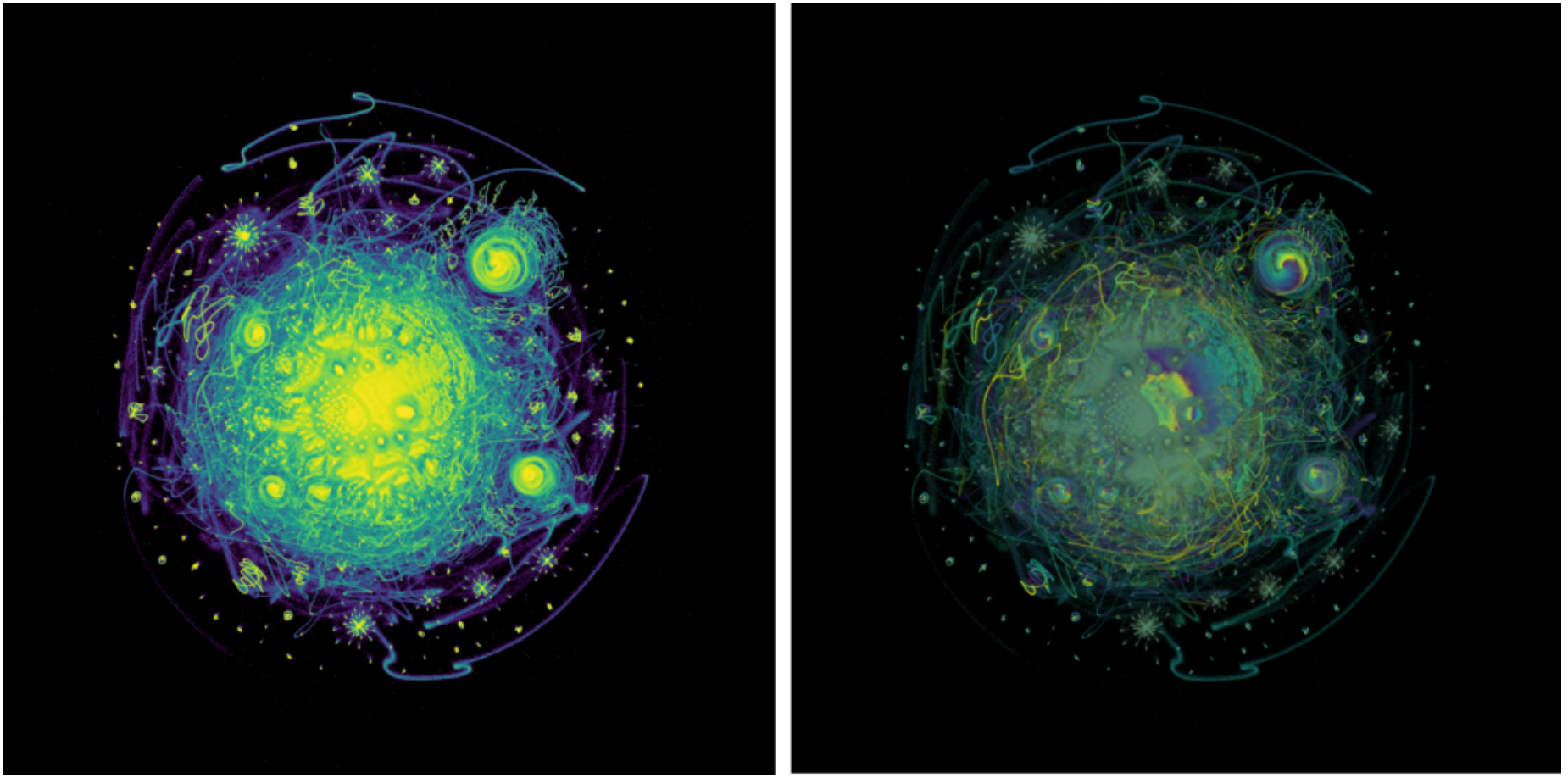
Dimension Reduction for OMICs Integration

OMICs Integration and Systems Biology course

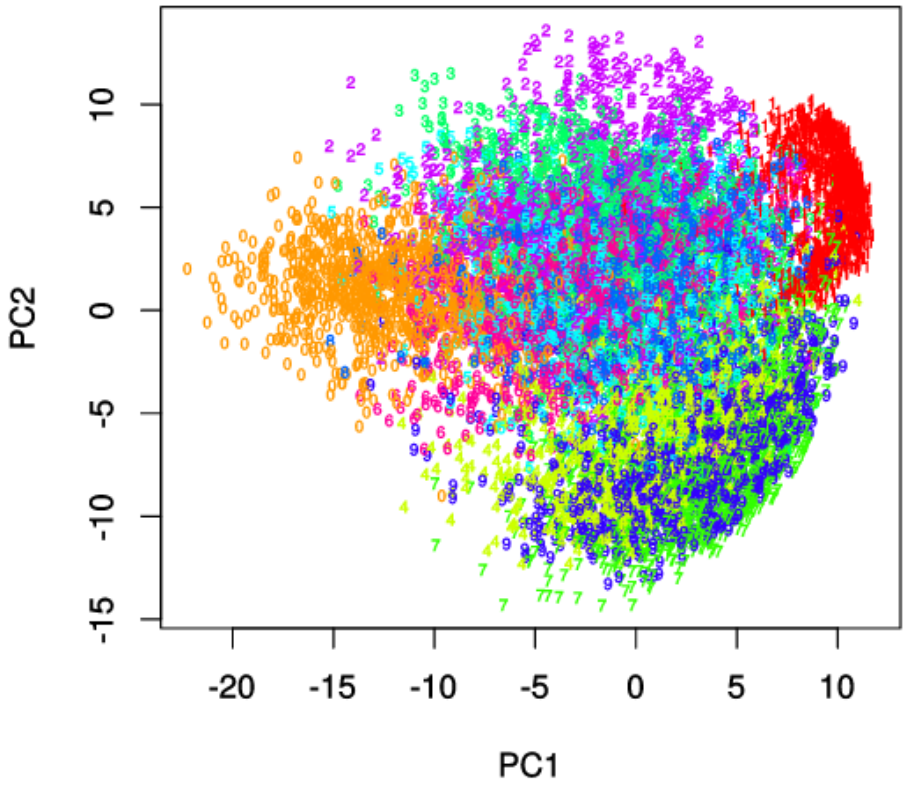
Nikolay Oskolkov, NBIS SciLifeLab

Lund, 5.10.2020

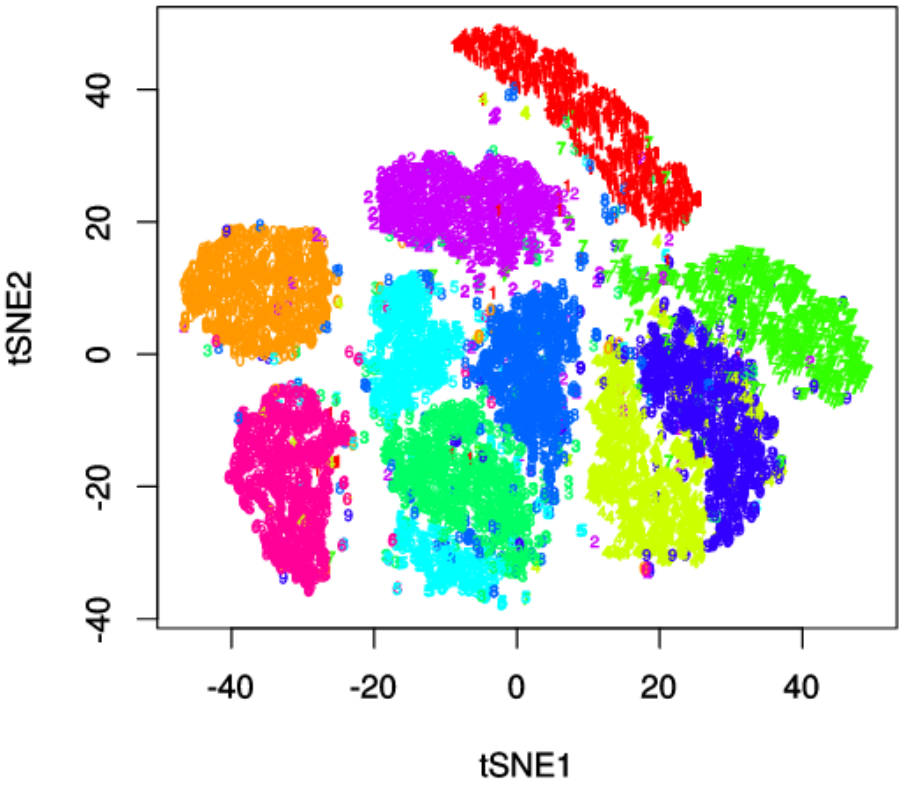


Dimension Reduction and Visualization

PCA PLOT WITH PRCOMP



tSNE MNIST



Dimension reduction is not 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

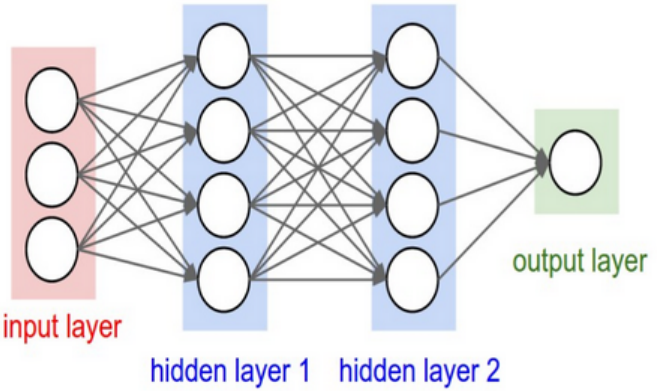
Bayesianism

P >> N

Frequentism

P ~ N

Deep Learning



P << N

Amount of Data

Ex.1

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

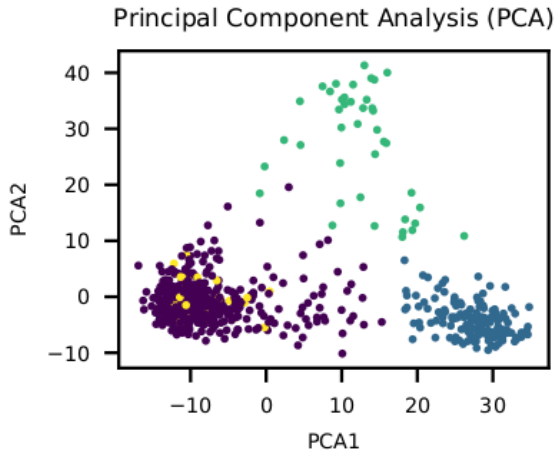
$$\beta = (X^T X)^{-1} X^T Y$$

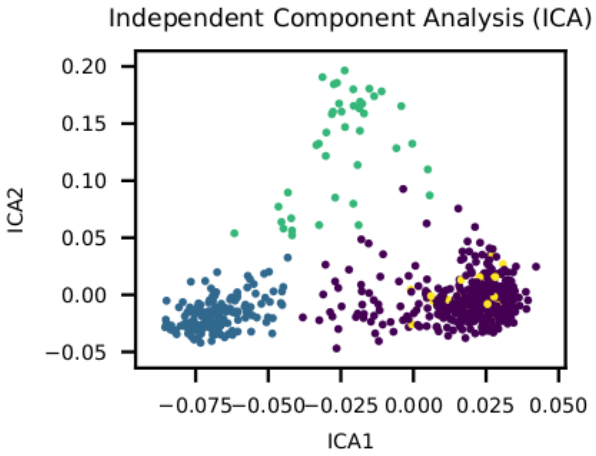
$$(X^T X)^{-1} \sim \frac{1}{\det(X^T X)} \cdots \to \infty, \quad n << p$$

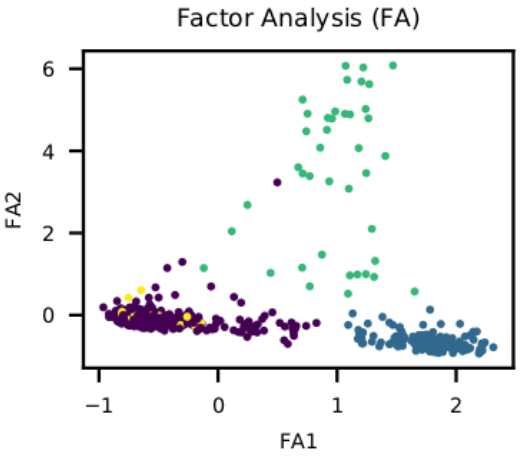
Ex.2 $$E[\hat{\sigma}^2] = \frac{n-p}{n}\sigma^2$$

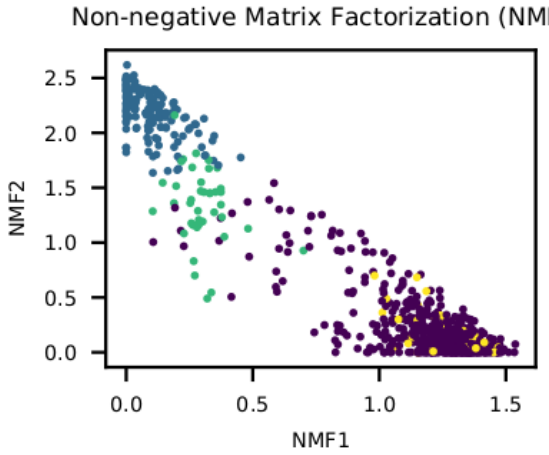
Biased ML variance estimator in HD-space

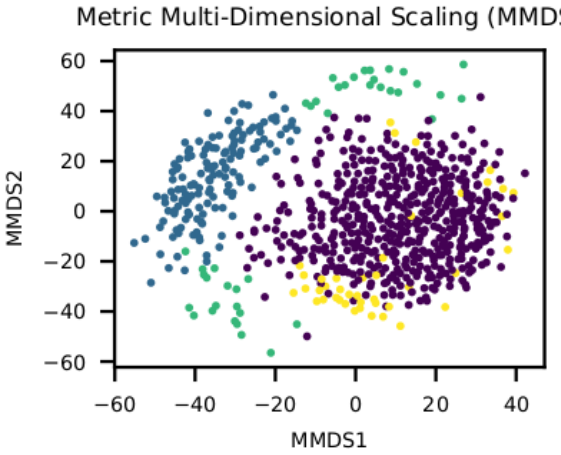
Linear Dimension Reduction

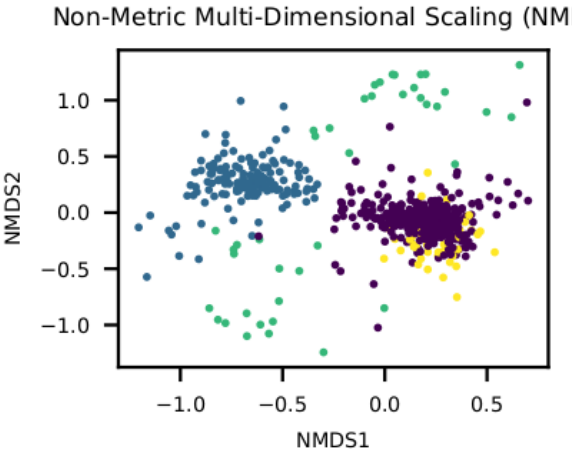


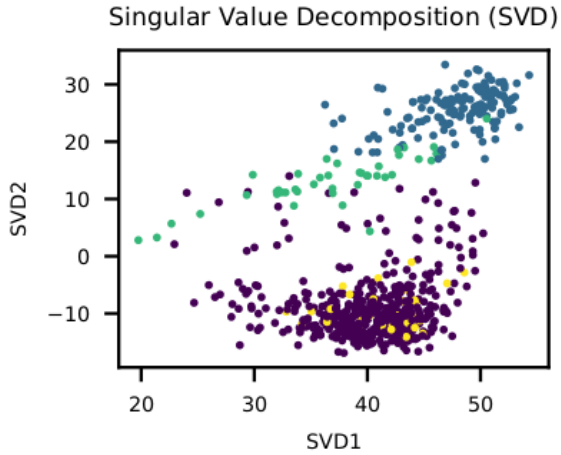


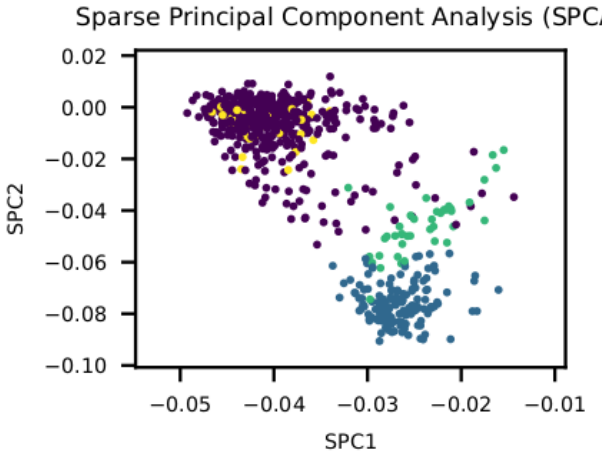


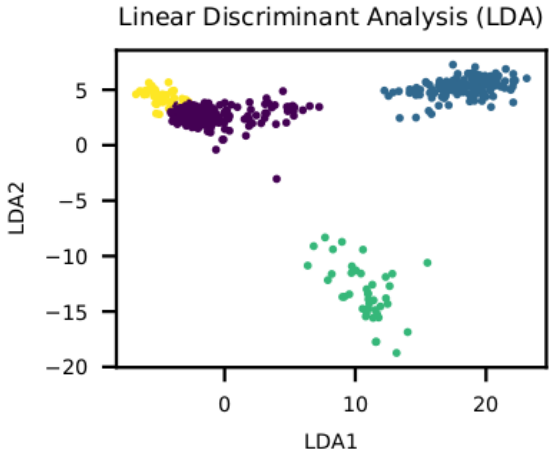




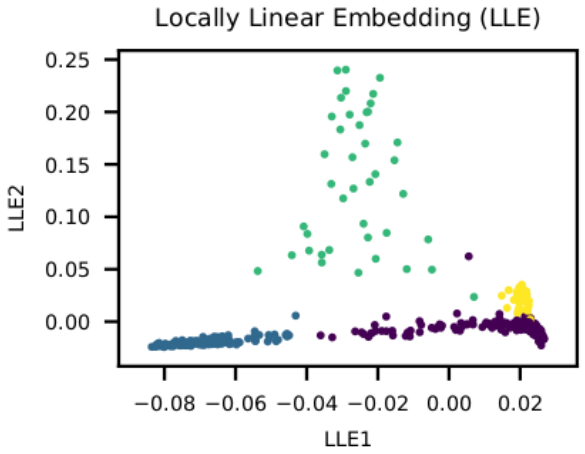


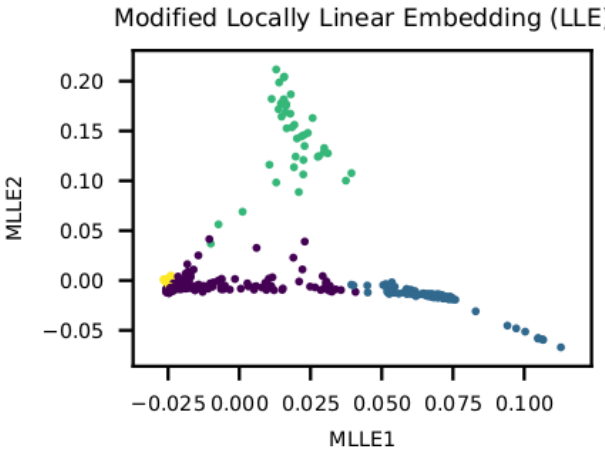


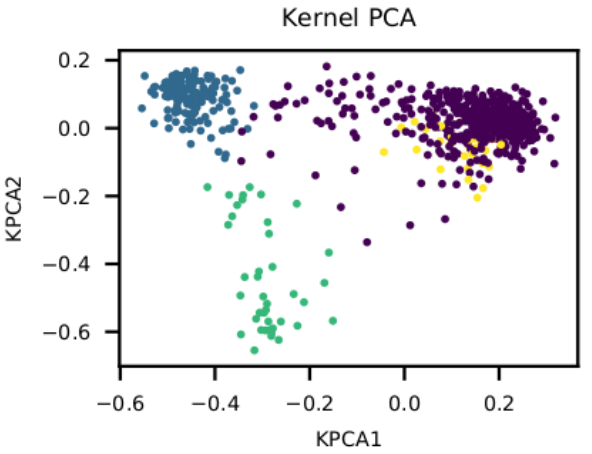


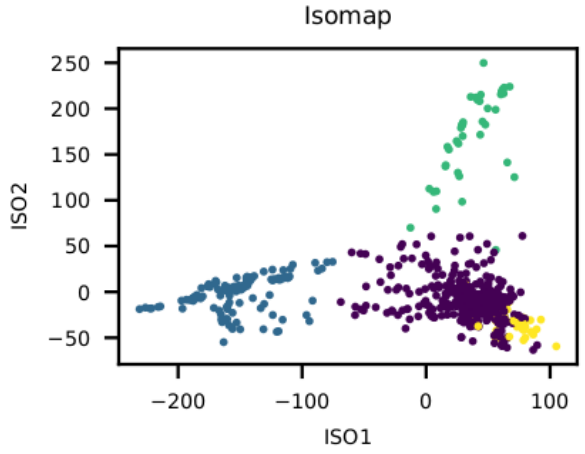


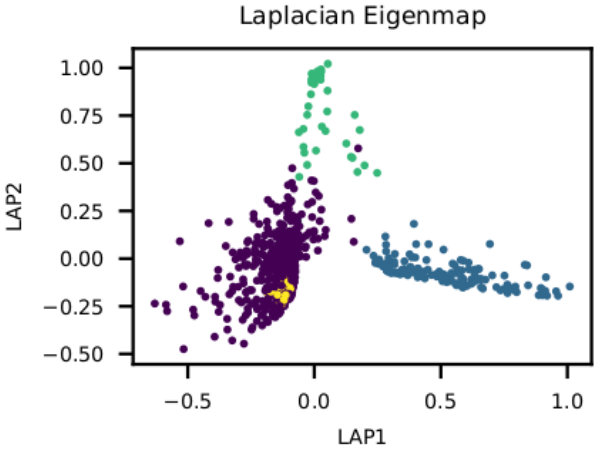
Non-Linear Dimension Reduction

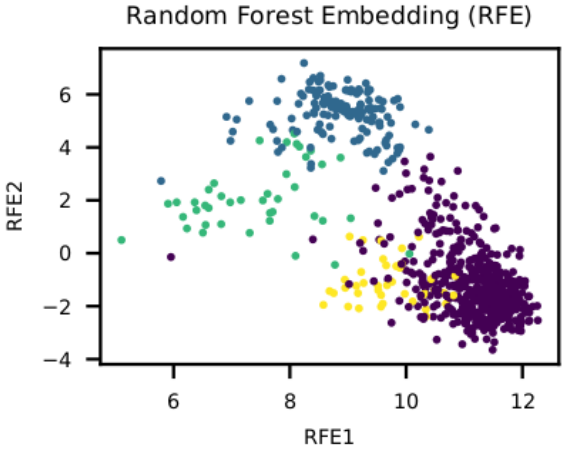


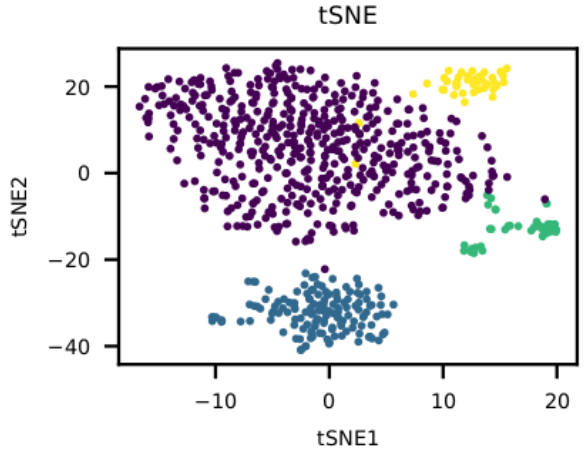


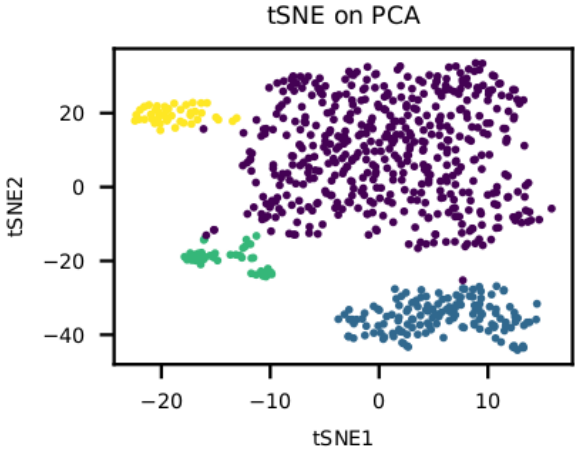


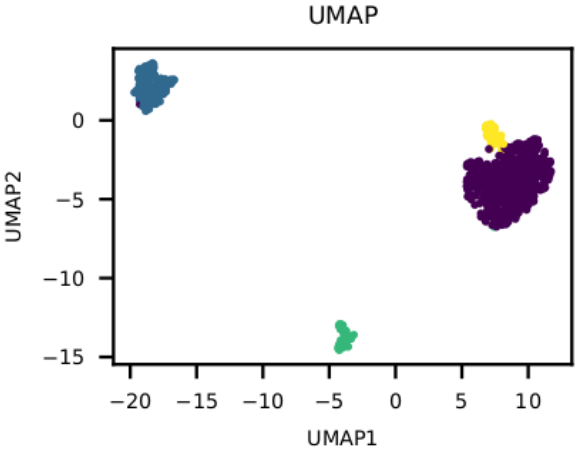






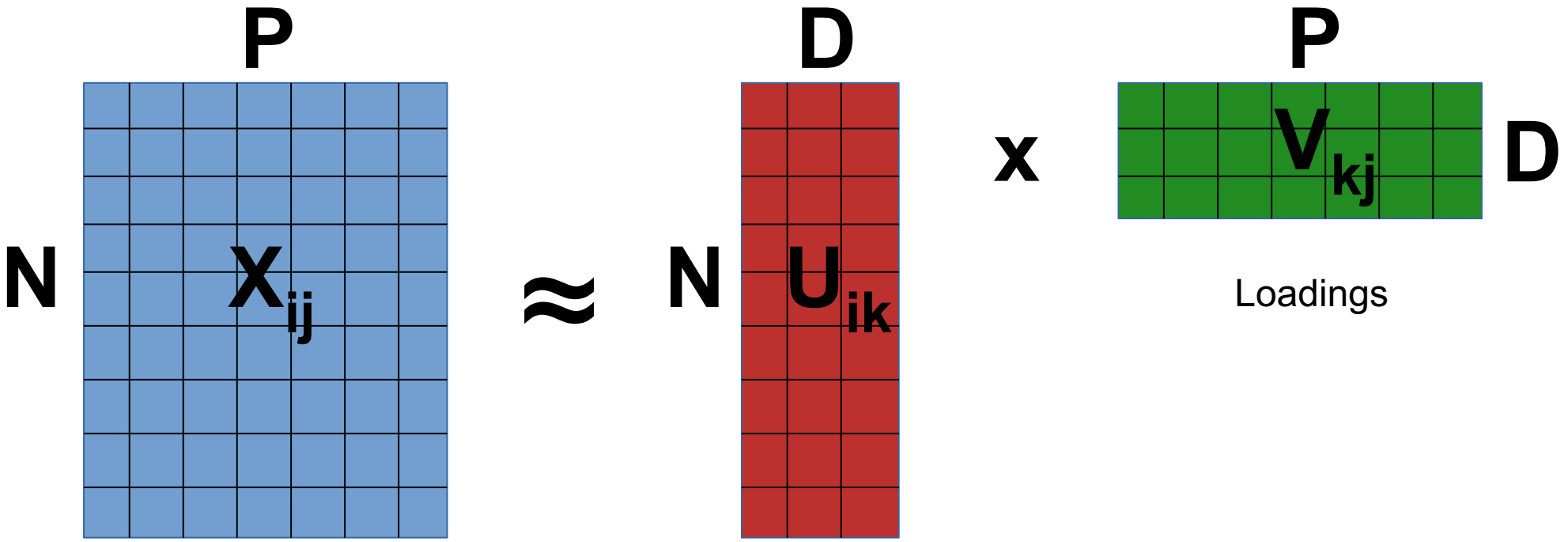






Matrix Factorization Dimension Reduction

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



Data

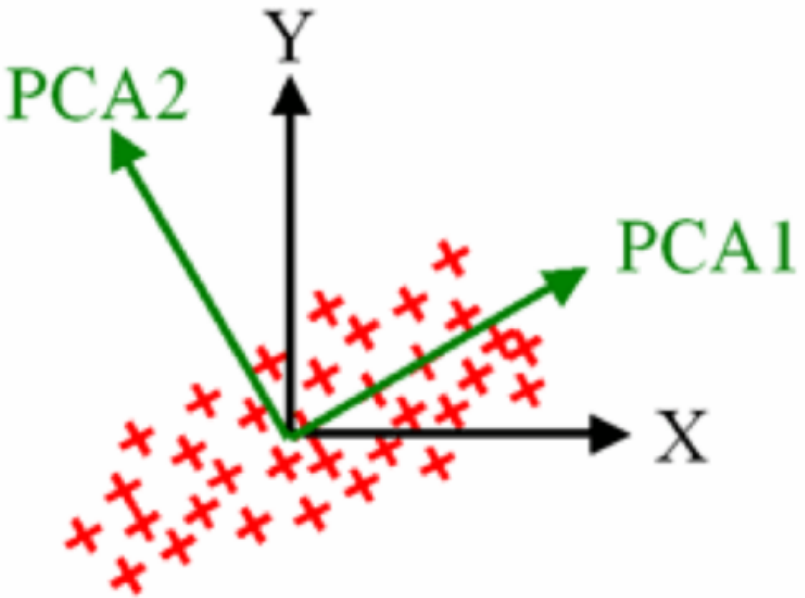
Low-dimensional data representation (embeddings)

Loadings

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

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 $\Longrightarrow < PC >= 0$

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

$$XX^T = A$$

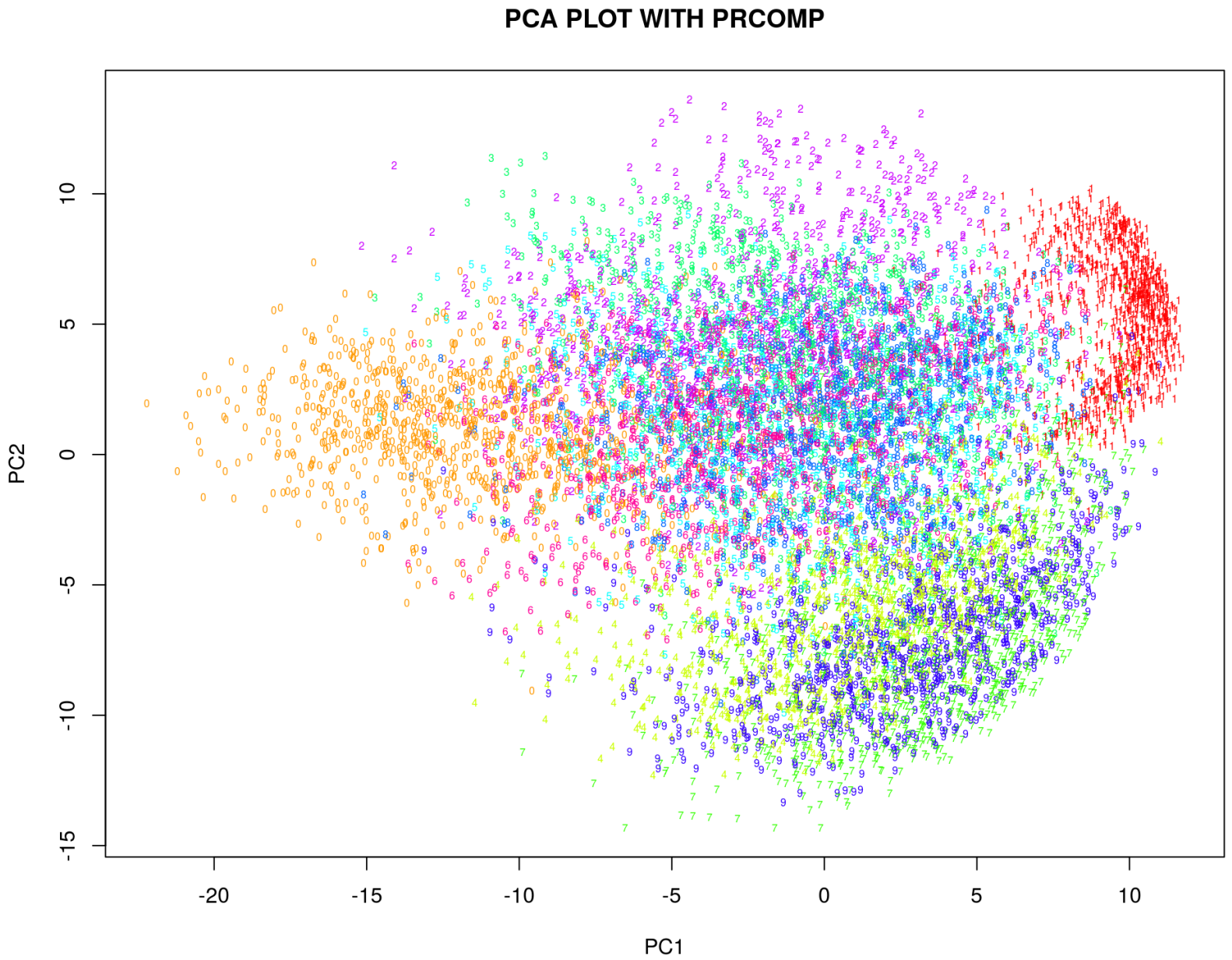
$$< PC^2 >= u^T A u$$

A is **variance-covariance** of X

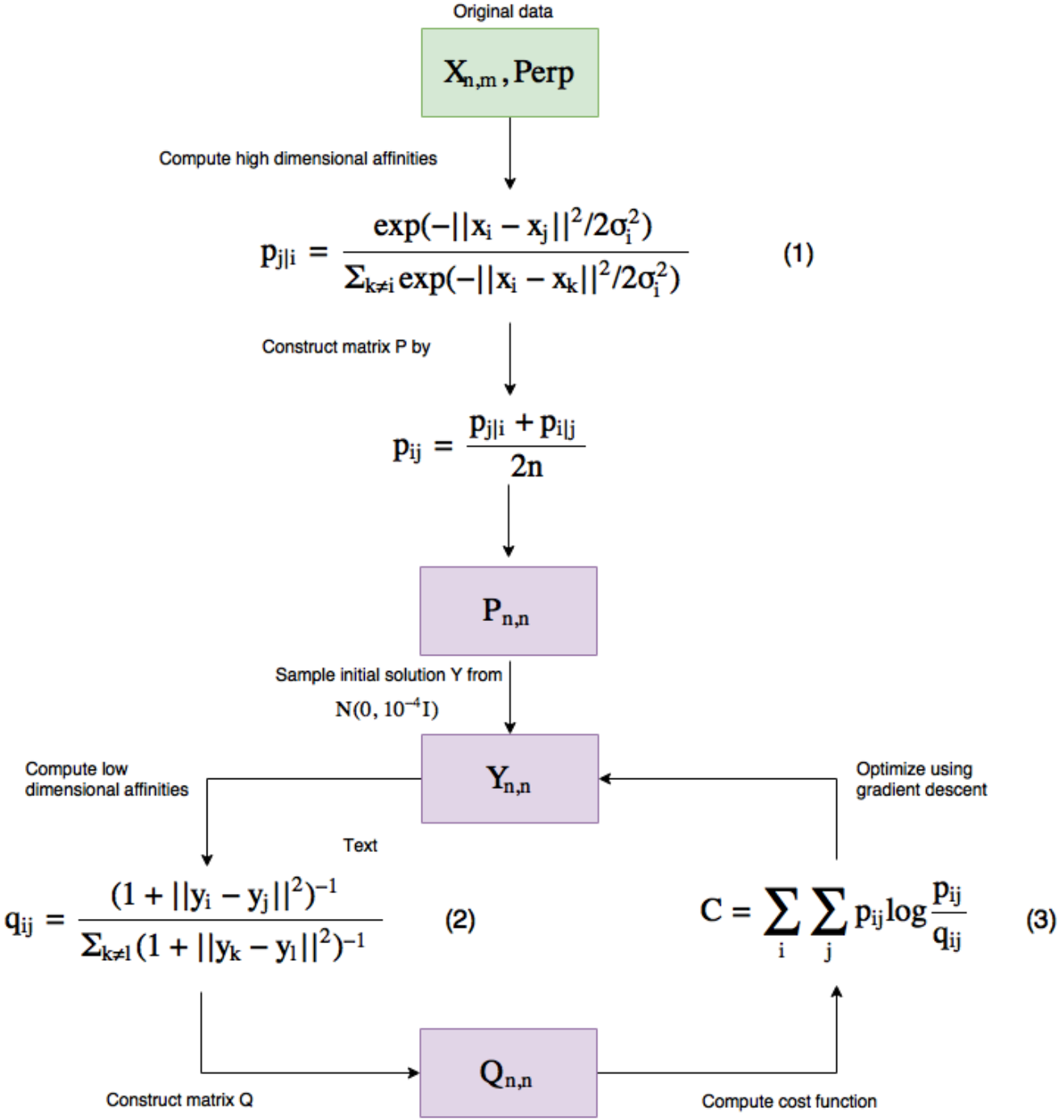
$$\max(u^T A u + \lambda(1 - u^T u)) = 0$$

$$Au = \lambda u$$

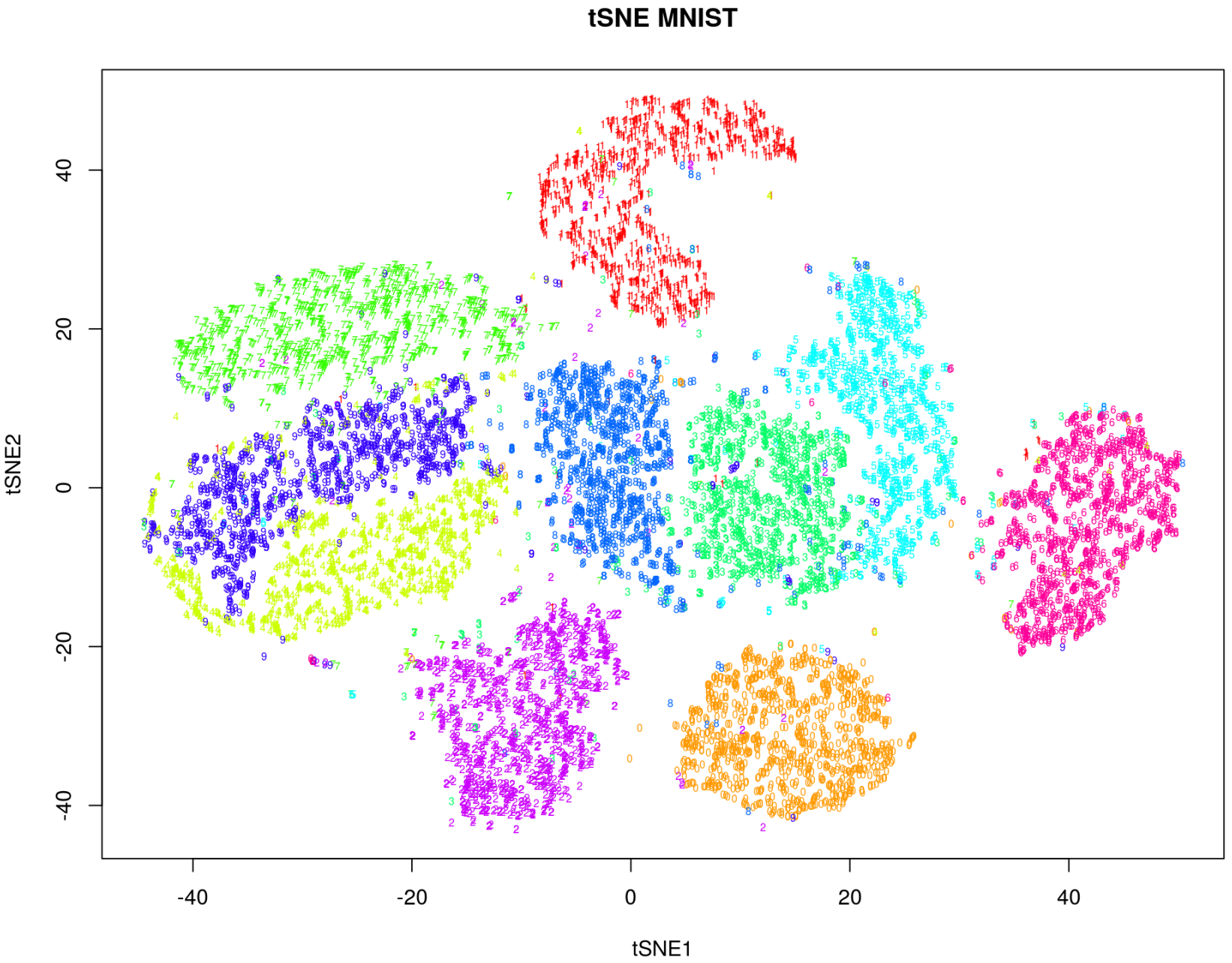
PCA on MNIST Data Set



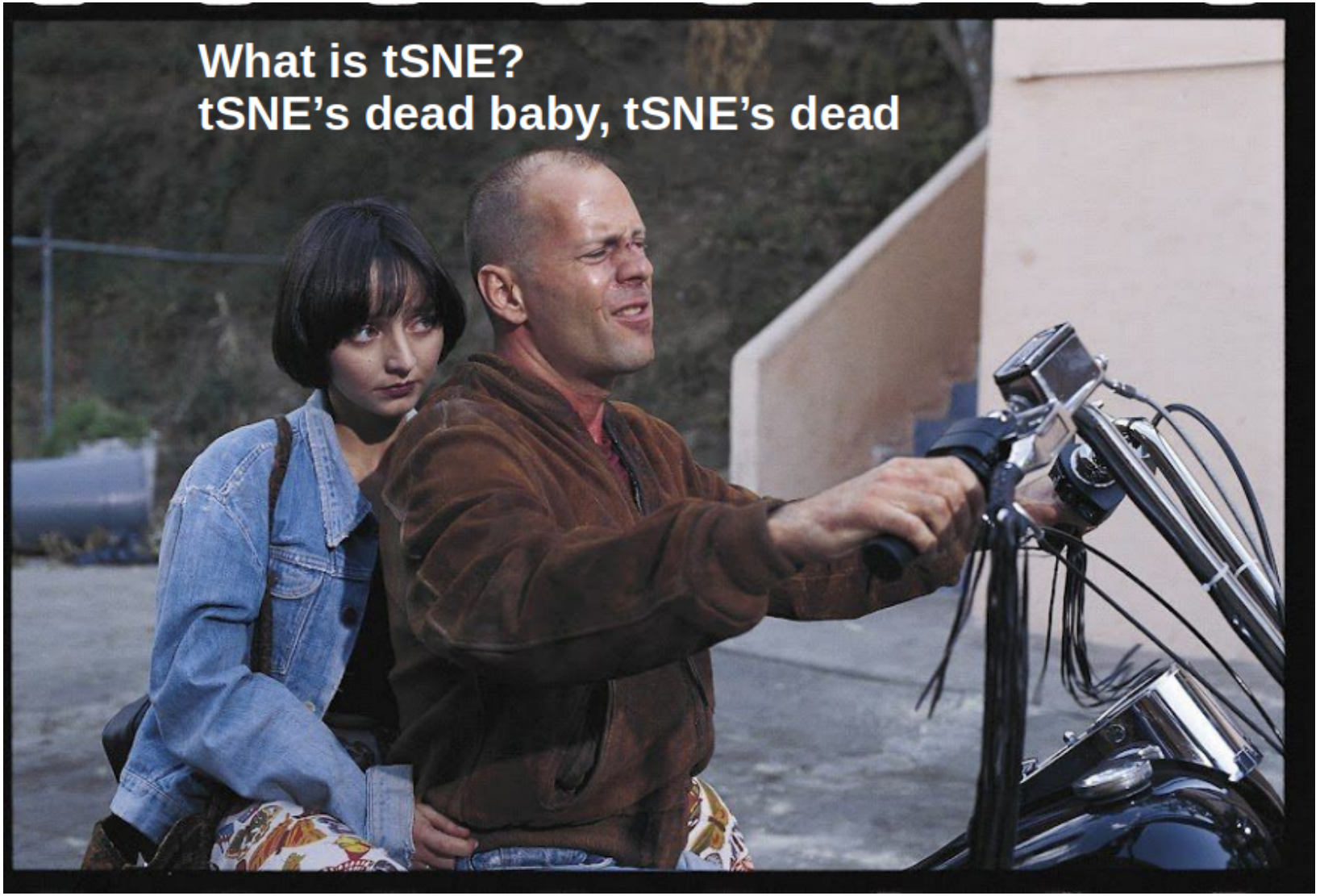
t-distributed Stochastic Neighbor Embedding (tSNE)



tSNE on MNIST Data Set



This is How you Might Feel if you are Doing scRNAseq Analysis



Is UMAP really superior over tSNE?

Why We are Uncomfortable Using tSNE

tSNE does not scale for large data sets?

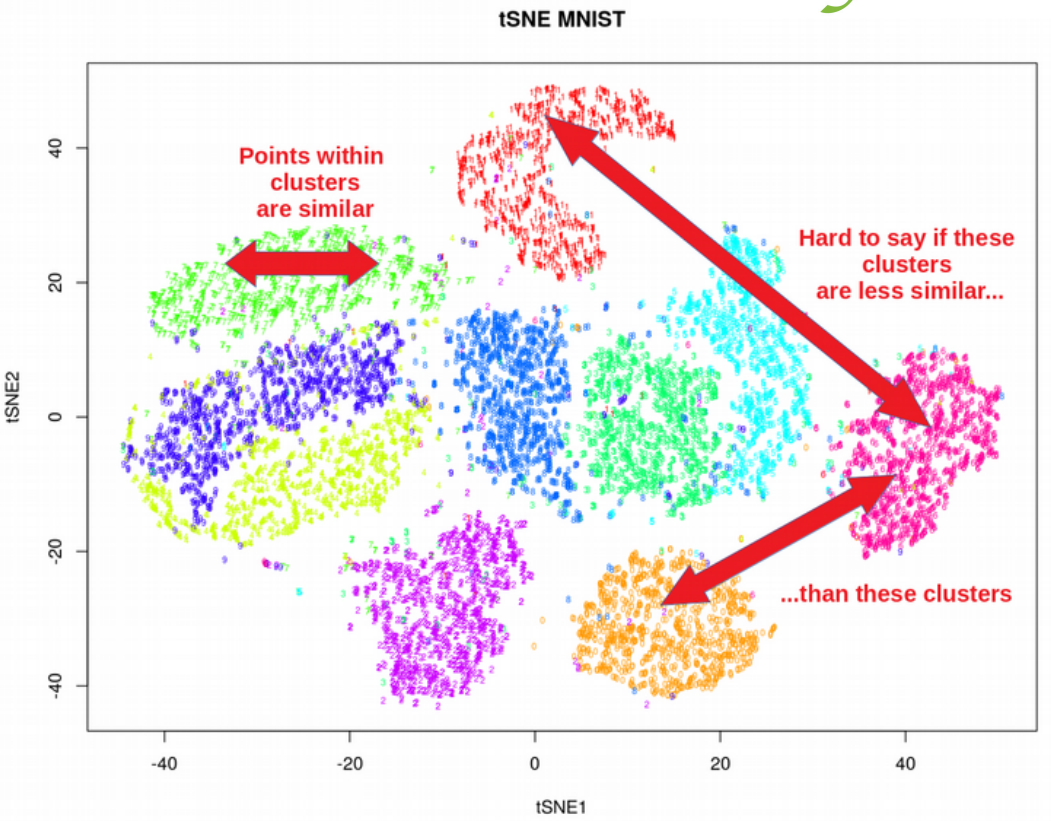
tSNE does not preserve global structure?

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)?



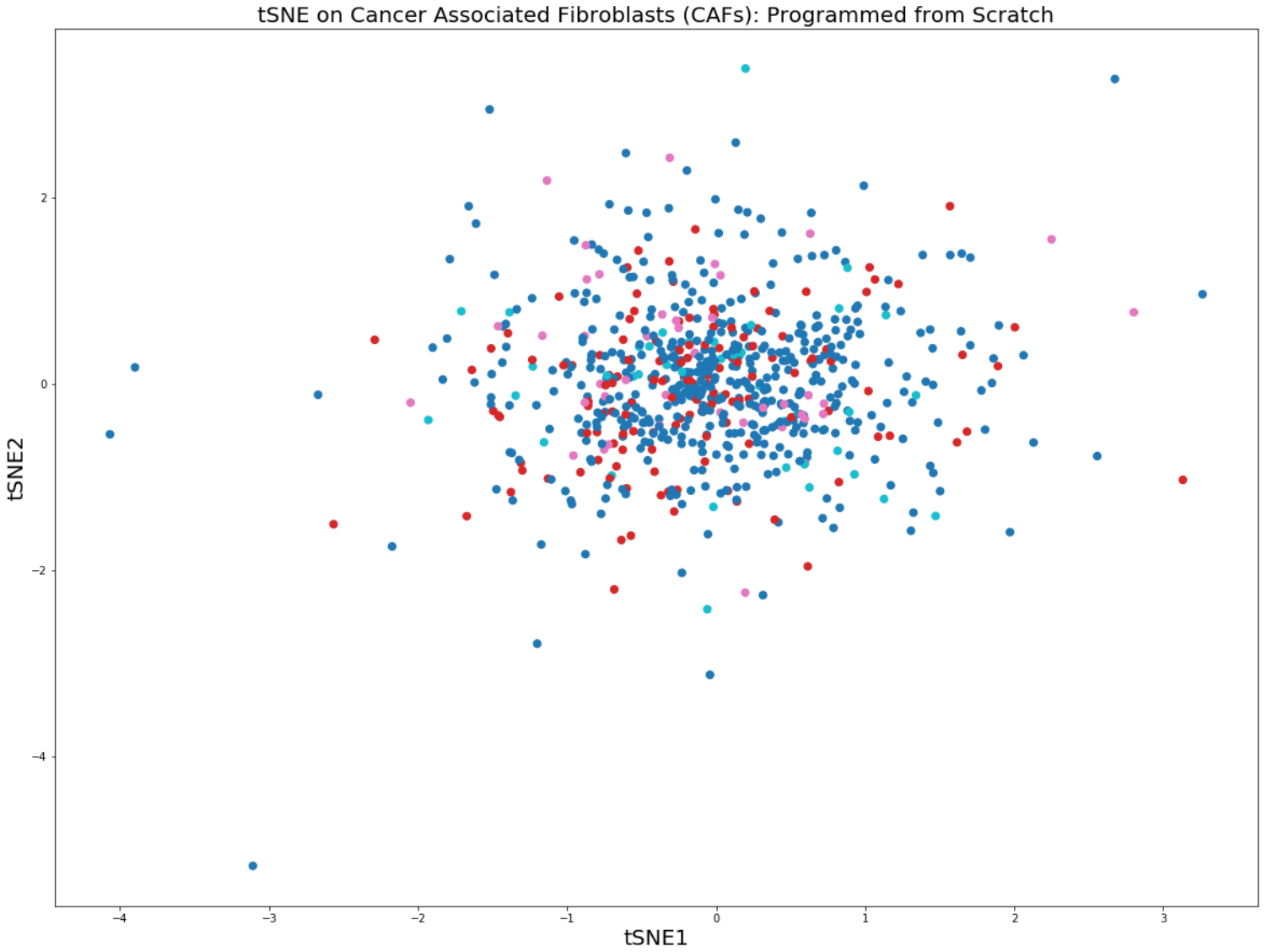
$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-||x_i - x_k||^2 / 2\sigma_i^2\right)}, \qquad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \qquad (1)$$

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

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

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

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

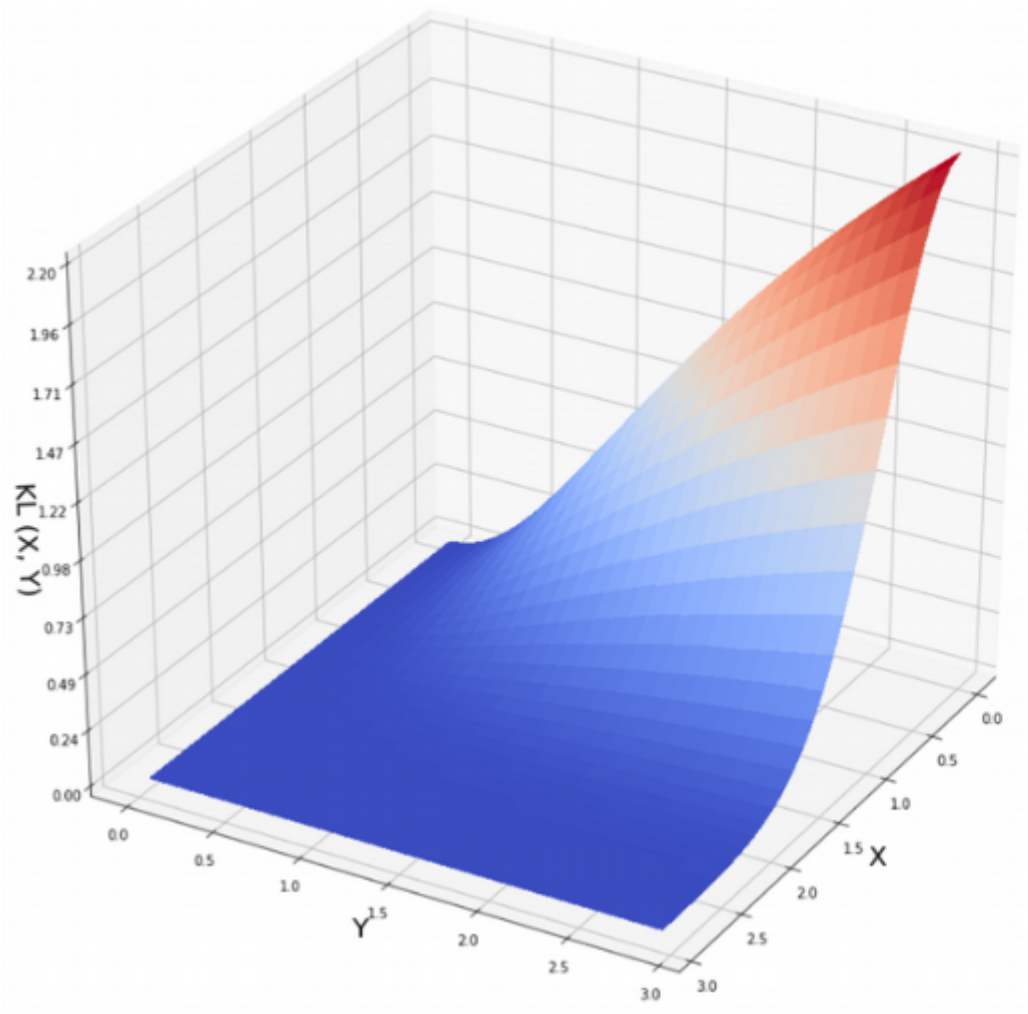
UMAP uses Cross-Entropy (not KL) as cost function

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

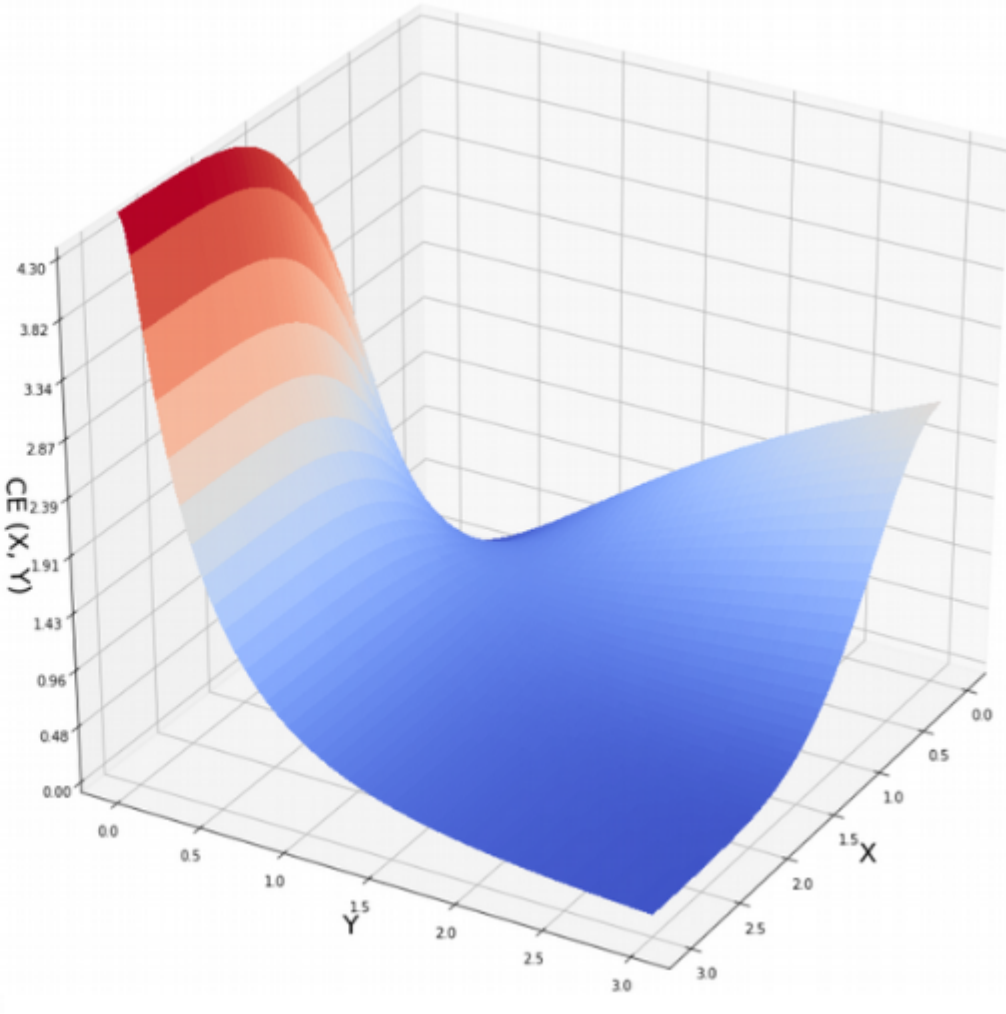
This is similar to tSNE cost function

This term is UMAP specific

Cost Function Makes UMAP Preserve More Global Structure than tSNE

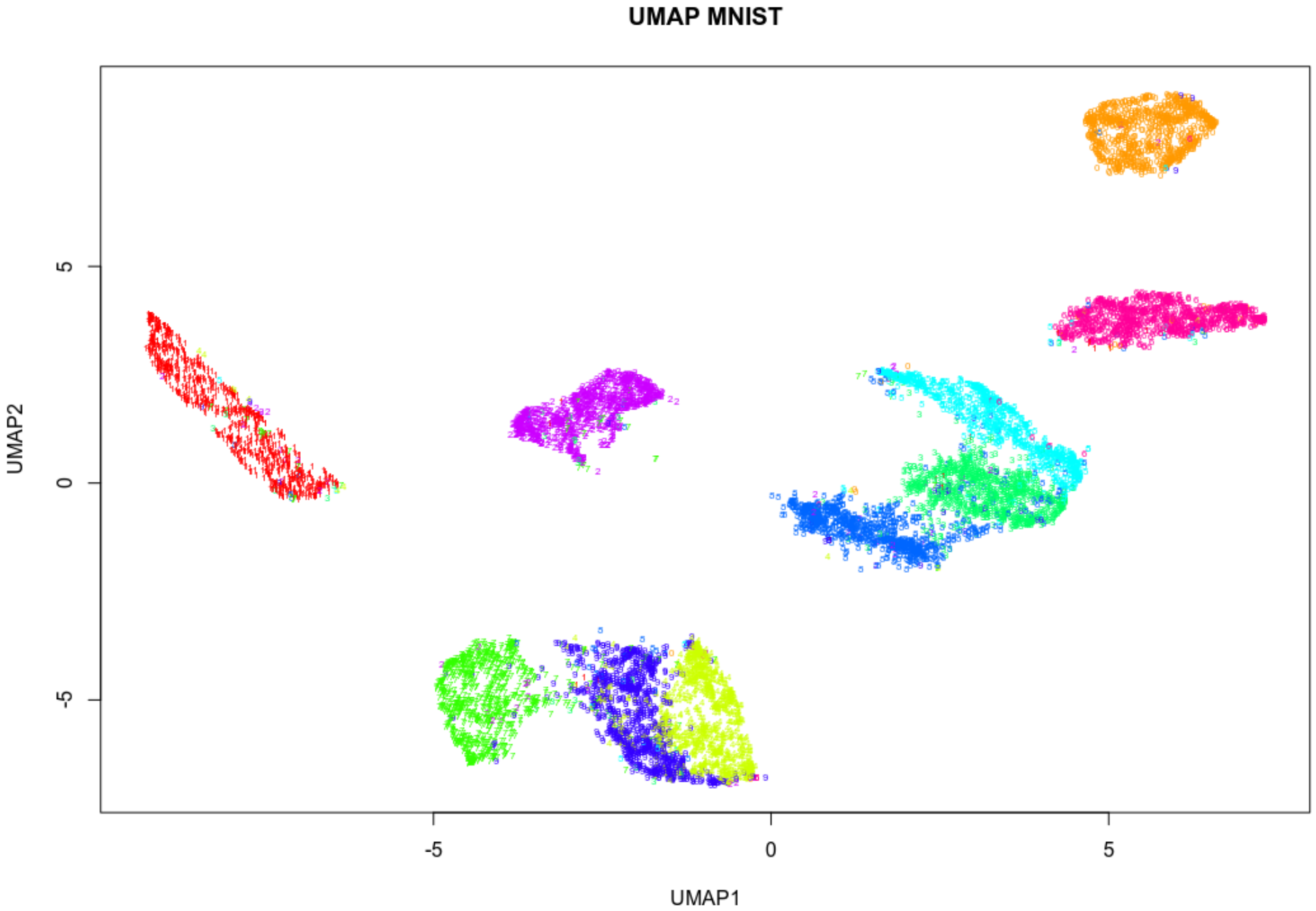


X → infinity, Y can be any

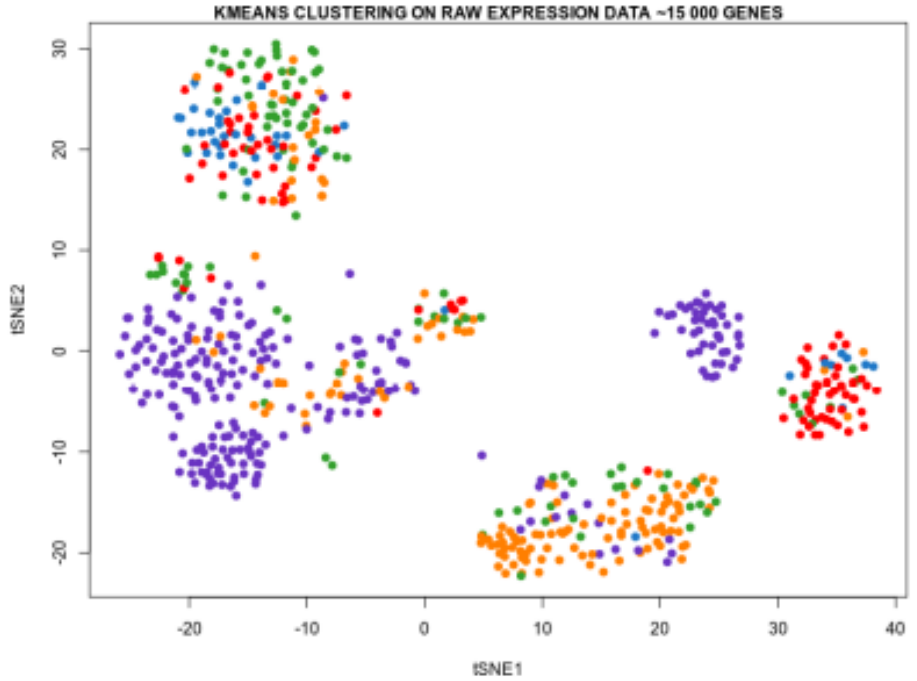


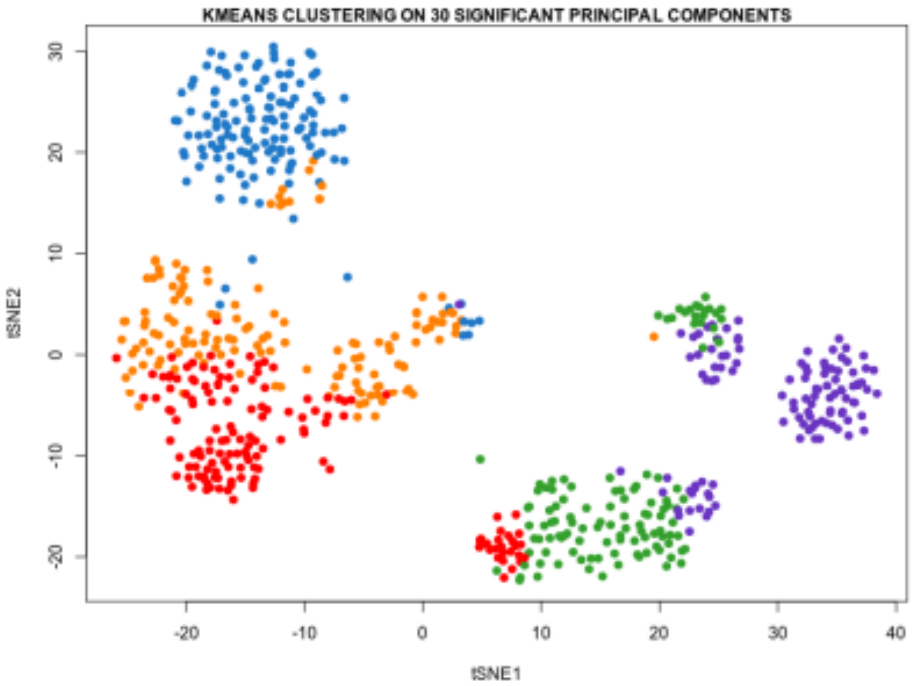
X → infinity, Y → infinity

UMAP on MNIST Data Set

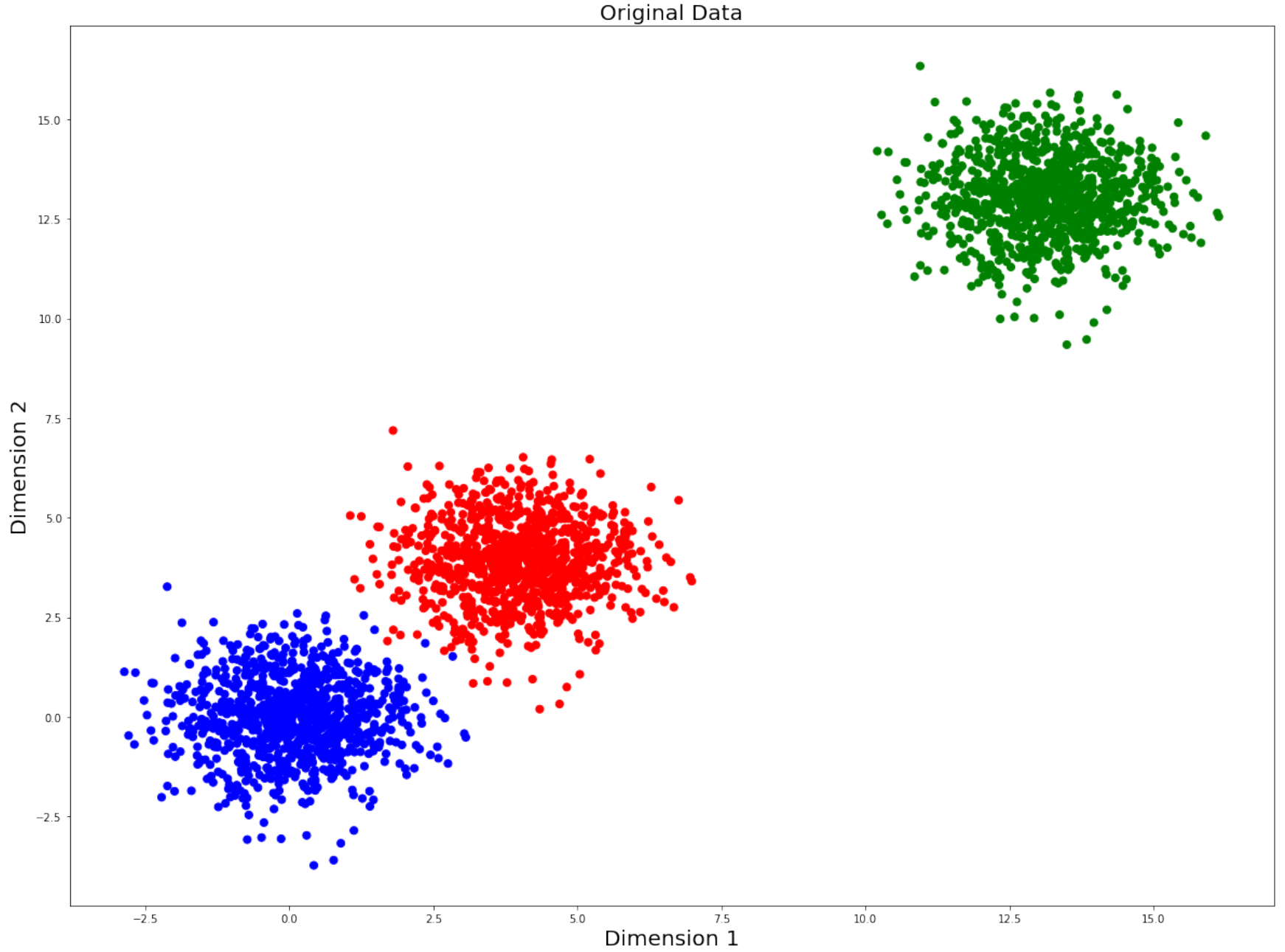


Why Global Structure Preservation is Important

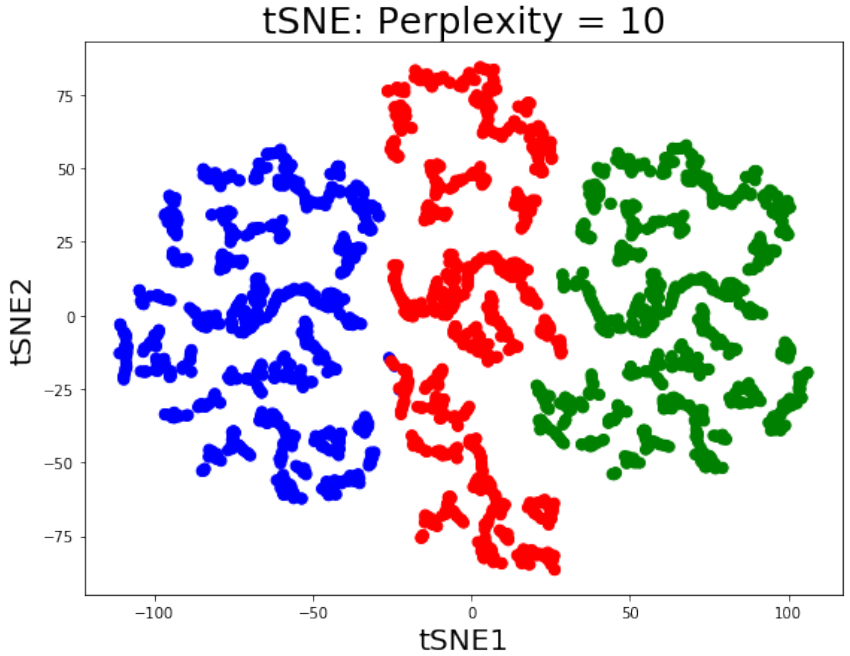


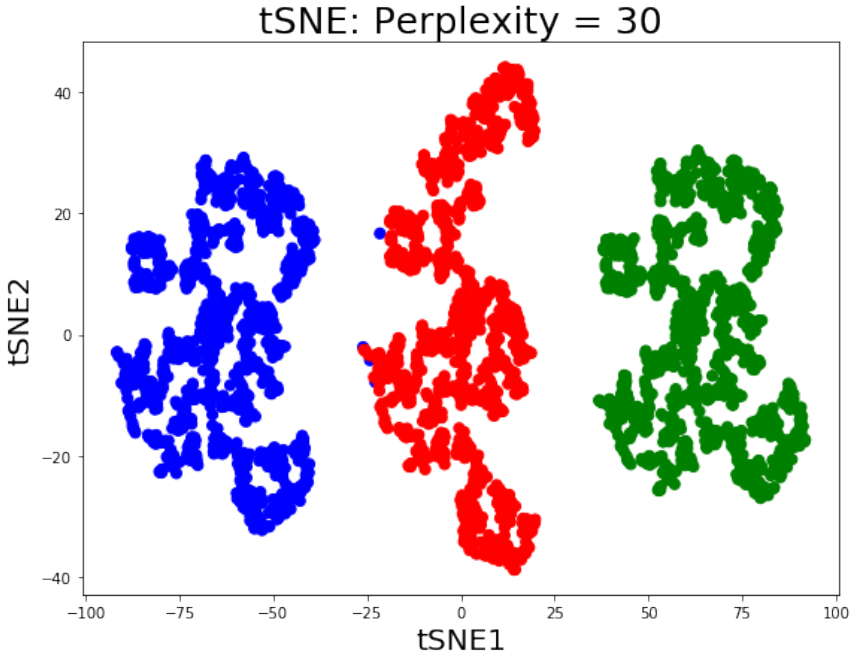


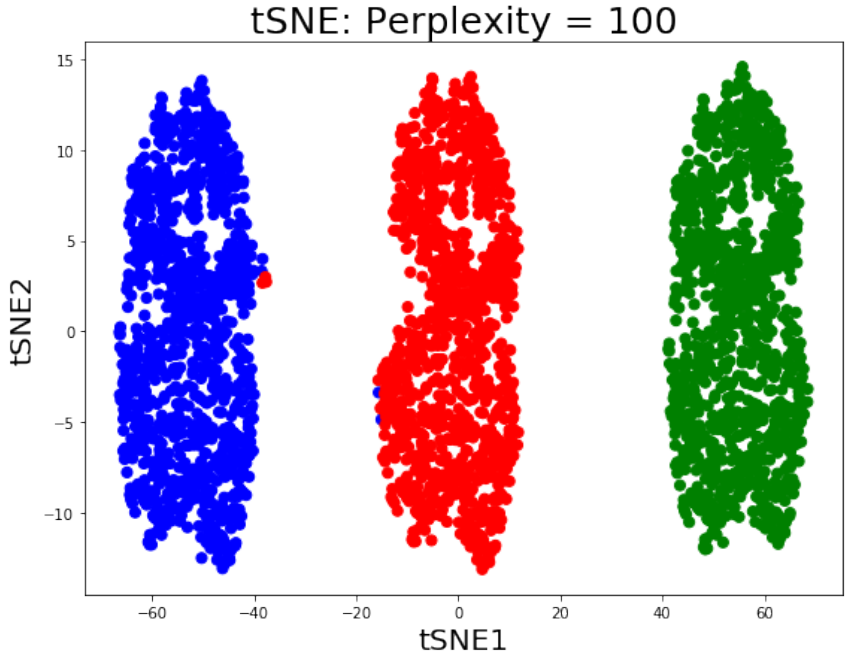
Why Global Structure Preservation is Important

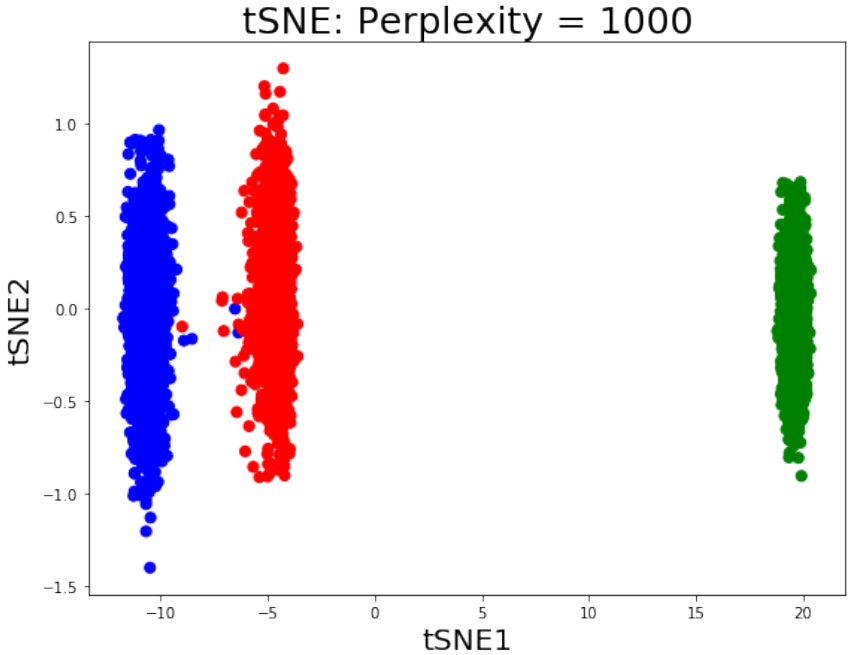


Why Global Structure Preservation is Important

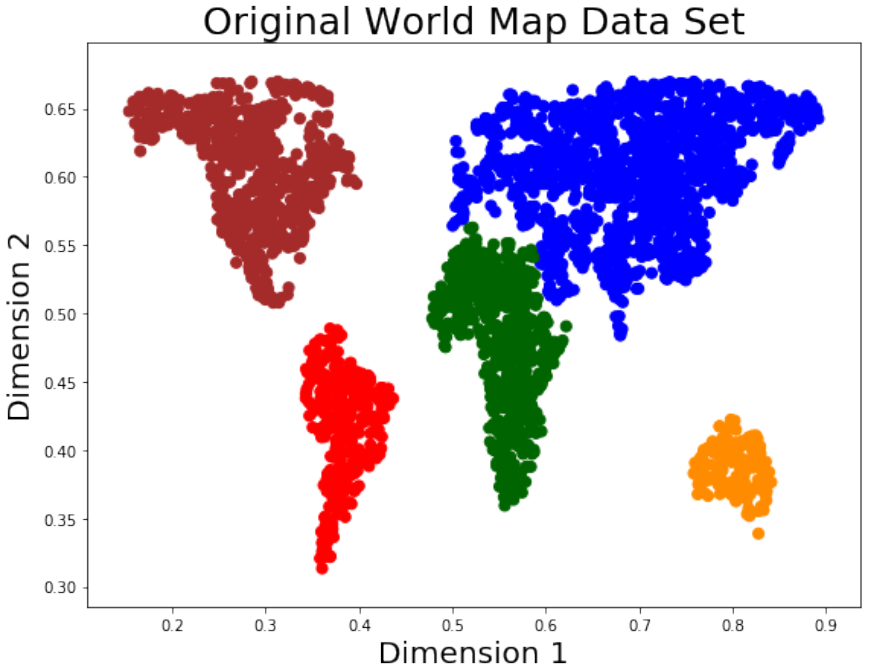


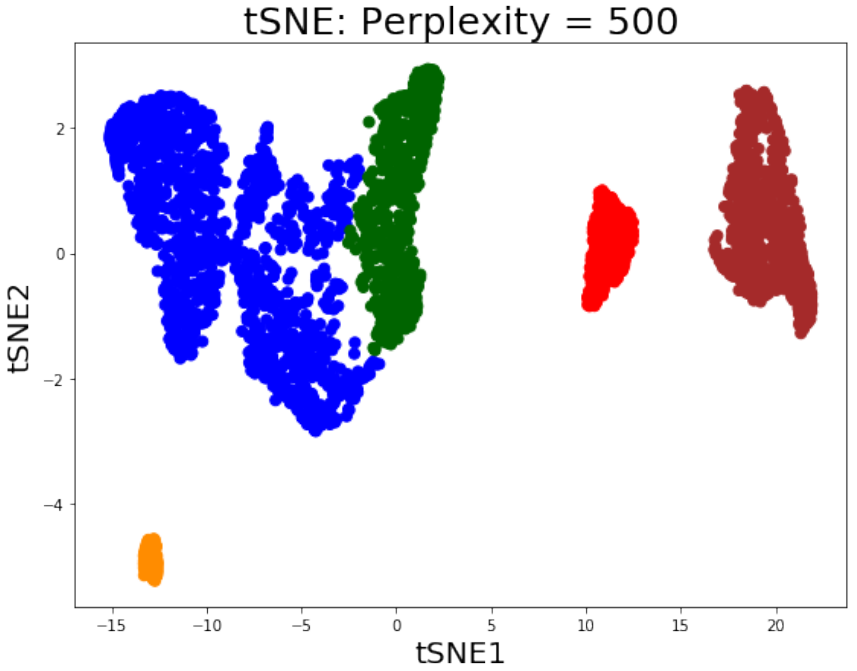


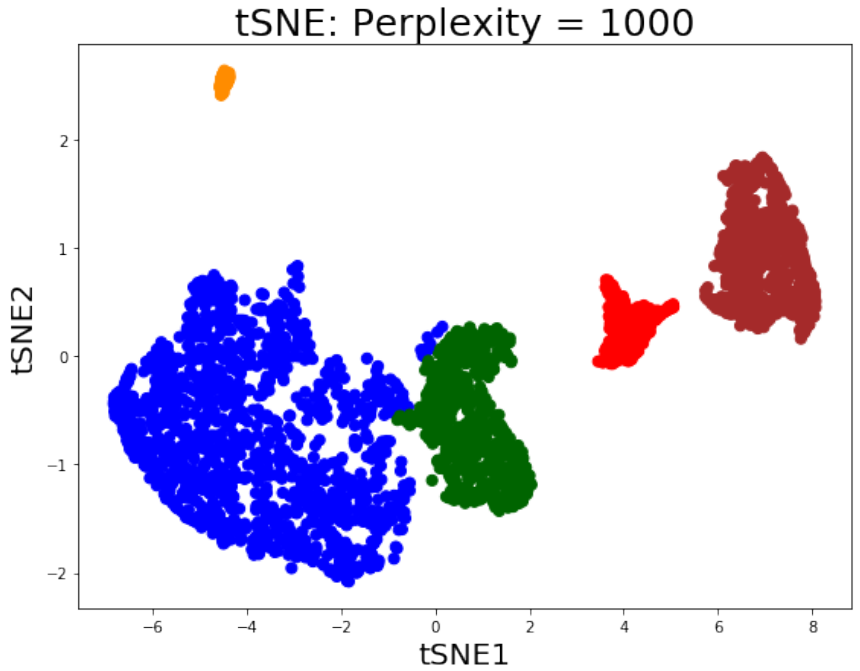


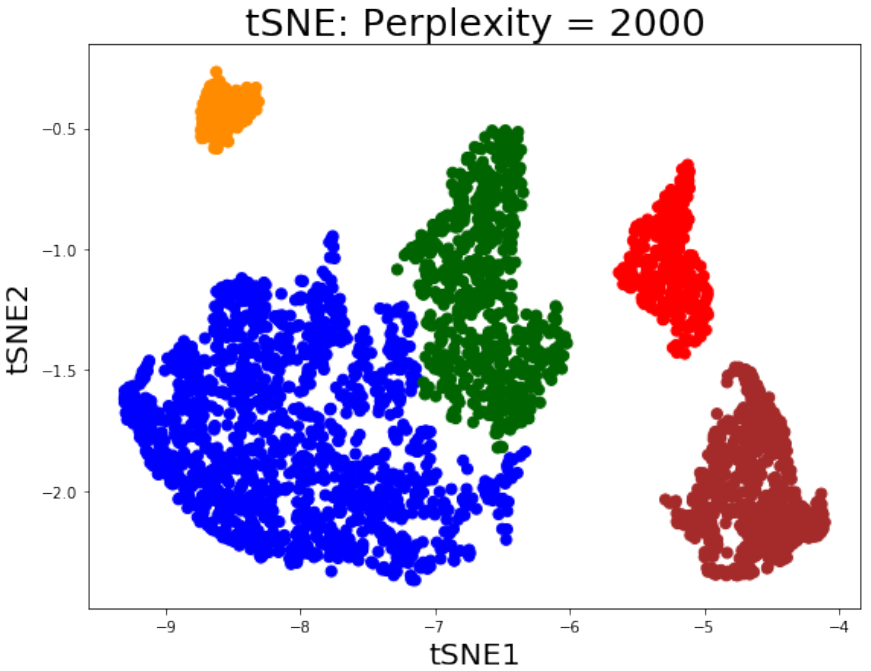


Why Global Structure Preservation is Important

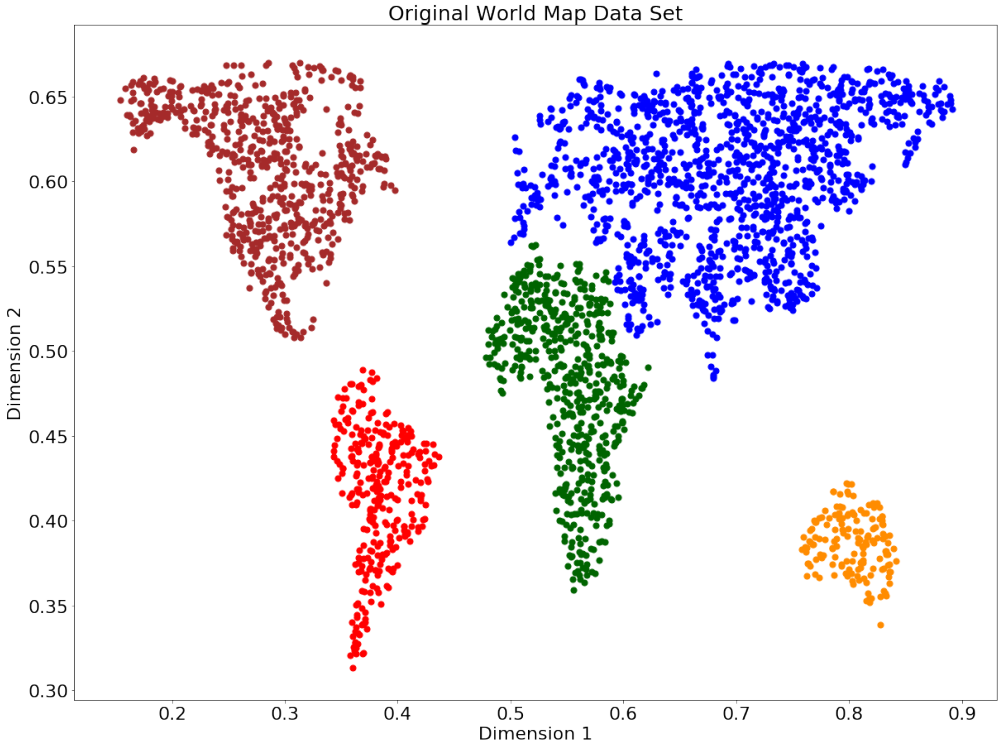


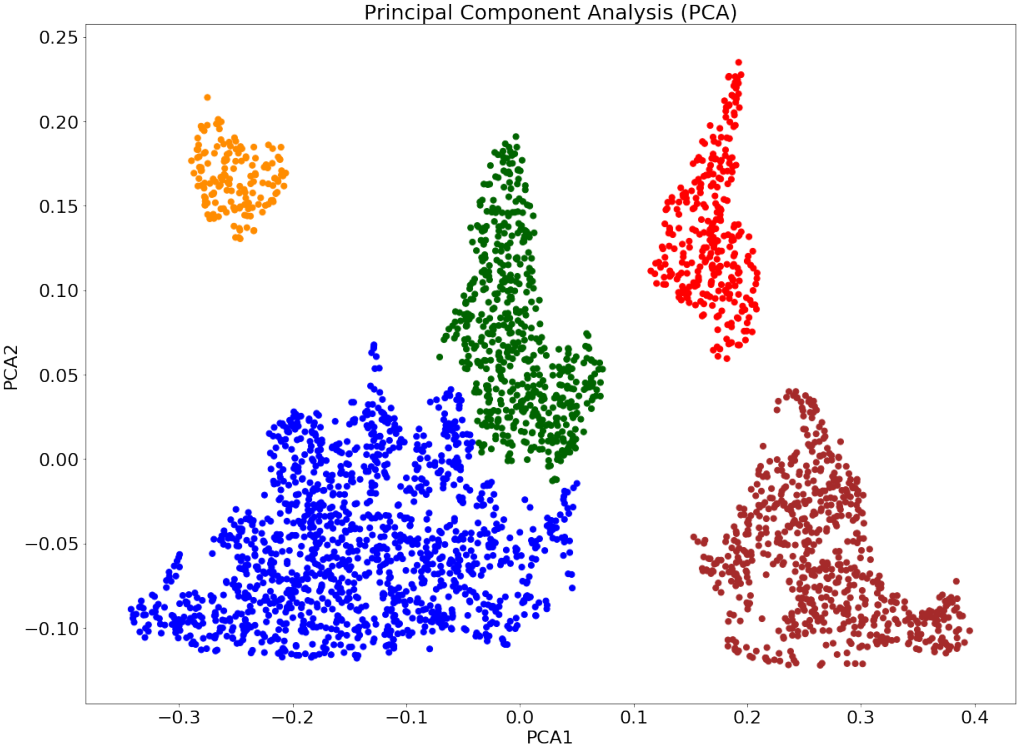


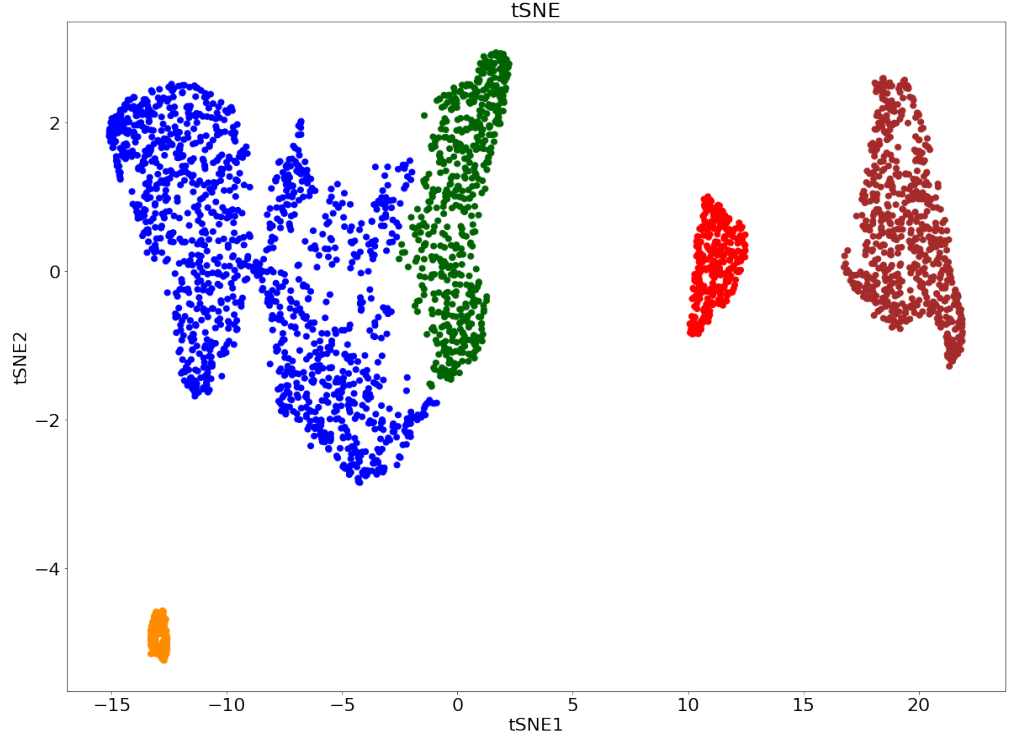


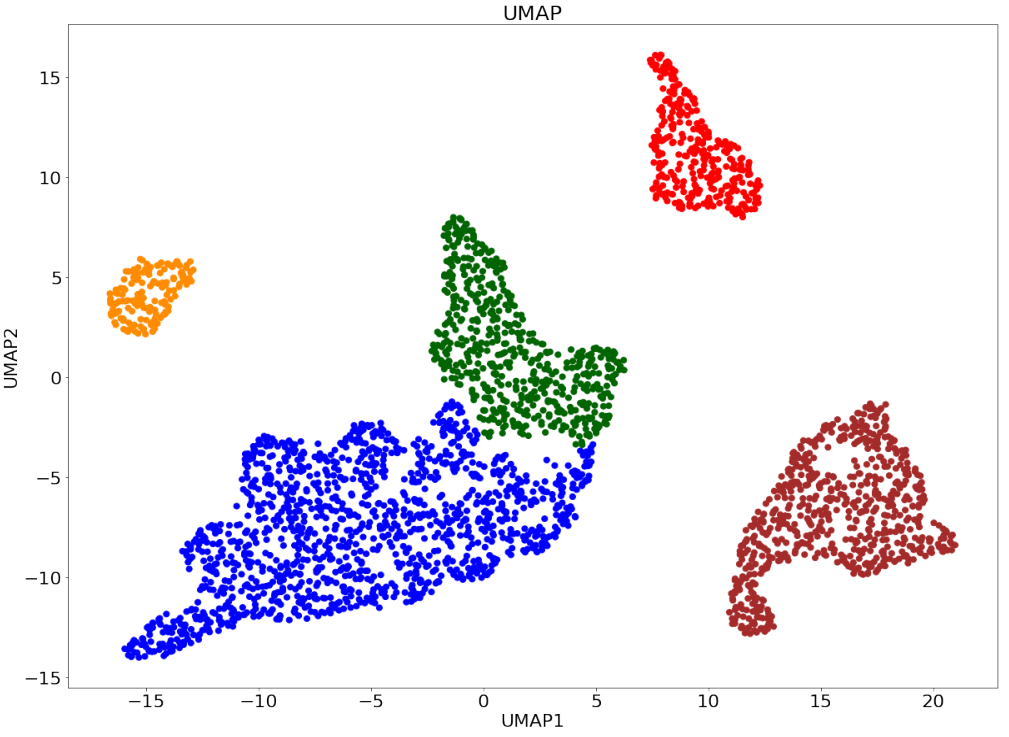


Global Structure Preservation for Linear Manifold

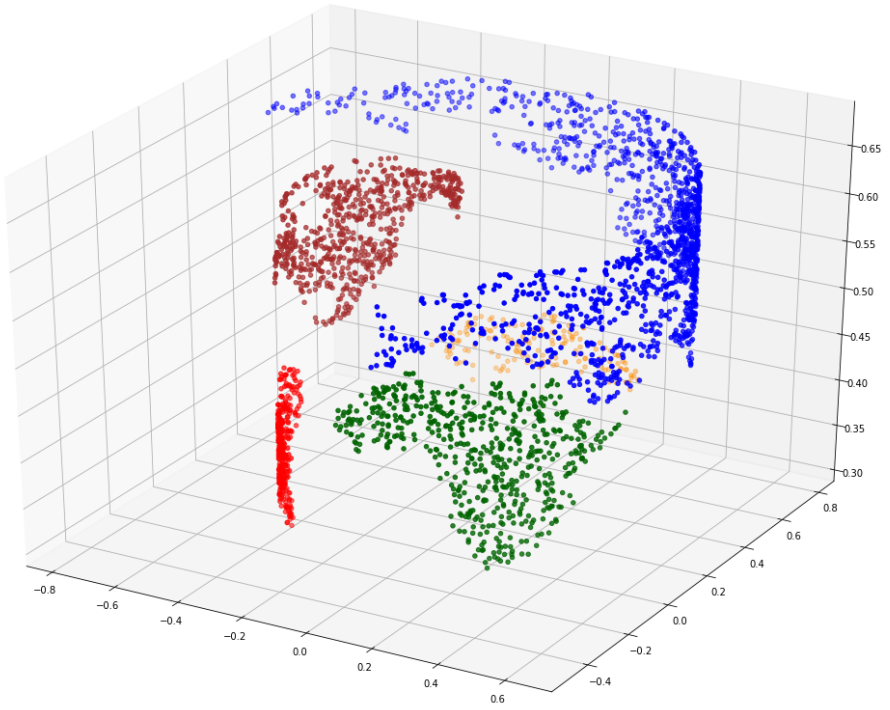


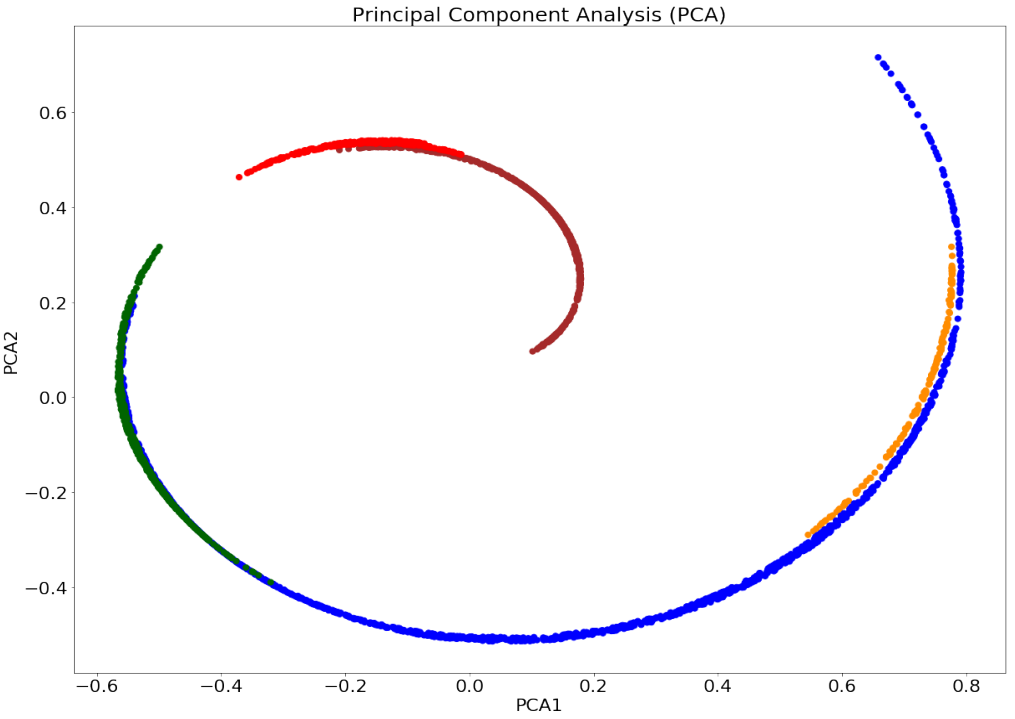


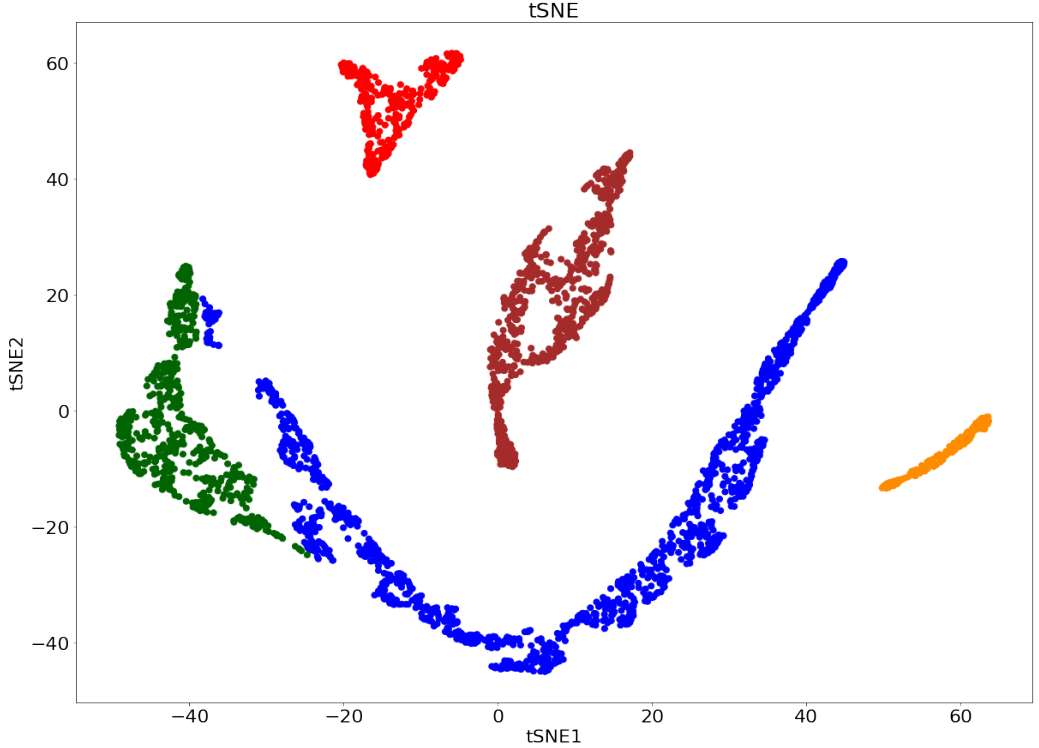


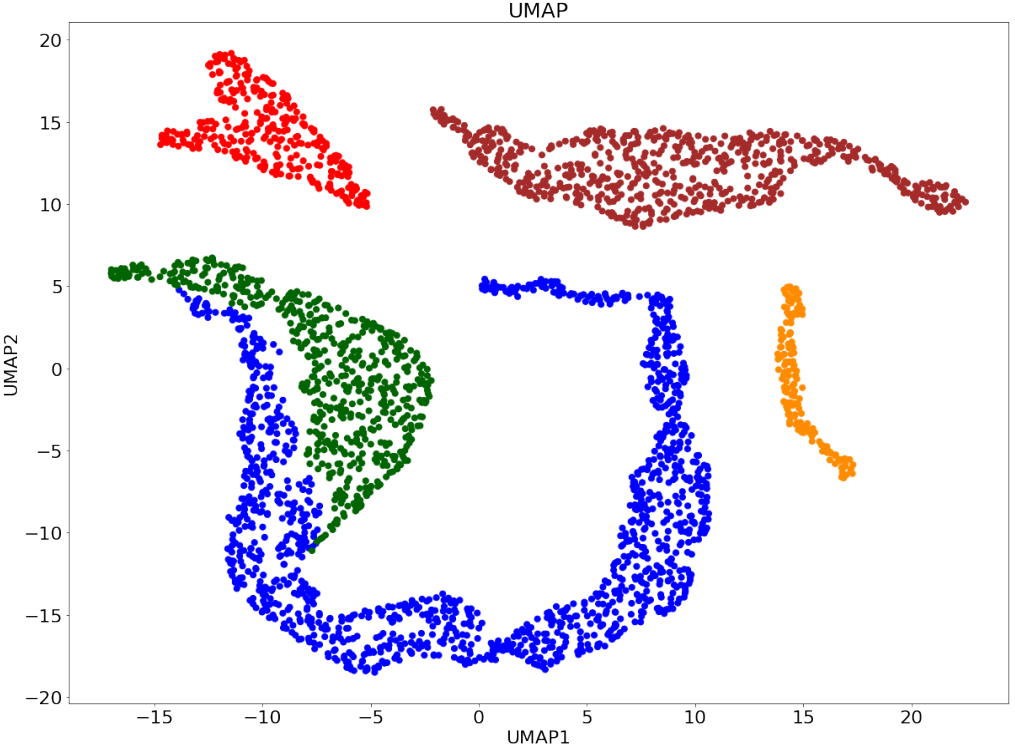


Global Structure Preservation for Non-Linear Manifold

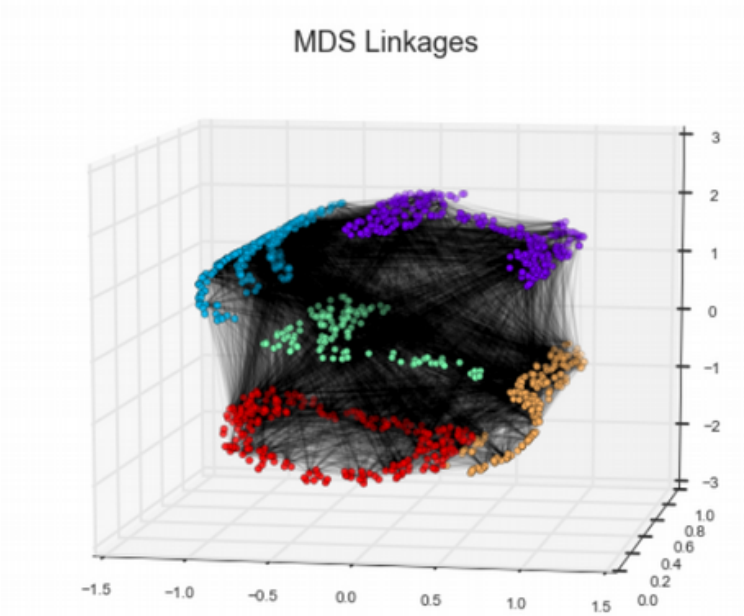


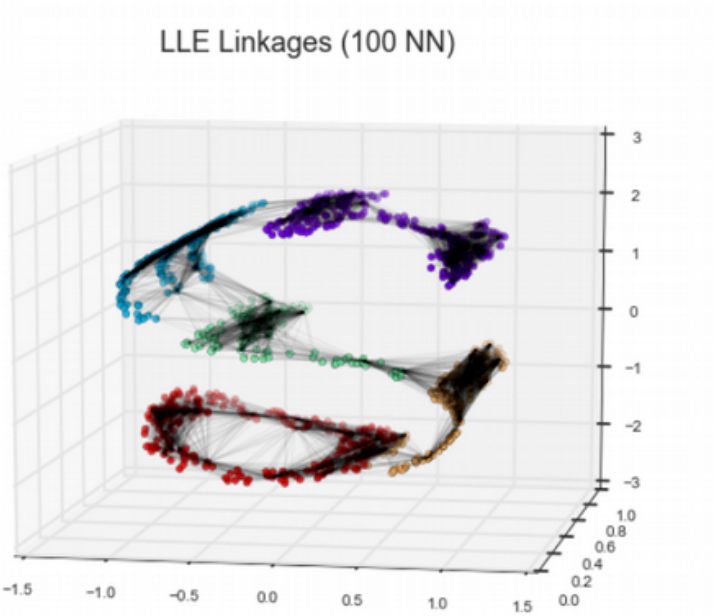


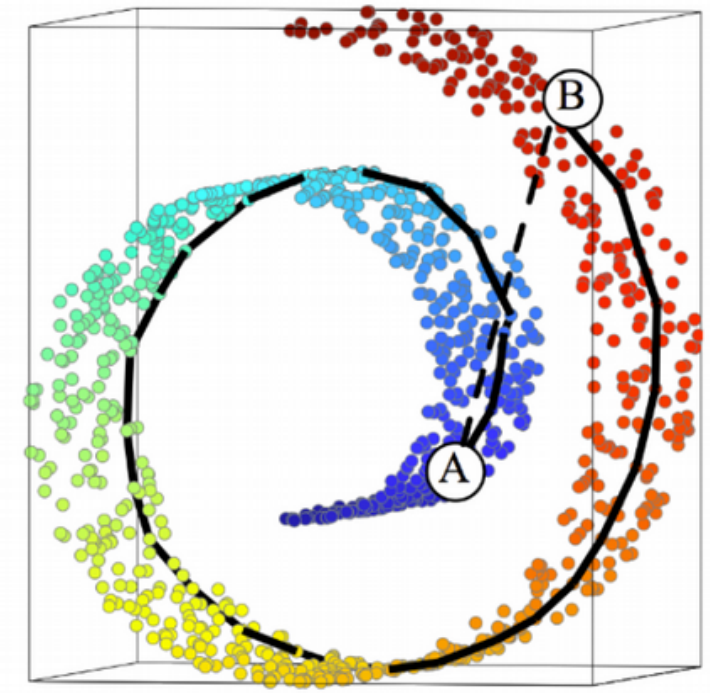


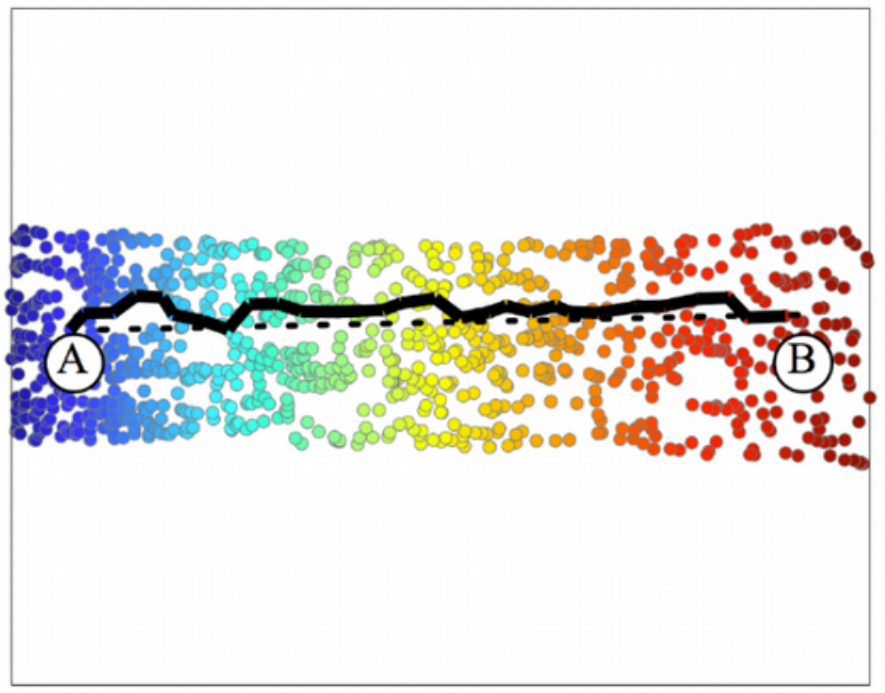


Why PCA Can't Unwrap Swiss Roll

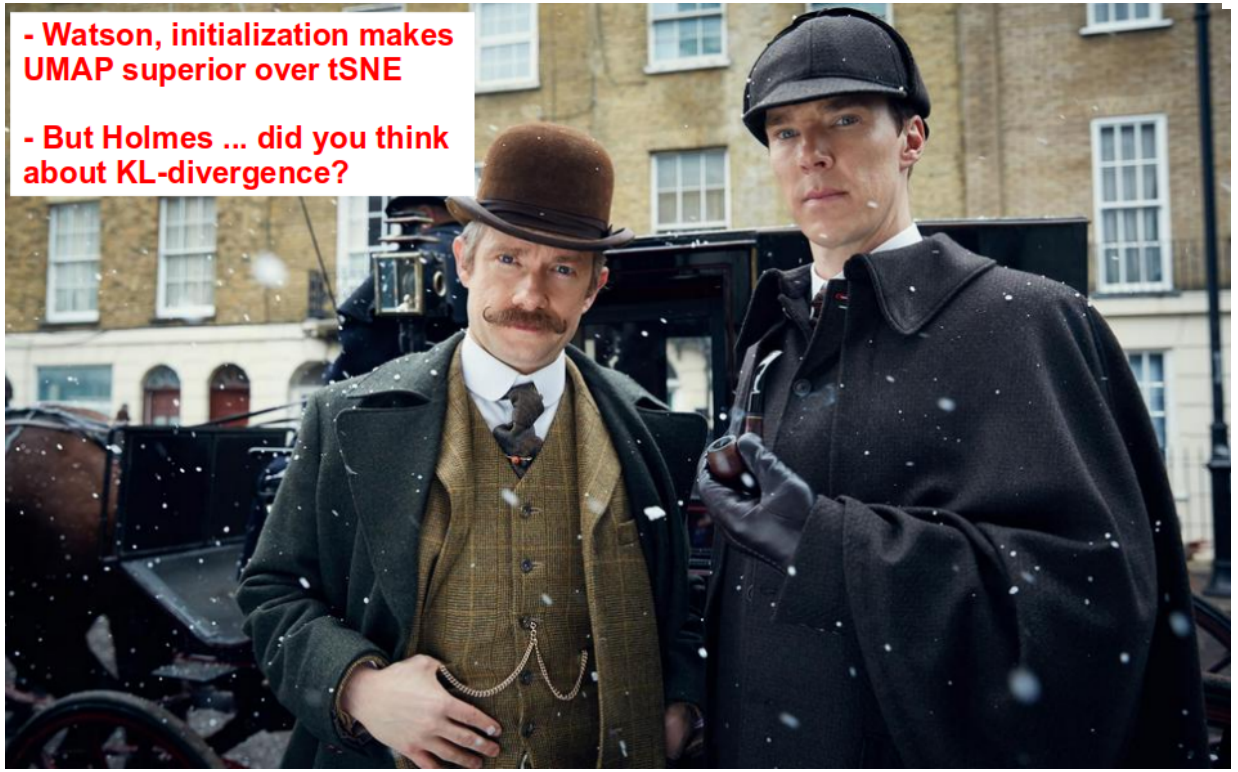








Initialization vs Cost Function



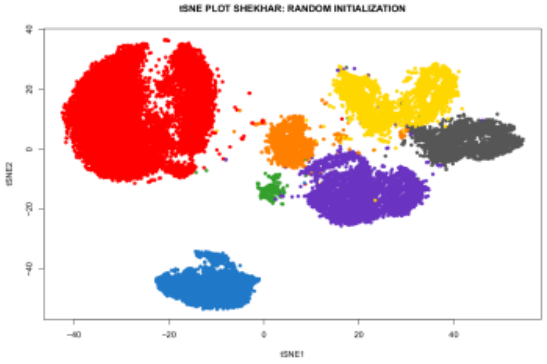
Initialization: PCA

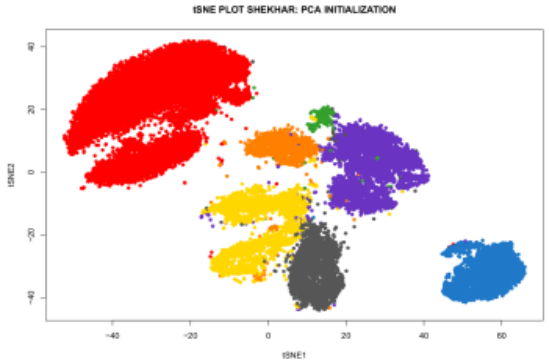
Cost function: KL / CE

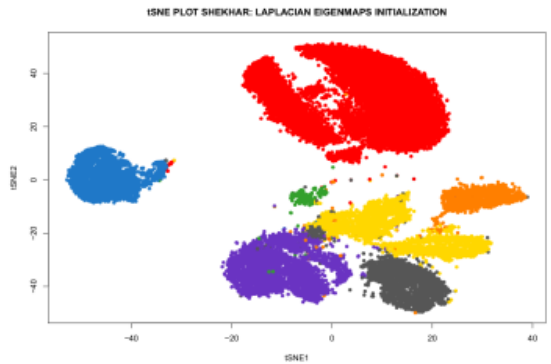
$$y_i = y_i - \mu \frac{\partial \text{Cost}}{\partial y_i}$$

Final embedding

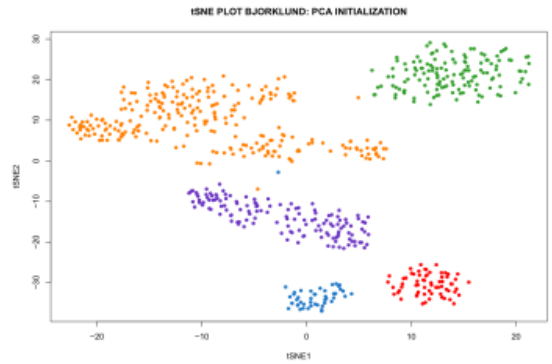
Learning rate

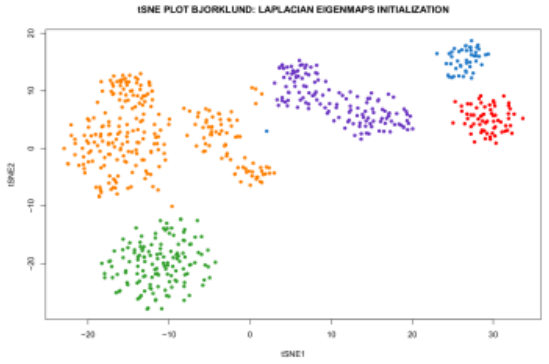

tSNE PLOT SHEKHAR: RANDOM INITIALIZATION
tSNE1
tSNE2

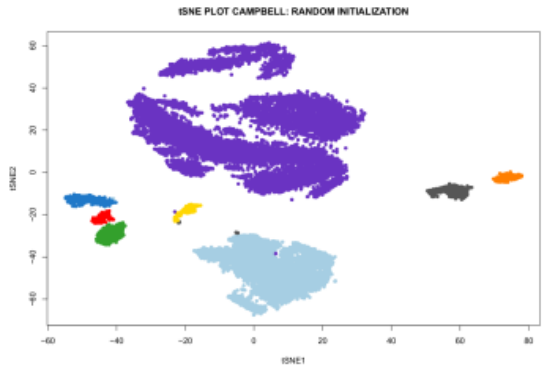

tSNE PLOT SHEKHAR: PCA INITIALIZATION
tSNE1
tSNE2

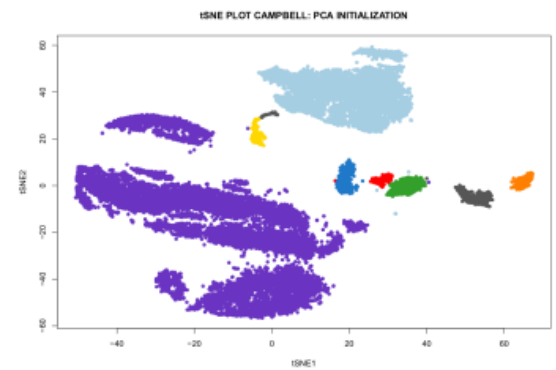

tSNE PLOT SHEKHAR: LAPLACIAN EIGENMAPS INITIALIZATION
tSNE1
tSNE2

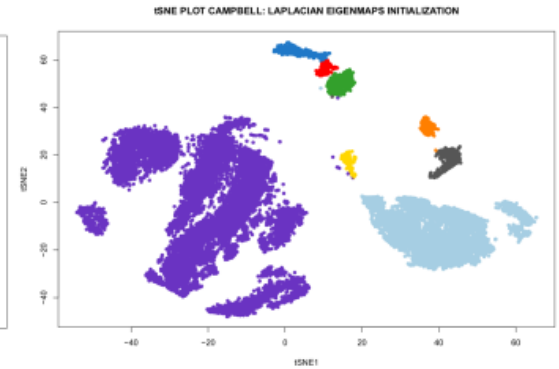

tSNE PLOT BJORKLUND: RANDOM INITIALIZATION
tSNE1
tSNE2

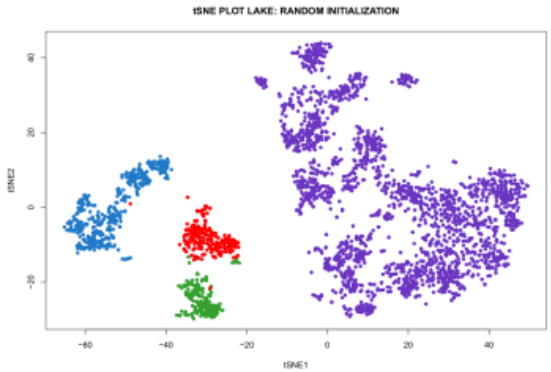

tSNE PLOT BJORKLUND: PCA INITIALIZATION
tSNE1
tSNE2

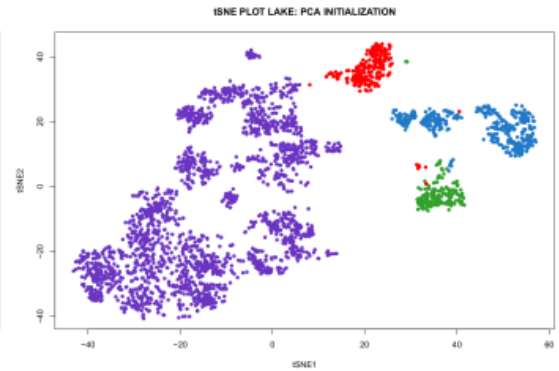

tSNE PLOT BJORKLUND: LAPLACIAN EIGENMAPS INITIALIZATION
tSNE1
tSNE2

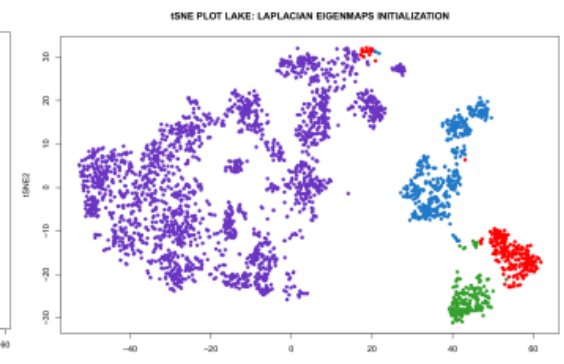

tSNE PLOT CAMPBELL: RANDOM INITIALIZATION
tSNE1
tSNE2

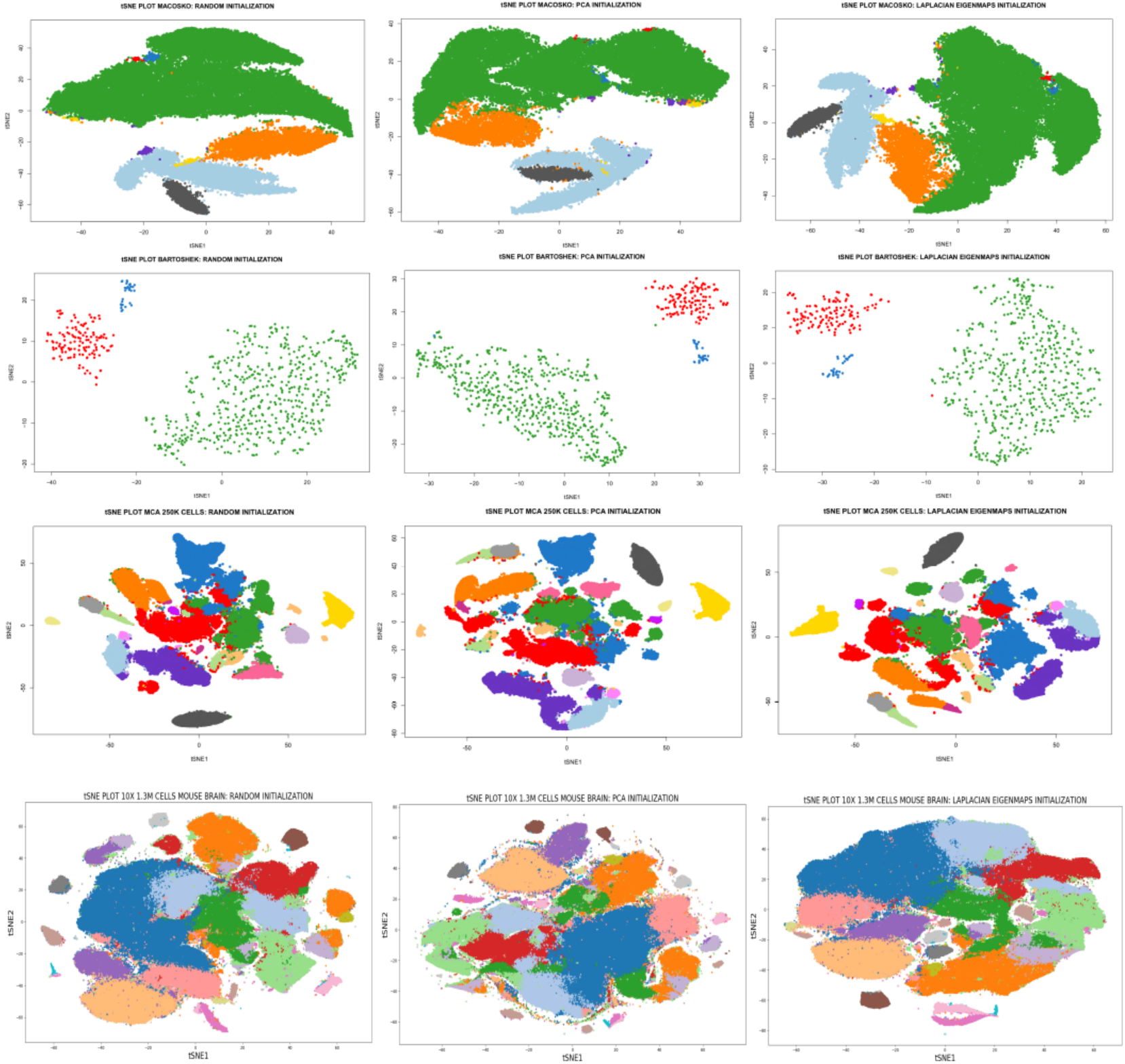

tSNE PLOT CAMPBELL: PCA INITIALIZATION
tSNE1
tSNE2


tSNE PLOT CAMPBELL: LAPLACIAN EIGENMAPS INITIALIZATION
tSNE1
tSNE2


tSNE PLOT LAKE: RANDOM INITIALIZATION
tSNE1
tSNE2

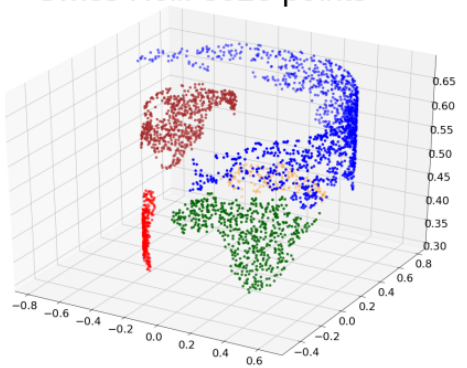

tSNE PLOT LAKE: PCA INITIALIZATION
tSNE1
tSNE2


tSNE PLOT LAKE: LAPLACIAN EIGENMAPS INITIALIZATION
tSNE1
tSNE2

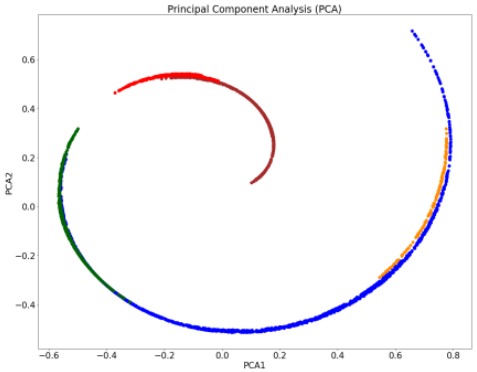

tSNE PLOT MACOSKO: RANDOM INITIALIZATION
tSNE PLOT MACOSKO: PCA INITIALIZATION
tSNE PLOT MACOSKO: LAPLACIAN EIGENMAPS INITIALIZATION
tSNE PLOT BARTOSHEK: RANDOM INITIALIZATION
tSNE PLOT BARTOSHEK: PCA INITIALIZATION
tSNE PLOT BARTOSHEK: LAPLACIAN EIGENMAPS INITIALIZATION
tSNE PLOT MCA 250K CELLS: RANDOM INITIALIZATION
tSNE PLOT MCA 250K CELLS: PCA INITIALIZATION
tSNE PLOT MCA 250K CELLS: LAPLACIAN EIGENMAPS INITIALIZATION
tSNE PLOT 10X 1.3M CELLS MOUSE BRAIN: RANDOM INITIALIZATION
tSNE PLOT 10X 1.3M CELLS MOUSE BRAIN: PCA INITIALIZATION
tSNE PLOT 10X 1.3M CELLS MOUSE BRAIN: LAPLACIAN EIGENMAPS INITIALIZATION
tSNE1
tSNE2

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

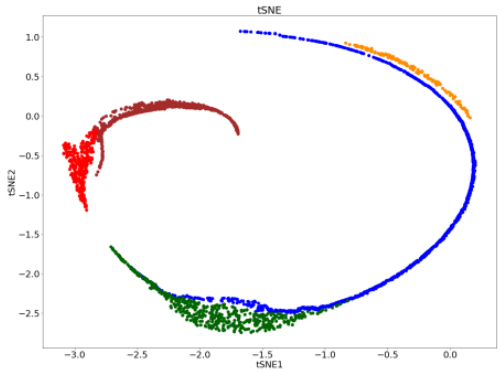
Swiss Roll: 3023 points



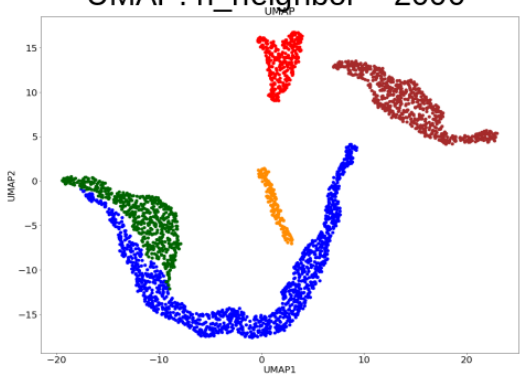
Principal Component Analysis



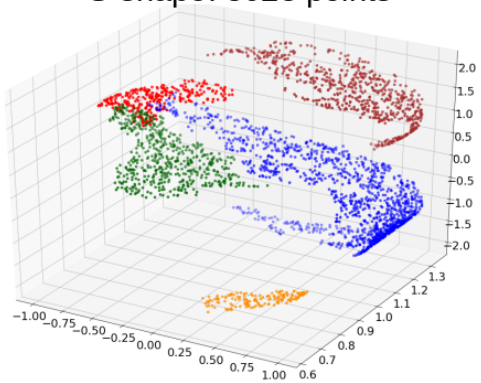
tSNE: perplexity = 2000

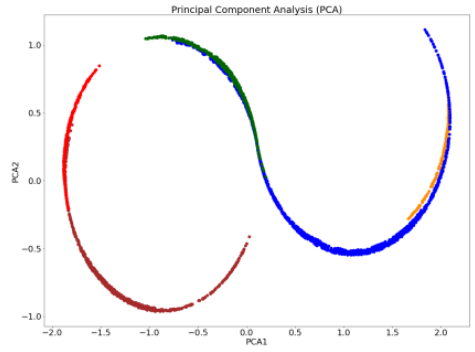


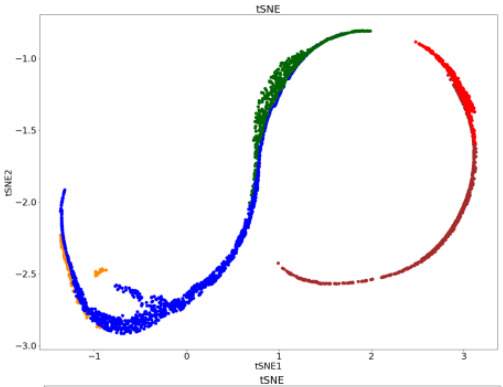
UMAP: n_neighbor = 2000

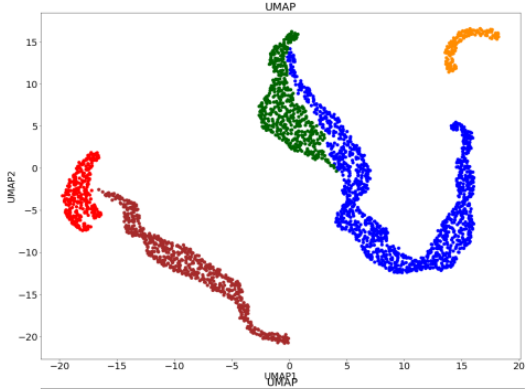


S-shape: 3023 points

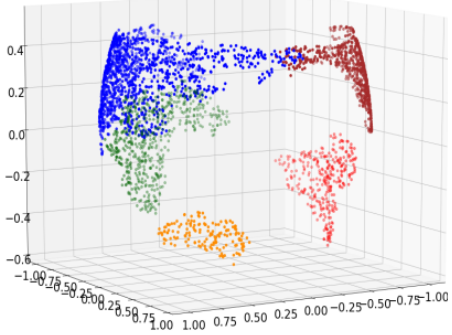


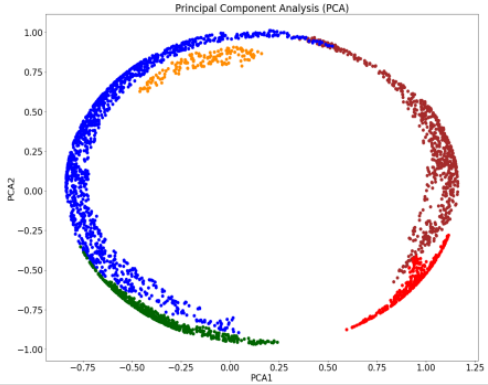


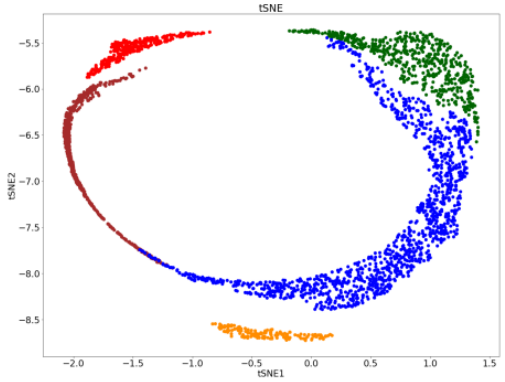


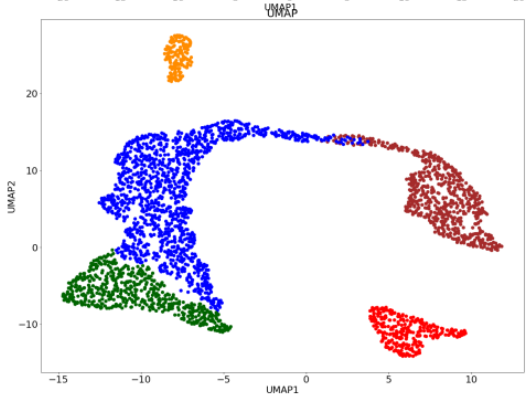


Sphere: 3023 points

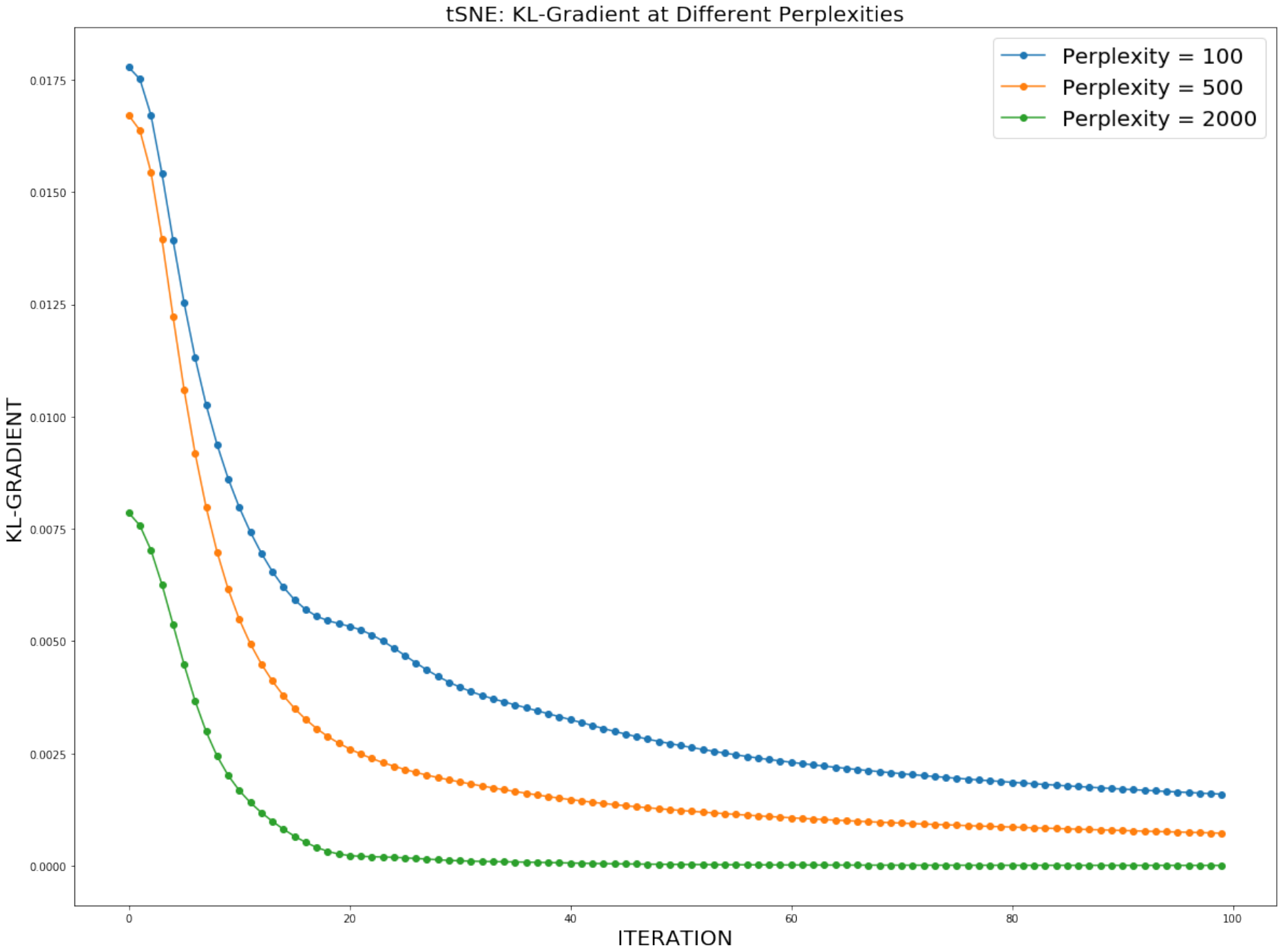




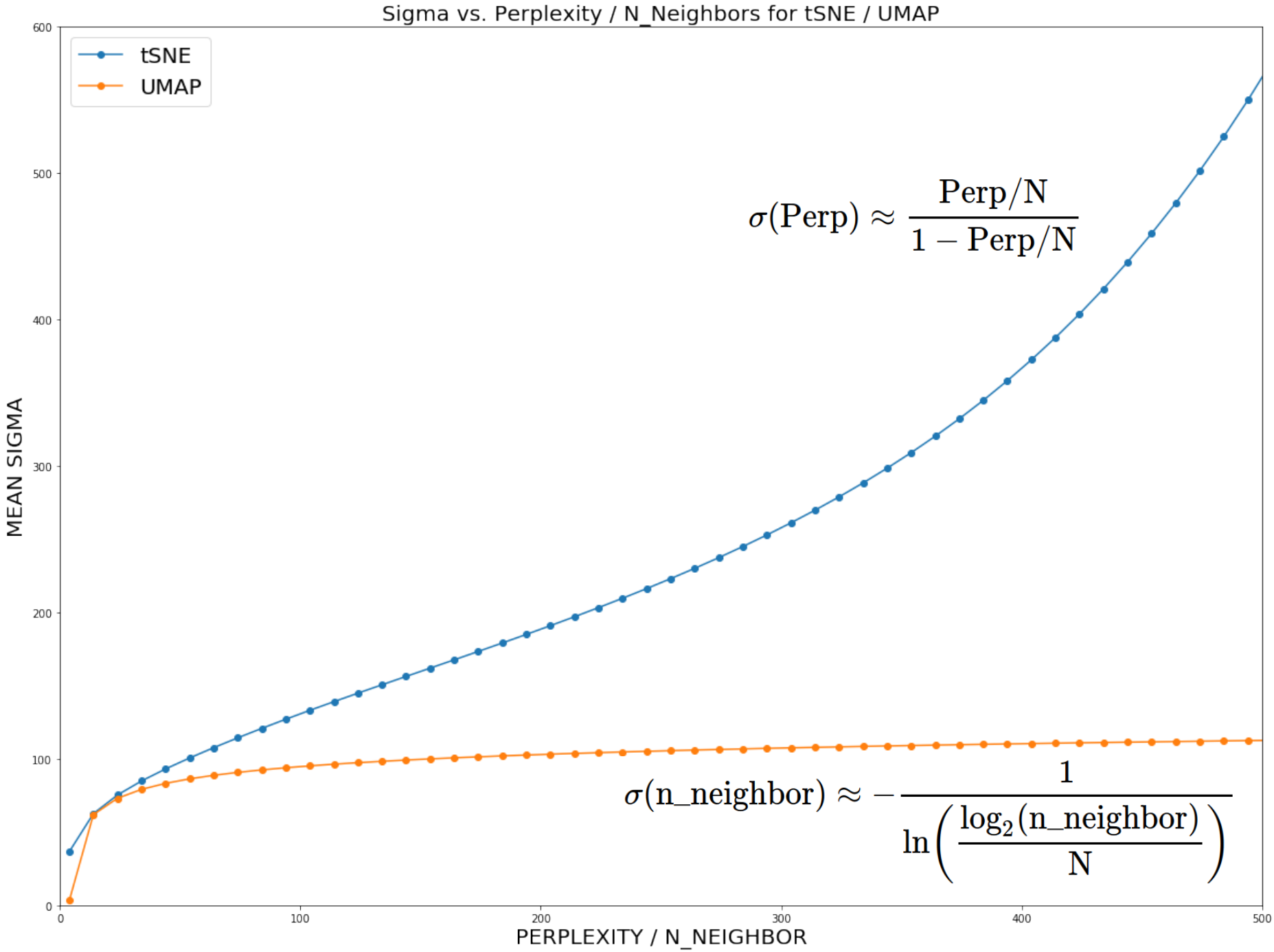




KL-Gradient Zero at Large Perplexity



Can Global Structure be Preserved at Large Perplexities?



Working with Large Perplexities: Gradient Disappears

$$y_i = y_i - \mu\frac{\partial KL}{\partial y_i}; \quad \frac{\partial KL}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)\frac{1}{1 + (y_i - y_j)^2}; \quad q_{ij} \approx \frac{1}{1 + (y_i - y_j)^2}; \quad p_{ij} \approx e^{-\frac{(x_i - x_j)^2}{2\sigma_i^2}}$$

In the limit $\sigma_i \to \infty$ the probability to observe points at a distance in high-dimensional space becomes $p_{ij} \to 1$. Therefore:

$$\frac{\partial KL}{\partial y_i} \approx 4\sum_j \left(1 - \frac{1}{1 + (y_i - y_j)^2}\right)(y_i - y_j)\frac{1}{1 + (y_i - y_j)^2} = 4\sum_j \frac{(y_i - y_j)^3}{(1 + (y_i - y_j)^2)^2}$$

In the limit of close embedding points:

$$y_i - y_j \to 0: \quad \frac{\partial KL}{\partial y_i} \approx 4\sum_i (y_i - y_j)^3 \to 0$$

In the limit of distant embedding points:

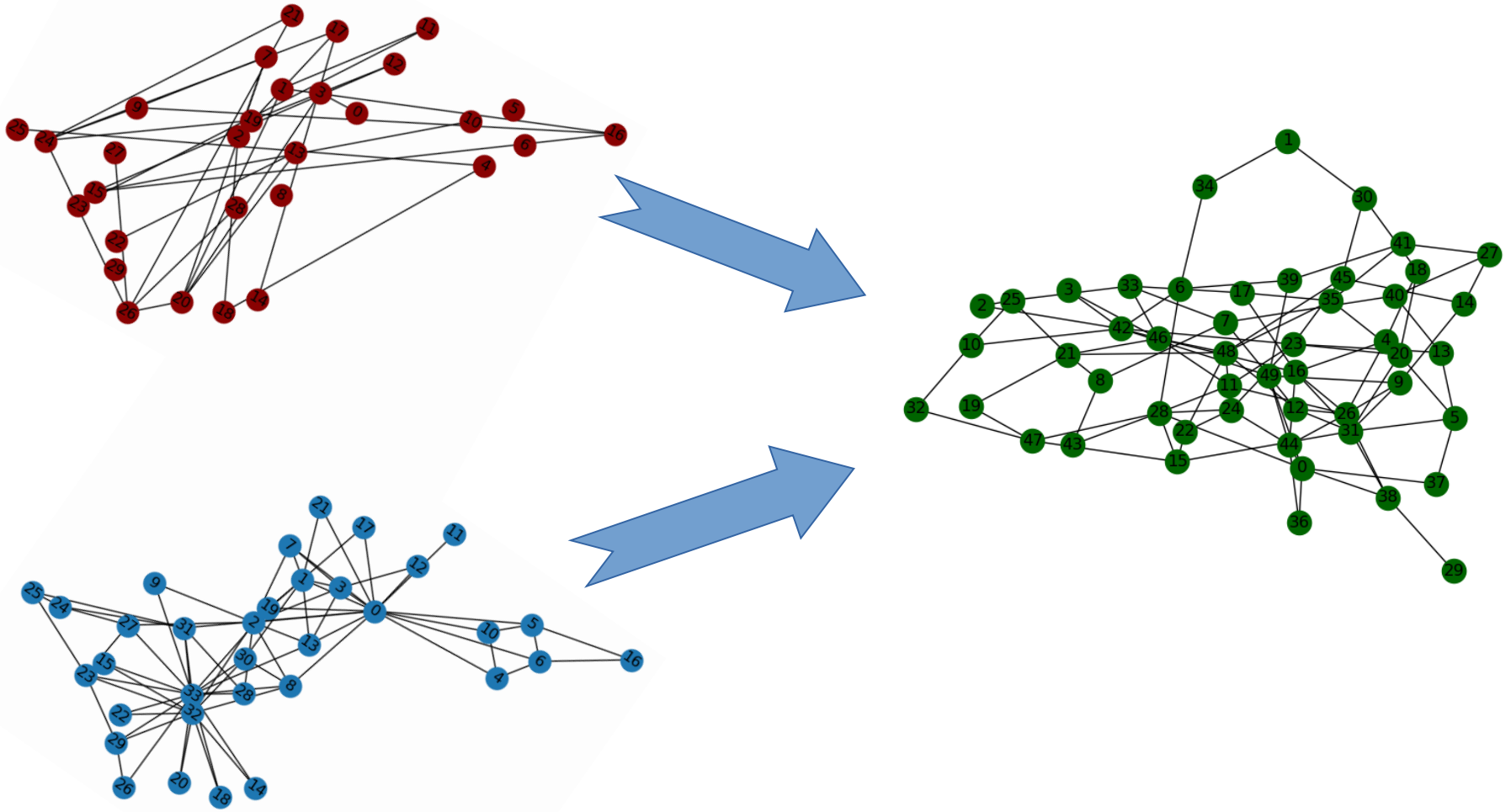
$$y_i - y_j \to \infty: \quad \frac{\partial KL}{\partial y_i} \approx 4\sum_i \frac{1}{y_i - y_j} \to 0$$

Large Perplexity might lead to disappearance of the KL-gradient,

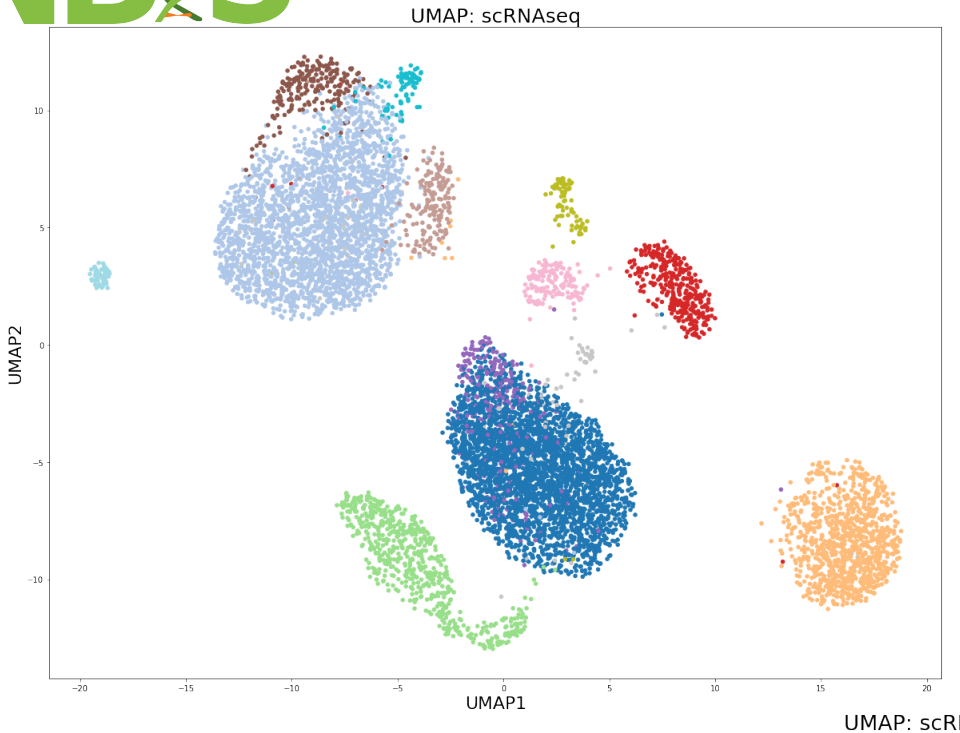
therefore **tSNE might degrade to something linear**,

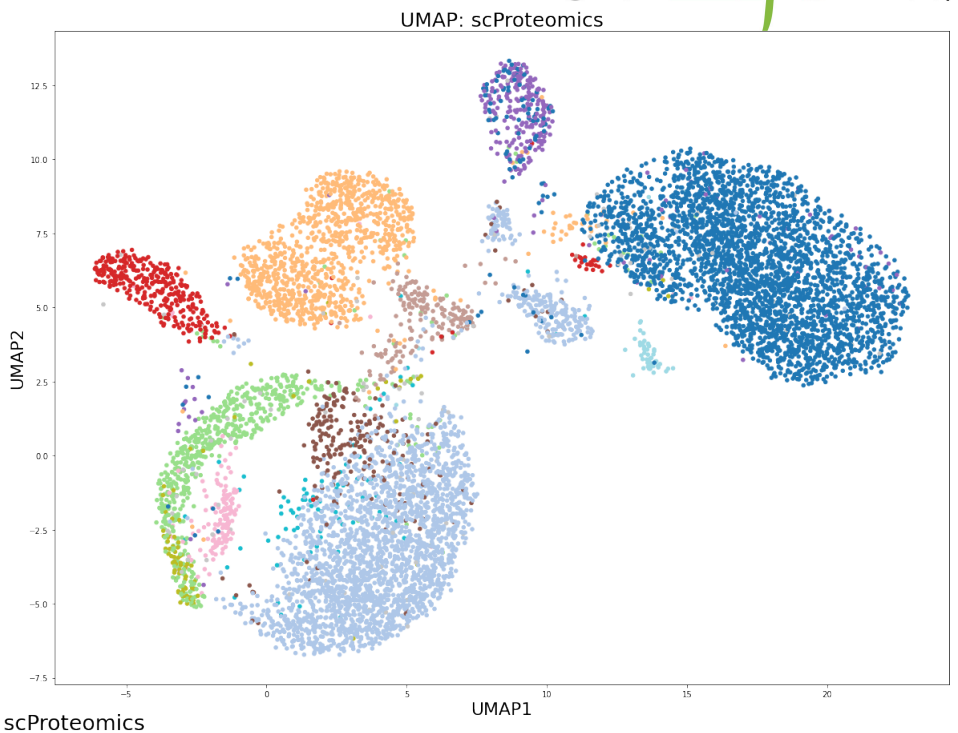
so if you start with PCA initialization and increase Perplexity, you end up with PCA

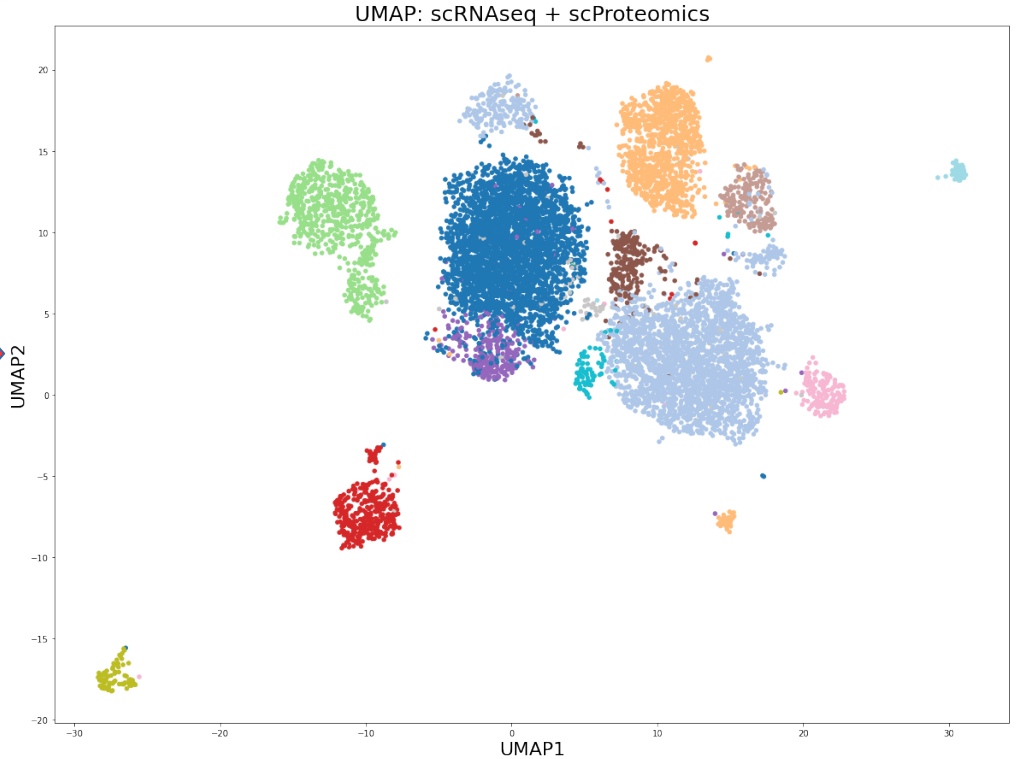
Graph-Based OMICs Integration



UMAP for Omics Integration







National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

Lunds Universitet