Single Cell RNA-seq Clustering

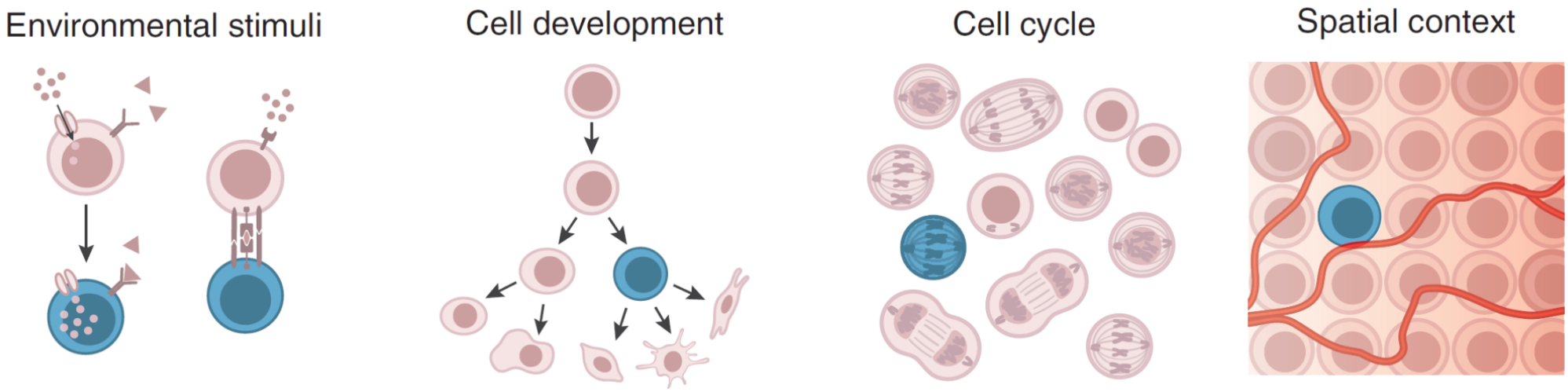
Marcel Reinders &
Lieke Michielsen & Ahmed Mahfouz
Delft Bioinformatics Lab, TU Delft
Leiden Computational Biology Center, LUMC

Leiden
Computational Biology Center

LUMC

TU Delft

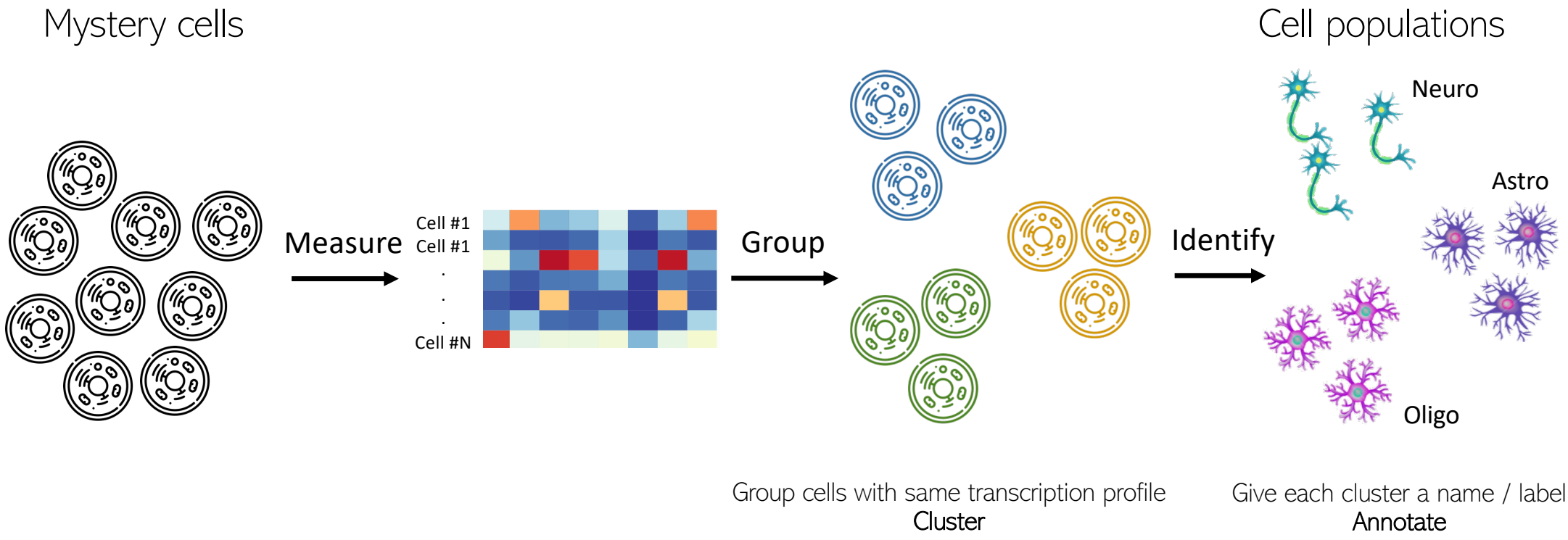
Cell Identity determined by diverse factors

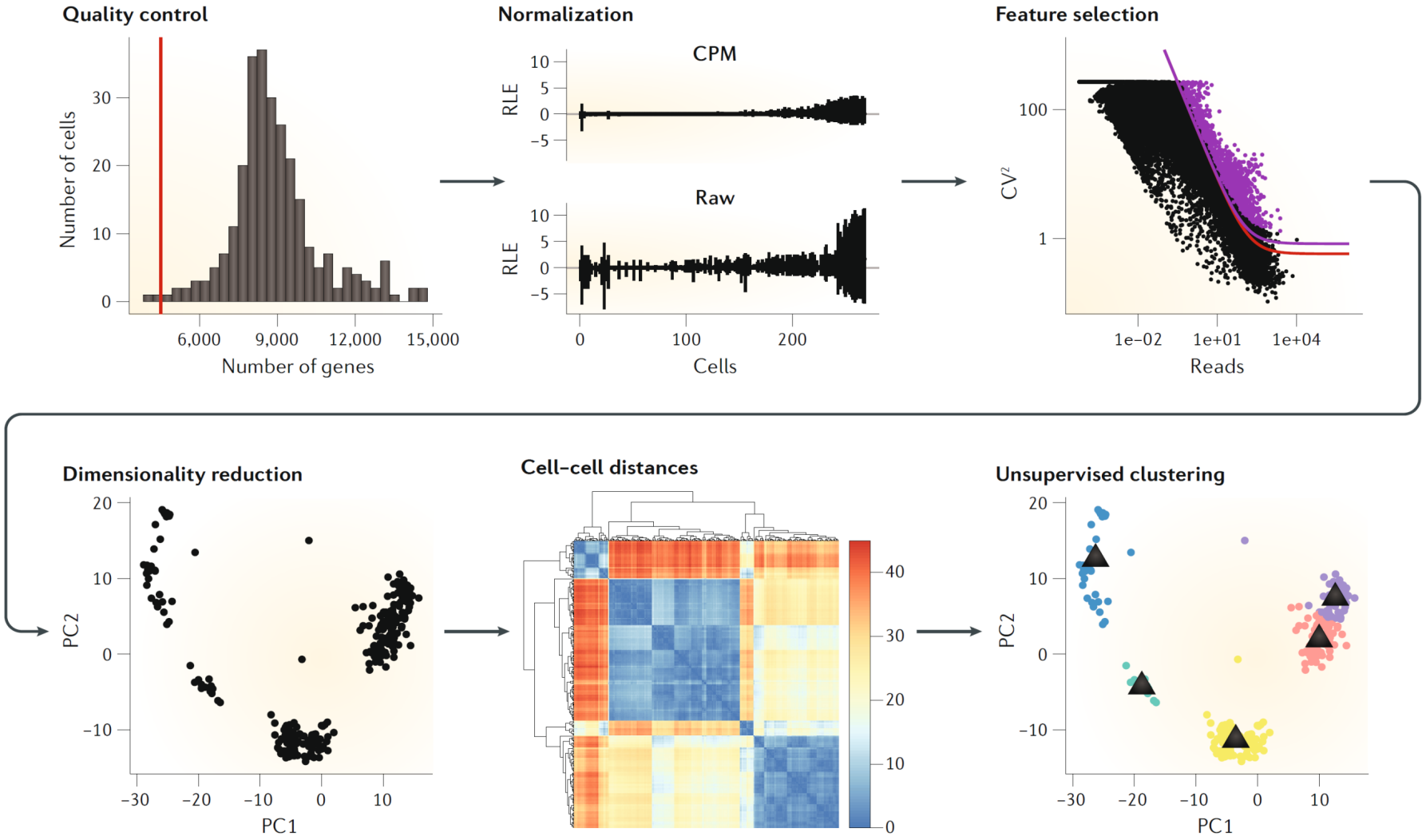


- **Cell type:** Permanent state (e.g., a hepatocyte typically cannot turn into a neuron).
- **Cell state:** Transiently state (e.g., during differentiation, stimulation, spatial patterning)

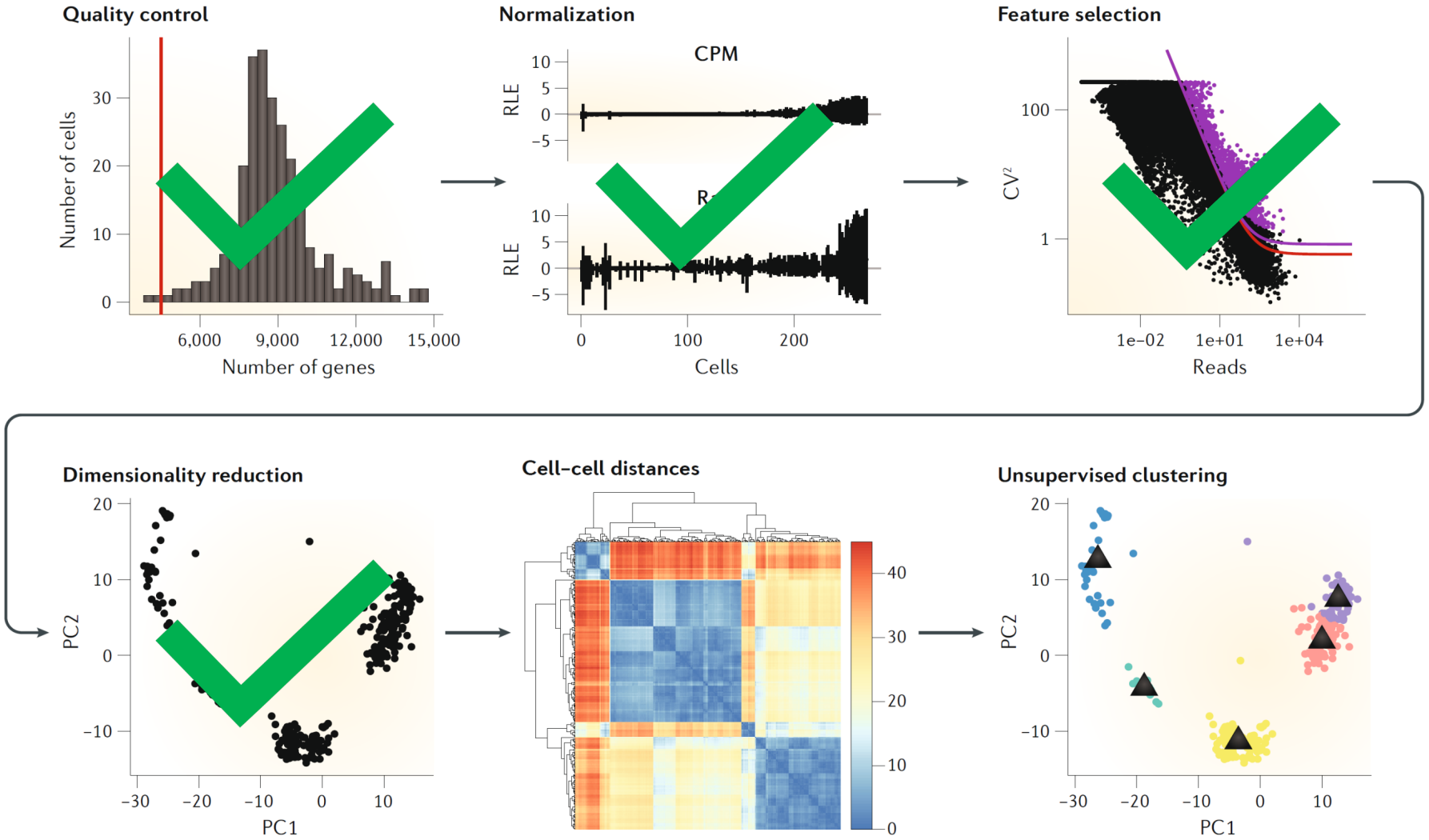
Wagner et al. (https://doi.org/10.1038/nbt.3711)

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)

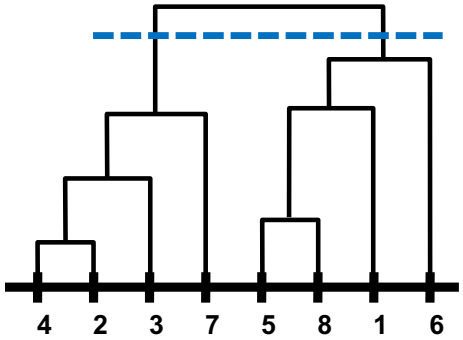
Outline

- Introduction to clustering
  - Hierarchical clustering
  - *k*-Means clustering
  - Graph-based clustering
- Cluster validation
- scRNA-seq clustering
- Annotating clusters

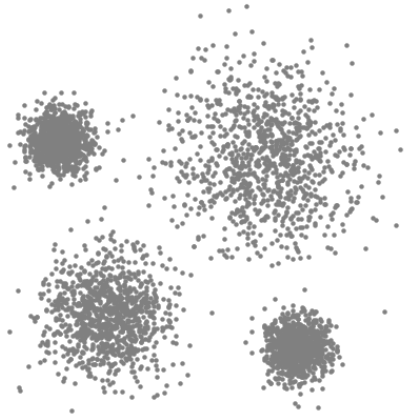
Outline

- Introduction to clustering
  - Hierarchical clustering
  - *k*-Means clustering
  - Graph-based clustering
- Cluster validation
- scRNA-seq clustering
- Annotating clusters

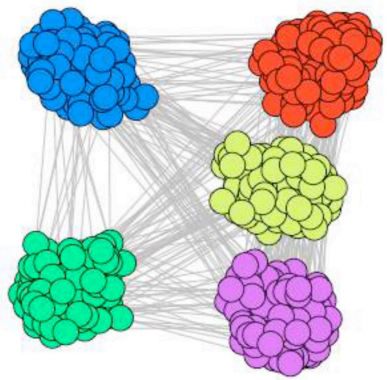
Many clustering approaches



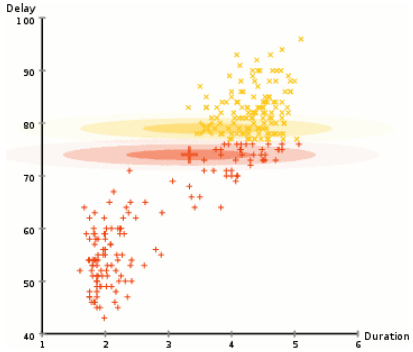
Hierarchical clustering



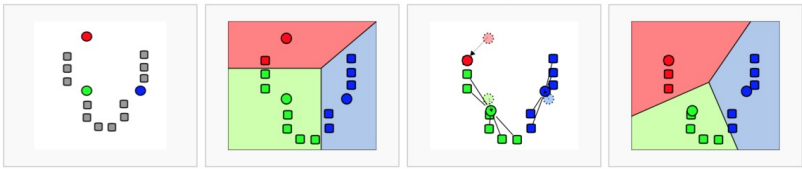
Mean shift clustering



Graph-based clustering



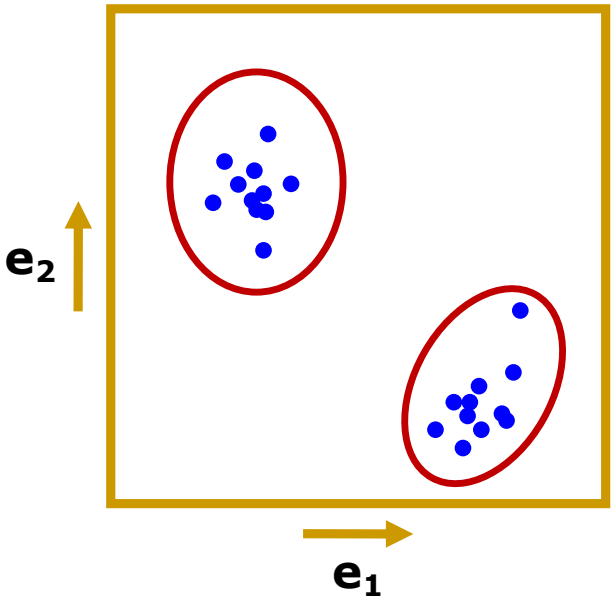
Gaussian mixture modeling

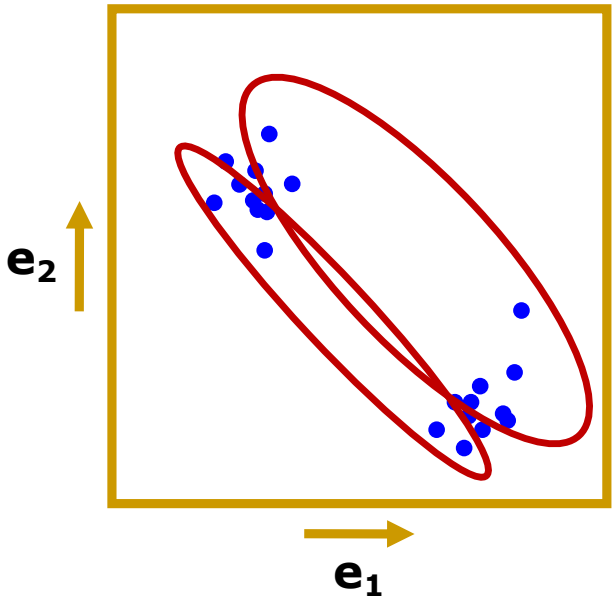


k-means clustering

Clustering

What defines a good clustering





Clustering

What defines a good clustering

Structure when:

1. Samples within cluster resemble each other

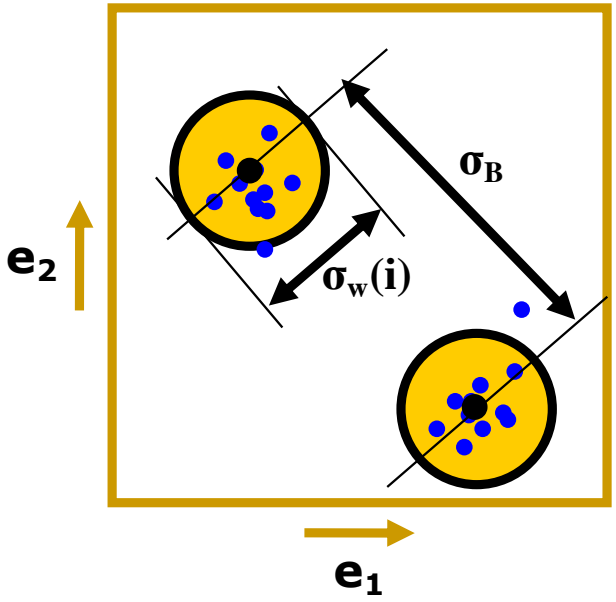
   (small pairwise distance between cells within cluster: small within variance, $\sigma_W(i)$, for each cluster i)

2. Clusters deviate from each other

   (large pairwise distance between cells of different clusters: large between variance, σ_B)

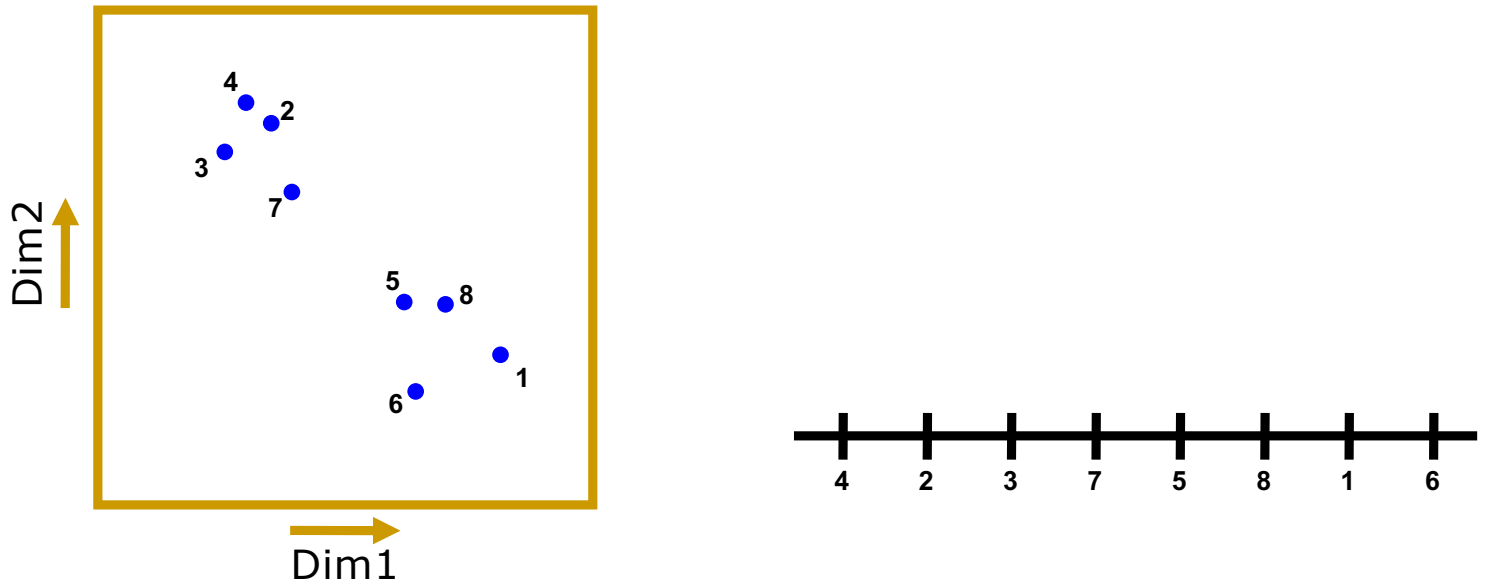
Group sa

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



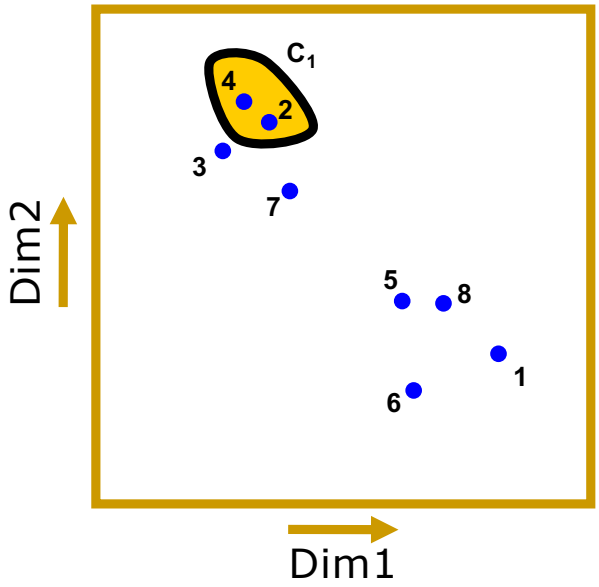
Hierarchical clustering

Hierarchical clustering

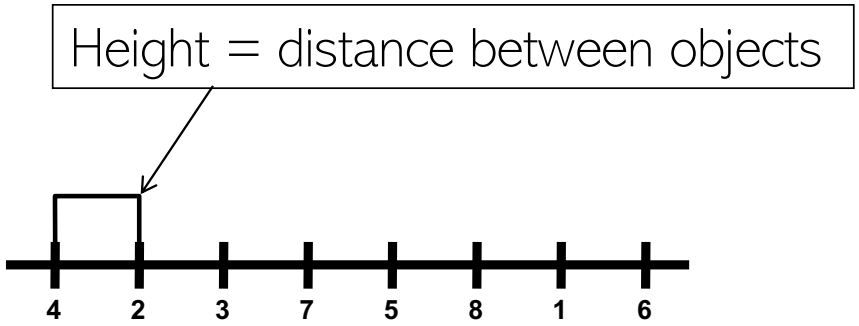


Find most similar objects (cells) and group them

Hierarchical clustering

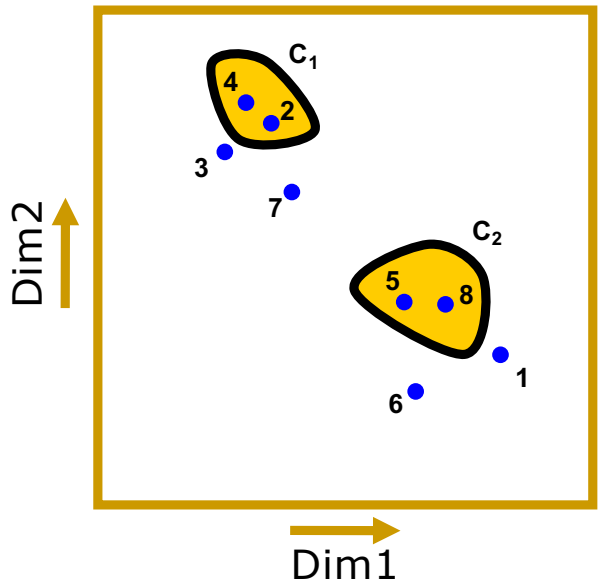


Dendrogram

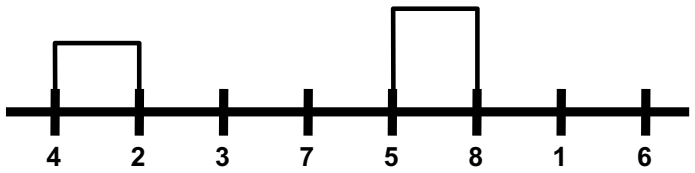


These are: objects 4 and 2
Again, find most similar objects (cells or clusters) and group them

Hierarchical clustering

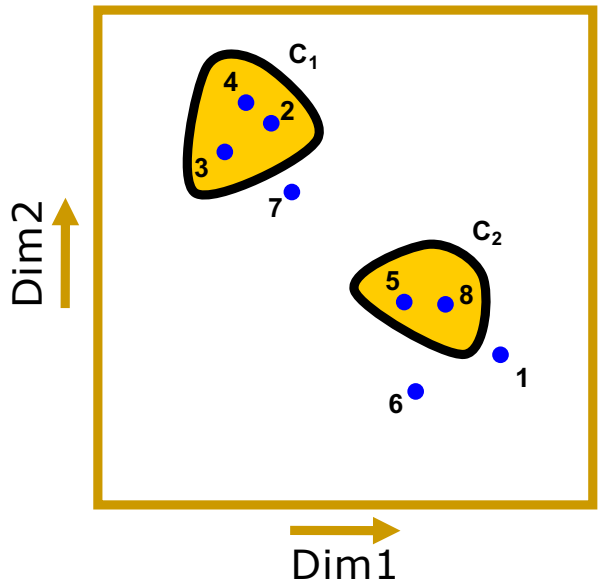


Dendrogram

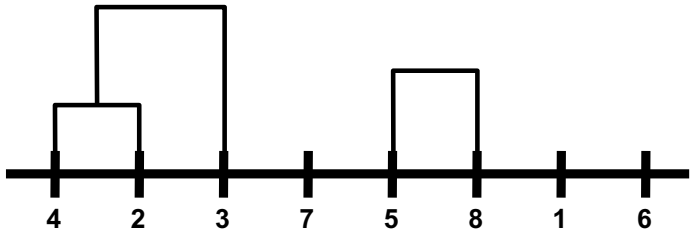


These are: objects 5 and 8
Repeat finding most similar objects (cells or clusters) and group them

Hierarchical clustering

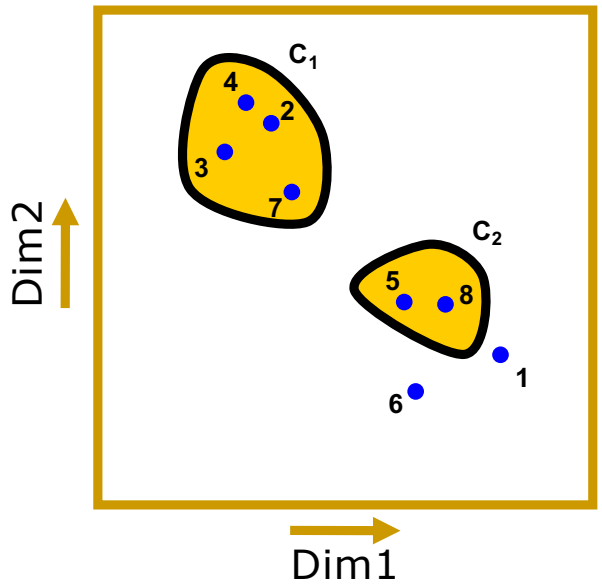


Dendrogram

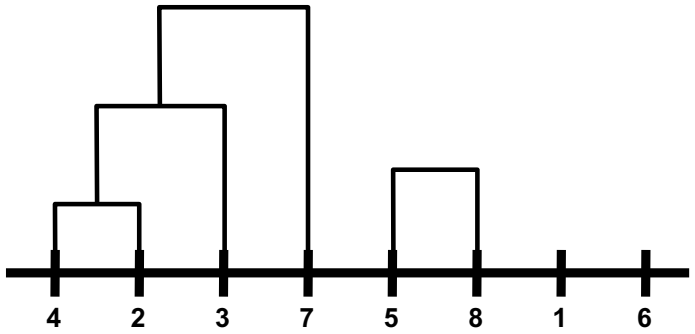


Join object 3 and cluster 1
Repeat process

Hierarchical clustering

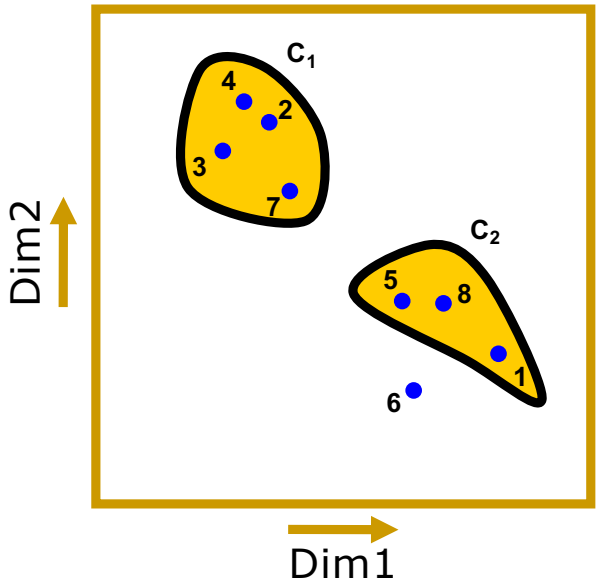


Dendrogram

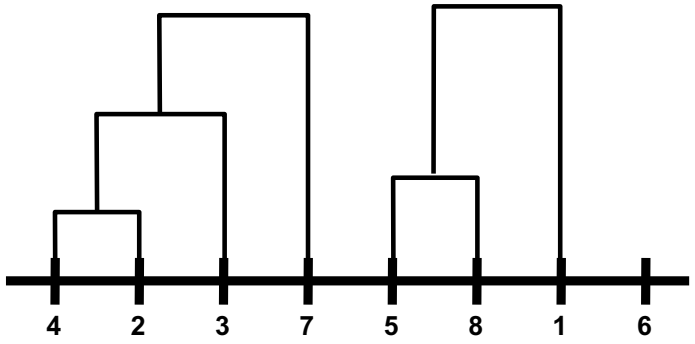


Join [object 7 and cluster 1] -> [cluster 1]
Repeat process

Hierarchical clustering

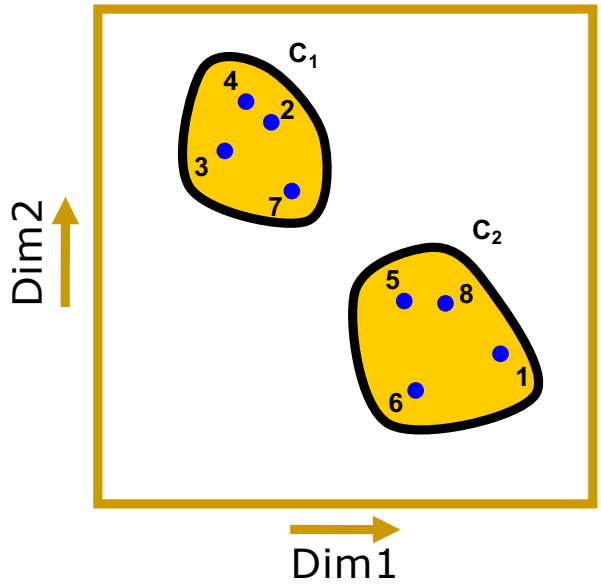


Dendrogram

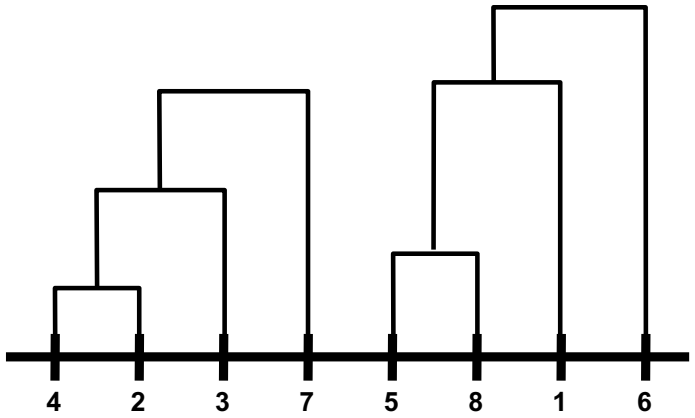


Join [object 1 and cluster 2] -> [cluster 2]
Repeat process

Hierarchical clustering

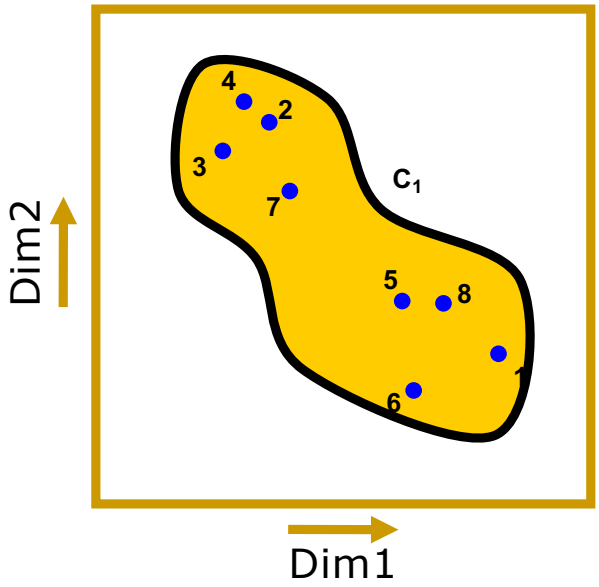


Dendrogram

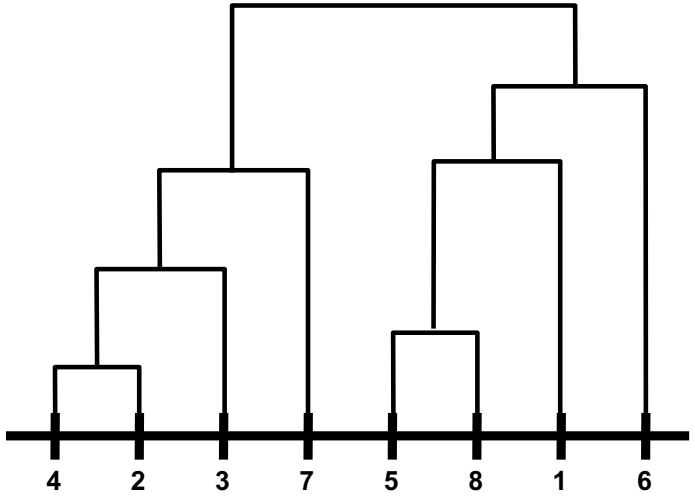


Join [object 6 and cluster 2] -> [cluster 2]
Repeat process

Hierarchical clustering

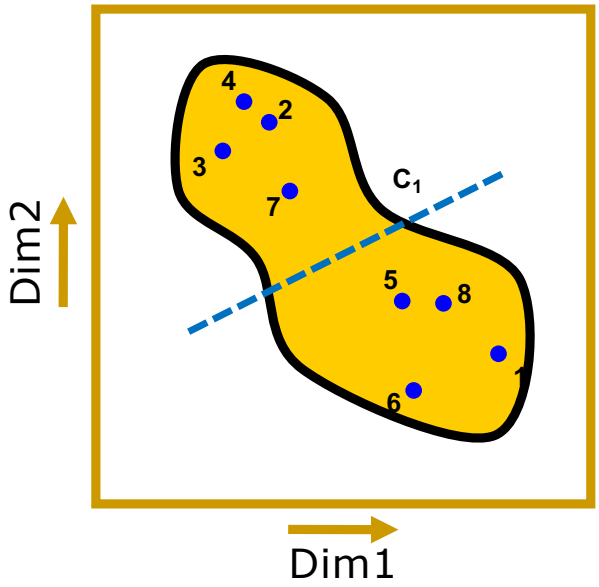


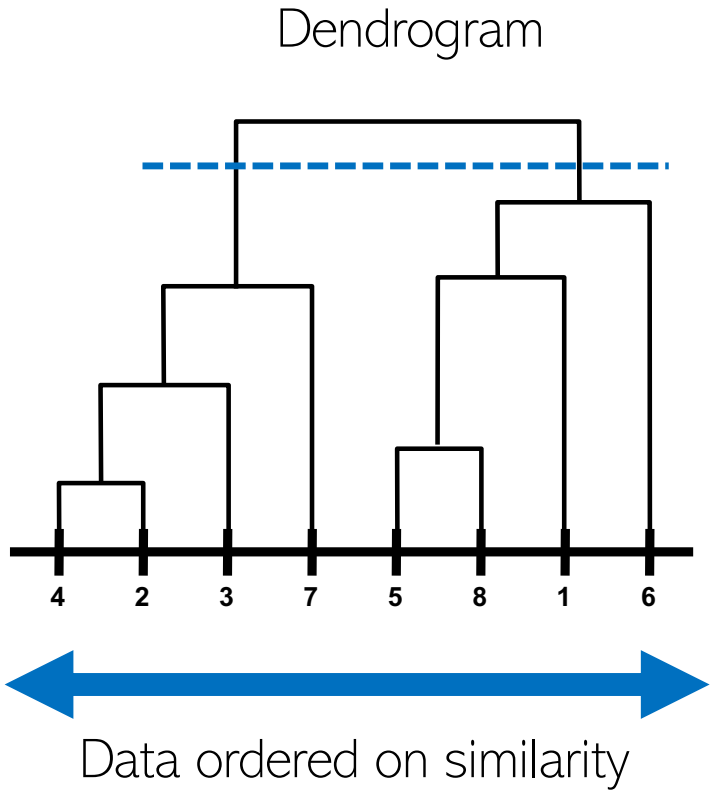
Dendrogram



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

Hierarchical clustering

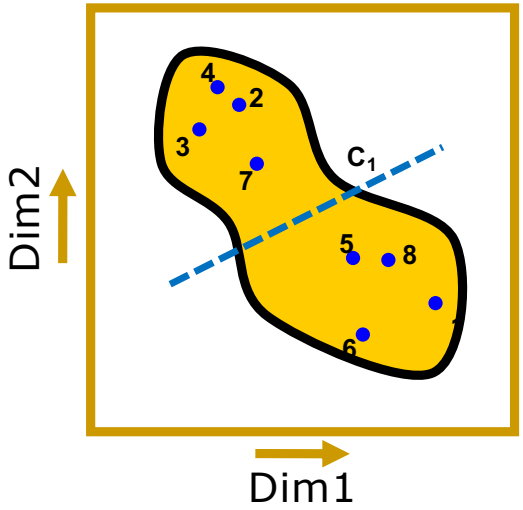


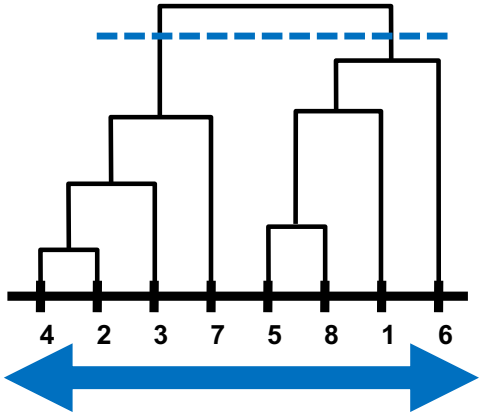


Hierarchical clustering

Need to know:

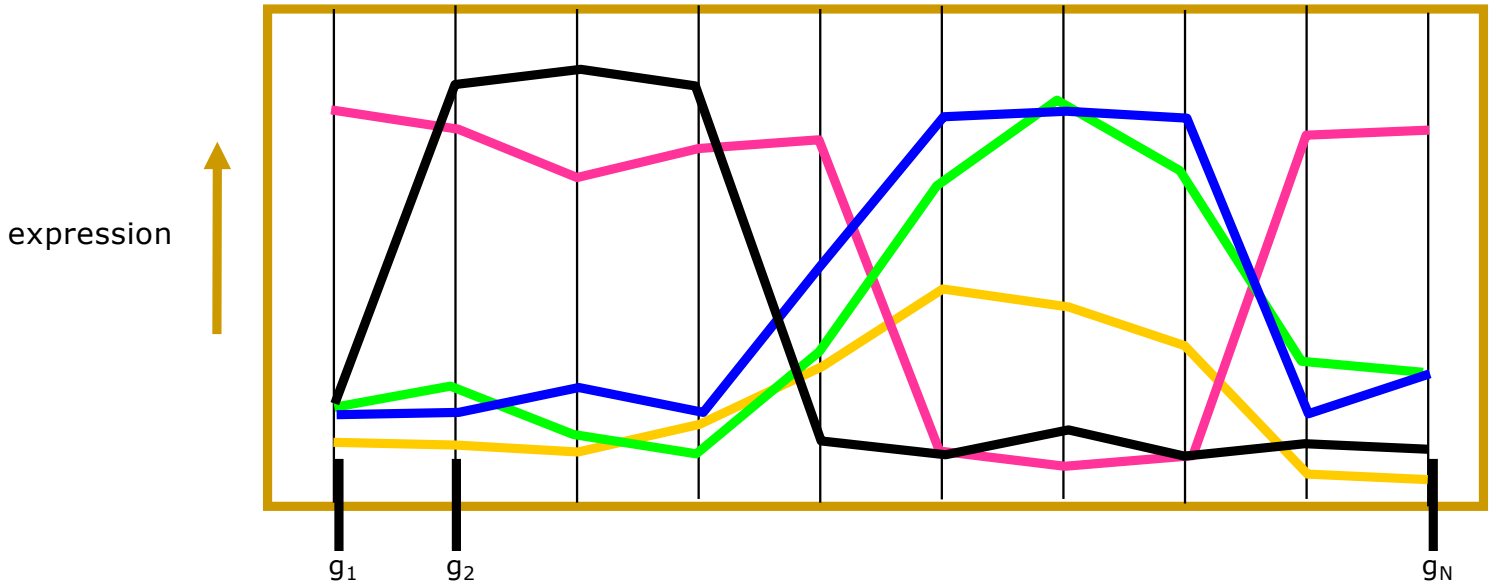
- Similarity between objects
- Similarity between clusters





Hierarchical clustering

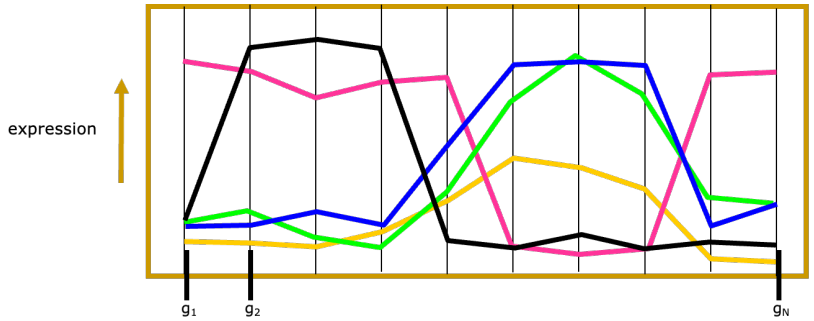
Similarity between objects



22

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

22

Mixed Pearson correlation

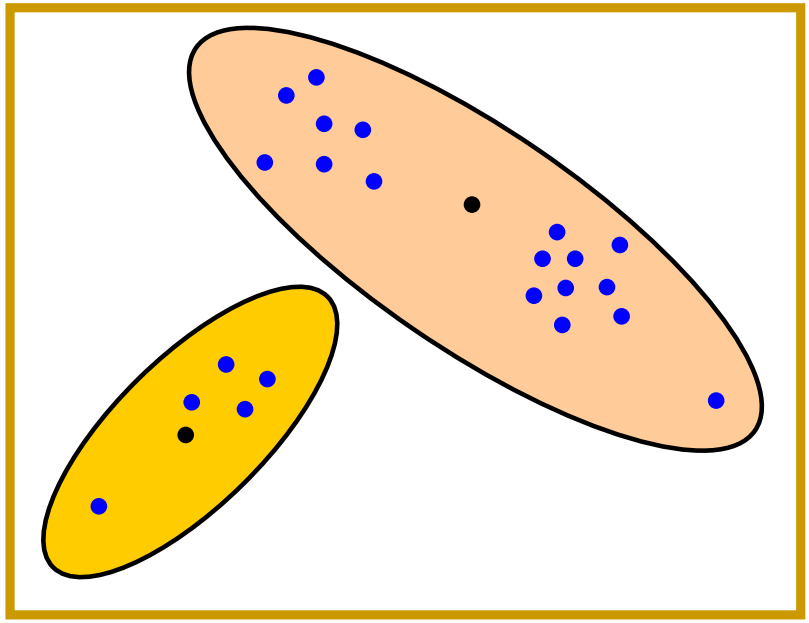
$$1 - |\rho_{ij}|$$

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

Ignore amplitude and sign

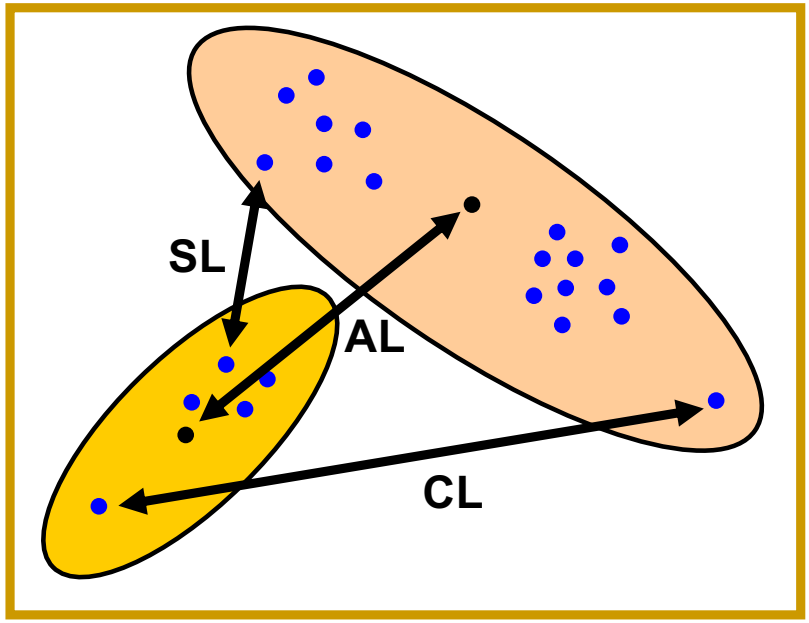
Hierarchical clustering

Similarity between clusters



Hierarchical clustering

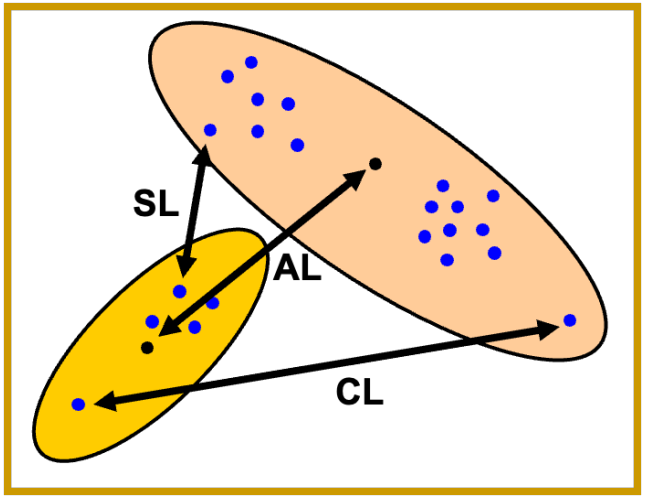
Similarity between clusters



Single linkage:	Closest objects
Complete linkage:	Furthest objects
Average linkage:	Average similarity

Hierarchical clustering

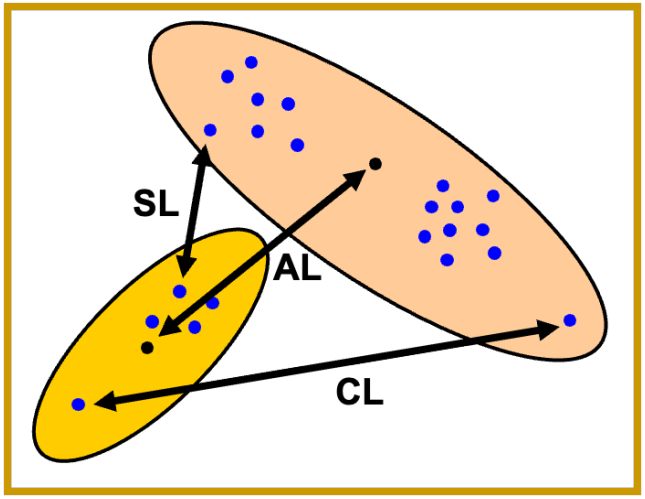
Similarity between clusters



- **Single linkage**: $d(X_i, X_j) = min(d(a,b) : a \in X_i, b \in X_j)$
- **Complete linkage**: $d(X_i, X_j) = max(d(a,b) : a \in X_i, b \in X_j)$
- **Average linkage**: $d(X_i, X_j) = \frac{1}{|X_i||X_j|} \sum_{a \in X_i} \sum_{b \in X_j} d(a,b)$

Hierarchical clustering

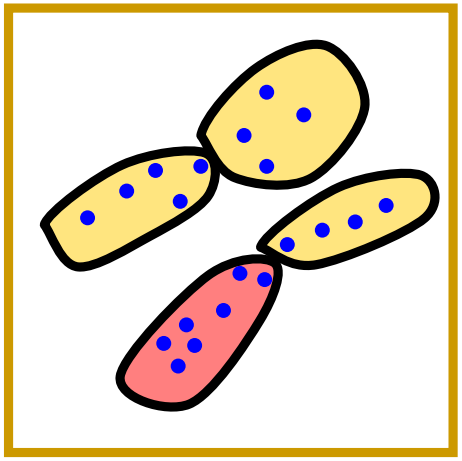
Similarity between clusters



- **Ward's method:** $d(X_i, X_j) = \|X_i - X_j\|^2$ (total variance)
- **Centroid distance:** $d(X_i, X_j) = \|c_s - c_t\|$

Hierarchical clustering

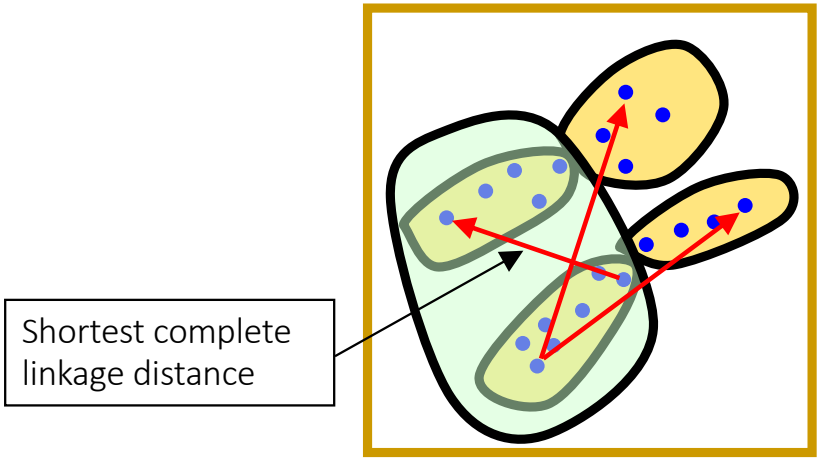
Similarity between clusters



Which cluster to merge with red cluster when using **complete linkage**?

Hierarchical clustering

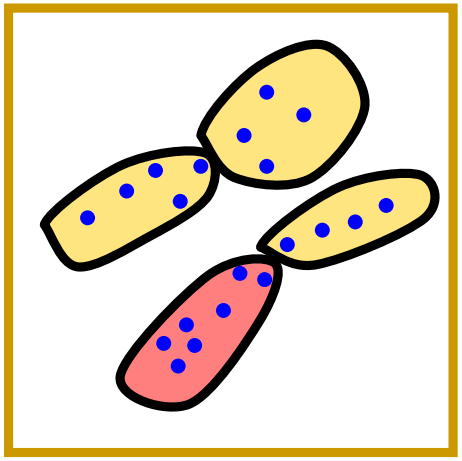
Similarity between clusters



Which cluster to merge with red cluster when using **complete linkage**?

Hierarchical clustering

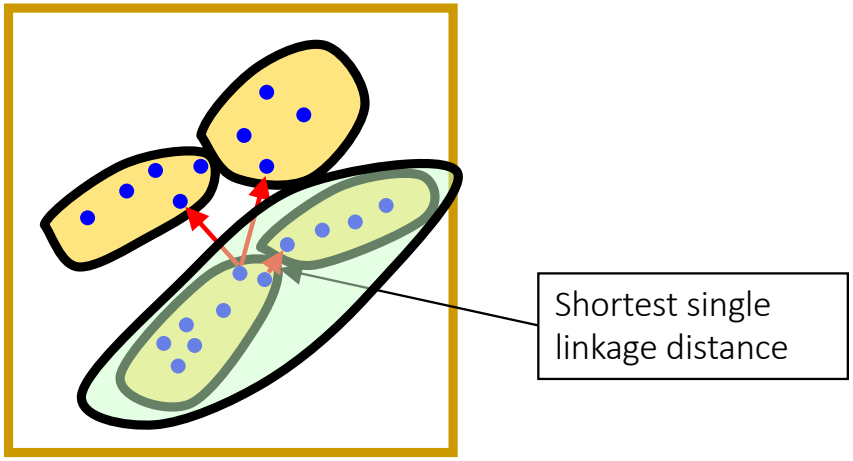
Similarity between clusters



Which cluster to merge with red cluster when using **single linkage**?

Hierarchical clustering

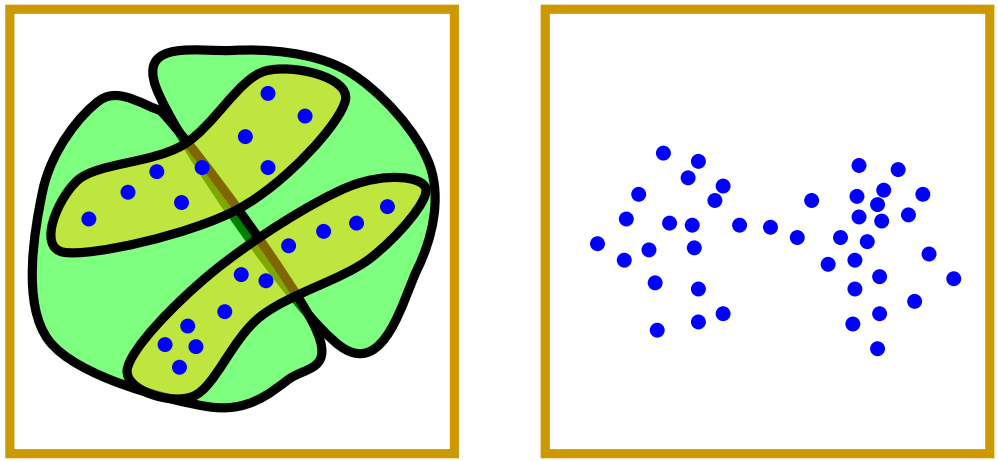
Similarity between clusters



Which cluster to merge with red cluster when using **single linkage**?

Hierarchical clustering

Similarity between clusters



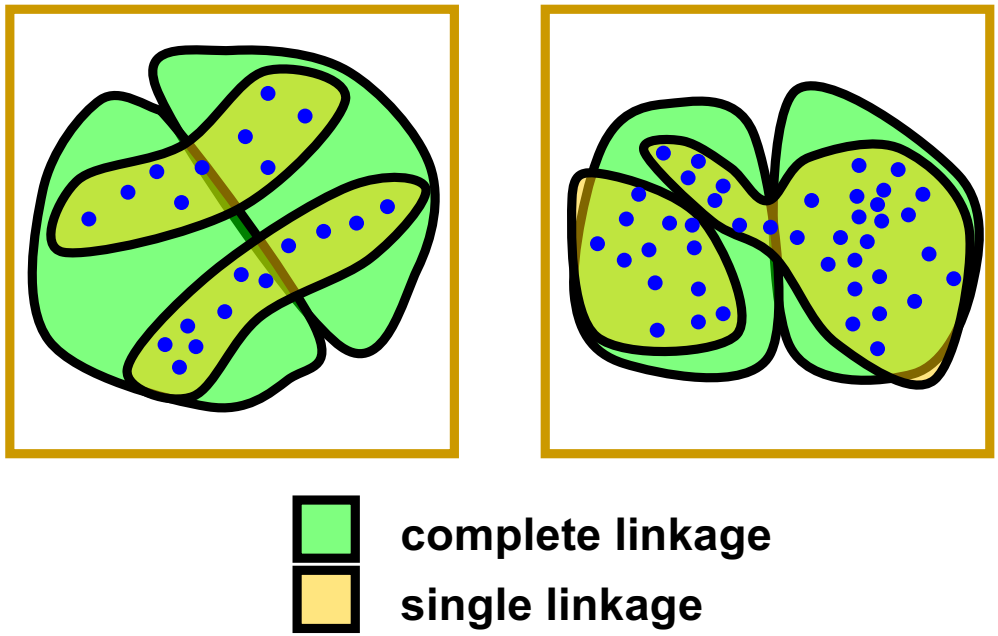
complete linkage

single linkage

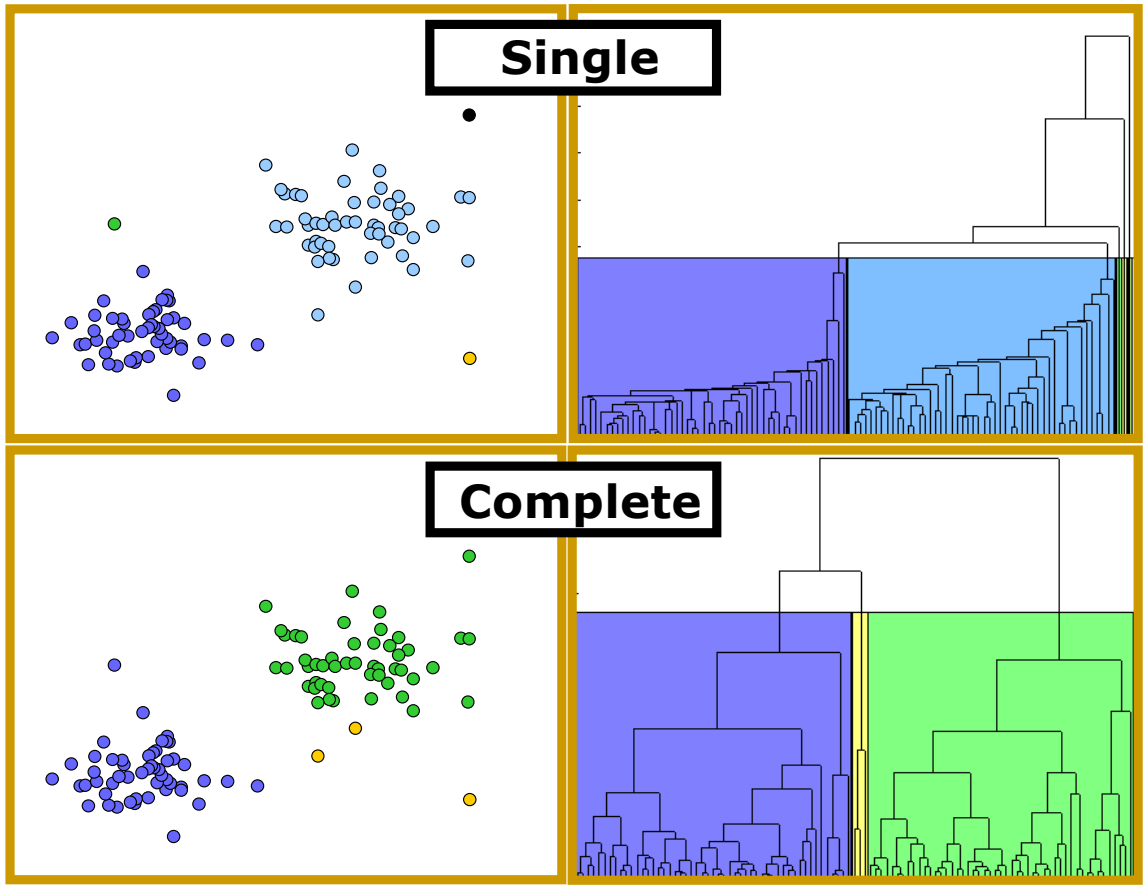
Hierarchical clustering

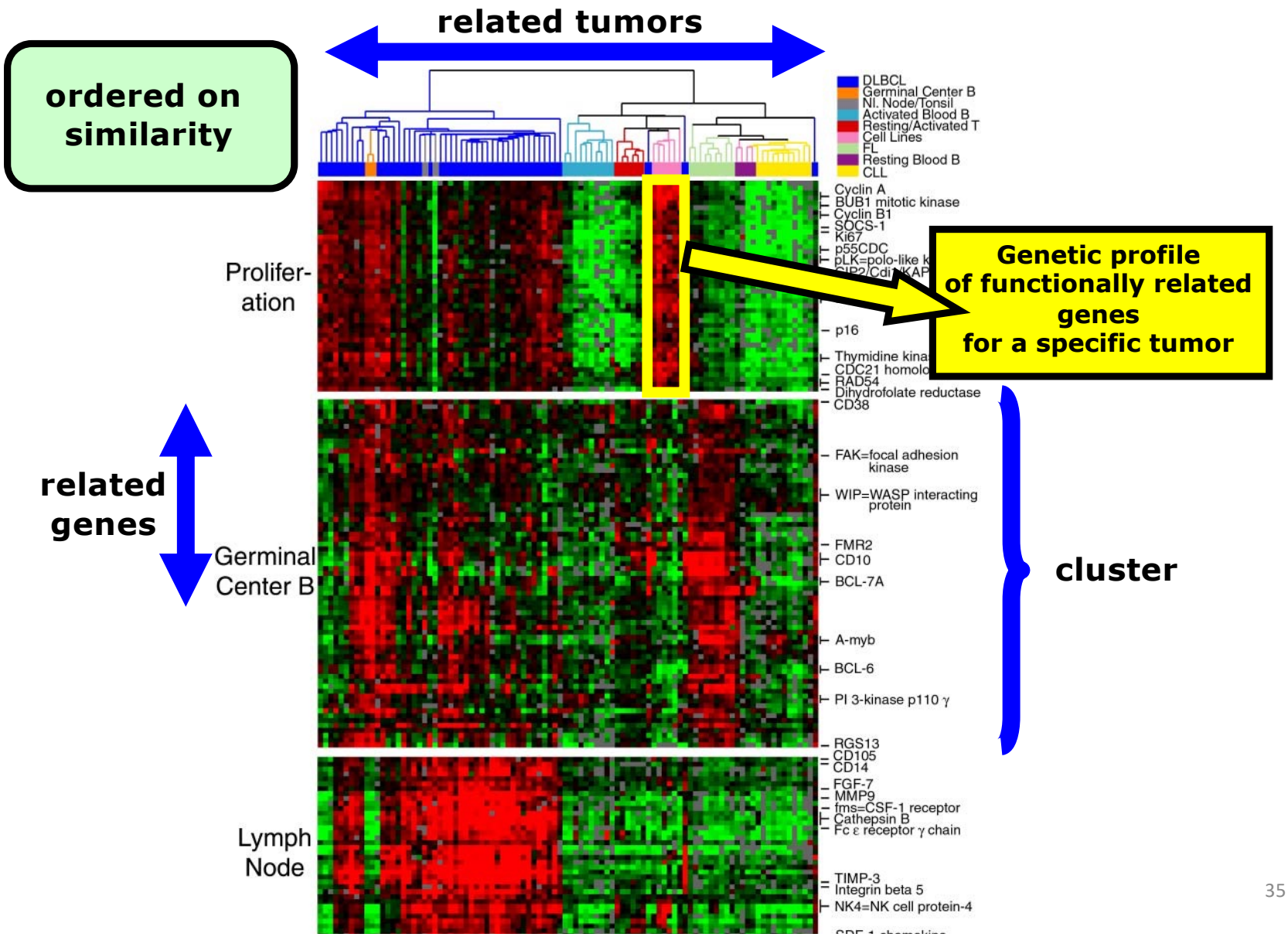
Similarity between clusters

- Single linkage → long and "loose" clusters
- Complete linkage → compact clusters



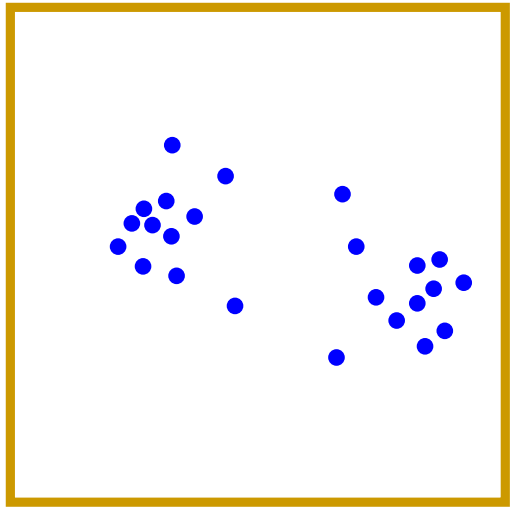
SL vs CL: Outlier influences



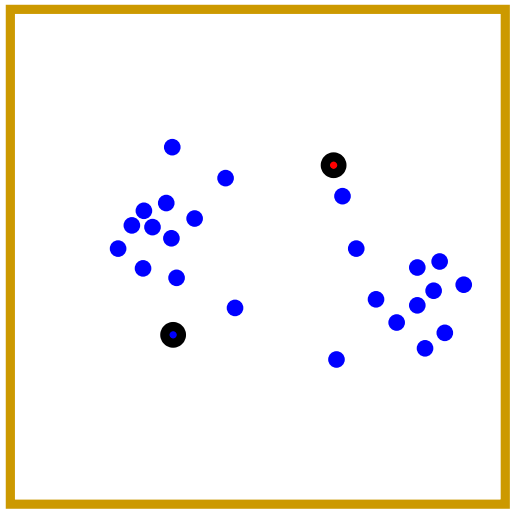

related tumors
ordered on similarity
DLBCL
Germinal Center B
Nl. Node/Tonsil
Activated Blood B
Resting/Activated T
Cell Lines
FL
Resting Blood B
CLL
Prolifer-ation
Cyclin A
BUB1 mitotic kinase
Cyclin B1
SOCS-1
Ki67
p55CDC
p16
Thymidine kina
CDC21 homolo
RAD54
Dihydrofolate reductase
CD38
Genetic profile of functionally related genes for a specific tumor
related genes
Germinal Center B
FAK=focal adhesion kinase
WIP=WASP interacting protein
FMR2
CD10
BCL-7A
A-myb
BCL-6
PI 3-kinase p110 γ
cluster
RGS13
CD105
CD14
FGF-7
MMP9
fms=CSF-1 receptor
Cathepsin B
Fc ε receptor γ chain
Lymph Node
TIMP-3
Integrin beta 5
NK4=NK cell protein-4

k-means clustering

k-Means clustering

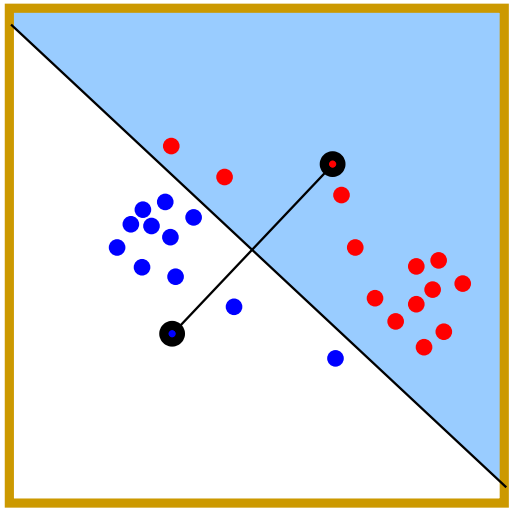


k-Means clustering



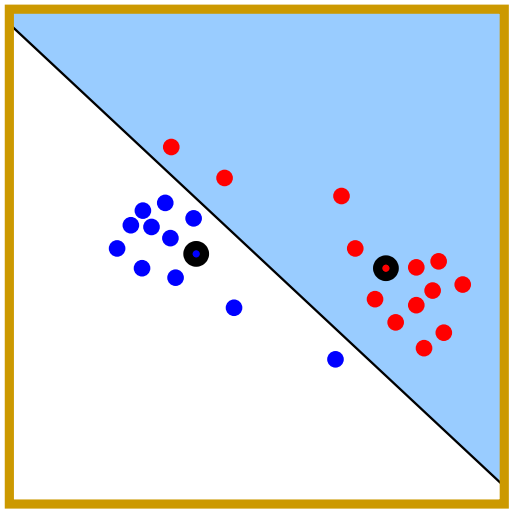
Choose randomly 2 prototypes

k-Means clustering



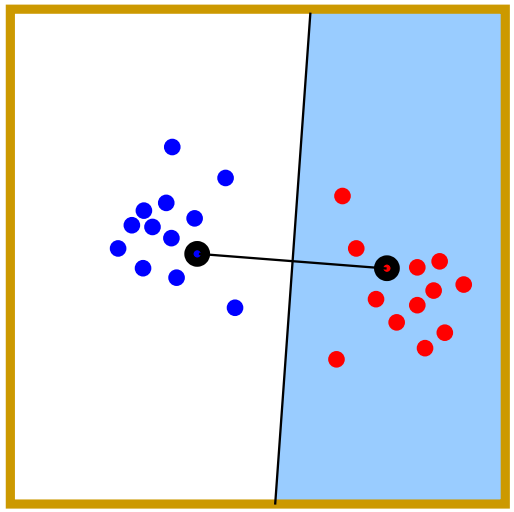
Assign objects to closest prototype
Blue area: cluster 1
White area: cluster 2

k-Means clustering



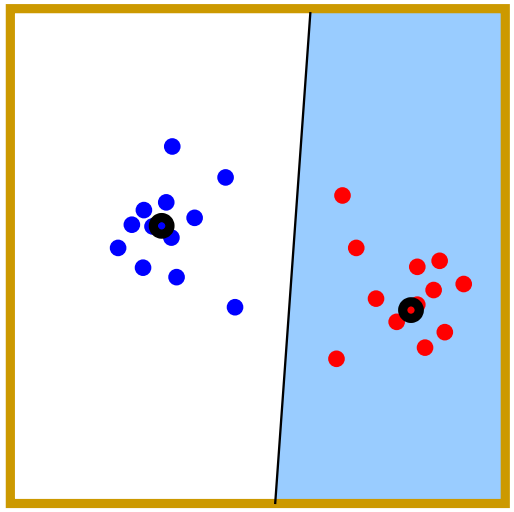
Calculate new cluster prototypes
By averaging objects

k-Means clustering



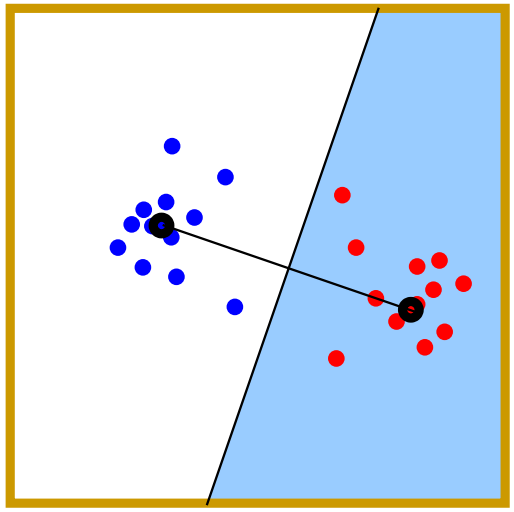
Re-assign objects to closest prototype
Blue area: cluster 1
White area: cluster 2

k-Means clustering



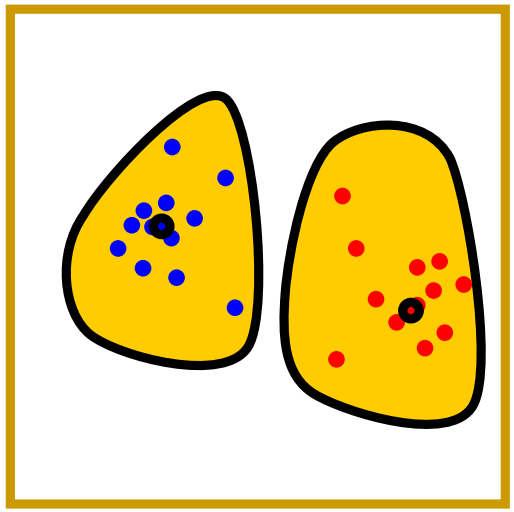
Re-calculate new cluster prototypes

k-Means clustering



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

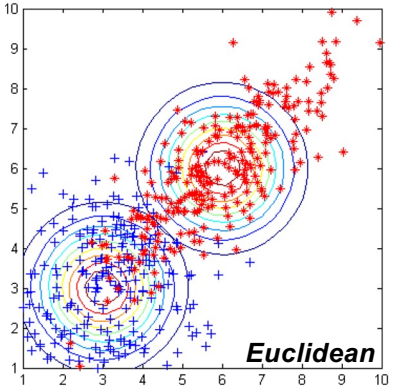
k-Means clustering

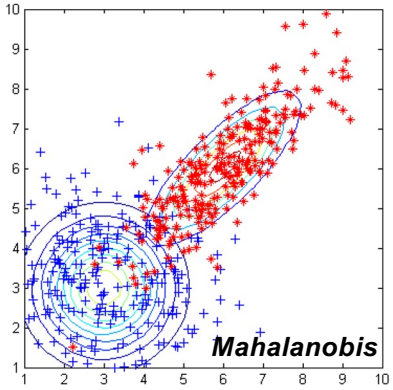


Establish clusters

k-Means clustering

Overview



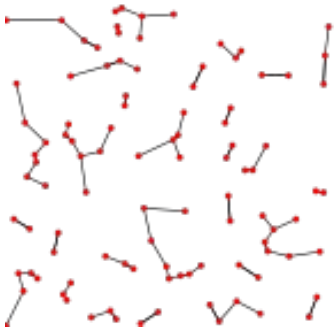


- *k*-Means
  - Fixed number of clusters (need to know a priori)
  - Choice of distance measure
  - Prototype choice
- Distance measure
  - Euclidean: Round clusters
  - Mahalanobis: Elongated clusters
- Prototype choice
  - Point
  - Line etc.
- Number of clusters
  - Validate clustering!

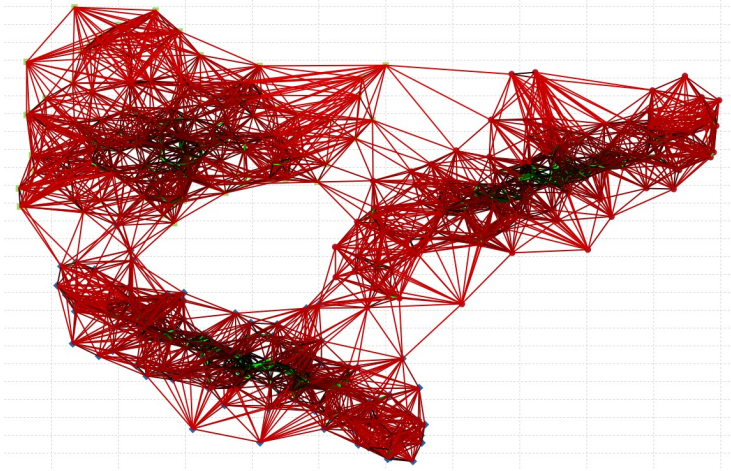
Graph-based Clustering

Graph-based clustering

- k-NN graph: connect every node to its k-nearest neighbors
- Find densely connected components (communities)



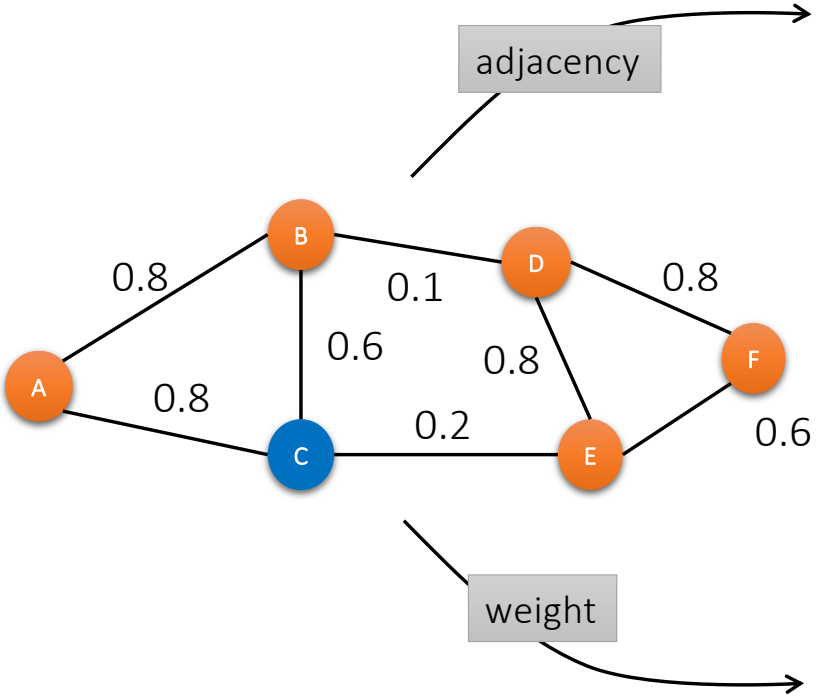
k=1



k=20

47

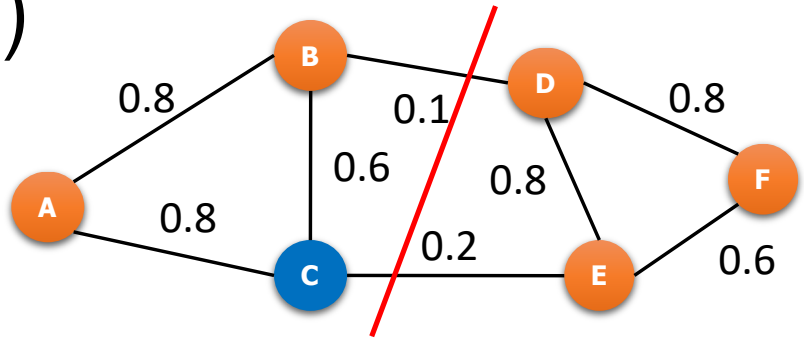
Graphs, adjacency and weight matrices



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

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

Spectral clustering (1)



- Minimise normalised cut
- Normalised cut
  between two clusters C_1 and C_2:

$$NC(C_1,C_2) = \frac{\text{cut}(C_1,C_2)}{assoc(C_1,V)} + \frac{\text{cut}(C_2,C_1)}{assoc(C_2,V)} = 2 - \left(\frac{assoc(C_1,C_1)}{assoc(C_1,V)} + \frac{assoc(C_2,C_2)}{assoc(C_2,V)}\right)$$

- cut(C_1, C_2) = weight of links between C_1 and C_2
- cut(C_2, C_1) = same
- assoc(C_1, V) = total weight of links from nodes in C_1 to entire graph
- assoc(C_2, V) = total weight of links from nodes in C_2 to entire graph

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 8, AUGUST 2000

Normalized Cuts and Image Segmentation

Jianbo Shi and Jitendra Malik, *Member, IEEE*

Louvain method for community detection

Maximize modularity

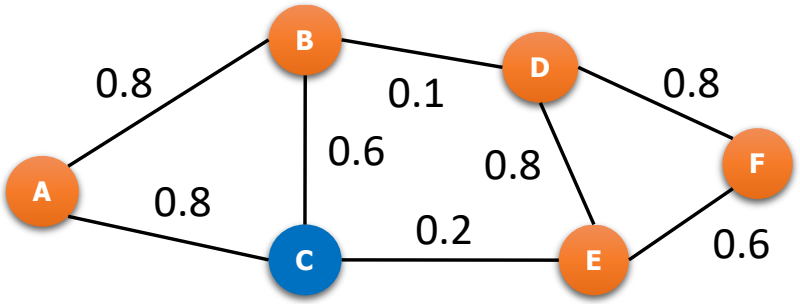
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

Expected edges in cluster c

Observed edges in cluster c

Q_{ij}

i , j	nodes
c_i , c_j	community of node i and j
A_{ij}	weight edge node i and j
k_i , k_j	total weight of edges of nodes i and j
$\delta(c_i,c_j)$	1 if c_i=c_j, zero otherwise
m	total weight of edges



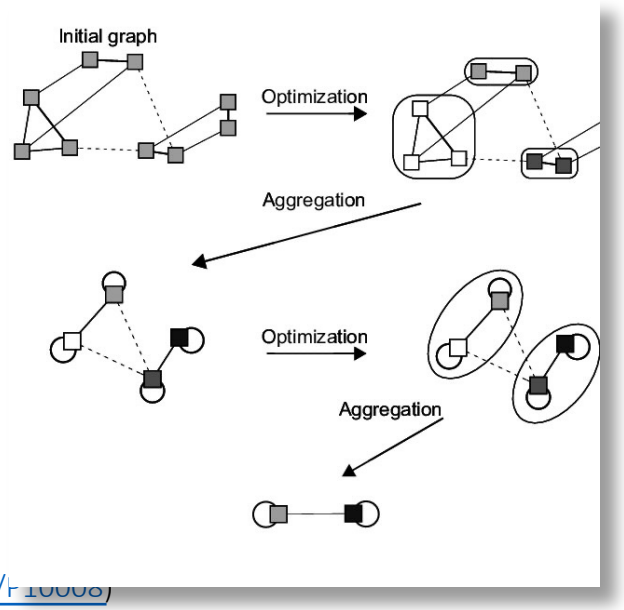
Blondel et al. (https://doi.org/10.1088/1742-5468/2008/10/P10008)

Louvain method for community detection

Modularity optimized in two phases, applied iteratively

Phase 1: Optimization (communities); moving nodes from own community to community of neighboring node (*improve Q*)

Phase 2: Aggregation; building new network by aggregating nodes in detected communities



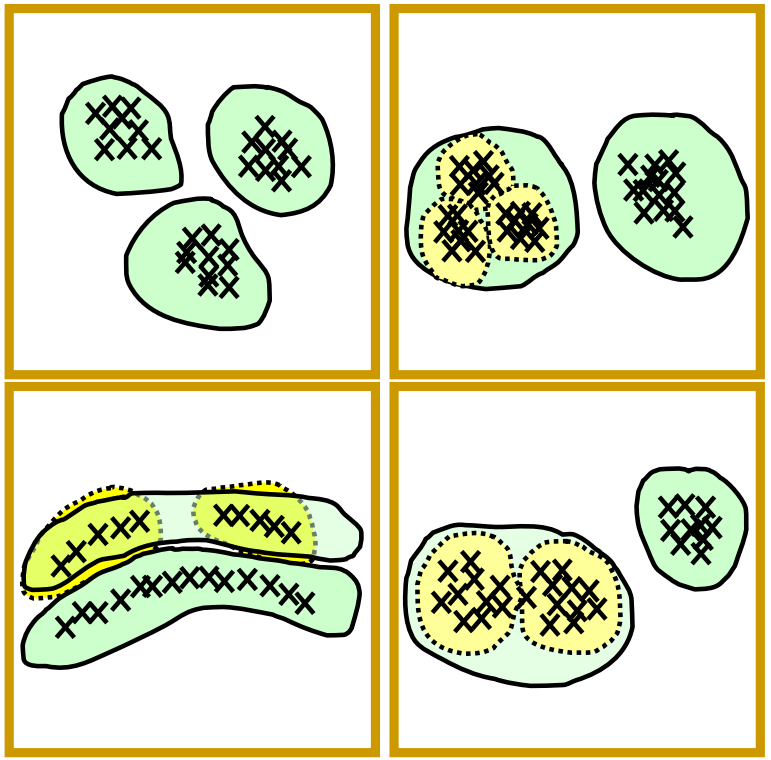
Blondel et al. (https://doi.org/10.1088/1742-5468/2008/10/P10008)

Outline

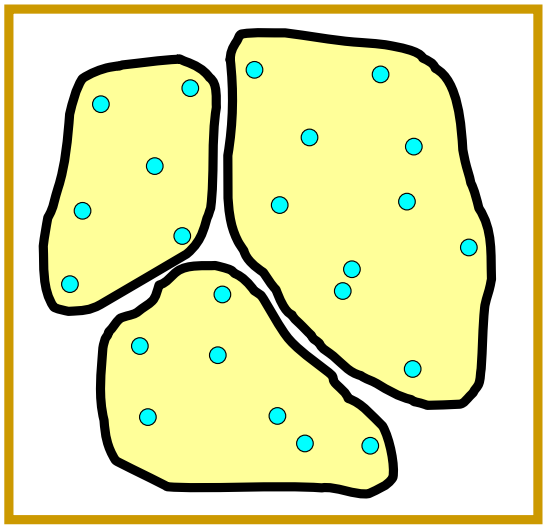
- Introduction to clustering
  - Hierarchical clustering
  - *k*-Means clustering
  - Graph-based clustering
- **Cluster validation**
- scRNA-seq clustering
- Annotating clusters

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



Cluster Validation



- Cluster tendency

  Clustering IMPOSES structure even though data may not posses it

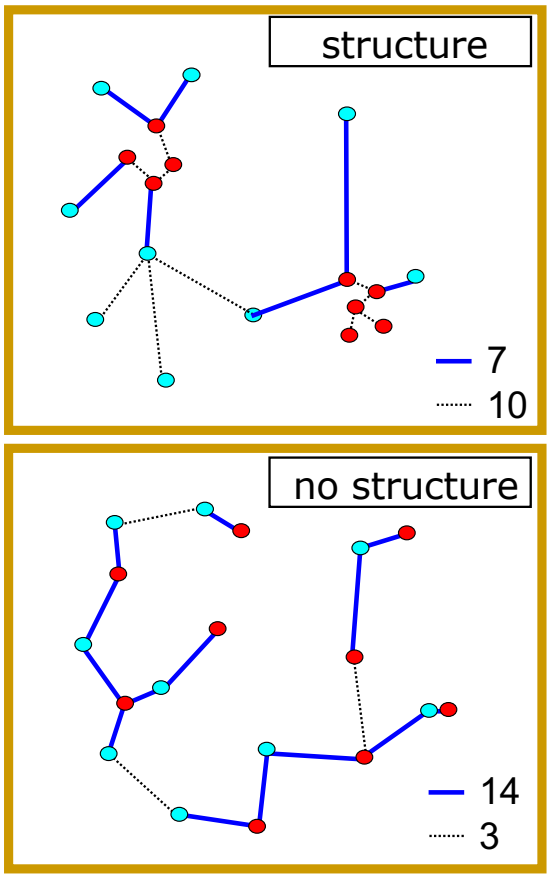
  Aim: Test whether data possesses structure

- Cluster validity

  Choices impose restrictions on for example shape

  Aim: Quantitative evaluation of the clustering results

Test for spatial randomness



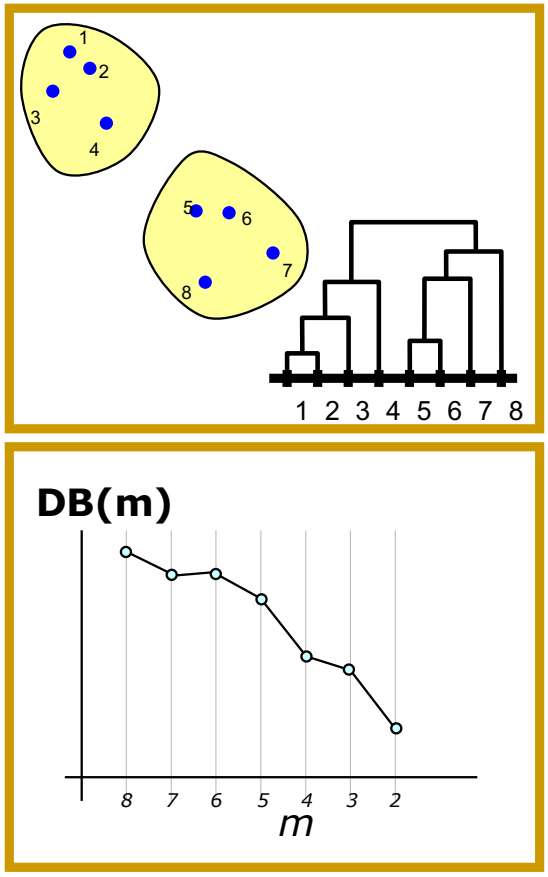
- Test

  If data (•) clusters frequently with random data (•) then data structureless

- Approach
  - Generate random vectors (**Y**) uniformly over observed region of data (**X**)
  - Find MST (single linkage HC) of **X** ∨ **Y**
  - Determine number of edges *q* that connect vectors of **X** with **Y**
  - If **X** contains clusters *q* should be small!

  (multiple random vs random measurements gives likelihood for *q*)

Davis-Bouldin index



- Test

  Select specific clustering according to a criteria

  For example: Davis-Bouldin index

- DB index

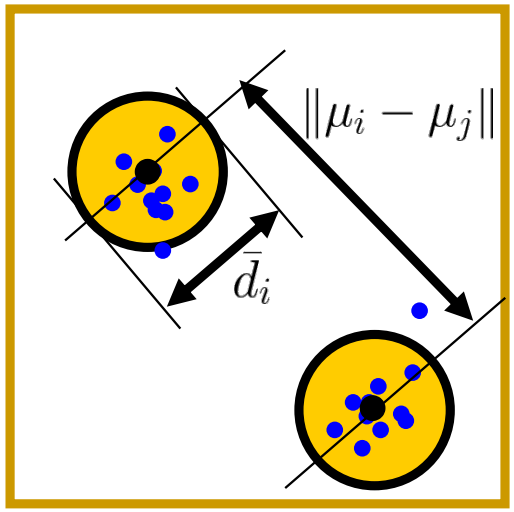
  For a specific clustering m, **DB**(**m**) : Average similarity of a cluster with its most similar cluster

- Approach

  Goal: Clusters to have minimal similarity

  Seek: Clustering that minimize *DB*(*m*) wrt *m*

Davis-Bouldin index



- Similarity cluster C_i and C_j

$$D_{i,j} = \frac{(\bar{d}_i + \bar{d}_j)}{\|\mu_i - \mu_j\|}$$

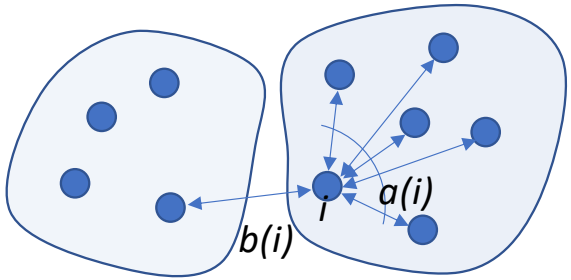
- $\bar{d}_i$:average distance within cluster *i*, μ_i :centroid of cluster *i*
- Most similar cluster to C_i

$$R_{i,j} = \max_{j \neq i} \{D_{i,j}\}$$

- DB index

$$DB = \frac{1}{k} \sum_{k=1}^{k} R_{i,j}$$

Silhouette score



- Measure similarity of object to its own cluster

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

with $a(i)$ being average distance to all objects in same cluster and $b(i)$ being closest object from all other clusters:

$$a(i) = \frac{1}{|C_i|} \sum_{\forall j} d(x_i, x_j) \qquad b(i) = min_{\forall j, j \notin C_i} d(x_i, x_j)$$

$-1 \leq s(i) \ll 1$; $s(i)$ is close to 1, if $a(i) \ll b(i)$; average distance within cluster much smaller than nearest objects

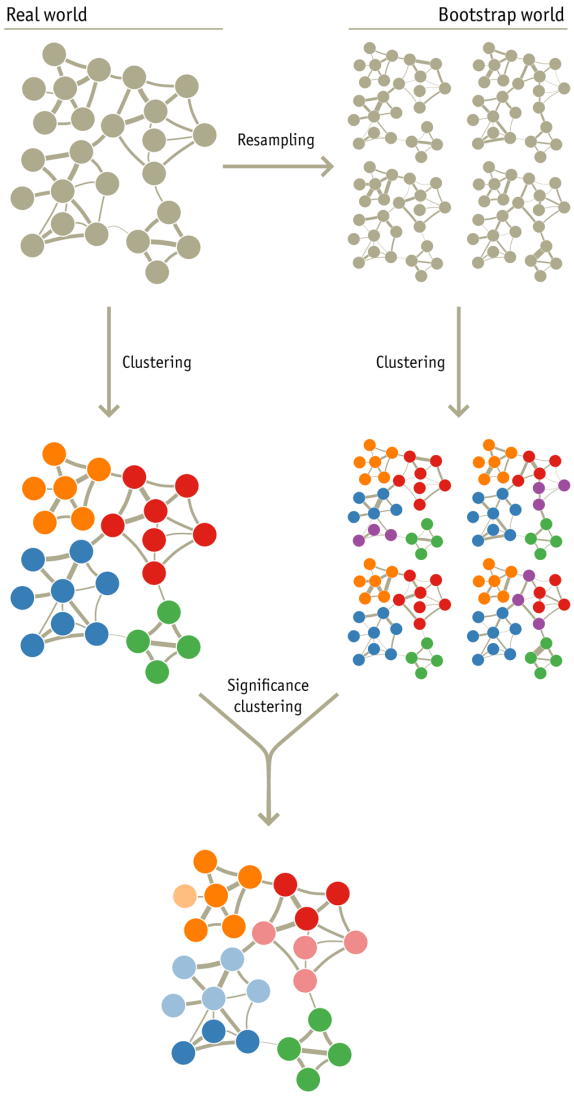
- Silhouette score is average of all these similarities

$$S = \frac{1}{N} \sum S(i)$$

Bootstrapping

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

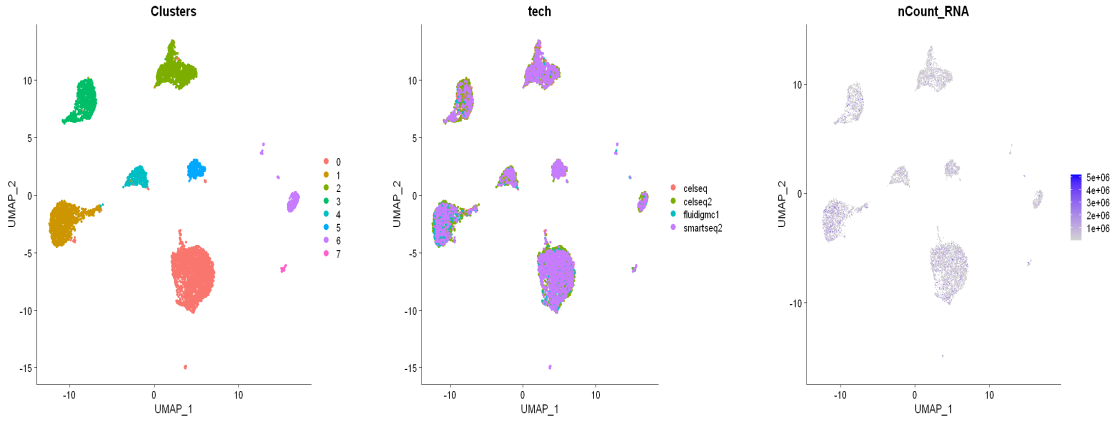
- Take a random set of cells
- Cluster these cells
- Comparing clusters to orginal clustering
- Estimate degree of support for clustering

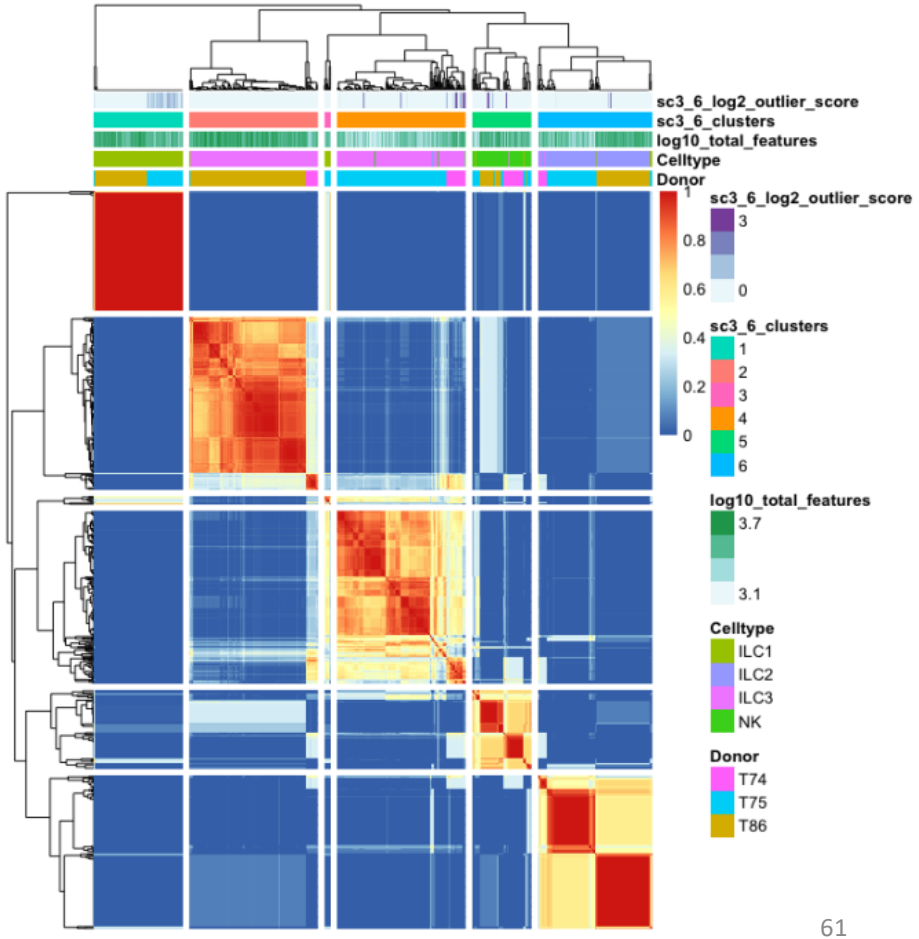


Rosvall and Bergstrom (https://doi.org/10.1371/journal.pone.0008694)

Always check QC data

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





Outline

- Introduction to clustering
  - Hierarchical clustering
  - *k*-Means clustering
  - Graph-based clustering
- Cluster validation
- **scRNA-seq clustering**
- Annotating clusters

scRNA-seq clustering methods

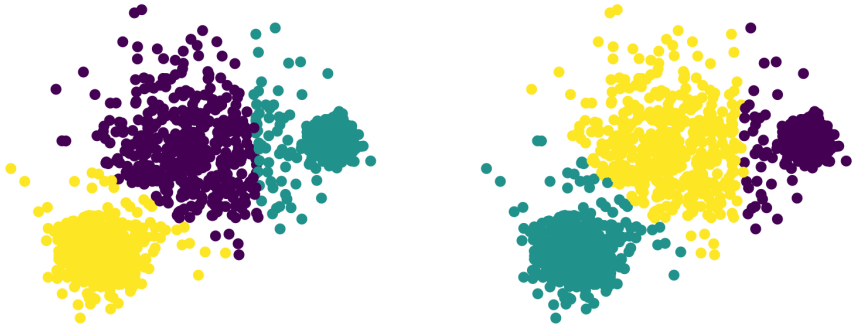
Name	Year	Method type	Strengths	Limitations
scanpy[4]	2018	PCA + graph-based	Very scalable	May not be accurate for small data sets
Seurat (latest)[3]	2016			
PhenoGraph[32]	2015			
SC3 (REF.[22])	2017	PCA + *k*-means	High accuracy through consensus, provides estimation of *k*	High complexity, not scalable
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)

Comparing different cluster labels

Adjusted Rand Index (ARI)

Measure of the similarity between two data clusterings



Given a set S of n elements, and two groupings or partitions (clusterings) of these elements

$X = \{X_1, X_2, \ldots, X_r\}, \quad Y = \{Y_1, Y_2, \ldots, Y_r\}$

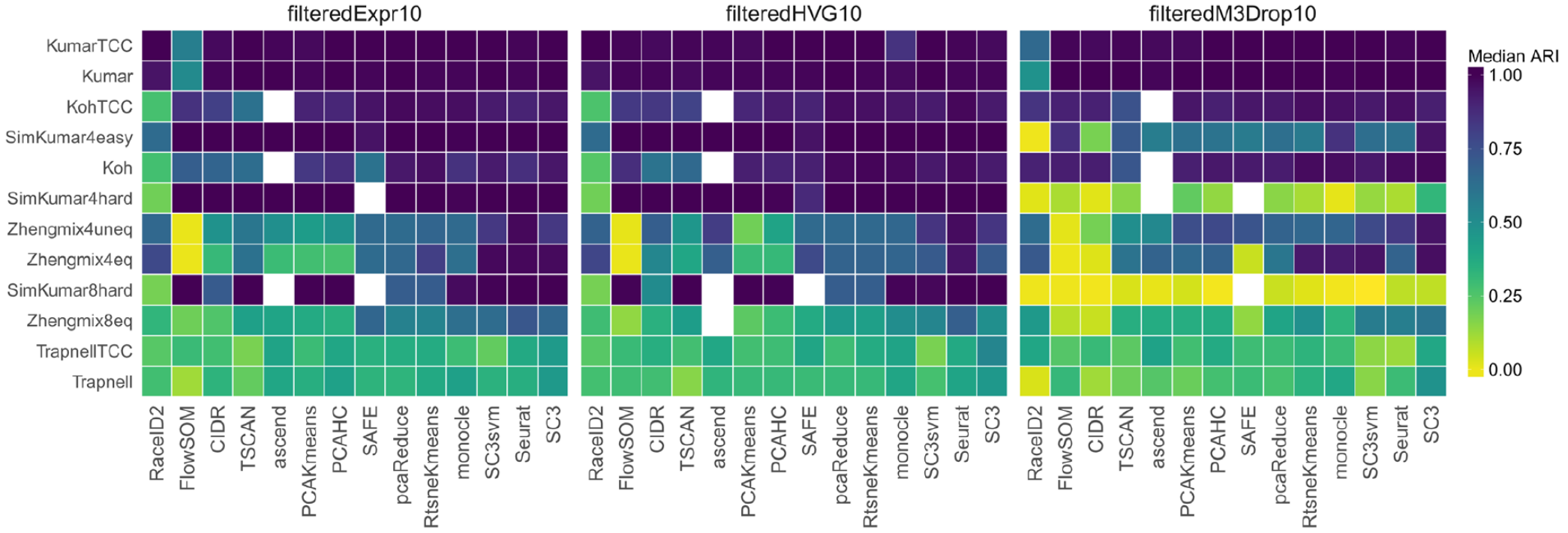
$X \backslash Y$	Y_1	Y_2	$\ldots$	Y_s	Sums
X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums	b_1	b_2	$\ldots$	b_s	

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

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

-1 ≤ ARI ≤ 1 (1: perfect overlap, 0: random overlap)

Benchmarking scRNA-seq clustering methods

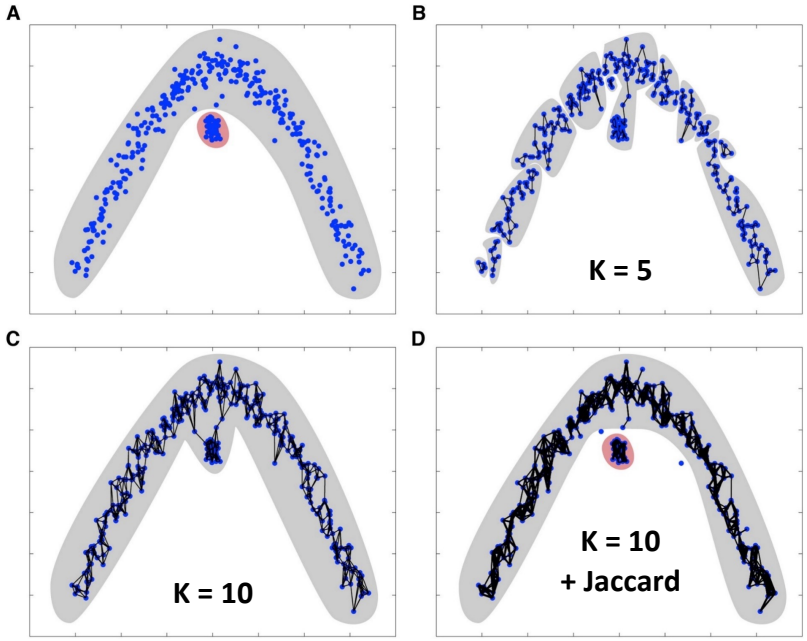


Row: different data set, Panel: different gene filtering, Column: different clustering method

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

Standard clustering approach

1. Select highly variable genes (~1000-5000 genes)
2. Reduce dimensions using PCA (~30-50 dimensions)
3. Construct kNN graph (~15-20 neighbors)
4. Louvain/Leiden community detection



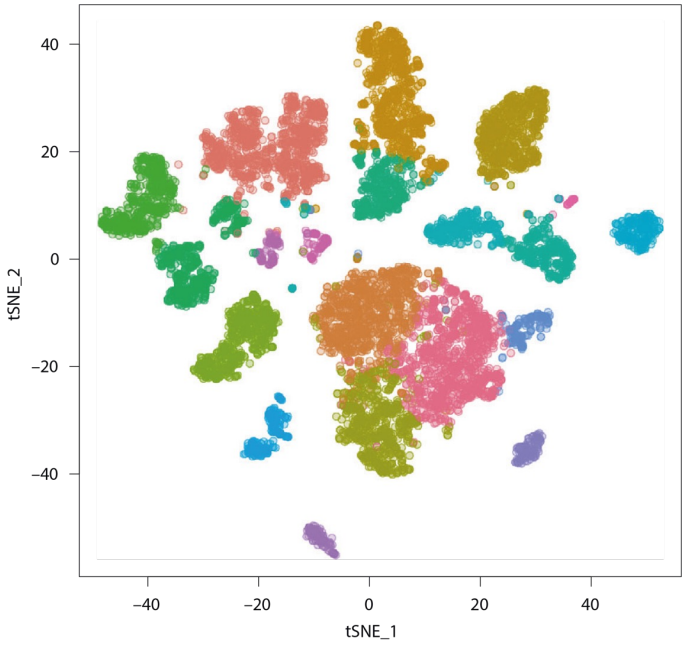
Levine et al. (https://doi.org/10.1016/j.cell.2015.05.047)

Outline

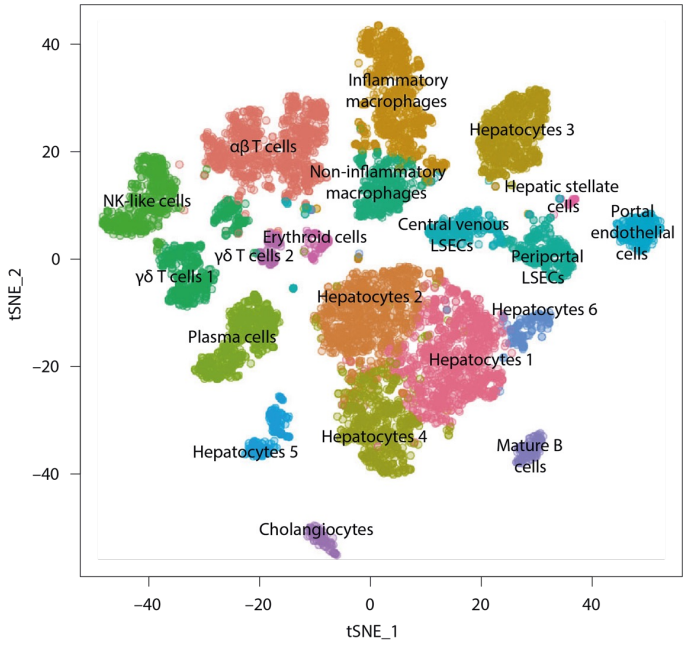
- Introduction to clustering
  - Hierarchical clustering
  - *k*-Means clustering
  - Graph-based clustering
- Cluster validation
- scRNA-seq clustering
- Annotating clusters

From clusters to annotations

Clustering



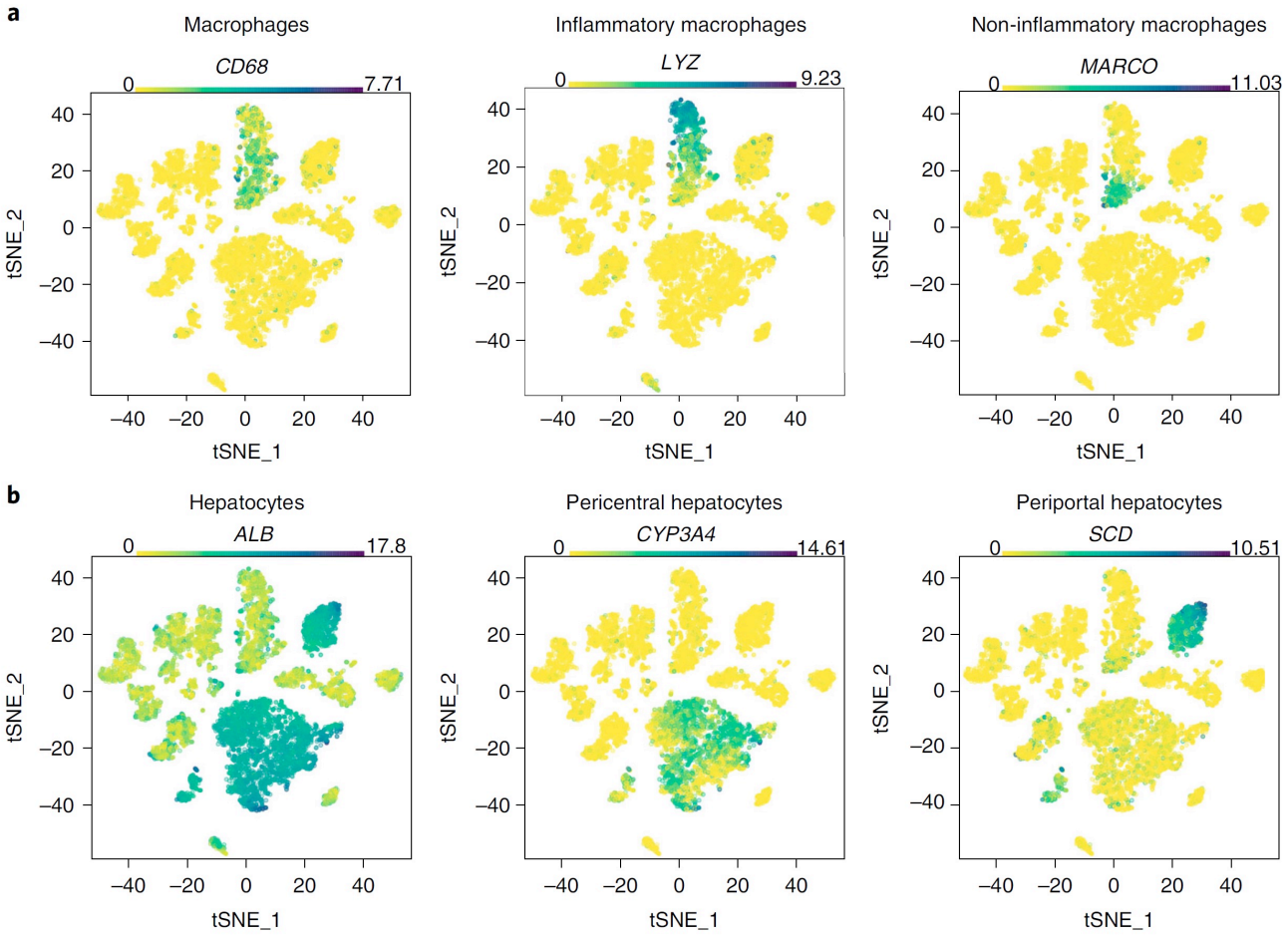
Annotation



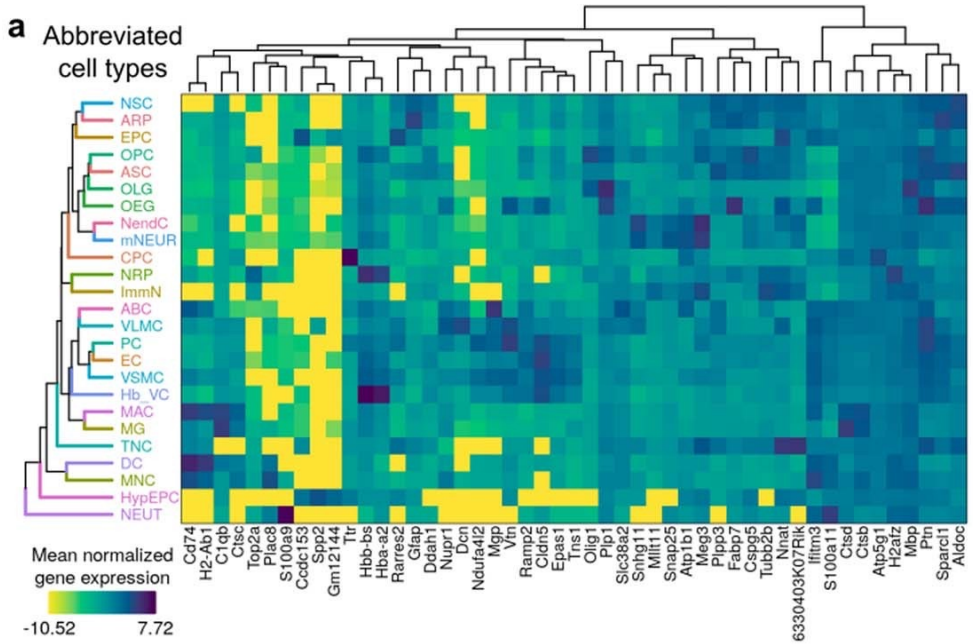
Gene expression overlay

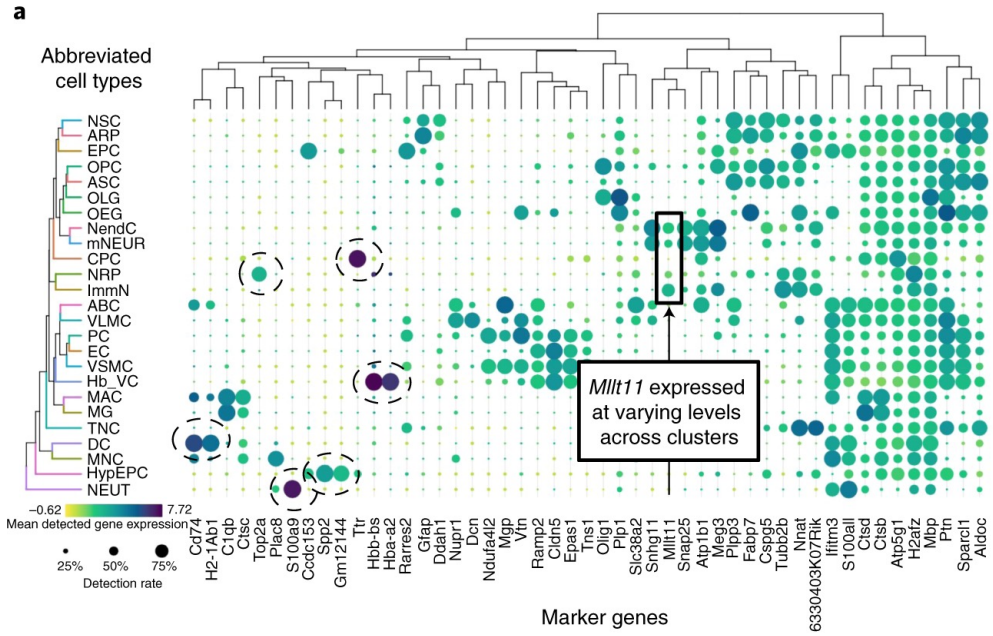
Easy

Challenging



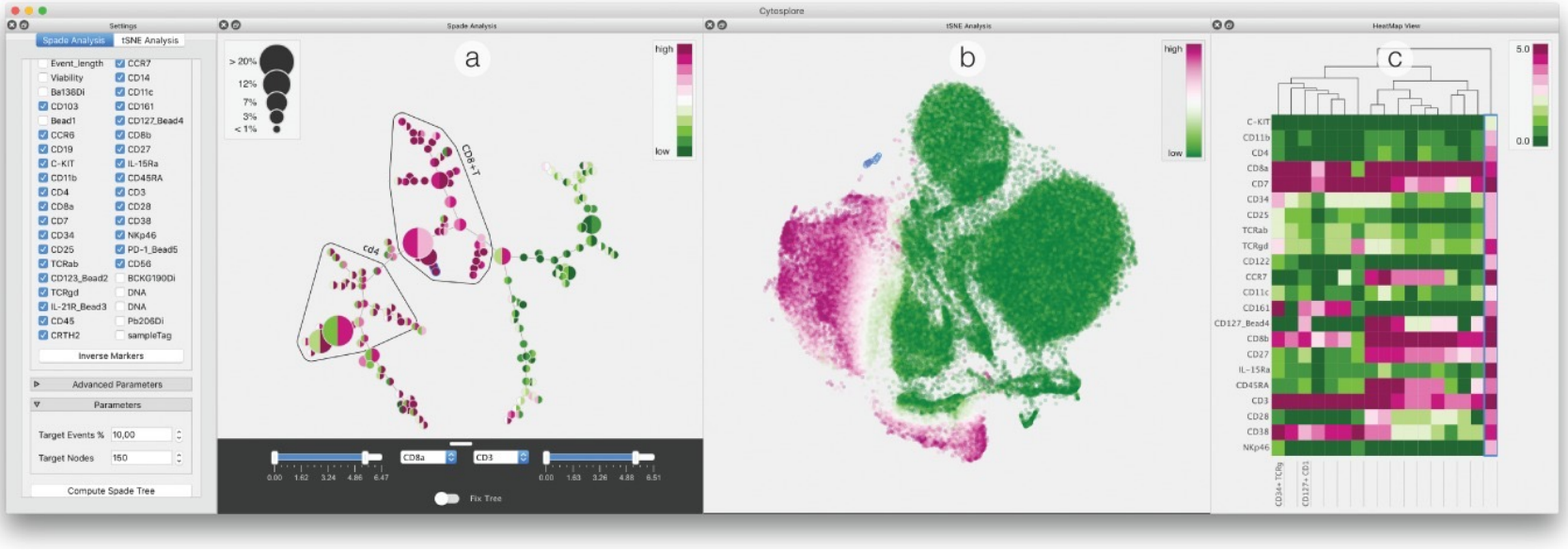
Alternatively: heatmaps & dot plots





Interactive visualization is important

- *Interactive* tools: Cytosplore, Loupe, cellxgene, …
- *Iterative* visualization: Seurat, scanpy,…



https://www.cytosplore.org/

Where do we get these marker genes?

Ideally: from a single cell atlas from a relevant organism, organ and disease context

- "Expert knowledge"
- Literature
- Other scRNA-seq data
- Marker databases: PangaloDB, CellMarker,…

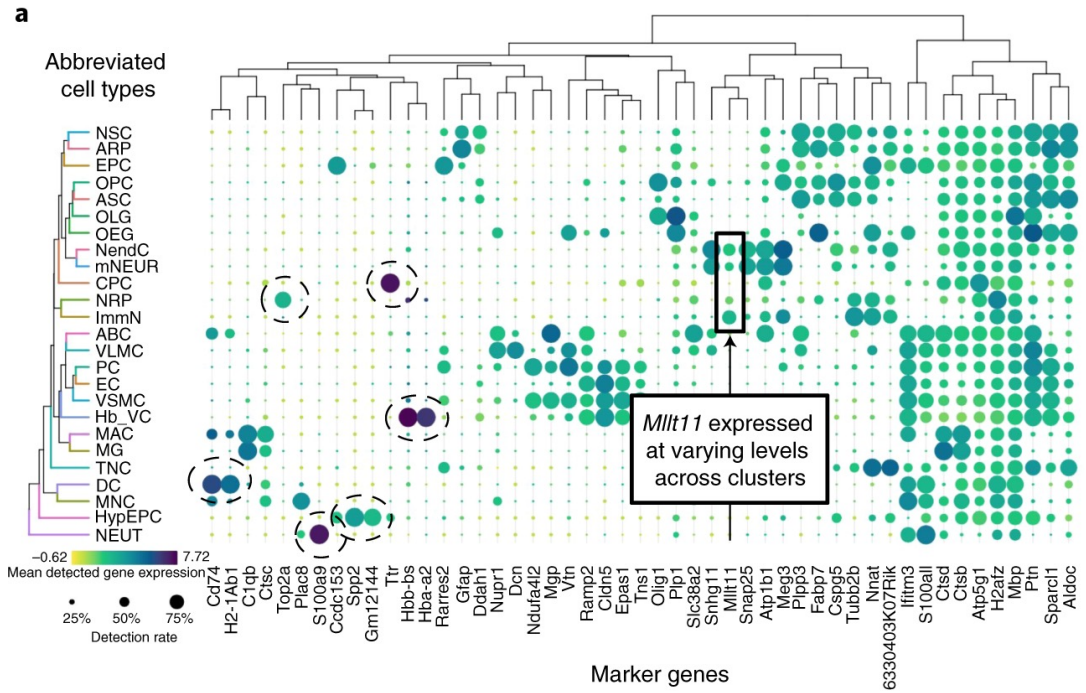
Challenges:

- Few well-known markers
- Some well-known markers may not be as specific as expected

What if I don’t have that many markers

- Identify “novel” markers by computing differential expression between a cluster and all other cells or between pairs of clusters

- Manually research markers to find functional information that may help identify the cell type

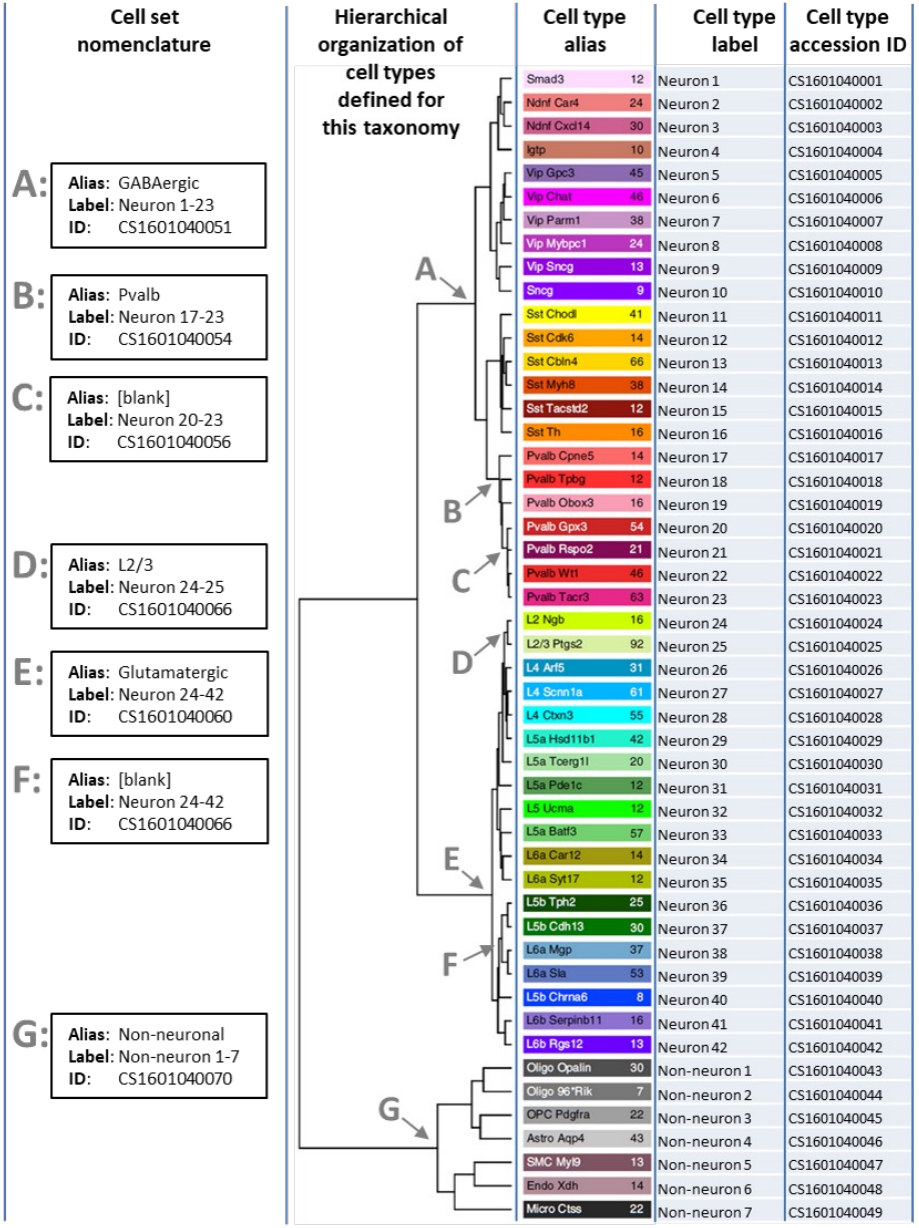


Annotation verification

1. Using *independent* data (e.g. fluorescence in situ hybridization)
2. Multi-modal single cell data
   - SNVs & CNVs
   - TCR/BCR
   - scRNA-seq+scATAC (mRNA + accessibility)
   - CITE-seq (surface proteins + mRNA)

Nomenclature

- How should we name cells?



Miller et al. (eLife 2020)
https://portal.brain-map.org/explore/classes/nomenclature

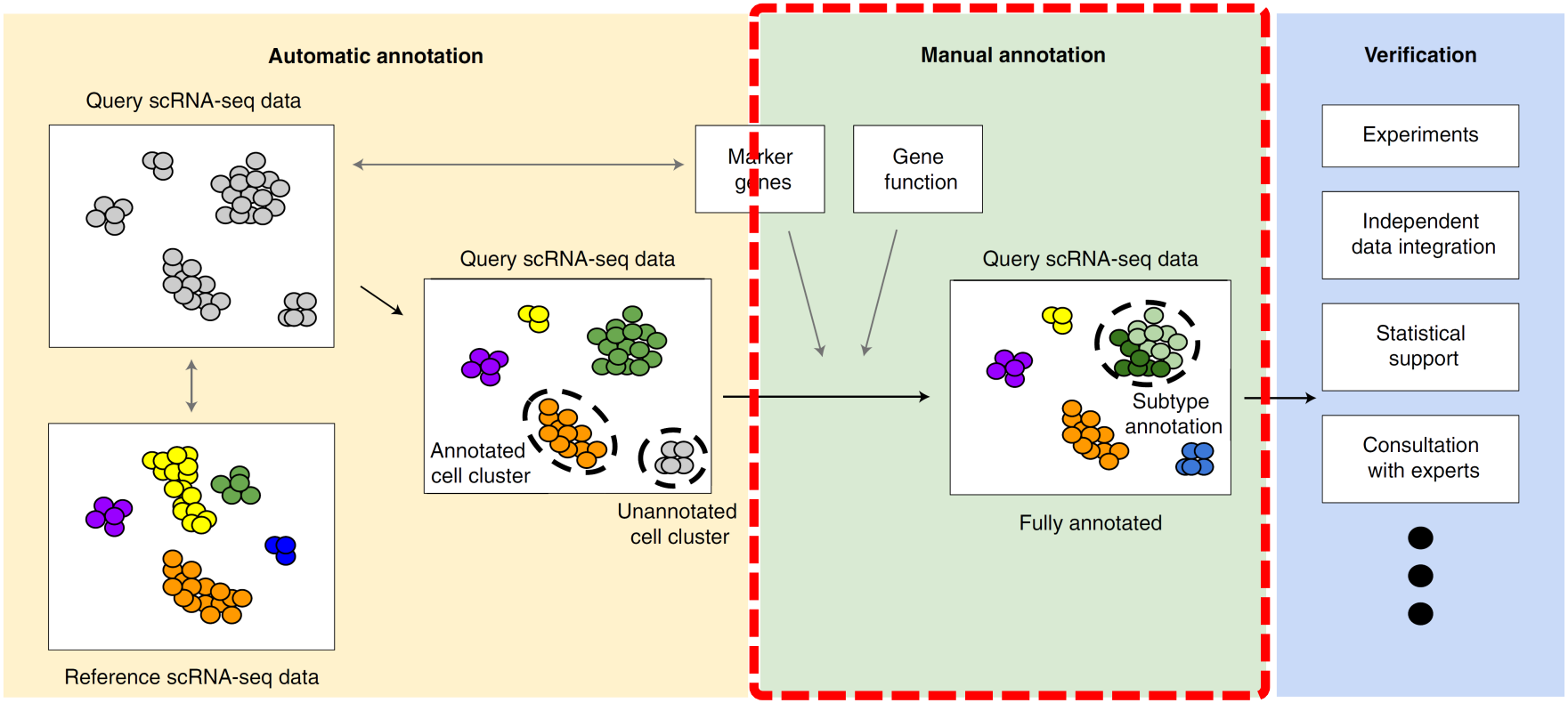
nature protocols

REVIEW ARTICLE

https://doi.org/10.1038/s41596-021-00534-0

Check for updates

Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods

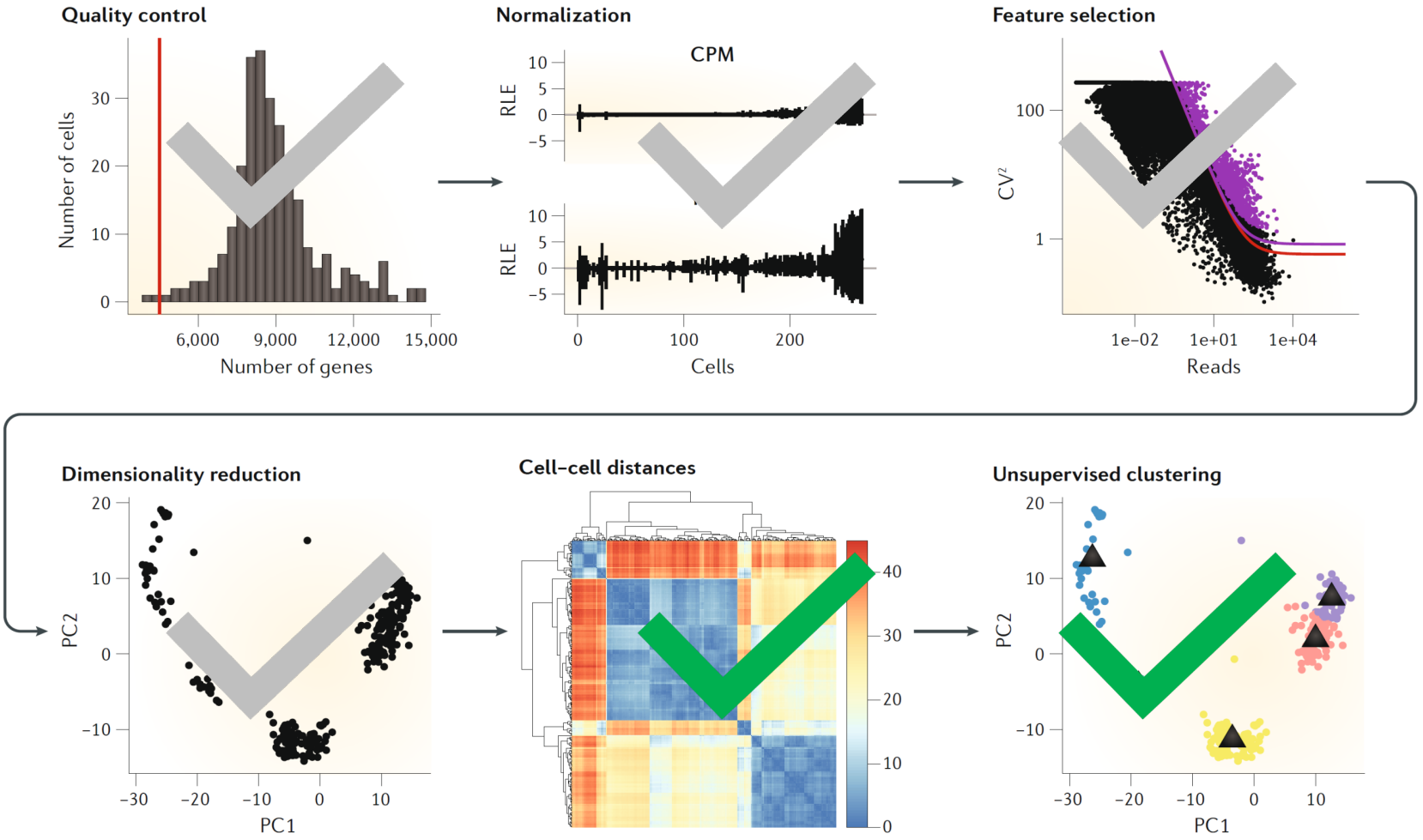


https://doi.org/10.1038/s41596-021-00534-0

Summary

- Start by identifying major well-known cell types (clearly defined, discrete cell clusters)
- Sometimes, it is useful to split the data into broad subsets (e.g., immune, endothelial and tumor) and analyze each separately
- Cell subtypes or poorly defined clusters are challenging
- Manual annotations heavily relies on marker genes and expert knowledge

Summary


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

Challenges

- **Subjectivity:** what is a cell type
  - *Different parameters will give different results*
  - *Validation is important*

- **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
  - *Computational efficiency*
  - *Visual exploration, crowding problem*

Clustering practical

- Hierarchical clustering: distances and linkage methods
- *k*-Means
- Graph-based clustering

- Annotating clusters

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

- Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods

https://doi.org/10.1038/s41596-021-00534-0