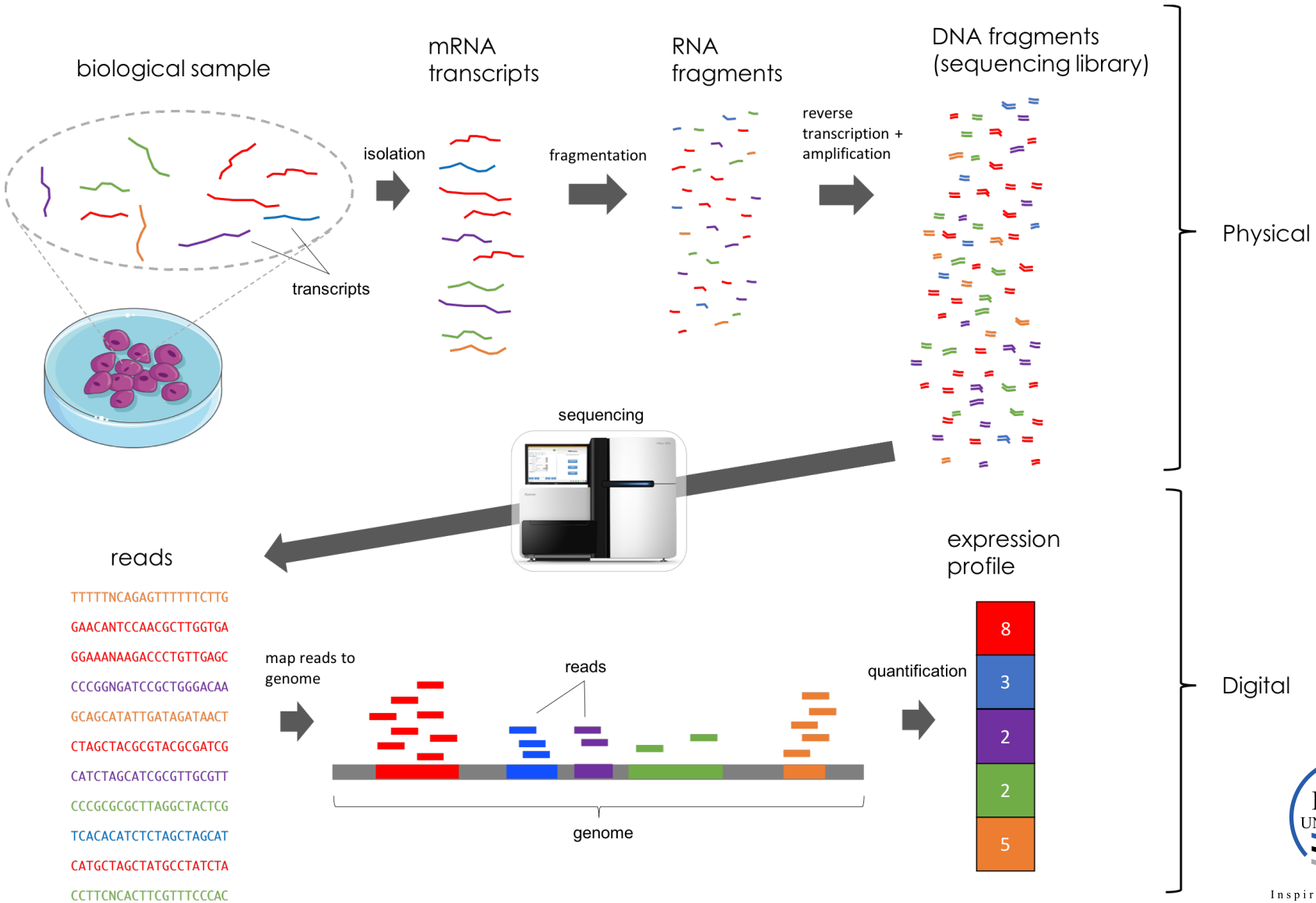
Bioinformatics: Gene Expression Analysis-II

Swakkhar Shatabda

Department of Computer Science and Engineering
BRAC University

Book Reference

What is RNA-Seq?



Sources of RNA-Seq Data

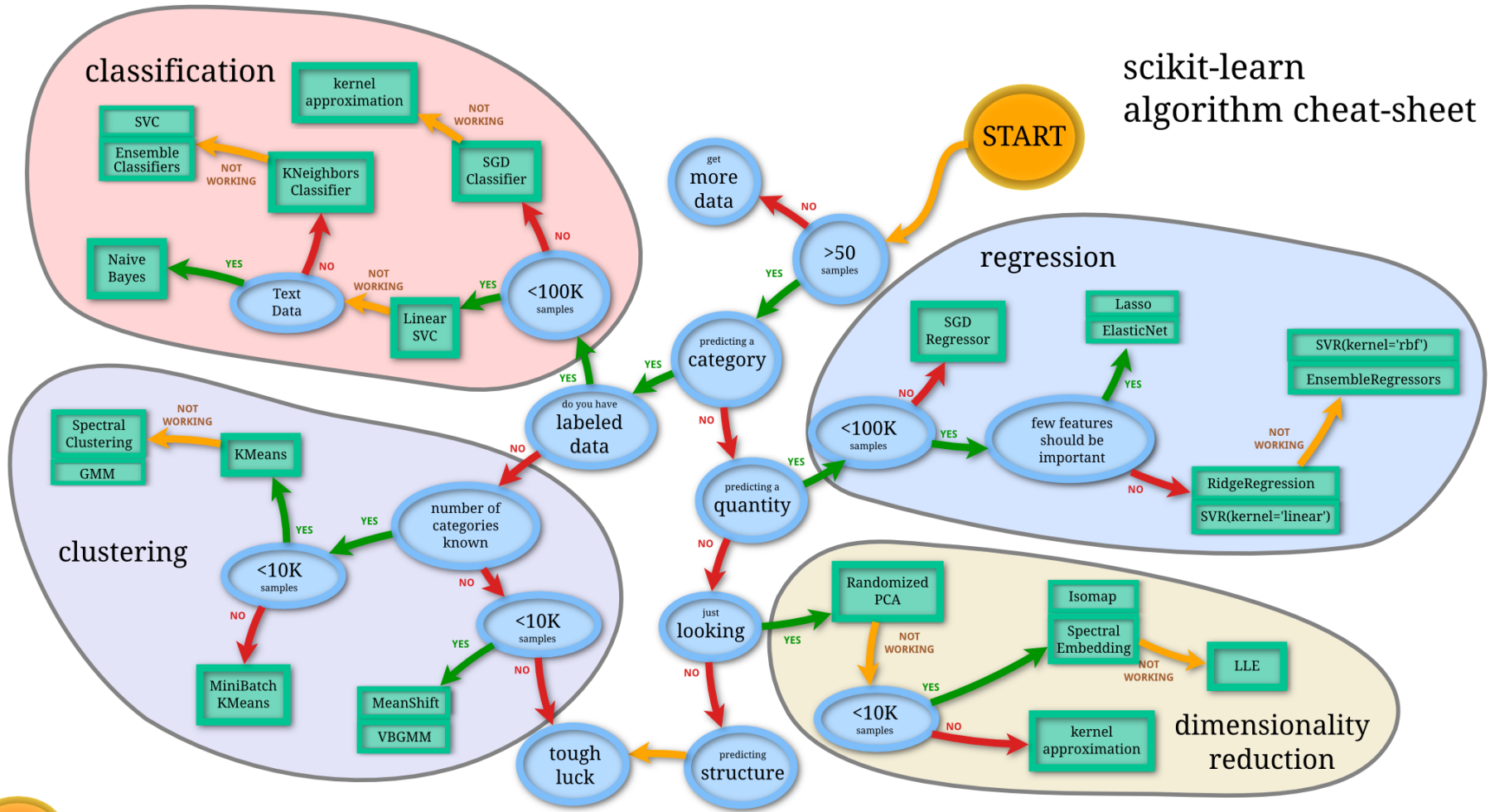
- Gene Expression Omnibus (GEO) (`http://www.ncbi.nlm.nih.gov/geo/`)
  - Both microarray and sequencing data
- Sequence Read Archive (SRA) (`http://www.ncbi.nlm.nih.gov/sra`)
  - All sequencing data (not necessarily RNA-Seq)
- ArrayExpress (`https://www.ebi.ac.uk/arrayexpress/`)
  - European version of GEO
- Homogenized data: MetaSRA, Toil, recount2, ARCHS4

Explore Parkinson's Disease

Microarray Experiment Data:
`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6613`

BRAC UNIVERSITY
Inspiring Excellence

ML Estimator Selection



Back

Dimensionality Reduction

- Seek and explore the inherent structure in data
- Unsupervised
- Data compression, summarization
- Pre-processing for visualization and supervised learning
- Can be adapted for classification and regression
- Well-known DR algorithms:
  - Principal Component Analysis (PCA)
  - Principal Component Regression (PCR)
  - Partial Least Squares Regression (PLSR)
  - Multidimensional Scaling (MDS)
  - Projection Pursuit
  - Linear Discriminant Analysis (LDA)
  - Mixture Discriminant Analysis (MDA)

BRAC UNIVERSITY
Inspiring Excellence

Linear vs Non-Linear

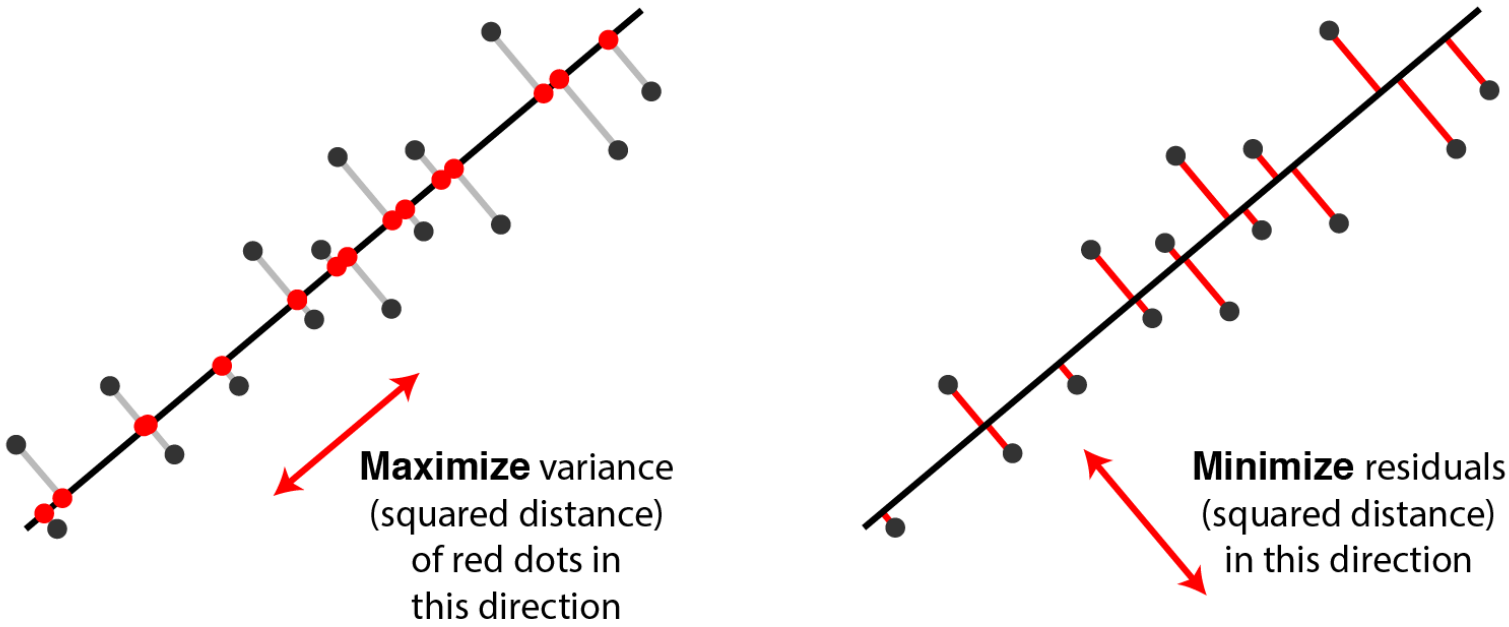
- Linear: Principal Component Analysis (PCA)
- Nonlinear DR, Manifold learning:
  - Isomap
  - Locally Linear Embedding (LLE)
  - Hessian Eigenmapping
  - Spectral Embedding
  - Multi-dimensional Scaling (MDS)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)

BRAC UNIVERSITY
Inspiring Excellence

PCA

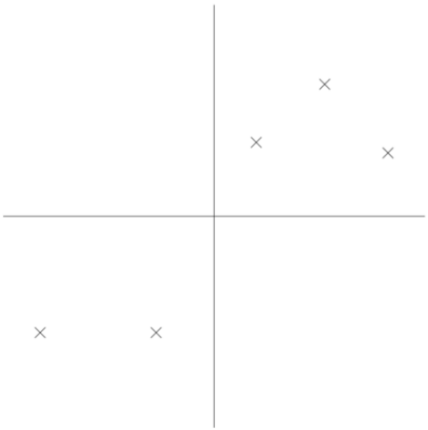
PCA can be interpreted in 2 different ways:

- Maximize the variance of projection along each component (dimension)
- Minimize the reconstruction error, that is, the squared distance between the original data and its projected coordinates

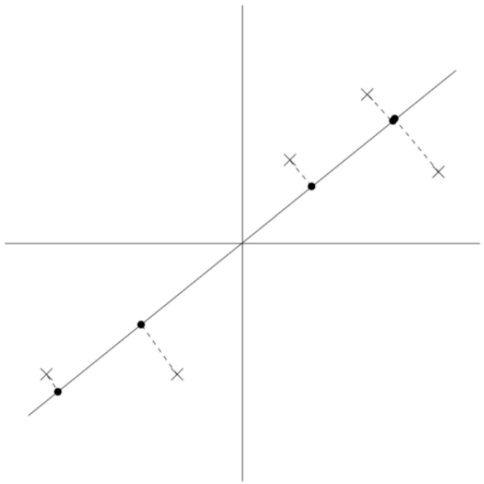


PCA at a glance

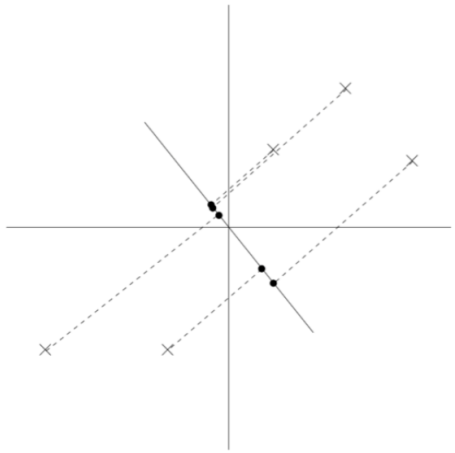
- PCA selects project direction that maximizes the variance
- The direction of maximum variance in the input space happens to be the same as the principal eigenvector of the covariance matrix of the data
- **PCA algorithm:** finding the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the dataset; this is the principle component.



Data after normalization



A projection with large variance



A projection with small variance

Eigen Values and Eigen Vectors

- For a given co-variance matrix, A, Eigen vectors, $\vec{v}$ are those vectors for which the product $A\vec{v}$ is a scalar multiple of $\vec{v}$. That is $\vec{v}$ satisfies the following Equation:

$$A\vec{v} = \lambda\vec{v}$$

- λ is a scalar, called eigen value.

BRAC UNIVERSITY
Inspiring Excellence

PCA Theorem

- Let $x_1, x_2, \cdots, x_m$ be a set of m $N \times 1$ vectors and let $\overline{x}$ be their mean:

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix} \qquad \overline{x} = \frac{1}{m} \sum_{i=1}^{i=m} \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix}$$

- Let X be a matrix with columns $x_1 - \overline{x}\ x_2 - \overline{x} \cdots\ x_m - \overline{x}$

$$X = \begin{bmatrix} x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_m - \overline{x} \end{bmatrix}$$

- Subtracting the mean is equivalent to translating the coordinate system to the location of the mean.

BRAC UNIVERSITY
Inspiring Excellence

PCA Theorem

- Let $Q = XX^T$ be the $n \times n$ matrix

$$Q = XX^T = \begin{bmatrix} x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_m - \overline{x} \end{bmatrix} \begin{bmatrix} (x_1 - \overline{x})^T \\ (x_2 - \overline{x})^T \\ \vdots \\ (x_m - \overline{x})^T \end{bmatrix}$$

- Note:
  - Q is square
  - Q is symmetric
  - Q is the covariance covariance matrix
  - Q can be very large (in vision, N is often the number of pixels in an image!)

BRAC UNIVERSITY
Inspiring Excellence

PCA Theorem

Theorem

Each x_j can be written as:

$$x_j = \overline{x} + \sum_{i=1}^{i=n} g_{ji} e_j$$

where e_i are the n eigenvectors of Q with non-zero eigenvalues.

- Expressing x in terms of $e_1 \cdots e_n$ has not changed the size of the data
- However, if the points are highly correlated many of the coordinates of x will be zero or closed to zero.

Demo: https://colab.research.google.com/drive/1td7cL4Y499eUOCbqN0ov9wKChN6D9JLy?usp=sharing

BRAC UNIVERSITY
Inspiring Excellence

Preserving distances

- Many DR methods focus on preserving distances, e.g. the above is the cost function for a particular DR method called metric MDS

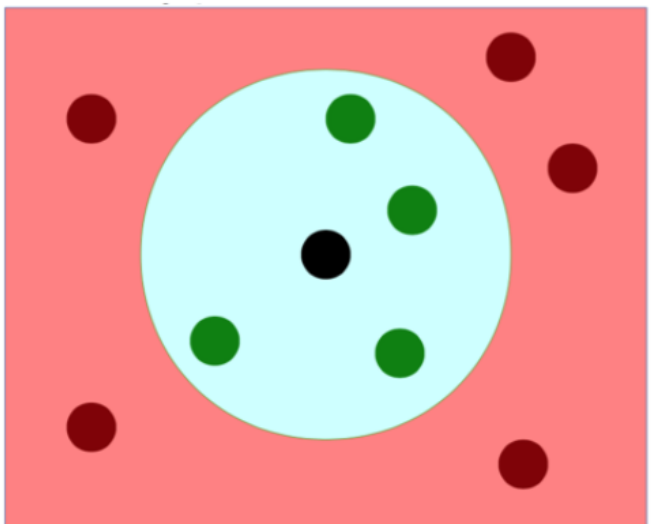
$$C = \frac{1}{a} \sum_{ij} w_{ij} (d_X(x_i, x_j) - d_Y(y_i, y_j))^2$$

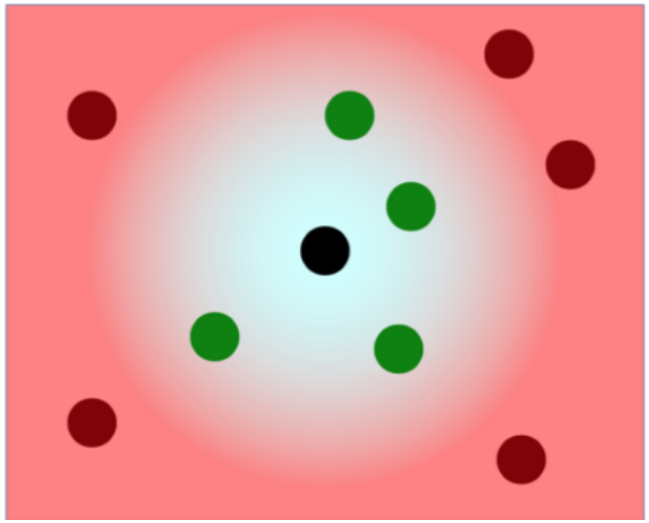
- An alternative idea is preserving neighborhoods.

BRAC UNIVERSITY
Inspiring Excellence

Preserving neighborhoods

- Neighbors are an important notion in data analysis, e.g.social networks, friends, twitter followers...
- Object nearby (in a metric space) are considered neighbors
- Consider hard neighborhood and soft neighborhood
- Hard: each point is a neighbor (green) or a non-neighbor (red)
- Soft: each point is a neighbor (green) or a non-neighbor (red) with some weight





BRAC UNIVERSITY
Inspiring Excellence

Probabilistic neighborhood

- Derive a probability of point j to be picked as a neighbor of i in the input space

$$p_{ij} = \frac{exp(-d_{ij}^2)}{\sum_{i \neq k} exp(-d_{ik}^2)}$$

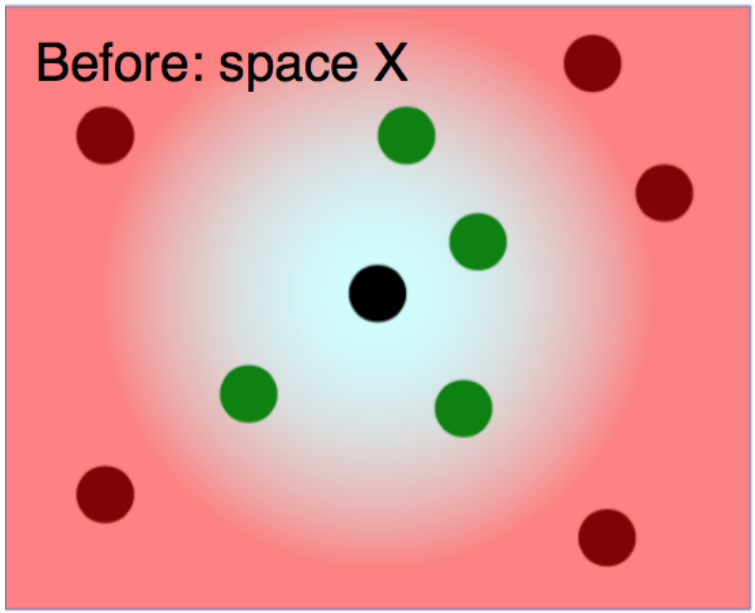
Probabilistic neighborhood

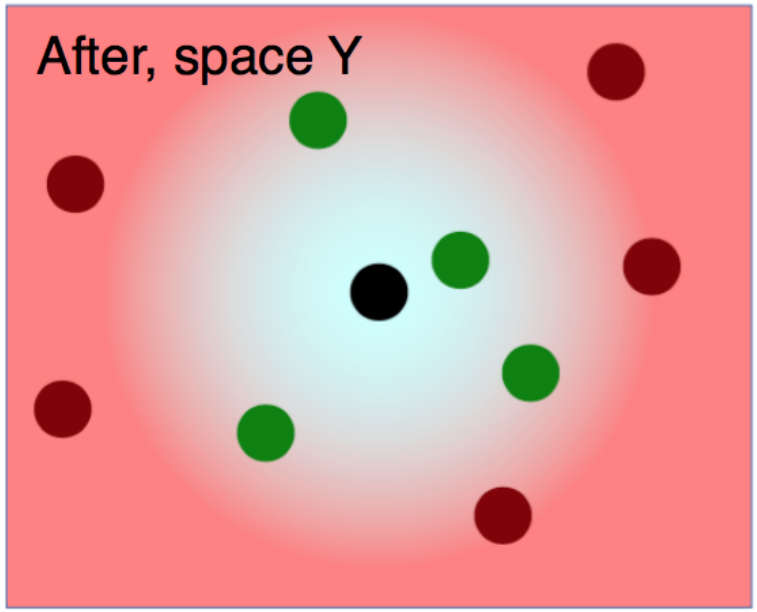
- Probabilistic Input Neighborhood

$$p_{ij} = \frac{exp(-||x_i - x_j||^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2)}$$

- Probabilistic output Neighborhood

$$q_{ij} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(-||y_i - y_k||^2)}$$





BRAC UNIVERSITY
nspiring Excellence

Stochastic Neighborhood Embedding

- Compare neighborhoods between the input and output!
- Using Kullback-Leibler (KL) divergence
- KL divergence: relative entropy (amount of surprise when encounter items from 1st distribution when they are expected to come from the 2nd)
- KL divergence is nonnegative and 0 iff the distributions are equal
- SNE: minimizes the KL divergence using gradient descent

$$C = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

BRAC UNIVERSITY
Inspiring Excellence

SNE: gradient descent

- Adjusting the output coordinates using gradient descent
- Gradient descent: iterative process to find the minimal of a function
- Start from a random initial output configuration, then iteratively take steps along the gradient
- Intuition: using forces to pull and push pairs of points to make input and output probabilities more similar x

BRAC UNIVERSITY
Inspiring Excellence

The crowding problem

- When embedding neighbors from a high-dim space into a low- dim space, there is too little space near a point for all of its close-by neighbors.
- Some points end up too far-away from each other
- Some points that are neighbors of many far-away points end up crowded near the center of the display.
- In other words, these points end up crowded in the center to stay close to all of the far-away points.
- t-SNE: using heavy-tailed distributions (i.e., t-distributions) to define neighbors on the display, to resolve the crowding problem

BRAC UNIVERSITY
Inspiring Excellence

t-distributed SNE

- Avoids crowding problem by using a more heavy-tailed neighborhood distribution in the low-dim output space than in the input space.
- Neighborhood probability falls off less rapidly; less need to push some points far off and crowd remaining points close together in the center.