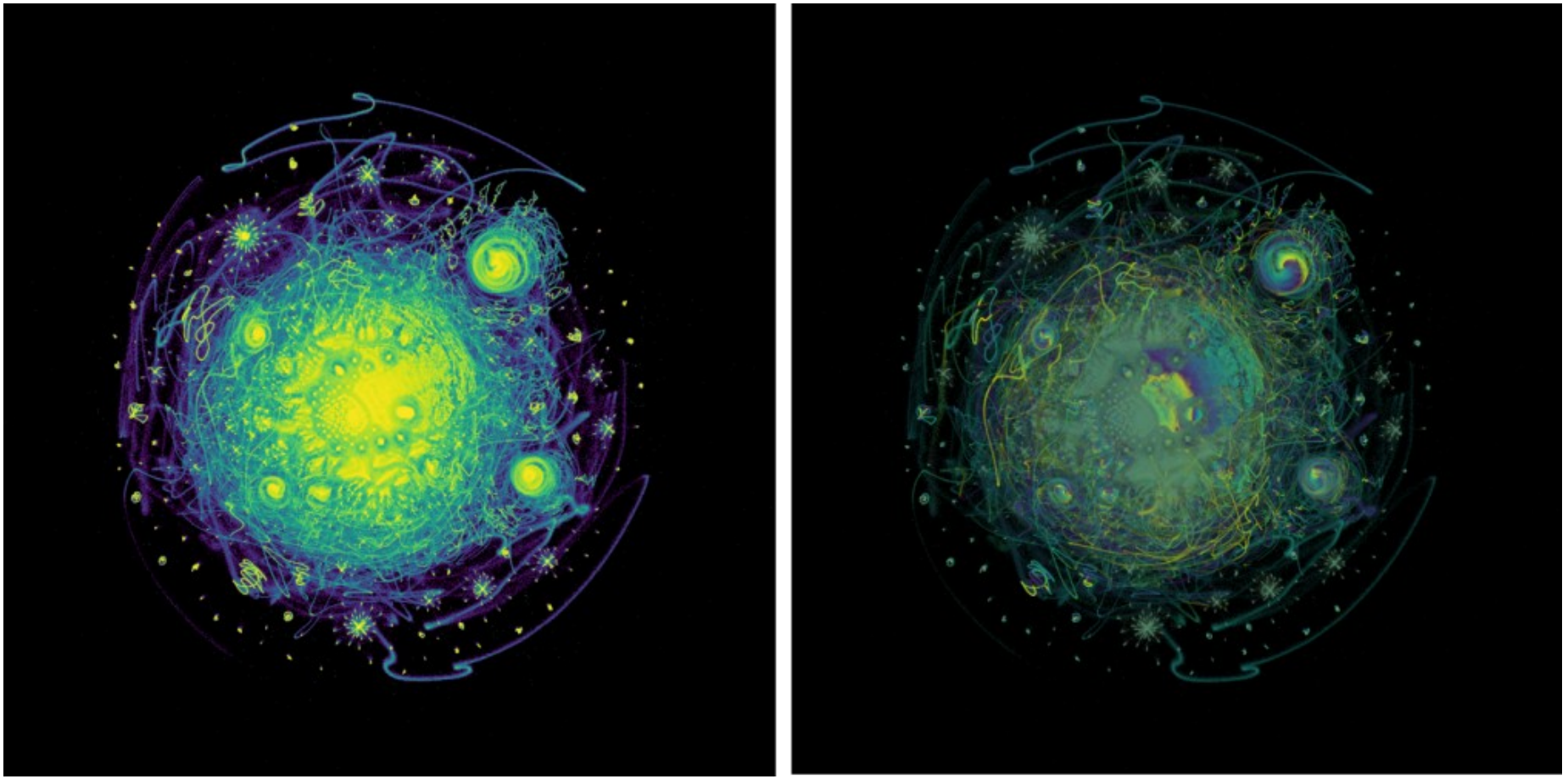
Dimension Reduction for OMICs Integration

OMICs Integration and Systems Biology course

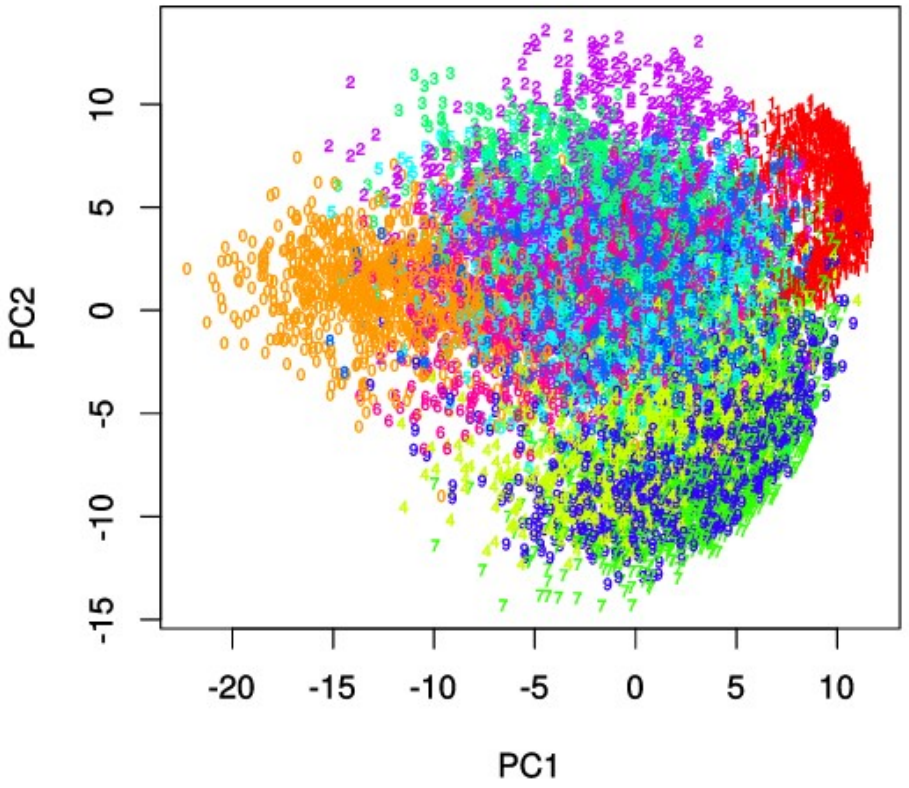
Nikolay Oskolkov, NBIS SciLifeLab

Lund, 5.10.2020

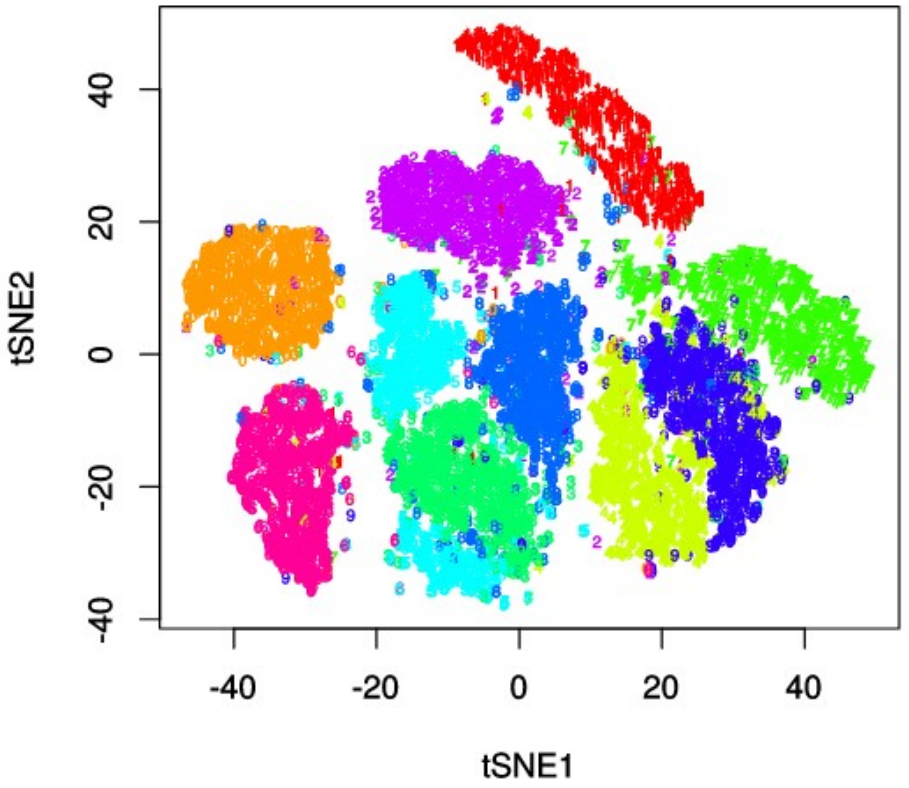


Dimension Reduction and Visualization

PCA PLOT WITH PRCOMP



tSNE MNIST



Dimension reduction is not only for visualization but overcoming the Curse of Dimensionality

The Curse of Dimensionality

P is the number of features (genes, proteins, genetic variants etc.)
N is the number of observations (samples, cells, nucleotides etc.)

Biomedicine

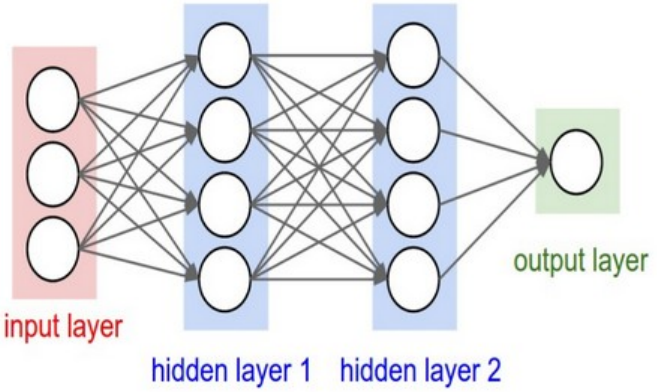
Bayesianism

P >> N

Frequentism

P ~ N

Deep Learning



P << N

Amount of Data

Ex.1

$$Y = \alpha + \beta X$$

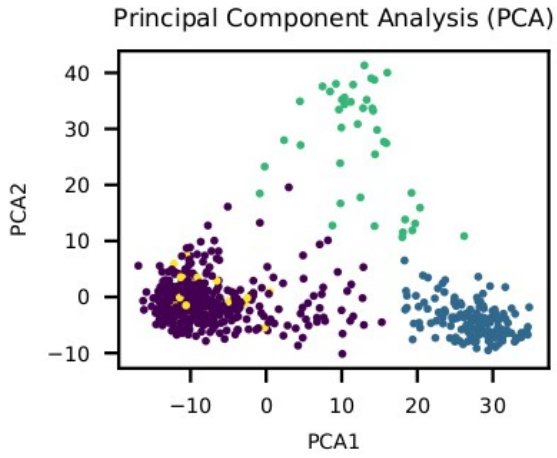
$$\beta = (X^T X)^{-1} X^T Y$$

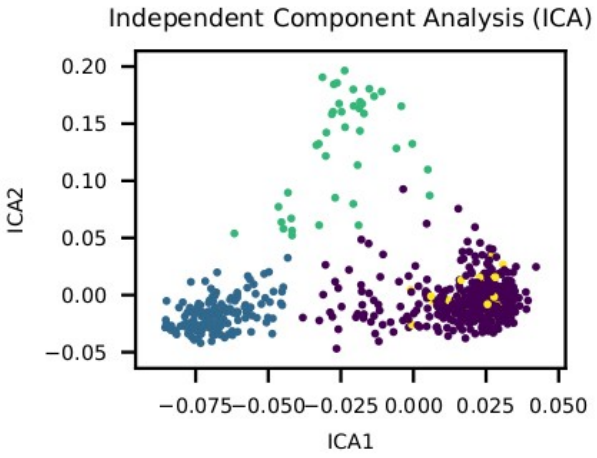
$$(X^T X)^{-1} \sim \frac{1}{\det(X^T X)} \cdots \to \infty, \quad n << p$$

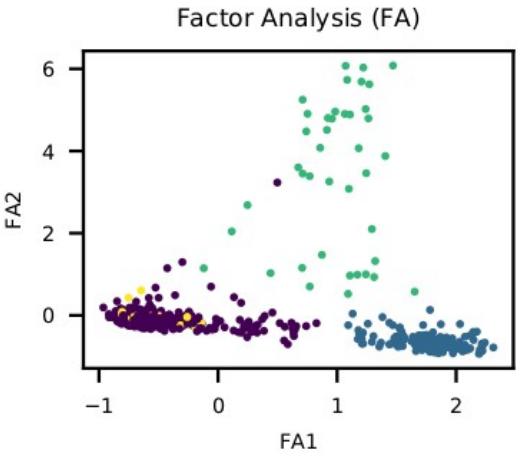
Ex.2 $$E[\hat{\sigma}^2] = \frac{n-p}{n}\sigma^2$$

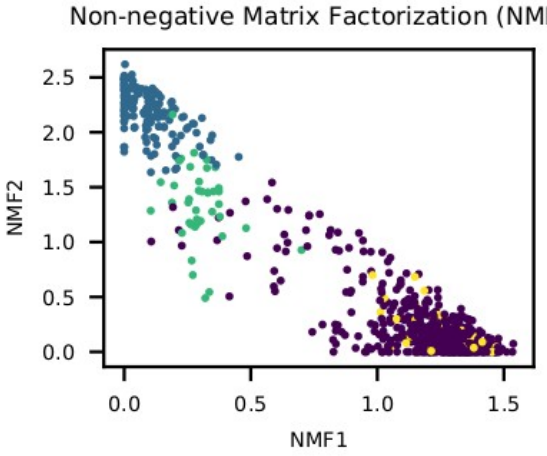
Biased ML variance estimator in HD-space

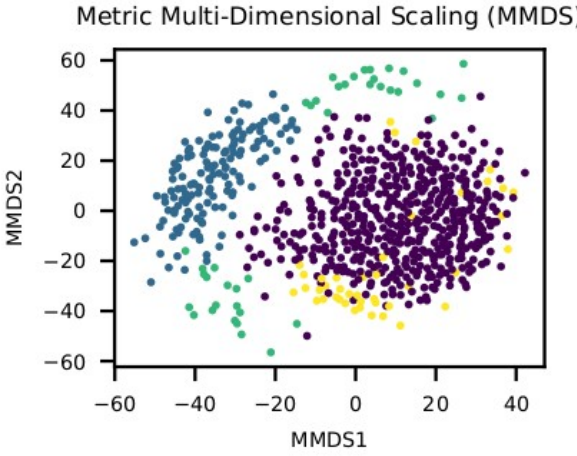
Linear Dimension Reduction

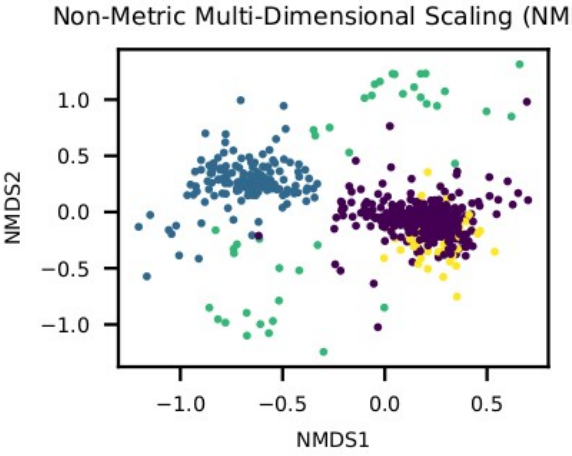


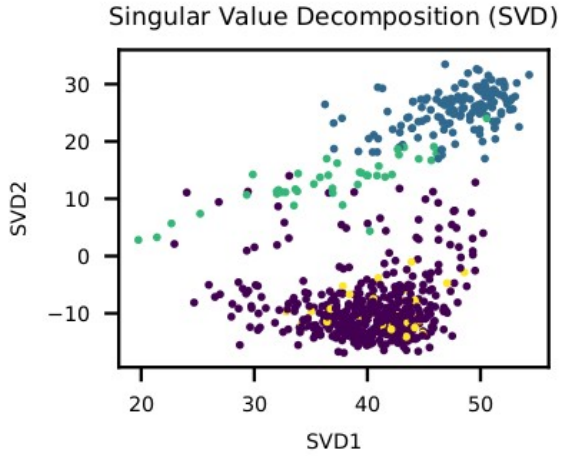


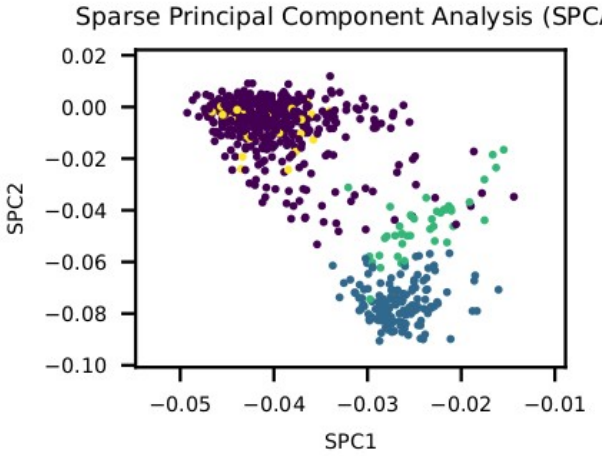


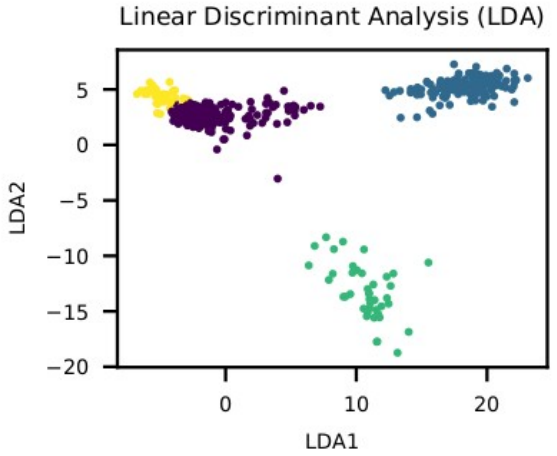




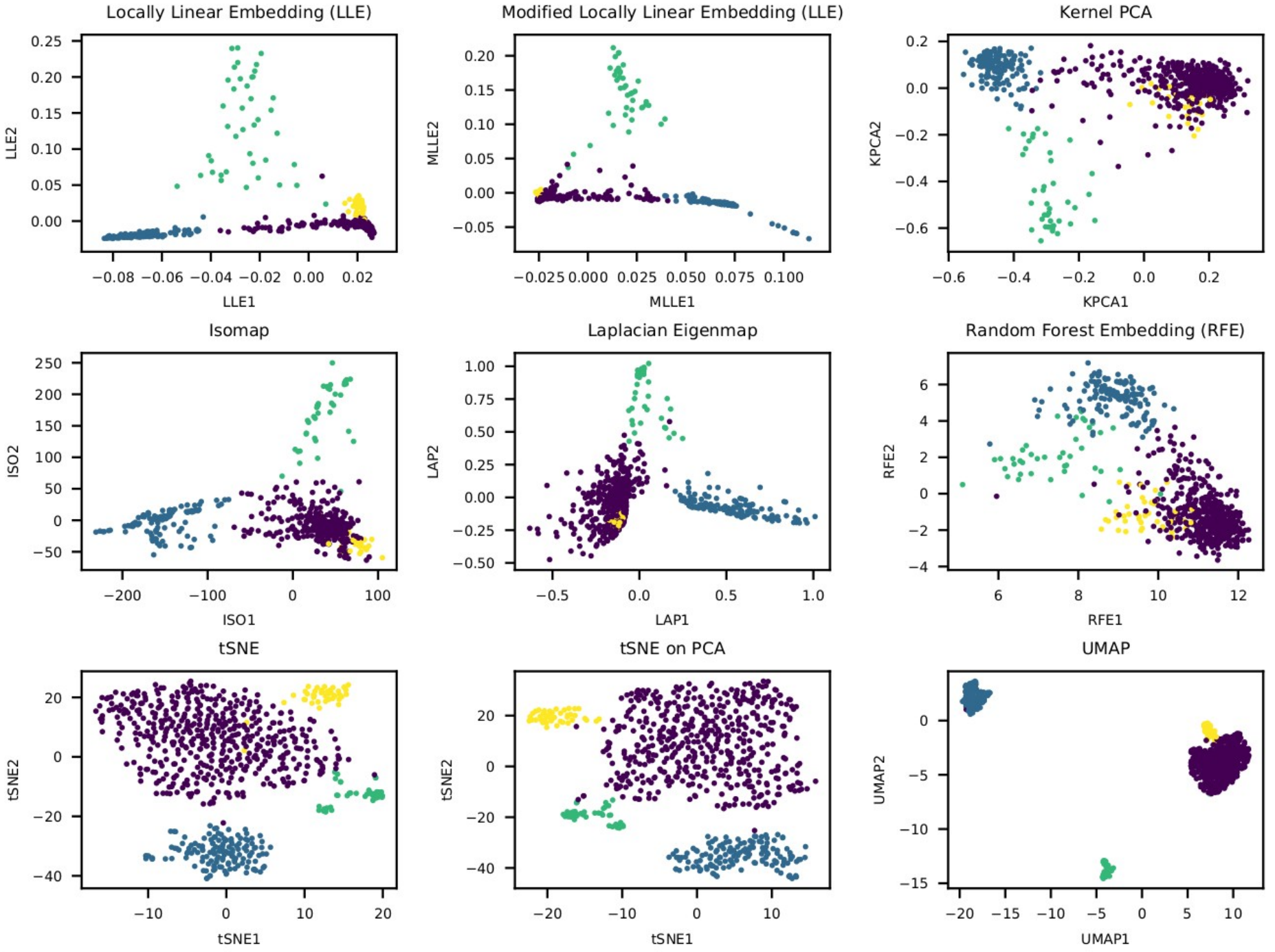






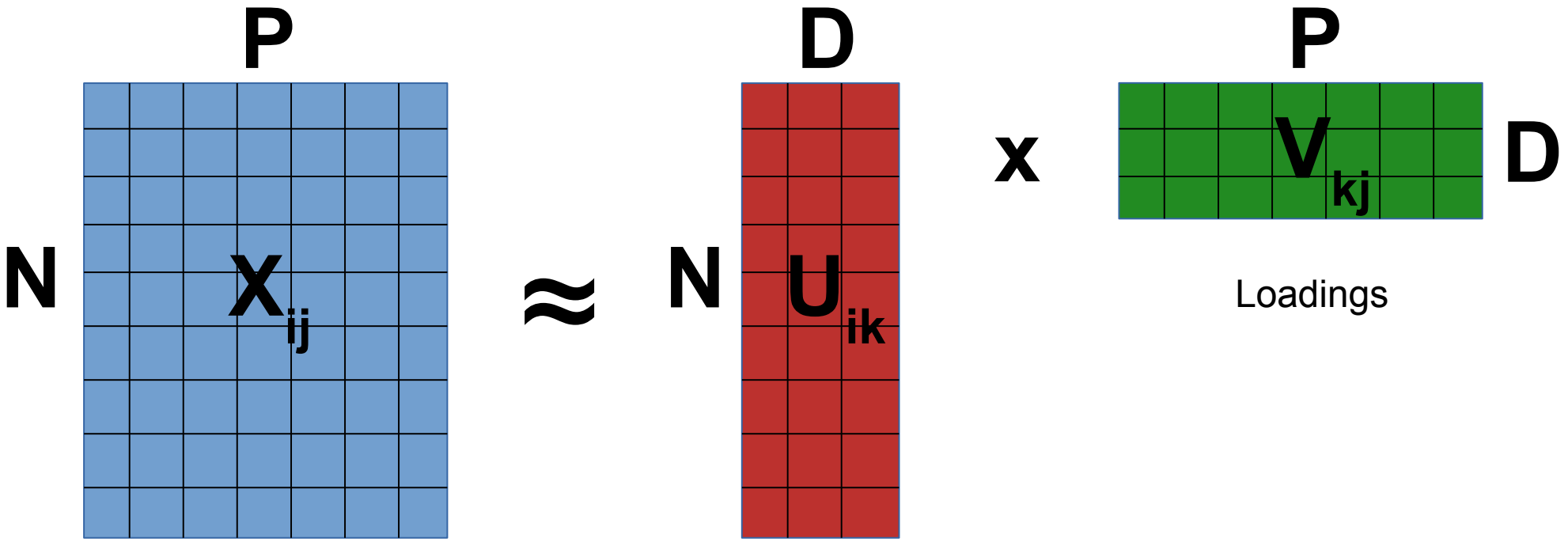


Non-Linear Dimension Reduction



Matrix Factorization Dimension Reduction

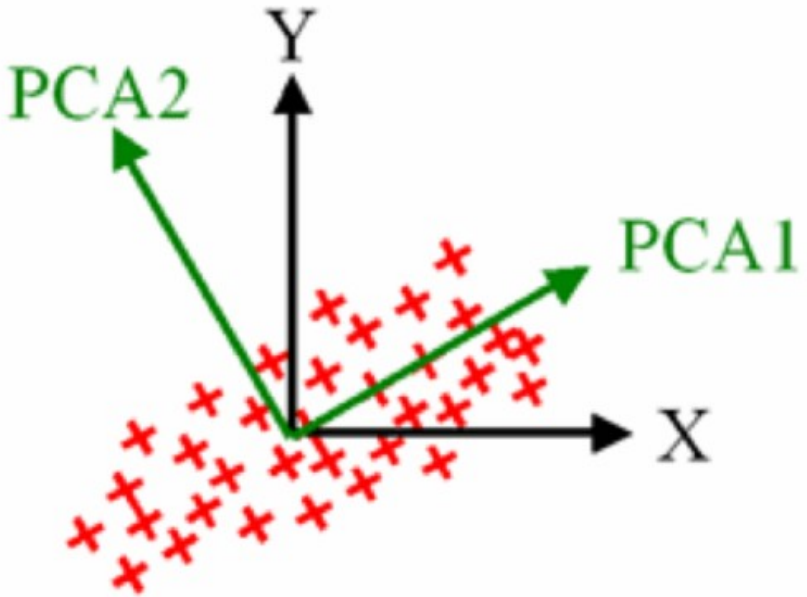
$$X_{ij} \approx U_{ik} V_{kj}$$



$$\text{Loss} = \sum_{i=1}^{N} \sum_{j=1}^{P} (X_{ij} - U_{ik} V_{kj})^2$$

Principal Component Analysis (PCA)

- Collapse p features (p>>n) to few latent features and keep variation
- Rotation and shift of coordinate system toward maximal variance
- PCA is an **eigen matrix decomposition** problem



$$PC = u^T X = X^T u$$

X is mean centered $\Longrightarrow < PC >= 0$

$$< (PC - < PC >)^2 >=< PC^2 >= u^T X X^T u$$

$$XX^T = A$$

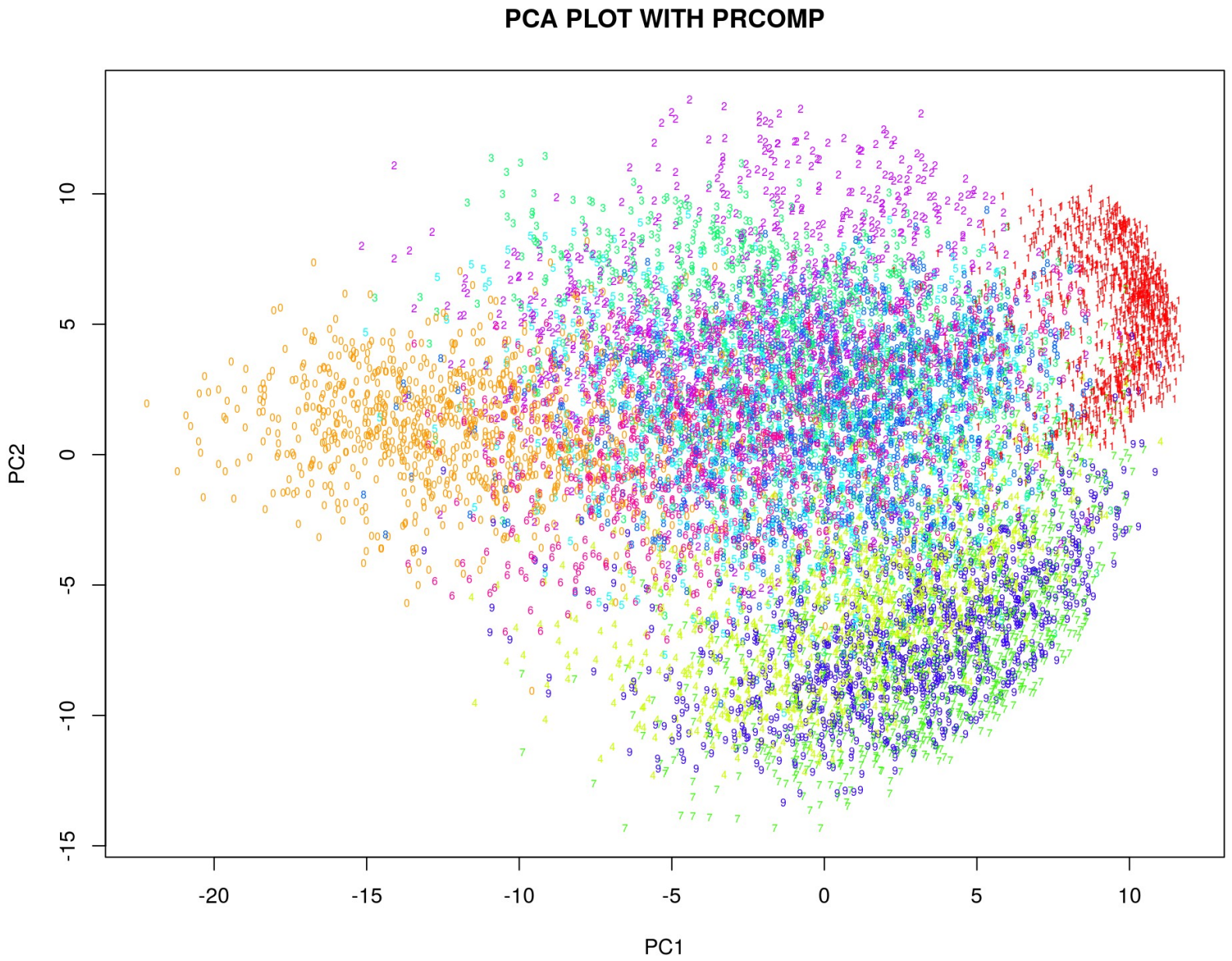
$$< PC^2 >= u^T A u$$

A is **variance-covariance** of X

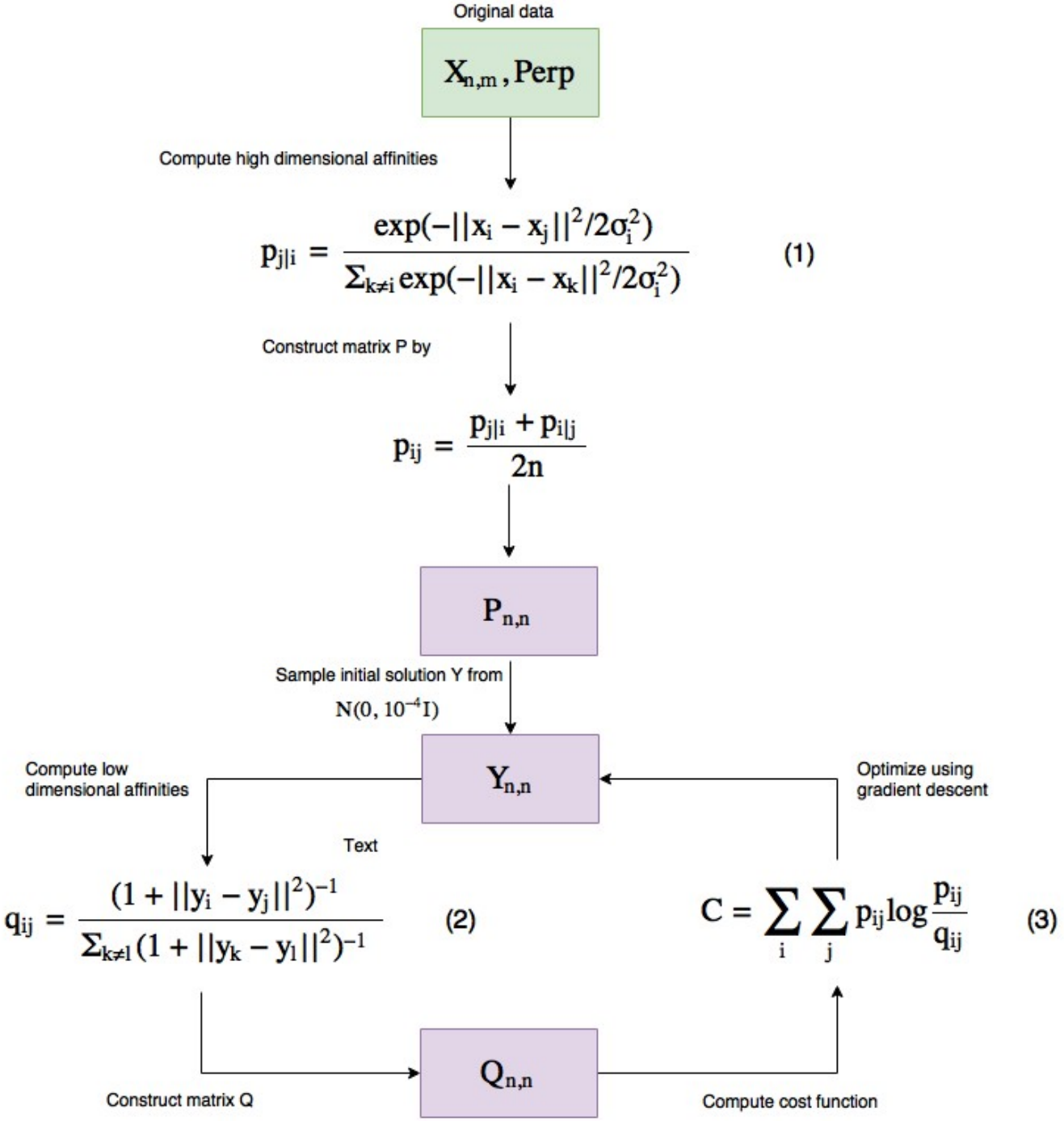
$$\max(u^T A u + \lambda(1 - u^T u)) = 0$$

$$Au = \lambda u$$

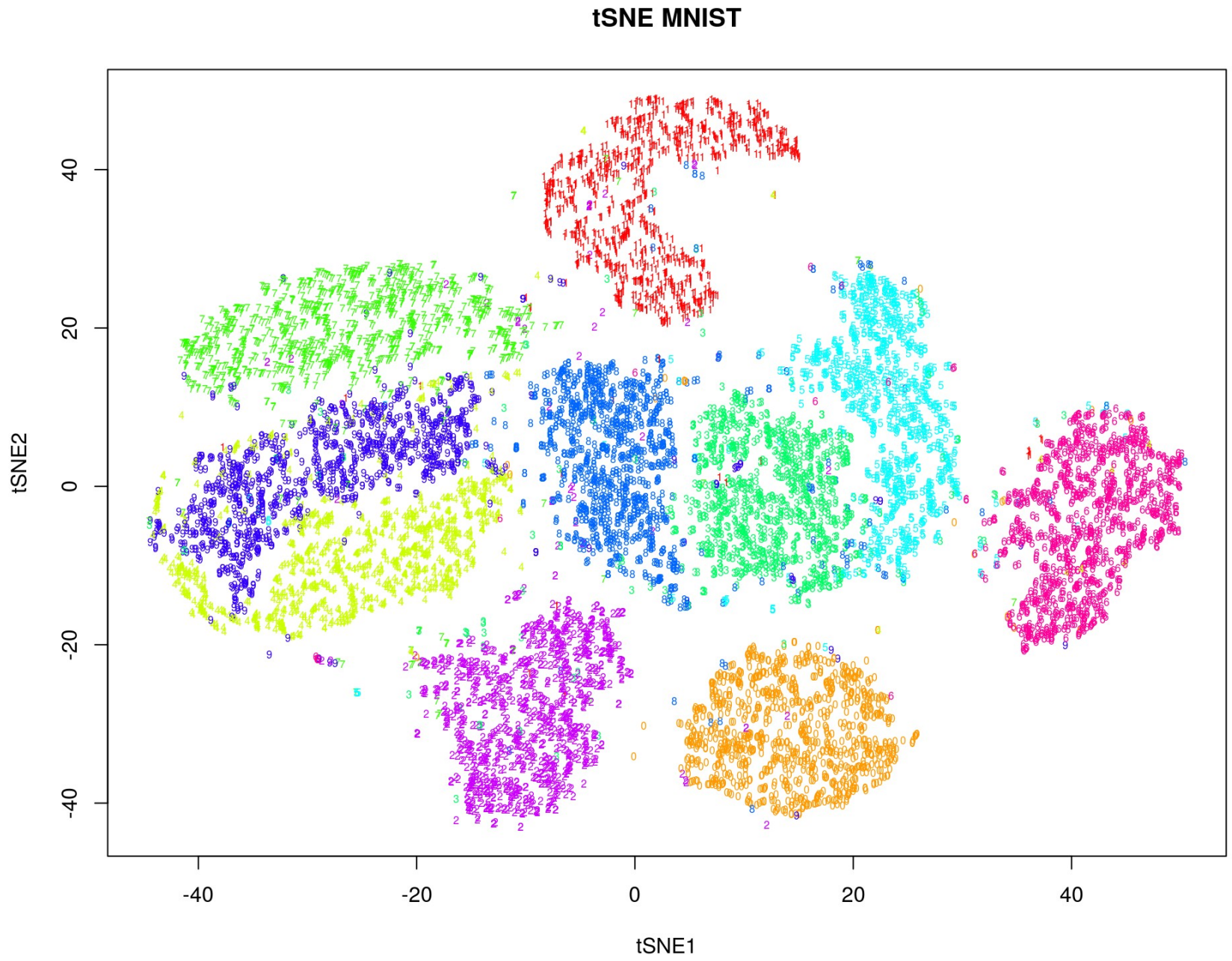
PCA on MNIST Data Set



t-distributed Stochastic Neighbor Embedding (tSNE)



tSNE on MNIST Data Set



Why We are Uncomfortable Using tSNE

tSNE does not scale for large data sets?

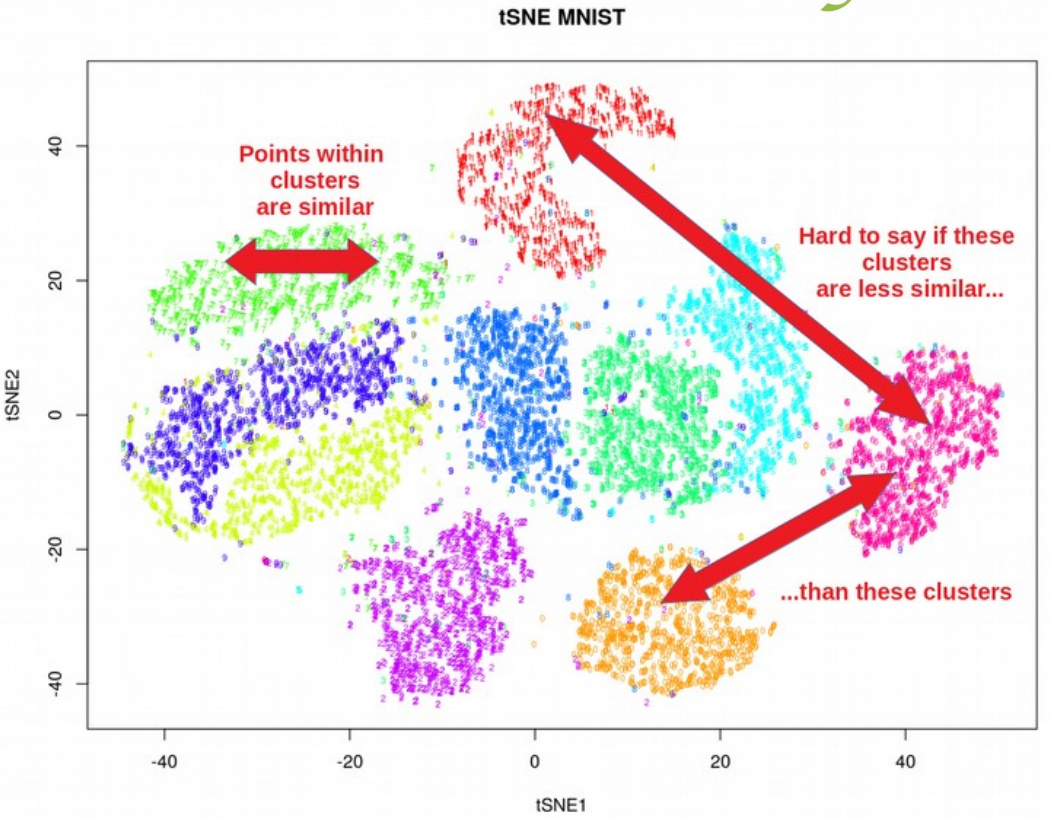
tSNE does not preserve global structure?

tSNE can only embed into 2-3 dims?

tSNE performs non-parametric mapping?

tSNE can not work with high-dimensional data directly (PCA needed)?

tSNE uses too much memory at large perplexities (FitSNE does not solve it)?



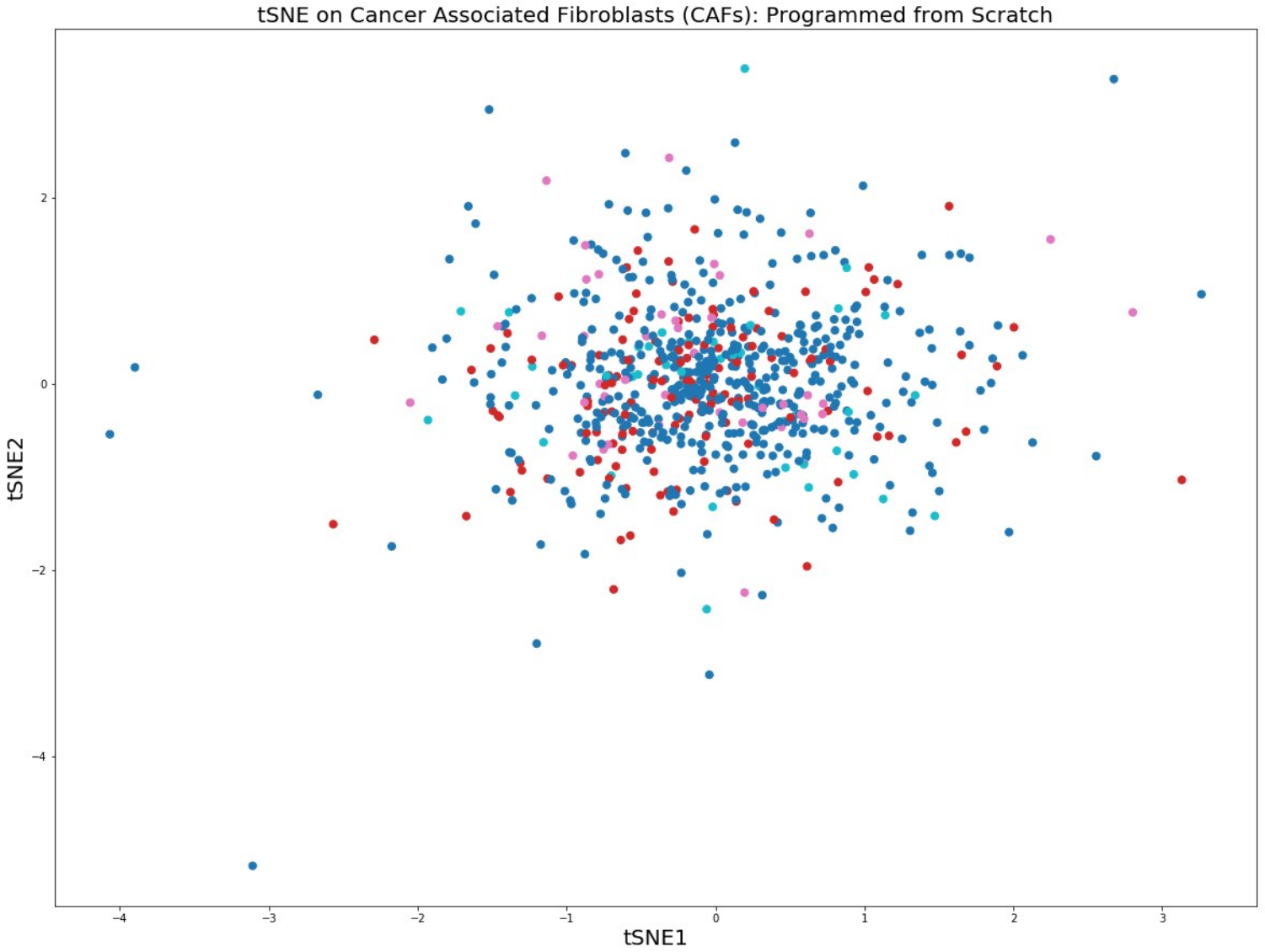
$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-||x_i - x_k||^2 / 2\sigma_i^2\right)}, \qquad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \qquad (1)$$

$$\text{Perplexity} = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}} \qquad (2)$$

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq l} \left(1 + ||y_k - y_l||^2\right)^{-1}} \qquad (3)$$

$$KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \qquad \frac{\partial KL}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + ||y_i - y_j||^2\right)^{-1} \qquad (4)$$

tSNE is Easy to Program from Scratch



How UMAP is Different from tSNE

UMAP uses local connectivity for high-dim probabilities

$$p_{i|j} = e^{-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}}$$

UMAP does not normalize probabilities (speed-up)

UMAP uses slightly different expression for nearest neighbors

$$k = 2^{\sum_i p_{ij}}$$

UMAP uses Laplacian Eigenmap for initialization

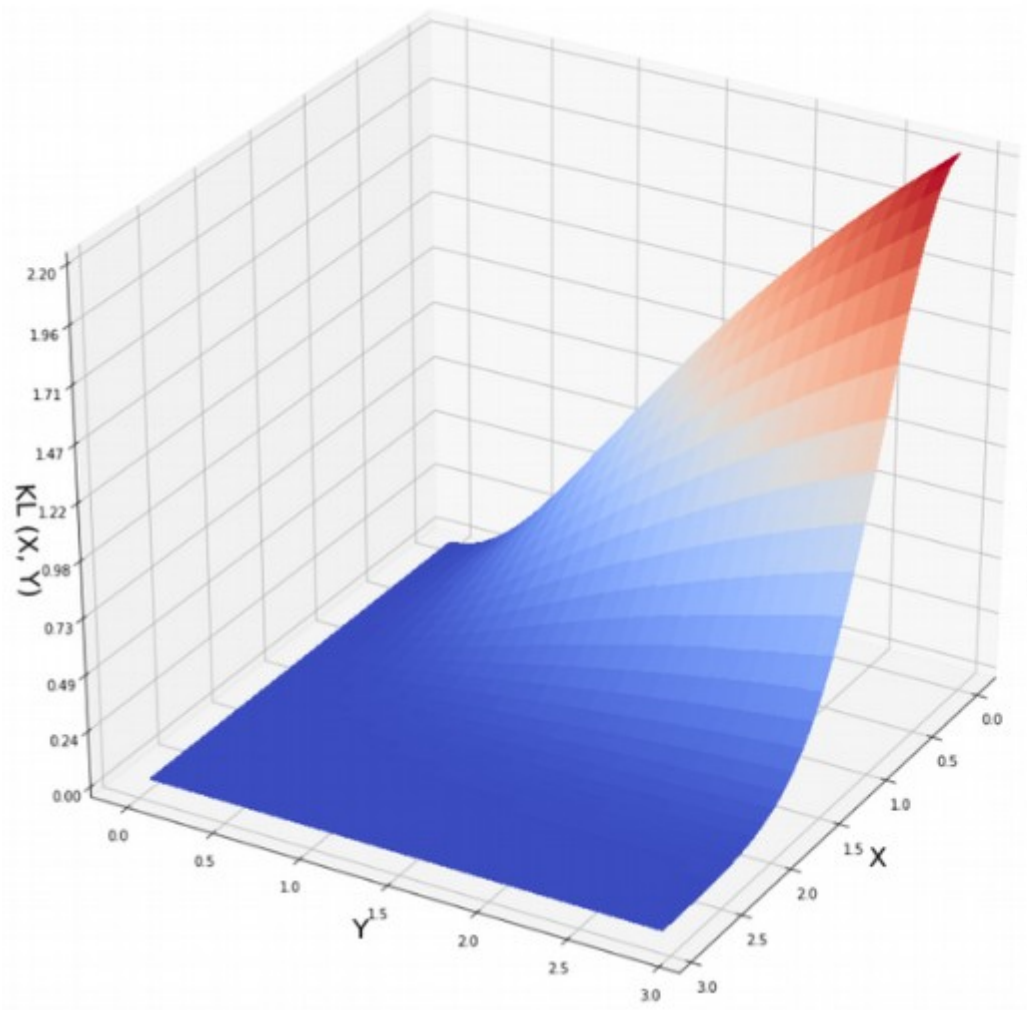
UMAP uses Cross-Entropy (not KL) as cost function

$$CE(X, Y) = \sum_i \sum_j \left[p_{ij}(X) \log\left(\frac{p_{ij}(X)}{q_{ij}(Y)}\right) + (1 - p_{ij}(X)) \log\left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)}\right) \right]$$

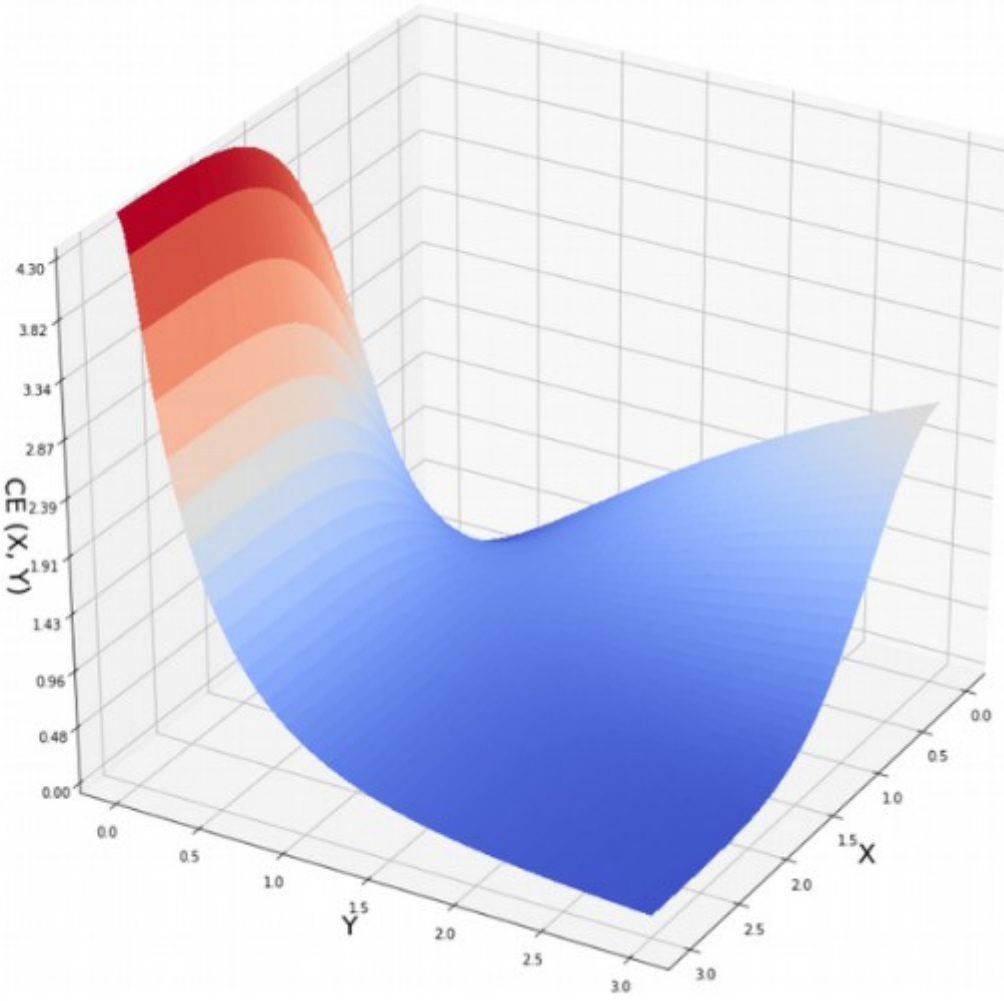
This is similar to tSNE cost function

This term is UMAP specific

Cost Function Makes UMAP Preserve More Global Structure than tSNE

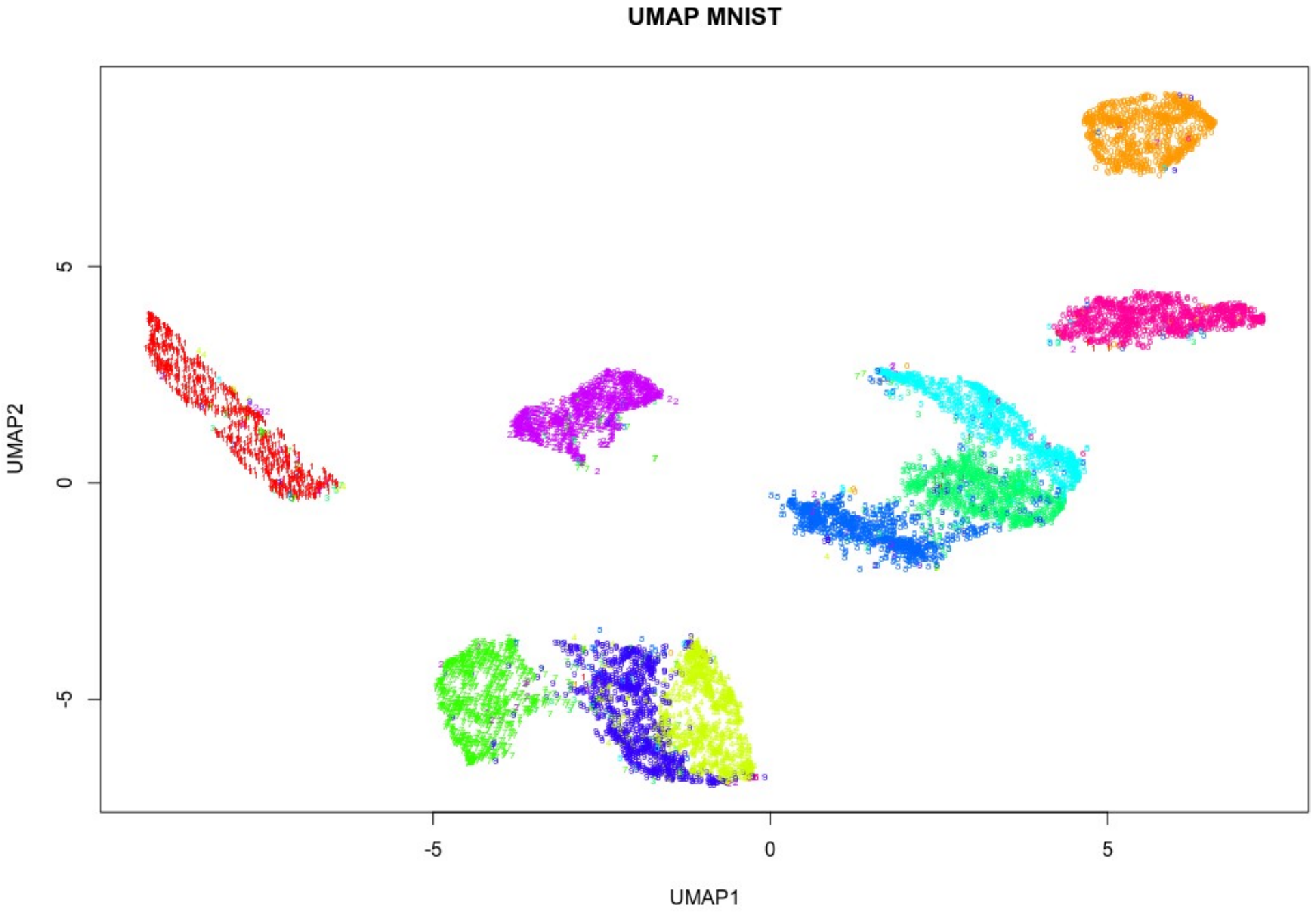


X → infinity, Y can be any

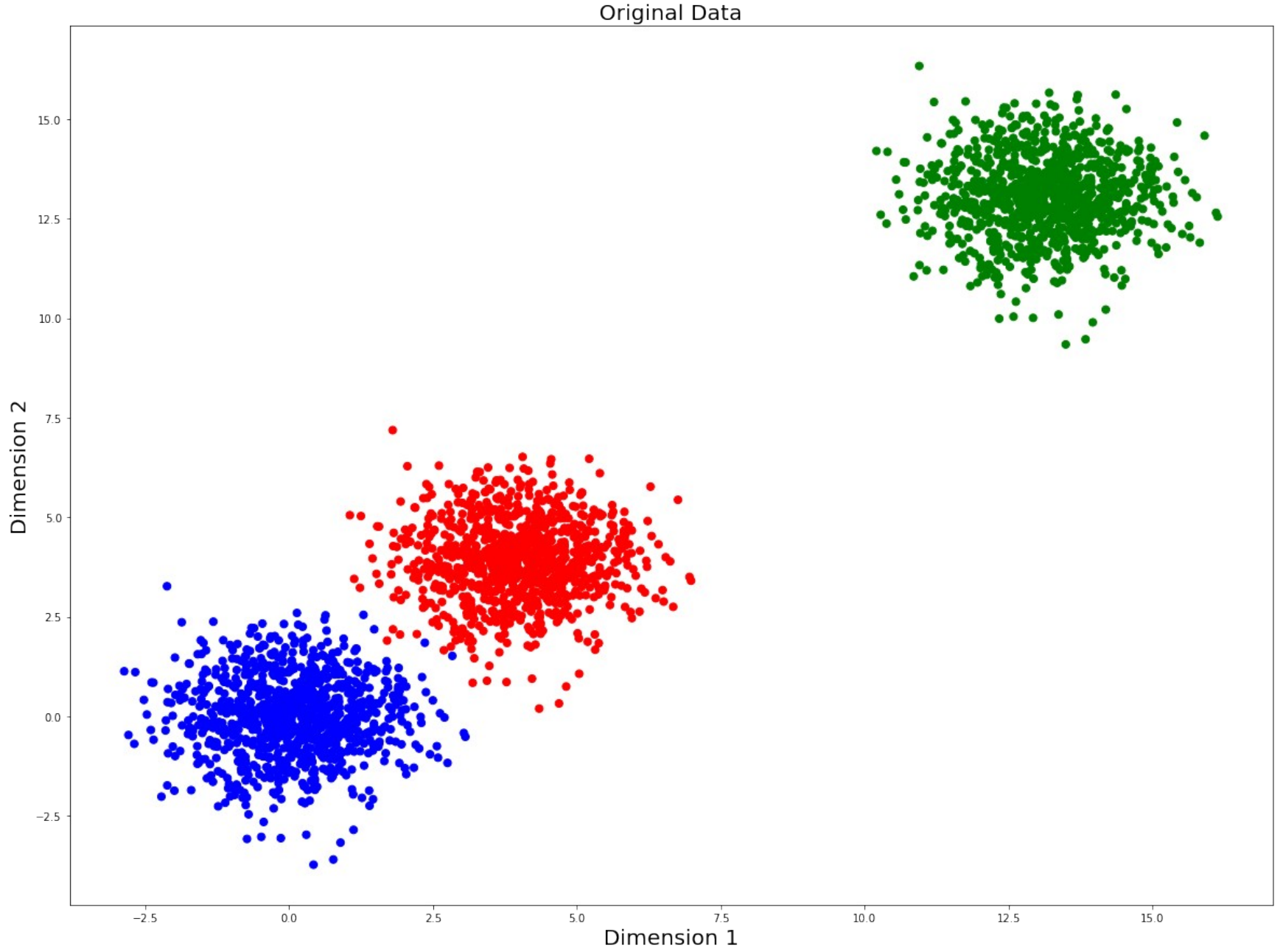


X → infinity, Y → infinity

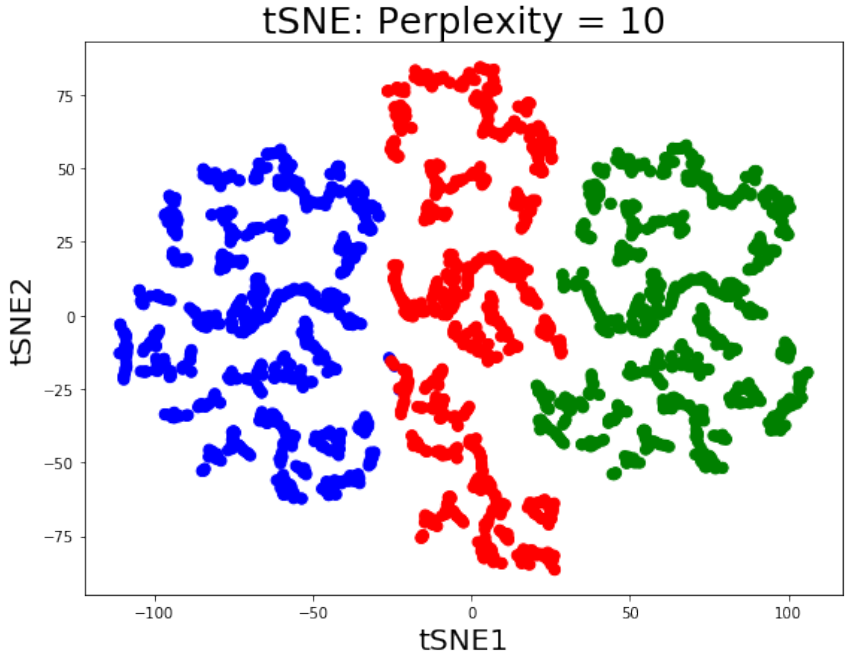
UMAP on MNIST Data Set

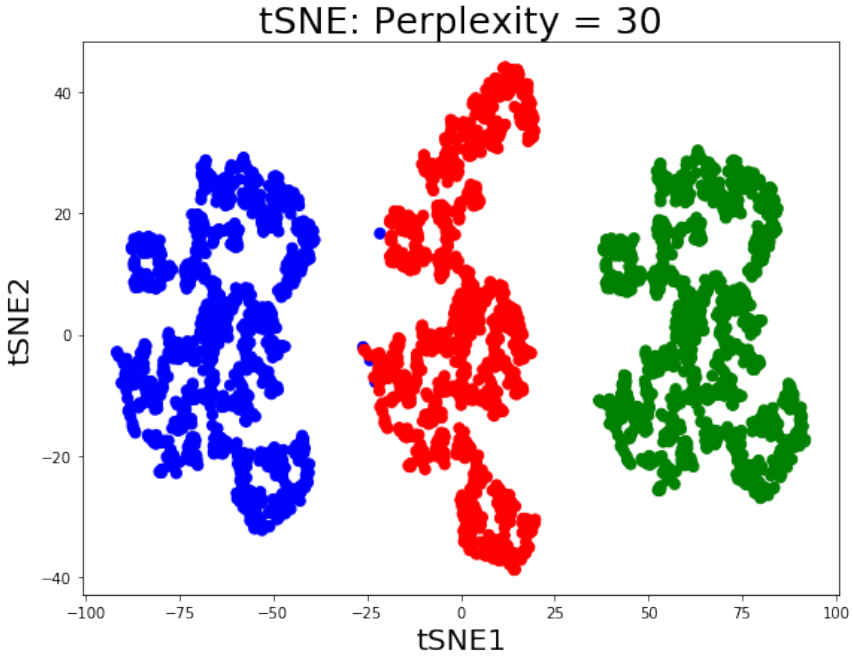


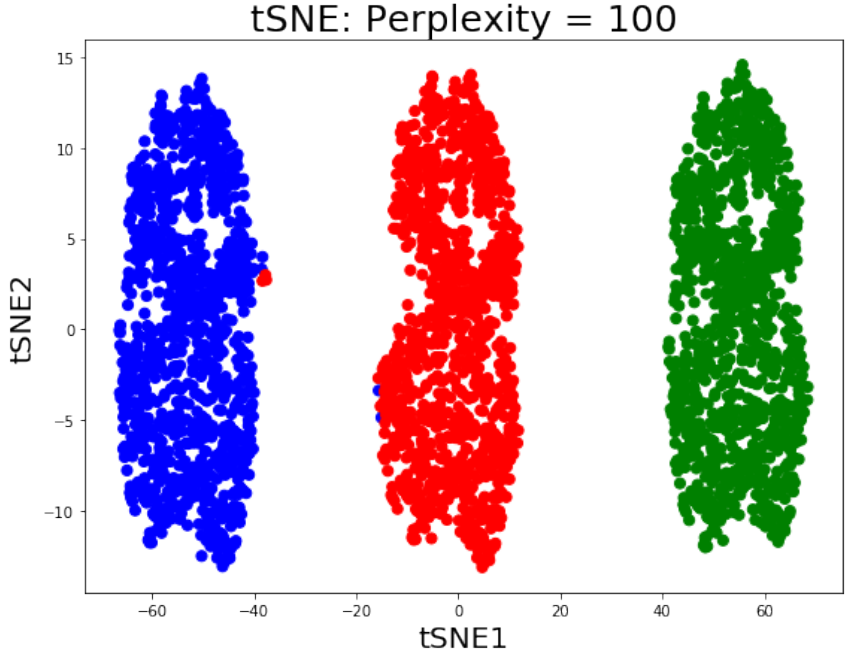
Why Global Structure Preservation is Important: Example #1

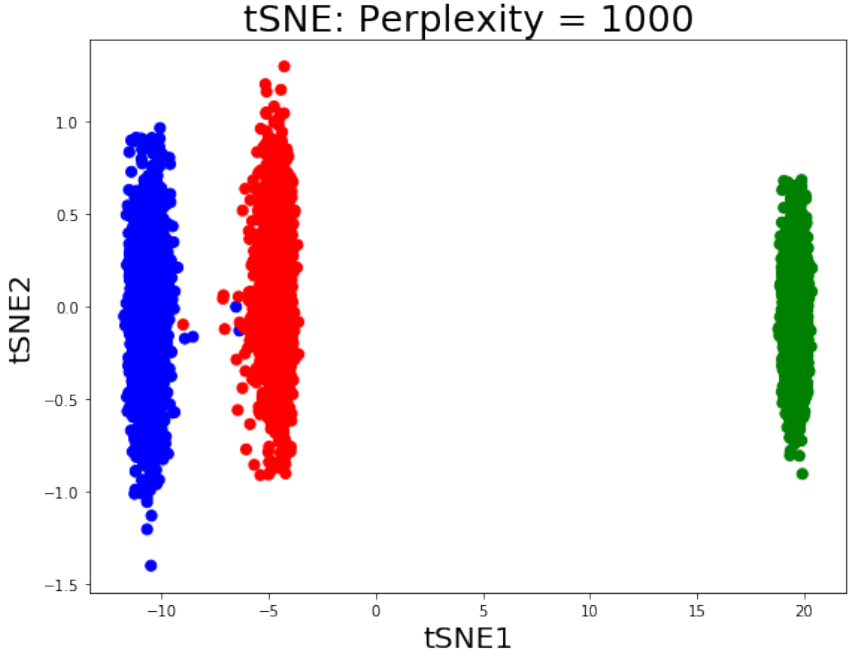


Why Global Structure Preservation is Important: Example #1

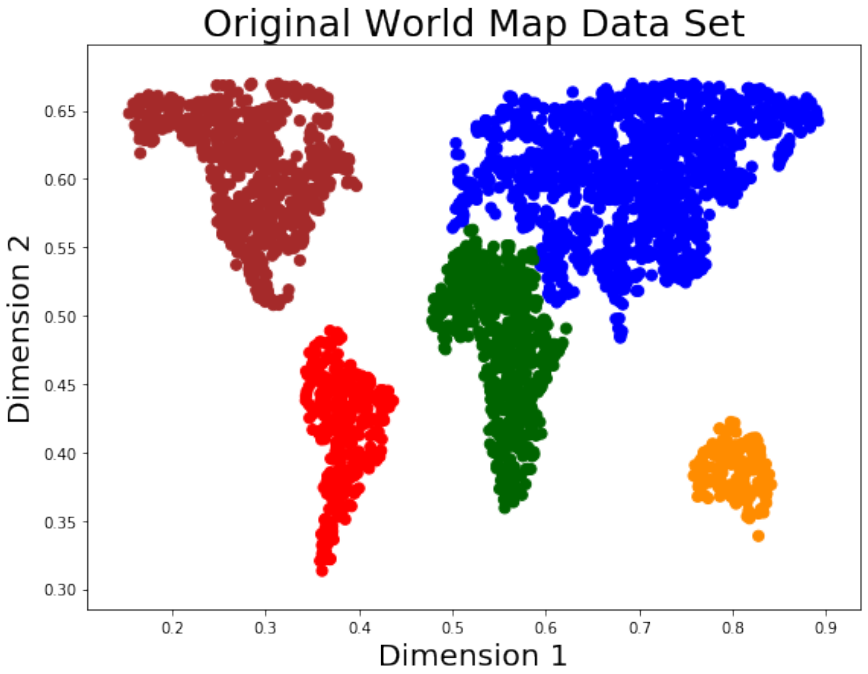


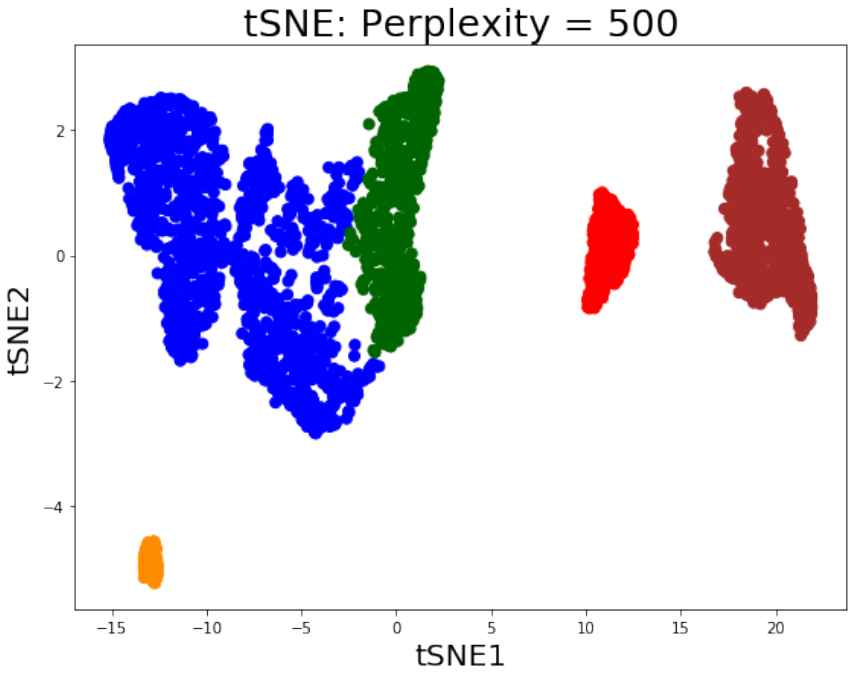


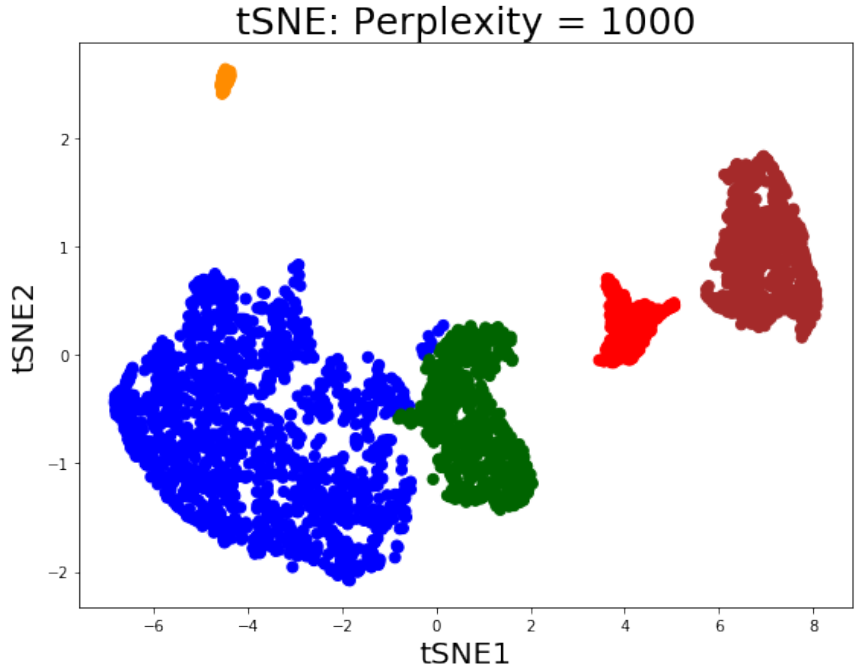


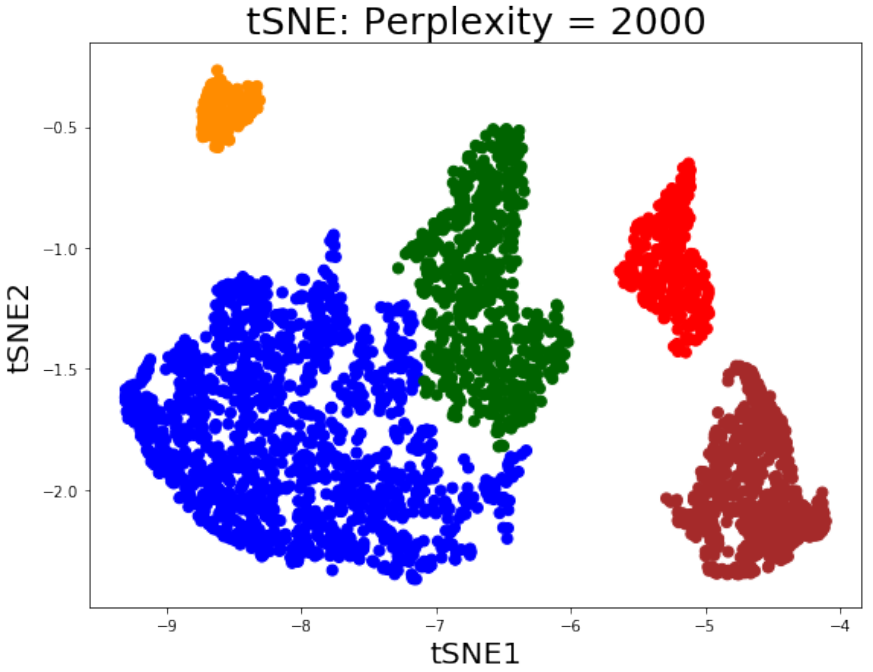


Why Global Structure Preservation is Important: Example #2

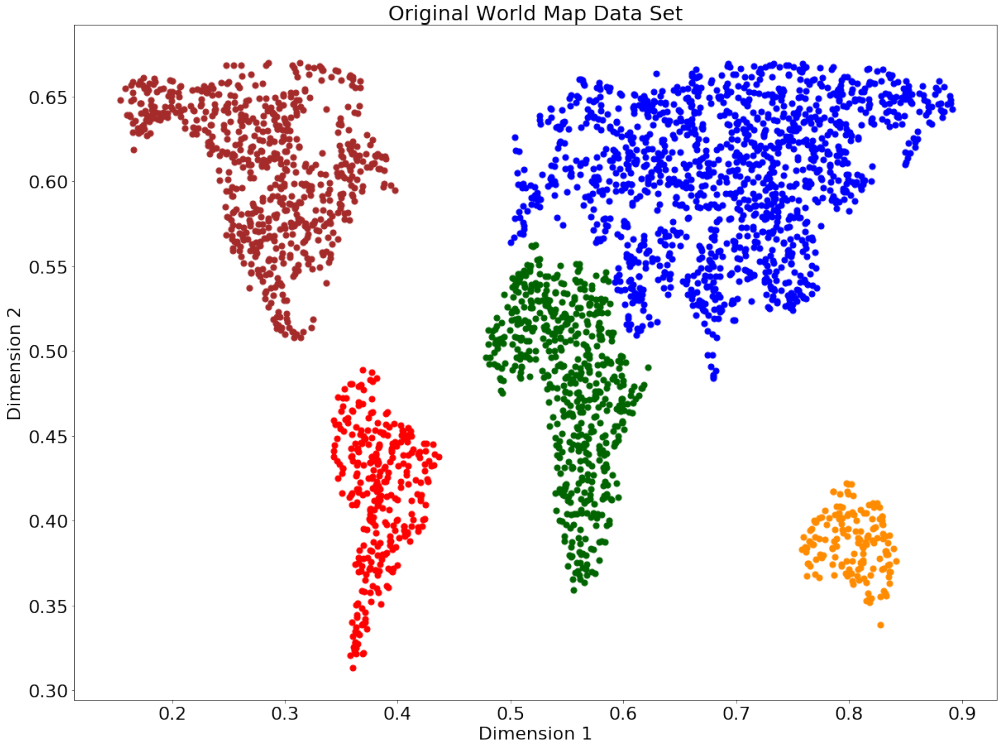


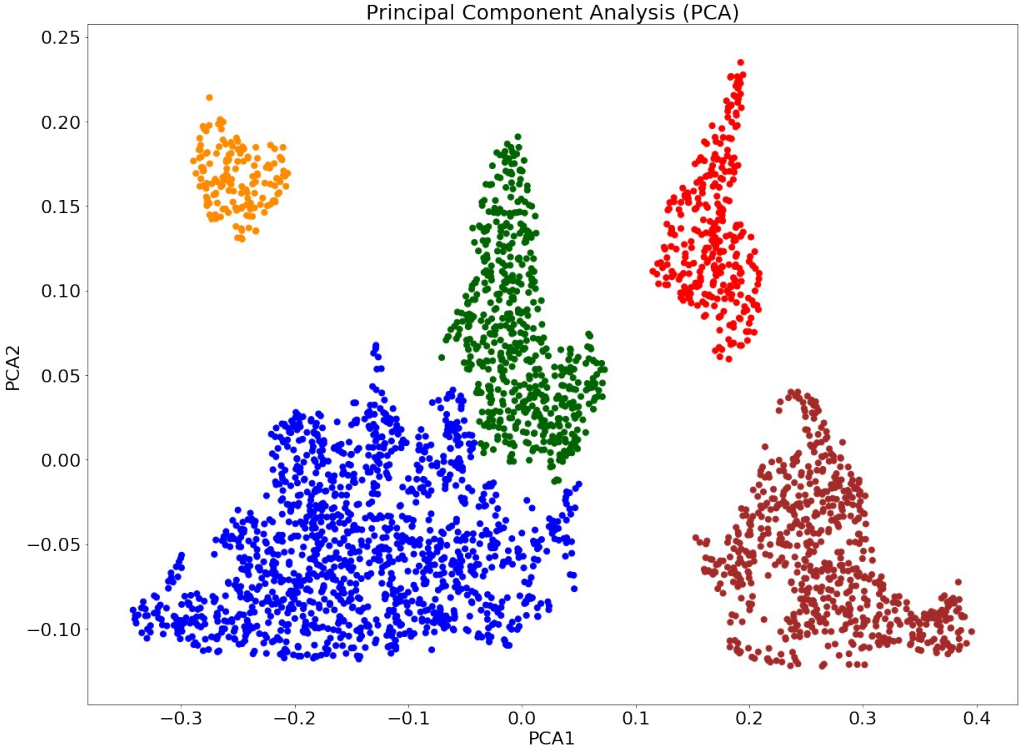


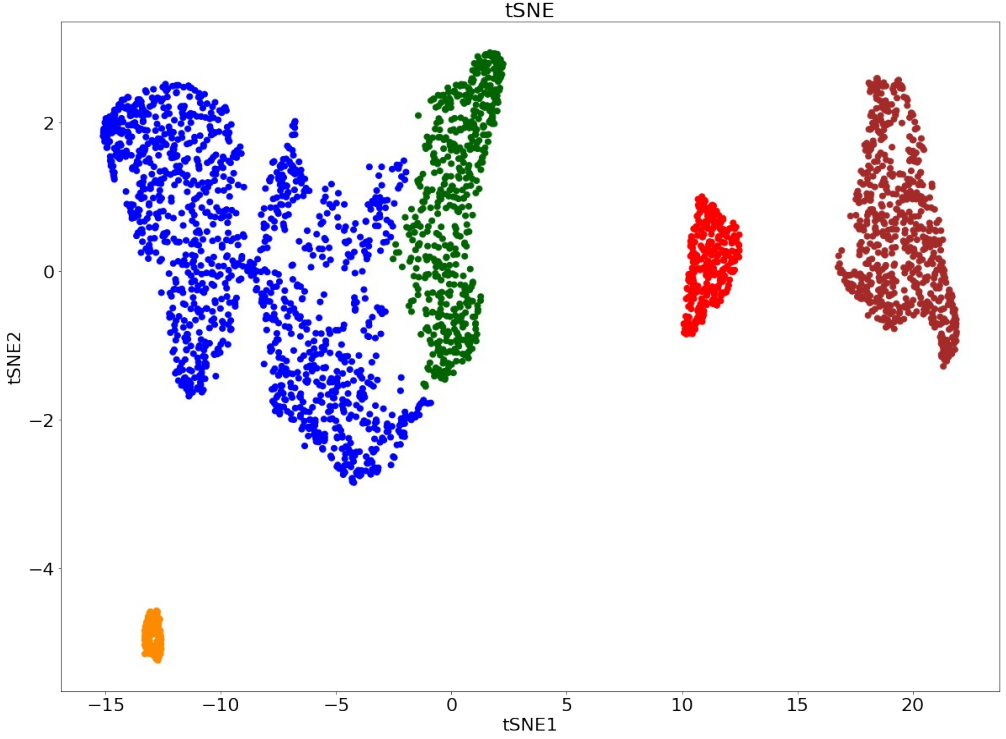


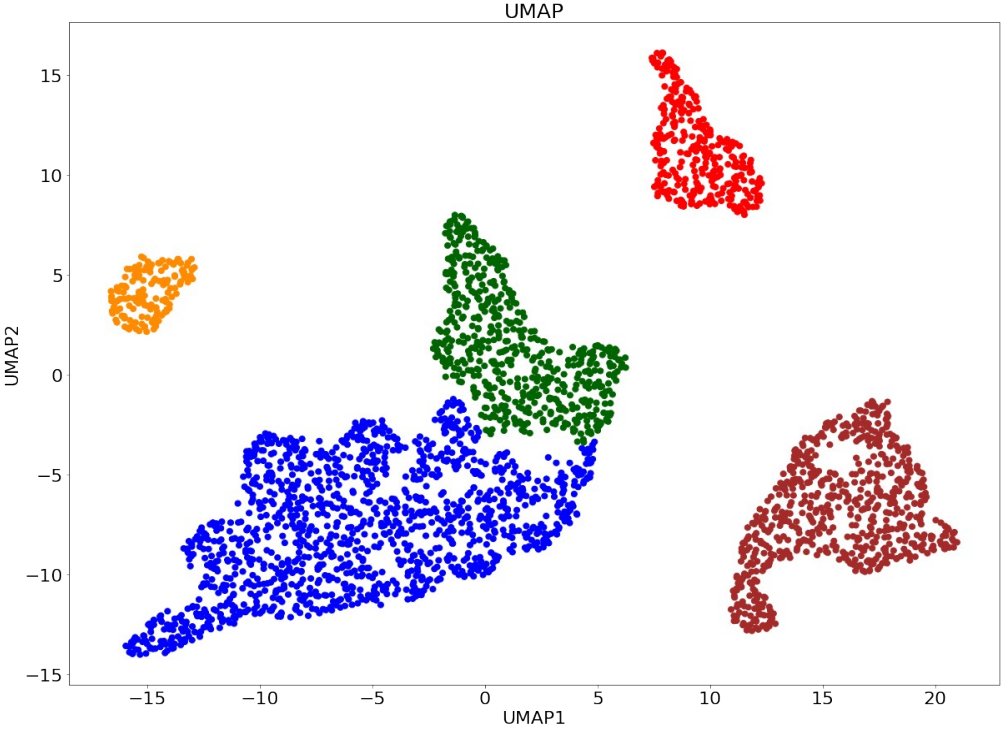


Global Structure Preservation for Linear Manifold

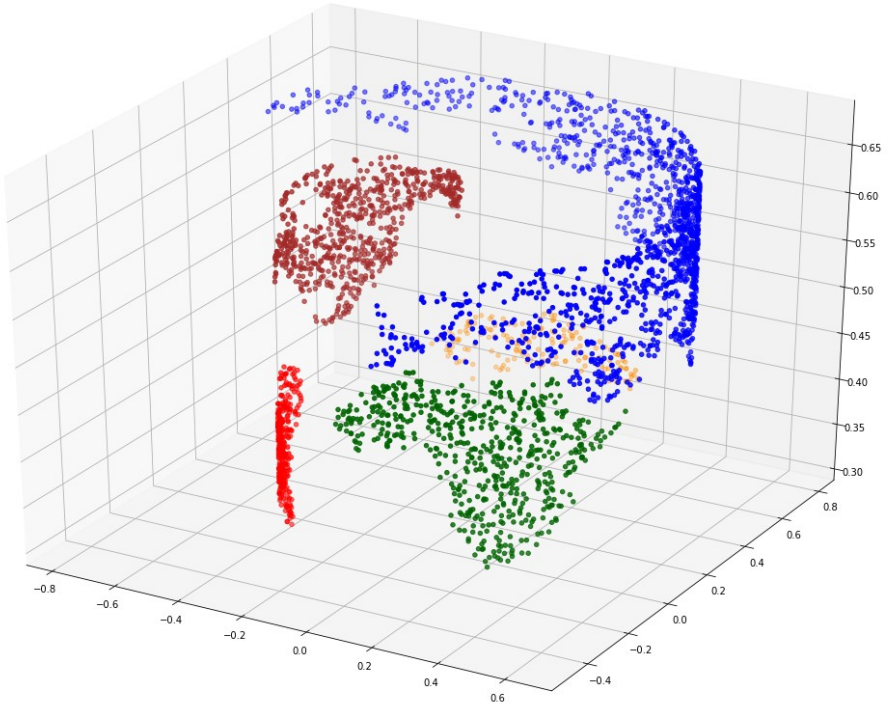


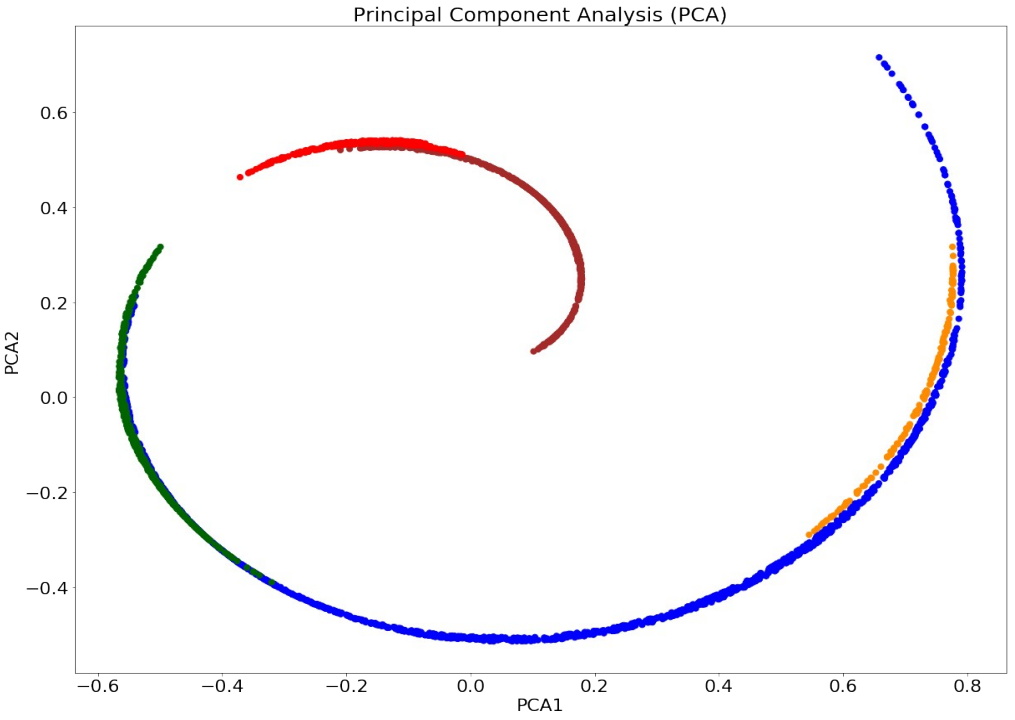


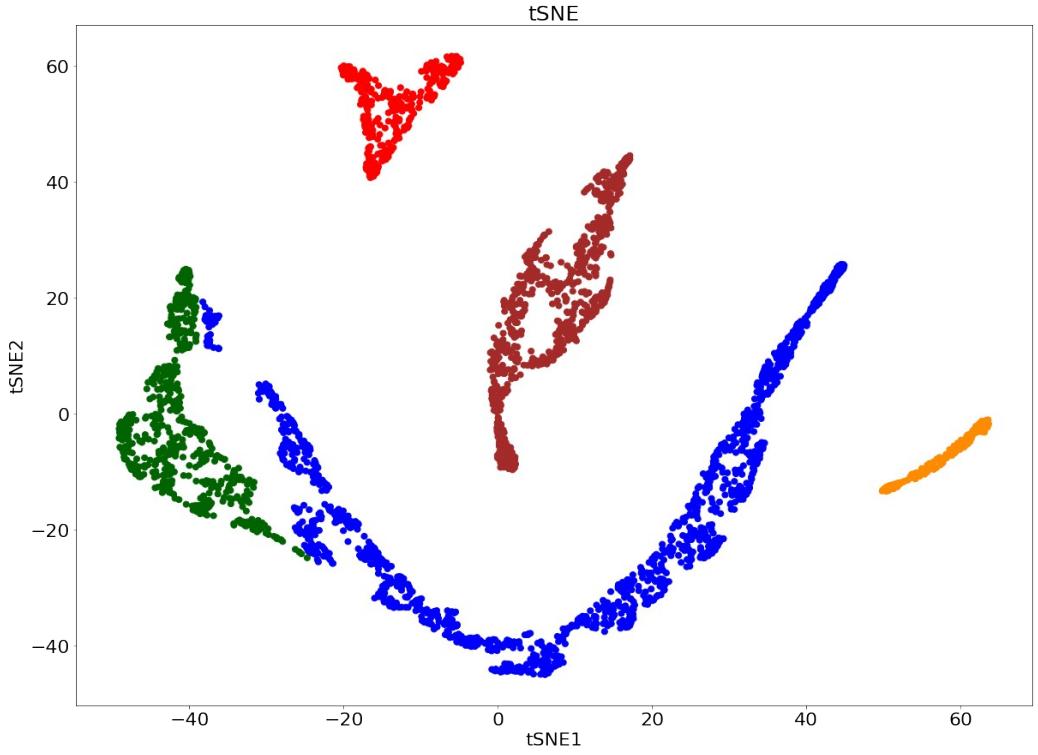


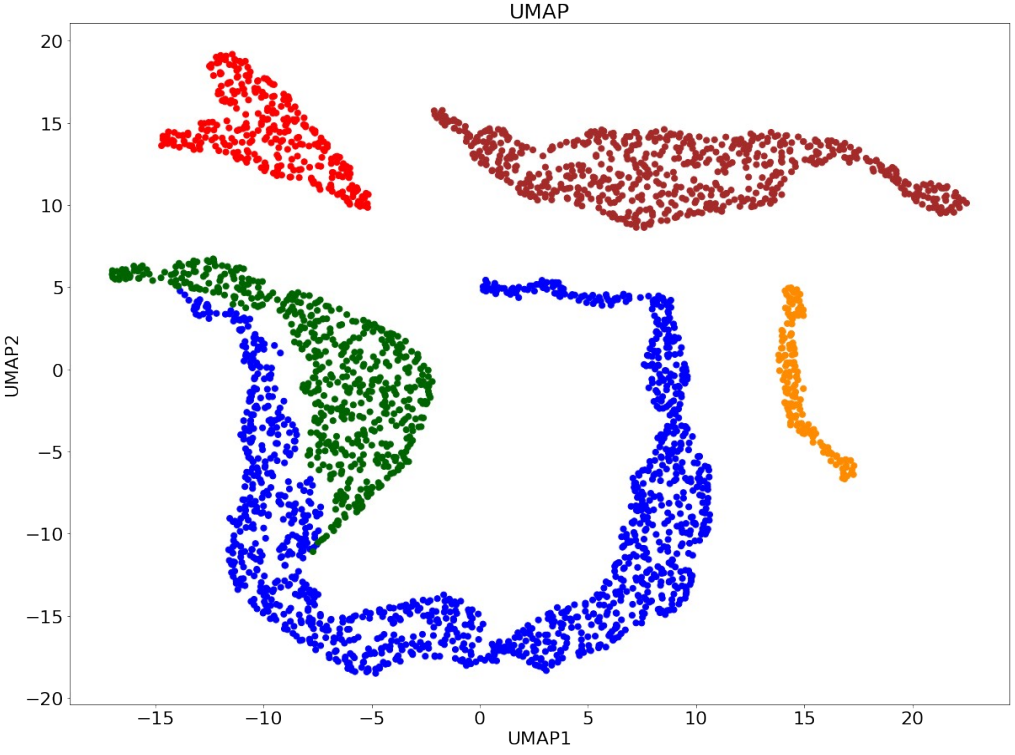


Global Structure Preservation for Non-Linear Manifold

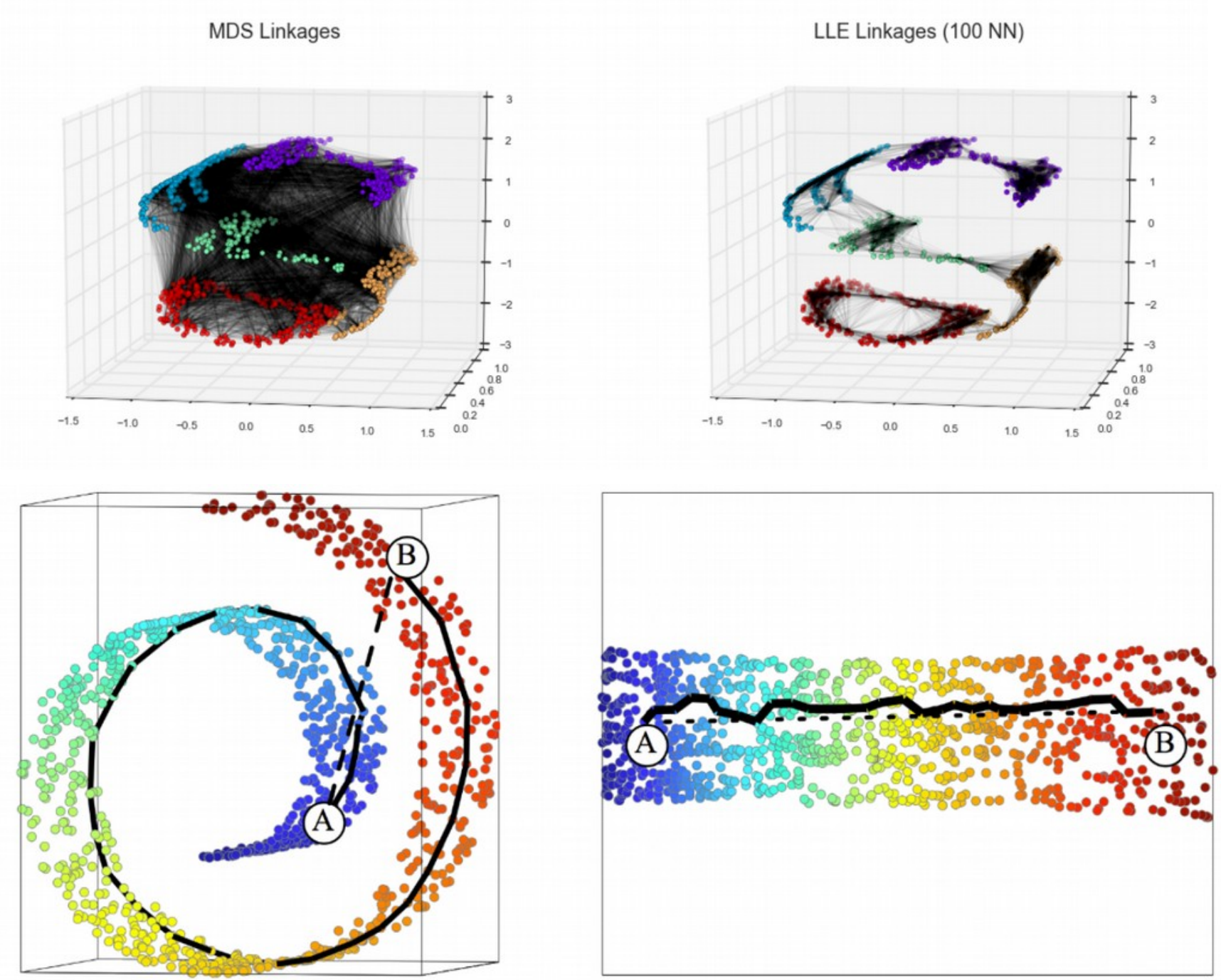




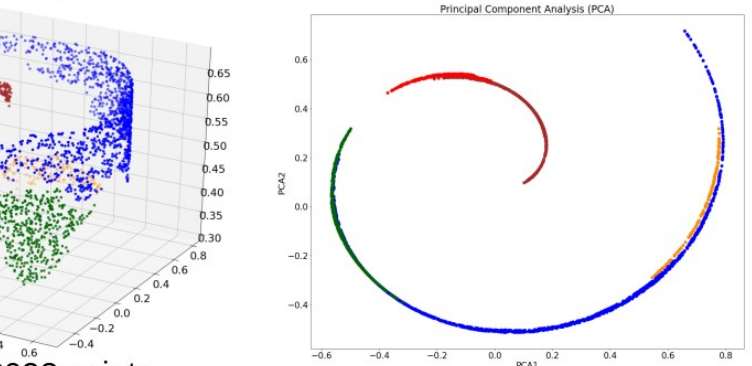
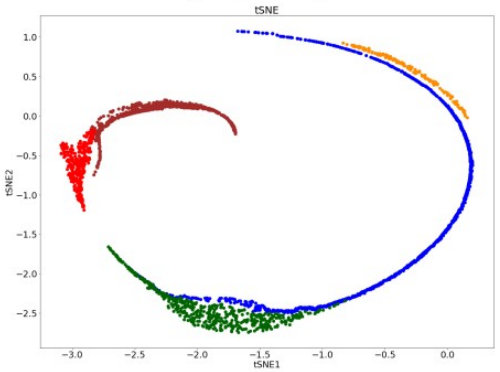
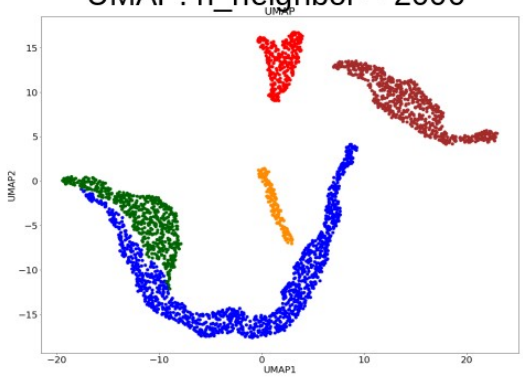
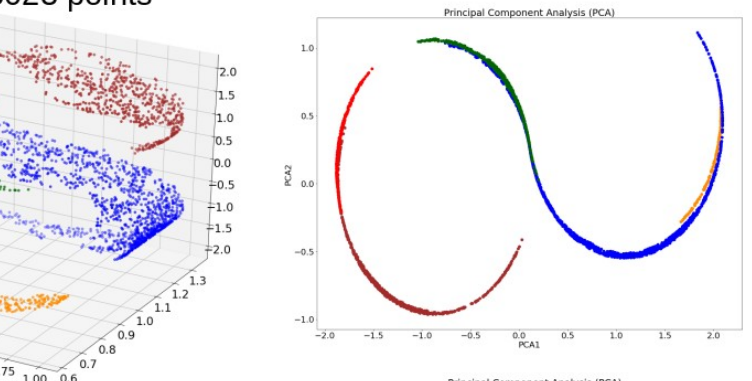
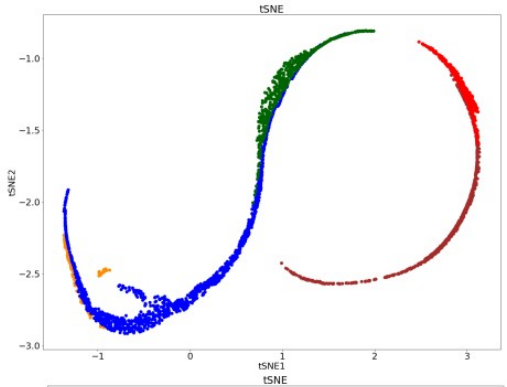
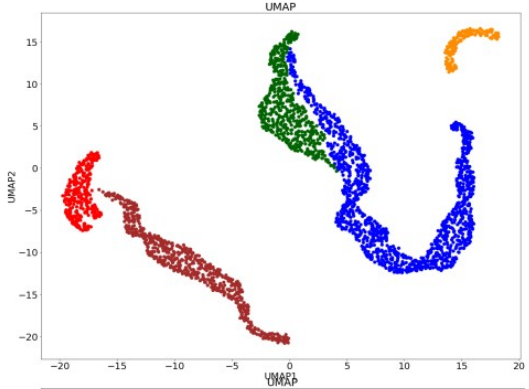
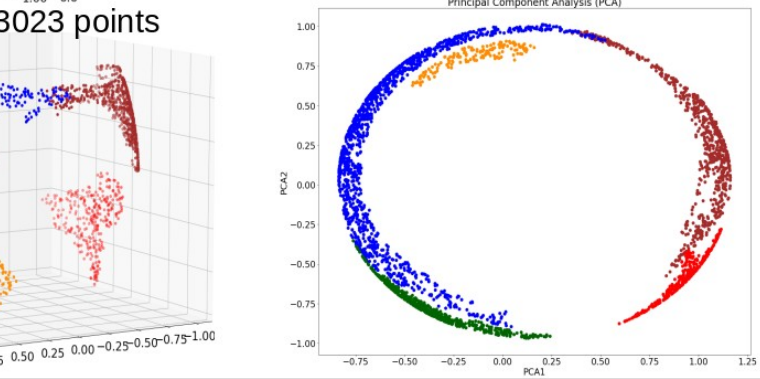
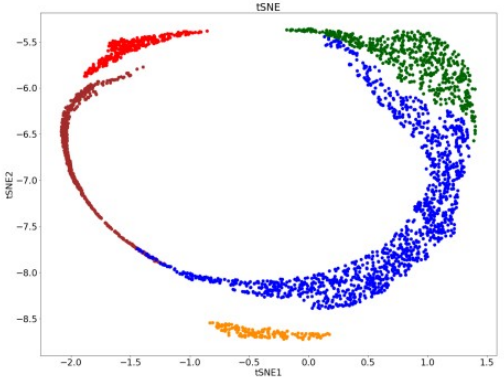
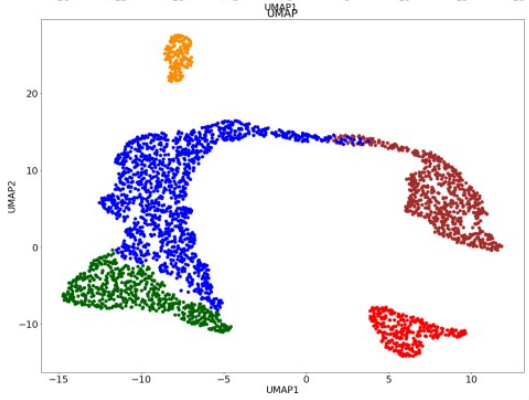




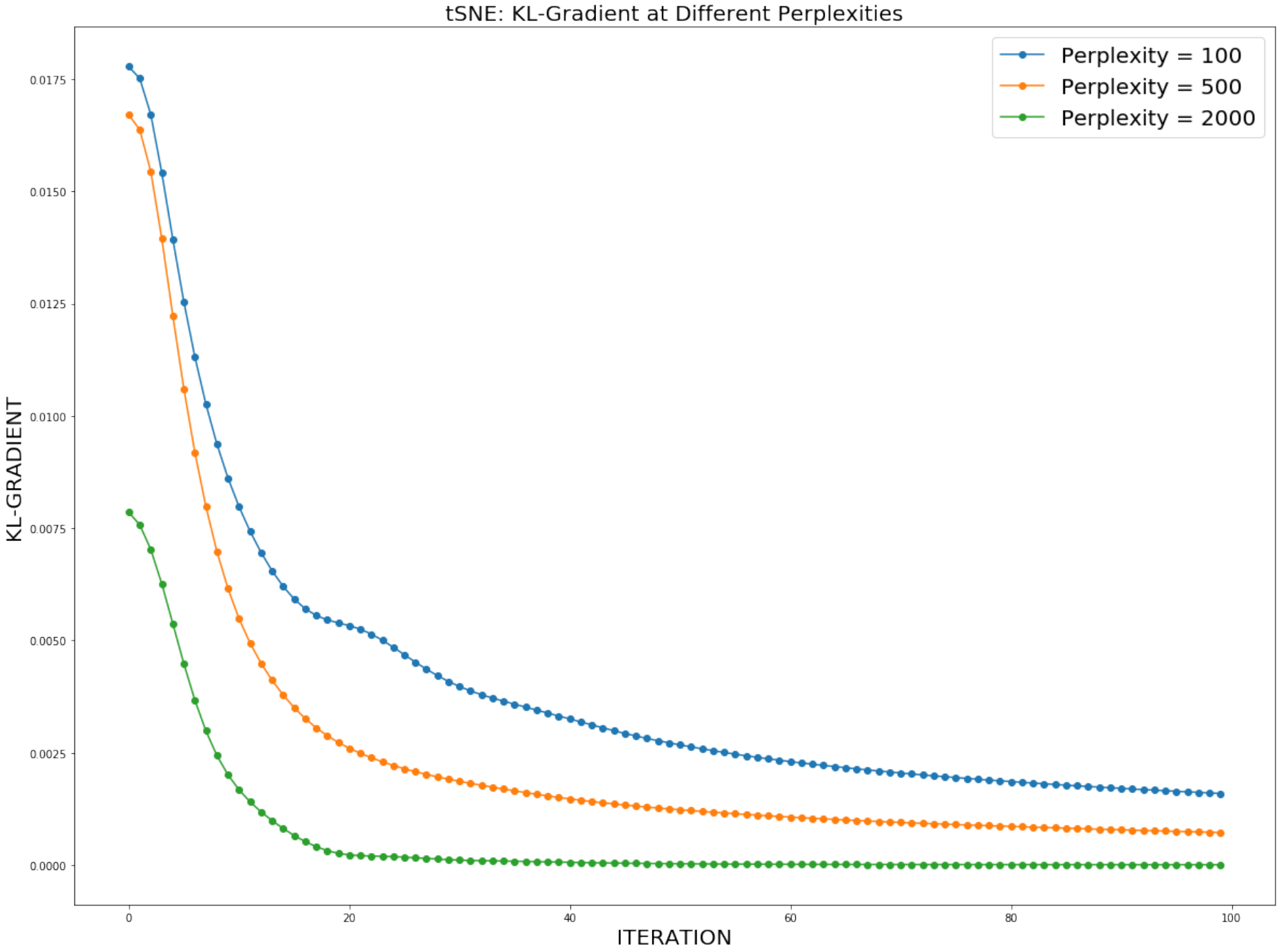
Why PCA Can't Unwrap Swiss Roll



Non-Linear Manifold: PCA vs. tSNE vs. UMAP

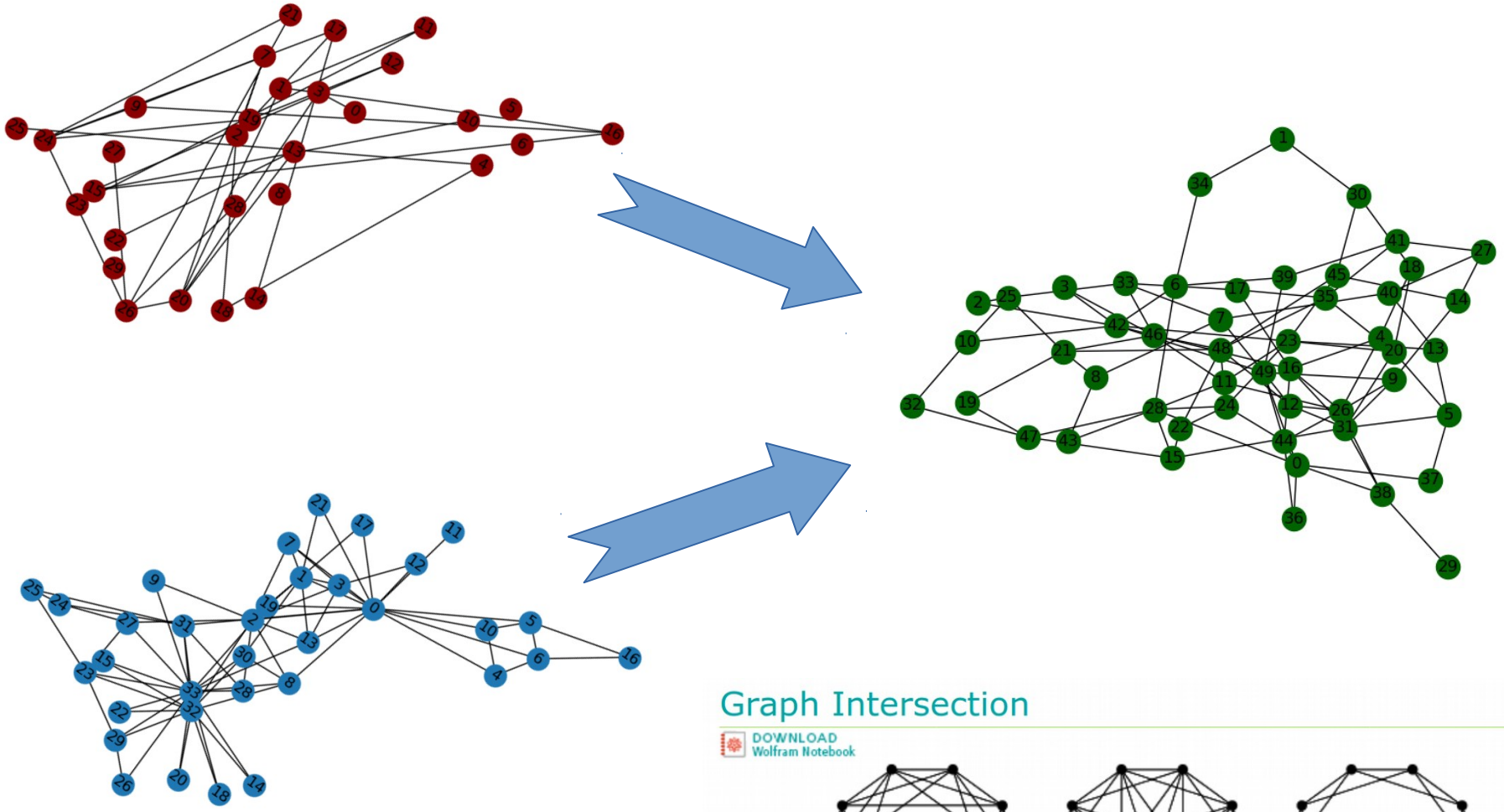
	Principal Component Analysis	tSNE: perplexity = 2000	UMAP: n_neighbor = 2000
Swiss Roll: 3023 points			
S-shape: 3023 points			
Sphere: 3023 points			

KL-Gradient Zero at Large Perplexity



UMAP for OMICs Integration

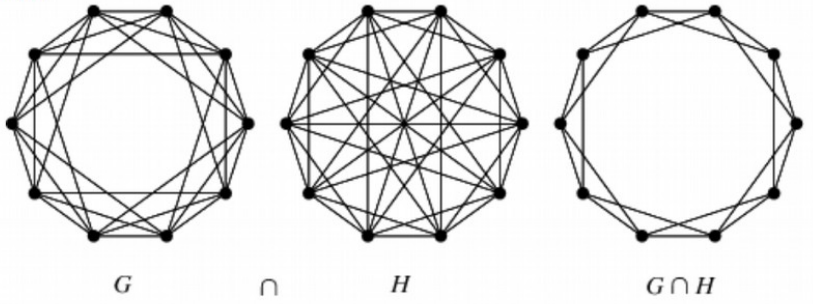
Graph-Based Omics Integration



Integration: keep edges consistently present across individual OMICs graphs

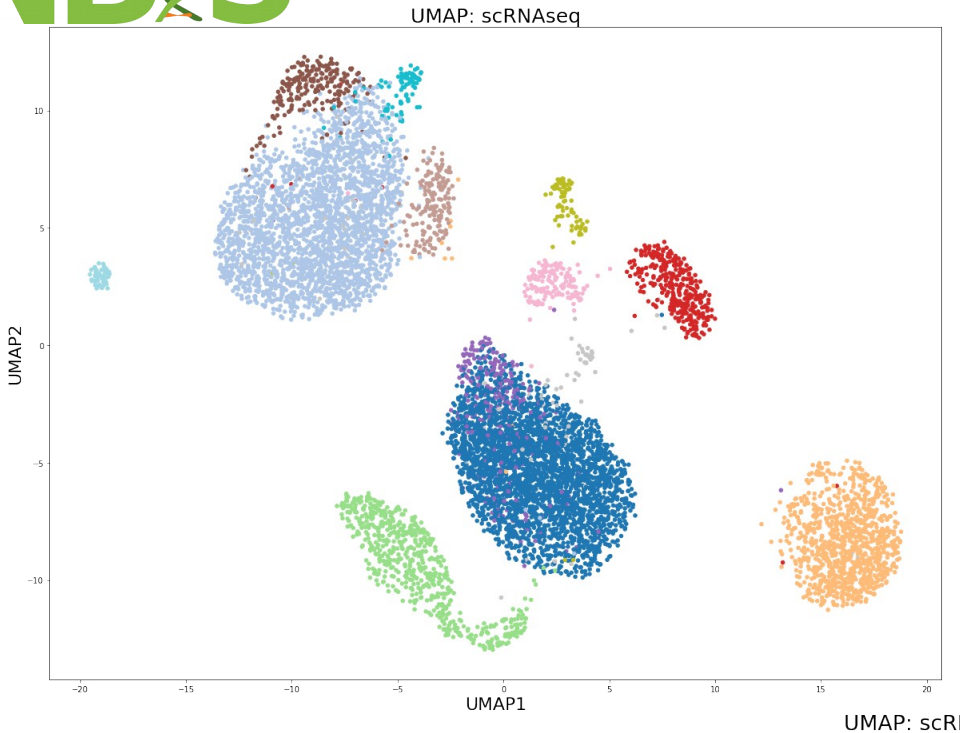
Graph Intersection

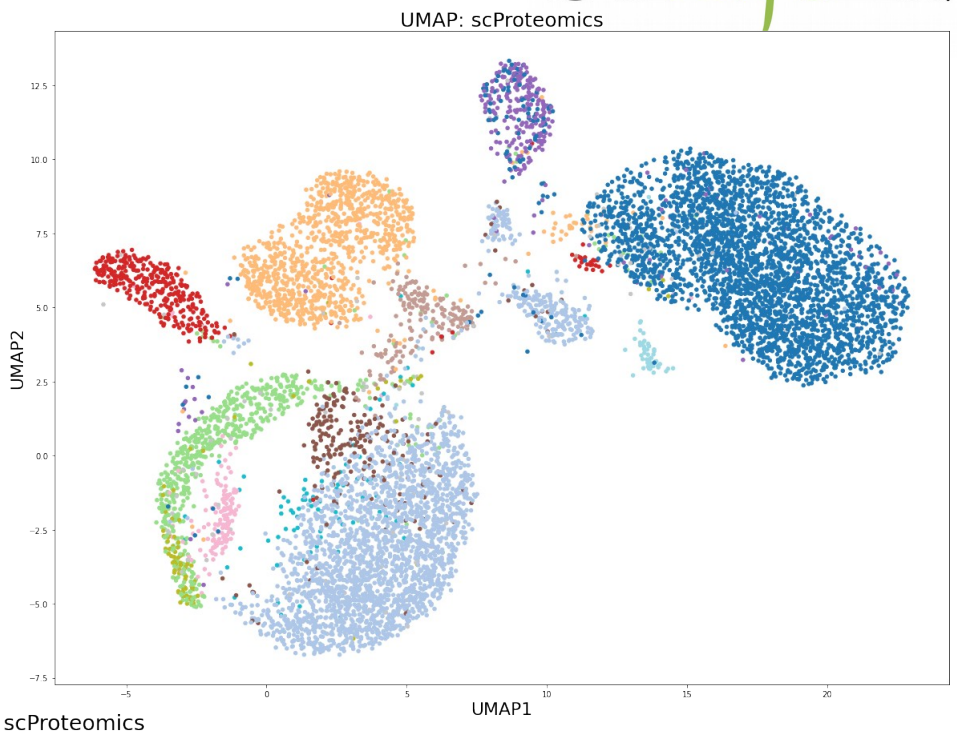
DOWNLOAD
Wolfram Notebook

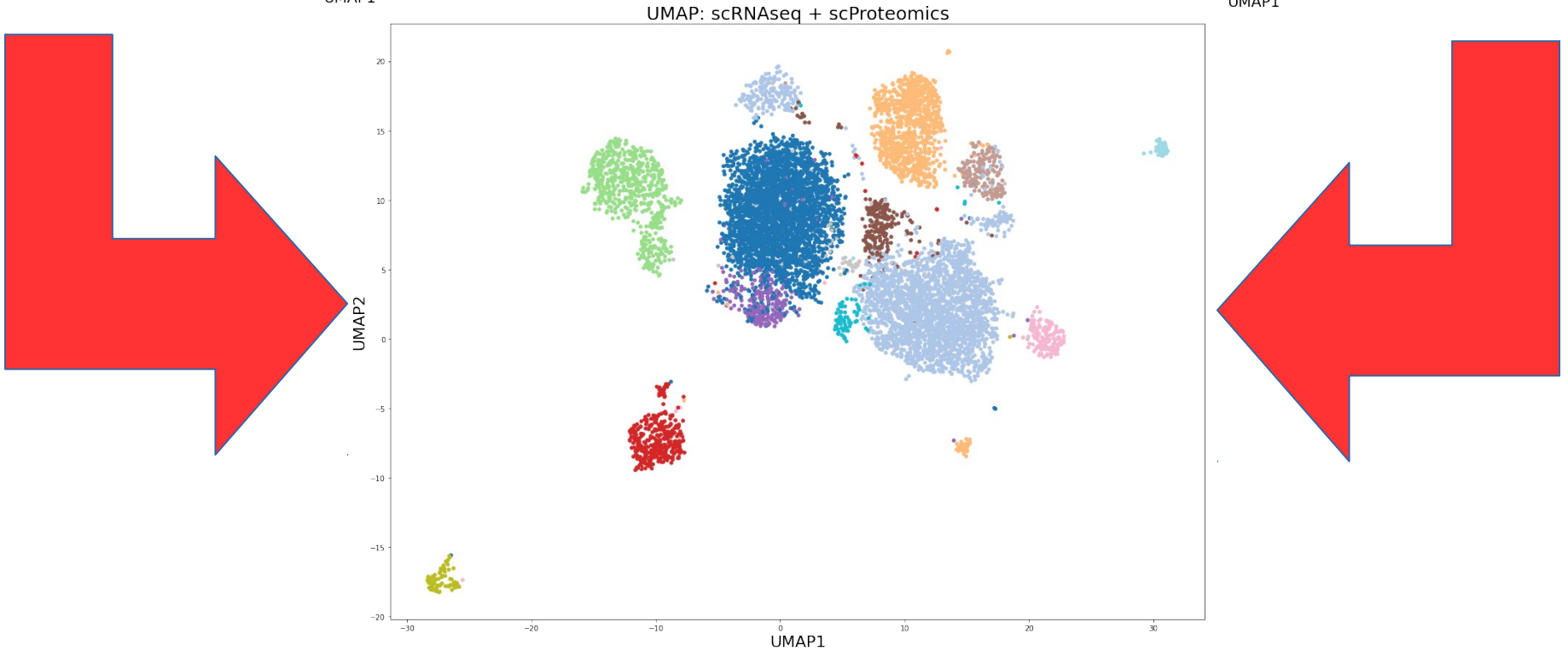


Let S be a set and $F = \{S_1, \ldots, S_p\}$ a nonempty family of distinct nonempty subsets of S whose union is $\bigcup_{i=1}^{p} S_i = S$. The intersection graph of F is denoted $\Omega(F)$ and defined by $V(\Omega(F)) = F$, with S_i and S_j adjacent whenever $i \neq j$ and $S_i \bigcap S_j \neq \emptyset$. Then a graph G is an intersection graph on S if there exists a family F of subsets for which G and $\Omega(F)$ are isomorphic graphs (Harary 1994, p. 19). Graph intersections can be computed in the Wolfram Language using `GraphIntersection`[*g*, *h*].

UMAP for Omics Integration







National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

Lunds Universitet