Interpretable and accessible Deep Learning for single-cell data analysis

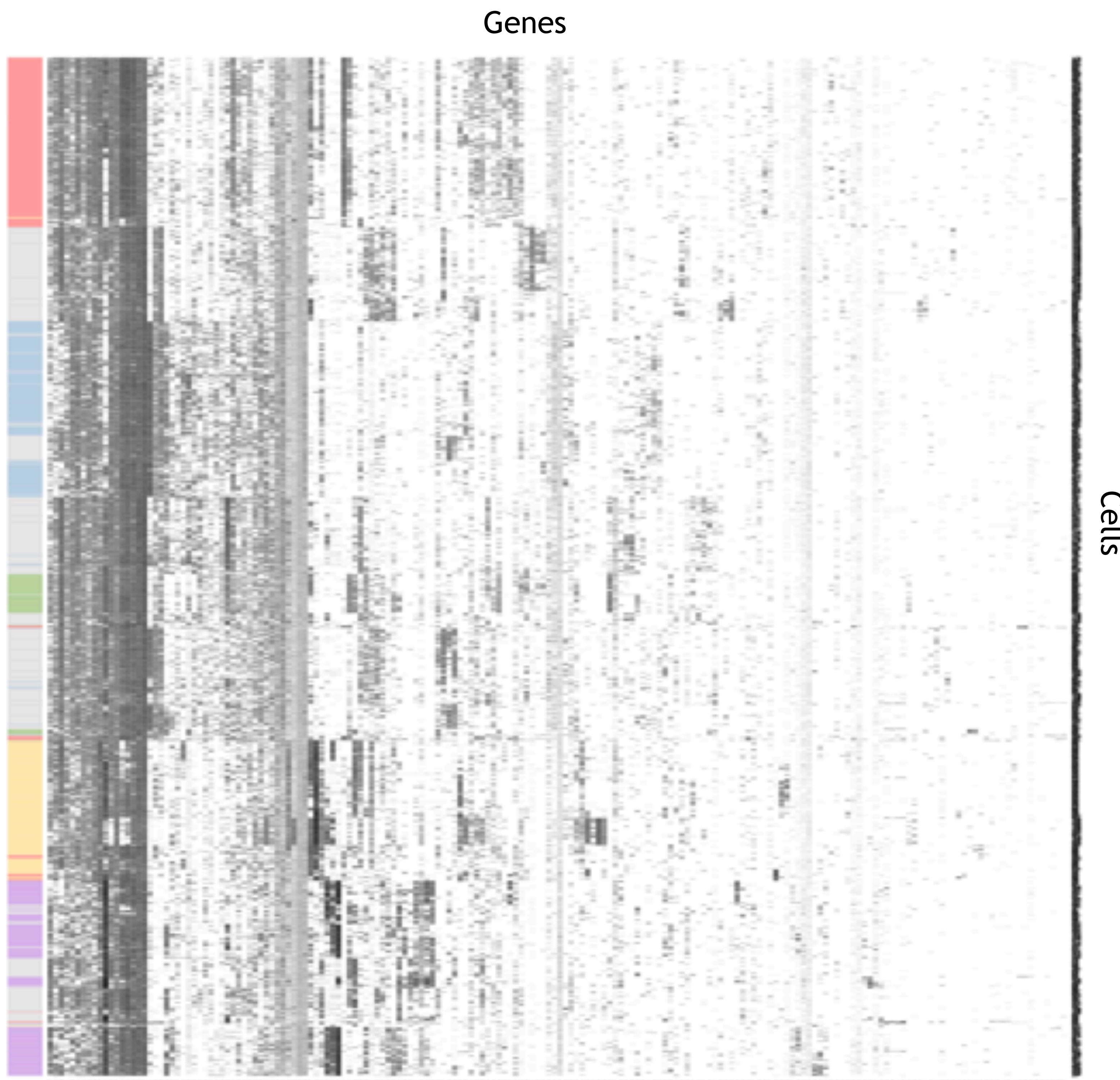
Sara Al-Rawi, Harald Binder, Maren Hackenberg, Moritz Hess, Martin Treppner

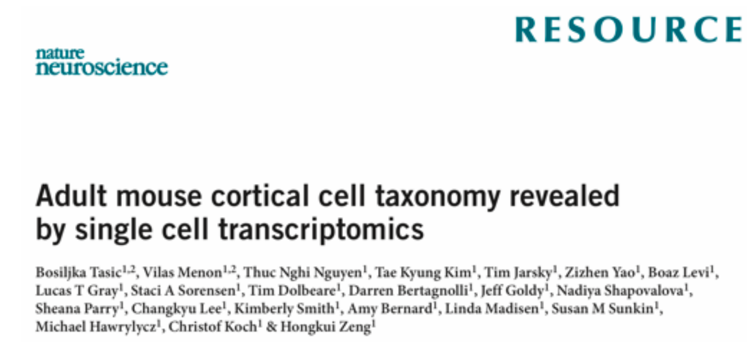
Institute of Medical Biometry and Statistics (IMBI)

SPONSORED BY THE

Federal Ministry of Education and Research

Gene expression - investigated in single cells



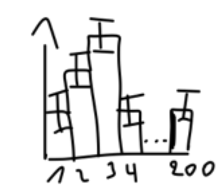

nature neuroscience

RESOURCE

Adult mouse cortical cell taxonomy revealed by single cell transcriptomics

Bosiljka Tasic[1,2], Vilas Menon[1,2], Thuc Nghi Nguyen[1], Tae Kyung Kim[1], Tim Jarsky[1], Zizhen Yao[1], Boaz Levi[1], Lucas T Gray[1], Staci A Sorensen[1], Tim Dolbeare[1], Darren Bertagnolli[1], Jeff Goldy[1], Nadiya Shapovalova[1], Sheana Parry[1], Changkyu Lee[1], Kimberly Smith[1], Amy Bernard[1], Linda Madisen[1], Susan M Sunkin[1], Michael Hawrylycz[1], Christof Koch[1] & Hongkui Zeng[1]

1525 cells



100 - 200 “interesting” genes

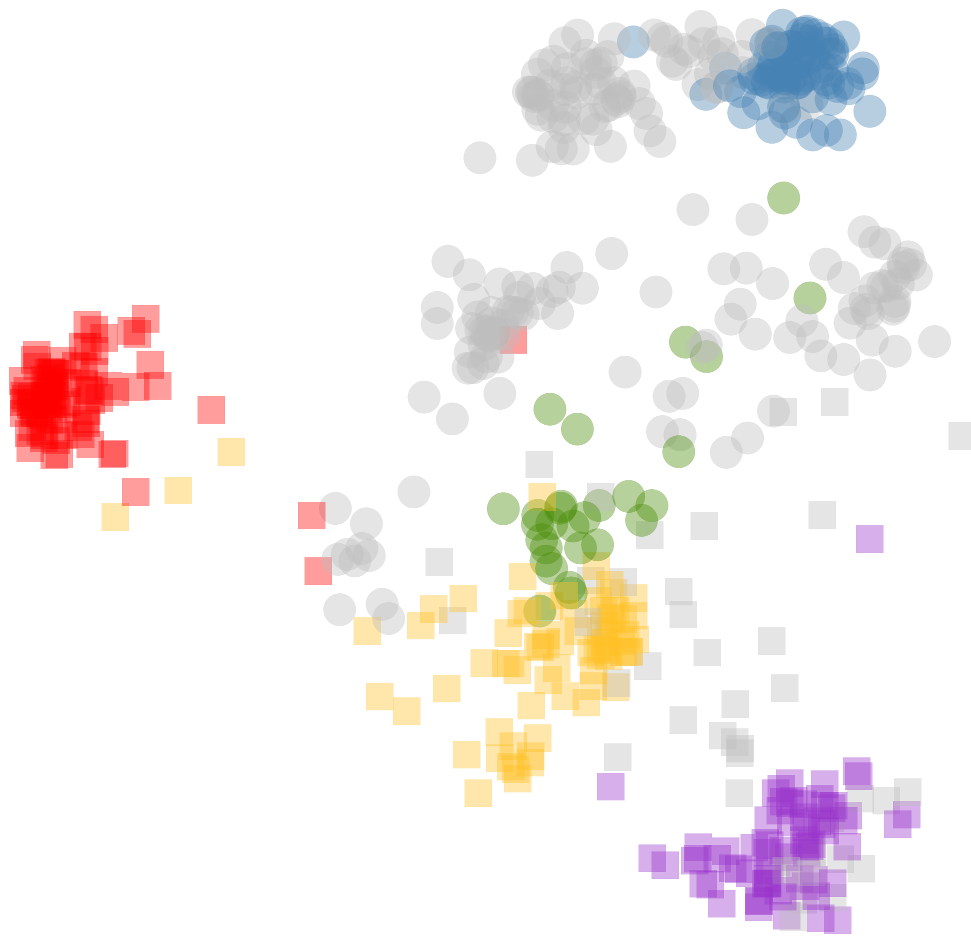
Part I: **Learning latent representations**

Part I: **Learning latent representations**

Principal components (PCs)
computed from raw data

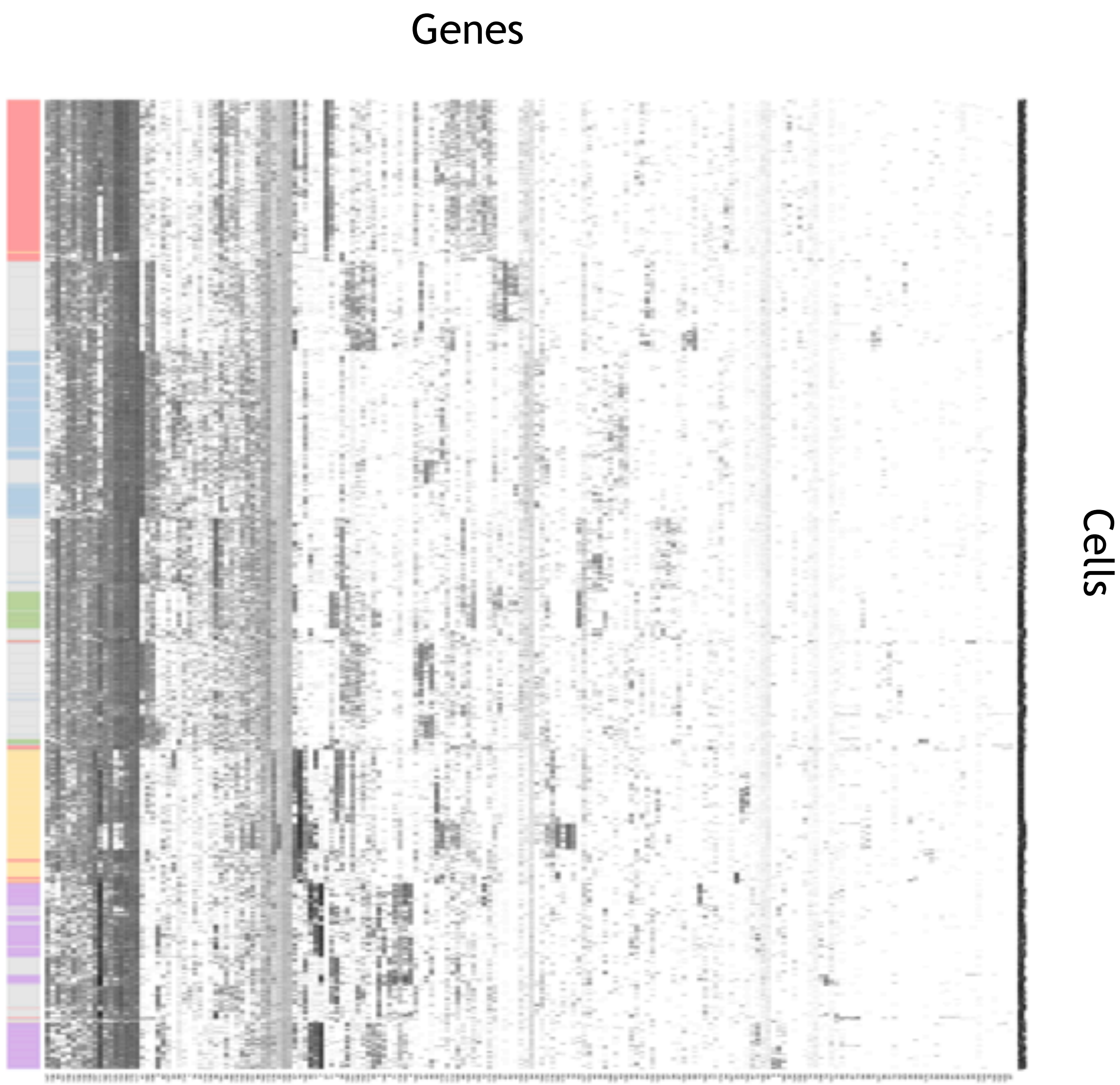
Part I: **Learning latent representations**

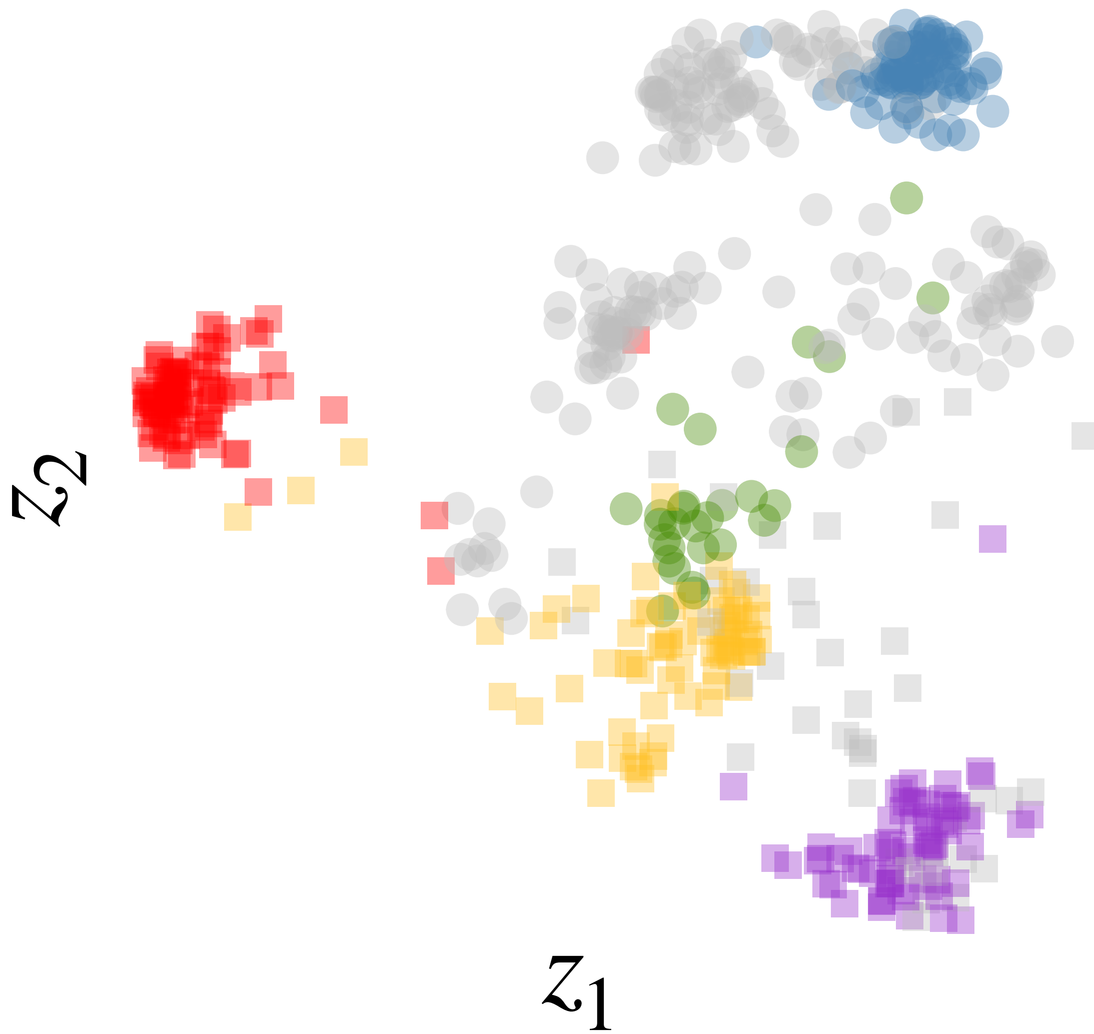
Principal components (PCs) computed from raw data



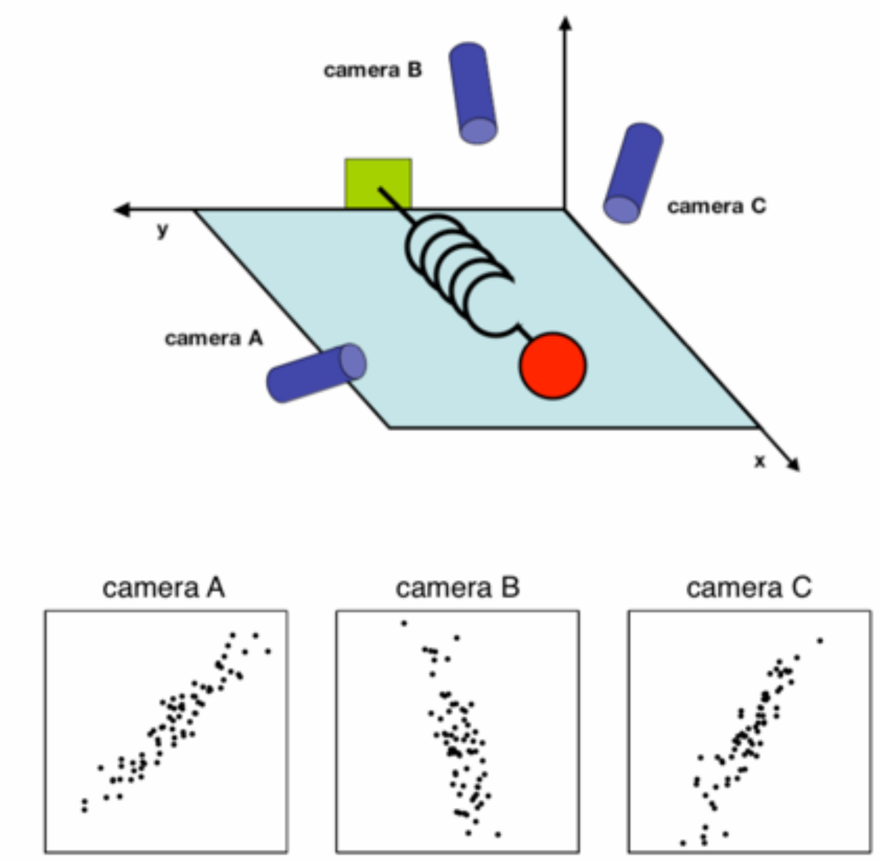
Latent representations, learned by a deep Boltzmann machine

Part II: Understanding the relationship between latent representations and the observed variables (genes).



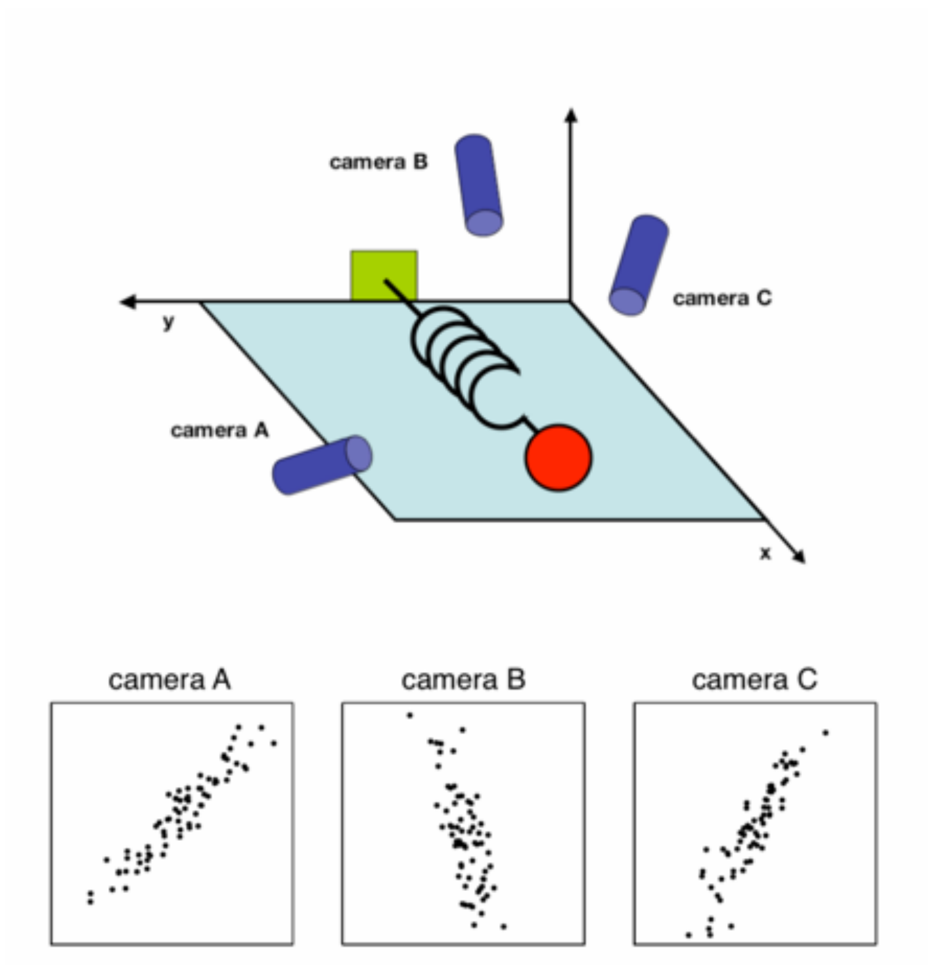


Principal component analysis (PCA)

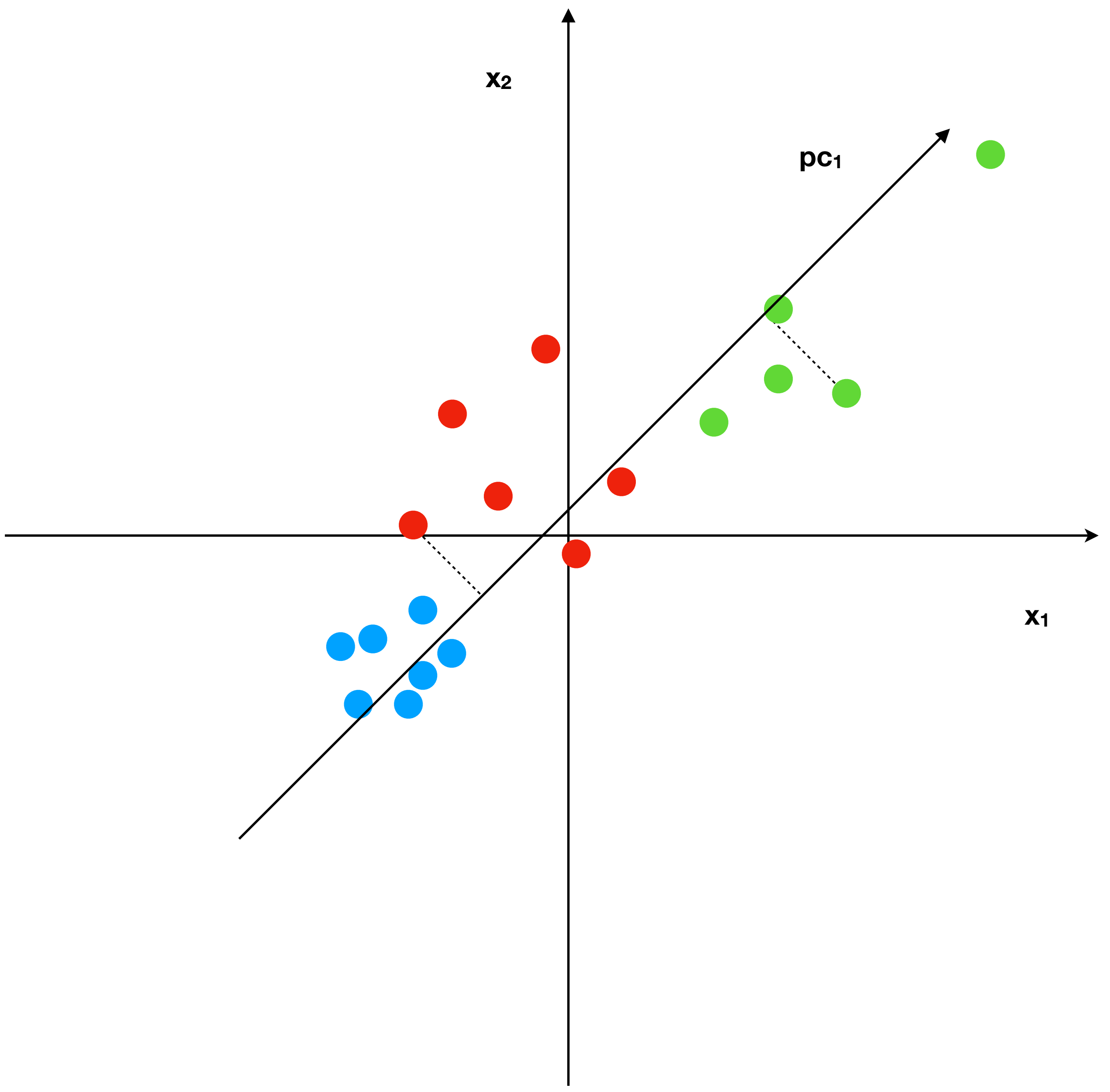


Shlens et al. 2014

Principal component analysis (PCA)

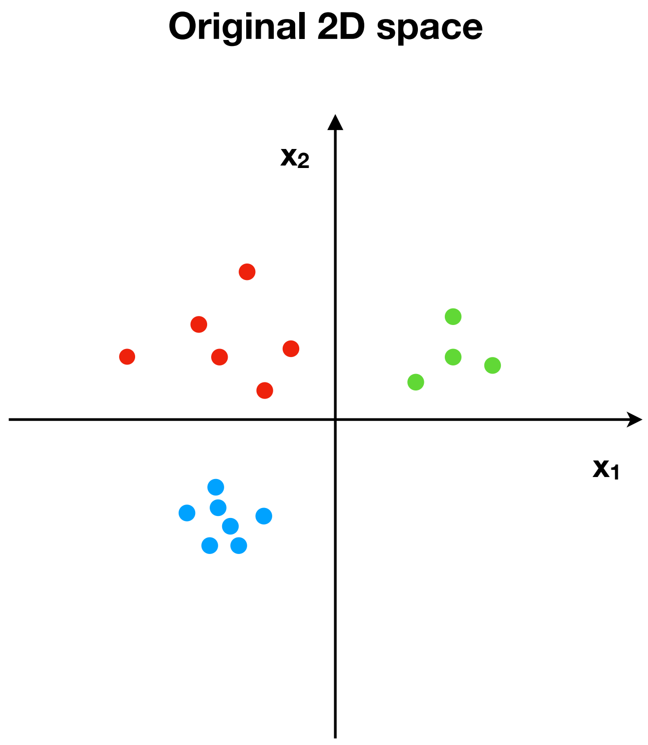


Shlens et al. 2014



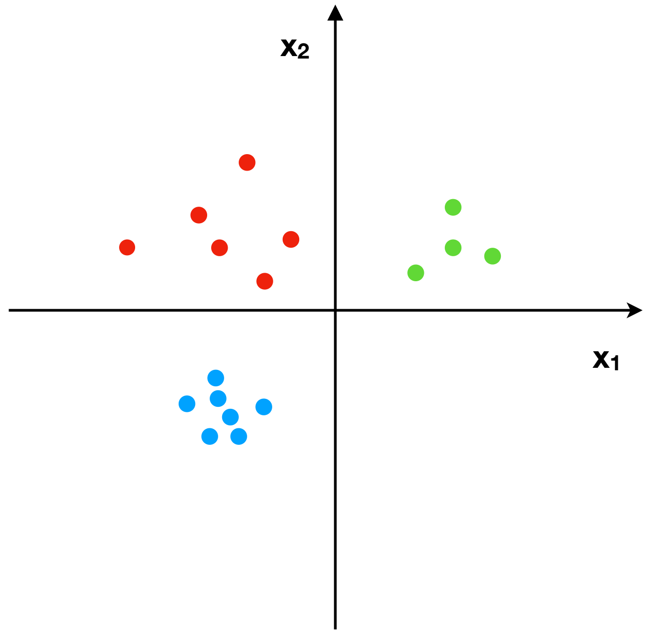
t-SNE and UMAP

t-SNE and UMAP

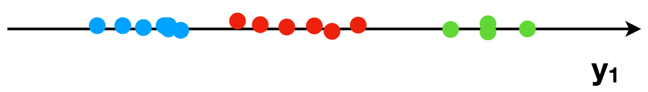


t-SNE and UMAP

Original 2D space

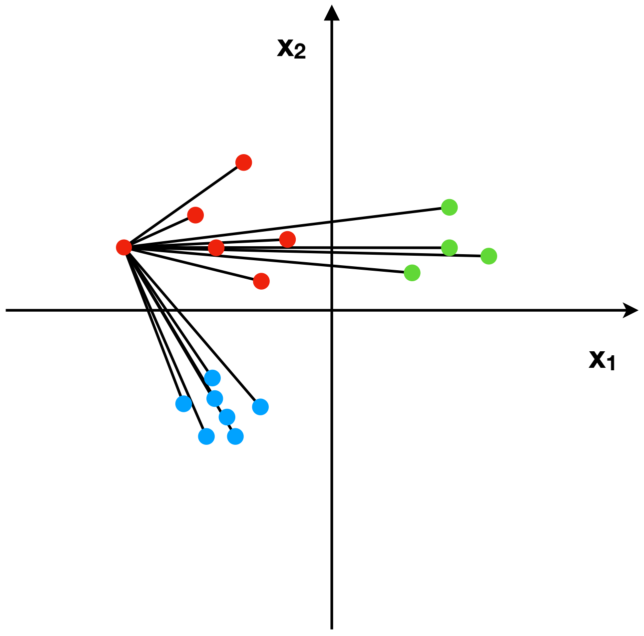


Dimension reduced 1D space

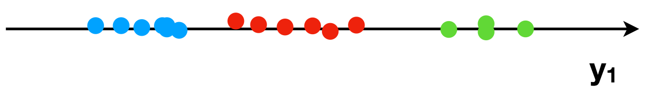


t-SNE and UMAP

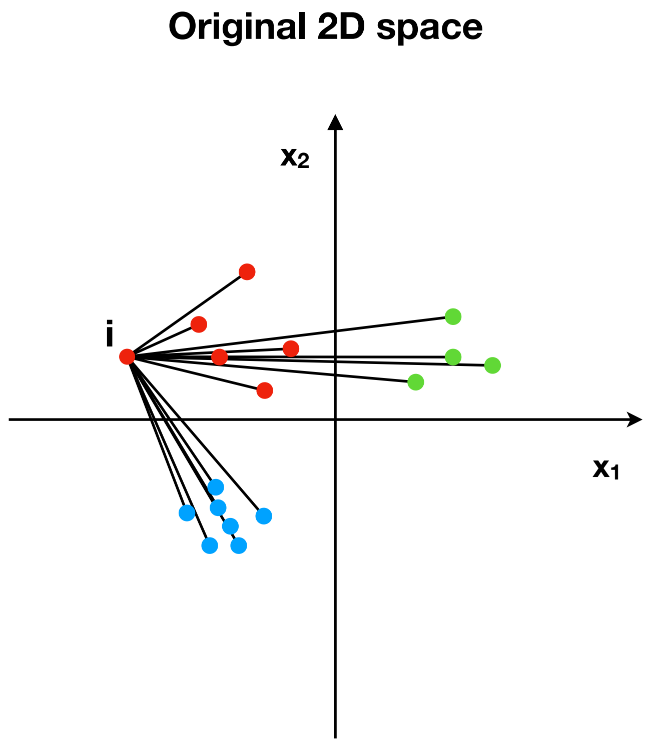
Original 2D space

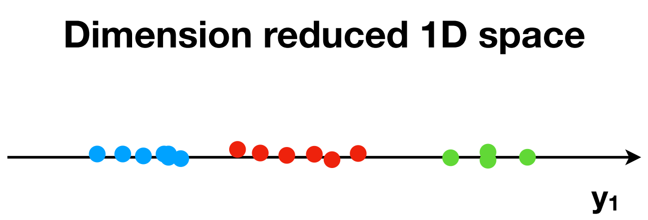


Dimension reduced 1D space



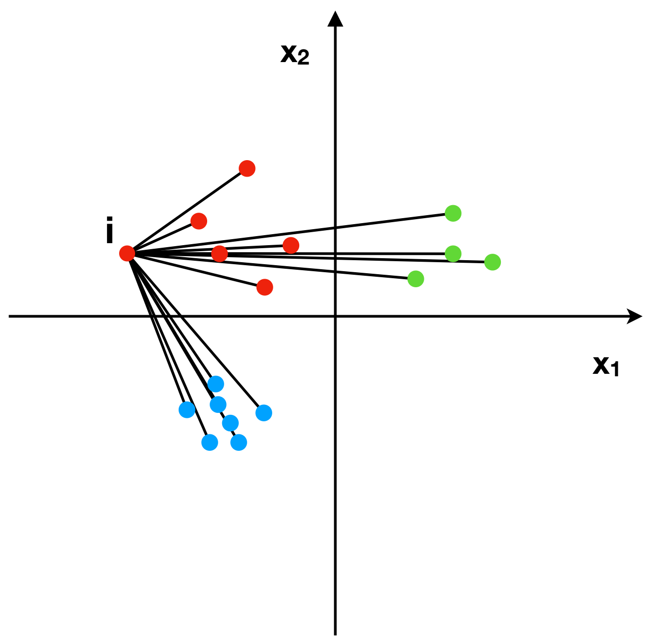
t-SNE and UMAP

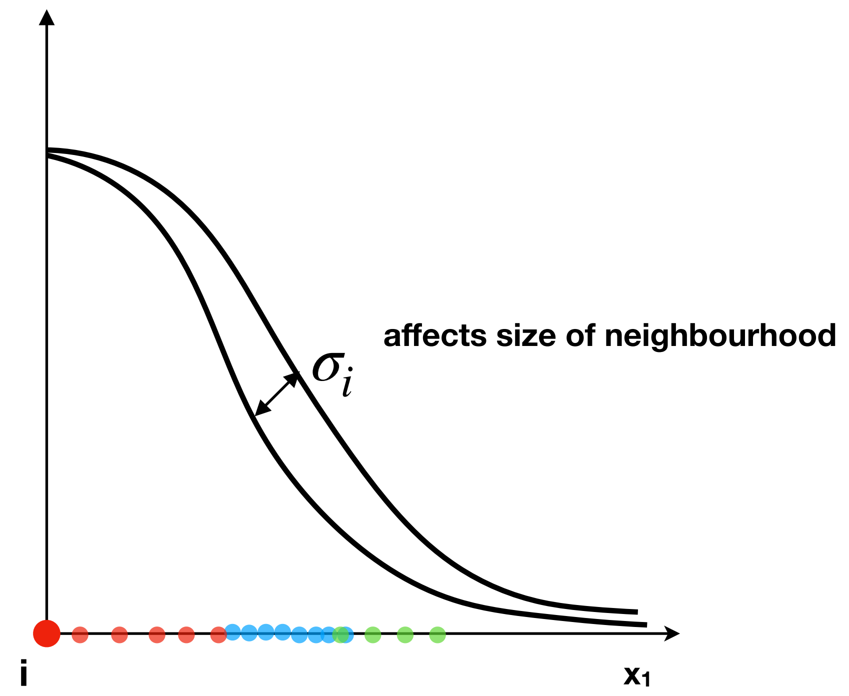




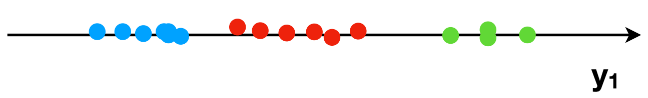
t-SNE and UMAP

Original 2D space



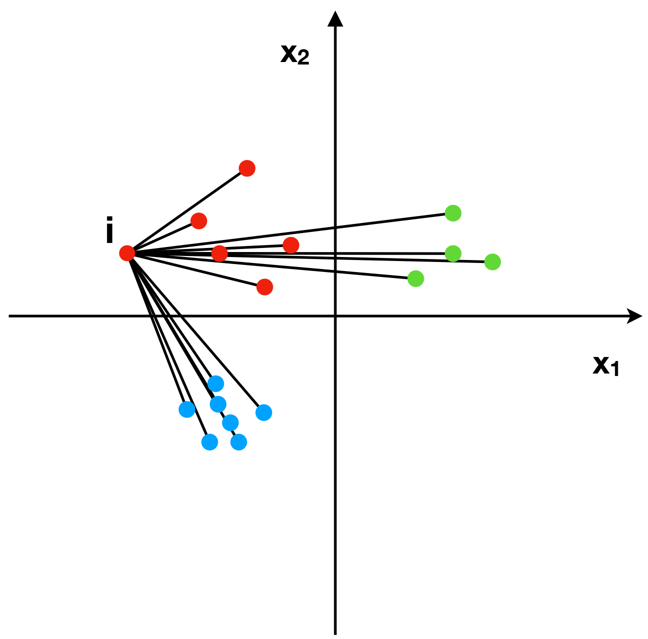


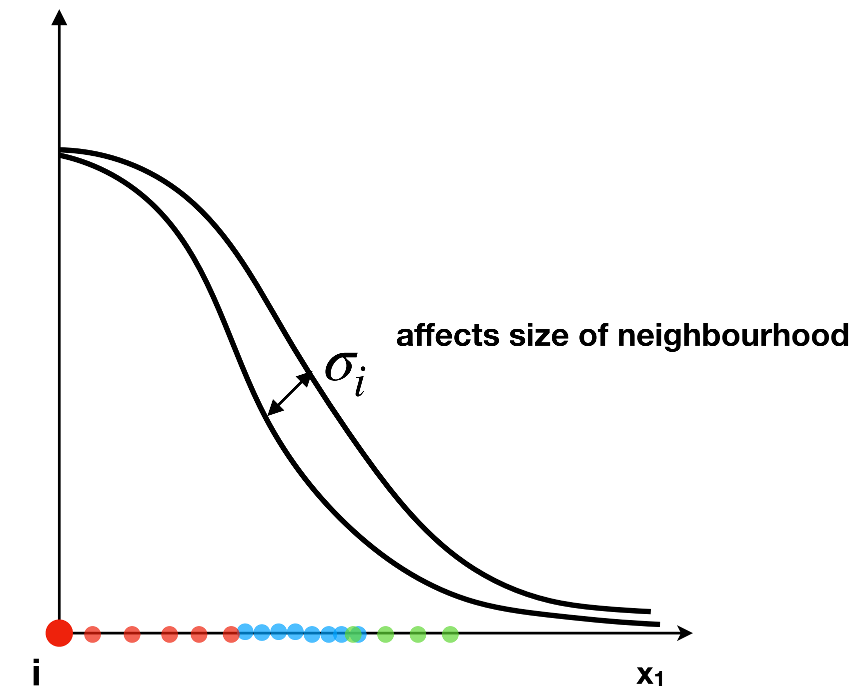
Dimension reduced 1D space



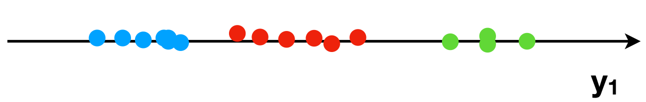
t-SNE and UMAP

Original 2D space





Dimension reduced 1D space



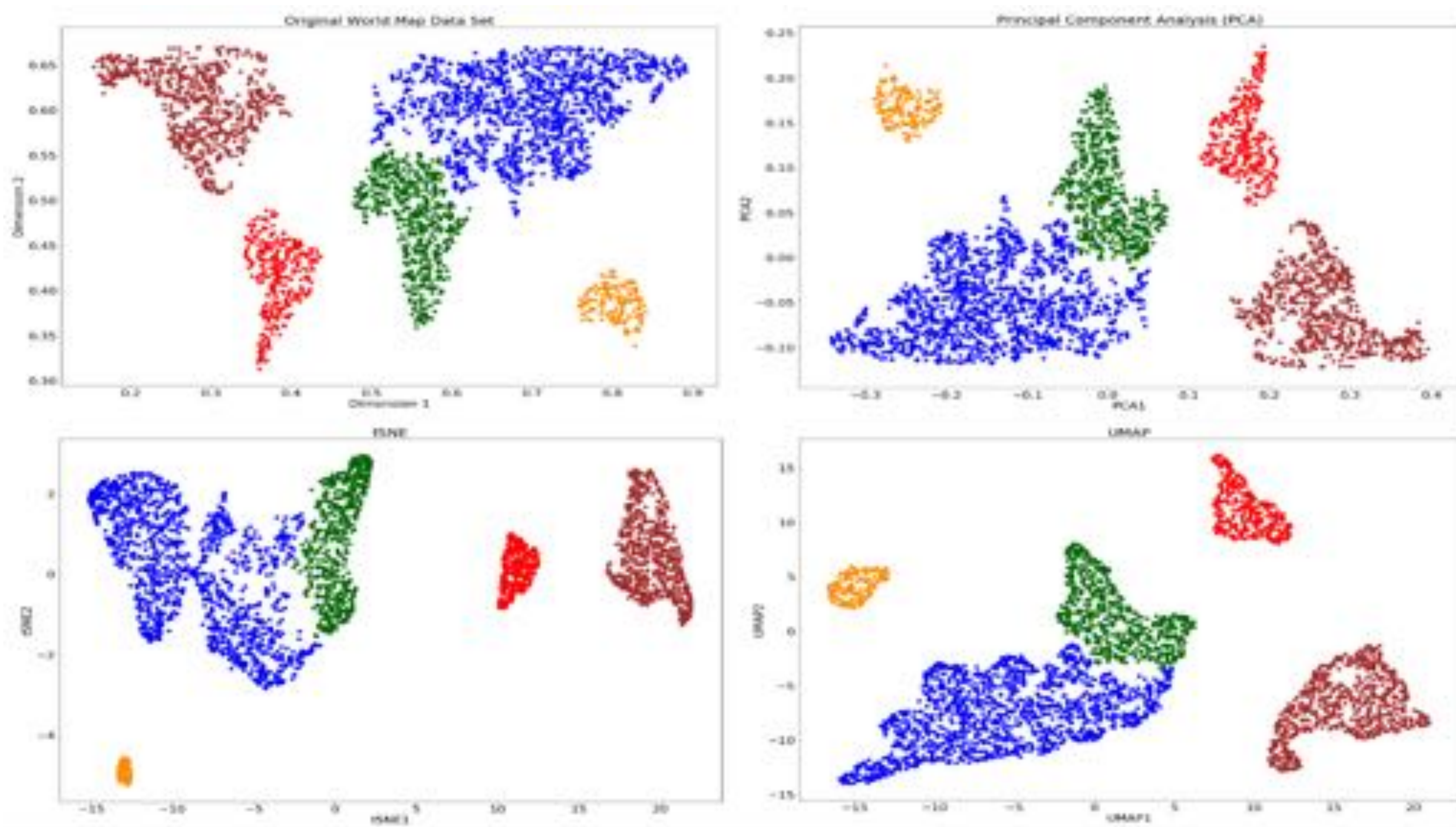
Objective function

$$C = D_{KL}(P\,||\,Q) = \sum_{x \in X} P(x) log \left(\frac{P(x)}{Q(x)} \right)$$

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - qij)(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1}$$

t-SNE and UMAP

Original **PCA**

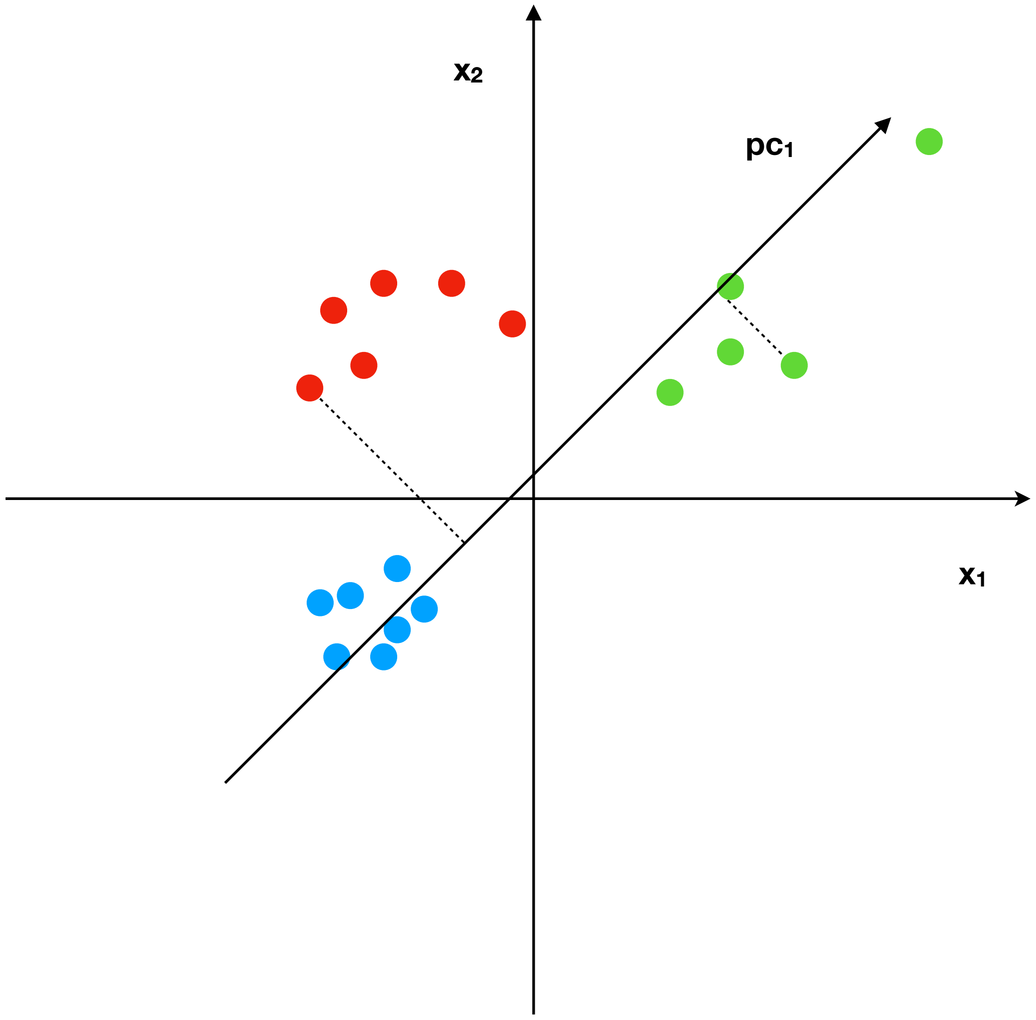


t-SNE **UMAP**

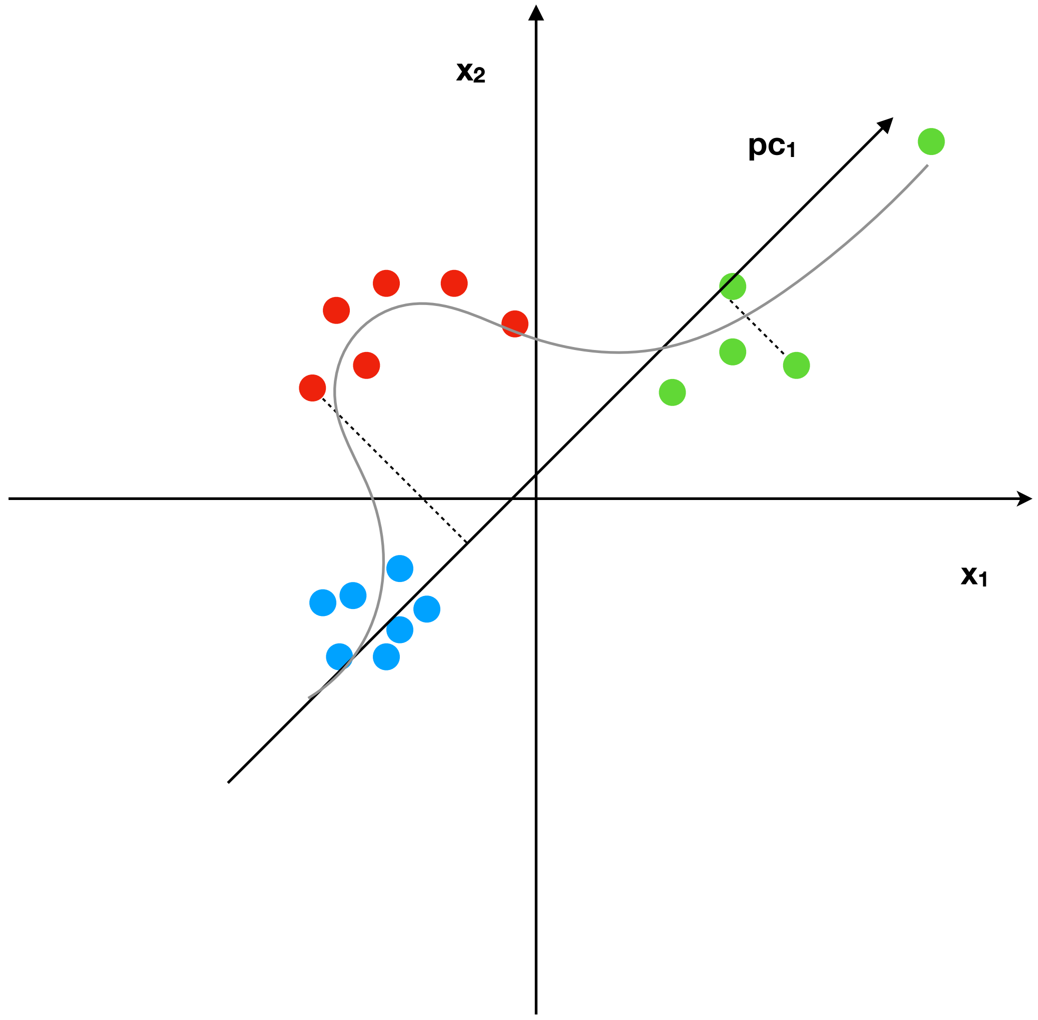
https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17

Autoencoder

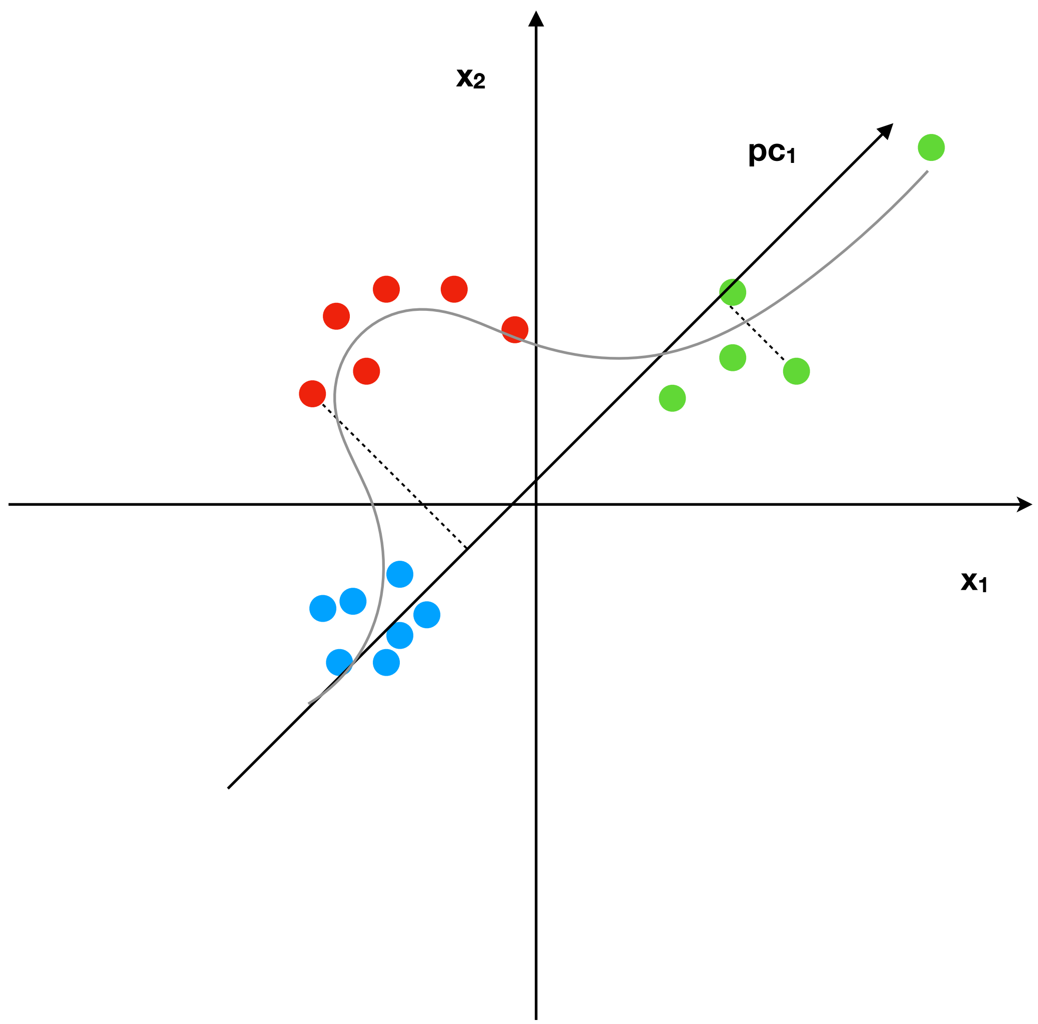
Autoencoder

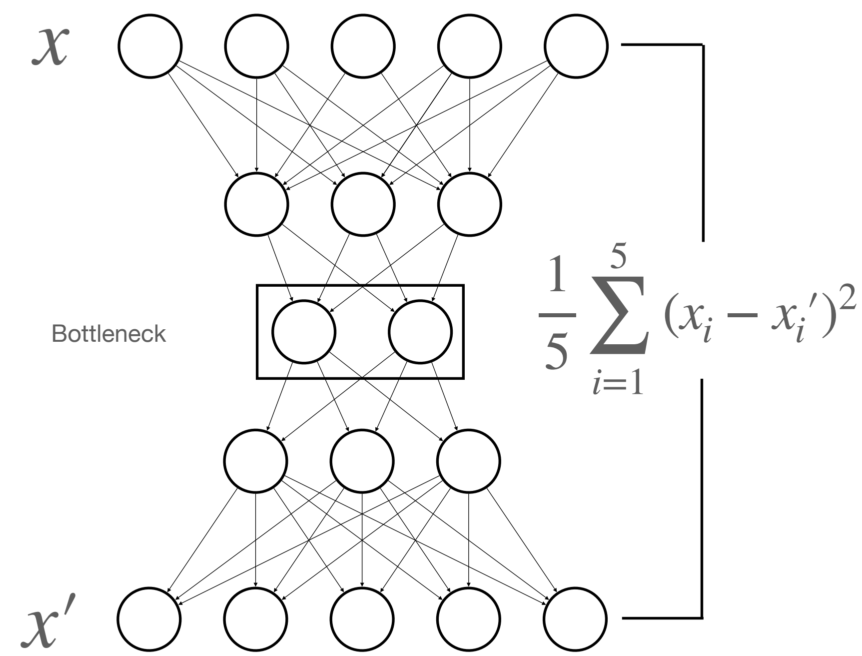


Autoencoder



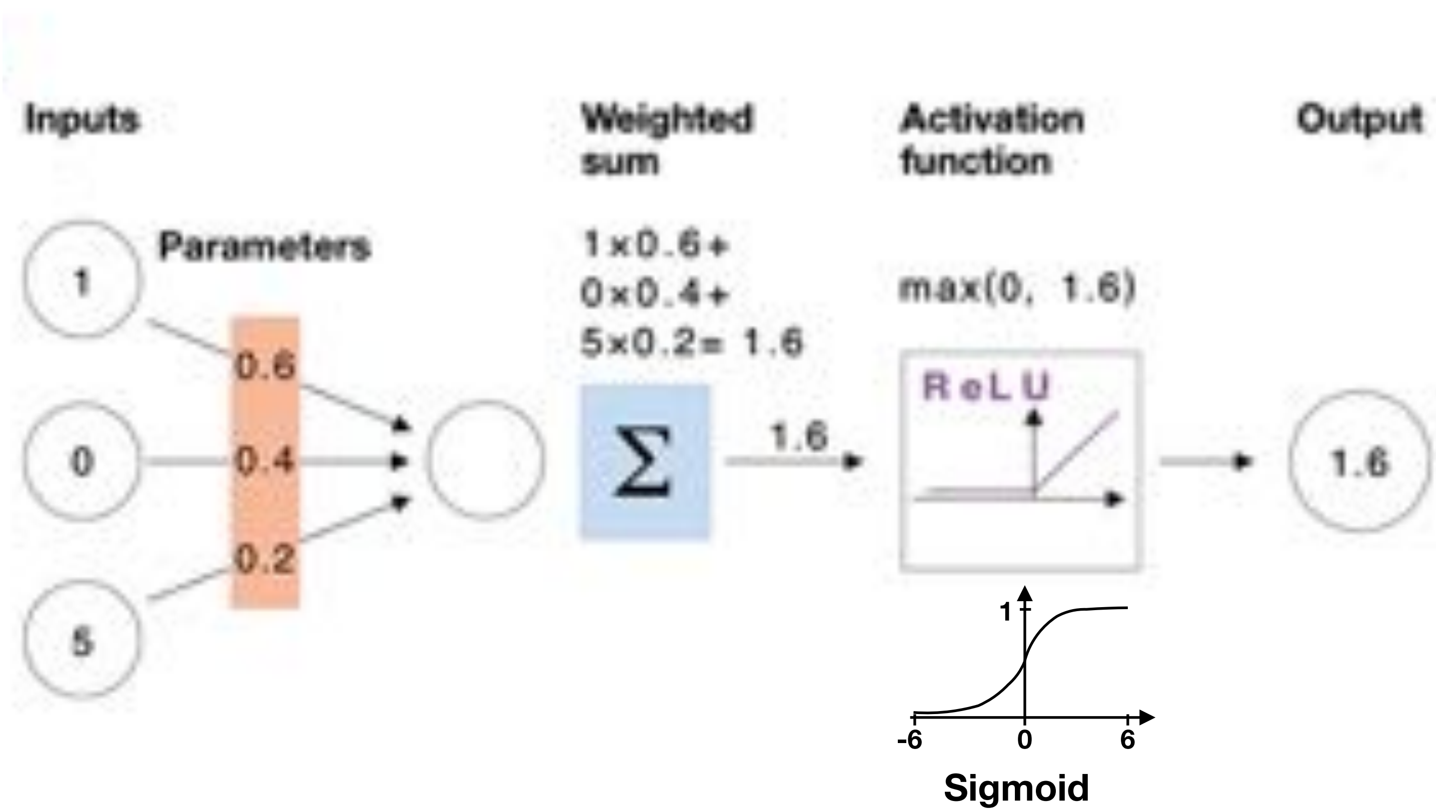
Autoencoder

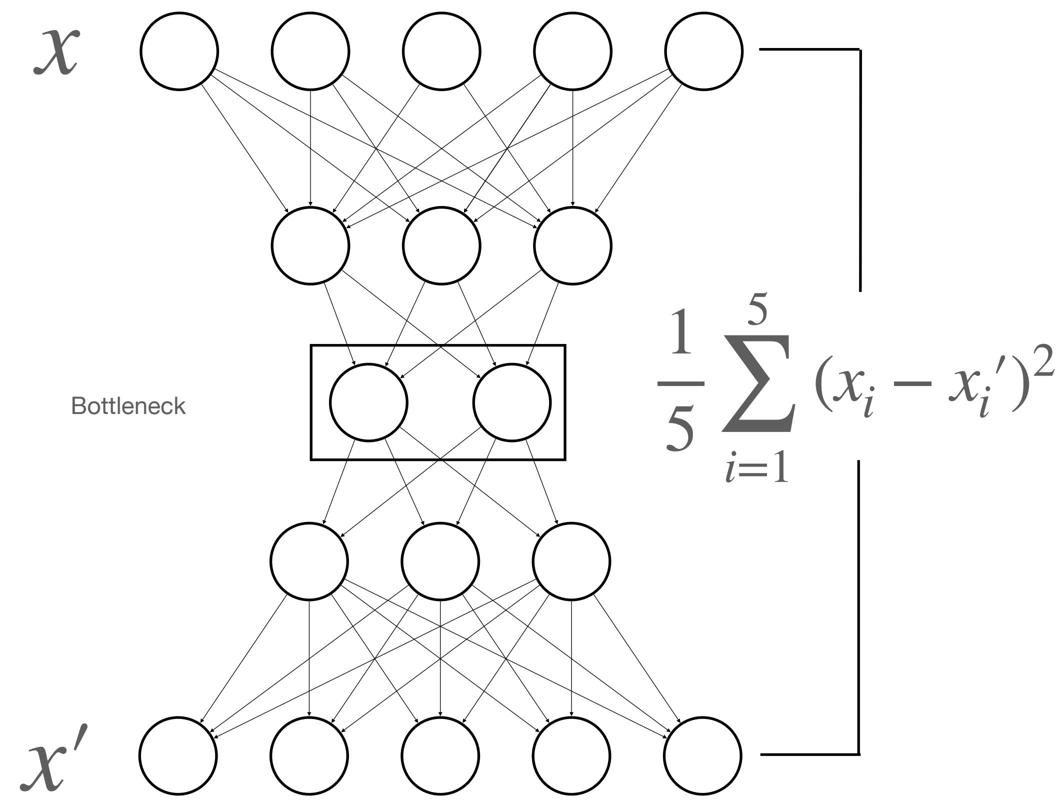




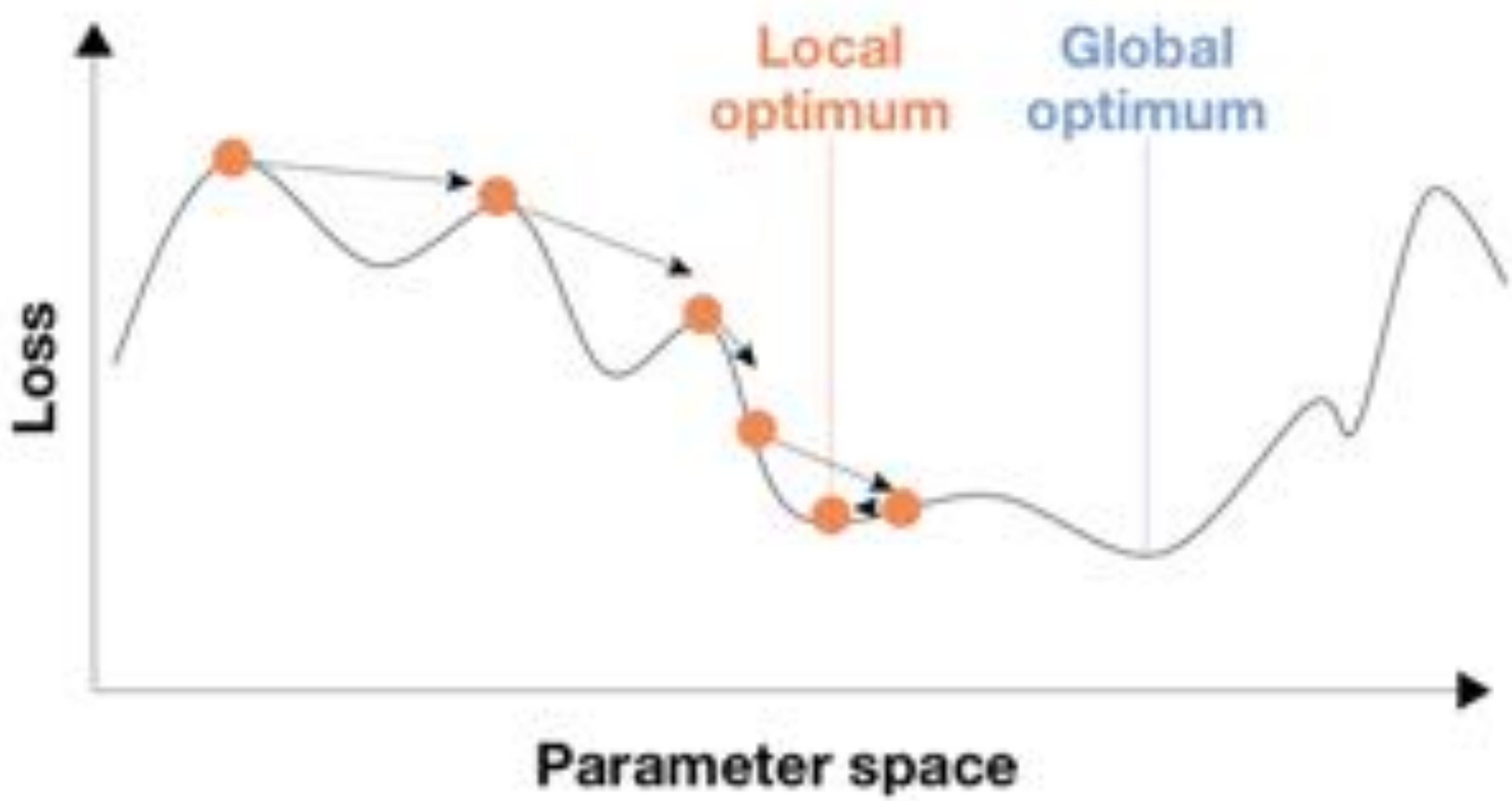
$x_1 =$	25	9	26	300	1
$x_1', \hat{x}_1 =$	20	4	20	310	1

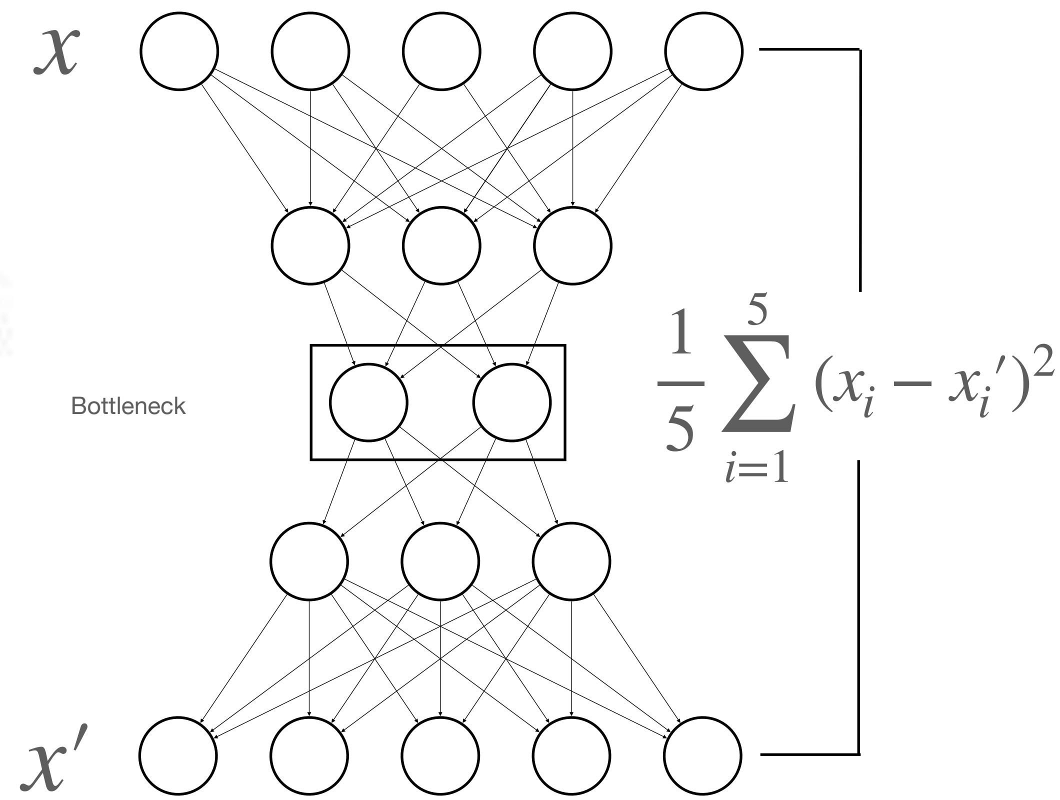
Learning non-linearities with deep neural networks



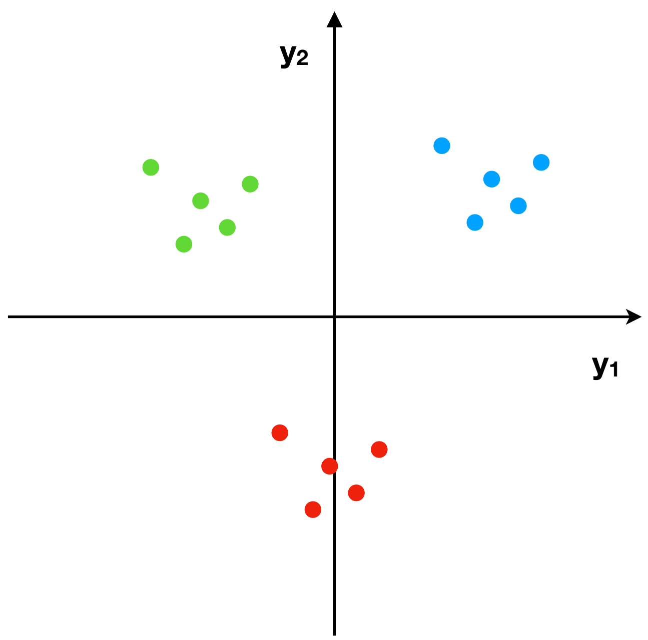


(Stochastic) gradient descent / Back-propagation

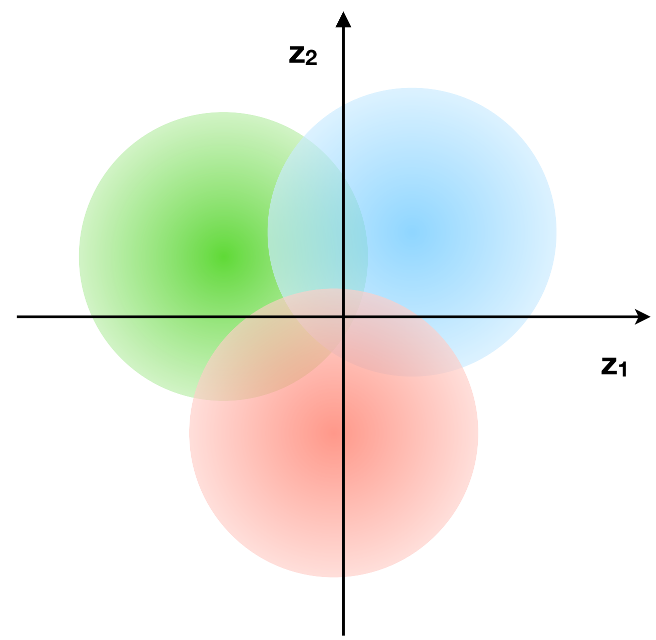




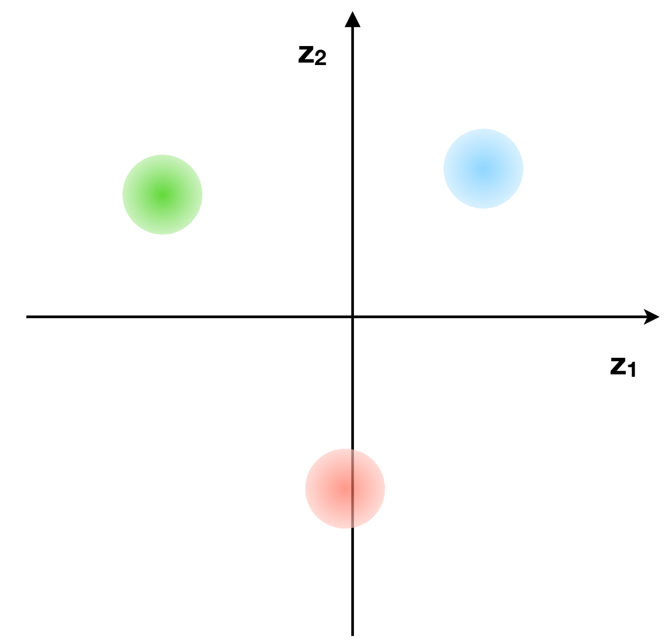
What is a good latent space ?



Ground truth

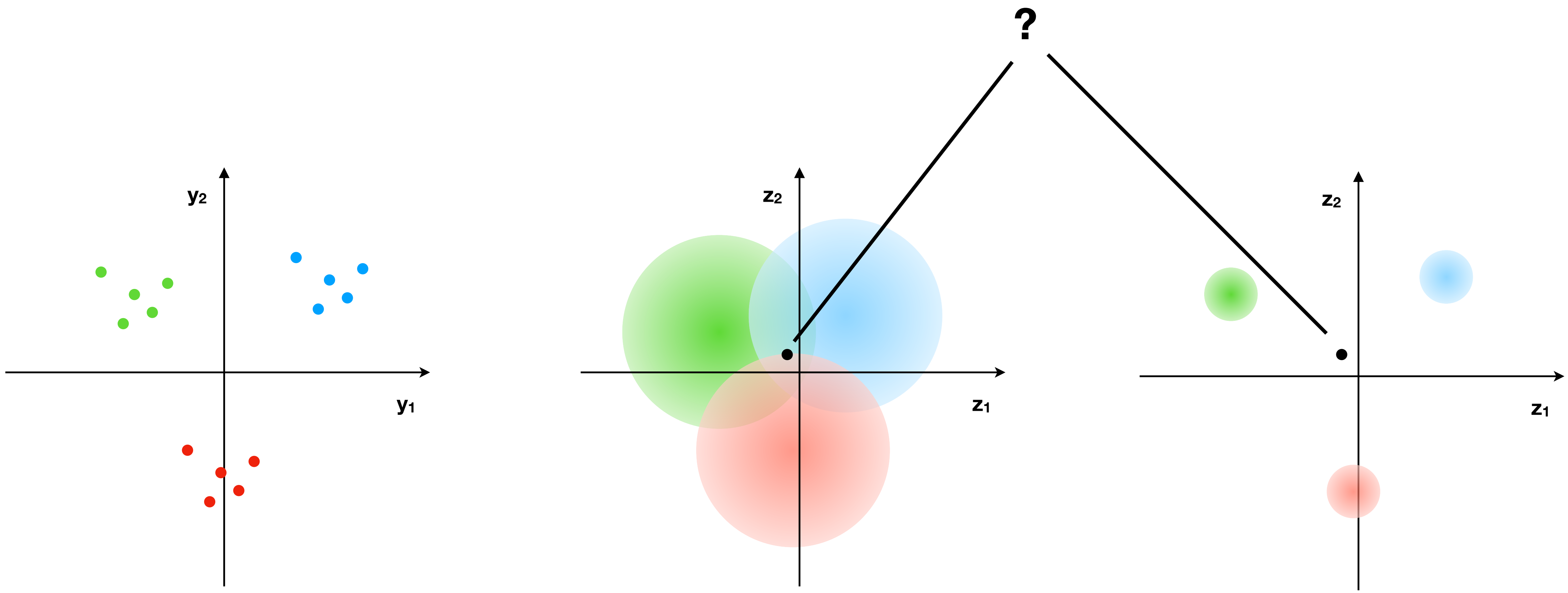


Latent space 1



Latent space 2

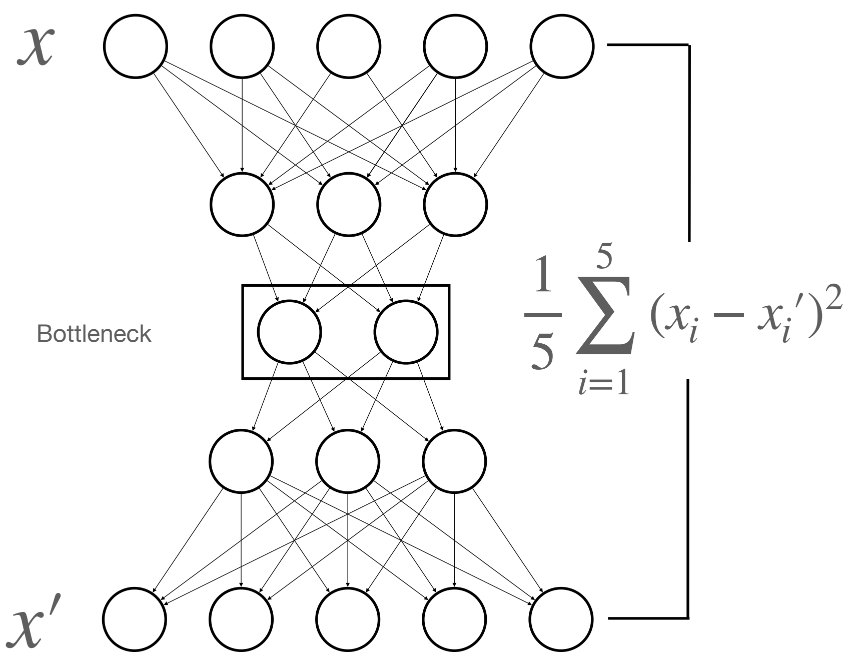
What is a good latent space ?



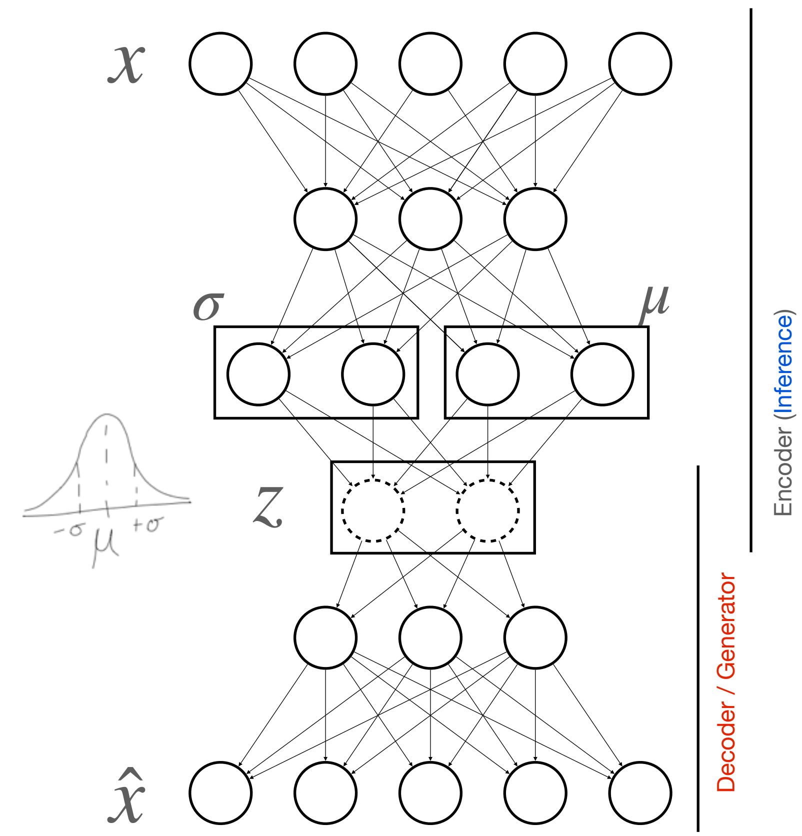
Ground truth **Latent space 1** **Latent space 2**

Variational Autoencoder (VAE)

Autoencoder



Variational Autoencoder

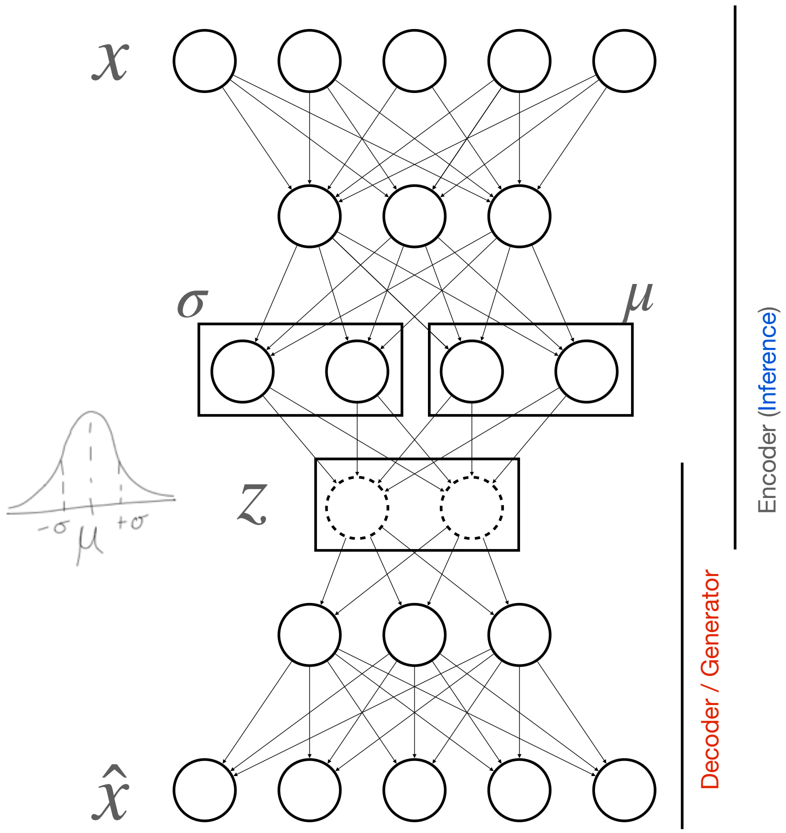


random

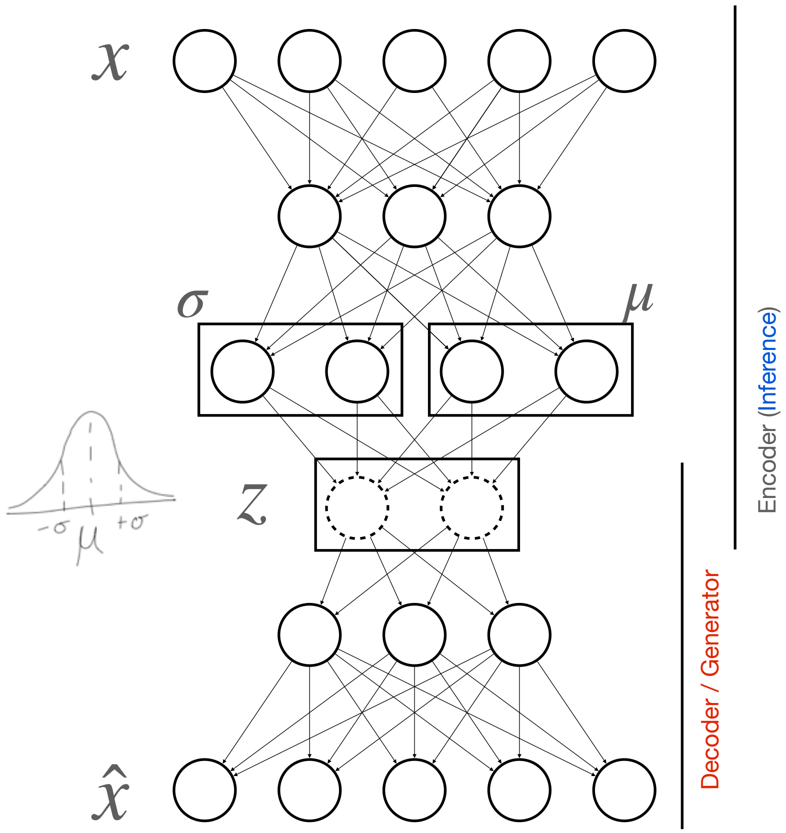
deterministic

$x_1 =$	25	9	26	300	1
$x_1', \hat{x}_1 =$	20	4	20	310	1

VAE viewed as a probabilistic random variable model



VAE viewed as a probabilistic random variable model



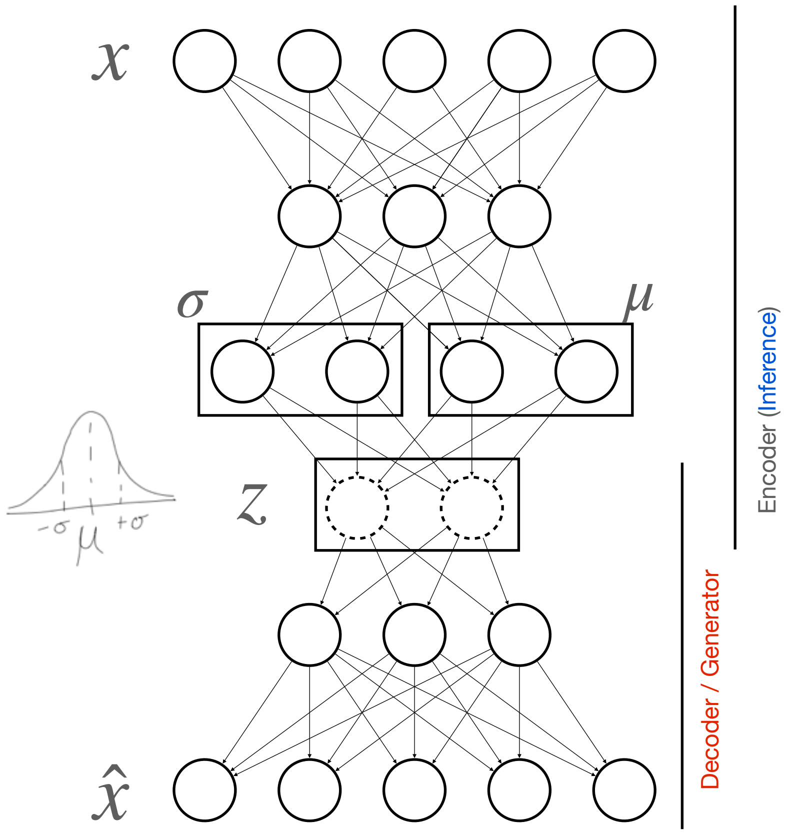
We want to learn the posterior distribution over the latent variables, given the data.

likelihood **prior**

$$p(z \mid x) = \frac{p(x \mid z) * p(z)}{p(x)}$$

evidence

VAE viewed as a probabilistic random variable model



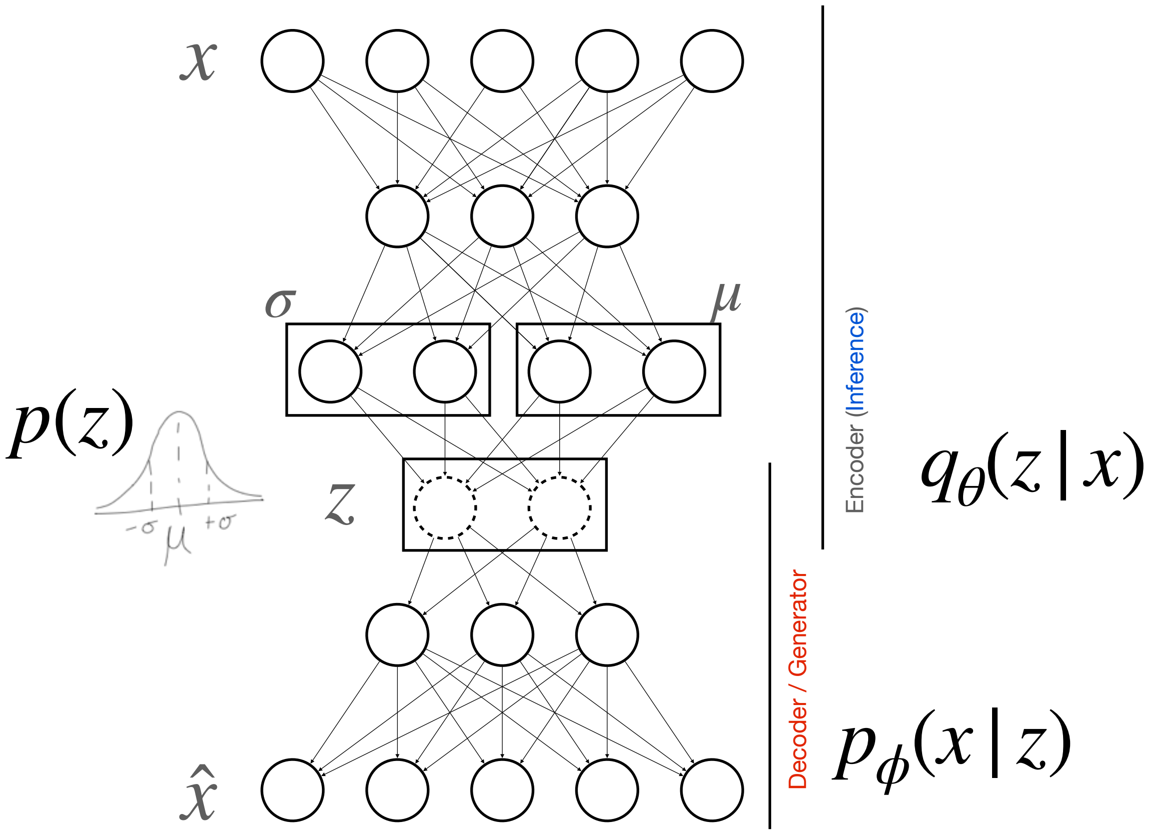
We want to learn the posterior distribution over the latent variables, given the data.

$$p(z|x) = \frac{\overbrace{p(x|z)}^{\textbf{likelihood}} * \overbrace{p(z)}^{\textbf{prior}}}{\underbrace{p(x)}_{\textbf{evidence}}}$$

Objective function: Evidence Lower BOund

$$log p_\theta(x_i) \geq \mathbb{E}_{q_\theta(z|x_i)}[log p_\phi(x_i|z)] - \mathbb{KL}(q_\theta(z|x_i)||p(z))$$

VAE viewed as a probabilistic random variable model



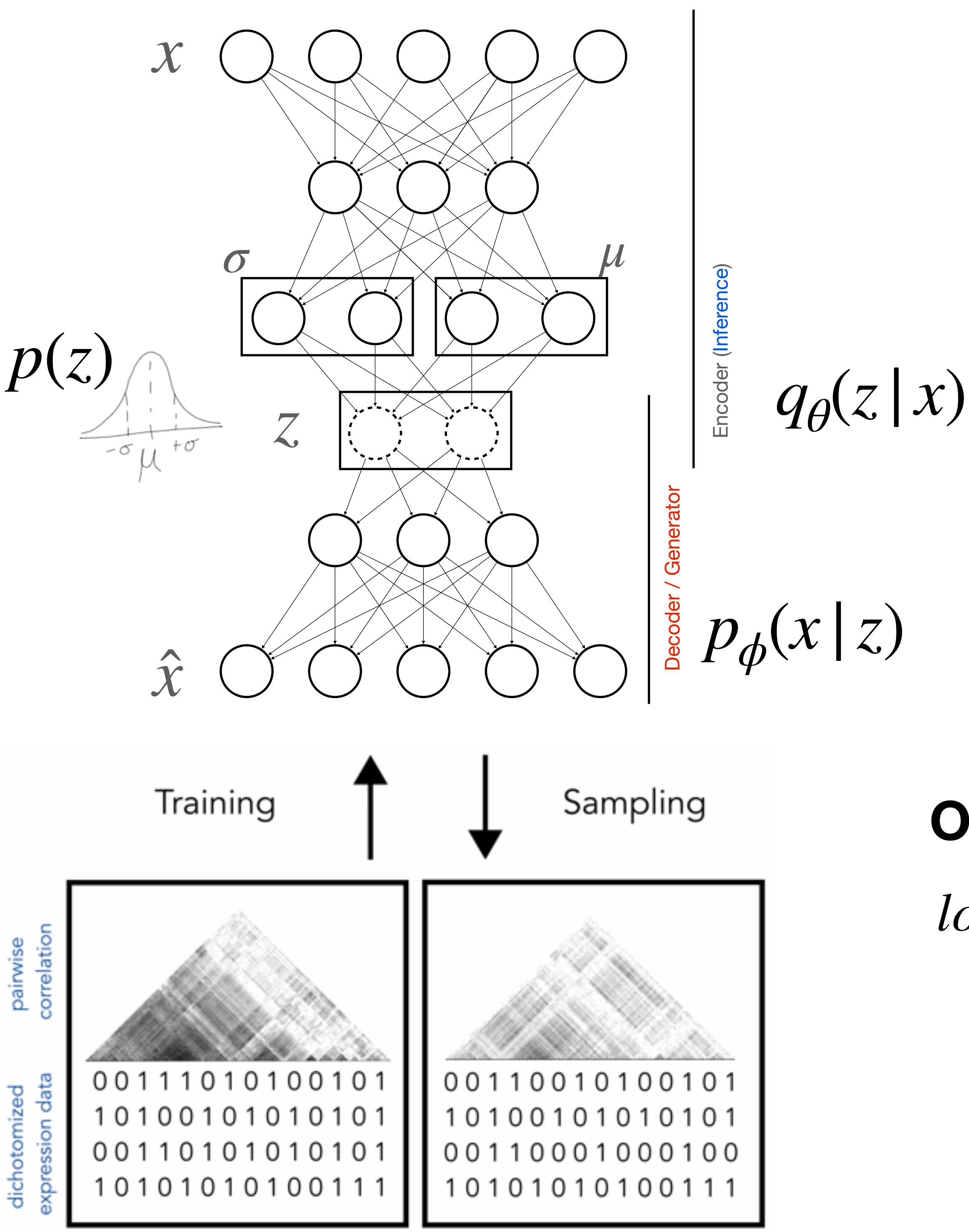
We want to learn the posterior distribution over the latent variables, given the data.

$$p(z|x) = \frac{\overbrace{p(x|z)}^{\text{likelihood}} * \overbrace{p(z)}^{\text{prior}}}{\underbrace{p(x)}_{\text{evidence}}}$$

Objective function: Evidence Lower BOund

$$logp_\theta(x_i) \geq \mathbb{E}_{q_\theta(z|x_i)}[logp_\phi(x_i|z)] - \mathbb{KL}(q_\theta(z|x_i)||p(z))$$

VAE viewed as a probabilistic random variable model



We want to learn the posterior distribution over the latent variables, given the data.

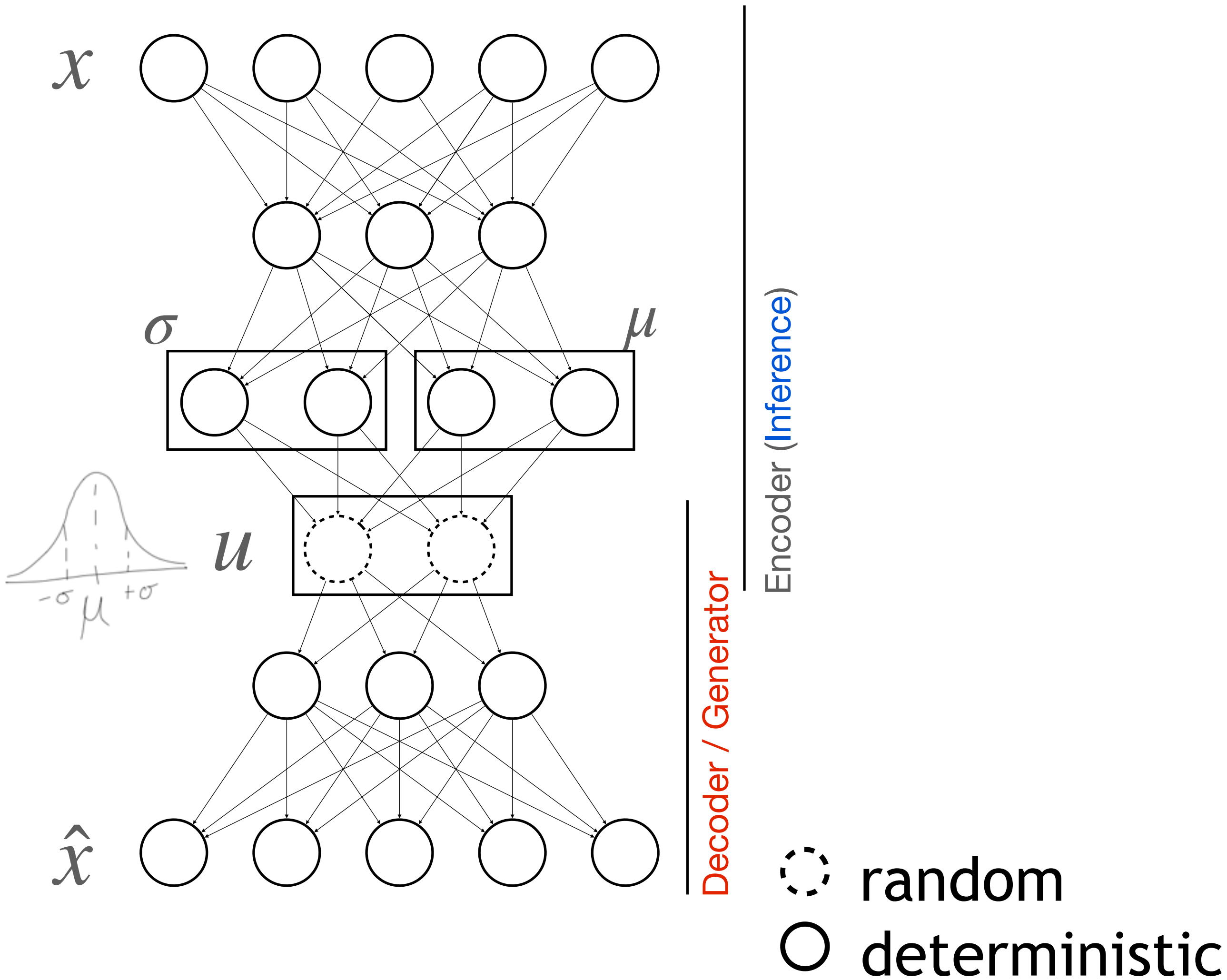
$$p(z|x) = \frac{\overbrace{p(x|z)}^{\text{likelihood}} * \overbrace{p(z)}^{\text{prior}}}{\underbrace{p(x)}_{\text{evidence}}}$$

Objective function: Evidence Lower BOund

$$log p_\theta(x_i) \geq \mathbb{E}_{q_\theta(z|x_i)}[log p_\phi(x_i|z)] - \mathbb{KL}(q_\theta(z|x_i)||p(z))$$

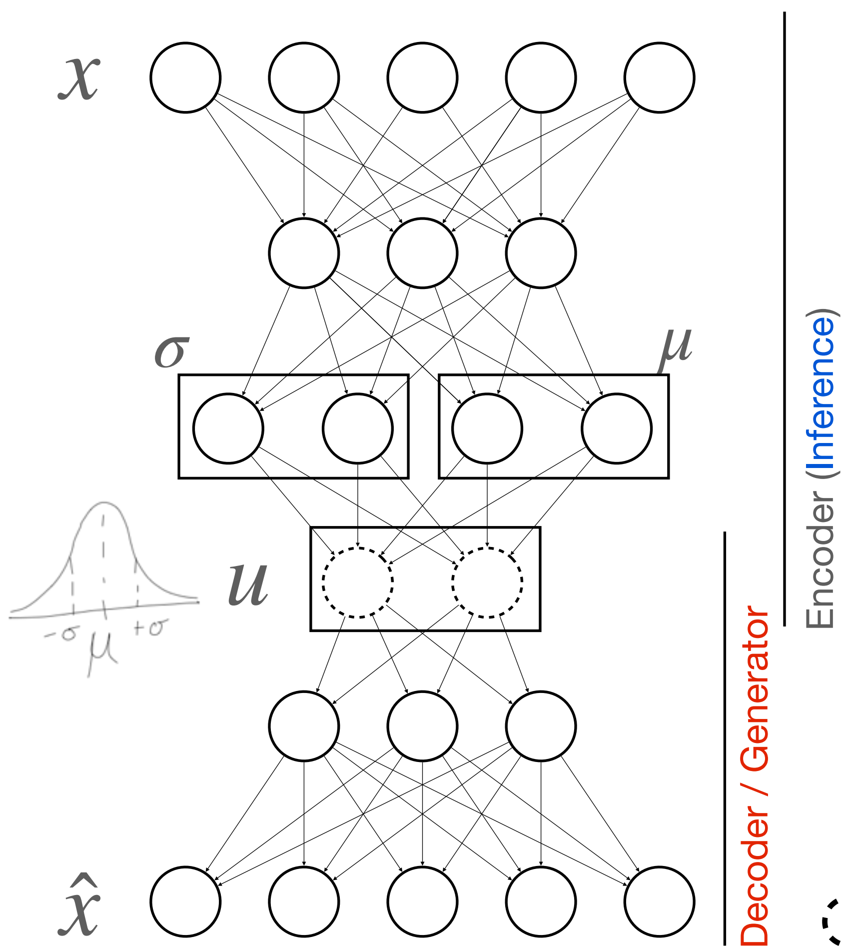
Other deep generative approaches …

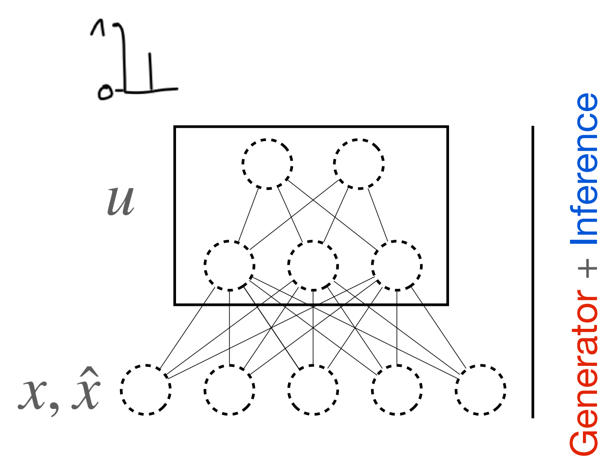
Variational Autoencoder



Other deep generative approaches …

Variational Autoencoder



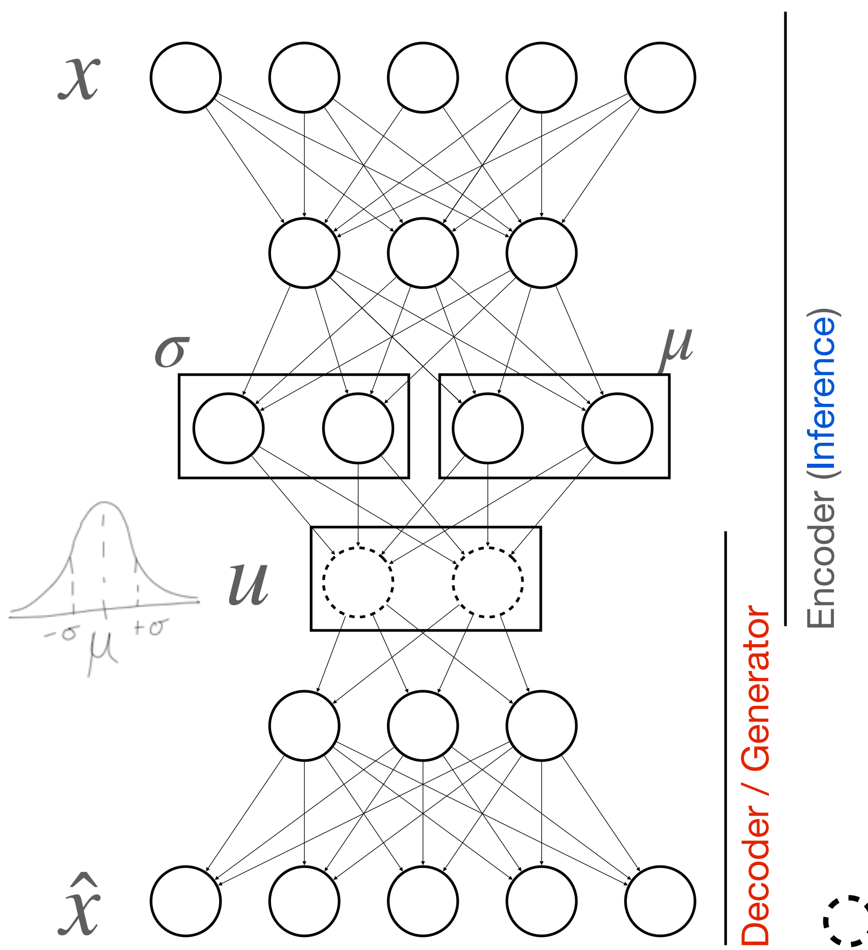


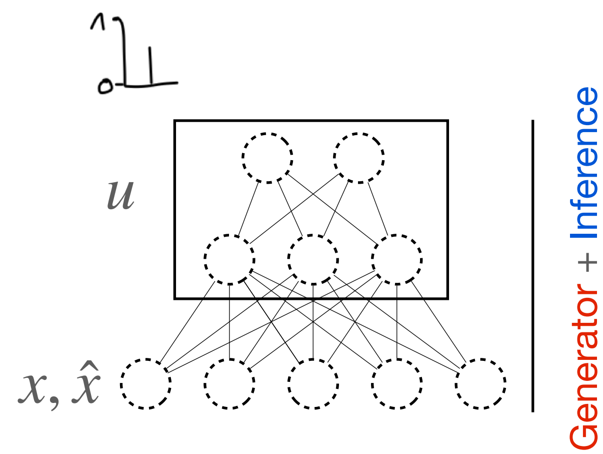
Deep Boltzmann
Machine

◌ random
○ deterministic

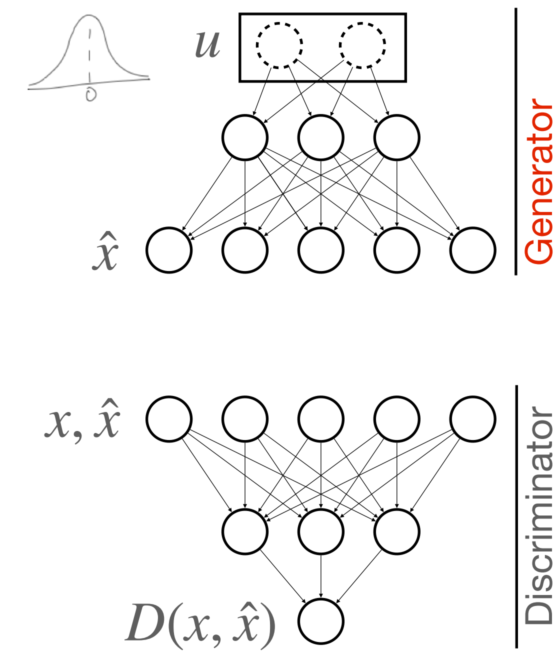
Other deep generative approaches …

Variational Autoencoder





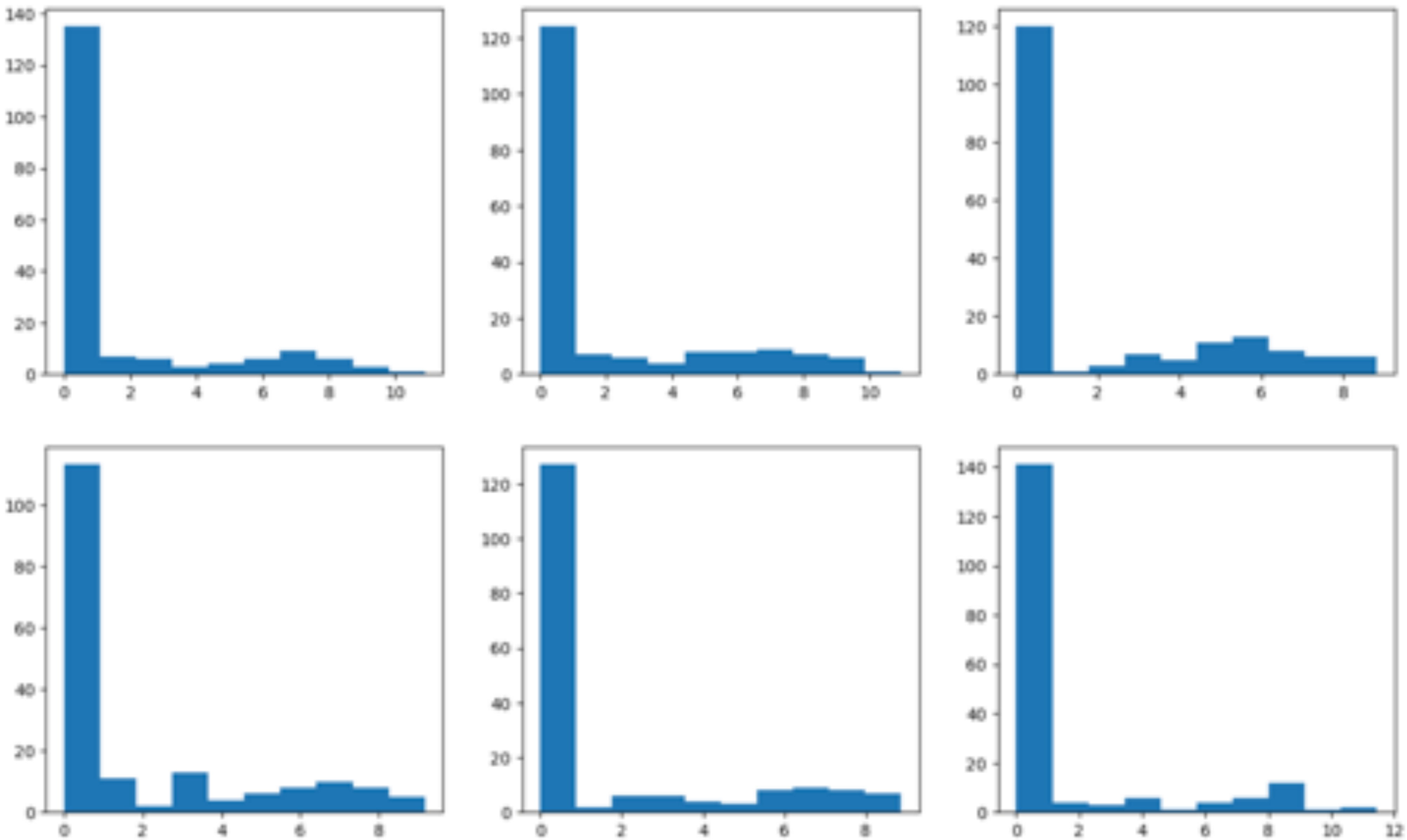
Deep Boltzmann Machine



Generative Adversarial Network

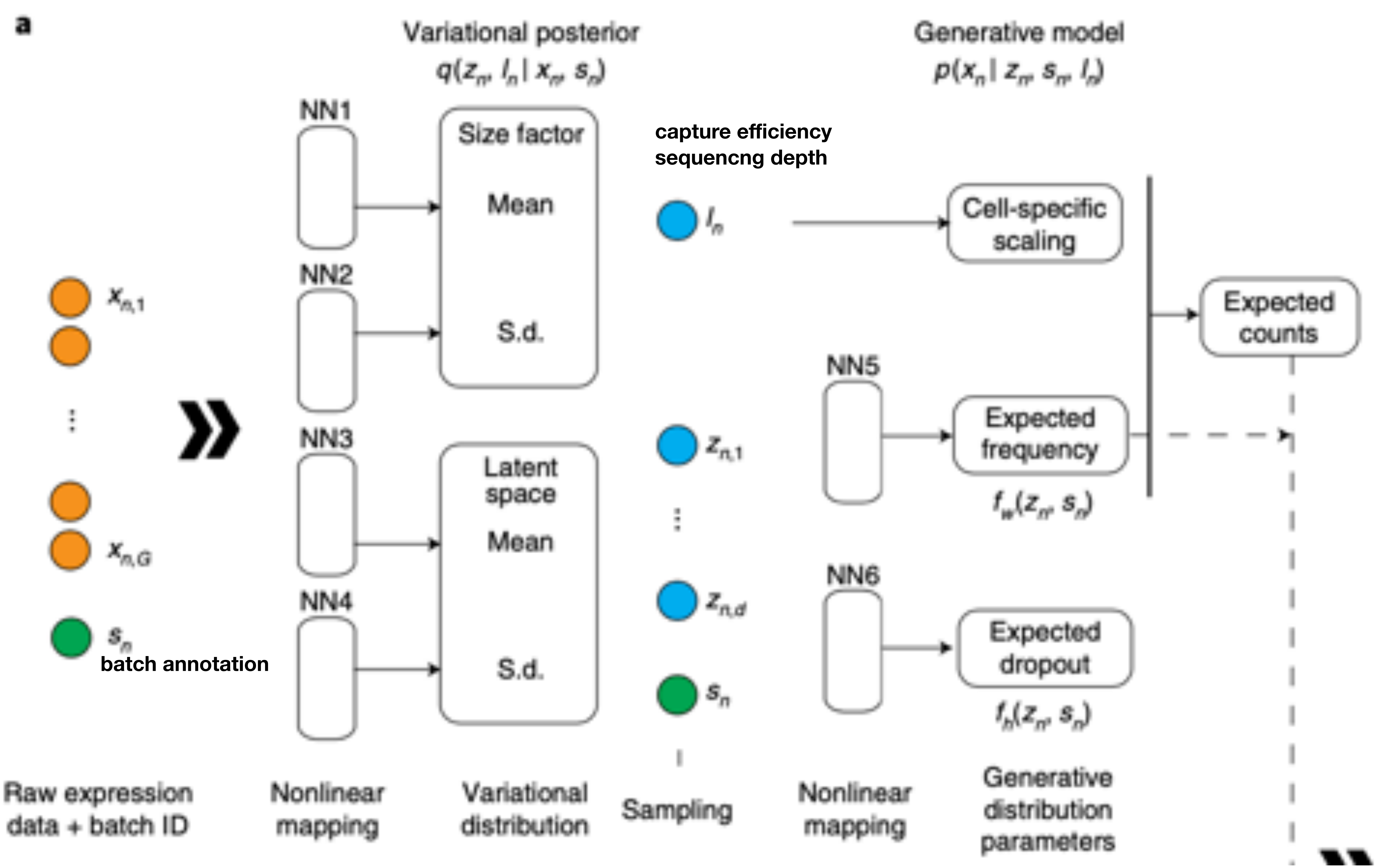
random
deterministic

Adaptations to single cell RNA-Seq data



Exemplary gene expression (log transformed) of six random genes from Tasic et al. (2016) data

Adaptations to single cell RNA-Seq data



From Lopez et al. 2018: "Deep generative modeling for single-cell transcriptomics"
"