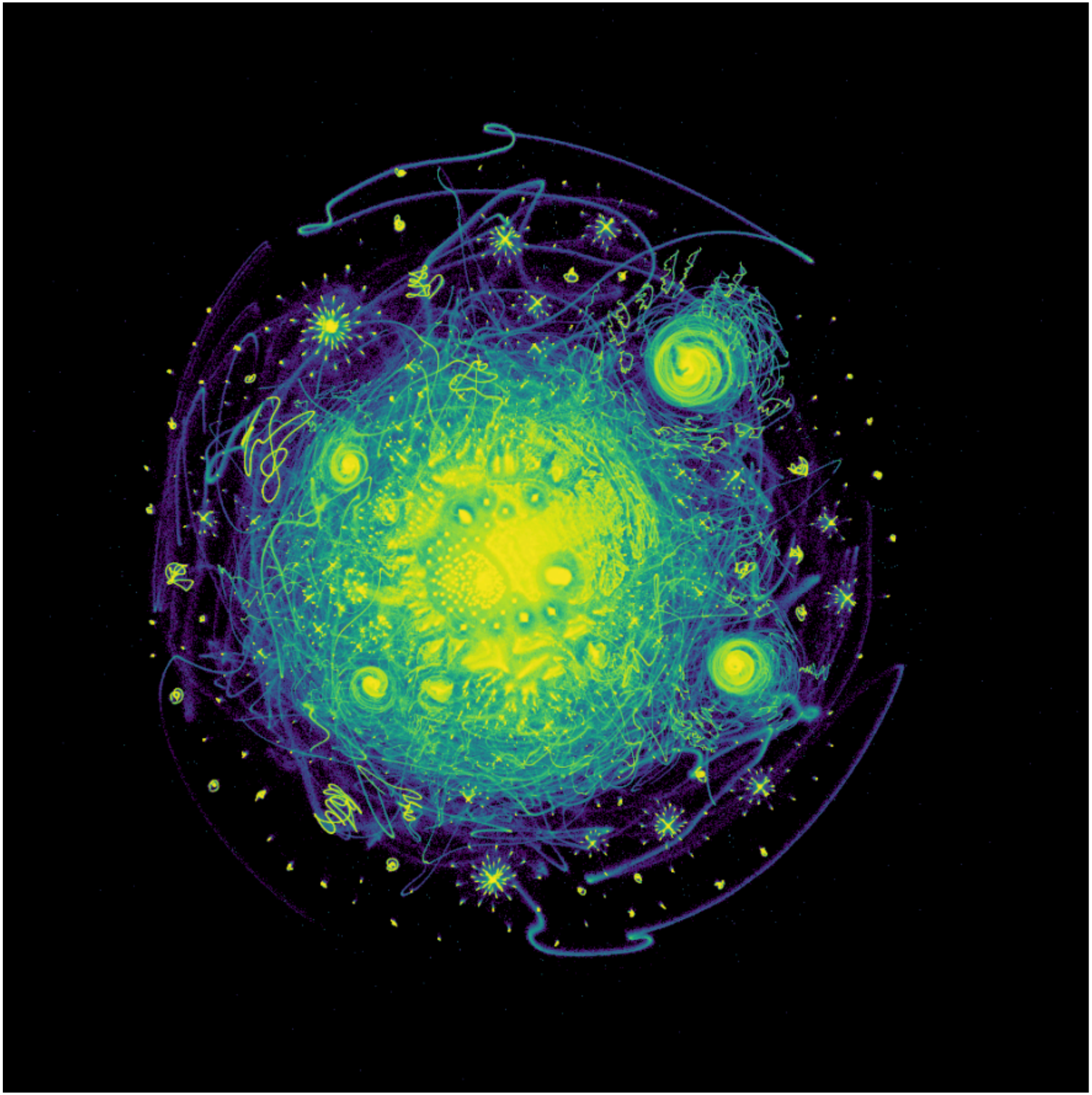


UMAP

UMAP

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Leland McInnes, John Healy, James Melville

https://doi.org/10.48550/arXiv.1802.03426

First submission: 2018

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Leland McInnes
Tutte Institute for Mathematics and Computing
leland.mcinnes@gmail.com

John Healy
Tutte Institute for Mathematics and Computing
jchealy@gmail.com

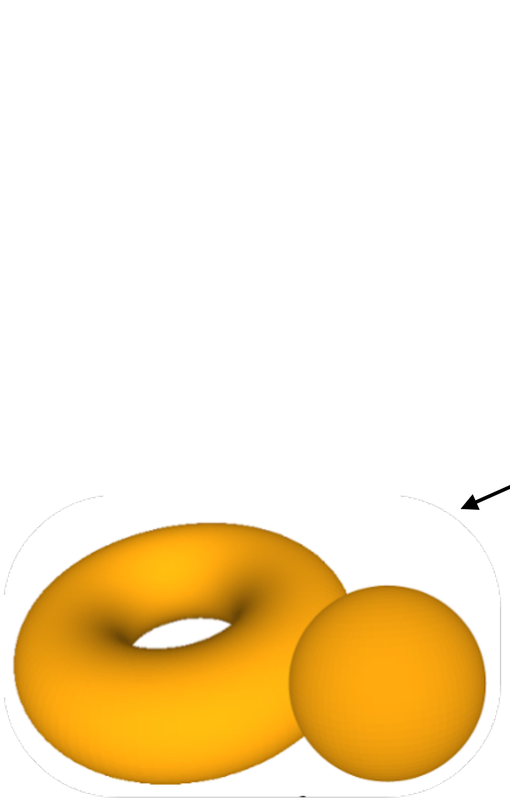
James Melville
jlmelville@gmail.com

September 21, 2020

Abstract

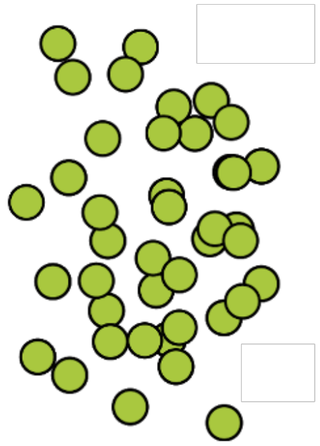
UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction. UMAP is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. The result is a practical scalable algorithm that is applicable to real world data. The UMAP algorithm is competitive with t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance. Furthermore, UMAP has no computational restrictions on embedding dimension, making it viable as a general purpose dimension reduction technique for machine learning.

UMAP - intuition



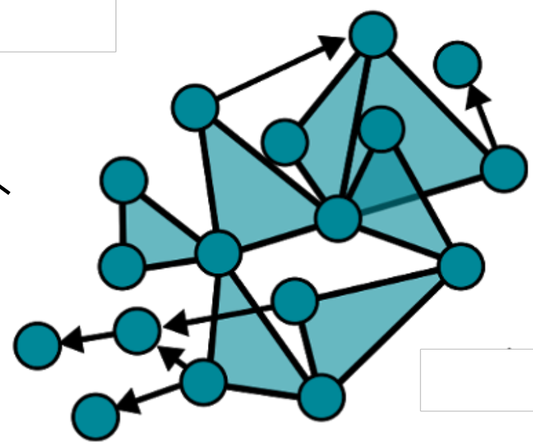
Manifold

lies on



Point Cloud

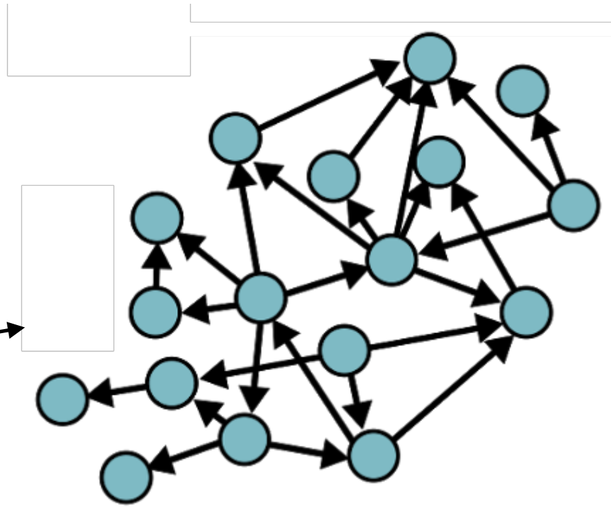
Čech or
Vietoris-Rips
complex



Simplicial Complex

homotopically
equivalent to

has a skeleton

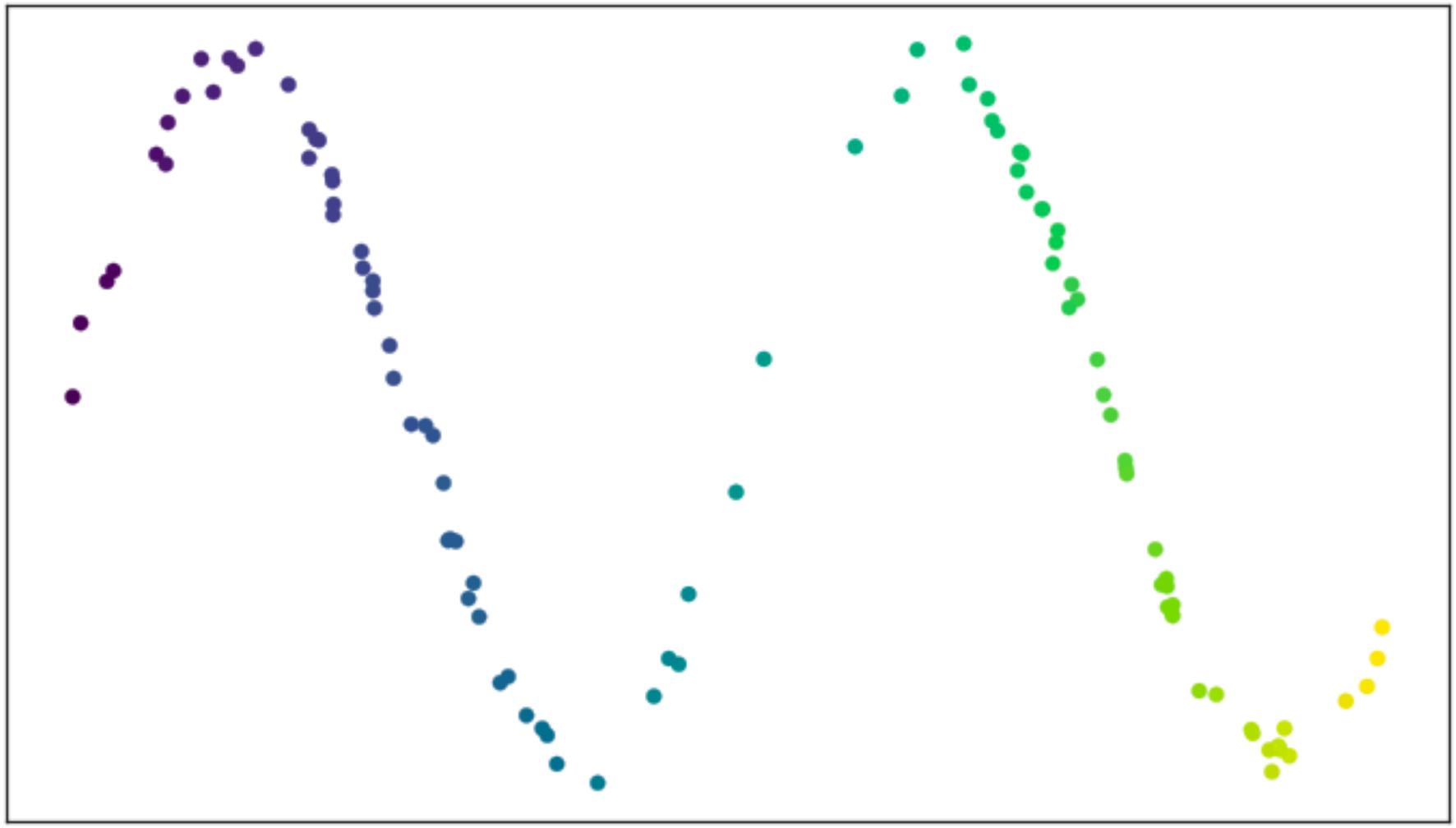


Network

Data Topology (menchelab.com)

UMAP

Step 1. Manifold approximation

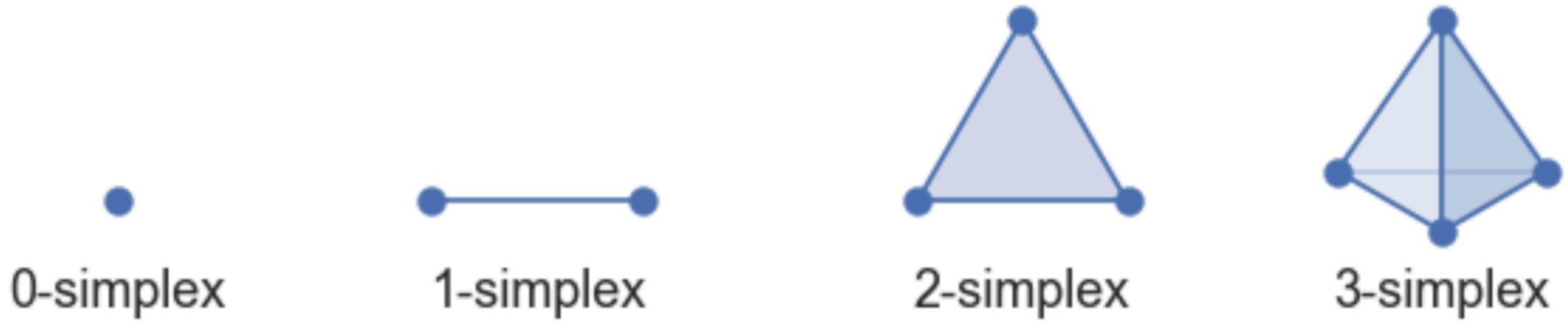


UMAP

Step 1. Manifold approximation

(k-)simplex

k-dim polytope that is the convex hull of its k+1 vertices;
a regular *n*-simplex can be constructed from a regular (*n* − 1)-simplex by connecting a new vertex to all original vertices by the common edge length



UMAP

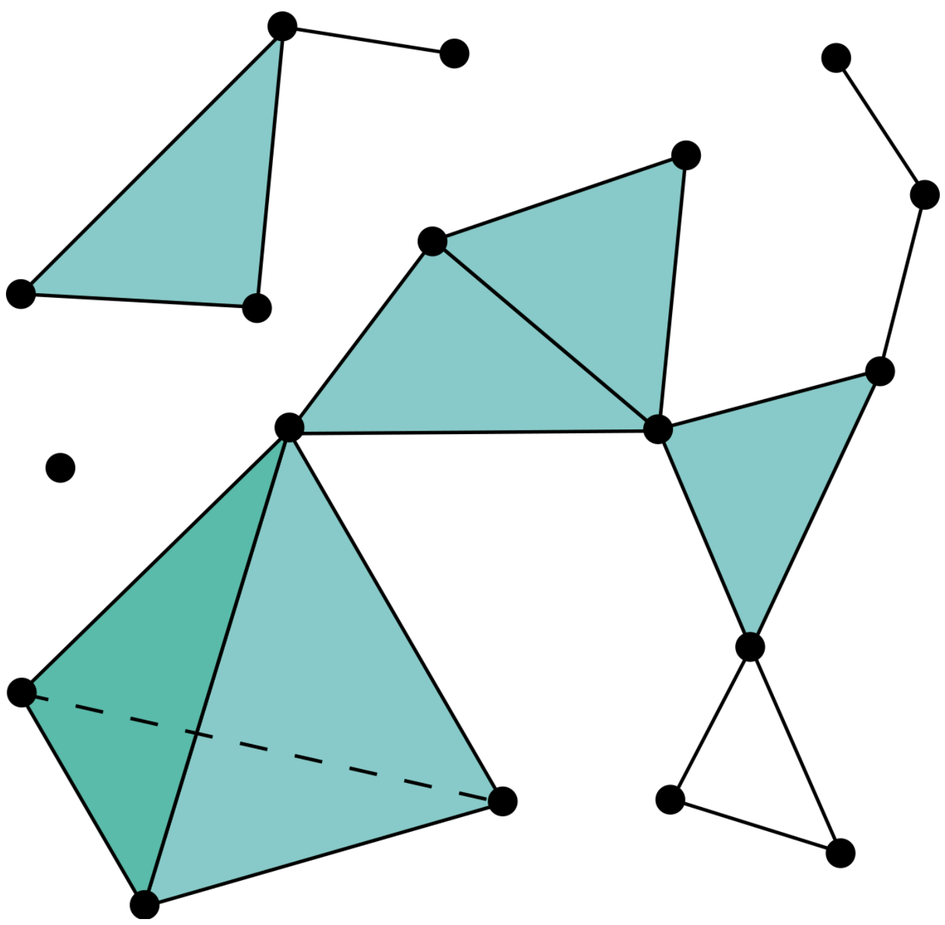
Step 1. Manifold approximation

(k-)simplex

k-dim polytope that is the convex hull of its k+1 vertices; a regular *n*-simplex can be constructed from a regular (*n* − 1)-simplex by connecting a new vertex to all original vertices by the common edge length

simplicial complex

set K of simplices glued together along faces: any face of any simplex in K is also in K and the intersection of two simplices in K is also in K



UMAP

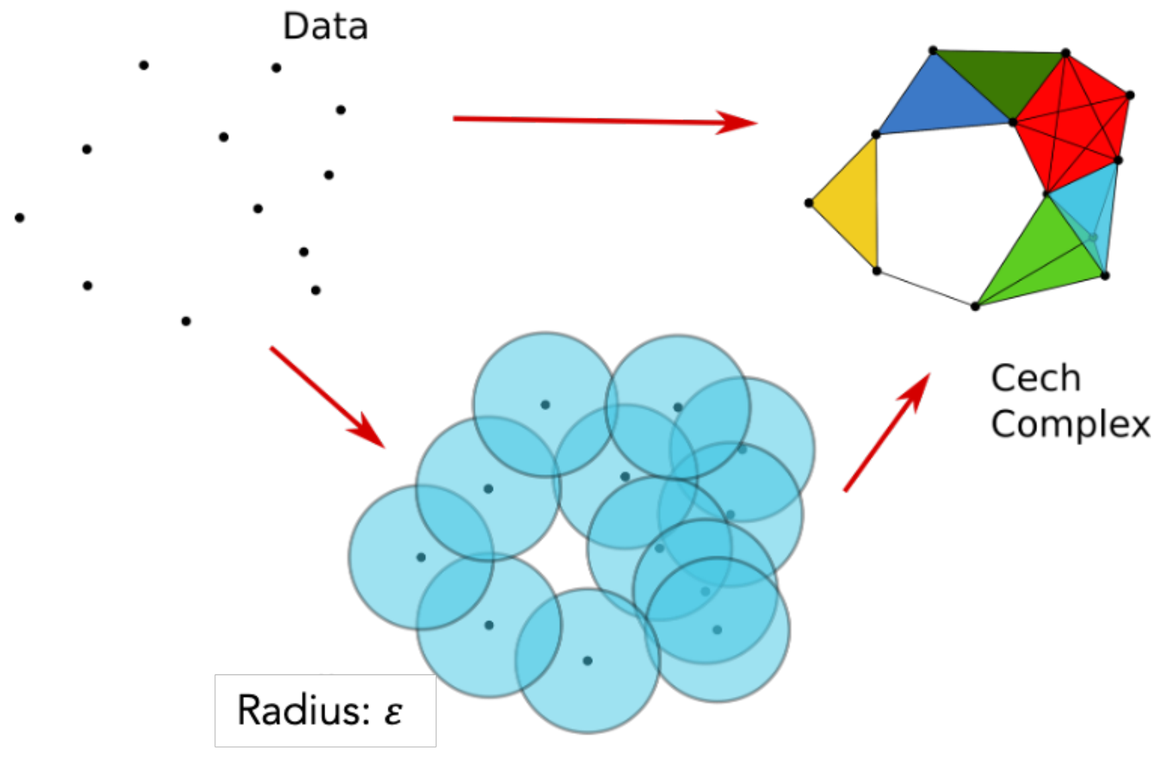
Step 1. Manifold approximation

Čech complex

Given a finite point cloud X and ε > 0, the Čech complex is the nerve of the set of the ε-balls centres at points of X.

Construction

- let each set in the cover be a 0-simplex;
- create a 1-simplex between two such sets if they have a non-empty intersection;
- create a 2-simplex between three such sets if the triple intersection of all three is non-empty;
- …



Does this simple process produce something that represents the original topological space in a meaningful way?

UMAP

Step 1. Manifold approximation

Nerve theorem

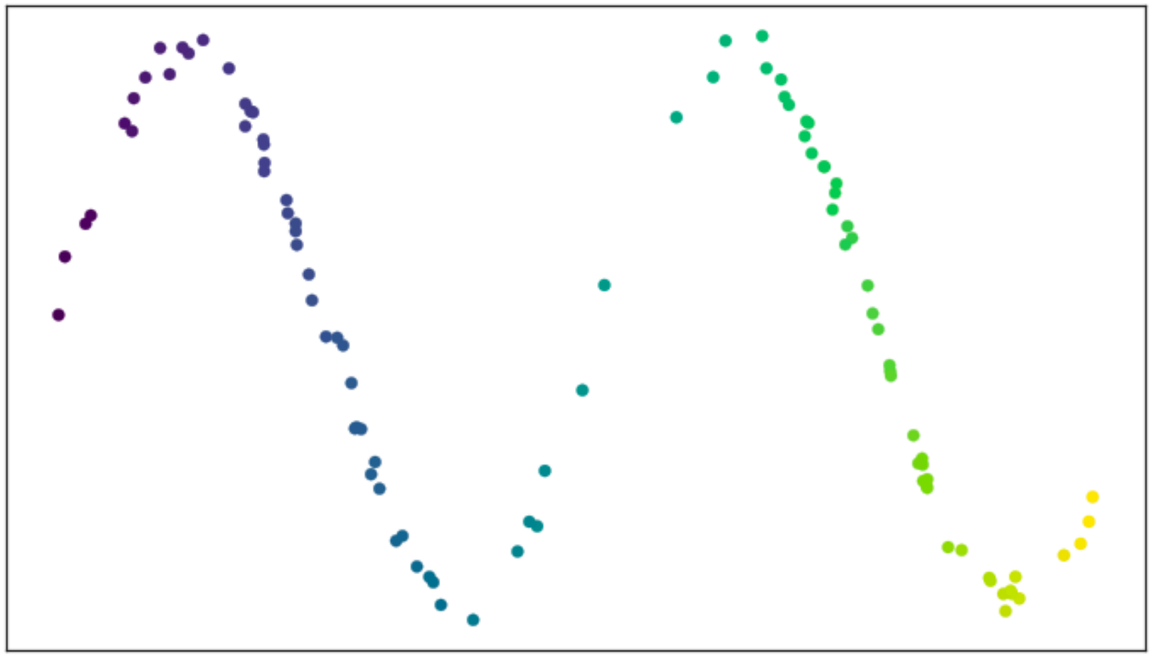
Let $\mathcal{U}=\{U_i\}_{i\in I}$ be an open cover of a topological space X. If, for each $\sigma\subseteq I$, the intersection $\cap_\sigma U$ is either contractible or empty, then $N(\mathcal{U})$ is homotopy equivalent to X

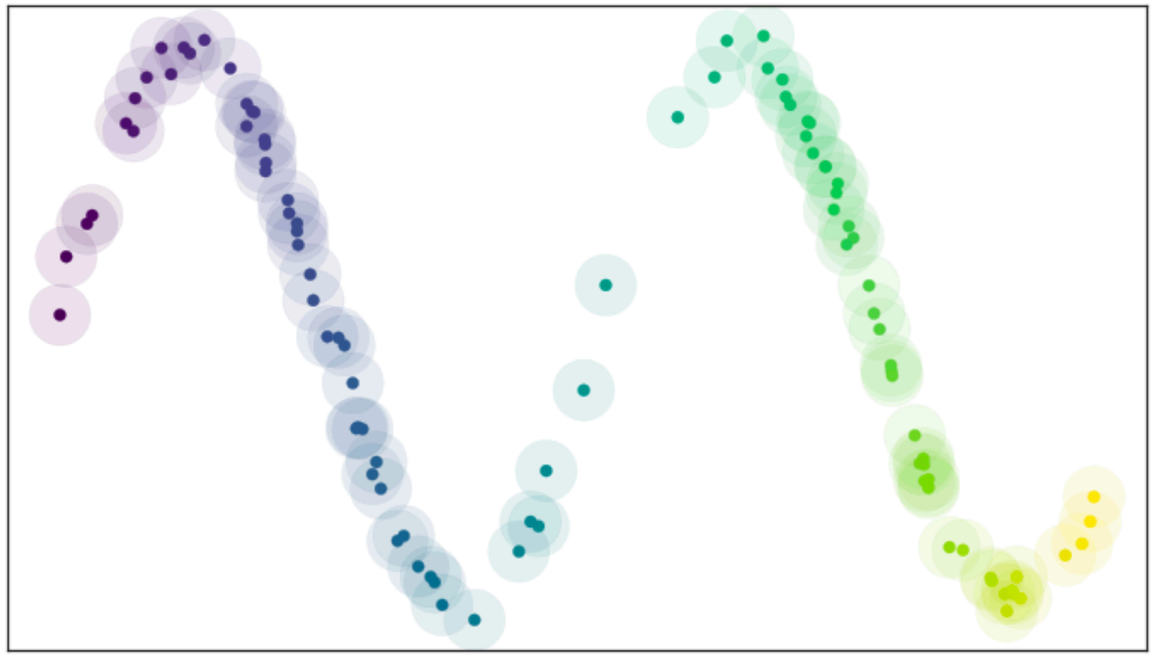
Thus, from the nerve of X we can actually recover all the key topological structures of the original space.

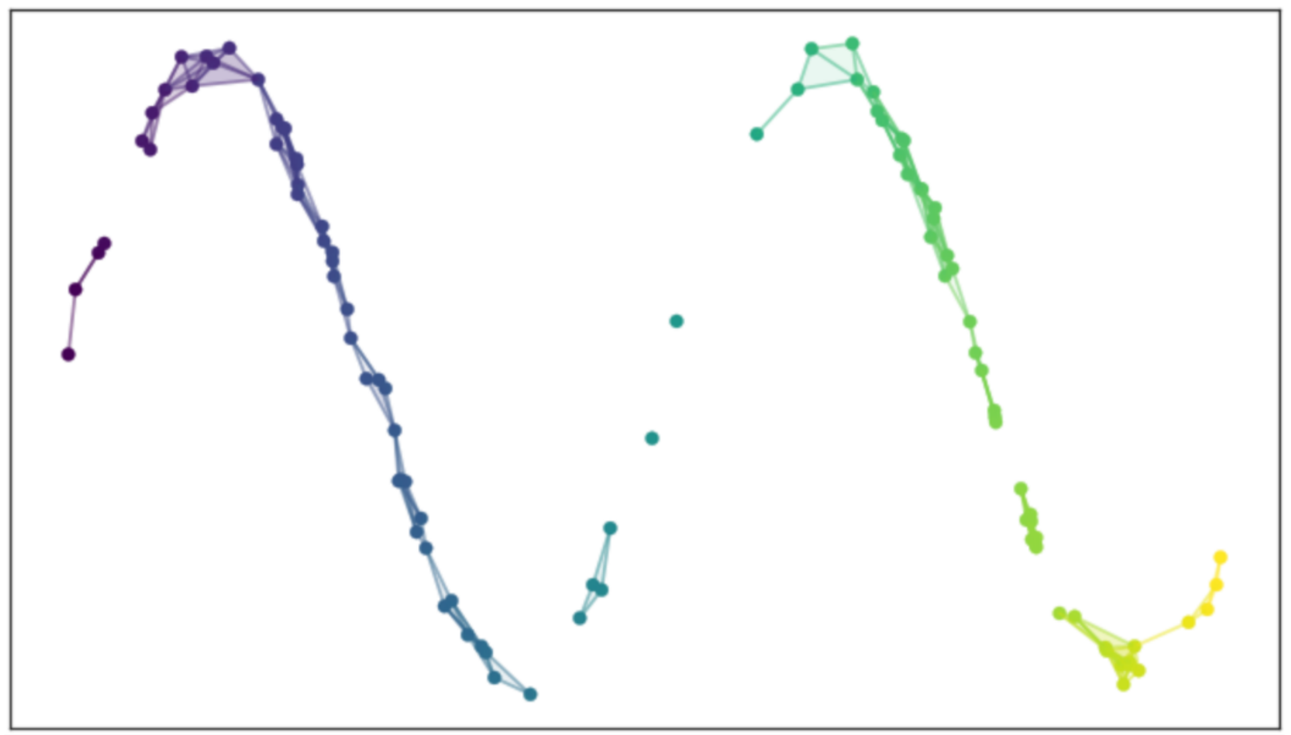
- Computationally, it's easier to work with simplicial complexes than with manifolds
- For computational purposes, we will use the Vietoris Rips complex $VR_\varepsilon(X)$, which is build considering only the 0- and 1- simplifies and can be represented as a graph. It can be shown that $\check{C}_\varepsilon(X)\subseteq VR_\varepsilon(X)\subseteq \check{C}_{2\varepsilon}(X)$

UMAP

Step 1. Manifold approximation







UMAP

Step 1. Manifold approximation

Čech complex

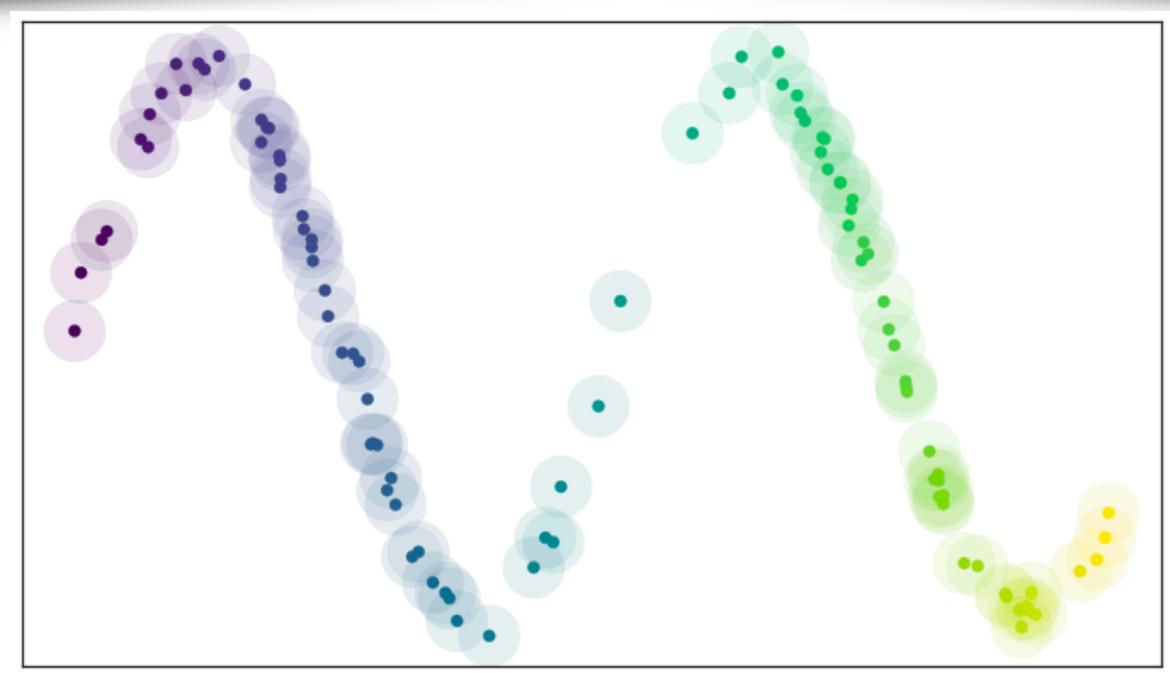
Given a finite point cloud X and ε > 0, the Čech complex is the nerve of the set of the ε-balls centres at points of X.

Nerve theorem

Let $\mathcal{U}=\{U_i\}_{i\in I}$ be an open cover of a topological space X. If, for each $\sigma\subseteq I$, the intersection $\cap_\sigma U$ is either contractible or empty, then $N(\mathcal{U})$ is homotopy equivalent to X

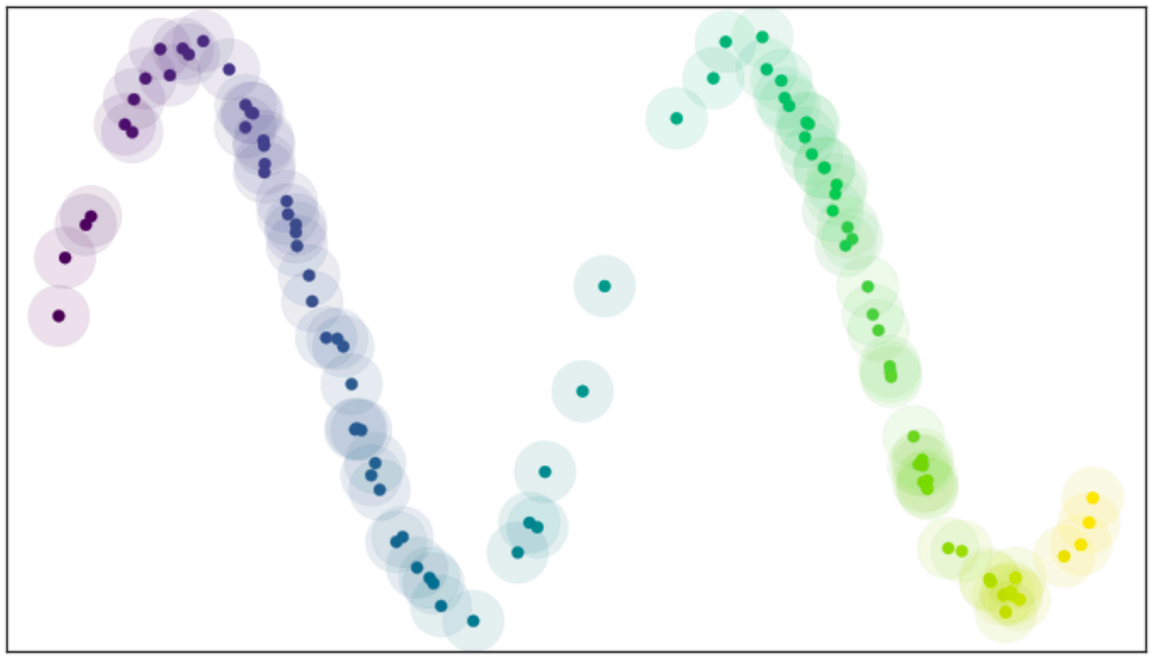
cover

a collection of subsets of X whose union is X



UMAP

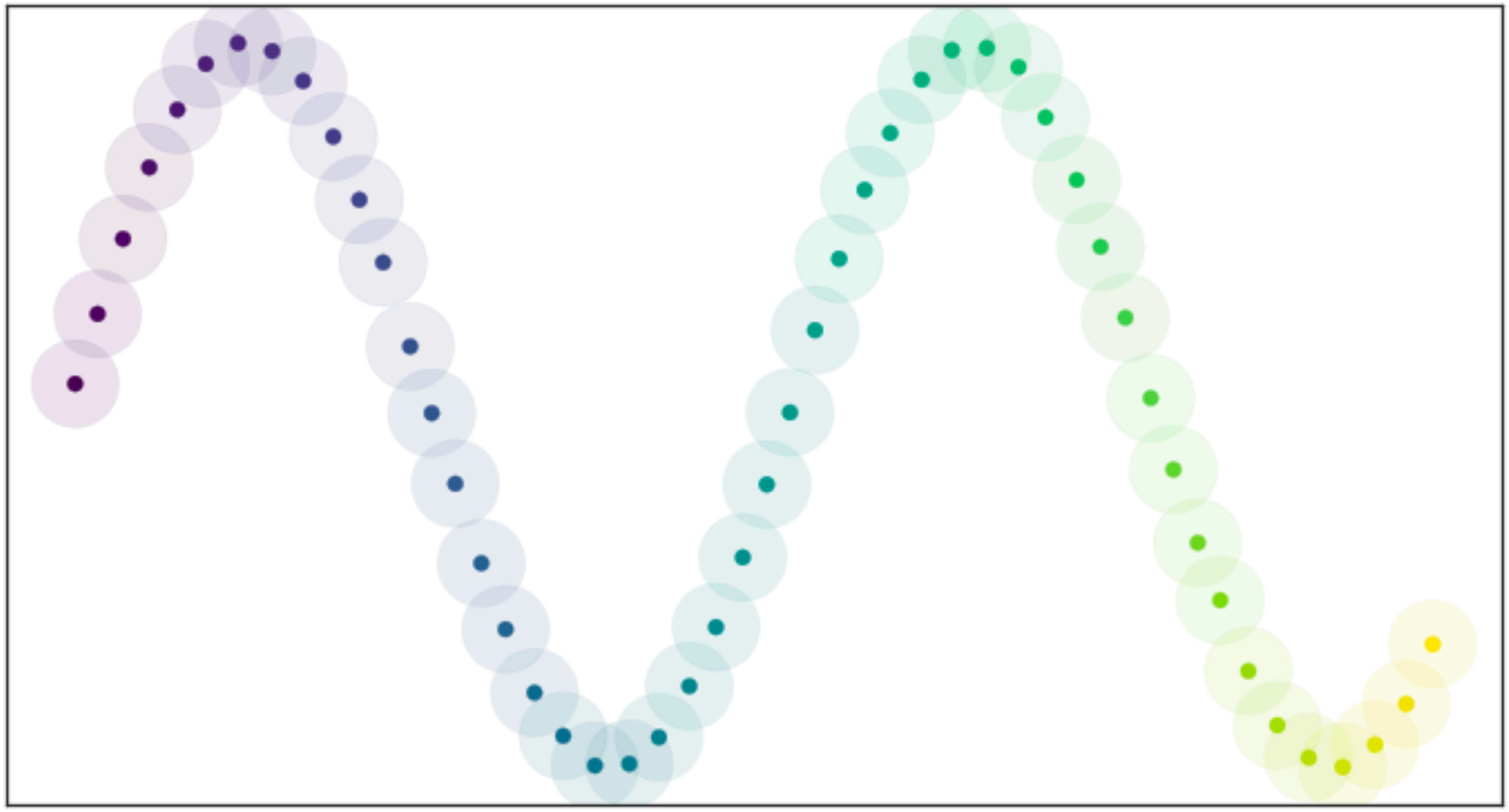
Step 1. Manifold approximation



Issue: how to find the open cover?
too small: too many connected components
too large: few very high dimensional simplices, and no structure

UMAP

Assumption 1. Uniform distribution



having data points uniformly distributed across the manifold

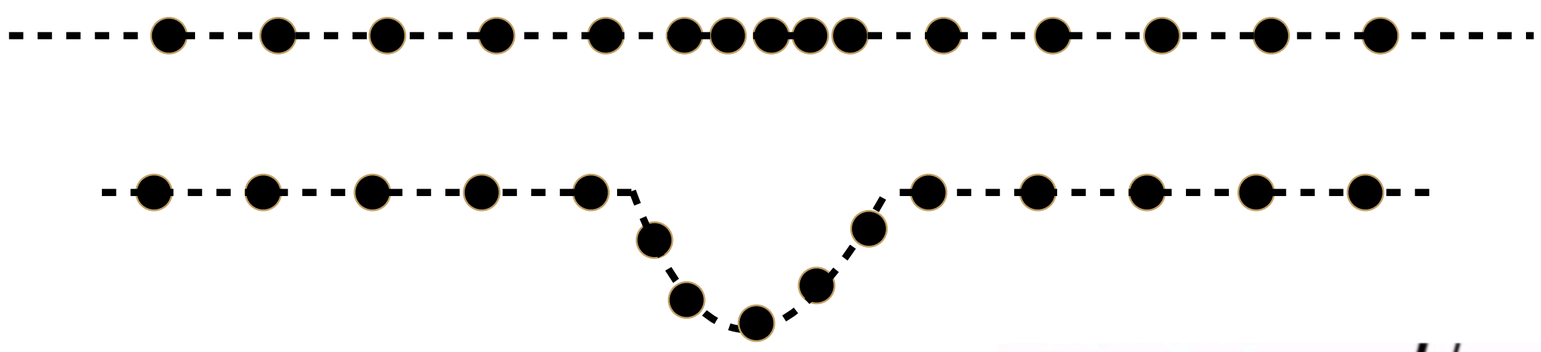
- radius ≅ half the average distance between points
- no gaps & no clumps in the cover

… but this does not happen with real world data!

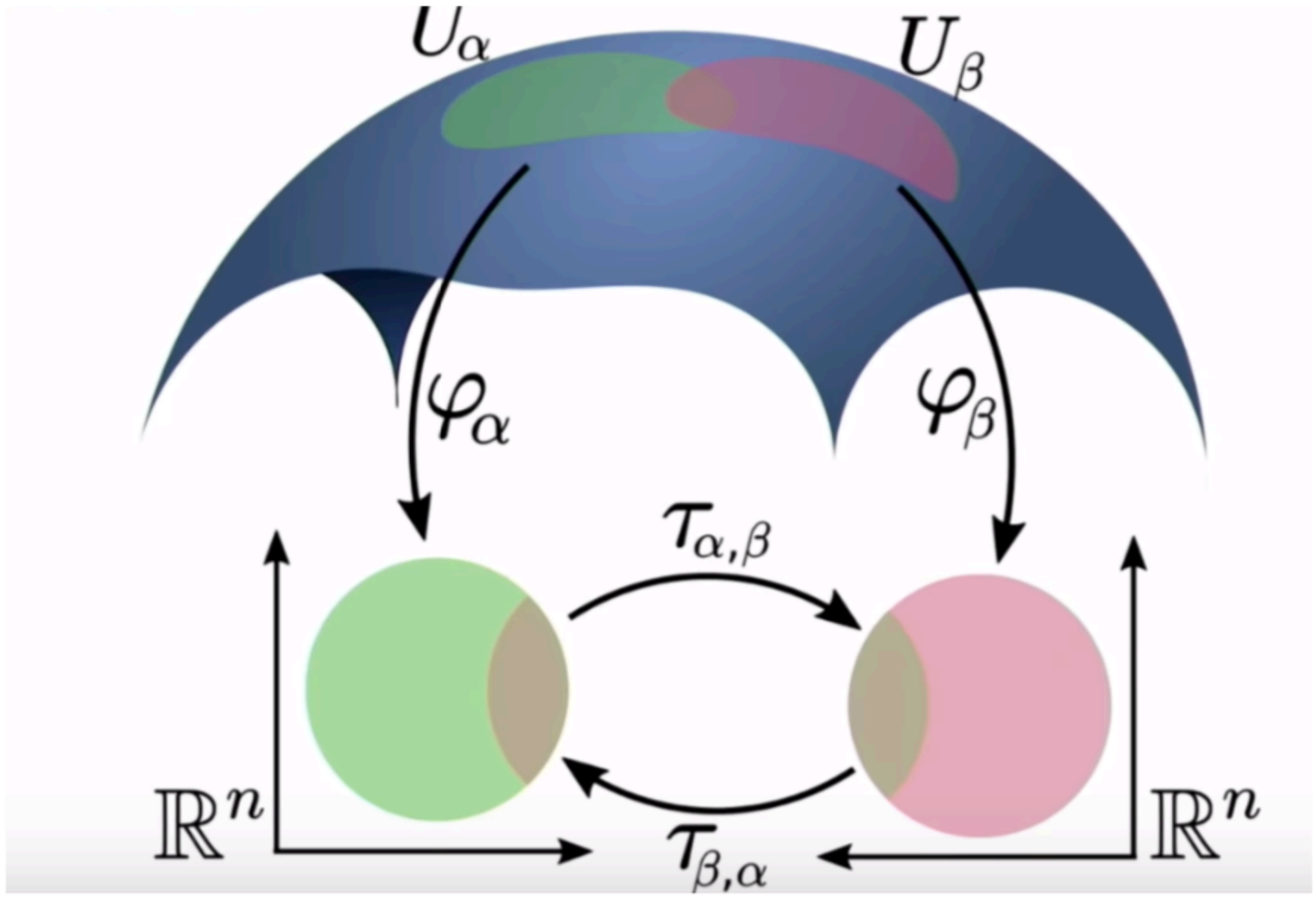
UMAP

Assumption 1. Uniform distribution

adapt the notion of distance on the manifold (stretch) so that all the points seem to be uniformly distributed



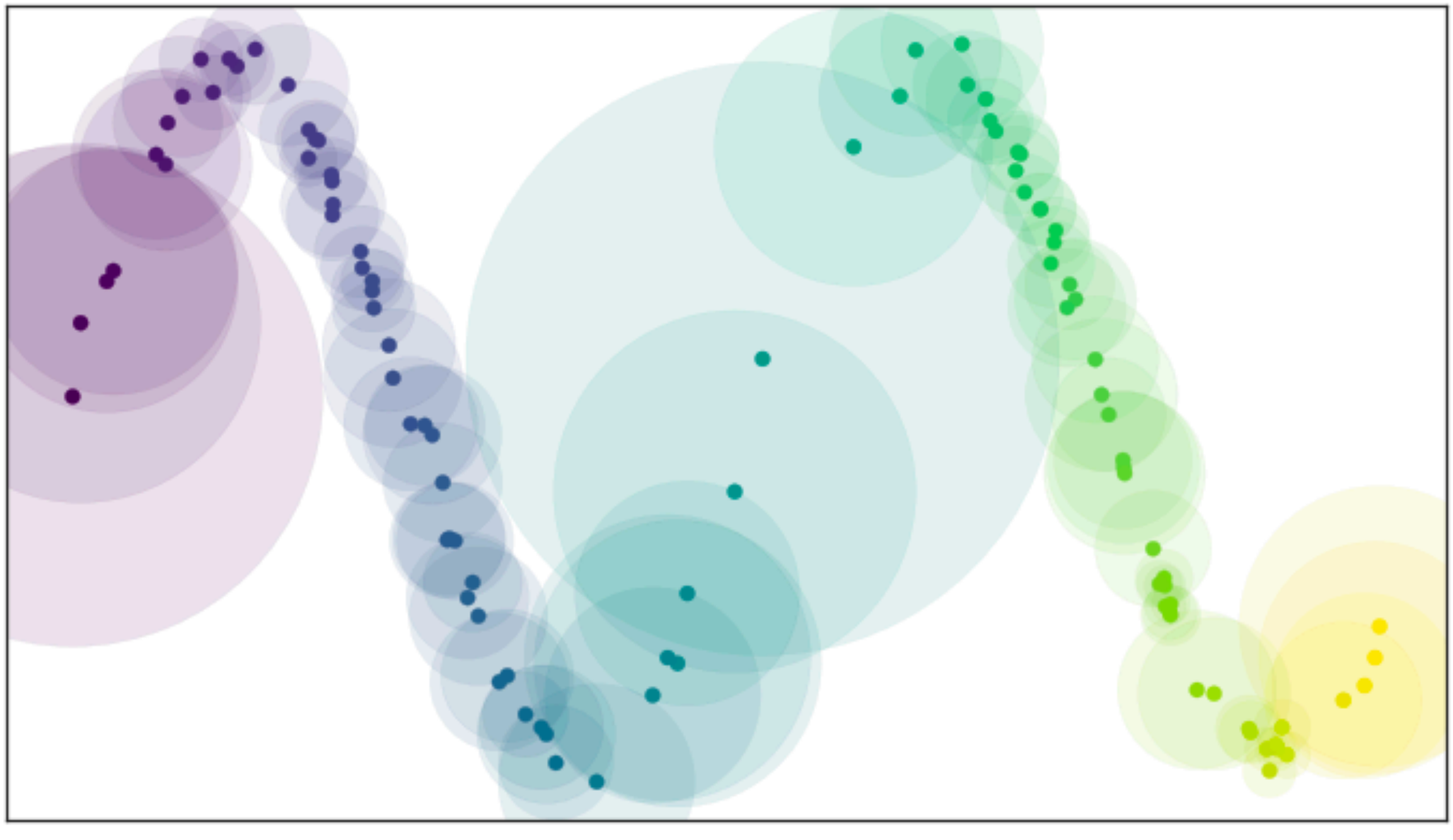
thus give to each point its own unique distance function, and select balls of radius one with respect to that local distance function



UMAP

Assumption 1. Uniform distribution

The unit ball about a point stretches to the k-th nearest neighbor of the point, where k is the sample size we are using to approximate the local sense of distance
(easier in terms of parameters)

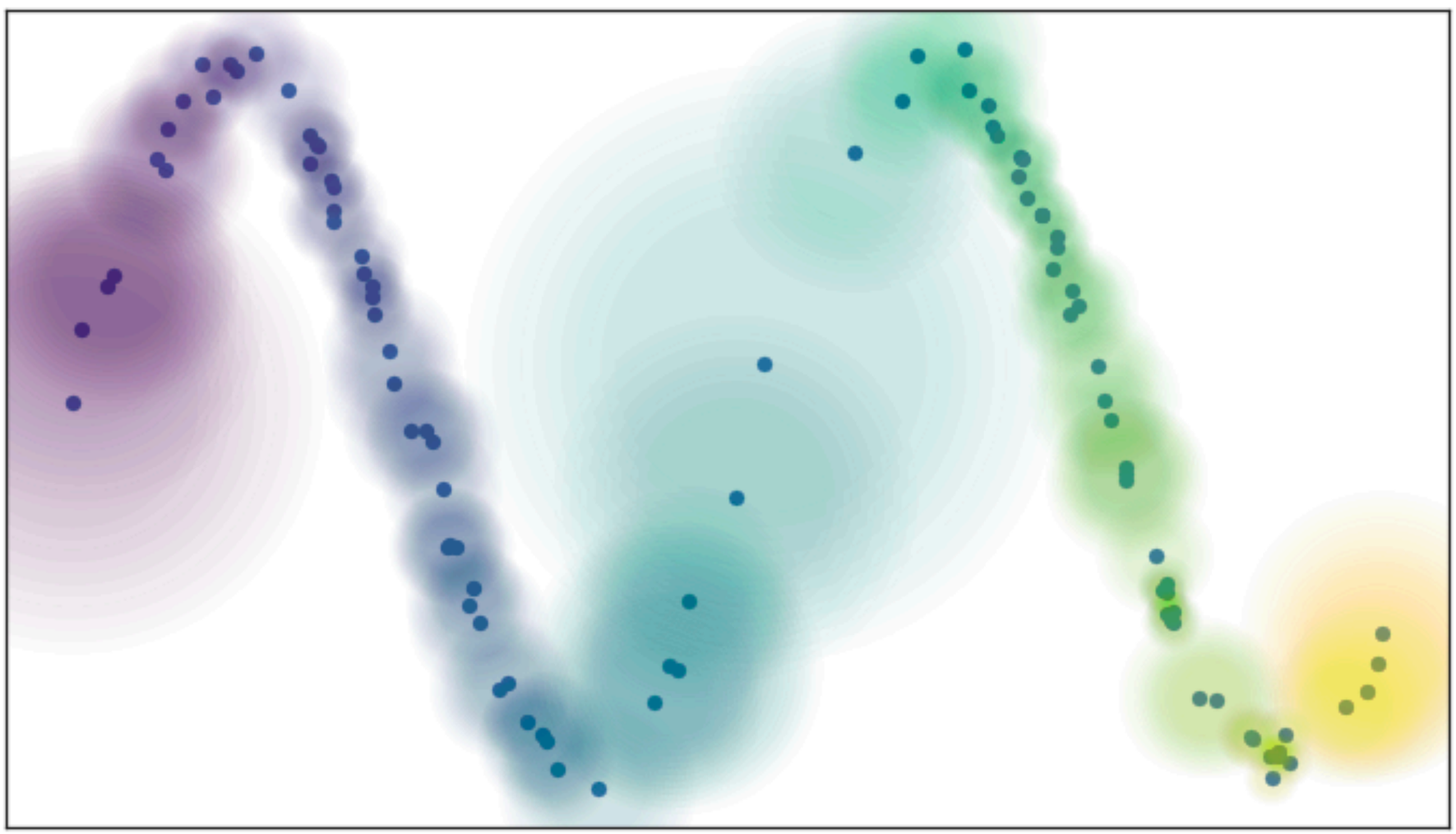


UMAP

Step 2. Fuzziness

given the local metric, we can weight the edges of the graph according to the edge' vertices distance

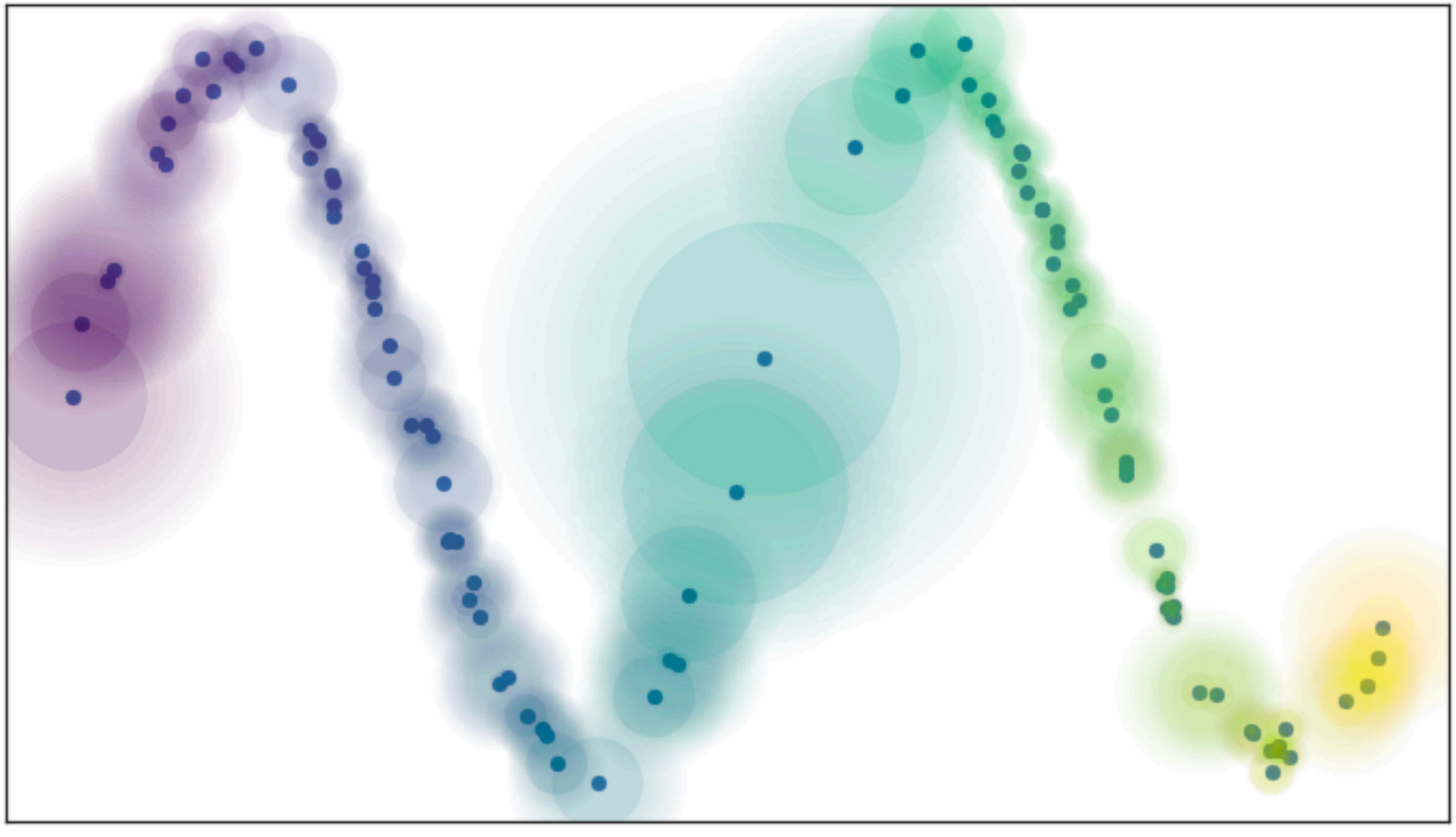
mathematically, we move from the simplicial complexes to the simplicial sets (category theory), which corresponds to move from classical balls to fuzzy balls, where ϵ becomes a function in [0,1] decreasing further away from the center $(1+ax^{2b})^{-1}$



UMAP

Assumption 2. Local connectivity

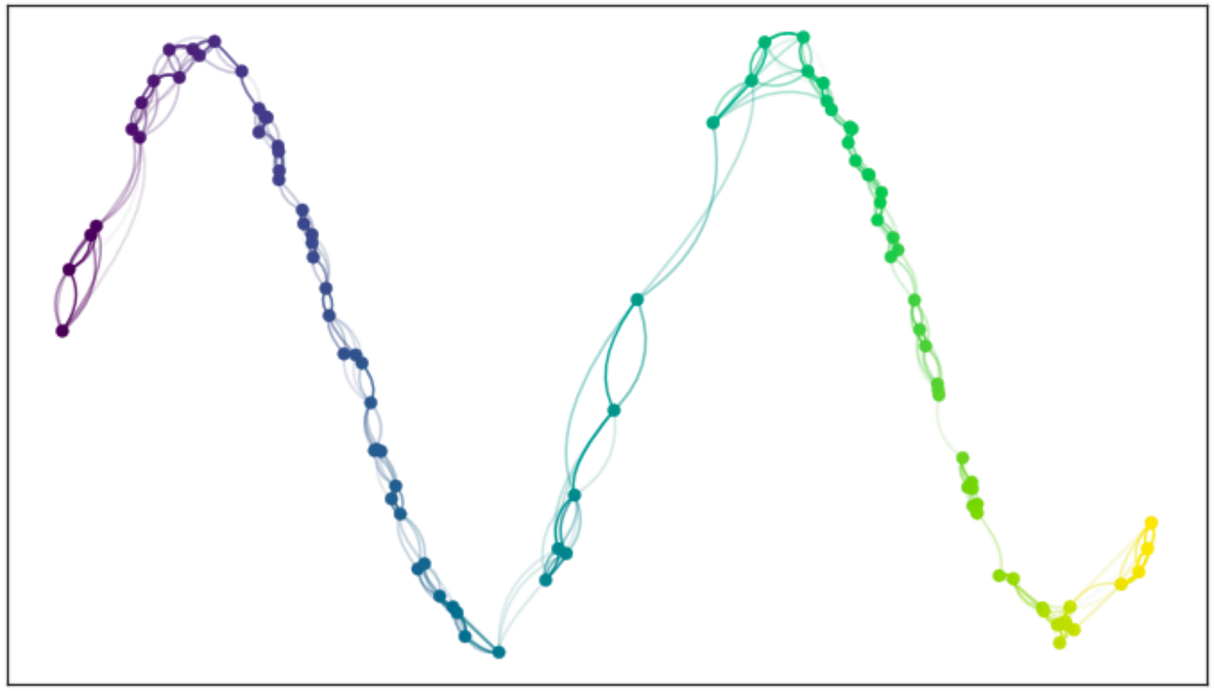
- points may end up being separated by the rest of the manifold
- local connectivity is introduced: the fuzziness decays only beyond the nearest neighbour
- the focus is on the difference in distances among nearest neighbors rather than the absolute distance, avoid the curse of dimensionality

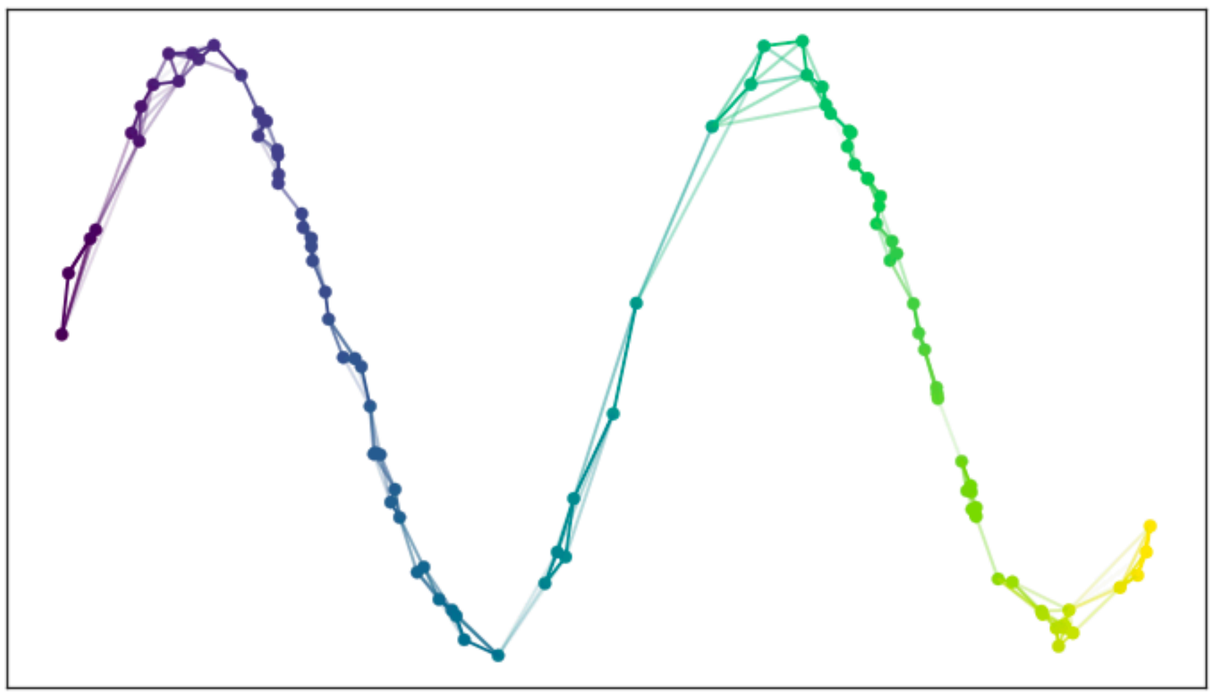


UMAP

Issue. Incompatibility

since each point has its own metric,
distance from a to b may be different
than distance from b to a





the theoretically grounded solution is correctly defining the fuzzy union of simplicial sets as the probability that at least one of the edge exists, thus ending with a single fuzzy simplicial complex (weighted graph)

UMAP

Step 3. Move to low dimensional space

we need now to faithfully embed the graph into a low-dim euclidean space so to preserve the original manifold structure

this means deciding which $f(w_h,w_l)$ to optimize, where w_h,w_l are the graph edges' weights in high and low dimension

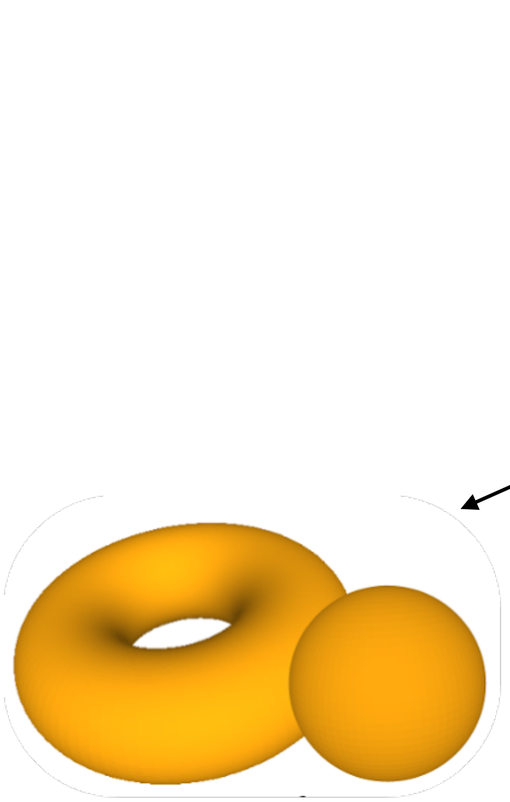
since w_h,w_l are bernoulli variables (w exists with prob. p and does not exist with prob. 1-p), the correct function is the cross-entropy

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)$$

attractive force between points when w is large in high-dimension
optimizing clumps

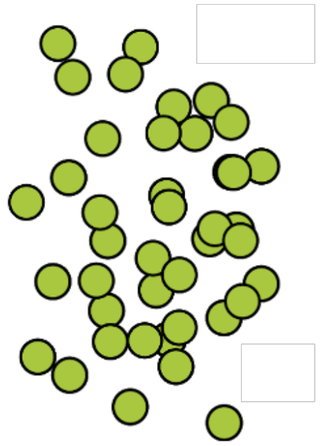
repulsive force between points when w is small in high-dimension:
optimizing gaps

UMAP - intuition



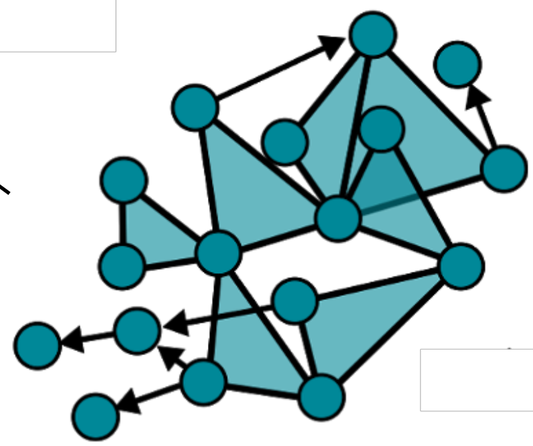
Manifold

lies on



Point Cloud

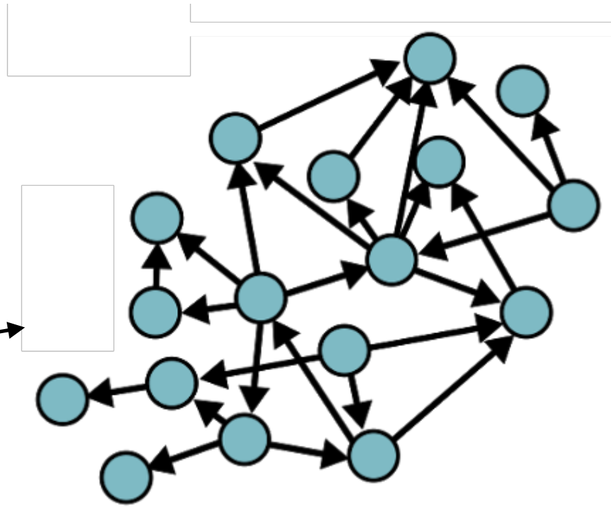
Čech or
Vietoris-Rips
complex



Simplicial Complex

homotopically
equivalent to

has a skeleton

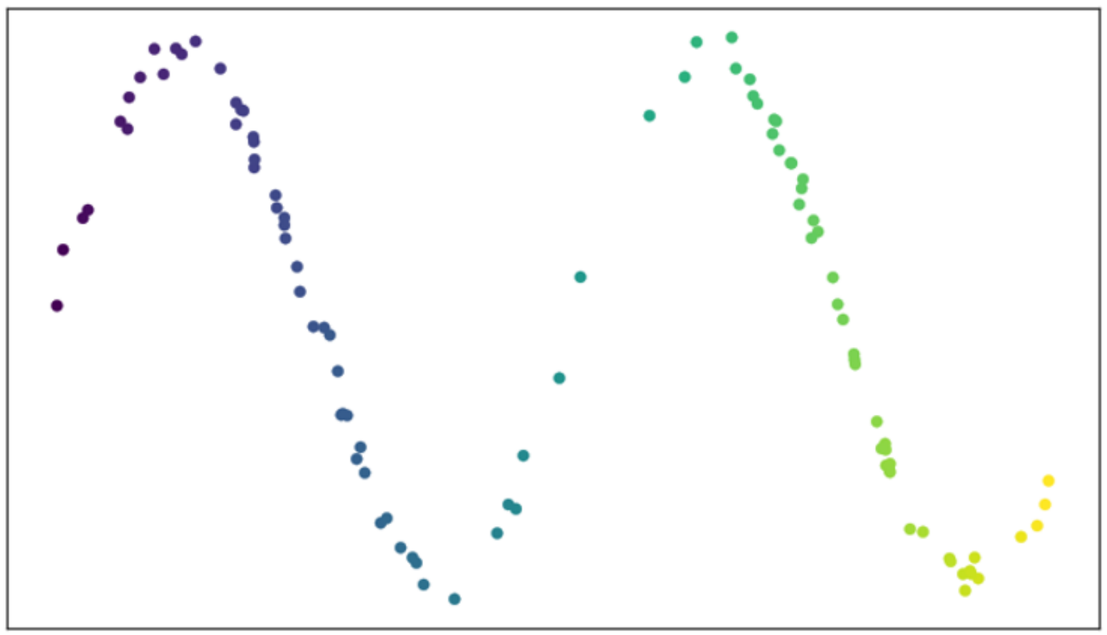


Network

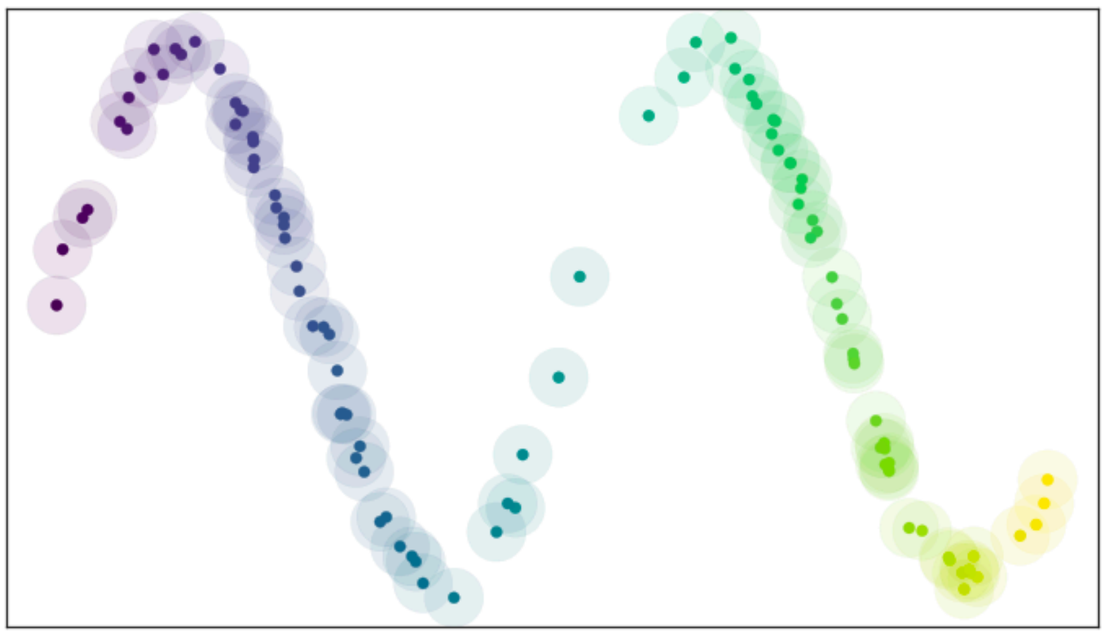
Data Topology (menchelab.com)

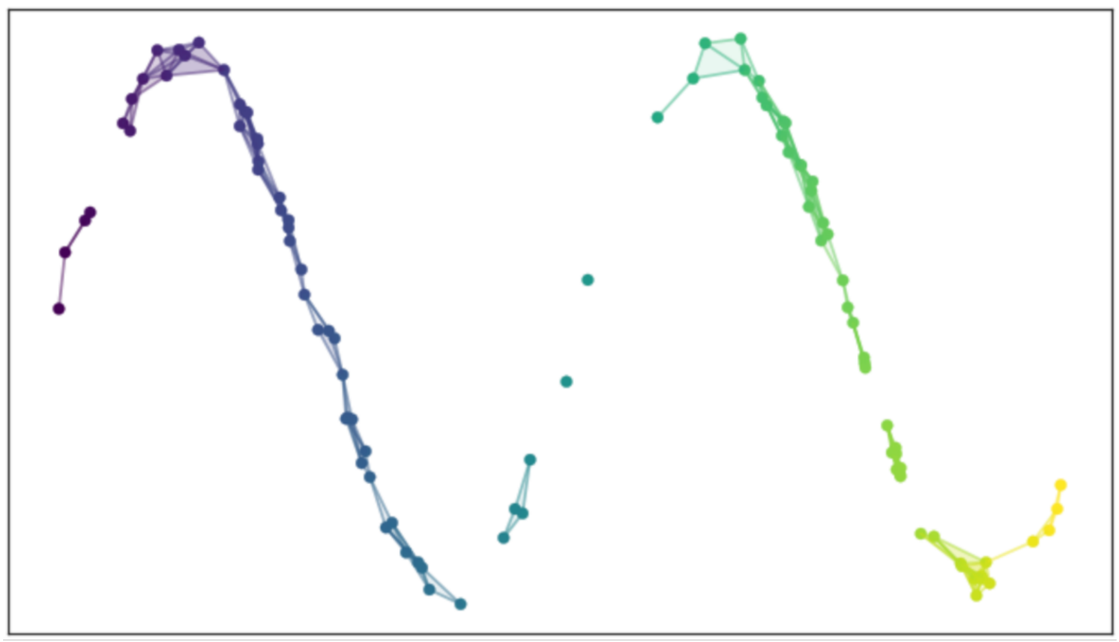
UMAP - recap

1. Finite points cloud in high dimensional space (on a low dimensional manifold by assumption)



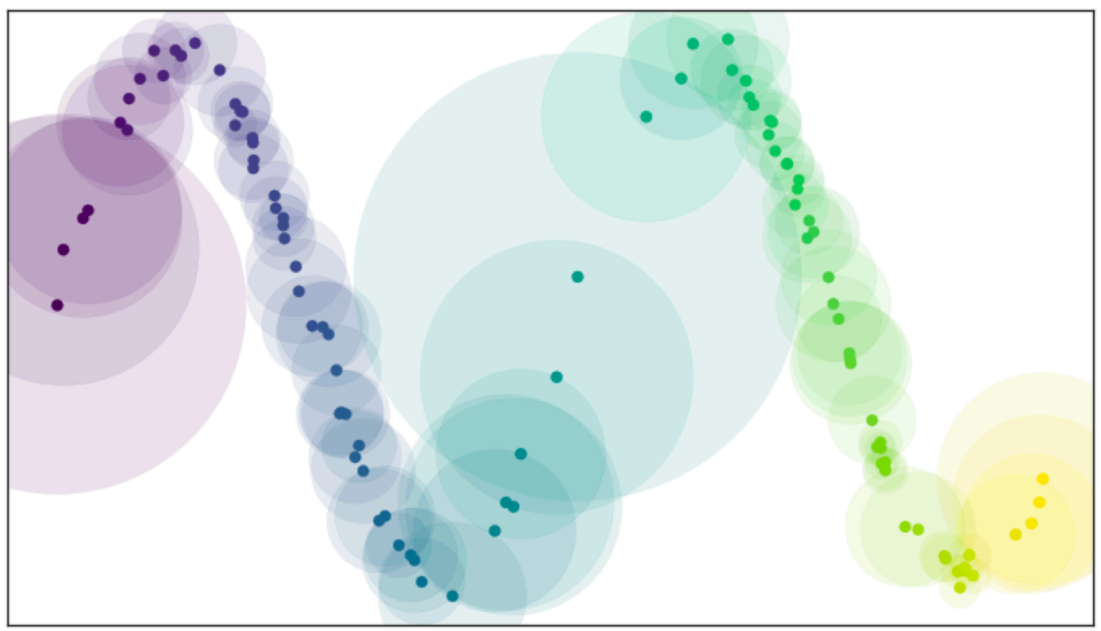
2. Consider the Čech complex of balls centred in the points of the cloud. [By the Nerve theorem it is homotopically equivalent to the original topological space, if we have a cover]



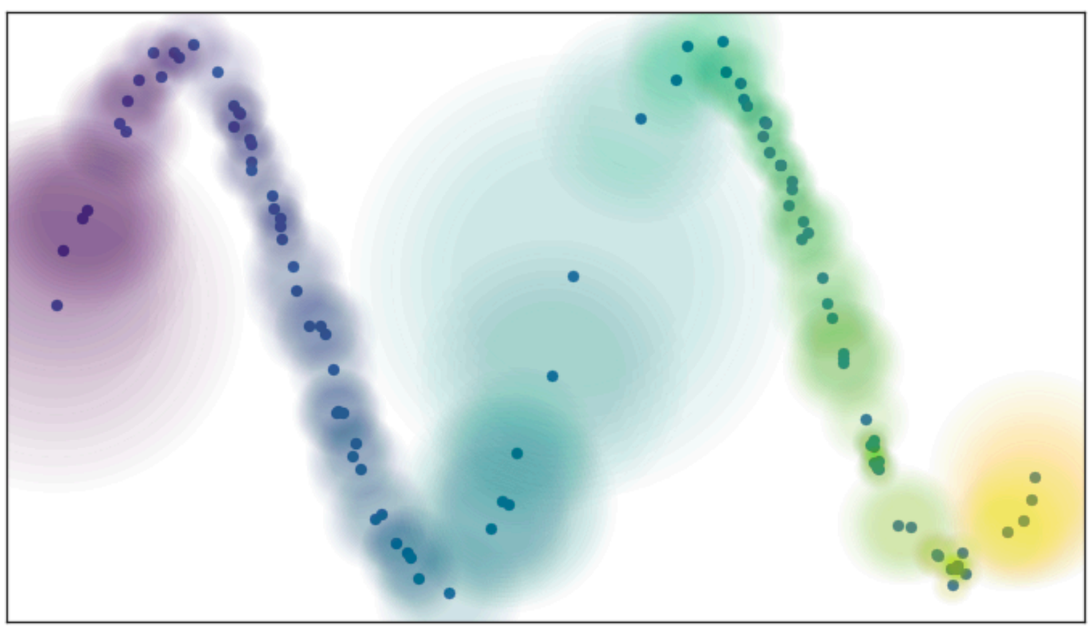


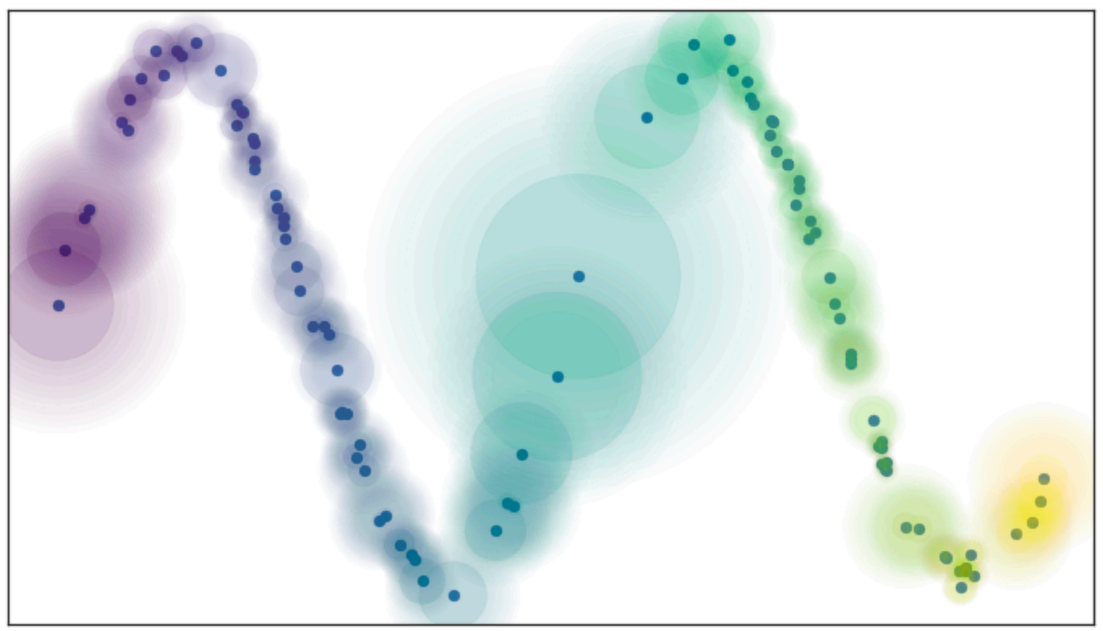
UMAP - recap

3. Use local metrics (uniformly distributed points) to build a cover. Each unit ball stretches to the k-th nearest neighbours.



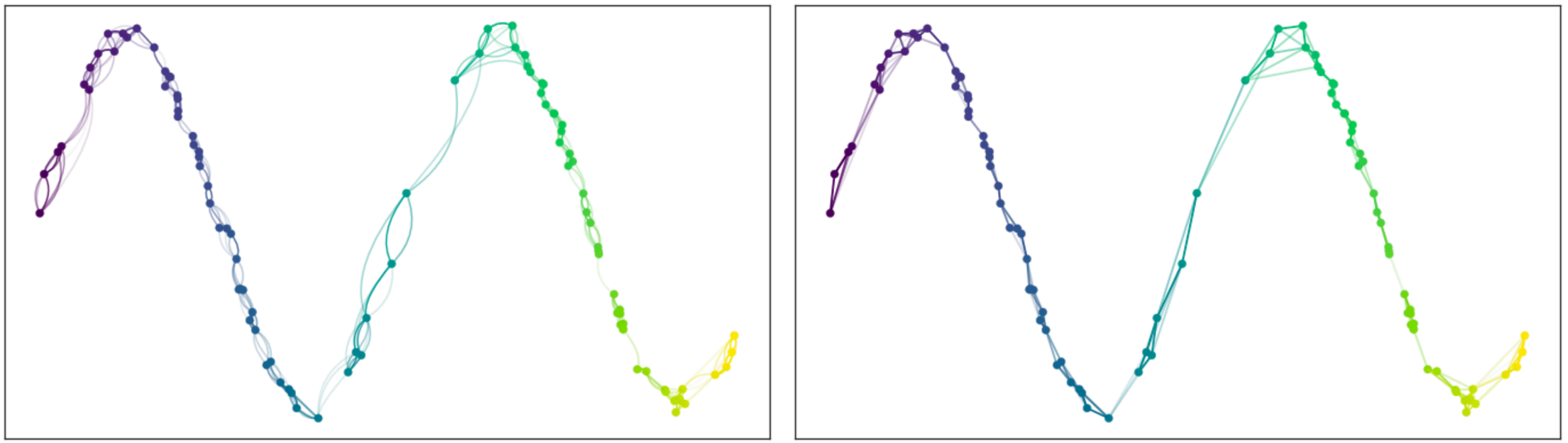
4. Consider fuzzy balls (value one to the 1st nearest neighbours to ensure connectivity, then decaying)





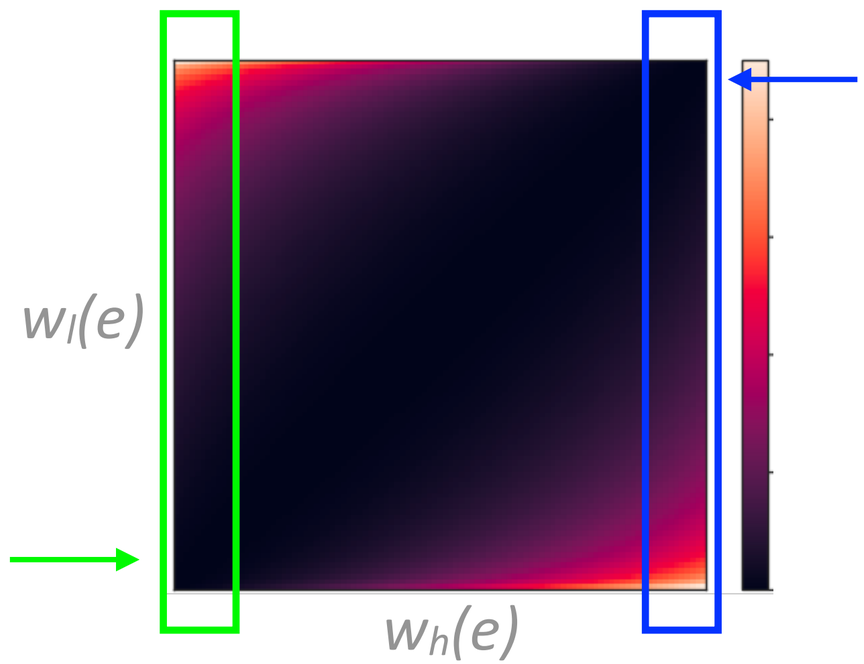
UMAP - recap

5. Build asymmetric weighted graph and find a way to make it symmetric.

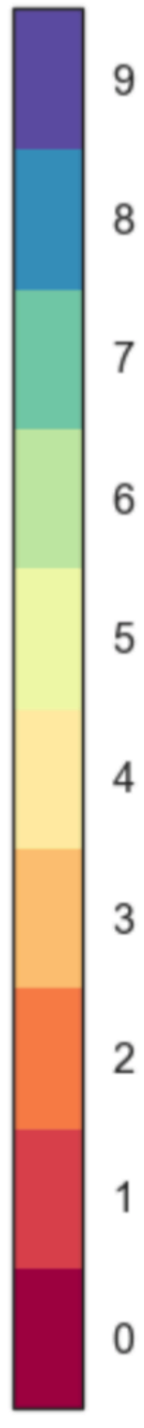


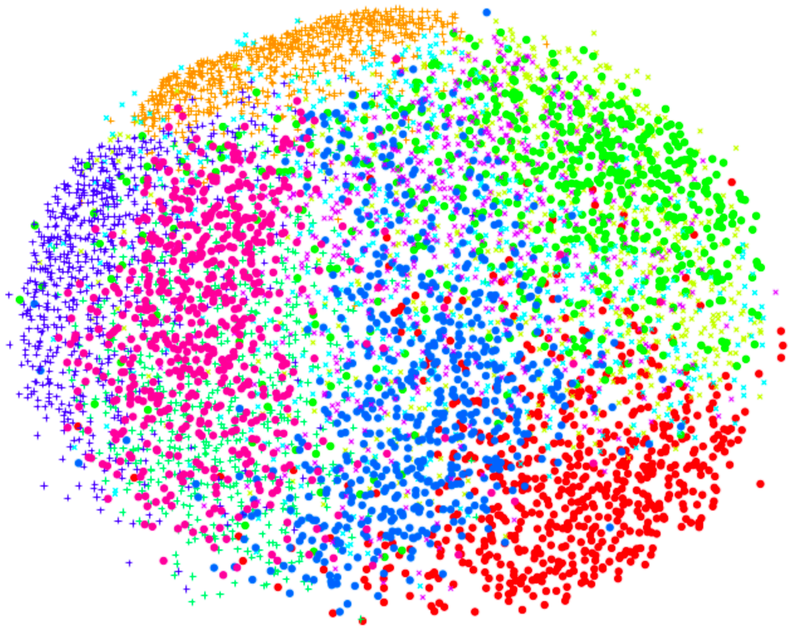
6. Project to low-dimensional space

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)$$

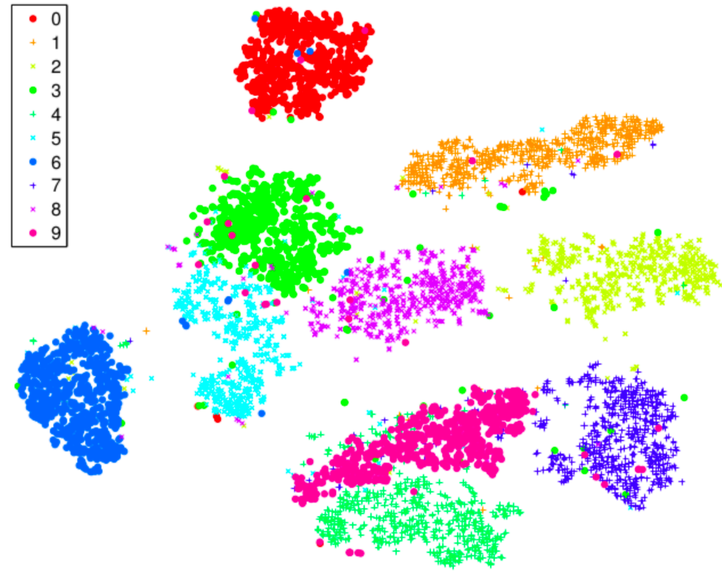


mnist





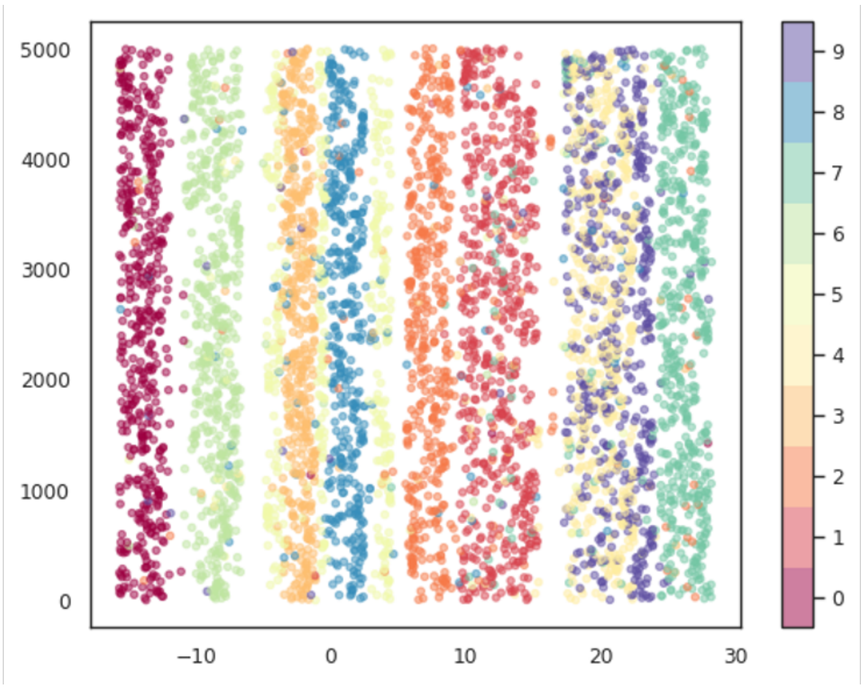
sammon



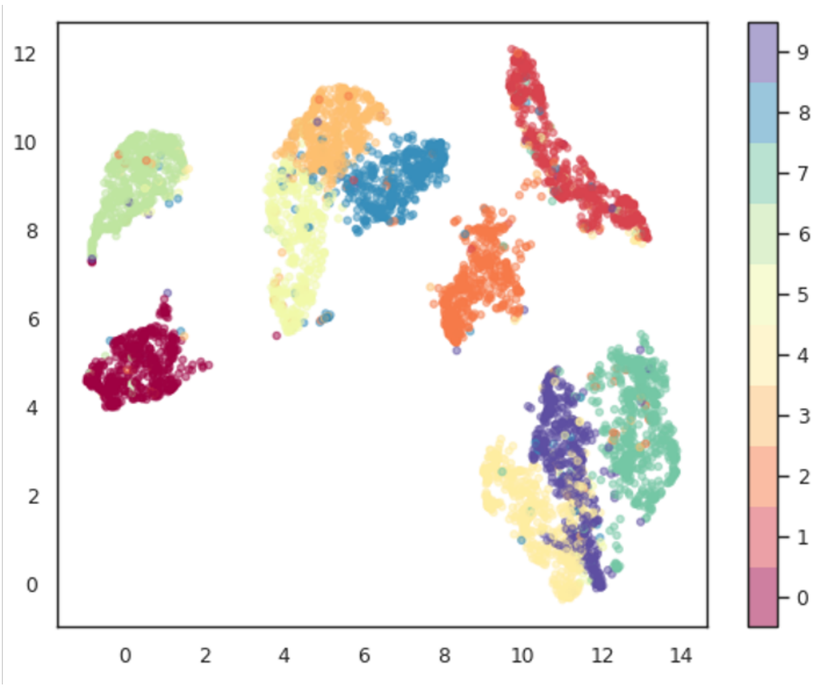
tsne

UMAP-Parameters

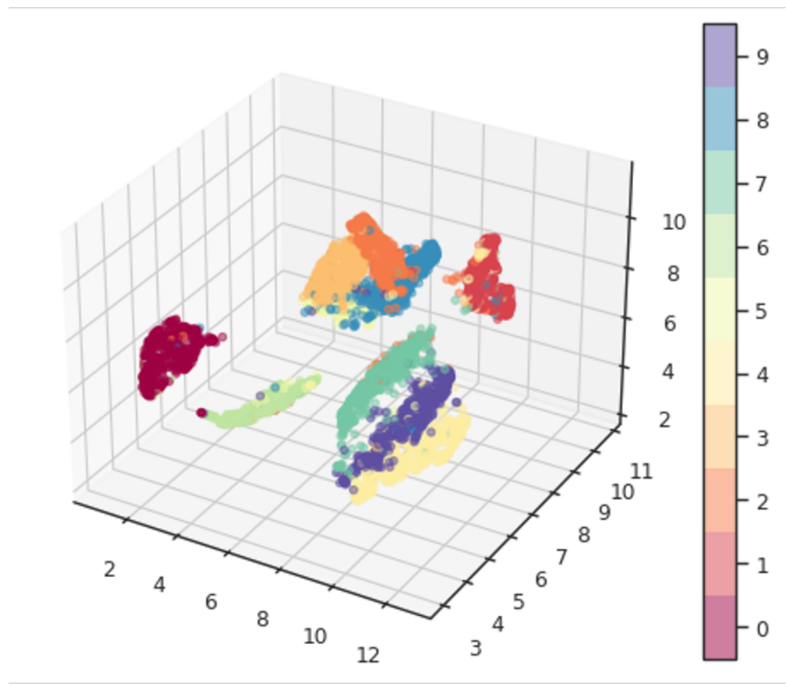
- d: dimensionality of the low-dimensional space (usually d=2 or d=3)



d=1



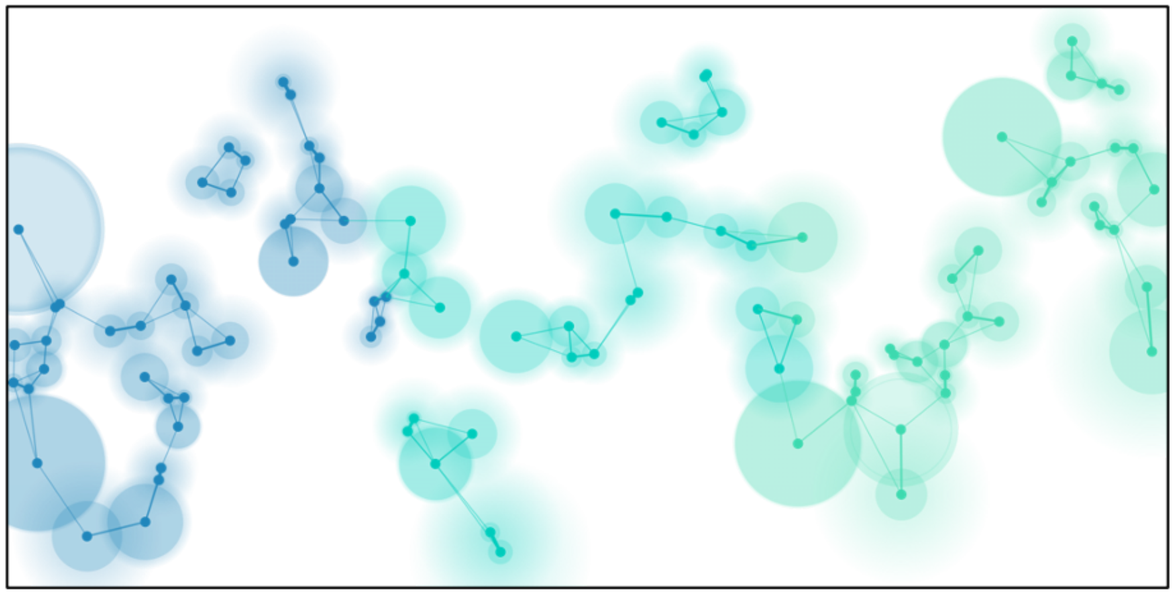
d=2



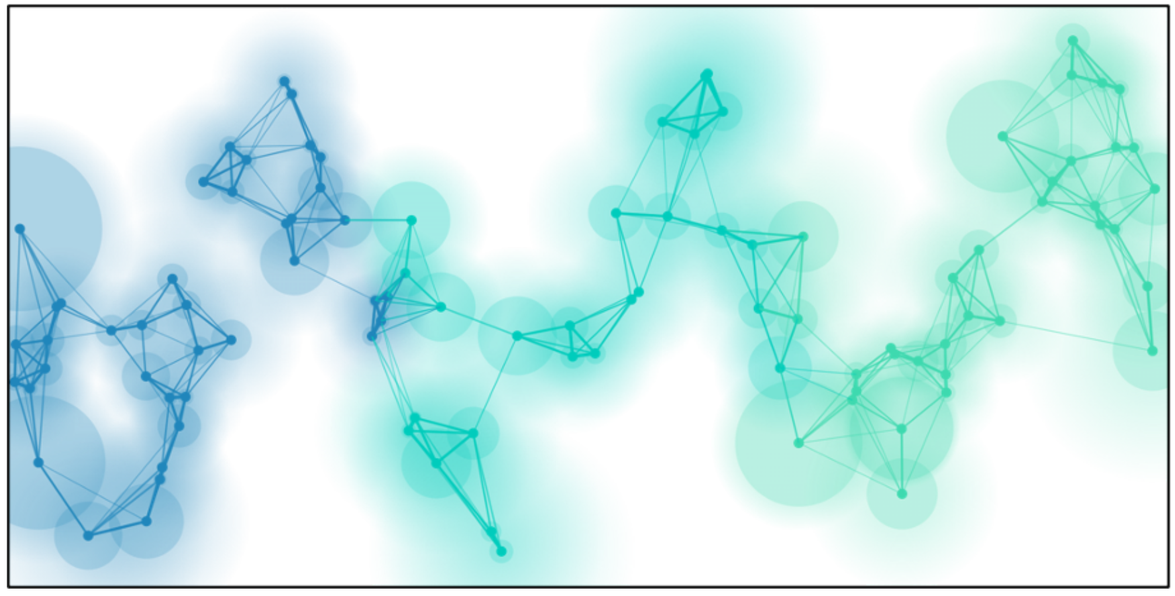
d=3

UMAP-Parameters

- d: dimensionality of the low-dimensional space (usually d=2 or d=3)
- n_neighbours: number of neighbours to construct the cover and initial graph



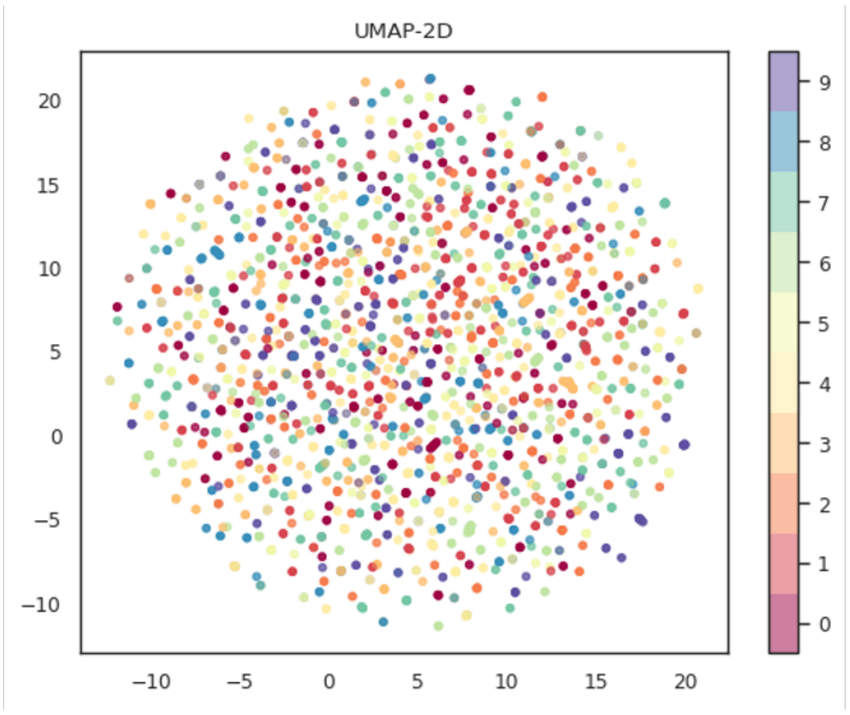
n_neighbours=2



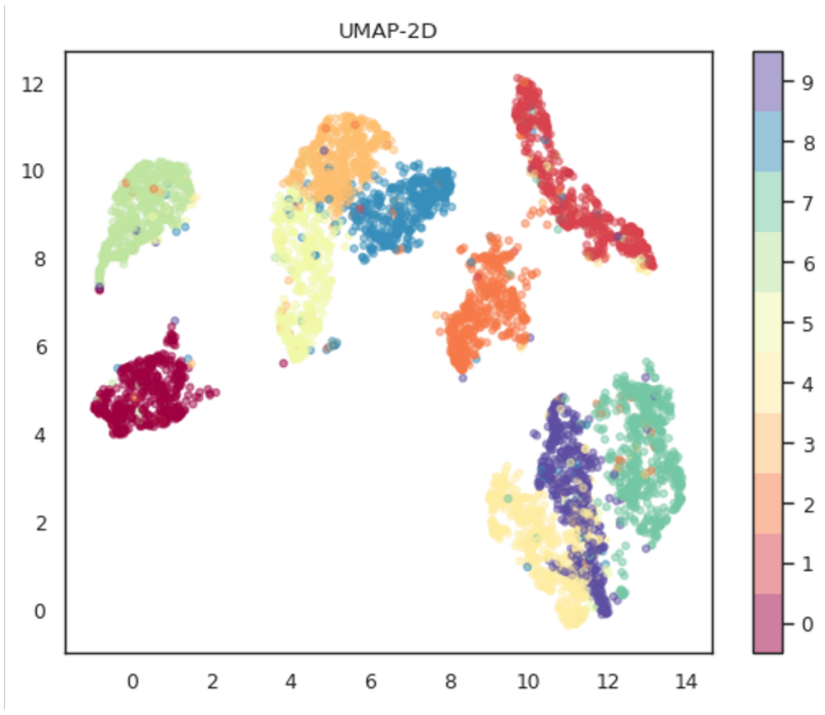
n_neighbours=5

UMAP-Parameters

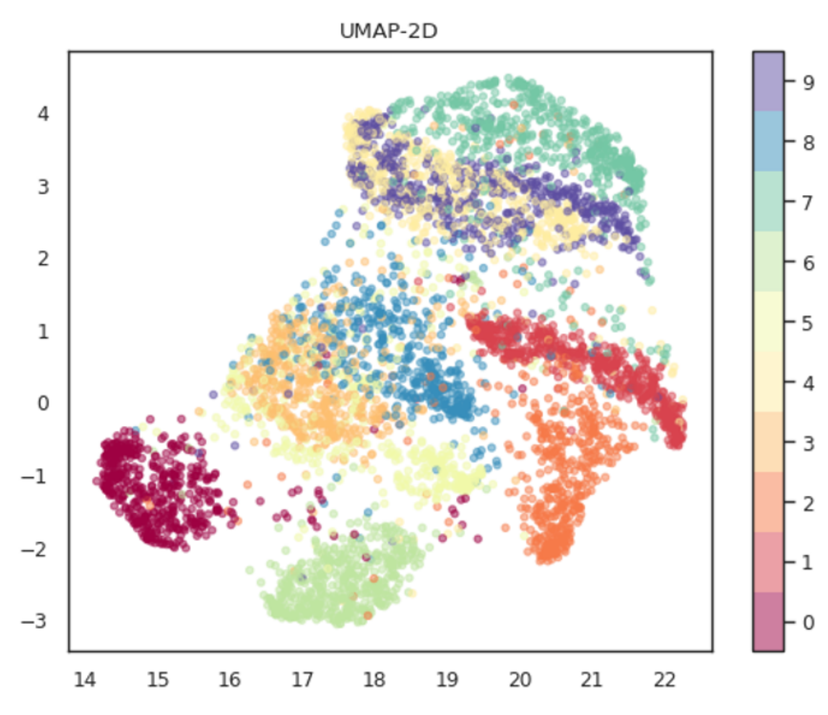
- d: dimensionality of the low-dimensional space (usually d=2 or d=3)
- n_neighbours: number of neighbours to construct the cover and initial graph



n_neighbours=2



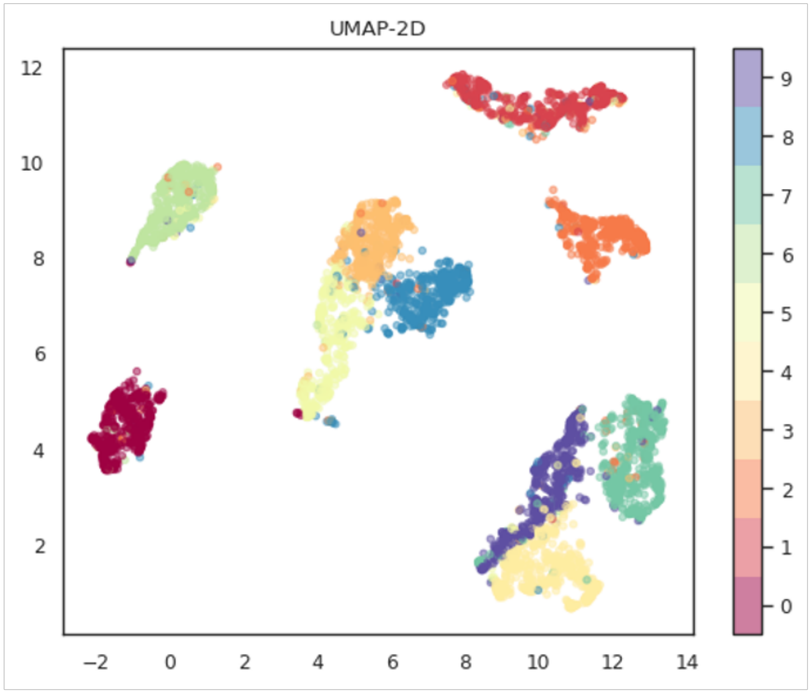
n_neighbours=15



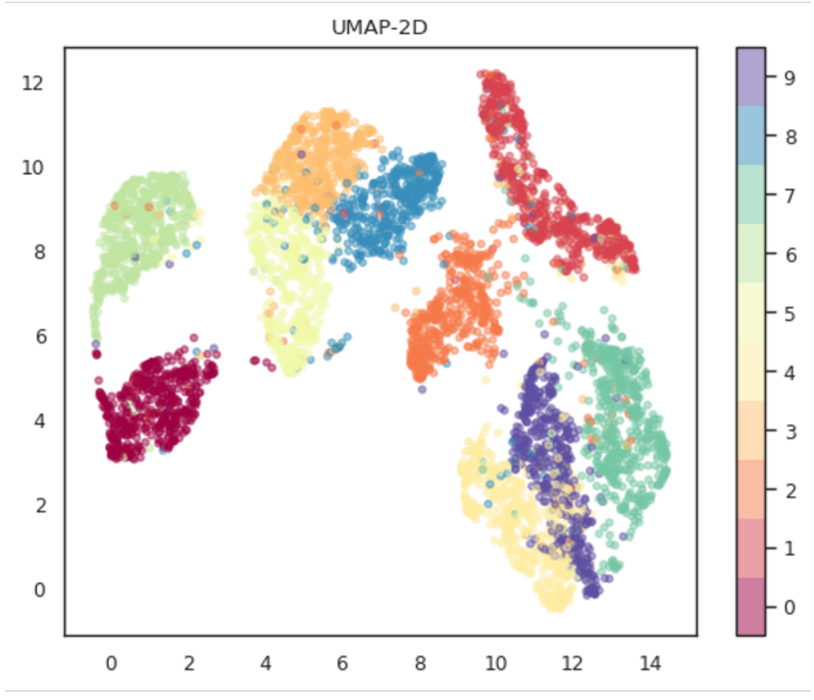
n_neighbours=400

UMAP-Parameters

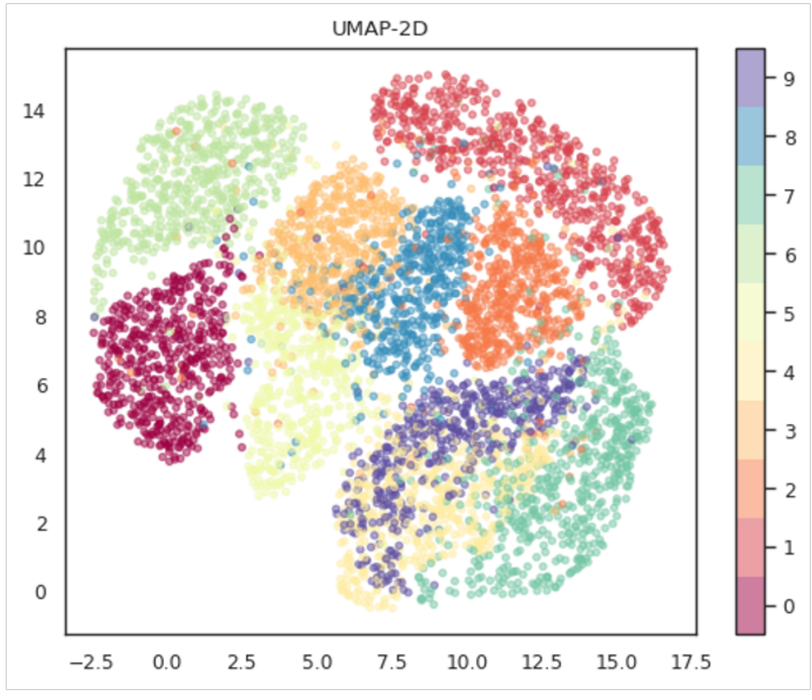
- d: dimensionality of the low-dimensional space (usually d=2 or d=3)
- n_neighbours: number of neighbours to construct the cover and initial graph
- min_dist: minimum distance between points in low-dimensional space



min_dist=0.0



min_dist=0.2



min_dist=0.8

UMAP-Parameters

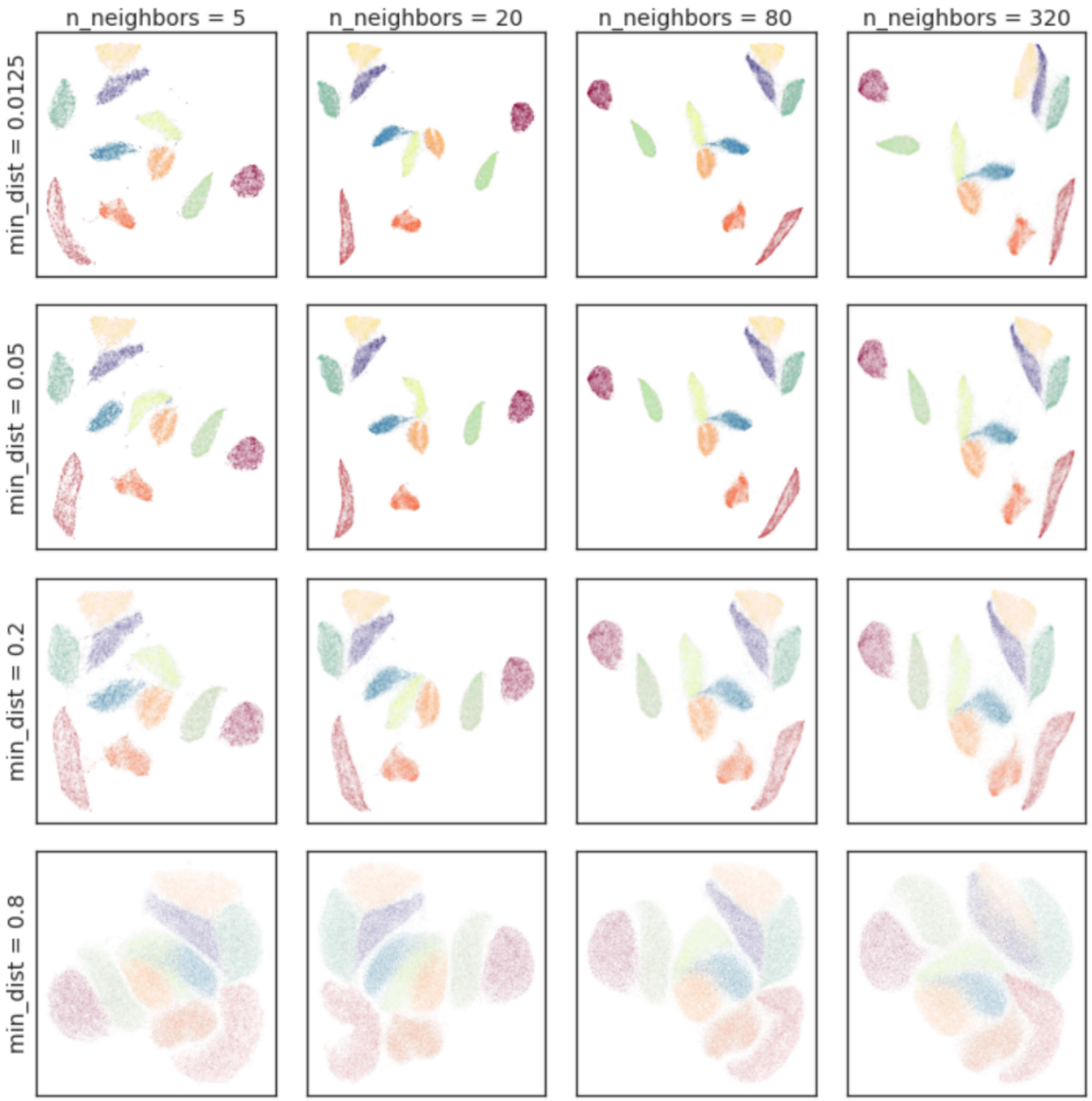
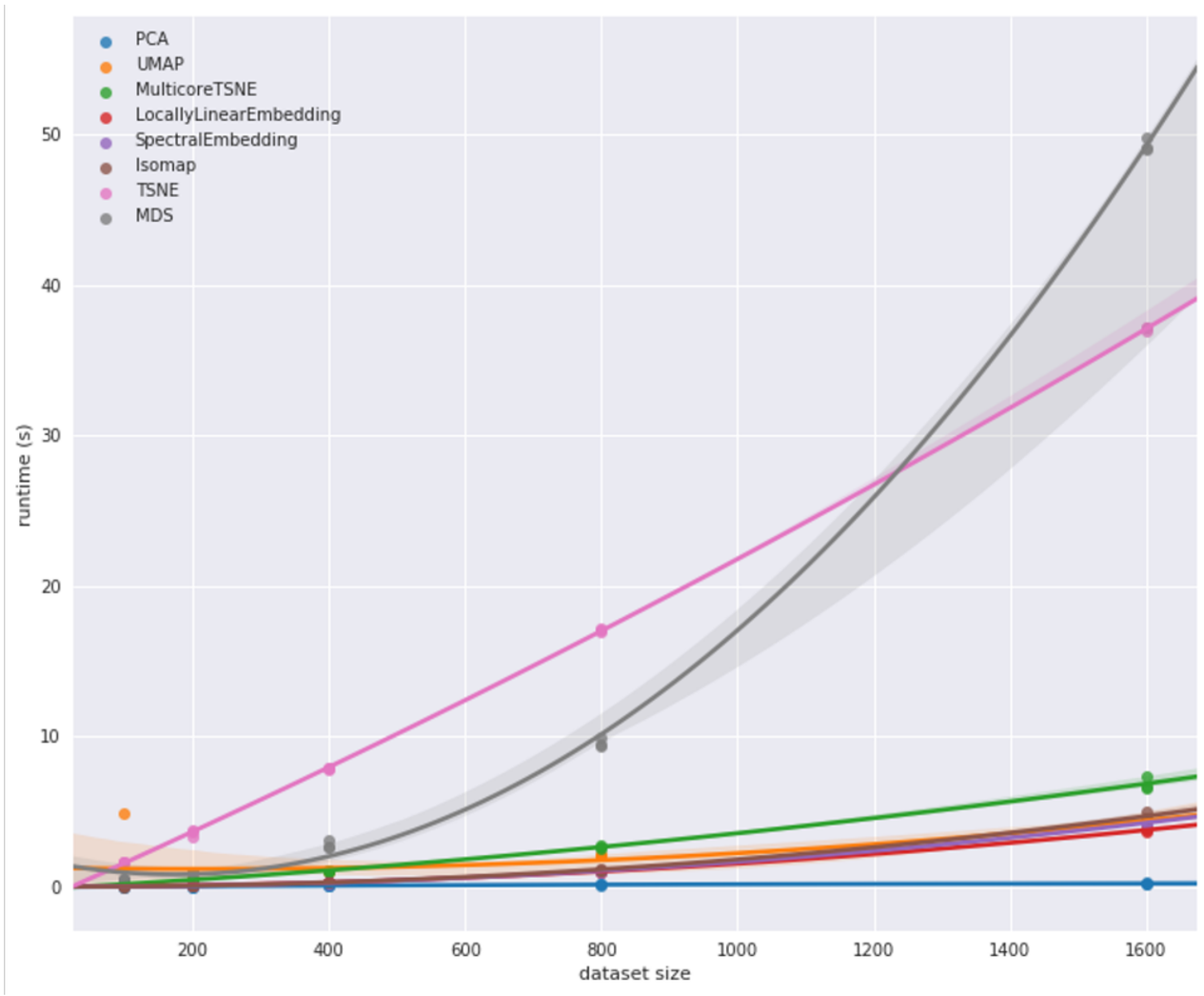
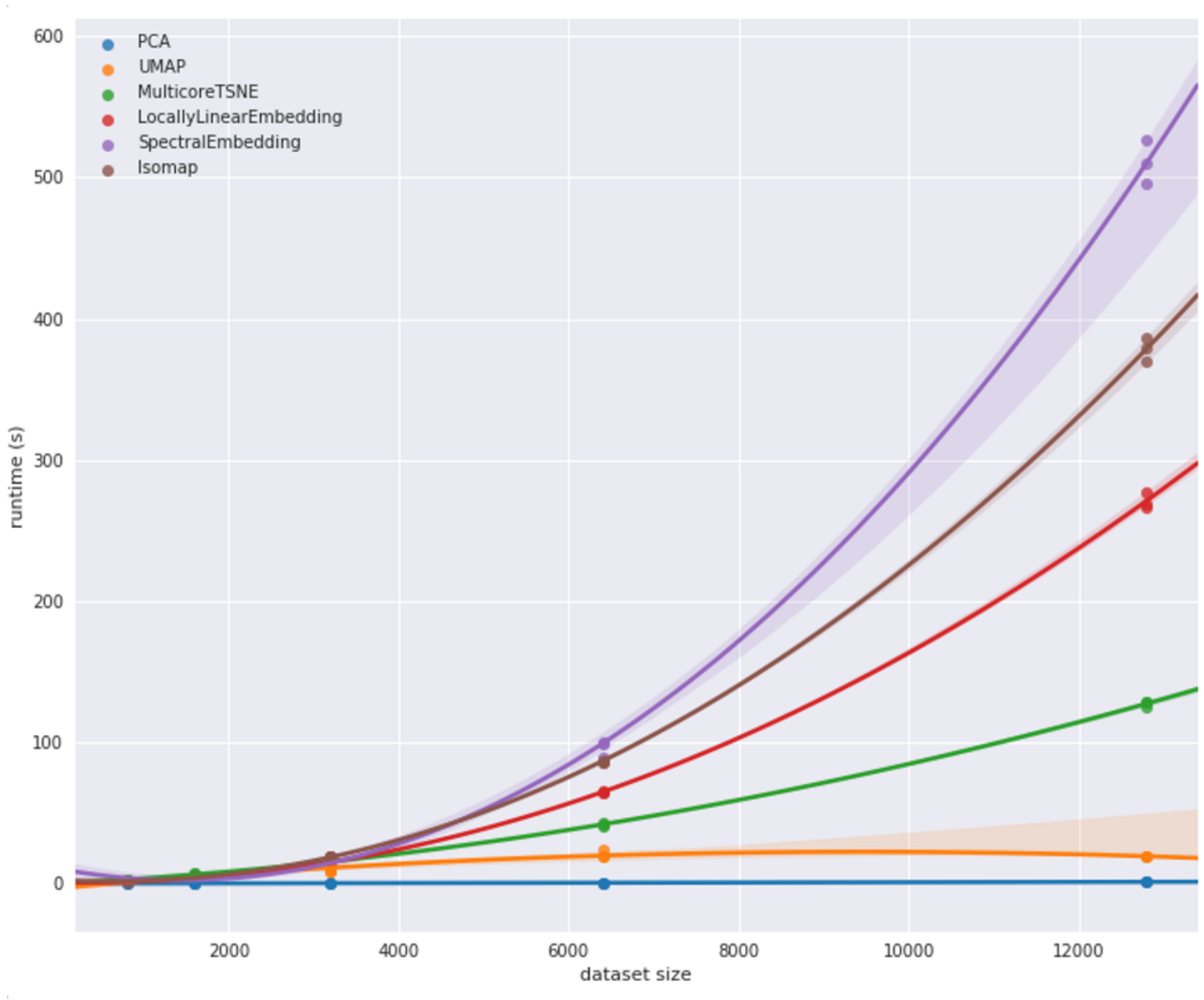


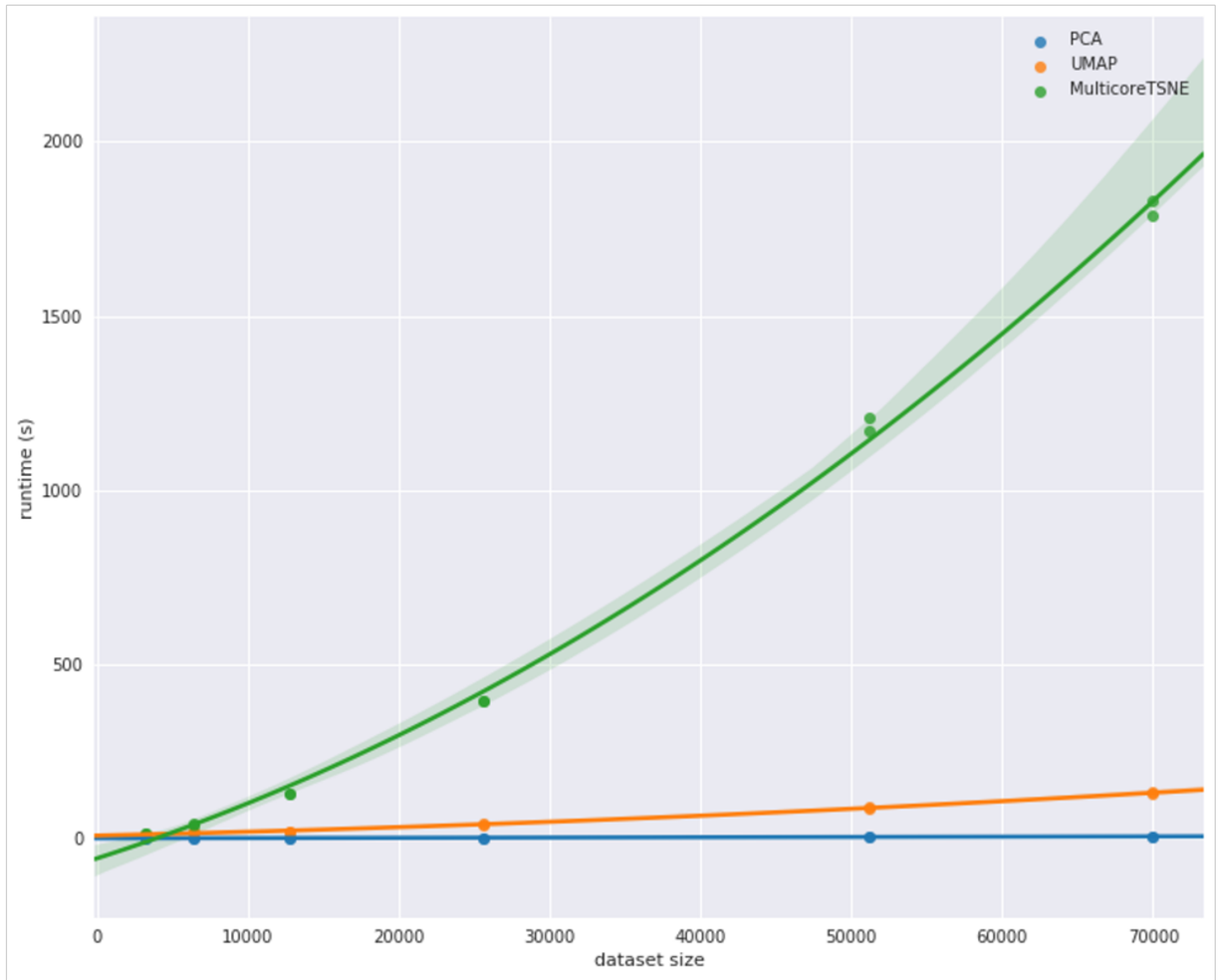
Figure 3: Variation of UMAP hyperparameters n and min-dist result in different embeddings. The data is the MNIST dataset, where each point is an 28x28 grayscale image of a hand-written digit.

UMAP - Resources





UMAP - Resources



	UMAP	FIt-SNE	t-SNE	LargeVis	Eigenmaps	Isomap
Pen Digits (1797x64)	9s	48s	17s	20s	**2s**	**2s**
COIL20 (1440x16384)	**12s**	75s	22s	82s	47s	58s
COIL100 (7200x49152)	**85s**	2681s	810s	3197s	3268s	3210s
scRNA (21086x1000)	**28s**	131s	258s	377s	470s	923s
Shuttle (58000x9)	**94s**	108s	714s	615s	133s	–
MNIST (70000x784)	**87s**	292s	1450s	1298s	40709s	–
F-MNIST (70000x784)	**65s**	278s	934s	1173s	6356s	–
Flow (100000x17)	**102s**	164s	1135s	1127s	30654s	–
Google News (200000x300)	**361s**	652s	16906s	5392s	–	–

Table 3: Runtime of several dimension reduction algorithms on various datasets. To allow a broader range of algorithms to run some of the datasets where subsampled or had their dimension reduced by PCA. The Flow Cytometry dataset was benchmarked on a 10% sample and the GoogleNews was subsampled down to 200,000 data points. Finally, the Mouse scRNA dataset was reduced to 1,000 dimensions via PCA. The fastest runtime for each dataset has been bolded.

UMAP - summary

Non linear method

Preserve local structure - Some info about global structure

Few interpretable hyperparameters

Efficient implementation

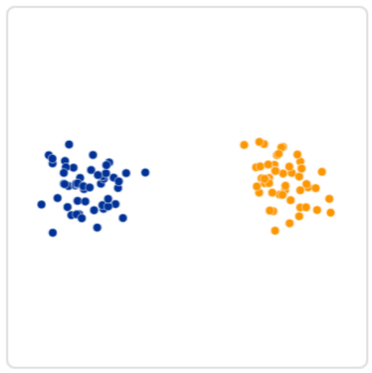
No interpretable dimensions

Hyperparameters choice is critical

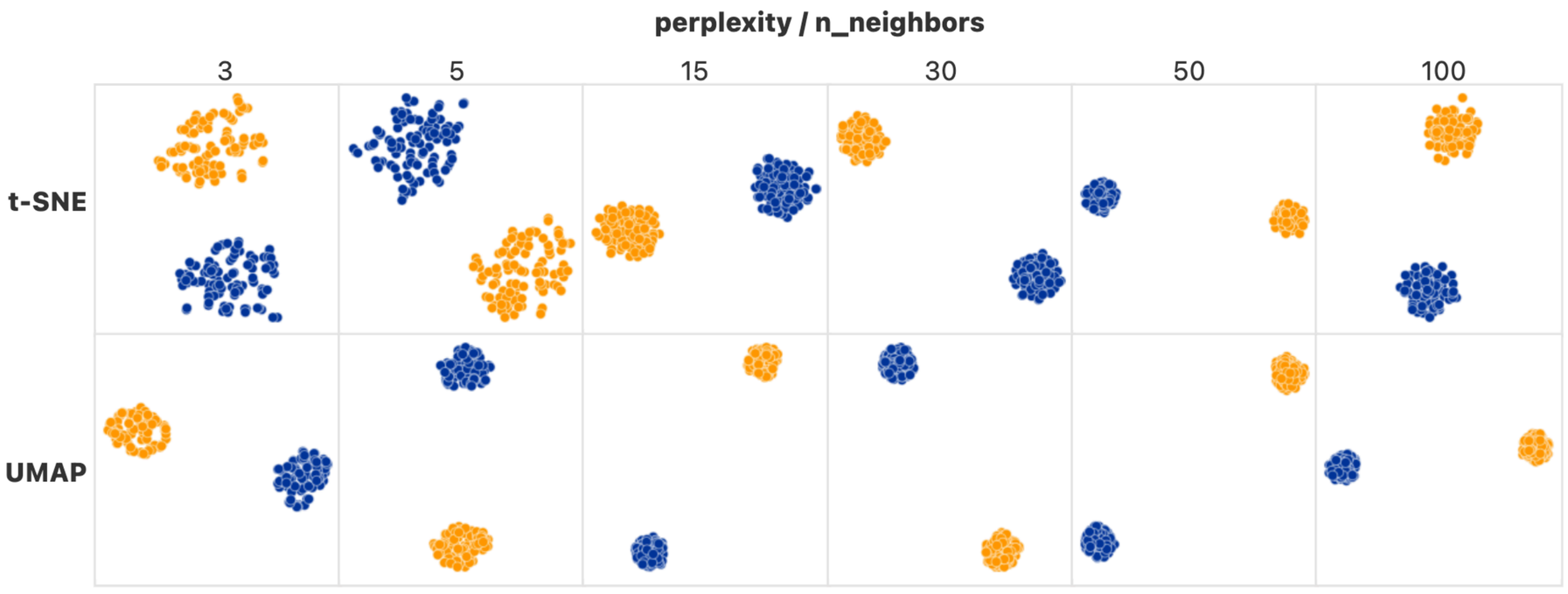
Assumption: existence of a manifold

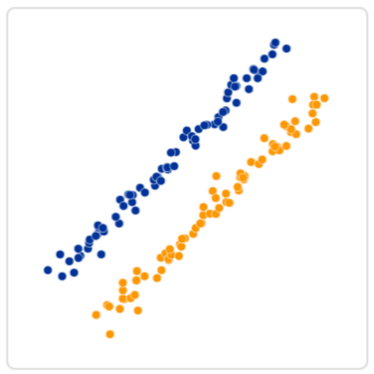
Based on topology, not on metrics

UMAP vs t-SNE

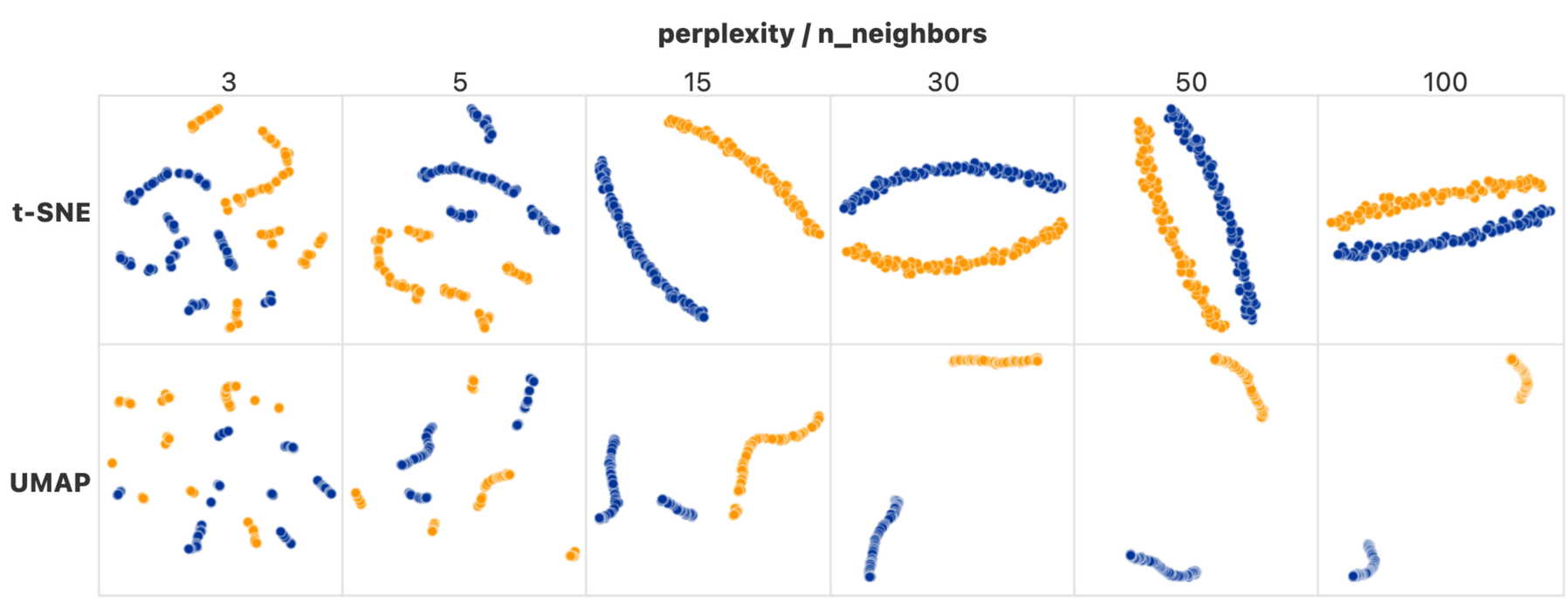


Original



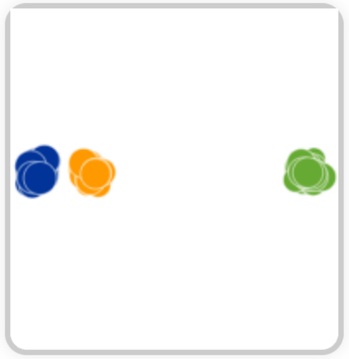


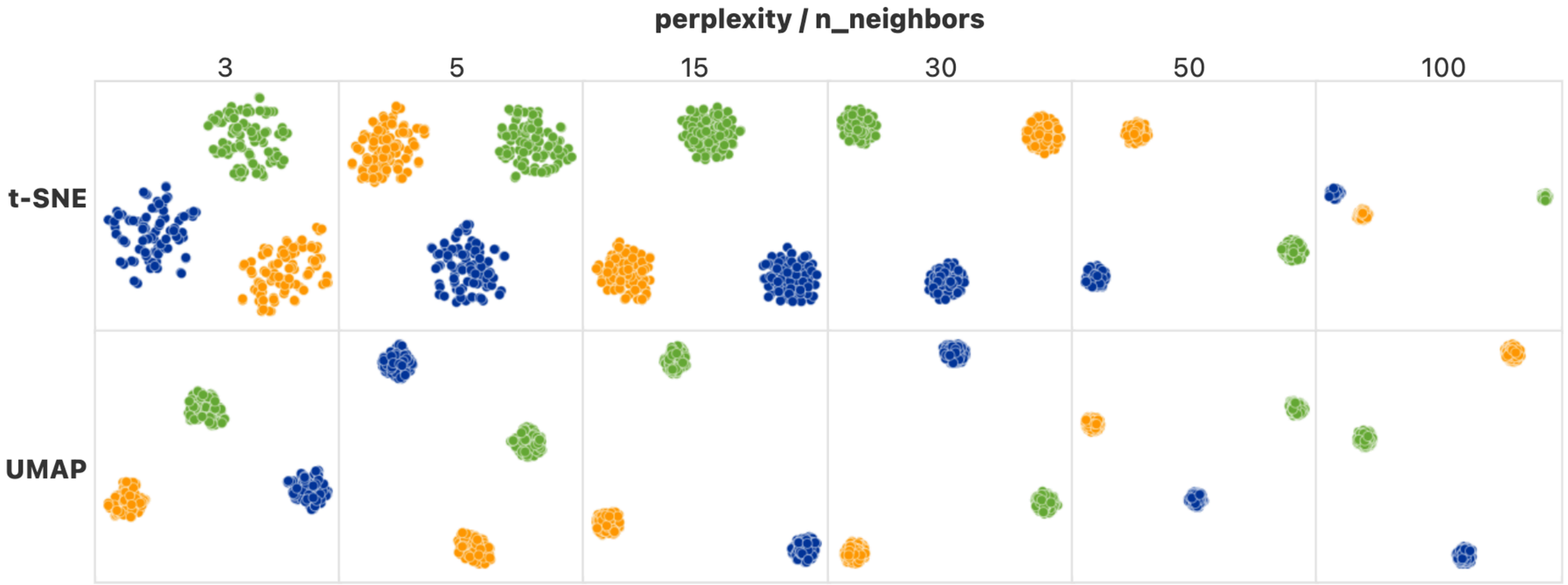
Original



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

UMAP vs t-SNE

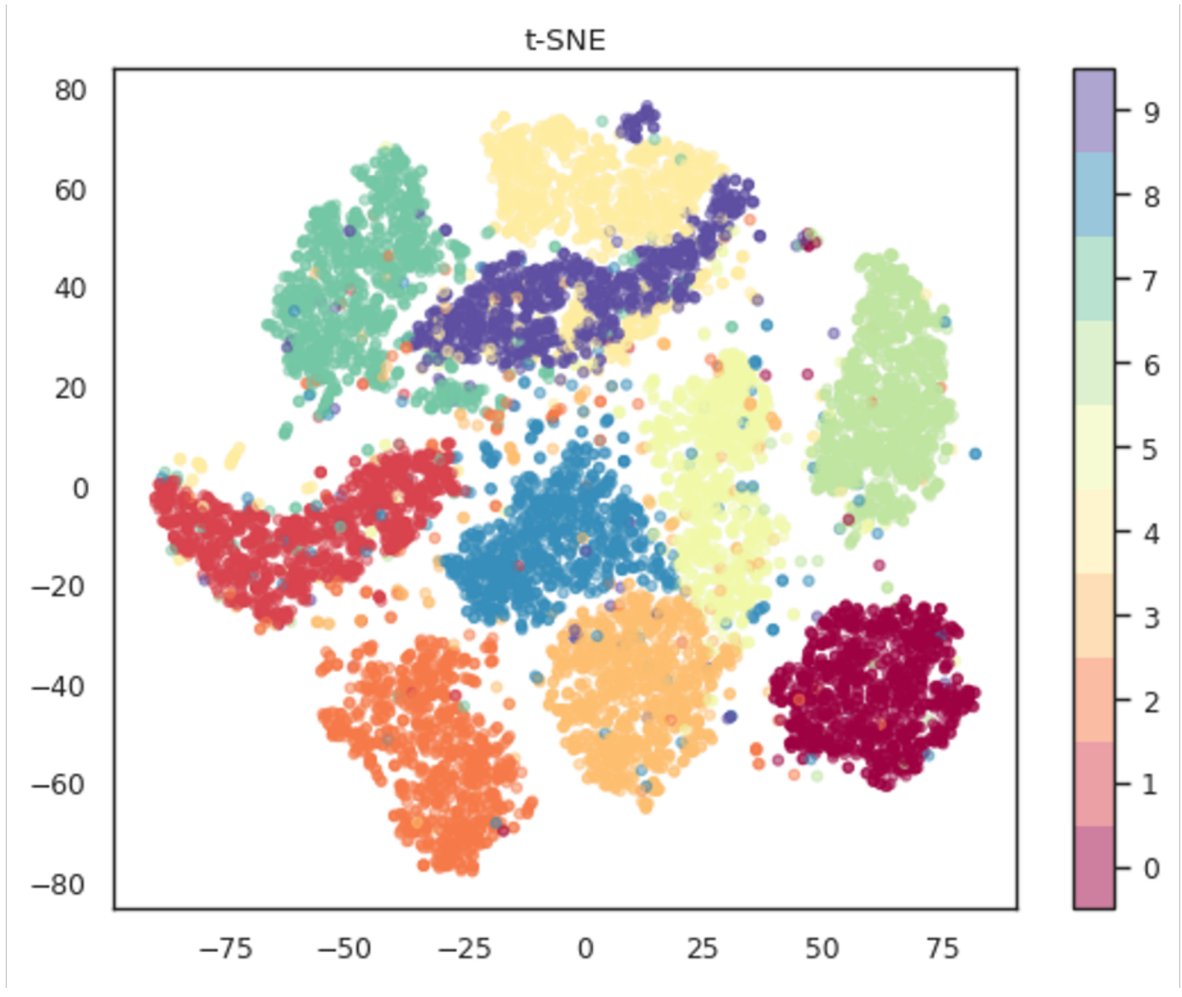


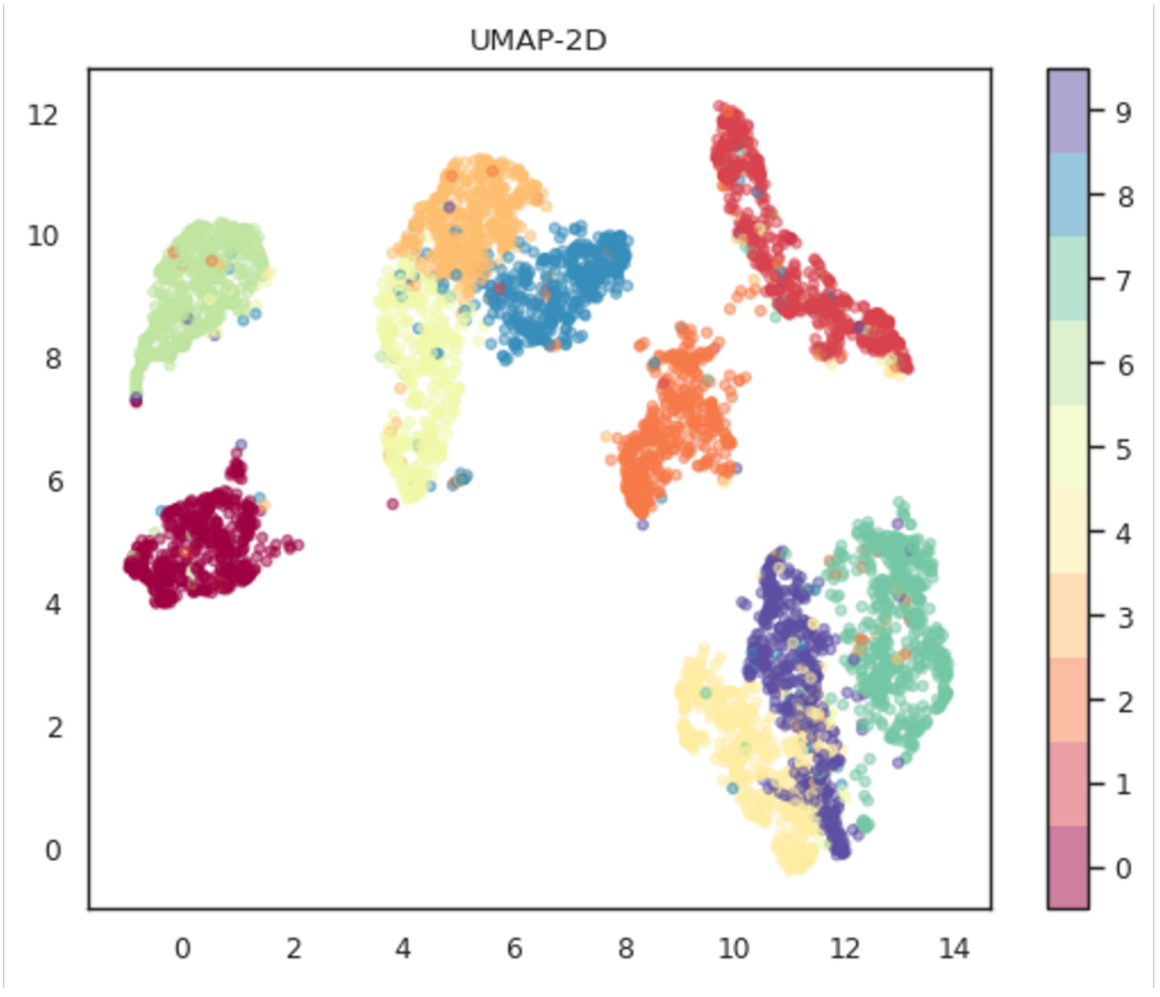


https://pair-code.github.io/understanding-umap/

UMAP vs t-SNE

mnist

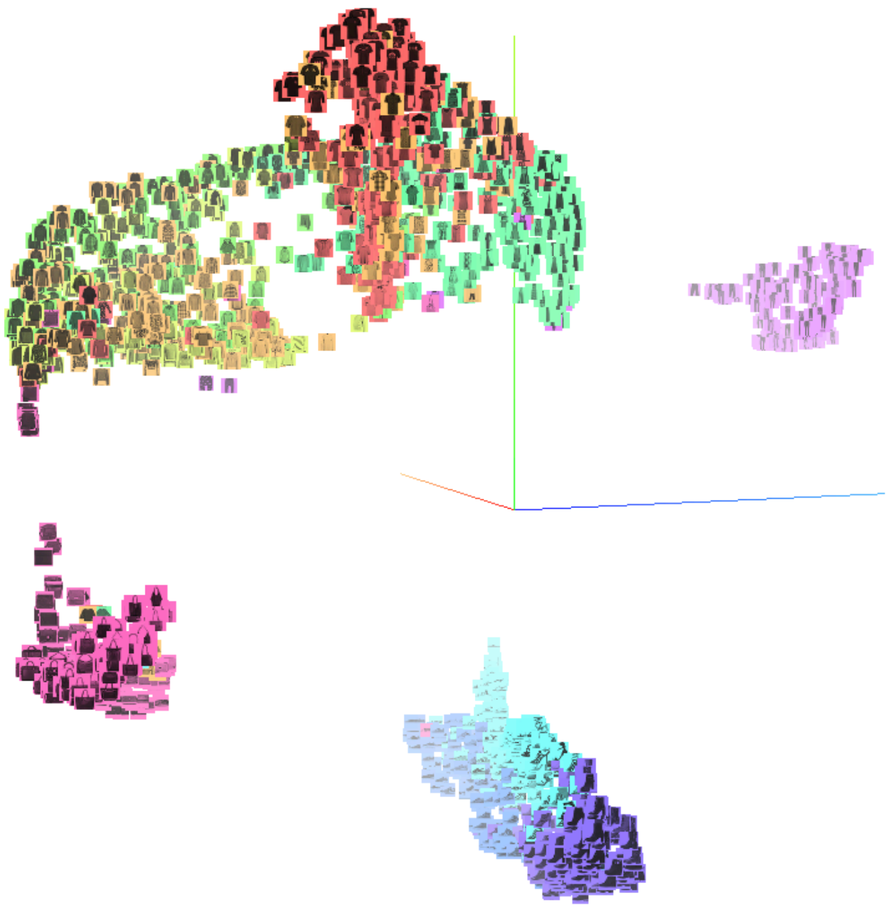




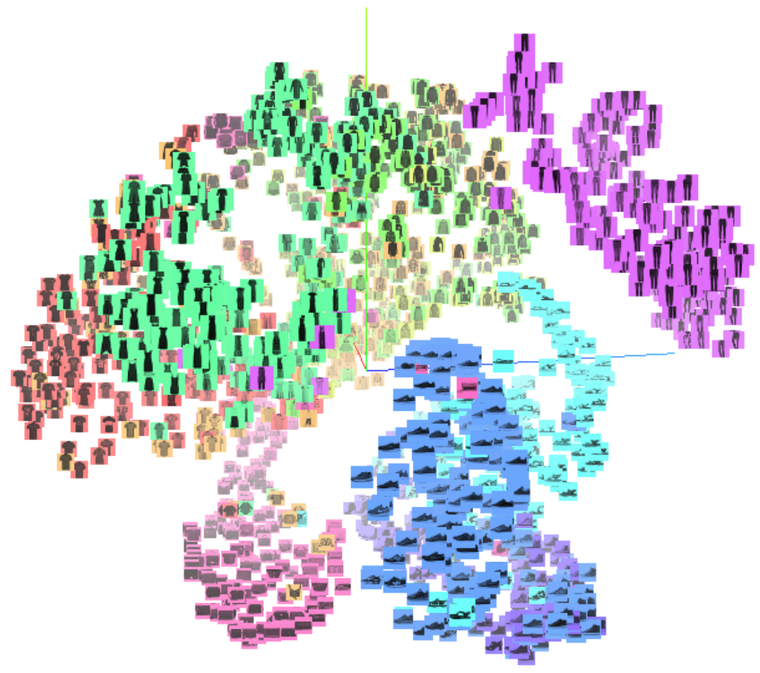
UMAP vs t-SNE

Fashion mnist

UMAP



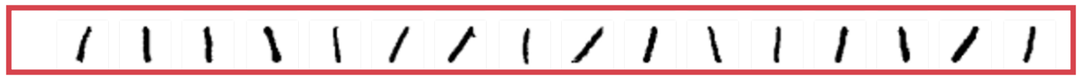
t-SNE

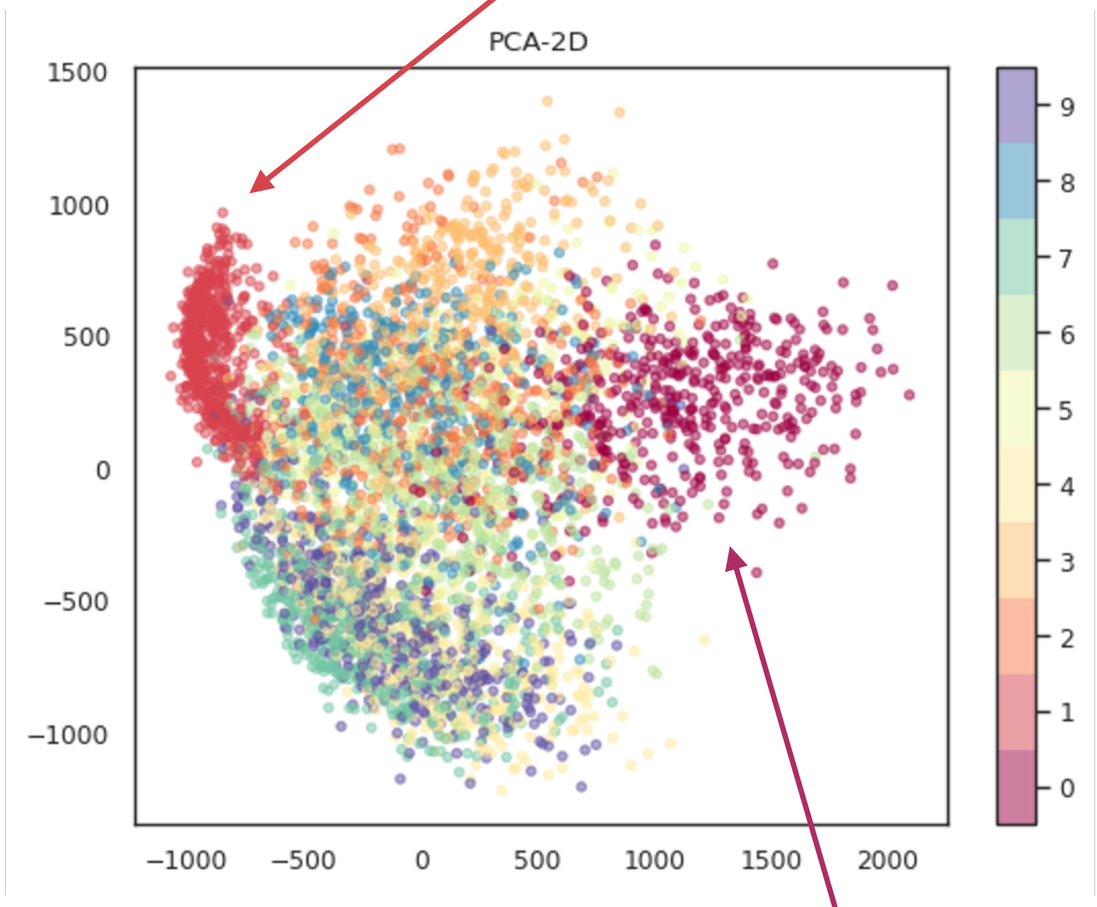


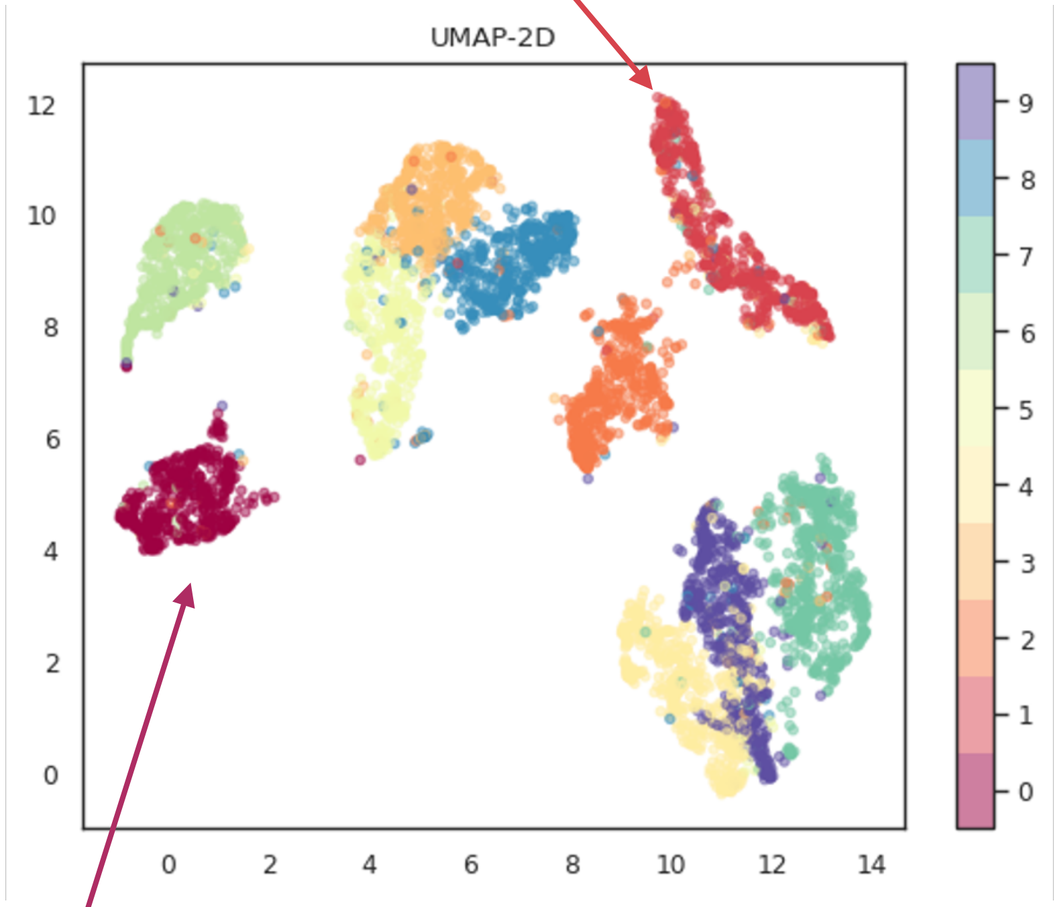
T-shirt/top | Shirt | Pullover | Coat | Dress | Sandal | Sneaker | Ankle boot | Trouser | Bag

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

DensMAP




PCA-2D
1500
1000
500
0
−500
−1000
−1000
−500
0
500
1000
1500
2000
9
8
7
6
5
4
3
2
1
0


UMAP-2D
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
9
8
7
6
5
4
3
2
1
0

DdensMAP

nature biotechnology

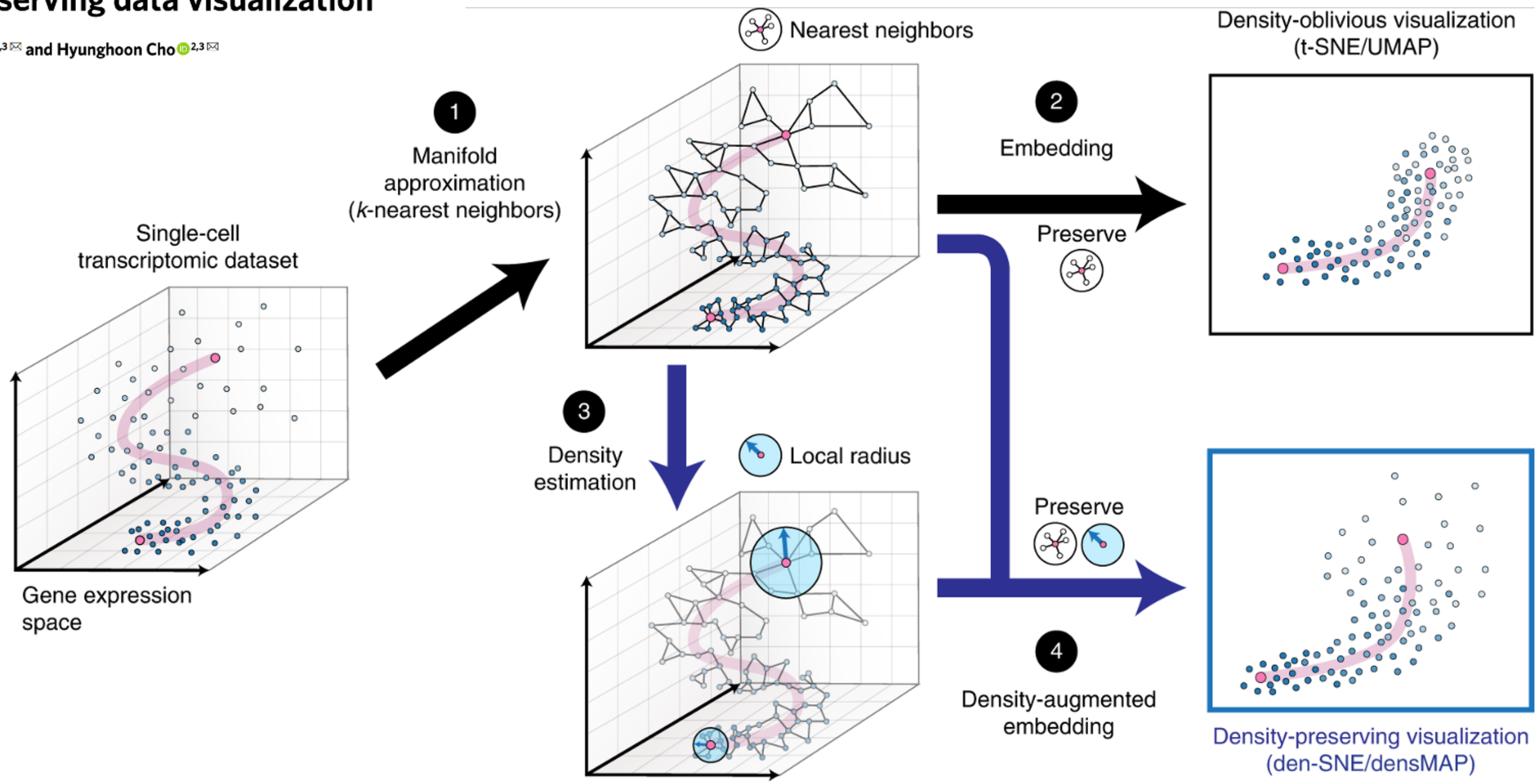
ARTICLES

https://doi.org/10.1038/s41587-020-00801-7

Check for updates

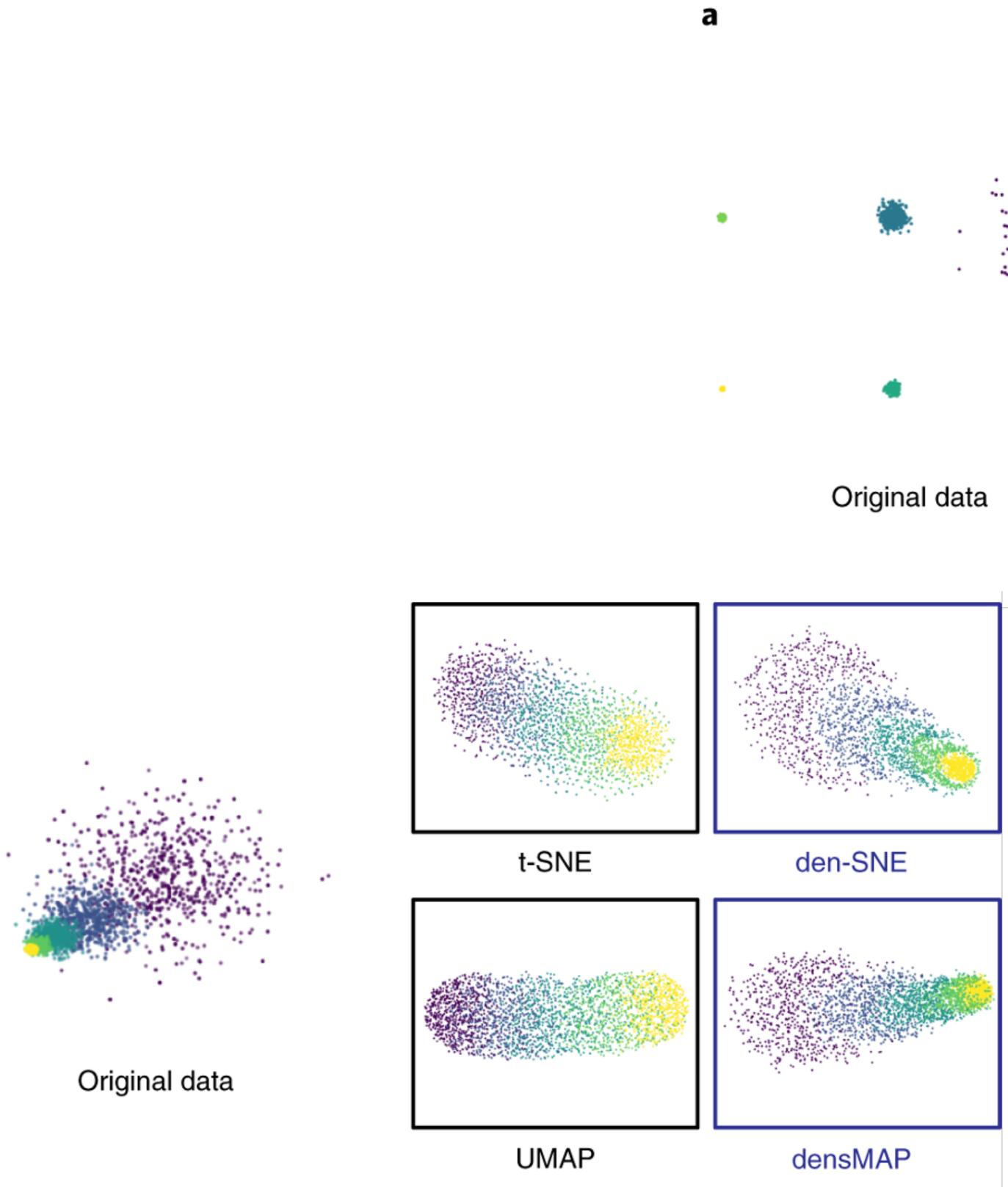
Assessing single-cell transcriptomic variability through density-preserving data visualization

Ashwin Narayan[1,2,3], Bonnie Berger[1,2,3] ✉ and Hyunghoon Cho[2,3] ✉



https://doi.org/10.1038/s41587-020-00801-7

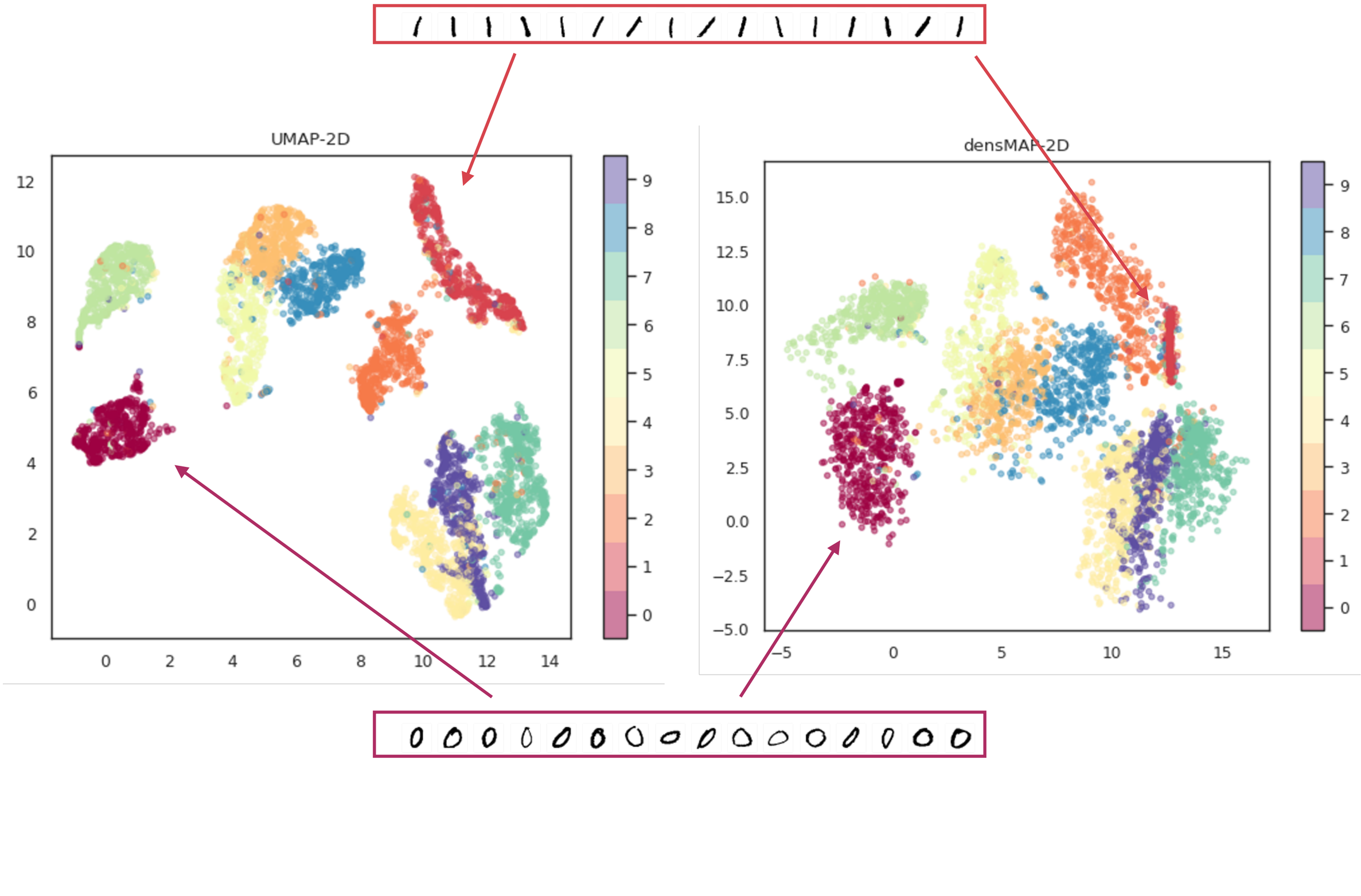
DensMAP



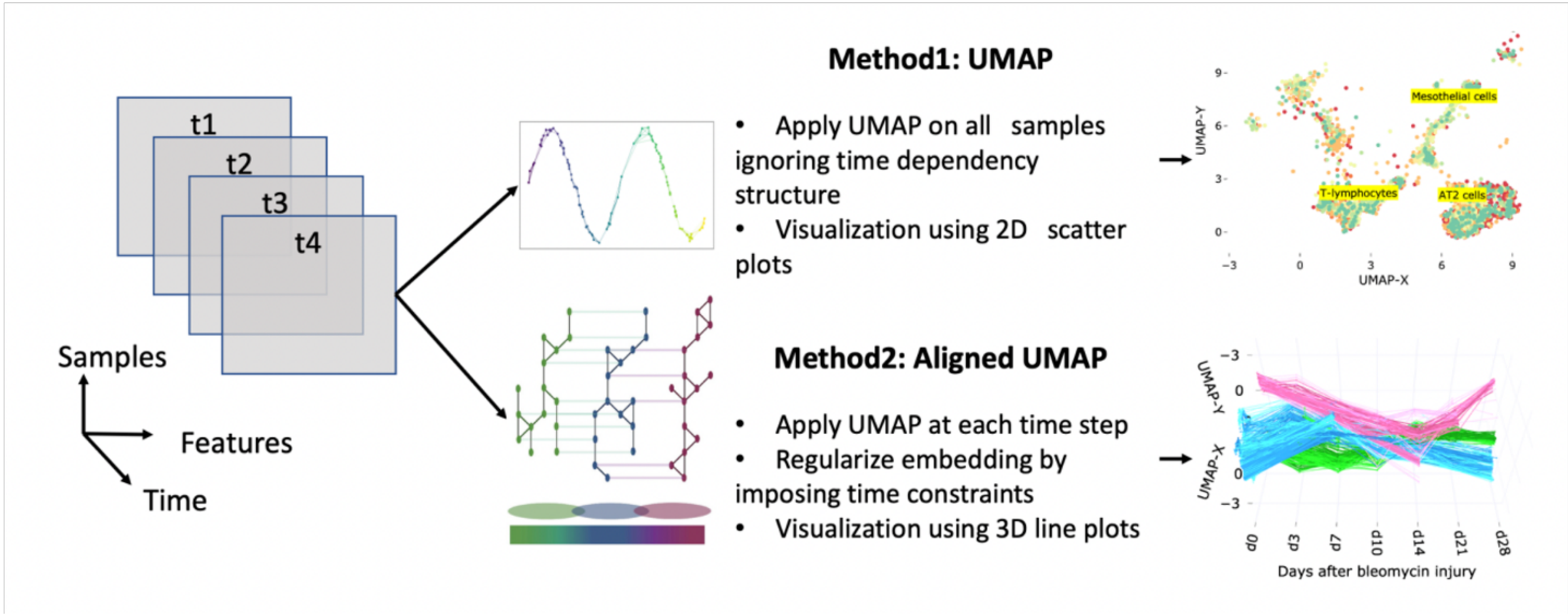
https://doi.org/10.1038/s41587-020-00801-7

DensMAP

mnist



Aligned UMAP



New parameters:

- alignment_regularisation: weigths importance of retaining alignment across different projections in time
- alignment_window_size: how far alignment is preserved across time (both in past and future)

Aligned UMAP

Patterns

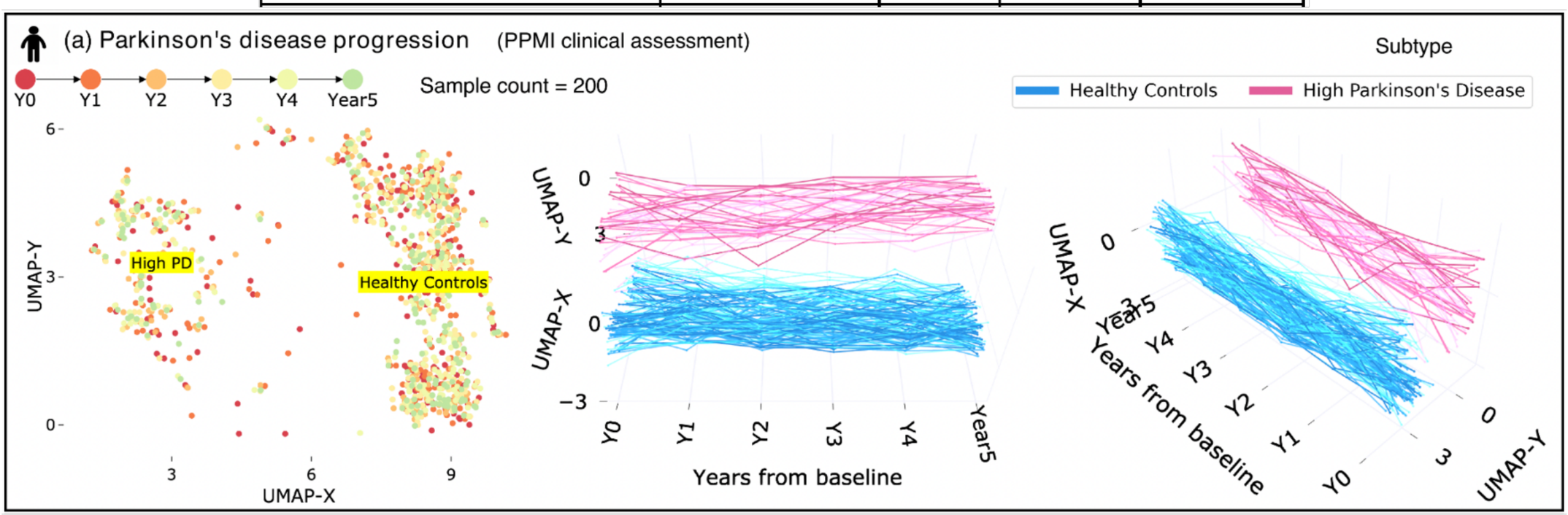
CellPress
OPEN ACCESS

Descriptor

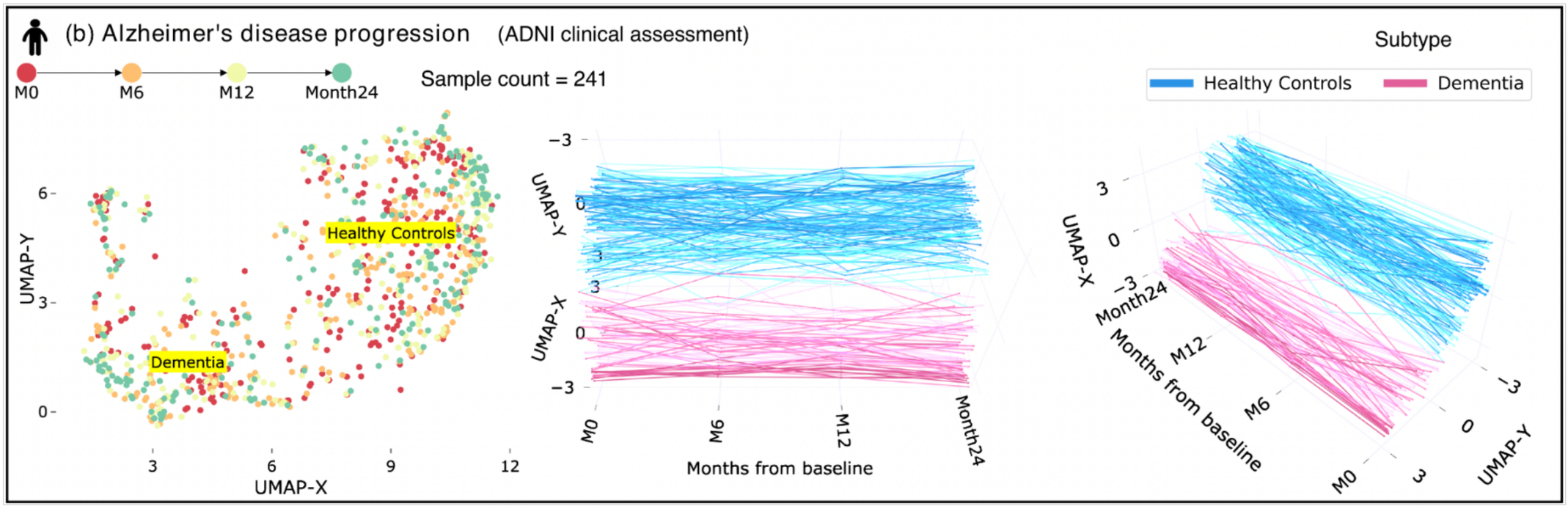
Application of Aligned-UMAP to longitudinal biomedical studies

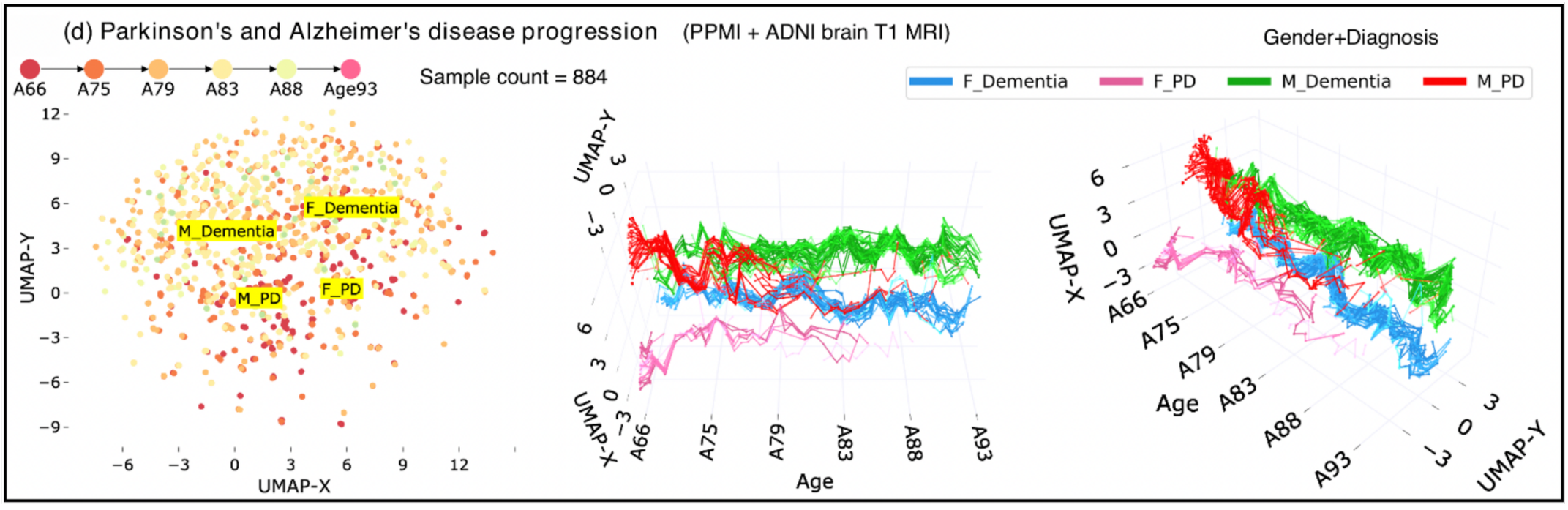
Anant Dadu,[1,2,3] Vipul K. Satone,[4] Rachneet Kaur,[4] Mathew J. Koretsky,[2,5] Hirotaka Iwaki,[2,3,5] Yue A. Qi,[2] Daniel M. Ramos,[2] Brian Avants,[6] Jacob Hesterman,[6] Roger Gunn,[6] Mark R. Cookson,[2,5] Michael E. Ward,[2,7] Andrew B. Singleton,[2,5] Roy H. Campbell,[1] Mike A. Nalls,[2,3,5] and Faraz Faghri[2,3,5,8,*]

Dataset	Modality	# samples	# features	# time sequences
PPMI clinical data[2]	Clinical Assessment	476	122	6
ADNI clinical data[3]	Clinical Assessment	435	78	4
PPMI-ADNI T1 MRI	MRI T1 Imaging	2,836	406	52



Aligned UMAP





…MAP

TriMAP (2019)

Link to paper (https://arxiv.org/abs/1910.00204)

A DR algorithm based on triplets

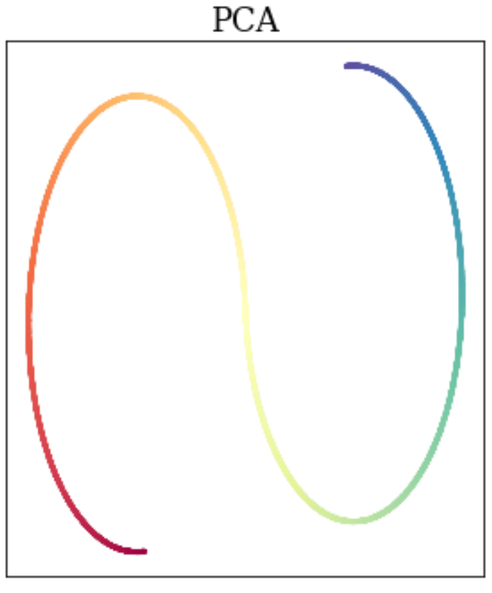
PacMAP (2021)

Link to paper (jmlr.org)

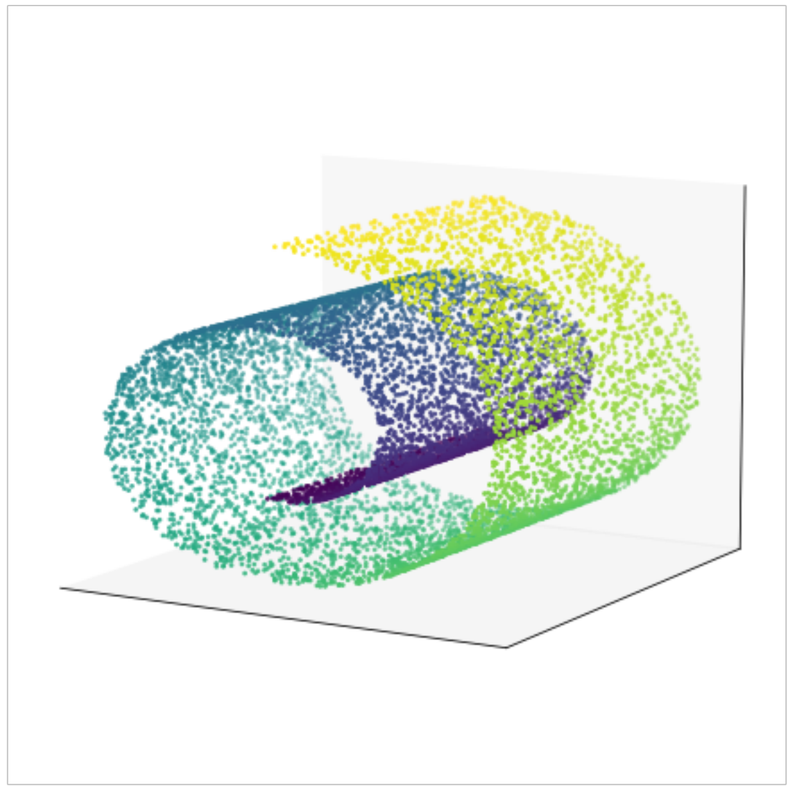
PacMAP description by the author (youtube.com)

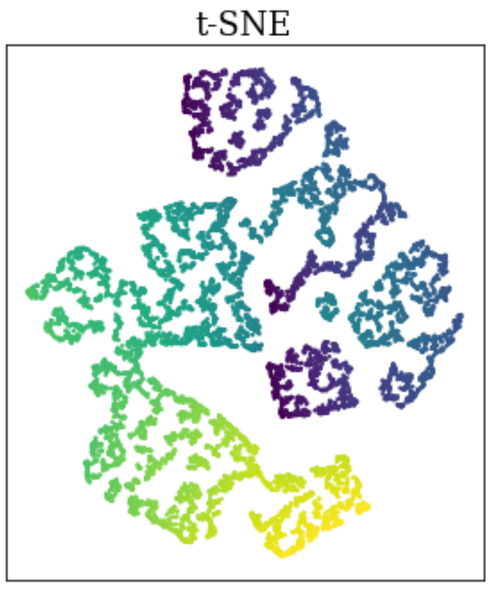
A DR algorithm derived from an empirical approach to deciphering t-SNE, UMAP, TriMAP for Data Visualization

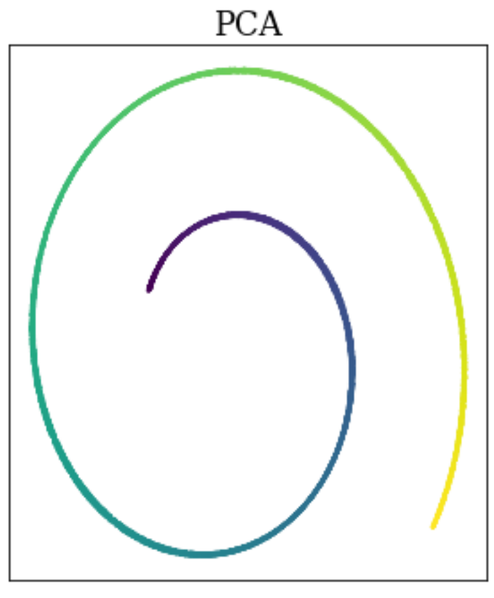
TriMAP



TriMAP

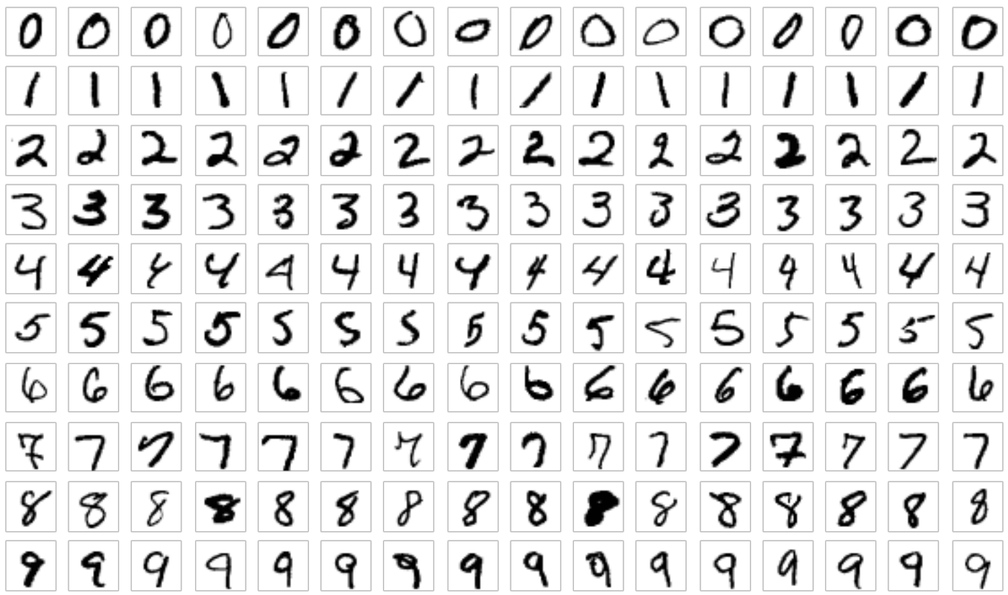


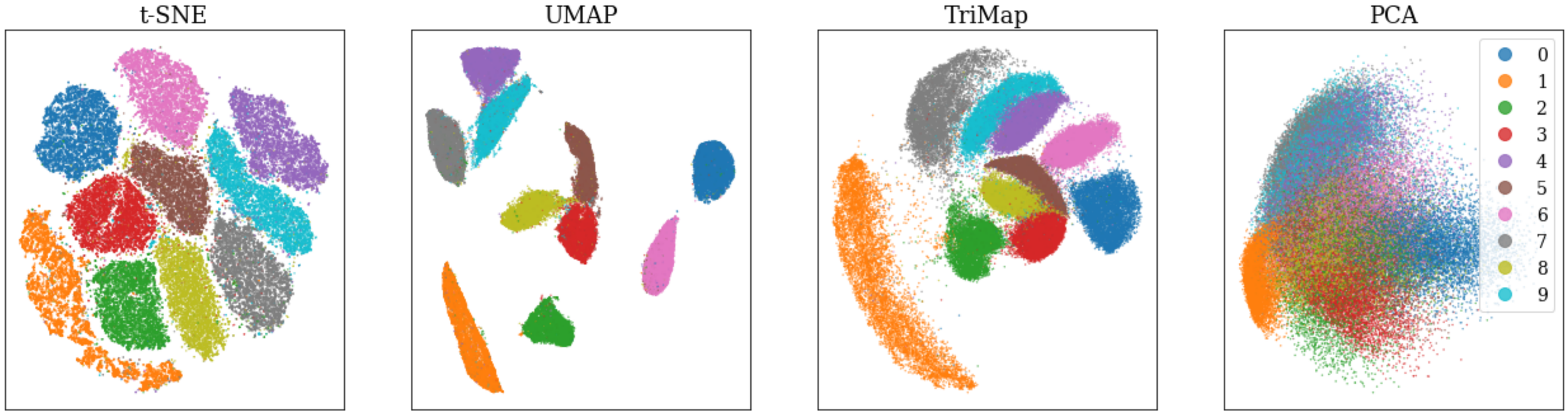




TriMAP

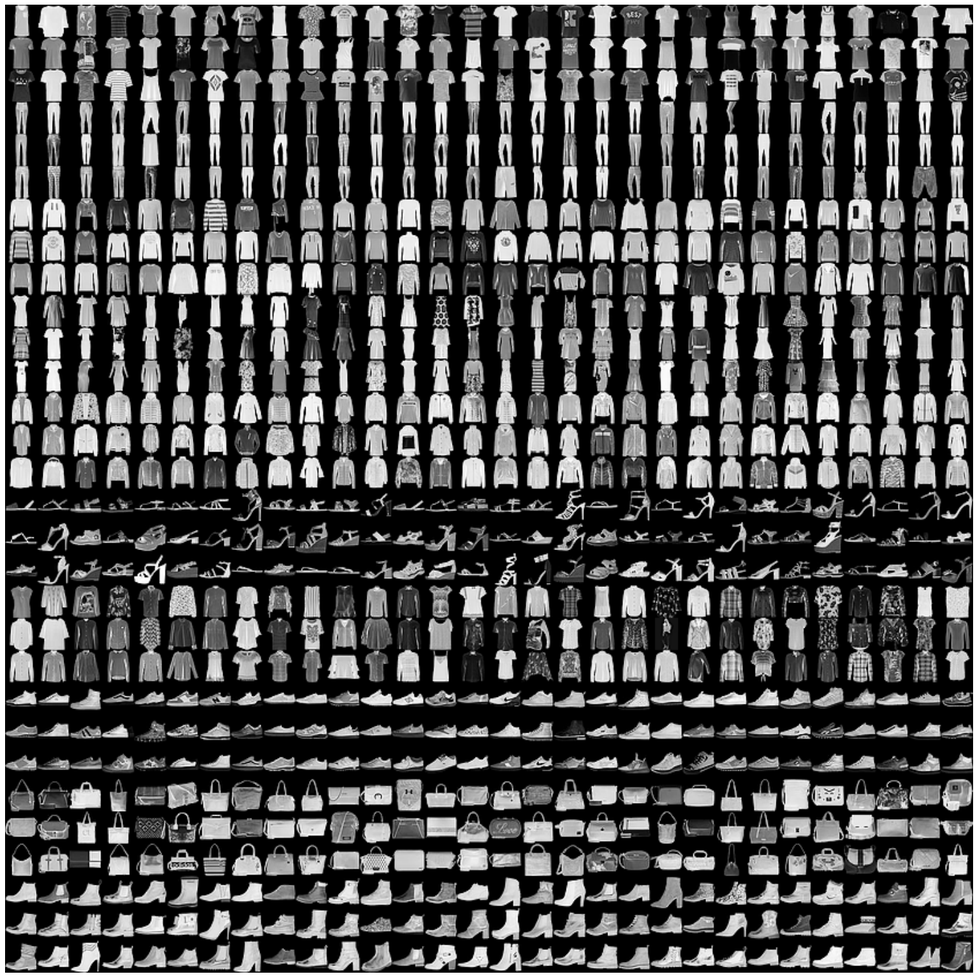
mnist

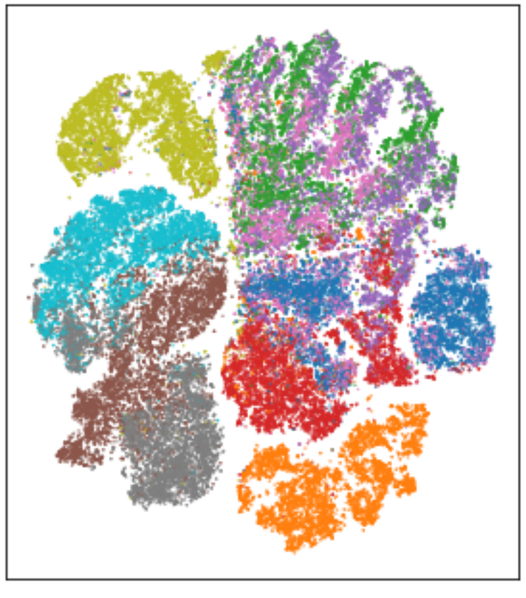


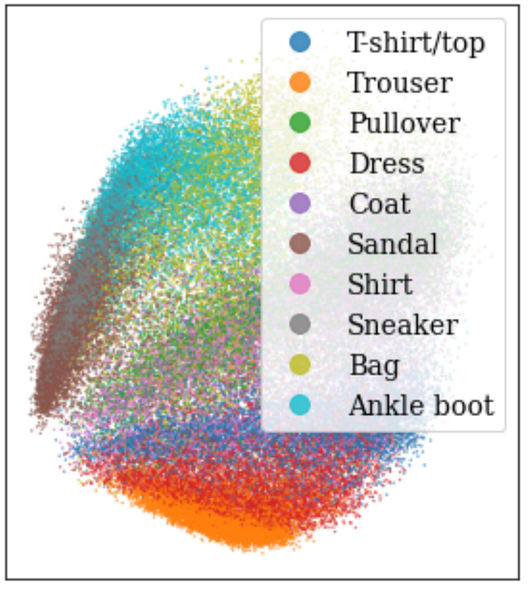


TriMAP

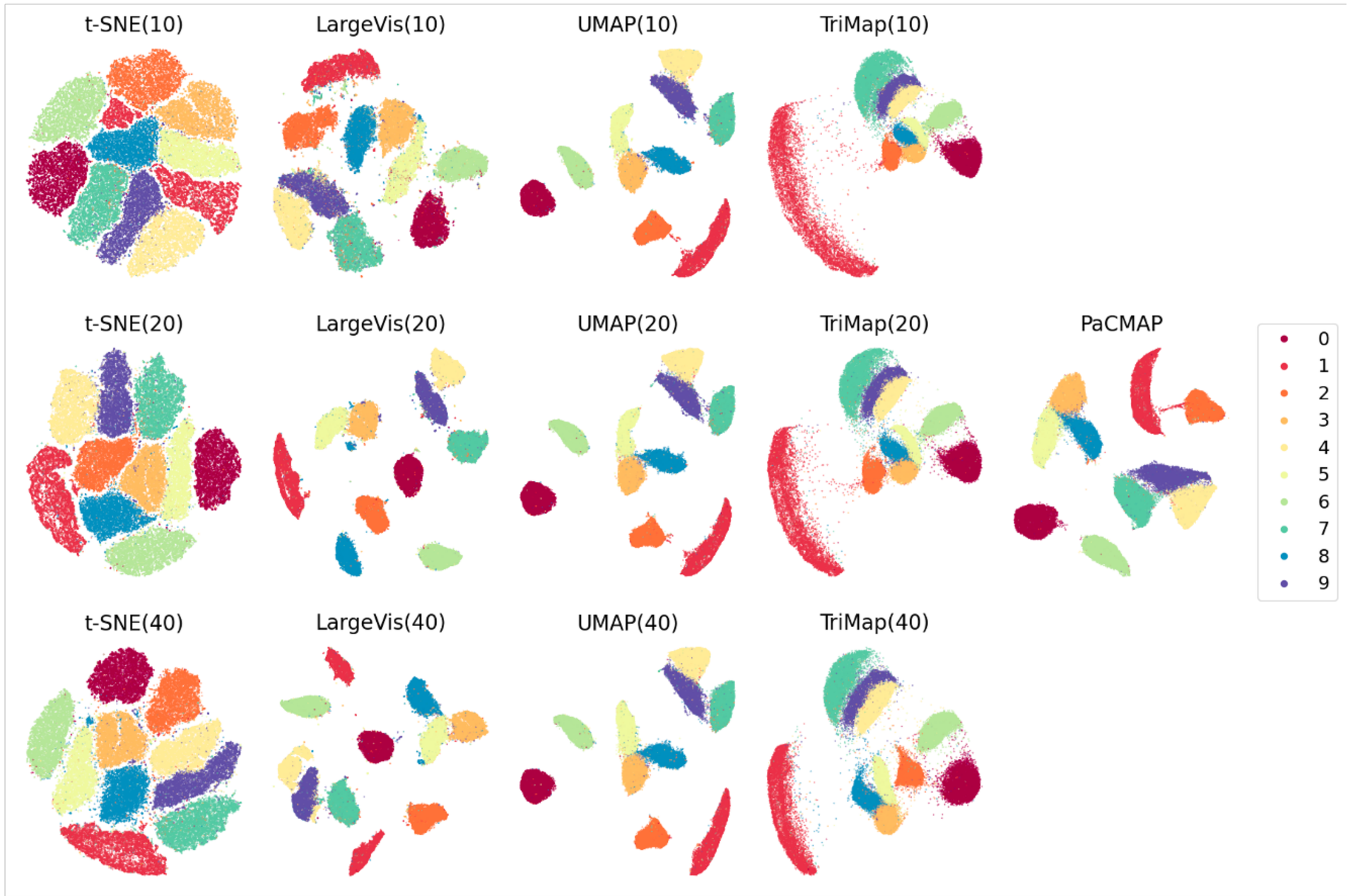
Fashion mnist





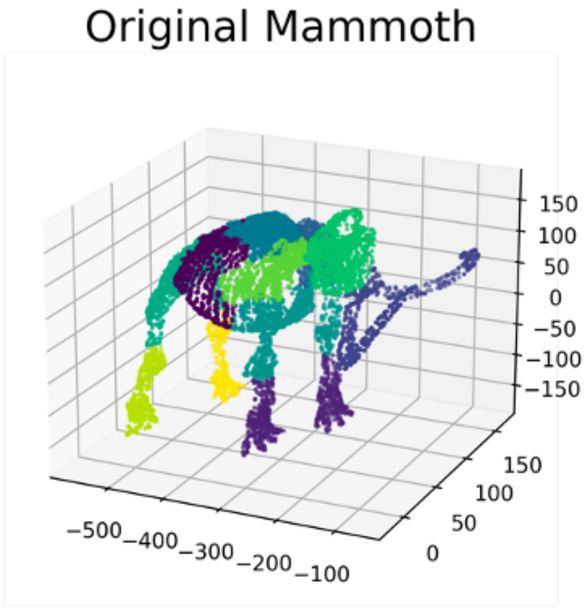


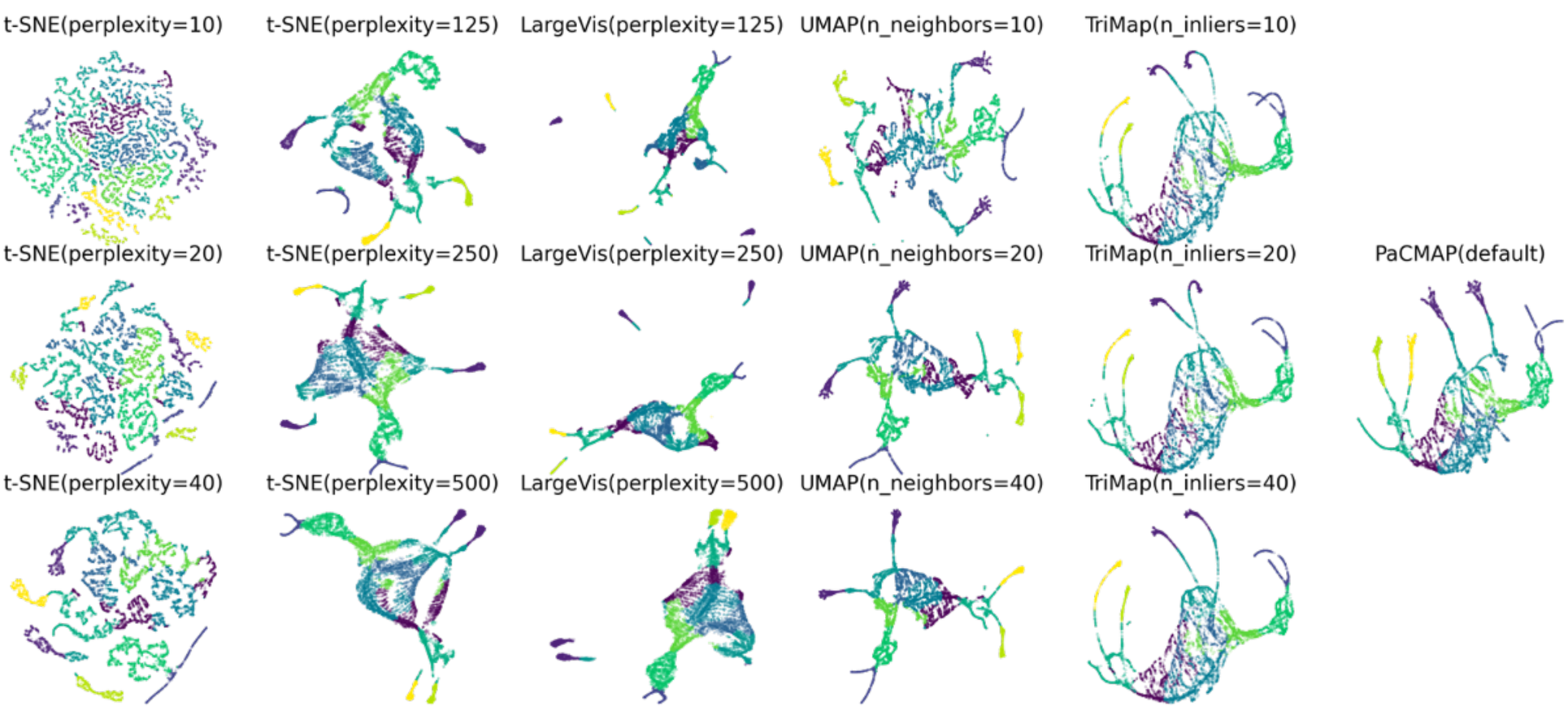
PaCMAP



Link to paper (jmlr.org)

PaCMAP





Link to paper (jmlr.org)