Posterior Collapse and Latent Variable Non-identifiability

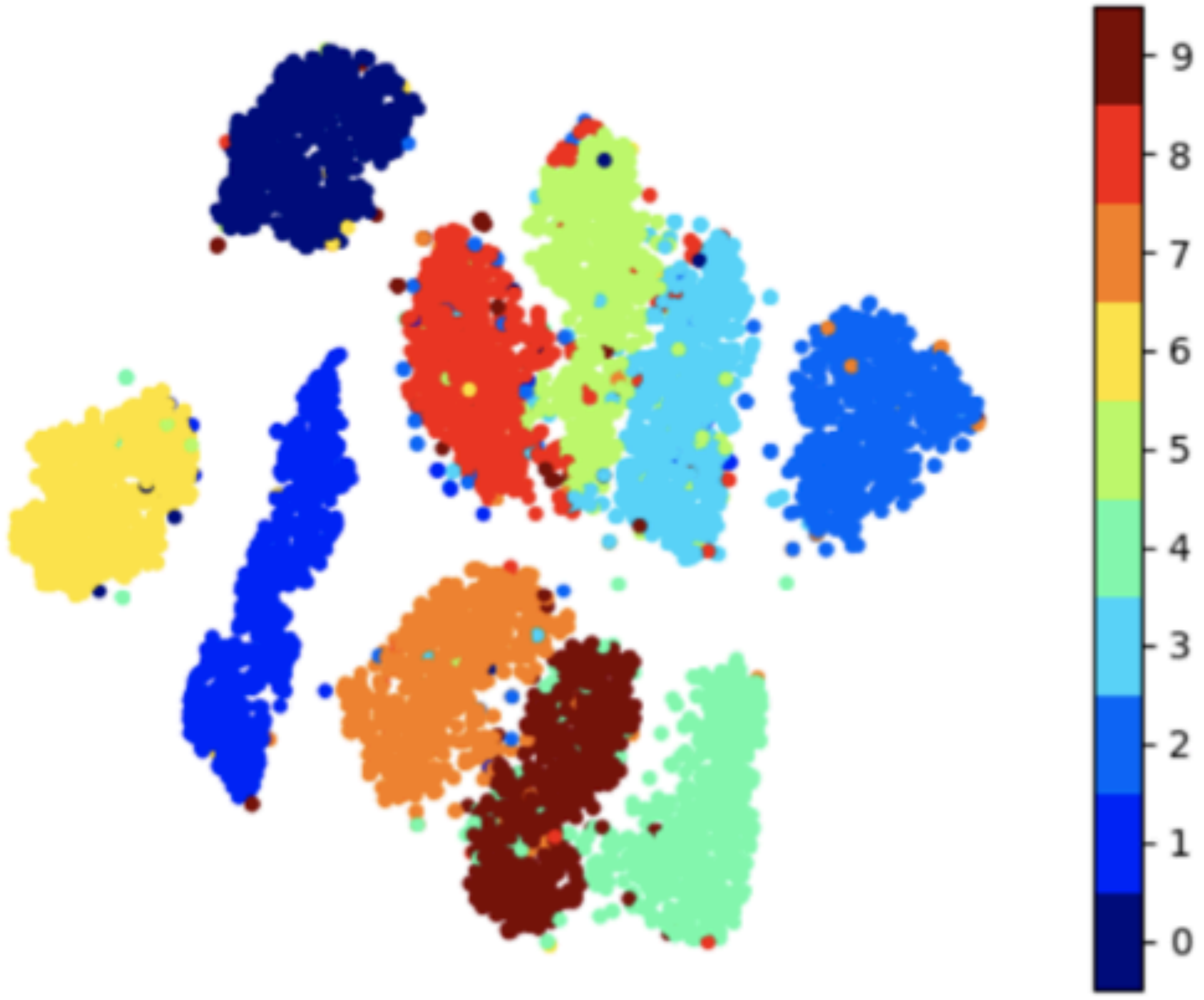
Yixin Wang, David Blei, John Cunningham

Modeling high-dimensional data with VAE

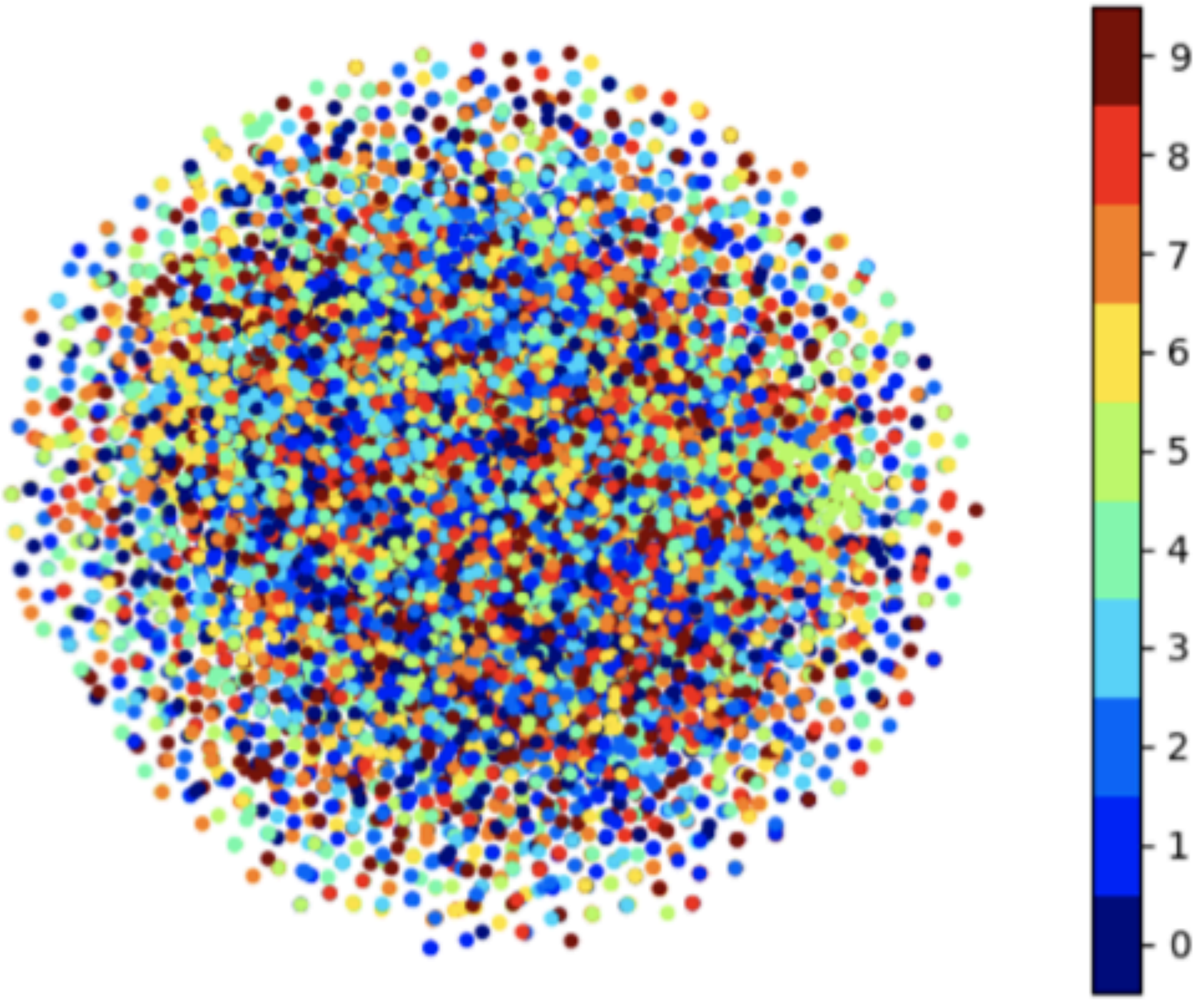
- Consider a dataset $\mathbf{x} = (x_1, \ldots, x_n)$; each datapoint m-dimensional.
- Positing n latent variables $\mathbf{z} = (z_1, \ldots, z_n)$; each latent K-dimensional
- A variational autoencoder (VAE) assumes each datapoint x_i is generated by the latent variable z_i,

$$z_i \sim p(z_i), \qquad x_i \mid z_i \sim p(x_i \mid z_i\,;\,\theta) = \mathrm{EF}(x_i \mid f_\theta(z_i))\,.$$

Posterior Collapse



Ideal

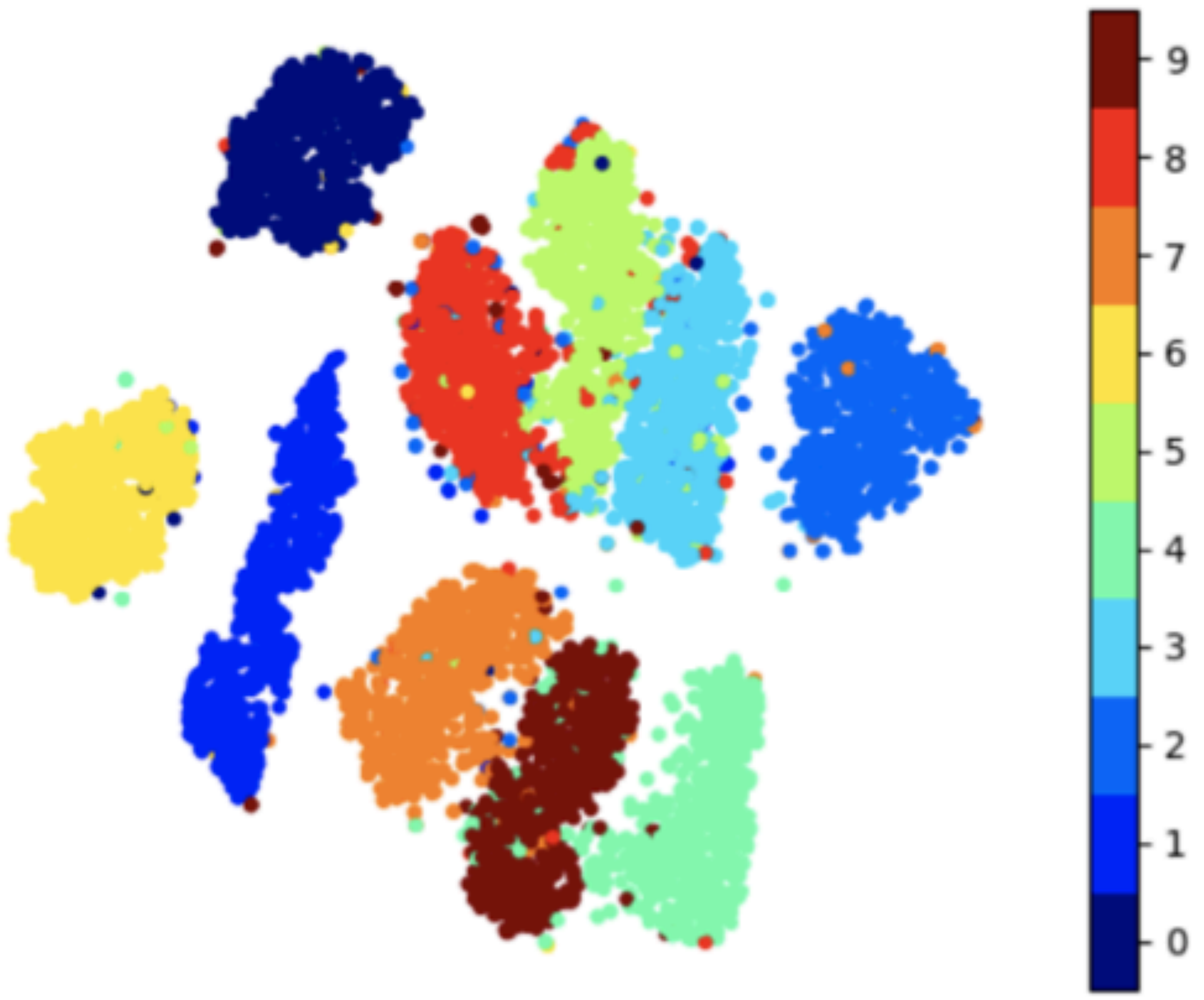


Reality

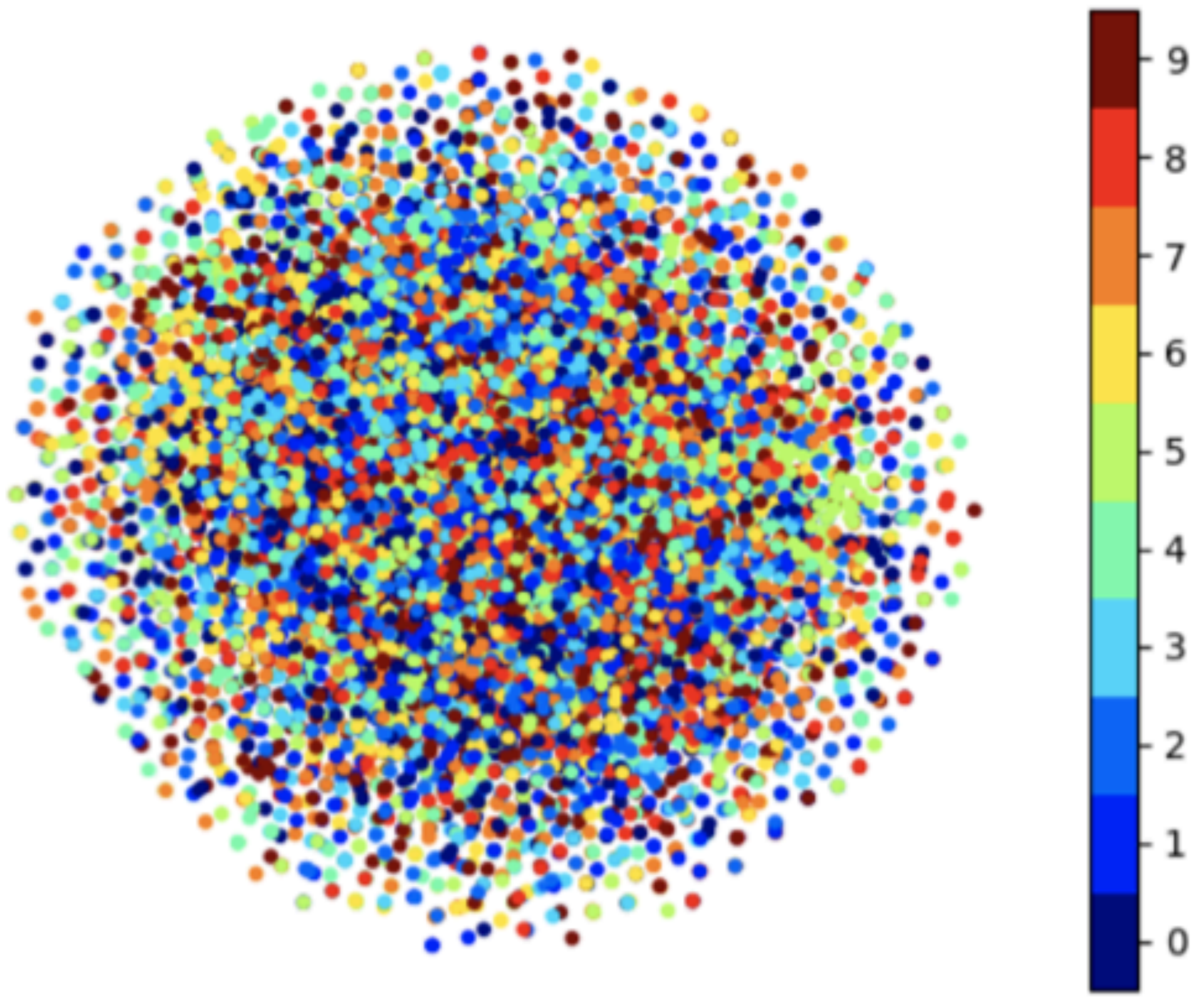
Jha et al. (CVPR, 2018)

- The model fits: Predictive likelihood high; Generate good new samples.
- Posterior is equal to the prior: Non-informative / useless as representations.

Posterior Collapse



Ideal



Reality

Jha et al. (CVPR, 2018)

- **Posterior collapse** is a phenomenon where the posterior of the latents in a VAE is equal to its uninformative prior

$$p(\boldsymbol{z} \,|\, \boldsymbol{x}\,;\, \theta^*) = p(\boldsymbol{z}).$$

This work

- Posterior collapse phenomenon is a problem of latent variable non-identifiability.
- It is not specific to the use of neural networks or particular inference algorithms in VAE. Rather, it is an intrinsic issue of the model and the dataset.
- We propose a class of IDVAE via Brenier maps to resolve latent variable non-identifiability and mitigate posterior collapse.

Posterior Collapse: Abstract away approximate inference

- We consider the ideal case where the variational approximation is exact.
- If the exact posterior suffers from posterior collapse, then so will the approximate posterior.
- A variational approximation cannot ``uncollapse'' a collapsed posterior.

Latent Variable Non-identifiability

- **Definition (Latent variable non-identifiability)**
  - Given a likelihood function $p(\mathbf{x}, \mathbf{z}; \theta)$, a parameter value $\theta = \hat{\theta}$, and a dataset $\mathbf{x} = (x_1, \ldots, x_n)$, the latent variable $\mathbf{z}$ is non-identifiable if

$$p(\mathbf{x} \,|\, \mathbf{z} = \tilde{\mathbf{z}}'; \hat{\theta}) = p(\mathbf{x} \,|\, \mathbf{z} = \tilde{\mathbf{z}}; \hat{\theta}) \qquad \forall \tilde{\mathbf{z}}', \tilde{\mathbf{z}} \in \mathcal{Z} \ .$$

Posterior Collapse iff Latent Variable Non-identifiability

- Theorem (Latent variable non-identifiability ⇔ Posterior collapse)
  - Consider a probability model $p(\mathbf{x}, \mathbf{z}; \theta)$, a dataset $\mathbf{x}$, and a parameter value $\theta = \hat{\theta}$. The latent variables $\mathbf{z}$ are non-identifiable at $\hat{\theta}$ if and only if the posterior of the latent variable $\mathbf{z}$ collapses, $p(\mathbf{z} \mid \mathbf{x}) = p(\mathbf{z})$.
- One line proof due to the Bayes rule
  - $p(\mathbf{z} \mid \mathbf{x}; \hat{\theta}) \propto p(\mathbf{z})p(\mathbf{x} \mid \mathbf{z}; \hat{\theta}) = p(\mathbf{z})p(\mathbf{x}; \hat{\theta}) \propto p(\mathbf{z})$

Posterior Collapse iff Latent Variable Non-identifiability

- It happens with exact inference.
- It happens in classical not-so-flexible models.
- It doesn't have to involve neural network.
- It happens with global optima.
- It happens with both local and global latent variables.

Posterior Collapse: Can we fix it?

- Make latent variables **identifiable** in VAE.
- Constructing identifiable VAE thus amounts to constructing an **injective likelihood function** for VAE.
- The construction is based on Brenier map / monotone transport map, which preserves flexibility but guarantees latent variable identifiability.

.