INSTITUT TUTTE INSTITUTE

UMAP

Uniform Manifold Approximation and Projection for dimension reduction

INSTITUT TUTTE INSTITUTE

Who am I?

I am a research mathematician at the Tutte Institute for Mathematics and Computing

My Ph.D. was in Profinite Lie Rings
(no, you don't care)

I now work on applying topological techniques to unsupervised learning problems

INSTITUT TUTTE INSTITUTE

What is Dimension Reduction?

INSTITUT TUTTE INSTITUTE

Find the “latent” features in your data

INSTITUT TUTTE INSTITUTE

INSTITUT
TUTTE
INSTITUTE

INSTITUT TUTTE INSTITUTE

Matrix Factorization

Neighbour Graphs

INSTITUT TUTTE INSTITUTE

Matrix Factorization

Principal Component Analysis

Linear Autoencoder

Latent Dirichlet Allocation

Non-negative Matrix Factorization

Generalised Low Rank Models

Word2Vec

GloVe

INSTITUT TUTTE INSTITUTE

Neighbour Graphs

Laplacian Eigenmaps

Hessian Eigenmaps

Local Tangent Space Alignment

JSE

Isomap

t-SNE

Locally Linear Embedding

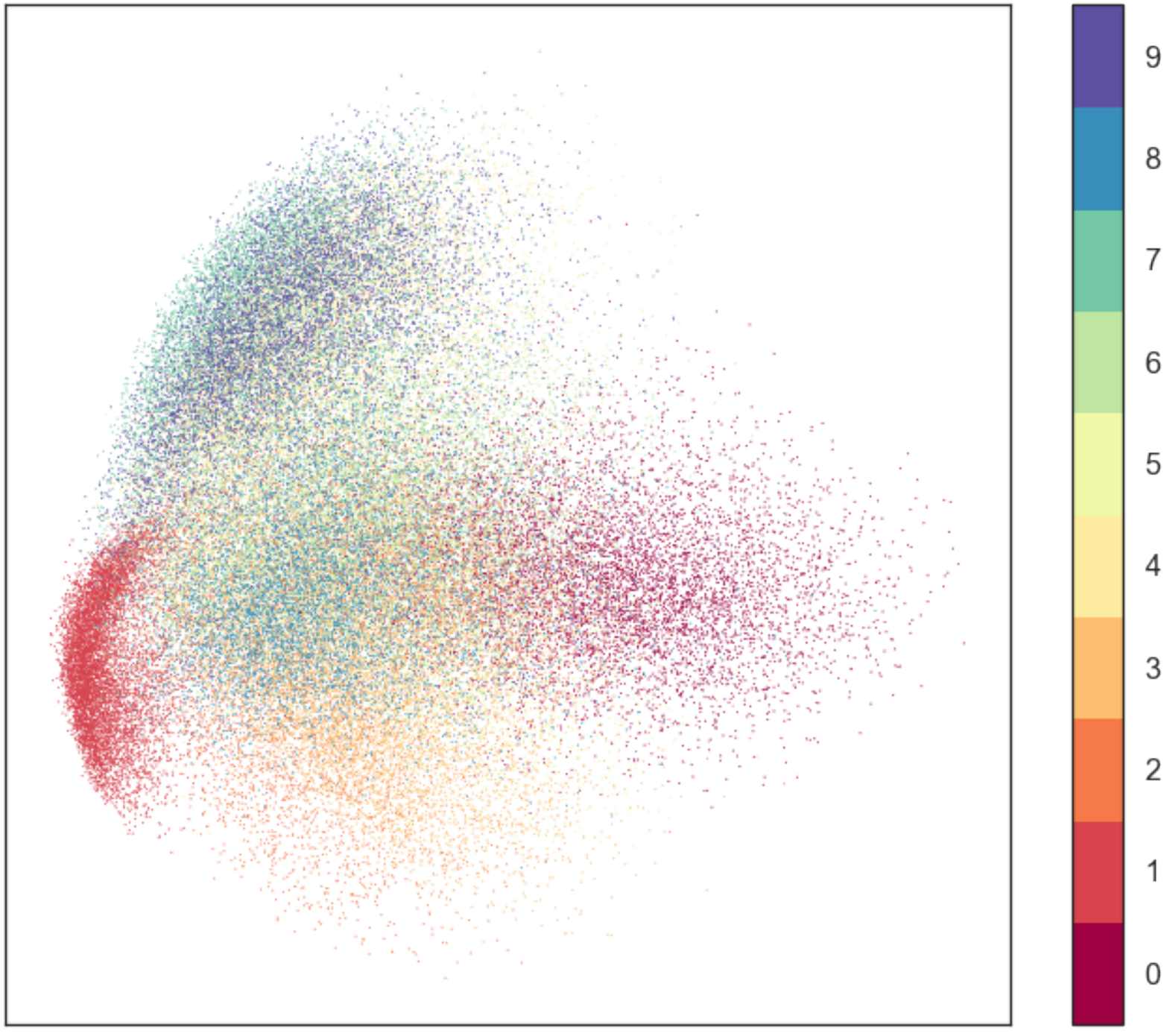
UMAP

INSTITUT TUTTE INSTITUTE

PCA is the prototypical matrix factorization

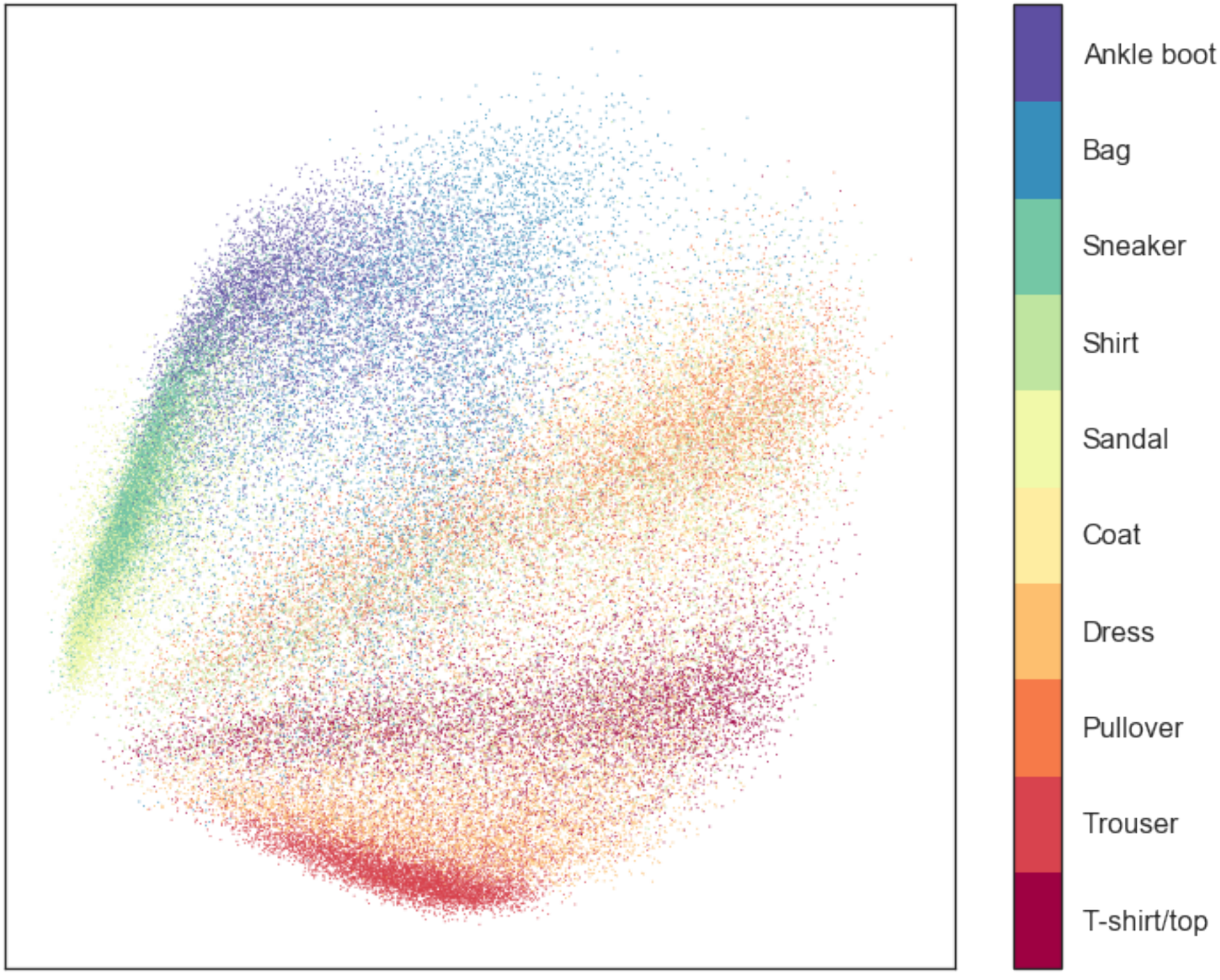
INSTITUT TUTTE INSTITUTE

PCA on MNIST digits



INSTITUT TUTTE INSTITUTE

PCA on Fashion MNIST

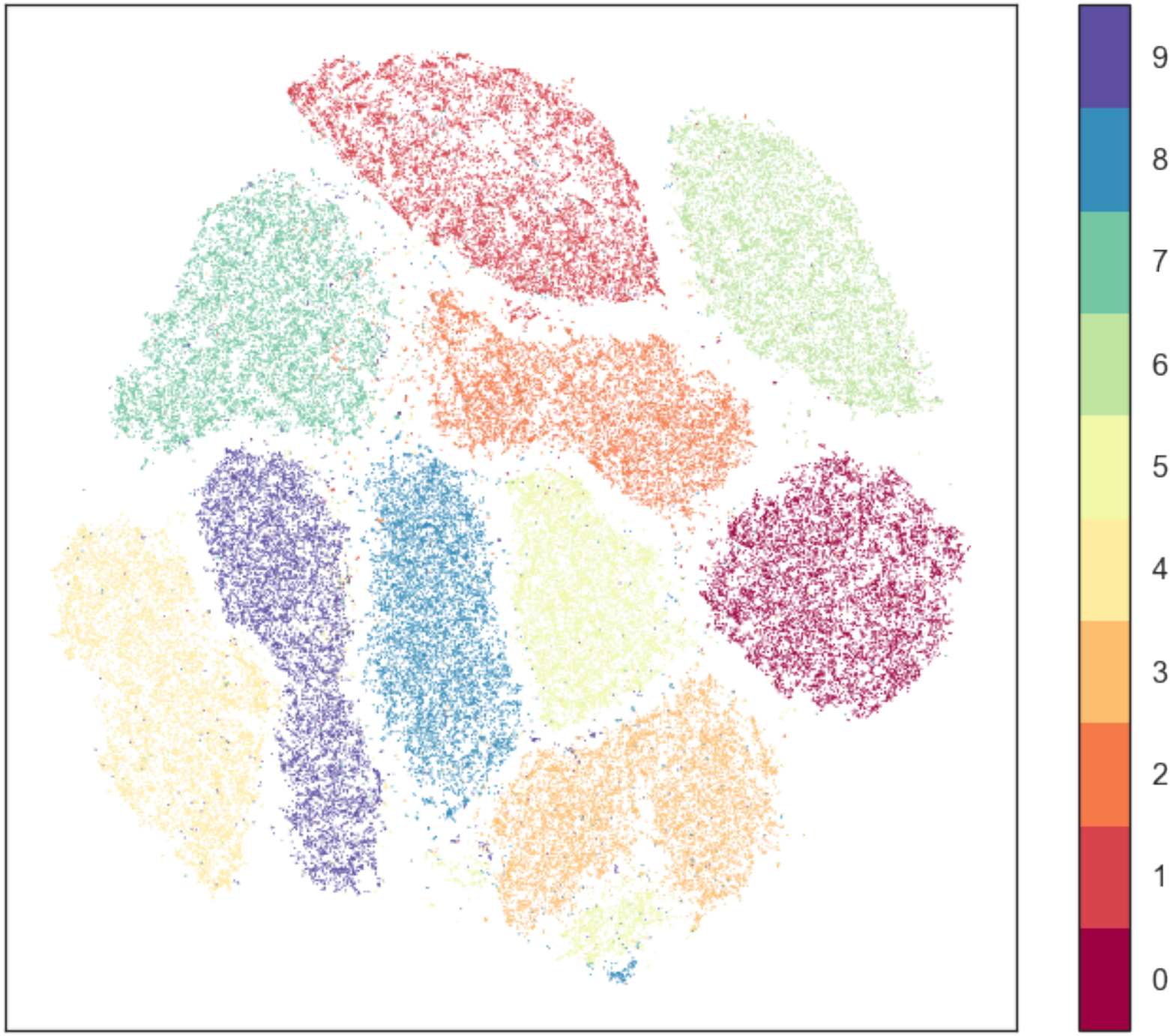


INSTITUT TUTTE INSTITUTE

t-SNE is the current state-of-the art for neighbour graphs

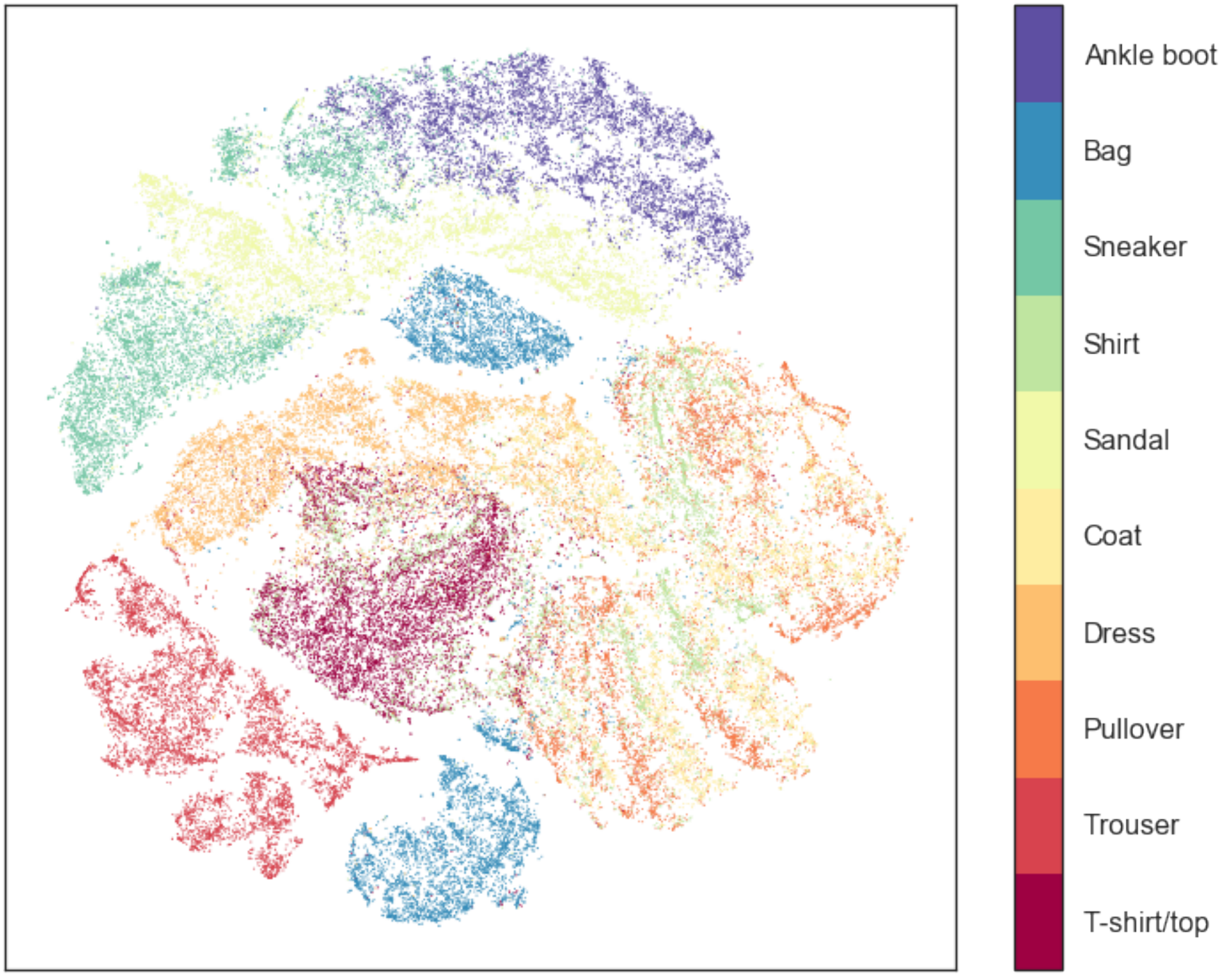
INSTITUT TUTTE INSTITUTE

t-SNE on MNIST digits



INSTITUT TUTTE INSTITUTE

t-SNE on Fashion MNIST



INSTITUT TUTTE INSTITUTE

Uniform Manifold Approximation and Projection

INSTITUT TUTTE INSTITUTE

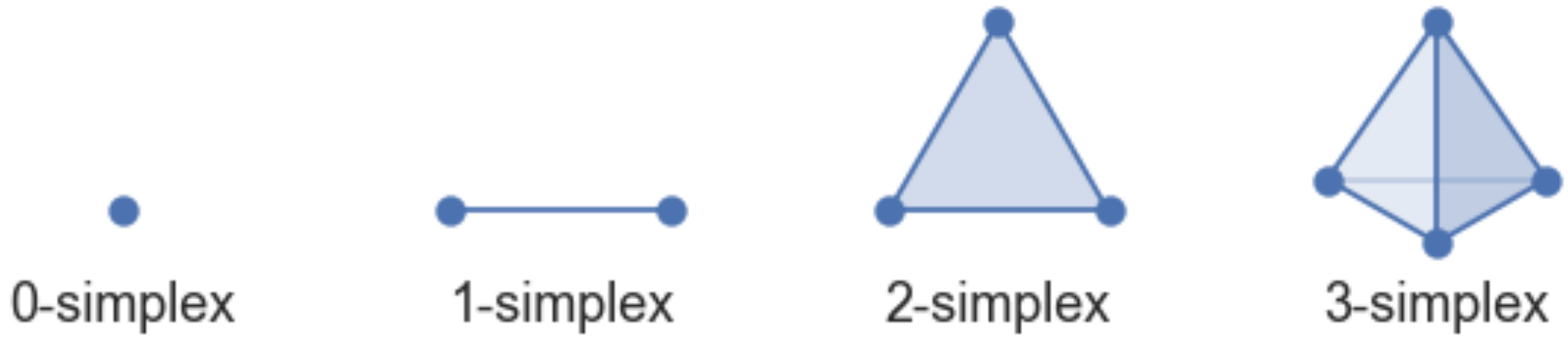
UMAP builds mathematical theory to justify the graph based approach

INSTITUT TUTTE INSTITUTE

First, a little bit of topological data analysis...

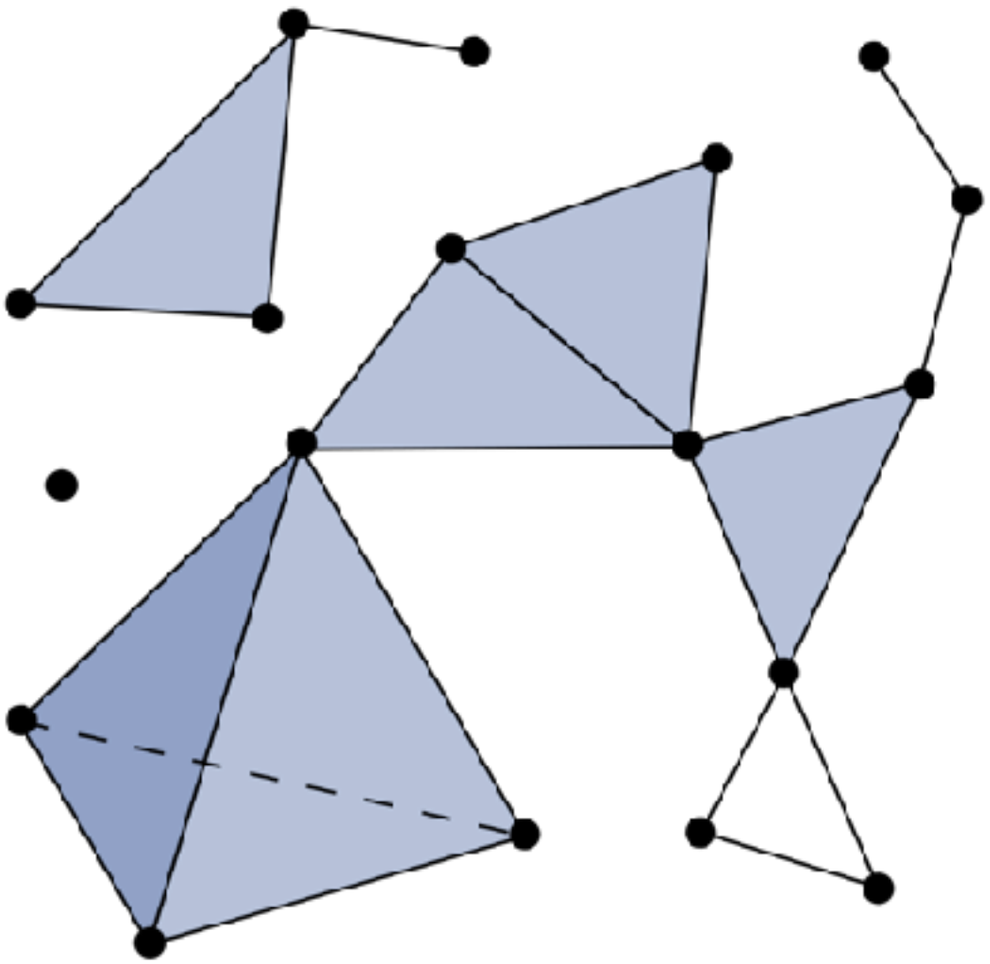
INSTITUT TUTTE INSTITUTE

Simplices

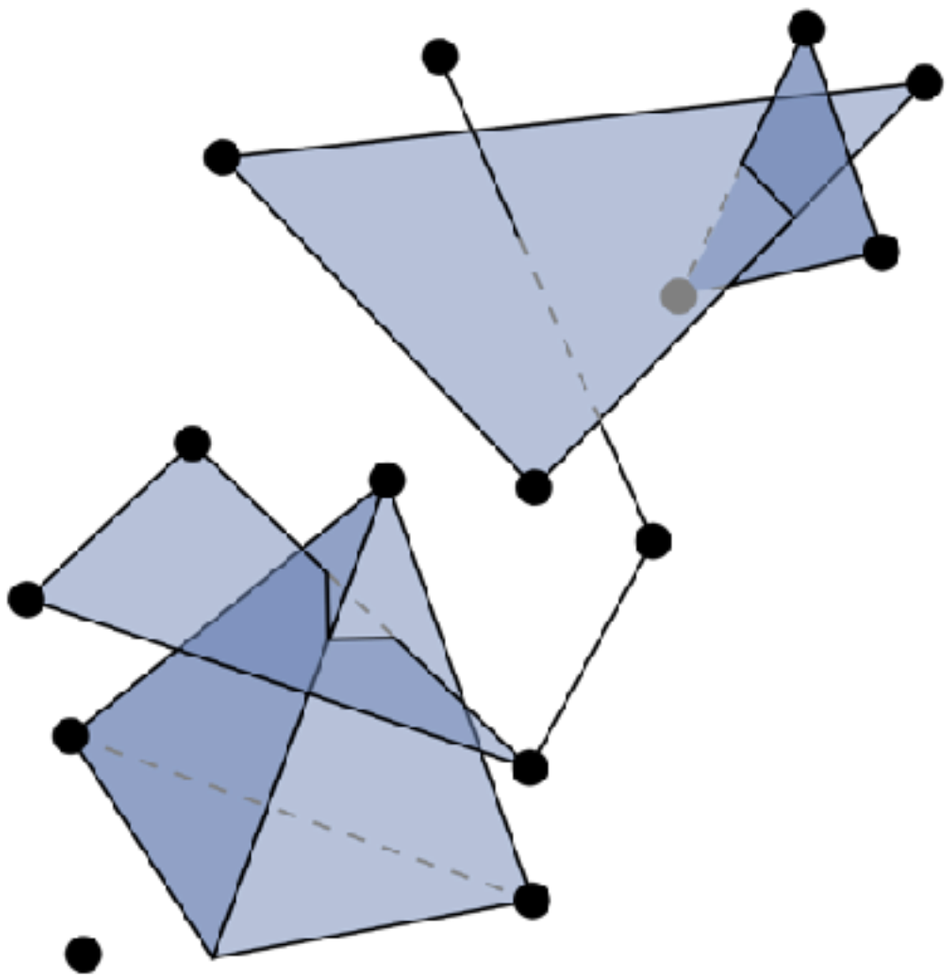


INSTITUT TUTTE INSTITUTE

Simplicial complex



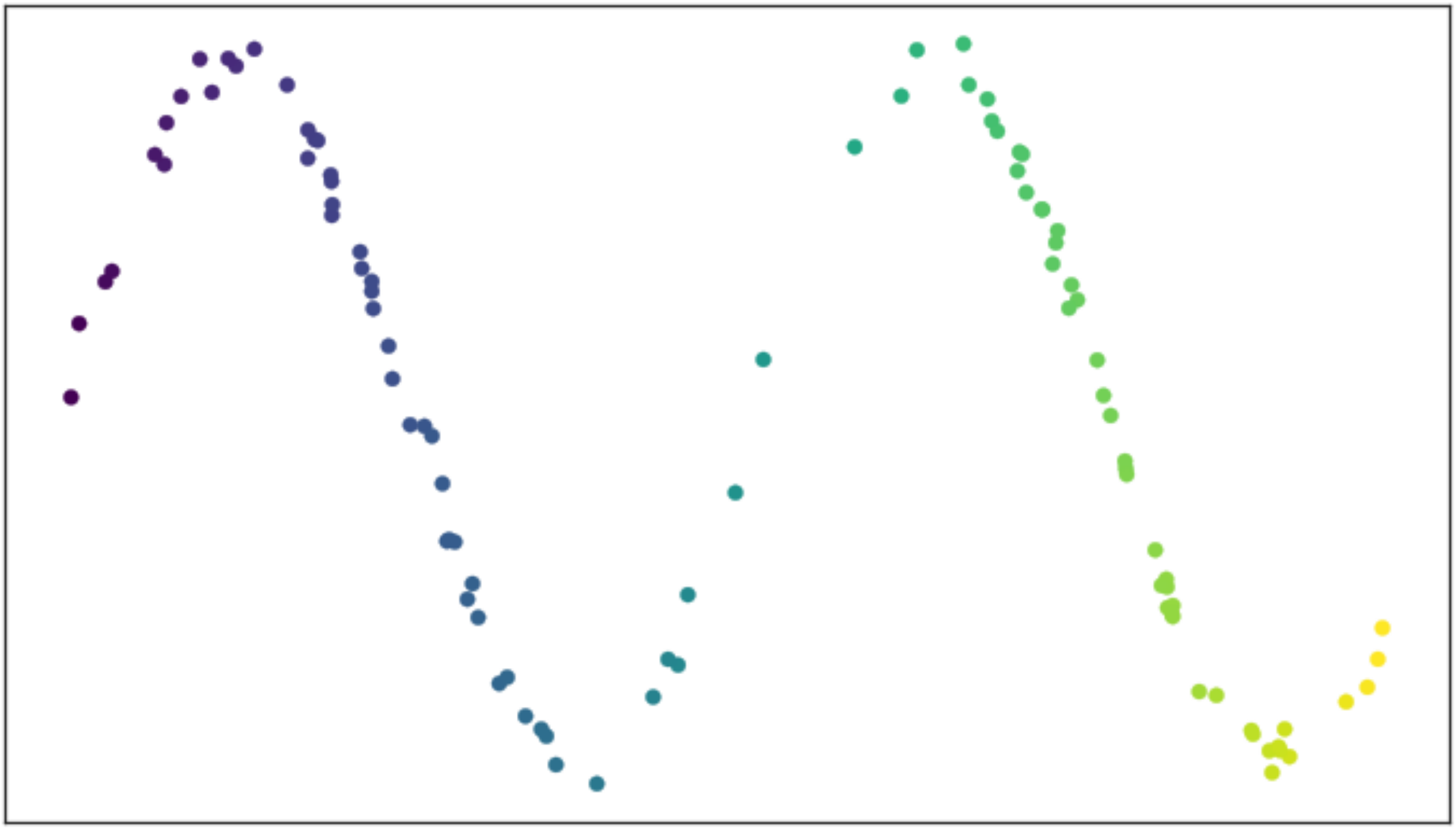
Collection of Simplices



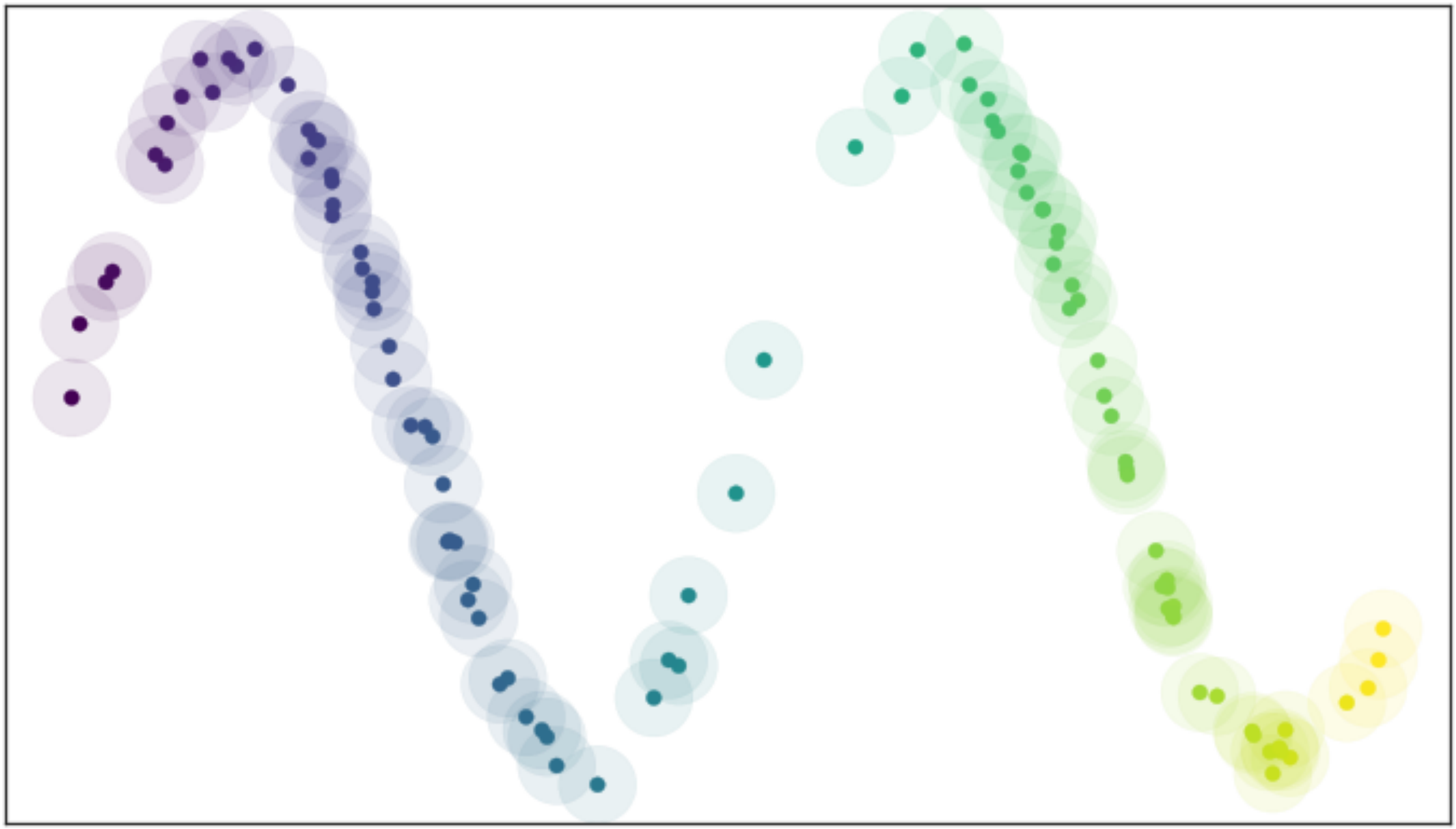
INSTITUT TUTTE INSTITUTE

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.*

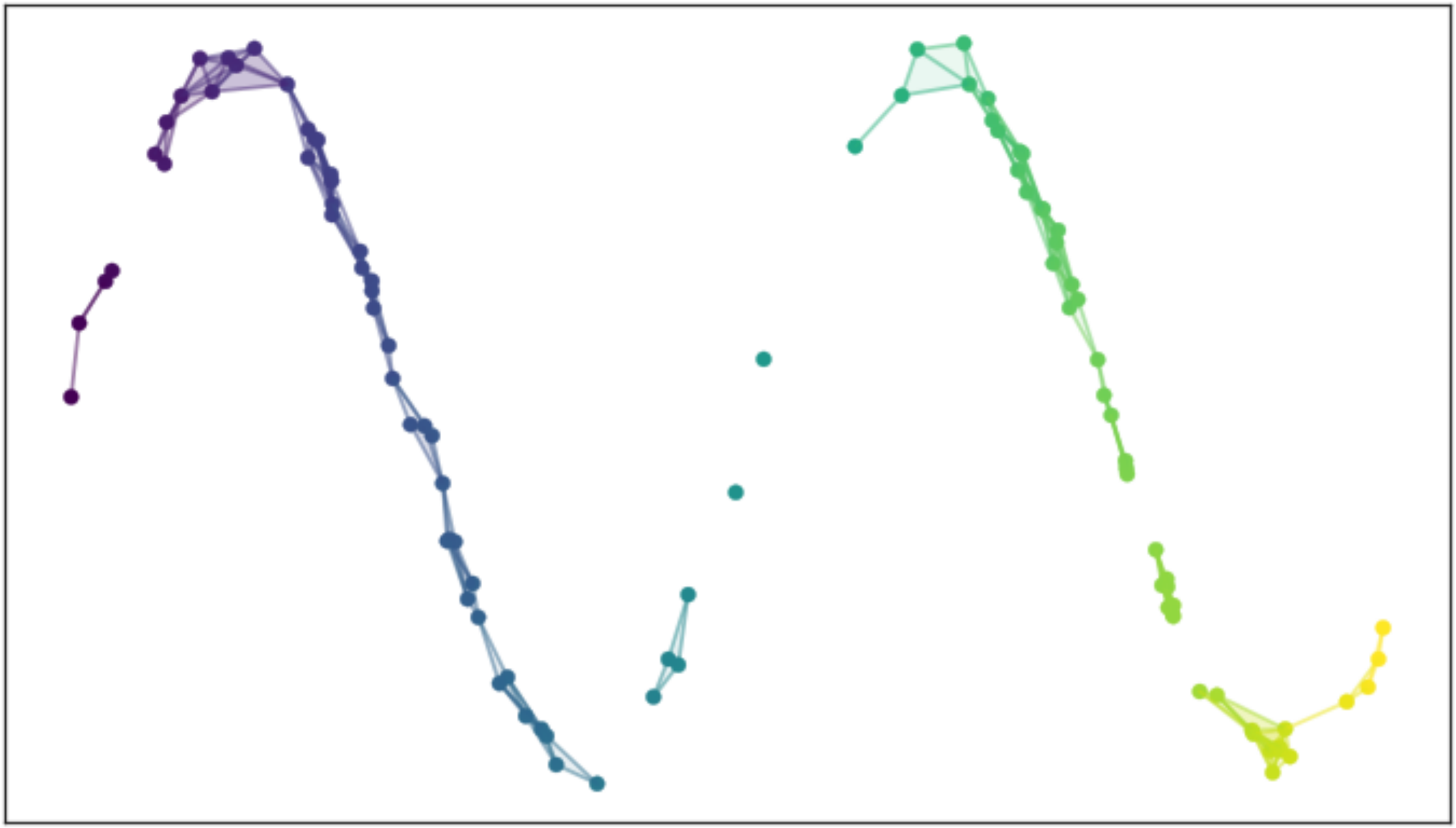
INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE



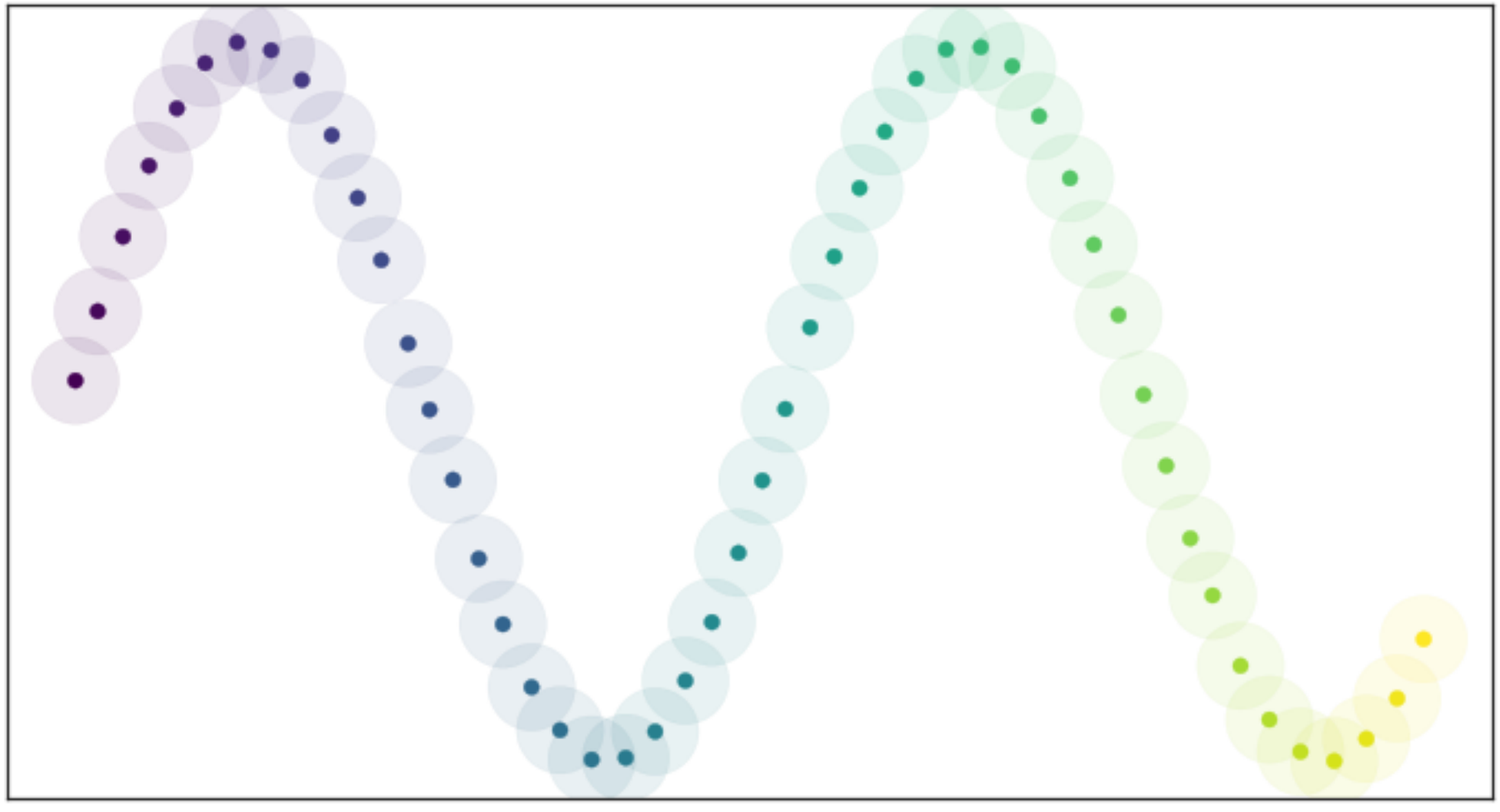
INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

If the data is uniformly distributed on the manifold then the cover will be “good”

INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

When is data that nicely behaved?

INSTITUT TUTTE INSTITUTE

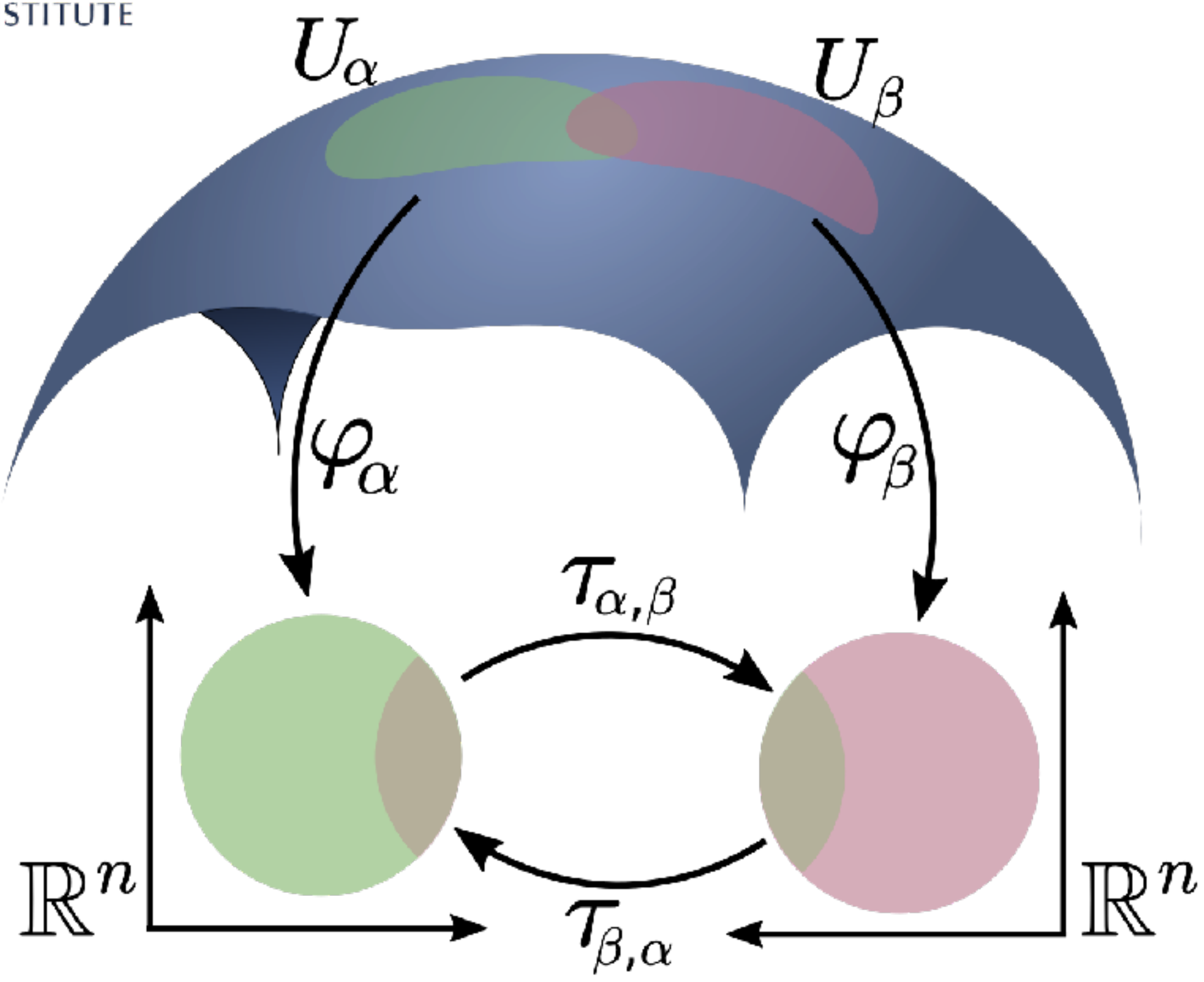
Assumption:

Data is uniformly distributed on the manifold

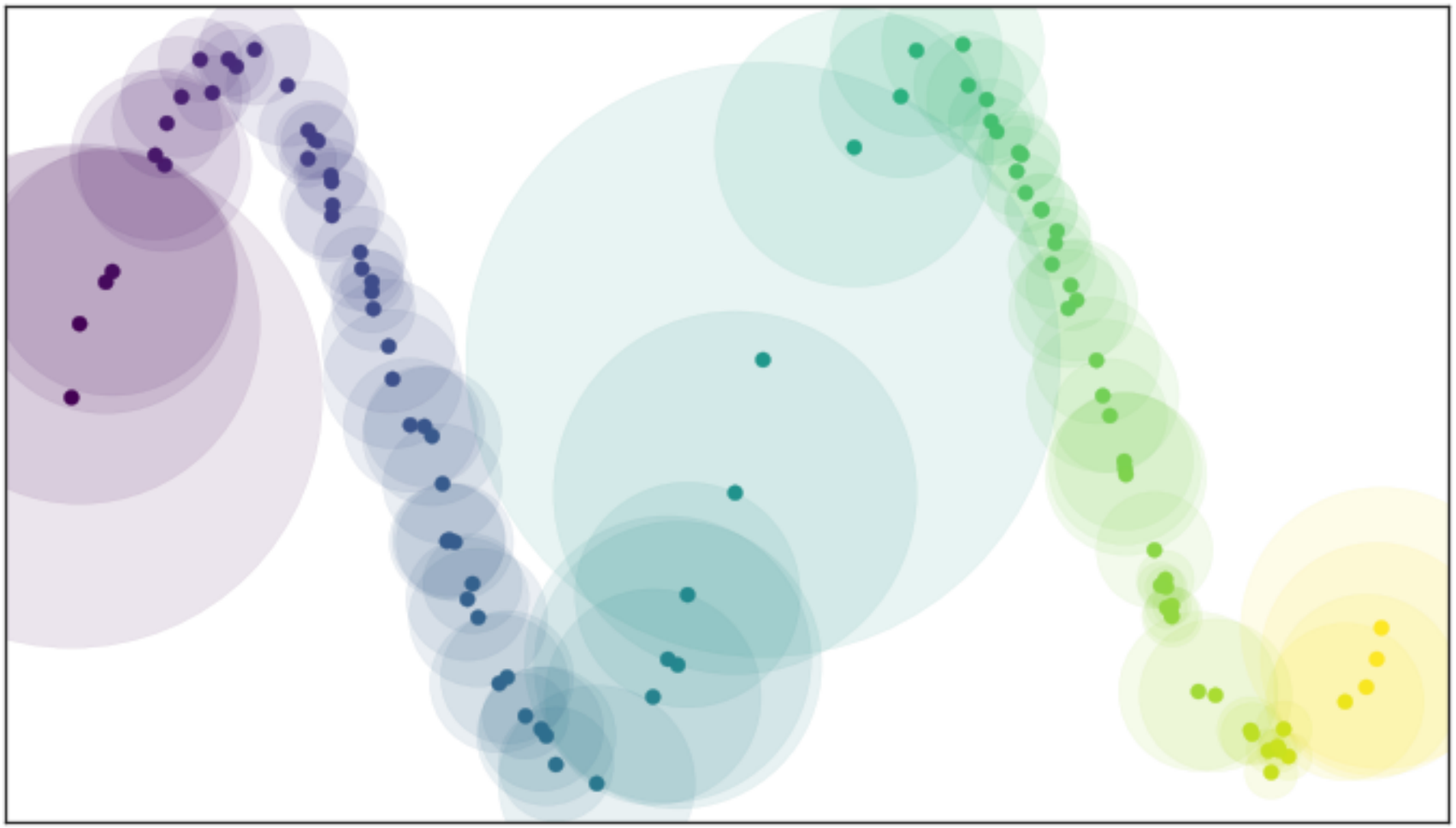
INSTITUT TUTTE INSTITUTE

Define a Riemannian metric on the manifold to make this assumption true

INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE



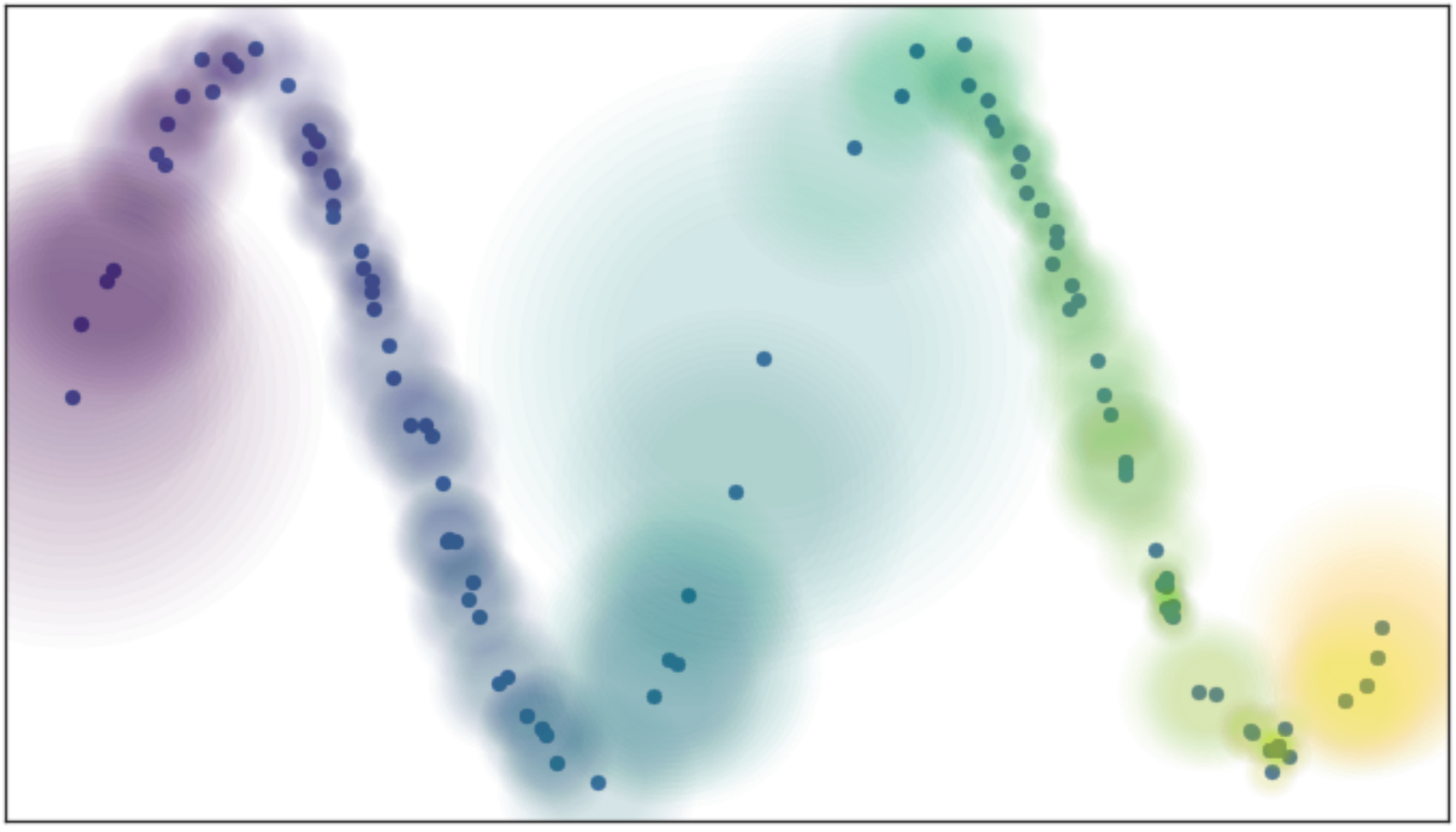
INSTITUT TUTTE INSTITUTE

Why choose a fixed radius? Why not have a fuzzy cover?

INSTITUT TUTTE INSTITUTE

Theorem 2 (UMAP Adjunction)**.** *The functors* FinReal : **sFuzz** $\to$ **FinEPMet** *and* FinSing : **FinEPMet** $\to$ **sFuzz** *form an adjunction* FinReal $\dashv$ FinSing.

INSTITUT TUTTE INSTITUTE

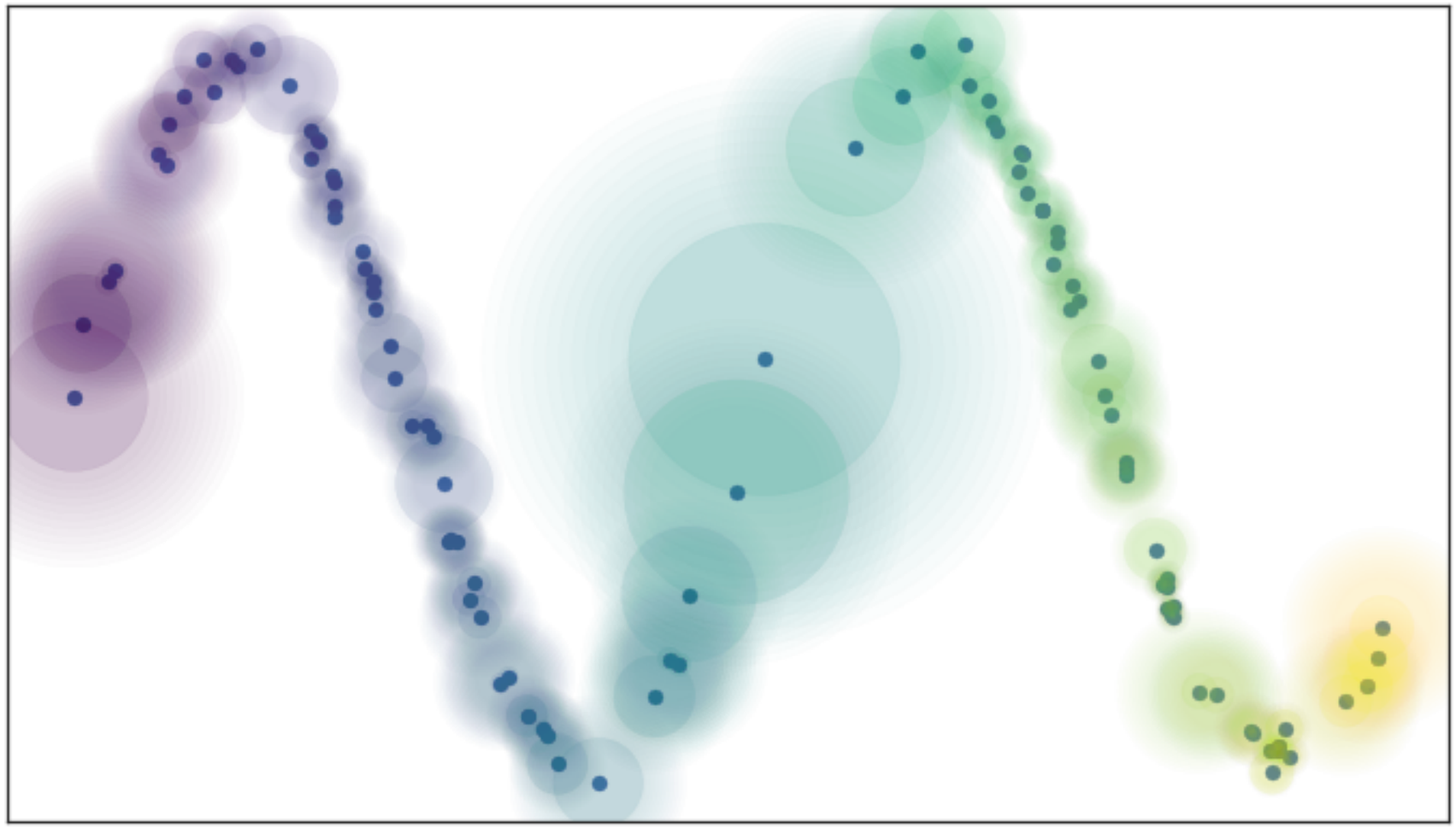


INSTITUT TUTTE INSTITUTE

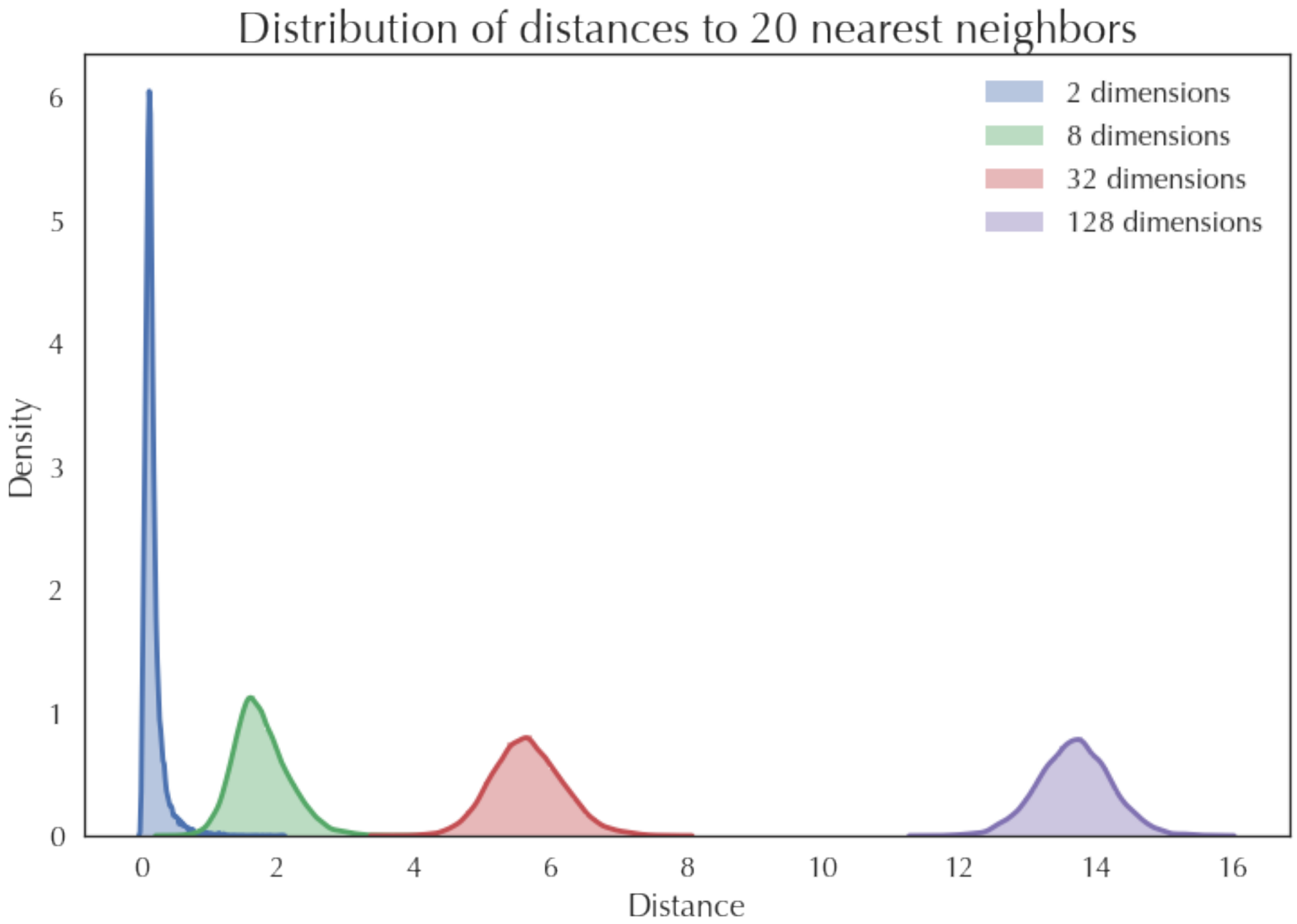
Assumption:

The manifold is locally connected

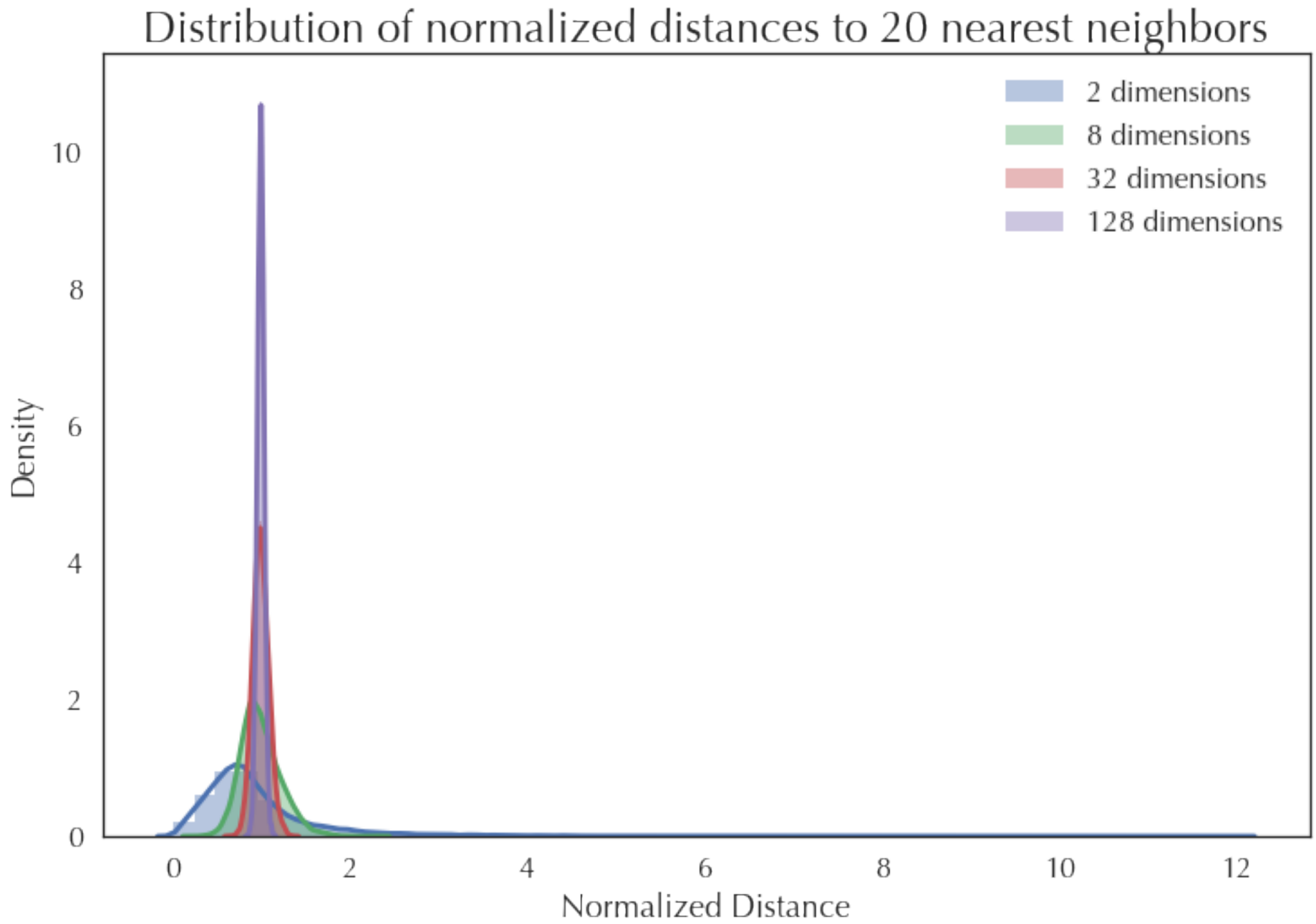
INSTITUT TUTTE INSTITUTE



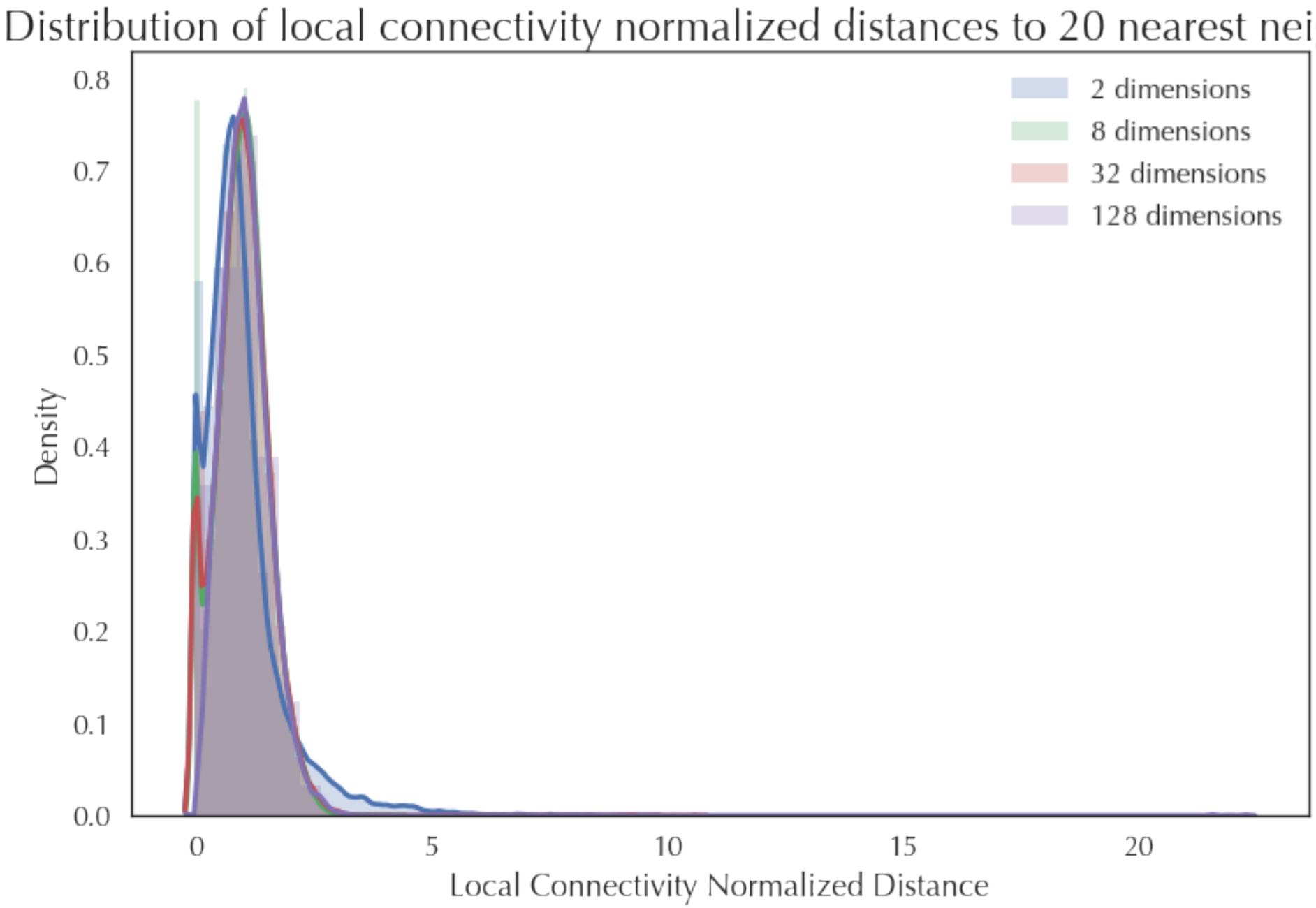
INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE



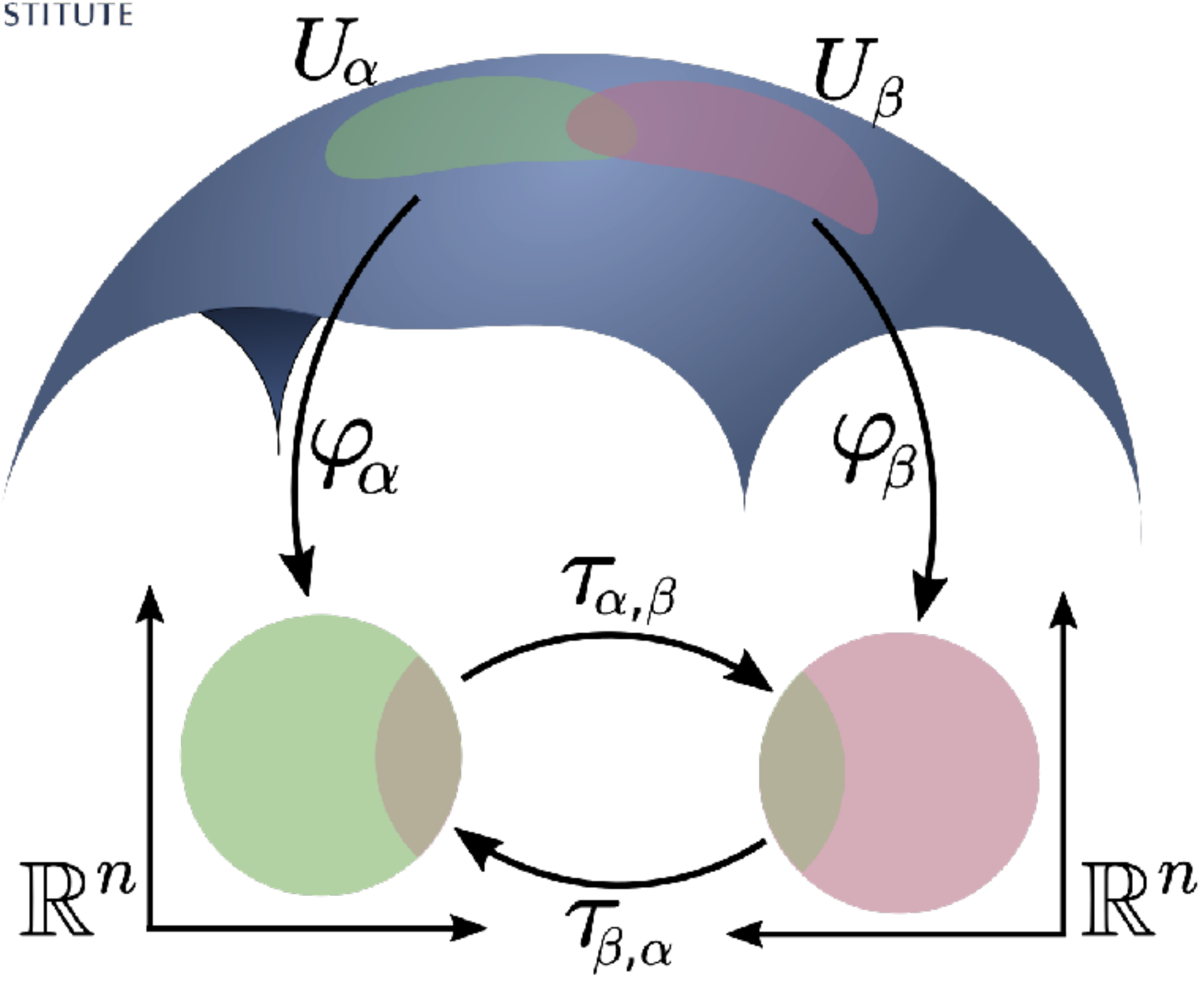
INSTITUT TUTTE INSTITUTE



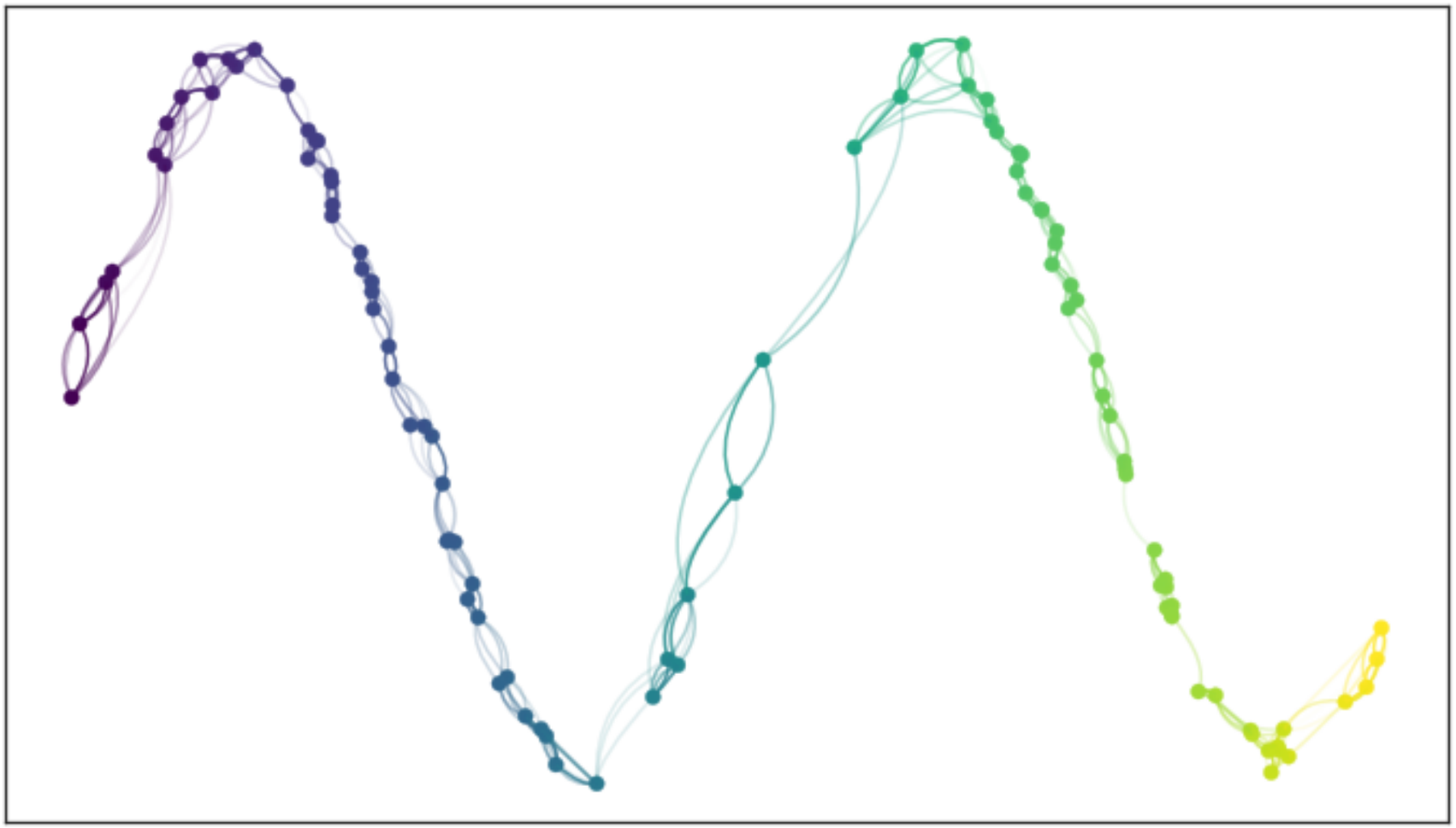
INSTITUT TUTTE INSTITUTE

But our local metrics are all incompatible!

INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

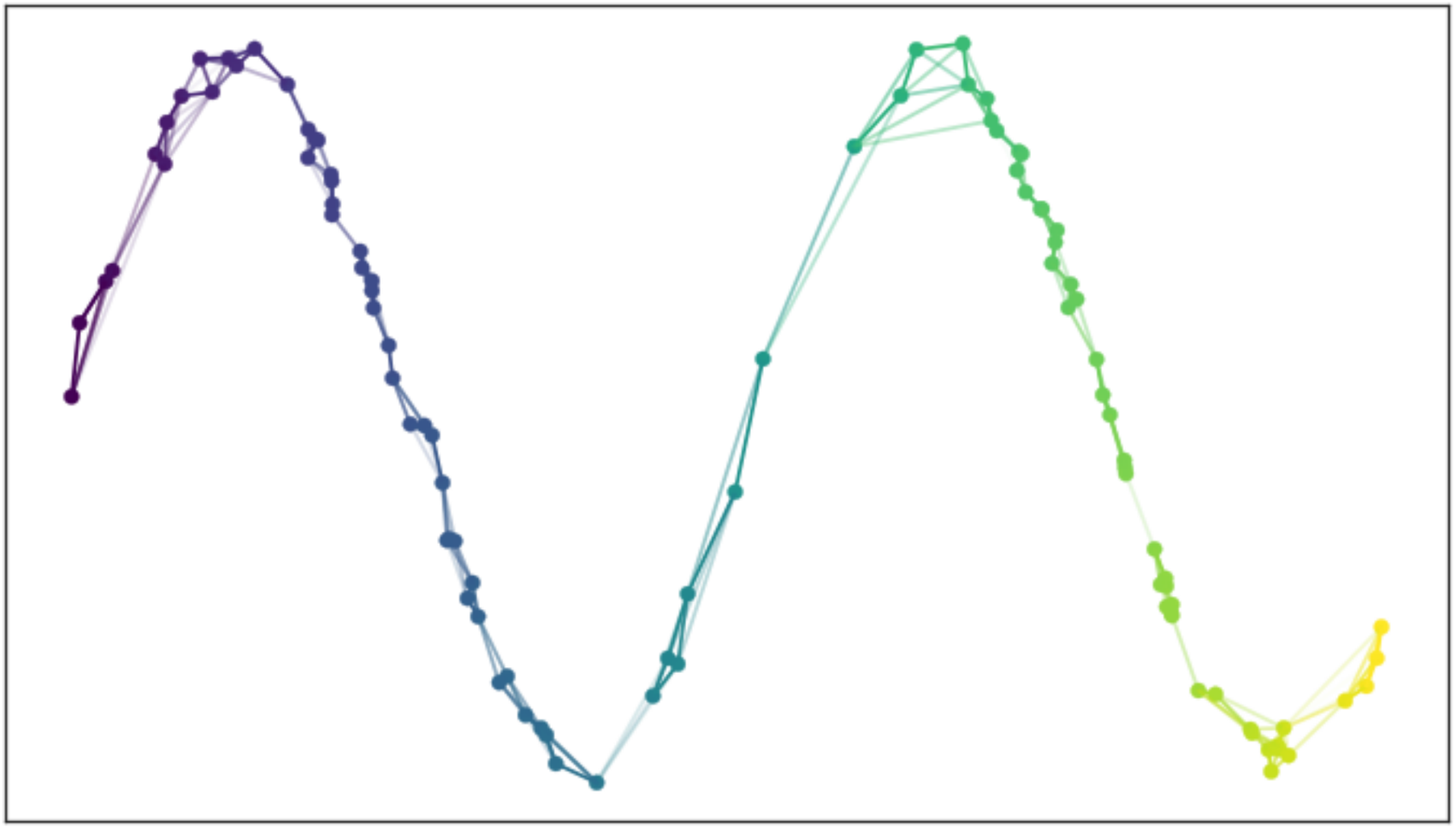
Theorem 2 (UMAP Adjunction)**.** *The functors* FinReal : **sFuzz** $\to$ **FinEPMet** *and* FinSing : **FinEPMet** $\to$ **sFuzz** *form an adjunction* FinReal $\dashv$ FinSing.

INSTITUT TUTTE INSTITUTE

Under a probabilistic fuzzy union the combination of weights on edges is given by

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

INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

Suppose we were given a low dimensional representation

INSTITUT TUTTE INSTITUTE

We can apply the same process to get a fuzzy graph!

INSTITUT TUTTE INSTITUTE

Except we *know* the manifold, and *don't know* the "correct" nearest neighbour distance

INSTITUT TUTTE INSTITUTE

Now measure the distance between the graphs using cross-entropy and optimize

INSTITUT TUTTE INSTITUTE

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

INSTITUT TUTTE INSTITUTE

We are just embedding the graph

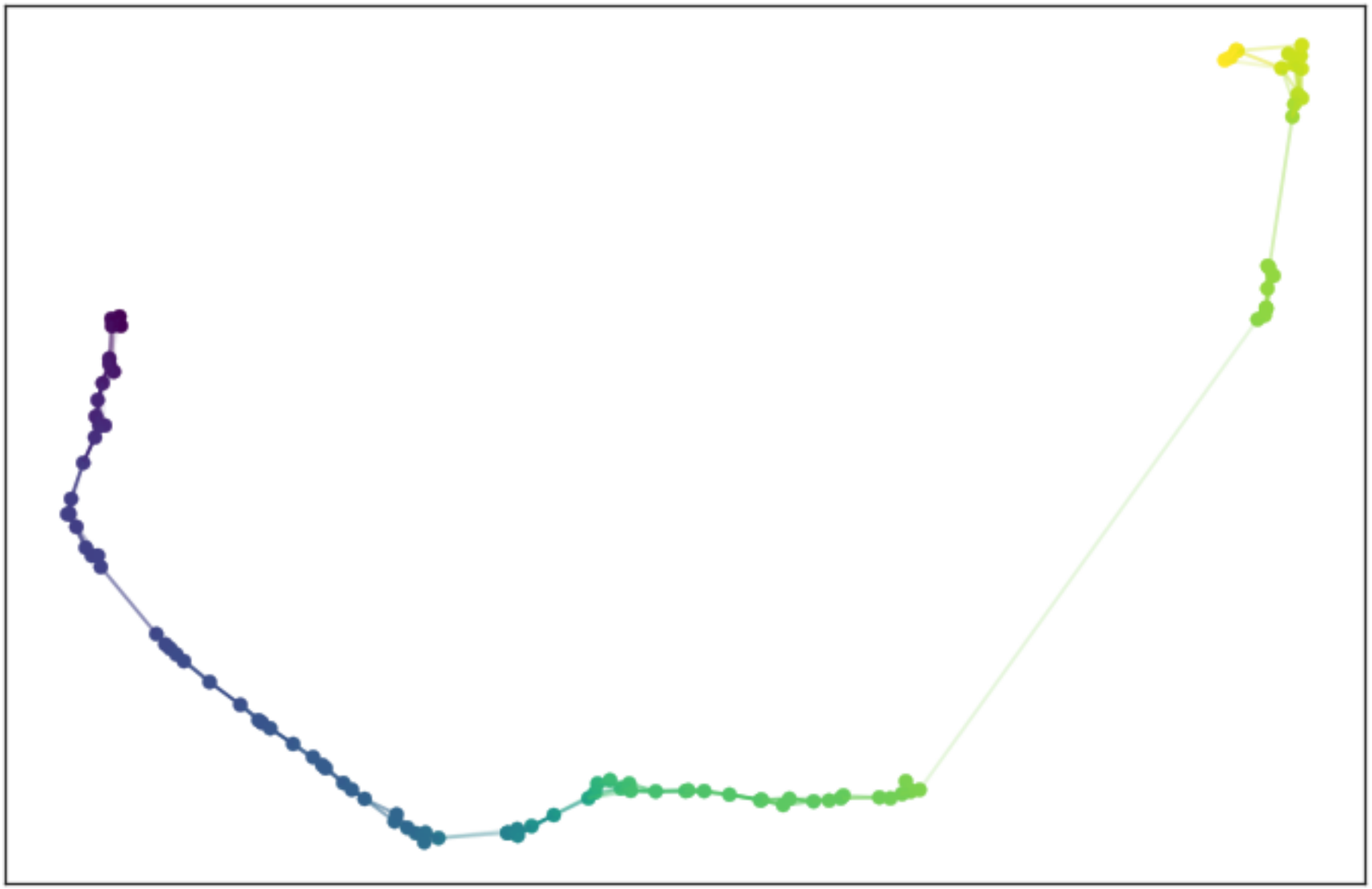
INSTITUT TUTTE INSTITUTE

Get the clumps right

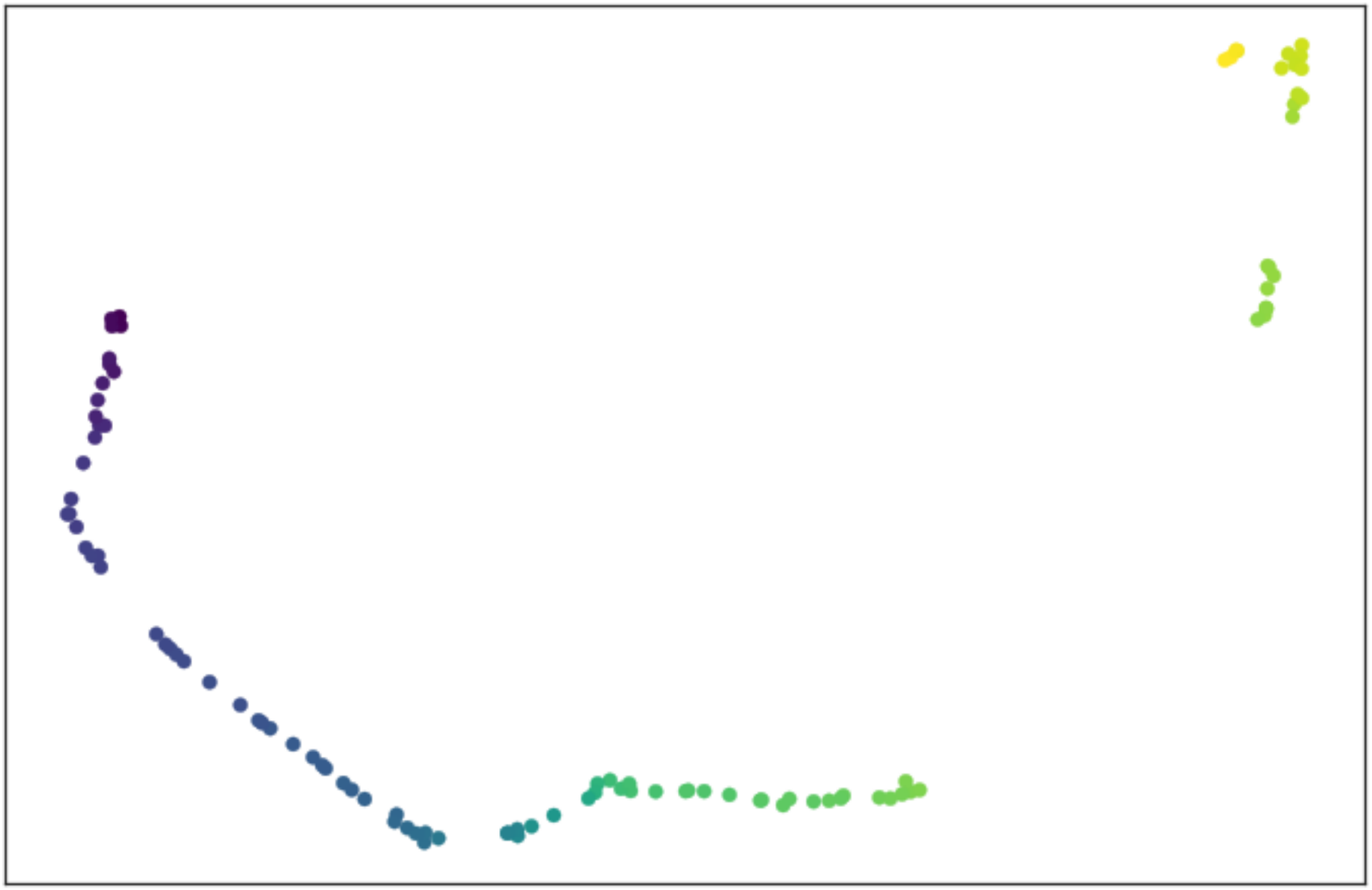
$$\sum_{a \in A} \mu(a) \log\left(\frac{\mu(a)}{\nu(a)}\right) + (1-\mu(a)) \log\left(\frac{1-\mu(a)}{1-\nu(a)}\right)$$

Get the gaps right

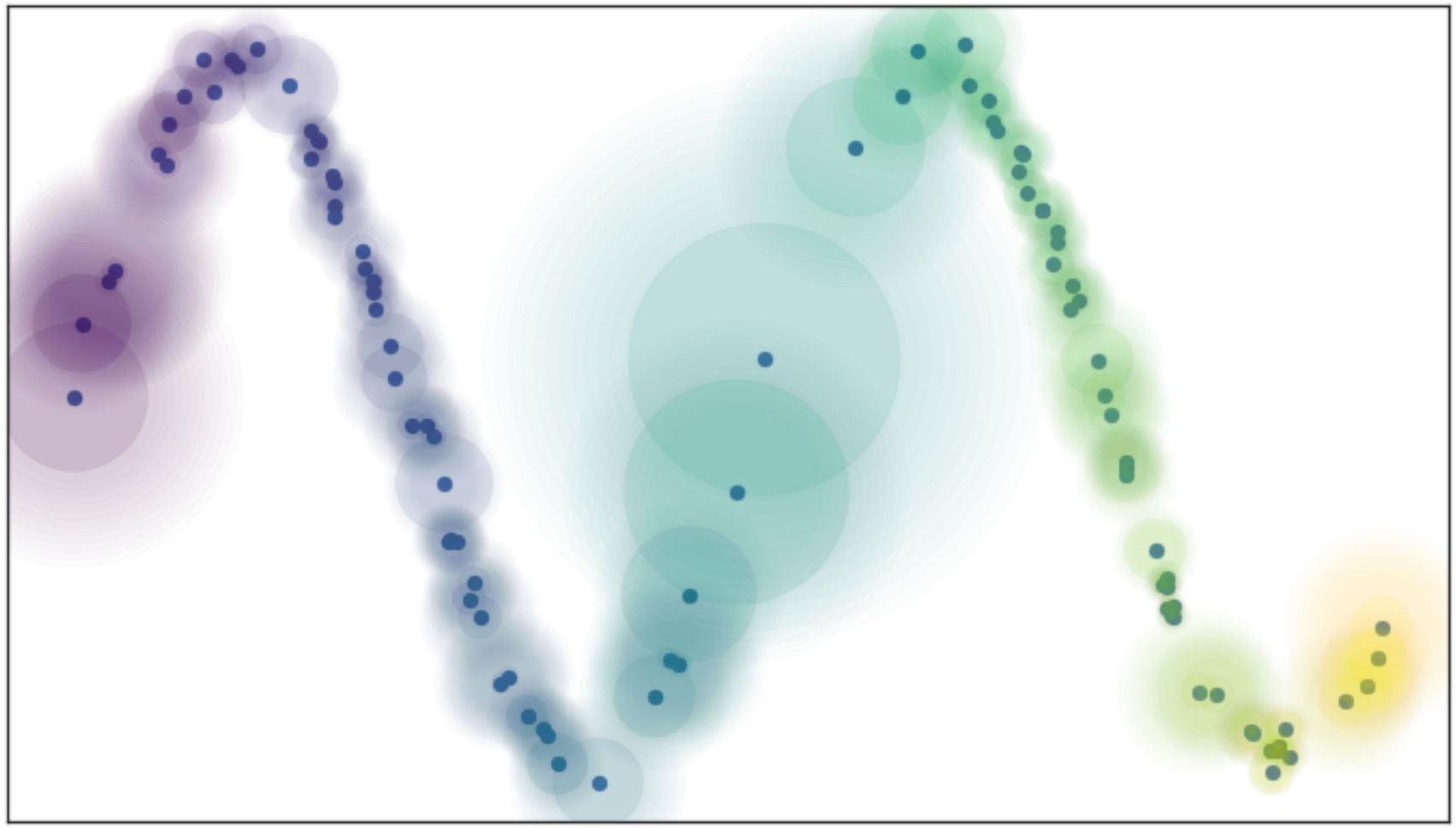
INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

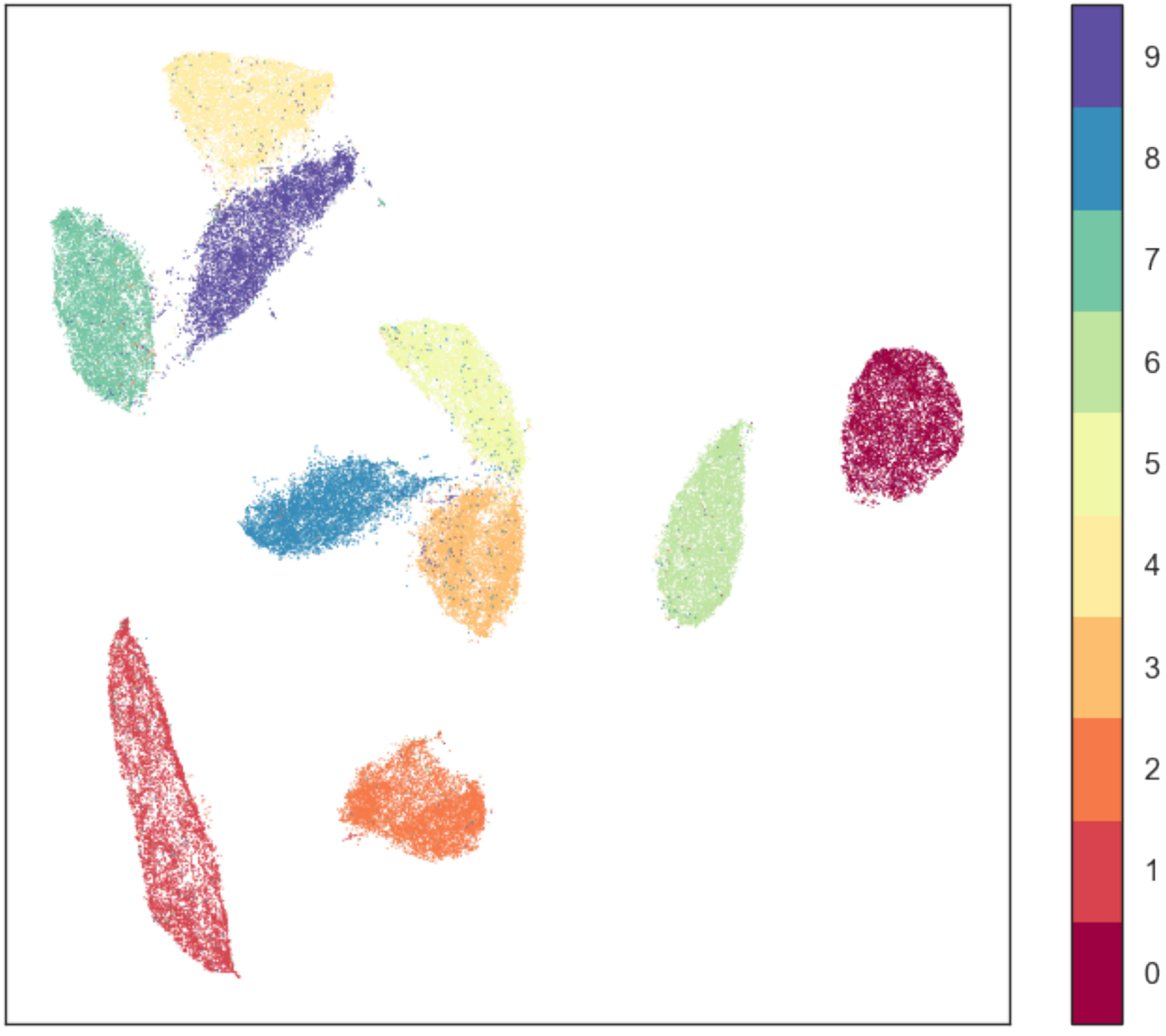


INSTITUT TUTTE INSTITUTE

On real data?

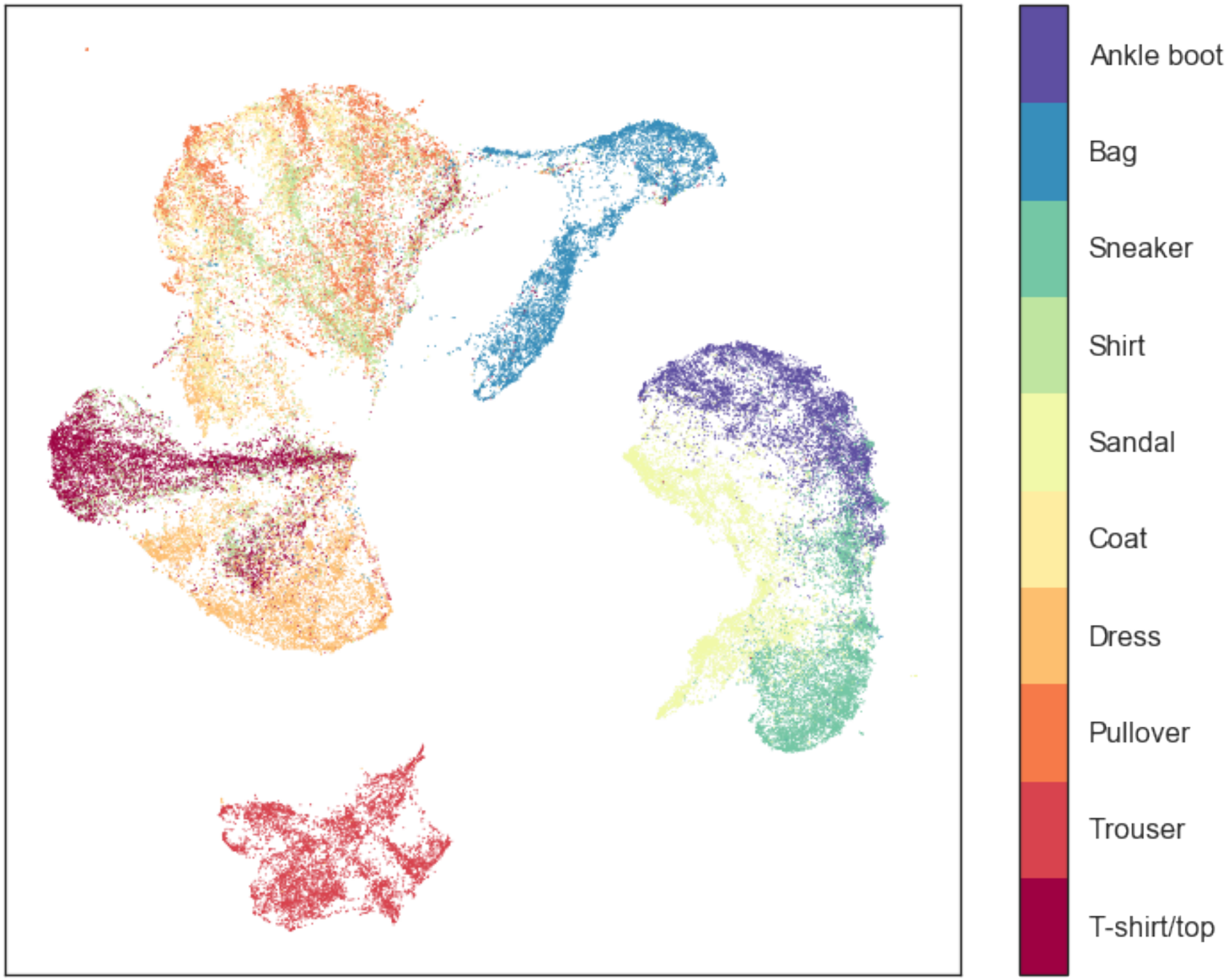
INSTITUT TUTTE INSTITUTE

UMAP on MNIST digits



INSTITUT TUTTE INSTITUTE

UMAP on Fashion MNIST



INSTITUT TUTTE INSTITUTE

Implementation

INSTITUT TUTTE INSTITUTE

Need to find (approximate) nearest neighbours very efficiently

Even in high dimensional space

INSTITUT TUTTE INSTITUTE

RP-trees + NN-descent

Dasgupta + Freund 2008

Dong, Charikar + Li 2011

INSTITUT TUTTE INSTITUTE

Need to optimize the layout subquadratically

INSTITUT TUTTE INSTITUTE

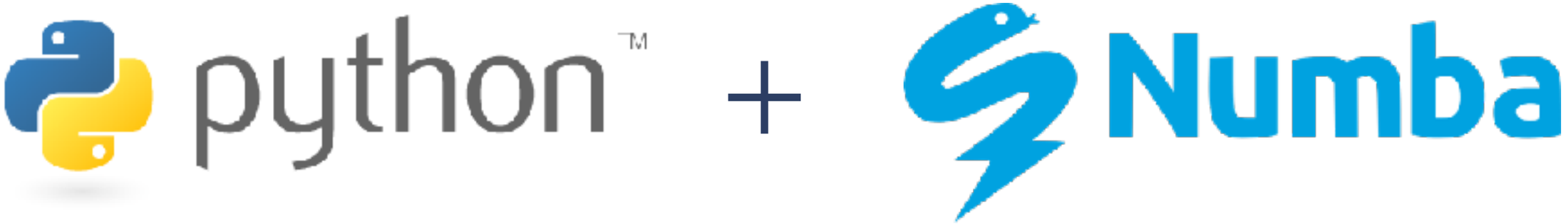
SGD + negative sampling

Mikolov et al 2013
Tang et al 2016

INSTITUT TUTTE INSTITUTE

Need to be high level but still fast

INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

Numba is awesome!

INSTITUT TUTTE INSTITUTE

- High performance
- Clean code
- Custom distance metrics

INSTITUT TUTTE INSTITUTE

Performance Comparison

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

INSTITUT TUTTE INSTITUTE

Where Next?

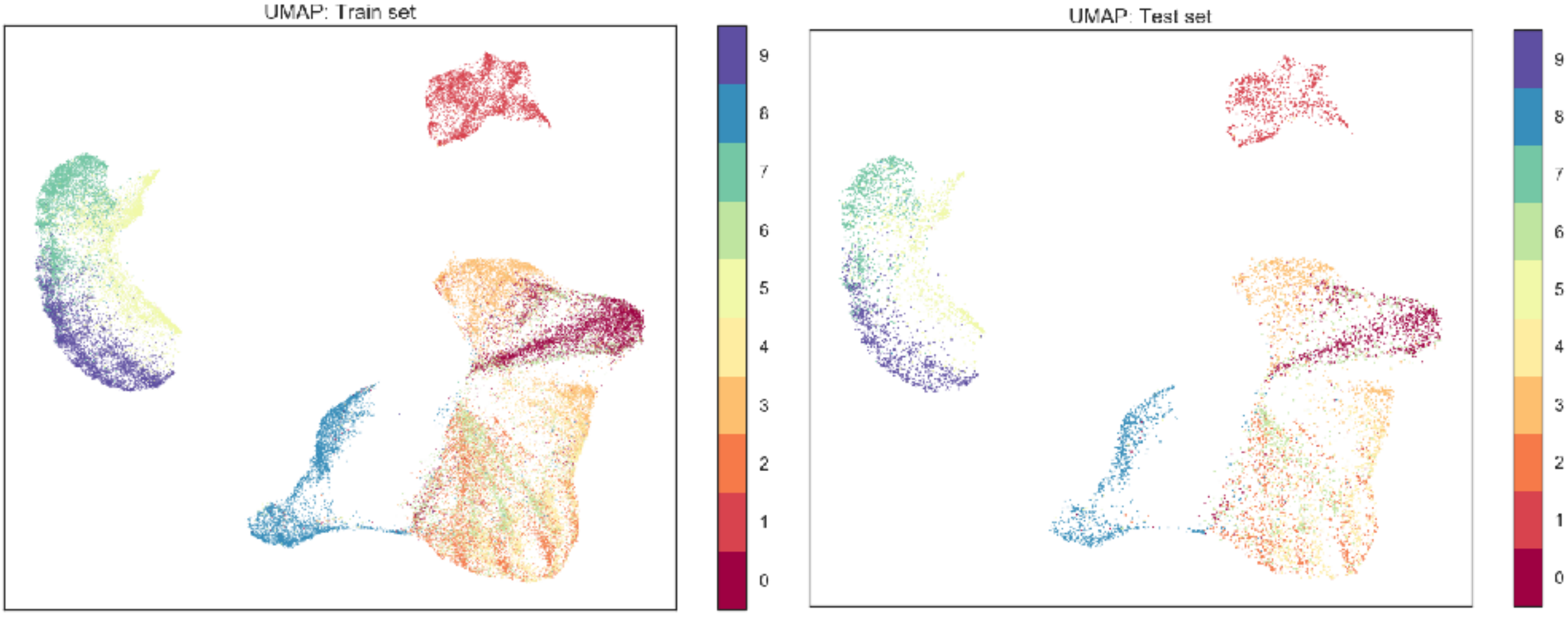
INSTITUT TUTTE INSTITUTE

Given the mathematical foundation, a number of options are available

INSTITUT TUTTE INSTITUTE

Embed new unseen points into an existing embedding

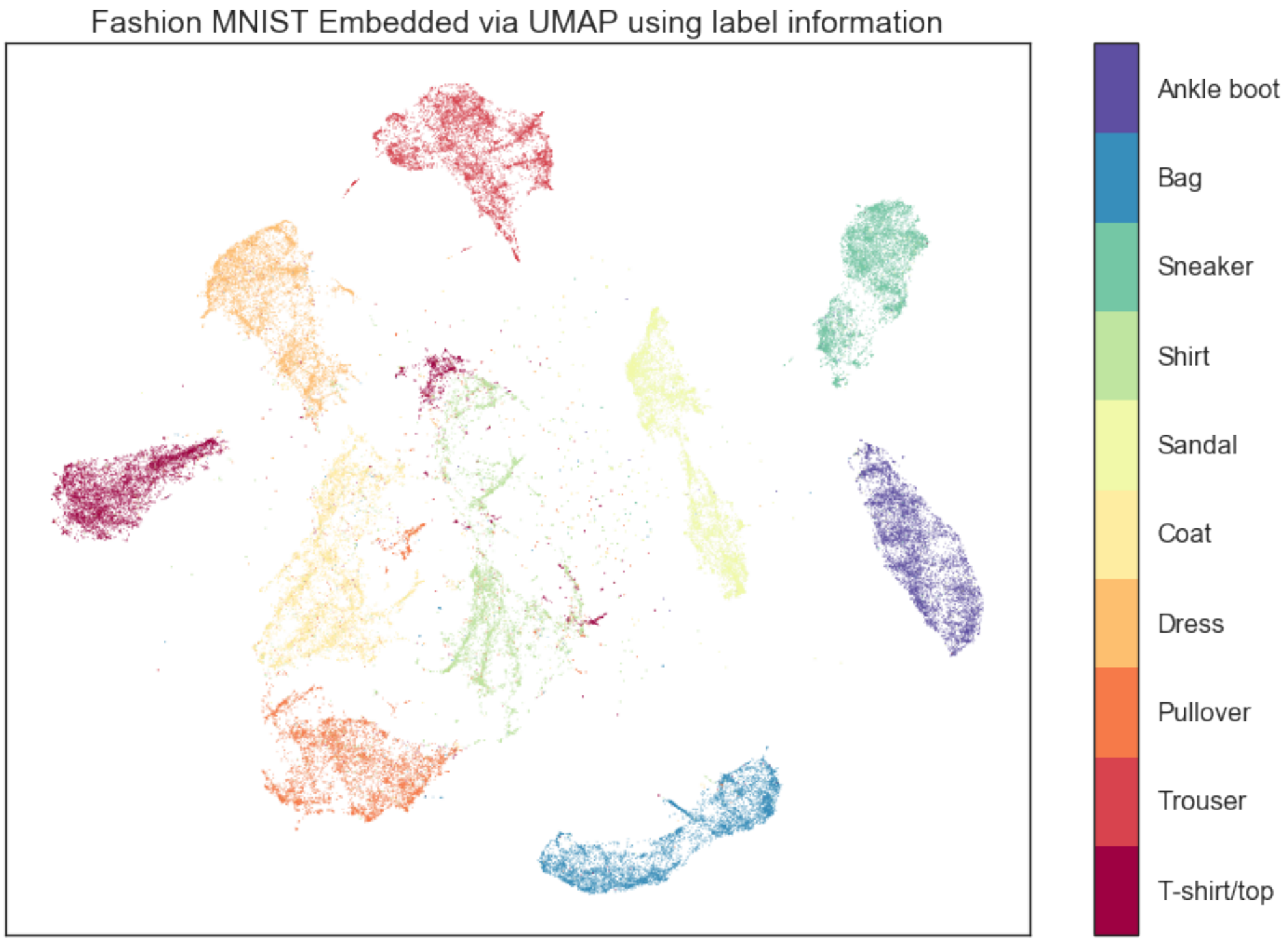
INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

Make use of labels for supervised dimension reduction

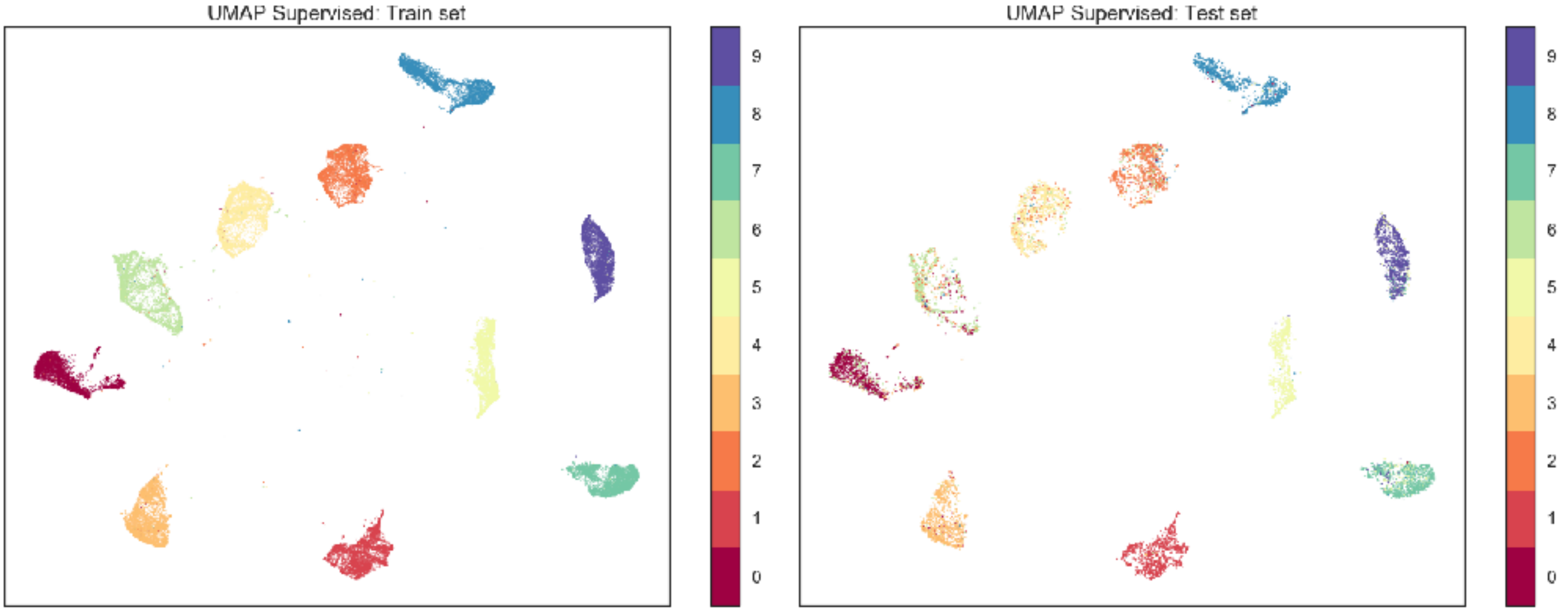
INSTITUT TUTTE INSTITUTE



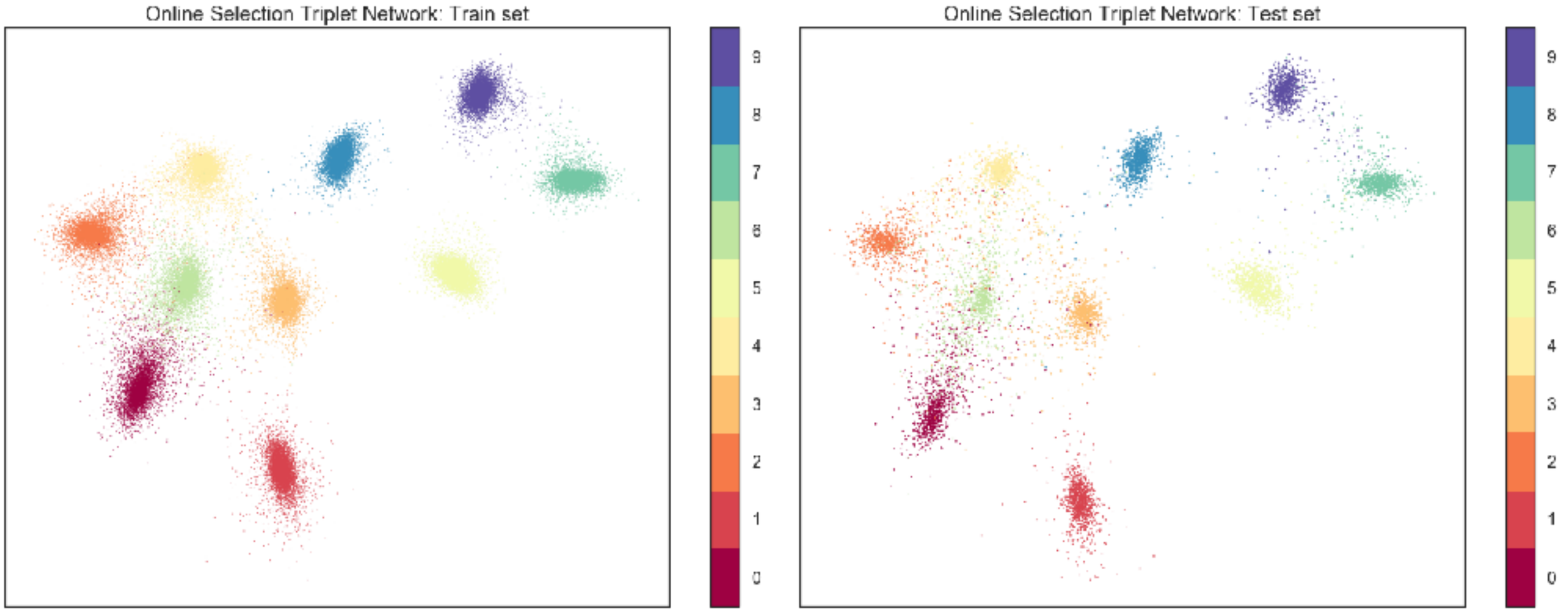
INSTITUT TUTTE INSTITUTE

And combine those for metric learning

INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE



Derived from Adam Bielski's Siamese/Triplet repository:
https://github.com/adambielski/siamese-triplet

INSTITUT TUTTE INSTITUTE

Adding one categorical variable is no harder theoretically than adding many

INSTITUT TUTTE INSTITUTE

Combine spaces with different metrics

Continuous, categorical, ordinal, Haversine, Levenstein, and more …

INSTITUT TUTTE INSTITUTE

As long as you can provide a metric for the datatype UMAP can combine it with other datatypes!

INSTITUT TUTTE INSTITUTE

UMAP for pandas dataframes!

INSTITUT TUTTE INSTITUTE

https://github.com/lmcinnes/umap

conda install -c conda-forge umap-learn

pip install umap-learn

leland.mcinnes@gmail.com

@leland_mcinnes