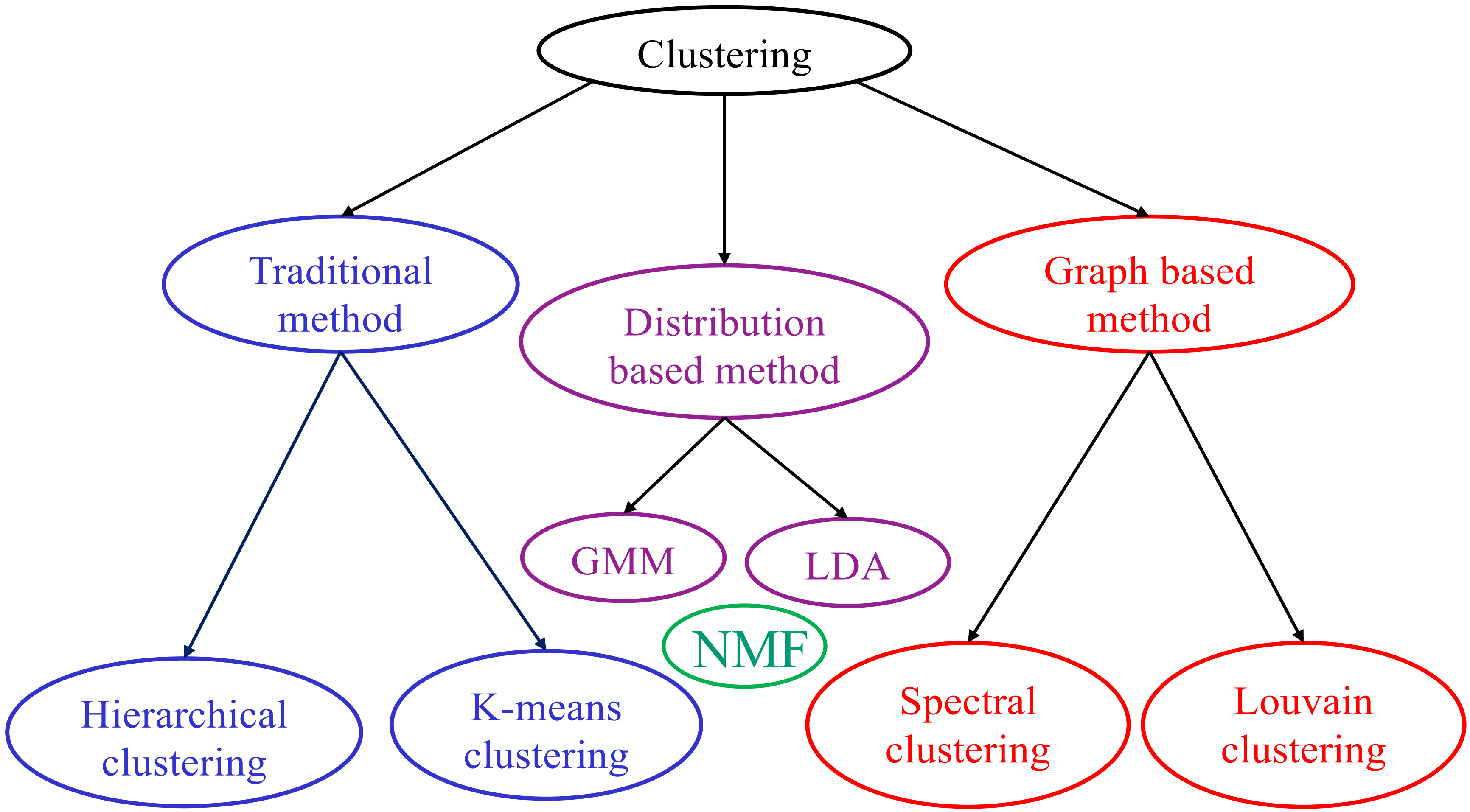
Clustering Methods:
From k-means to Gaussian Mixture Model and Louvain Algorithm

Maxwell Lee

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

November 16, 2020

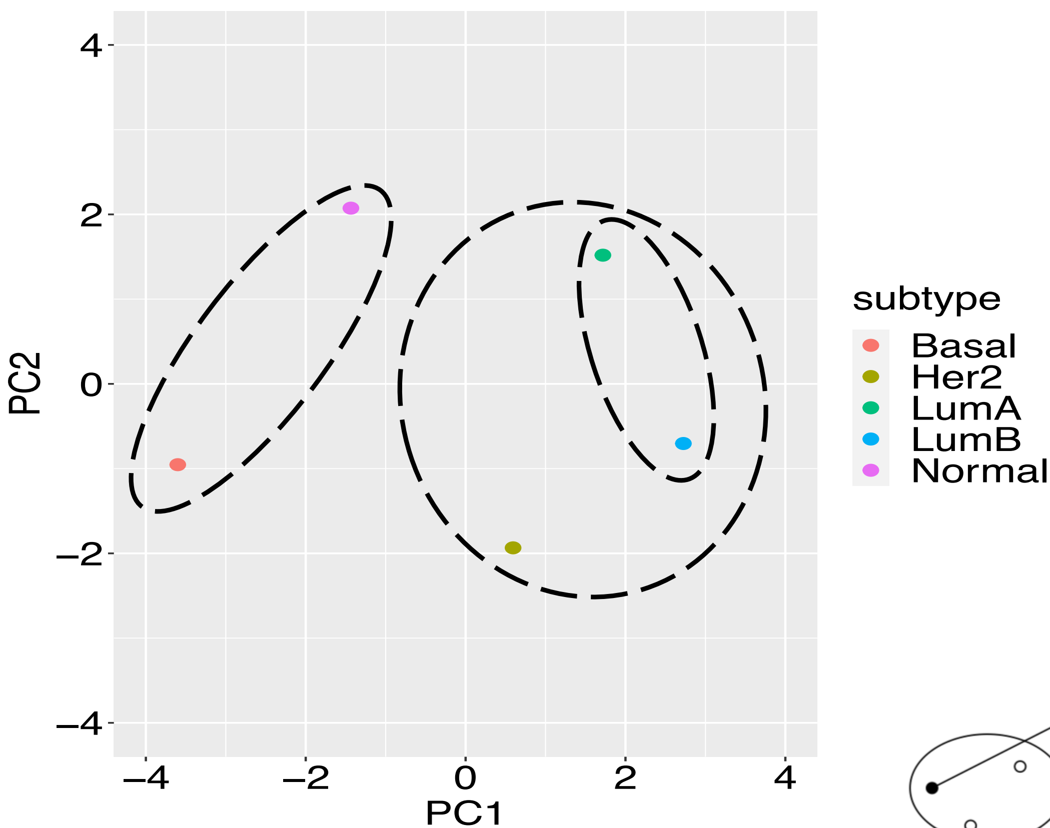
Outline of Clustering Methods

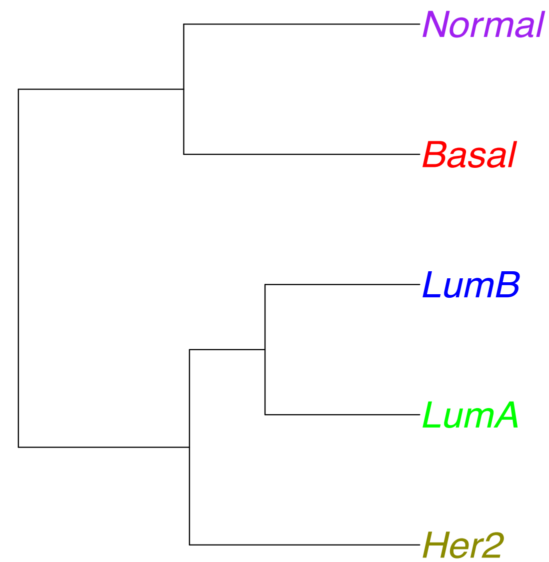


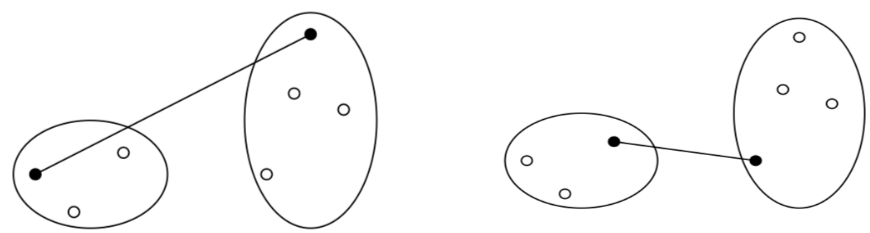
GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai

Hierarchical Agglomerative Clustering Algorithm





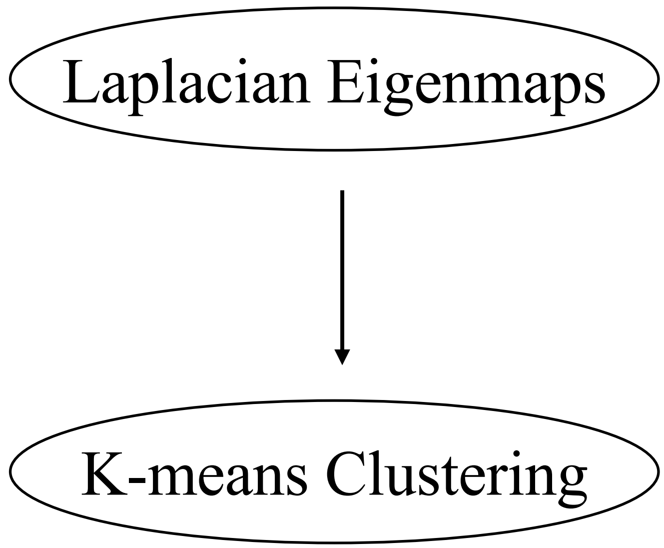


complete-linkage single-linkage

- Start with n leaf nodes
- Sequentially merge a pair of nodes with the smallest distance or minimal variance
- End with a single cluster

Spectral Clustering

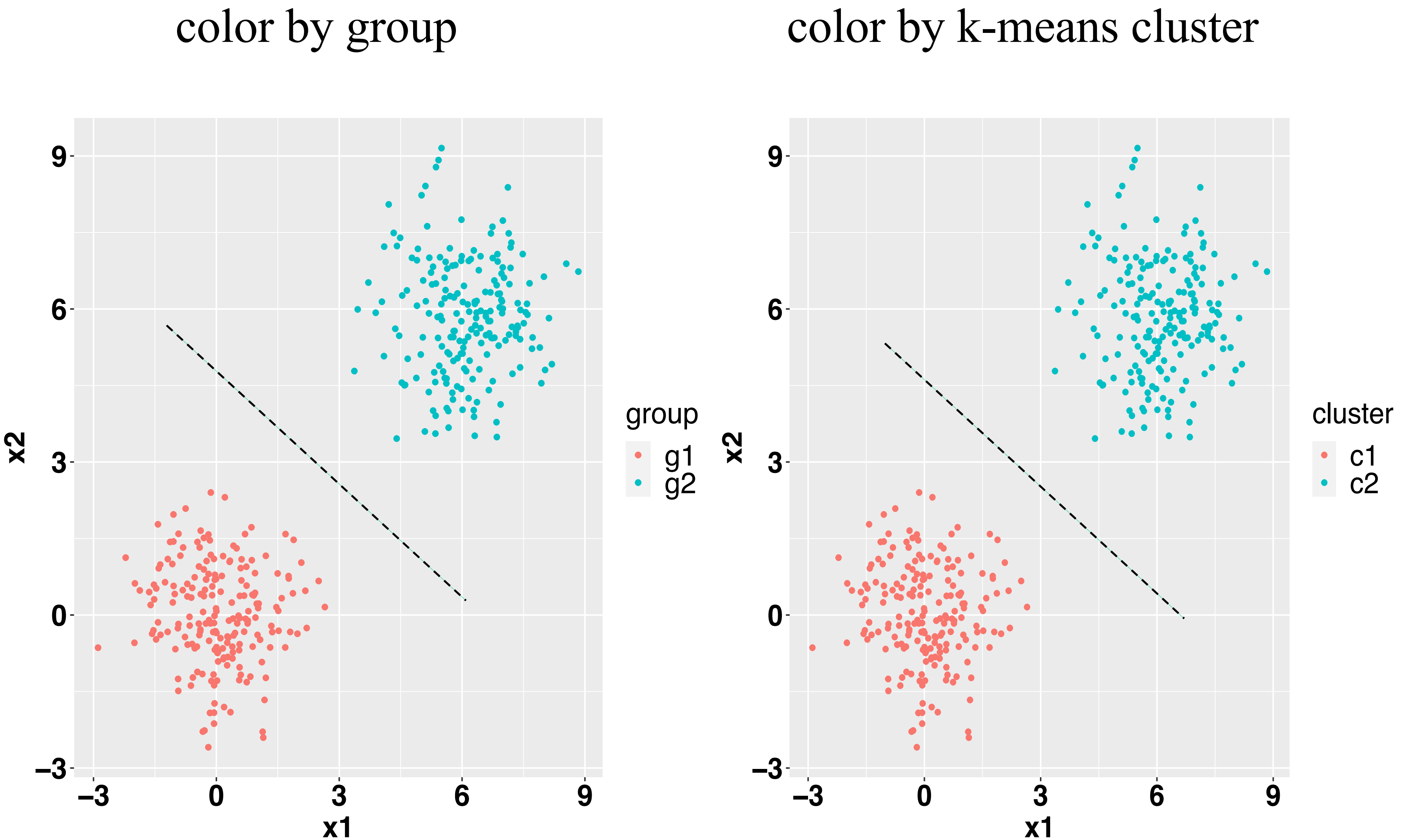
Spectrum: set of its eigenvalues



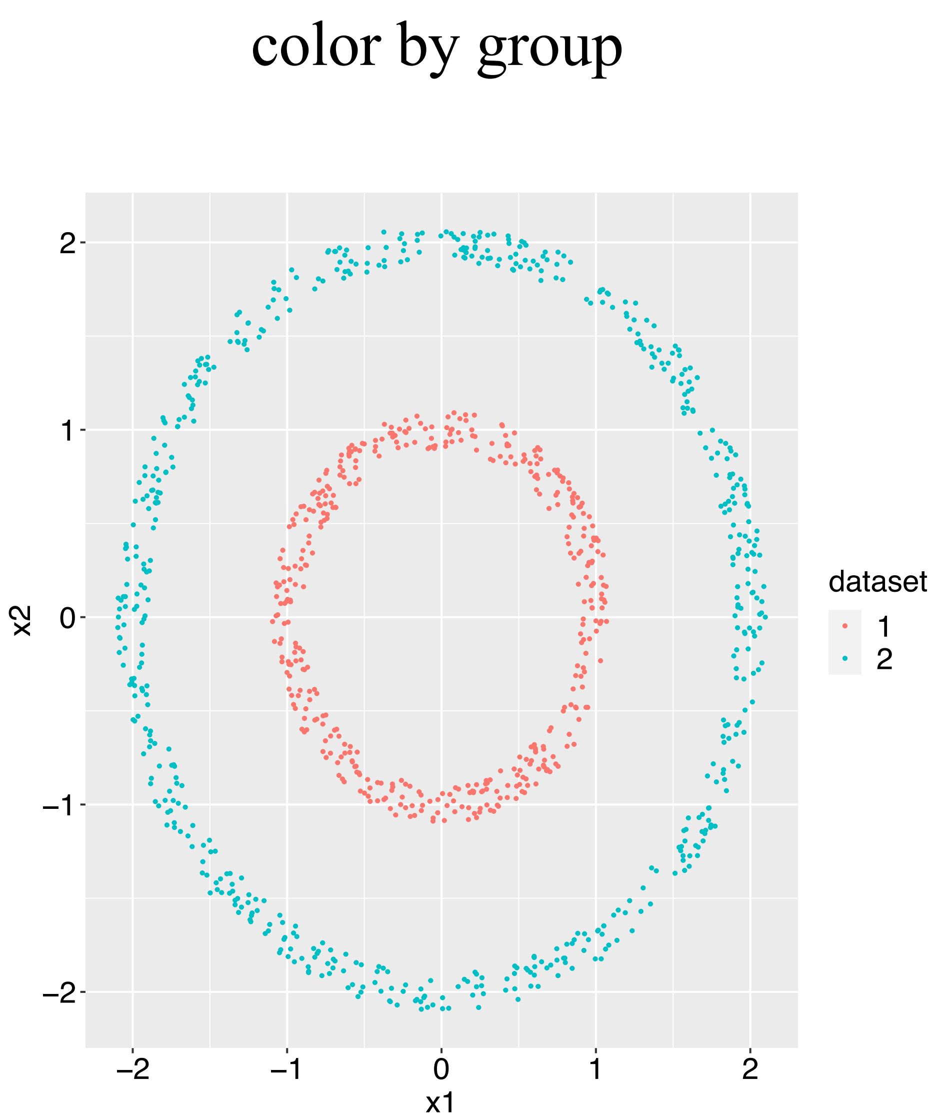
Euclidean distance is not appropriate in HD or in non-Euclidean space

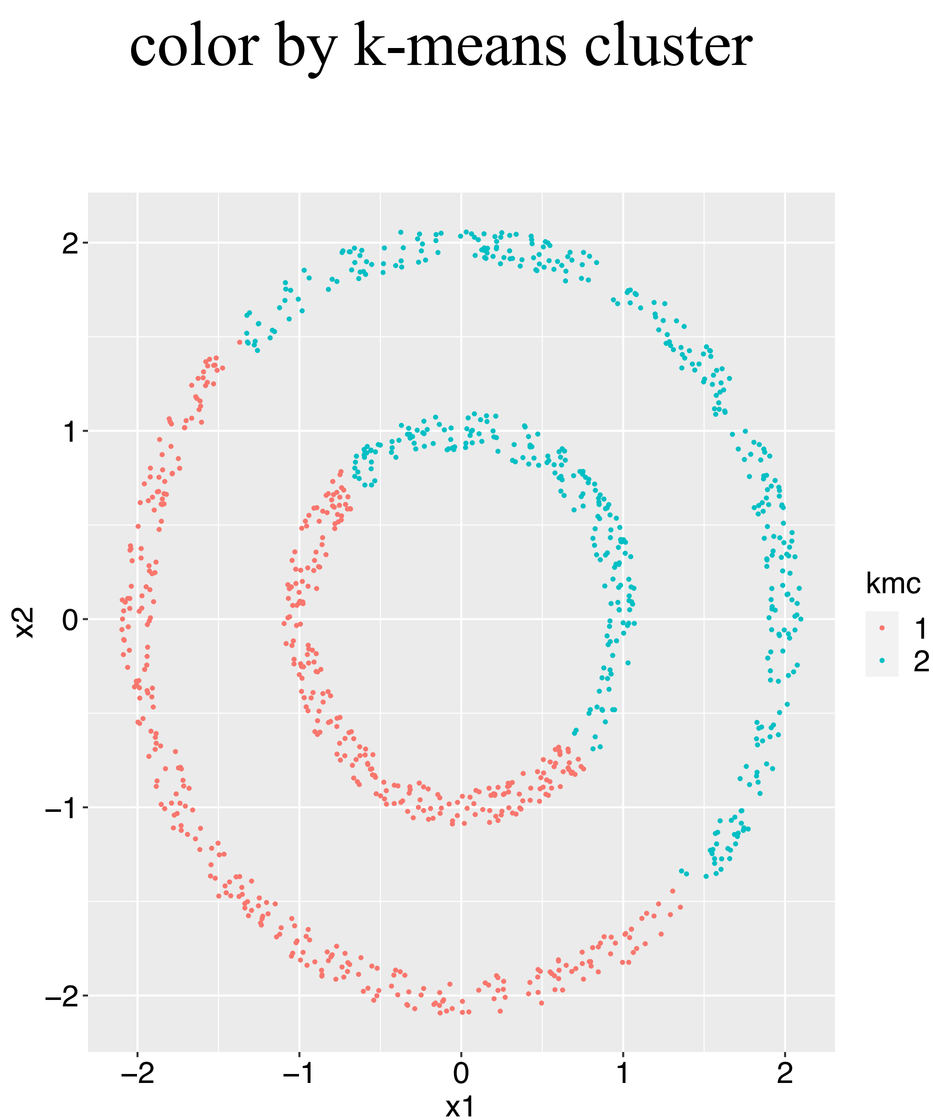
- Curse of dimensionality
- Laplacian Eigenmaps is a non-linear dimension reduction method

K-means Clustering

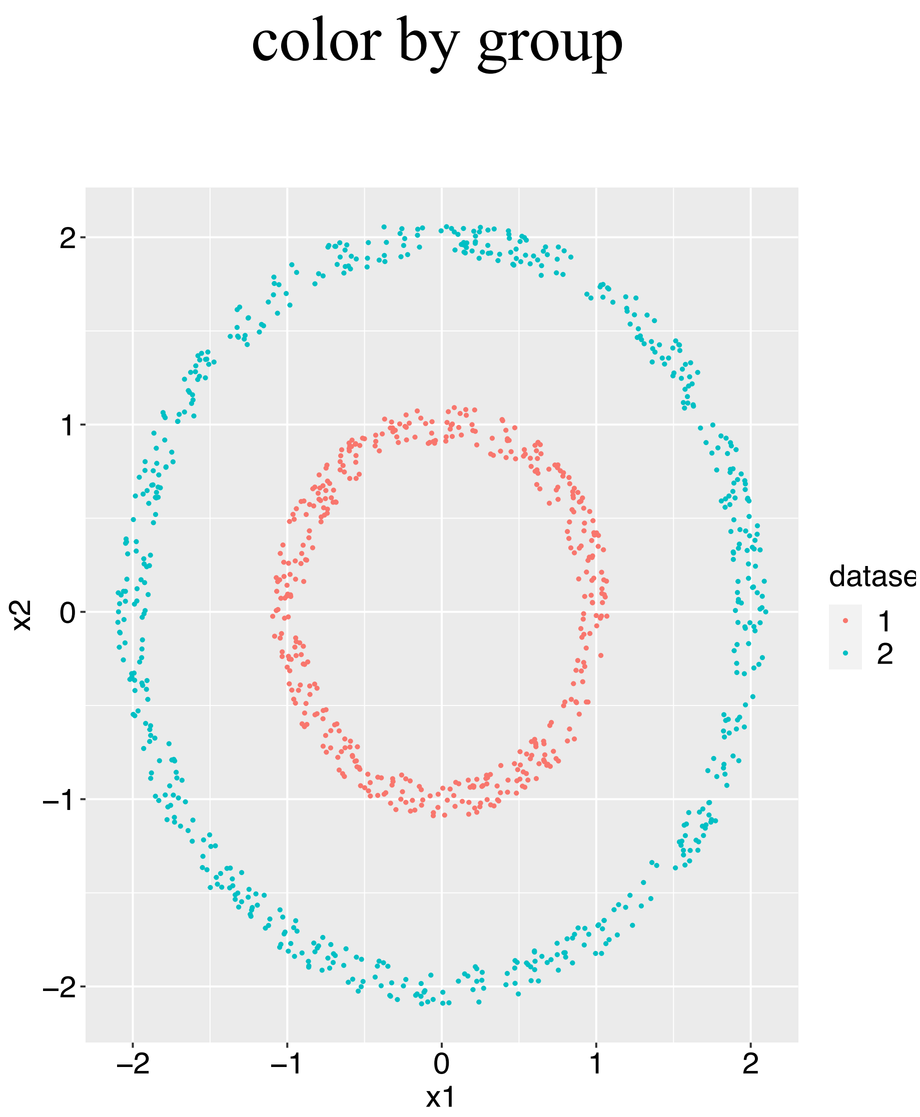


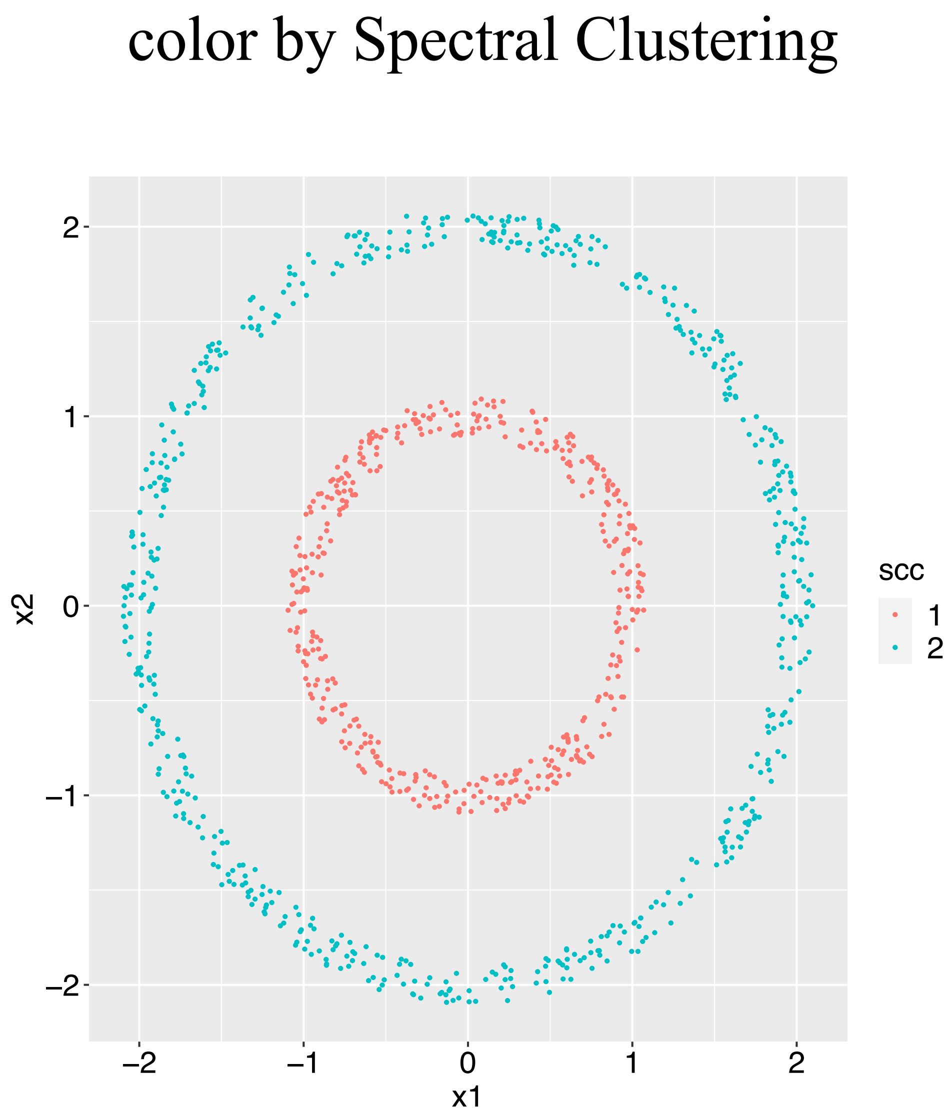
Two Circles and K-means Clustering



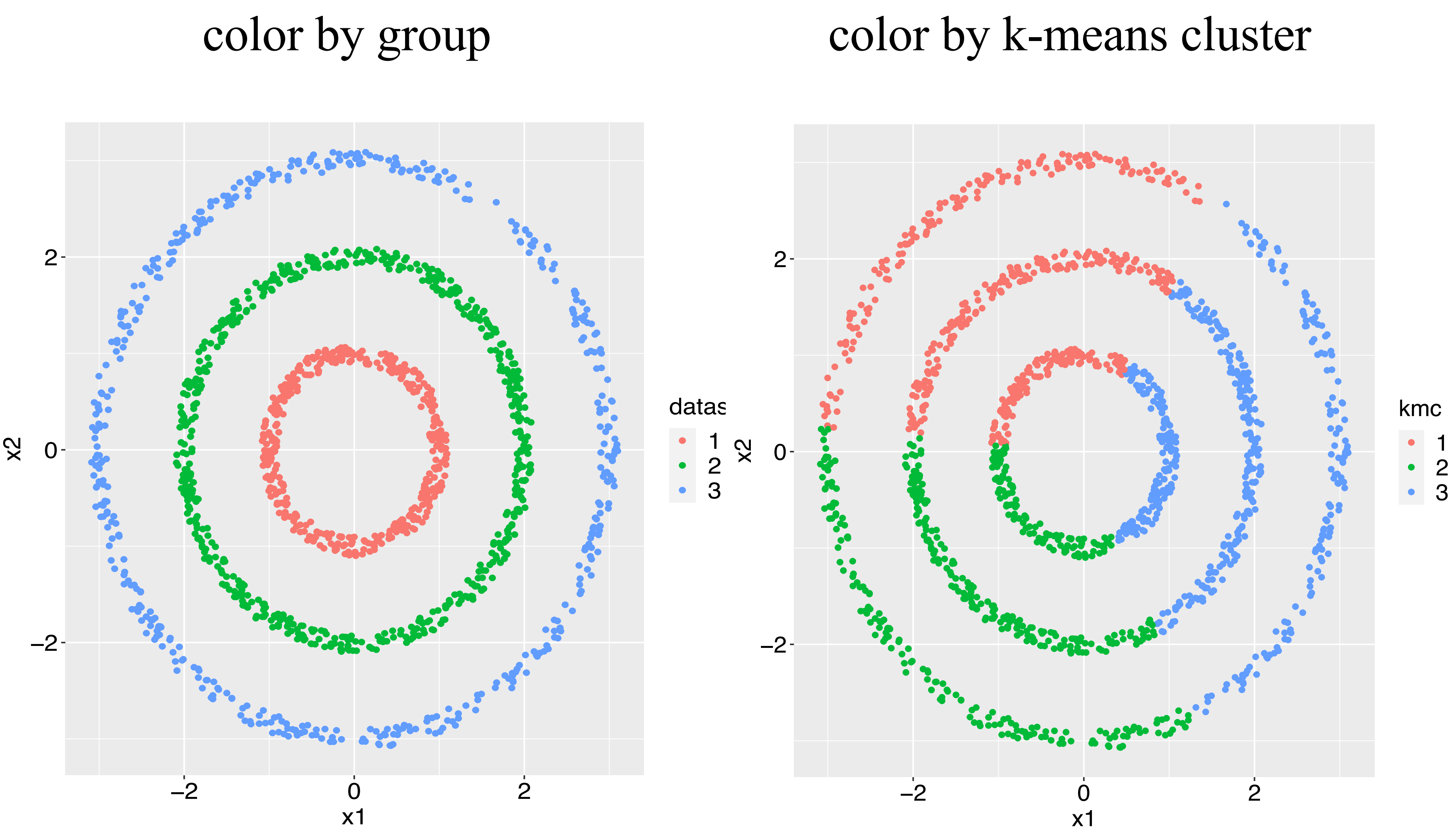


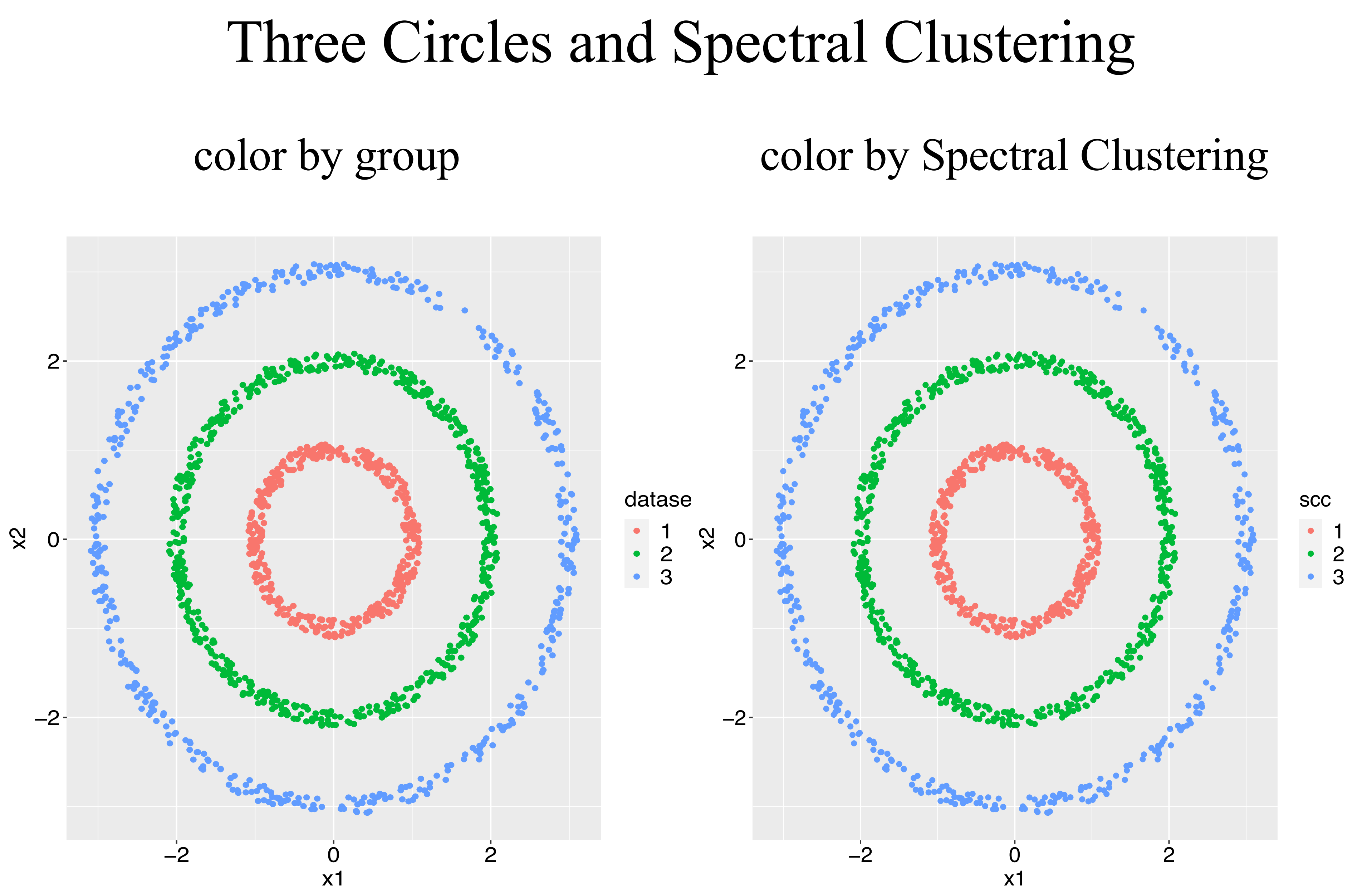
Two Circles and Spectral Clustering





Three Circles and K-means Clustering



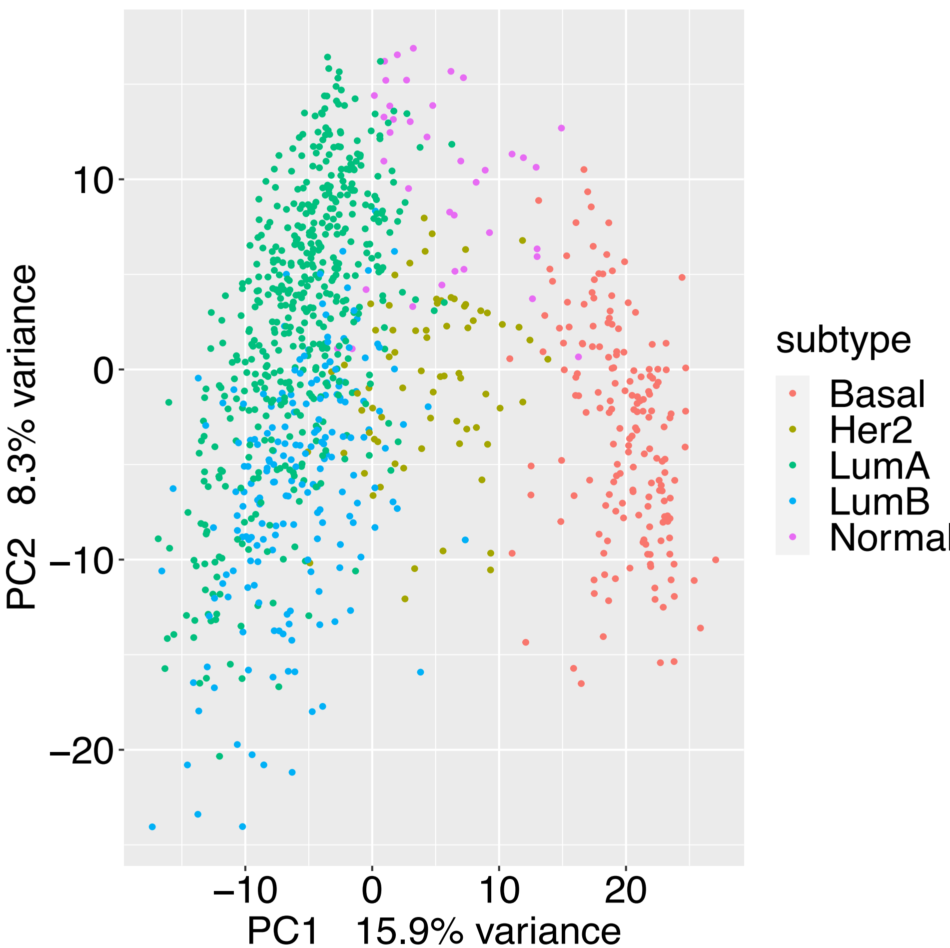

Three Circles and Spectral Clustering
color by group
color by Spectral Clustering
x1
x2
−2
0
2
datase
1
2
3
scc
1
2
3

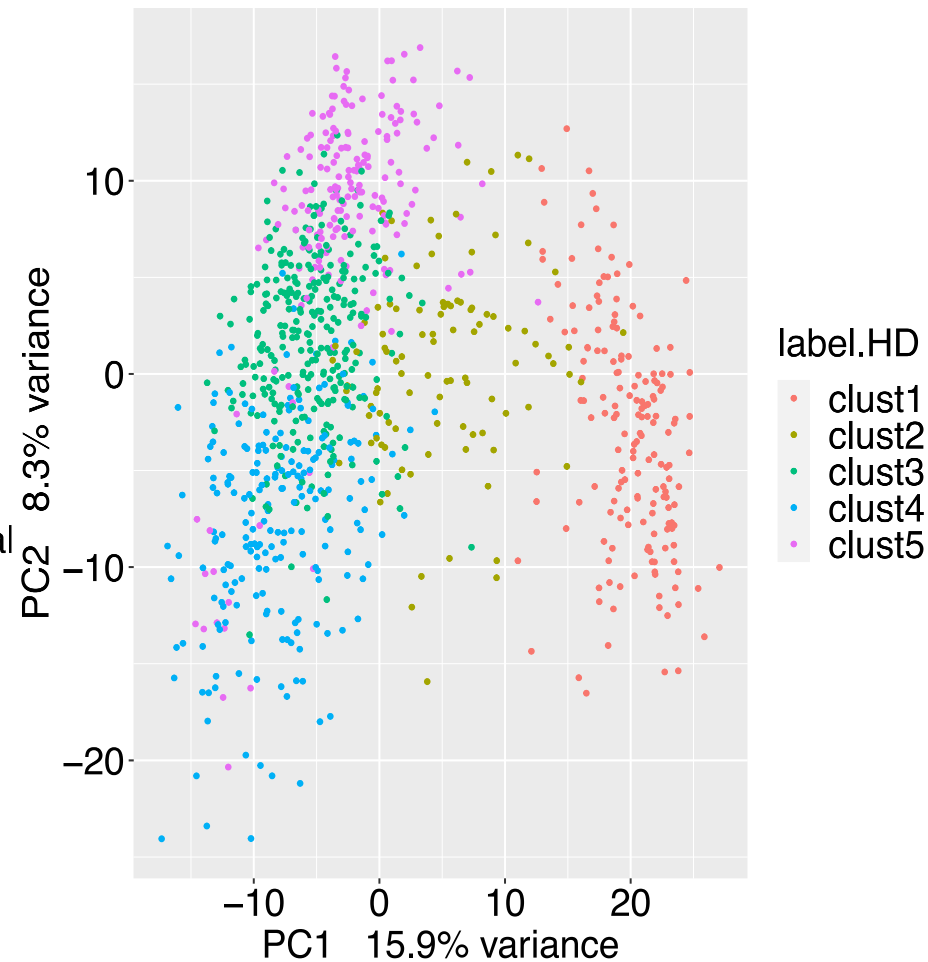
PCA Label by Subtype vs. Spectral Clustering

Label by subtype

Spectral Clustering in HD

977 samples accuracy 64.5%





HD: high dimension, 5000 genes

Spectral Clustering vs. k-means Cluster (HD)

5000 genes

Spectral Clustering

	Basal	Her2	LumA	LumB	Normal
clust1	165	0	0	0	5
clust2	8	64	8	13	7
clust3	0	5	245	47	2
clust4	0	4	78	**132**	0
clust5	0	0	**169**	1	**24**

Accuracy = (165 + 64 + 245 + 132 + 24) / 977 = 64.5%

Match
Mismatch

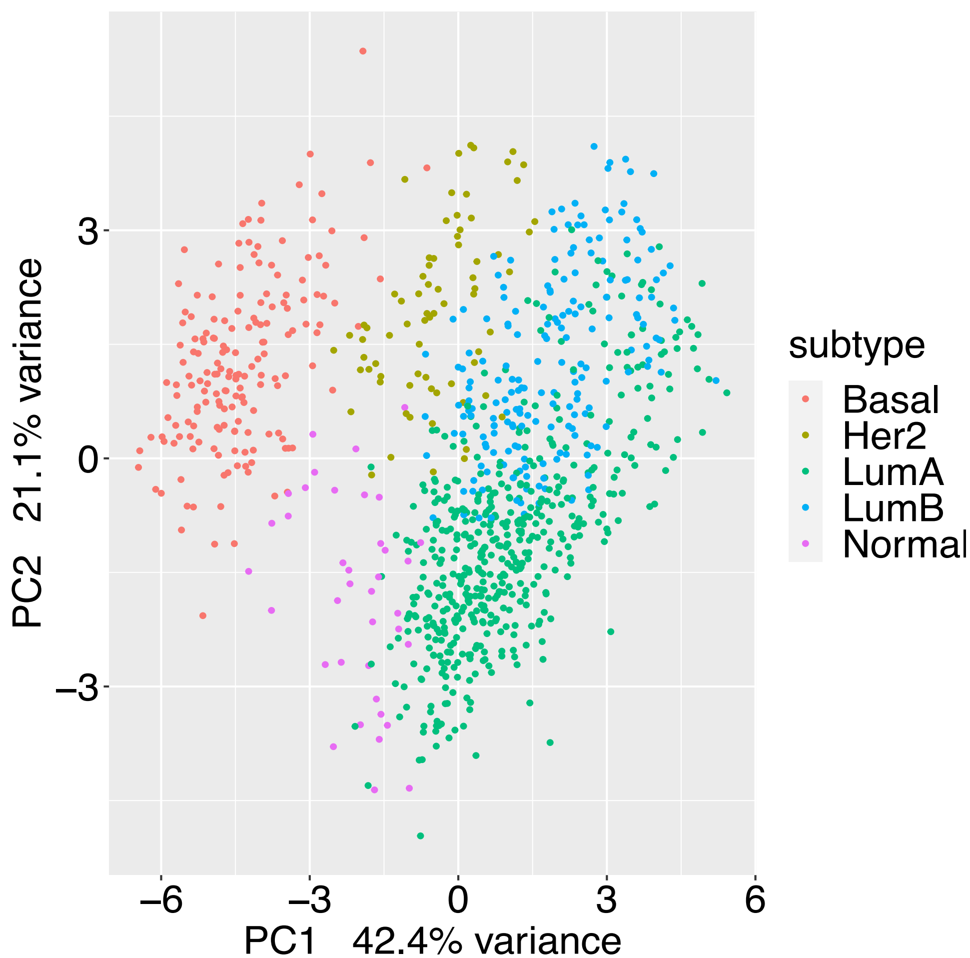
K-means

	Basal	Her2	LumA	LumB	Normal
clust1	169	0	0	0	6
clust2	4	69	17	40	5
clust3	0	0	**268**	11	**21**
clust4	0	0	**125**	119	0
clust5	0	4	90	23	6

Accuracy = (169 + 69 + 268 + 119 + 6) / 977 = 64.6%

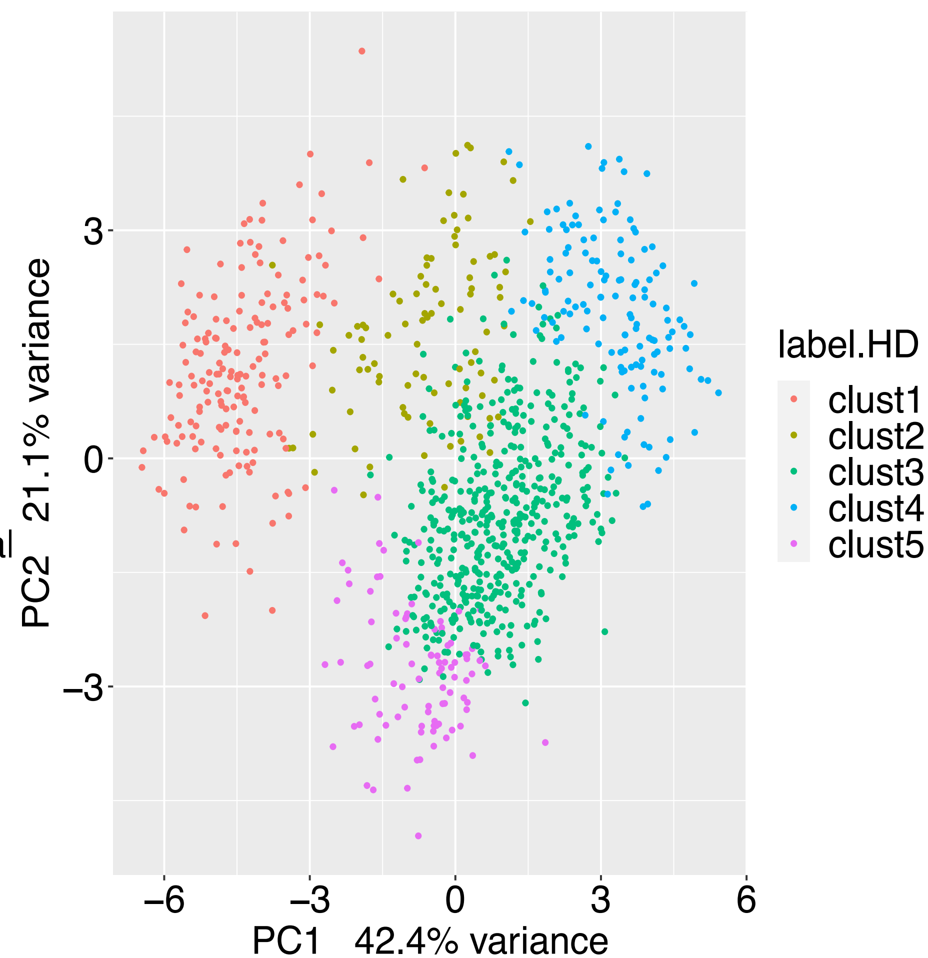
PCA with pam50: Label by Subtype vs. Spectral Clustering

Label by subtype



Label by Spectral Clustering

accuracy 70%



HD: high dimension, 39 genes

Spectral Clustering vs. k-means Clustering (pam50)

Spectral Clustering

	Basal	Her2	LumA	LumB	Normal
clust1	167	0	0	0	6
clust2	6	64	6	12	5
clust3	0	6	**361**	**114**	2
clust4	0	3	67	67	0
clust5	0	0	66	0	25

Accuracy = (167 + 64 + 361 + 67 + 25) / 977 = 70%

Match
Mismatch

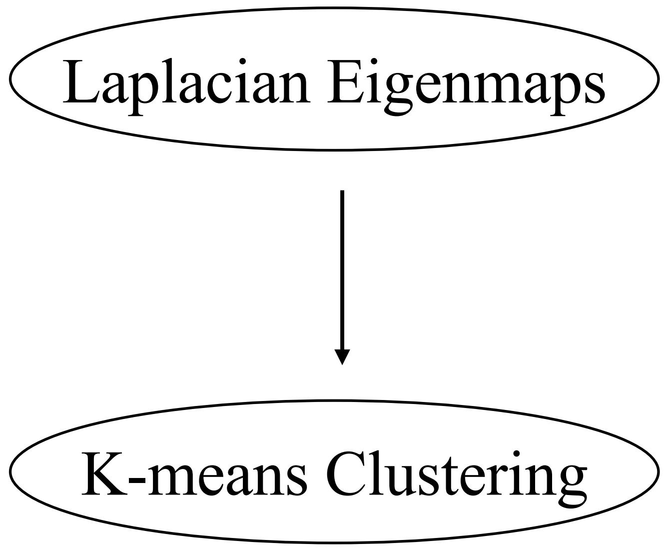
K-means

	Basal	Her2	LumA	LumB	Normal
clust1	170	0	0	0	8
clust2	2	72	11	31	4
clust3	0	0	221	75	0
clust4	1	1	81	**87**	0
clust5	0	0	**187**	0	26

Accuracy = (170 + 72 + 221 + 87 + 26) / 977 = 59%

Spectral Clustering

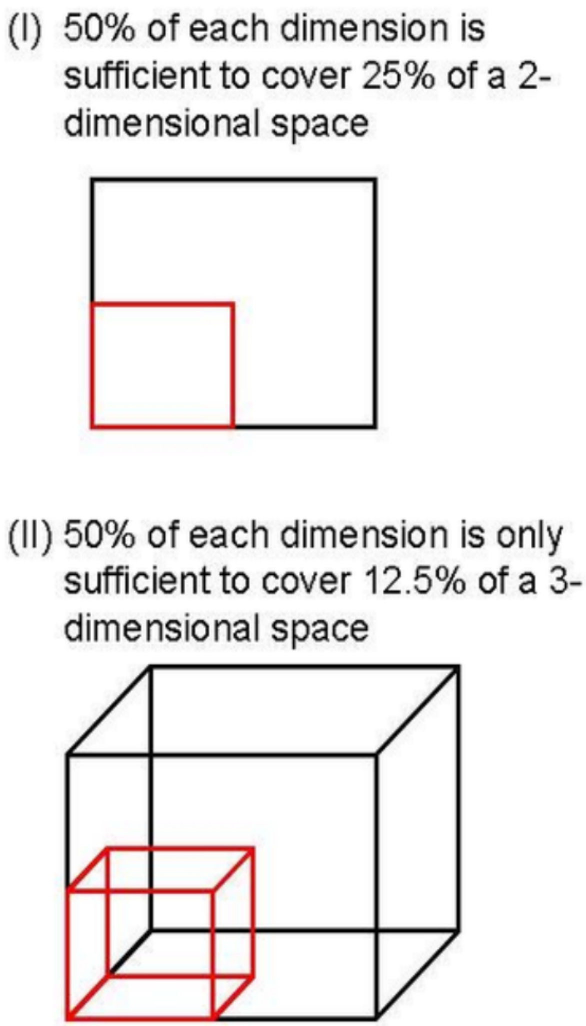
Spectrum: set of its eigenvalues

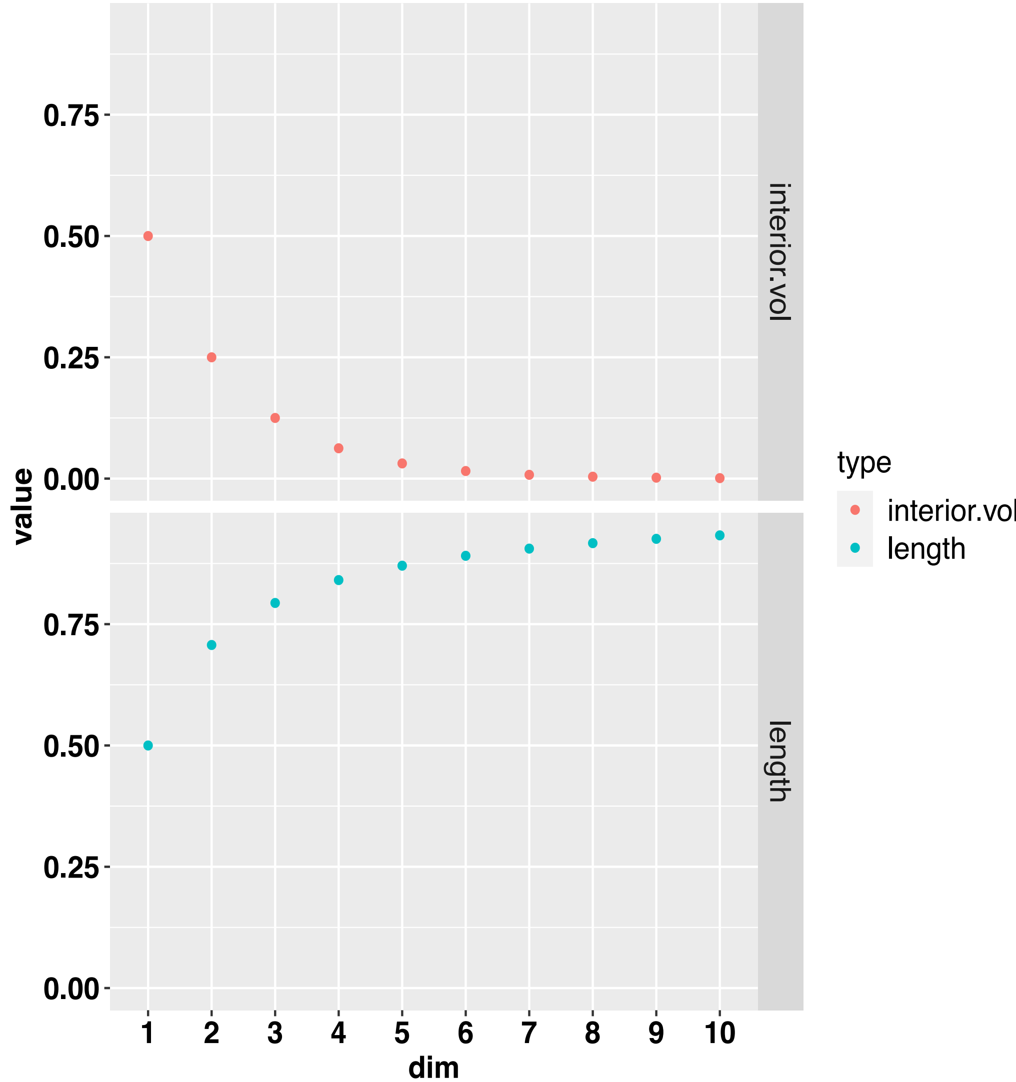


Euclidean distance is not appropriate in HD or in non-Euclidean space

- Curse of dimensionality
- Laplacian Eigenmaps is a non-linear dimension reduction method

Curse of Dimensionality



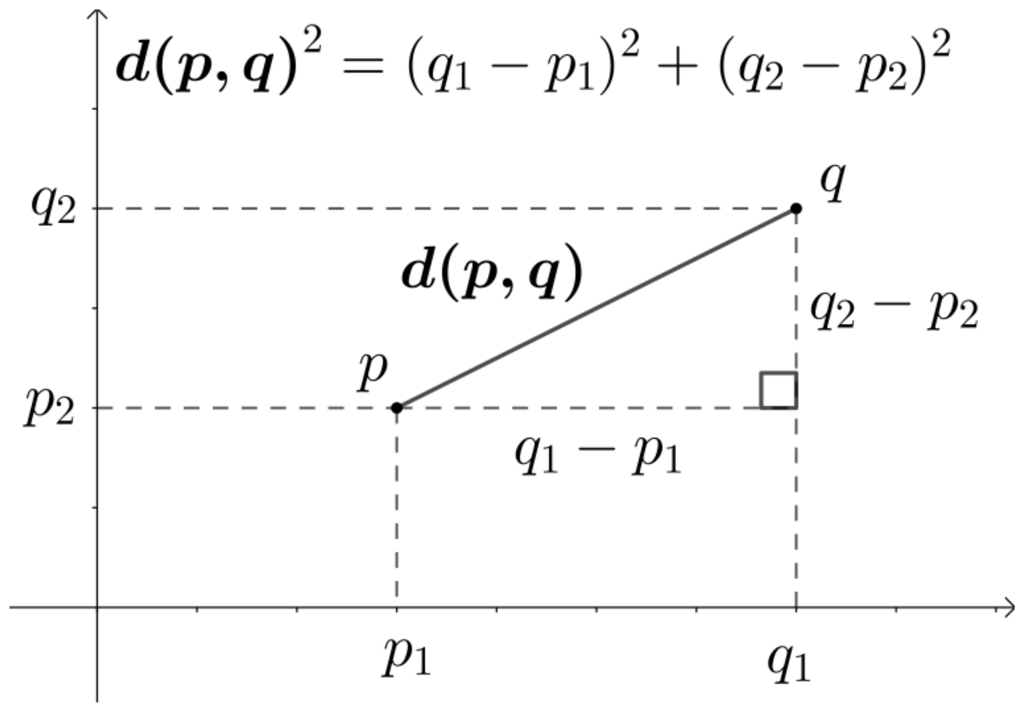


In high dimension, for any given point, all other points are on the surface of n-ball (hypersphere).

The Problems of Curse of Dimensionality

High density of local data points is required to estimate parameters in k-means and GMM

Euclidean Distance

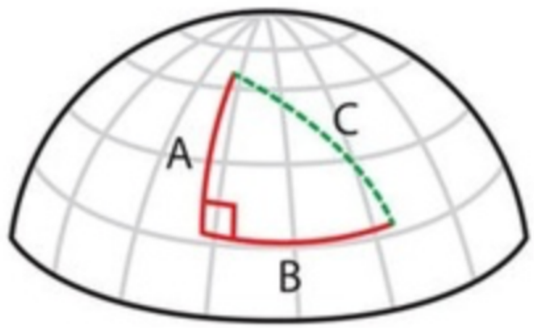


$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

Triangle on a Curved Surface vs on a Plane

Non-Euclidean

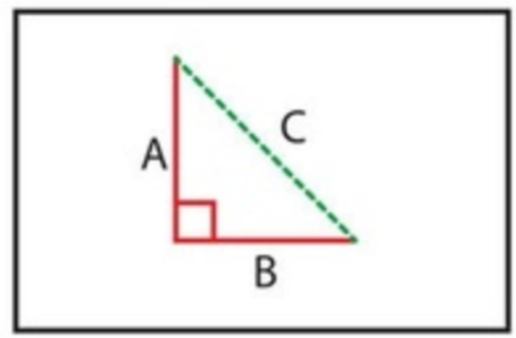
$A^2 + B^2 > C^2$



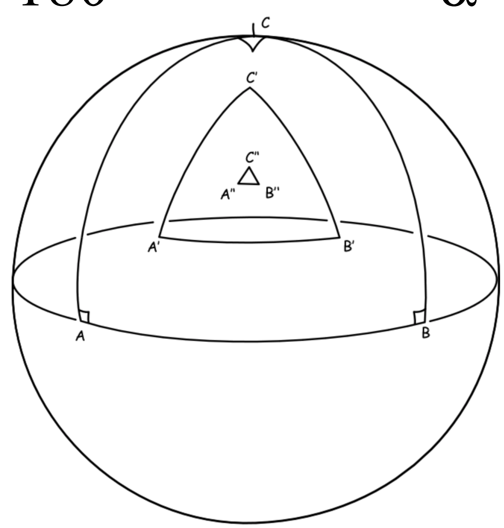
$\alpha + \beta + \gamma > 180^0$

Euclidean

$A^2 + B^2 = C^2$

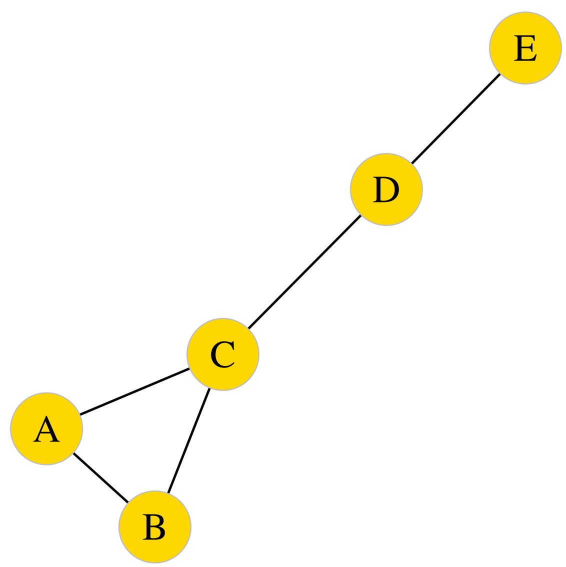


$\alpha + \beta + \gamma = 180^0$



Matrix Representation of Graph

undirected graph

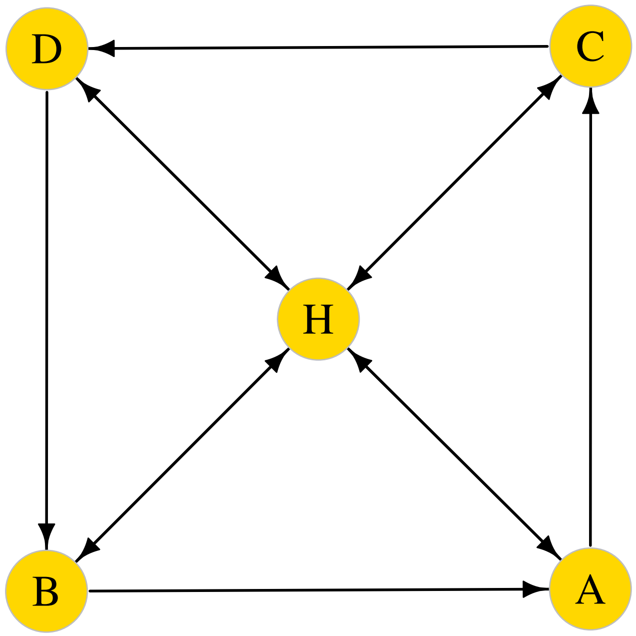


adjacency matrix

	A	B	C	D	E
A	0	1	1	0	0
B	1	0	1	0	0
C	1	1	0	1	0
D	0	0	1	0	1
E	0	0	0	1	0

Finding One-Stop Flight by Matrix Multiplication

directed graph



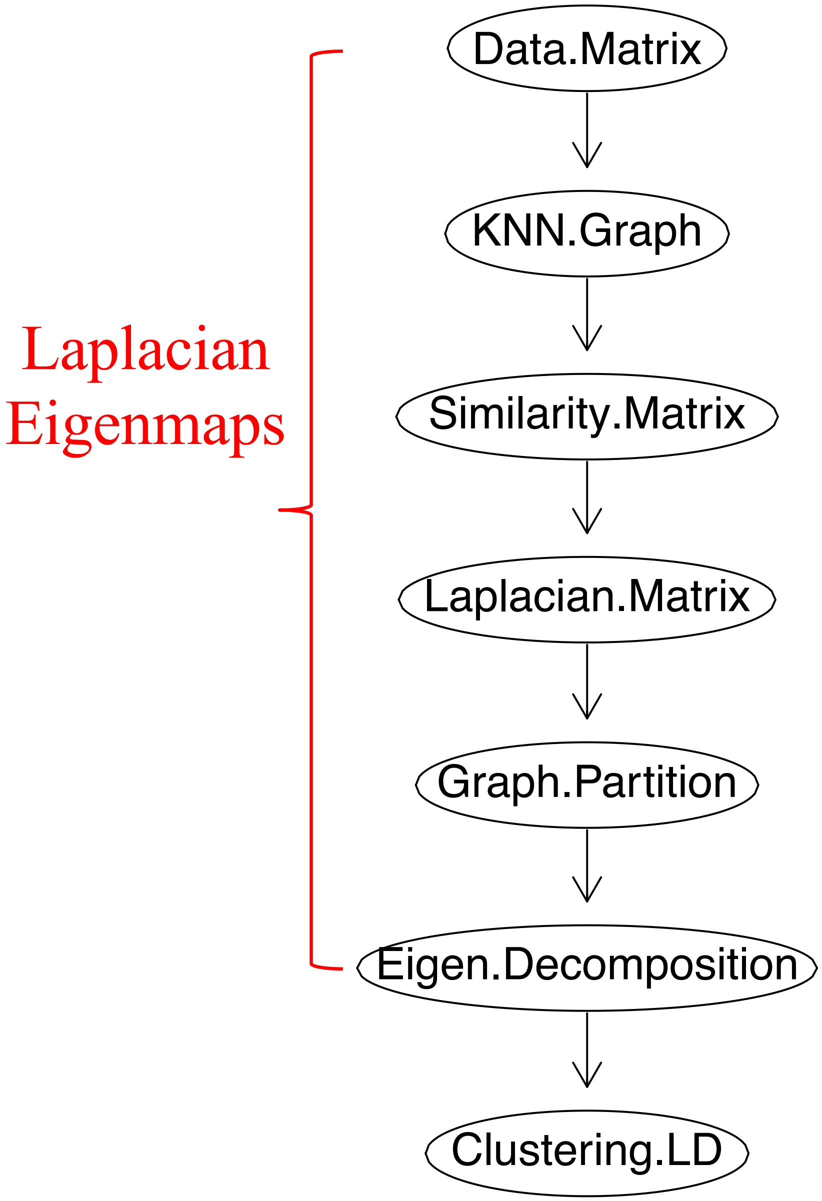
adjacency matrix (M)

	A	B	C	D	H
A	0	0	1	0	1
B	1	0	0	0	1
C	0	0	0	1	1
D	0	1	0	0	1
H	1	1	1	1	0

one-stop flight (M%*%M)

	A	B	C	D	H
A	1	1	1	2	1
B	1	1	2	1	1
C	1	2	1	1	1
D	2	1	1	1	1
H	1	1	1	1	4

Algorithm of Spectral Clustering



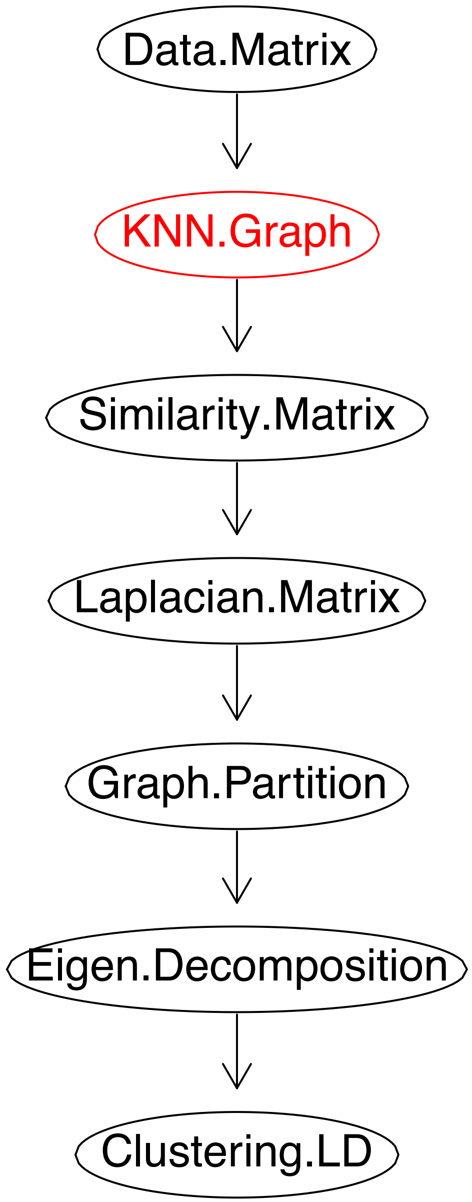
KNN: k nearest neighbor

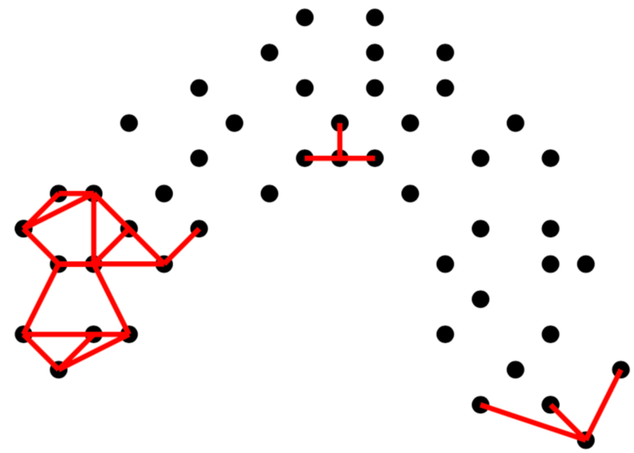
Weight matrix: $w_{ij}=\exp((x_i - x_j)^{\wedge 2}/\sigma)$

L: Laplacian Matrix (D – W)

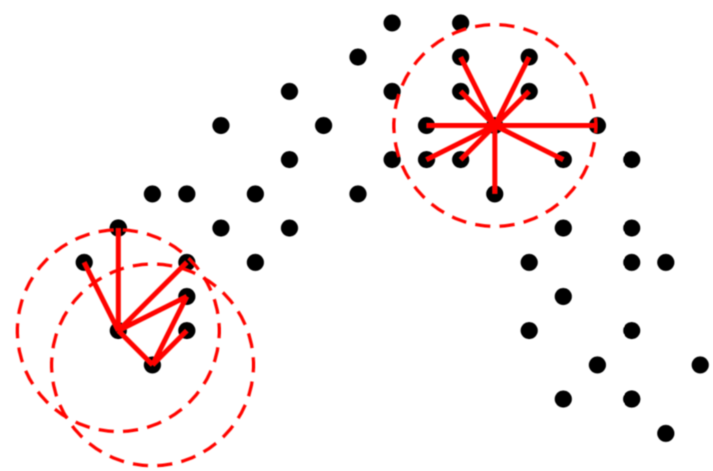
LD: low dimension

K-Nearest Neighbor Graph



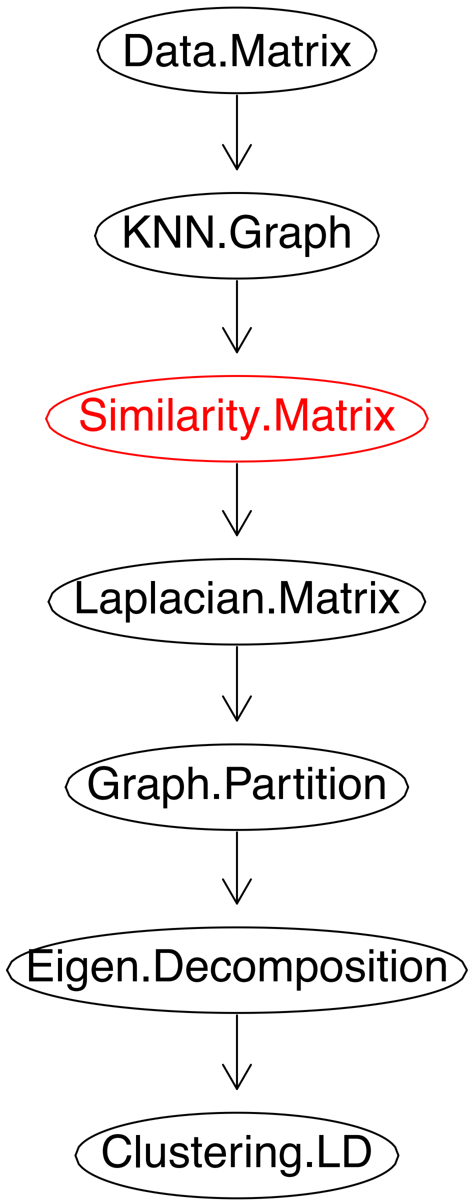


kNN graph $(k = 3)$

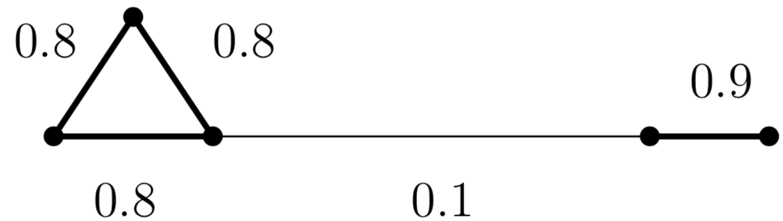


ϵ-ball graph

Similarity Matrix



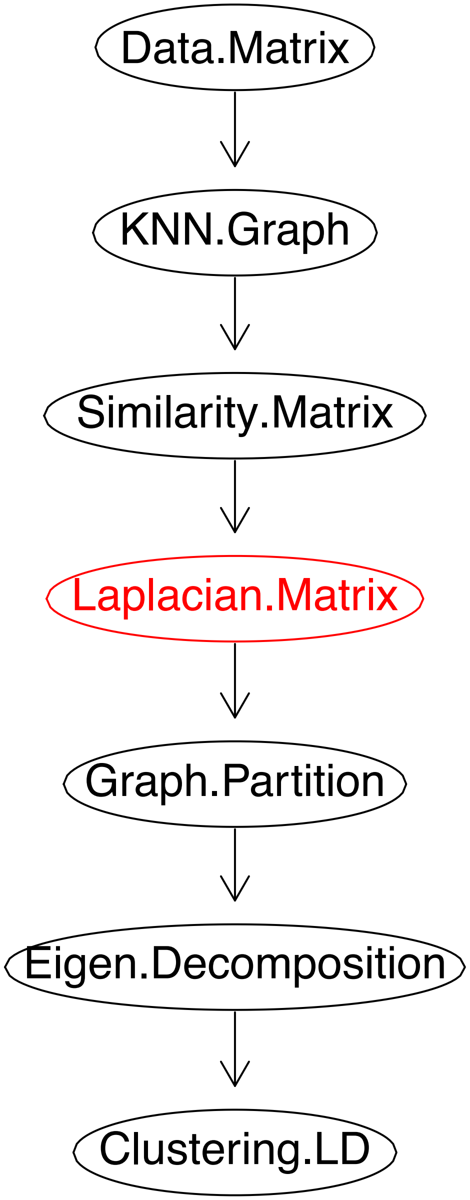
Weight matrix: $w_{ij}=\exp(-(x_i - x_j)^{\wedge 2}/2\sigma^{\wedge 2})$

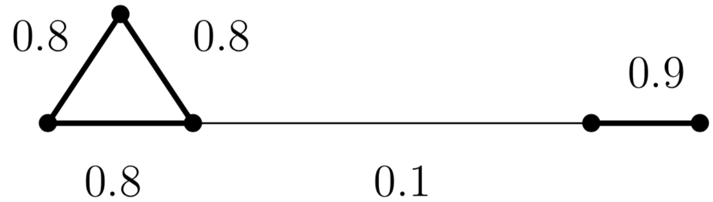


$$\mathbf{W} = \begin{pmatrix} 0 & .8 & .8 & 0 & 0 \\ .8 & 0 & .8 & 0 & 0 \\ .8 & .8 & 0 & .1 & 0 \\ 0 & 0 & .1 & 0 & .9 \\ 0 & 0 & 0 & .9 & 0 \end{pmatrix}$$

diag(W) = (0,0,0,0,0)

Laplacian Matrix



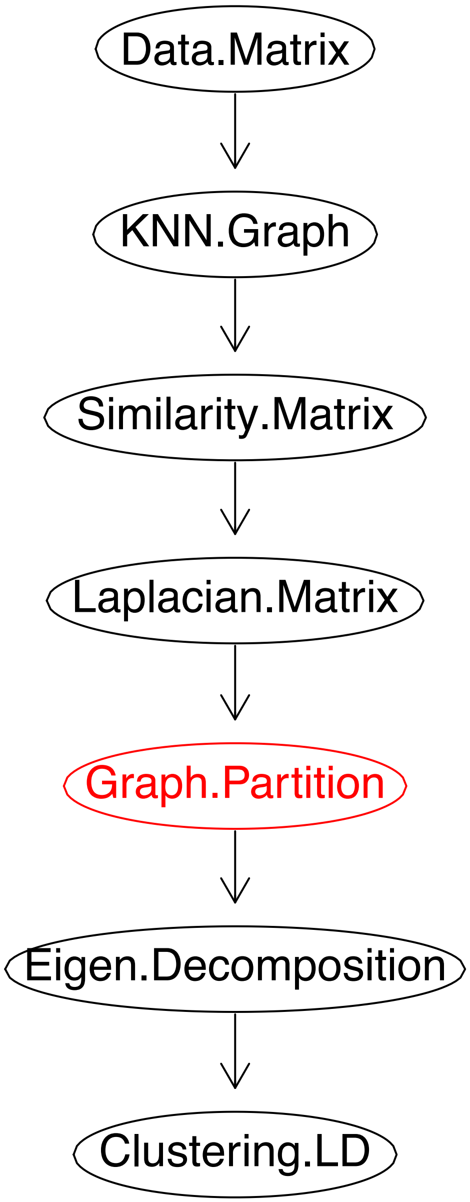


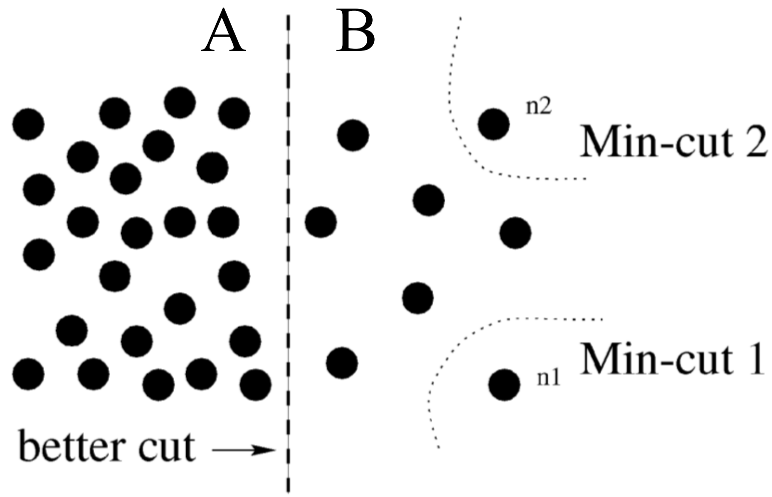
$$\mathbf{W} = \begin{pmatrix} 0 & .8 & .8 & 0 & 0 \\ .8 & 0 & .8 & 0 & 0 \\ .8 & .8 & 0 & .1 & 0 \\ 0 & 0 & .1 & 0 & .9 \\ 0 & 0 & 0 & .9 & 0 \end{pmatrix}$$

D is 5 by 5 square matrix
diag(D) = (1.6,1.6,1.7,1,0.9)

L: Laplacian Matrix = (D – W)

Graph Partitioning with Ratio Cut





$$\mathrm{Cut}(A,B) = \sum_{i \in A,\, j \in B} w_{ij}$$

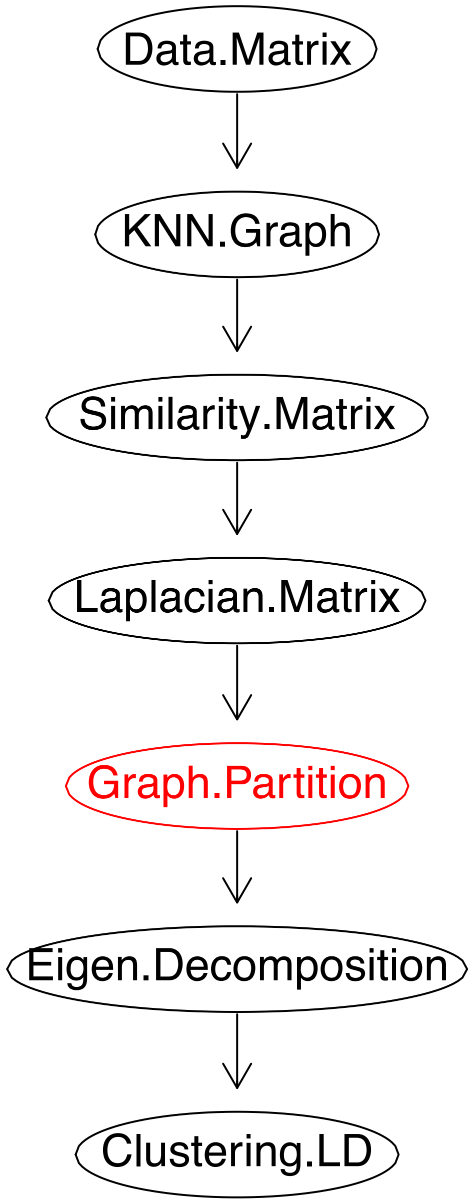
RatioCut(A,B) = cut(A,B)(1/a + 1/b)

a: number of nodes in A

b: number of nodes in B

Jianbo Shi and Jitendra Malik, 2000

Graph Partitioning with Ratio Cut



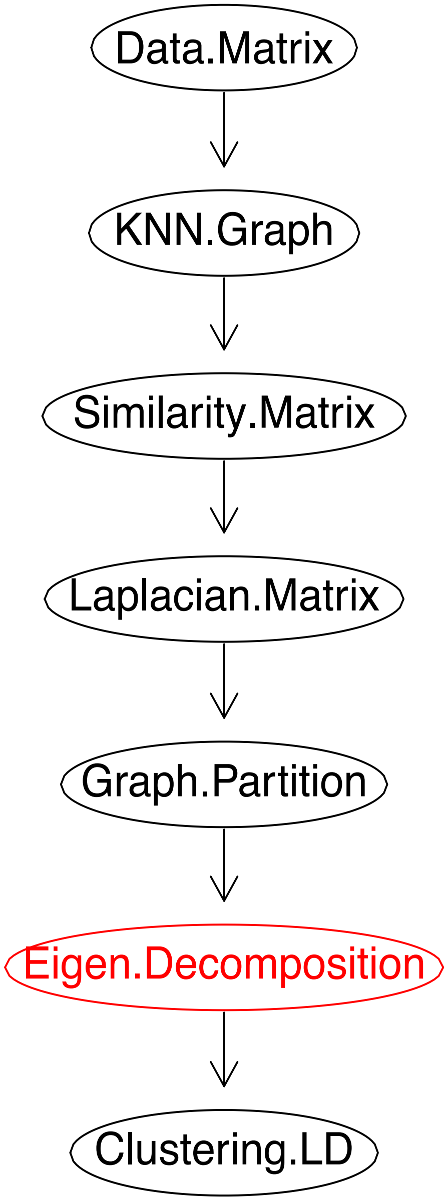
$$f_i = \begin{cases} (b/a)^{1/2}, & i \in A \\ -(a/b)^{1/2}, & i \in B \end{cases}$$

$$\sum_{i,j} w_{ij}(f_i - f_j)^2$$

$$=2(b/a + a/b + 2)\text{cut}(A,B)$$
$$=2n(1/a + 1/b)\text{cut}(A,B)$$
$$=2n\text{RatioCut}(A,B)$$

n: total number of nodes (a+b)

Eigen Decomposition of Laplacian Matrix



$$\mathbf{f} = (f_1, f_2, \ldots, f_n)$$

$$\mathbf{f^T L f} = \mathbf{f^T (D\text{-}W) f} = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2$$

L is positive semidefinite

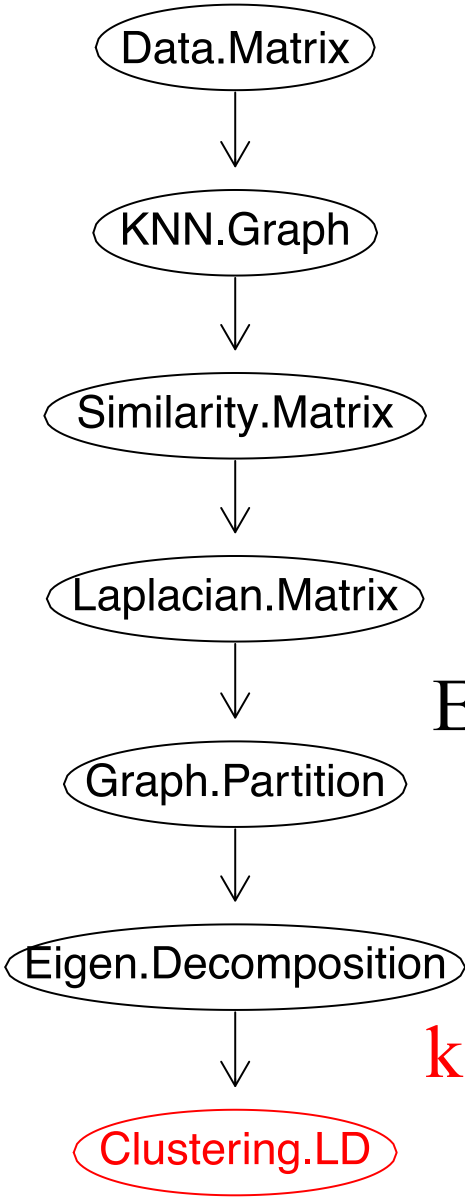
$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

$$\text{Cost function} = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2$$

subject to $\mathbf{f^T f} = 1$

Eigen decomposition of Laplacian matrix L

K-means Clustering in Low Dimension

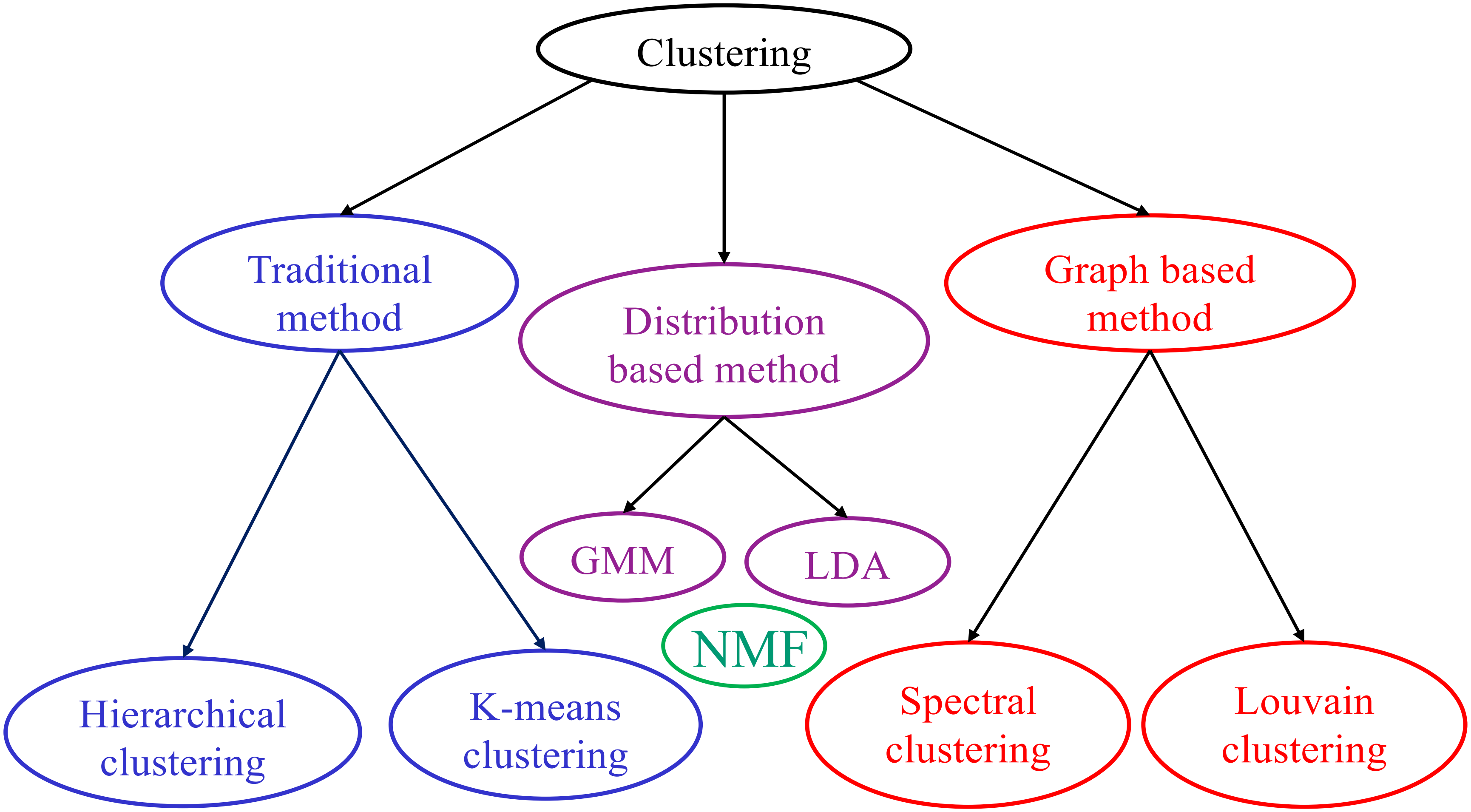


$$0 = \lambda_1 \leq \lambda_2 \leq \cdot \cdot \cdot \leq \lambda_n$$

Eigen decomposition of Laplacian matrix L

k-means clustering/GMM in low-dimension
with eigen vectors $\mathbf{f}_2, \ldots, \mathbf{f}_k$

Outline of Clustering Methods



GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai