Università
della
Svizzera
italiana

Institute
of
Oncology
Research

Dimensionality Reduction

Luciano Cascione, PhD
Bioinformatics Core Unit

SIB Swiss Institute of Bioinformatics | iOR Bioinformatics Core Unit | USI Faculty of Biomedical Sciences

Why Dimensionality Reduction in scRNA-Seq?

scRNA-seq data = thousands of genes × thousands of cells
high noise, redundancy, and sparsity

Challenge	Solution
High-dimensional gene space contains noise & redundancy	Focus on the most informative gene patterns
Most genes are not useful to define cell identity	Highlight key features that distinguish cell types
Many algorithms break with too many variables	Simplify to a manageable feature space
Data too complex to interpret visually	Enable 2D/3D visualization (UMAP, t-SNE)

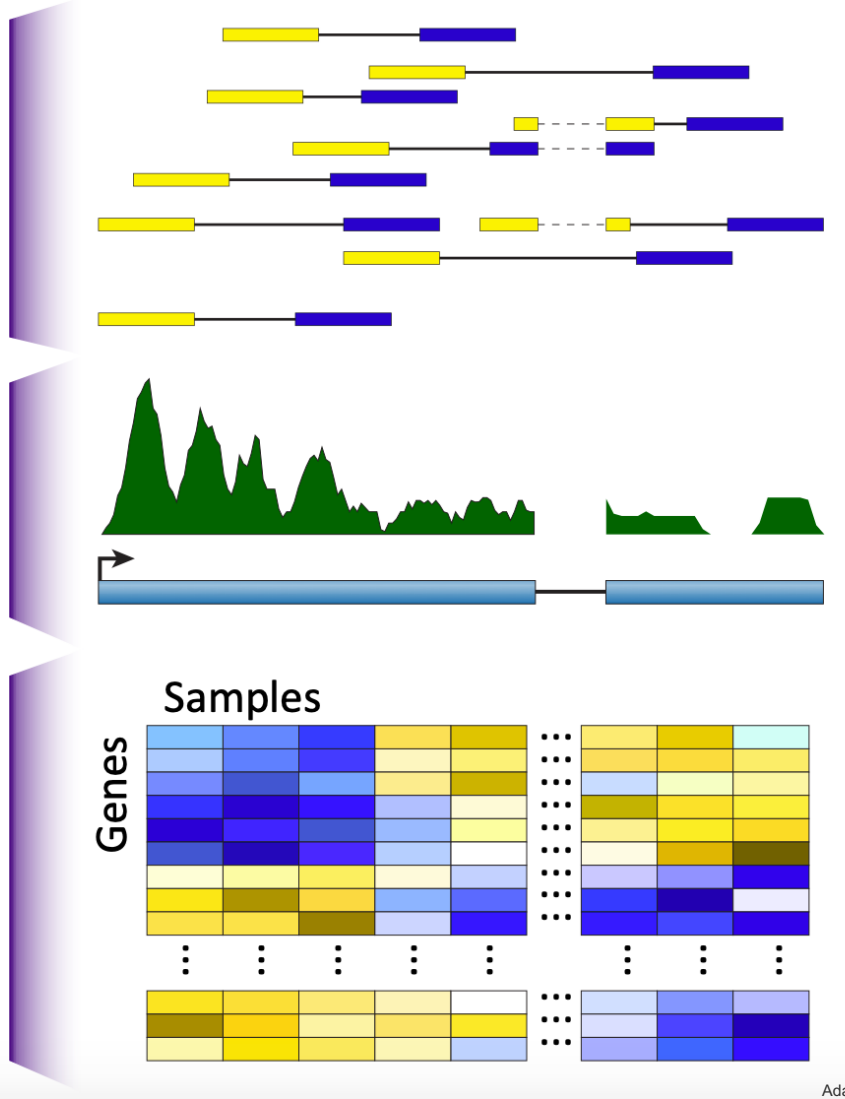
Dimensionality reduction helps us find the real biological signal

It is all about matrix

Raw RNA-Seq reads

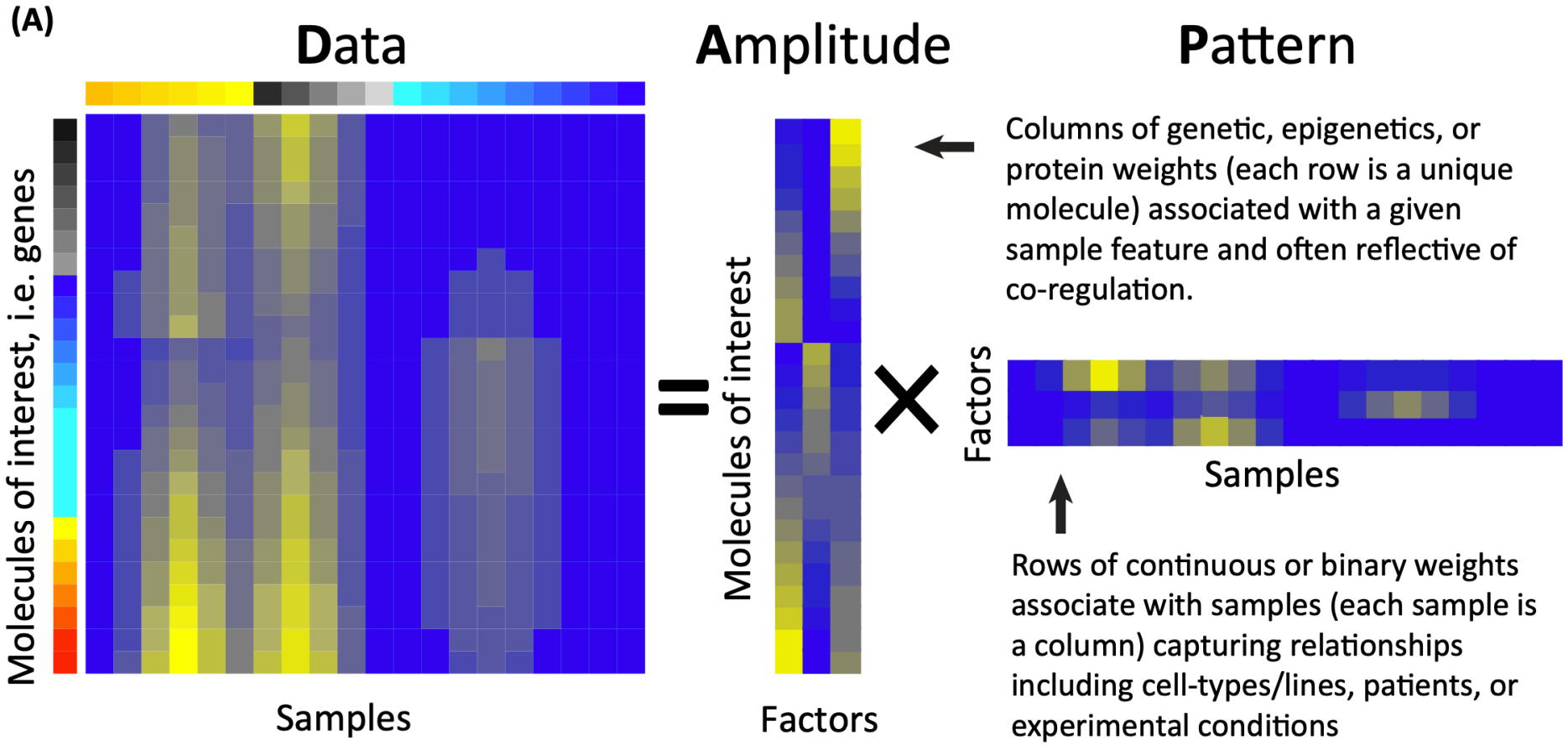
Alignment and quantification

Normalization and log-transformation



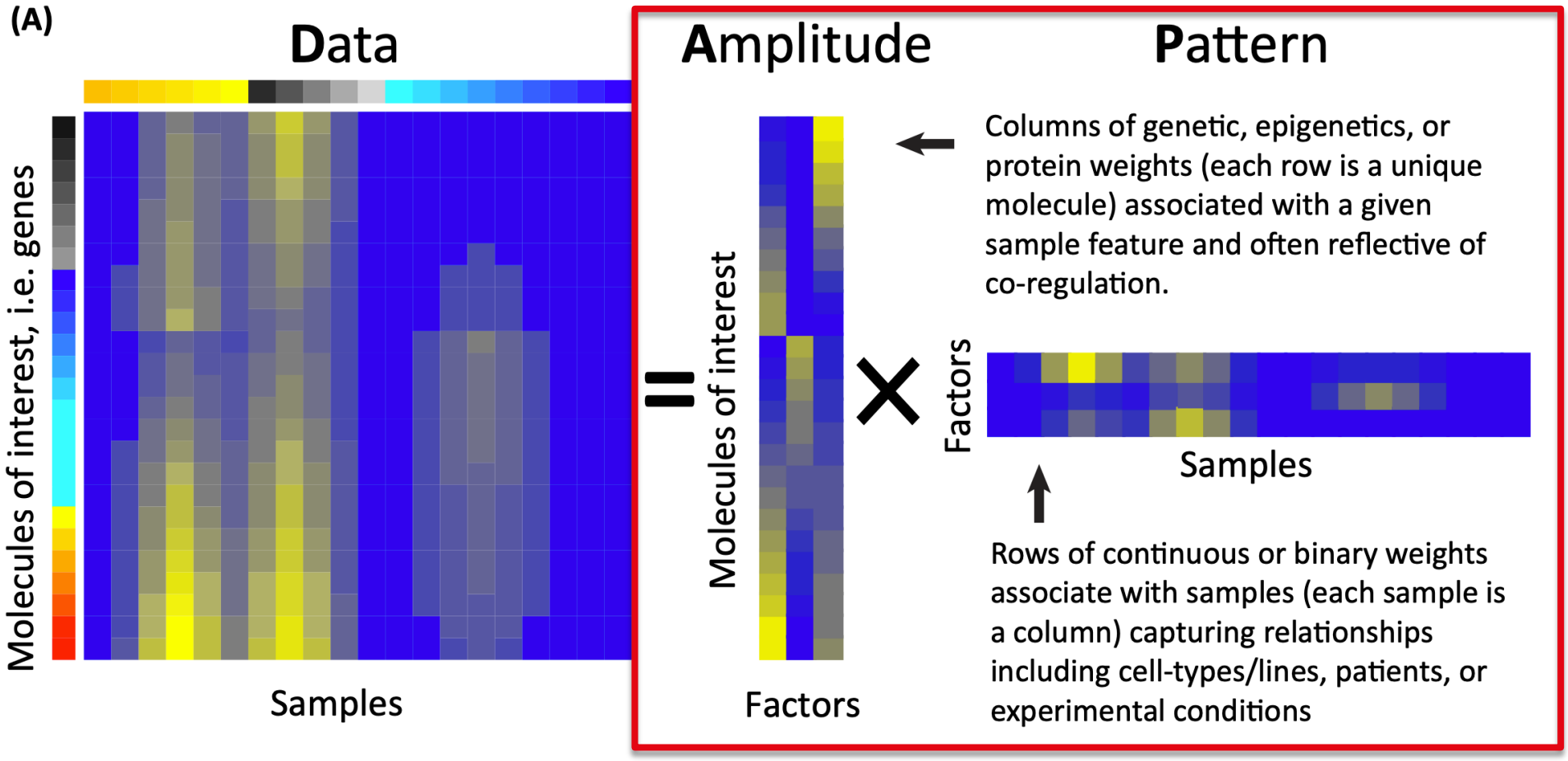
Adapted from : Stein-O'Brien, et al. Trends in Genetics (2018)

Matrix Factorization



Adapted from : Stein-O'Brien, et al. Trends in Genetics (2018)

Matrix Factorization



From the Amplitude and Pattern matrices we can derive biological insights

Adapted from : Stein-O'Brien, et al. Trends in Genetics (2018)

PCA recipe

Calculate the covariance matrix

- How each gene's expression correlates with every other gene's expression across cells
- High covariance suggests that two genes have similar patterns across cells

Eigen Decomposition

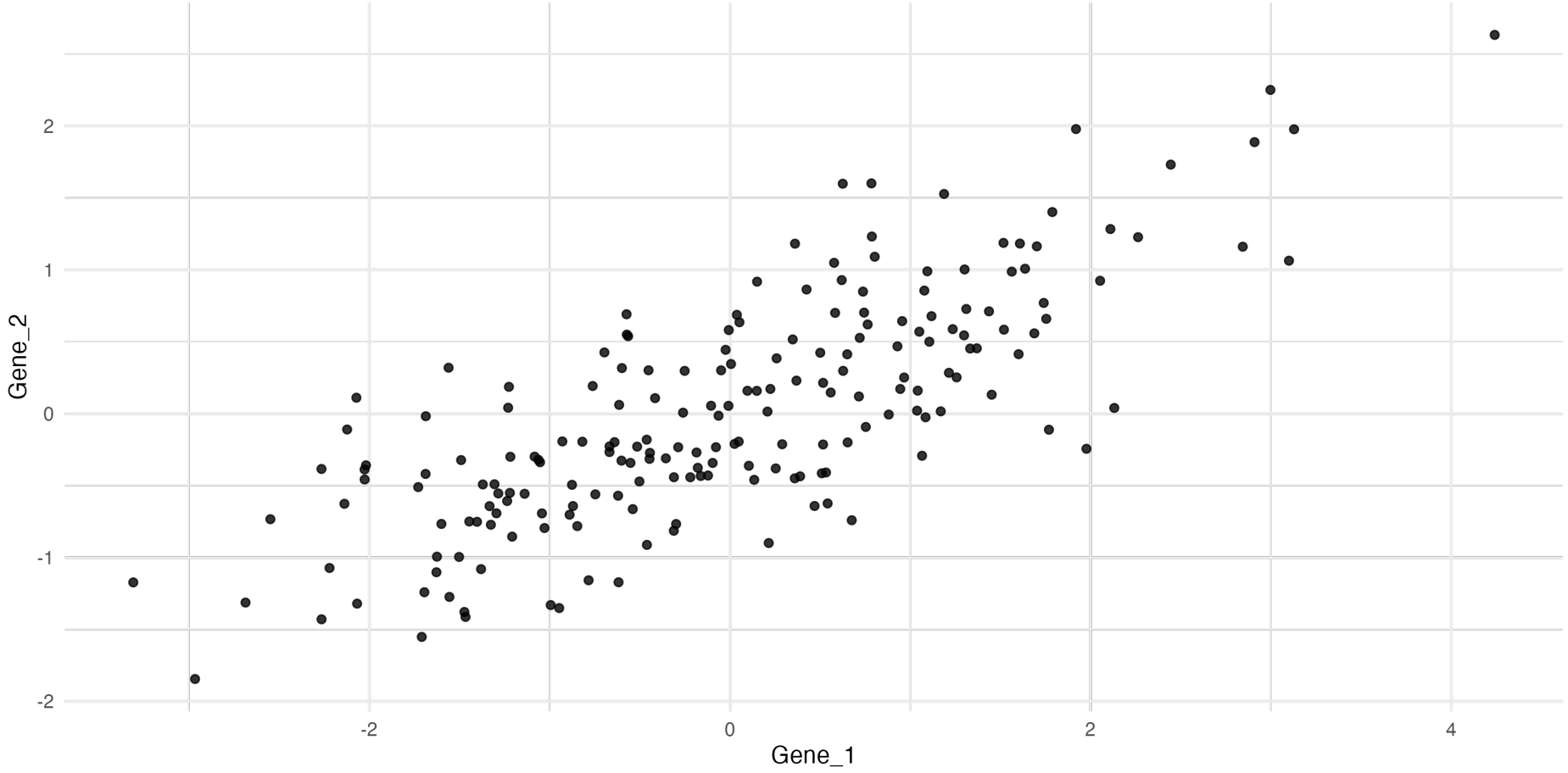
- **Eigenvectors** (Principal Components, PCs): Represent new axes (or directions) in the data space along which the variation is maximized (aka gene weights)
- **Eigenvalues**: Indicate the amount of variance explained by each PC

Projection into the eigenvectors

- Genes are projected onto the new set of axes (PCs).
- Each cell now has a score (coordinate) on each PC, representing its position in the reduced-dimension space.

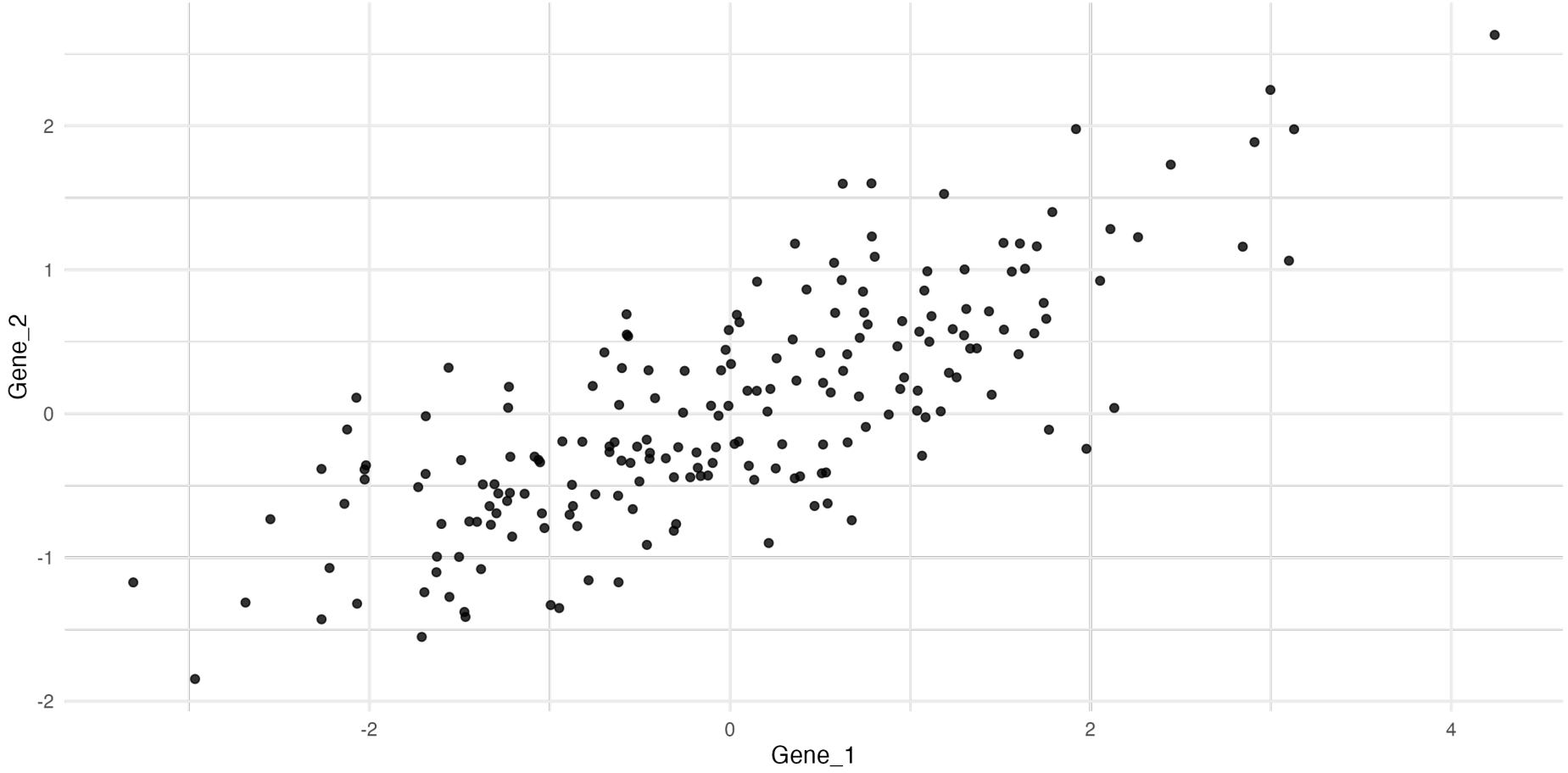
Principal Component Analysis

PCA learns orthogonal factors ordered by the relative amount of variation of the data that they explain



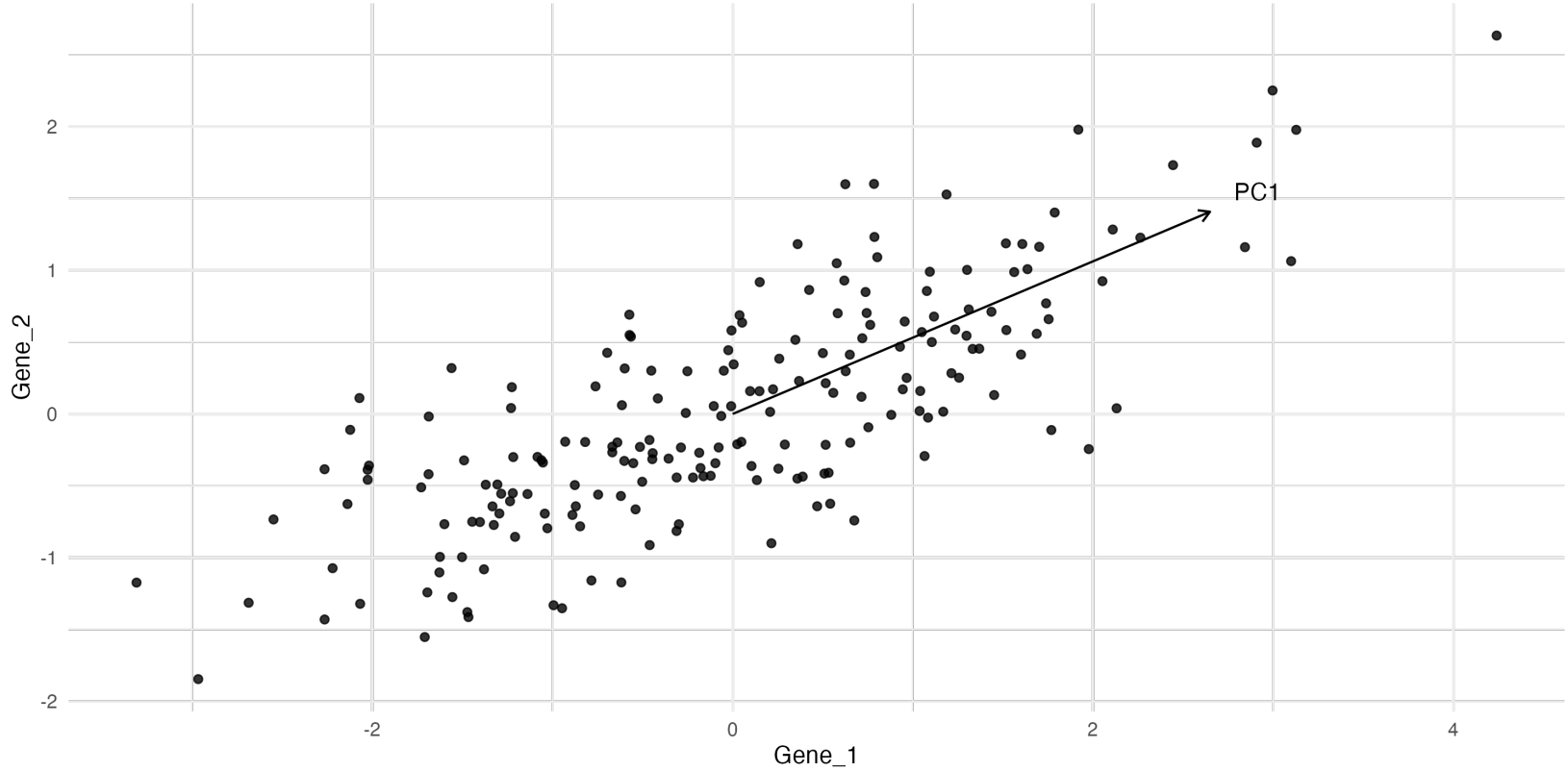
Principal Component Analysis

PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread.



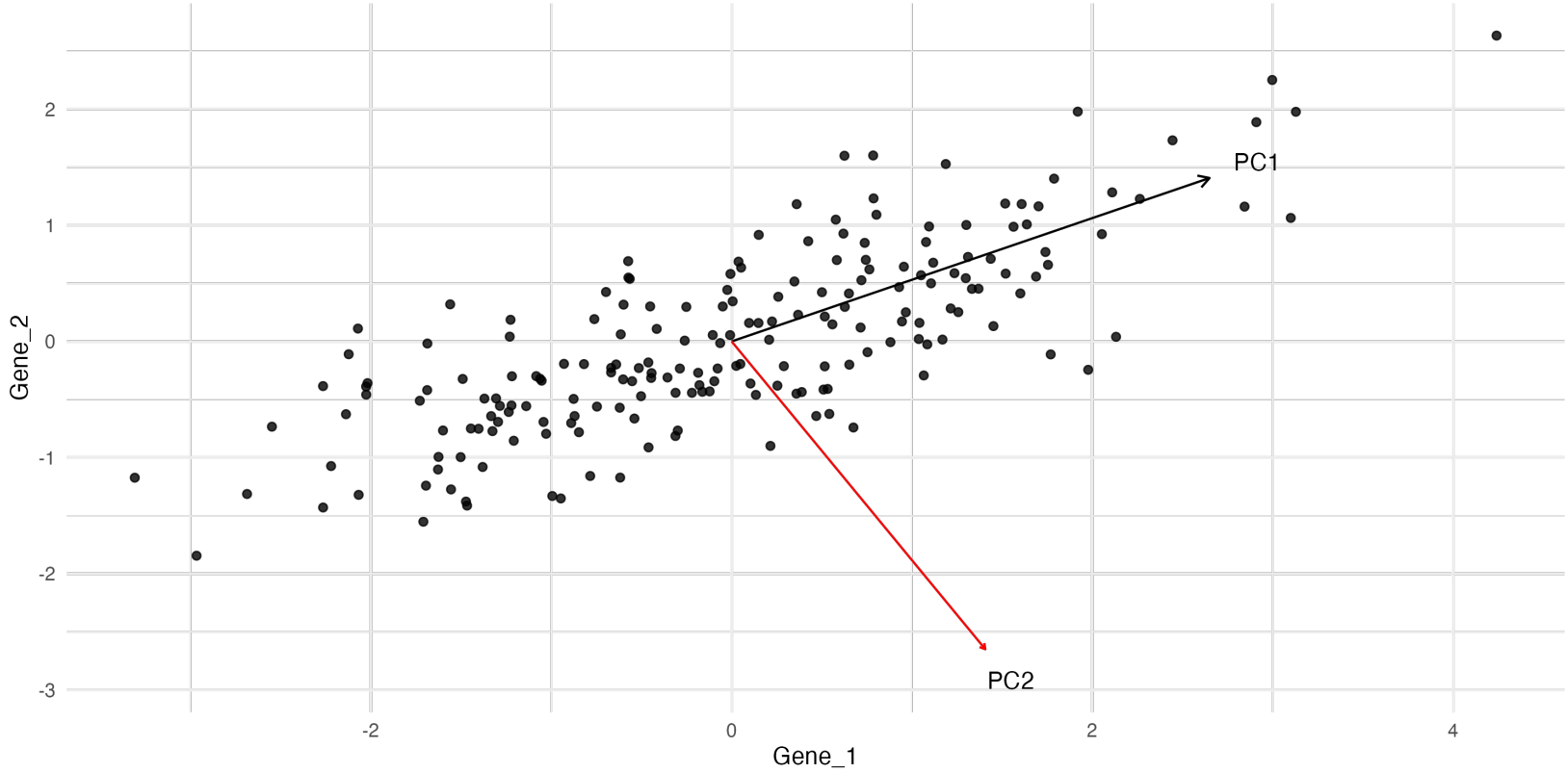
Principal Component Analysis

PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread.



Principal Component Analysis

PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread.



Principal Component Analysis - Rotation

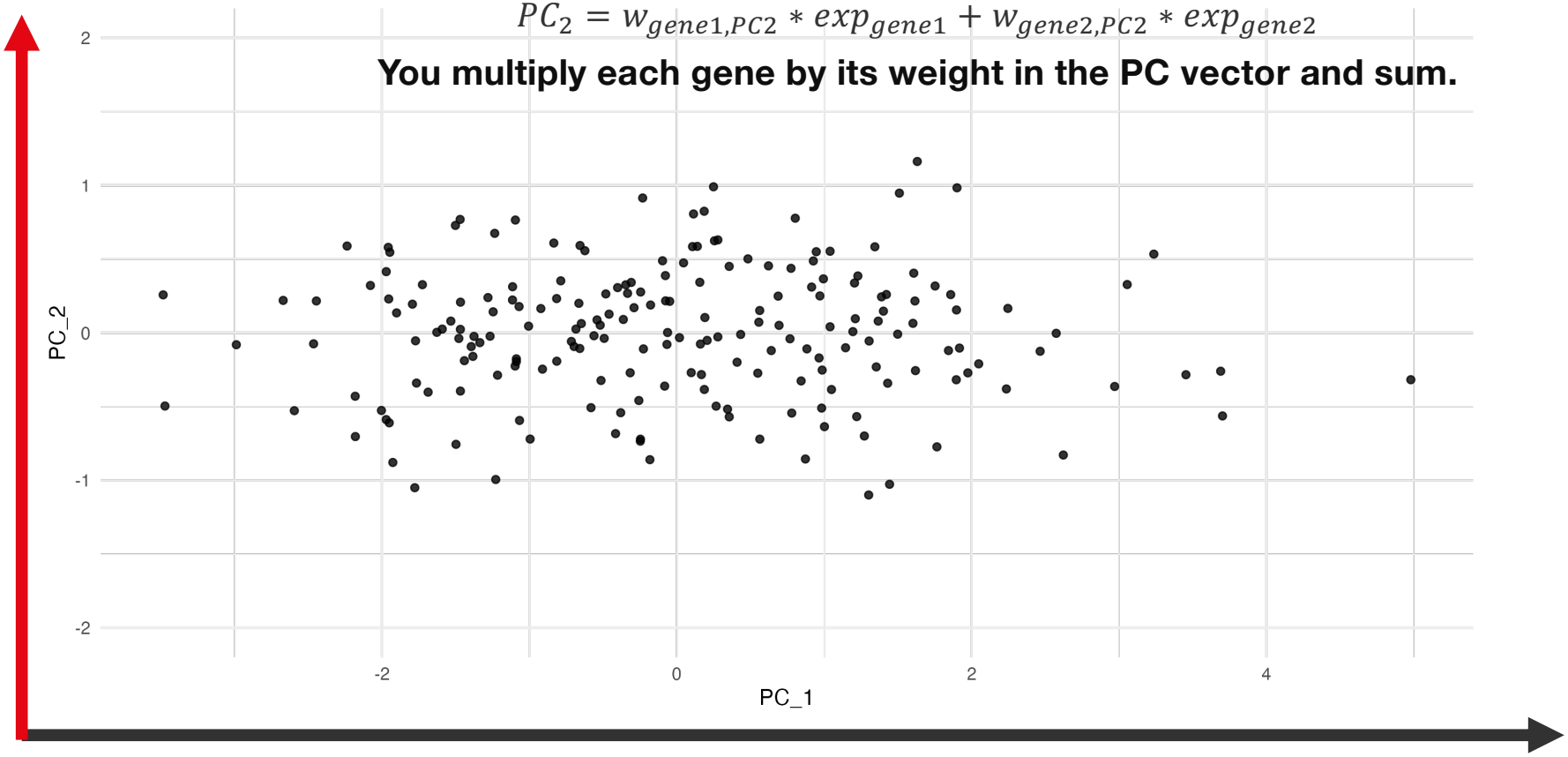
New axis that are linear combination of the original axes

Then the **projection (PC score)** of a cell onto PC1 is:

$$PC_1 = w_{gene1,PC1} * exp_{gene1} + w_{gene2,PC1} * exp_{gene2}$$

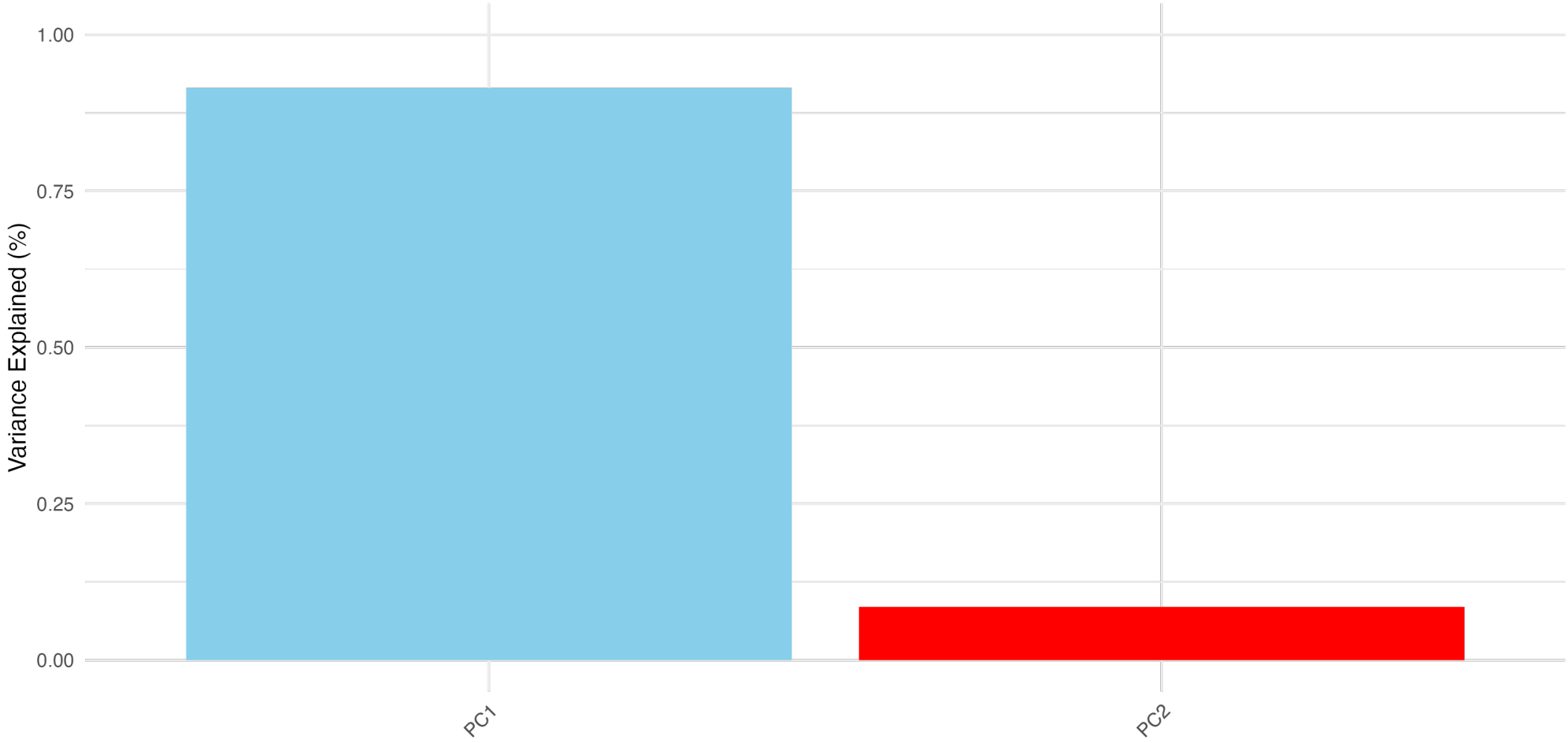
$$PC_2 = w_{gene1,PC2} * exp_{gene1} + w_{gene2,PC2} * exp_{gene2}$$

You multiply each gene by its weight in the PC vector and sum.



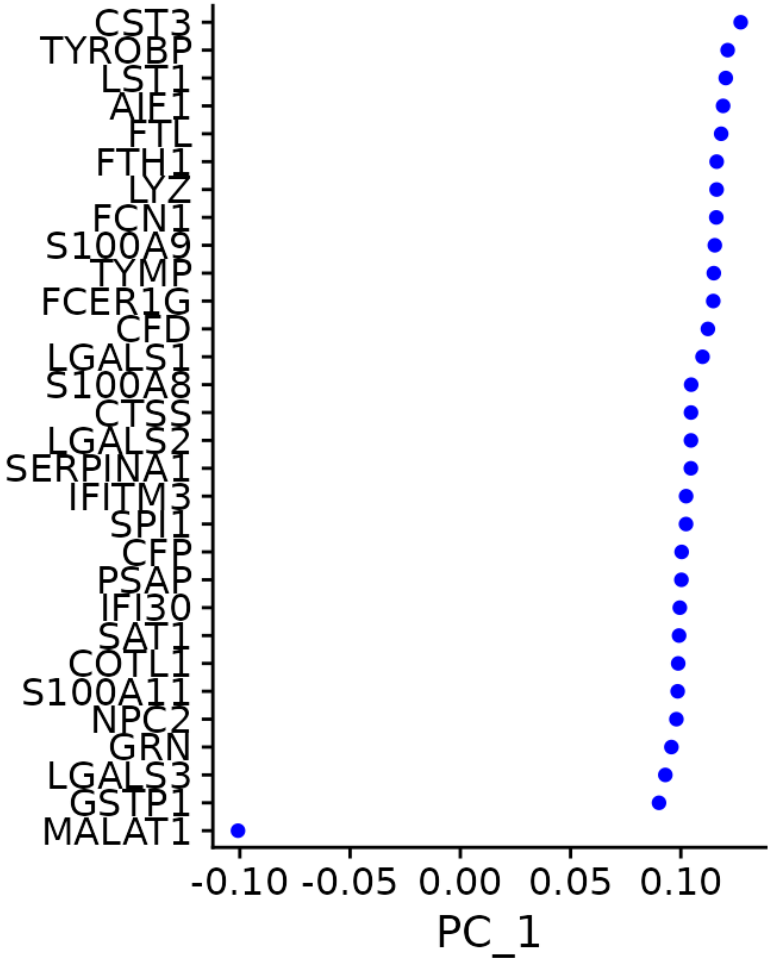
Principal Component Analysis

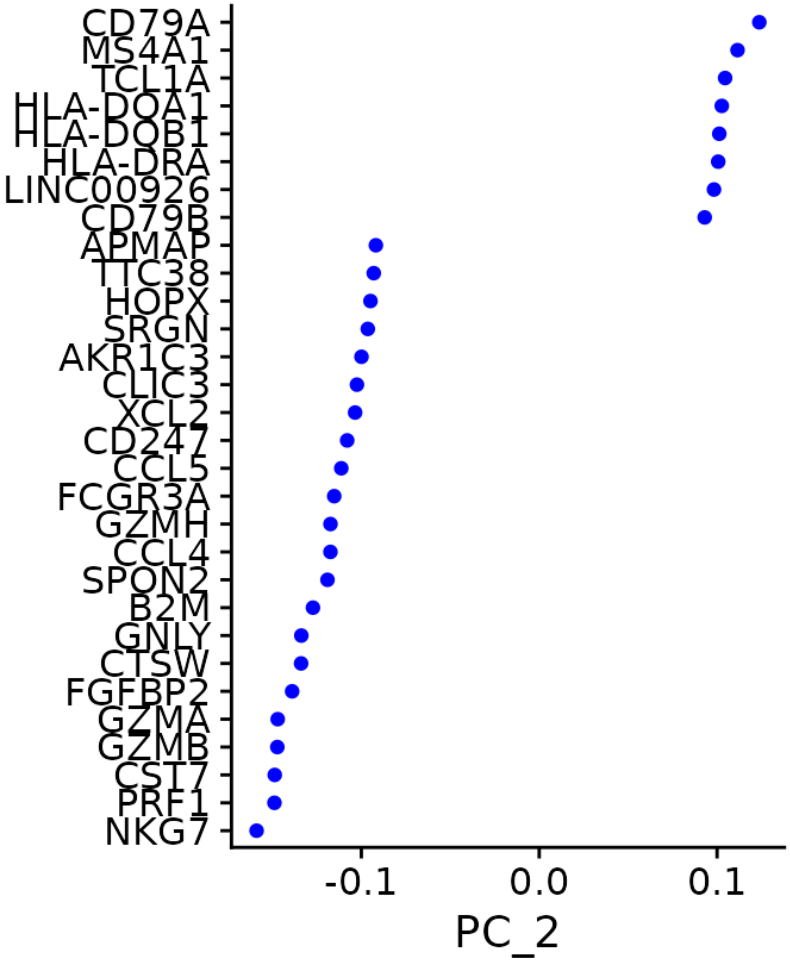
New axis that are linear combination of the original axes



Choosing the number of PCs

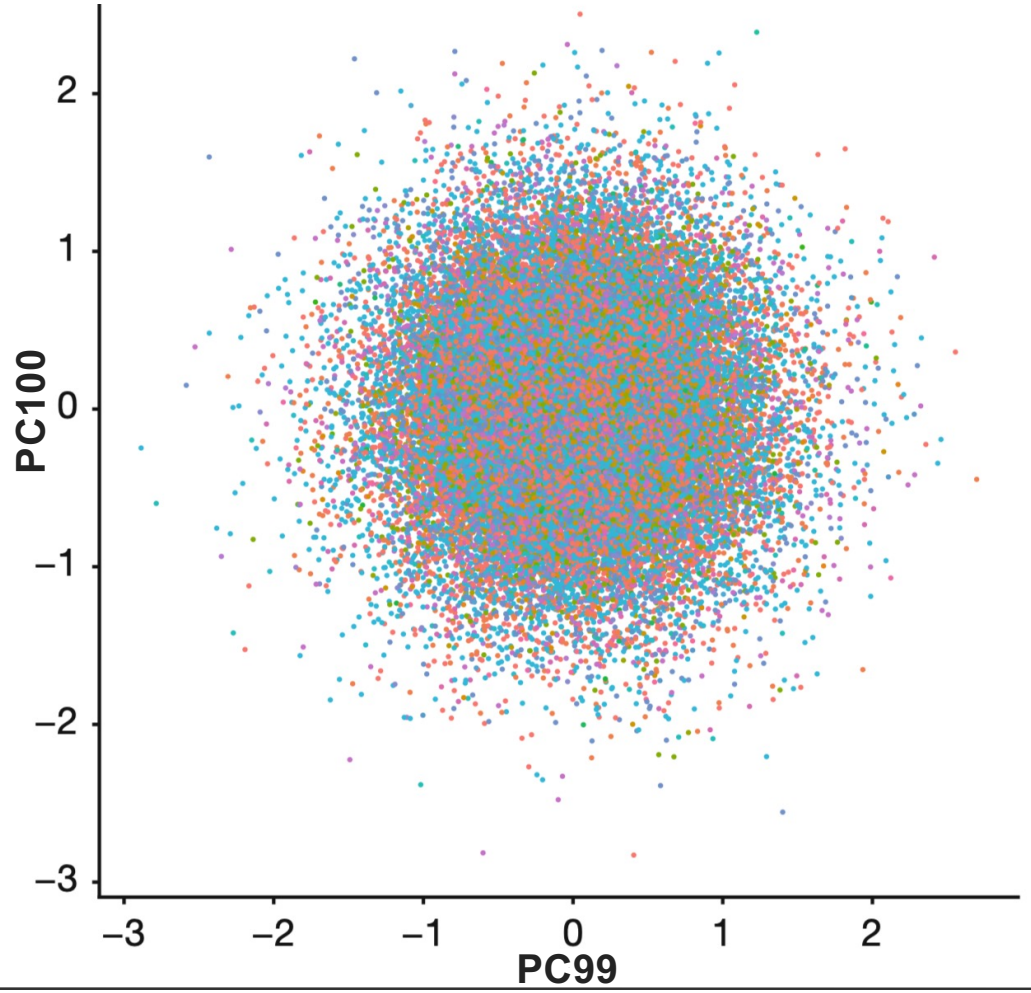
The **top PCs** contain higher variance from the data and could help identifying interesting biological aspects of your sample, but we can not include all the PCs





Choosing the number of PCs

The **top PCs** contain higher variance from the data and could help identifying interesting biological aspects of your sample, but we can not include all the PCs



Late PCs are just noise

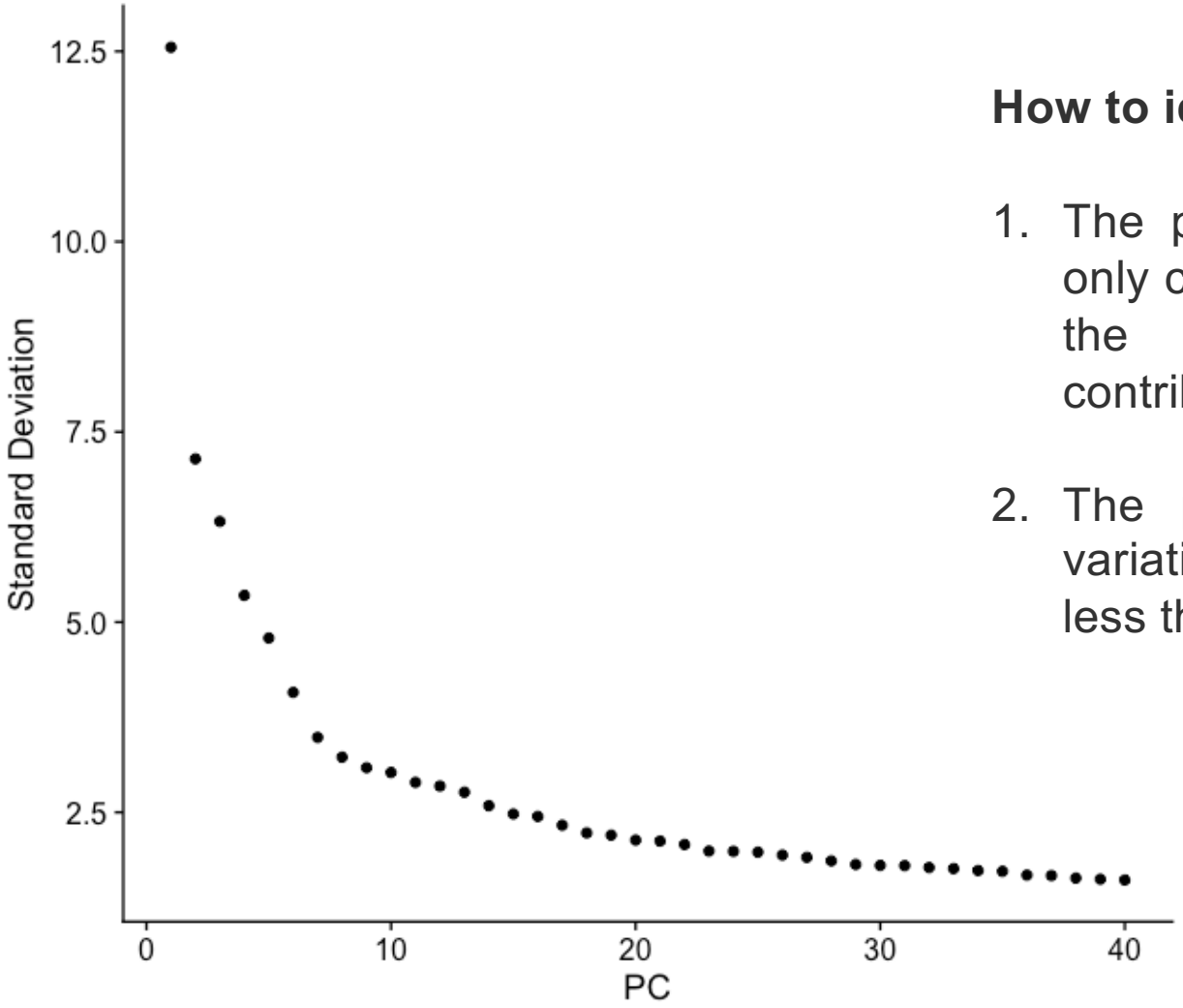
Adapted from: Linderman GC. Methods Mol Biol. 2021

Choosing the number of PCs

We could use some heuristic approaches:

- **PCs that explain at least 1% of variance**
- **The first 5-10 PCs**
- **Elbow-Plot**

The Elbow-point

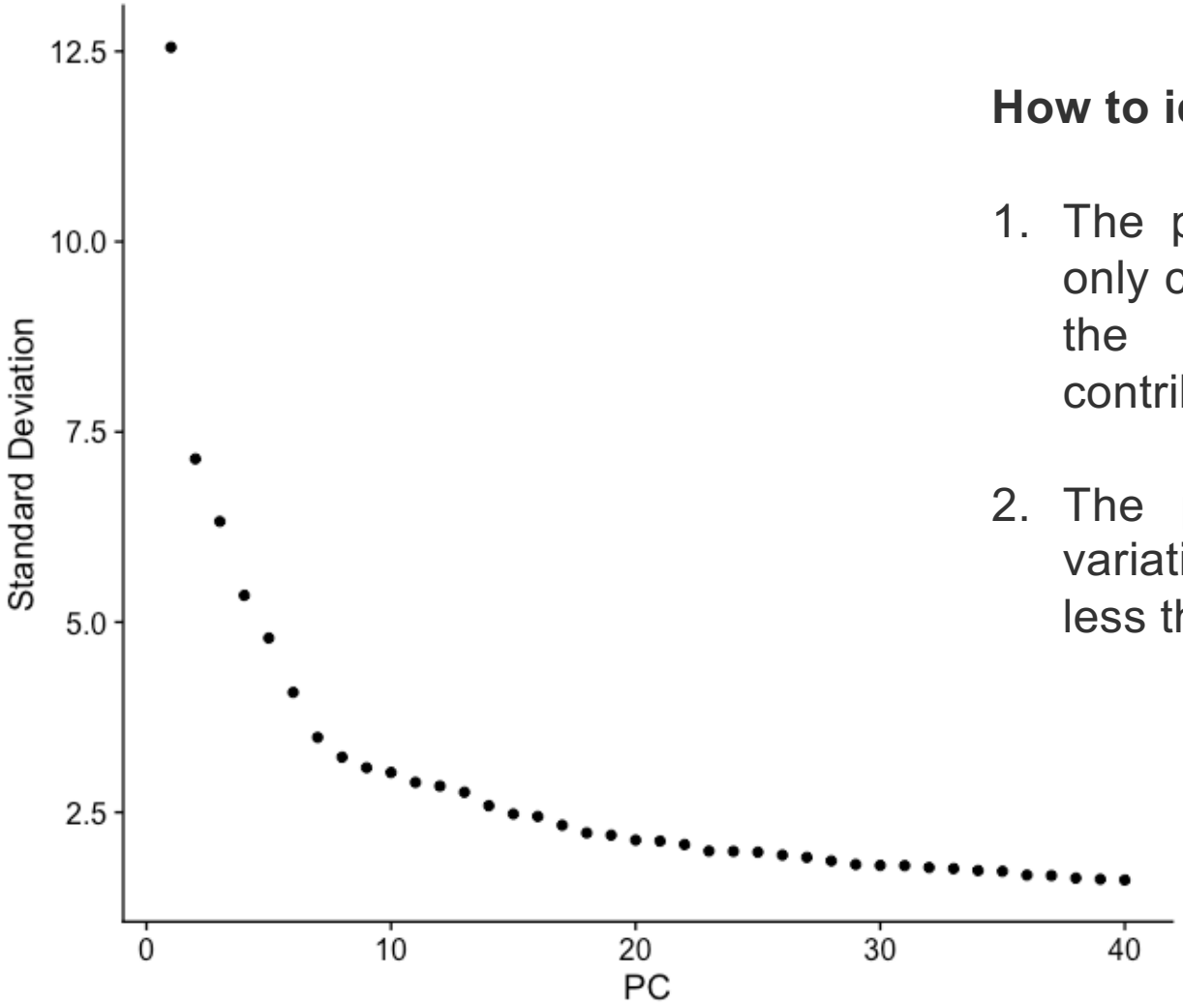


How to identify the Elbow point:

1. The point where the principal components only contribute 5% of standard deviation and the principal components cumulatively contribute 90% of the standard deviation

2. The point where the percent change in variation between the consecutive PCs is less than 0.1%.

The Elbow-point



How to identify the Elbow point:

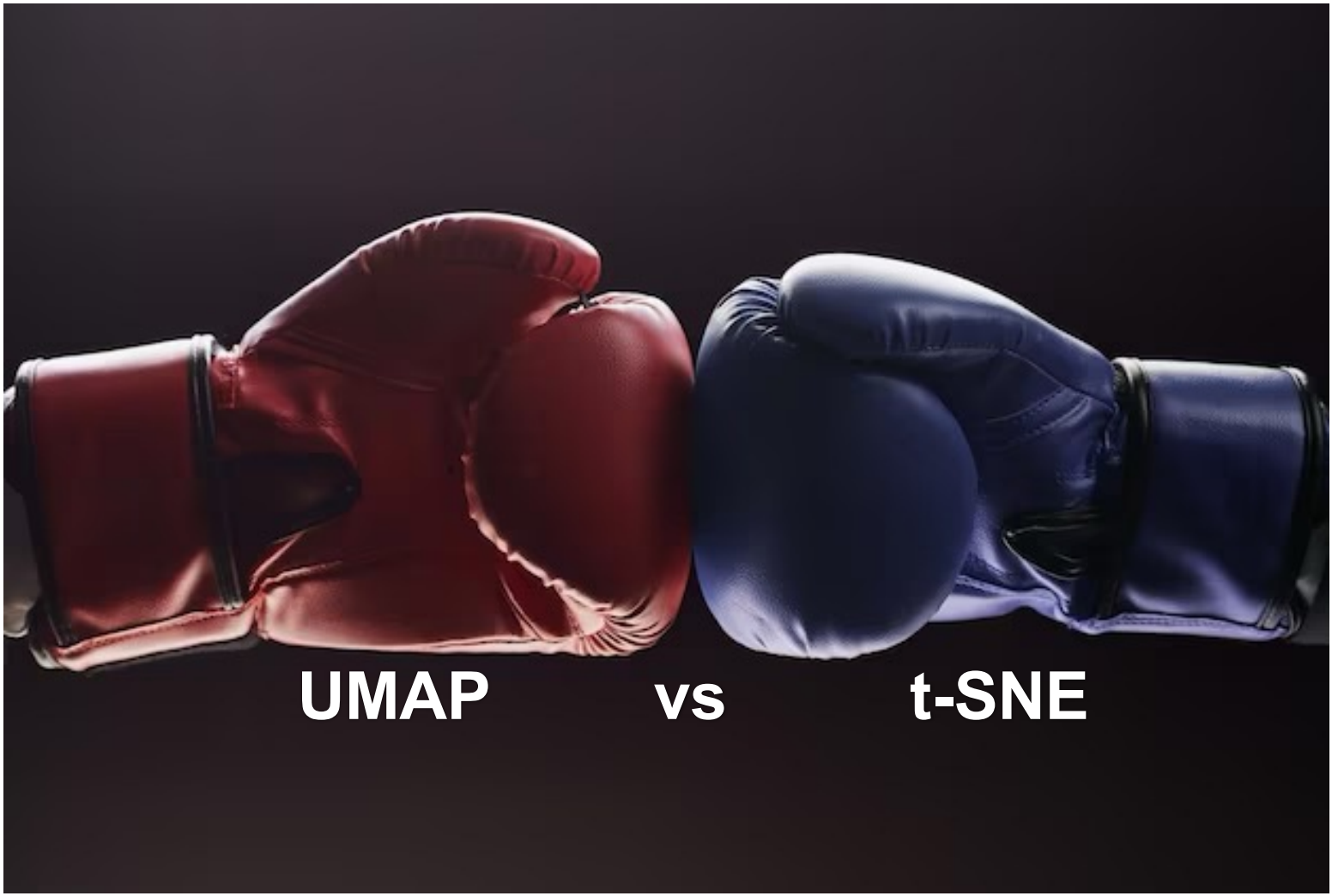
1. The point where the principal components only contribute 5% of standard deviation and the principal components cumulatively contribute 90% of the standard deviation

2. The point where the percent change in variation between the consecutive PCs is less than 0.1%.

Practical Considerations

- Cell sizes and sequencing depth are usually captured in the top principal components
- Repeat downstream analyses with a different number of PCs: 10, 15, or even 50. As you will observe, the results often do not differ dramatically.
- Late PCs may define rare subsets of cells.
- When selecting the number of PCs, it's better to choose more rather than Performing downstream analyses with only 5 PC seriously weaken the analysis..

Non-linear Methods for Dimensionality Reduction



t-SNE: t-distributed Stochastic Neighbourhood Embedding

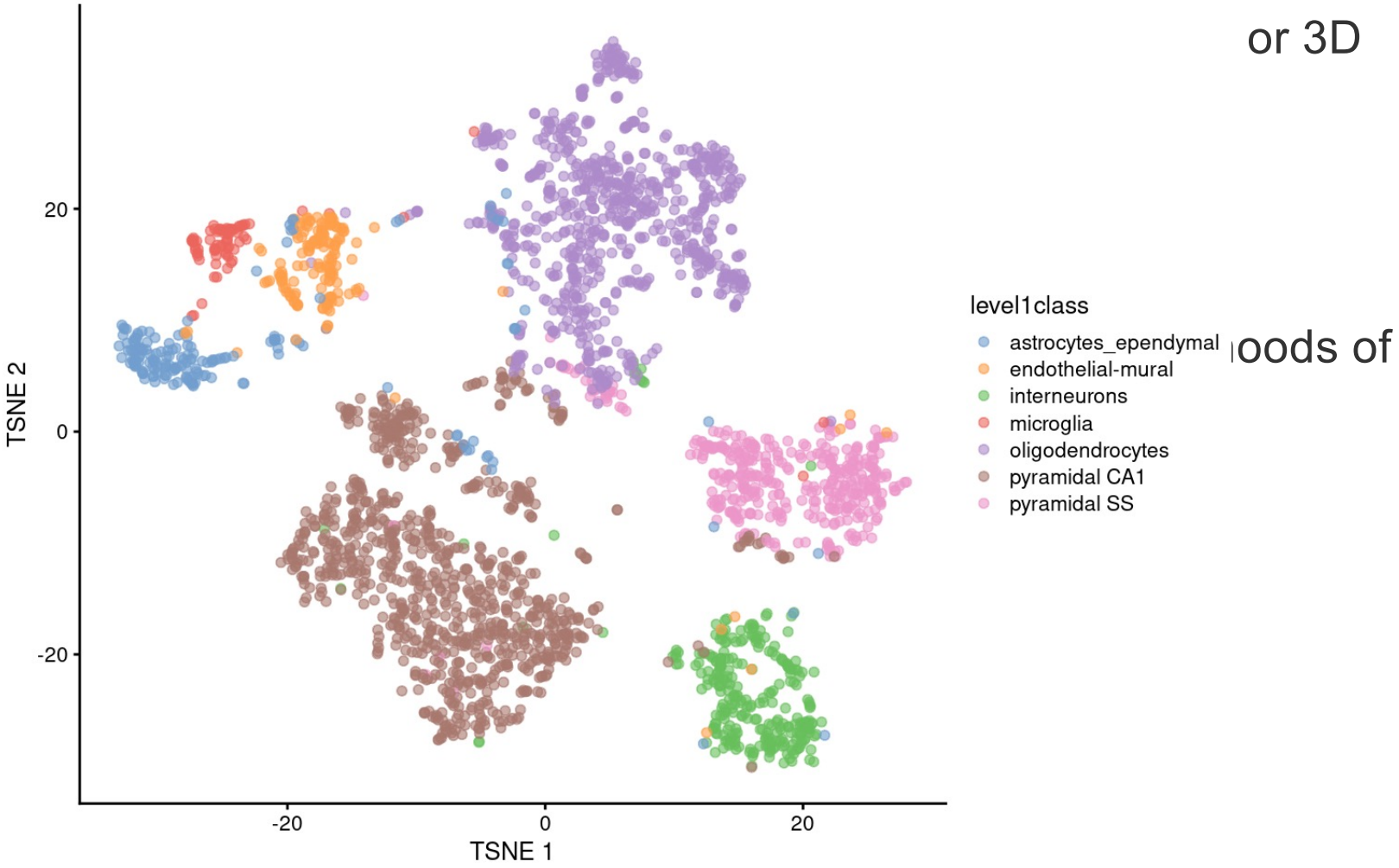
t-SNE take
map, wher
- Cells wit
- Cells wit

or 3D

It focuses
similar cell

oods of

Step-by-step
1.Measure si
2.Like saying
3.Create **prob**
4.Start placing
5.Move points
6.Cells that we
7.Cells that we
8.Repeat until



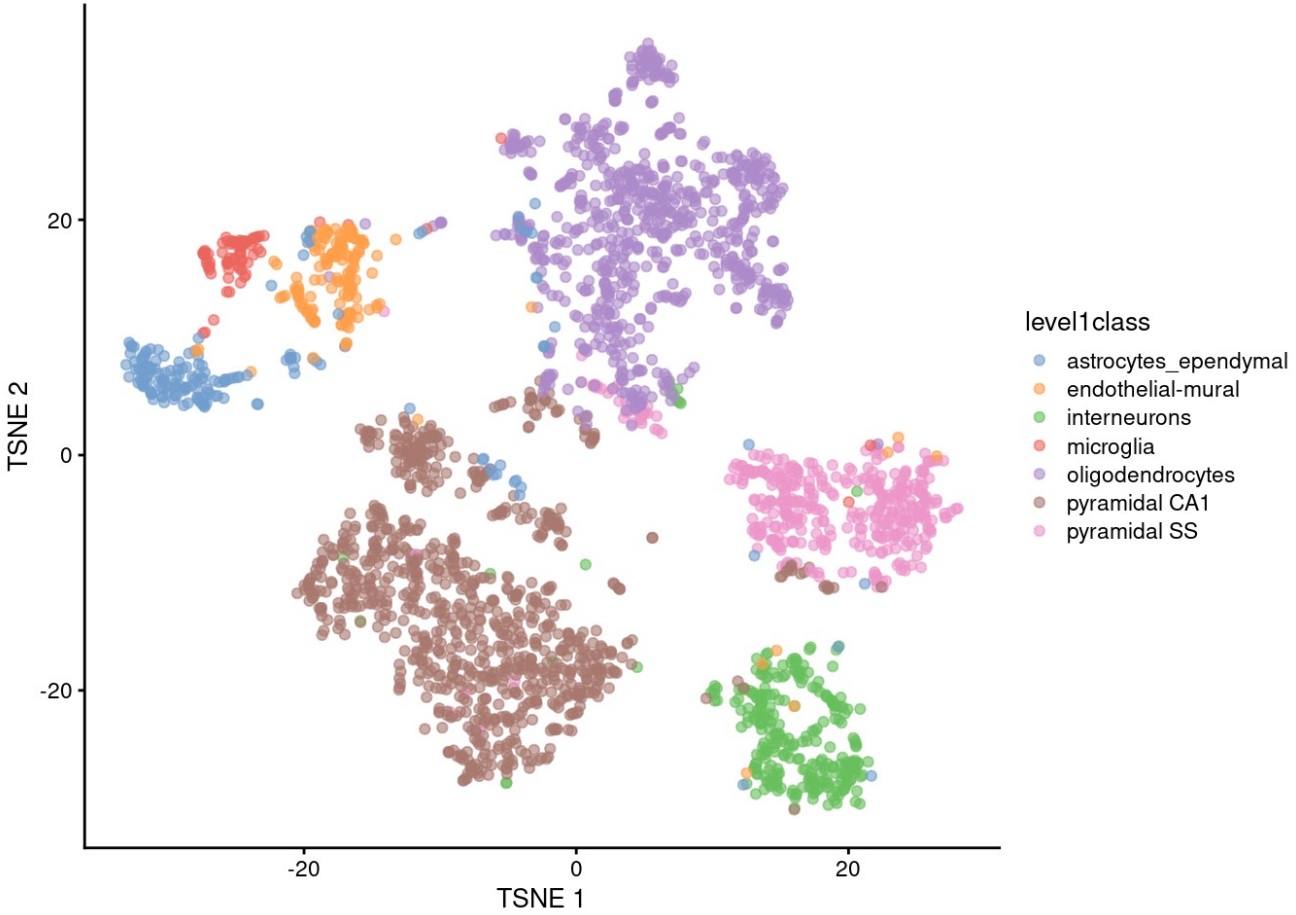
Paper: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

t-SNE: t-distributed Stochastic Neighbourhood Embedding

t-SNE takes high-dimensional gene expression data and creates a 2D or 3D map, where:

- **Cells** with very **similar** expression profiles → **placed close together**
- **Cells** with **different** programs → **placed far apart**

It focuses on preserving local structure → meaning it keeps neighborhoods of similar cells intact.



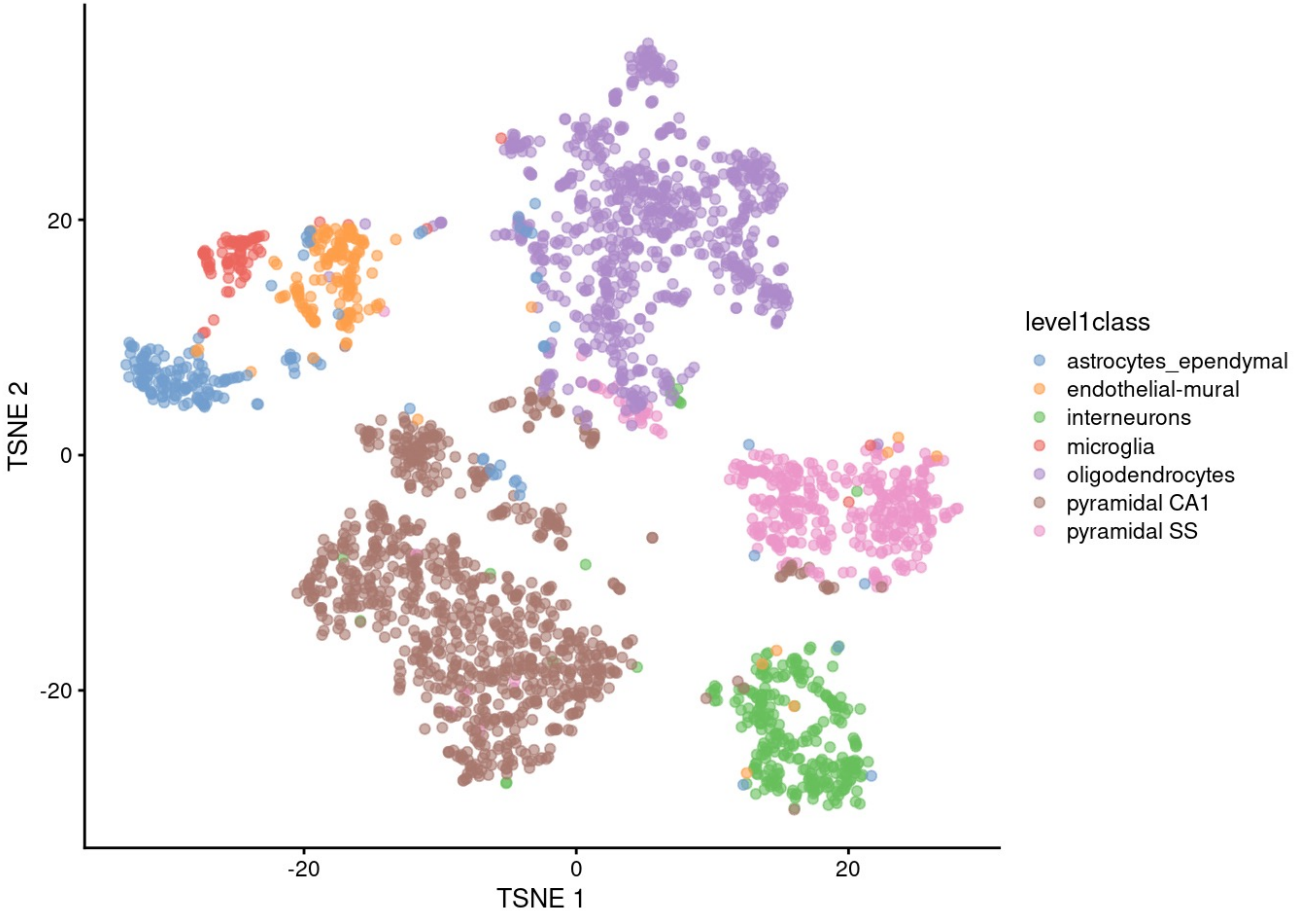
Paper: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

t-SNE: t-distributed Stochastic Neighbourhood Embedding

t-SNE takes high-dimensional gene expression data and creates a 2D or 3D map, where:

- **Cells** with very **similar** expression profiles → **placed close together**
- **Cells** with **different** programs → **placed far apart**

It focuses on preserving local structure → meaning it keeps neighborhoods of similar cells intact.



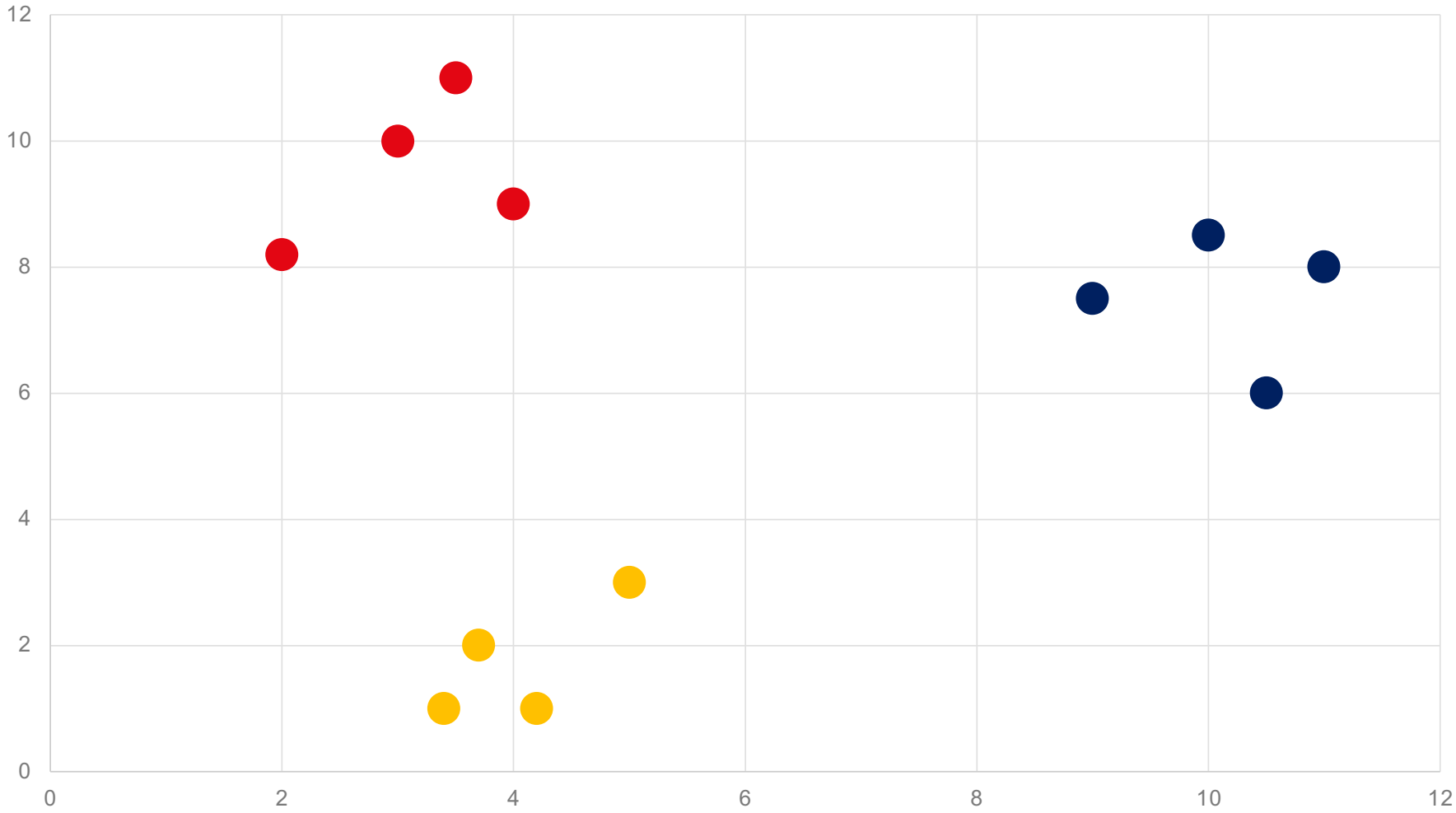
cluster distances cannot be interpreted quantitatively

Paper: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

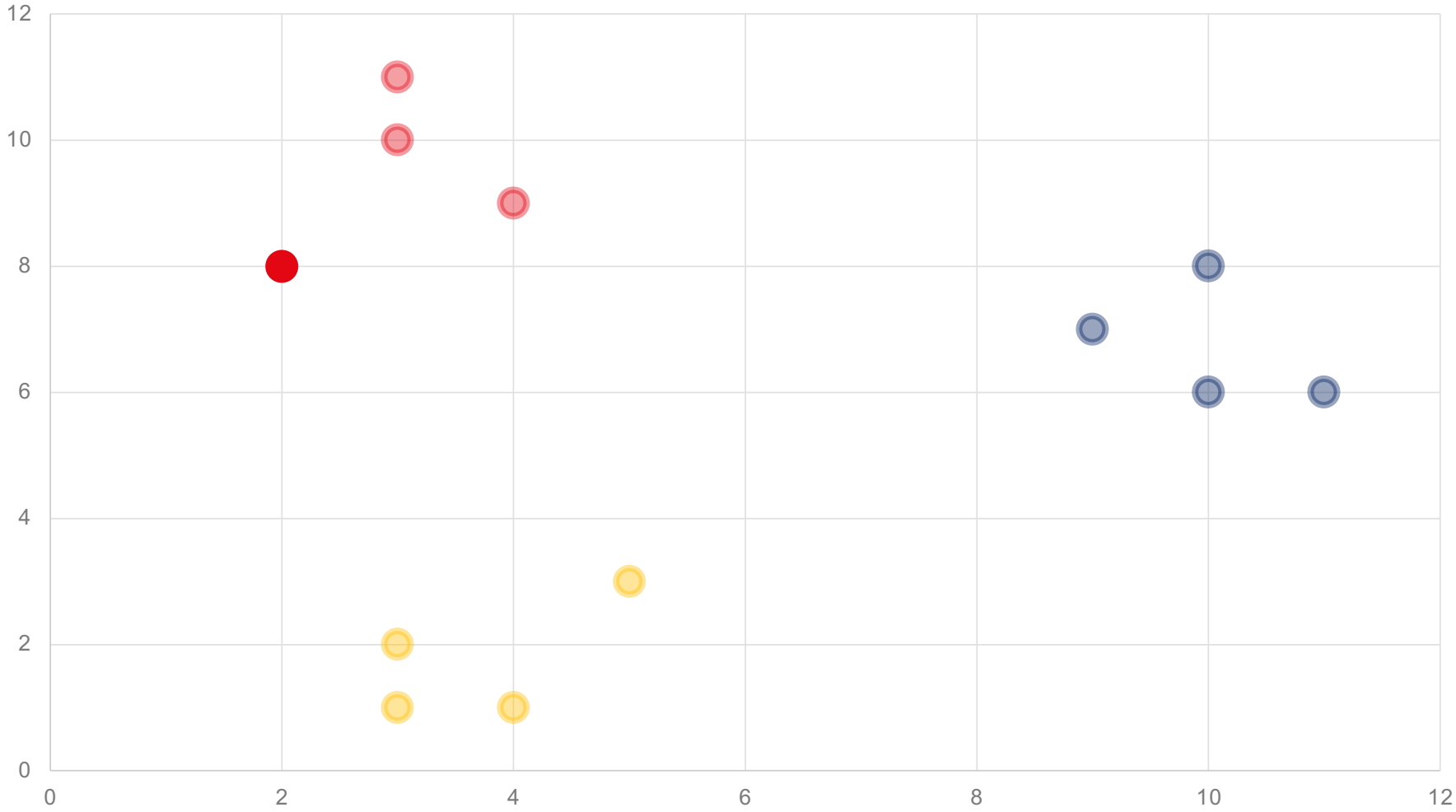
T-SNE recipe

1. Measure similarity between every pair of cells
   - Like saying “who are your closest neighbors?”
2. Create probabilities of being neighbors in high-dimensional space
3. Start placing points randomly in 2D
4. Move points gradually so that:
   - Cells that were similar in gene expression → stay close
   - Cells that were different → push apart
5. Repeat until stable

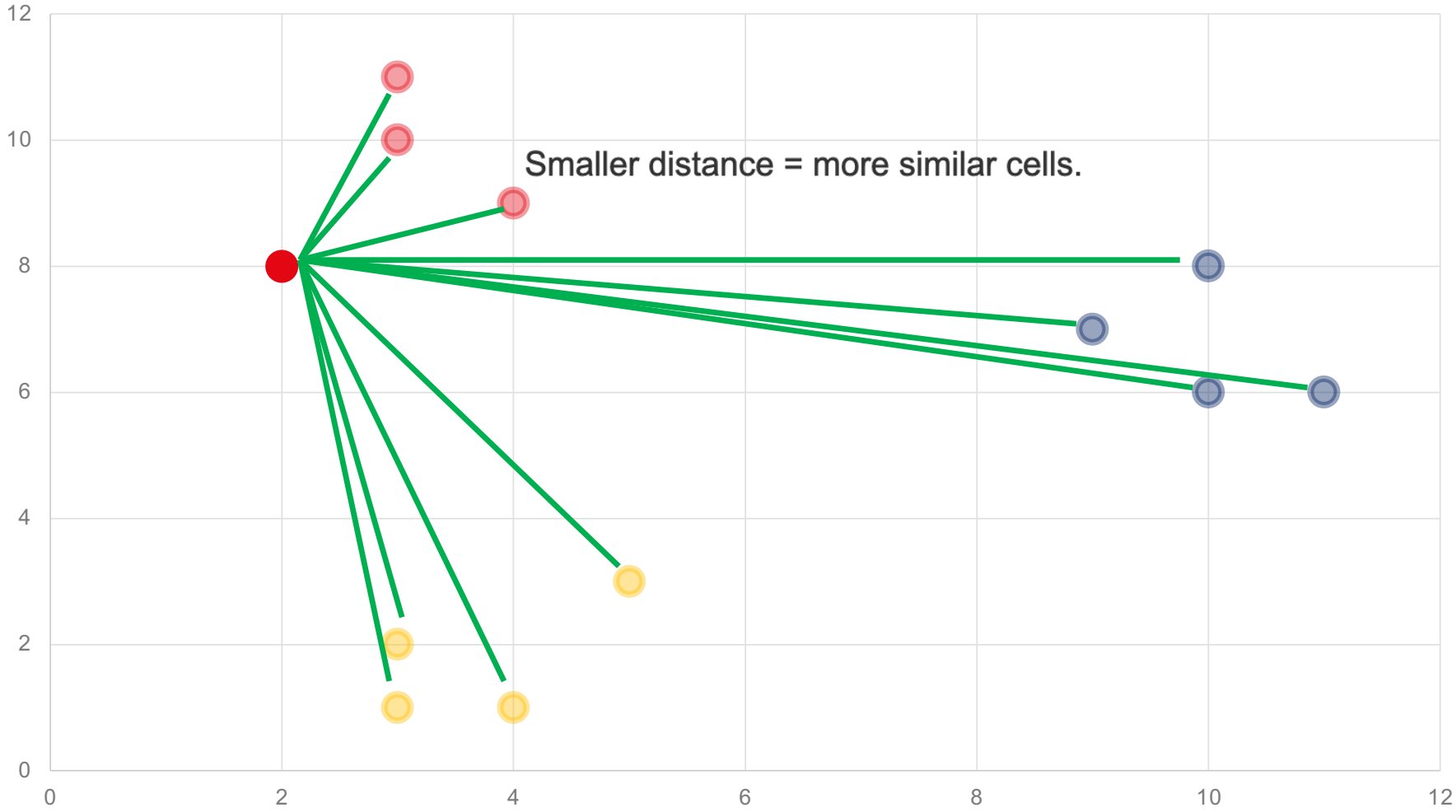
t-SNE - Example



First-Step

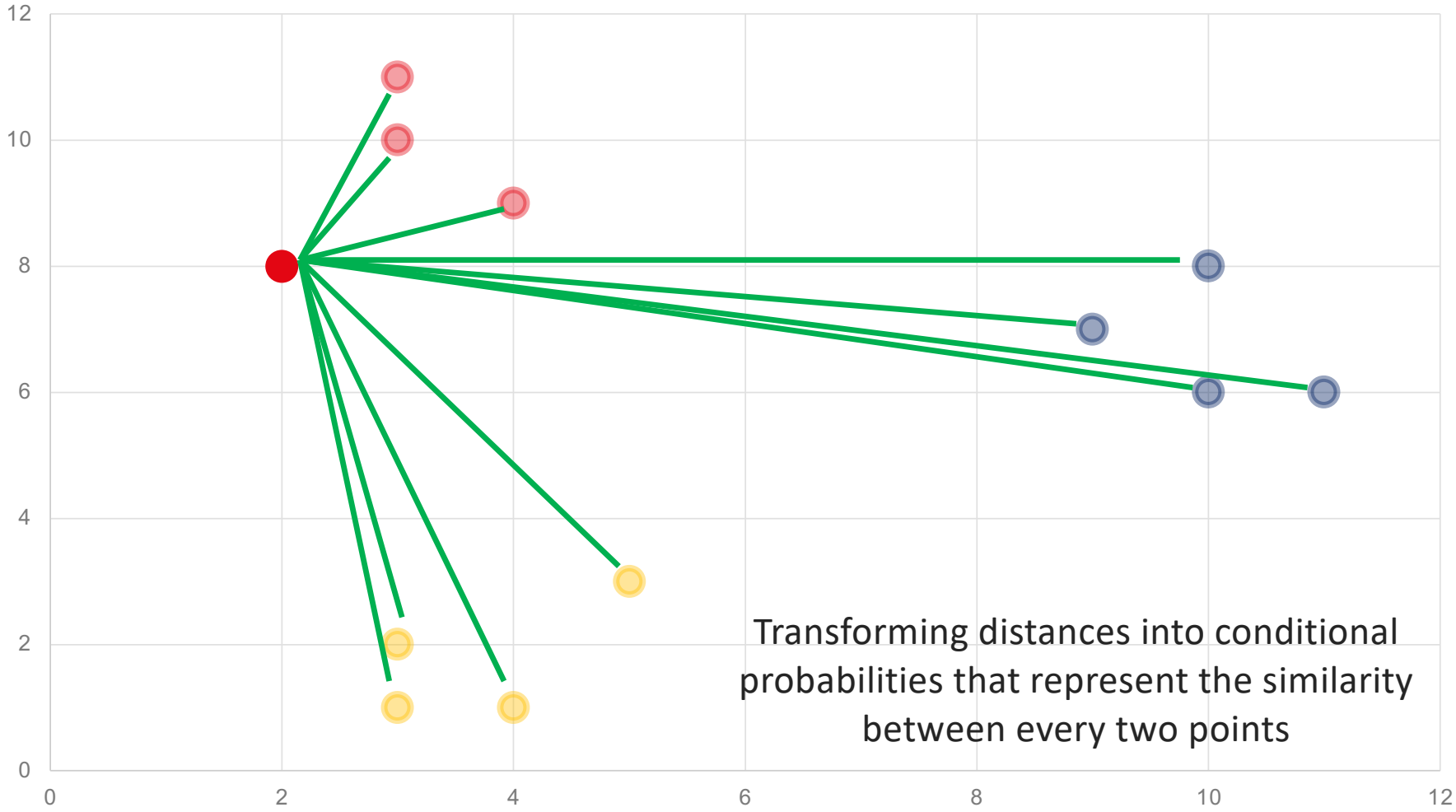


Computing distances



For each cell i, compute the distance to another cell j (usually Euclidean distance in PCA space)

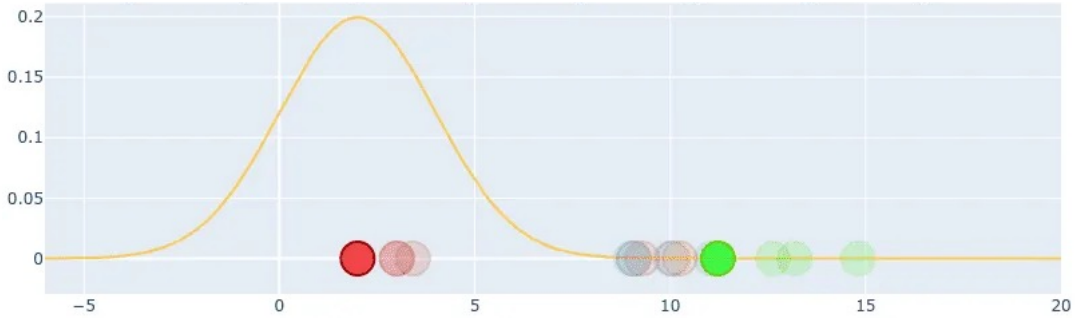
From distance to probability



Conditional Probability

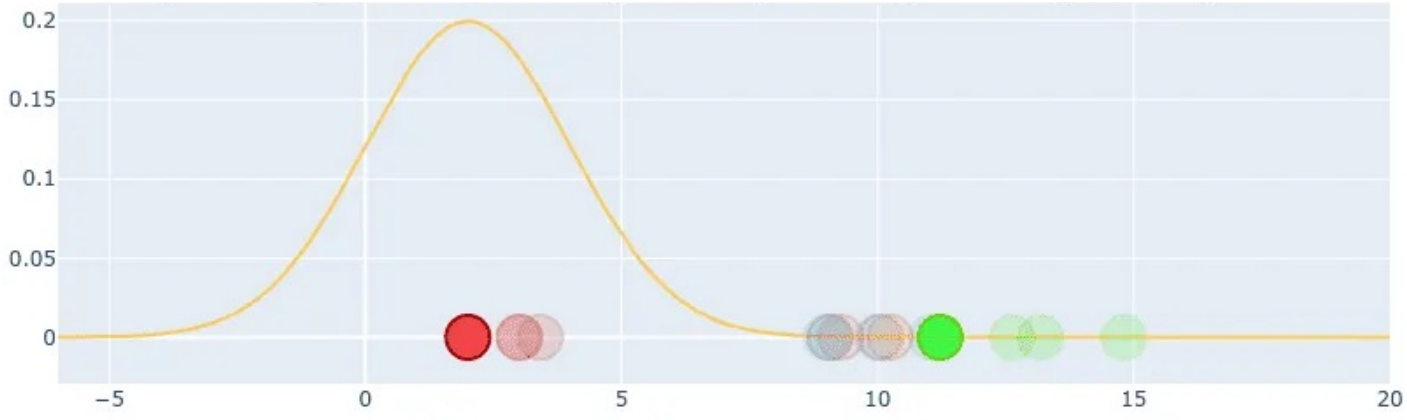
The conditional probability of point x_j to be next to point x_i is represented by a Gaussian centered at x_i with a standard deviation of σ_i

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$



From conditional probability to joint-probability

The conditional probability of point x_j to be next to point x_i is represented by a Gaussian centred at x_i with a standard deviation of σ_i

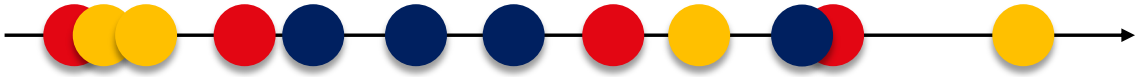


$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

joint probability distribution: $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$

Creating data in a low dimension

A random set of points in 1D

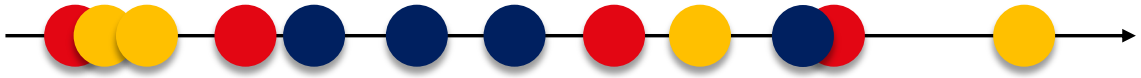


For this set of points, we will create their joint probability distribution but this time we will be using the t-distribution and not the Gaussian

Kullback-Leiber divergence to make the joint probability distribution of the data points in the low dimension as similar as possible to the one from the original dataset

Creating data in a low dimension

A random set of points in 1D

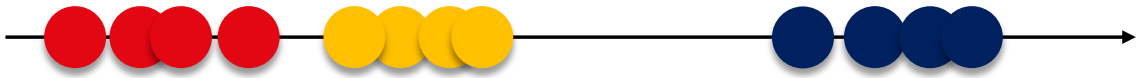


For this set of points, we will create their joint probability distribution but this time we will be using the t-distribution and not the Gaussian

Kullback-Leiber (KL) divergence to make the joint probability distribution of the data points in the low dimension as similar as possible to the one from the original dataset.

Creating data in a low dimension

t-SNE uses **gradient descent** to minimize is the KL divergence of the joint probability distribution P from the high-dimensional space and Q from the low-dimensional space.



Key parameters:

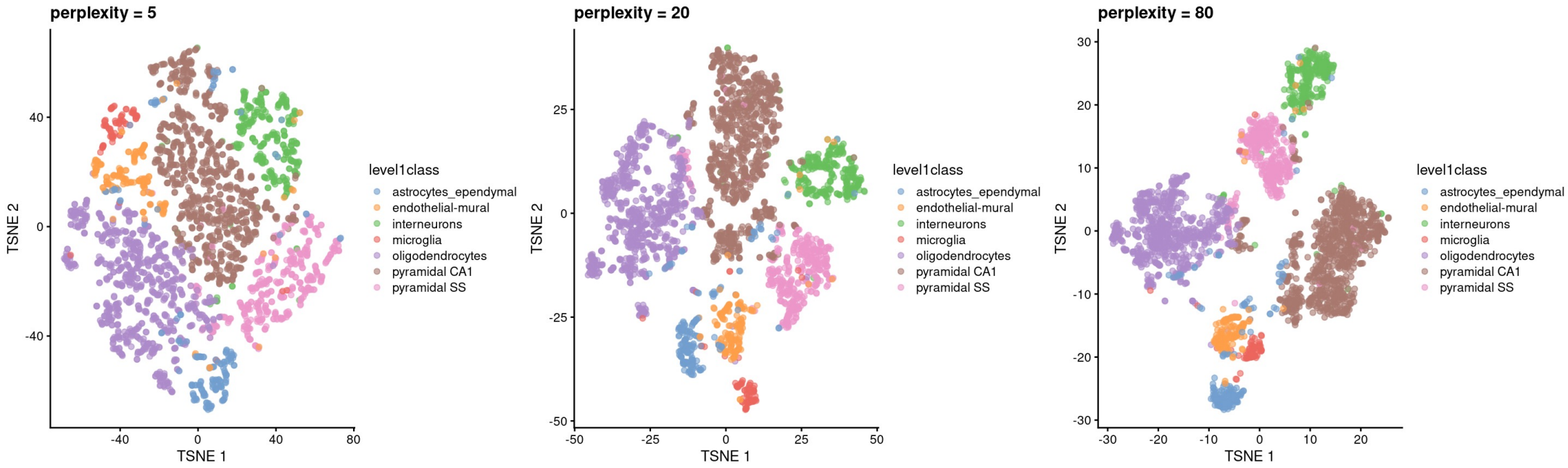
Gradient descent:

- learning rate
- number of iterations

Perplexity: It is used for choosing the standard deviation σ_i of the Gaussian representing the conditional distribution in the high-dimensional space. The model is rather robust for perplexities between 5 to 50, but it has a huge impact on the final plot.

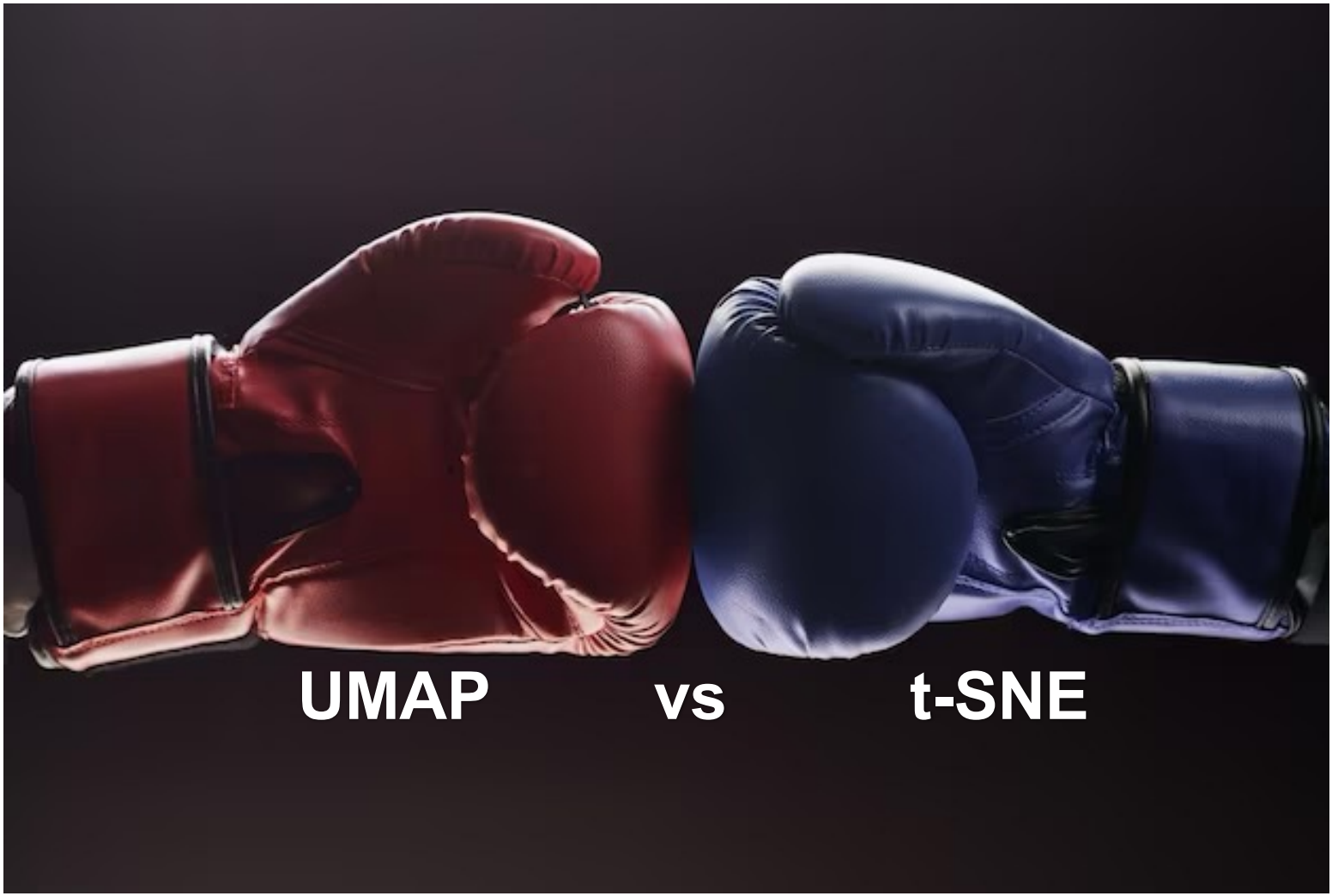
Perplexity

The “perplexity” is an important parameter that determines the granularity of the visualization.



Note: t-SNE involves a random initialization, so we need to set the seed to ensure that the chosen results are reproducible

Non-linear Methods for Dimensionality Reduction



UMAP

Manifold Approximation and Projection

Authors: McInnes L. and Healy J.
Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018

Non-linear dimensionality reduction approach. It offers several advantages over t-SNE:

PRO:

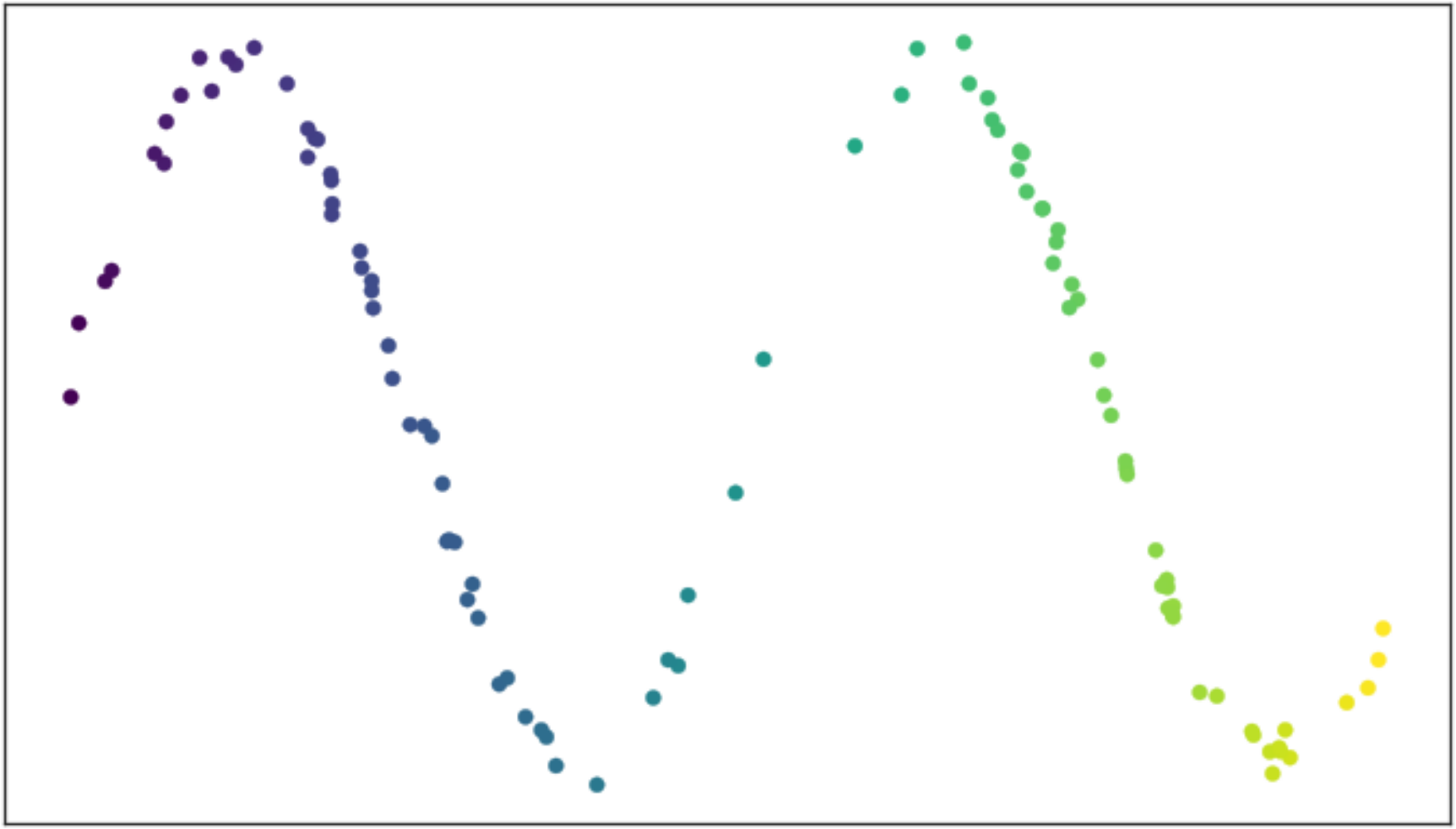
- increased speed
- It can use any distance metrics
- better preservation of the data's global structure
- Defines both LOCAL and GLOBAL distances
- Can be applied to new data points
- Works on original data, but best on PCA reduced dimension (default in Seurat)

UMAP Theory

Step 1: construct the initial high-dimensional graph, UMAP builds something called a "**fuzzy simplicial complex**". This is really just a representation of a weighted graph, with edge weights representing the likelihood that two points are connected.

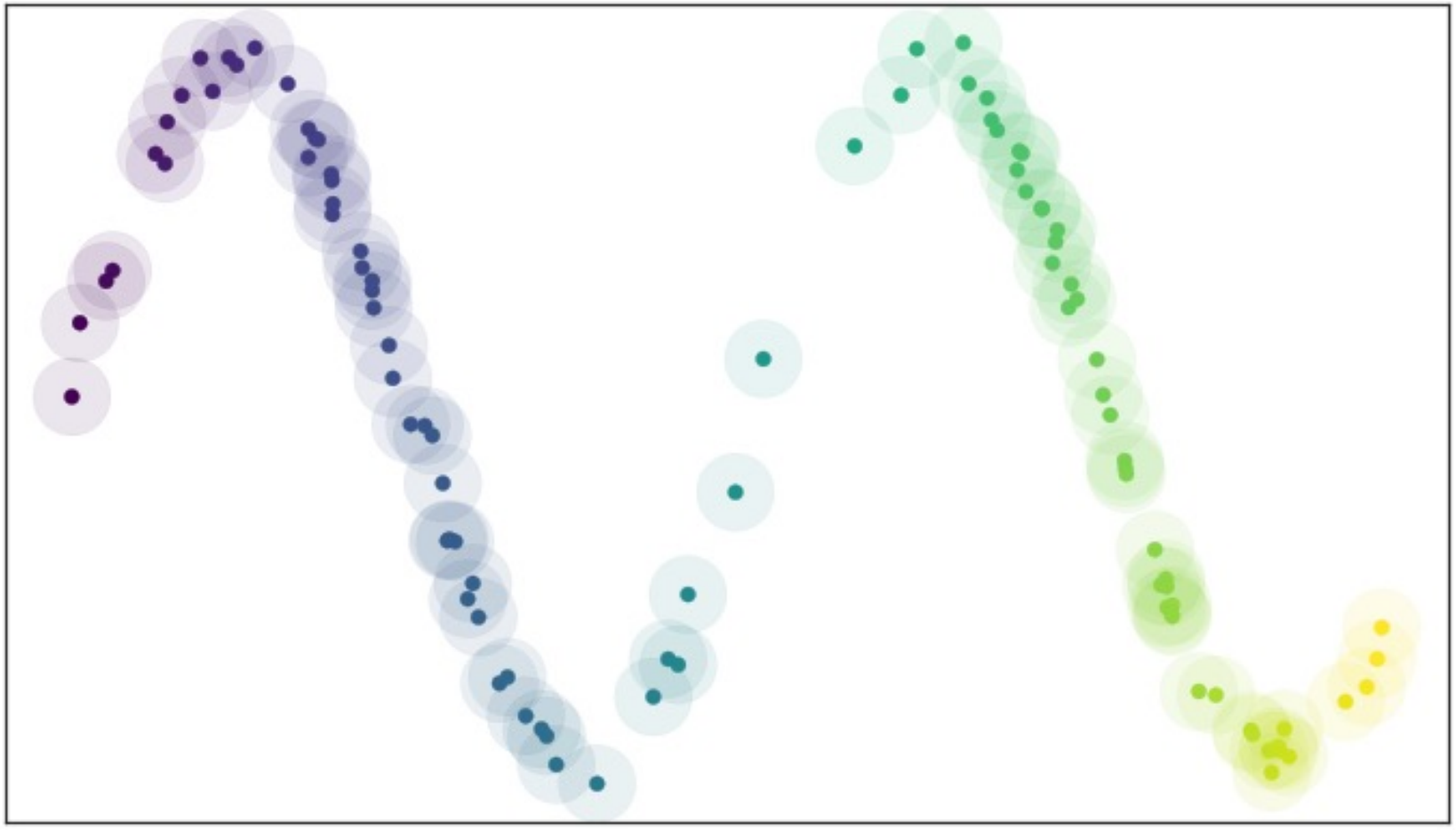
UMAP Theory

Step 1: UMAP extends a radius outwards from each point



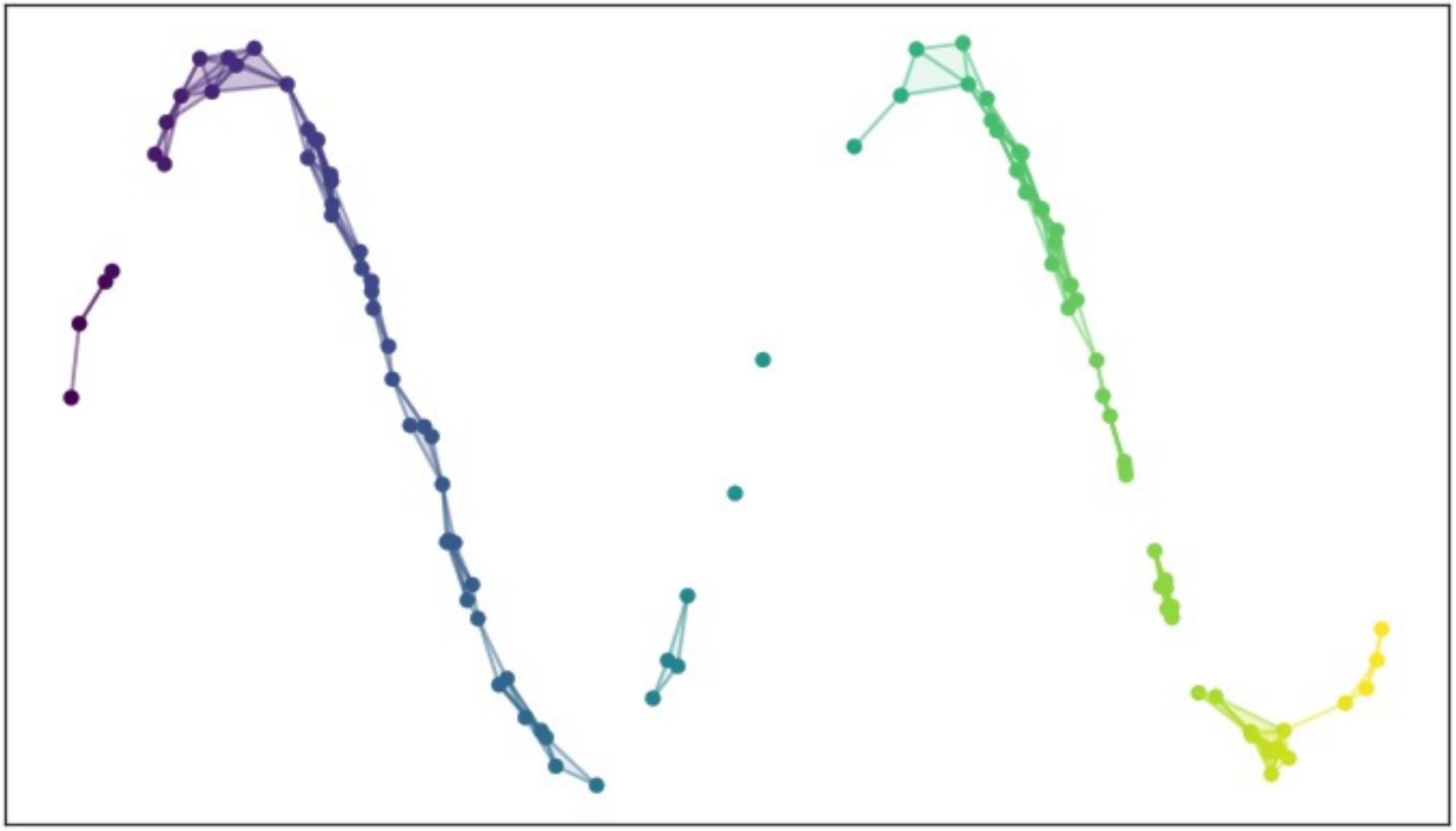
UMAP Theory

Step 1: UMAP extends a radius outwards from each point



UMAP Theory

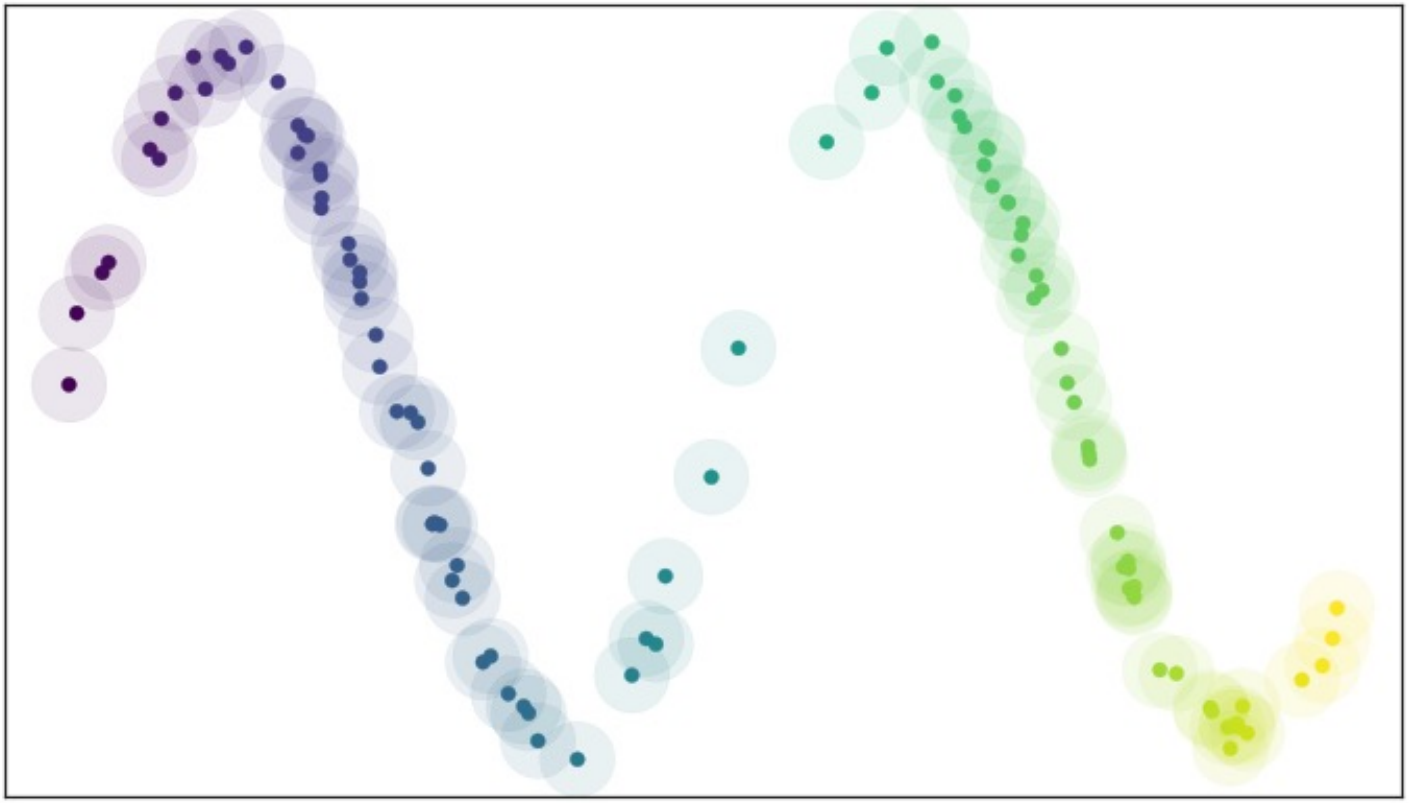
Step 1: Connect points when those radii overlap



UMAP Theory

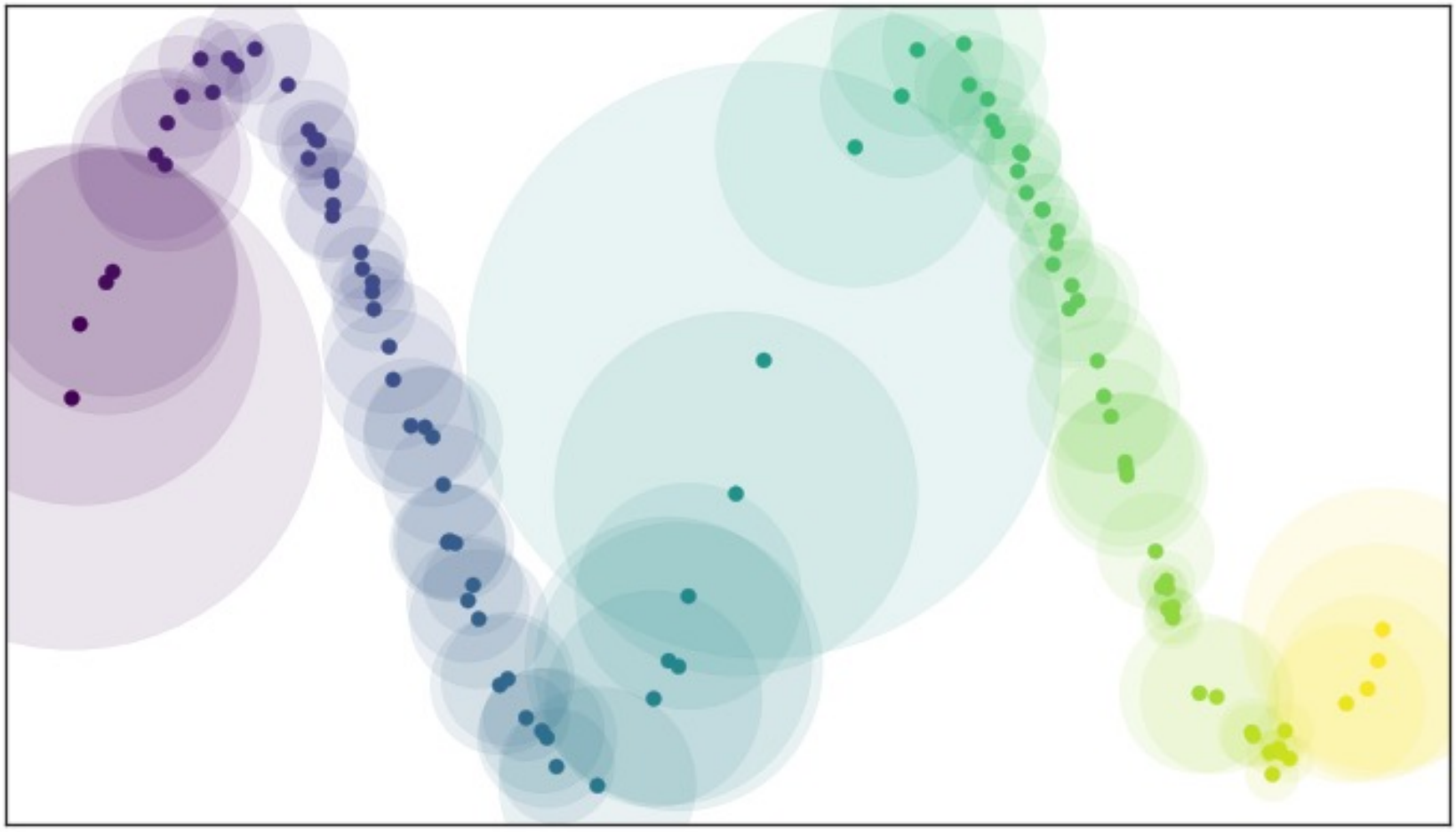
Choosing this radius is critical:

- too small a choice will lead to small, isolated clusters
- too large a choice will connect everything together



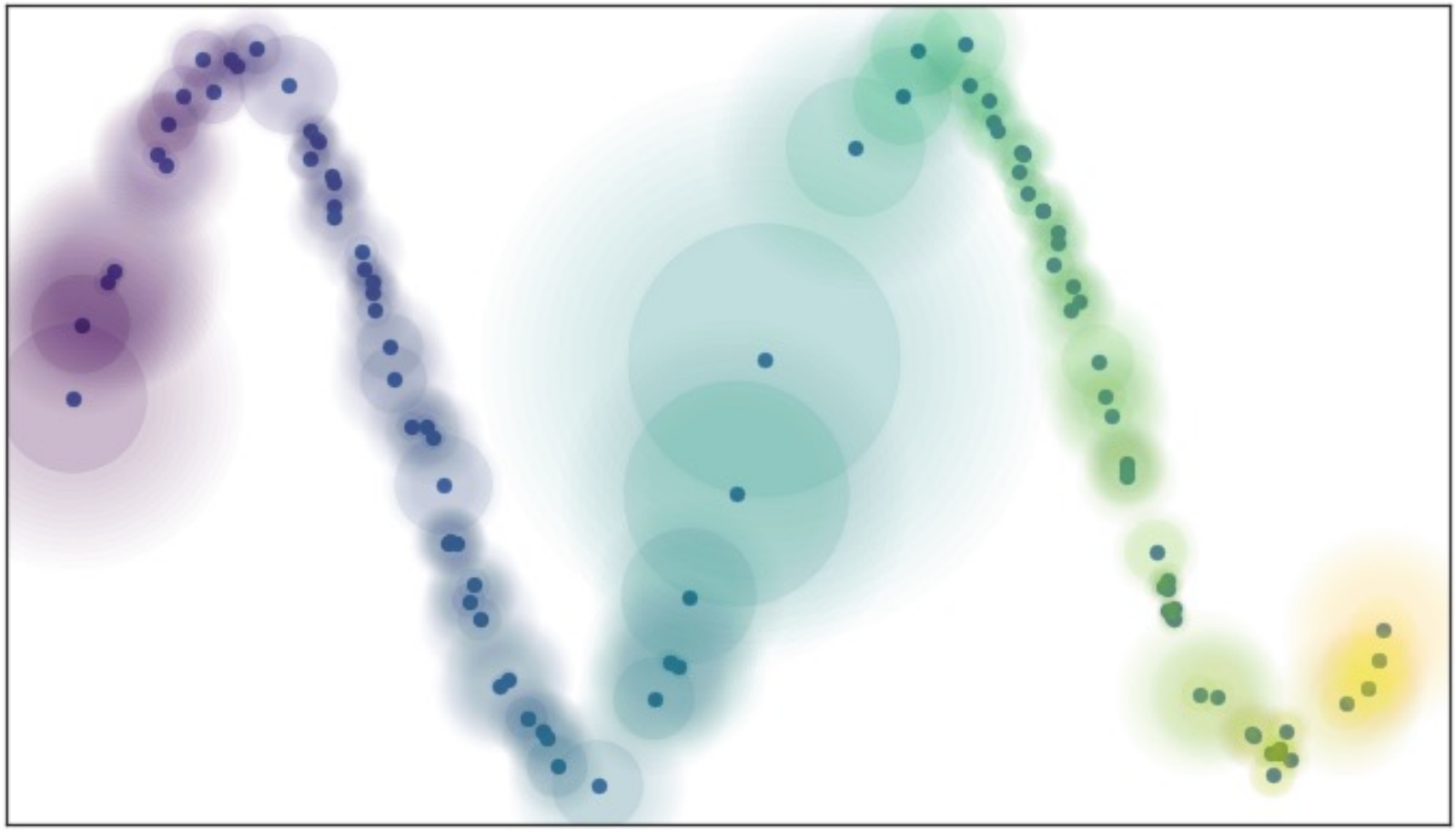
UMAP Theory

Rather than using a fixed radius, UMAP uses a variable radius determined for each point based on the distance to its **k-th nearest neighbours**.



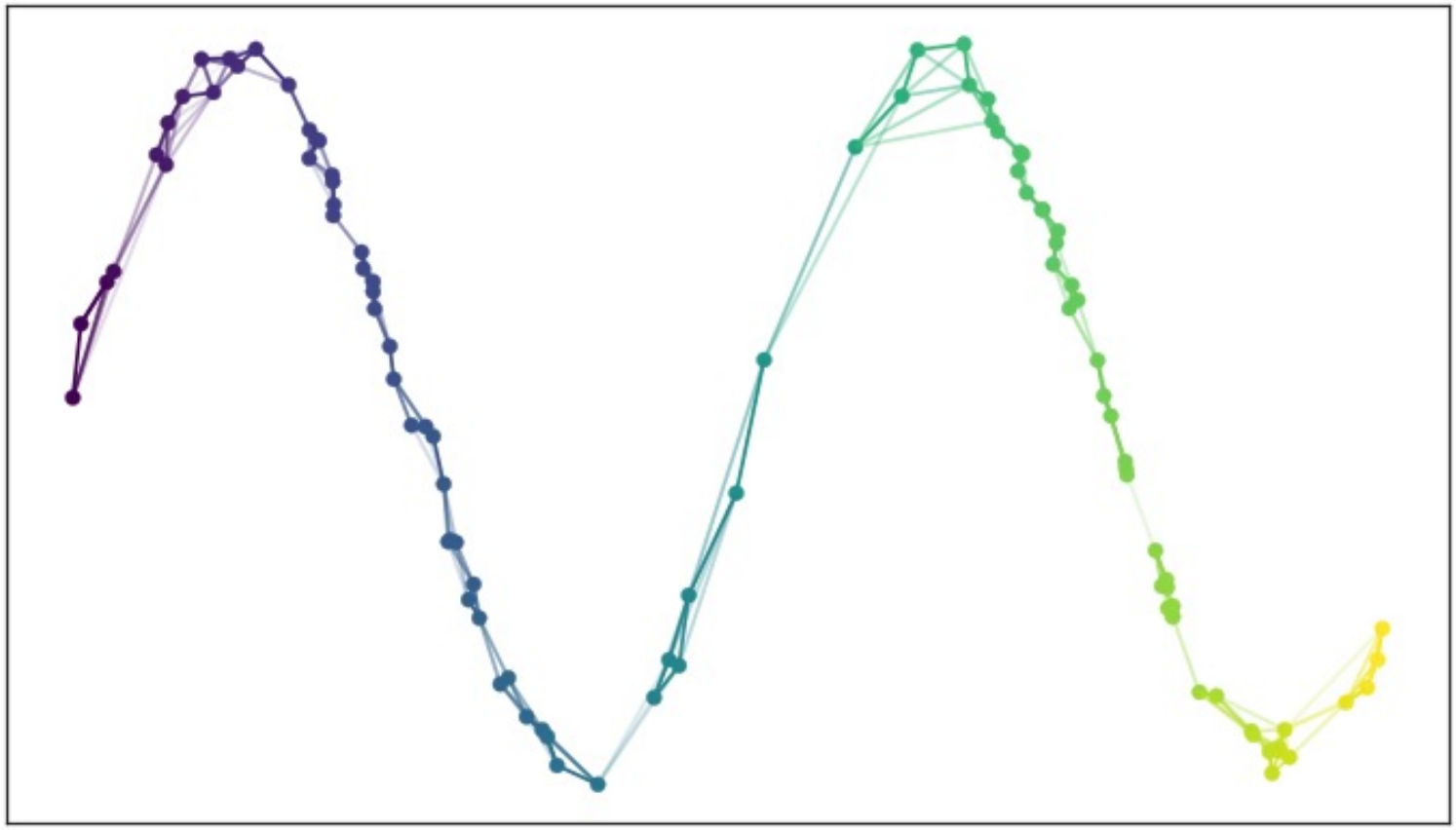
UMAP Theory

Within this local radius, connectedness is then made “fuzzy” by making each connection a probability, with further points less likely to be connected.



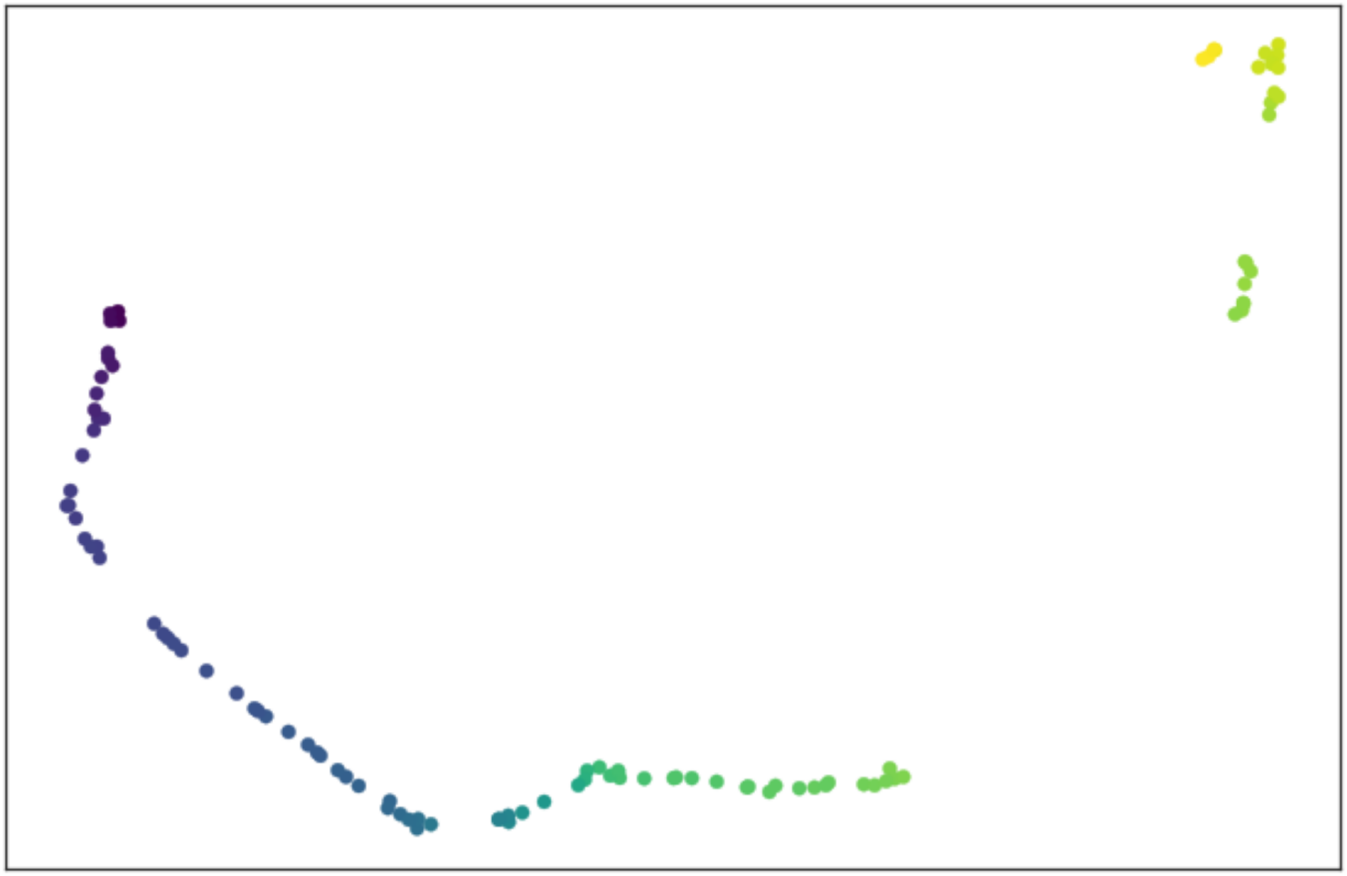
UMAP Theory

All points must be connected to at least its closest neighbouring point.
The final output of this process is a weighted graph, with edge weights representing the likelihood that two points are “connected” in our high-dimensional manifold.



Final Step

Once the final, fuzzy simplicial complex is constructed, UMAP projects the data into lower dimensions essentially via a force-directed graph layout algorithm



Key hyper-parameters

1. **n_neighbors:** Determines the number of neighboring points considered when computing the local structure of the data. It defines the balance between local and global structure in the UMAP embedding.

   - **Typical Values: Ranges from 5 to 50. For scRNA-Seq data, values around 10-30 are often used.**

     Lower values focus on capturing the local structure (more fine-grained clusters).
     Higher values provide a more global view of the data, potentially merging cluster

2. **min_dist:** Controls how tightly UMAP packs points together in the low-dimensional space. It sets the minimum distance between points in the embedded space.

   - **Typical Values: Between 0.001 and 0.5. For scRNA-Seq, a common default is around 0.1.**

   Lower values (e.g., 0.001) will result in more compact clusters, making it easier to identify tight groupings.

   Higher values (e.g., 0.5) allow for more spread-out points, which can reveal broader patterns but may blur smaller clusters.

3. **metric**: Defines the distance metric used to measure how similar or dissimilar two data points are. Common metrics include ‘euclidean,’ ‘manhattan,’ ‘cosine,’ and more.
4. **4. n_components:** Specifies the number of dimensions in the output space. For visualization, this is typically set to 2 (for 2D plots) or 3 (for 3D plots).

Notes on UMAP

1. Hyperparameters really matter

Run UMAP multiple times with a variety of hyperparameters, how is the projection affected by its parameters?

2. Cluster sizes in a UMAP plot mean nothing

The size of clusters relative to each other is essentially meaningless

3. Distances between clusters might not mean anything

The distances between clusters is likely to be meaningless

4. Spurious clustering can be observed

Due to Random noise that doesn't always look random (e.g. low values of n_neighbors)

5. UMAP is stochastic

Different runs with the same hyperparameters can yield different results

Consideration

Choosing one over the other depends heavily on the dataset and the goals of your analysis

UMAP is more time-saving due to the clever solution in creating a rough estimation of the high dimensional graph instead of measuring every point

UMAP gives a better balance between local versus global structure, thus overall gives a more accurate presentation of the global structure. This will come in handy in trajectory analysis

Summary

UMAP:

- **Better for preserving global structure**: UMAP often provides a better representation of larger structures and relative distances between clusters, making it more effective for capturing hierarchical relationships.
- **More interpretable distances**: UMAP's distances between clusters are more meaningful, so it is commonly used when comparing clusters and observing relationships at a broader level.
- **Fast and scalable**: UMAP is faster and scales well with large datasets, making it a preferred choice for single-cell datasets with many thousands of cells.

t-SNE:

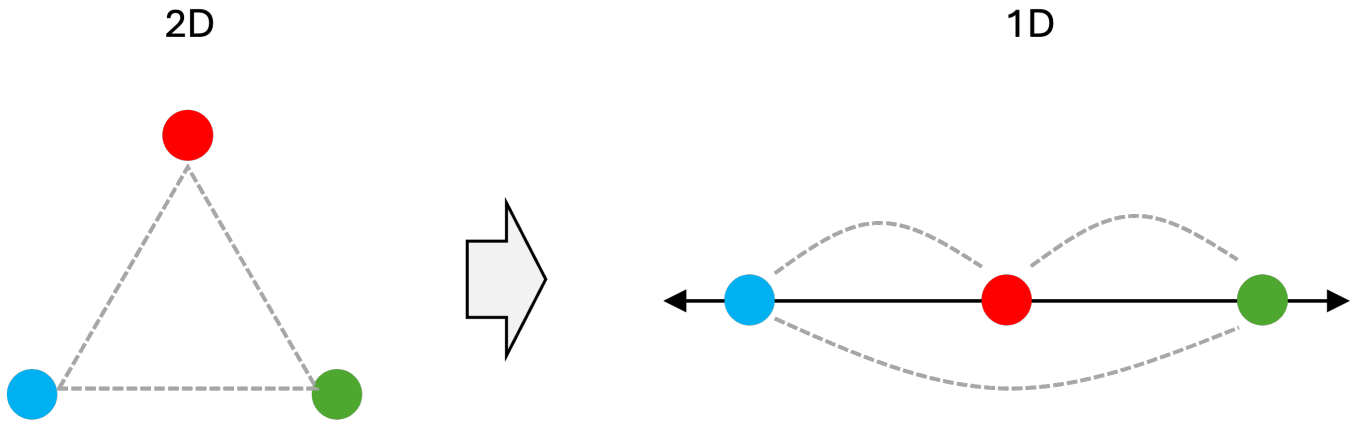
- **Better for local structure**: t-SNE is optimized to preserve local structure, so it excels at highlighting small or subtle differences within clusters.
- **More reliable for smaller datasets**: For smaller datasets or when the primary interest is identifying fine-grained details in cell subpopulations, t-SNE can sometimes give clearer, tighter clusters.
- **Limited interpretability of distances**: The distance between clusters in t-SNE plots may not be very meaningful, so it's less ideal for analyzing relationships across clusters.

In summary:

- Use **UMAP** if you need a broad overview of the dataset and want to capture global patterns.
- Use **t-SNE** when focusing on identifying distinct subpopulations or fine-grained differences.

Consideration

it is mathematically impossible to avoid losing information when mapping data from high to low dimensions, these algorithms inevitably lose some aspect of the data, either by distortion or ommision, when plotting it in lower dimensions.



conclusions one draws from a dimensionality reduction plot have some probability of not actually being true of the data

Skepticism about this methods

PLOS COMPUTATIONAL BIOLOGY

PERSPECTIVE

The specious art of single-cell genomics

Tara Chari[1], **Lior Pachter**[1,2]*

1 Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, United States of America, **2** Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, United States of America

* lpachter@caltech.edu

In UMAP and t-SNE plots specific **cluster shapes, separations, and proximities** can appear different depending on algorithm parameters

Clusters can sometimes be artifacts of the method rather than true biological distinctions.

Groups or clusters that appear well-separated in the 2D plot might not actually be as distinct in the high-dimensional space

UMAP and t-SNE are valuable tools for exploratory analysis, but let's use them with caution and validation

Chari, T., Banerjee, J. & Pachter, L. Plos Comp Biology (2023)