

# Auto-Encoding Variational Bayes

Diederik P (Durk) Kingma, Max Welling  
University of Amsterdam  
Ph.D. Candidate, advised by Max



Durk Kingma



Max Welling

# Problem class

- Directed graphical model:

$\mathbf{x}$  : observed variable

$\mathbf{z}$  : latent variables (continuous)

$\theta$  : model parameters

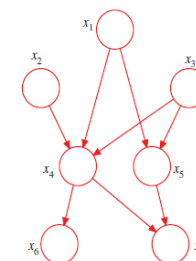
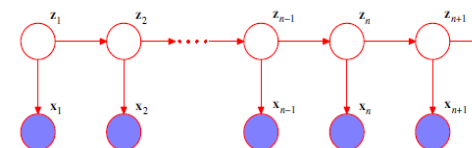
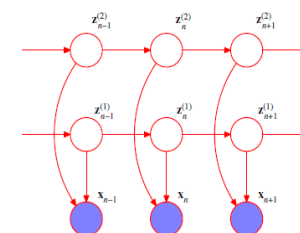
$p_{\theta}(\mathbf{x}, \mathbf{z})$ : joint PDF

- Factorized, differentiable

- Hard case: **intractable posterior distribution**  $p_{\theta}(\mathbf{z}|\mathbf{x})$

e.g. neural nets as components

- We want **fast approximate posterior inference** per datapoint
  - After inference, learning params is easy



# Approximate Inference/Learning methods

- MCMC / Monte Carlo EM
  - often too slow / scaling issues
- Wake-Sleep
  - Improper
- Why not pure MAP / Maximization?
  - Heavily overfits with high dimensionality

# Auto-Encoding Variational Bayes

Idea:

- **Learn neural net to approximate the posterior**
  - $q_{\phi}(z|x)$  with 'variational parameters'  $\phi$
  - one-shot approximate inference
  - akin to the recognition model in Wake-Sleep
- **Construct estimator of the variational lower bound**  
which we can optimize jointly w.r.t.  $\phi$  jointly with  $\theta$ 
  - > Stochastic gradient ascent

# Variational Lower Bound of the marg. lik.

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

where  $\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})]$

# Monte Carlo estimator of the variational bound

Shorthand:

$$f_{\theta, \phi}(\mathbf{z}) = \log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})$$

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [f_{\theta, \phi}(\mathbf{z})] \simeq \frac{1}{L} \sum_{l=1}^L f_{\theta, \phi}(\mathbf{z}^{(l)})$$

where  $\mathbf{z}^{(l)} \sim q_{\phi}(\mathbf{z}|\mathbf{x})$  (samples)

Can we differentiate through the sampling process w.r.t.  $\phi$  ?

# Key reparameterization trick

**Construct samples  $z \sim q_\varphi(z|x)$  in two steps:**

1.  $\varepsilon \sim p(\varepsilon)$       (*random seed independent of  $\varphi$* )

2.  $z = g(\varphi, \varepsilon, x)$  (differentiable perturbation)

such that  $z \sim q_\varphi(z|x)$  (the correct distribution)

Examples:

- if  $q(z|x) \sim N(\mu(x), \sigma(x)^2)$   
     $\varepsilon \sim N(0, I)$   
     $z = \mu(x) + \sigma(x) * \varepsilon$
- (approximate) Inverse CDF
- Much more possibilities (see paper)

# SGVB estimator

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) &= \int q_{\boldsymbol{\phi}}(\mathbf{z}) [\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{z})] d\mathbf{z} \\ &\simeq \frac{1}{L} \sum_{l=1}^L \left( \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}^{(l)}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}^{(l)}) \right)\end{aligned}$$

where  $\boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon})$  (samples from noise variable)

$$\mathbf{z}^{(l)} = g(\boldsymbol{\epsilon}^{(l)}, \boldsymbol{\phi})$$

(such that  $\mathbf{z}^{(l)} \sim q_{\boldsymbol{\phi}}(\mathbf{z})$ )

Really simple and appropriate for differentiation w.r.t.  $\boldsymbol{\phi}$  and  $\boldsymbol{\theta}$ !



# Auto-Encoding Variational Bayes

## Online algorithm

repeat

$\mathbf{x} \leftarrow$  random datapoint or minibatch

$\epsilon \leftarrow$  sample from  $p(\epsilon)$

$g_{\theta}, g_{\phi} \leftarrow \nabla_{\theta, \phi} \tilde{\mathcal{L}}(\theta, \phi; \mathbf{x}, g(\epsilon, \phi))$

$\theta \leftarrow \theta + \alpha \cdot g_{\theta}$

$\phi \leftarrow \phi + \alpha \cdot g_{\phi}$

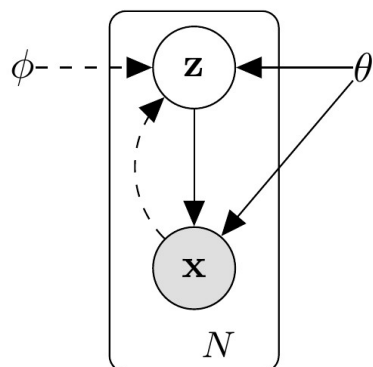
until convergence

Backprop  
(Torch7 / Theano)

e.g. Adagrad

**Scales to very large datasets!**

# Model used in experiments



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

$$p_{\theta}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mu(\mathbf{z}), \sigma(\mathbf{z})\mathbf{I})$$

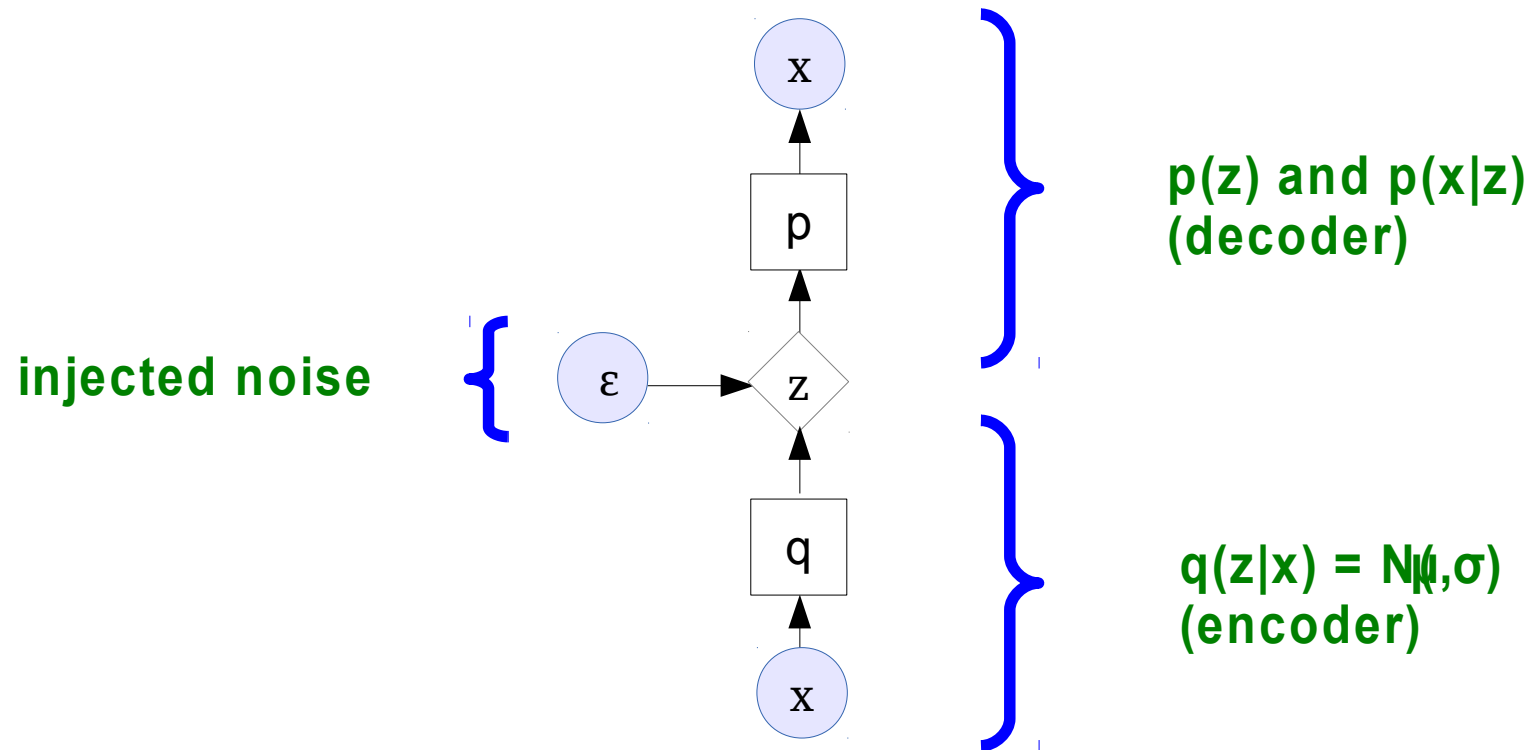
$$q_{\phi}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu(\mathbf{x}), \sigma(\mathbf{x})\mathbf{I})$$

$$\tilde{\mathcal{L}}(\theta, \phi; \mathbf{x}) = \underbrace{\log p_{\theta}(\mathbf{x}|\mathbf{z}^{(l)})}_{\text{(noisy) negative reconstruction error}} + \underbrace{\log p_{\theta}(\mathbf{z}^{(l)}) - \log q_{\phi}(\mathbf{z}^{(l)}|\mathbf{x})}_{\text{regularization terms}}$$

(noisy) negative reconstruction error      regularization terms

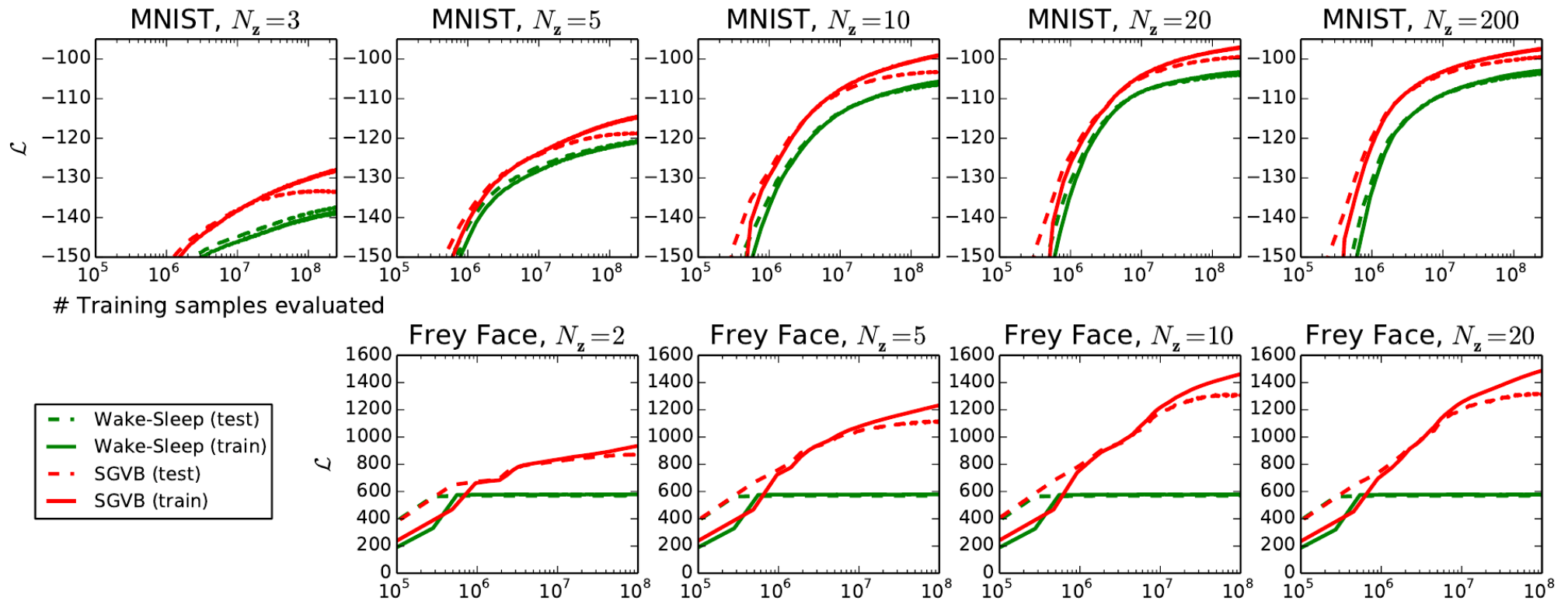
where  $\mathbf{z}^{(l)} \sim q_{\phi}(\mathbf{z}|\mathbf{x})$

# Variational auto-encoder

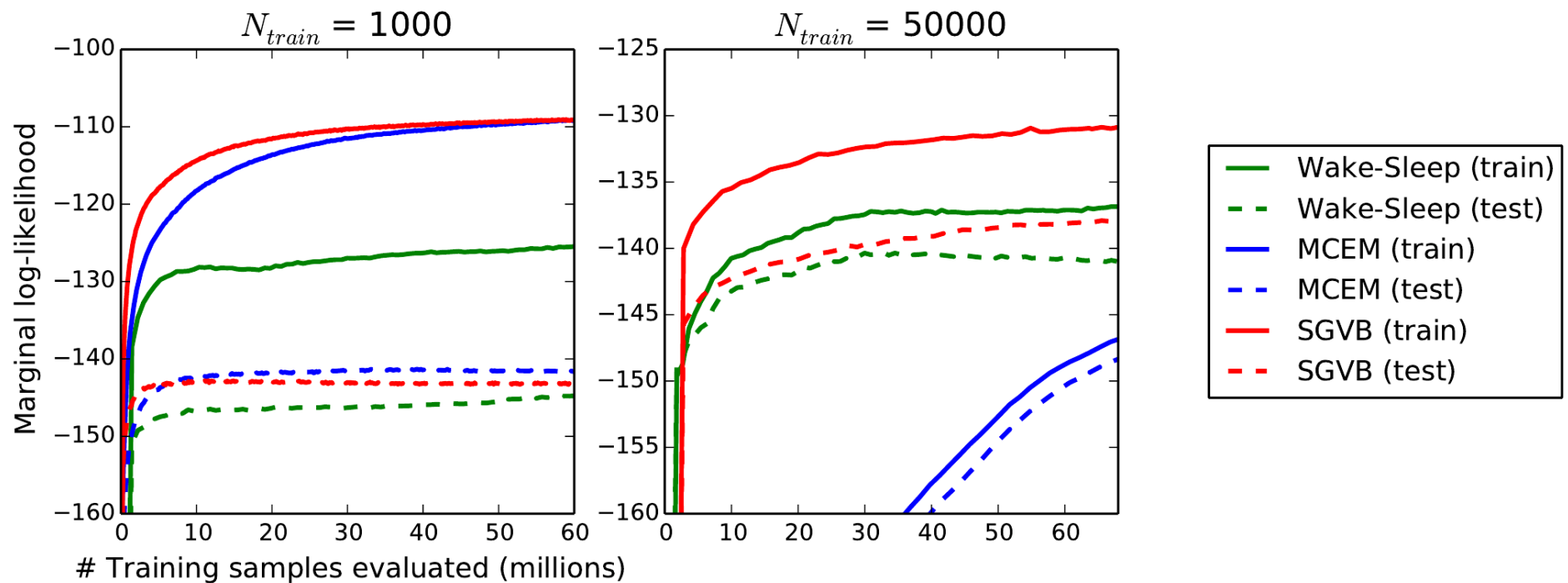


# Results:

## Marginal likelihood lower bound

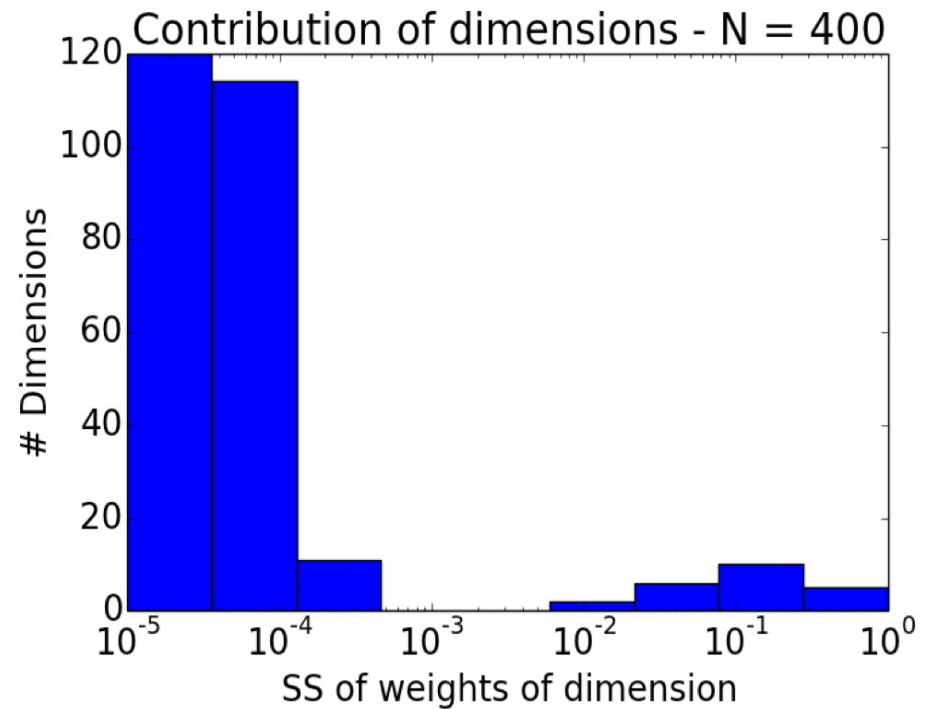
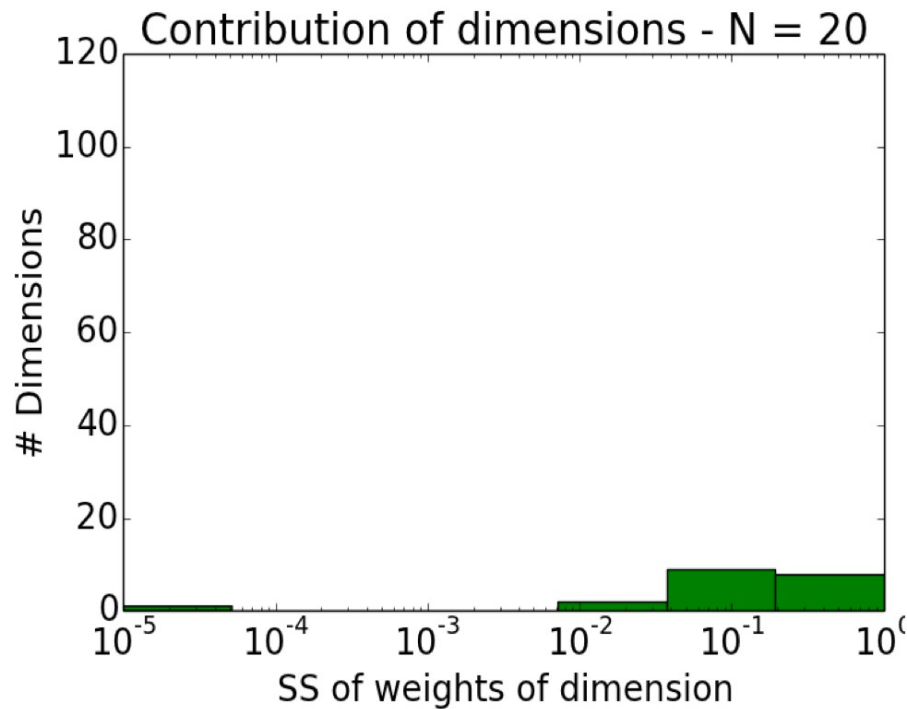


# Results: Marginal log-likelihood



**Monte Carlo EM does not scale well to large datasets**

# Robustness to high-dimensional latent space



# Samples from MNIST (simple ancestral sampling)



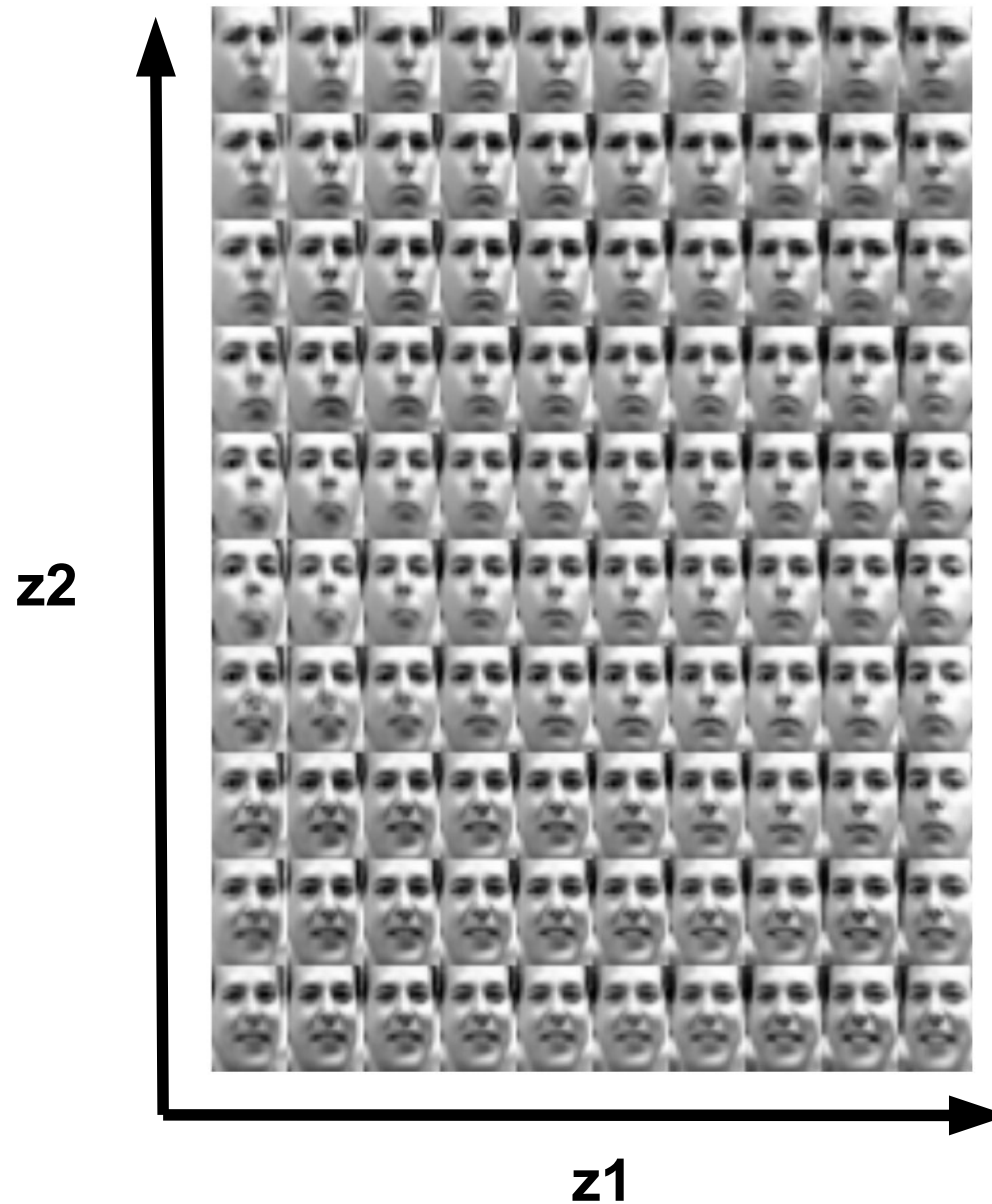
(a) 2-D latent space

(b) 5-D latent space

(c) 10-D latent space

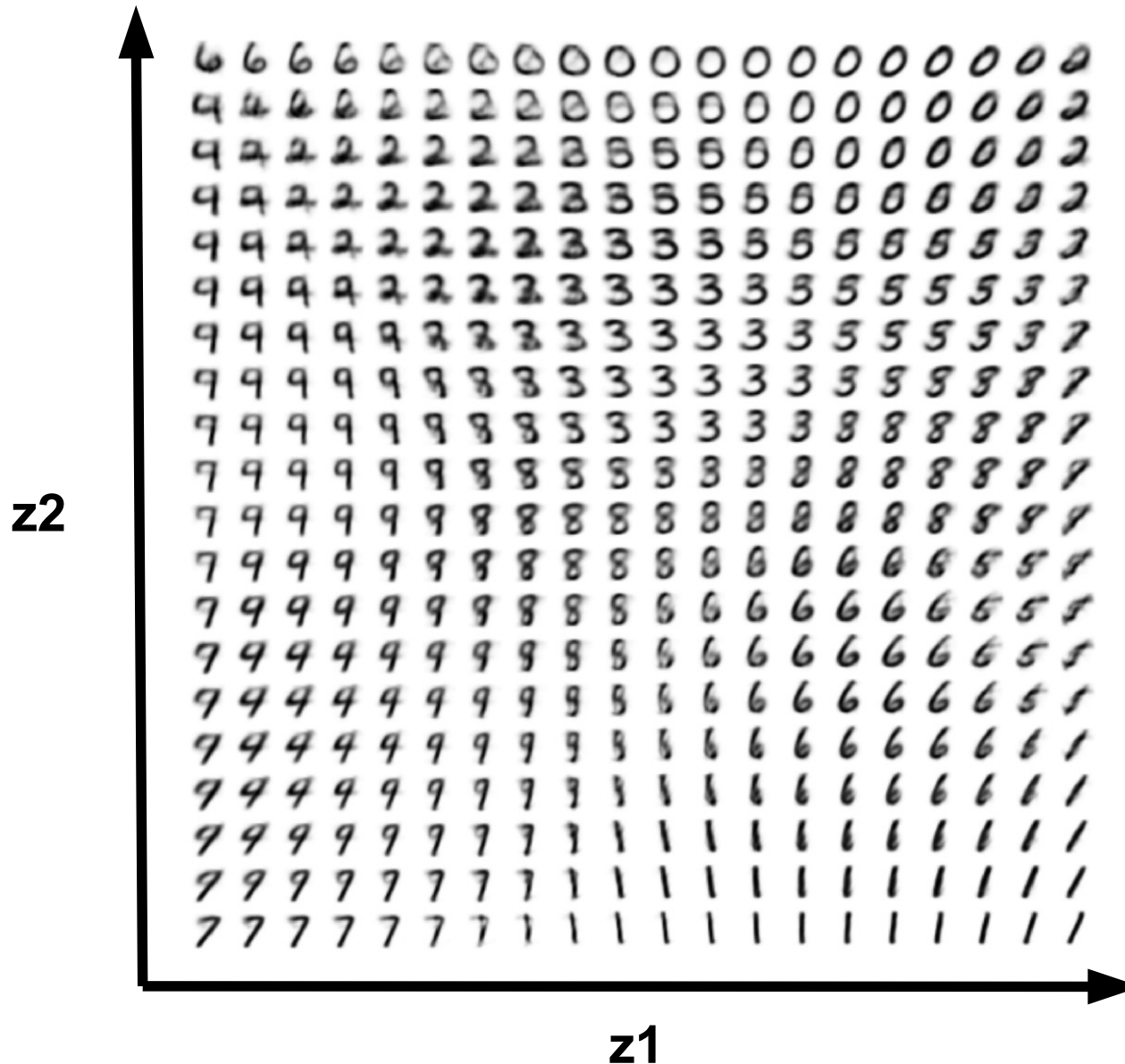
(d) 20-D latent space

# 2D Latent space: Frey Face

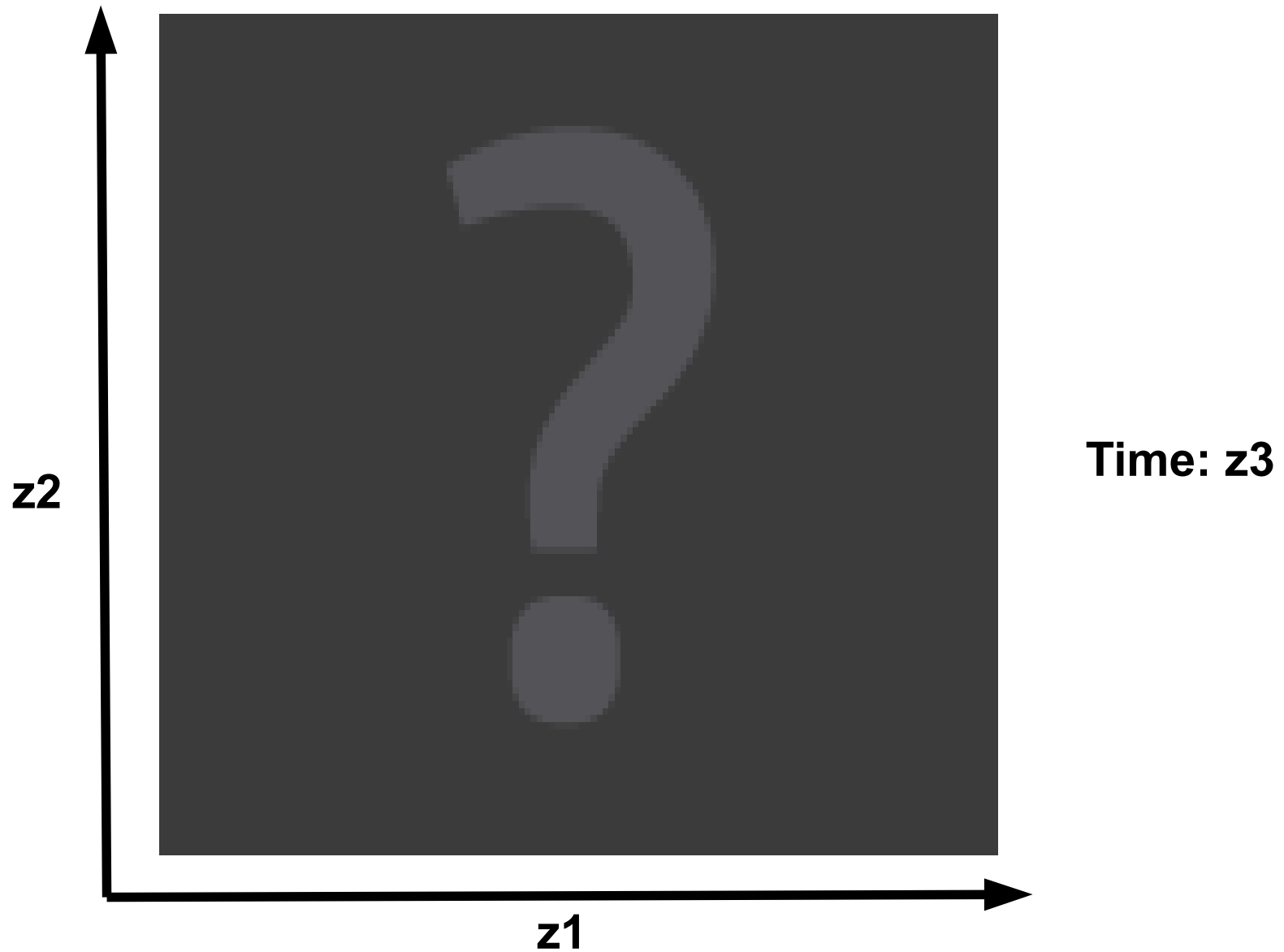




# 2D Latent space: MNIST



# 3D latent space: MNIST



# Labeled Faces in the Wild (random samples from generative model)



# Potential applications

- Representation learning
- Deep generative models of images, video, audio
- Optimal compression (bits-back coding)
- Broader applications of SGVB estimator:  
e.g. learning posterior of the global parameters
- Also see very recent paper:  
“Stochastic Back-propagation and Variational Inference in Deep Latent Gaussian Models”  
[\[Danilo J. Rezende, Shakir Mohamed, Daan Wierstra, 2014\]](#)



# Conclusion

- **Auto-Encoding Variational Bayes**
  - Applies to almost any directed model with continuous latent variables
  - Optimizes a lower bound of the marginal likelihood
  - Scales to very large datasets
  - Simple
  - Fast



**Thanks!**

<https://github.com/y0ast/Variational-Autoencoder.git>