Dimensionality reduction and UMAP

Jan Brabec

Dimensionality reduction

- Goal: ***represent m-dimensional data in n-dimensional space, where m > n***
- Applications:
  - Data analysis and visualisation
  - Feature selection
  - Feature extraction: *Find latent features in your data*.
  - Reduces the time (training/testing) and storage space required

Two high-level aproaches

- **Matrix factorization**

  PCA

  Linear Autoencoder

  Word2Vec [1]

  GloVE [2]

  Variational Autoencoders

- **Neighbour graphs**

  IsoMap

  Laplacian Eigenmaps

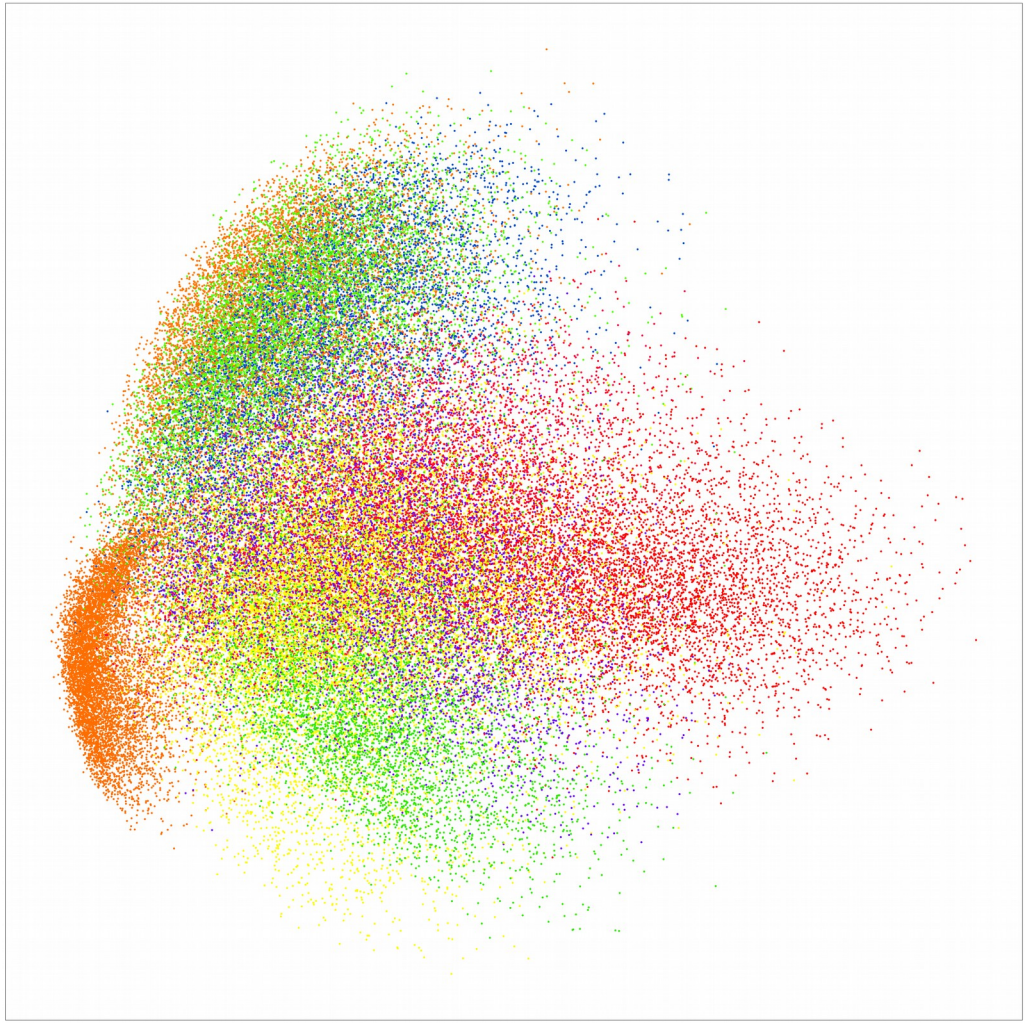
  t-SNE

  UMAP

[1] Levy, Omer, and Yoav Goldberg. "Neural word embedding as implicit matrix factorization." *Advances in neural information processing systems*. 2014.

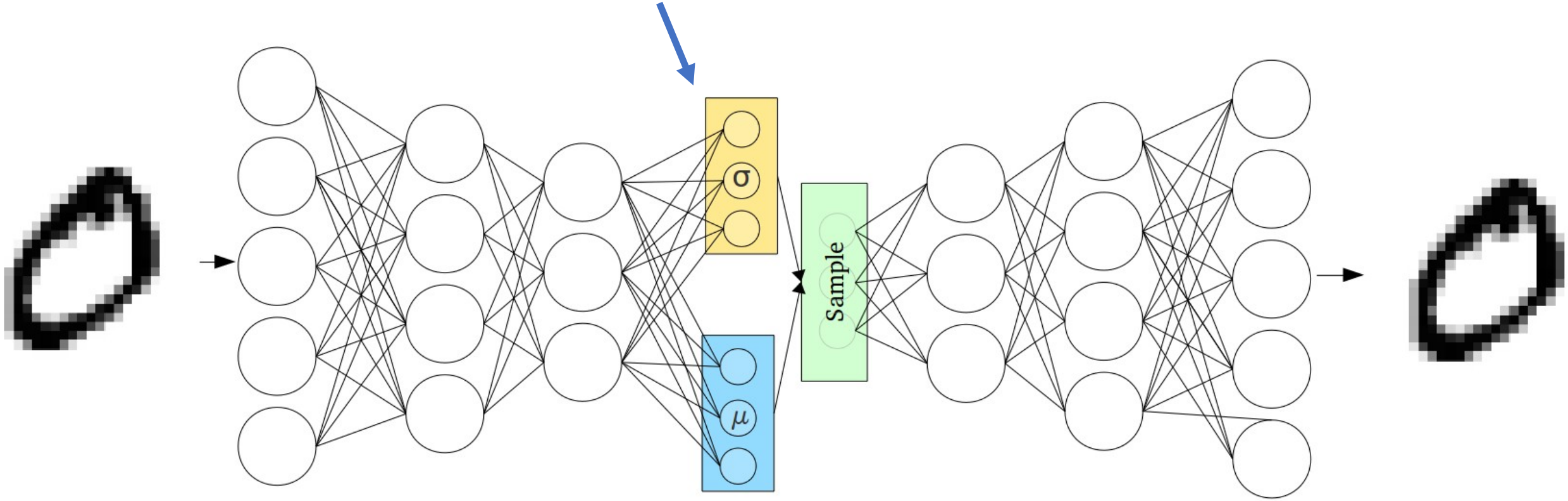
[2] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation

PCA on MNIST

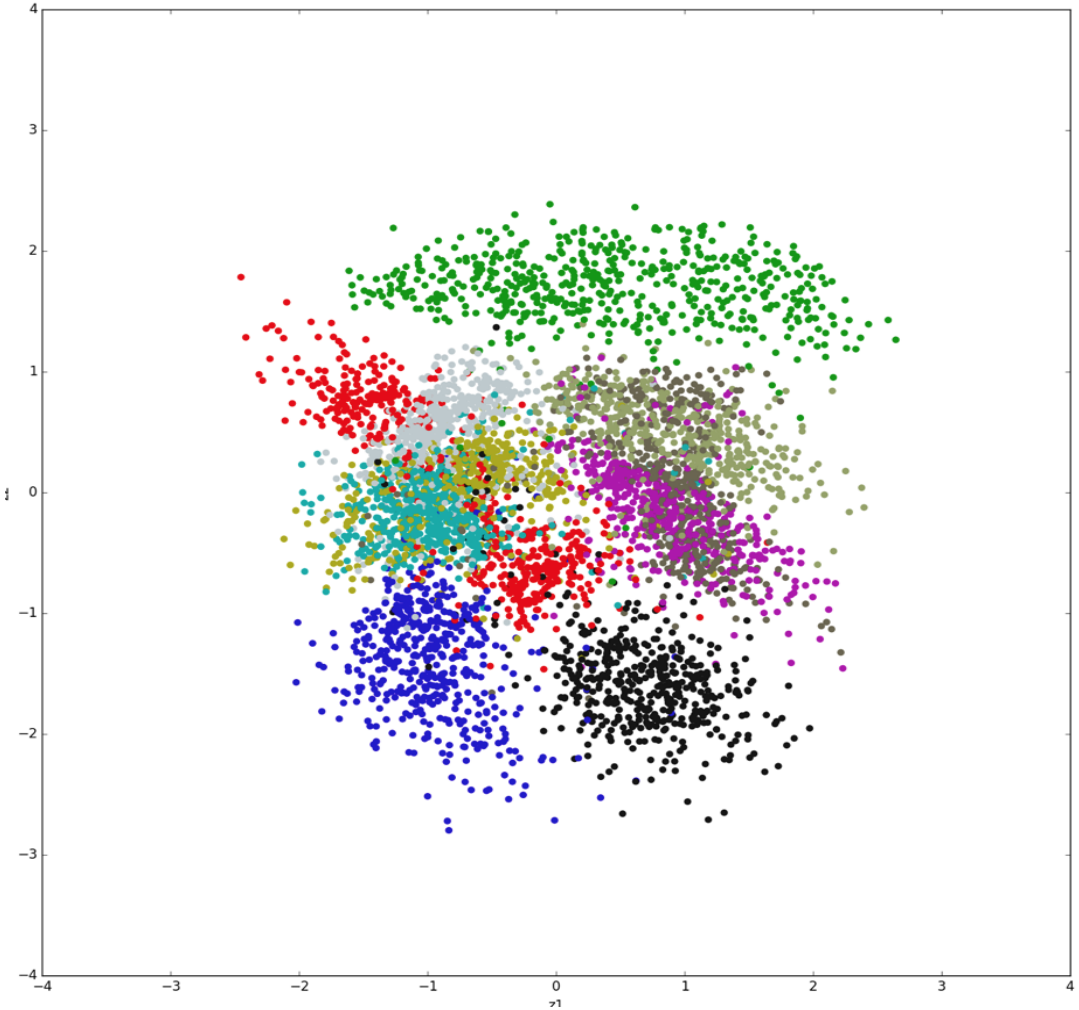


Variational Autoencoder

- Autoencoders but the latent space is continuous



VAE on MNIST



https://github.com/musyoku/variational-autoencoder

t-SNE

1. Compute pair-wise similarities between high-dimensional points (**P**)
2. Distribute the low-dimensional representations randomly and compute pair-wise similarities between them too (**Q**)
3. Minimize KL-divergence between **P** and **Q** with gradient descent

$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{i} P(i) \log\left(\frac{Q(i)}{P(i)}\right),$$

Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.Nov (2008): 2579-2605.

t-SNE: Similarities

$$p_{j|i} = \frac{\exp\left(-|x_i - x_j|^2 \Big/ 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-|x_i - x_k|^2 \Big/ 2\sigma_i^2\right)}$$

$$q_{ij} = \frac{f(|x_i - x_j|)}{\sum_{k \neq i} f(|x_i - x_k|)} \quad \text{with} \quad f(z) = \frac{1}{1+z^2}$$

t-SNE on MNIST

https://distill.pub/2016/misread-tsne/

Pros & cons

PCA

- + Fast (probably fastest overal)
- + Requires no parameter tuning
- + Outputs transformation
- + The output space is meaningful
- + Survived test of time. Well understood
- - Linear

VAE

- - Still mostly experimental
- - Require lots of tuning
- - Only applied to ‚toy‘ datasets so far [1]

t-SNE

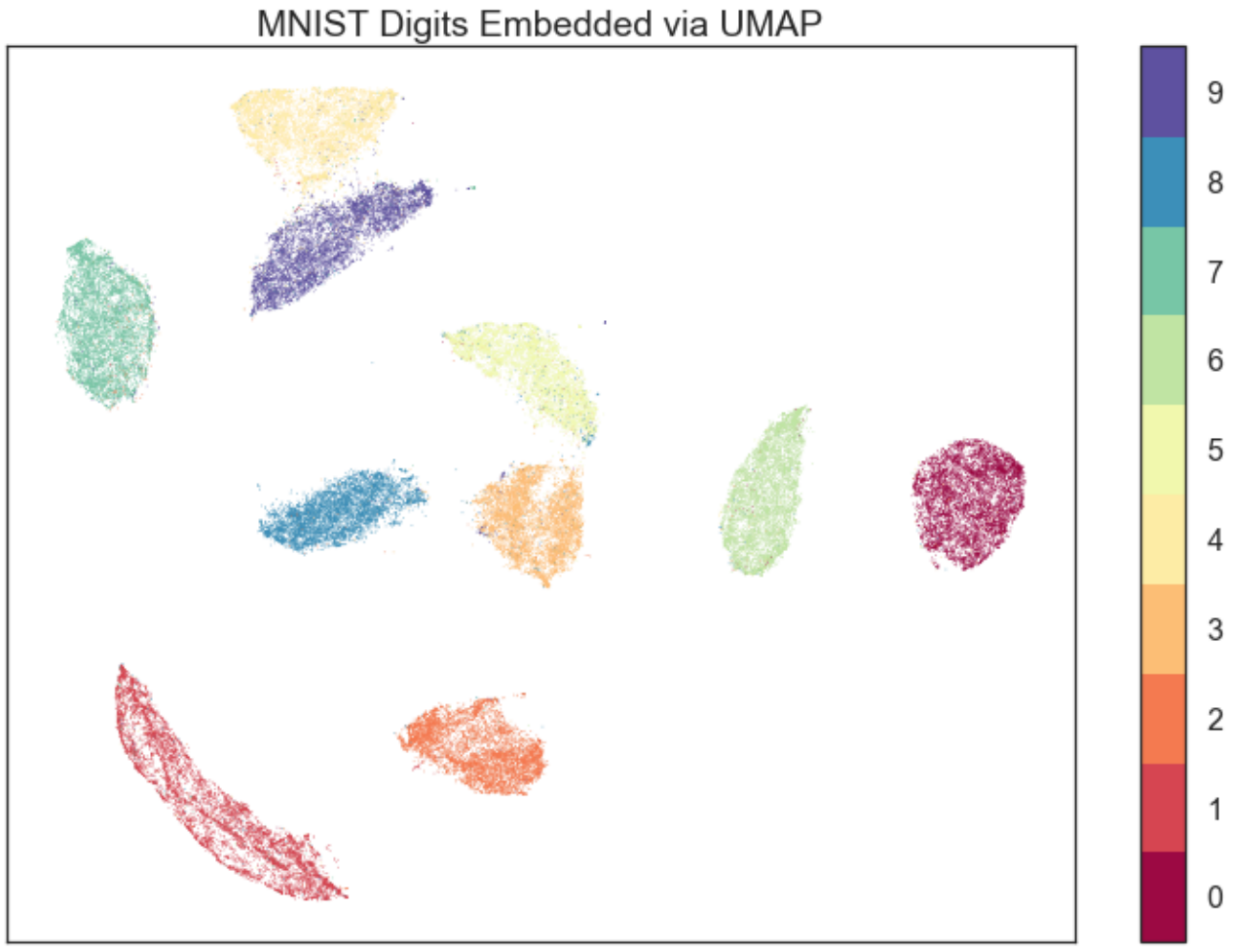
- + SOTA for visualization
- - Slow
- - Does not output transformation
- - Stochastic
- - Usually max 3 dimensions
- - Output space not meaningful [2]
- - Hyperparameters really matter

[1] https://umap-learn.readthedocs.io/en/latest/faq.html

[2] https://distill.pub/2016/misread-tsne/

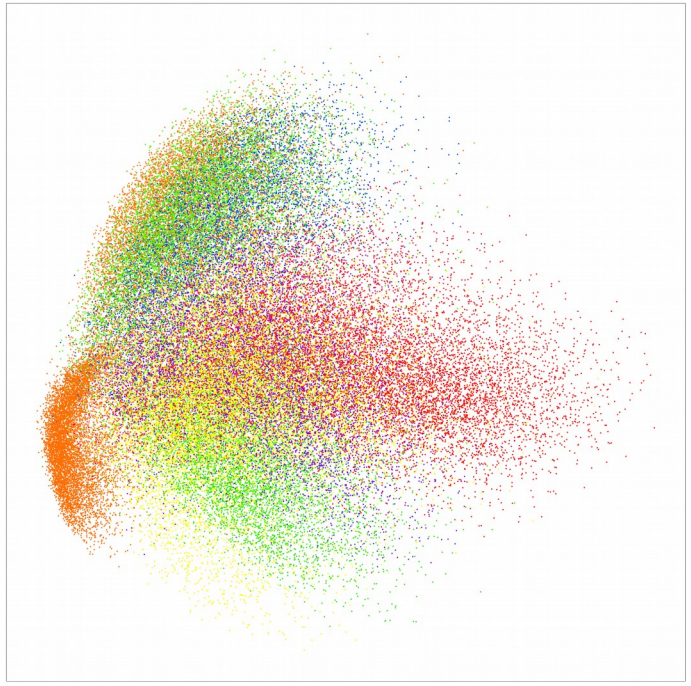
UMAP - Uniform Manifold Approximation and Projection

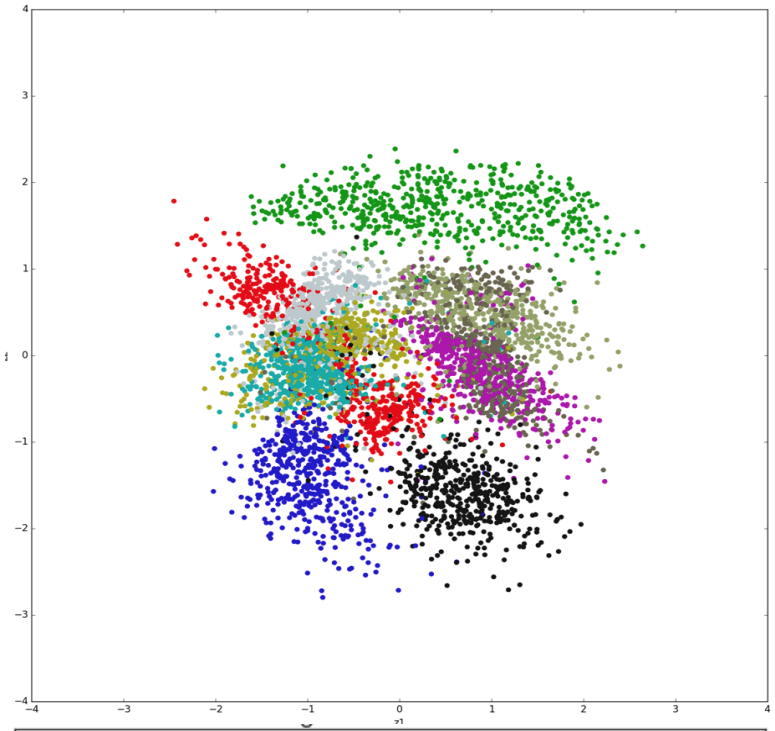
- + SOTA for visualisation
- + Scales well with embedding dimensions
- + Fast
- + Outputs transformation (but it has to remember input dataset…)
- + More meaningful output space than t-SNE, preserves more global structure
- + Strong theoretical foundation
- ? Manifold assumption
- - Slower than PCA

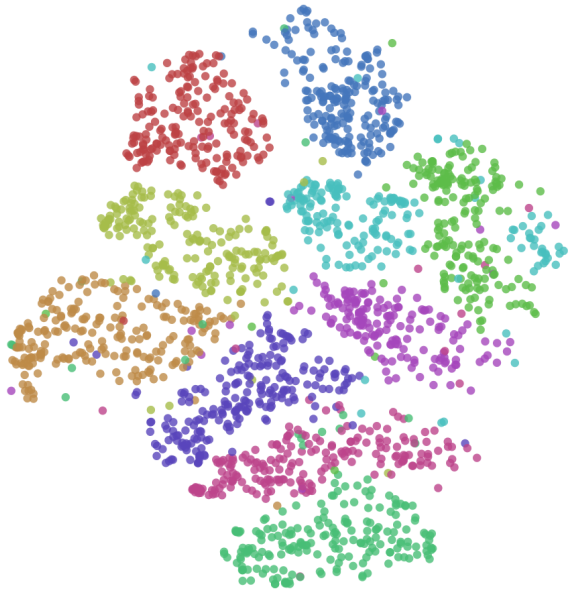


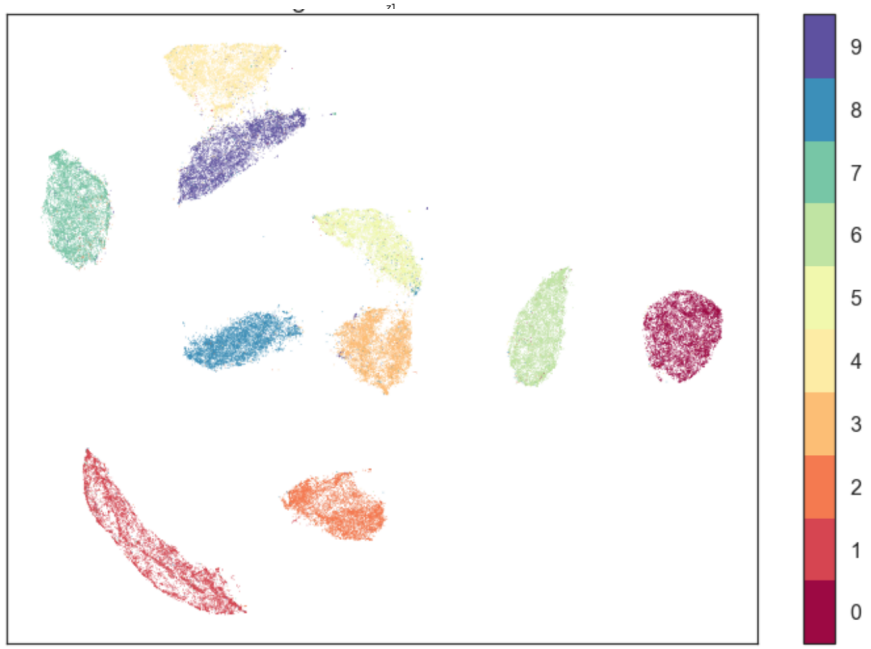
[1] https://github.com/lmcinnes/umap

https://arxiv.org/abs/1802.03426



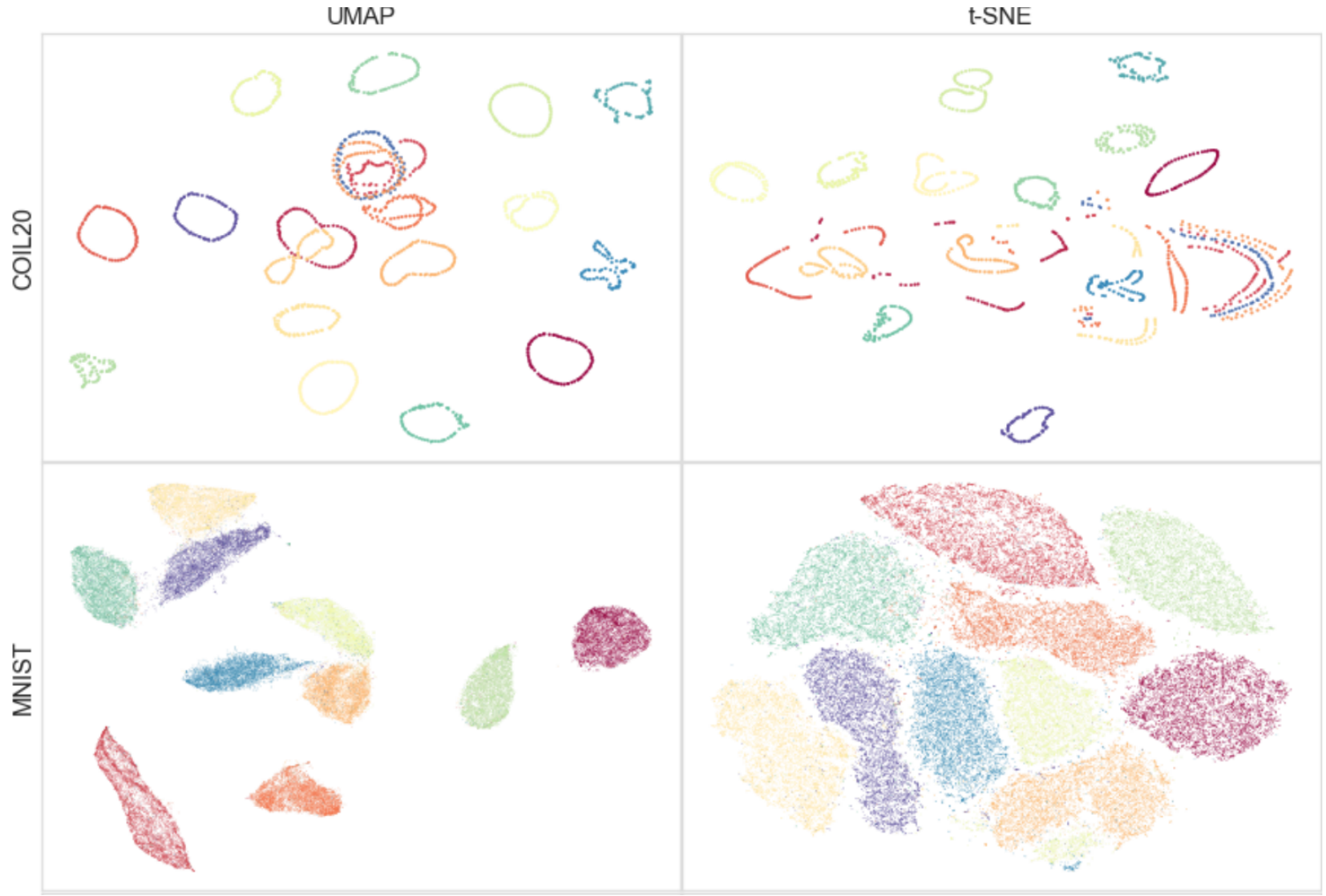

4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4





9
8
7
6
5
4
3
2
1
0

UMAP vs t-SNE



UMAP vs t-SNE

Performance

dataset	dataset size	t-SNE	UMAP
COIL20	1440x16384	20s	**7s**
COIL100	72000x49152	683s	**121s**
Shuttle	58000x9	741s	**140s**
MNIST	70000x784	1337s	**98s**
F-MNIST	70000x784	906s	**78s**
GoogleNews	200000x300	16214s	**821s**

https://arxiv.org/abs/1802.03426

Math background - topology

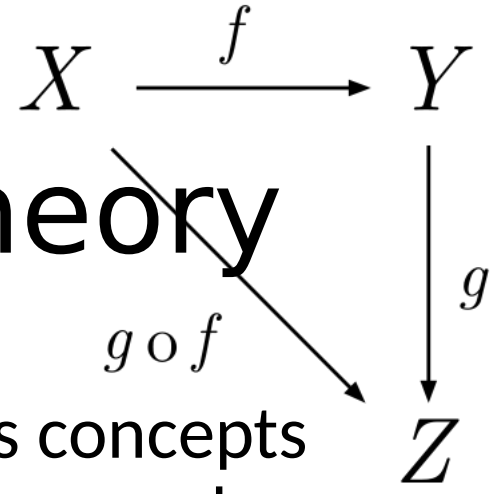
- **Topological space:** Defined as a set of points, along with a set of neighbourhoods for each point, satisfying a set of axioms relating points and neighbourhoods.
- **Homeomorphism:** Is a continuous function between topological spaces that has a continuous inverse function.
- **Manifold:** Is a topological space that locally resembles Euclidean space near each point.
- **Cover:** Cover of a set X is a collection of sets whose union [illegible] X
- Two functions are **homotopic** if one can be „continuously deformed“ into the other

Math background: Homotopic equivalence

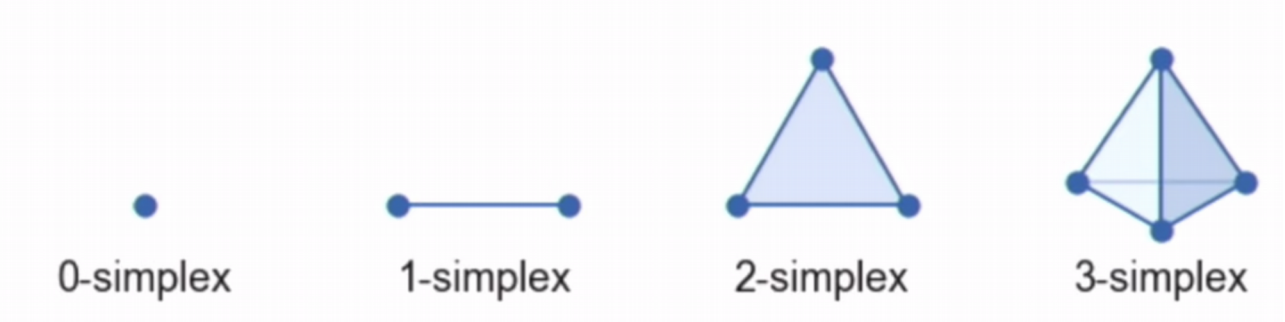
- Given two spaces X and Y, we say they are **homotopy equivalent** if there exist continuous maps $f : X \to Y$ and $g : Y \to X$ such that $g \circ f$ is homotopic to the identity map idX and $f \circ g$ is homotopic to idY

- In short: Every homeomorphism is homotopy equivalence but not the other way around.

Math background – category theory



- **Category theory** formalizes mathematical structure and its concepts in terms of a labeled directed graph called a **category**, whose nodes are called objects, and whose labelled directed edges are called arrows (or morphisms).
- **Functor** is a map between categories.
- A **simplicial set [1]** is a categorical model capturing those topological spaces that can be built up from simplices and their incidence relations.



[1] https://arxiv.org/pdf/0809.4221.pdf

UMAP - assumptions

1. The data is uniformly distributed on a Riemannian manifold;
2. The Riemannian metric is locally constant (or can be approximated as such);
3. The manifold is locally connected (can‘t have a point that is completely separated from everything else).

UMAP - overview

- „UMAP uses **local manifold approximations** and patches together their local **fuzzy simplicial set representations**. This constructs a topological representation of the high dimensional data. Given a low dimensional representation of the data, a similar process can be used to construct an equivalent topological representation. UMAP then optimizes the layout of the data representation in the low dimensional space, minimizing the cross-entropy between the two topological representations“.

[1] https://arxiv.org/abs/1802.03426

UMAP – construction of high-dimensional fuzzy topological representation

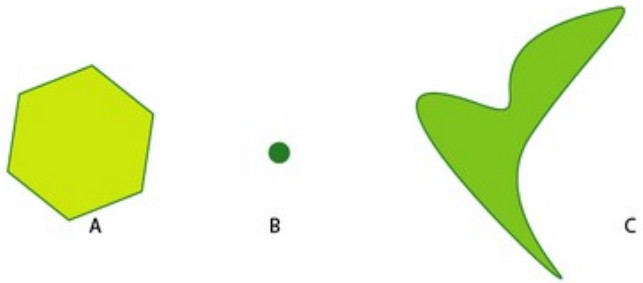
1. Approximate manifold on which the data is supposed to lie
2. Construct a fuzzy simplicial set representation of the approximated manifold (to have continuous distance. We will see in a moment…)

Nerve theorem

Theorem 1 (Nerve theorem). *Let $\mathcal{U} = \{U_i\}_{i \in I}$ be a cover of a topological space X. If, for all $\sigma \subset I$ $\bigcap_{i \in \sigma} U_i$ is either contractible or empty, then $\mathcal{N}(\mathcal{U})$ is homtopically equivalent to X.*

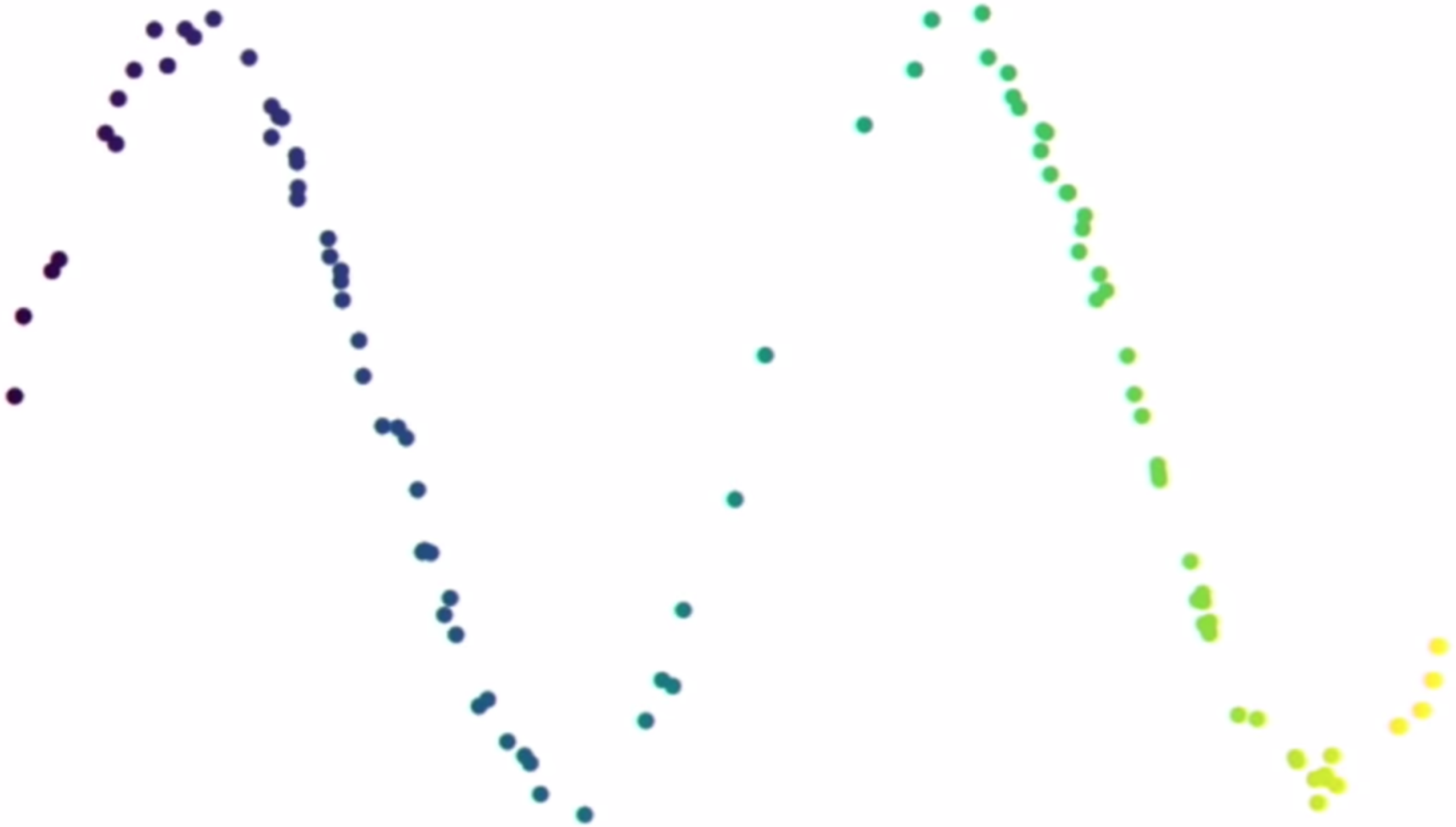
Nerve

finite set $J \subseteq I$ belongs to N if and only if the intersection of the U_i whose subindices are in J is non-empty,



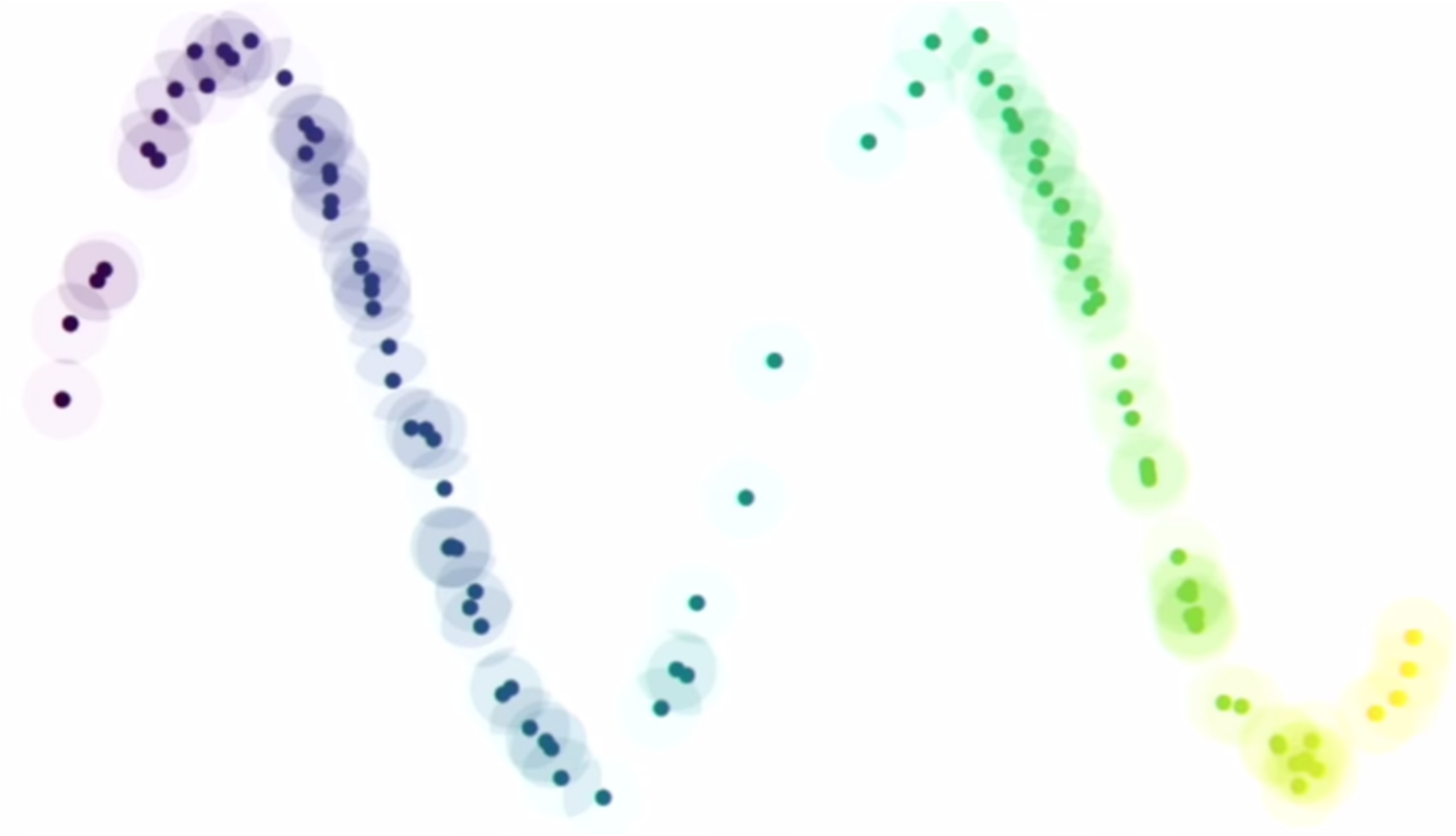
Consequence: If we build a simplificial complex from Points (uniformly distributed) in a certain way, we can actually recover all the important topology of the original space.

Example

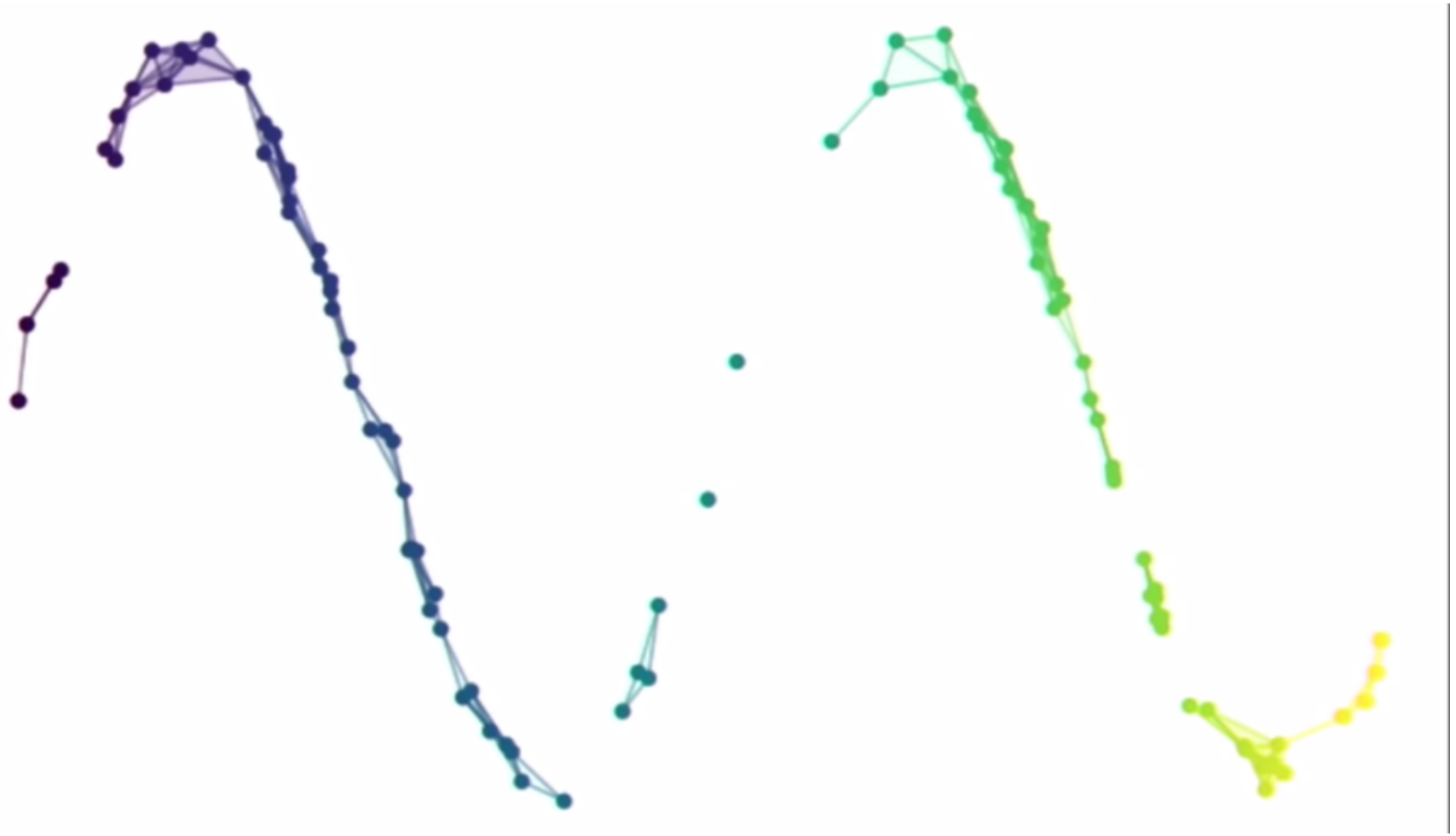


https://www.youtube.com/watch?v=nq6iPZVUxZU

Example



Example

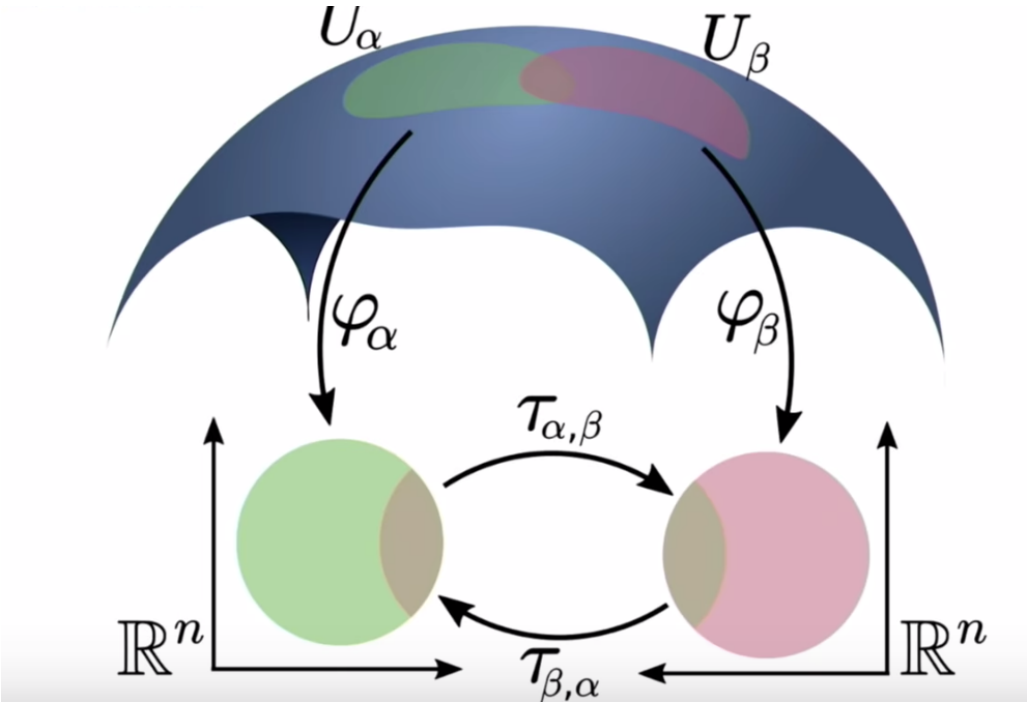


Nice but we haven‘t exactly recovered the topology. What went wrong? The data are not uniformly distributed on the manifold

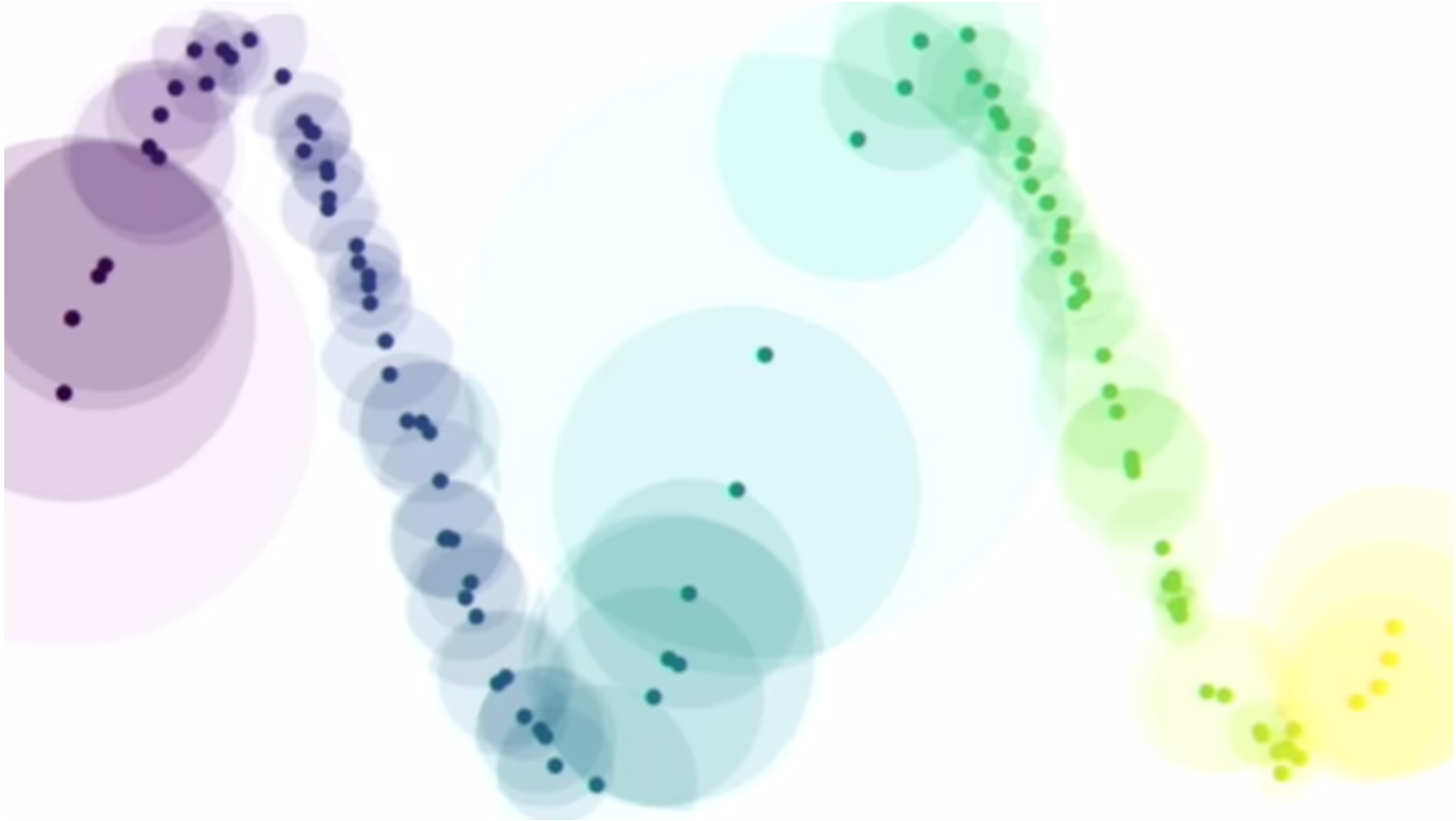
This would work perfectly

Time to use our assumptions

- Riemannian manifold.
- There are patches of data that are uniformly distributed. In each of these patches we introduce a different notion of distance



Result



Why fixed radius? Why not fuzzy cover?

Theorem 2 (UMAP Adjunction). *The functors* FinReal : **sFuzz** → **FinEPMet** *and* FinSing : **FinEPMet** → **sFuzz** *form an adjunction* FinReal ⊣ FinSing.

- **sFuzz**: Category of fuzzy simplicial sets
- **FinEPMet:** Category of finite extended pseudo-metric spaces

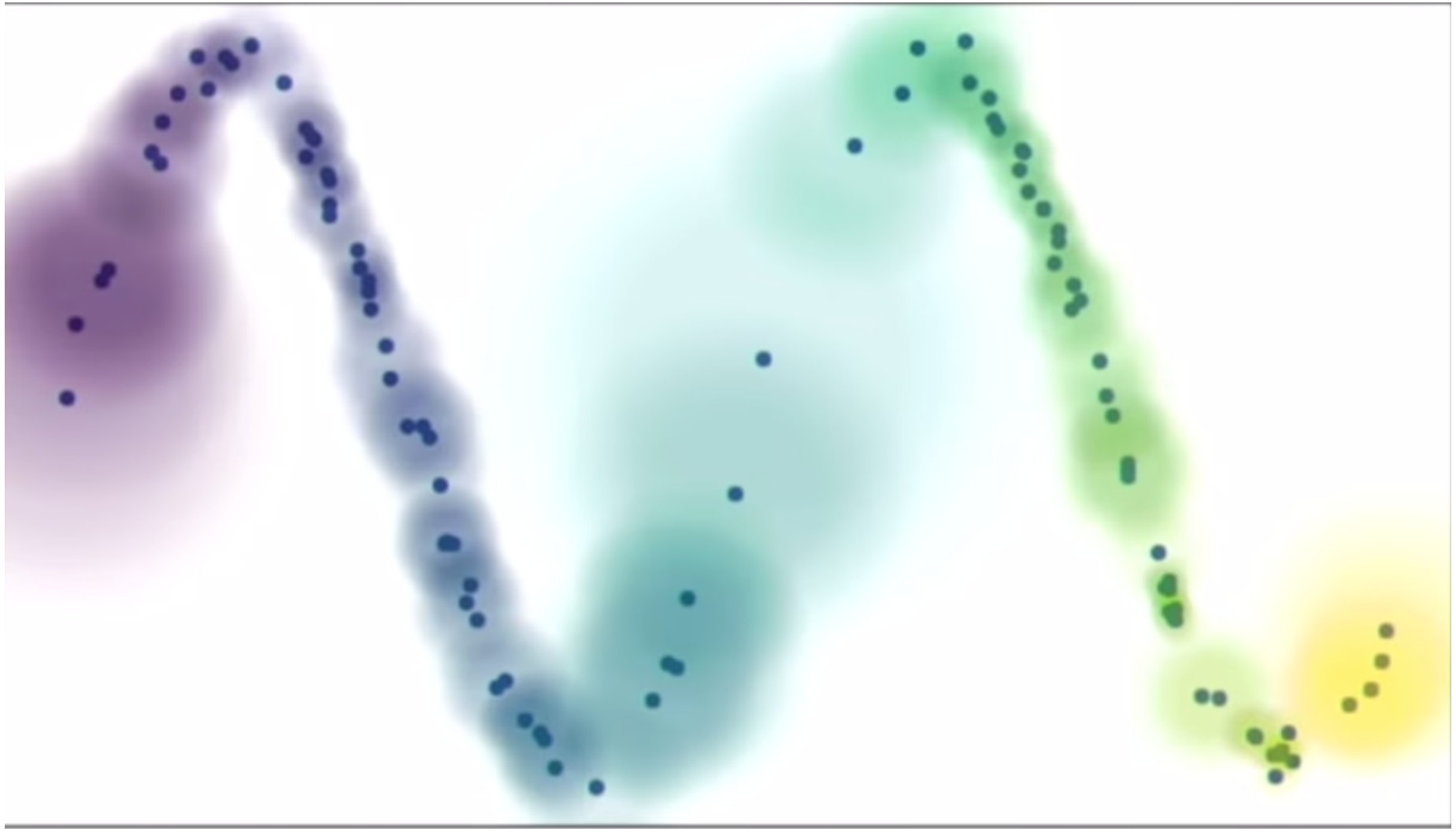
„Extended metric spaces are a cute way of working with topological spaces whose connected components are ordinary metric spaces.“ [1]

„An adjunction between categories C and D is somewhat akin to a "weak form" of an equivalence between C and D, and indeed every equivalence is an adjunction.“ [2]

[1] https://math.stackexchange.com/questions/1964378/differences-between-extended-metric-space-and-metric-space

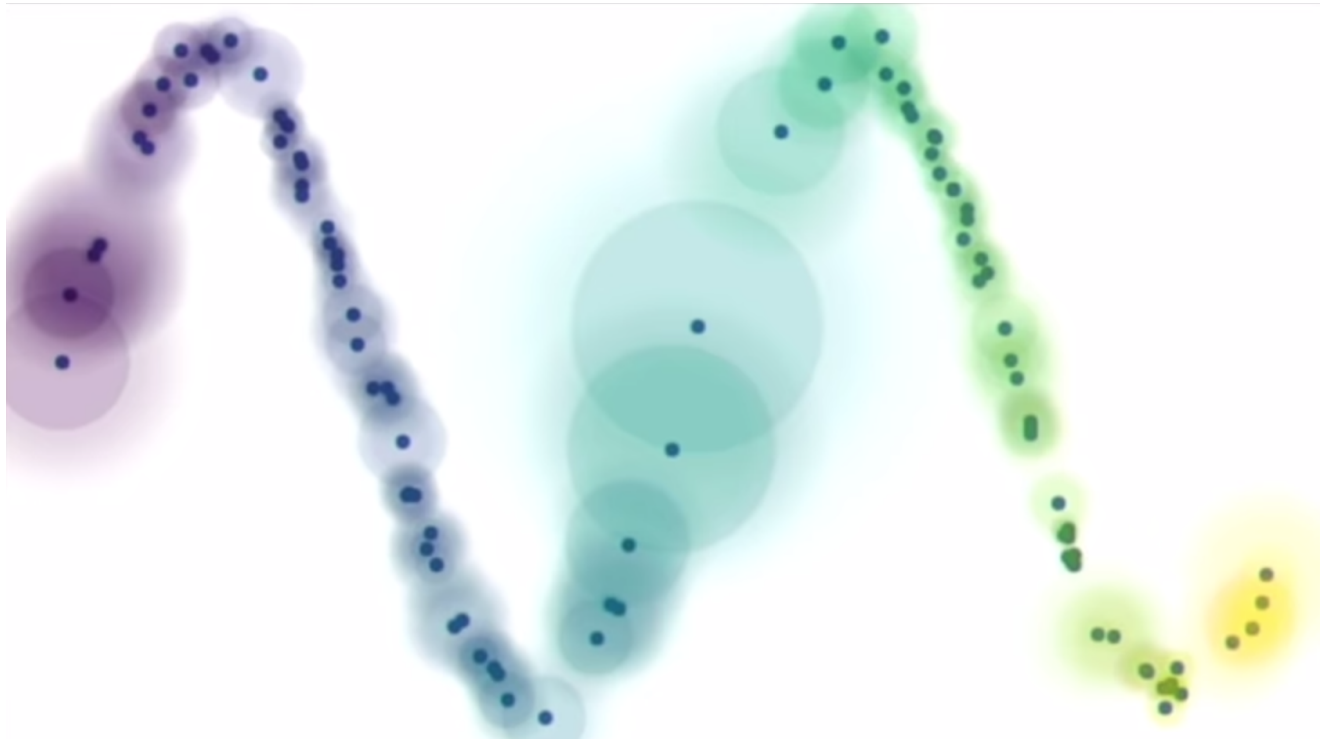
[2] https://www.wikiwand.com/en/Adjoint_functors

Fuzzy cover



Fuzzy cover – manifold locally connected

(we always have 1 neighbour in the uniformly distributed neighbourhood)

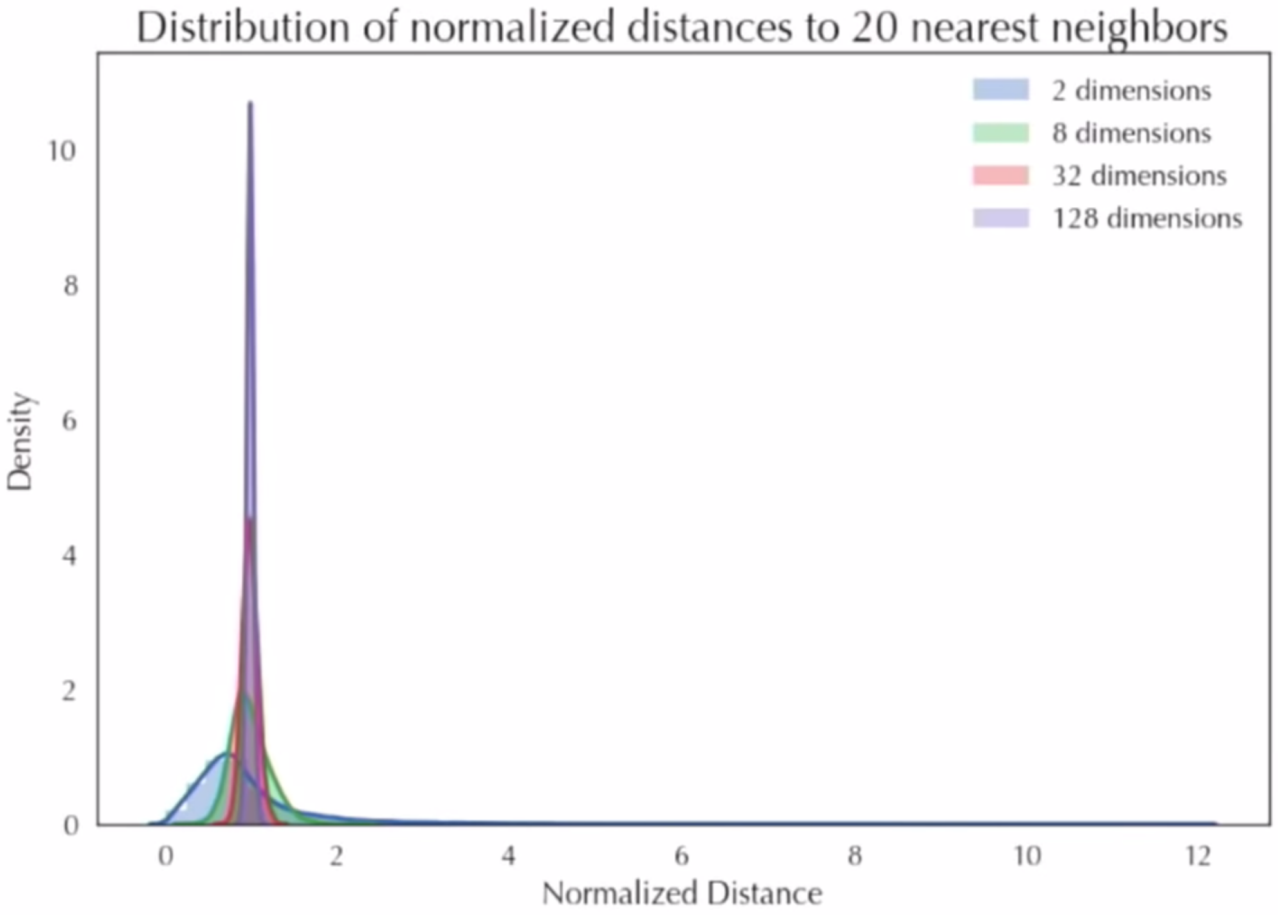


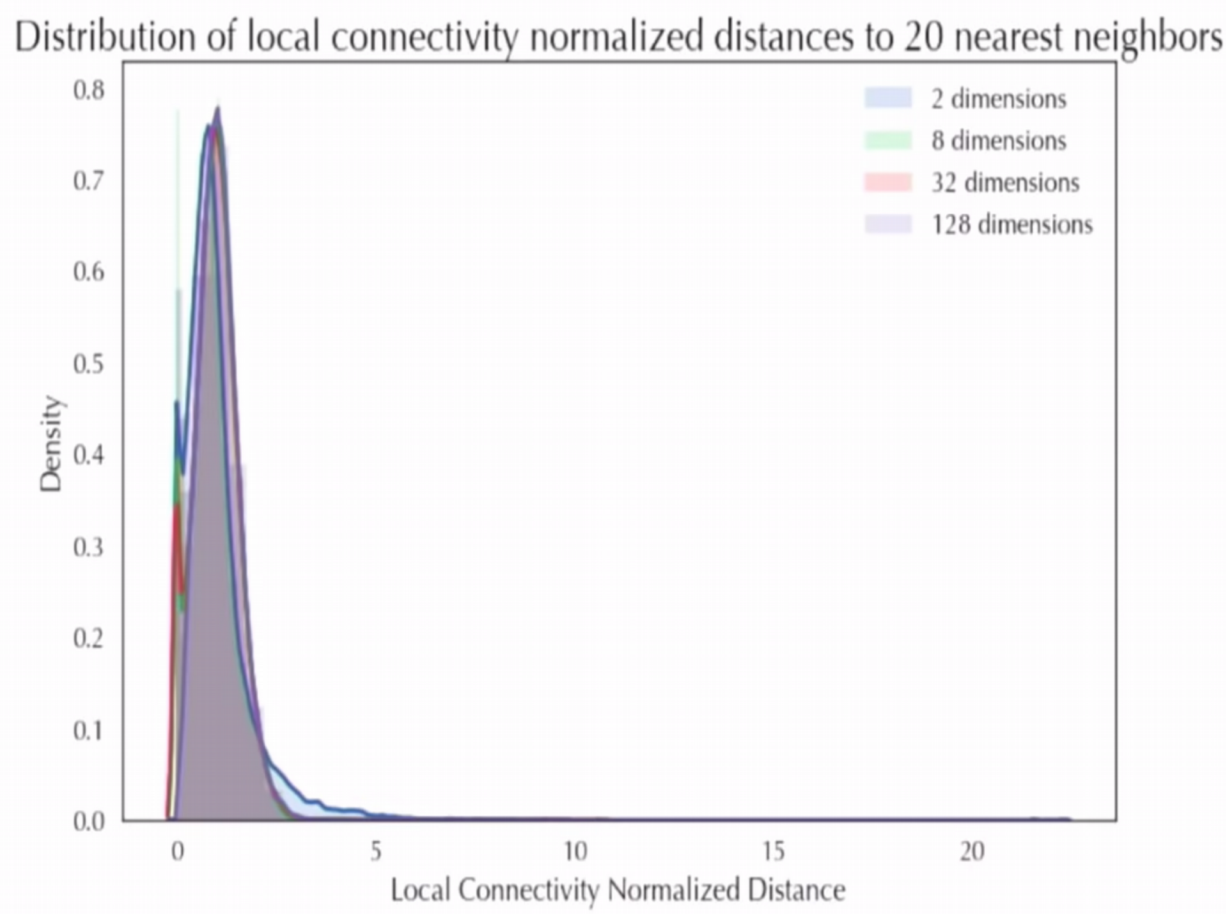
We define (in neighbourhood of C):

- Distance between A and B in neighbourhood of C is infinite.
- Distance from C to nearest neighbour is 0.
- Other distances from C are geodesic distance beyond the first neighbor

Geodesic distance is approximated via Lemma 1
In the UMAP paper [1]

This adresses the curse of dimensionality

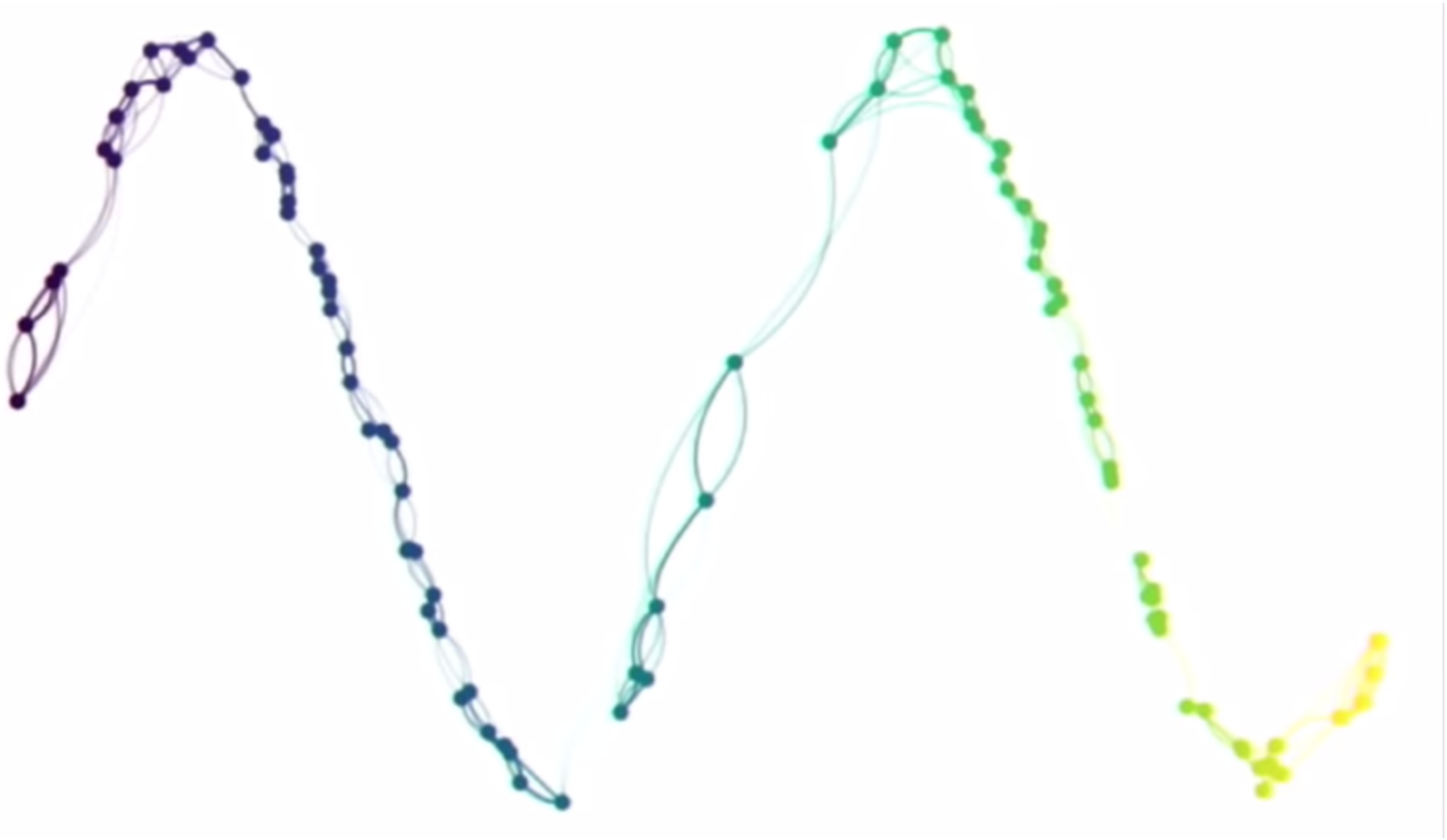




In high dimensions everything is the same distance away

Distributions are dimension invariant ⏪

Nerve - multigraph



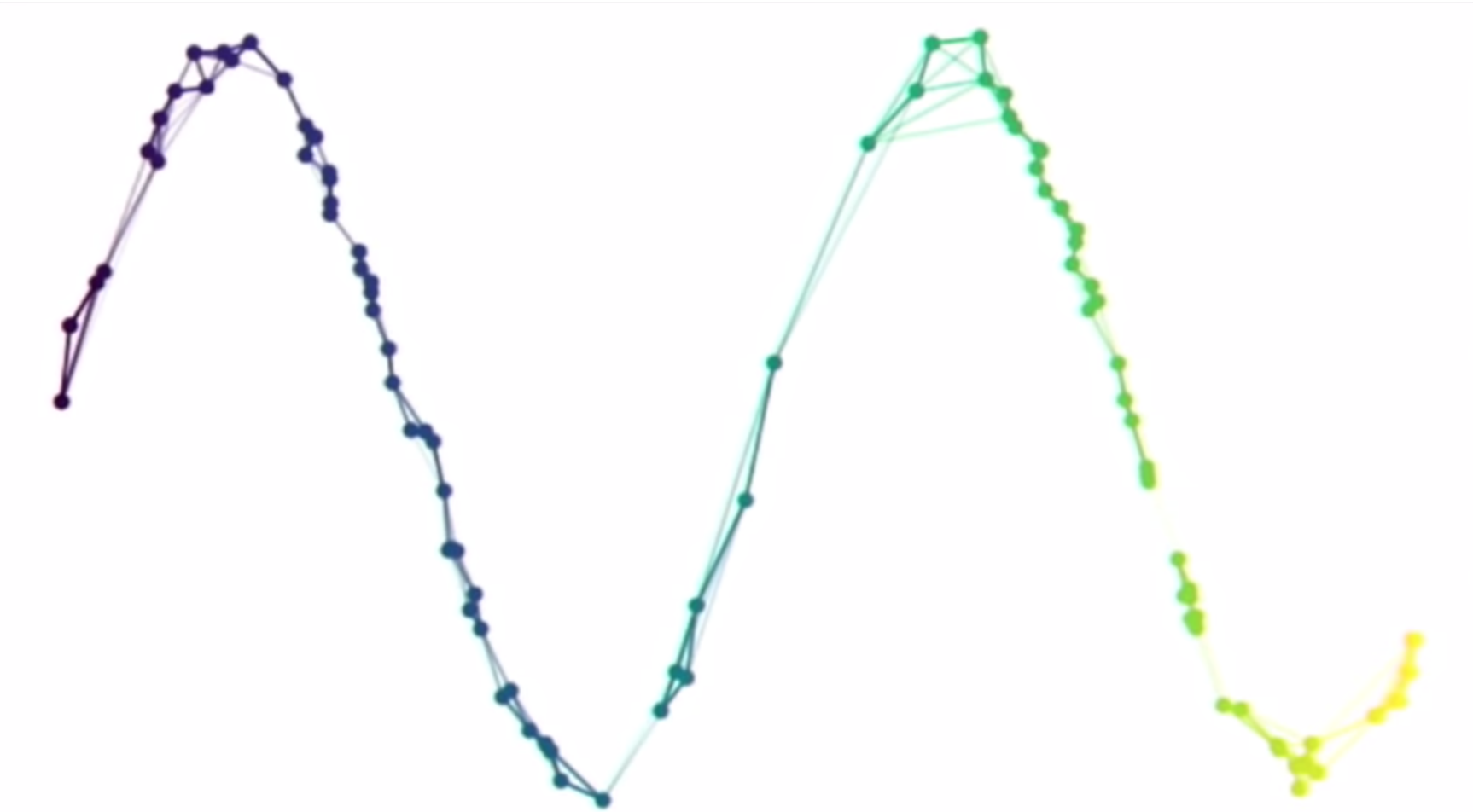
How to glue different metrics together?

Theorem 2 (UMAP Adjunction)**.** *The functors* FinReal : **sFuzz** → **FinEPMet** *and* FinSing : **FinEPMet** → **sFuzz** *form an adjunction* FinReal ⊣ FinSing.

We are working in the FinEPMet category now (each point has its space). But since it is „equivalent“ to sFuzz we can use theorem about fuzzy sets to glue them together:

$$f(\alpha, \beta) = \alpha + \beta - \alpha \cdot \beta$$

Result



Low dimensional representation

- We want to embed into R^d
- We know the manifold => R^d
- We know the distance metric
- Hyperparameter: expected dist between nearest neighbours

We can now construct a fuzzy simplicial set in R^d

Graph => fuzzy set of edges

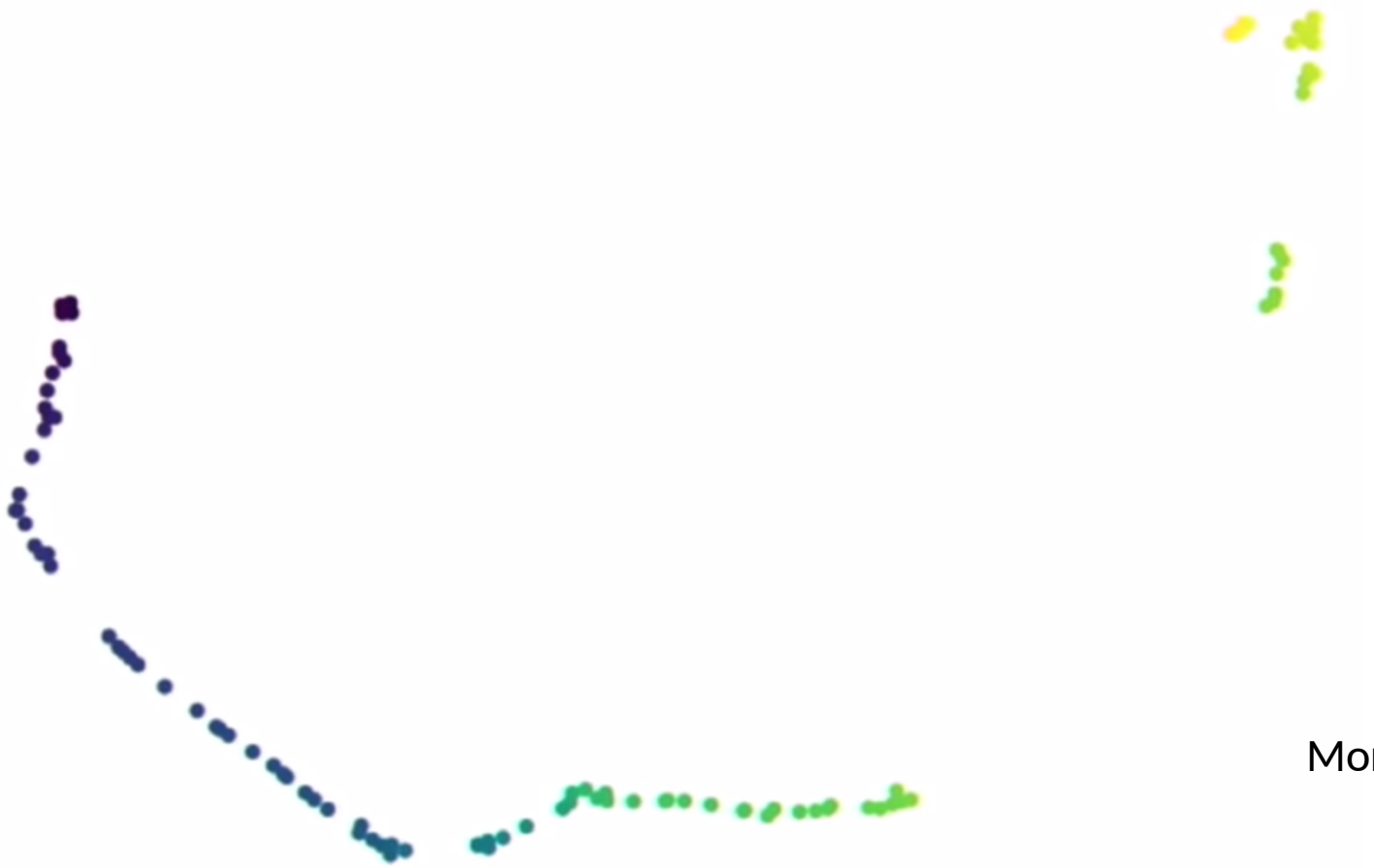
membership strength function $\mu : A \to [0, 1]$.

Definition 10. *The cross entropy C of two fuzzy sets (A, μ) and (A, ν) is defined as*

$$C((A,\mu),(A,\nu)) \triangleq \sum_{a \in A} \mu(a) \log \left(\frac{\mu(a)}{\nu(a)} \right) + (1 - \mu(a)) \log \left(\frac{1 - \mu(a)}{1 - \nu(a)} \right).$$

Optimize via stochastic gradient descent

Result - gaps due to locally dense areas



More data should fix this

Implementation details

- k-nearest neighbour effectively computed with Nearest-Neighbor-Descent algorithm [1]
- Optimizer utilizes probabilistic edge sampling [2] and negative sampling [3]

[1] Wei Dong, Charikar Moses, and Kai Li. E·cient k-nearest neighbor graph construction for generic similarity measures. In Proceedings of the 20th International Conference on World Wide Web, WWW '11, pages 577–586, New York, NY, USA, 2011. ACM.

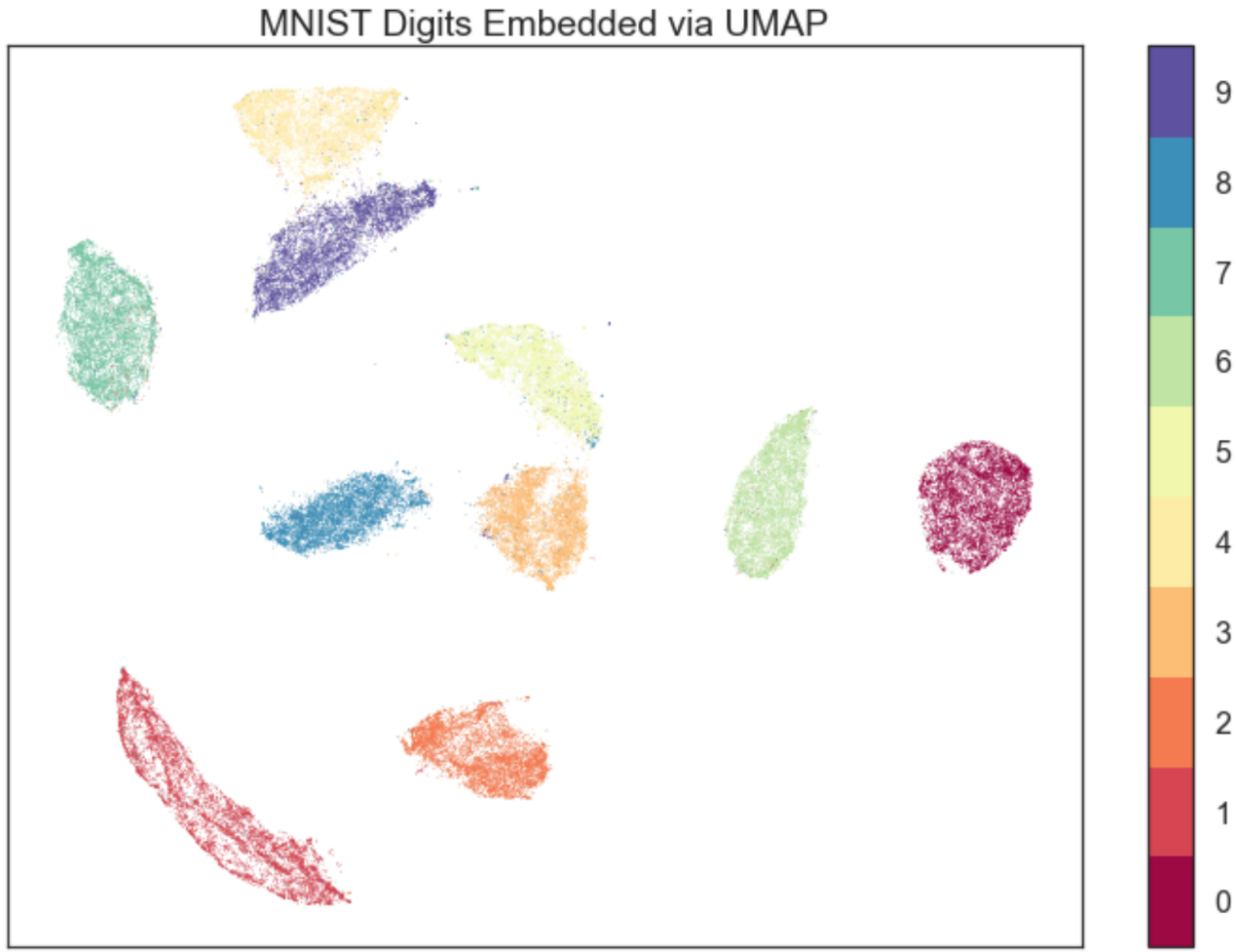
[2] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing largescale and high-dimensional data. In Proceedings of the 25th International Conference on World Wide Web, pages 287–297. International World Wide Web Conferences Steering Commi□ee, 2016.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Je– Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.

Hyperparameters

- n_neighbors: This determines the number of neighboring points used in local approximations of manifold structure. Larger values will result in more global structure being preserved at the loss of detailed local structure. In general this parameter should often be in the range 5 to 50, with a choice of 10 to 15 being a sensible default.
- min_dist: This controls how tightly the embedding is allowed compress points together. Larger values ensure embedded points are more evenly distributed, while smaller values allow the algorithm to optimise more accurately with regard to local structure. Sensible values are in the range 0.001 to 0.5, with 0.1 being a reasonable default.
- metric: This determines the choice of metric used to measure distance in the input space. A wide variety of metrics are already coded, and a user defined function can be passed as long as it has been JITd by numba.

https://github.com/lmcinnes/umap


MNIST Digits Embedded via UMAP
9
8
7
6
5
4
3
2
1
0