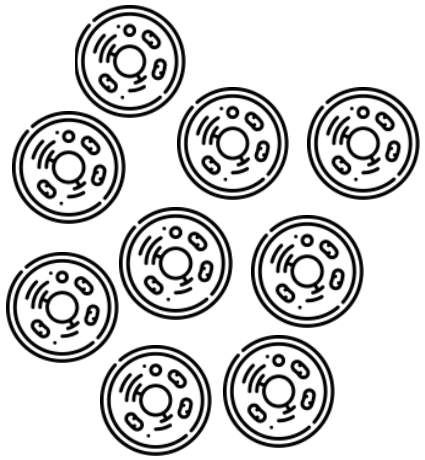
Clustering & cell annotation

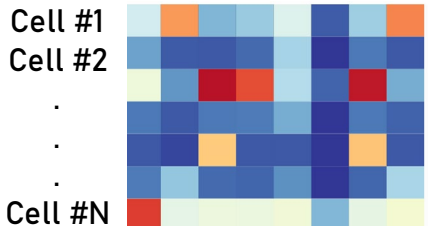
Lieke Michielsen
Department of Human Genetics, LUMC
Delft Bioinformatics Lab, TU Delft

How can we identify cell populations?

Mystery cells

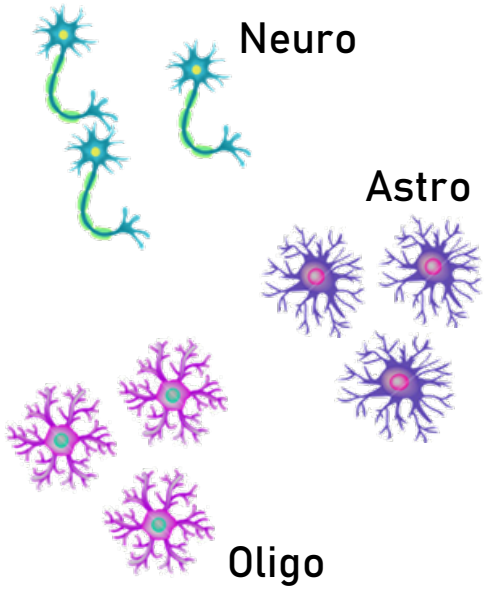


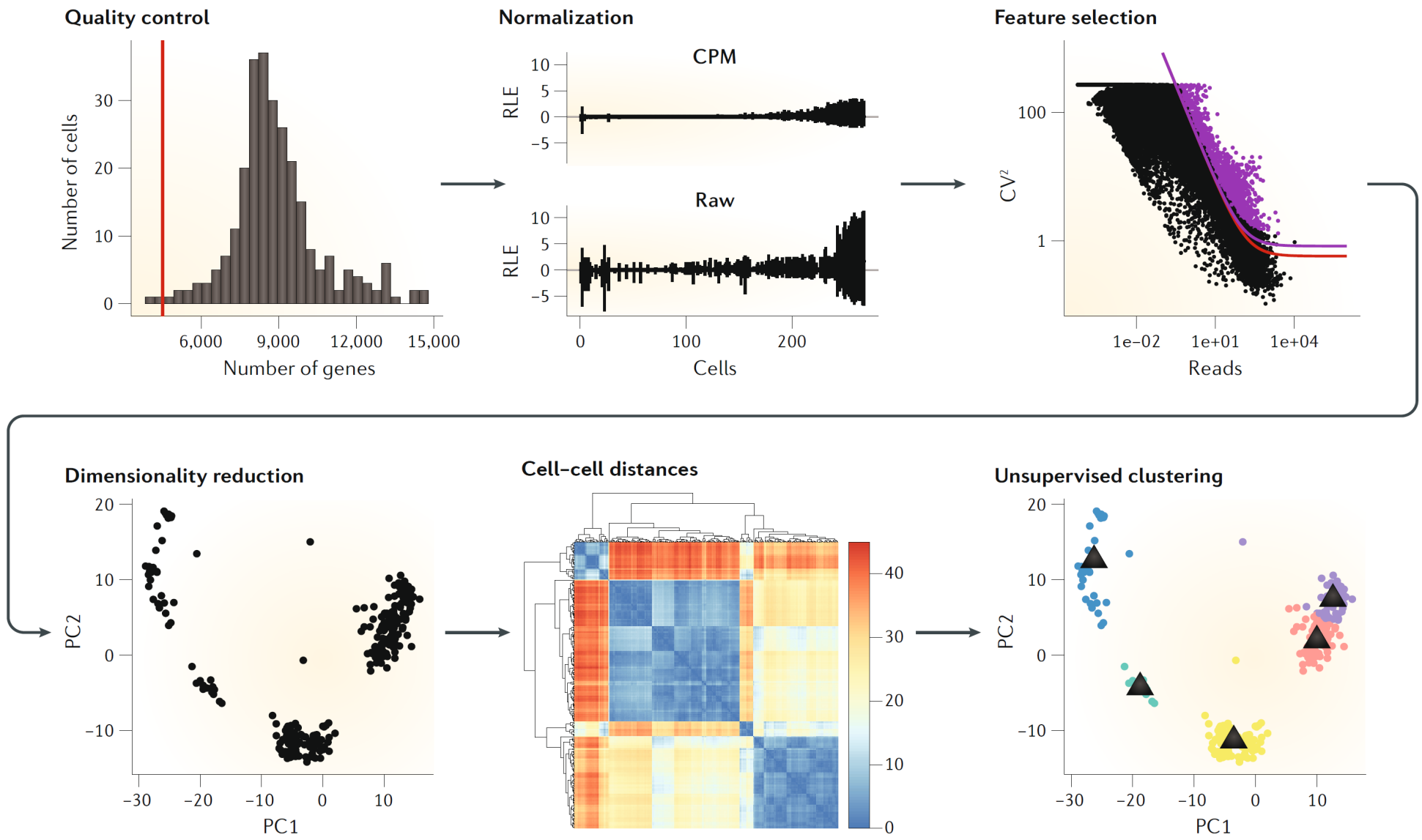
Measure



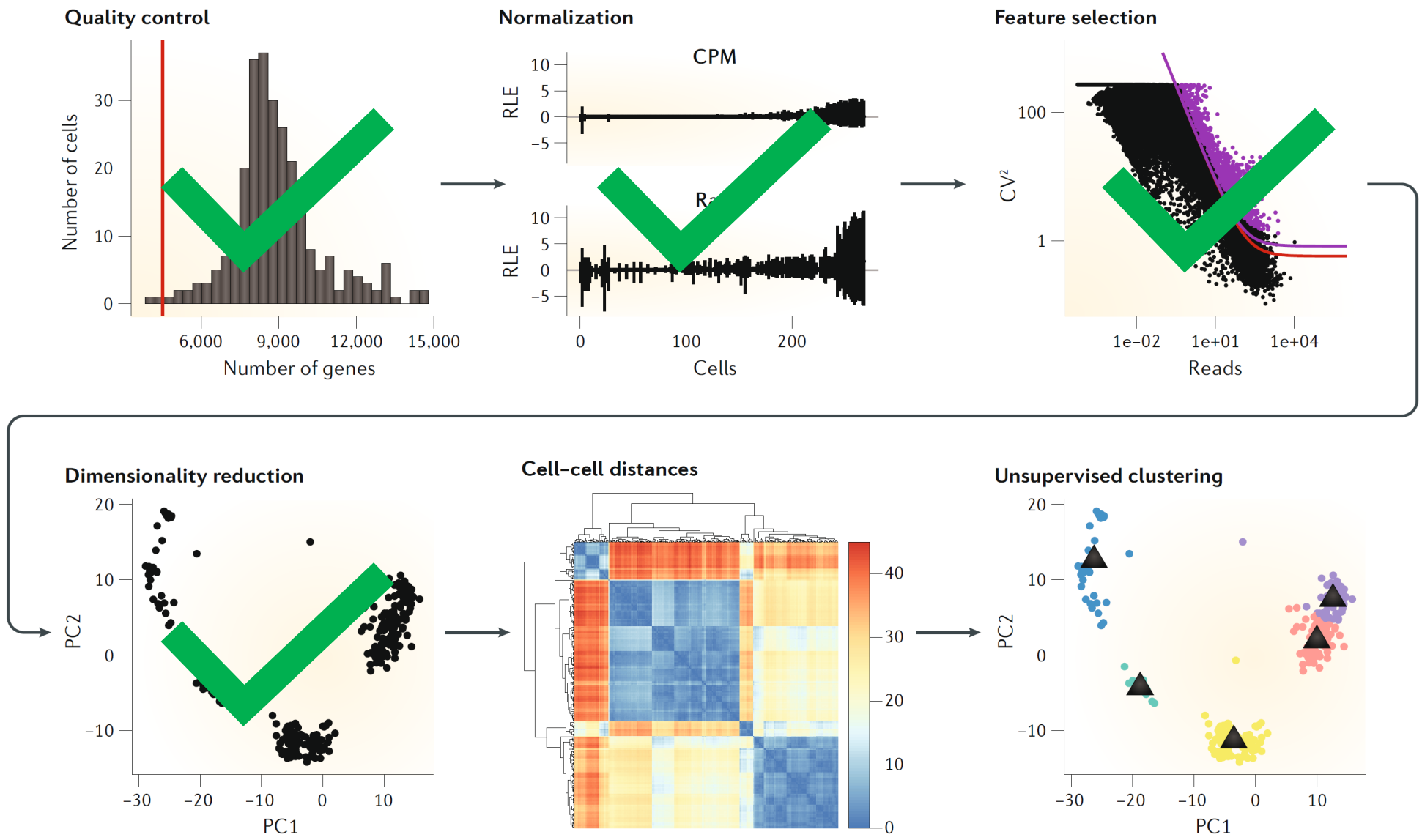
Identify

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

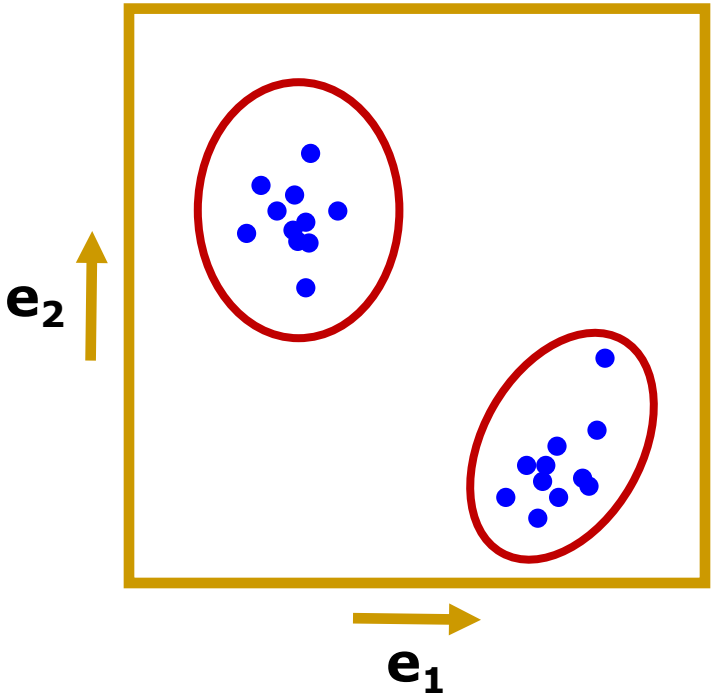
- Introduction to clustering
  - Hierarchical clustering
  - *k*-Means clustering
  - Graph-based clustering
- scRNA-seq clustering
- Annotating clusters
- Cluster validation
- Challenges & outlook

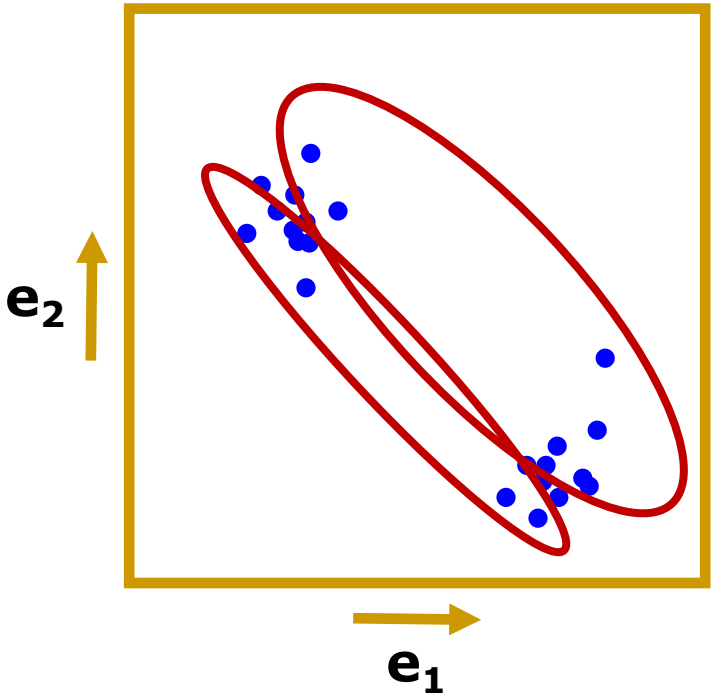
Outline

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

Clustering

- What defines a good clustering?

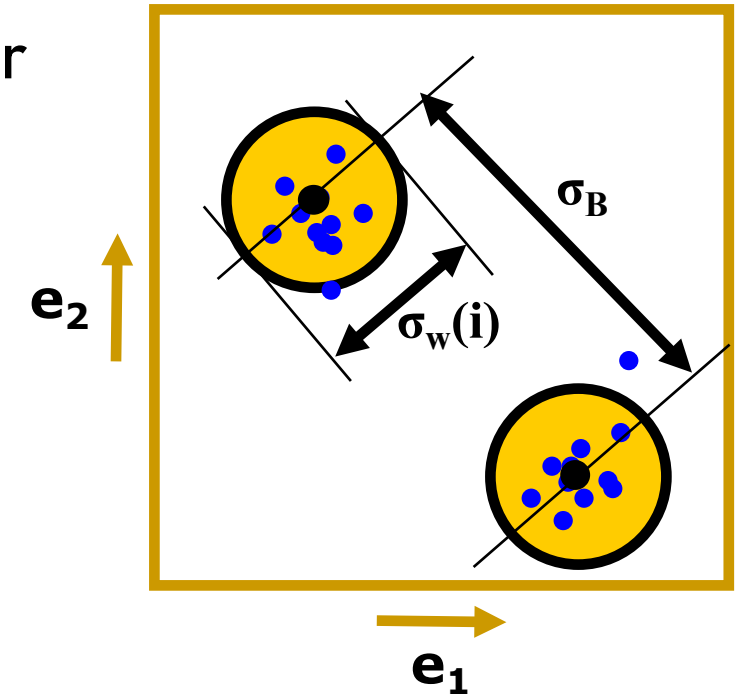




Clustering

Structure when:

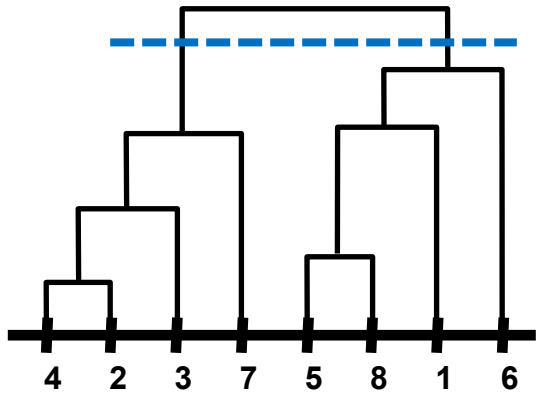
1. Samples within cluster resemble each other *(small within variance, $\sigma_W(i)$)*
2. Clusters deviate from each other *(large between variance, σ_B)*



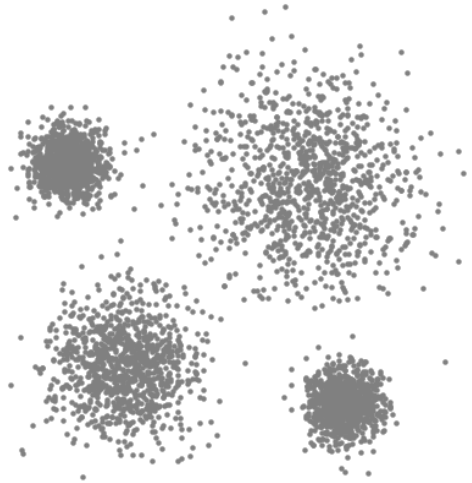
Group samples such that:

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

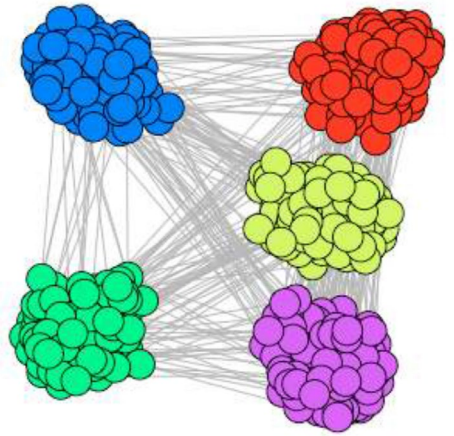
Many clustering approaches



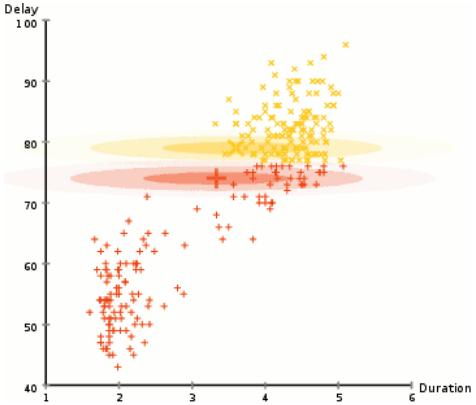
Hierarchical clustering



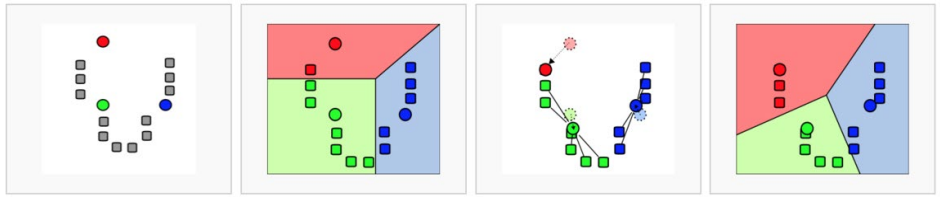
Mean shift clustering



Graph-based clustering



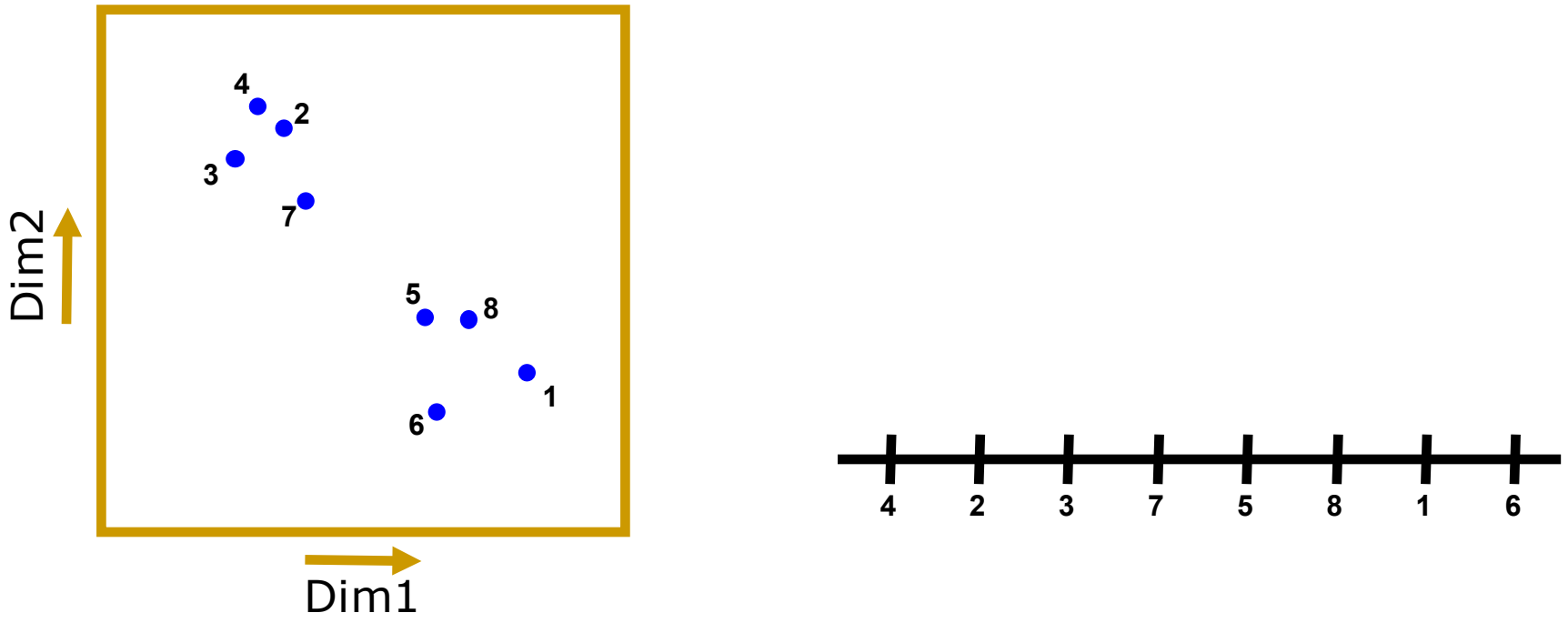
Gaussian mixture modeling



k-means clustering

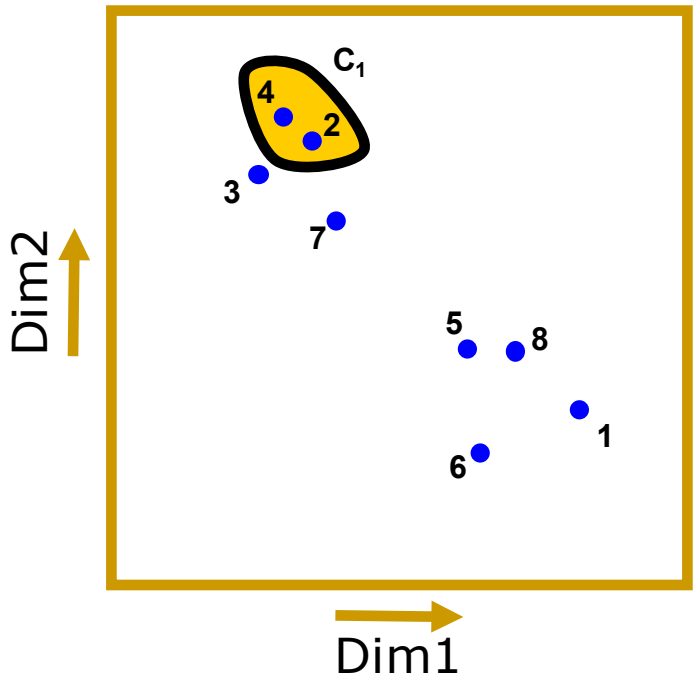
Herarchical clustering

Herarchical clustering

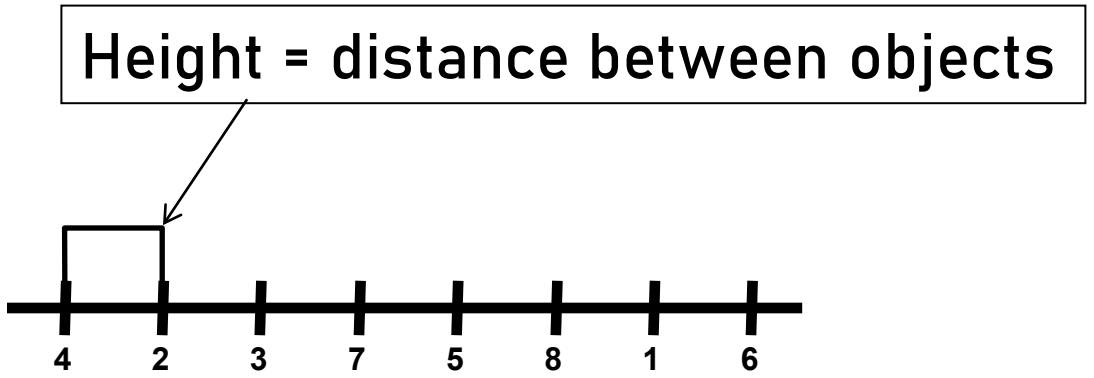


Find most similar objects (cells) and group them

Herarchical clustering

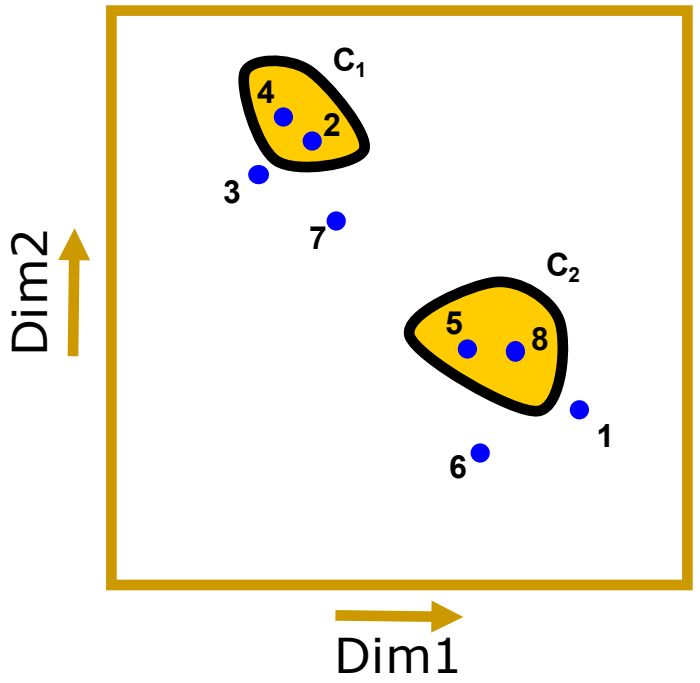


Dendrogram

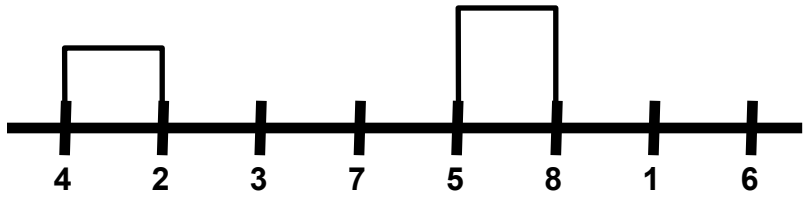


These are: objects 4 and 2
Again, find most similar objects (cells or clusters) and group them

Herarchical clustering

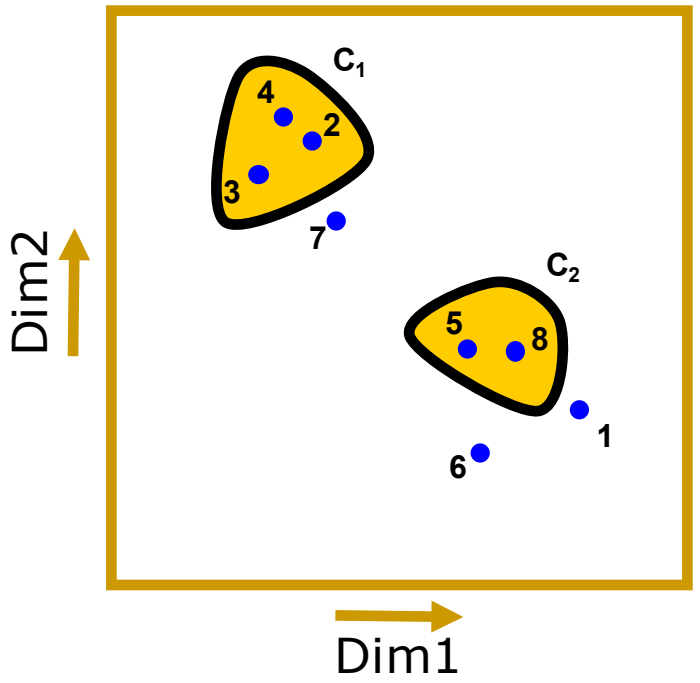


Dendrogram

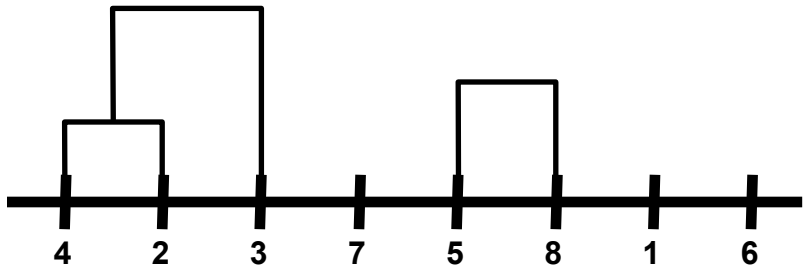


These are: objects 5 and 8
Repeat finding most similar objects (cells or clusters) and group them

Herarchical clustering

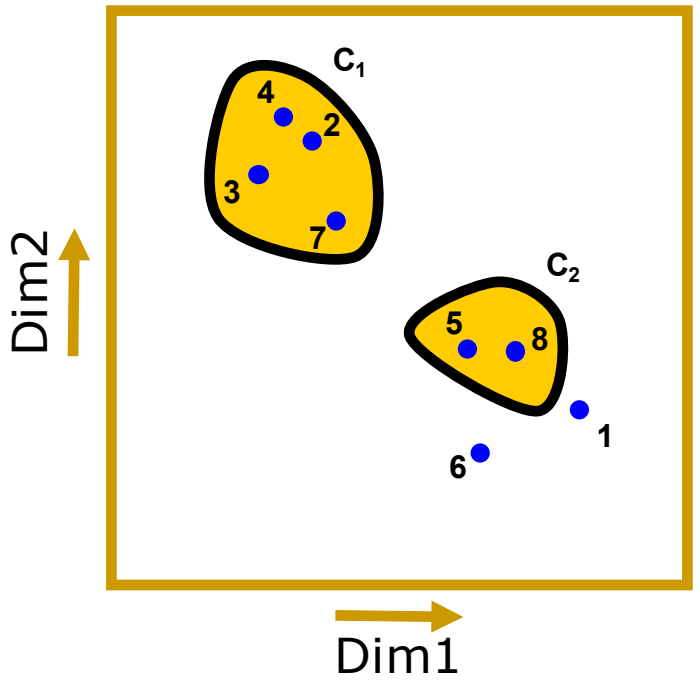


Dendrogram

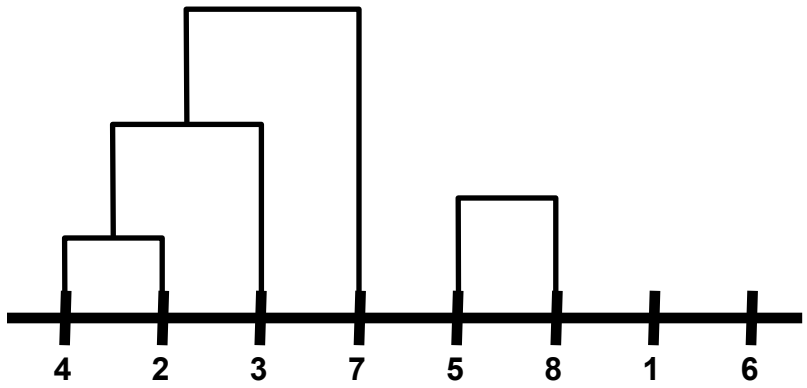


Join object 3 and cluster 1
Repeat process

Herarchical clustering

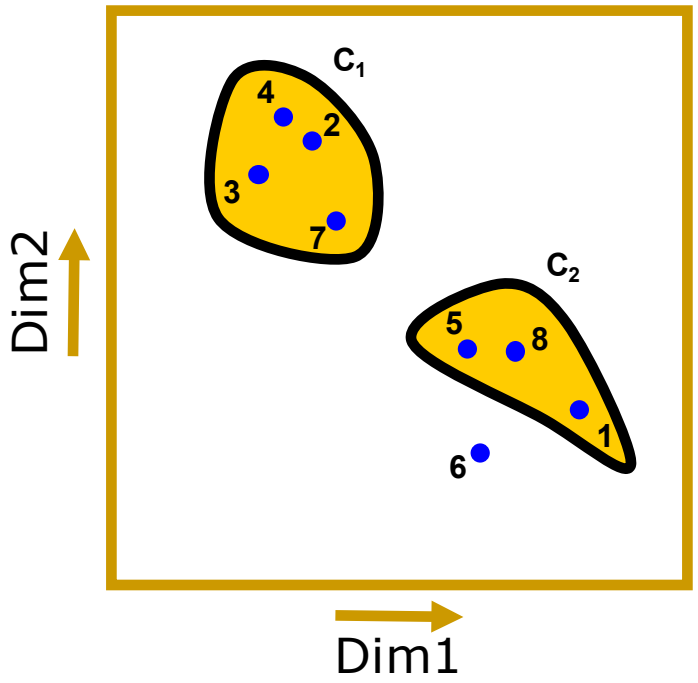


Dendrogram

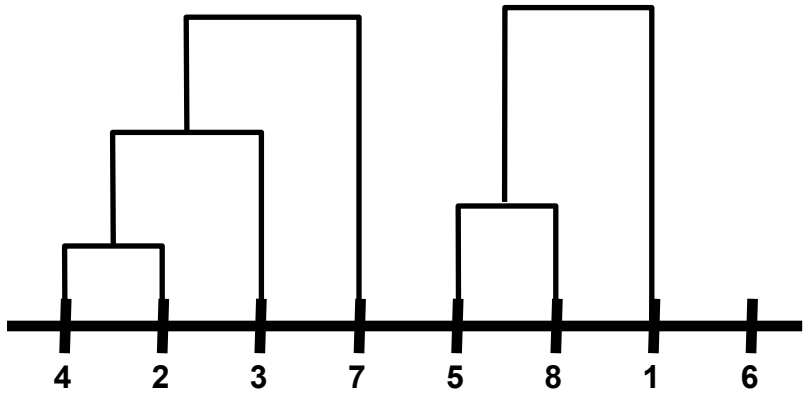


Join object 7 and cluster 1
Repeat process

Herarchical clustering

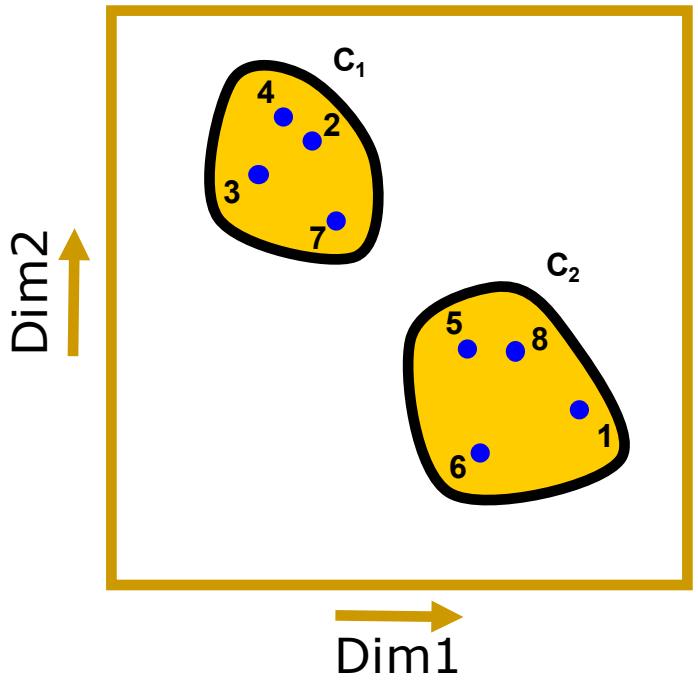


Dendrogram

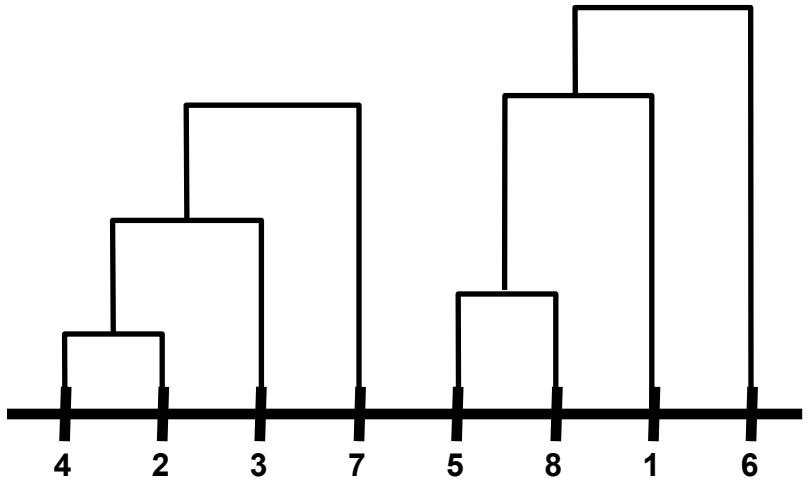


Join object 1 and cluster 2
Repeat process

Herarchical clustering

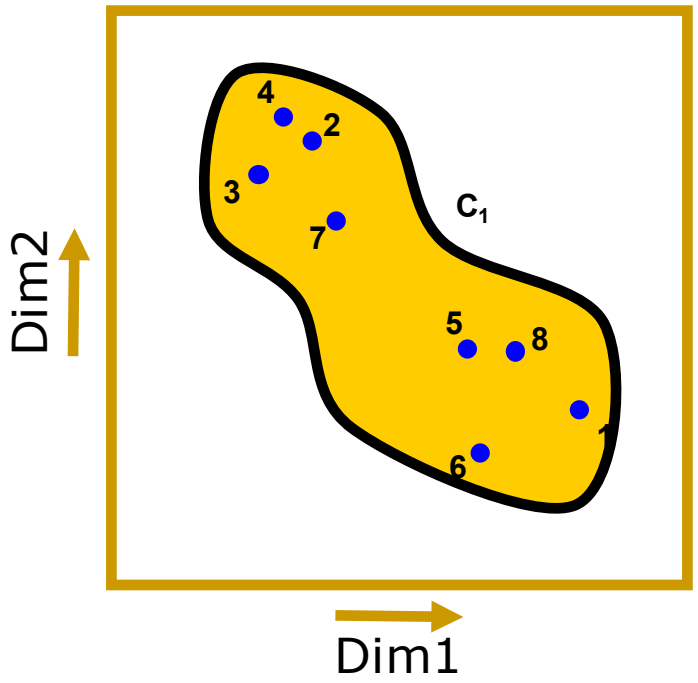


Dendrogram

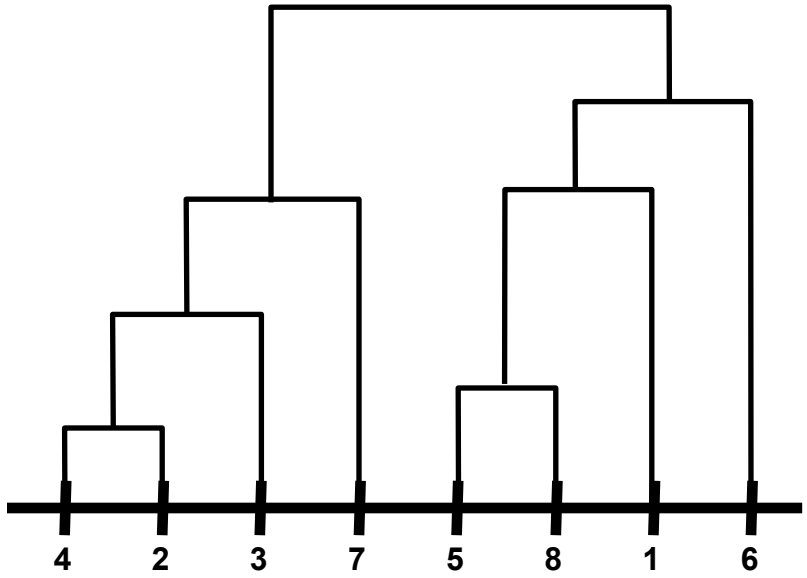


Join object 6 and cluster 2
Repeat process

Herarchical clustering

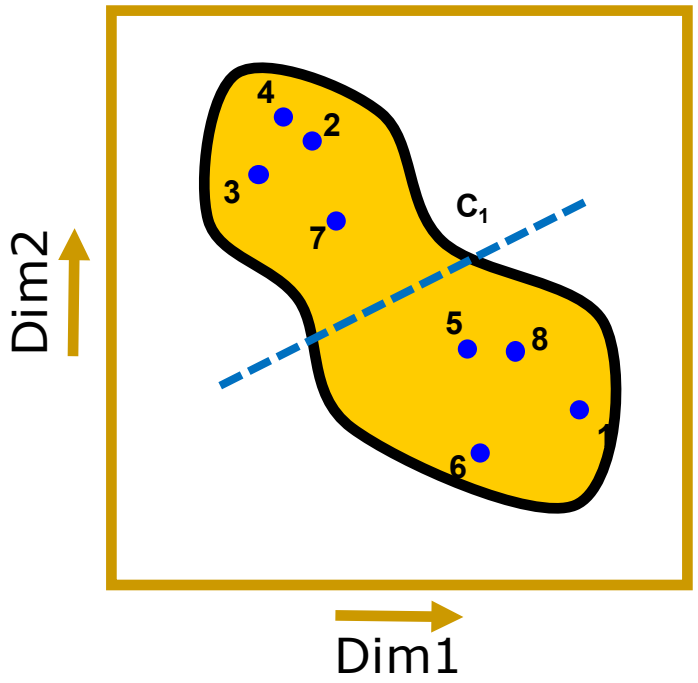


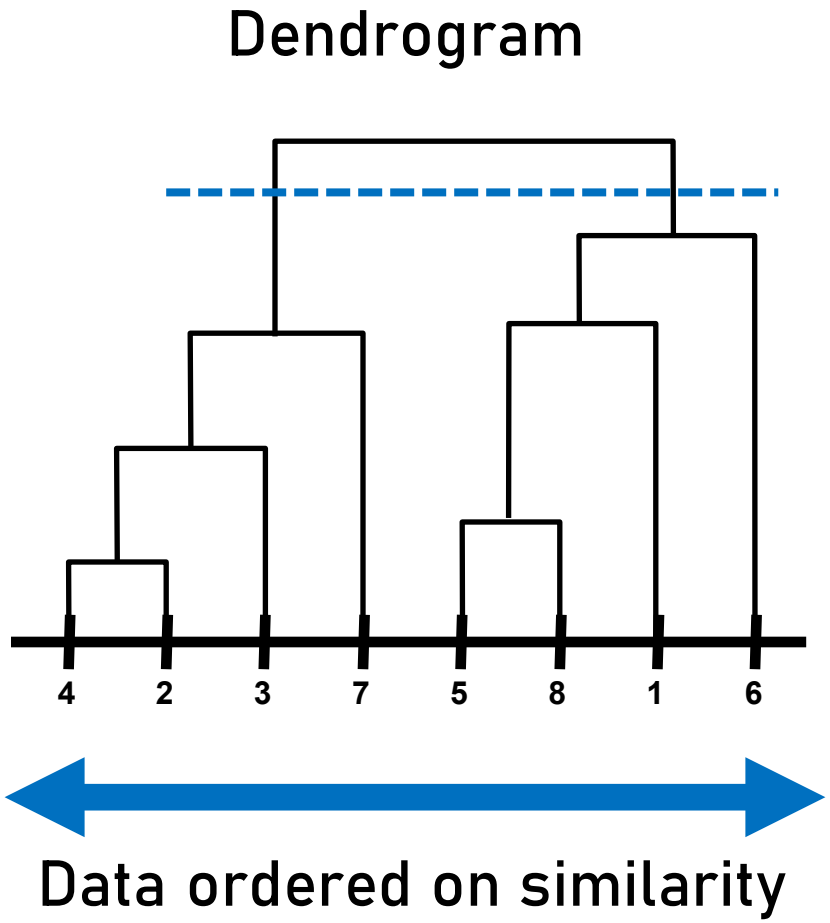
Dendrogram



Join cluster 1 and cluster 2
All in one cluster: FINISHED!

Herarchical clustering

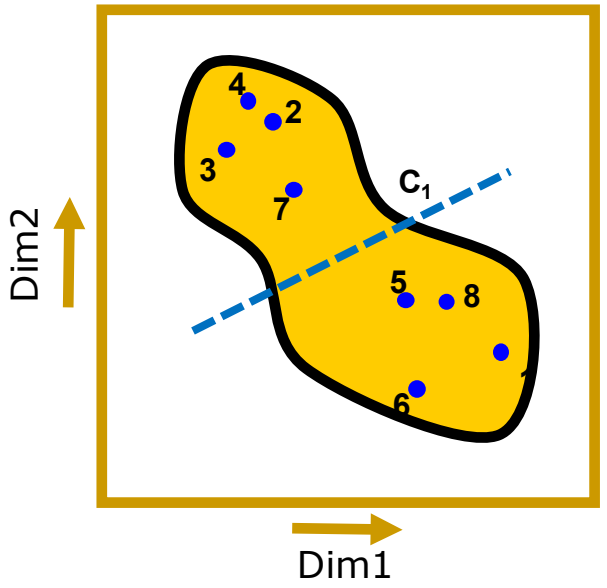


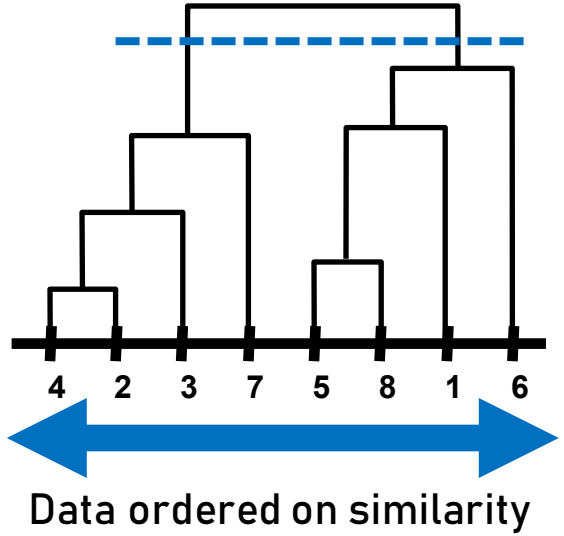


Herarchical clustering

Need to know:

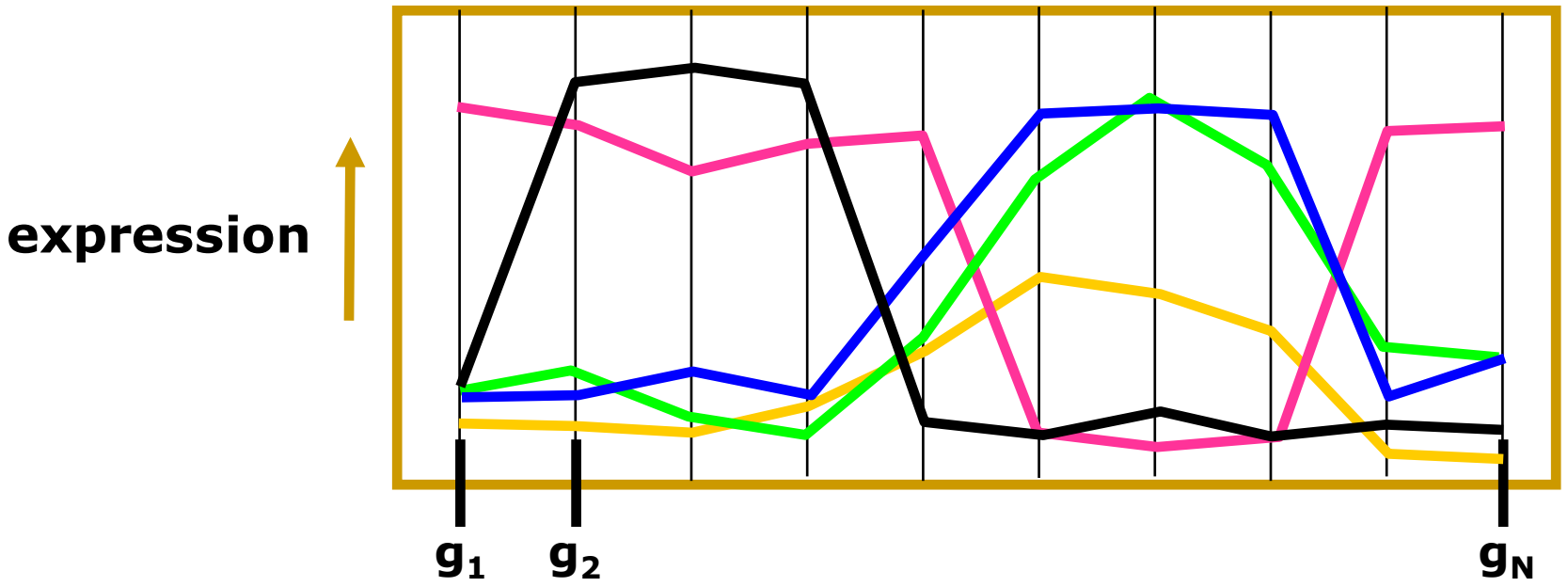
- Similarity between objects
- Similarity between clusters





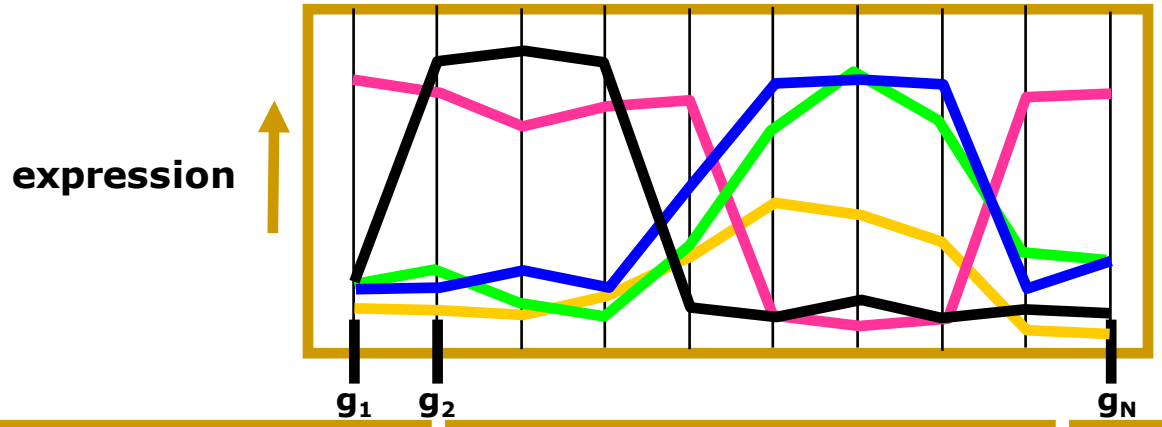
Herarchical clustering

Similarity between objects



Herarchical clustering

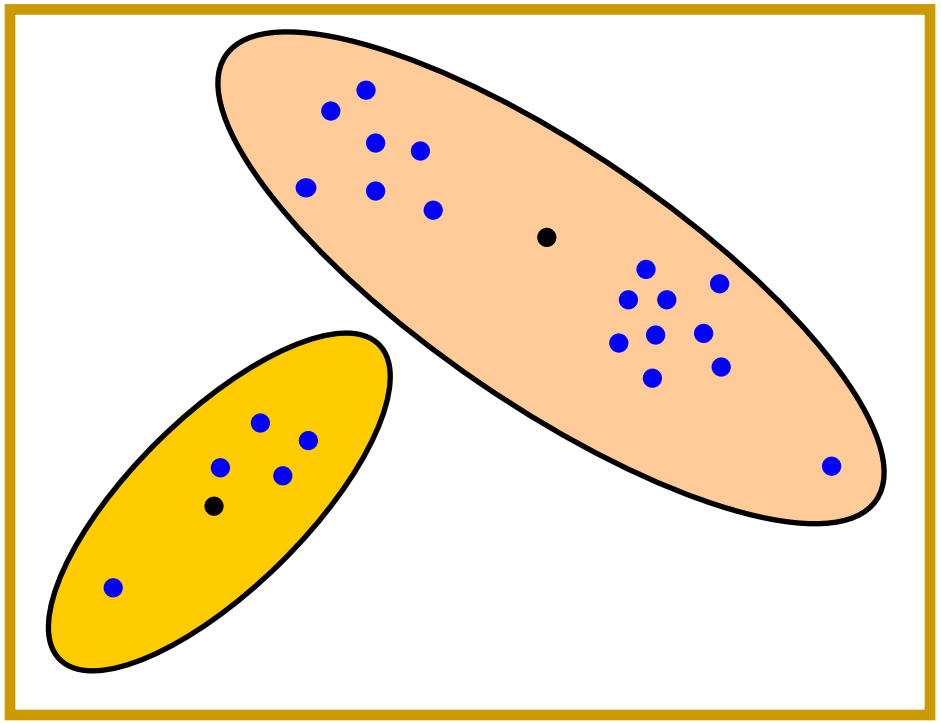
Similarity between objects



Euclidean distance	Pearson correlation	Mixed Pearson correlation
$\sqrt{(\sum((x_i - x_j)^2))}$	$1 - \rho_{ij}$	$1 - \lvert\rho_{ij}\rvert$
d(●,●) < d(●,●)	d(●,●) ≈ d(●,●)	d(●,●) ≈ d(●,●)
d(●,●) << d(●,●)	d(●,●) << d(●,●)	d(●,●) ≈ d(●,●)
d(●,●) << d(●,●)	d(●,●) << d(●,●)	d(●,●) << d(●,●)
Match exact shape	Ignore amplitude	Ignore amplitude and sign

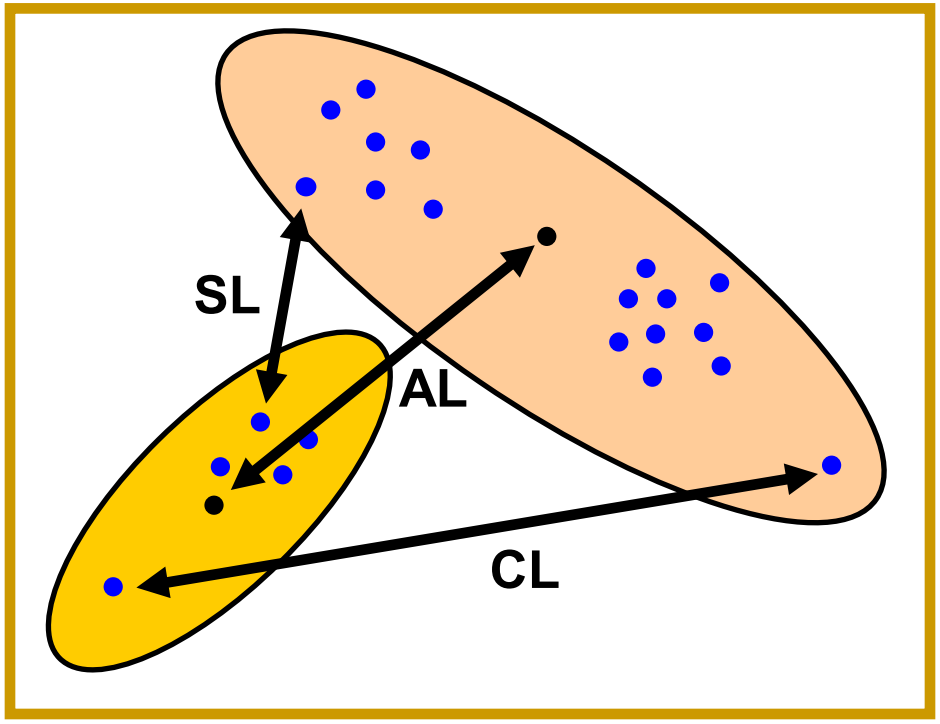
Herarchical clustering

Similarity between clusters



Herarchical clustering

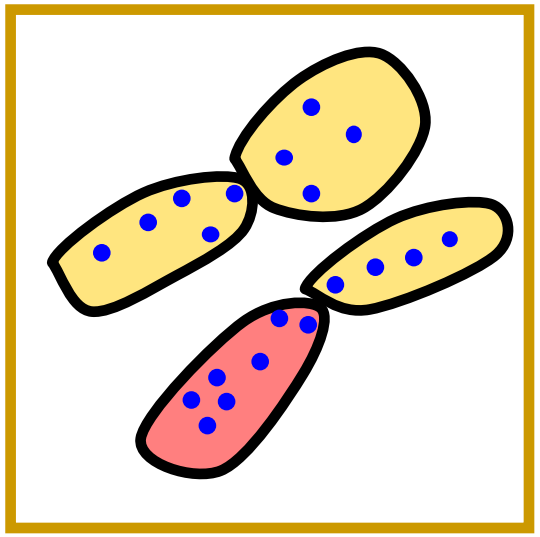
Similarity between clusters



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

Herarchical clustering

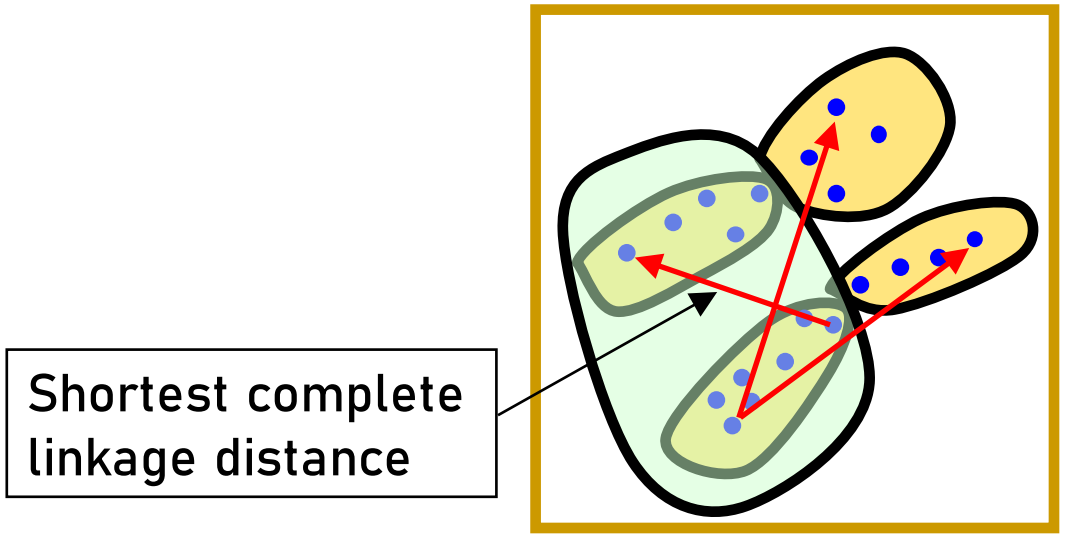
Similarity between clusters



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

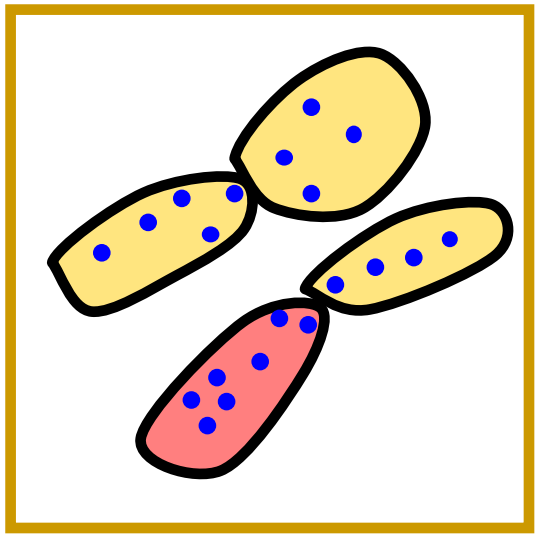
Herarchical clustering

Similarity between clusters



Herarchical clustering

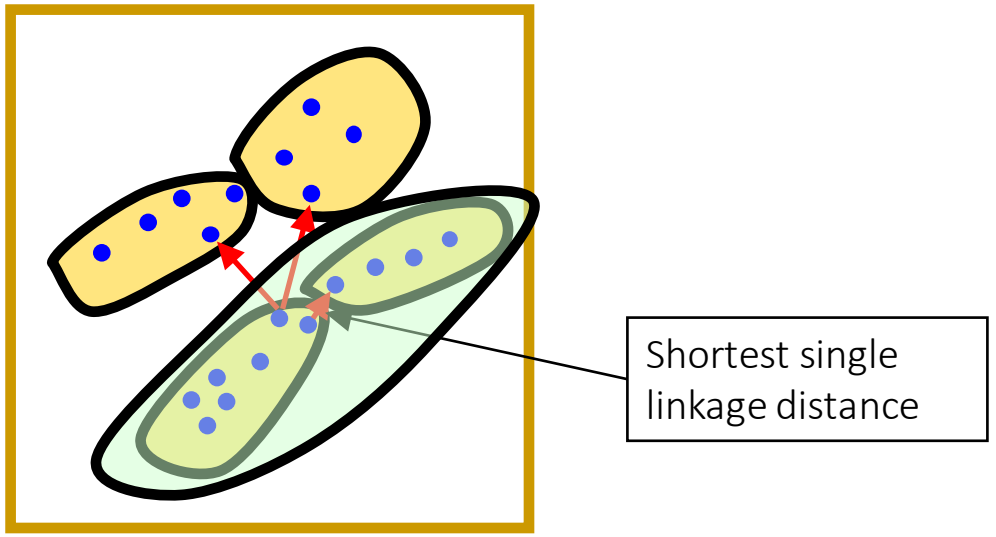
Similarity between clusters



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

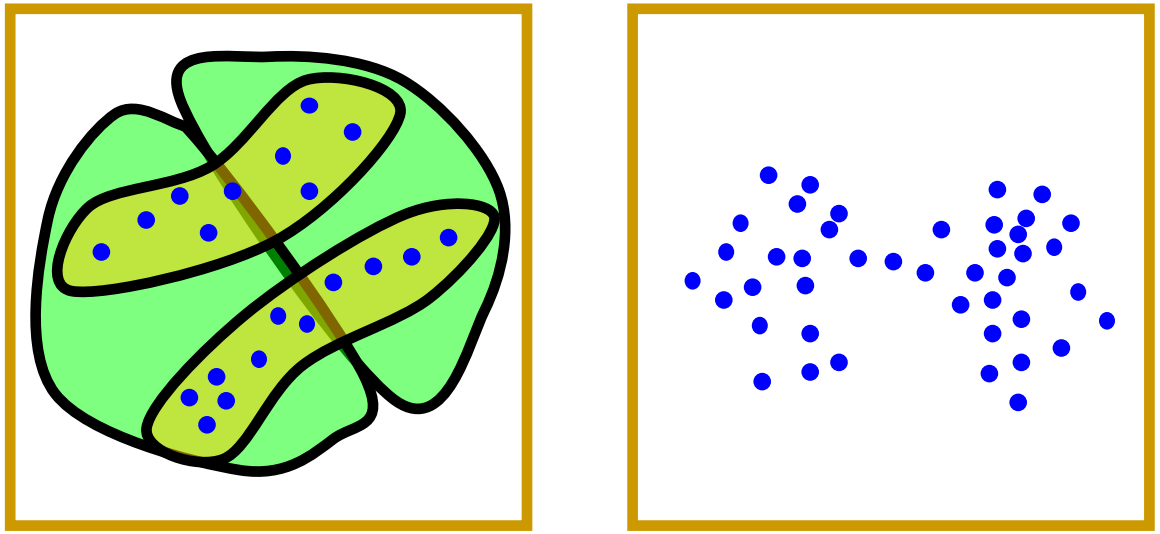
Herarchical clustering

Similarity between clusters



Herarchical clustering

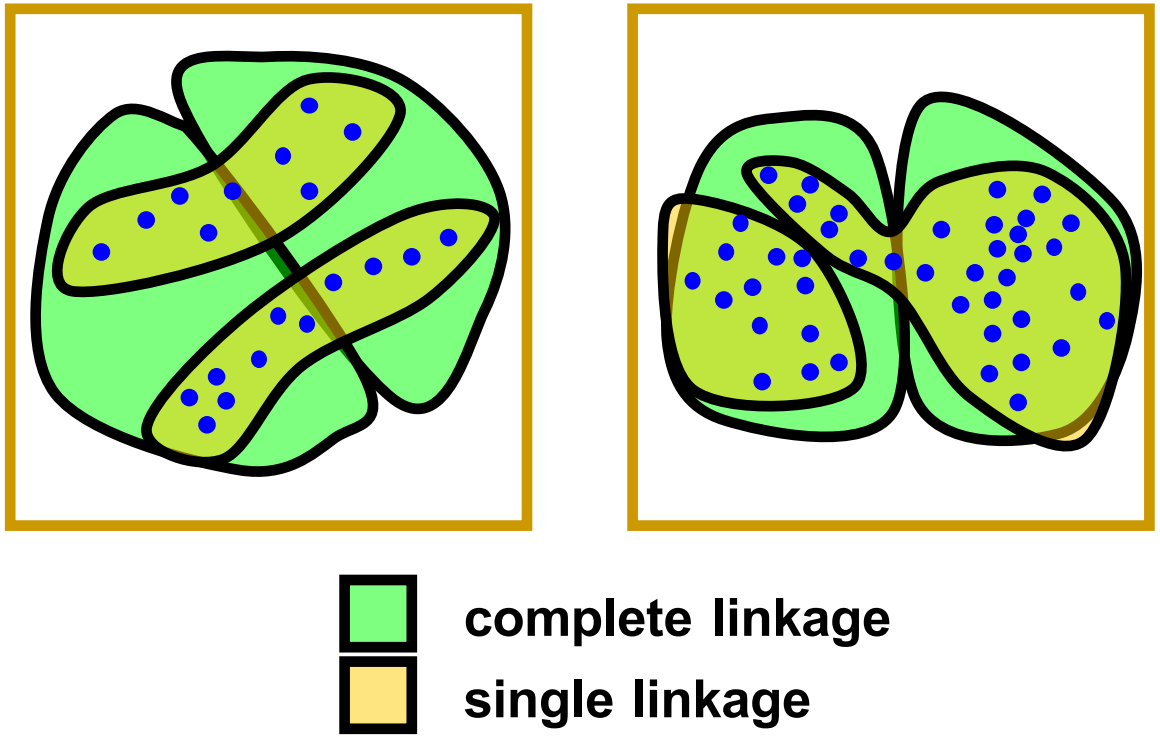
Similarity between clusters



Herarchical clustering

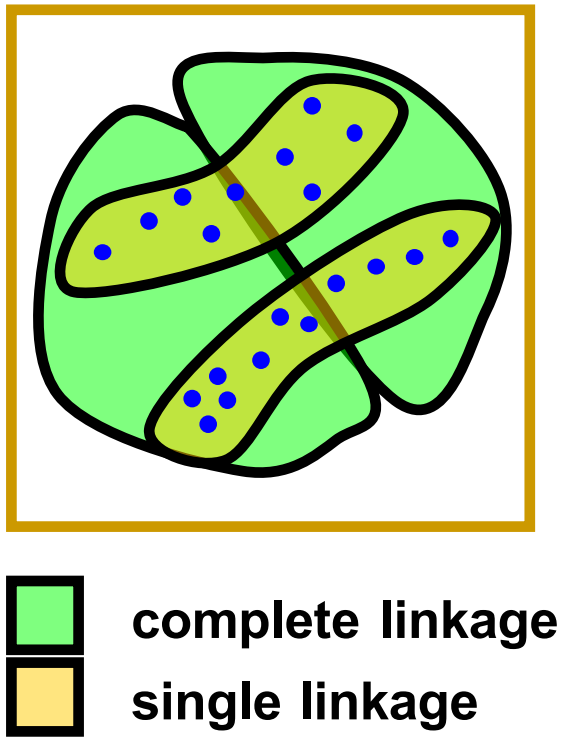
Similarity between clusters

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

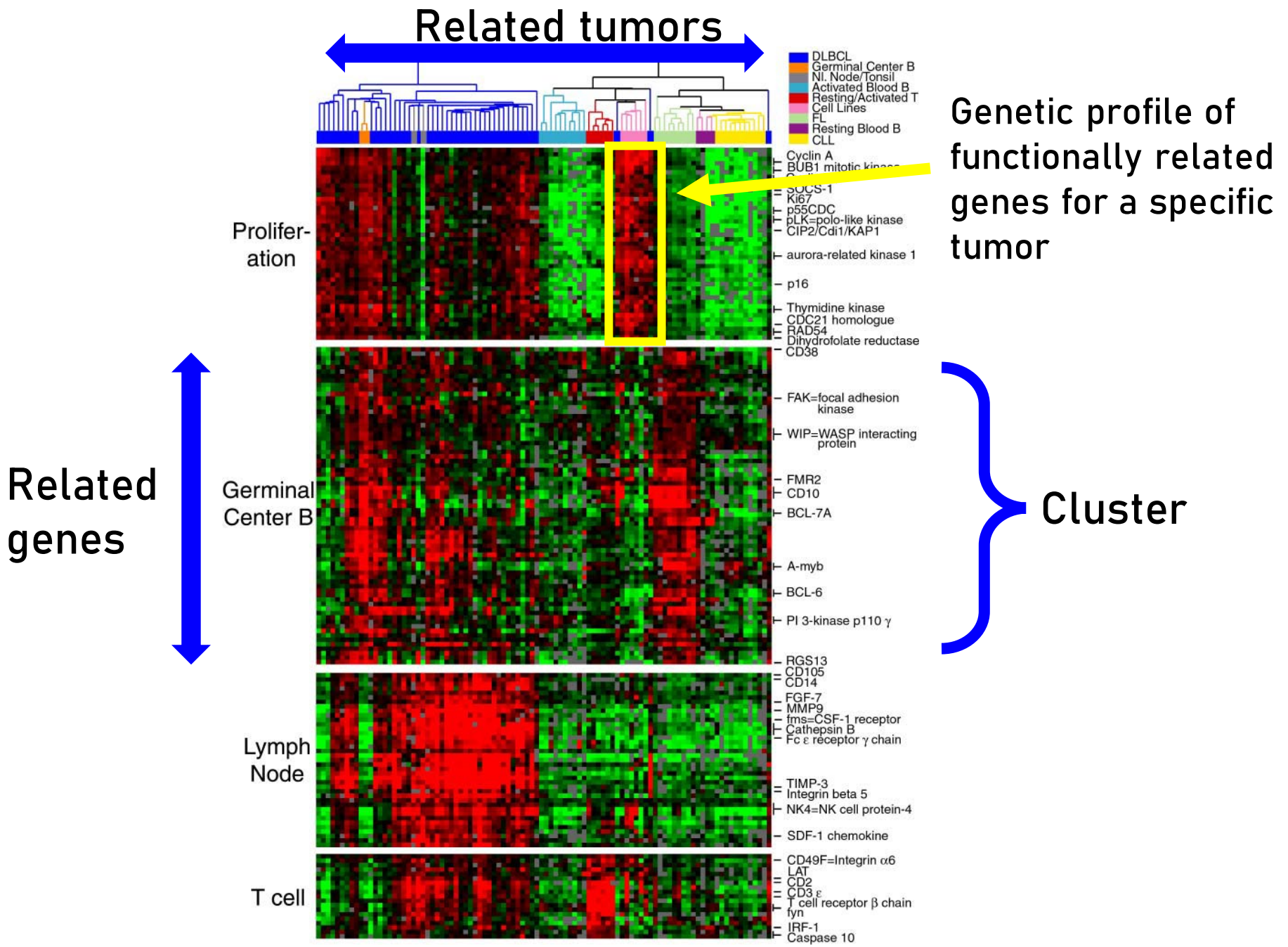


Herarchical clustering

Overview

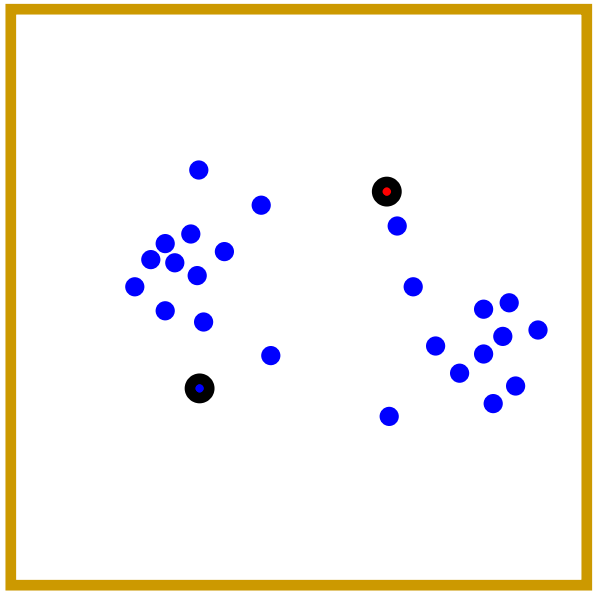


- Hierarchical clustering
  - Choice of distance measure
  - Choice of linkage type
- Distance measure
  - Euclidean
  - Correlation
- Linkage
  - Single
  - Average
  - Complete
- Number of clusters
  - Predefined or based on a cut-off in the dendrogram
  - Validate clustering!


Related tumors
DLBCL
Germinal Center B
Nl. Node/Tonsil
Activated Blood B
Resting/Activated T
Cell Lines
FL
Resting Blood B
CLL
Genetic profile of functionally related genes for a specific tumor
Prolifer-ation
Cyclin A
BUB1 mitotic kinase
SOCS-1
Ki67
p55CDC
pLK=polo-like kinase
CIP2/Cdi1/KAP1
aurora-related kinase 1
p16
Thymidine kinase
CDC21 homologue
RAD54
Dihydrofolate reductase
CD38
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
SDF-1 chemokine
CD49F=Integrin α6
LAT
CD2
CD3 ε
T cell receptor β chain
fyn
IRF-1
Caspase 10
T cell

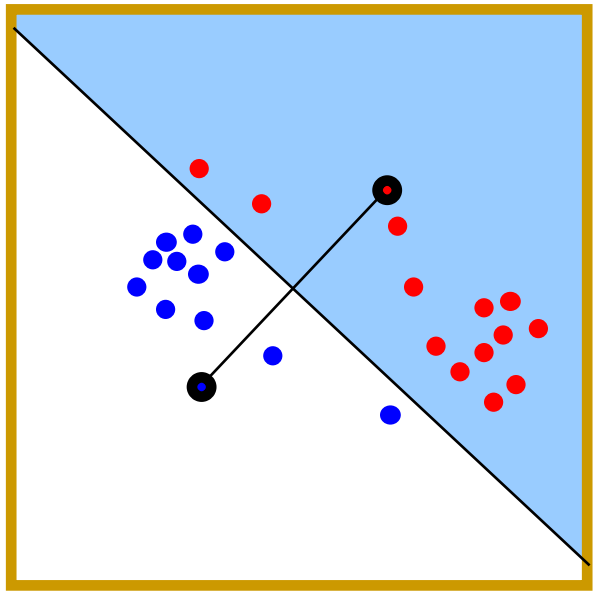
k-Means clustering

k-Means clustering



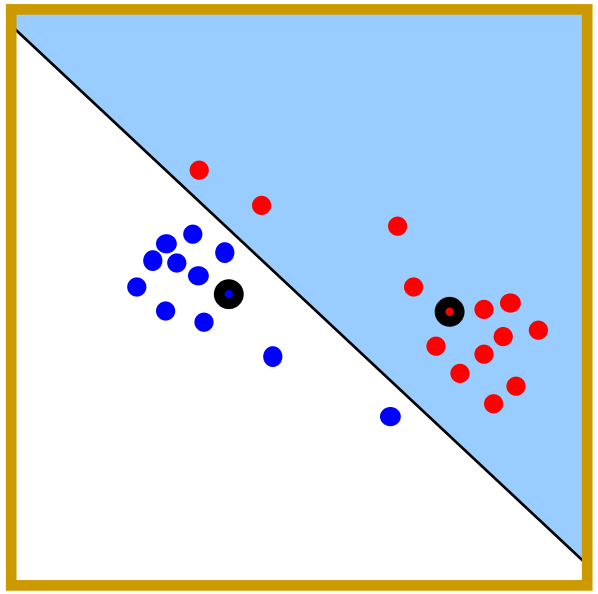
Choose randomly *k* prototypes

k-Means clustering



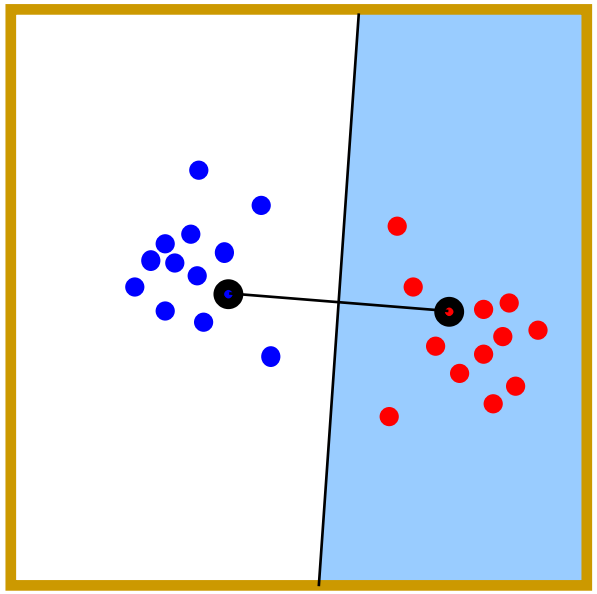
Assign objects to the closest prototype
Blue area: cluster 1
White area: cluster 2

k-Means clustering



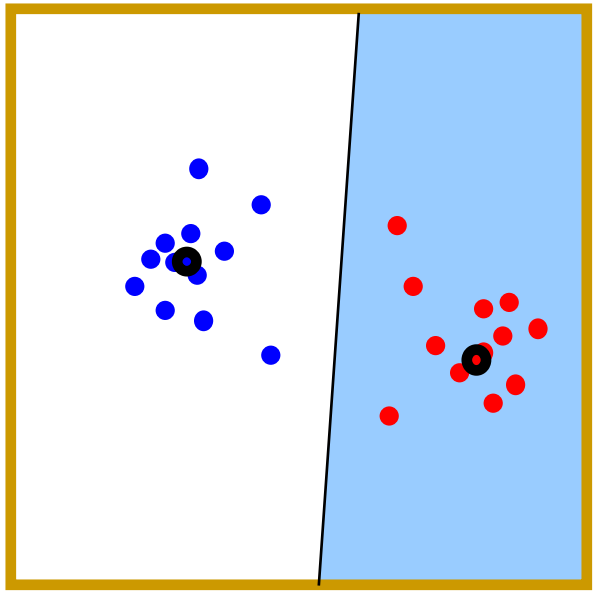
Calculate new cluster prototypes
By averaging objects

k-Means clustering



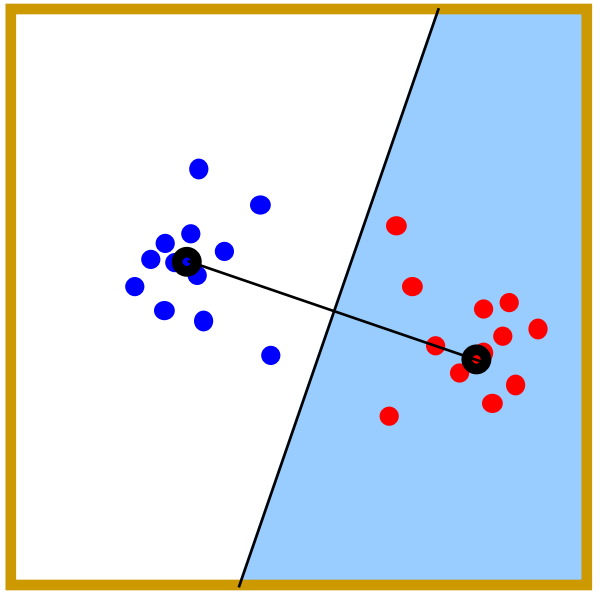
Re-assign objects to the closest prototype
Blue area: cluster 1
White area: cluster 2

k-Means clustering



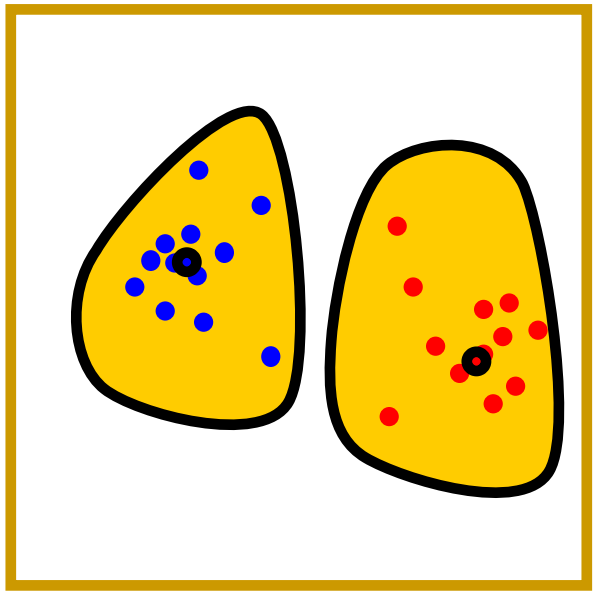
Re-calculate new cluster prototypes

k-Means clustering



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

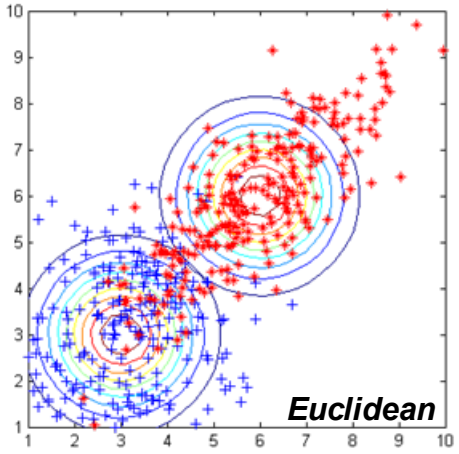
k-Means clustering

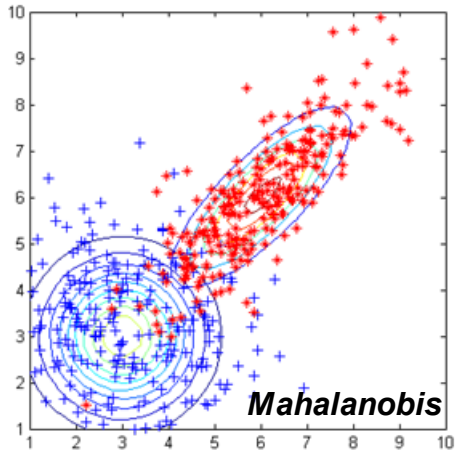


Establish clusters

k-Means clustering

Overview



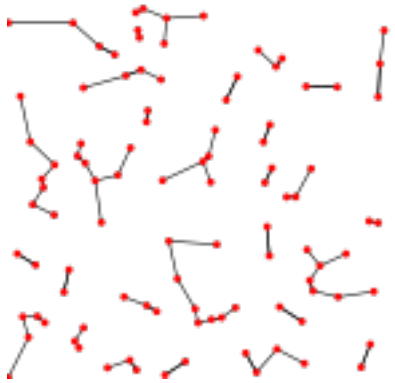


- *k*-Means
  - Fixed number of clusters (need to know a priori)
  - Choice of distance measure
  - Prototype choice
- Distance measure
  - Euclidean: Round clusters
  - Mahalanobis: Elongated clusters
- Prototype choice
  - Point
  - Line etc.
- Number of clusters
  - Predefined by *k*
  - Validate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

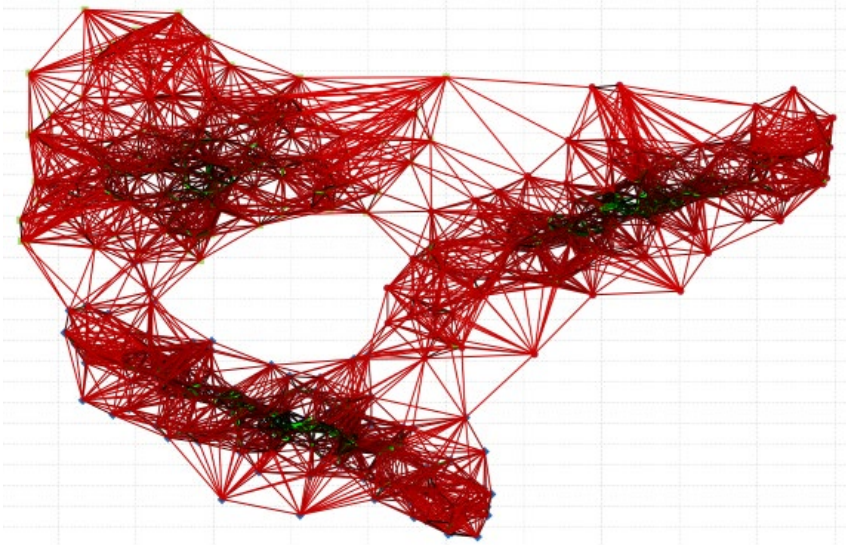
Graph-based clustering

Graph-based clustering

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



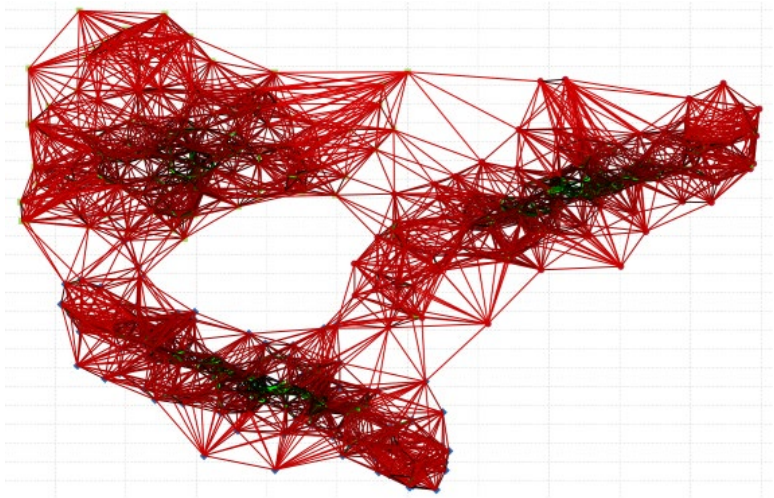
k=1



k=20

Graph-based clustering

- Maximize modularity score
  - Dense connections between nodes within clusters
  - Sparse connections between nodes in different clusters



Observed edges in cluster c

Expected edges in cluster c

$$H = \frac{1}{2m}\sum_c (e_c - \gamma \frac{K_c^2}{2m})$$

$m = edges\ in\ the\ graph$

$e_c = edges\ in\ cluster\ c$

$K_c = \sum_{n \in c} degree(n)$

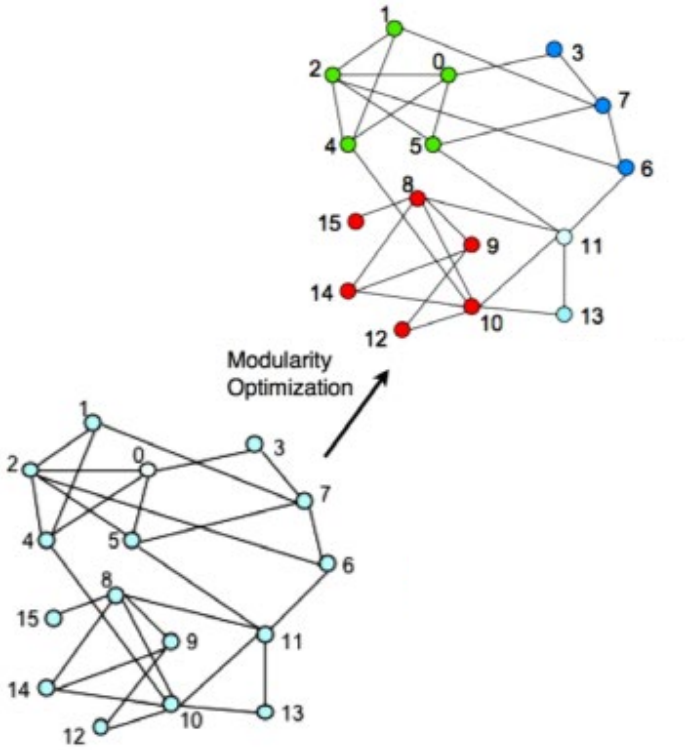
$\gamma = resolution$

Graph-based clustering

Louvain algorithm

Two steps

1. Local moving of nodes: move node i to community of neighbor j , if this increases H
2. Aggregate nodes



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

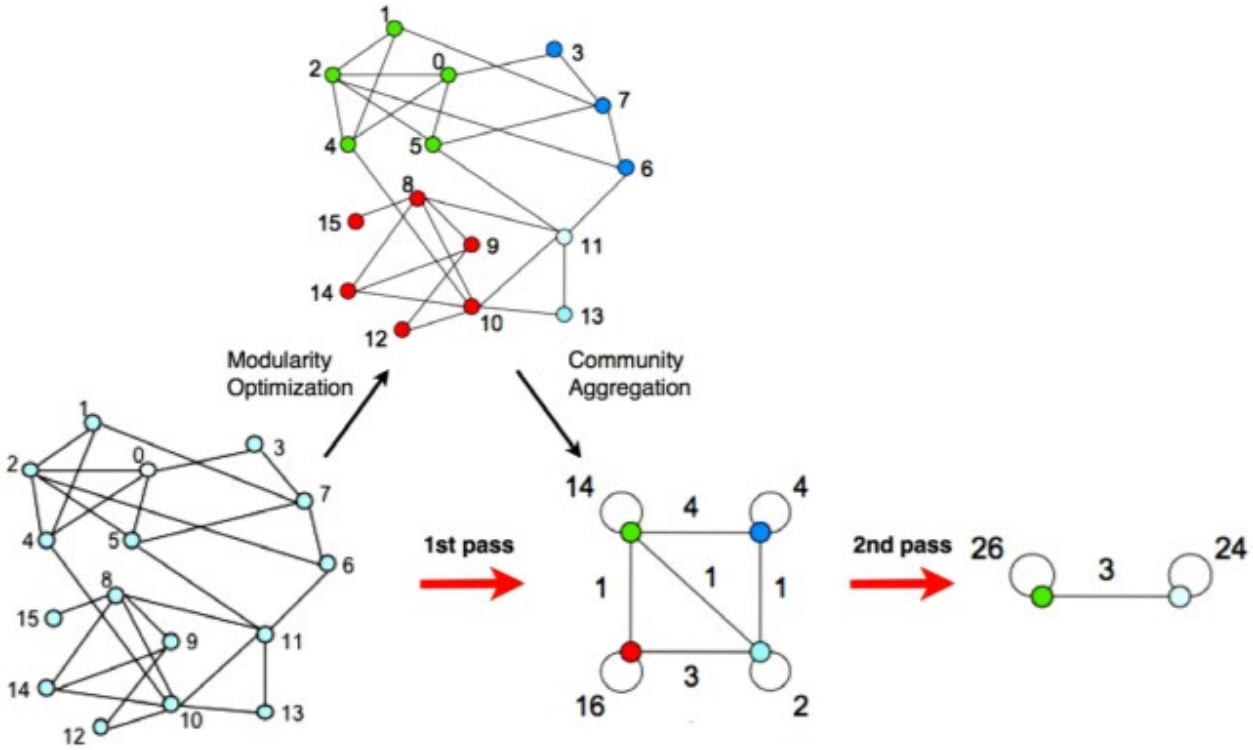
Graph-based clustering

Louvain algorithm

Two steps

1. Local moving of nodes: move node i to community of neighbor j , if this increases H
2. Aggregate nodes

Iterate until no more changes

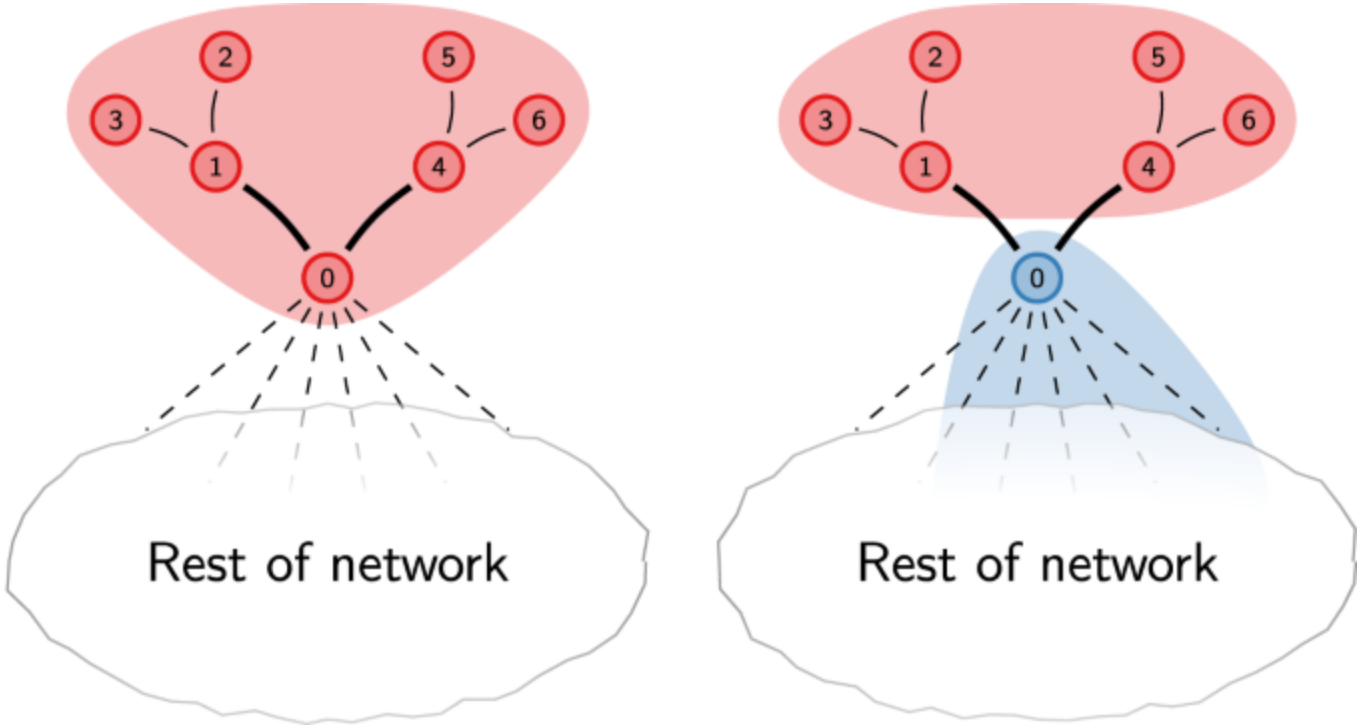


Blondel et al. (https://doi.org/10.1088/1742-5468/2008/10/P10008)

Graph-based clustering

Louvain algorithm

- During the ‘moving step’, nodes can become internally disconnected
- Nodes 1-6 still locally optimal assigned

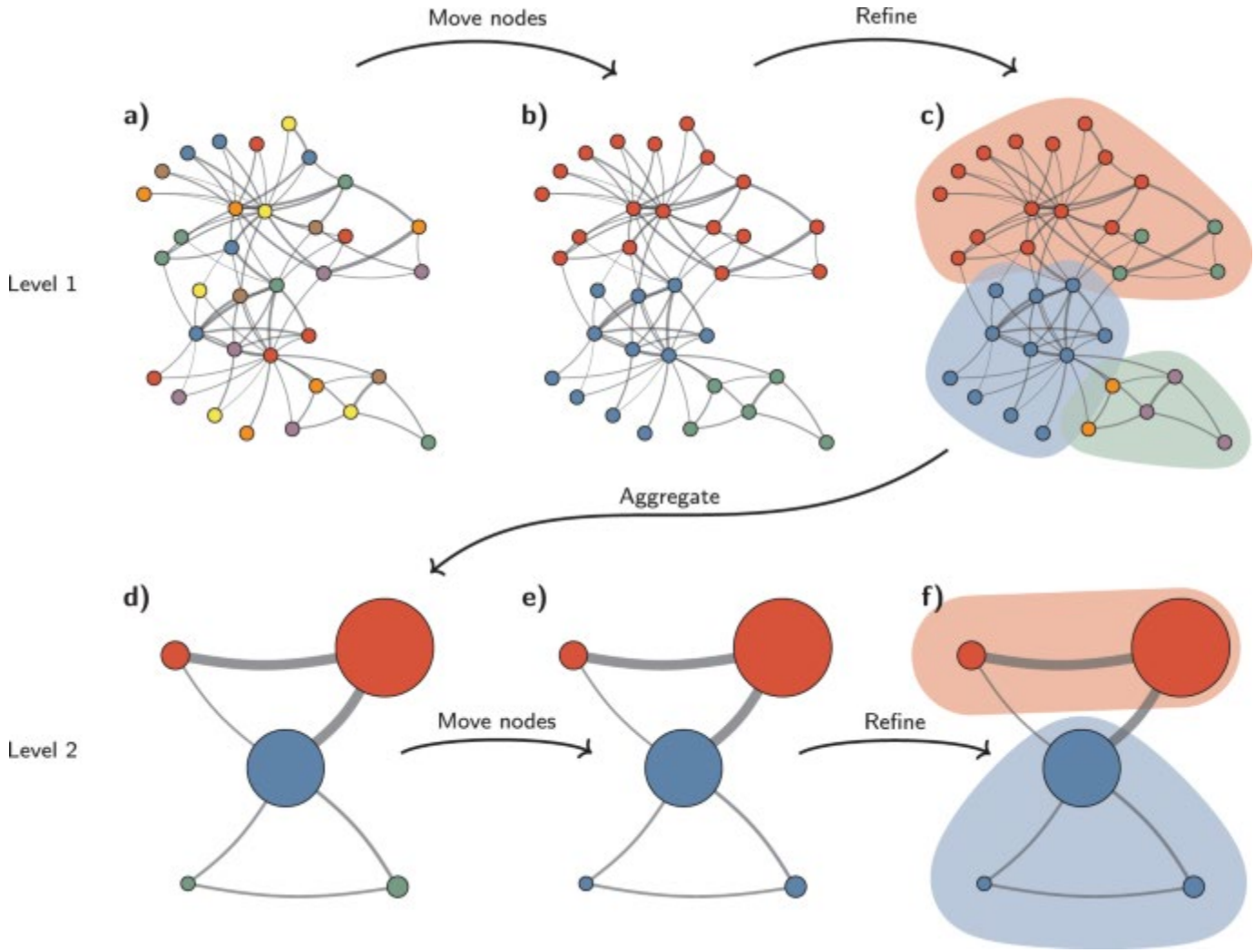


Traag et al. (https://doi.org/10.1038/s41598-019-41695-z)

Graph-based clustering

Leiden algorithm

- **Solution: add refinement step**

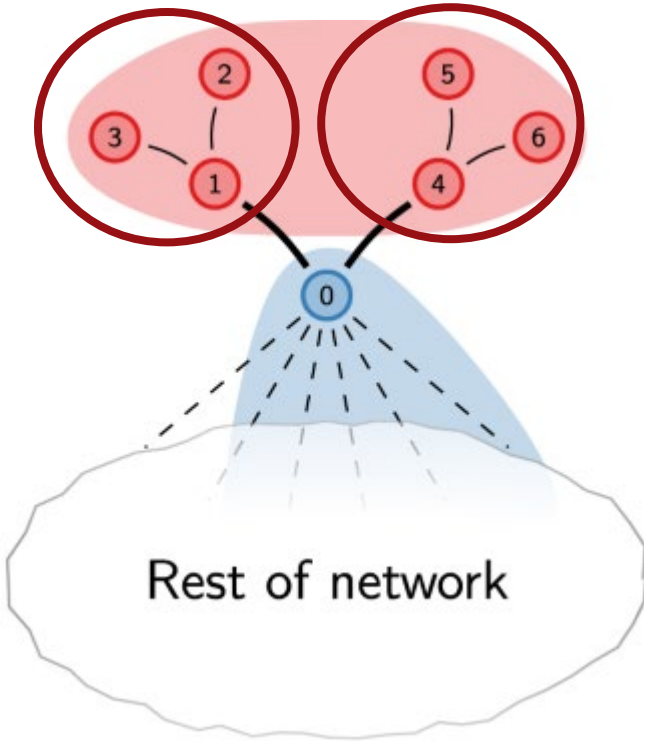


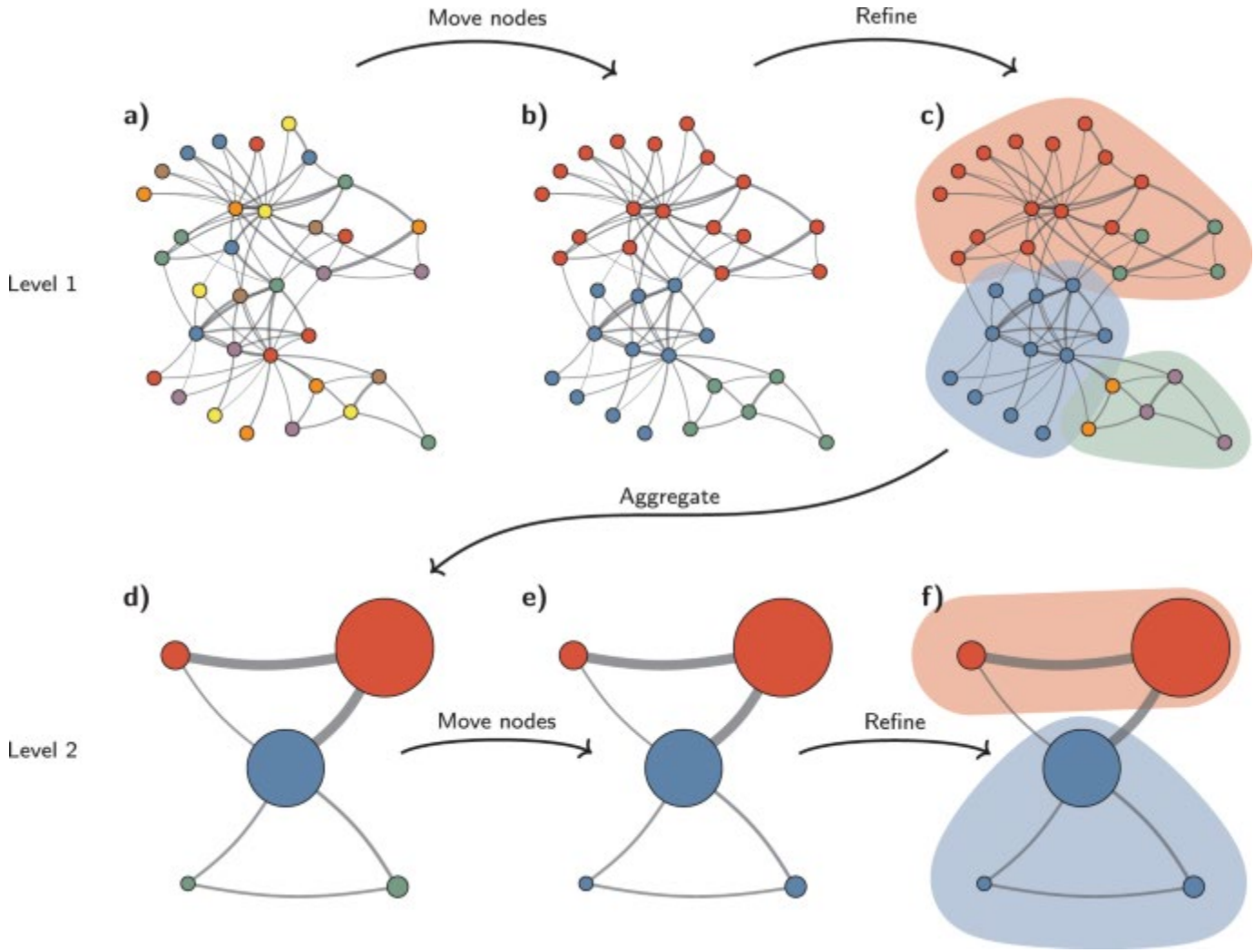
Traag et al. (https://doi.org/10.1038/s41598-019-41695-z)

Graph-based clustering

Leiden algorithm

- **Solution: add refinement step**

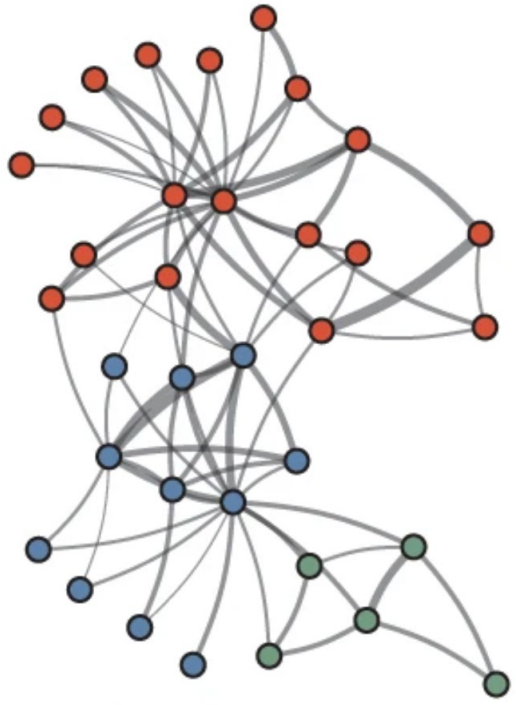




Traag et al. (https://doi.org/10.1038/s41598-019-41695-z)

Graph-based clustering

Overview

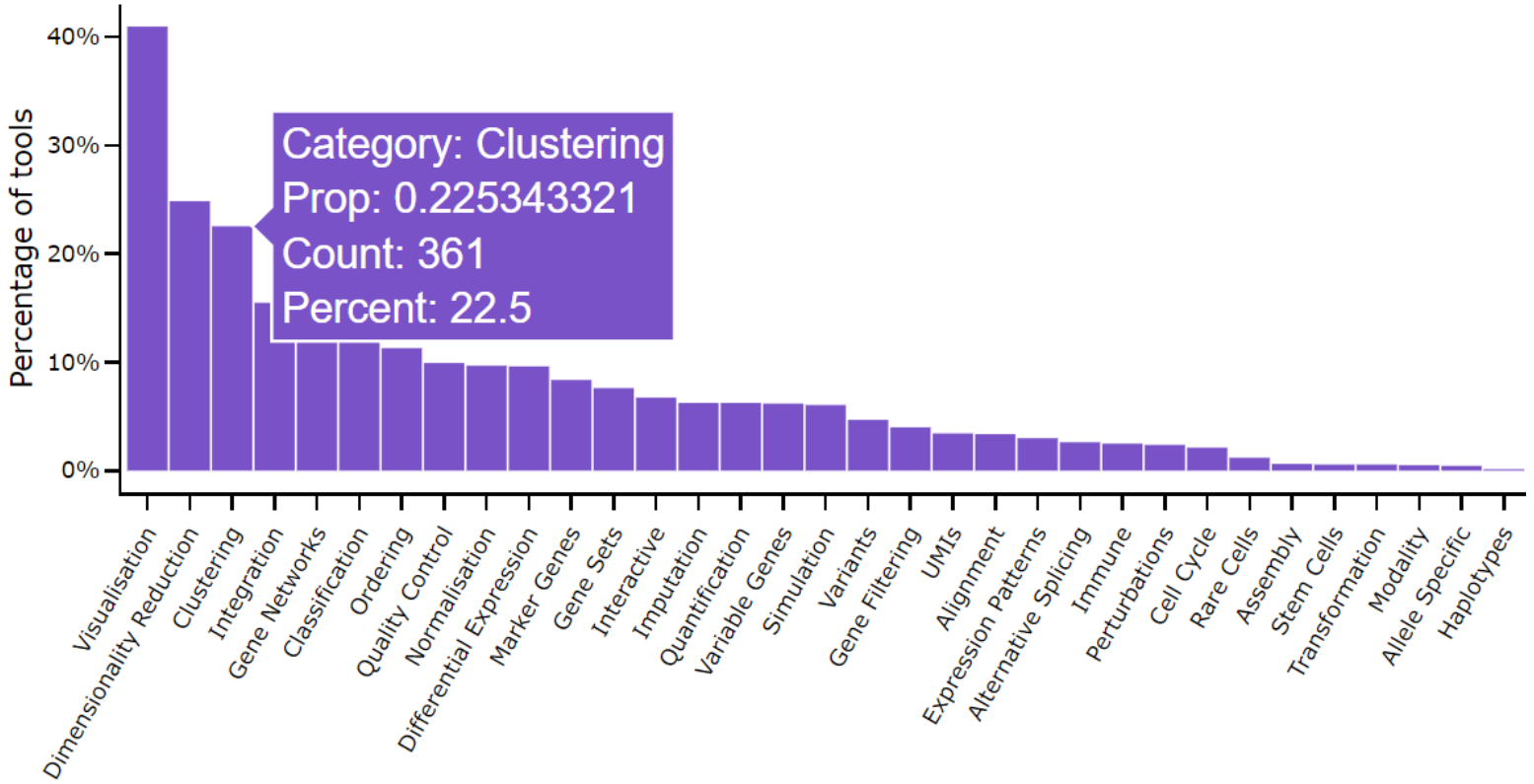


- Graph-based clustering
  - Number of neighbors when constructing the graph
  - Resolution parameters
- Resolution
  - High → less clusters
  - Low → more clusters
- Number of clusters
  - Determined using resolution parameter
  - Validate clustering!

Outline

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

>300 scRNA-seq clustering methods available



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

scRNA-seq clustering methods

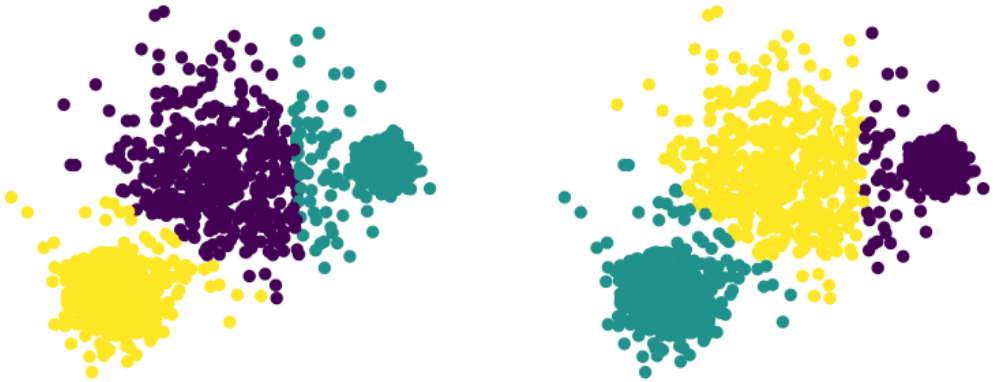
Name	Year	Method type	Strengths	Limitations
scanpy[4]	2018	PCA + graph-based	Very scalable	May not be accurate for small data sets
Seurat (latest)[3]	2016			
PhenoGraph[32]	2015			
SC3 (REF.[22])	2017	PCA + *k*-means	High accuracy through consensus, provides estimation of *k*	High complexity, not scalable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

How to compare different cluster labels?

Adjusted Rand Index (ARI)

Measure of the similarity between two data clusterings



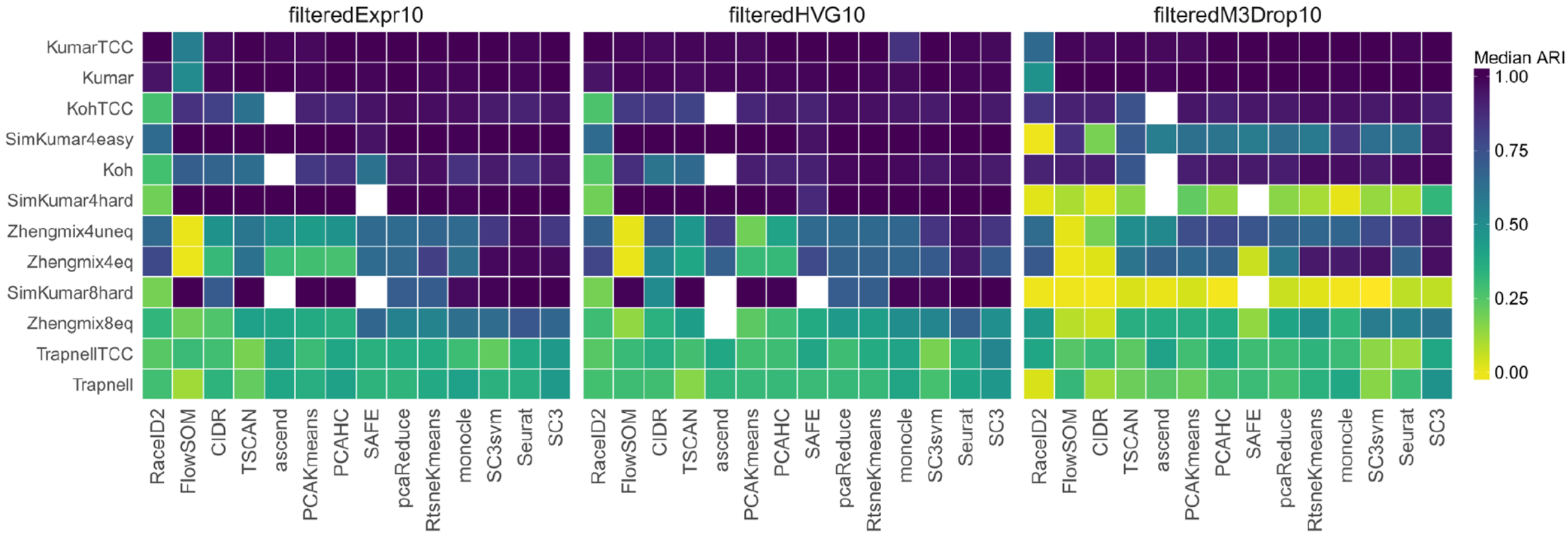
Given a set S of n elements, and two groupings or partitions of these elements

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

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

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

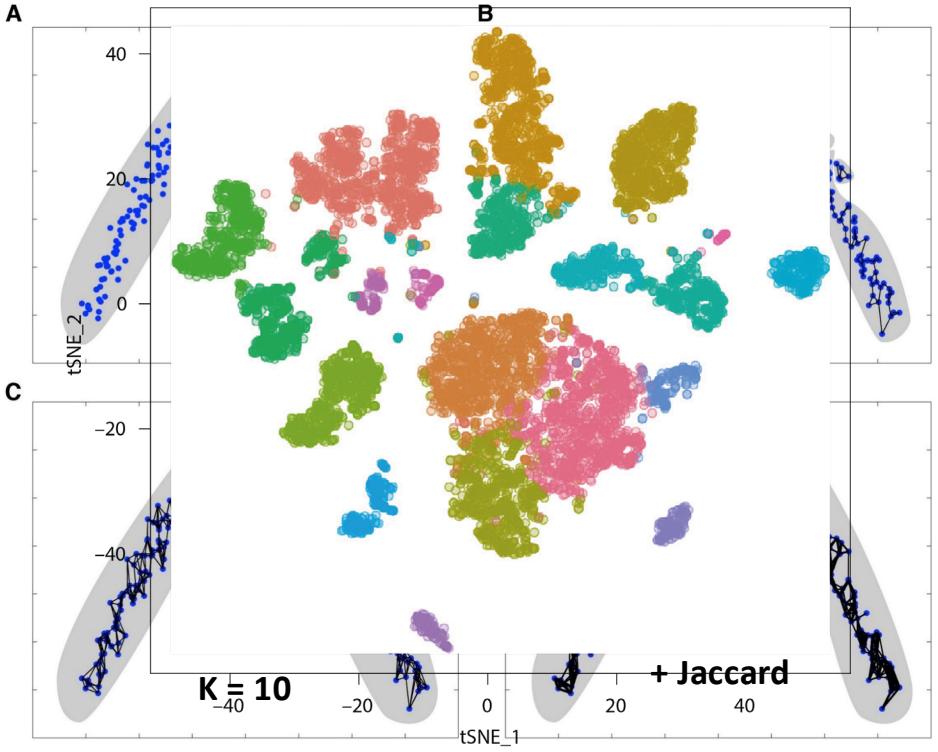
Benchmarking scRNA-seq clustering methods



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



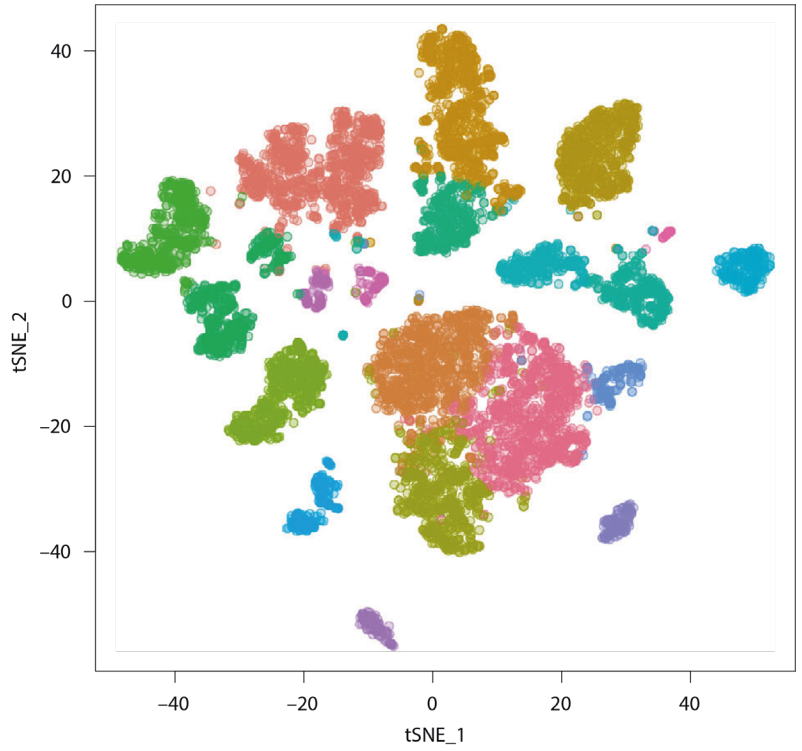
Levine et al. (https://doi.org/10.1016/j.cell.2015.05.047)

Outline

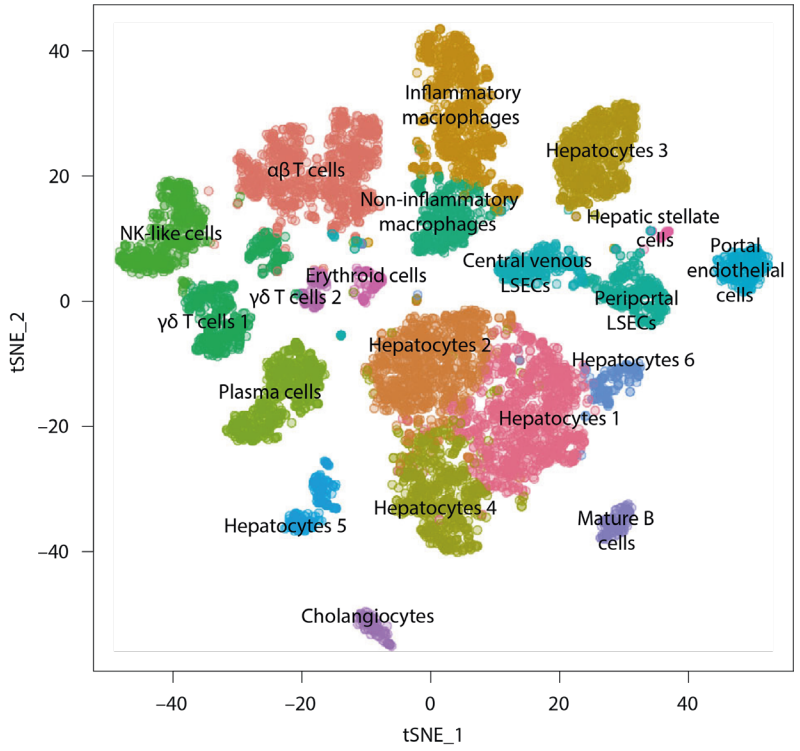
- Introduction to clustering
  - Hierarchical clustering
  - *k*-Means clustering
  - Graph-based clustering
- scRNA-seq clustering
- **Annotating clusters**
- Cluster validation
- Challenges & outlook

From clusters to annotations

Clustering



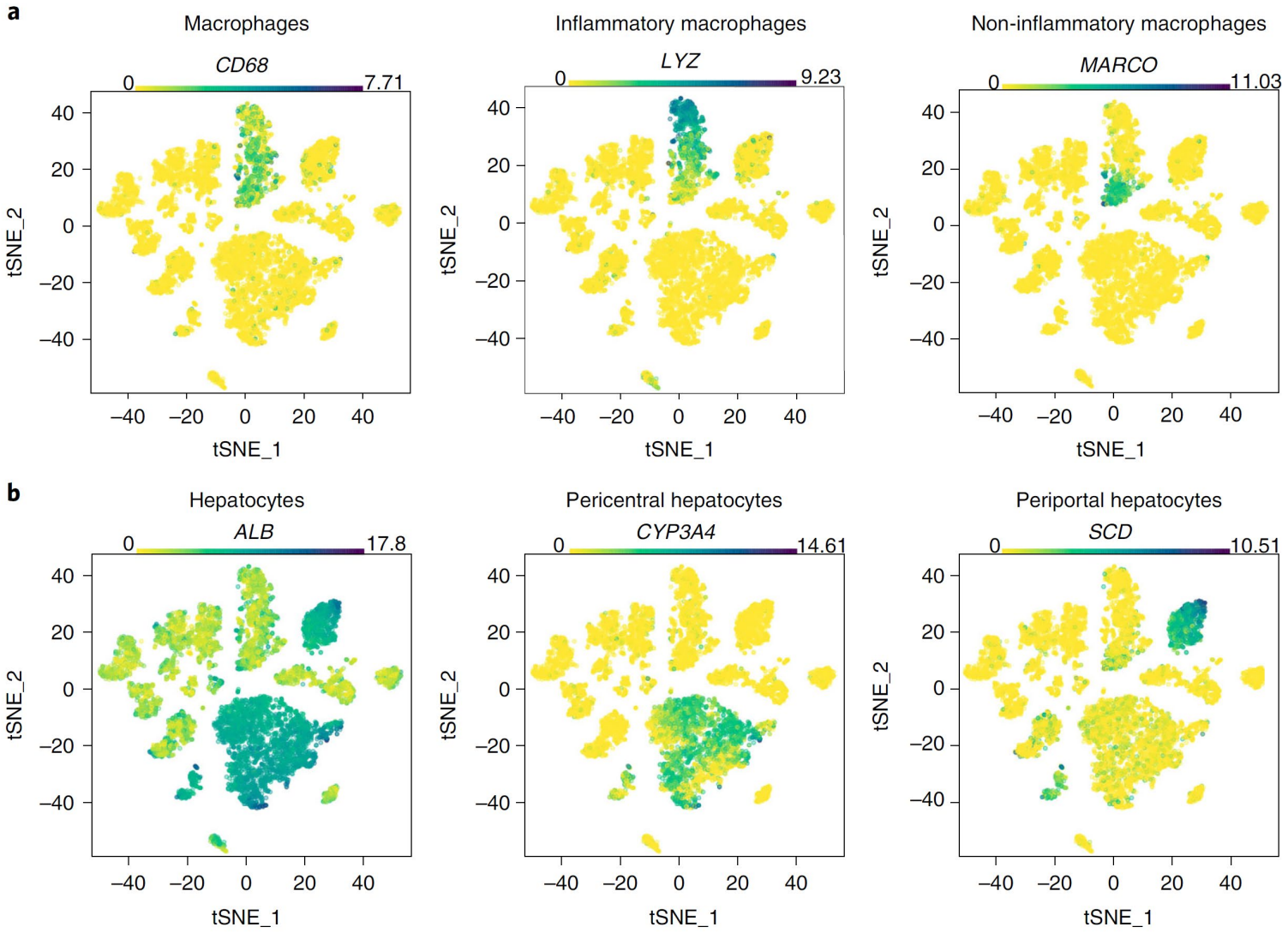
Annotation



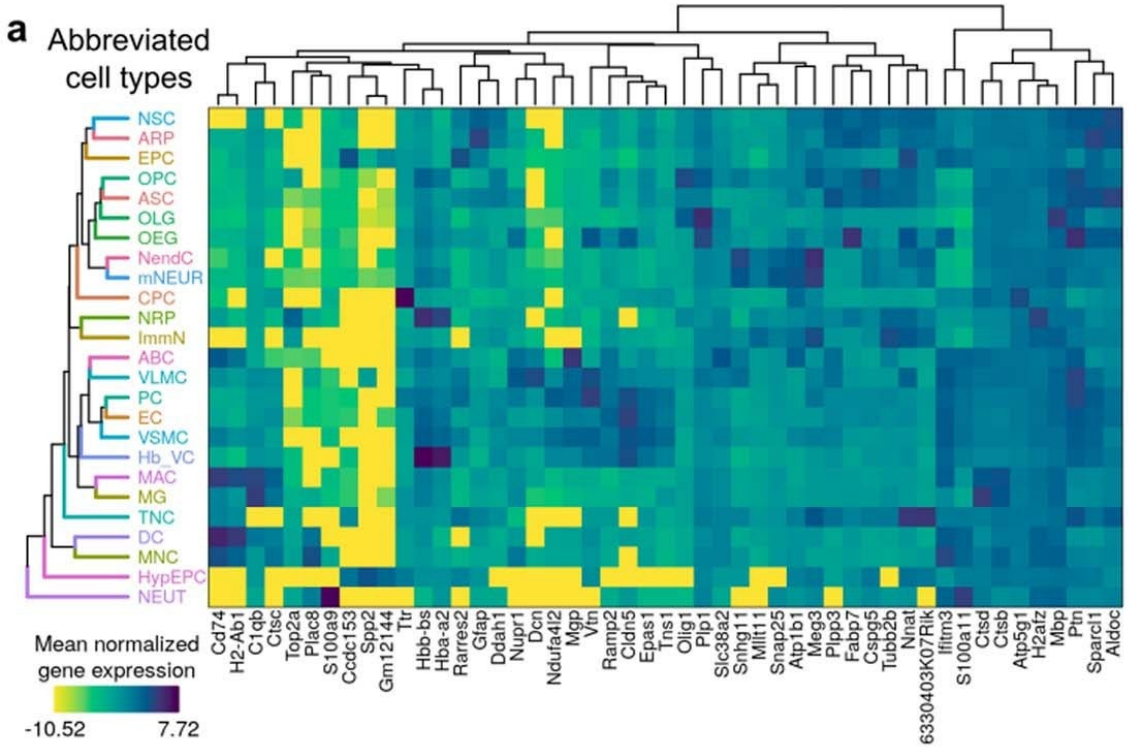
Gene expression overlay

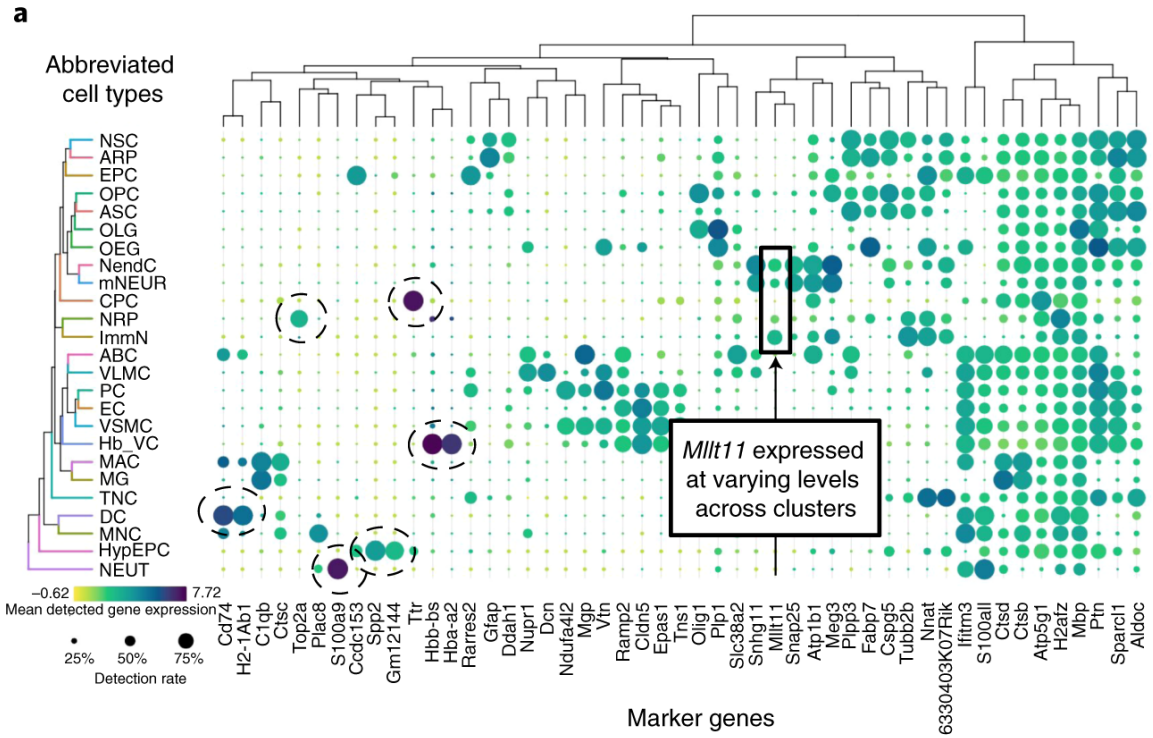
Easy

Challenging



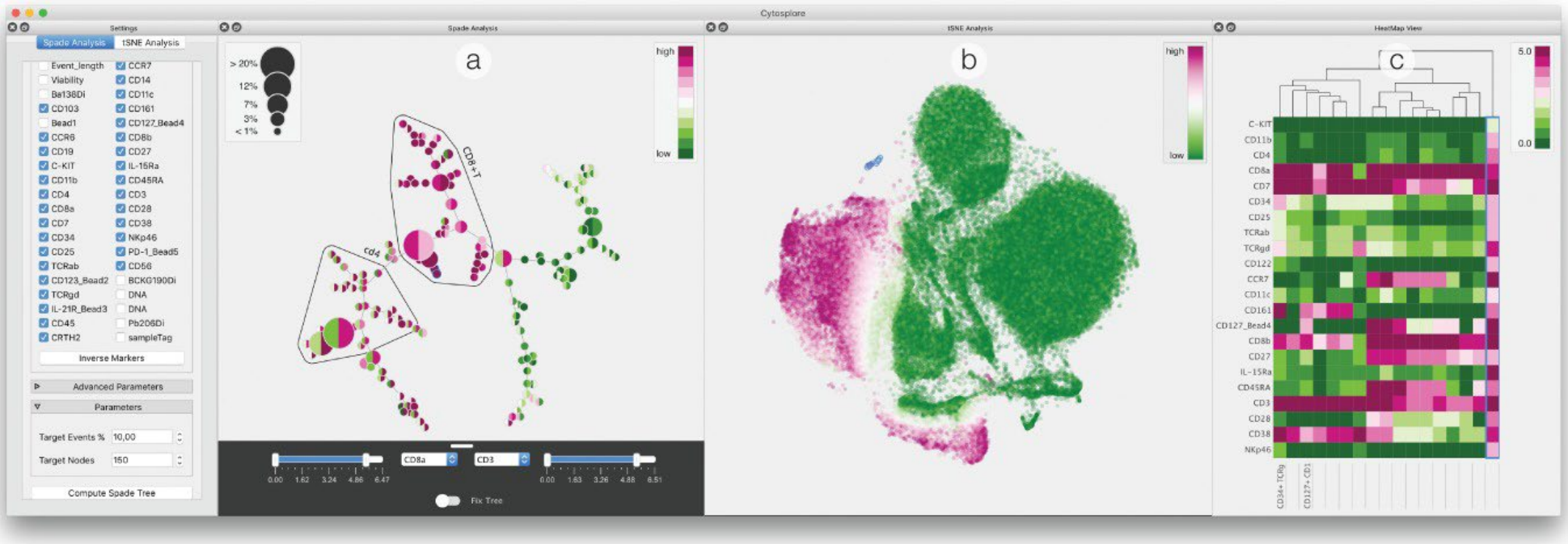
Alternatively: heatmaps & dot plots





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

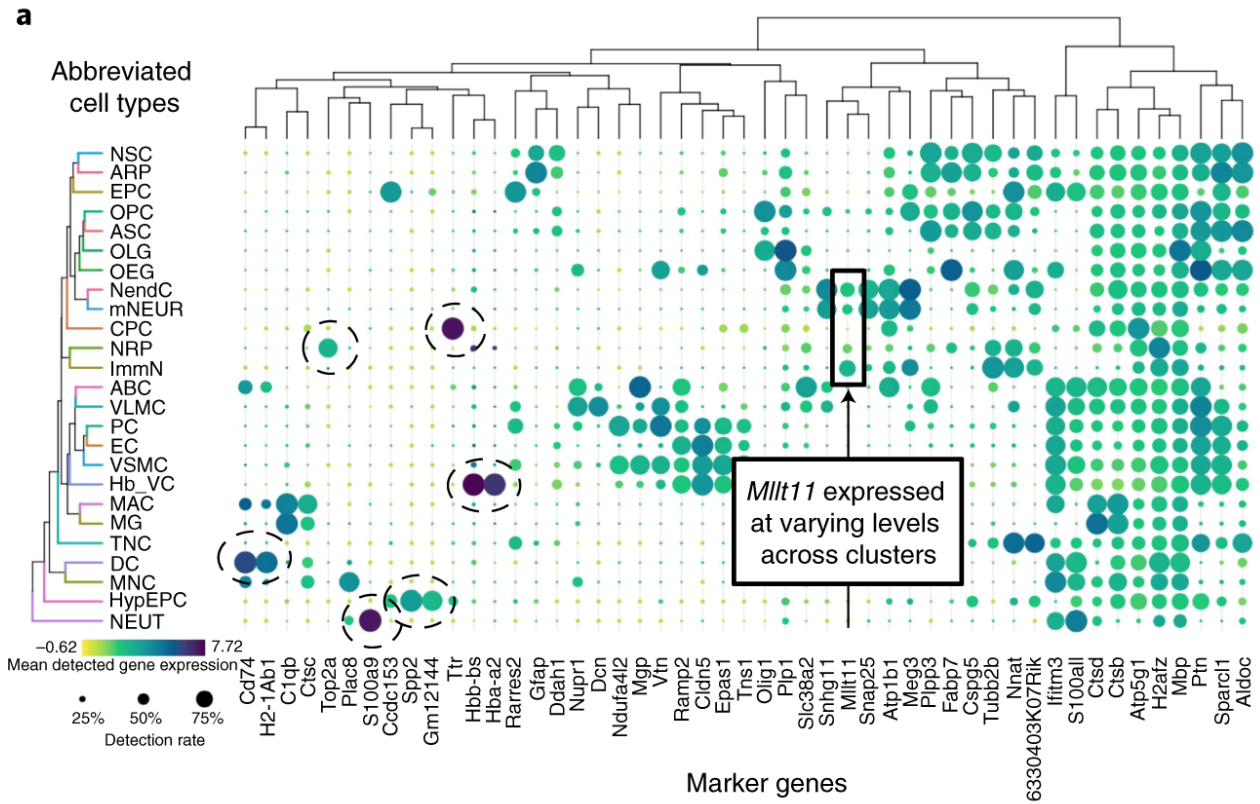
- "Expert knowledge"
- Literature
- Other scRNA-seq data
- Marker databases: PanglaoDB, CellMarker,…

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 differentially expressed genes to find functional information that may help identify the cell type



Complicating factors

1. Clusters that express markers of more than one cell type
   - Doublets?
   - Likely small, higher-than-average genes and UMIs per cell
   - Doublet detection tools: Scrublet, DoubletFinder, scds

2. Ambient RNA
   - RNA derived from one or more cell types that are sensitive to tissue dissociation
   - Markers of the contaminating cell types may be spread to all other cell clusters
   - Ambient RNA correction tools: SoupX, CellBlender

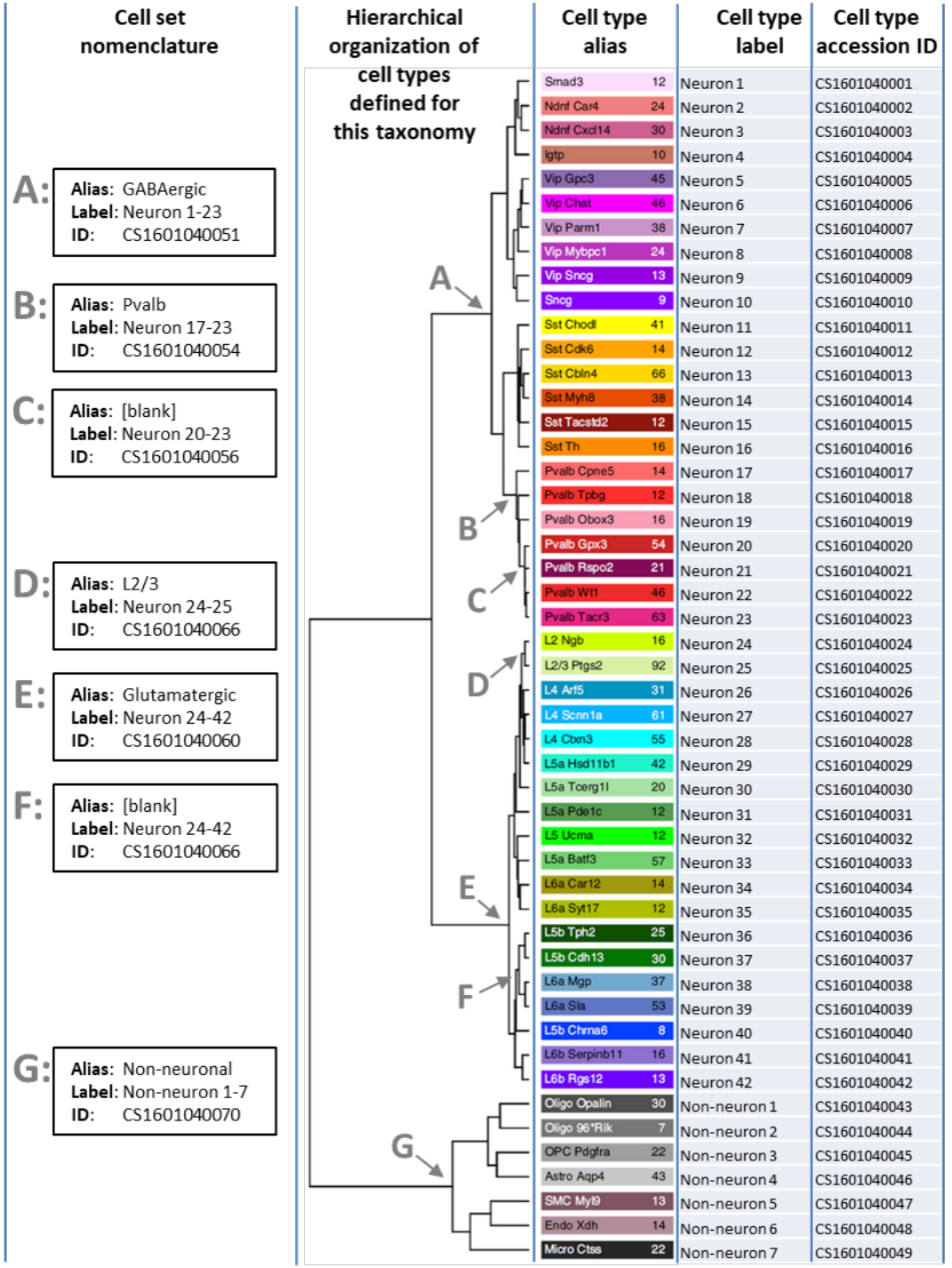
Watch out not to remove rare "interesting" cells!

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. (https://doi.org/10.7554/eLife.59928)
https://portal.brain-map.org/explore/classes/nomenclature

Summary

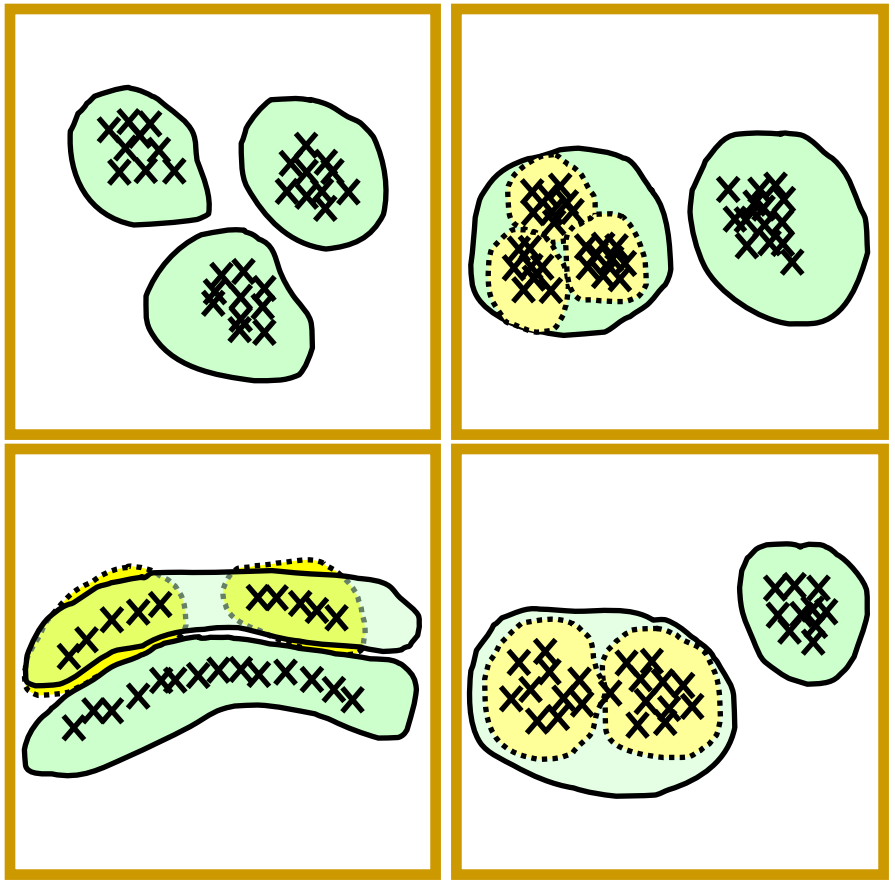
- Start by identifying major well-known cell types (clearly defined, discrete cell clusters)
- Split the data into broad subsets (e.g., immune, endothelial and tumor) and analyze each separately
- Cell subtypes or poorly defined clusters are challenging
- Manual annotations heavily rely on marker genes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

Outline

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

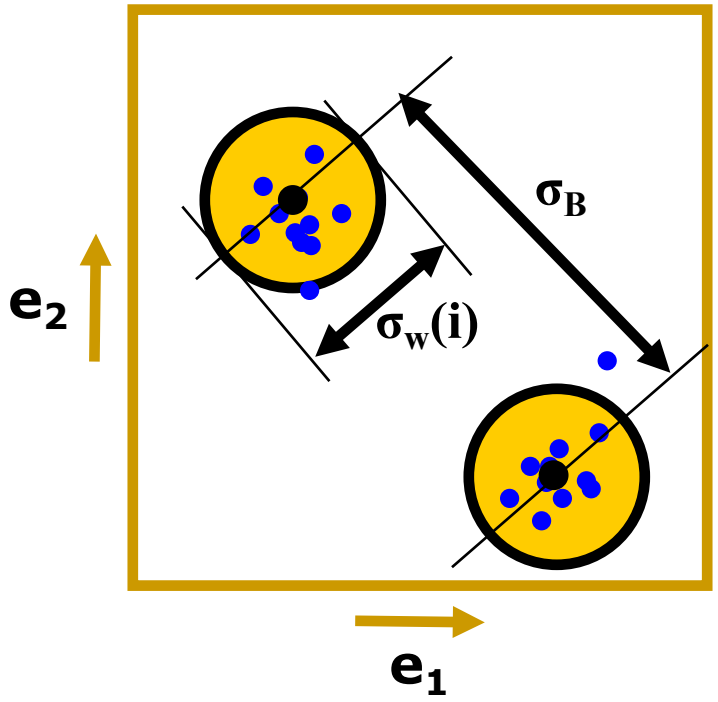
Clustering is subjective!

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



Cluster criteria

- Silhouette score
  - Goal: optimize cohesion within a cluster and separation between clusters
  - Seek: clustering that maximizes SI



Silhouette score

1. Mean distance between i and all other points in cluster C_i

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

2. Mean nearest cluster distance of i

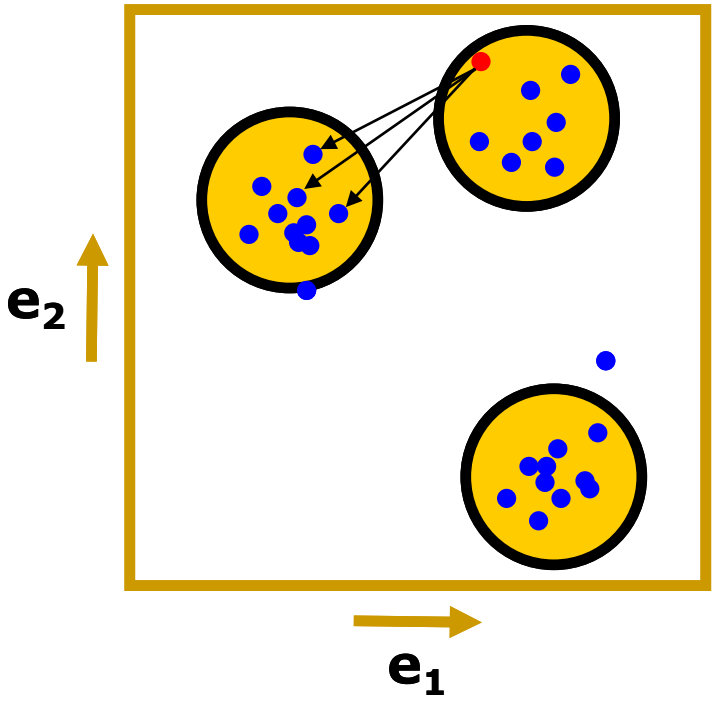
$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

3. Silhouette score for i

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

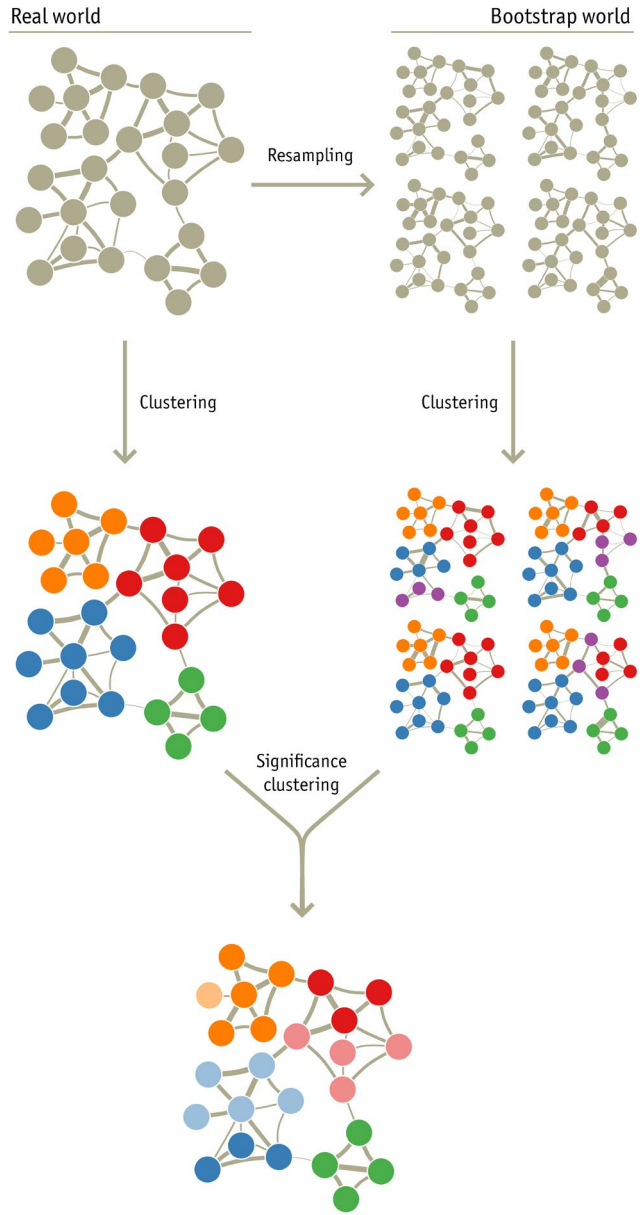
4. Total silhouette score

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



Bootstrapping

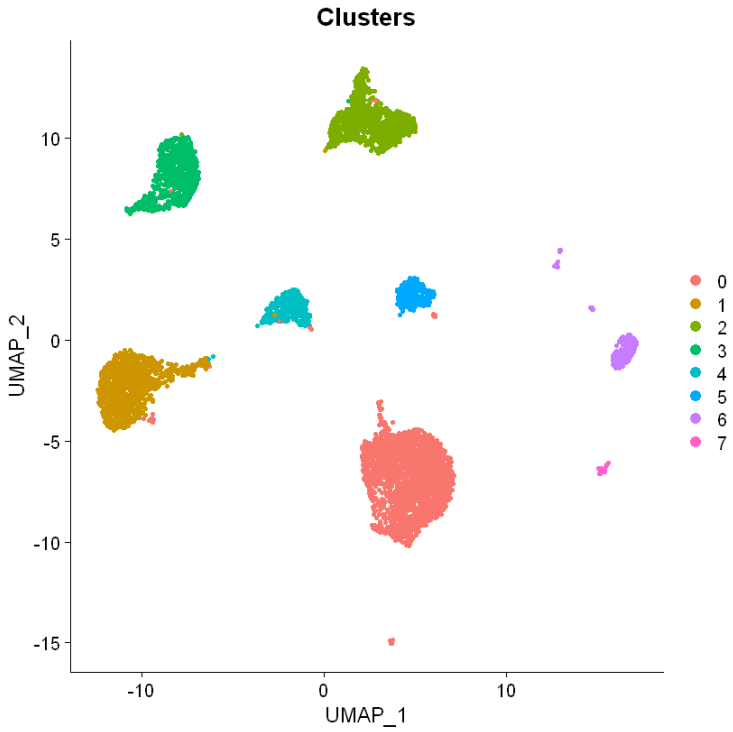
- How confident can you be that the clusters you see are real?
  - Take a random set of cells
  - Cluster
  - Compare to original clustering
  - Estimate support for clustering

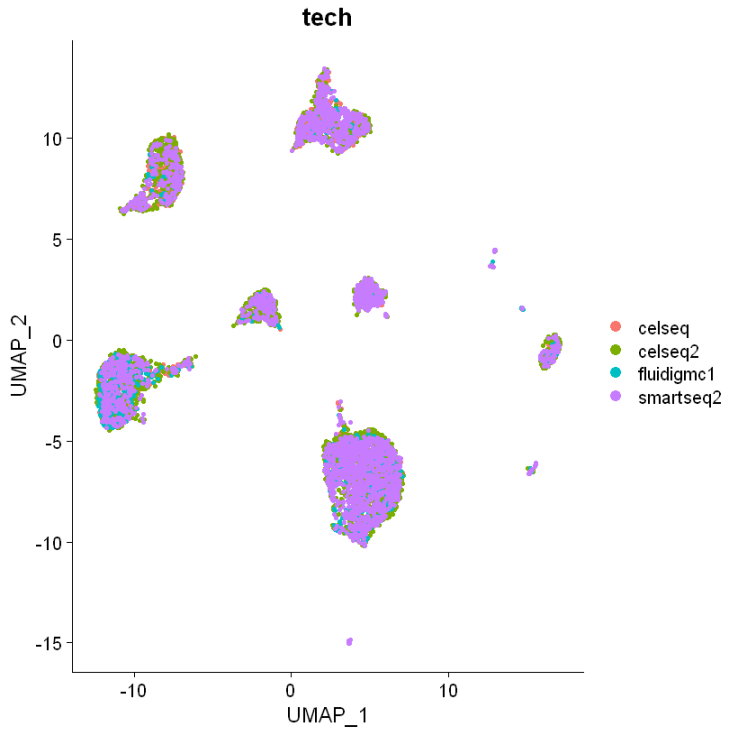


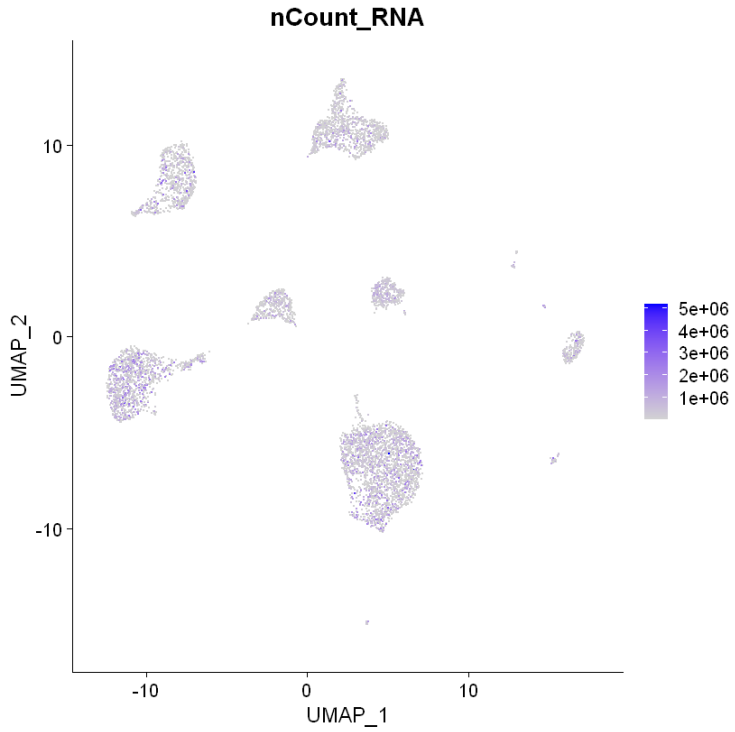
Rosvall & Bergstrom (https://doi.org/10.1371/journal.pone.0008694)

Always check QC data

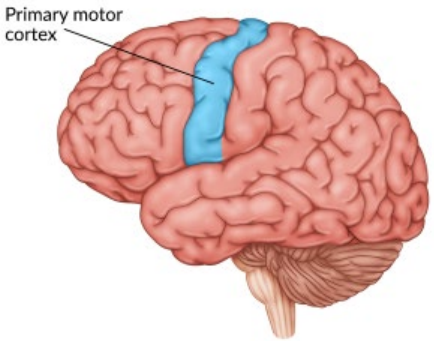
- Are your clusters mainly related to batches, qc-measures (especially detected genes)?







Example: annotating human brain cells



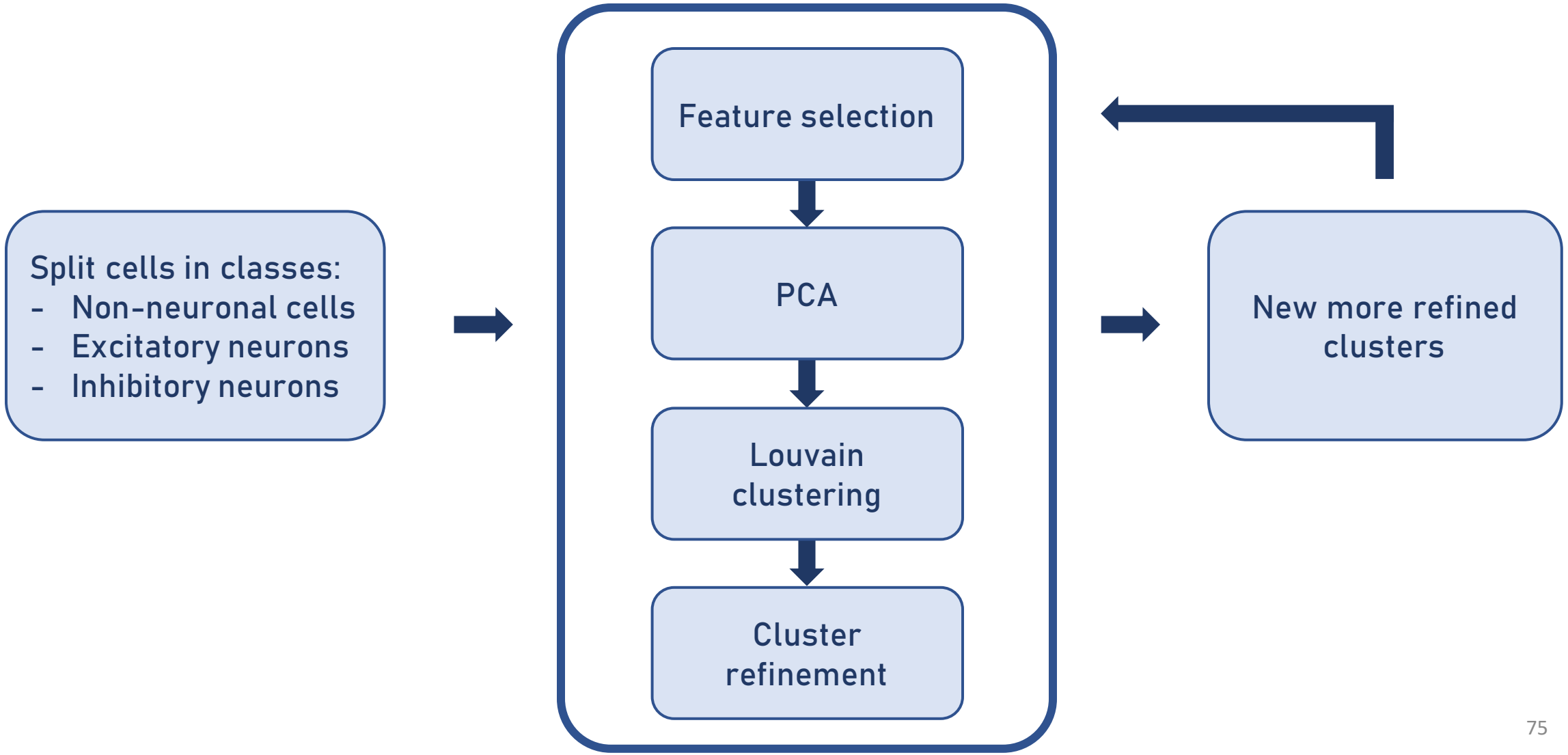
72,621 cells

32,991 genes

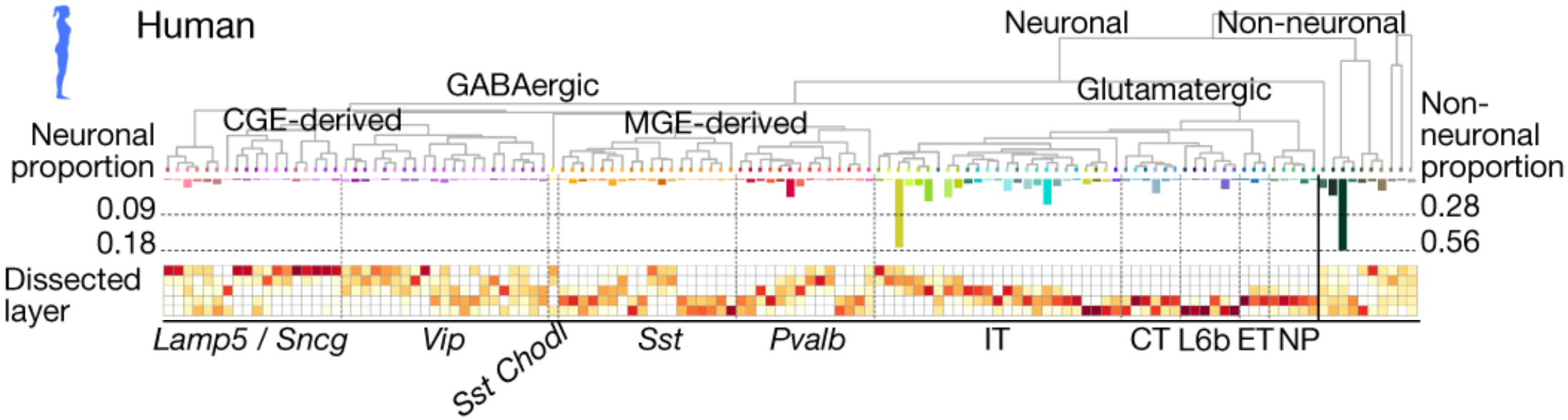
127 clusters

Bakken et al. (https://doi.org/10.1038/s41586-021-03465-8)

Iterative clustering approach



Example: annotating human brain cells



Bakken et al. (https://doi.org/10.1038/s41586-021-03465-8)

Outline

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

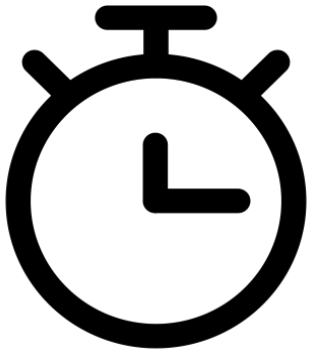
Challenges

- **Subjectivity:** what is a cell type?
  - Different parameters yield different results
  - Validation is important

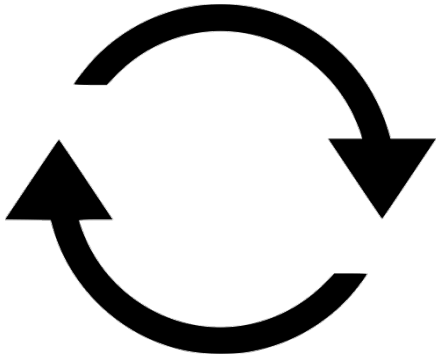
- **Scalability:** number of cells has grown from ~10^2 to ~10^6
  - Computational efficiency
  - Visual exploration, crowding problem

Downside of clustering

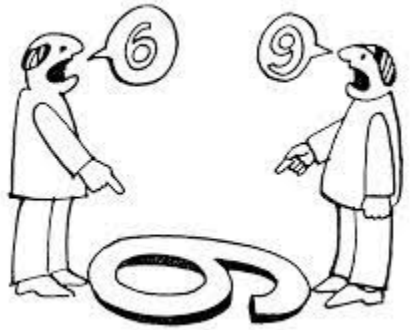
Time consuming



Not reproducible



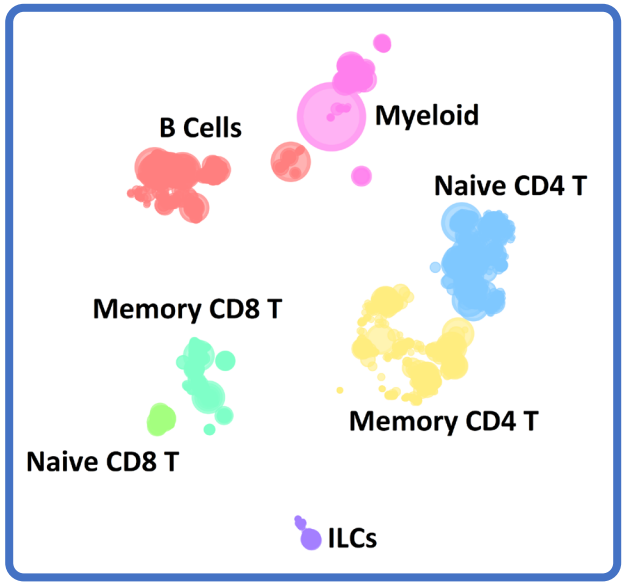
Subjective



Lots of single-cell data is available nowadays!
Can we use that to annotate our cells?

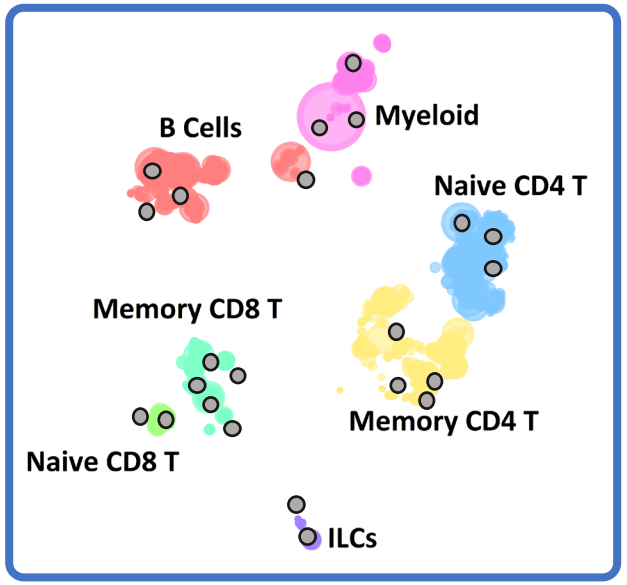
Supervised approach

Annotated cells (e.g. atlas)



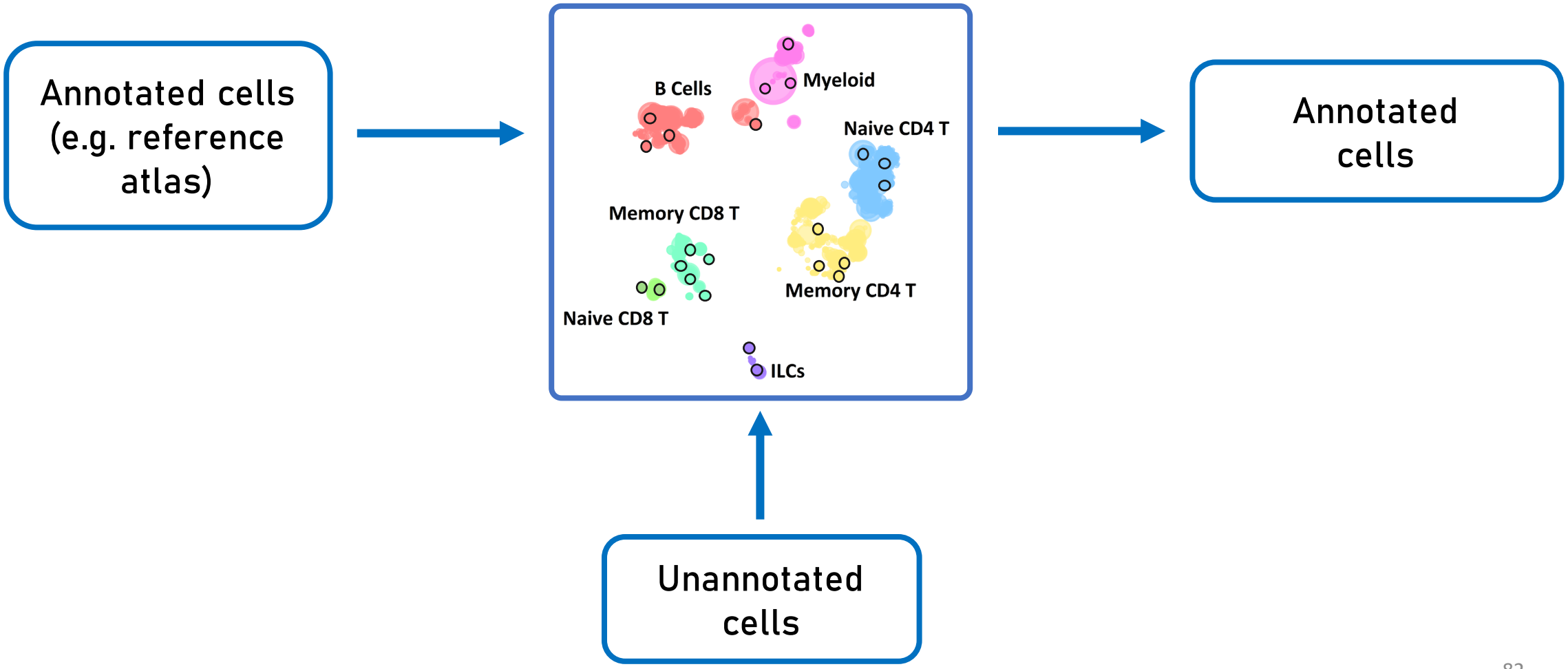
Supervised approach

Annotated cells (e.g. reference atlas)



Unannotated cells

Supervised approach



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