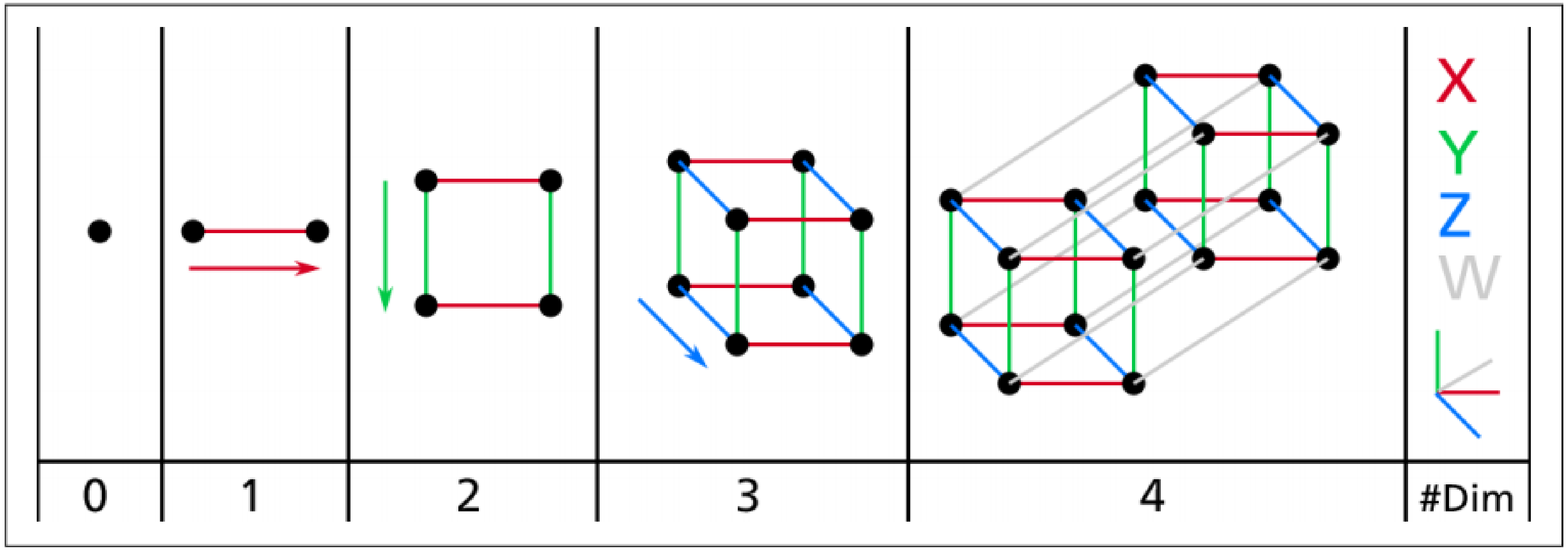
ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG VIỆT - HÀN
VIETNAM - KOREA UNIVERSITY OF INFORMATION AND COMMUNICATION TECHNOLOGY
한-베정보통신기술대학교
Nhân bản – Phụng sự – Khai phóng

Chapter 7

Dimensionality Reduction

Machine Learning

- **Data have many features**
  - Training extremely slow,
  - Harder to find a good solution.
- **This problem is often referred to as the curse of dimensionality**
- **Possible to reduce the number of features**



- **The Curse of Dimensionality**

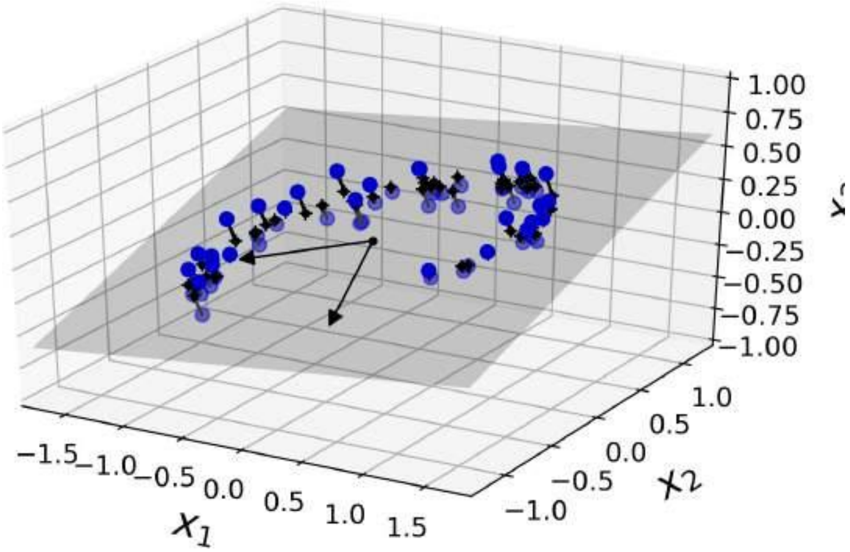
  - Handling the high-dimensional data is very difficult in practice, commonly known as the curse of dimensionality.

  - If the dimensionality of the input dataset increases, any machine learning algorithm and model becomes more complex.

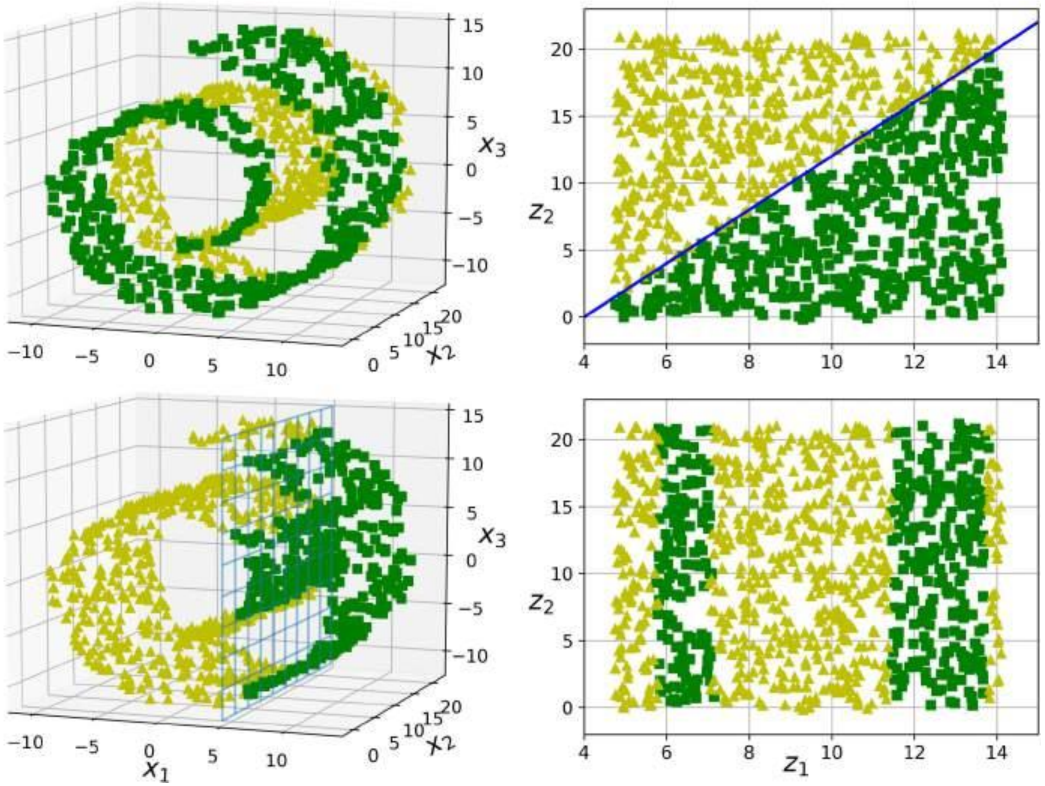
- **What is Dimensionality Reduction?**
  - Dimensionality reduction means reducing feature
  - is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information

- **Why Dimensionality Reduction is Important?**
  - Few features mean less complexity
  - Less storage space because you have fewer data
  - Fewer features require less computation time
  - Model accuracy improves due to less misleading data
  - Algorithms train faster
  - Reducing the data set's feature dimensions helps visualize the data faster
  - It removes noise and redundant features

- **Two main approaches to reducing dimensionality:**
  - Projection (phép chiếu)
  - Manifold Learning (Học đa tạp)

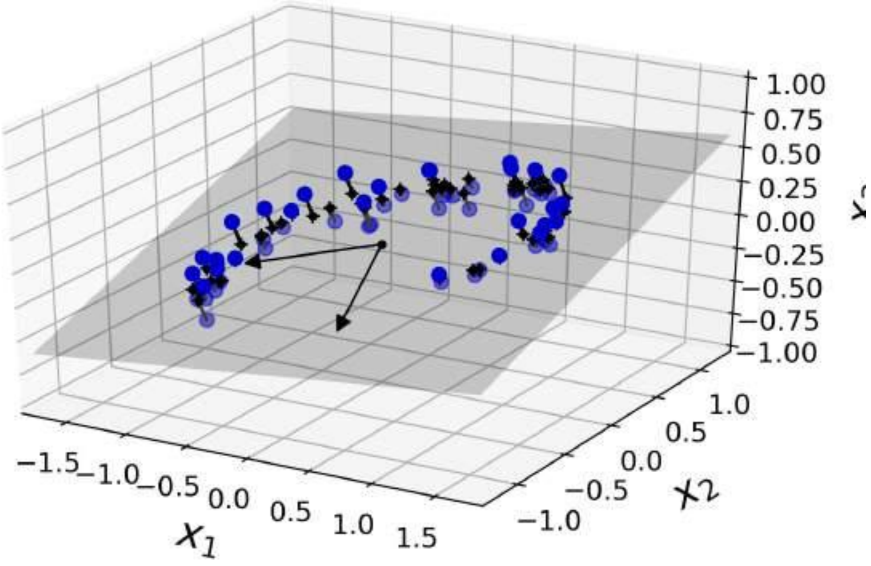


A 3D dataset lying close to a 2D subspace

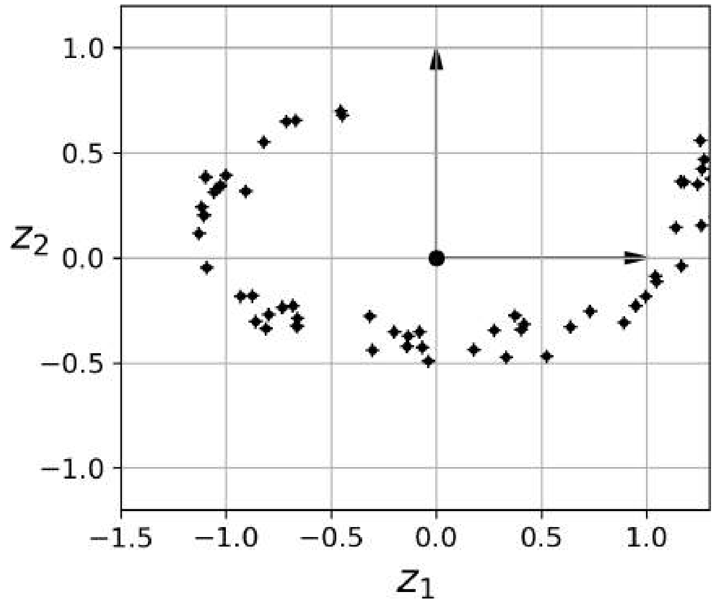


The decision boundary may not always be simpler with lower dimensions

- **Projection** (phép chiếu)

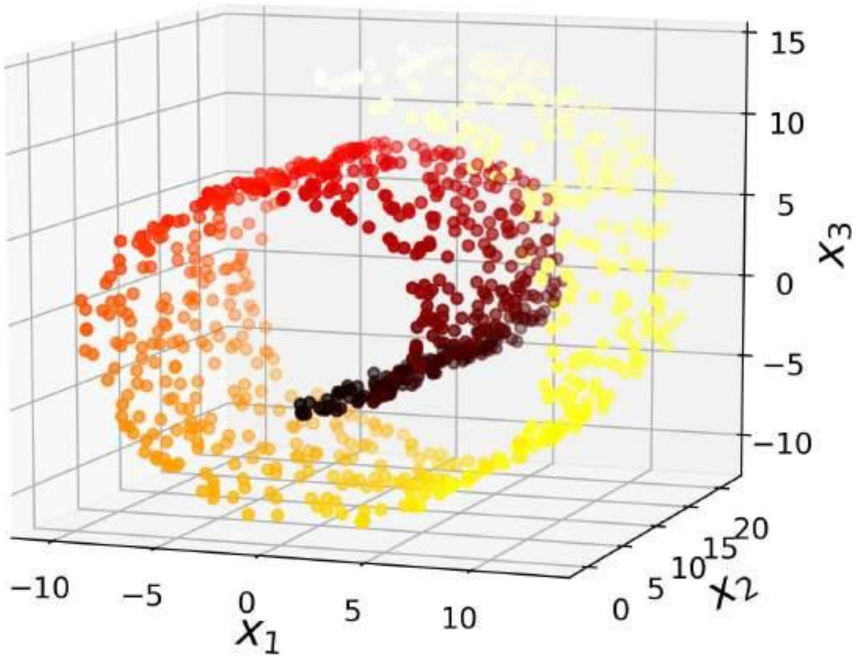


A 3D dataset lying close to a 2D subspace

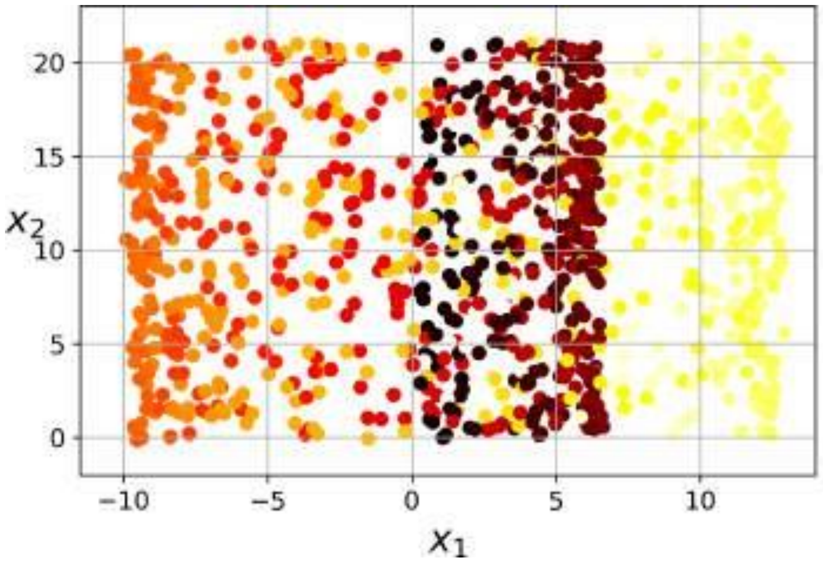


The new 2D dataset after projection

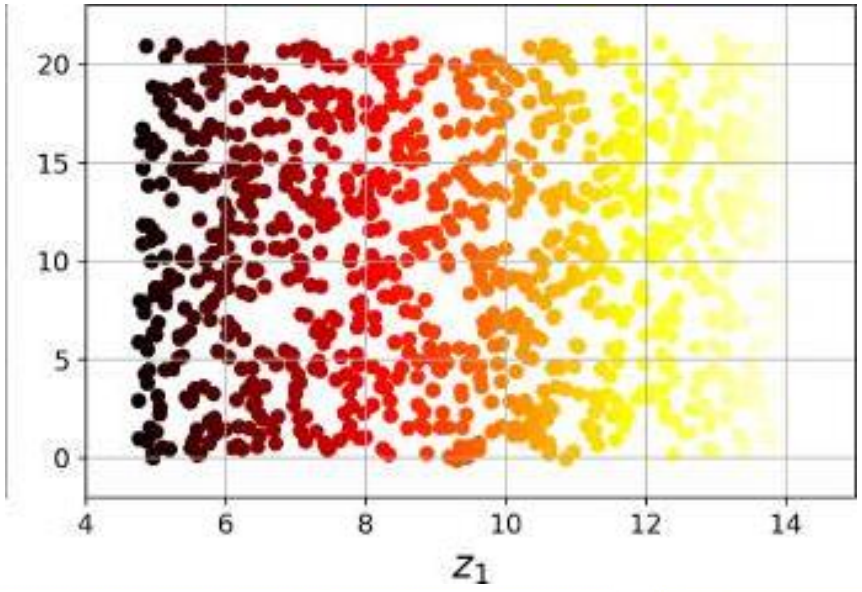
- **Projection** (phép chiếu)



Swiss roll dataset

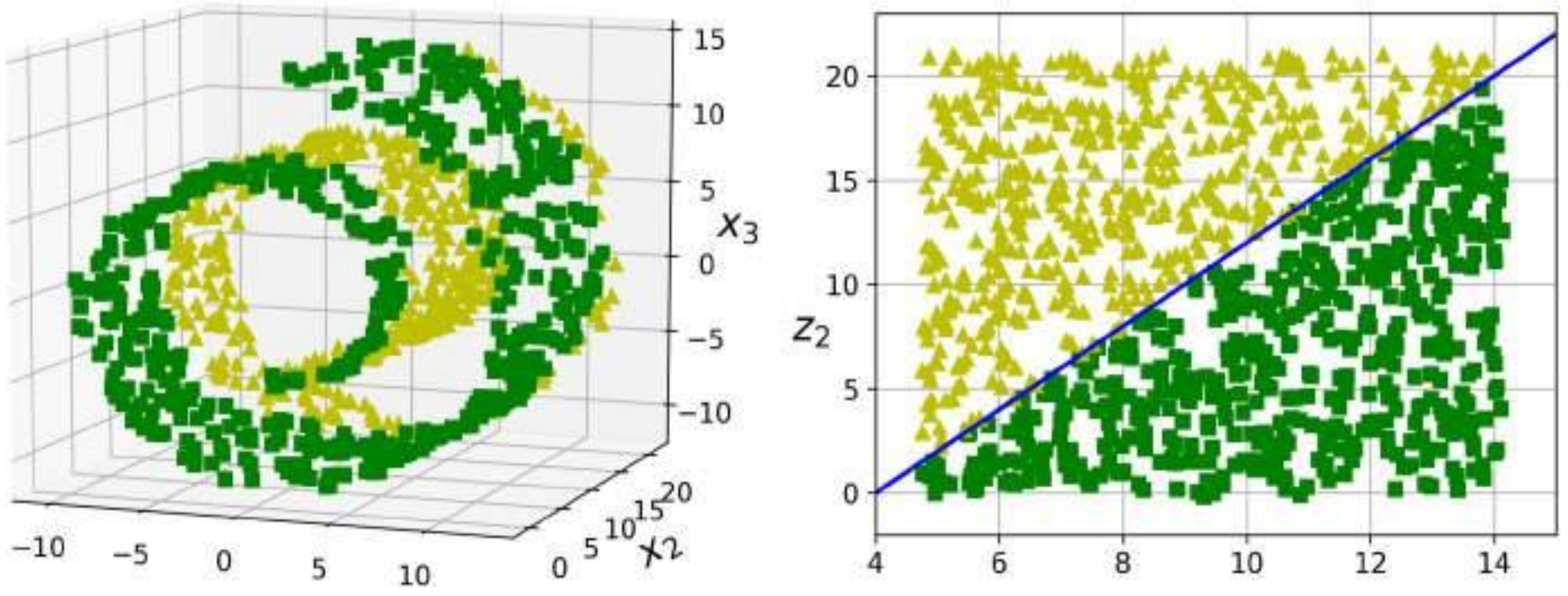


Squashing *by projecting onto a plane*



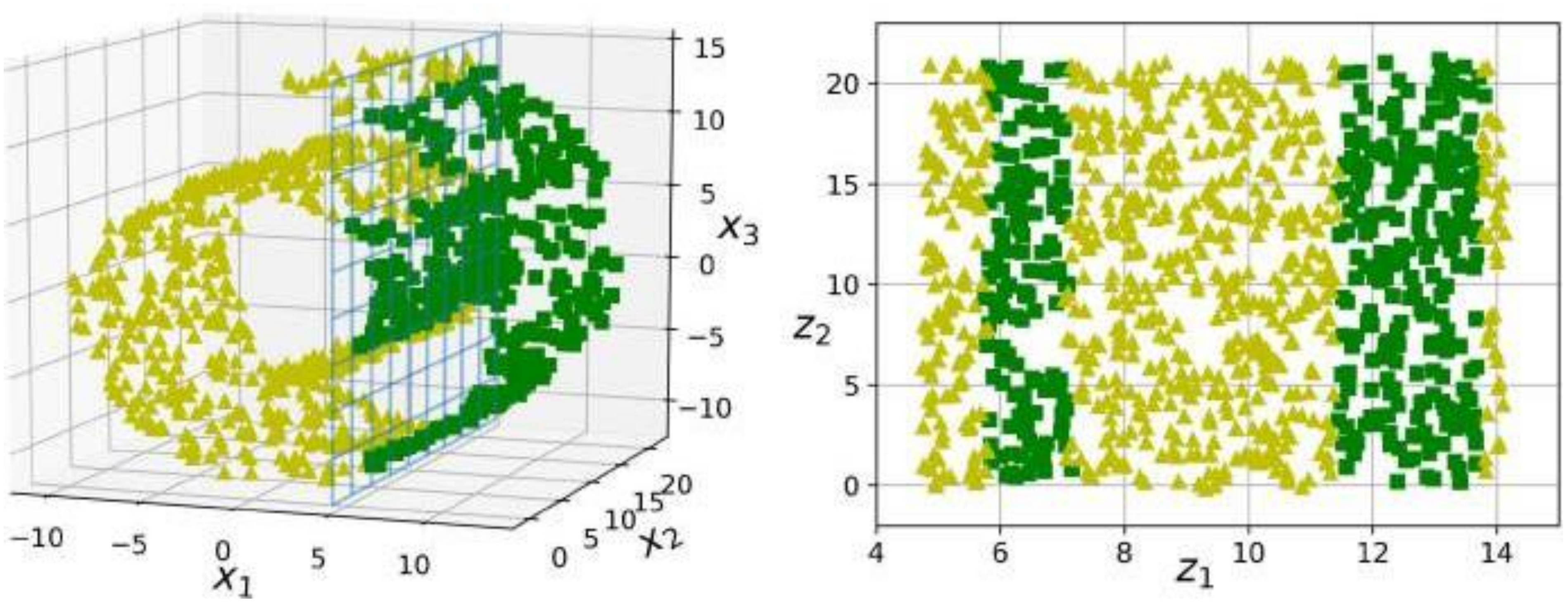
unrolling *the Swiss roll*

- **Manifold Learning** (Học đa tạp)



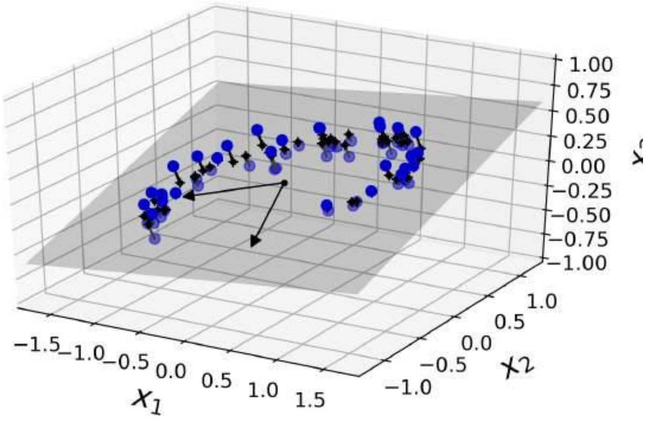
the Swiss roll is split into two classes:
in the 3D space (on the left), the decision boundary would be fairly complex,
but in the 2D unrolled manifold space (on the right), the decision boundary is a simple straight line.

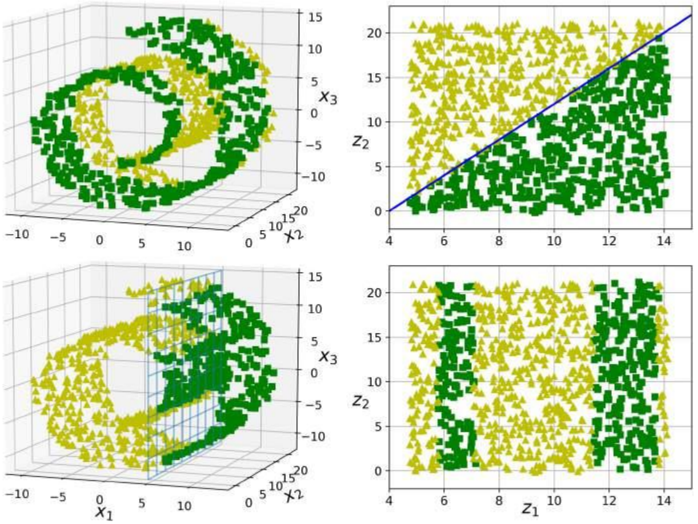
- **Manifold Learning** (Học đa tạp)



the decision boundary is located at x = 5.
This decision boundary looks very simple in the original 3D space (a vertical plane),
but it looks more complex in the unrolled manifold (a collection of four independent line segments).

- **Two main approaches to reducing dimensionality:**
  - Projection (phép chiếu)
  - Manifold Learning (Học đa tạp)





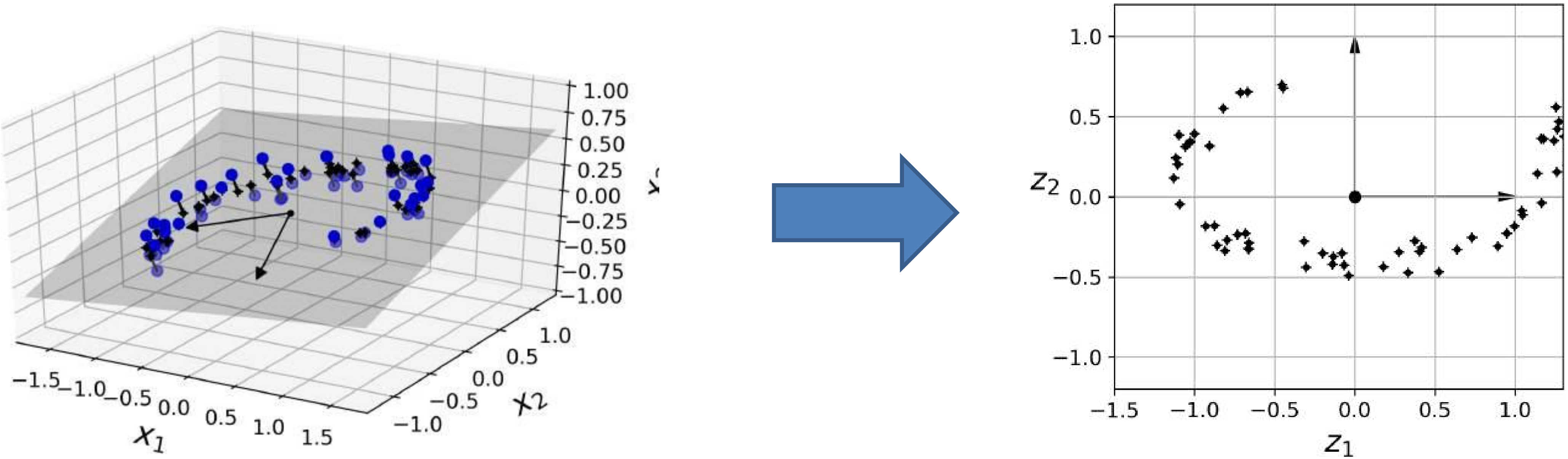
reduce the dimensionality ⇨ usually speed up training,
but it may not always lead to a better or simpler solution;
⇨ **it all depends on the dataset**

The most popular dimensionality reduction techniques

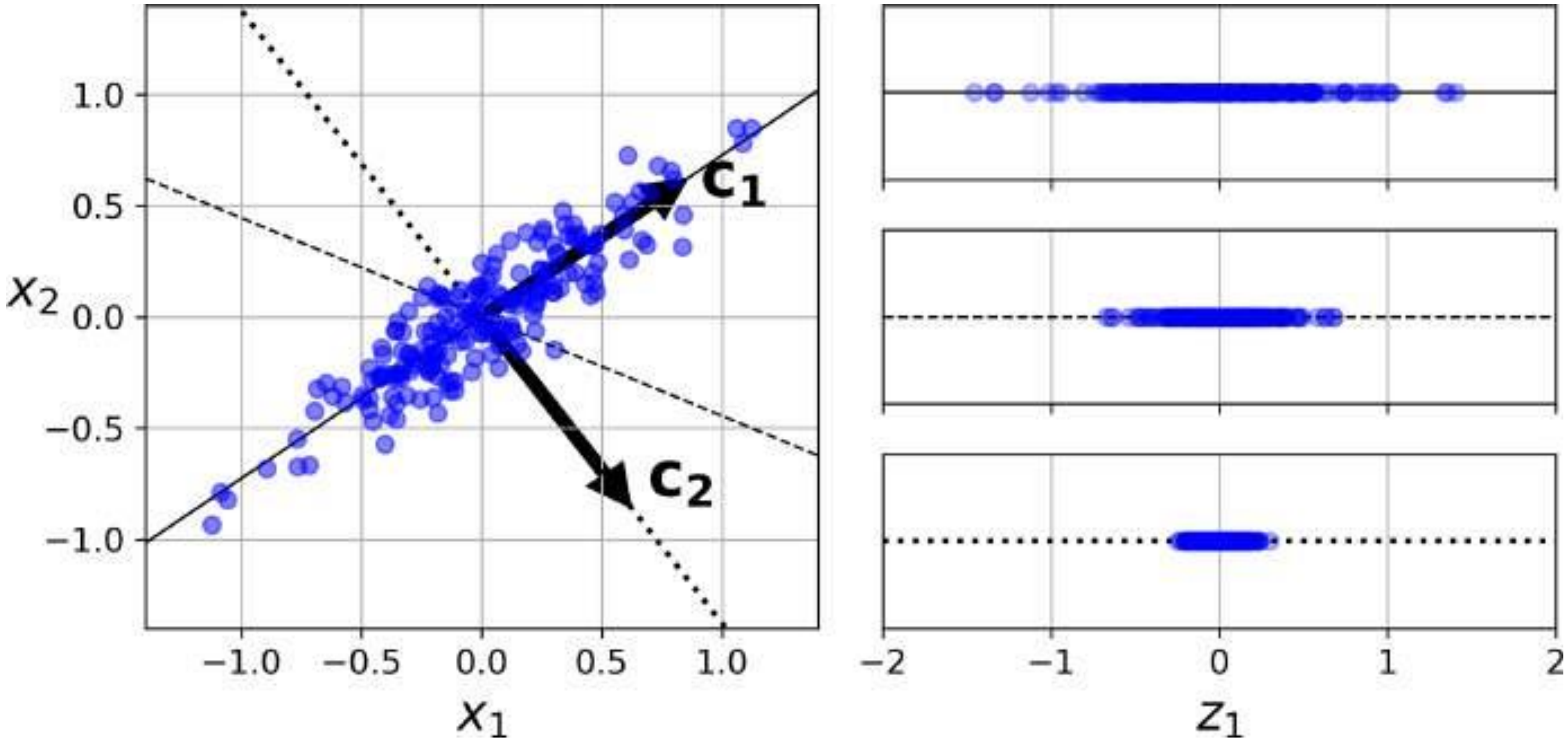
- **Three of the most popular dimensionality reduction techniques:**
  - PCA (Principal Component Analysis – *Phân tích tích thành phần chính*)
  - Kernel PCA
  - LLE (Locally Linear Embedding - Embedding tuyến tính cục bộ)

The most popular dimensionality reduction techniques

- **PCA** (Principal Component Analysis – *Phân tích tích thành phần chính*)
  - First it identifies the hyperplane that lies closest to the data,
  - and then it projects the data onto it

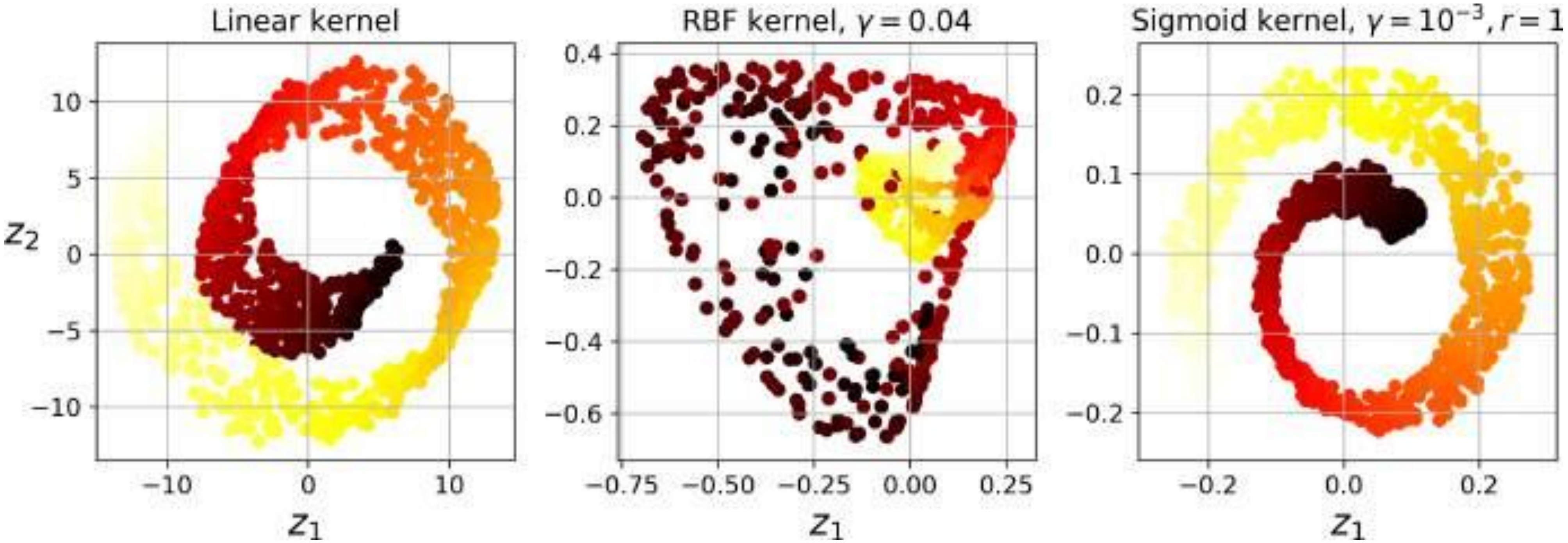


- **PCA** (Principal Component Analysis – *Phân tích tích thành phần chính*)
  - First it identifies the hyperplane that lies closest to the data,
  - and then it projects the data onto it



The most popular dimensionality reduction techniques

- Kernel PCA
  - making it possible to perform complex nonlinear projections for dimensionality reduction

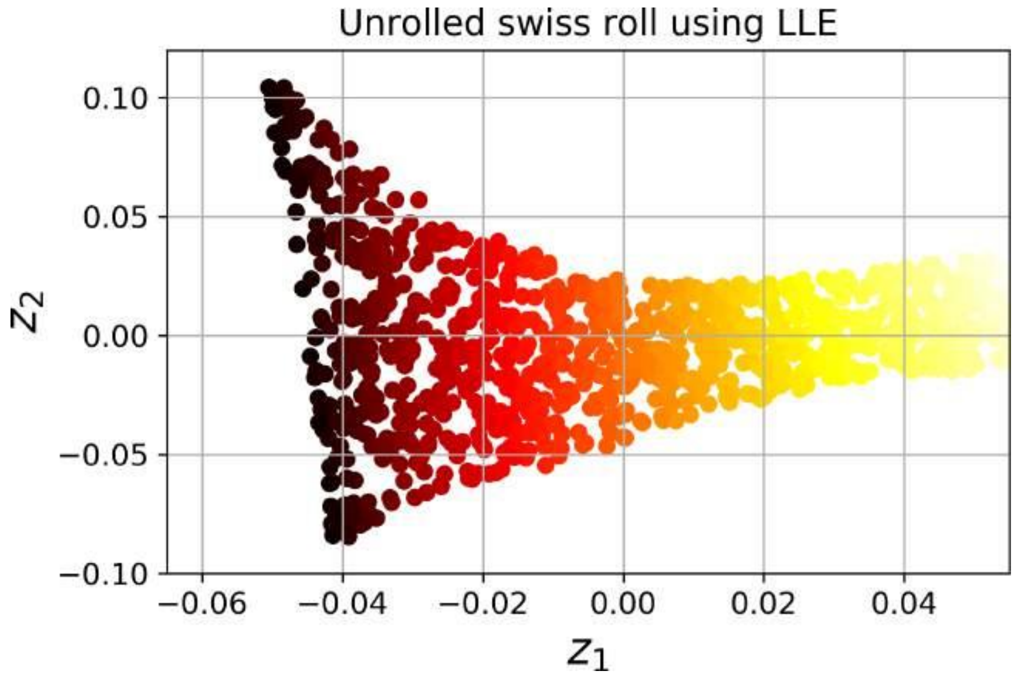


the Swiss roll, reduced to two dimensions using a linear kernel
(equivalent to simply using the PCA class), an RBF kernel, and a sigmoid kernel (Logistic).

The most popular dimensionality reduction techniques

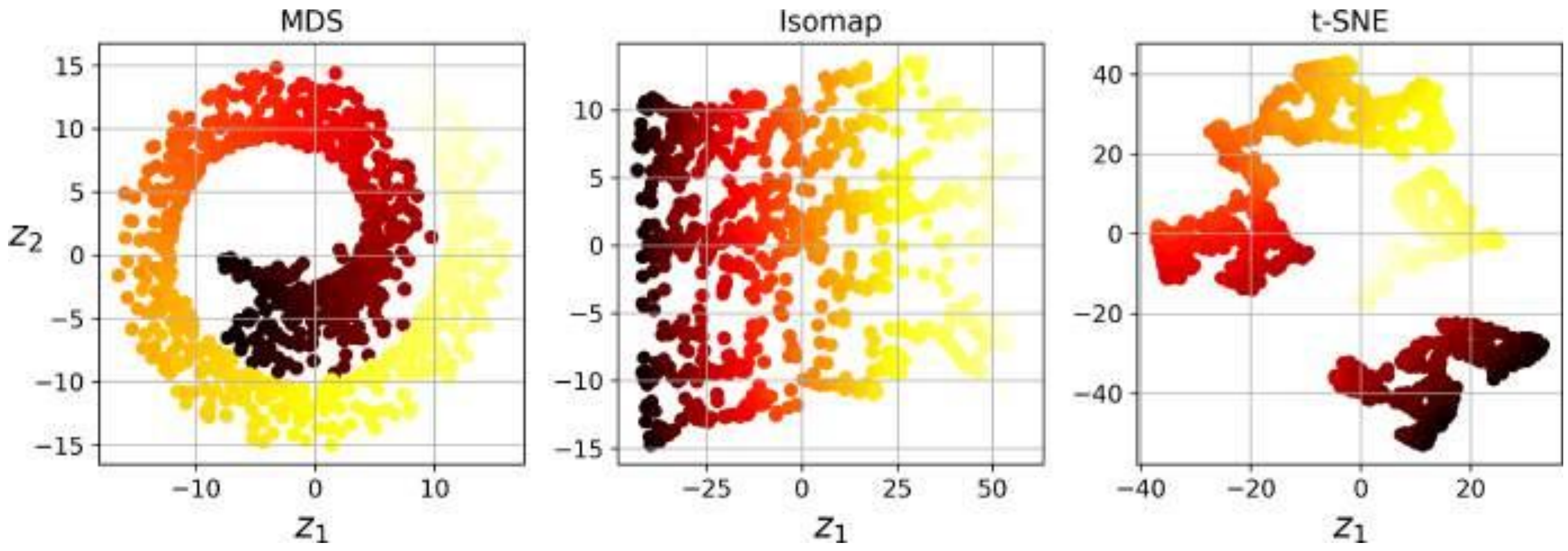
- LLE (Locally Linear Embedding - Embedding tuyến tính cục bộ)
  - LLE works by first measuring how each training instance linearly relates to its closest neighbors (c.n.),
  - and then looking for a low-dimensional representation of the training set where these local relationships are best preserved

Unrolled Swiss roll using LLE



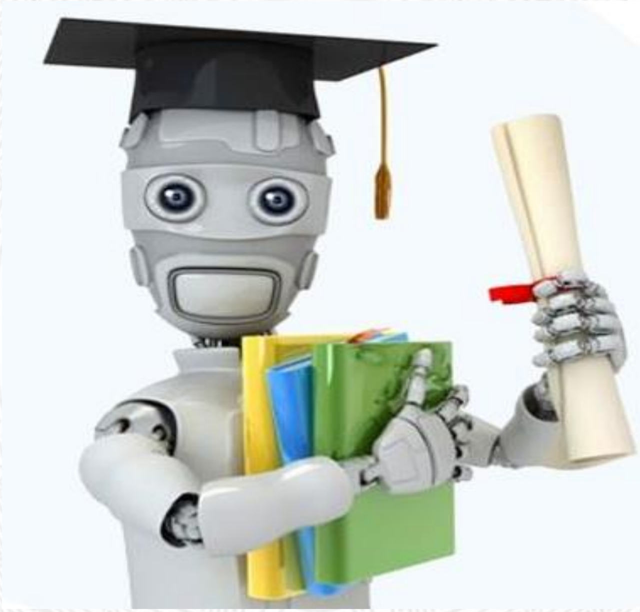
- Other Dimensionality Reduction Techniques
  - **Isomap:** creates a graph by connecting each instance to its nearest neighbors, then reduces dimensionality while trying to preserve the geodesic distances between the instances
  - **t-SNE** (t-Distributed Stochastic Neighbor Embedding) reduces dimensionality while trying to keep similar instances close and dissimilar instances apart.
  - **LDA** (Linear Discriminant Analysis) is actually a classification algorithm, but during training it learns the most discriminative axes between the classes, and these axes can then be used to define a hyperplane onto which to project the data
    ⇨ LDA is a good technique to reduce dimensionality before running another classification algorithm such as an SVM classifier.

- **Other Dimensionality Reduction Techniques**
  - **MDS** (Multidimensional Scaling - co dãn đa chiều) reduces dimensionality while trying to preserve the distances between the instances



Reducing the Swiss roll to 2D using various techniques

Nhân bản – Phụng sự – Khai phóng



Enjoy the Course…!