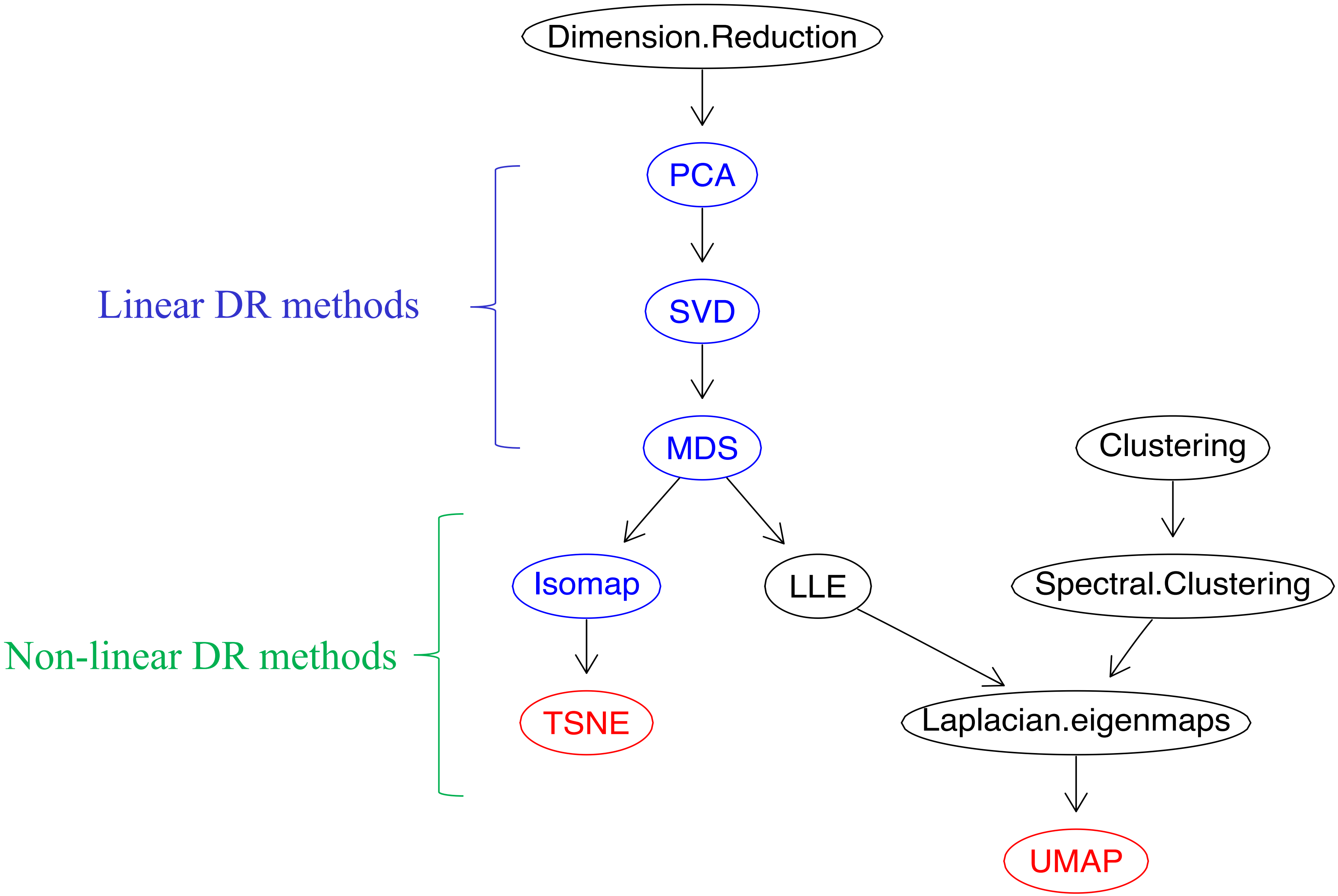
Dimension Reduction Methods:
From PCA to TSNE and UMAP

Maxwell Lee

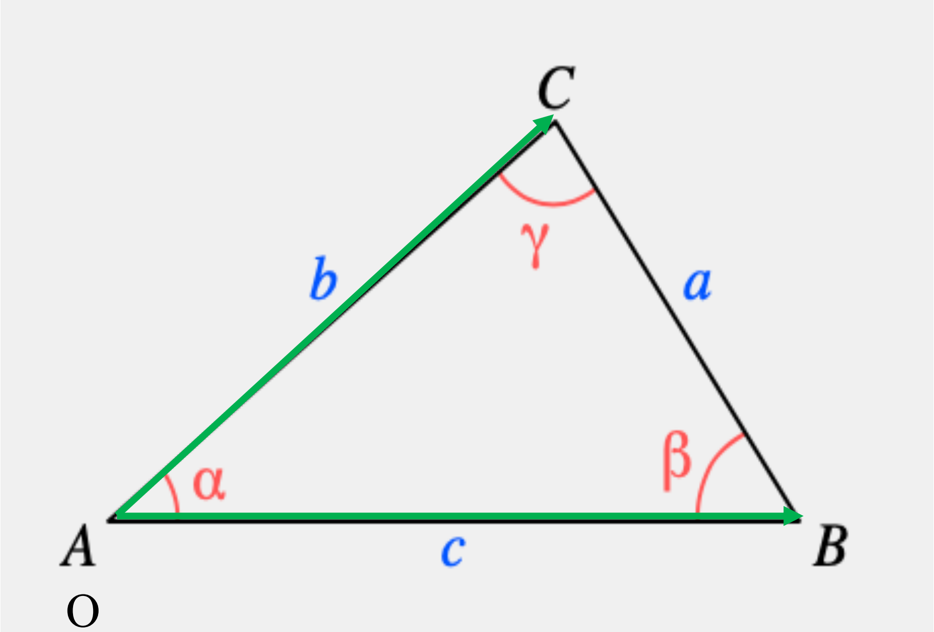
High-dimension Data Analysis Group
Laboratory of Cancer Biology and Genetics
Center for Cancer Research
National Cancer Institute

May 7, 2020

Road Map for Dimension Reduction Methods



The Dot Product of Two Vectors is the Difference Between the Squared Distances (Law of Cosines)



$$a^2 = b^2 + c^2 - 2bc\cos(\alpha)$$

$$bc\cos(\alpha) = -\tfrac{1}{2}(a^2 - b^2 - c^2)$$

$$\boxed{\mathbf{b} \bullet \mathbf{c} = -1/2(a^2 - b^2 - c^2)}$$

Warren Torgerson in 1958

Eigen Decomposition of Gram Matrix (Similarity Matrix)

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ . & & & . \\ . & & & . \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix}$$

g_{ij} is dot product between element i and j which captures similarity or relatedness

$$G = U\Lambda U^T$$

$$Z = U\Lambda^{1/2}$$

G: Gram matrix or kernel matrix

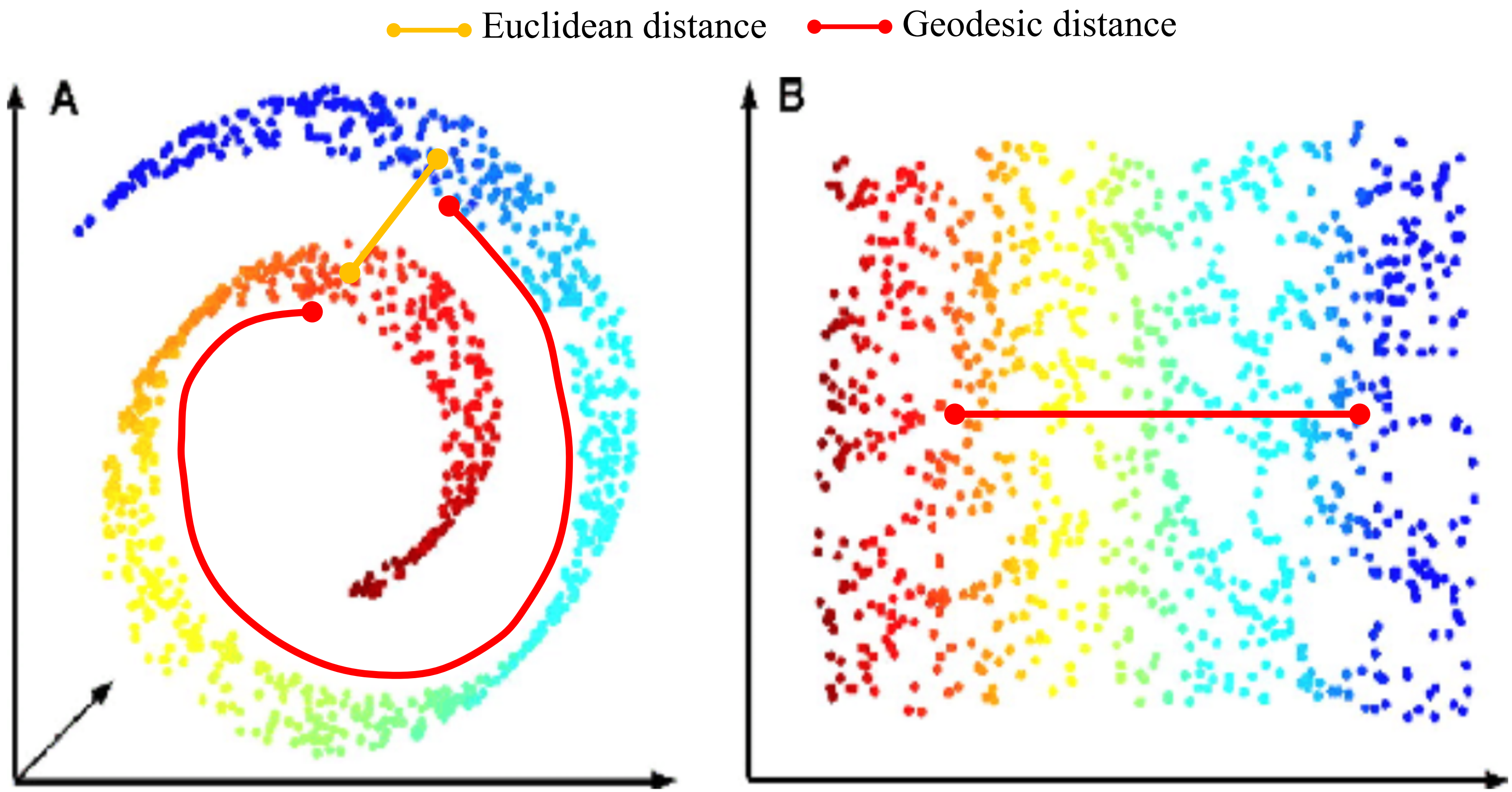
U: Eigen vector

Λ: Eigen value

Z: principal component

Nonlinear Dimension Reduction of Swiss Roll Dataset

Swiss roll manifold in 3D $\xrightarrow{\text{unfolding}}$ 2D sheet

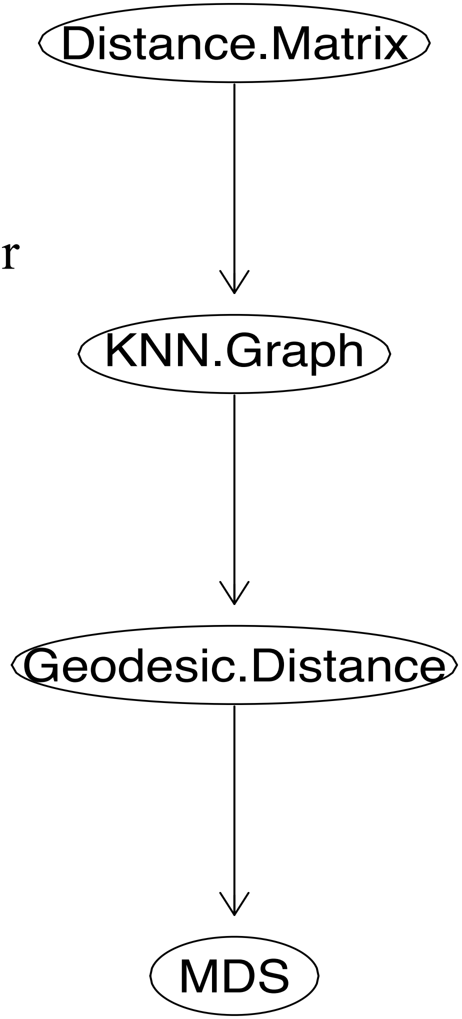


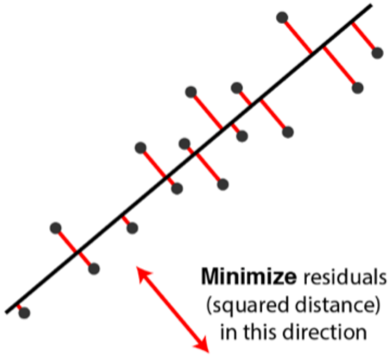
3-dimension

2-dimension

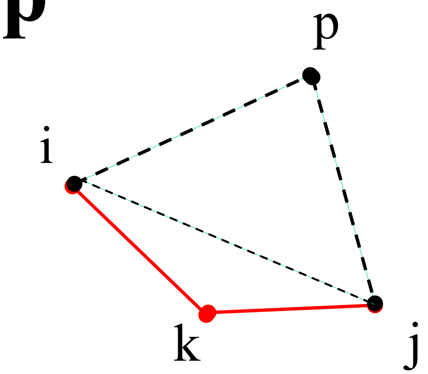
Algorithm of Isomap

KNN: k nearest neighbor





Joshua Tenenbaum et al Science 2000



$d_{ij} = d_{ik} + d_{kj}$

d_{ij} through k is the shortest path.

$$E = \|\tau(D_G) - \tau(D_Y)\|_{L^2}$$

$\tau(D) = -1/2(HDH)$

$H = I - 1/n(ee^T)$

E: cost function

D_G: distance matrix in high dim

D_Y: distance matrix in low dim

τ: transform D to Gram matrix

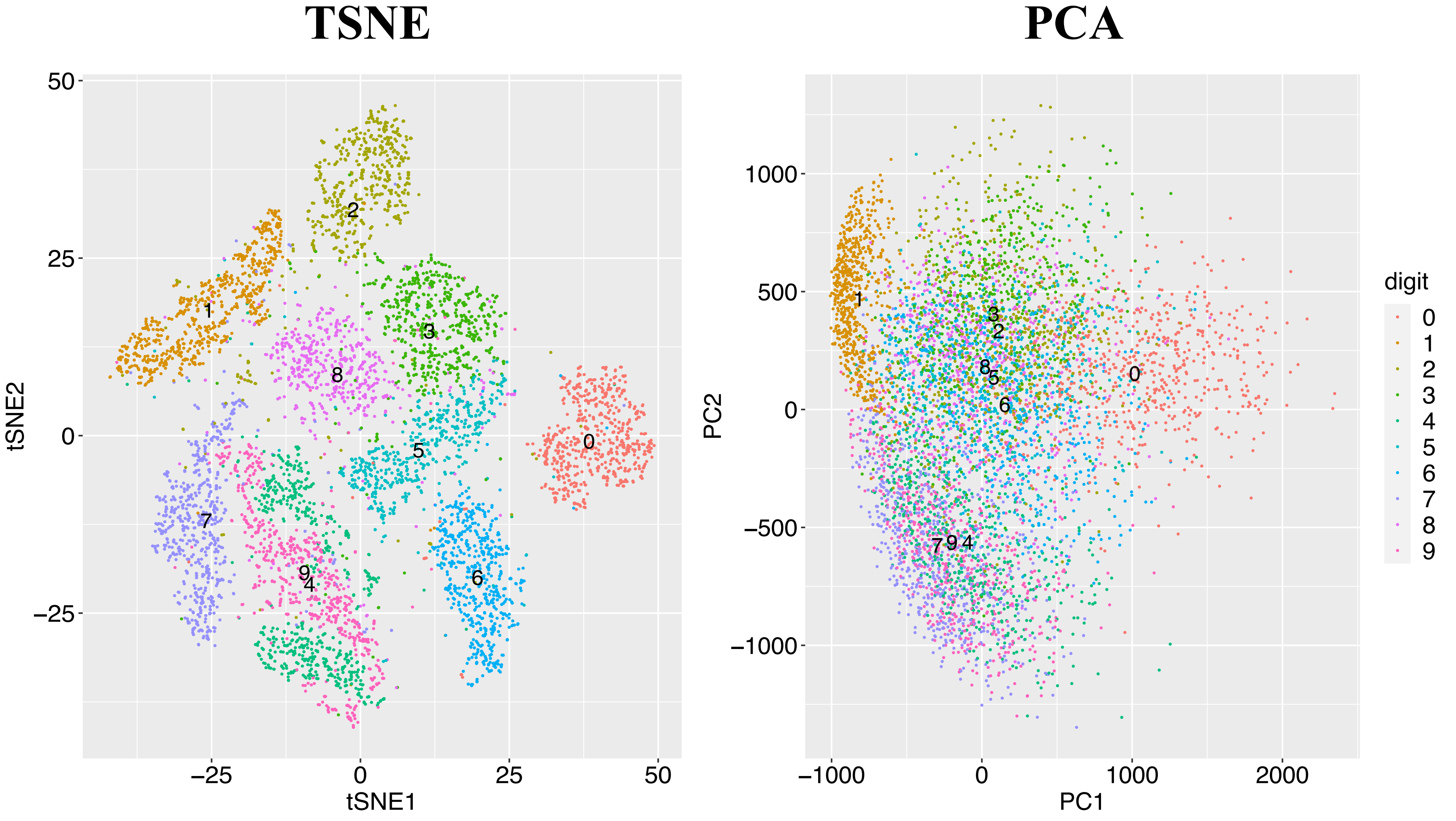
H: centering matrix

I: identity matrix

e: vector of 1

TSNE Versus PCA of the Same MNIST Dataset

sample size n=6000

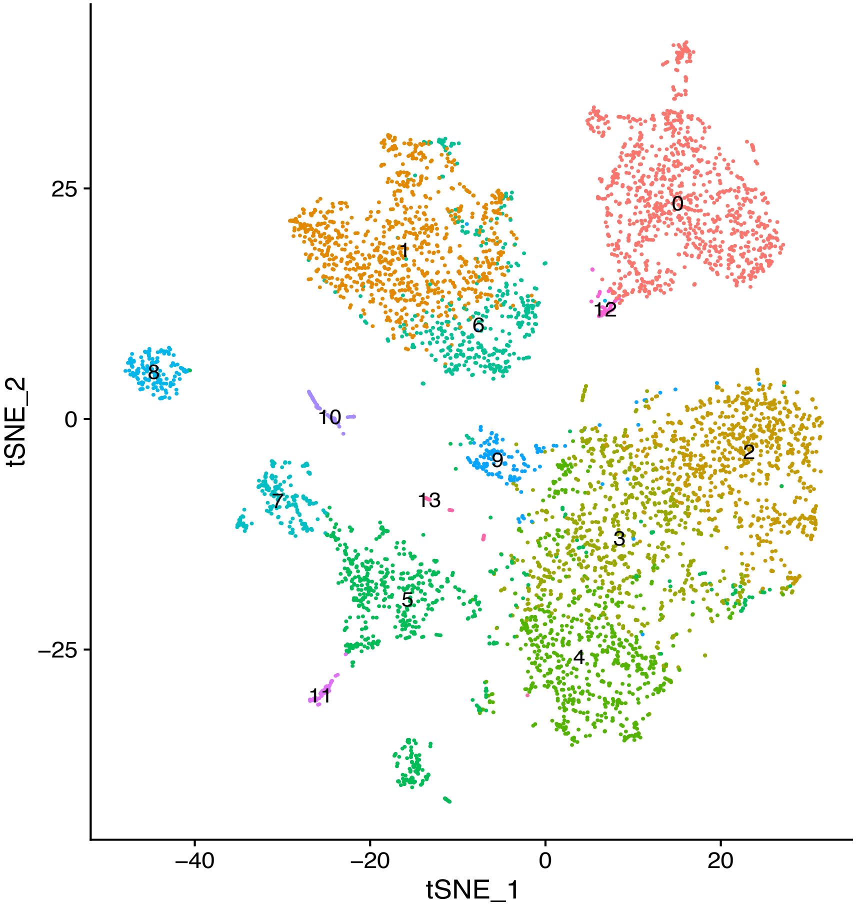


Laurens van der Maaten and Geoffrey Hinton, JMLR 2008

TSNE vs. PCA of a Single Cell RNAseq Data

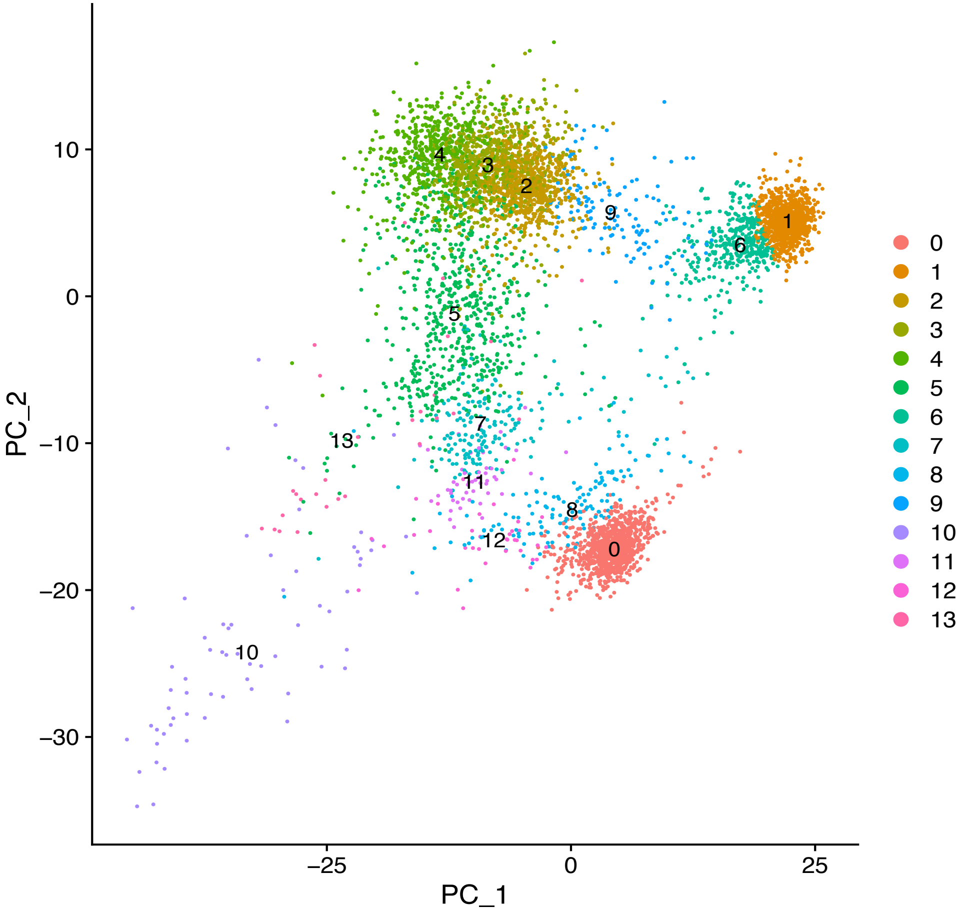
cell number n~6000
Clusters were identified before TSNE and PCA analysis

TSNE



Cells in cluster are more spread out.

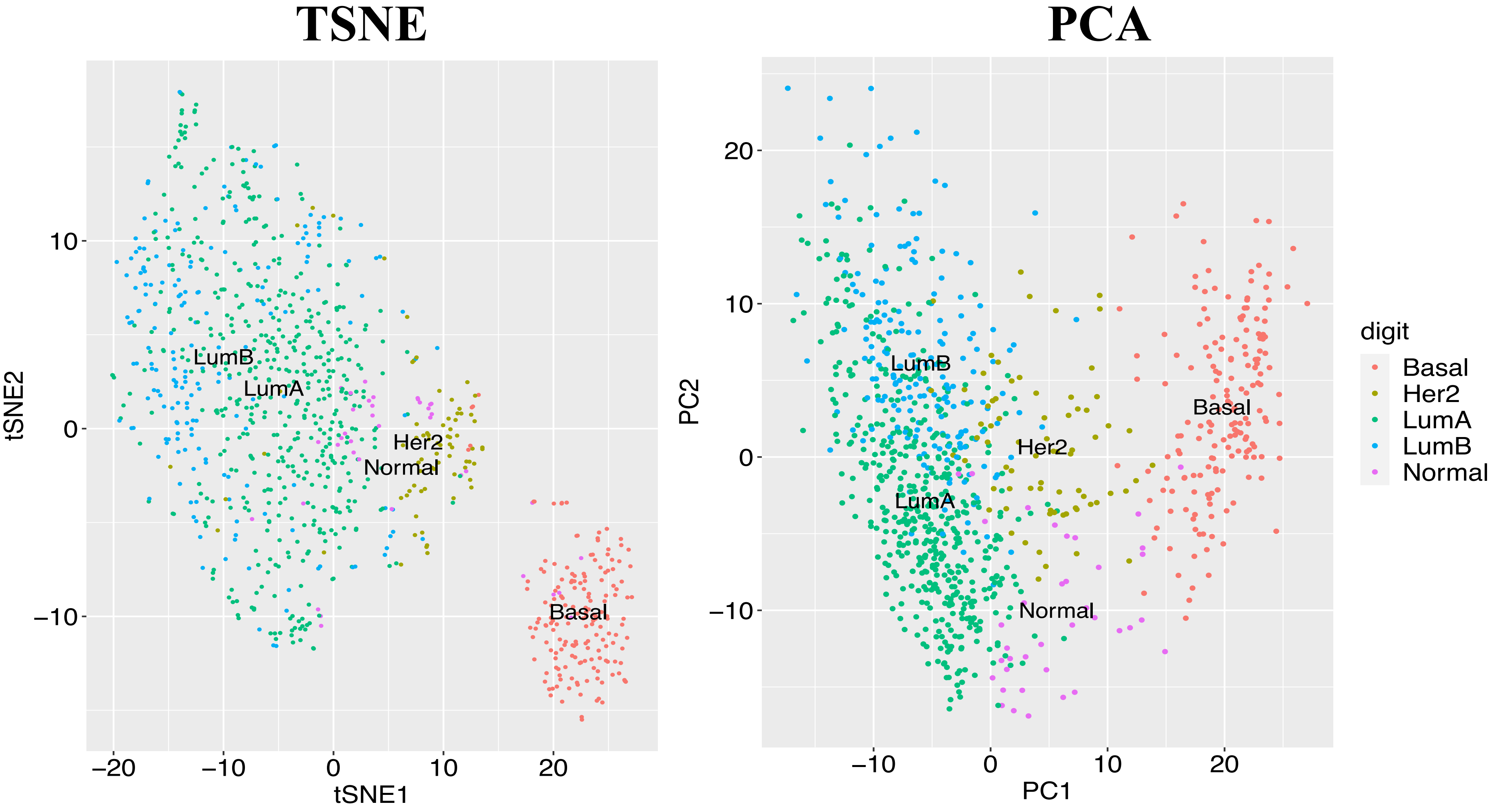
PCA



Variance of PC is driven by outliers.

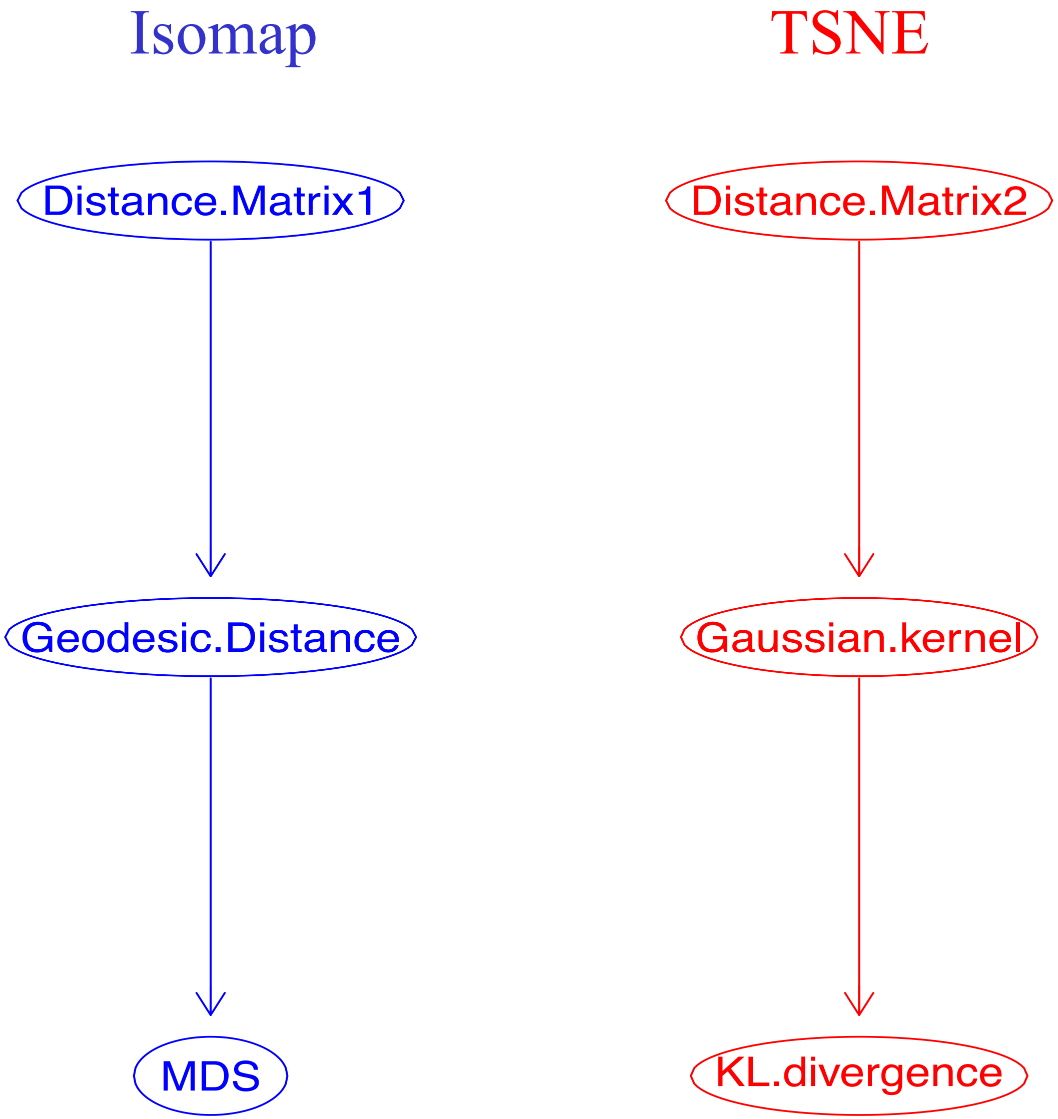
TSNE vs. PCA of TCGA Breast Cancer Data

sample size n=977


TSNE

PCA

Isomap vs. TSNE



Isomap

Distance.Matrix1

Geodesic.Distance

MDS

TSNE

Distance.Matrix2

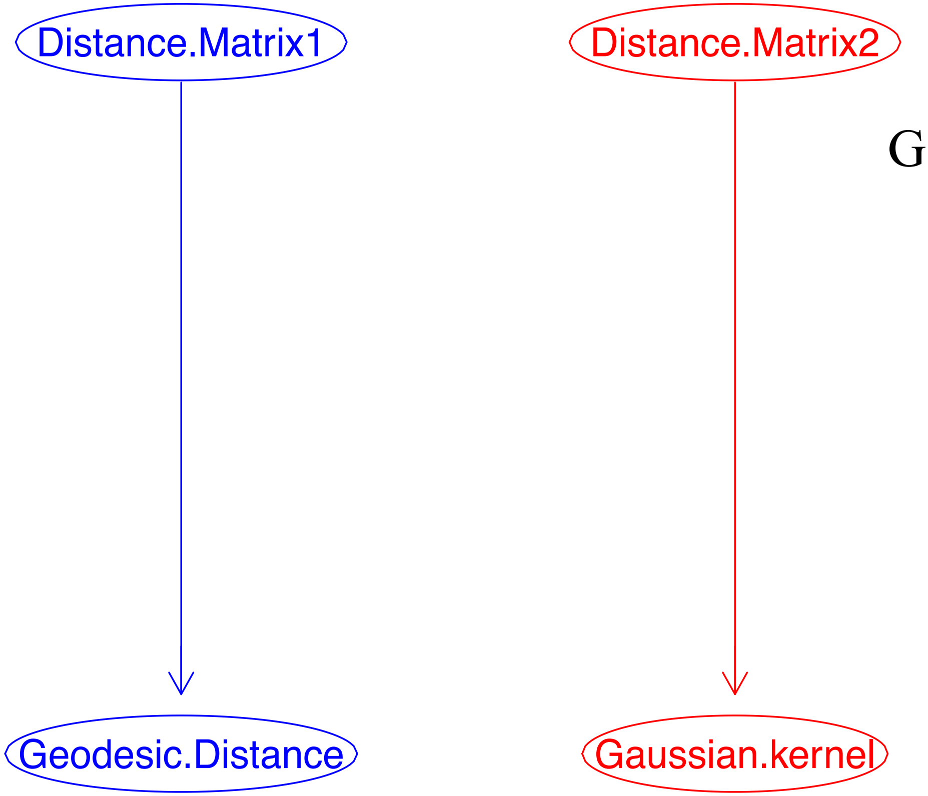
Gaussian.kernel

KL.divergence

Isomap vs. TSNE

Isomap

TSNE



Distance.Matrix1

Geodesic.Distance

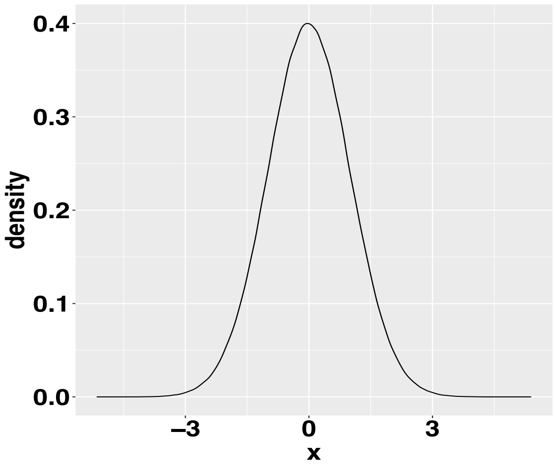
Distance.Matrix2

Gaussian.kernel

Gaussian density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$z^2 = (x - \mu)^2/\sigma^2$$

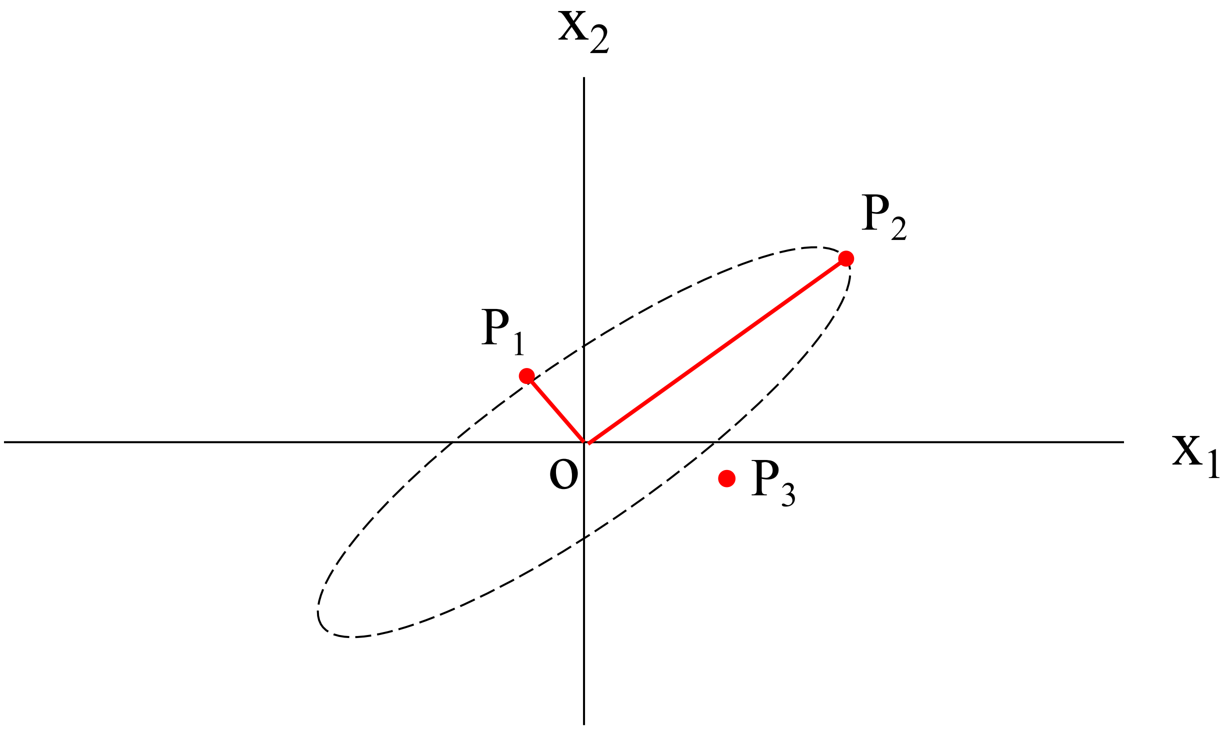


D: distance matrix

G: Gram matrix

$$D \longrightarrow G$$

Euclidean Distance vs Mahalanobis Distance



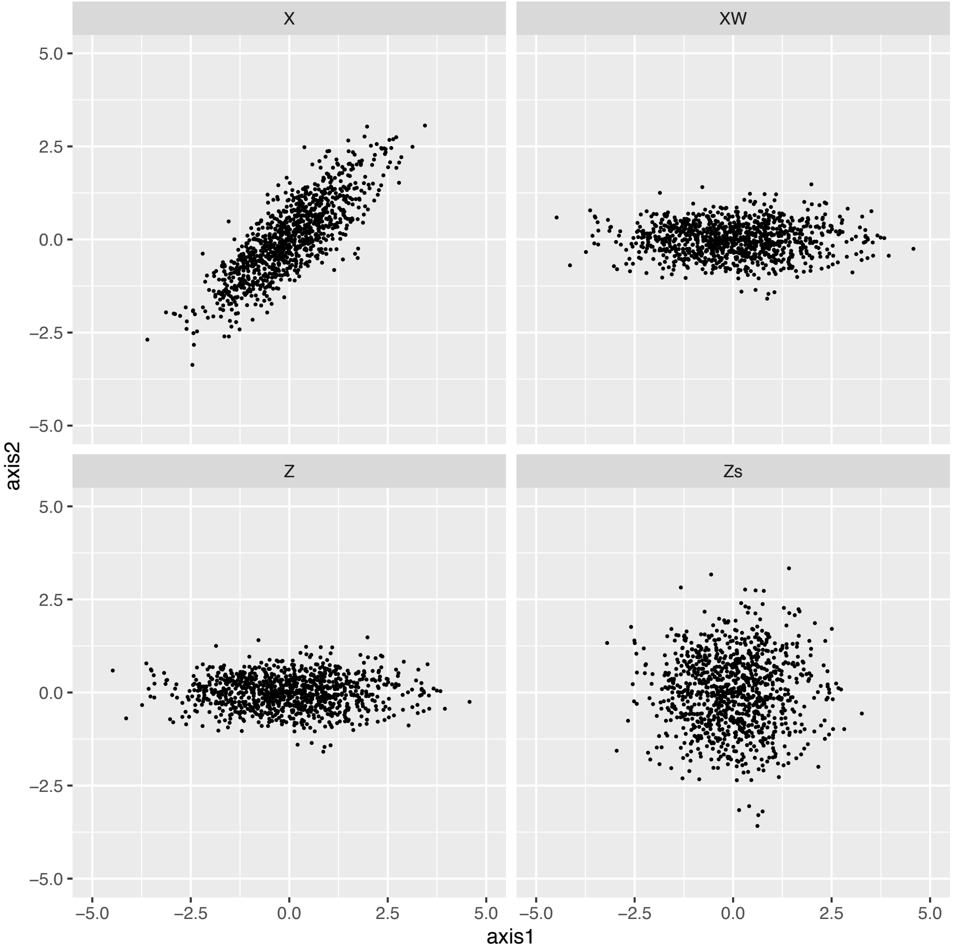
Euclidean distance: $P_1 < P_3 < P_2$

Probability: $p_1 = p_2 > p_3$

Mahalanobis distance is a statistical distance related to probability

Prasanta Chandra Mahalanobis in 1936

Multivariate Gaussian Distribution



Σ: covariance matrix

Σ^{-1}: inverse of Σ

Λ: Diagonal matrix with Eigen values

W: Eigen vectors

Z: Principal Components

Z_s: Standardized Z

z: a sample from Z_s

T: Transposition

μ: mean vector

$$Z = XW$$

$$Z_s = XW\Lambda^{-1/2}$$

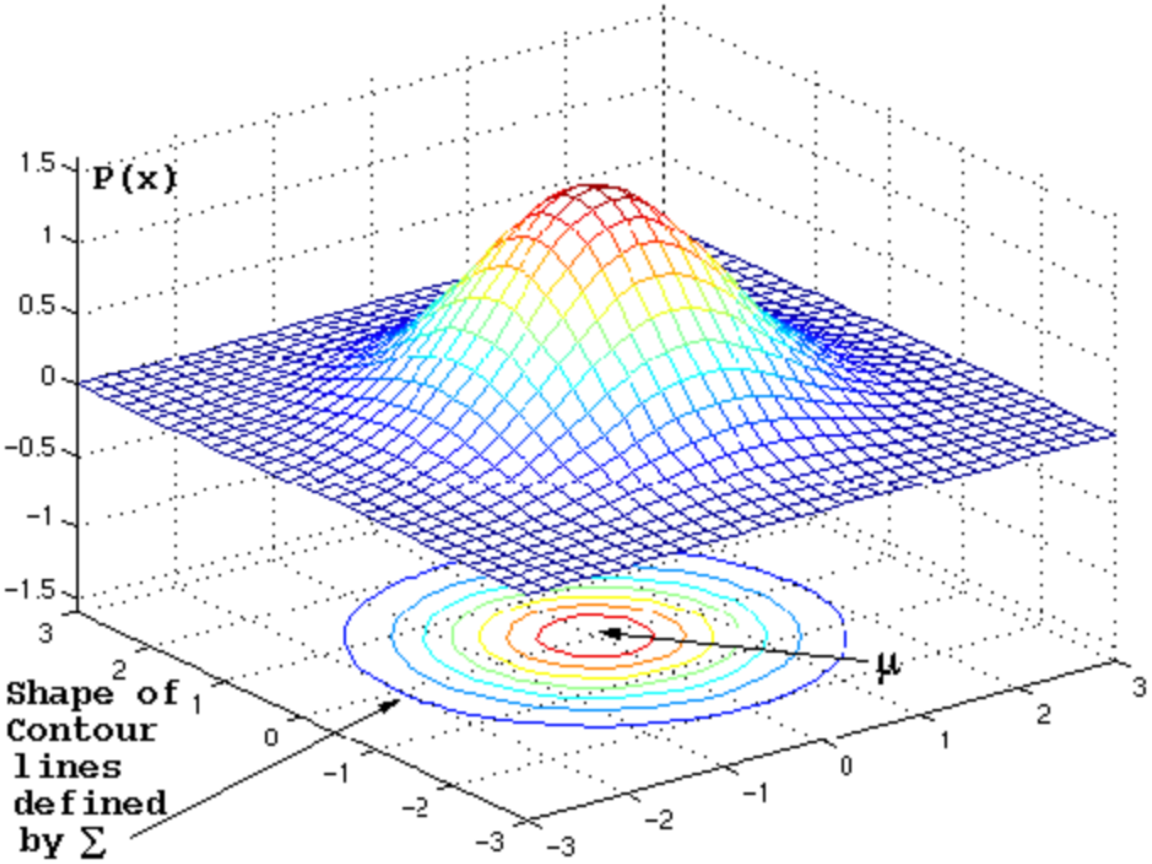
$$z = \Lambda^{-1/2}W^T x$$

$$z^T z = x^T W\Lambda^{-1/2}\Lambda^{-1/2}W^T x$$

$$z^T z = x^T \Sigma^{-1} x$$

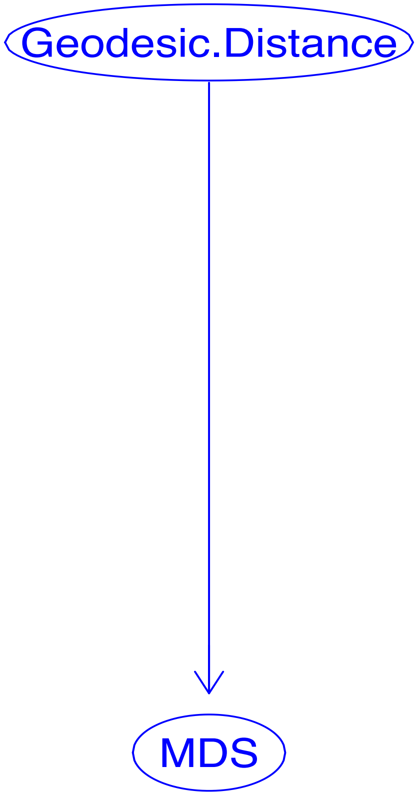
$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Multivariate Gaussian Distribution



T-distributed Stochastic Neighbor Embedding (TSNE)

Isomap



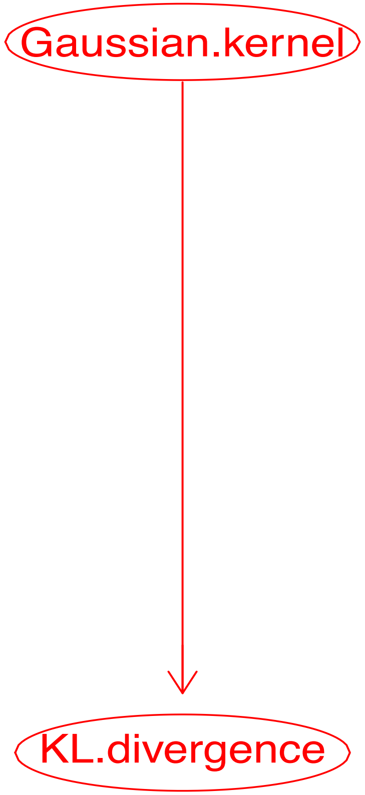
Geodesic.Distance

MDS

$$G = U \Lambda U^{T}$$

$$Z = U \Lambda^{1/2}$$

TSNE



Gaussian.kernel

KL.divergence

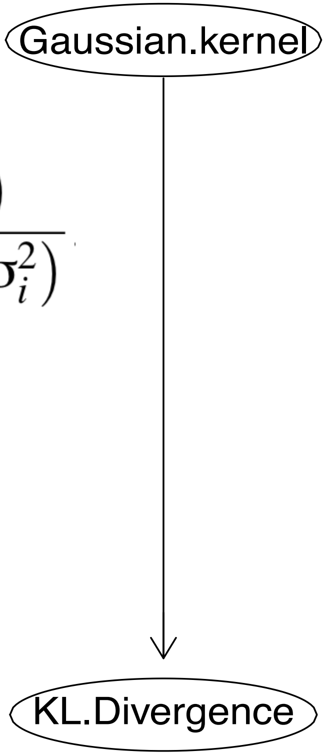
$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

T-distributed Stochastic Neighbor Embedding (TSNE)

High dimension

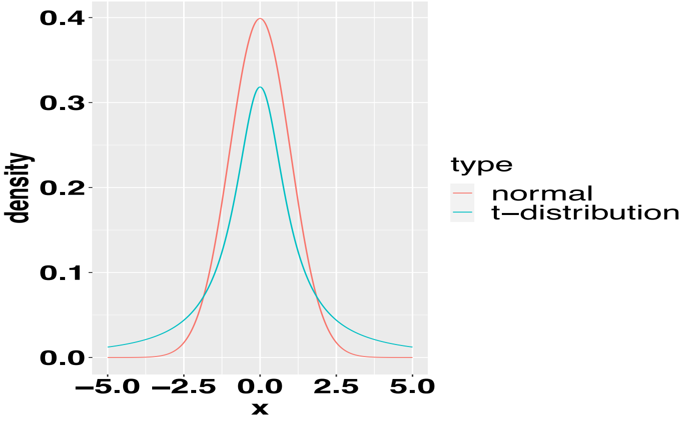
$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)}$$

$$p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$$



Low dimension

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



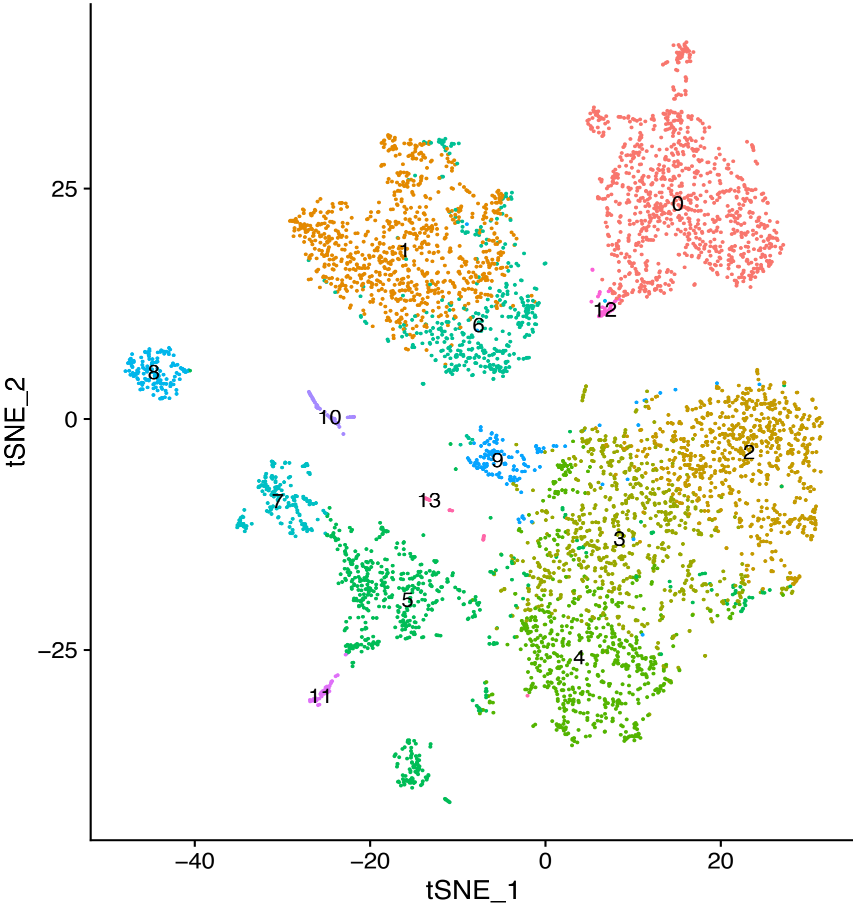
$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

$$\frac{\delta C}{\delta y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1+\|y_i - y_j\|^2\right)^{-1}$$

TSNE vs. PCA of a Single Cell RNAseq Data

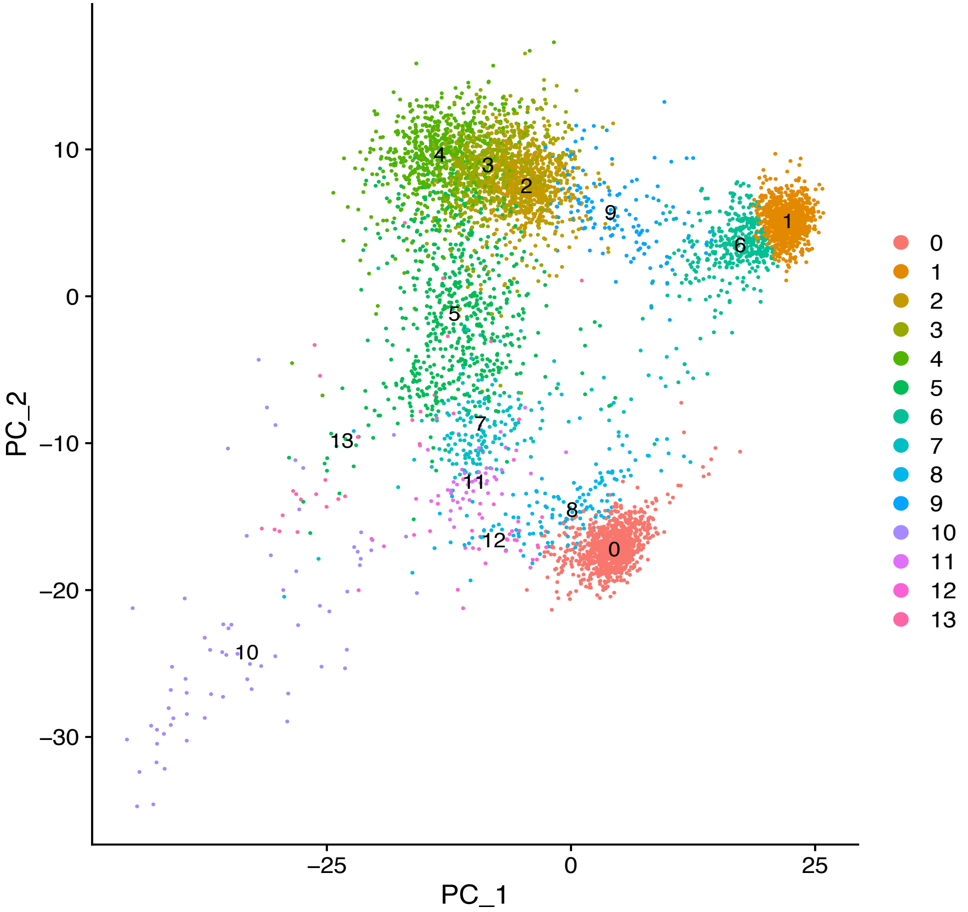
cell number n~6000
Clusters were identified before TSNE and PCA analysis

TSNE



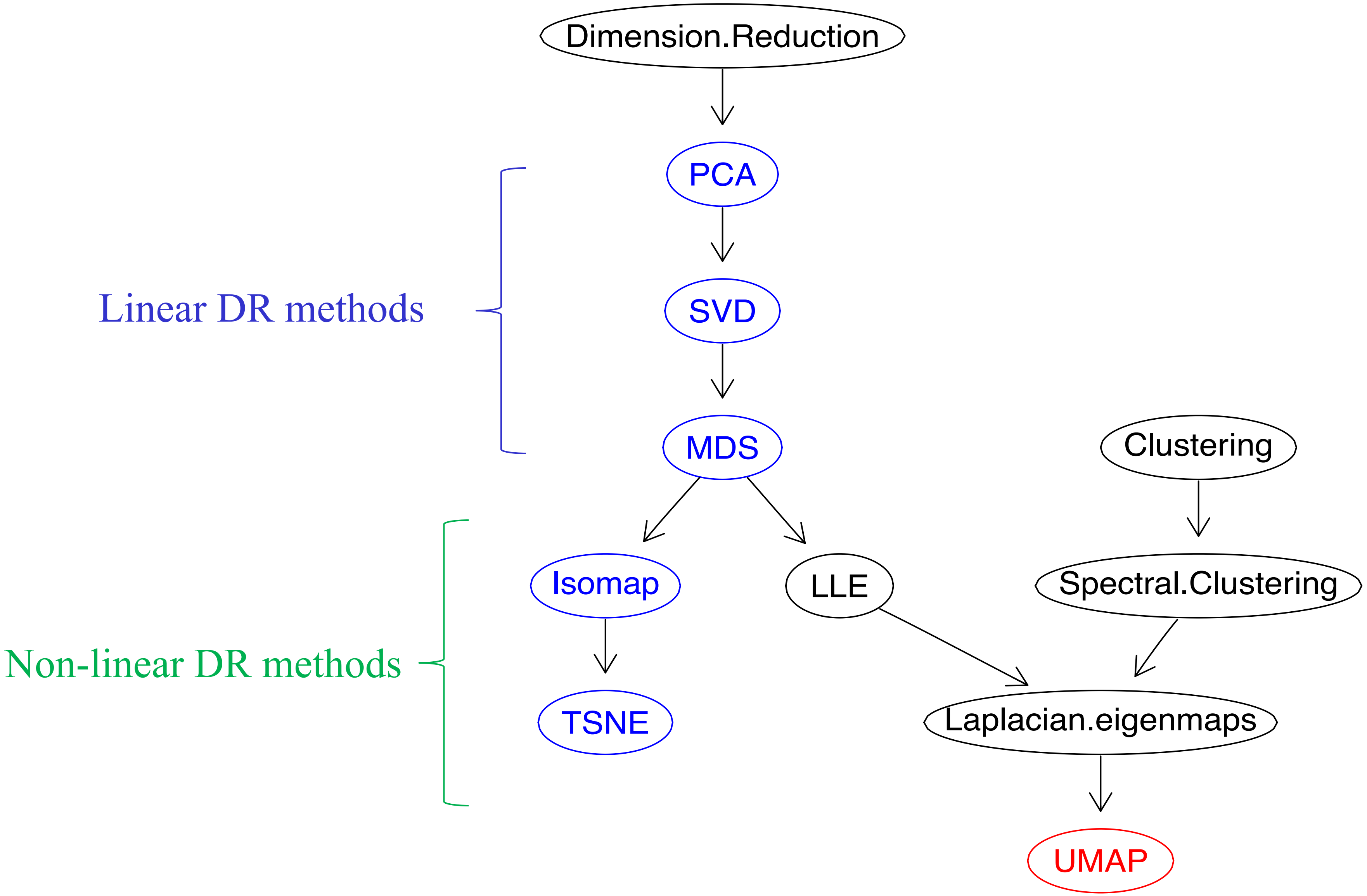
Cells in cluster are more spread out.

PCA



Variance of PC is driven by outliers.

Road Map for Dimension Reduction Methods



Dimension.Reduction

PCA

SVD

MDS

Clustering

Isomap

LLE

Spectral.Clustering

TSNE

Laplacian.eigenmaps

UMAP

Linear DR methods

Non-linear DR methods

Uniform Manifold Approximation and Projection (UMAP)

TSNE is pretty good.

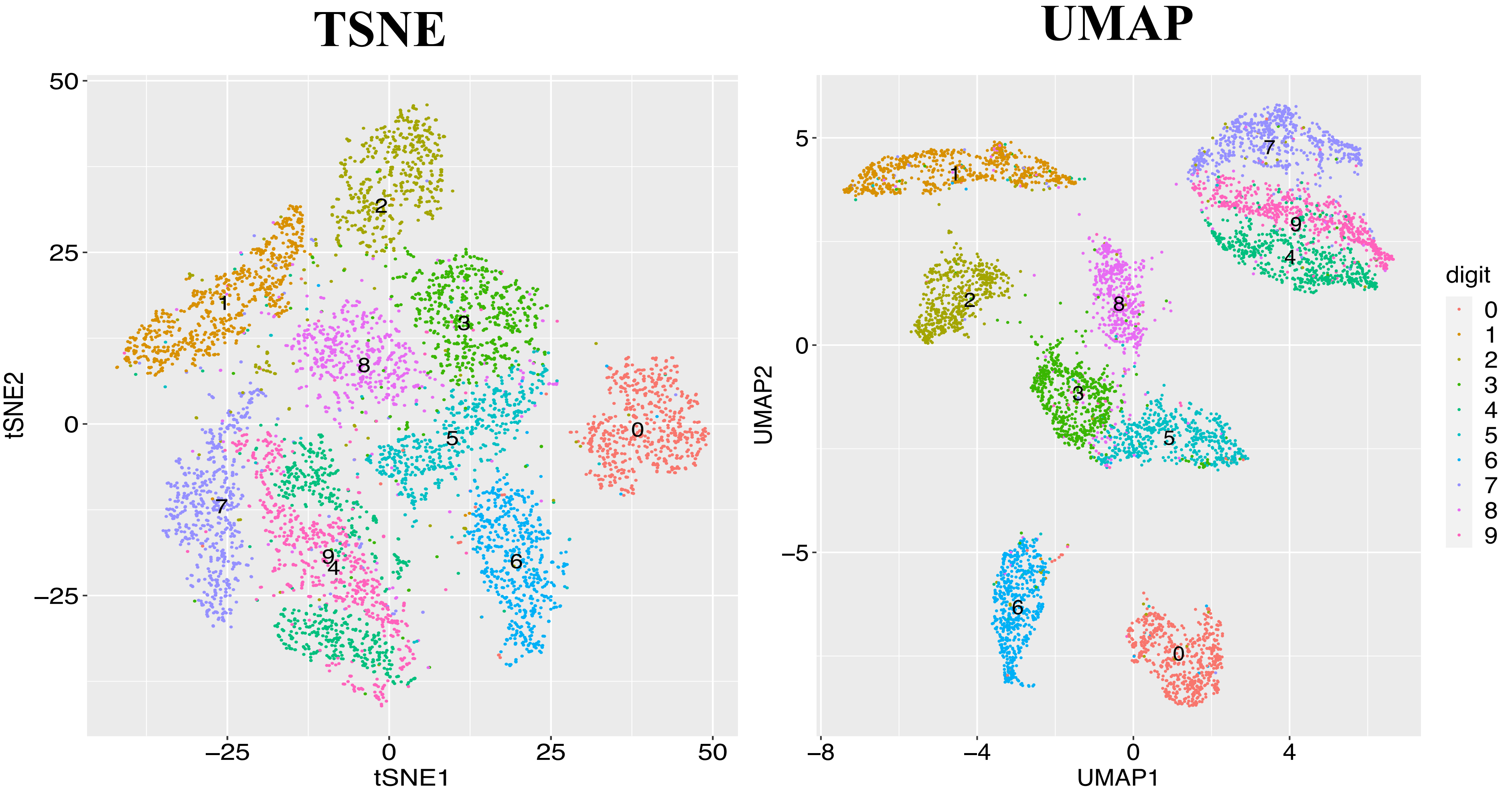
Why do we need UMAP?

	TSNE	UMAP
speed	moderate	fast
Structure preserved	local and global	local and global
Number of components	2	2 or more

Leland McInnes, John Healy, and James Melville arXiv 2018

TSNE Versus UMAP of the MNIST Dataset

sample size n=6000

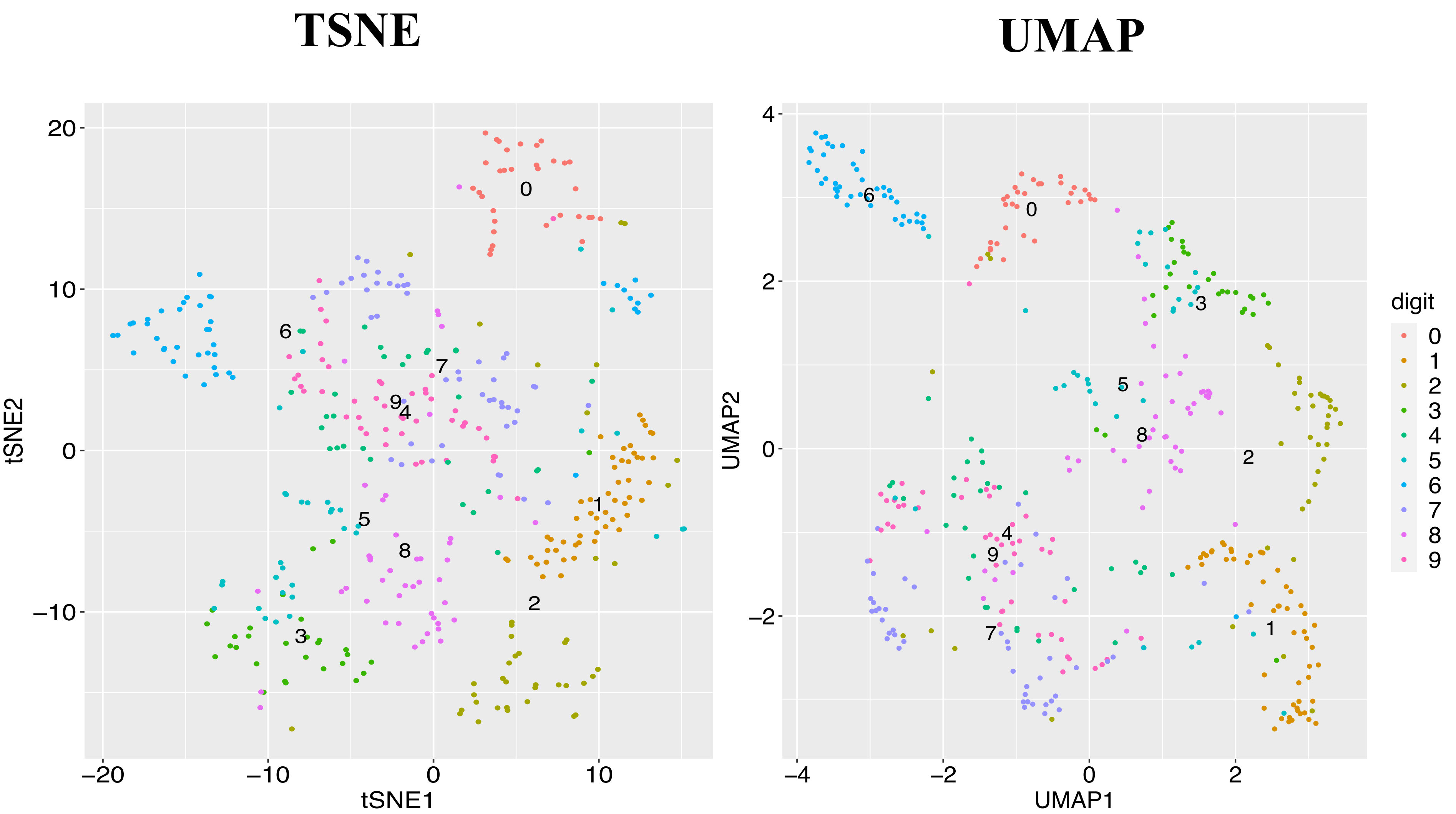


TSNE

UMAP

TSNE Versus UMAP of the Same MNIST Dataset

sample size n=400

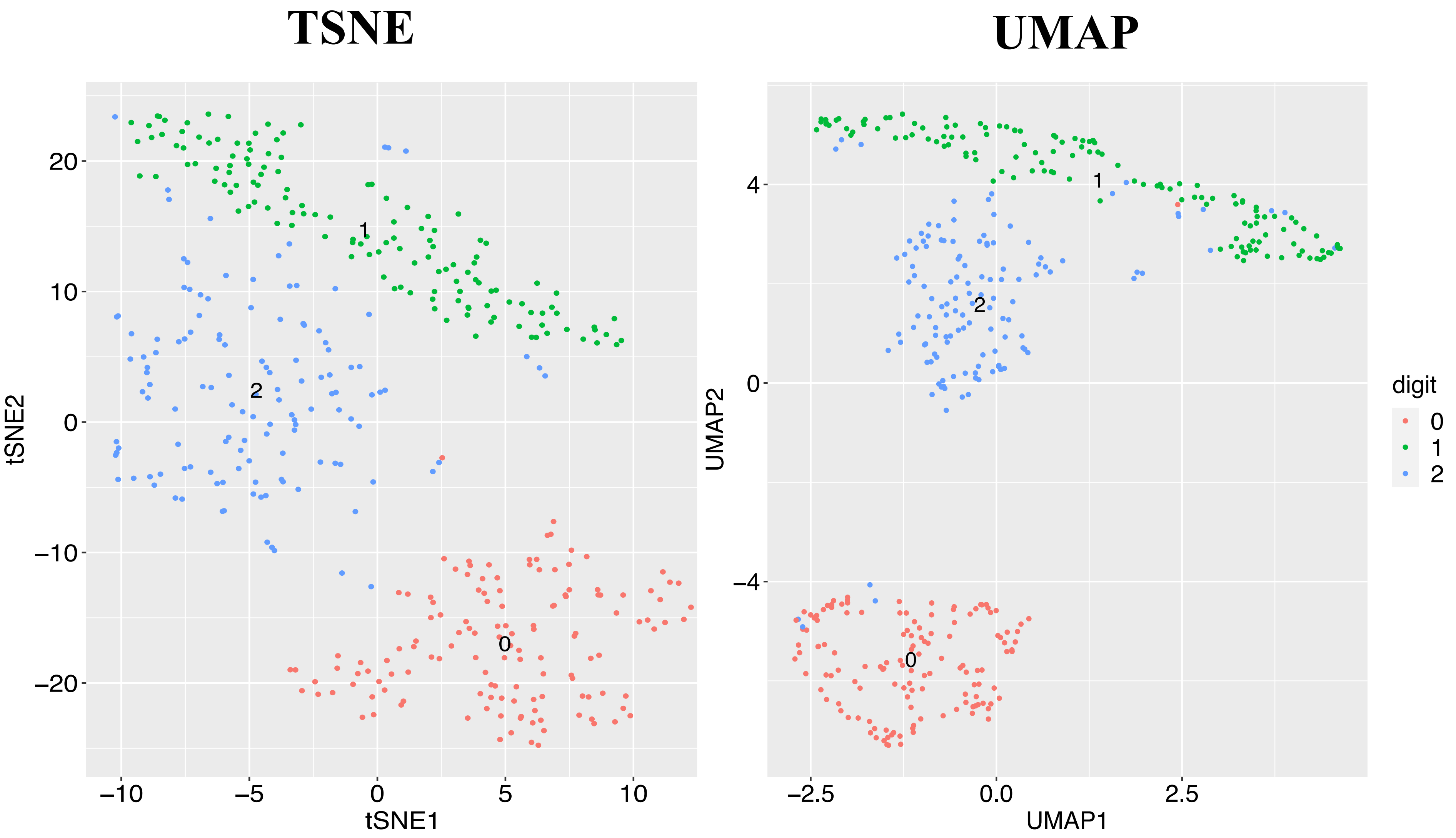


TSNE

UMAP

TSNE Versus UMAP of the Same MNIST Dataset

sample size n=400 and digits of 0, 1, and 2



TSNE

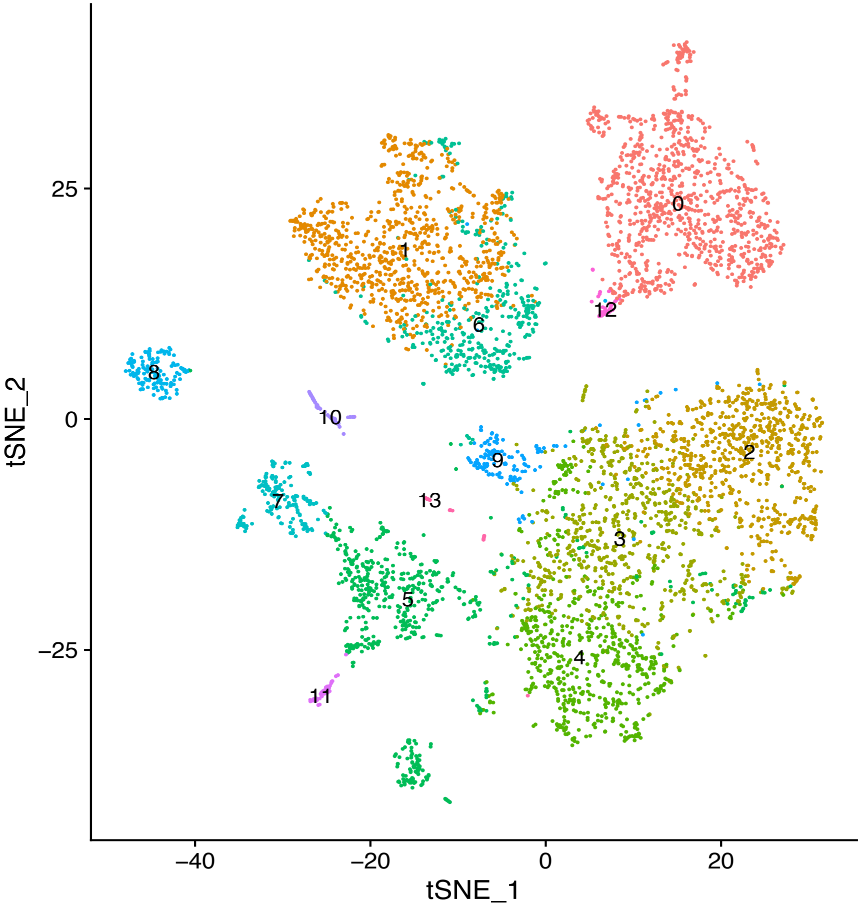
UMAP

TSNE vs. UMAP of a Same Single Cell RNAseq Data

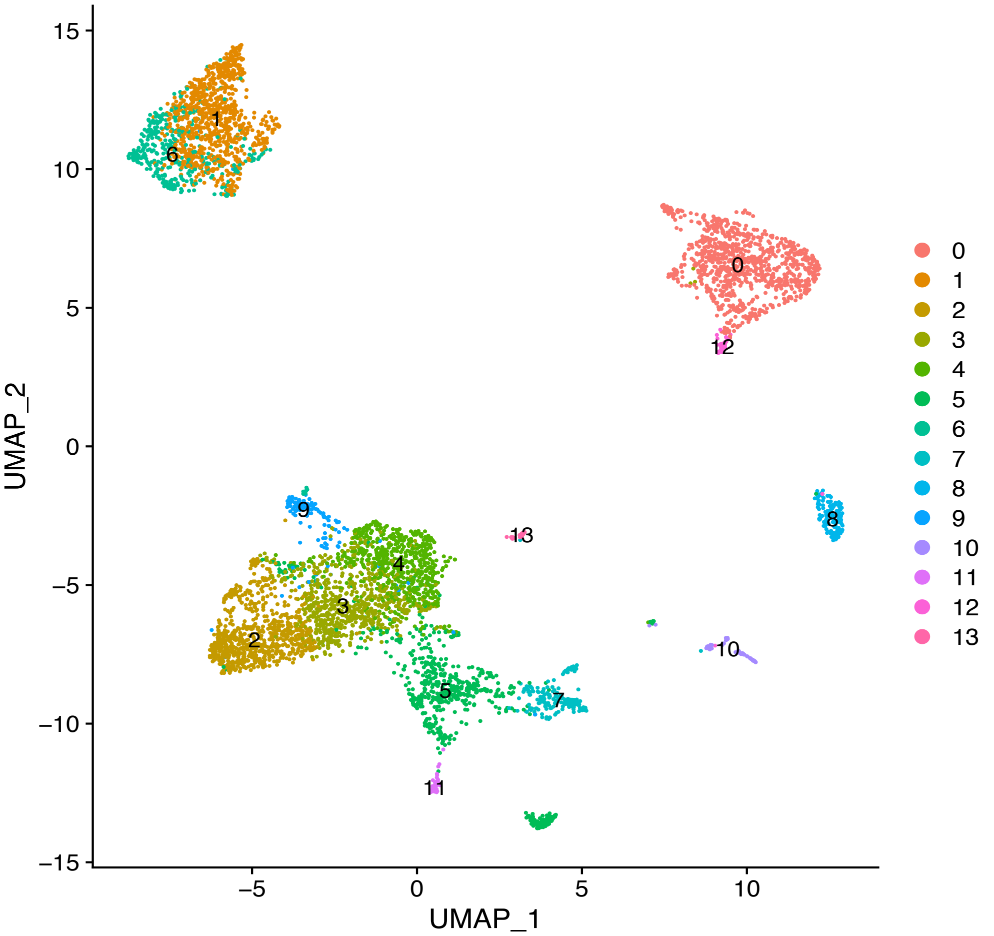
cell number n~6000

Clusters were identified before TSNE and UMAP analysis

TSNE

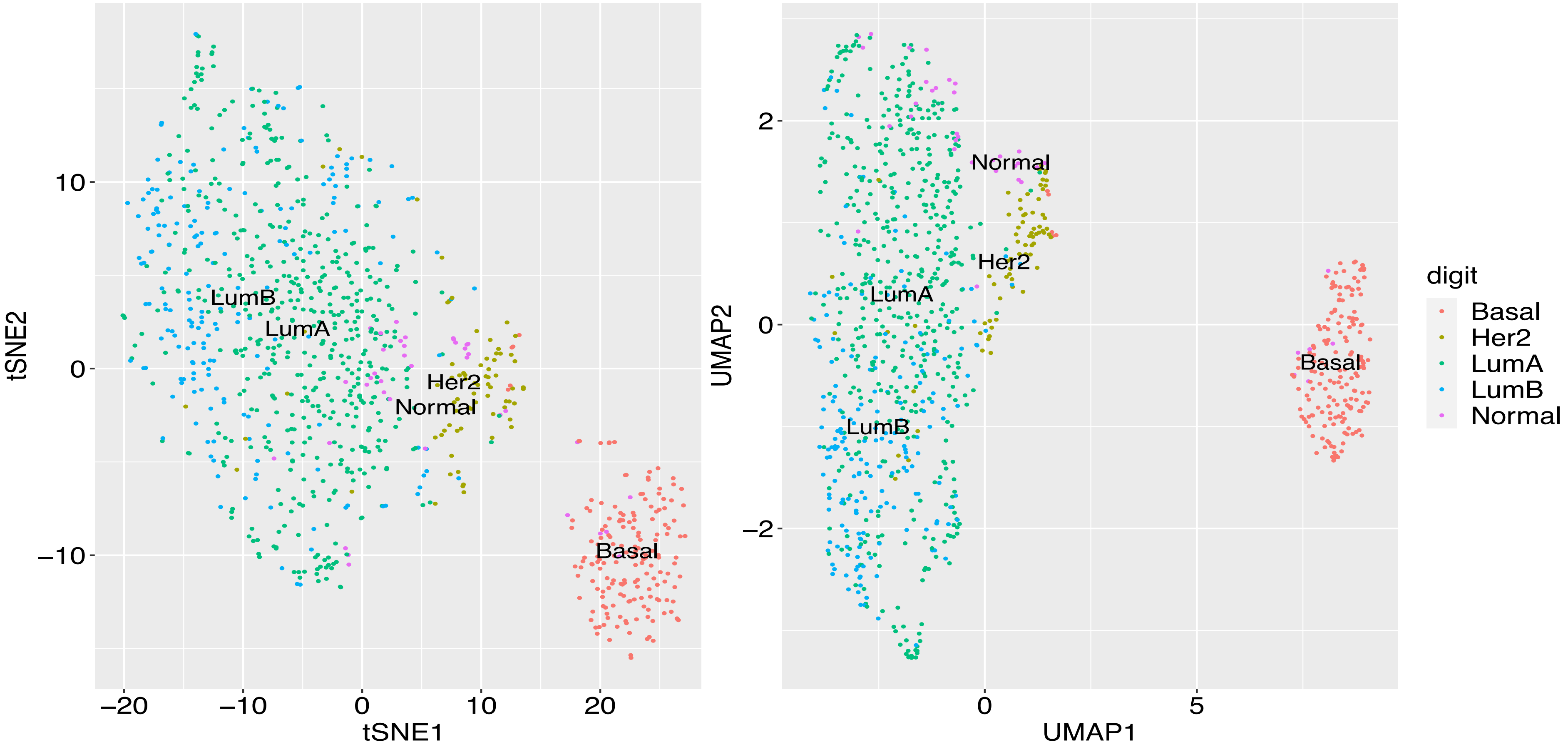


UMAP



TSNE vs. UMAP of TCGA Breast Cancer Data

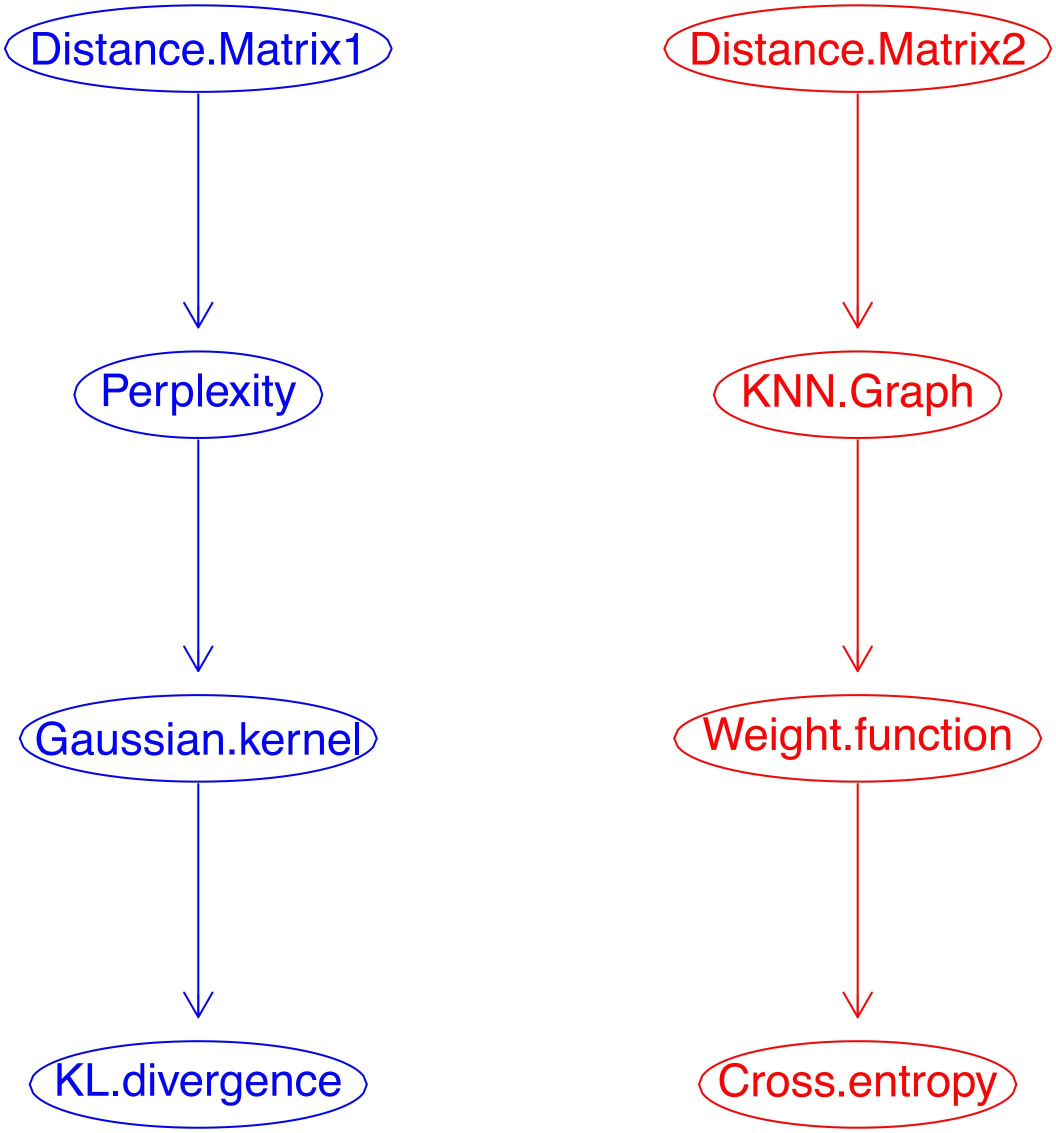
sample size n=977



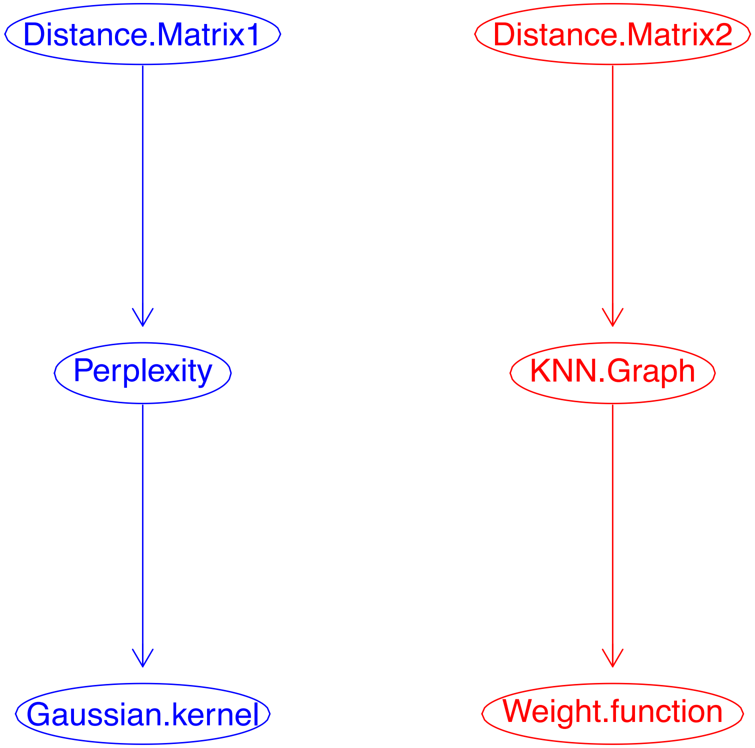
Comparison of PCA, TSNE, and UMAP

	Data type	Sample size	complexity	Performance
MNIST	image	6000	High	UMAP > TSNE > PCA
ScRNAseq	ScRNAseq	~6000	High?	UMAP ~ TSNE > PCA
TCGA	Bulk RNAseq	~1000	moderate	UMAP ~ TSNE ~ PCA

TSNE vs. UMAP


Distance.Matrix1
Perplexity
Gaussian.kernel
KL.divergence
Distance.Matrix2
KNN.Graph
Weight.function
Cross.entropy

TSNE vs. UMAP



Distance.Matrix1

Perplexity

Gaussian.kernel

Distance.Matrix2

KNN.Graph

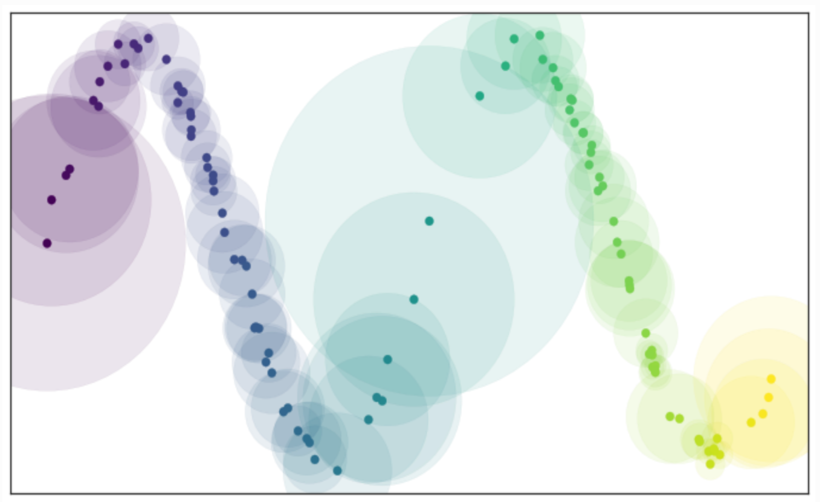
Weight.function

$$p(x;\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$$w((x_i,x_{i_j}))=\exp\left(\frac{-\max(0,d(x_i,x_{i_j})-\rho_i)}{\sigma_i}\right)$$

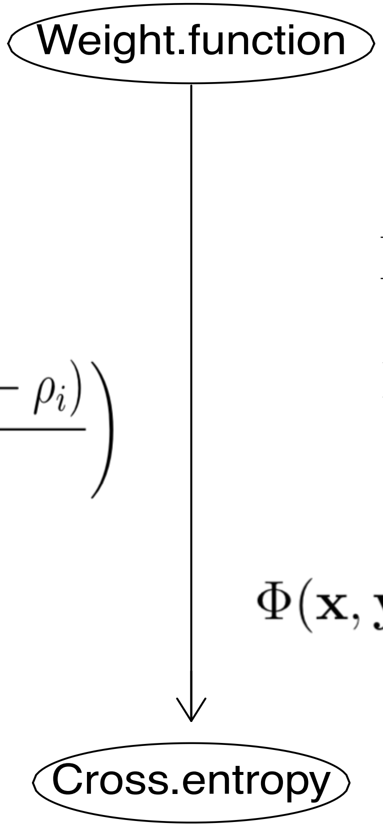
$$B=A+A^\top-A\circ A^\top$$

$$\sum_{j=1}^{k}\exp\left(\frac{-\max(0,d(x_i,x_{i_j})-\rho_i)}{\sigma_i}\right)=\log_2(k)$$



ρ_i: shortest distance of x_i neighbors

Uniform Manifold Approximation and Projection (UMAP)



Weight.function

Cross.entropy

High-dimension

$$w((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right)$$

$$B = A + A^\top - A \circ A^\top$$

Low-dimension

Laplacian Eigenmaps

[redacted]

$$\Phi(\mathbf{x}, \mathbf{y}) = \left(1 + a(\|\mathbf{x} - \mathbf{y}\|_2^2)^b\right)^{-1}$$

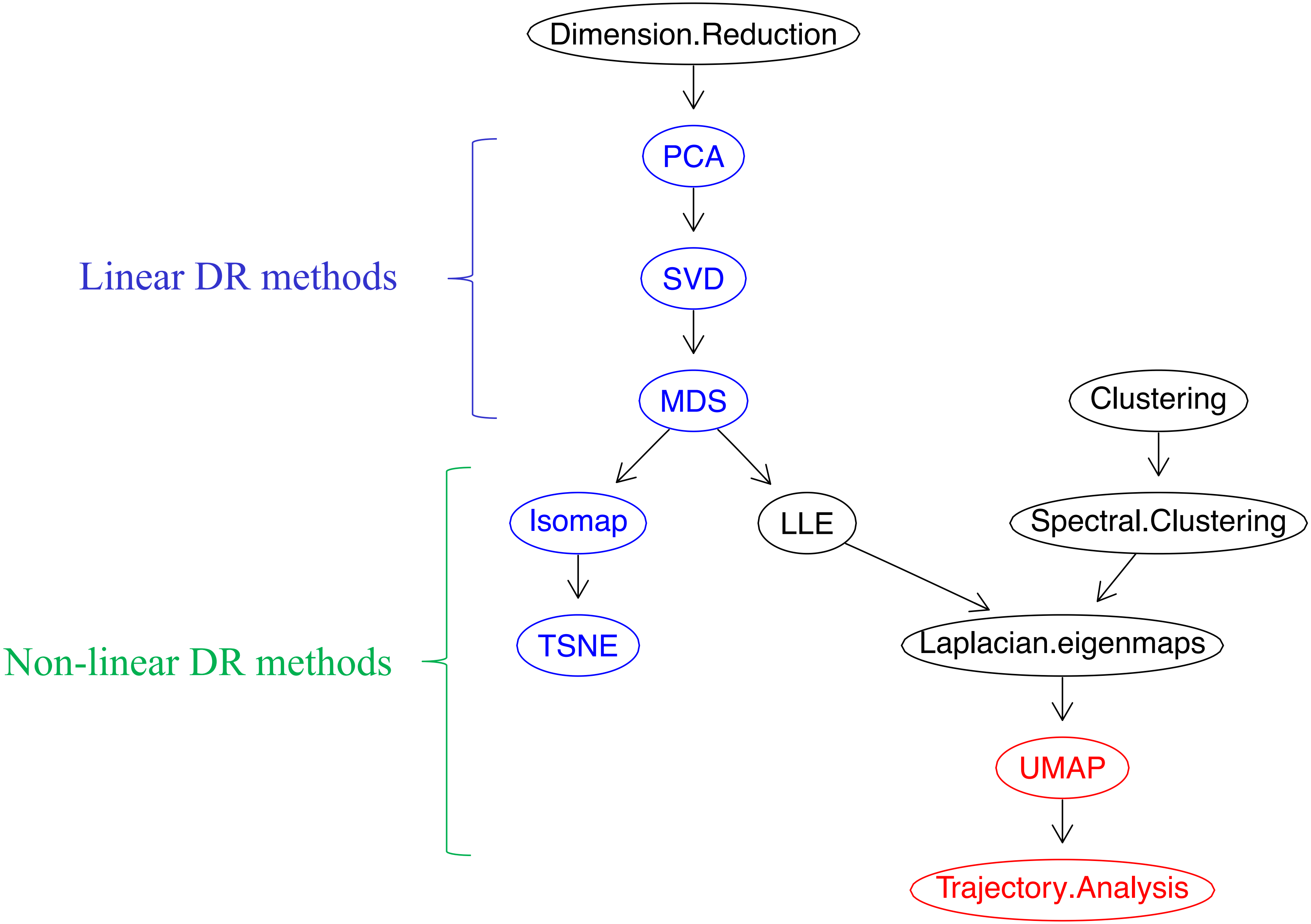
TSNE cost function

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

UMAP cost function

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

Road Map for Dimension Reduction Methods



Dimension.Reduction

Linear DR methods

PCA

SVD

MDS

Clustering

Non-linear DR methods

Isomap

LLE

Spectral.Clustering

TSNE

Laplacian.eigenmaps

UMAP

Trajectory.Analysis