E9: 309 ADL 25-11-2020

http://leap.ee.iisc.ac.in/sriram/teaching/ADL2020/

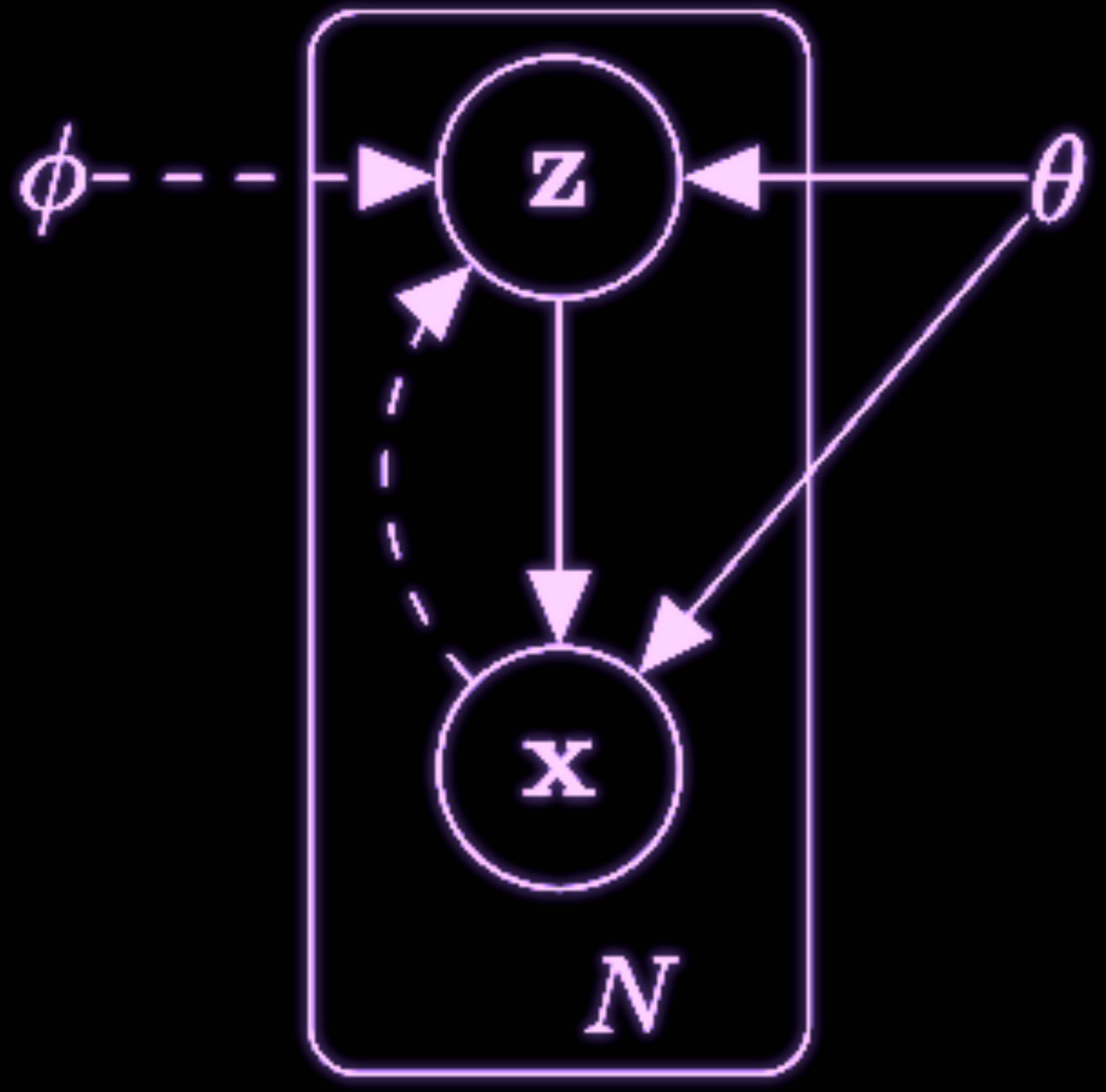
Housekeeping

- Mid-term exam
  - December 5th (Saturday) [Topics covered up to Dec 2nd]
  - Mode of exam
    - ✓ Time to respond - 3 hours
      - Exam paper uploaded in Teams Channel and response (photo-scanned and uploaded in your folder).
      - Open book, open notes
        - ★ Strictly no online communication or help sought.
          - ★ Academic integrity and ethics strongly followed.

Recap of previous class

Variational auto encoders

- The data $\mathbf{x}$ and latent variable $\mathbf{z}$
- The forward model
  - ✓ Sample the latent variable
  - ✓ Sample the data given the latent $p_\theta(\mathbf{x}|\mathbf{z})$
- The marginal distribution $\int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}$
  - ✓ maybe intractable
- The posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$
  - ➡️ may also be intractable

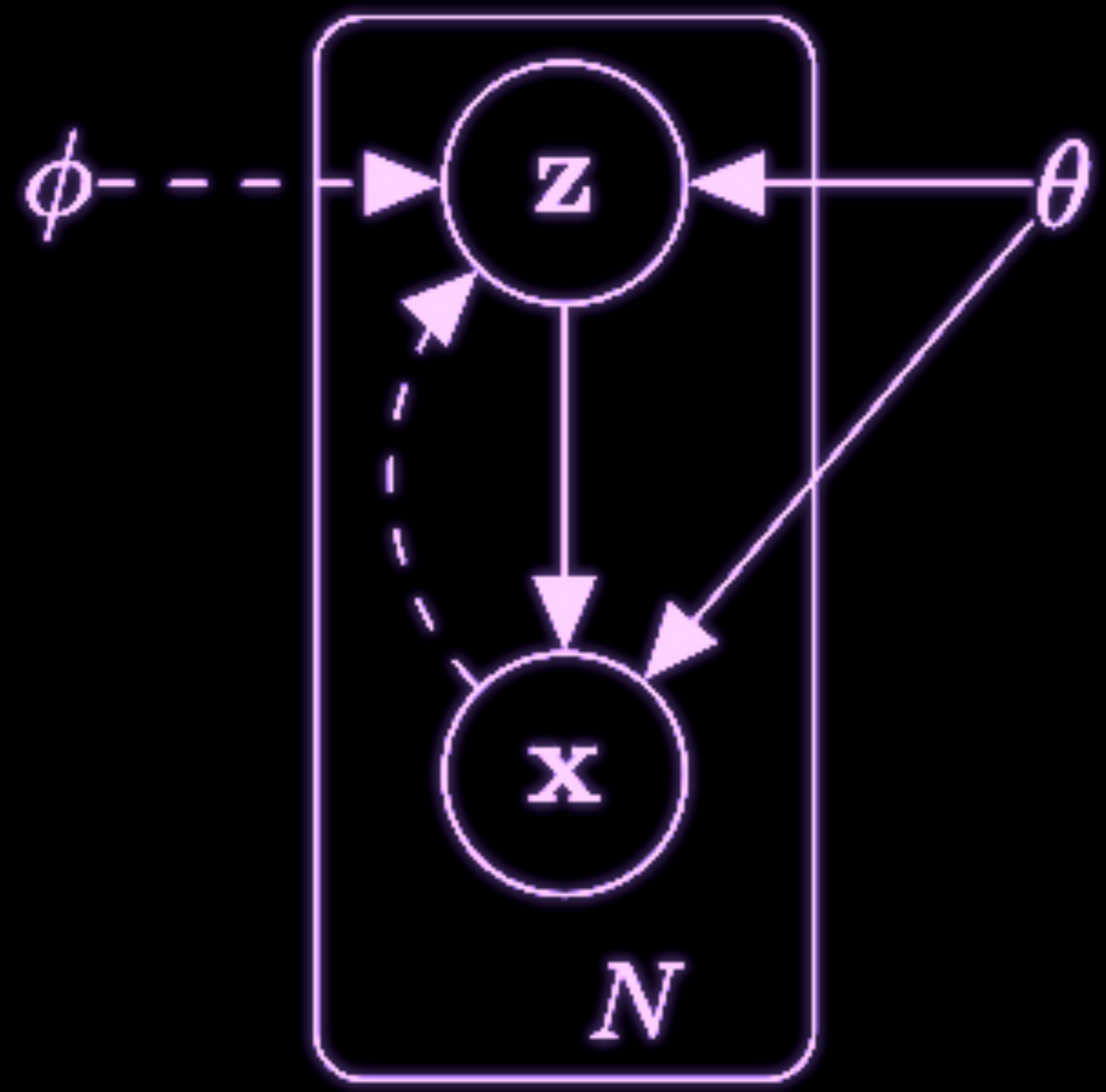


Variational auto encoders

- Approximate the posterior

$$q_\phi(\mathbf{z}|\mathbf{x}) \sim p_\theta(\mathbf{z}|\mathbf{x})$$

  - Using variational lower bound.



Variational lower bound

(Variational approx) - I

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x},\mathbf{z})d\mathbf{z}$$

log-likelihood

$$= \log \int p_\theta(\mathbf{x},\mathbf{z})\frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}d\mathbf{z}$$

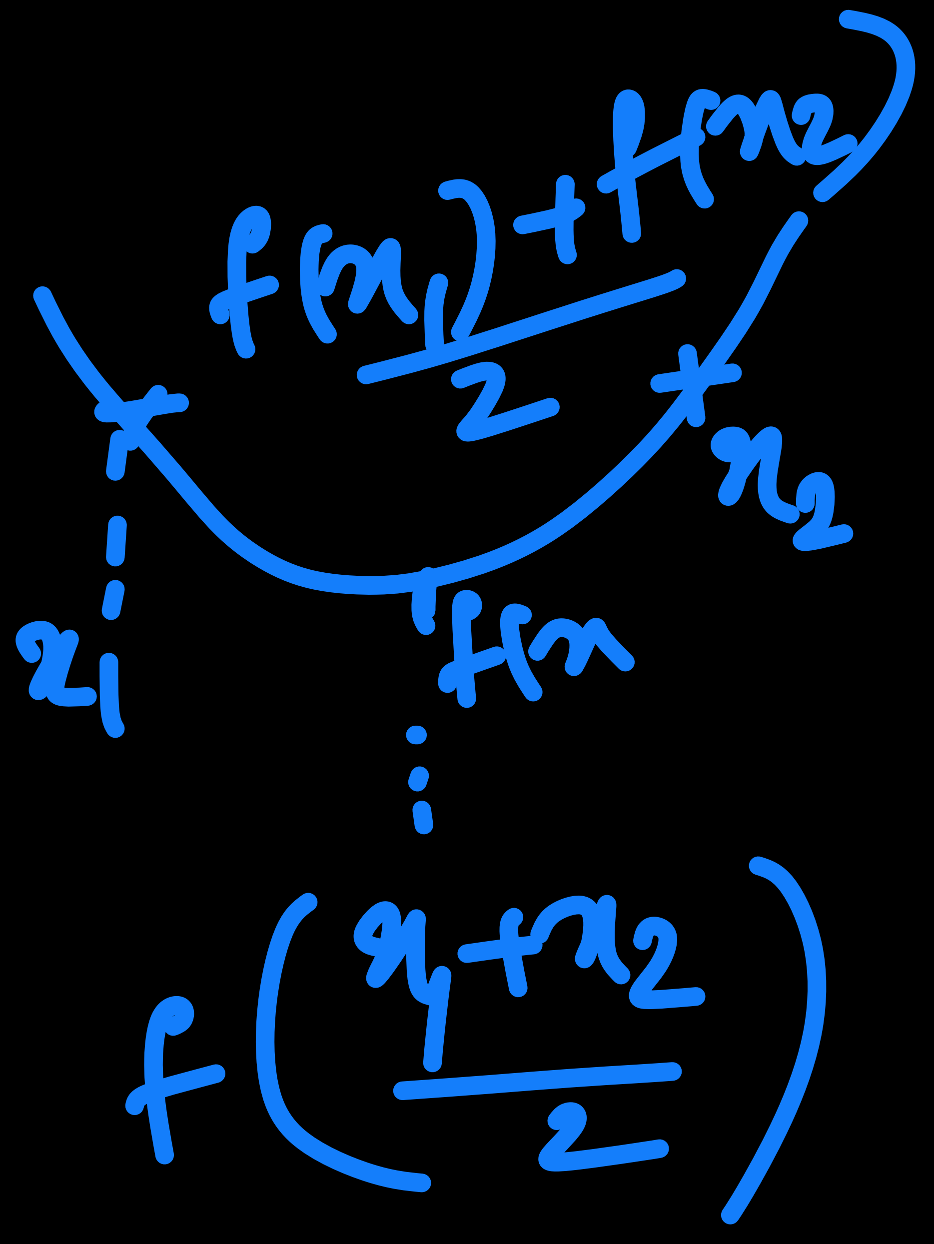
$q_\phi(z|x)$
↳ density function

$$= \log\left(\mathbb{E}_q\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right)$$

$$\geq \mathbb{E}_q\left(\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right)$$

[Jensen's inequality]

$$\cancel{= \mathbb{E}_q\left(\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right)}$$



$$= \mathbb{E}_q\left(\log\left[p_\theta(\mathbf{x},\mathbf{z})\right]\right) + H_q(\mathbf{z}|\mathbf{x})$$

Entropy

Shannon entropy

$$H_q(z|x) = -\int_z q_\phi(z|x) \log q_\phi(z|x)\, d$$

$$\geq 0$$

$$\underbrace{\log p_\theta(x) \geq E_q[\log p_\theta(x,z)]}_{\text{Variational lower bound}}$$

Variational lower bound

(Proof - II)

$$0 \leq \underbrace{KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))}\ \overset{\Delta}{=} \int q_\phi(\mathbf{z}|\mathbf{x}) log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

$$= - \int q_\phi(\mathbf{z}|\mathbf{x}) log \frac{p_\theta(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

$$p_\theta(z/x) = \frac{p_\theta(z,x)}{p_\theta(x)}$$

$$q_\phi(z/x) \sim p_\theta(z/x)$$

$$= - \int q_\phi(\mathbf{z}|\mathbf{x}) log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z} + \boxed{\int q_\phi(\mathbf{z}|\mathbf{x}) log\ p_\theta(\mathbf{x}) d\mathbf{z}}$$

$$= - \underbrace{\int q_\phi(\mathbf{z}|\mathbf{x}) log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z}} + \underbrace{log\ p_\theta(\mathbf{x})}$$

log-likelihood

✴ But $\underbrace{KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})) \geq 0}$ … thus we get $\boxed{log\ p_\theta(\mathbf{x}) \geq \int q_\phi(\mathbf{z}|\mathbf{x}) log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z}}$

Variational lower bound

$\log p_\theta(x) \geq L(x; \phi, \theta)$

- Defining the variational lower bound

$$L(\mathbf{x};\phi,\theta) = \int q_\phi(\mathbf{z}|\mathbf{x}) log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} dz = \mathbb{E}_q\left(log \left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \right)$$

✓ Called ELBO (evidence lower bound of the data)

$$\boxed{L(\mathbf{x};\phi,\theta) = log\ p_\theta(\mathbf{x}) - KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))}$$

- Maximizing the lower bound ← for neural model
  - ★ Maximizes the log likelihood of the data
  - ★ Minimized the KL divergence between the true posterior distribution and appromixation.

Variational lower bound

- Variational lower bound properties - can be split two terms

$$\boxed{\underbrace{L(\mathbf{x};\phi,\theta)}} = \mathbb{E}_q\left(log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right)$$

$$= \mathbb{E}_q\left(log\left[\frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right)$$

$$= -\underbrace{KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))} + \mathbb{E}_q\left(\underbrace{log\ p_\theta(\mathbf{x}|\mathbf{z})}\right)$$

$\uparrow L \le \underline{\log p_\theta(x)}$

- Can be split into two terms

$\underbrace{-KL(q_\phi(z|x), p_\theta(z))}$

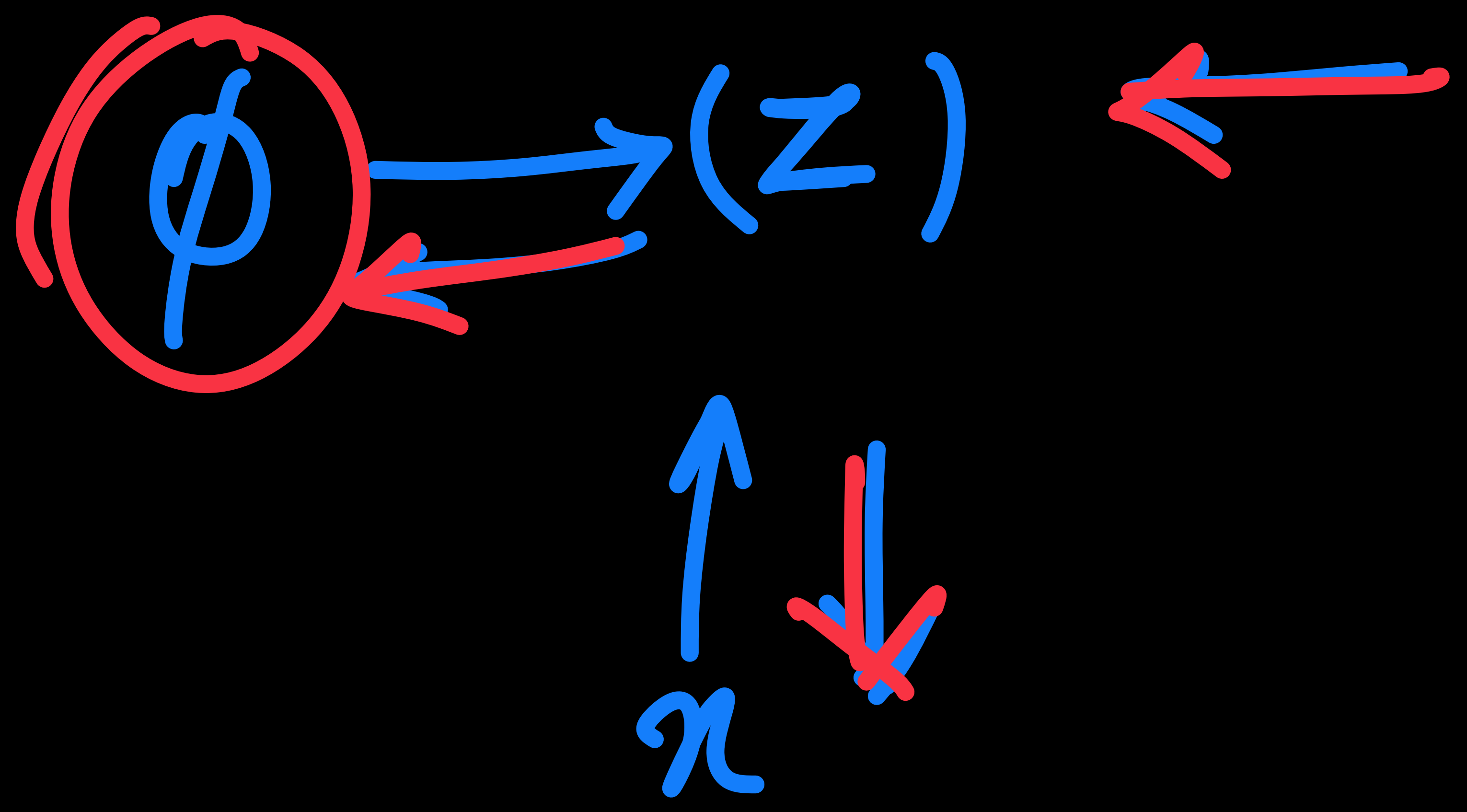
$$\uparrow \underline{L} = \min KL\left(q_\phi(z/x) \,||\, P_\theta(z)\right) + \max E_q\left(\log P_\theta(z/x)\right)$$

How?

(θ, ϕ)

$\boxed{z - \text{random variable.}}$

$$\frac{\partial L}{\partial \theta}\uparrow, \quad \frac{\partial L}{\partial \phi}\uparrow$$



Goal: How can we decouple randomness of variable and the dependence on the learnable parameters

$\epsilon \sim p(\epsilon) \xrightarrow{g_{\phi, x}(\cdot)} z/x \sim q_\phi(z/x) \approx p_\theta(z/x)$

has no parameter

learnable parameters

Reparameterization [Change of variables]

- Take a random variable independent of the data and parameters

$$\epsilon \sim p(\epsilon)$$

$p(\epsilon)$ [is not dependent on θ, ϕ]

- Transform the random variable to the latent variable with parameters

$$\mathbf{z} = g(\epsilon, \phi, \mathbf{x})$$

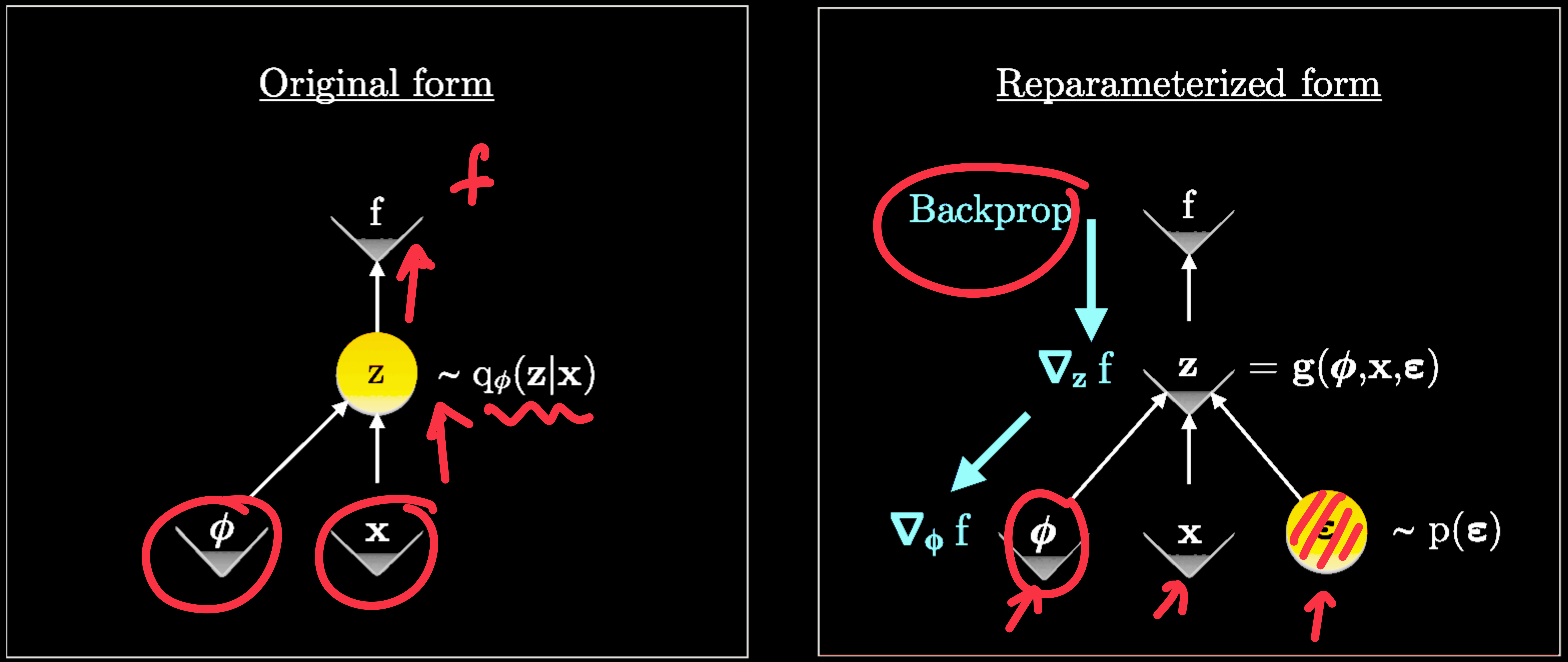
function

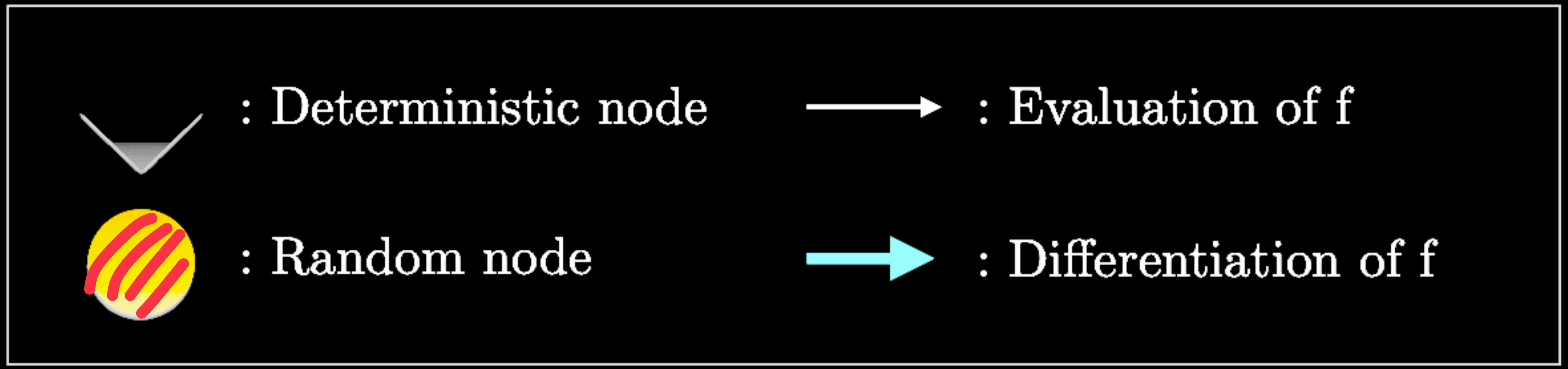
$q_\phi(z/x) = f(\phi, x)$

- Now the gradient computation of the expectation can be simplified.

  ✓ Decoupling the sampling and the gradient computations on the term involving $\mathbb{E}_q\left(log\ p_\theta(\mathbf{x}|\mathbf{z})\right)$

Reparameterization





Assumptions and Approximations for VAE

$$L(\mathbf{x};\phi,\theta) = KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(z)) + \mathbb{E}_q\left(log\ p_\theta(\mathbf{x}|\mathbf{z})\right)$$

- Assume

$$p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I})$$ [Standard Gaussian]

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu}(\mathbf{x}),\boldsymbol{\sigma}(\mathbf{x})^2\mathbf{I})$$ [function of data]

$$\Lambda = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots\ \sigma_d^2 \end{bmatrix}$$

d- latent dimensions

- The conditional distribution

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x};\boldsymbol{f}_\theta(\mathbf{z}),\mathbf{I})$$

covariance

Assumptions and Approximations

$$L(\mathbf{x};\phi,\theta) = \boxed{KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))} + \mathbb{E}_q\left(\underbrace{log\ p_\theta(\mathbf{x}|\mathbf{z})}\right)$$

- Assume

$$p_\theta(z) = \mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I})$$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu}(\mathbf{x}),\boldsymbol{\sigma}(\mathbf{x})^2\mathbf{I})$$

Prove the KL divergence expression for 2 Gaussians

- The KL divergence

$$\mathcal{D}[\mathcal{N}(\mu_0,\Sigma_0)\|\mathcal{N}(\mu_1,\Sigma_1)] = \underbrace{\frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1-\mu_0)^\top\Sigma_1^{-1}(\mu_1-\mu_0) - k + \log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)\right)}$$

dimension

✓ Between two Gaussians

Assumptions and Approximations

MSE

$$L(\mathbf{x};\phi,\theta) = \underbrace{KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(z))}_{\text{regularization}} + \underbrace{\mathbb{E}_q\left(log\ p_\theta(\mathbf{x}|\mathbf{z})\right)}_{\text{MSE}}$$

↑ regularization

- Computing the lower bound
  - ✓ Expectation $\boxed{\mathbb{E}_q\left(log\ p_\theta(\mathbf{x}|\mathbf{z})\right)} \approx \frac{1}{L}\sum_{l=1}^{L} \log P_\theta(x/z^l)$

    $z^l \sim q_\phi(z/x)$
    - approximated as single sample estimate from the Gaussian.
    - Will be equivalent of the reconstruction error

Training samples $\underline{x}_1, \underline{x}_2, \ldots x_N$

$$\underline{x}_i \xrightarrow[\phi]{\text{Encoder}} \mu_d(x) \quad \log$$

$$\underbrace{\log \Lambda = \begin{bmatrix} \sigma_1(x) & & 0 \\ & \ddots & \\ 0 & & \log \sigma_d(x) \end{bmatrix}}$$

$q_\phi(z/x)$

$$\underbrace{E_{q_\phi}\left[\log p_\theta(x/z)\right]} \approx \underbrace{\frac{1}{L}\sum_{l=1}^{L} \log \text{———}}_{L = 1} \leftarrow (\hat{z}_{i,l}) \sim q_\phi(z/x)$$

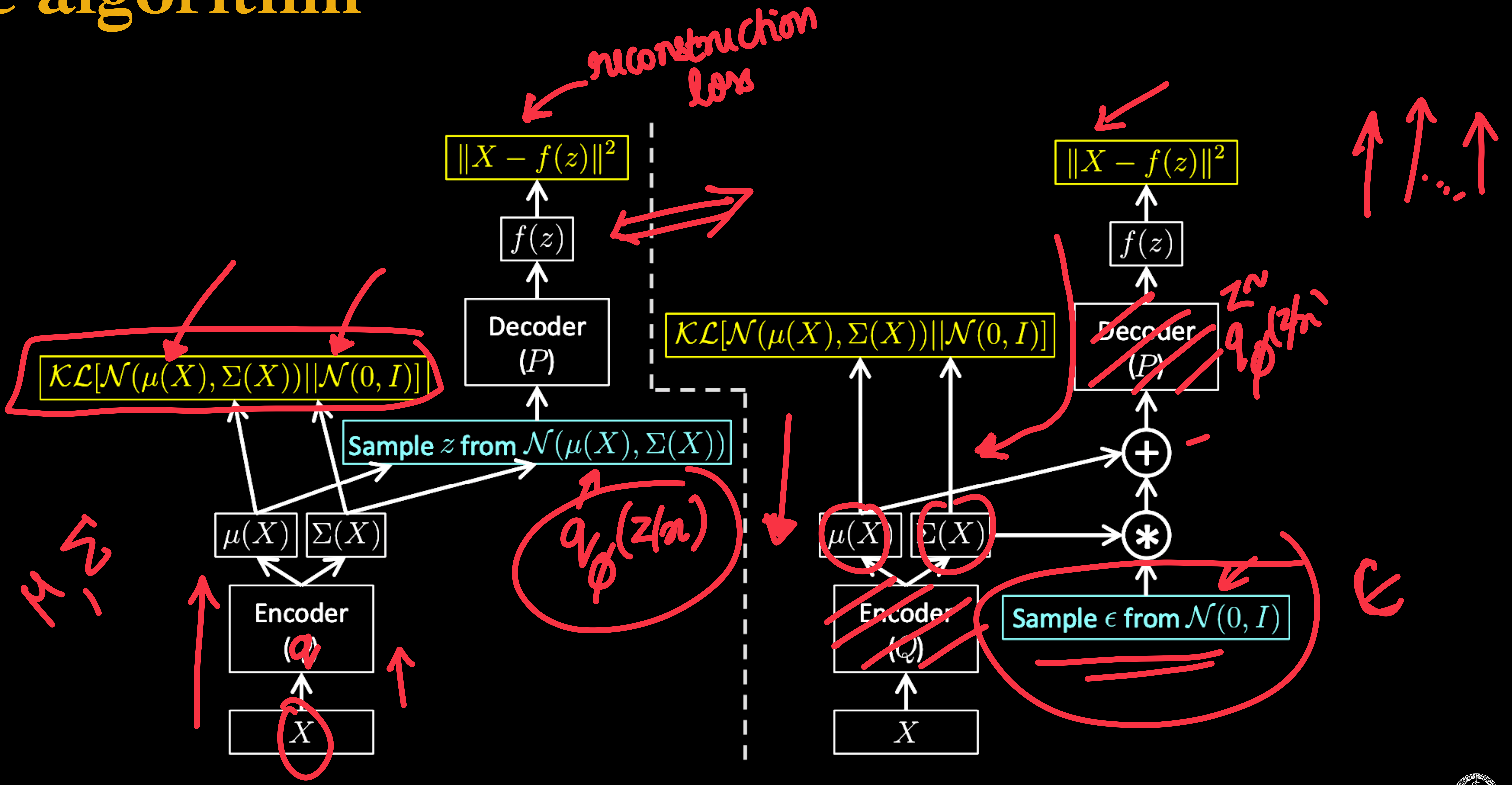
$$E_q[\log p_\theta(x/z)] \approx \log p_\theta[x^i/z^i] \uparrow$$

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

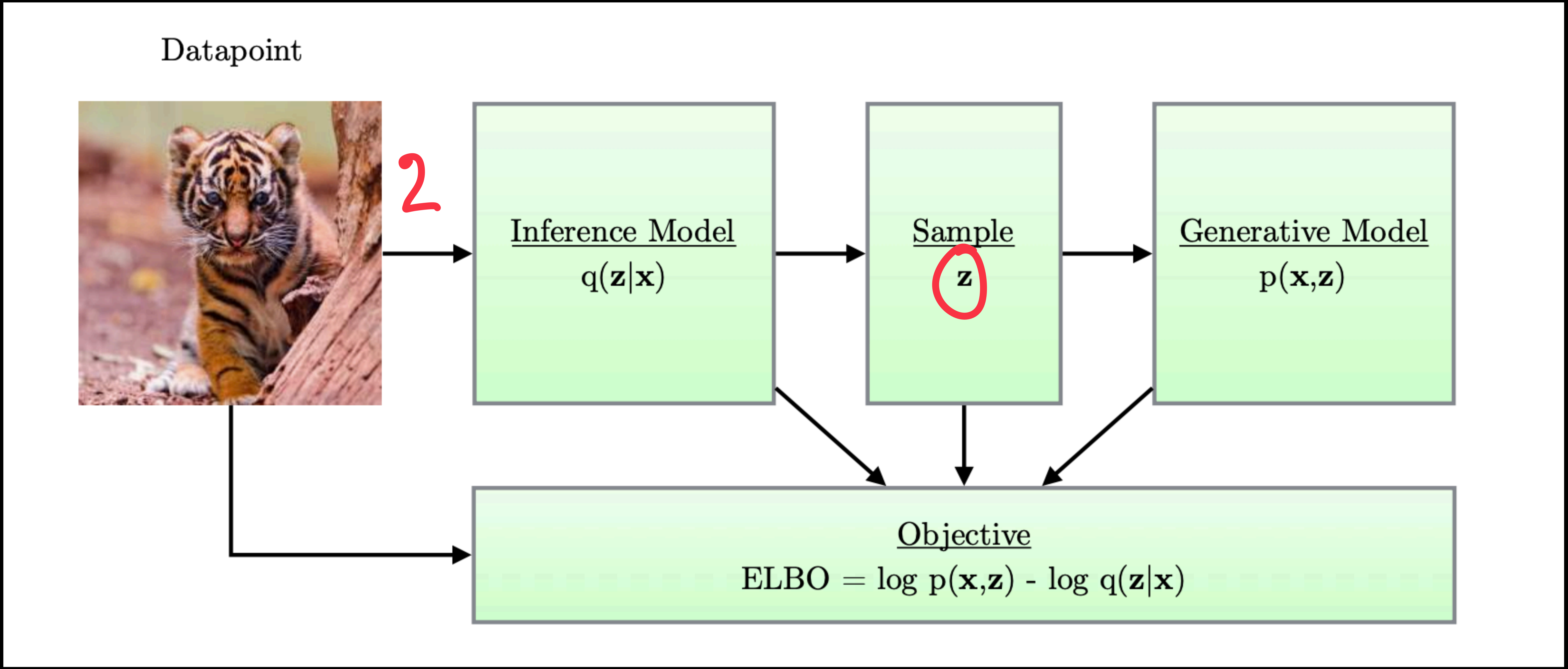
$$\sum_i - \| \underbrace{x^i}_{\text{data}} - \underbrace{f_\theta(z^i)}_{\substack{\text{decoder} \\ \text{neural} \\ \text{network}}} \|^2 \downarrow$$

$$L = \underbrace{\boxed{KL\left(\text{divergence } q_\phi(z|x) \,\|\, P_\theta(z)\right)}}_{\text{Regularized } \underline{\text{MSE}}} + \underbrace{E_q\left(\underbrace{\|x - f_\theta(z)\|^2}_{\text{AE}}\right)}$$

The algorithm

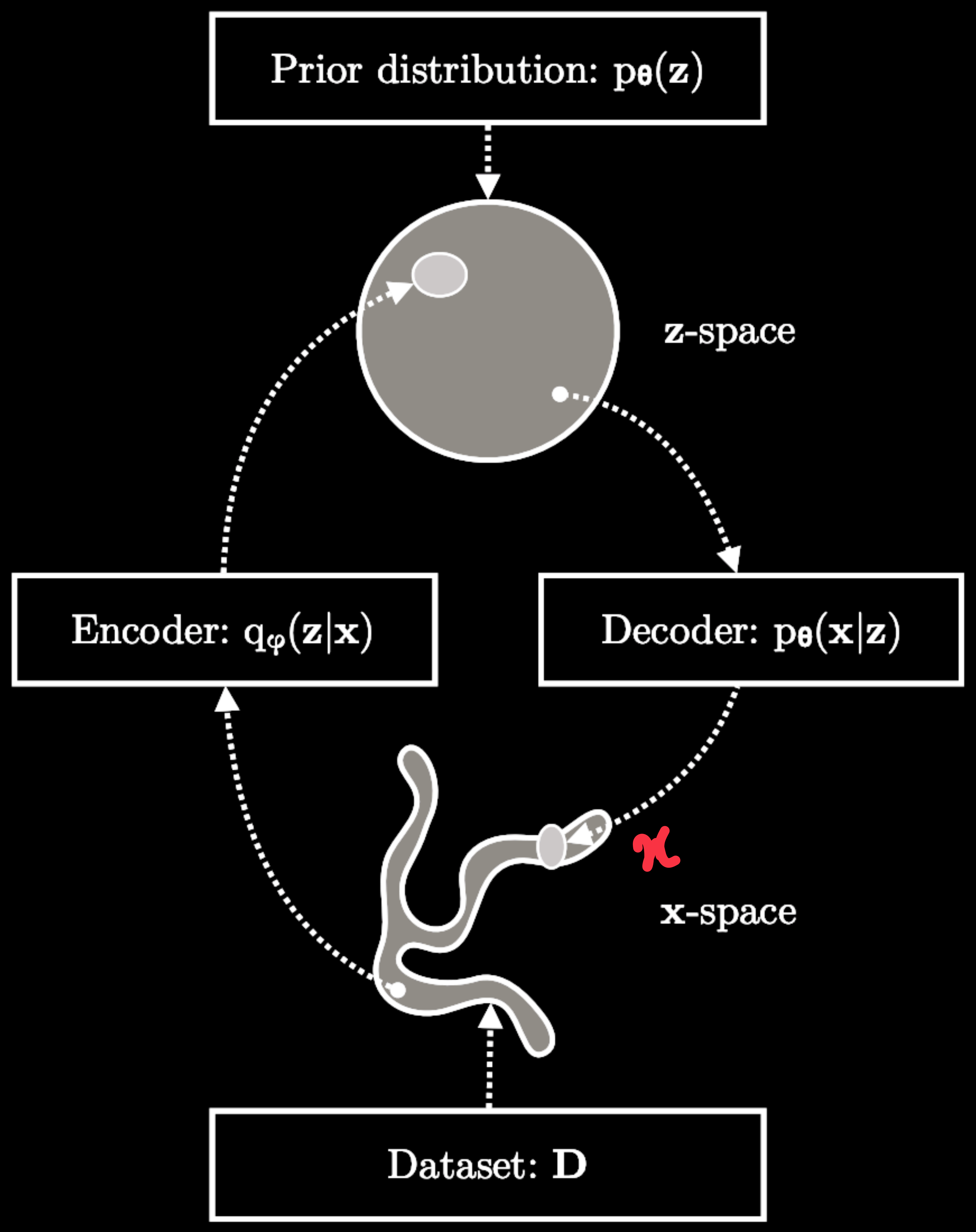


The VAE model

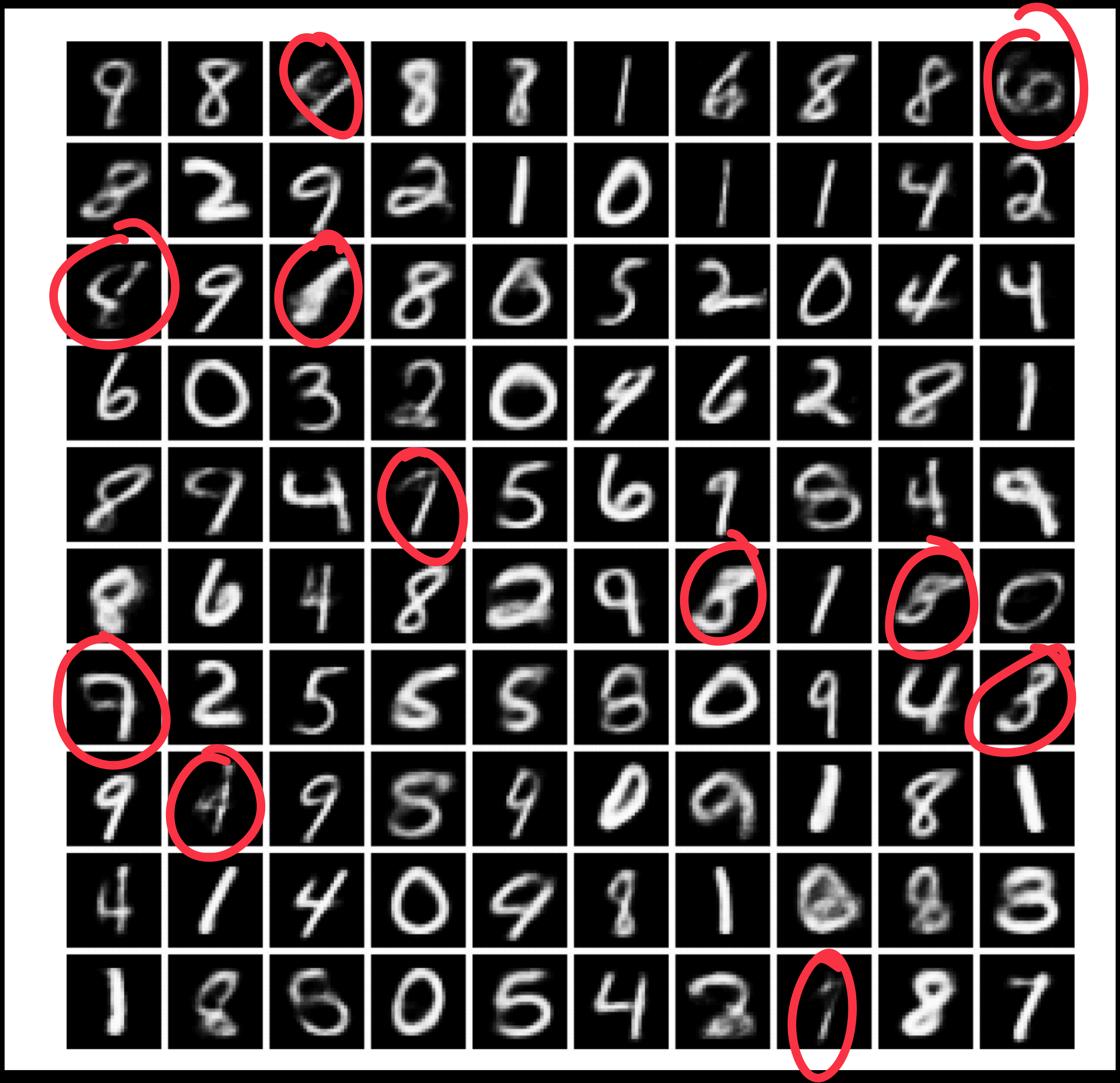


The big picture

VAE



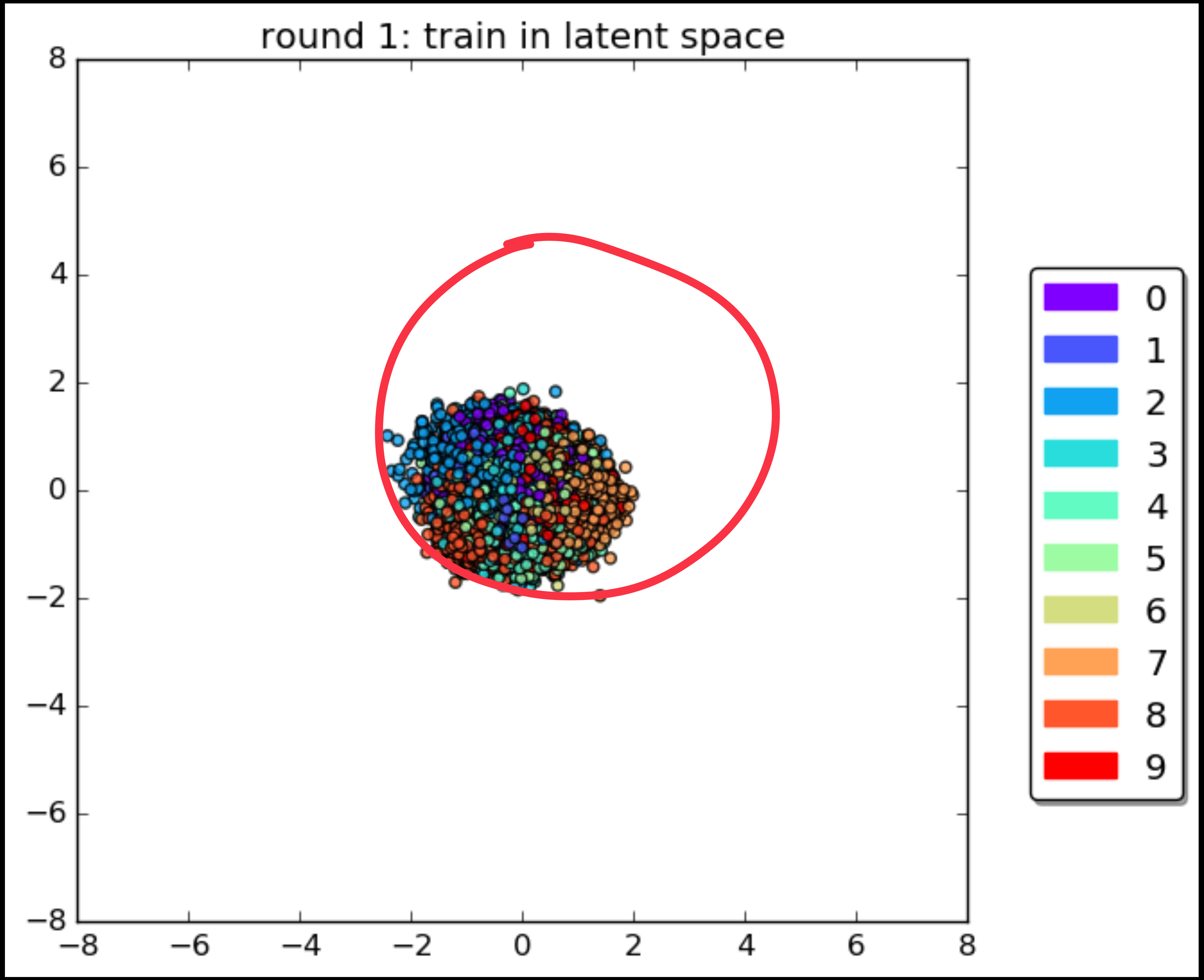
Some examples



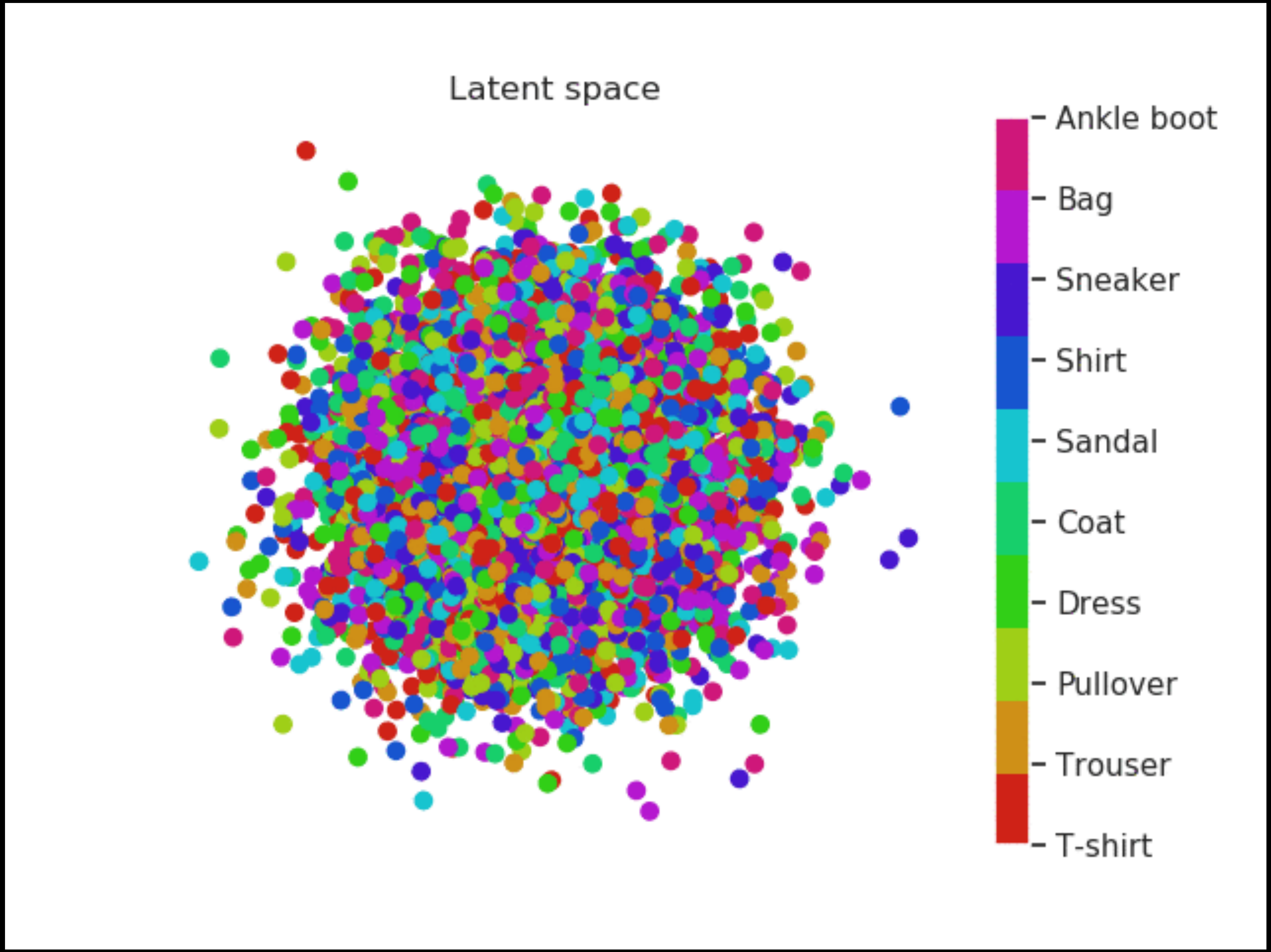
Visualizing the latent space

$d = 2$

Convolutional VAE



Visualizing the latent space



More examples

(a) 2-D latent space (b) 5-D latent space (c) 10-D latent space (d) 20-D latent space

Reading Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).