DLI Accelerated Data Science Teaching Kit

Lecture 15.6 - UMAP

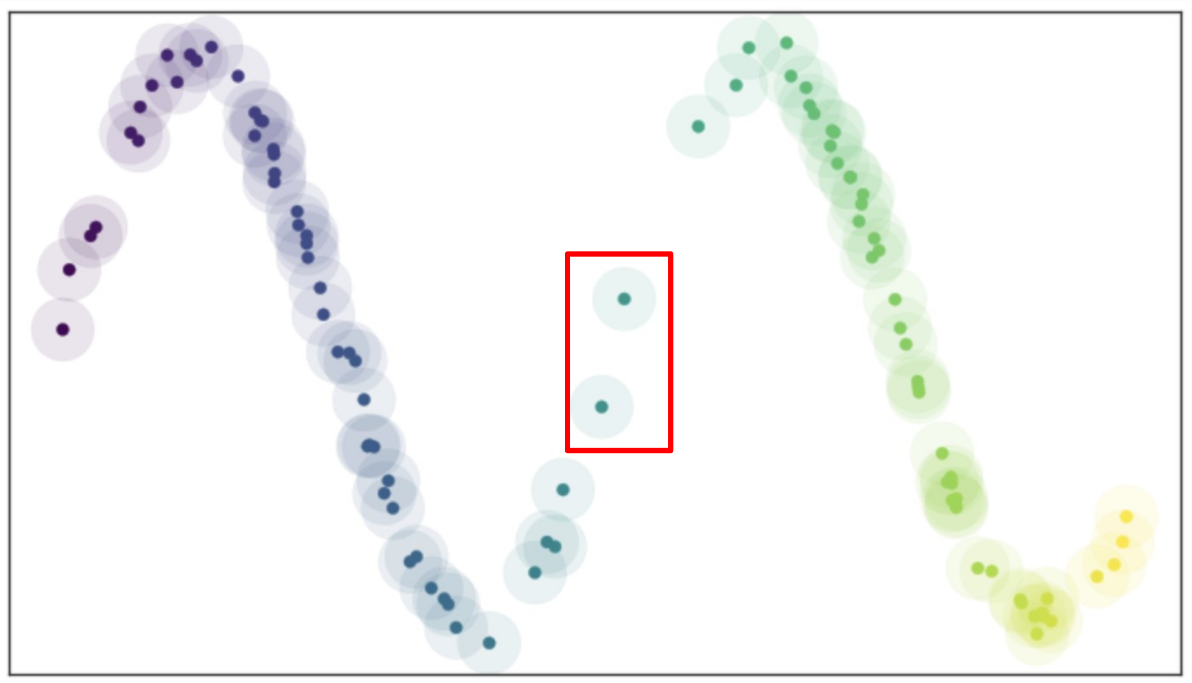
UMAP for Dimensional Reduction

- Matrix Factorization
  - Example: Principle Component Analysis
  - Good at capturing the Global Structure of the data
  - Only keeping the principle component, meaning there is a loss in information
- Neighbor Graph
  - Example: **UMAP**, t-SNE
  - Good at capturing the Local Strcutre of the data
  - Simplices: Topological structure in multi dimentional space
  - Nerve Theorem: We can keep all information in the topological space

UMAP Overview

Uniform Manifold Approximation and Projection

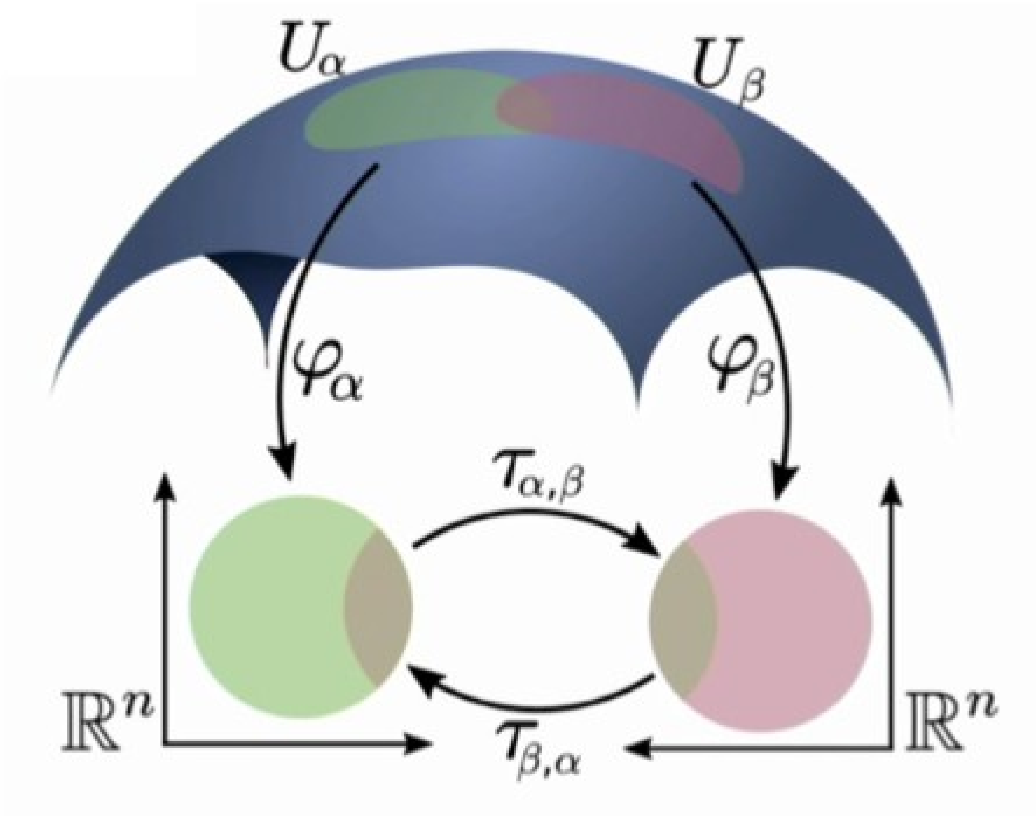
- Based on creating simplex in high-dimensional space
  - Points are connected with a line if the distance between them is under a certain threshold
  - We can use different distance metrics (e.g., Euclidean)
- **Problem**: Data are not usually uniformly distanced
  - We can have points that are disconnected from other points



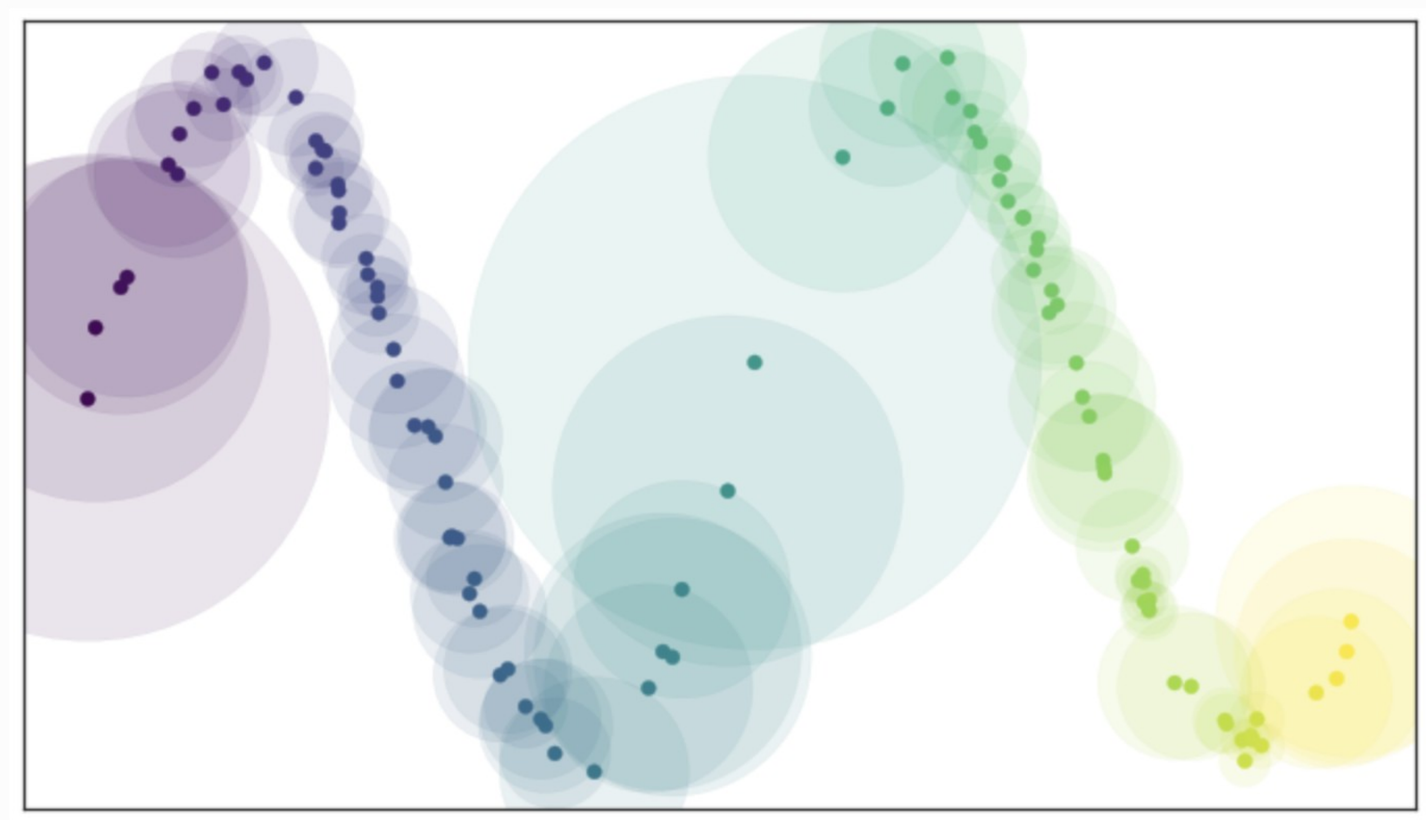
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

Uniform Manifold

- **Solution**: Uniform Manifold & Riemannian Metrics
- Stretching or shrinking according to where the data appear sparser or denser
- We define a **Uniform Manifold** where each points are equally distanced from each other



Manifold & Riemannian Metrics

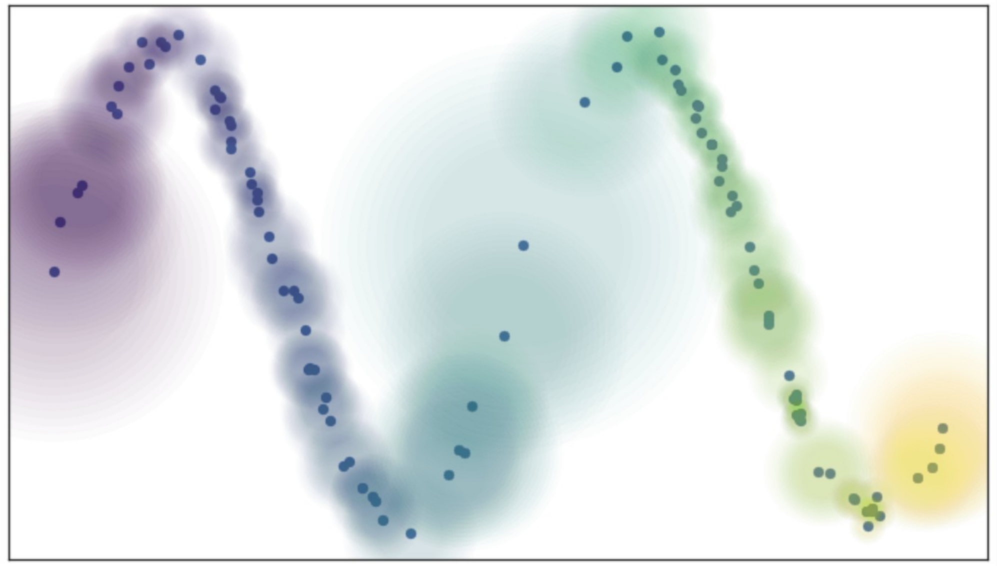


Distance in the manifold projected onto the real space

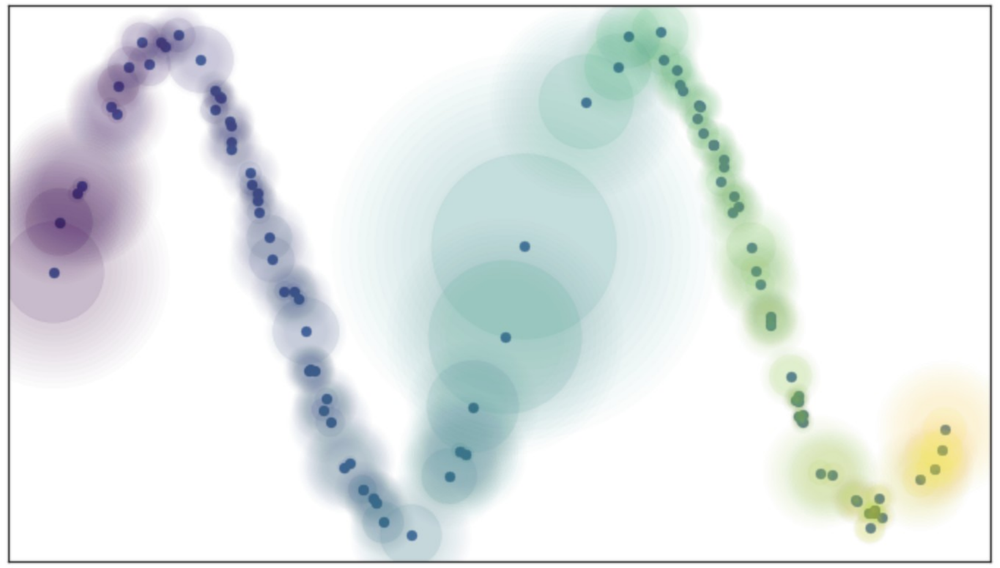
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

Fixed Radius vs. Fuzzy cover

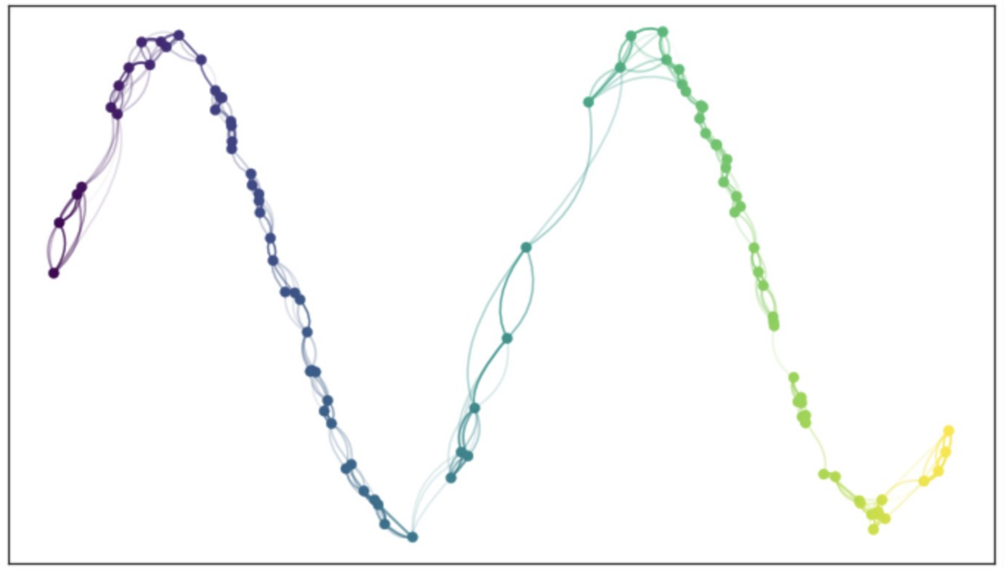
- We can now generate a simplex where all data points are connected
- **Problem**: Cannot differentiate distance in this simplex.
  - We are using a fixed radius to determine if two data points should be connected.
- **Solution**: Fuzzy cover
  - We still need the manifold to be locally connected



Fuzzy Cover



Fuzzy Cover + Locally connected

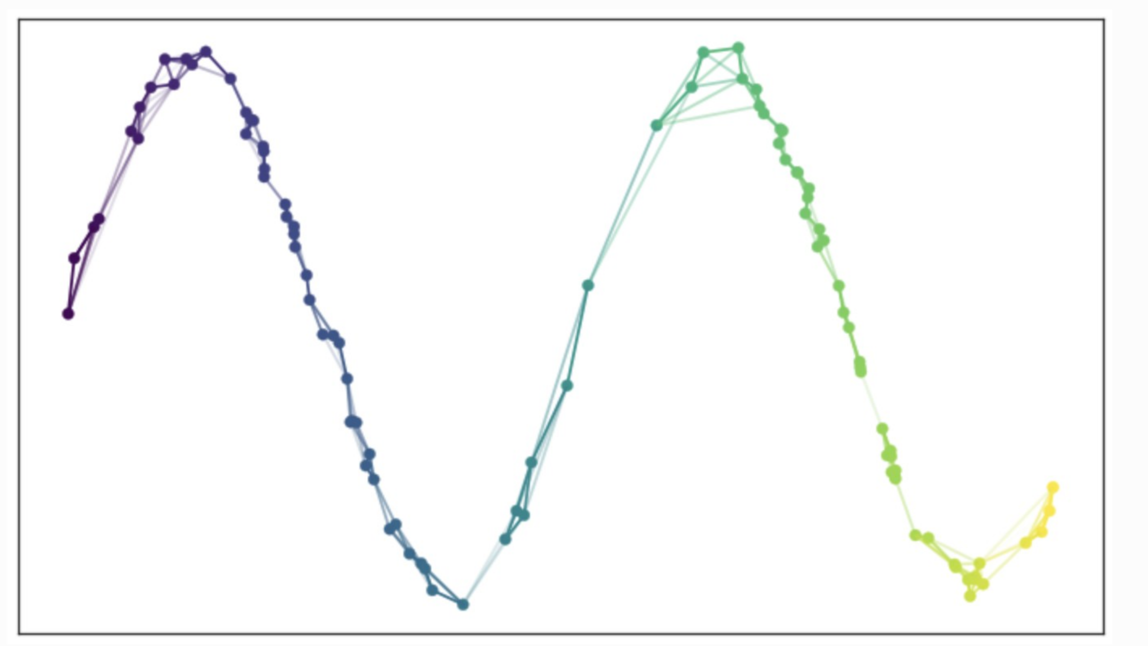


Edges with incompatible weight
(Differentiate by different color)

https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

UMAP Adjunction

- **Problem**: Local metrics are not compatible
- **Solution**: UMAP Adjunction
- We can combine weights in different edges in this form: $f(\alpha, \beta) = \alpha + \beta - \alpha\beta$



Graph with combined weight

https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

UMAP Hyperparameter

- n_neighbors
  - The number of approximate nearest neighbors used to construct the initial high-dimensional graph
  - Most important
  - Local versus global structure
  - Low: focus more on local structure
  - High: focus more on global structure
- min_dist
  - The minimum distance between points in low-dimensional space
  - How tightly UMAP clumps points together
  - Low: More tightly packed embeddings
  - High: More loosely packed embeddings

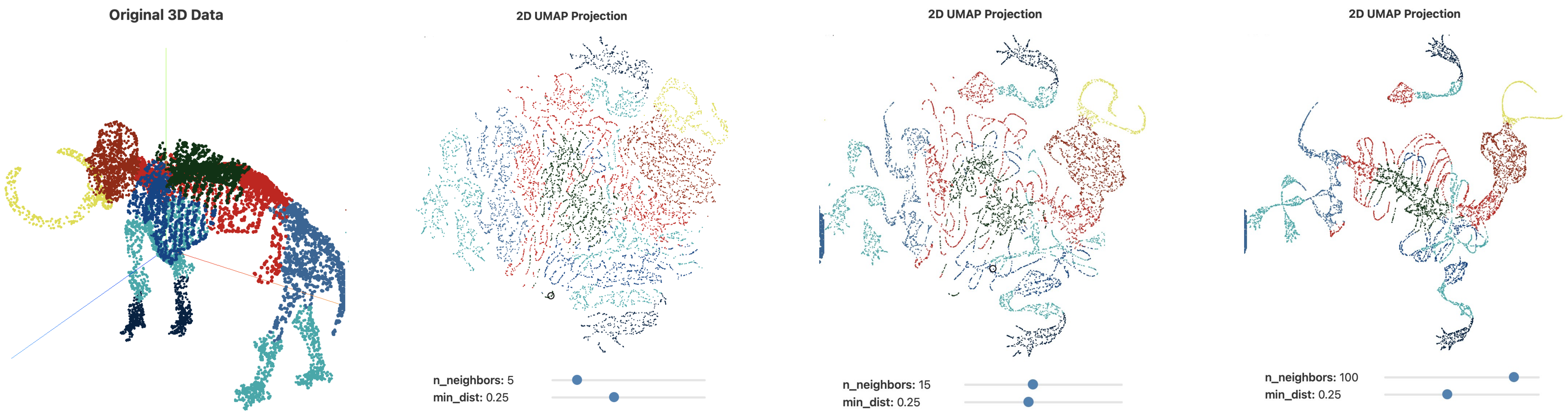
https://pair-code.github.io/understanding-umap/index.html

UMAP Hyperparameter

- n_neighbors
  - The number of approximate nearest neighbors used to construct the initial high-dimensional graph
  - Most important
  - Local versus global structure
  - Low: focus more on local structure
  - High: focus more on global structure
- min_dist
  - The minimum distance between points in low-dimensional space
  - How tightly UMAP clumps points together
  - Low: More tightly packed embeddings
  - High: More loosely packed embeddings

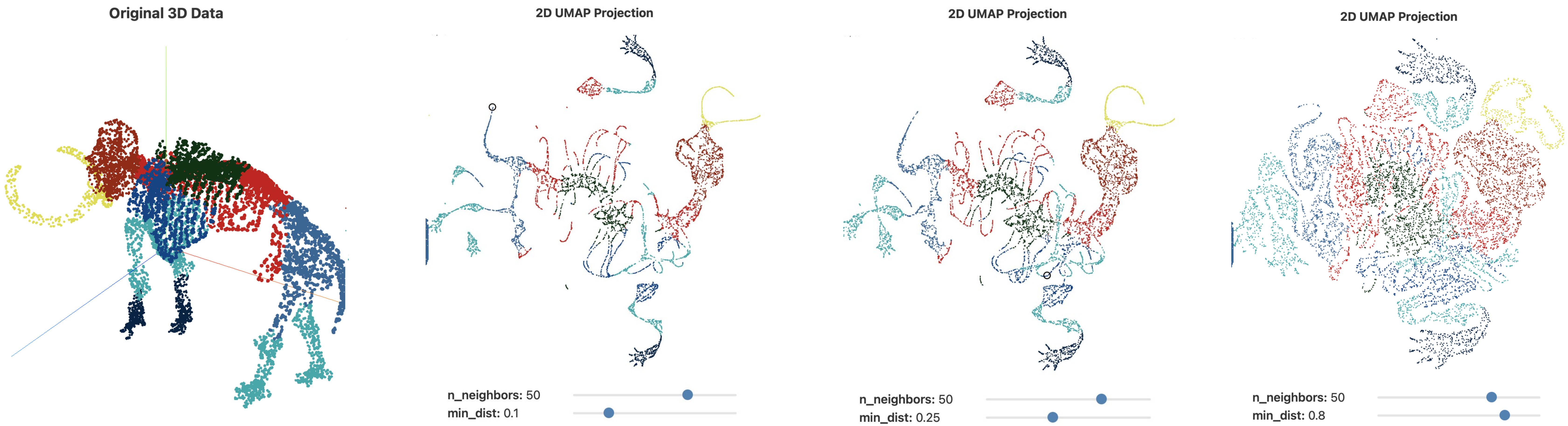
https://pair-code.github.io/understanding-umap/index.html

UMAP Hyperparameter (n_neighbors)



https://pair-code.github.io/understanding-umap/index.html

UMAP Hyperparameter (min_dist)



https://pair-code.github.io/understanding-umap/index.html

Performance

	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

	UMAP speed up over t-SNE
COIL20	3x
MNIST	13x
Fashion MNIST	11x
GoogleNews	19x

https://www.youtube.com/watch?v=nq6iPZVUxZU

DLI Accelerated Data Science Teaching Kit

Thank You