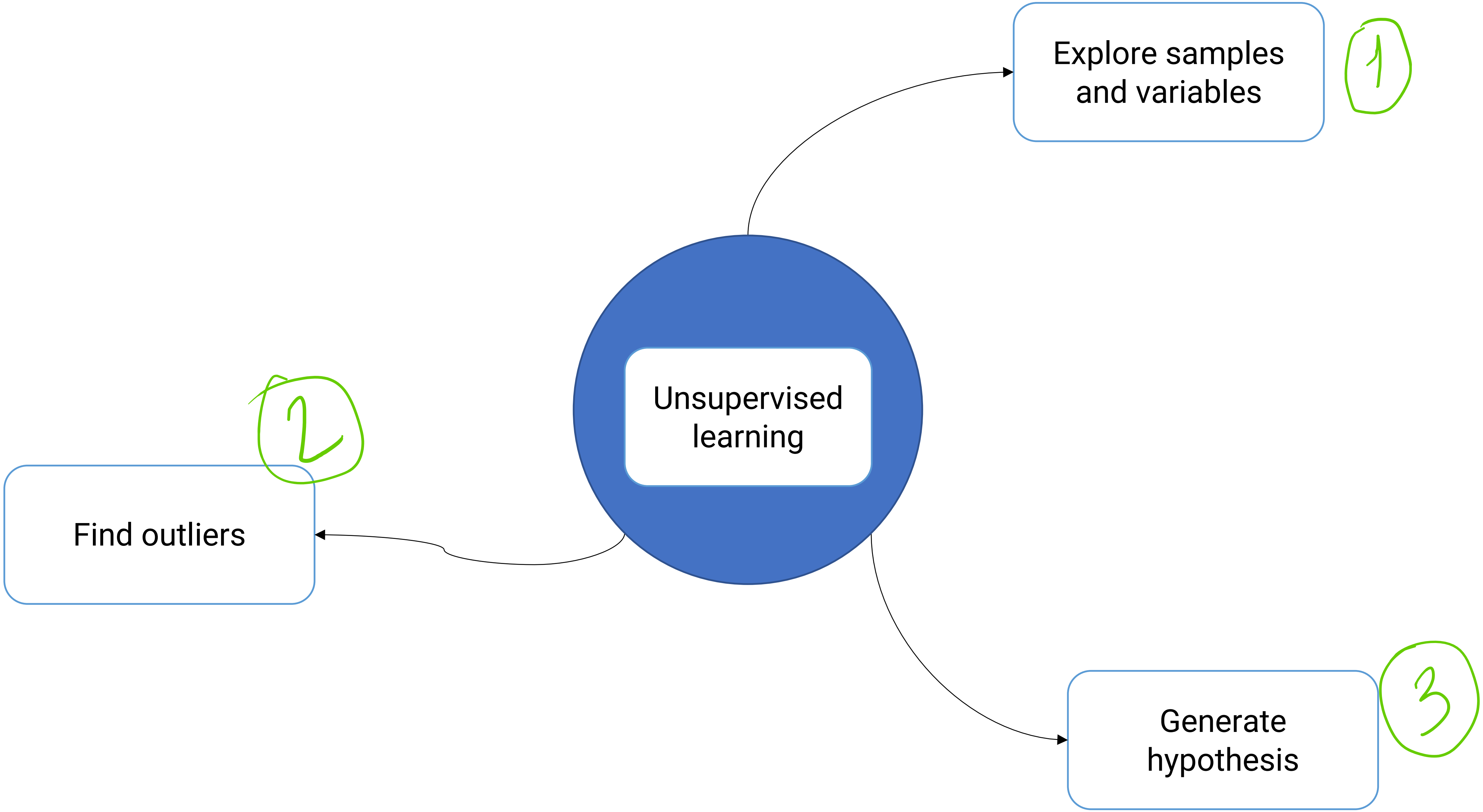
Computational genomics: hands on course

Unsupervised learning



Unsupervised learning

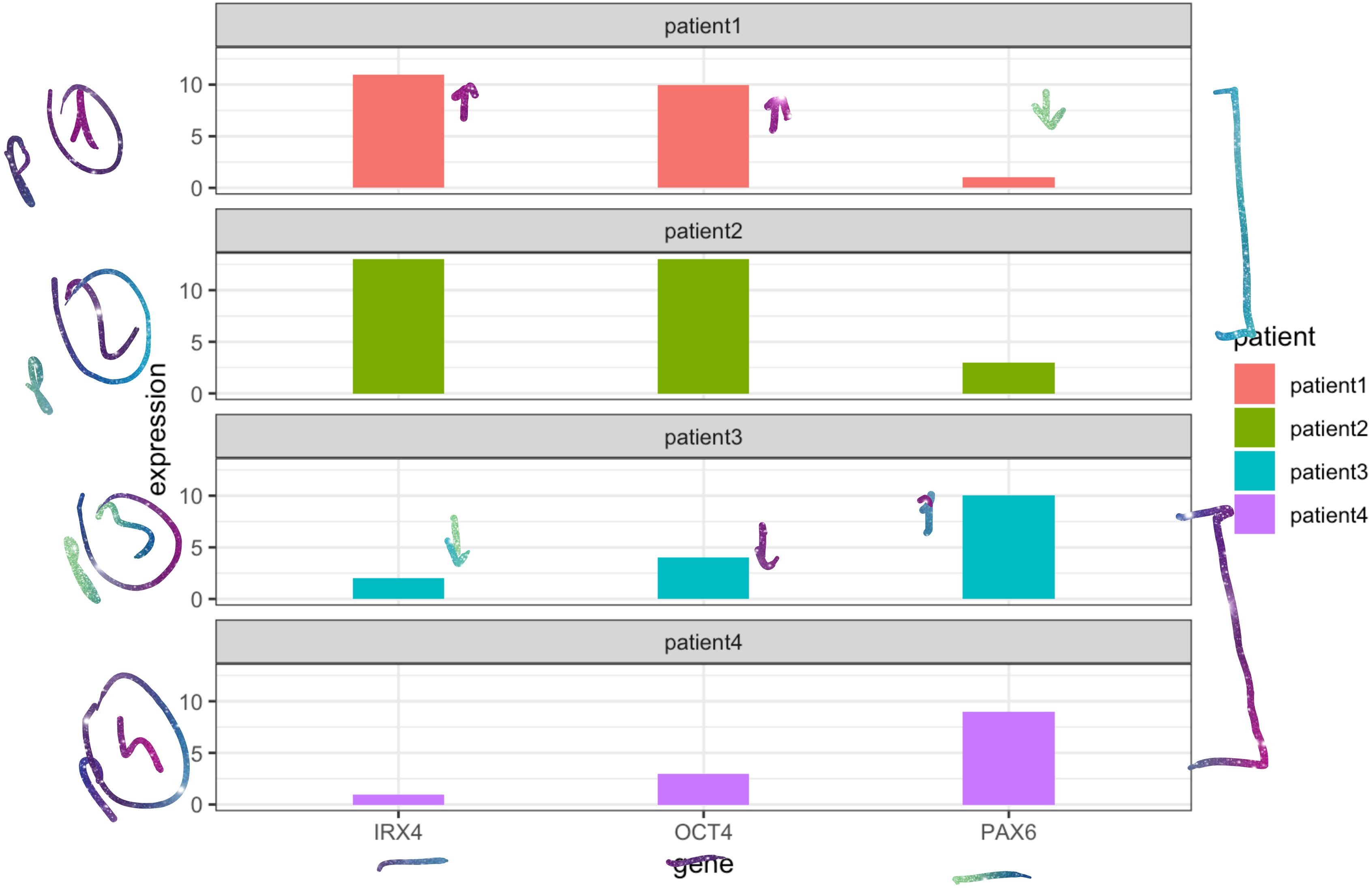
Explore samples and variables 1

Find outliers 2

Generate hypothesis 3

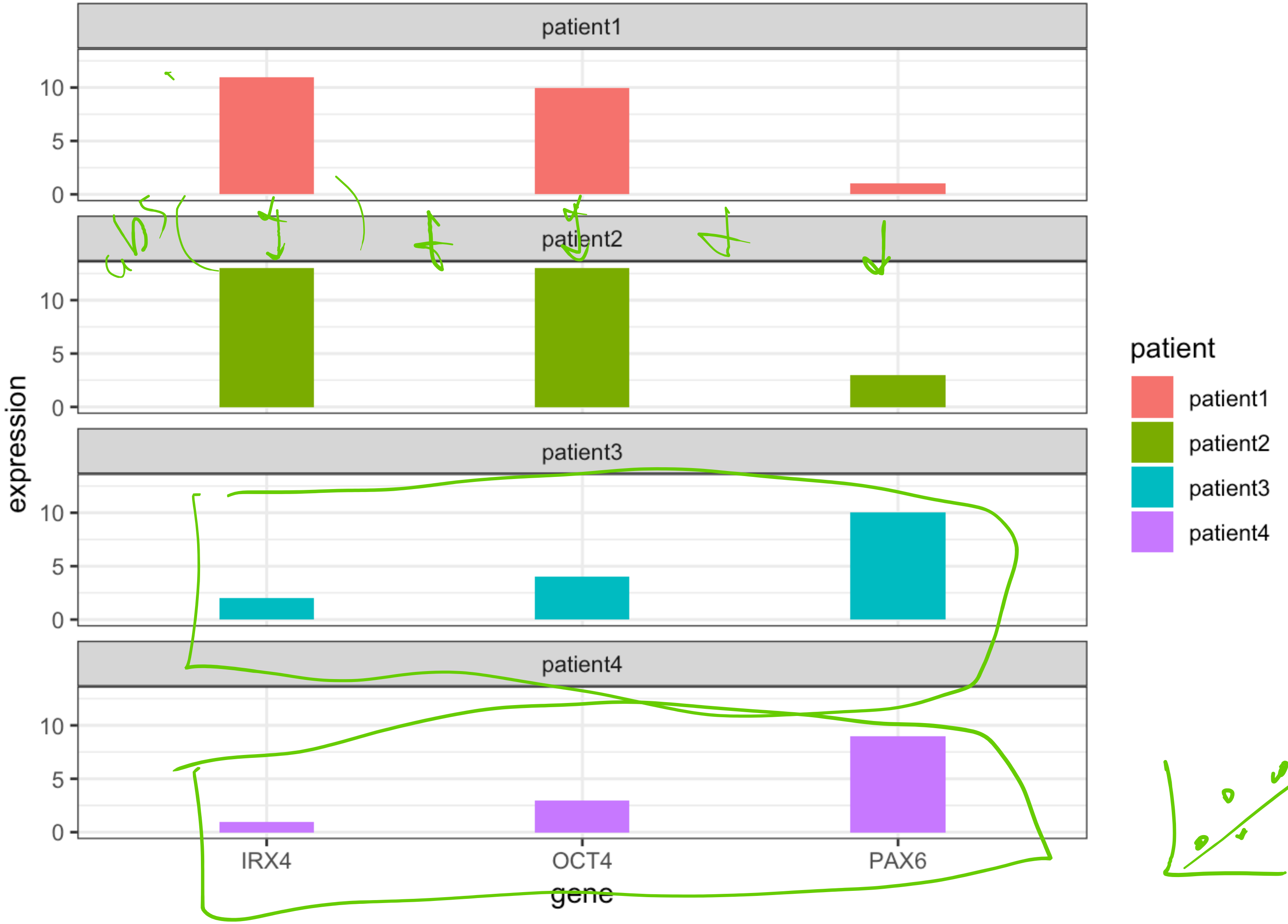
Clustering: grouping samples

distance metrics



Clustering: grouping samples

distance metrics



$$d_{AB} = \sum_{i=1}^{n} |e_{Ai} - e_{Bi}|$$

Manhattan distance" or "L1 norm"

$$d_{AB} = \sqrt{\sum_{i=1}^{n} (e_{Ai} - e_{Bi})^2}$$

"Euclidean Distance" or "L2 norm"

0,9 =

$$d_{AB} = 1 - \rho$$

1-rho

→ 0,1

"Correlation Distance"

Cor. dist

0,1 ⇒ 0,9

Data processing

Scaling

standardization

```
(x-mean(x))/sd(x)
```

$\Rightarrow \bar{x}=0$

sd=1

scaling

✓

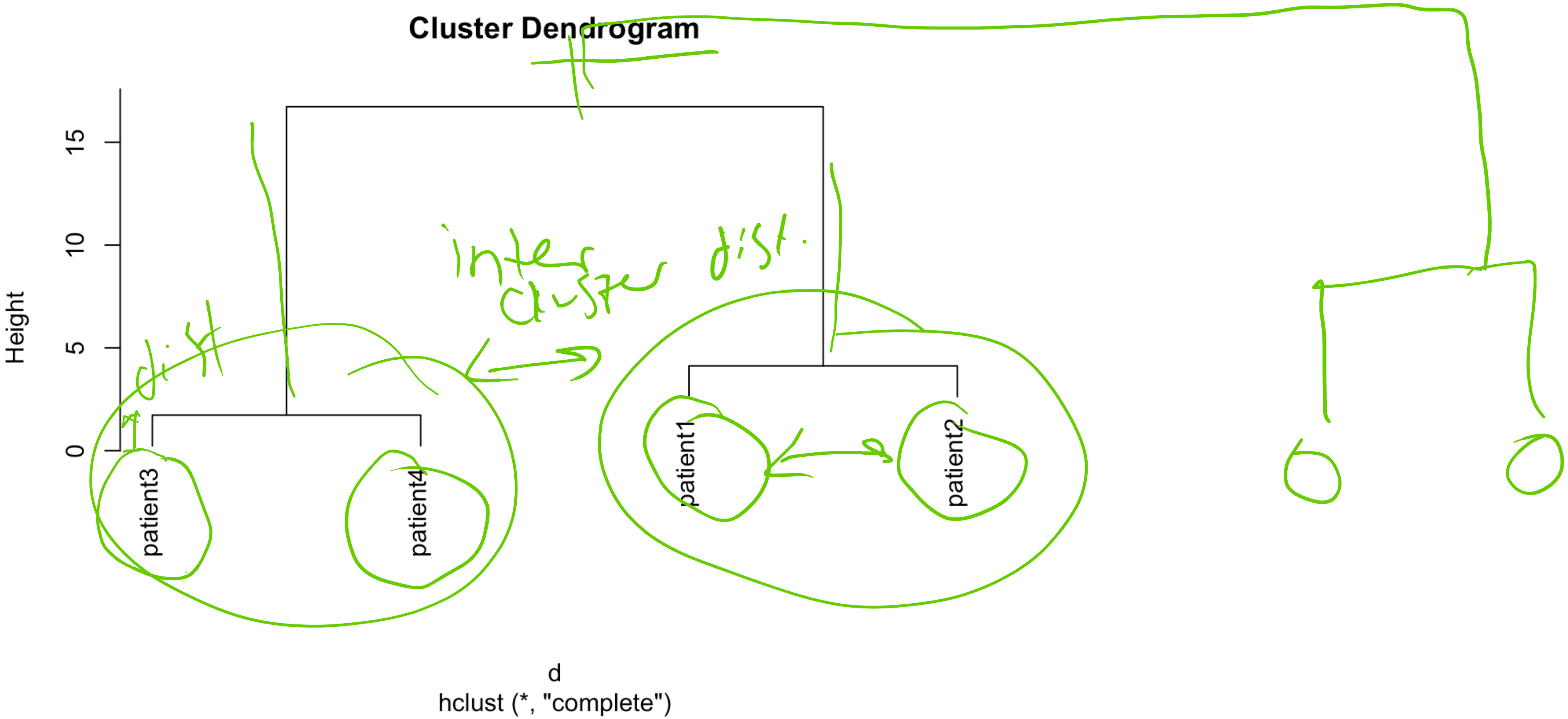
gene. exp

scale()

More on this later in the supervised learning section.

Clustering: grouping samples

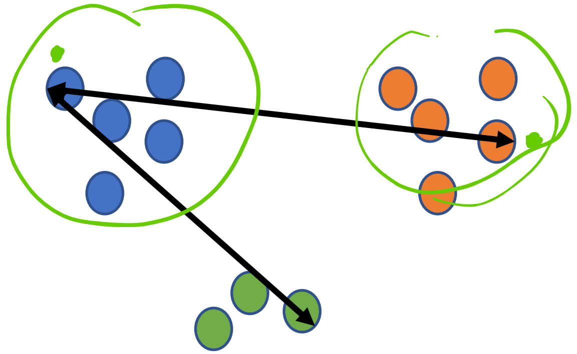
Hierarchical clustering



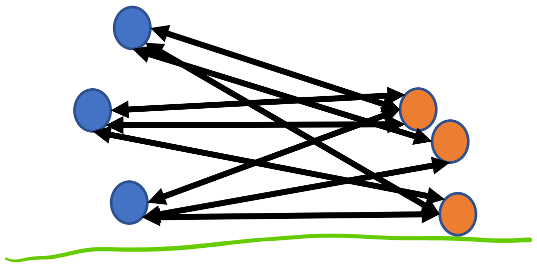
Clustering: grouping samples

different approaches for merging clusters

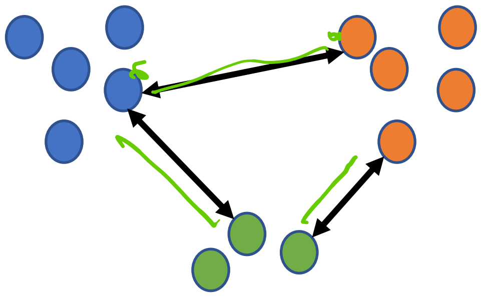
Complete Linkage



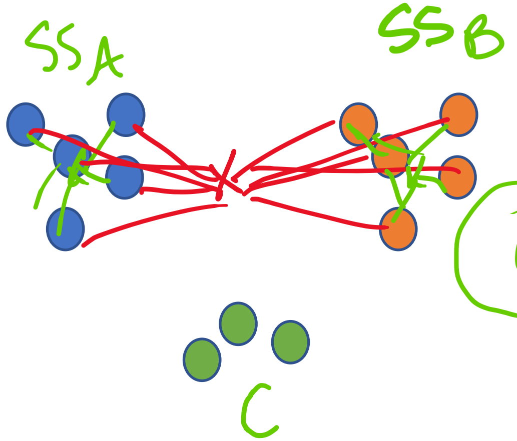
Average Linkage



Single Linkage



Ward's method



$C_{AB} = SS_{AB} - SS_A - SS_B$ → 10

$C_{AC} = 5$

$C_{BC} = 20$

Clustering: grouping samples

Real world example

Gene expression profiles from leukaemia patients and healthy controls

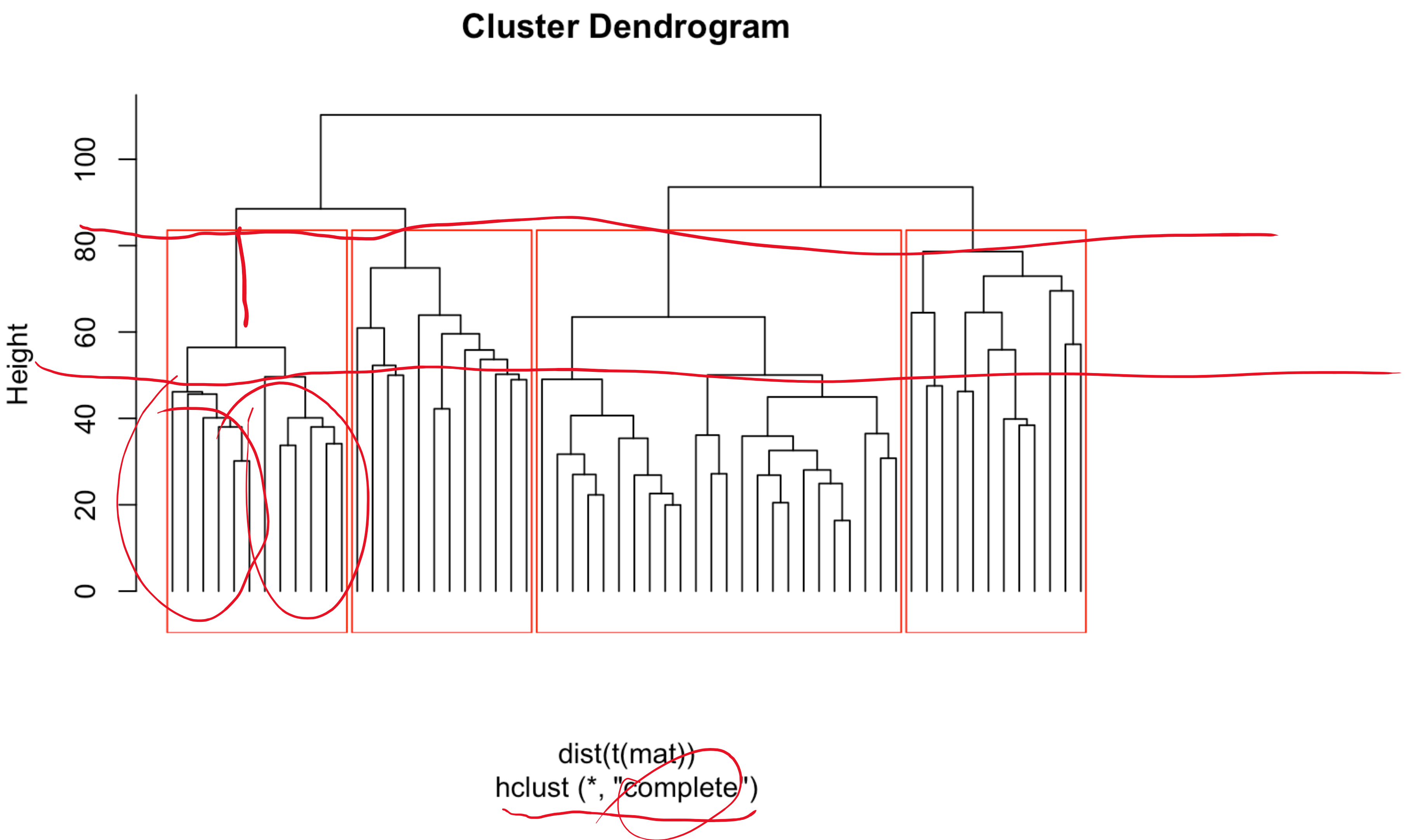
ALL
AML
CLL
CML

no Leukamia

1000 genes

Clustering: grouping samples

cutting the tree

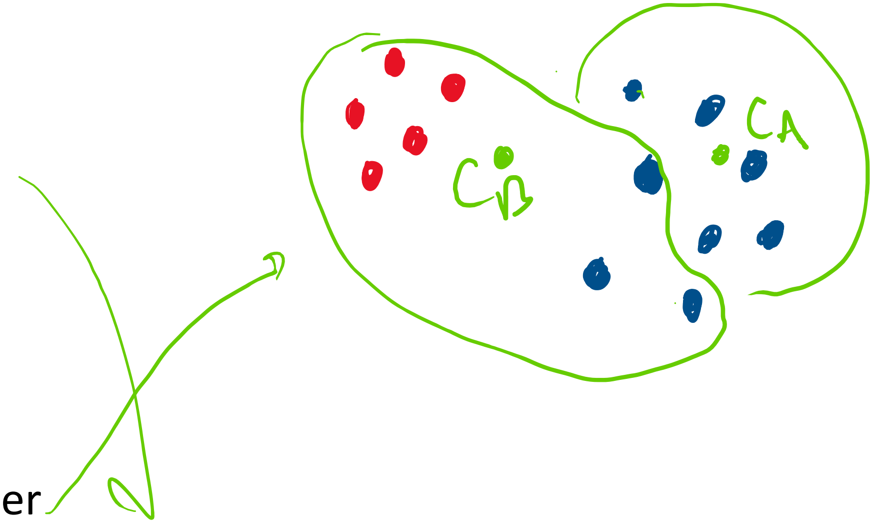


Clustering: grouping samples

K-means clustering

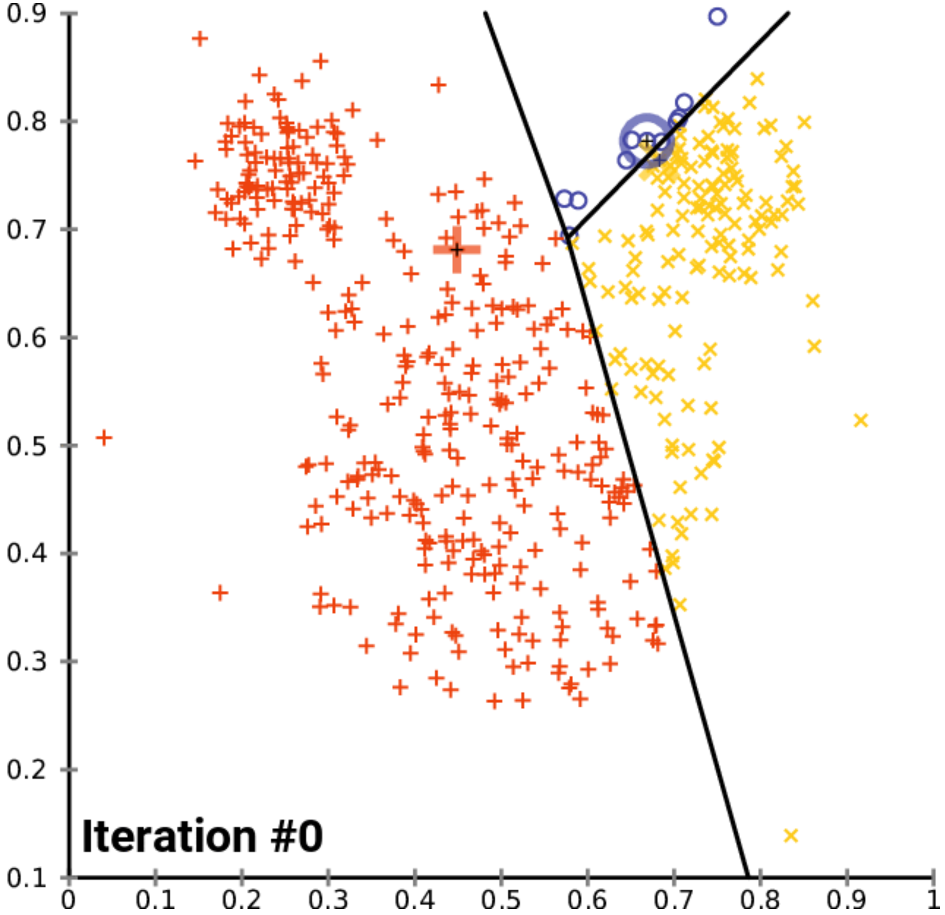
k=2

1. randomly choose k centers
2. Assign each data point to nearest center
3. Update centroid as the mean value of data points in the cluster
4. Repeat steps 2-3 until until sum of squared distances to cluster minimized



Clustering: grouping samples

K-means clustering



Clustering: grouping samples

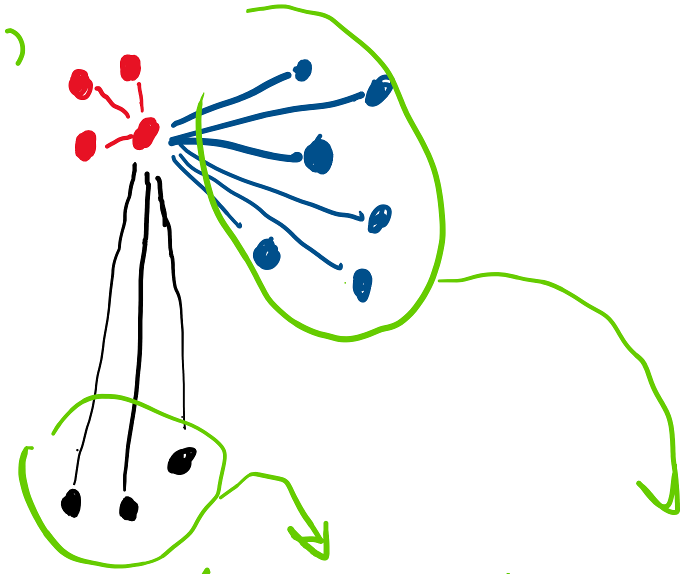
how to define the best k ? -> Silhouette score

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

+

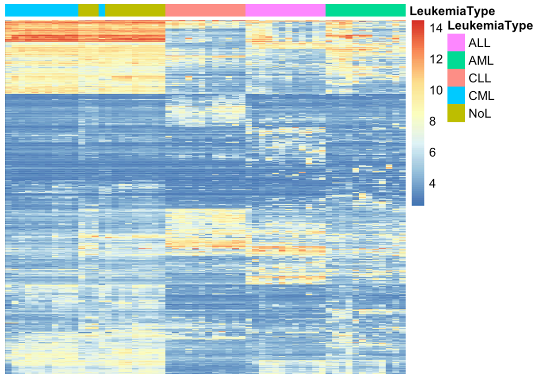
[-1, 1]

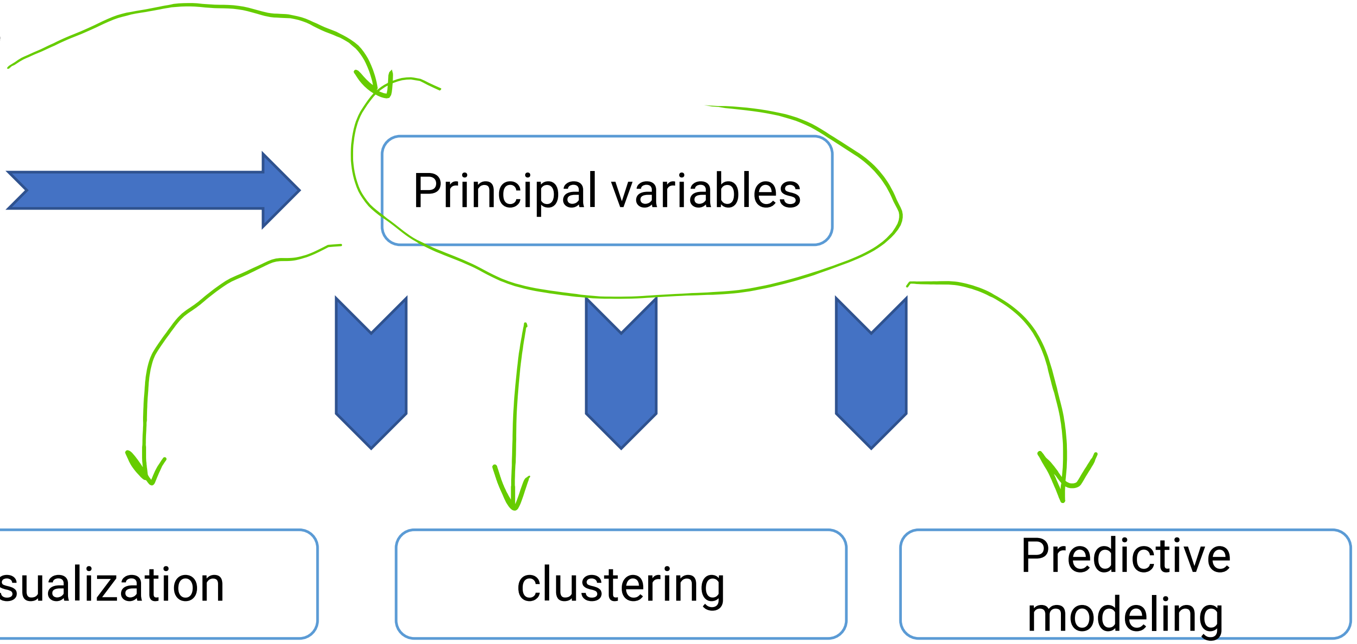
$a(i) = \text{mean}(-)$



$b(i) = \min(\text{mean}(-), \text{mean}(-))$

Dimensionality reduction techniques





Principal variables

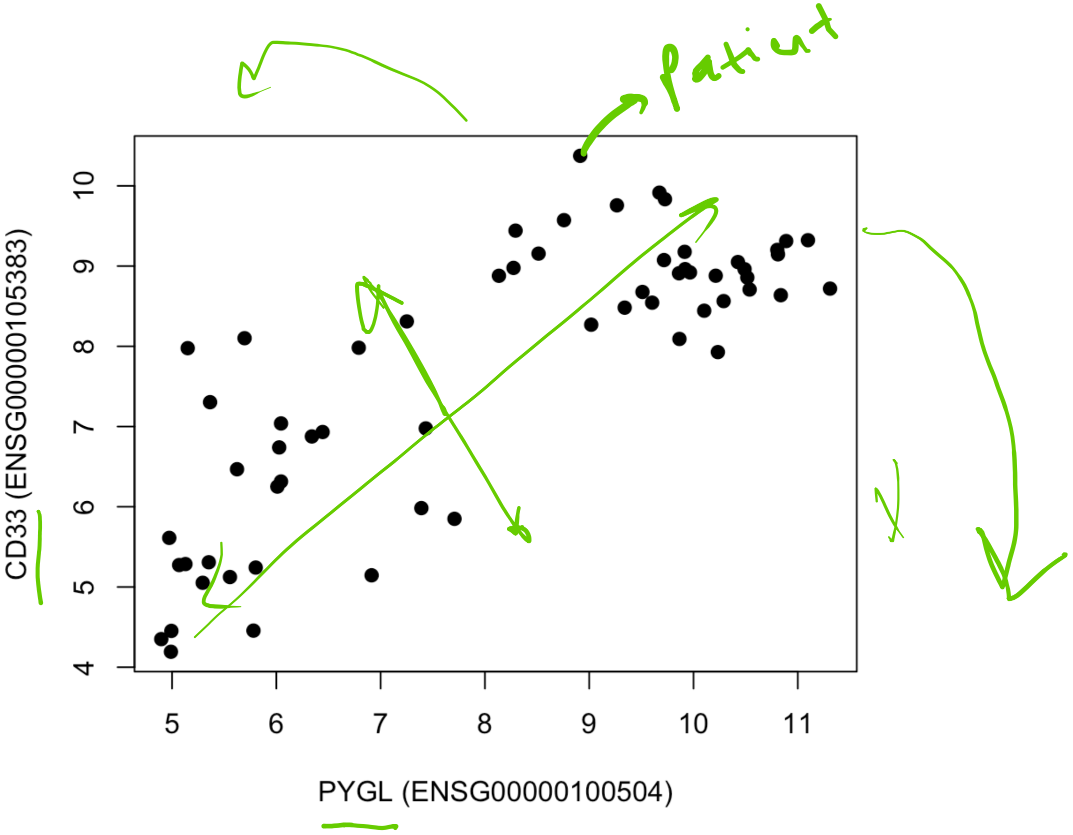
visualization

clustering

Predictive modeling

Dimensionality reduction techniques

Principal component analysis (PCA)

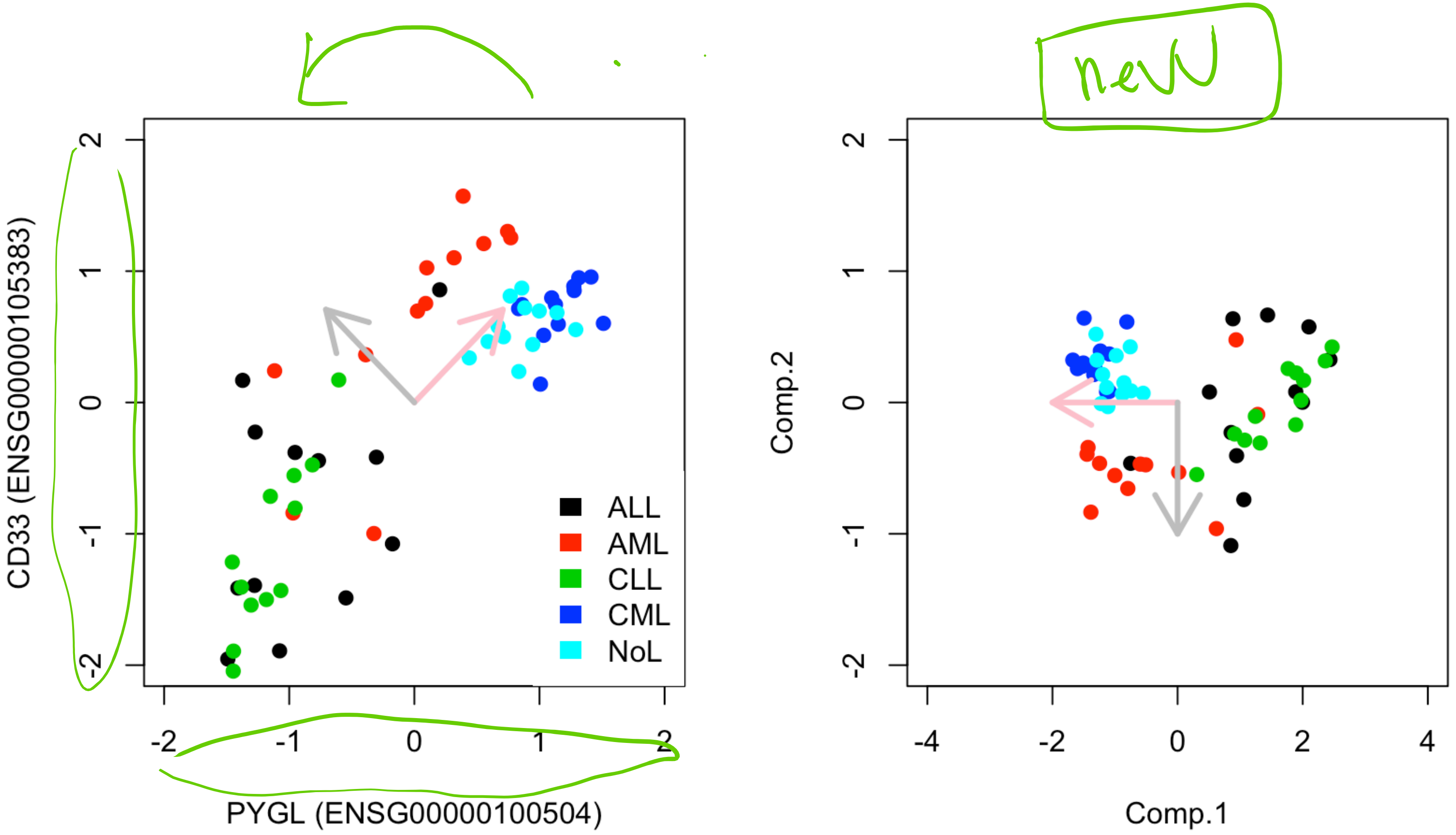


princomp()

scale()

Dimensionality reduction techniques

Principal component analysis (PCA)



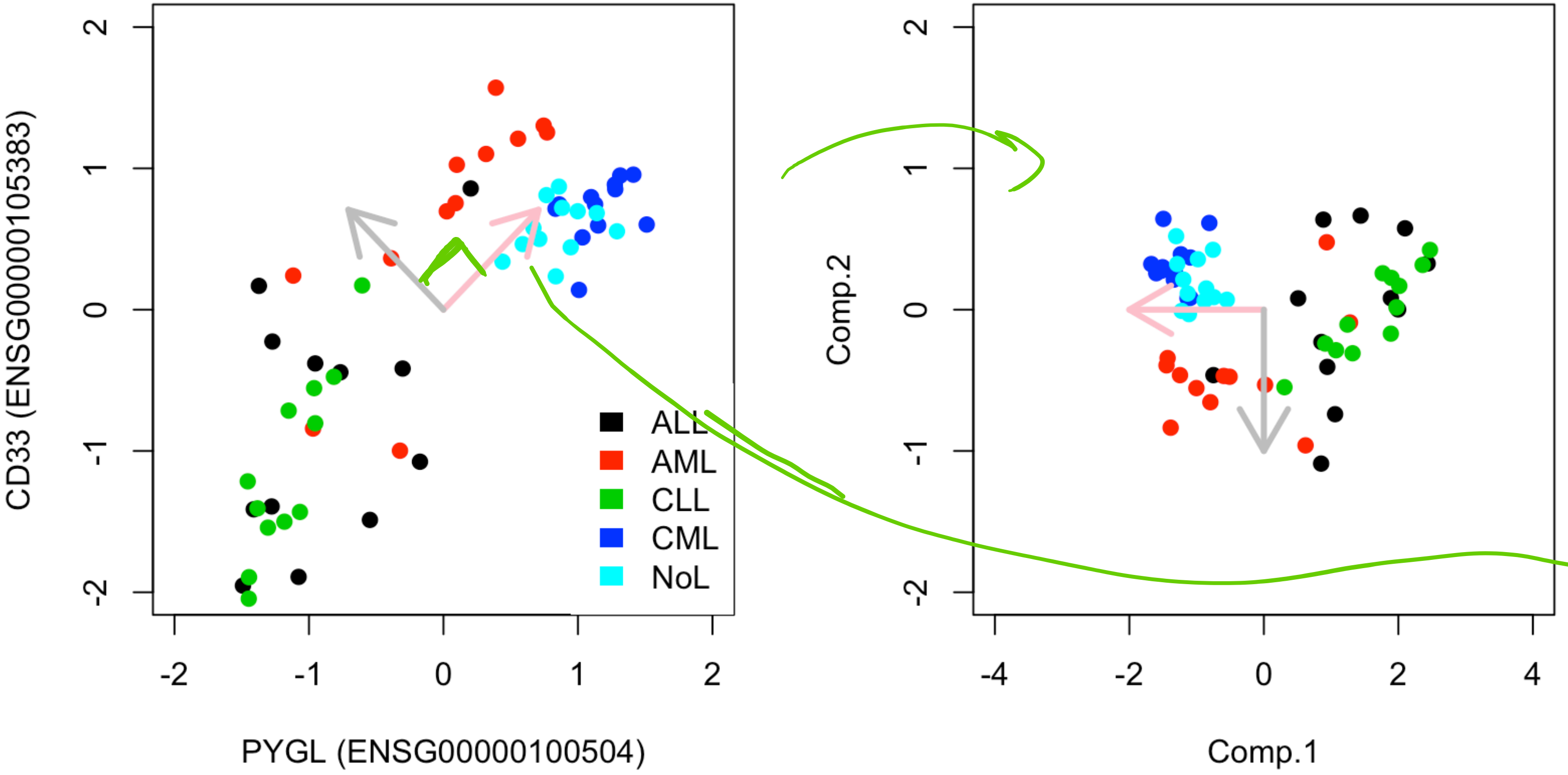
x̄ = 0

sd = 1

Dimensionality reduction techniques

Principal component analysis (PCA)

$$\boxed{\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y),}$$



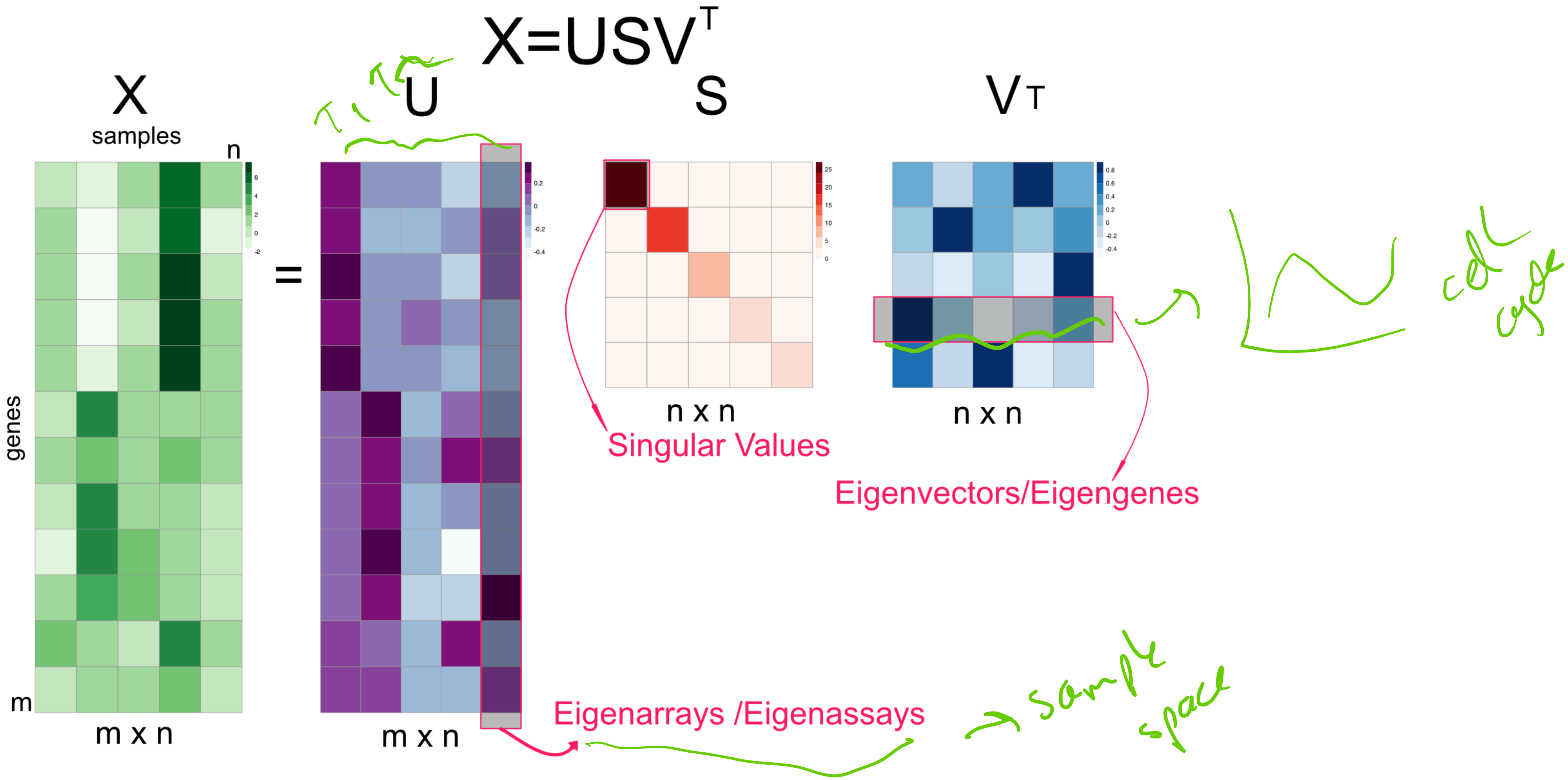
eigen
↓
cov. mat

eigenvector

eigenvalues

Dimensionality reduction techniques

PCA with singular value decomposition (SVD)



$X=USV^T$

Singular Values

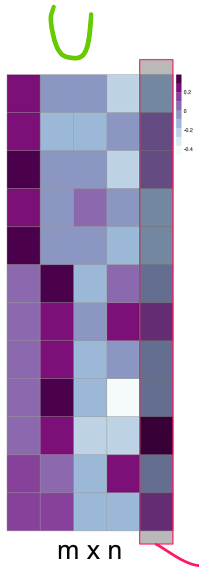
Eigenvectors/Eigengenes

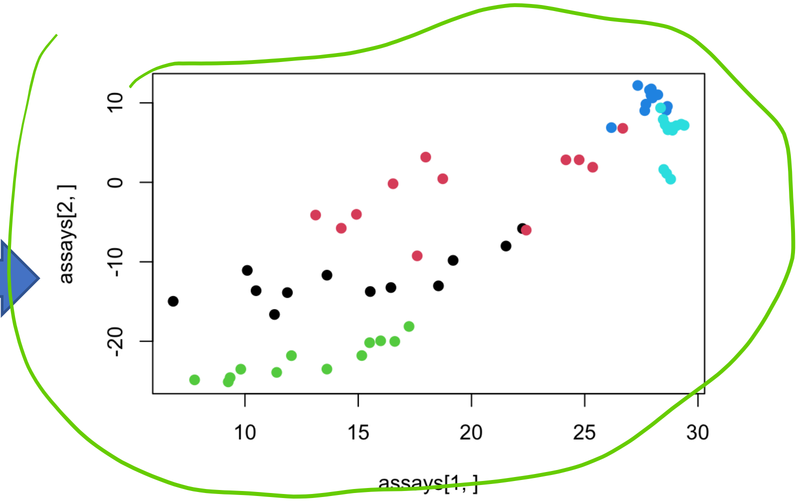
Eigenarrays /Eigenassays

Dimensionality reduction techniques

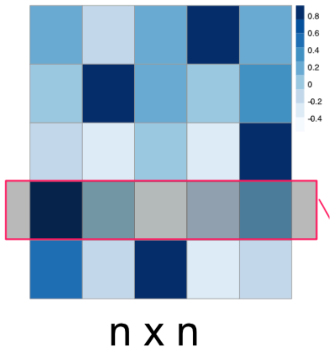
PCA with singular value decomposition (SVD)

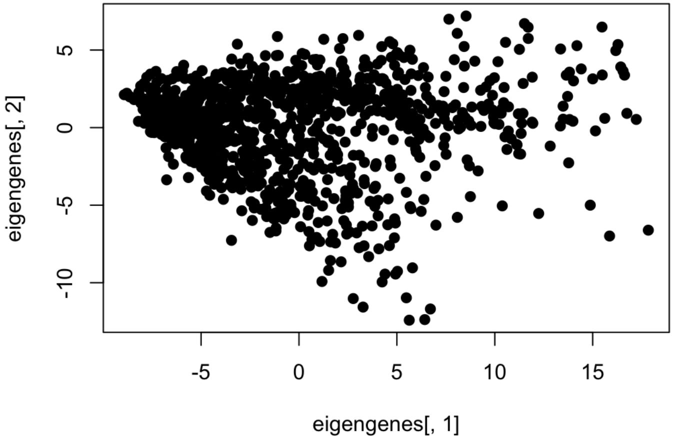
Project samples/assays eigenassays





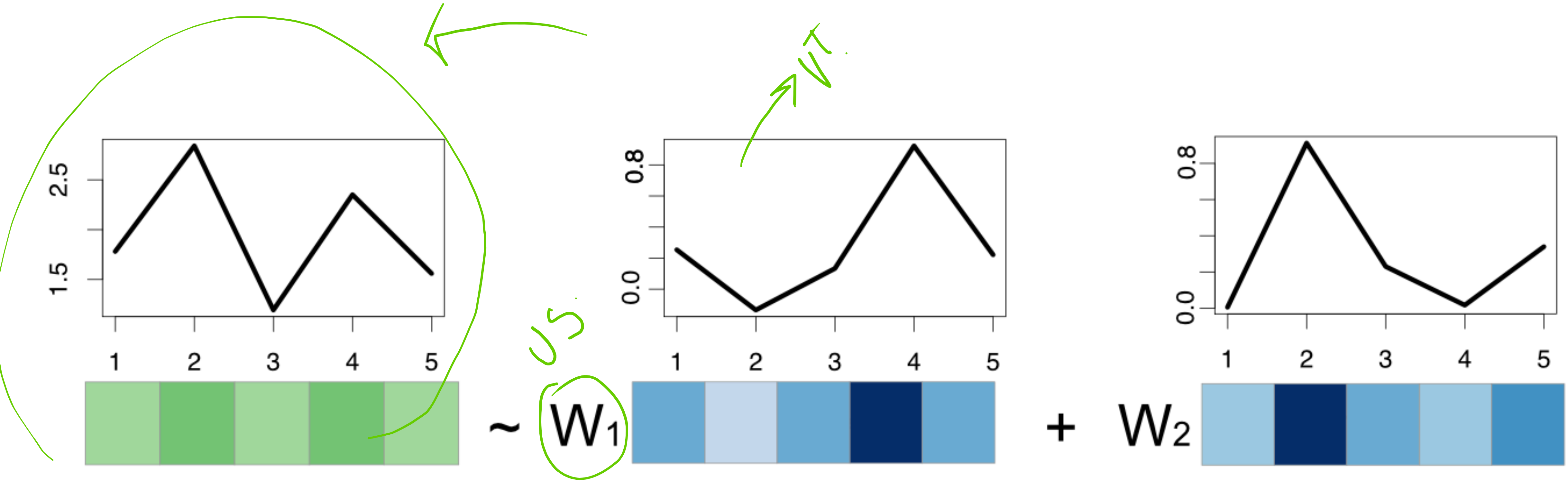
Project genes on eigengenes





Dimensionality reduction techniques

eigenvectors (latent factors) as expression programs



Gene 7 expression

Eigenvector 1

Eigenvector 2

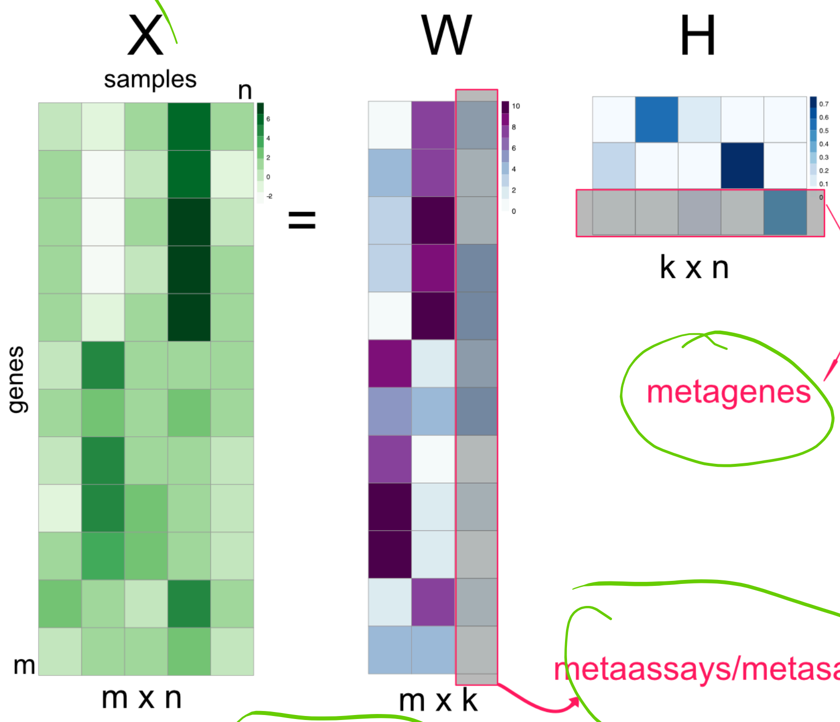
Dimensionality reduction techniques

Matrix factorization

deep learning

PCA/SVD | ICA | NMF | autoencoders

X=WH



metagenes

metaassays/metasamples

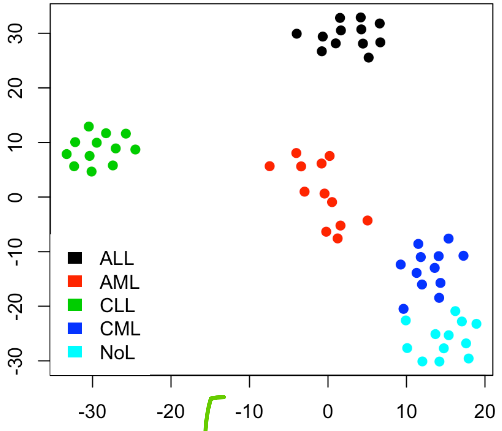
Bonus: latent factors

Distance compression

t-SNE | MDS | UMAP

Distance in high dimensions → Distance in 2D/3D

$$\begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & \cdots & d_{1n}^2 \\ d_{21}^2 & 0 & d_{23}^2 & \cdots & d_{2n}^2 \\ d_{31}^2 & d_{32}^2 & 0 & \cdots & d_{3n}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1}^2 & d_{n2}^2 & d_{n3}^2 & \cdots & 0 \end{bmatrix}$$

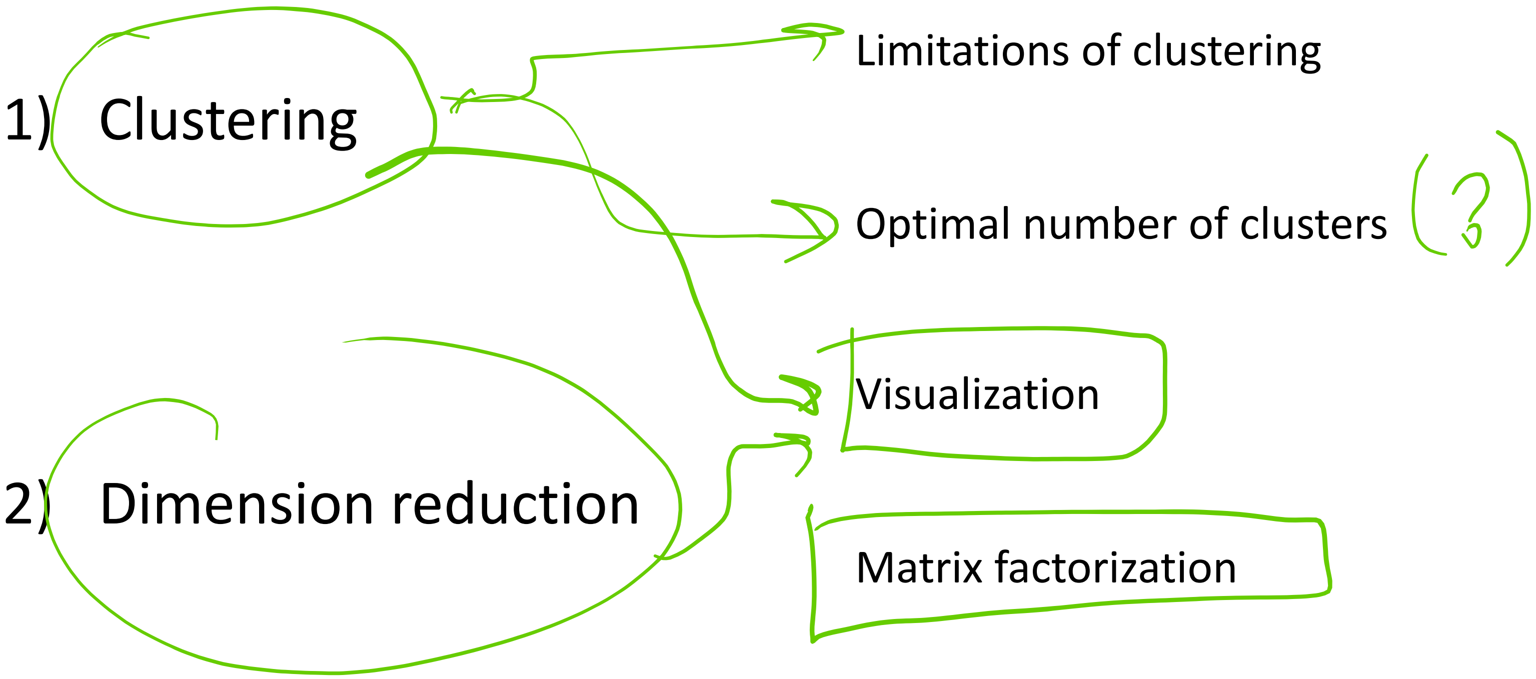


t-SNE

Unsupervised learning for Genomics:

Recap

Key concepts:



1) Clustering

2) Dimension reduction

Limitations of clustering

Optimal number of clusters (?)

Visualization

Matrix factorization