Big Data

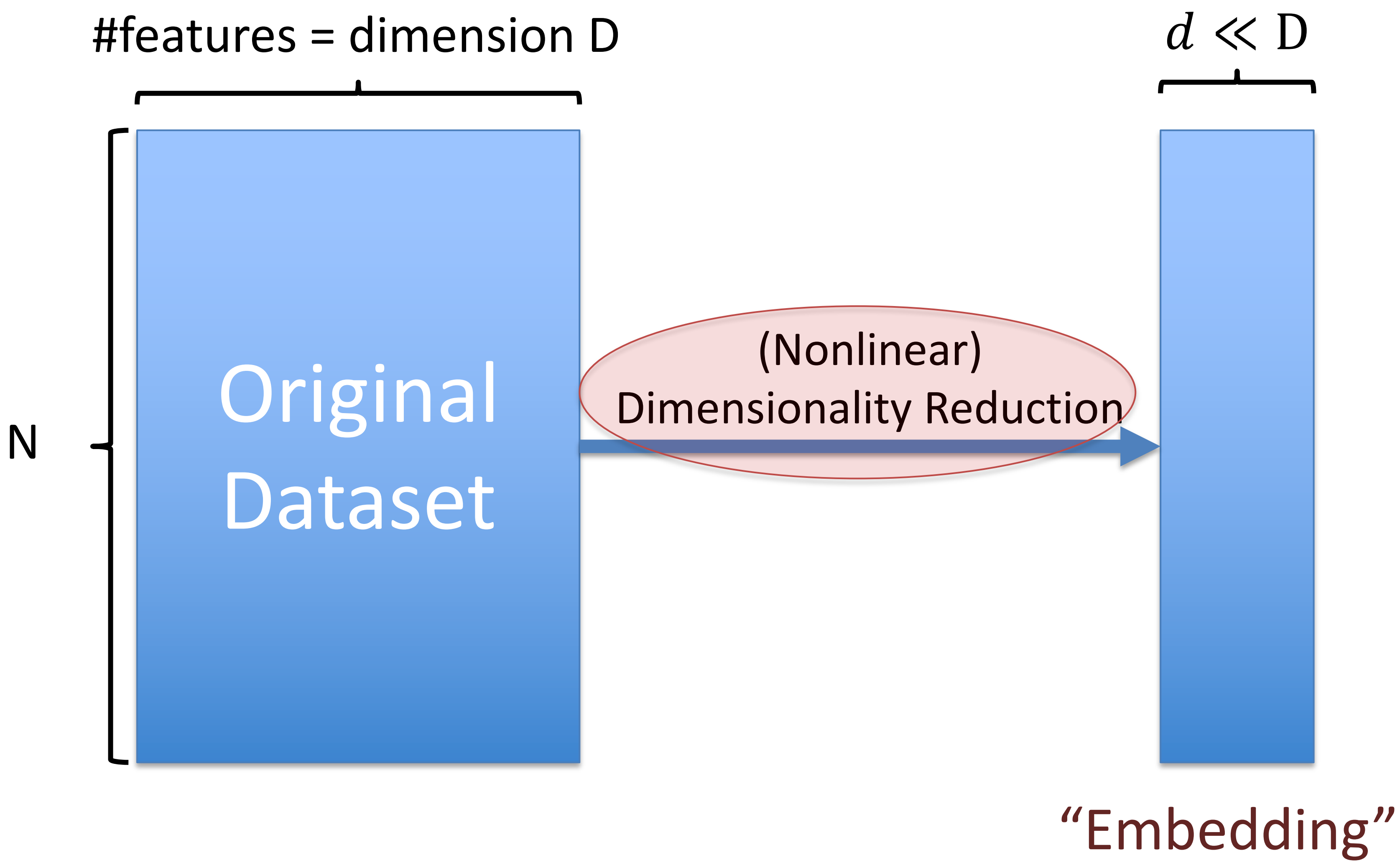
VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

AIML427

Manifold Learning/
Nonlinear Dimensionality Reduction (1)

Dr Bach Nguyen
(with thanks to Dr Andrew Lensen)
Bach.Nguyen@vuw.ac.nz

Dimensionality Reduction



How do we do the (NL)DR?

- We've discussed a lot of supervised methods
  - Filter: information gain, mutual information, correlation, …
  - Wrapper: Training a classifier given a feature subset
  - Embedded: Decision Trees, Genetic Programming

- But what about if you have **no labels?**
- Or what if you want a (filter) method that considers the entire set of feature relationships?
  - Pairwise redundancies feel "naïve"
  - What about the topology of the data?

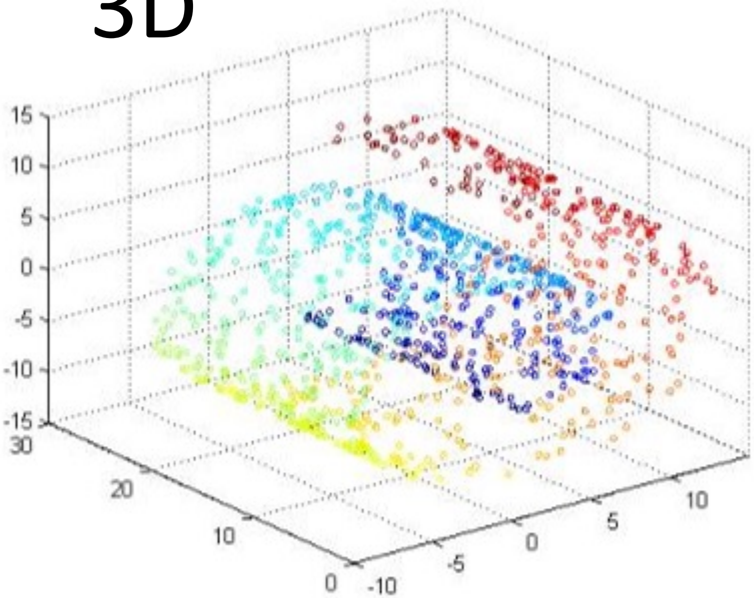
- Let us make an "assumption": our data is **actually** (*intrinsically*) lower-dimensional than the D features we have
  - We say it lies on a lower-dimensional **manifold (embedding)**

Swiss Rolls

This Photo by Unknown Author is licensed under CC BY-NC-ND

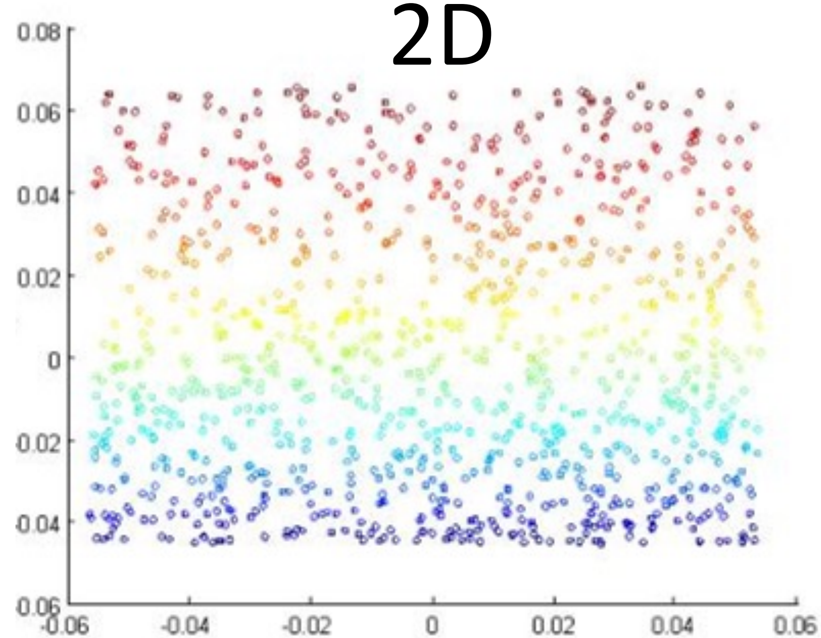
…mathematically speaking

3D



Nonlinear Transformation

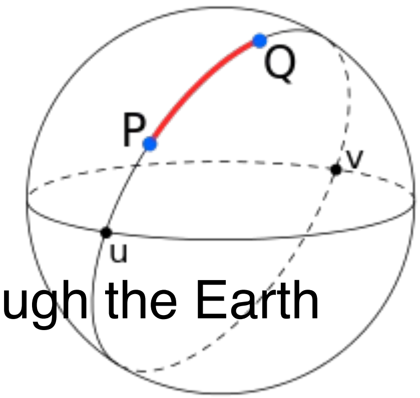
2D

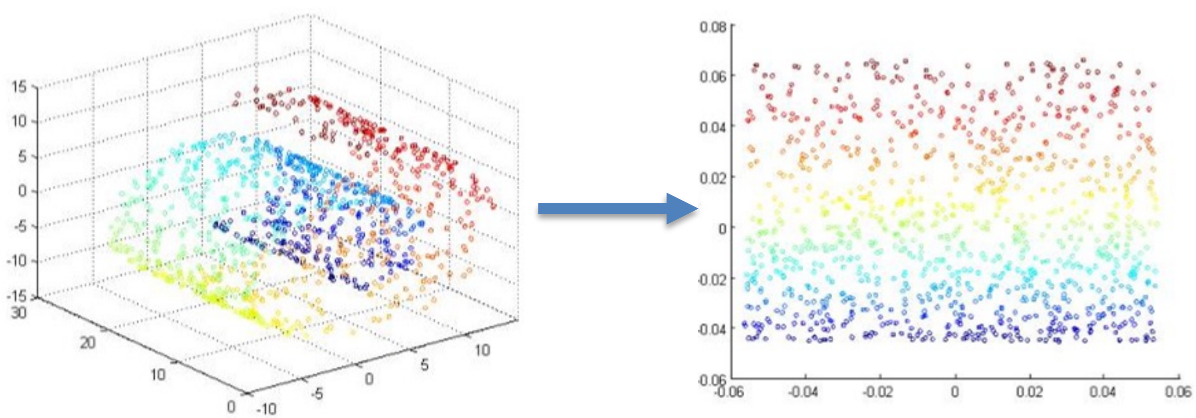


Did we lose any structure?

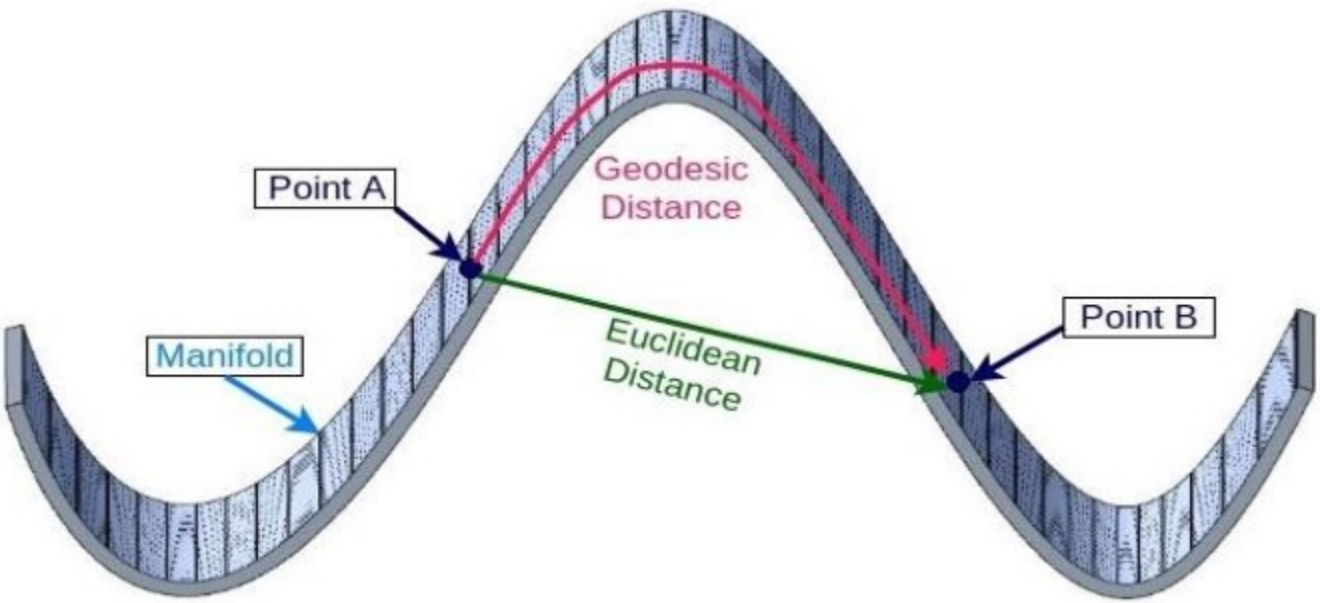
Well…

- The red and blue points were close in Euclidean (straight-line) space in 3D – but aren't in 2D
- …but this is desirable, as there is a "gap" in the topology
- We have preserved the **geodesic distance**
  - Geodesic Distance = shortest path in a graph
  - E.g. distance from Auckland to London flying vs through the Earth





Geodesic Distance



Manifolds

- In mathematics, a manifold is a topological space that locally resembles Euclidean space near each point.
  - Small portions of a circle look like a line
  - “Small” portions of the Earth look flat
  - Topology ignores *bending*

I thought this was an AI class…

- It is!

- The key takeaway: a manifold is a d-dimensional object "living" in a D-dimensional world.

- If we can find or *approximate* this manifold, we can represent our data in d dimensions instead of D!

- We call this d-dimensional space an embedding – it is embedded within the D-dimensional space.

Manifold Learning

- We want to *find* a smaller embedding of our data
- Manifold learning (MaL): using machine learning to learn an embedding
- Nonlinear dimensionality reduction (NLDR): a broader term: any approach to reducing dimensionality through nonlinear transformations
- In practice, the two terms are used pretty interchangeably

Manifold Learning

- How do we **know** our data lies on a manifold?
- Ways to estimate the intrinsic dimensionality of data
- In practice, an *approximation* of the original topology is perfectly useable.
- As with PCA – if we retain the majority of the variance/structure, we likely retain the important patterns
- Feature construction? Yes…and no. We'll come back to that

“CLASSIC” MANIFOLD LEARNING ALGORITHMS

Multidimensional Scaling (MDS)

- Dates from as early as 1938 (!)

- Minimises the difference between the *pairwise* distance (between instances) in the high-dimensional vs embedding space:

$$E_M = \sum_{k \neq l} [d(k,l) - d'(k,l)]^2$$

where *k* and *l* are two instances, $d(k,l)$ is the distance in embedded space and $d'(k,l)$ the distance in high-dim space

Multidimensional Scaling (MDS)

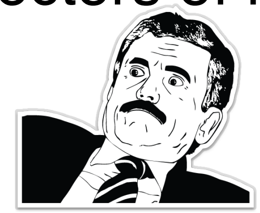
- *Metric* MDS:
  - Preserve magnitudes and orders
  - Assume triangle inequality
  - Sensitive to outliers
  - *Sammon's Mapping* normalises each distance

- *Nonmetric* MDS uses ranks instead of raw distance (*why?*)
  - Preserve orders
  - Less sensitive to triangle inequality and outliers
  - Lost information

- **How** does it find the embedding?
  - **Optimisation problem…**
  - Numerical optimisation techniques
  - Or Eigen-decomposition…

Isomap (Isometric Mapping)

1. Construct neighbourhood graph (connect $a, b\ if\ d(a, b) < \epsilon$)
2. Compute shortest paths (geodesic distance matrix)
3. Construct d-dimensional embedding (eigenvectors of matrix)



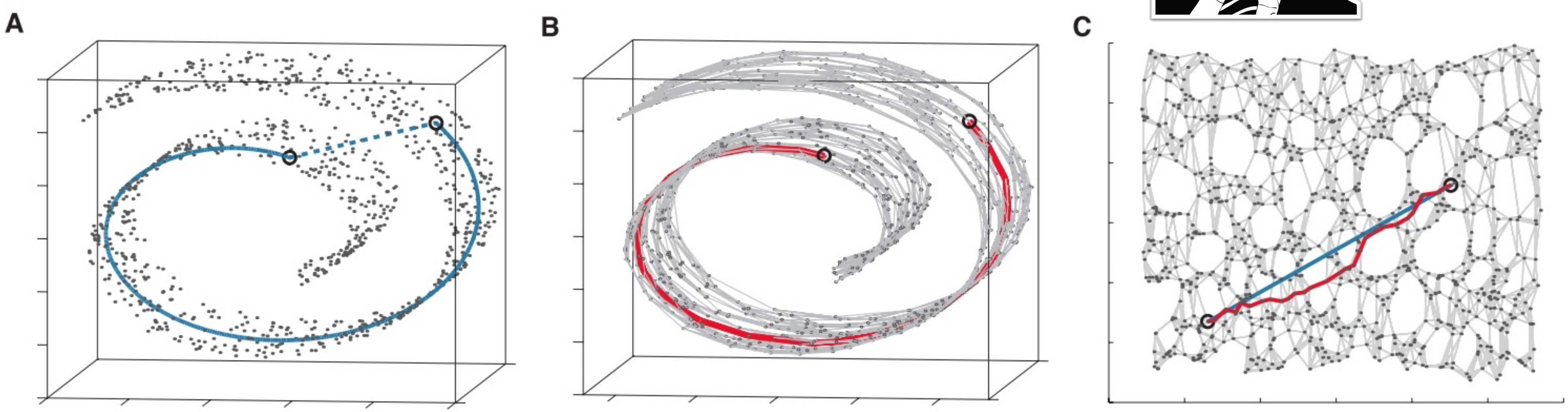


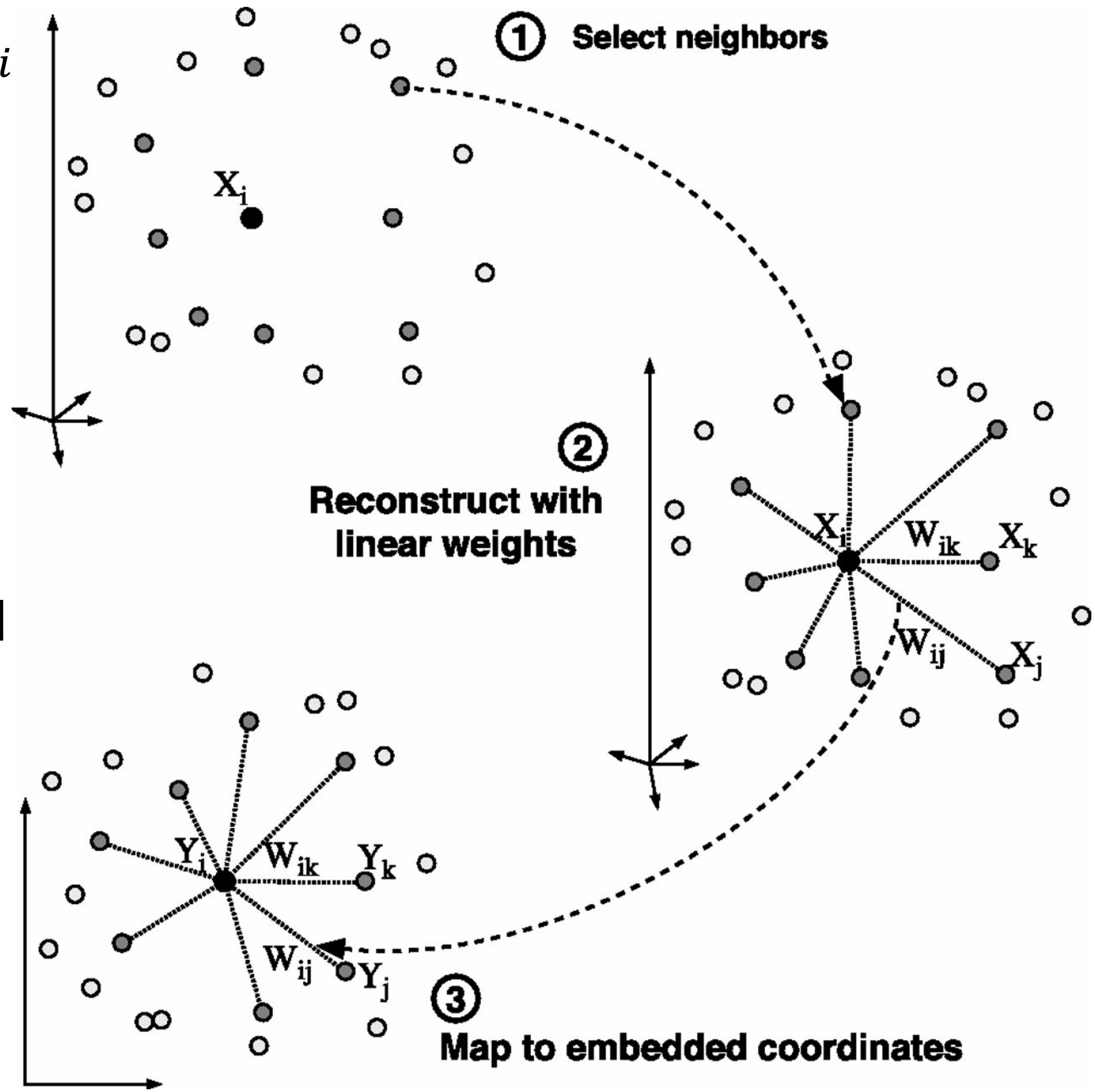
Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (**A**) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold (length of solid curve). (**B**) The neighborhood graph *G* constructed in step one of Isomap (with $K = 7$ and $N = 1000$ data points) allows an approximation (red segments) to the true geodesic path to be computed efficiently in step two, as the shortest path in *G*. (**C**) The two-dimensional embedding recovered by Isomap in step three, which best preserves the shortest path distances in the neighborhood graph (overlaid). Straight lines in the embedding (blue) now represent simpler and cleaner approximations to the true geodesic paths than do the corresponding graph paths (red).

Tenenbaum, J. B., Silva, V. D., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319-2323. https://www.science.org/doi/10.1126/science.290.5500.2319

Locally Linear Embedding (LLE)

1. Select (K) neighbours for each datapoint X_i
2. Compute weights W_{ij} to reconstruct X_i from its neighbours (least-squares)
3. Do some *fancy linear algebra* to find the embedding that best minimises the reconstruction error.

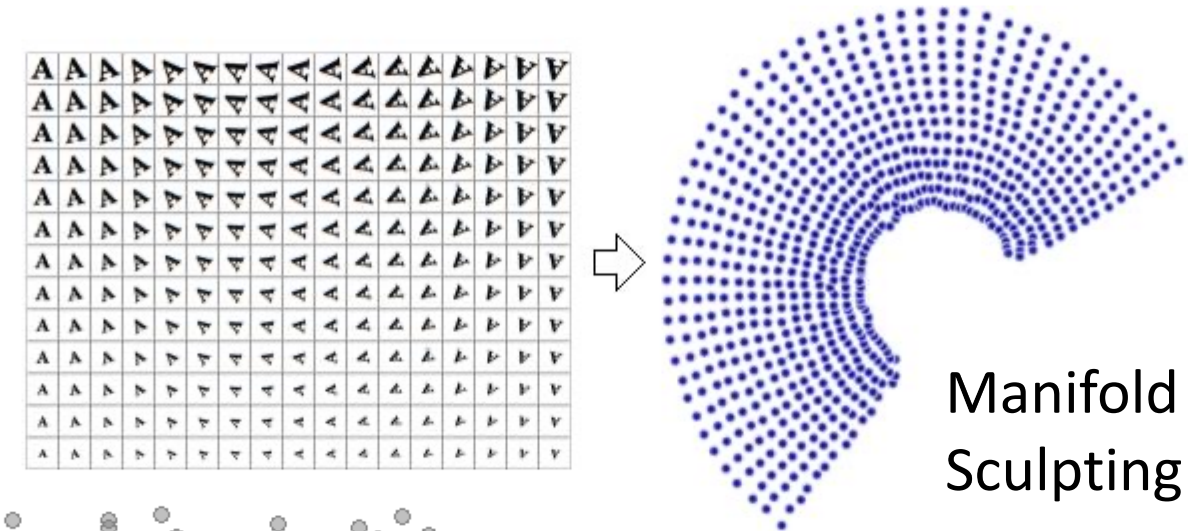
NB W_{ij} is unchanged



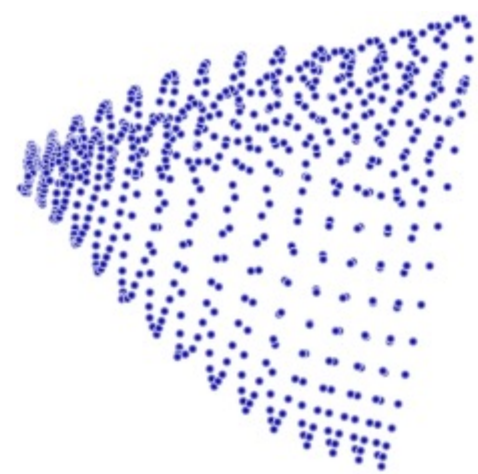
Types of Manifold Learning

- Mapping vs non-mapping methods
  - Embedding = f(high-dim space)? “See” the functional mapping; vs
  - Embedding is optimised “with respect to” high-dim space
- Local vs global methods
  - Preserve local neighbourhoods more; vs
  - Ensure global structure is maintained
  - (Both?)
  - Isomap vs LLE?
- Deterministic/analytical vs (stochastic) search
  - Eigen-decomposition/numeric optimisation; vs
  - Gradient descent/EC/…

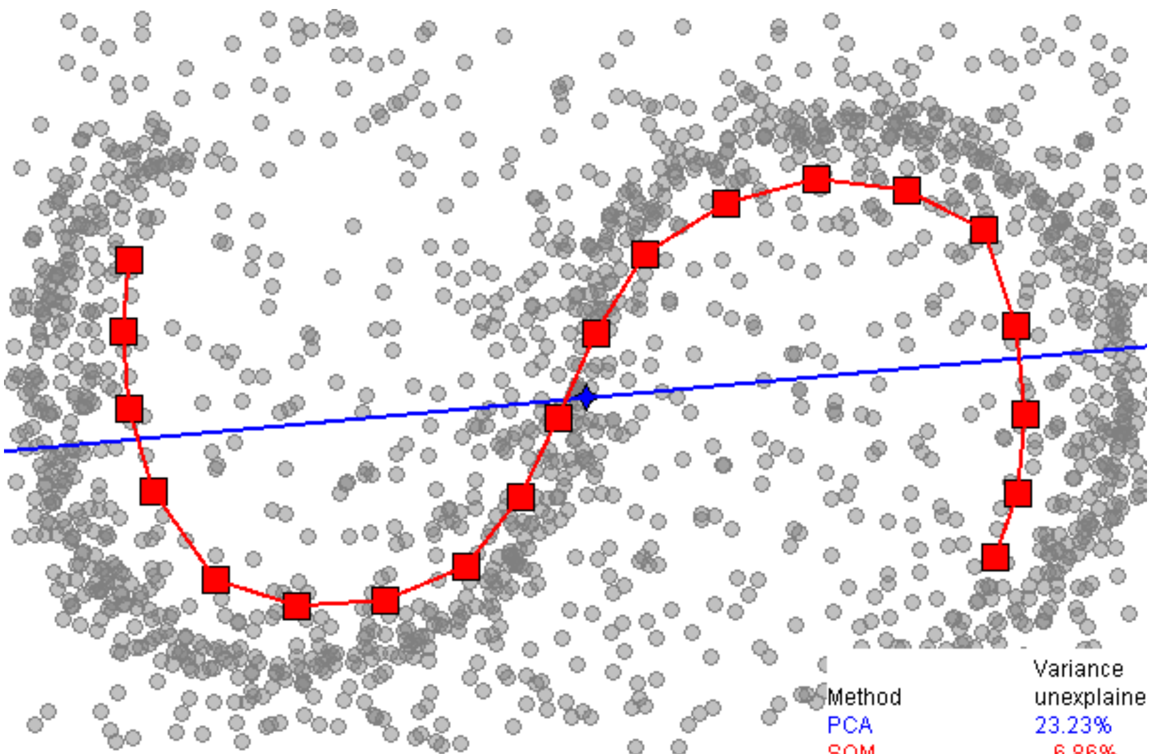
PCA vs NLDR?



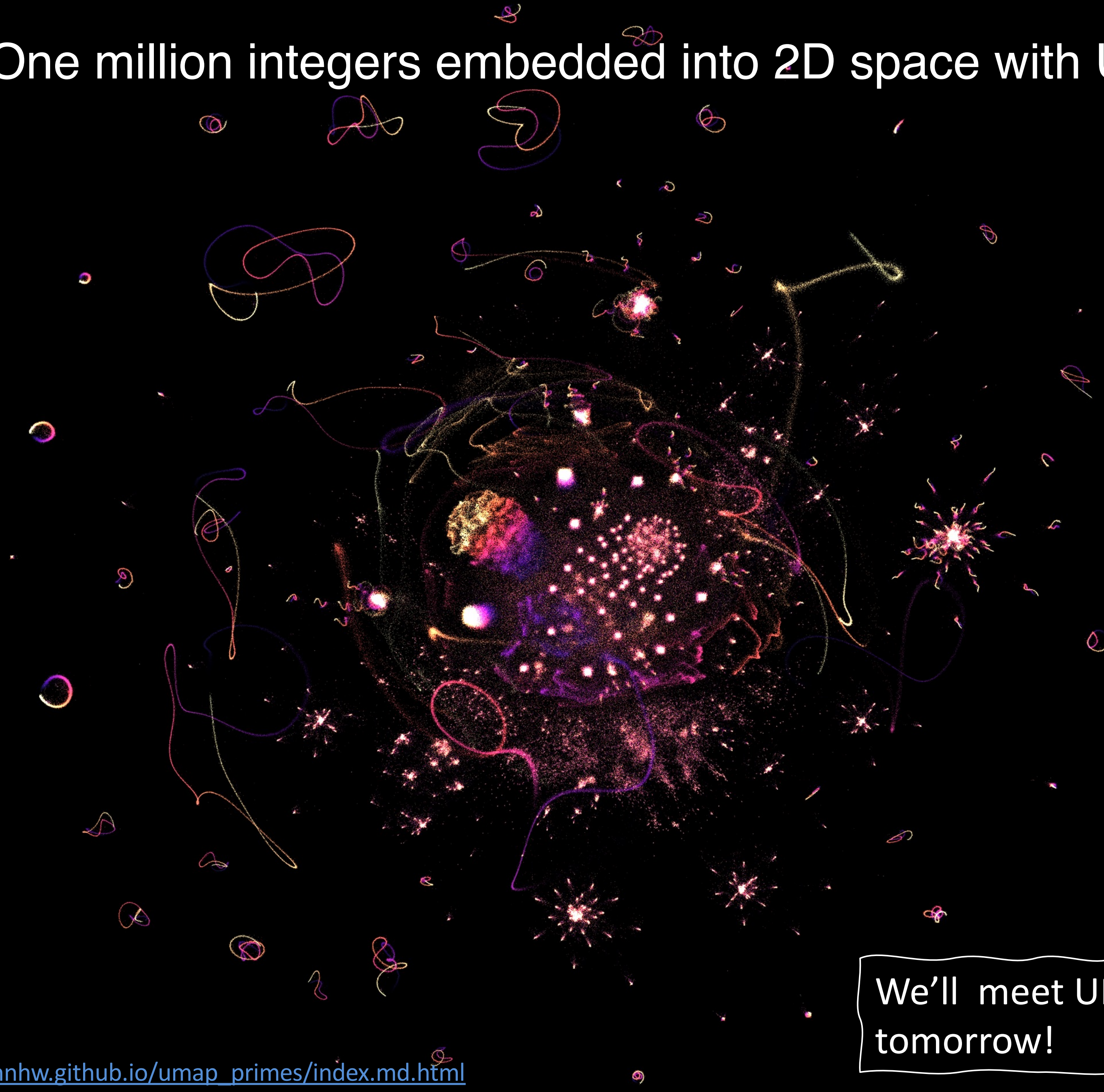
Manifold
Sculpting



PCA



One million integers embedded into 2D space with UMAP



We'll meet UMAP tomorrow!

https://johnhw.github.io/umap_primes/index.md.html