Non-Linear Dimensionality Reduction

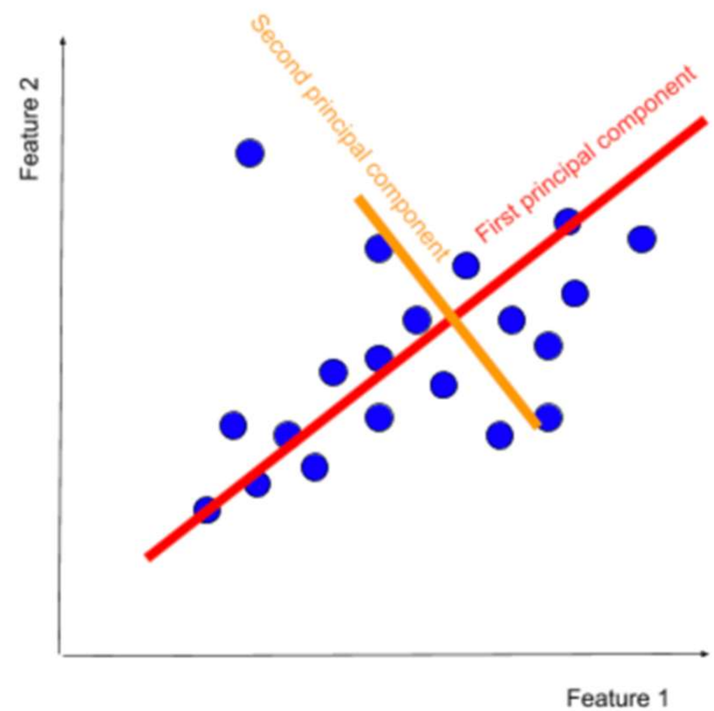
ISOMAP

Principal Component Analysis (PCA)

- **Goal:**

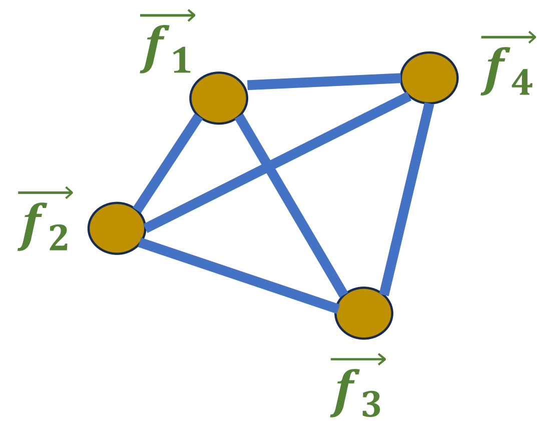
To reduce the dimensionality of the data, while retaining as much as possible of the variation present in the dataset

- **Limitations:**
    - Low interpretability of principal components
    - Not robust against outliers
    - Assumes a linear relationship between features
    - Is sensitive to the scale of the features



Multi-dimensional Scaling (MDS)

- PCA maps input features from d dimensional feature space to k dimensional latent features, with k<<d.
- An alternative approach to PCA is based on preserving pairwise distances.
- MDS helps us visualize data in low dimensions by creating a mapping that will also preserve the relative distance between data
- Given $n(n-1)/2$ pairwise distances $d_{ij} = |f_i - f_j|$, find a low-dimensional embedding $\mathcal{D} \rightarrow \mathcal{X}$ such that $x_i - x_j \approx d_{ij}$



$$\mathcal{D} = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} \\ d_{12} & 0 & d_{23} & d_{24} \\ d_{13} & d_{23} & 0 & d_{34} \\ d_{14} & d_{24} & d_{34} & 0 \end{bmatrix}$$

Classical Approach to MDS

$\mathcal{B}$ is also known as centred inner-product matrix/centred-Gram matrix

- An $\boldsymbol{n} \times \boldsymbol{n}$ dissimilarity matrix $\mathcal{D}$ is provided, with matrix elements $\boldsymbol{d_{ij}}$ showing the Euclidean distance between points $\boldsymbol{i}$ and $\boldsymbol{j}$.

- A double-centered matrix $\mathcal{B}$ (the mean of all rows and columns are zero) is formed, with $\mathcal{B} = -\frac{1}{2}\mathcal{C}\mathcal{D}^2\mathcal{C},\ \boldsymbol{with}\ \mathcal{C} = \mathbb{I} - \frac{1}{n}\mathbb{l}$.

- Do eigen decomposition of $\mathcal{B}$, and select the **m-largest** eigenvalues and eigenvectors, $\mathcal{X}_{n\times m} = \boldsymbol{E}_{n\times m}\, \boldsymbol{\Lambda}^{1/2}_{m\times m}$,
  - $\boldsymbol{E}_{n\times m}$, a $n \times m$ matrix of the m largest eigenvectors
  - $\boldsymbol{\Lambda}^{1/2}_{m\times m}$ is the corresponding $m \times m$ matrix of the square-root of the m largest eigenvalues on the main diagonal

Classical Approach to MDS

- The eigenpair approach exploits the dissimilarity matrix $\mathcal{D}$ to obtain spatial coordinates $\mathcal{X}$

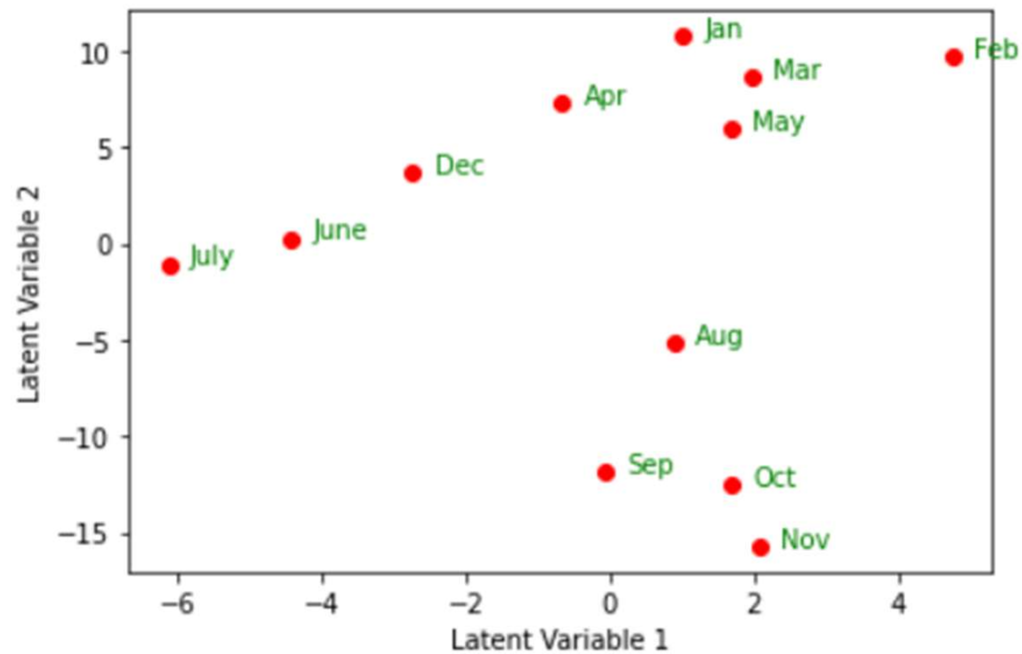
$$Z = \sqrt{\sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(d_{ij} - \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2} \right)^2}$$

- The value Z can be thought of as an objective function value, used to simply find a solution (x_{ij}) resulting in an aggregate minimized distance equal to the given distance matrix

- MDS decomposes an $n \times n$ dissimilarity matrix $\mathcal{D}$ and outputs as $n \times 2$ $(or\ n \times 3)$.

Example

months	sales	profit	ordered	rebounds
Jan	4	3	7	4
Feb	4	2	3	5
Mar	6	2	6	5
Apr	7	5	7	6
May	8	4	5	5
June	14	8	8	8
July	16	7	8	10
Aug	19	6	4	4
Sep	25	8	2	3
Oct	25	10	2	2
Nov	28	11	1	2
Dec	11	5	6	9

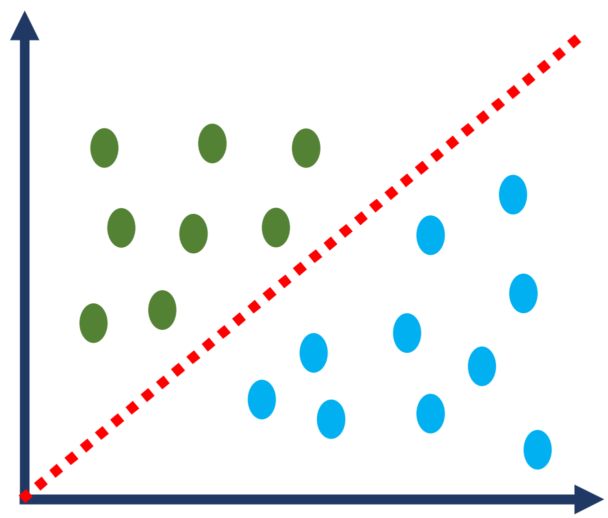


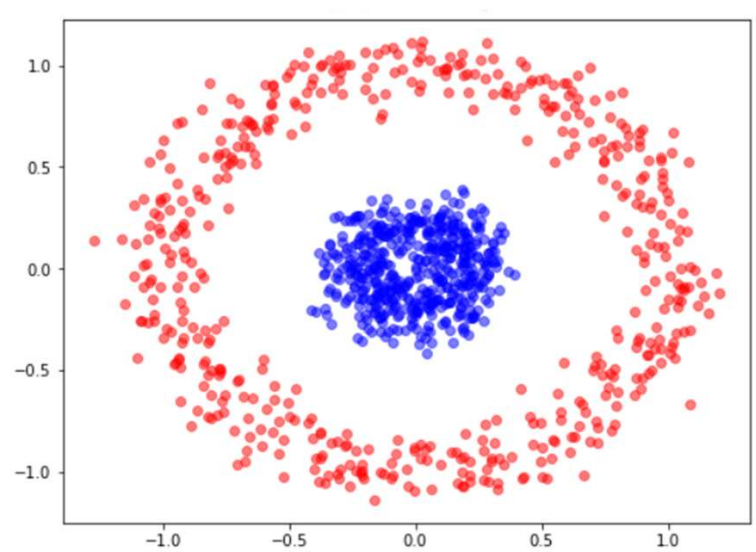
PCA vs. MDS

- The two methods are in some sense "similar" to each other
  - In PCA, we compute the $\boldsymbol{N} \times \boldsymbol{N}$ covariance matrix, with $\boldsymbol{N}$ being the number of features of the dataset
  - In MDS, we compute the $\boldsymbol{D} \times \boldsymbol{D}$ disimilarity matrix, with $\boldsymbol{D}$ being the number of samples of the dataset
- For Euclidean distances $\boldsymbol{d_{ij}}$ in MDS, MDS finds an embedding that preserves the interpoint distances, equivalent to PCA
- Both PCA and MDS have similar strengths, but limited to linear projections

Linear vs. Nonlinear Problems

- The PCA approach is a linear projection technique that works well if the data is linearly separable





- How can we generalize to arbitrary manifolds?

Non-Linear Data

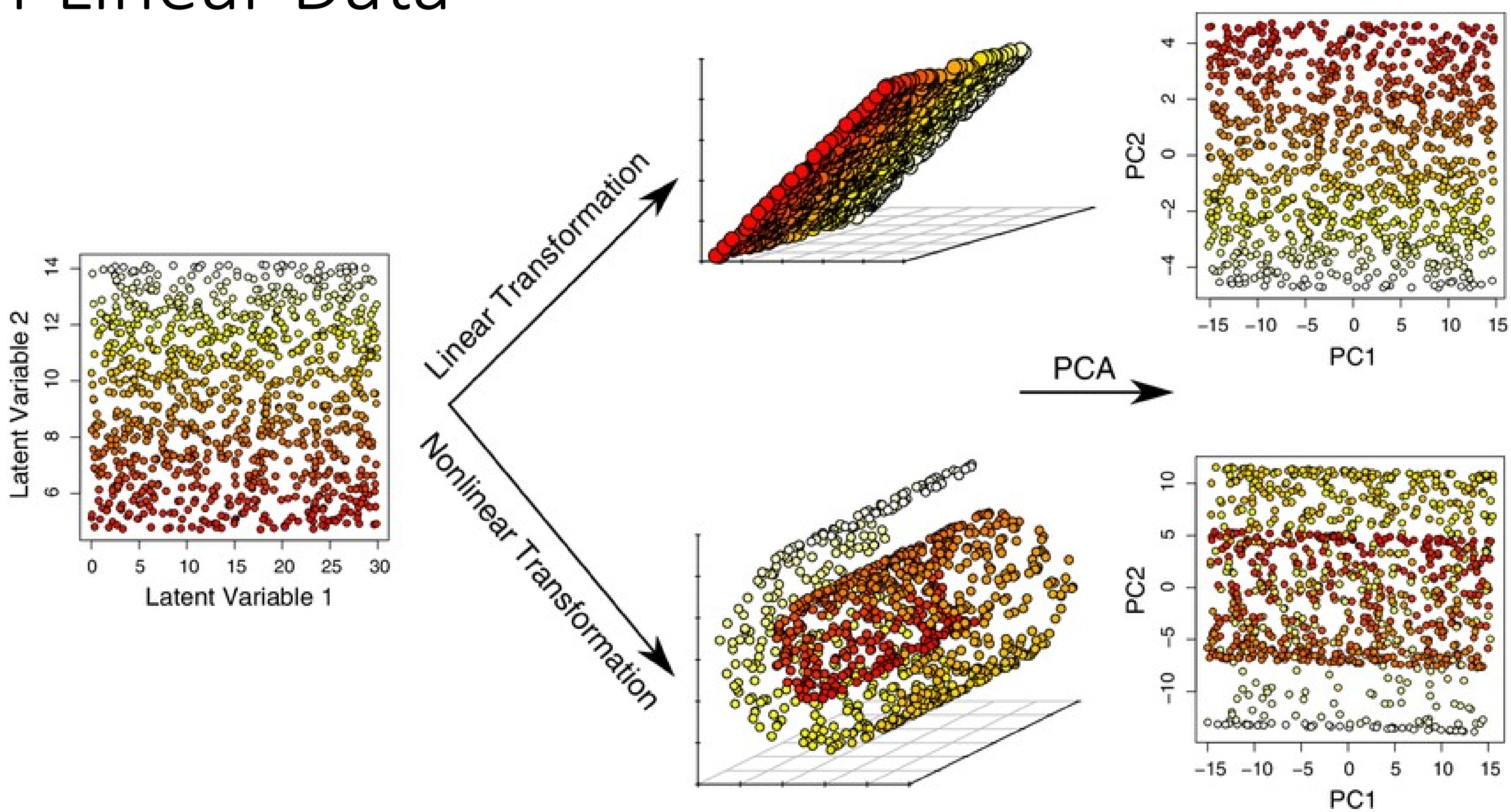


Image Source: [Dimensionality Reduction Techniques for Visualizing Morphometric Data: Comparing Principal Component Analysis to Nonlinear Methods](), Trina Y Du

Isometric Feature Mapping (IsoMap)

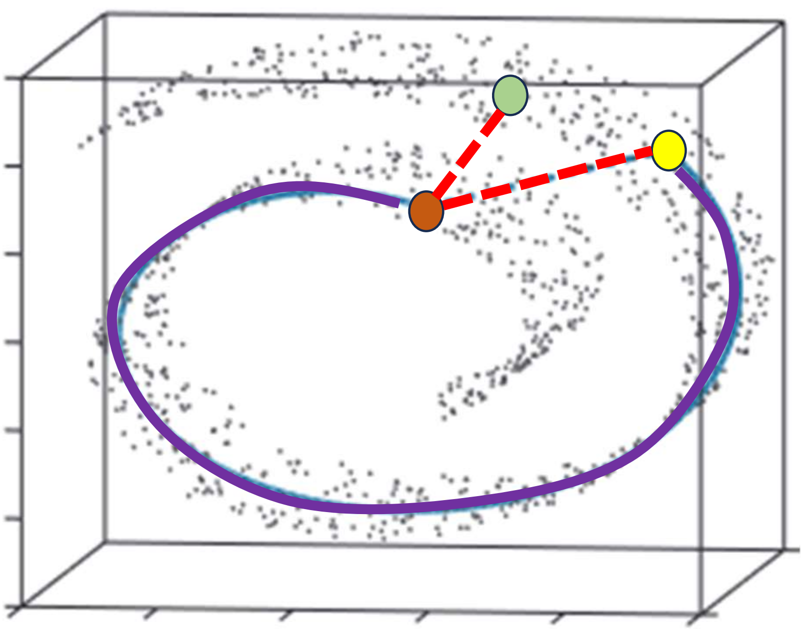
Joshua B. Tenenbaum, Vin de Silva, John C. Langford

Science, Vol 290, Dec 2000, 2319-2323

https://wearables.cc.gatech.edu/paper_of_week/isomap.pdf

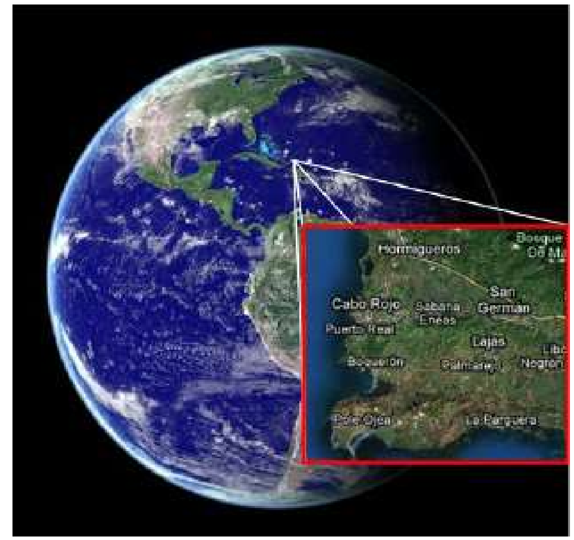
Isometric Feature Mapping (IsoMap)

- MDS as we know seeks an embedding that preserves pairwise distances between data points



- What is the distance between the brown dot and the yellow dot, in the “Swiss Roll” dataset?
- Is it the red dotted line, or the violet solid line?
- Is d( ) > d( )
- How to robustly measure distances along the manifold?

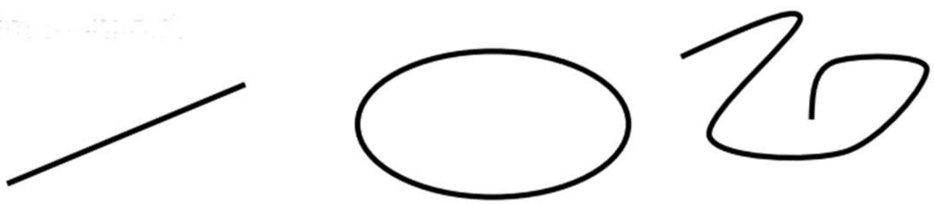
Manifold



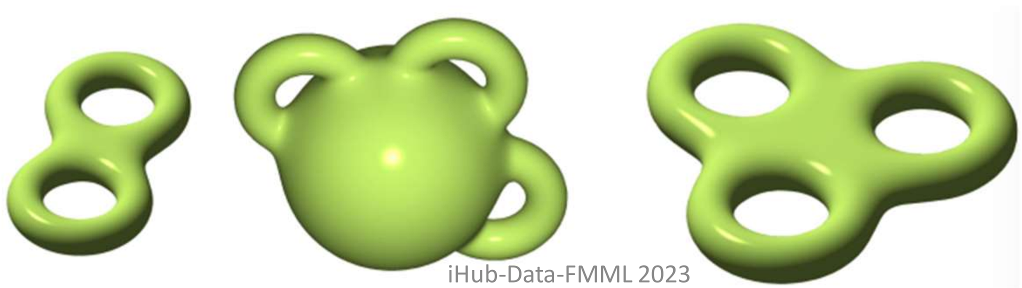
- A manifold is a space or shape that locally resembles Euclidean space near each point, but globally may be more complicated

- A manifold is called a 2-manifold when its surface appears to look relatively flat, but in reality, is not. Our Earth is a 2-manifold

1-dim manifolds:

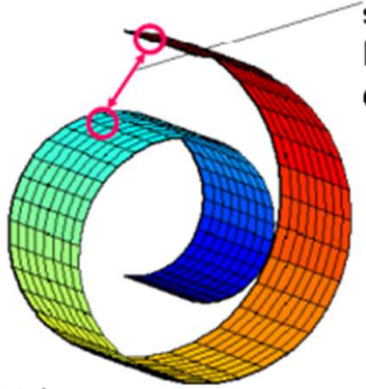


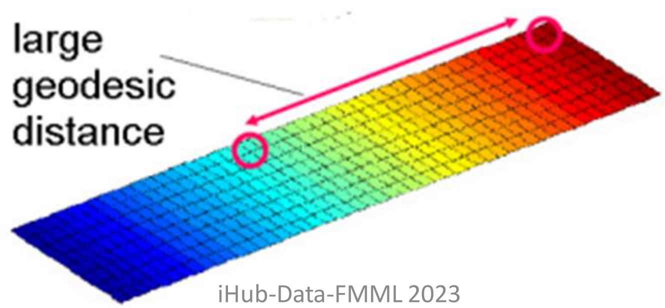
2-dim manifolds:



Geodesics

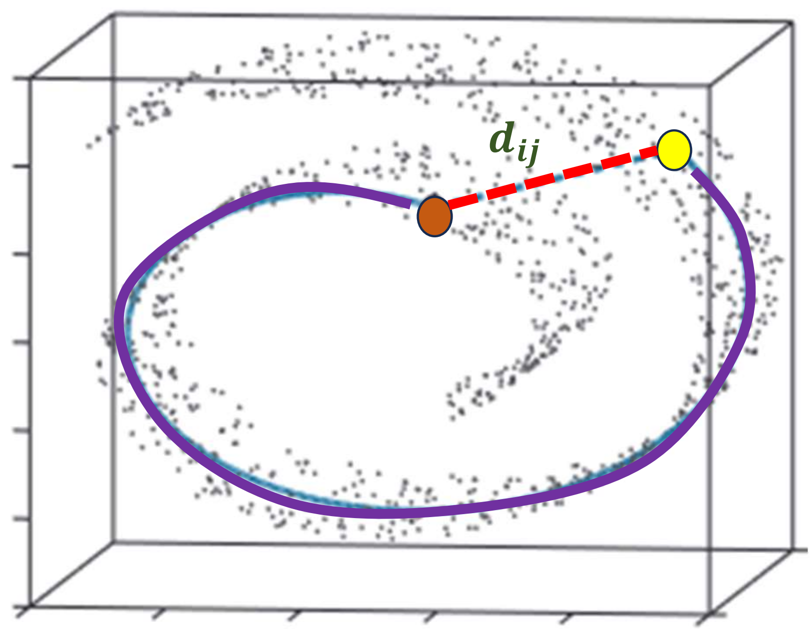
- Geodesics on manifolds define locally length minimising curves, which minimizes the distance between the endpoints
- Geodesics generalize the notion of a straight line in Euclidean space to curved spaces





Isometric Feature Mapping (IsoMap)

- Geodesic distances measured on the manifold may be longer than the corresponding Euclidean straight-line distance d_{ij}
- The geodesic distances reflect the true low-dimensional geometry of the manifold, which the Euclidean distance fail to capture
- Isomap seeks to preserve the intrinsic geometry of data, as captured in the geodesic manifold distances between all pairs of data points

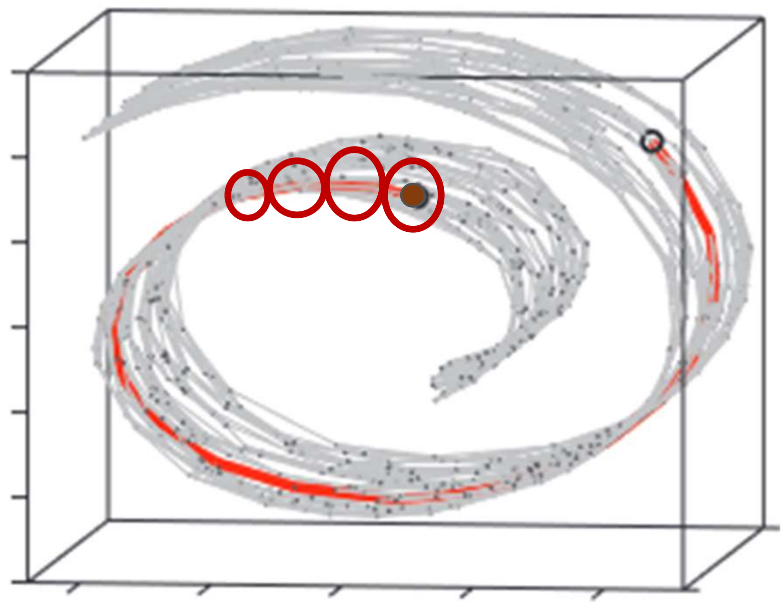


Isomap Algorithm – Step 1

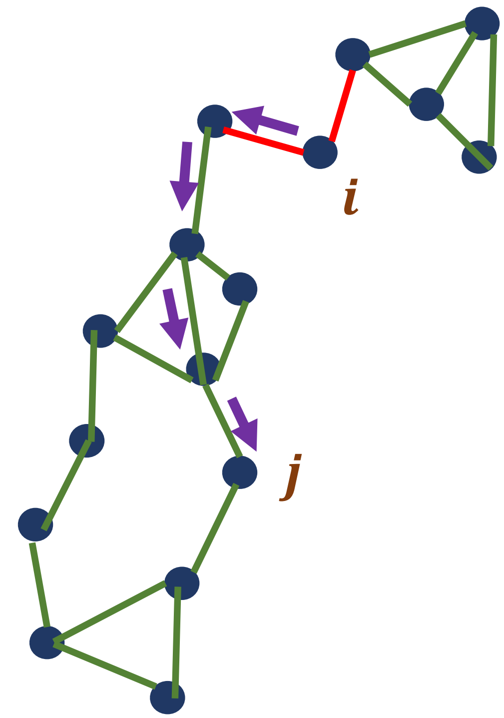
- How to compute geodesics rather than Euclidean distances without knowing the manifold?

<u>Build the adjacency graph over high-dim points X</u>

- Construct neighbourhood graph:
  - Use K-Nearest Neighbours
  - Neighbours within a fixed radius (ε)
- The graph G over all data points is defined by connecting points $\boldsymbol{i}$ and $\boldsymbol{j}$ $\left(\boldsymbol{d_{ij}^X}\right)$, provided they are closer than $\boldsymbol{\varepsilon}$, or if $\boldsymbol{i}$ is one if the K nearest neighbours of $\boldsymbol{j}$
- The edge length is equal to $\boldsymbol{d_{ij}^X}$



Isomap Algorithm – Step 2



Compute shortest paths using Floyd's Algorithm

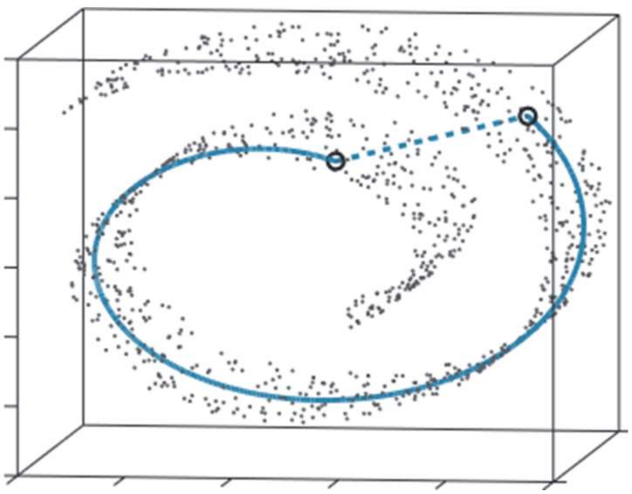
- If points i and j are linked by an edge; $d_{ij}^{G} = d_{ij}^{X}$, else $d_{ij}^{G} = \infty$
- For each value of $k = 1, 2, \ldots, N$ in turn, replace all entries d_{ij}^{G} by $\min\{d_{ij}^{G}, d_{ik}^{G} + d_{kj}^{G}\}$
- The matrix of final values $D^{G} = \{d_{ij}^{G}\}$, will contain the shortest path distance between all pairs of points in G

Isomap Algorithm – Step 3

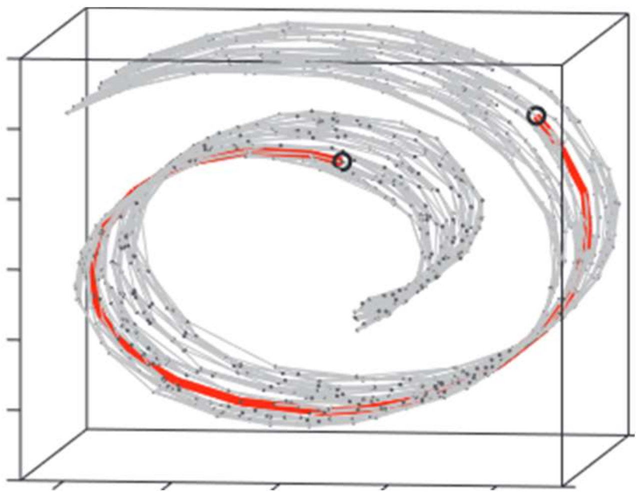
Construct d-dimensional embedding by applying MDS to geodesic distances

- Construct the double centered matrix, $\boldsymbol{\tau}\left(\boldsymbol{D}^{\boldsymbol{G}}\right) = -\frac{1}{2}\boldsymbol{H}\ \boldsymbol{S}\ \boldsymbol{H}$, with $\boldsymbol{S}_{ij} = \ (\boldsymbol{D}_{ij}^{\boldsymbol{G}})^{2}$ and $\boldsymbol{H}$ is the "centering matrix" $\{\boldsymbol{H}_{ij} = \ \boldsymbol{\delta}_{ij} - \boldsymbol{1}/\boldsymbol{N}\}$
- Do eigen decomposition of $\boldsymbol{\tau}$, and select the **d-largest** eigenvalues and eigenvectors, $\boldsymbol{D}^{\boldsymbol{Y}} = \ \boldsymbol{E}_{\boldsymbol{n}\times\boldsymbol{d}}\ \boldsymbol{\Lambda}_{\boldsymbol{d}\times\boldsymbol{d}}^{\boldsymbol{1}/\boldsymbol{2}}$,
- The coordinate vectors $\boldsymbol{y}_{\boldsymbol{i}}$ for the points in $\boldsymbol{Y}$ are chosen to minimize the cost function, $\boldsymbol{E} = |\boldsymbol{\tau}\left(\boldsymbol{D}^{\boldsymbol{G}}\right) - \boldsymbol{\tau}(\boldsymbol{D}^{\boldsymbol{Y}})|_{\boldsymbol{L}^{2}}$

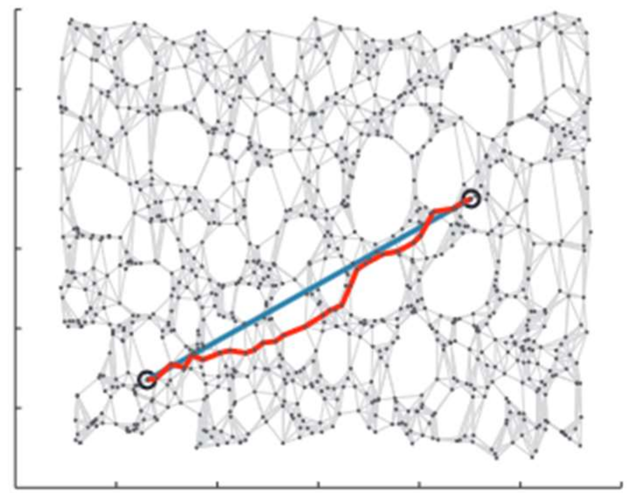
Isomap Algorithm – Steps



The “Swiss roll” dataset



The graph G constructed with K = 7 and N =1000 data points allows an approximation (red segments) to the true geodesic path to be computed efficiently as the shortest path in G.



The two-dimensional embedding recovered by Isomap. Straight lines in the embedding (blue) represent simpler approximations to the true geodesic paths than do the corresponding graph paths (red).

Dimensionality Reduction

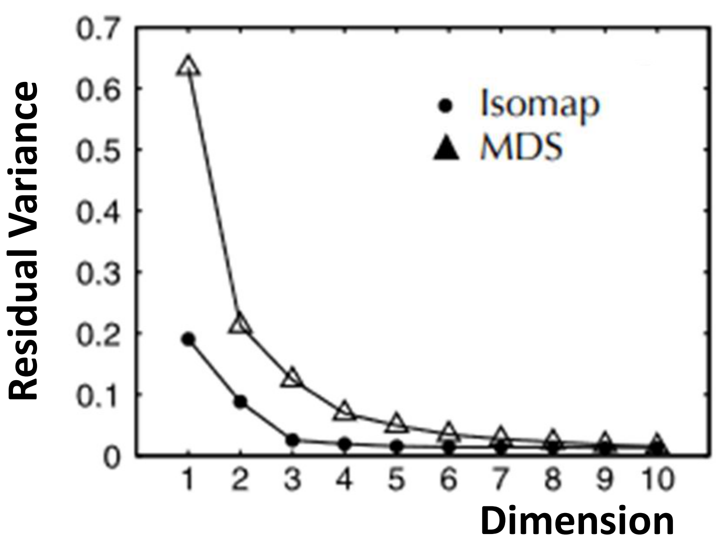
698 samples of 64 pixel by 64-pixel images of a face

- The images are thought of as points in a high-dimensional (4096) vector space, with each input dimension corresponding to the brightness of one pixel in the image or the firing rate of one retinal ganglion cell.

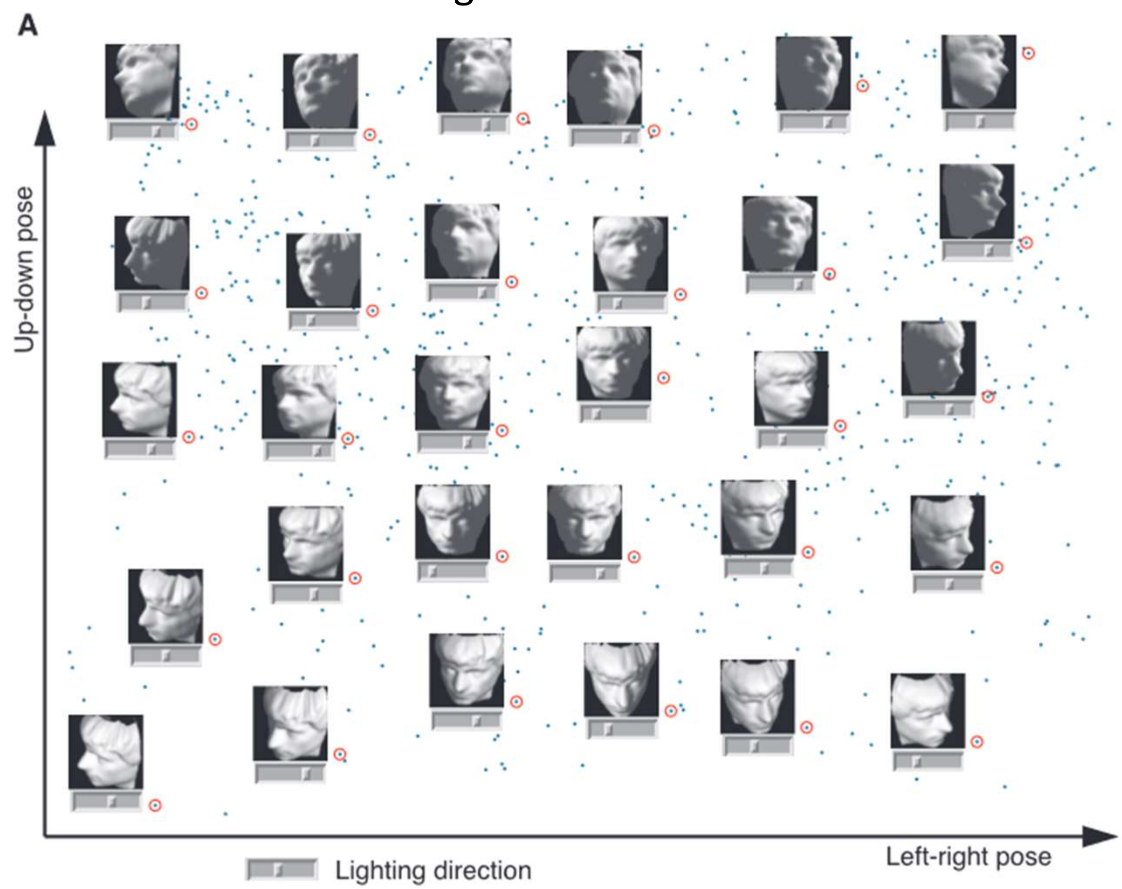
IsoMap Examples

Input: 698 raw images,

$64 \times 64 - \text{pixel}$ images of a face rendered with different poses and lighting directions



Isomap k = 6, learns a three-dimensional embedding, of the data's intrinsic geometric structure

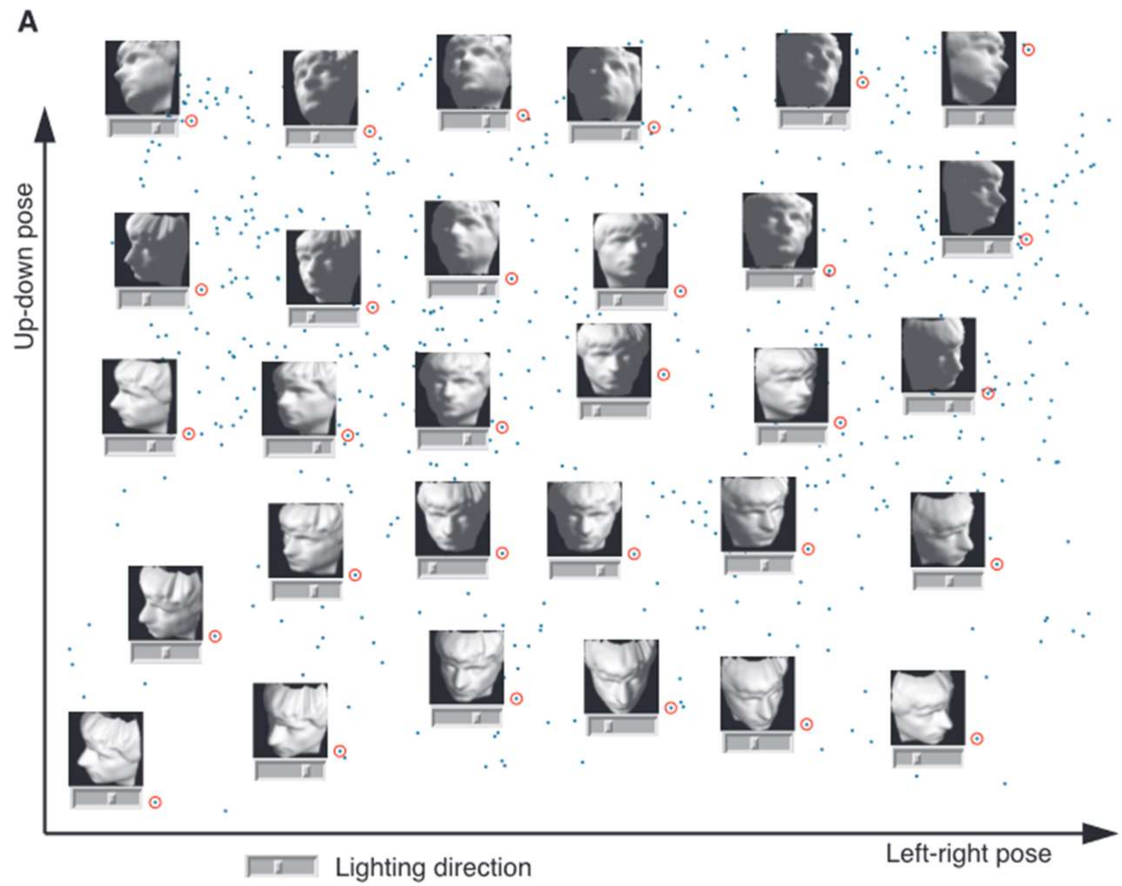


IsoMap Examples

Input: 698 raw images,

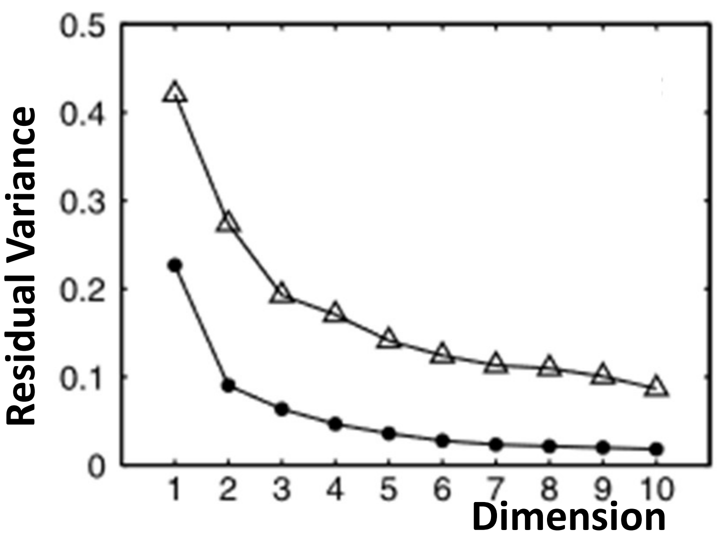
$64 \times 64 - \text{pixel}$ images of a face rendered with different poses and lighting directions

- IsoMap recovers the low-dimensional structure in the data
- Coordinates in the embedding correspond to meaningful modes of variation in the image

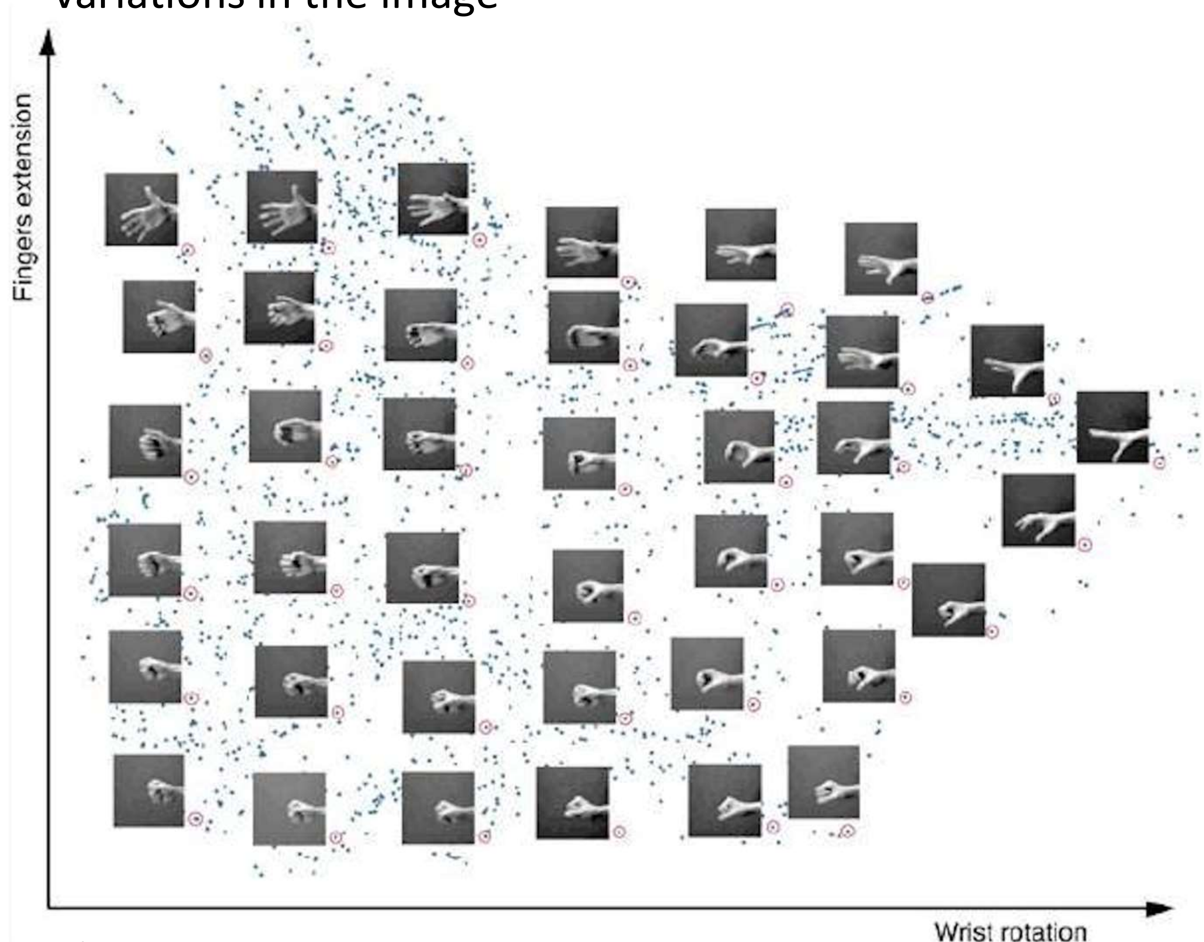


IsoMap Examples

- Hand Images - varying wrist rotation and finger extension
- 2000, $64 \times 64 - \text{pixel}$ images, with Isomap neighbors, k=6

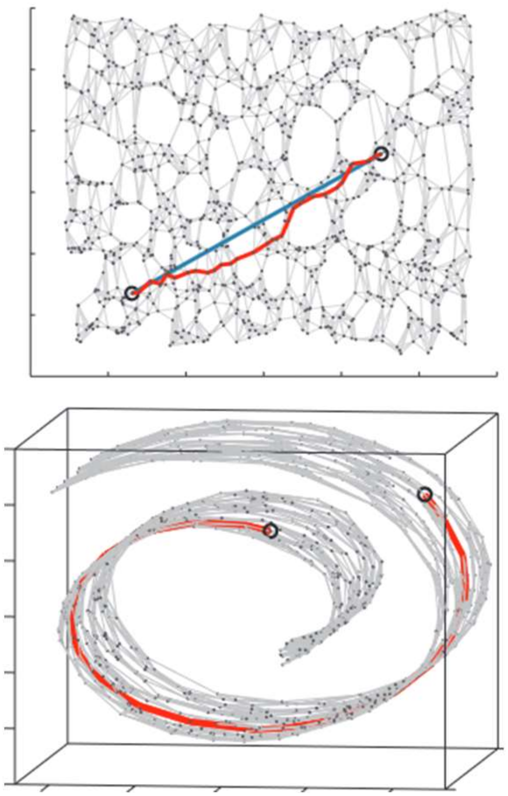


Trajectories in the embedding correspond to meaningful variations in the image



IsoMap Coordinate Space

- Interpolations along “straight” lines in the embedding space yield realistic, though highly nonlinear, transitions in the image



Convergence of IsoMap

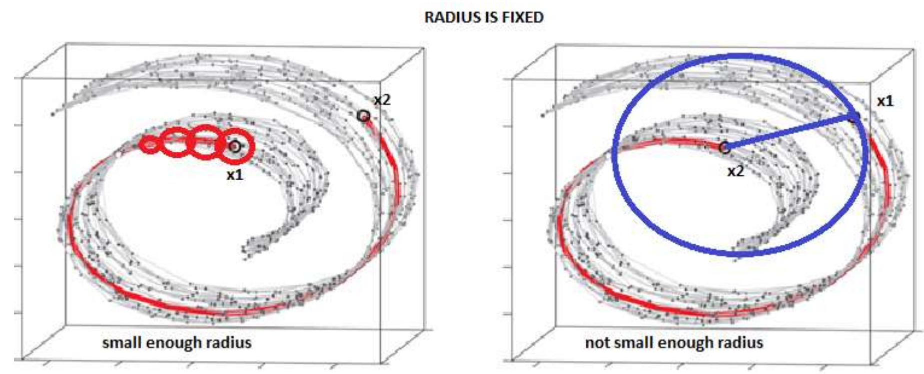
- IsoMap asymptotically recovers the true dimensionality and geometric structure of a strictly larger class of nonlinear manifolds

- The graph distances $\boldsymbol{d_{ij}^{G}}$ provide increasingly better approximations to the intrinsic geodesic distances $\boldsymbol{d_{ij}^{M}}$, in the limit of infinite data points

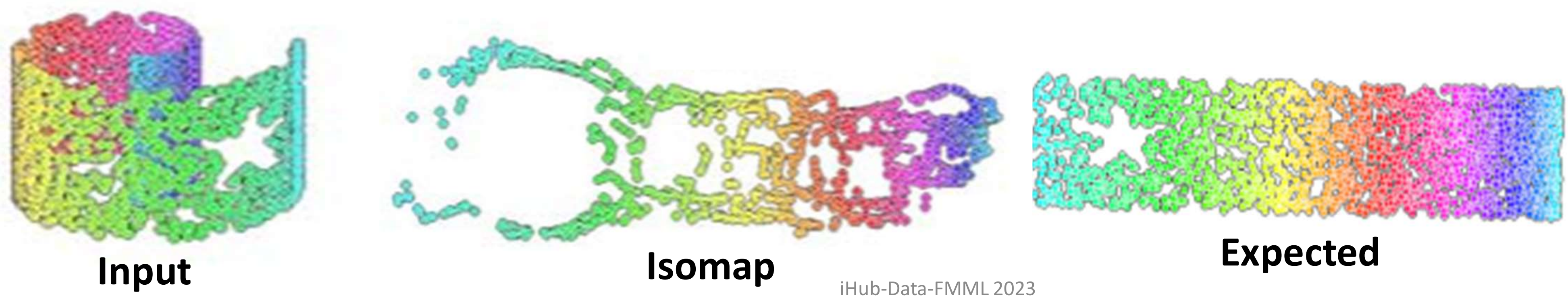
- For non-Euclidean manifolds, such as a hemisphere or the surface of a doughnut, Isomap produces a globally optimal low-dimensional Euclidean representation

IsoMap Shortcomings

- Sensitive to the radius, ϵ and the number of neighbours, k
- Does not scale well with large $\boldsymbol{N}$, as $\boldsymbol{N} \times \boldsymbol{N}$ eigenvector calculation is computationally expensive



- Isomap suffers from nonconvexity such as holes on manifolds



Notes

- Isomap preserves non-linear manifold, by finding the graph that preserves the global, nonlinear geometry of the data

- Isomap keeps the advantage of PCA and MDS

  - Non-iterative procedure
  - Global Optimality
  - Guaranteed convergence

- Isomap represents the global structure of a dataset within a single coordinate system

Overall Summary

	PCA	MDS	ISOMAP
Speed	Extremely fast	Very slow	Extremely slow
Predicts Geometry	No	No	Yes
Handles non-convex data?	No	No	No
Handles curvature?	No	No	Yes
Sensitive to parameters	No	No	Yes