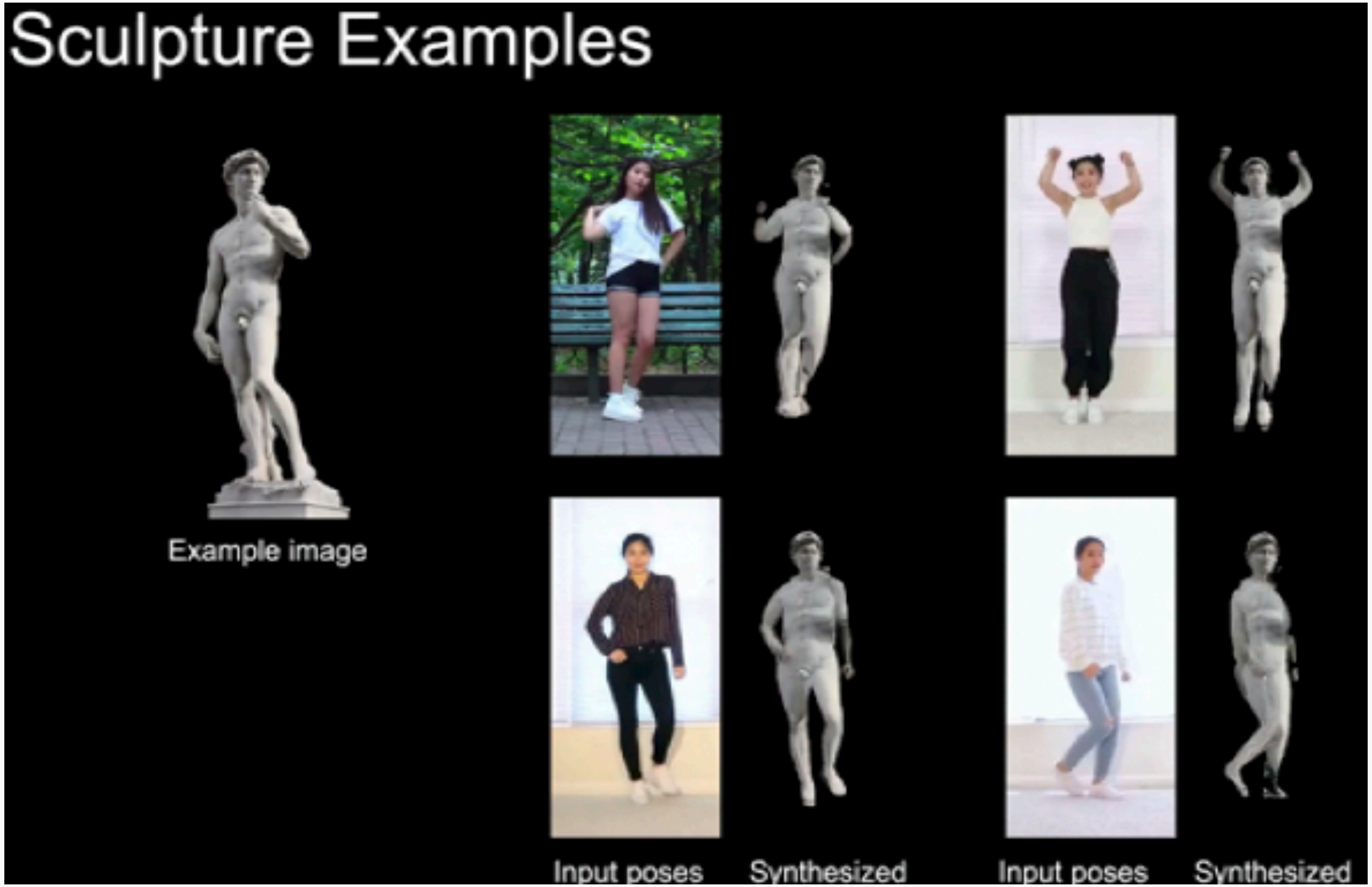
Training and the math behind everything

State of the Art in AI



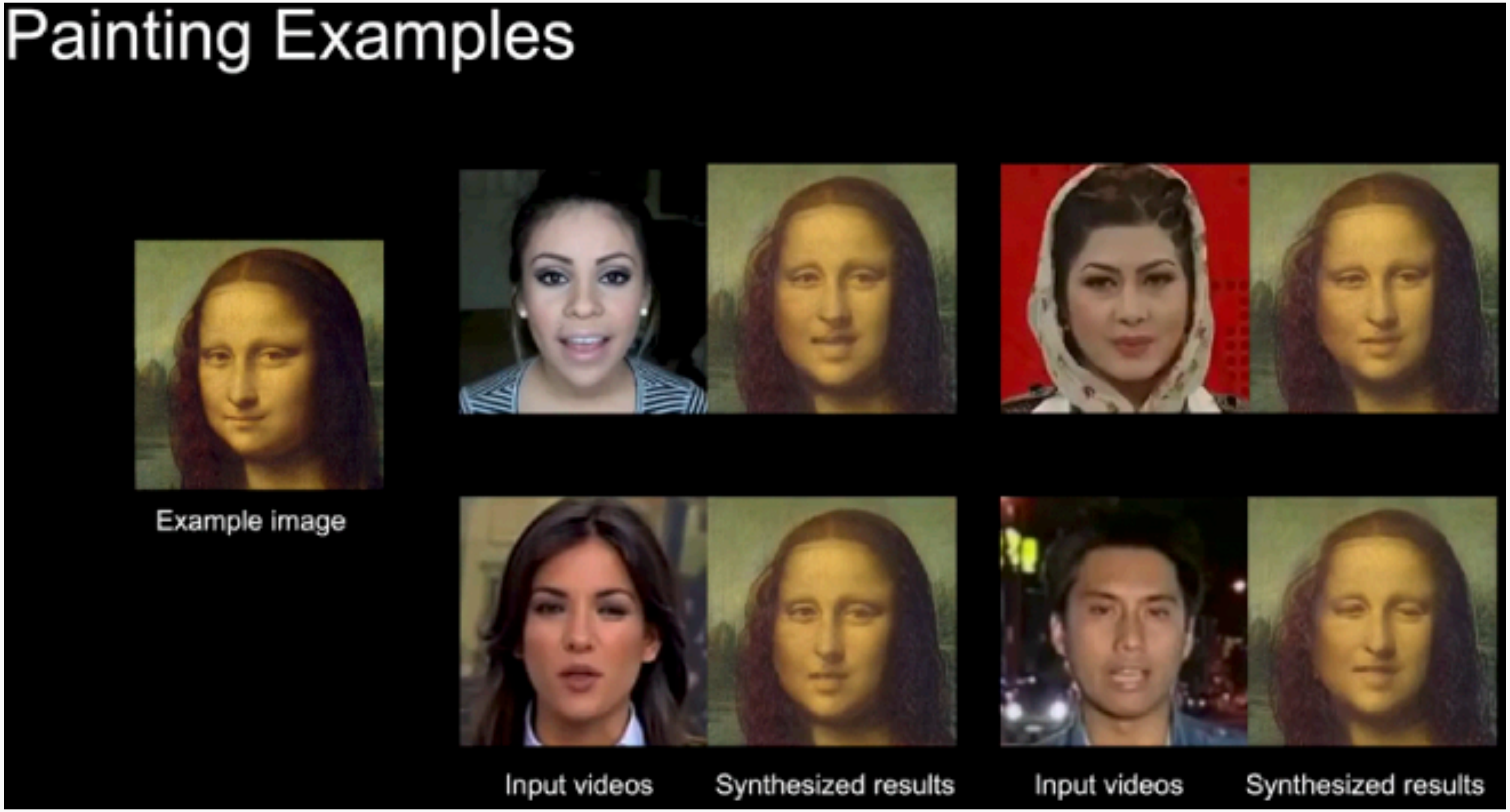
https://openai.com/blog/solving-rubiks-cube/

State of the Art in AI



https://nvlabs.github.io/few-shot-vid2vid/

State of the Art in AI



https://nvlabs.github.io/few-shot-vid2vid/

Generative Modeling

"What I cannot create, I do not understand."
- Richard Feynman

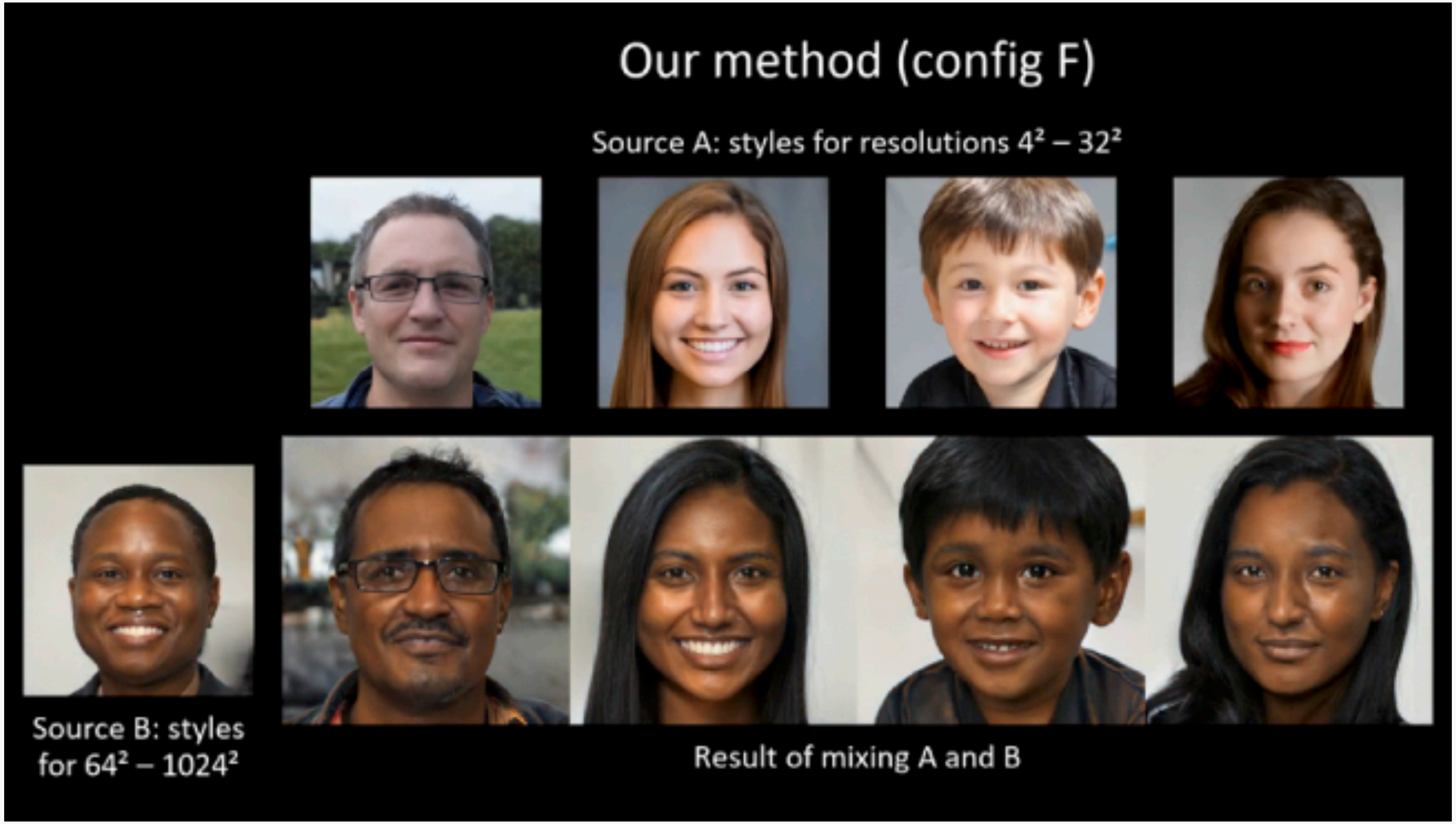
Generative Modeling

https://github.com/tkarras/progressive_growing_of_gans

Generative Modeling

https://github.com/tkarras/progressive_growing_of_gans

Generative Modeling



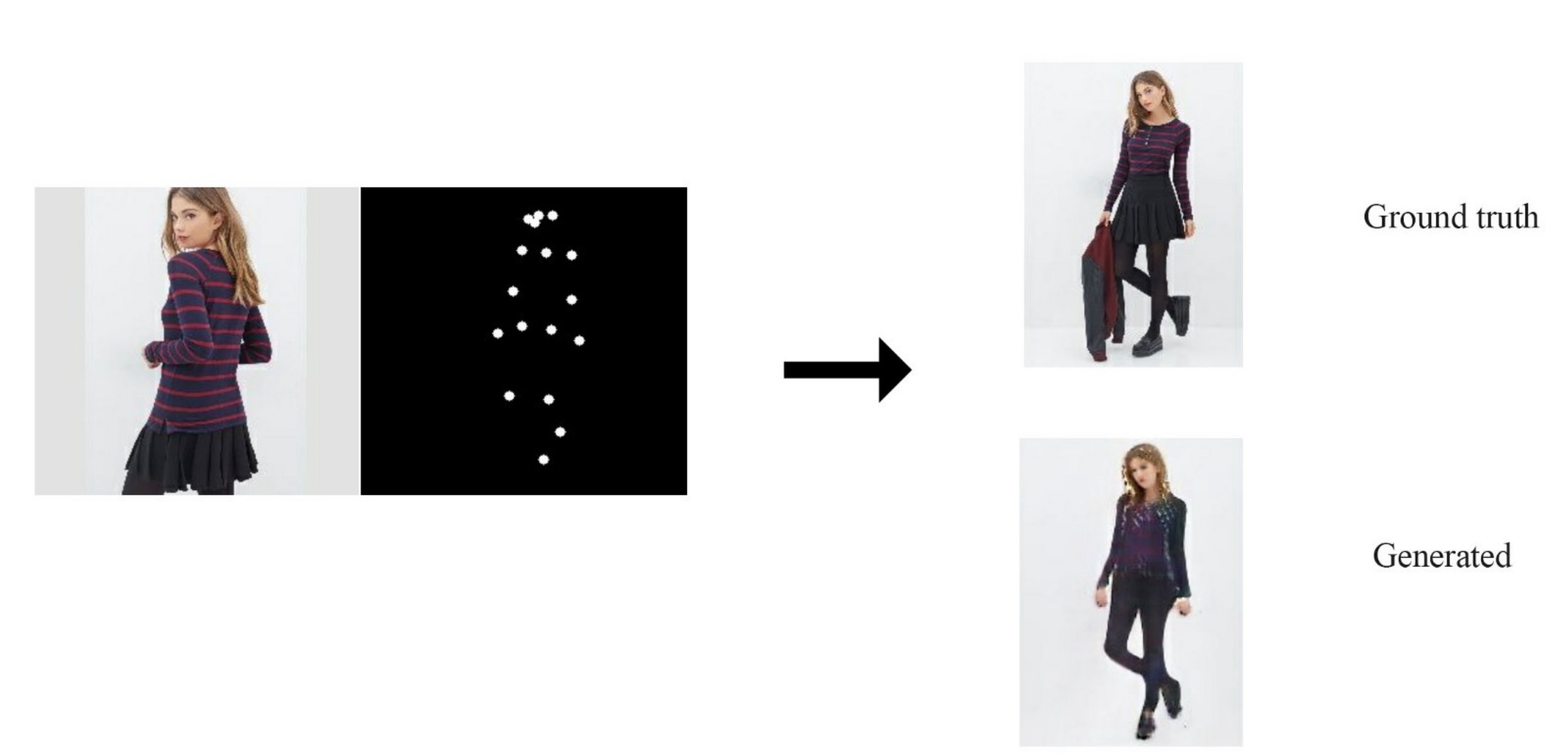
https://github.com/NVlabs/stylegan2

Generative Modeling

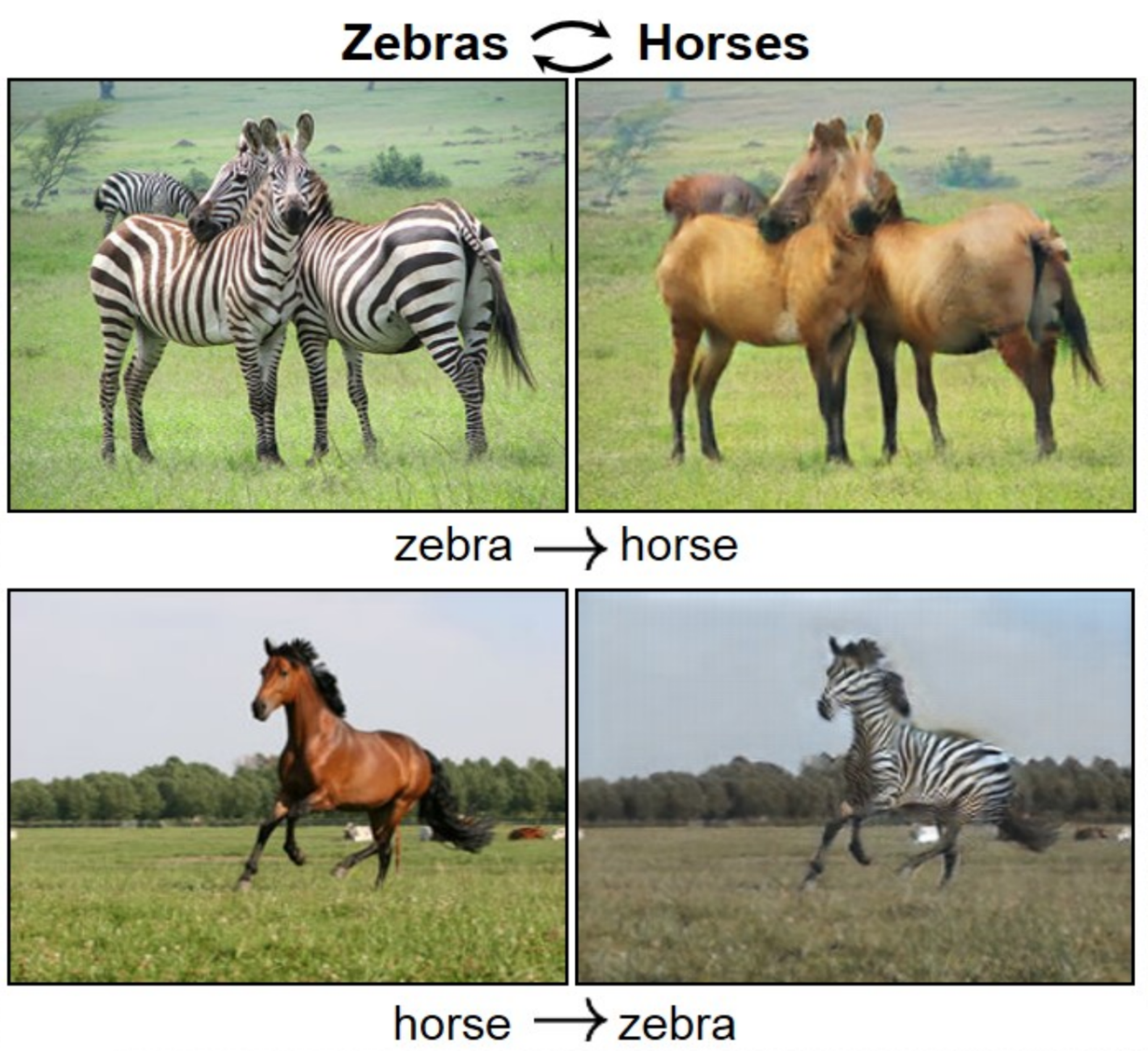
Figure 7: Generated samples

https://arxiv.org/pdf/1708.05509.pdf

Generative Modeling



Generative Modeling



(a) MidiNet model 1

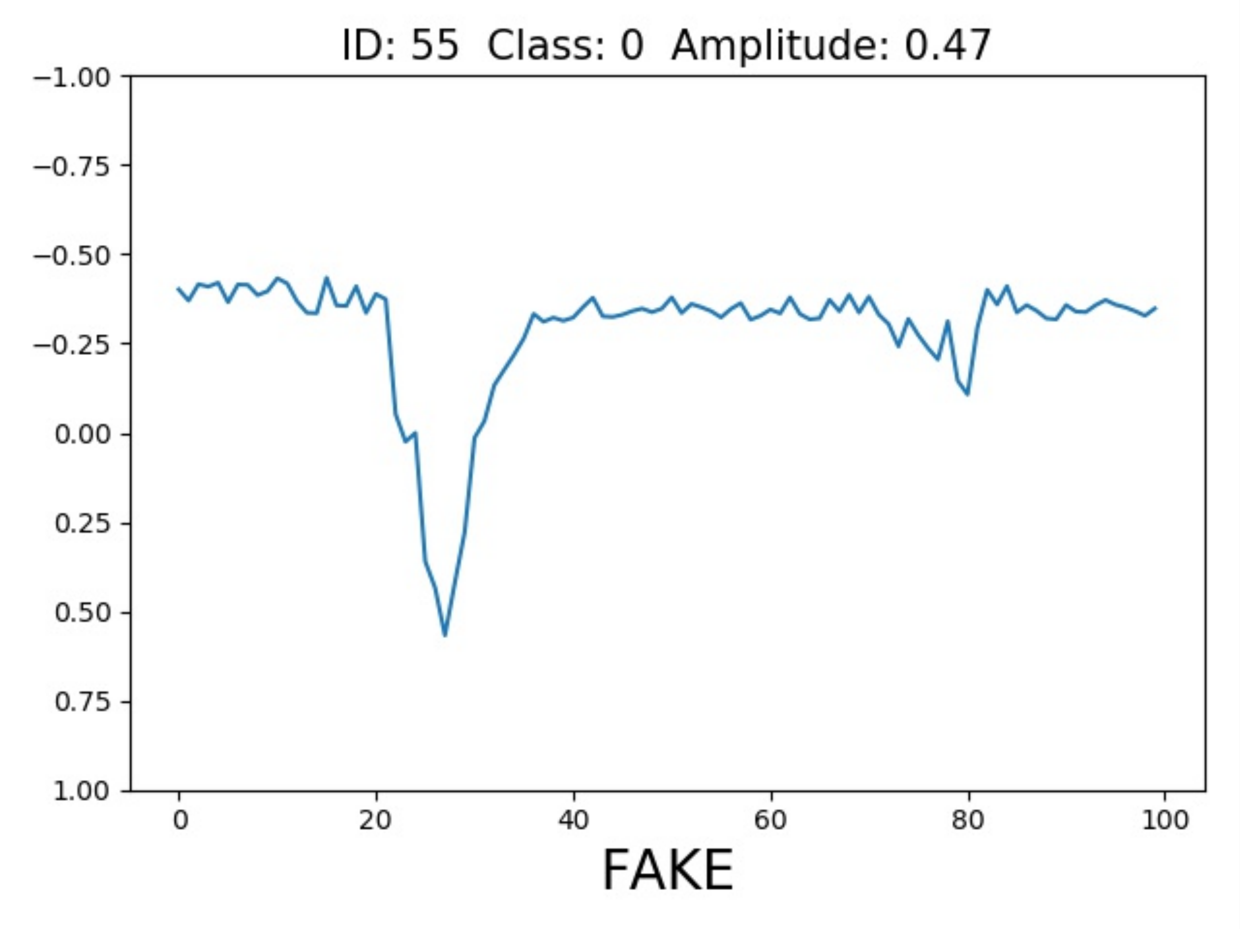
(b) MidiNet model 2

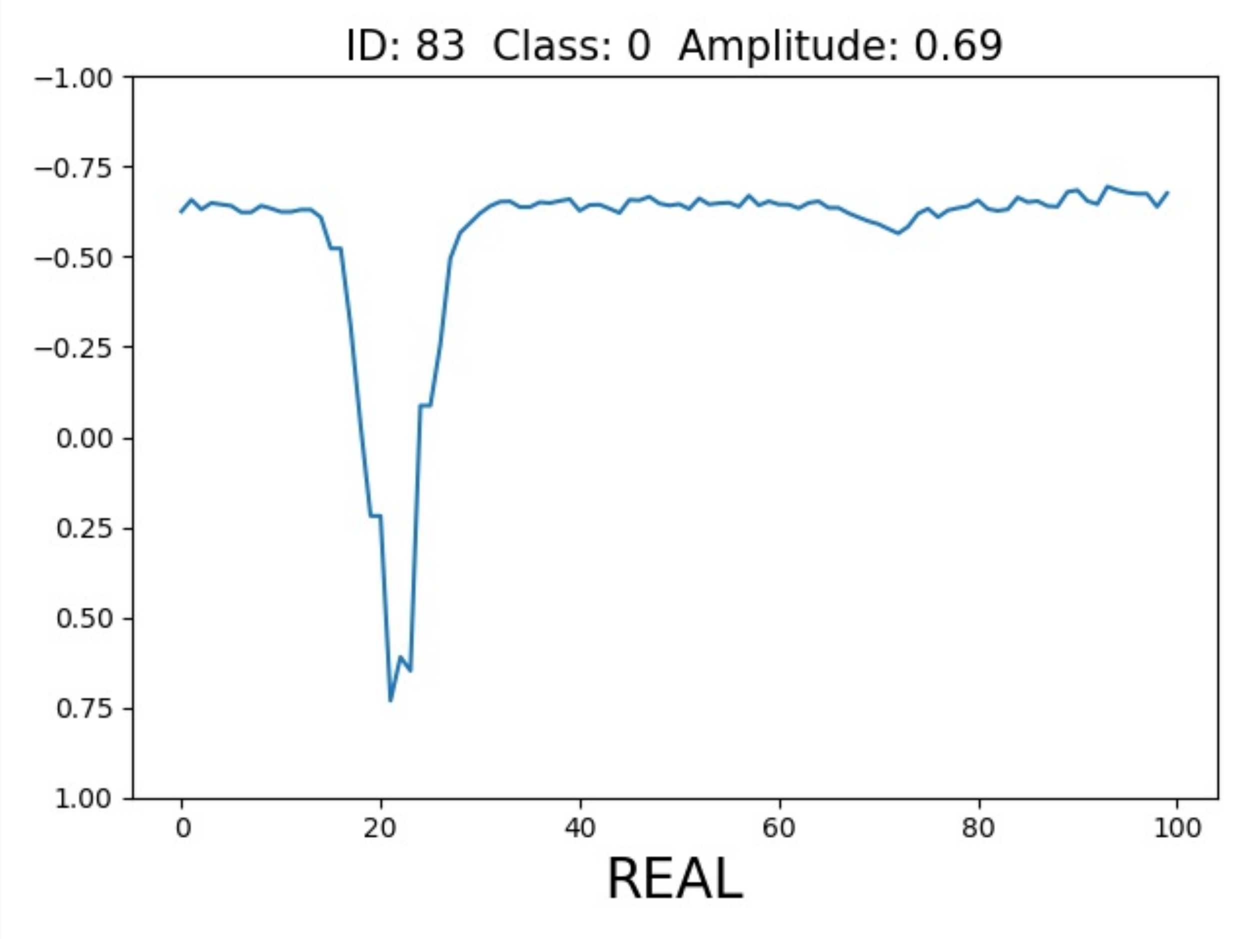
(c) MidiNet model 3

Figure 3. Example result of the melodies (of 8 bars) generated by different implementations of MidiNet.

Another use of generating new data is to give us ideas and options. Suppose we're planning a house. We can give the computer the space we have available, and its location. From this, the computer can can give us some ideas.

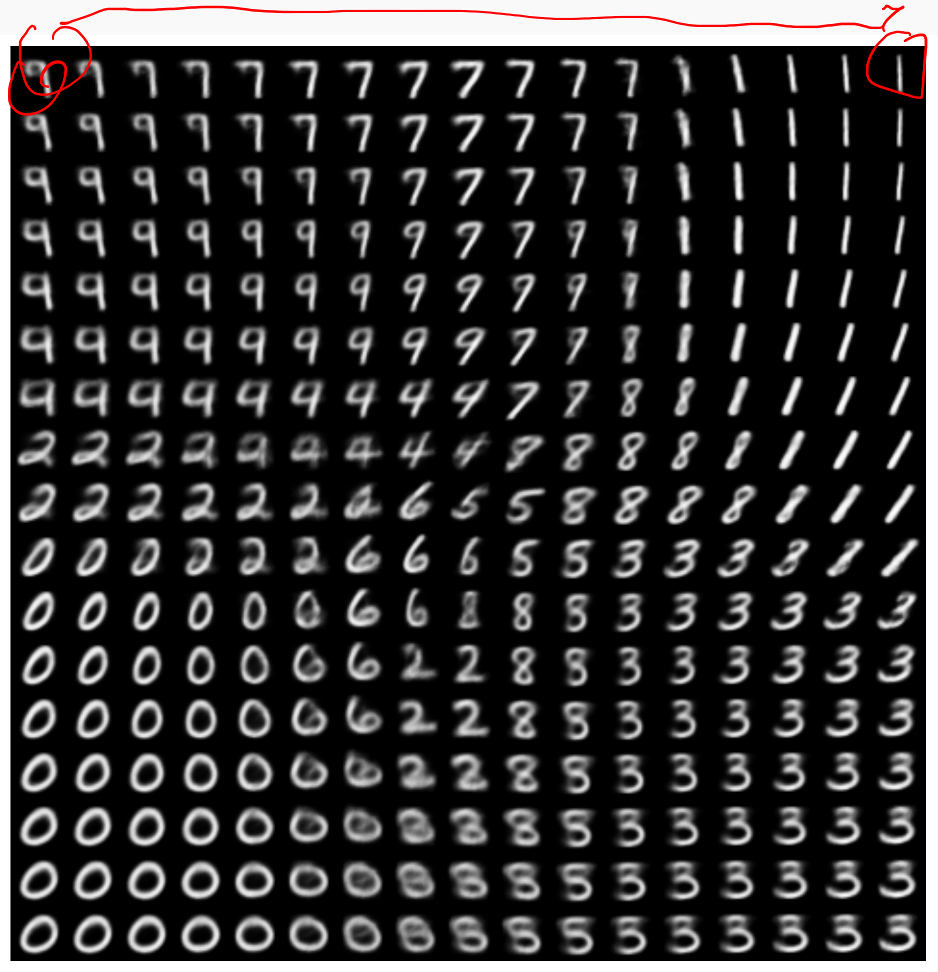
Big networks require big data, and getting high-quality, labeled data is difficult. If we're generating that data our selves, we can make as much of it as we like.





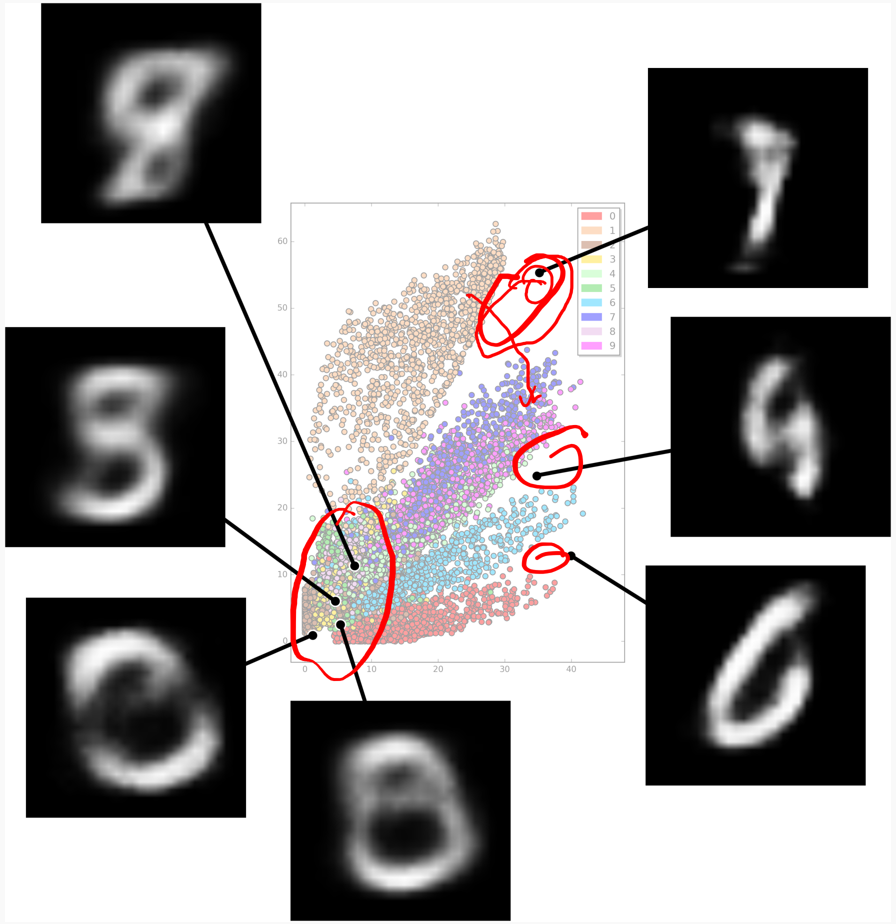
Generating Data

We saw how to generate new data with a AE in Lecture 12.



Problems with Autoencoders (from lecture 12)

- Gaps in the latent space
- Discrete latent space
- Separability in the latent space



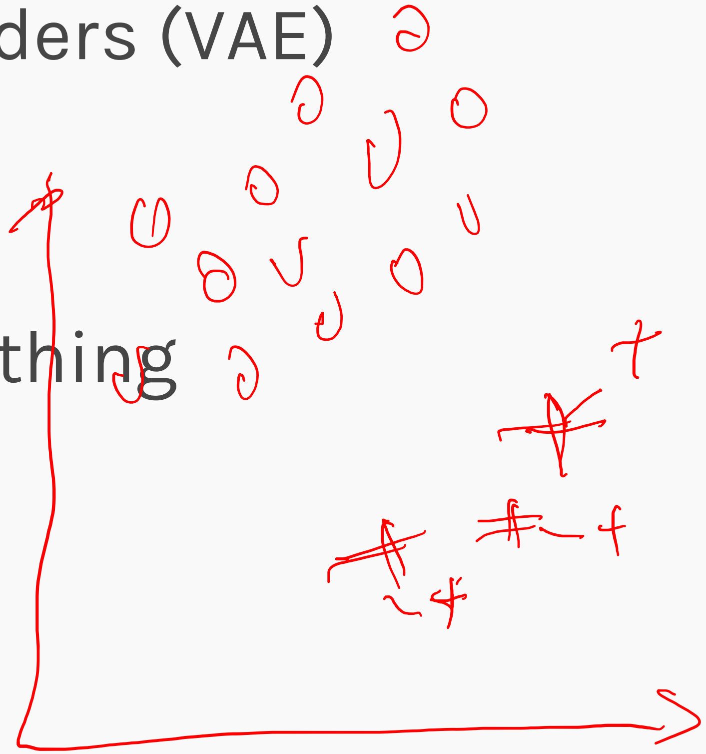
Outline

Motivation for Variational Autoencoders (VAE)

Mechanics of VAE

Separability of VAE

Training and the math behind everything



Generative models

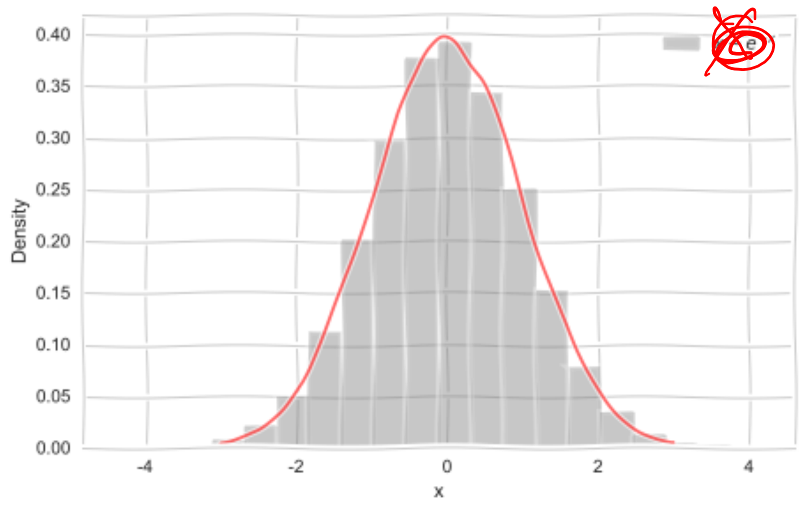
Imagine we want to generate data from a distribution,

$$x \sim p(x)$$

e.g.

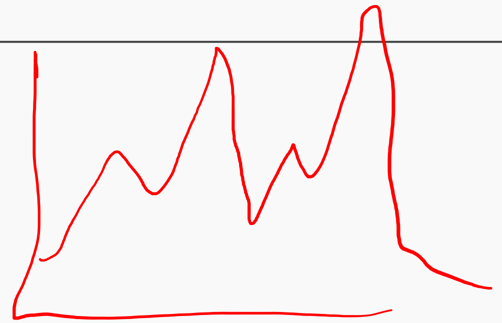
$$x \sim \mathcal{N}(\mu, \sigma)$$

$x \sim N(0,1)$

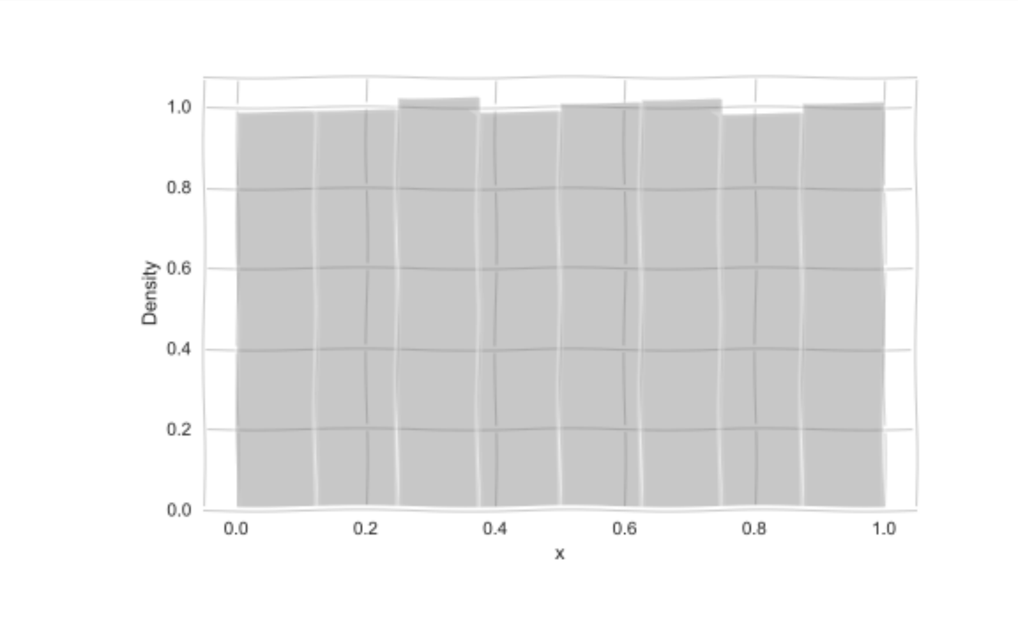


Generative models

But how do we generate such samples?



$$z \sim \text{Unif}(0, 1)$$

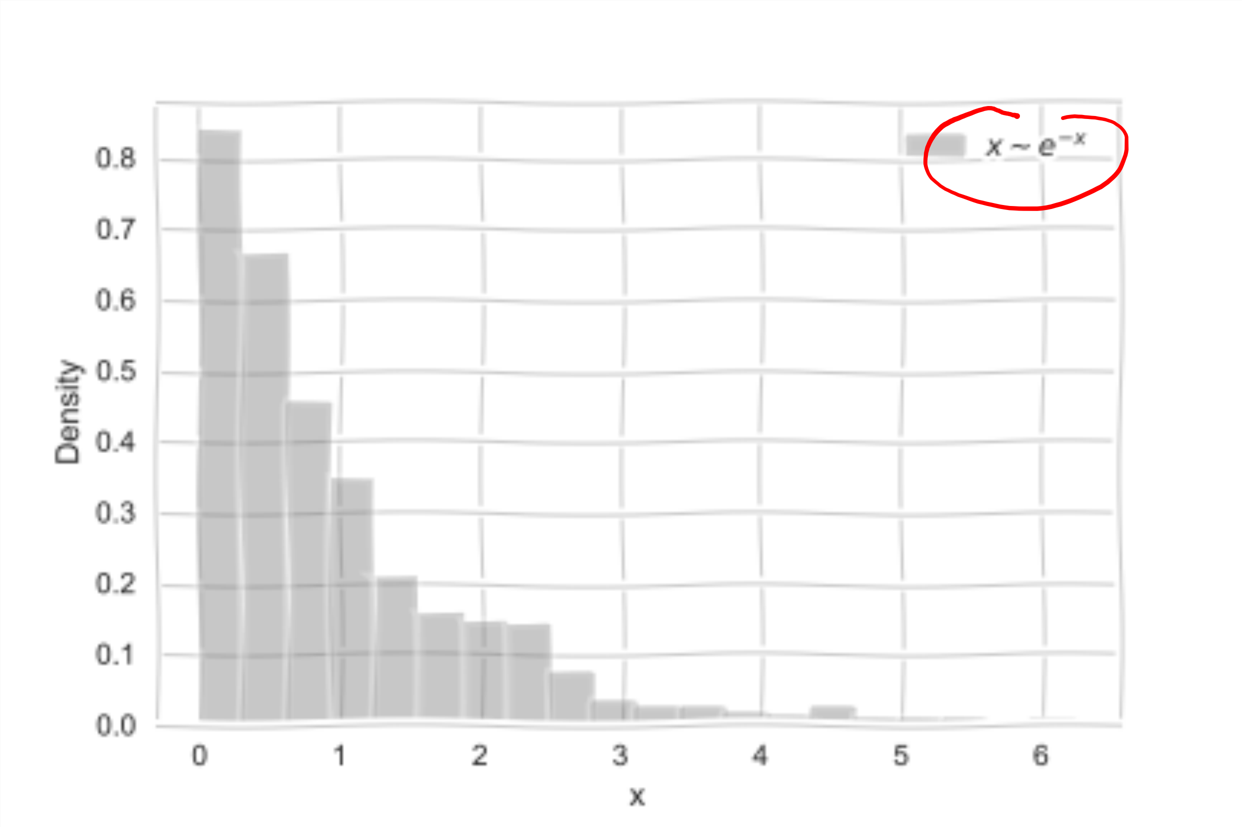


Generative models

But how do we generate such samples?

$$z \sim \mathrm{Unif}(0,1) \quad \boxed{\mathrm{x} = \ln \mathrm{z}}$$

$x \sim e^{-x}$



Generative models

In other words we can think that if we choose $\mathbf{z} \sim \boldsymbol{Uniform}$ then there is a mapping:

$$\boxed{x = f(z)}$$

z ~ unif

0.1
0.5
0.3
0.2

-1
-0.5

such as:

$$x \sim p(x)$$

where in general f is some complicated function.

We already know that **Neural Networks are great in learning complex functions.**

SPECIFY

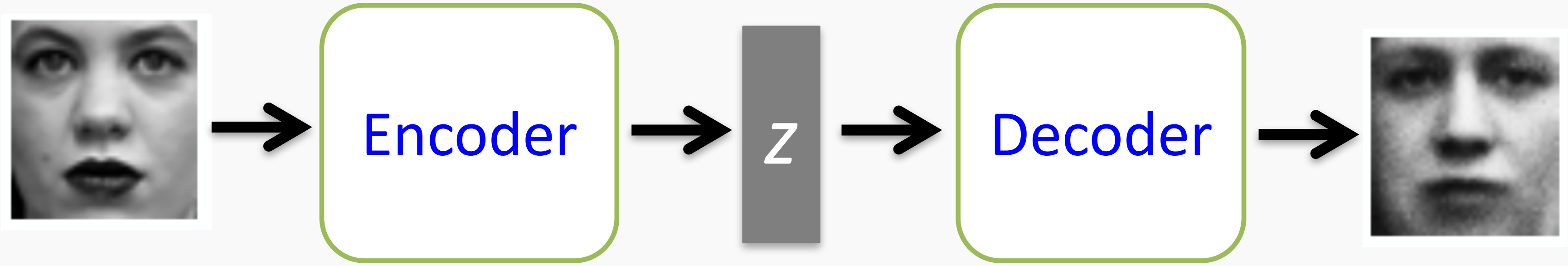
unif(z)

I WANT

$$z \sim g(z) \longrightarrow x = f(z) \longrightarrow x \sim p(x)$$

Traditional Autoencoders

In traditional autoencoders, we can think of encoder and decoders as some function mapping.

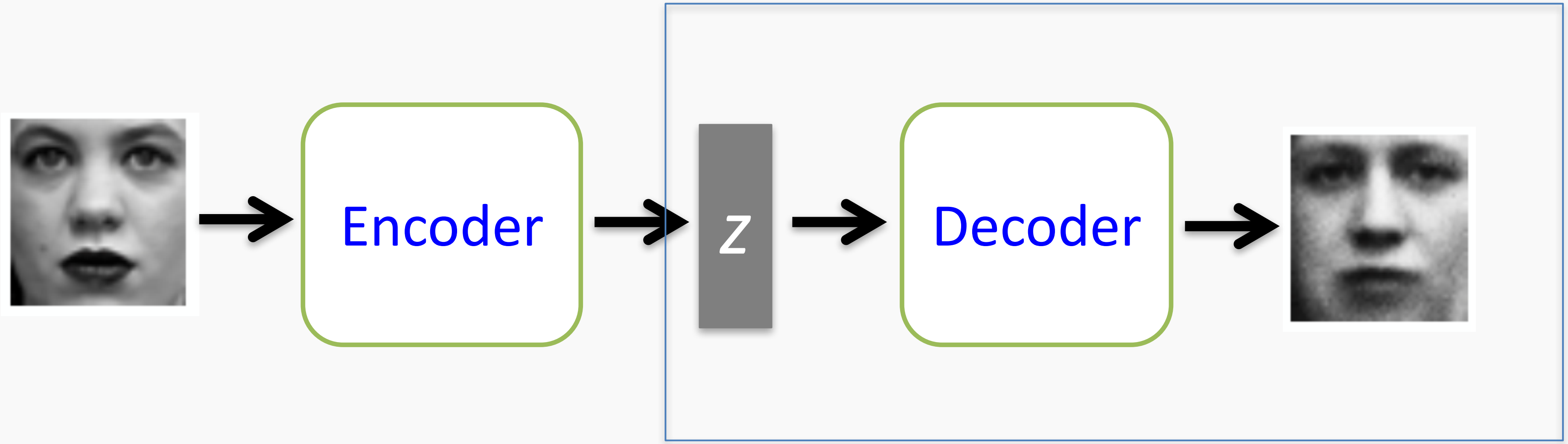


$$\boxed{z = h(x)}$$

$$\boxed{\hat{x} = f(z)}$$

Variational Autoencoders

To go to variational autoencoders, we need to first add some stochasticity and think of it as a probabilistic modeling.

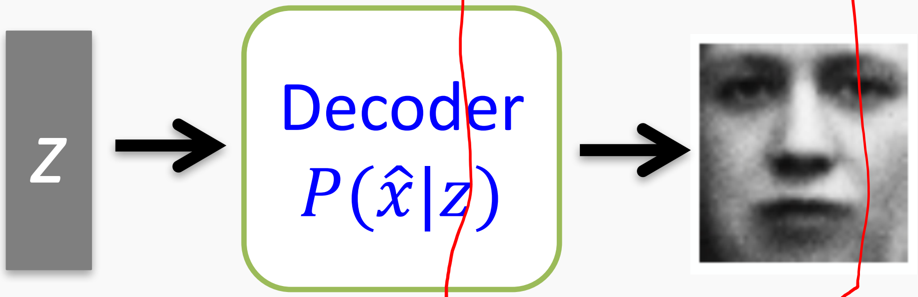


Variational Autoencoders

z ~ unif   x = f(z)   x ~ P(x)

Sample from $g(z)$

e.g. Standard Gaussian



$z \sim g(z)$ $\hat{x} = f(z)$ $\hat{x} \sim P(x|z)$

Variational Autoencoders

Traditional AE

Encoder → z

Latent space 1-D

$X_1 \rightarrow \text{VAE} \rightarrow \hat{X}_1$

$X_1 \rightarrow \text{VAE} \rightarrow \tilde{X}$

np.random.normal(loc=μ, scale=σ, size=1)

2-D

x

Encoder

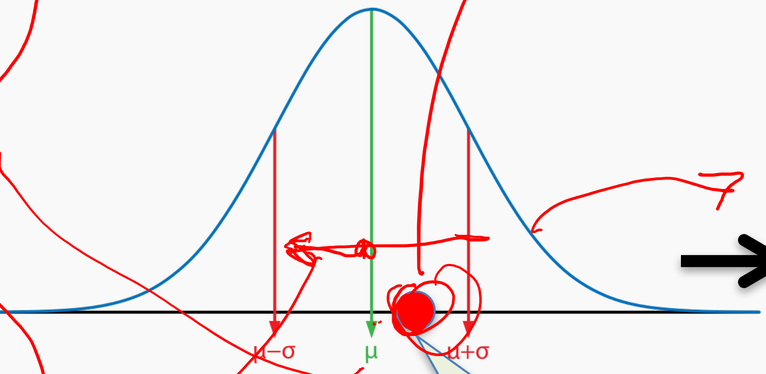
W_e

z_1

Consider this to be the mean of a normal μ

z_2

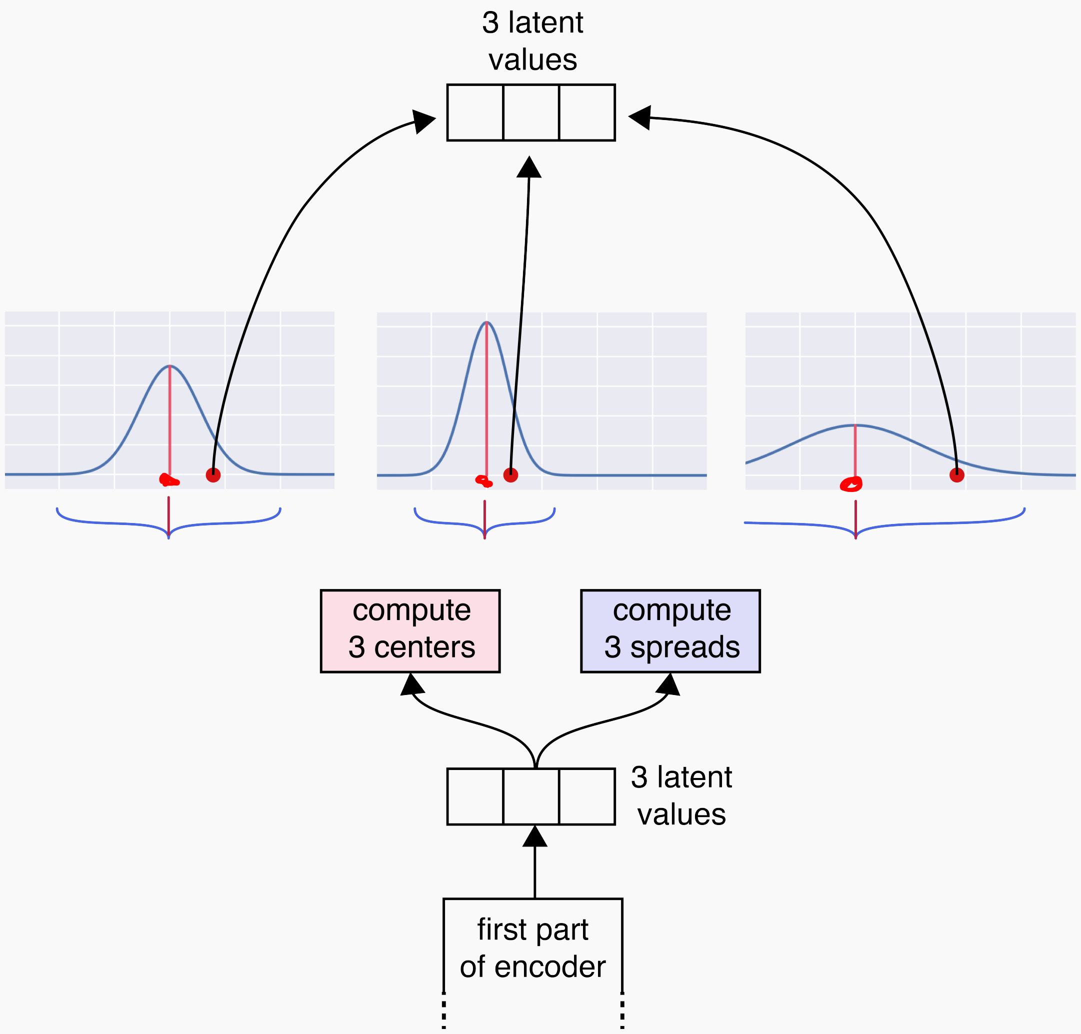
Consider this to be the std of a normal σ



→ Decoder

Randomly chosen value
Latent value, z

Variational Autoencoders



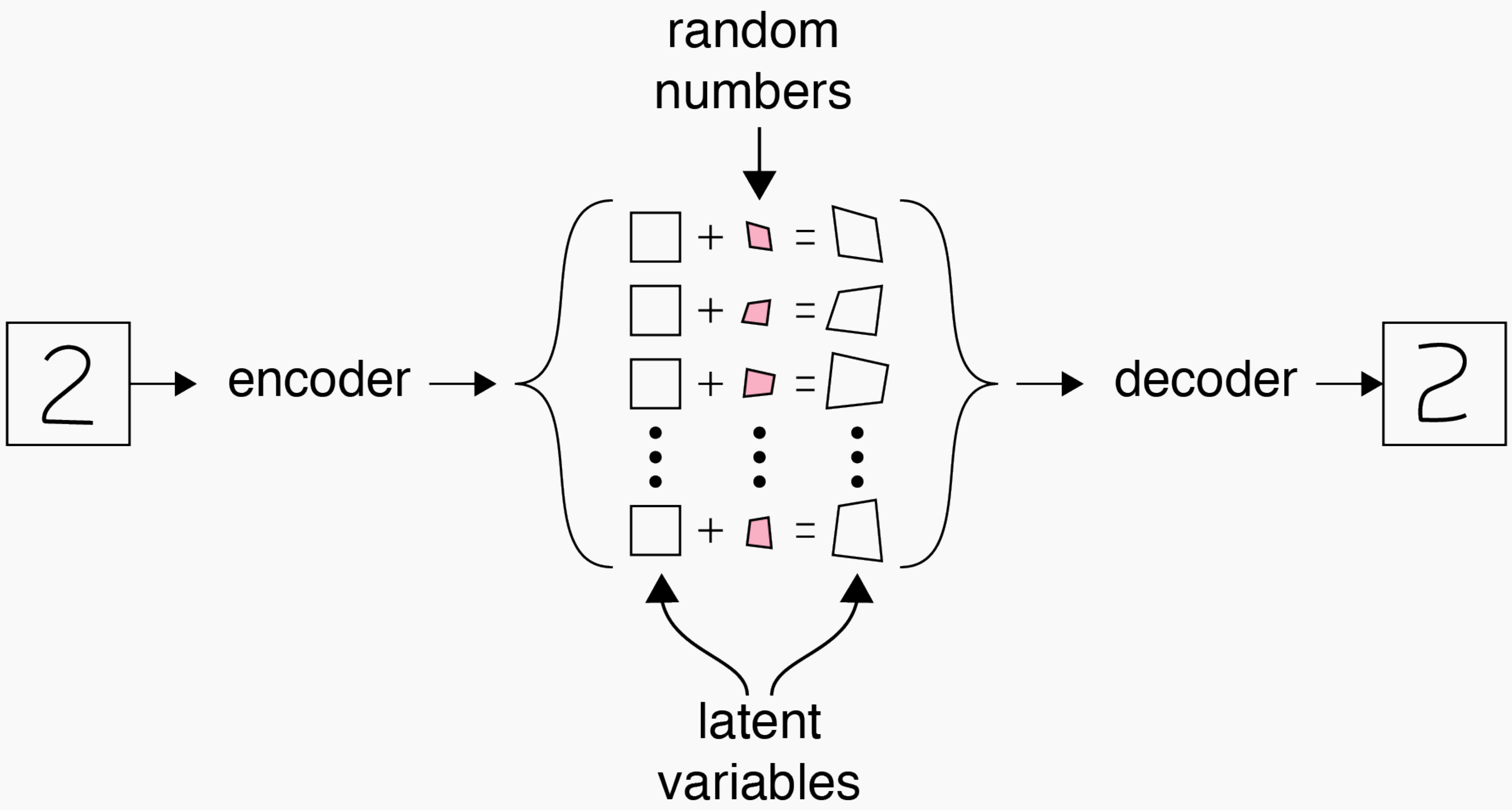
traditional AE

$x \to z \sim 30 \to$

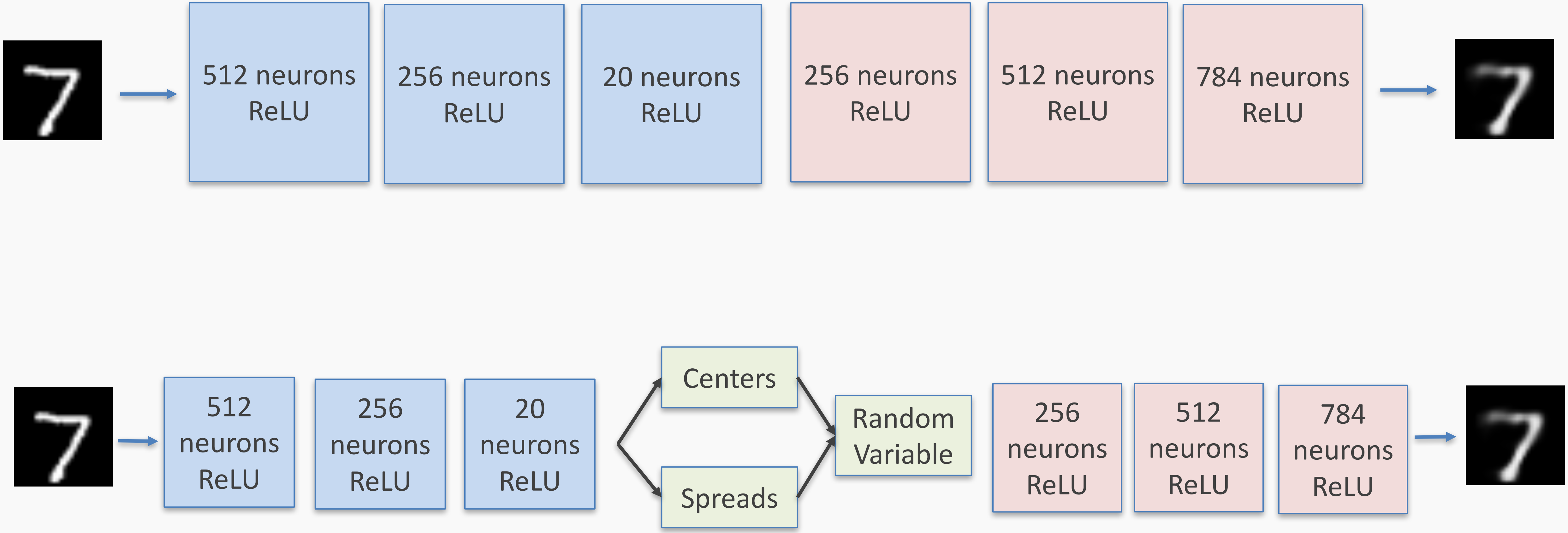
$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & \emptyset \\ & \sigma_2^2 & \\ \emptyset & & \sigma_3 \end{pmatrix}$$

Variational Autoencoders



Variational Autoencoders



Outline

Motivation for Variational Autoencoders (VAE)

Mechanics of VAE

Separability of VAE

Training and the math behind everything

Separability in Variational Autoencoders

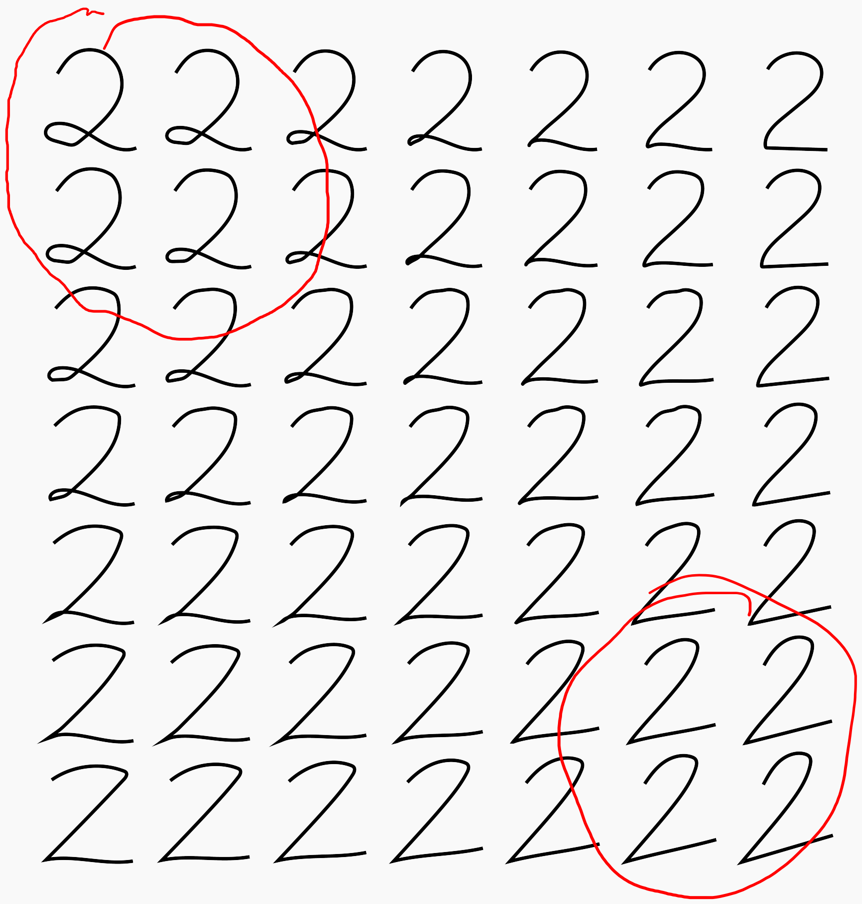
7 7

Separability is not only between classes but we also want similar items in the same class to be near each other.

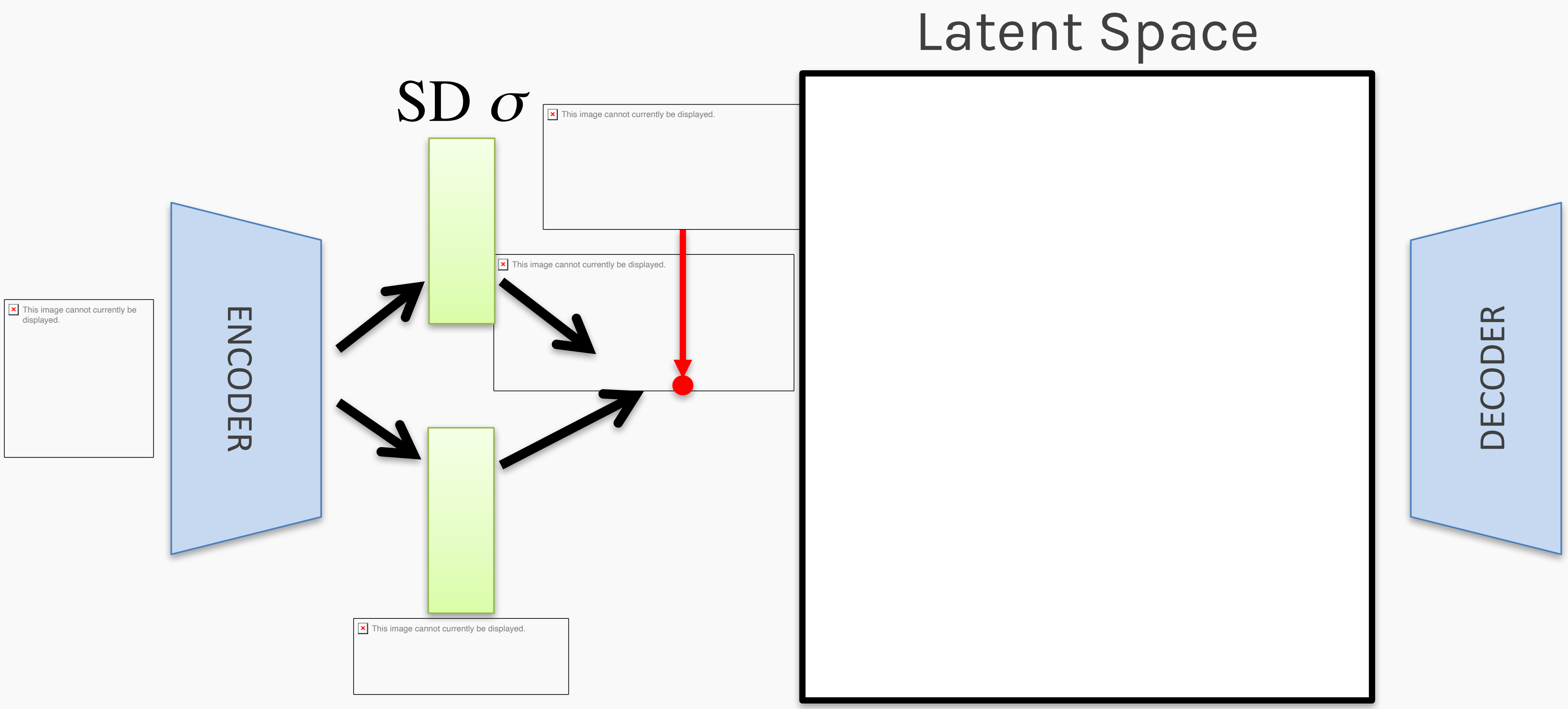
This is similar to word encoding we have talked in the previous lecture.

For example, there are different ways of writing “2”, we want similar styles to end up near each other.

Let’s examine VAE, there is something magic happening once we add stochasticity in the latent space.

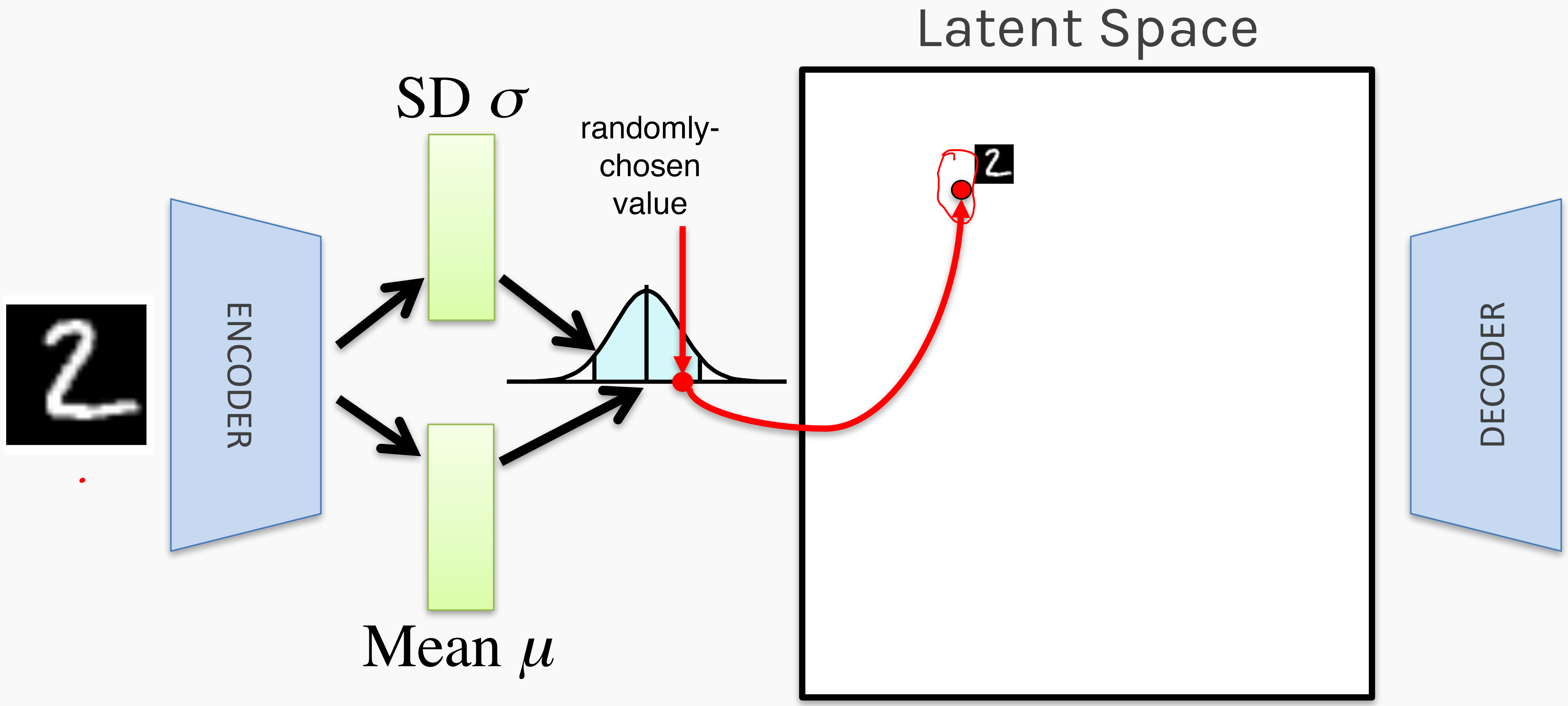


Separability in Variational Autoencoders



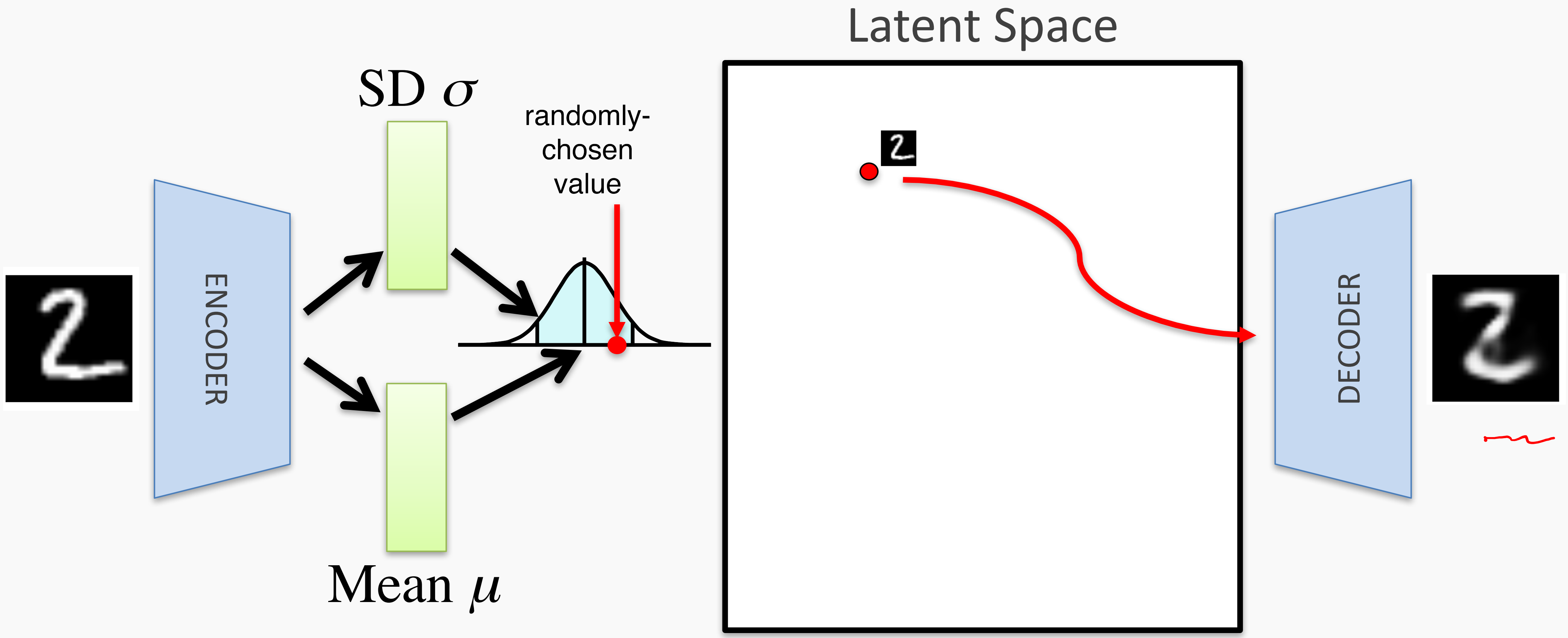
Encode the first sample (a "2") and find μ_1, σ_1

Separability in Variational Autoencoders



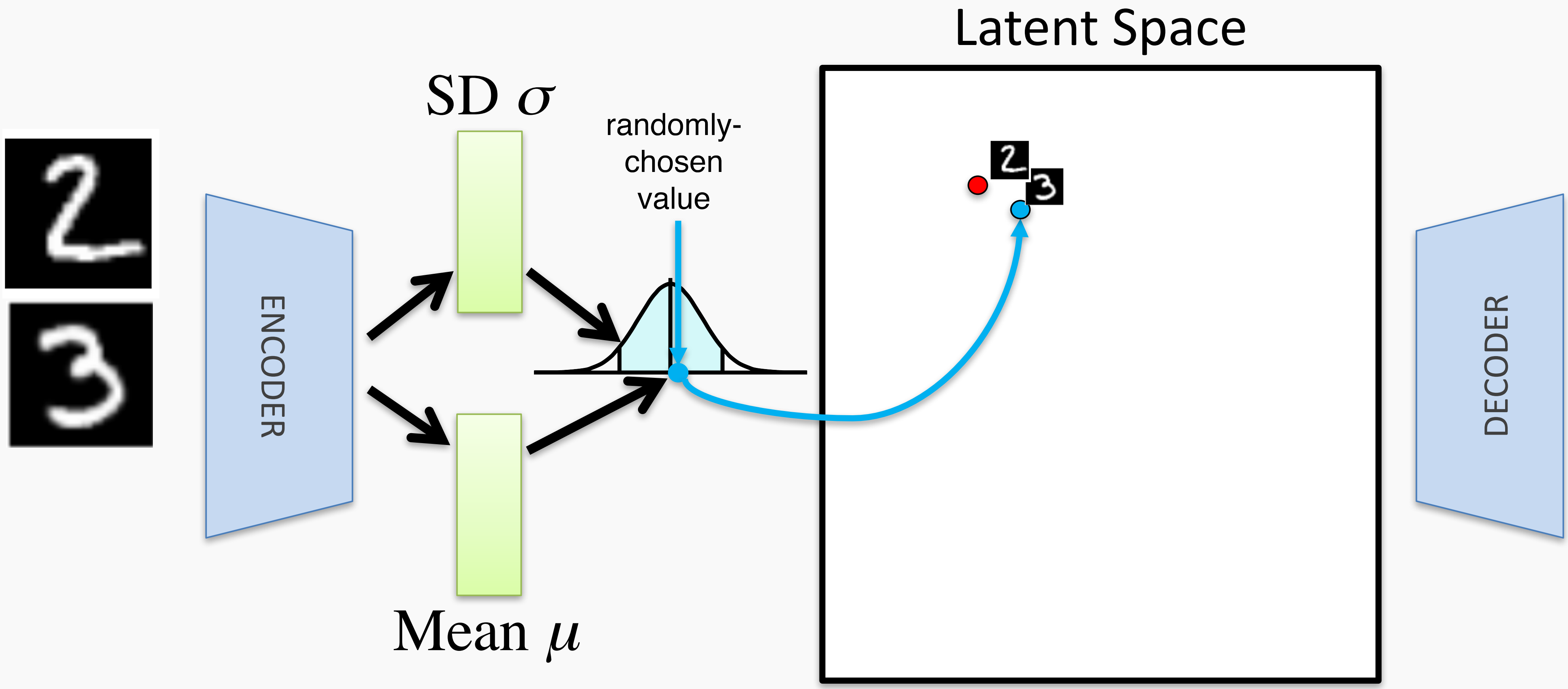
Sample $z_1 \sim N(\mu_1, \sigma_1)$

Blending Latent Variables



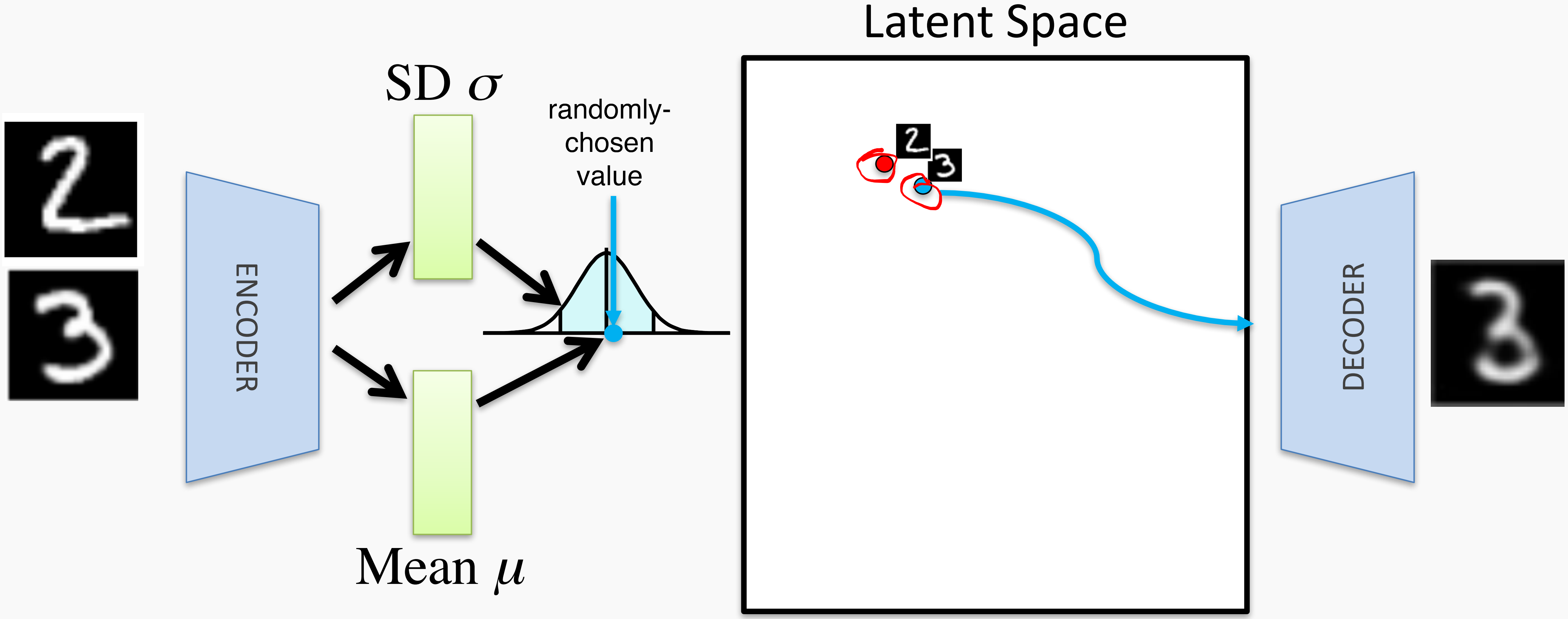
Decode to $\hat{x}_1$

Separability in Variational Autoencoders



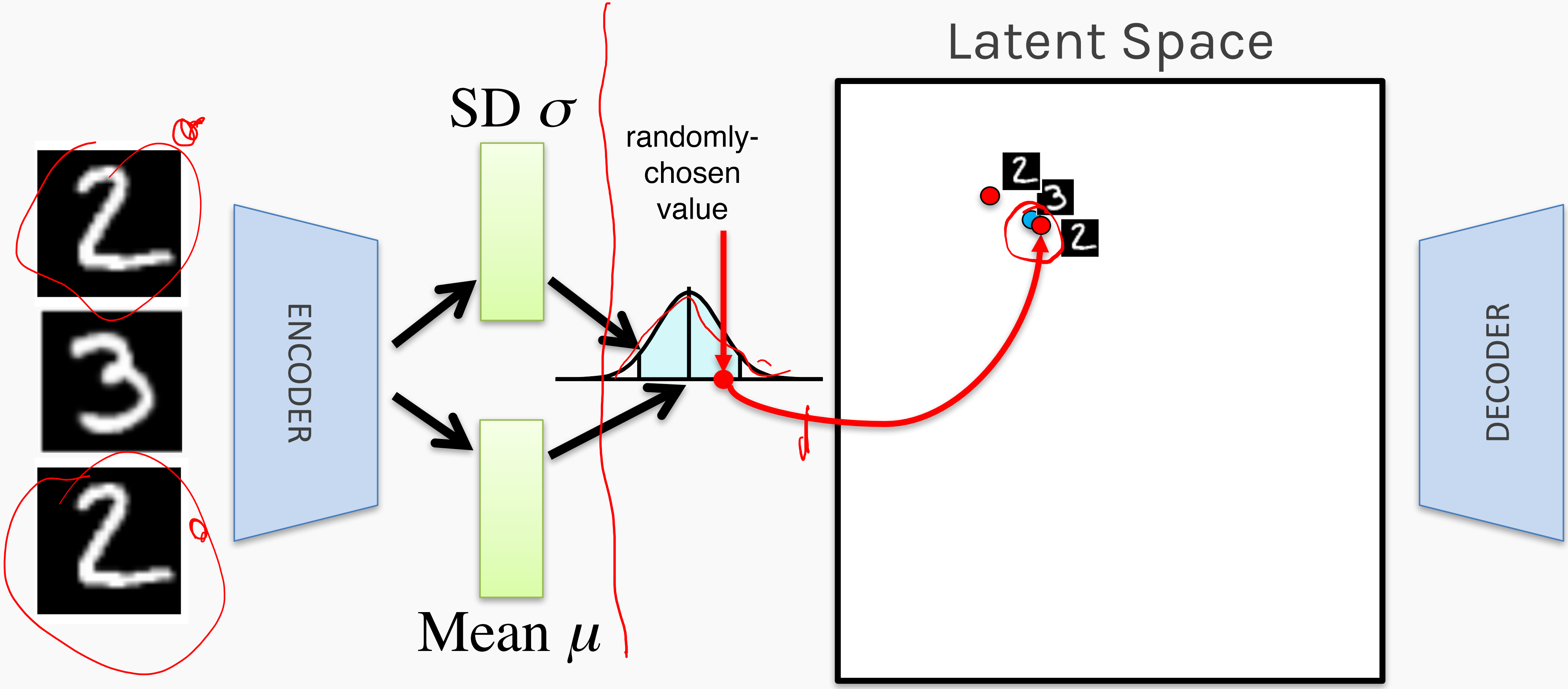
Encode the second sample (a "3") find μ_2, σ_2. Sample $z_2 \sim N(\mu_2, \sigma_2)$

Separability in Variational Autoencoders



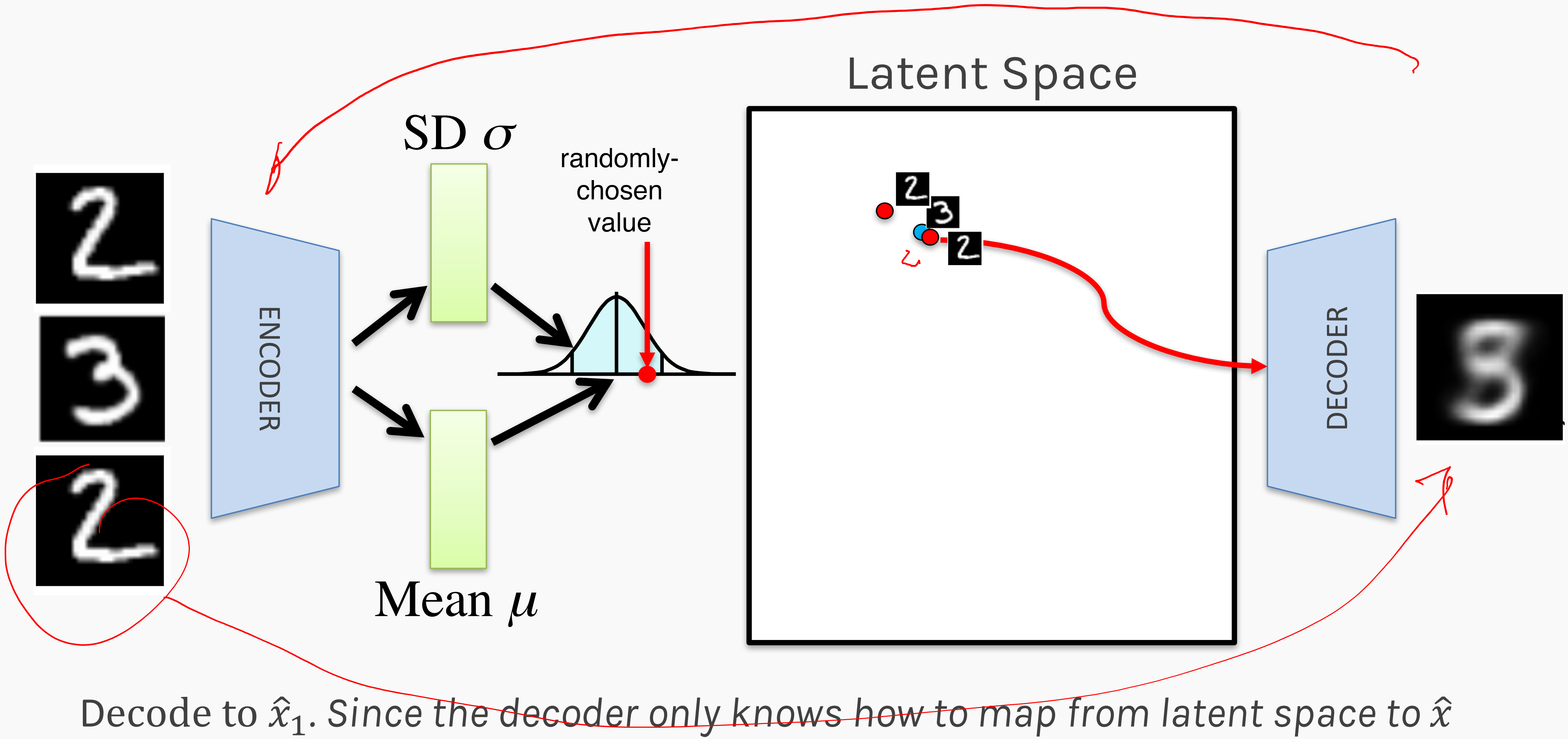
Decode to $\hat{x}_2$

Separability in Variational Autoencoders



Train with the first sample (a “2”) again and find μ_1, σ_1. However $z_1 \sim N(\mu_1, \sigma_1)$ **will not be the same**. *It can happen to be close to the “3” in latent space.*

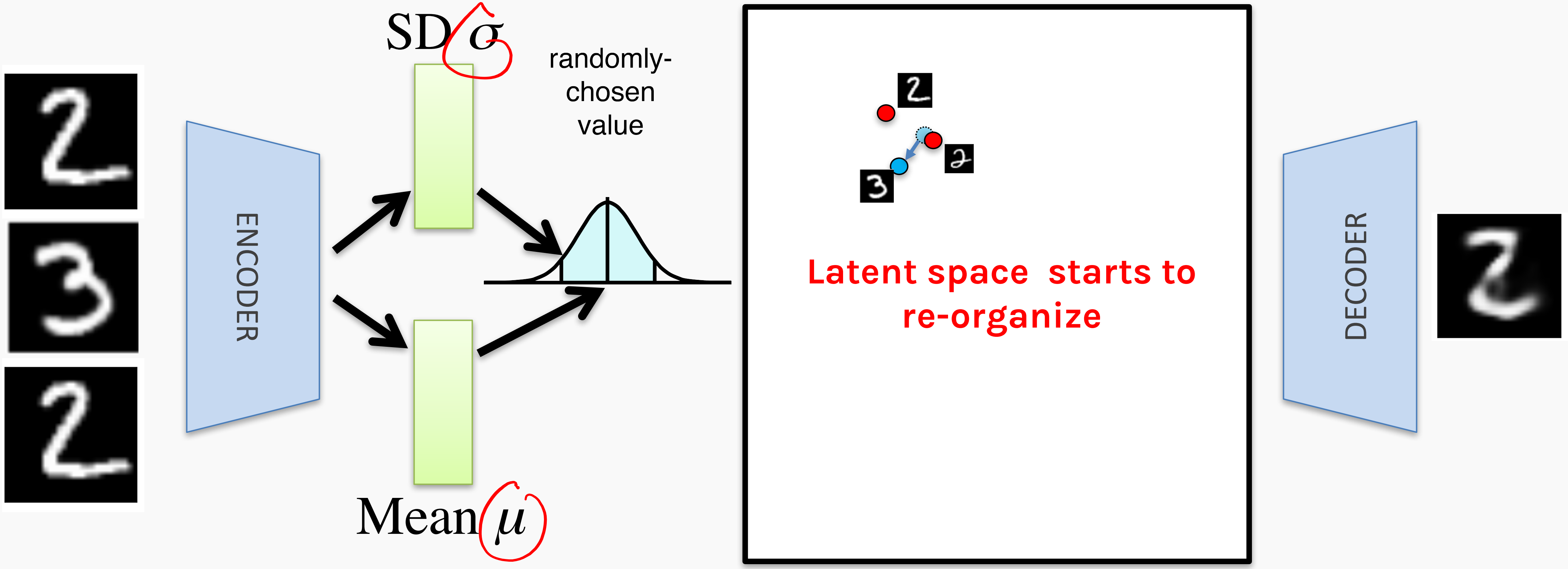
Separability in Variational Autoencoders



Decode to $\hat{x}_1$. *Since the decoder only knows how to map from latent space to $\hat{x}$ space, it will return a "3".*

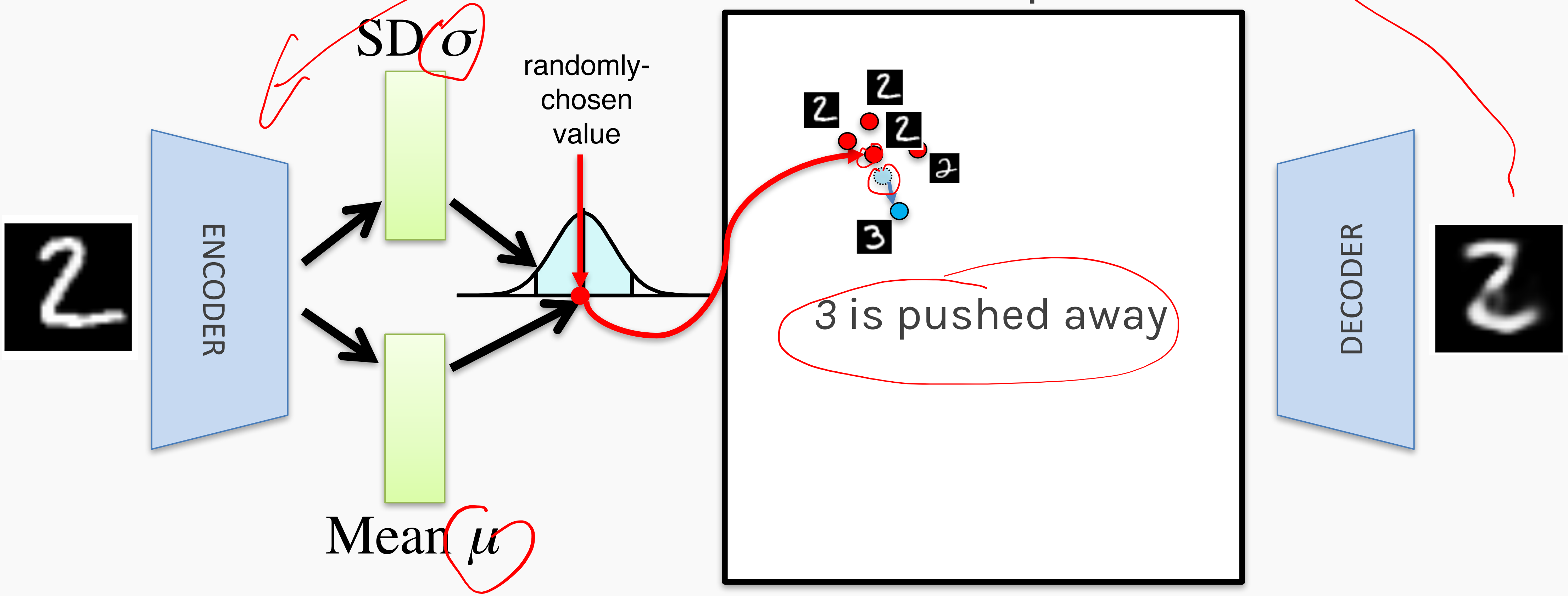
Separability in Variational Autoencoders

Train with 1st sample again



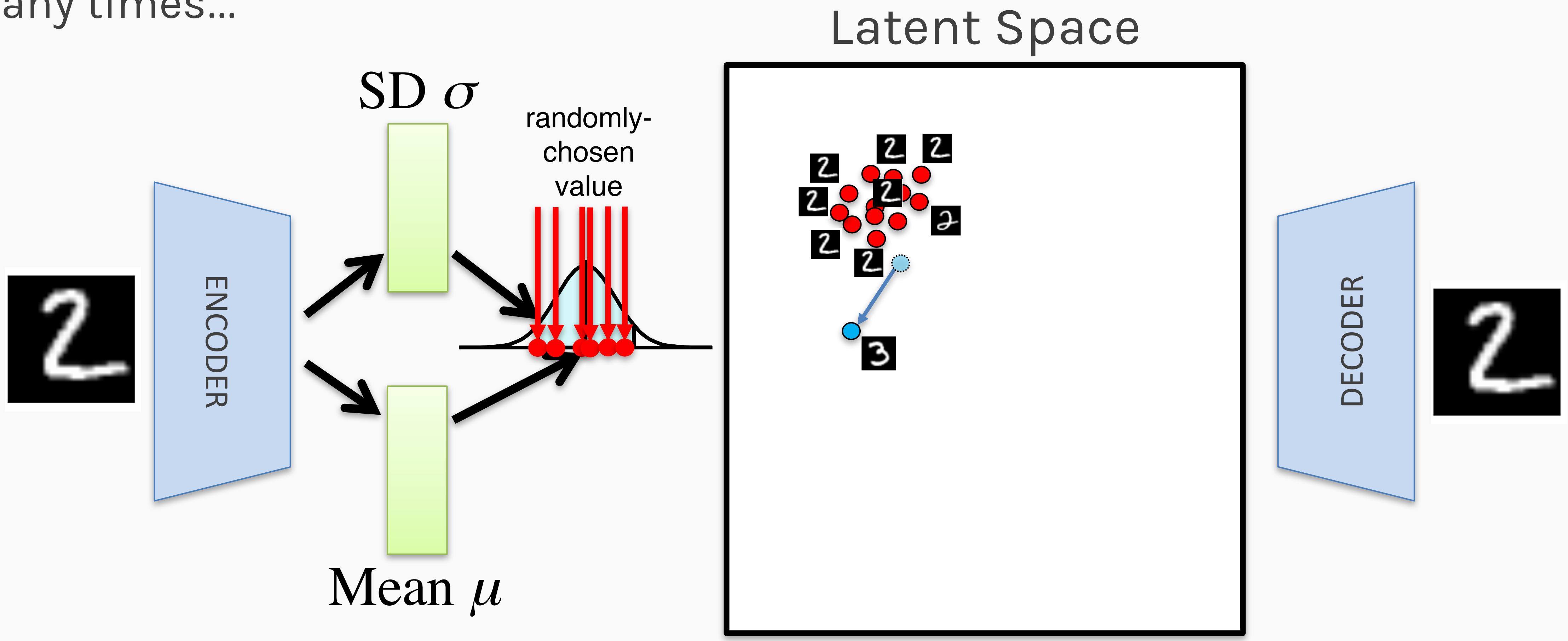
Separability in Variational Autoencoders

And again…



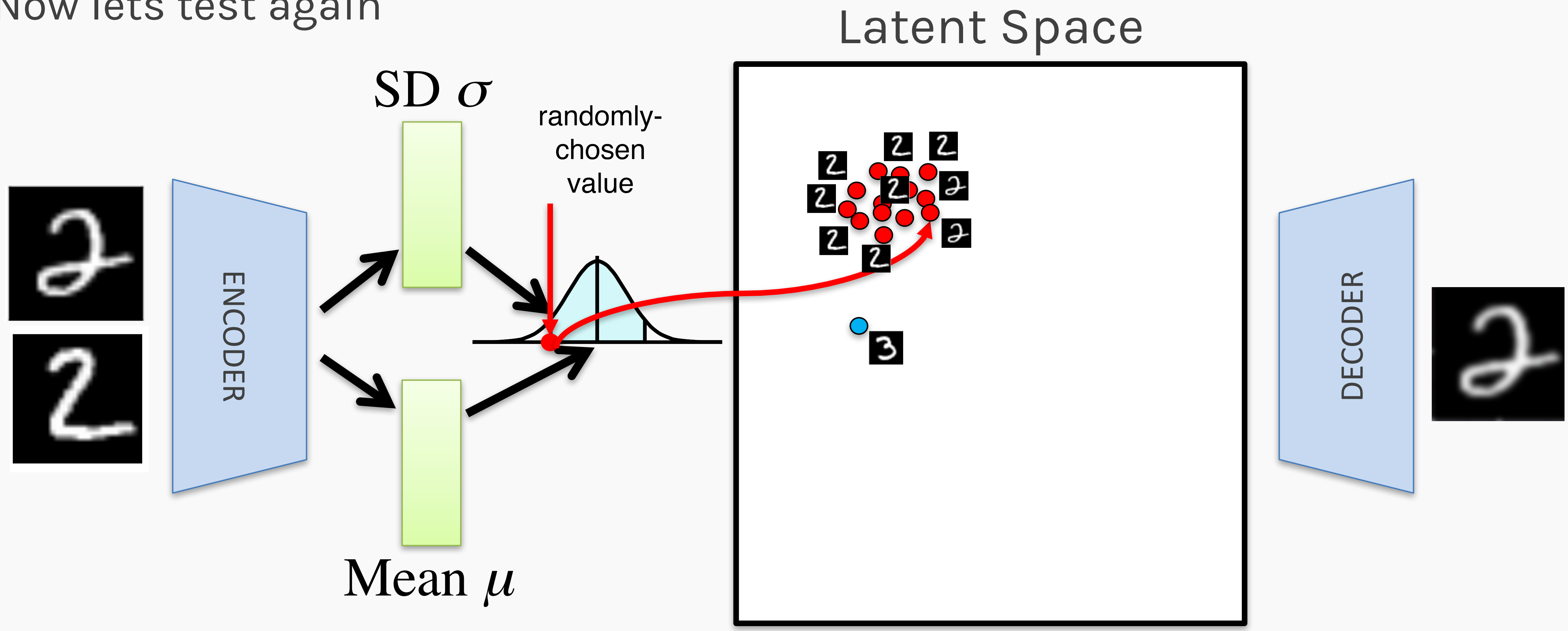
Separability in Variational Autoencoders

Many times…



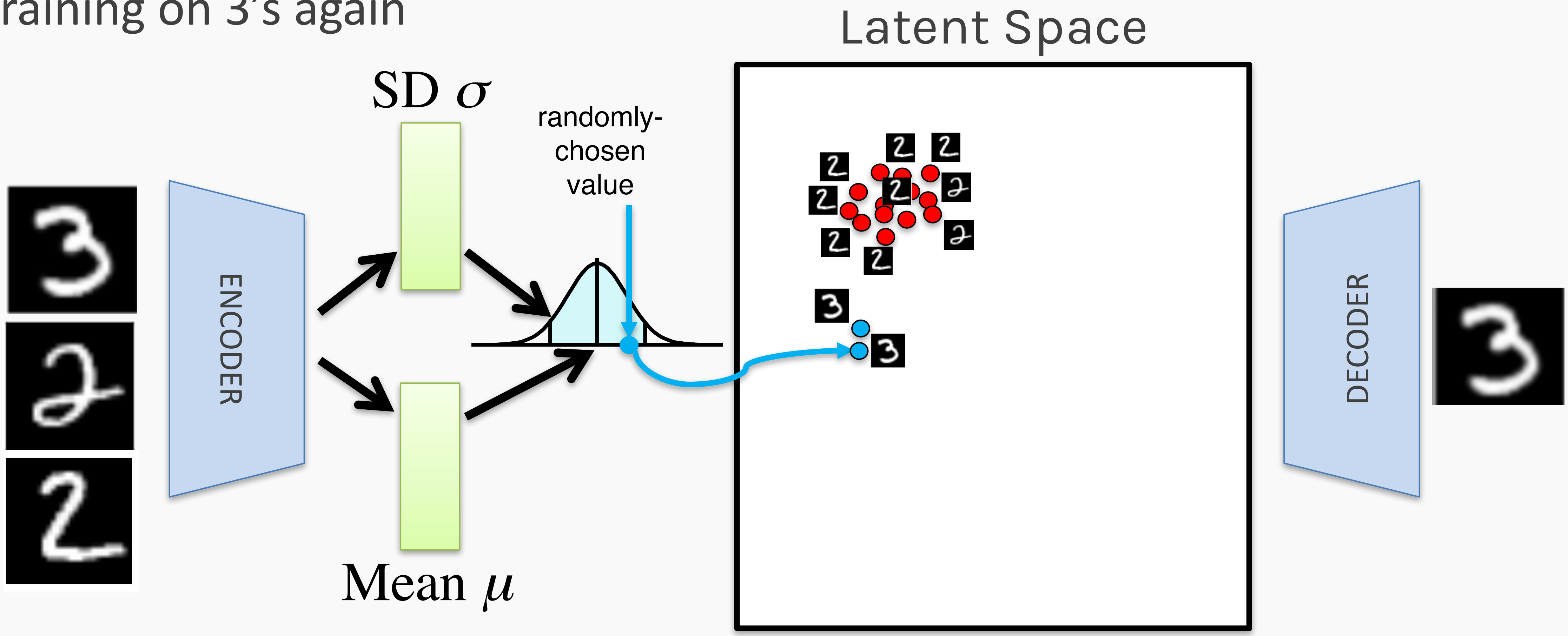
Separability in Variational Autoencoders

Now lets test again



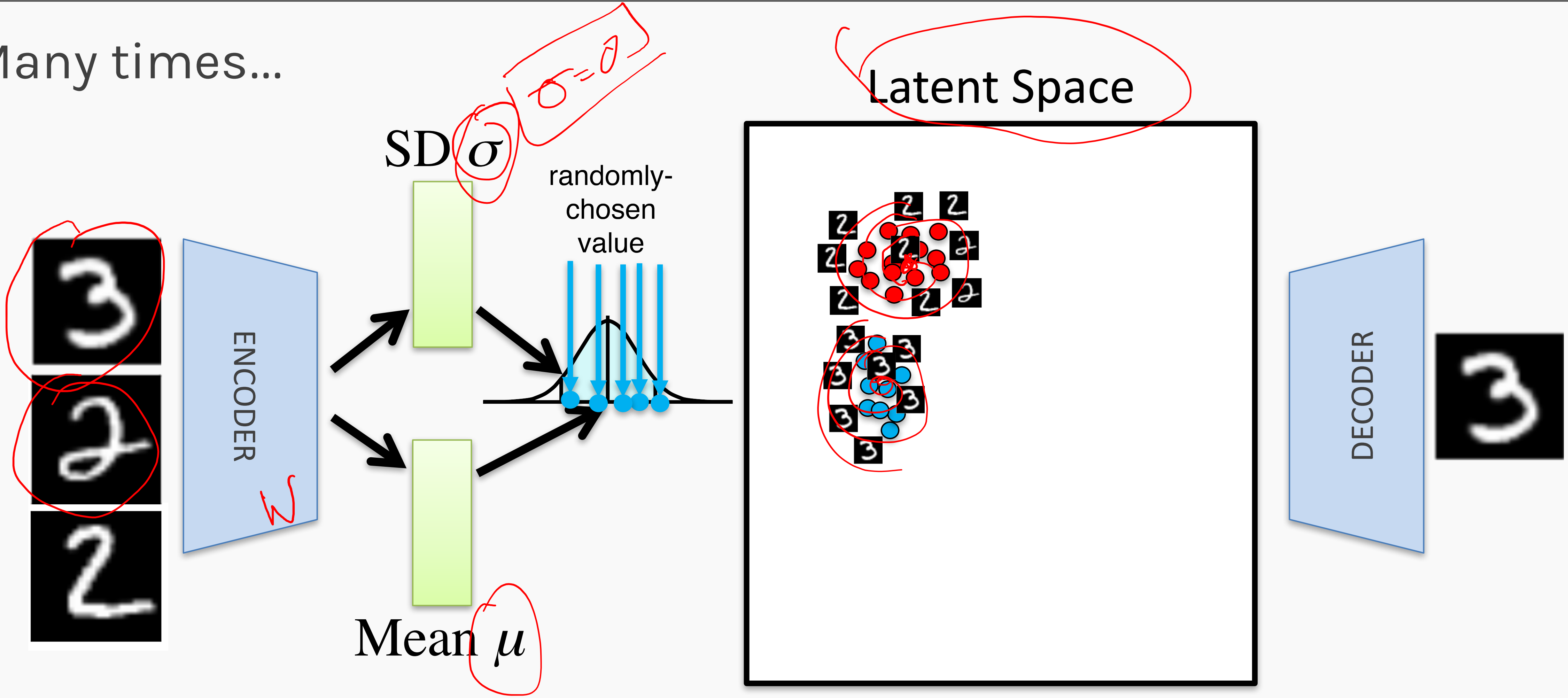
Separability in Variational Autoencoders

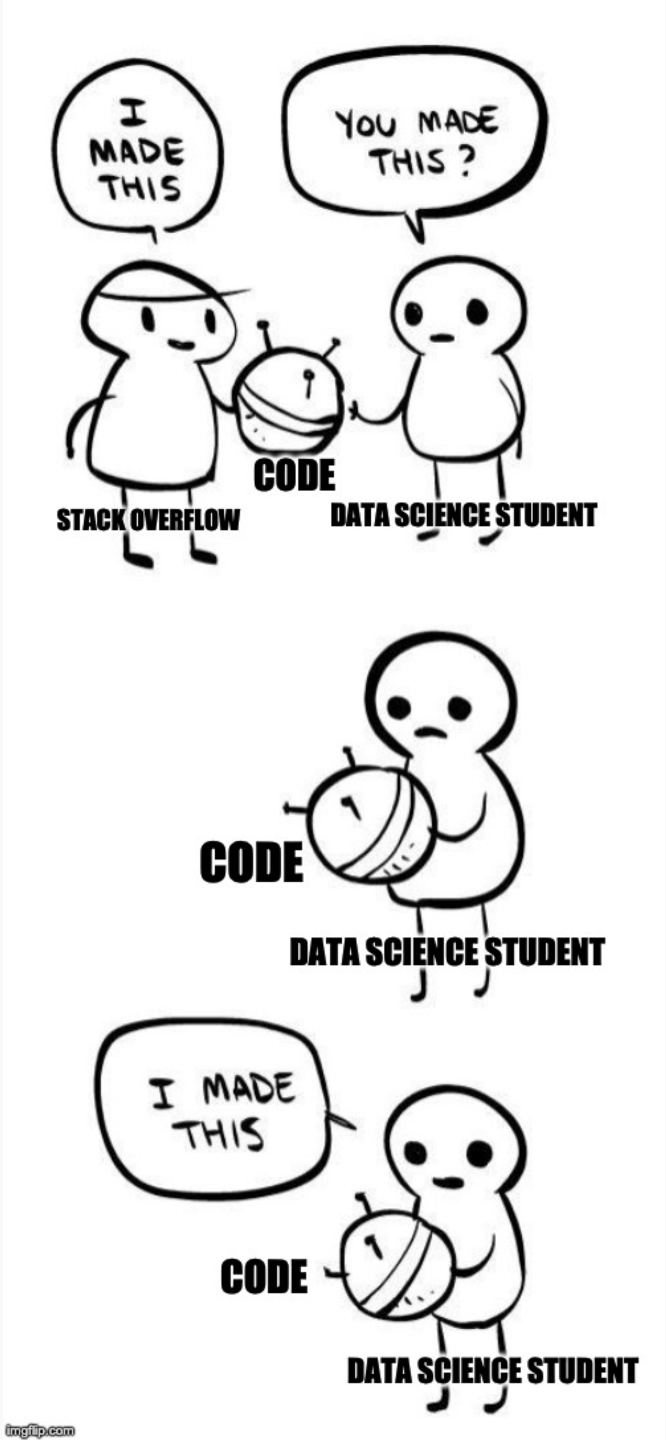
Training on 3's again



Separability in Variational Autoencoders

Many times...




I MADE THIS
YOU MADE THIS ?
CODE
STACK OVERFLOW
DATA SCIENCE STUDENT
CODE
DATA SCIENCE STUDENT
I MADE THIS
CODE
DATA SCIENCE STUDENT
imgflip.com

Outline

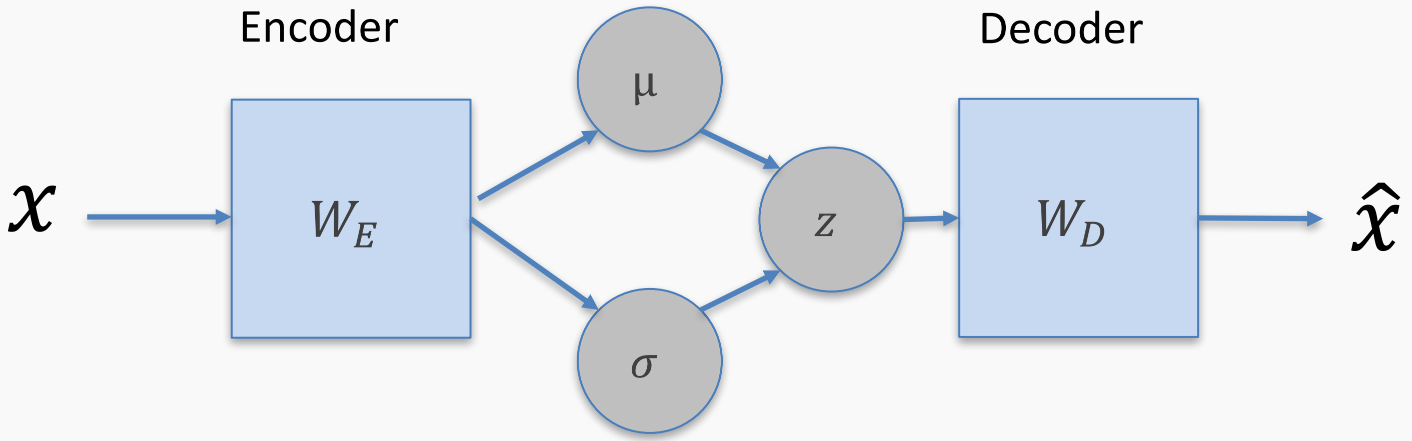
Motivation for Variational Autoencoders (VAE)

Mechanics of VAE

Separability of VAE

Training and the math behind everything

Training



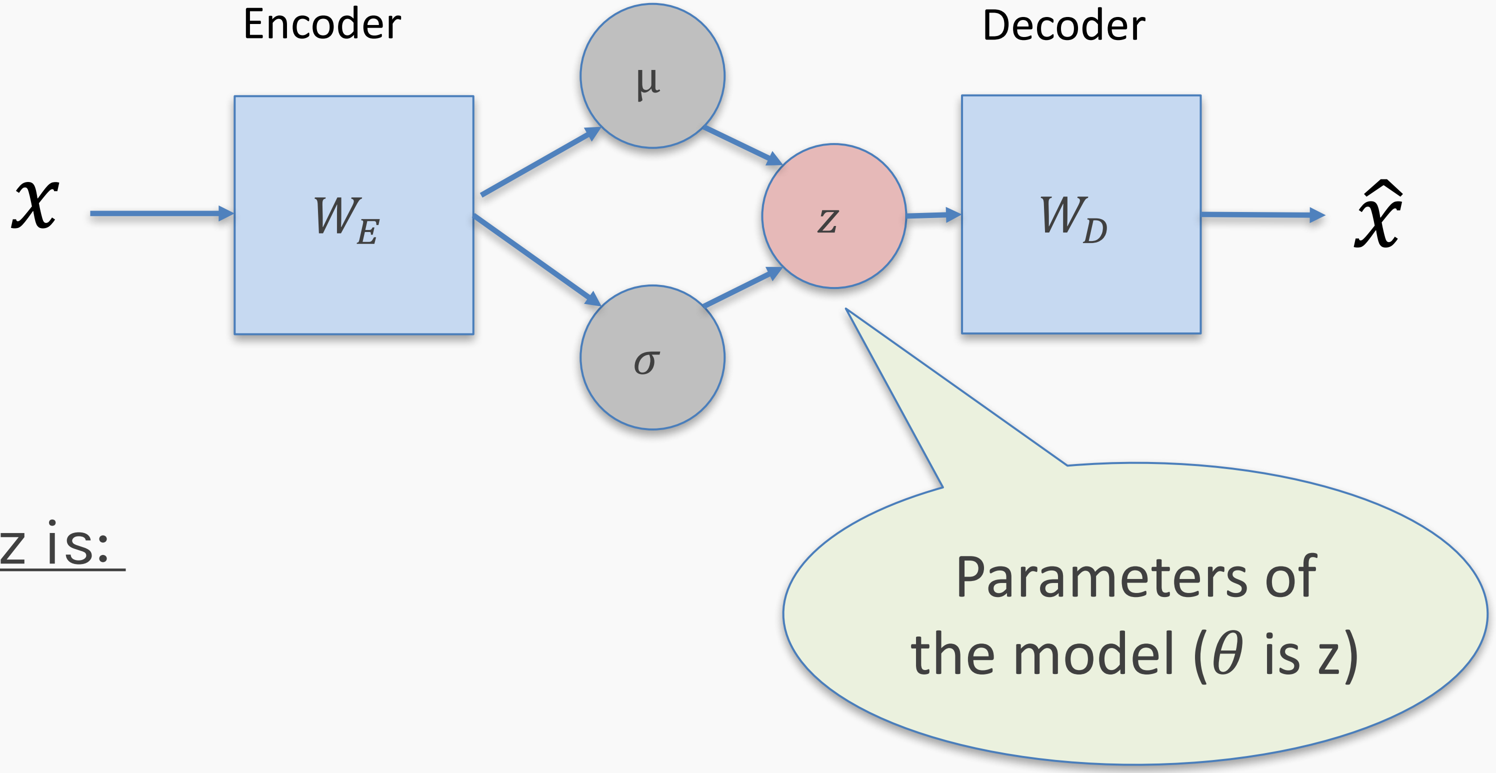
Training means learning W_E and W_E.

- Define a loss function $\mathcal{L}$
- Use stochastic gradient descent (or Adam) to minimize $\mathcal{L}$

The Loss function:

- Reconstruction error: $\mathcal{L}_R = \frac{1}{n}\sum_i (x_i - \hat{x}_i)^2$
- Similarity between the probability of z give x and p$(z|x)$ and some predefine probability distribution p(z), which can be given by Kullback-Leibler divergence (KL): $KL(p(z|x)||p(z))$ [Note: we will get into the details in the a-section]

Bayesian AE



Bayes rule:

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

Posterior for our parameters, z is:

$$p(z|x,\hat{x}) \propto p(\hat{x}|z,x)p(z)$$

Posterior predictive, probability to see $\hat{x}$ given x; **this is INFERENCE**:

$$p(\hat{x}|x) = \int p(\hat{x}|z,x)p(z|x)dz$$

Decoder: NN

Posterior

Bayesian AE

The posterior, $P(z|x,\hat{x})$, can be sampled with MCMC, i.e. no minimization of Loss function. How?

1. Set the priors, $p(z)$
2. Define the likelihood, $P(\hat{x}|z,x)$
3. Propose a new z* and:
   a. check if $P(z^*|x,\hat{x})/P(z|x,\hat{x})$ >1: accept, z^*
   b. If $P(z^*|x,\hat{x})/P(z|x,\hat{x})$ <1 throw a random coin and accept/reject z^*
4. This will converge to true $P(z|x,\hat{x})$!
5. Calculate $P(\hat{x}|x) = \int P(\hat{x}|z,x)P(z|x)dz$ (Note: this is easily done with sample from z and re-weight given the likelihood)

DOABLE!

Variational AE

Problem: z is the dimensionality of your latent space, which can be too large. In other words this $\int p(\hat{x}|z,x)p(z|x)dz$ becomes intractable.

Instead we turn this into a minimization problem – Variational Calculus

Find a q$(z|x)$ that is similar to $p(z|x)$ by minimizing their difference.

After some math:

$$\underbrace{-\mathbf{E}_{z\sim q_\phi(z|x)} \log\left(p_\theta\left(x|z\right)\right)}_{\text{Reconstruction Loss}} + \underbrace{KL\left(q_\phi\left(z|x\right)\|p_\theta(z)\right)}_{\text{Proposal distribution should resemble a Gaussian}}$$

```python
def make_VAE(input_dim, latent_dim, bottleneck_dim, reg=0.01, dropout_rate=0.0):
    flat_dim = latent_dim[0] * latent_dim[1] * latent_dim[2]

    x = layers.Input(shape=input_dim, batch_size=batch_size)
    xe = ConvEncoder(input_shape=input_dim, dropout_rate=dropout_rate)(x)
    xe = layers.Flatten()(xe)
    z_mean = layers.Dense(bottleneck_dim[0], activation='linear')(xe)
    z_log_var = layers.Dense(bottleneck_dim[0], activation='linear')(xe)
    z = Sampling()([z_mean, z_log_var])
    xr = layers.Dense(flat_dim, activation='relu')(z)
    xr = layers.Reshape(latent_dim)(xr)
    xr = ConvDecoder(input_shape=latent_dim, dropout_rate=dropout_rate)(xr)

    encoder = models.Model(inputs=x, outputs=z)
    VAE = models.Model(inputs=x, outputs=xr)

    if reg > 0.0:
        kl_loss = - 0.5 * tf.reduce_mean(z_log_var - tf.square(z_mean) - tf.exp(z_log_var) + 1)
        VAE.add_loss(reg * kl_loss)

    opt = optimizers.Adam(learning_rate=1e-4)
    loss = losses.MeanSquaredError()

    VAE.compile(optimizer=opt, loss=loss)
    VAE.summary()

    return VAE, encoder
```

μ, σ

$z \sim N(\mu, \sigma)$

$L = L_R + \lambda L_{KL}$

Training VAE

Apply stochastic gradient descent (SGD)

Problem:

Sampling step not differentiable

Use a re-parameterization trick

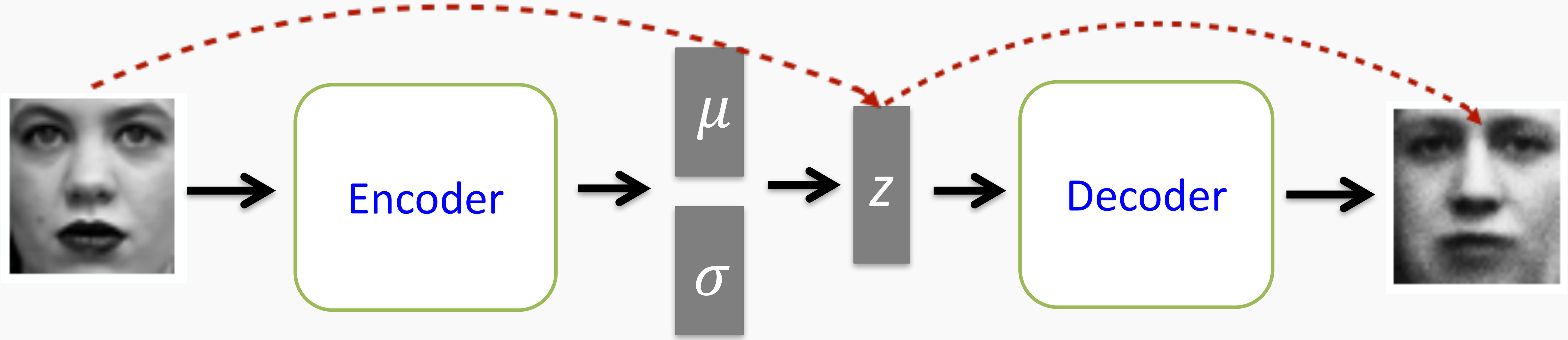
- Move sampling to input layer, so that the sampling step is independent of the model

$x - \boxed{w_E} \; \mu, \sigma \to z - \boxed{} \to \hat{x} \to L(x, \hat{x}, \theta)$

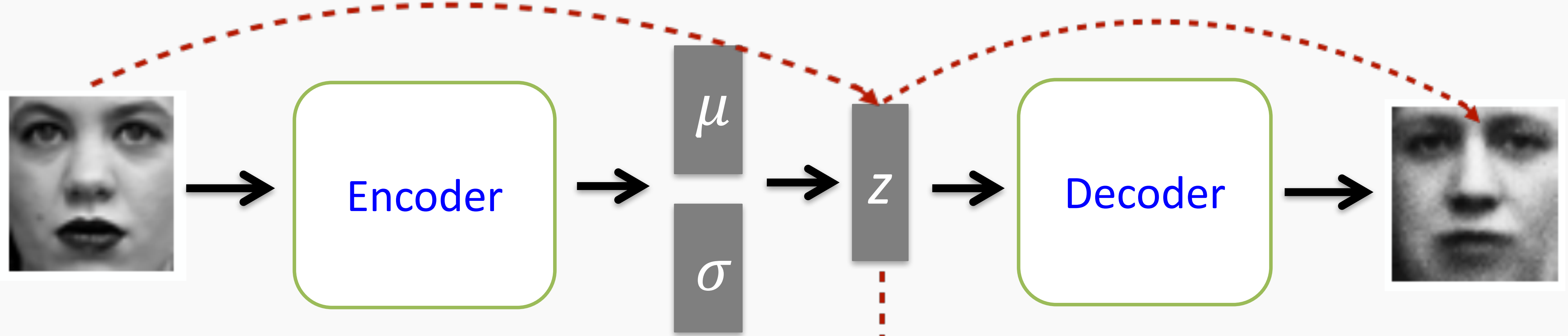
$z \sim N(\mu, \sigma)$

$\boxed{\frac{\partial z}{\partial w_E}}$?

Reparametrization Trick



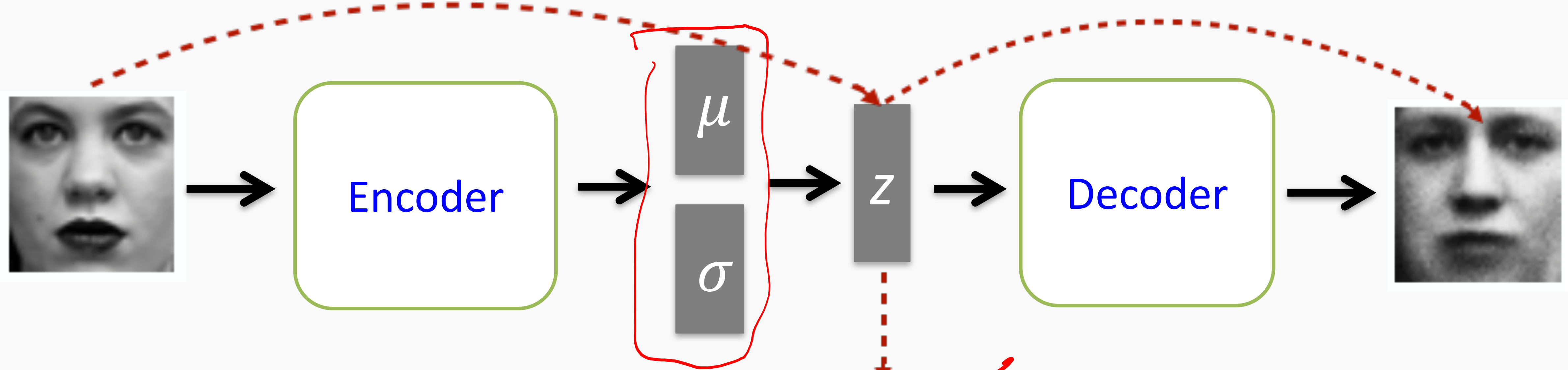
Reparametrization Trick



$$Z = \mu + \varepsilon \circ \sigma$$

1

Reparametrization Trick



$z \sim N(\mu, \sigma)$

$$Z = \mu + \varepsilon \circ \sigma$$

$$\varepsilon \sim N(0, I)$$

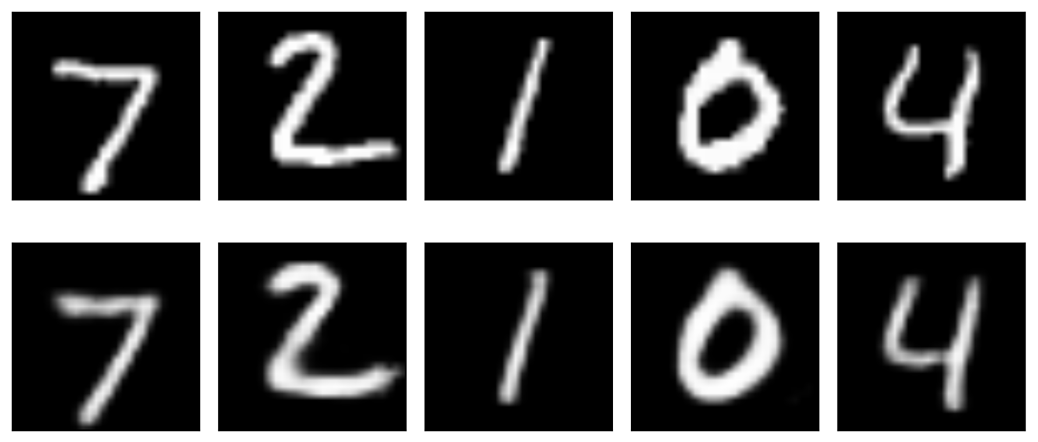
$\frac{\partial \mu}{\partial W_E}$, $\frac{\partial \sigma}{\partial W_E}$

Training VAE

Traditional AE:

Input Image:

Output Images:

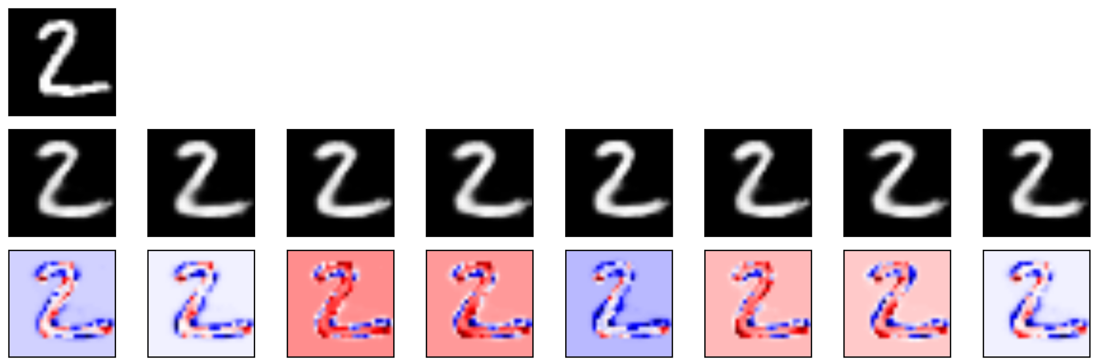


Variational AE:

Input Image:

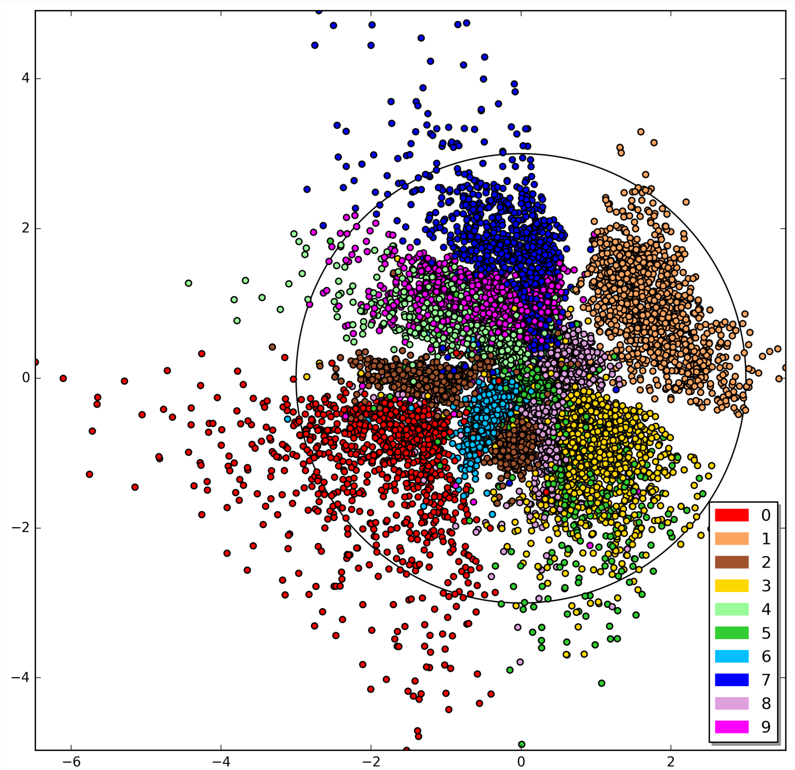
Output Images:

Difference:



Latent space of VAE

- More separable than AE
- Because of the prior N(0,1) everything is center at (0,0) with spread of approx 1.



- Blending is more continuous because latent space is continuous