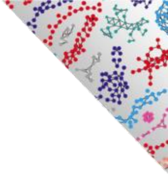
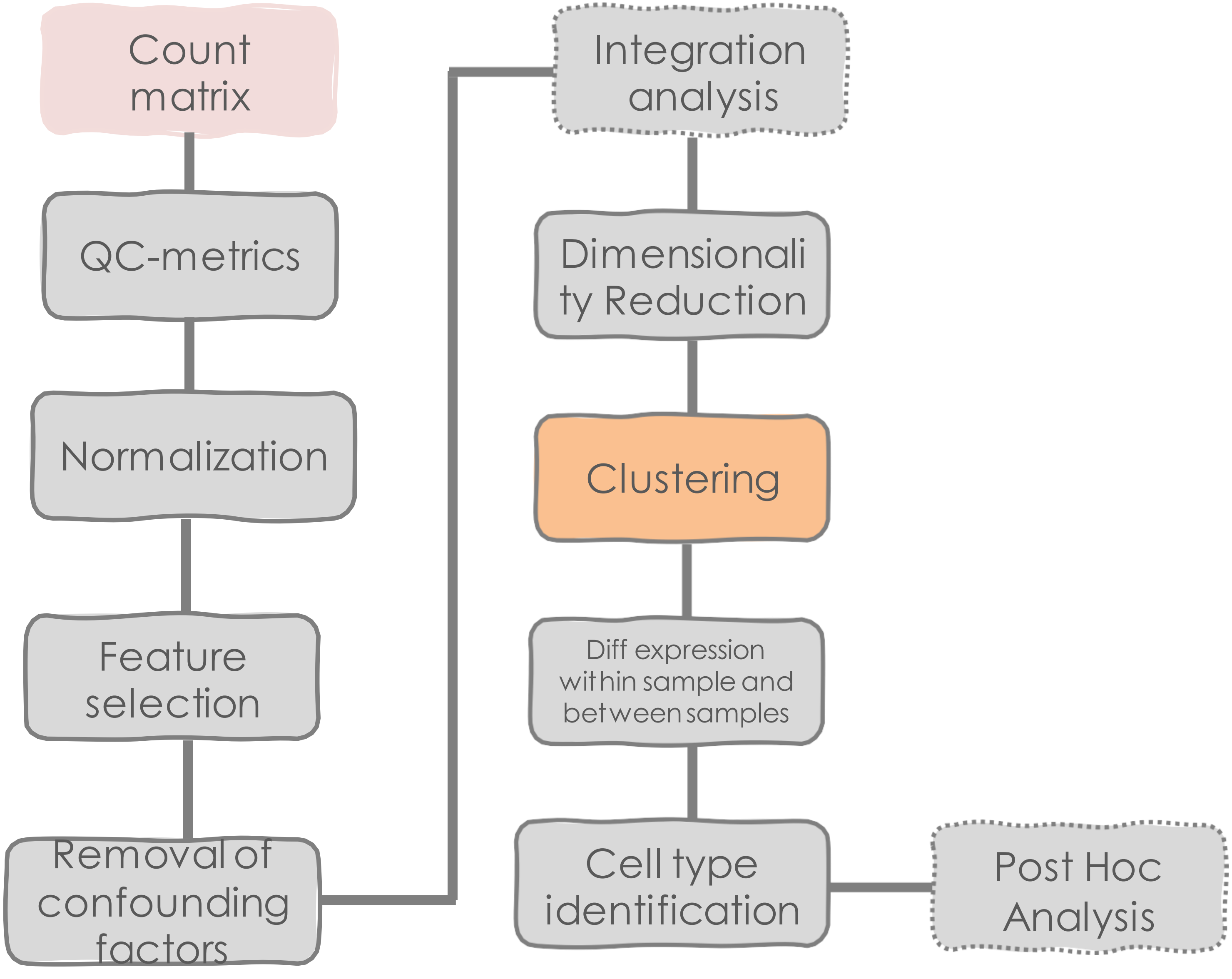




Count matrix

QC-metrics

Normalization

Feature selection

Removal of confounding factors

Integration analysis

Dimensionality Reduction

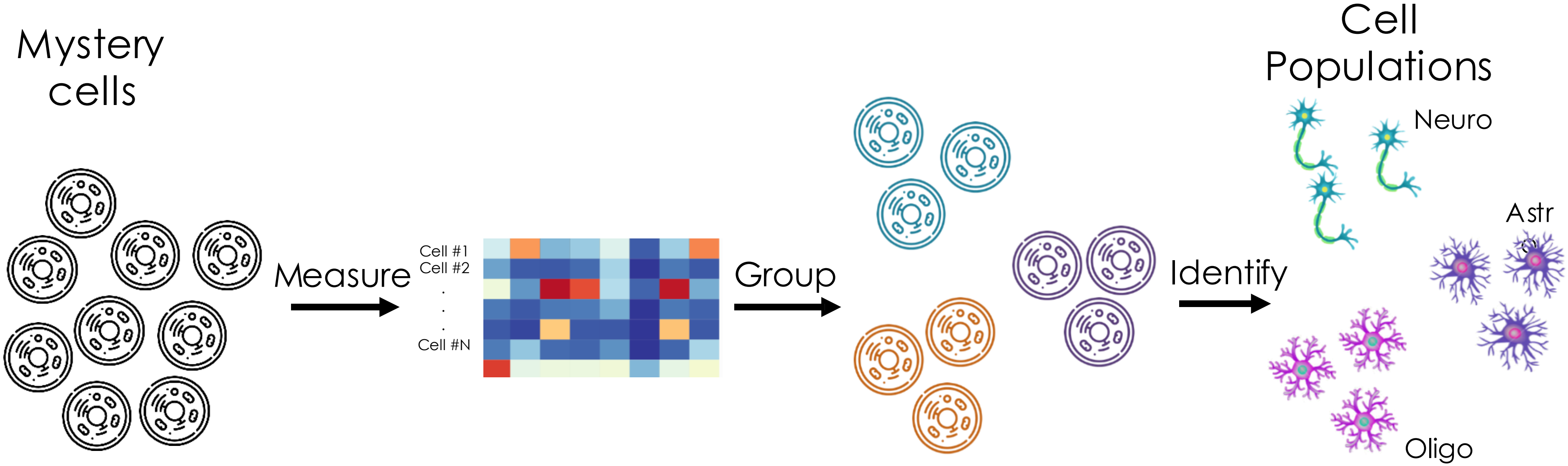
Clustering

Diff expression within sample and between samples

Cell type identification

Post Hoc Analysis

Clustering: Cell Identity



Mystery cells

Measure

Cell #1
Cell #2
.
.
.
Cell #N

Group

Identify

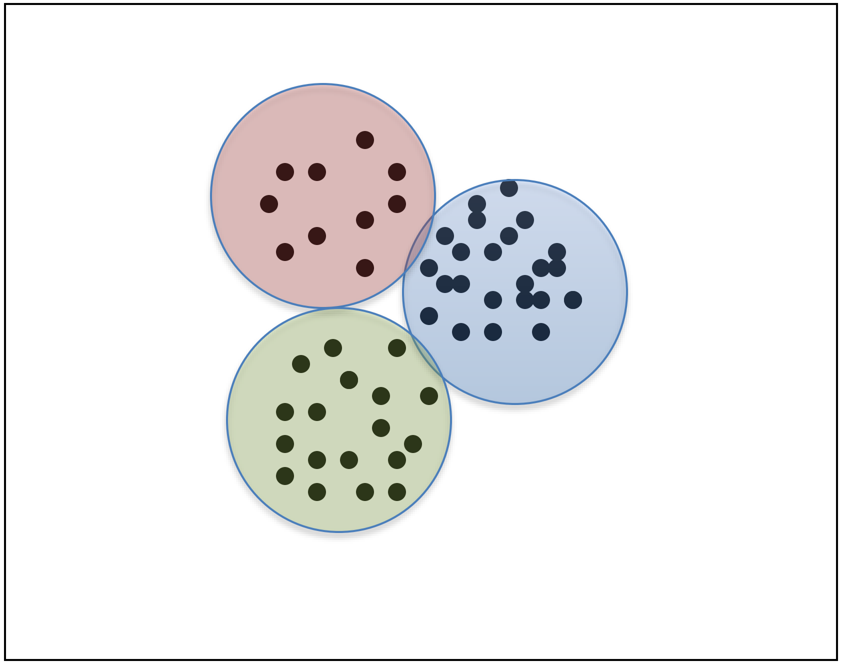
Cell Populations

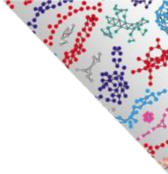
Neuro

Astro

Oligo

Clustering: Cell Identity

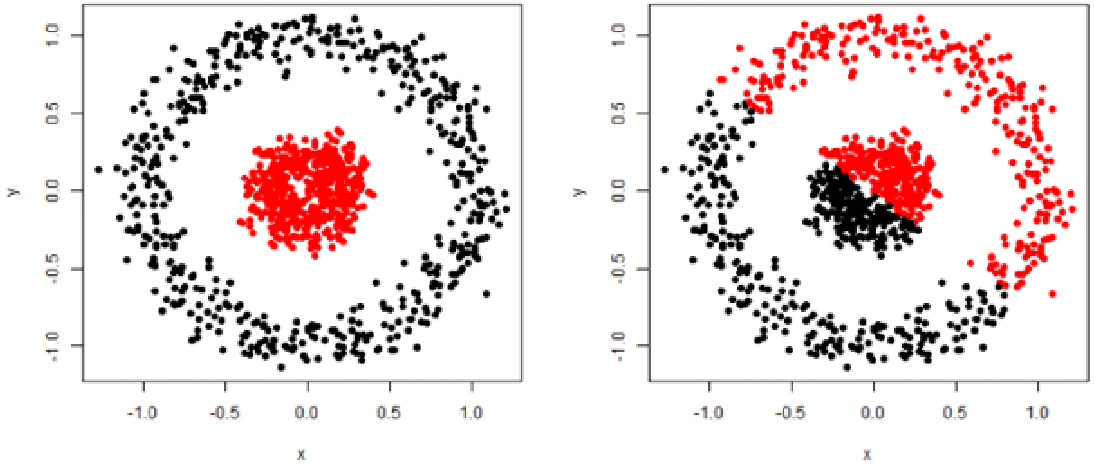




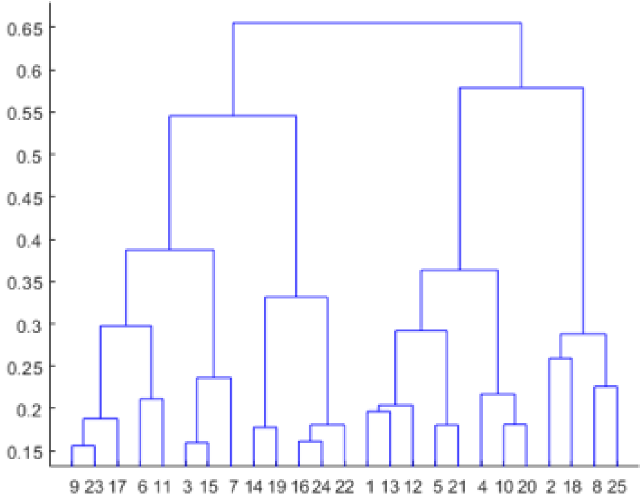
Clustering

Clustering method are divided into two categories* :

Partitioning clustering



Hierarchical clustering



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

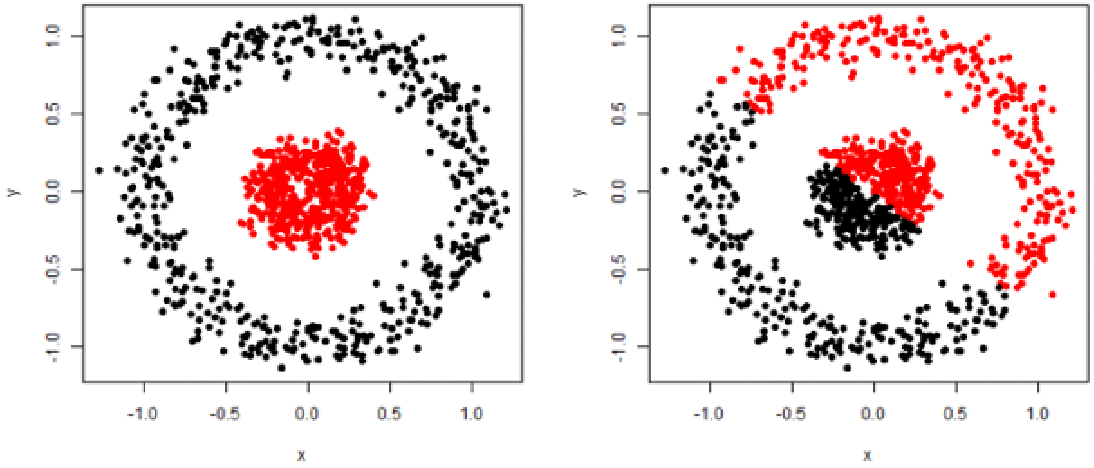
Partitioning clustering

Convex partitioning. Example: K-means

Density based approaches. Example: DBSCAN

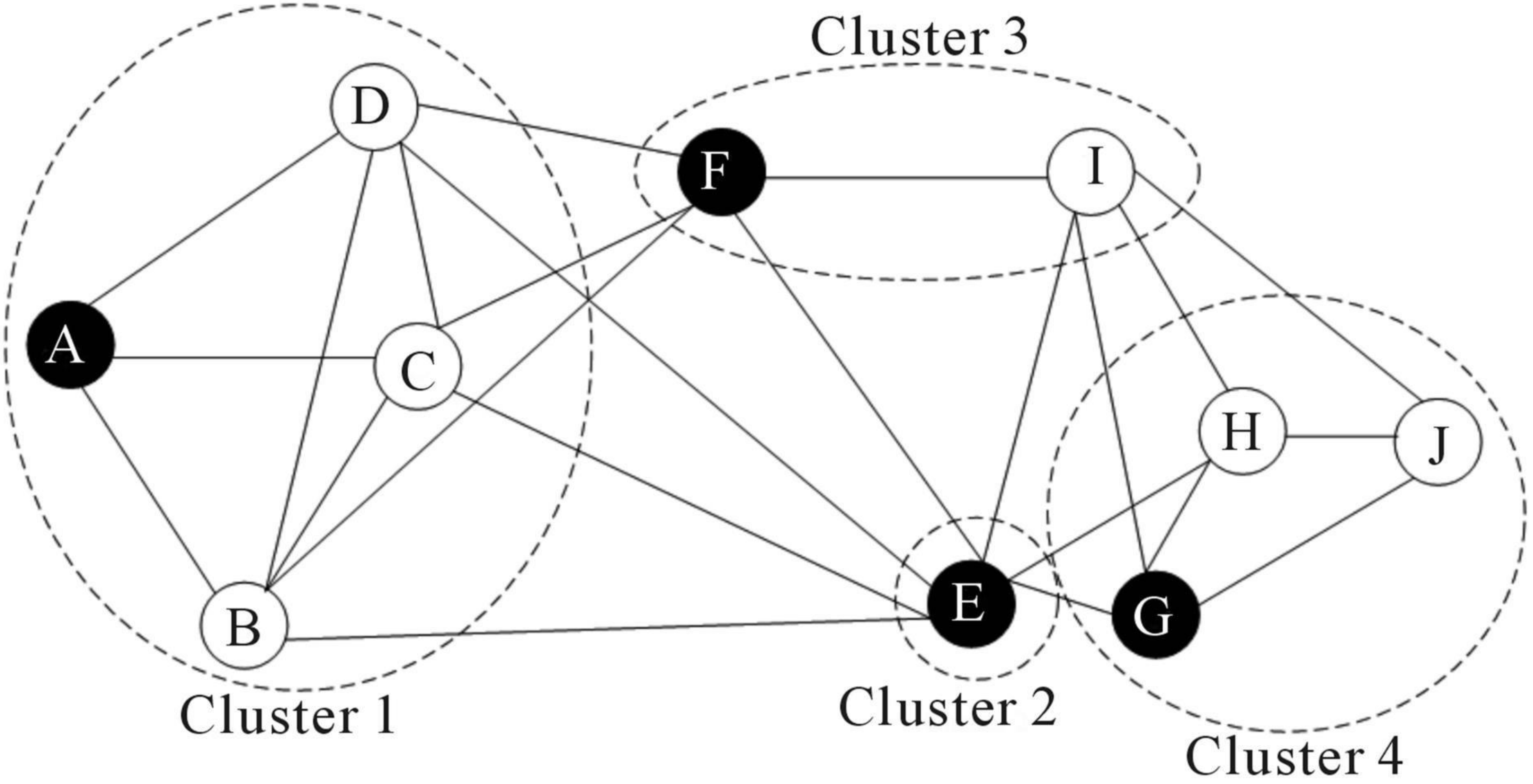
Model-based approaches. Example: Mclust

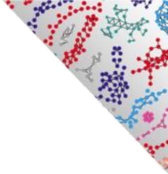
Graph based approaches : Example to follow



Graph-based

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





Graph-based: types

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

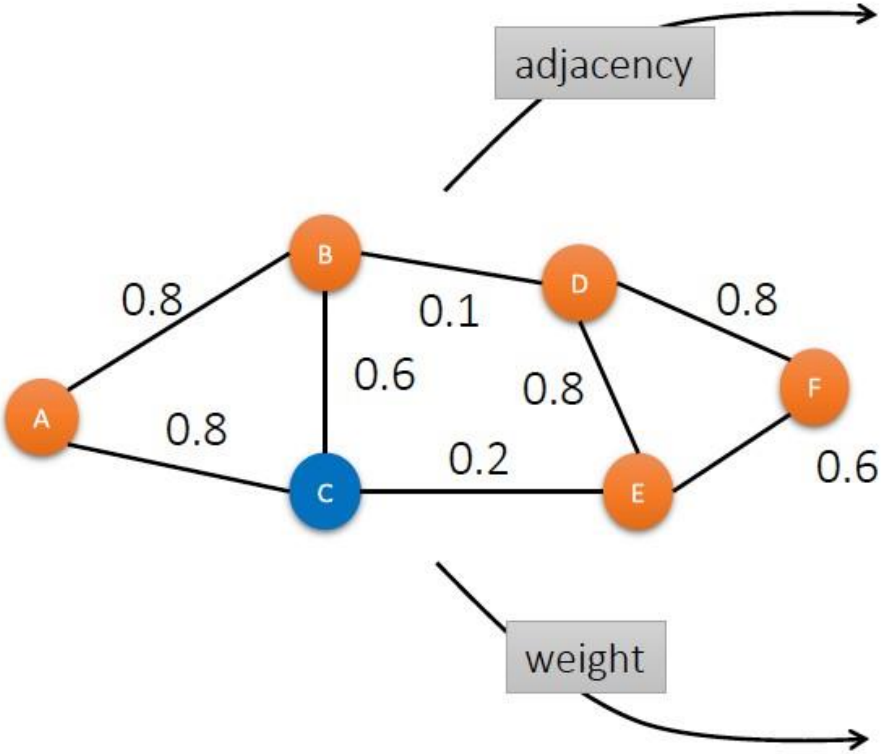
A graph in which two vertices *p* and *q* are connected by an edge, if the distance between *p* and *q* is among the *k*-th smallest distances from *p* to other objects from *P*.

- 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.

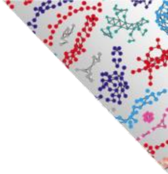
```
obj <- FindNeighbors(obj)
```

Graph-based: types



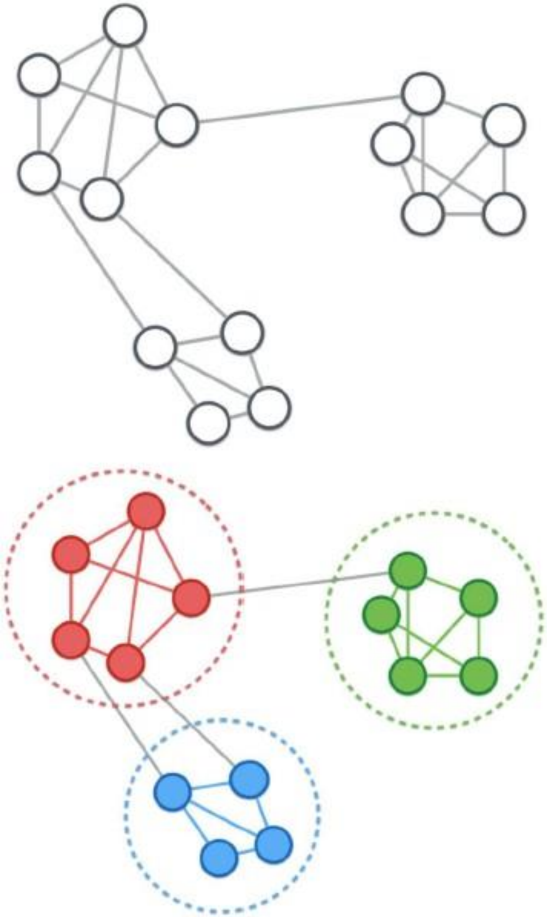
$$A = \begin{array}{c|cccccc|} & A & B & C & D & E & F \\ A & 0 & 1 & 1 & 0 & 0 & 0 \\ B & 1 & 0 & 1 & 1 & 0 & 0 \\ C & 1 & 1 & 0 & 0 & 1 & 0 \\ D & 0 & 1 & 0 & 0 & 1 & 1 \\ E & 0 & 0 & 1 & 1 & 0 & 1 \\ F & 0 & 0 & 0 & 1 & 1 & 0 \end{array}$$

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



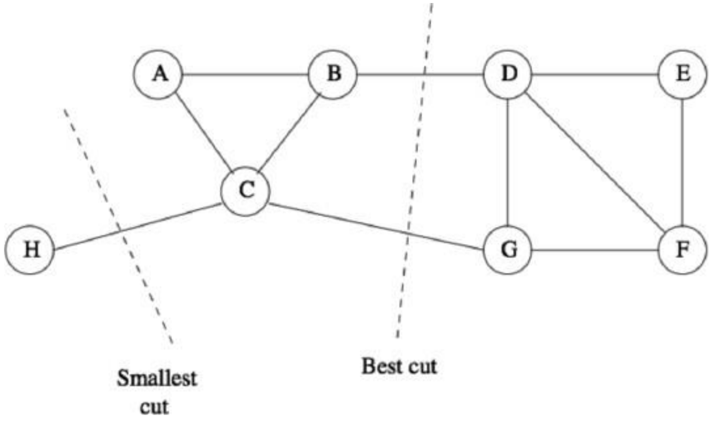
Graph-based: communities

- 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.

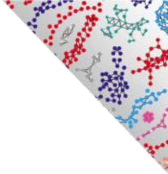


Graph-based: Cuts

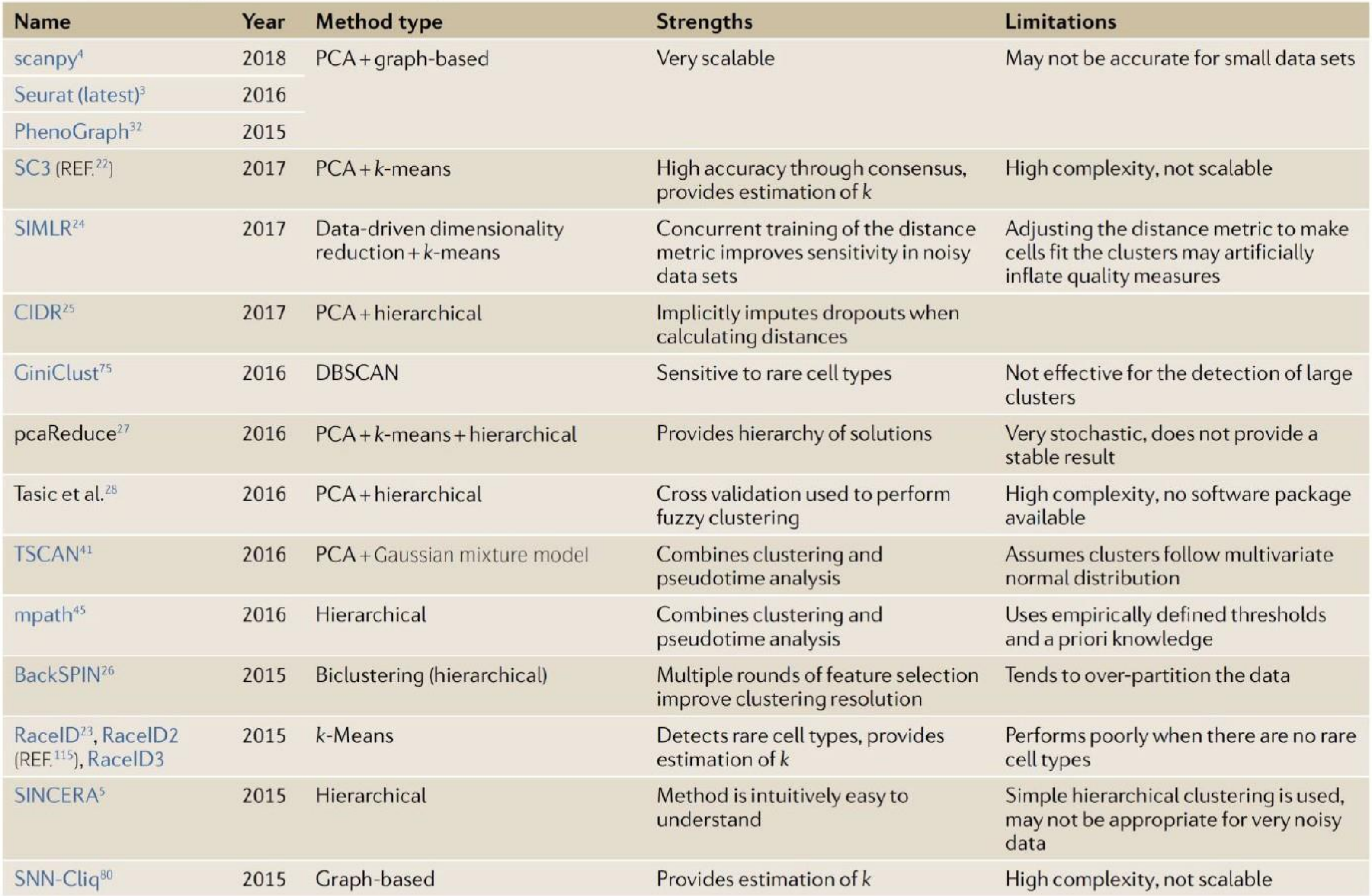
- Graph cut partitions a graph into subgraphs
- Cut size is the number of cut edges
- Clustering by graph cuts: find the smallest cut that bi-partitions the graph
- The smallest cut is not always the best cut
- NP-hard
  - Heuristic methods applied e.g. Louvain



*Shi & Malik (IEEE PAMI 2000)



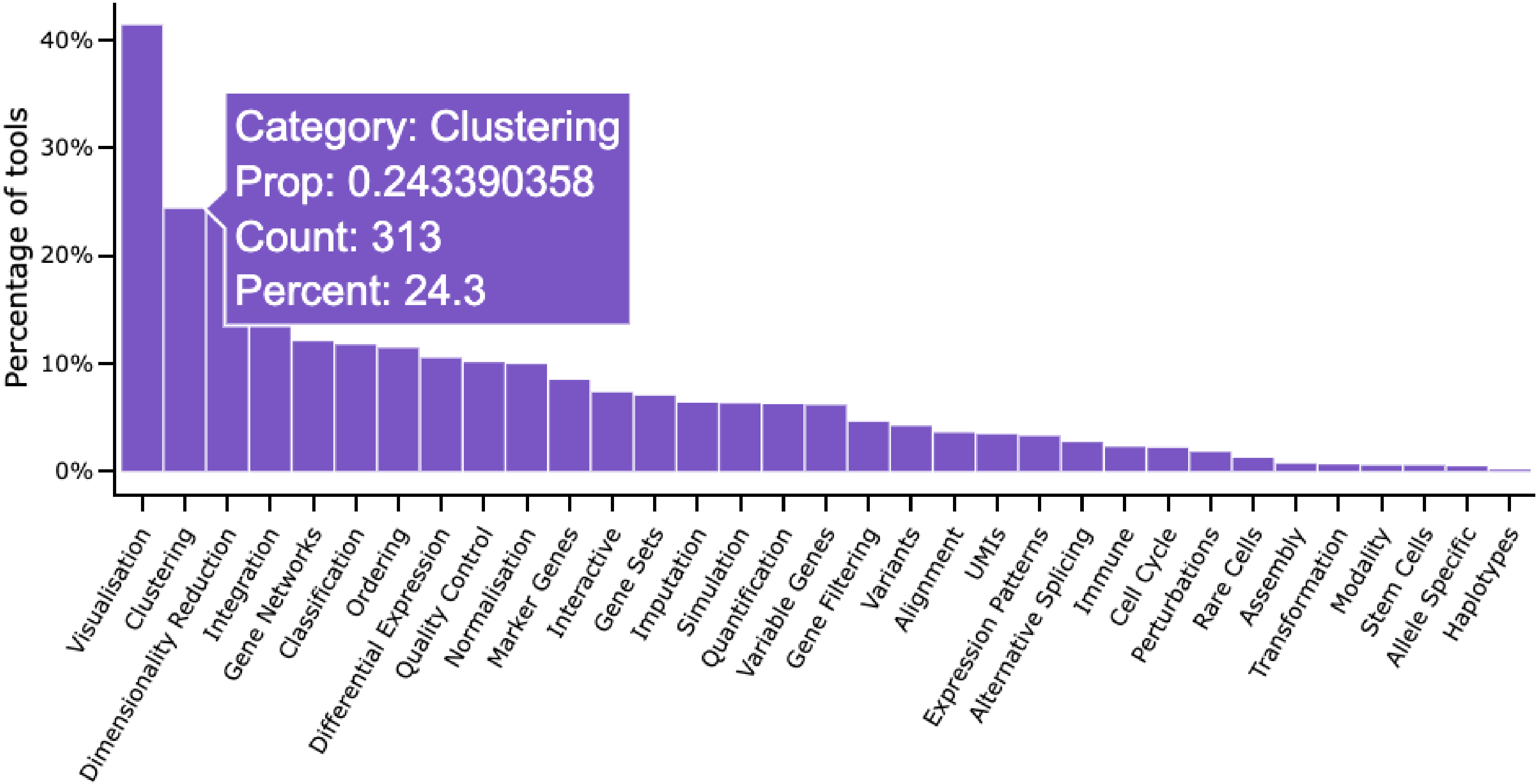
scRNA-seq clustering methods



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

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

scRNA-seq clustering methods



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

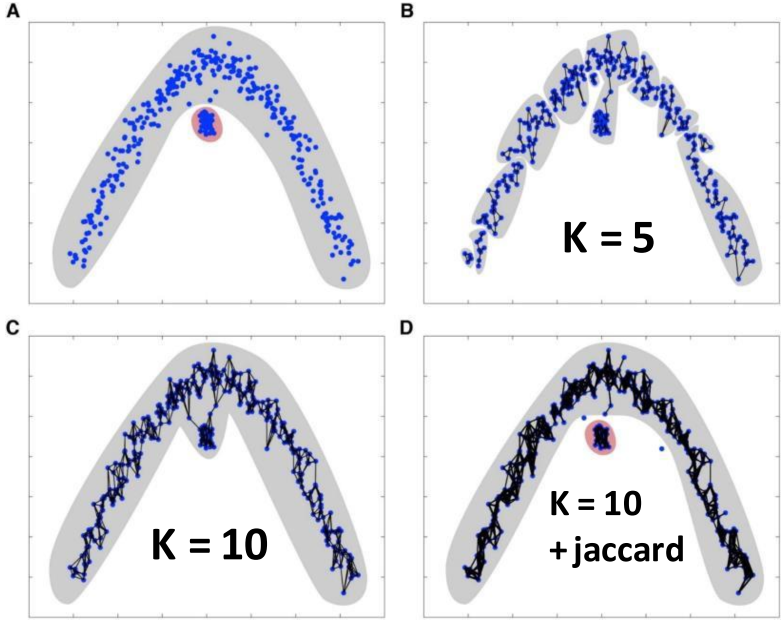
Seurat

1. Construct SNN graph based on the Euclidean distance in PCA space.
(Default, but could be also kNN)

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

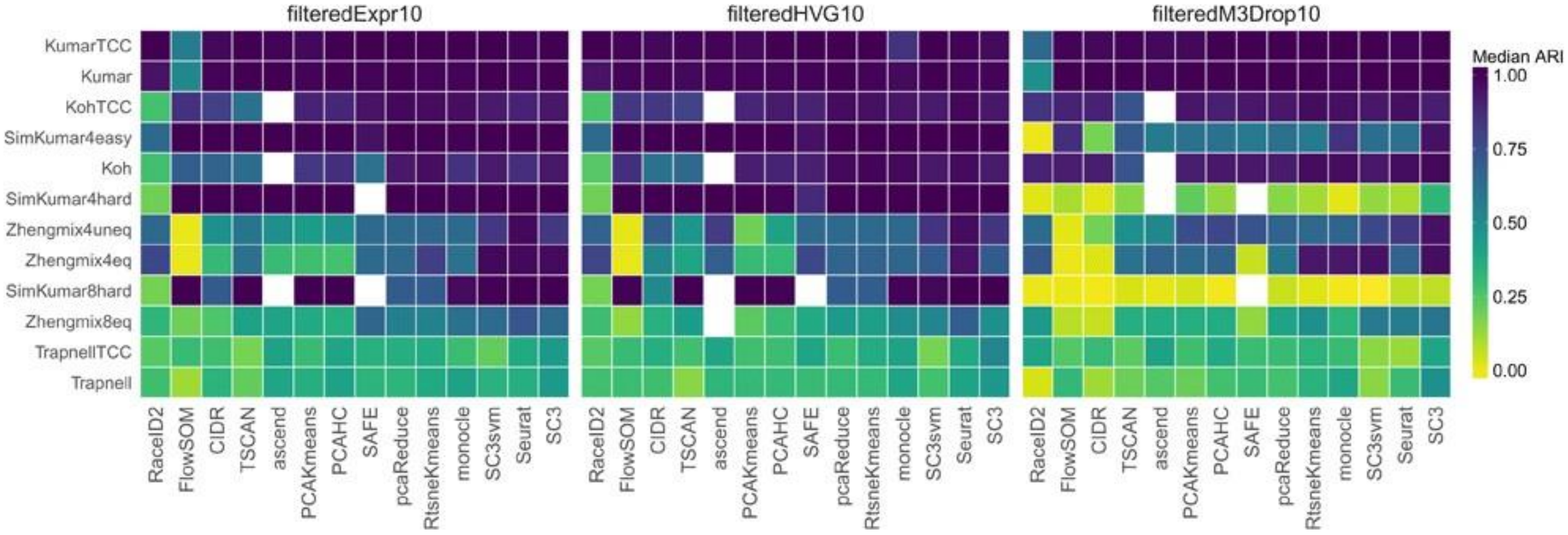
3. Cluster cells by optimizing for modularity (cuts) (Louvain algorithm)
(Modularity is a cost function, resolution is a parameter used to calculate the modularity)

```
obj <- FindClusters(obj)
```

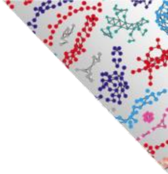


Xu and Su (https://doi.org/10.1093/bioinformatics/btv088)
Levine et al. (https://doi.org/10.1016/j.cell.2015.05.047)

Benchmarking



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



Clustering: Challenges

- What is a cell type?
- What is the number of clusters *k*?
- Bootstrapping
- Check QC after clustering to see if no biases are constituting your clusters
- Clustering is subjective – No ground truth
- How stable are the clusters
- How dependent are the clusters on the surrounding cells
- Scalability: in the last few years the number of cells in scRNA-seq experiments has grown by several orders of magnitude from $\sim 10^2$ to $\sim 10^6$