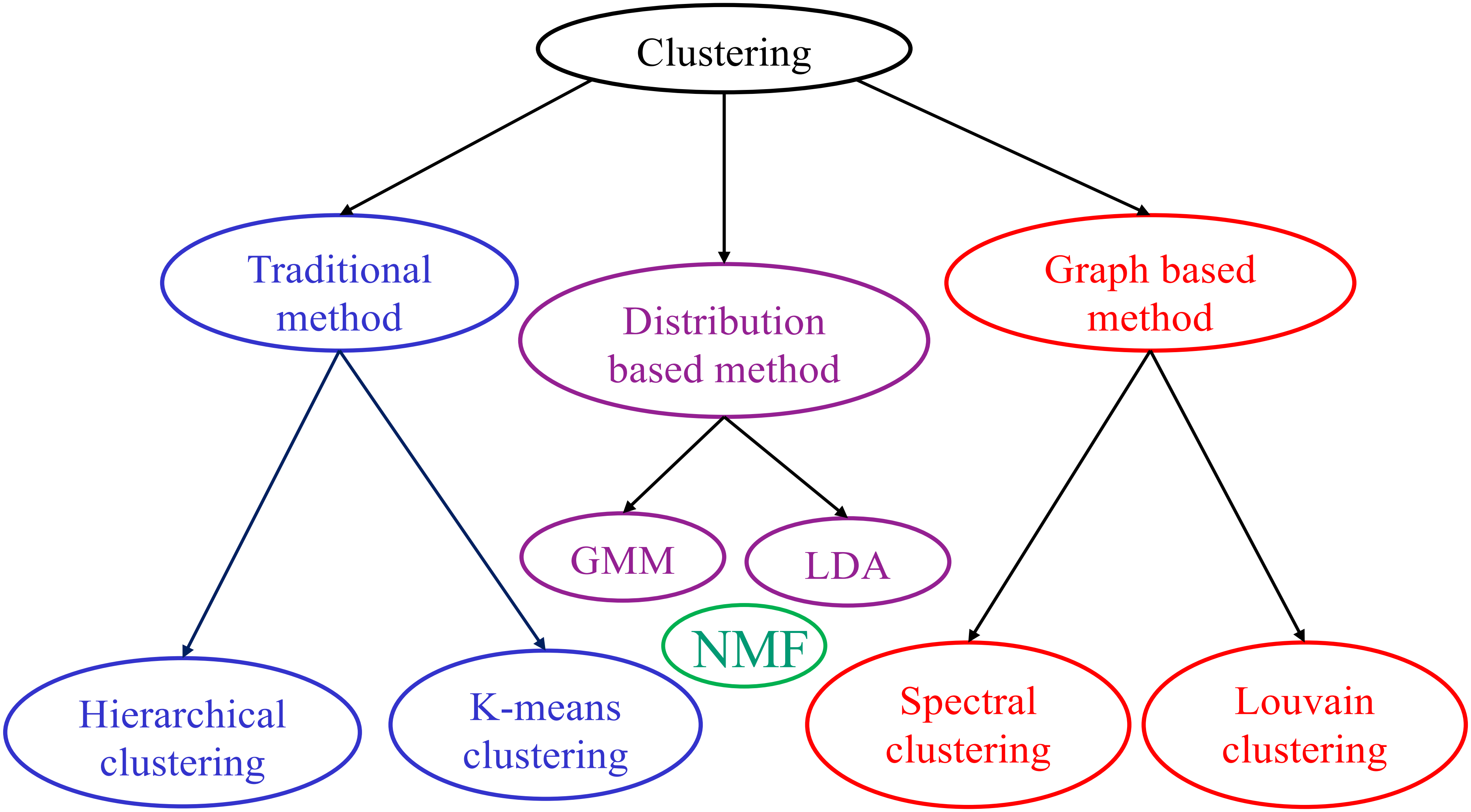
# Clustering Methods:
# From k-means to Gaussian Mixture Model and Louvain Algorithm

Maxwell Lee

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

October 19, 2020

# Outline of Clustering Methods

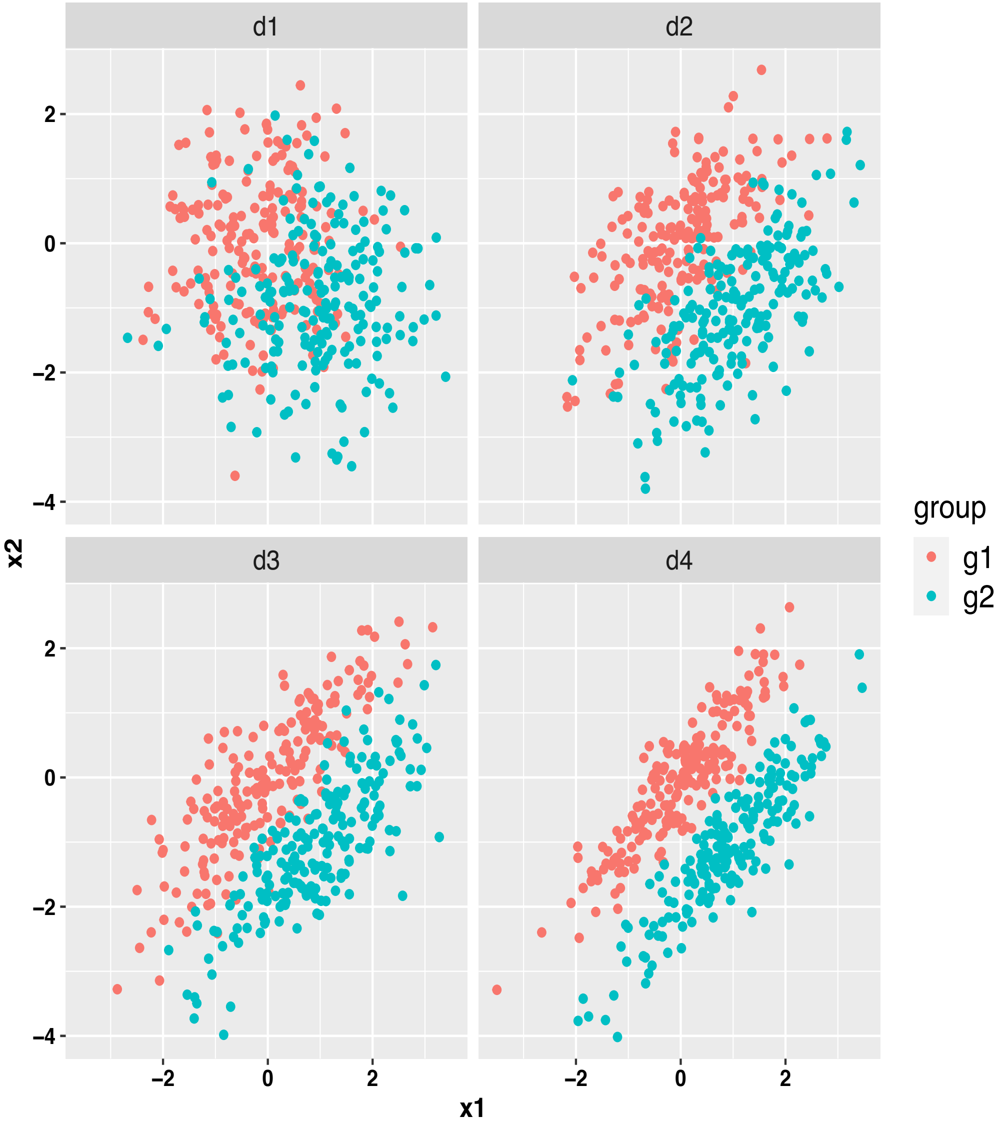


GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

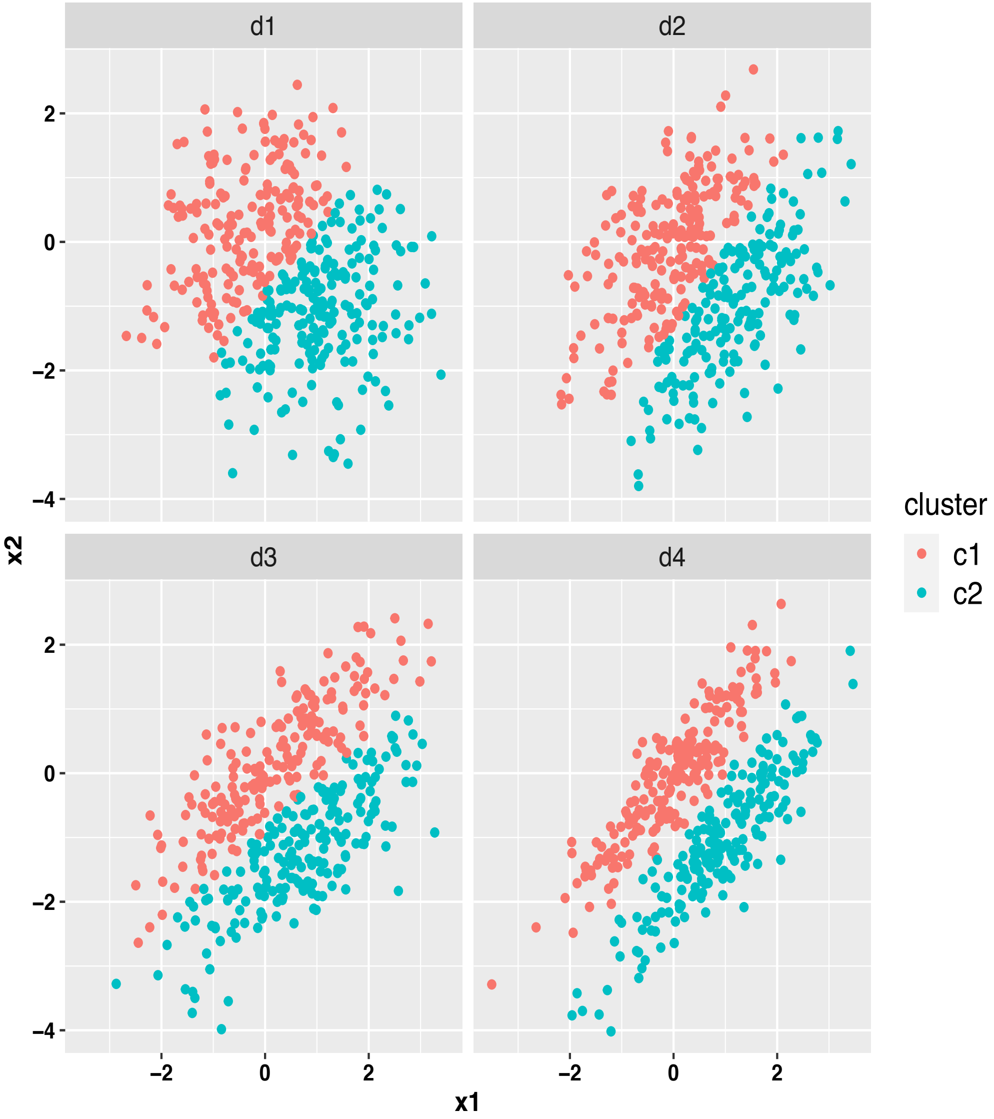
Contributed by Emily Tai

# Effect of Covariance Structure on GMM Clustering

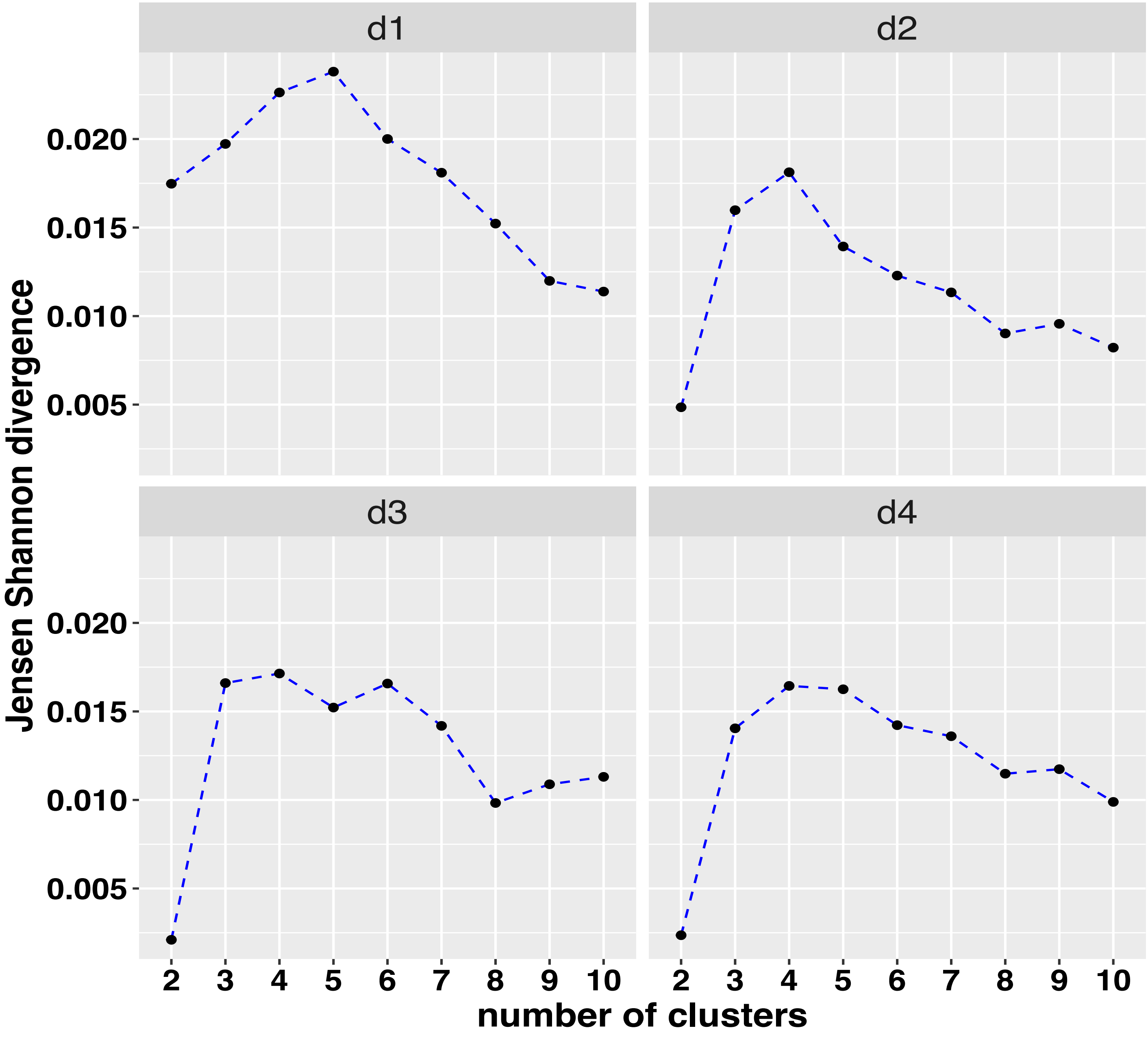
## Label by group



## Label by GMM cluster

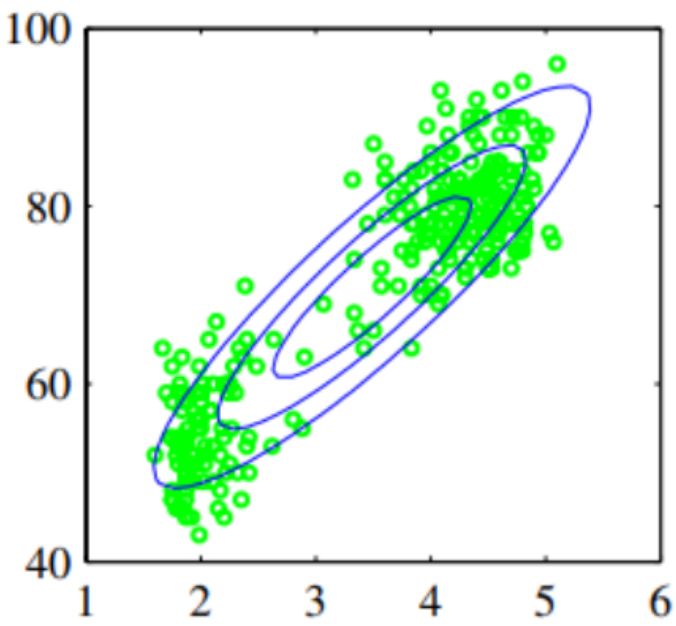


# Choose the Number of Clusters with Jensen-Shannon Divergence

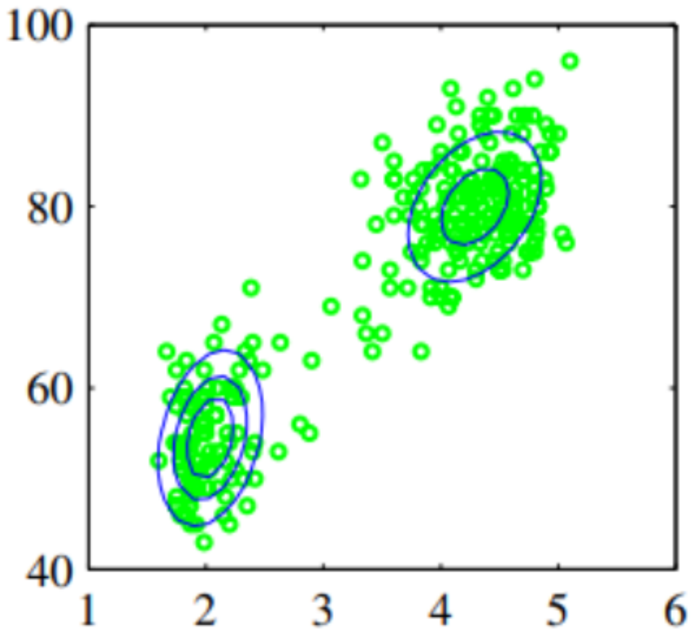


# Mixture of Bivariate Gaussian Distributions

Single Gaussian



Mixture of two Gaussians



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

$$p(\mathbf{x}) = \sum_{k=1}^{K} \underset{\substack{\uparrow \\ \text{Mixing coefficient}}}{\pi_k} \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$

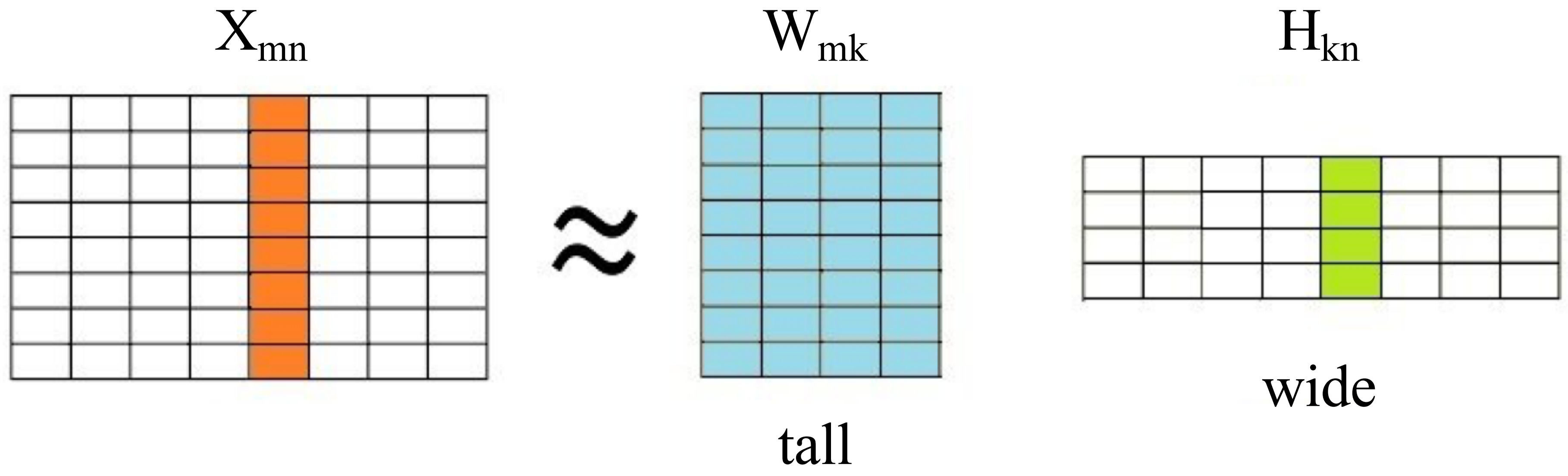
# Non-Negative Matrix Factorization (NMF)

NMF is a dimension reduction method. Why are we talking about it here?

Primary reason:
NMF is closely related to k-means and GMM. We can gain better understanding of clustering analysis through the lens of dimension reduction.

Secondary reason:
NMF is also closely related to Latent Dirichlet Allocation (LDA). It helps to understand LDA.

# Mathematic Model of Non-Negative Matrix Factorization



$X_{mn}$: m features; n samples
$W_{mk}$: m features; k latent variables
$H_{km}$: k latent variables; n encodings

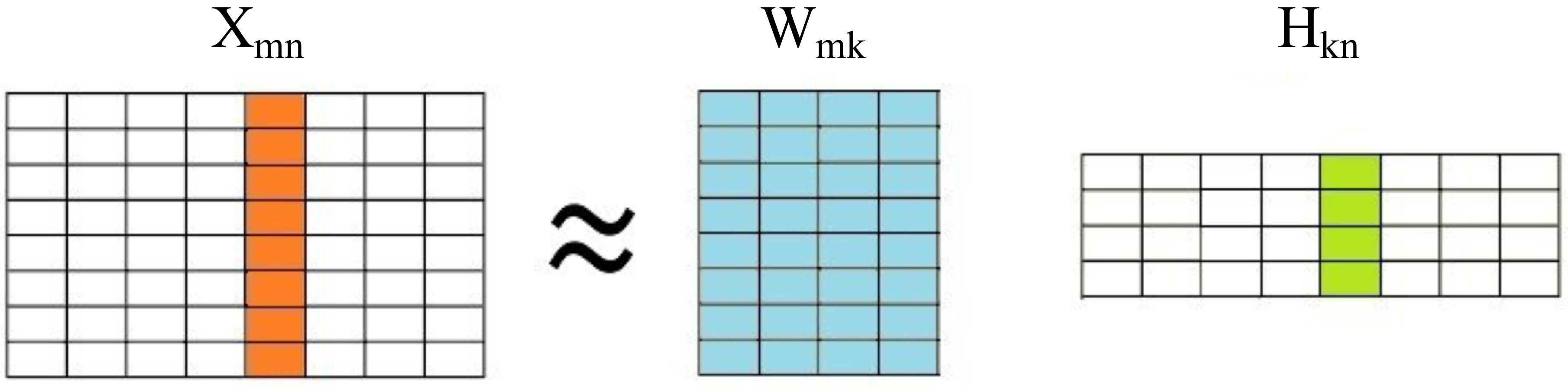
latent variables:
basis images
topics
centroids
signatures

Each element of matrix is non-negative
X >= 0; W >= 0; H >= 0

$k << \min(m,n) \longrightarrow$ Dimension reduction

Lee and Seung, Nature 1999; 401:788–791

# NMF is Related to PCA



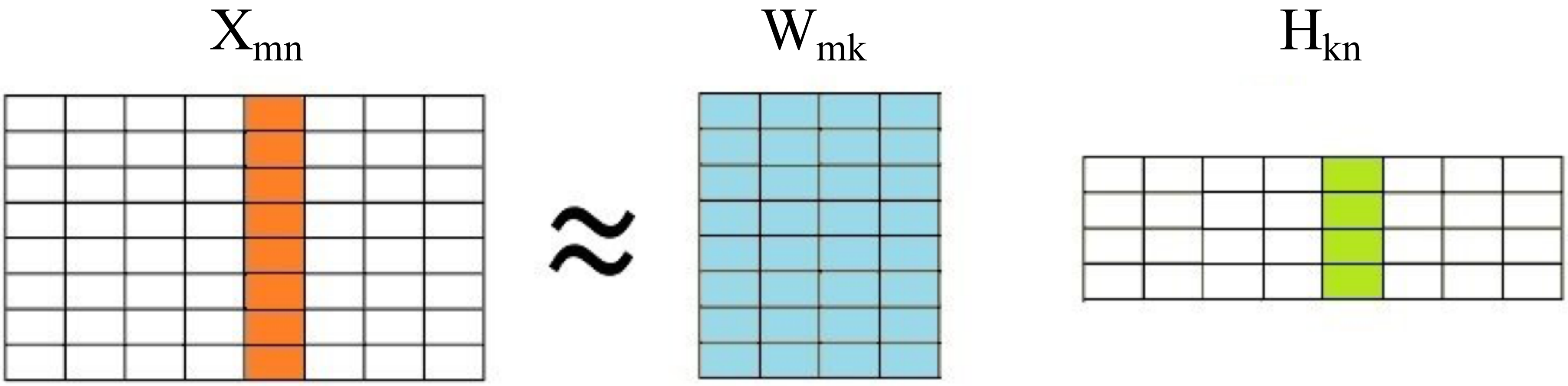
$$X \sim WH$$

PCA

$$XE = Z$$
$$X = ZE^T$$
$$X^T = EZ^T$$
$$X^T \sim E_{mk}Z^T$$

But PCA can have negative values!

# Understanding NMF from Topic Modeling (Mixture Model)



$X_{mn}$: m words; n documents
$W_{mk}$: m words; k topics
$H_{km}$: k topics; n documents

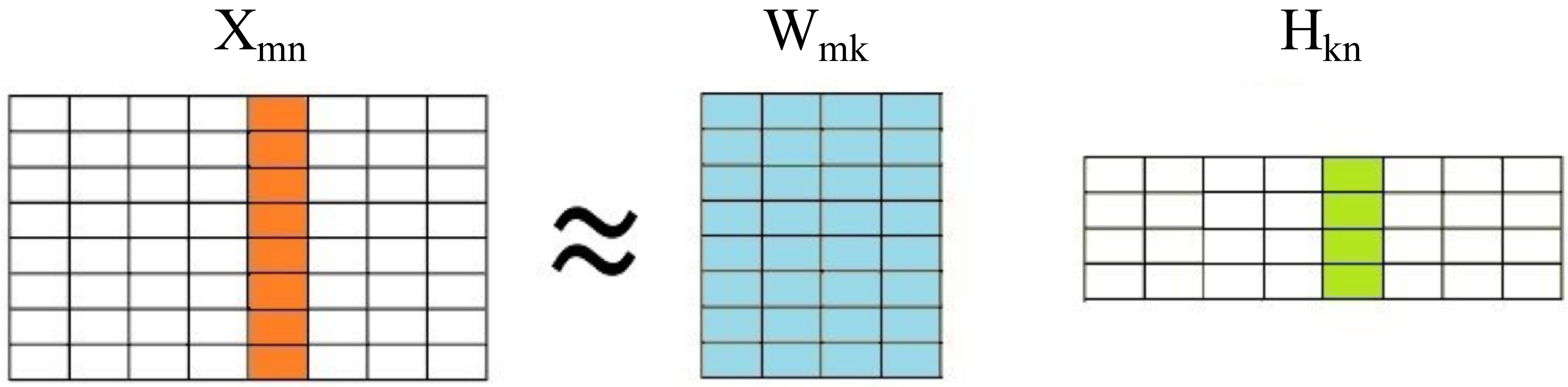
Topic modeling:
$\Sigma_i X_{ij} = 1$ (column sums to 1)
$\Sigma_i W_{ij} = 1$ and $\Sigma_i H_{ij} = 1$

$x_{ij} \sim \Sigma_k w_{ik} h_{kj}$

# Understanding NMF from Topic Modeling (Mixture Model)

I will talk about spectral clustering, which is a graph-based method and consists of dimension reduction with Laplacian Eigenmap and k-means clustering in the reduced dimension space. I will also talk about Louvain algorithm, which is used in Seurat package to cluster single cell RNAseq data. Louvain algorithm is a network community approach. It is very fast and has capacity to do clustering analysis for million nodes in a network. I will provide practical examples to illustrate how each method works and how to interpret the results of clustering analysis and explain the pros and cons of each method.

# Understanding NMF from Topic Modeling (Mixture Model)

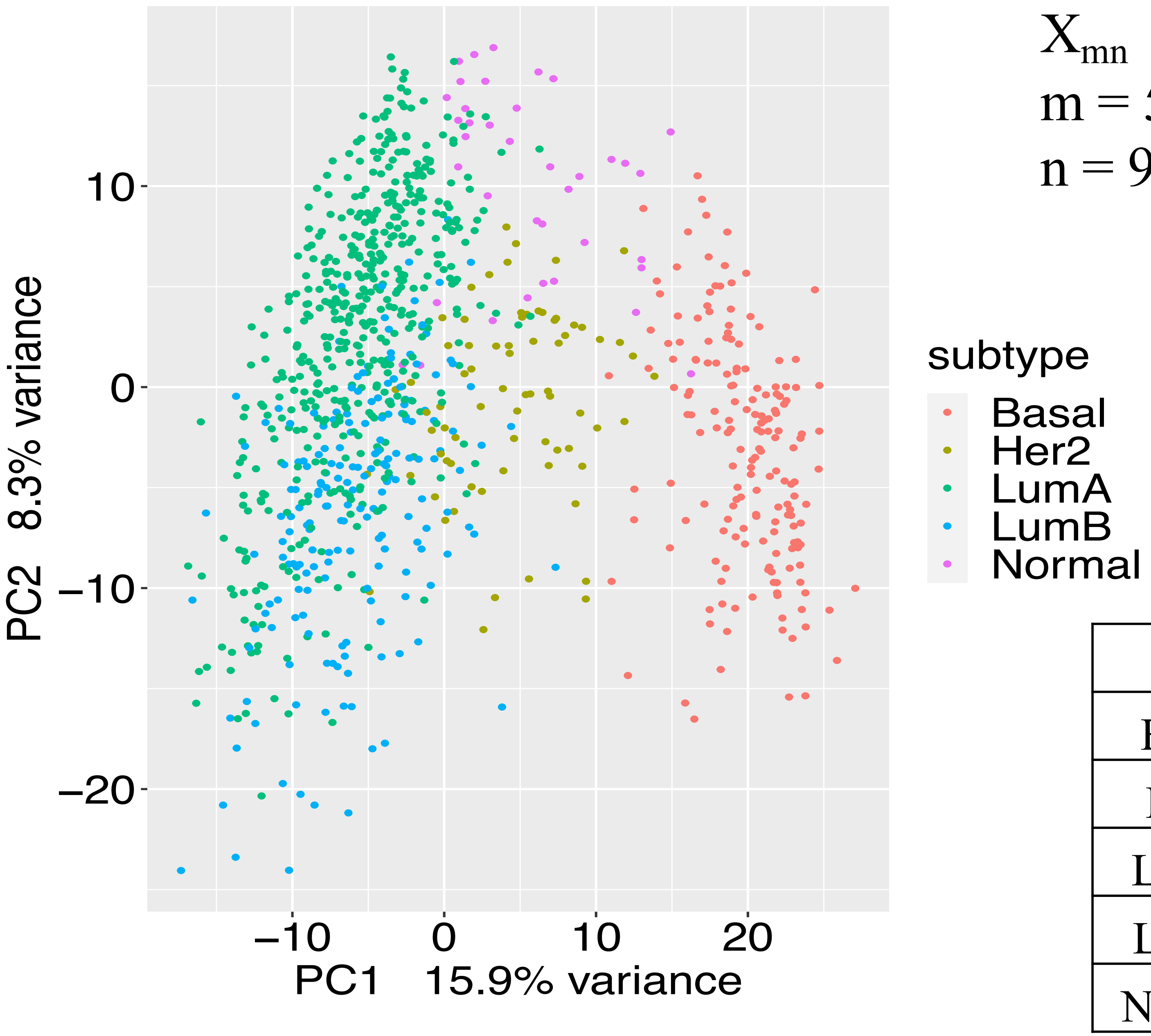


$$x_j \sim W\, h_j$$

Clustering:
label document with the topic having the highest frequency

nsNMF: nonsmooth NMF (sparse NMF)

# PCA of TCGA BRCA Samples



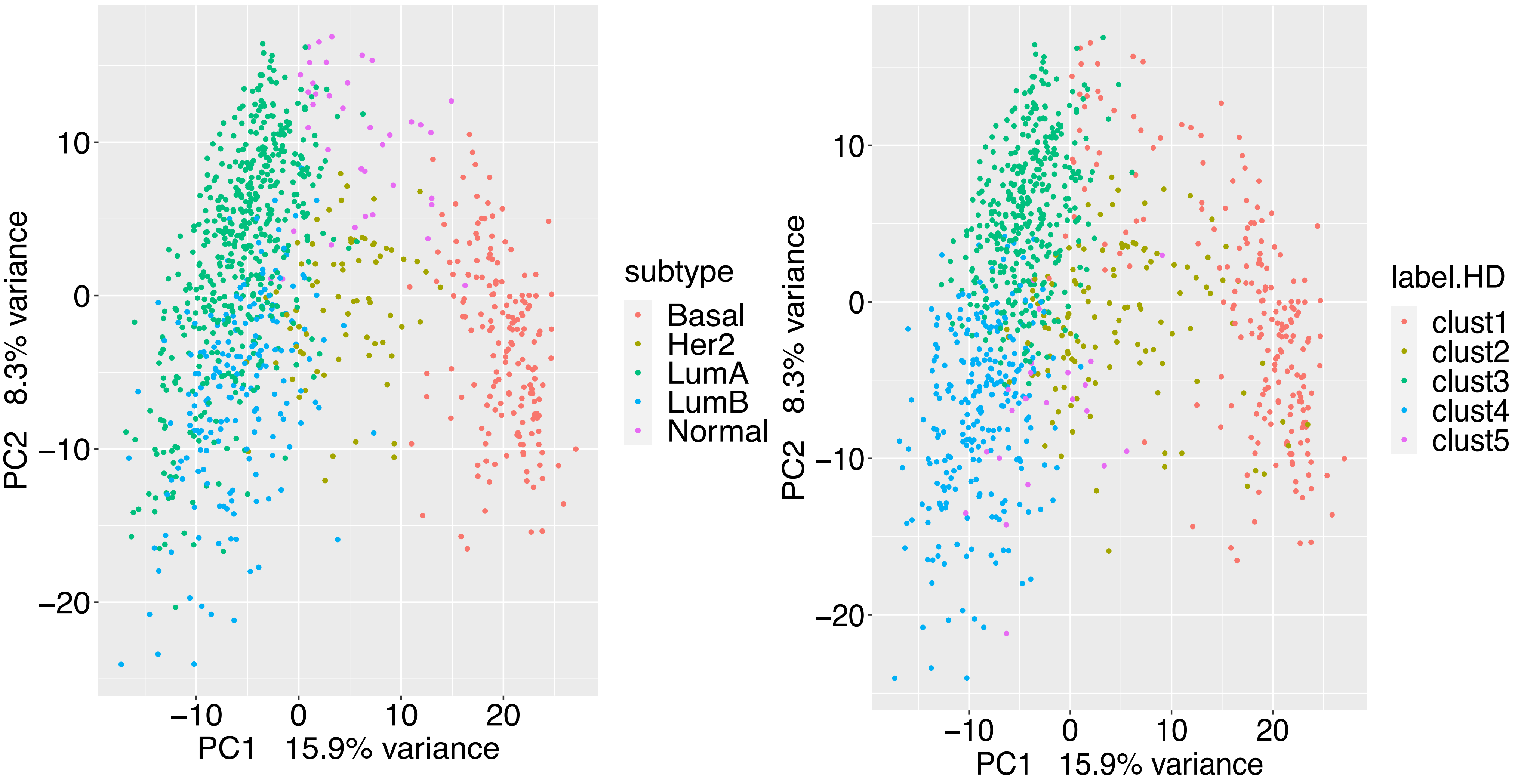
$X_{mn}$

m = 5000 genes

n = 977 samples

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

# PCA: Label by Subtype vs. by NMF Cluster



Label by subtype

Label by NMF clusters in high-dimension

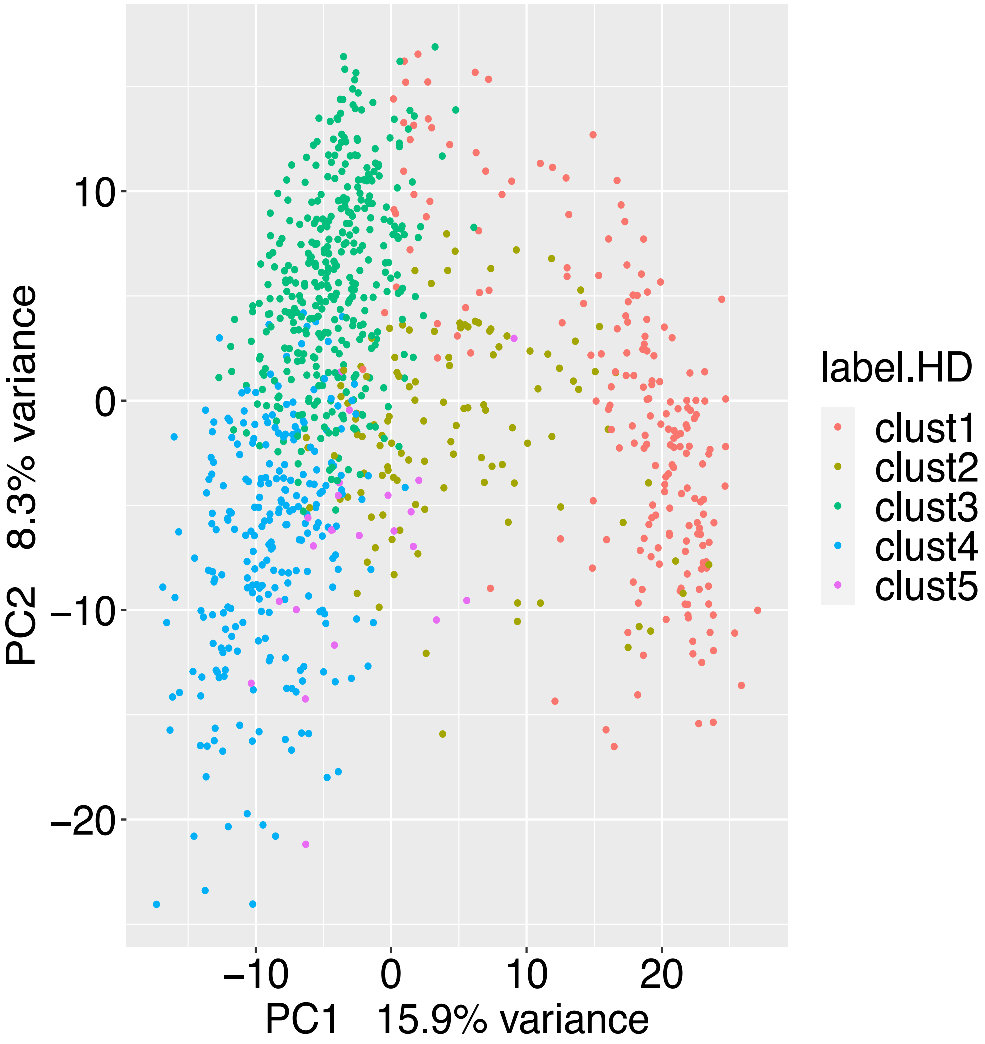
accuracy 68.5%

HD: high dimension, 5000 genes

# PCA: Label by NMF Cluster vs. by k-means Cluster

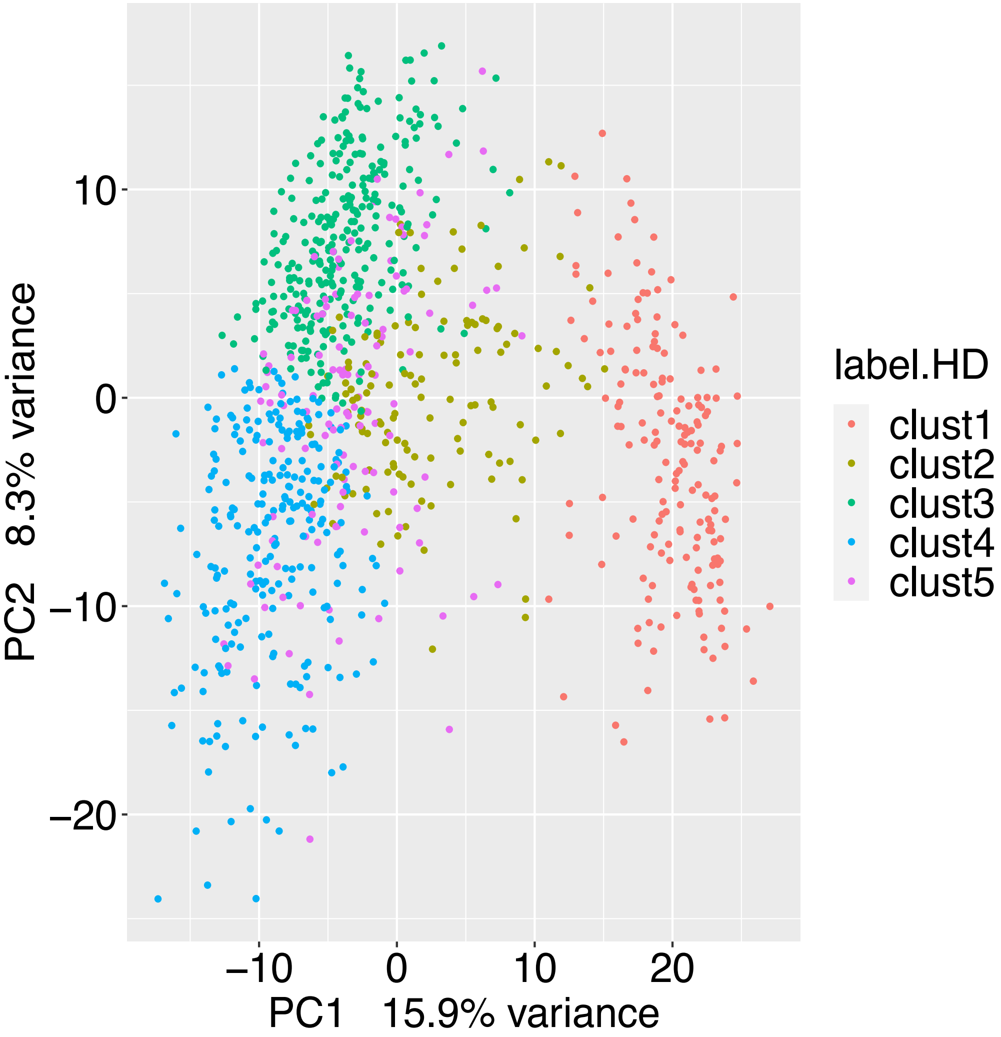
## Label by NMF clusters in high-dimension

accuracy 68.5%



## Label by k-means clusters in high-dimension

accuracy 65%



HD: high dimension, 5000 genes

# Comparison Between Subtype and NMF vs. k-means Cluster (HD)

NMF

| | Basal | Her2 | LumA | LumB | Normal |
|---|---|---|---|---|---|
| clust1 | 155 | 2 | 11 | 1 | **30** |
| clust2 | 18 | 62 | 7 | 32 | 2 |
| clust3 | 0 | 4 | **338** | 38 | 6 |
| clust4 | 0 | 2 | 132 | 114 | 0 |
| clust5 | 0 | 3 | 12 | 8 | 0 |

Accuracy = (155 + 62 + 338 + 114) / 977 = 68.5%

Match

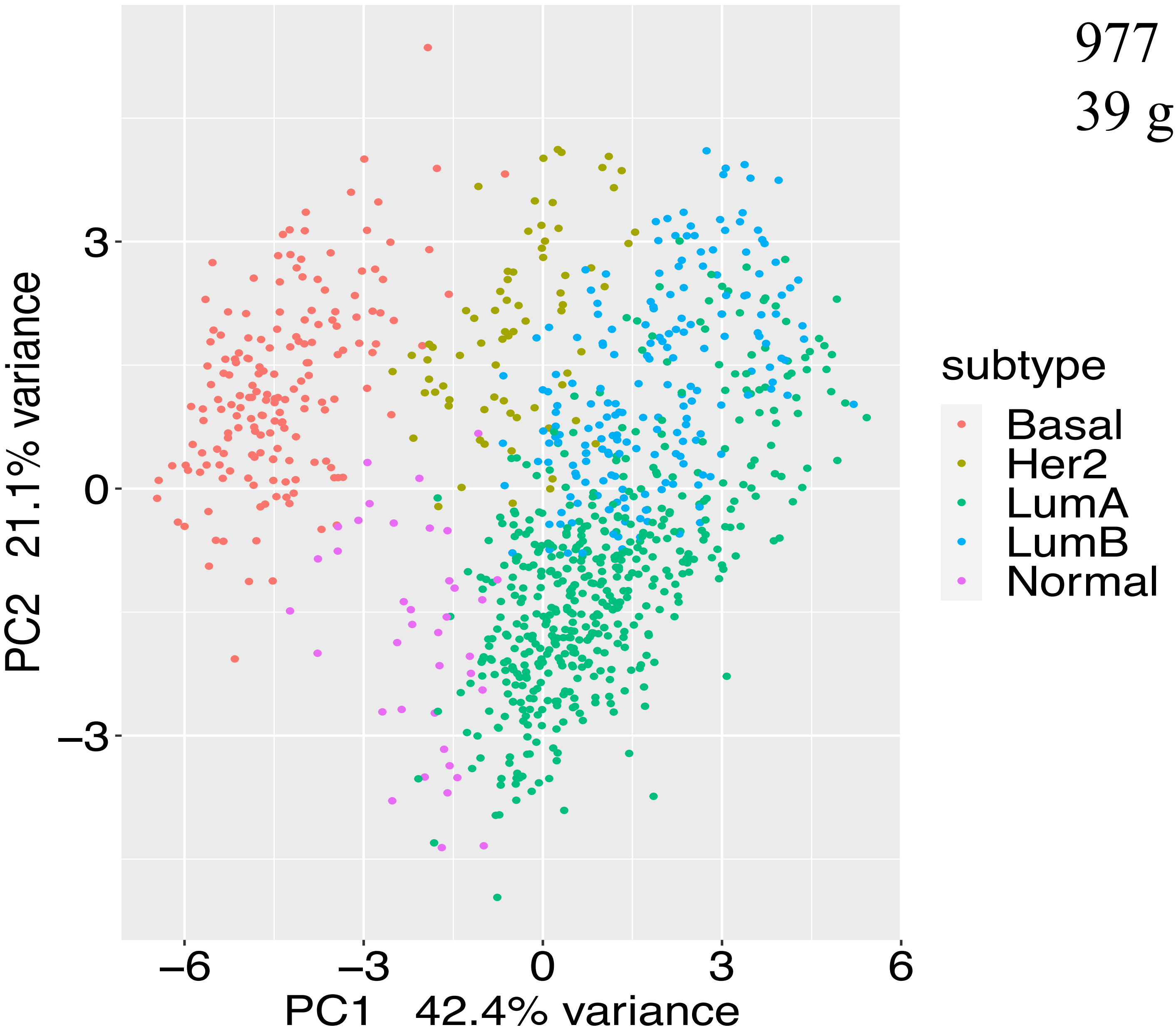
Mismatch

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 |

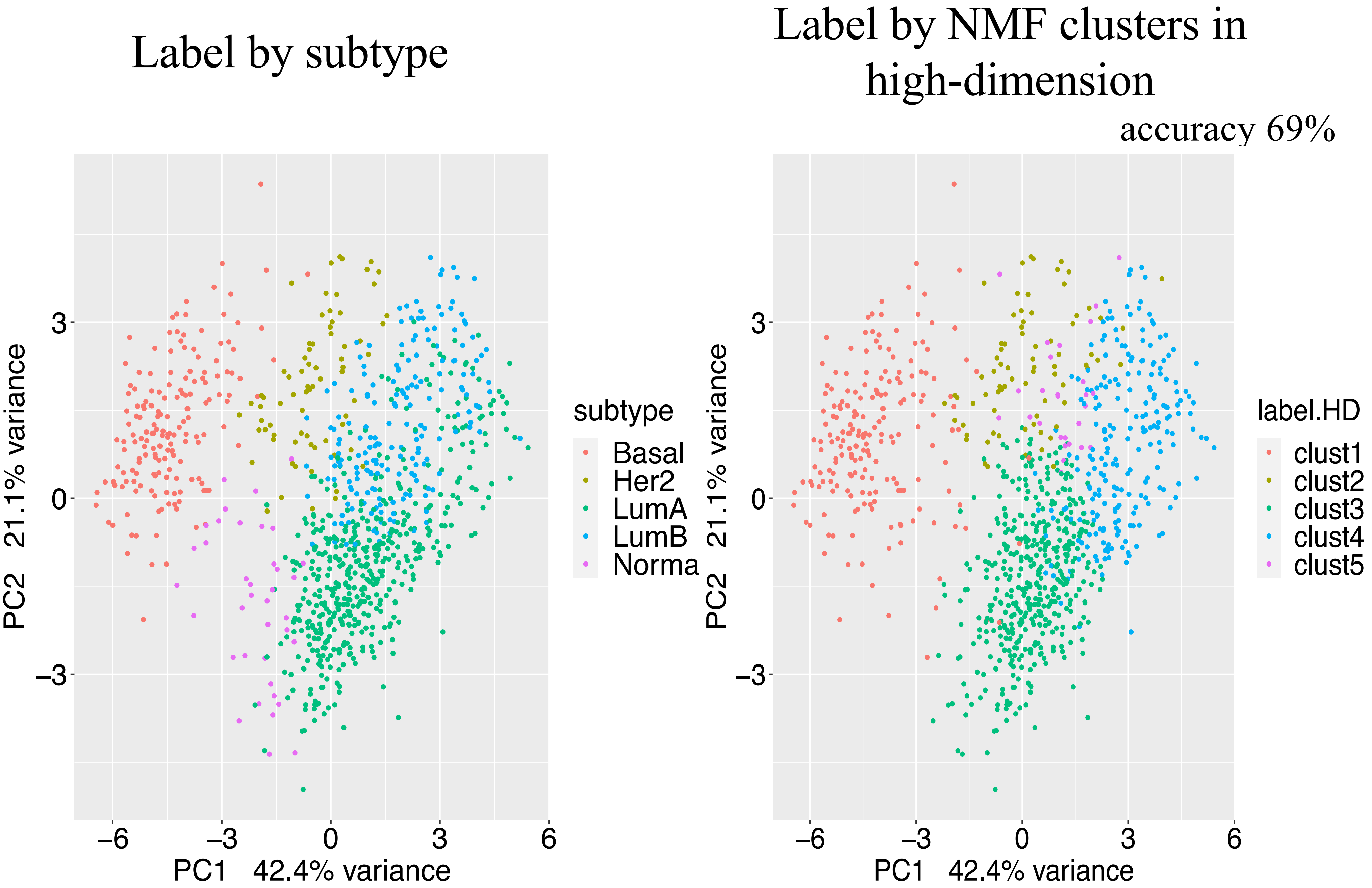
Accuracy = (169 + 69 + 268 + 119 + 6) / 977 = 64.6%

# PCA of TCGA BRCA Samples with Pam50 Genes



977 samples
39 genes

# PCA with pam50: Label by Subtype vs. by NMF Clusters

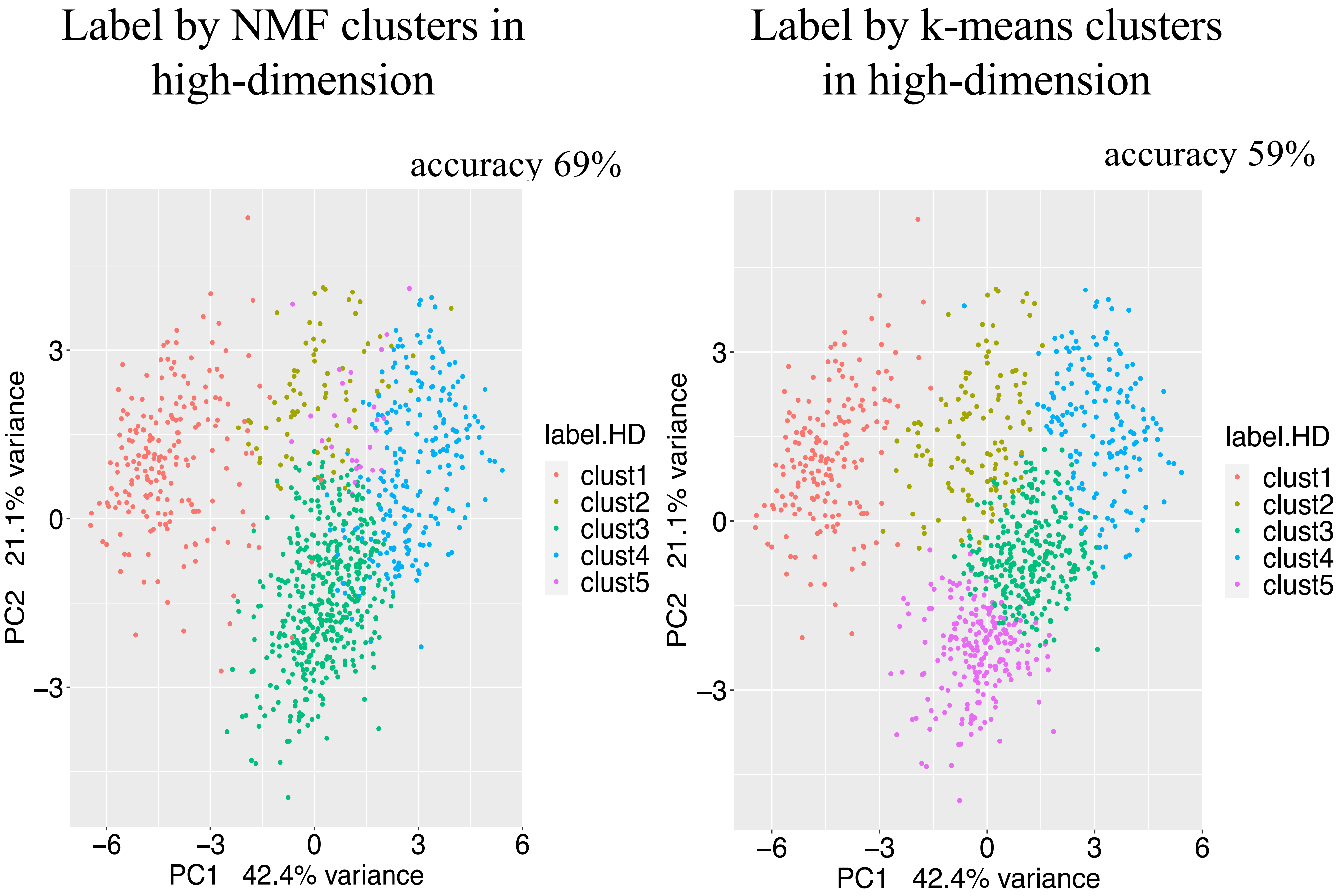


Label by subtype

Label by NMF clusters in high-dimension

accuracy 69%

# PCA with pam50: Label by NMF vs. k-means Clusters



Label by NMF clusters in high-dimension

accuracy 69%

Label by k-means clusters in high-dimension

accuracy 59%

# Comparison Between Subtype and NMF vs. k-means Cluster (HD)

NMF

| | Basal | Her2 | LumA | LumB | Normal |
|---|---|---|---|---|---|
| clust1 | 172 | 6 | 5 | 0 | 15 |
| clust2 | 0 | 58 | 3 | 15 | 0 |
| clust3 | 0 | 9 | **356** | 61 | 23 |
| clust4 | 0 | 0 | **136** | 91 | 0 |
| clust5 | 1 | 0 | 0 | **26** | 0 |

Accuracy = (172 + 58 + 356+ 91) / 977 = 69%

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%

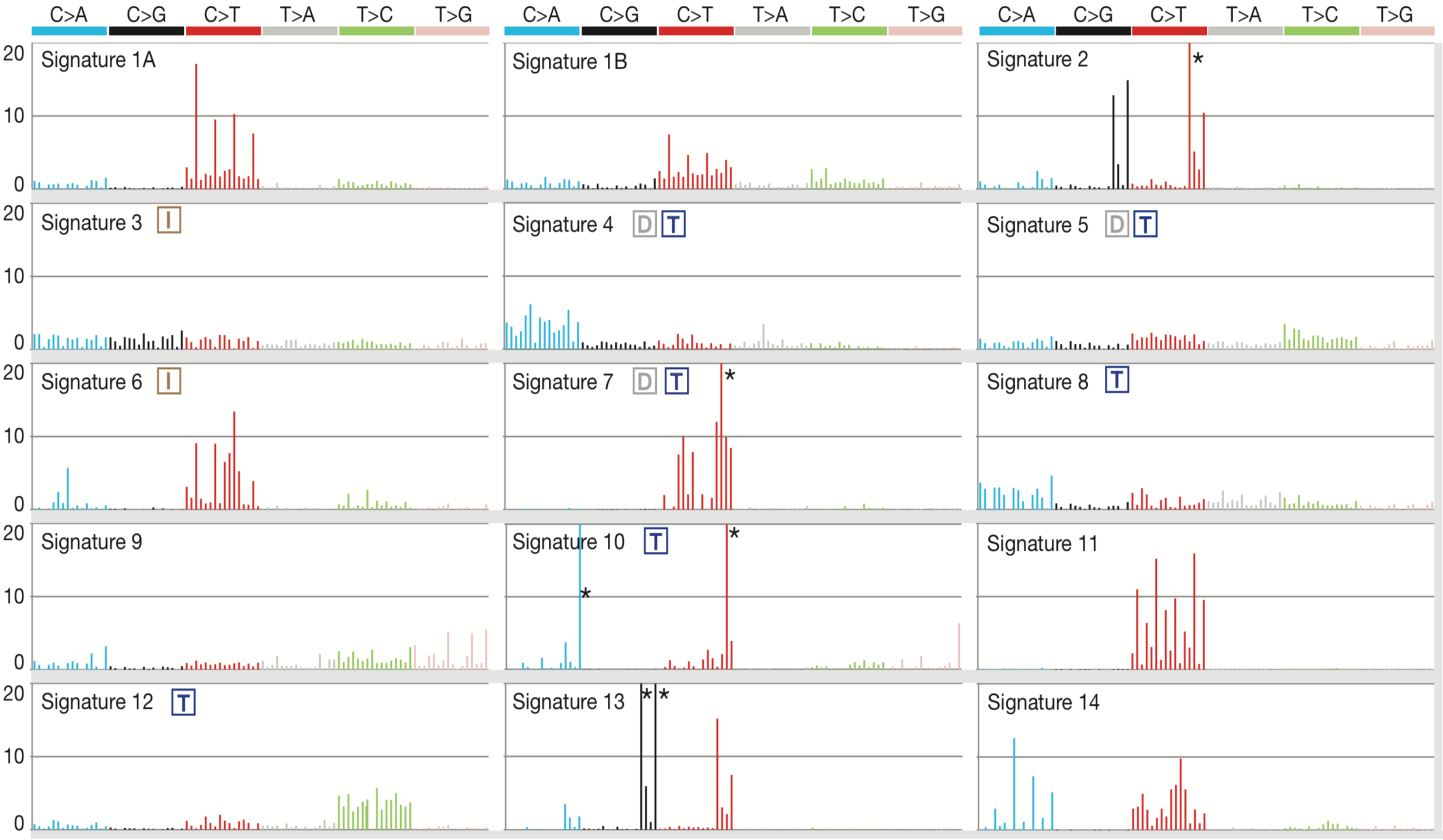
# Application of NMF to COSMIC Mutation Signature

| | A | C | G | T |
|---|---|---|---|---|
| A | 0 | ***1*** | **1** | **1** |
| C | **1** | 0 | **1** | **1** |
| G | 1 | 1 | 0 | 1 |
| T | 1 | 1 | 1 | 0 |

$N_1[A>C]_2N_3$

4*6*4=96

# COSMIC Mutation Signatures

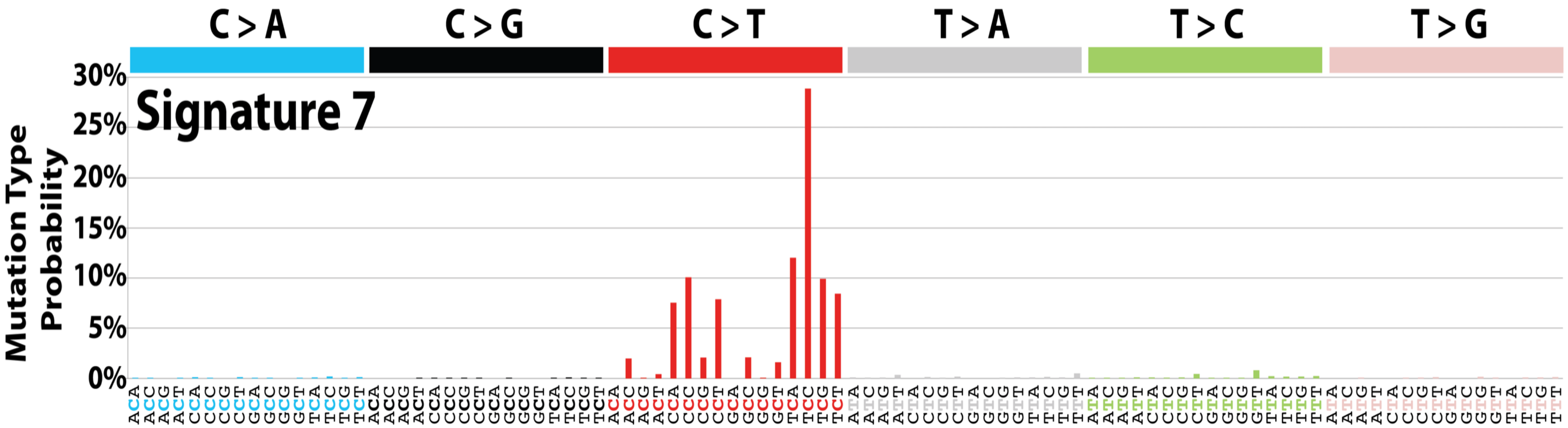


NMF analysis of 7042 tumors

15 out of 22 signatures

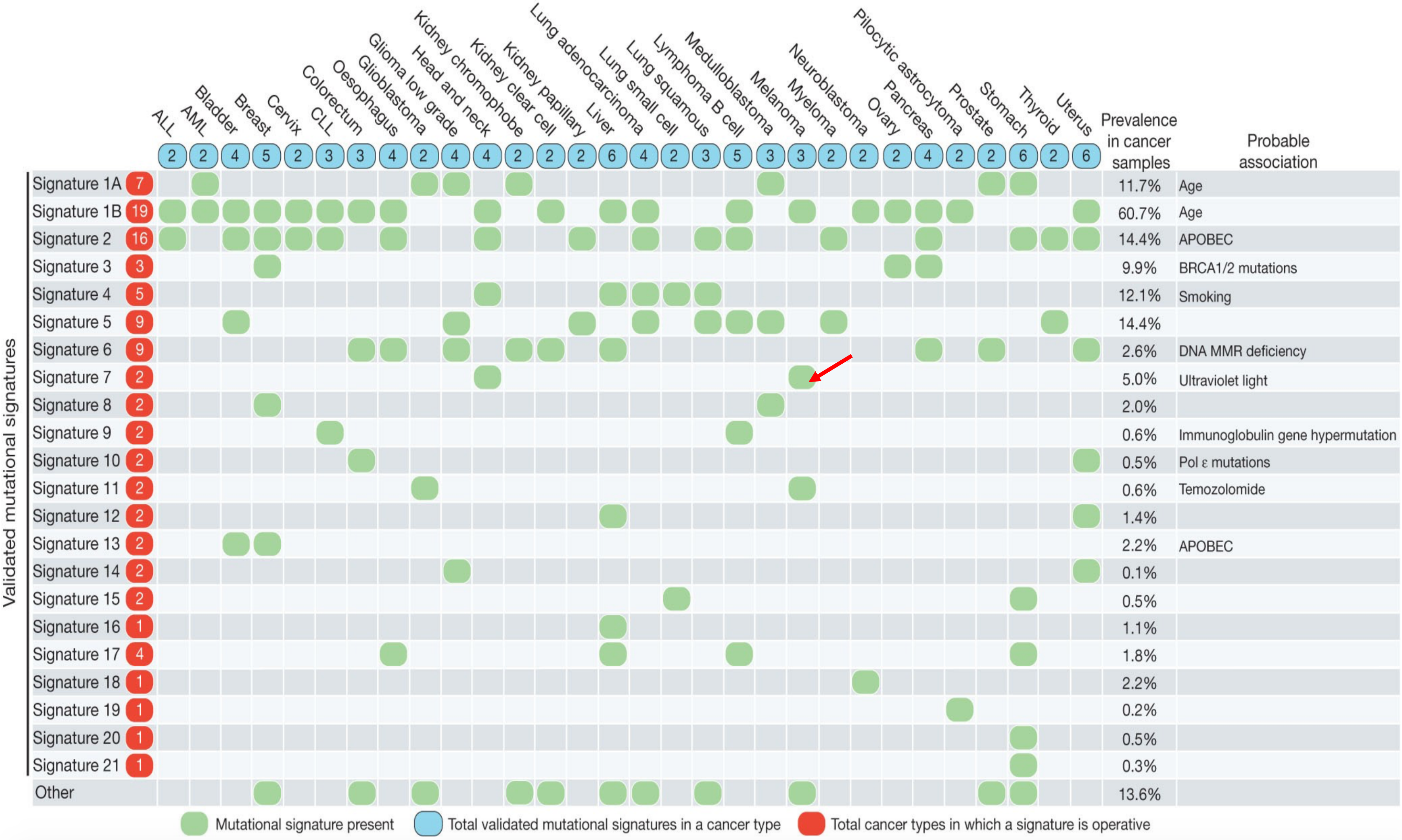
Ludmil B. Alexandrov, …, Michael R. Stratton Nature 500:415–421(2013)

# COSMIC Mutation Signatures



Ludmil B. Alexandrov, …, Michael R. Stratton Nature 500:415–421(2013)

# COSMIC Mutation Signatures



Validated mutational signatures

| | | ALL | AML | Bladder | Breast | Cervix | CLL | Colorectum | Oesophagus | Glioblastoma | Glioma low grade | Head and neck | Kidney chromophobe | Kidney clear cell | Kidney papillary | Liver | Lung adenocarcinoma | Lung small cell | Lung squamous | Lymphoma B cell | Medulloblastoma | Melanoma | Myeloma | Neuroblastoma | Ovary | Pancreas | Pilocytic astrocytoma | Prostate | Stomach | Thyroid | Uterus | Prevalence in cancer samples | Probable association |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2 | 2 | 4 | 5 | 2 | 3 | 3 | 4 | 2 | 4 | 4 | 2 | 2 | 2 | 6 | 4 | 2 | 3 | 5 | 3 | 3 | 2 | 2 | 2 | 4 | 2 | 2 | 6 | 2 | 6 | | |
| Signature 1A | 7 | | 🟩 | | | | | | | 🟩 | 🟩 | | 🟩 | | | | | | | | 🟩 | | | | | | | 🟩 | 🟩 | | | 11.7% | Age |
| Signature 1B | 19 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | 🟩 | | 🟩 | | 🟩 | 🟩 | | | 🟩 | | 🟩 | | 🟩 | 🟩 | 🟩 | 🟩 | | | | 🟩 | 60.7% | Age |
| Signature 2 | 16 | 🟩 | | 🟩 | 🟩 | 🟩 | 🟩 | | 🟩 | | | 🟩 | | | 🟩 | | 🟩 | | 🟩 | 🟩 | | | 🟩 | | | 🟩 | | | 🟩 | 🟩 | 🟩 | 14.4% | APOBEC |
| Signature 3 | 3 | | | | 🟩 | | | | | | | | | | | | | | | | | | | | 🟩 | 🟩 | | | | | | 9.9% | BRCA1/2 mutations |
| Signature 4 | 5 | | | | | | | | | | | 🟩 | | | | 🟩 | 🟩 | 🟩 | 🟩 | | | | | | | | | | | | | 12.1% | Smoking |
| Signature 5 | 9 | | | 🟩 | | | | | | | 🟩 | | | | 🟩 | | 🟩 | | 🟩 | 🟩 | 🟩 | | 🟩 | | | | | | | 🟩 | | 14.4% | |
| Signature 6 | 9 | | | | | | | 🟩 | 🟩 | | 🟩 | | 🟩 | 🟩 | | 🟩 | | | | | | | | | | 🟩 | | 🟩 | | | 🟩 | 2.6% | DNA MMR deficiency |
| Signature 7 | 2 | | | | | | | | | | | 🟩 | | | | | | | | | | 🟩 | | | | | | | | | | 5.0% | Ultraviolet light |
| Signature 8 | 2 | | | | 🟩 | | | | | | | | | | | | | | | | 🟩 | | | | | | | | | | | 2.0% | |
| Signature 9 | 2 | | | | | | 🟩 | | | | | | | | | | | | | 🟩 | | | | | | | | | | | | 0.6% | Immunoglobulin gene hypermutation |
| Signature 10 | 2 | | | | | | | 🟩 | | | | | | | | | | | | | | | | | | | | | | | 🟩 | 0.5% | Pol ε mutations |
| Signature 11 | 2 | | | | | | | | | 🟩 | | | | | | | | | | | | 🟩 | | | | | | | | | | 0.6% | Temozolomide |
| Signature 12 | 2 | | | | | | | | | | | | | | | 🟩 | | | | | | | | | | | | | | | 🟩 | 1.4% | |
| Signature 13 | 2 | | | 🟩 | 🟩 | | | | | | | | | | | | | | | | | | | | | | | | | | | 2.2% | APOBEC |
| Signature 14 | 2 | | | | | | | | | | 🟩 | | | | | | | | | | | | | | | | | | | | 🟩 | 0.1% | |
| Signature 15 | 2 | | | | | | | | | | | | | | | | | 🟩 | | | | | | | | | | | 🟩 | | | 0.5% | |
| Signature 16 | 1 | | | | | | | | | | | | | | | 🟩 | | | | | | | | | | | | | | | | 1.1% | |
| Signature 17 | 4 | | | | | | | | 🟩 | | | | | | | 🟩 | | | | 🟩 | | | | | | | | | 🟩 | | | 1.8% | |
| Signature 18 | 1 | | | | | | | | | | | | | | | | | | | | | | | 🟩 | | | | | | | | 2.2% | |
| Signature 19 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | 🟩 | | | | | 0.2% | |
| Signature 20 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 🟩 | | | 0.5% | |
| Signature 21 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 🟩 | | | 0.3% | |
| Other | | | | | 🟩 | | | 🟩 | | 🟩 | | | 🟩 | 🟩 | | 🟩 | 🟩 | | 🟩 | | | 🟩 | | | | | | 🟩 | 🟩 | | | 13.6% | |

🟩 Mutational signature present ● Total validated mutational signatures in a cancer type ● Total cancer types in which a signature is operative

Ludmil B. Alexandrov, …, Michael R. Stratton Nature 500:415–421(2013)

# Algorithm of NMF

$$X \approx WH$$

$$\min_{W,H\geq 0} ||X - WH||_F^2 = \sum_{i,j} (X - WH)_{ij}^2$$

Alternating multiplicative update

$H \leftarrow H \odot W^T X / W^T W H$

$W \leftarrow W \odot X H^T / W H H^T$

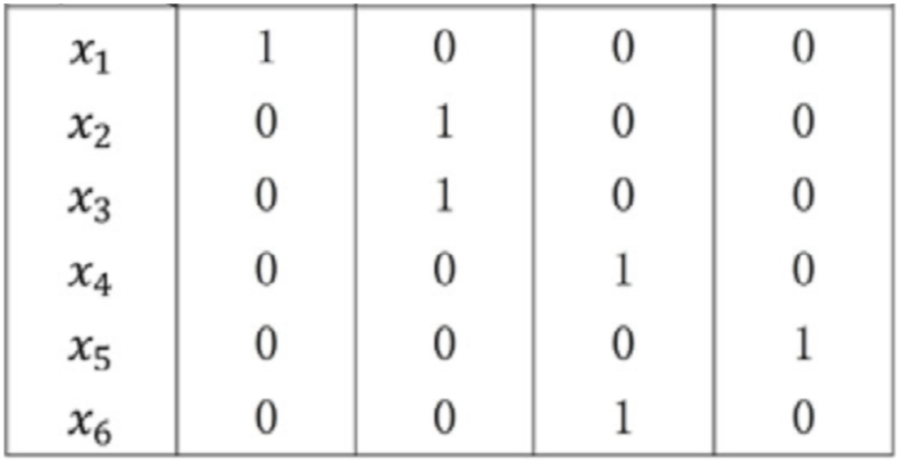
Hadamard product operator $\odot$
element-wise multiplication of matrices

# Matrix Representation of k-means Clustering

$$\min_{\mathcal{C}_1,\mathcal{C}_2,\ldots,\mathcal{C}_k} \sum_{i=1}^{k} \sum_{\mathbf{x}\in\mathcal{C}_i} ||\mathbf{x} - \mu_i||^2$$

$\mu_i$ is the centroid of data points in $C_i$

$B_{nk} =$

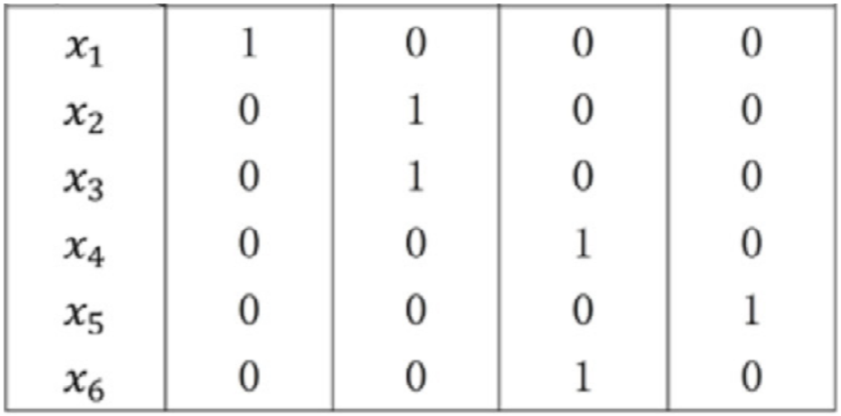


|  | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| $x_1$ | 1 | 0 | 0 | 0 |
| $x_2$ | 0 | 1 | 0 | 0 |
| $x_3$ | 0 | 1 | 0 | 0 |
| $x_4$ | 0 | 0 | 1 | 0 |
| $x_5$ | 0 | 0 | 0 | 1 |
| $x_6$ | 0 | 0 | 1 | 0 |

Class membership matrix B:
each row has only one 1; the others are 0
sum of column is the size of the cluster
columns are orthogonal

# Matrix Representation of k-means Clustering

$B_{nk} =$



| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| $x_1$ | 1 | 0 | 0 | 0 |
| $x_2$ | 0 | 1 | 0 | 0 |
| $x_3$ | 0 | 1 | 0 | 0 |
| $x_4$ | 0 | 0 | 1 | 0 |
| $x_5$ | 0 | 0 | 0 | 1 |
| $x_6$ | 0 | 0 | 1 | 0 |

$\sum_i B_{ij} = |\mathcal{C}_j|$ sum of column is the size of the cluster

$$D = diag(1/|\mathcal{C}_1|, 1/|\mathcal{C}_2|, ..., 1/|\mathcal{C}_k|)$$

$$(BD^{\frac{1}{2}})^T BD^{\frac{1}{2}} = D^{\frac{1}{2}} B^T BD^{\frac{1}{2}} = I$$

# Matrix Representation of k-means Clustering

$$\min_{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k} \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} ||\mathbf{x} - \mu_i||^2$$

$$\mu_i = \frac{1}{|\mathcal{C}_i|} \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x}$$

M = XBD

M: k means (centroids) in columns

X: n samples in columns

B: k clusters in columns

D: diagonal matrix; 1/cluster size

XBDB$^T$: each sample selects its corresponding centroid in columns

$$\min_{B} ||X - XBDB^T||_F^2$$

# k-means Clustering is Equivalent to Sparse NMF

k-means $$\min_{B} ||X - XBDB^T||_F^2$$

NMF $$\min_{W,H\geq 0} ||X - WH||_F^2$$

W ~ XBD

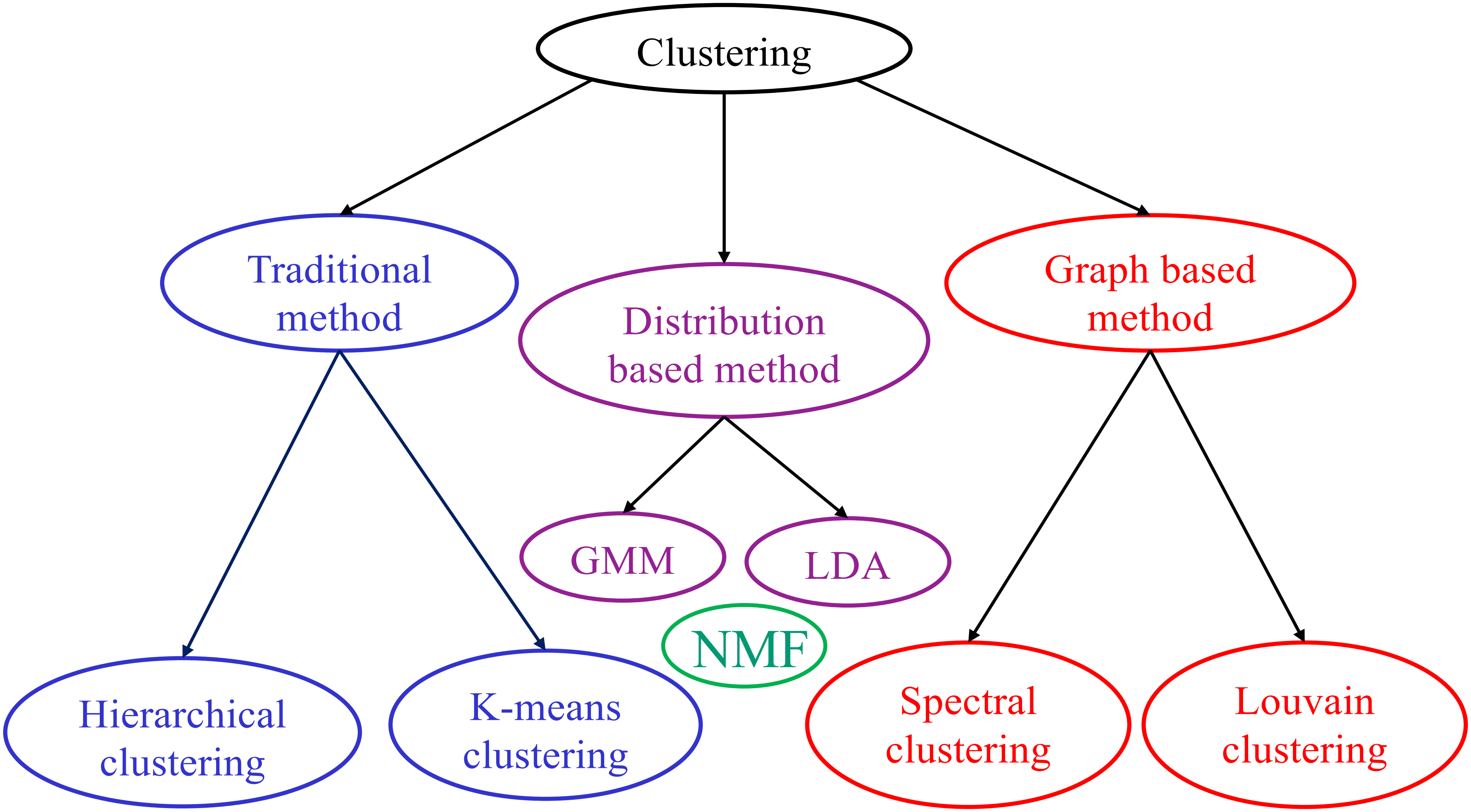
H ~ $B^T$

B is orthogonal and sparse

Sparse NMF or non-smooth NMF

NMF is related to GMM and mixture model

# Outline of Clustering Methods



Clustering

Traditional method

Distribution based method

Graph based method

GMM

LDA

NMF

Hierarchical clustering

K-means clustering

Spectral clustering

Louvain clustering

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

Contributed by Emily Tai