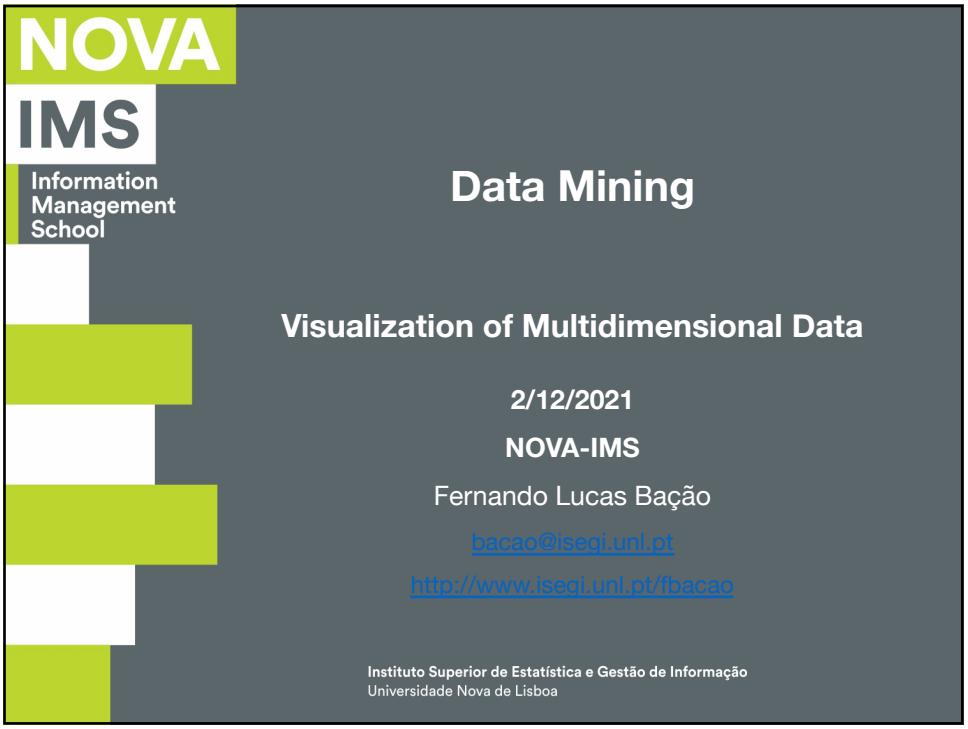
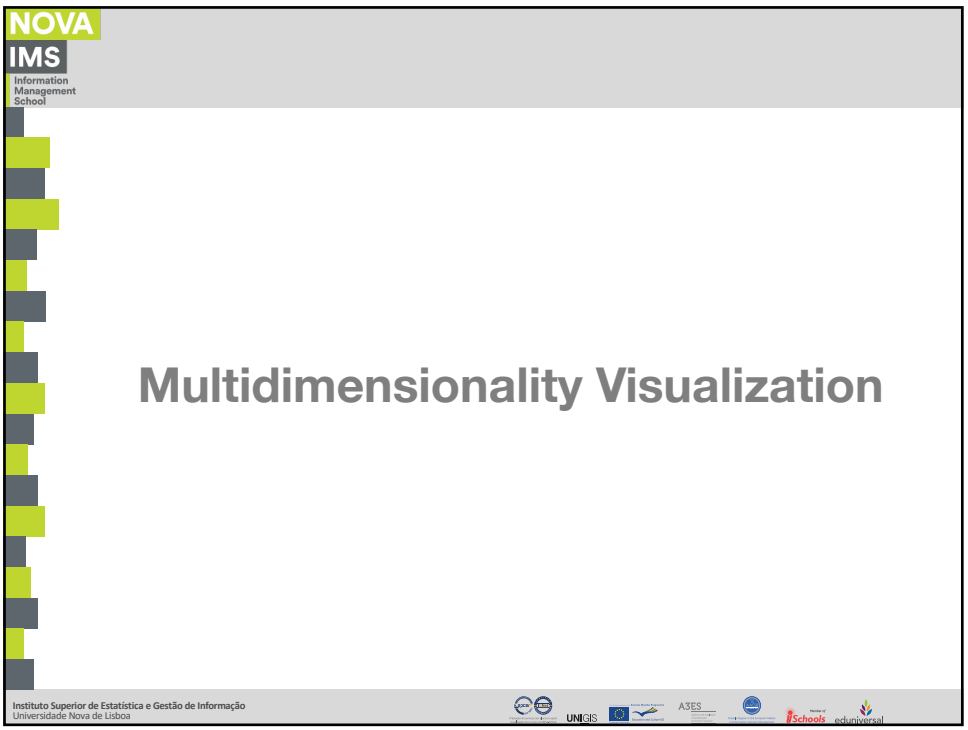
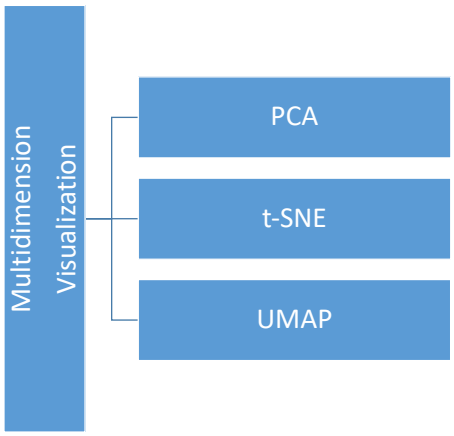

NOVA
IMS
Information
Management
School
Data Mining
Visualization of Multidimensional Data
2/12/2021
NOVA-IMS
Fernando Lucas Bação
bacao@isegi.unl.pt
http://www.isegi.unl.pt/fbacao
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1


NOVA
IMS
Information
Management
School
Multidimensionality Visualization
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2

Multidimensionality Visualization



3

Principal Component Analysis

PCA

4

NOVA IMS Information Management School

Principal Component Analysis

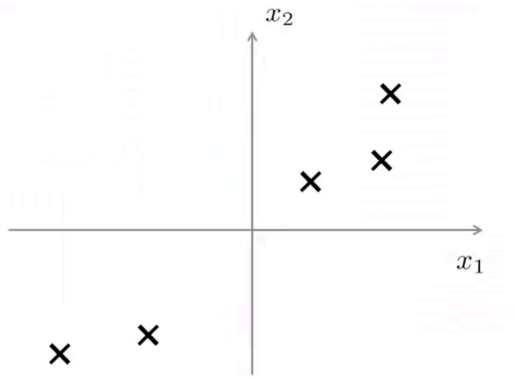
- **Size Reduction of the Input Space:**
  - Principal Component Analysis
    - A procedure that uses an **orthogonal transformation** to convert a set of observations of possibly correlated variables into a set of values of **linearly uncorrelated variables called principal components**.
    - The number of principal components is **equal to the number of original variables**.
    - This transformation is defined in such a way that the first principal component has the **largest possible variance** (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance.

5

NOVA IMS Information Management School

Principal Component Analysis

- **Size Reduction of the Input Space:**
  - Principal Component Analysis

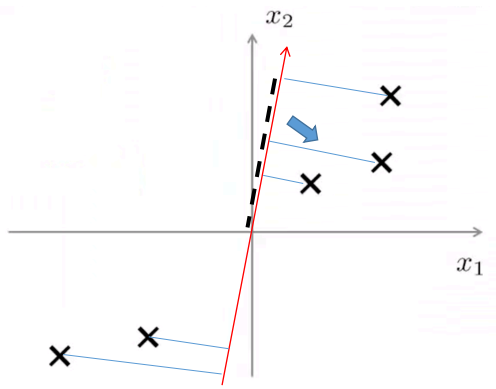


6

NOVA IMS Information Management School

Principal Component Analysis

- **Size Reduction of the Input Space:**
  - Principal Component Analysis (SVD)

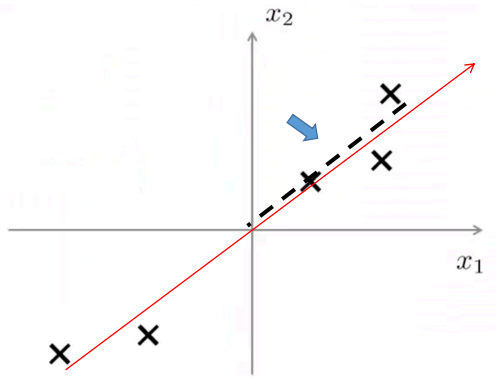


7

NOVA IMS Information Management School

Principal Component Analysis

- **Size Reduction of the Input Space:**
  - Principal Component Analysis (SVD)

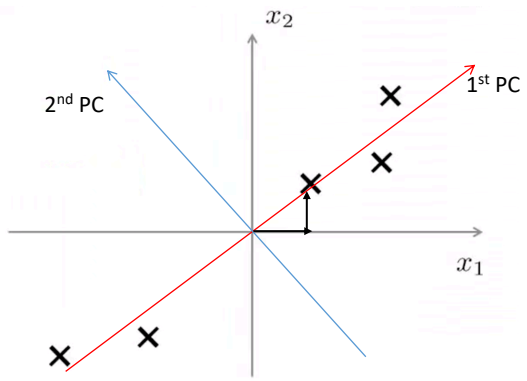


8

NOVA IMS Information Management School

Principal Component Analysis

- **Size Reduction of the Input Space:**
  - Principal Component Analysis (SVD)

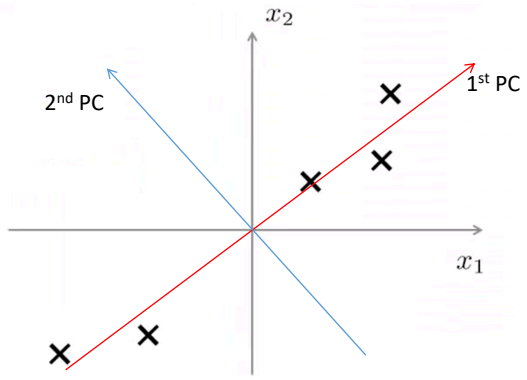


9

NOVA IMS Information Management School

Principal Component Analysis

- **Size Reduction of the Input Space:**
  - Principal Component Analysis (SVD)

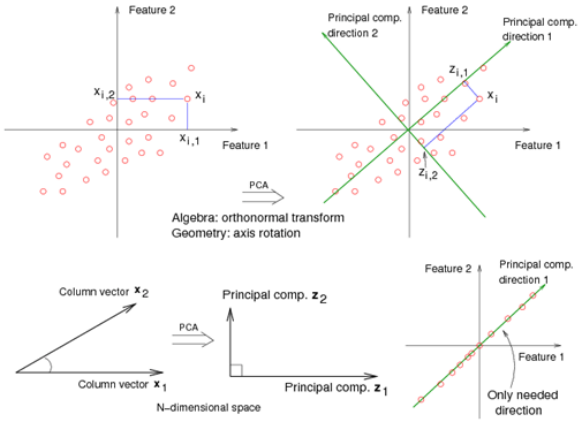


10

NOVA IMS Information Management School

Principal Component Analysis

- **Size Reduction of the Input Space:**
  - Principal Component Analysis

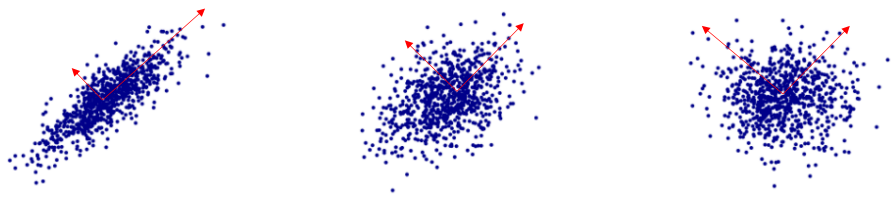


11

NOVA IMS Information Management School

Principal Component Analysis

- **Size Reduction of the Input Space:**
  - Principal Component Analysis



12

t-distributed Stochastic Neighbor Embedding

t-SNE

13

- **t-SNE:**
  - t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by **giving each datapoint a location in a two or three-dimensional map**.
  - It is a **nonlinear dimensionality reduction technique** well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions.
  - Specifically, it **models each high-dimensional object** by a two- or three-dimensional point in such a way that **similar objects are modeled by nearby points** and dissimilar objects are modeled by distant points with high probability.

14

NOVA IMS Information Management School

t-SNE

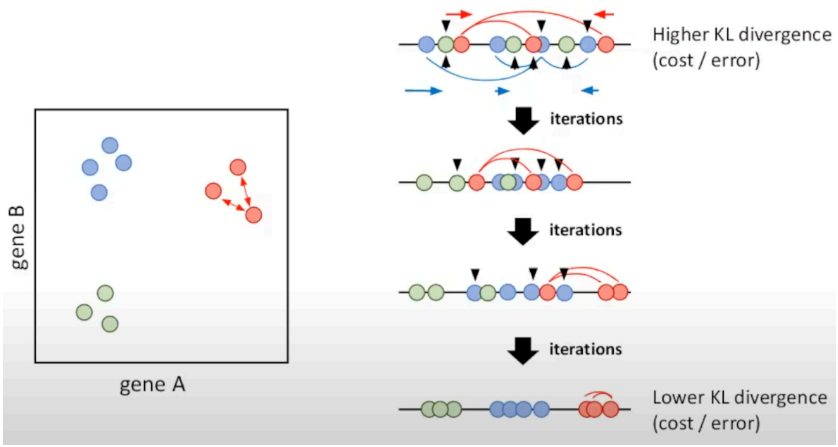
- **t-SNE:**
  - The t-SNE algorithm comprises two stages:
    - First, t-SNE **constructs a probability distribution over pairs of high-dimensional objects** in such a way that similar objects are assigned a higher probability while dissimilar points are assigned a lower probability.
    - Second, t-SNE defines a **similar probability distribution over the points in the low-dimensional map**, and it minimizes the Kullback–Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map.
  - While the original algorithm uses the Euclidean distance between objects as the base of its similarity metric, this can be changed as appropriate.

15

NOVA IMS Information Management School

t-SNE

- **t-SNE:**

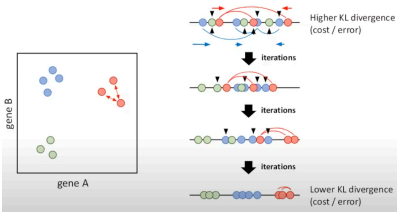


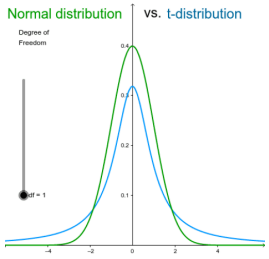
16

NOVA IMS Information Management School

t-SNE

- **t-SNE:**
  - Optimizing the KL divergence is a measure of how one probability distribution is different from a second, reference probability distribution





17

NOVA IMS Information Management School

t-SNE

- **t-SNE:**
  - t-SNE has been used for **visualization in a wide range of applications**, including genomics, computer security research, natural language processing, music analysis, cancer research, bioinformatics, etc
  - While t-SNE plots often seem to display clusters, **the visual clusters can be influenced strongly by the chosen parameterization** and therefore a good understanding of the parameters for t-SNE is necessary.
  - Interactive exploration may thus be necessary to choose parameters and validate results.

18

NOVA IMS Information Management School

t-SNE

- **t-SNE:**
  - Perplexity is the main t-SNE parameter, "perplexity,"
    - Basically defines how to balance attention between local and global aspects of your data.
    - Perplexity is a measure for information that is defined as 2 to the power of the Shannon entropy.
    - In t-SNE, the perplexity may be viewed as a knob that sets the number of effective nearest neighbors.
    - It is comparable with the number of nearest neighbors k that is employed in many manifold learners.
    - The original paper says, "The performance of SNE is fairly robust to changes in the perplexity, and typical values are between 5 and 50."

19

NOVA IMS Information Management School

t-SNE

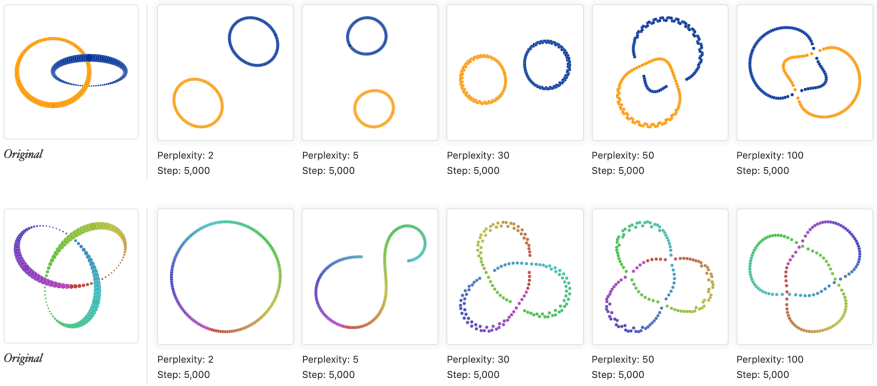
- **t-SNE:**
  - Perplexity is the main t-SNE parameter, "perplexity,"

20

NOVA IMS Information Management School

t-SNE

- **t-SNE:**
  - Perplexity is the main t-SNE parameter, "perplexity,"



21

NOVA IMS Information Management School

Uniform Manifold Approximation and Projection

UMAP

22

NOVA IMS Information Management School

UMAP

- **UMAP:**
  - Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE.
  - The algorithm is founded on three assumptions about the data
    - The data is uniformly distributed on Riemannian manifold (a topological space that locally resembles Euclidean space near each point);
    - The manifold is locally connected.
    - From these assumptions it is possible to model the manifold with a fuzzy topological structure. The embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure.
    - For a more detailed explanation see https://youtu.be/nq6iPZVUxZU

23

NOVA IMS Information Management School

UMAP

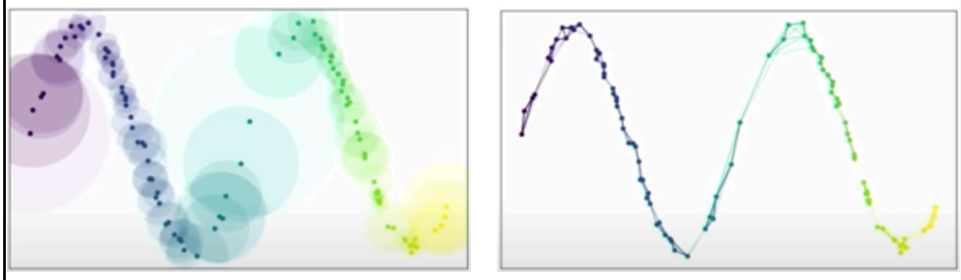
- **UMAP:**
  - Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE.
  - Two phases in the UMAP algorithm
    - The first phase consists of constructing a fuzzy topological representation in the original space;
    - The second phase is simply optimizing (through stochastic gradient descent) the low dimensional representation to have as close a fuzzy topological representation as possible as measured by cross entropy.

24

NOVA IMS Information Management School

UMAP

- **UMAP:**
  - The first phase consists of constructing a fuzzy topological representation;



25

NOVA IMS Information Management School

UMAP

- **UMAP:**
  - The second phase is simply optimizing (through stochastic gradient descent) the low dimensional representation to have as close a fuzzy topological representation as possible as measured by cross entropy.

Get the clumps right

$$\sum_{a \in A} \mu(a) \log\left(\frac{\mu(a)}{\nu(a)}\right) + (1-\mu(a)) \log\left(\frac{1-\mu(a)}{1-\nu(a)}\right)$$

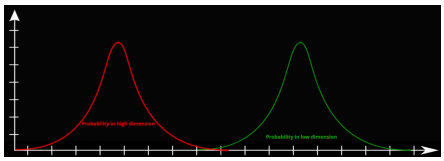
Get the gaps right

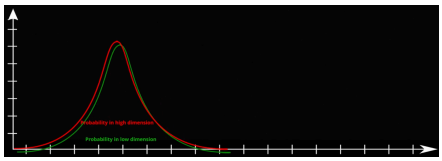
26

NOVA IMS Information Management School

UMAP

- **UMAP:**
  - The second phase is simply optimizing (through stochastic gradient descent) the low dimensional representation to have as close a fuzzy topological representation as possible as measured by cross entropy.





27

NOVA IMS Information Management School

UMAP

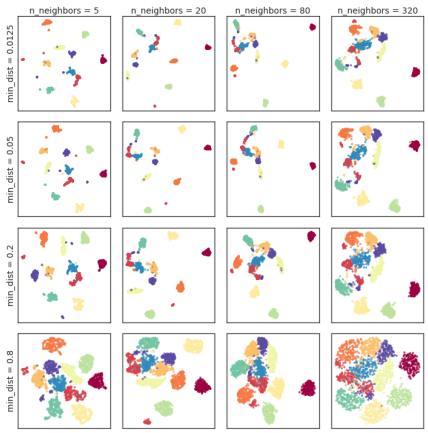
- **UMAP:**
  - Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE.
  - Parameters
    - Number of nearest neighbors - controls how UMAP balances local versus global structure in the data;
    - Minimum distance - controls how tightly UMAP is allowed to pack points together. It provides the minimum distance apart that points are allowed to be in the low dimensional representation

28

NOVA IMS Information Management School

UMAP

- **UMAP:**



29

NOVA IMS Information Management School

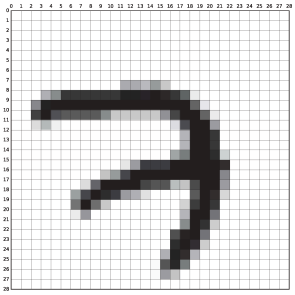
Comparison Between the 3 Methods

30

NOVA IMS Information Management School

Multidimensionality Visualization

- **Mnist Digits:**
  - 28x28 image (784 dimensions)
  - grayscale images of handwritten single digits
  - 60,000 examples, and a test set of 10,000 examples



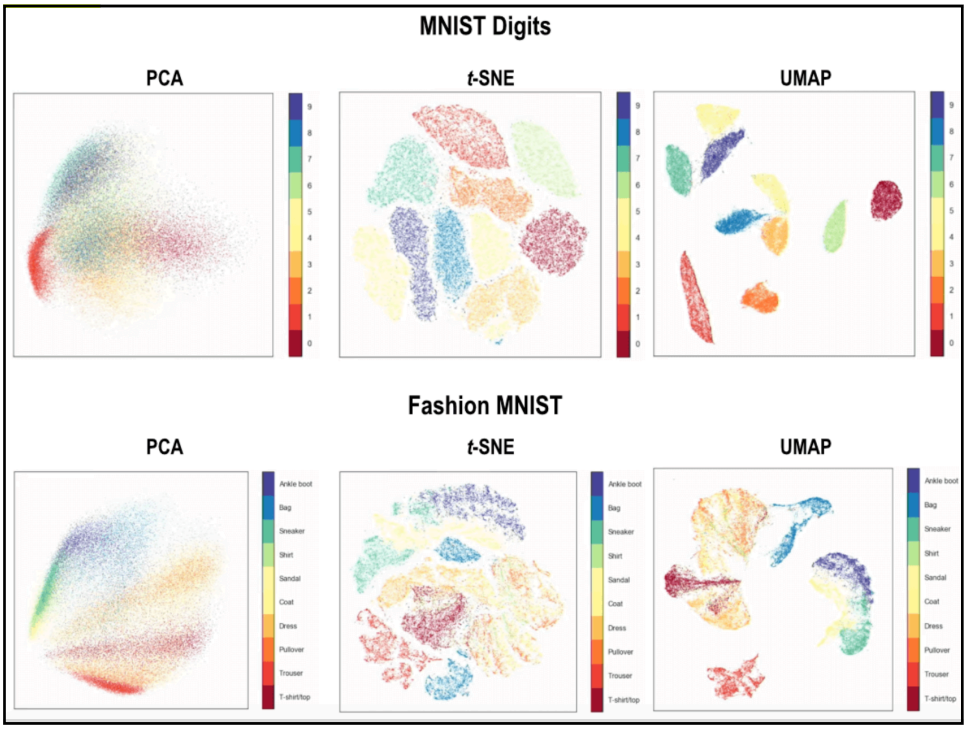
(a) MNIST sample belonging to the digit '7'.

(b) 100 samples from the MNIST training set.

NOVA IMS Information Management School

Multidimensionality Visualization

- **Mnist Fashion:**
  - 28x28 image (784 dimensions)
  - grayscale images of handwritten single digits
  - The dataset has 60,000 images



33

NOVA IMS Information Management School

Multidimensionality Visualization

- **Mnist Digits:**
  - 3D woolly mammoth skeleton
  - https://pair-code.github.io/understanding-umap/

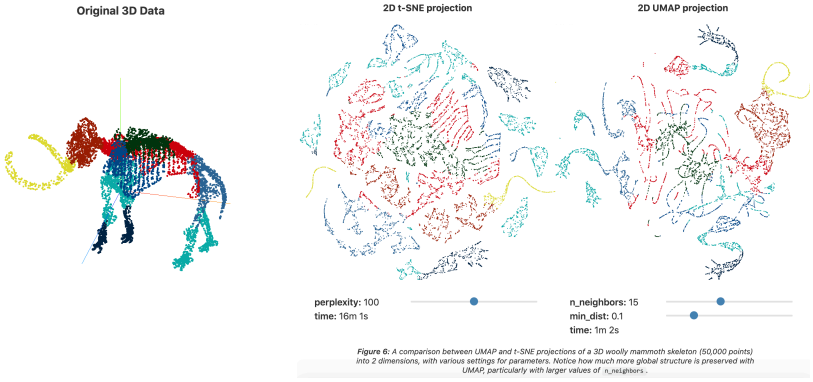


Figure 6: A comparison between UMAP and t-SNE projections of a 3D woolly mammoth skeleton (50,000 points) into 2 dimensions, with various settings for parameters. Notice how much more global structure is preserved with UMAP, particularly with larger values of `n_neighbors`.

34

NOVA IMS Information Management School

UMAP

- **UMAP:**

	t-SNE	UMAP
COIL20	20 seconds	7 seconds
MNIST	22 minutes	98 seconds
Fashion MNIST	15 minutes	78 seconds
GoogleNews	4.5 hours	14 minutes

35

Questions?

36