Tutorial On

PROBABILISTIC U-NETs

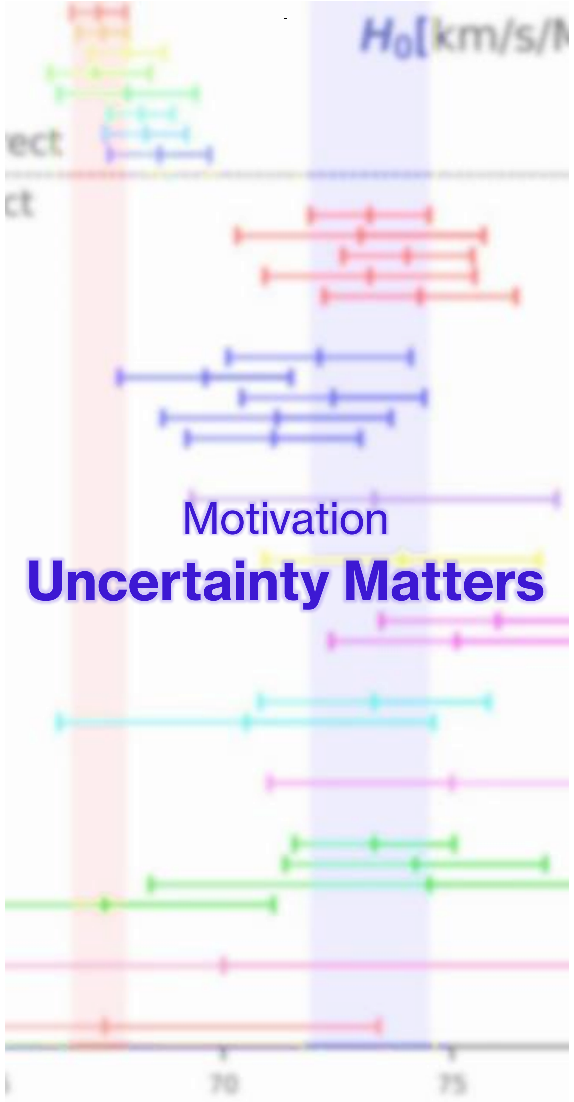
Hadi Sotoudeh

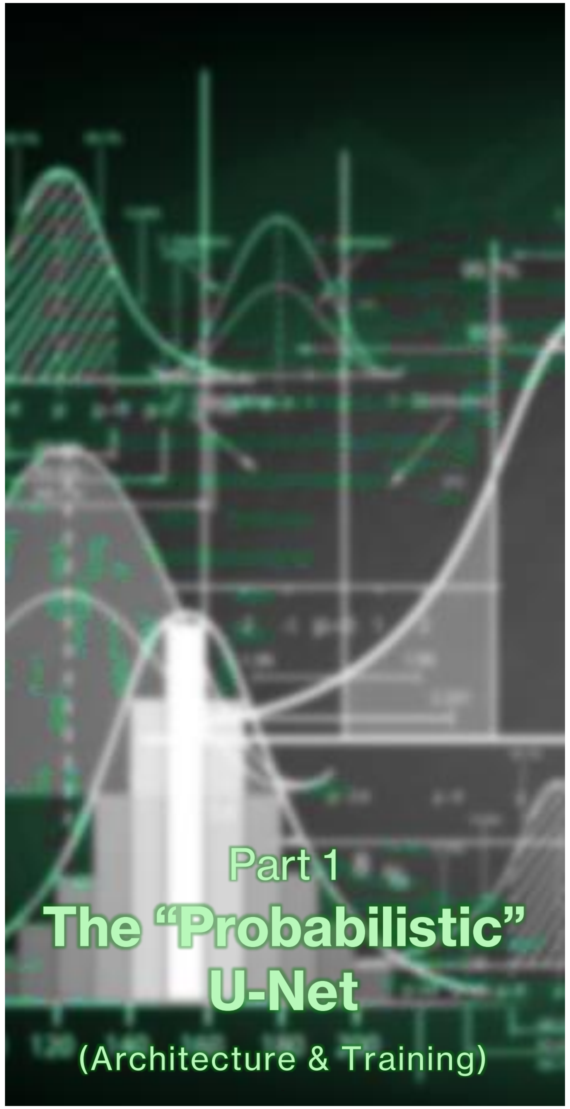
Université de Montréal   Mila   Ciela

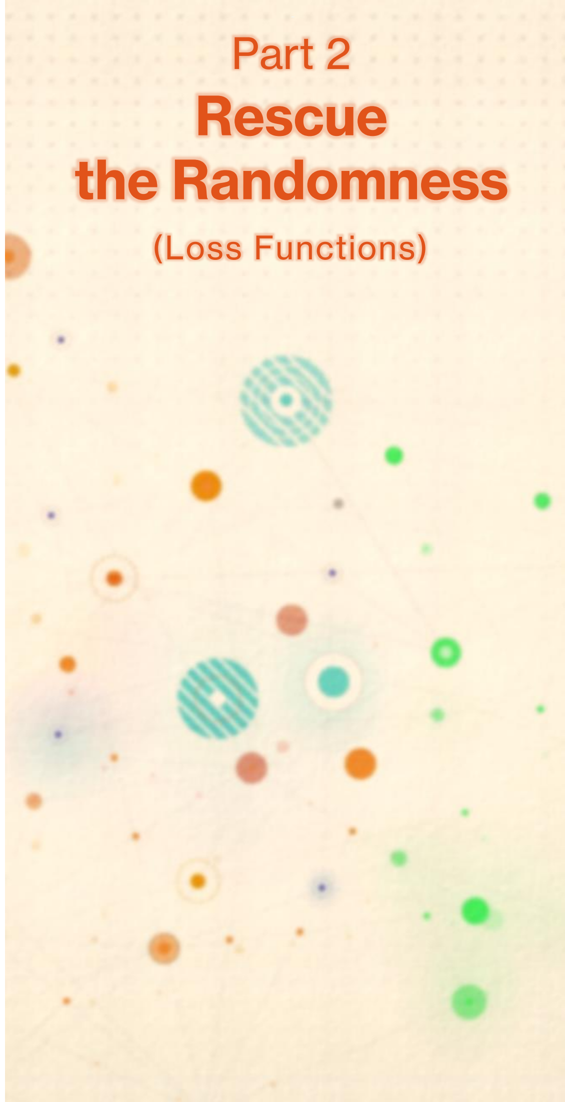
KITP Program on Building a Physical Understanding of Galaxy Evolution with Data-driven Astronomy   23 Feb 2023

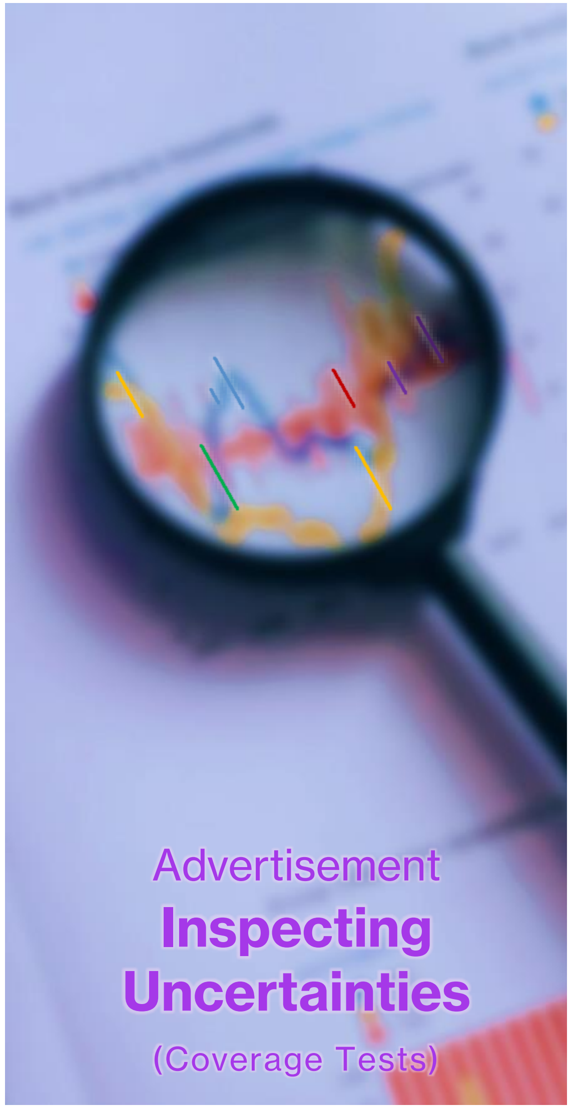
Access Repo

Outline


Motivation
Uncertainty Matters

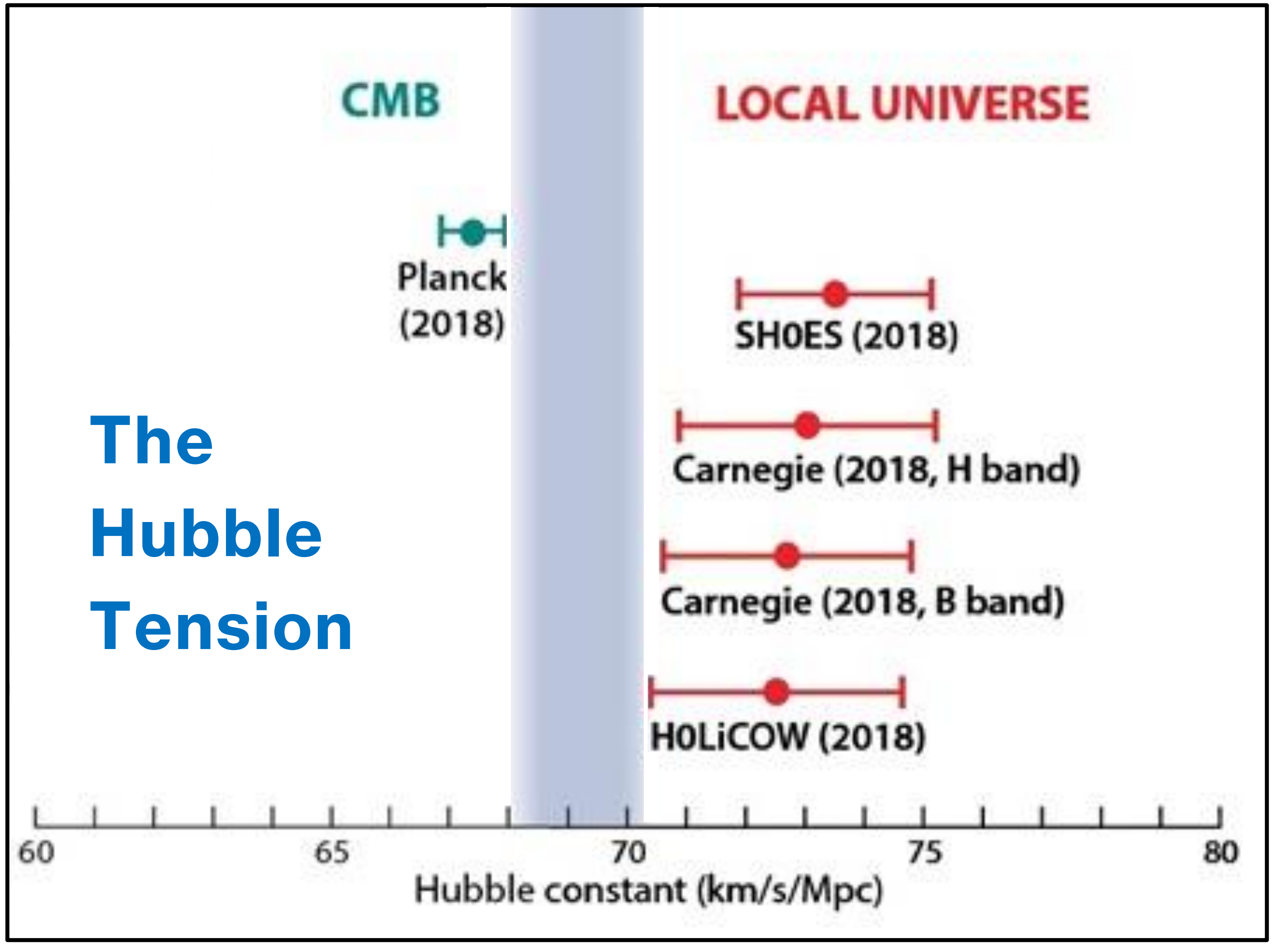

Part 1
The "Probabilistic"
U-Net
(Architecture & Training)


Part 2
Rescue
the Randomness
(Loss Functions)


Advertisement
Inspecting
Uncertainties
(Coverage Tests)

Uncertainty Matters

- Every physical measurement is meaningful with an uncertainty estimate.

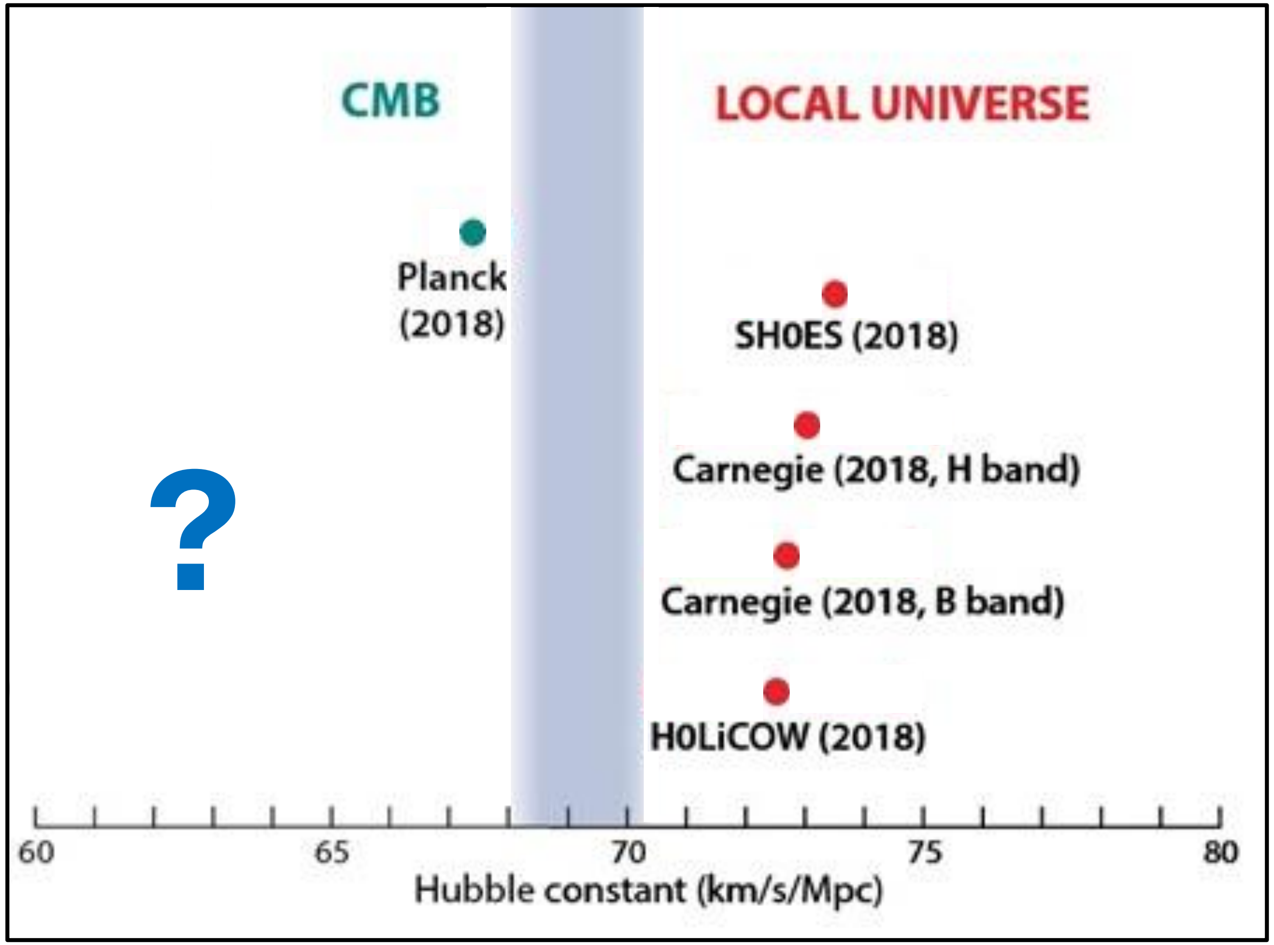


Credit: Roen Kelly, *Astronomy* Magazine – Figure taken with modification.

Deep Learning for Physical Discoveries

- Traditional deep learning methods → No uncertainty estimate

- Physical applications require models capable of quantifying uncertainties.



#uncertainty_quantification

Credit: Roen Kelly, *Astronomy* Magazine – Figure taken with modification.

Deep Generative Models

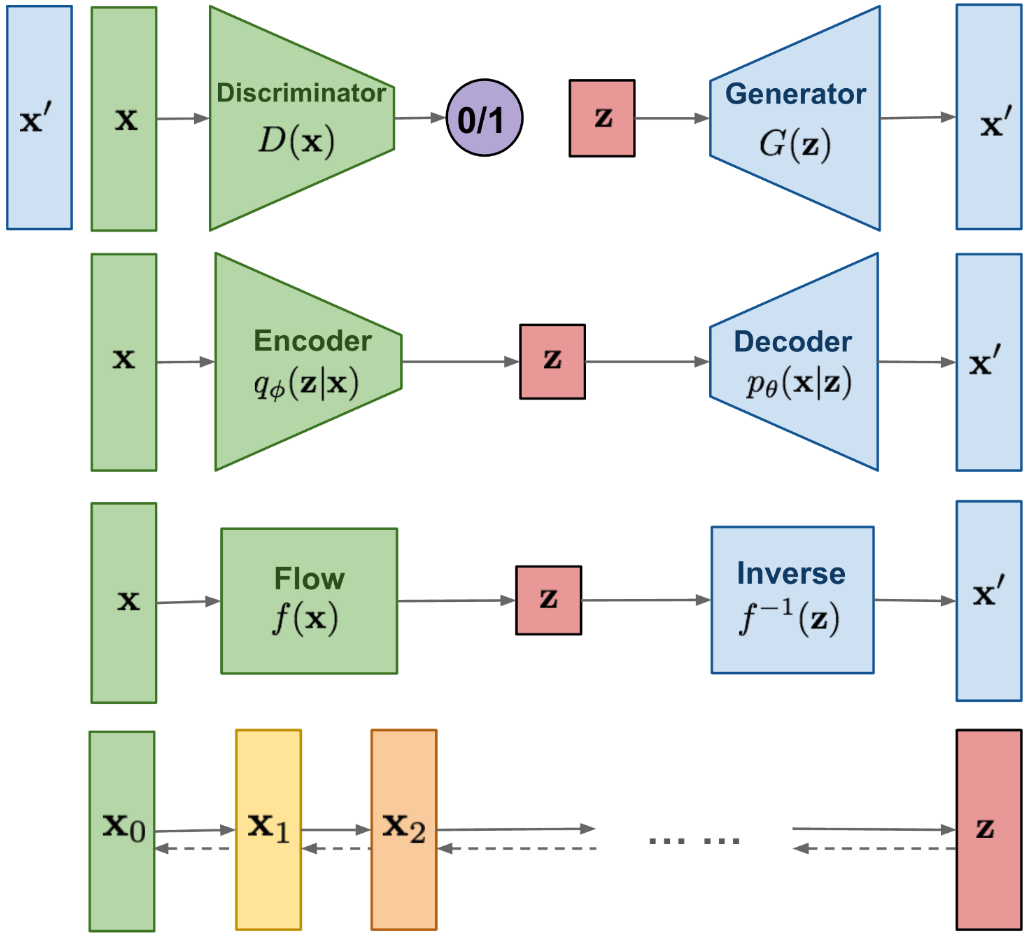
- Can learn to generate new data that resembles a distribution → Can encode uncertainties

GAN: Adversarial training

VAE: maximize variational lower bound

Flow-based models: Invertible transform of distributions

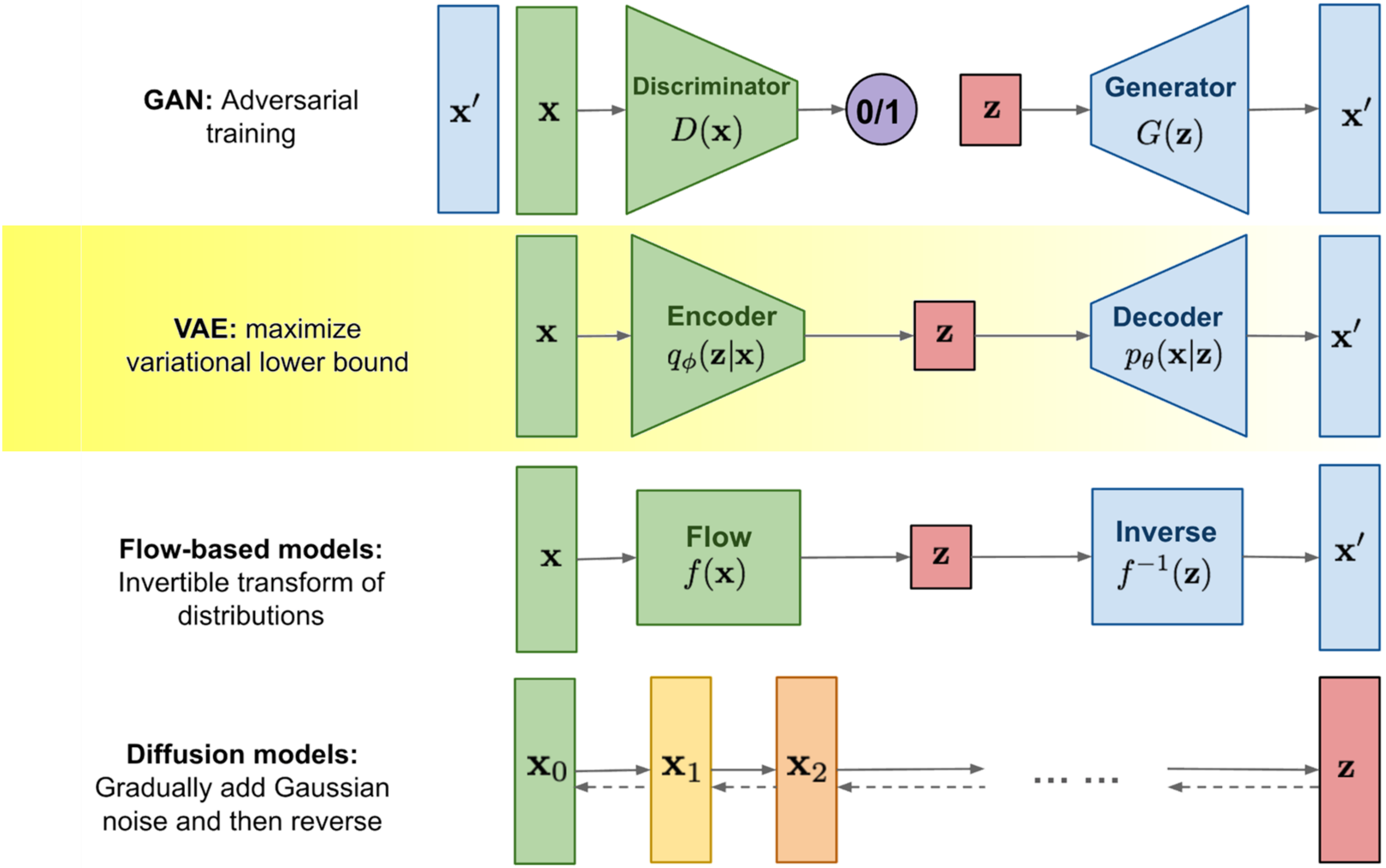
Diffusion models: Gradually add Gaussian noise and then reverse



Credit: Lilian Weng's Personal Blog: lilianweng.github.io

Prob. U-Net

- Advantages:
  - Relatively lower computational cost during both training & inference
  - Well-understood theoretical framework



Credit: Lilian Weng's Personal Blog: lilianweng.github.io

Part 1

The “Probabilistic” U-Net

- Autoencoder, VAE & cVAE
- U-Net
- Probabilistic U-Net
- Training
- Toy Problem 1: Source Reconstruction

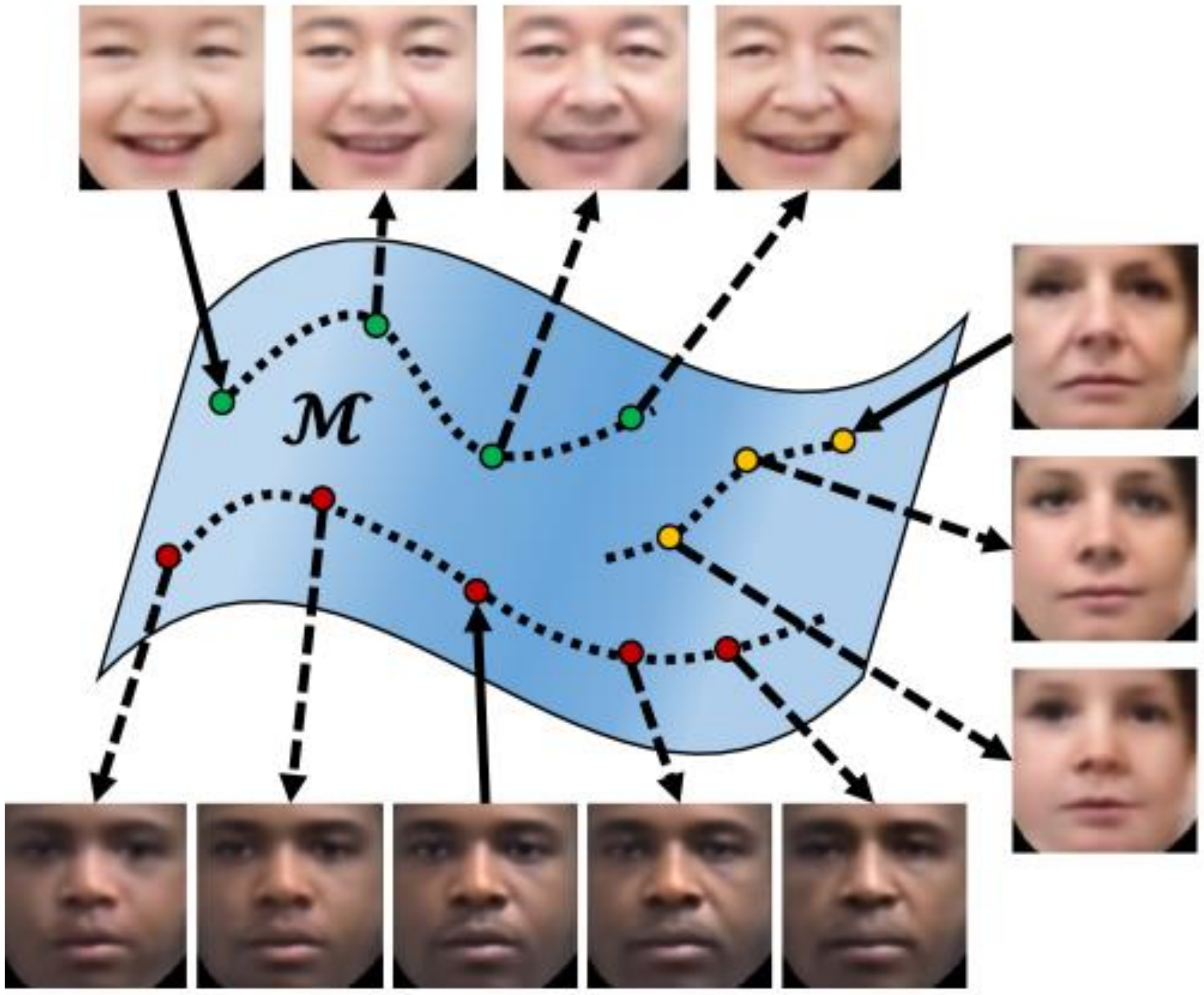
Curse of Dimensionality

pixel space is overparameterized!

Manifold Hypothesis

- Many real-world high-dim datasets lie along low-dim **latent manifolds** inside that space

- Manifold of valid human faces
    - If accessible, can easily draw samples from the distribution of valid faces

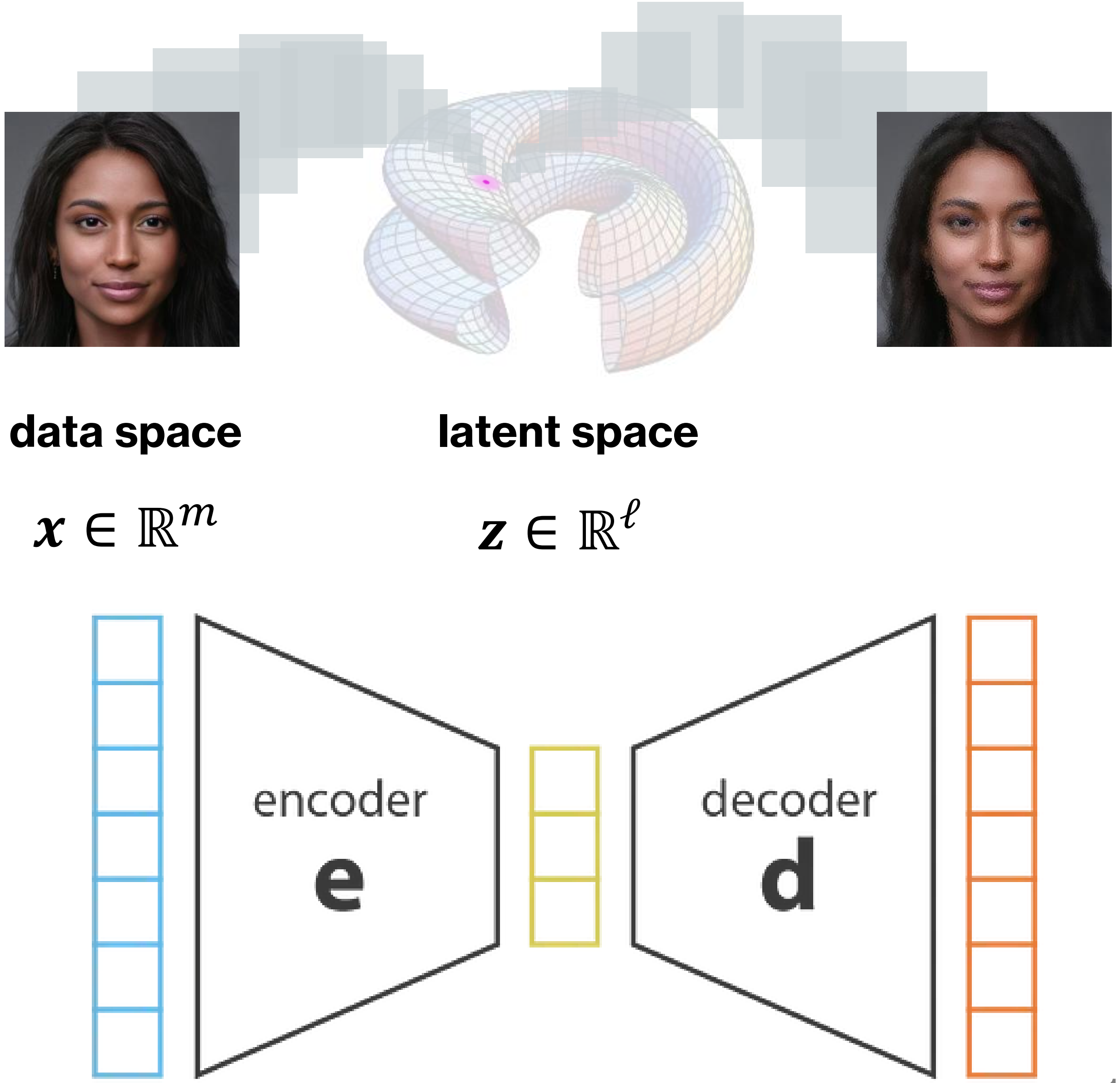


Credit: Zhifei Zhang et al. (2017)

Latent Space

- Dimensionality lower than data space ($\ell < m$)

- Defined by the **encoder** & **decoder**



Latent Space

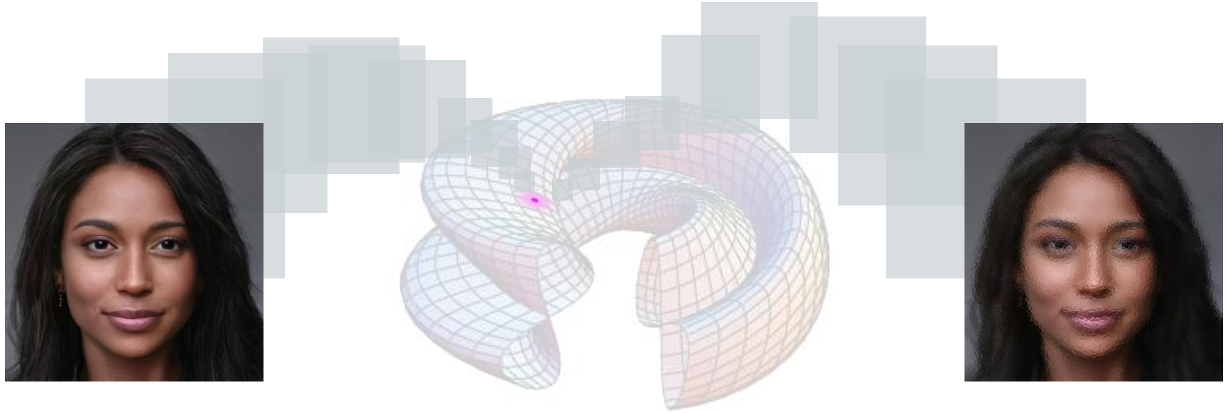
- **Goal:** Best encoder-decoder pair
  - keep maximum information
  - minimize reconstruction error

- Reconstructed Image: $\hat{\boldsymbol{x}} \coloneqq d\big(e(\boldsymbol{x})\big)$

- Examples of reconstruction error:
  - $\mathrm{MSE}(\boldsymbol{x},\ \hat{\boldsymbol{x}}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2$
  - $\mathrm{BCE}(\boldsymbol{x},\ \hat{\boldsymbol{x}}) = -\sum x_i \log \hat{x}_i + (1 - x_i) \log(1 - \hat{x}_i)$

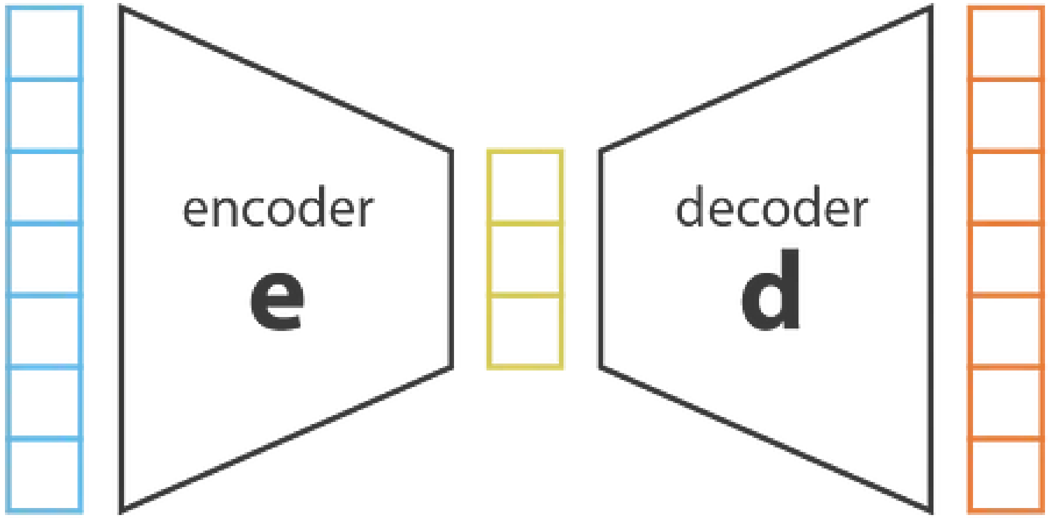
#dimensionality_reduction

#representation_learning

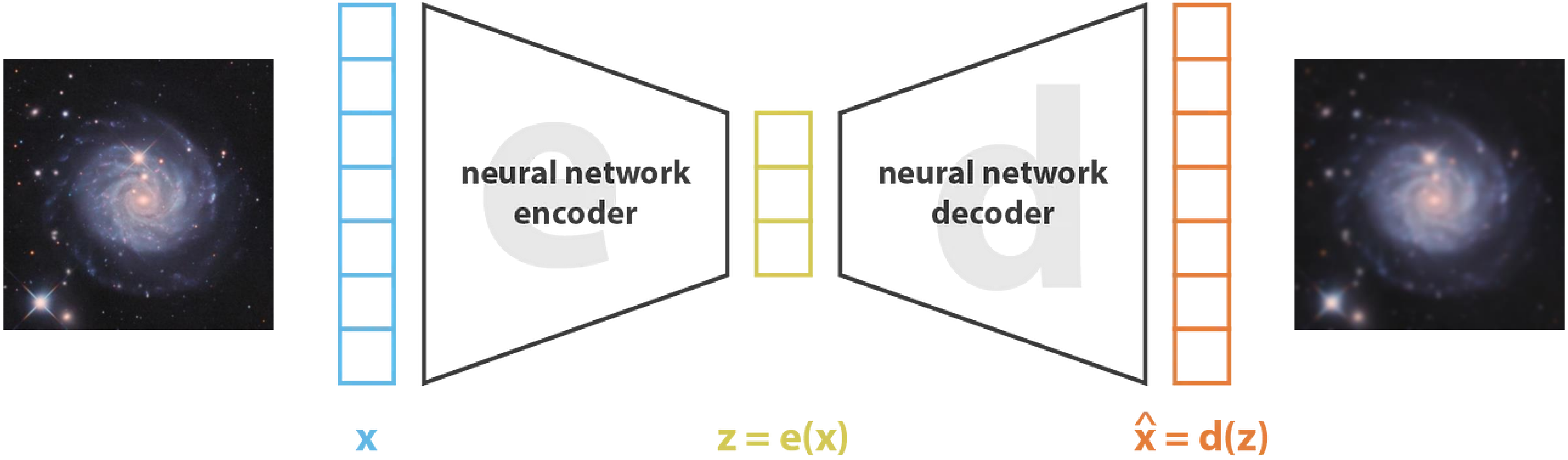


data space **latent space**

$\boldsymbol{x} \in \mathbb{R}^m$ $\boldsymbol{z} \in \mathbb{R}^\ell$



Autoencoder



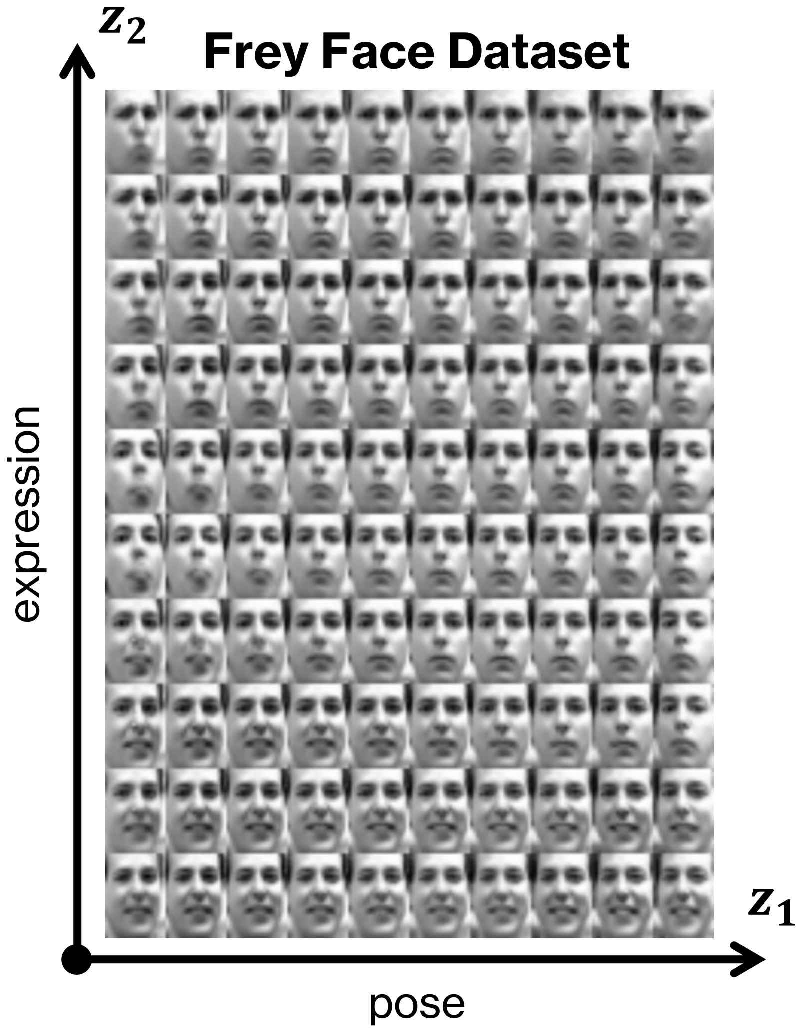
$$\mathcal{L}_{\text{rec}} = \text{MSE}(\boldsymbol{x}, \hat{\boldsymbol{x}})$$

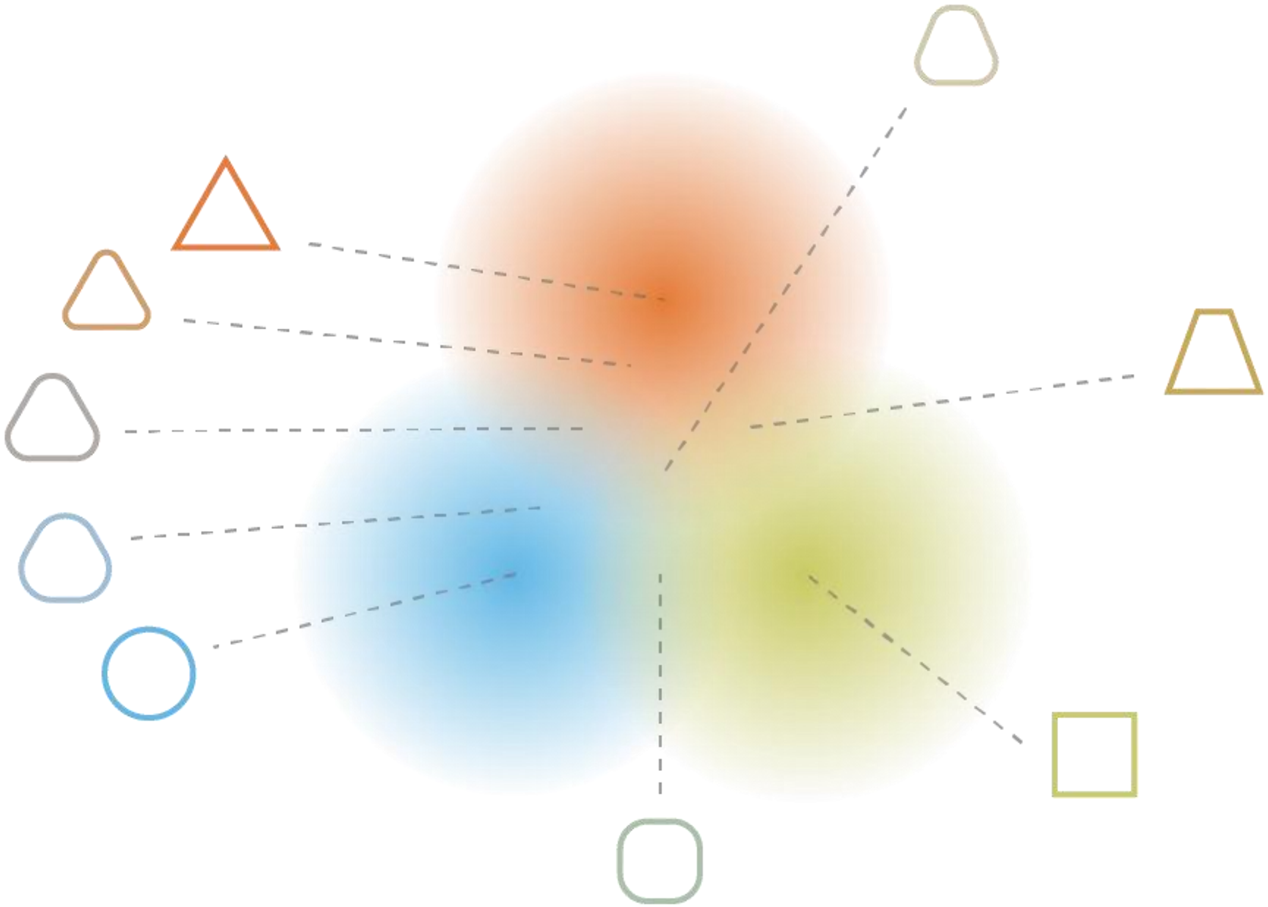
Problem: Irregular Latent Space

- Autoencoders only focus on reconstruction → Don't care about the structure of latent space
  - Tend to learn **punctual distributions**
- Latent space should be **continuous** and **complete**

Credit: Joseph Rocca's Post: towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Ideal Latent Space

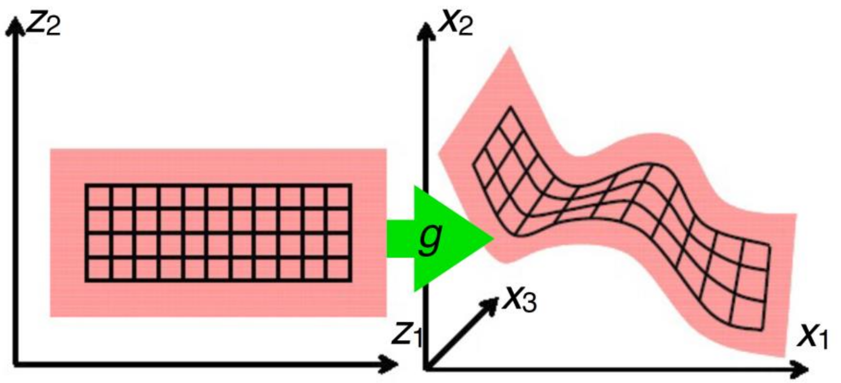




Credit: Joseph Rocca's Post:
towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Probabilistic Setup

- Learn a mapping from some latent distribution on $\boldsymbol{z}$ to a complicated distribution on $\boldsymbol{x}$



- Sample from the prior distribution in latent space → Map the sample to data space

$p(\boldsymbol{z})$ = something simple $\qquad\qquad$ $p(\boldsymbol{x}|\boldsymbol{z})$ modeled by generator

- Learn representation such that the marginal **data likelihood** (evidence) is maximized:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z})\, d\boldsymbol{z} \qquad \text{where} \qquad p(\boldsymbol{x}, \boldsymbol{z}) = \underbrace{p(\boldsymbol{x}|\boldsymbol{z})}_{\text{likelihood}}\ \underbrace{p(\boldsymbol{z})}_{\text{prior}}$$

likelihood prior

Variational Inference

Problem: $p(x) = \int p(x,z)\, dz$ is intractable

- Variational inference approach: Find a lower bound for the integral using an auxiliary distribution (q)

variational posterior — true posterior

$$\ln p(\boldsymbol{x}) = \underbrace{\int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z}}_{\text{Evidence Lower BOund (ELBO)}} - \underbrace{\int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z}}_{\text{Variational Gap}}$$

$$\ln p(\boldsymbol{x}) \geq \text{ELBO}$$

This is what we will try to maximize!

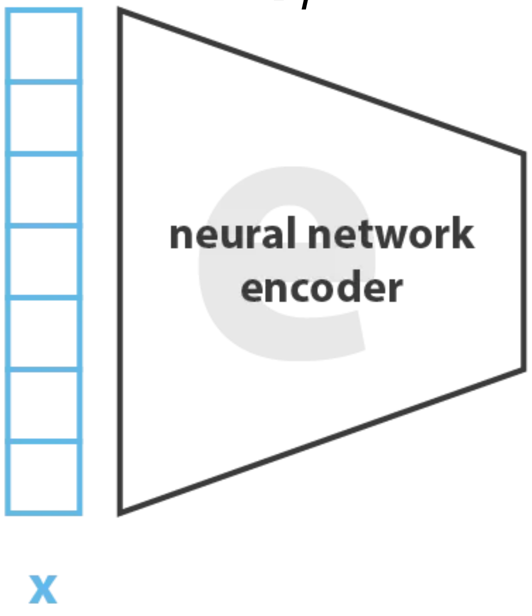
#approximate_inference

Variational Autoencoder

- Based on variational inference

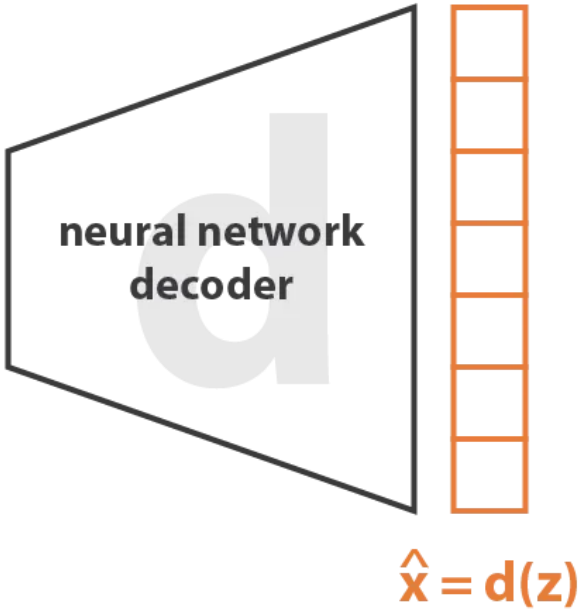
"implicitly"
models variational posterior

$q_\phi(\boldsymbol{z}|\boldsymbol{x})$

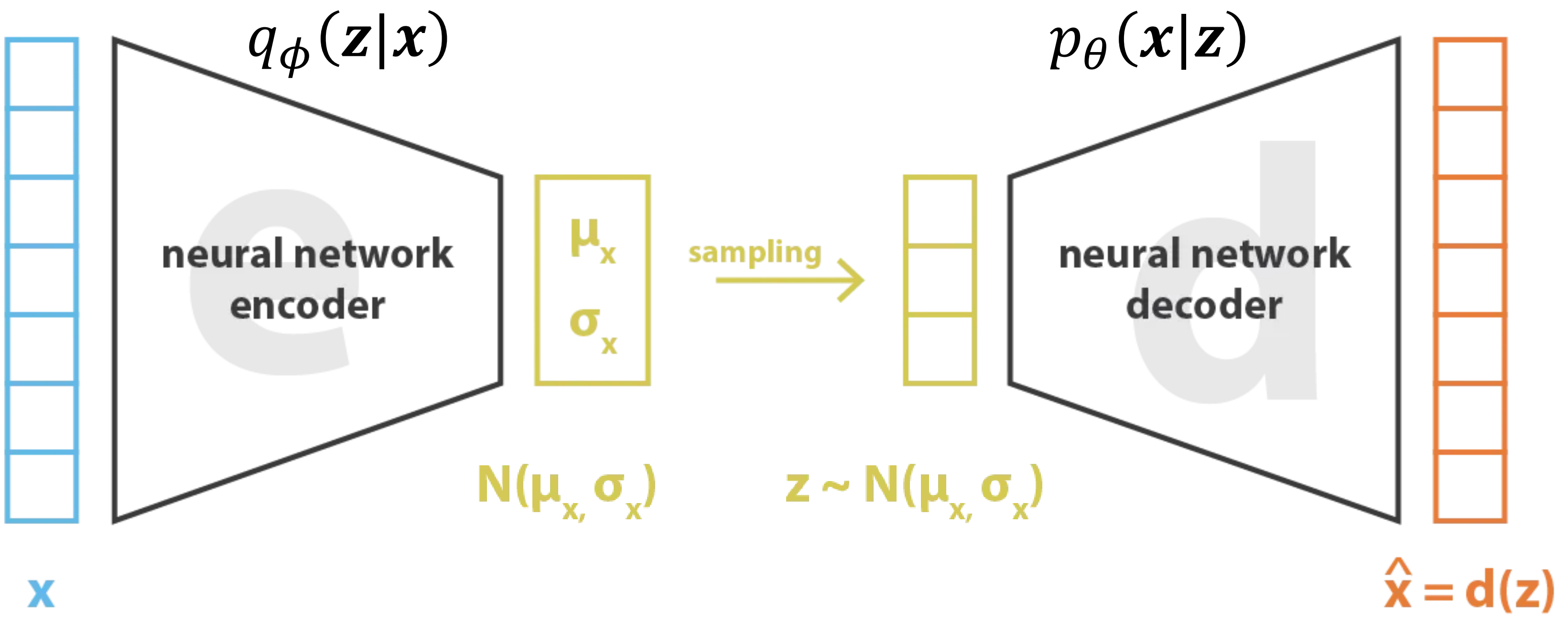


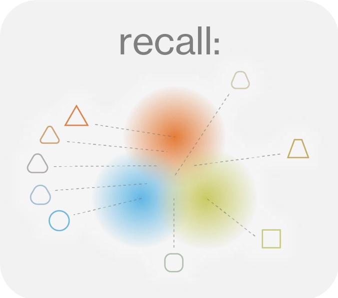
"implicitly"
models likelihood

$p_\theta(\boldsymbol{x}|\boldsymbol{z})$



Variational Autoencoder





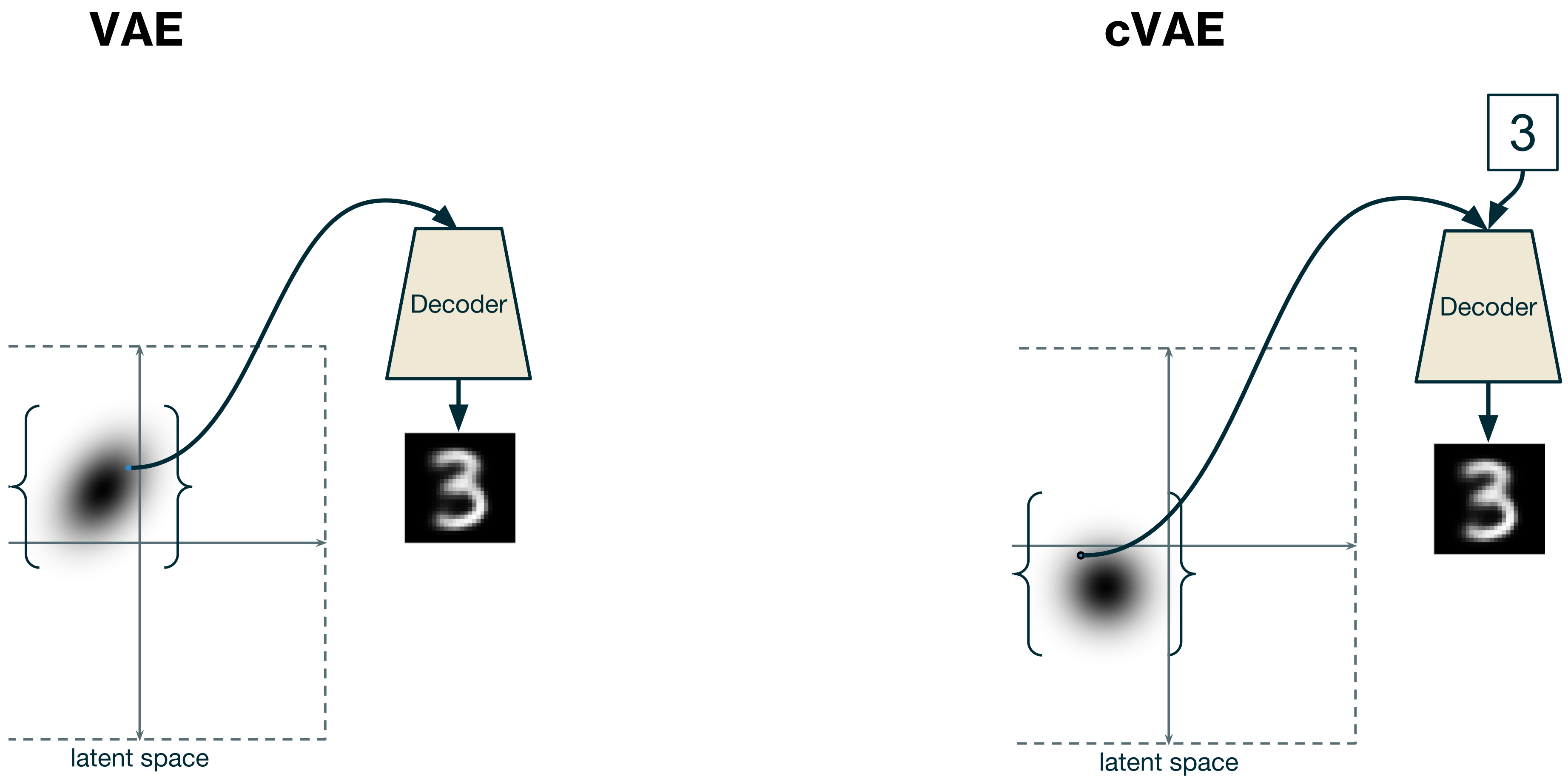
$$\text{ELBO}(\theta, \phi, \boldsymbol{x}) = \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \ \ln\left(\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right) dz$$

$$\mathcal{L}_{\text{ELBO}} = -\text{ELBO}(\theta, \phi, \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[-\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})] + D_{\text{KL}}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})\right)$$

$\mathcal{L}_{\text{rec}}$ (*reconstruction* term) $\mathcal{L}_{\text{KL}}$ (*regularization* term)

Conditional VAE

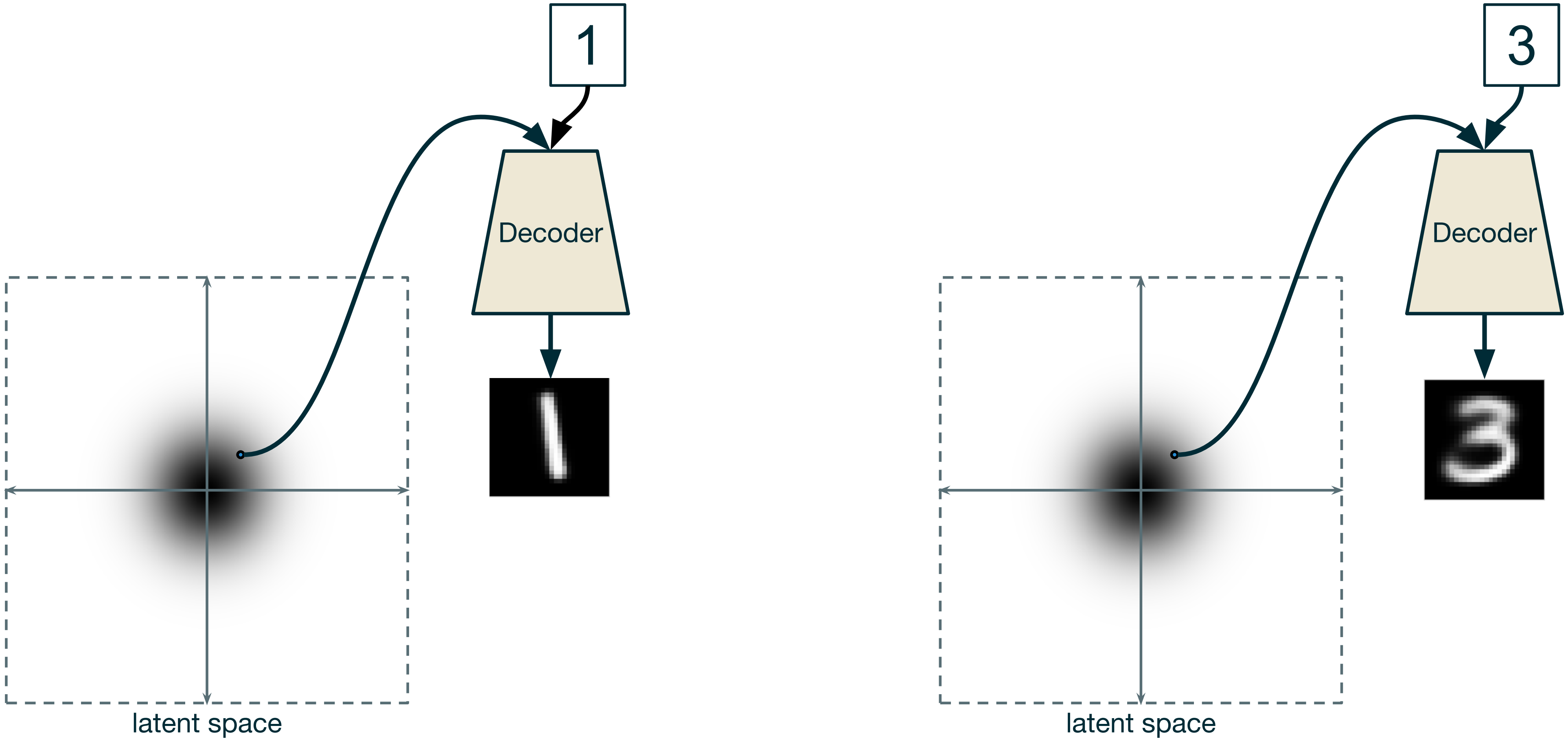
- How to generate samples of a particular class?



Credit: Isaac Dykeman's Personal Blog: ijdykeman.github.io

Conditional VAE

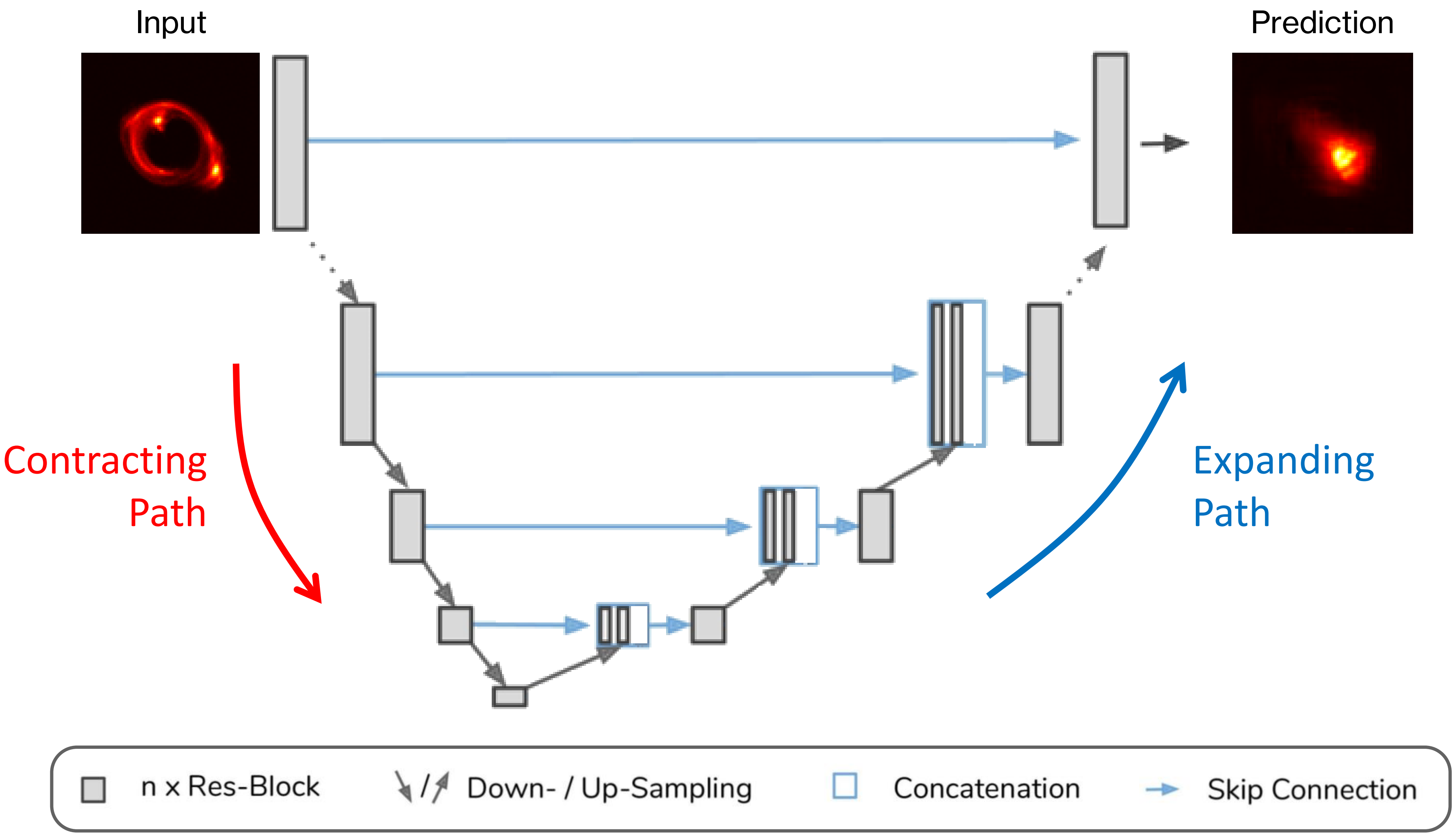
- Used to generate samples of a particular class on demand



Credit: Isaac Dykeman's Personal Blog: ijdykeman.github.io

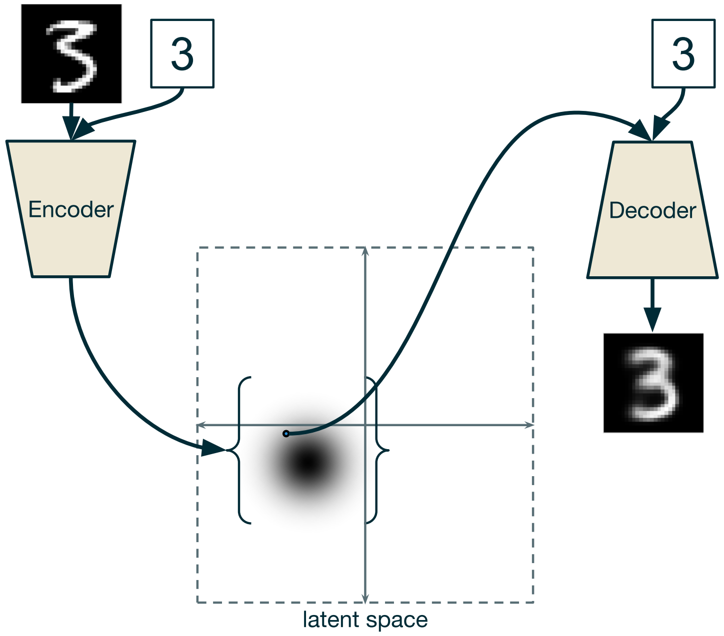
U-Net

- A type of convolutional neural network architecture → learn image to image mapping



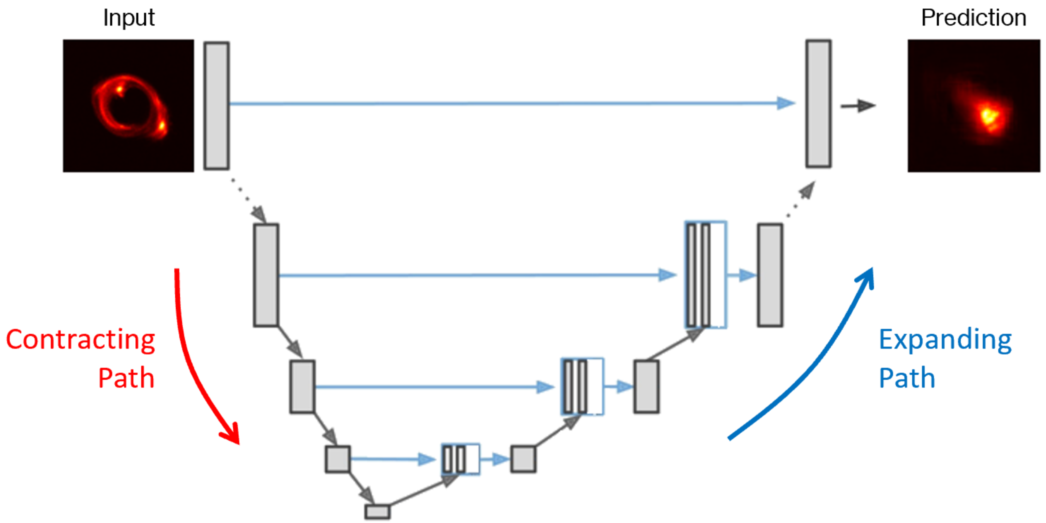
What We Have So Far

cVAE



a deep generative model to **generate new data** based on a noise vector and a set of conditional inputs

U-Net

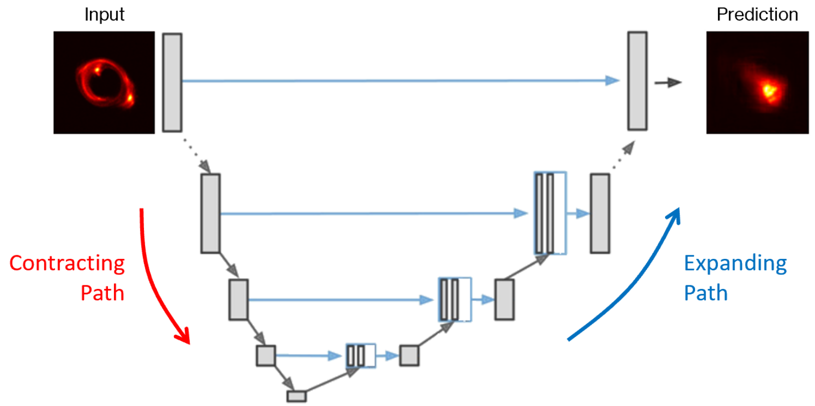


a convolutional neural network to learn **image to image mapping**

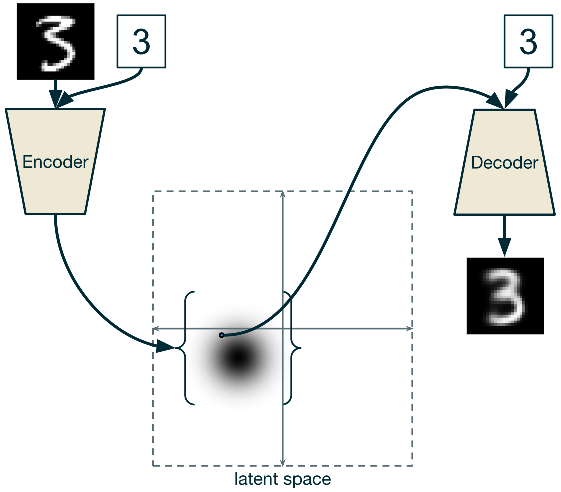
What We Need

- Problem Space...
  - high-dimensional observations and parameters
  - noisy observations
  - a **manifold of parameters** consistent with a given observation instead of a deterministic prediction (underconstrained problem)

- We Need...
  - high-dimensional inference
  - quantify uncertainties
  - model variability



+?

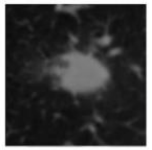


Applications

Model Output Variability

Training Set: 1 observation ↔ multiple predictions

Example: Different doctors assign different lesion areas on lung CT scans



CT Scan

$\boldsymbol{x}$

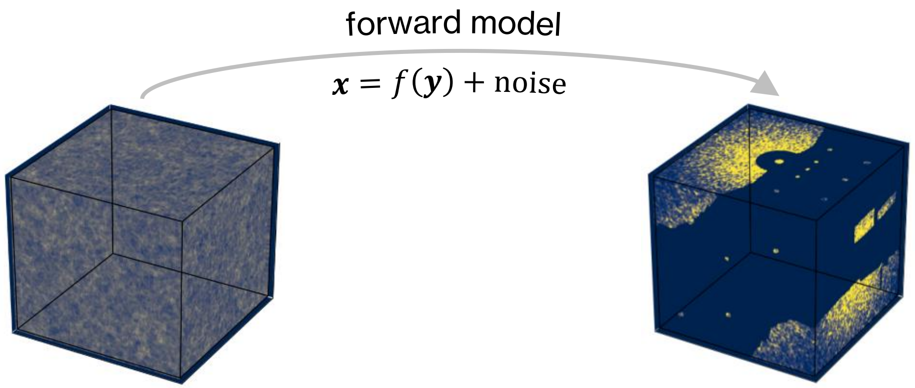
Segmentation Samples

$\boldsymbol{y}$

Inverse Problems

Training Set: multiple observations ↔ 1 prediction

Example: Reconstruct the initial conditions of the Universe



Initial Conditions

$\boldsymbol{y}$

Observed Galaxy Distribution

$\boldsymbol{x}$

Applications

Model Output Variability

Training Set: 1 observation ↔ multiple predictions

Inverse Problems

Training Set: multiple observations ↔ 1 prediction

in both cases, we are interested in modelling

lesion areas on lung CT scans

the Universe

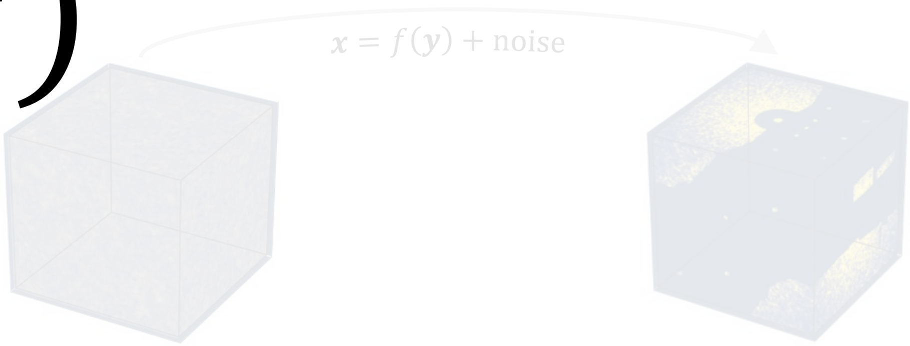
$$p(y|x)$$

CT Scan

x

Segmentation Samples

y



Initial Conditions

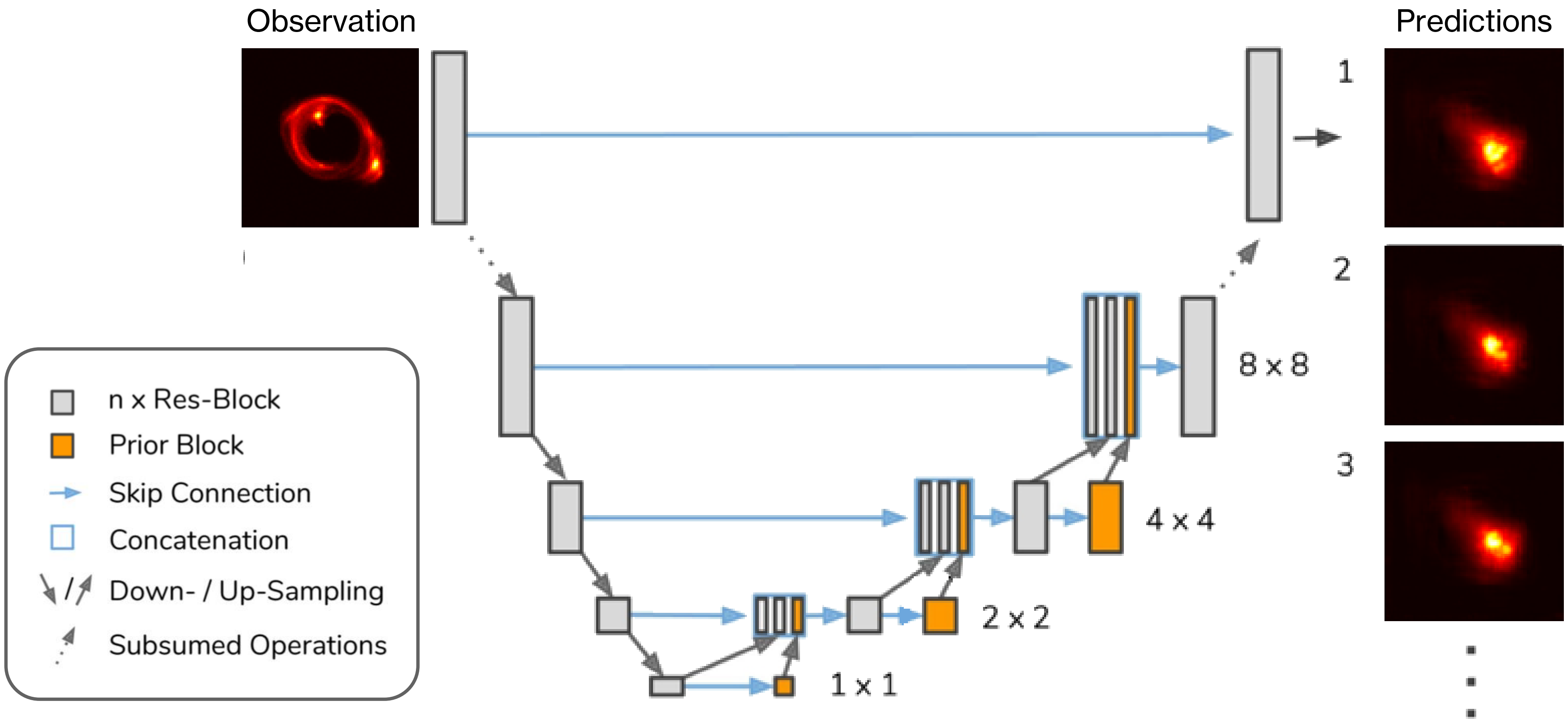
y

Observed Galaxy Distribution

x

Prob. U-Net

- Combination of cVAE & U-Net
- Latent spaces at several “scales” of the expanding path



Credit: Simon Kohl et al. (2019) – Figure taken with modification.

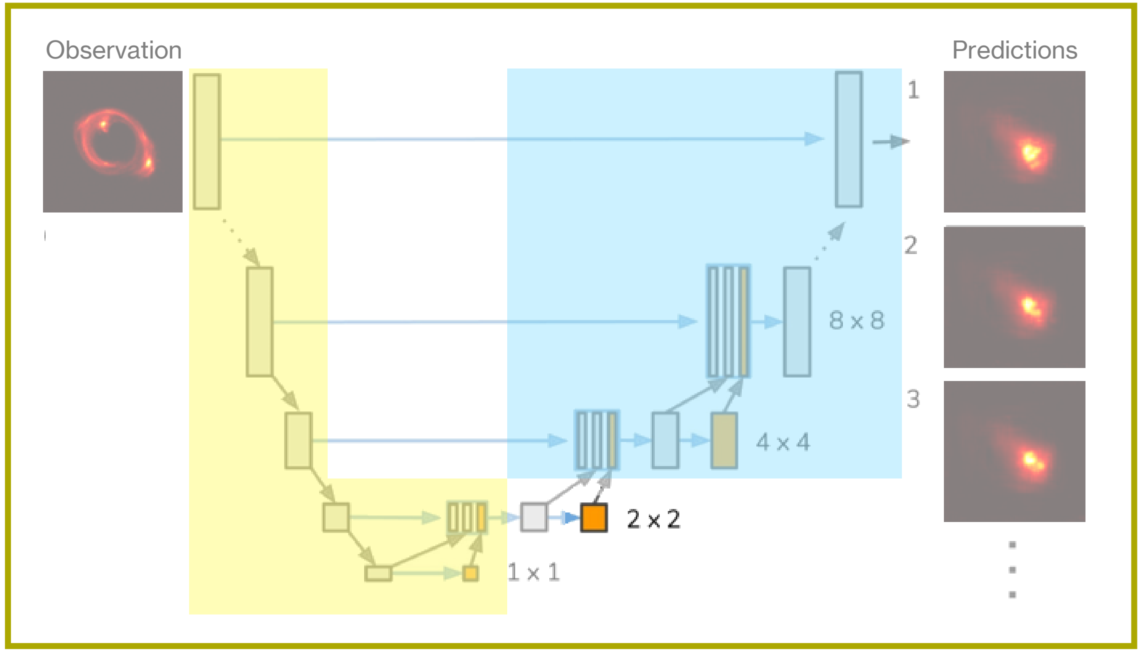
Prob. U-Net

- Prior "conditioned" on the observation and latents of previous scales

$$\boldsymbol{z}_i \sim p(\boldsymbol{z}_i | \boldsymbol{z}_{<i}, \boldsymbol{x})$$

- Joint prior decomposes into priors of each scale

$$p(\boldsymbol{z}_0, \dots, \boldsymbol{z}_L \mid \boldsymbol{x}) = p(\boldsymbol{z}_L | \boldsymbol{z}_{<L}, \boldsymbol{x}) \cdot \dots \cdot p(\boldsymbol{z}_0 | \boldsymbol{x})$$

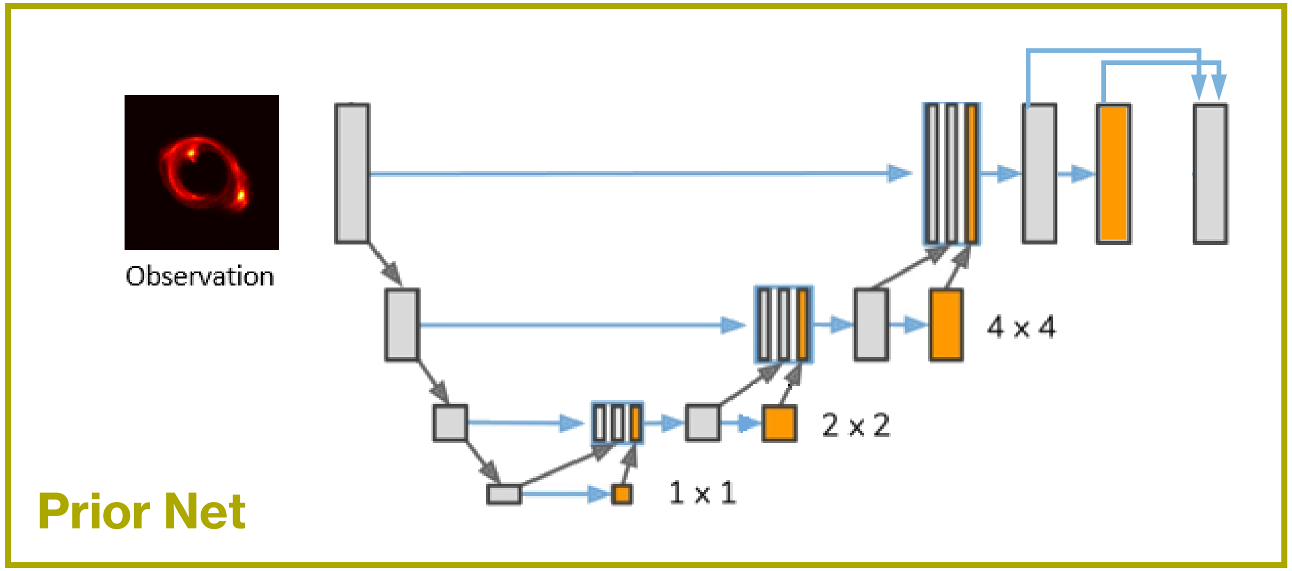


Prior Net

Question: Which part(s) resemble the VAE component that models

prior / likelihood / variational posterior?

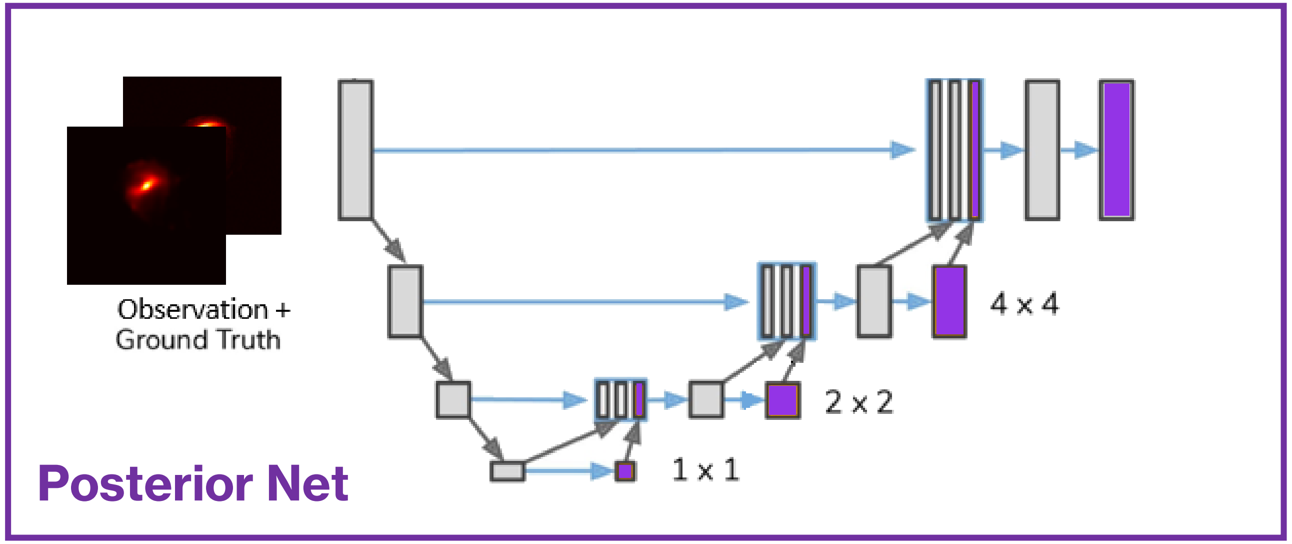
Prob. U-Net



Used in Training & Inference

$$\boldsymbol{z}_i \sim p(\boldsymbol{z}_i | \boldsymbol{z}_{<i}, \boldsymbol{x})$$

$$p(\boldsymbol{z}_0, \ldots, \boldsymbol{z}_L \mid \boldsymbol{x}) = p(\boldsymbol{z}_L \mid \boldsymbol{z}_{<L}, \boldsymbol{x}) \cdot \ldots \cdot p(\boldsymbol{z}_0 \mid \boldsymbol{x})$$

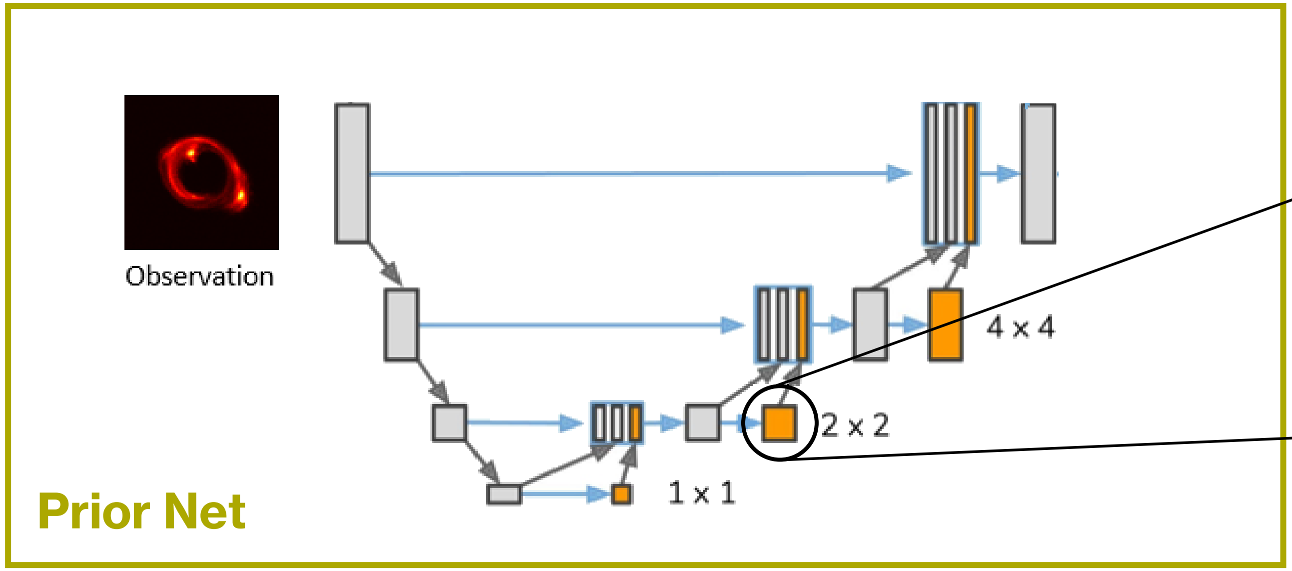


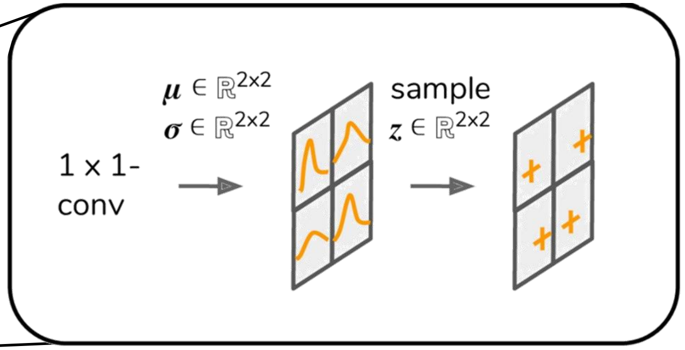
Used in Training

$$\boldsymbol{z}_i \sim q(\boldsymbol{z}_i | \boldsymbol{z}_{<i}, \boldsymbol{x}, \boldsymbol{y})$$

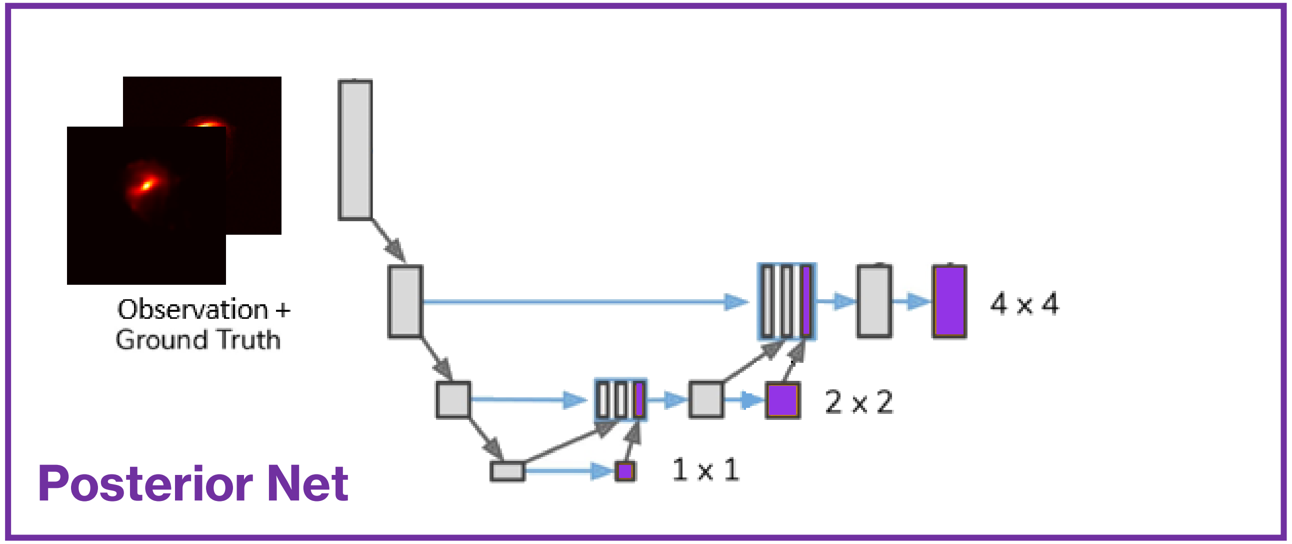
$$q(\boldsymbol{z}_0, \ldots, \boldsymbol{z}_L \mid \boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{z}_L \mid \boldsymbol{z}_{<L}, \boldsymbol{x}, \boldsymbol{y}) \cdot \ldots \cdot q(\boldsymbol{z}_0 \mid \boldsymbol{x}, \boldsymbol{y})$$

Prob. U-Net





Latents are pixelwise Gaussians



Posterior Net has a “truncated” decoder

Training

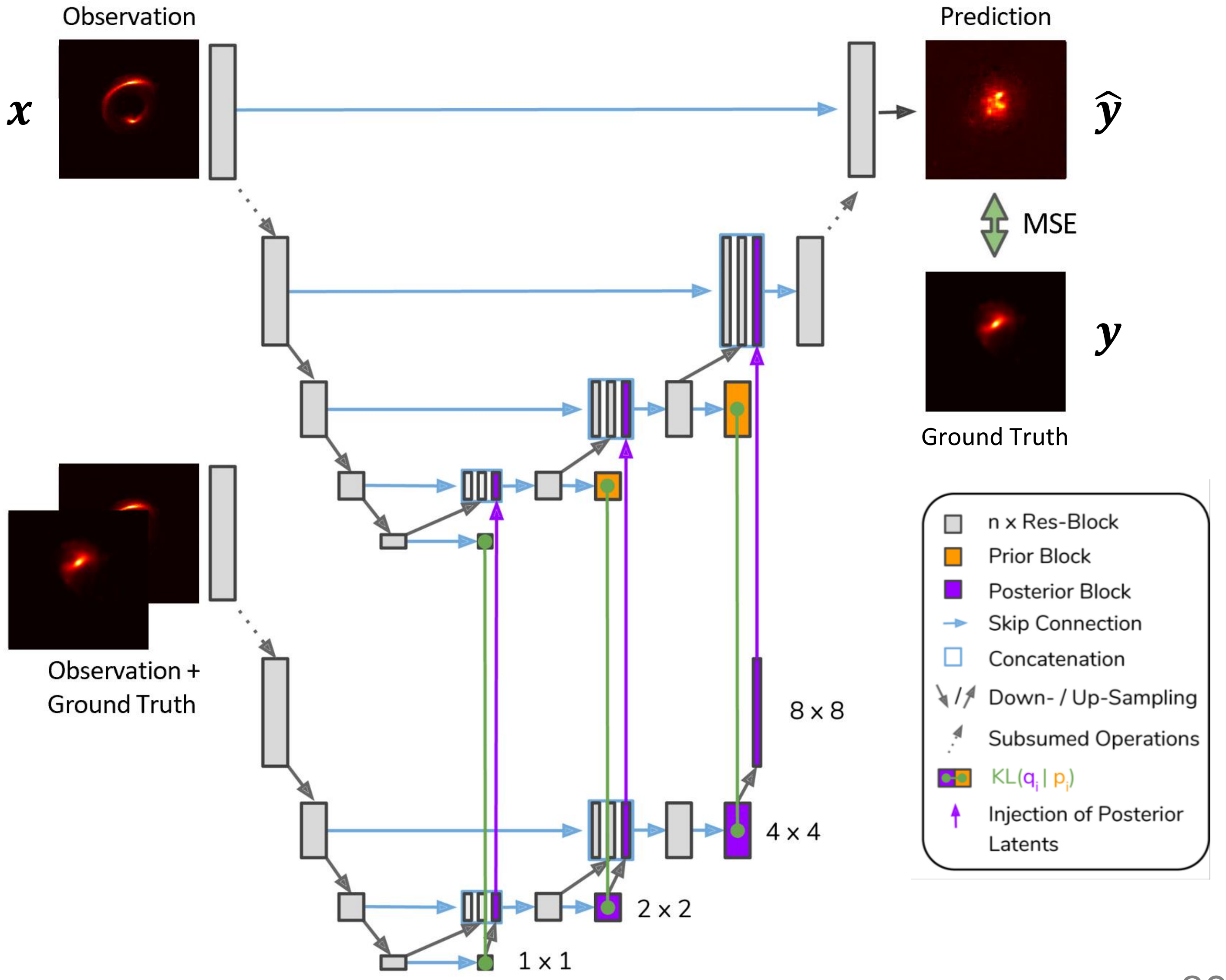
- Means and STDs predicted using both networks
  - Used to calculate KL

- Samples drawn from Posterior Net latents and inserted into the Prior Net

- **Objective:** Maximize evidence

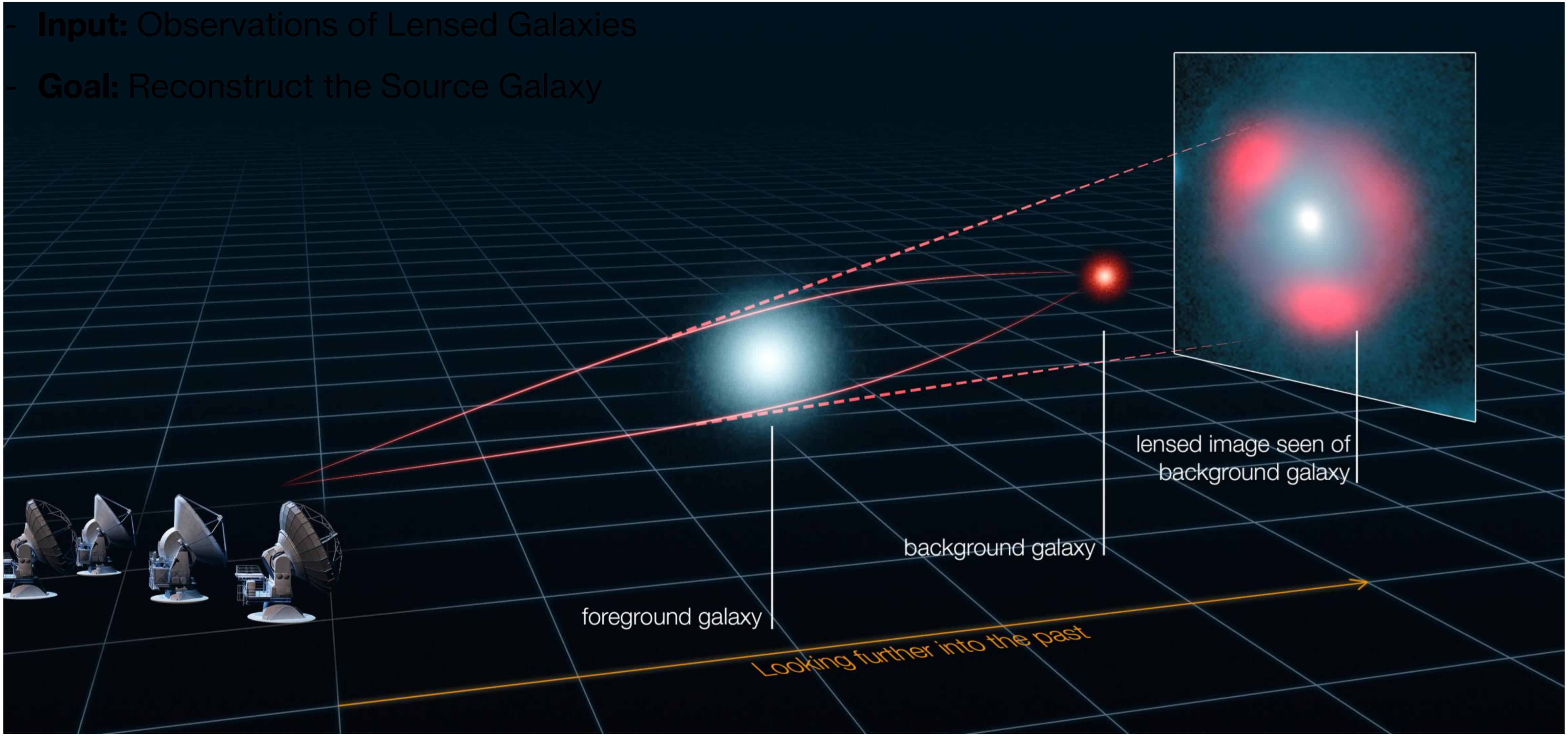
$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{z\sim Q}[-\ln p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z})] + \sum_{i=0}^{L} D_{\text{KL}}\big(q_i(\boldsymbol{z}_i \mid \boldsymbol{z}_{<i}, \boldsymbol{x}, \boldsymbol{y}) \parallel p_i(\boldsymbol{z}_i \mid \boldsymbol{z}_{<i}, \boldsymbol{x})\big)$$

$$\mathcal{L}_{\text{ELBO}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}$$



Toy Problem 1

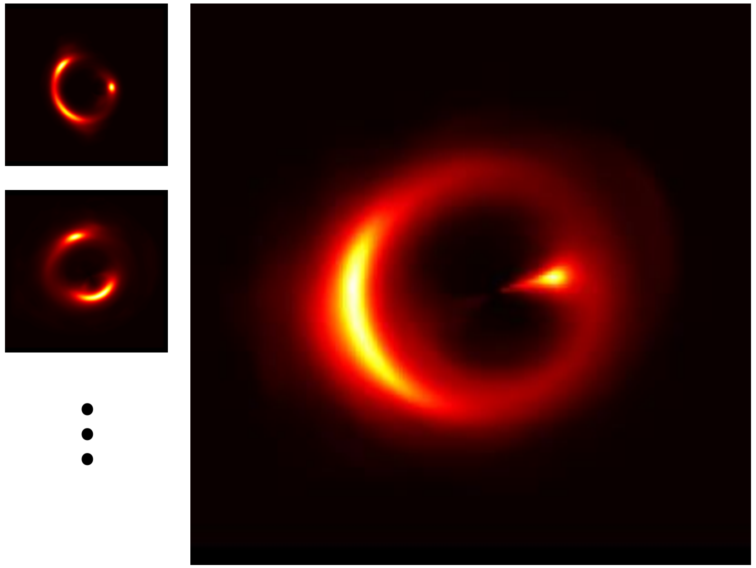
- **Input:** Observations of Lensed Galaxies
- **Goal:** Reconstruct the Source Galaxy



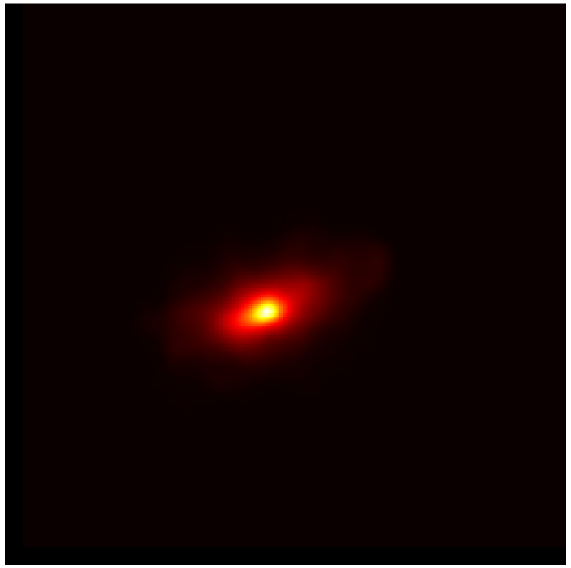
Credit: ALMA (ESO/NRAO/NAOJ), Luis Calçada (ESO), Yashar Hezaveh et al.

Toy Problem 1

Input: Observation of a ~~Lens~~-Source System



Goal: Find the Undistorted Image of the Source Galaxy



- Different ways to lens the source galaxy → Problem is underconstrained
- ***More Precise*** **Goal:** Draw samples from the posterior distribution of reconstructed source images

#source_reconstruction

#posterior_sampling

A Subtle Difference!

Variational Posterior

Defined in Latent Space

$$p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$$

we mean this when we say **posterior network**!

Parameters Posterior

Defined in Parameter Space

$$p(\boldsymbol{y}|\boldsymbol{x})$$

we mean this when we say **posterior sampling**!

$\boldsymbol{x}$ ∈ data space $\boldsymbol{y}$ ∈ parameter space $\boldsymbol{z}$ ∈ latent space

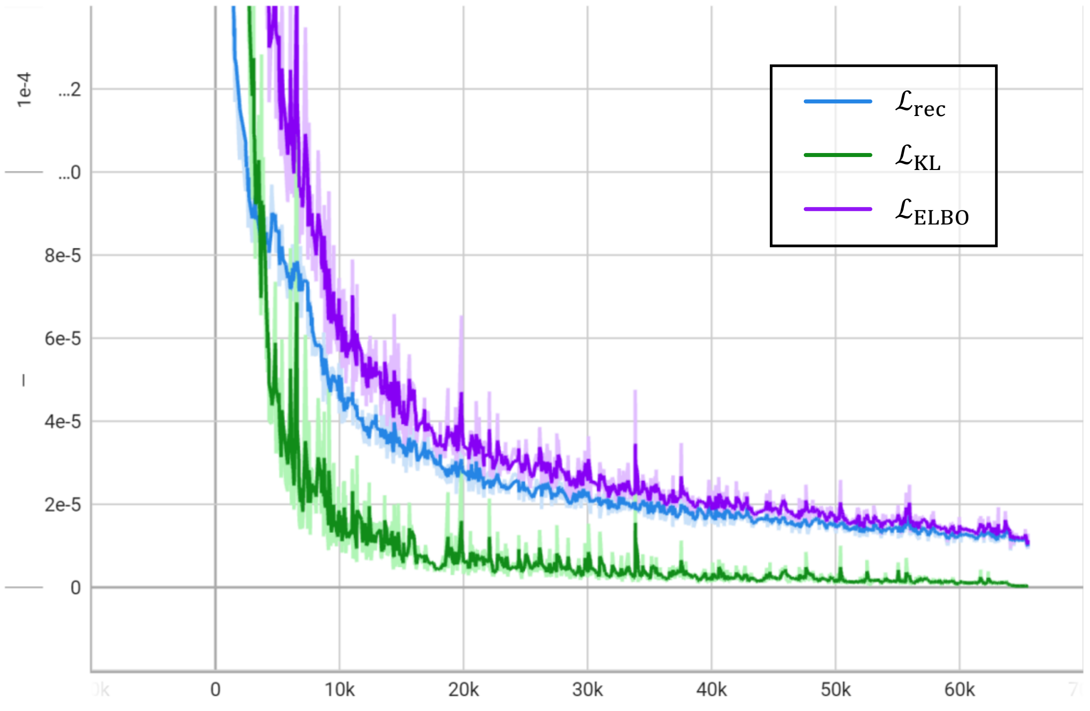
Part 2

Rescue the Randomness

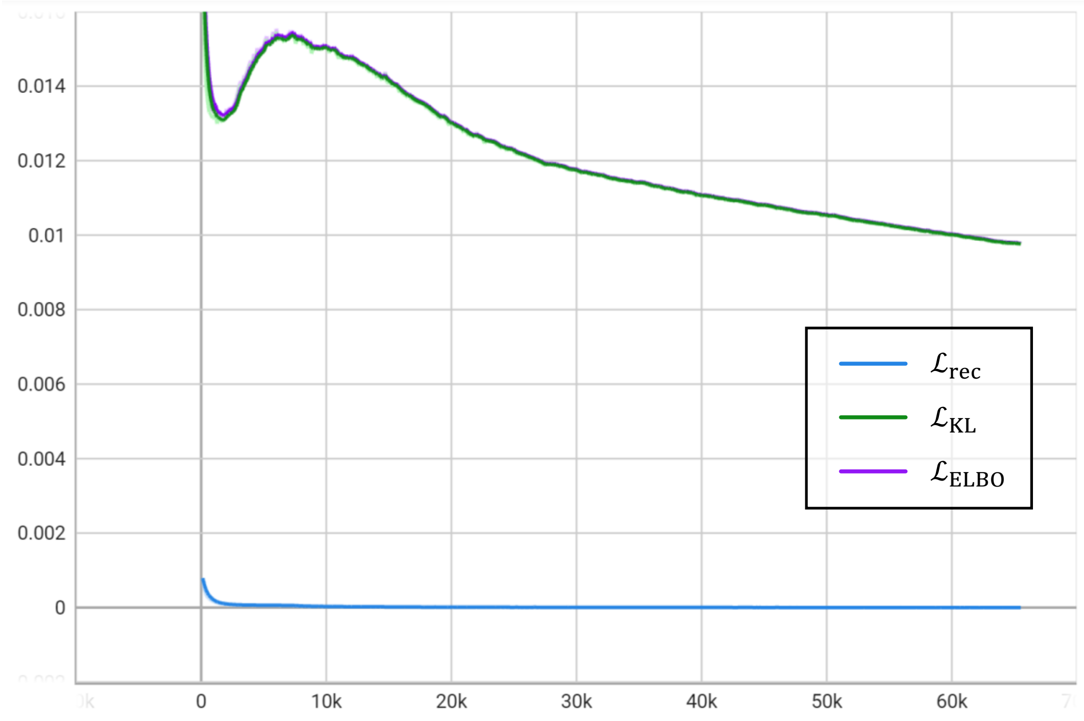
- KL Vanishing Problem
- ELBO Loss with β
- GECO Loss
- Toy Problem 2: One-hot Flipping

KL Vanishing Problem

Training



Validation



- KL Term vanishes early in the training due to non-informative latents
- Model ignores the cause of probabilistic behavior by setting respective weights to 0 → Deterministic

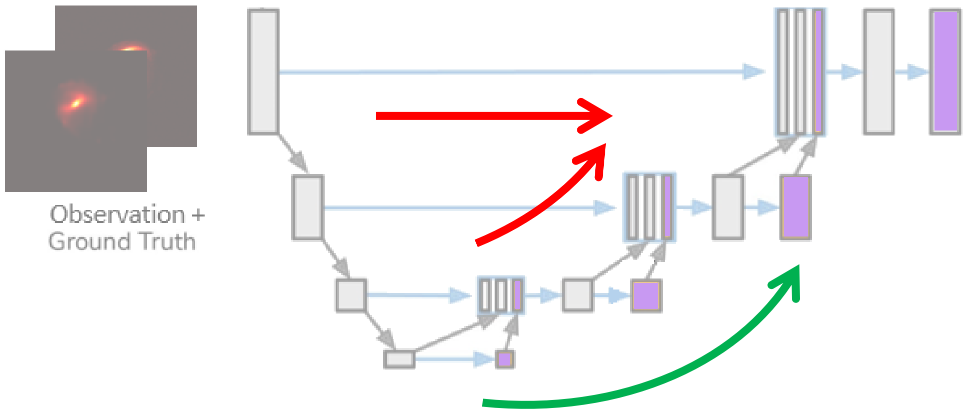
KL Vanishing Problem

- KL Term vanishes early in the training due to non-informative latents
- Model ignores the cause of probabilistic behavior by setting respective weights to 0 → Deterministic

- Happens when two types of paths exist:

  A. Latent Path: Conditioned on the latent space (same as VAEs)

  B. Leaky Path: Does not pass through latent spaces;
  Leaks the ground truth information



ELBO with β

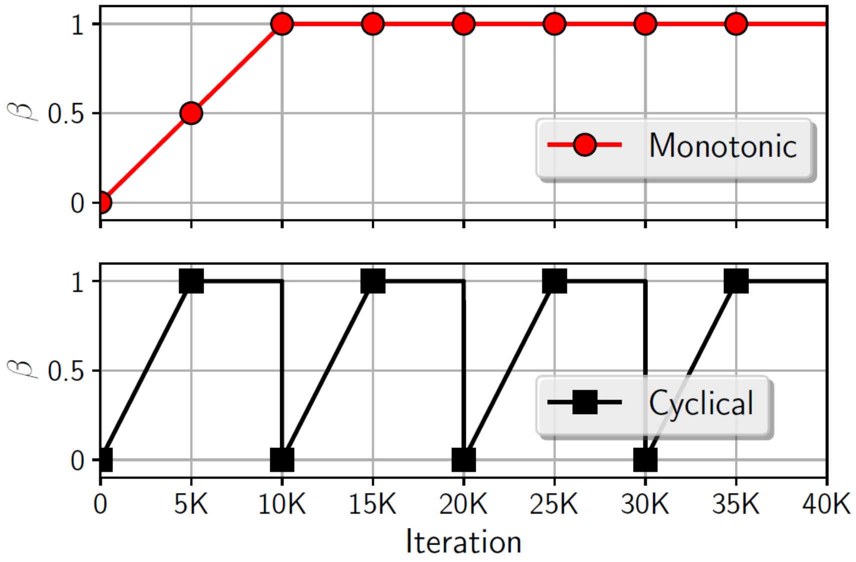
- **Idea:** Prevent the optimization scheme from caring too much about the KL term before having meaningful latents.

- Possible Approaches:
  - Set $0 < \beta < 1$
  - Start with $\beta = 0$ and Gradually Increase it (**Beta Annealing**)
  - Other ways of scheduling β (e.g., **Cyclical Schedule**)

- What is the best way to schedule β?
  - Variety of choices
  - Depends on the specific problem

$$\mathcal{L}_{\text{ELBO}} = \mathcal{L}_{\text{rec}} + \beta\, \mathcal{L}_{\text{KL}}$$

governs the amount of regularization



Credit: D. Rezende, F. Viola (2018)

GECO

- **G**eneralized **E**LBO with **C**onstrained **O**ptimization
- Constrained Optimization Framework
  - Minimize the KL Term under a set of reconstruction constraints
- λ plays the role of β → automatically updated during training
  - (Usually) tend to focus on the reconstruction loss early in the training until it reaches κ;
  - Then moves the pressure over on the KL Term.
- Advantages:
  - Hyperparameter (κ) defined in data space → More intuitive
  - β is updated automatically

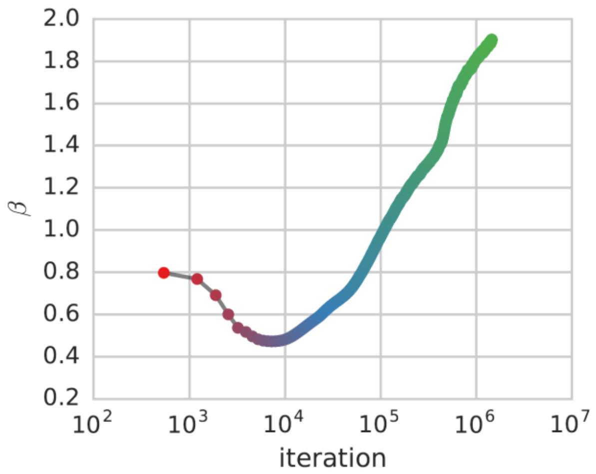
$$\mathcal{L}_{\text{ELBO}} = \mathcal{L}_{\text{rec}} + \beta \, \mathcal{L}_{\text{KL}}$$

$$\mathcal{L}_{\text{GECO}} = \lambda \, (\mathcal{L}_{\text{rec}} - \kappa) + \mathcal{L}_{\text{KL}}$$

Lagrange multiplier

reconstruction threshold

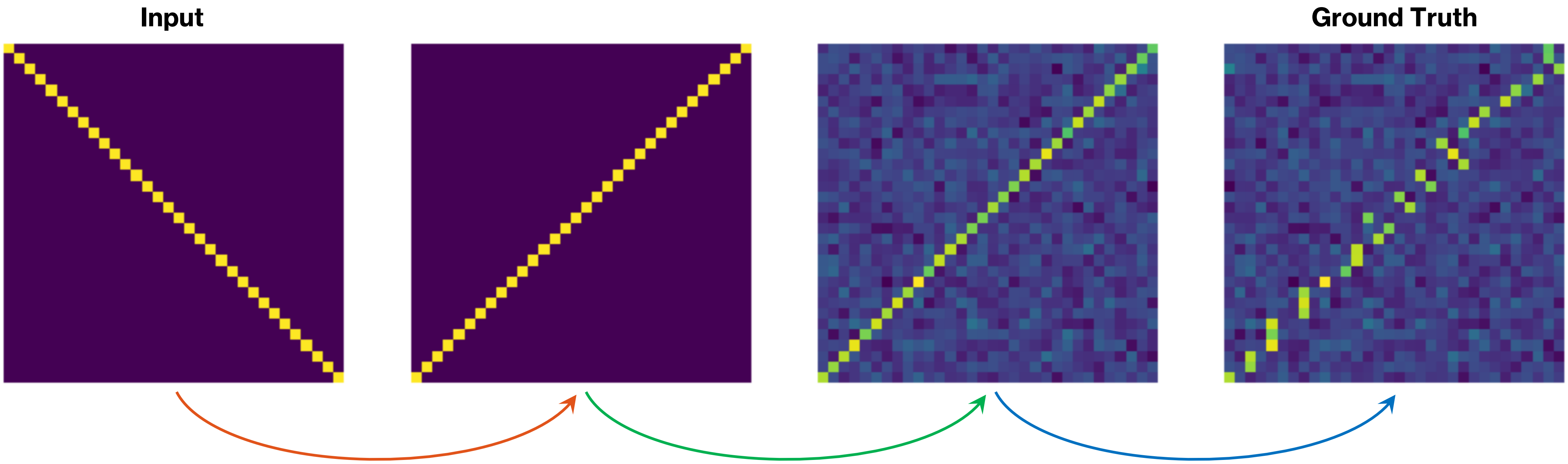
$$\lambda \equiv \frac{1}{\beta}$$



Credit: H. Fu, C. Li et al. (2019)

Toy Problem 2

- Training Set: all 32-bit **one-hot** vectors



Flip All Sequences

Global Uncertainty

Add Noise ~ $\mathcal{N}(0, \sigma)$

Local Uncertainty

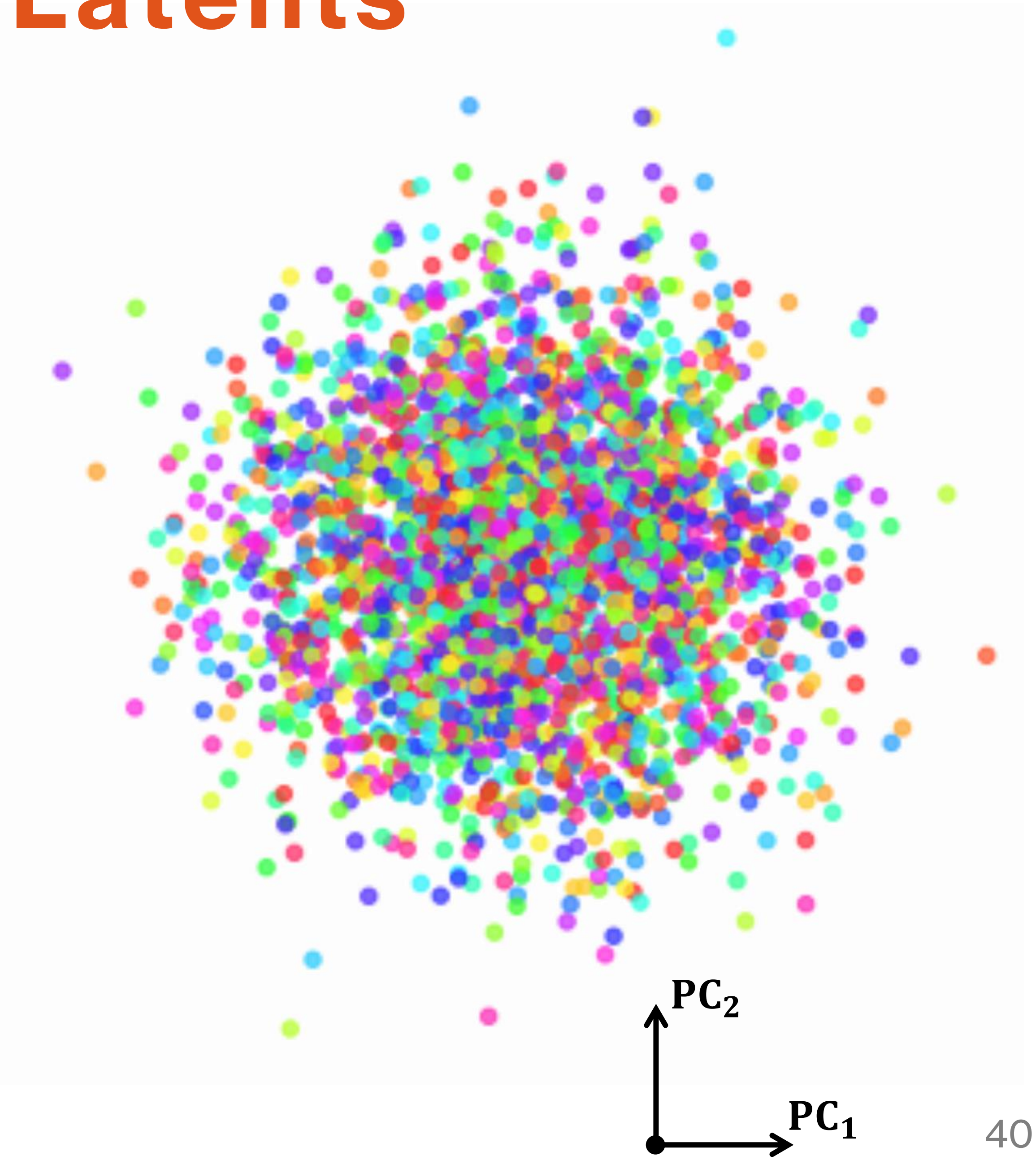
Roll Each Sequence by r Pixels

$$p(r = 0) = 0.4$$
$$p(+1) = p(-1) = 0.2$$
$$p(+2) = p(-2) = 0.1$$

- New realizations generated at each training step

Visualizing Latents

- Assign a unique color to each input
- Sample a bunch of latent representations for each input
  - For an arbitrary scale of the Prob. U-Net
  - Can have many dimensions
- Plot the first two **principal components** (orthogonal directions with most variability)




“Advertisement”
Inspecting Uncertainties
Coverage Probability Test

Don’t Get Too Excited!

- Having a model with probabilistic behavior is not enough!
- Require comprehensive statistical analysis that goes beyond the model's assumptions
- To make sure uncertainties are appropriately quantified

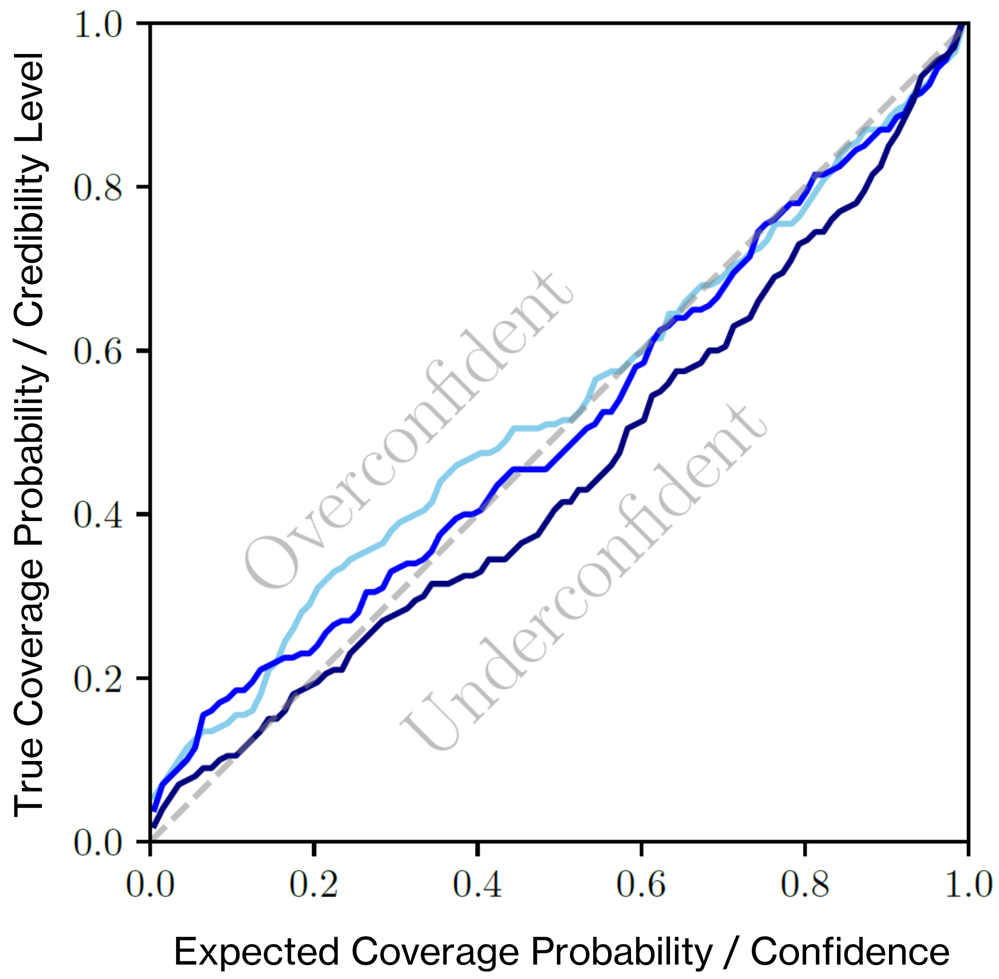
DALL· E ' s impression of

A Robot Thinking About Statistics

Coverage Probability Test

High-level explanation:

1. Repeatedly sample from the model
2. Calculate a confidence interval using samples (expected coverage)
3. Check if the true value falls within the interval
4. Repeat steps 1-3 for multiple “samples - true value” combinations
5. Calculate the fraction of times that the true value falls within each confidence interval (true coverage)
6. Plot true coverage vs. expected coverage



Credit: Pablo Lemos et al. (2022)

Advertisement

Sampling-Based Accuracy Testing of Posterior Estimators for General Inference

Pablo Lemos [1 2 3 4 *] Adam Coogan [1 2 3 *] Yashar Hezaveh [1 2 3] Laurence Perreault-Levasseur [1 2 3]

arXiv: 2302.03026

- Method to estimate coverage probabilities of generative posterior estimators without posterior evaluations (by just using samples).
- Necessary and sufficient to show that a posterior estimator is optimal.
- `pip`-installable package on the way!

Thank You For Your Attention! 💐

Contact: hadi.sotoudeh@umontreal.ca

Collaborators: Laurence Perreault-Levasseur, Pablo Lemos, Ève Campeau-Poirier, Charles Wilson, Alexandre Adam

This work was made possible through the generous support of:

Université de Montréal, Mila, Ciela Institute, Simons Foundation, Digital Research Alliance of Canada, IVADO

To Read More...

- **Prob. U-Net:** Kohl, S. A. A., “A Probabilistic U-Net for Segmentation of Ambiguous Images”, arXiv: 1806.05034 🔗

- **Hierarchical Prob. U-Net:** Kohl, S. A. A., “A Hierarchical Probabilistic U-Net for Modeling Multi-Scale Ambiguities”, arXiv: 1905.13077 🔗

- **KL Vanishing and Cyclical β:** Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L., “Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing”, arXiv: 1903.10145 🔗

- **GECO:** Jimenez Rezende, D. and Viola, F., “Taming VAEs”, arXiv: 1810.00597 🔗

- **Coverage Test:** Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L., “Sampling-Based Accuracy Testing of Posterior Estimators for General Inference”, arXiv: 2302.03026 🔗

- **VAEs:** Rocca, J., Blog Post on “Understanding Variational Autoencoders (VAEs)”, 🔗

- **Conditional VAEs:** Dykeman, I., Blog Post on “Conditional Variational Autoencoders”, 🔗