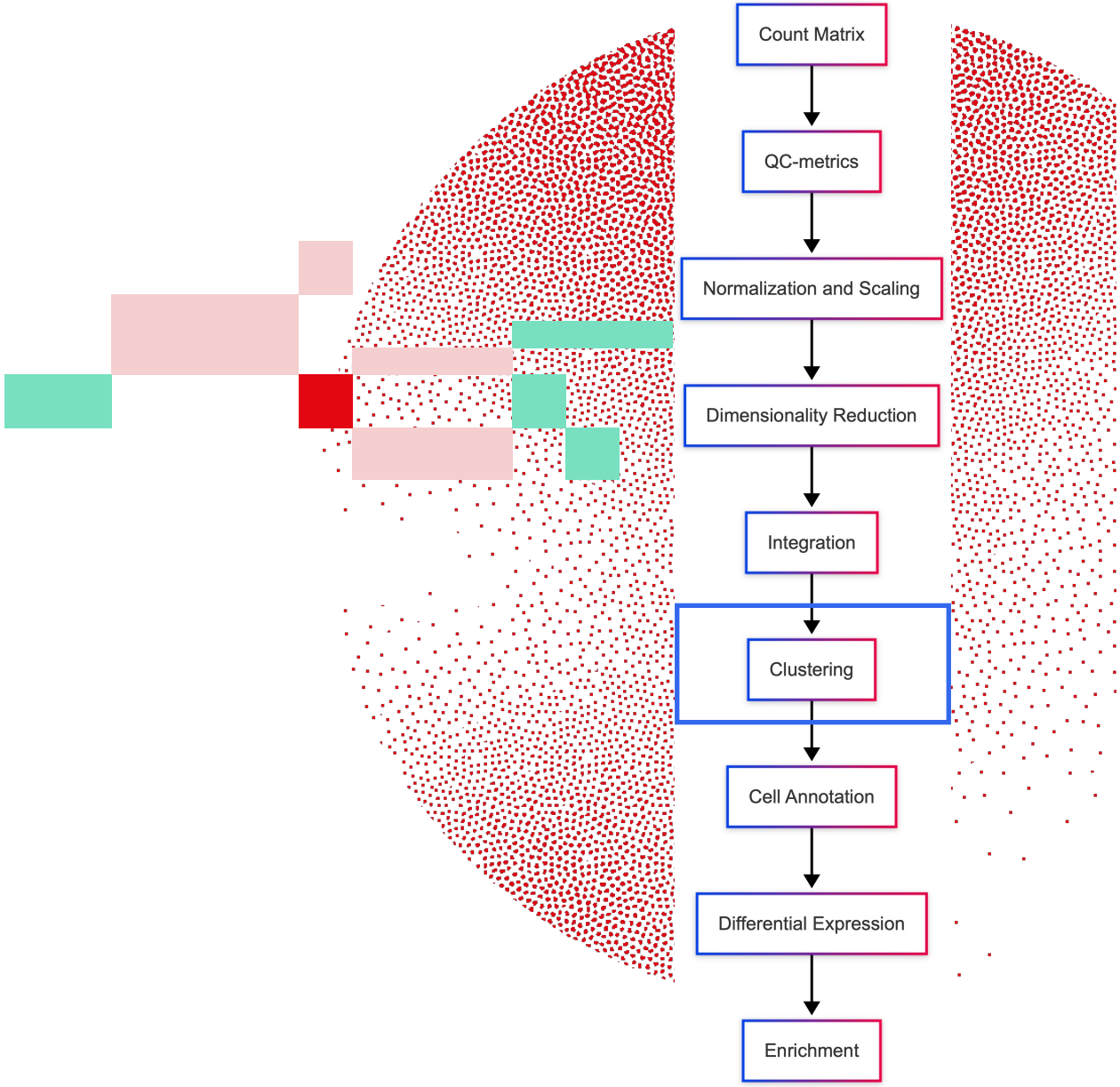
Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Clustering



Deepak Tanwar

March 18-20, 2025

Adapted from previous year courses

Feedback from Geert van Geest

elixir SWITZERLAND

Learning objectives

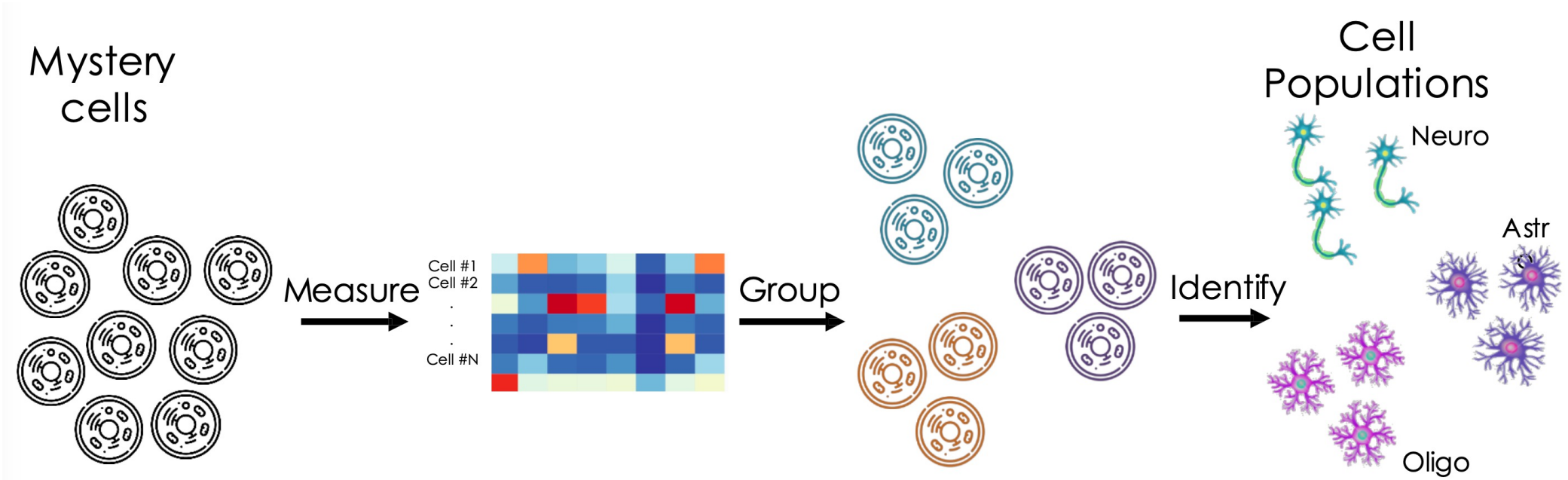
Understand the importance of clustering in single-cell RNA-seq data analysis

Identify different types of clustering methods and their applications

Apply graph-based clustering techniques, including the Louvain algorithm, to define cell populations

In single-cell RNA-seq, we use unsupervised clustering to empirically define groups of cells with similar expression profiles. Its primary purpose is to summarize complex data into a digestible format for human interpretation.

Clustering


Mystery cells
Measure
Cell #1
Cell #2
Cell #N
Group
Identify
Cell Populations
Neuro
Astr
Oligo

Why is clustering important?

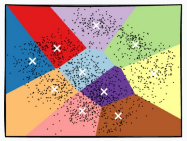
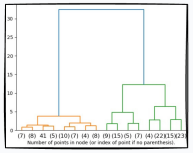
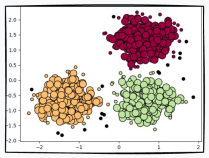
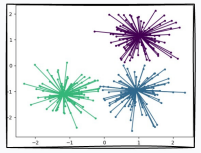
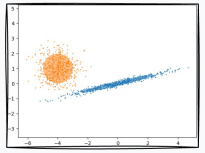
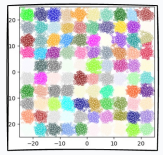
1. Identify cell populations
   Example: distinguish T cells, B cells, and macrophages in immune system studies

2. Detect novel cell types
   Example: identifying rare stem cell populations or new subtypes of neurons in brain tissue

3. Understand cellular heterogeneity
   Example: analyzing tumor samples to reveal subpopulations of malignant cells with distinct gene expression profiles

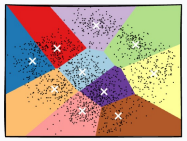
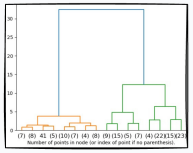
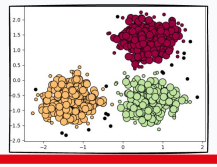
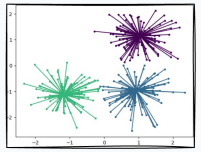
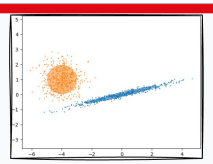
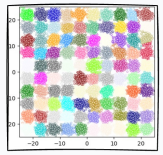
4. Track cell state transitions
   Example: Following the differentiation trajectory of stem cells into mature blood cells during hematopoiesis (blood cell production process)

Different types of clustering

Clustering Algorithm Type		Clustering Methodology	Algorithm(s)
	Centroid-based	Cluster points based on proximity to centroid	KMeans KMeans++ KMedoids
	Connectivity-based	Cluster points based on proximity between clusters	Hierarchical Clustering (Agglomerative and Divisive)
	Density-based	Cluster points based on their density instead of proximity	DBSCAN OPTICS HDBSCAN
	Graph-based	Cluster points based on graph distance	Affinitiy Propagation Spectral Clustering
	Distribution-based	Cluster points based on their likelihood of belonging to the same distribution.	Gaussian Mixture Models (GMMs)
	Compression-based	Transform data to a lower dimensional space and then perform clustering	BIRCH

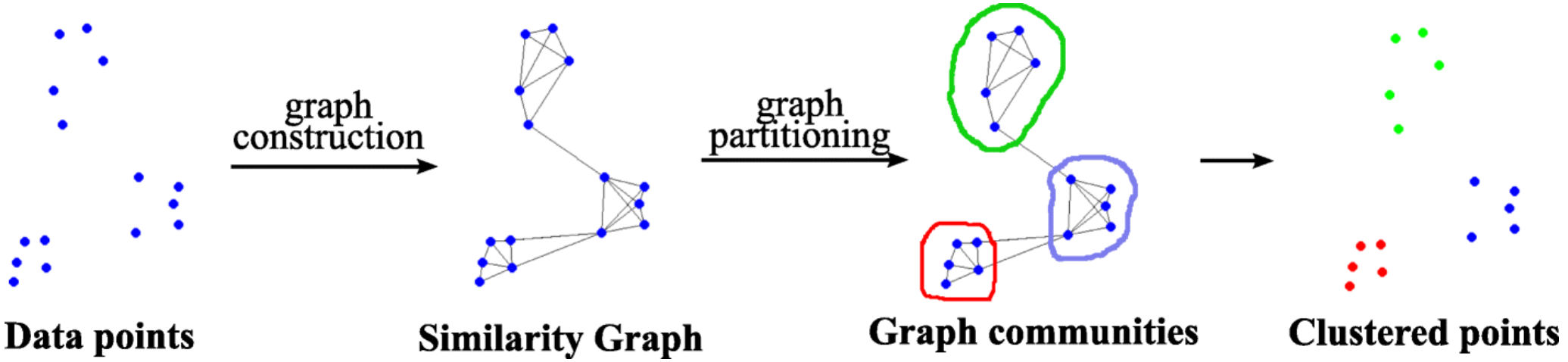
https://blog.dailydoseofds.com/p/beyond-kmeans-6-must-know-types-of

Different types of clustering

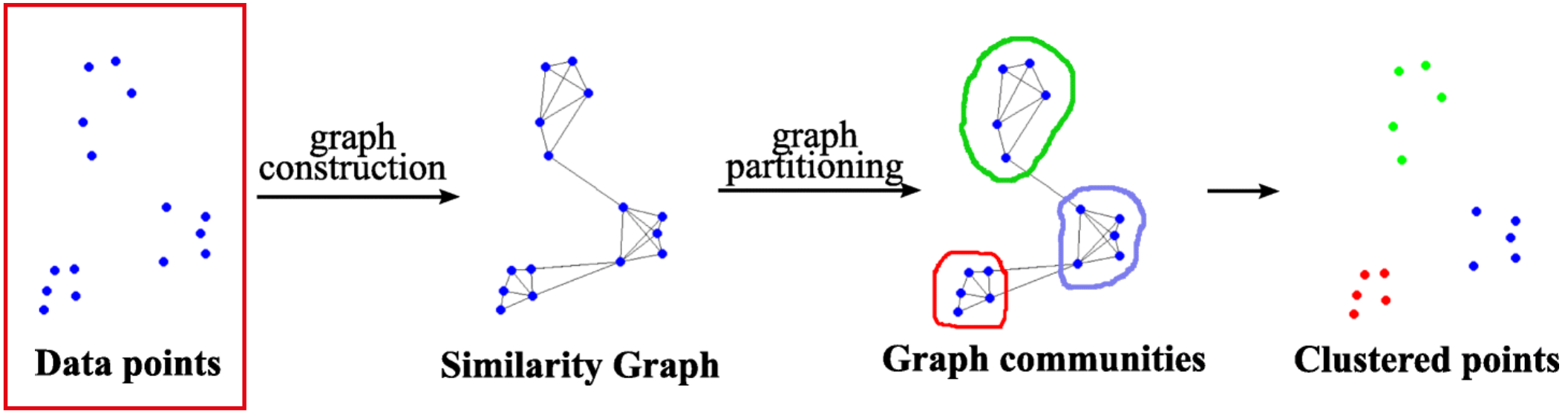
Clustering Algorithm Type		Clustering Methodology	Algorithm(s)
	Centroid-based	Cluster points based on proximity to centroid	KMeans KMeans++ KMedoids
	Connectivity-based	Cluster points based on proximity between clusters	Hierarchical Clustering (Agglomerative and Divisive)
	Density-based	Cluster points based on their density instead of proximity	DBSCAN OPTICS HDBSCAN
	Graph-based	Cluster points based on graph distance	Affinitiy Propagation Spectral Clustering
	Distribution-based	Cluster points based on their likelihood of belonging to the same distribution.	Gaussian Mixture Models (GMMs)
	Compression-based	Transform data to a lower dimensional space and then perform clustering	BIRCH

https://blog.dailydoseofds.com/p/beyond-kmeans-6-must-know-types-of

Graph-based clustering

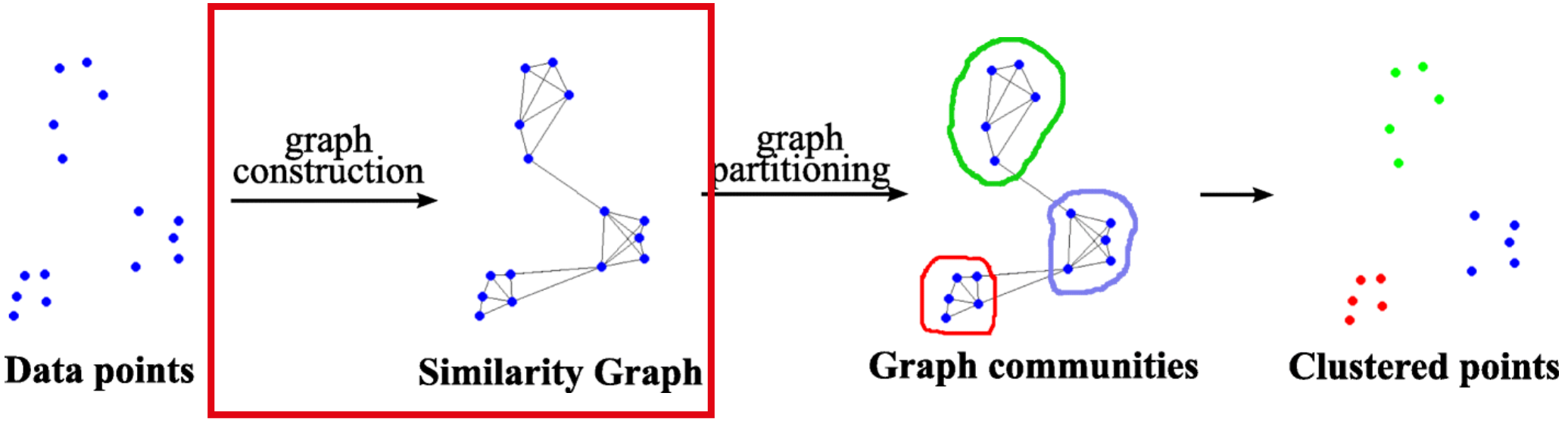


Graph-based clustering



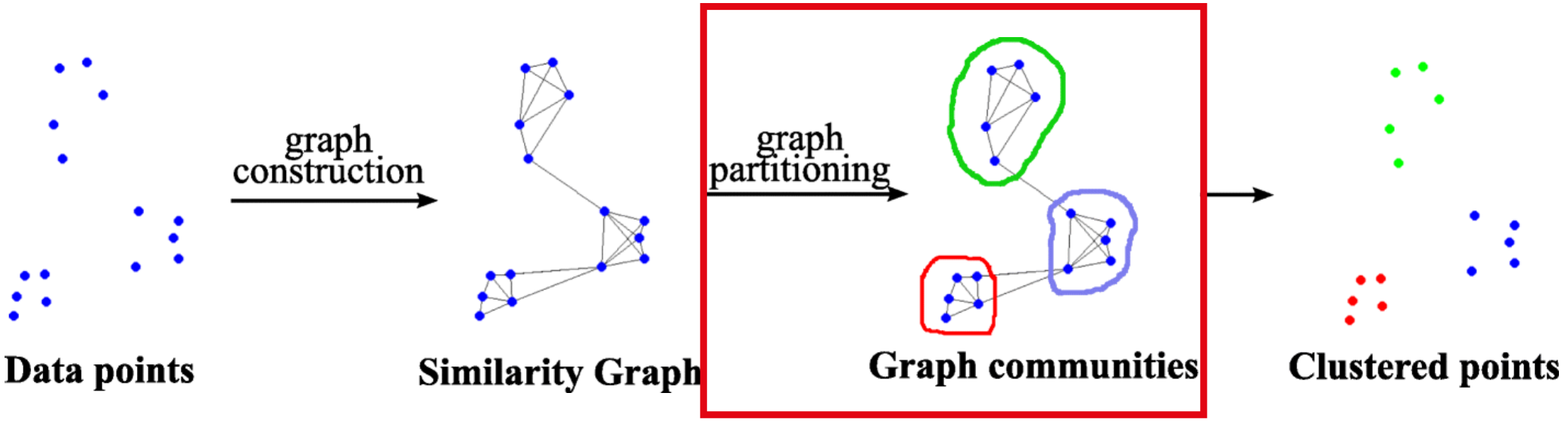
single cells after dimensionality reduction (like PCA)

Graph-based clustering



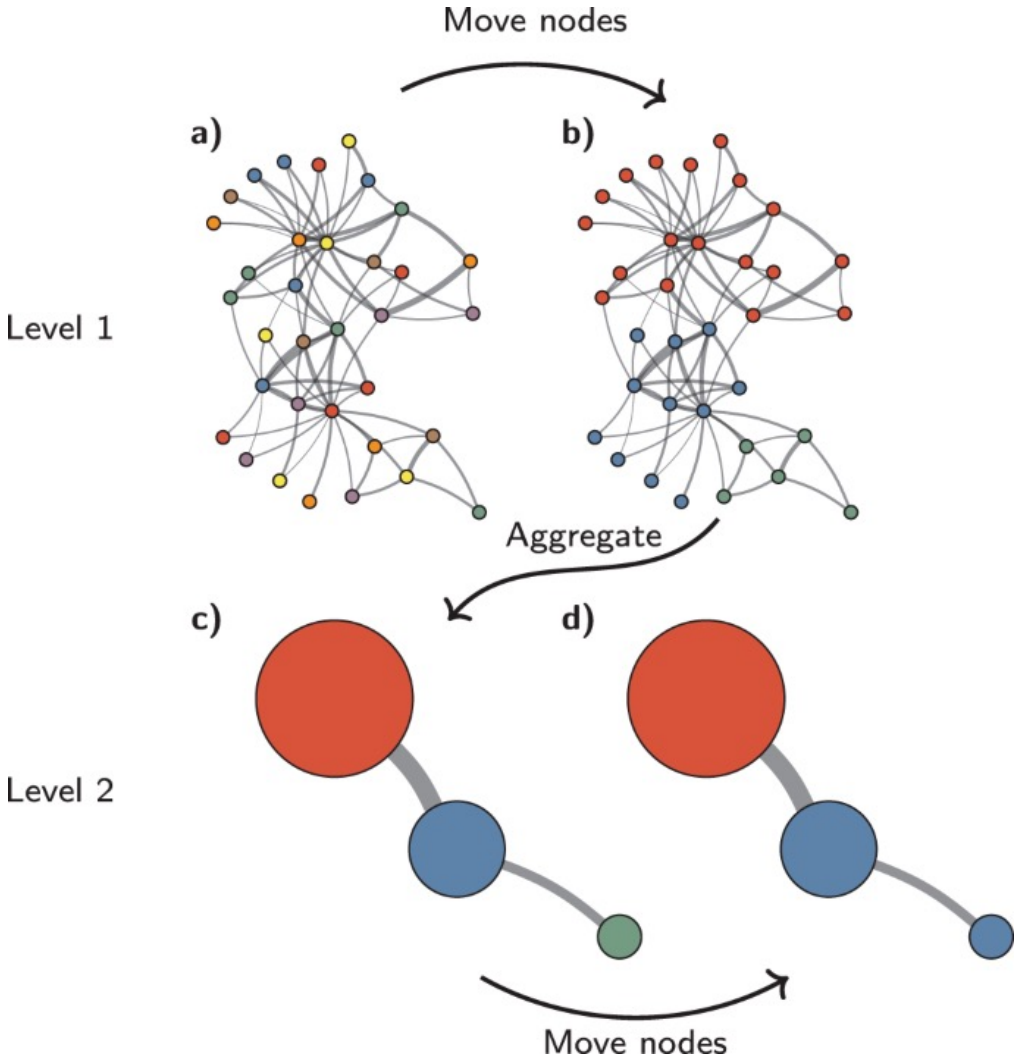
For each cell, the K-nearest neighbors (kNN) are identified

Graph-based clustering



The graph is partitioned into communities (clusters) based on modularity optimization

Graph communities by **Louvain** algorithm



a) The Louvain algorithm starts from a singleton partition in which each node is in its own community.

b) The algorithm moves individual nodes from one community to another to find a partition.

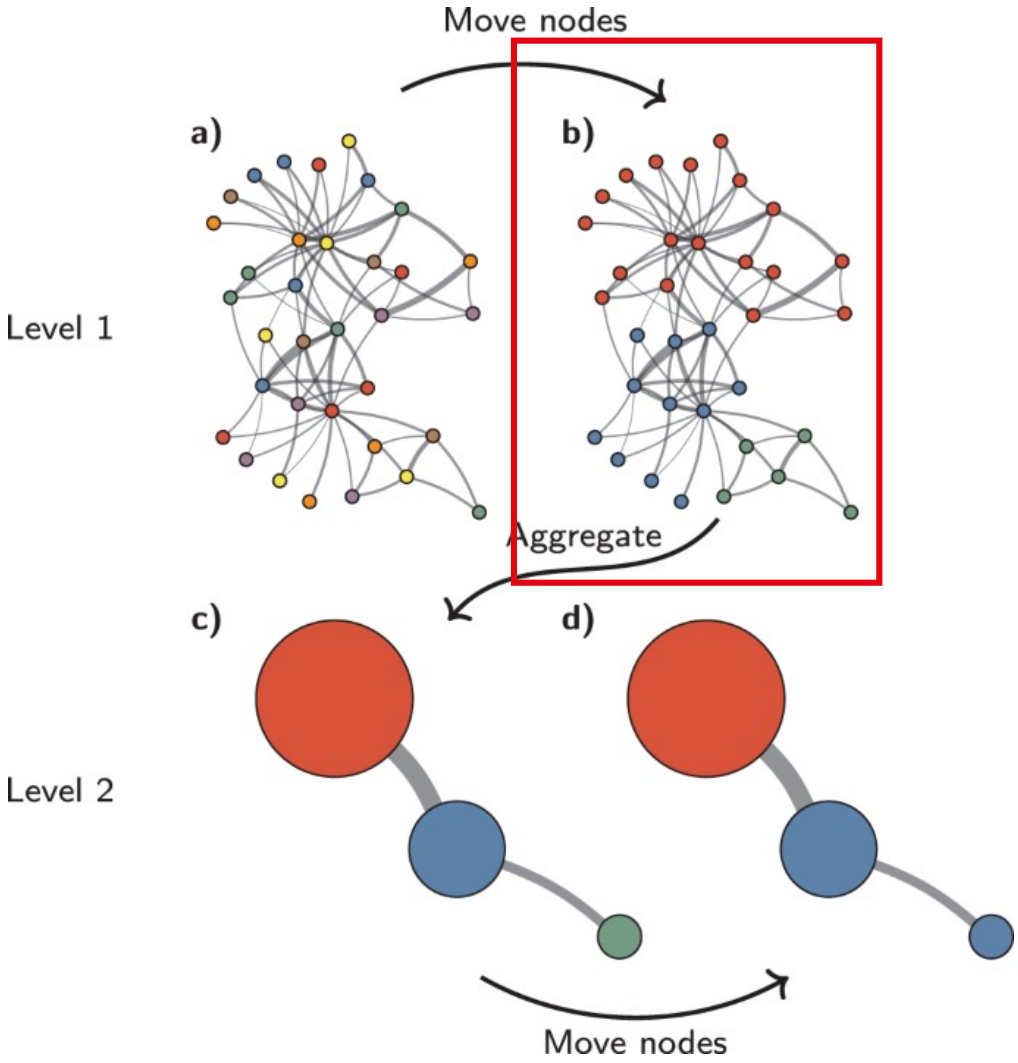
c) Based on this partition, an aggregate network is created.

d) The algorithm then moves individual nodes in the aggregate network.

These steps are repeated until the quality cannot be increased further.

https://www.nature.com/articles/s41598-019-41695-z

Louvain algorithm resolution parameter



a) The Louvain algorithm starts from a singleton partition in which each node is in its own community.

b) The algorithm moves individual nodes from one community to another to find a partition.

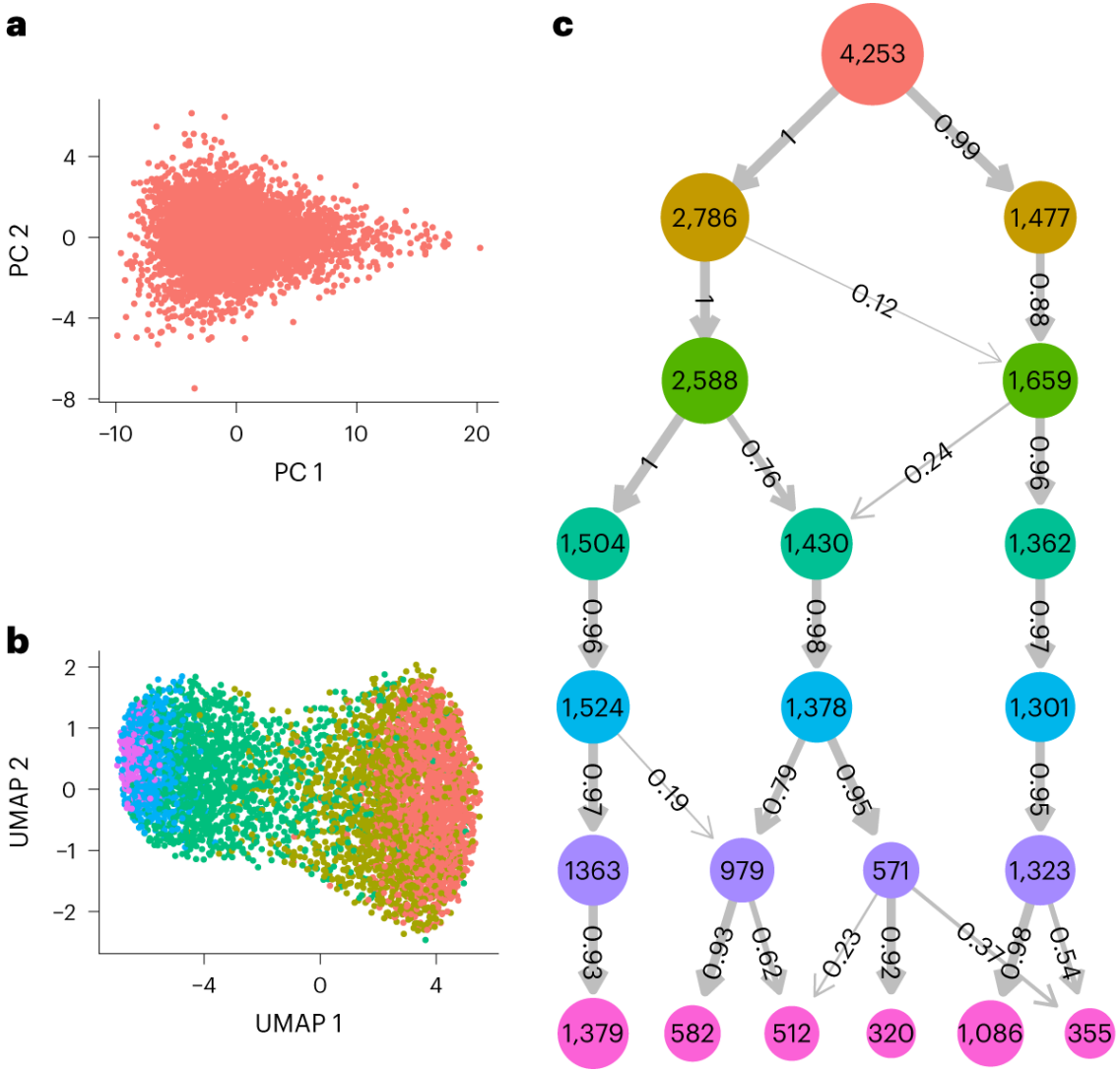
c) Based on this partition, an aggregate network is created.

d) The algorithm then moves individual nodes in the aggregate network.

These steps are repeated until the quality cannot be increased further.

https://www.nature.com/articles/s41598-019-41695-z

Louvain algorithm resolution parameter

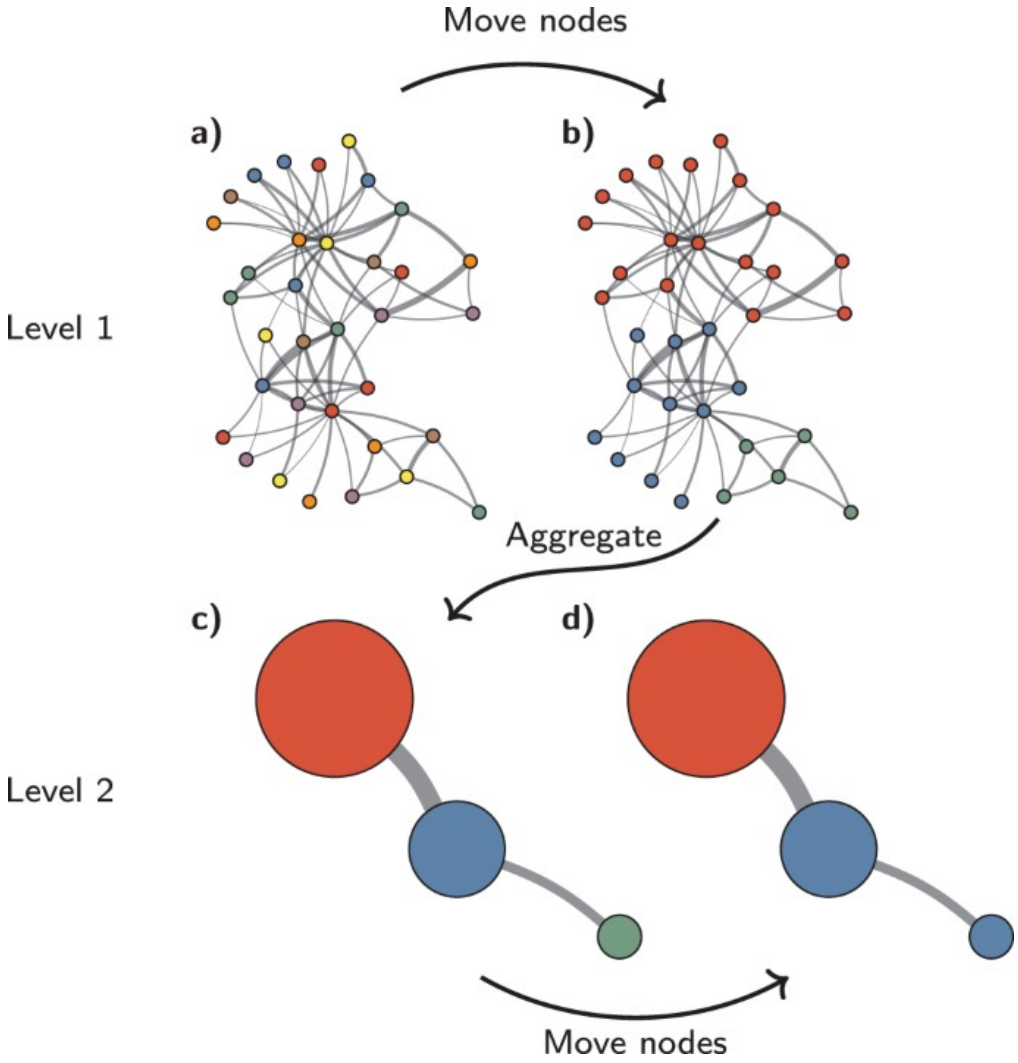


Louvain algorithm resolution parameter

Observation	Meaning
Clusters are stable for a few steps	**Optimal resolution**
Sudden splitting of clusters	Over-clustering
Merging of distinct clusters	Under-clustering

Resolution	Number of Clusters	Stability
0.2	5 clusters	Stable
0.4	7 clusters	Stable
0.6	8 clusters	Stable
0.8	15 clusters	Over-clustering

Graph communities by **Louvain** algorithm



a) The Louvain algorithm starts from a singleton partition in which each node is in its own community.

b) The algorithm moves individual nodes from one community to another to find a partition.

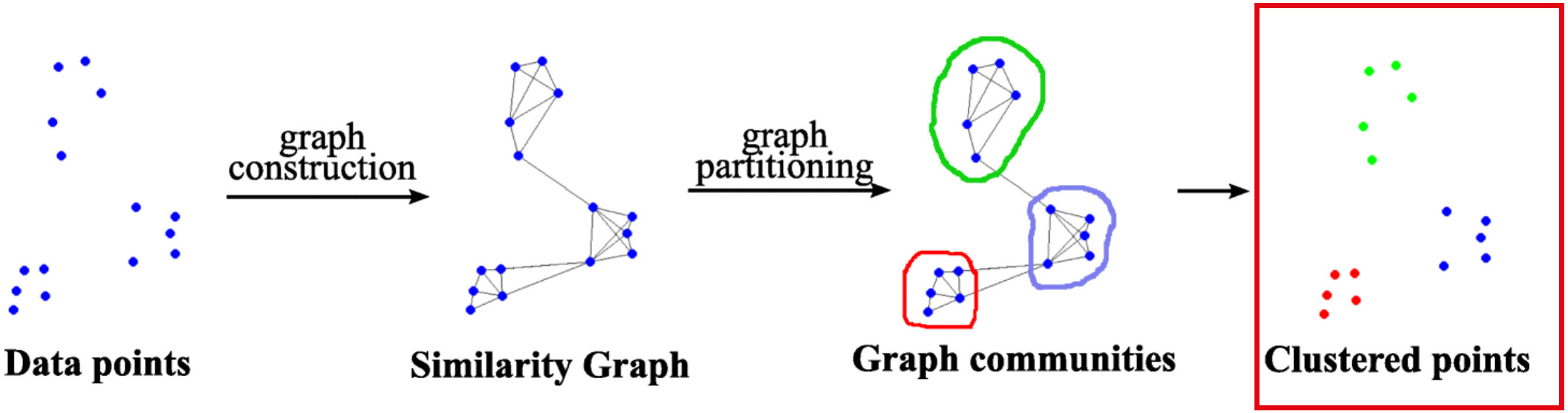
c) Based on this partition, an aggregate network is created.

d) The algorithm then moves individual nodes in the aggregate network.

These steps are repeated until the quality cannot be increased further.

https://www.nature.com/articles/s41598-019-41695-z

Graph-based clustering



Cells within the same community are assigned to the same cluster

Quiz

1. In graph-based clustering, what is constructed after identifying the k-nearest neighbors (KNN)?

a) Heatmap
b) Similarity graph
c) UMAP plot
d) Differential expression matrix

Quiz

1. In graph-based clustering, what is constructed after identifying the k-nearest neighbors (KNN)?

a) Heatmap
b) Similarity graph
c) UMAP plot
d) Differential expression matrix

2. Which algorithm is commonly used for modularity optimization in graph-based clustering?

a) K-means
b) Louvain
c) PCA
d) t-SNE

Summary

Importance of Clustering:

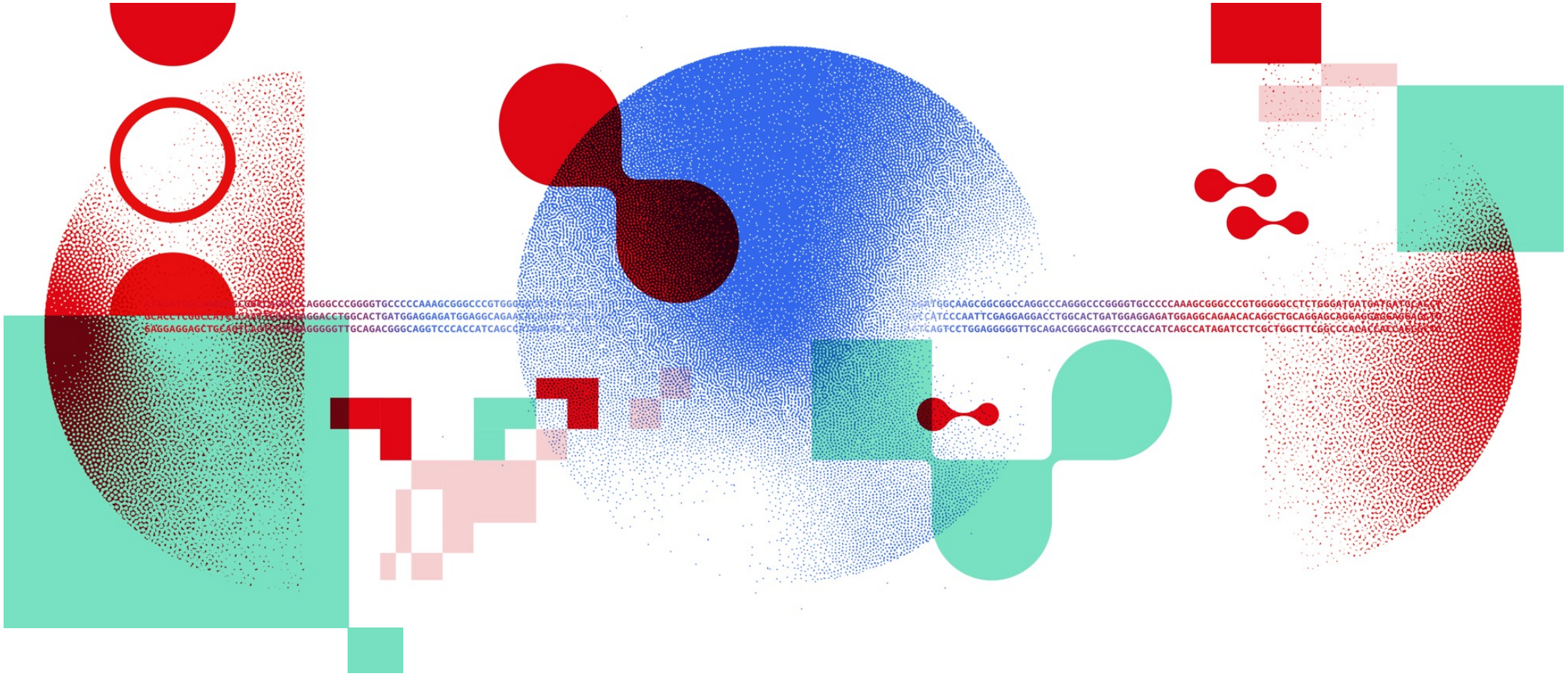
- Identify Cell Populations
- Detect Novel Cell Types

Types of Clustering:

- Various methods beyond traditional k-means clustering, including graph-based clustering

Graph-Based Clustering:

- Dimensionality Reduction: Use techniques like PCA to reduce data complexity
- K-Nearest Neighbors (kNN): Identify kNN for each cell
- Modularity Optimization: Partition the graph into communities (clusters) using the Louvain algorithm
- Louvain Algorithm: Iteratively move nodes to optimize community structure



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss