Single Cell RNA-seq Clustering

Ahmed Mahfouz
Leiden Computational Biology Center, LUMC
Delft Bioinformtaics Lab, TU Delft

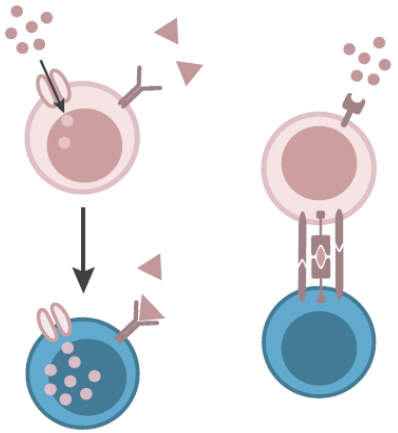
Leiden
Computational Biology Center

LUMC

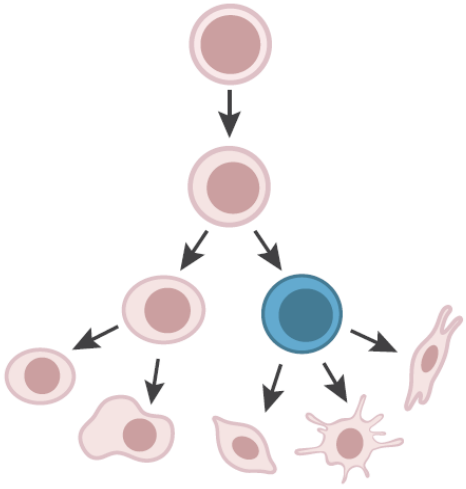
TU Delft

Cell Identity

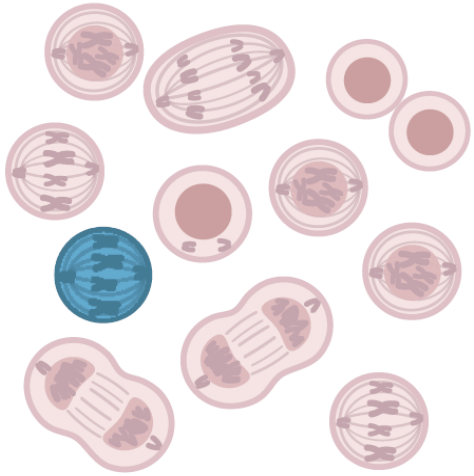
Environmental stimuli



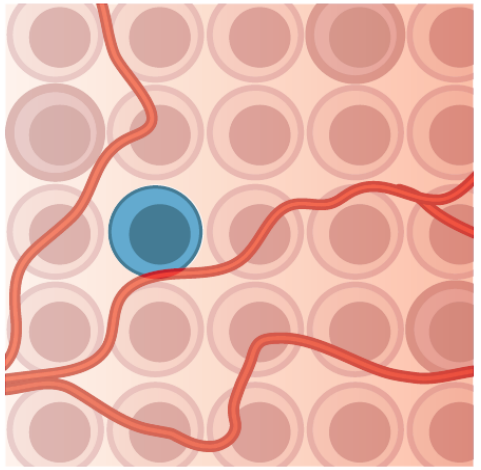
Cell development



Cell cycle

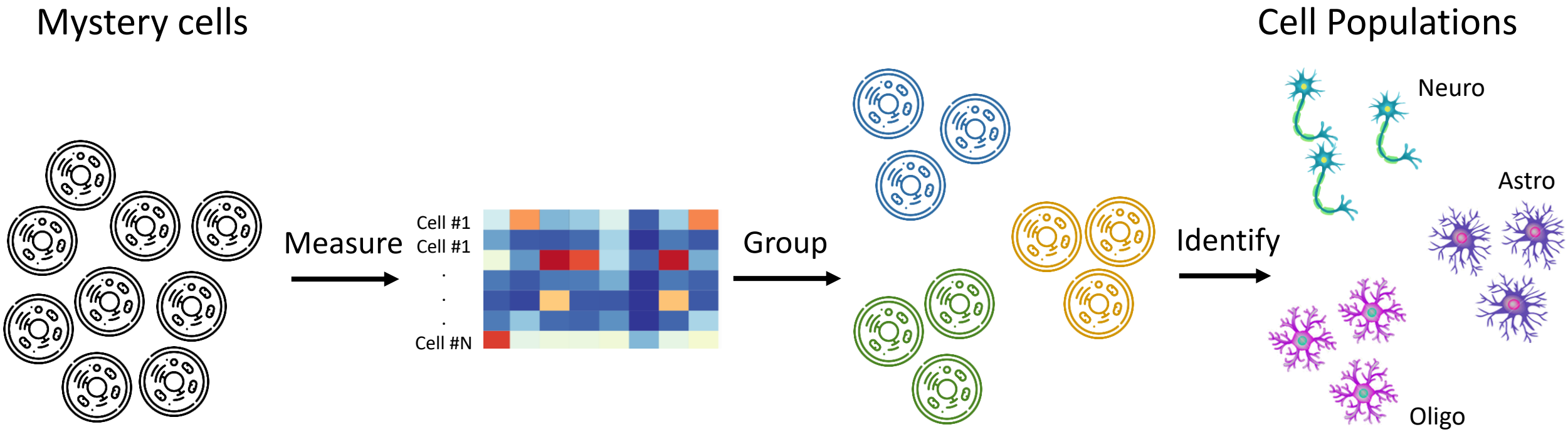


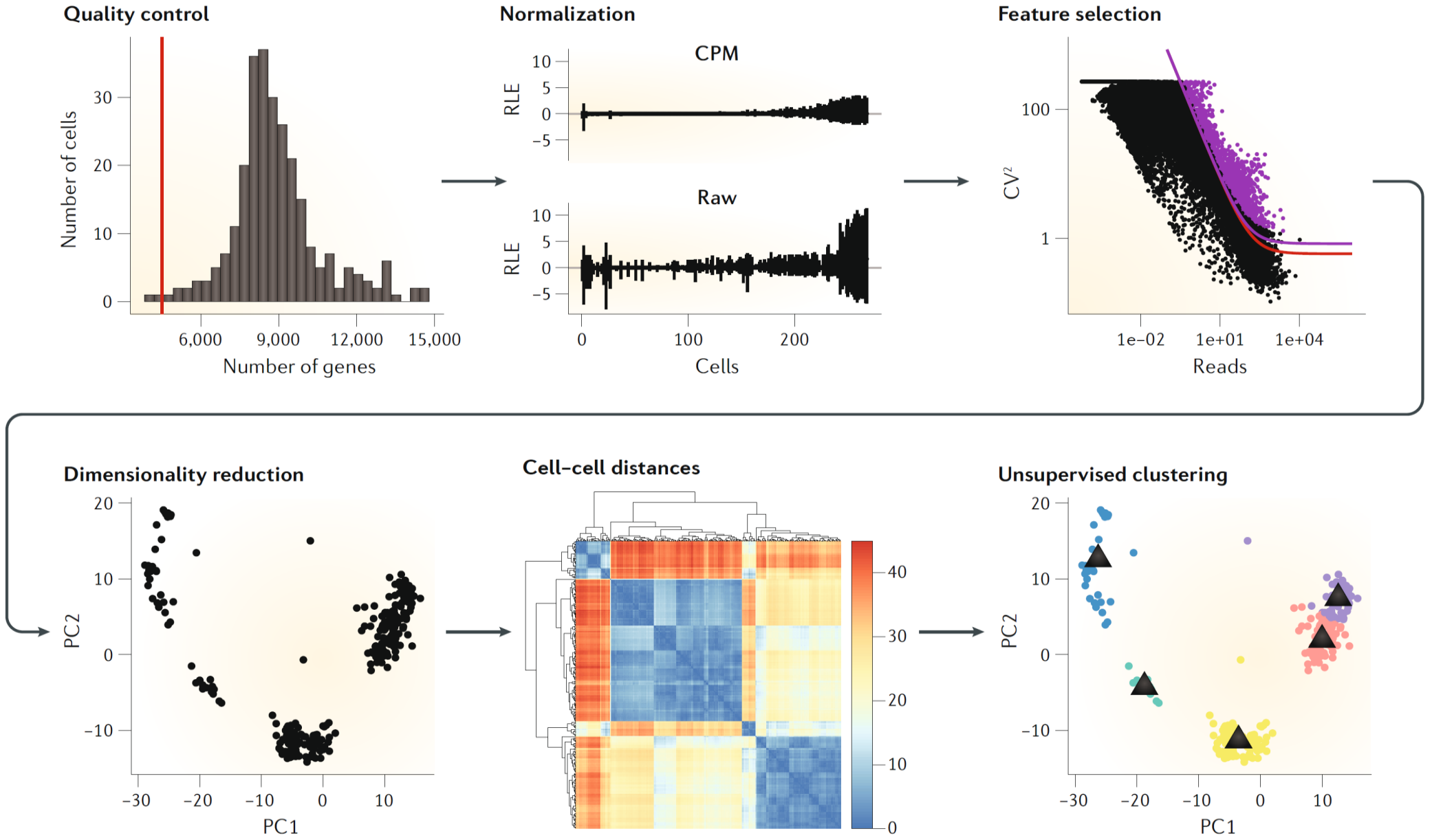
Spatial context



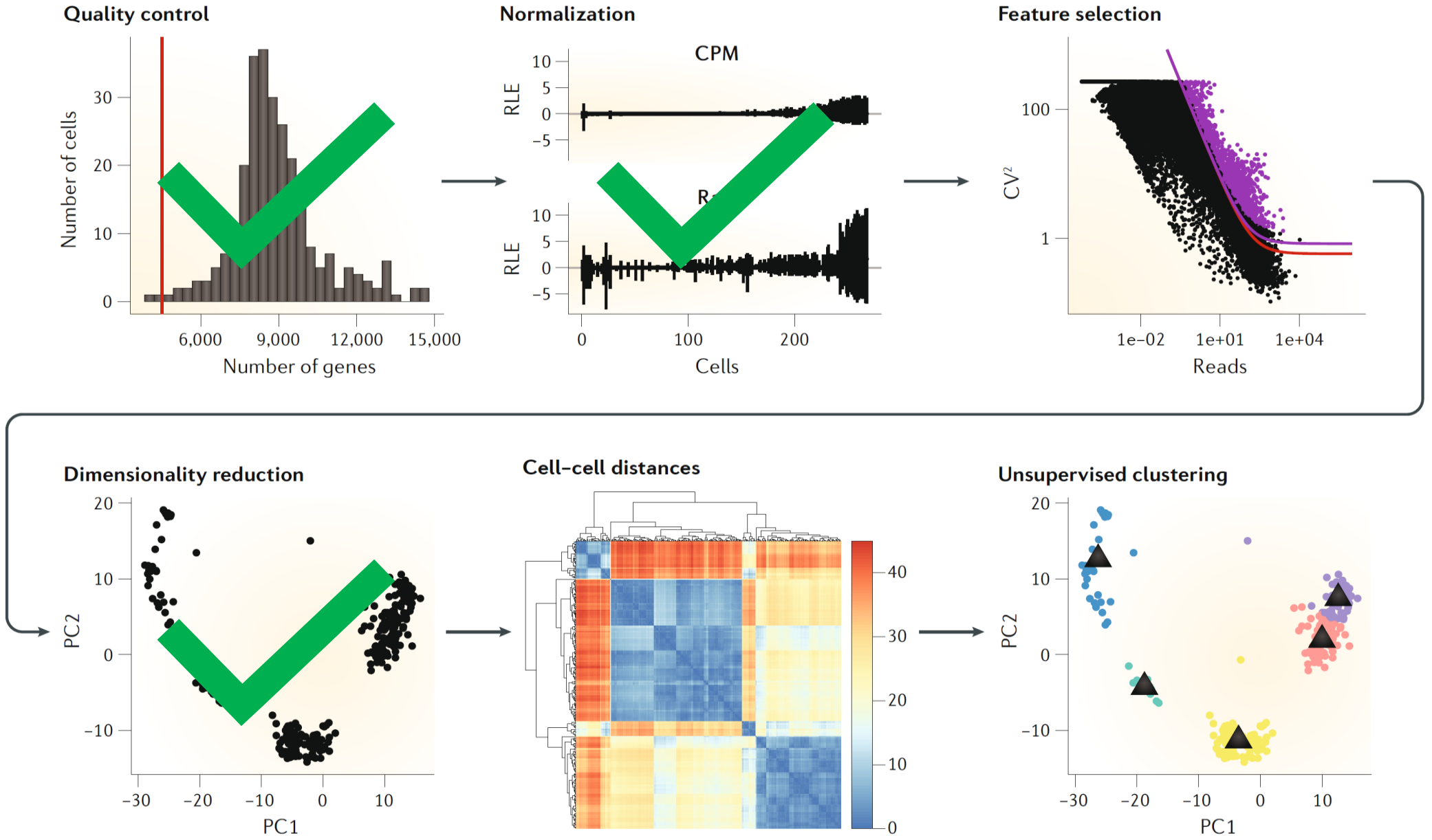
Wagner et al. (https://doi.org/10.1038/nbt.3711)

How can we identify cell populations?





Kiselev et al. (https://doi.org/10.1038/s41576-018-0088-9)



Kiselev et al. (https://doi.org/10.1038/s41576-018-0088-9)

Outline

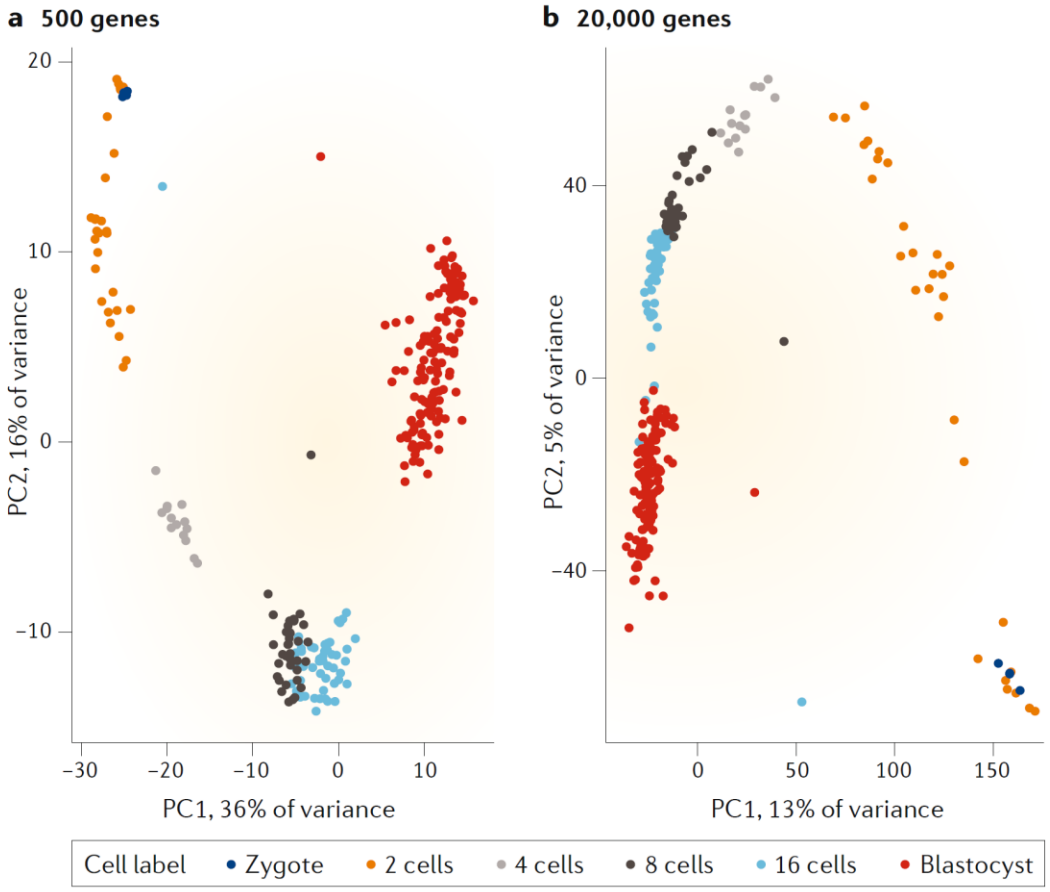
- Feature selection
- Introduction to clustering
  - Hierarchical clustering
  - *k*-Means clustering
  - Graph-based clustering
- scRNA-seq clustering
  - Single Cell Consensus Clustering (SC3)
  - Seurat
- Validation

Feature selection

- Curse of dimensionality:

More features (genes) -> smaller distances between samples (cells)

- Remove genes which only exhibit technical noise
  - Increase the signal:noise ratio
  - Reduce the computational complexity



Kiselev et al. (https://doi.org/10.1038/s41576-018-0088-9)

Feature selection

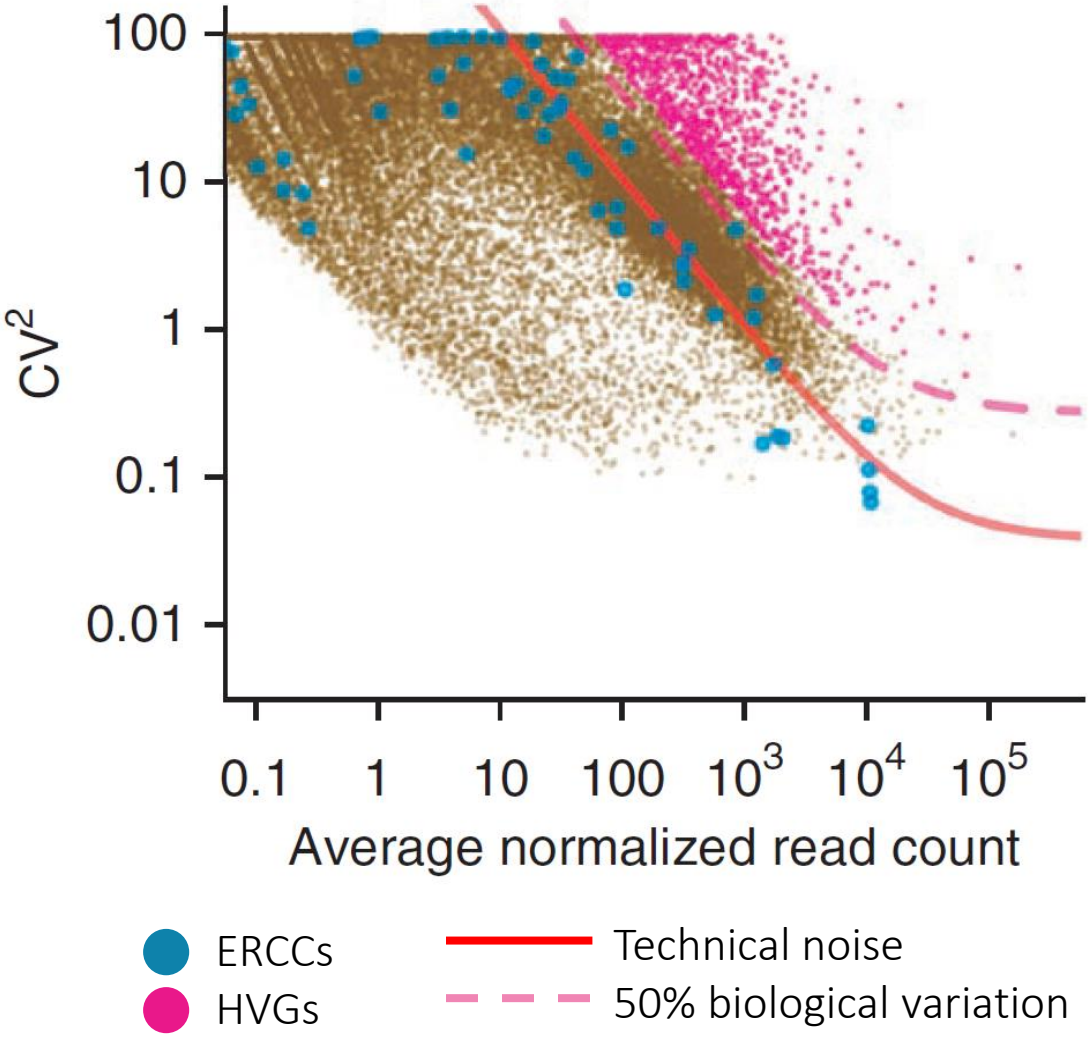
Highly Variable Genes (HVG)

- $CV = \frac{var}{mean} = \frac{\sigma}{\mu}$

- Fit a gamma generalized linear model

- No ERCCs?

  -> estimate technical noise based on all genes



Brennecke et al. (https://doi.org/10.1038/nmeth.2645)

Feature selection

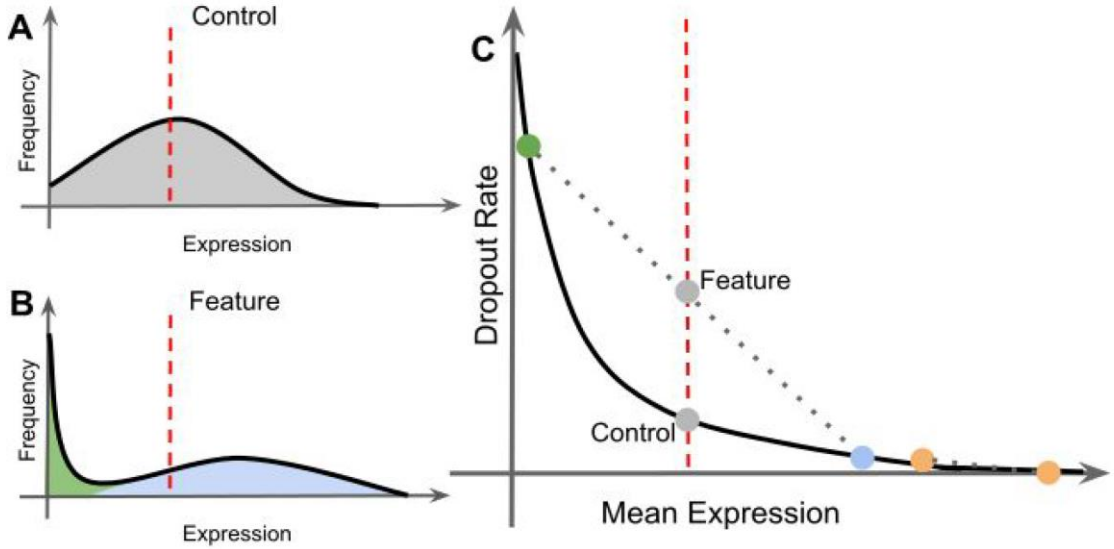
M3Drop: Dropout-based feature selection

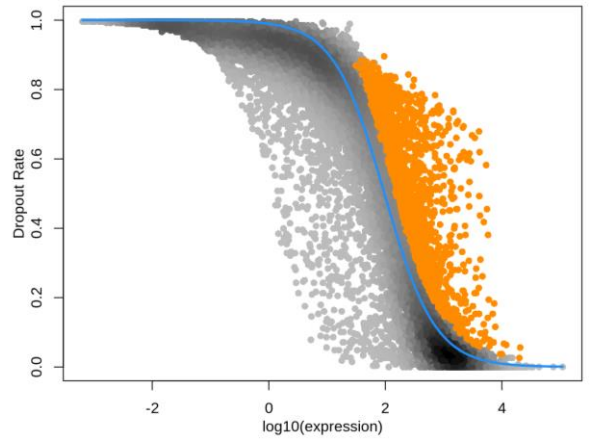
- Reverse transcription is an enzyme reaction thus can be modelled using the Michaelis-Menten equation:

$$P_{dropout} = 1 - \frac{S}{K_M + S}$$

S: average expression

K_M: Michaelis-Menten constant

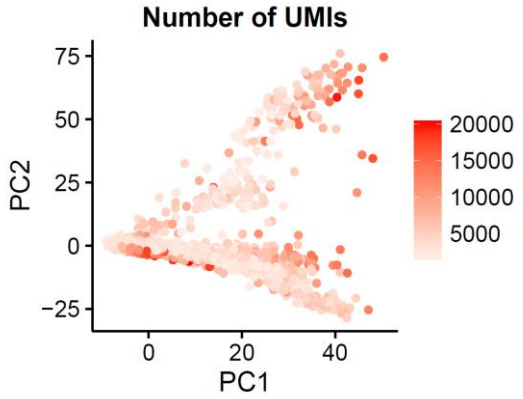


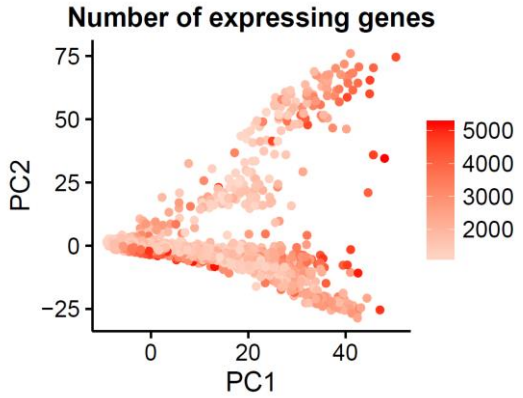


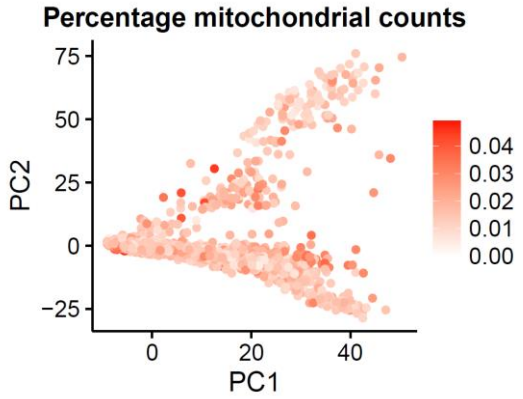
Brennecke et al. (https://doi.org/10.1038/nmeth.2645)

Dealing with confounders

- Confounding factors:
  - number of detected molecules
  - number of expressing genes
  - mitochondrial gene expression
  - cell cycle
  - …
- Solution: use a linear model to regress them out





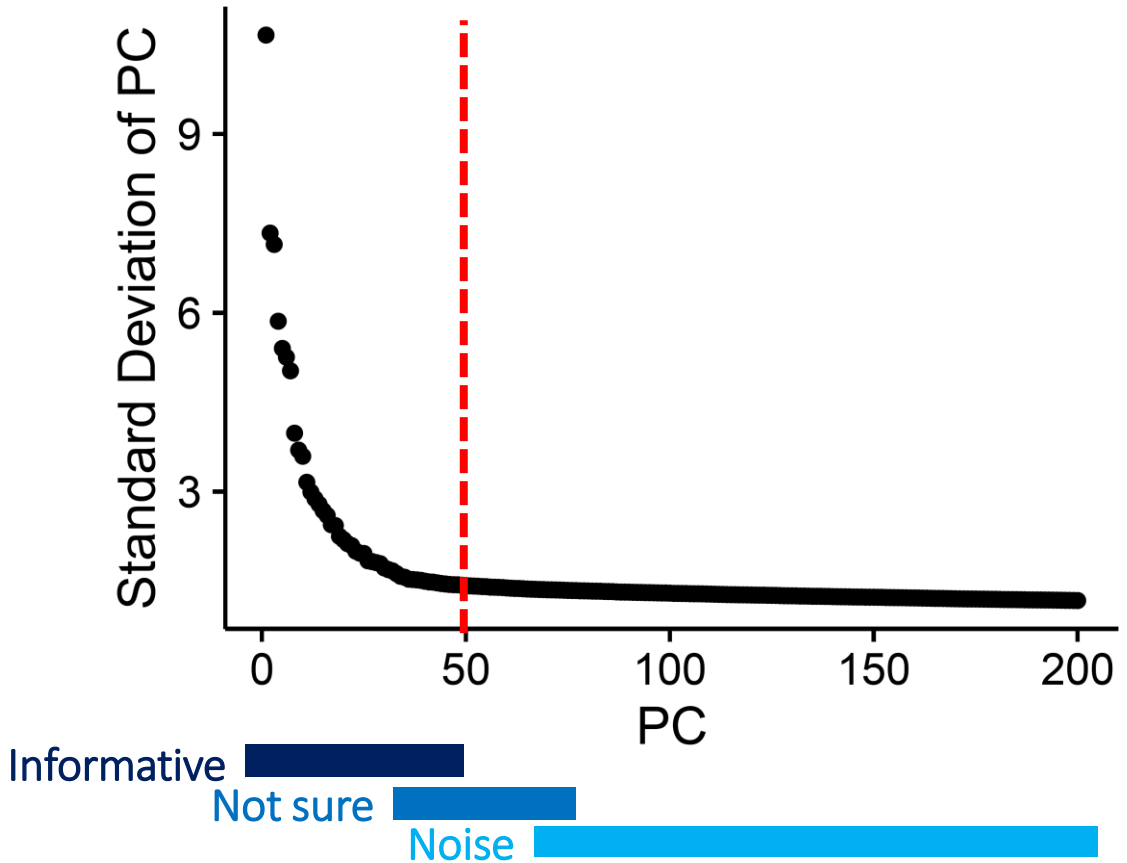


Selecting principal components

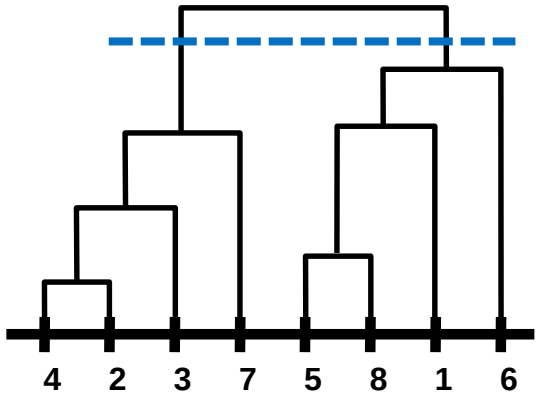
- To overcome the extensive technical noise in scRNA-seq data, it is common to cluster cells based on their PCA scores

- Each PC represents a ‘metagene’ that (linearly) combines information across a correlated gene set

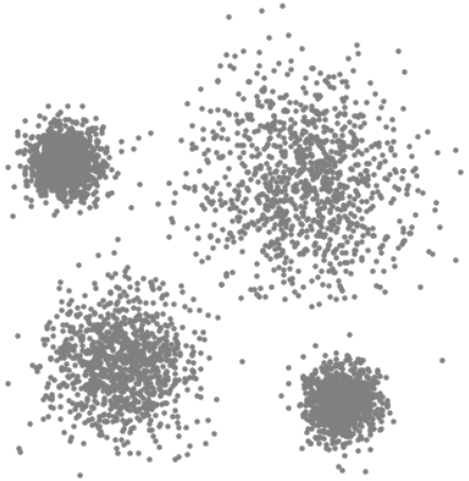
Scree/Elbow plot



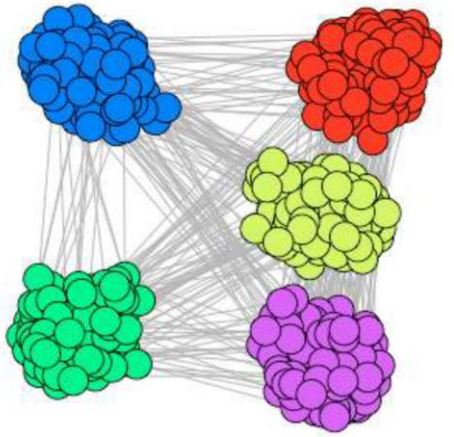
Many clustering approaches



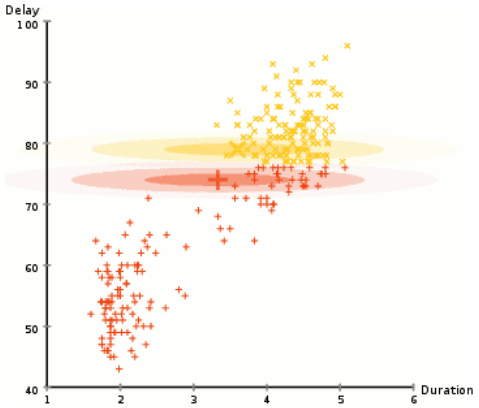
Hierarchical Clustering



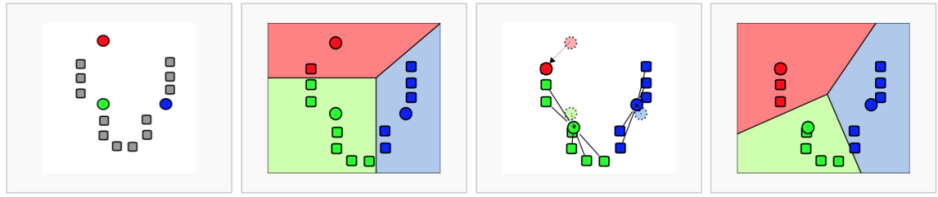
Mean shift clustering



Graph-based clustering

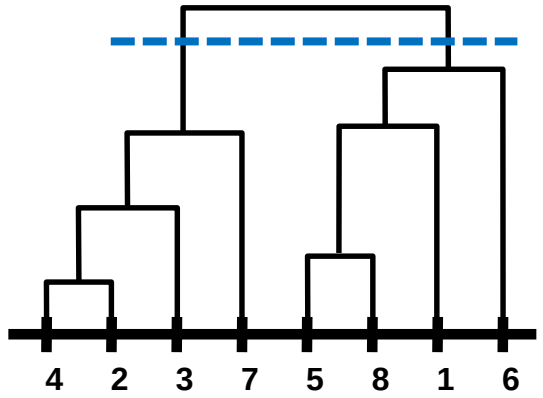


Gaussian mixture modeling

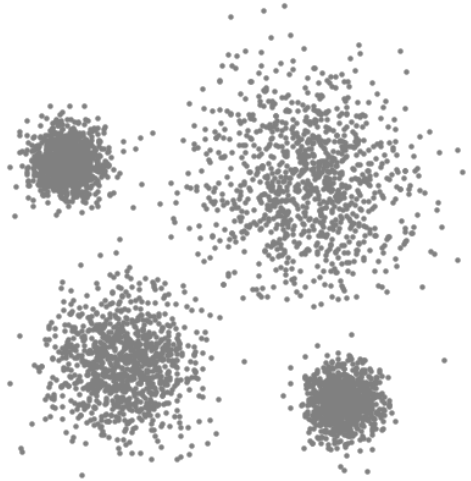


k-means clustering

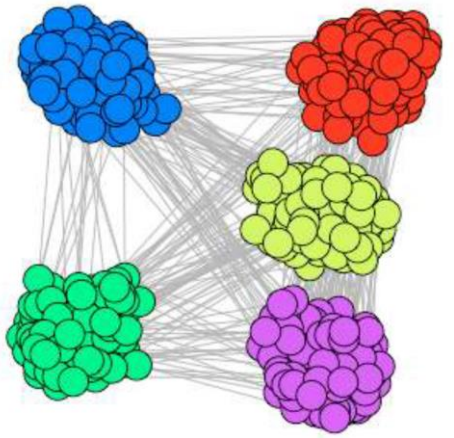
Many clustering approaches



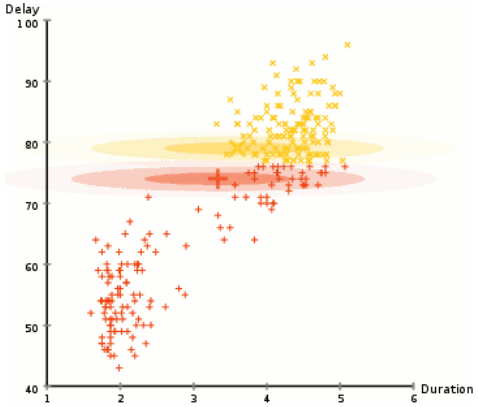
Hierarchical Clustering



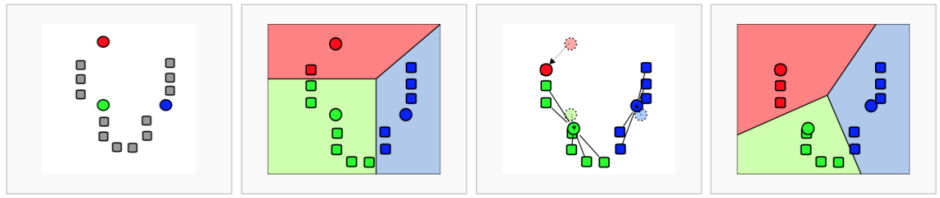
Mean shift clustering



Graph-based clustering

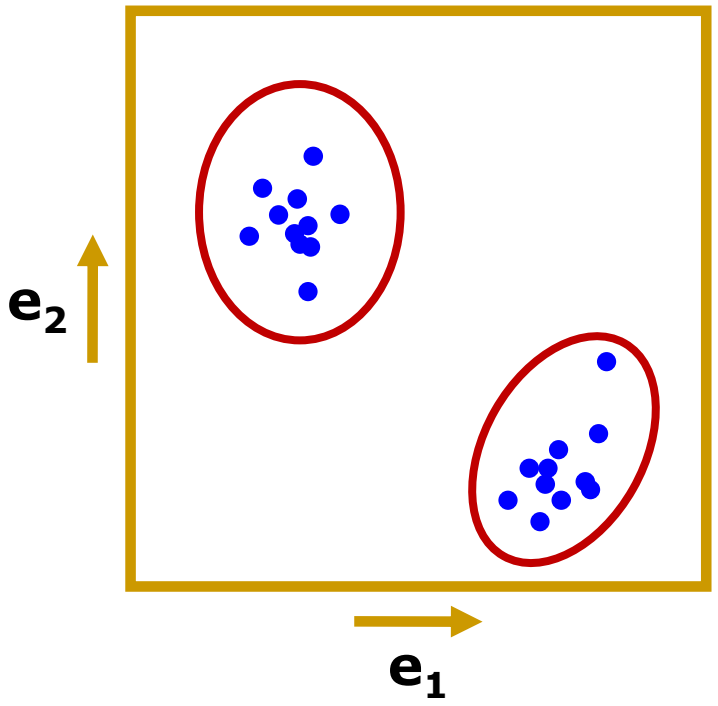


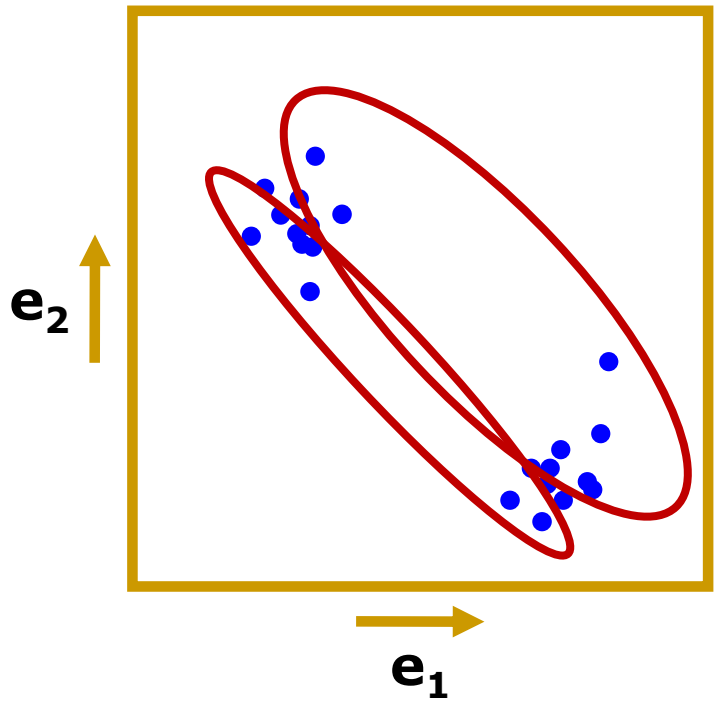
Gaussian mixture modeling



k-means clustering

Clustering

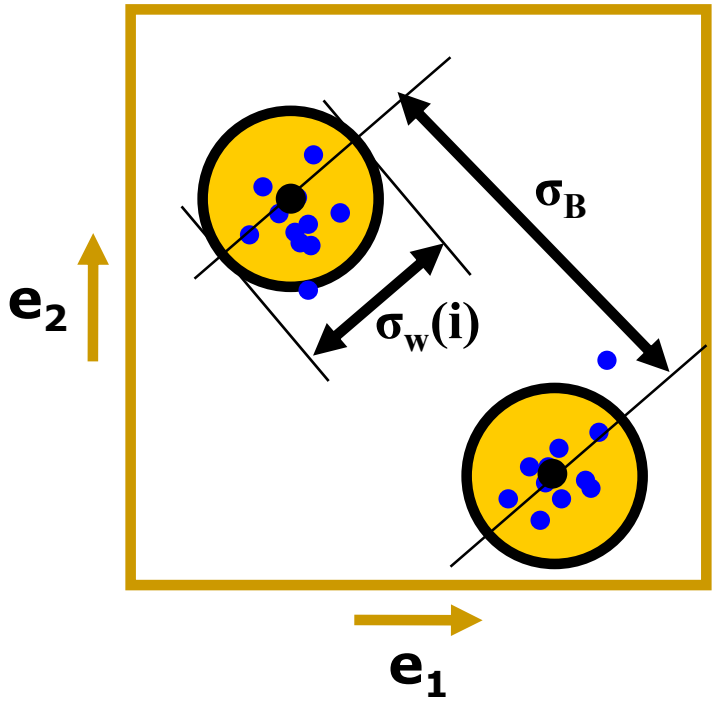




Clustering

- Structure when:

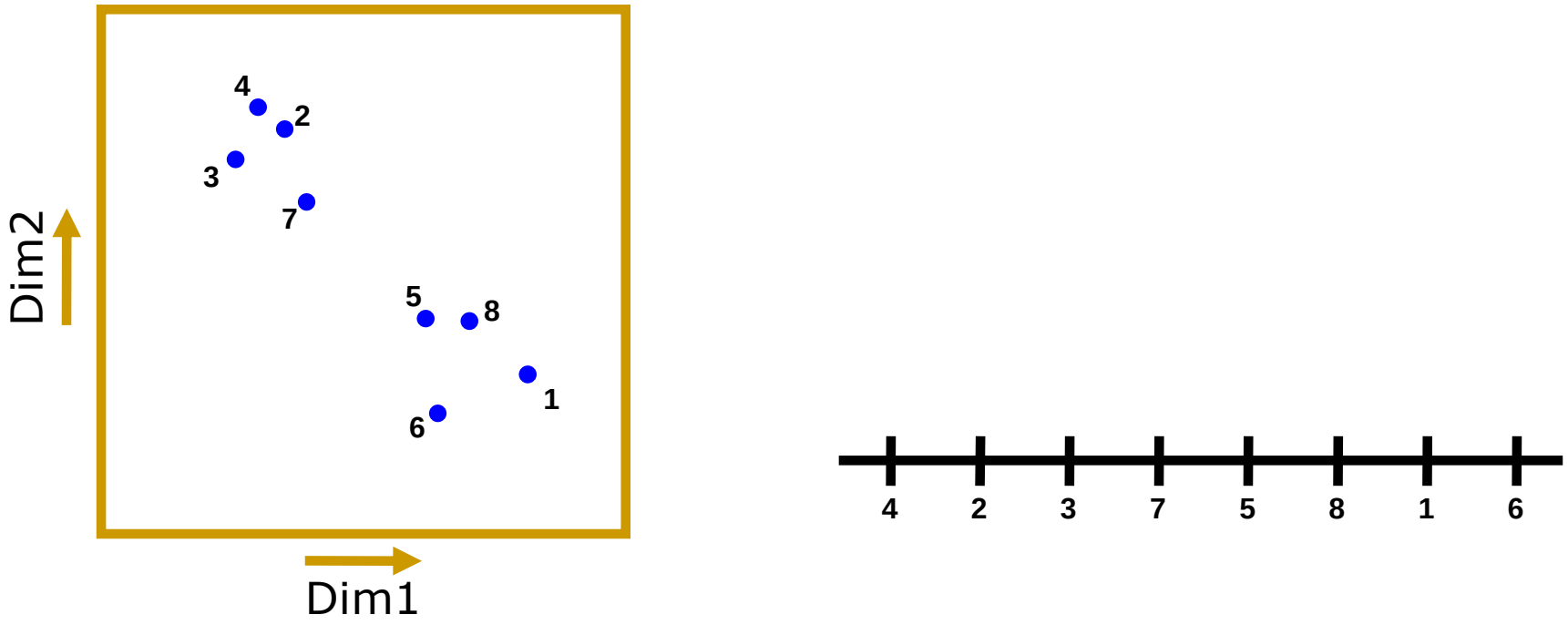
1) Samples within cluster resemble each other *(within variance, $\sigma_W(i)$)*
2) Clusters deviate from each other *(between variance, σ_B)*



- Group samples such that:

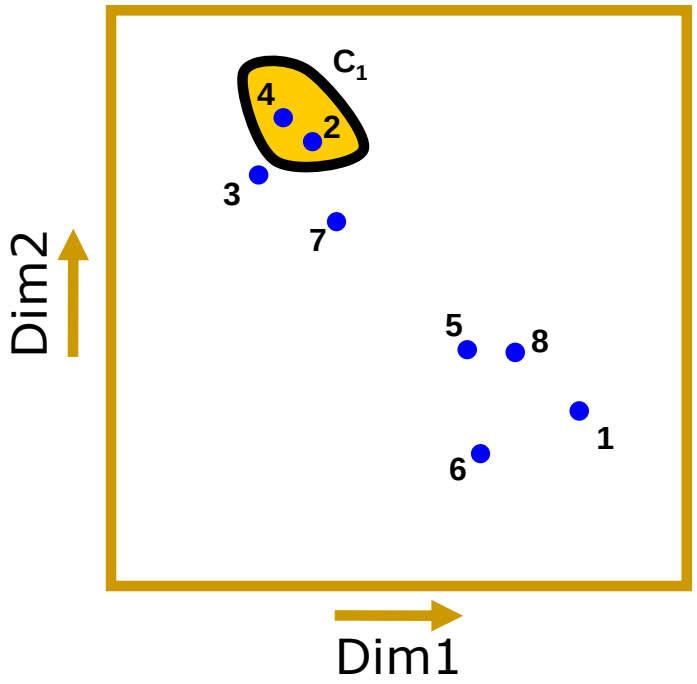
$$\min\left(\frac{\sum_{\forall clusters}\sigma_W(i)}{\sigma_B}\right) \rightarrow \begin{matrix}\sigma_W\text{: small \&} \\ \sigma_B\text{ : large}\end{matrix}$$

Hierarchical clustering

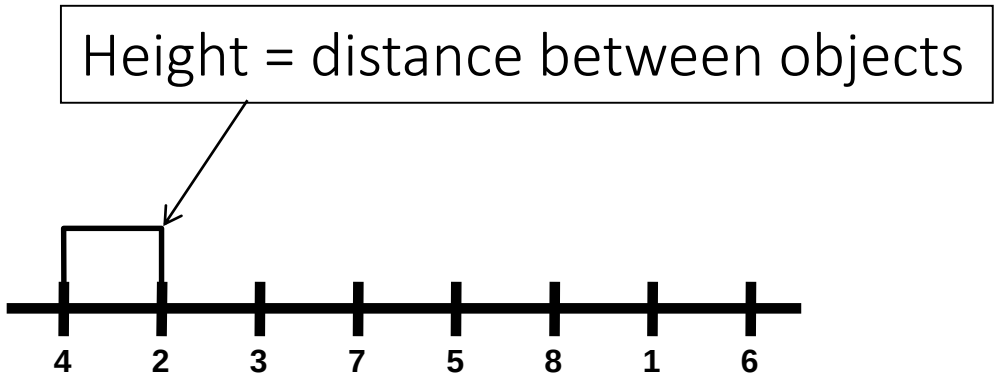


Find most similar objects (genes) and group them

Hierarchical clustering



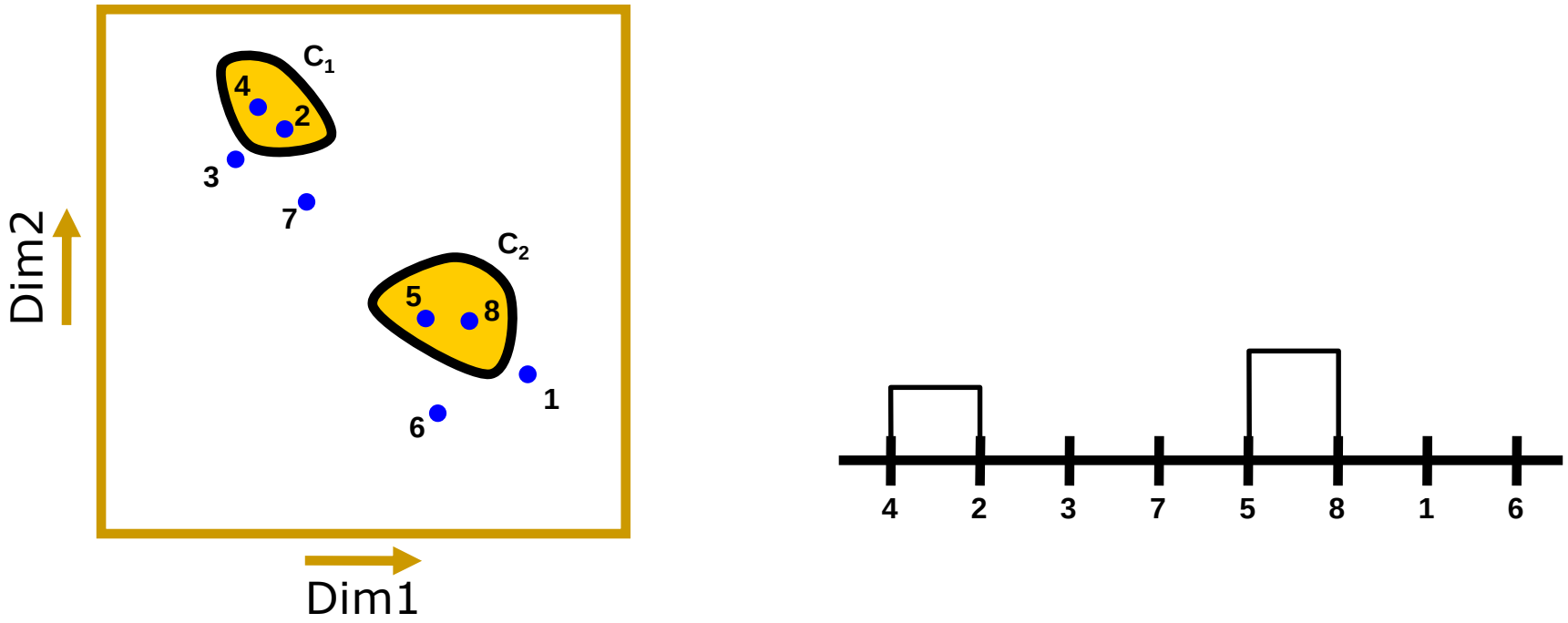
dendrogram



These are: objects 4 and 2
Again, find most similar objects (genes or clusters) and group them

Hierarchical clustering

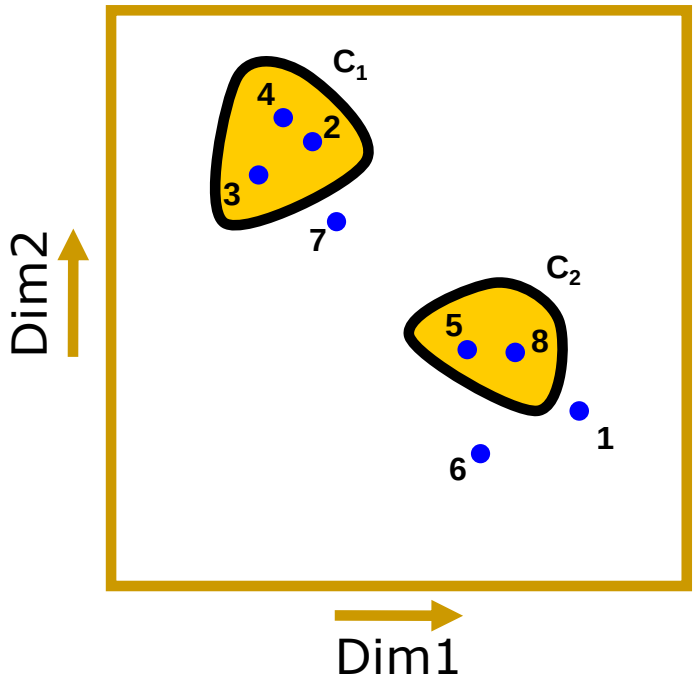
dendrogram



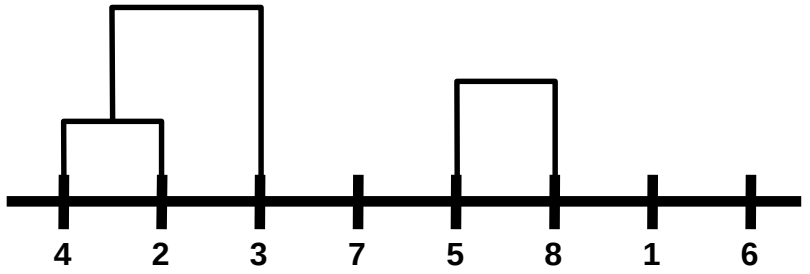
These are: objects 5 and 8

Repeat finding most similar objects (genes or clusters) and grouping them

Hierarchical clustering

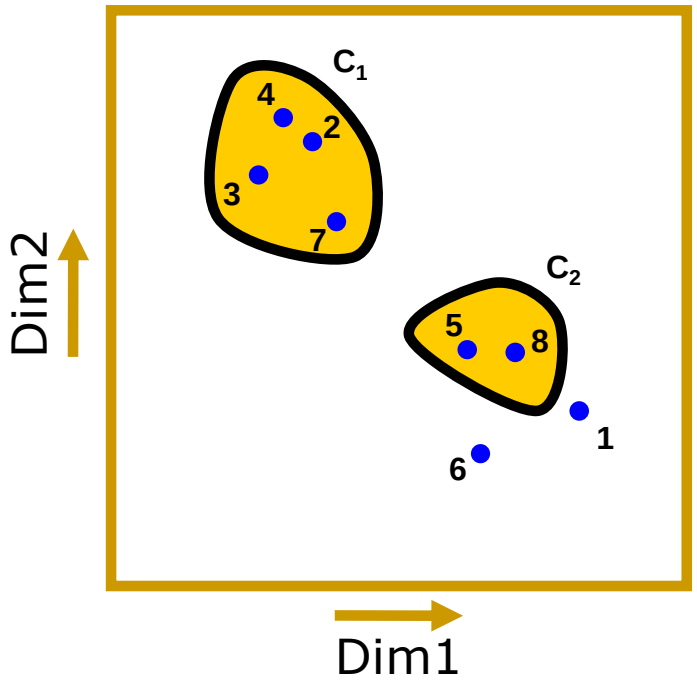


dendrogram

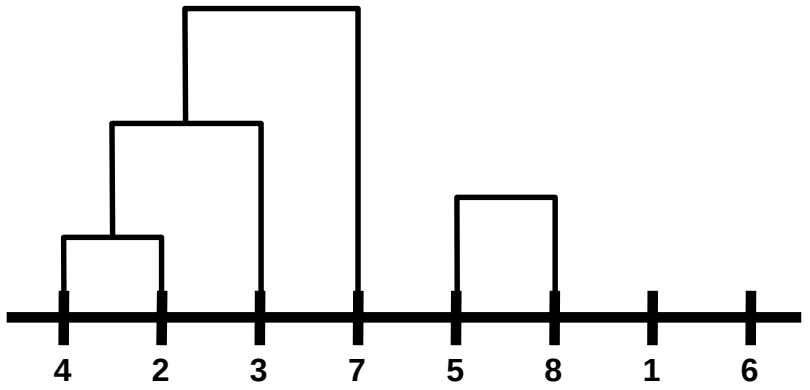


Join object 3 and cluster 1
Repeat process

Hierarchical clustering

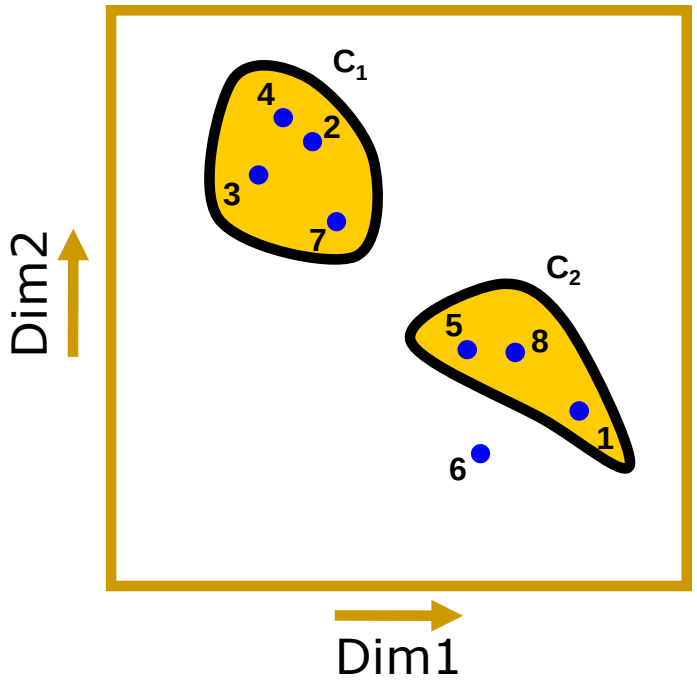


dendrogram

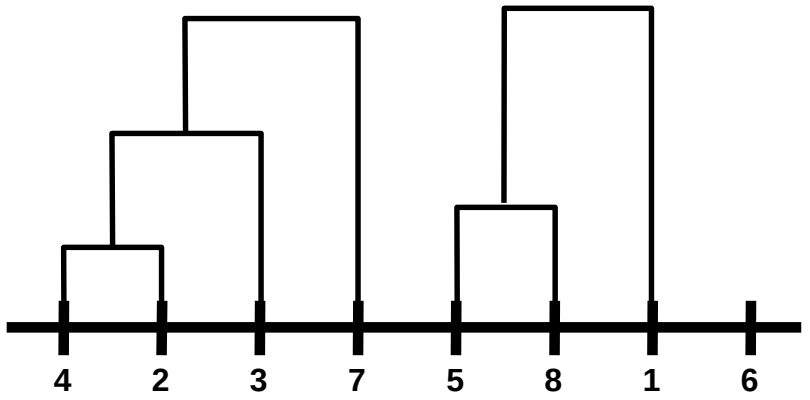


Join [object 7 and cluster 1] -> [cluster 1]
Repeat process

Hierarchical clustering

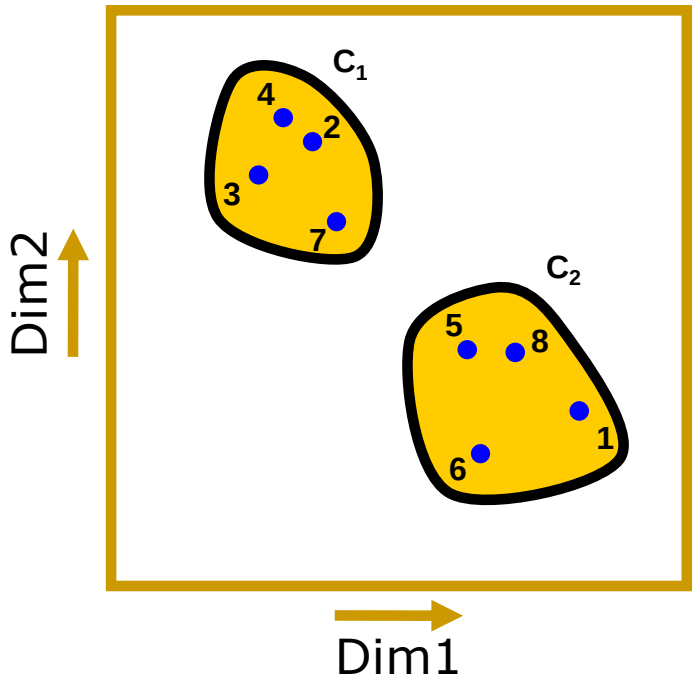


dendrogram

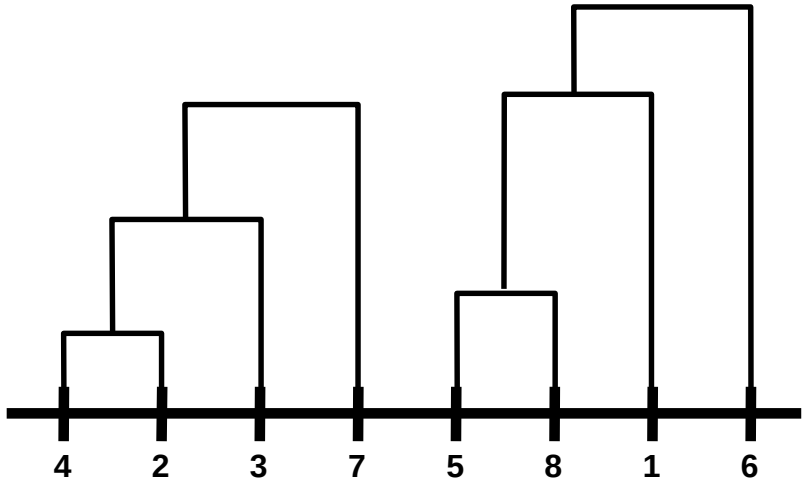


Join [object 1 and cluster 2] -> [cluster 2]
Repeat process

Hierarchical clustering

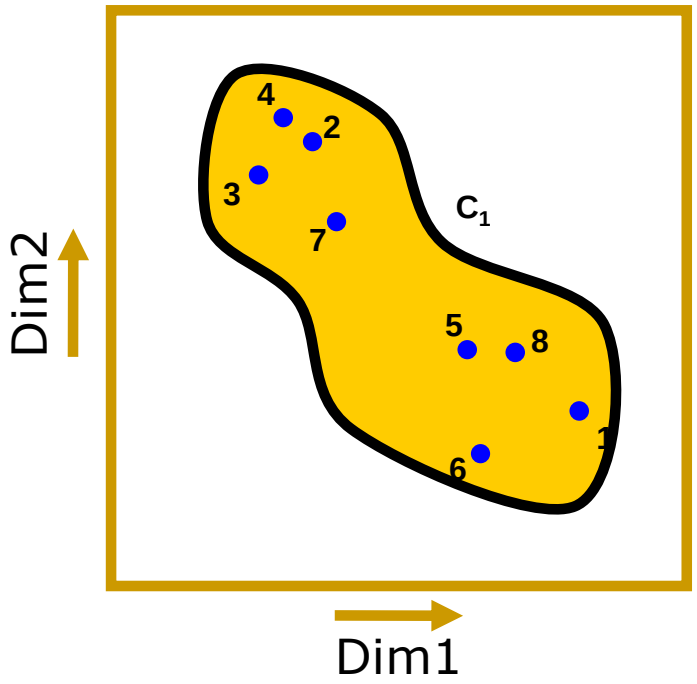


dendrogram

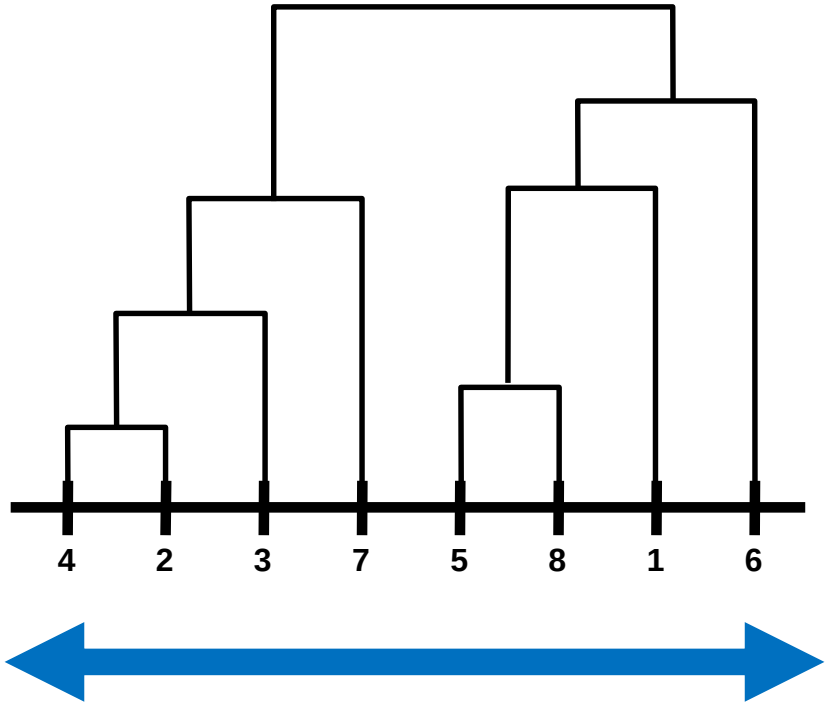


Join [object 6 and cluster 2] -> [cluster 2]
Repeat process

Hierarchical clustering

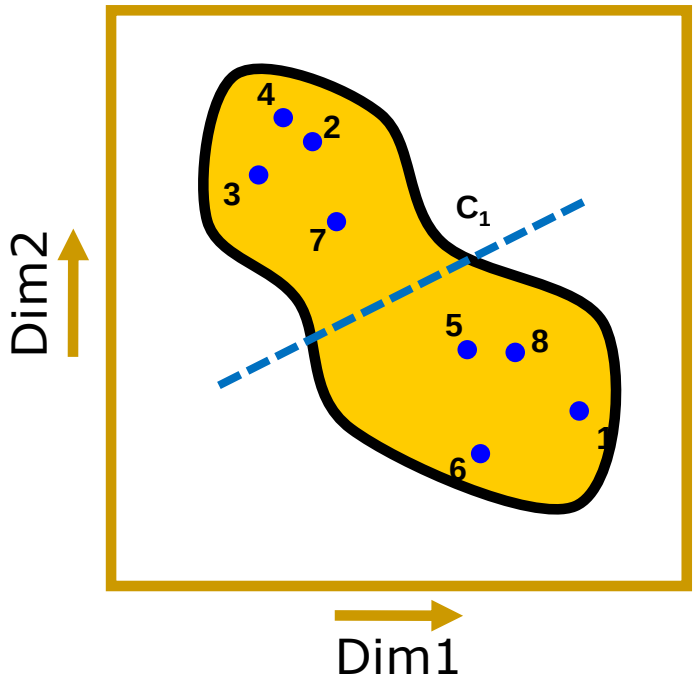


dendrogram

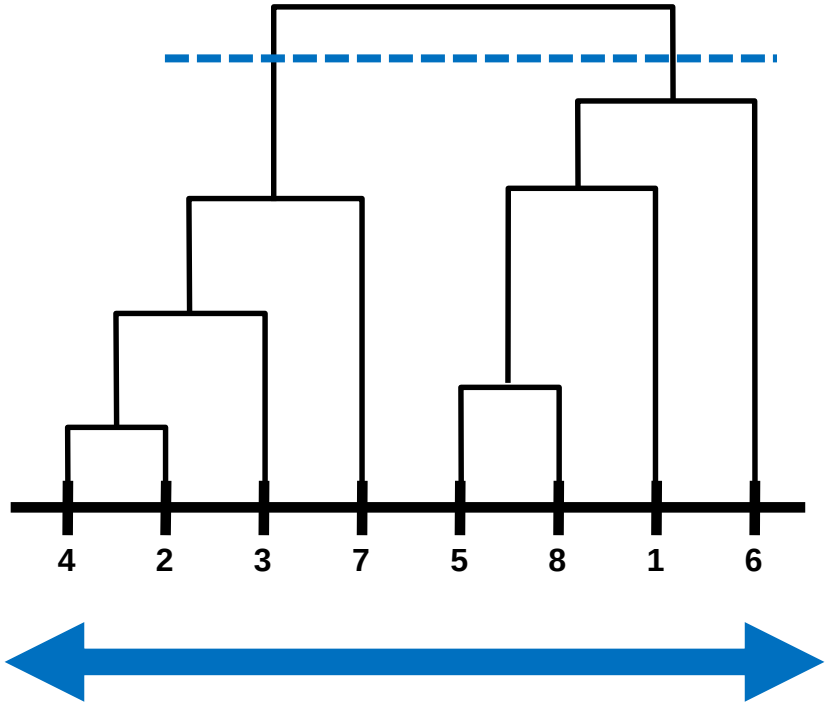


Join [cluster 1 and cluster 2] -> [cluster 1]
All in one cluster: FINISHED!

Hierarchical clustering



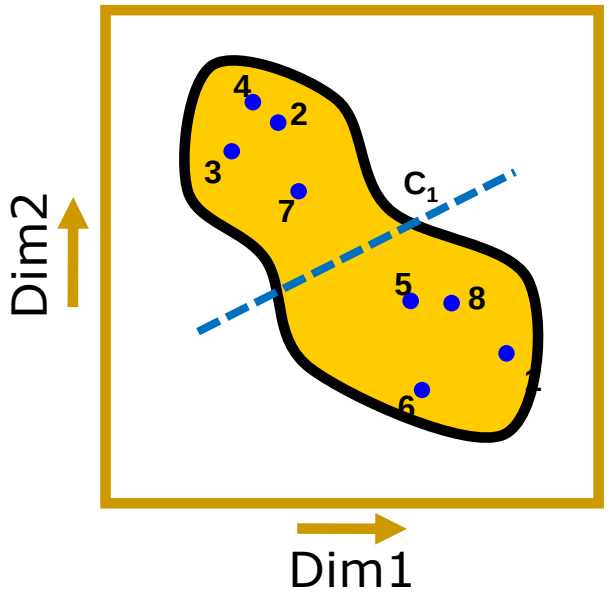
dendrogram

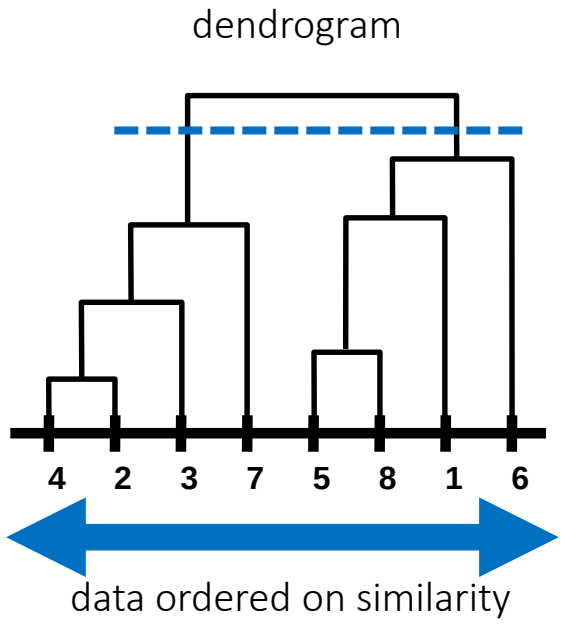


Hierarchical clustering

Need to know:

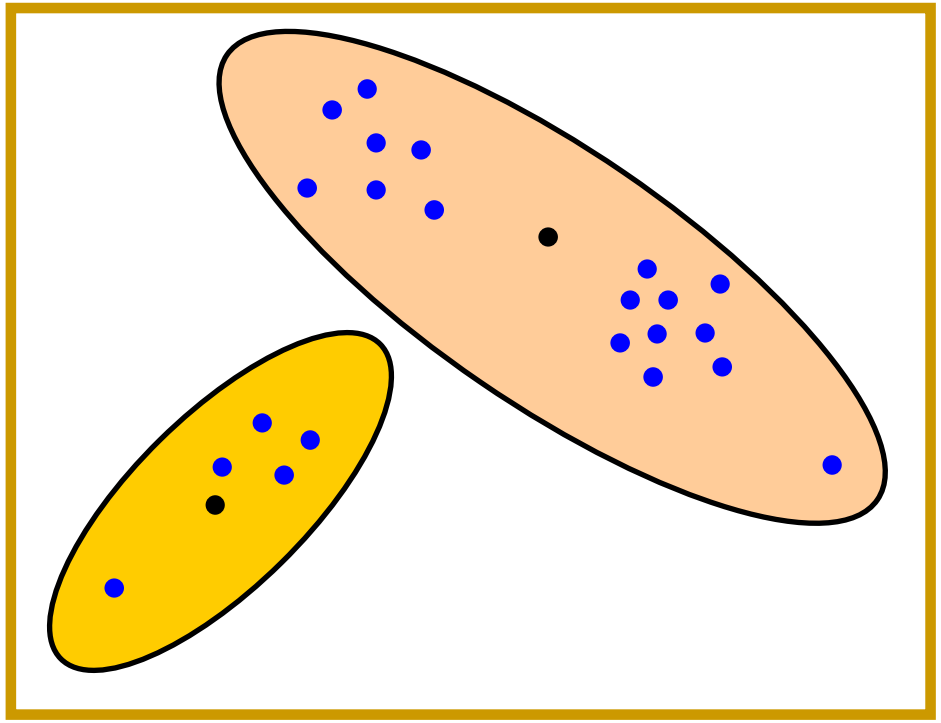
- Similarity between objects
- Similarity between clusters





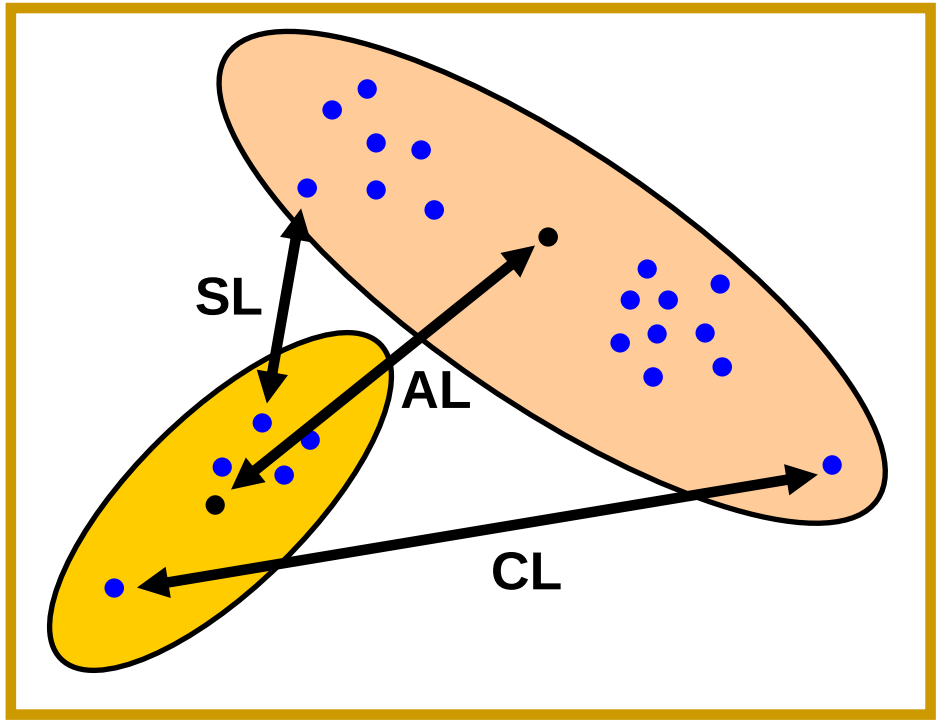
Hierarchical clustering

Similarity between clusters



Hierarchical clustering

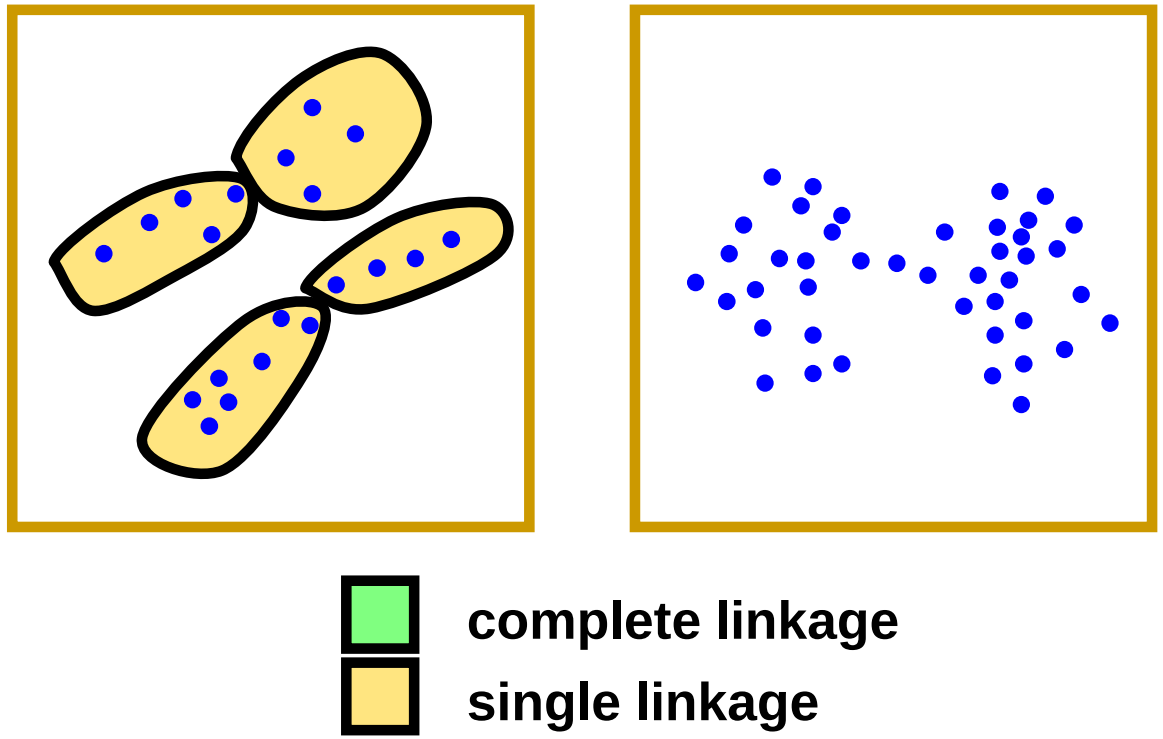
Similarity between clusters



- **Single linkage:** Closest objects
- **Complete linkage:** Furthest objects
- **Average linkage:** Average dissimilarity

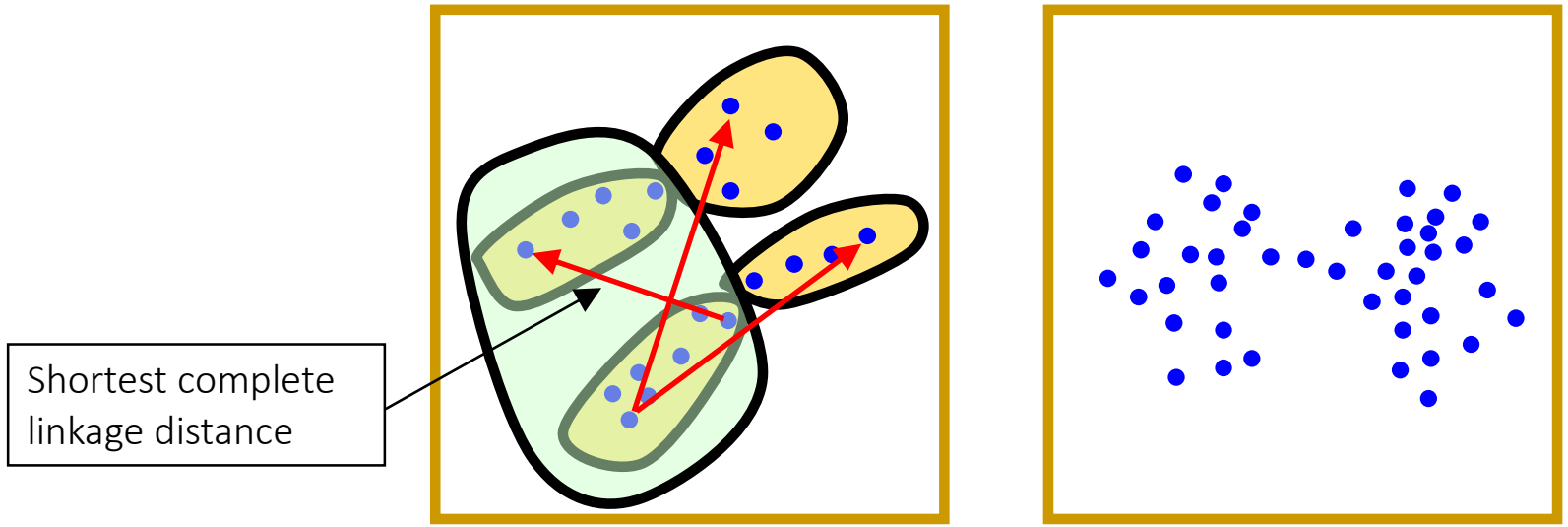
Hierarchical clustering

Similarity between clusters



Hierarchical clustering

Similarity between clusters

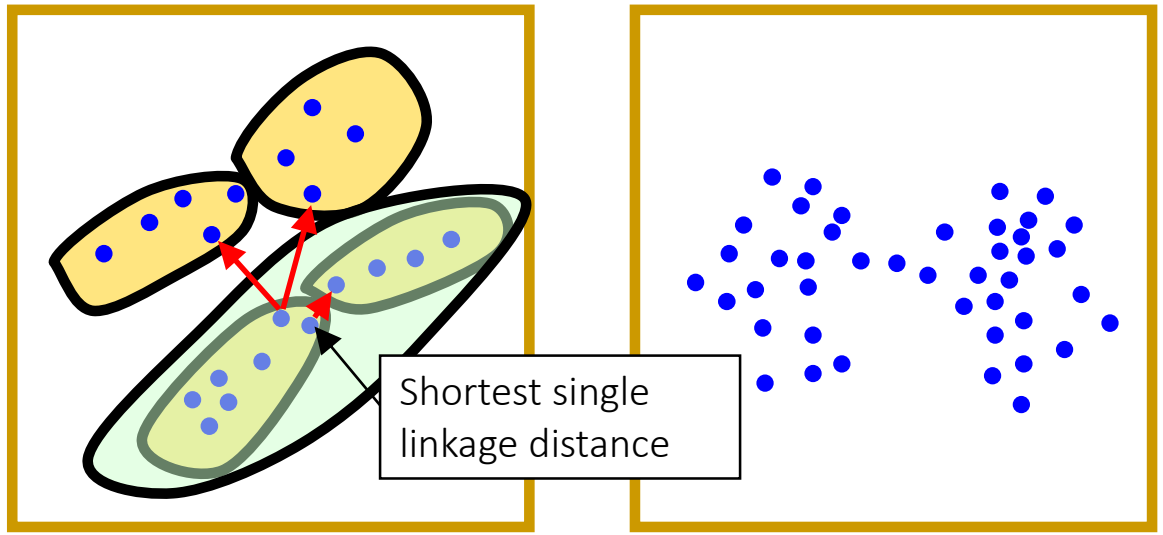


complete linkage

single linkage

Hierarchical clustering

Similarity between clusters



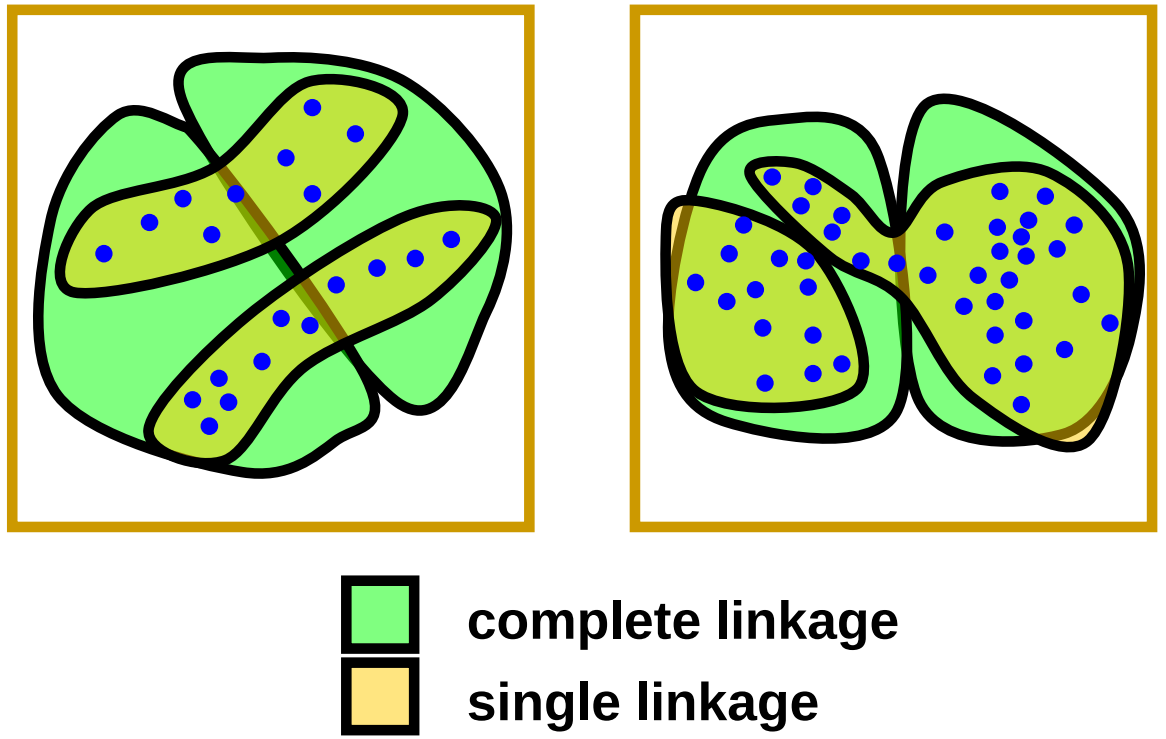
complete linkage

single linkage

Hierarchical clustering

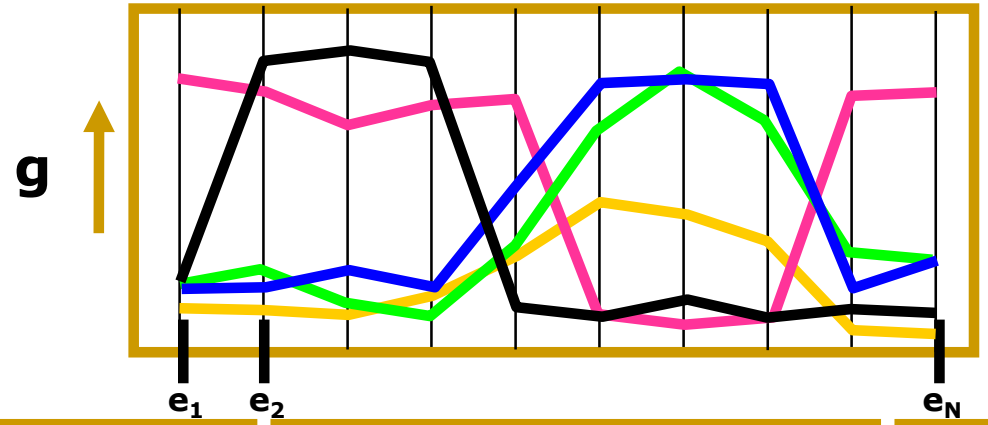
Similarity between clusters

- Single linkage -> long and “loose” clusters
- Complete linkage -> compact clusters



Hierarchical clustering

Similarity between objects



Euclidean distance

$$d(g_i, g_j) = \sqrt{(\sum((x_i - x_j)^2))}$$

d(●,●) < d(●,●)
d(●,●) << d(●,●)
d(●,●) << d(●,●)

Match exact shape

Pearson correlation

$$1 - \rho_{ij}$$

d(●,●) ≈ d(●,●)
d(●,●) << d(●,●)
d(●,●) << d(●,●)

Ignore amplitude

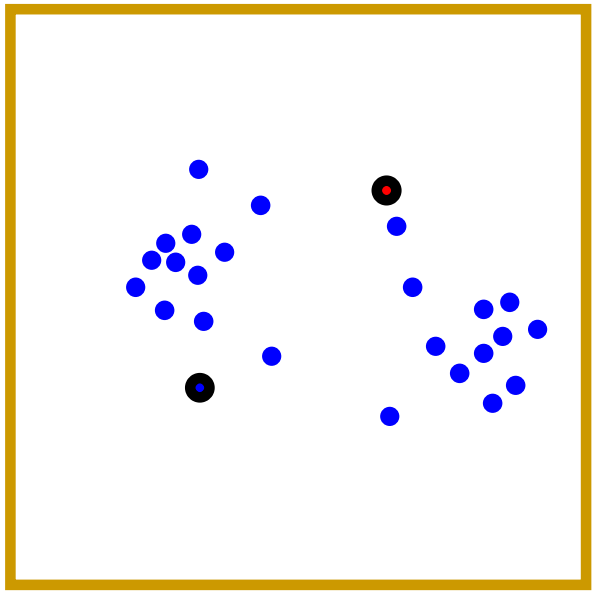
Mixed Pearson correlation

$$1 - |\rho_{ij}|$$

d(●,●) ≈ d(●,●)
d(●,●) ≈ d(●,●)
d(●,●) << d(●,●)

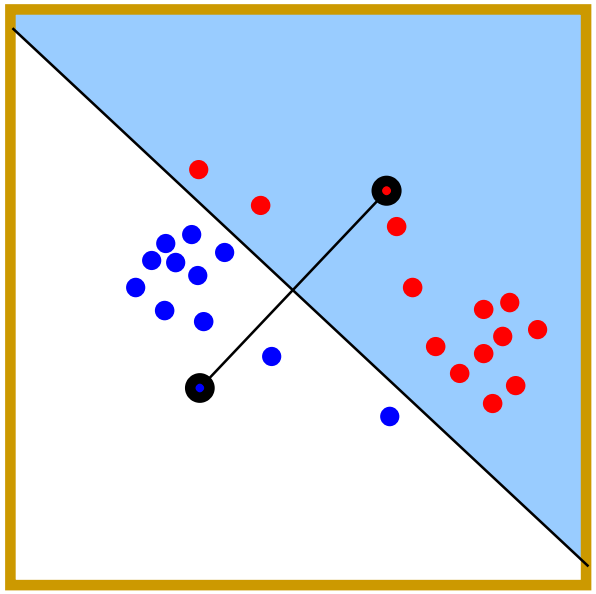
Ignore amplitude and sign

k-Means clustering



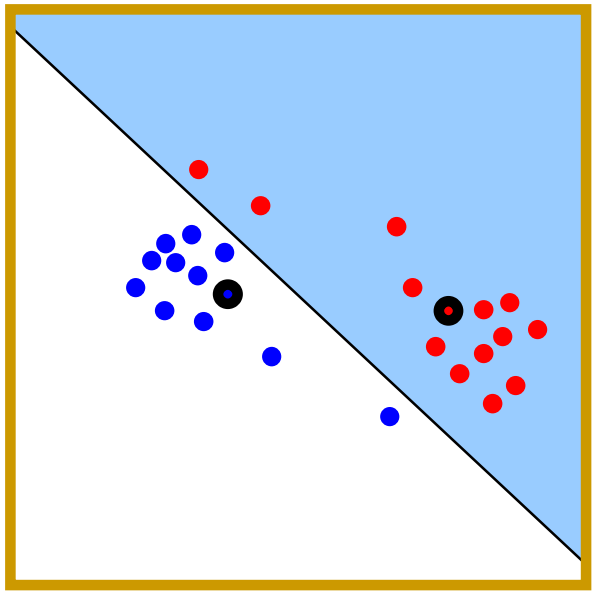
Choose randomly 2 prototypes

k-Means clustering



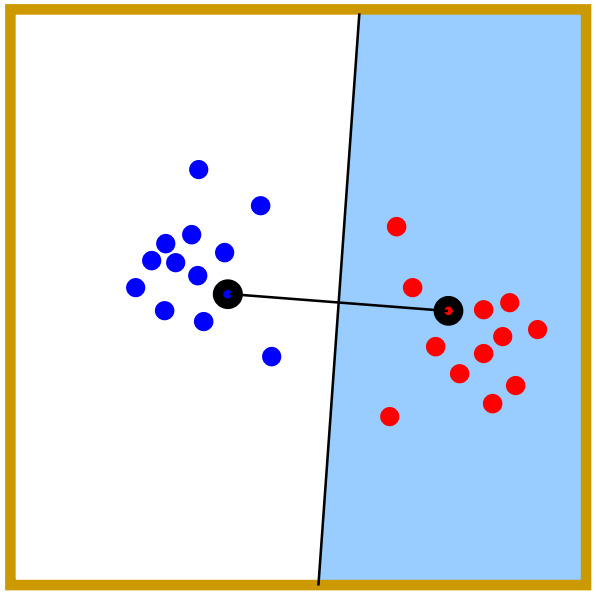
Assign objects to closest prototype
Blue area: cluster 1
White area: cluster 2

k-Means clustering



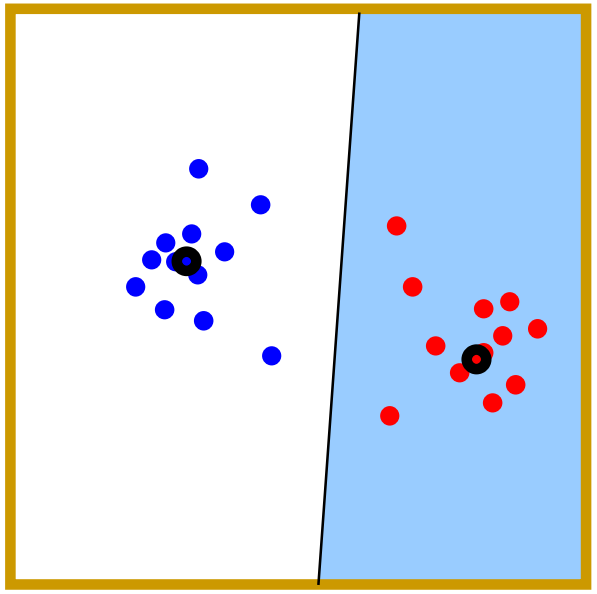
Calculate new cluster prototypes
By averaging objects

k-Means clustering



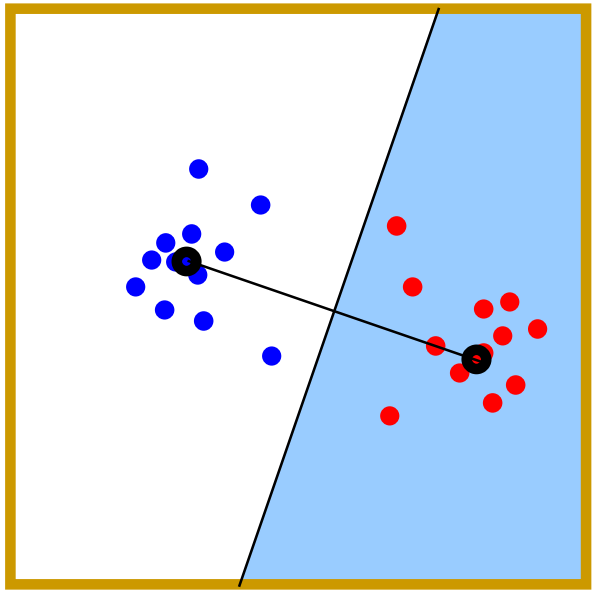
Re-assign objects to closest prototype
Blue area: cluster 1
White area: cluster 2

k-Means clustering



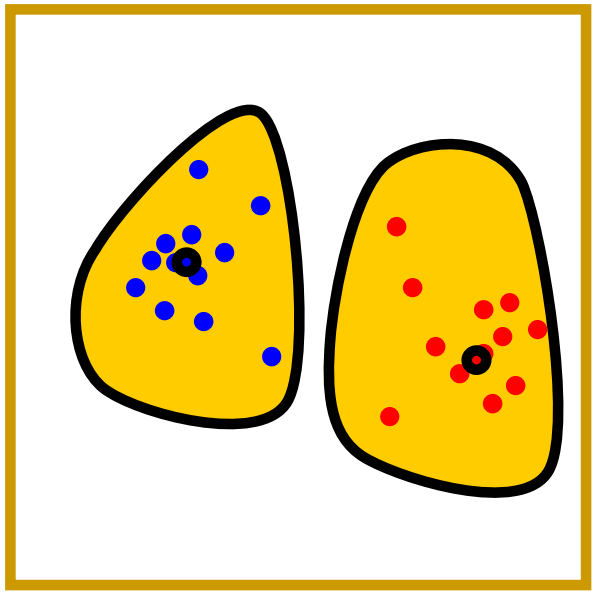
Re-calculate new cluster prototypes

k-Means clustering



Re-assign objects to closest prototype
If no objects change cluster then finished

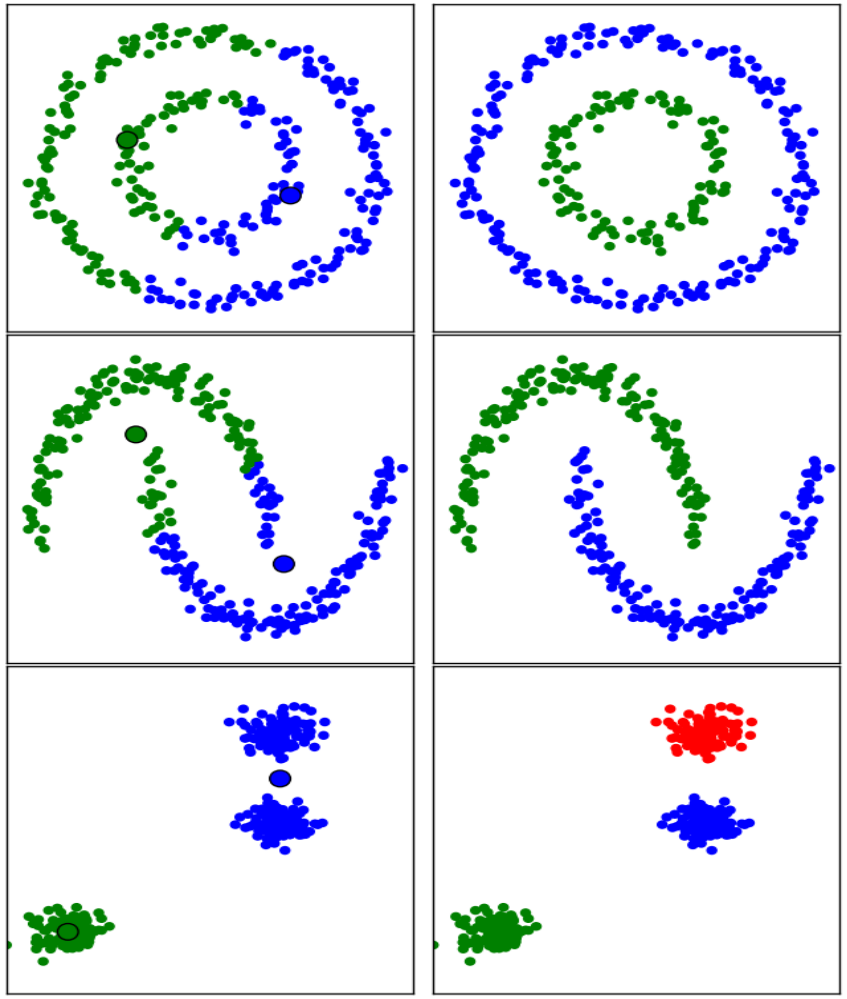
k-Means clustering



Establish clusters

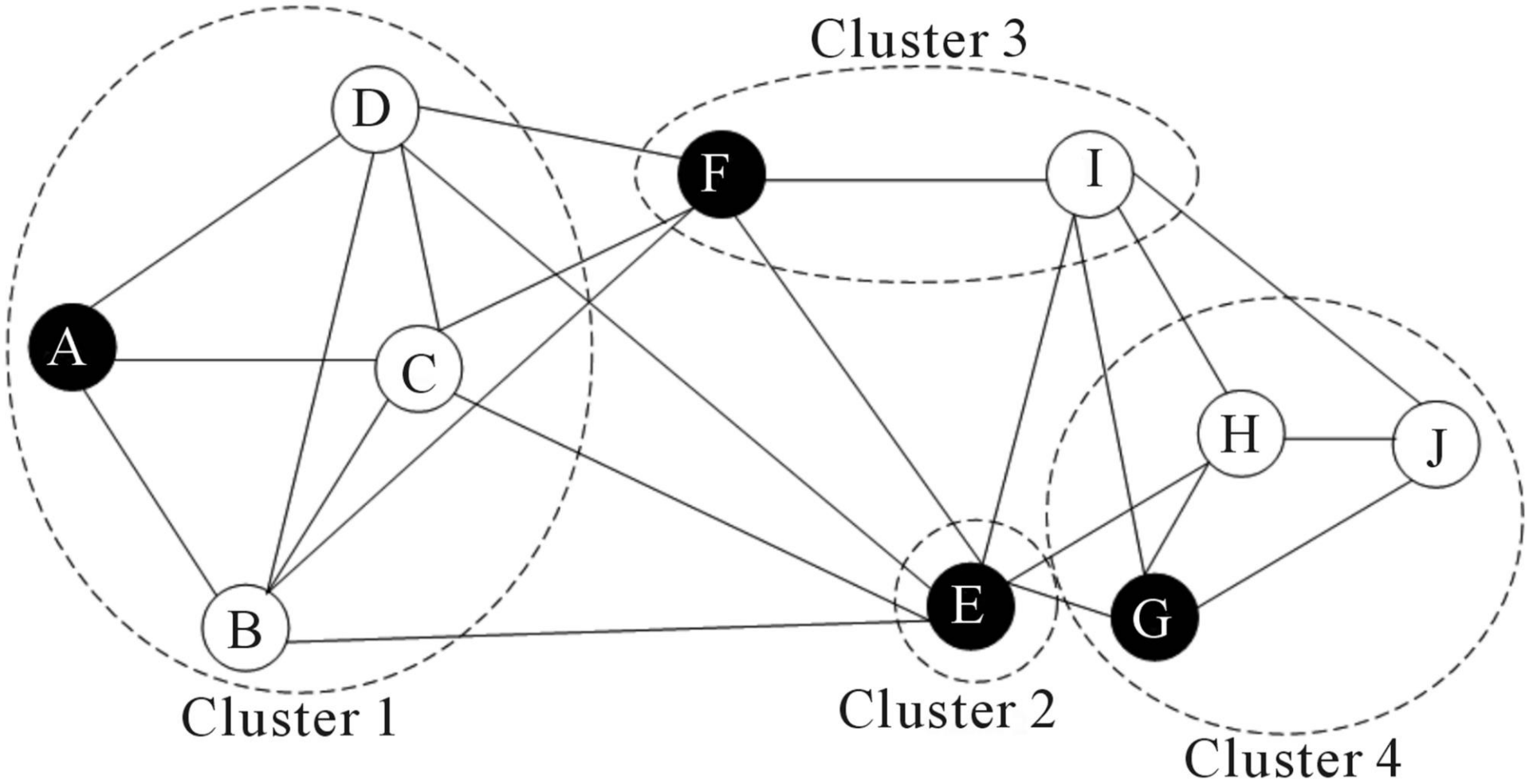
Limitations of *k*-Means

- World contains more than circles
- May take forever to converge
- Need to specify K



Graph-based clustering

Nodes -> cells
Edges -> similarity



Graph Types

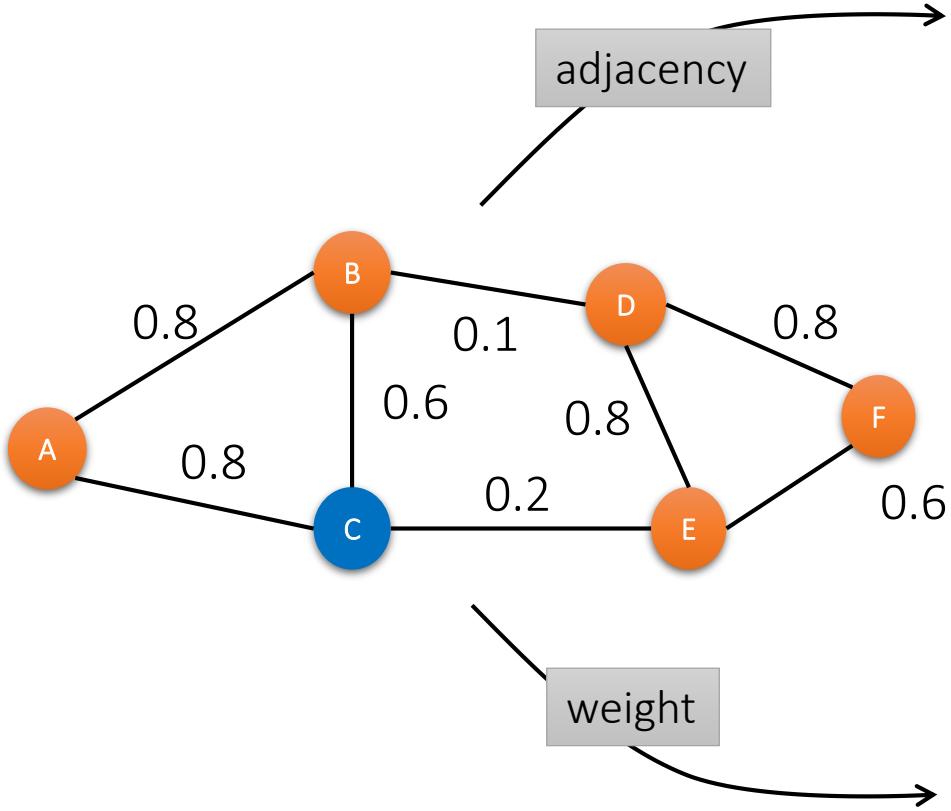
- **k-Nearest Neighbor (kNN) graph**

A graph in which two vertices *p* and *q* are connected by an edge, if the distance between *p* and *q* is among the *k*-th smallest distances from *p* to other objects from *P.*

- **Shared Nearest Neighbor (SNN) graph**

A graph in which weights define proximity, or similarity between two nodes in terms of the number of neighbors (i.e., directly connected nodes) they have in common.

Graphs, adjacency and weight matrices



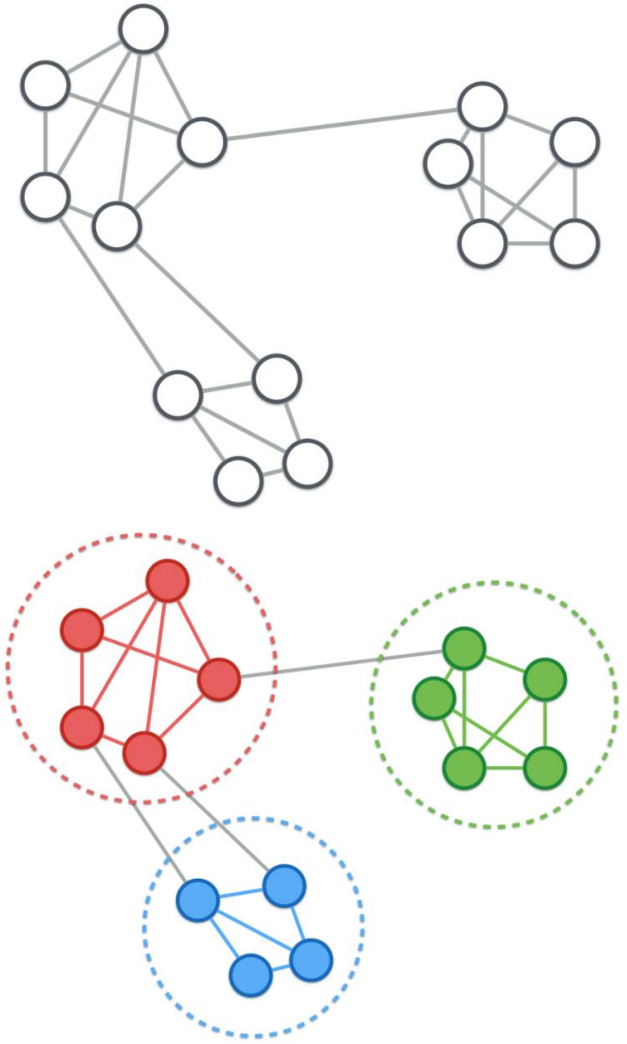
$$A = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

$$W = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 0.8 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0 & 0.6 & 0.1 & 0 & 0 \\ 0.8 & 0.6 & 0 & 0 & 0.2 & 0 \\ 0 & 0.1 & 0 & 0 & 0.8 & 0.8 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0.6 \\ 0 & 0 & 0 & 0.8 & 0.6 & 0 \end{pmatrix} \end{array}$$

Graph clustering (Community detection)

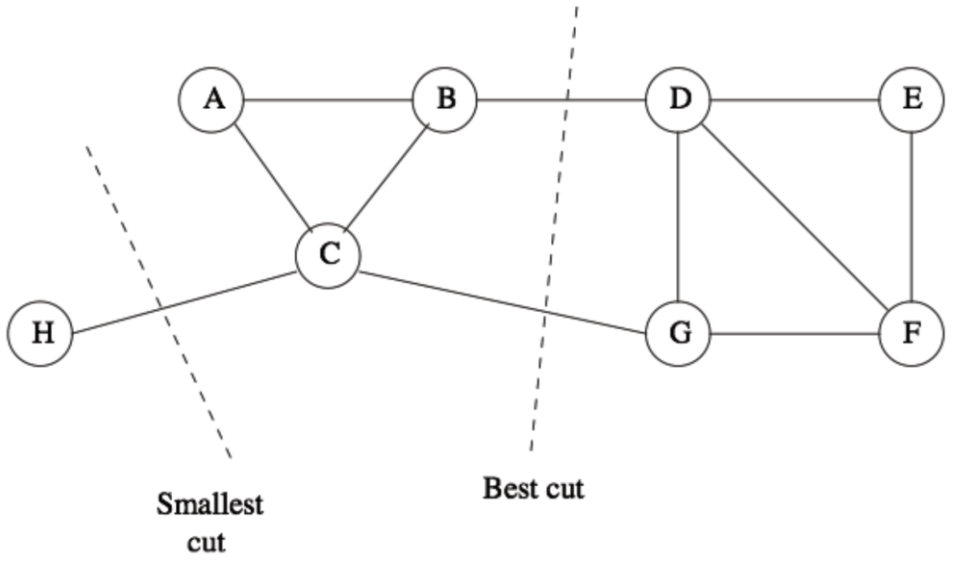
- **Communities (clusters):** groups of nodes with higher probability of being connected to each other than to members of other groups

- **Community detection:** find a group (community) of nodes with more edges inside the group than edges linking nodes of the group with the rest of the graph.



Graph cuts

- **Graph cut** partitions a graph into subgraphs
- **Cut size** is the number of cut edges
- Clustering by graph cuts: find the smallest cut that bi-partitions the graph
- The smallest cut is not always the best cut



Normalized cut

- The following way provides a good measure for the quality of a cut:
  - Denote $vol(S)$ the number of nodes in (sub)graph S
  - Denote $cut(S,T)$ the number of edges that connects nodes in S with those in T
  - The normalized cut value is:

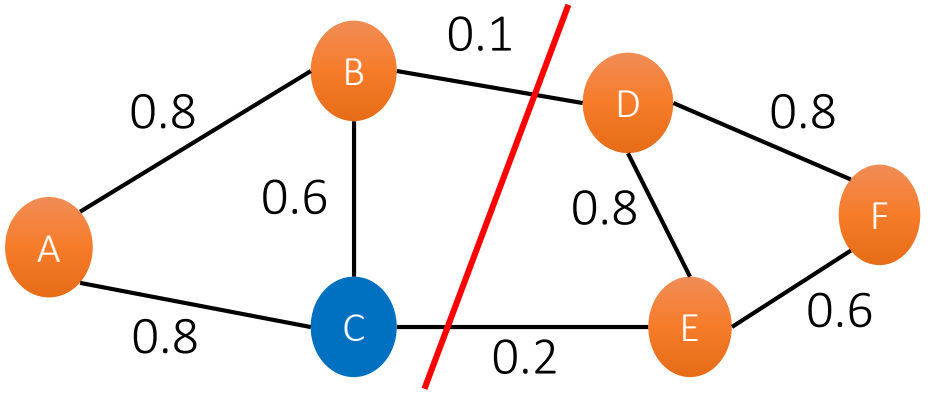
$$Ncut(S,T) = \frac{cut(S,T)}{vol(S)} + \frac{cut(S,T)}{vol(T)}$$

- The normalized cut dislikes cuts that generate very small subgraphs

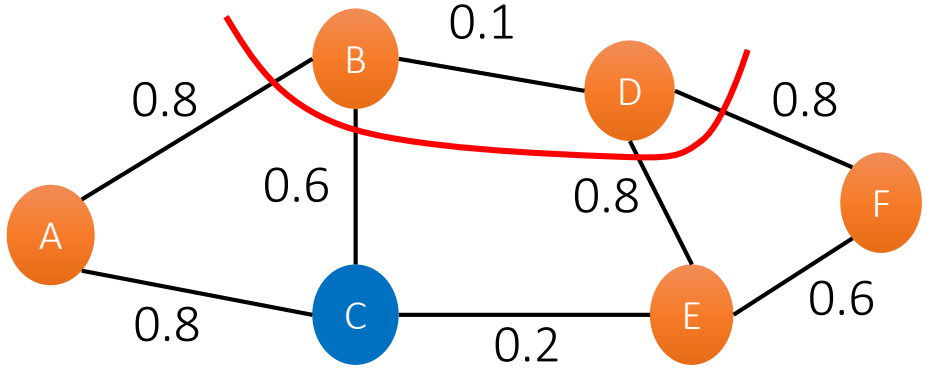
Shi & Malik (IEEE PAMI 2000)

Normalized cut (example)

- cut(S,T) = 0.1 + 0.2 = 0.3
- vol(S)= 0.3 + 0.6 + 0.8 + 0.8 = 2.5
- vol(T)= 0.3 + 0.8 + 0.8 + 0.6 = 2.5
- Ncut(S,T) = 0.3/2.5 + 0.3/2.5 = 0.24



- cut(S,T) = 0.8 + 0.6 + 0.8 + 0.8 = 3.0
- vol(S) = 3.0 + 0.1 = 3.1
- vol(T) = 3.0 + 0.8 + 0.2 + 0.6 = 4.6
- Ncut(S,T) = 3.0/3.1 + 3.0/4.6 = 1.62



Normalized cut

- Searching for the best normalized cut is NP-hard

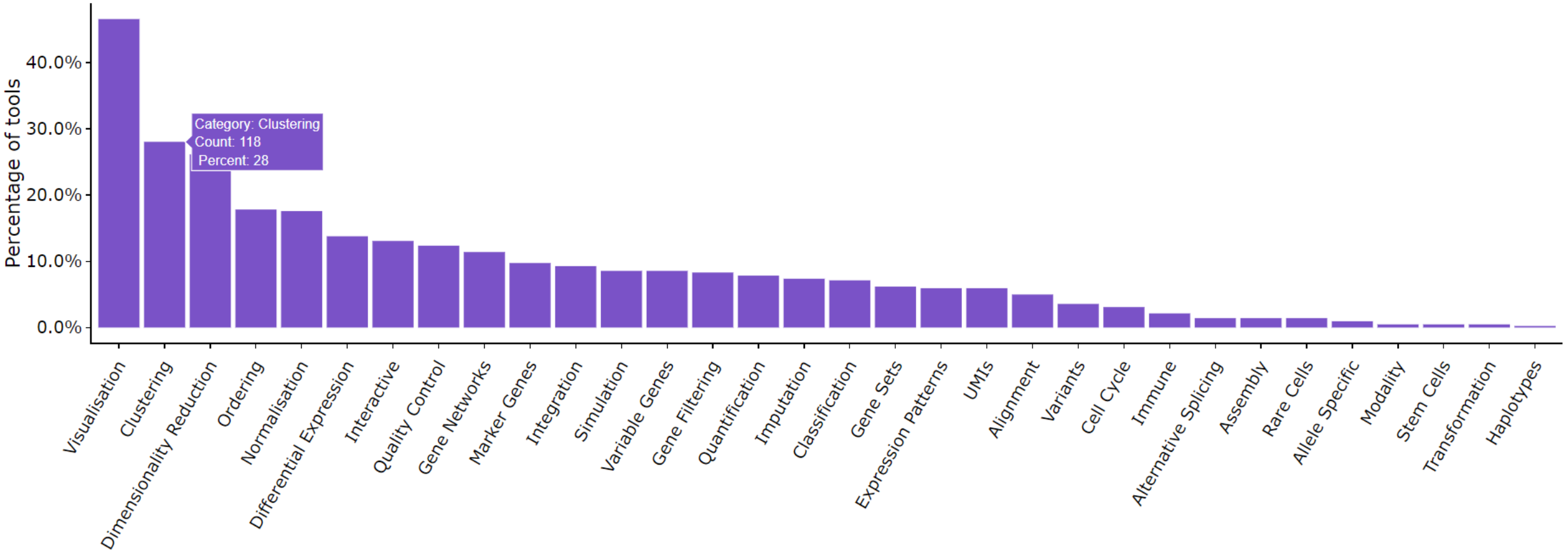
- We need a heuristic method to solve the problem:
  - Spectral clustering
  - Louvain
  - Markov clustering
  - …

scRNA-seq clustering methods

Name	Year	Method type	Strengths	Limitations
scanpy[4]	2018	PCA + graph-based	Very scalable	May not be accurate for small data sets
Seurat (latest)[3]	2016			
PhenoGraph[32]	2015			
SC3 (REF.[22])	2017	PCA + *k*-means	High accuracy through consensus, provides estimation of *k*	High complexity, not scalable
SIMLR[24]	2017	Data-driven dimensionality reduction + *k*-means	Concurrent training of the distance metric improves sensitivity in noisy data sets	Adjusting the distance metric to make cells fit the clusters may artificially inflate quality measures
CIDR[25]	2017	PCA + hierarchical	Implicitly imputes dropouts when calculating distances	
GiniClust[75]	2016	DBSCAN	Sensitive to rare cell types	Not effective for the detection of large clusters
pcaReduce[27]	2016	PCA + *k*-means + hierarchical	Provides hierarchy of solutions	Very stochastic, does not provide a stable result
Tasic et al.[28]	2016	PCA + hierarchical	Cross validation used to perform fuzzy clustering	High complexity, no software package available
TSCAN[41]	2016	PCA + Gaussian mixture model	Combines clustering and pseudotime analysis	Assumes clusters follow multivariate normal distribution
mpath[45]	2016	Hierarchical	Combines clustering and pseudotime analysis	Uses empirically defined thresholds and a priori knowledge
BackSPIN[26]	2015	Biclustering (hierarchical)	Multiple rounds of feature selection improve clustering resolution	Tends to over-partition the data
RaceID[23], RaceID2 (REF.[115]), RaceID3	2015	*k*-Means	Detects rare cell types, provides estimation of *k*	Performs poorly when there are no rare cell types
SINCERA[5]	2015	Hierarchical	Method is intuitively easy to understand	Simple hierarchical clustering is used, may not be appropriate for very noisy data
SNN-Cliq[80]	2015	Graph-based	Provides estimation of *k*	High complexity, not scalable

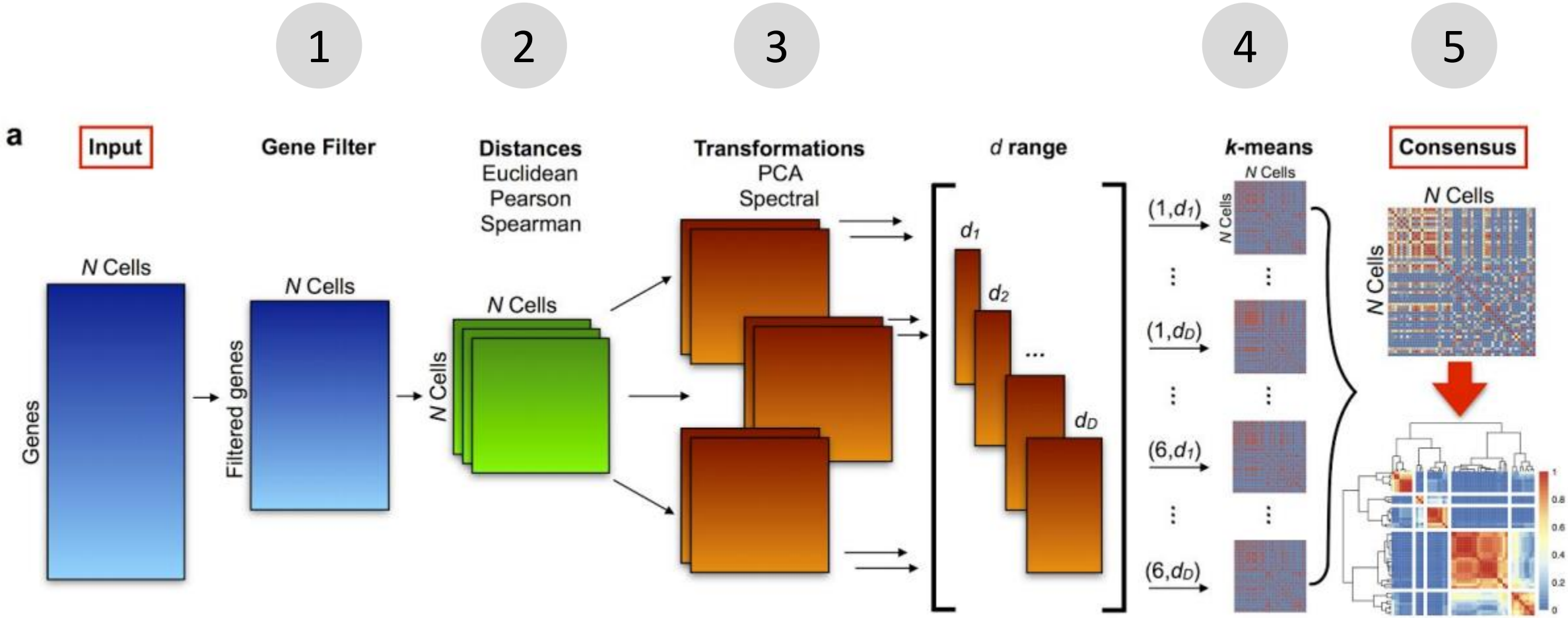
Kiselev et al. (https://doi.org/10.1038/s41576-018-0088-9)

scRNA-seq clustering methods



https://www.scrna-tools.org/

Single Cell Consensus Clustering – SC3



Kiselev et al. (https://doi.org/10.1038/nmeth.4236)

Single Cell Consensus Clustering – SC3

1) Gene filtering – rare and ubiquitous genes

2) Distance matrices (DM) – Euklidean, Spearman, Pearson

3) Transformation of DM with PCA or Laplacian

4) K-means clustering with first *d* eigenvectors

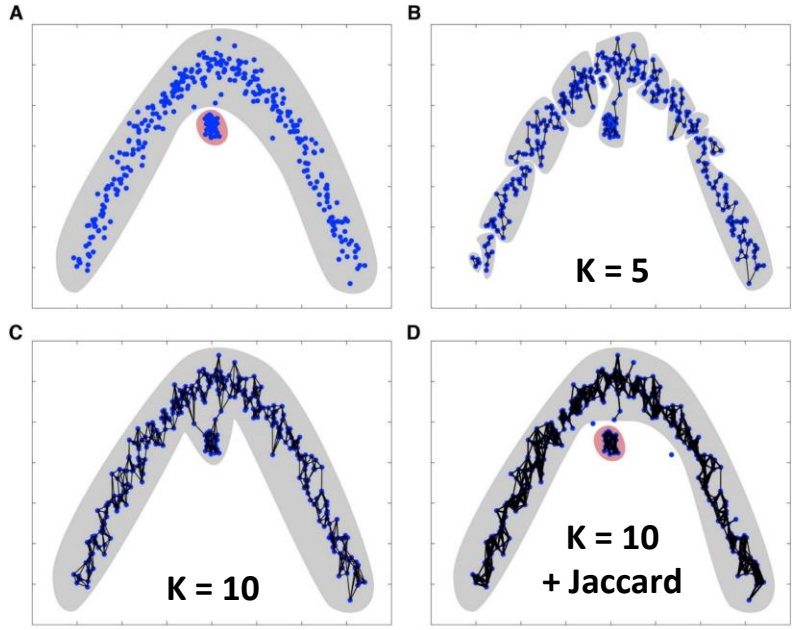
5) Consensus clustering – distance 1/0 for cells in same/different clusters -> hierarchical clustering on average distances.

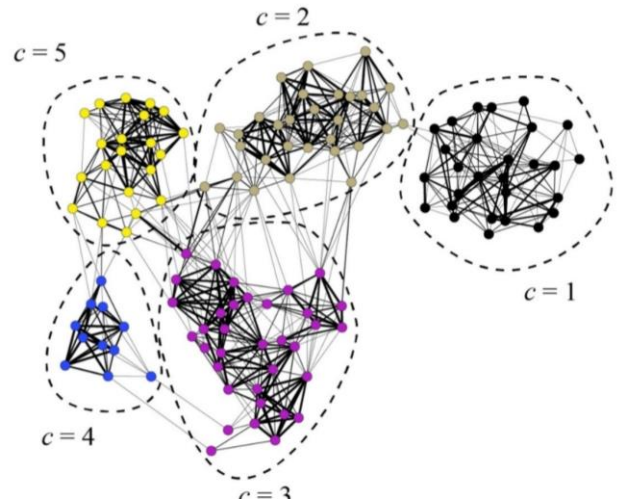
Seurat

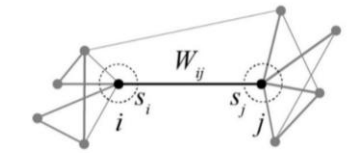
1) Construct KNN (k-nearest neighbor) graph based on the euclidean distance in PCA space.

2) Refine the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard distance).

3) Cluster cells by optimizing for modularity (Louvain algorithm)







$$Q = \frac{1}{2m}\sum_{i,j}\left[W_{ij} - \frac{s_i s_j}{2m}\right]\delta(c_i, c_j)$$

Xu and Su (https://doi.org/10.1093/bioinformatics/btv088)
Levine et al. (https://doi.org/10.1016/j.cell.2015.05.047)

Comparing different clusterings

- Adjusted Rand Index (ARI)

Given a set S of n elements, and two groupings or partitions (e.g. clusterings) of these elements $X = \{X_1, X_2, \dots, X_r\}$ and $Y = \{Y_1, Y_2, \dots, Y_r\}$

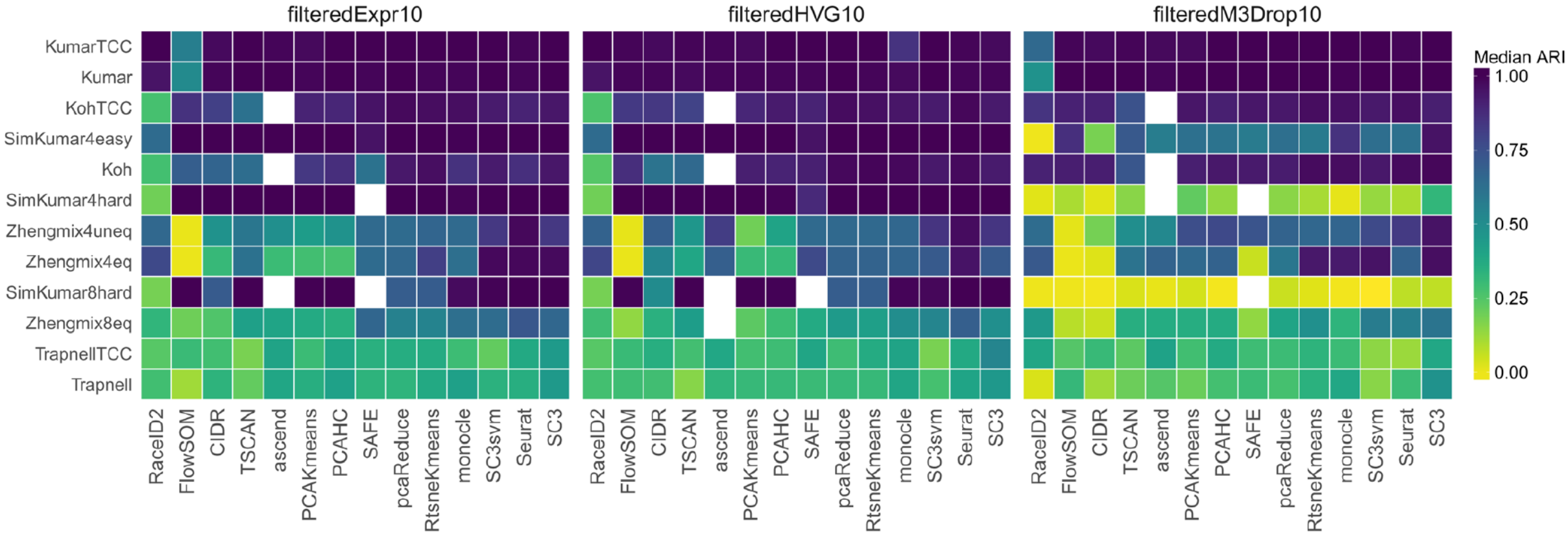
Confusion matrix/contingency table

$X \backslash Y$	Y_1	Y_2	$\dots$	Y_s	Sums
X_1	n_{11}	n_{12}	$\dots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\dots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	a_r
Sums	b_1	b_2	$\dots$	b_s	

$$n_{ij} = |X_i \cap Y_j|$$

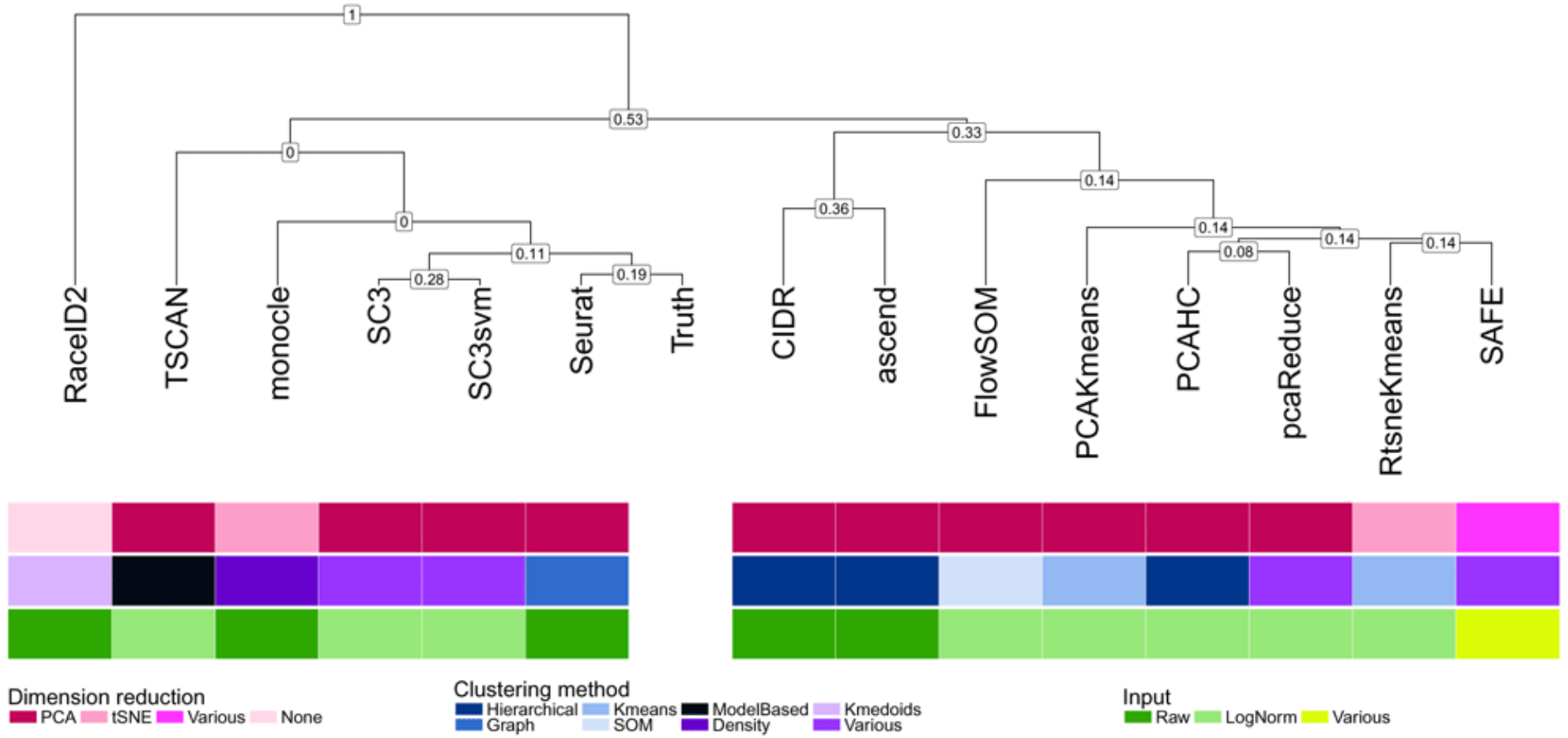
$$\overbrace{ARI}^{\text{Adjusted Index}} = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}]}_{\text{Max Index}} - \underbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}_{\text{Expected Index}}}$$

Benchmarking scRNA-seq clustering methods



Duò et al. (https://doi.org/10.12688/f1000research.15666.2)

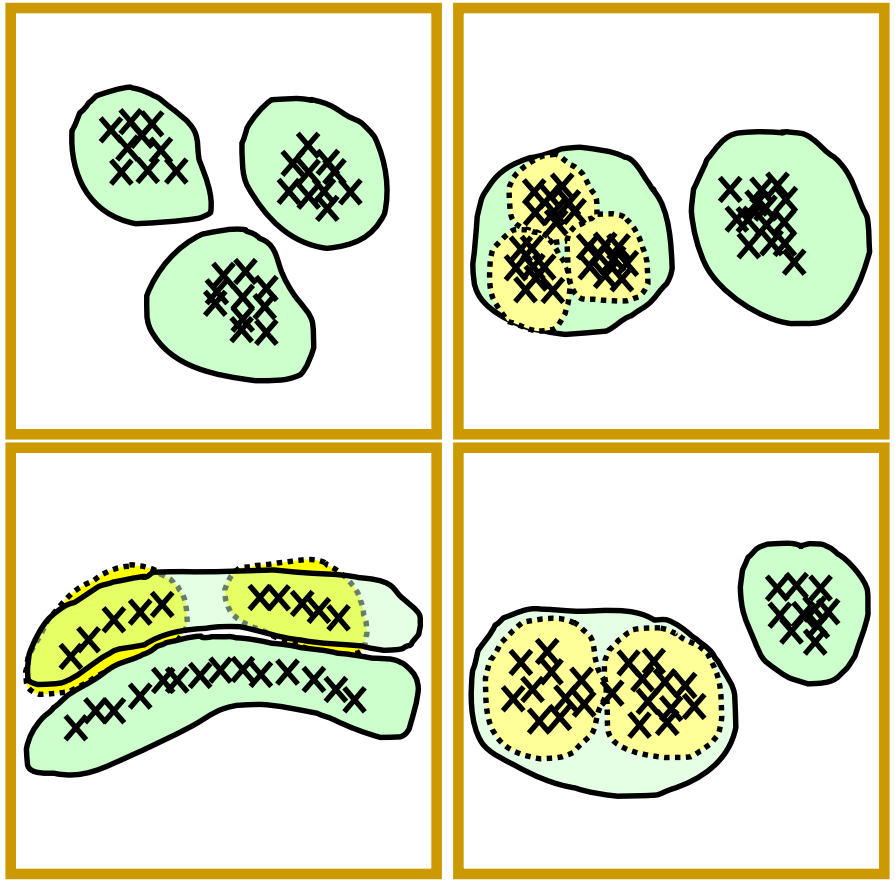
Benchmarking scRNA-seq clustering methods



Duò et al. (https://doi.org/10.12688/f1000research.15666.2)

Clustering is subjective!

- Principle choices
  - Similarity measure
  - Algorithm
- Different choice leads to different results
  - Subjectivity becomes reality
- Cluster process
  - Validate, interpret (generate hypothesis), repeat steps



How many clusters do you really have?

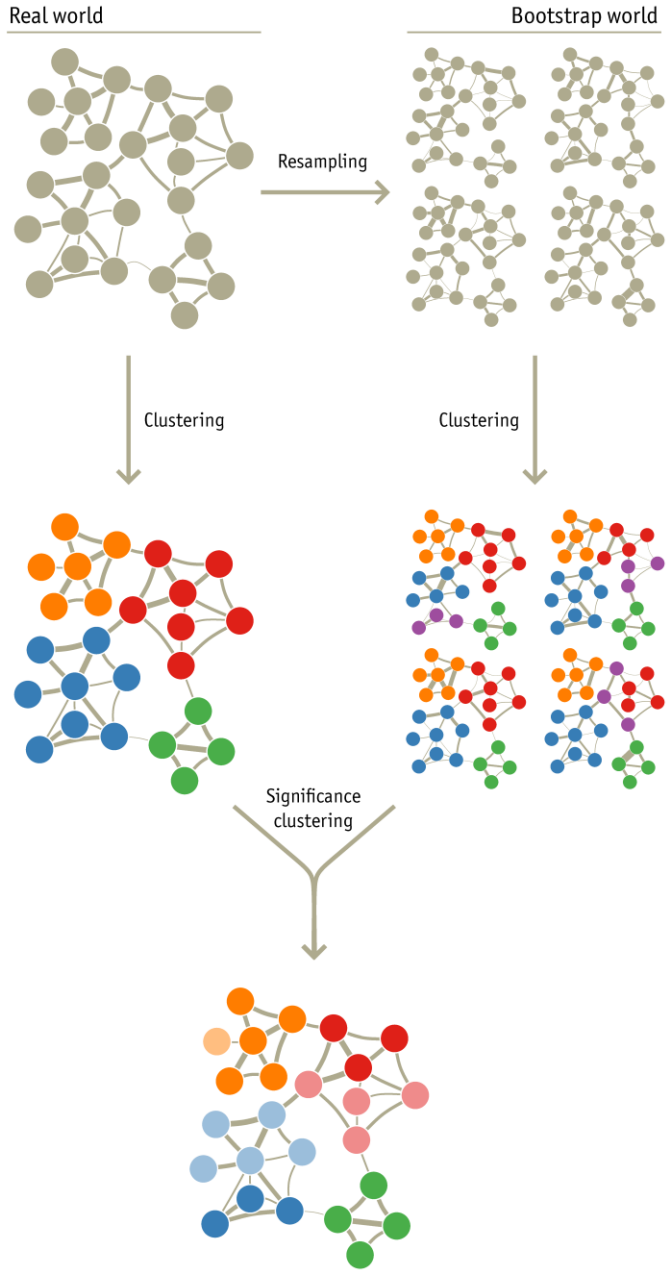
- It is hard to know when to stop clustering – you can always split the cells more times.

- Can use:
  - Do you get any/many significant DE genes from the next split?
  - Some tools have automated predictions for number of clusters – may not always be biologically relevant

Bootstrapping

- How confident can you be that the clusters you see are real?

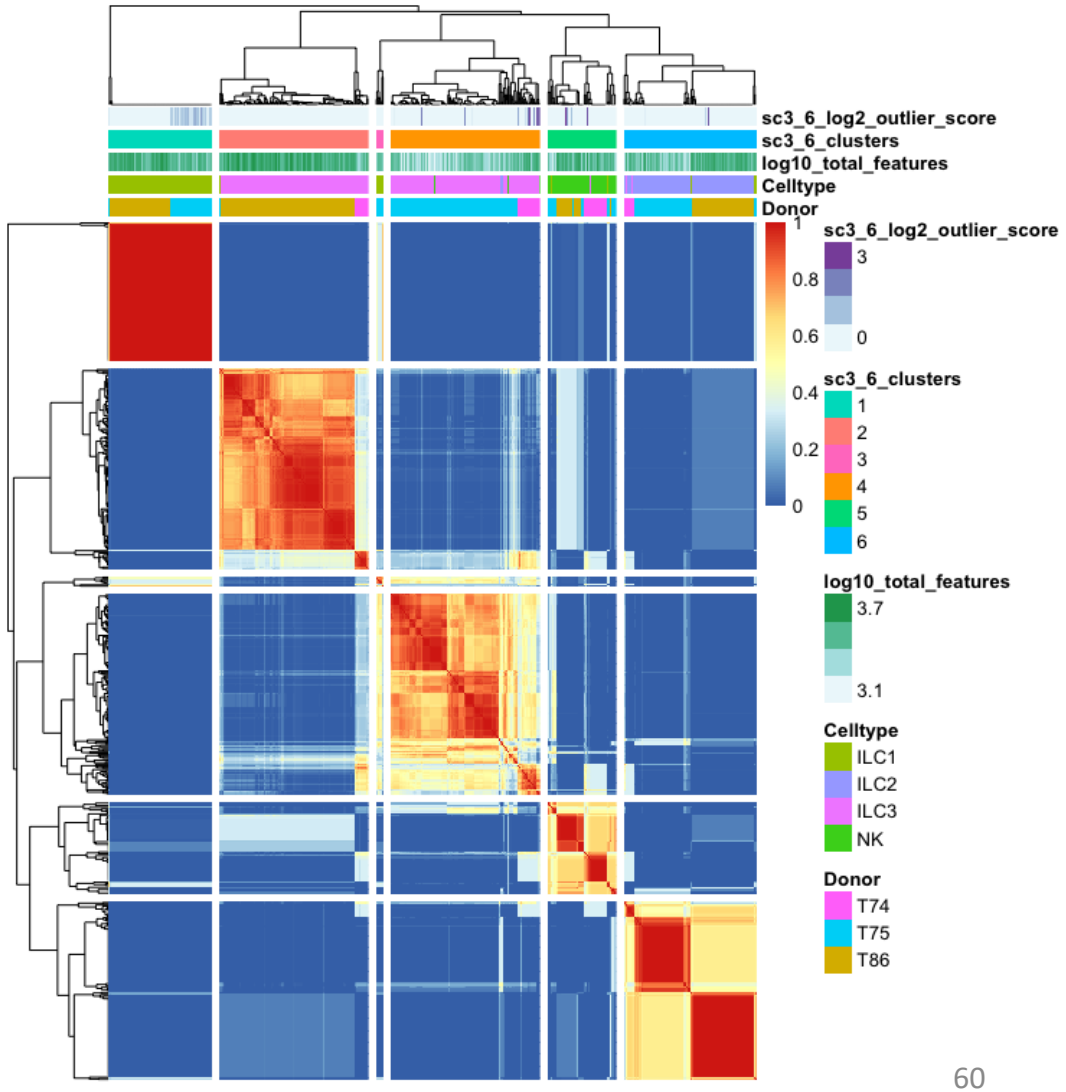
- You can always take a random set of cells from the same cell type and manage to split them into clusters.



Rosvall and Bergstrom (https://doi.org/10.1371/journal.pone.0008694)

Always check QC data

- Is what your splitting mainly related to batches, qc-measures (especially detected genes)?



From clusters to cell identities

- Using lists of DE genes and prior knowledge of the biology

- Using lists of DE genes and comparing to other scRNAseq data or sorted cell populations

Databases with celltype gene signatures

- PanglaoDB (https://panglaodb.se/)
  - Human: 295 samples, 72 tissues, 1.1 M cells
  - Mouse: 976 samples, 173 tissues, 4 M cells
  - Franzén et al (https://doi.org/10.1093/database/baz046)

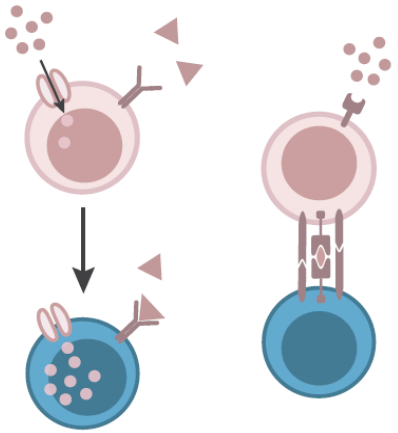
- CellMarker (http://biocc.hrbmu.edu.cn/CellMarker/)
  - Human: 13,605 cell markers of 467cell types in 158 tissues
  - Mouse: 9,148 cell makers of 389 cell types in 81 tissues
  - Zhang et al. (https://doi.org/10.1093/nar/gky900)

Challenges in clustering

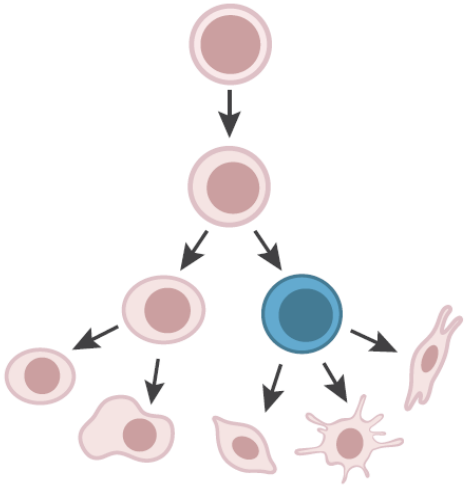
- What is a cell type?
- What is the number of clusters *k*?
- **Scalability**: in the last few years the number of cells in scRNA-seq experiments has grown by several orders of magnitude from ~10^2 to ~10^6

Cell identity

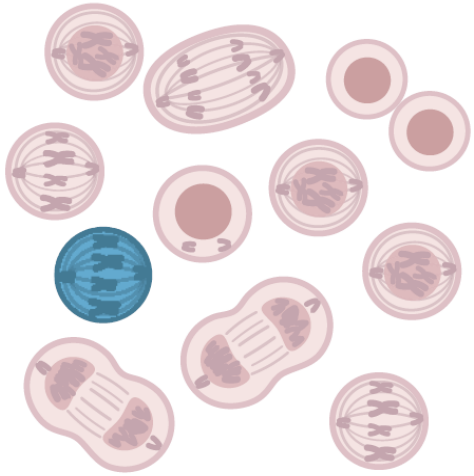
Environmental stimuli



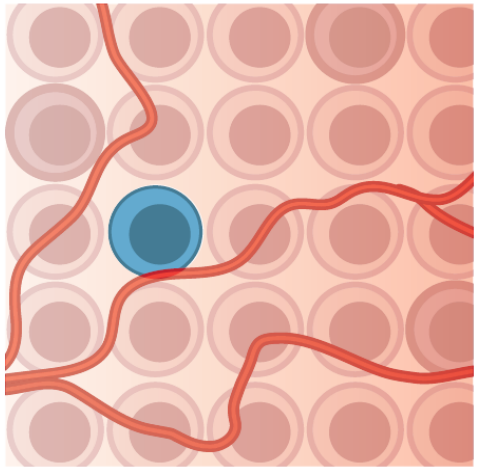
Cell development



Cell cycle

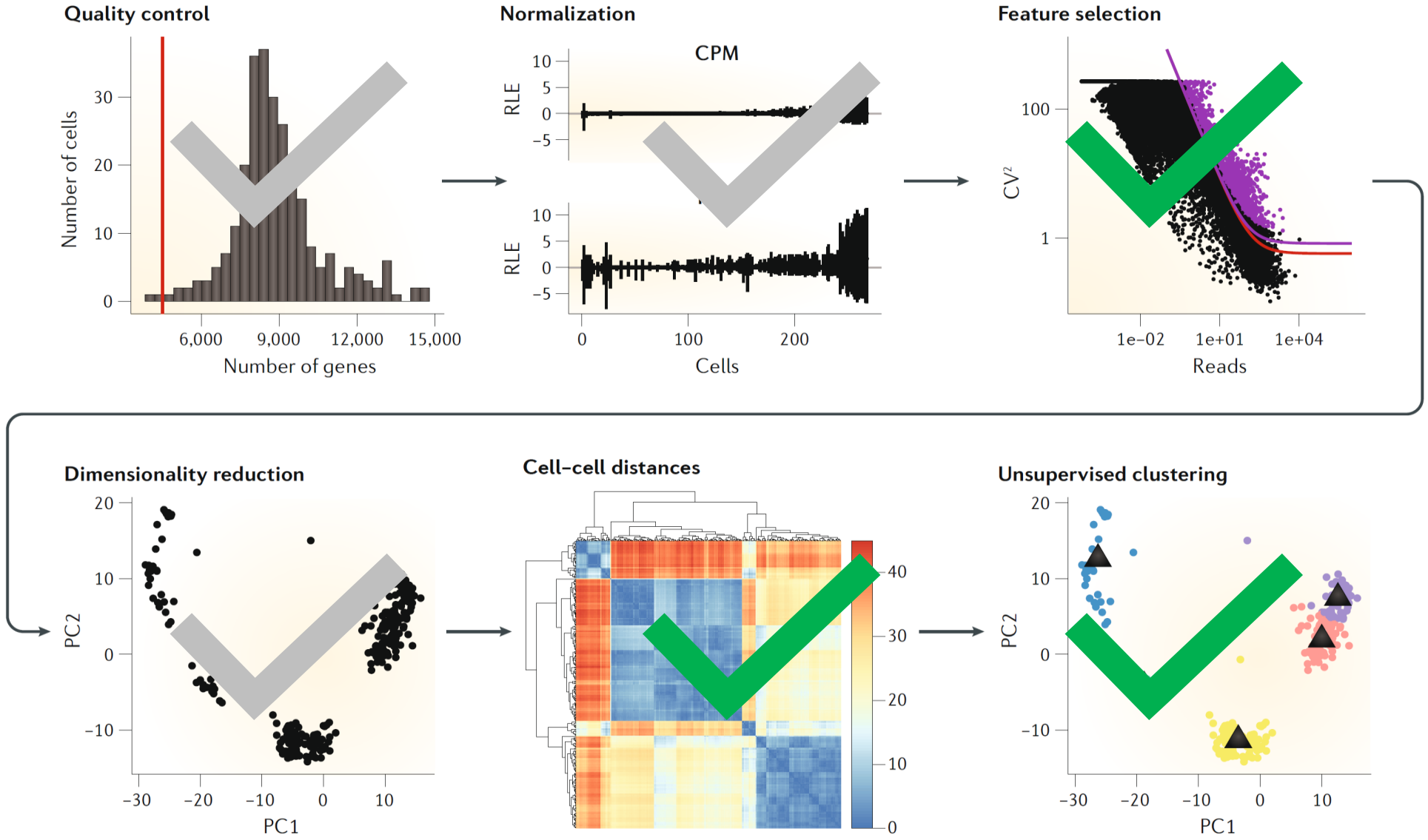


Spatial context



Wagner et al. (https://doi.org/10.1038/nbt.3711)

Summary


Quality control
Number of cells
30
20
10
0
6,000
9,000
12,000
15,000
Number of genes
Normalization
CPM
RLE
10
5
0
−5
0
100
200
Cells
Feature selection
CV²
100
1
1e−02
1e+01
1e+04
Reads
Dimensionality reduction
PC2
20
10
0
−10
−30
−20
−10
0
10
PC1
Cell–cell distances
40
30
20
10
0
Unsupervised clustering
PC2
20
10
0
−10
−30
−20
−10
0
10
PC1

Clustering practical

- Feature selection (HVG)
- Dimensionality reduction: select principal components
- Hierarchical clustering: distances and linkage methods
- tSNE + *k*-Means
- Graph-based clustering

Resources

- Kiselev et al. ”Challenges in unsupervised clustering of single- cell RNA- seq data”

https://doi.org/10.1038/s41576-018-0088-9

- Duò et al. ” A systematic performance evaluation of clustering methods for single-cell RNA-seq data”

https://doi.org/10.12688/f1000research.15666.2

- Orchestrating Single-Cell Analysis with Bioconductor

https://osca.bioconductor.org/

- Hemberg single cell course: Analysis of single cell RNA-seq data

https://scrnaseq-course.cog.sanger.ac.uk/website/index.html

- Slides Åsa Björklund (NBIS, SciLifeLab)

https://github.com/NBISweden/workshop-scRNAseq/tree/master/slides2019