Generative Models for Images

Yuchen Liang

REU Summer 2025

The Ohio State University
COLLEGE OF ENGINEERING

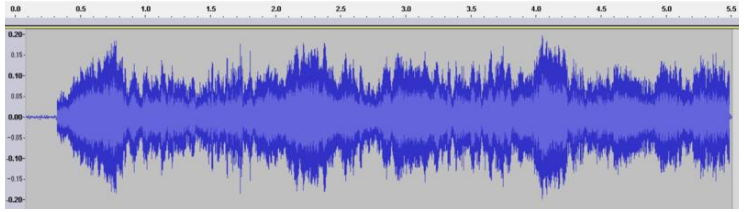
Agenda

1. **What are generative models? Examples?**
2. How to evaluate generative models for images?
3. Variational Auto-Encoder (VAE)

Introduction

Challenge: understand complex, unstructured inputs

Computer Vision



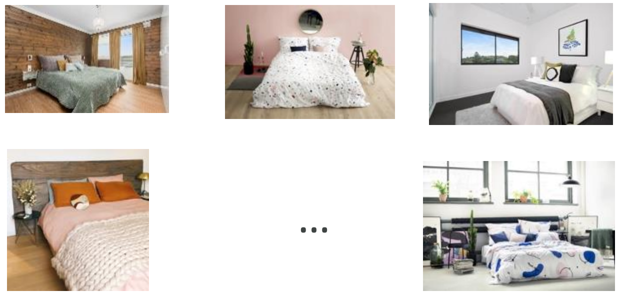
Computational Speech

Natural Language Processing

Robotics

Statistical Generative Models

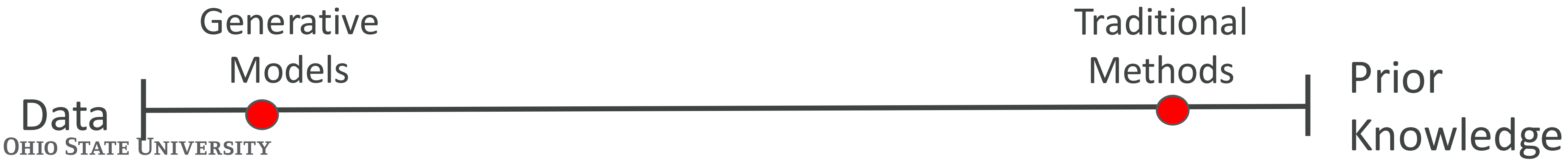
Statistical generative models are **learned from data**



Data
(e.g., images of bedrooms)

Prior Knowledge
(e.g., physics, materials, ..)

Priors are always necessary, but there is a spectrum



Statistical Generative Models

A statistical generative model is a **probability distribution** p(x)

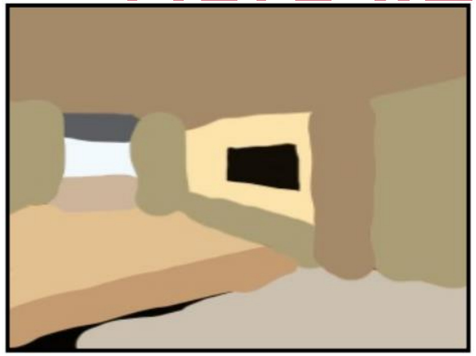
- **Data:** samples (e.g., images of bedrooms)
- **Prior knowledge:** parametric form (e.g., Gaussian?), loss function, optimization algorithm, etc.

Image x  → A probability distribution p(x) → scalar probability p(x)

It is generative because **sampling from p(x) generates new images**

 … 

Data generation in the real world



Generate

Generative model
of realistic images

Stroke paintings to realistic images
[Meng, He, Song, et al., ICLR 2022]

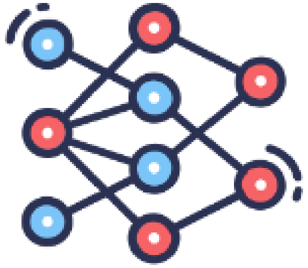
“Ace of Pentacles”

Generate

Generative model
of paintings

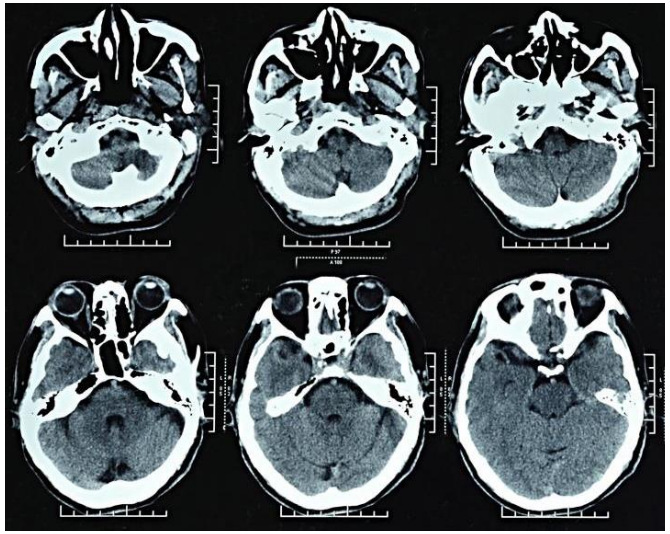
Language-guided artwork creation
https://chainbreakers.kath.io @RiversHaveWings

Solving inverse problems with generative models



Generative model
of medical images

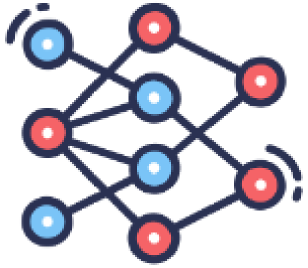
Generate



Medical image reconstruction
[Song et al., ICLR 2022]

Outlier detection with generative models

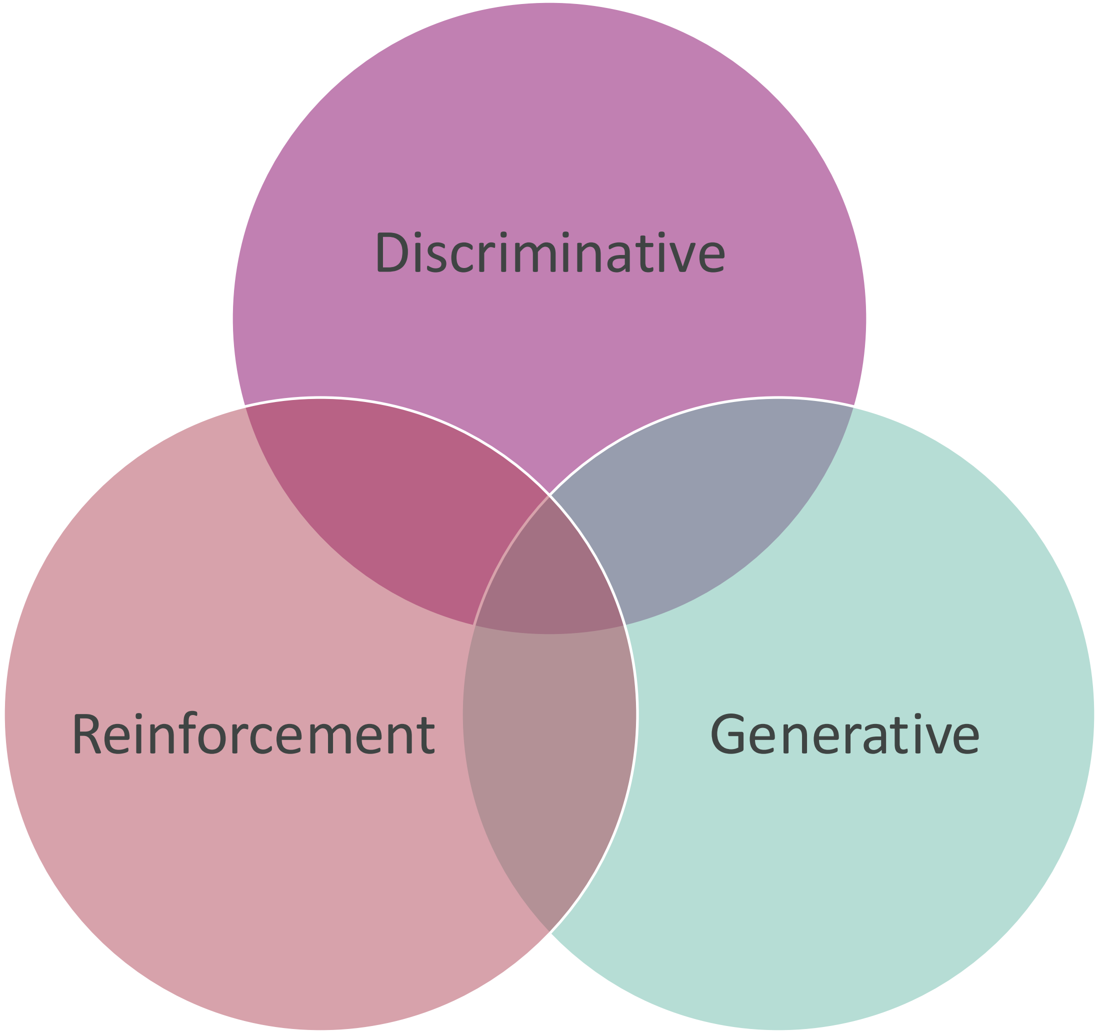
High probability



Generative model of traffic signs

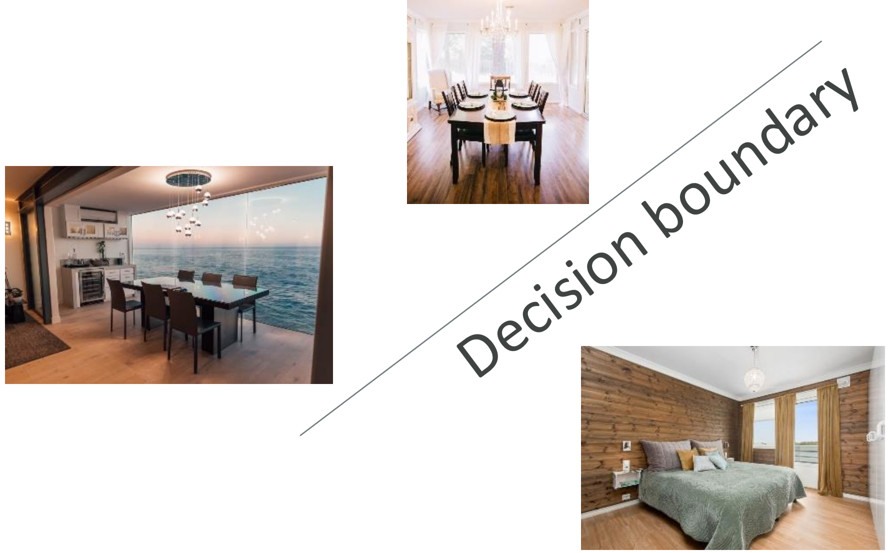
Outlier detection
[Song et al., ICLR 2018]

Category of ML Problems



Discriminative vs. generative

Discriminative: classify bedroom vs. dining room



The image X is given. **Goal**: some decision boundary

- Requires **conditional distribution over label Y**: **p(Y|X)**
- E.g.: logistic regression, convolutional neural net, etc.

Generative: generate X

Y=Dining , X= 

Y=Bedroom , X= 

The input X is **not** given. **Goal**: generate X based on label

- Requires a model of the **joint distribution over both X and Y**: **p(Y, X)**

Discriminative vs. generative

Joint and conditional are related via **Bayes Rule**:

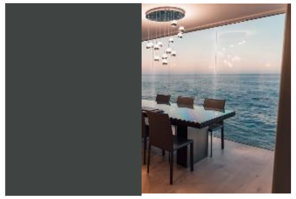
P(Y = Bedroom | X= ) =

P(Y = Bedroom, X= ) **Joint**

P(X = ) **Marginal**

Conditional

Discriminative: no need to model: P(X= )

Therefore, it cannot handle missing data: P(Y = Bedroom | X = ) ???

Images and Text

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

P(image | caption)

TEXT PROMPT a store front that has the word ‘openai’ written on it. . . .

AI-GENERATED IMAGES

Text2Image Diffusion Models

User input:

An astronaut riding a horse

Text2Image Diffusion Models

User input:

A perfect Italian meal

Text2Image Diffusion Models

User input:

泰迪熊穿着戏服，站在太和殿前唱京剧

A teddy bear, wearing a costume, is standing in front of the Hall of Supreme Harmony and singing Beijing opera

Dalle3

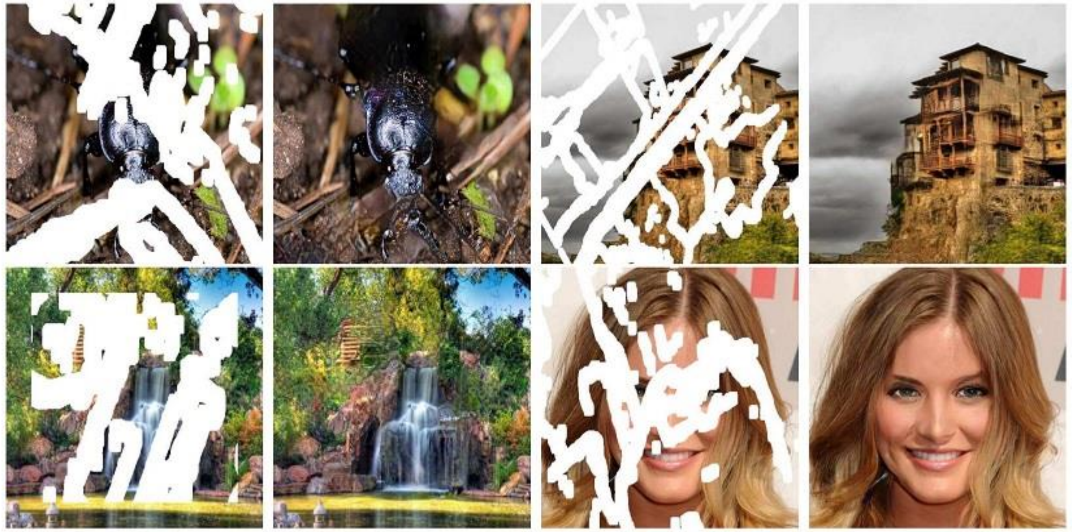
A minimap diorama of a cafe adorned with indoor plants. Wooden beams crisscross above, and a cold brew station stands out with tiny bottles and glasses

Progress in Inverse Problems

P(high resolution | low resolution)

Menon et al, 2020

P(full image| mask)



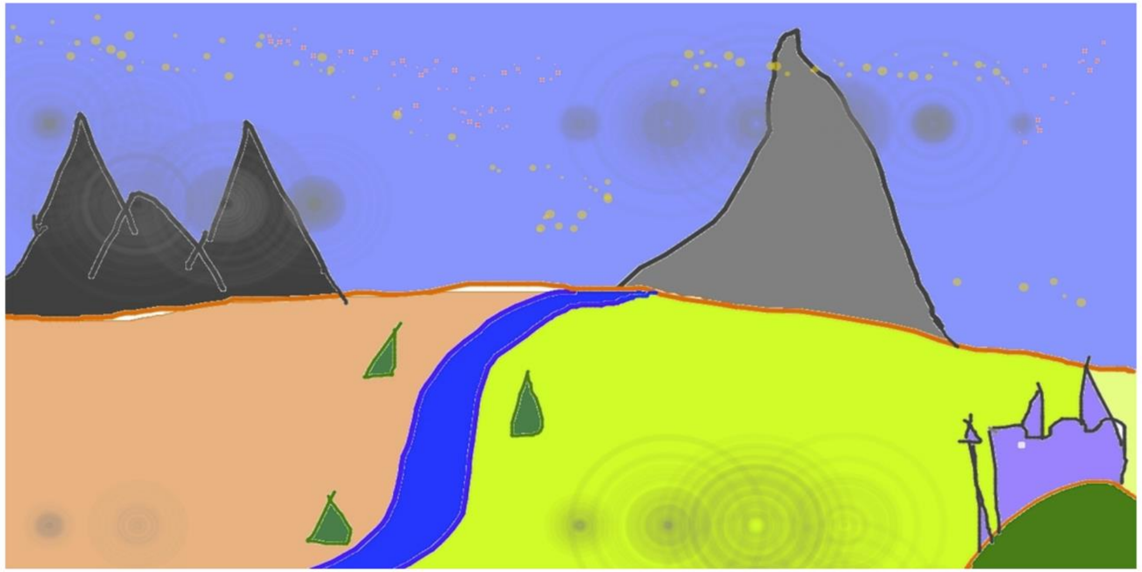
Liu al, 2018

P(color image| greyscale)

Antic, 2020

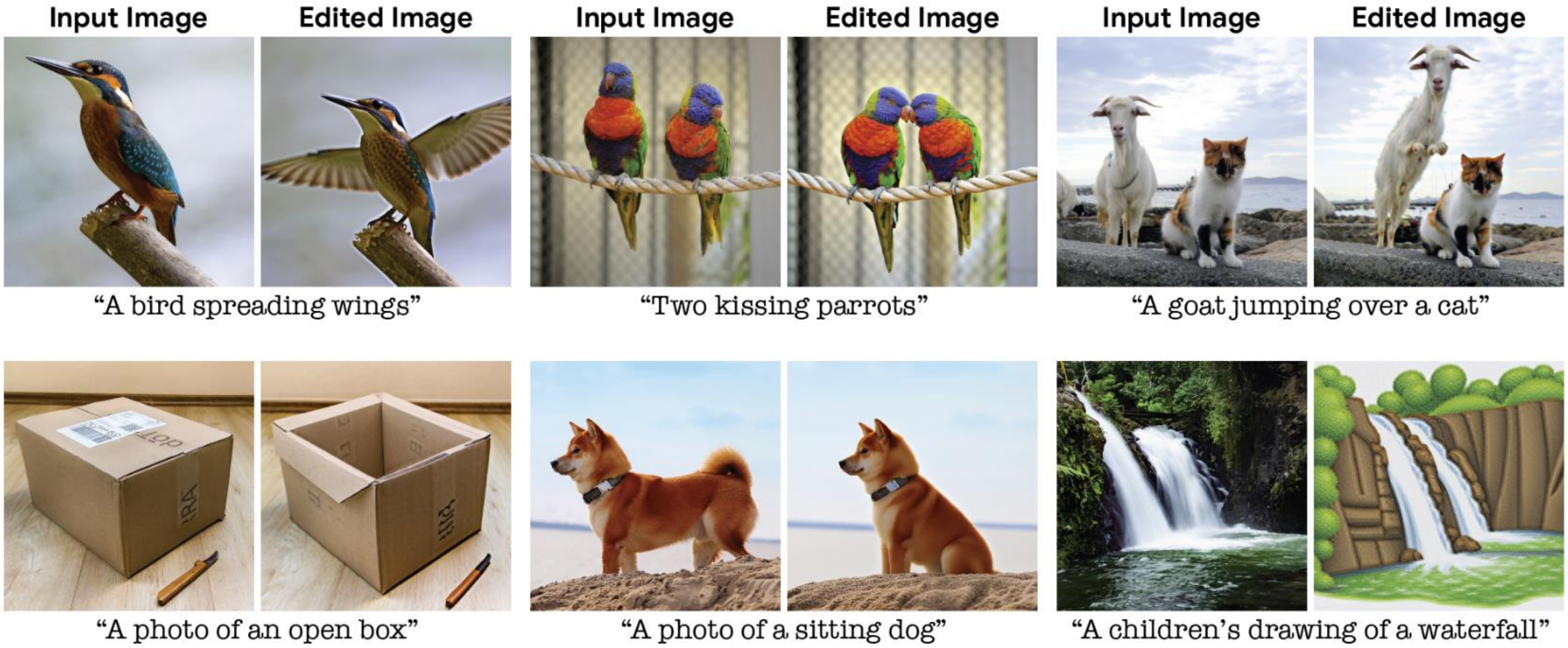
Progress in Inverse Problems

User input:



Progress in Inverse Problems

Progress in Inverse Problems



Kawar et al., 2023

Language Generation

Custom prompt

To get an A+ in deep generative models, students have to

P(next word | previous words)

Completion

To get an A+ in deep generative models, students have to be willing to work with problems that are a whole lot more interesting than, say, the ones that most students work on in class. If you're a great student, the question above can be avoided and you'll be able to do great work, but if you're not, you will need to go beyond the basics before getting good.

Now to be clear, this advice is not just for the deep-learning crowd; it is good advice for any student who is taking his or her first course in machine learning.

The key point is that if you have a deep, deep brain of a computer scientist, that's just as important to you.

Radford et al., 2019
Demo from talktotransformer.com

Machine Translation

Conditional generative model P(English text| Chinese text)

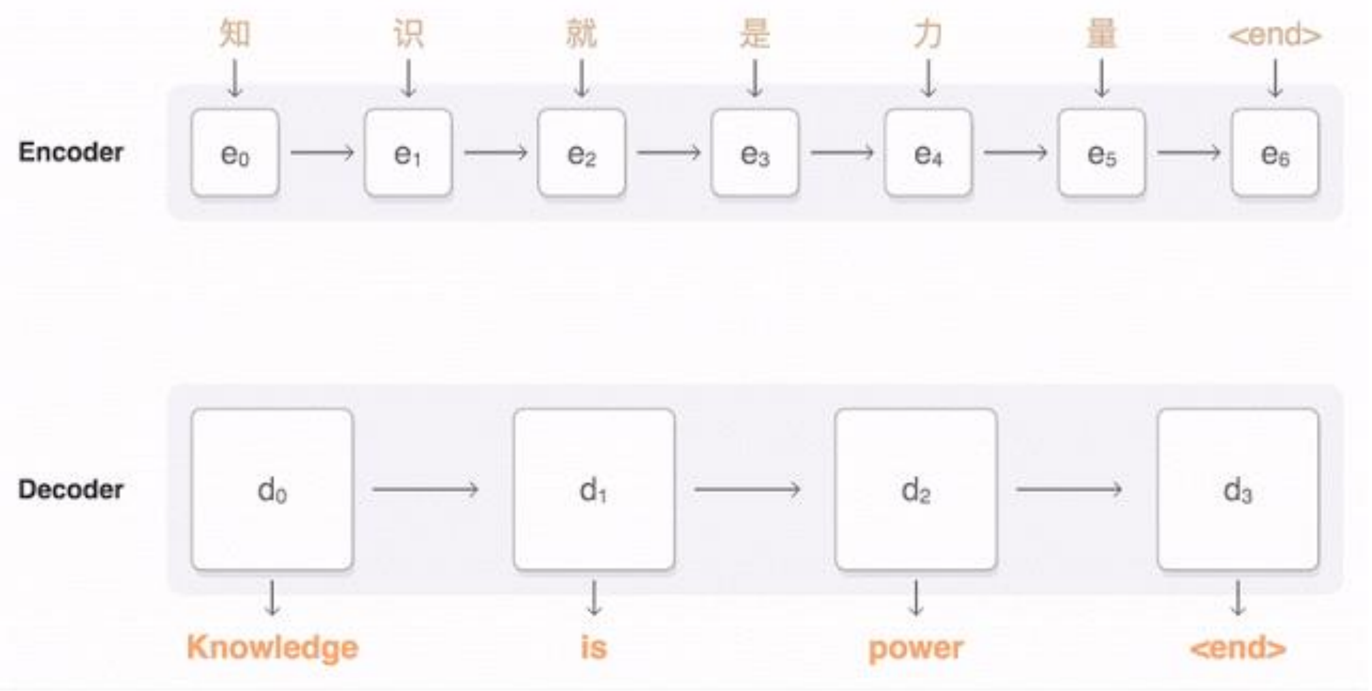


Figure from Google AI research blog.

Code Generation

parse_expenses.py | write_sql.go | sentiment.ts | addresses.rb

```python
import datetime

def parse_expenses(expenses_string):
    """Parse the list of expenses and return the list of triples (date, va
```

OpenAI Codex

Video Generation

Suddenly, the walls of the embankment broke and there was a huge flood

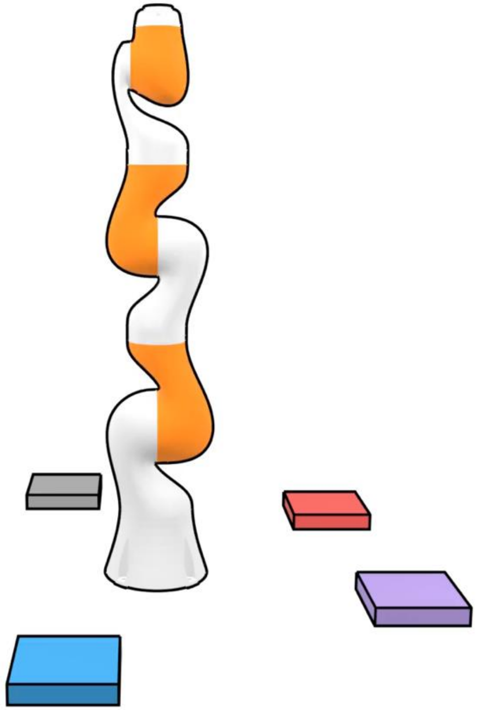
Video Generation

a couple sledding down a snowy hill on a tire roman chariot style

Imitation Learning

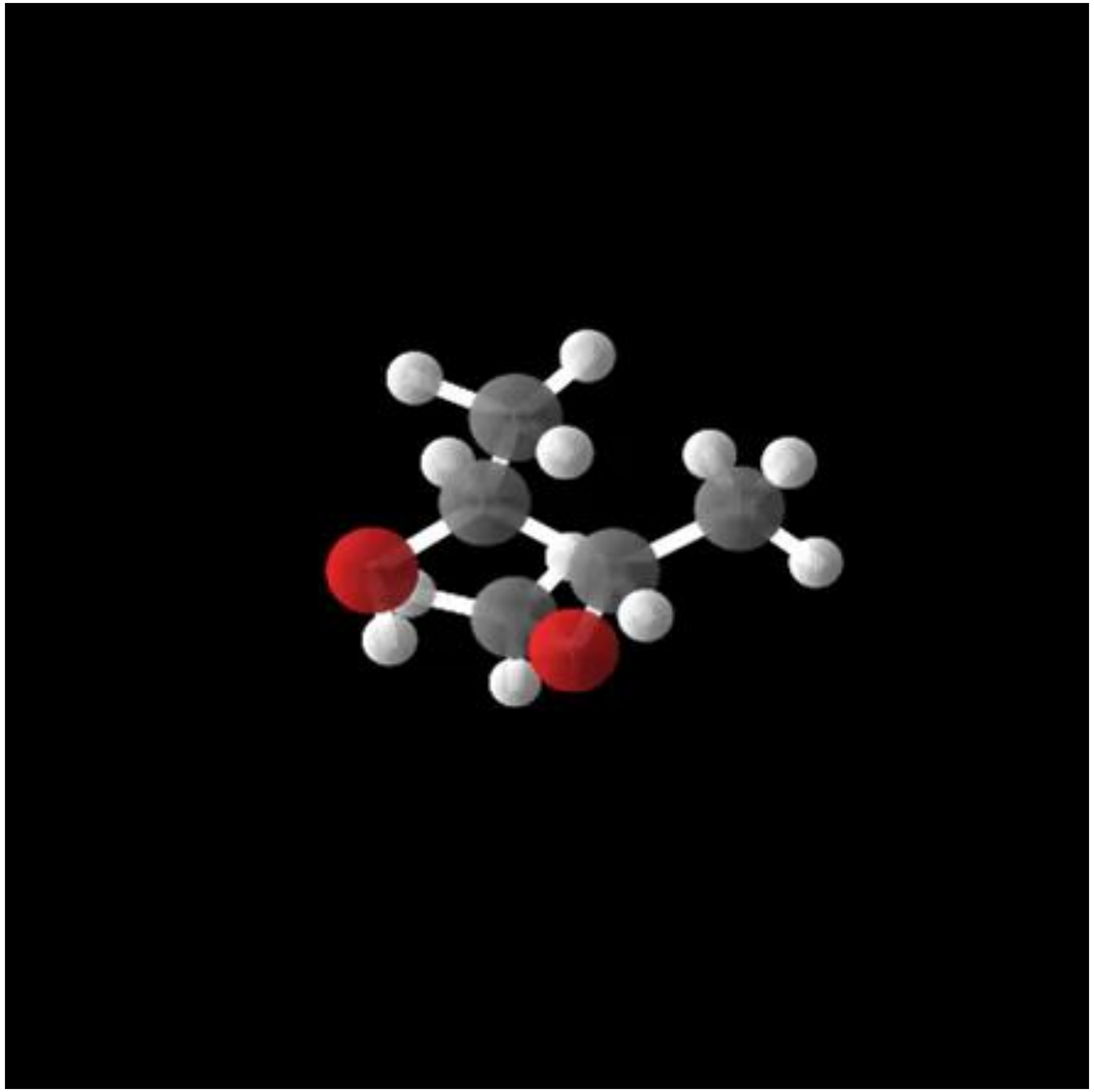
Conditional generative model P(actions | past observations)

Li et al., 2017



Janner et al., 2022

Molecule generation



From Discrimitive to Generative Modeling

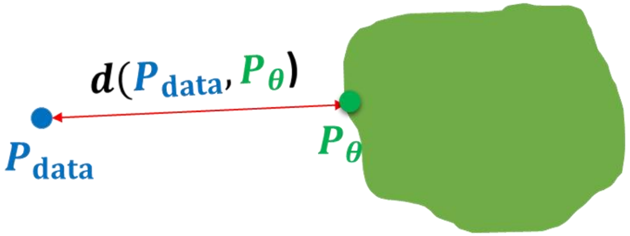
- Discrimitive ML: Classification & Regression
- What are the components for ML discrimitive models?

1. **Representation**: Need a discrimitive function from x -> y
   - E.g.: SVM/XGBoost? RNN/CNN? Transformer?
2. **Learning**: Need a loss function and a training algorithm
   - Loss: Cross-ent, Euclidean distance…
   - Algorithm: GD, SGD, mini-batches…

Key Challenges

1. **Representation**: How do we model the **marginal/joint distribution** of many random variables?
2. **Learning**: What is the right way to **compare probability distributions**?

$\mathbf{x}_i \sim P_{\text{data}}$
$i = 1, 2, \dots, n$



$\theta \in M$

Model family

- **Other challenges:** How to obtain a sample? How to evaluate the performance?

Agenda

1. What are generative models? Examples?
2. **How to evaluate generative models for images?**
3. Variational Auto-Encoder (VAE)

Typical strategy for discriminative models

- Includes classification and regression

1. Obtain a test dataset with data $x_1, \ldots, x_N$ and labels $y_1, \ldots, y_N$
2. Evaluate the model for the estimated labels $\hat{y}_i = f_\theta(x_i)$
3. Compare $y_1, \ldots, y_N$ and $\hat{y}_1, \ldots, \hat{y}_N$ with respect to some distance metric
   - Binary/Categorical -> cross-entropy; Real-valued -> Avg Euclidean distance, etc.

Difficulty for generative models

1. We can still obtain a test dataset with data $x_1, \dots, x_N$, but **no labels**
2. We can generate $\tilde{x}_1, \dots, \tilde{x}_N$, but **no matching guarantee** that $x_i \approx \tilde{x}_i$

- Attempt 1: Use human judgement
  - Humans are **costly**
  - No uniform criteria…
- Attempt 2: Evaluate $\mathbb{E}_{X \sim p_{test}}[\log p_\theta(X)]$
  - Hard to compute in real-world…

Evaluations for Images

- Candidate 1: Inception Scores
- Assumptions
  1. The generative model is trained on some **labelled** dataset
  2. We have a well-trained **classifier** $c(y|x)$
- Inception Score = Sharpness (S) * Diversity (D)
  - …the higher the better

$$IS = S * D$$

Inception Score Example

Low sharpness

High sharpness

$$S = \exp\left(E_{\mathbf{x}\sim p}\left[\int c(y|\mathbf{x}) \log c(y|\mathbf{x}) \mathrm{d}y\right]\right)$$

Low diversity

High diversity

$$D = \exp\left(-E_{\mathbf{x}\sim p}\left[\int c(y|\mathbf{x}) \log c(y) \mathrm{d}y\right]\right)$$

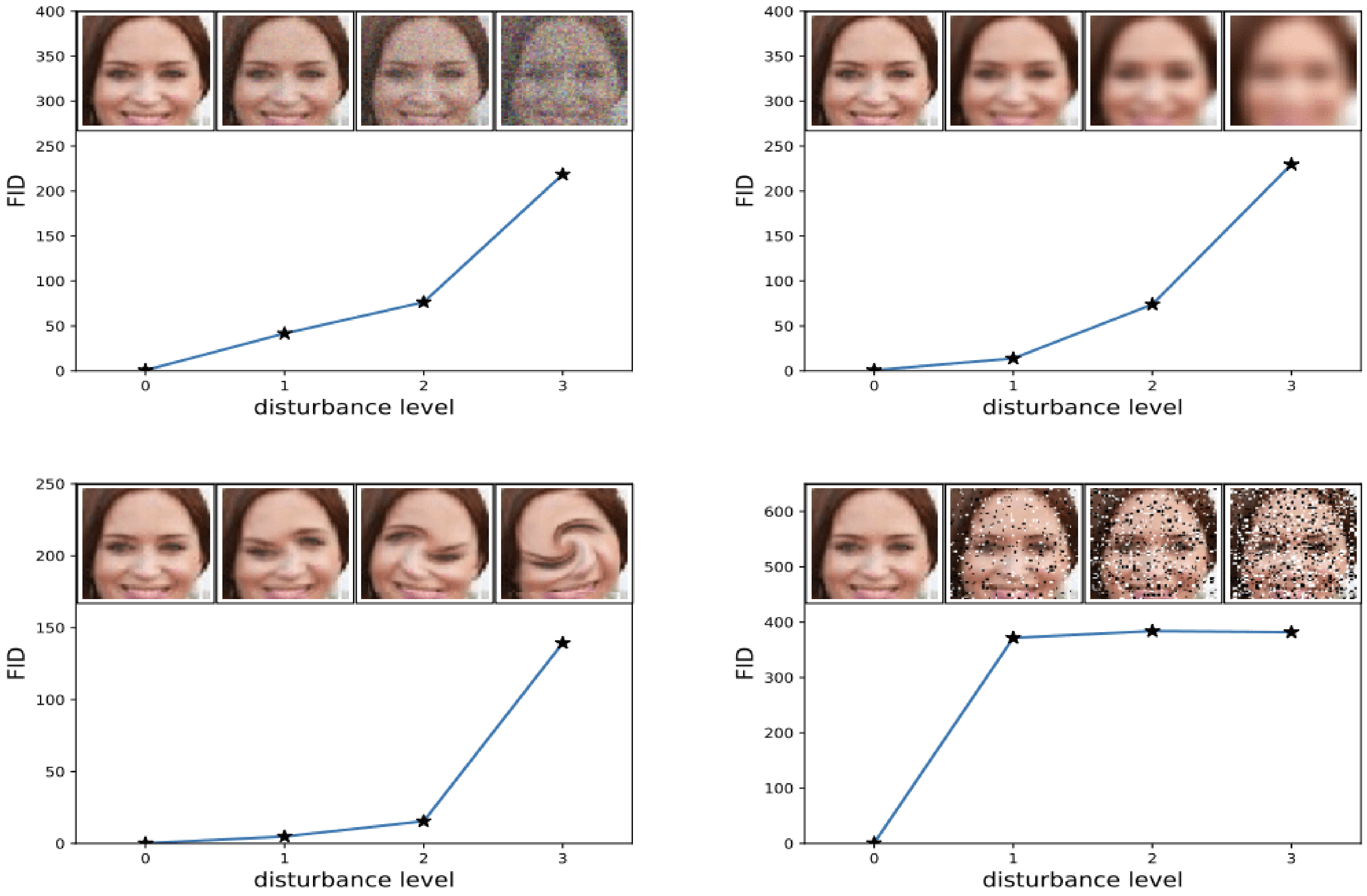
S. Ermon, Stanford CS 236 Lecture 15.

Evaluations for Images (cont.)

- Candidate 2: Fréchet Inception Distance (FID)
  - …the lower the better

1. Step 1: Let $\mathcal{G}$ denote the generated samples, and let $\mathcal{T}$ denote the test dataset
2. Step 2: Compute the feature vectors $\mathrm{F}_{\mathcal{G}}$ and $\mathrm{F}_{\mathcal{T}}$ (usually the last pooling layer of Inception v3, which is 2048-dim)
3. Step 3: Fit a multivariate Gaussian for each as $(\mu_{\mathcal{G}}, \Sigma_{\mathcal{G}})$ and $(\mu_{\mathcal{T}}, \Sigma_{\mathcal{T}})$
4. Step 4: The FID is equal to **Wasserstein-2** distance between them

FID e



https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/

Evaluations for Images (cont.)

- Candidate 3: (specifically for **image recovery**)
  1. MSE (and RMSE): pixel-level matchness
  2. PSNR: MSE weighted by the maximal pixel value
  3. SSIM: product of luminance, contrast, and structure (or correlation)
- General procedure
  1. original image O -> noisy image N-> recovered image R
  2. Compare O and R with the metrics above

Recovery I

10.4236/jcc.2019.73002

Agenda

1. What are generative models? Examples?

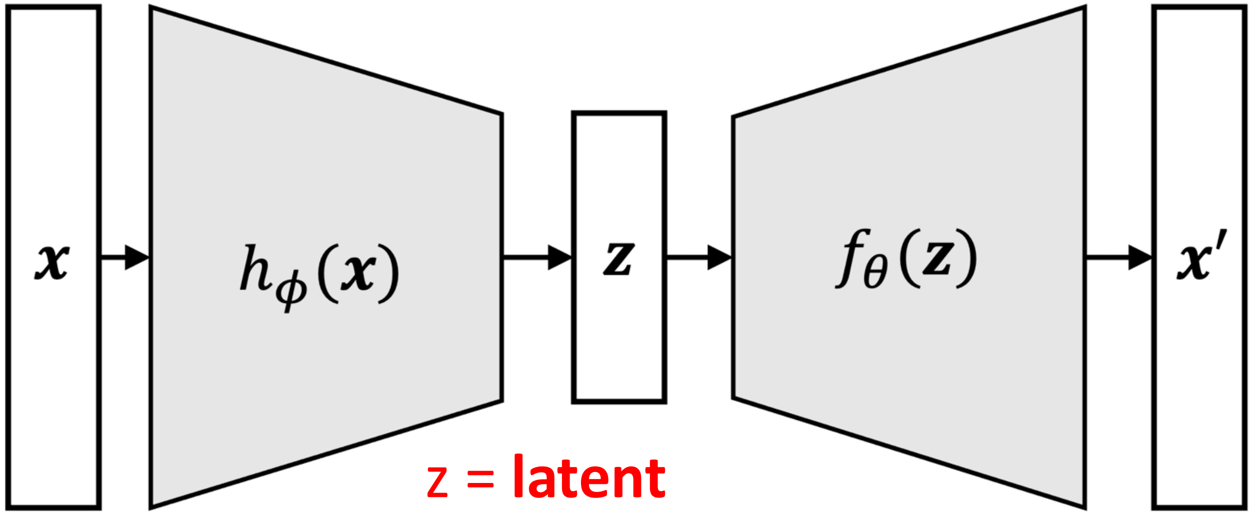
2. How to evaluate generative models for images?

3. **Variational Auto-Encoder (VAE)**

VAE: Distribution Representation

- Simply use feed-forward neural network (FFNN)
- Auto-Encoder Structure: **Encoder** network h_ϕ + **Decoder** network f_θ

Say, an image of 28x28=784 dimensions

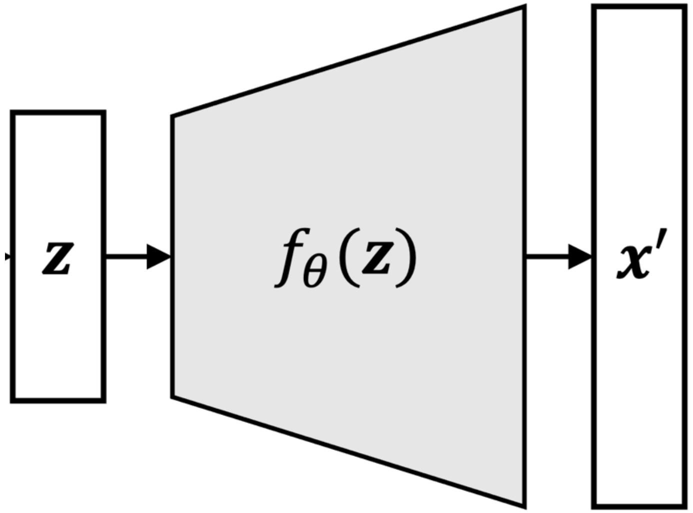


z = **latent** (compressed) data, say 50 dimensions

VAE Sampling

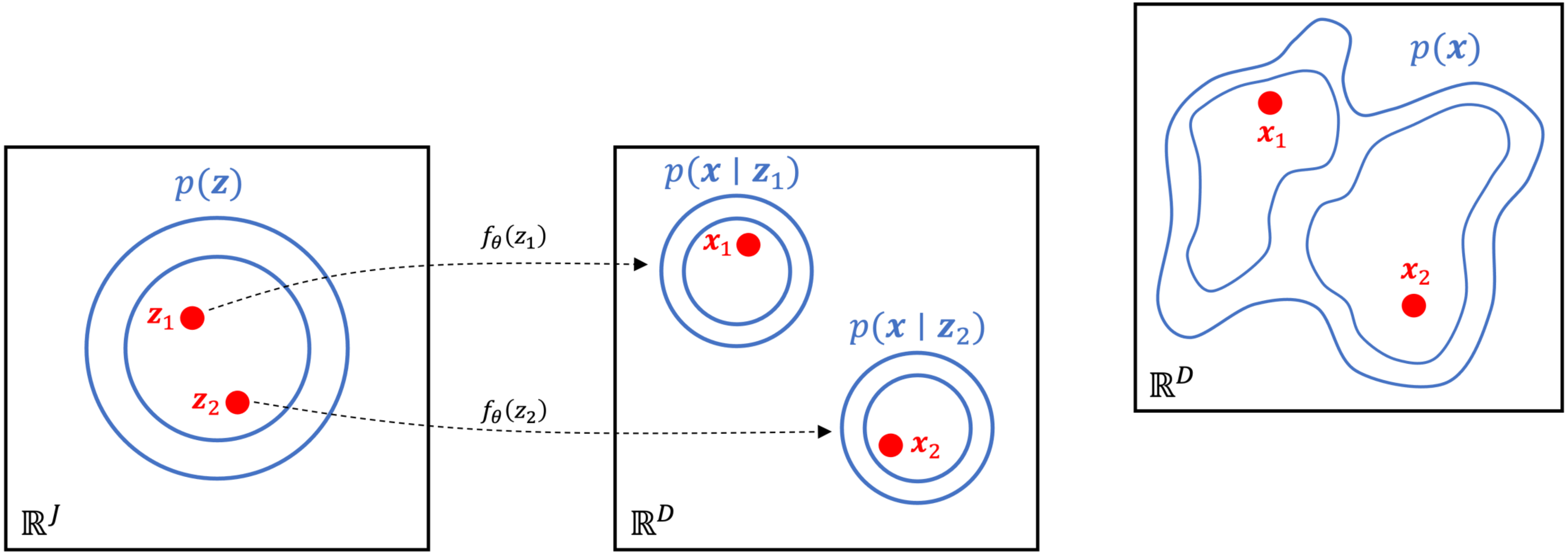
1. Generate a **random** variable $z \sim \mathcal{N}(\mu, \Sigma)$
2. One forward-pass of the **decoder** network $\mathrm{x}' = f_\theta(z)$
3. (Optional): Re-sample again using x' (say, from a localized Gaussian)

Initial **random** variable



Generated parameters (or data)

VAE Sampling



https://mbernste.github.io/posts/vae/

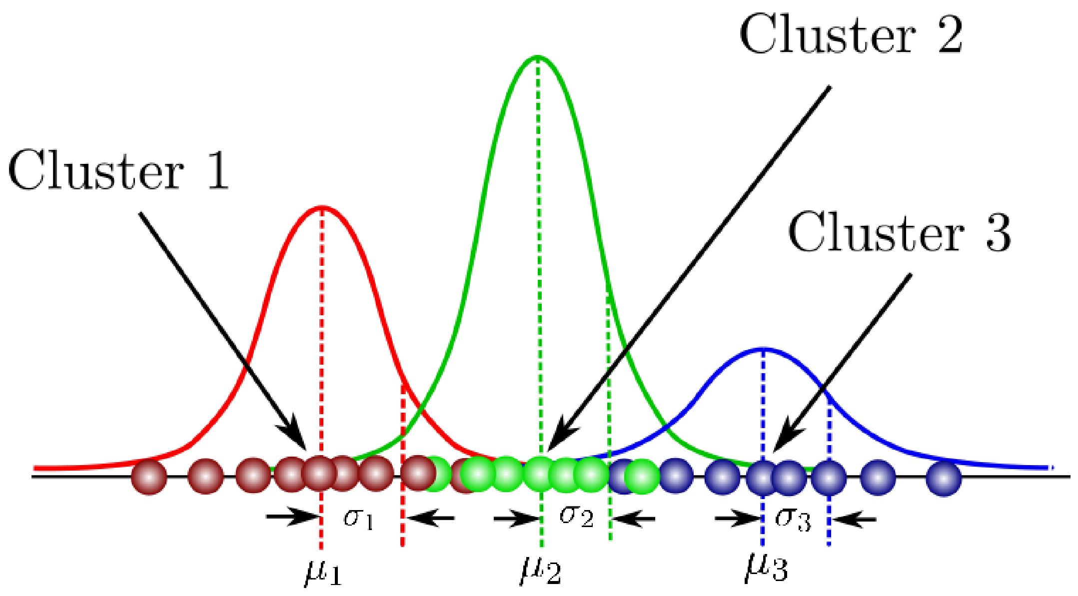
VAE: How to learn?

- By law of total probability:

Decoder (Gaussian) Latent Dist (also Gaussian)

$$p_\theta(x) = \int p_\theta(x,z)\,dz = \int p_\theta(x|z)\,p_\theta(z)\,dz$$

  - This is **Gaussian mixture**, which is very complex



VAE: How to learn? (cont.)

- Suppose we are given N data (images): $x_1, \dots, x_N$
- Goal: maximize (log-)**likelihood**:

$$\log p_\theta(x_1, \dots, x_N) = \sum_i \log p_\theta(x_i)$$

VAE: How to learn? (cont.)

- Let’s do some math…

$$\log p_\theta(x) = \log\left(\int p_\theta(x,z)dz\right)$$ (See previous slide)

$$= \log\left(\int \frac{p_\theta(x,z)}{q_\phi(z|x)} q_\phi(z|x)dz\right)$$ (Multiply and divide the same quantity)

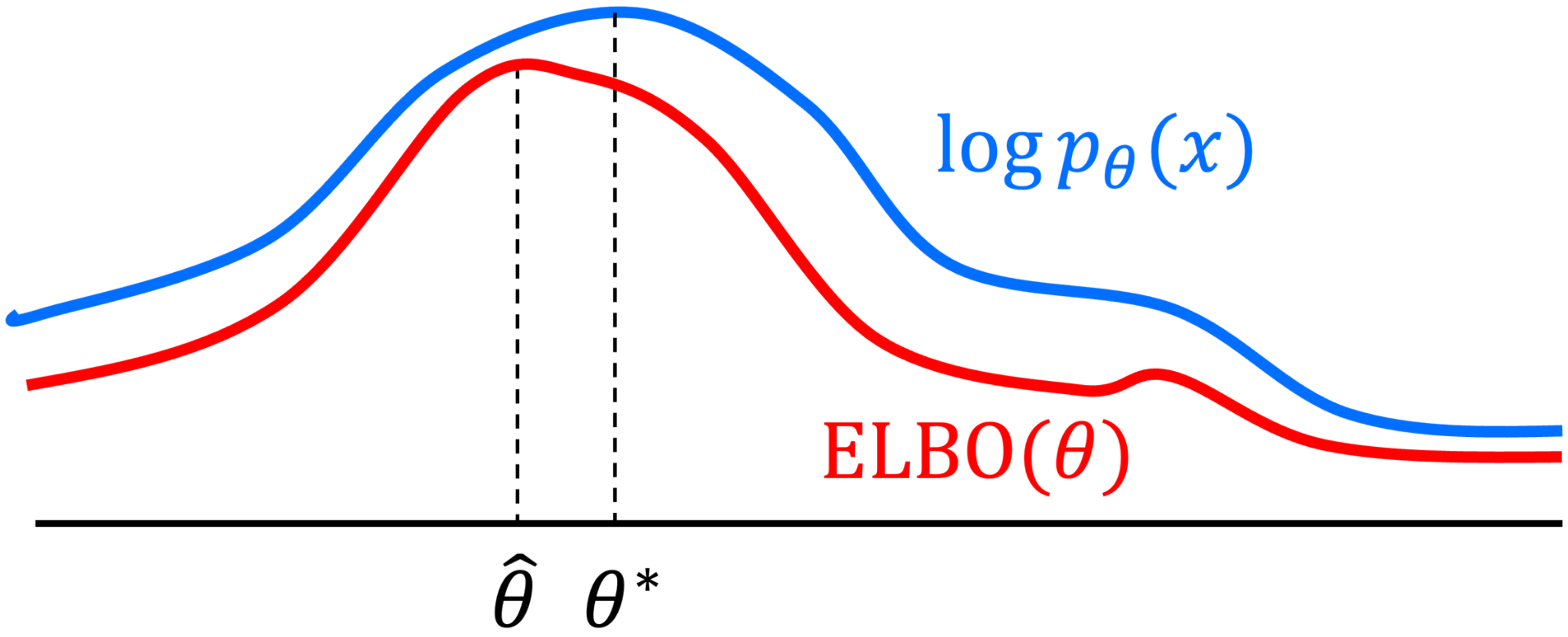
$$\geq \underbrace{\int \log \frac{p_\theta(x,z)}{q_\phi(z|x)} q_\phi(z|x)dz}_{\text{ELBO}}$$ (Jensen’s inequality)

p_θ = Decoder/Generator

q_ϕ = Encoder

- The last line is the **Evidence Lower BOund (ELBO)**

ELBO visualized



Optimization Algorithm for ELBO

- Step 1: Obtain N datapoints (possibly from a random minibatch)
- Step 2: Compute the ELBO
$$ELBO = \int \left(\log p_\theta(x, z) - \log q_\phi(z|x)\right) q_\phi(z|x)\, d\, z$$
- Step 3: Calculate the gradient w.r.t. θ and ϕ
- Step 4: Gradient Update
- Iterate Steps 2-4 until convergence…

Optimizating VAE ELBO (cont.)

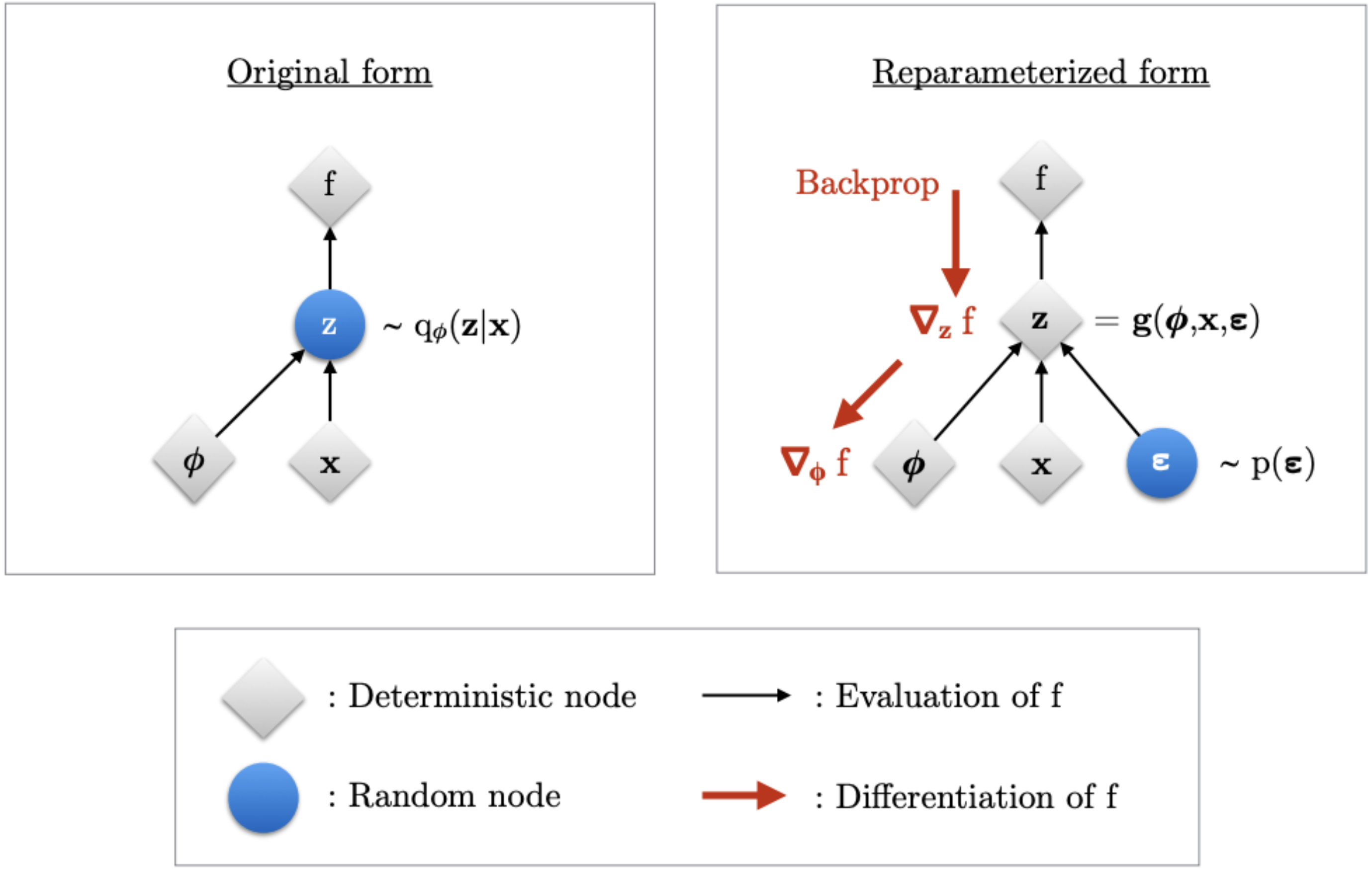
$$ELBO = \int \left(\log p_\theta(x,z) - \log q_\phi(z|x)\right) q_\phi(z|x)\, d\, z$$

- It's easy to take derivative w.r.t. θ:

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

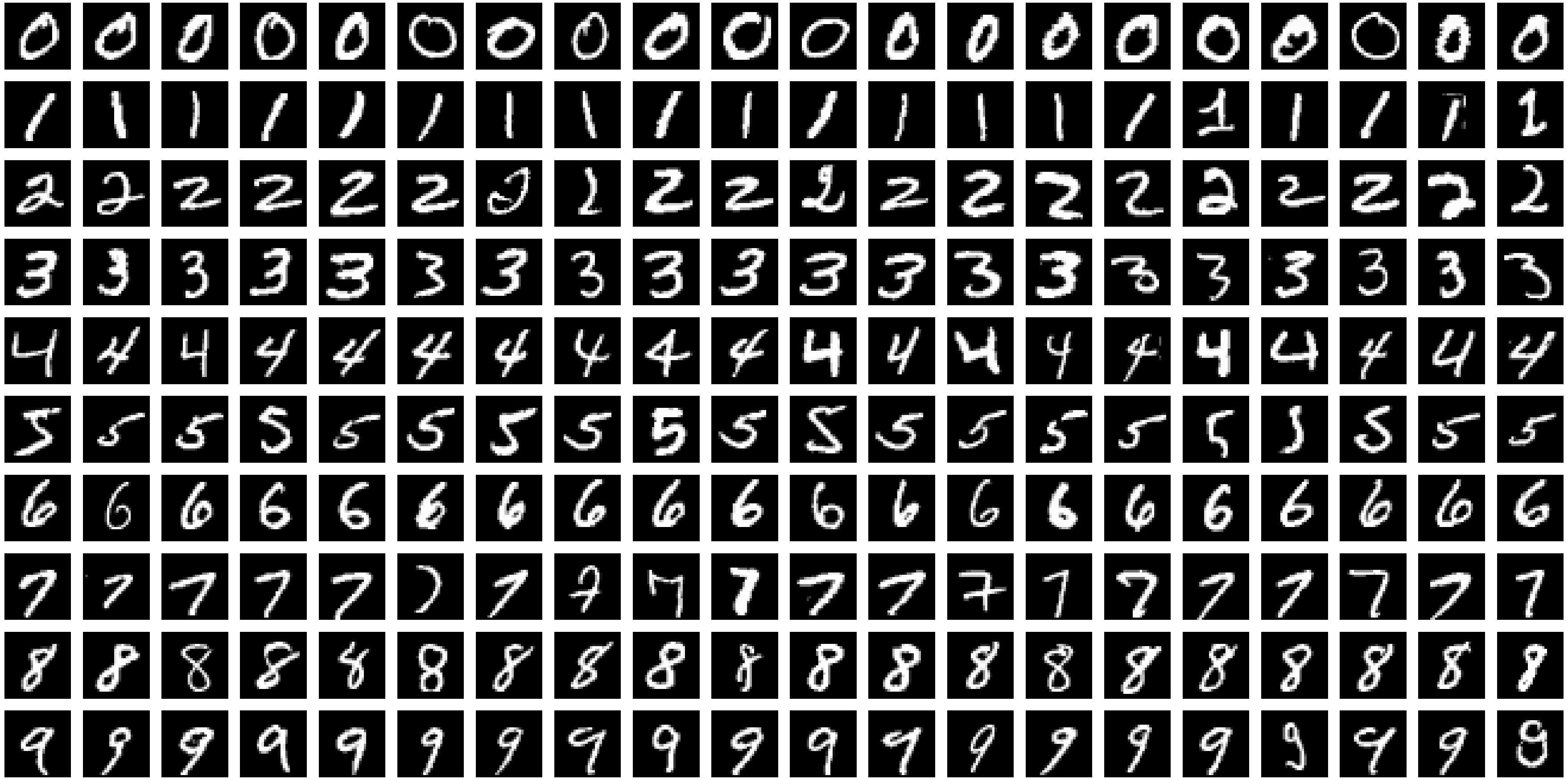
- It's very hard to take derivative w.r.t. ϕ! There is an integral!
- How to solve for this? **Reparameterization trick**
- Now, we can proceed in the previous algorithm…

Reparameterization Trick

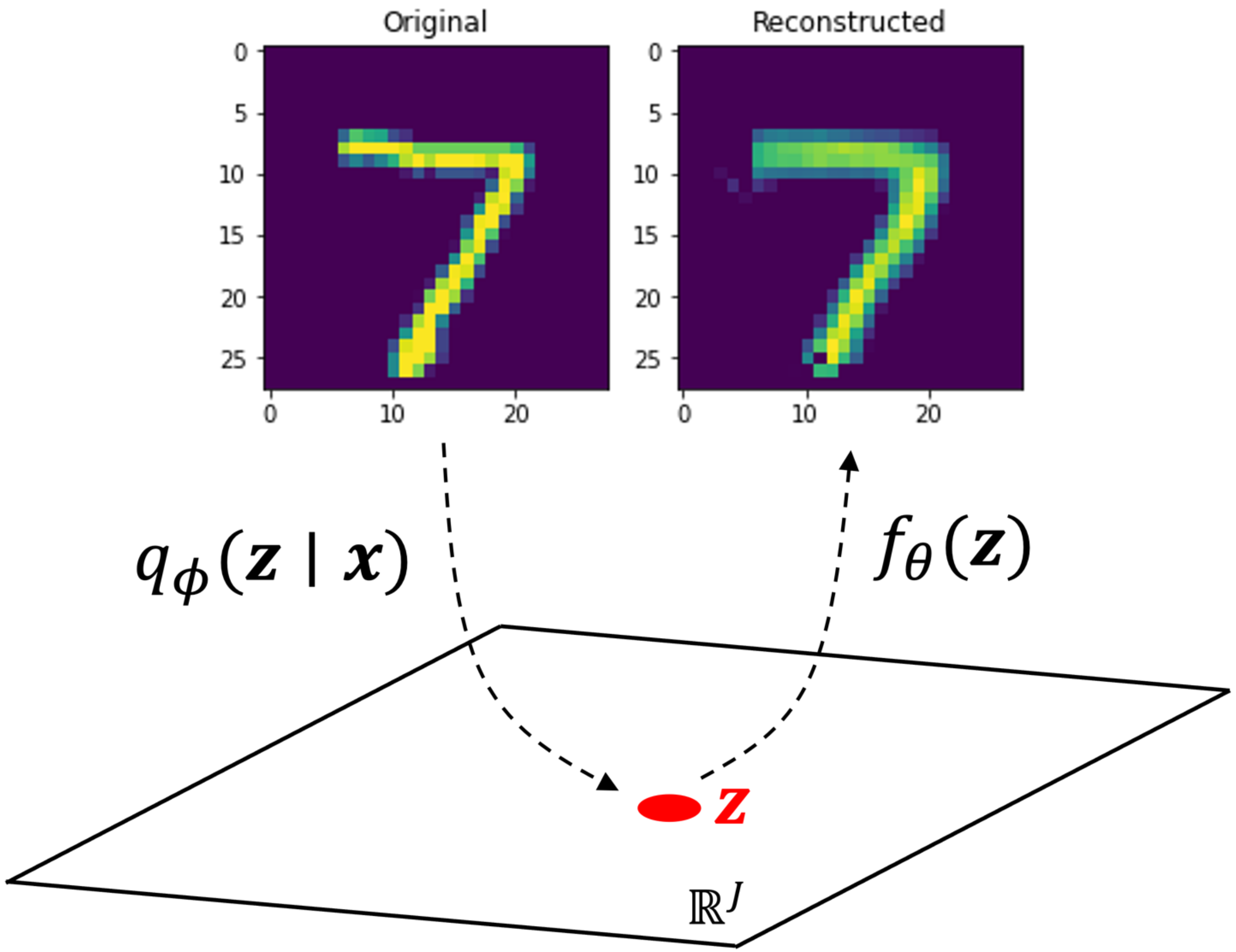


https://arxiv.org/pdf/1906.02691

Applying VAE on MNIST

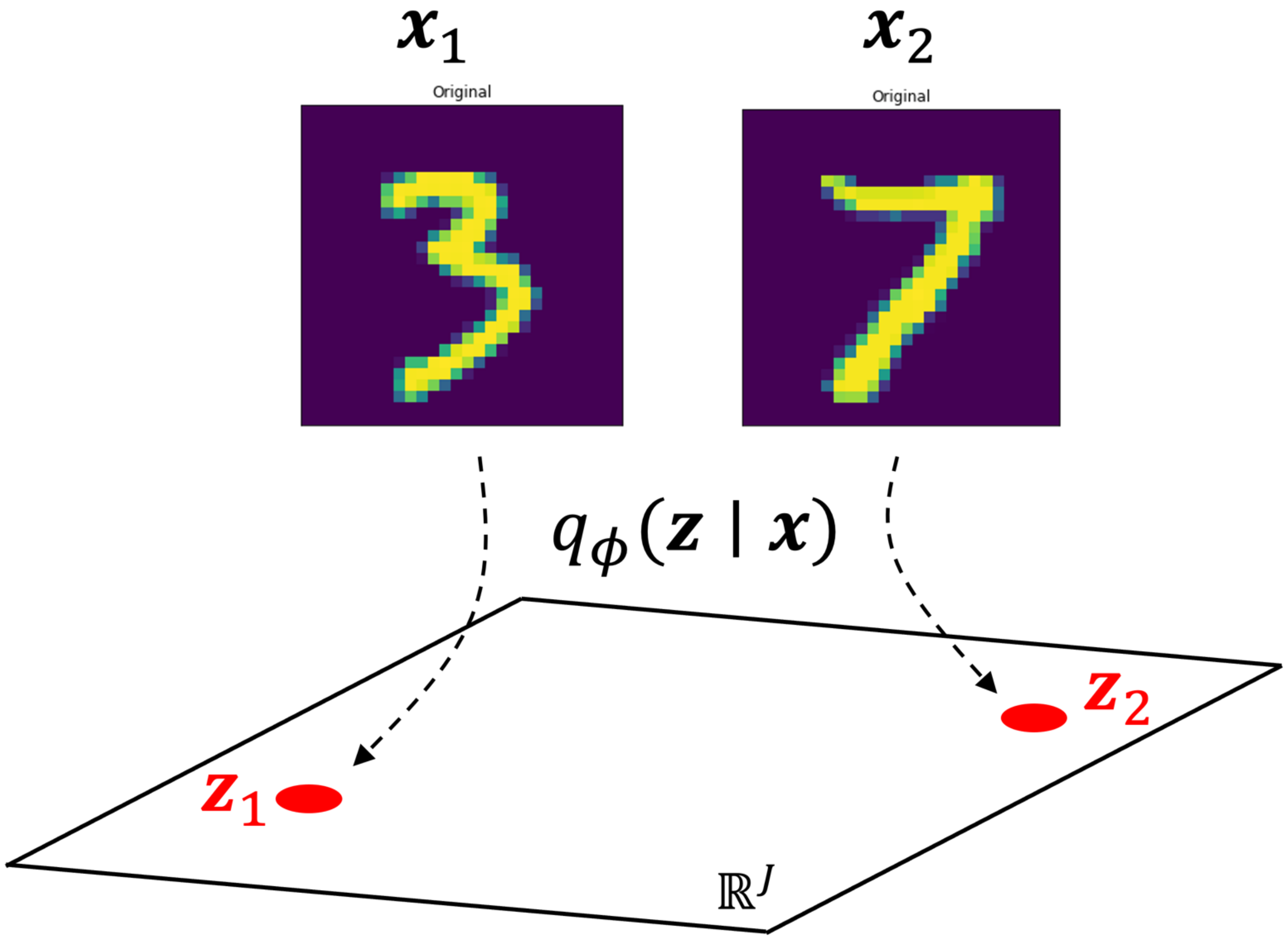


Reconstruction Example



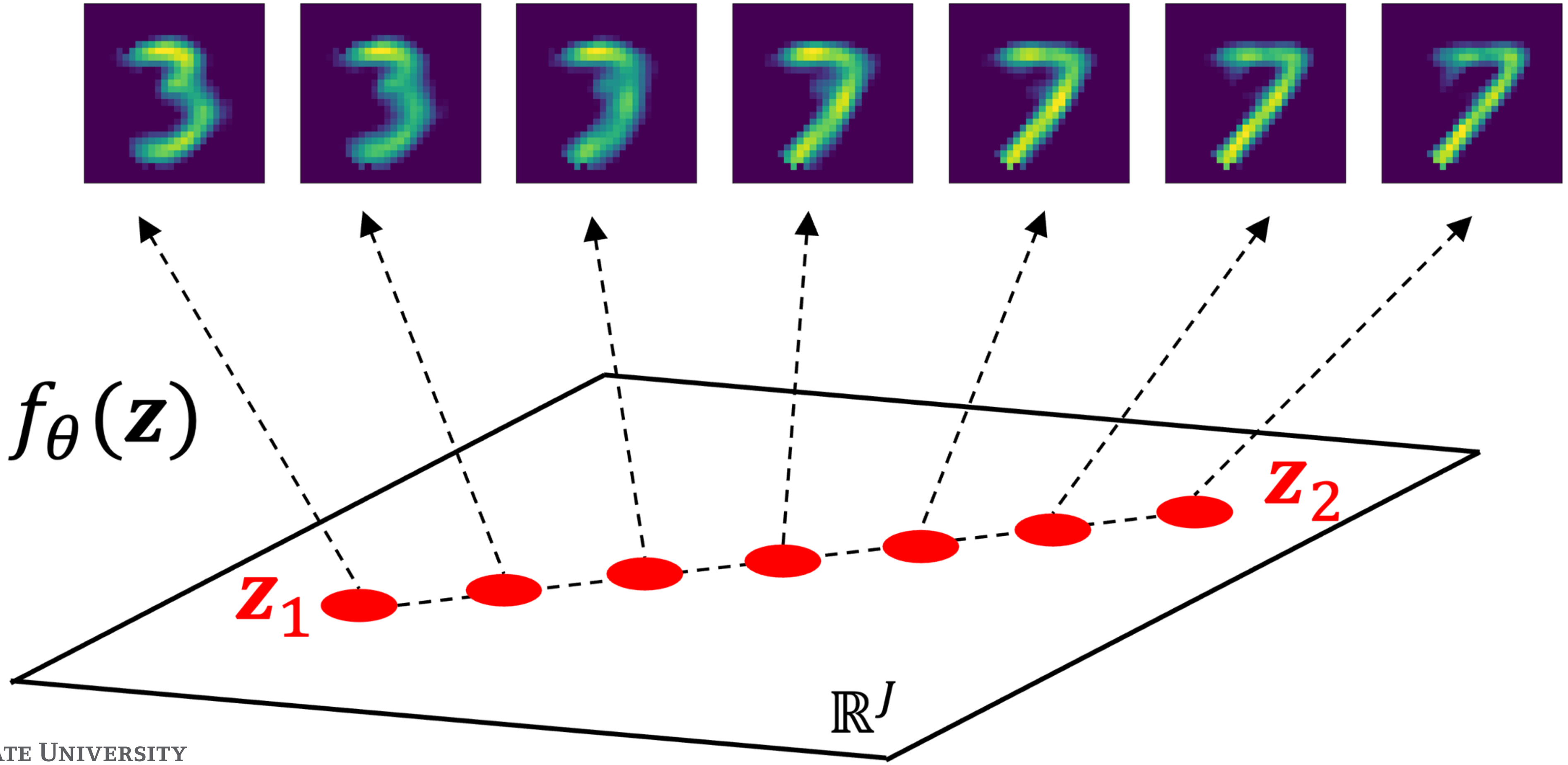
https://mbernste.github.io/posts/vae/

Exploring the Latent Space...



https://mbernste.github.io/posts/vae/

Exploring the Latent Space... (cont.)



https://mbernste.github.io/posts/vae/

Simulation

- Check out https://colab.research.google.com/drive/1tz_aNMLJjDqOtfK7MdGgylapDwf5FC-T?usp=sharing

Homework

For 2 of 3 models below, change at least 3 parameters of the model in class; examine any differerence (quantitatively or qualitatively)

1. VAE
2. DCGAN
3. Diffusion Model

- Send your report to chen.11020@buckeyemail.osu.edu

Questions?

References

- S. Ermon, Stanford CS 236 Course, URL: deepgenerativemodels.github.io
- D. P. Kingma and M. Welling, An Introduction to Variational Autoencoder, https://arxiv.org/pdf/1906.02691
- M. N. Bernstein, Variational autoencoders, URL: https://mbernste.github.io/posts/vae/