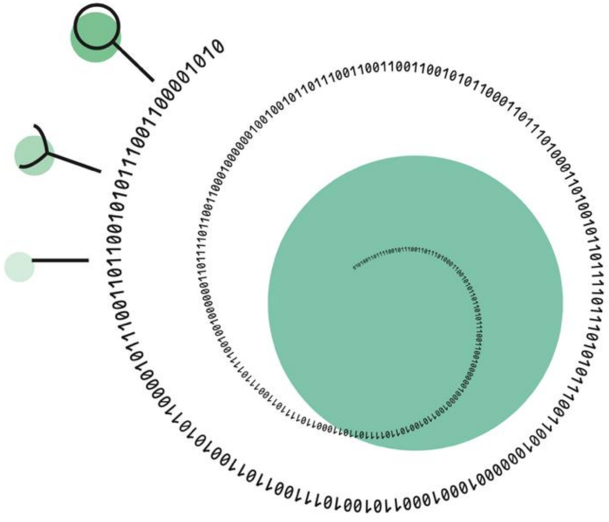
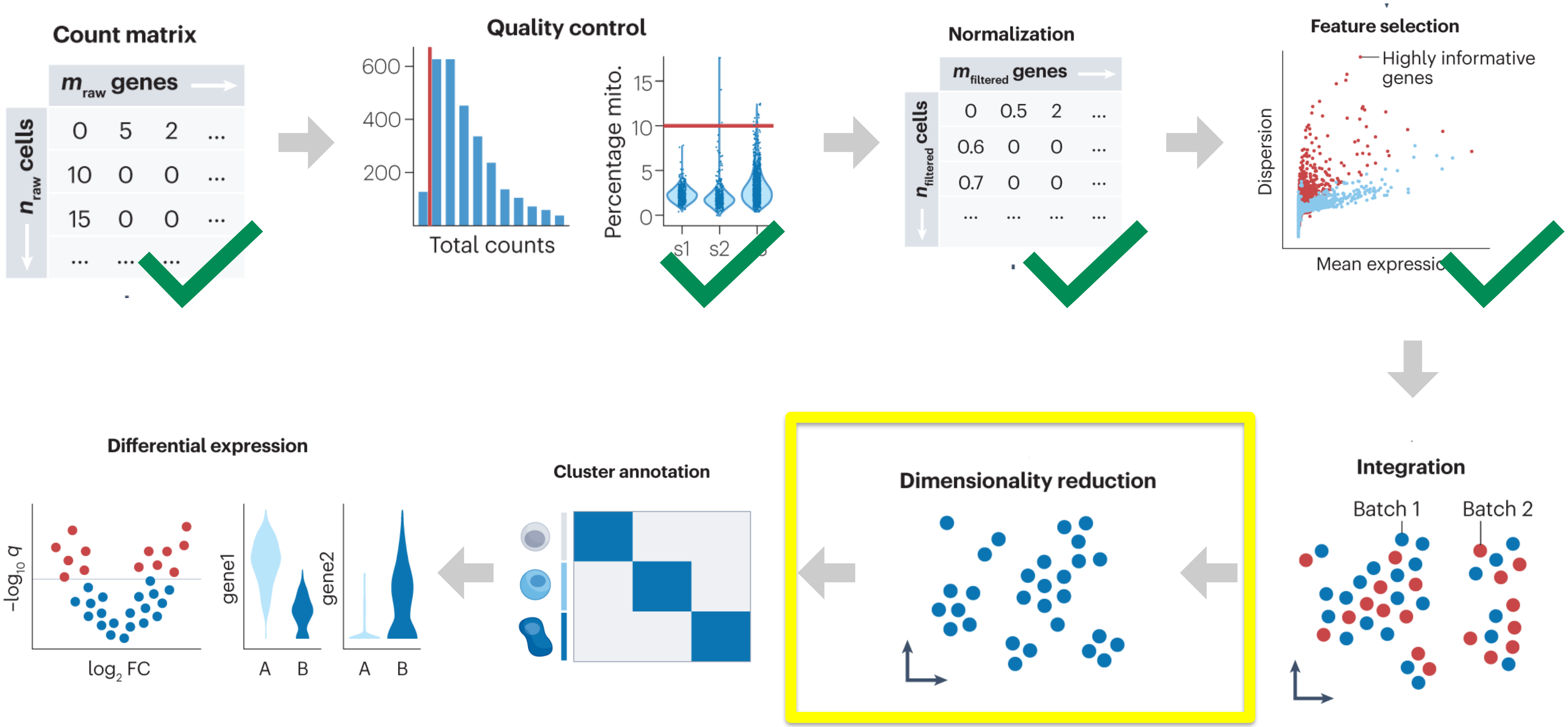


Single Dell Data Analysis Course

Dimensionality reduction 2: UMAP and graph-based clustering

Lisa Buchauer
Professor of Systems Biology of Infectious Diseases
Department of Infectious Diseases and Intensive Care
Charité - Universitätsmedizin Berlin

Today

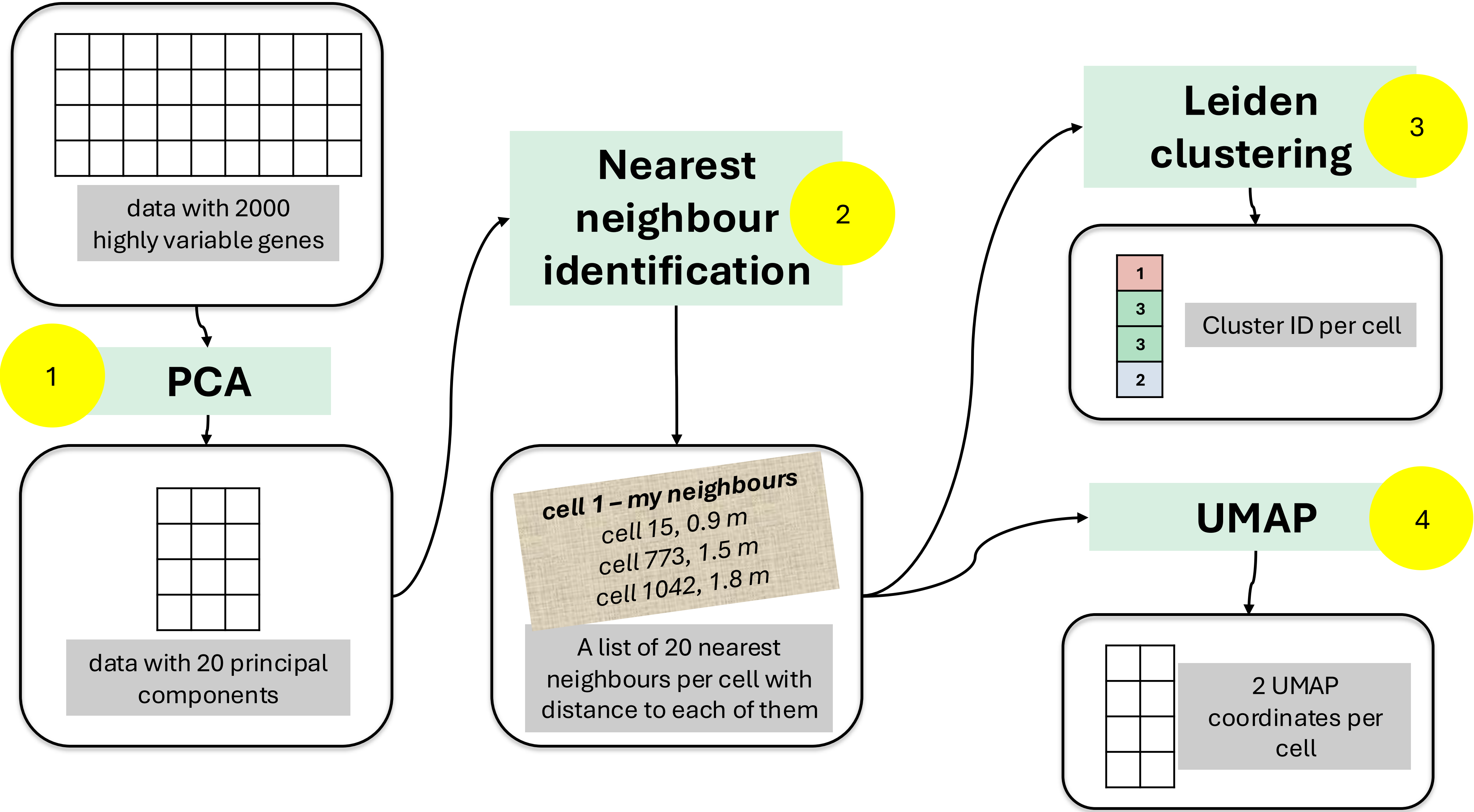


0	5	2	...
10	0	0	...
15	0	0	...
...	...	...	

0	0.5	2	...
0.6	0	0	...
0.7	0	0	...
...	...	...	...

Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w

Data types along the processing path



data with 2000 highly variable genes

1 PCA

data with 20 principal components

2 Nearest neighbour identification

cell 1 – my neighbours
cell 15, 0.9 m
cell 773, 1.5 m
cell 1042, 1.8 m

A list of 20 nearest neighbours per cell with distance to each of them

3 Leiden clustering

1
3
3
2

Cluster ID per cell

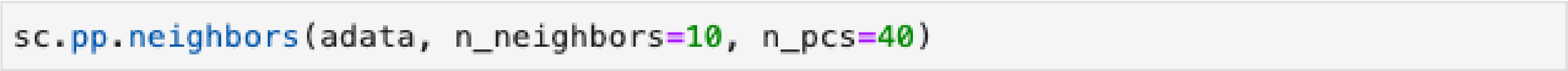
4 UMAP

2 UMAP coordinates per cell

2

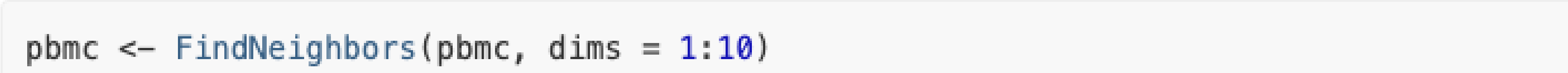
Finding Nearest neighbours

python



```
sc.pp.neighbors(adata, n_neighbors=10, n_pcs=40)
```

R



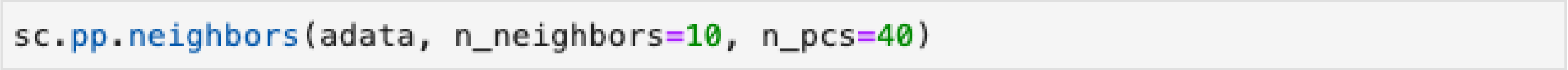
```
pbmc <- FindNeighbors(pbmc, dims = 1:10)
```

https://satijalab.org/seurat/articles/pbmc3k_tutorial
https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html

2

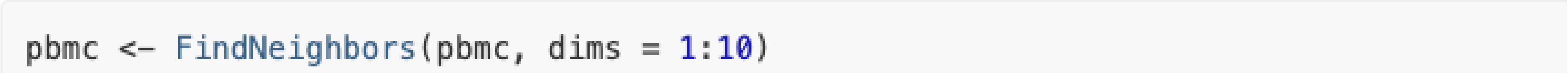
Finding Nearest neighbours

python

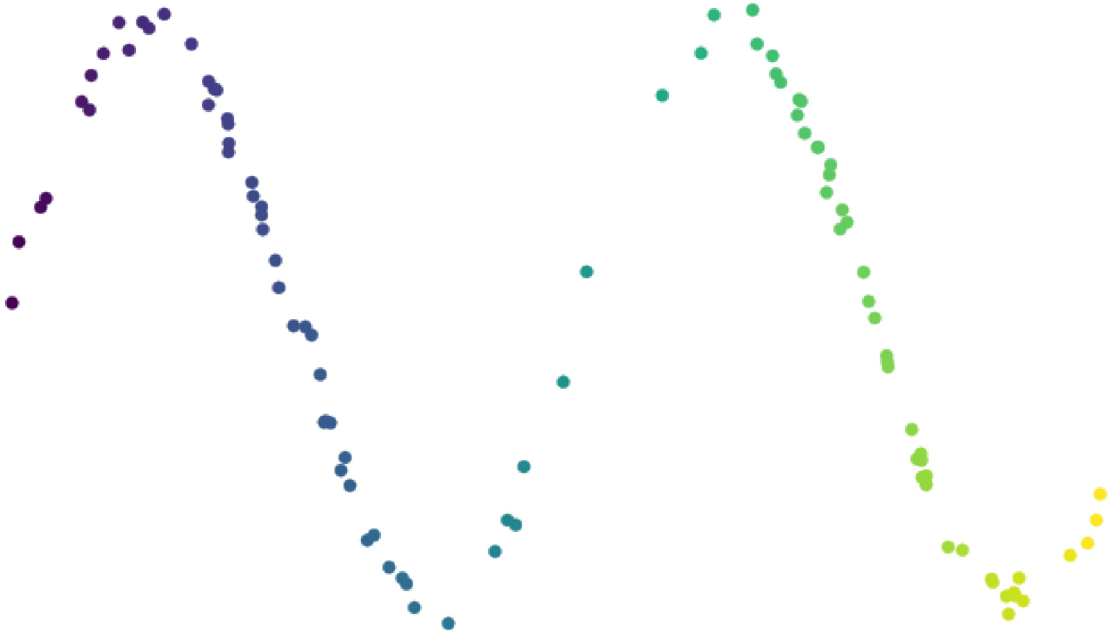


```
sc.pp.neighbors(adata, n_neighbors=10, n_pcs=40)
```

R



```
pbmc <- FindNeighbors(pbmc, dims = 1:10)
```

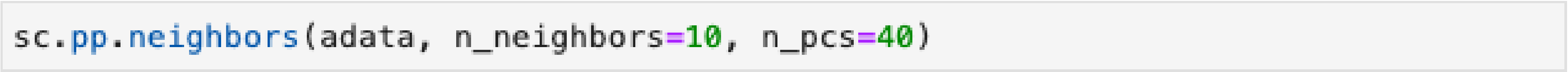


https://satijalab.org/seurat/articles/pbmc3k_tutorial
https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

2

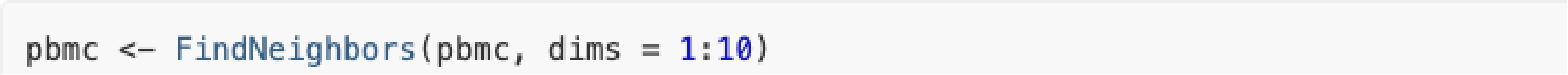
Finding Nearest neighbours

python

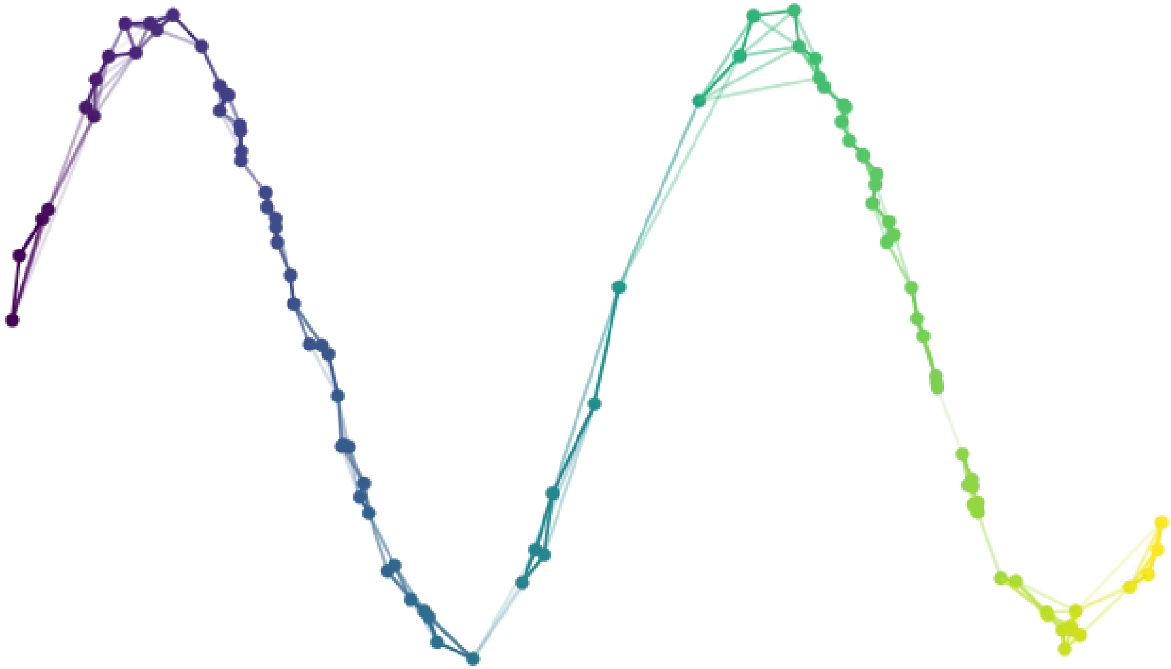


```
sc.pp.neighbors(adata, n_neighbors=10, n_pcs=40)
```

R



```
pbmc <- FindNeighbors(pbmc, dims = 1:10)
```

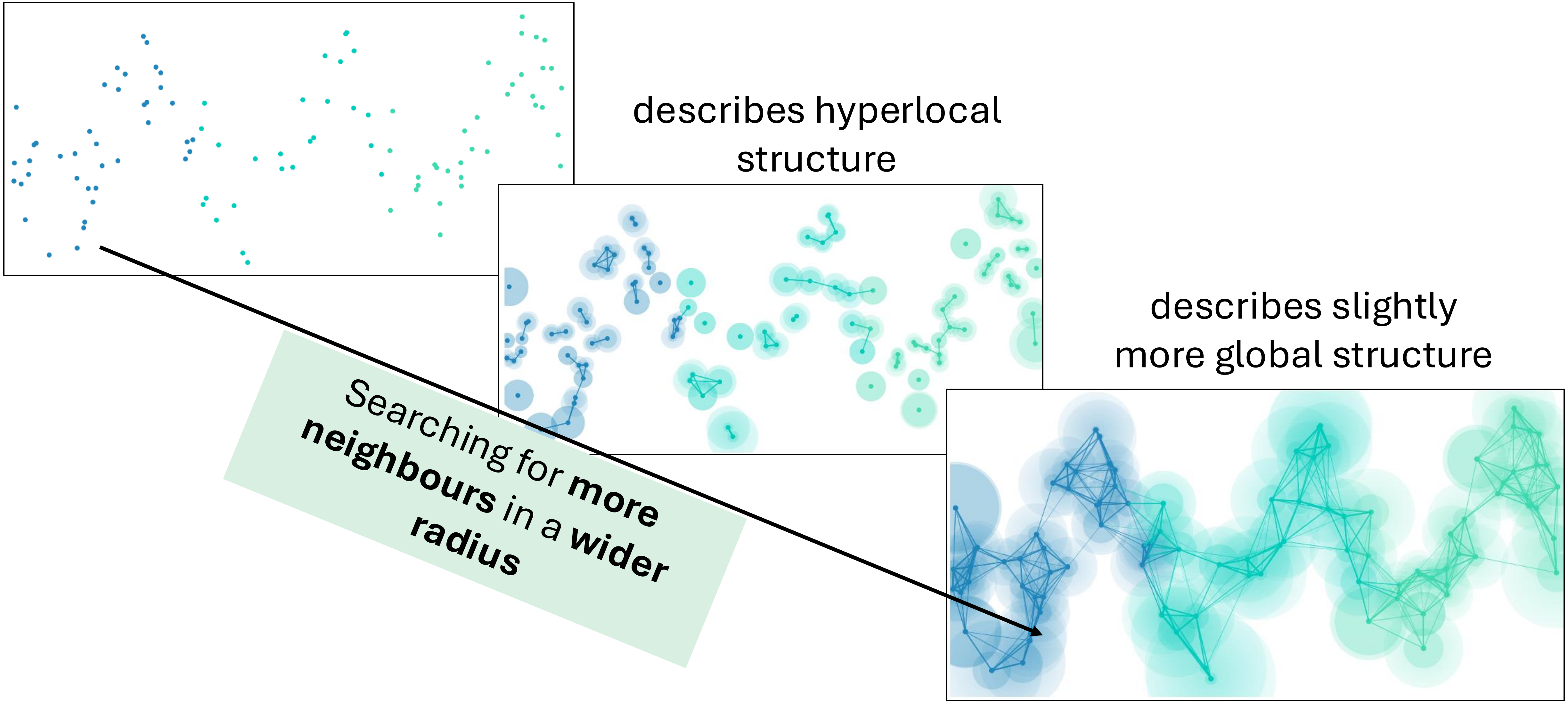


Idea: describing a dataset by way of each data point’s k nearest neighbours retains relevant structural information and allows efficient computations

https://satijalab.org/seurat/articles/pbmc3k_tutorial
https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

2

How many neighbours are selected impacts what information is retained in the graph



describes hyperlocal structure

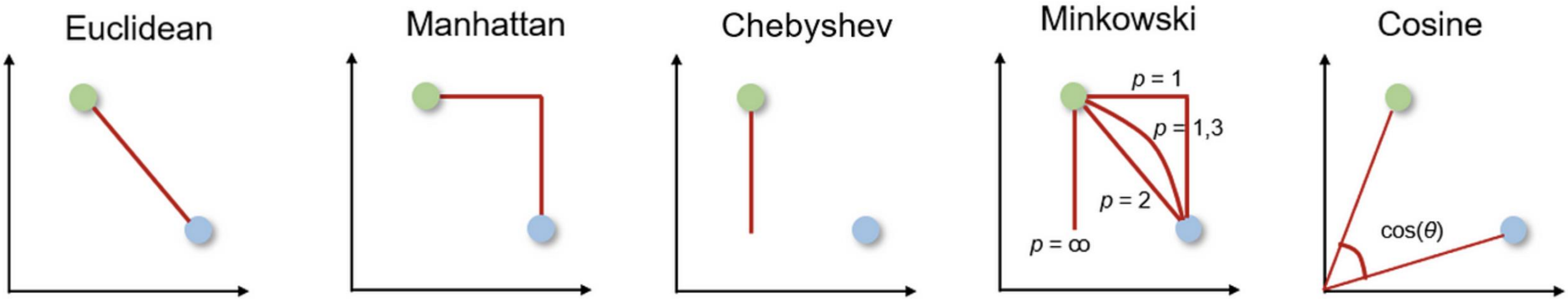
describes slightly more global structure

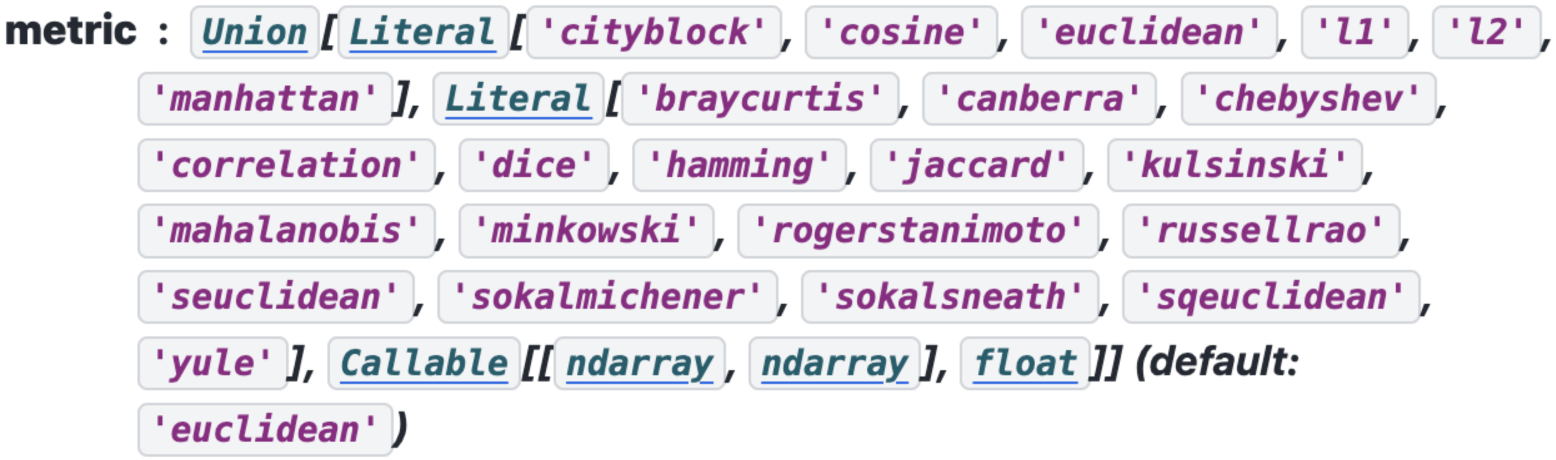
Searching for **more neighbours** in a **wider radius**

https://pair-code.github.io/understanding-umap/

2

Side topic: The role of the distance metrics

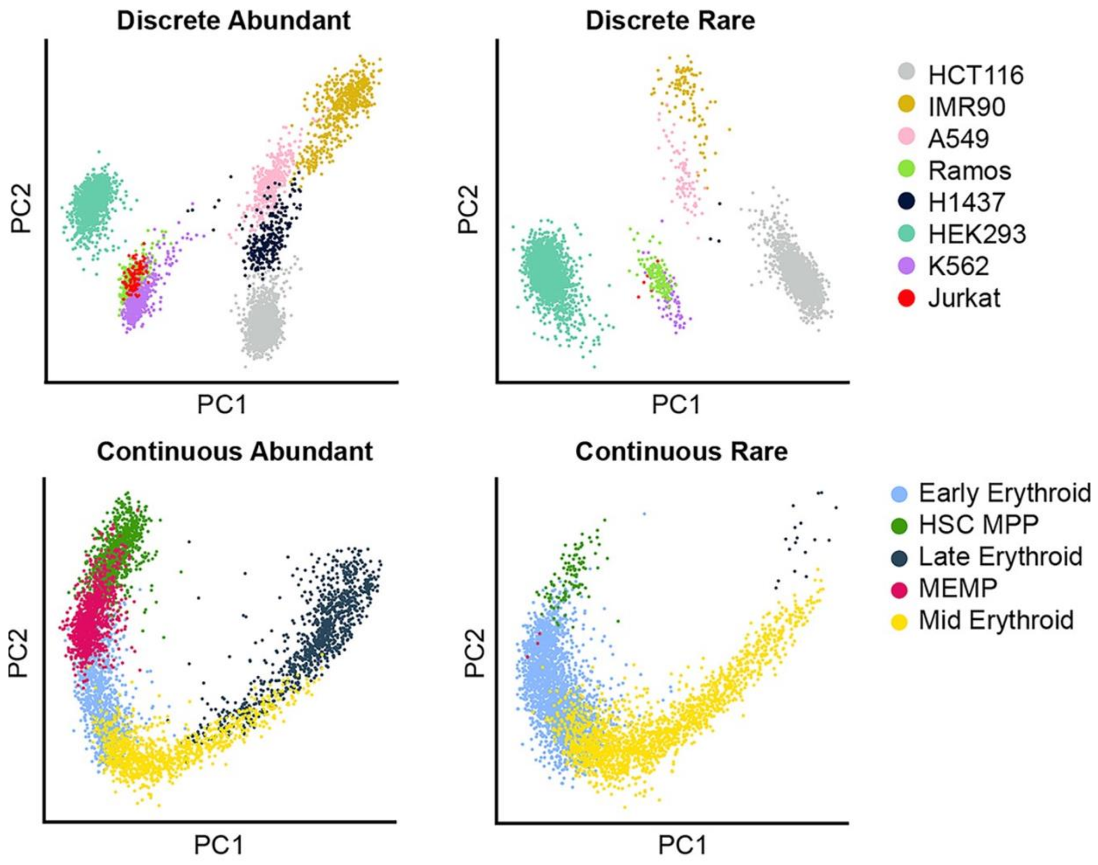


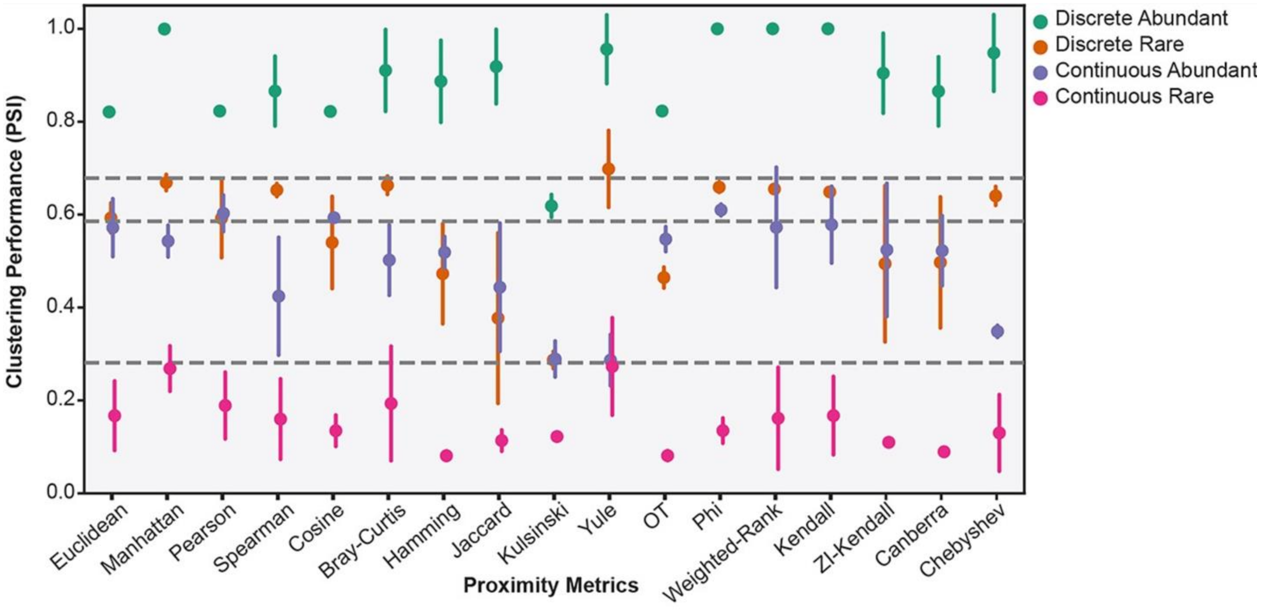

metric : Union[Literal['cityblock', 'cosine', 'euclidean', 'l1', 'l2', 'manhattan'], Literal['braycurtis', 'canberra', 'chebyshev', 'correlation', 'dice', 'hamming', 'jaccard', 'kulsinski', 'mahalanobis', 'minkowski', 'rogerstanimoto', 'russellrao', 'seuclidean', 'sokalmichener', 'sokalsneath', 'sqeuclidean', 'yule'], Callable[[ndarray, ndarray], float]] (default: 'euclidean')

https://scanpy.readthedocs.io/en/stable/api/generated/scanpy.pp.neighbors.html

2

Side topic: The role of the distance metrics

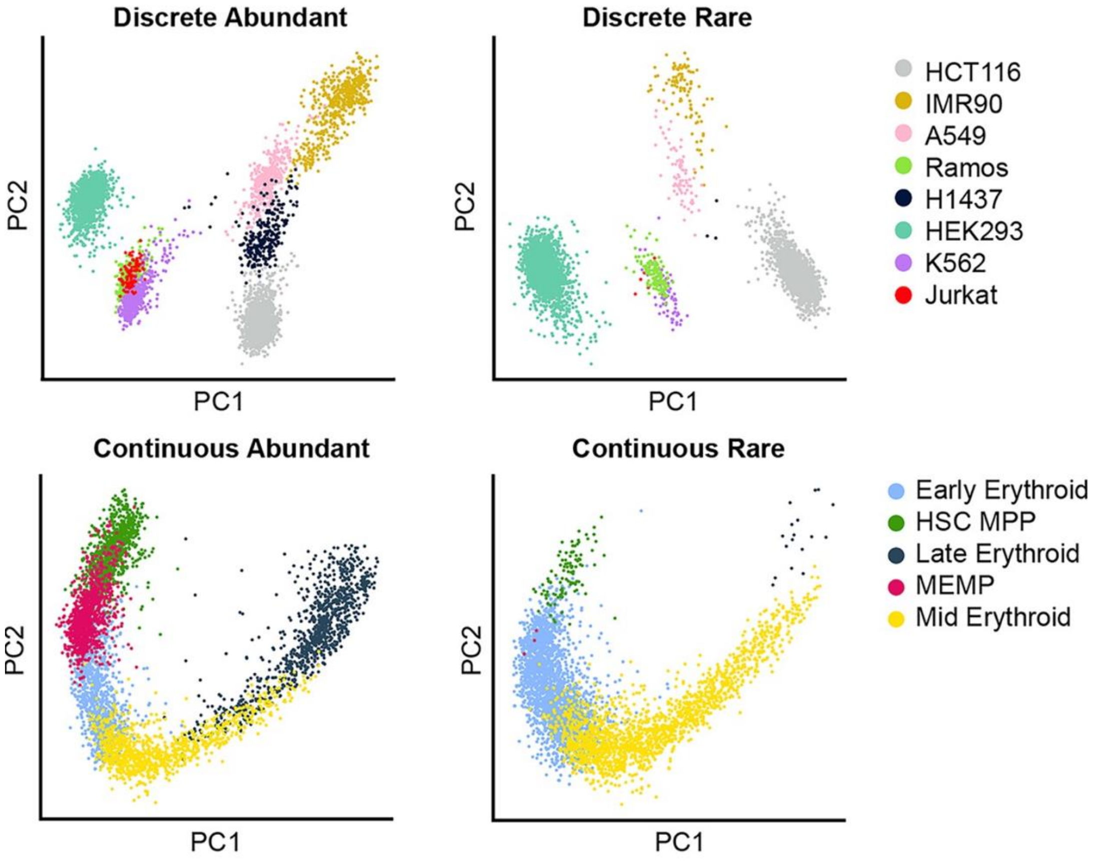


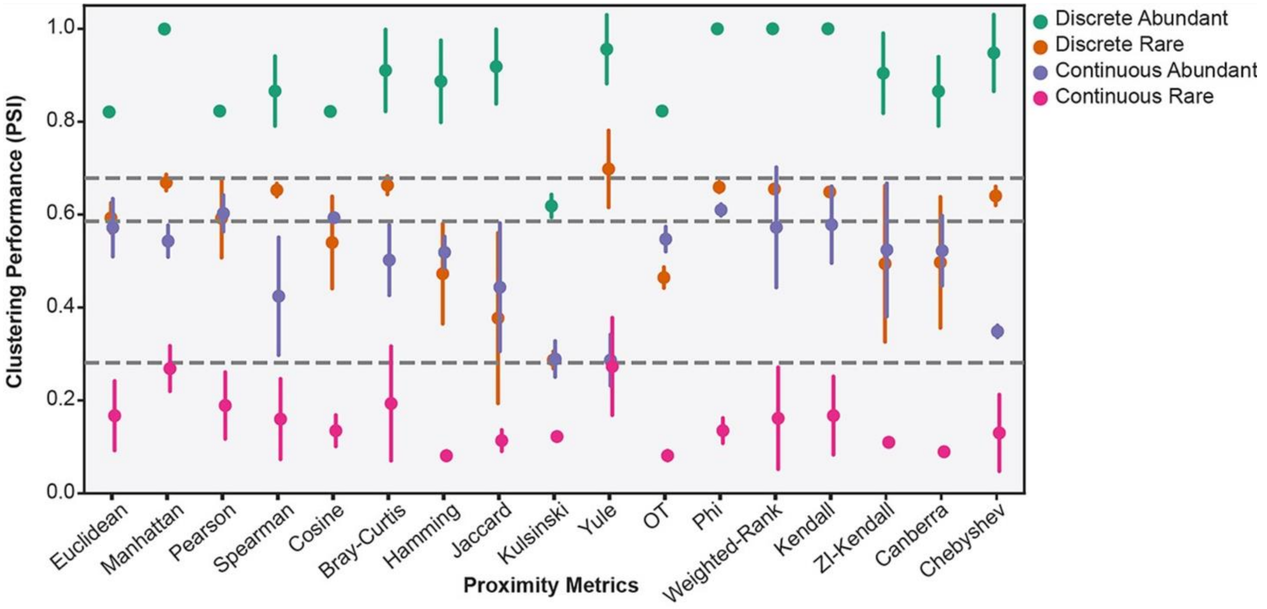


https://academic.oup.com/bib/article/23/6/bbac387/6712300

2

Side topic: The role of the distance metrics





Choice of distance metric has a bearing on results – but in practice, the default (Euclidean in PC-space) is almost always used

https://academic.oup.com/bib/article/23/6/bbac387/6712300

2

Nearest neighbours search results in an adjacency matrix

cell 1 – my neighbours
cell 15, 0.9 m
cell 773, 1.5 m
cell 1042, 1.8 m

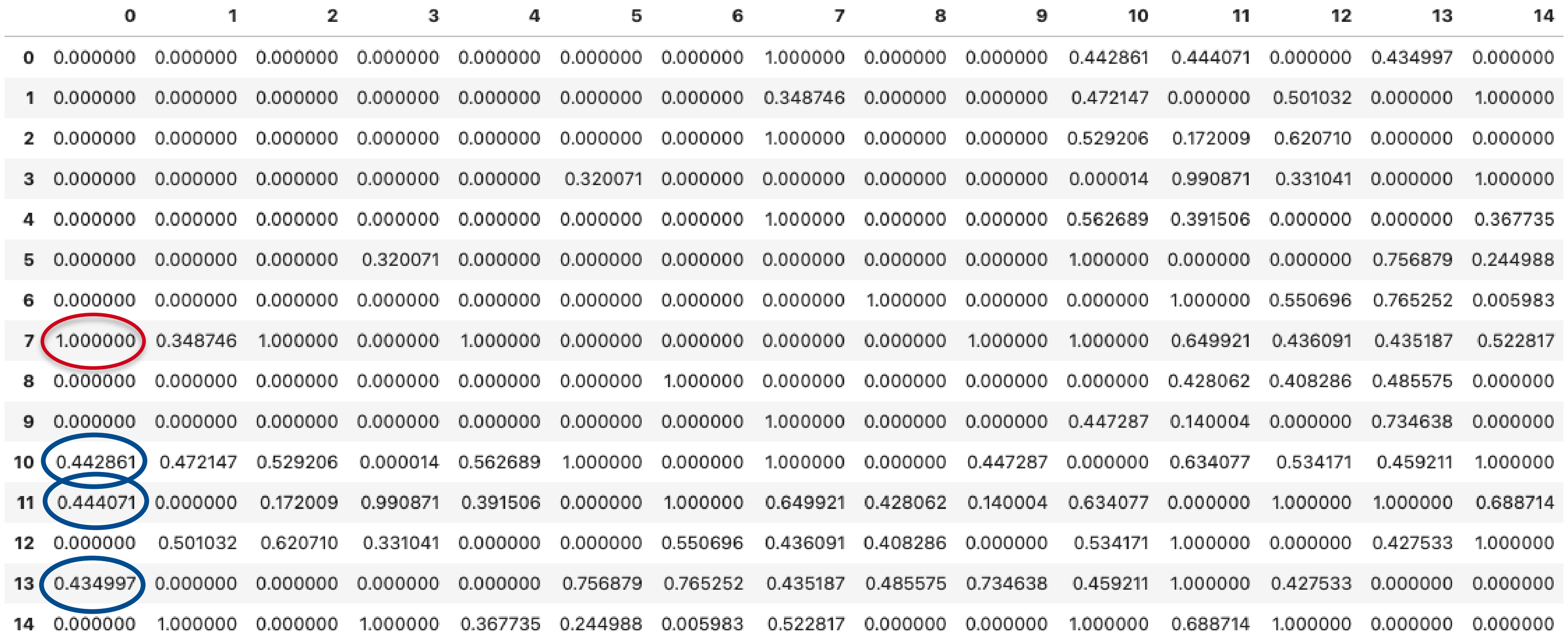
adjacency matrix for
2 nearest neighbours

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5
Cell 1	1	1	0	1	0
Cell 2	1	1	1	0	0
Cell 3	0	1	1	0	1
Cell 4	1	0	0	1	1
Cell 5	0	0	1	1	1

2

Finding Nearest neighbours

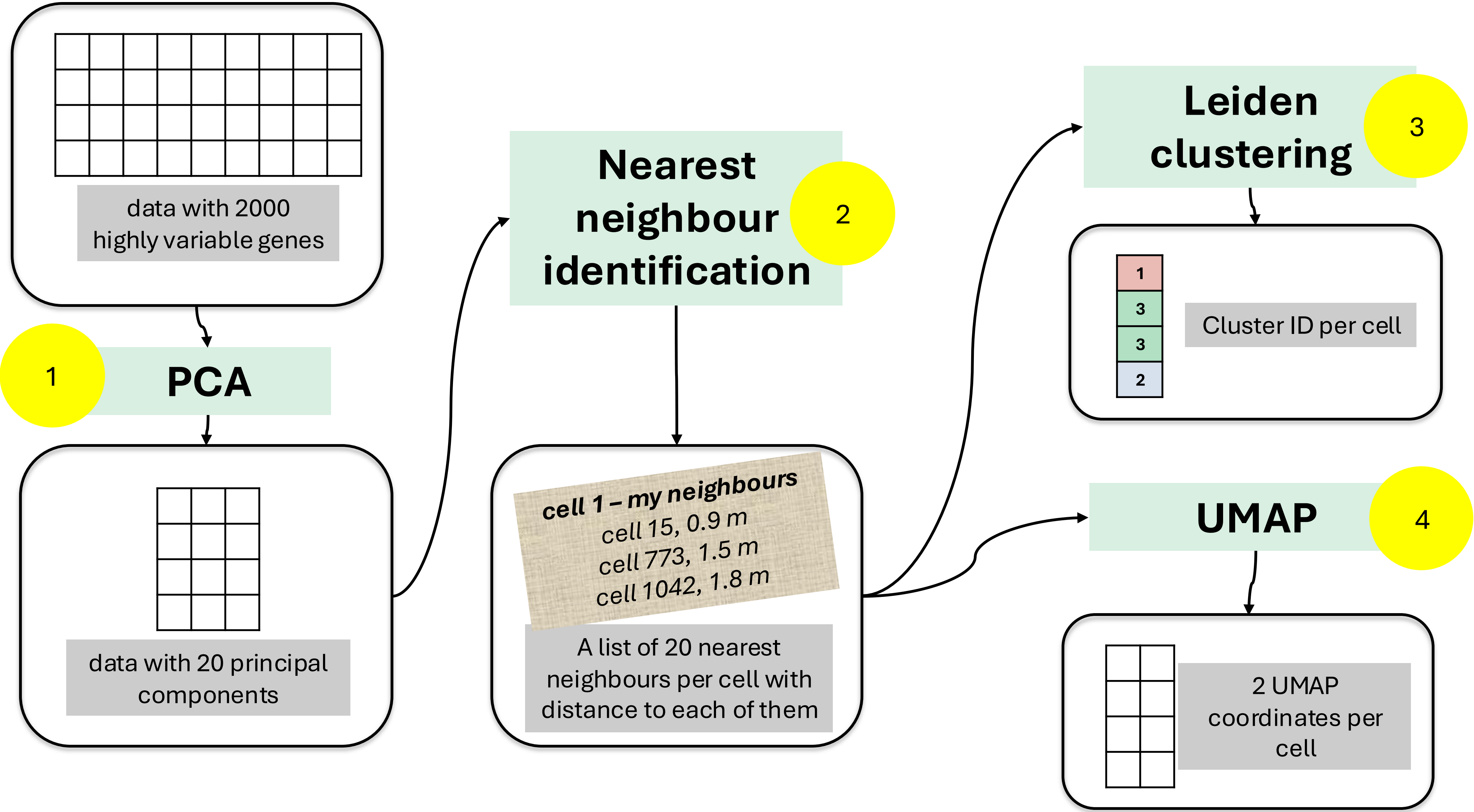
Slightly more information can be retained with a “fuzzy adjacency matrix”, in which the nearest neighbour gets connectivity 1 and additional neighbours get lower connectivities based on distances.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.442861	0.444071	0.000000	0.434997	0.000000
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.348746	0.000000	0.000000	0.472147	0.000000	0.501032	0.000000	1.000000
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.529206	0.172009	0.620710	0.000000	0.000000
3	0.000000	0.000000	0.000000	0.000000	0.000000	0.320071	0.000000	0.000000	0.000000	0.000000	0.000014	0.990871	0.331041	0.000000	1.000000
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.562689	0.391506	0.000000	0.000000	0.367735
5	0.000000	0.000000	0.000000	0.320071	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.756879	0.244988
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	0.550696	0.765252	0.005983
7	1.000000	0.348746	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	0.649921	0.436091	0.435187	0.522817
8	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.428062	0.408286	0.485575	0.000000
9	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.447287	0.140004	0.000000	0.734638	0.000000
10	0.442861	0.472147	0.529206	0.000014	0.562689	1.000000	0.000000	1.000000	0.000000	0.447287	0.000000	0.634077	0.534171	0.459211	1.000000
11	0.444071	0.000000	0.172009	0.990871	0.391506	0.000000	1.000000	0.649921	0.428062	0.140004	0.634077	0.000000	1.000000	1.000000	0.688714
12	0.000000	0.501032	0.620710	0.331041	0.000000	0.000000	0.550696	0.436091	0.408286	0.000000	0.534171	1.000000	0.000000	0.427533	1.000000
13	0.434997	0.000000	0.000000	0.000000	0.000000	0.756879	0.765252	0.435187	0.485575	0.734638	0.459211	1.000000	0.427533	0.000000	0.000000
14	0.000000	1.000000	0.000000	1.000000	0.367735	0.244988	0.005983	0.522817	0.000000	0.000000	1.000000	0.688714	1.000000	0.000000	0.000000

3k PBMC dataset → downsampled to 15 cells → computed 4 nearest neighbours with the UMAP neighbour algorithm (via scanpy)

Data types along the processing path



data with 2000 highly variable genes

PCA

data with 20 principal components

Nearest neighbour identification

cell 1 – my neighbours
cell 15, 0.9 m
cell 773, 1.5 m
cell 1042, 1.8 m

A list of 20 nearest neighbours per cell with distance to each of them

Leiden clustering

Cluster ID per cell

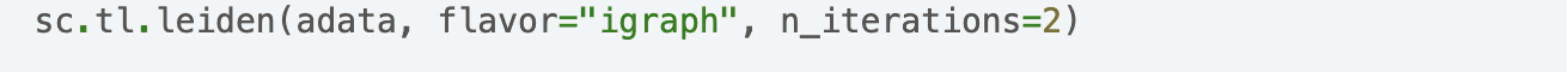
UMAP

2 UMAP coordinates per cell

3

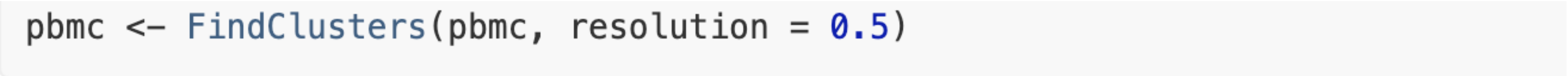
Clustering / community-detection with the Leiden algorithm

python

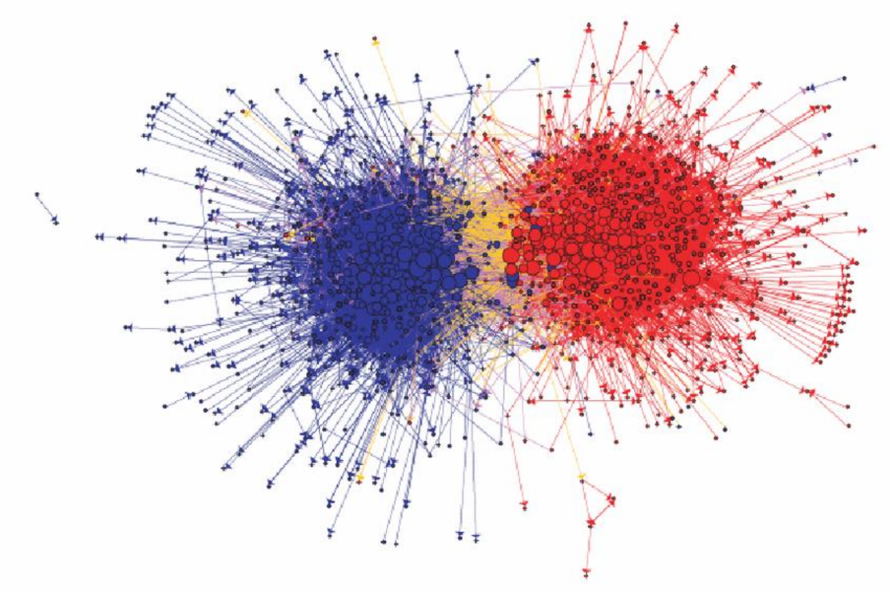


```
sc.tl.leiden(adata, flavor="igraph", n_iterations=2)
```

R



```
pbmc <- FindClusters(pbmc, resolution = 0.5)
```

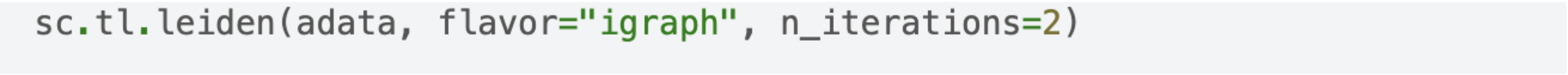


Arms et al.: “From Wayback Machine to Yesternet: New Opportunitites for Social Science”, https://www.cs.cornell.edu/wya/papers/Yesternet.pdf

3

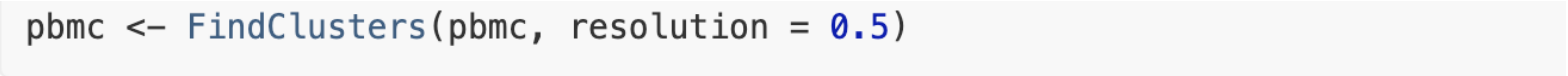
Clustering / community-detection with the Leiden algorithm

python



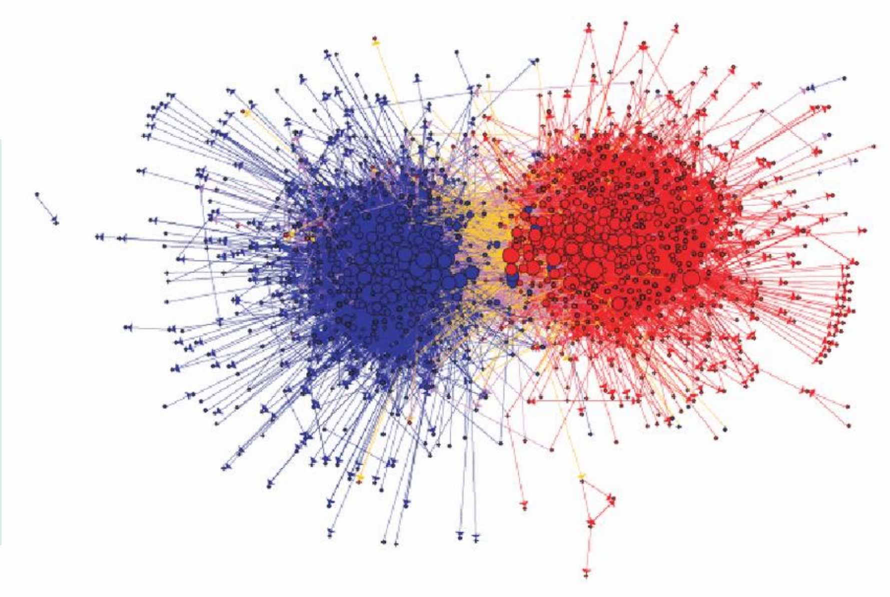
```python
sc.tl.leiden(adata, flavor="igraph", n_iterations=2)
```

R



```r
pbmc <- FindClusters(pbmc, resolution = 0.5)
```

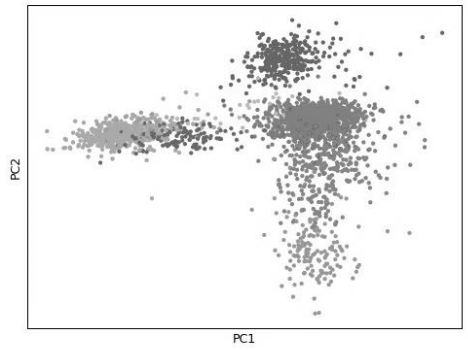
US political blogs
red – conservative,
blue – liberal,
edges represent direct hyperlinks



Arms et al.: “From Wayback Machine to Yesternet: New Opportunitites for Social Science”, https://www.cs.cornell.edu/wya/papers/Yesternet.pdf

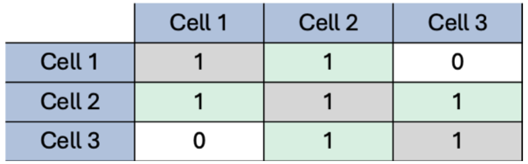
3

Clustering / community-detection with the Leiden algorithm



“raw” data

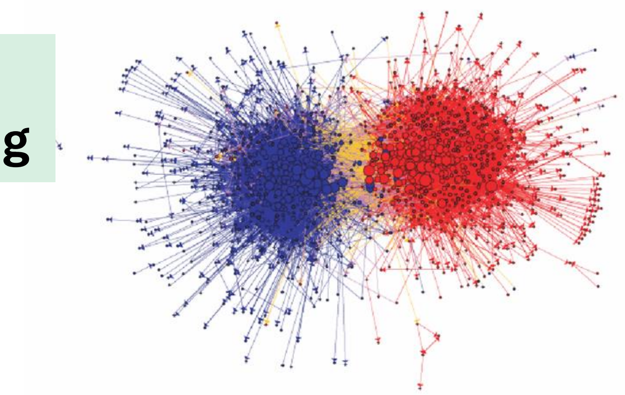
neighbour search



	Cell 1	Cell 2	Cell 3
Cell 1	1	1	0
Cell 2	1	1	1
Cell 3	0	1	1

adjacency matrix
(graph edges)

Leiden clustering



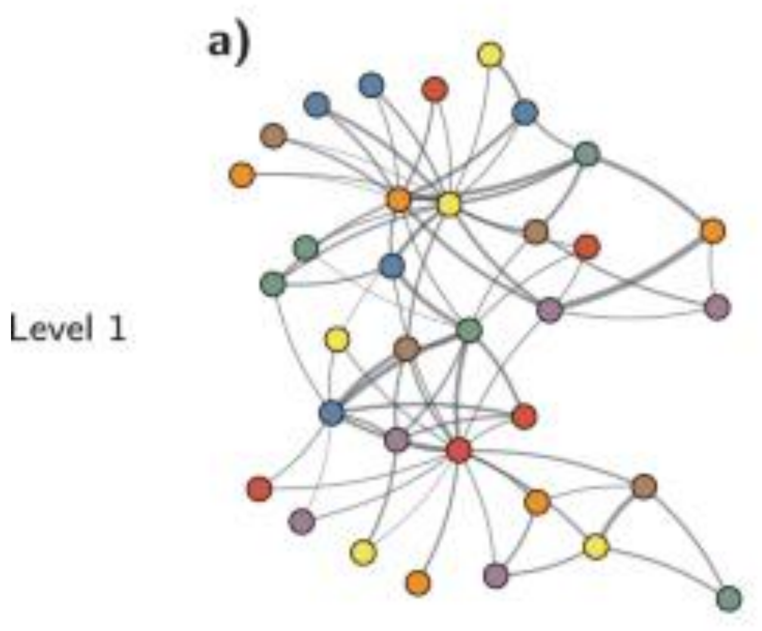
community assignment
(colors)

Leiden (and Louvain) algorithms optimize **modularity** –
the density of connections within a community compared to
between communitites

3

Clustering / community-detection with the Leiden algorithm

modularity – the density of connections within a community compared to between communitites

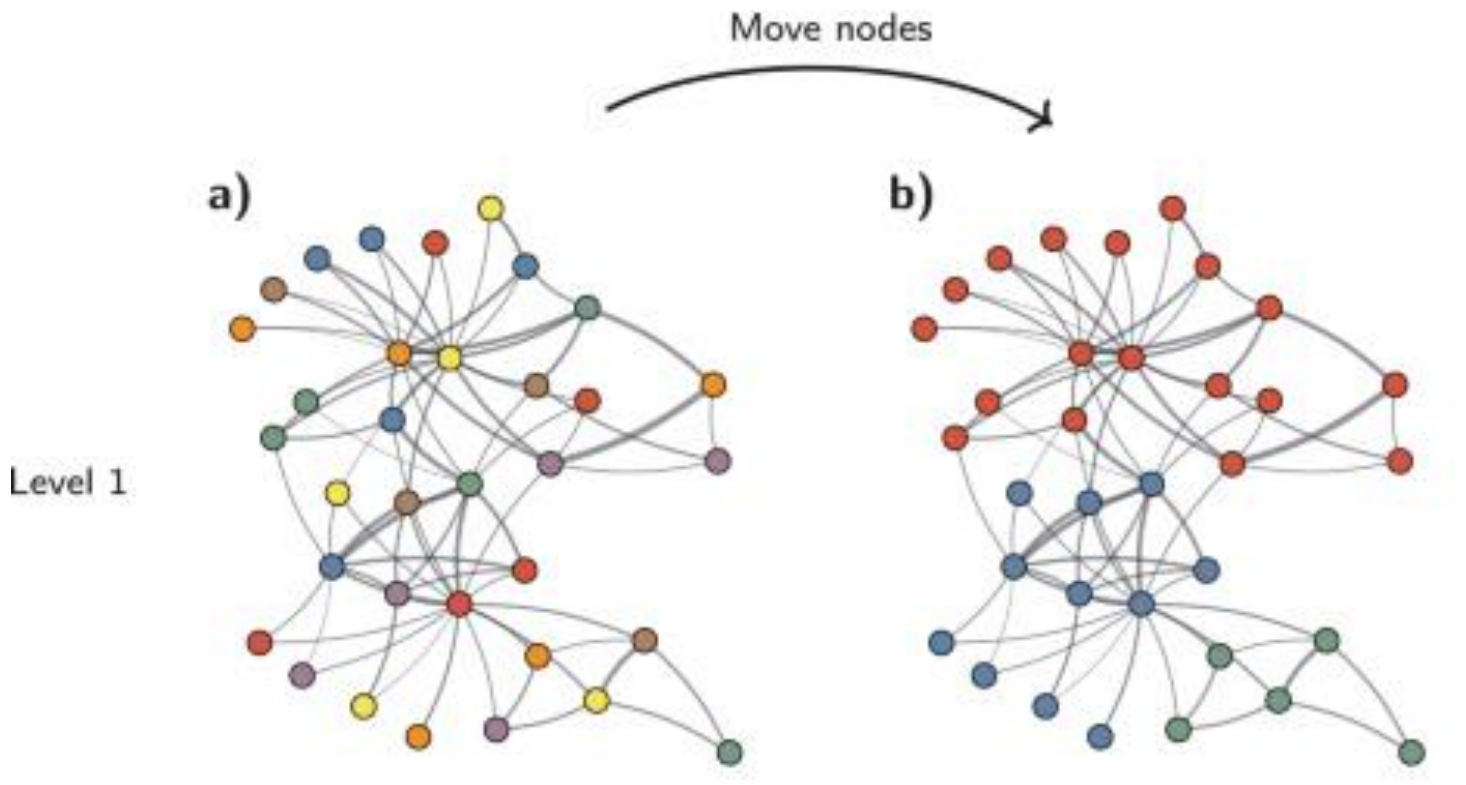


Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

3

Clustering / community-detection with the Leiden algorithm

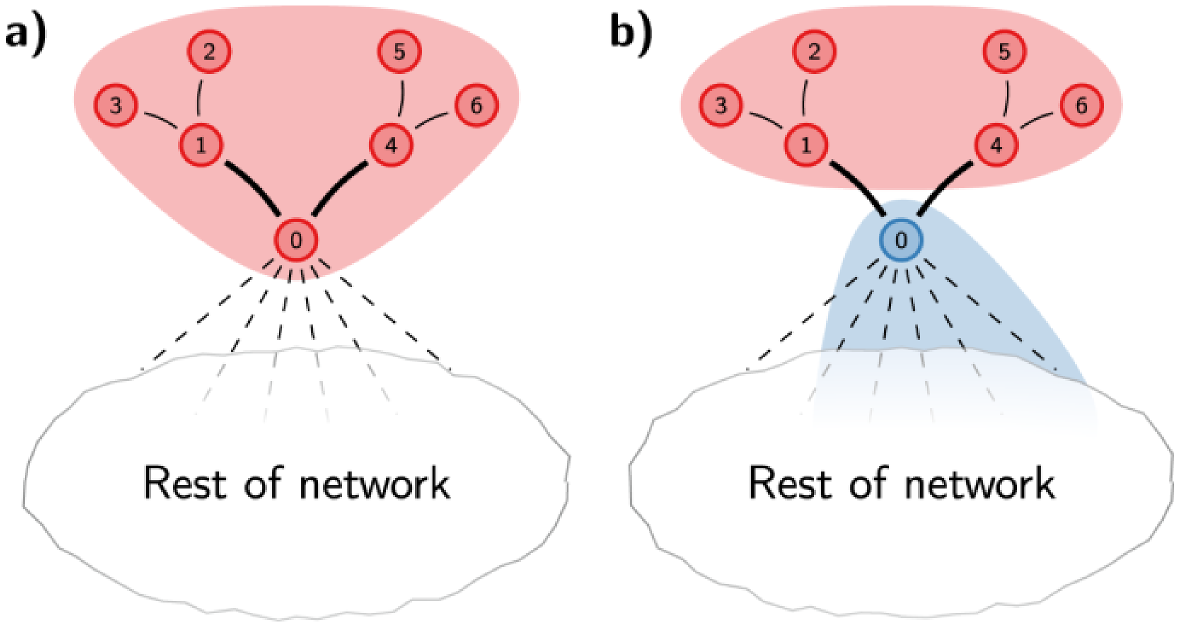
modularity – the density of connections within a community compared to between communitites



Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

3

Choose Leiden over Louvain.



*Louvain: after node 0 moves to another community, 1-6 stay in one cluster even though they are no longer connected

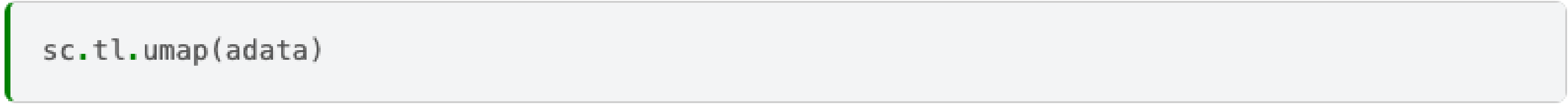
- Louvain risks returning “disconnected communities”*, fixed in Leiden
- Leiden runs faster

Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

4

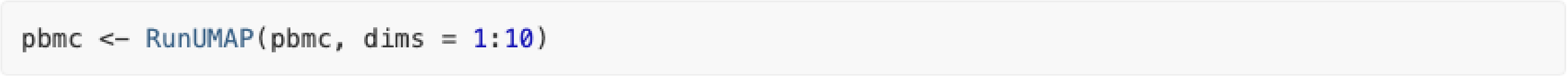
Non-linear dimension reduction – example: UMAP (Uniform Manifold Approximation and Projection)

python



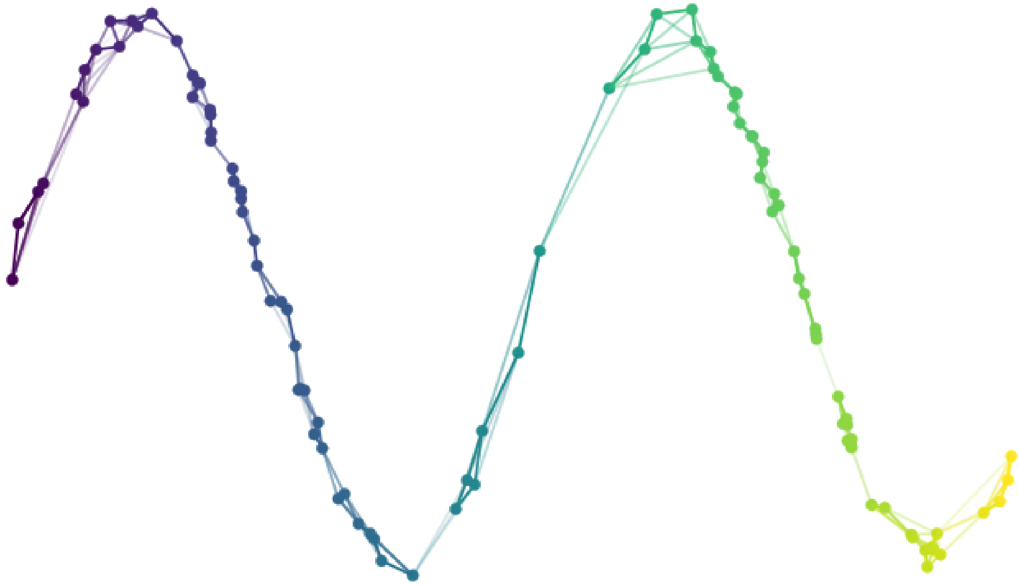
```
sc.tl.umap(adata)
```

R



```
pbmc <- RunUMAP(pbmc, dims = 1:10)
```

Goal
Embed the cellular graph into lower dimensions in such a way that neighbourhoods (= **local structure**) are preserved.



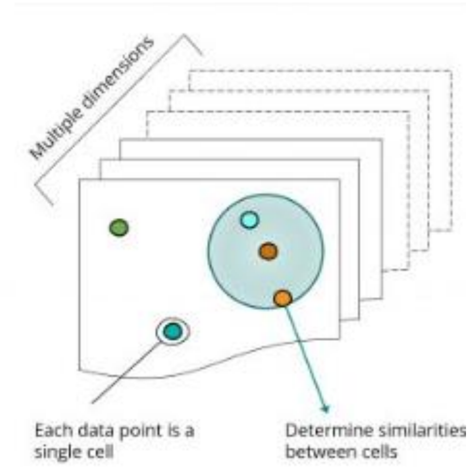
https://satijalab.org/seurat/articles/pbmc3k_tutorial
https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

4

Non-linear dimension reduction – example: UMAP (Uniform Manifold Approximation and Projection)

Phase 1 – in high D

- Determine similarities between cells in high-dim space (= fuzzy adjacency matrix)


Multiple dimensions
Each data point is a single cell
Determine similarities between cells

https://www.scdiscoveries.com/blog/knowledge/what-is-a-umap-plot/

4

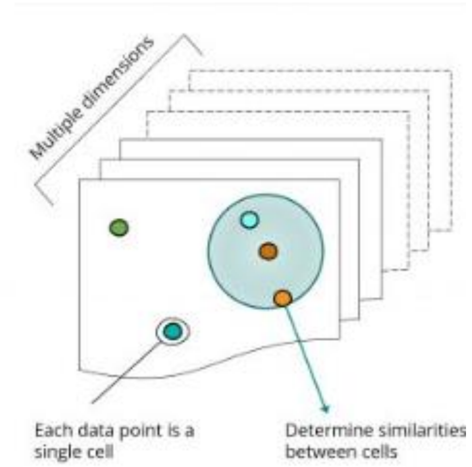
Non-linear dimension reduction – example: UMAP (Uniform Manifold Approximation and Projection)

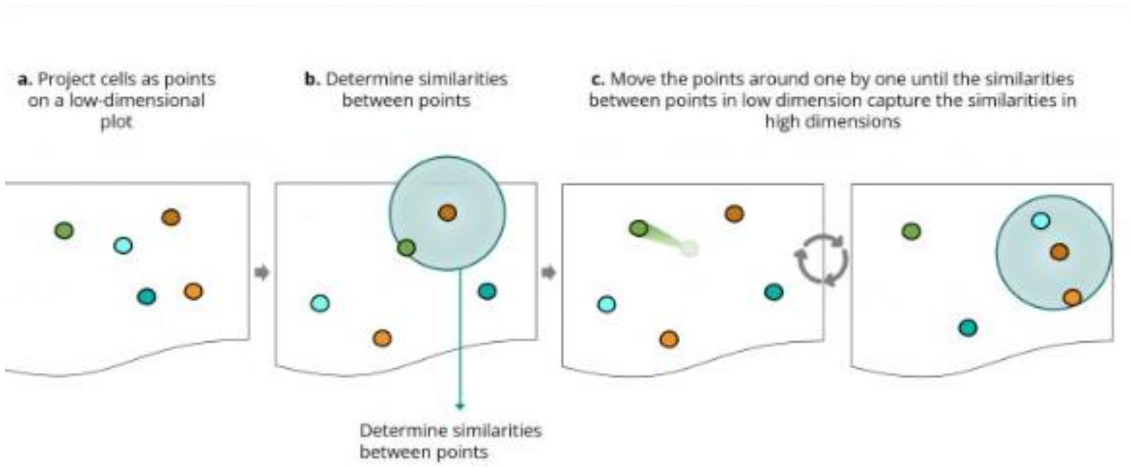
Phase 1 – in high D

- Determine similarities between cells in high-dim space (= fuzzy **adjacency matrix**)

Phase 2 – in 2/3 D

- Project cells into low D
- Determine similarities between cells in low D
- Move cells around until adjacency matrix resembles the one from high D

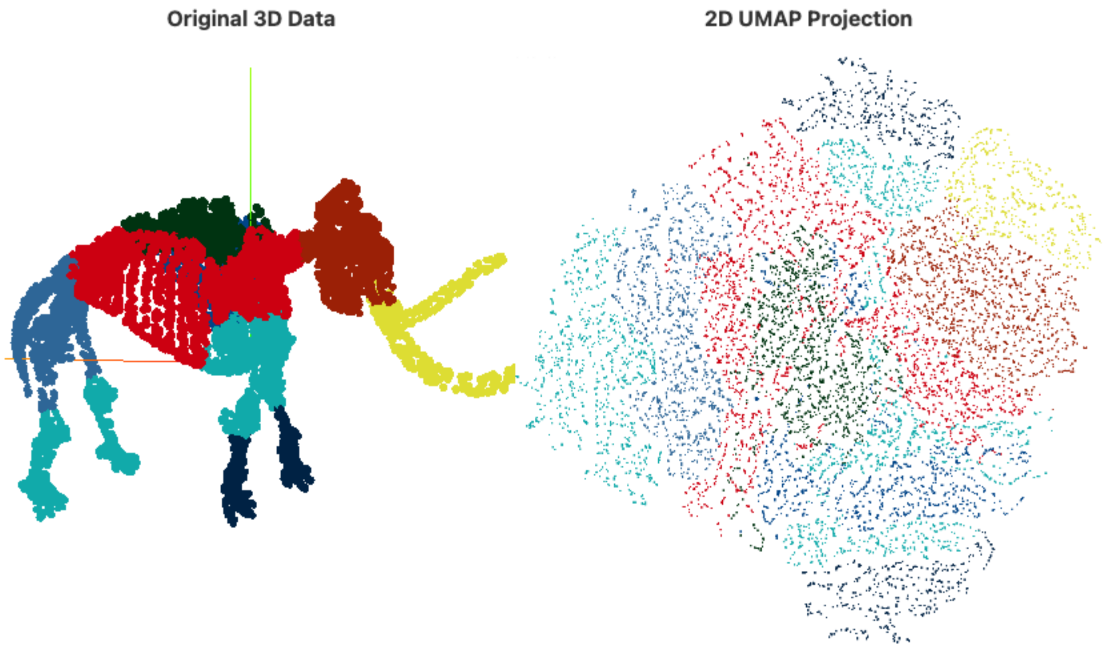

Multiple dimensions
Each data point is a single cell
Determine similarities between cells


a. Project cells as points on a low-dimensional plot
b. Determine similarities between points
c. Move the points around one by one until the similarities between points in low dimension capture the similarities in high dimensions
Determine similarities between points

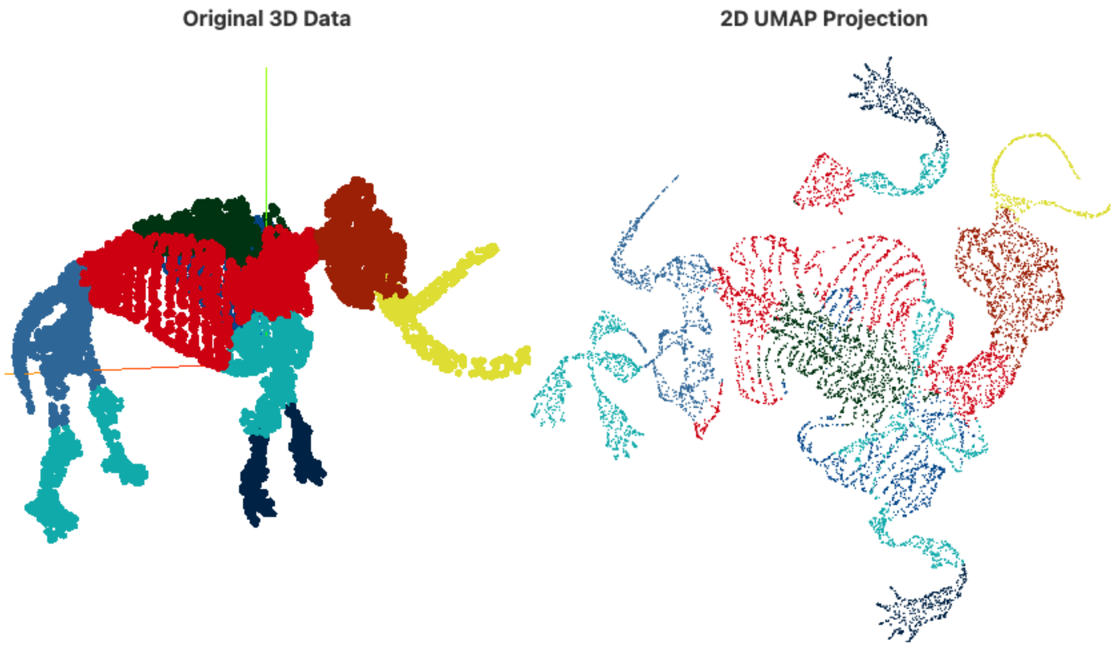
https://www.scdiscoveries.com/blog/knowledge/what-is-a-umap-plot/

4

The most important UMAP parameter:
the number of neighbours with which to build the graph



n_neighbors = 5



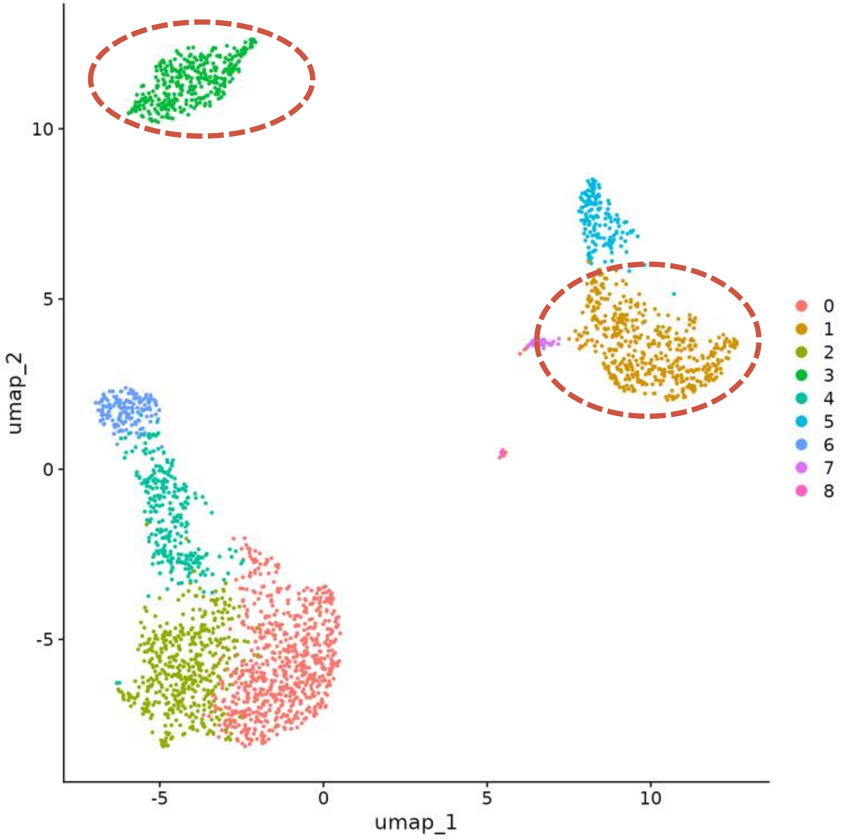
n_neighbors = 200

https://pair-code.github.io/understanding-umap/
VISIT THIS BLOG POST!

4

Interpreting tSNE and UMAP in the single-cell world

The green cluster is less variable than the beige-brown cluster because it is smaller.

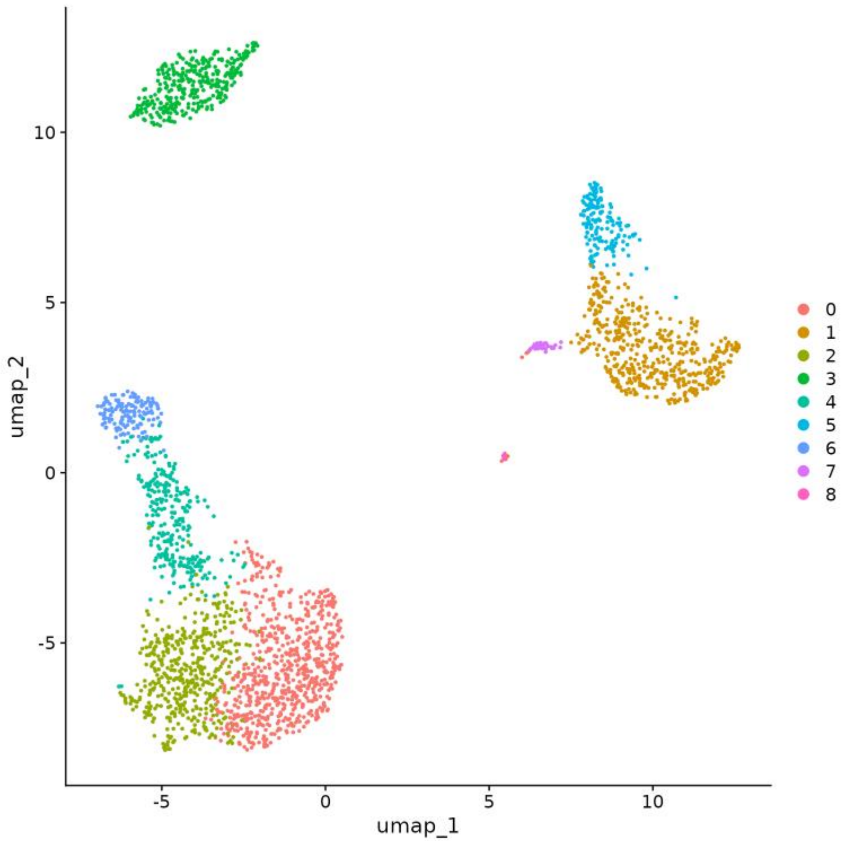


https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#perform-linear-dimensional-reduction

4

Interpreting tSNE and UMAP in the single-cell world

The green cluster is less variable than the beige-brown cluster because it is smaller.



You shall not

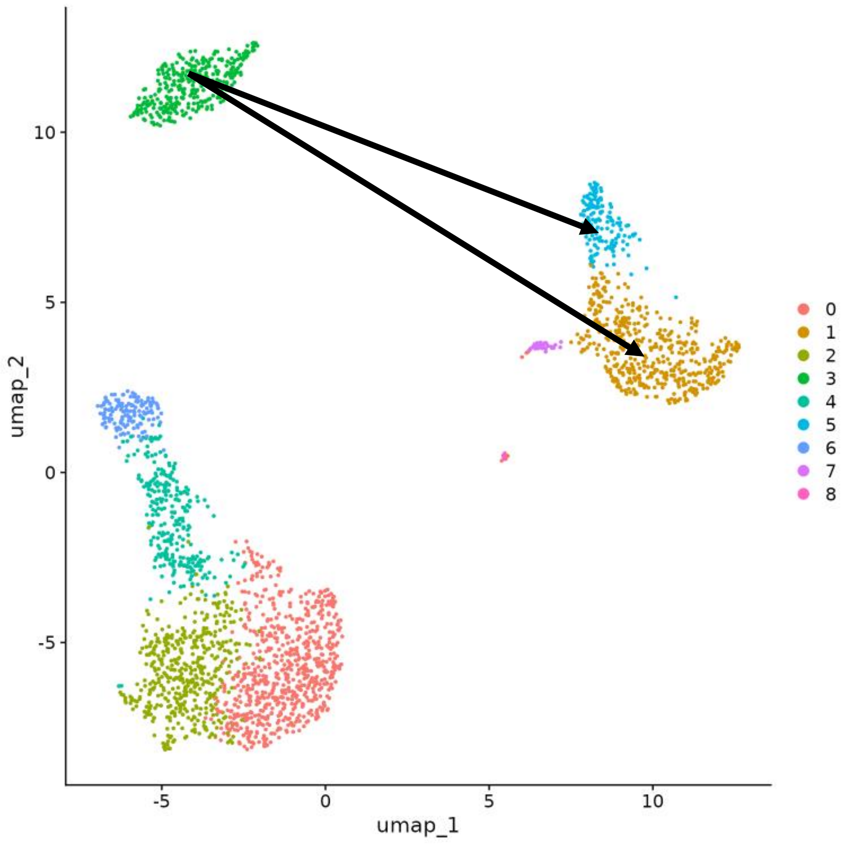
Interpret visual cluster size ✗

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#perform-linear-dimensional-reduction

4

Interpreting tSNE and UMAP in the single-cell world

The green cluster is more similar to the blue cluster than to the brown cluster because it is closer.



You shall not

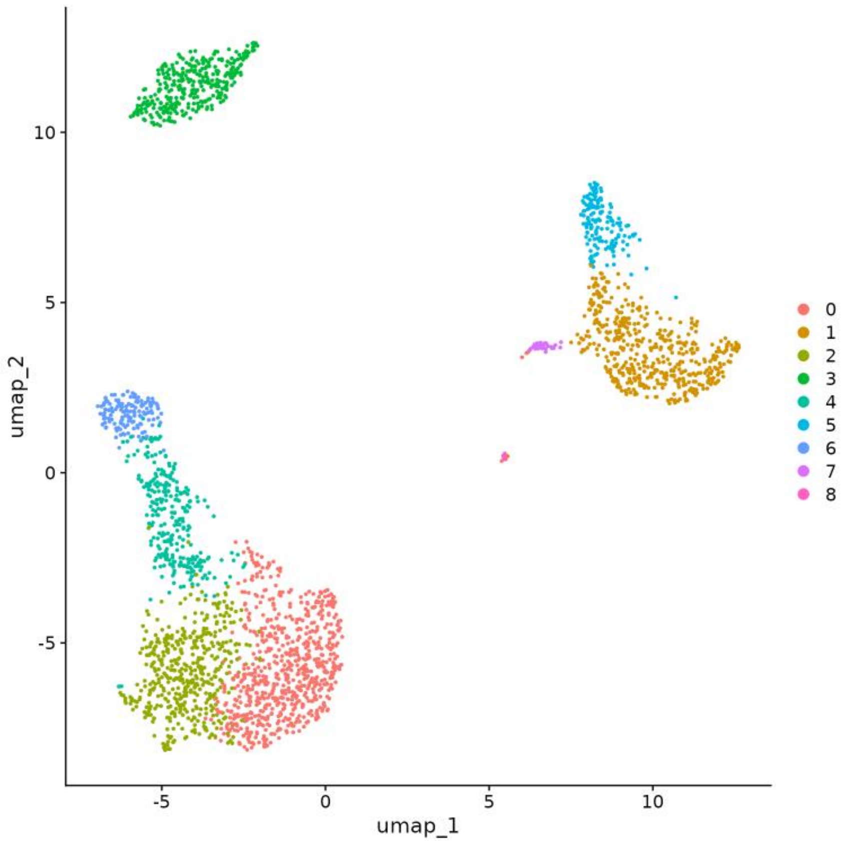
Interpret visual cluster size ✗

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#perform-linear-dimensional-reduction

4

Interpreting tSNE and UMAP in the single-cell world

The green cluster is more similar to the blue cluster than to the brown cluster because it is closer.



You shall not

Interpret visual cluster size ✗

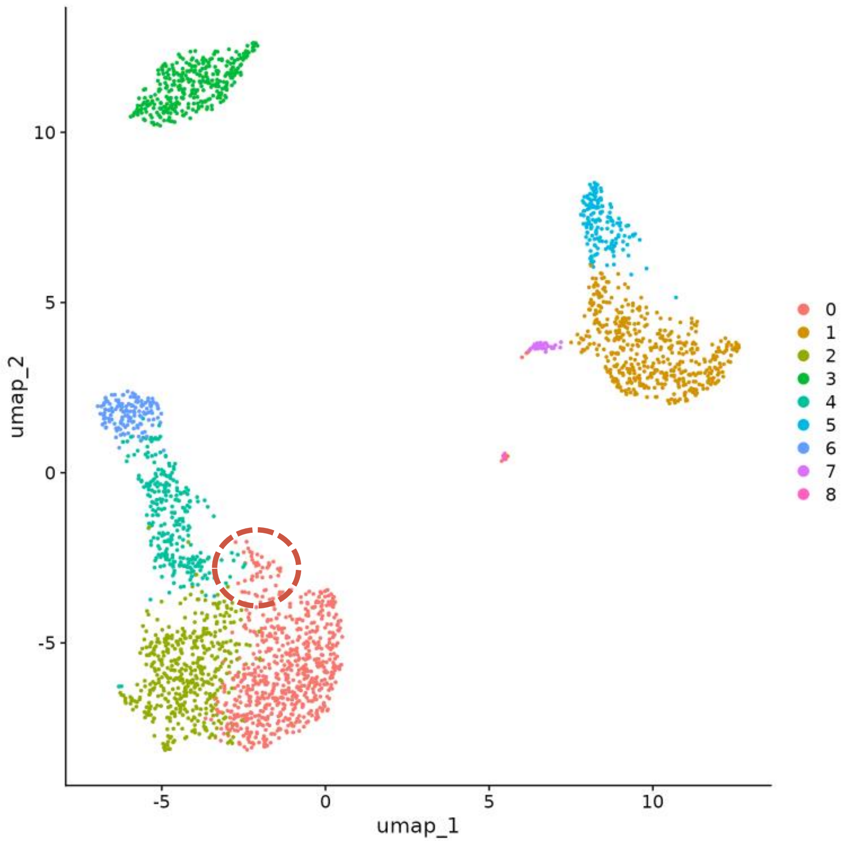
Interpret distances between cluster ✗

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#perform-linear-dimensional-reduction

4

Interpreting tSNE and UMAP in the single-cell world

There are clearly several subpopulations within the red cluster.



You shall not

Interpret visual cluster size ✗

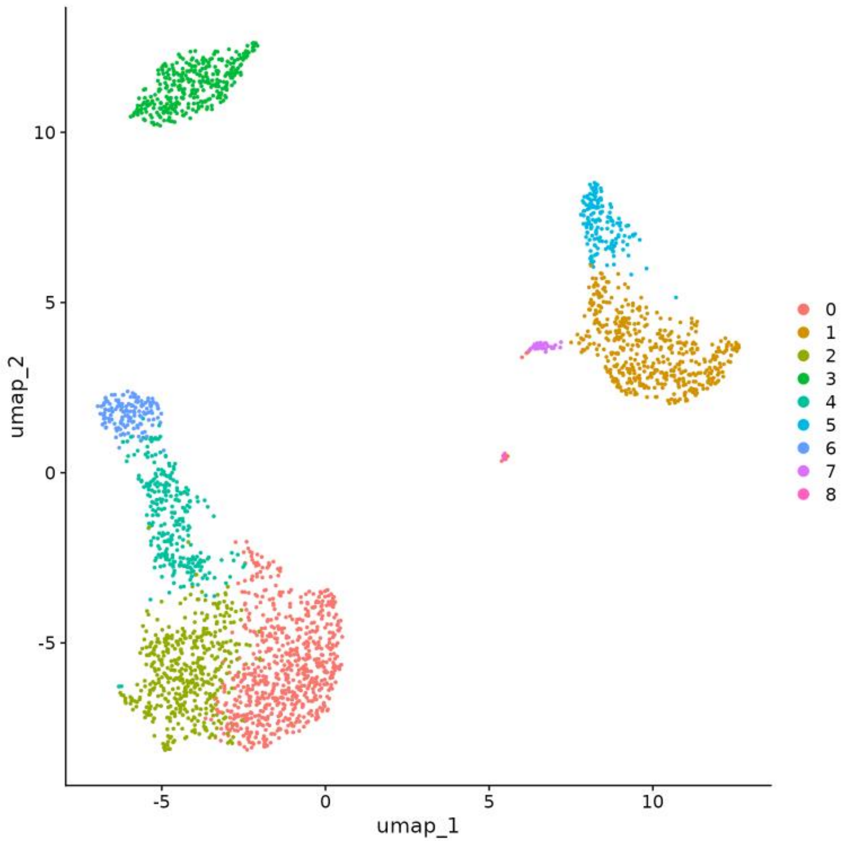
Interpret distances between cluster ✗

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#perform-linear-dimensional-reduction

4

Interpreting tSNE and UMAP in the single-cell world

There are clearly several subpopulations within the red cluster.



You shall not

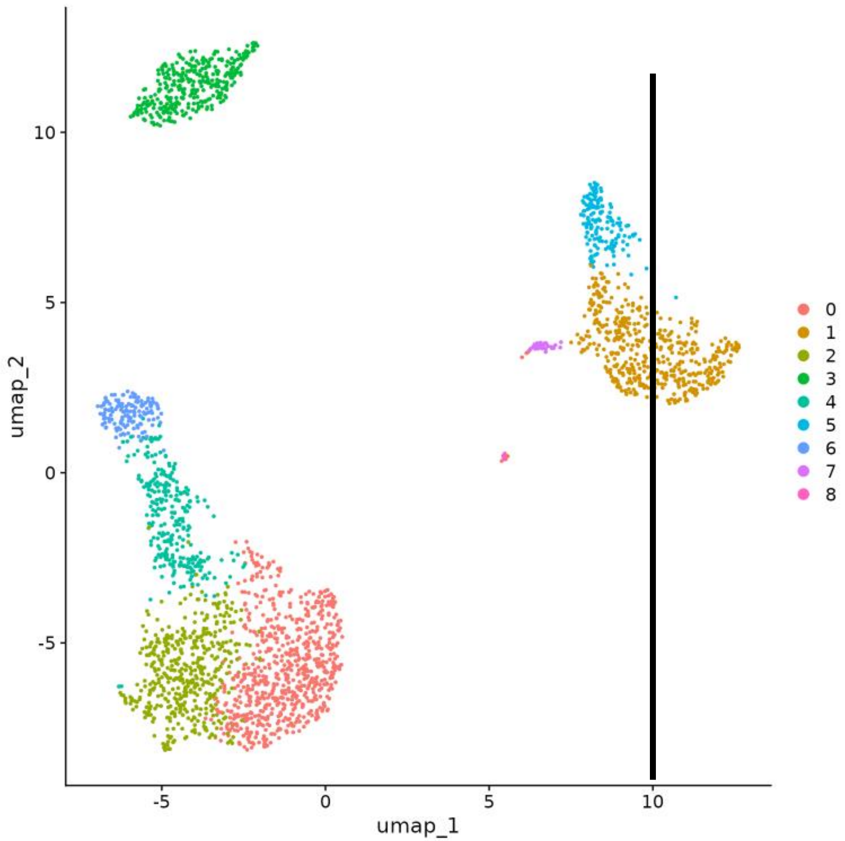
- Interpret visual cluster size ✗
- Interpret distances between cluster ✗
- Perform clustering (manual or automatic) on UMAP/tSNE ✗

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#perform-linear-dimensional-reduction

4

Interpreting tSNE and UMAP in the single-cell world

We considered all cells with umap_1>10 as marker-positive for the analysis.



You shall not

- Interpret visual cluster size ✗
- Interpret distances between cluster ✗
- Perform clustering (manual or automatic) on UMAP/tSNE ✗

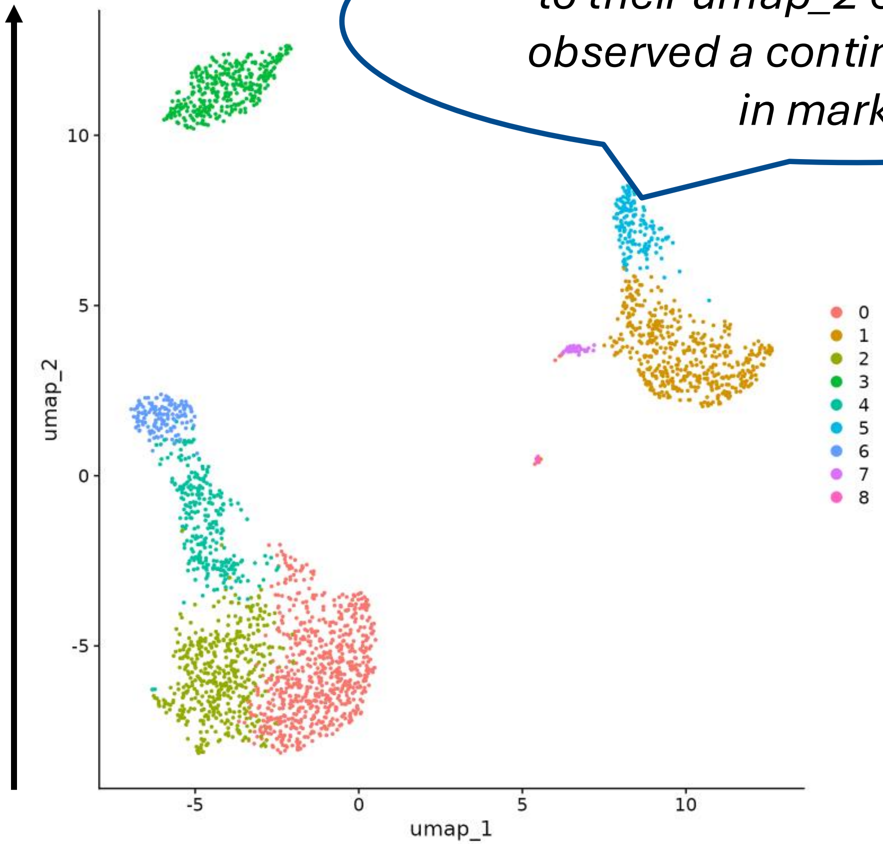
https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#perform-linear-dimensional-reduction

4

Interpreting tSNE and UMAP in the single-cell world

We considered all cells with umap_1>10 as marker x-positive for the analysis.

When we ordered cells according to their umap_2 coordinate, we observed a continuous increase in marker y.



You shall not

Interpret visual cluster size ☒

Interpret distances between cluster ☒

Perform clustering (manual or automatic) on UMAP/tSNE ☒

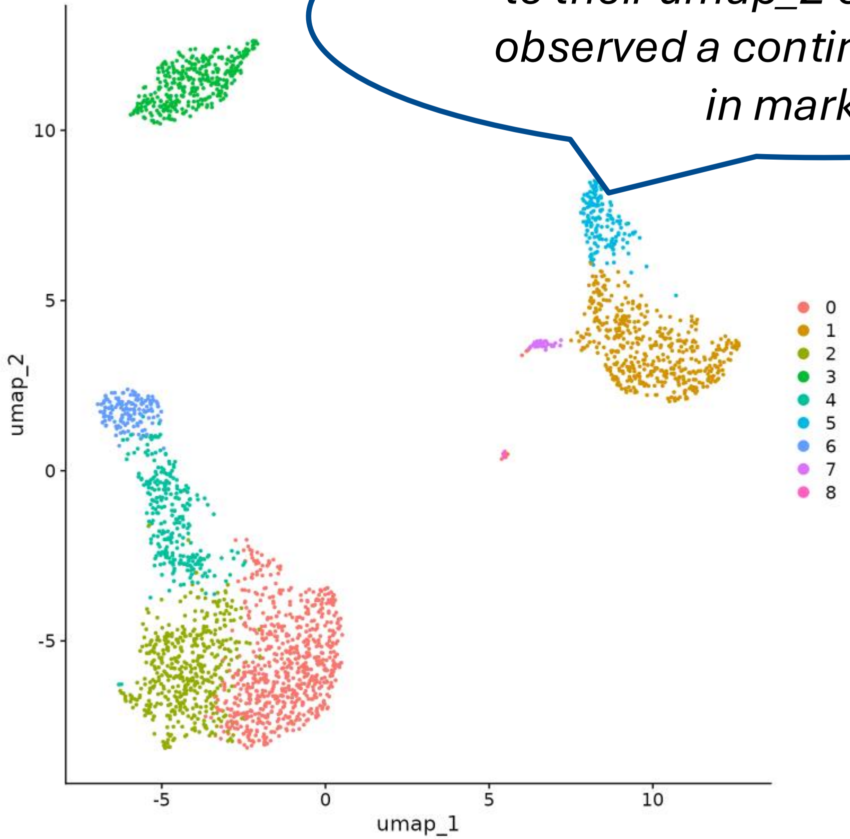
https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#perform-linear-dimensional-reduction

4

Interpreting tSNE and UMAP in the single-cell world

We considered all cells with umap_1>10 as marker x-positive for the analysis.

When we ordered cells according to their umap_2 coordinate, we observed a continuous increase in marker y.



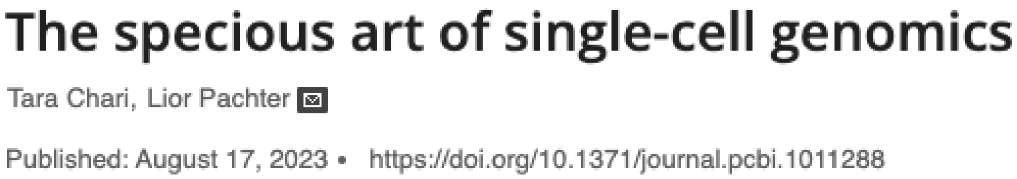
You shall not

- Interpret visual cluster size ✗
- Interpret distances between cluster ✗
- Perform clustering (manual or automatic) on UMAP/tSNE ✗
- Perform gating or ordering based on UMAP/tSNE coordinates. ✗

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#perform-linear-dimensional-reduction

4

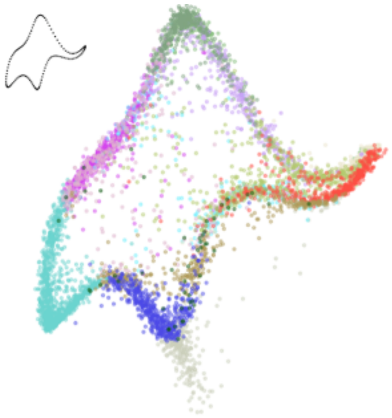
The “UMAP is wrong and useless” controversy

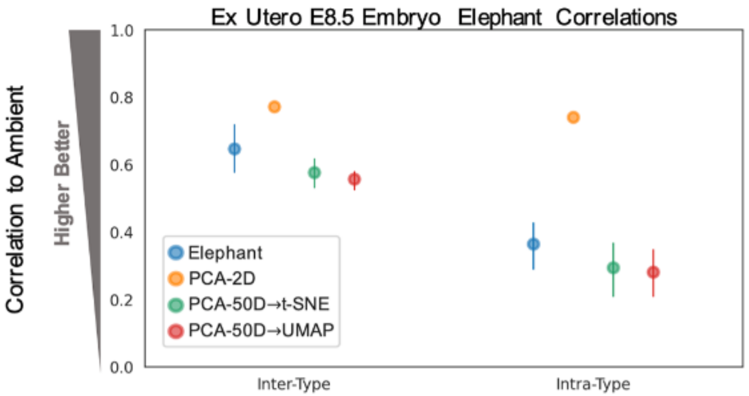


The specious art of single-cell genomics

Tara Chari, Lior Pachter

Published: August 17, 2023 • https://doi.org/10.1371/journal.pcbi.1011288

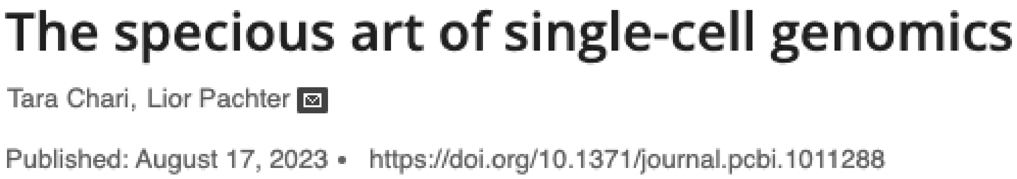




https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011288

4

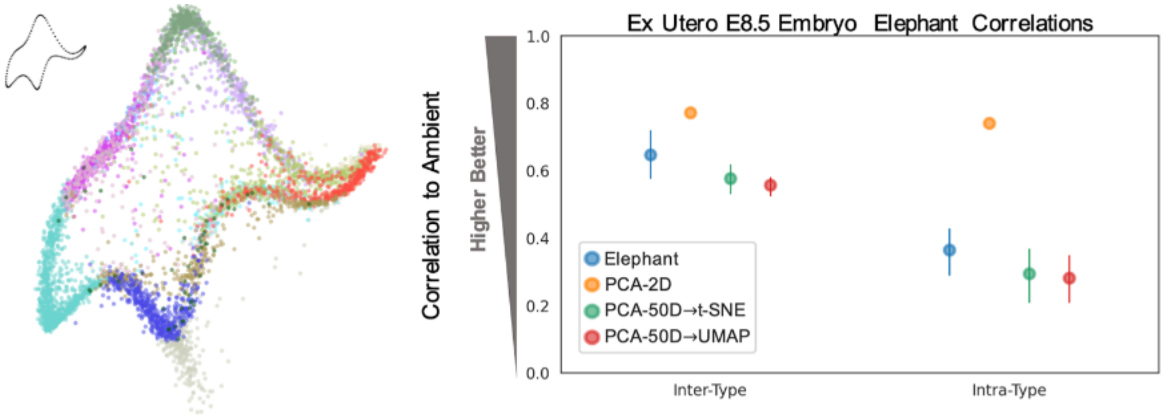
The “UMAP is wrong and useless” controversy

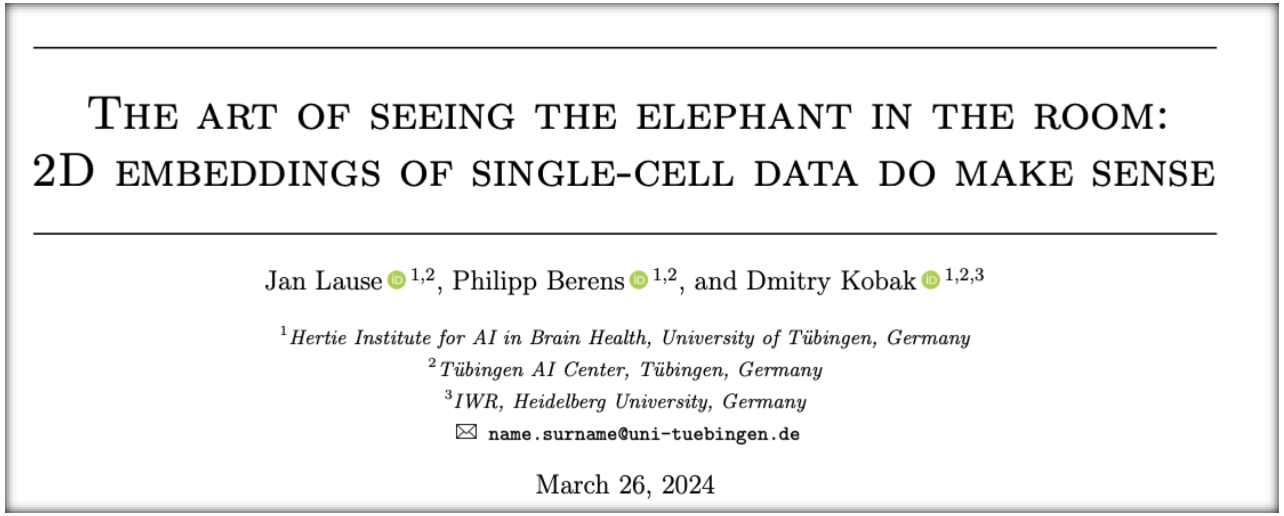


The specious art of single-cell genomics

Tara Chari, Lior Pachter ✉

Published: August 17, 2023 • https://doi.org/10.1371/journal.pcbi.1011288





The art of seeing the elephant in the room: 2D embeddings of single-cell data do make sense

Jan Lause [1,2], Philipp Berens [1,2], and Dmitry Kobak [1,2,3]

[1] *Hertie Institute for AI in Brain Health, University of Tübingen, Germany*
[2] *Tübingen AI Center, Tübingen, Germany*
[3] *IWR, Heidelberg University, Germany*
✉ name.surname@uni-tuebingen.de

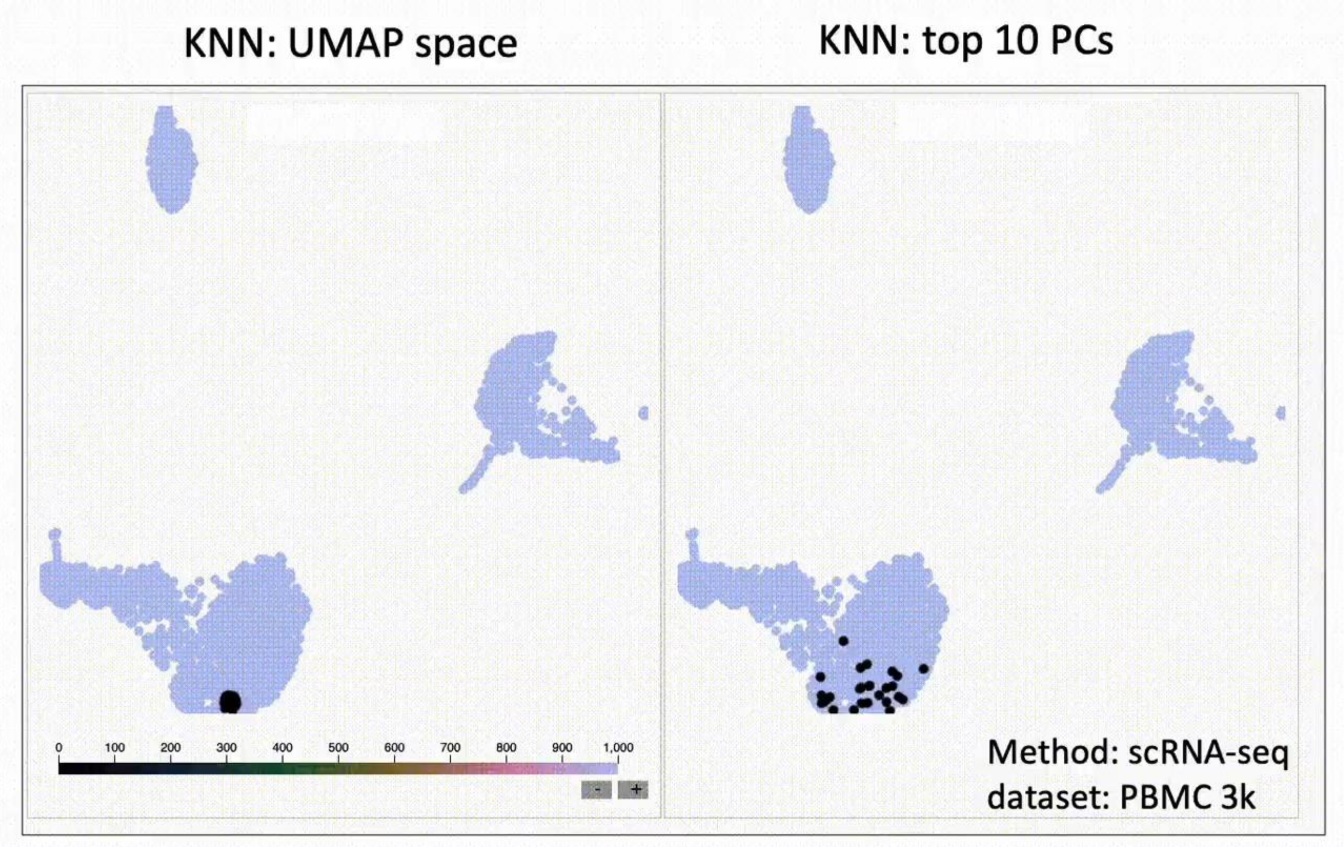
March 26, 2024

https://www.biorxiv.org/content/10.1101/2024.03.26.586728v1

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011288

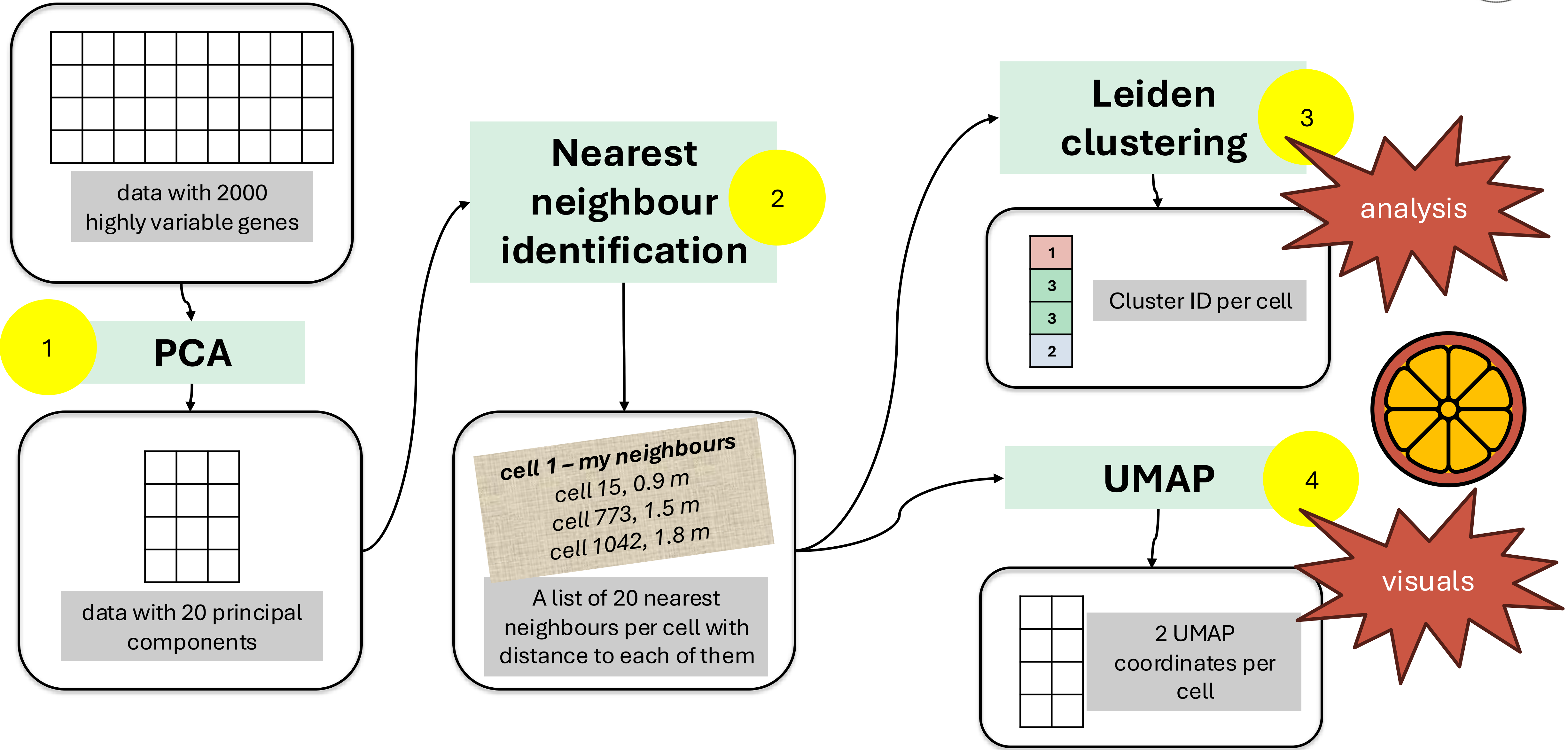
4

While 50D can approximate the full information reasonably, **2D cannot preserve the full information.**



https://twitter.com/i/status/1775934737917567360

Data types along the processing path



data with 2000 highly variable genes

PCA

data with 20 principal components

Nearest neighbour identification

cell 1 – my neighbours

cell 15, 0.9 m

cell 773, 1.5 m

cell 1042, 1.8 m

A list of 20 nearest neighbours per cell with distance to each of them

Leiden clustering

Cluster ID per cell

analysis

UMAP

2 UMAP coordinates per cell

visuals