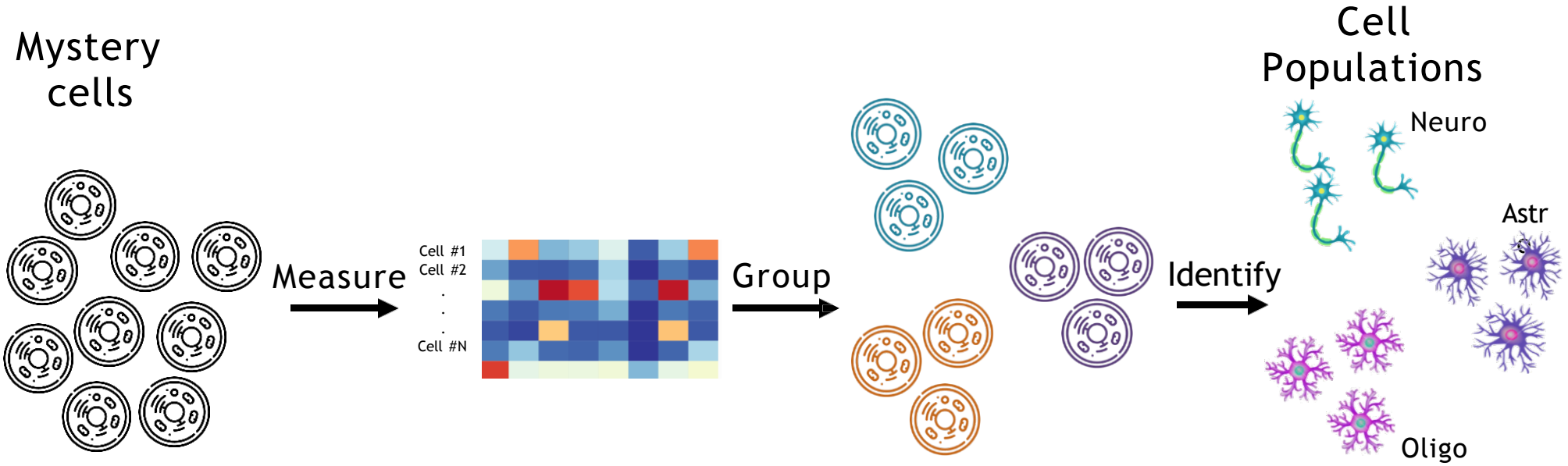
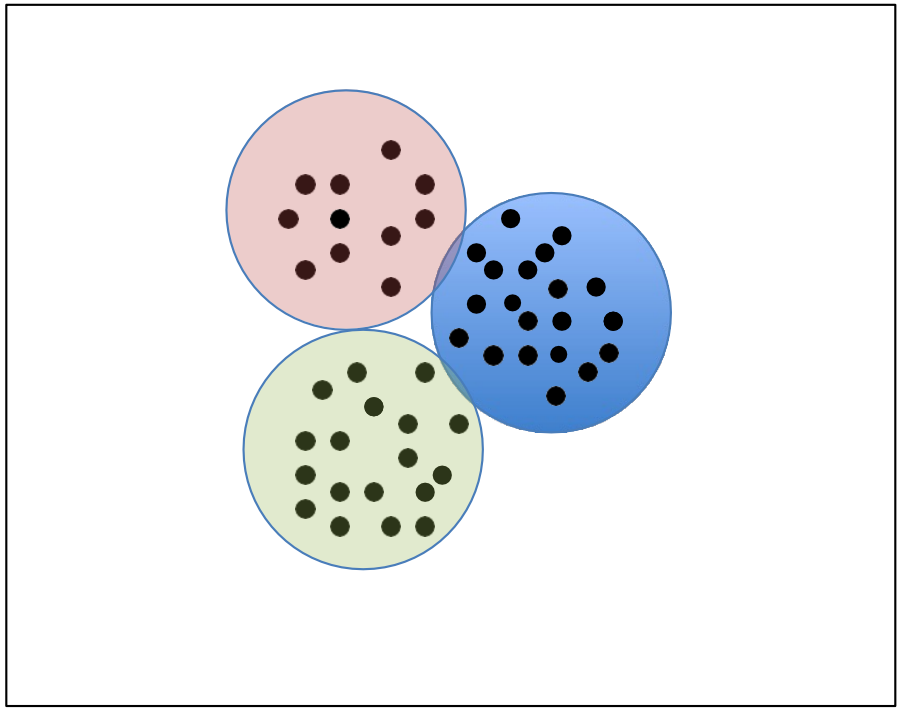
# Clustering -> Cell Identity



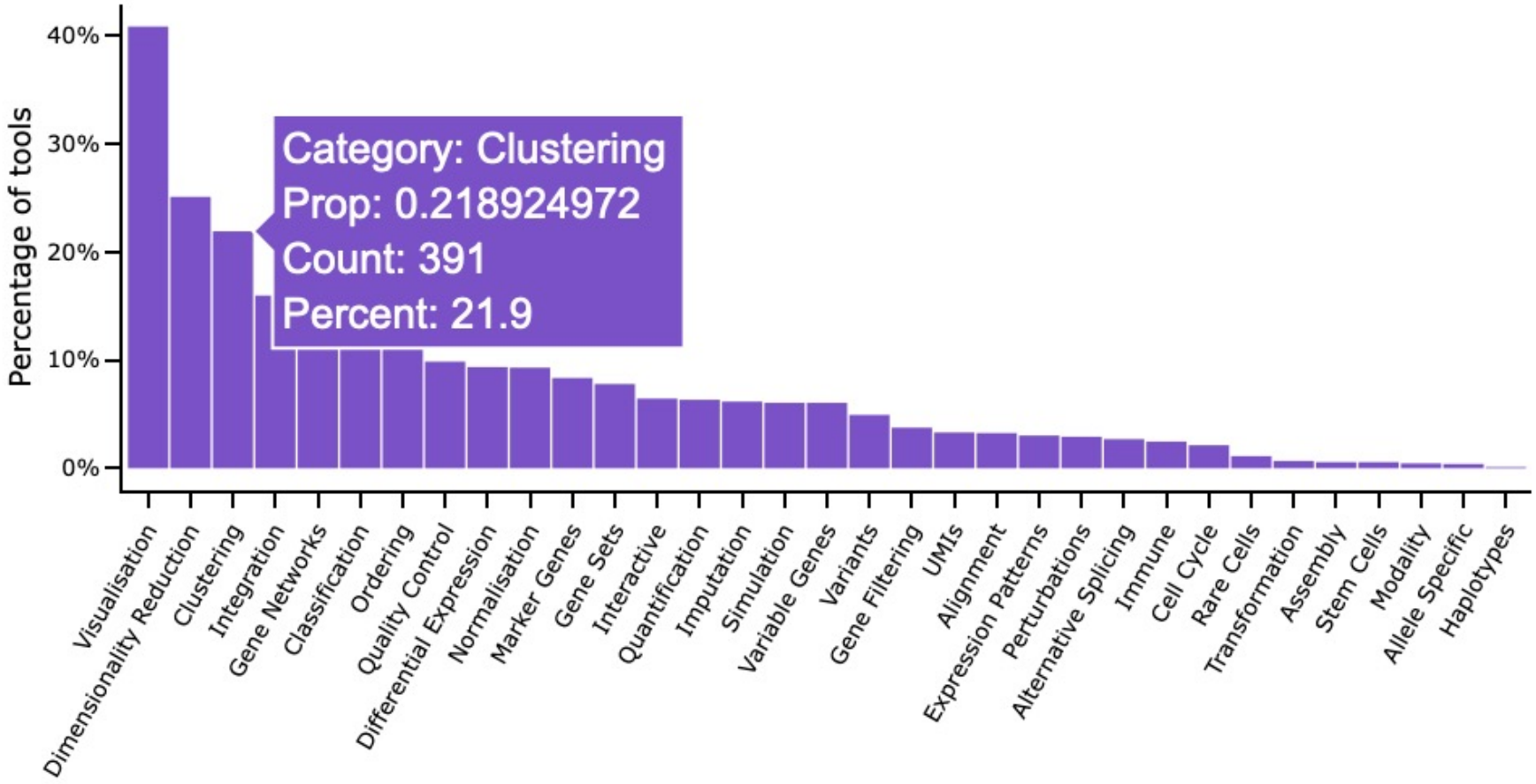
# Goal of clustering

Group similar cells together so that we can identify populations.

Decide where to partition the groups



# scRNA-seq clustering methods

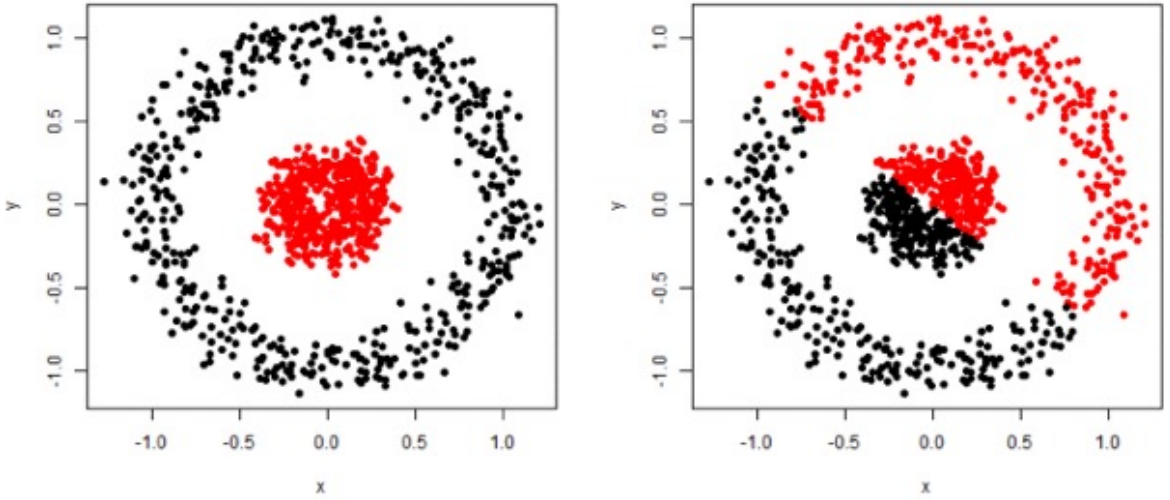



May 24th, 2024
https://www.scrna-tools.org/

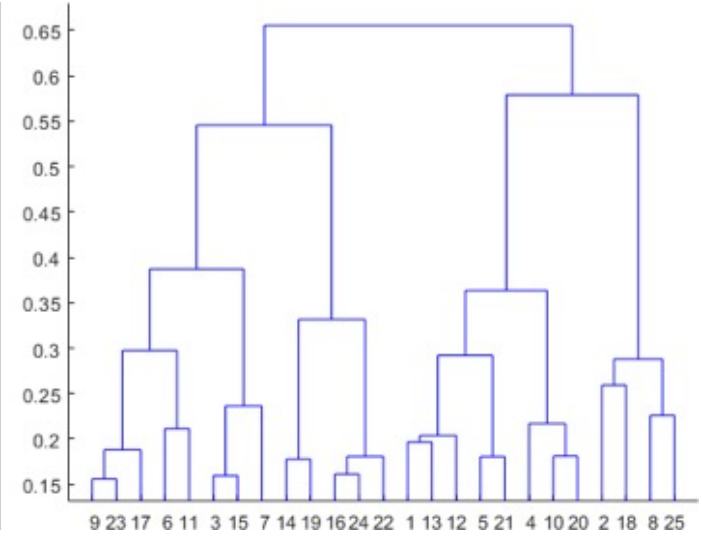
# Clustering methods

Clustering methods are divided into two categories*

Partitioning clustering




Hierarchical clustering




*Handbook of cluster analysis, Hennig C. et al.

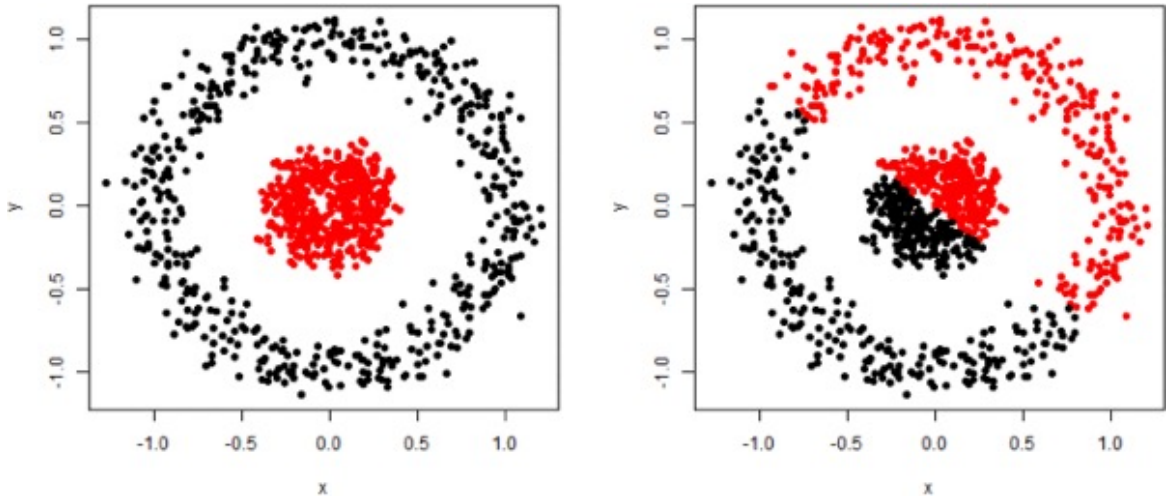
# Partitioning clustering

Convex partitioning. Example: K-means

Density based approaches. Example: DBSCAN

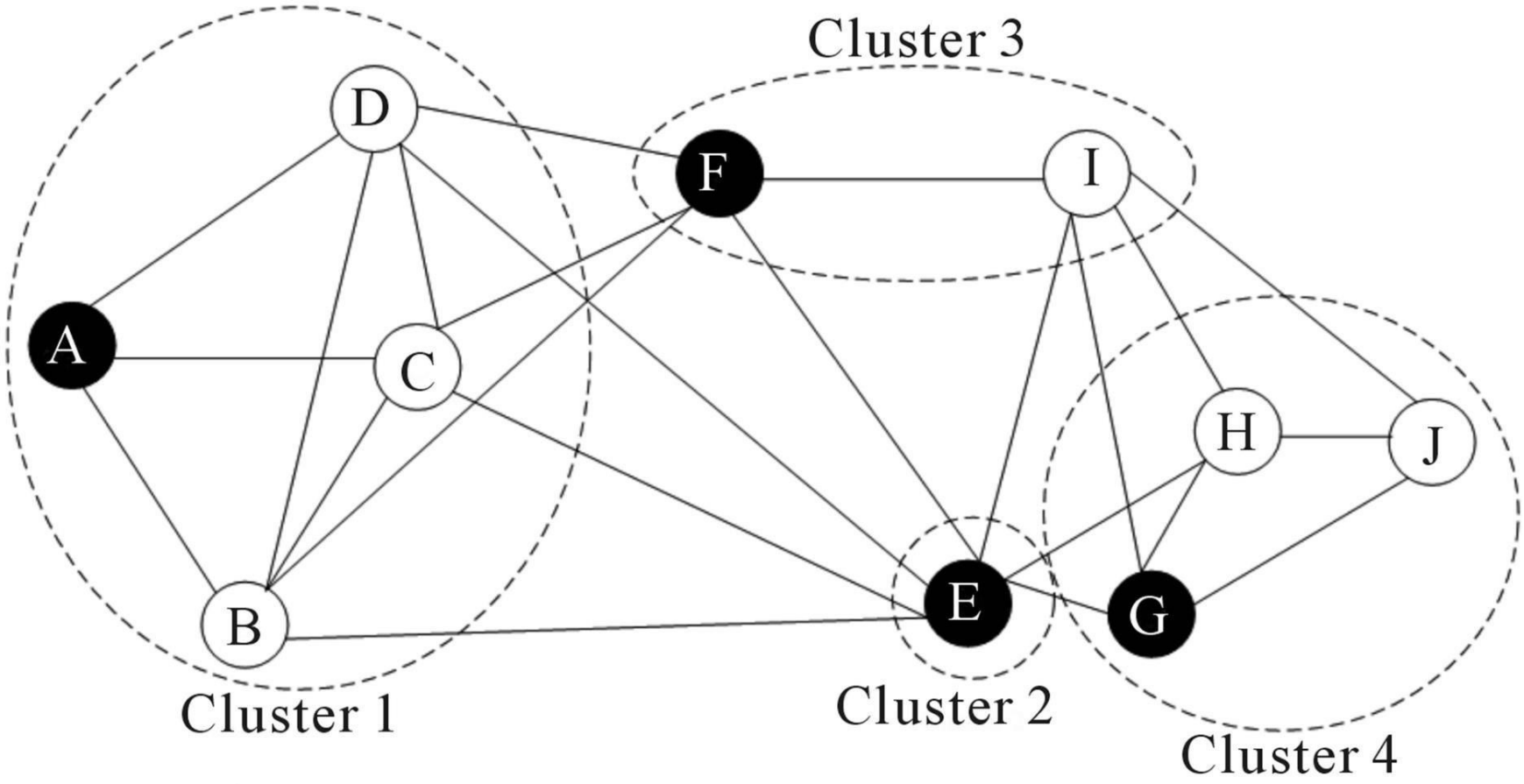
Model-based approaches. Example: Mclust

Graph based approaches : Example to follow



# Graph-based

- Nodes -> cells
- Edges -> similarity ()



# Graph-based: 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.

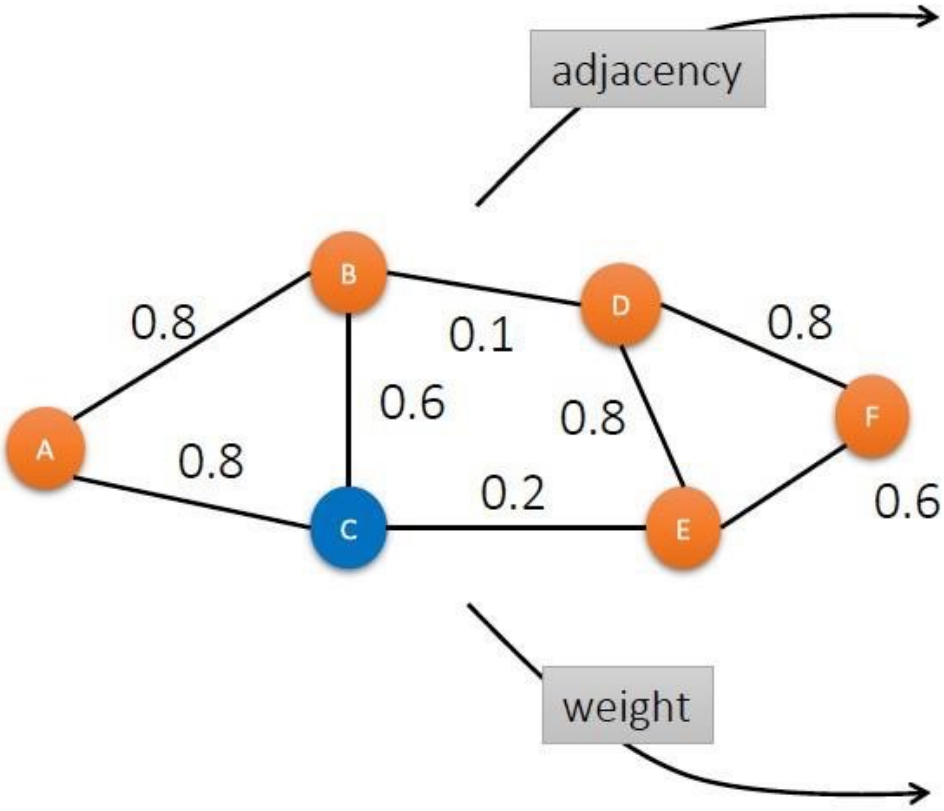
```
sc.pp.neighbors(adata, n_pcs = 15, n_neighbors = 15, knn=True)
```

specify the number of neighbors (2-100, eg. 30) and number of PCs to use

Too few neighbors: not enough connections for frequent, related cell types

Too many neighbors: may connect small groups to distant, unrelated cell types

# Graph-based: types

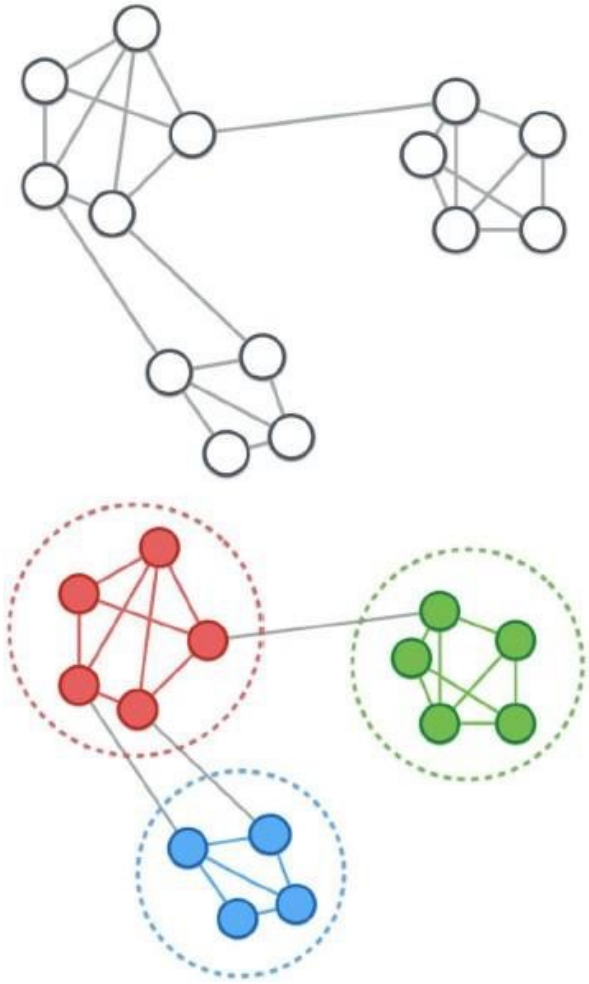



$$A = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 &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

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

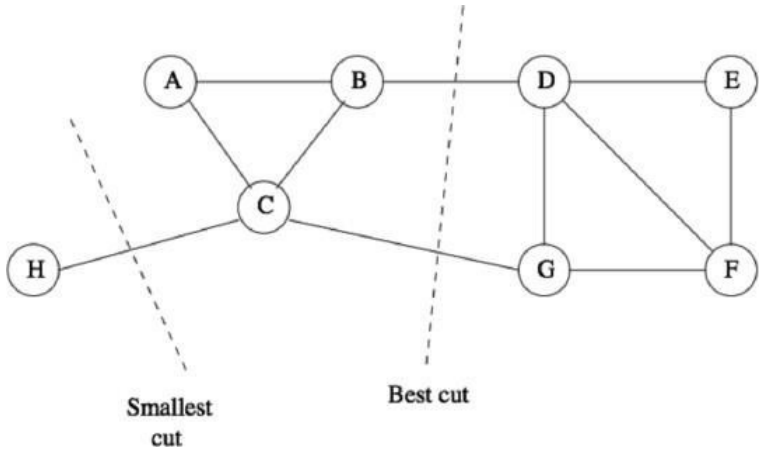
# Graph-based: communities

- 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-based: 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




*Shi & Malik (IEEE PAMI 2000)

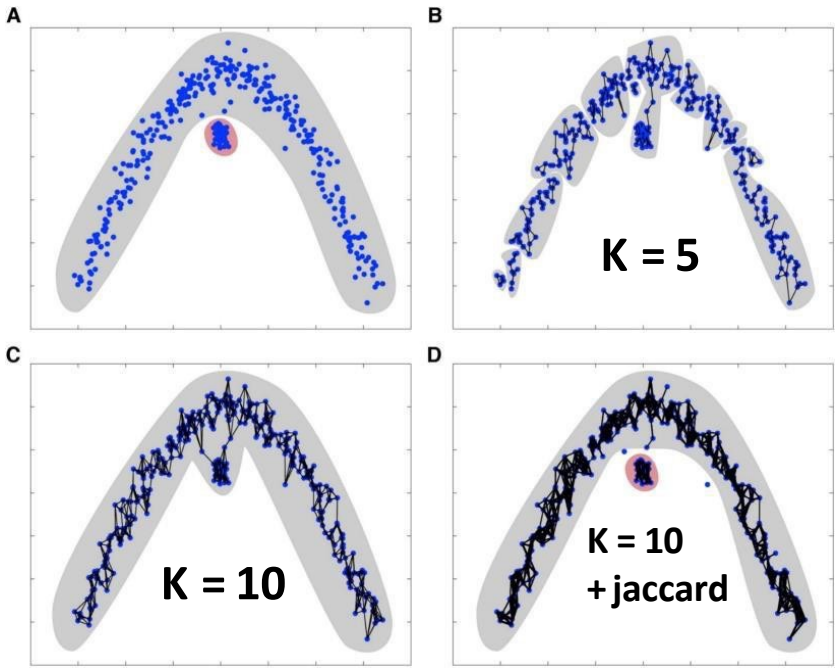
# Louvain algorithm

1. Construct kNN graph based on the Euclidean distance in PCA space. (Could also be SNN)

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

3. Cluster cells by optimizing for modularity (cuts)

**Modularity** is a cost function, resolution is a parameter used to calculate the modularity.

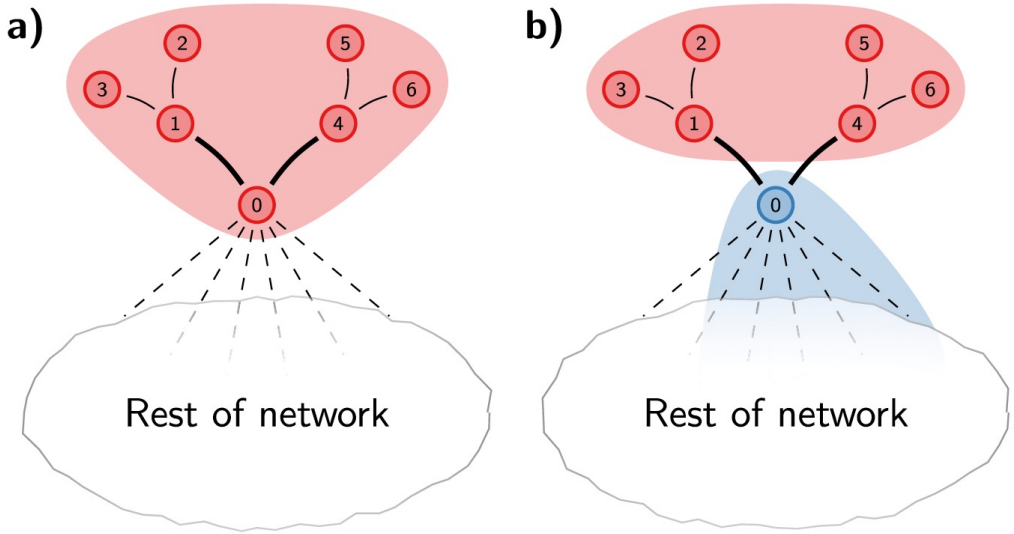



Xu and Su (https://doi.org/10.1093/bioinformatics/btv088)
Levine et al. (https://doi.org/10.1016/j.cell.2015.05.047)

# Scanpy: Louvain or Leiden clustering

Leiden clustering is an improvement of the Louvain clustering, it improves upon it to produce more reliable and meaningful clusters. Faster algorithm.

Leiden guarantees well-connected communities by refining edges within sub-communities.

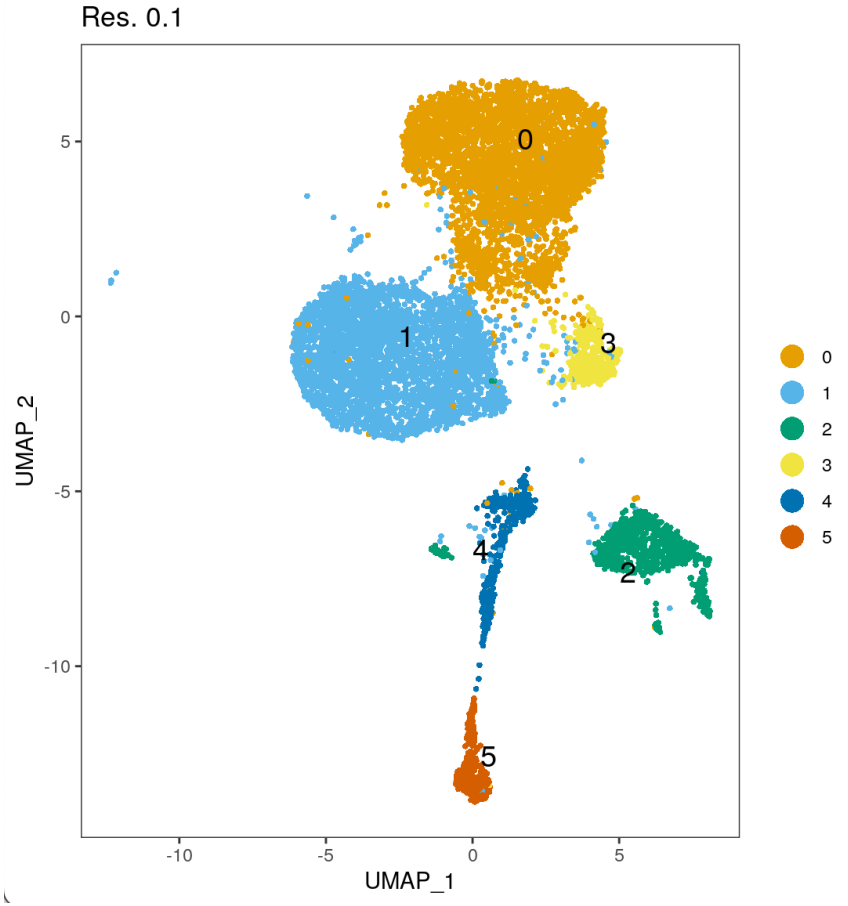


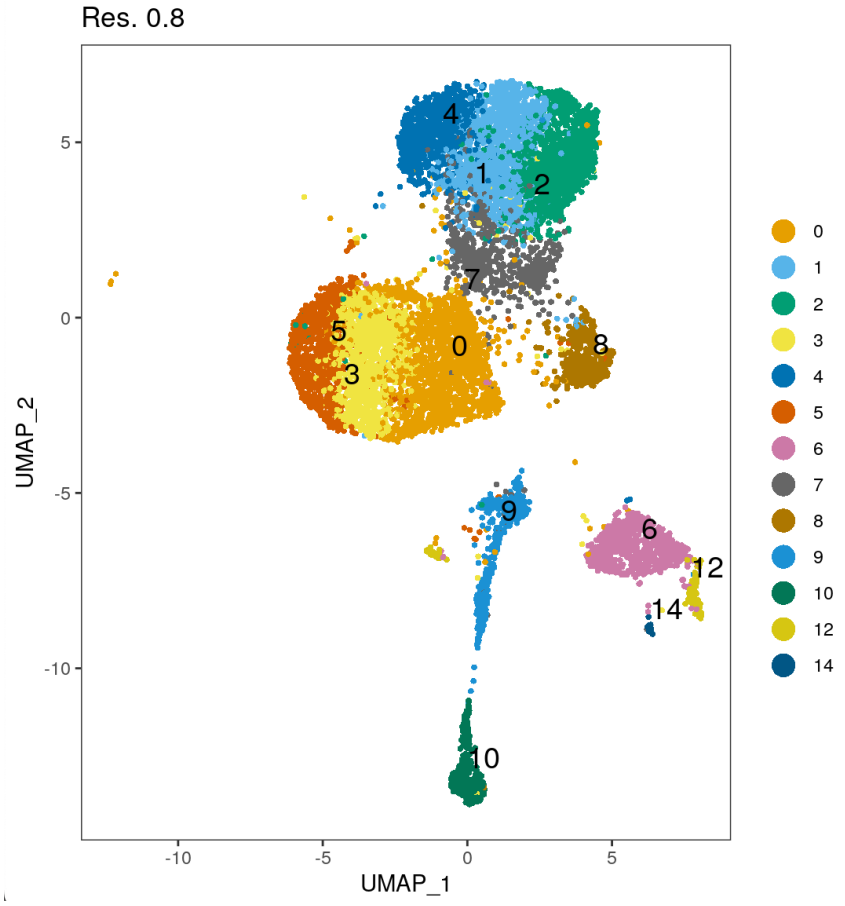

```
sc.tl.leiden(adata, resolution=1)
sc.tl.louvain(adata, resolution=1)
```

Traag et al, 2019 https://doi.org/10.1038/s41598-019-41695-z

# Resolution parameter

Resolution: A parameter value controlling the coarseness of the clustering.
Overlay cluster IDs on the UMAP (visualization)
Check which resolution matches your biological knowledge / expectations



# Clustering: Challenges

- What is a cell type?
- What is the number of clusters *k*?
- Check QC after clustering to see if no biases are constituting your clusters
- How stable are the clusters, how dependent are the clusters on the surrounding cells
- Clustering is subjective – No ground truth
- Justify clusters: known marker genes, eg dotplot
- 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 ~$10^6$

# Question on clustering