logistics

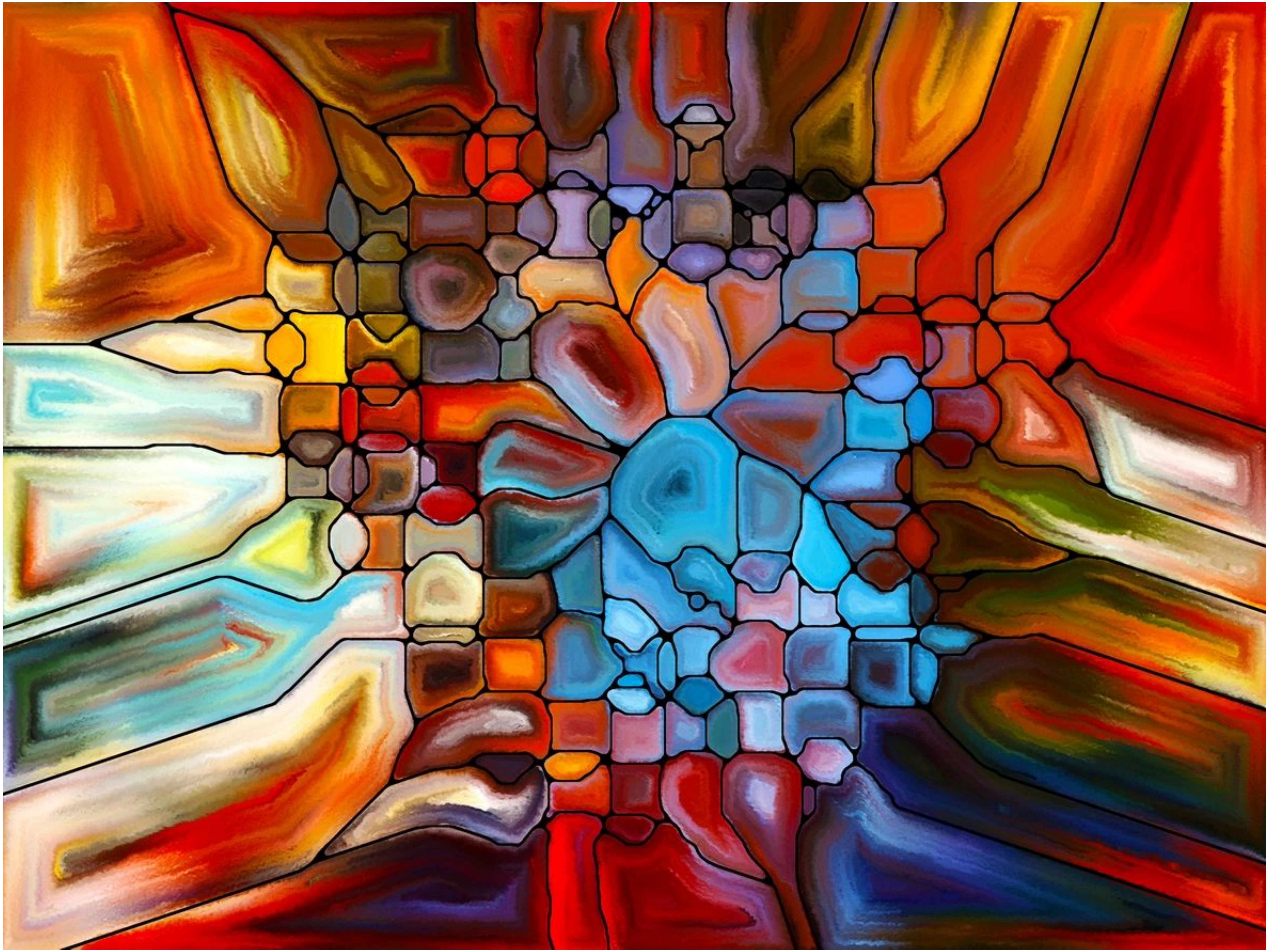
exam:

- written questions + practical Python exercises
- to be held in presence during exam time
- access to online resources allowed

Bring your own laptop!

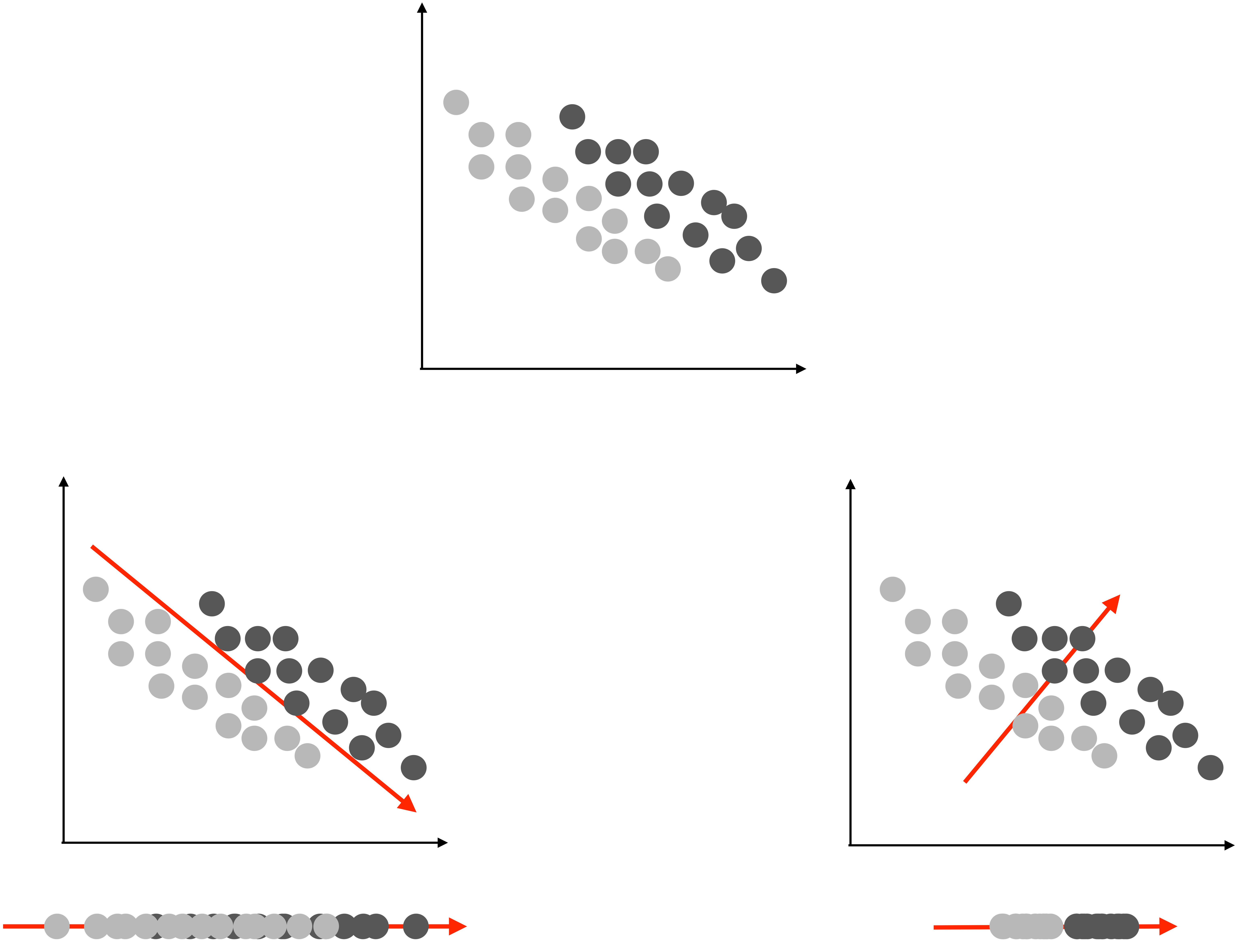
exam enrollment:

- If you are attending ONLY the **Data Visualisation Lab [145683]**:
  - Register ONLY to the exam in one of the following dates:
    12th June
    04th July
    11th September
- If you are attending the **Big Data Technologies and Visualisation [136069]**:
  - Register to the exam in one of the following dates:
    12th June
    04th July
    11th September
  - Register ALSO to the corresponding exam verbalisation
    30th June
    31st July
    13th September



Supervised DR

Example



Unsupervised vs supervised

Unsupervised	Supervised
Retain meaningful properties of the original data	Maximise discriminative power wrt target variables
Unlabelled data	Labelled data
More versatile, used for a wide range of tasks	More specific, usually for prediction tasks
PCA, t-SNE, UMAP, …	LDA, NCA, PLS,…

Supervised DR

- Linear Discriminant Analysis
- Neighbourhood Components Analysis
- Supervised UMAP
- Centroid-Encoder

Supervised DR

- Linear Discriminant Analysis
- Neighbourhood Components Analysis
- Supervised UMAP
- Centroid-Encoder

LDA

Linear Discriminant Analysis

Aim: find a straight line or plane that best separates classes while minimizing overlap within each class

LDA

Linear Discriminant Analysis

Aim: find a straight line or plane that best separates classes while minimizing overlap within each class

Theoretical assumptions

- Data follows a normal distribution
- Each class has identical covariance matrix

LDA

Linear Discriminant Analysis

Aim: find a straight line or plane that best separates classes while minimizing overlap within each class

Theoretical assumptions

- Data follows a normal distribution
- Each class has identical covariance matrix

- 2 classes: Fisher's Linear Discriminant (Fisher, 1936)
- Multiple classes: Linear Discriminant Analysis (Rao, 1948)

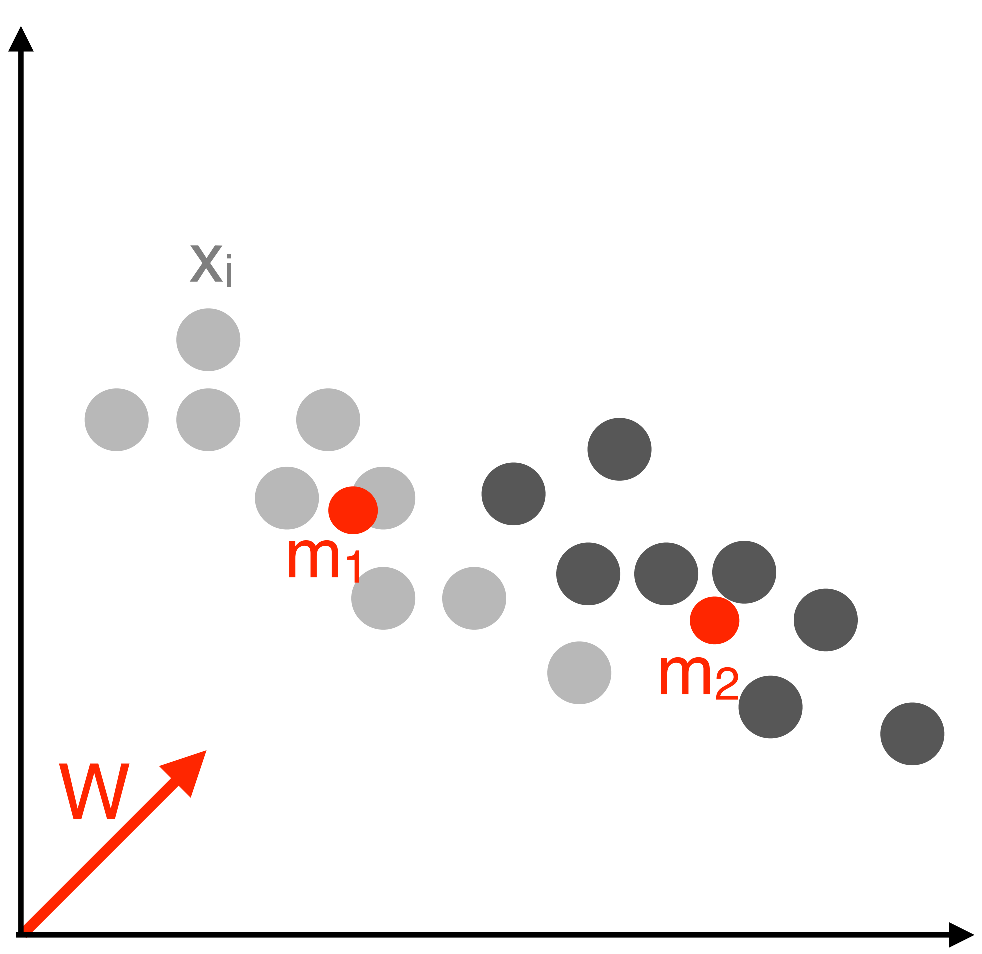
FLD

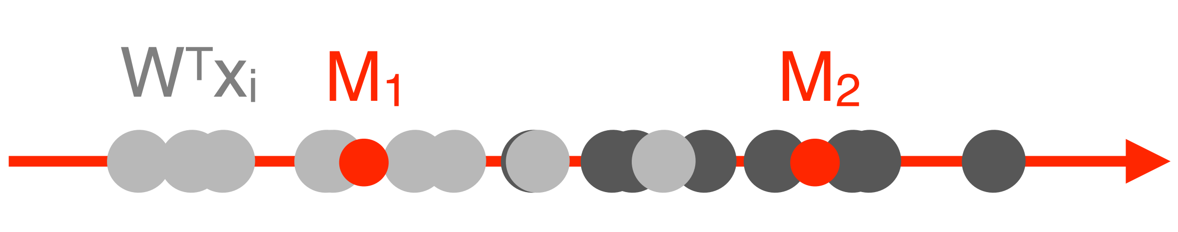
Fisher's Linear Discriminant

Idea: project data points onto a line to maximise between-class scatter and minimise within-class scatter

- D-dimensional space
- x_1, x_2, …, x_N data samples
- 2 classes C_1 with N_1 samples and C_2 with N_2 samples ($N = N_1 + N_2$)

- m_i mean in high-dimensional space
- M_i mean in low-dimensional space



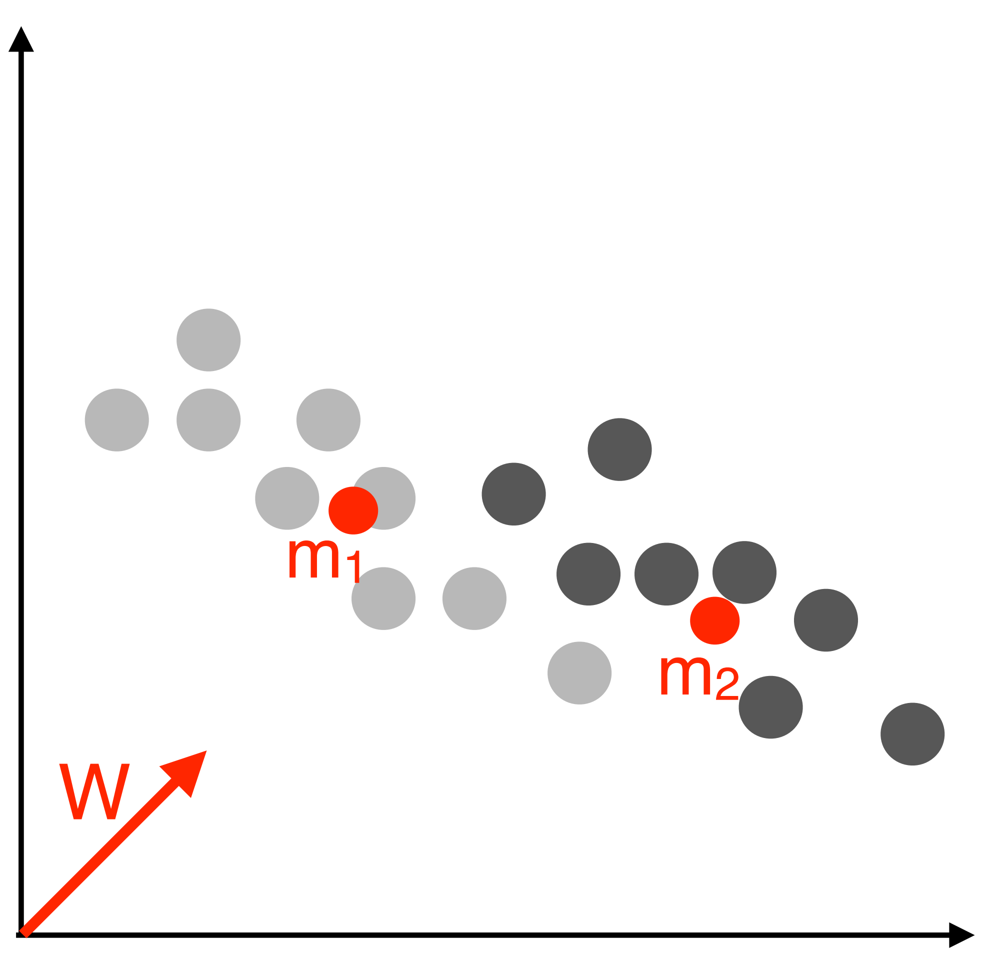


FLD

Fisher's Linear Discriminant

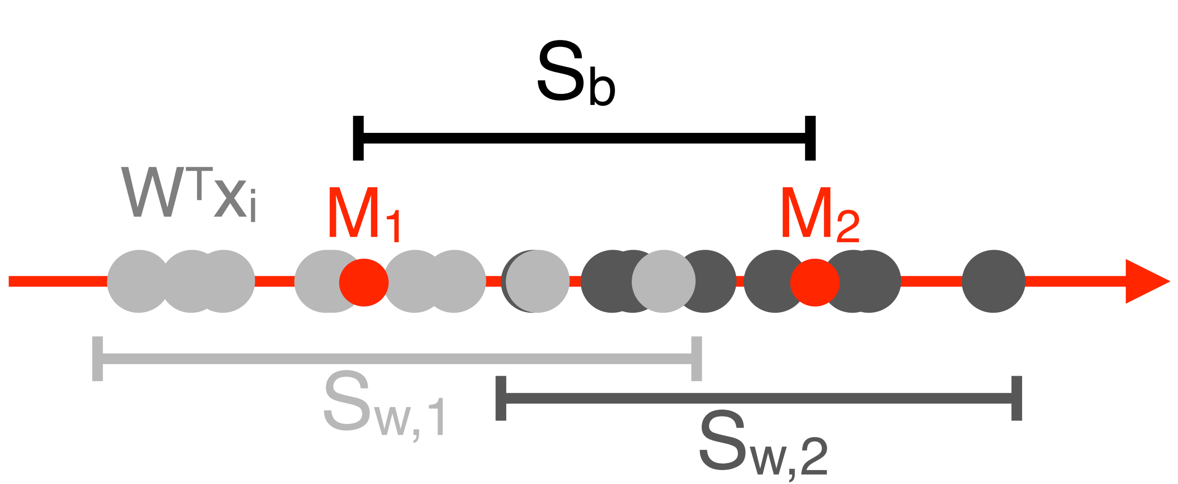
Idea: project data points onto a line to maximise between-class scatter and minimise within-class scatter

- d-dimensional space
- x_1, x_2, …, x_n data samples
- 2 classes C_1 with N_1 samples and C_2 with N_2 samples ($N = N_1 + N_2$)



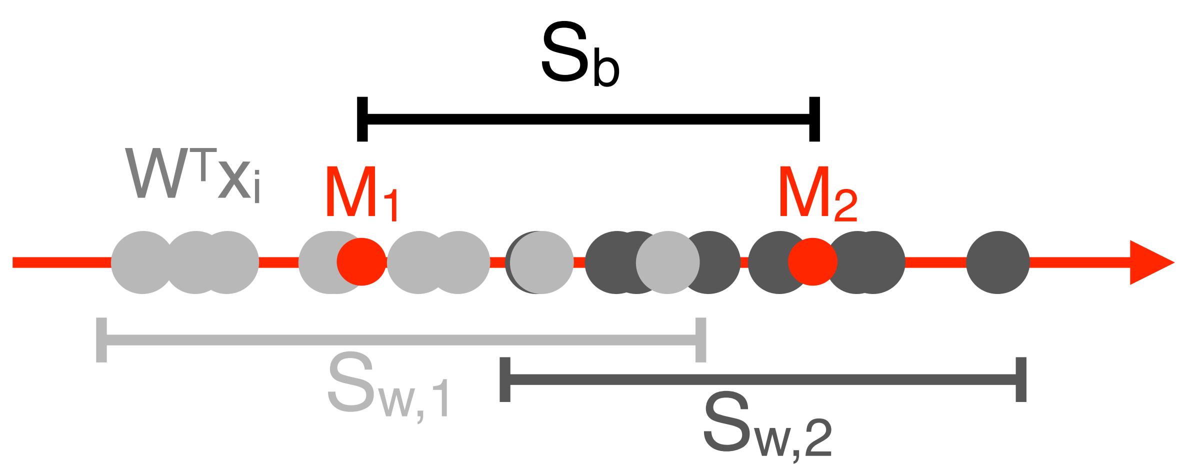
- m_i mean in high-dimensional space
- M_i mean in low-dimensional space

- S_w within class scatter
- S_b between class scatter



FLD

Fisher's Linear Discriminant



$$\frac{S_b^2}{S_{w,1}^2 + S_{w,2}^2} = \frac{(M_1 - M_2)^2}{S_{w,1}^2 + S_{w,2}^2} \rightarrow \frac{\text{ideally large}}{\text{ideally small}}$$

Numerator $$(M_1 - M_2)^2 = (W^T m_1 - W^T m_2)(W^T m_1 - W^T m_2)^T = W^T (m_1 - m_2) W = W^T s_b W$$

Between class scatter in high-dim

Denominator $$s_{w,i} = \sum_{x_n \in C_i} (x_n - m_i)(x_n - m_i)^T$$

$$S_{w,i} = \sum_{x_n \in C_i} (W^T x_n - M_i)(W^T x_n - M_i)^T = \sum_{x_n \in C_i} (W^T x_n - W^T m_i)(W^T x_n - W^T m_i)^T =$$

$$= W^T \left(\sum_{x_n \in C_i} (x_n - m_i)(x_n - m_i)^T \right) W = W^T s_{w,i} W$$

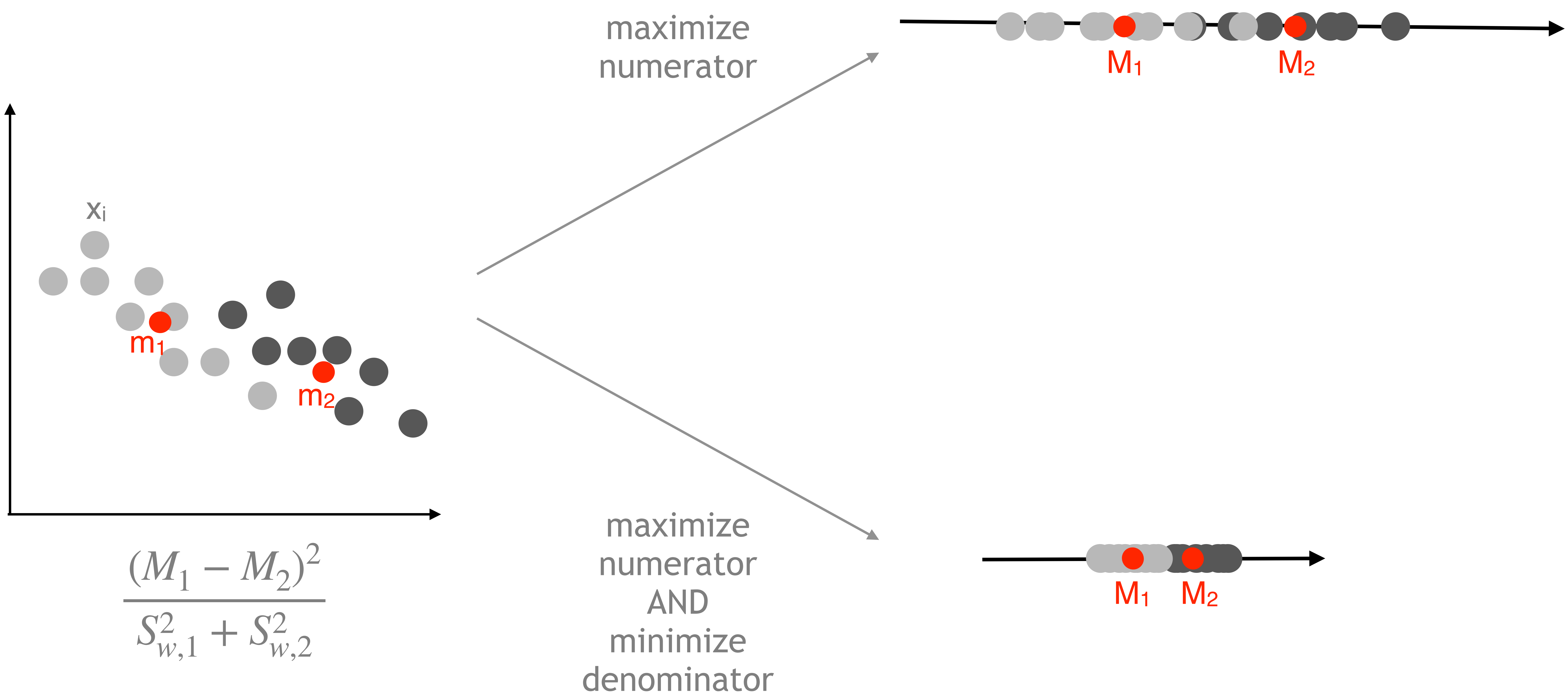
$$S_{w,1}^2 + S_{w,2}^2 = W^T (s_{w,1}^2 + s_{w,2}^2) W = W^T s_w W$$

Within class scatter in high-dim

Solution: eigenvector of the matrix $s_w^{-1} s_b$

FLD

Fisher's Linear Discriminant



$$\frac{(M_1 - M_2)^2}{S_{w,1}^2 + S_{w,2}^2}$$

LDA

Linear Discriminant Analysis

Aim: find a straight line or plane that best separates classes while minimizing overlap within each class

- d-dimensional space
- $x_1, \; x_2, \; \dots, \; x_N$ data samples
- C number of distinct classes

$$m_i = \frac{1}{N_i} \sum_{x_n \in C_i} x_n$$

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i$$

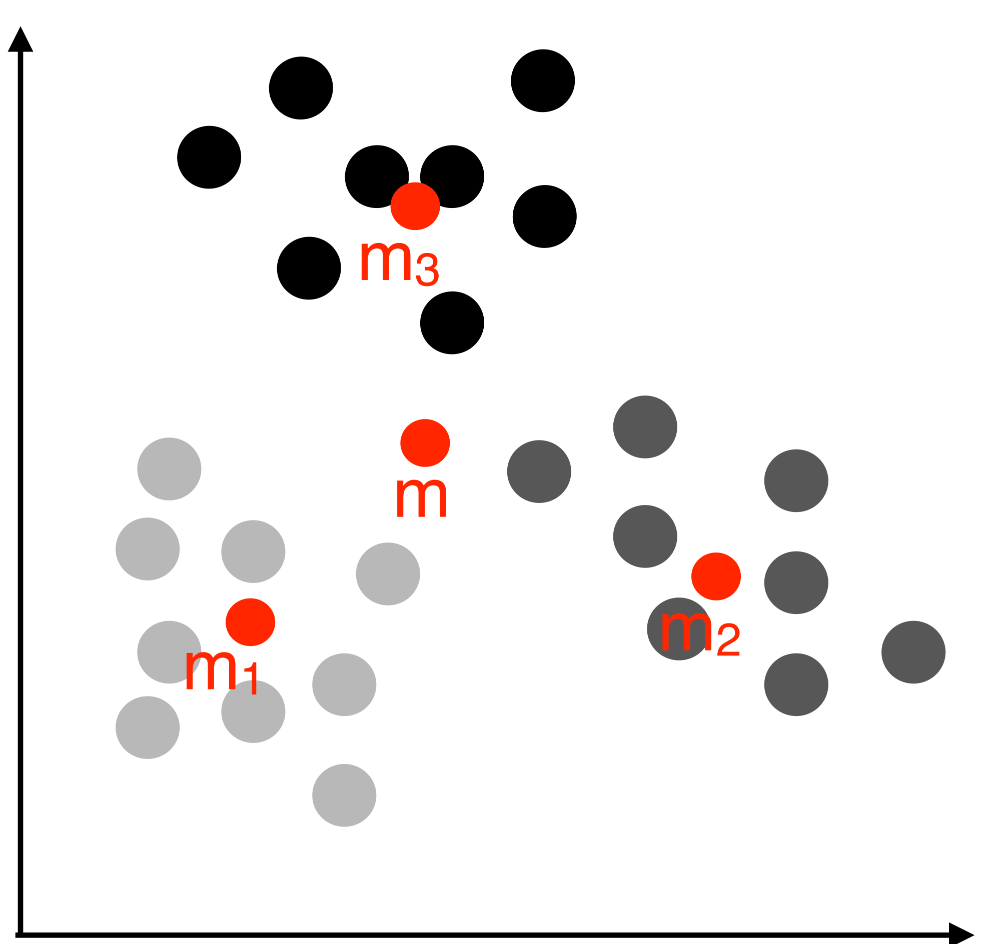
$$s_b = \sum_{i=1}^{C} N_i (m_i - m)(m_i - m)^T$$

between class scatter

$$s_{w,i} = \sum_{x_n \in C_i} (x_n - m_i)(x_n - m_i)^T$$

$$s_w = \sum_{k=1}^{C} s_{w,k}$$

within class scatter



$$W = eig(s_w^{-1} s_b)$$

LDA

Linear Discriminant Analysis

Aim: find a straight line or plane that best separates classes while minimizing overlap within each class

Limitations:

- Linear (extension with non-linear kernels)
- Small sample problem
  Space dimensionality much larger than number of samples
- At most C-1 feature projection

Supervised DR

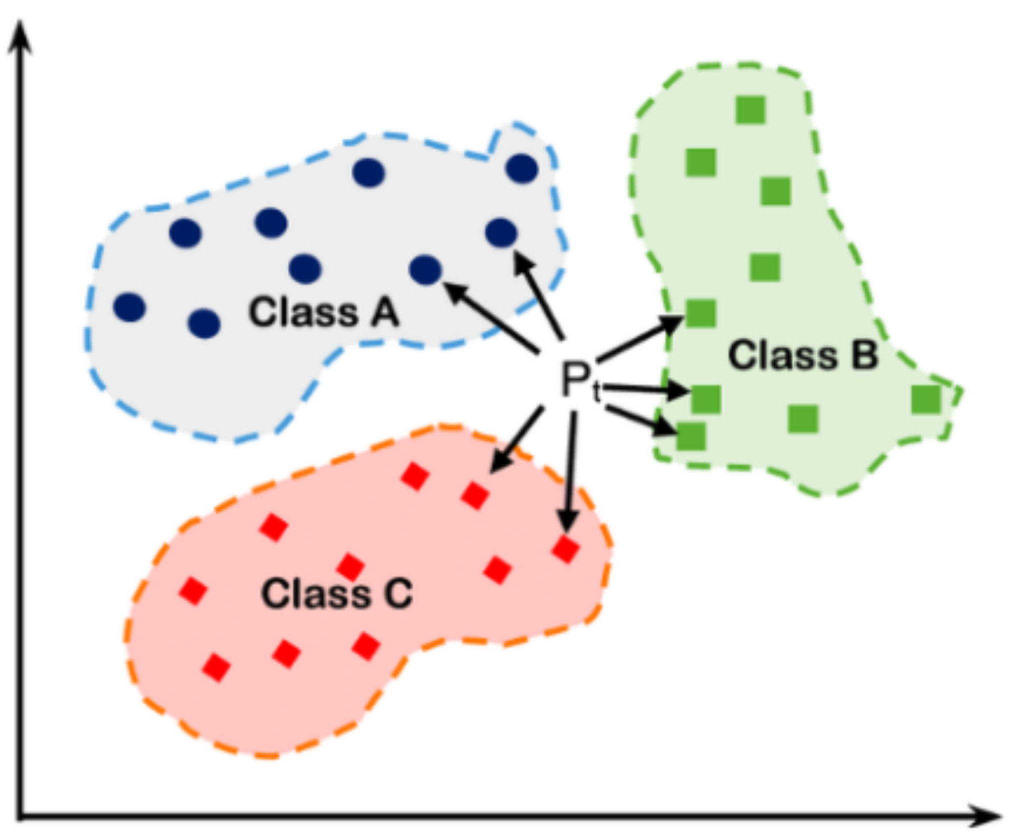
- Linear Discriminant Analysis
- Neighbourhood Components Analysis
- Supervised UMAP
- Centroid-Encoder

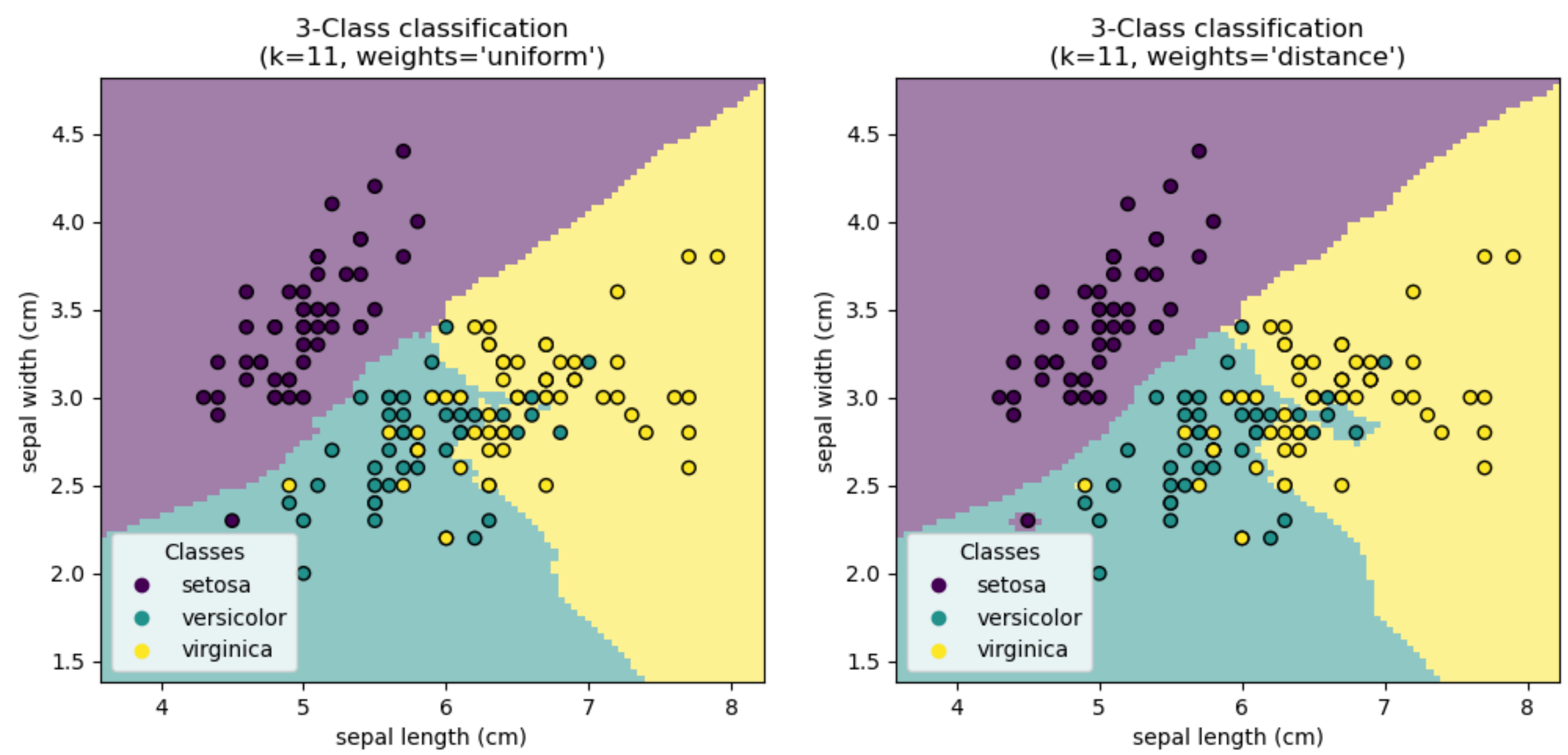
NCA

Neighborhood Component Analysis

Aim: find a linear transformation of data such that the average leave-one-out KNN classification performance is maximized in the low dimensional space

K Nearest Neighbors





K Nearest neighbours classifier

- simple yet effective
- Non linear decision boundaries
- Non parametric
- Only one parameter (number of neighbours k)

- What does “near” mean? What are neighbours?
- Computational complexity

NCA

Neighborhood Component Analysis

- What does "near" mean? What are neighbours?

Given a family of metrics, the "best" metric is the one maximising the leave-one-out kNN cross validation error on the training set

What if the family of metrics is continuously parametrised?
Leave-one-out kNN cross validation error highly discontinuous

Solution: Stochastic Neighbour Selection

Neighbour selection probability $p_{ij} = \frac{e^{-d_{ij}^2}}{\sum_{k \neq i} e^{-d_{ik}^2}}, p_{ii} = 0$

Probability of correct classification $p_i = \sum_{j \in C_i} p_{ij}$

Objective function $L = \sum_i p_i$

Quadratic distance metrics $d_{ij} = (x_i - x_j)^T Q (x_i - x_j) = (Ax_i - Ax_j)^T (Ax_i - Ax_j), Q = A^T A$

NCA

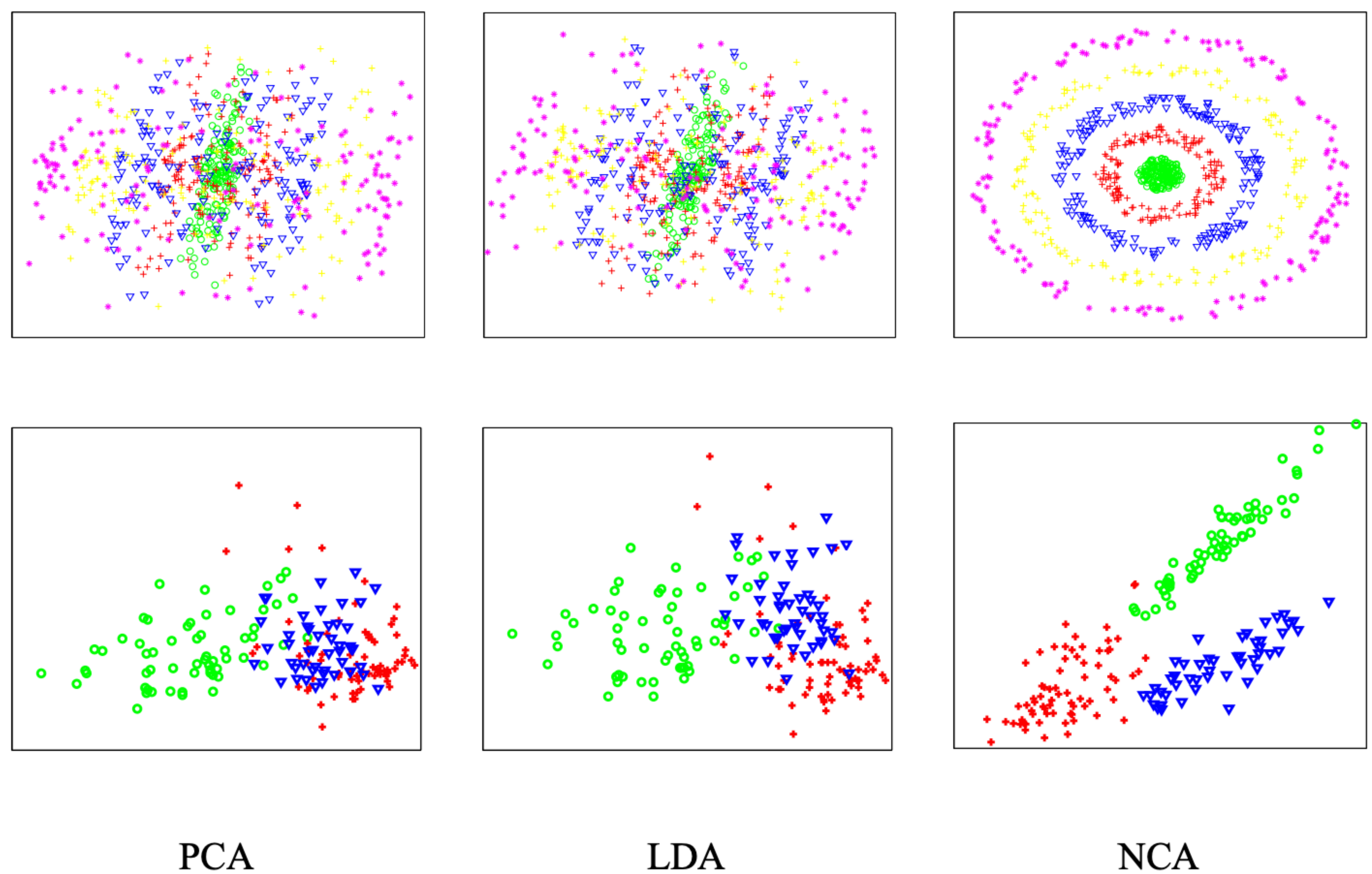
Neighborhood Component Analysis

- What about dimensionality reduction?

  Nonsquare matrix A of size d x D (d<D)
  If d=2 or d=3, data visualisation is possible.

5 concentric circles
In two dim: concentric circles
Third dim: Gaussian noise

Wine dataset from UCI
D=13



NCA

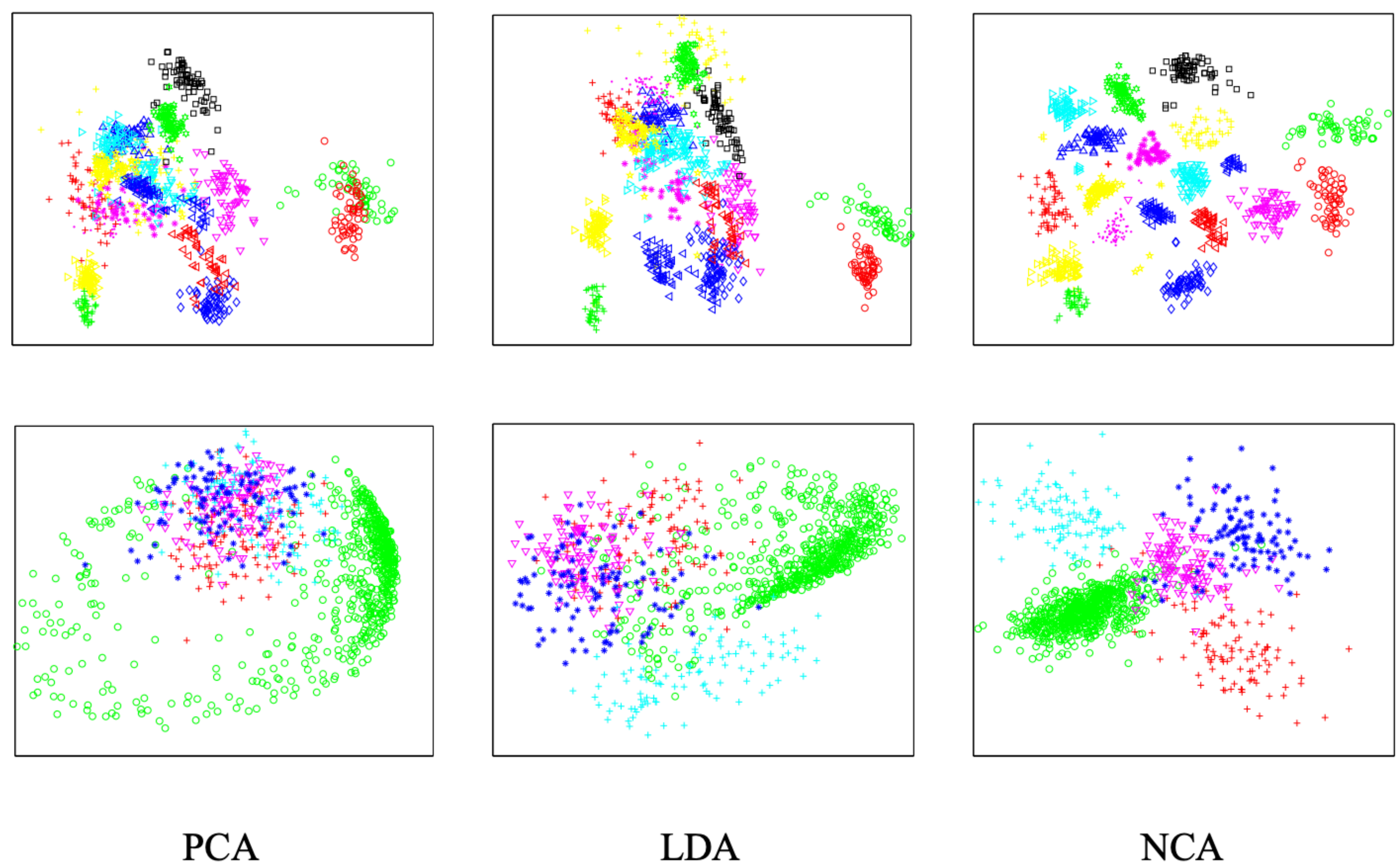
Neighborhood Component Analysis

- What about dimensionality reduction?

Nonsquare matrix A of size d x D (d<D)
If d=2 or d=3, data visualisation is possible.

Faces dataset
D=560

5 digits from USPS dataset
D=64



NCA

Neighborhood Component Analysis

Aim: find a linear transformation of data such that the average leave-one-out KNN classification performance is maximized in the low dimensional space

- Linear trasformation
- No assumptions on classification boundaries (linear, non linear)
- Non parametric
- Computationally efficient

Neighbourhood Components Analysis

Jacob Goldberger, Sam Roweis, Geoff Hinton, Ruslan Salakhutdinov
Department of Computer Science, University of Toronto
{jacob,roweis,hinton,rsalakhu}@cs.toronto.edu

link to paper

Supervised DR

- Linear Discriminant Analysis
- Neighbourhood Components Analysis
- Supervised UMAP
- Centroid-Encoder

Supervised UMAP

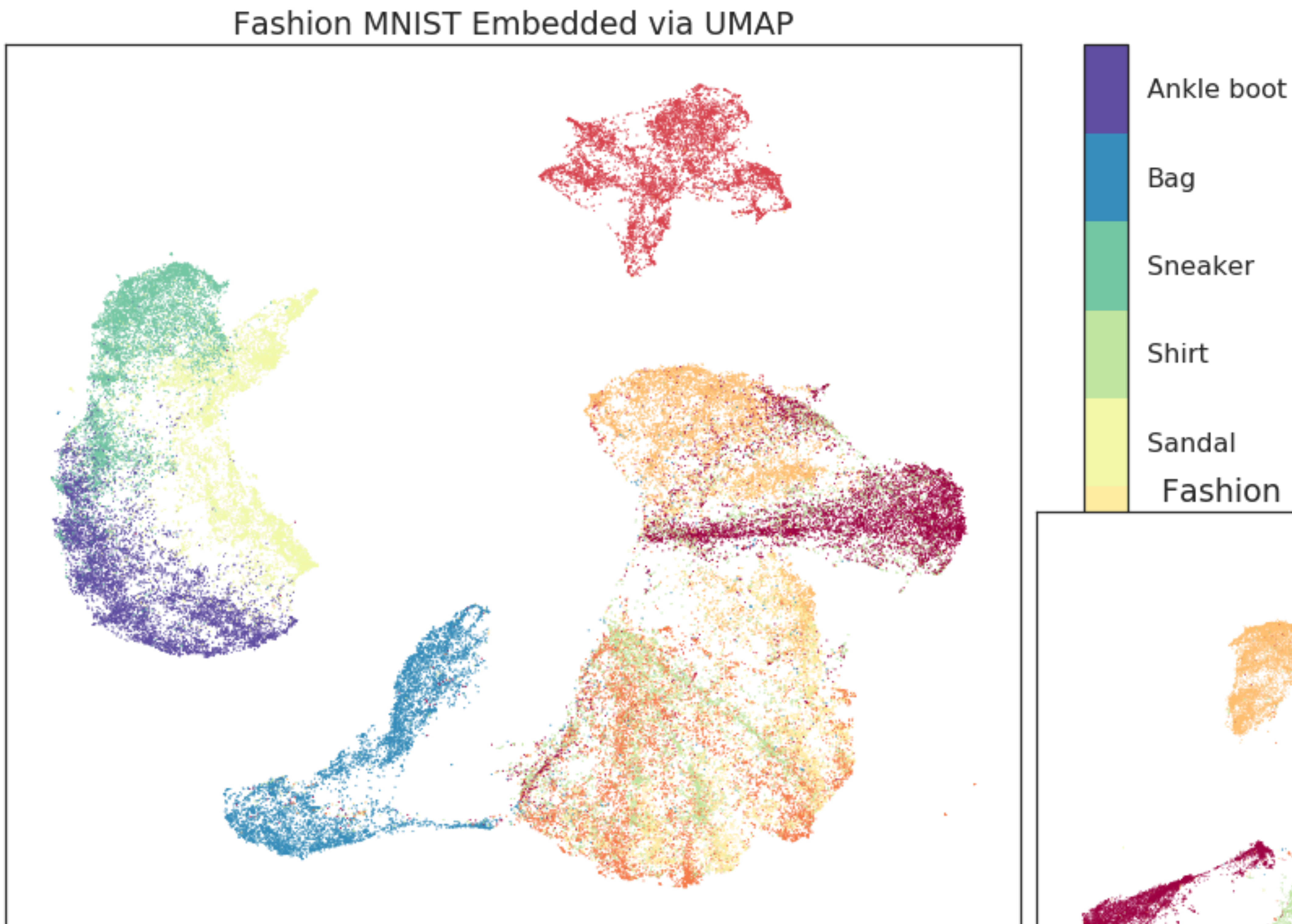
Aim: use of categorical label information to do supervised dimension reduction

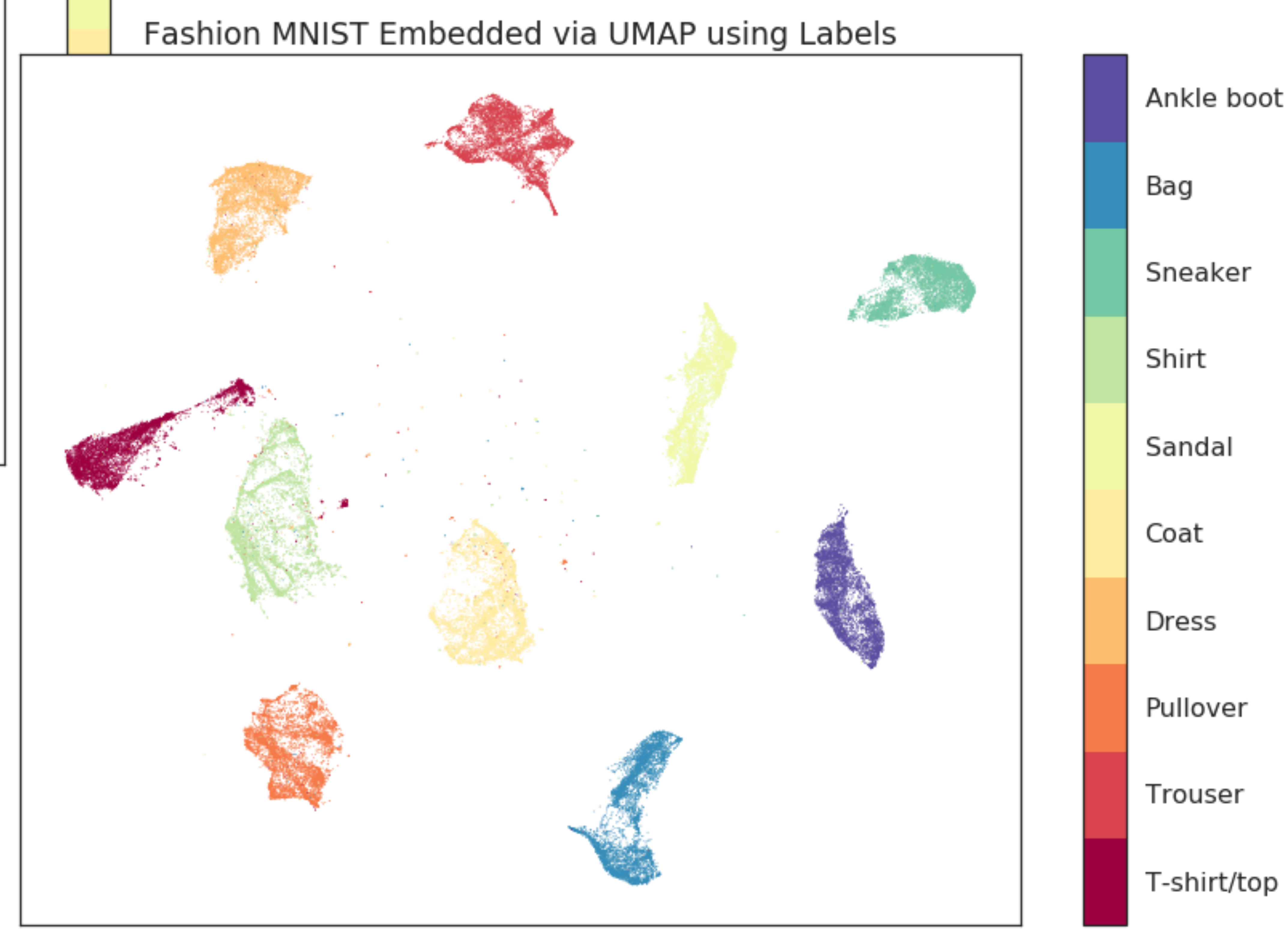
Idea:

- Use labels to learn a new metric space
- Build two representations (one using data and one using labels)
- Merge the two representations

Supervised UMAP

Fashion MNIST



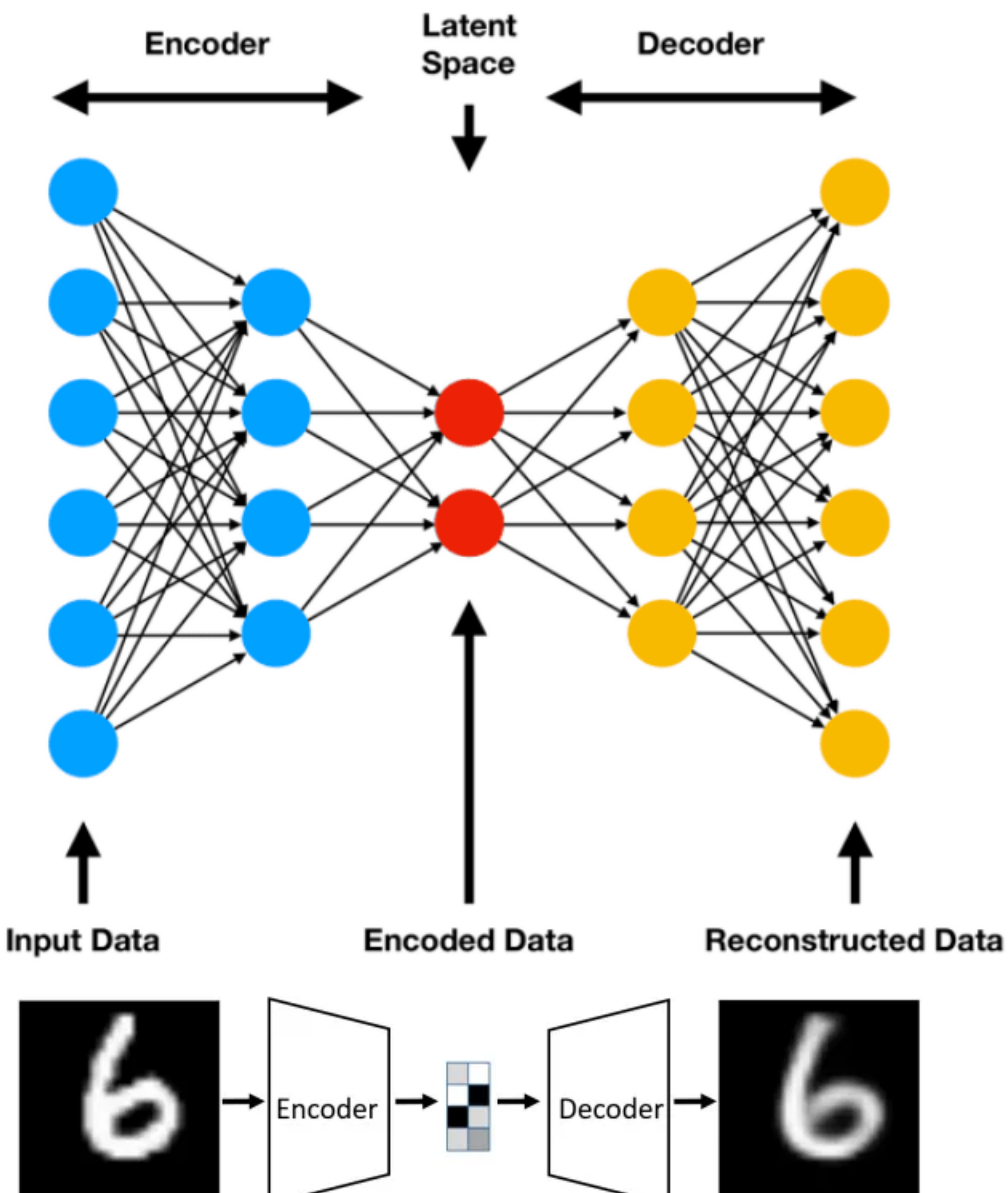


Supervised DR

- Linear Discriminant Analysis
- Neighbourhood Components Analysis
- Supervised UMAP
- Centroid-Encoder

Centroid-Encoder

Aim: add label information to autoencoder architecture



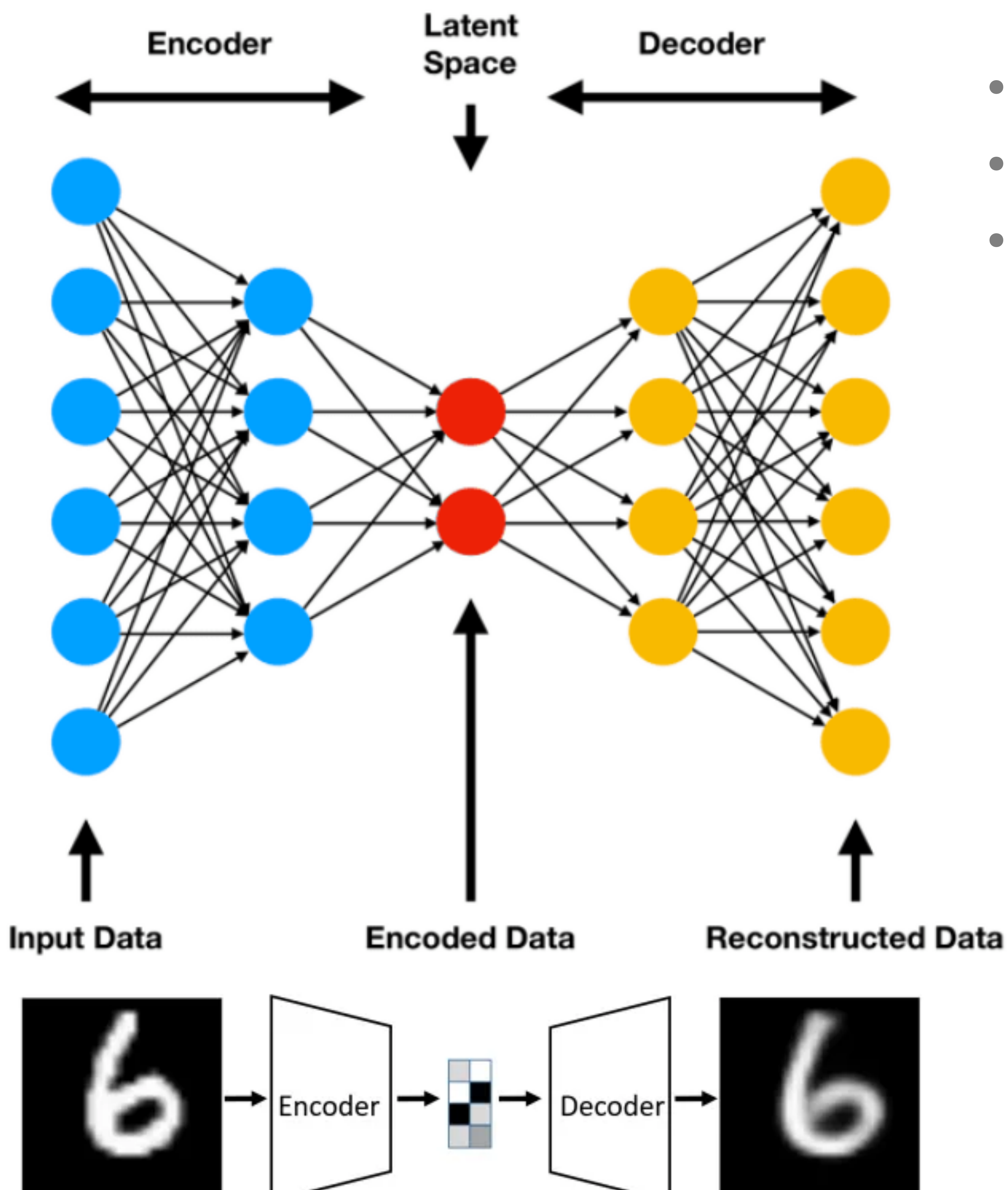
Original AE:
Minimize reconstruction error

Centroid-Encoder Idea:

- Class centroid in original space
- Map samples to output space centroids

Centroid-Encoder

Aim: add label information to autoencoder architecture



- d-dimensional space
- D = {x_1, x_2, …, x_N} data samples
- C number of distinct classes

Class centroid $$c_i = \frac{1}{N_i} \sum_{x_n \in C_i} x_n$$

Input-output pairs $$\{(x_n, c_j)\}_{i=1}^{N}, x_n \in C_j$$

AE loss function

$$L_C(D, \theta) = \frac{1}{2N} \sum_{i=1}^{N} \|x_i - f(x_i, \theta)\|_2^2$$

Centroid-Encoder loss function

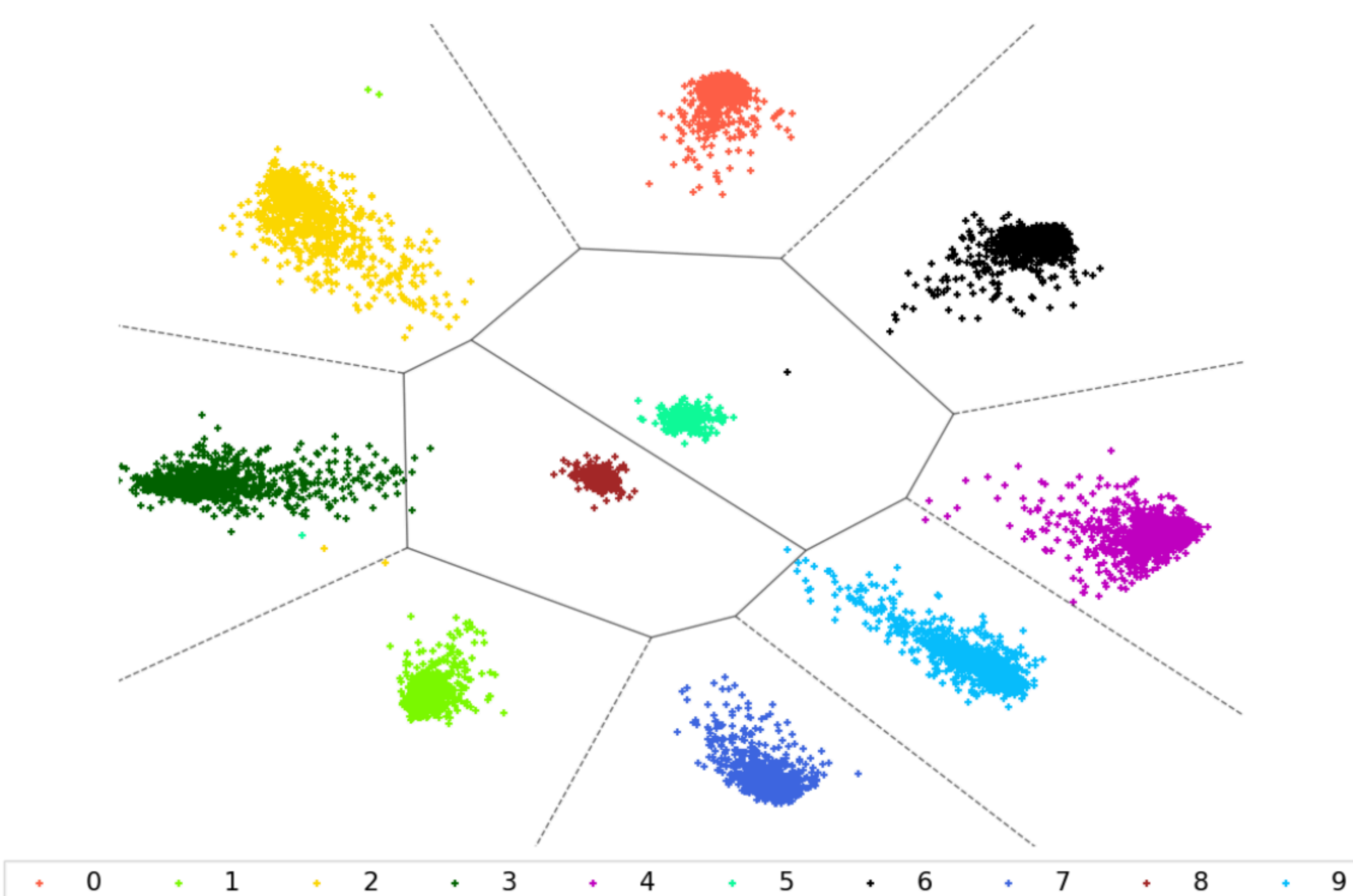
$$L_C(D, \theta) = \frac{1}{2N} \sum_{j=1}^{C} \sum_{x_n \in C_j} \|c_j - f(x_n, \theta)\|_2^2$$

Centroid-Encoder

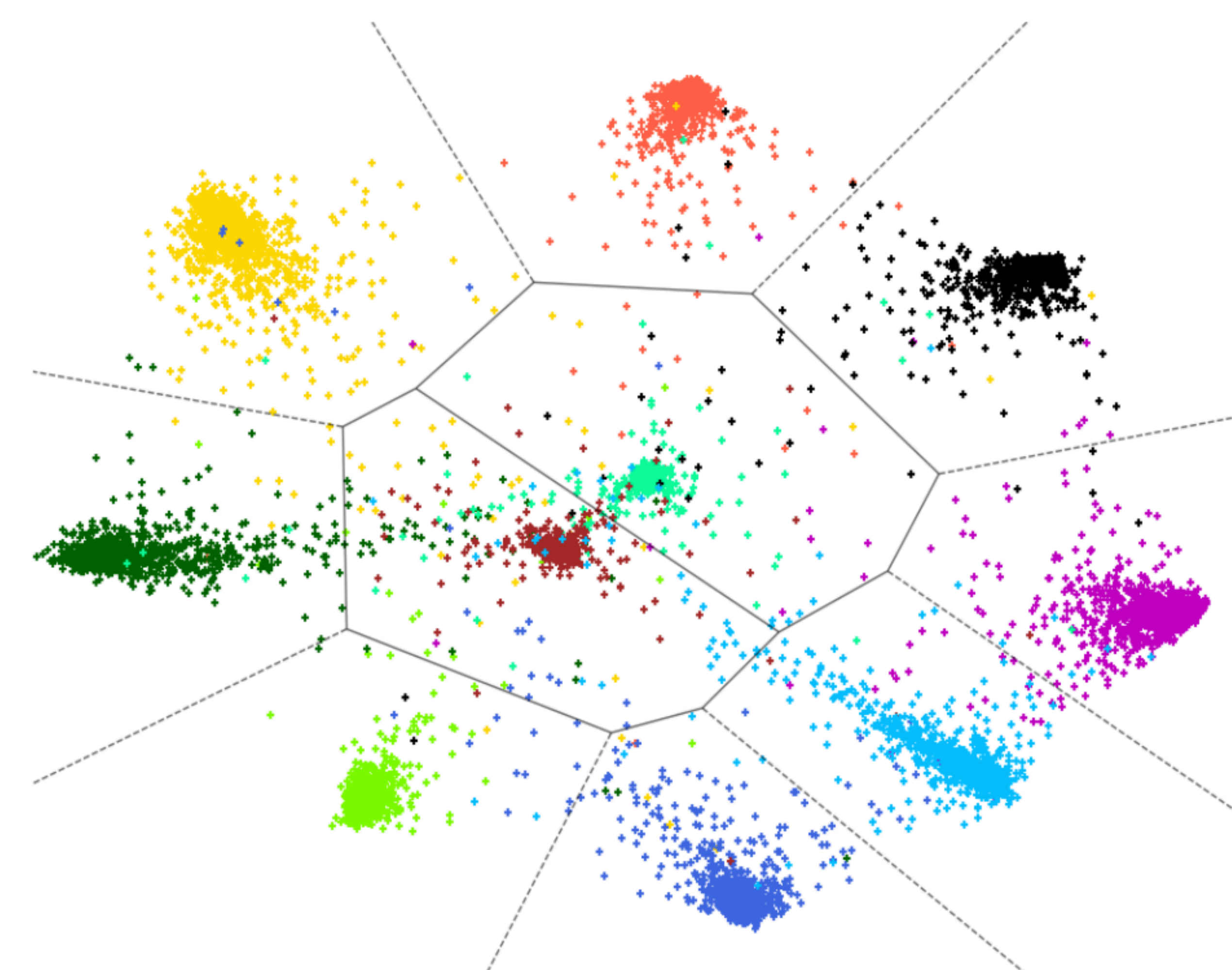
MNIST

Network architecture:
784-1000-500-125-2-125-500-1000-784

Training set



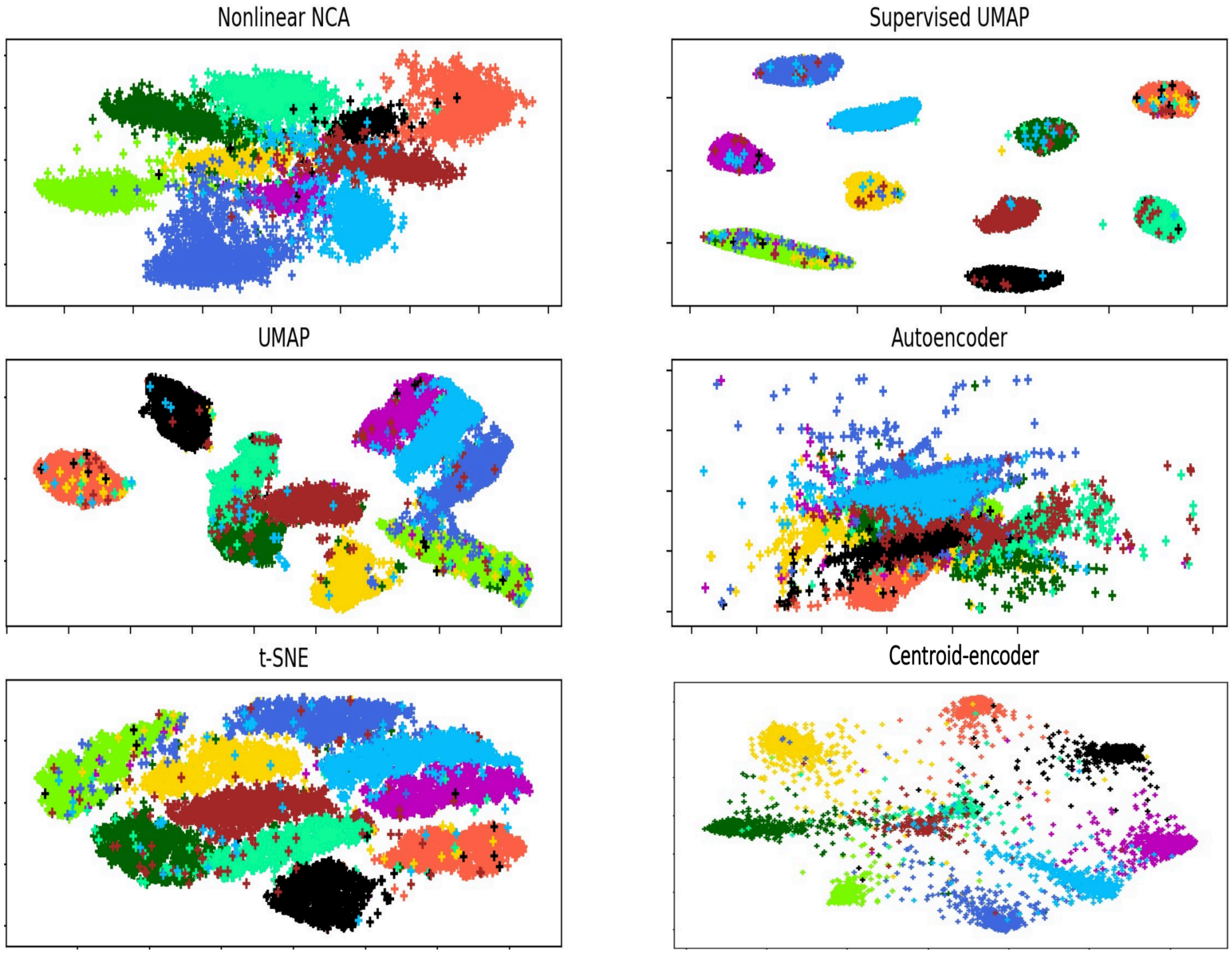
Test set



Centroid-Encoder

MNIST

Network architecture:
784-1000-500-125-2-125-500-1000-784



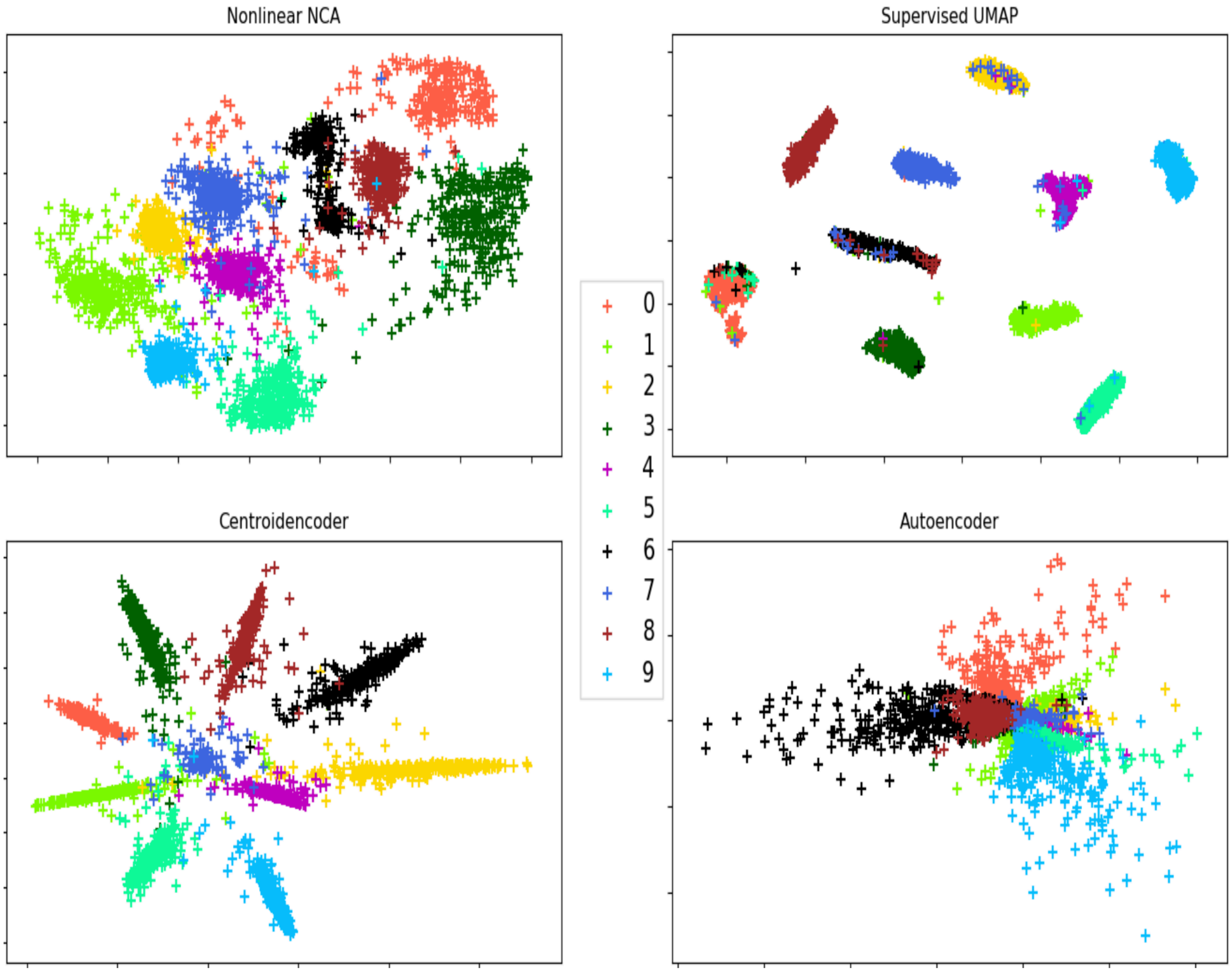
Method	Dataset	
	MNIST	USPS
Centroid-encoder	2.61 ± 0.09	2.91 ± 0.31
Supervised UMAP	3.45 ± 0.03	6.17 ± 0.23
NNCA	4.71 ± 0.57	6.58 ± 0.80
Autoencoder	22.04 ± 0.78	$16.49 \pm .91$
dt-MCML	**2.03**	$\mathbf{2.46 \pm 0.35}$
dG-MCML	2.13	3.37 ± 0.18
GerDA	3.2	NA
dt-NCA	3.48	5.11 ± 0.28
dG-NCA	7.95	10.22 ± 0.76
AAE	4.20	NA
pt-SNE	9.90	NA

Error rate of 5NN classifier in 2D

Centroid-Encoder

USPS

Network architecture:
64-2000-1000-500-2-500-1000-2000-784



Method	Dataset	
	MNIST	USPS
Centroid-encoder	2.61 ± 0.09	2.91 ± 0.31
Supervised UMAP	3.45 ± 0.03	6.17 ± 0.23
NNCA	4.71 ± 0.57	6.58 ± 0.80
Autoencoder	22.04 ± 0.78	$16.49 \pm .91$
dt-MCML	**2.03**	$\mathbf{2.46 \pm 0.35}$
dG-MCML	2.13	3.37 ± 0.18
GerDA	3.2	NA
dt-NCA	3.48	5.11 ± 0.28
dG-NCA	7.95	10.22 ± 0.76
AAE	4.20	NA
pt-SNE	9.90	NA

Error rate of 5NN classifier in 2D

Centroid-Encoder

Non linear

Customizable architecture

Training complexity

Journal of Machine Learning Research 23 (2022) 1-34 Submitted 2/20; Revised 12/21; Published 1/22

Supervised Dimensionality Reduction and Visualization using Centroid-Encoder

Tomojit Ghosh TOMOJIT.GHOSH@COLOSTATE.EDU
Department of Computer Science
Colorado State University
Fort Collins, CO 80523 ,USA

Michael Kirby KIRBY@MATH.COLOSTATE.EDU
Department of Mathematics
Colorado State University
Fort Collins, CO 80523, USA

Link to paper