Machine learning ideas in Seurat packages

Jan/16/2020

Sunho Park

Outline

- **Dimensionality reduction (visualization)**
  - t-SNE
  - Uniform Manifold Approximation and Projection (UMAP)

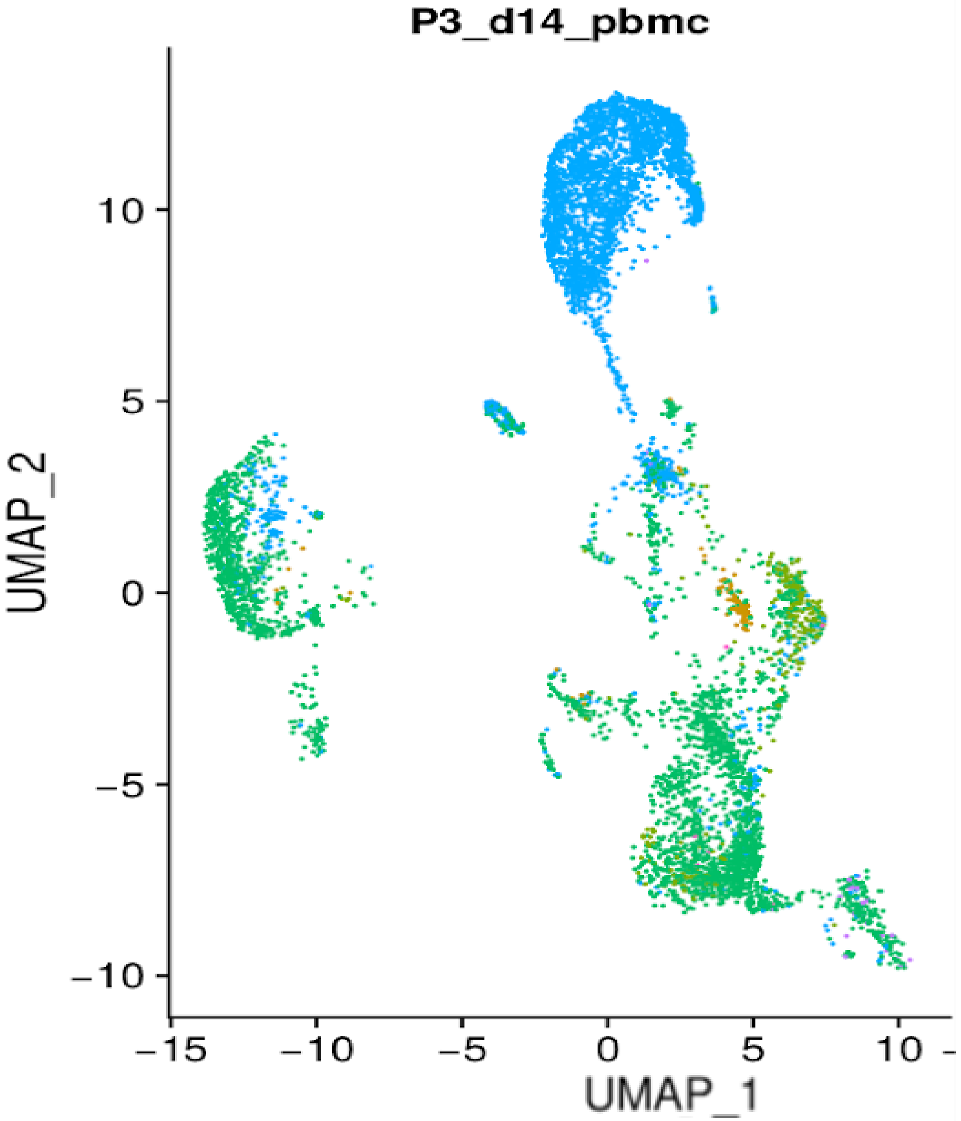
- **Clustering approaches**
  - Graph based methods: e.g., Louvain algorithm
  - Cell-type identification

- **Multiple Dataset Integration**
  - Canonical correlation analysis & L2-norm normalization
  - Anchoring

Outline

- **Dimensionality reduction (visualization)**
  - t-SNE
  - Uniform Manifold Approximation and Projection (UMAP)
- Clustering approaches
  - Graph based methods: Louvain algorithm
  - Cell-type identification
- Multiple Dataset Integration
  - Canonical correlation analysis & L2-norm normalization
  - Anchoring

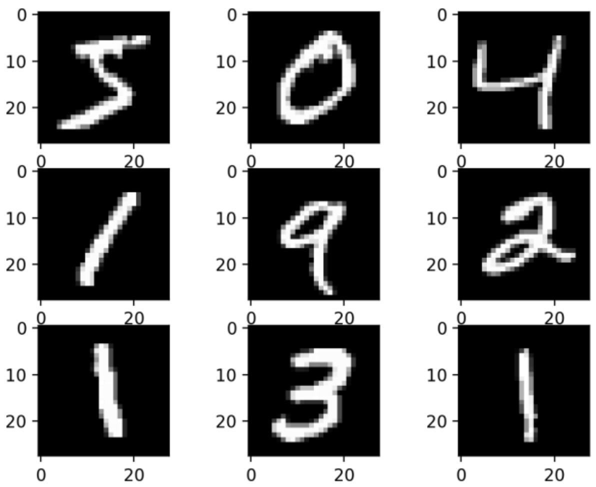
Dimensionality reduction for single-cell data



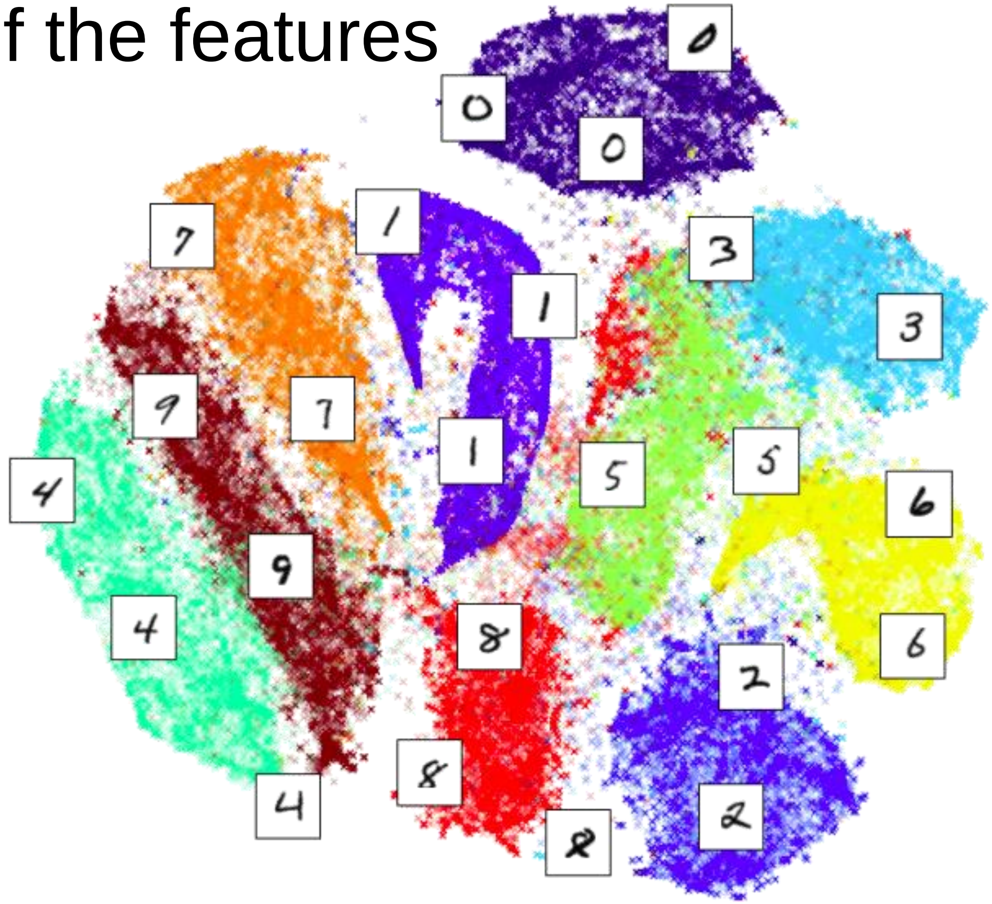
Dimensionality reduction

Figure credit: liam schoneveld

- A process of reducing the number of the features



Original image (MNIST): $\mathbf{x}_i \in \mathbb{R}^{784}$ (28×28)



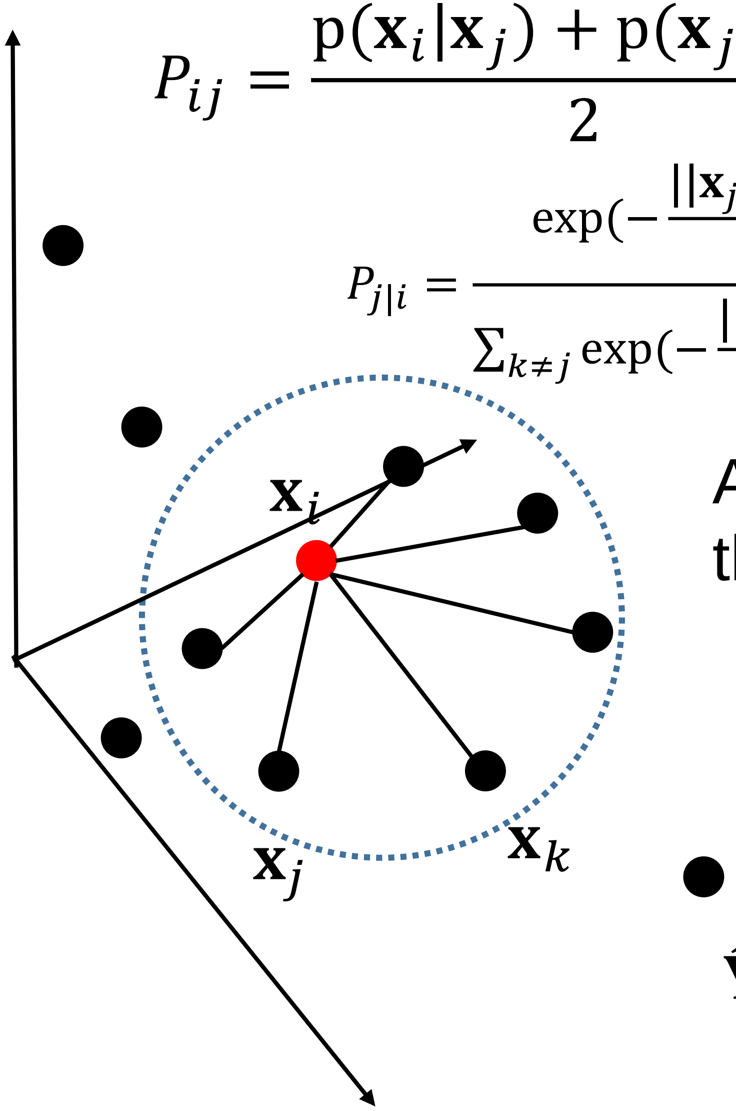
Reduced space (by t-SNE): $\mathbf{x}_i \in \mathbb{R}^2$

Dimensionality reduction: t-SNE

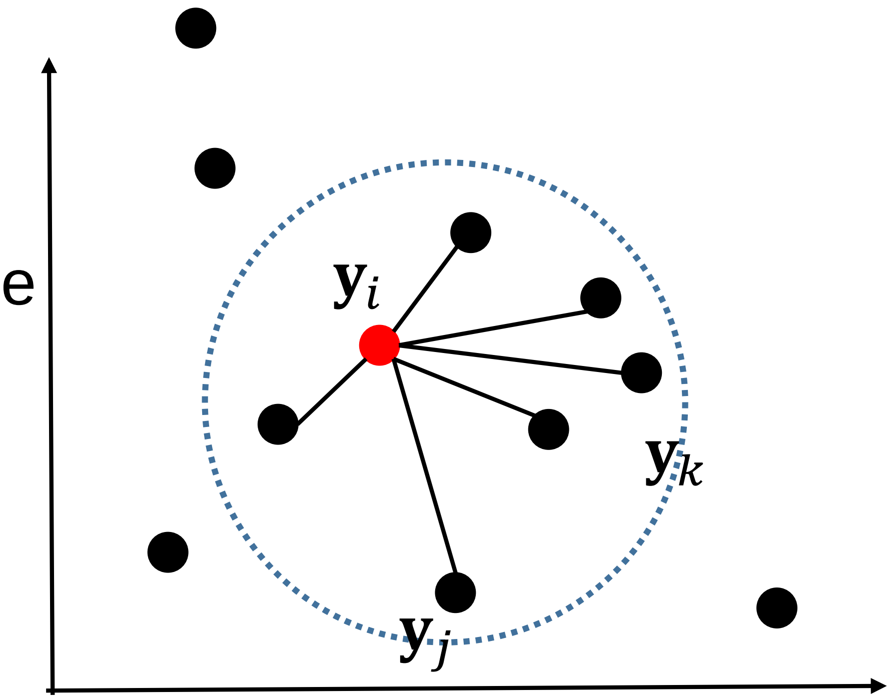
$$P_{ij} = \frac{\mathrm{p}(\mathbf{x}_i|\mathbf{x}_j) + \mathrm{p}(\mathbf{x}_j|\mathbf{x}_i)}{2}$$

$$Q_{ij} = \frac{(1 + ||\mathbf{y}_j - \mathbf{y}_i||_2^2)^{-1}}{\sum_{k\neq j}(1 + ||\mathbf{y}_k - \mathbf{y}_i||_2^2)^{-1}}$$

$$P_{j|i} = \frac{\exp(-\frac{||\mathbf{x}_j - \mathbf{x}_i||_2^2}{2\sigma_i^2})}{\sum_{k\neq j}\exp(-\frac{||\mathbf{x}_k - \mathbf{x}_i||_2^2}{2\sigma_i^2})}$$



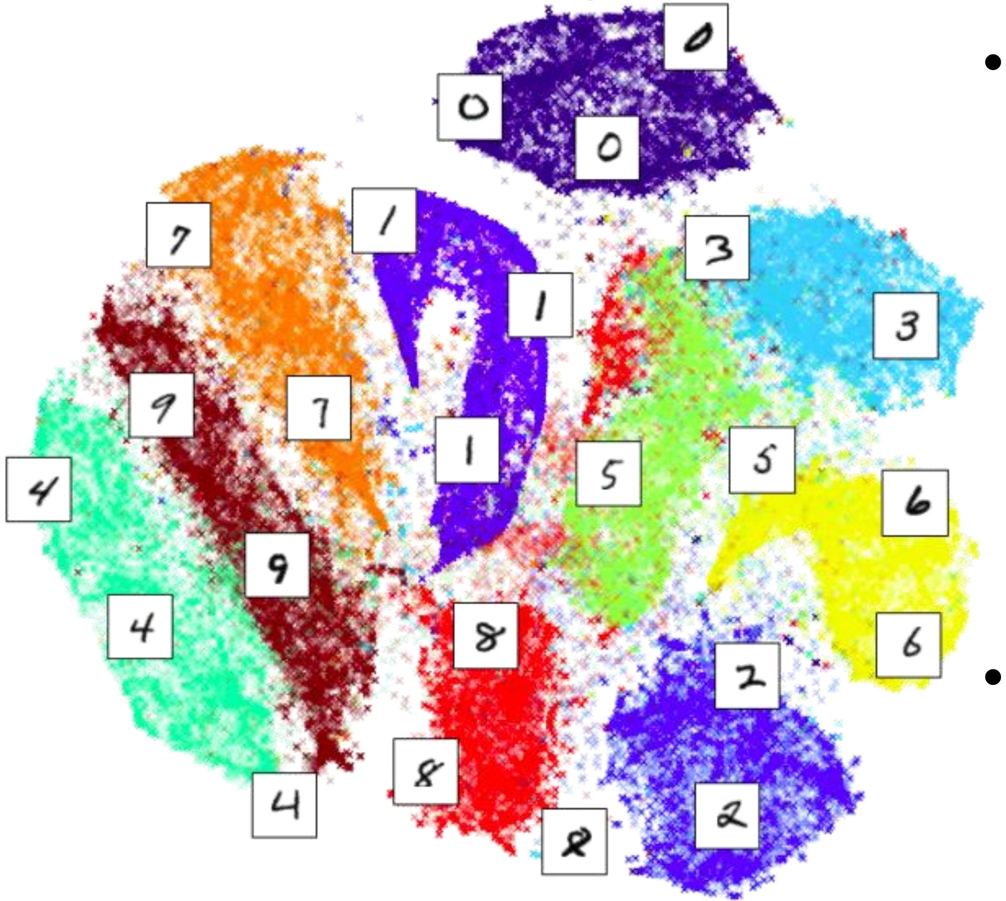
A mapping to preserve
the local neighbor structure
$\{P_{ij}\} \approx \{Q_{ij}\}$



$$\hat{\mathbf{Y}} = \min_{\{\mathbf{y}_\mathrm{i}\}} \mathrm{KL}(\mathrm{P}||\mathrm{Q})$$
$$= \min_{\{\mathbf{y}_\mathrm{i}\}} \sum_\mathrm{i}\sum_\mathrm{j} P_{ij}\log\frac{P_{ij}}{Q_{ij}}$$

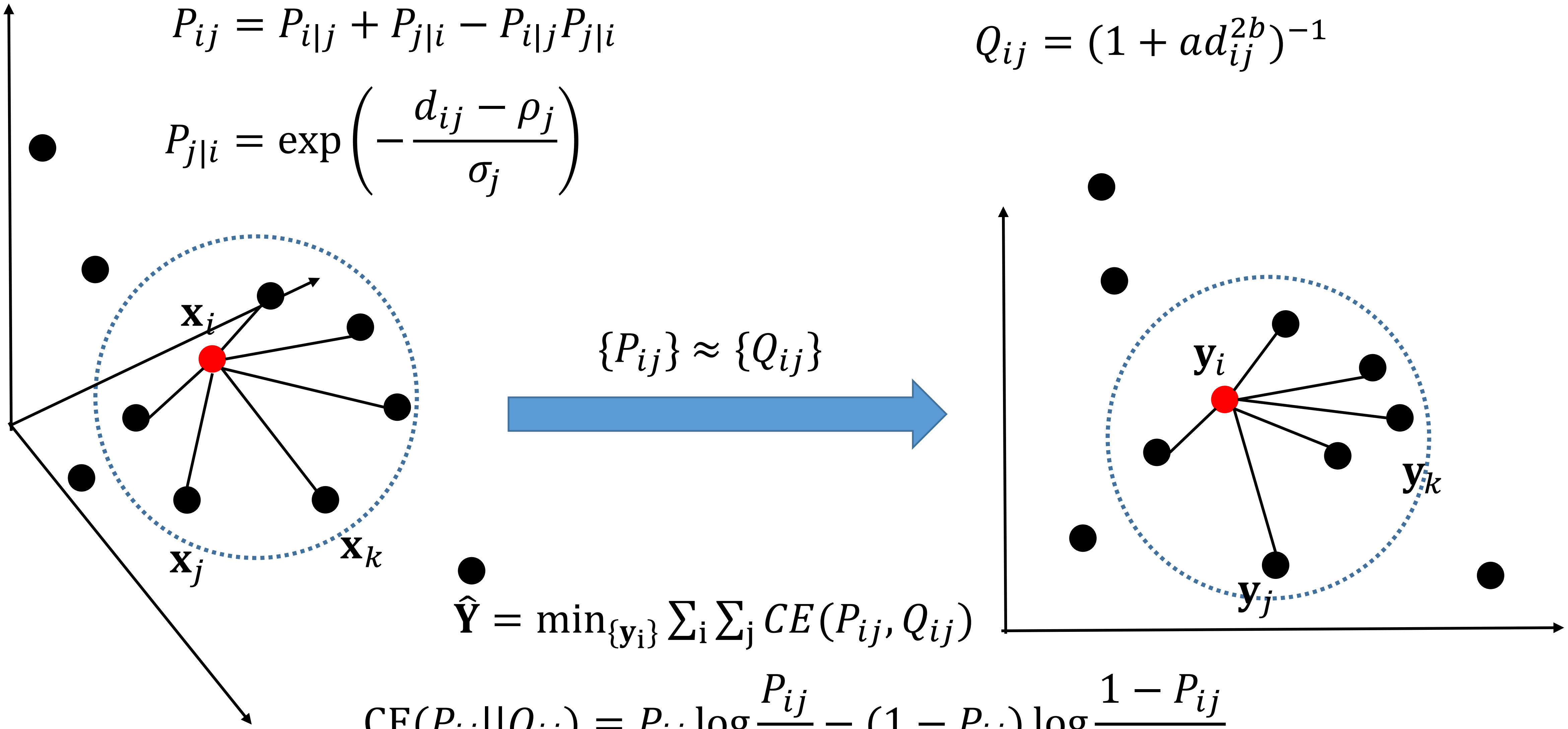
Dimensionality reduction

- t-SNE has been widely used for visualization of high dimensional data, but it has some limitations:



- **The global structure is not preserved**
  - points in the same cluster are close to each other (local structure)
  - Similarities between clusters might not be accurate (global structure, e.g., consider clusters (0 and 3), (0 and 6))
- **It is not scalable to handle fast growing sample sizes in single cell data.**

Uniform Manifold Approximation and Projection (UMAP)



$$P_{ij} = P_{i|j} + P_{j|i} - P_{i|j}P_{j|i}$$

$$P_{j|i} = \exp\left(-\frac{d_{ij} - \rho_j}{\sigma_j}\right)$$

$$Q_{ij} = (1 + a d_{ij}^{2b})^{-1}$$

$$\{P_{ij}\} \approx \{Q_{ij}\}$$

$$\hat{\mathbf{Y}} = \min_{\{\mathbf{y}_\mathrm{i}\}} \sum_\mathrm{i} \sum_\mathrm{j} CE(P_{ij}, Q_{ij})$$

$$\mathrm{CE}(P_{ij}||Q_{ij}) = P_{ij} \log \frac{P_{ij}}{Q_{ij}} - (1 - P_{ij}) \log \frac{1 - P_{ij}}{1 - Q_{ij}}$$

Dimensionality reduction (t-SNE vs UMAP)

- UMAP works better for preserving the global structure of the data than t-SNE
  - t-SNE

$$\mathrm{KL}(P_{ij}||Q_{ij}) \approx \exp\left(-d_{ij}^X\right)\log(1+(d_{ij}^Y)^2)$$

– d_{ij}^X is large: d_{ij}^Y can be any value (i.e., the global structure is not guaranteed)

  - UMAP

$$\mathrm{CE}(P_{ij}||Q_{ij}) \approx \exp\left(-d_{ij}^X\right)\log\left(1+(d_{ij}^Y)^2\right) + \left(1-\exp\left(-d_{ij}^X\right)\right)\log\left(\frac{1+(d_{ij}^Y)^2}{(d_{ij}^Y)^2}\right)$$

$-d_{ij}^X$ is small: $\mathrm{CE}(P_{ij}||Q_{ij}) \approx \log\left(1+(d_{ij}^Y)^2\right)$

$-d_{ij}^X$ is large: $\mathrm{CE}(P_{ij}||Q_{ij}) \approx \log\left(\frac{1+(d_{ij}^Y)^2}{(d_{ij}^Y)^2}\right)$ (i.e., it gives a high penalty for a small d_{ij}^Y and thus the global structure can be preserved)

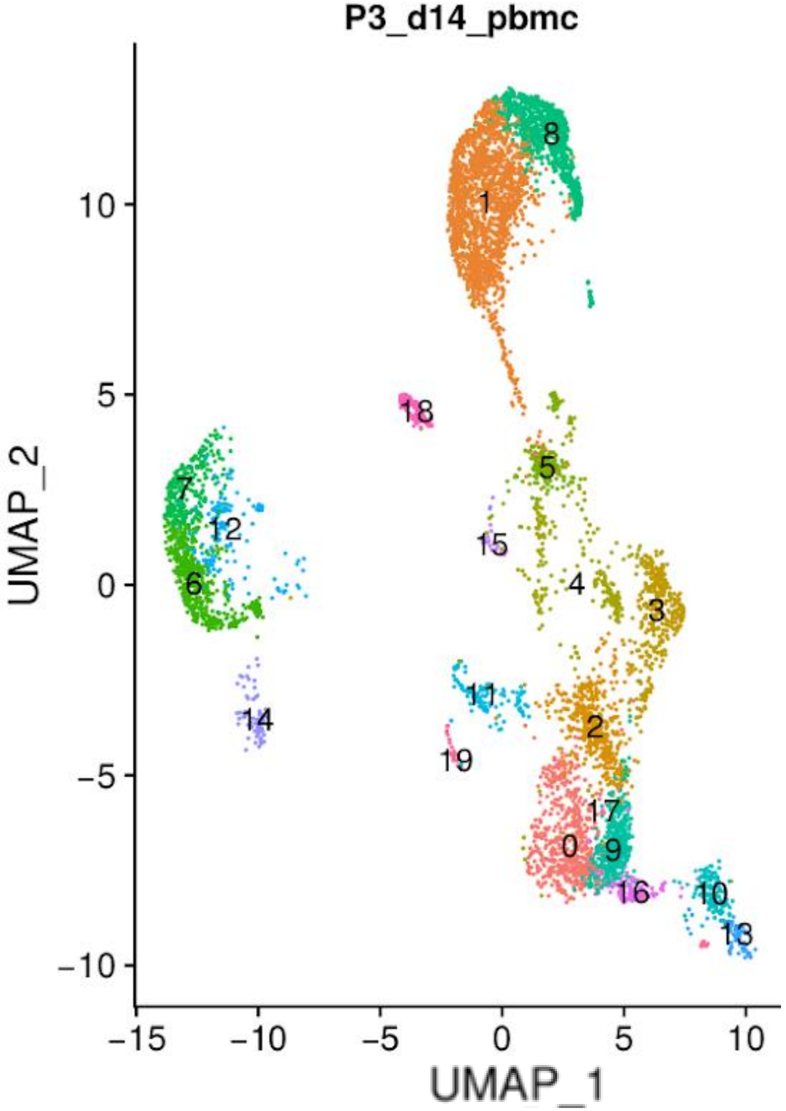
Outline

- Dimensionality reduction (visualization)
  - t-SNE
  - Uniform Manifold Approximation and Projection (UMAP)

- **Clustering approaches**
  - Graph based methods: e.g., Louvain algorithm
  - Cell-type identification

- Multiple Dataset Integration
  - Canonical correlation analysis & L2-norm normalization
  - Anchoring

Clustering approaches



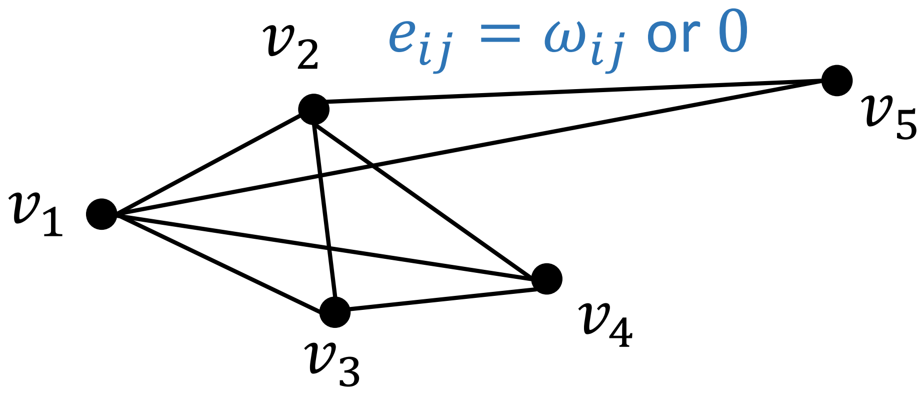
Community detection (clustering) in networks (graphs)

- Graph

  It is used to represent complex systems (e.g., friend networks on Facebook, gene-gene interaction networks)

  - $V = \{v_1, v_2, \dots, v_N\}$: nodes (vertices)
  - $E = \{e_{ij}\}$: links (edges)

$$G = \{V, E\}$$



Clustering on graphs

- Goal: divide a graph into multiple clusters where the nodes in the same cluster are more close to each other than to those in other clusters

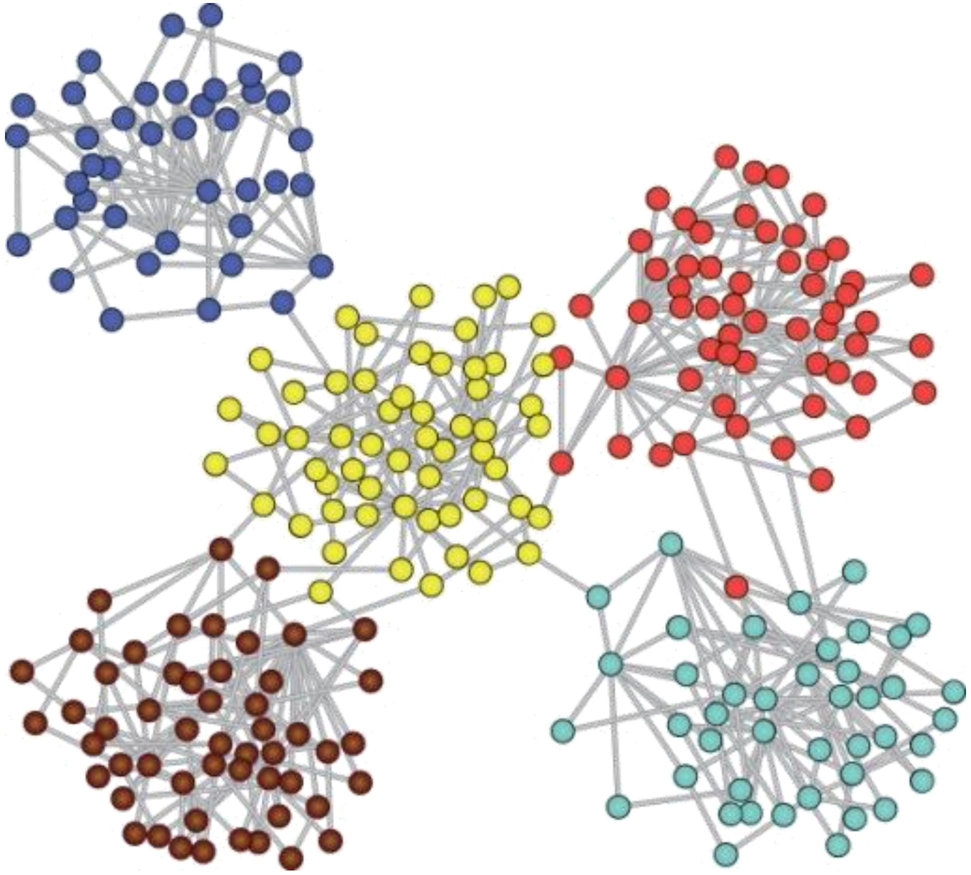


Figure credit: https://github.com/benedekrozemberczki/awesome-community-detection

Graph clustering

- Modularity
  - Measurement of the strength of partitioning modules into modules (clusters)

$$Q(\gamma) = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{d_i d_j}{2m}\right]\delta(\gamma(v_i), \gamma(v_j))$$

    - d_i: degree of the node i ($d_i = \sum_j A_{ij}$) and $m = \frac{1}{2}\sum_{ij} A_{ij}$
    - $\frac{d_i d_j}{2m}$: probability of an edge existing between v_i and v_j in the random null model
    - the fraction of the edges inside cluster minus the expected fraction if edges were distributed at random

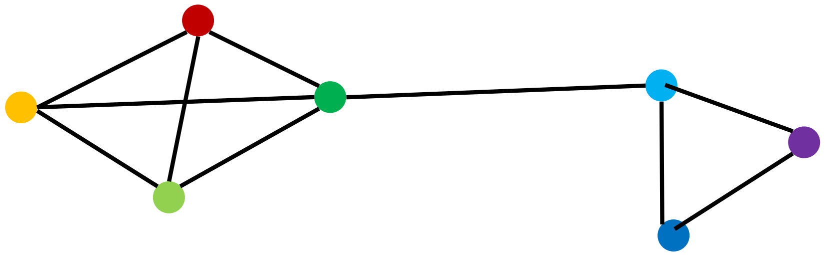
  - The high modularity means that there are dense connections between the nodes in the same cluster but sparse connections between nodes in different clusters

Graph clustering

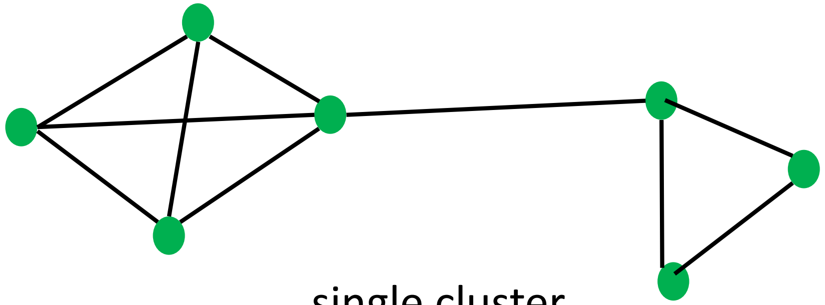
- Modularity

$$Q(\gamma) = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{d_i d_j}{2m}\right]\delta(\gamma(v_i),\gamma(v_j))$$

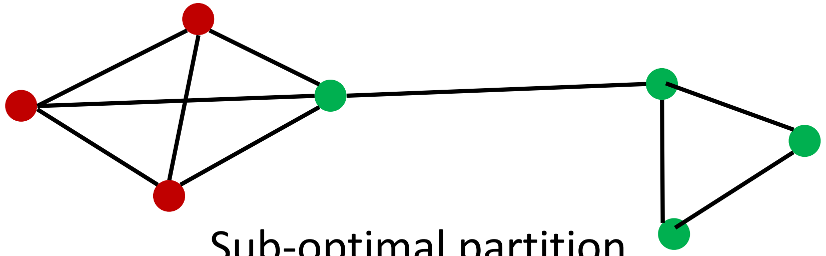
The high modularity means that there are dense connections between the nodes in the same cluster but sparse connections between nodes in different clusters



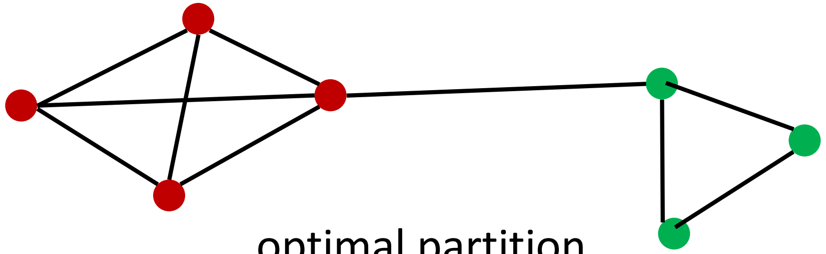
# of clusters = # of nodes



single cluster



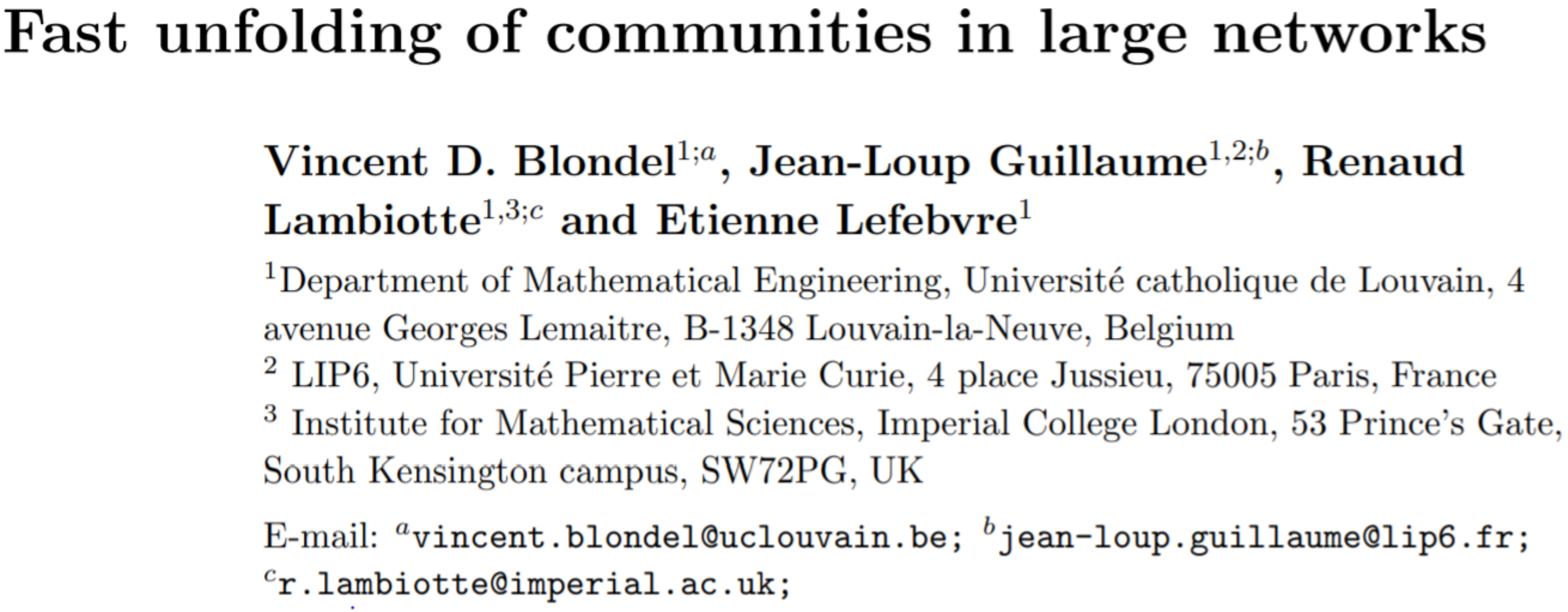
Sub-optimal partition



optimal partition

Clustering on graphs

- Louvain algorithm
  - greedily maximizes the modularity score by using an agglomerative approach
  - is one of the fastest modularity-based algorithms and scalable



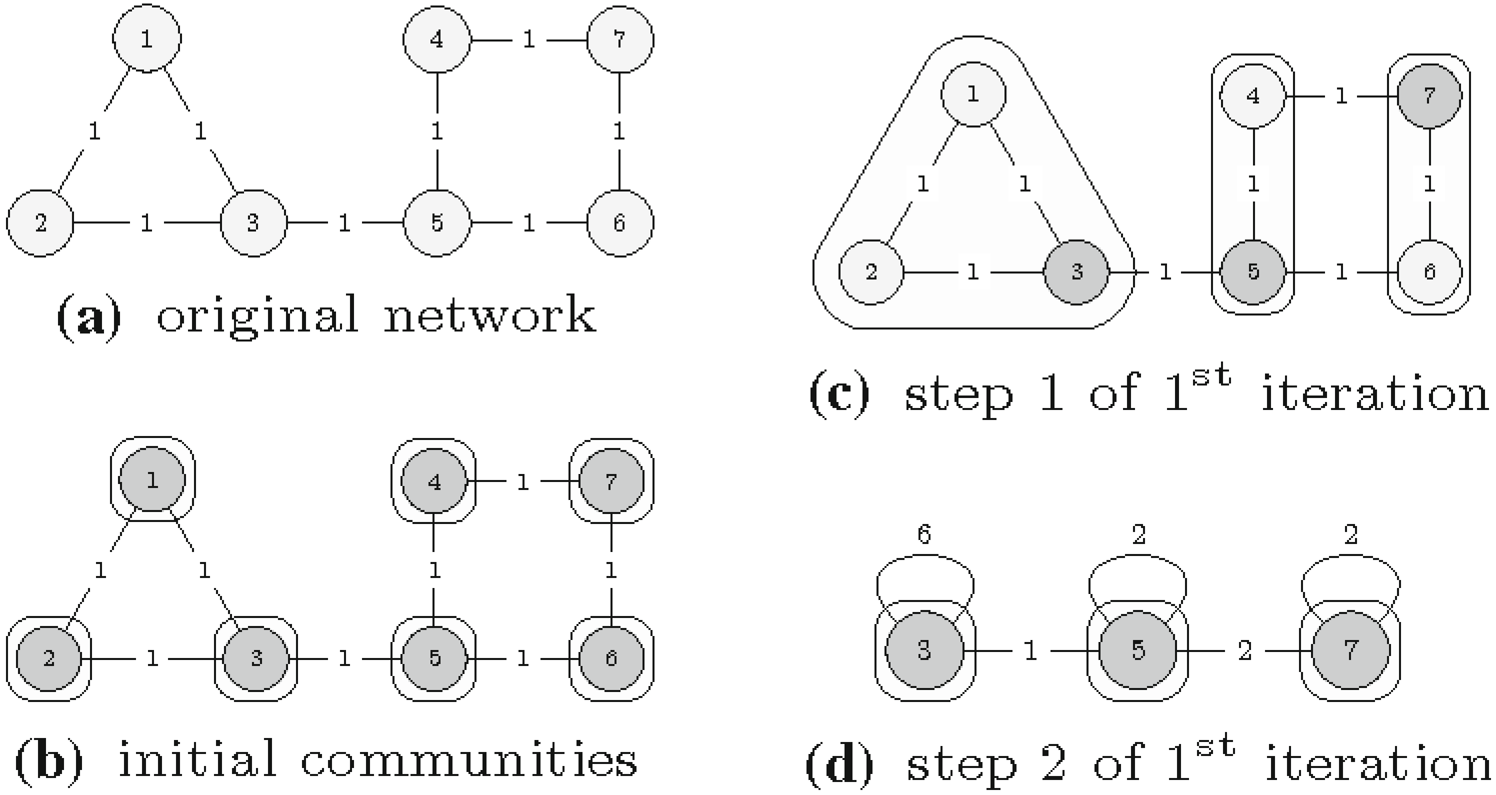
Fast unfolding of communities in large networks

Vincent D. Blondel[1;a], **Jean-Loup Guillaume**[1,2;b], **Renaud Lambiotte**[1,3;c] **and Etienne Lefebvre**[1]

[1]Department of Mathematical Engineering, Université catholique de Louvain, 4 avenue Georges Lemaitre, B-1348 Louvain-la-Neuve, Belgium
[2] LIP6, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France
[3] Institute for Mathematical Sciences, Imperial College London, 53 Prince's Gate, South Kensington campus, SW72PG, UK

E-mail: [a]vincent.blondel@uclouvain.be; [b]jean-loup.guillaume@lip6.fr; [c]r.lambiotte@imperial.ac.uk;

Clustering on graphs (Louvain algorithm)



(a) original network

(b) initial communities

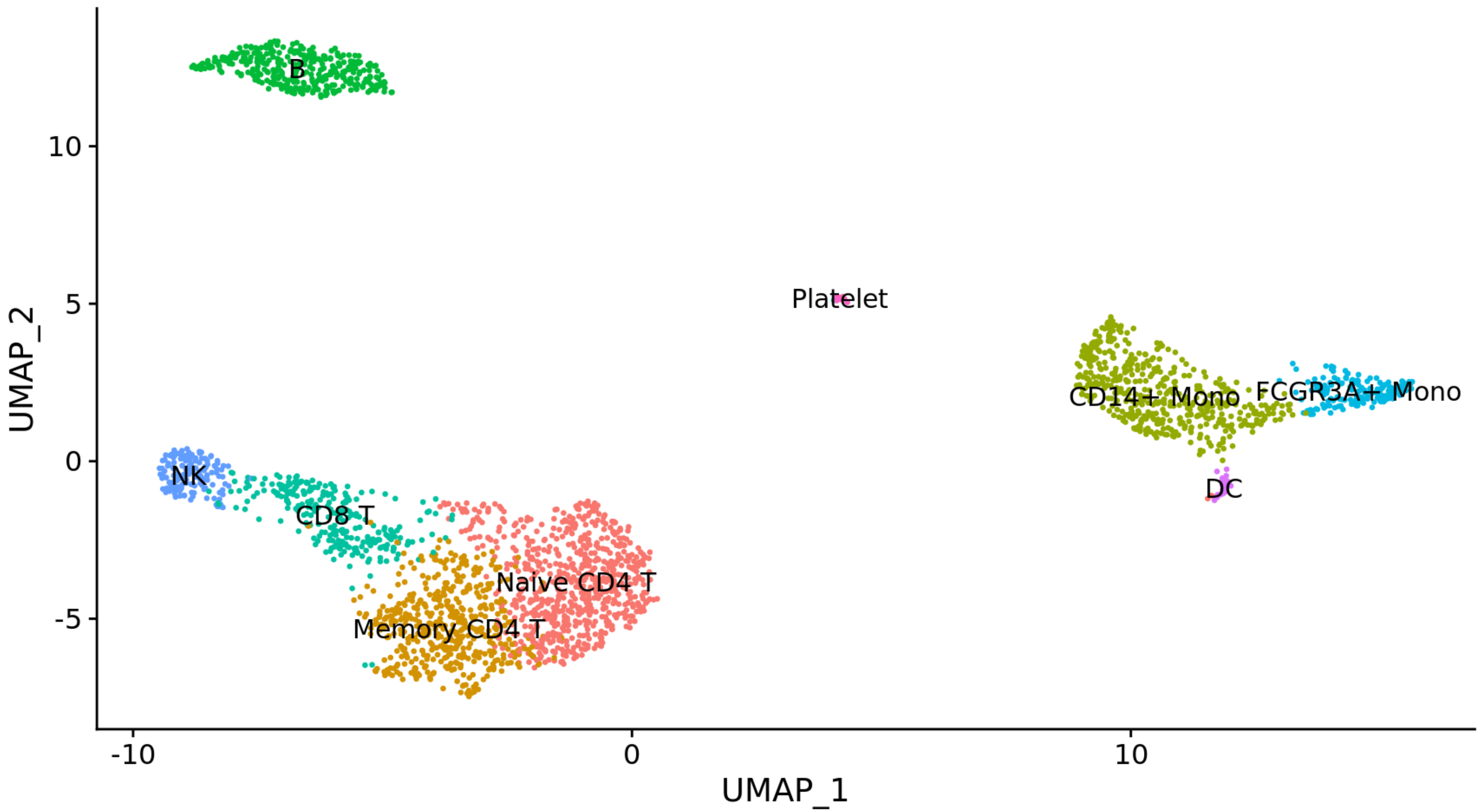
(c) step 1 of 1st iteration

(d) step 2 of 1st iteration

Figure: Mário Cordeiro et al., “Dynamic community detection in evolving networks using locality modularity optimization”, Social Network Analysis and Mining, 2016

Cell-type identification

- Identify cell-types from single-cell RNA sequencing data



Cell-type identification

- **Clustering step**
  - Find sub-groups in single-cell data using any clustering methods (e.g., Louvain algorithm)

- **Assignment step**
  - Marker identification: Find differently expressed genes in each cluster compared to other cells (i.e., points in other clusters)

  - Assign cell-type identities to the clusters by matching the marker genes with the members in known cell-type signatures

Cell-type identification

1. Clustering

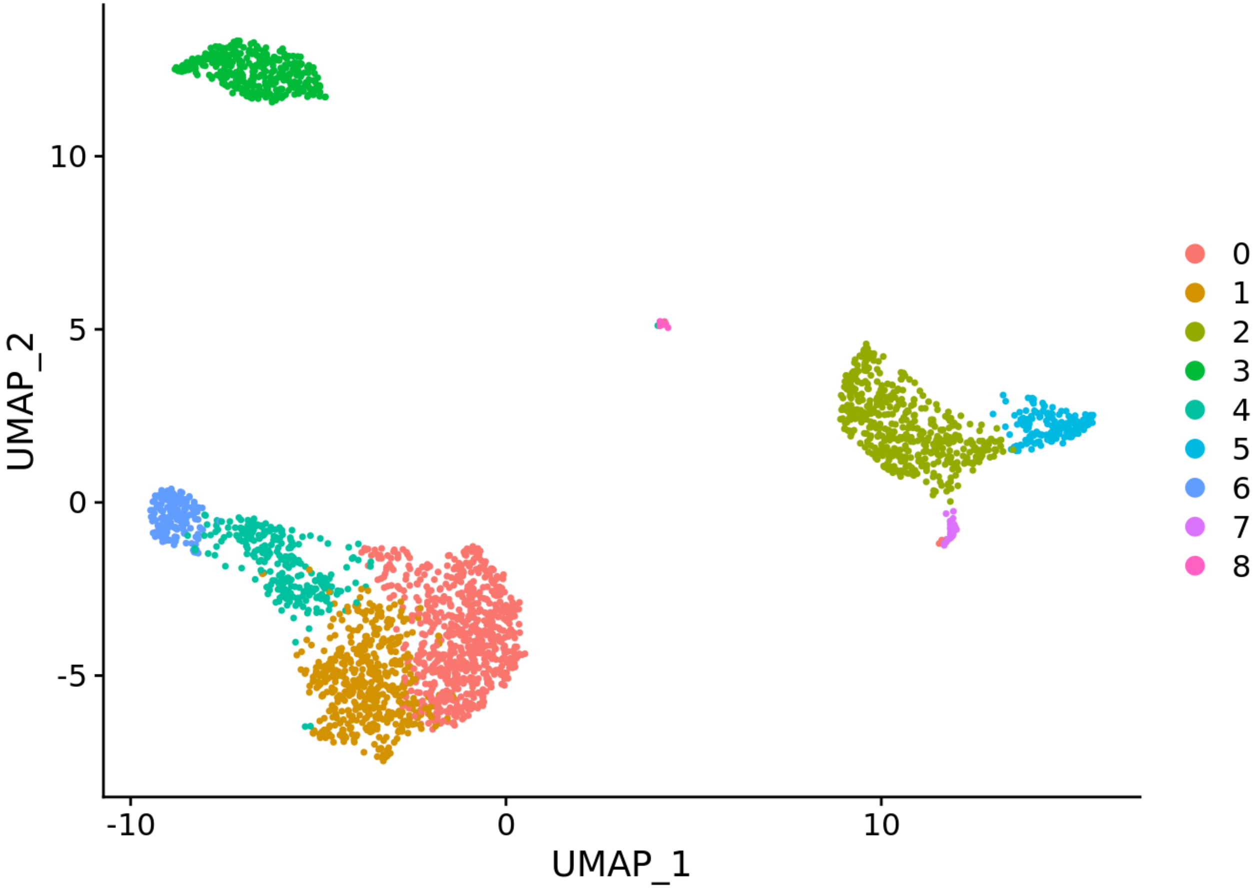
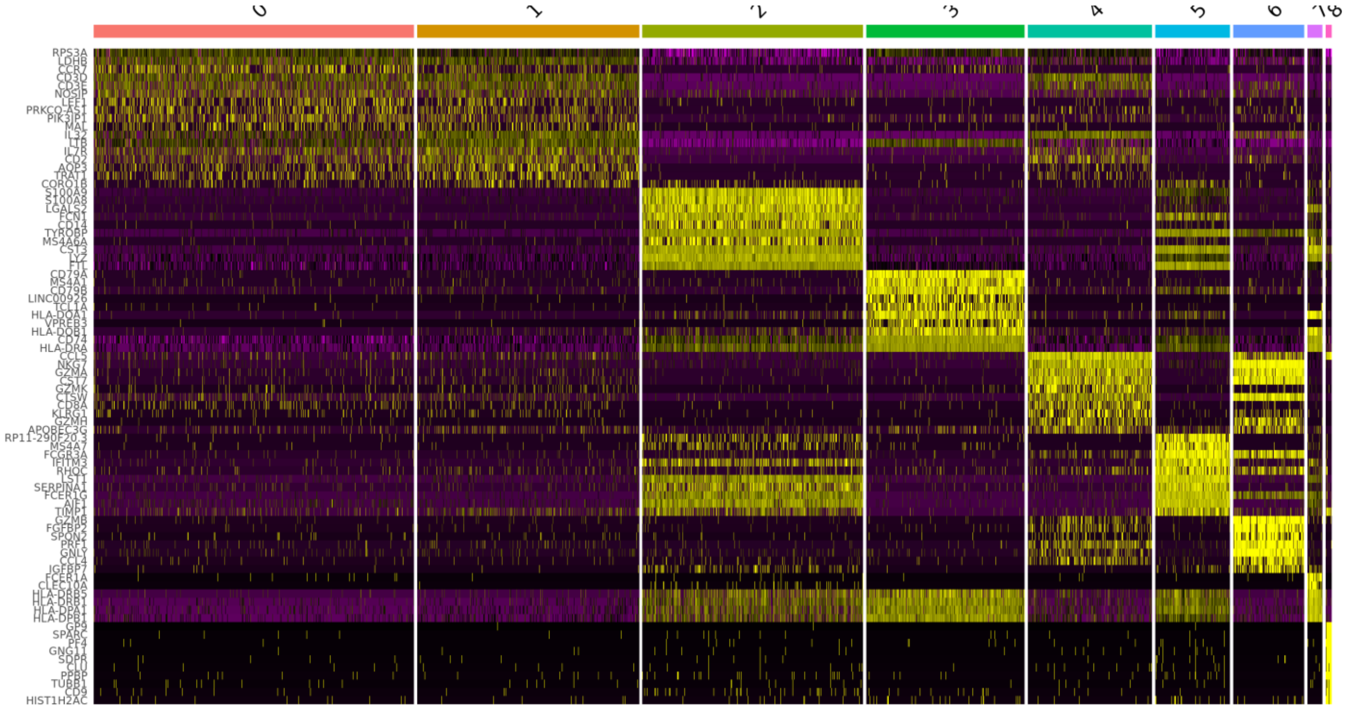


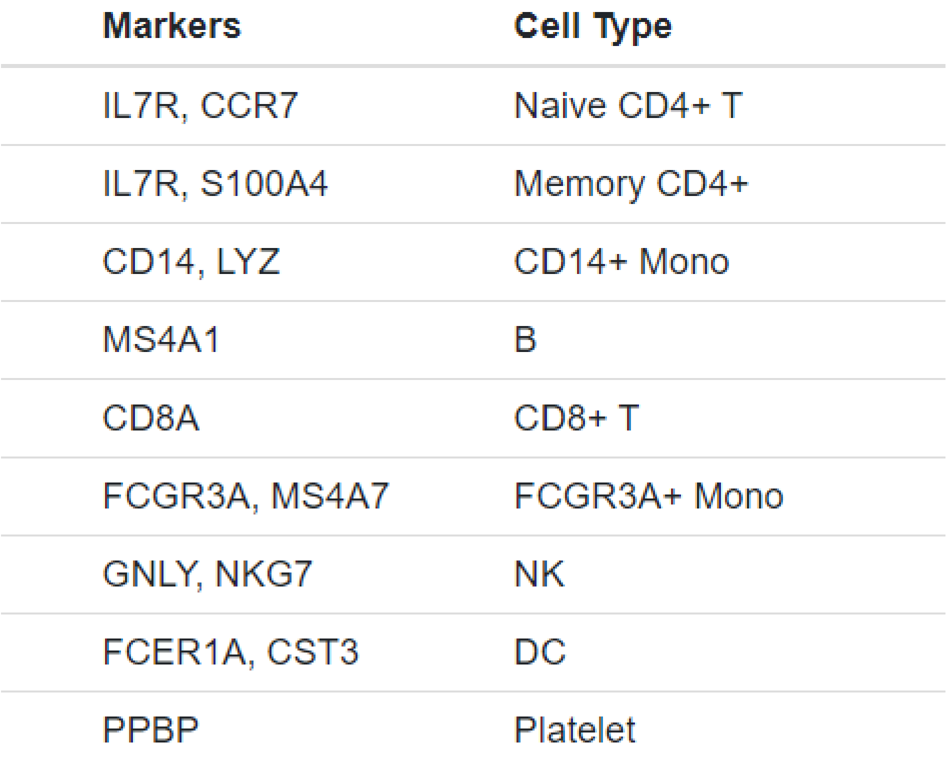
Figure credit: Seurat package

Cell-type identification

2-1. Marker gene identification: finding differently expressed genes

2-2. Cell-type assignment



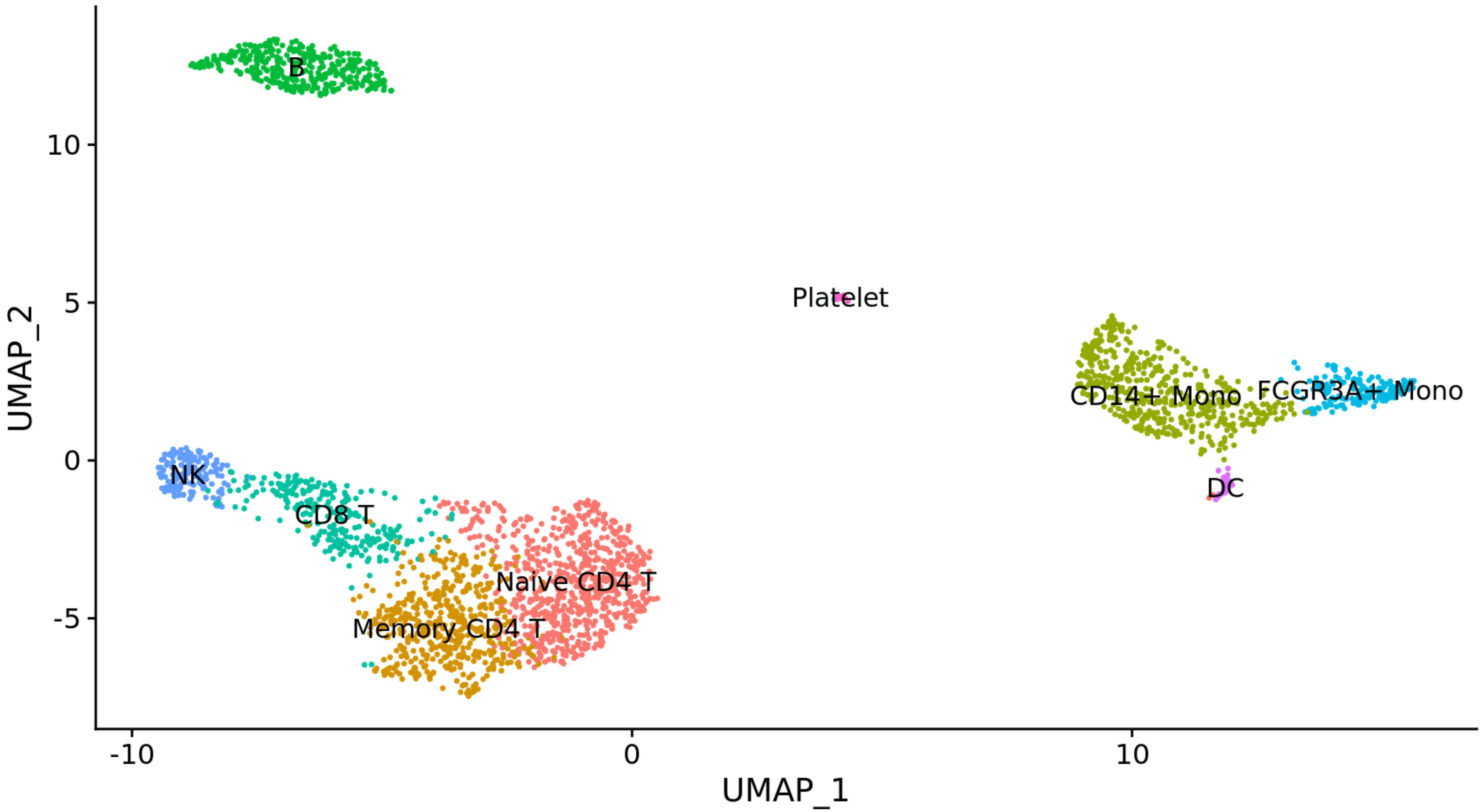


Markers	Cell Type
IL7R, CCR7	Naive CD4+ T
IL7R, S100A4	Memory CD4+
CD14, LYZ	CD14+ Mono
MS4A1	B
CD8A	CD8+ T
FCGR3A, MS4A7	FCGR3A+ Mono
GNLY, NKG7	NK
FCER1A, CST3	DC
PPBP	Platelet

Figure credit: Seurat package

Cell-type identification

- Final results



Outline

- Dimensionality reduction (visualization)
  - t-SNE
  - Uniform Manifold Approximation and Projection (UMAP)

- Clustering approaches
  - Graph based methods: e.g., Louvain algorithm
  - Cell-type identification

- **Multiple Dataset Integration**
  - Canonical correlation analysis & L2-norm normalization
  - Anchoring

Multiple Dataset Integration

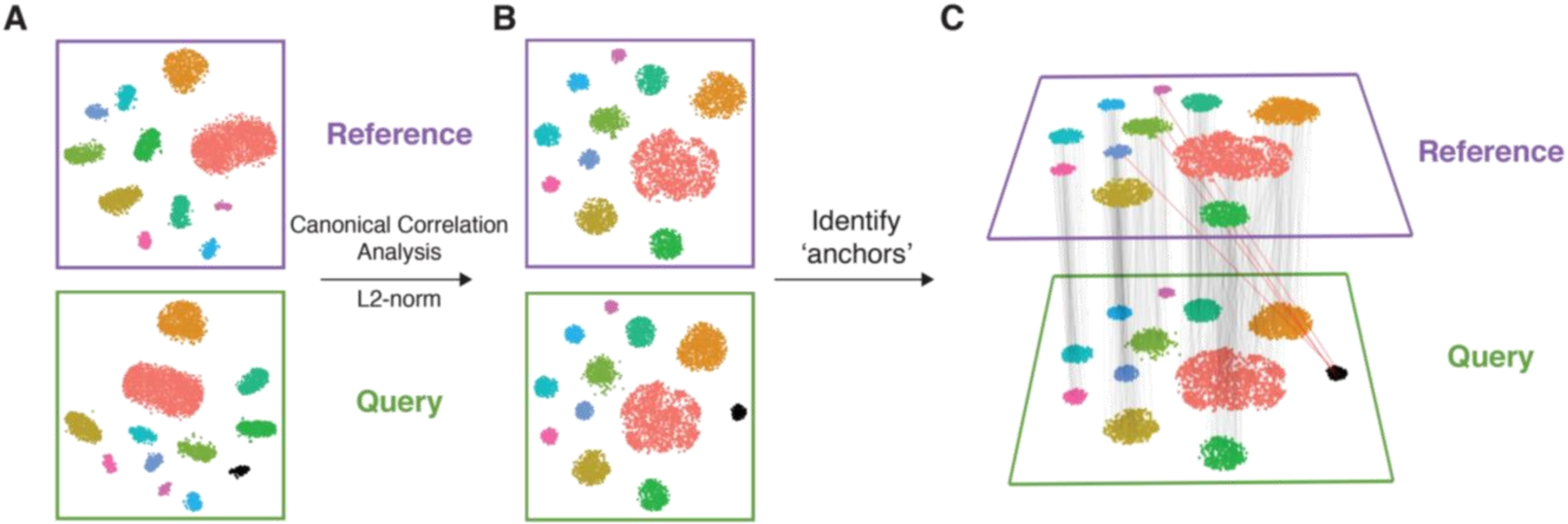


Figure: Tim Stuart et al. “Comprehensive Integration of Single-Cell Data”, Cell 2019

Multiple Dataset Integration

- Canonical correlation analysis (CCA)

- finds a common space of two sets of data

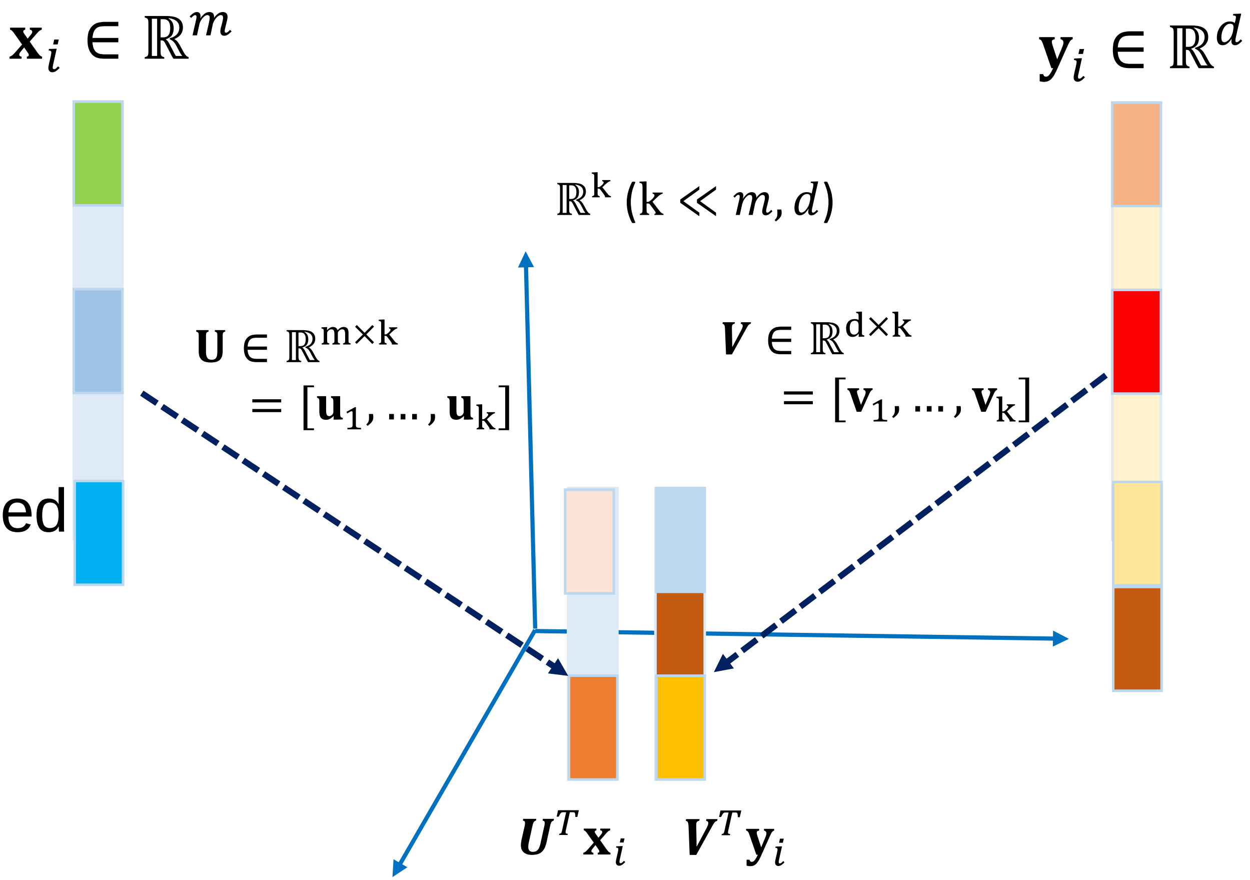
$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^{\boldsymbol{T}}$$

$$\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N]^{\boldsymbol{T}}$$

- finds a linear transformation of variables

that makes the two variables maximally correlated

$$\hat{\mathbf{u}}, \hat{\mathbf{v}} = \mathrm{argmax}_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{X}^T \mathbf{Y} \mathbf{v}$$
$$\text{s.t. } \mathbf{u}^\mathrm{T} \mathbf{X}^\mathrm{T} \mathbf{X} \mathbf{u} \le 1, \mathbf{v}^\mathrm{T} \mathbf{Y}^\mathrm{T} \mathbf{Y} \mathbf{u} \le 1$$



$\mathbf{x}_i \in \mathbb{R}^m$
$\mathbf{y}_i \in \mathbb{R}^d$
$\mathbb{R}^\mathrm{k}\ (\mathrm{k} \ll m, d)$
$\mathbf{U} \in \mathbb{R}^{\mathrm{m}\times\mathrm{k}}$
$= [\mathbf{u}_1, \dots, \mathbf{u}_\mathrm{k}]$
$\boldsymbol{V} \in \mathbb{R}^{\mathrm{d}\times\mathrm{k}}$
$= [\mathbf{v}_1, \dots, \mathbf{v}_\mathrm{k}]$
$\boldsymbol{U}^T \mathbf{x}_i$
$\boldsymbol{V}^T \mathbf{y}_i$

Multiple Dataset Integration

- Canonical correlation analysis (CCA)

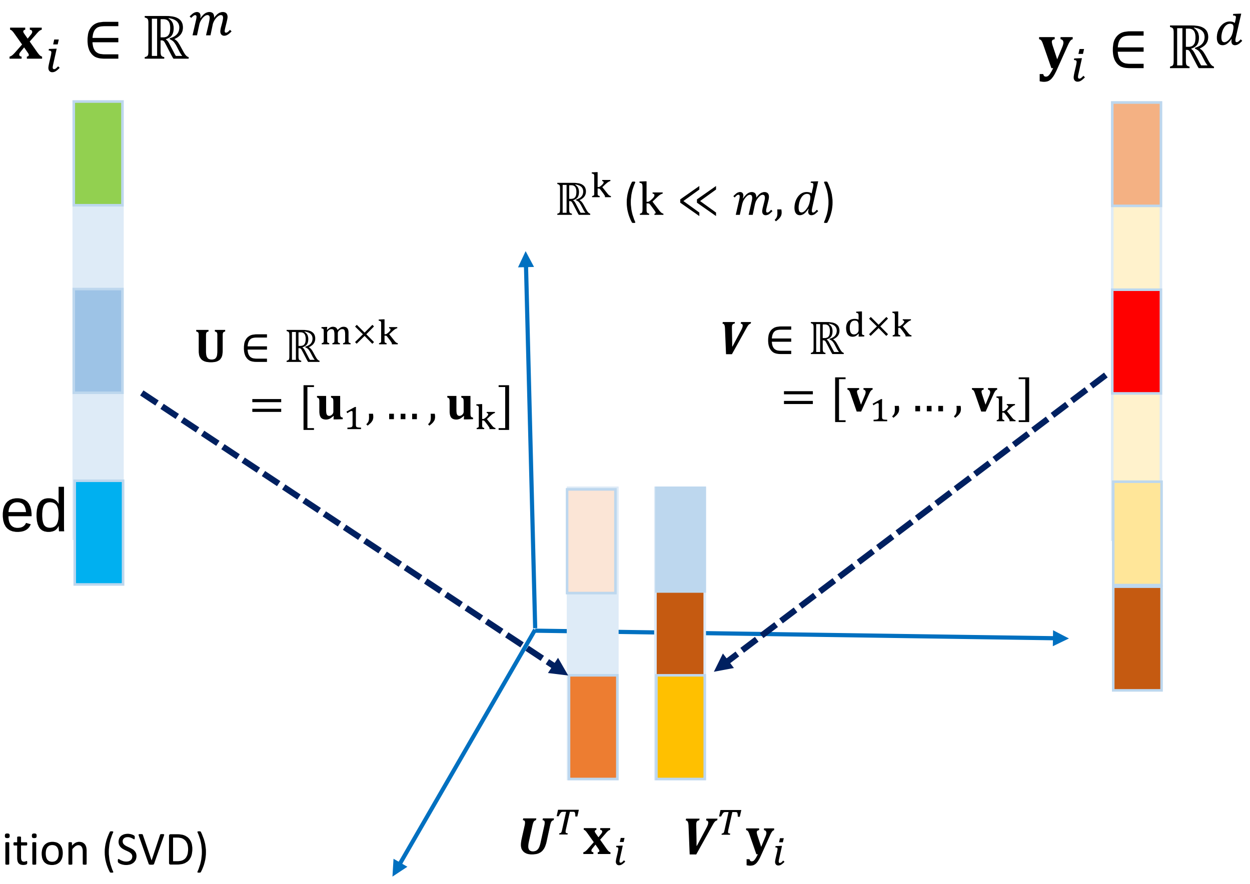
- finds a common space of two sets of data

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^{\boldsymbol{T}}$$

$$\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N]^{\boldsymbol{T}}$$

- finds a linear transformation of variables

that makes the two variables maximally correlated


$\mathbf{x}_i \in \mathbb{R}^m$
$\mathbf{y}_i \in \mathbb{R}^d$
$\mathbb{R}^{\mathrm{k}}$ $(\mathrm{k} \ll m, d)$
$\mathbf{U} \in \mathbb{R}^{\mathrm{m \times k}} = [\mathbf{u}_1, \dots, \mathbf{u}_{\mathrm{k}}]$
$\boldsymbol{V} \in \mathbb{R}^{\mathrm{d \times k}} = [\mathbf{v}_1, \dots, \mathbf{v}_{\mathrm{k}}]$
$\boldsymbol{U}^T \mathbf{x}_i$
$\boldsymbol{V}^T \mathbf{y}_i$

In Seurat, the maximization problem is simplified as

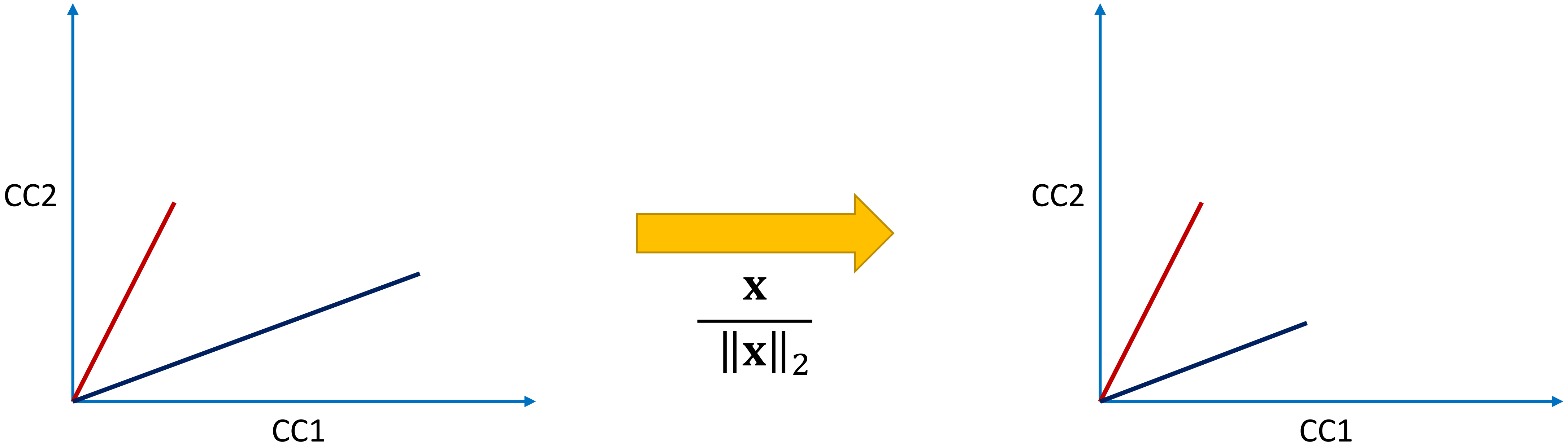
$$\hat{\mathbf{u}}, \hat{\mathbf{v}} = \mathrm{argmax}_{\mathbf{u},\mathbf{v}} \mathbf{u}^{\boldsymbol{T}} \mathbf{X}^{\boldsymbol{T}} \mathbf{Y} \mathbf{v}$$

$$\text{s.t. } ||\mathbf{u}||_2^2 \leq 1, ||\mathbf{v}||_2^2 \leq 1$$

The problem can be solved using singular value decomposition (SVD)

Multiple Dataset Integration

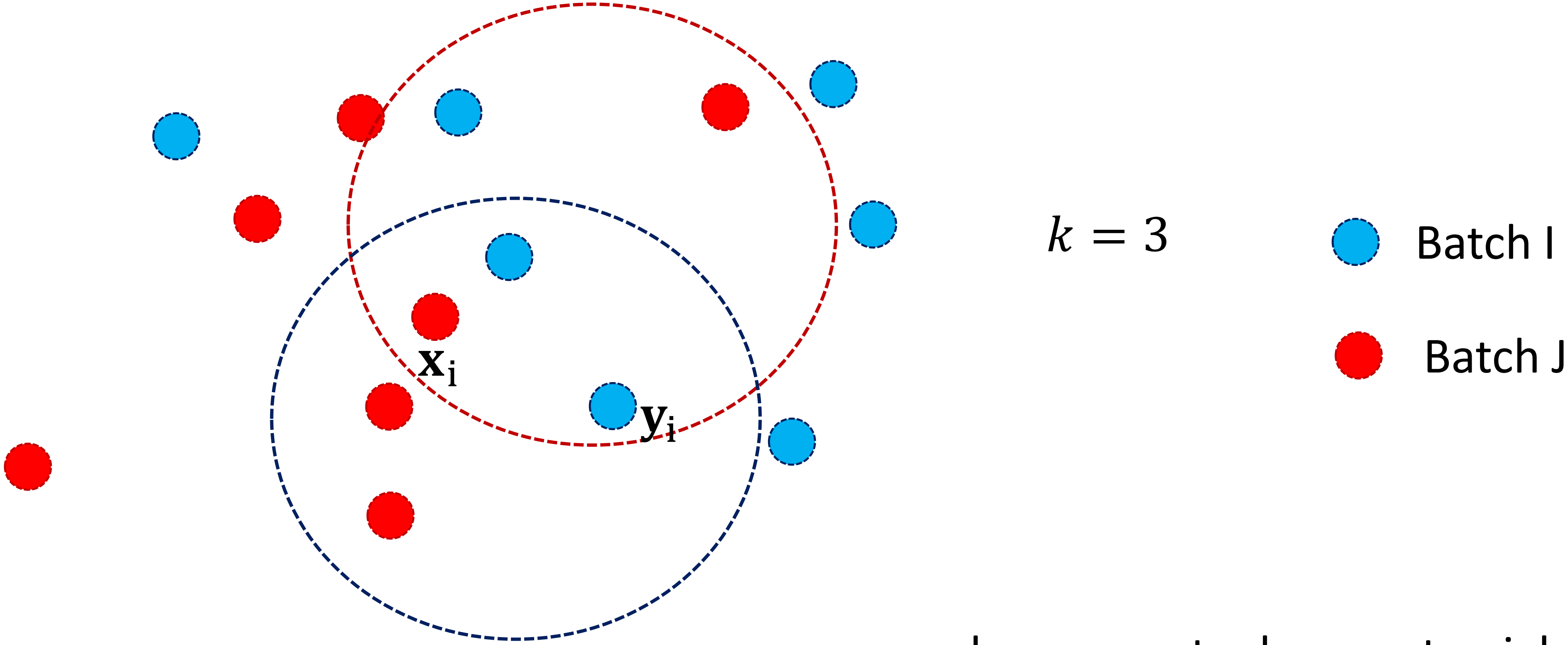
- L_2-norm normalization



$$\frac{\mathbf{x}}{\|\mathbf{x}\|_2}$$

Multiple Dataset Integration

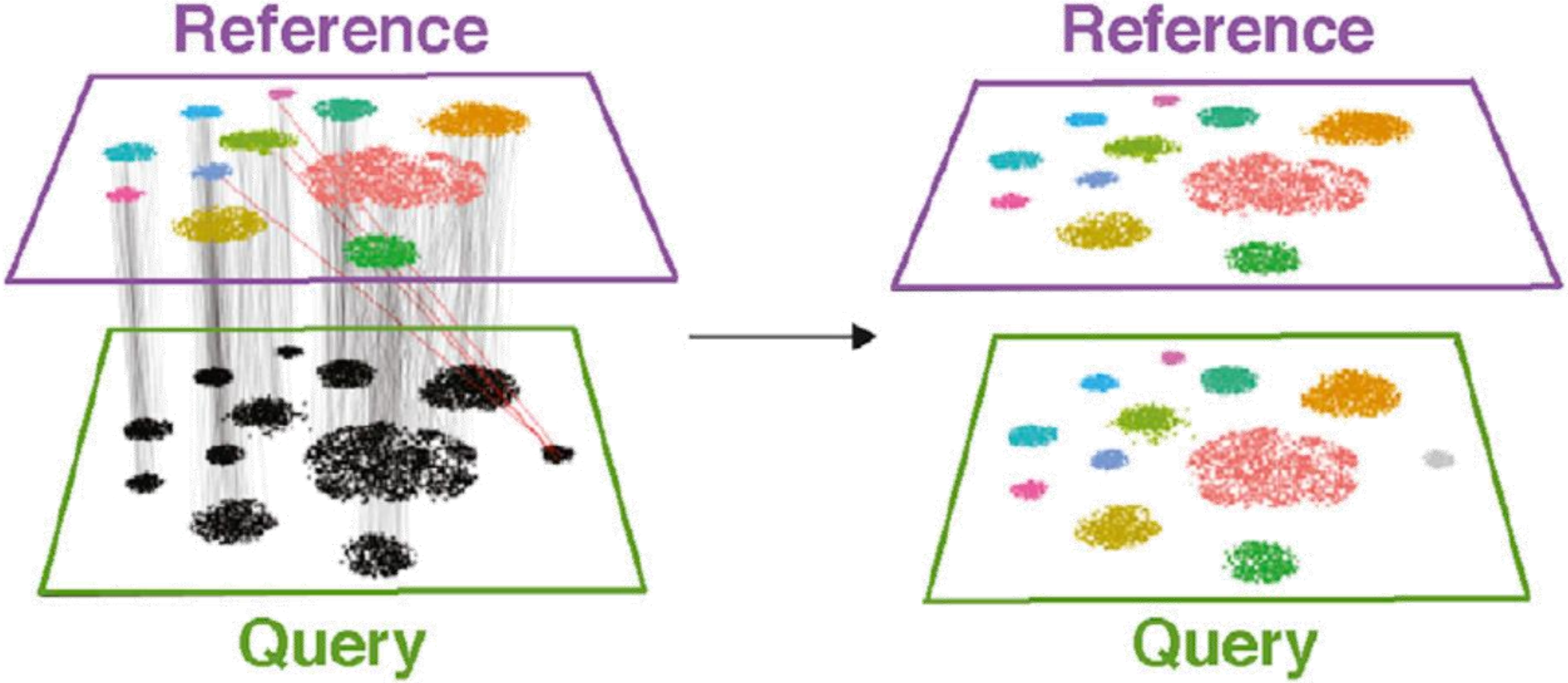
- Anchors (Mutual nearest neighbors)



$\mathbf{x}_i$ and $\mathbf{y}_i$ are mutual nearest neighbors

Multiple Dataset Integration

- Label transfer



References

- Seurat package: https://satijalab.org/seurat/
- Stuart, Tim et al. “Comprehensive Integration of Single-Cell Data” Cell, Volume 177, Issue 7, 1888 - 1902.e21, 2019
- How Exactly UMAP Works: https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668
- V. Blondel et al., “Fast unfolding of communities in large networks”, Journal of Statistical Mechanics: Theory and Experiment 2008