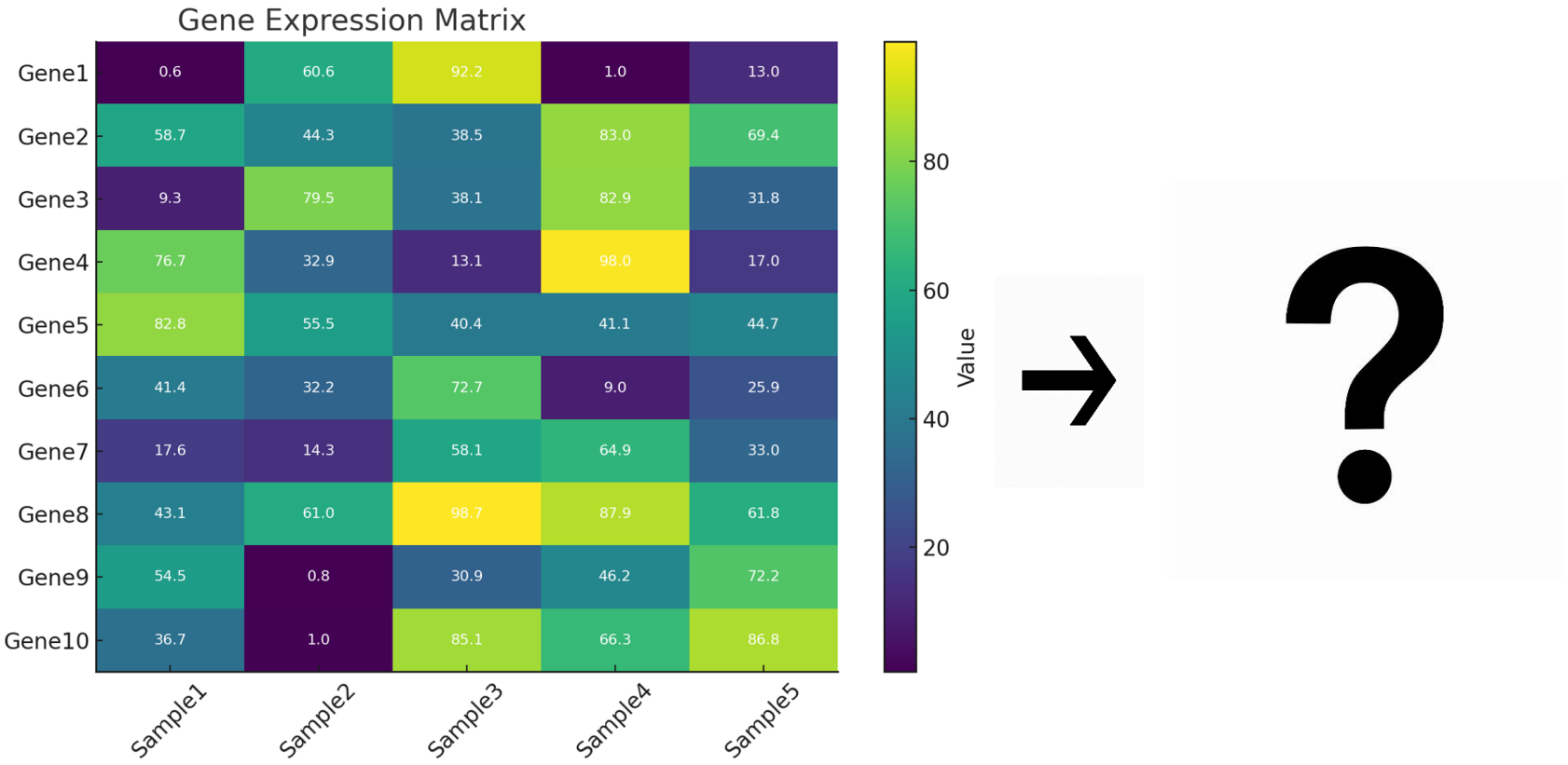
RNA-seq 데이터로부터.



컴퓨터가 데이터를 통해 **스스로 패턴을 학습**하고,
새로운 데이터에 대해 **예측하거나 분류**하는 알고리즘을 만들 수 있을까

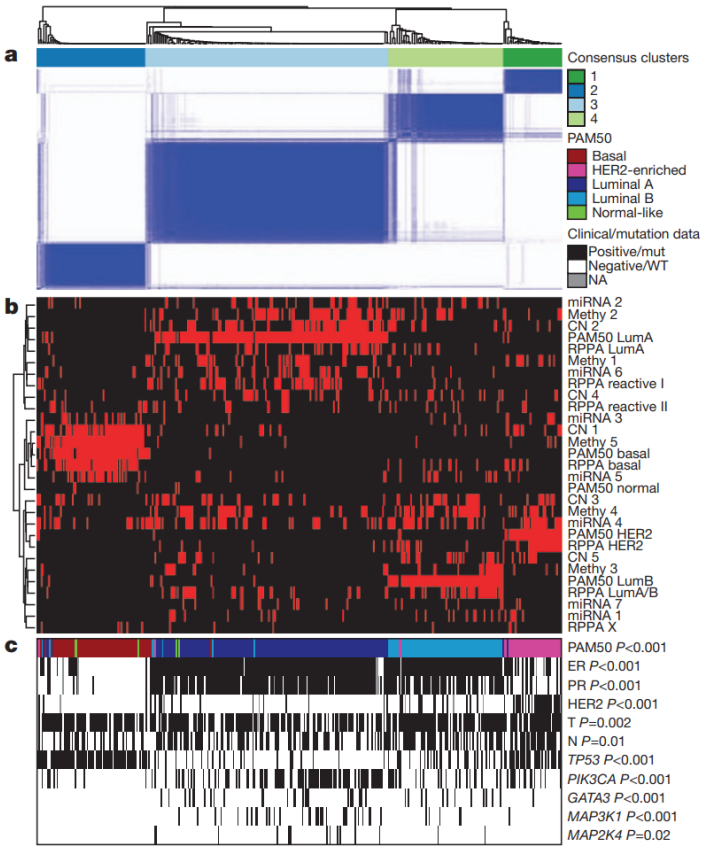
RNA-seq 데이터로부터.

ARTICLE

doi:10.1038/nature11412

Comprehensive molecular portraits of human breast tumours

The Cancer Genome Atlas Network*



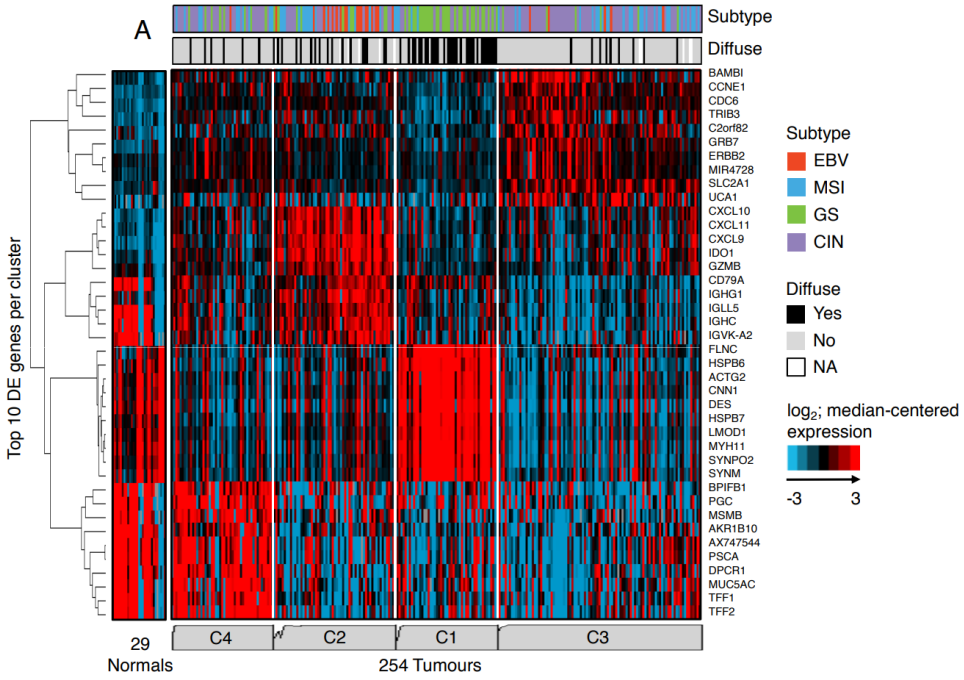
ARTICLE

OPEN

doi:10.1038/nature13480

Comprehensive molecular characterization of gastric adenocarcinoma

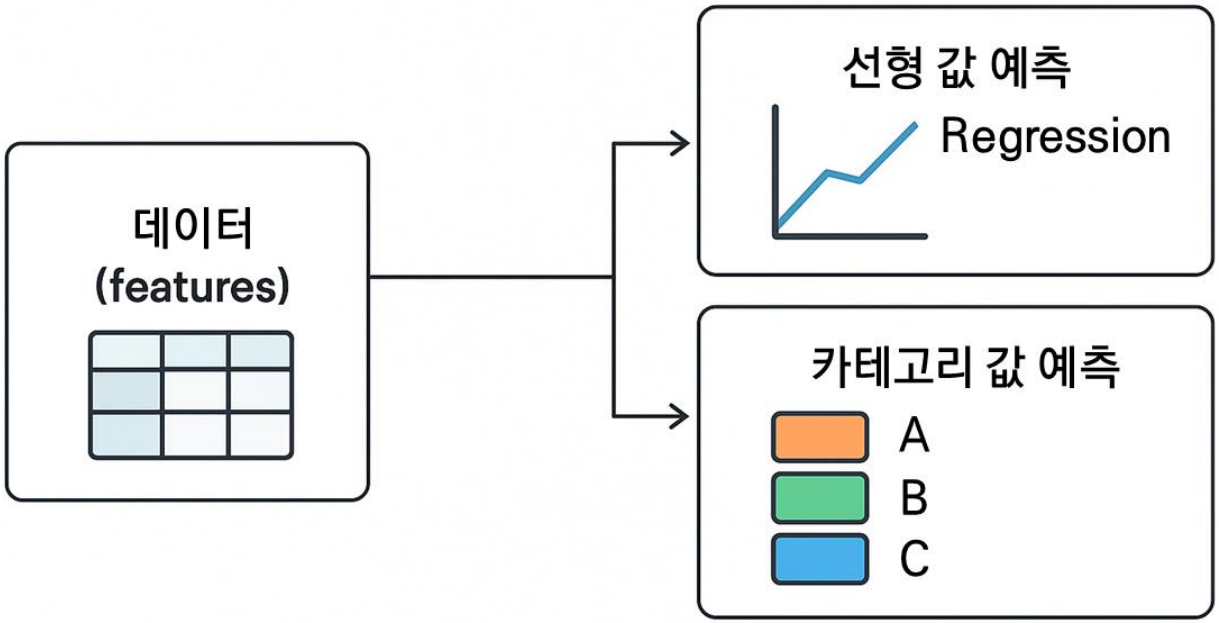
The Cancer Genome Atlas Research Network*



Supervised vs Unsupervised

Supervised Learning (지도학습)

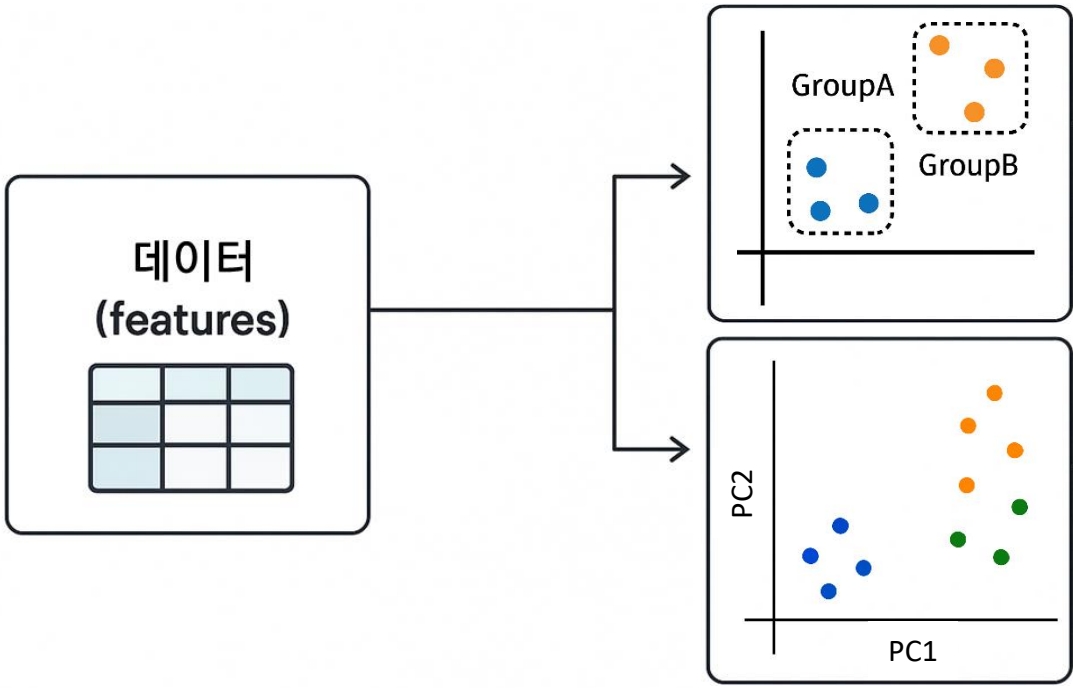
입력 X와 레이블 Y(정답)가 쌍(pair)으로 주어질 때, 학습을 통해 새로운 입력에 대한 **NEW Y**(예측값)를 생성



Supervised vs Unsupervised

Unsupervised Learning (비지도학습)

레이블 없이 순수 입력 데이터만 있을 때, 데이터 구조나 패턴(잠재 요인)을 찾아내는 기법



Clustering

Clustering: 레이블(정답) 없이 데이터 자체의 유사성(거리나 밀도 등)을 기반으로 샘플들을 "그룹(클러스터)"으로 묶는 비지도 학습

K-means clustering

사전에 군집 개수 K를 지정하고, "클러스터 중심(centroid)"과 샘플 간 거리를 최소화!!

Goal : $$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{x \in C_k} \| x - \mu_k \|^2 \quad \text{where } \mu_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$$

Clustering

Clustering: 레이블(정답) 없이 데이터 자체의 유사성(거리나 밀도 등)을 기반으로 샘플들을 "그룹(클러스터)"으로 묶는 비지도 학습

Non-negative matrix factorization clustering

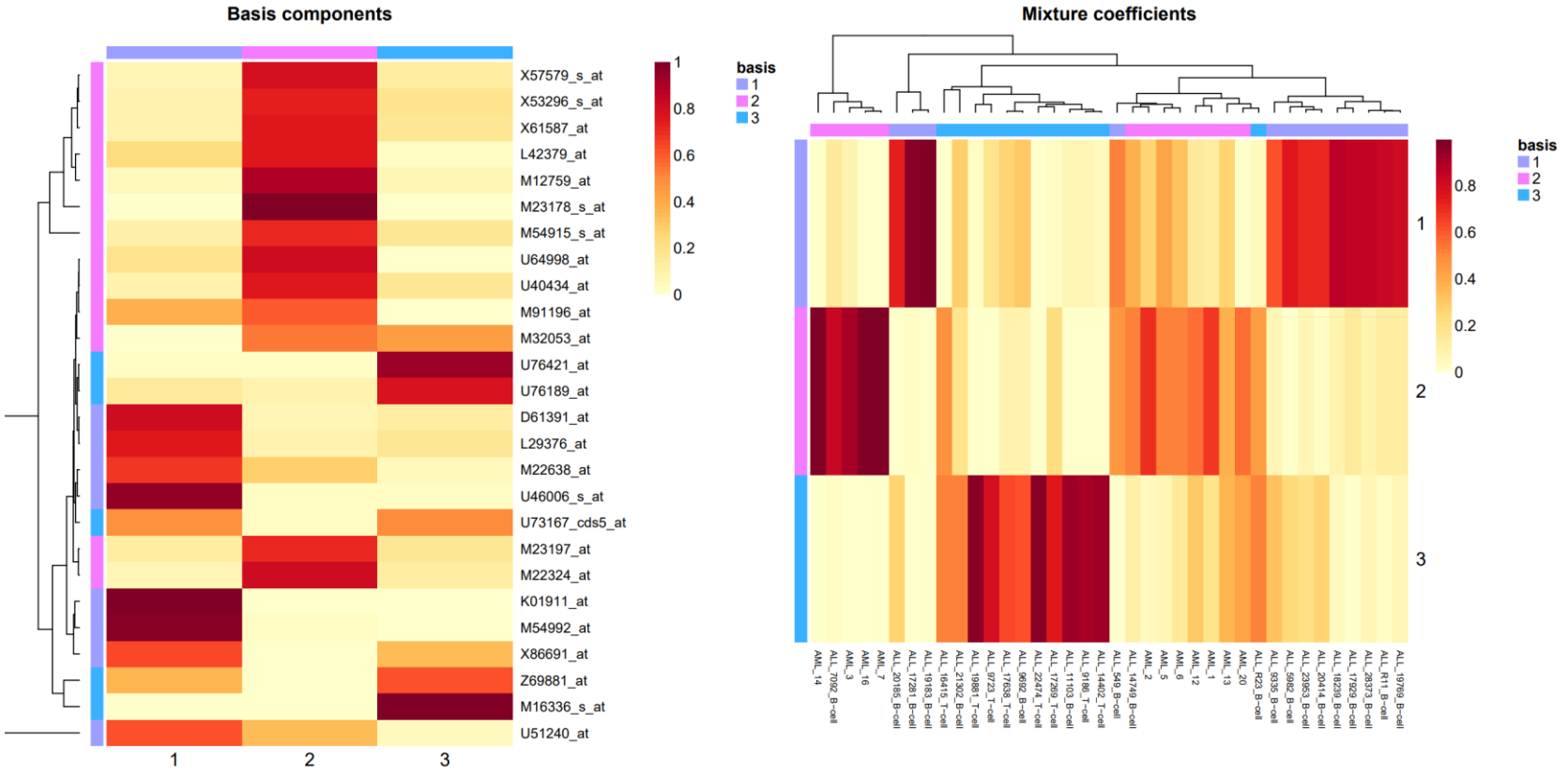
NMF 자체는 '군집화 알고리즘'이라기보다는 비음수 행렬 분해 기법이지만, 그 결과로 얻어지는 **Basis 행렬 W**과 **Coefficient 행렬 H** 를 활용하여 샘플(또는 유전자)을 클러스터링

Goal : $X \approx WH$, $\min_{W,H \geq 0} \underbrace{[D(X, WH) + R(W, H)]}_{=F(W,H)}$

NMF

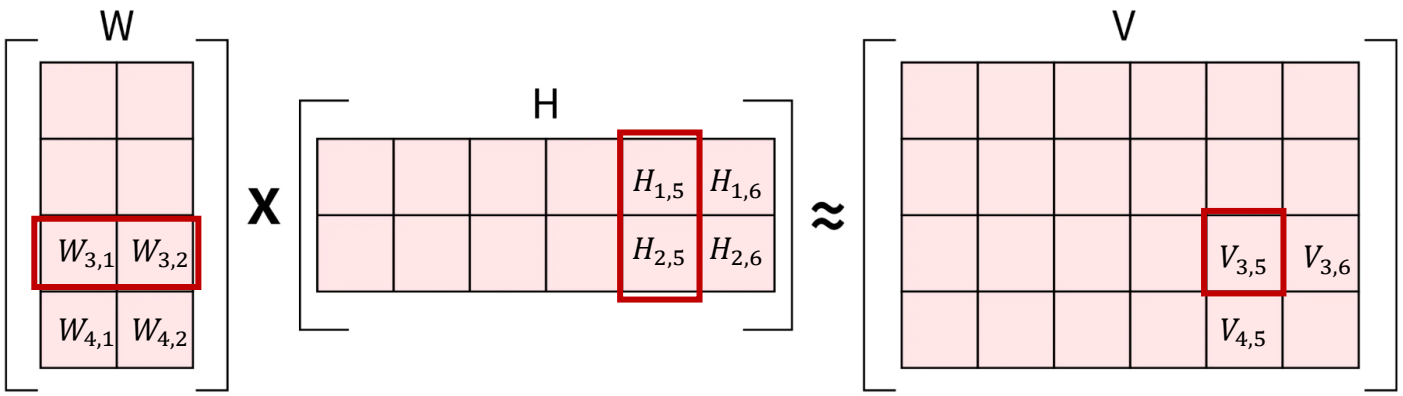
유전자 X 샘플 데이터 매트릭스를

유전자를 나타내는 (유전자) X (N) 매트릭스와
(N) X (샘플) 매트릭스로 행렬 분해할 수 있다면?



NMF

행렬의 인수분해가 의미하는 것은?



$$W_{3,1} * H_{1,5} + W_{3,2} * H_{2,5} = V_{3,5} \qquad W_{3,1} * H_{1,6} + W_{3,2} * H_{2,6} = V_{3,6}$$

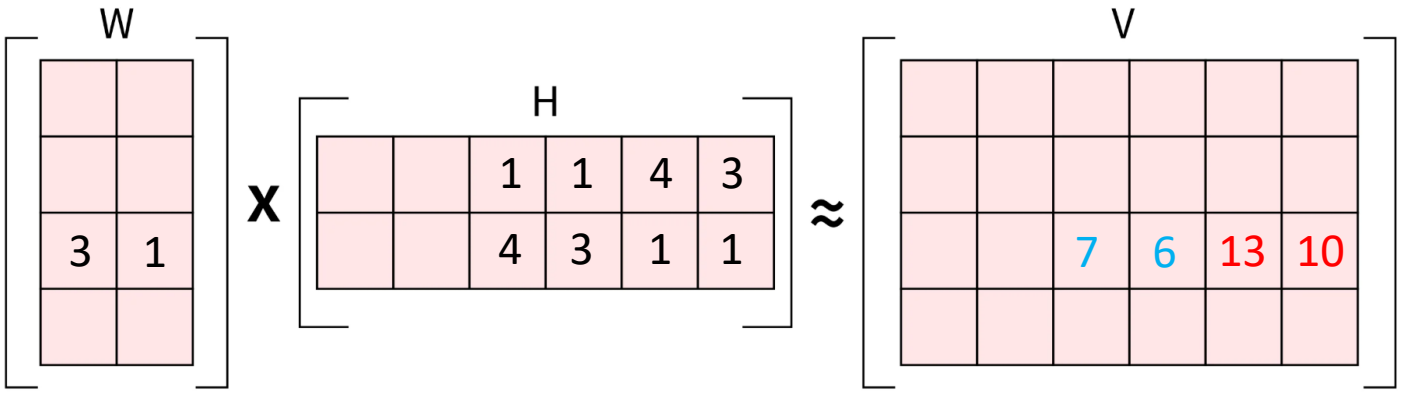
$$W_{4,1} * H_{1,5} + W_{4,2} * H_{2,5} = V_{4,5}$$

다섯번째 샘플과 여섯번째 샘플이 비슷하다?
-> $V_{3,5}$ 와 $V_{3,6}$ 이 비슷하다!!! -> $H_{\sim,5}$ 와 $H_{\sim,6}$ 가 비슷하다!!

그렇다면 W, H 크기는 무엇을 의미하는가?

NMF

H 행렬 정보를 어떻게 사용할까?



$W_{3,1} * H_{1,3} + W_{3,2} * H_{2,3} = V_{3,3}$
$3 * 1 + 1 * 4 = 7$

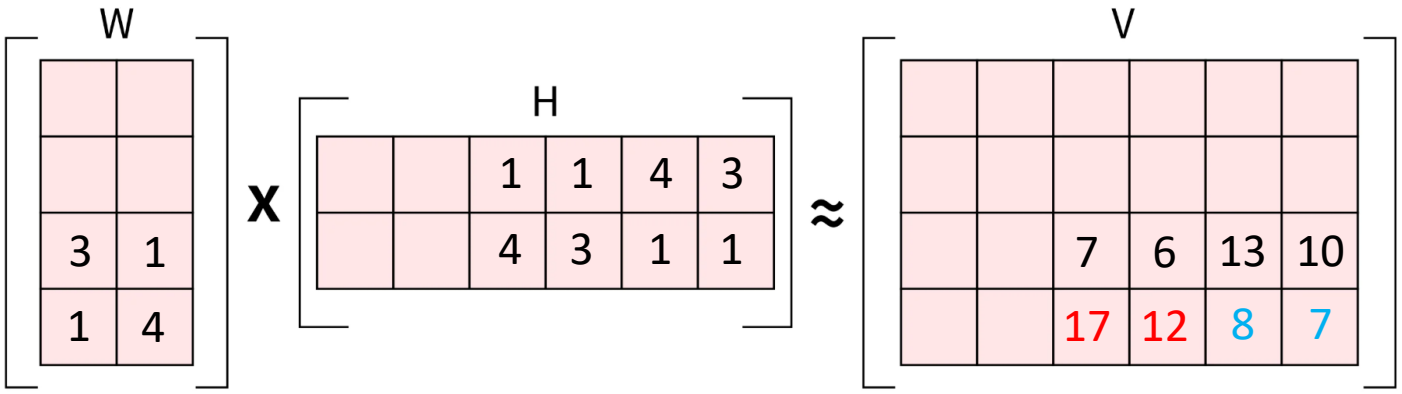
$W_{3,1} * H_{1,4} + W_{3,2} * H_{2,4} = V_{3,4}$
$3 * 1 + 1 * 3 = 6$

$W_{3,1} * H_{1,5} + W_{3,2} * H_{2,5} = V_{3,5}$
$3 * 4 + 1 * 1 = 13$

$W_{3,1} * H_{1,6} + W_{3,2} * H_{2,6} = V_{3,6}$
$3 * 3 + 1 * 1 = 10$

NMF

W 행렬 정보를 어떻게 사용할까?



$$W_{4,1} * H_{1,3} + W_{4,2} * H_{2,3} = V_{4,3}$$
$$1 * 1 + 4 * 4 = 17$$

$$W_{4,1} * H_{1,4} + W_{4,2} * H_{2,4} = V_{4,4}$$
$$1 * 1 + 4 * 3 = 12$$

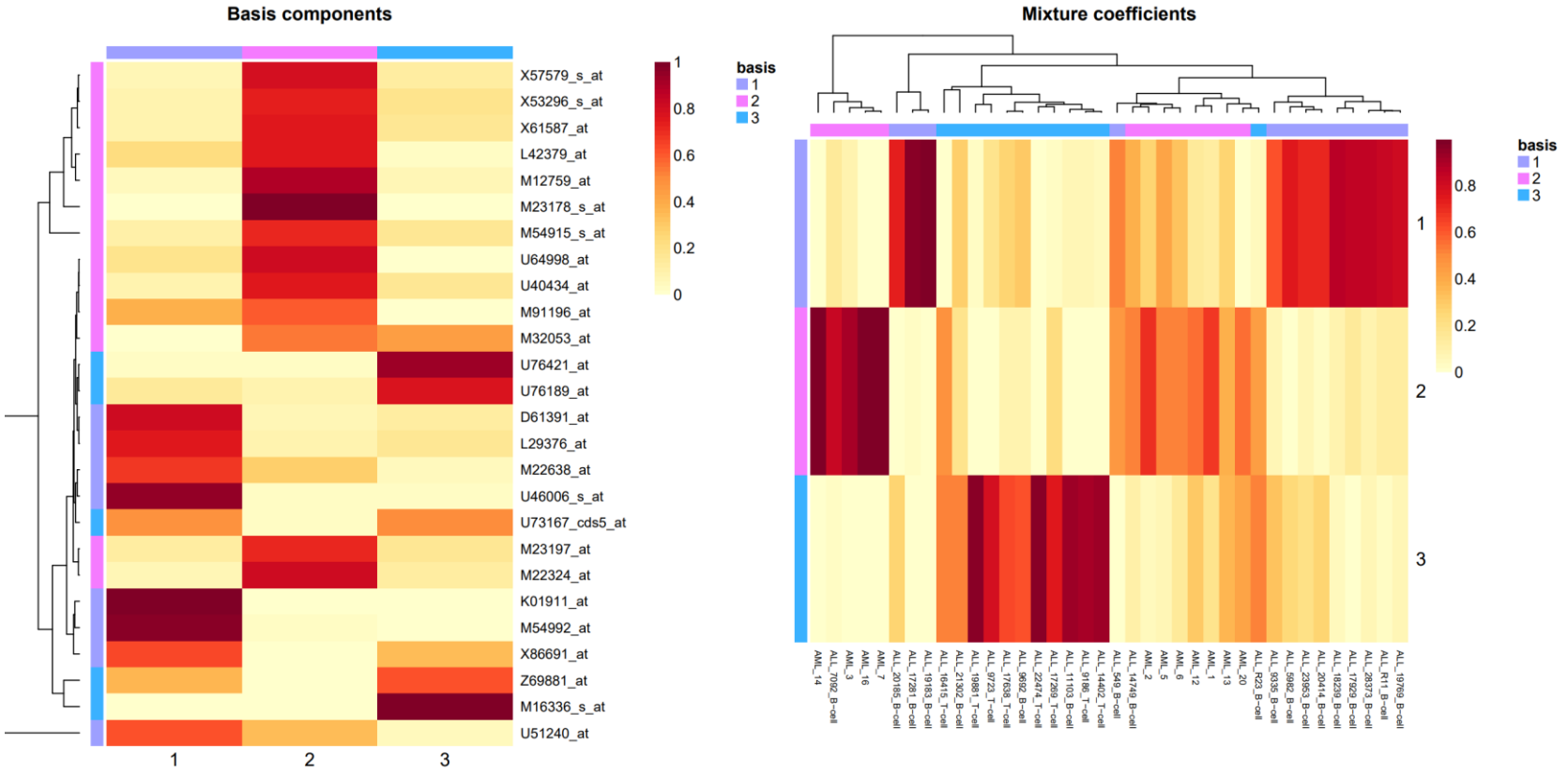
$$W_{4,1} * H_{1,5} + W_{4,2} * H_{2,5} = V_{4,5}$$
$$1 * 4 + 4 * 1 = 8$$

$$W_{4,1} * H_{1,6} + W_{4,2} * H_{2,6} = V_{4,6}$$
$$1 * 3 + 4 * 1 = 7$$

NMF

유전자 X 샘플 데이터 매트릭스를

유전자를 나타내는 (유전자) X (N) 매트릭스와
(N) X (샘플) 매트릭스로 행렬 분해할 수 있다면?



NMF

H 행렬 정보를 어떻게 사용할까?

0.8147	0.9134	0.2785	0.9649	0.9572	0.1419	0.7922	0.03571	0.6787	0.3922
0.9058	0.6324	0.5469	0.1576	0.4854	0.4218	0.9595	0.8491	0.7577	0.6555
0.127	0.09754	0.9575	0.9706	0.8003	0.9157	0.6557	0.934	0.7431	0.1712

샘플 j는 클러스터 $\arg\max_k H_{k,j}$

2 1 3 3 1 3 2 3 2 2

NMF

Connectivity Matrix

클러스터가 결정되면, **연결 행렬 Connectivity Matrix)** 을 만든다

$$C_{ij}^{(t)} = \begin{cases} 1, & \text{샘플 i 와 j 가 같은 클러스터면} \\ 0, & \text{다르면} \end{cases}$$

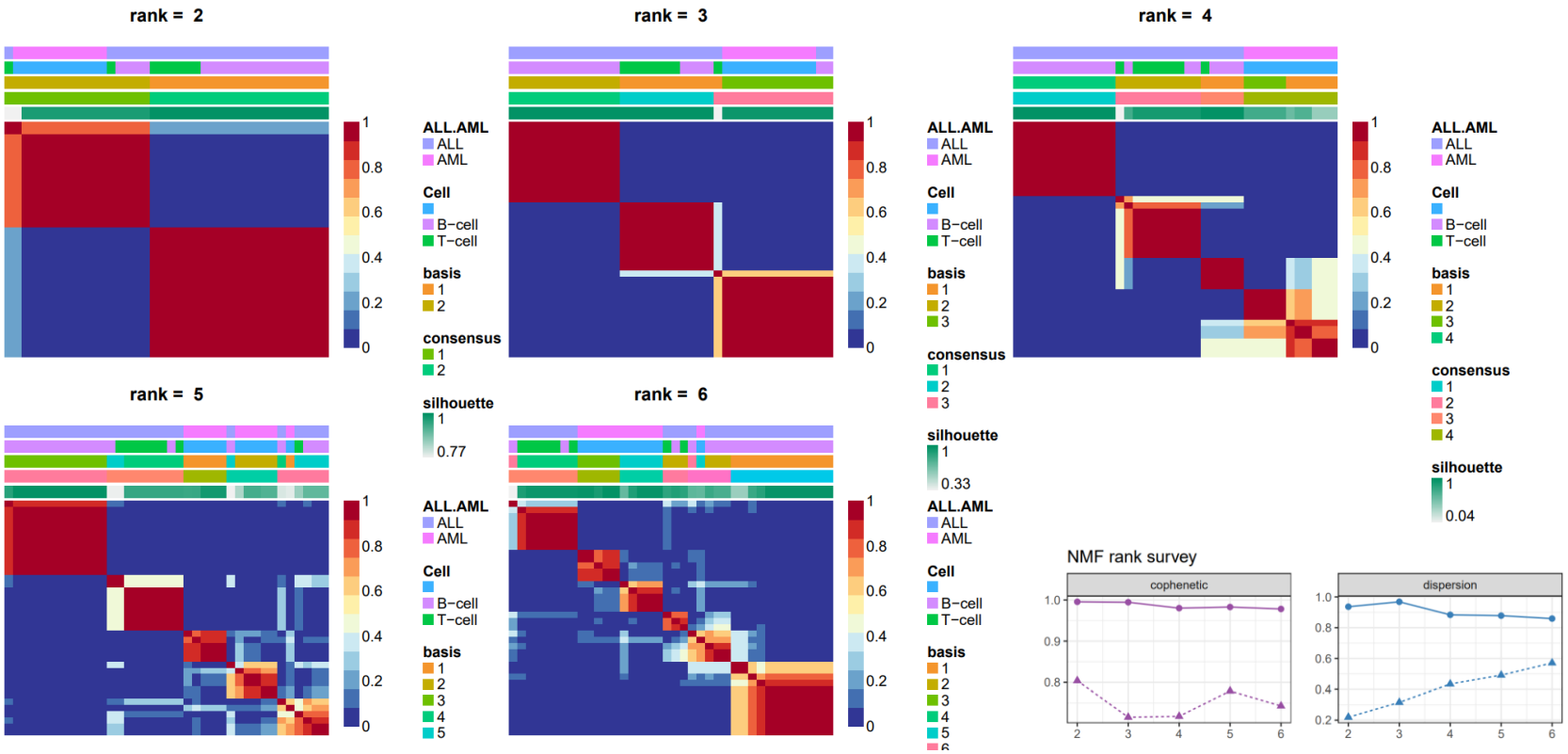
$C_{i,j}$

1	0	0	0	0	0	1	0	1	1
0	1	0	0	1	0	0	0	0	0
0	0	1	1	0	1	0	1	0	0
0	0	1	1	0	1	0	1	0	0
0	1	0	0	1	0	0	0	0	0
0	0	1	1	0	1	0	1	0	0
1	0	0	0	0	0	1	0	1	1
0	0	1	1	0	1	0	1	0	0
1	0	0	0	0	0	1	0	1	1
1	0	0	0	0	0	1	0	1	1

	2	5	1	7	9	10	3	4	6	8
2	1	1	0	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0	0
1	0	0	1	1	1	1	0	0	0	0
7	0	0	1	1	1	1	0	0	0	0
9	0	0	1	1	1	1	0	0	0	0
10	0	0	1	1	1	1	0	0	0	0
3	0	0	0	0	0	0	1	1	1	1
4	0	0	0	0	0	0	1	1	1	1
6	0	0	0	0	0	0	1	1	1	1
8	0	0	0	0	0	0	1	1	1	1

NMF

Cluster 그룹은 몇 개가 적당할까?



Clustering

실습