Swiss Institute of
Bioinformatics

Dimensionality Reduction

Luciano Cascione, PhD
Bioinformatics Core Unit

LUCIANO CASCIONE, PHD

BELLINZONA, OCT. 30TH 2024

What for?

scRNA-Seq data is composed by thousand of genes:

- “Remove” **redundancies** in the data
- Identify the **most relevant** information (find and filter noise)
- Reduce **computational time** for downstream procedures
- **Facilitate clustering**, since some algorithms struggle with too many dimensions
- Data **visualization**

DR: Don’t

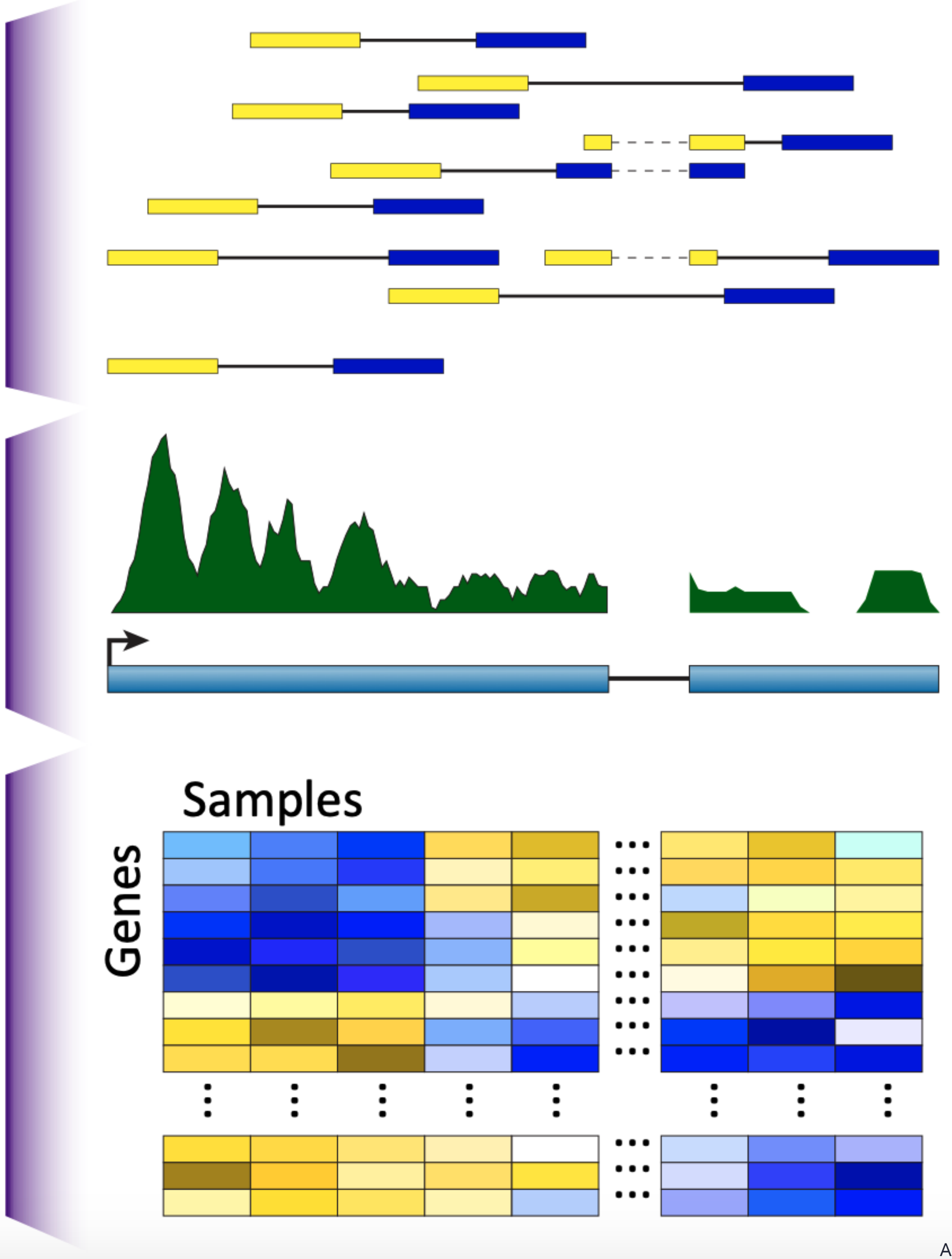
- They are not perfect representation of the high dimension
- One does loose information
- What is close in the projection might actually be far.
- What is far might actually be close
- Conclusions (specially biologically relevant conclusions) should NOT be drawn baed on the dimensionality reduction.

It is all about matrix

Raw RNA-Seq reads

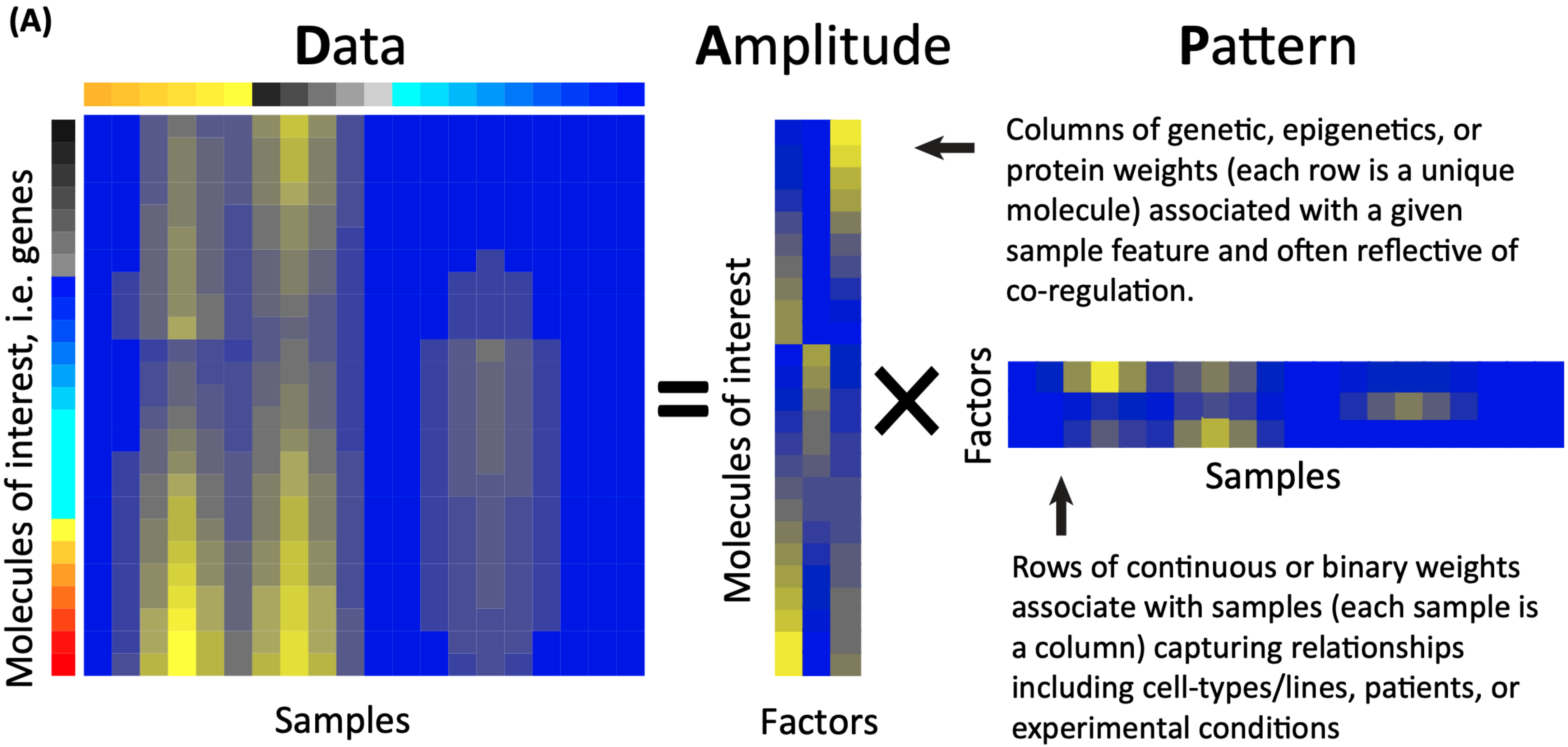
Alignment and quantification

Normalization and log-transformation



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

Matrix Factorization

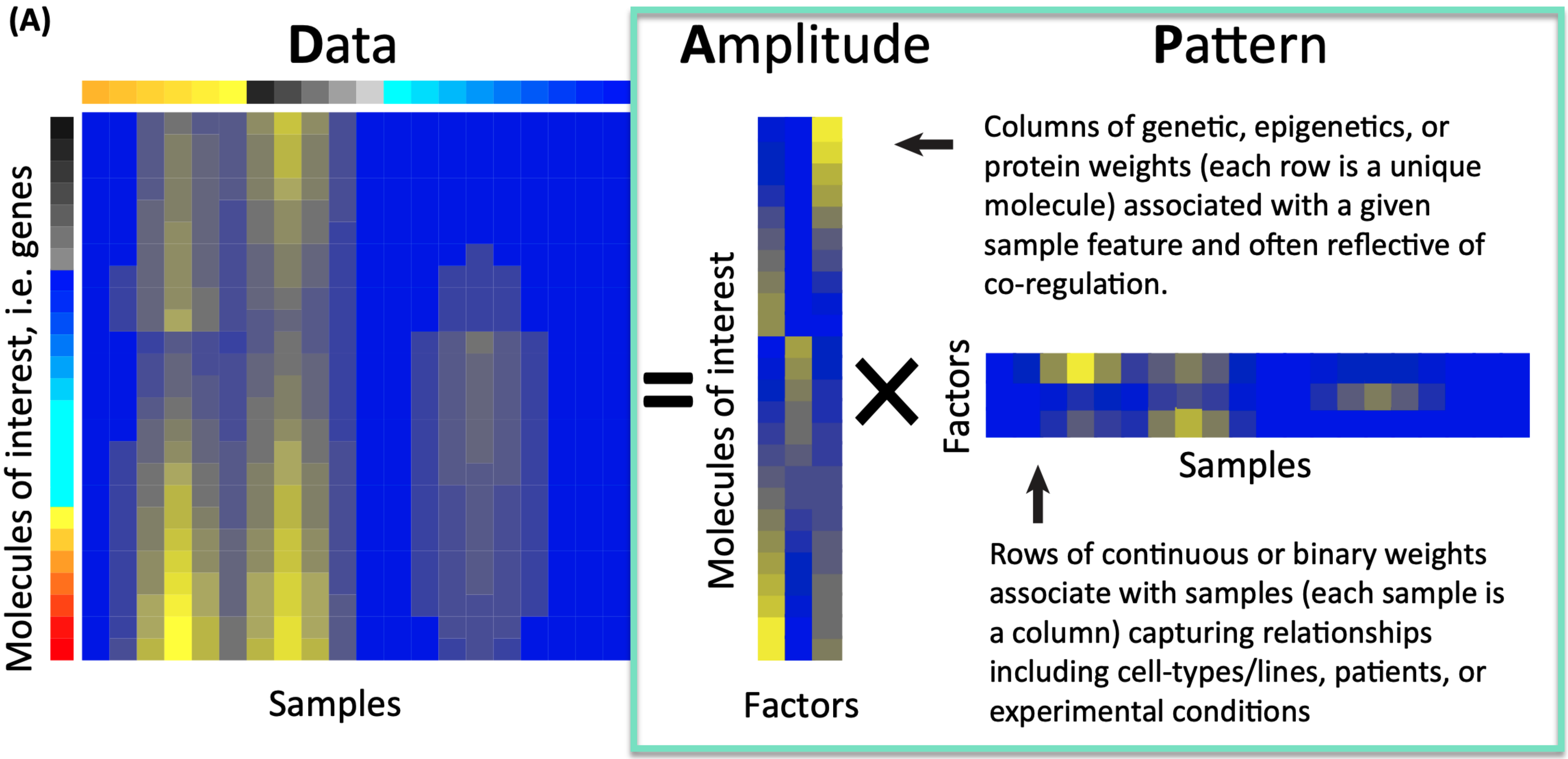


Columns of genetic, epigenetics, or protein weights (each row is a unique molecule) associated with a given sample feature and often reflective of co-regulation.

Rows of continuous or binary weights associate with samples (each sample is a column) capturing relationships including cell-types/lines, patients, or experimental conditions

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

Matrix Factorization



Columns of genetic, epigenetics, or protein weights (each row is a unique molecule) associated with a given sample feature and often reflective of co-regulation.

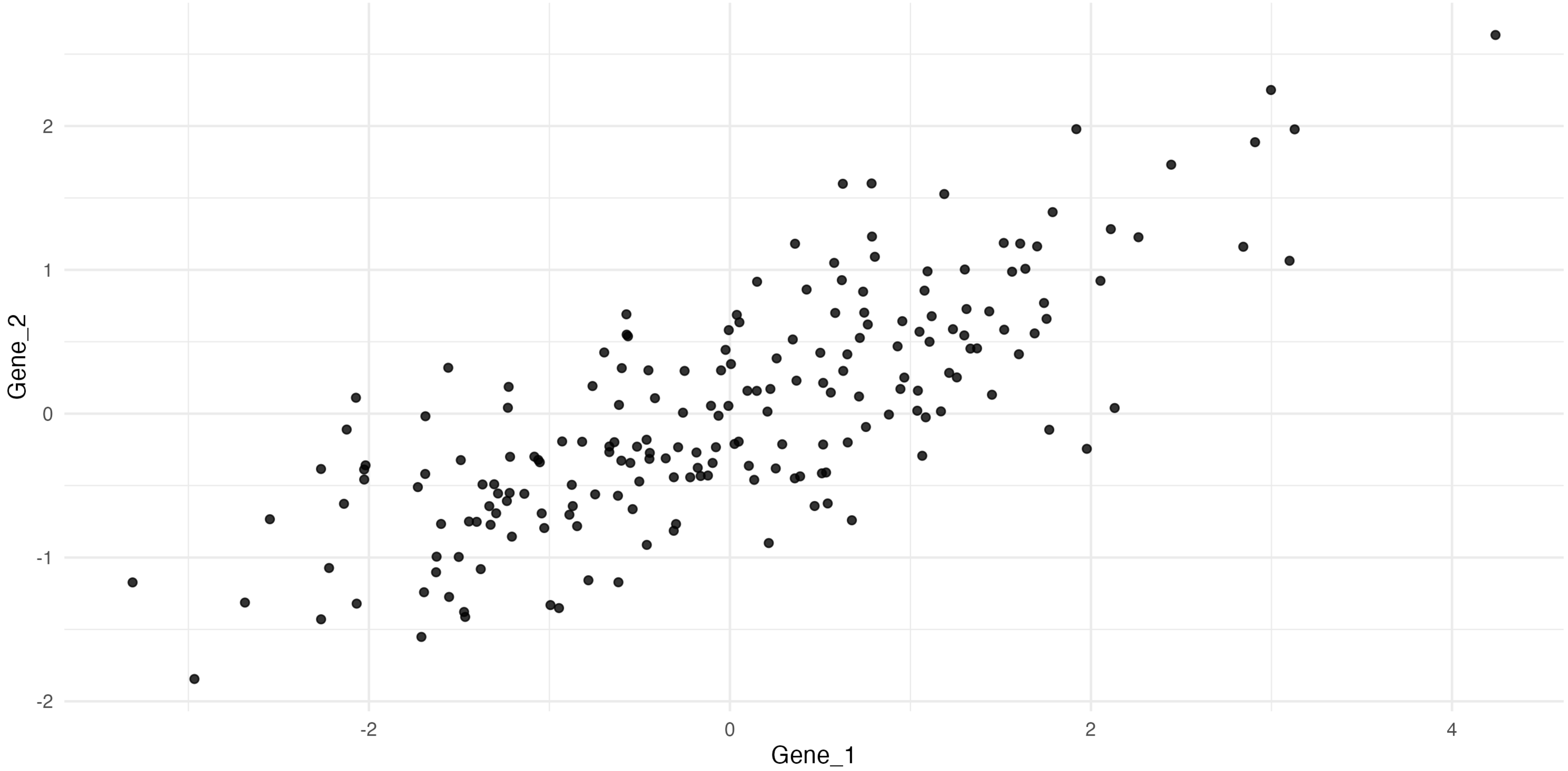
Rows of continuous or binary weights associate with samples (each sample is a column) capturing relationships including cell-types/lines, patients, or experimental conditions

From the Amplitude and Pattern matrices derive biological insights

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

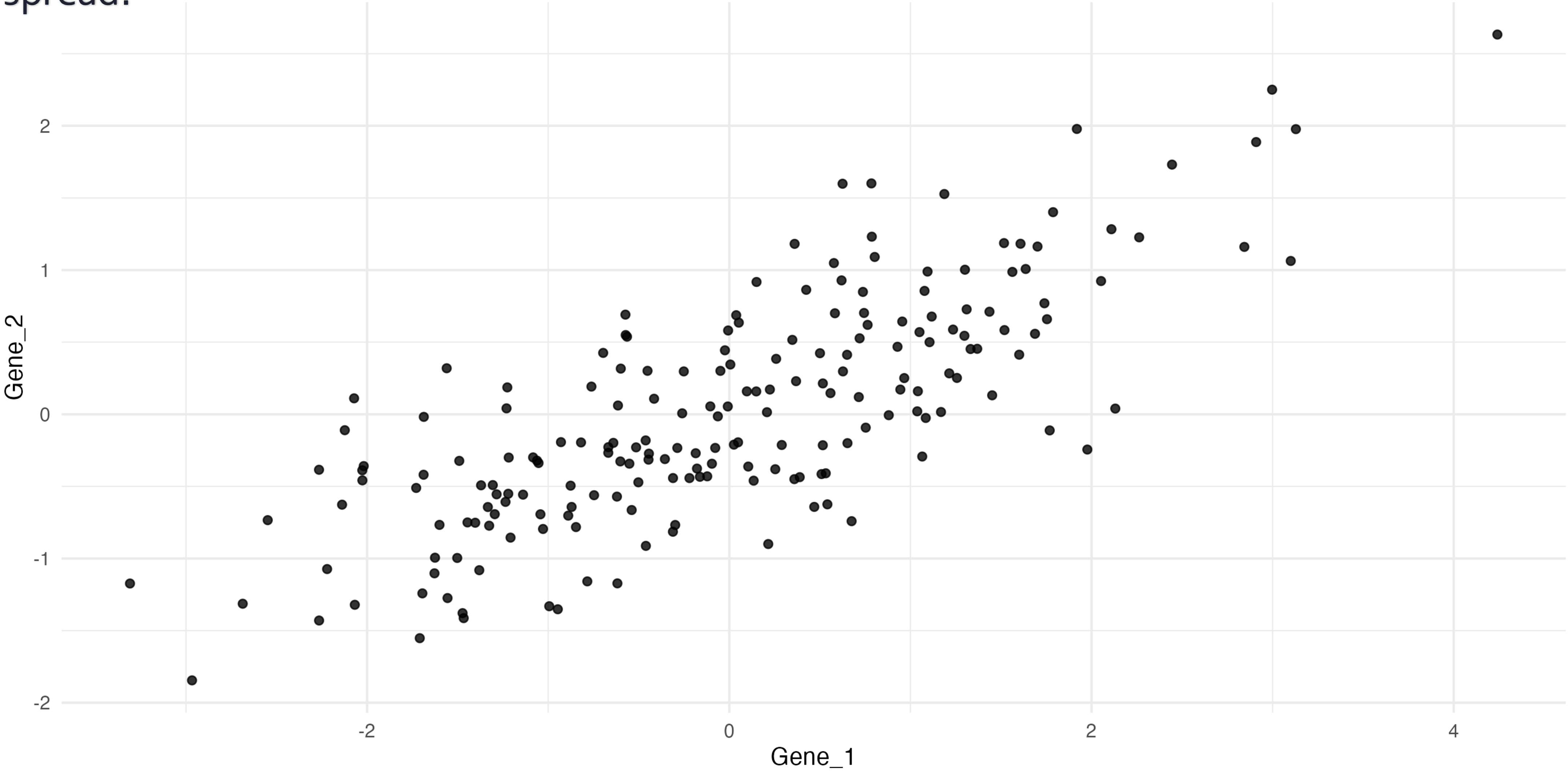
Principal Component Analysis

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



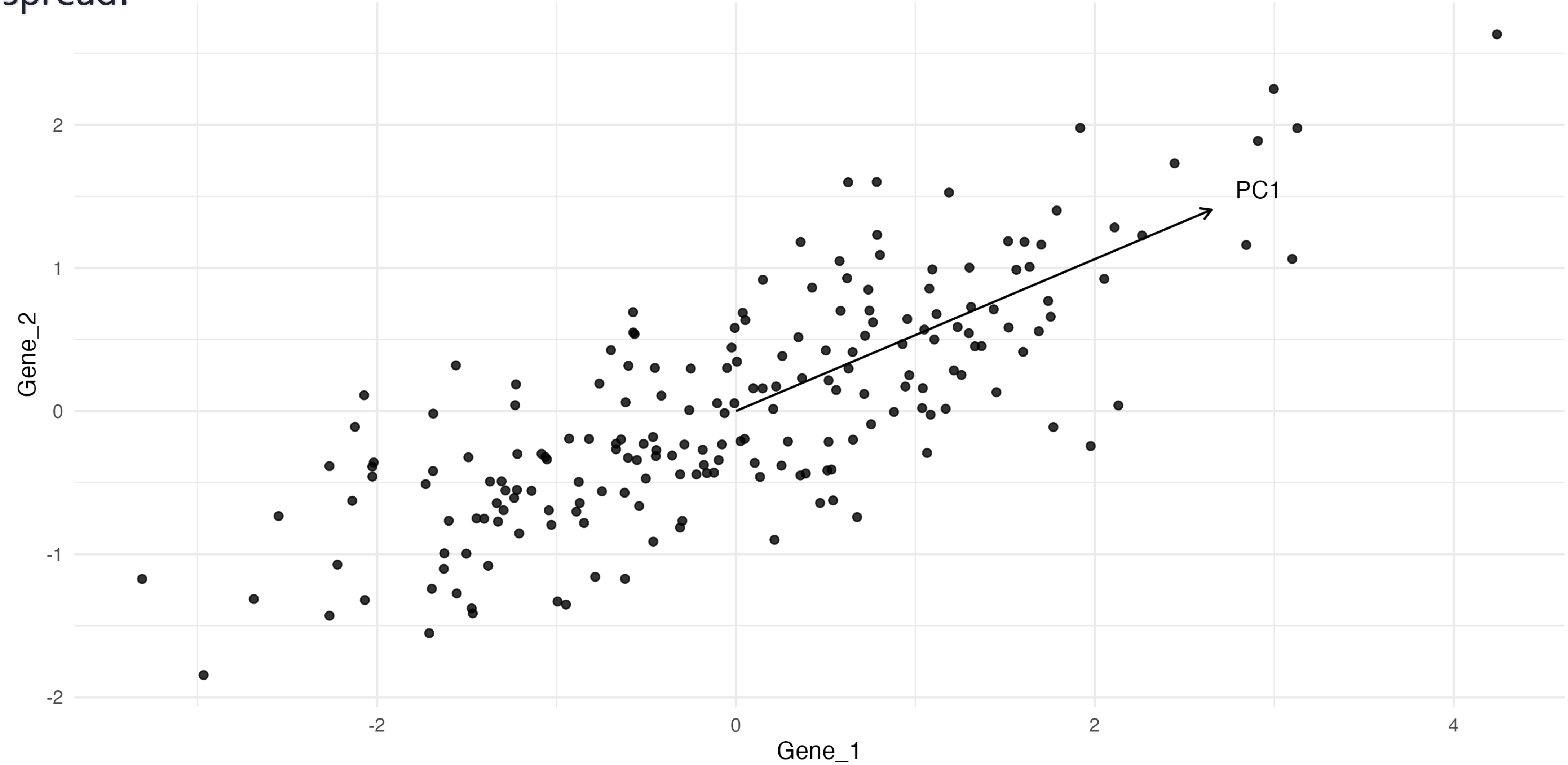
Principal Component Analysis

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



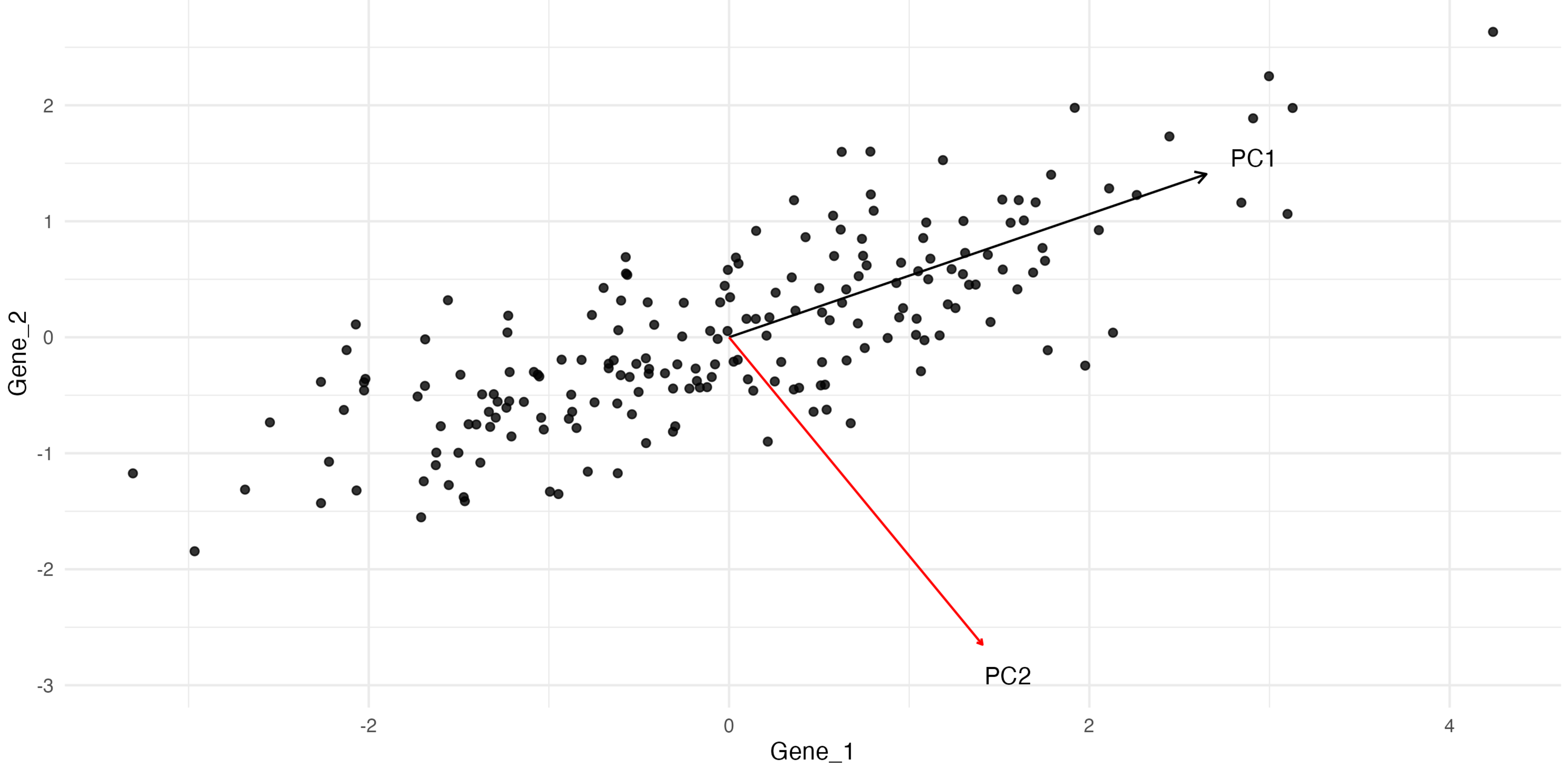
Principal Component Analysis

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



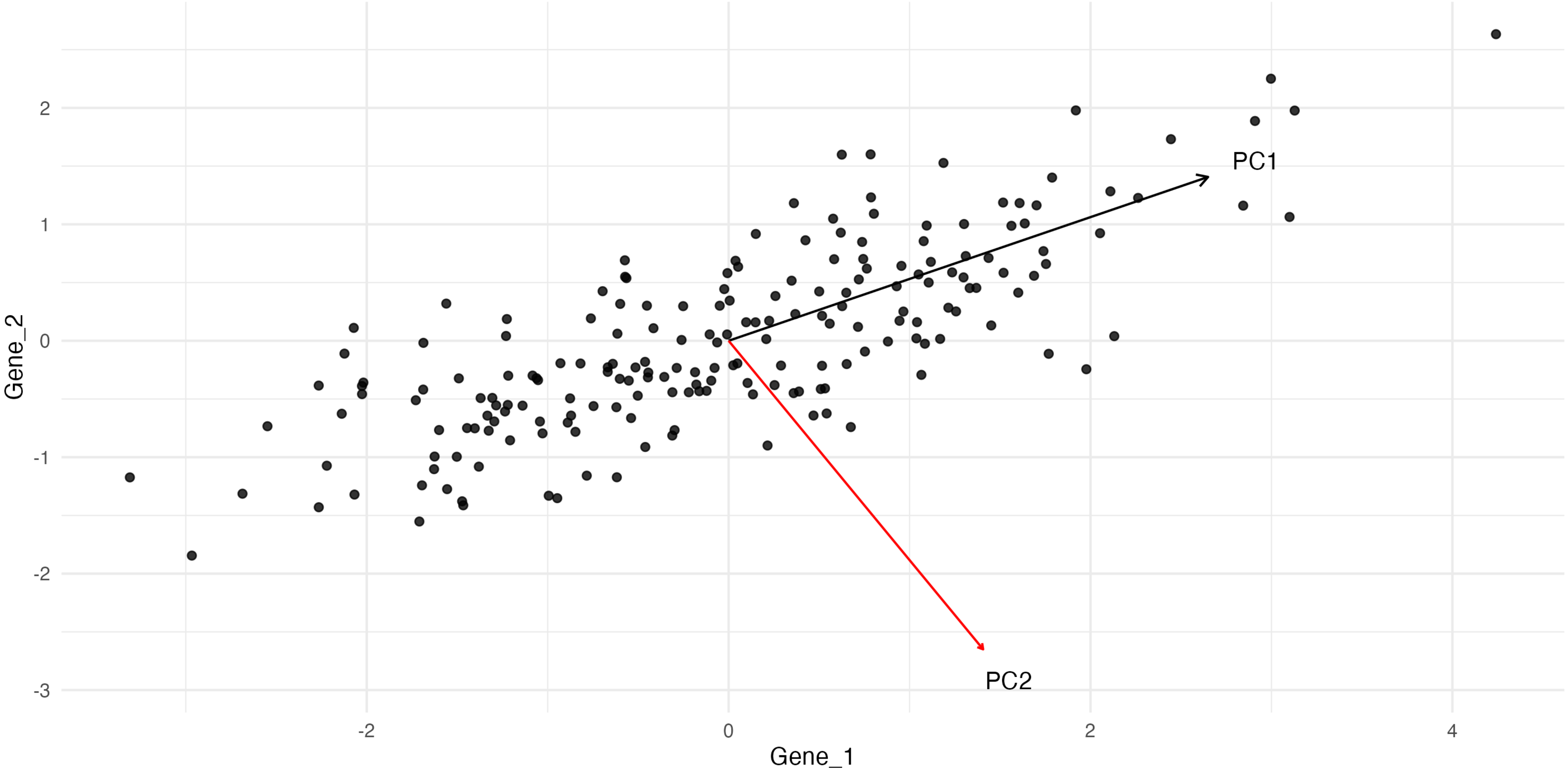
Principal Component Analysis

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



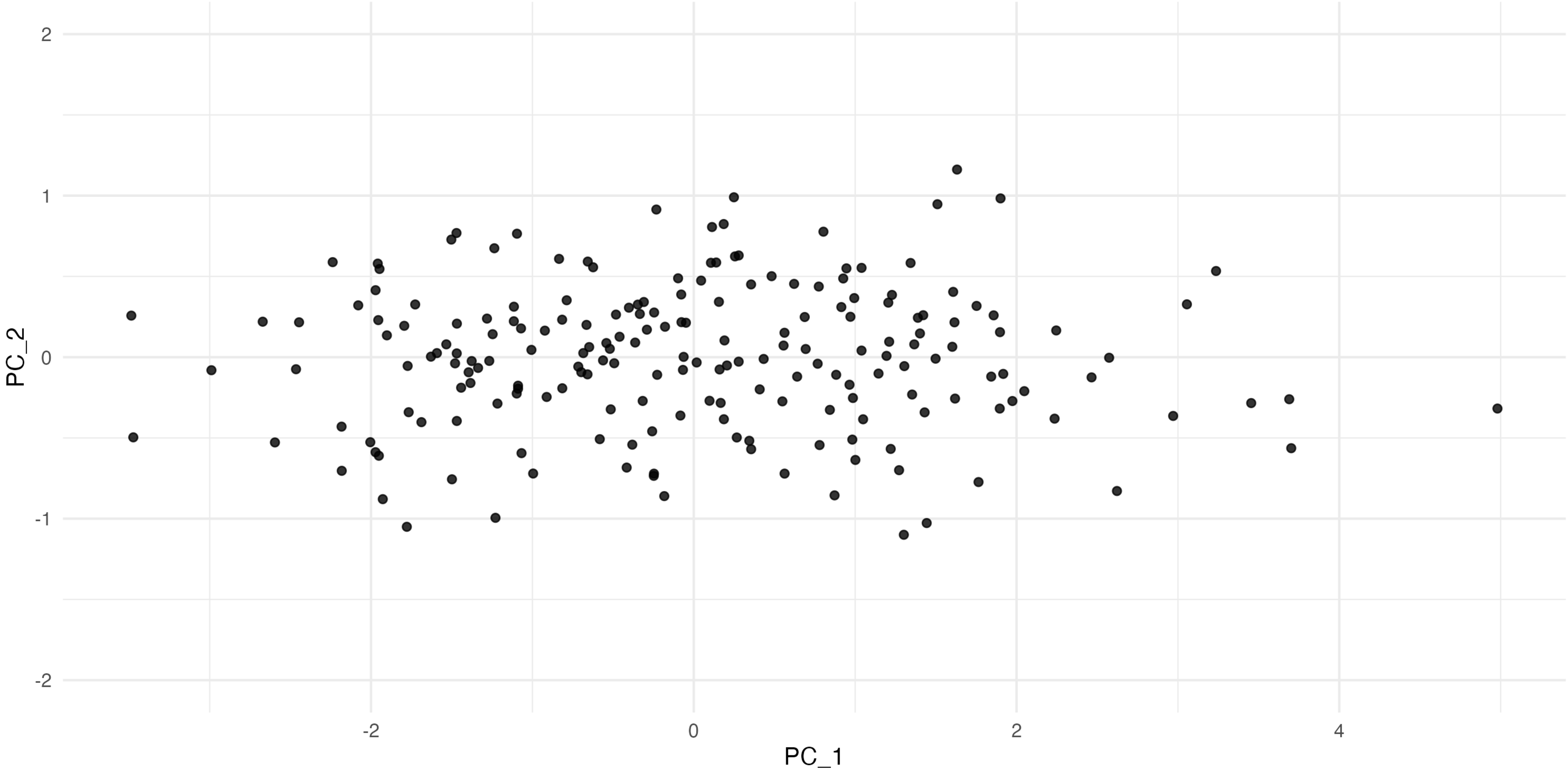
Principal Component Analysis

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



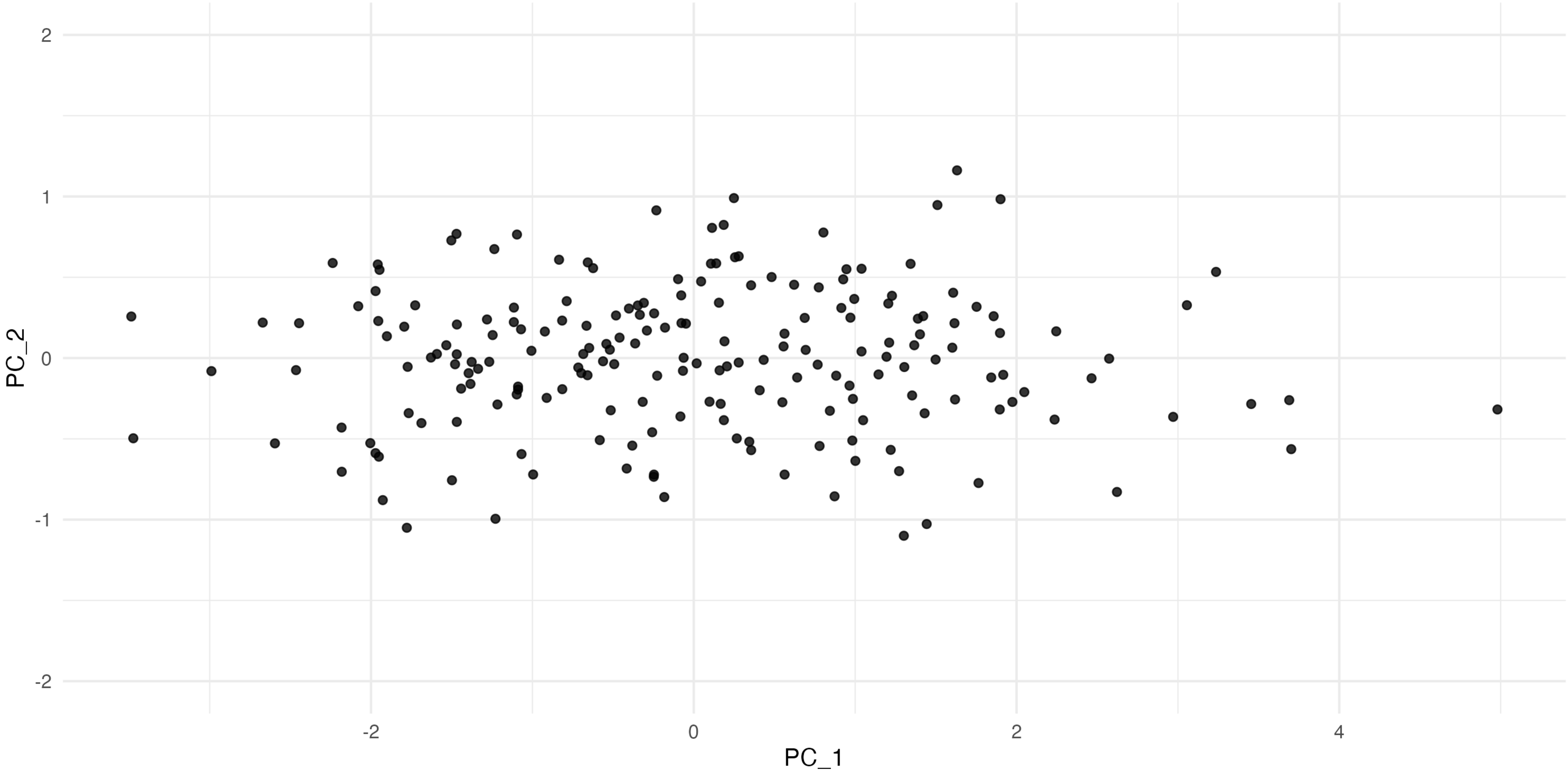
Principal Component Analysis

New axis that are linear combination of the original axes



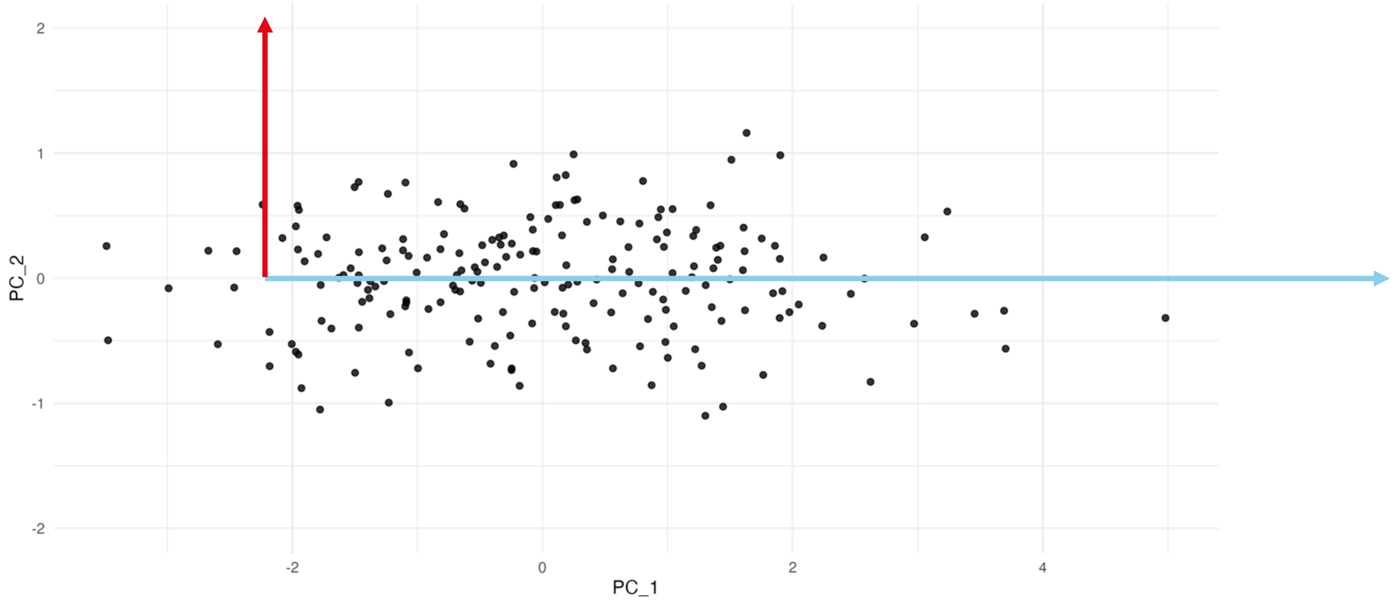
Principal Component Analysis

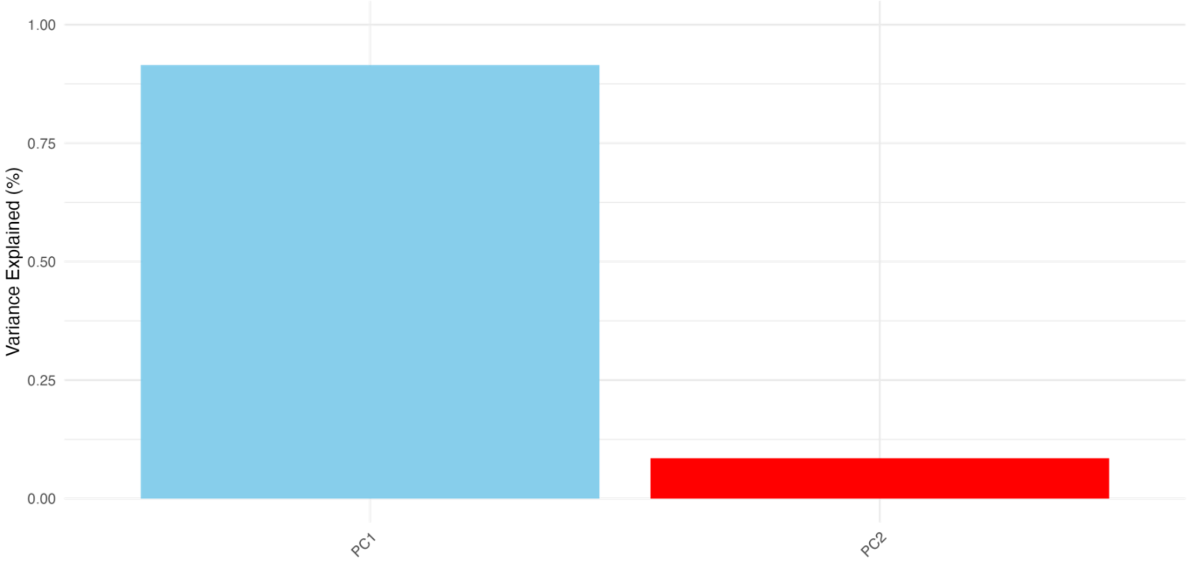
New axis that are linear combination of the original axes



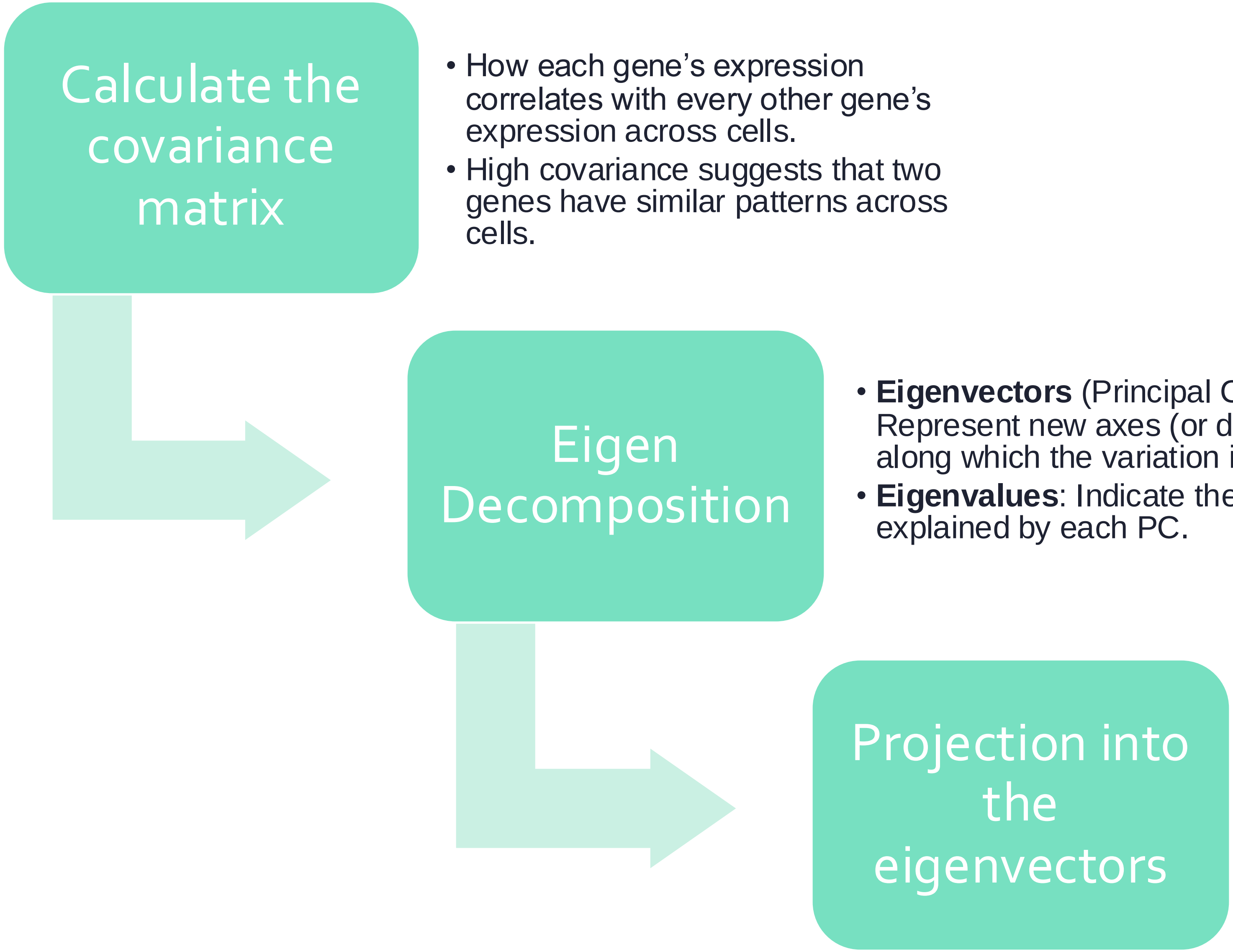
Principal Component Analysis

New axis that are linear combination of the original axes





Mathematically



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.
- **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.

Choosing the number of PCs

A **LINEAR** dimensionality reduction technique and the **TOP** principal components contain higher variance from the data.

PCA finds **dominant sources** of variation in high-dimensional datasets, inferring genes that distinguish between samples. Maximizing the variability captured in specific factors, as opposed to spreading relatively evenly among factors, may mix different biological signal in a single component.

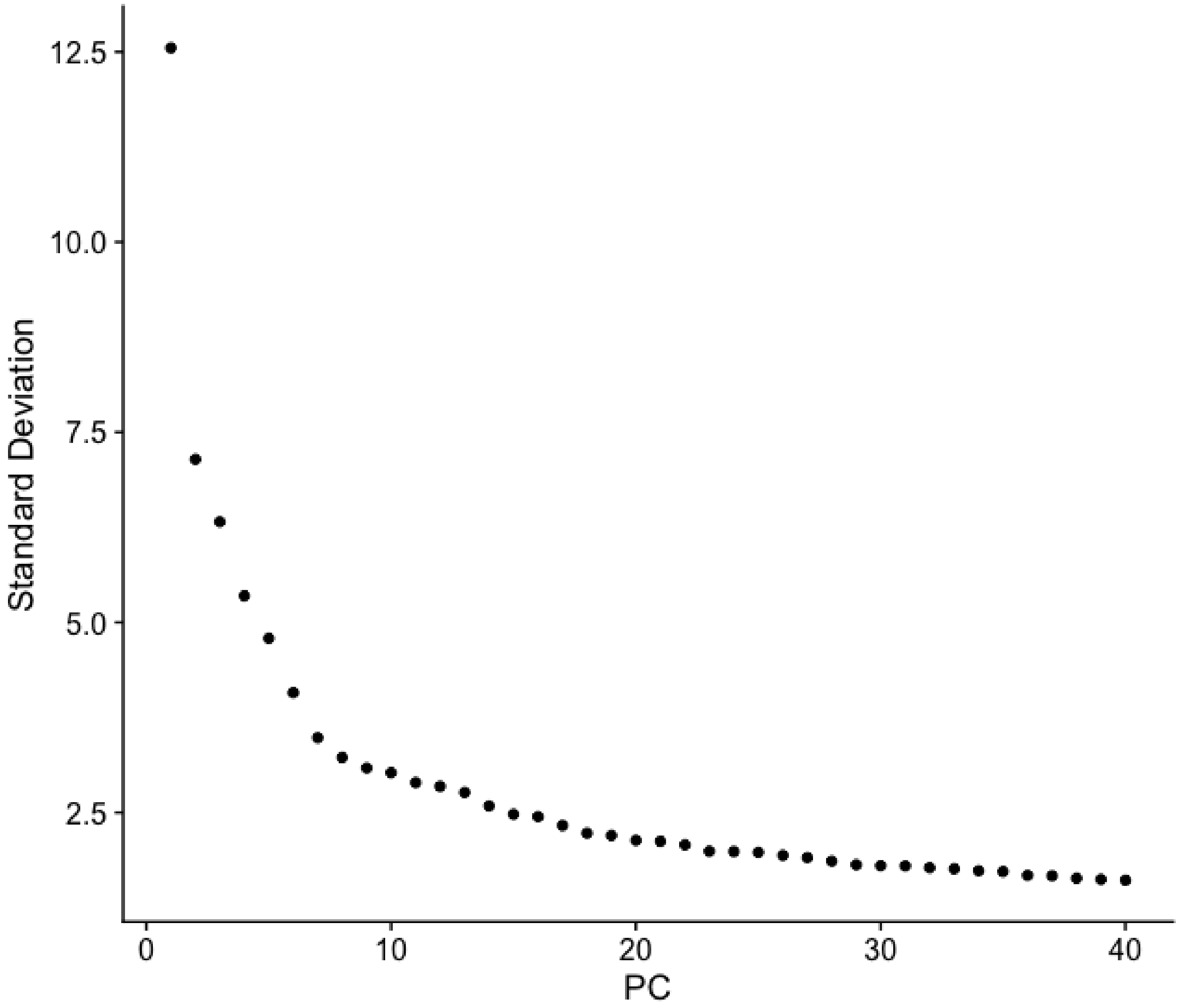
We could select:

- PCs that explain at least 1% of variance
- Jackstraw of significant p-values
- The first 5-10 PCs

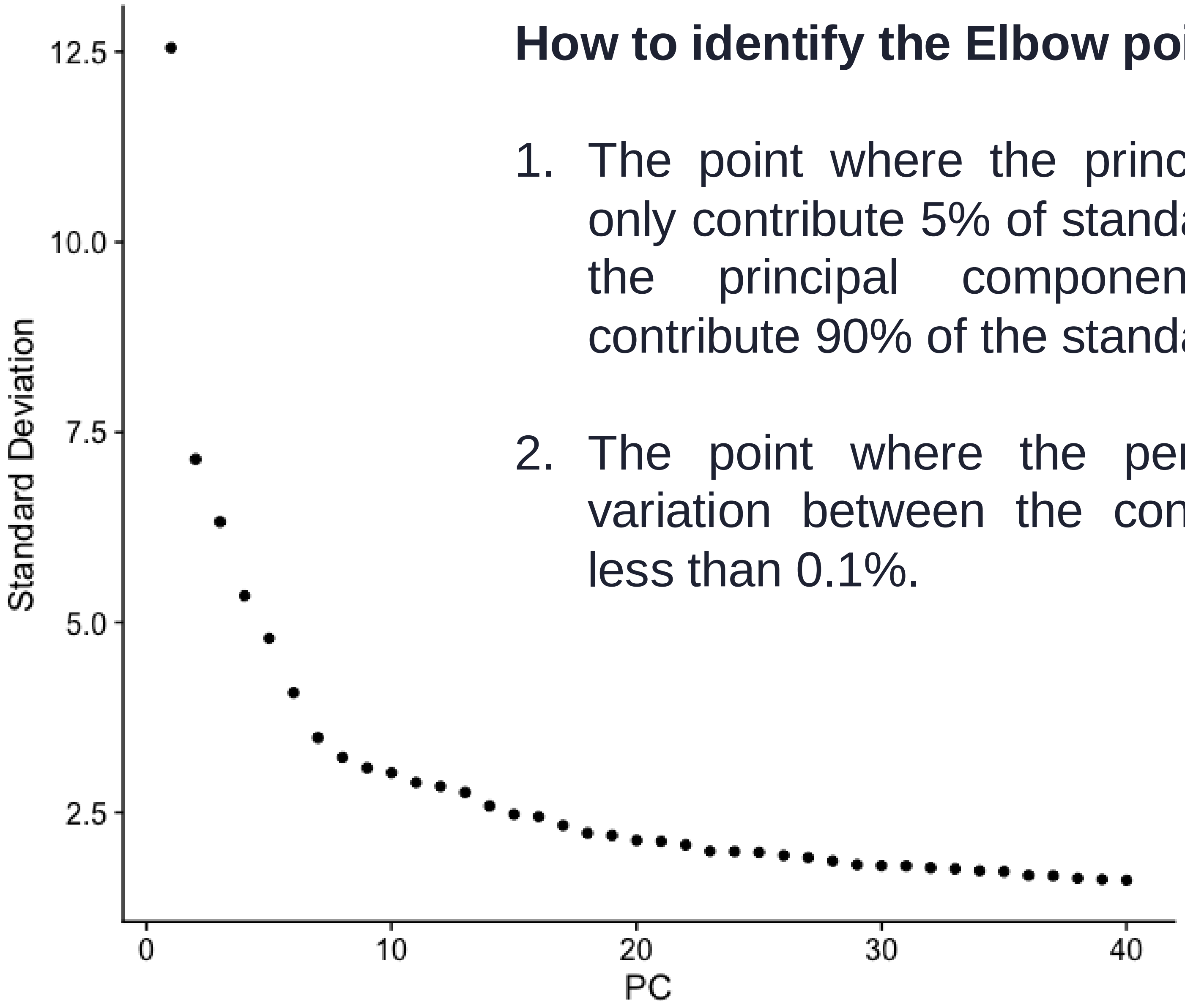
Issue:

- Cell sizes and sequencing depth are usually captured in the top principal components

In real-life



The Elbow-point

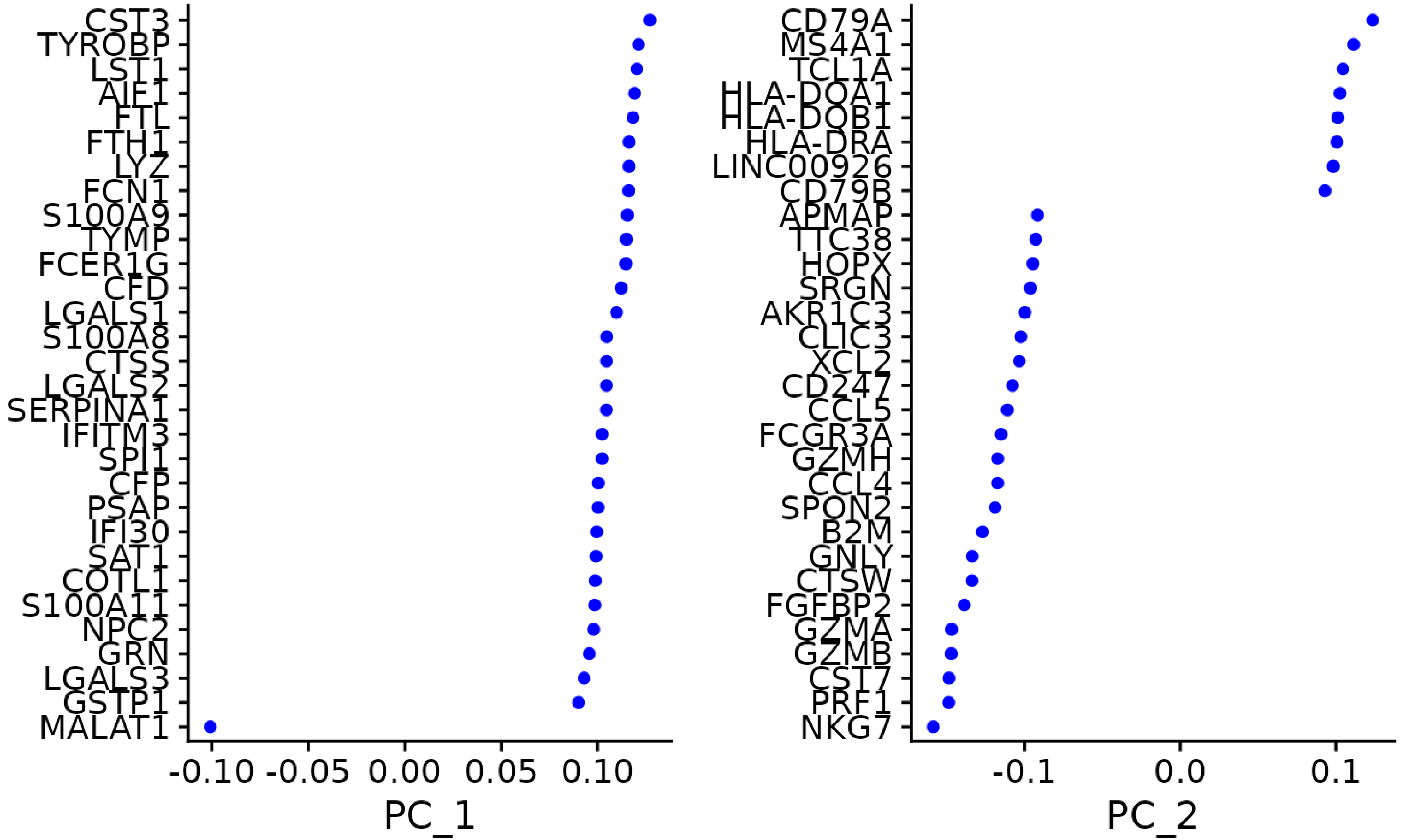


How to identify the Elbow point:

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%.

Take a look at PCs



Non-linear Methods for Dimensionality Reduction

t-SNE

t-distributed Stochastic Neighbourhood Embedding

Authors: Laurens van der Maaten, Geoffrey Everest Hinton
http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

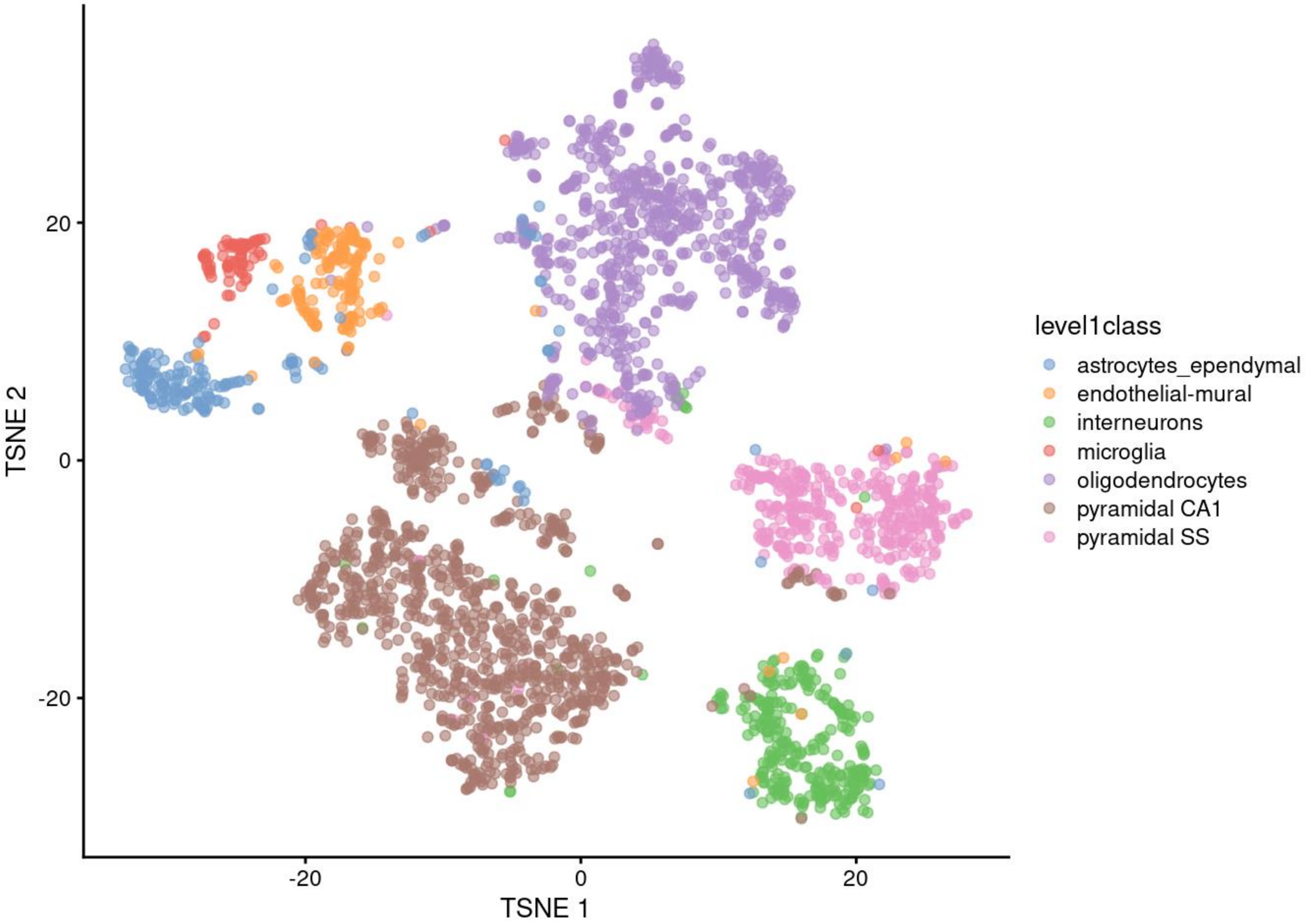
Non-linear dimensionality reduction approach. It reduces the high-dimensional data into a two- or three-dimensional space in such a way that:

- Similar objects/samples/cells are modelled by nearby points
- Dissimilar objects/samples/cells are modelled by distant points

Minimize the divergence between:

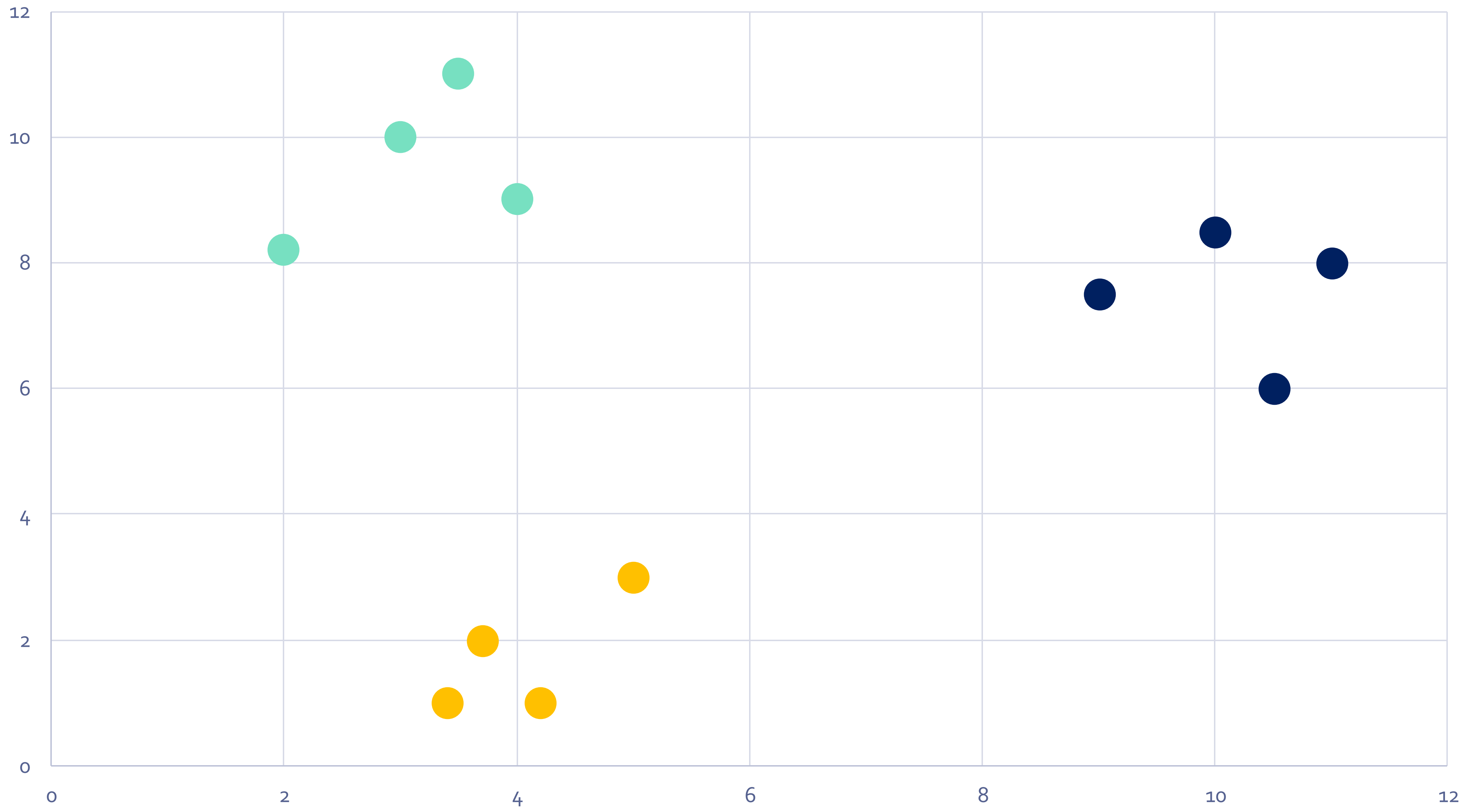
- Distribution of the **pairwise similarities of the input** objects/samples/cells
- Distribution of the **pairwise similarities of corresponding low-dimensional points** in embedding

t-SNE

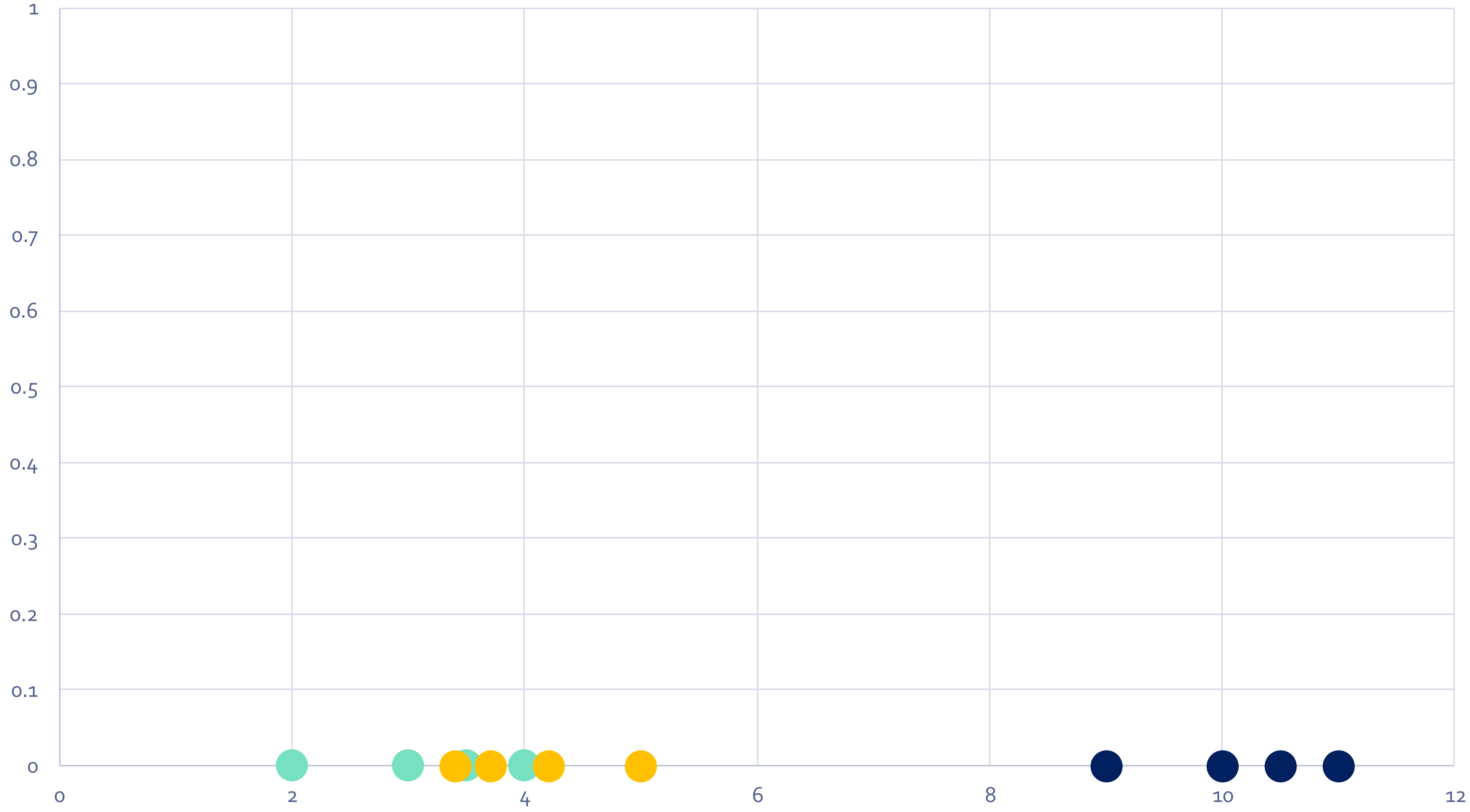


level1class
● astrocytes_ependymal
● endothelial-mural
● interneurons
● microglia
● oligodendrocytes
● pyramidal CA1
● pyramidal SS
TSNE 2
20
0
-20
TSNE 1
-20
0
20

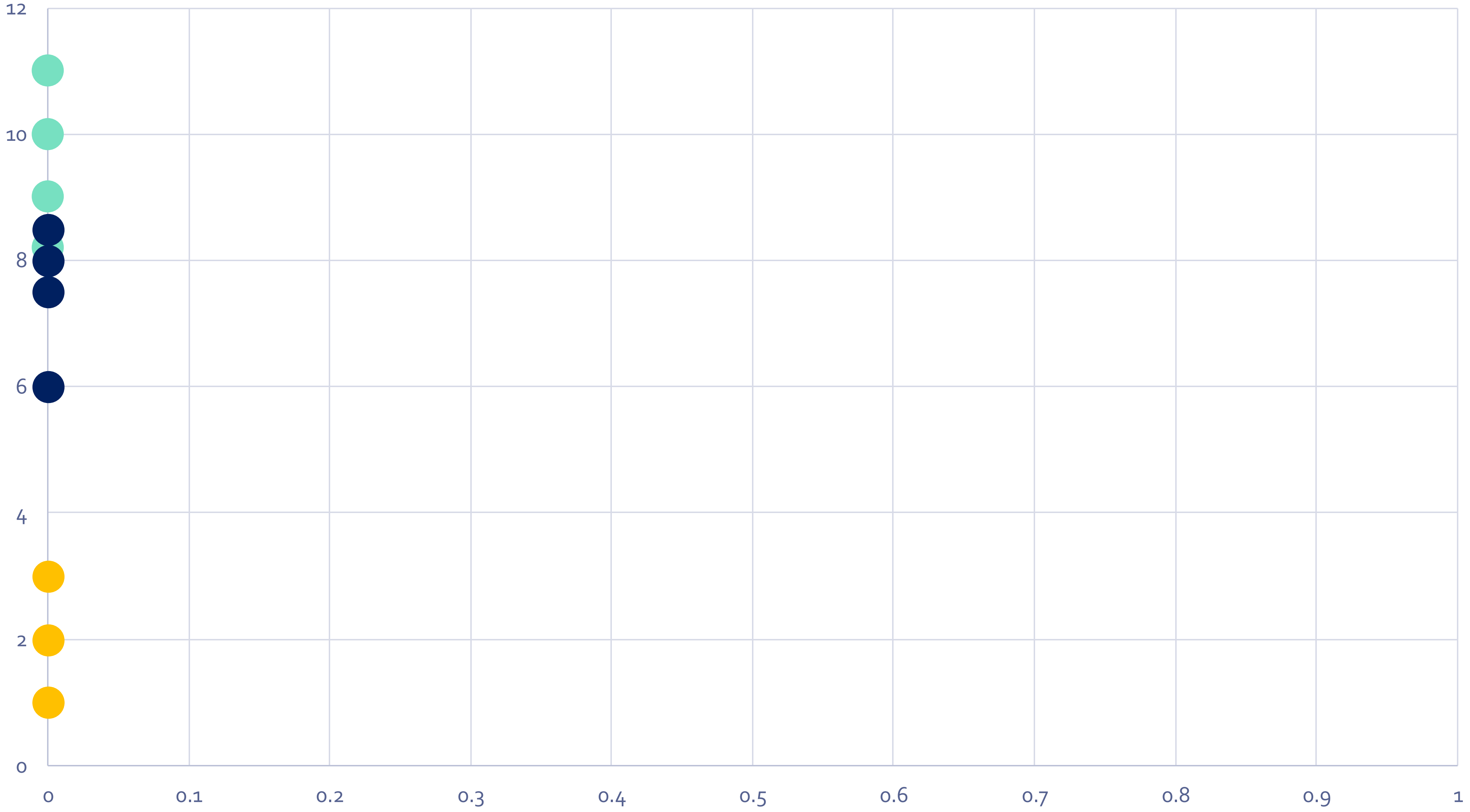
t-SNE - Example



Projection on x-axes



Projection on y-axes



t-SNE

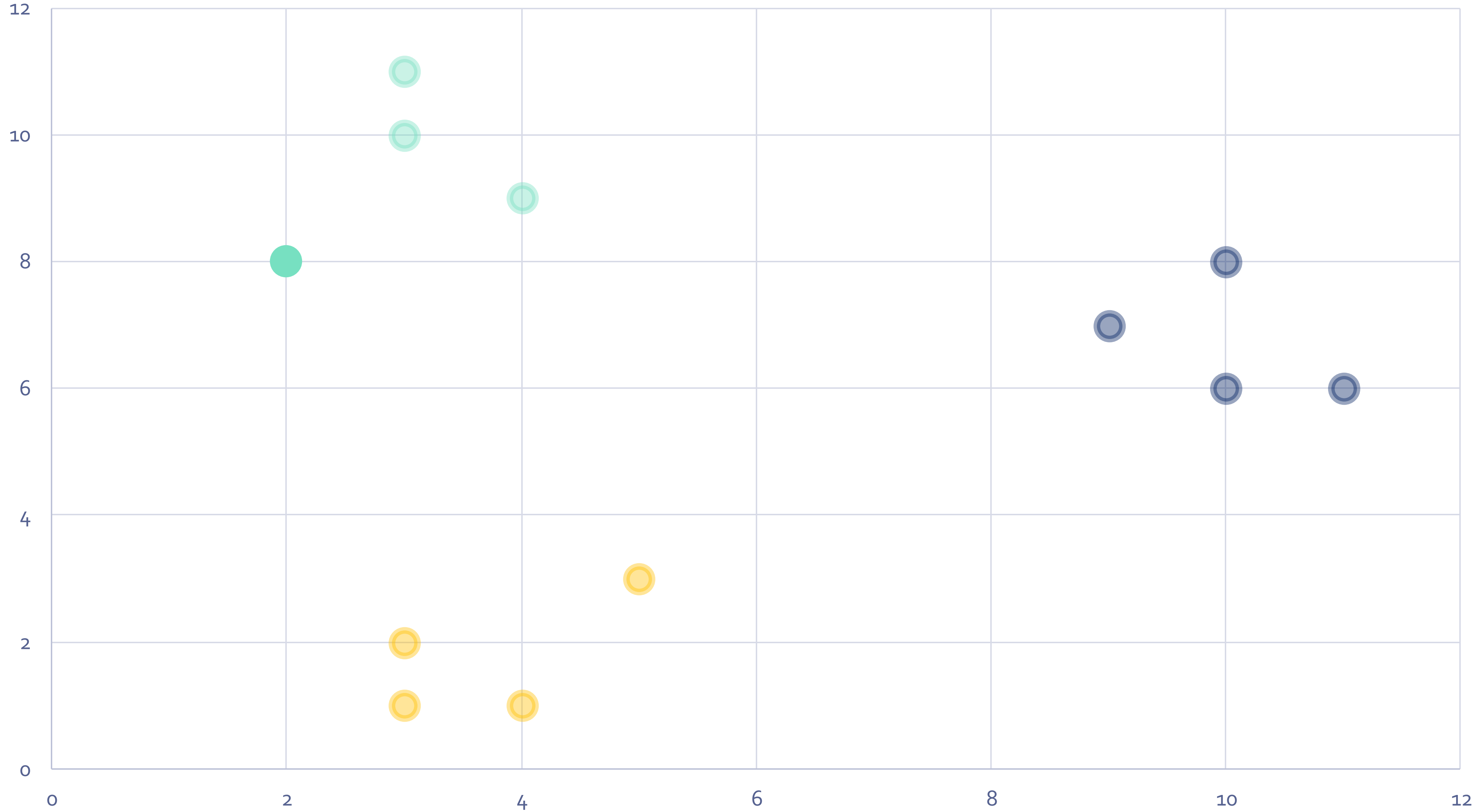
Minimize the divergence between:

- Distribution of the **pairwise similarities of the input** objects/samples/cells
- Distribution of the **pairwise similarities of corresponding low-dimensional points** in embedding

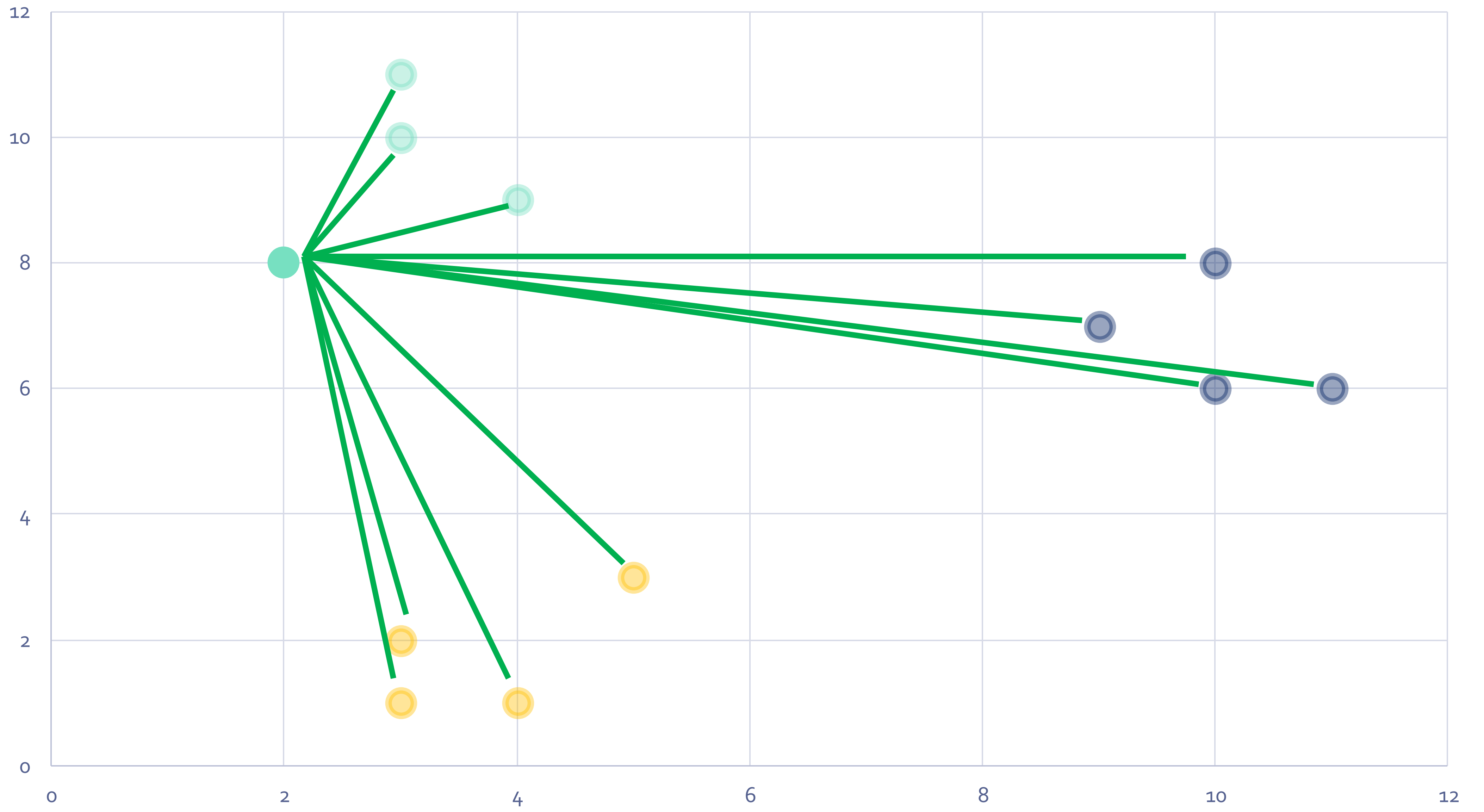
Three stages:

- Calculating a joint probability distribution that represents the similarities between the data points
- Creating a dataset of points in the target dimension and then calculating the joint probability distribution for them as well
- Using gradient descent to change the dataset in the low-dimensional space so that the joint probability distribution representing it would be as similar as possible to the one in the high dimension

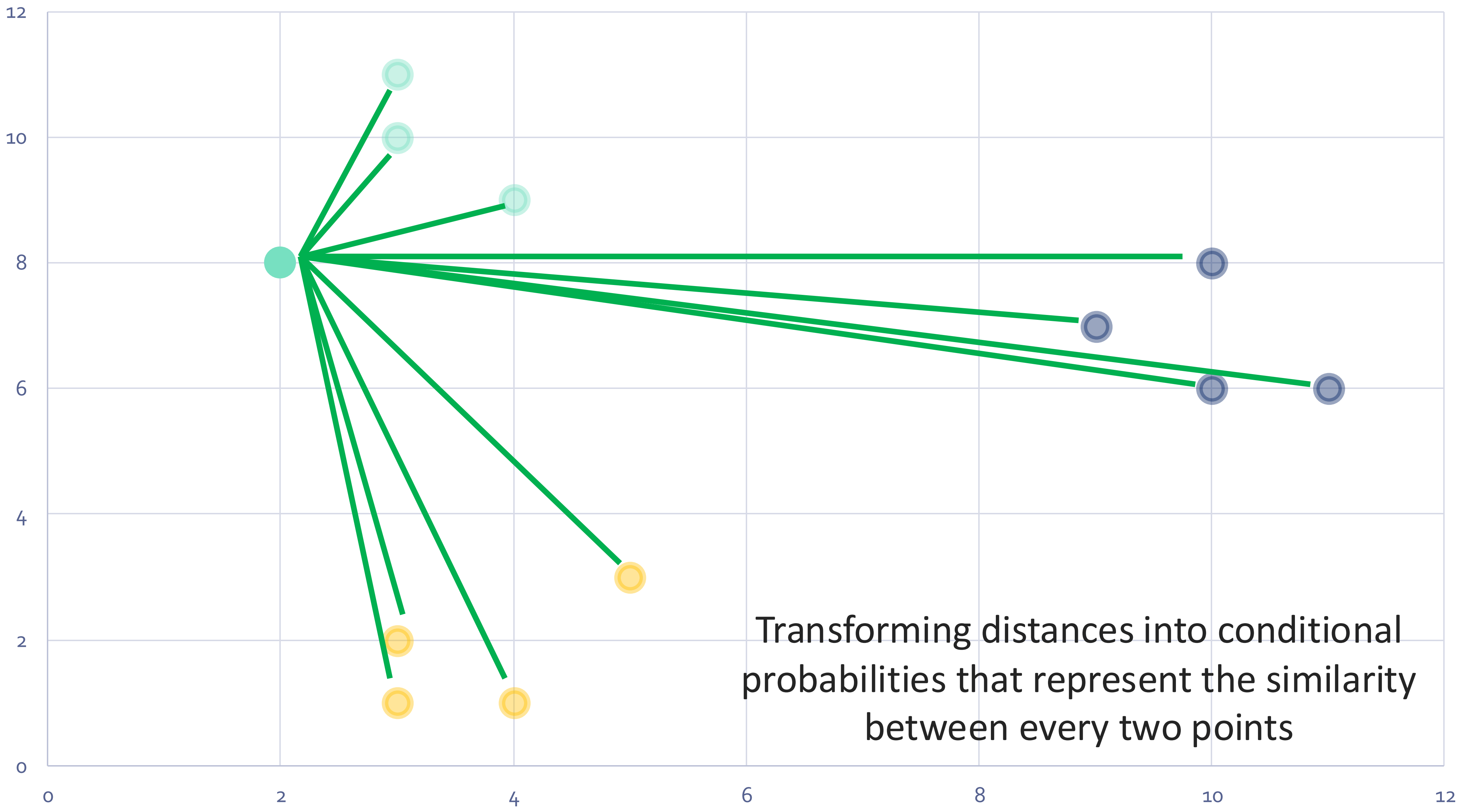
First-Step



Computing distances



From distance to probability

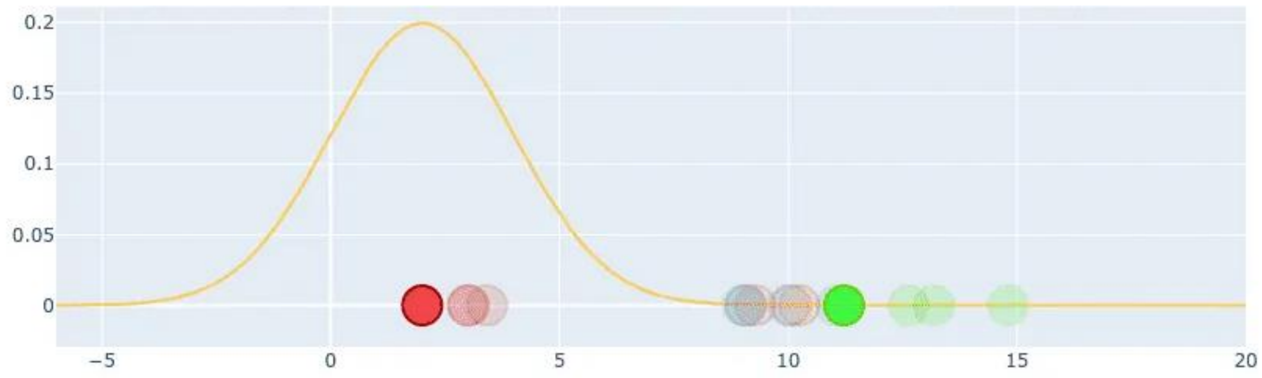


Transforming distances into conditional probabilities that represent the similarity between every two points

Coditional Probability

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

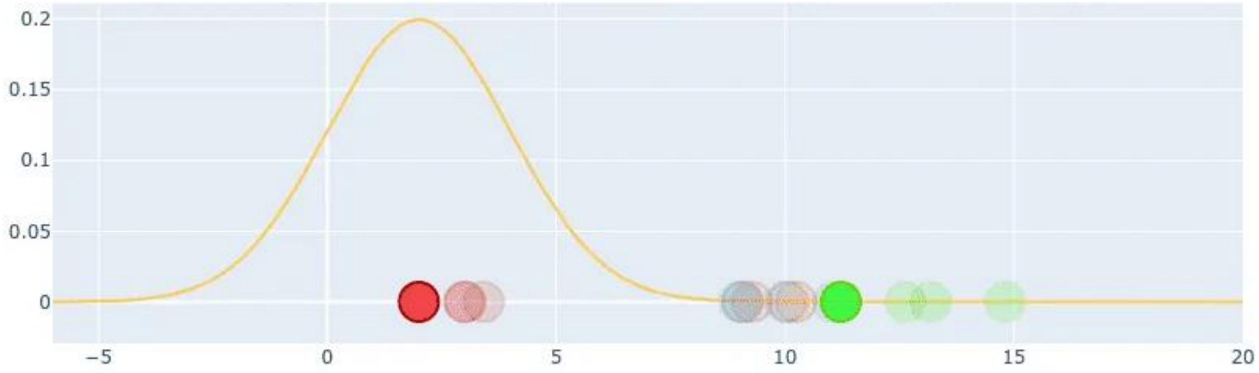
$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$



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 cantered at x_i with a standard deviation of σ_i

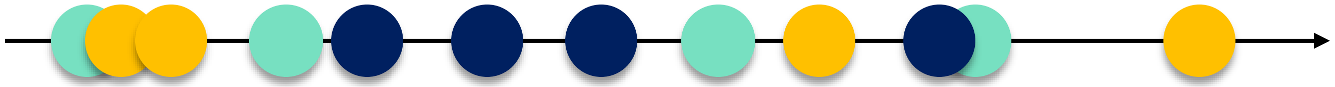
$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$



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

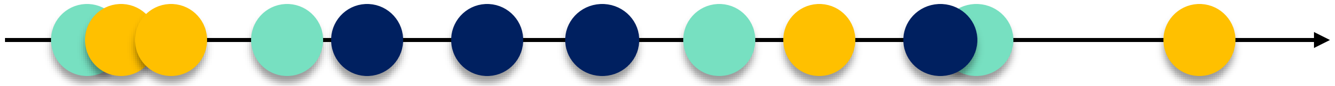


For this set of points, we will create their joint probability distribution but this time we will be using the t-distribution and not the Gaussian

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.

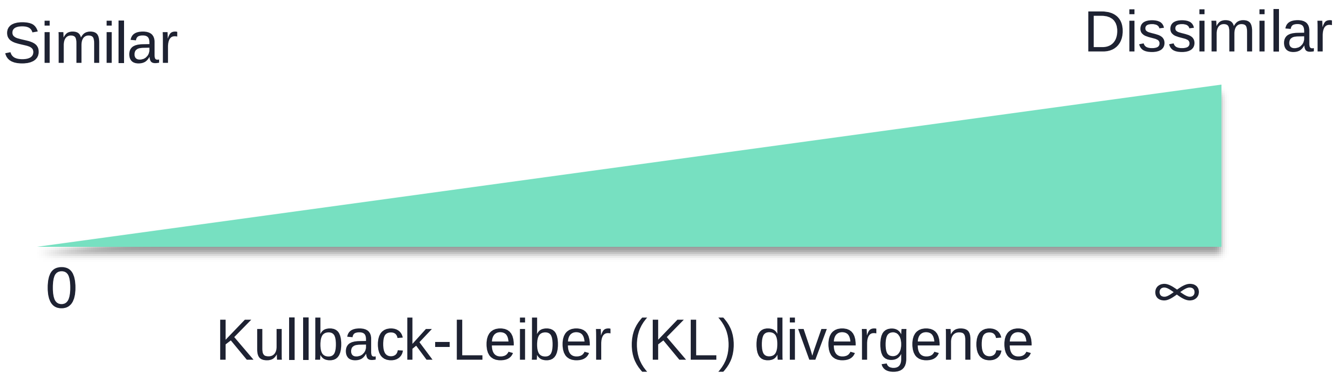
Creating data in a low dimension

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.



Similar

Dissimilar

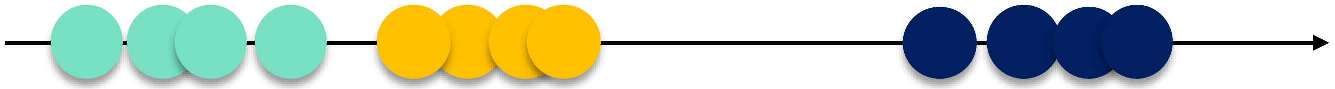
0

∞

Kullback-Leiber (KL) divergence

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:

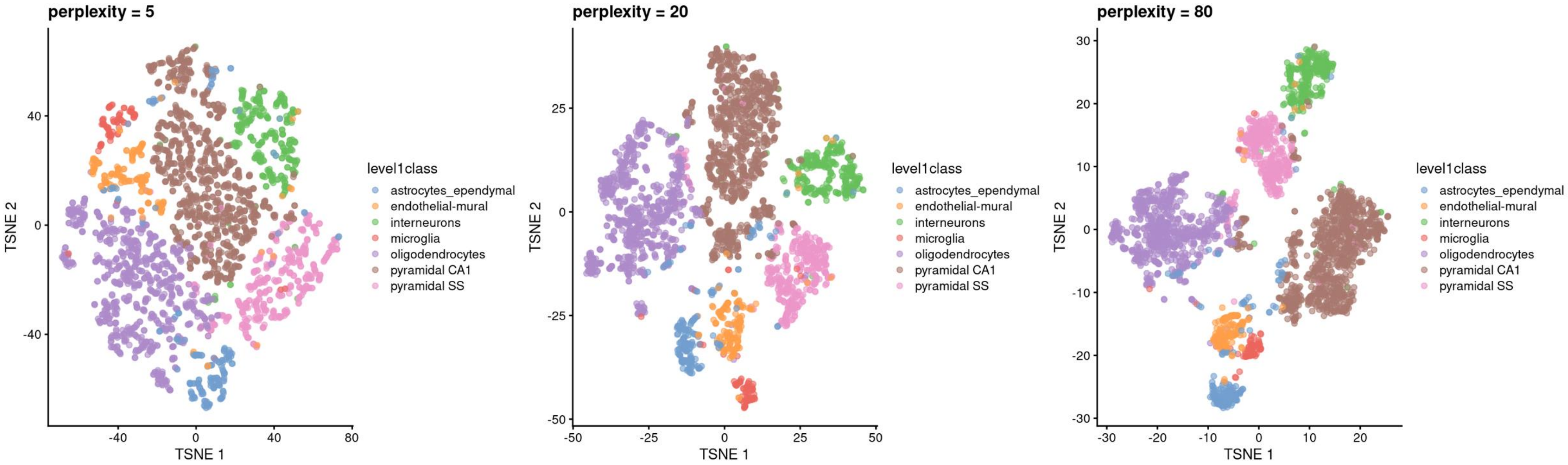
Gradient descent:

- 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

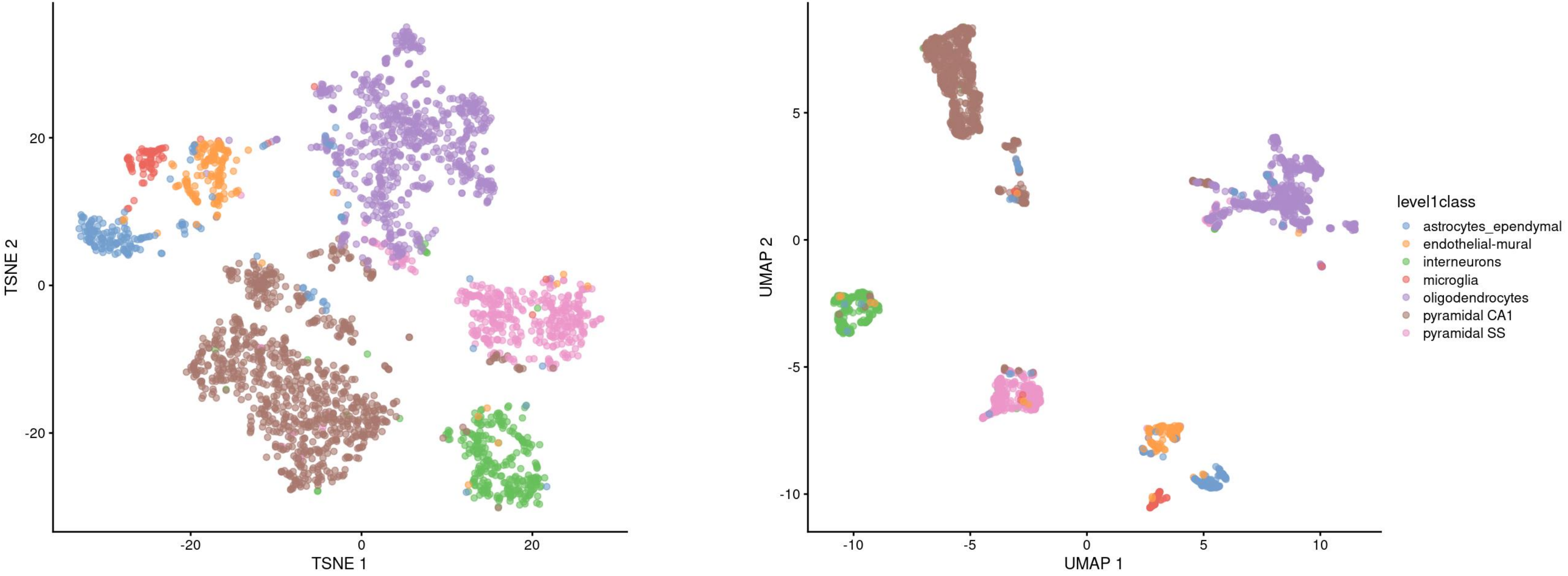
Authors: McInnes L. and Healy J.

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:

- increased speed
- better preservation of the data's global structure
- It can use any distance metrics
- 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)

T-SNE vs UMAP

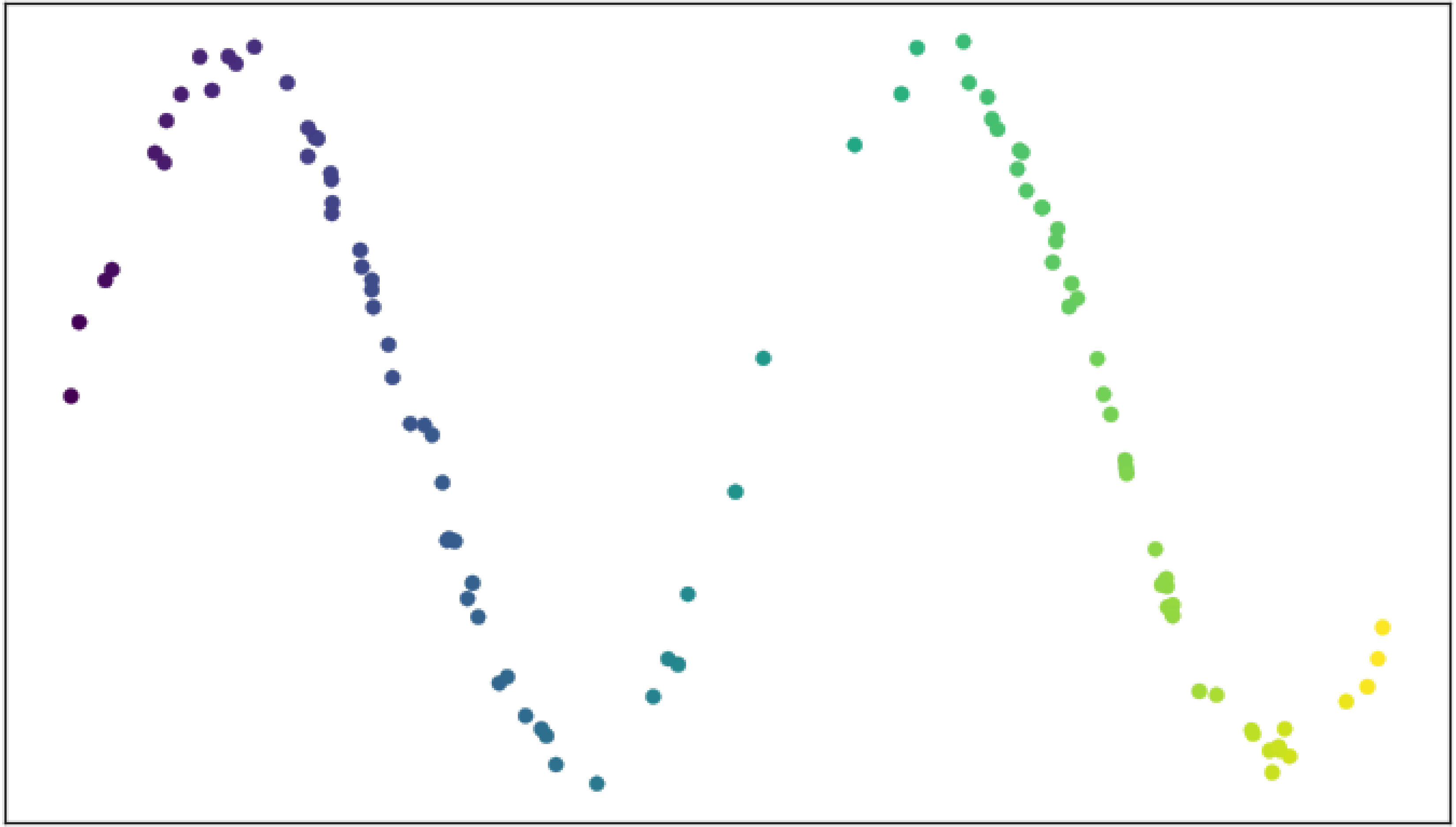

TSNE 2
20
0
-20
-20
0
20
TSNE 1
UMAP 2
5
0
-5
-10
-10
-5
0
5
10
UMAP 1
level1class
astrocytes_ependymal
endothelial-mural
interneurons
microglia
oligodendrocytes
pyramidal CA1
pyramidal SS

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.

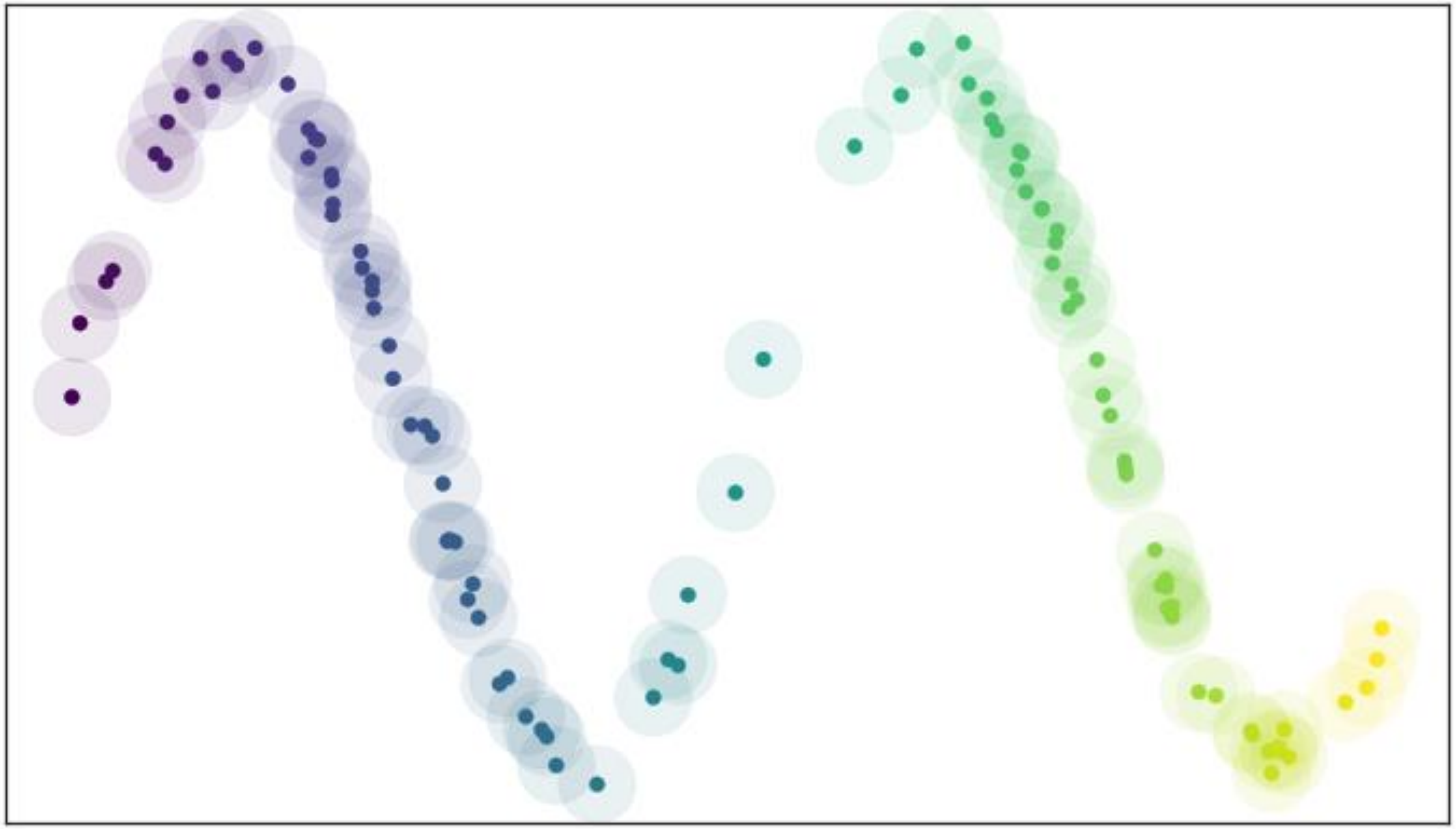
UMAP Theory

Step 1: UMAP extends a radius outwards from each point



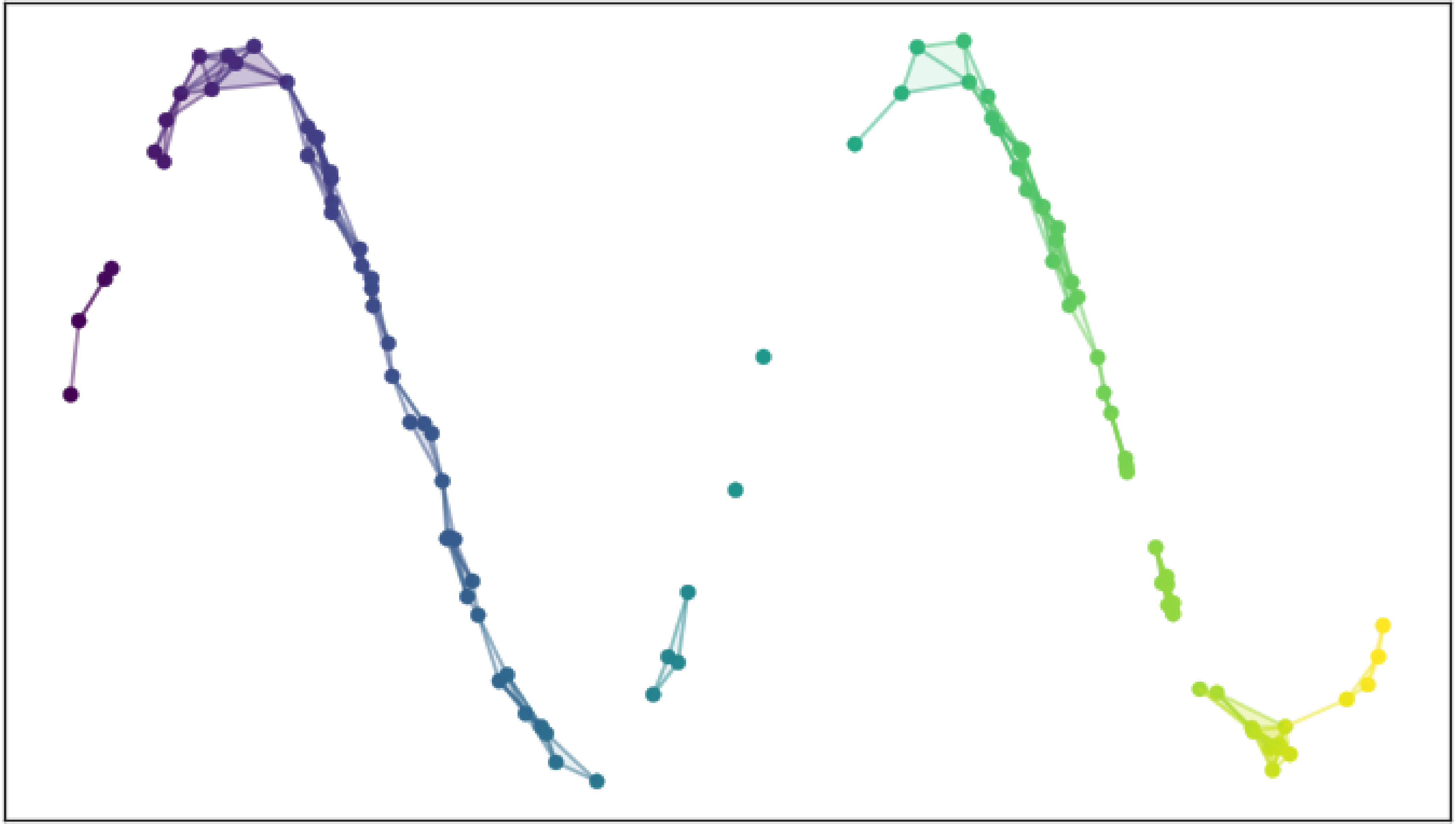
UMAP Theory

Step 1: UMAP extends a radius outwards from each point



UMAP Theory

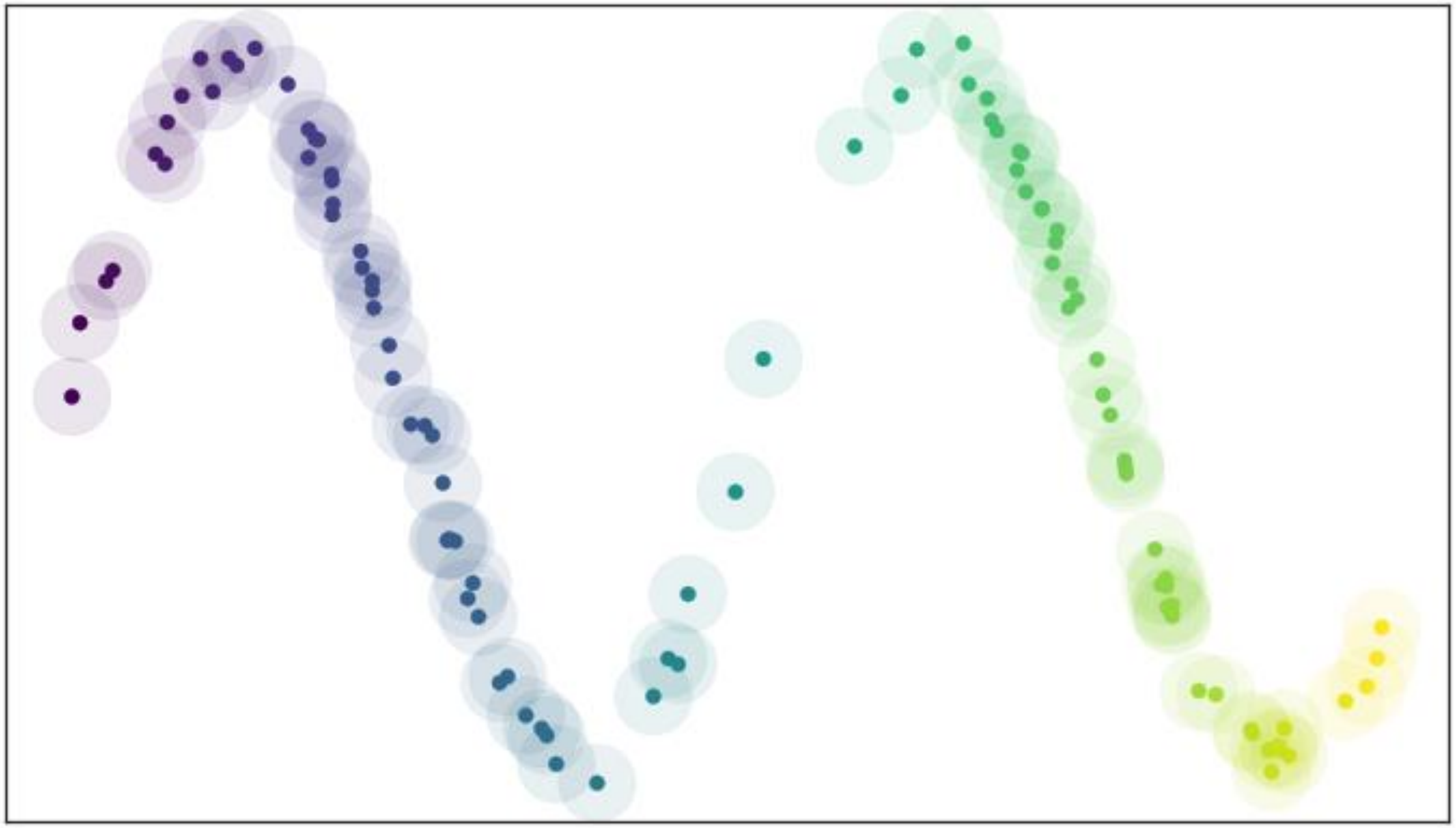
Step 1: UMAP extends a radius outwards from each point, connecting points when those radii overlap



UMAP Theory

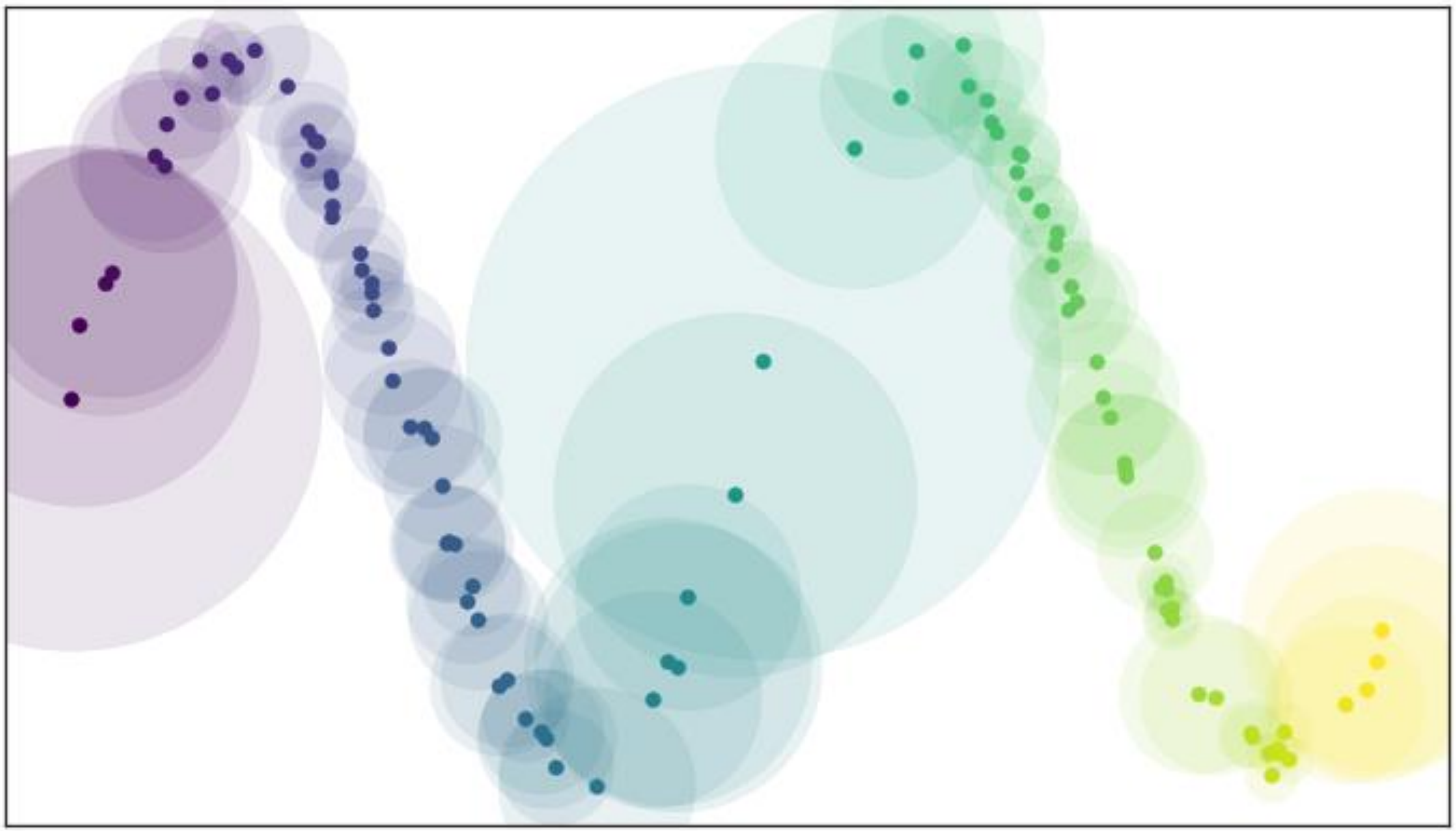
Choosing this radius is critical:

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



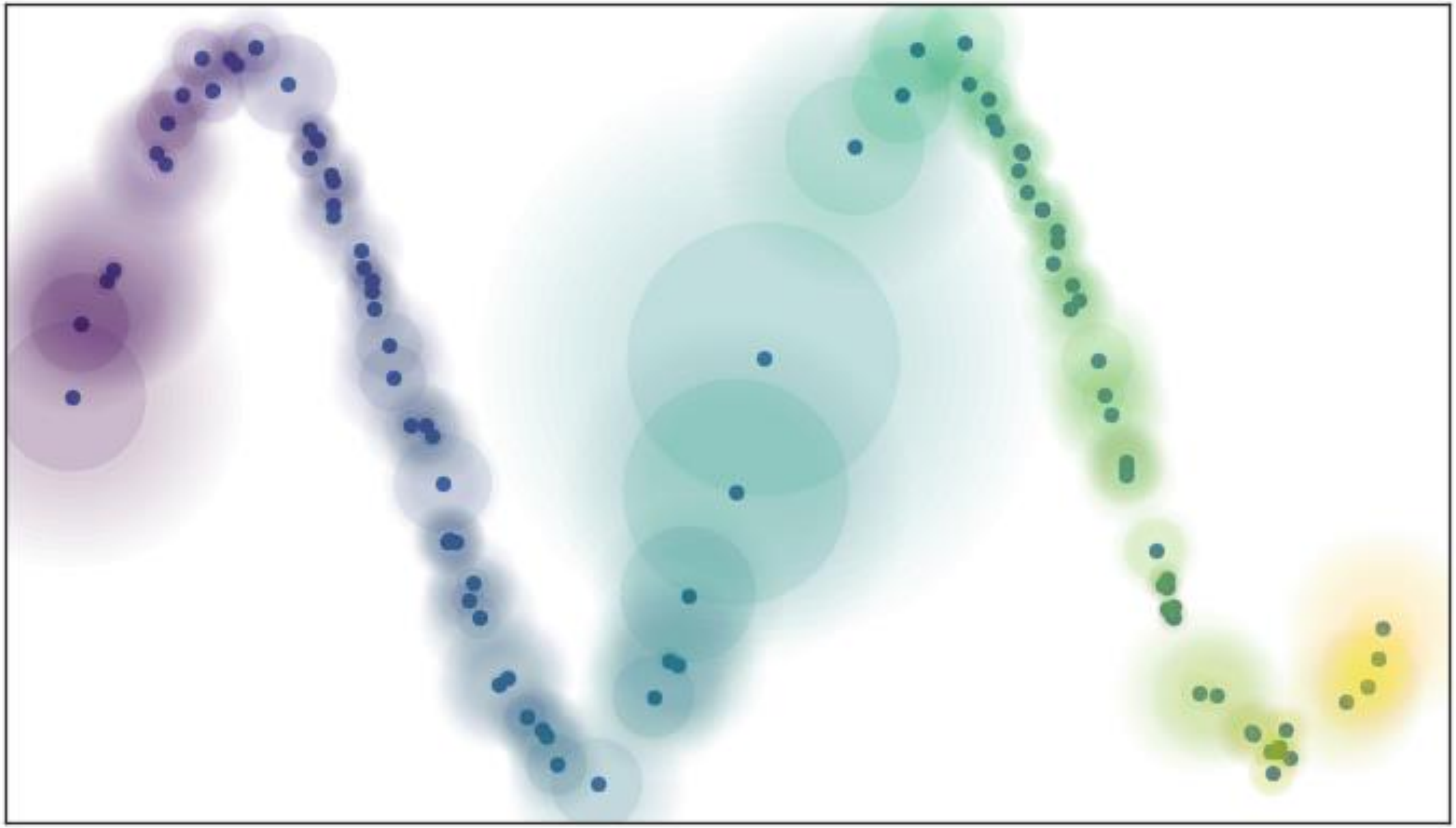
UMAP Theory

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



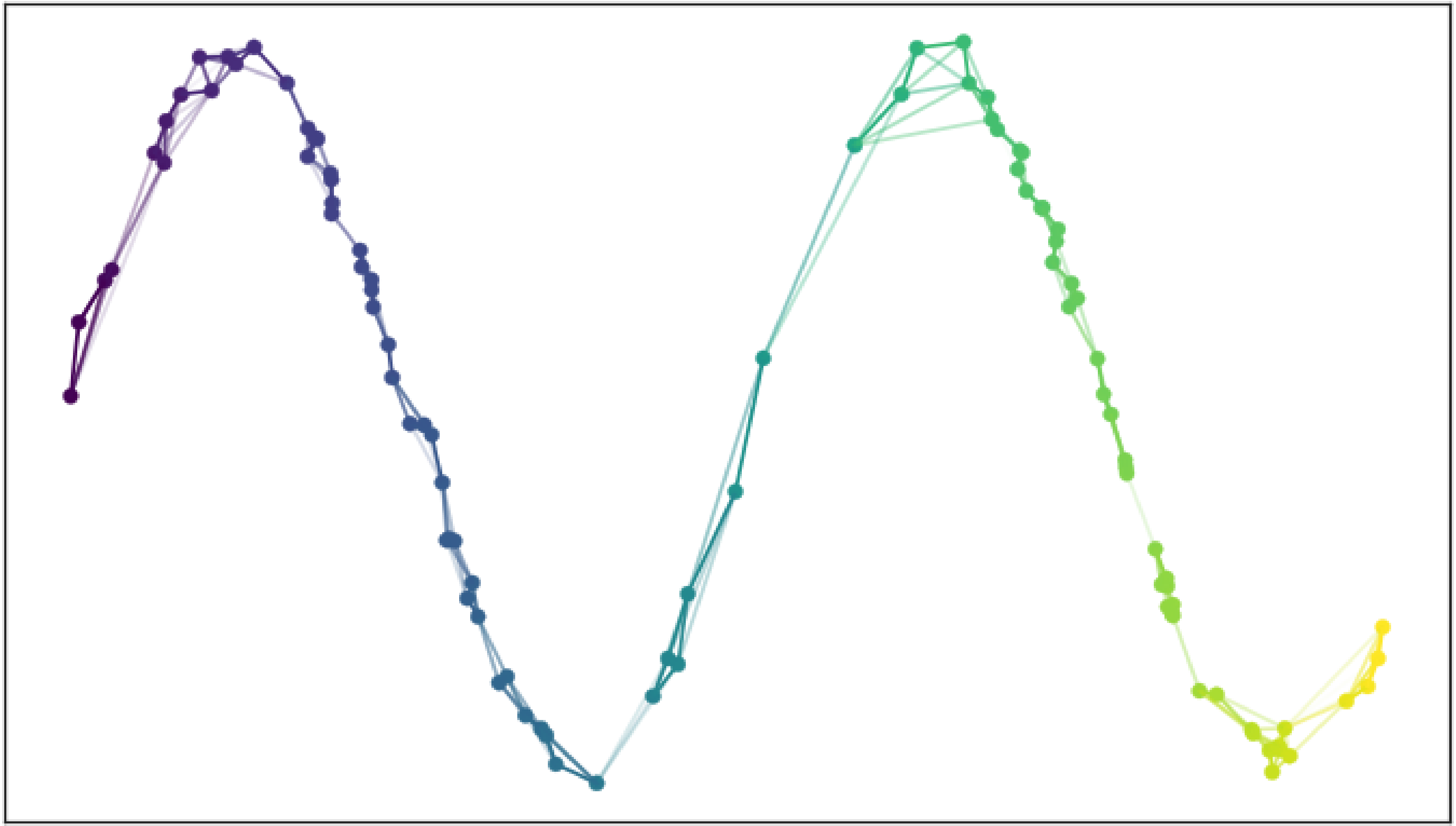
UMAP Theory

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



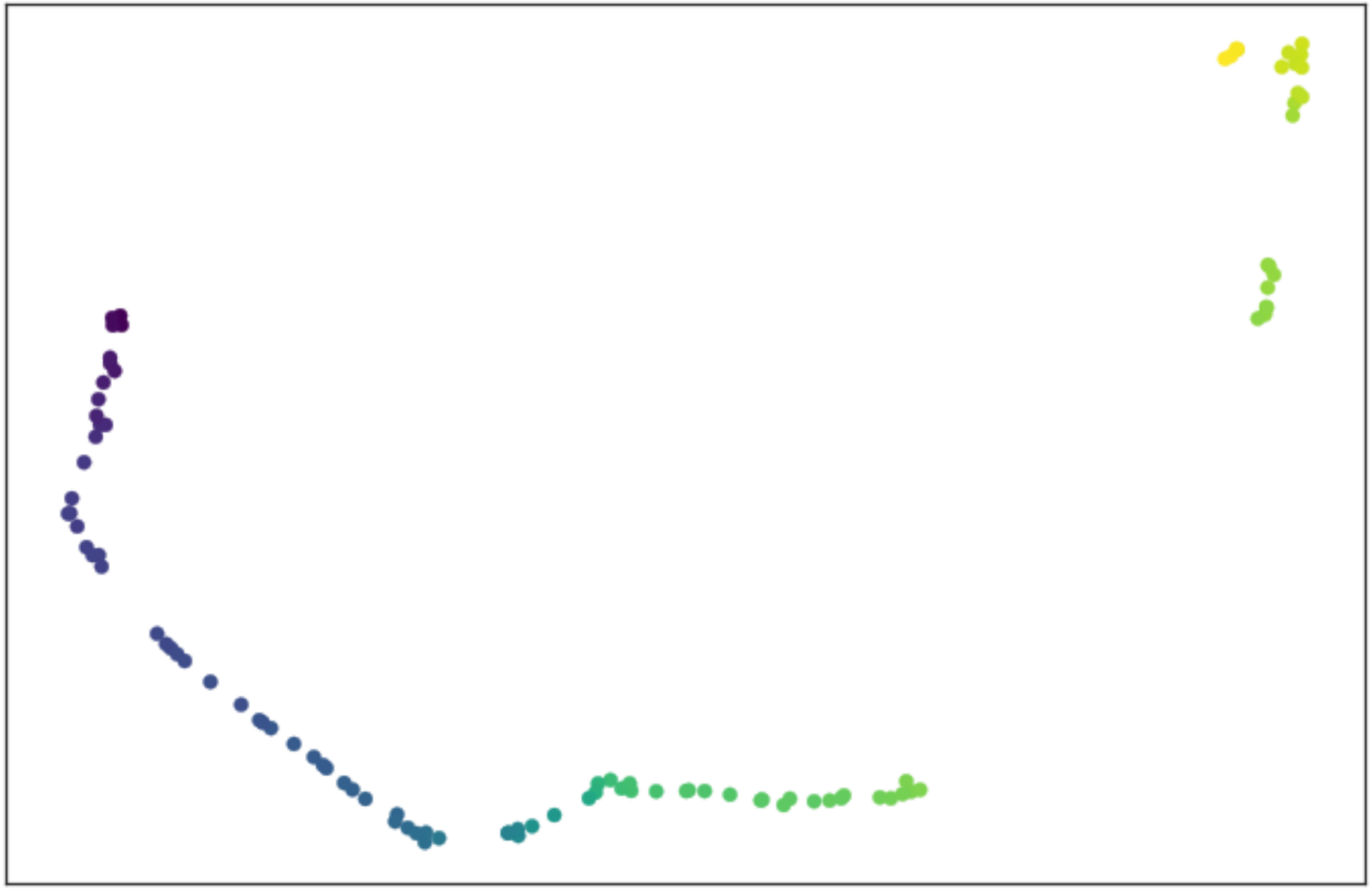
UMAP Theory

All points must be connected to at least its closest neighboring 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 clusters.

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. 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

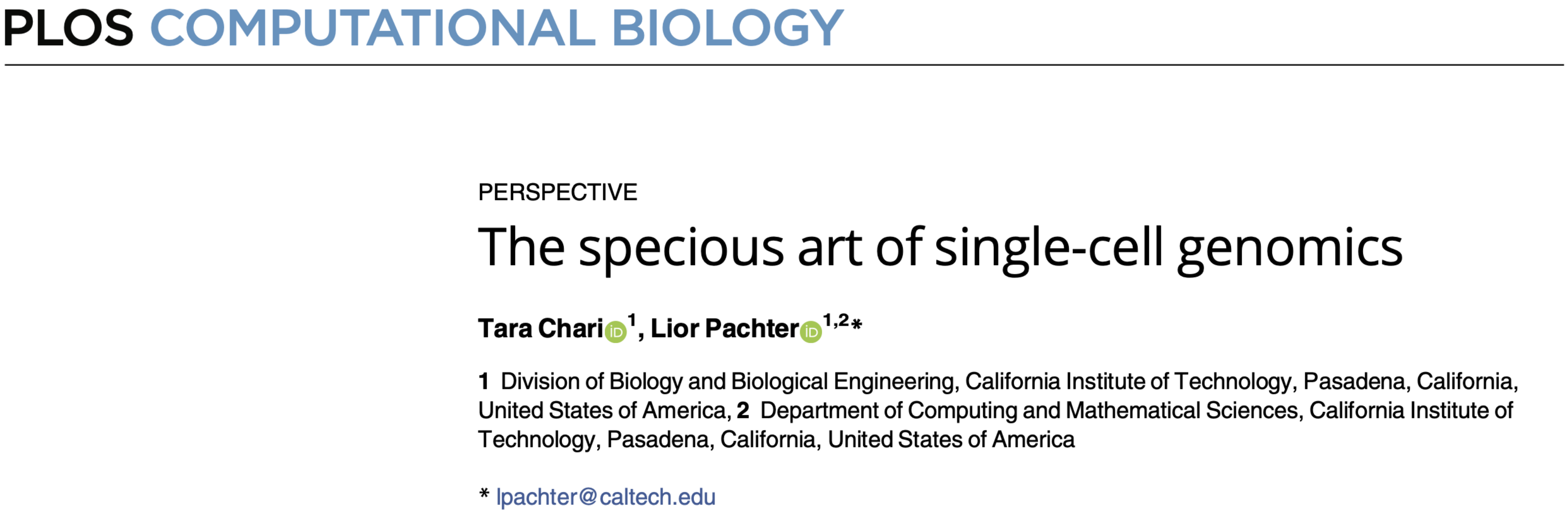
4. Spurious clustering can be observed
Due to Random noise that doesn’t always look random (e.g. low values of n_neighbors)

5. UMAP is stochastic
Different runs with the same hyperparameters can yield different results

Consideration

- t-SNE and UMAP are both for data visualization.
- t-SNE and UMAP are both non-linear, graph-based methods for dimensionality reduction in scRNA-seq analysis.
- t-SNE moves the high dimensional graph to a lower dimensional space points by points. UMAP compresses that graph.
- Key parameters for t-SNE and UMAP are the perplexity and number of neighbors, respectively.
- 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.

Skepticism about this methods



PLOS COMPUTATIONAL BIOLOGY

PERSPECTIVE

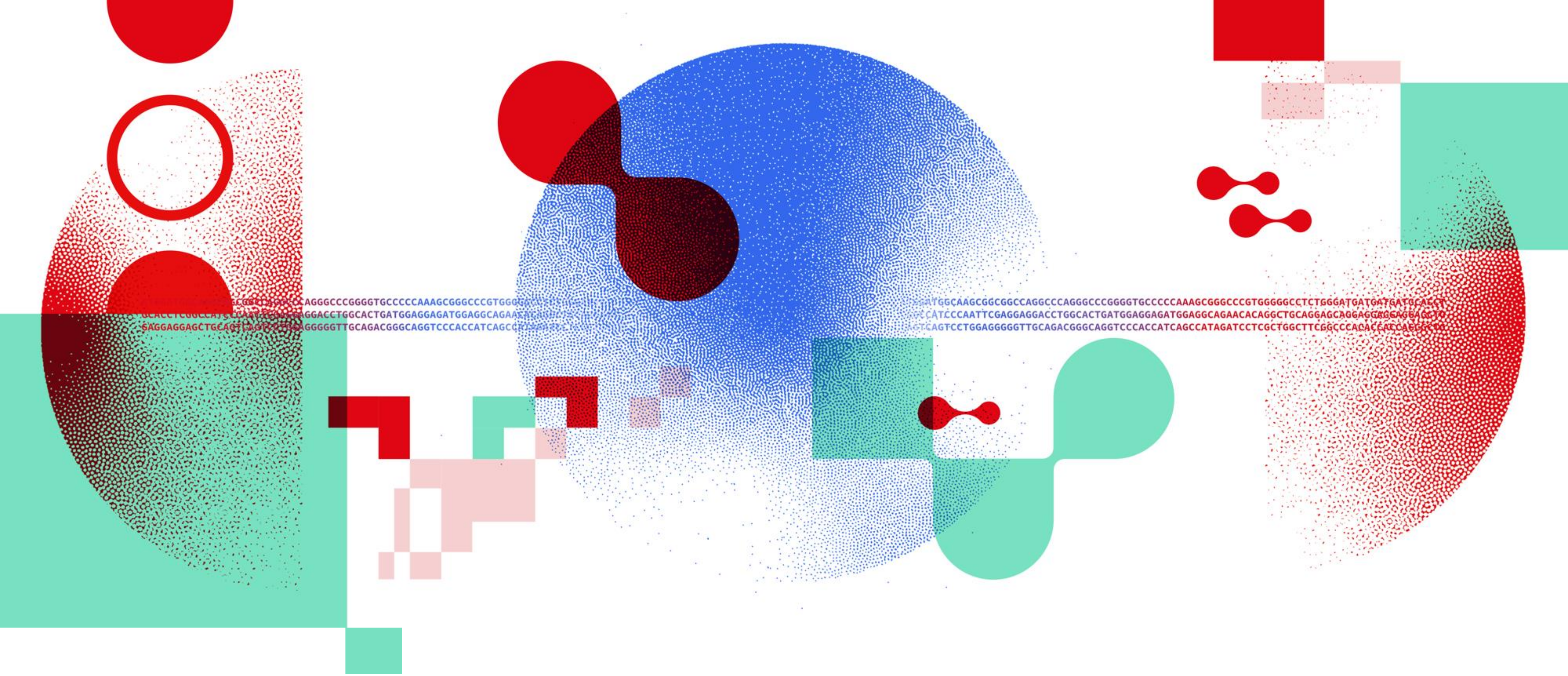
The specious art of single-cell genomics

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

Chari, T., Banerjee, J. & Pachter, L. The specious art of single-cell genomics. Plos Comp Biology (2023)



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss