UC San Diego

Data Embedding:
A Geometric Perspective

Instructor: Hao Su

Feb 7, 2019

Agenda

- General theories of embedding
- Algorithms of computing data embedding

Agenda

- **General theories of embedding**

- Algorithms of computing data embedding

Many Names

- Dimensionality reduction
- Embedding
- Multidimensional scaling
- Manifold learning

…

Basic Task

Given pairwise distances extract an embedding.

Is it always possible?
What dimensionality?

Metric Space

Ordered pair (M, d) where M is a set and d satisfies

$$d(x,y) \geq 0$$

$$d(x,y) = 0 \iff x = y$$

$$d(x,y) = d(y,x)$$

$$d(x,z) \leq d(x,y) + d(y,z)$$

$$\forall x, y, z \in M$$

Many Examples of Metric Spaces

$$\mathbb{R}^n, d(x, y) := \|x - y\|_p$$

$$S \subset \mathbb{R}^3, d(x, y) := \text{geodesic}$$

$$C^\infty(\mathbb{R}), d(f, g)^2 := \int_{\mathbb{R}} (f(x) - g(x))^2 \, dx$$

Isometry [ahy-**som**-i-tree]:

A map between metric spaces that preserves pairwise distances.

Q:

Can you always embed a metric space isometrically in ?

Q:

Can you always embed a **finite** metric space isometrically in ?

Disappointing Example

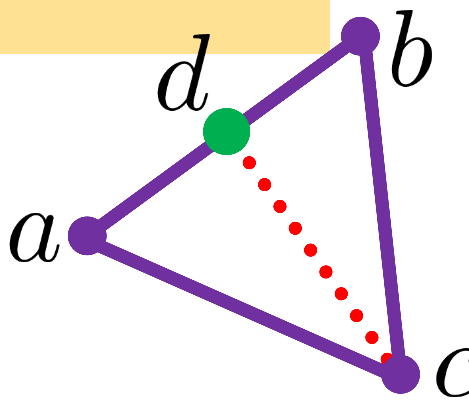
$$X := \{a, b, c, d\}$$

$$d(a, d) = d(b, d) = 1$$

$$d(a, b) = d(a, c) = d(b, c) = 2$$

$$d(c, d) = 1.5$$

Cannot be embedded in Euclidean space!

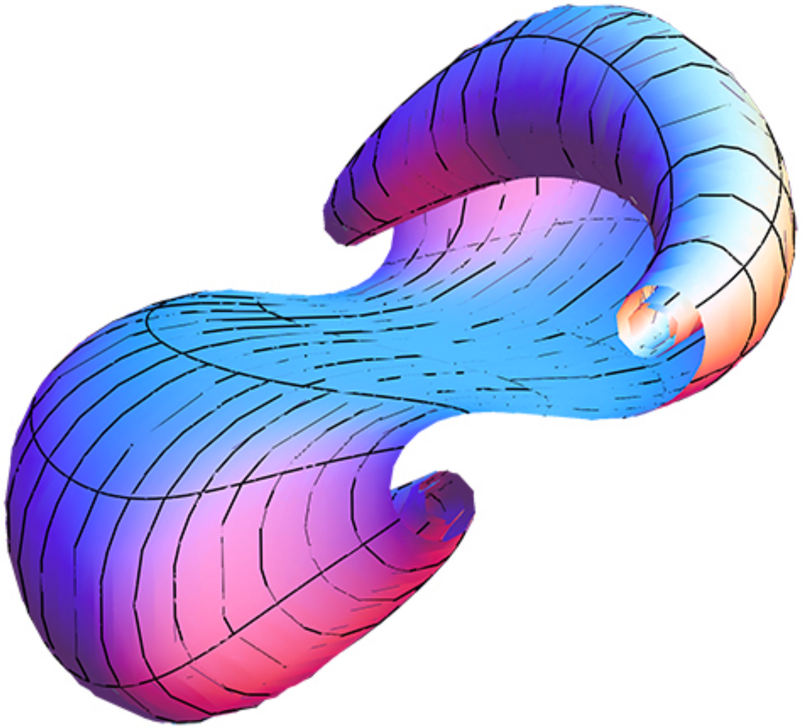


https://chiasme.wordpress.com/2013/10/07/when-does-a-finite-metric-space-embed-isometrically-into-an-euclidean-space/

Embedding Manifold to Euclidean Space

Riemannian manifold

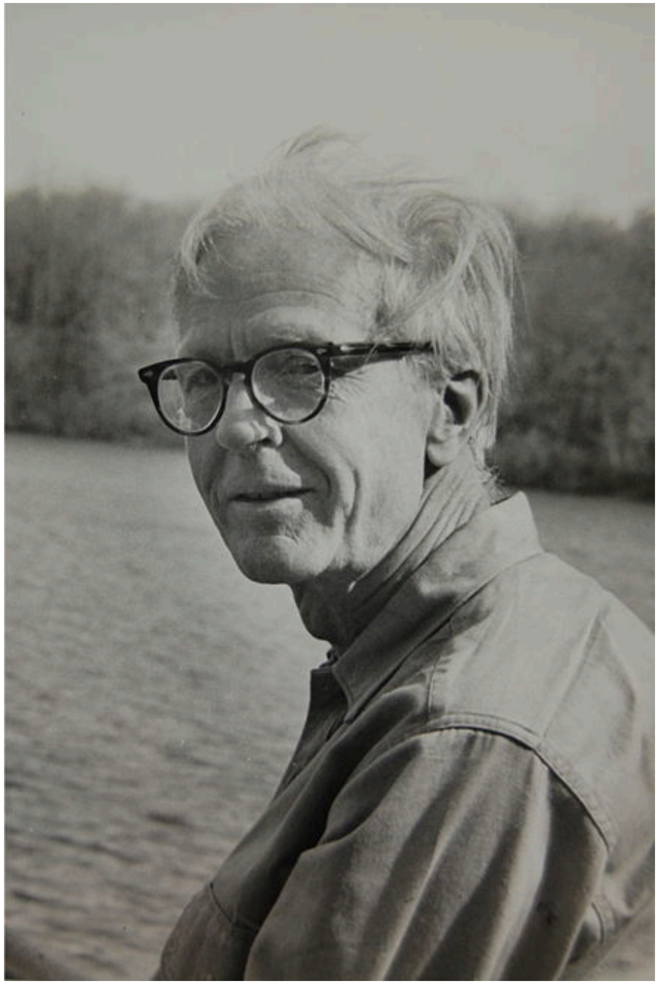
A **Riemannian space** (M, g) is a real, smooth manifold M equipped with an inner product g_p on the tangent space T_pM at each point p that varies smoothly from point to point.



Embedding of Riemannian Manifold

Strong Whitney embedding theorem:

Any smooth real m-manifold (required also to be Hausdorff and second-countable) can be smoothly embedded in the real $2m$-space ($\mathbf{R}^{2m}$), if $m > 0$.



- This is the **best linear** bound on the smallest-dimensional Euclidean space that all m-dimensional manifolds embed in.
- Although every n-manifold embeds in $\mathbf{R}^{2n}$, one can frequently do better.

Embedding of Riemannian Manifold

Theorem. Let (M,g) be a Riemannian manifold and f: $M^m \to \mathbf{R}^n$ a short C^∞-embedding (or immersion) into Euclidean space $\mathbf{R}^n$, where $n \geq m+1$. Then for arbitrary $\varepsilon > 0$ there is an embedding (or immersion) f_ε: $M^m \to \mathbf{R}^n$ which is

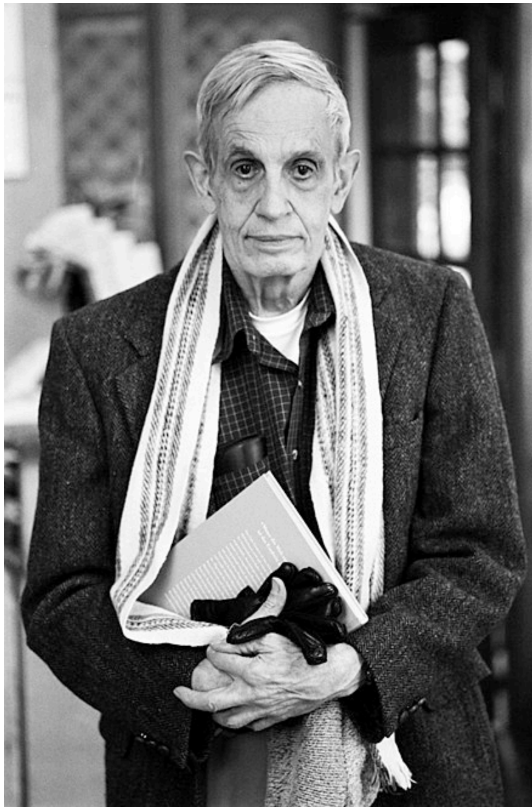
i. in class C^1,

ii. isometric: for any two vectors $v, w \in T_x(M)$ in the tangent space at $x \in M$,

$$g(v,w) = \langle df_\epsilon(v), df_\epsilon(w) \rangle,$$

iii. ε-close to f:

$$|f(x) - f_\epsilon(x)| < \epsilon \ \forall\, x \in M.$$



In particular, any m-dimensional Riemannian manifold admits an isometric C^1-embedding into an *arbitrarily small neighborhood* in $2m$-dimensional Euclidean space.

For example, it follows that any closed oriented Riemannian surface can be C^1 isometrically embedded into an arbitrarily small ε-ball in Euclidean 3-space (for small there is no such C^2-embedding since from the formula for the Gauss curvature an extremal point of such an embedding would have curvature $\geq \varepsilon^{-2}$).

Embedding in ℓ_p

Approximate Embedding

$$\text{expansion}(f) := \max_{x,y} \frac{\mu(f(x), f(y))}{\rho(x,y)}$$

$$\text{contraction}(f) := \max_{x,y} \frac{\rho(x,y)}{\mu(f(x), f(y))}$$

$$\text{distortion}(f) := \text{expansion}(f) \times \text{contraction}(f)$$

http://www.cs.toronto.edu/~avner/teaching/S6-2414/LN1.pdf

Well-Known Result

Theorem (Bourgain, 1985).
Let (X,d) be a metric space on n points. Then,

$$(X, d) \xhookrightarrow{O(\log n)} \ell_p^{O(\log^2 n)}$$

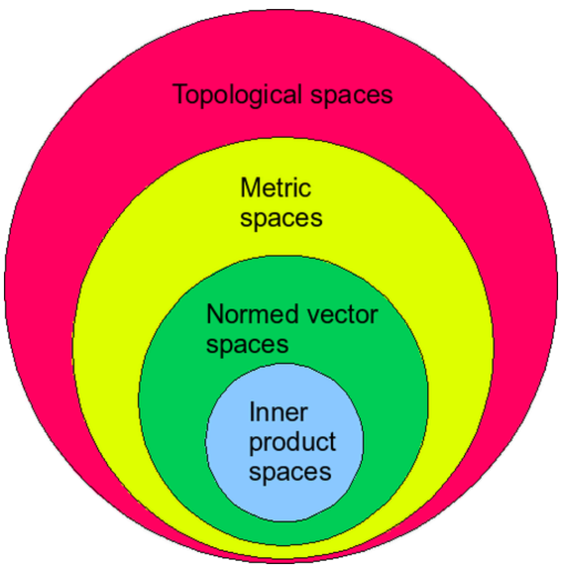
Any n-point metric space (X, D) can be embedded in ℓ_2 (in fact, in every ℓ_p) with distortion $O(\log n)$.

http://ttic.uchicago.edu/~harry/teaching/pdf/lecture3.pdf

THE EUCLIDEAN SPACES ℓ_2^d

metric, the simplest in many respects, and the most restricted. Every finite ℓ_2 metric embeds isometrically in ℓ_p for all p. More generally, we have the following

THEOREM 8.1.1 *Dvoretzky's theorem (a finite quantitative version)*
For every d and every $\varepsilon > 0$ there exists $n = n(d, \varepsilon) \leq 2^{O(d/\varepsilon^2)}$ such that ℓ_2^d can be $(1+\varepsilon)$-embedded in every n-dimensional normed space.



THE EUCLIDEAN SPACES ℓ_2^d

- Dimension reduction:

THEOREM 8.2.3 *Johnson and Lindenstrauss* [JL84]

For every $\varepsilon > 0$, *any* n-*point* ℓ_2 *metric can be* $(1+\varepsilon)$-*embedded in* $\ell_2^{O(\log n/\varepsilon^2)}$.

THE EUCLIDEAN SPACES ℓ_∞^d

The spaces ℓ_∞^d are the richest (and thus generally the most difficult to deal with); *every* n-point metric space (X, D) embeds isometrically in ℓ_∞^n. To see this, write $X = \{x_1, x_2, \dots, x_n\}$ and define $f\colon X \to \ell_\infty^n$ by $f(x_i)_j = D(x_i, x_j)$.

THEOREM 8.2.2

For an integer $b > 0$ set $c = 2b-1$. Then any n-point metric space can be embedded in ℓ_∞^d with distortion c, where $d = O(bn^{1/b} \log n)$.

Graph Embedding

THEOREM 8.3.2 *Rao* [Rao99]

Any n-point planar-graph metric can be embedded in ℓ_2 with distortion $O(\sqrt{\log n})$.

TABLE 8.5.1 A summary of approximate embeddings

FROM	TO	DISTORTION	REFERENCE
any	ℓ_p, $1 \le p < \infty$	$O(\log n)$	[Bou85]
constant-degree expander	ℓ_p, $p < \infty$ fixed	$\Omega(\log n)$	[LLR95]
k-reg. graph, $k \ge 3$, girth g	ℓ_2	$\Omega(\sqrt{g})$	[LMN02]
any	$\ell_\infty^{O(bn^{1/b}\log n)}$	$2b{-}1$, $b{=}1,2,\ldots$	[Mat96]
some	$\Omega(n^{1/b})$-dim'l. normed space	$2b{-}1$, $b{=}1,2,\ldots$ (Erdős's conj.!)	[Mat96]
any	ℓ_1^1	$\Theta(n)$	[Mat90]
any	ℓ_p^d, d fixed	$O(n^{2/d}\log^{3/2} n)$, $\Omega\left(n^{1/\lfloor (d+1)/2\rfloor}\right)$	[Mat90]
ℓ_2 metric	$\ell_2^{O(\log n/\varepsilon^2)}$	$1+\varepsilon$	[JL84]
ℓ_1 metric	$\ell_1^{n^\alpha}$, $0 < \alpha < 1$	$\Omega(\alpha^{-1/2})$	[BC03]
planar or forbidden minor	ℓ_2	$O(\sqrt{\log n})$	[Rao99]
series-parallel	ℓ_2	$\Omega(\sqrt{\log n})$	[NR02]
planar	$\ell_\infty^{O(\log^2 n)}$	$O(1)$	implicit in [Rao99]
outerplanar or series-parallel	ℓ_1	$O(1)$	[GNRS99]
tree	ℓ_1	1	(folklore)
tree	$\ell_\infty^{O(\log n)}$	1	[LLR95]
tree	ℓ_2	$\Theta((\log\log n)^{1/2})$	[Bou86, Mat99]
tree	ℓ_2^d	$O(n^{1/(d-1)})$	[Gup00]
tree, unit edges	ℓ_2^2	$\Theta(\sqrt{n})$	[BMMV02]
Hausdorff metric over (X,D)	$\ell_\infty^{\lvert X\rvert}$	1	[FCI99]
Hausd. over s-subsets of (X,D)	$\ell_\infty^{s^{O(1)}\lvert X\rvert^\alpha \log\Delta}$	$c(\alpha)$	[FCI99]
Hausd. over s-subsets of ℓ_p^k	$\ell_\infty^{s^2(1/\varepsilon)^{O(k)}\log\Delta}$	$1+\varepsilon$	[FCI99]
EMD over (X,D)	ℓ_1	$O(\log\lvert X\rvert)$	[Cha02, FRT03]
Levenshtein metric	ℓ_1	$\ge 3/2$	[ADG+03]
block-edit metric over Σ^d	ℓ_1	$O(\log d \cdot \log^* d)$	[MS00, CM02]
(1,2)-B metric	$\ell_\infty^{O(B\log n)}$	1	[GI03]; for ℓ_p cf. [Tre01]
any	convex comb. of dom. trees (HSTs)	$O(\log n)$	[FRT03]
any	convex comb. of spanning trees	$2^{O(\sqrt{\log n\log\log n})}$	[AKPW95]

Agenda

- General theories of embedding

- **Algorithms of computing data embedding**

Euclidean Case

$$D_{ij} = \|x_i - x_j\|_2^2, D \in \mathbb{R}^{n \times n}$$

Proposition. .

Proof:

$$D = -2X^\top X + \mathrm{diag}(X^\top X)\mathbf{1}^\top + \mathbf{1}\mathrm{diag}(X^\top X)^\top$$

Embedding via eigenvalue problem (take $x_1 = 0$):

$$\|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2x_i \cdot x_j$$

$$\implies x_i \cdot x_j = \frac{1}{2}\left[\|x_i\|_2^2 + \|x_j\|_2^2 - \|x_i - x_j\|_2^2\right]$$

Gram Matrix [gram **mey**-triks]:

A matrix of inner products

$$X^\top X$$

Classical Multidimensional Scaling

1. Double centering: $B := -\frac{1}{2}JDJ$
   Centering matrix $J := I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$

2. Find m largest eigenvalues/eigenvectors

3. $X = E_m \Lambda_m^{1/2}$

"MDS"

Torgerson, Warren S. (1958). *Theory & Methods of Scaling*.

Stress Majorization

$$\min_{X} \sum_{ij} \left(d_{ij}^{0} - \|x_i - x_j\|_2\right)^2$$

Nonconvex!

SMACOF:

Scaling by Majorizing a Complicated Function

de Leeuw, J. (1977), "Applications of convex analysis to multidimensional scaling" *Recent developments in statistics*, 133–145.

SMACOF Potential Terms

$$\min_X \sum_{ij} \left(d_{ij}^0 - \|x_i - x_j\|_2\right)^2$$

$$\sum_{ij} (d_{ij}^0)^2 = \text{const.}$$

$$\sum_{ij} \|x_i - x_j\|_2^2 = \operatorname{tr}(XVX^\top), \text{ where } V = 2nI - 2\mathbf{1}\mathbf{1}^\top$$

$$\sum_{ij} d_{ij}^0 \|x_i - x_j\|_2 = \operatorname{tr}(XB(X)X^\top)$$

$$\text{where } b_{ij}(X) := \begin{cases} -\frac{2d_{ij}^0}{\|x_i - x_j\|_2} & \text{if } x_i \neq x_j, i \neq j \\ 0 & \text{if } x_i = x_j, i \neq j \\ -\sum_{j \neq i} b_{ij} & \text{if } i = j \end{cases}$$

SMACOF Lemma

$$\sum_{ij} (d_{ij}^0)^2 = \text{const.}$$

$$\sum_{ij} \|x_i - x_j\|_2^2 = \text{tr}(XVX^\top)$$

$$\sum_{ij} d_{ij}^0 \|x_i - x_j\|_2 = \text{tr}(XB(X)X^\top)$$

$$\text{where } b_{ij}(X) := \begin{cases} -\frac{2d_{ij}^0}{\|x_i - x_j\|_2} & \text{if } x_i \neq x_j, i \neq j \\ 0 & \text{if } x_i = x_j, i \neq j \\ -\sum_{j \neq i} b_{ij} & \text{if } i = j \end{cases}$$

Lemma. Define

$$\tau(X, Z) := \text{const.} + \text{tr}(XVX^\top) - 2\text{tr}(XB(Z)Z^\top)$$

Then,

$$\tau(X, X) \leq \tau(X, Z) \; \forall Z$$

with equality exactly when X=Z.

See *Modern Multidimensional Scaling* (Borg, Groenen)

SMACOF: Single Step

$$X^{k+1} \leftarrow \min_X \tau(X, X^k)$$

$$\tau(X,Z) := \text{const.} + \operatorname{tr}(XVX^\top) - 2\operatorname{tr}(XB(Z)Z^\top)$$

$$\begin{aligned}
\implies 0 &= \nabla_X[\tau(X, X^k)] \\
&= 2XV - 2X^k B(X^k) \\
\implies X^{k+1} &= X^k B(X^k) V^+ \\
V^+ &= (2nI - 2\mathbf{1}\mathbf{1}^\top)^+ \\
&= \frac{1}{2n}\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)^+ \\
&= \frac{1}{2n}\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)
\end{aligned}$$

Majorization-Minimization algorithm

Objective convergence:
$\tau(X^{k+1}, X^{k+1}) \leq \tau(X^k, X^k)$

More General Metric MDS

- In general, we minimize directly the square loss on distances

$$\text{stress} = \mathcal{L}(\hat{d}_{ij}) = \left(\sum_{i<j} (\hat{d}_{ij} - f(d_{ij}))^2 / \sum d_{ij}^2 \right)^{\frac{1}{2}}$$

- Sammon mapping

$$\text{Sammon's stress}(\hat{d}_{ij}) = \frac{1}{\sum_{\ell<k} d_{\ell k}} \sum_{i<j} \frac{(\hat{d}_{ij} - d_{ij})^2}{d_{ij}}$$

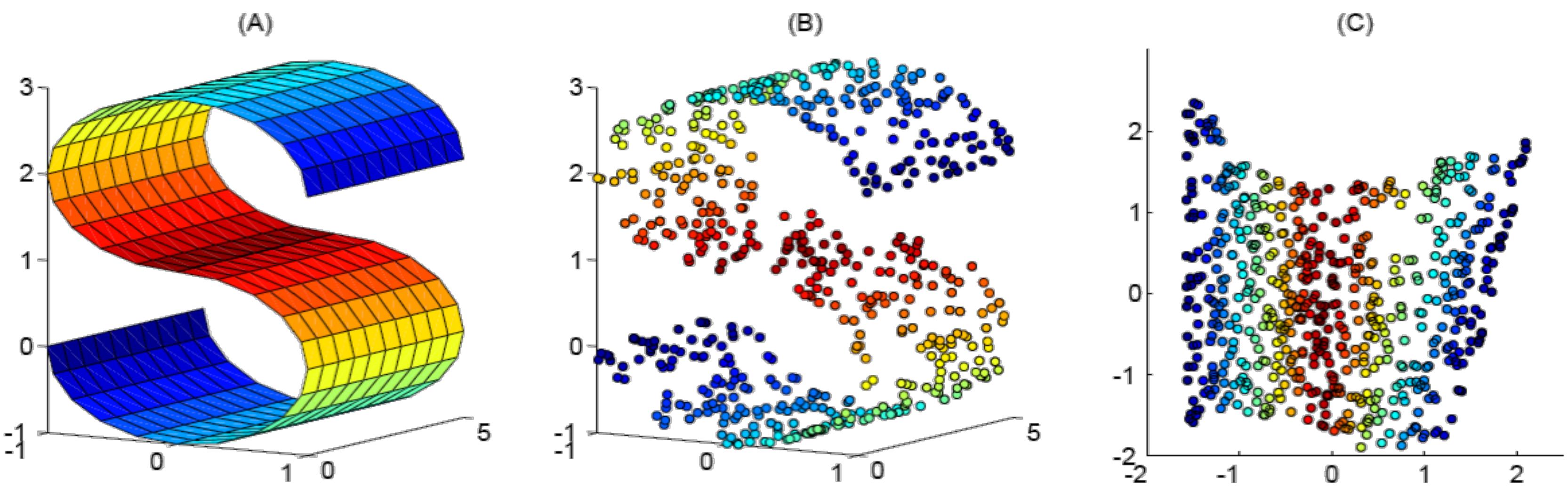
  - This weighting system normalizes the squared-errors in pairwise distances by using the distance in the original space. As a result, Sammon mapping preserves the small d_{ij}, giving them a greater degree of importance in the fitting procedure than for larger values of d_{ij}

Generally solved by gradient descent

Non-Linear Dimensionality Reduction

Non-Linear Dimensionality Reduction

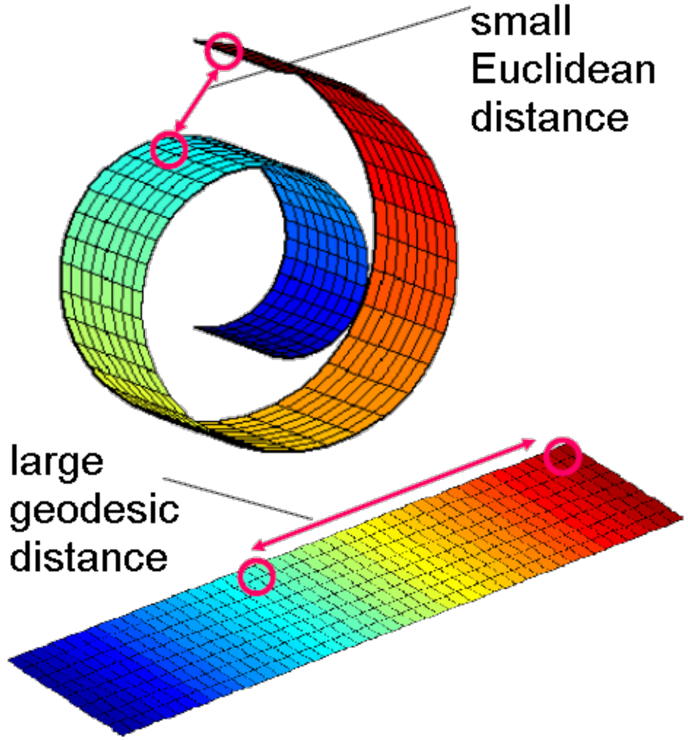
- Many data sets contain essential nonlinear structures that "invisible" to PCA and MDS
- Must resort to some nonlinear dimensionality reduction approaches



Data sampled from a non-linear manifold

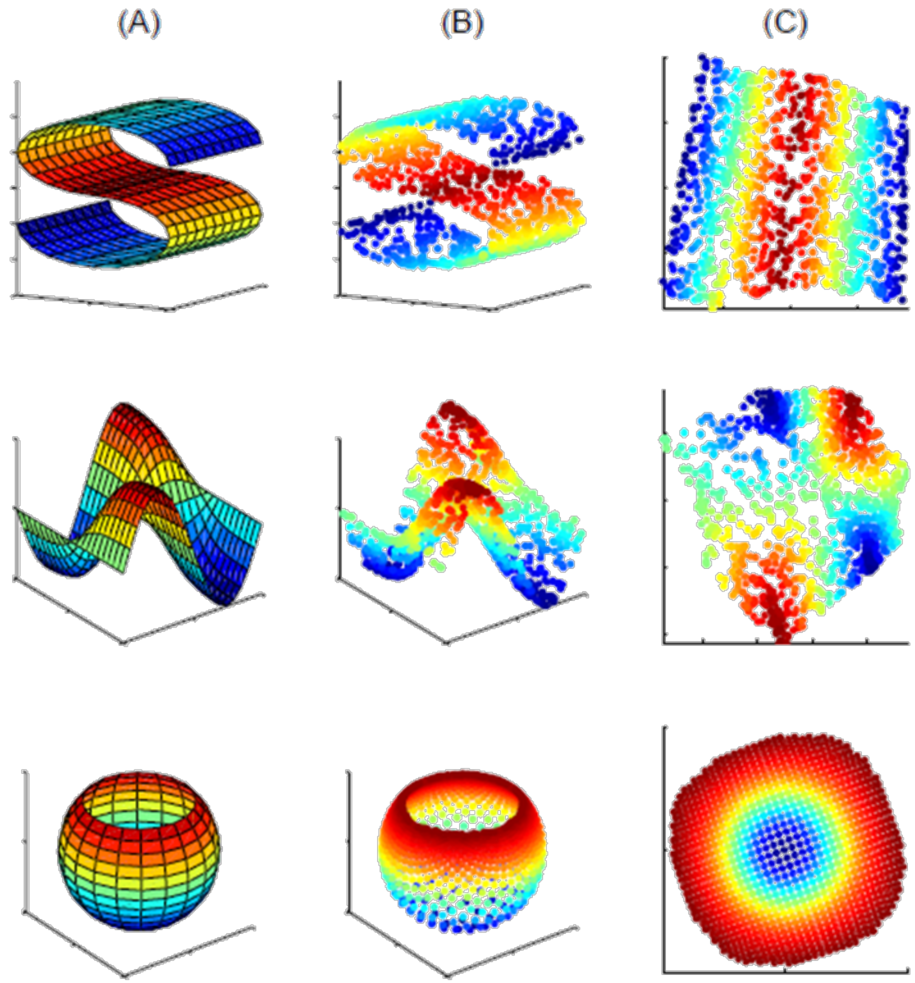
The Choice of Distance

- We can try to capture the manifold structure through the right notion of distance directly on the manifold (geodesic distance)



The Challenge of NLDR

- An unsupervised learning algorithm must discover the global internal coordi-nates of the manifold without external signals that suggest how the data should be embedded in low dimensions

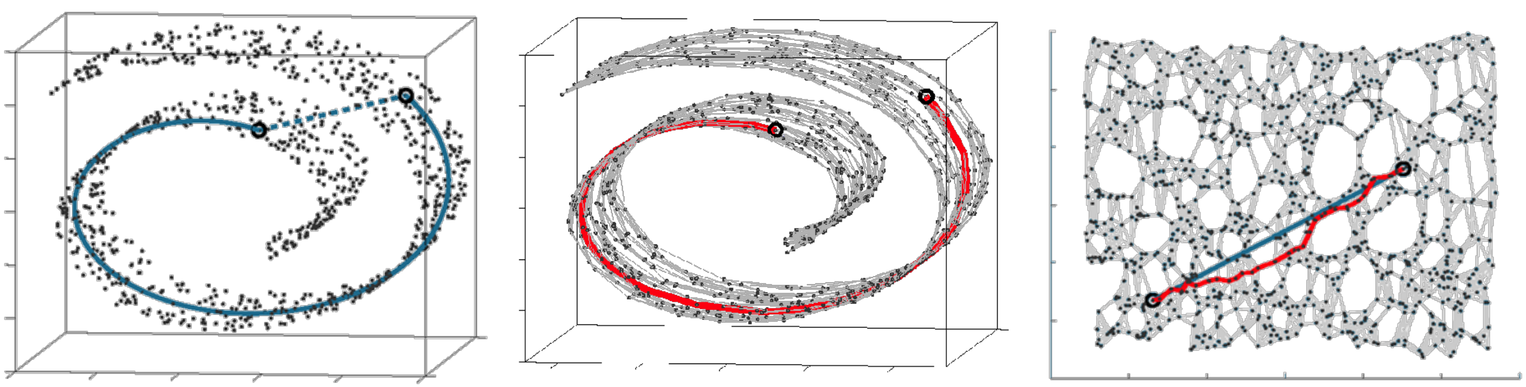


Isomap

ISOMAP
(J. B. Tenenbaum, V. de Silva and J. C. Langford)

- Example of non-linear structure (Swiss roll)
  - Only the geodesic distances reflect the true low-dimensional geometry of the manifold

- ISOMAP (Isometric Feature Mapping)
  - Preserves the intrinsic geometry of the data
  - Uses the geodesic manifold distances between all pairs



ISOMAP (Algorithm Description)

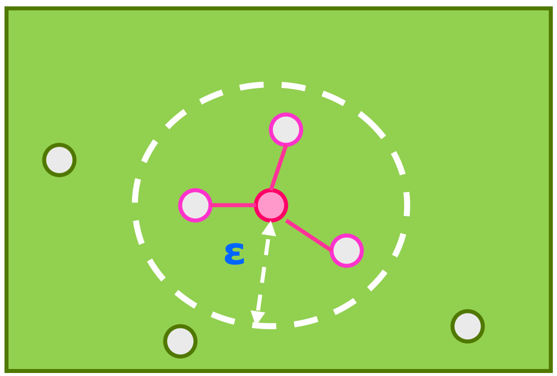
- **Step 1**
  - Form a near-neighbor graph G on the original data points, weighing the edges based on their original distances $d_X(i, j)$
- **Step 2**
  - Estimate the geodesic distances $d_G(i, j)$ between all pairs of points on the sampled manifold by computing their shortest path distances in the graph G.
- **Step 3**
  - Construct an embedding of the data in *d*-dimensional Euclidean space *Y* that best preserves the distances (MDS).

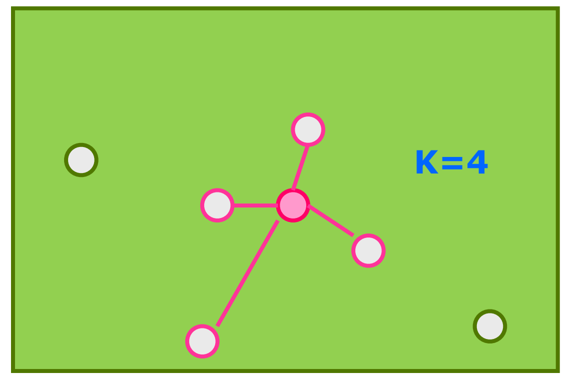
Near Neighbor Graph

- **Step 1**
  - Determining neighboring points within a fixed radius based on the input space distance

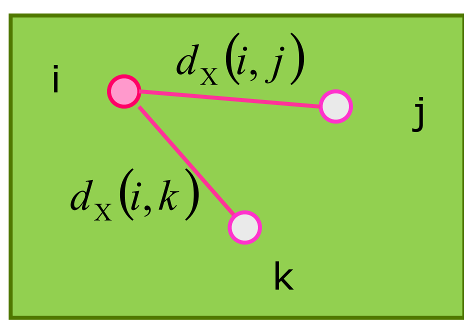
 # ε-radius **# K-nearest neighbors**

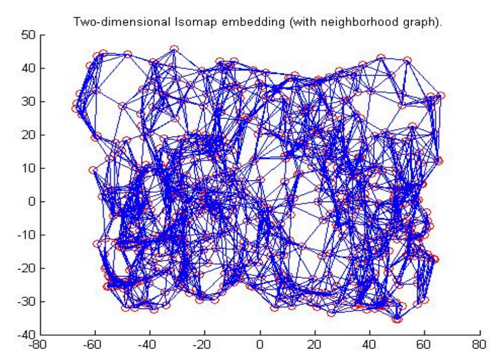
$$d_X(i, j)$$





  - These neighborhood relations are represented as a weighted graph G over the data points.





Shortest Path Computation

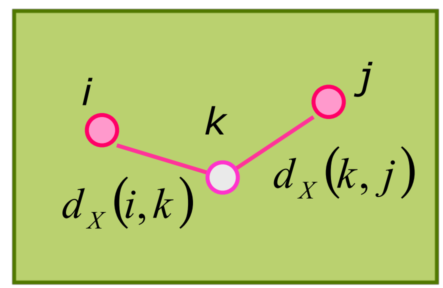
- **Step 2**
  - Estimating the geodesic distances $\boxed{d_G(i,j)}$ between all pairs of points on the manifold by computing their shortest path distances in the graph G.
  - Can be done using Floyd/Warshall's algorithm or Dijkstra's algorithm

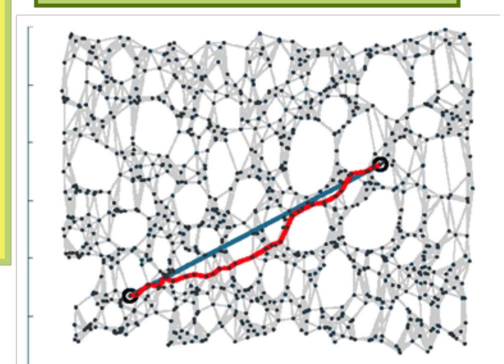
$$d_G(i,j) = d_X(i,j) \quad \text{neighboring } i, j$$

$$d_G(i,j) = \infty \qquad \text{othewise}$$

$$\mathit{for}\ \mathrm{k} = 1,2,\ldots,\mathrm{N}$$

$$d_G(i,j) = \min\{\, d_X(i,j),\ d_X(i,k) + d_X(k,j)\}$$

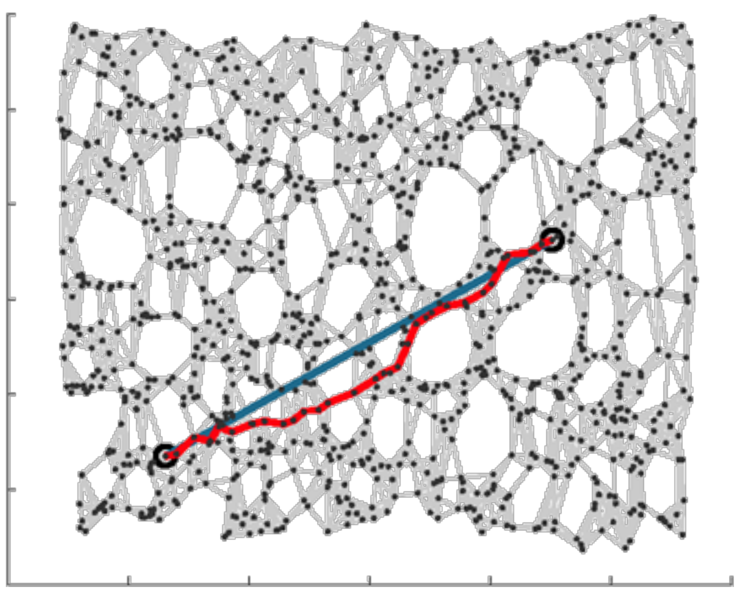




Euclidean Embedding

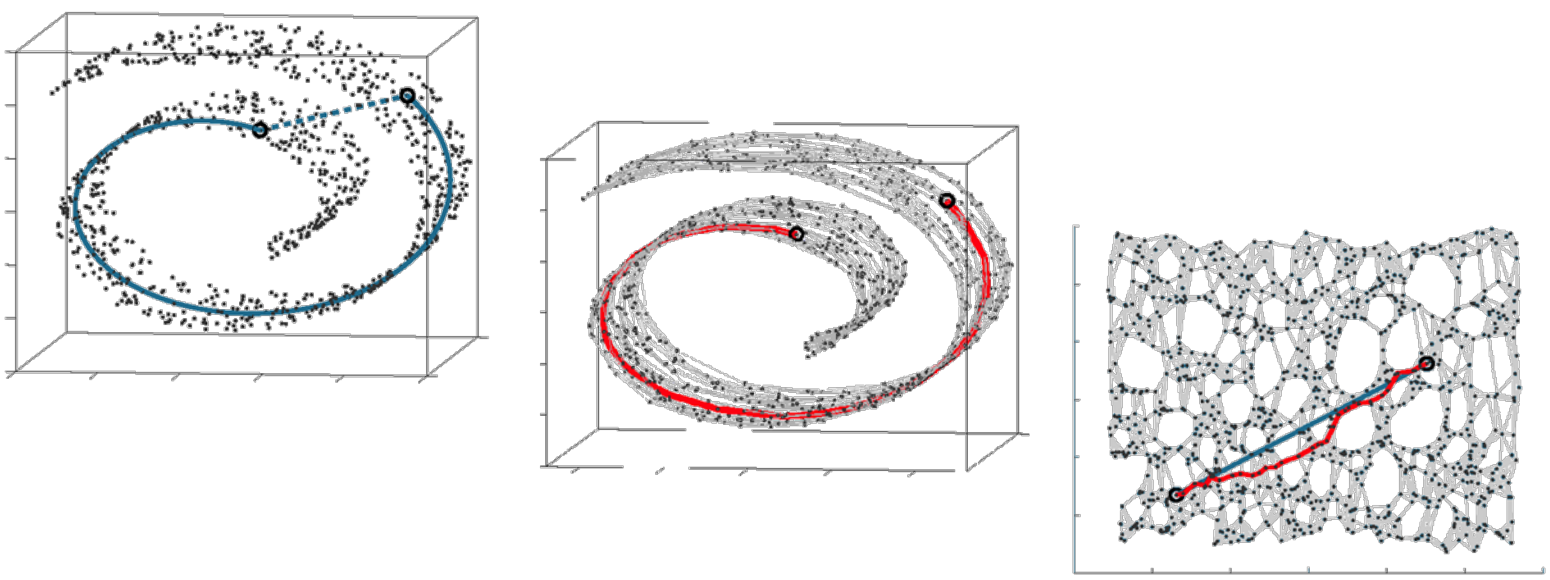
- **Step 3**
  - Constructing an embedding of the data in *d*-dimensional Euclidean space *Y* that best preserves the inter point distances

- This is of course nothing but an MDS problem



Recovery Guarantees

- Isomap is guaranteed asymptotically to recover the true dimensionality and geometric structure of nonlinear manifolds.
- As the sample data points increases, the graph distances provide increasingly better approximations to the intrinsic geodesic distances.



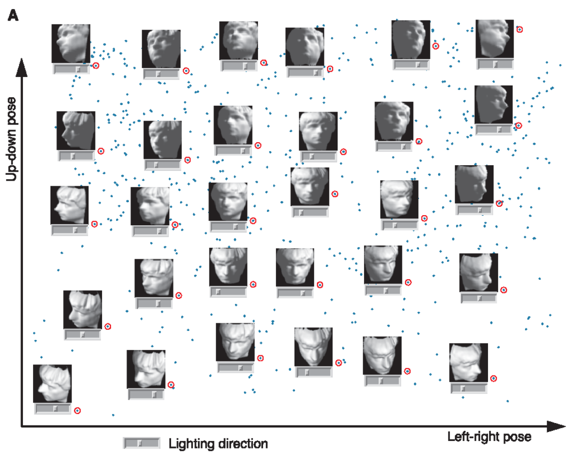
ISOMAP Examples

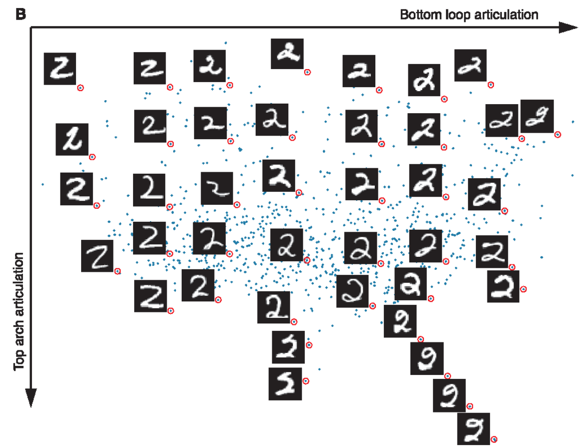
 # Face

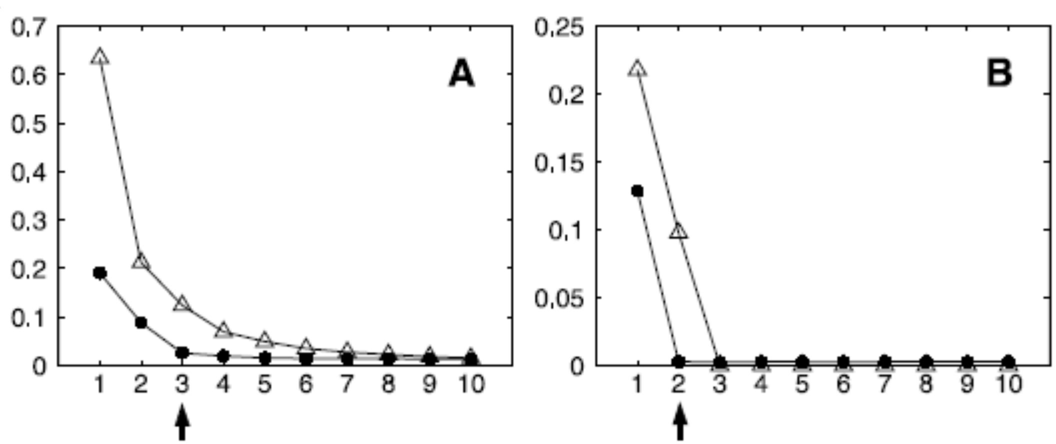
: face pose and illumination

Hand writing

: bottom loop and top arch

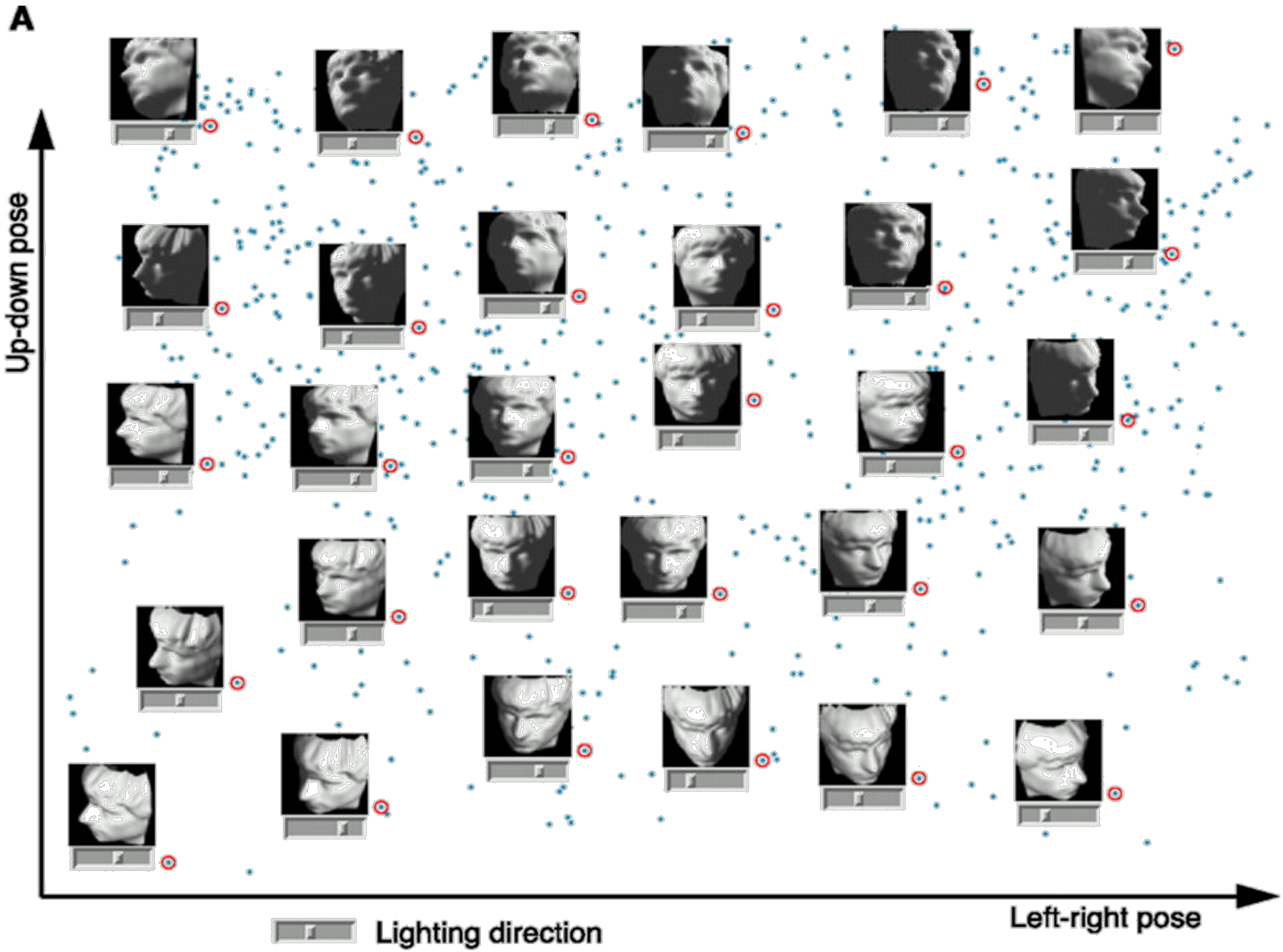






MDS : open triangles

Isomap : filled circles

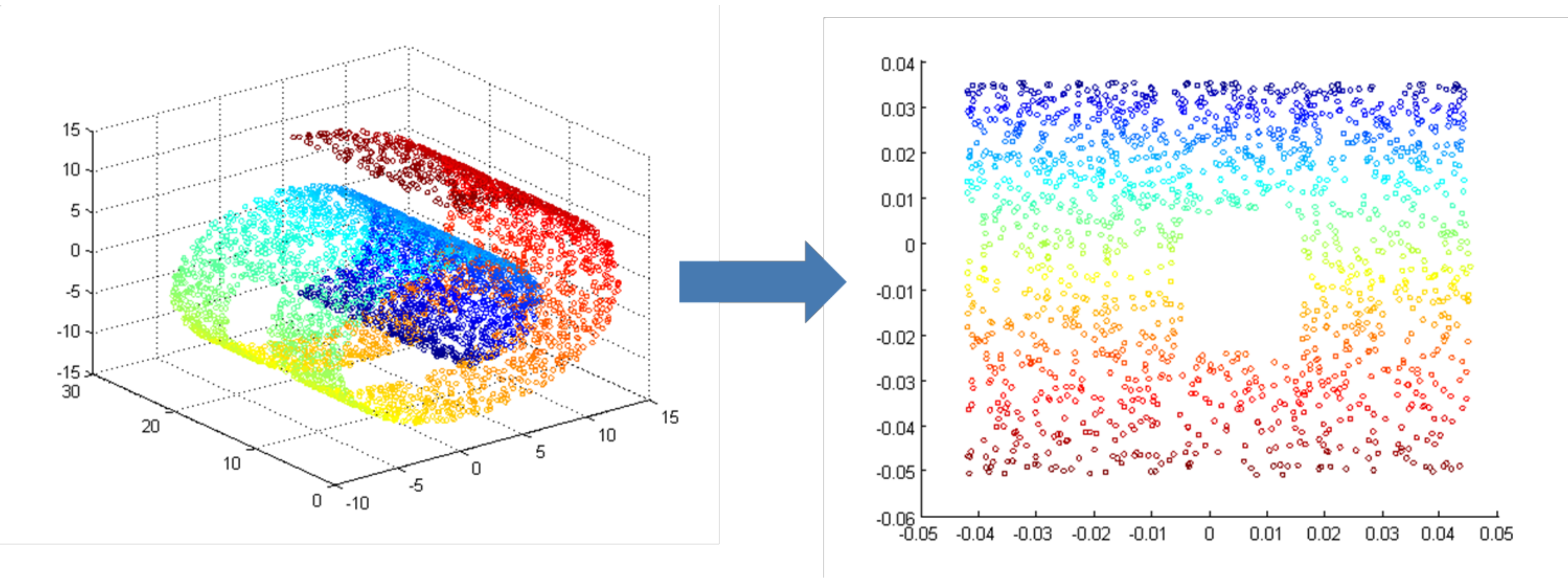

A
Up-down pose
Left-right pose
Lighting direction

UC San Diego

Laplacian Eigenmaps

Laplacian Eigenmaps
(M. Belkin, P. Niyogi)

- Start same as Isomap, but use a spectral embedding in lieu of MDS



Hole distorts long geodesic distances

Locally Linear Embeddings

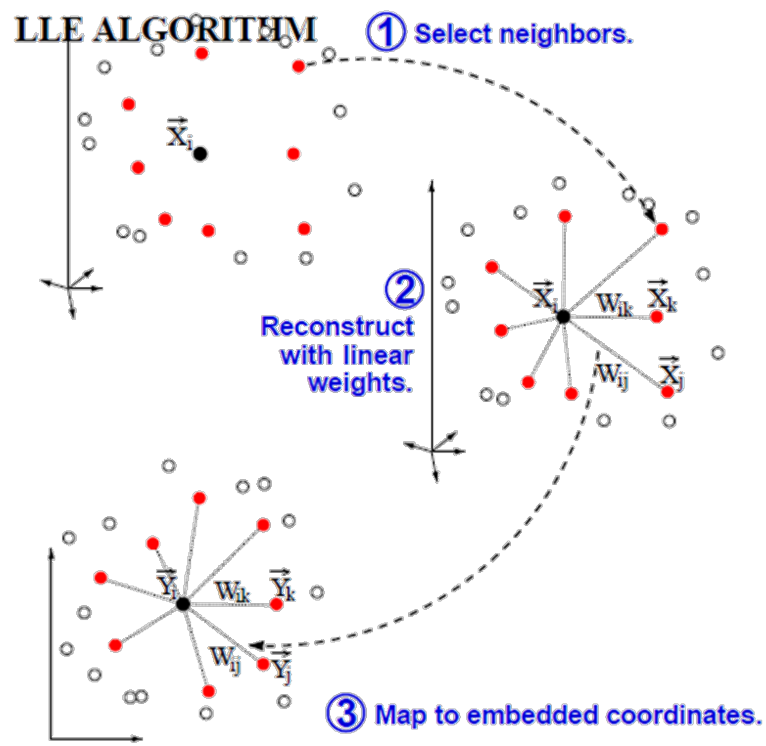
Locally Linear Embeddings (LLE) (S. T. Roweis and L. K. Saul)

- Define neighborhood relations between points (build NN graph)
  - *k* nearest neighbors
  - ε-balls
- Find weights that reconstruct each data point from its neighbors:

$$\min_{\sum_j w_{ij}=1} \left\| \mathbf{x_i} - \sum_{j \in N(i)} w_{ij} \mathbf{x_j} \right\|^2$$

- Find low-dimensional coordinates so that the same weights hold: $\mathbf{x_1}', \ldots, \mathbf{x_n}' \in R^d$

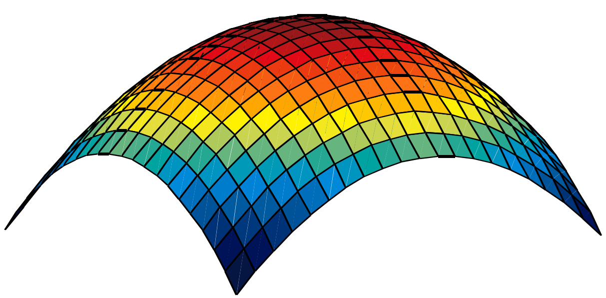
$$\min_{\mathbf{x_1}', \ldots, \mathbf{x_n}'} \sum_i \left\| \mathbf{x_i}' - \sum_{j \in N(i)} w_{ij} \mathbf{x_j}' \right\|^2$$



$$x_i' = Y_i$$

From Local to Global

- The weights w_{ij} capture the local shape
  - Invariant to translation, rotation and scale of the neighborhood
  - If the neighborhood lies on a manifold, the *local* mapping from the global coordinates (R^D) to the surface coordinates (R^d) is almost linear
  - Thus, the weights w_{ij} should hold also for manifold (R^d) coordinate system!



$$\min_{\sum_j w_{ij}=1} \left\| \mathbf{x_i} - \sum_{j \in N(i)} w_{ij} \mathbf{x_j} \right\|^2$$

$$\min_{\mathbf{x_1}',\ldots,\mathbf{x_n}'} \sum_i \left\| \mathbf{x_i}' - \sum_{j \in N(i)} w_{ij} \mathbf{x_j}' \right\|^2$$

Solving the Minimizations

- Linear least squares (using Lagrange multipliers)

$$\min_{\sum_j w_{ij}=1} \left\| \mathbf{x_i} - \sum_{j \in N(i)} w_{ij} \mathbf{x_j} \right\|^2$$

- To find $\mathbf{x_1}', \ldots, \mathbf{x_n}' \in R^d$ that minimize,

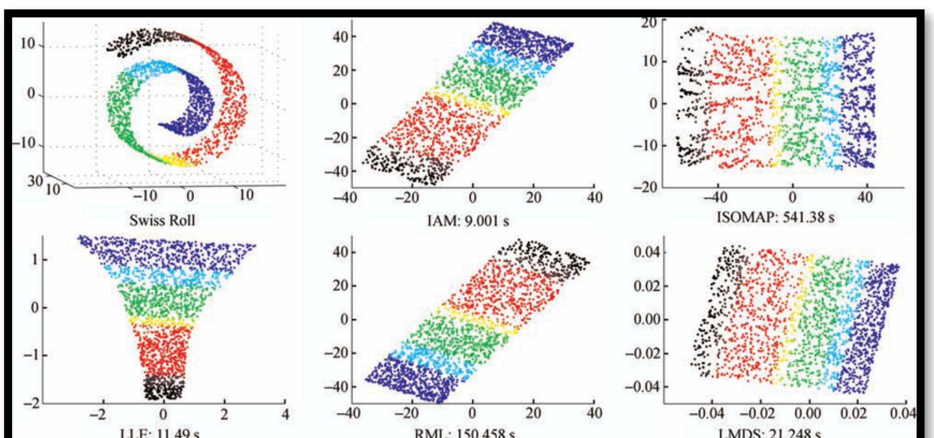
$$\min_{\mathbf{x_1}', \ldots, \mathbf{x_n}'} \sum_i \left\| \mathbf{x_i}' - \sum_{j \in N(i)} w_{ij} \mathbf{x_j}' \right\|^2$$

a sparse eigenvalue problem is solved. Additional constraints are added for conditioning:

$$\sum_i \mathbf{x_i}' = 0, \qquad \frac{1}{n} \sum_i \mathbf{x_i}' \mathbf{x_i}'^T = I$$

Comparison: ISOMAP vs. LLE

ISOMAP	LLE
Global distances	Local averaging
k-NN graph distances	*k*-NN graph weighting
Largest eigenvectors	Smallest eigenvectors
Dense matrix	Sparse matrix



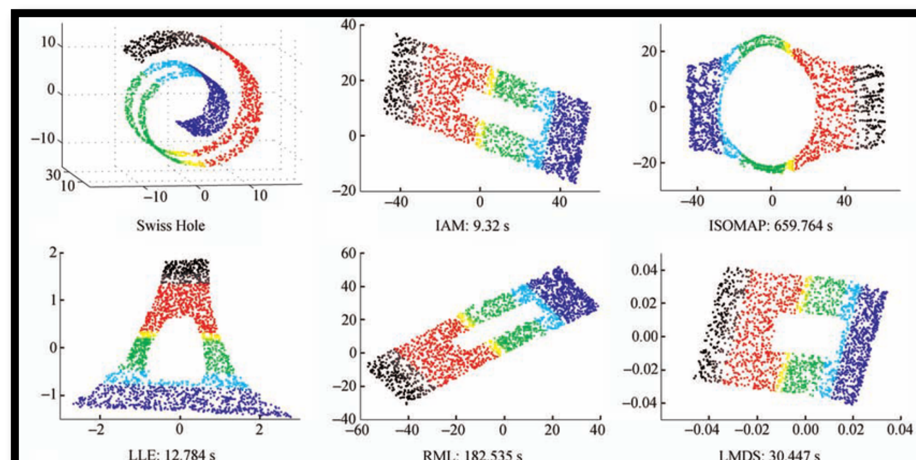


Image from "Incremental Alignment Manifold Learning." Han et al. JCST 26.1 (2011).

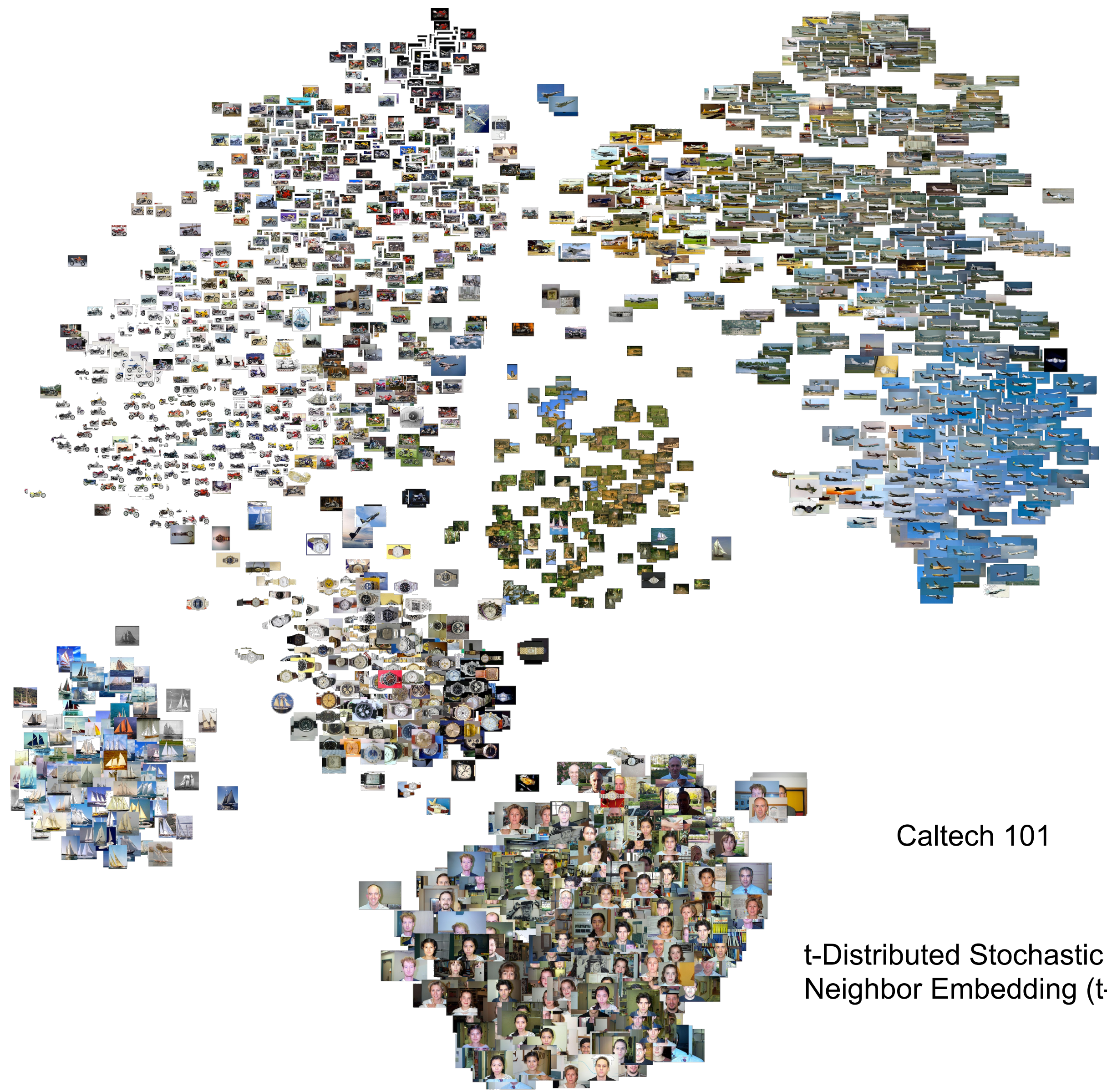
Many More Methods

Many NLDR Methods

-

Contents [hide]

From Wikipedia



Caltech 101

t-Distributed Stochastic Neighbor Embedding (t-SNE)