



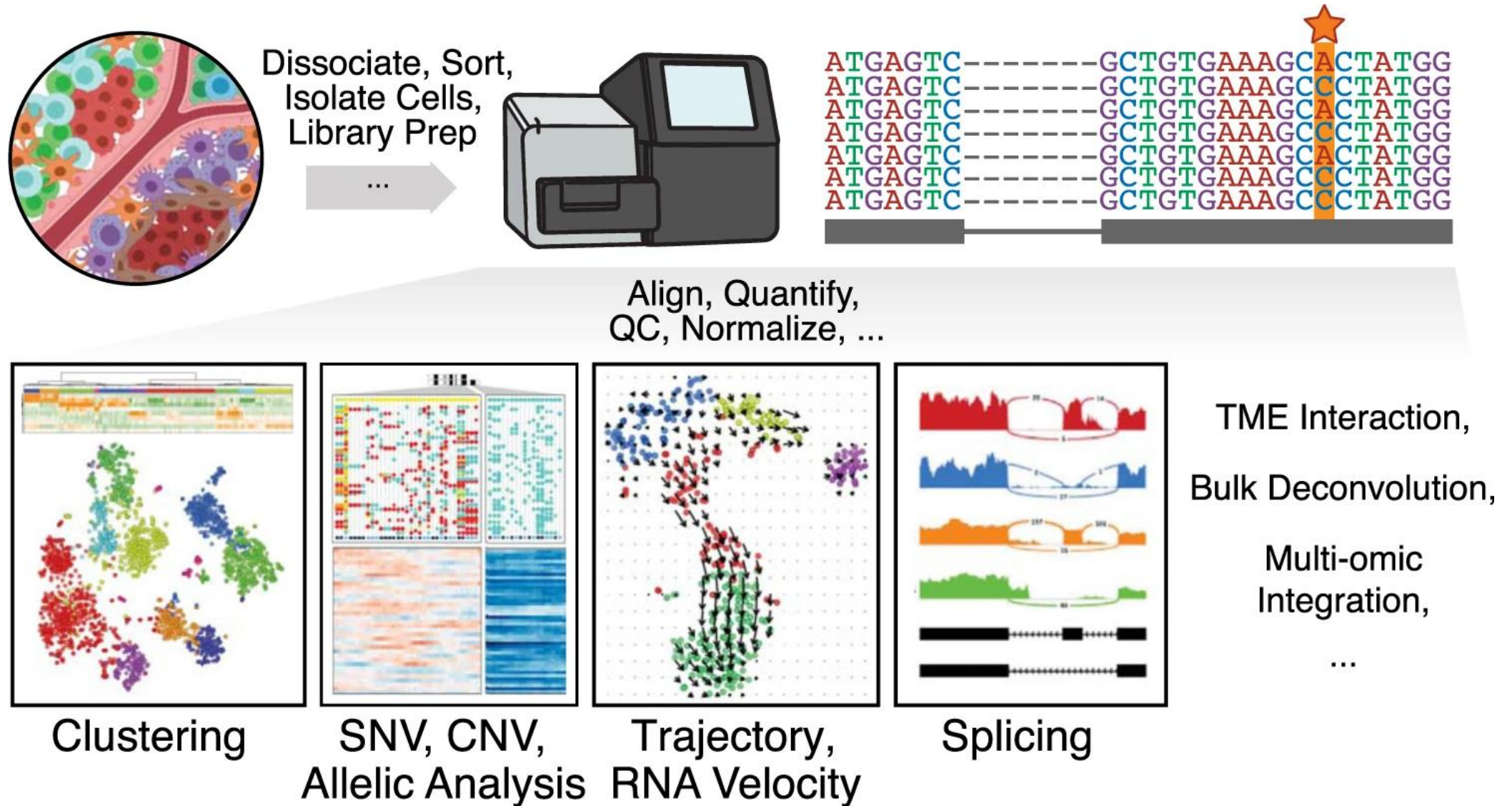
Swiss Institute of
Bioinformatics

Clustering

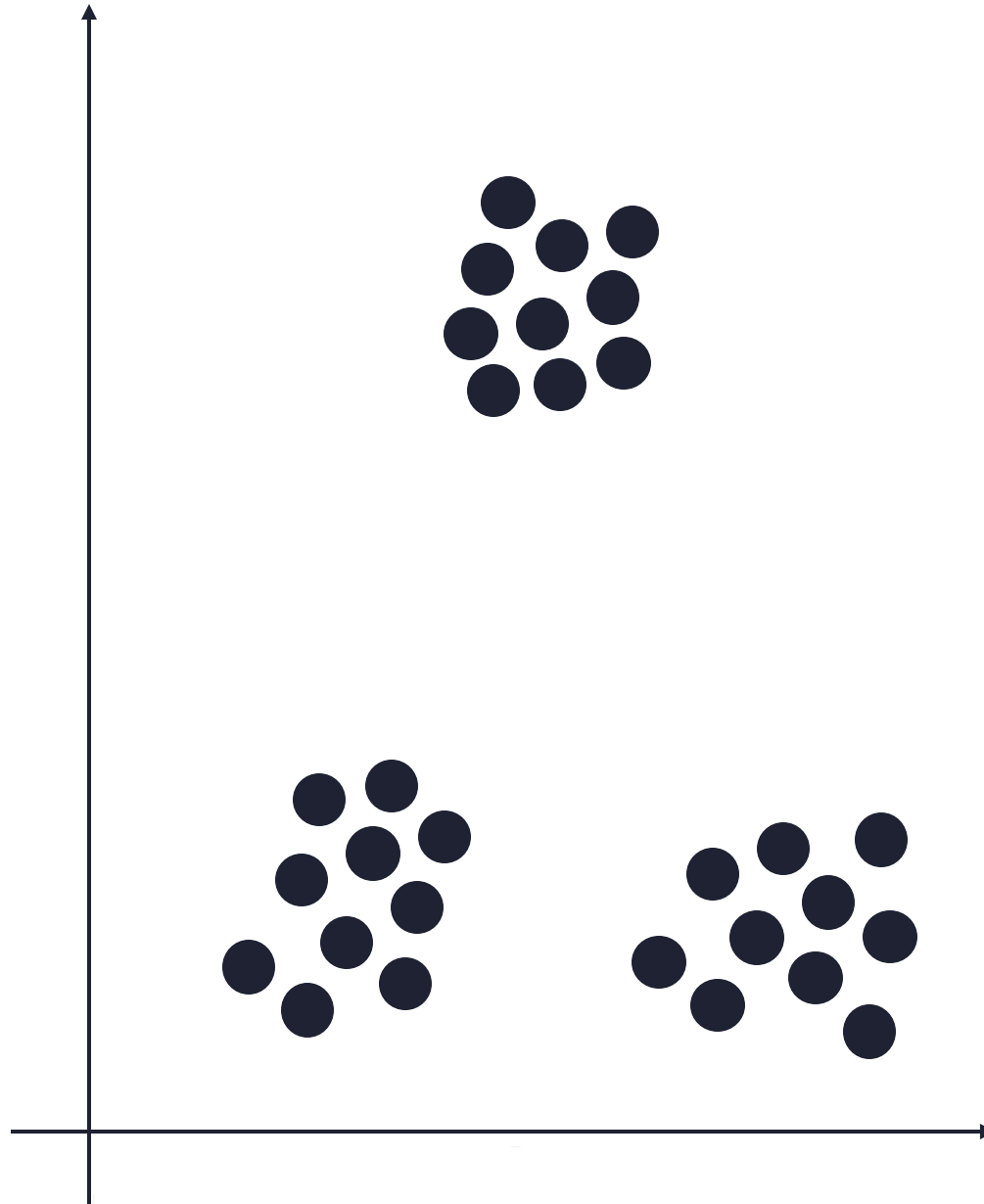
Luciano Cascione, PhD
Bioinformatics Core Unit

LUCIANO CASCIONE, PHD
BELLINZONA, OCT. 30TH 2024

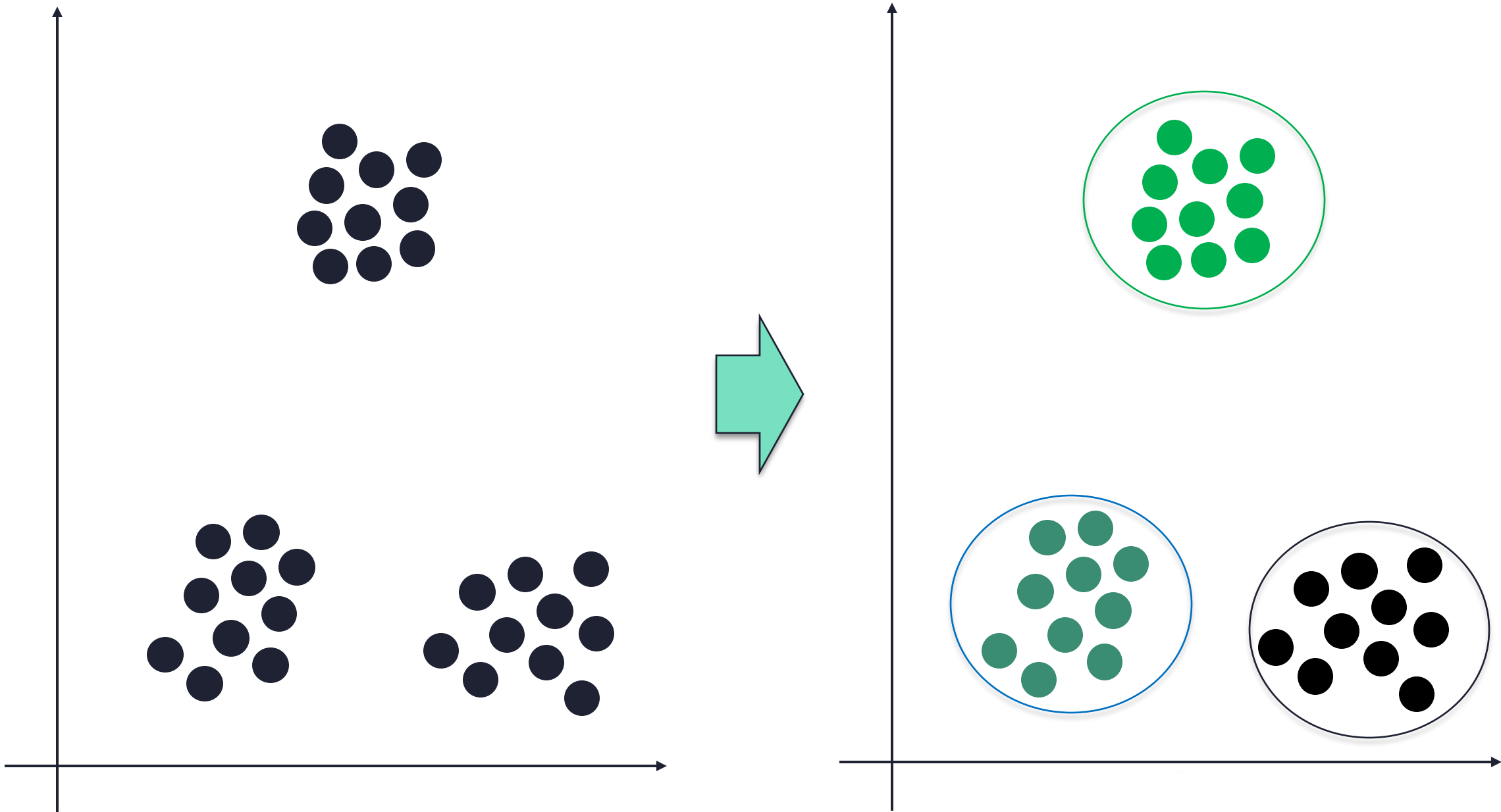
Singlecell RNA-Seq workflow



Clustering

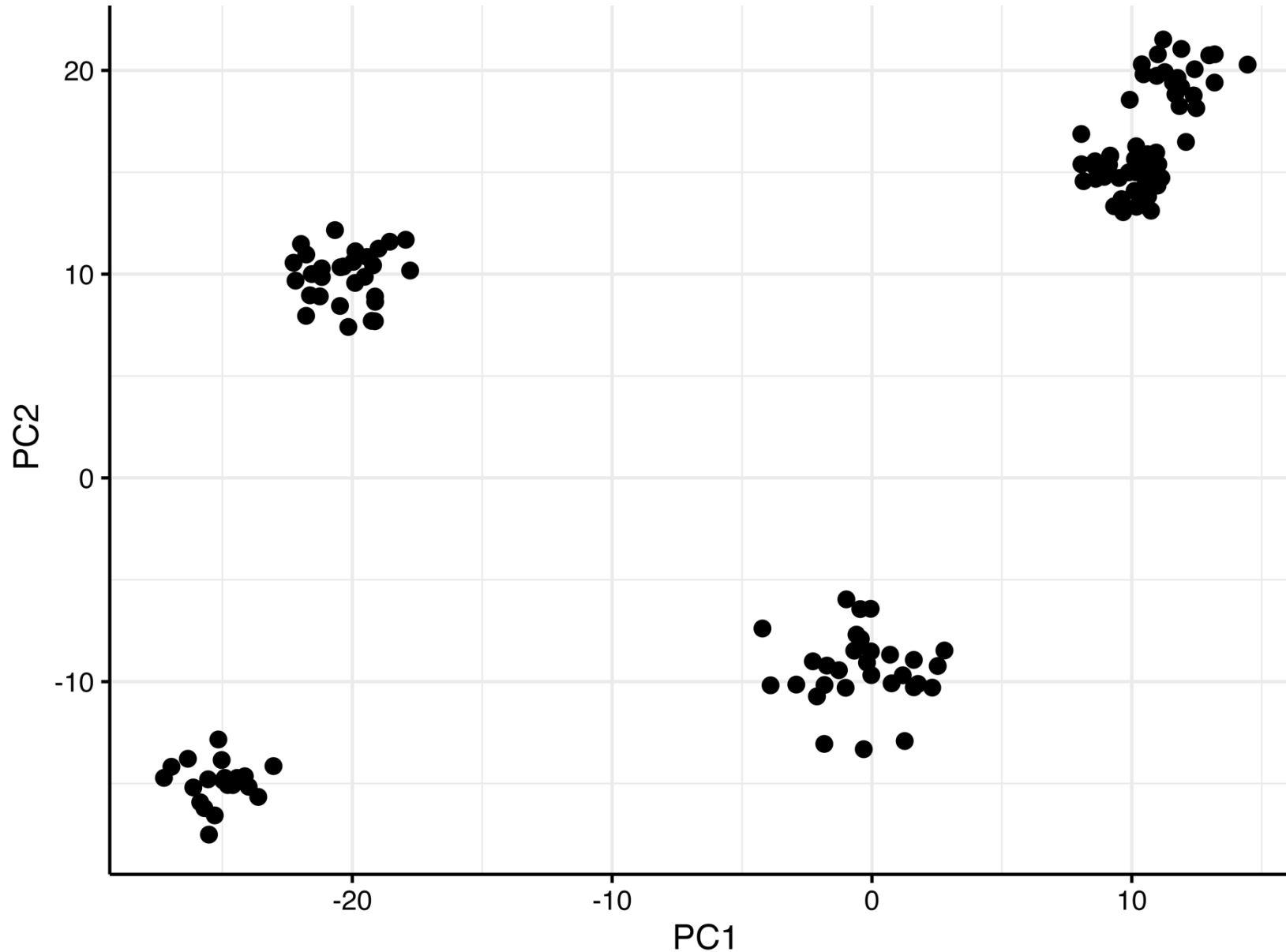


Clustering : discovering the natural groupings of a set of objects



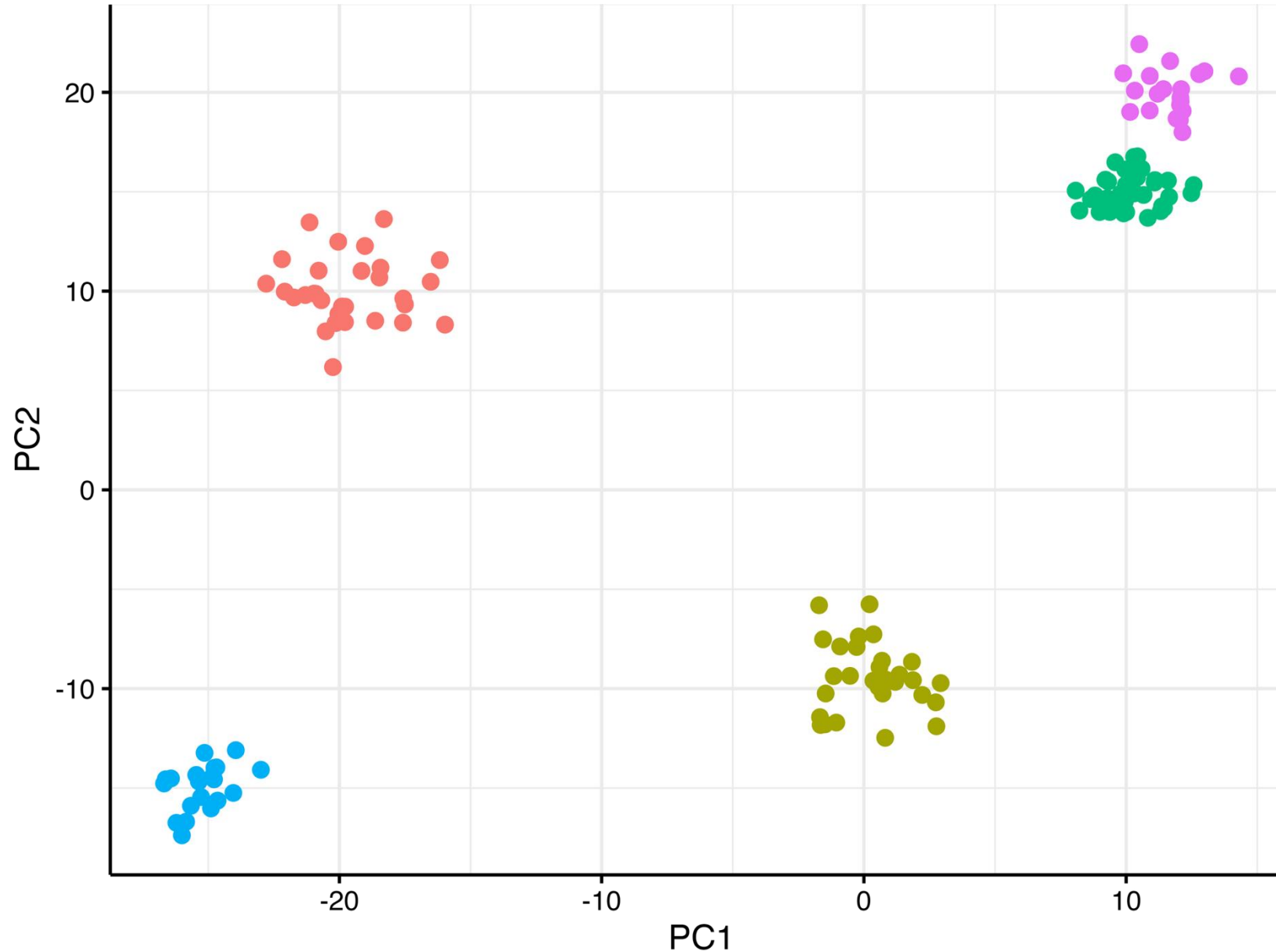
Clustering scRNA-Seq data

The goal of clustering is to group cells with similar gene expression profiles



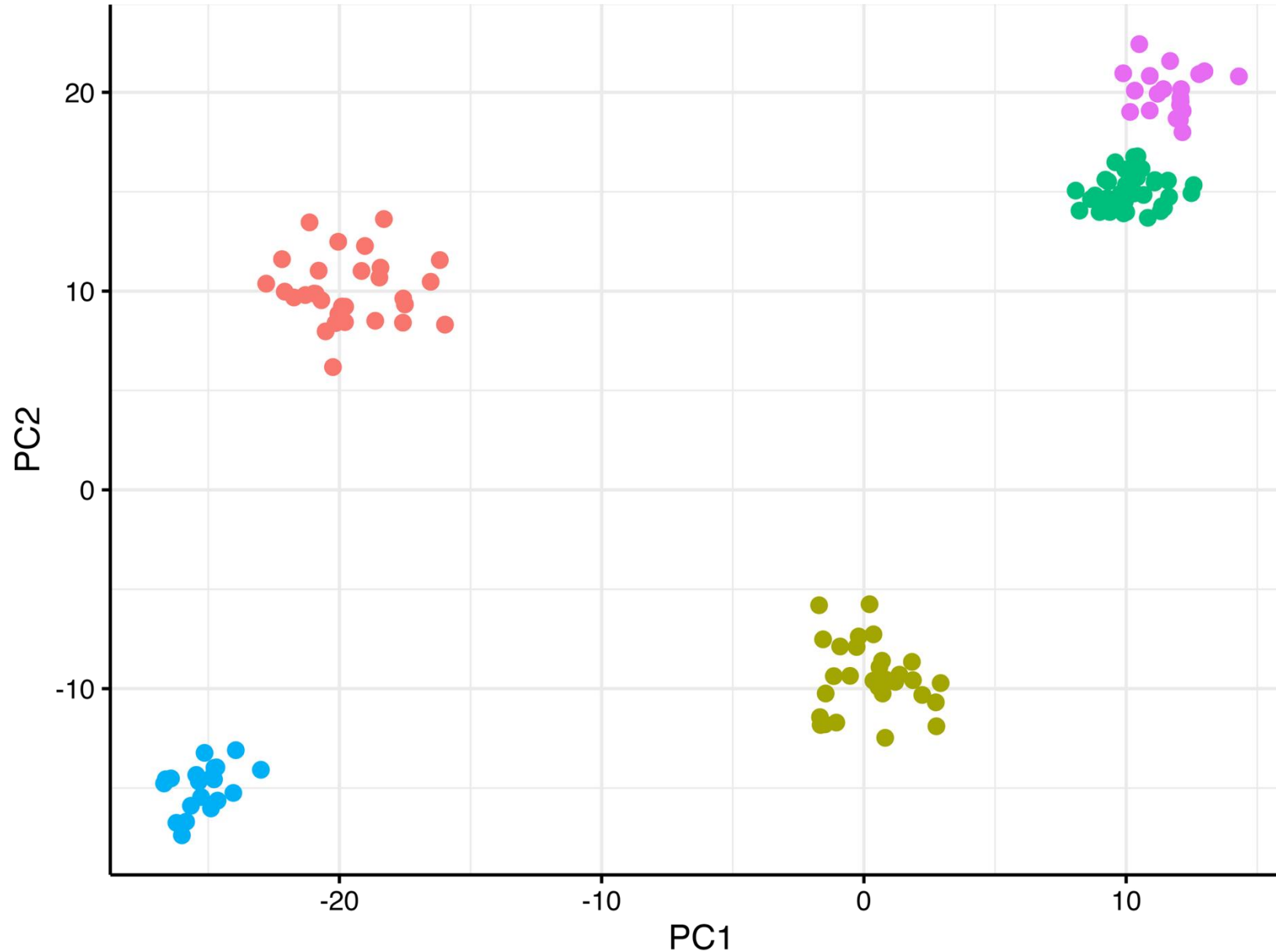
Clustering scRNA-Seq data

The goal of clustering is to group cells with similar gene expression profiles

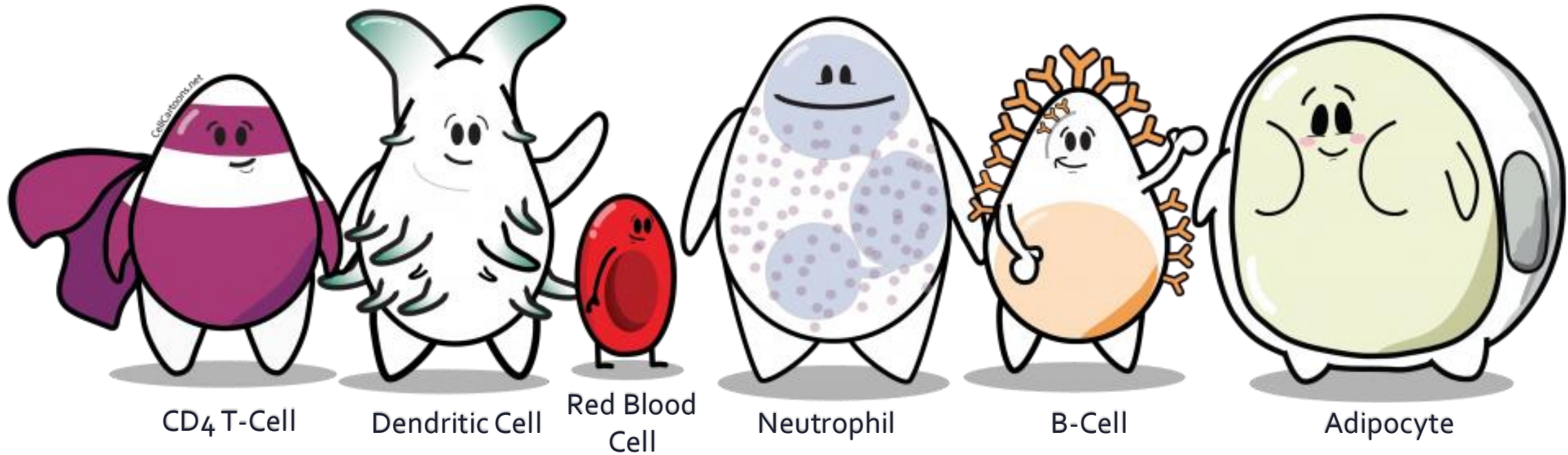


Clustering scRNA-Seq data

The goal of clustering is to group cells with **similar gene expression profiles**



Understanding cell types and states



Gene Expression Profile similarity

My gene expression is
like a one-hit wonder:
hemoglobin,
hemoglobin,
hemoglobin



I have a whole playlist
of immune responses.
My genes are ready for
action



Our Input

Gene	Cell_1	Cell_2	...	Cell_N
Gene_1	25	8	4	14
Gene_2	27	61	32	28
Gene_3	16	0	0	12
Gene_4	1	1	1	100
Gene_5	0	15	31	78
...	...	...	...	...
Gene_M	14	41	87	16

Its dimensionality-reduced representation

Cell	PC_1	PC_2	...	PC_X
Cell_1	5	12	...	9
Cell_2	8	14	...	7
Cell_3	0	1	...	9
...	...	...	...	...
Cell_N	-15	-3	...	12

Similarity scores (e.g., using Euclidean distance) are calculated in the PCA-reduced expression space.

Clustering

There are several scRNA-Seq clustering methods, but they can all be grouped into one of the following classes, or a combination of two or more:

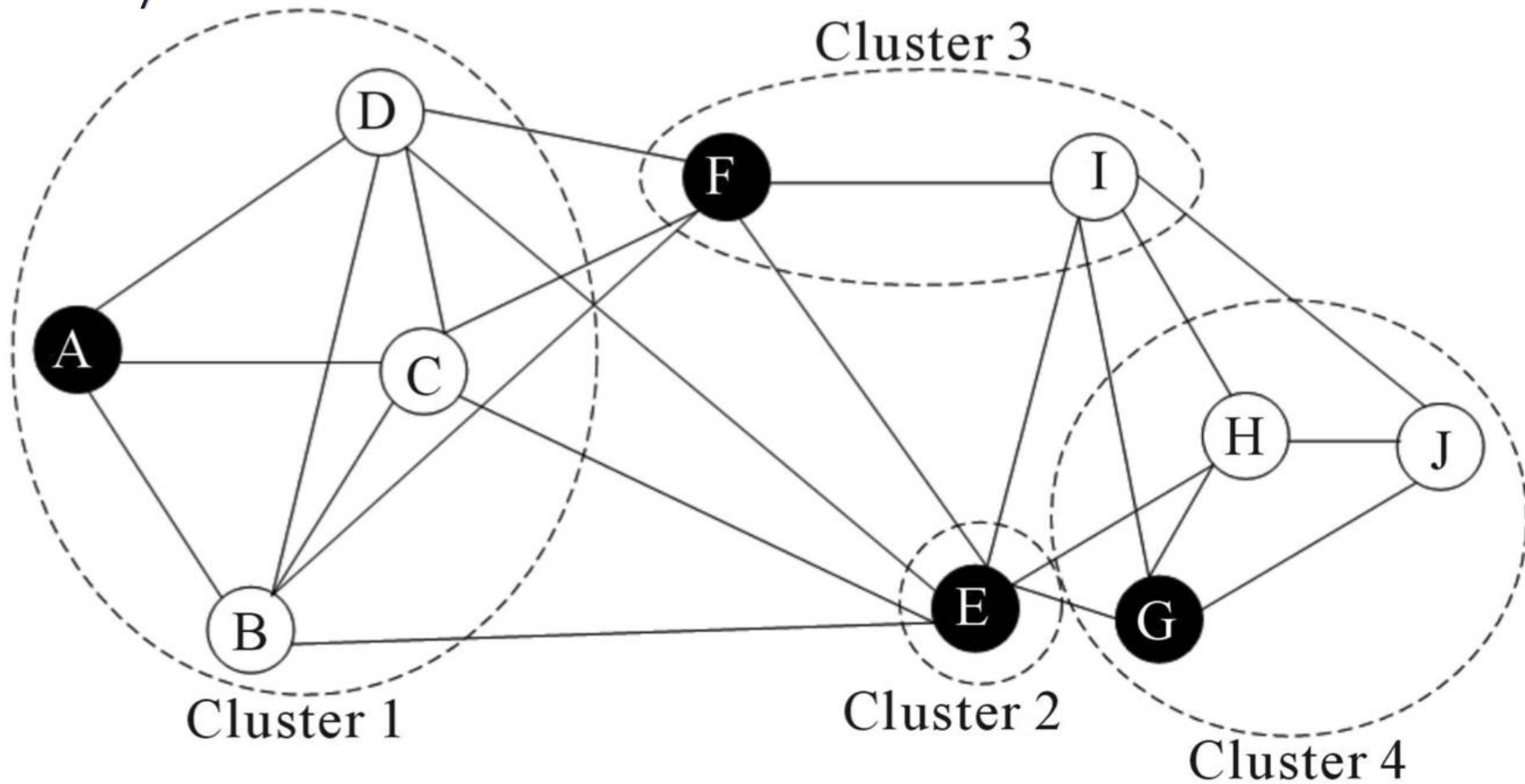
- Hierarchical
- K-means
- Graph-based
- Gaussian mixture
- Mean shift

excellent review: [Challenges in unsupervised clustering of single-cell RNA-seq data from Kiselev et al., \(2019\).](#)

Graph-based methods

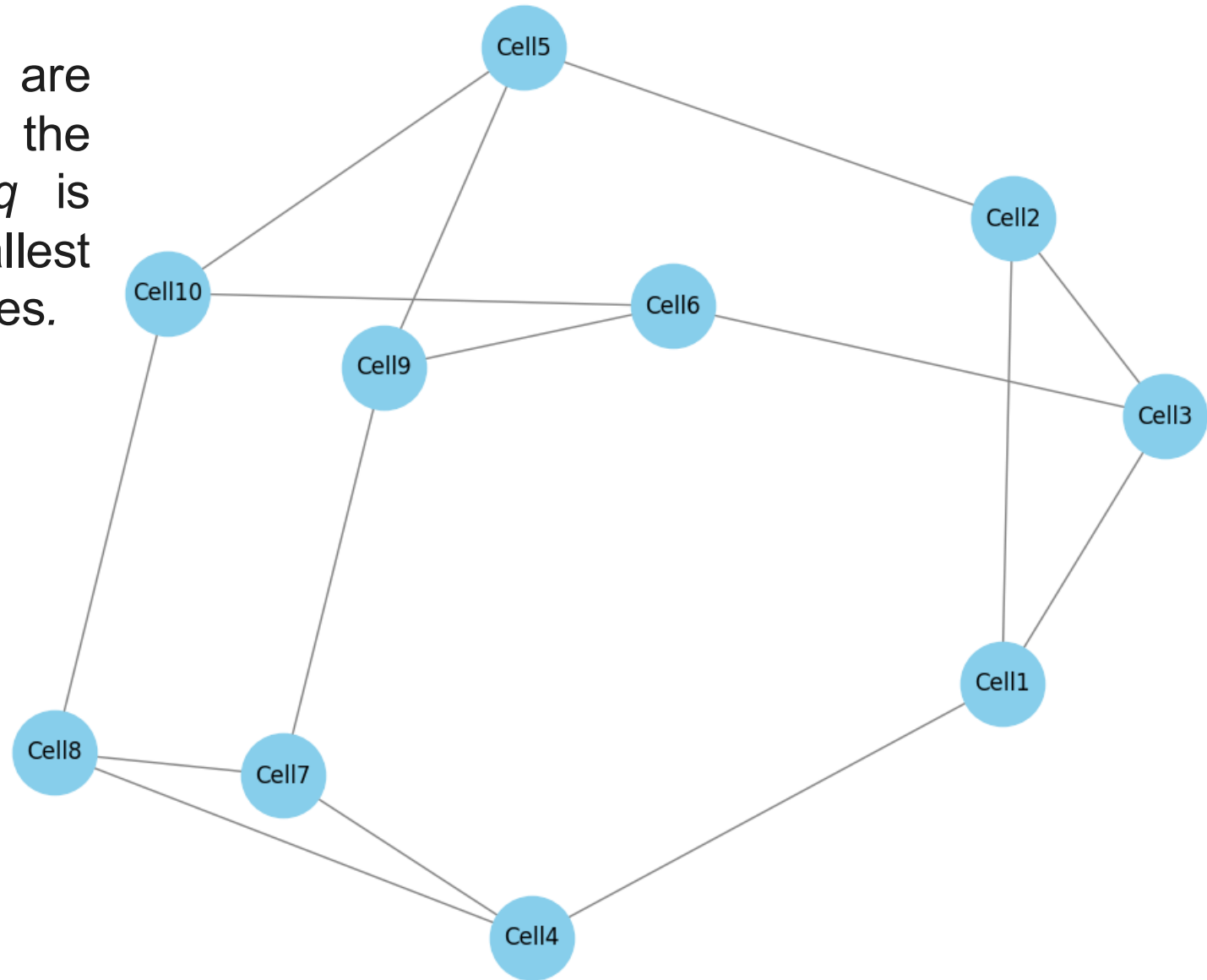
Nodes -> Cells

Edges -> Similarity



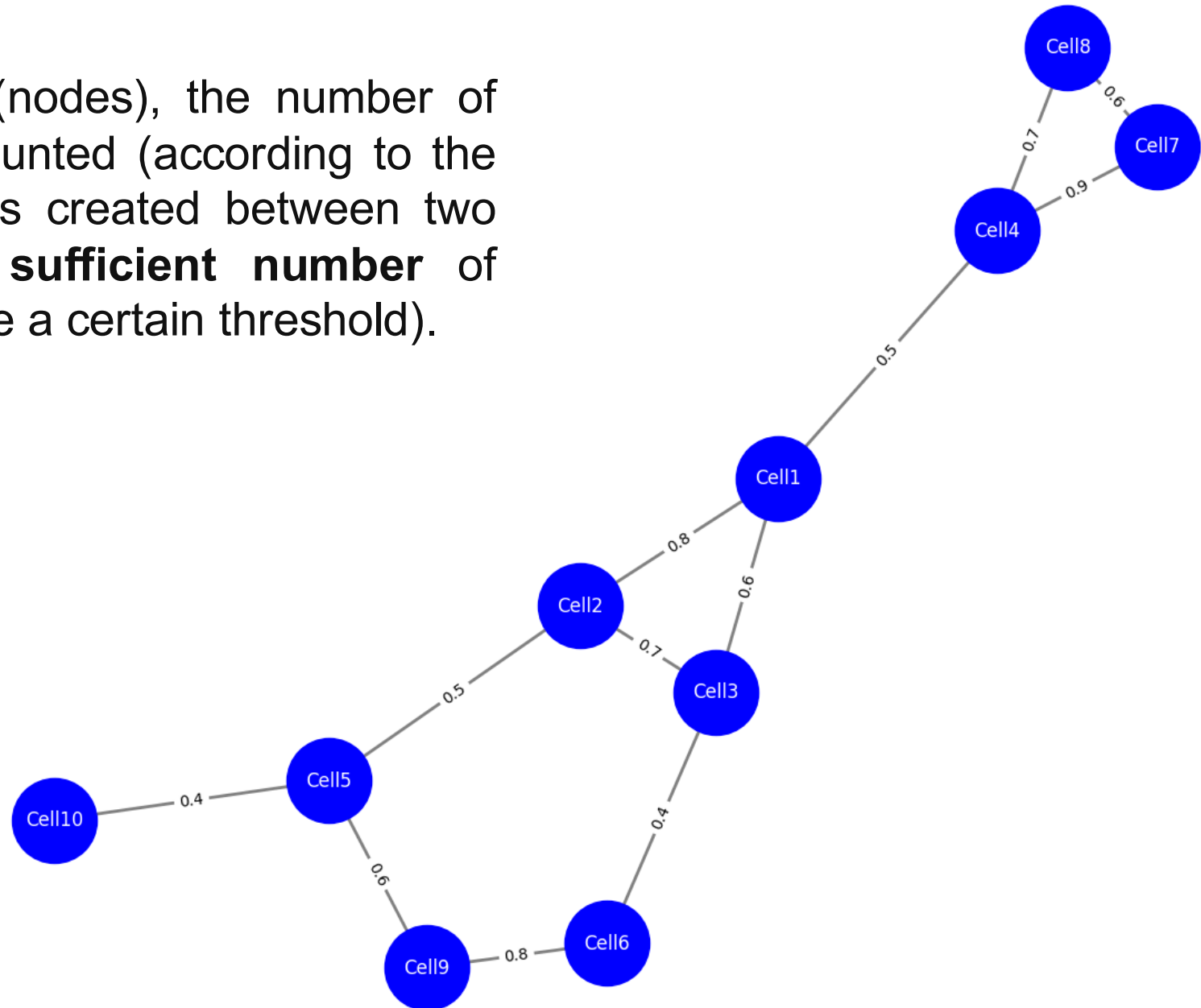
K-Nearest Neighbor

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 nodes.



Shared Nearest Neighbor

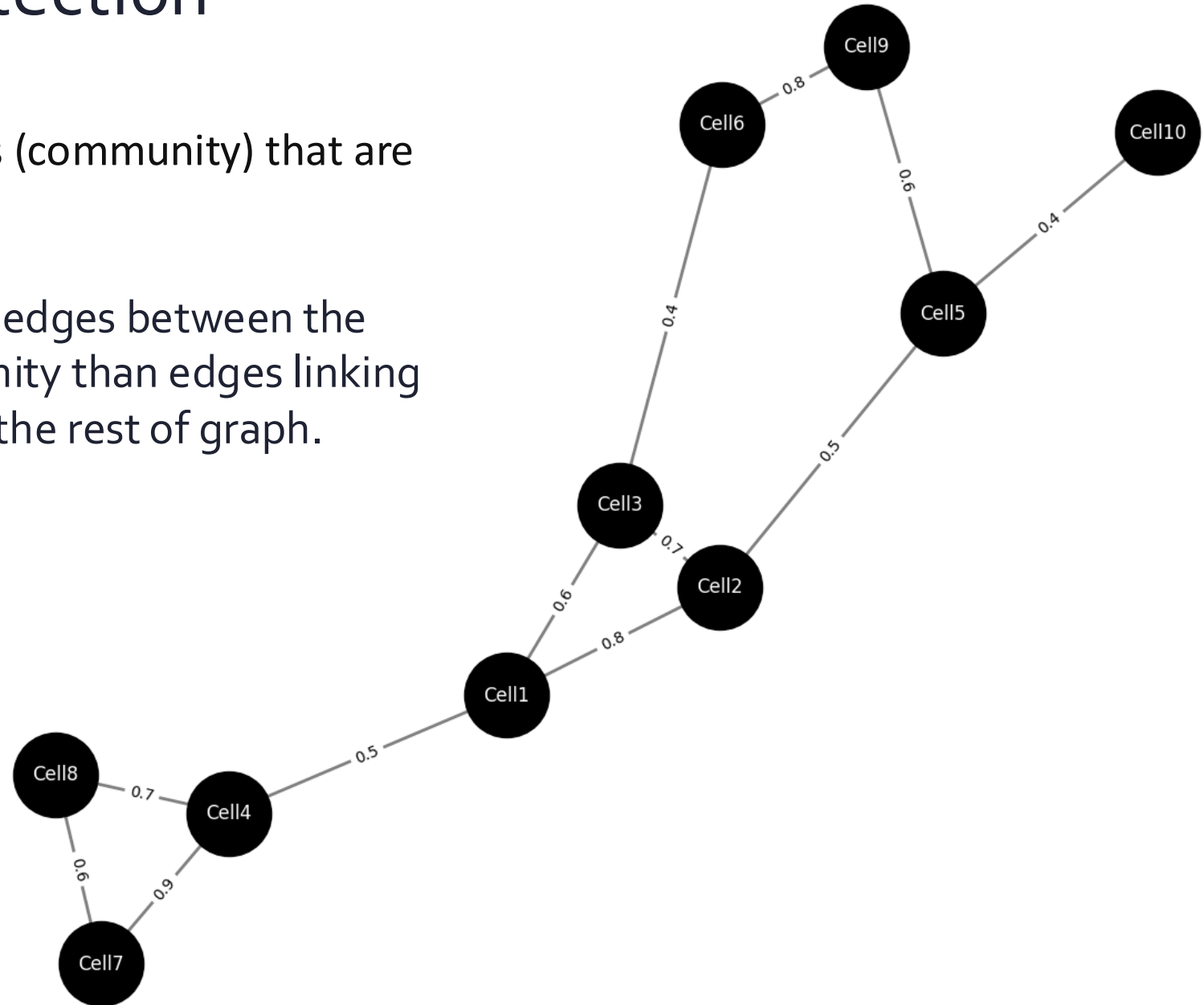
For each pair of cells (nodes), the number of shared neighbours is counted (according to the KNN graph). An edge is created between two cells if they share a **sufficient number** of nearest neighbors (above a certain threshold).



Community Detection

Identify clusters of nodes (community) that are densely connected.

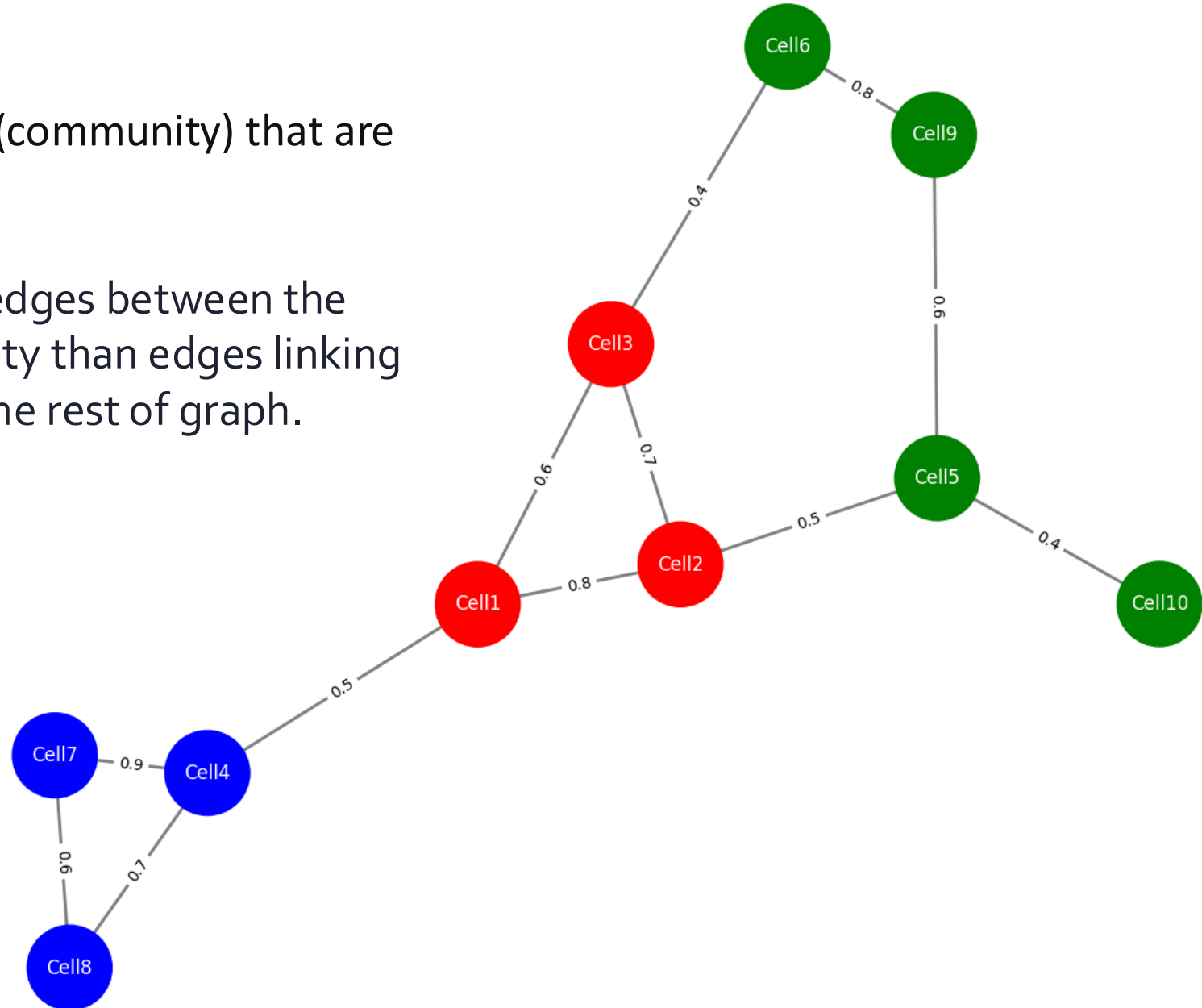
A community has more edges between the members of the community than edges linking nodes of the group with the rest of graph.



Community detection

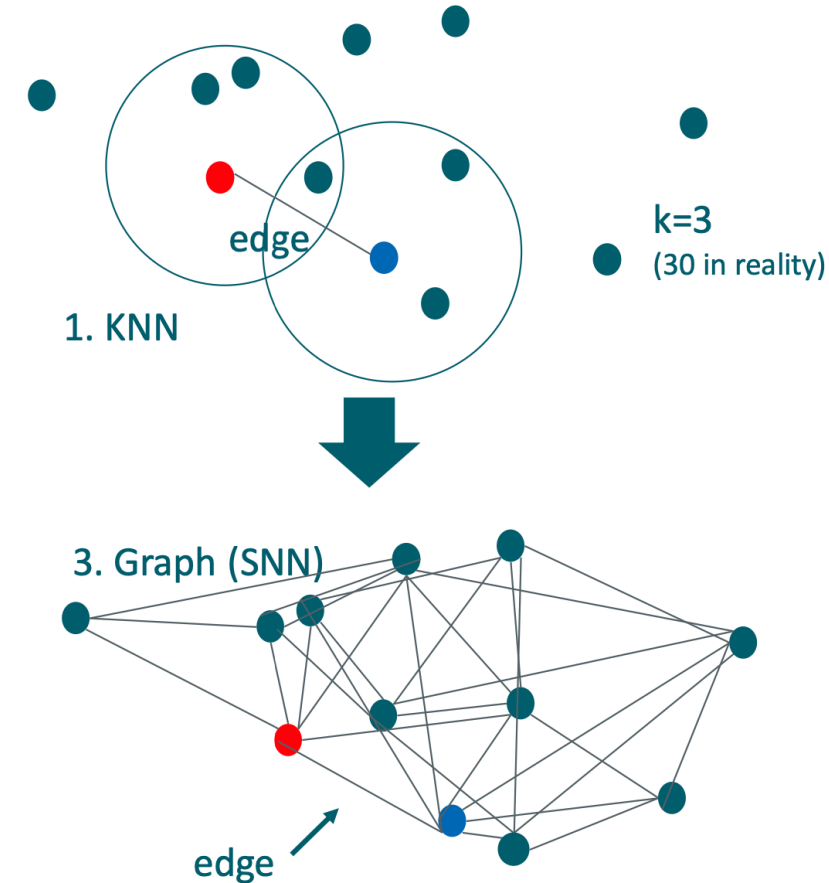
Identify clusters of nodes (community) that are densely connected.

A community has more edges between the members of the community than edges linking nodes of the group with the rest of graph.



Louvain Algorithm

1. Identify k nearest neighbours of each cell
2. Rank the neighbours based on distance
3. Build the graph: add an edge between cells if they have a shared nearest neighbour (SNN)
 - Give edge weights based on ranking
4. Cut the graph to subgraphs (clusters) by optimizing modularity:
 - Louvain algorithm by default



Clustering: Issues and obstacles

- 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
- 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 $\sim 10^6$

