Variational Autoencoders

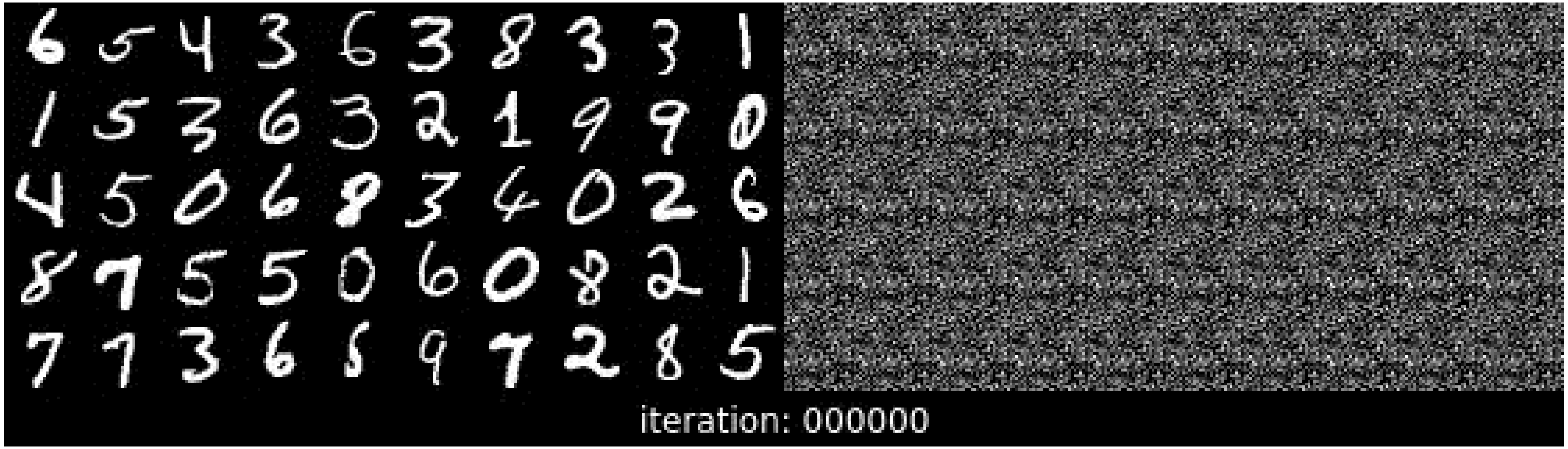
Shangsong Liang

Sun Yat-sen University

Orignially produced by Alon Oring

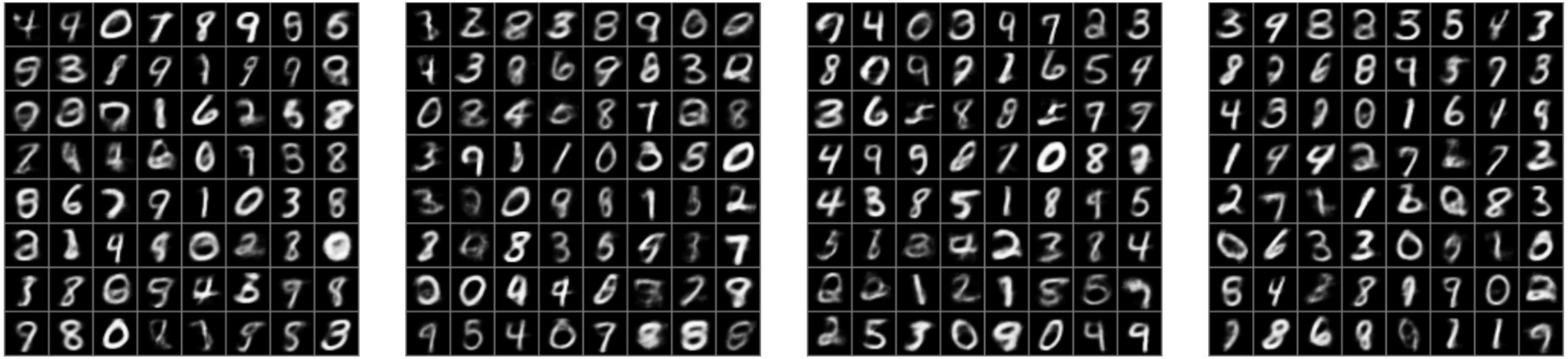
Recap - Autoencoders

- Traditional AE are models designed to output a reconstruction of their input by deconstructing input data into hidden representations and reconstructing them into the original input
- The appeal of this setup is that the model learns its own definition of a salient representation based only on data – no labels or heuristics



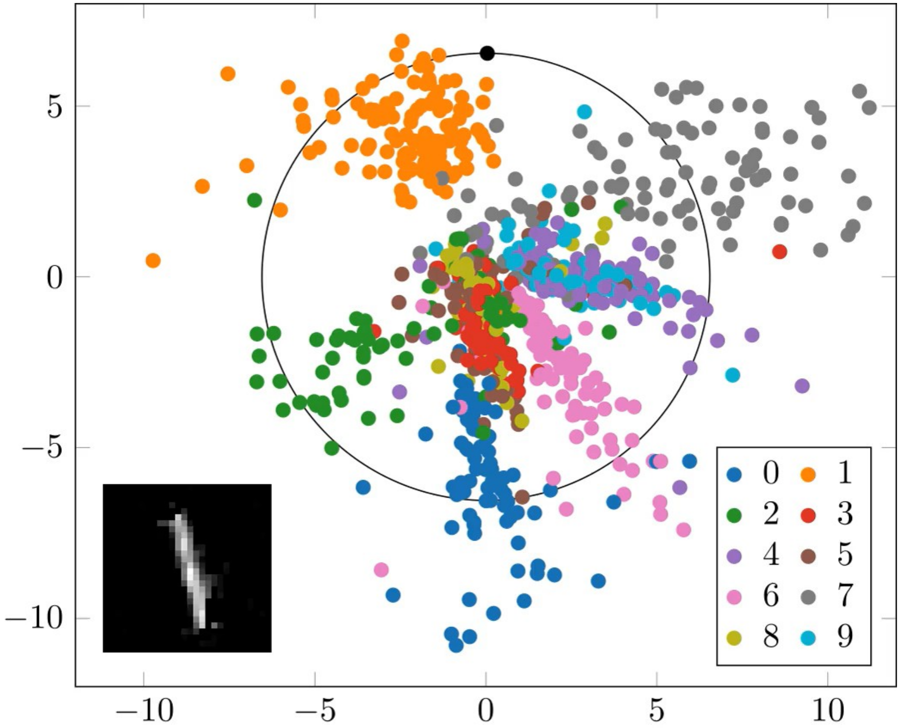
Variational Autoencoders

- A probabilistic twist on autoencoders that enables:
  - Novel image synthesis from random samples
  - Transition from image to image or from mode to mode
  - Aggregation of similar images to close locations in the latent space

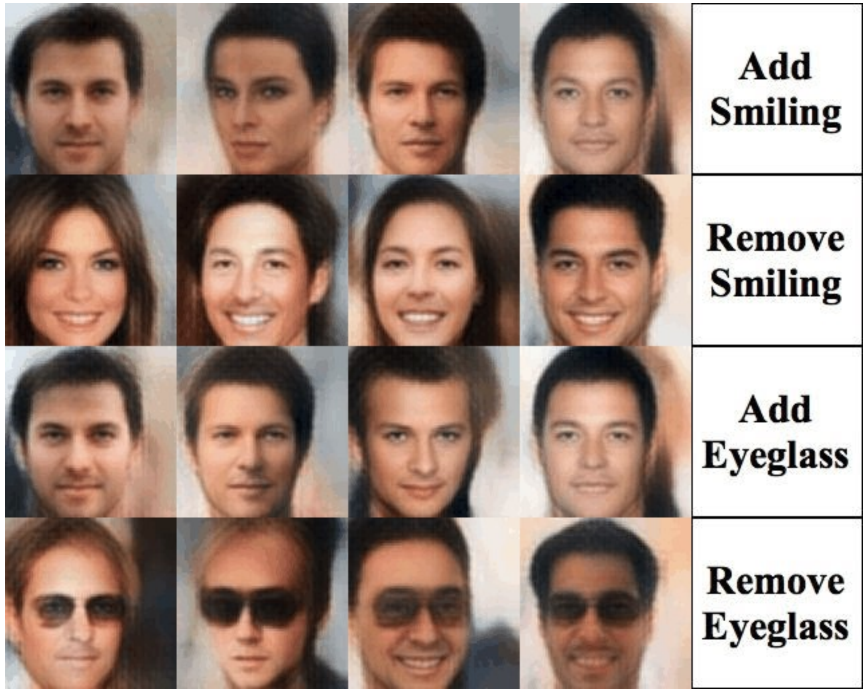


Motivation

- Variational Autoencoders are a deep learning technique for learning useful latent representations



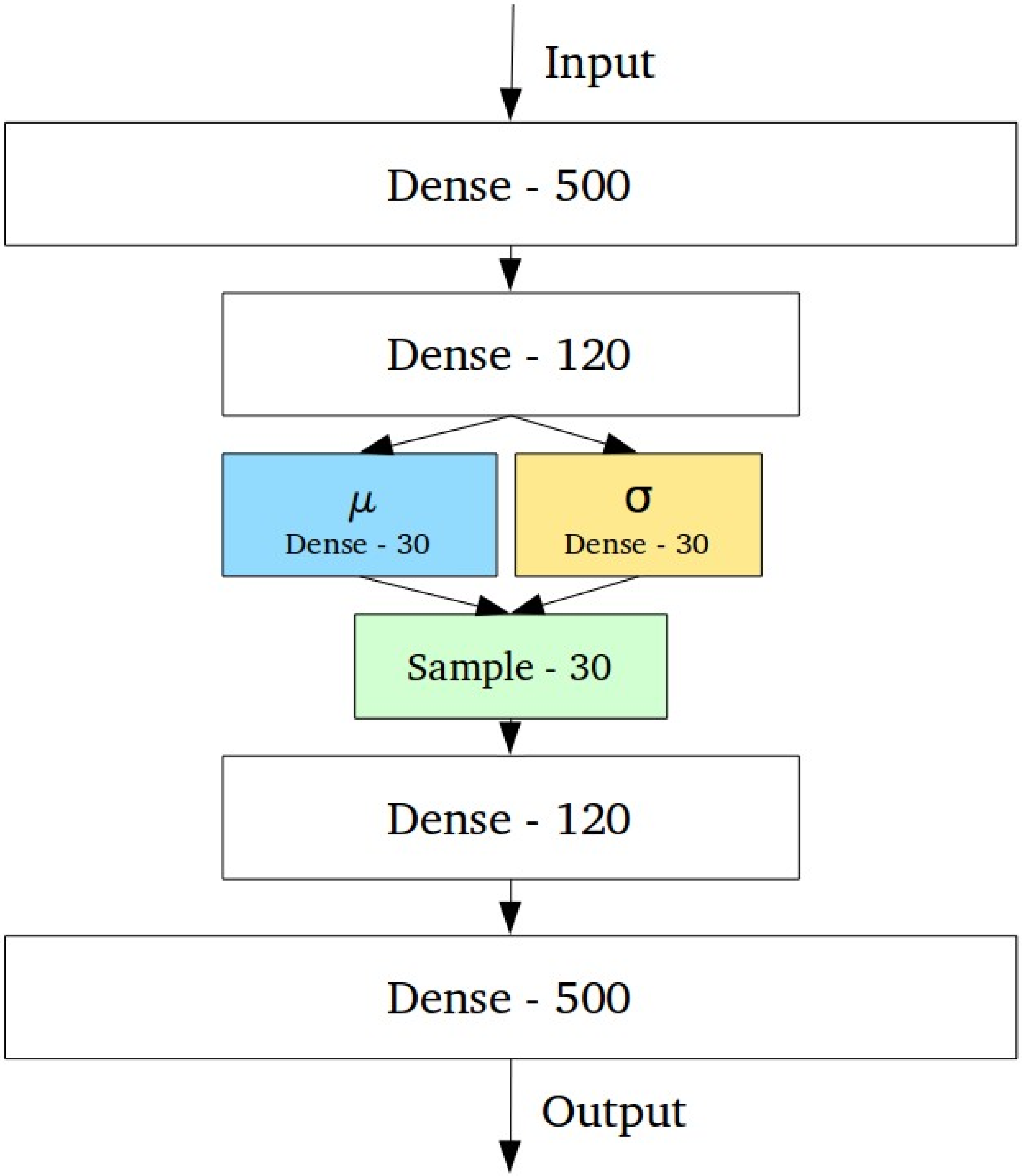
Regular Autoencoder on MNIST dataset



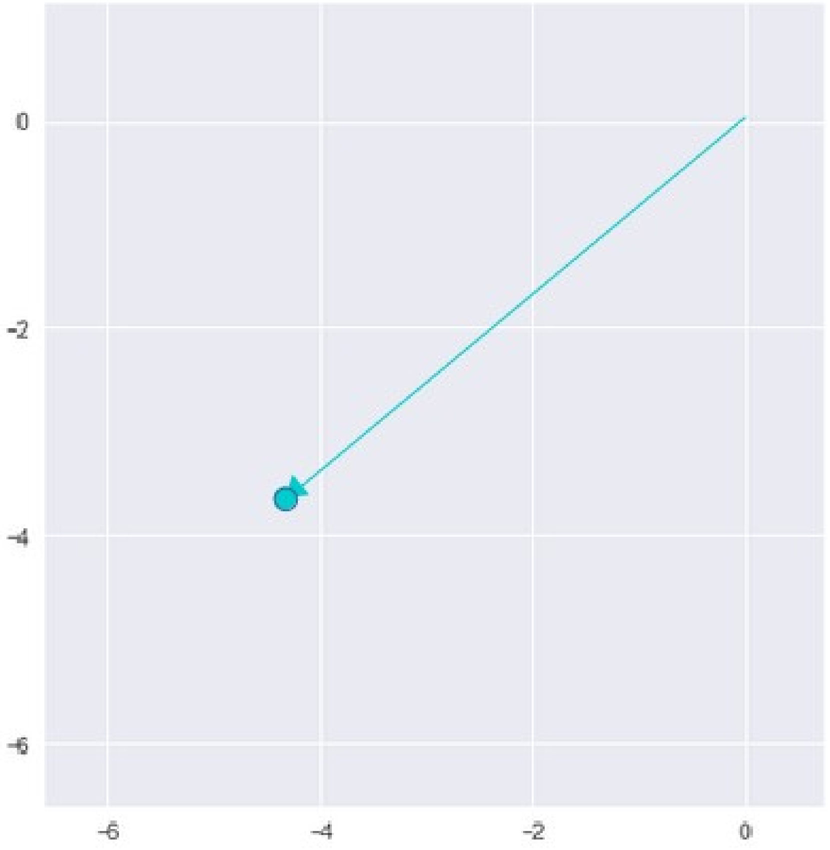
Variational Autoencoder on CelebA dataset

Architecture

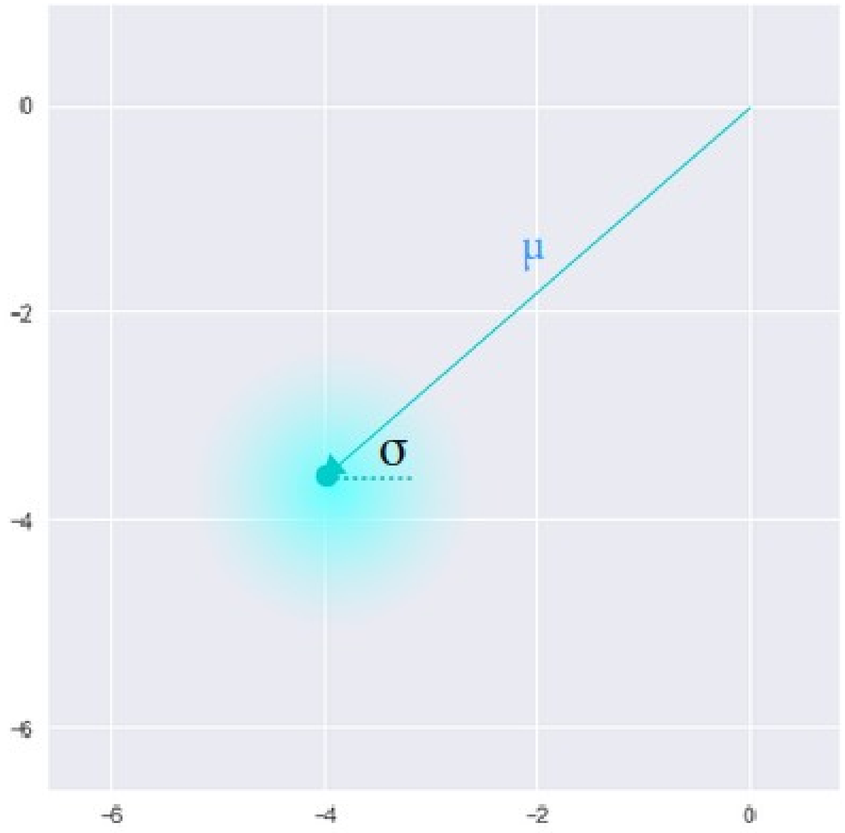
- Very similar to the regular autoencoder
- The probabilistic nature of the VAE is enabled using a sampling layer



VAE – Probabilistic Intuition

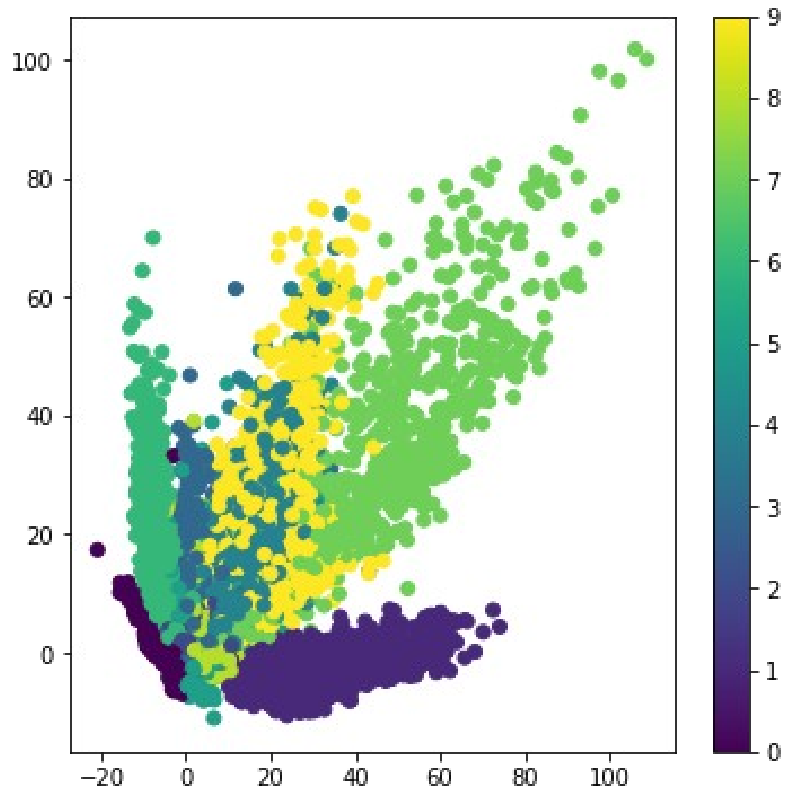


Standard Autoencoder
(direct encoding coordinates)

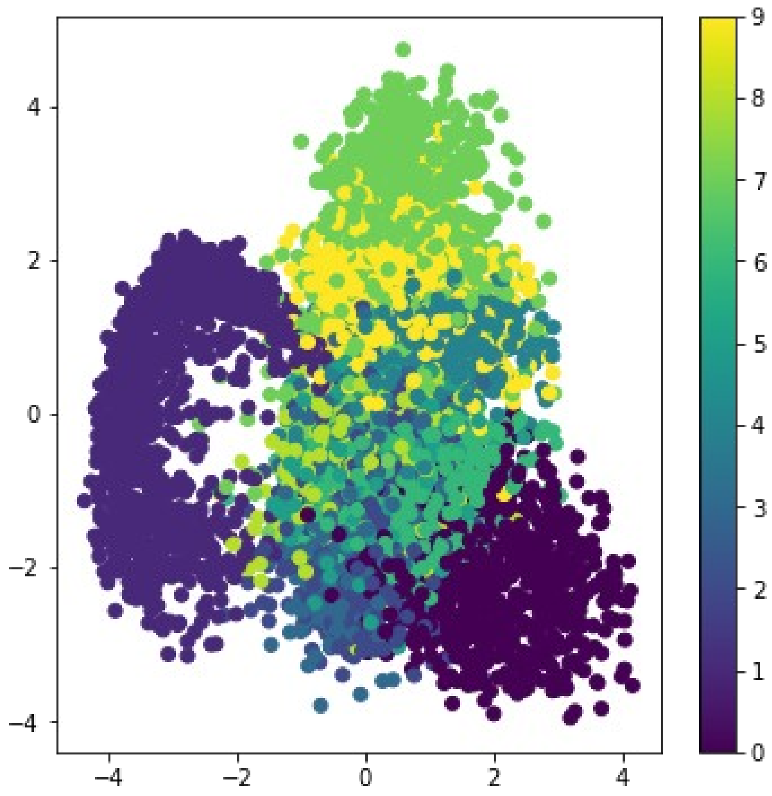


Variational Autoencoder
(μ and σ initialize a probability distribution)

Latent Space Representation



Autoencoder



Variational Autoencoder

Information Theory Recap

Information – Definition and Intuition

- Lets define an information function, , in terms of an event with probability . What should be its properties?

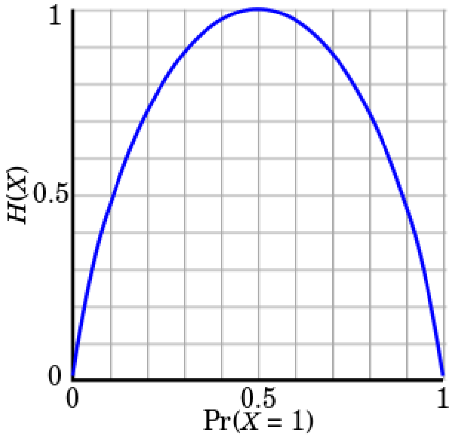
  - is monotonically decreasing – an increase in the probability decreases the information from an observed event

  - - Information due to independent events is additive

- We can guess entropy is

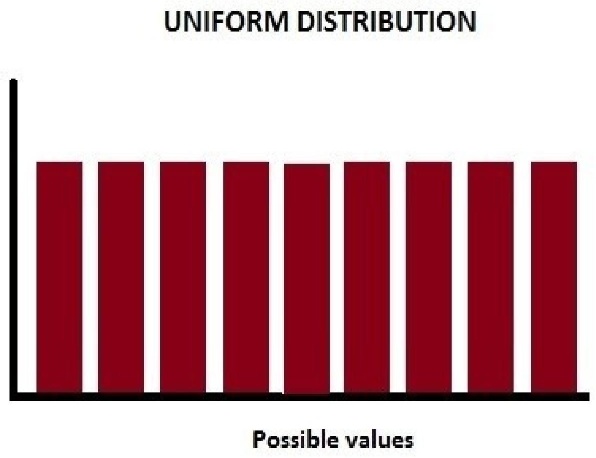
Entropy – Definition

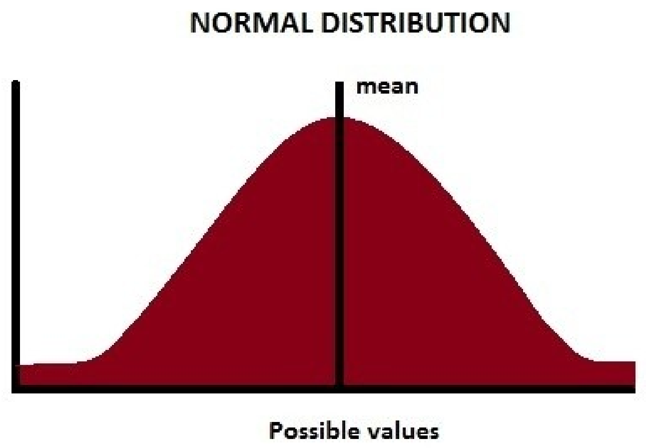
- The entropy is defined as the expected value of the information of a random variable:



Entropy - Intuition

- The maximum entropy is the one that corresponds to the least amount of knowledge defined by the probability density function
- When can we expect maximum entropy for the following cases:
  - Among probability distributions over a finite range of values?
  - Among probability distributions over a infinite range of values?





Kullback–Leibler Divergence

- Let’s define a measure of similarity between distributions and
- How about:

- However, we take the expectation with respect to and obtain

Kullback–Leibler Divergence Properties

- KL measure is a divergence, not a distance
- A condition of a measure to be a metric is to be symmetric

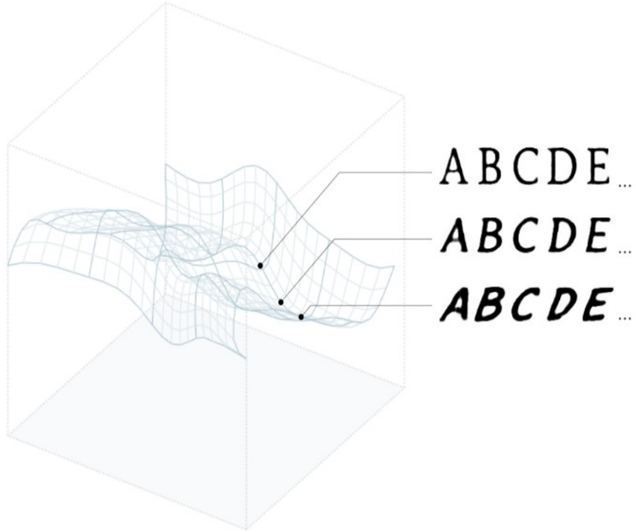
Information Theory - Summary

- The entropy of a distribution gives the **minimum** number of bits per message that would be needed on average to losslessly encode events drawn from
- The cross entropy is the **total** number of bits per message needed to encode events drawn from true distribution if using an optimal code for
- KL Divergence measures the **average** number of extra bits per message

Information Theory Perspective

Latent Variable Models – General Case

- We can only see our **training data,**
- We assume the data is governed by some **unobserved random variable,**
- The image generation process consists of two steps: a value of is generated from some **prior distribution** , and an image is generated from a **conditional distribution**
- We can now obtain two important expressions:

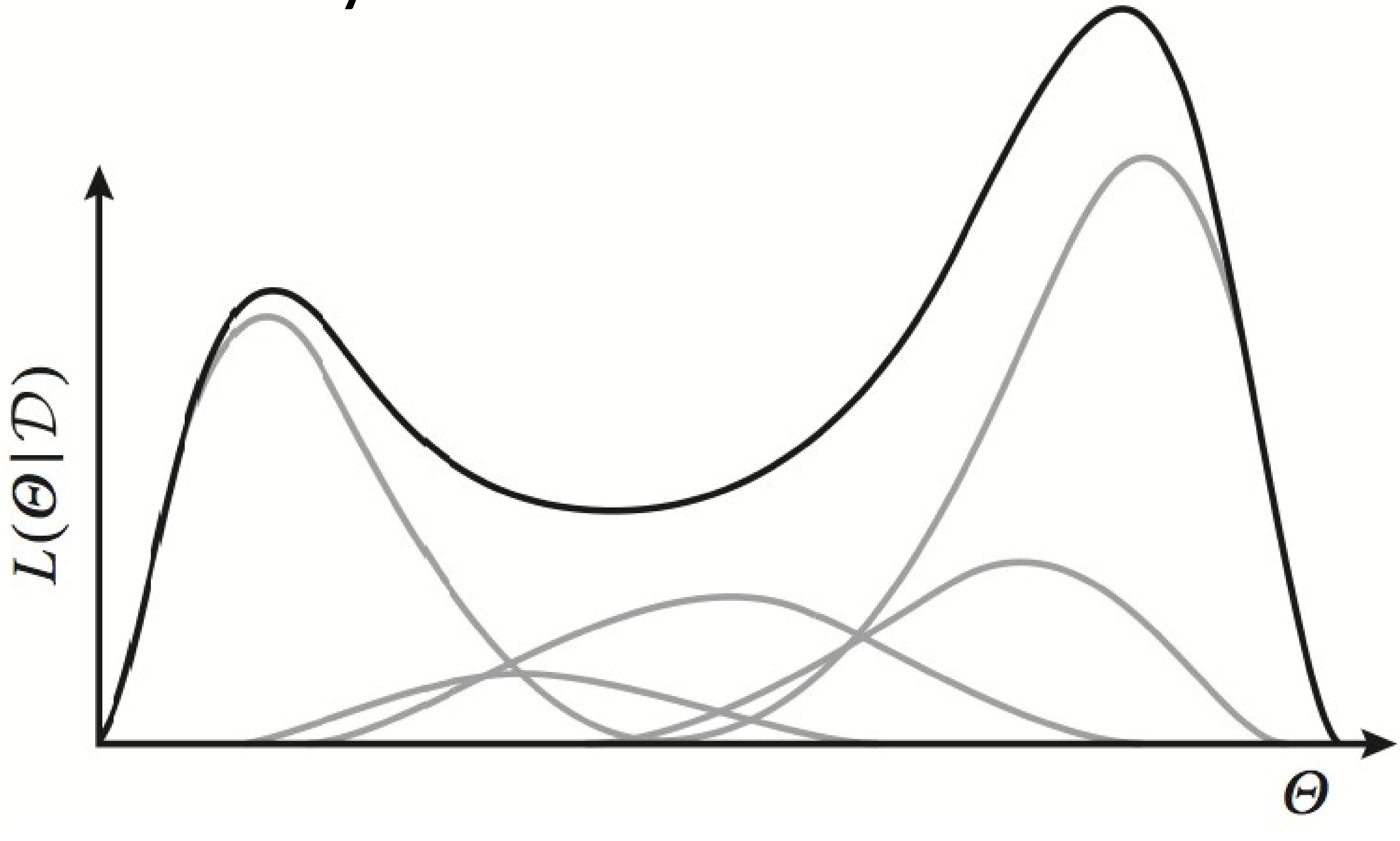


Learning Statistical Models

- Using expectation-maximization algorithm we can iteratively find a maximum likelihood or maximum a posteriori estimates of the parameters in our statistical model and we are done:

- For many models, this evidence integral is unavailable in closed form or requires exponential time to compute. The evidence is what we need to compute the conditional posterior using Bayes

Intractability - Intuition



Intractability

- We can "solve" the intractable part in two ways
  - Variational Inference
  - Markov Chain Monte Carlo (Does not scale well with large datasets)

Variational Inference

Variational Inference

- Suppose we are given an intractable probability distribution
- We can approximate the intractable distribution using some other tractable distribution,
- What will help us choose a distribution that will best approximate the intractable posterior ?

Information Theory Revisited

- We interpret the unobserved variables z as a latent representation or **code**, therefore, we shall refer to the model as a **probabilistic encoder**, since given a datapoint it produces a **distribution** over the possible values of the code

- Similarly, we refer to as a **probabilistic decoder**, since given a code it produces a **distribution** over the possible values of

Back to the optimization problem

- Lets substitute our intractable optimization problem:

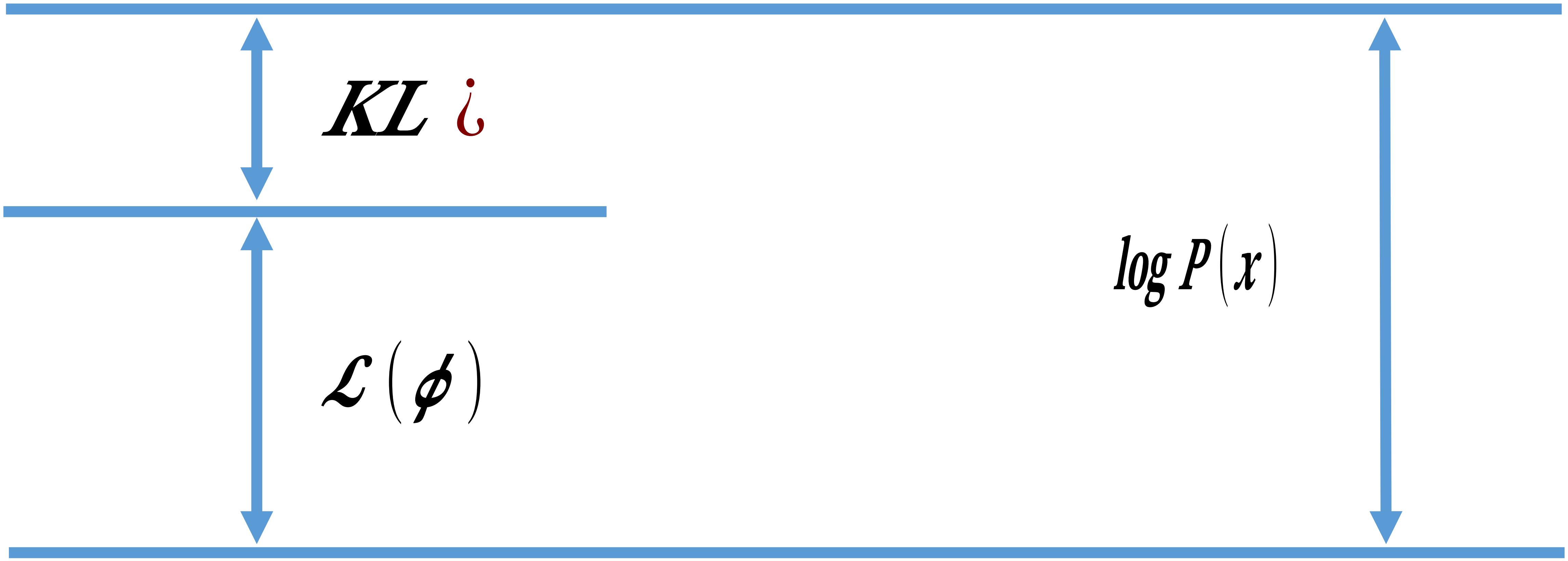
- With a variational optimization problem:

- Our goal is to find the closest in divergence to the exact conditional

- Does this transition really help us?

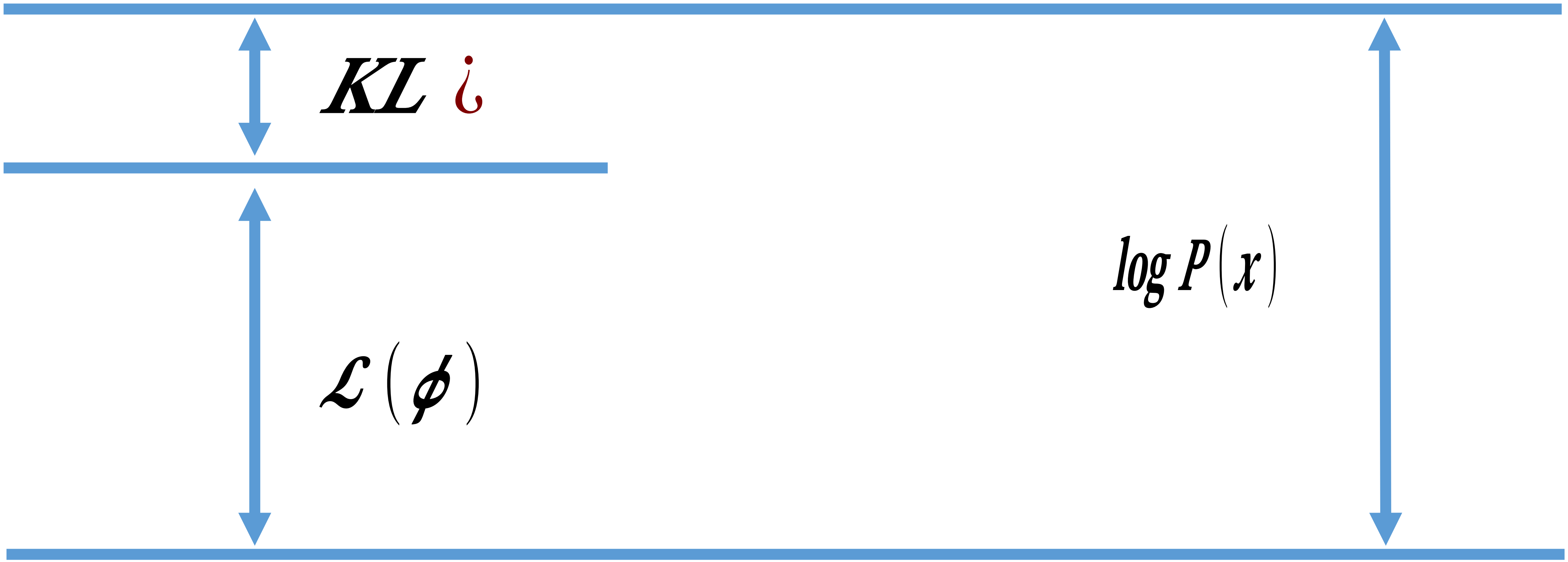
Let the derivations begin

“Minimizing” KL Divergence



$$\log p(x) = KL(q(z \vee x) \vee ¿ \, p(z|x)) + \mathcal{L}(\phi)$$

“Minimizing” KL Divergence



$$\log p(x) = KL\left(q(z \vee x) \vee ¿\, p(z|x)\right) + \mathcal{L}(\phi)$$

Evidence Lower BOund (ELBO)

Putting it all together

Assume P is Gaussian. What does this mean? What about Bernoulli?

Does not depend on z

$$\geq 0$$

We push the approximate posterior to the prior

$$\mathcal{L}(\phi)$$

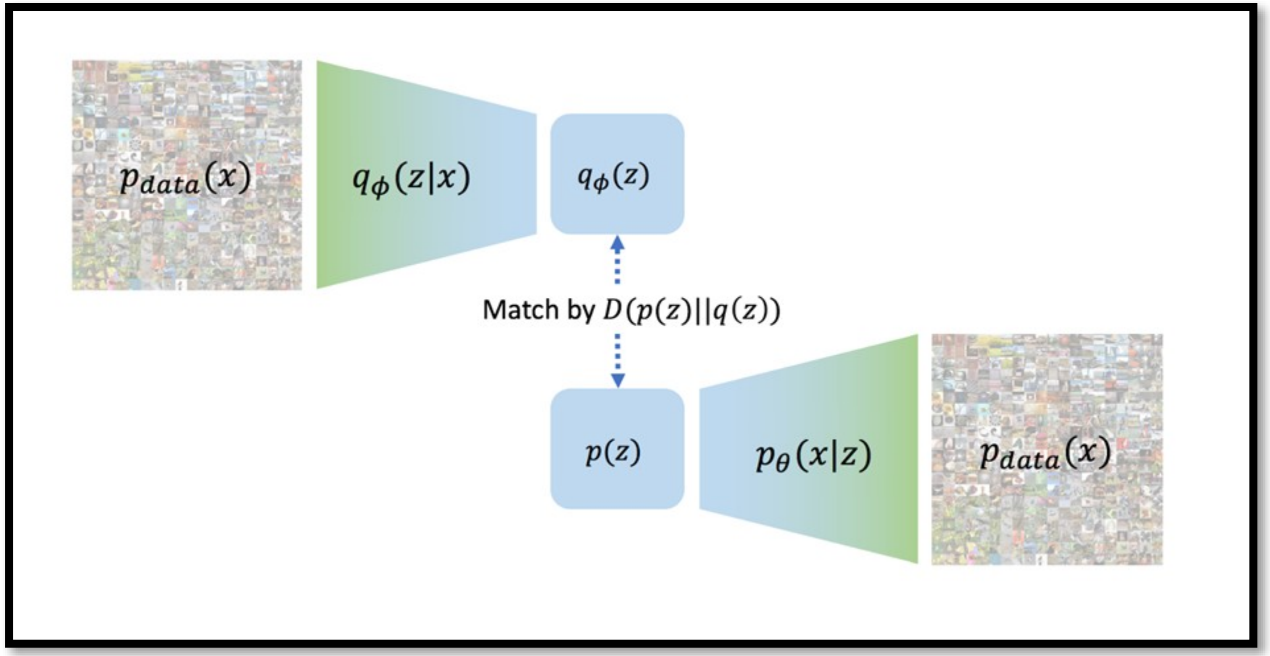
What we have so far

- We want to learn a latent variable model
- The likelihood and posterior are intractable and we can’t use EM
- We approximate the posterior using a tractable function and use KL divergence to pick the best possible approximation
- Because we cannot compute the KL, we optimize an alternative objective that is equivalent to the KL up to an added constant
- The ELBO has similar properties to a regularized autoencoder

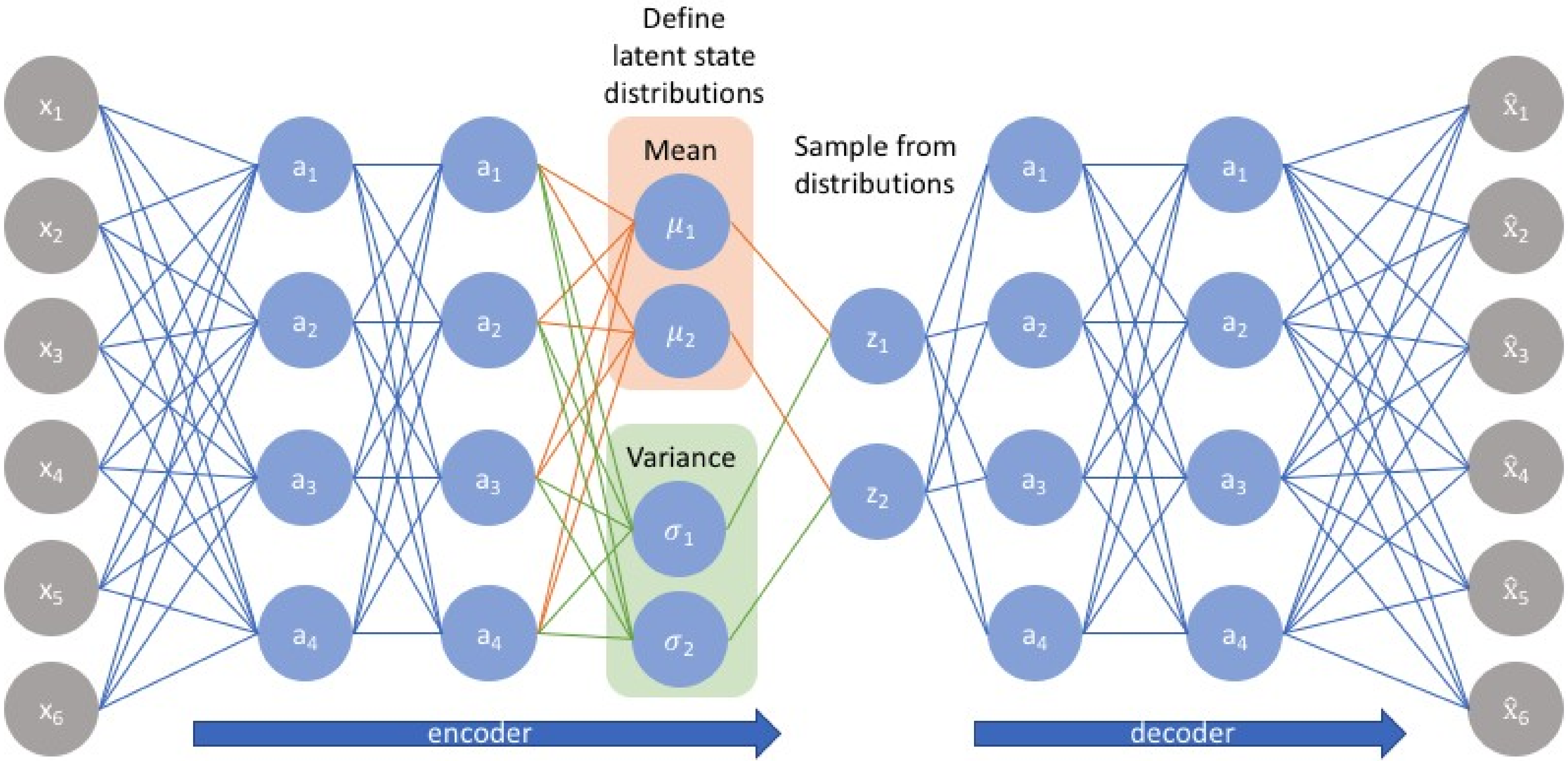
Neural Network Perspective

In practice

- How should we pick the approximating functions?
- A **probabilistic encoder** , approximating the true (intractable) posterior distribution
- A **generative decoder** , which notably does not rely on any input



The Variational Autoencoder



Features as Probability Distributions

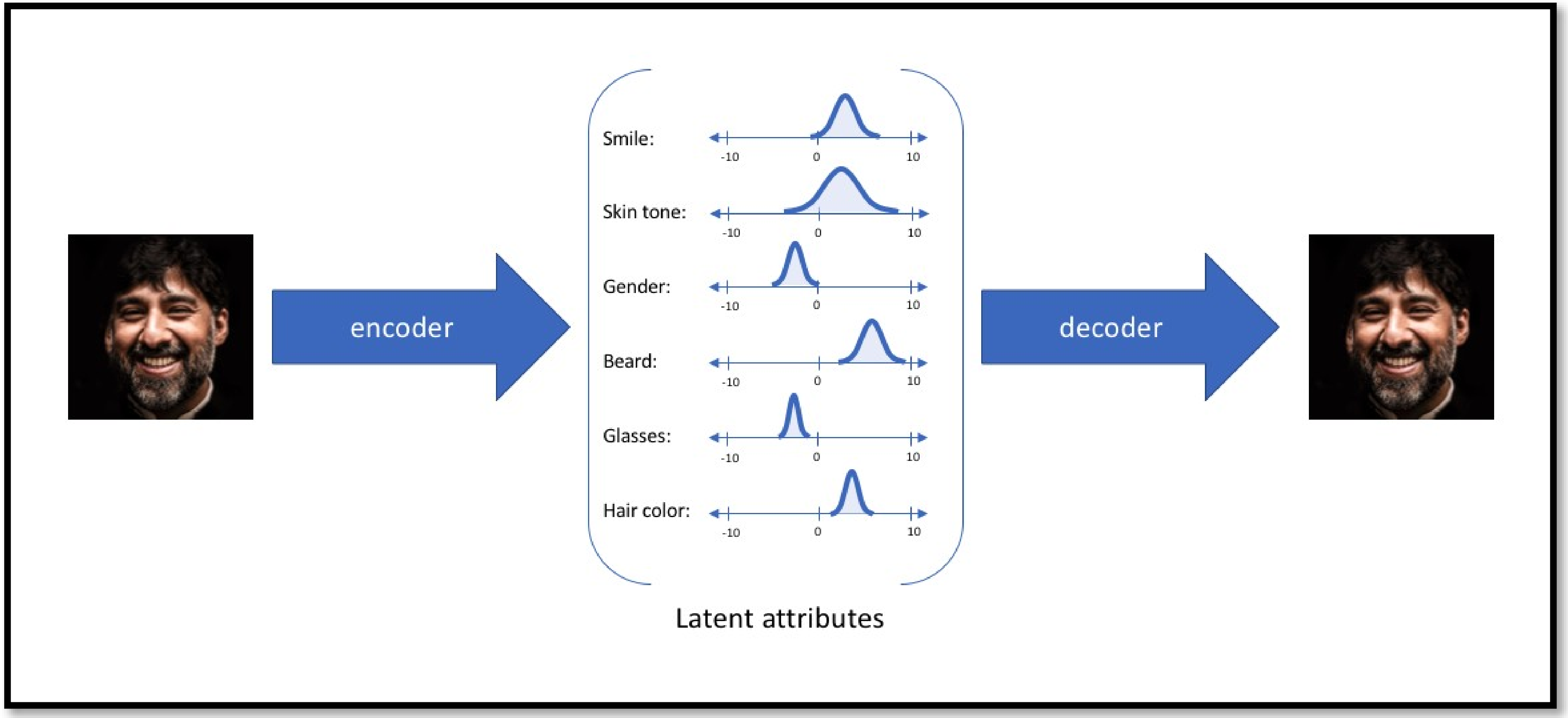
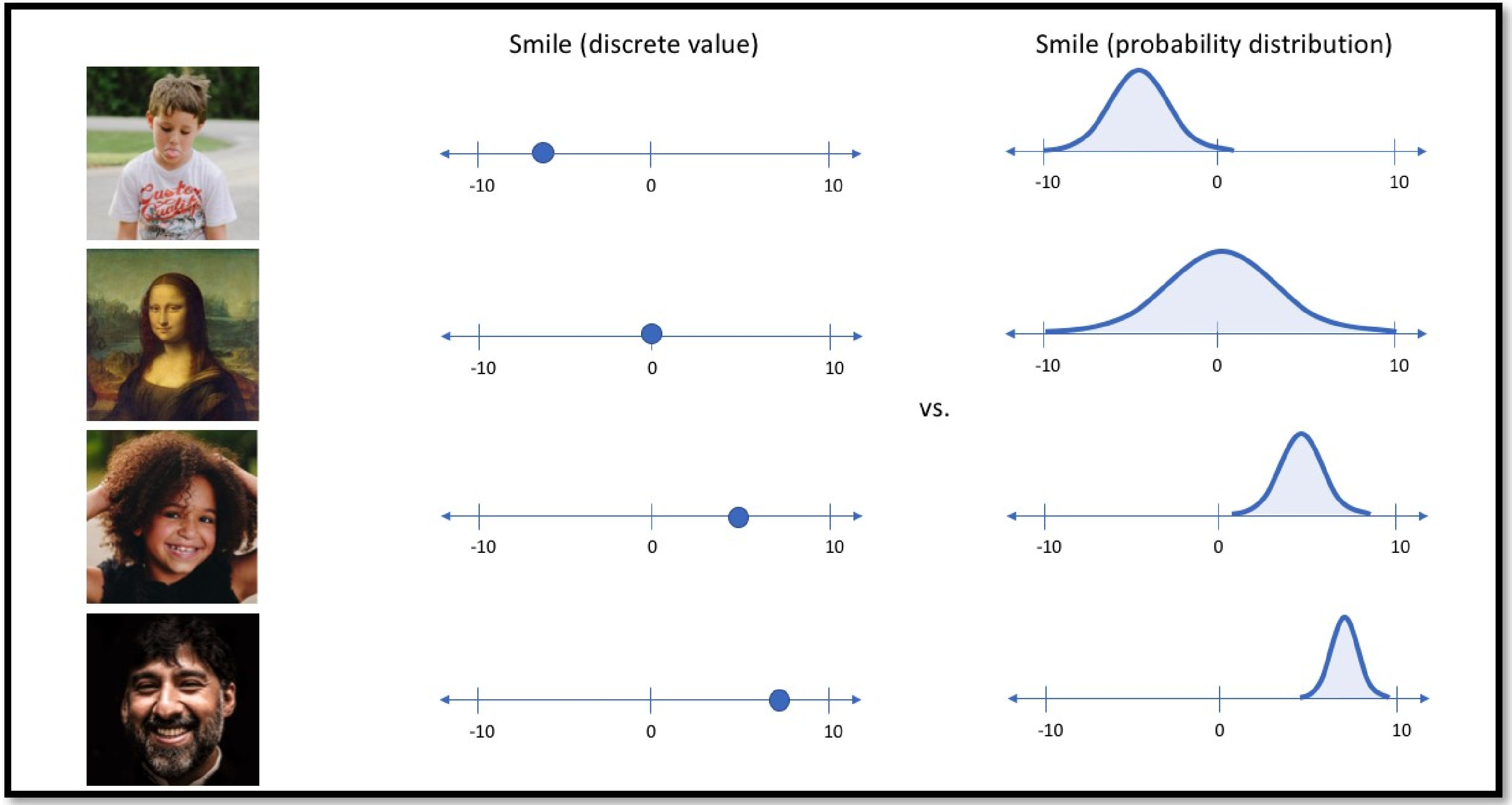


Image taken from jeremyjordan.me/variational-autoencoders/

Features as Probability Distributions



Features as Probability Distributions

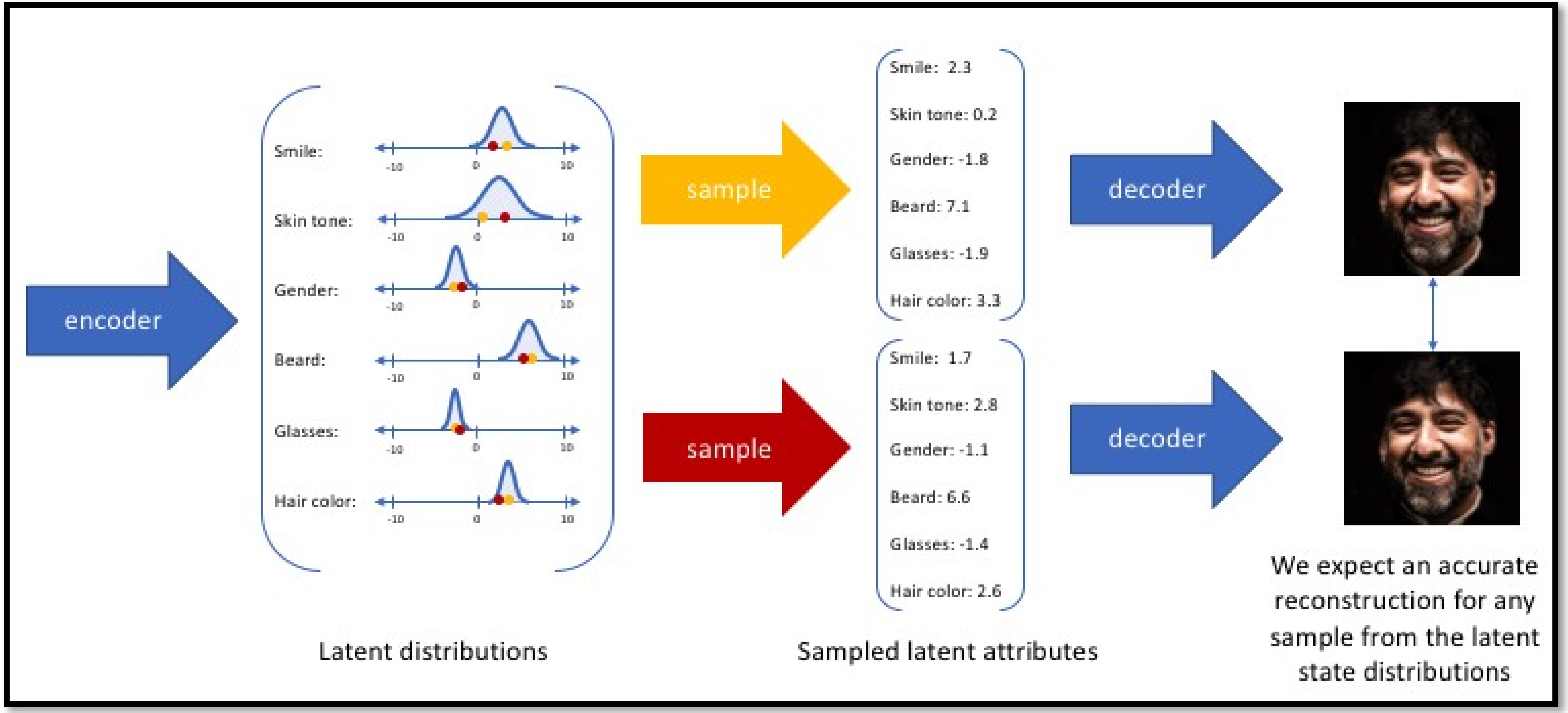
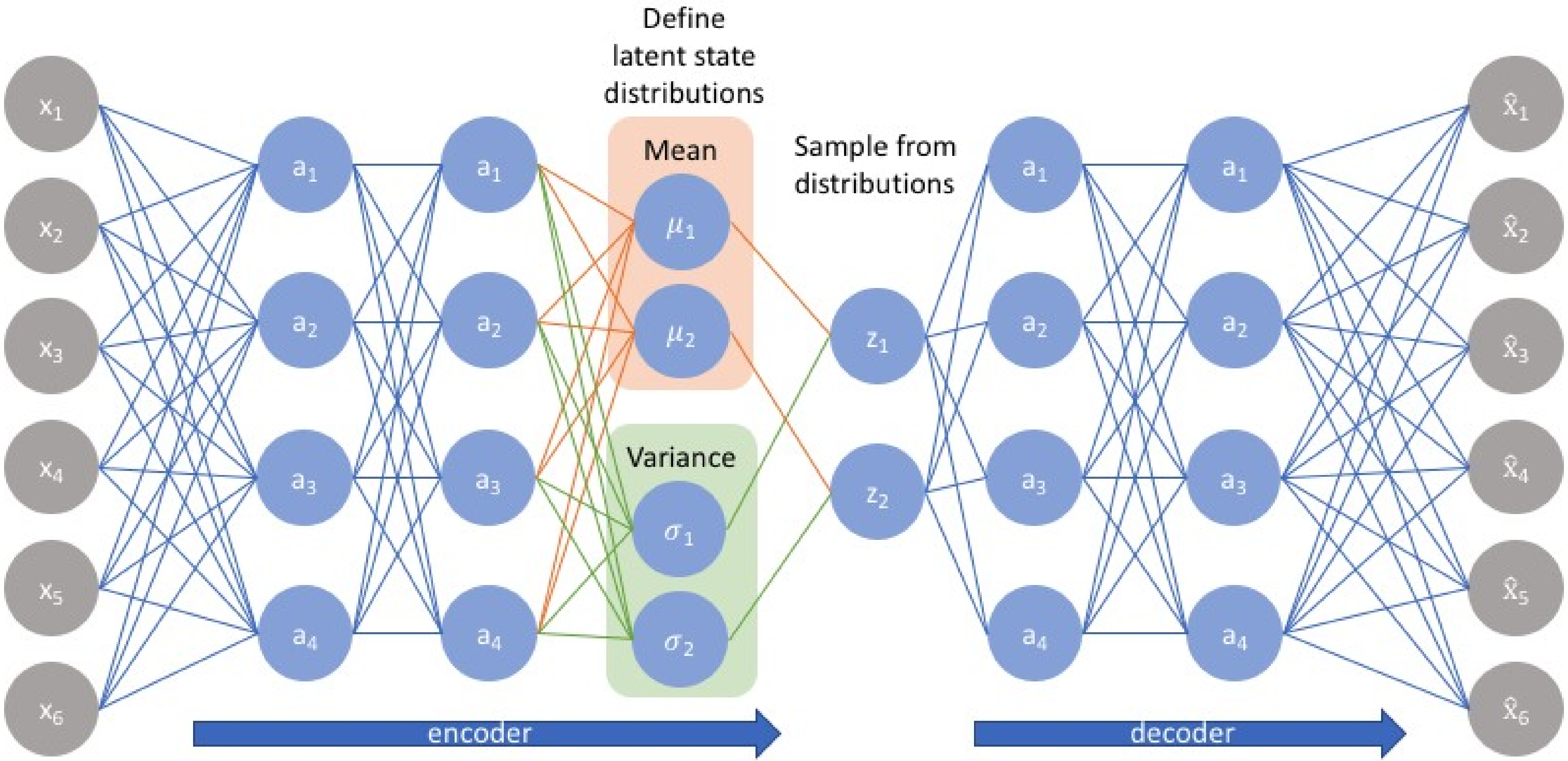
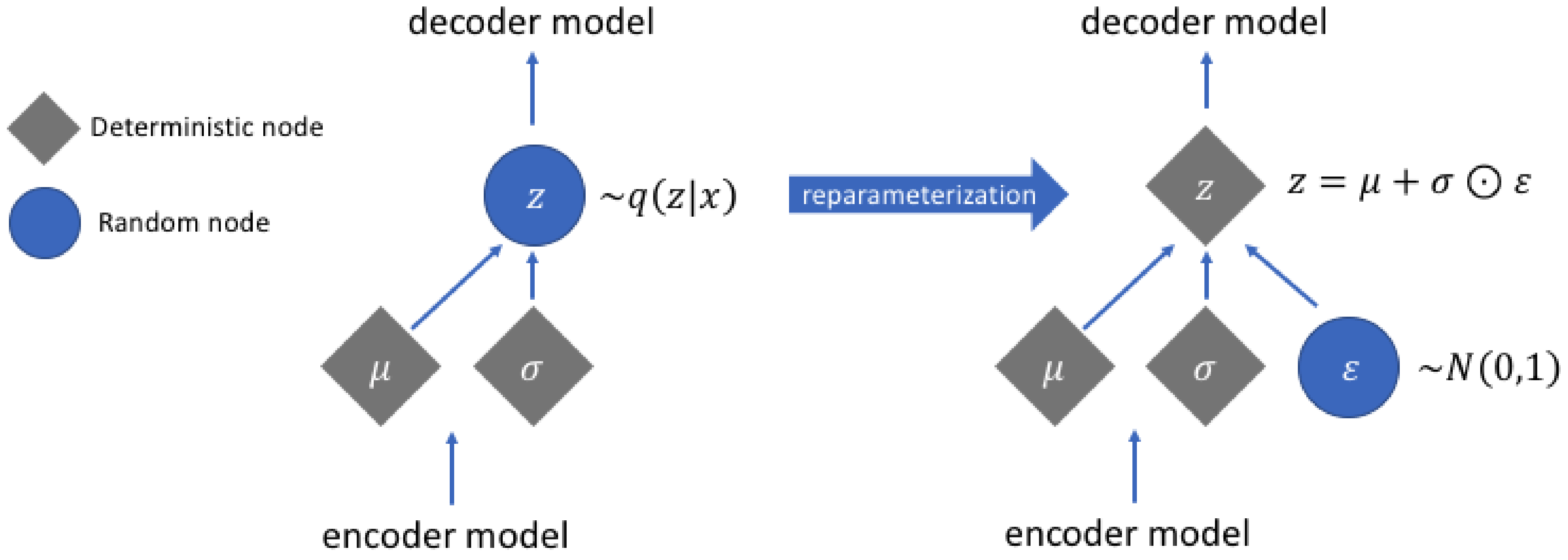


Image taken from jeremyjordan.me/variational-autoencoders/

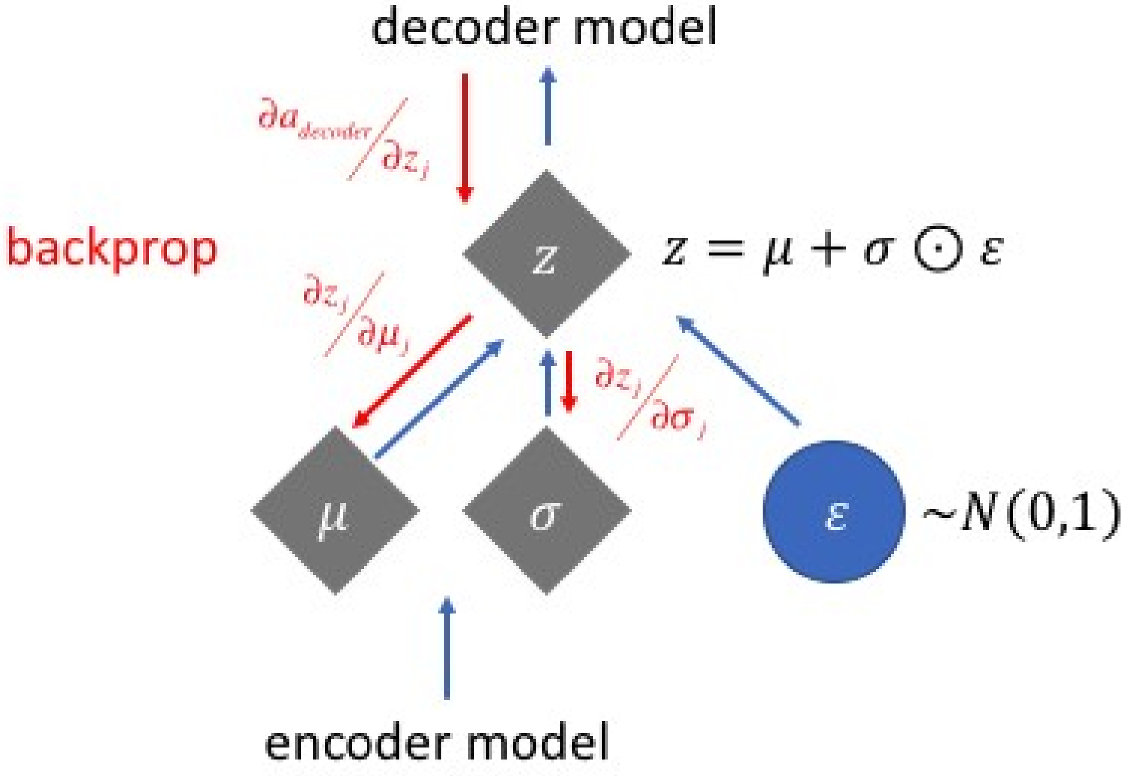
The Variational Autoencoder



Sample Layer



Back-Propagation through Random Operations



KL Divergence for Gaussian Distributions

- Recall that the density function for a multivariate Gaussian is:

- Consider two such distributions and compute

KL Divergence in code

$$KL\left[N(\mu,\Sigma) \vee \text{¿} N(0,1)\right] = -\frac{1}{2}\left[n + \log(det(\Sigma)) - \mu^T \mu - tr(\Sigma)\right]$$

```
vae = Model(x, x_recon)
recon_loss = metrics.binary_crossentropy(x, x_recon)
kl_loss = -0.5 * K.sum(1 + z_log_var – K.square(z_mean) – K.exp(z_log_var), axis=-
1)
vae_loss = K.mean(kl_loss(
vae.add_loss(vae_loss)
vae.compile(optimizer='rmsprop', loss=None)
```

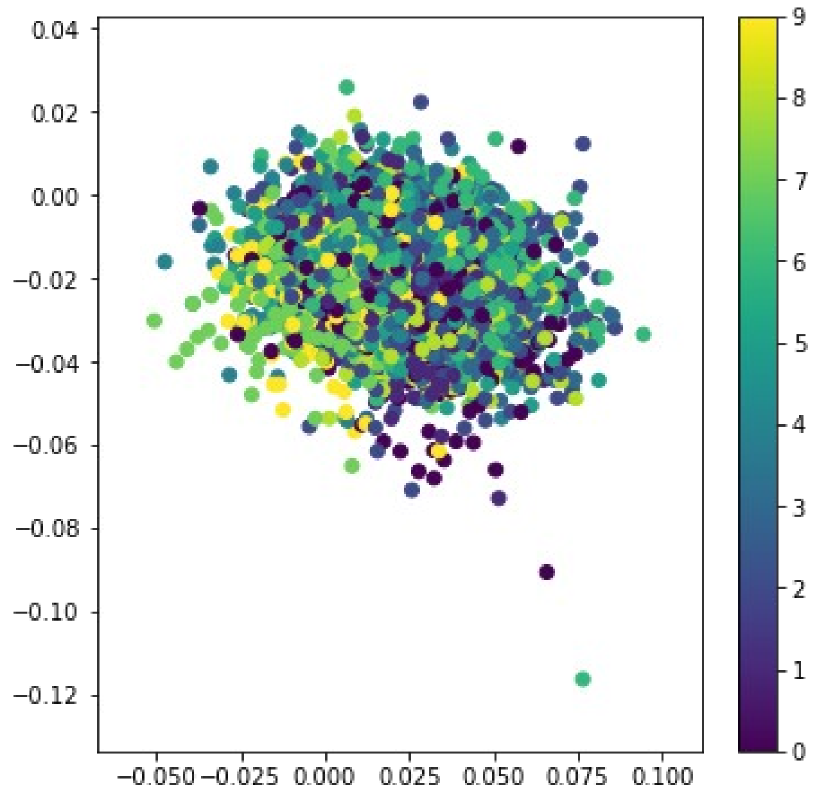
Back to VAE motivation

- VAE are a deep learning technique for learning useful latent representations
  - Image Generation
  - Latent Space Interpolation
  - Latent Space Arithmetic

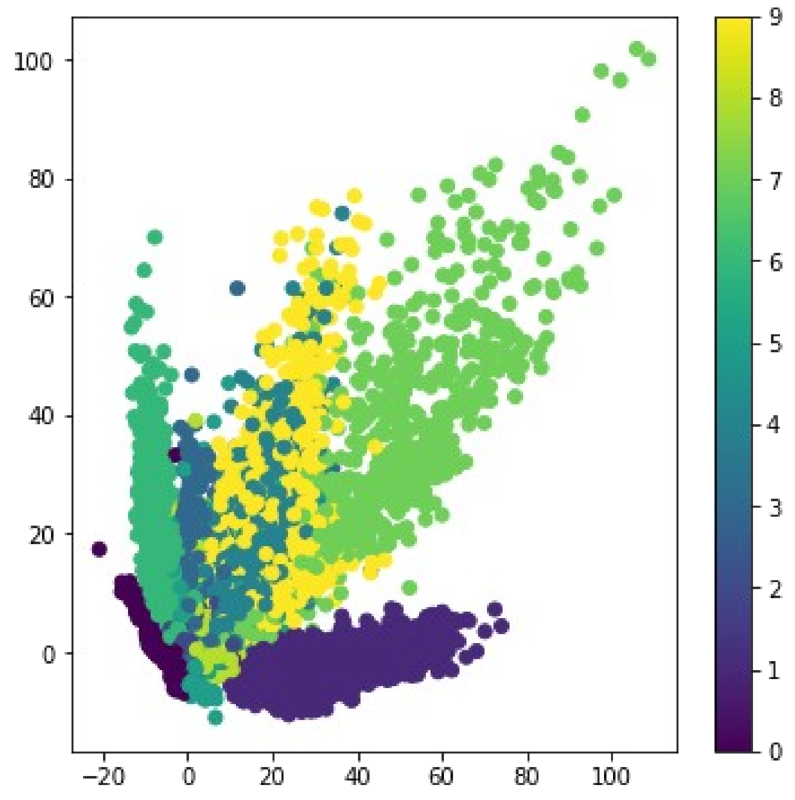
- Is the new learned latent space useful?

Latent Space Visualizations

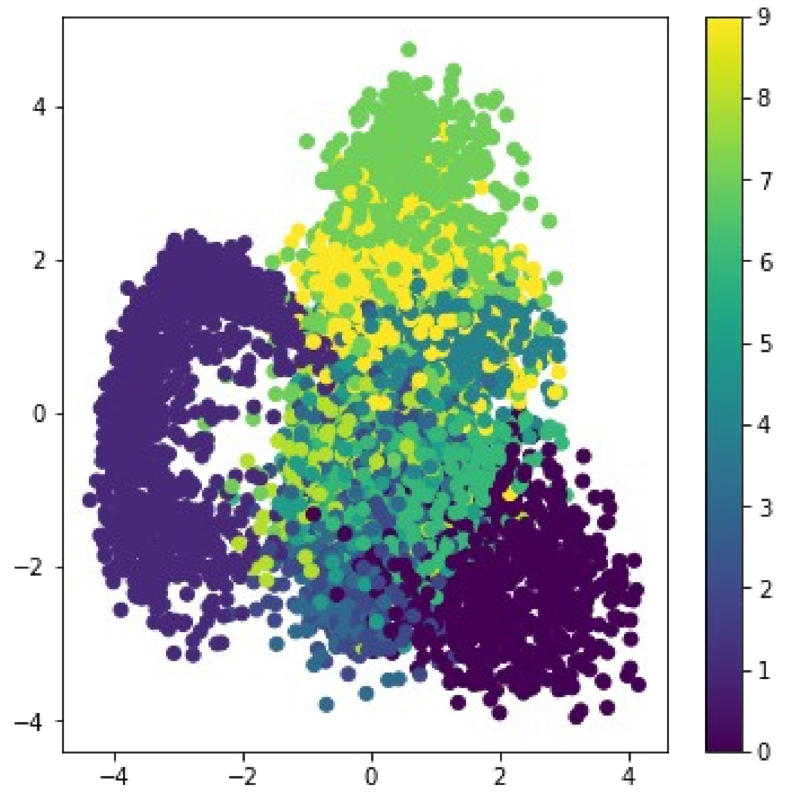
KL Divergence



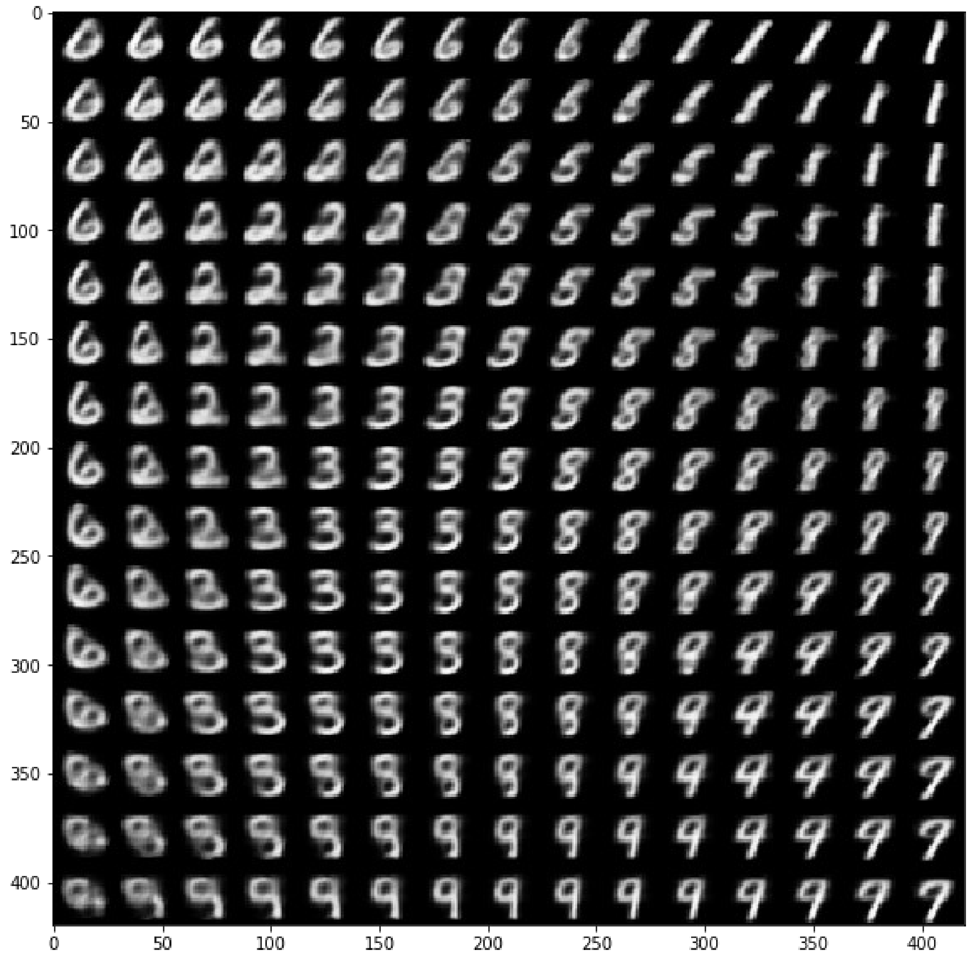
Reconstruction



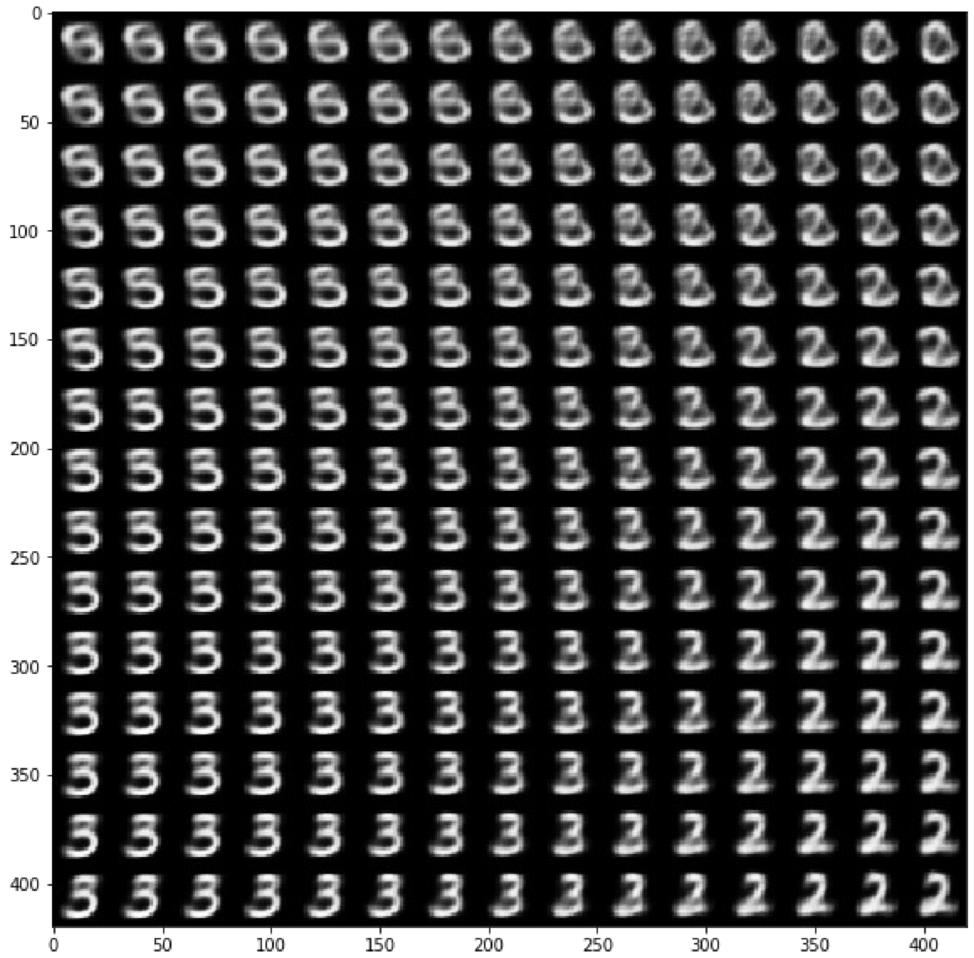
Reconstruction + KL Divergence



Generating Numbers (MNIST)



Generating Images from VAE

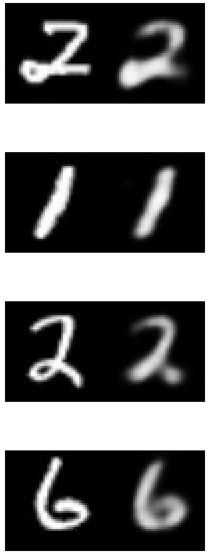


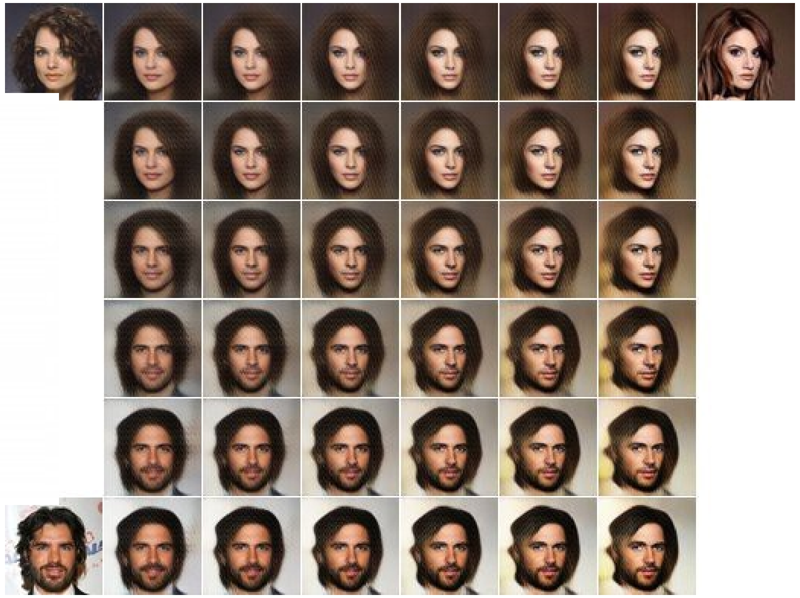
"Generating" Images from Traditional AE

Generating Faces (CelebA)

Latent Space Interpolation

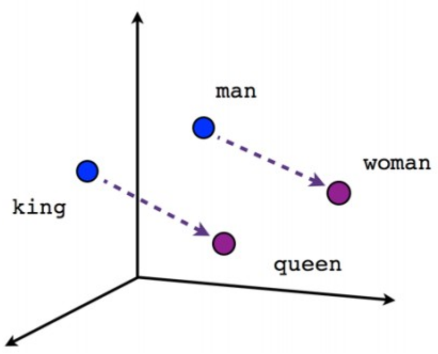
- If the latent space representation is useful, maybe we could take two different images, represent them as points in latent space, and create images from the line connecting the two points?



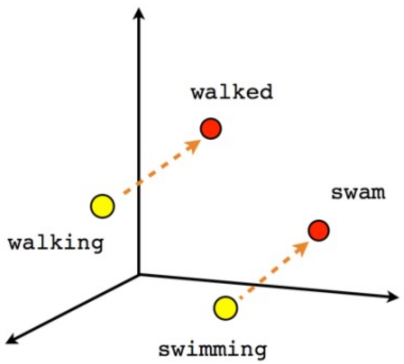


Latent Space Arithmetic

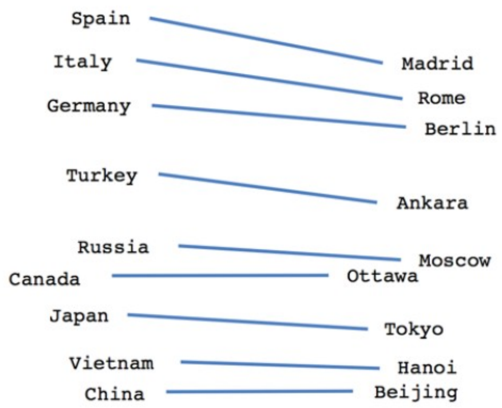
- Instead of interpolation, could we extract the latent vector responsible for a specific attribute?



Male-Female

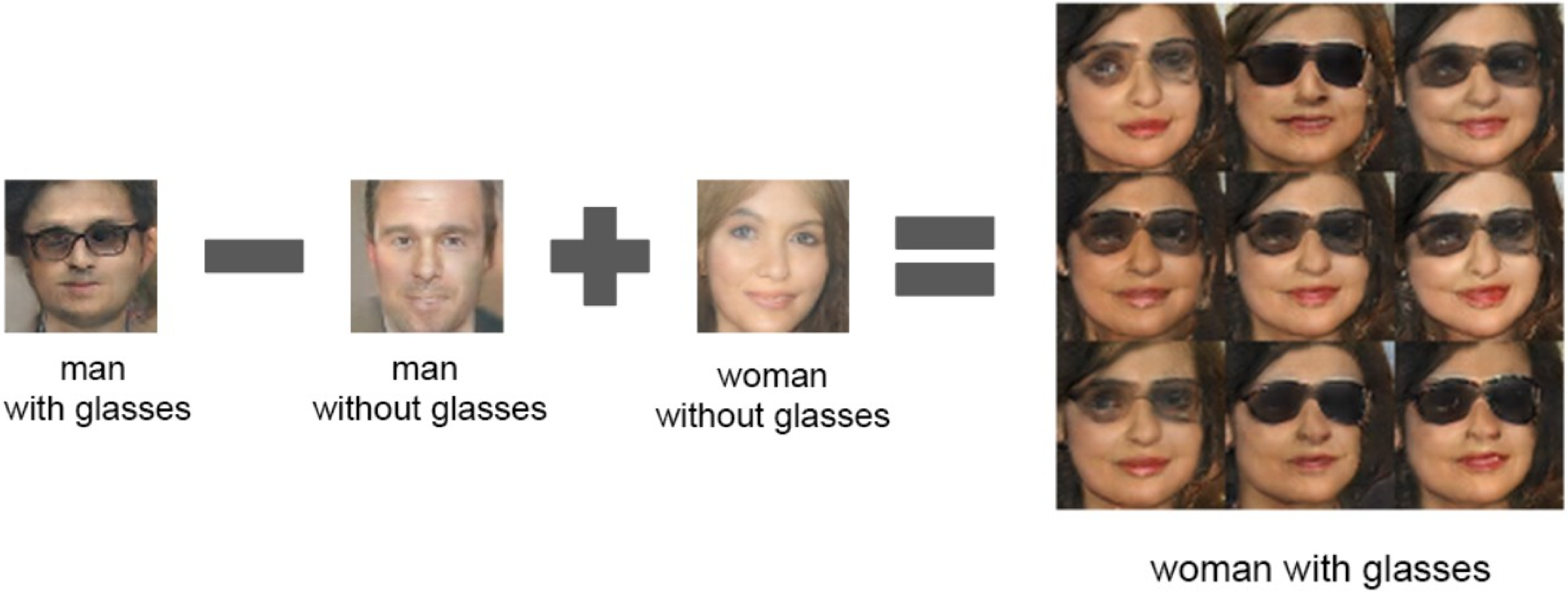


Verb tense



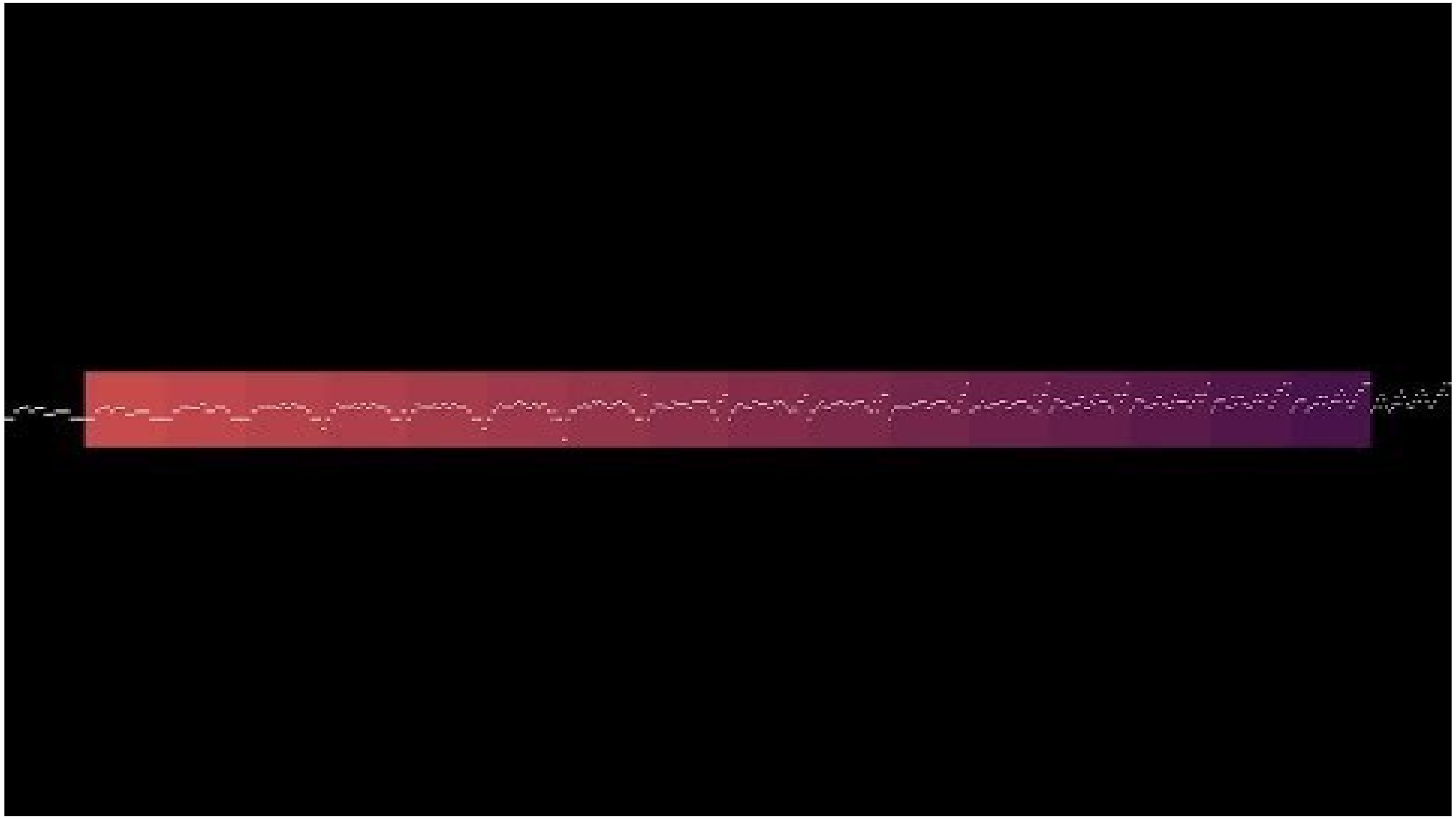
Country-Capital

Latent Space Arithmetic



Latent Space Arithmetic

Music VAE



Summary

- Probabilistic spin to traditional autoencoders
- Defines an intractable distribution and optimizes a variational lower bound
- Allows data generation and a useful latent representation
- Samples blurrier and lower quality images compared to state-of-the-art techniques

Bibliography

- https://arxiv.org/pdf/1312.6114.pdf
- https://arxiv.org/pdf/1606.05908.pdf
- https://arxiv.org/pdf/1502.04623.pdf
- http://blog.fastforwardlabs.com/2016/08/12/introducing-variational-autoencoders-in-prose-and.html
- http://blog.fastforwardlabs.com/2016/08/22/under-the-hood-of-the-variational-autoencoder-in.html
- https://ermongroup.github.io/cs228-notes/extras/vae/
- http://kvfrans.com/variational-autoencoders-explained/
- https://stats.stackexchange.com/questions/267924/explanation-of-the-free-bits-technique-for-variational-autoencoders
- http://szhao.me/2017/06/10/a-tutorial-on-mmd-variational-autoencoders.html
- https://www.cs.princeton.edu/courses/archive/spring17/cos598E/Ghassen.pdf
- https://www.jeremyjordan.me/variational-autoencoders/