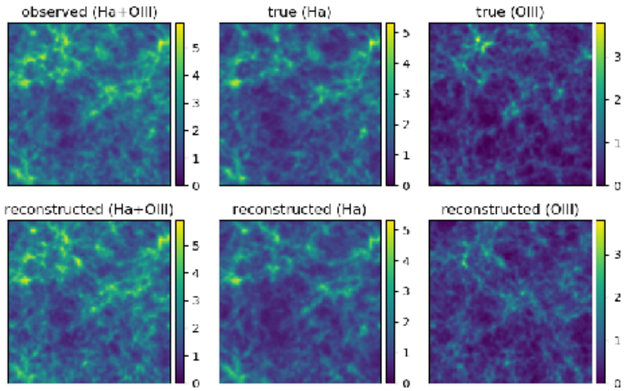
Generative model

Kana Moriwaki (UTokyo)

A^3-Net Summer School
2 - 6 Sep. 2024

Kana Moriwaki



https://www-utap.phys.s.u-tokyo.ac.jp/~moriwaki/

- **Assistant Professor (the University of Tokyo, RESCEU/UTAP)**
- **Research interest:**
  - **High-redshift galaxy formation, line-emitting galaxies**
  - **The large-scale structure of the Universe**
  - **Cosmic reionization**

- **I have been developing conditional GANs for signal reconstruction from large-scale line intensity maps: link**
- **We are developing surrogate models for speeding up galaxy formation simulations: link**
- **Our recent review on machine learning for cosmology: link**

Outline

plan

- **What is a good generative model?**
- **What is latent space?**

30 min

- **Variational auto encoder (VAE):**
  - Hands-on: VAE for galaxy spectrum

40 min (incl. break)

- **Generative adversarial network (GAN)**

10 min

  - Hands-on: GAN for galaxy spectrum

30 min (incl. break)

- **Flow-based models**

20 min

  - (Hands-on: Flow for galaxy spectrum)

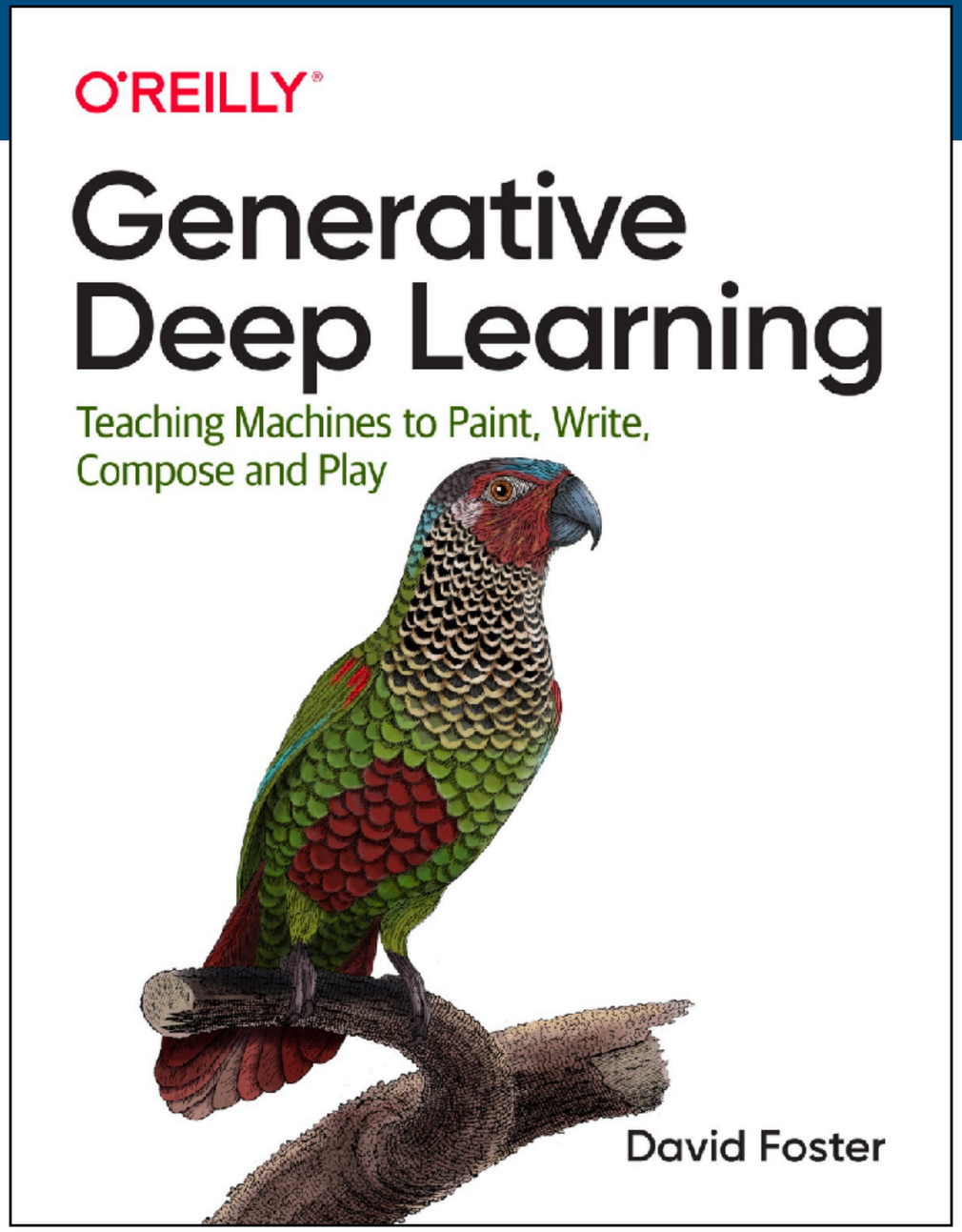
(if we have time)

- **Diffusion models**

20 min

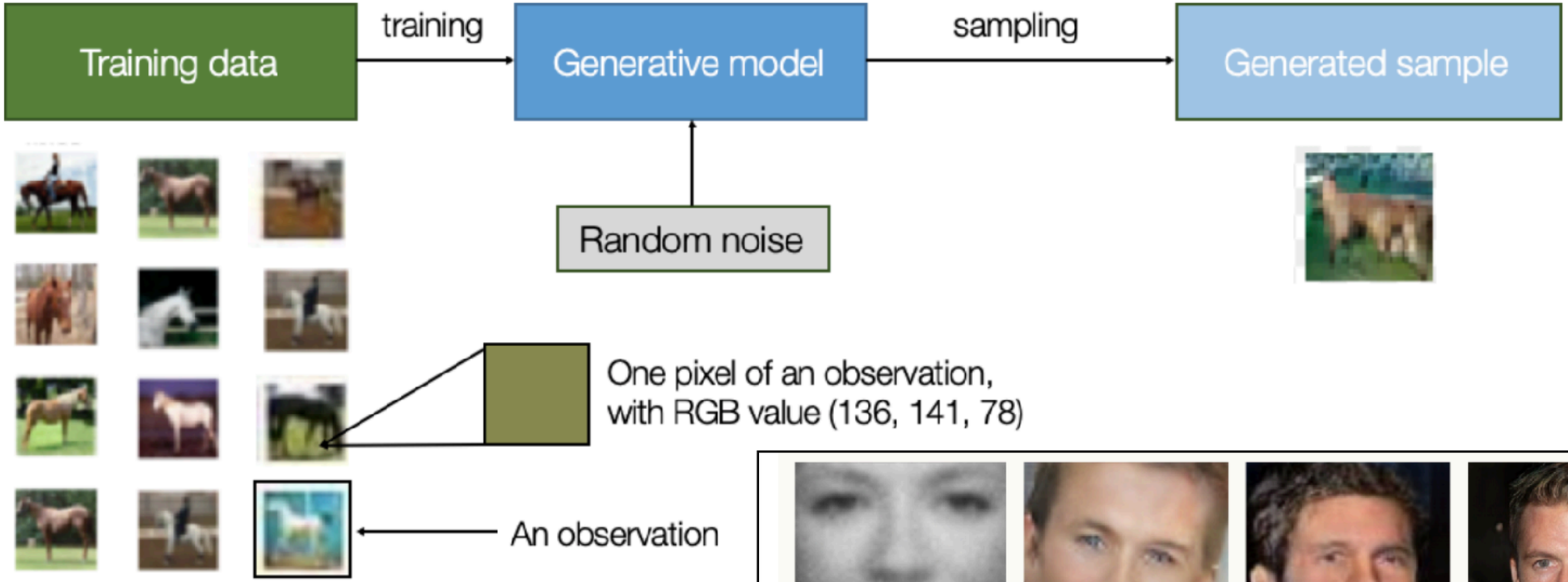
- **Summary**

Textbook


O'REILLY
Generative Deep Learning
Teaching Machines to Paint, Write, Compose and Play
David Foster

What is generative model?

Generative model: models that generate new data that are similar to the training data



Foster 2019

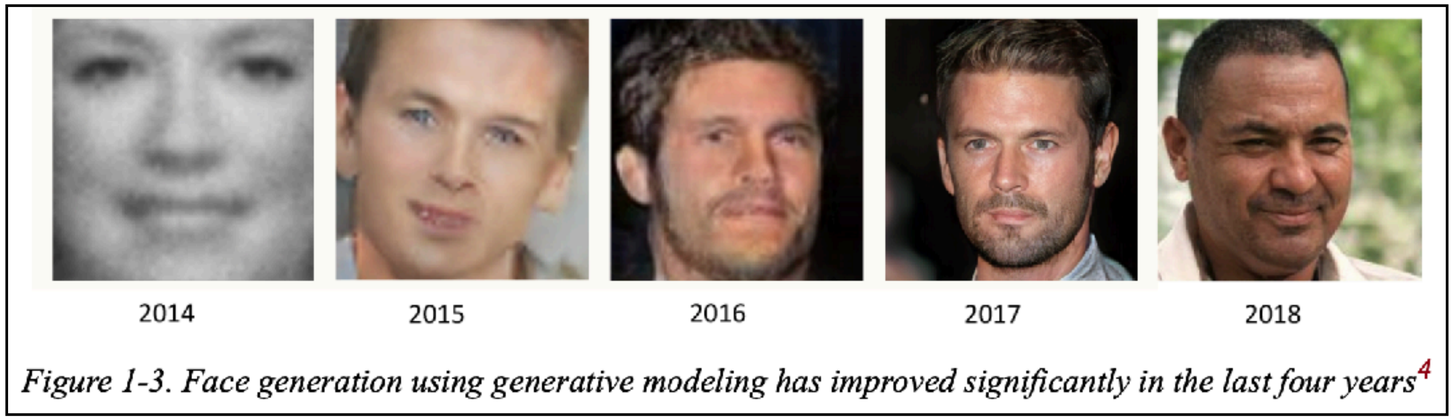


Figure 1-3. Face generation using generative modeling has improved significantly in the last four years[4]

Stable Diffusion XL

Create and inspire using the worlds fastest growing open source AI platform.

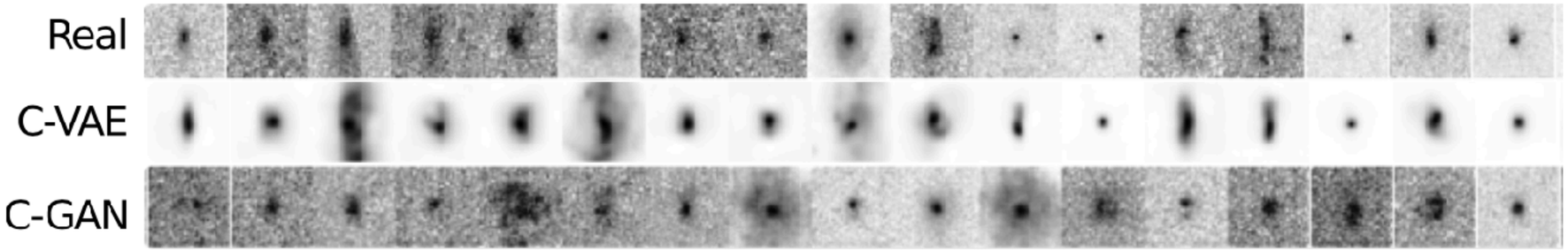
With Stable Diffusion XL, you can create descriptive images with shorter prompts and generate words within images. The model is a significant advancement in image generation capabilities, offering enhanced image composition and face generation that results in stunning visuals and realistic aesthetics.

Stable Diffusion XL is currently in beta on DreamStudio and other leading imaging applications. Like all of Stability AI's foundation models, Stable Diffusion XL will be released as open source for optimal accessibility in the near future.

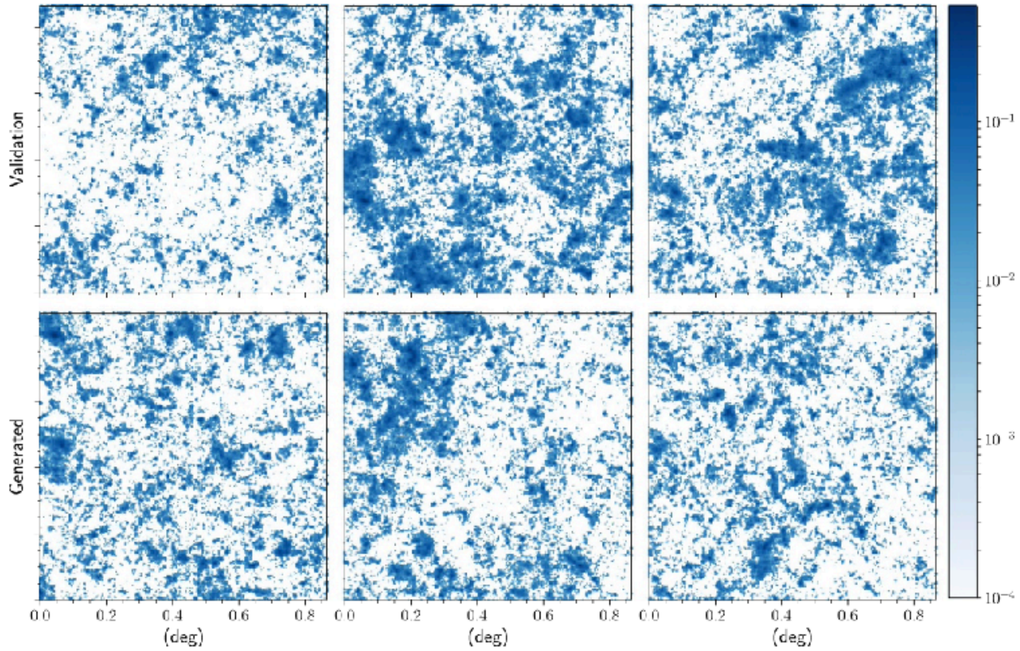
DreamStudio

Generative models in astrophycis

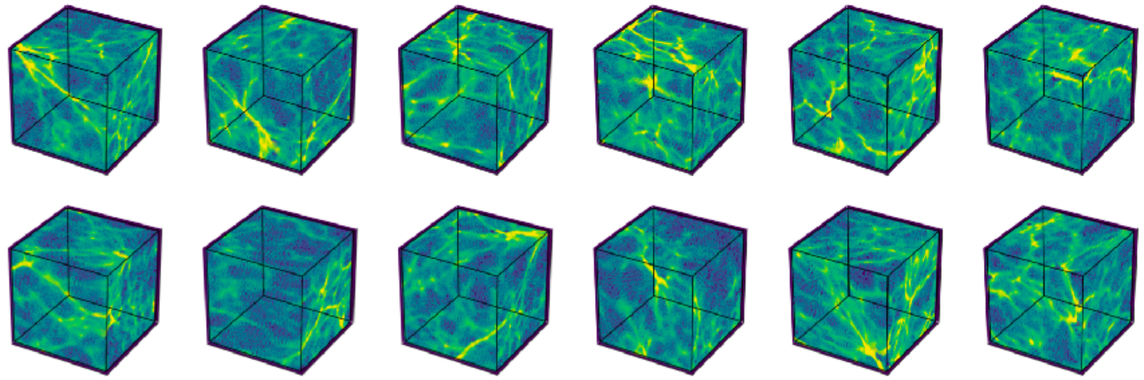
Mock galaxy images used for gravitational lens analysis (Ravanbakhsh+16)



Weak lensing convergence maps generated by CosmoGAN (Mustafa+2019)



Neutral hydrogen maps generated by HIGAN (Zamudio-Fernandez et al. 2019)

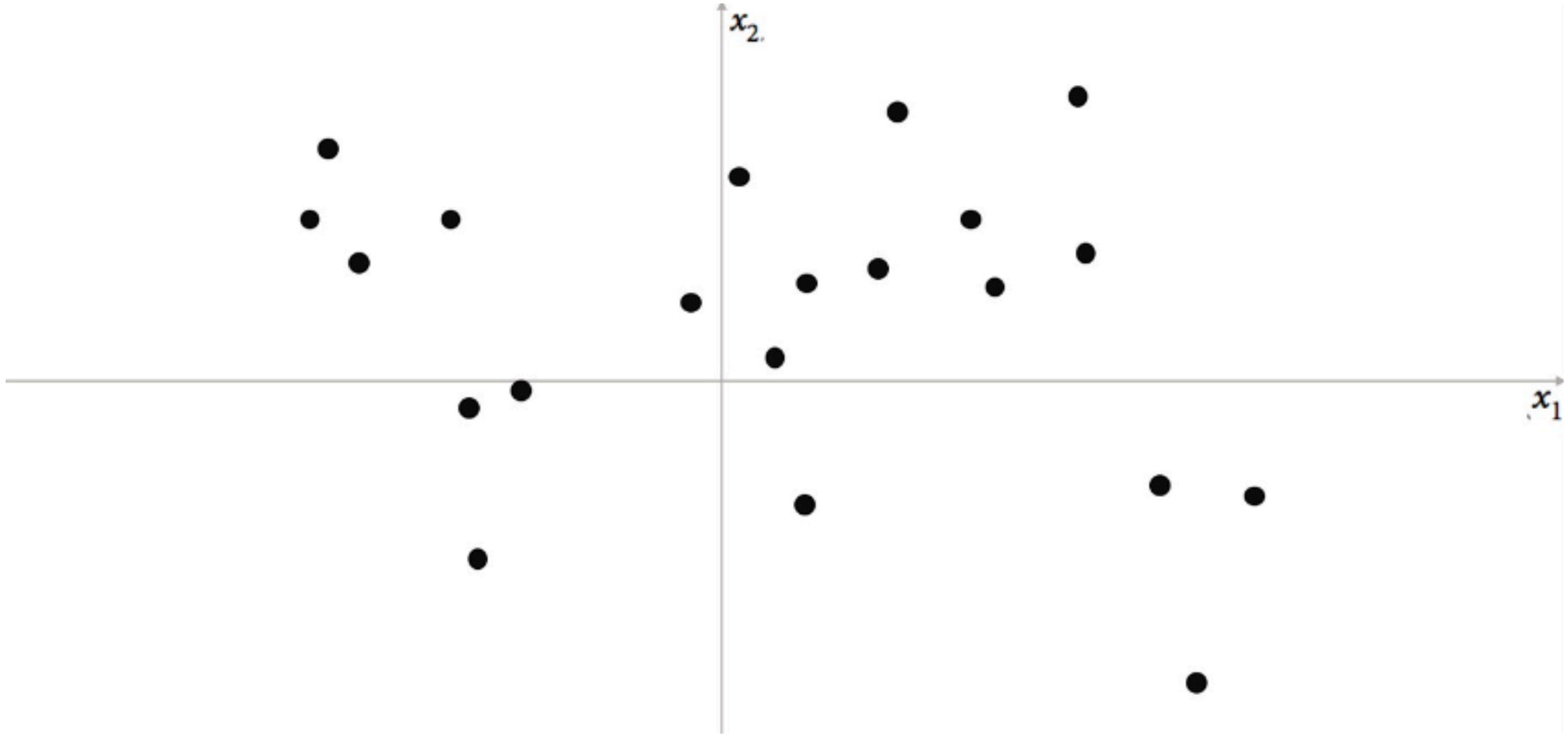


What is a good generative model?

What is a good generative model?

Let's start by playing a generative modeling game in just two dimensions. I've chosen a rule that has been used to generate the set of points $\mathbf{X}$ in Figure 1-4. Let's call this rule p_{data}. Your challenge is to choose a different point $\mathbf{x} = (x_1, x_2)$ in the space that looks like it has been generated by the same rule.

Foster 2019

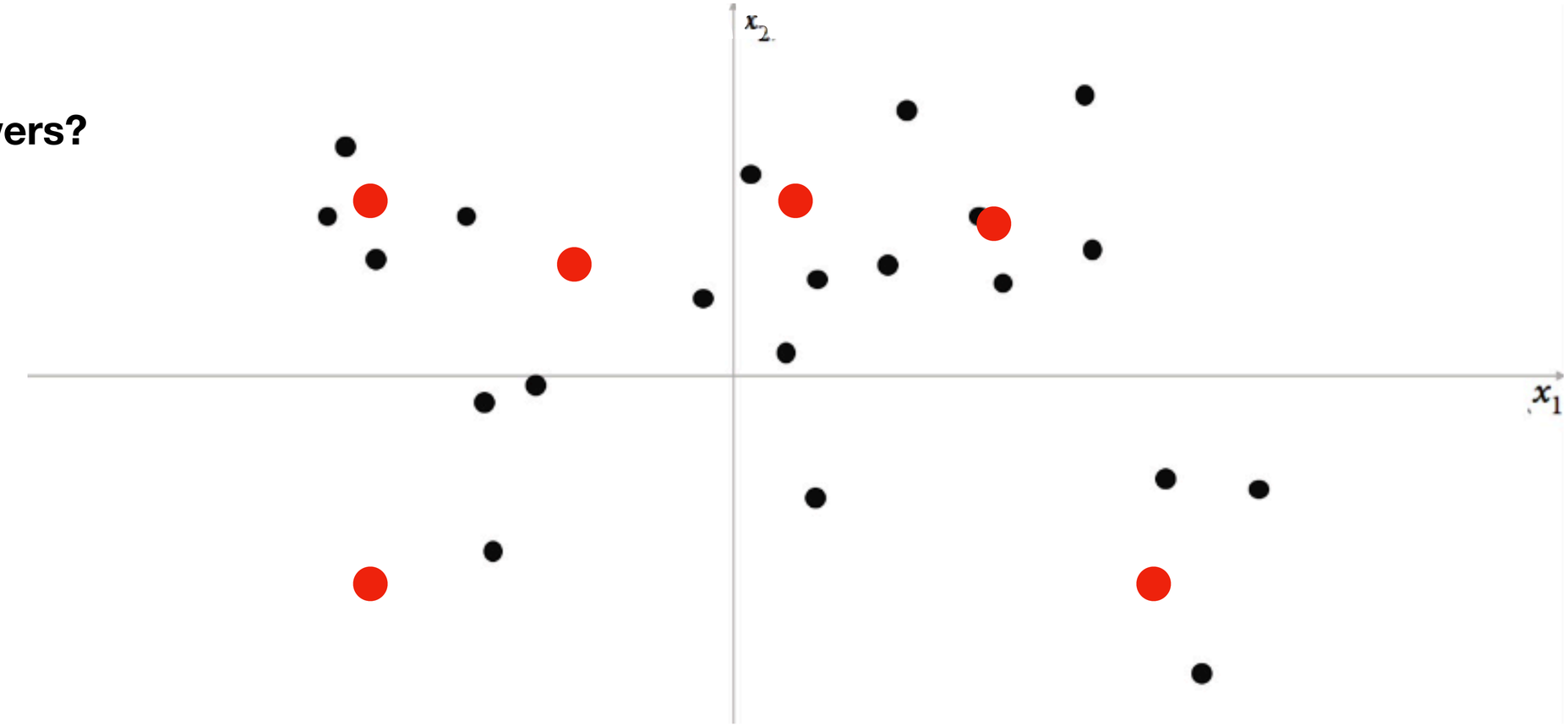


What is a good generative model?

Let's start by playing a generative modeling game in just two dimensions. I've chosen a rule that has been used to generate the set of points $\mathbf{X}$ in Figure 1-4. Let's call this rule p_{data}. Your challenge is to choose a different point $\mathbf{x} = (x_1, x_2)$ in the space that looks like it has been generated by the same rule.

Foster 2019

- **Possible answers?**

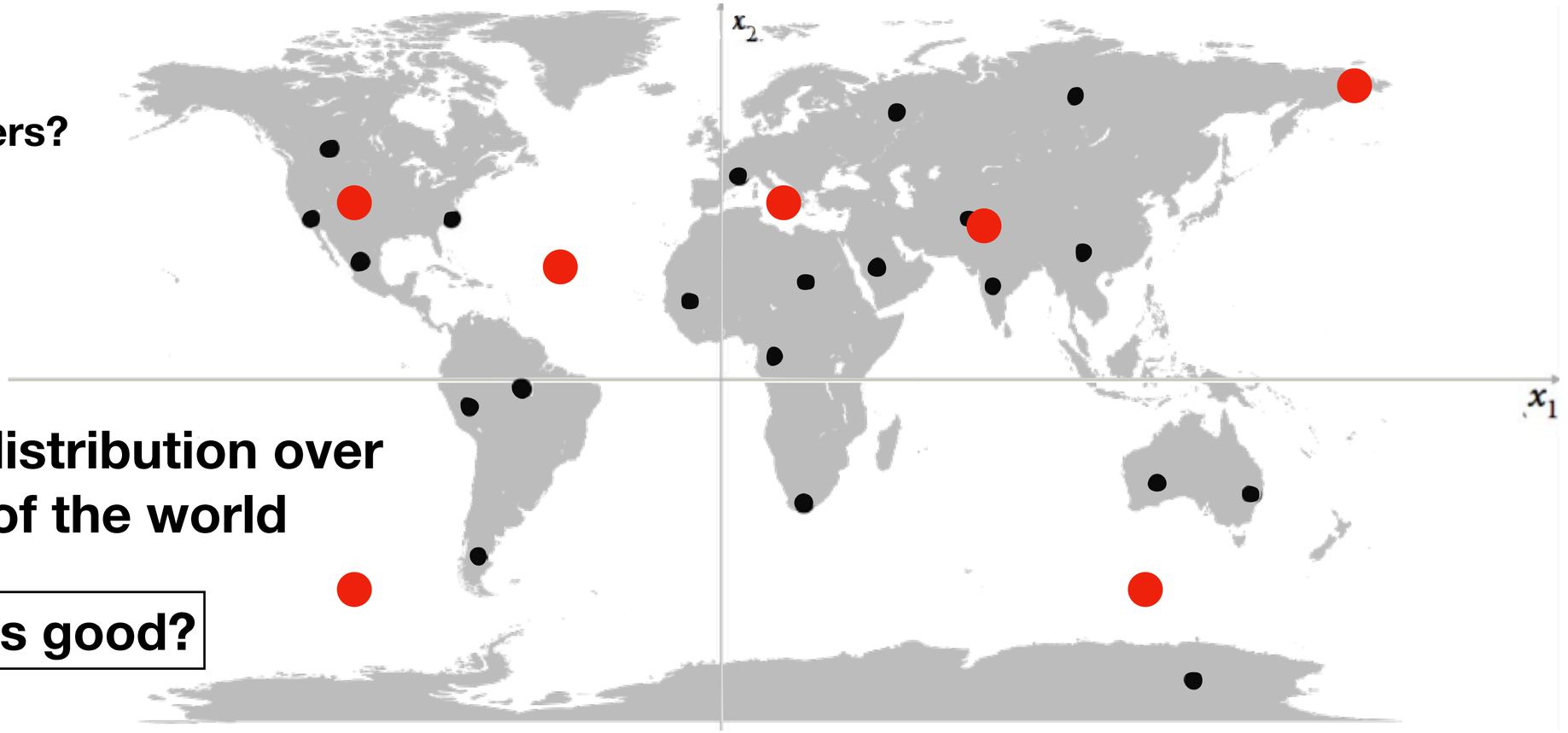


What is a good generative model?

Let's start by playing a generative modeling game in just two dimensions. I've chosen a rule that has been used to generate the set of points $\mathbf{X}$ in Figure 1-4. Let's call this rule p_{data}. Your challenge is to choose a different point $\mathbf{x} = (x_1, x_2)$ in the space that looks like it has been generated by the same rule.

Foster 2019

- Possible answers?



Rule: uniform distribution over the land mass of the world

Which one looks good?

What is a good generative model?

Key requirements

- **High quality:**

  Data should be generated following underlying rules

  Generated data resembles real data both individually and statistically

- **Diversity:**

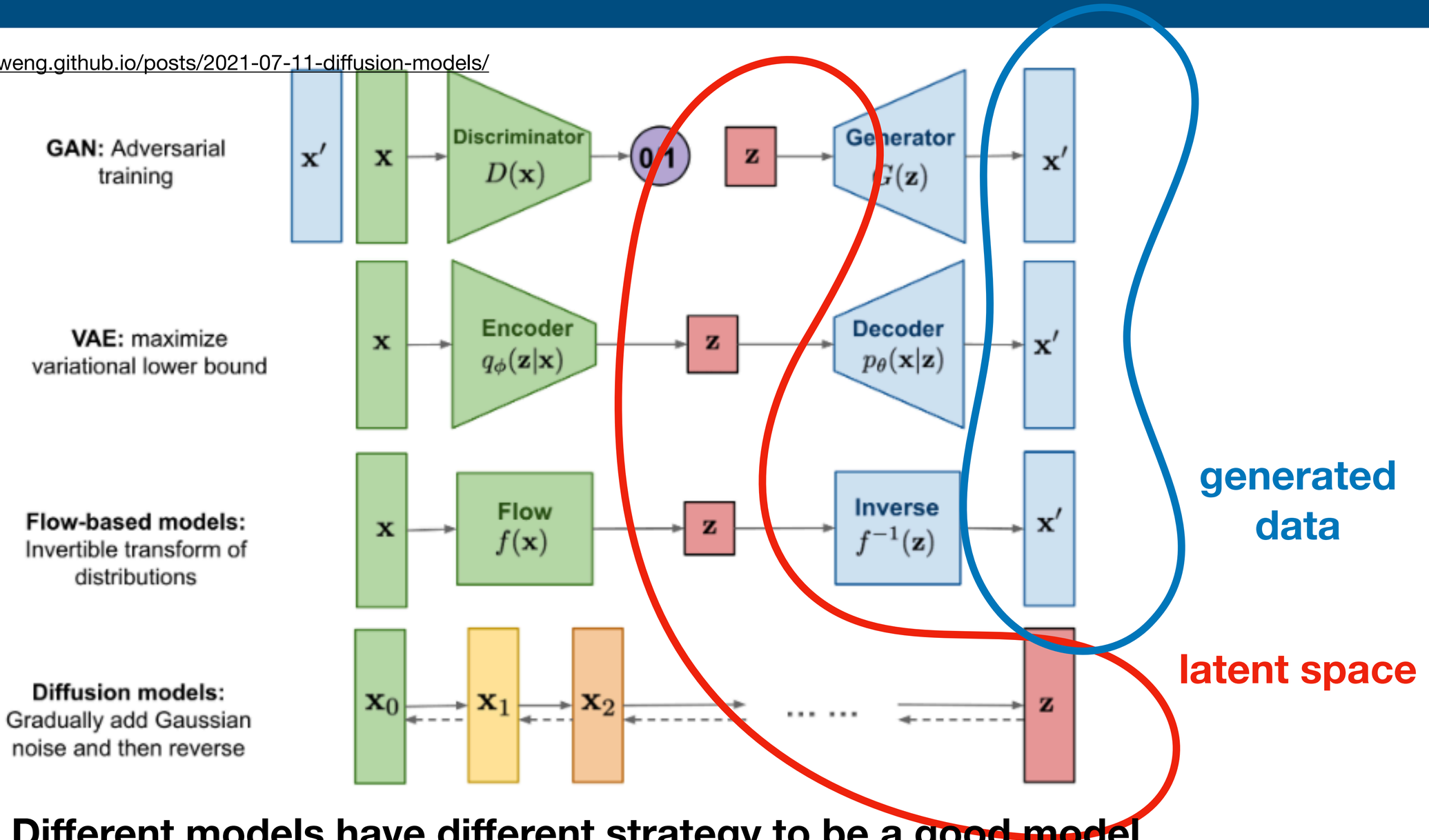
  the model can generate many different variations

Optional features

- **Training stability**
- **Computational efficiency**
- **Sampling speed**
- **etc.**

Generative models

Image from https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

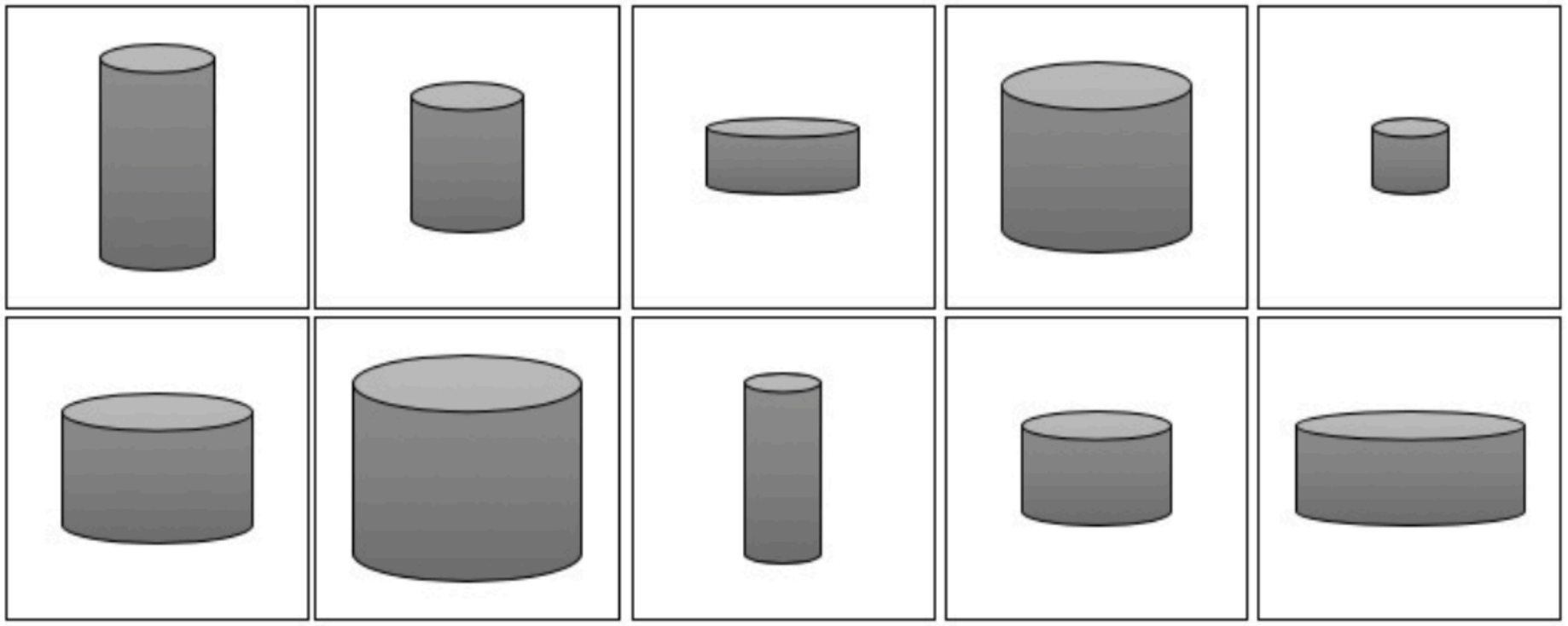


Different models have different strategy to be a good model

What is latent space?

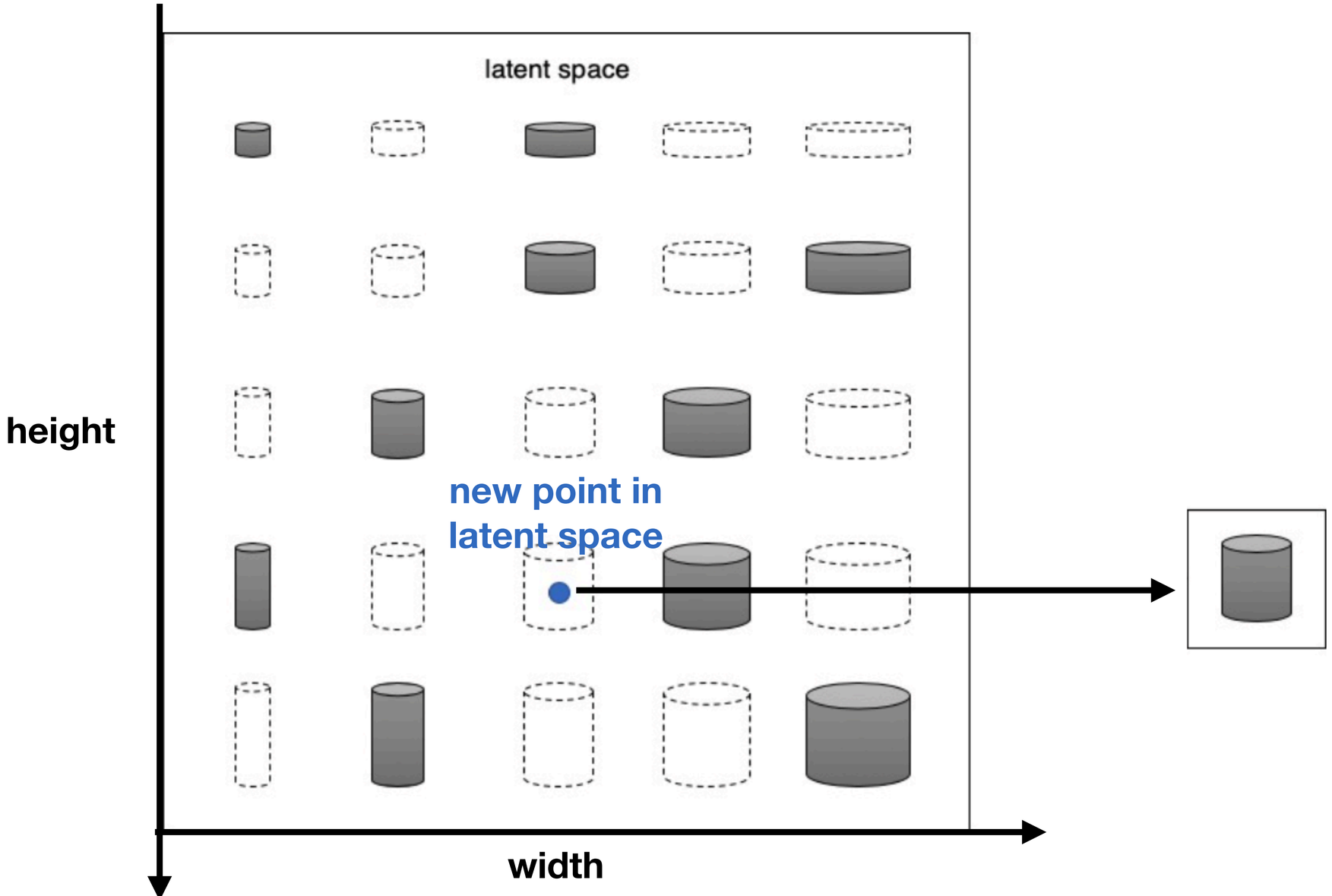
What is Latent Space?

Training dataset

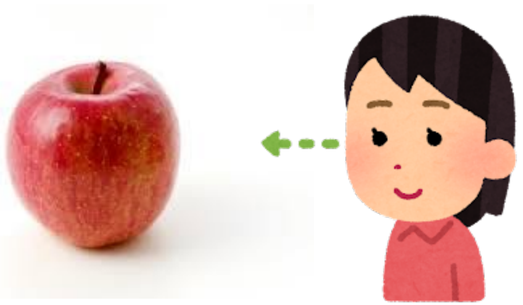


What are the representative features?

What is Latent Space?



Why is Latent Space Needed?



pixel 1: (255, 255, 255)
pixel 2: (255, 10, 10)
...

Representation learning allows us to generate new realistic data

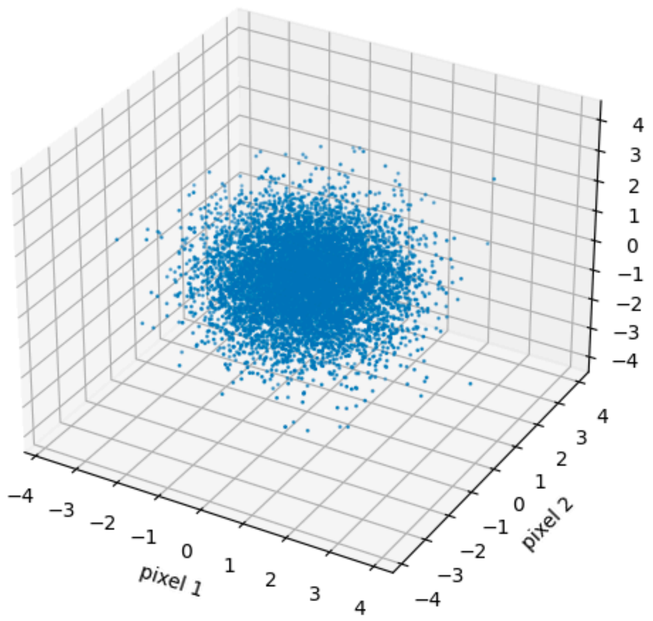
- round
- red
- a stem at the top
...

What about green instead of red?

Data space ⇔ Latent space

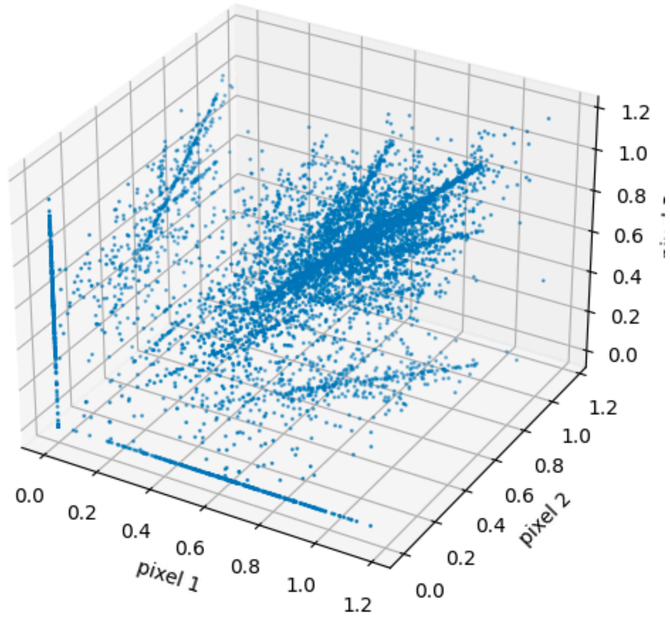
All popular generative models learn how to map from latent space to data space

latent space:

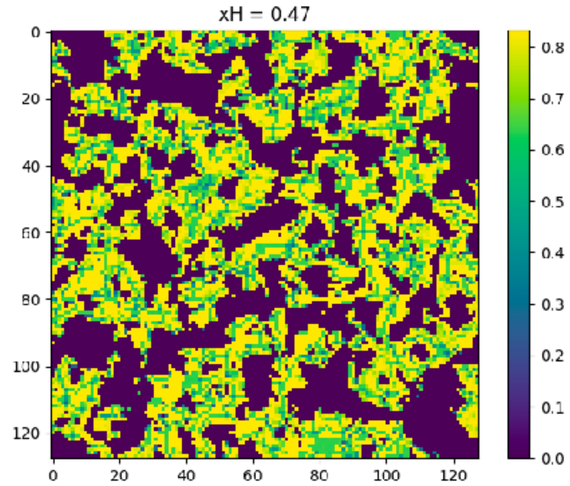


?

data space: 256×256 dimension

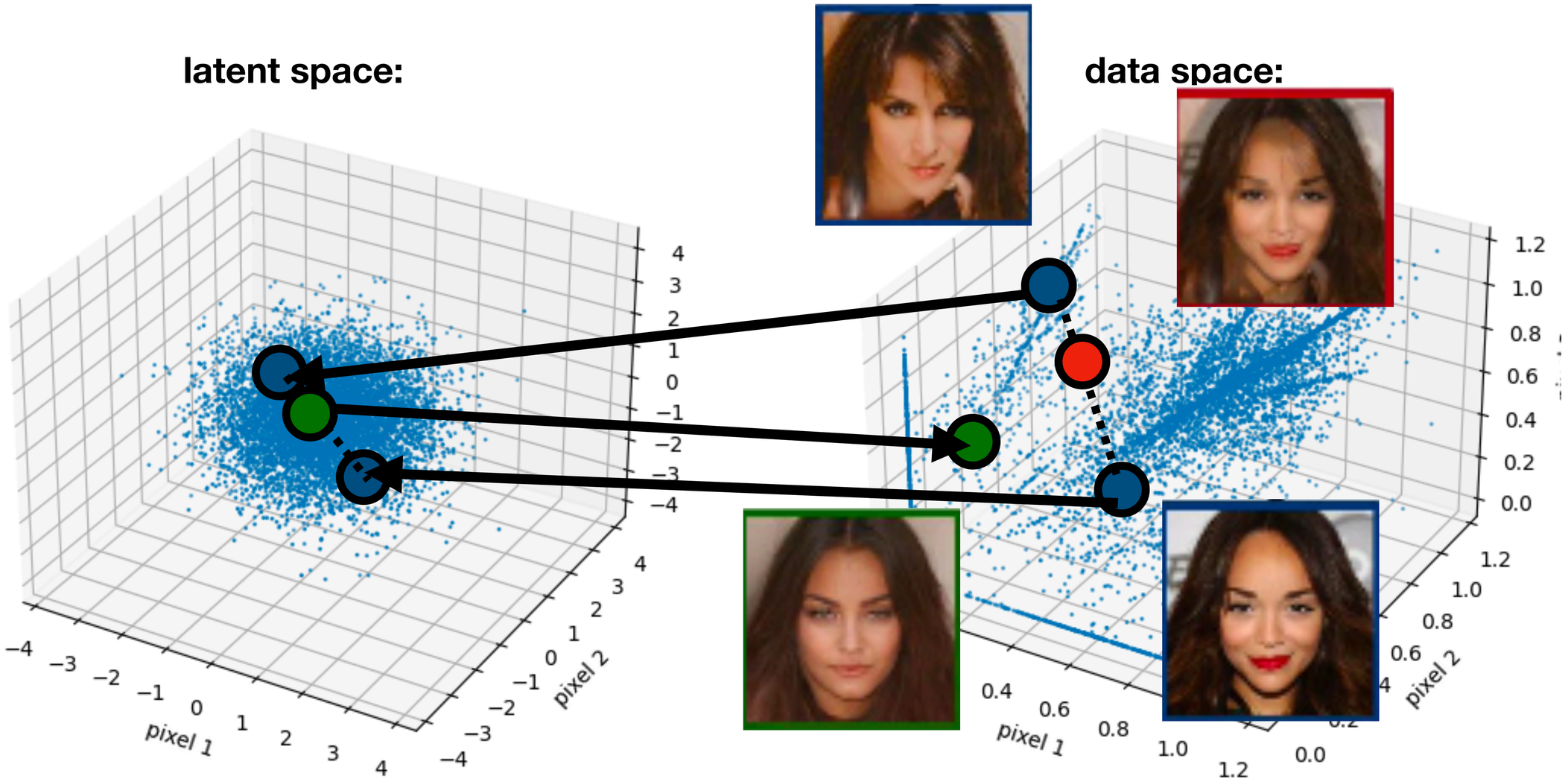


← distribution of pixel values of 21cm map



Different models adopt different mapping / training strategies

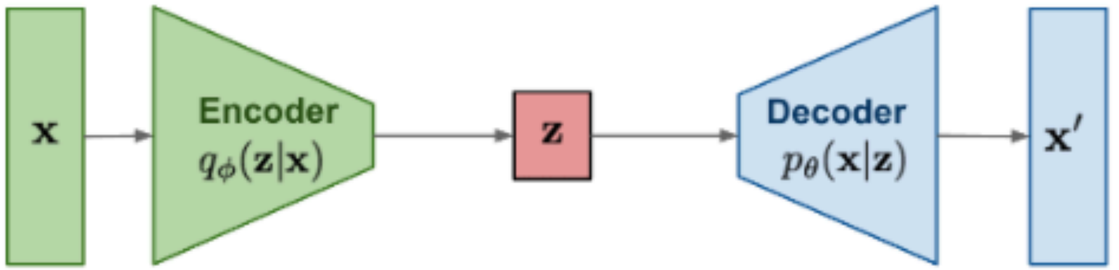
Importance of latent space



Images from Ho et al. 2020

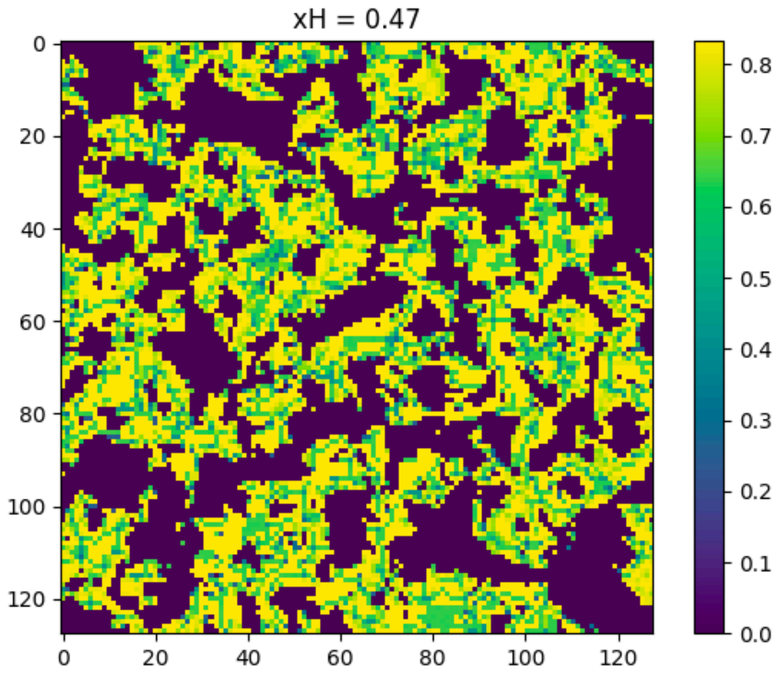
Auto Encoder and Variational Auto Encoder

VAE: maximize variational lower bound

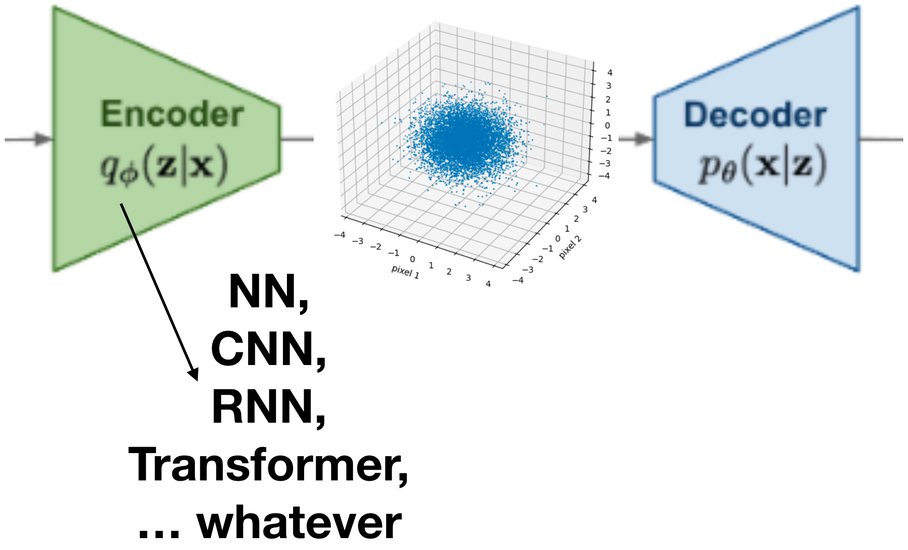


Auto Encoder

Data
high dimension

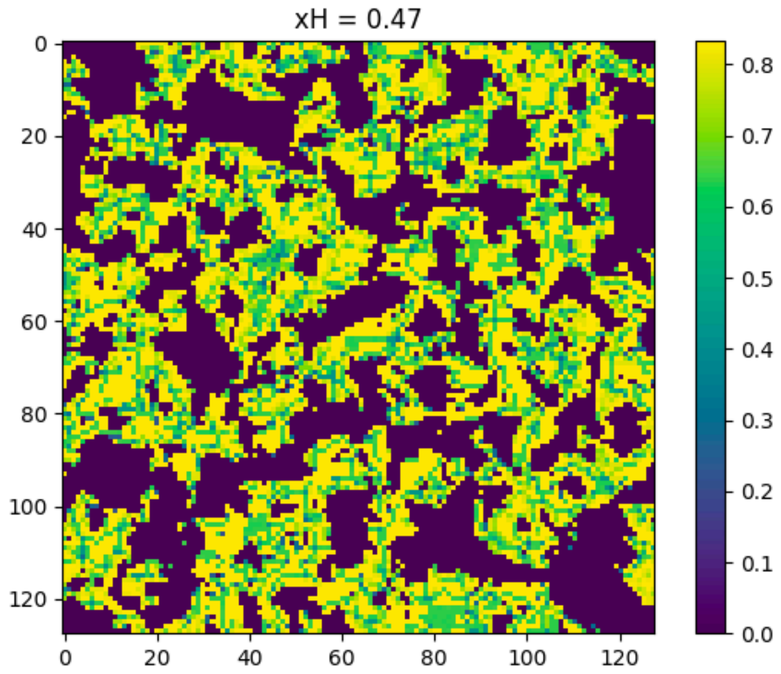


Latent space
low dimension



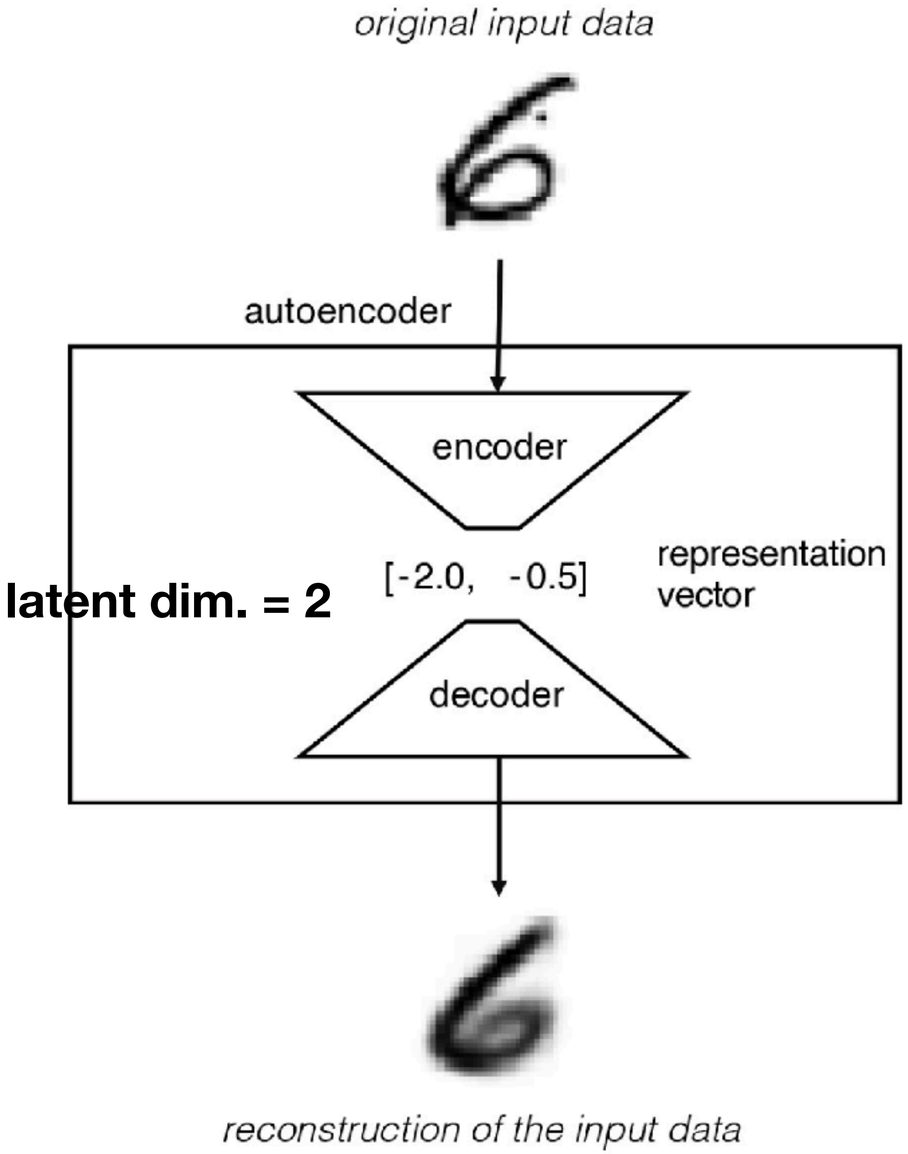
NN,
CNN,
RNN,
Transformer,
… whatever

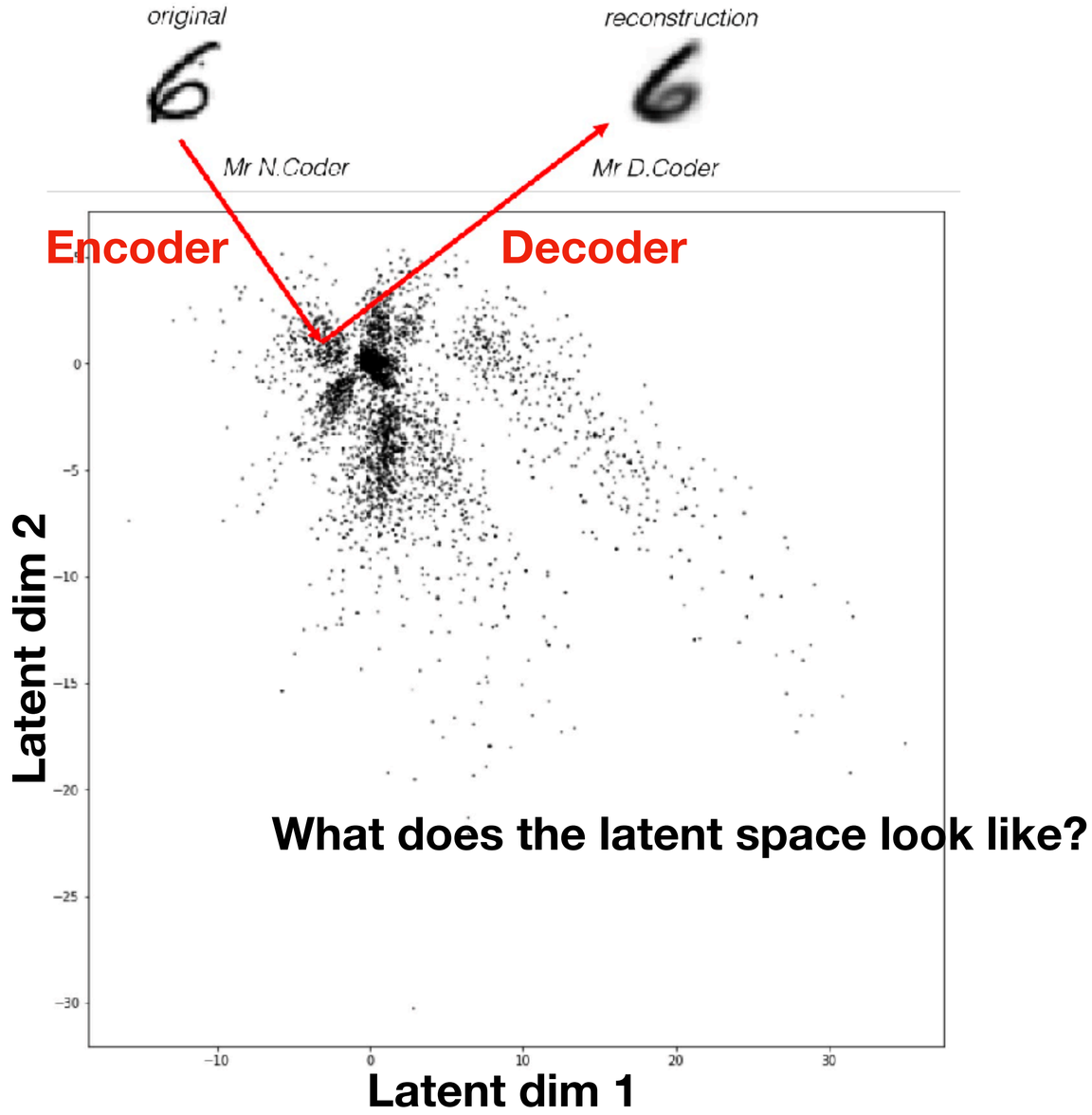
Generated data
high dimension



Train two networks so that generated data become similar to input data
Loss: differences between input and generated data (e.g., MSE)

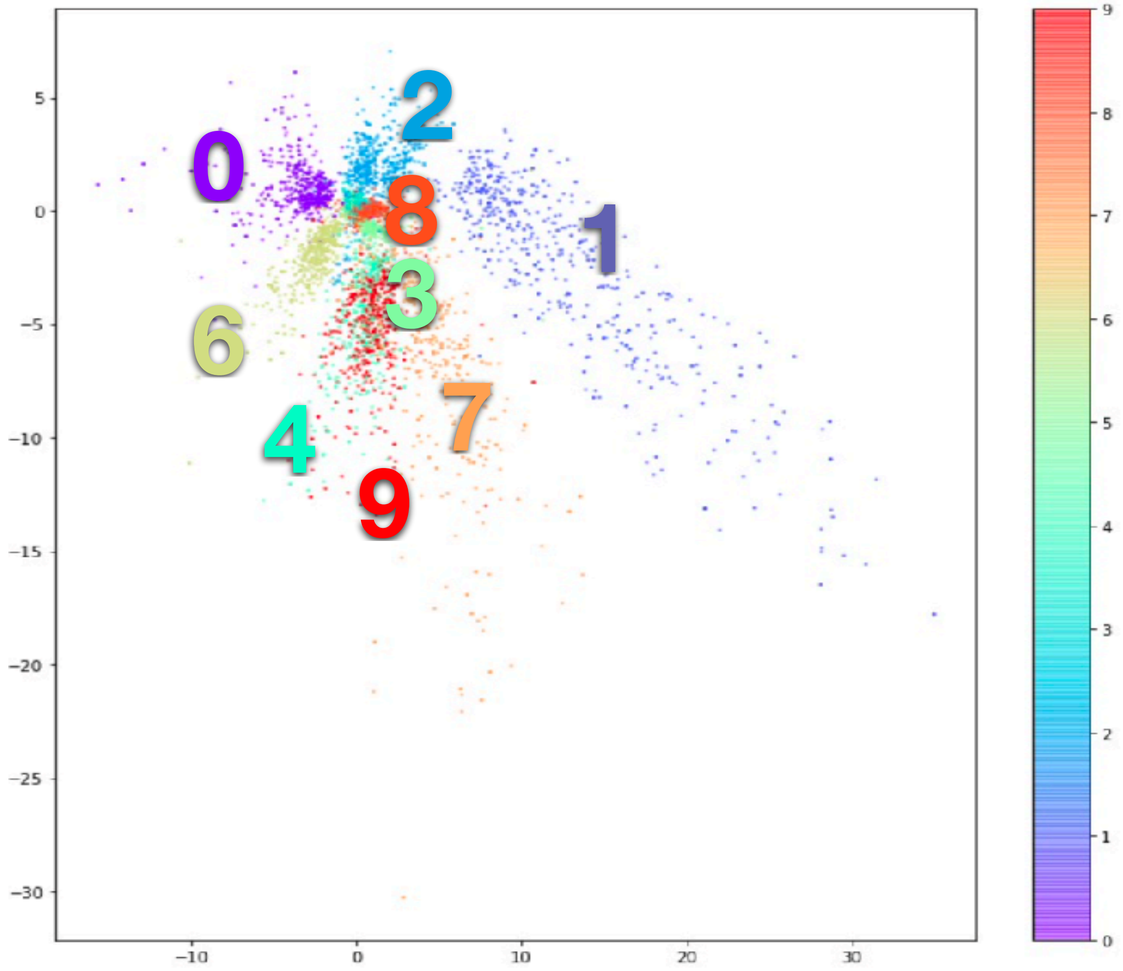
Example results on handwritten digits (MNIST)





What does the latent space look like?

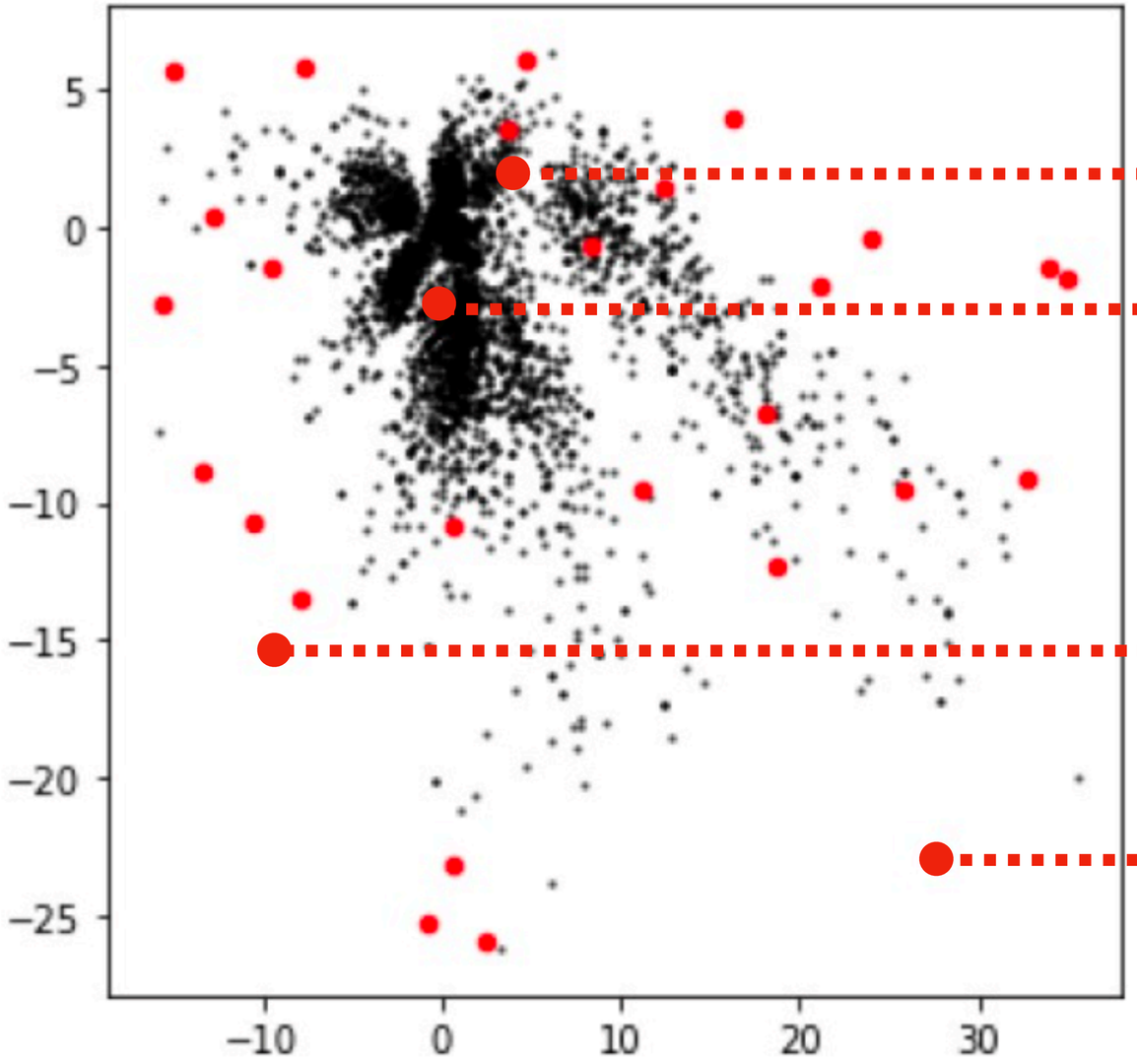
What does the latent space look like?

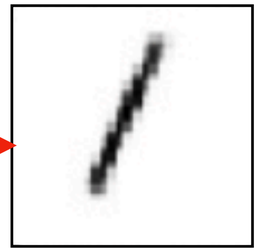


→ AE can also be used for clustering analysis as a byproduct

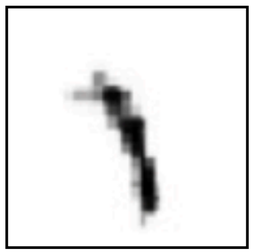
The similar (same) digits are close to each other in the latent space even though the digit labels are never shown to the model during training!

A Problem with Auto Encoder

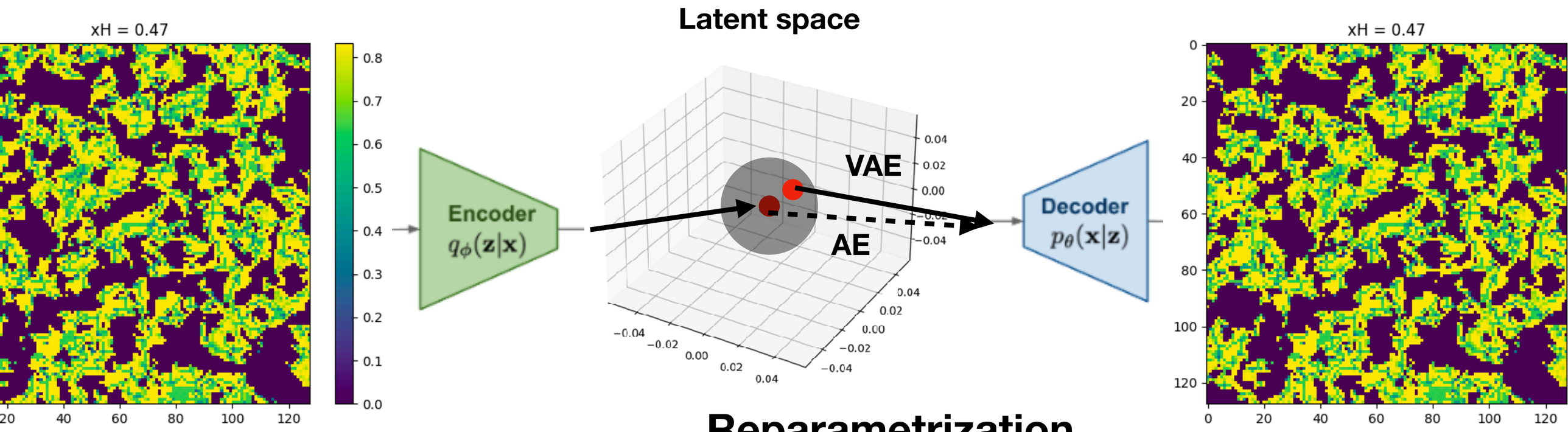




When sampled from out-of-data region,
AE cannot generate data well

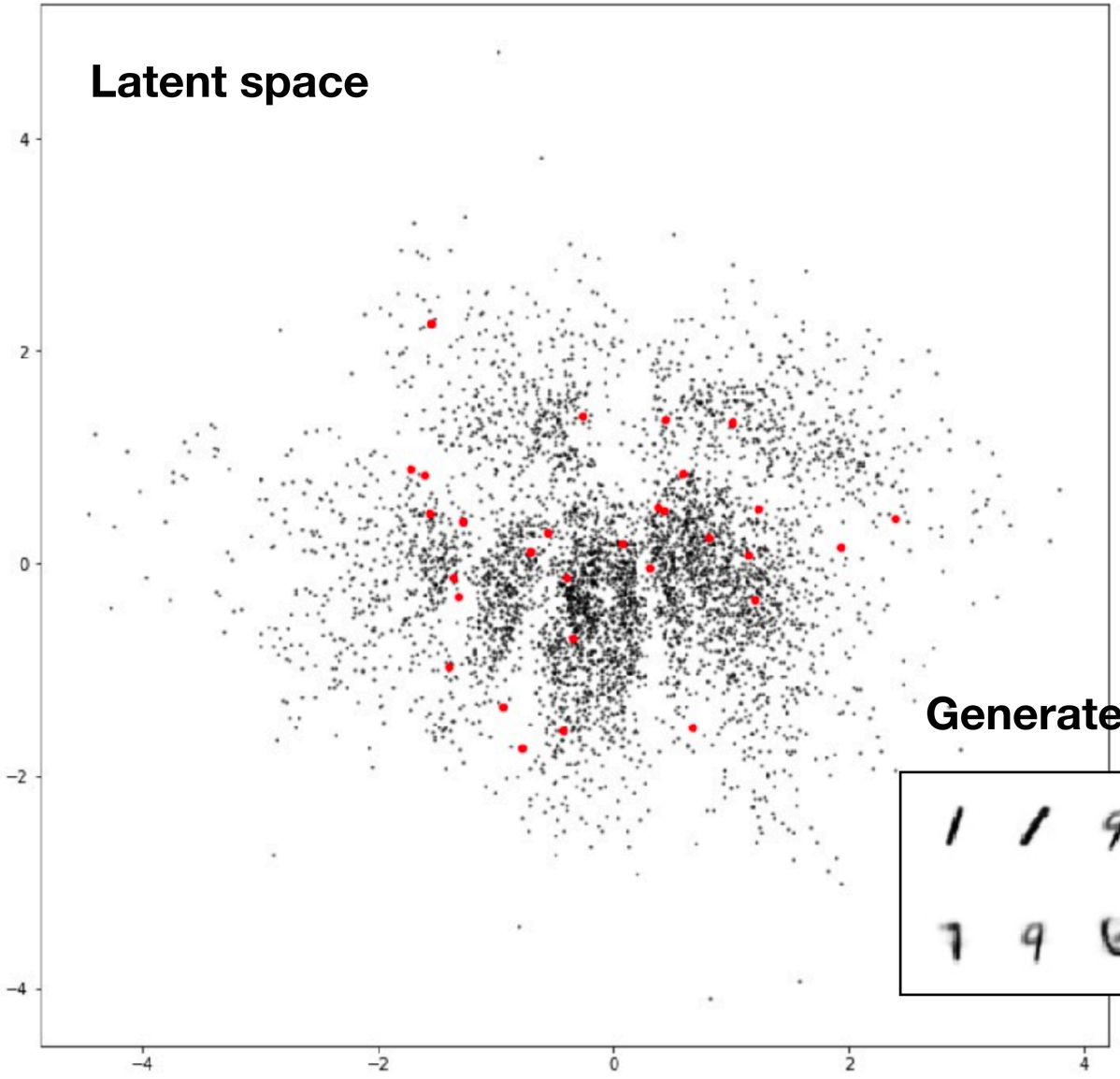


Variational Auto Encoder

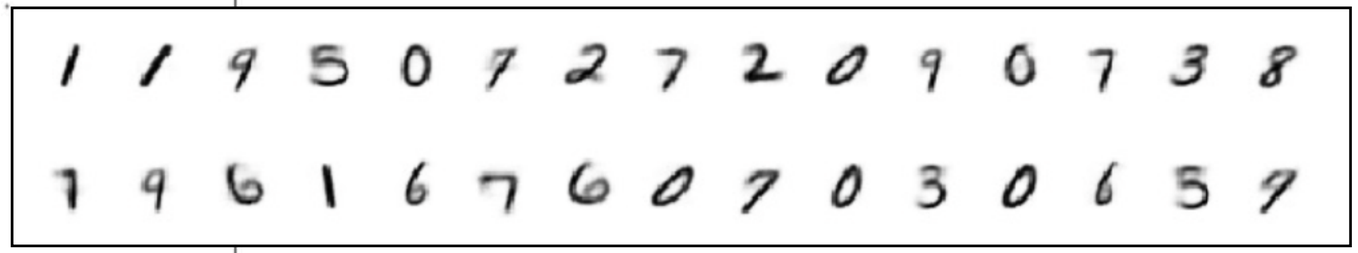


Reparametrization
(add random small shift)

Variational Auto Encoder



Generated samples



Comparison

	Latent space dim.	Data Quality	Prob. Dist.	Sampling speed	Note
AE	low	△	×	○	
VAE	low	○	△	○	
GAN					
Flow					
Diffusion					

○: Good, △: so-so, ×: not good

Hands-on: VAE for SDSS galaxy spectrum

Hack Resources

google drive for data

It is most straightforward to directly work with the above Google drive from colab, since the files don't need to leave Google's servers in that case. The way I figured out how to do this is as follows (there might be a better method):

1. open the above Google drive.
2. right-click on the file or folder you need. → **sdss_galaxy_spec.hdf5**
3. click "Organize" -> "Add shortcut".
4. in "All locations", choose "My Drive" and click "Add".
5. in the colab instance, open the "Files" explorer on the left.
6. click "Mount Drive" icon.
7. you should be able to see the file now. To switch the colab working directory to your drive, type: `cd "/content/drive/MyDrive/"`

github: a3net_2024/Lecture_Day4_Moriwaki

Day 4: generative models

Hands-on exercises

Example codes use SDSS spectrum dataset (sdss_galaxy_spec.hdf5) in Google drive.

VAE → **copy the cells to your notebook**

GAN

Flow

Hands-on: VAE for SDSS galaxy spectrum

```
z = self.reparameterize(mu, logvar) # Use this for VAE
#z = mu                             # Use this for AE
```

change this

Questions:

- Is there any difference between AE and VAE?

- Are features in data properly reproduced? If not, which hyper parameters to change?

- What happens if you increase the latent-space dimension?

- What properties correspond to the latent space? (See an example plot for redshift in the last cell)

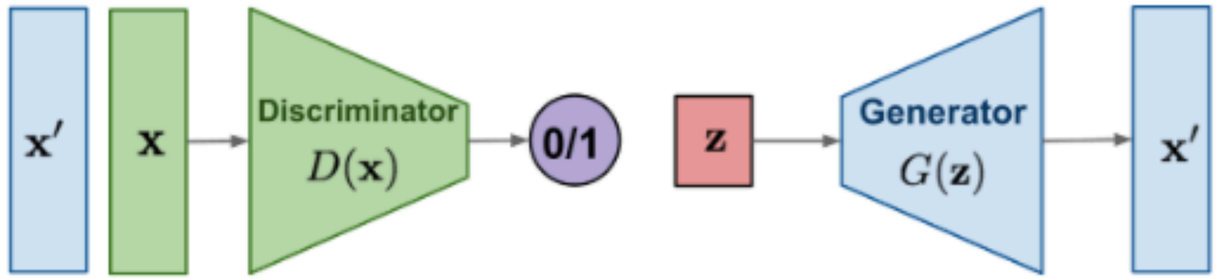
- How does the generated data change when you move a point in the latent space in a certain direction?

- etc.

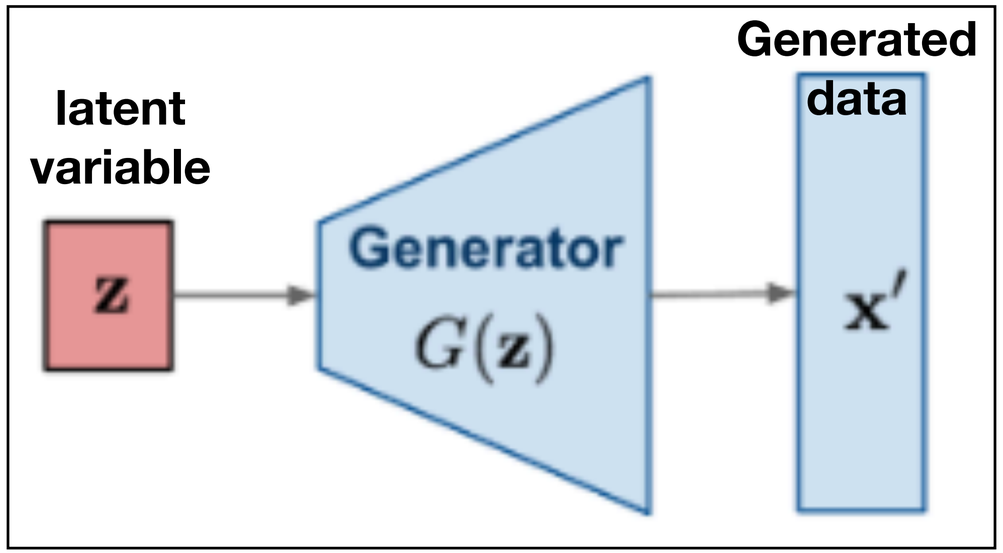
~ 30 min incl. break (10:10 - ~10:40)

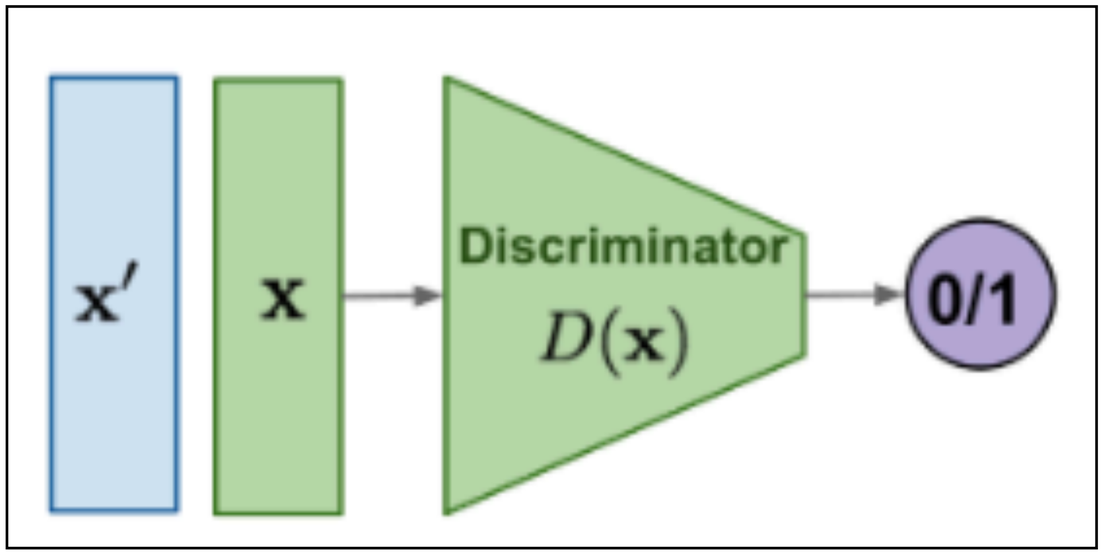
Generative Adversarial Network

GAN: Adversarial training



Generative Adversarial Network (GAN)



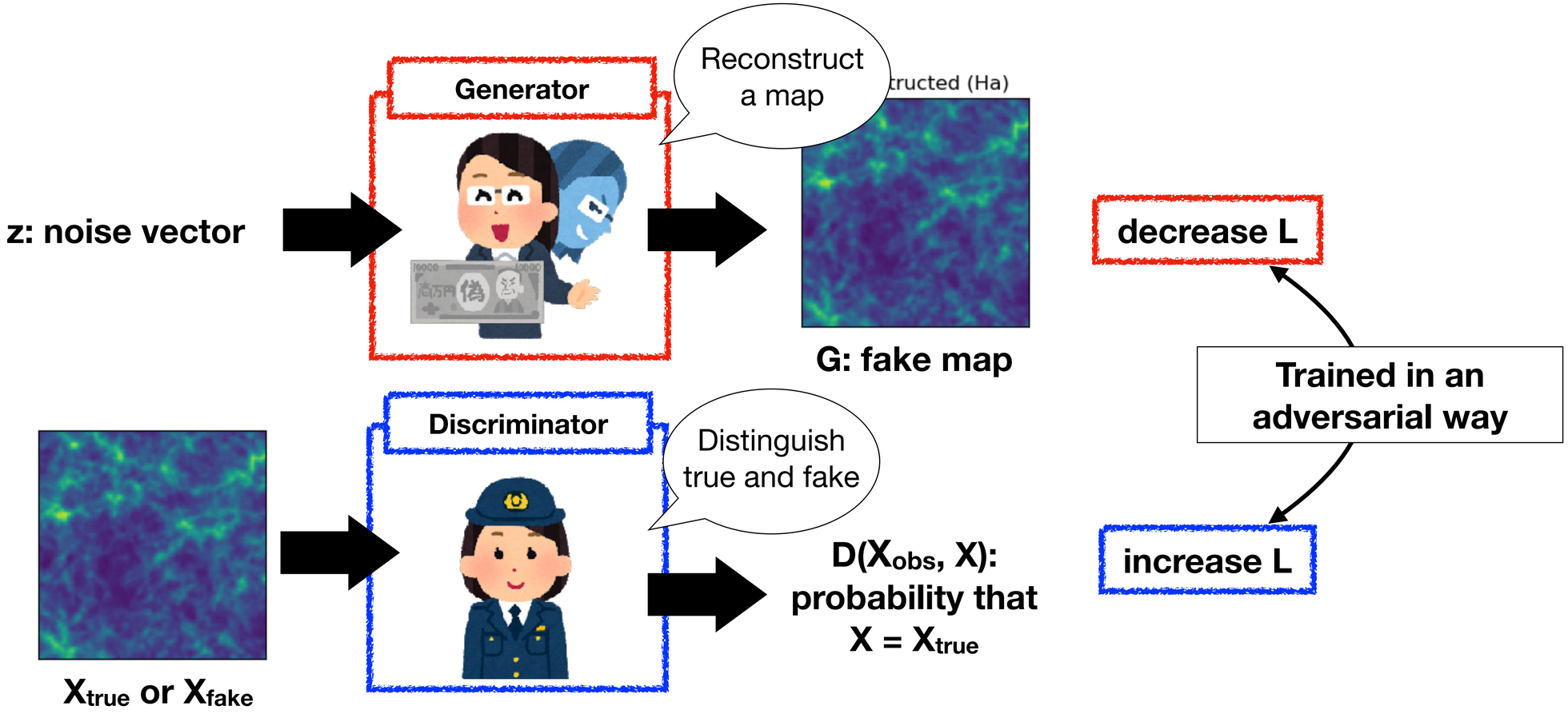


Very similar to decoder in VAE, but there is no encoder
→ How to train the model without corresponding *ground truth*?

Prepare another network!

Generative Adversarial Network (GAN)

- GAN: Generator and Discriminator are updated in an adversarial way.



Loss function: $L = \log D(X_{\text{true}}) + \log[1 - D(G)]$

Comparison

	Latent space dim.	Data Quality	Prob. Dist.	Sampling speed	Note
AE	low	×	×	○	
VAE	low	△	△	○	
GAN	low	○	×	○	Sometimes unstable
Flow					
Diffusion					

○: Good, △: so-so, ×: not good

Hands-on: GAN for SDSS galaxy spectrum

Questions:

- Is there any difference in generated data between VAE and GAN?

- How does the loss function look like? Does anyone observe training's instability?

- What happens if you adopt a large hidden dimension?

~30 min incl. break (10:50-11:20)

Comparison

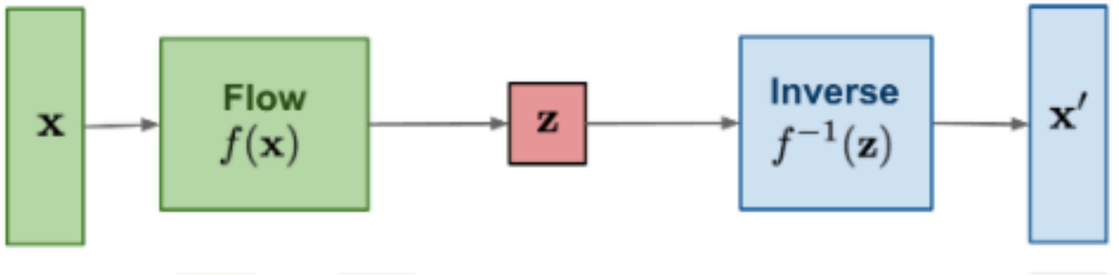
	Latent space dim.	Data Quality	Prob. Dist.	Sampling speed	Note
AE	low	×	×	○	
VAE	low	△	△	○	
GAN	low	○	×	○	Sometimes unstable
Flow					
Diffusion					

Can we use a latent space with high dimensionality to preserve the information in the data ?
→ Yes, but an appropriate data transformation must be adopted

○: Good, △: so-so, ×: not good

Flow-based Model

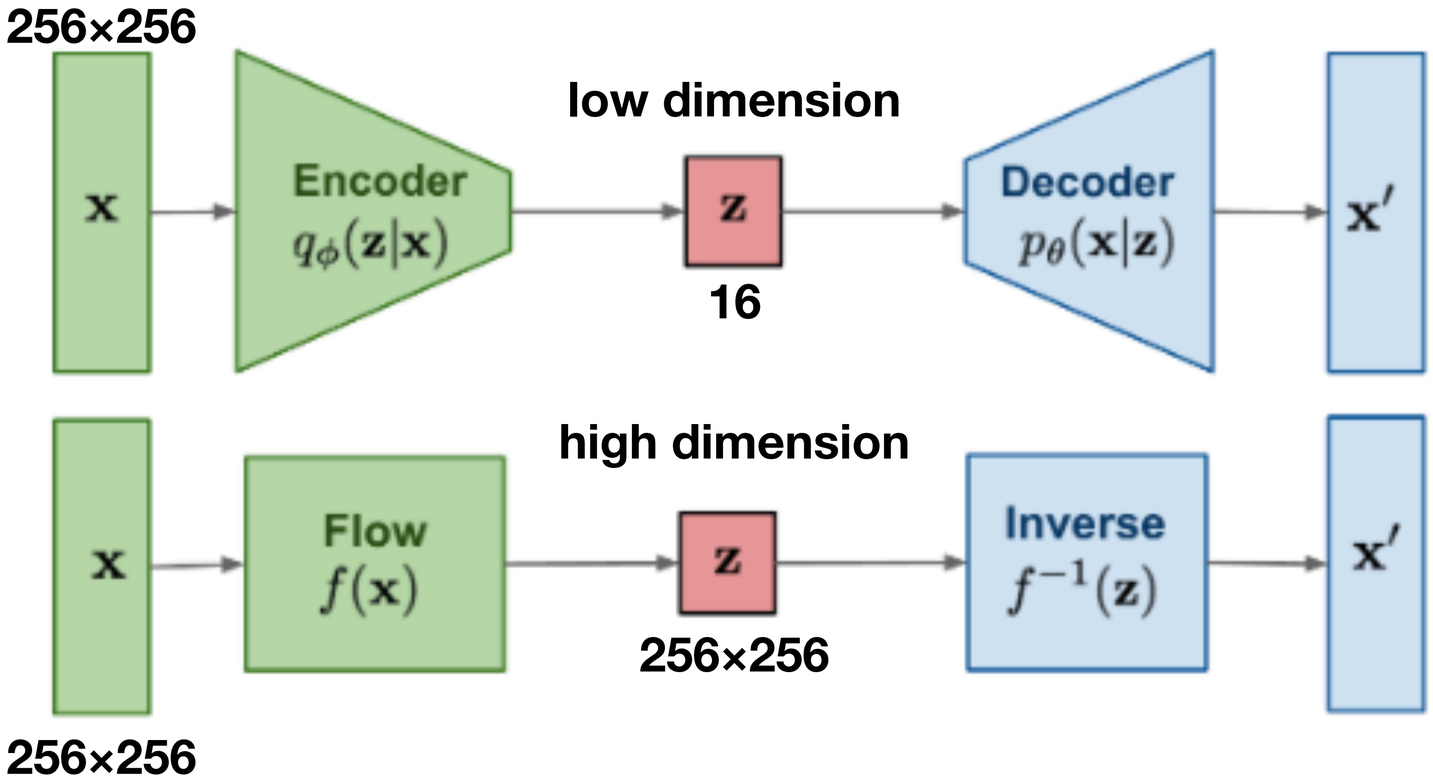
Flow-based models:
Invertible transform of distributions



Flow-based models

e.g., Dinh et al. 2014

VAE: maximize variational lower bound



Flow-based models: Invertible transform of distributions

Flow-based model:

- **Latent space has the same dimension the data space**
- **Transformation(s) are invertible and their determinants can be (easily) computed**

Flow-based models

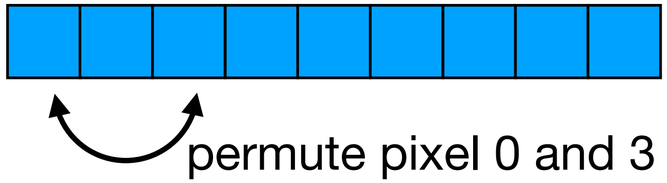
Transformations are invertible and their determinants can be (easily) computed

Examples:

$$f(x) = \alpha x \quad (\alpha \neq 0) \longrightarrow f^{-1}(x) = \frac{x}{\alpha}, \quad \det\left(\frac{\partial f(x)}{\partial x}\right) = \alpha$$

$$f(x) = \begin{pmatrix} 3 & 2 \\ 0 & 4 \end{pmatrix} x \longrightarrow f^{-1}(x) = \begin{pmatrix} 1/3 & -1/6 \\ 0 & 1/4 \end{pmatrix} x, \quad \det\left(\frac{\partial f(x)}{\partial x}\right) = 3 \times 4$$

f : permutation $\longrightarrow$ f^{-1} : inverse permutation, $\det\left(\frac{\partial f(x)}{\partial x}\right) = 1$

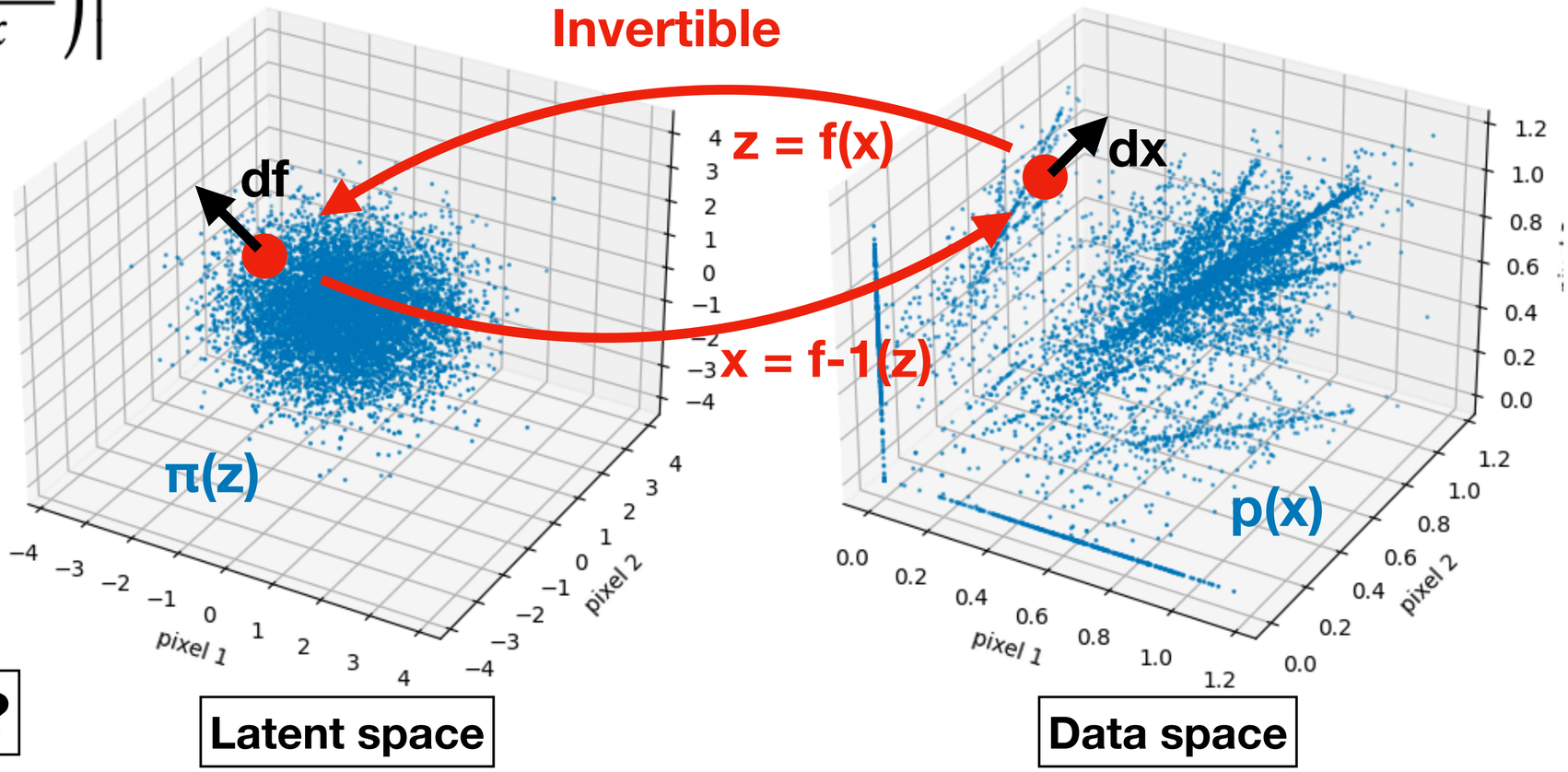


Flow-based models

Transformations are invertible and their determinants can be (easily) computed

$$p(x) = \pi(f(x)) \left| \det \left(\frac{\partial f(x)}{\partial x} \right) \right|$$

→ **data probability distribution p(x) can be computed if latent space distribution is known**

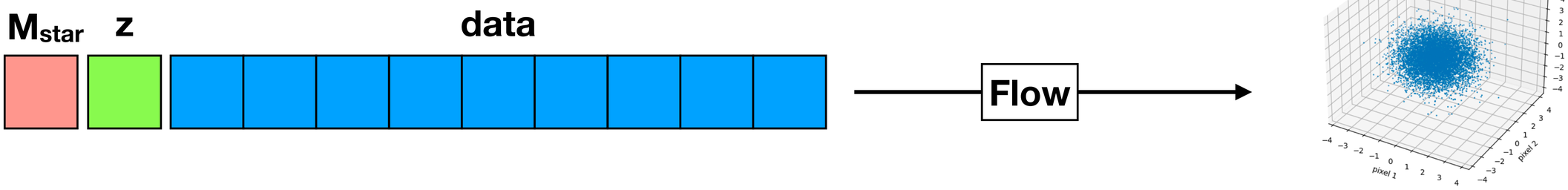


When useful in science?

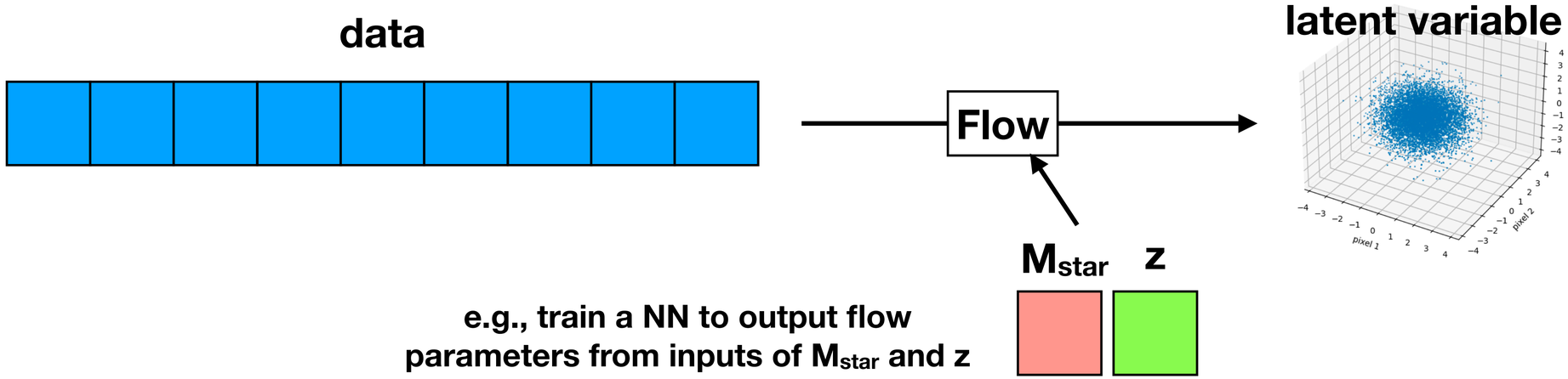
Parameter Inference with Flow-based Models

One can build a *conditional* flow model

approach 1: add parameters to data



approach 2: use parameters-dependent flows

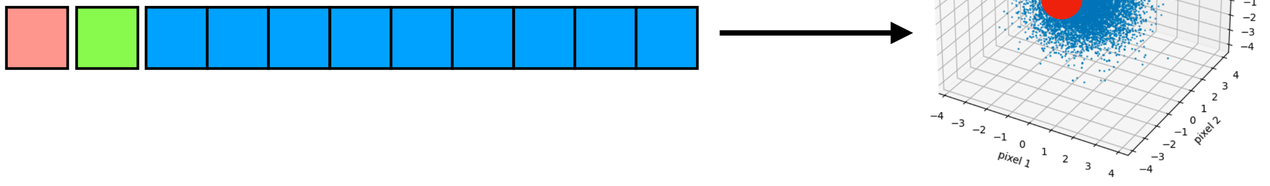


e.g., train a NN to output flow parameters from inputs of M_{star} and z

Parameter Inference with Flow-based Models

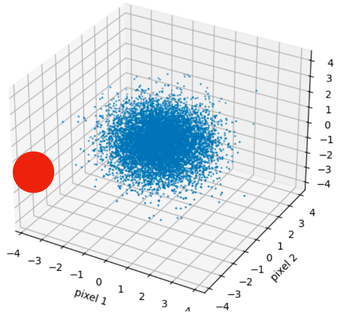
Correct parameters

M_{star} z data

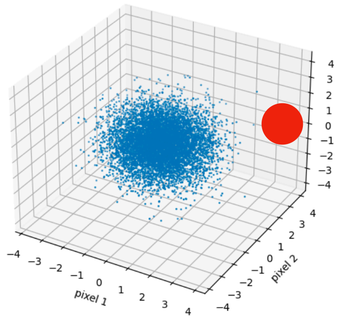


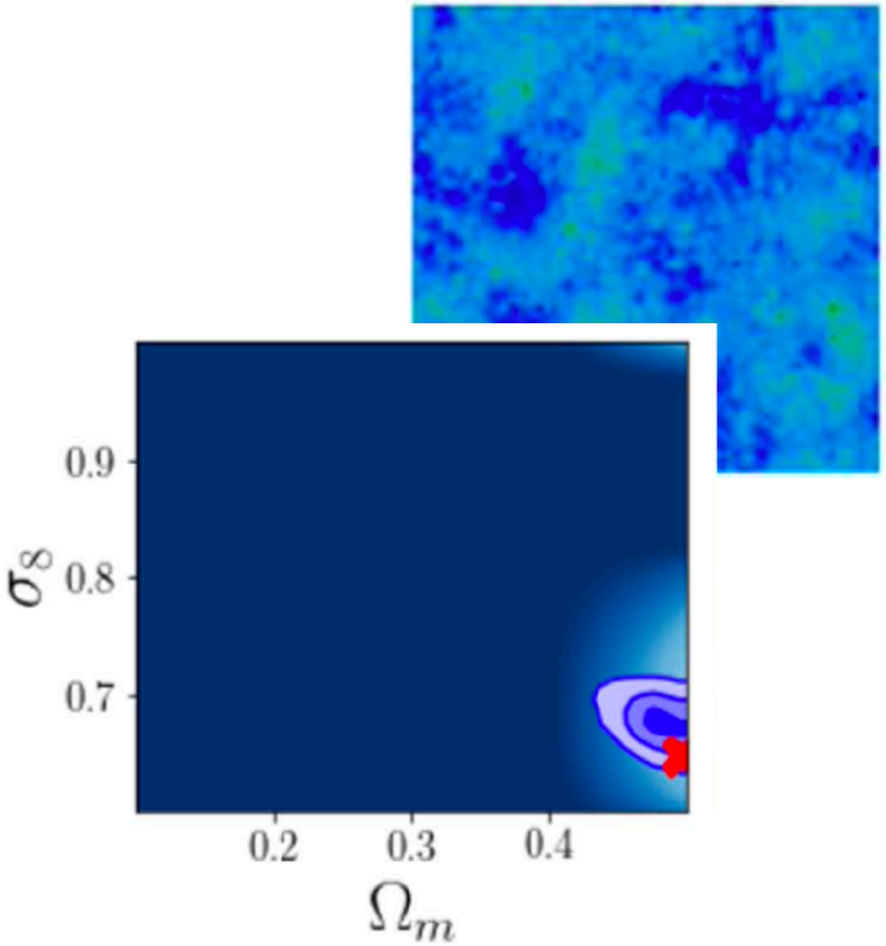
Wrong parameters

M_{star} z data



M_{star} z data





e.g., Hassan+2022: generate HI density field conditioned on cosmological parameters

Comparison

	Latent space dim.	Data Quality	Prob. Dist.	Sampling speed	Note
AE	low	×	×	○	
VAE	low	△	△	○	
GAN	low	○	×	○	Sometimes unstable
Flow	high	×/△	○	○	Less applicability for high-dim. data
Diffusion					

○: Good, △: so-so, ×: not good

Hands-on: Flows for SDSS galaxy spectrum

Questions:

- Can we reproduce the data?

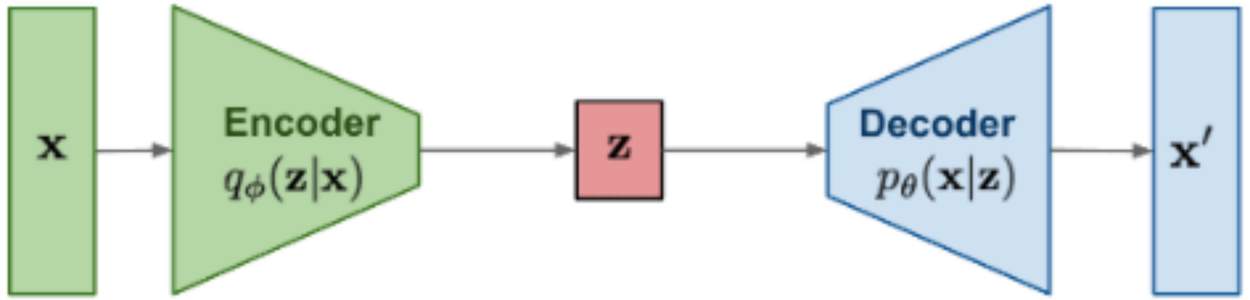
- Can we make the model conditional?

Diffusion model

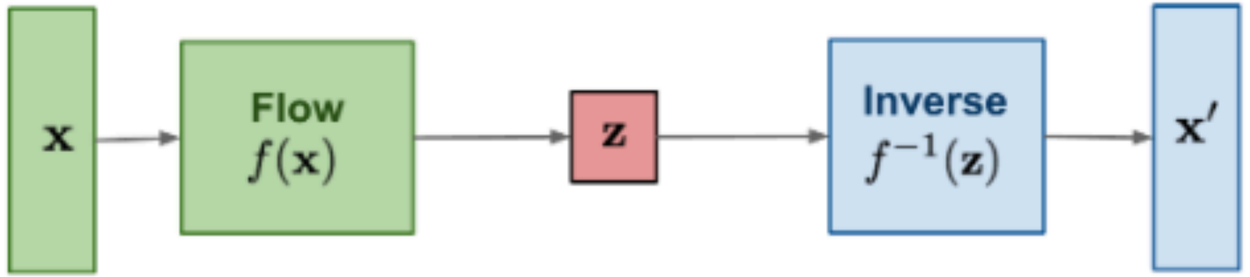
Diffusion models:
Gradually add Gaussian noise and then reverse

Diffusion Model

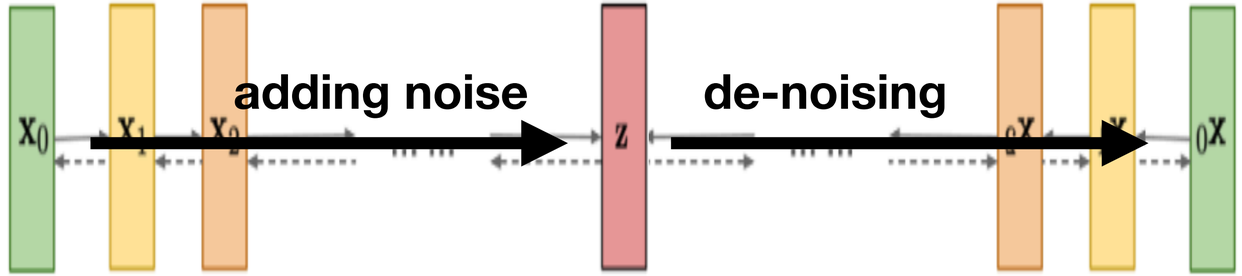
VAE: maximize variational lower bound



Flow-based models: Invertible transform of distributions



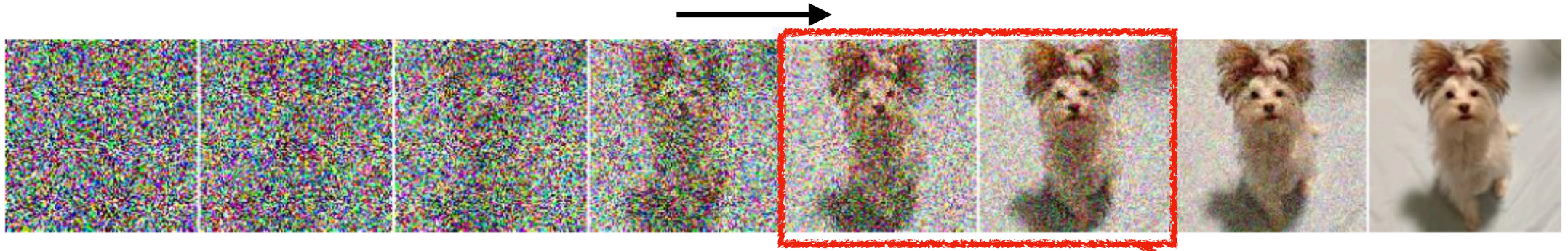
Diffusion models: Gradually add Gaussian noise and then reverse



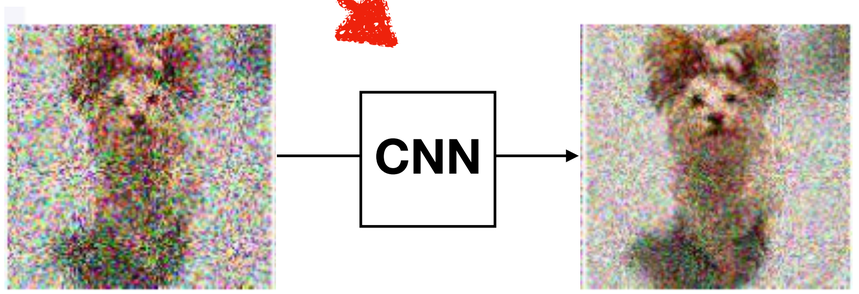
Diffusion Model

- Add simple Gaussian noises to encode images
- Relies on training a neural network under a simple denoising task

Backward process: learned



Forward process: add noise (fixed)



+ Simpler mechanism than GAN → training is more stable
- Slow sampling (partly solved; Song+2020)
- Could require more computational resources

Diffusion Models vs GANs!

Figure 6: Samples from BigGAN-deep with truncation 1.0 (FID 6.95, left) vs samples from our diffusion model with guidance (FID 4.59, middle) and samples from the training set (right).

Dhariwal & Nichol 2021

Applications of Diffusion Models to Astronomical Data

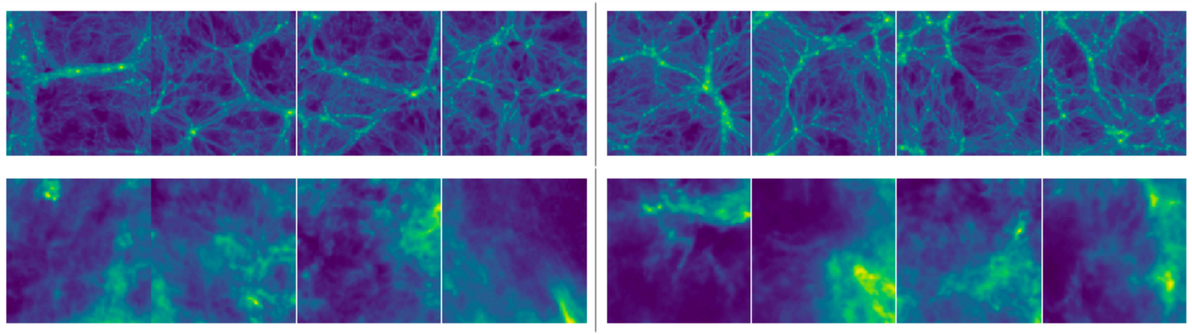
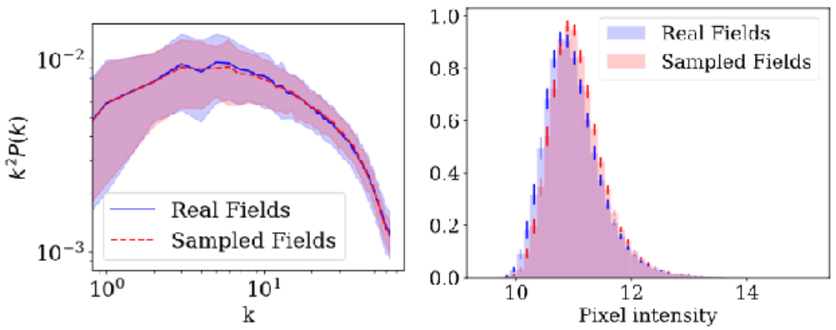
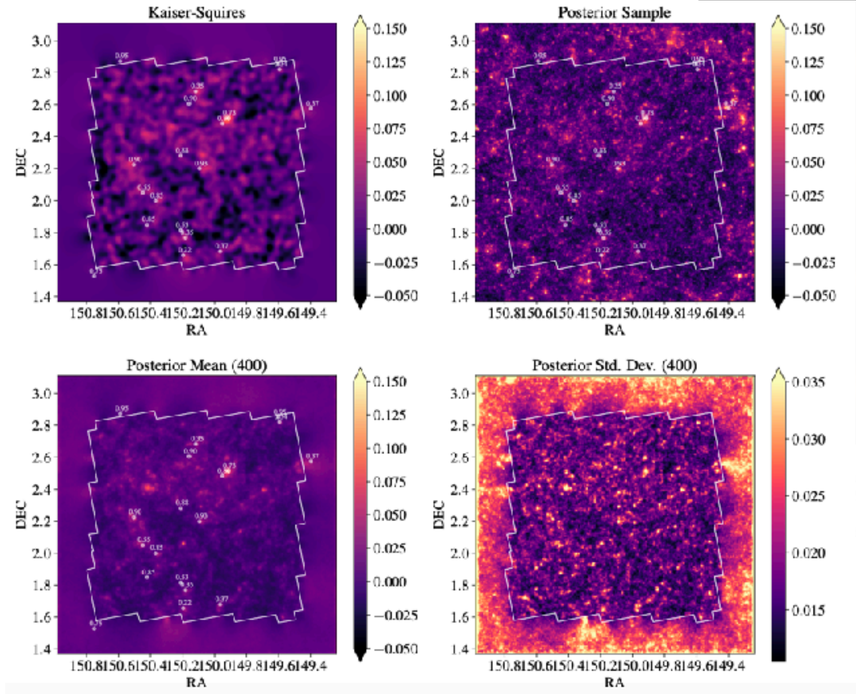


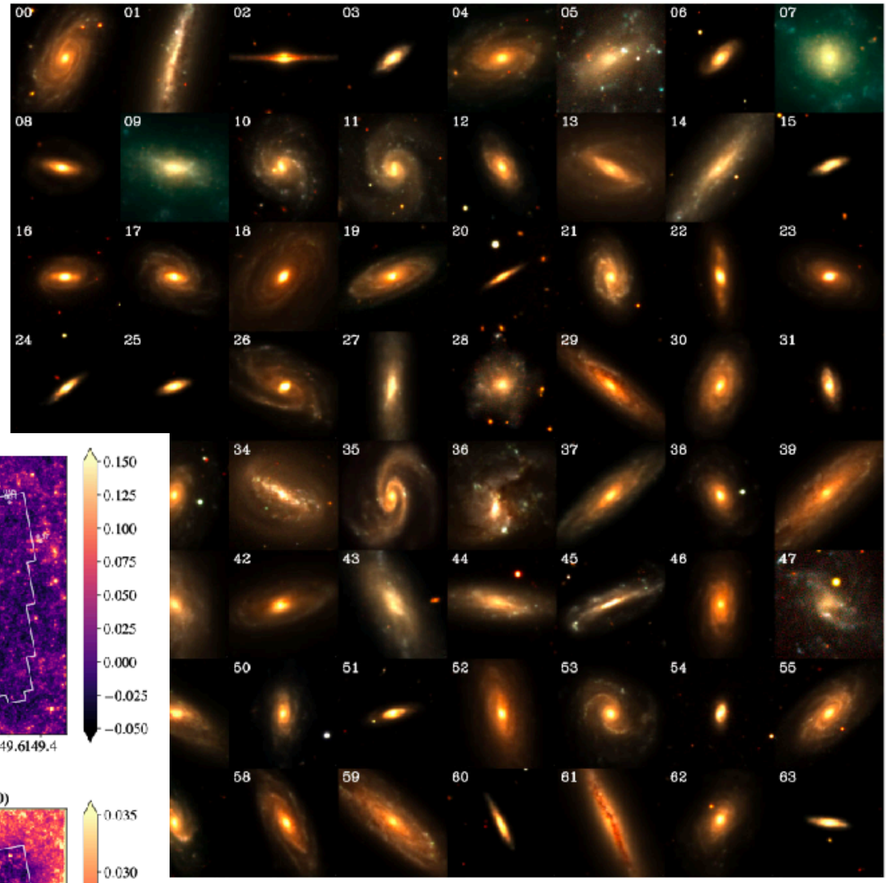
Figure 1: Four log cold dark matter mass density fields from the training data (top left) and from the sampled model (top right) at 128x128. Four samples of dust from the training data (bottom left) and from the trained model (bottom right).



Mudur & Finkbeiner 2022



Remy et al. 2023



ated galaxies designed to mimic the PROBES data set, interspersed with real examples from the data set itself. The images have real data split is 50/50. All images are *grz* RGB composites with identical scaling (we have performed a 99.5 per cent percentile brightness features). A key stating which galaxies are real and which are generated is provided at the end of this paper. More l at http://mjjsmith.com/thisisnotagalaxy.

Smith et al. 2022

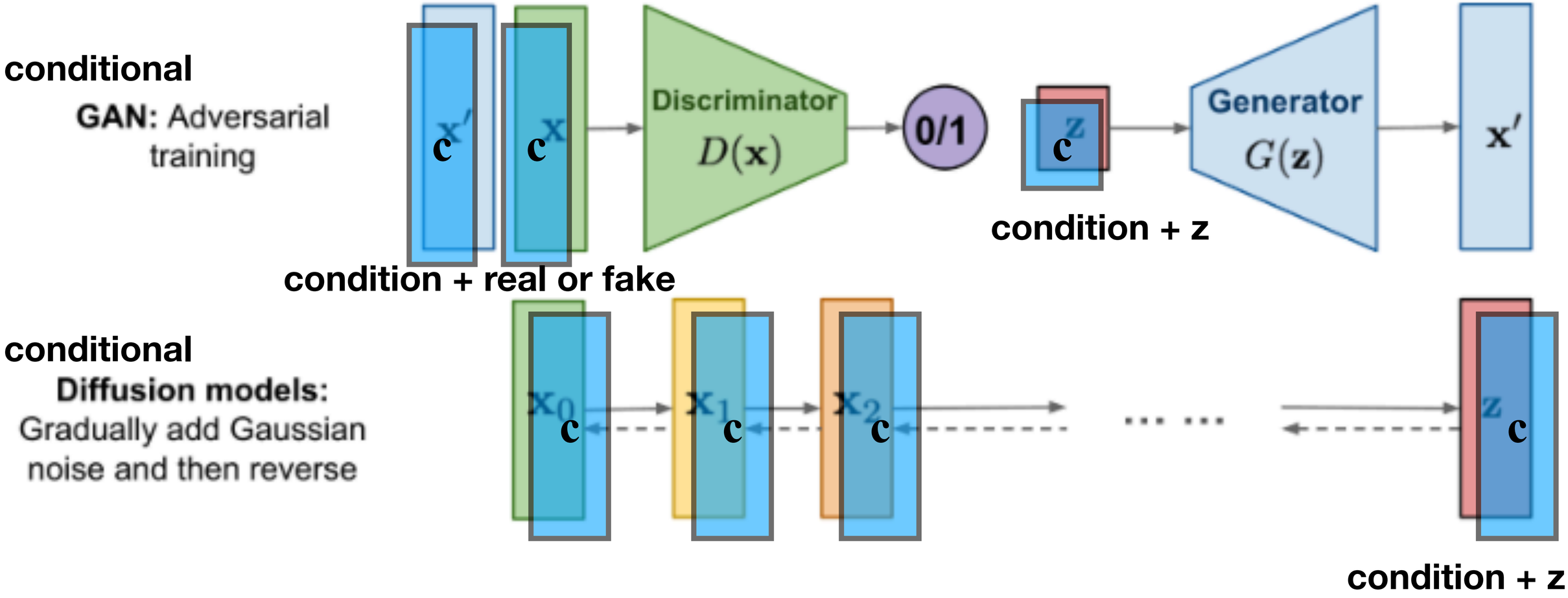
Comparison

	Latent space dim.	Data Quality	Prob. Dist.	Sampling speed	Note
AE	low	×	×	○	
VAE	low	△	△	○	
GAN	low	○	×	○	Sometimes unstable
Flow	high	×/△	○	○	Less applicability for high-dim. data
Diffusion	high	○	△	△	Could need more computational resources

○: Good, △: so-so, ×: not good

Conditional Generative Model

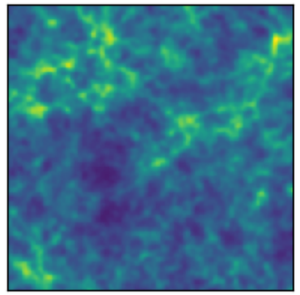
I introduced flow-based models conditioned on a few parameters, but one can make a model conditioned on higher-dimensional data (e.g., image-to-image model)



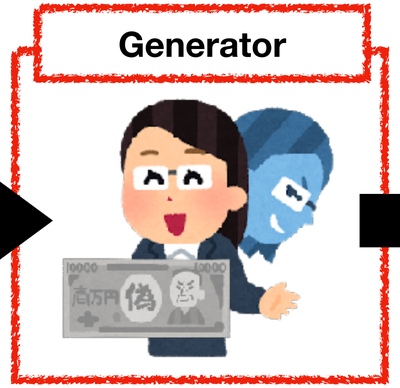
Conditional GAN for Line Intensity Mapping

Noisy observation → **Clean signal**

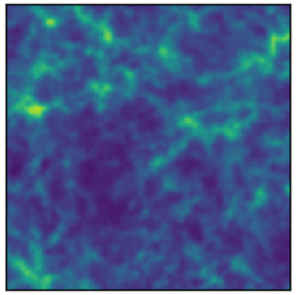
observed (Ha+OIII)



X_{obs}: observed map

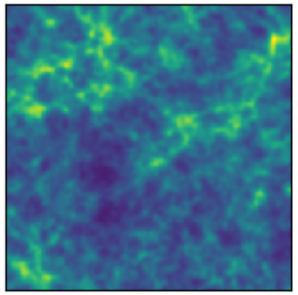


reconstructed (Ha)

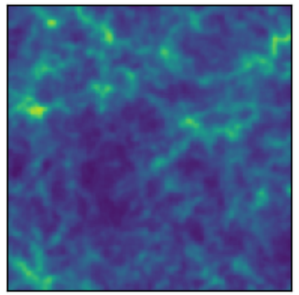


$G(X_{obs})$: fake map

observed (Ha+OIII)



X_{obs}: observed map



X_{true} or X_{fake}

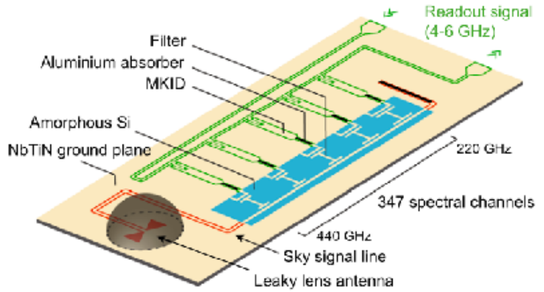
$D(X_{obs}, X)$: probability that $X = X_{true}$

See Moriwaki+2020 for more details

Loss function:

$$L[G,D] = \log D(X_{\rm obs}, X_{\rm true}) + \log[1 - D(X_{\rm obs}, G(X_{\rm obs}))] + \lambda \langle |X_{\rm true} - G(X_{\rm obs})| \rangle$$

A new LIM project is just launched! Contact me if you are interested in joining our project



DESHIMA 2.0

Taniguchi et al., J. Low Temp. Phys. (2022)

https://sites.google.com/view/sublime-tifuun/

Summary

Which models to use?

- Want to do data compression or clustering analysis not only data generation?
→ VAE

- Want generate high-quality data?
→ GAN or diffusion models

- Want to do parameter inference?
→ Flow-based models

	Latent space dim.	Data Quality	Prob. Dist.	Sampling speed	Note
AE	low	×	×	○	
VAE	low	△	△	○	
GAN	low	○	×	○	Unstable?
Flow	high	×/△	○	○	Less applicabilit
Diffusion	high	○	△	△	More resources?