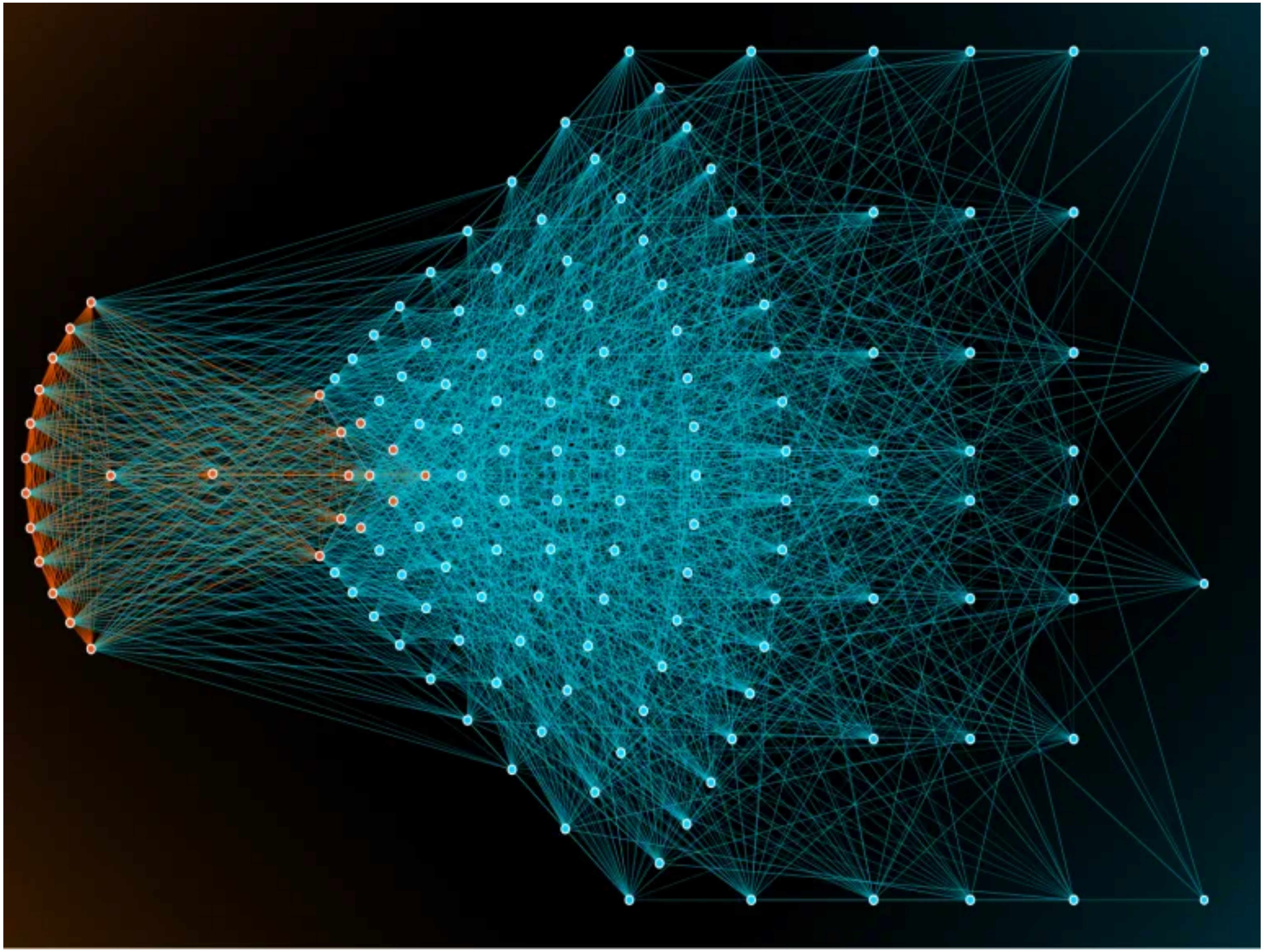
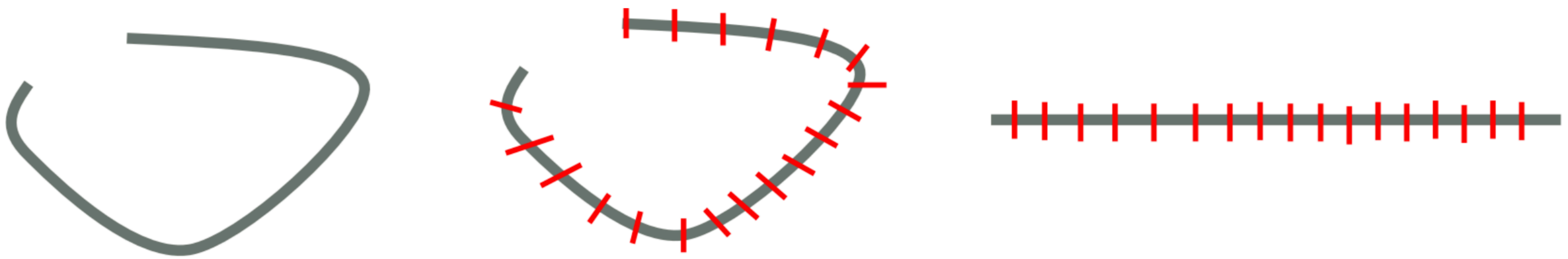


Other DR

LLE

Locally Linear Embedding

Aim: preserve local (linear) structure of data in the embedding space



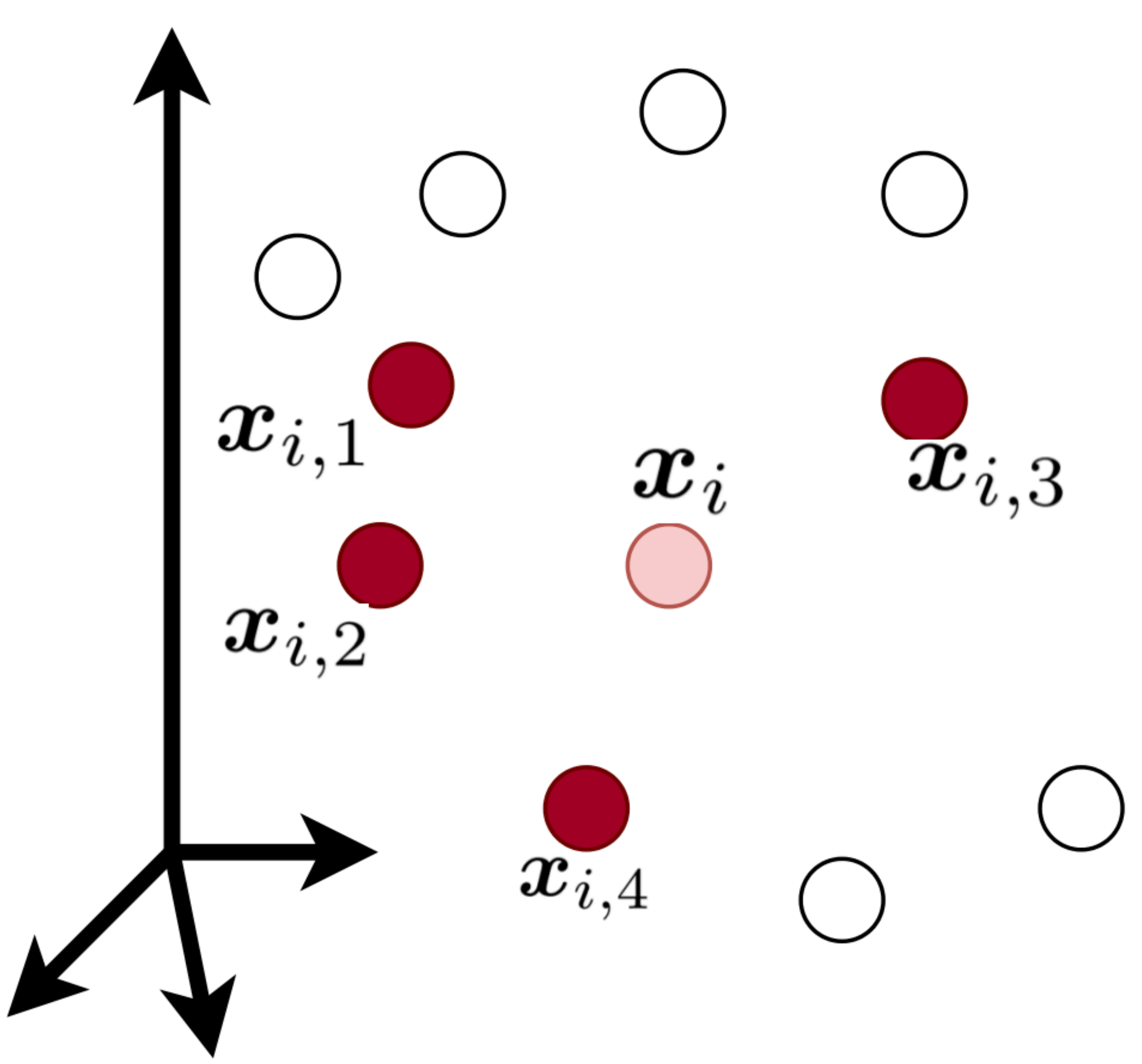
Ghojogh et al, 2020

LLE

Locally Linear Embedding

Aim: preserve local structure of data in the embedding space

Step 1. Neighbour selection



Ghojogh et al, 2020

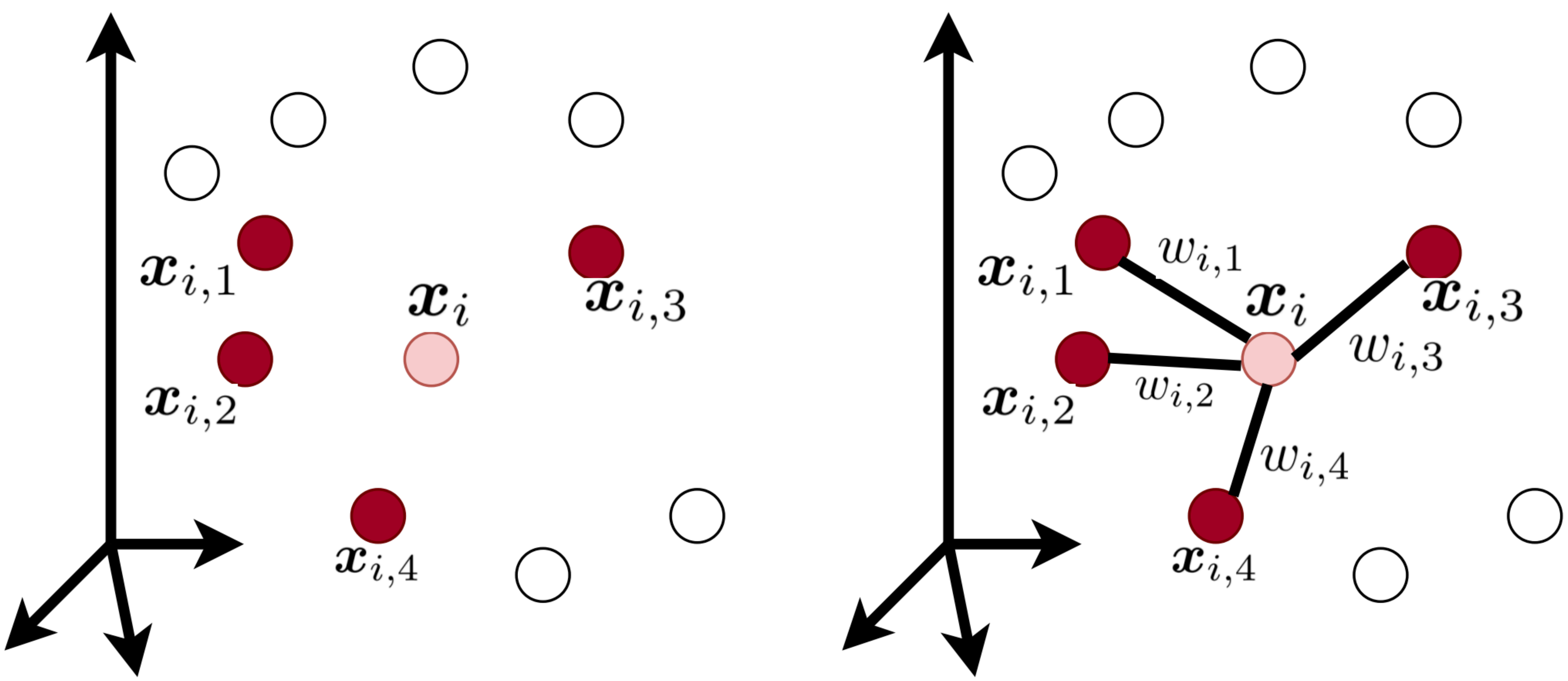
LLE

Locally Linear Embedding

Aim: preserve local structure of data in the embedding space

Step 1. Neighbour selection

Step 2. Weights calculation



Ghojogh et al, 2020

LLE

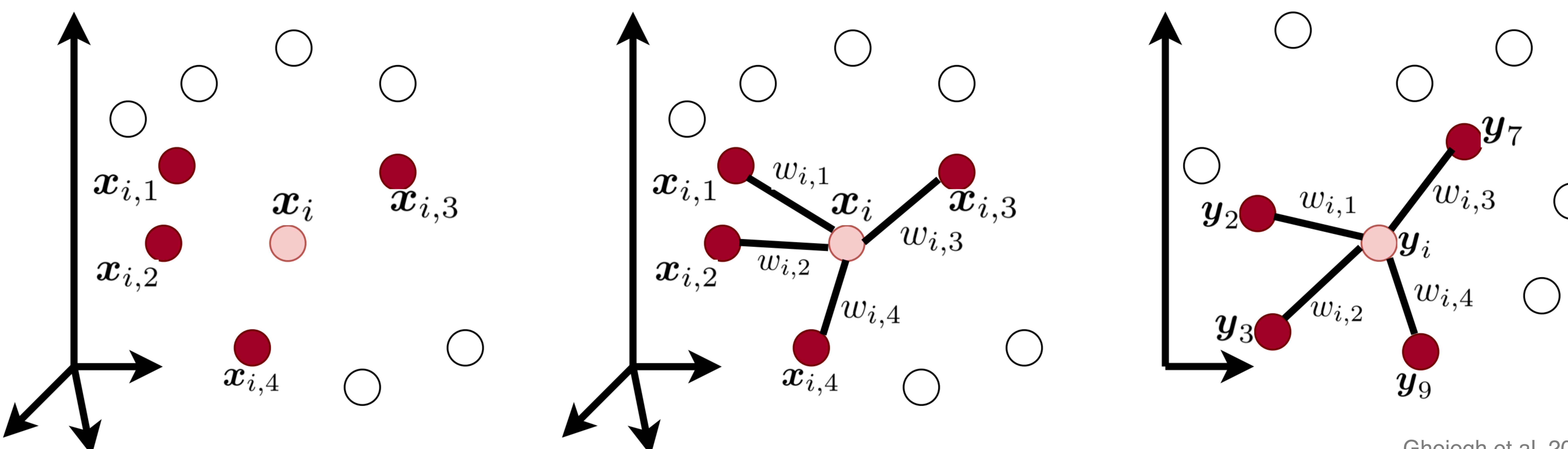
Locally Linear Embedding

Aim: preserve local structure of data in the embedding space

Step 1. Neighbour selection

Step 2. Weights calculation

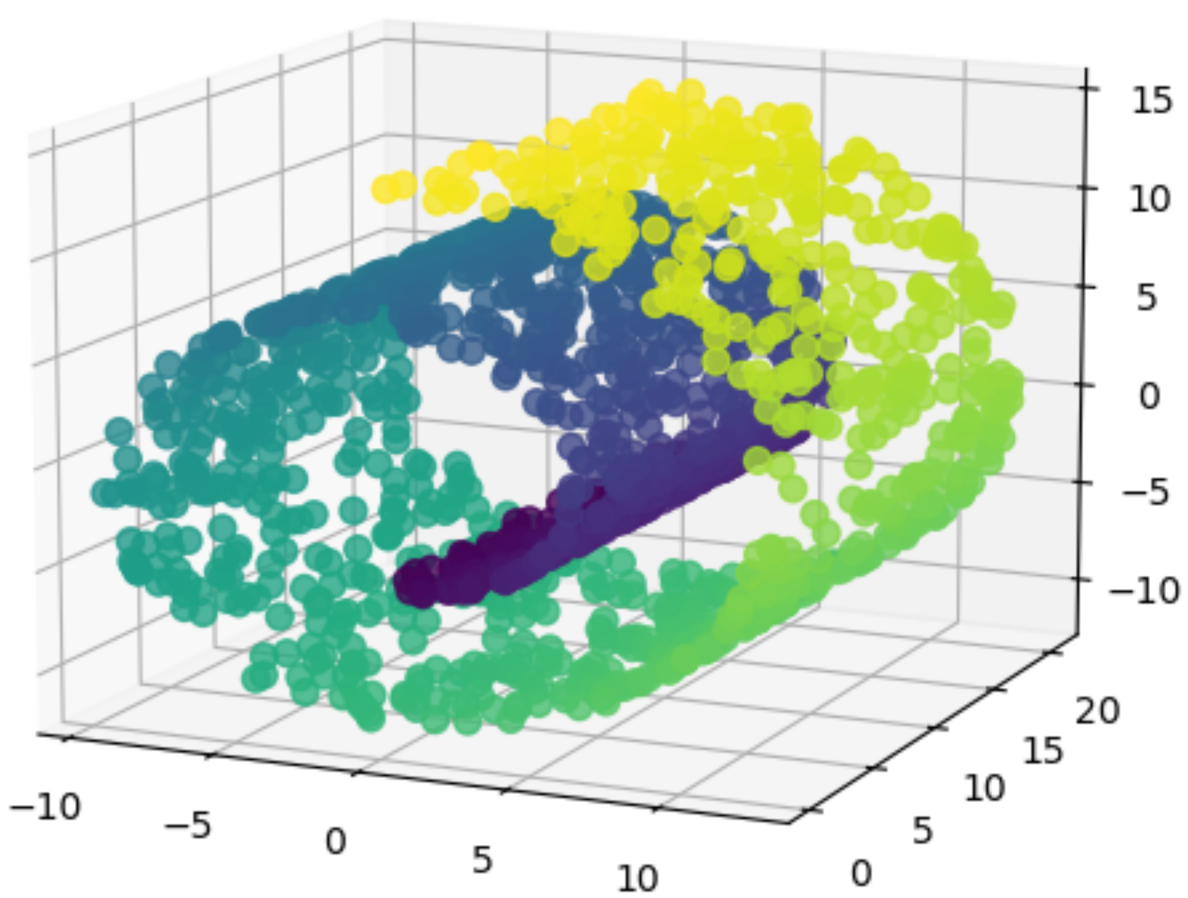
Step 3. Low dimensional embedding

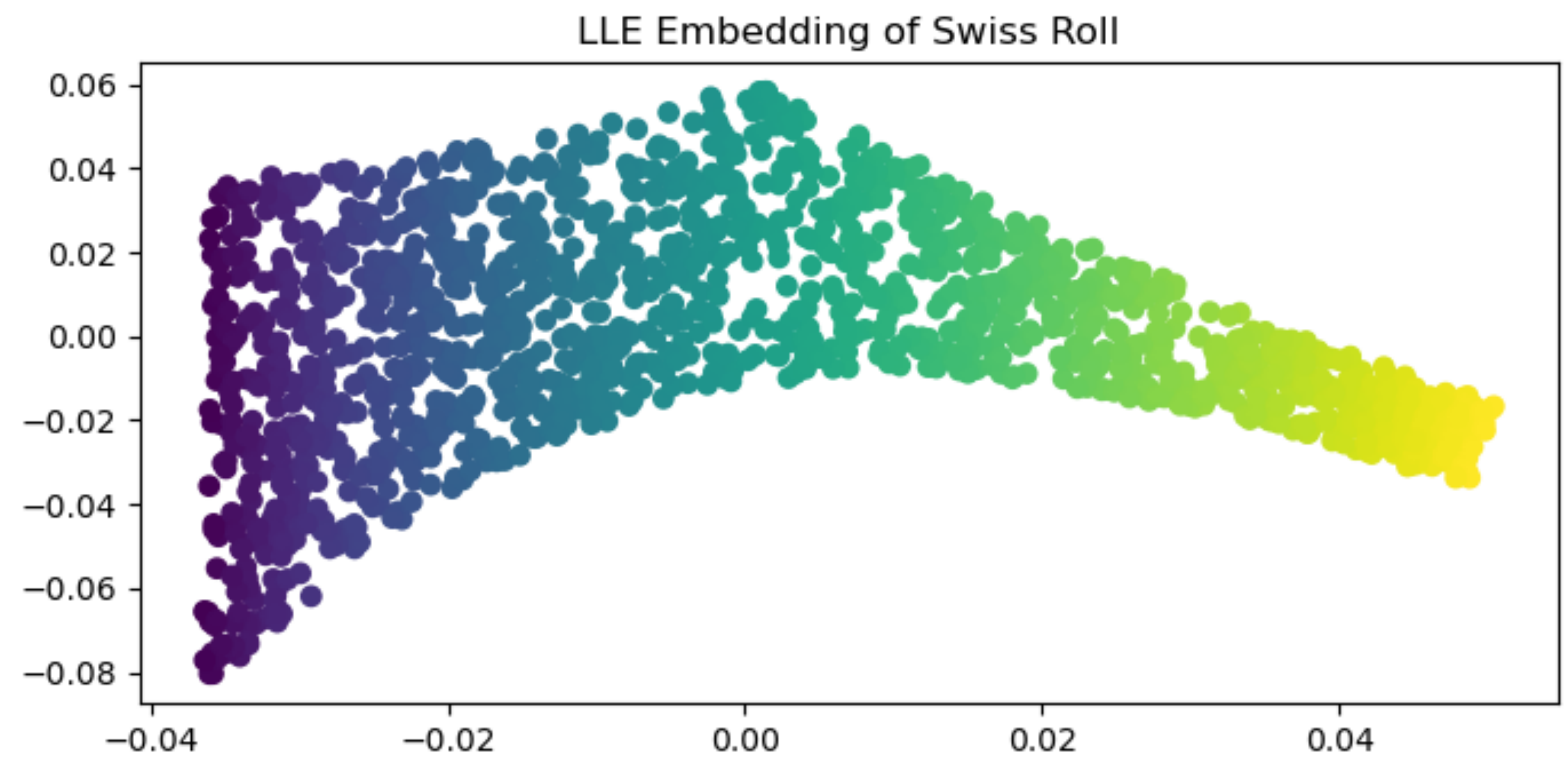


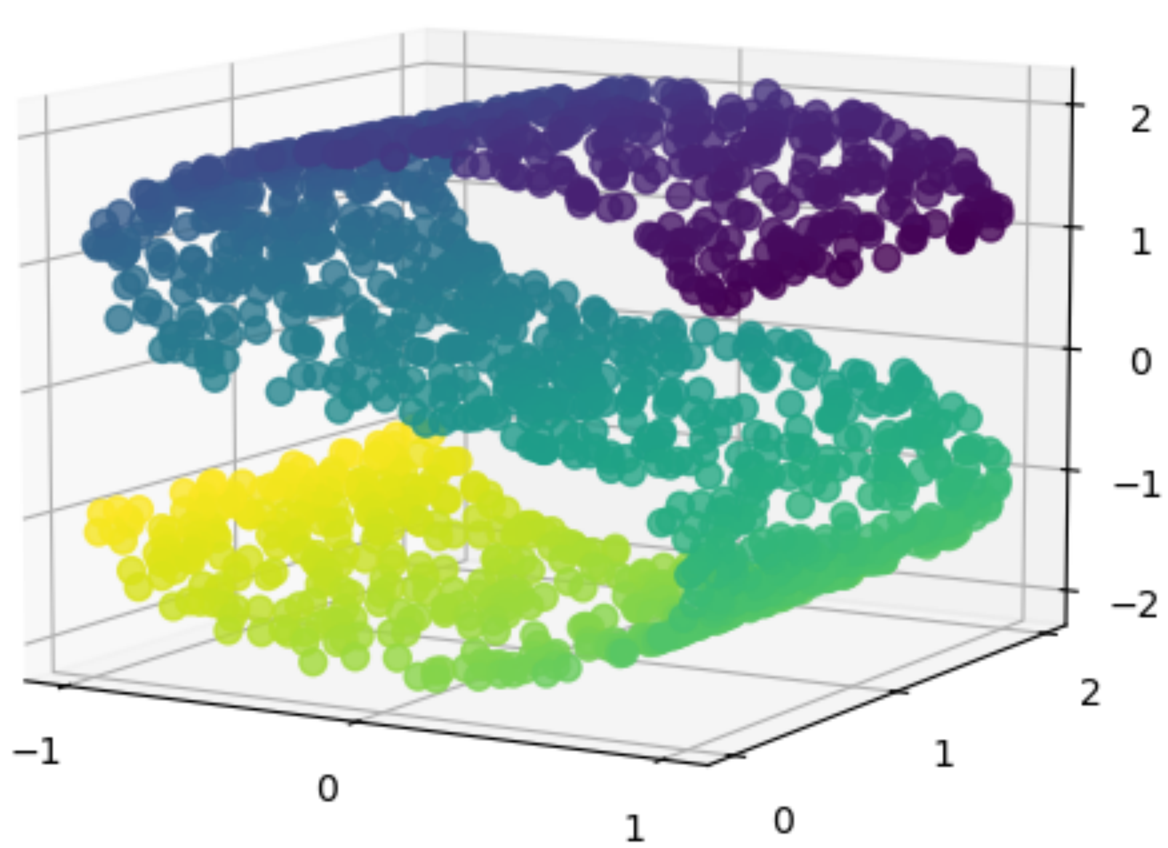
Ghojogh et al, 2020

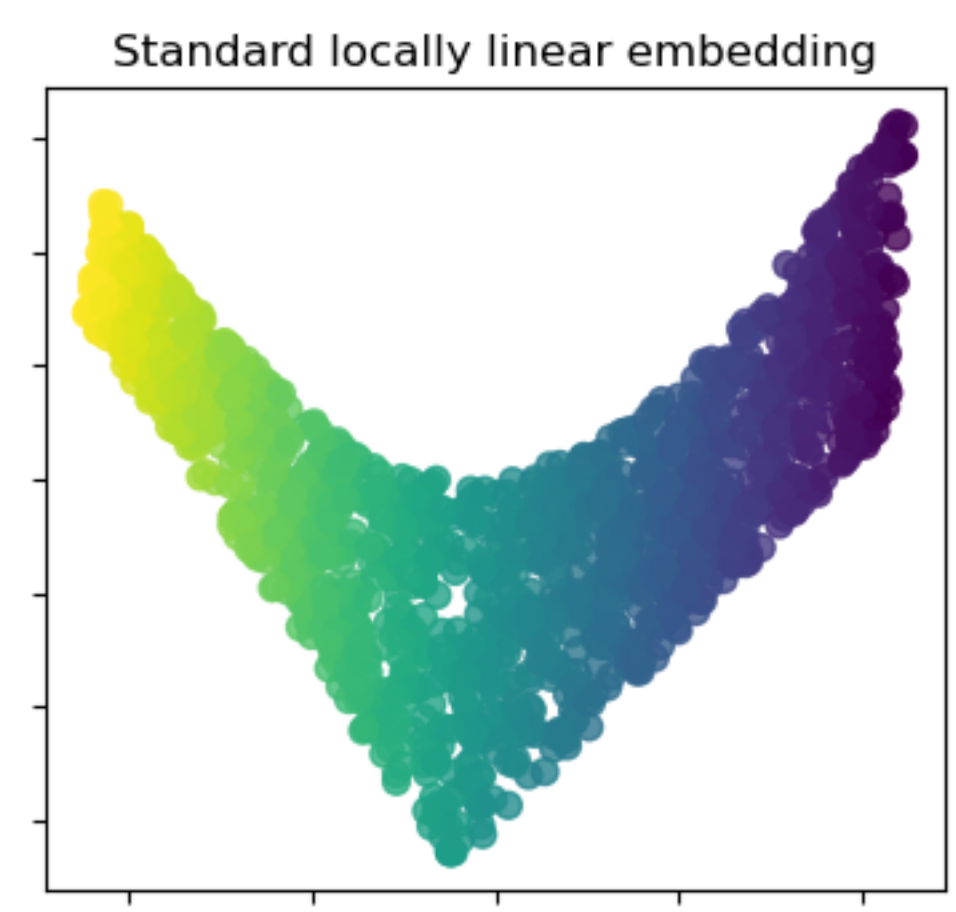
LLE

Locally Linear Embedding



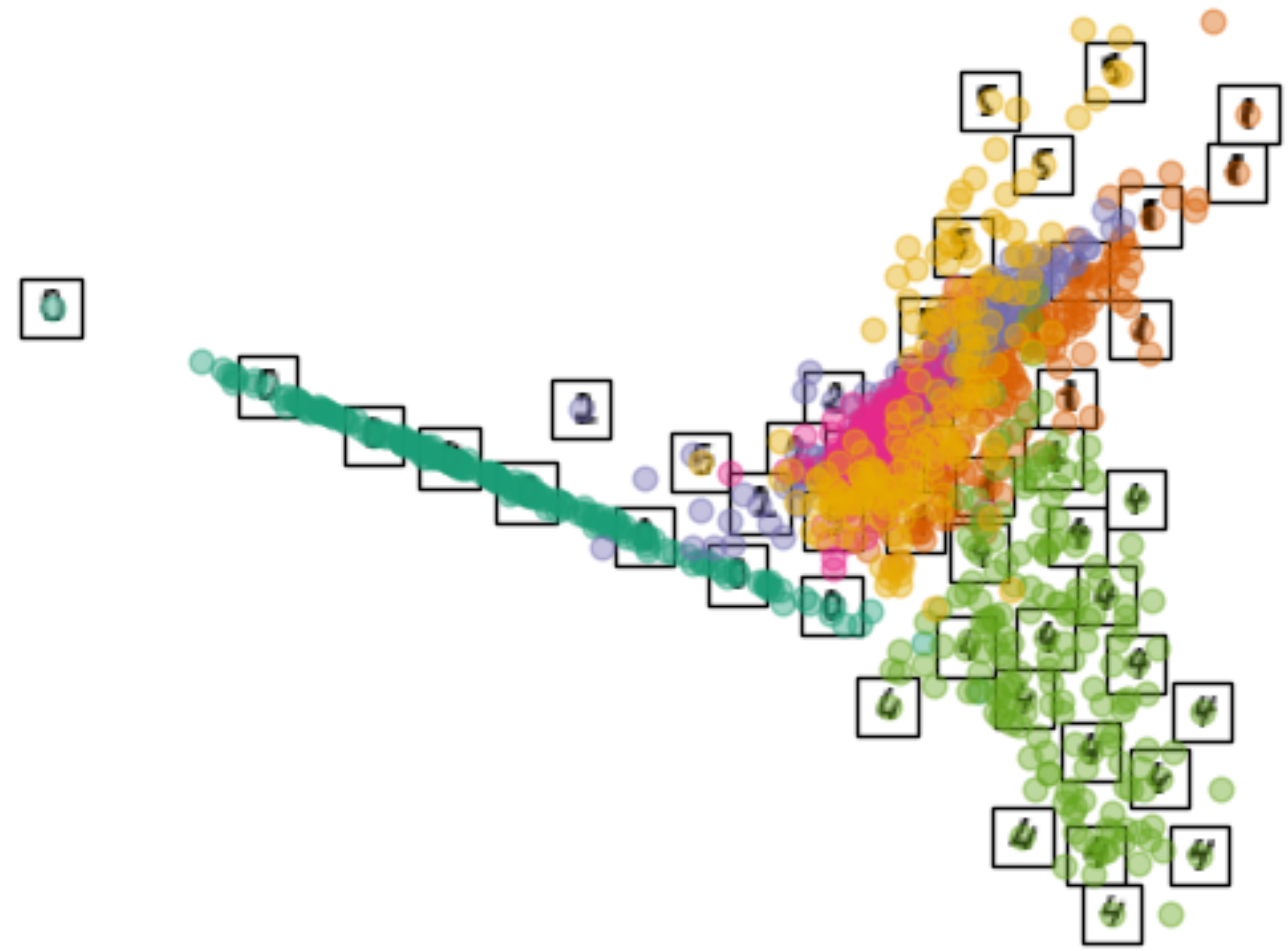






LLE

Locally Linear Embedding



LLE

Locally Linear Embedding

Non linear

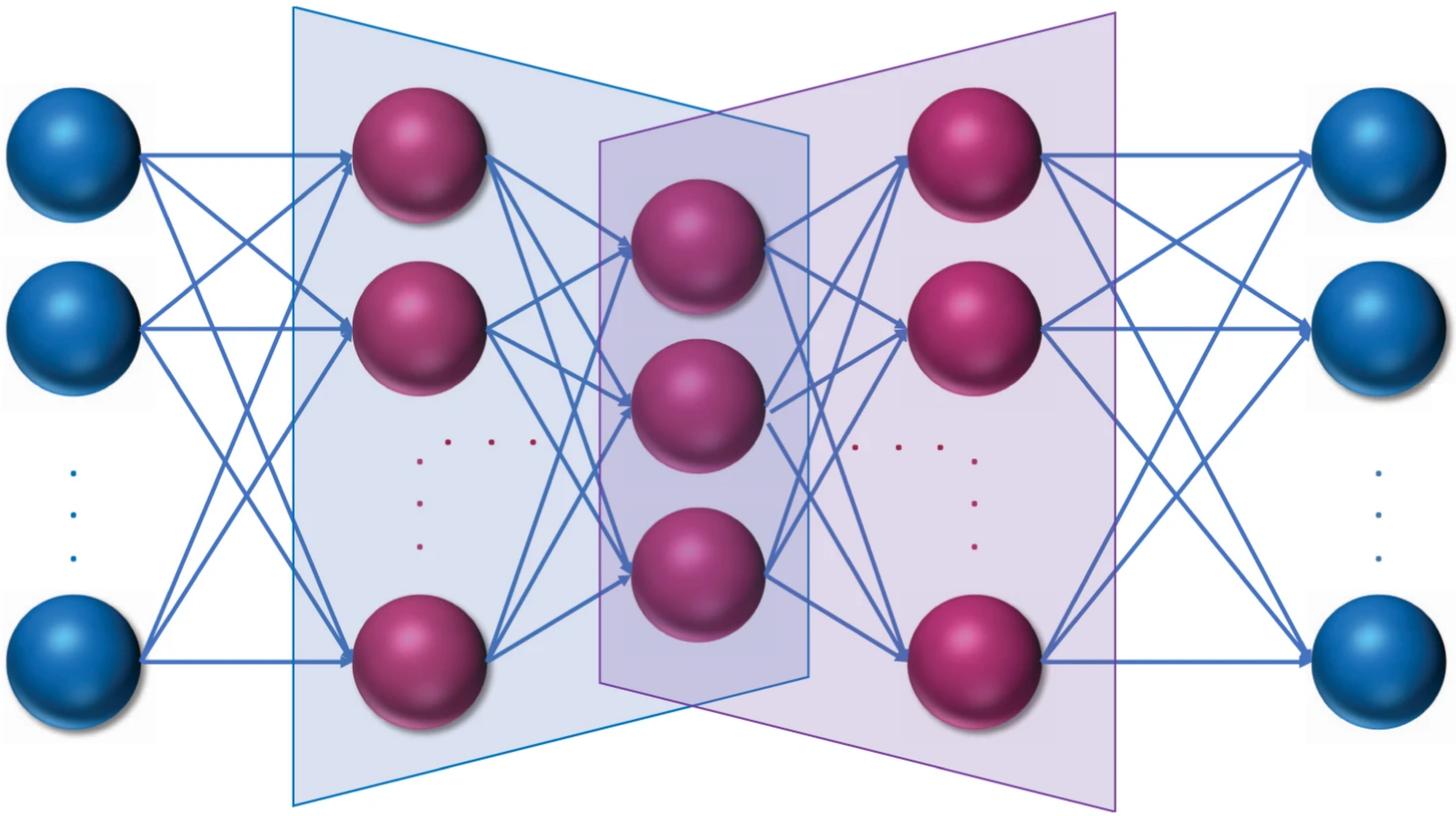
Assumption: manifold locally linear

Number of neighbours choice is critical

Autoencoders

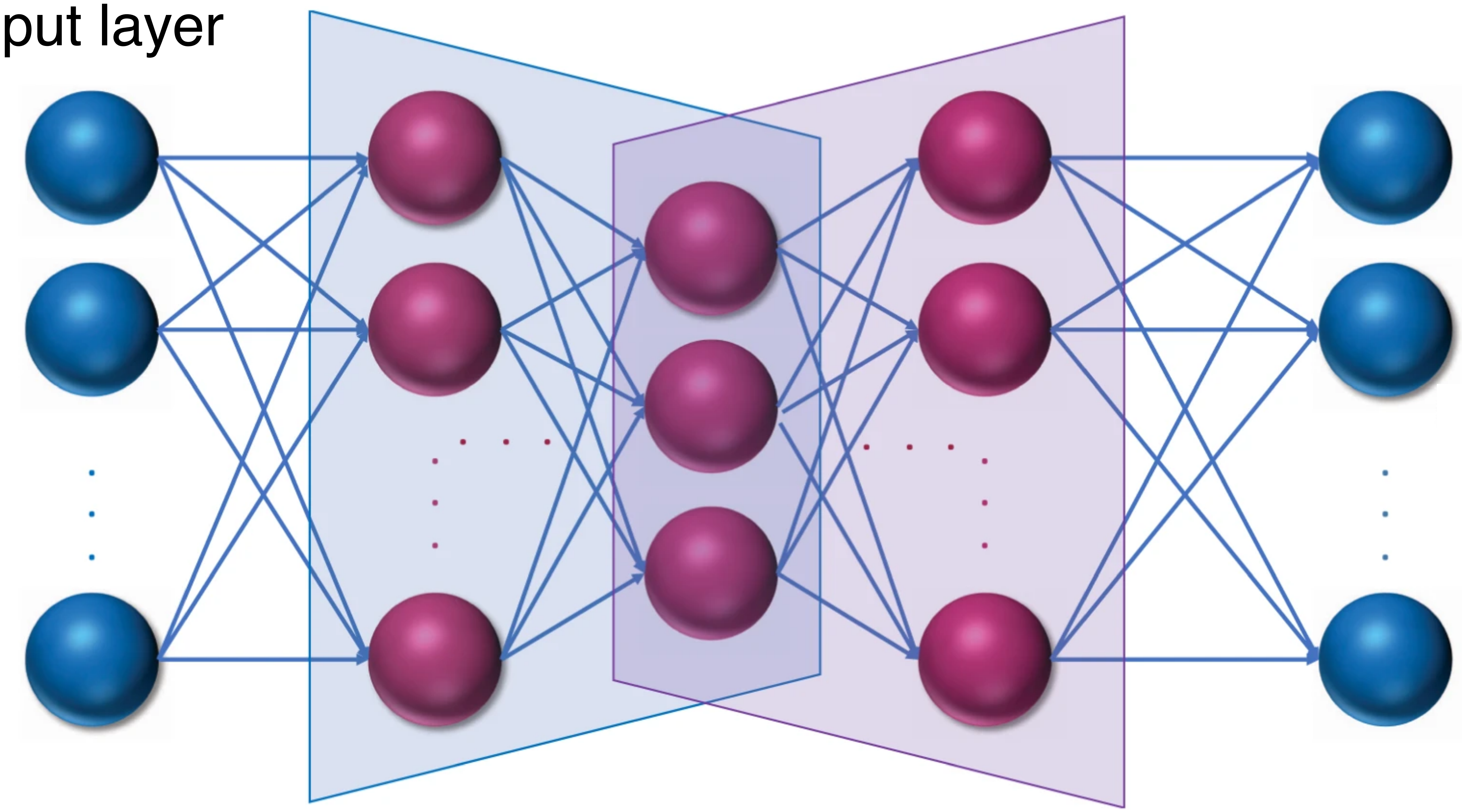
Deep Learning based DR

Aim: learn a compress representation of data



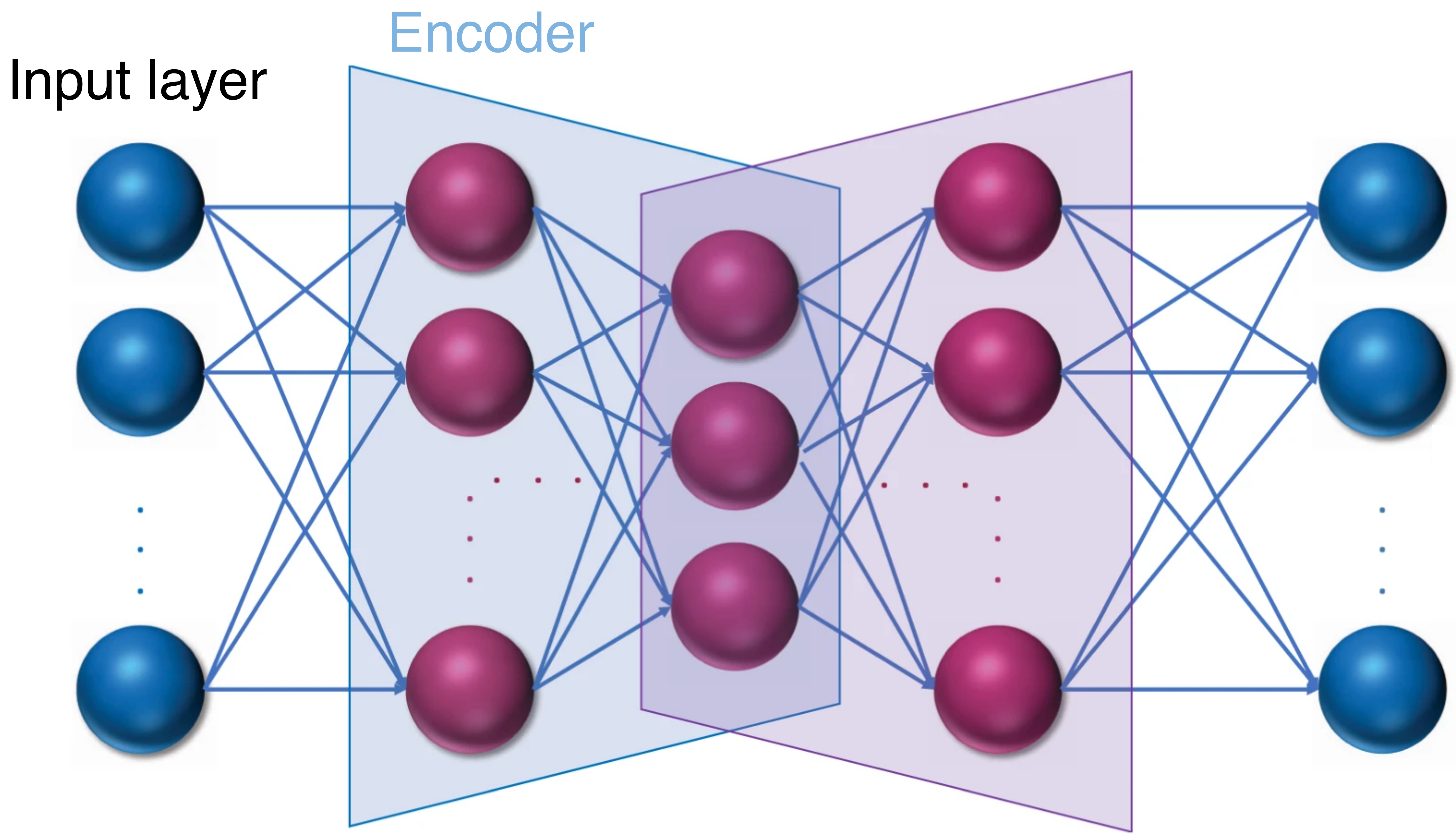
Kiarashinejad et al, 2020

Autoencoders



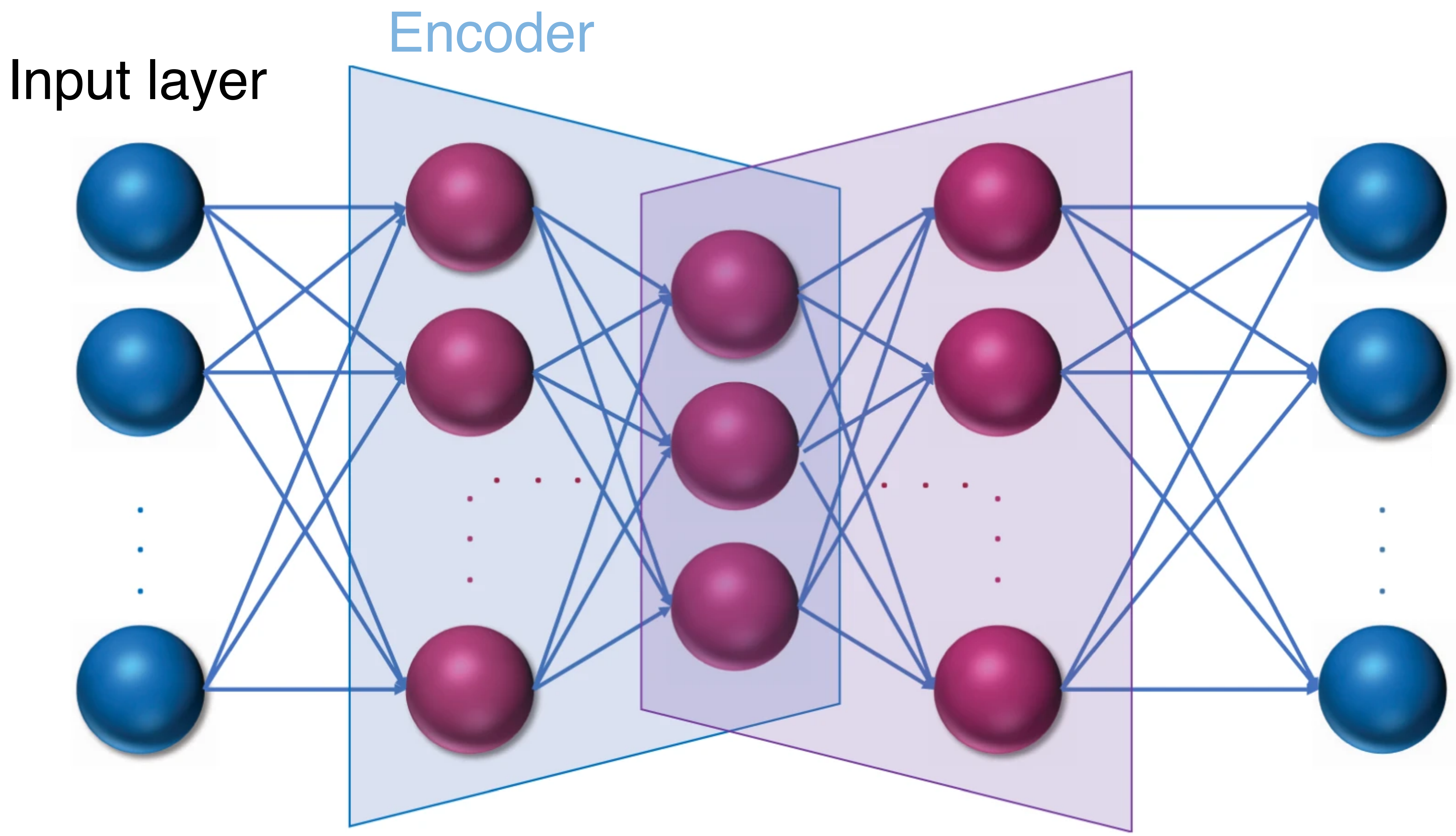
Input layer: how data enter the network. It could be high-dimensional vector, like a pixel array for image or a list of features for a dataset

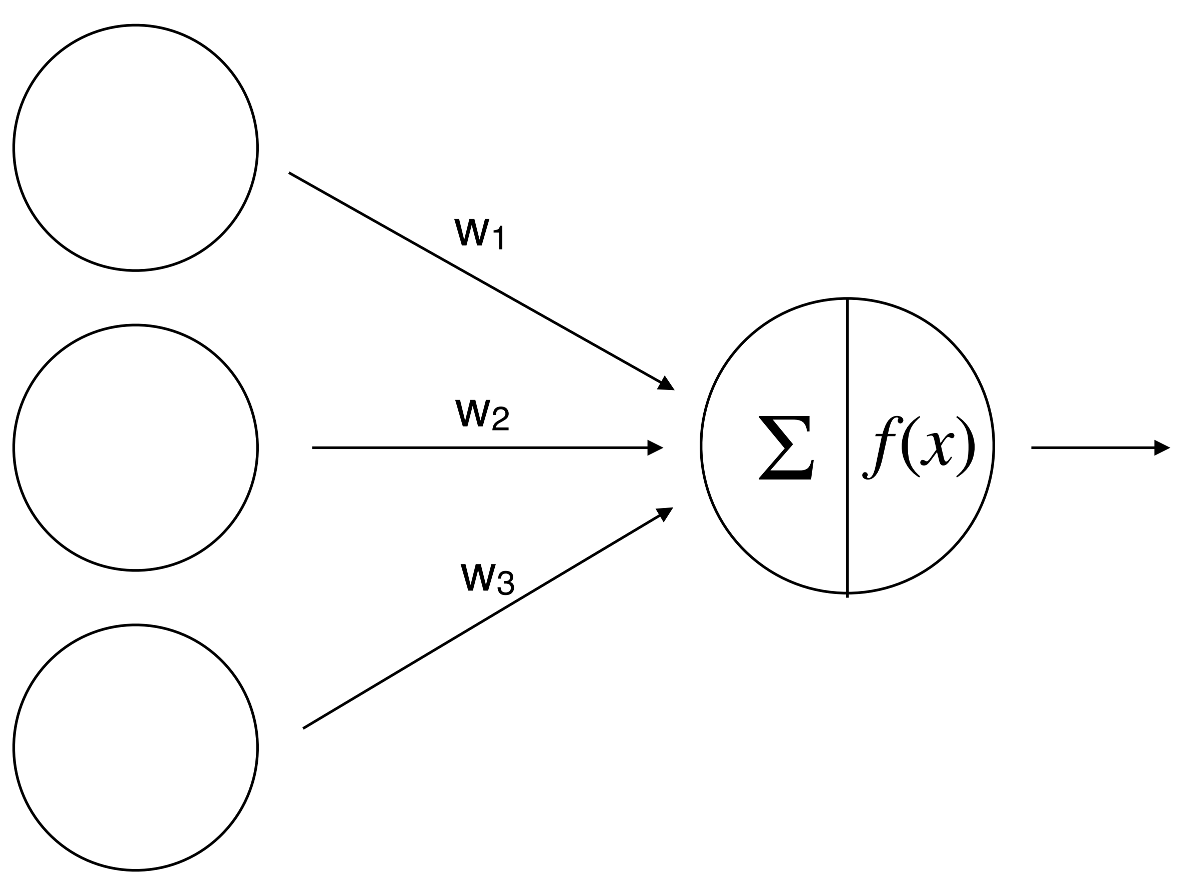
Autoencoders

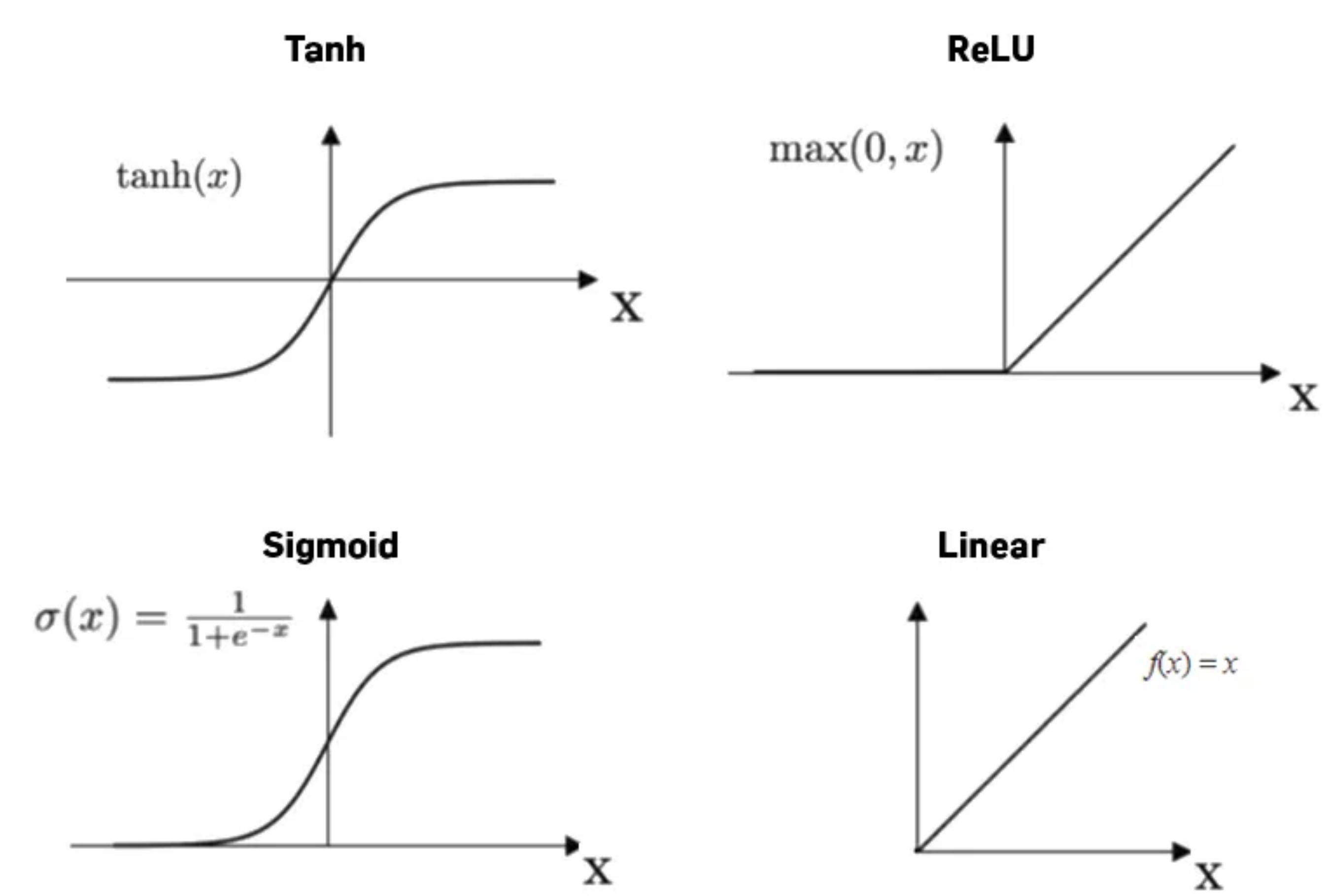


Encoder: set of layers in the autoencoder architecture responsible for compressing the dimensions of input space

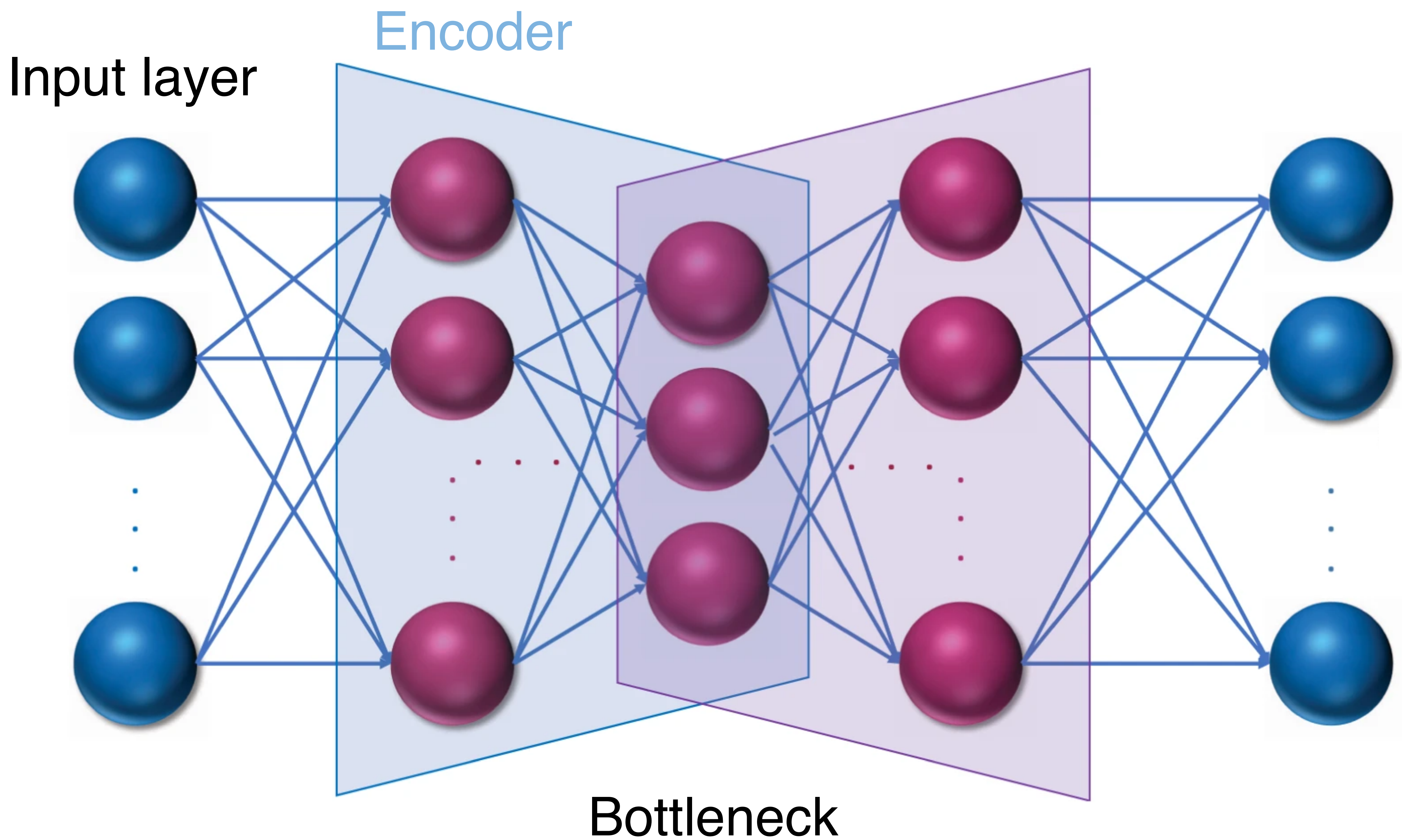
Autoencoders





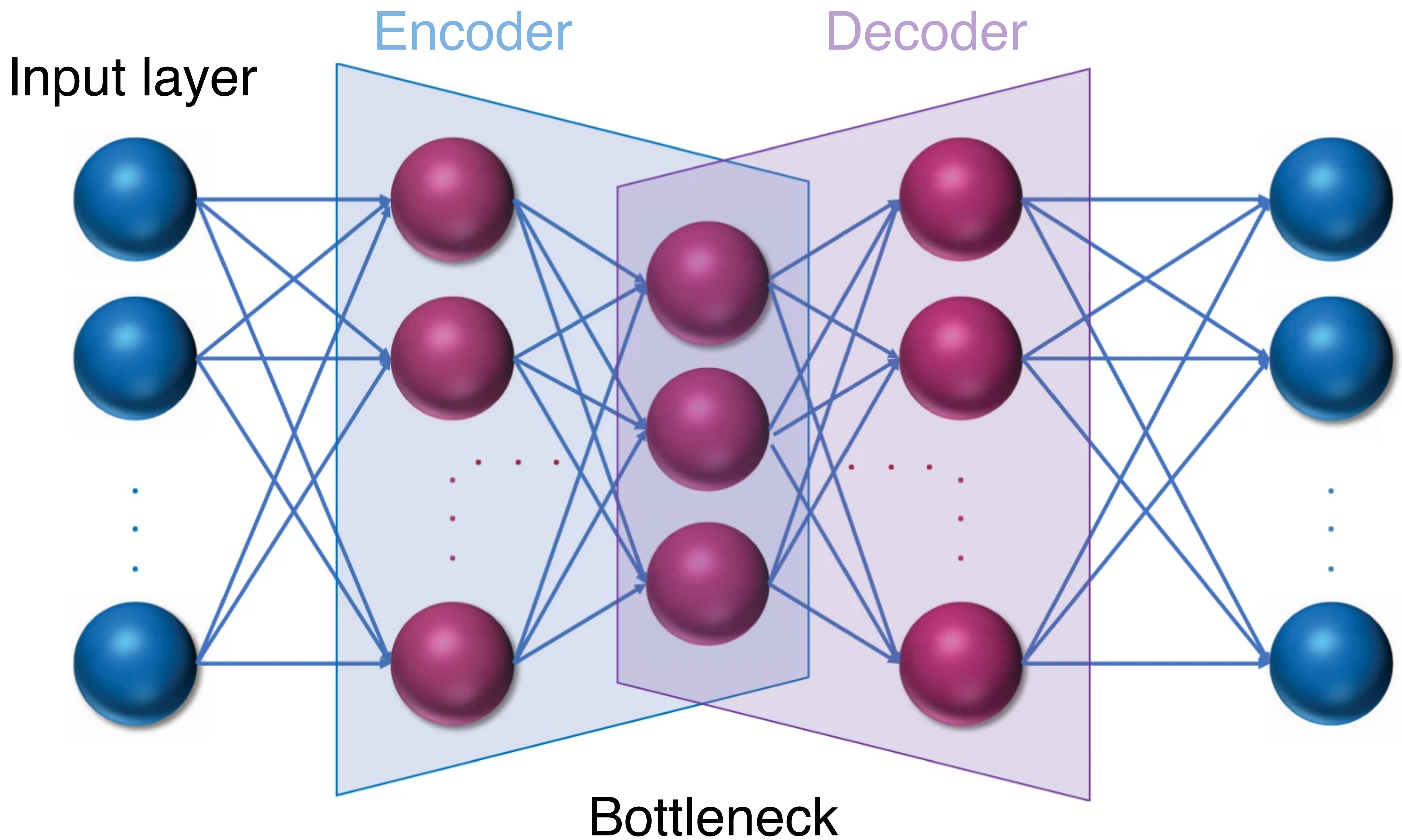


Autoencoders



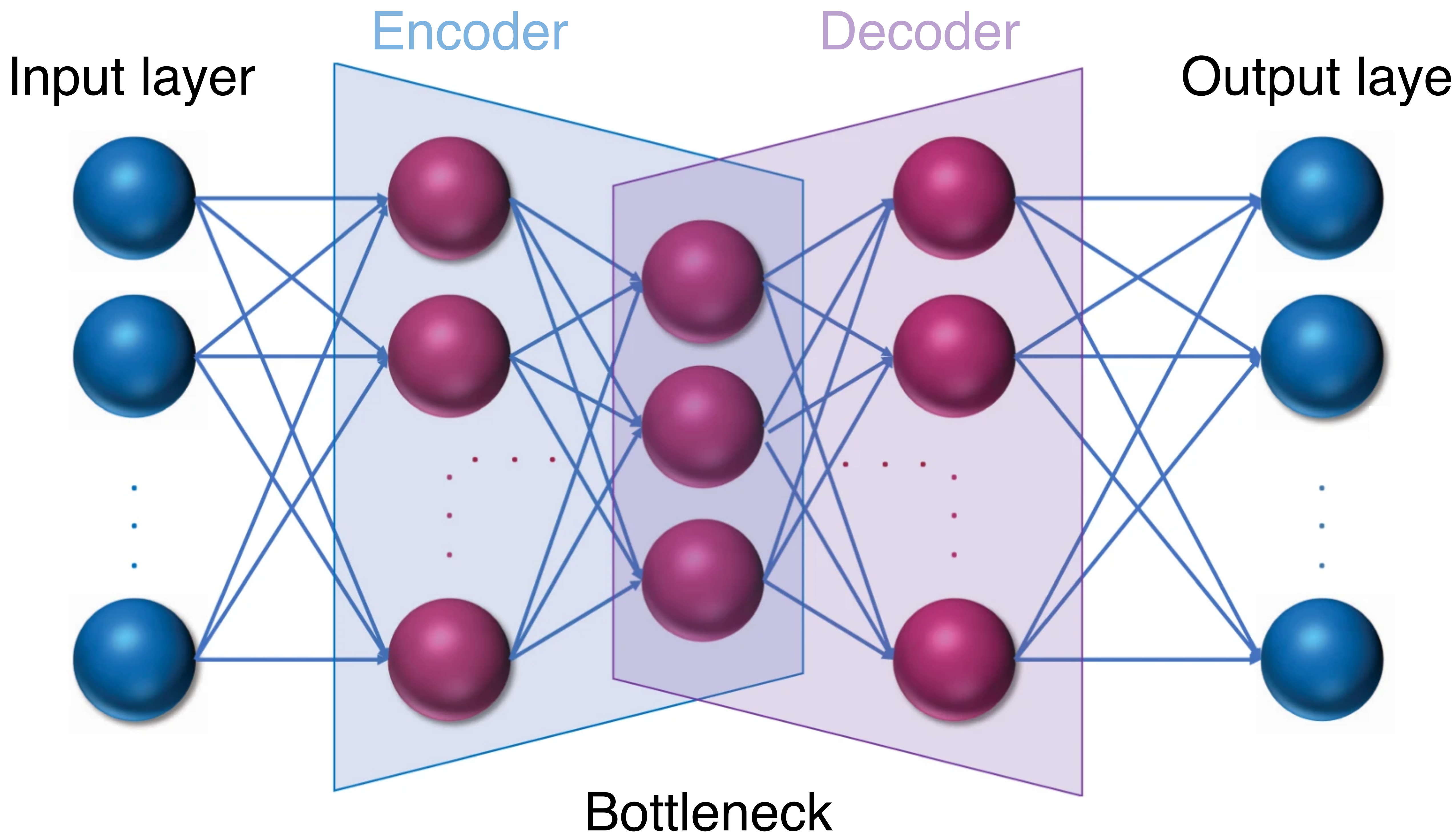
Bottleneck: the layer with number of nodes equal to the desired number of features to extract from an autoencoder model

Autoencoders



Decoder: set of layer in autoencoder retreiving information from low dimensional latent space to the output layer

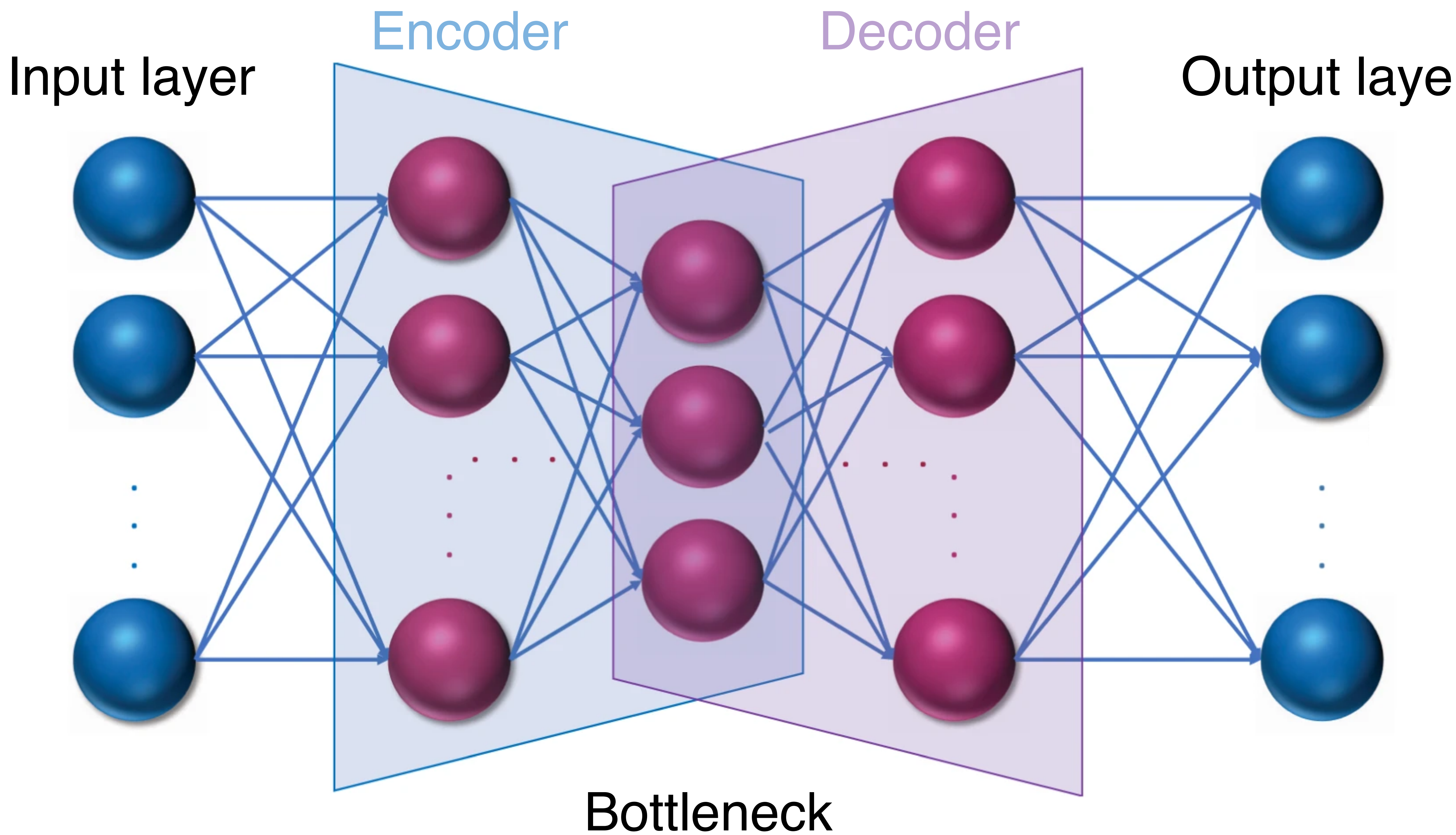
Autoencoders



Output layer: produces the final reconstructed data

Reconstruction error: MSE (mean squared error) or Cross-Entropy

Autoencoders



Training through backgropagation:

1. Forward pass
2. Loss calculation
3. Backpropagation through gradient descent

Autoencoders

“Hyperparameters”

Net size
- bottleneck size
- number of layers
- number of nodes per layer

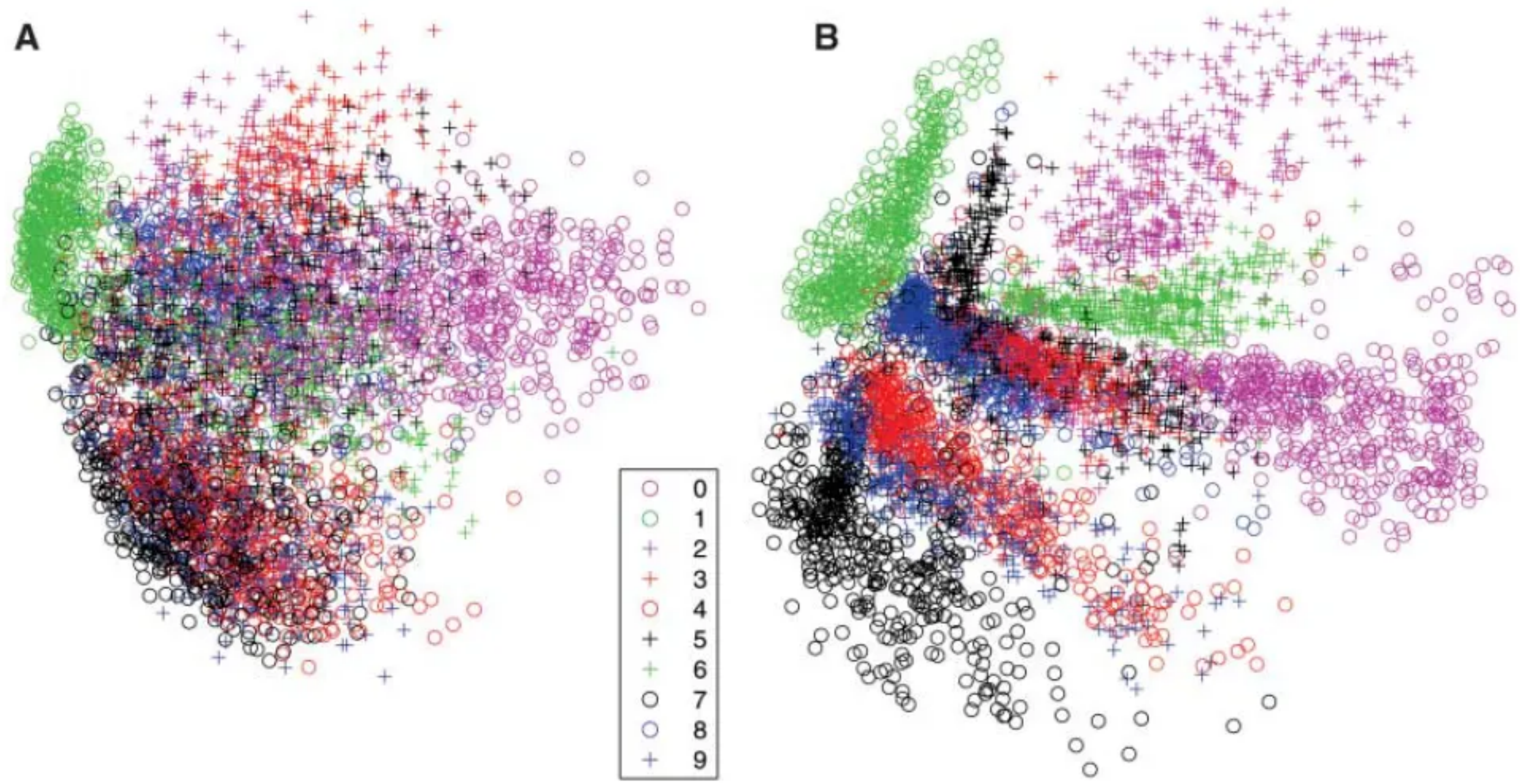
Activation function

Cost function

Learning parameters

Autoencoders

mnist



PCA

784-1000-500-250-2 autoencoder

Autoencoders

Non linear

Customizable architecture

Training complexity

Prone to overfitting

Latent space interpretability

Autoencoders

Are auto encoders preserving local or global structure?

Chemometrics and Intelligent Laboratory Systems 235 (2023) 104758

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Nonlocal, local and global preserving stacked autoencoder based fault detection method for nonlinear process monitoring

Jingchao Yang, Li Wang *

School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai, 200230, China

Cost function original AE

$$J_{AE}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\|x_i - \widehat{x}_i\|^2$$

Proposed cost function

Local structure

$$\min J_{NLGPAE}(\theta) = \min J_{AE}(\theta) + \lambda \frac{J_{LP}(\theta)}{J_{NLP}(\theta) + J_{GP}(\theta)}$$

Non local structure

Global structure

Accepted to the ICML 2022 Workshop on Topology, Algebra, and Geometry in Machine Learning

LOCAL DISTANCE PRESERVING AUTO-ENCODERS USING CONTINUOUS kNN GRAPHS

Nutan Chen, Patrick van der Smagt & Botond Cseke
Machine Learning Research Lab
Volkswagen Group, Germany
`{nutan.chen,botond.cseke}@volkswagen.de`

Neurocomputing 160 (2015) 250–260

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Laplacian Auto-Encoders: An explicit learning of nonlinear data manifold

Kui Jia [a,b,*], Lin Sun [c], Shenghua Gao [d], Zhan Song [e], Bertram E. Shi [c]

[a] *Department of Electrical and Computer Engineering, Faculty of Science and Technology, University of Macau, Macau SAR, China*
[b] *Advanced Digital Sciences Center, 08-10 Connexis North Tower, 1 Fusionopolis Way, Singapore 138632, Singapore*
[c] *Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR*
[d] *School of Information Science and Technology, ShanghaiTech University, Shanghai, China*
[e] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*

DR methods

A qualitative comparison

Method			Parameters	Metric
Principal Component Analysis	Linear	Global	-	y
Multidimensional scaling	Linear	Global	-	y

DR methods

A qualitative comparison

Method			Parameters	Metric
Principal Component Analysis	Linear	Global	-	y
Multidimensional scaling	Linear	Global	-	y
Isometric mapping	Non linear	Local	K nearest neighbours	y
Locally Linear Embedding	Non linear	Local	K nearest neighbours	y

DR methods

A qualitative comparison

Method			Parameters	Metric
Principal Component Analysis	Linear	Global	-	y
Multidimensional scaling	Linear	Global	-	y
Isometric mapping	Non linear	Local	K nearest neighbours	y
Locally Linear Embedding	Non linear	Local	K nearest neighbours	y
t-distributed Stochastic Neighbor Embedding	Non linear	Local	Perplexity	n

DR methods

A qualitative comparison

Method			Parameters	Metric
Principal Component Analysis	Linear	Global	-	y
Multidimensional scaling	Linear	Global	-	y
Isometric mapping	Non linear	Local	K nearest neighbours	y
Locally Linear Embedding	Non linear	Local	K nearest neighbours	y
t-distributed Stochastic Neighbor Embedding	Non linear	Local	Perplexity	n
Uniform Manifold Approximation and Projection	Non linear	Local	K-nearest neighbours Minimum distance	n

DR methods

A qualitative comparison

Method			Parameters	Metric
Principal Component Analysis	Linear	Global	-	y
Multidimensional scaling	Linear	Global	-	y
Isometric mapping	Non linear	Local	K nearest neighbours	y
Locally Linear Embedding	Non linear	Local	K nearest neighbours	y
t-distributed Stochastic Neighbor Embedding	Non linear	Local	Perplexity	n
Uniform Manifold Approximation and Projection	Non linear	Local	K-nearest neighbours Minimum distance	n
Autoencoders	Non linear	-	Net size	N

DR methods

A quantitative comparison

October 26, 2009

TiCC TR 2009–005

Dimensionality Reduction: A Comparative Review

Laurens van der Maaten **Eric Postma**

Jaap van den Herik
TiCC, Tilburg University

Journal of Machine Learning Research 22 (2021) 1-73 Submitted 9/20; Revised 4/21; Published 7/21

Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization

Yingfan Wang* YINGFAN.WANG@DUKE.EDU
Duke University, USA

Haiyang Huang* HAIYANG.HUANG@DUKE.EDU
Duke University, USA

Cynthia Rudin CYNTHIA@CS.DUKE.EDU
Duke University, USA

Yaron Shaposhnik YARON@SIMON.ROCHESTER.EDU
University of Rochester, USA

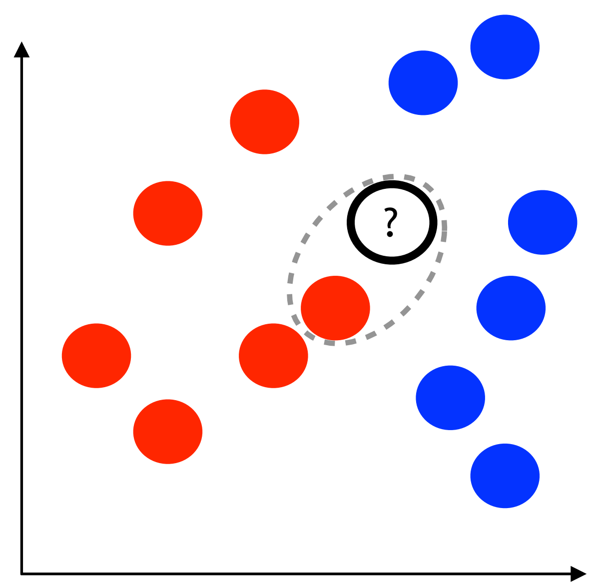
TriMap: Large-scale Dimensionality Reduction Using Triplets

Ehsan Amid & Manfred K. Warmuth
Google Research, Brain Team
{eamid, manfred}@google.com

DR methods

A quantitative comparison

- Local structure
  1-nearest neighbour generalisation error



- Continuity

Rank of low-dim j according to pairwise dist

$$C(k) = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in V_i^{(k)}} (\hat{r}(i,j) - k)$$

High-dim but not low-dim neighbours

- Trustworthiness

Rank of high-dim j according to pairwise dist

$$T(k) = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in U_i^{(k)}} (r(i,j) - k)$$

Low-dim but not high-dim neighbours

DR methods

A quantitative comparison

- Global structure
  Global score

Minimum Reconstruction Error $\mathcal{E}(\boldsymbol{Y} \mid \boldsymbol{X}) := \min_{\boldsymbol{A} \in \mathbb{R}^{m \times d}} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Y}\|_{\mathrm{F}}^2$

$$\mathrm{GS}(\boldsymbol{Y} \mid \boldsymbol{X}) := \exp\left(-\frac{\mathcal{E}(\boldsymbol{Y} \mid \boldsymbol{X}) - \mathcal{E}_{\mathrm{PCA}}}{\mathcal{E}_{\mathrm{PCA}}}\right) \in [0, 1]$$

- Random Triplet Accuracy
  Percentage of random triplets maintaining ordering in high and low dim

- Centroid Triplet Accuracy (labelled data)
  Percentage of class centroids triplets maintaining ordering in high and low dim

DR methods

A quantitative comparison

October 26, 2009

TiCC TR 2009–005

Dimensionality Reduction: A Comparative Review

Laurens van der Maaten **Eric Postma**

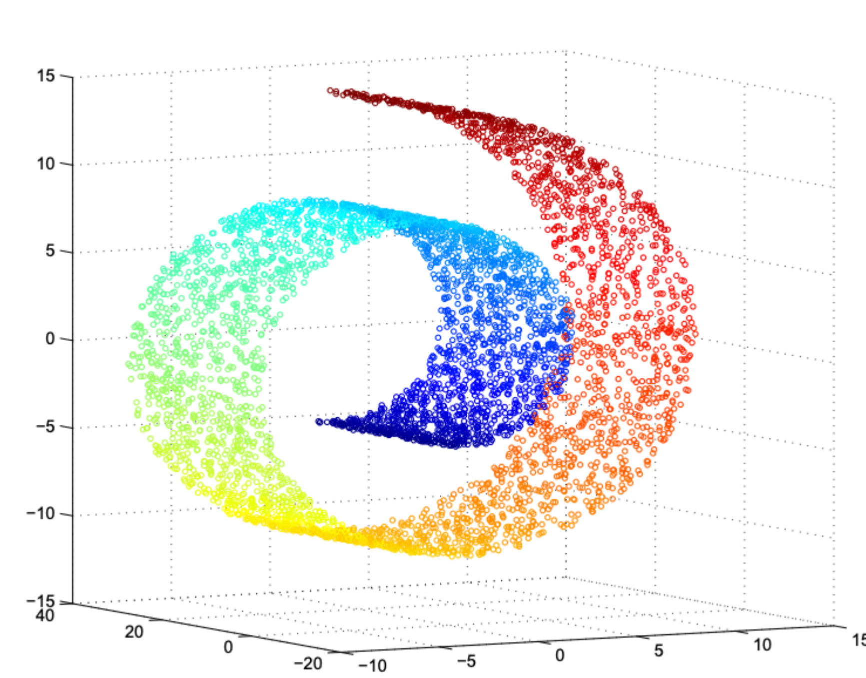
Jaap van den Herik
TiCC, Tilburg University

DR methods

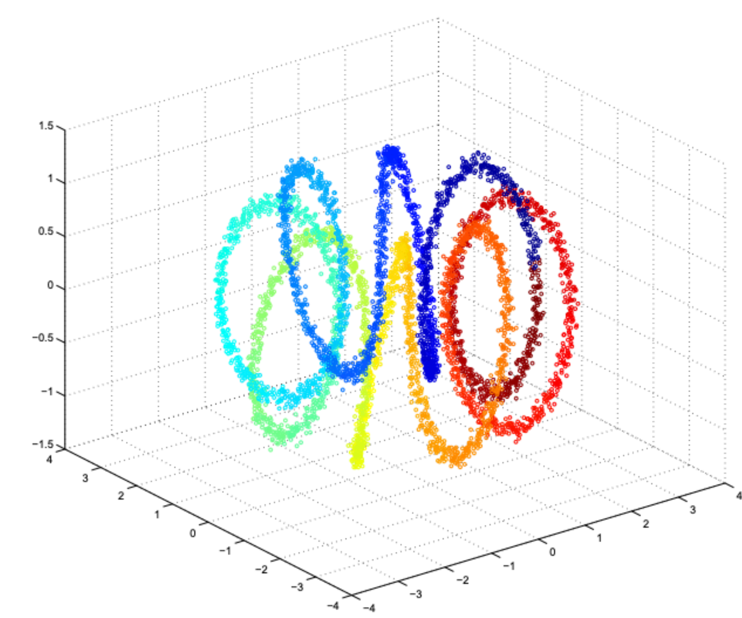
A quantitative comparison

Artificial datasets (n=5000 samples)

Low-dim manifold isometric to Euclidean space



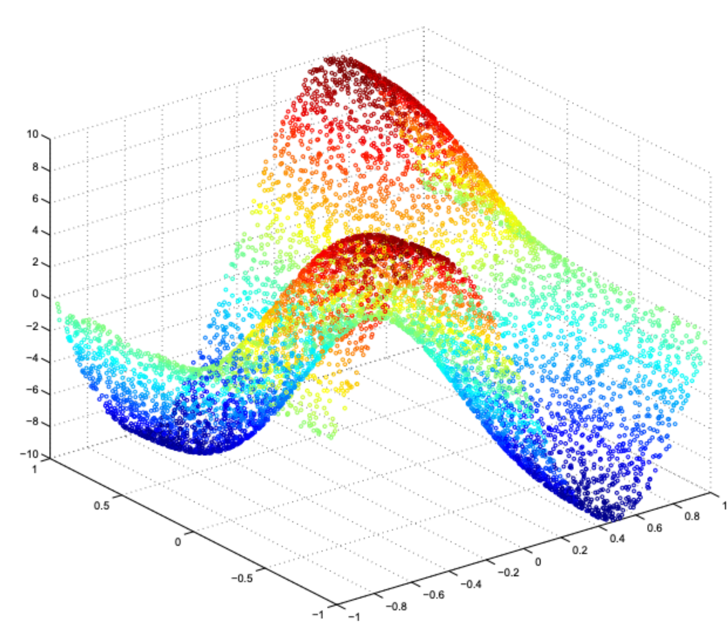
(a) Swiss roll dataset.



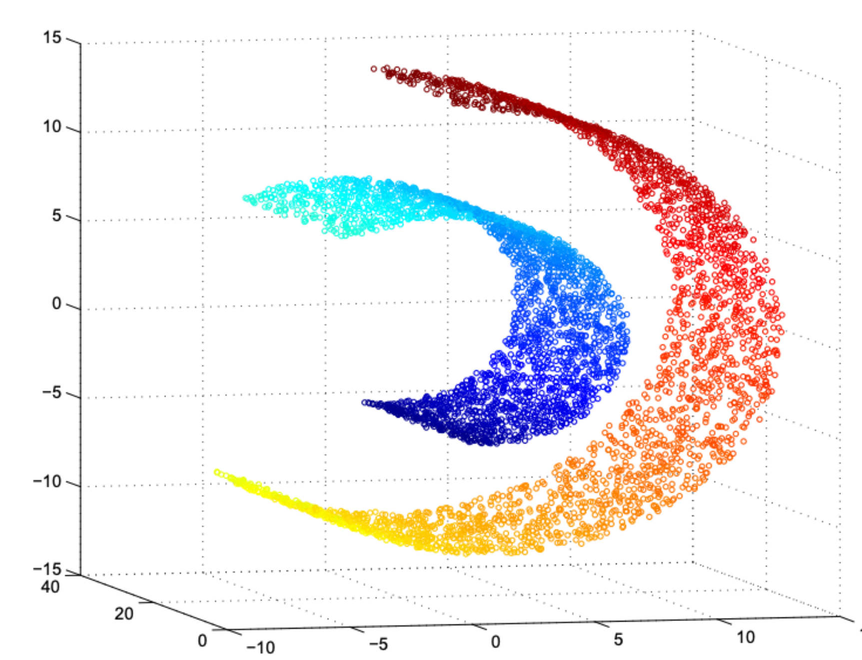
(b) Helix dataset.

Low-dim manifold not isometric to Euclidean space

Low-dim manifold not isometric to Euclidean space



(c) Twinpeaks dataset.



(d) Broken Swiss roll dataset.

Disconnected manifold

+ HD artificial dataset (5-dim manifold in 10-dim space)

DR methods

A quantitative comparison

Natural datasets

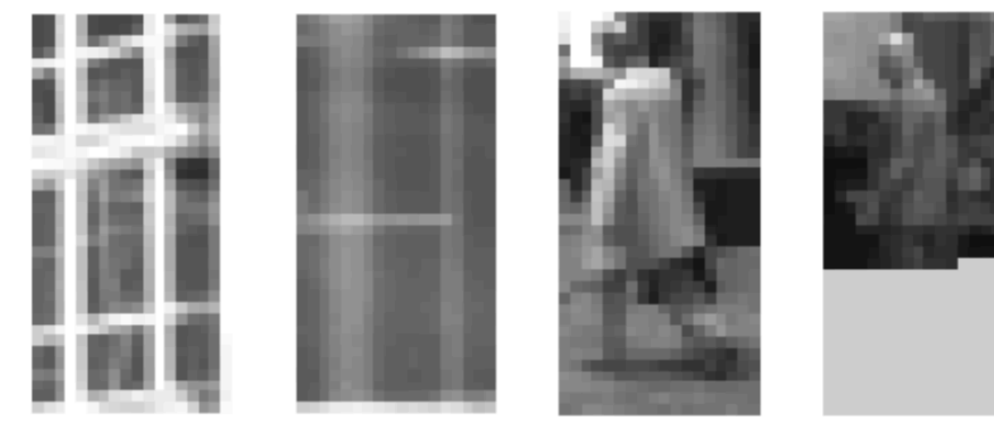
MNIST

5000 images
28 x 28 pixels (D=784)
10 classes

COIL20

1440 images
32 x 32 pixels (D=1024)
20 classes

NiSIS dataset



3675 images
36 x 18 pixels (D=648)

ORL dataset

400 images
112 x 92 pixels (D=10304)
40 faces

HIVA dataset

Drug discovery (active and non active components against HIV/AIDS)
3845 datapoints
D=1617
2 classes

DR methods

A quantitative comparison

Artificial datasets

Generalisation error 1-NN (↓)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2	3.68%	29.76%	**3.40%**	3.74%	22.34%	49.00%

Continuity C(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2		**1.00**	0.99	**1.00**	**1.00**	0.50

Trustworthiness T(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2		0.88	0.99	**1.00**	0.89	0.46

DR methods

A quantitative comparison

Artificial datasets

Generalisation error 1-NN (↓)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2	3.68%	29.76%	**3.40%**	3.74%	22.34%	49.00%
Helix	1	1.24%	35.50%	**13.18%**	32.32%	52.22%	52.22%
Twin Peaks	2	0.40%	0.26%	**0.22%**	0.94%	0.32%	49.06%

Continuity C(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2		**1.00**	0.99	**1.00**	**1.00**	0.50
Helix	1		0.98	0.97	**0.99**	0.50	0.75
Twin Peaks	2		**1.00**	0.99	0.99	**1.00**	0.50

Trustworthiness T(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2		0.88	0.99	**1.00**	0.89	0.46
Helix	1		0.78	0.74	**0.83**	0.35	0.64
Twin Peaks	2		0.98	0.98	0.99	**1.00**	0.52

DR methods

A quantitative comparison

Artificial datasets

Generalisation error 1-NN (↓)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2	3.68%	29.76%	**3.40%**	3.74%	22.34%	49.00%
Helix	1	1.24%	35.50%	**13.18%**	32.32%	52.22%	52.22%
Twin Peaks	2	0.40%	0.26%	**0.22%**	0.94%	0.32%	49.06%
Broken Swiss	2	2.14%	25.96%	**14.48%**	36.94%	27.40%	87.86%

Continuity C(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2		**1.00**	0.99	**1.00**	**1.00**	0.50
Helix	1		0.98	0.97	**0.99**	0.50	0.75
Twin Peaks	2		**1.00**	0.99	0.99	**1.00**	0.50
Broken Swiss	2		**1.00**	0.98	0.98	**1.00**	0.73

Trustworthiness T(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2		0.88	0.99	**1.00**	0.89	0.46
Helix	1		0.78	0.74	**0.83**	0.35	0.64
Twin Peaks	2		0.98	0.98	0.99	**1.00**	0.52
Broken Swiss	2		0.96	**0.97**	0.94	**0.97**	0.70

DR methods

A quantitative comparison

Artificial datasets

Generalisation error 1-NN (↓)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2	3.68%	29.76%	**3.40%**	3.74%	22.34%	49.00%
Helix	1	1.24%	35.50%	**13.18%**	32.32%	52.22%	52.22%
Twin Peaks	2	0.40%	0.26%	**0.22%**	0.94%	0.32%	49.06%
Broken Swiss	2	2.14%	25.96%	**14.48%**	36.94%	27.40%	87.86%
HD	5	24.19%	22.18%	23.26%	20.74%	**20.70%**	48.18%

Continuity C(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2		**1.00**	0.99	**1.00**	**1.00**	0.50
Helix	1		0.98	0.97	**0.99**	0.50	0.75
Twin Peaks	2		**1.00**	0.99	0.99	**1.00**	0.50
Broken Swiss	2		**1.00**	0.98	0.98	**1.00**	0.73
HD	5		**1.00**	0.98	0.99	**1.00**	0.89

Trustworthiness T(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
Swiss Roll	2		0.88	0.99	**1.00**	0.89	0.46
Helix	1		0.78	0.74	**0.83**	0.35	0.64
Twin Peaks	2		0.98	0.98	0.99	**1.00**	0.52
Broken Swiss	2		0.96	**0.97**	0.94	**0.97**	0.70
HD	5		**1.00**	0.98	**1.00**	**1.00**	0.68

DR methods

A quantitative comparison

Natural datasets

Generalisation error 1-NN (↓)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20	5.11%	**6.74%**	12.64%	10.02%	6.90%	7.18%

Continuity C(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.94	0.96	**1.00**	**1.00**

Trustworthiness T(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.96	0.96	**1.00**	**1.00**

DR methods

A quantitative comparison

Natural datasets

Generalisation error 1-NN (↓)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20	5.11%	**6.74%**	12.64%	10.02%	6.90%	7.18%
COIL20	5	0.14%	3.82%	15.69%	22.29%	**0.83%**	51.11%

Continuity C(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.94	0.96	**1.00**	**1.00**
COIL20	5		**1.00**	0.90	0.95	**1.00**	0.92

Trustworthiness T(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.96	0.96	**1.00**	**1.00**
COIL20	5		**0.99**	0.89	0.93	**0.99**	0.88

DR methods

A quantitative comparison

Natural datasets

Generalisation error 1-NN (↓)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20	5.11%	**6.74%**	12.64%	10.02%	6.90%	7.18%
COIL20	5	0.14%	3.82%	15.69%	22.29%	**0.83%**	51.11%
ORL	8	2.50%	4.75%	27.50%	11.00%	**2.75%**	6.25%

Continuity C(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.94	0.96	**1.00**	**1.00**
COIL20	5		**1.00**	0.90	0.95	**1.00**	0.92
ORL	8		**0.99**	0.76	0.95	**0.99**	0.98

Trustworthiness T(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.96	0.96	**1.00**	**1.00**
COIL20	5		**0.99**	0.89	0.93	**0.99**	0.88
ORL	8		**0.99**	0.78	0.95	**0.99**	**0.99**

DR methods

A quantitative comparison

Natural datasets

Generalisation error 1-NN (↓)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20	5.11%	**6.74%**	12.64%	10.02%	6.90%	7.18%
COIL20	5	0.14%	3.82%	15.69%	22.29%	**0.83%**	51.11%
ORL	8	2.50%	4.75%	27.50%	11.00%	**2.75%**	6.25%
NiSIS	15	8.24%	**7.95%**	13.36%	15.48%	48.98%	9.22%

Continuity C(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.94	0.96	**1.00**	**1.00**
COIL20	5		**1.00**	0.90	0.95	**1.00**	0.92
ORL	8		**0.99**	0.76	0.95	**0.99**	0.98
NiSIS	15		**1.00**	0.84	0.91	0.47	**1.00**

Trustworthiness T(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.96	0.96	**1.00**	**1.00**
COIL20	5		**0.99**	0.89	0.93	**0.99**	0.88
ORL	8		**0.99**	0.78	0.95	**0.99**	**0.99**
NiSIS	15		**0.99**	0.89	0.92	0.47	**0.99**

DR methods

A quantitative comparison

Natural datasets

Generalisation error 1-NN (↓)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20	5.11%	**6.74%**	12.64%	10.02%	6.90%	7.18%
COIL20	5	0.14%	3.82%	15.69%	22.29%	**0.83%**	51.11%
ORL	8	2.50%	4.75%	27.50%	11.00%	**2.75%**	6.25%
NiSIS	15	8.24%	**7.95%**	13.36%	15.48%	48.98%	9.22%
HIVA	15	4.63%	5.05%	4.92%	4.97%	**3.51%**	5.12%

Continuity C(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.94	0.96	**1.00**	**1.00**
COIL20	5		**1.00**	0.90	0.95	**1.00**	0.92
ORL	8		**0.99**	0.76	0.95	**0.99**	0.98
NiSIS	15		**1.00**	0.84	0.91	0.47	**1.00**
HIVA	15		**0.99**	0.84	0.80	0.51	**0.99**

Trustworthiness T(12) (↑)

Dataset	Target dim	None	PCA	Isomap	LLE	Sammon	Autoencoder
MNIST	20		**1.00**	0.96	0.96	**1.00**	**1.00**
COIL20	5		**0.99**	0.89	0.93	**0.99**	0.88
ORL	8		**0.99**	0.78	0.95	**0.99**	**0.99**
NiSIS	15		**0.99**	0.89	0.92	0.47	**0.99**
HIVA	15		0.97	0.87	0.80	0.42	**0.98**

DR methods

A quantitative comparison

TriMap: Large-scale Dimensionality Reduction Using Triplets

Ehsan Amid & Manfred K. Warmuth
Google Research, Brain Team
{eamid, manfred}@google.com

Journal of Machine Learning Research 22 (2021) 1-73 Submitted 9/20; Revised 4/21; Published 7/21

Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization

Yingfan Wang* YINGFAN.WANG@DUKE.EDU
Duke University, USA

Haiyang Huang* HAIYANG.HUANG@DUKE.EDU
Duke University, USA

Cynthia Rudin CYNTHIA@CS.DUKE.EDU
Duke University, USA

Yaron Shaposhnik YARON@SIMON.ROCHESTER.EDU
University of Rochester, USA

DR methods

A quantitative comparison

Datasets

Olivetti Faces

COIL20

COIL100

S-curve with a hole

Mammoth

USPS

MNIST

Mouse scRNA-seq

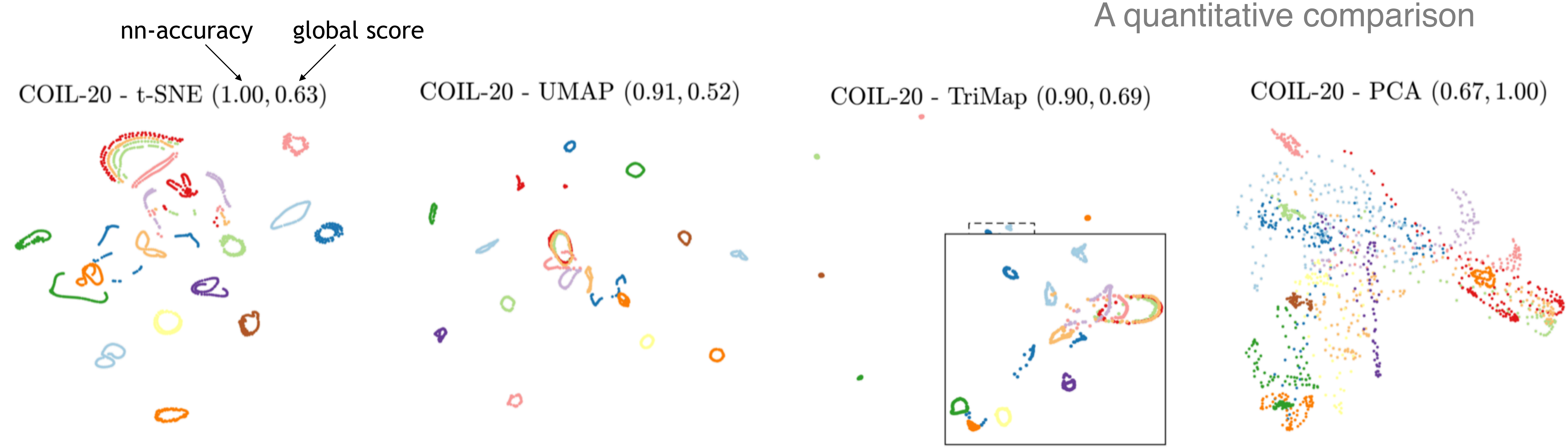
20 NewsGroups

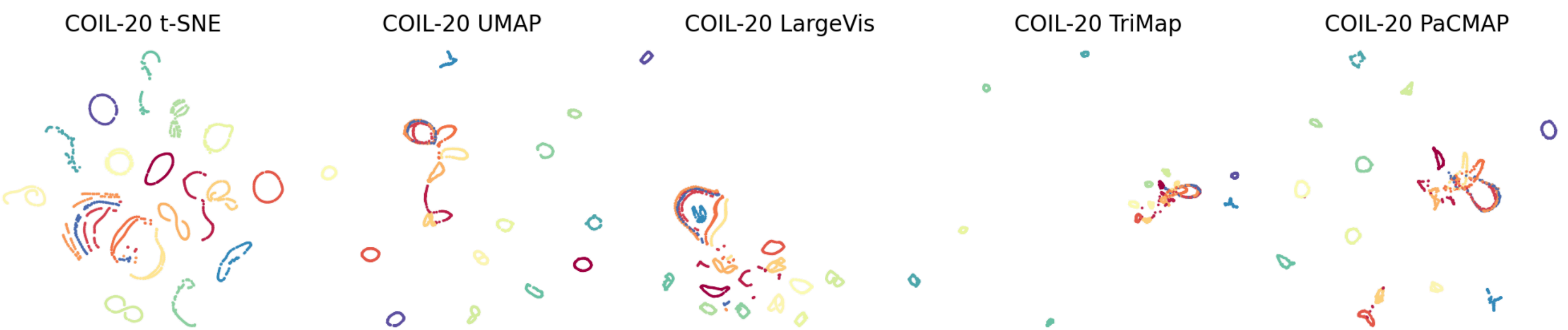
Flow Cytometry

KDD Cup 99

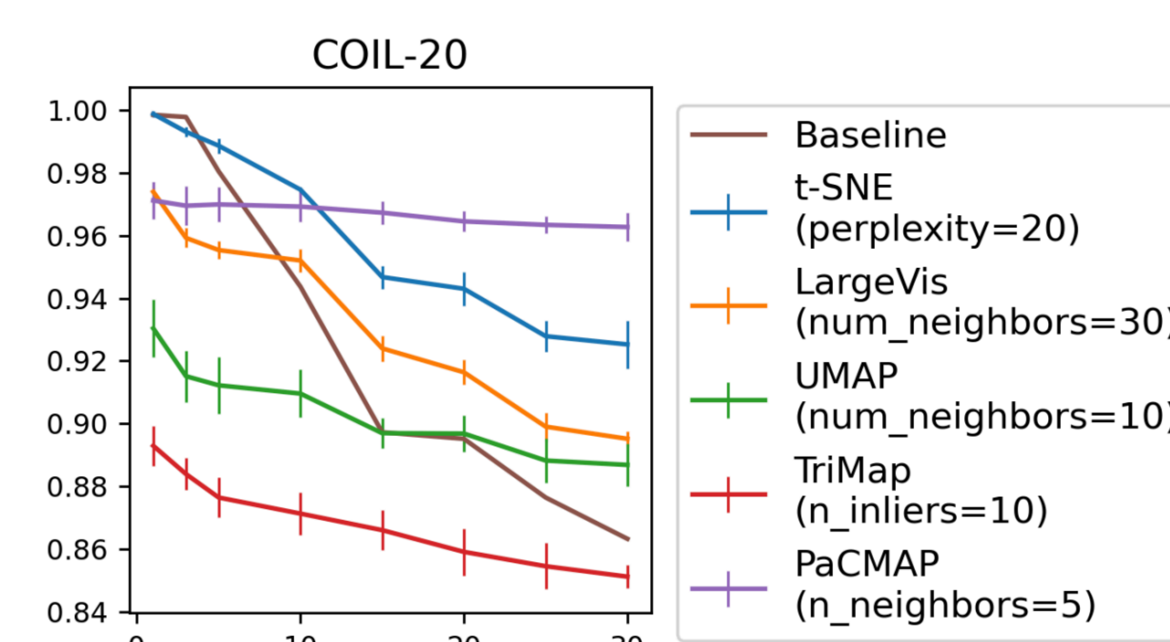
DR methods

A quantitative comparison



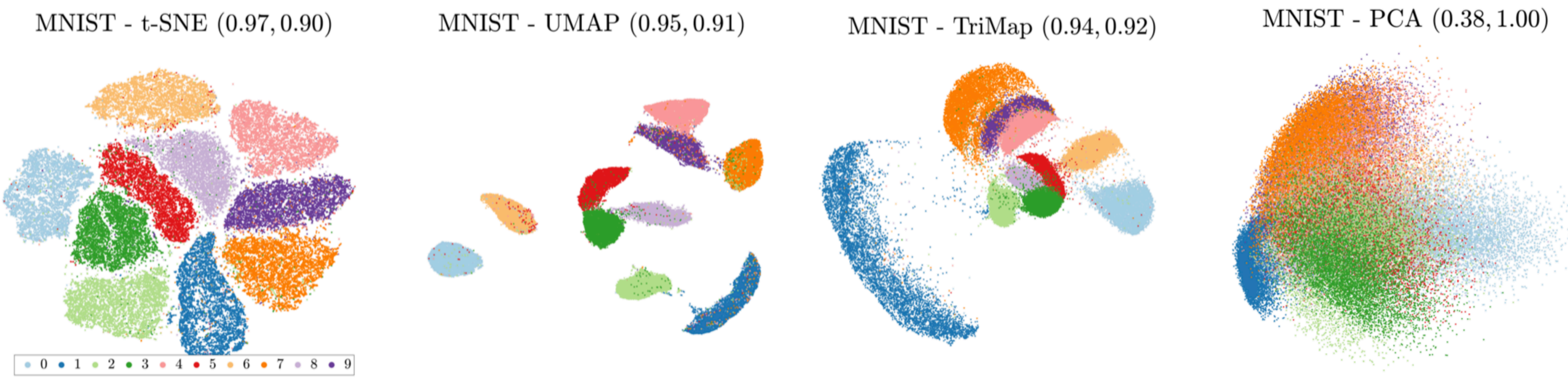


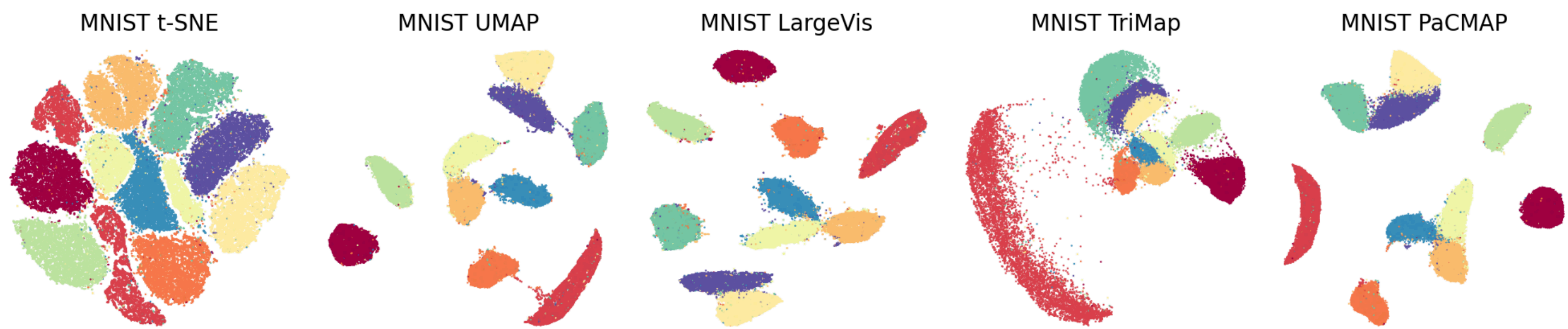
Measure	Baseline	t-SNE	LargeVis	UMAP	TriMAP	PaCMAP	PCA
SVM accuracy	0.972	0.909	0.799	0.844	0.778	**0.942**	
nn-Accuracy		**1.00**	0.96	0.91	0.90		0.67
Random Triplet accuracy		0.698	**0.735**	0.649	0.659	0.699	
Centroids Triplet Accuracy		0.751	**0.799**	0.705	0.715	0.758	
Global score		0.63	0.74	0.52	**0.69**		**1.00**



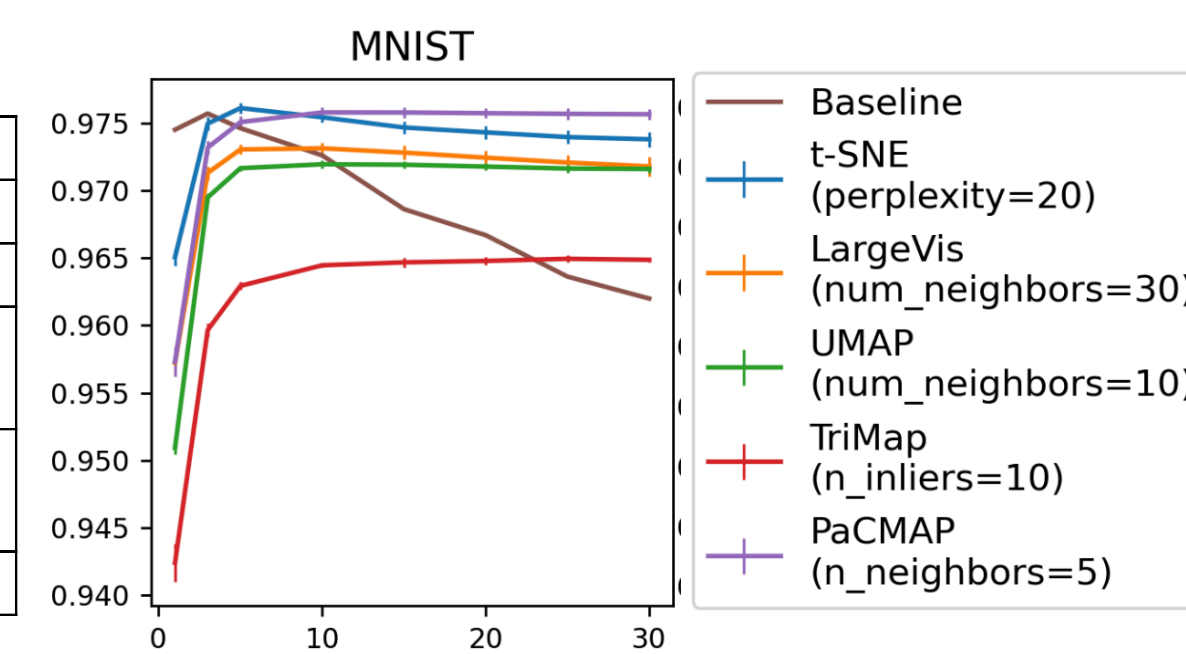
DR methods

A quantitative comparison



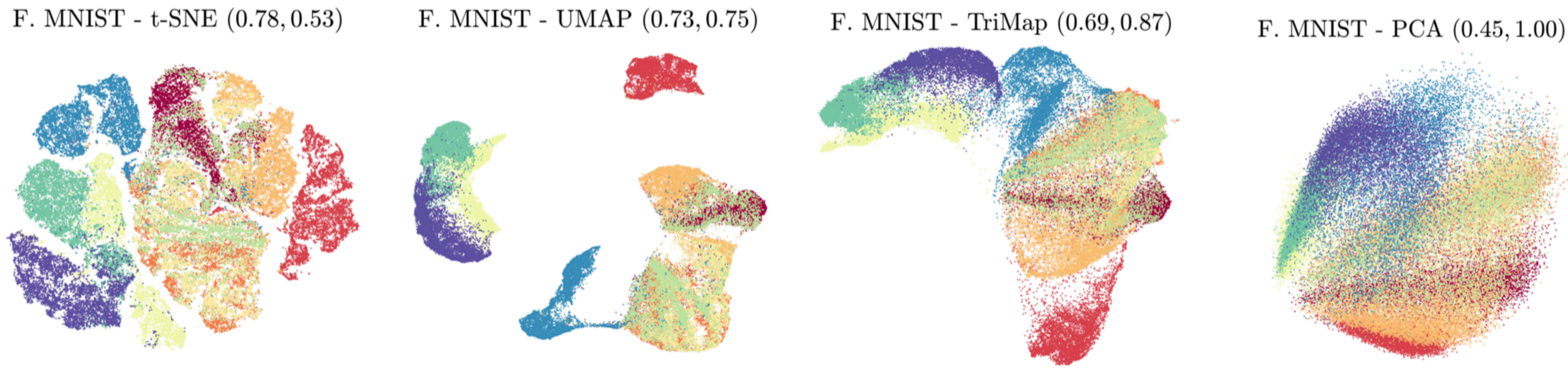


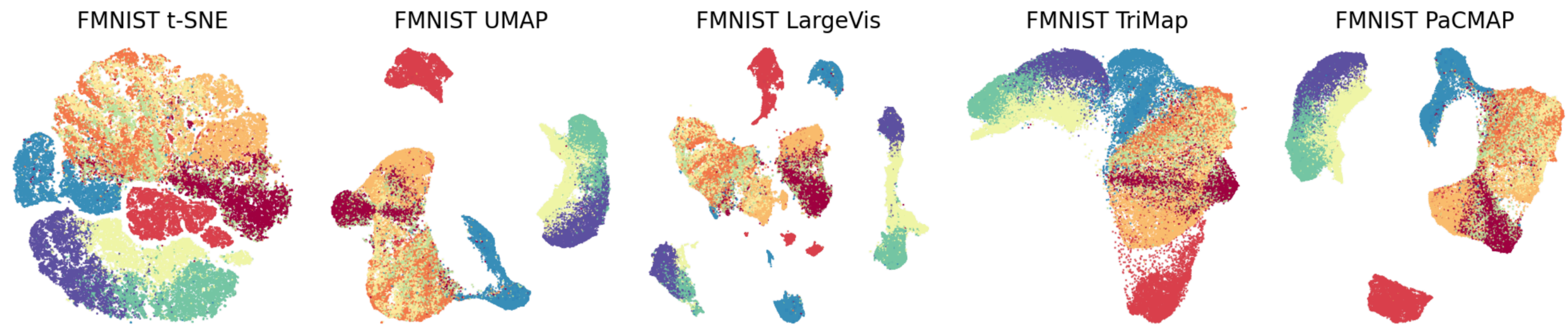
Measure	Baseline	t-SNE	LargeVis	UMAP	TriMAP	PaCMAP	PCA
SVM accuracy	0.926	0.967	0.965	0.970	0.960	**0.974**	
nn-Accuracy		**0.97**	0.95	0.95	0.94		0.38
Random Triplet accuracy		0.600	0.601	0.614	0.600	**0.619**	
Centroids Triplet Accuracy		0.650	0.668	0.793	**0.806**	0.772	
Global score		0.90	0.87	0.91	**0.92**		**1.00**



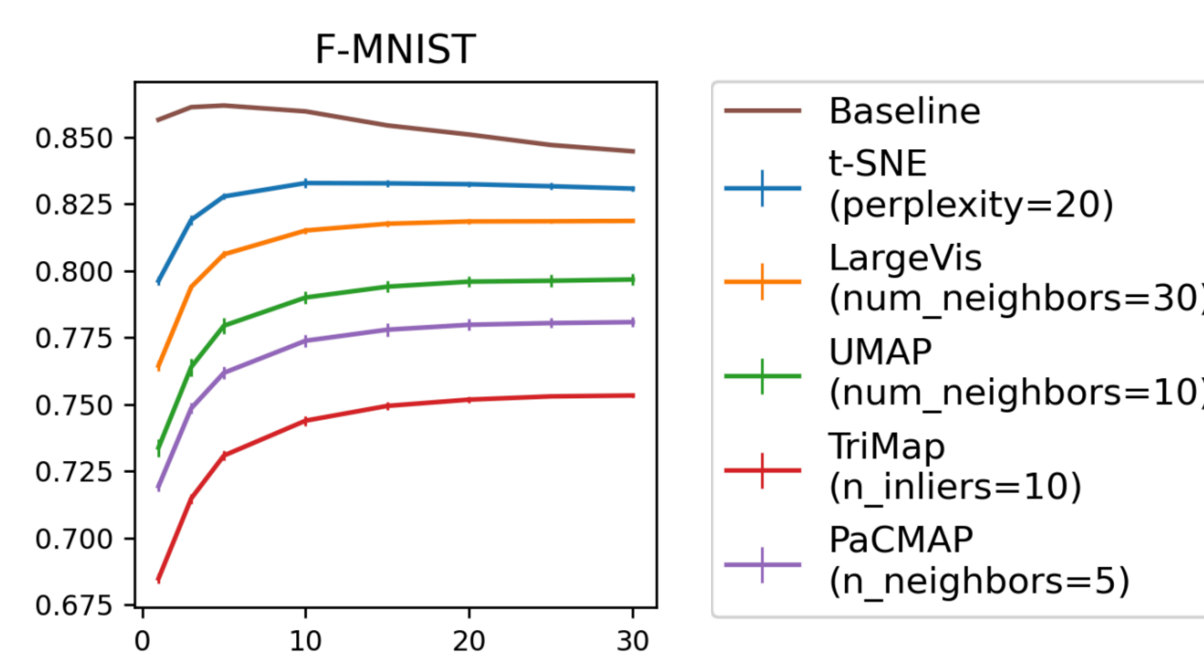
DR methods

A quantitative comparison





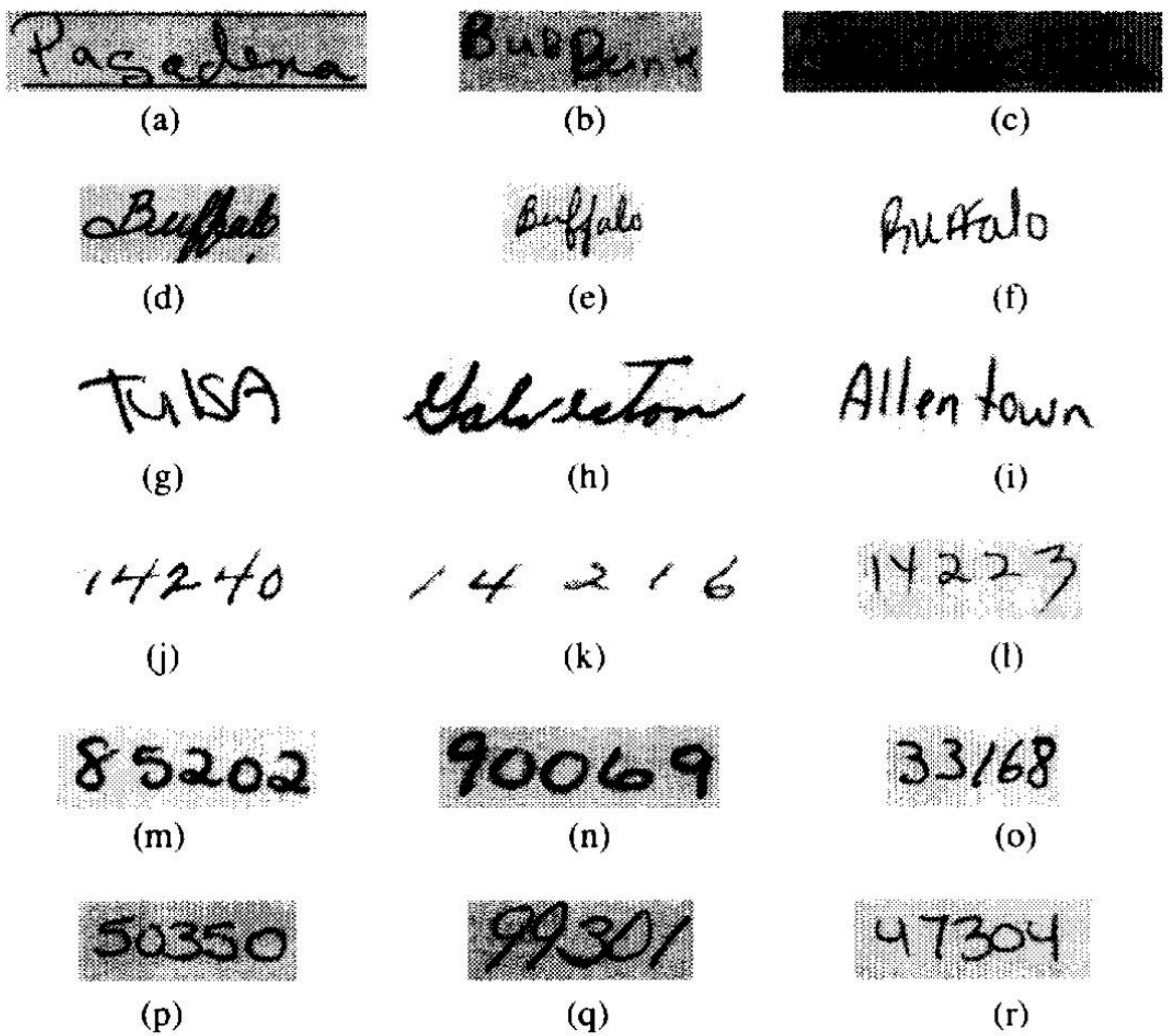
Measure	Baseline	t-SNE	LargeVis	UMAP	TriMAP	PaCMAP	PCA
SVM accuracy	**0.854**	0.754	0.748	0.742	0.729	0.752	
nn-Accuracy		**0.78**	0.75	0.73	0.69		0.45
Random Triplet accuracy		0.679	0.657	0.740	**0.777**	0.741	
Centroids Triplet Accuracy		0.726	0.749	0.869	**0.895**	0.858	
Global score		0.53	0.49	0.75	**0.87**		**1.00**



DR methods

A quantitative comparison

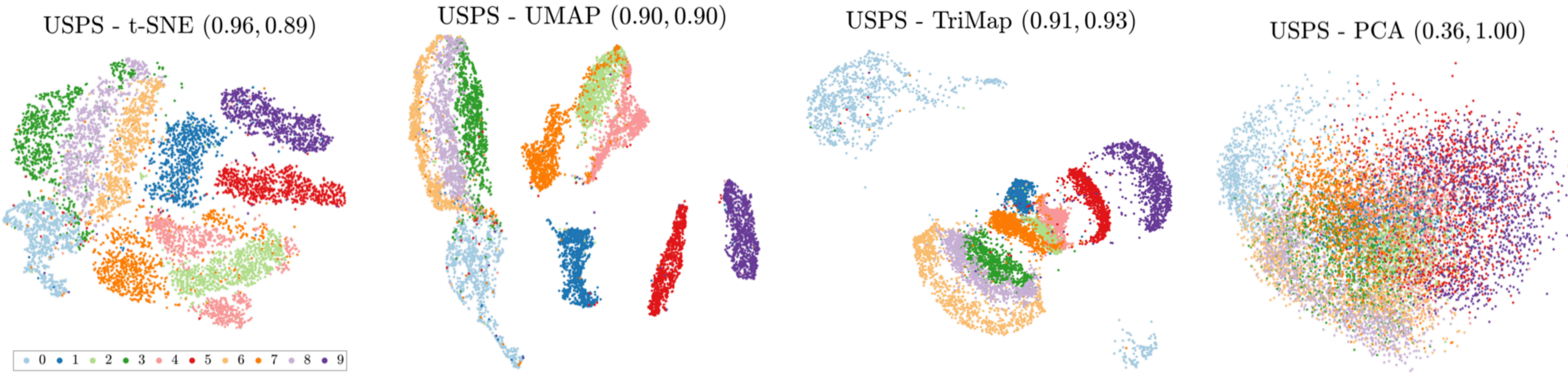
Datasets

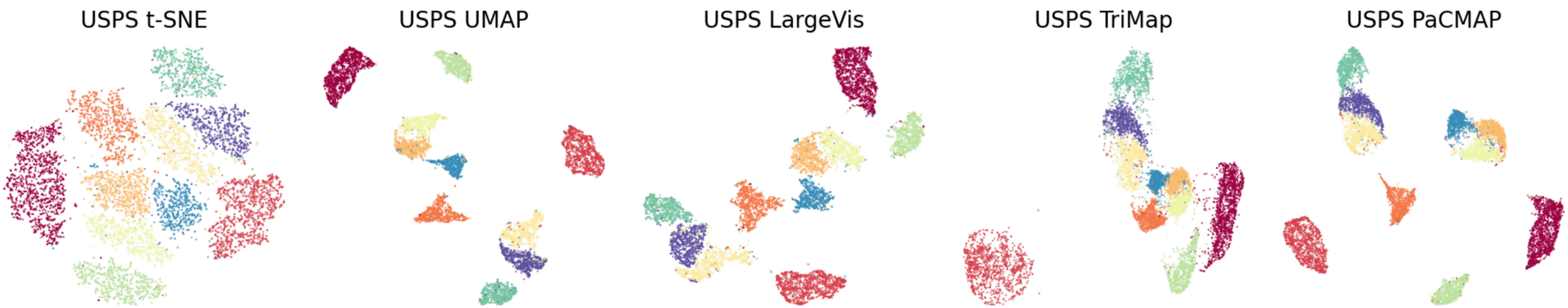


- Olivetti Faces
- COIL20
- COIL100
- S-curve with a hole
- Mammoth
- **USPS**
- MNIST
- Mouse scRNA-seq
- 20 NewsGroups
- Flow Cytometry
- KDD Cup 99

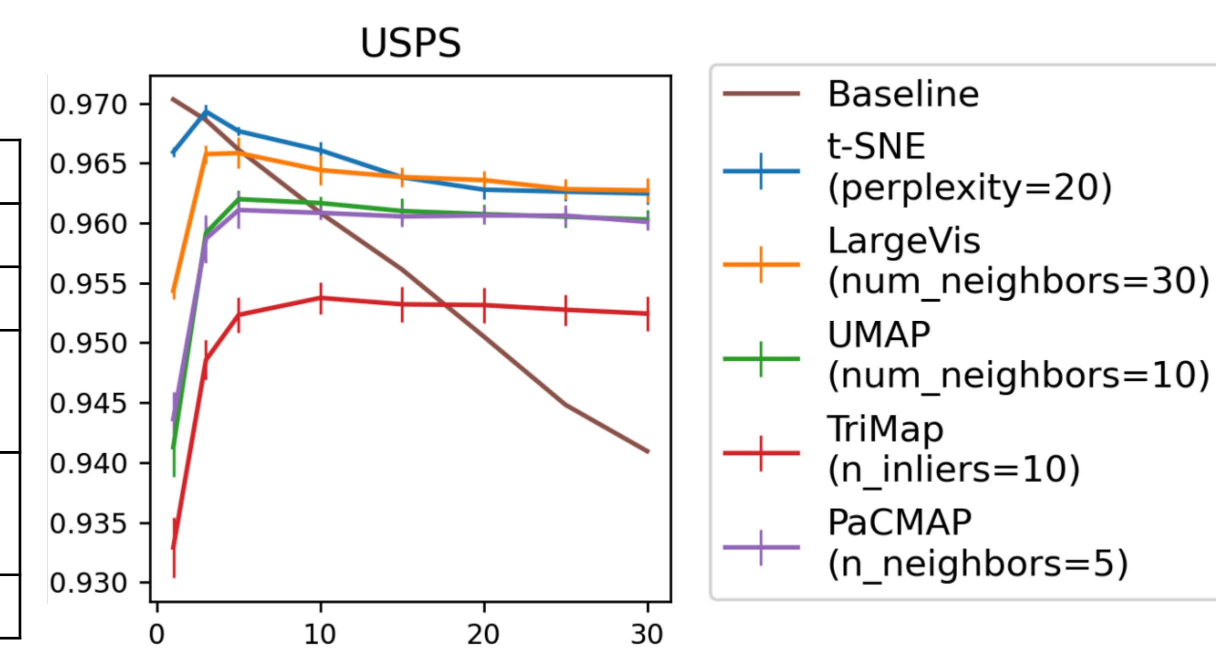
DR methods

A quantitative comparison





Measure	Baseline	t-SNE	LargeVis	UMAP	TriMAP	PaCMAP	PCA
SVM accuracy	0.949	0.959	0.957	0.956	0.946	**0.958**	
nn-Accuracy		**0.96**	0.94	0.90	0.93		0.36
Random Triplet accuracy		0.654	0.668	**0.669**	0.640	0.665	
Centroids Triplet Accuracy		0.756	0.811	0.814	**0.874**	0.852	
Global score		0.89	0.87	0.90	**0.93**		**1.00**



DR methods

A quantitative comparison

Datasets

20000 documents
20 newsgroups

- Computer graphics
- Windows OS Miscellaneous
- IMB PC Hardware
- Windows X

- Automobiles
- Motorcycles
- Baseball
- Hockey

- Cryptography
- Electronics
- Medicine
- Space

- Miscellaneous Sales

- Miscellaneous Politics
- Politics and Guns
- Middle East Politics

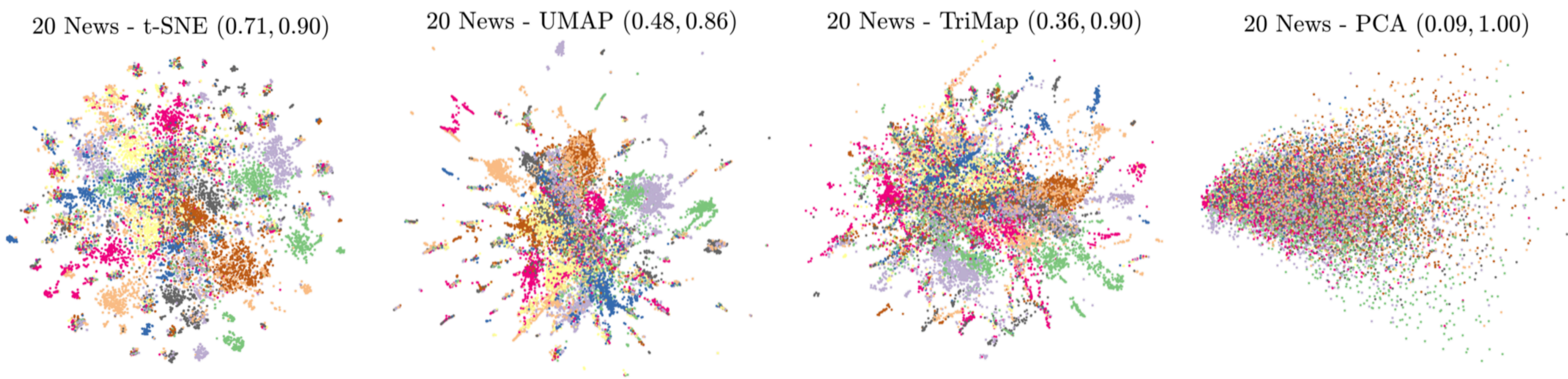
- Miscellaneous Religion
- Atheism
- Christianity

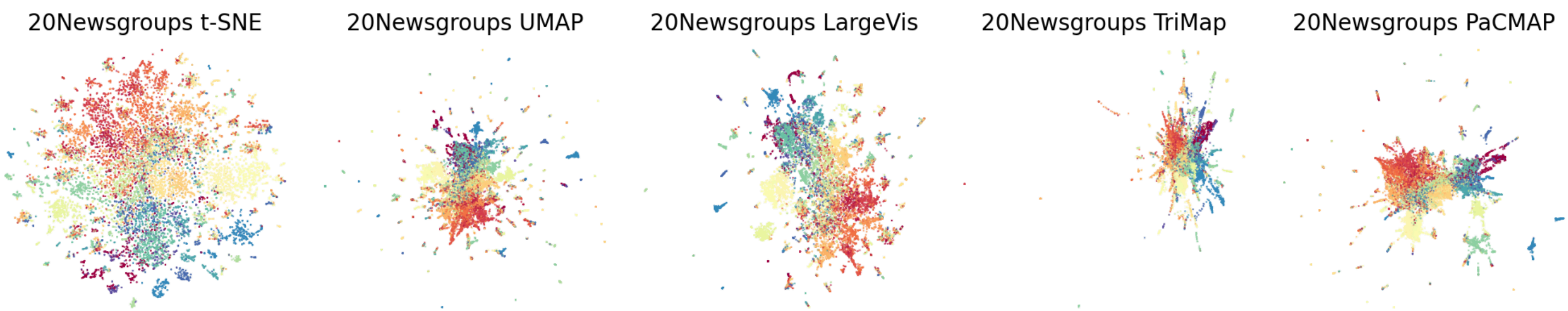
```
From: Bill.Kayser@delft.SGp.slb.COM (Bill Kayser)
Subject: Re: TeleUse, UIM/X, and C++
Article-I.D.: parsival.199304060629.AA00339
Organization: The Internet
Lines: 25
NNTP-Posting-Host: enterpoop.mit.edu
To: xpert@expo.lcs.mit.edu
Cc: Bill.Kayser@delft.sgp.slb.com


>
> Does anyone have any good ideas on how to integrate C++ code elegantly
> with TeleUse, UIM/X / Interface Architect generated code?
>
> Source would be great, but any suggestions are welcome.

It's my understanding that the next release of UIM/X, due out
last February :-) has full support for C++.

I use XDesigner which does not have the interpreter or UI meta languages
of these other tools but does fully support C++ code generation,
reusable templates via C++ classes which are generated, a variety of
other handy features for using C++ and layout functions in different
ways, and generates Motif 1.2 code (including drag 'n drop,
internationalization, etc.).  Fits in quite nicely with Doug Young's
paradigm for C++/Motif.

Available in the US from VI Corp, in Europe from Imperial Software,
London (see FAQ for details).

Bill
------------------------------------------------------------------------
                                           Schlumberger Geco Prakla
                                           kayser@delft.sgp.slb.com
```

Olivetti Faces
COIL20
COIL100
S-curve with a hole
Mammoth
USPS
MNIST
Mouse scRNA-seq
20 NewsGroups
Flow Cytometry
KDD Cup 99

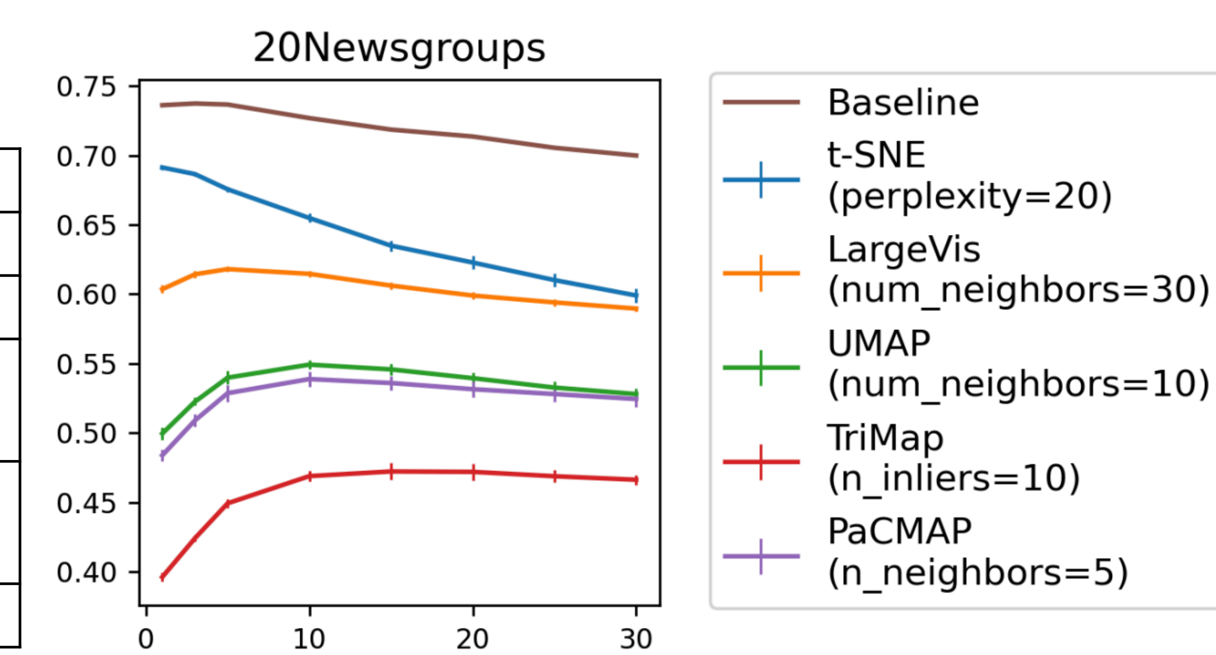
DR methods

A quantitative comparison



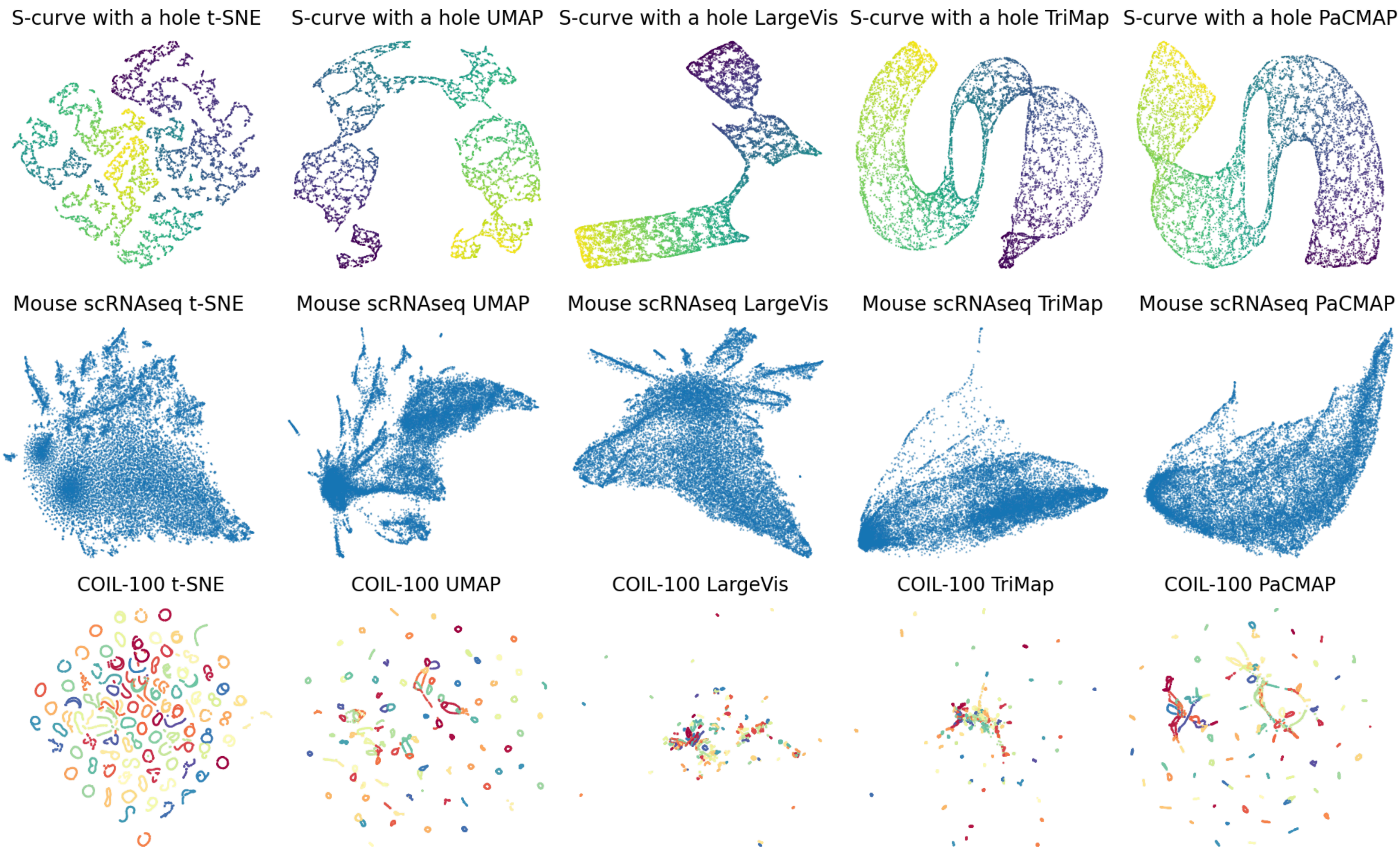


Measure	Baseline	t-SNE	LargeVis	UMAP	TriMAP	PaCMAP	PCA
SVM accuracy	**0.792**	0.435	0.444	0.431	0.410	0.447	
nn-Accuracy		**0.71**	0.59	0.48	0.36		0.09
Random Triplet accuracy		0.645	0.632	0.664	**0.704**	0.666	
Centroids Triplet Accuracy		0.787	0.779	0.774	**0.794**	0.773	
Global score		**0.90**	0.89	0.86	**0.90**		**1.00**



DR methods

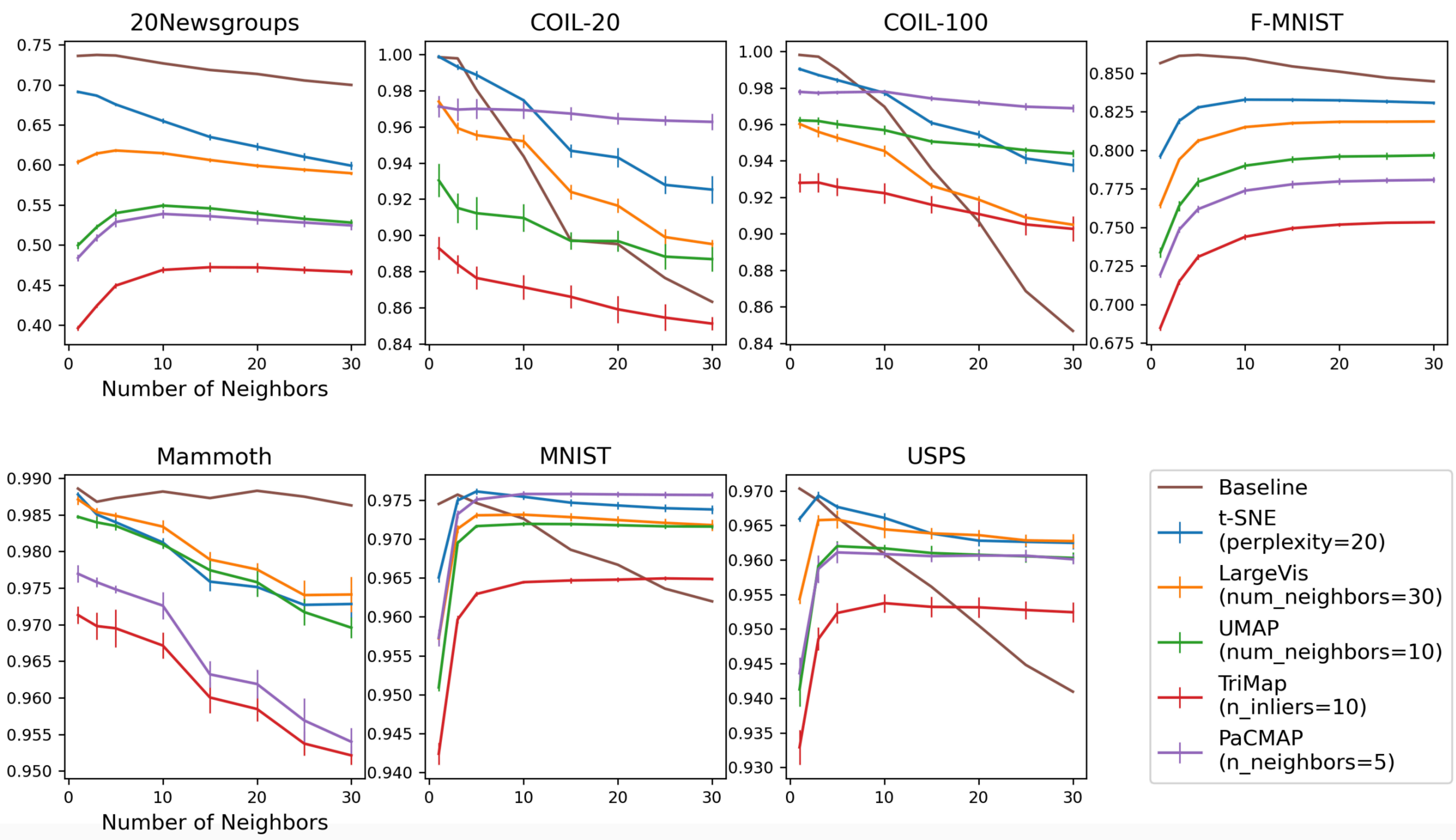
A quantitative comparison



DR methods

A quantitative comparison

K-nearest neighbours classifier accuracy



DR methods
A quantitative comparison

SVM classifier accuracy

Dataset (Size)	Baseline	t-SNE	LargeVis	UMAP	TriMap	PaCMAP
Olivetti Faces (0.4K)	0.965	*0.590 ± 0.039*	0.048 ± 0.004	0.562 ± 0.021	0.572 ± 0.014	***0.614 ± 0.013***
COIL-20 (1.4K)	0.972	0.909 ± 0.015	0.799 ± 0.020	0.844 ± 0.004	0.778 ± 0.010	**0.942 ± 0.009**
COIL-100 (7.2K)	0.989	0.911 ± 0.004	0.707 ± 0.014	0.879 ± 0.007	0.737 ± 0.019	**0.933 ± 0.009**
USPS (9k)	0.949	***0.959 ± 0.002***	*0.957 ± 0.001*	*0.956 ± 0.002*	0.946 ± 0.001	*0.958 ± 0.001*
Mammoth (10K)	0.961	0.927 ± 0.009	0.923 ± 0.011	***0.941 ± 0.003***	0.900 ± 0.004	*0.933 ± 0.004*
20Newsgroups (18K)	0.792	0.435 ± 0.014	*0.444 ± 0.012*	0.431 ± 0.013	0.410 ± 0.007	***0.447 ± 0.006***
MNIST (70K)	0.926	0.967 ± 0.002	0.965 ± 0.004	0.970 ± 0.001	0.960 ± 0.001	**0.974 ± 0.001**
F-MNIST (70K)	0.854	***0.754 ± 0.003***	*0.748 ± 0.003*	0.742 ± 0.003	0.729 ± 0.001	*0.752 ± 0.004*
KDD Cup 99 (4M)	0.956	–	–	–	0.767 ± 0.034	**0.909 ± 0.035**

DR methods

A quantitative comparison

Random Triplet accuracy

Dataset (Size)	t-SNE	LargeVis	UMAP	TriMap	PaCMAP
Olivetti Faces (0.4K)	0.739 ± 0.012	0.532 ± 0.013	*0.754 ± 0.010*	0.739 ± 0.012	***0.761 ± 0.003***
COIL-20 (1.4K)	0.698 ± 0.016	**0.735 ± 0.011**	0.649 ± 0.014	0.659 ± 0.006	0.699 ± 0.007
COIL-100 (7.2K)	0.577 ± 0.012	0.630 ± 0.021	0.568 ± 0.011	0.633 ± 0.002	**0.718 ± 0.005**
USPS (9K)	0.654 ± 0.013	*0.668 ± 0.011*	***0.669 ± 0.002***	0.640 ± 0.002	*0.665 ± 0.002*
S-Curve with Hole (9.5K)	0.722 ± 0.045	0.834 ± 0.041	0.800 ± 0.013	0.838 ± 0.004	**0.866 ± 0.010**
Mammoth (10K)	0.701 ± 0.038	0.766 ± 0.024	0.816 ± 0.001	***0.874 ± 0.001***	*0.872 ± 0.003*
20Newsgroups (18K)	0.645 ± 0.002	0.632 ± 0.001	0.664 ± 0.002	**0.704 ± 0.002**	0.666 ± 0.003
Mouse scRNA-seq (20K)	0.715 ± 0.002	0.719 ± 0.003	*0.727 ± 0.002*	***0.728 ± 0.001***	*0.727 ± 0.001*
MNIST (70K)	0.600 ± 0.007	0.601 ± 0.007	0.614 ± 0.001	0.600 ± 0.001	**0.619 ± 0.001**
F-MNIST (70K)	0.679 ± 0.019	0.657 ± 0.011	0.740 ±0.001	**0.777 ± 0.001**	0.741 ± 0.002
Flow Cytometry (3M)	–	–	–	0.857 ± 0.001	**0.894 ± 0.005**
KDD Cup 99 (4M)	–	–	–	0.660 ± 0.007	**0.752 ± 0.002**

Centroid triplets accuracy

Dataset (Size)	t-SNE	LargeVis	UMAP	TriMap	PaCMAP
Olivetti Faces (0.4K)	0.788 ± 0.008	0.534 ± 0.006	**0.804 ± 0.002**	0.786 ± 0.007	0.797 ± 0.003
COIL-20 (1.4K)	0.751 ± 0.017	**0.799 ± 0.015**	0.705 ± 0.017	0.715 ± 0.004	0.758 ± 0.012
COIL-100 (7.2K)	0.625 ± 0.014	0.680 ± 0.015	0.630 ± 0.001	0.704 ± 0.001	**0.756 ± 0.005**
USPS (9K)	0.756 ± 0.029	0.811 ± 0.025	0.814 ± 0.018	**0.874 ± 0.002**	0.852 ± 0.002
Mammoth (10K)	0.654 ± 0.067	0.714 ± 0.036	0.818 ± 0.009	0.837 ± 0.003	**0.877 ± 0.002**
20Newsgroups (18K)	*0.787 ± 0.004*	*0.779 ± 0.026*	*0.774 ± 0.016*	***0.794 ± 0.007***	0.773 ± 0.008
MNIST (70K)	0.650 ± 0.026	0.668 ± 0.030	0.793 ± 0.009	**0.806 ± 0.001**	0.772 ± 0.008
F-MNIST (70K)	0.726 ± 0.052	0.749 ± 0.070	0.869 ± 0.002	**0.895 ± 0.001**	0.858 ± 0.002
KDD Cup99 (4M)	–	–	–	0.536 ± 0.010	**0.572 ± 0.004**

DR methods

A quantitative comparison

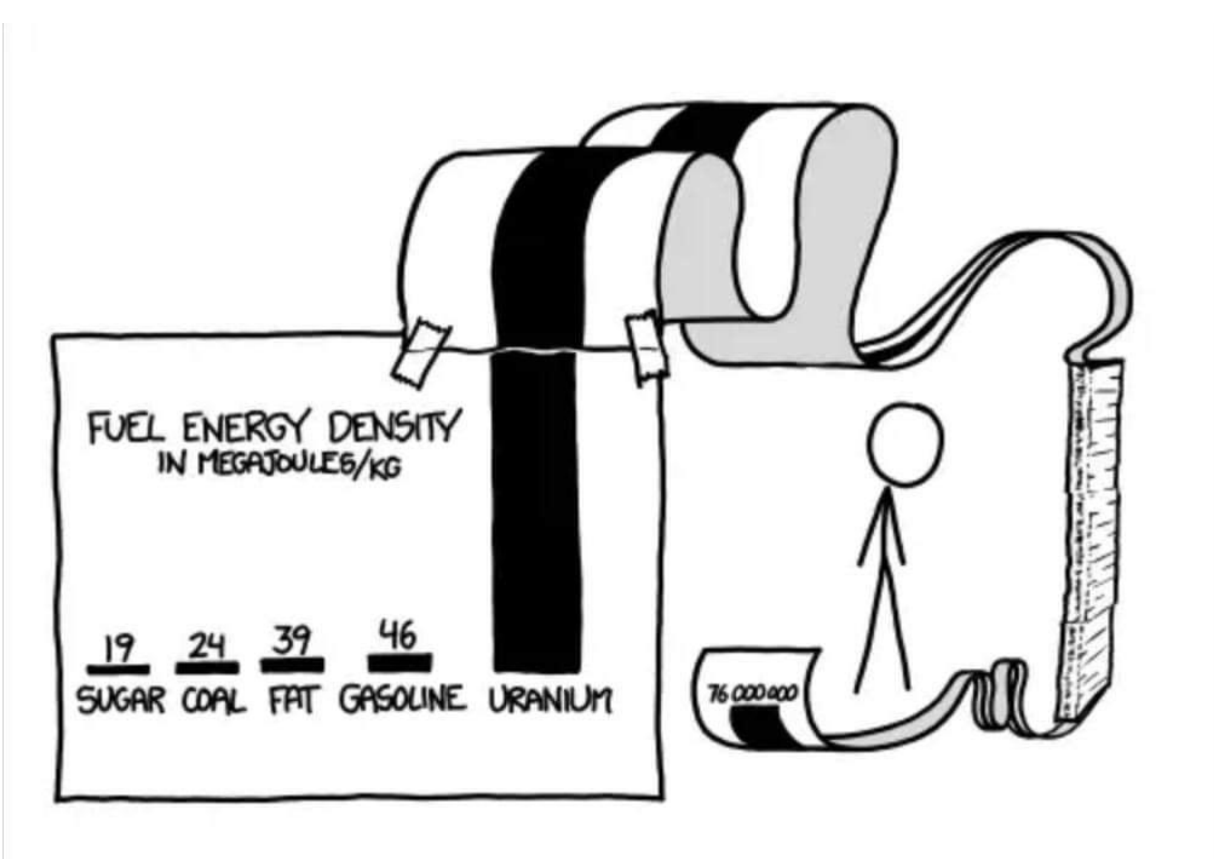
Quantitative assessment useful but not generalizable

No “one solution fits all”

“Best” method depends on goal

“Best” method depends on data

Other features to consider: computational resources (time and memory), robustness to parameters adjustment, robustness to initialisation, interpretability



SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*