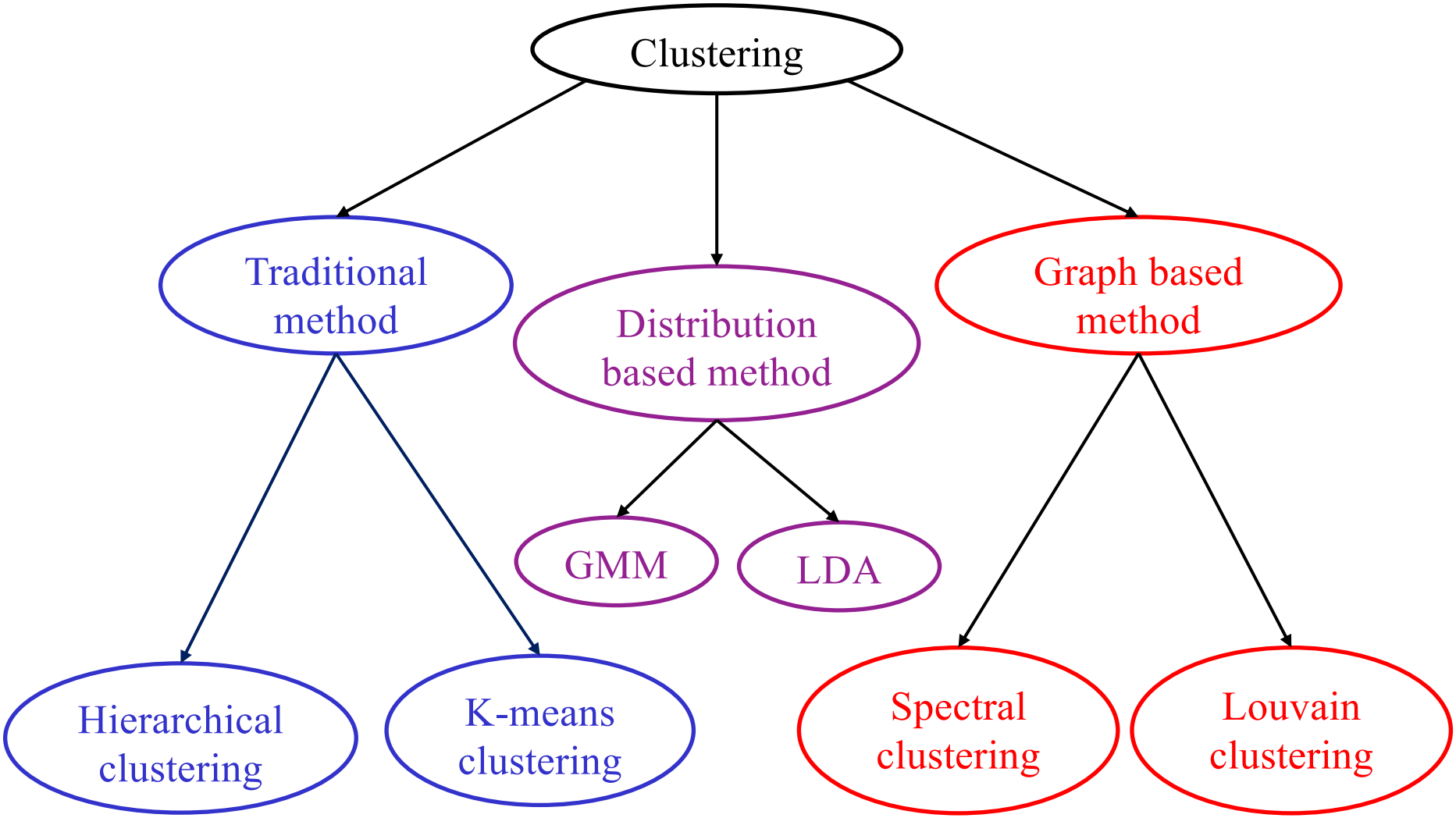
Clustering Methods:
From k-means to Gaussian Mixture Model and Louvain Algorithm

Maxwell Lee

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

September 21, 2020

Outline of Clustering Methods



GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation

Contributed by Emily Tai

Data Matrix (Table)

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ . & & & . \\ . & & & . \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

X_{np}

n observations and p variables

Multivariate Linear Regression Model

y is response variable or
dependent variable

$x_1 \ldots x_p$ are independent variables

$$\left[\begin{array}{c|ccc} y_1 & x_{11} & x_{12} \ \ldots & x_{1p} \\ y_2 & x_{21} & x_{22} \ \ldots & x_{2p} \\ . & & & . \\ . & & & . \\ y_n & x_{n1} & x_{n2} \ \ldots & x_{np} \end{array}\right]$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ \ldots + \beta_p x_p + \varepsilon$$

$$y = X\beta + \varepsilon$$

Application of Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y	x	application
Tumor size	Gene expression	correlation
Gene expression	Treatment vs control	t-test
Treatment response	Gene expression	Classification (glm)

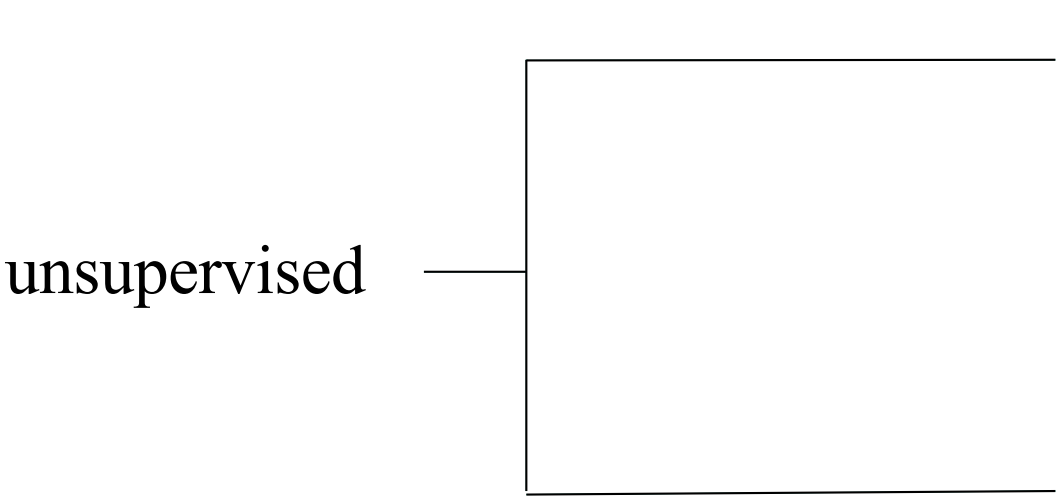
glm: generalized linear model

Unsupervised Analysis

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- We do not have data for response variable y or sample label
- We are more interested in intrinsic relationship among samples and variables without influenced by the response variable

Unsupervised Statistical Learning



Clustering analysis:
k-means, GMM, LDA
hierarchical clustering,
spectral clustering
Louvain clustering

Dimension reduction:
PCA, MDS, TSNE, UMAP

GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation

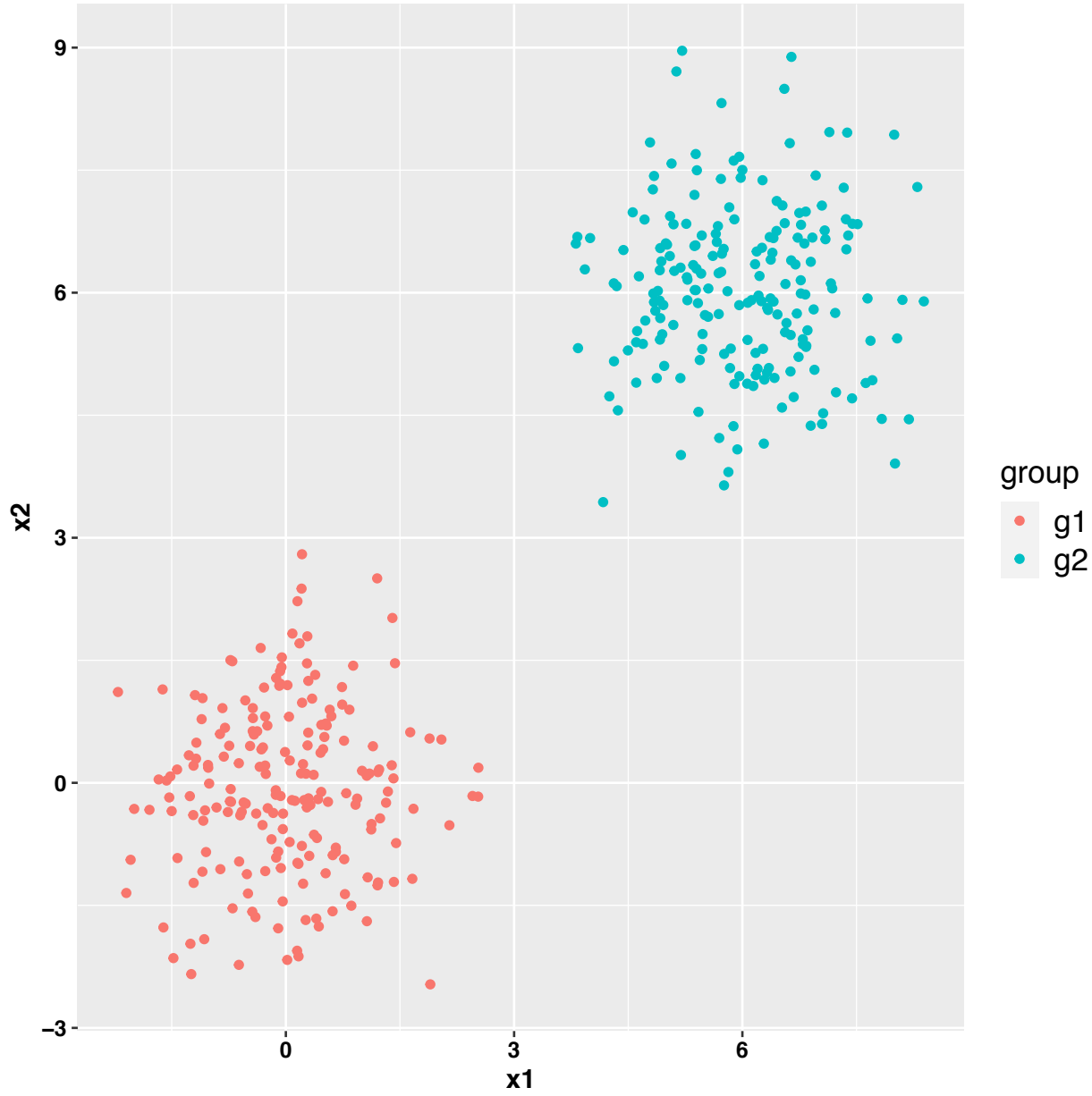
PCA: Principal Component Analysis
MDS: Multidimensional scaling
TSNE: T-distributed Stochastic Neighbor Embedding
UMAP: Uniform Manifold Approximation and Projection

Cluster Structure Reflects Sample Heterogeneity

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

$\rho = 0$

Group1		Group2	
μ_1	μ_2	μ_1	μ_2
0	0	6	6

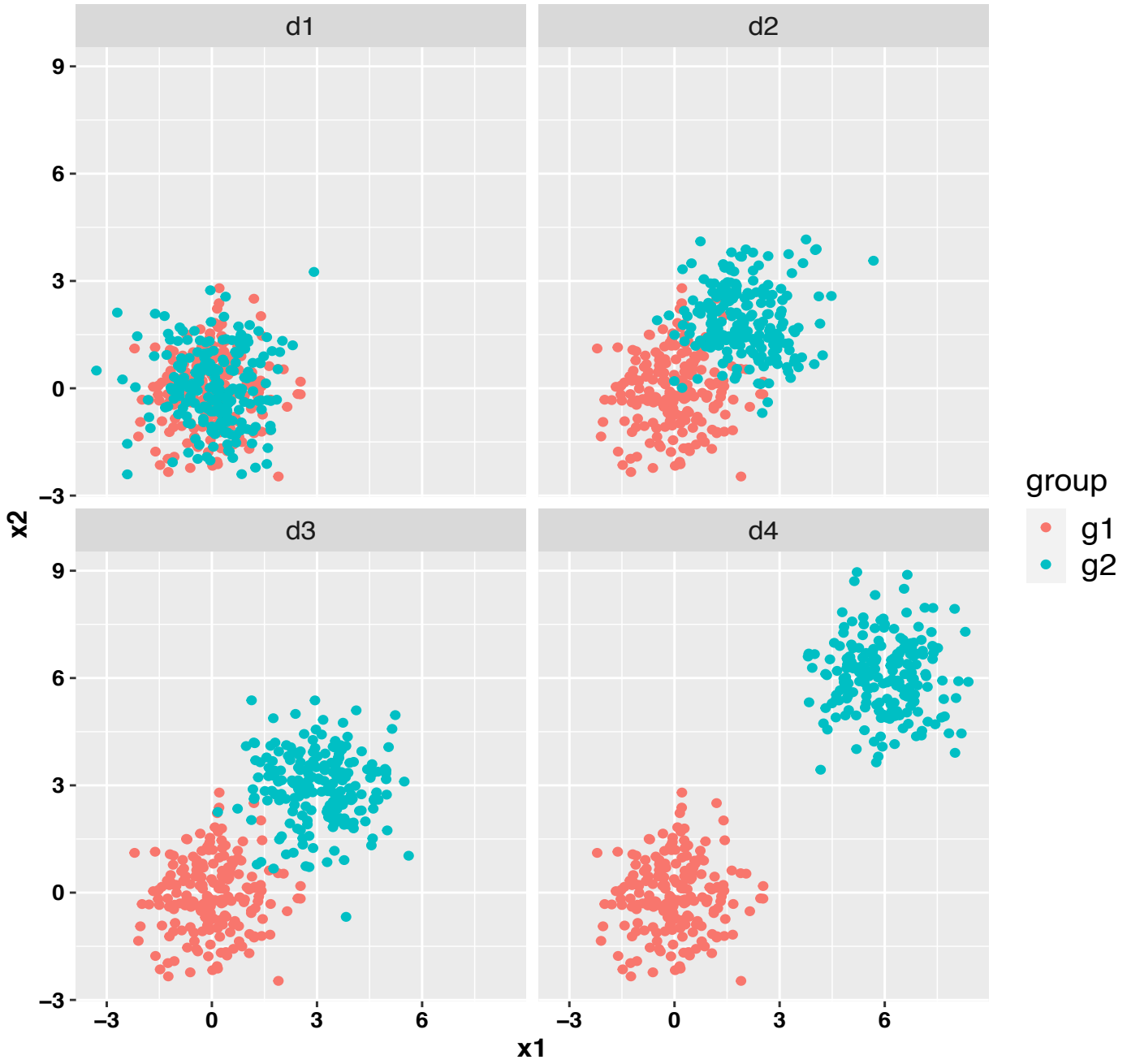


Increased Separation Between Clusters Is Related to Increased Distance Between the Groups

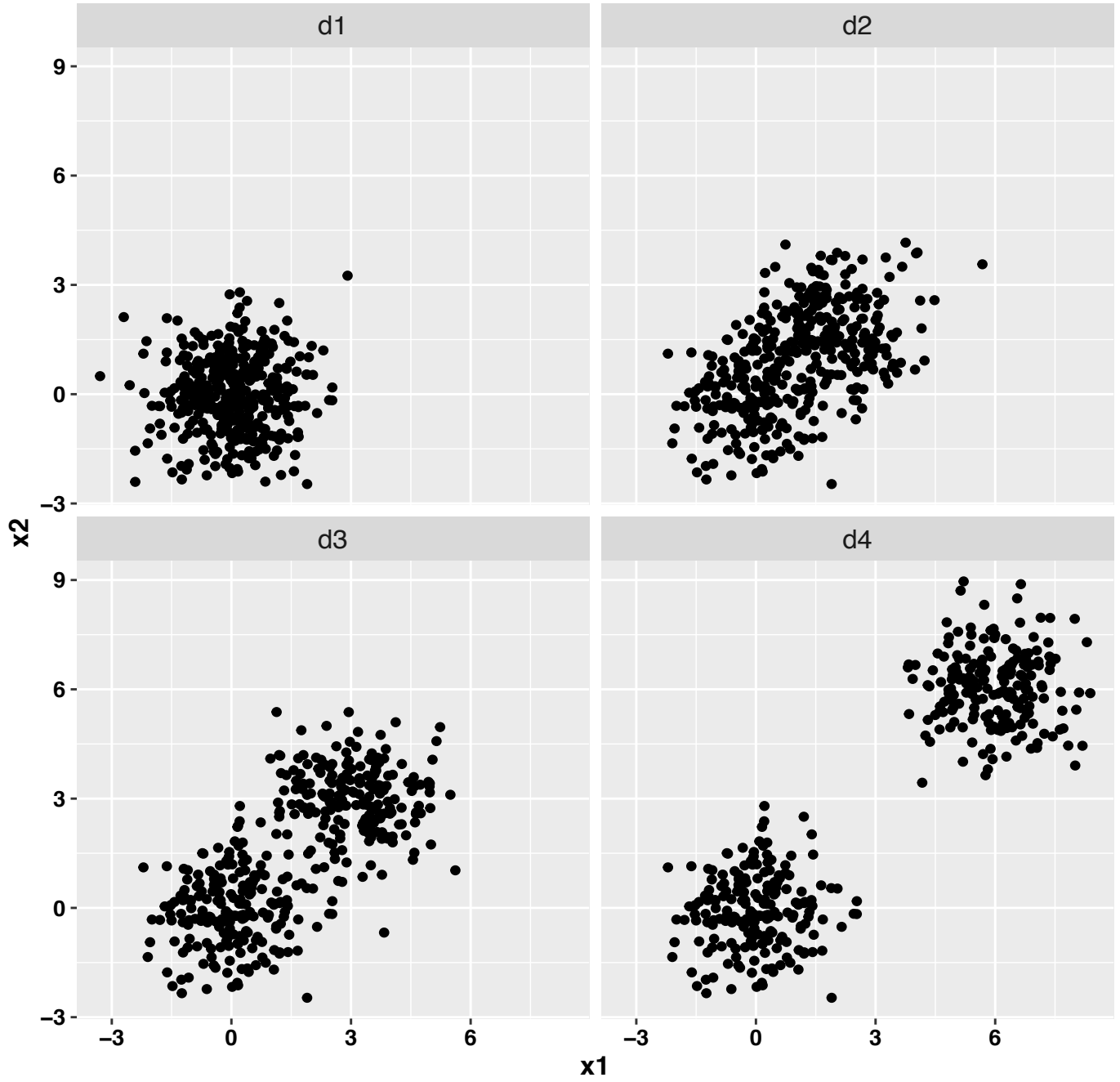
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

$\rho = 0$

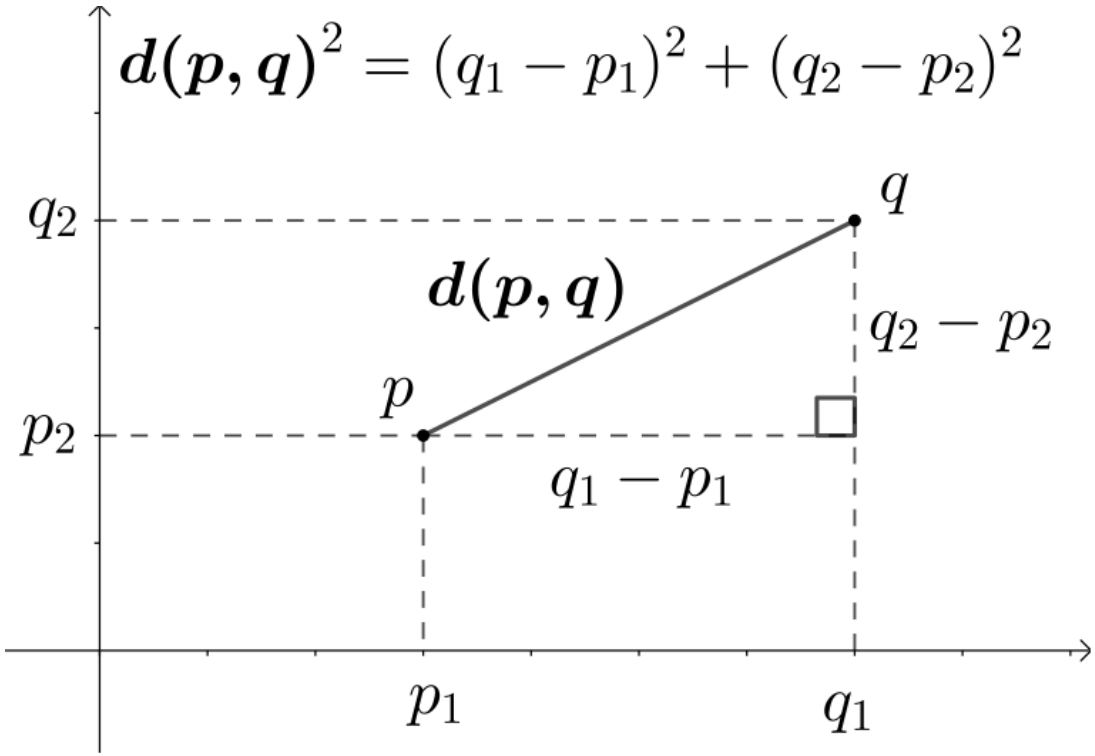
	Group1		Group2	
	μ_1	μ_2	μ_1	μ_2
d1	0	0	0	0
d2	0	0	2	2
d3	0	0	3	3
d4	0	0	6	6



How To Determine Cluster Structure Without Sample Label?

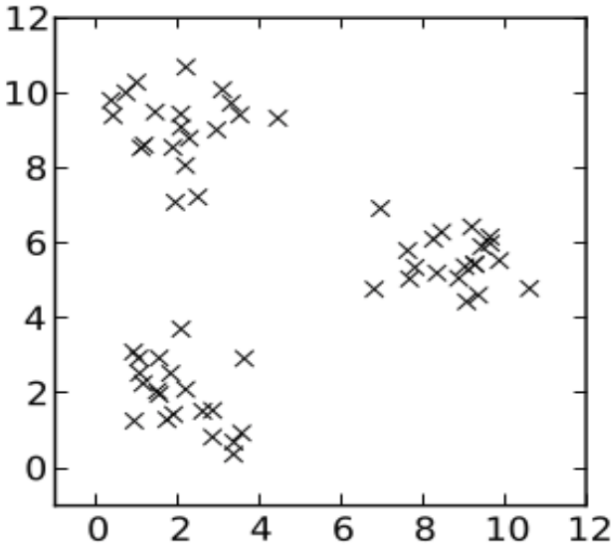


Euclidean Distance

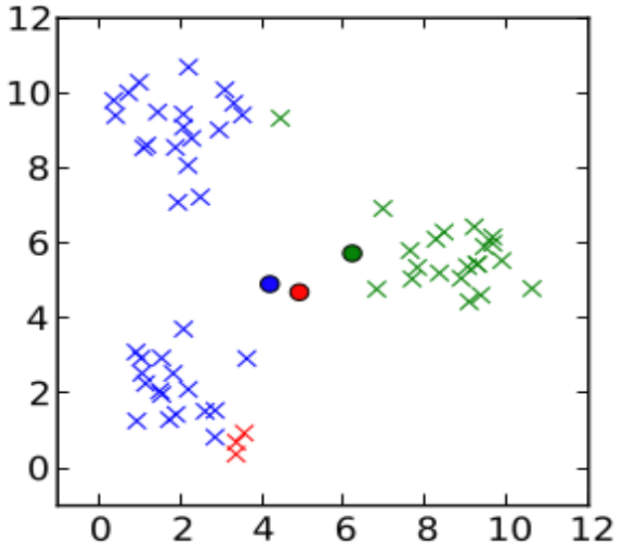


$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

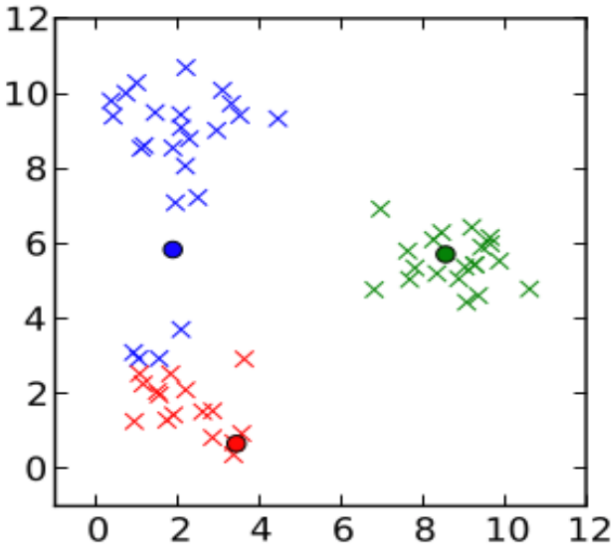
Outline of k-means Clustering Method



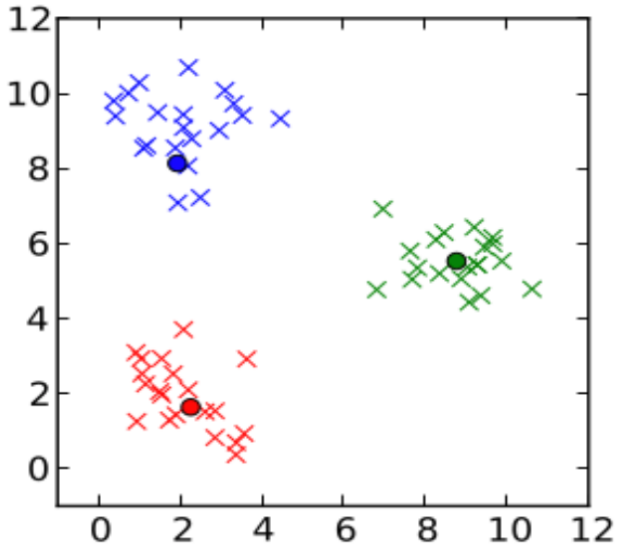
(a) dataset.



(b) step 1.



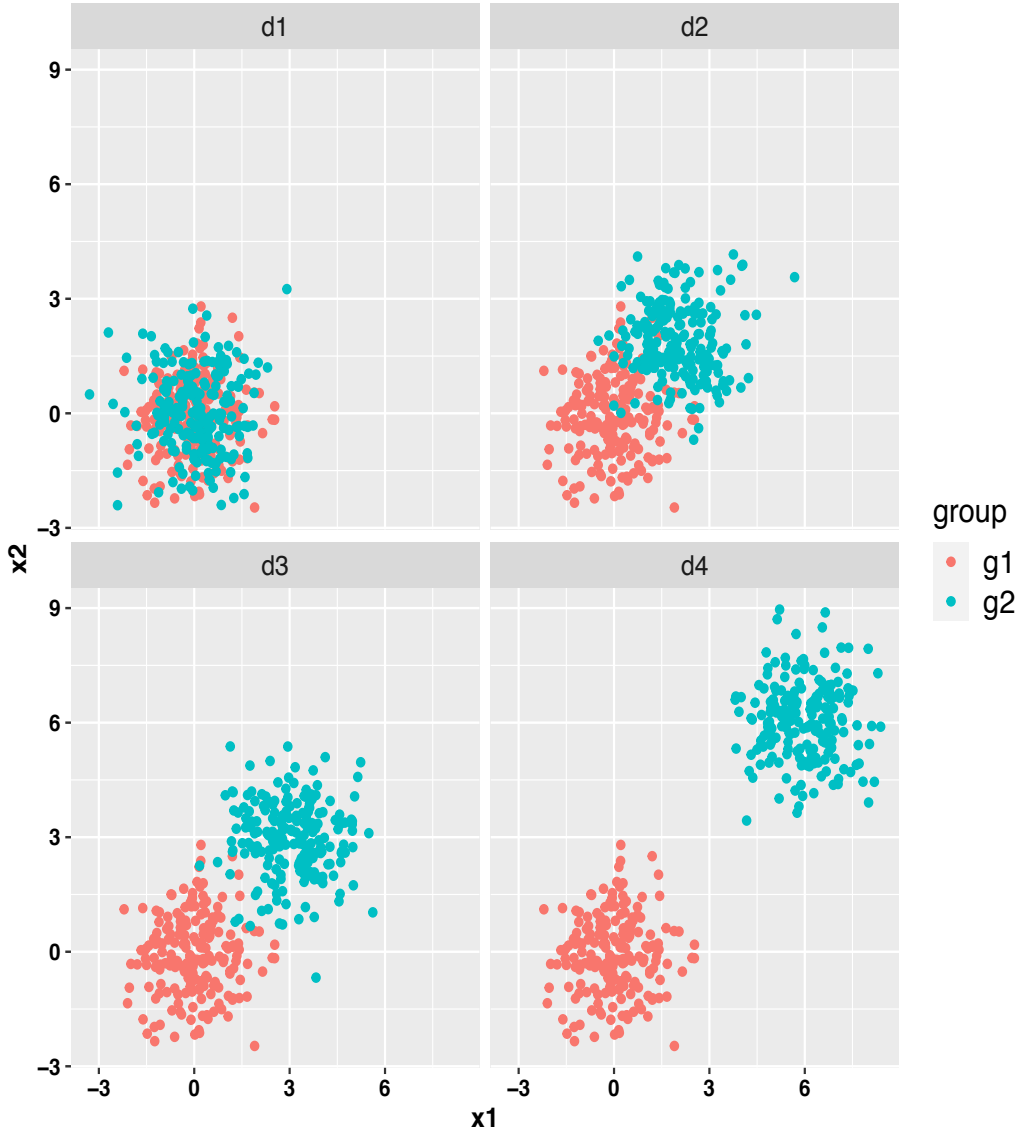
(c) step 2.



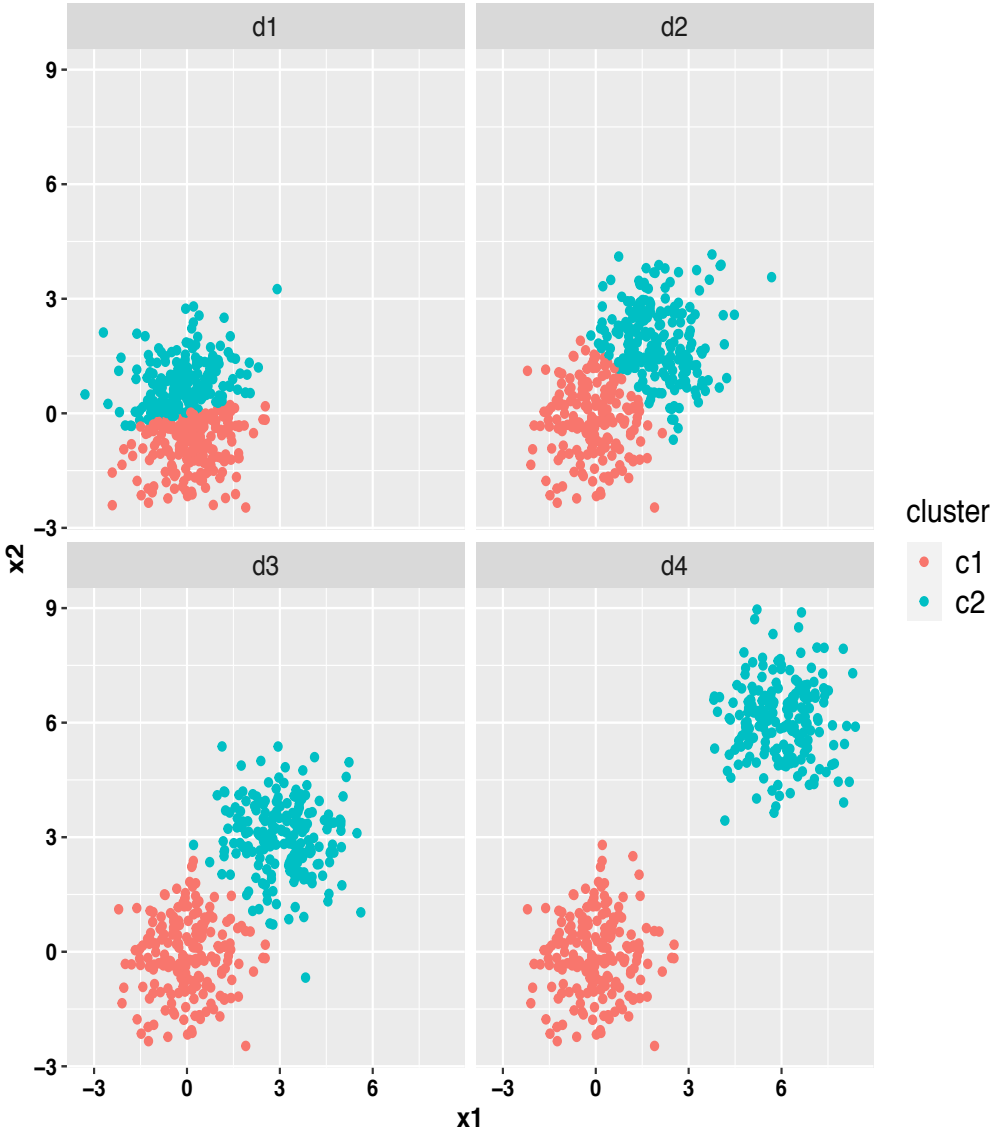
(d) step 3.

K-means Clustering

color by group



color by k-means cluster



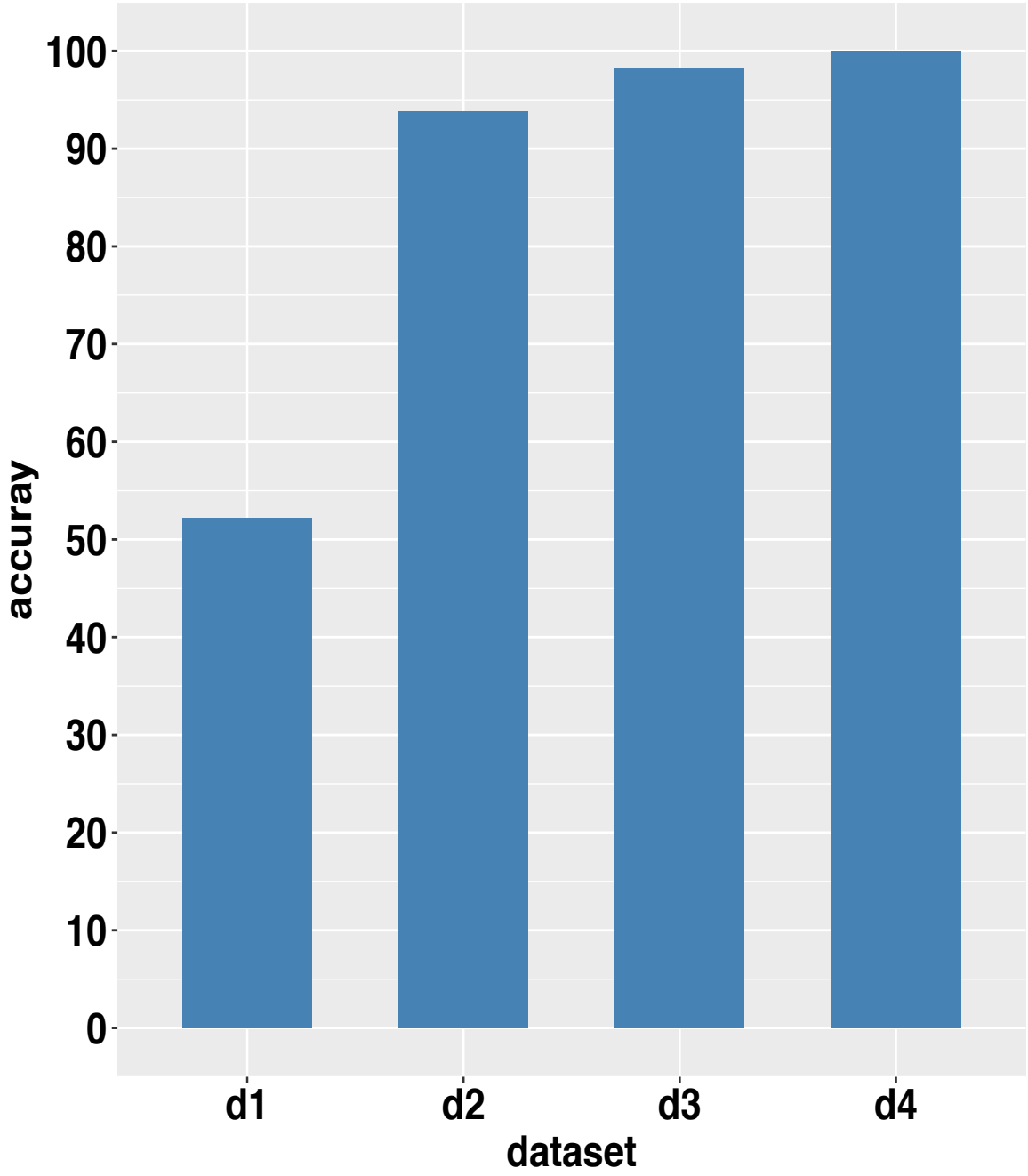
Accuracy of k-means Clustering

Confusion matrix
Column: actual category
Row: assigned category

	g1	g2
c1	183	8
c2	17	192

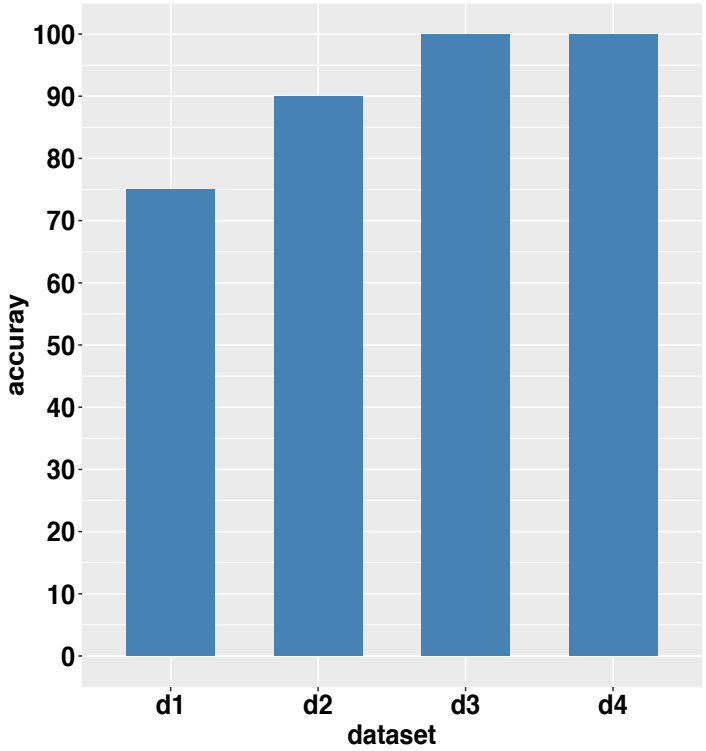
accuracy of dataset d2
Match: diagonal elements (red)
Mismatch: off diagonal elements (green)

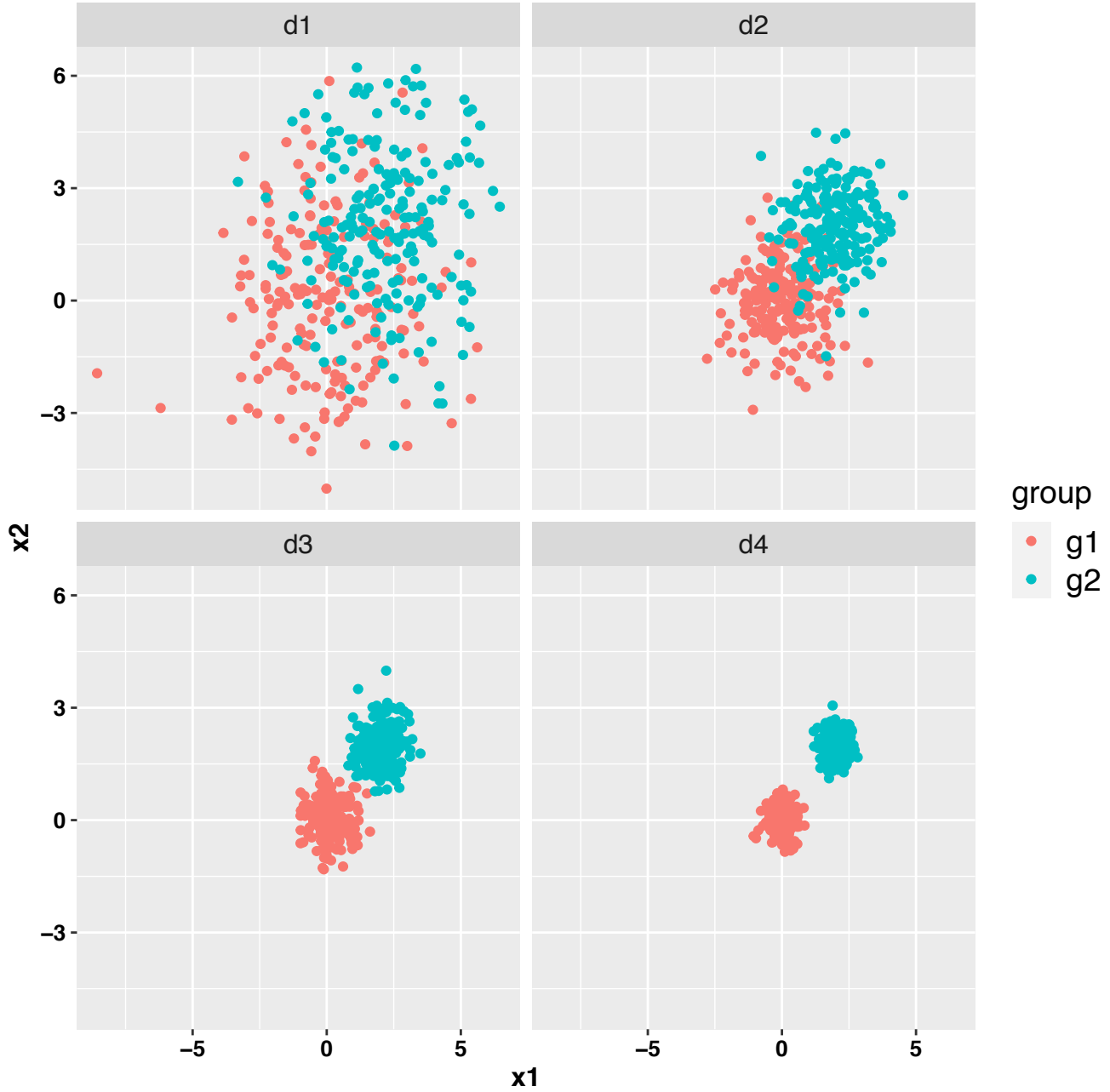
accuracy=(183+192)/400
=93.75%



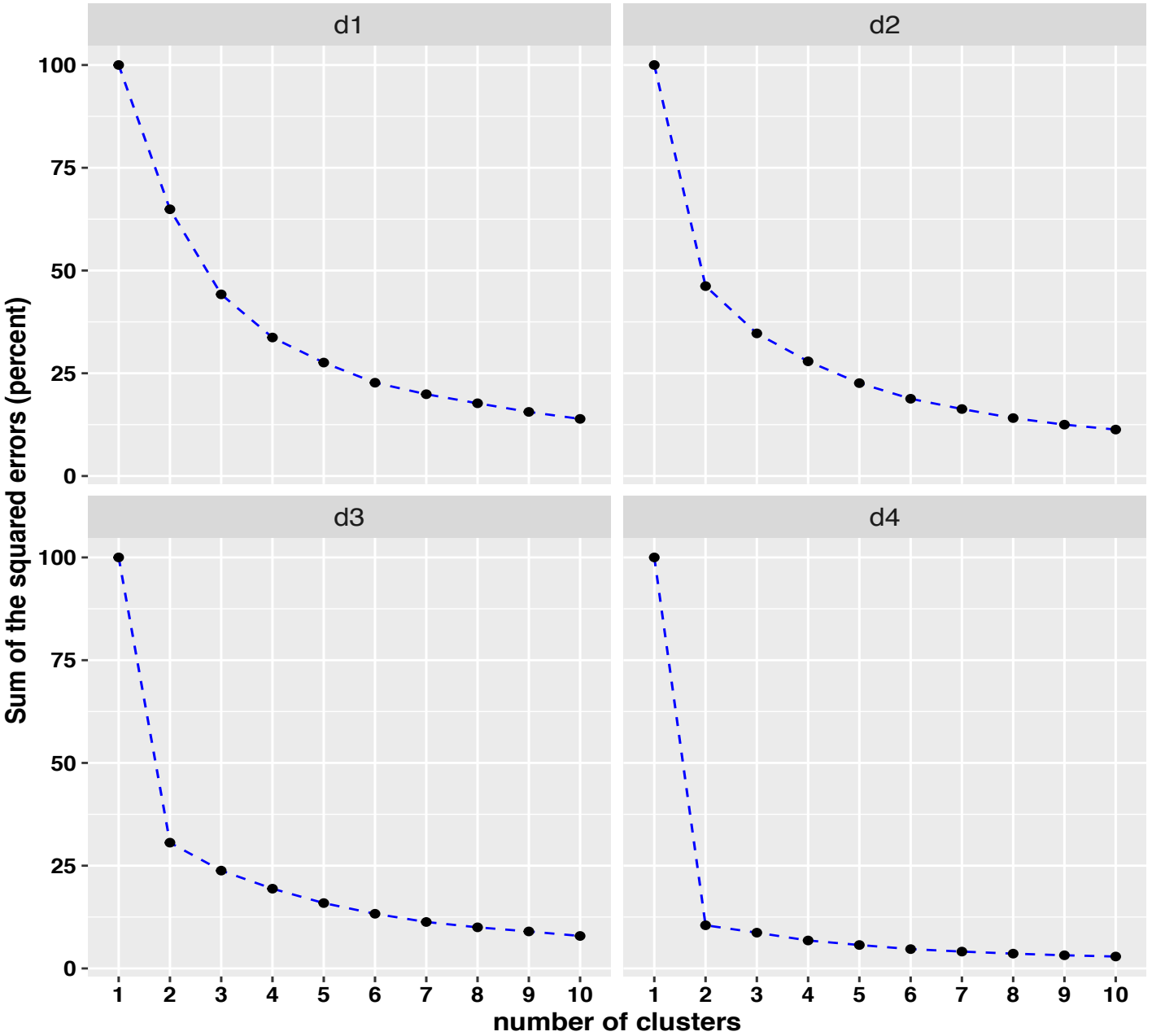
Better Separation Between Clusters If Within Group Variance is Reduced

	variance
d1	4
d2	1
d3	0.3
d4	0.1





How To Choose the Number of Clusters?

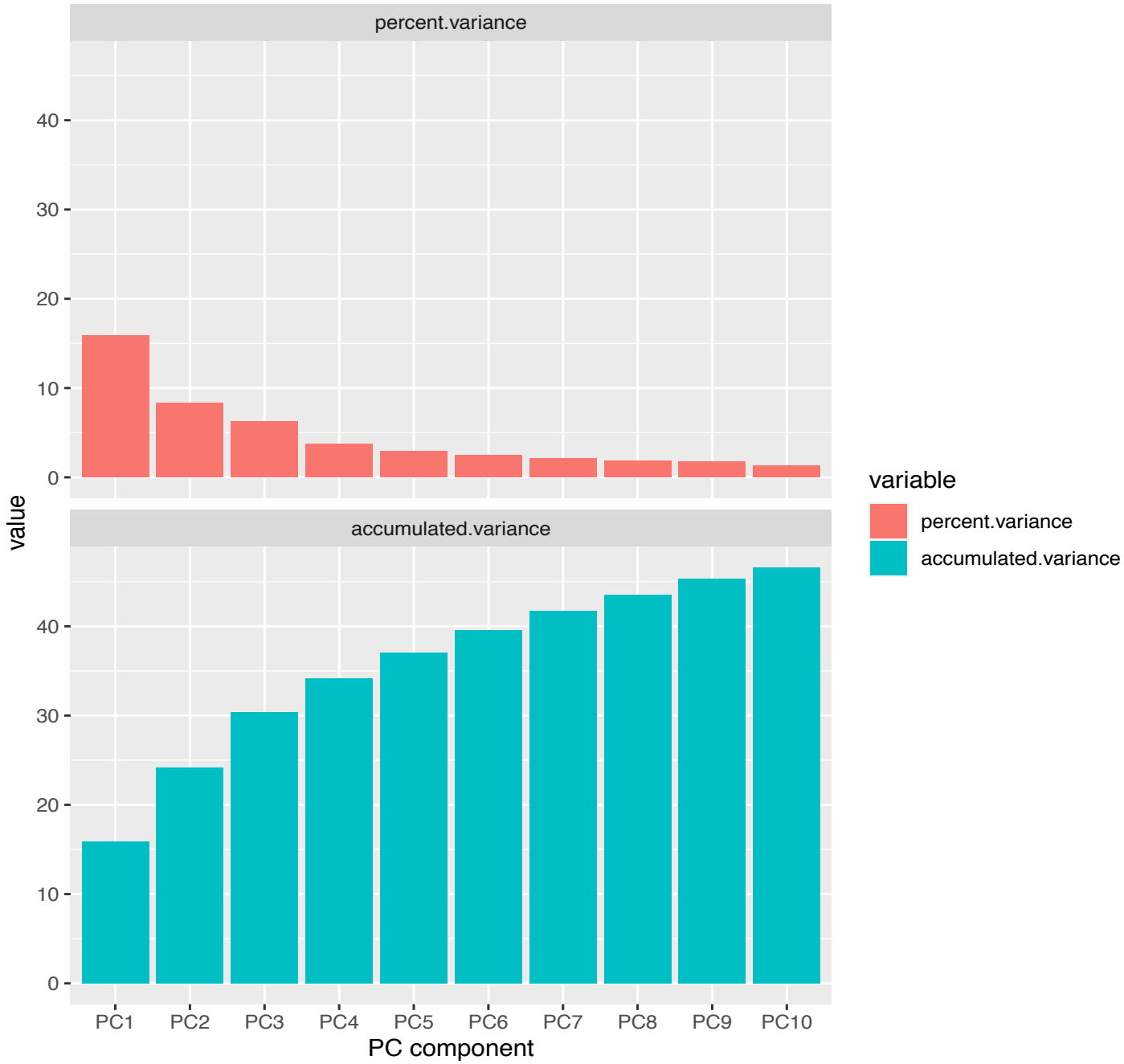


PCA of TCGA BRCA Samples

977 samples
5000 genes

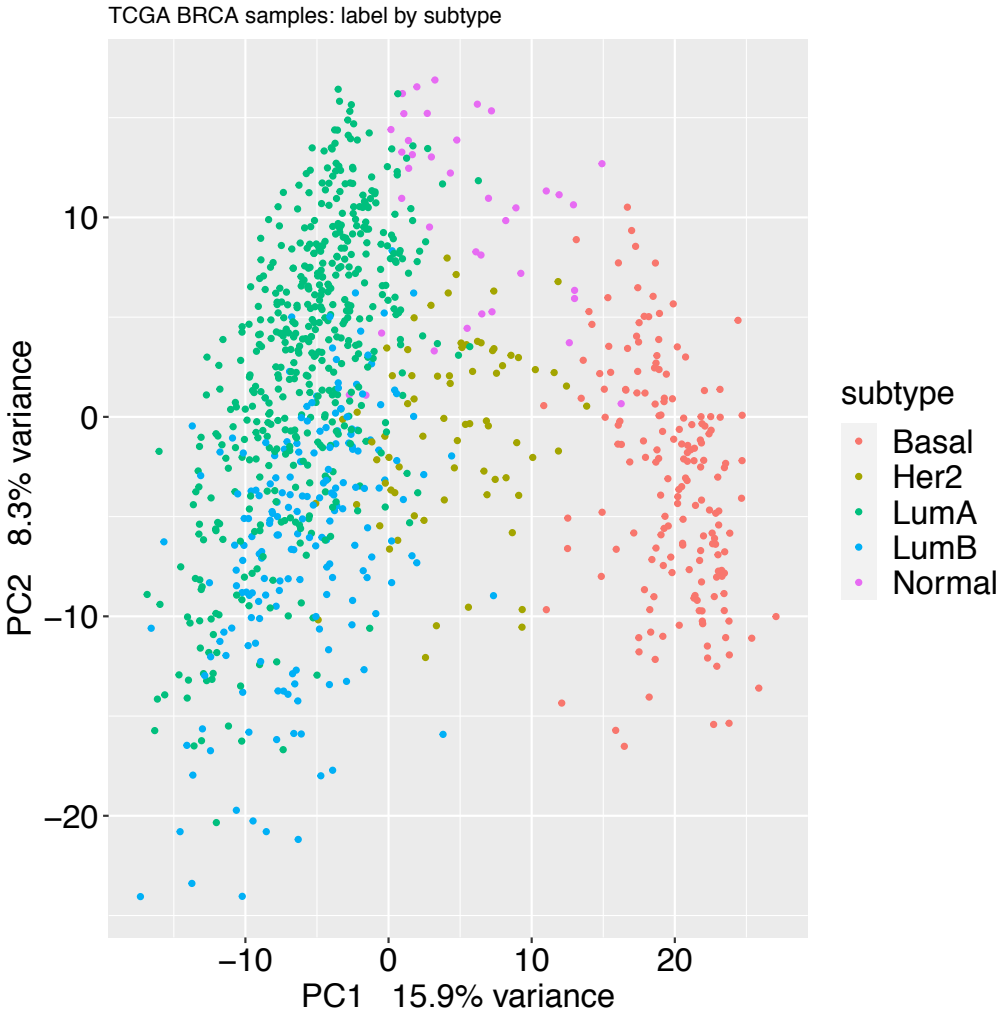
x	freq
Basal	173
Her2	73
LumA	500
LumB	193
Normal	38

Percent of Variance Accounted for by PCA Components

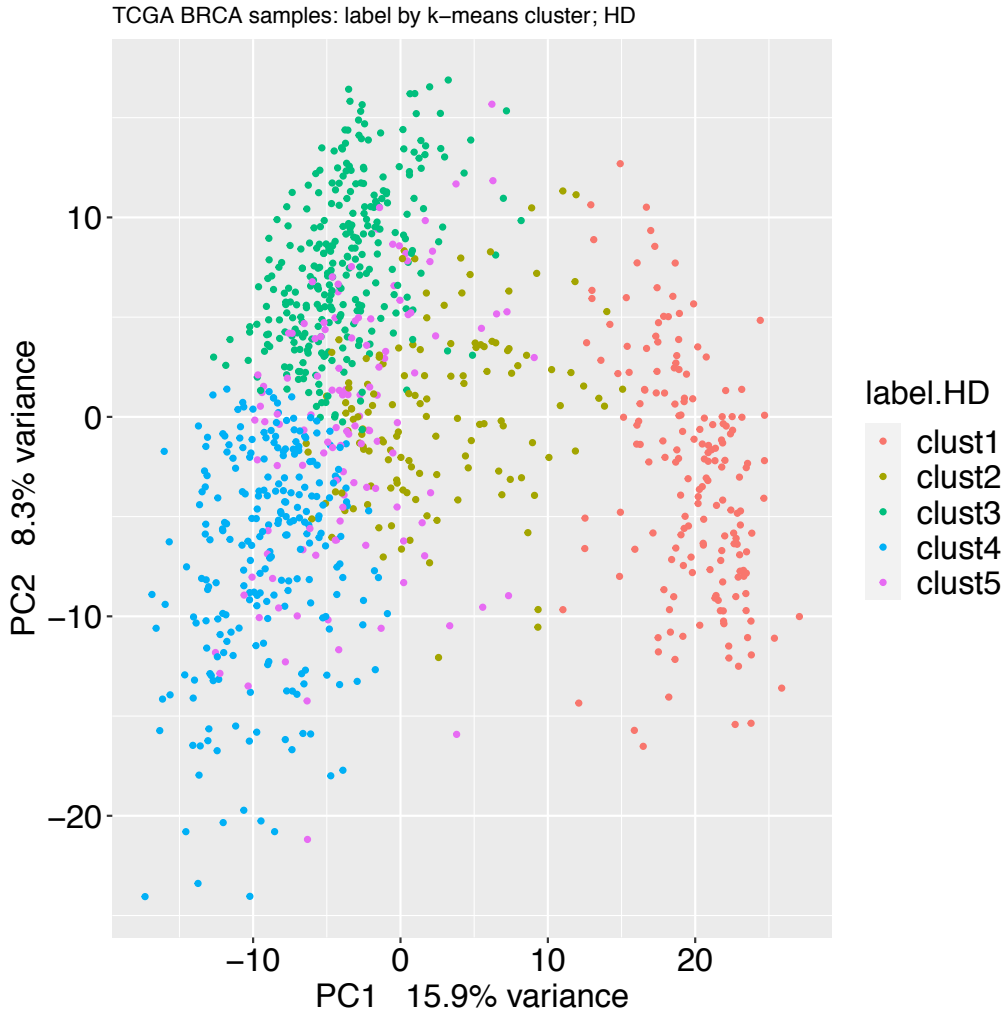


PCA: Label by Subtype vs. by k-means Cluster

Label by subtype



Label by k-means clusters in high-dimension



HD: high dimension, 5000 genes

Comparison Between Subtype and Cluster (HD)

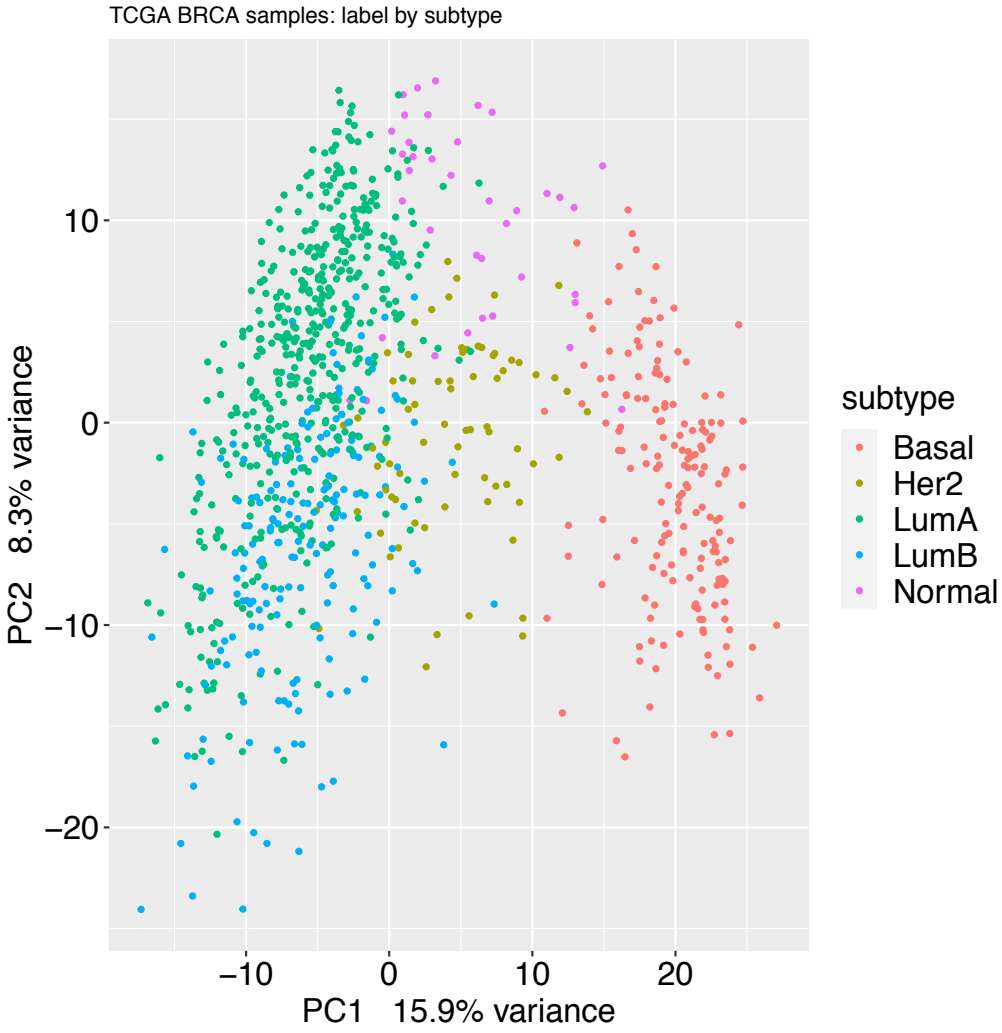
Confusion matrix
Column: actual category
Row: assigned category

	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

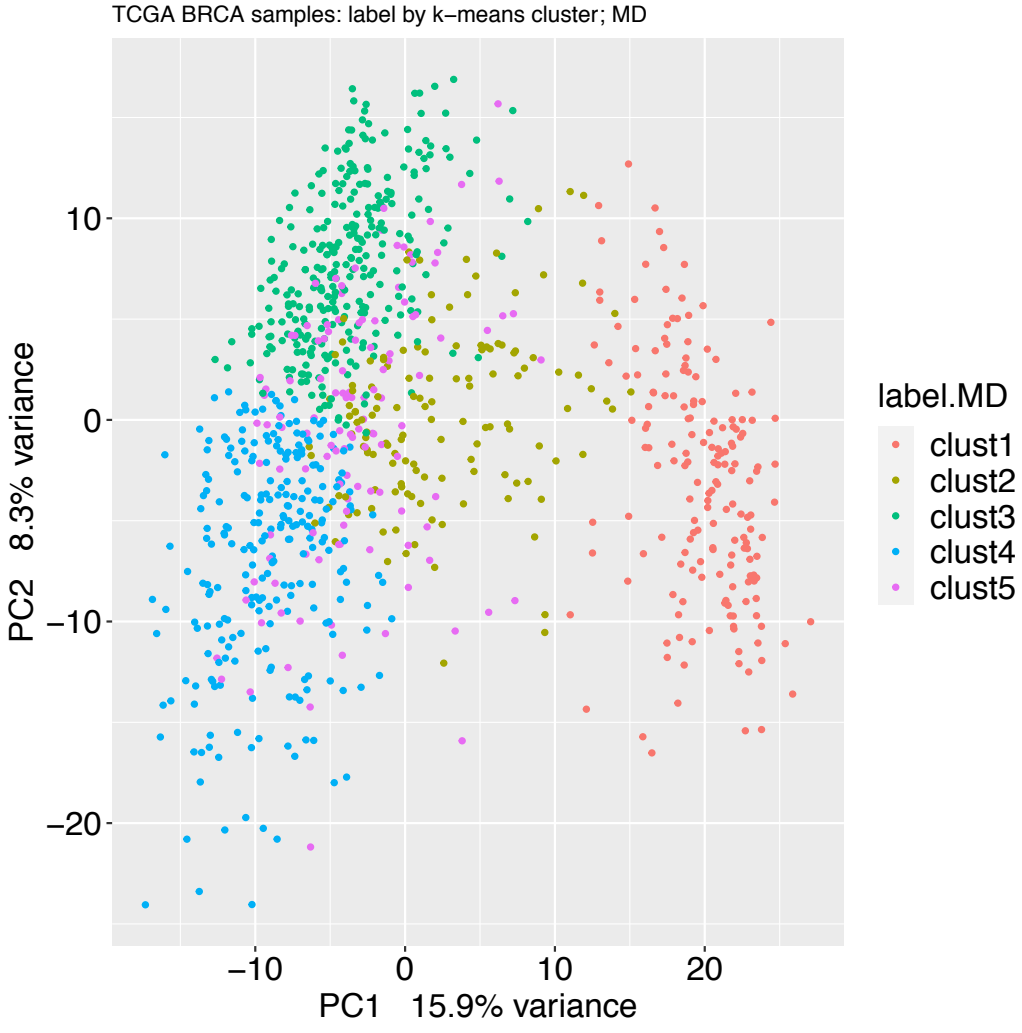
Match: diagonal elements (red)
Mismatch: off diagonal elements (green)
Accuracy = (169 + 69 + 268 + 119 + 6) / 977 = 65%

PCA: Label by Subtype vs. by k-means Clusters

Label by subtype



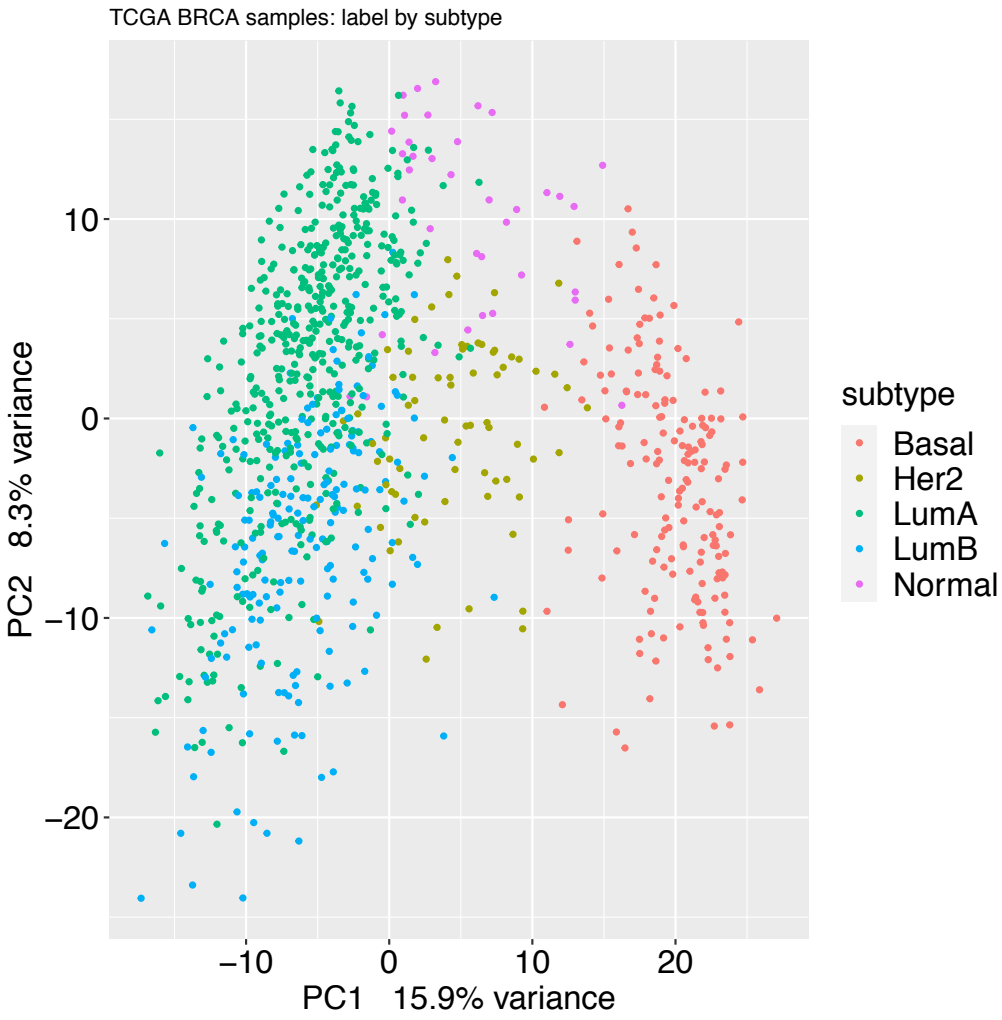
Label by k-means clusters with 10 PCs



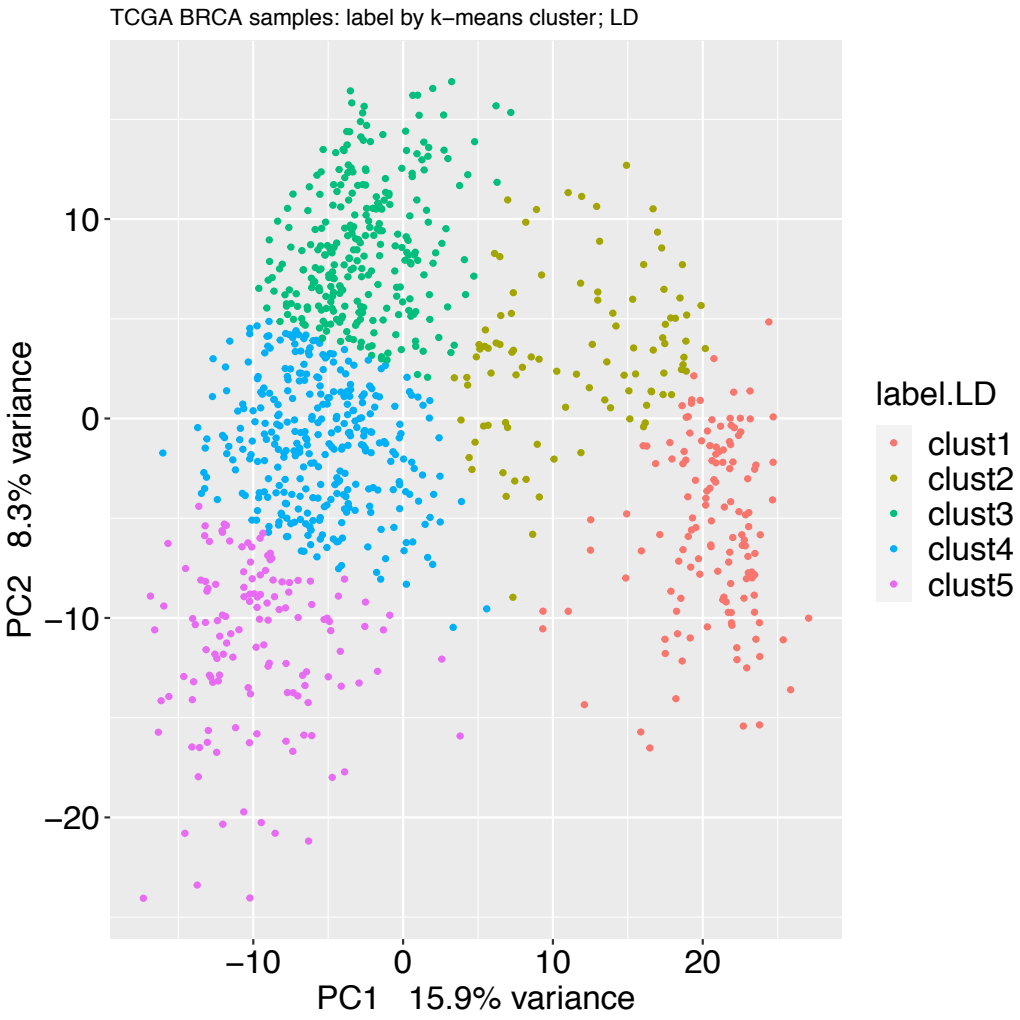
MD: mid dimension

PCA: Label by Subtype vs. by k-means Clusters

Label by subtype



Label by k-means clusters with 2 PCs



LD: low dimension

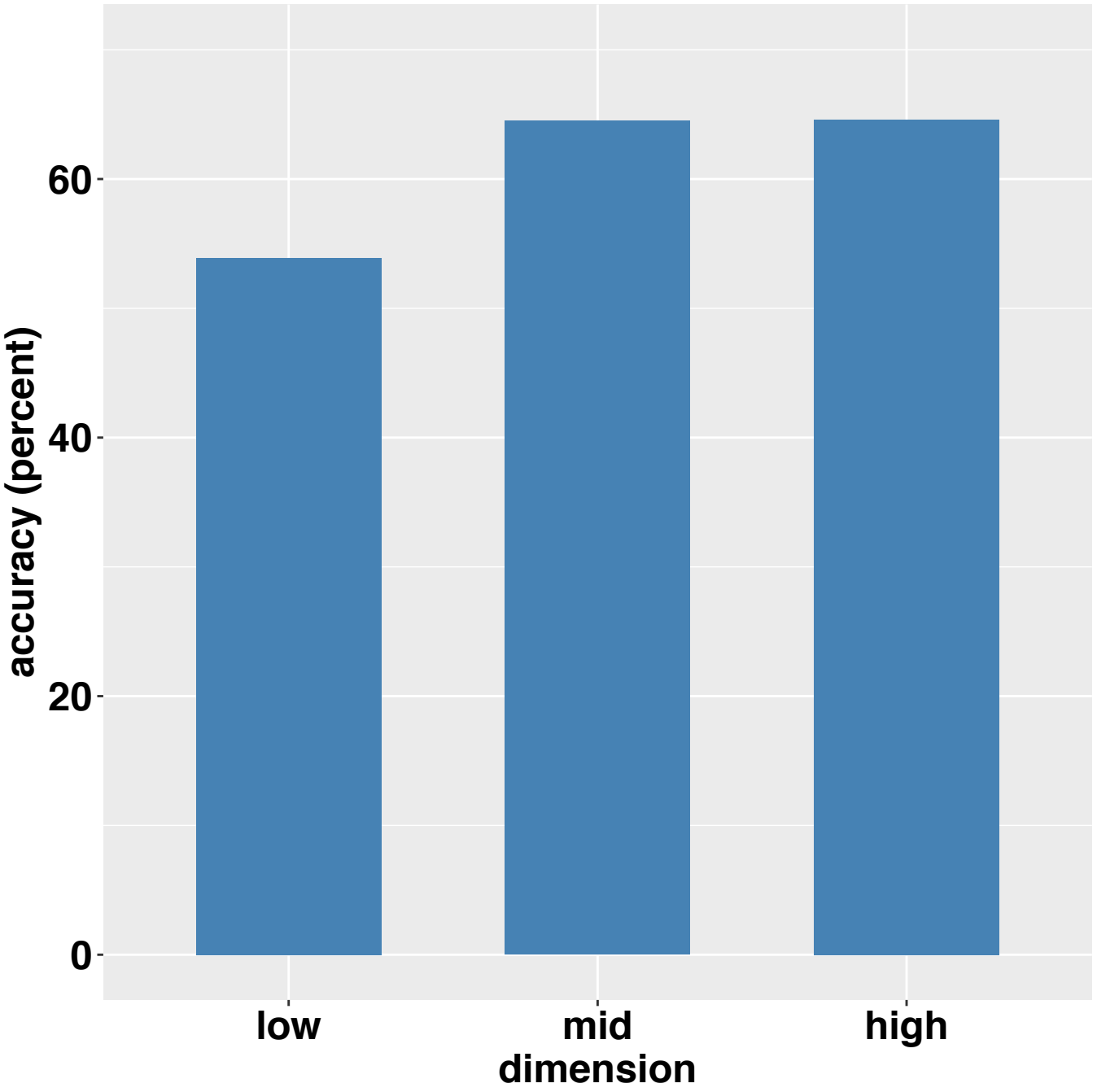
Comparison Between Subtype and Cluster (2 PCs)

Confusion matrix
Column: actual category
Row: assigned category

	Basal	Her2	LumA	LumB	Normal
clust1	133	2	0	0	0
clust2	40	37	2	2	17
clust3	0	8	247	8	19
clust4	0	24	190	110	2
clust5	0	2	61	73	0

Match: diagonal elements (red)
Mismatch: off diagonal elements (green)
Accuracy = (133 + 37 + 247 + 110) / 977 = 54%

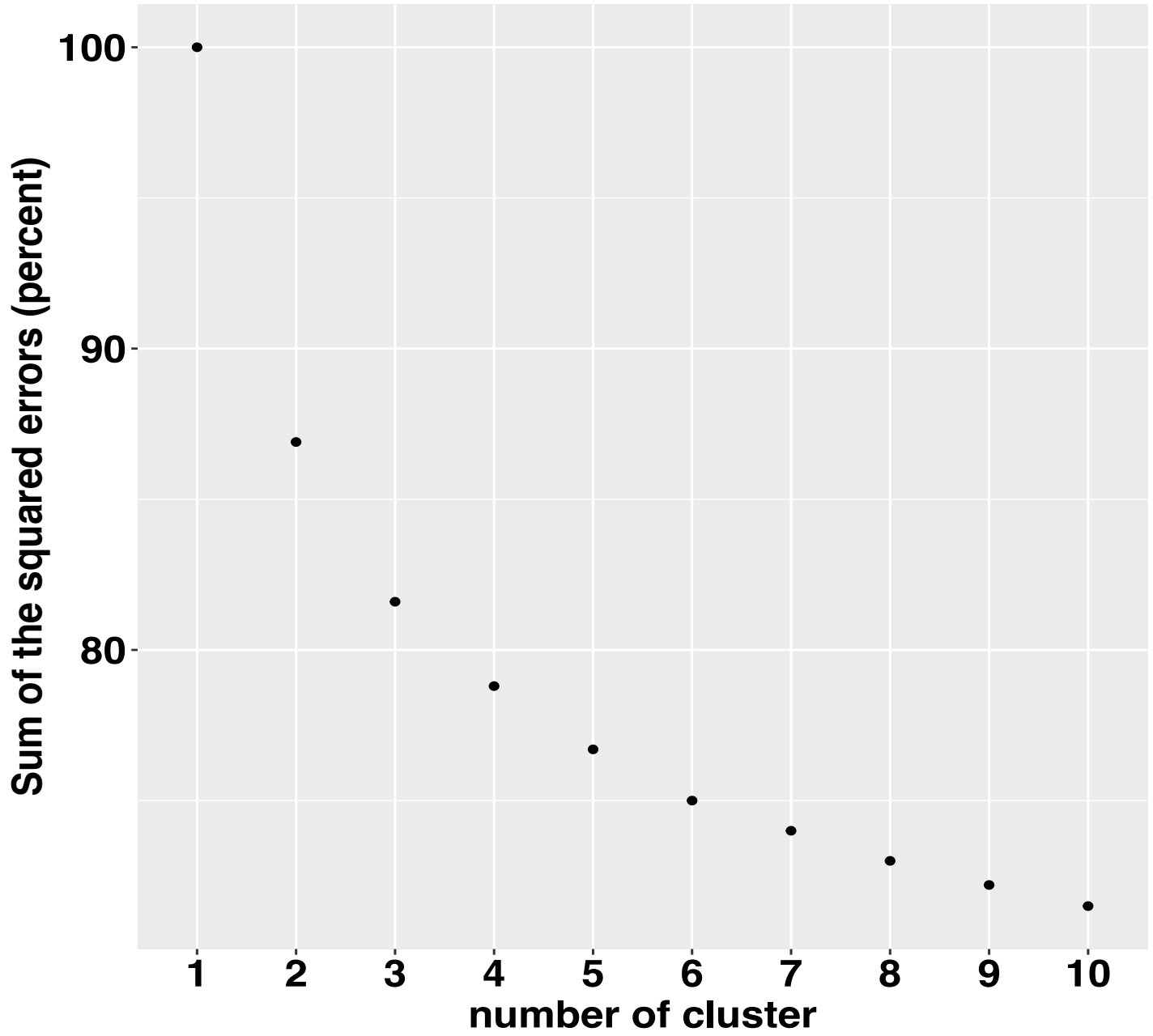
Accuracy of k-means Clustering Determined with Different Dimension



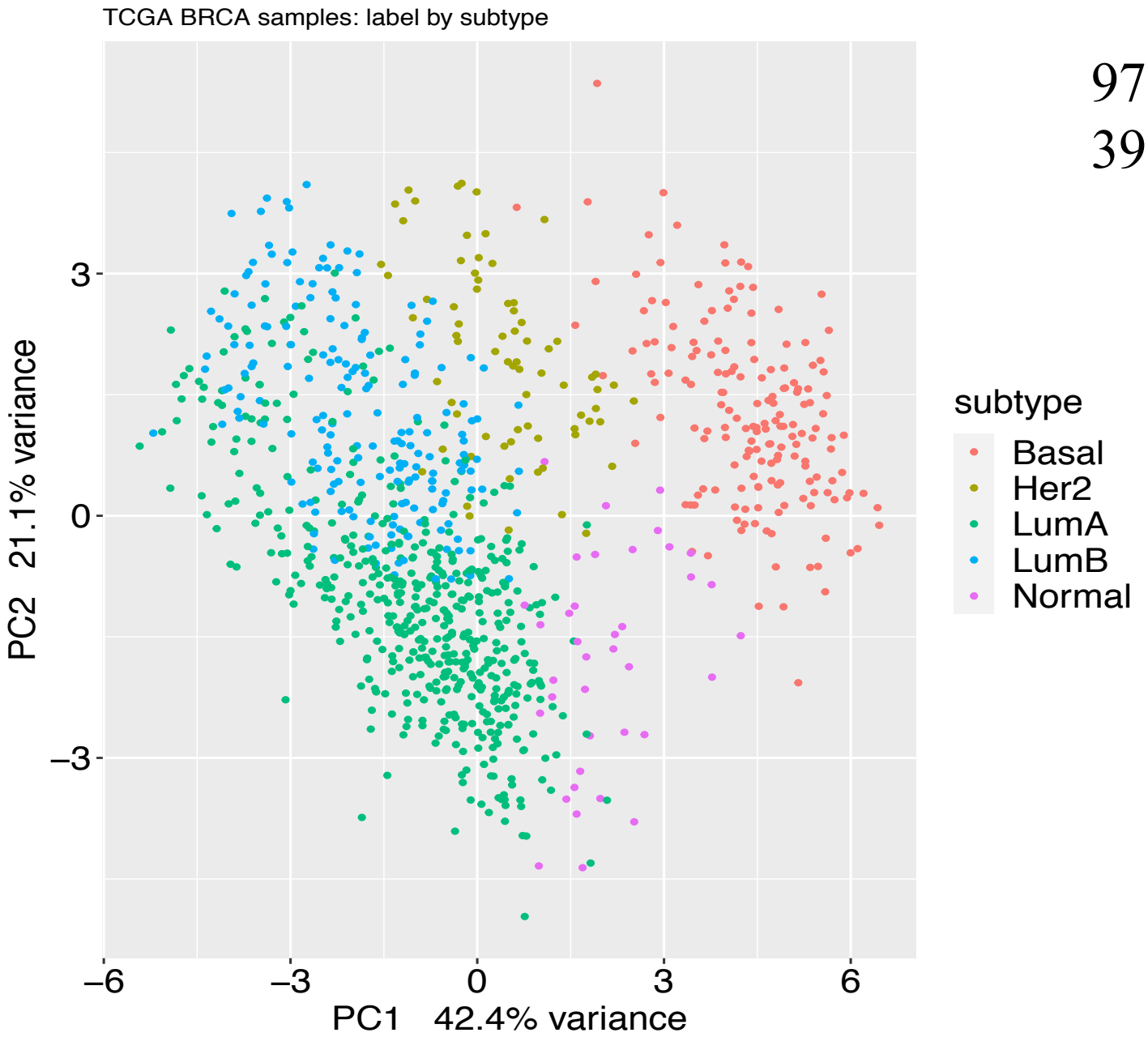
low: 2 PCs
mid: 10 PCs
high: 5000 genes

Three Dimensional Object and Its Two dimensional Image: Image of Two Tennis Balls Shadow Partially Overlaps

Sum of the Squared Errors of k-means Clustering

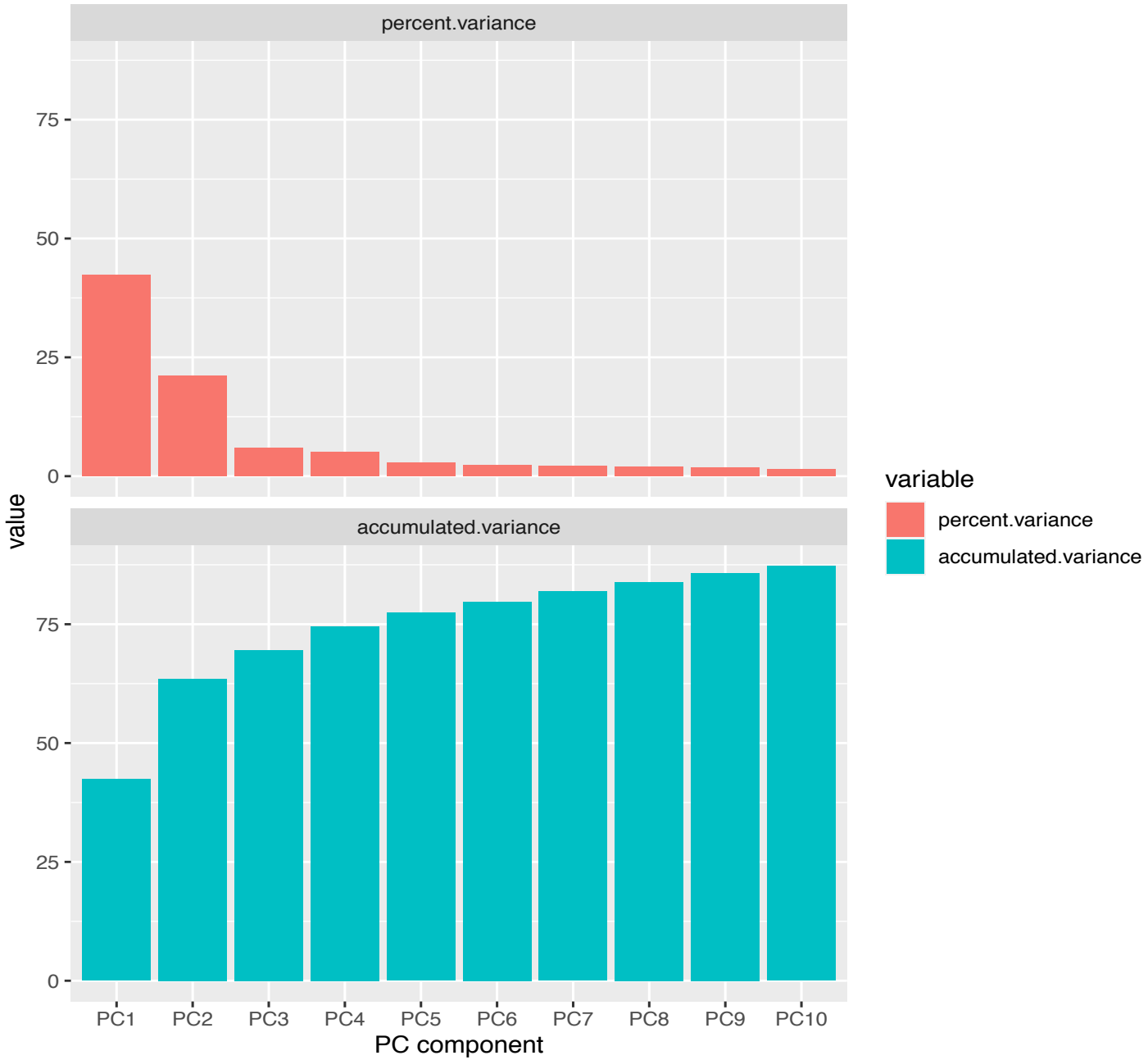


PCA of TCGA BRCA Samples with Pam50 Genes



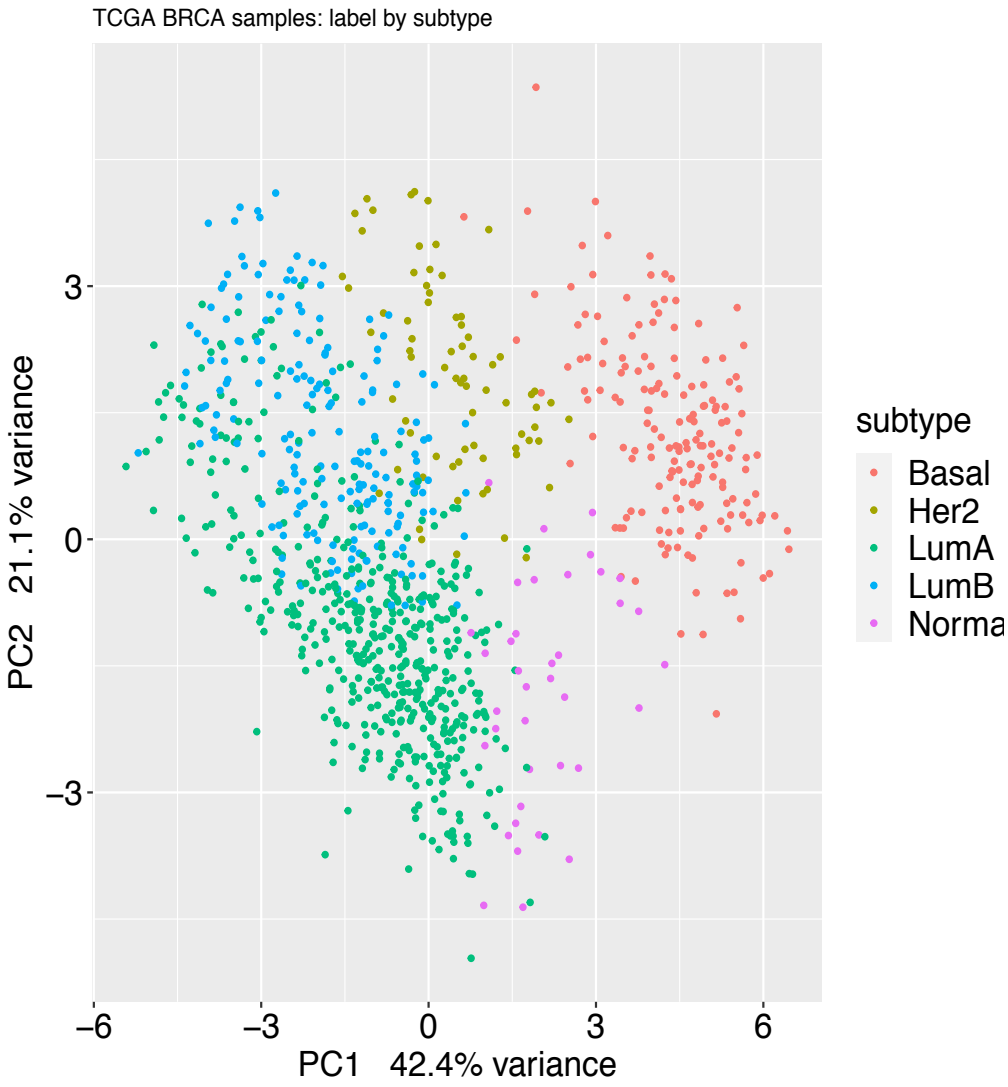
977 samples
39 genes

Percent of Variance Accounted for by PCA Components

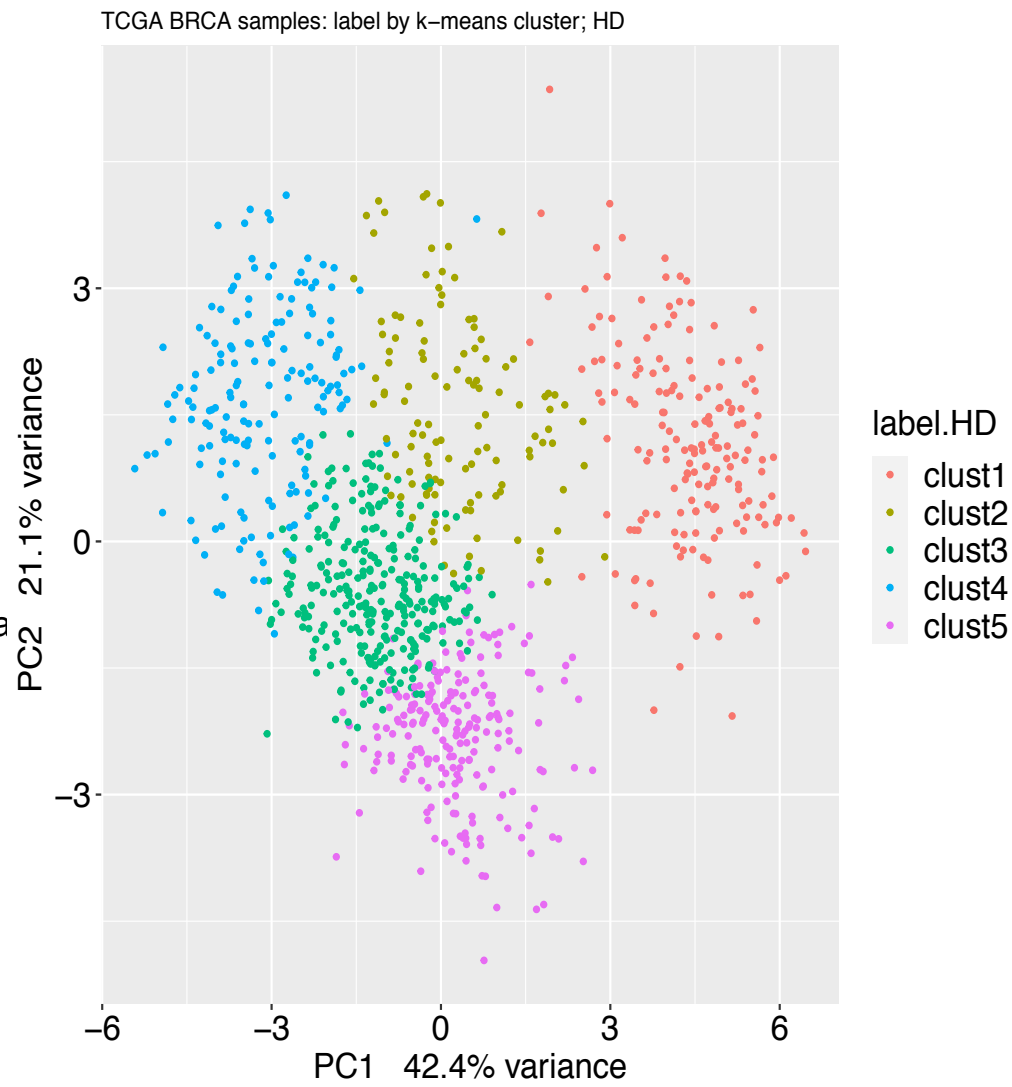


PCA: Label by Subtype vs. by k-means Clusters

Label by subtype



Label by k-means clusters in high-dimension



Comparison Between Subtype and Cluster (HD)

Confusion matrix
Column: actual category
Row: assigned category

	Basal	Her2	LumA	LumB	Normal
clust1	170	0	0	0	8
clust2	2	72	11	31	4
clust3	0	0	214	76	0
clust4	1	1	82	86	0
clust5	0	0	193	0	26

Match: diagonal elements (red)
Mismatch: off diagonal elements (green)
Accuracy = (170 + 72 + 221 + 87 + 26) / 977 = 59%

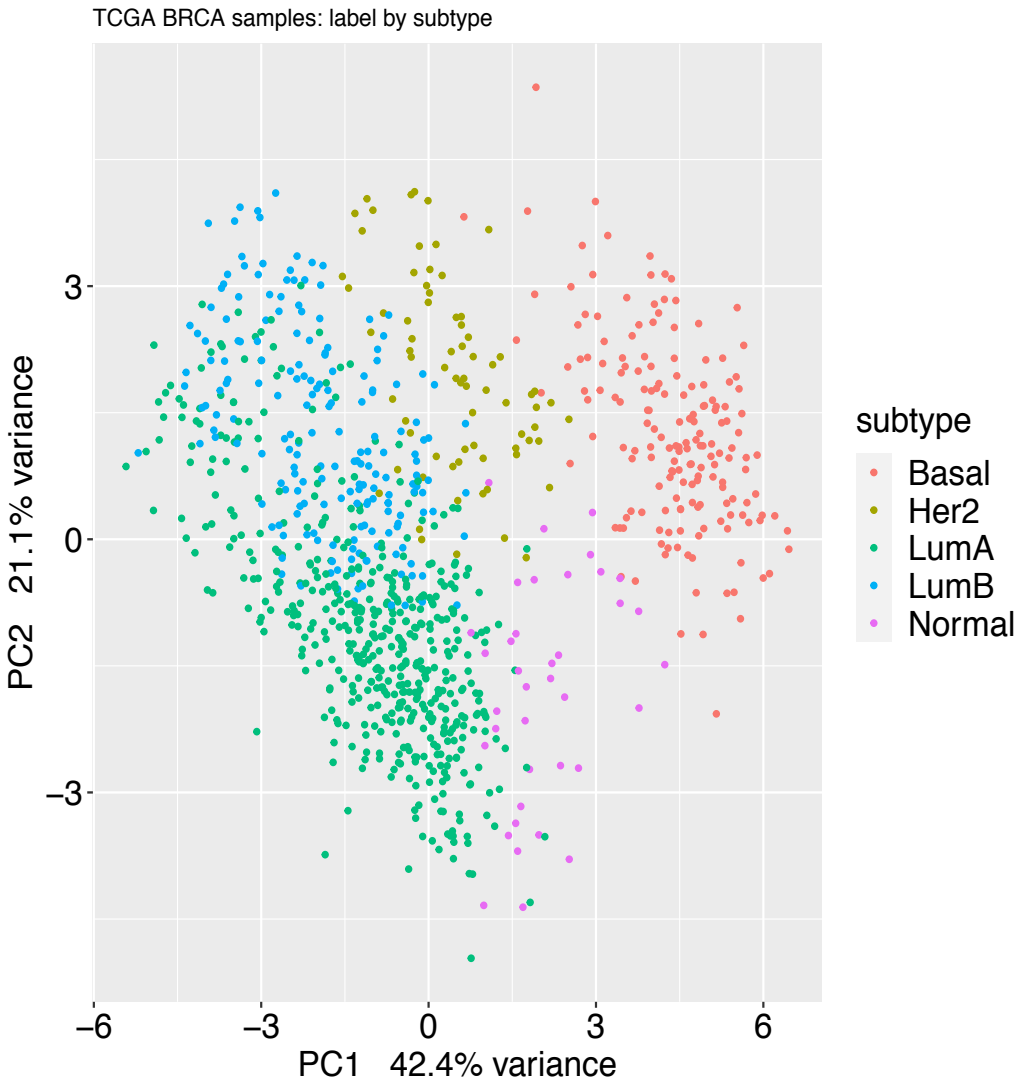
PCA: Label by Subtype vs. by k-means Clusters

Label by subtype

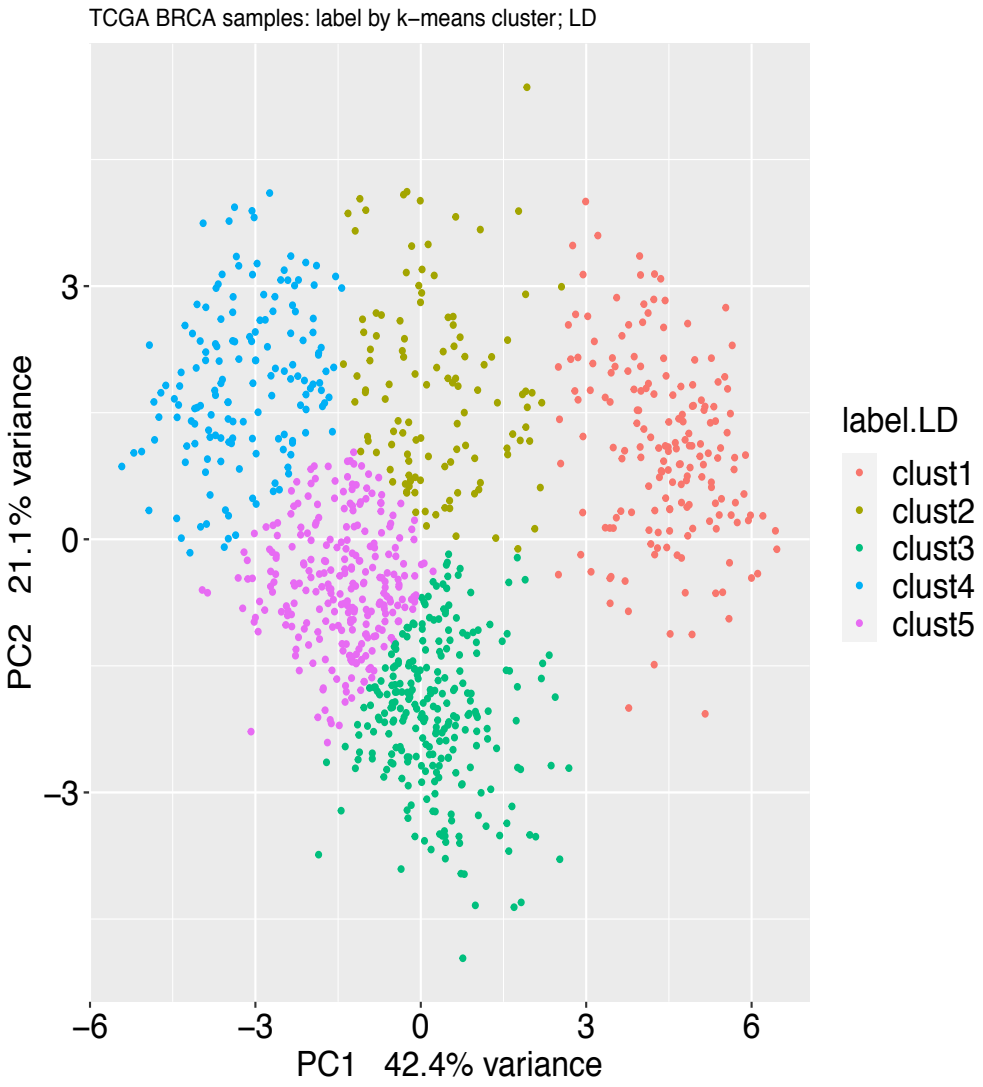
Label by k-means clusters with 10 PCs

PCA: Label by Subtype vs. by k-means Clusters

Label by subtype



Label by k-means clusters with 2 PCs



Comparison Between Subtype and Cluster (2 PCs)

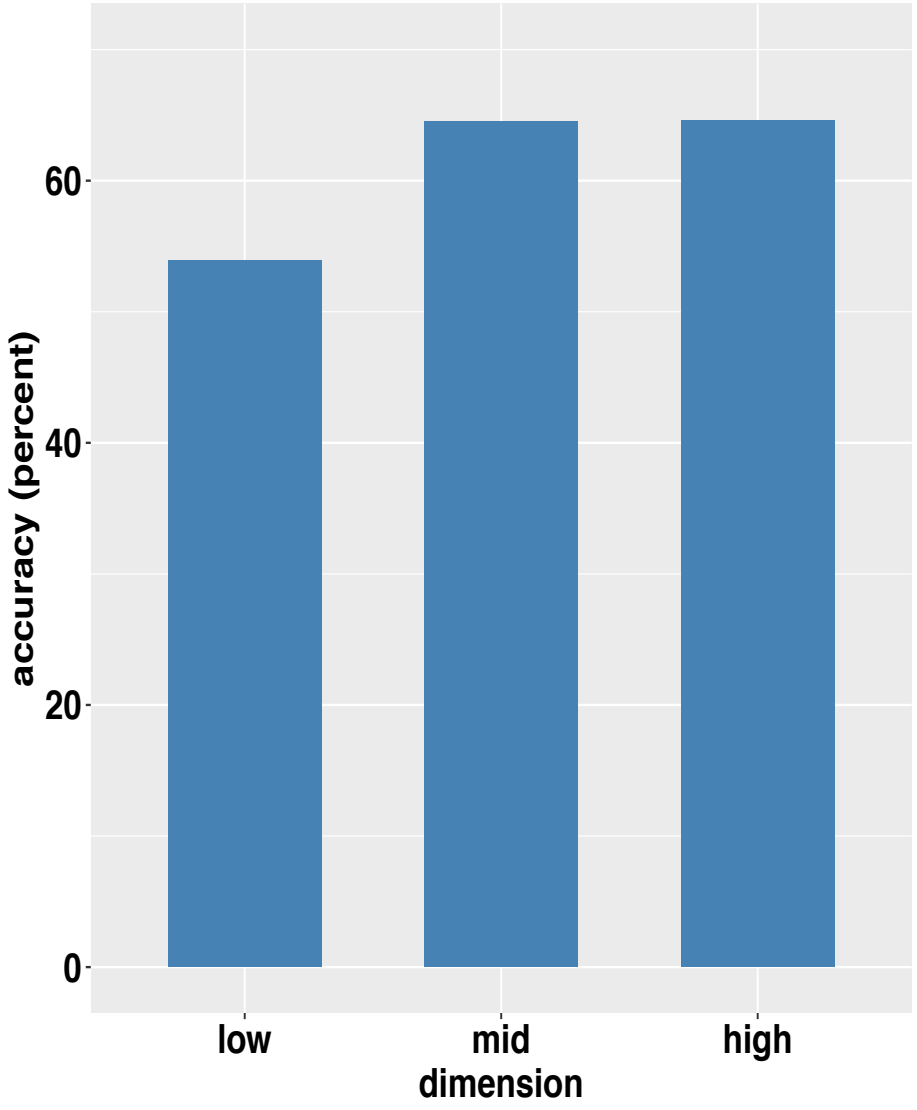
Confusion matrix
Column: actual category
Row: assigned category

	Basal	Her2	LumA	LumB	Normal
clust1	166	1	0	0	9
clust2	7	65	11	33	2
clust3	0	2	227	1	27
clust4	0	2	66	87	0
clust5	0	3	196	72	0

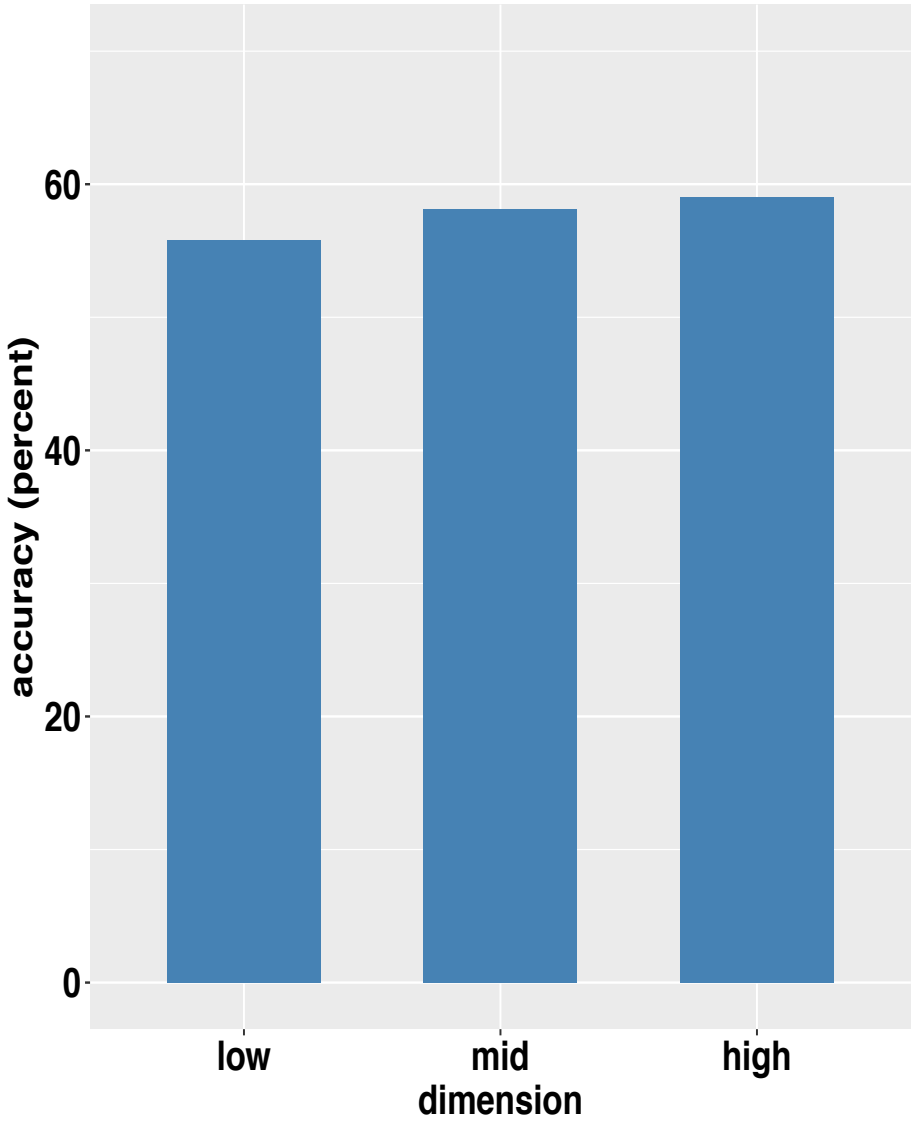
Match: diagonal elements (red)
Mismatch: off diagonal elements (green)
Accuracy = (166 + 65 + 227 + 87) / 977 = 56%

Accuracy of k-means Clustering Determined with Different Dimension

low: 2 PCs
mid: 10 PCs
high: 5000 genes



low: 2 PCs
mid: 10 PCs
high: 39 genes



K-means Clustering Algorithm

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where,

$x_i^{(j)}$ = data point

c_j = cluster center

n = Number of data points

k = Number of cluster

$||x_i^{(j)} - c_j||^2$ = distance between a data point $x_i^{(j)}$ and cluster centre c_j.

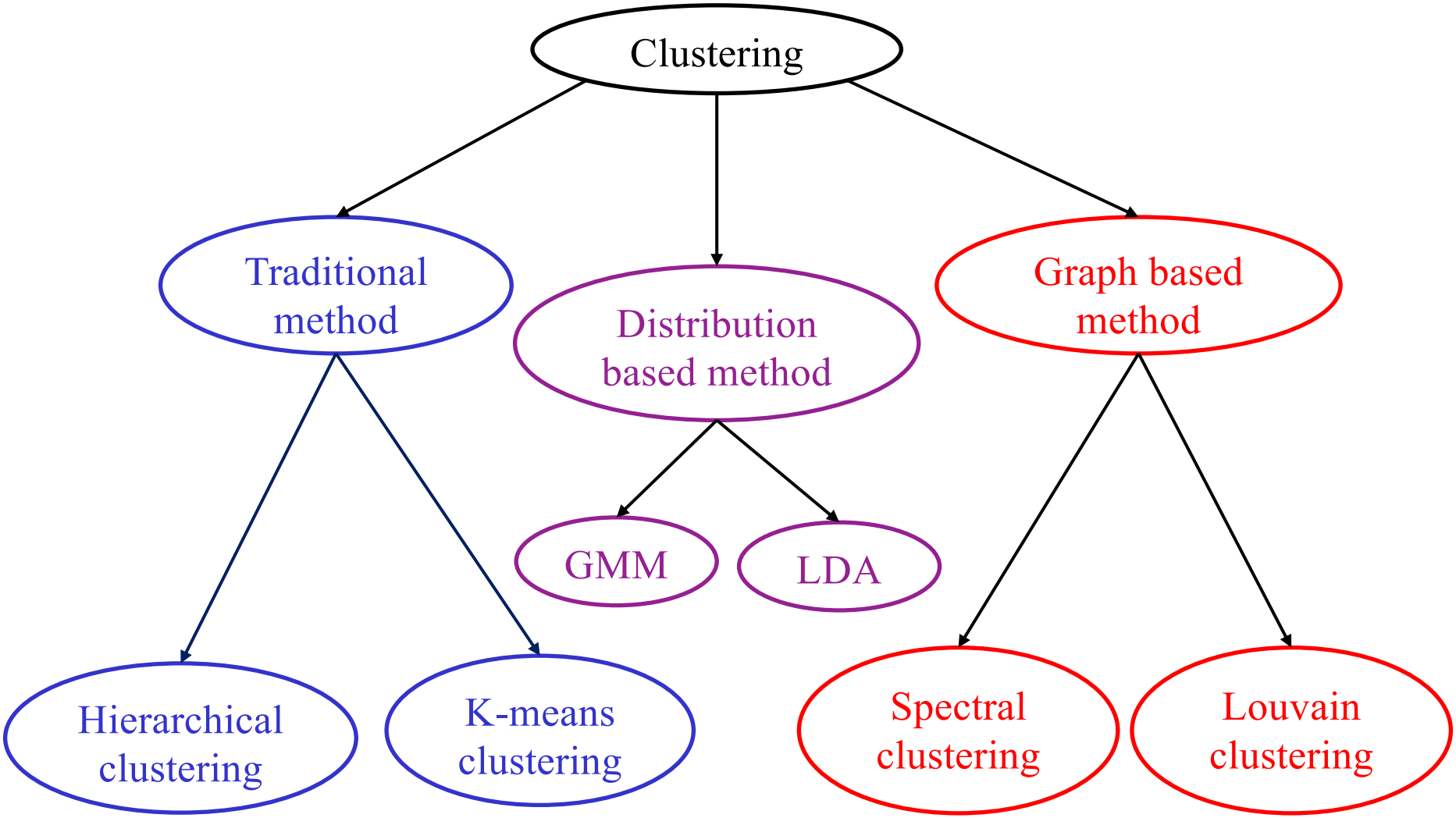
Initialization: randomly select k data points as k centroids

Alternating between two steps until converge

Assignment step: assign each data point to its nearest centroid

Centroid update step: update centroid based on assignment of data point

Outline of Clustering Methods



GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation

Contributed by Emily Tai