Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Dimensionality reduction

Deepak Tanwar

March 18-20, 2026

Adapted from previous year courses

elixir SWITZERLAND

Learning objectives

Understand the Curse of Dimensionality

Identify and Apply Dimensionality Reduction techniques

Understand PCA and UMAP

Determine appropriate scenarios for using UMAP in data analysis.

scRNA-seq data = thousands of genes

× thousands of cells

high noise, redundancy, and sparsity

scRNA-seq data = thousands of genes
× thousands of cells

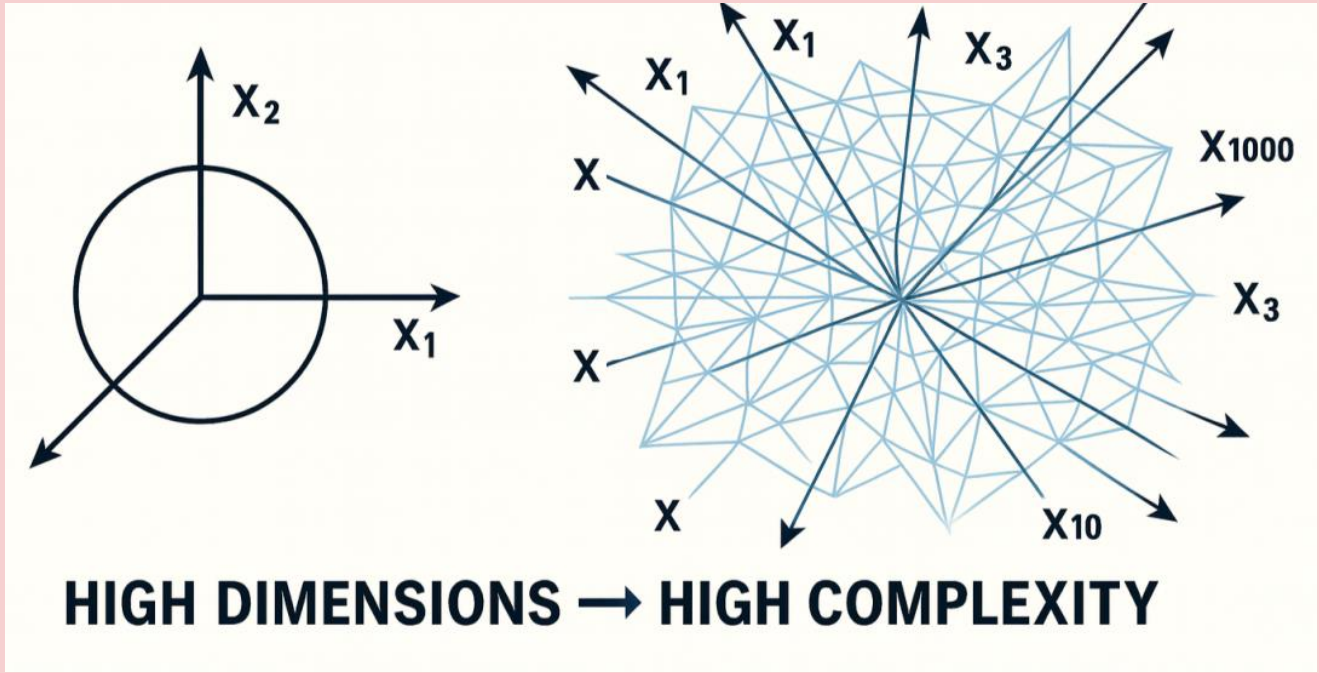
high noise, redundancy, and sparsity

Noise → wrong values (e.g., false zeros)
Redundancy → repeated information
Sparsity → too many zeros

scRNA-seq data = thousands of genes

× thousands of cells

high noise, redundancy, and sparsity



https://www.linkedin.com/pulse/curse-dimensionality-why-more-features-arent-always-better-ahamed-gpbie/

Curse of dimensionality: In theory high-dimensional data contains more information, but in practice this is not the case. Higher dimensional data often contains more noise and redundancy and therefore adding more information does not provide benefits for downstream analysis steps.

Challenge	Solution
High-dimensional gene space is noisy, redundant, and sparse	Select informative features (e.g., highly variable genes)
Only a small subset of genes defines cell identity	Highlight key features that distinguish cell types
Many algorithms struggle in high dimensions	Reduce dimensionality to a manageable feature space
High-dimensional data is hard to visualize	Project data into 2D/3D (UMAP, t-SNE)

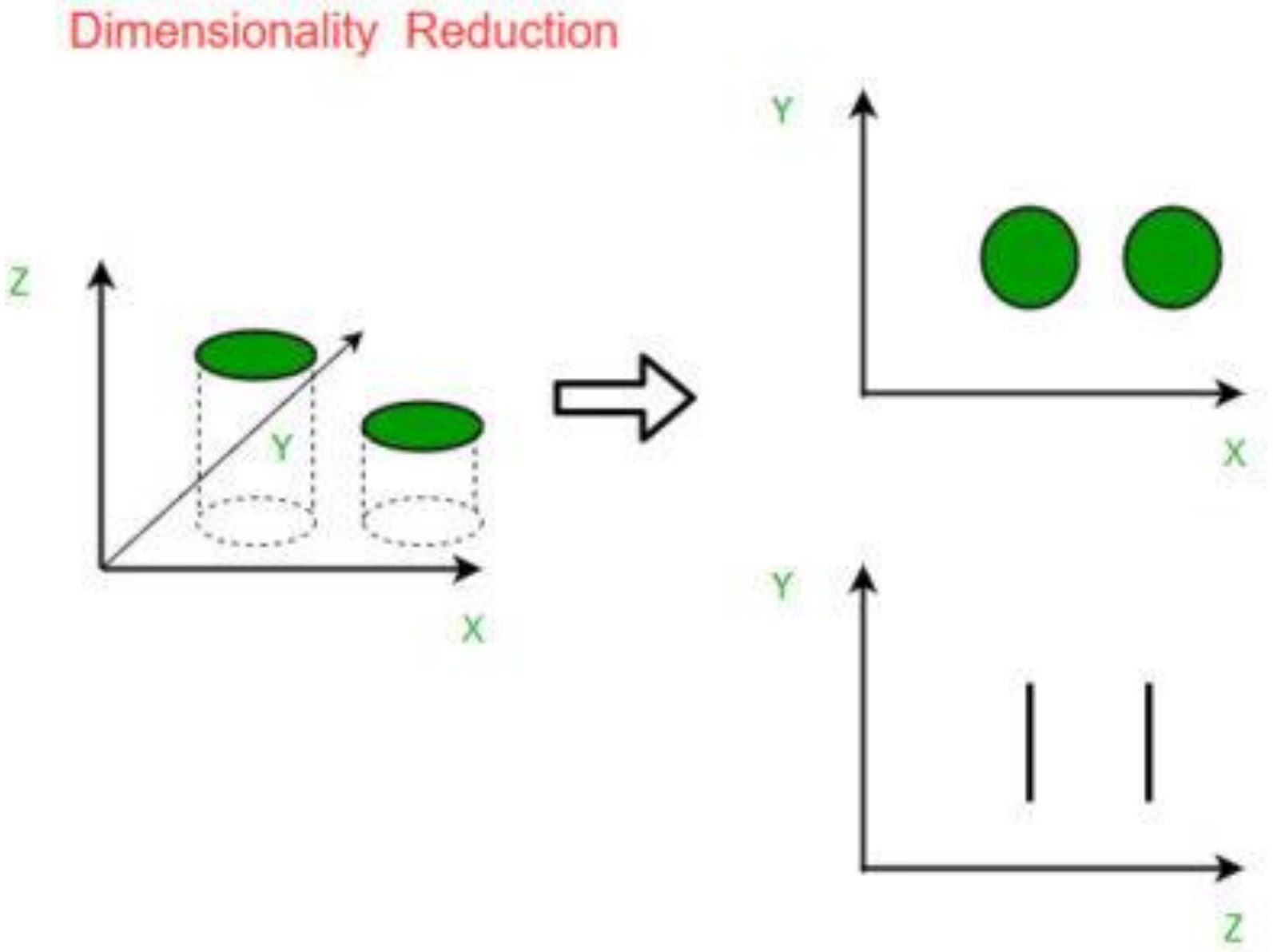
Need for dimensionality reduction in scRNA-seq data

scRNA-seq data is composed of 1000s of genes. Dimensionality reduction would be helpful in:

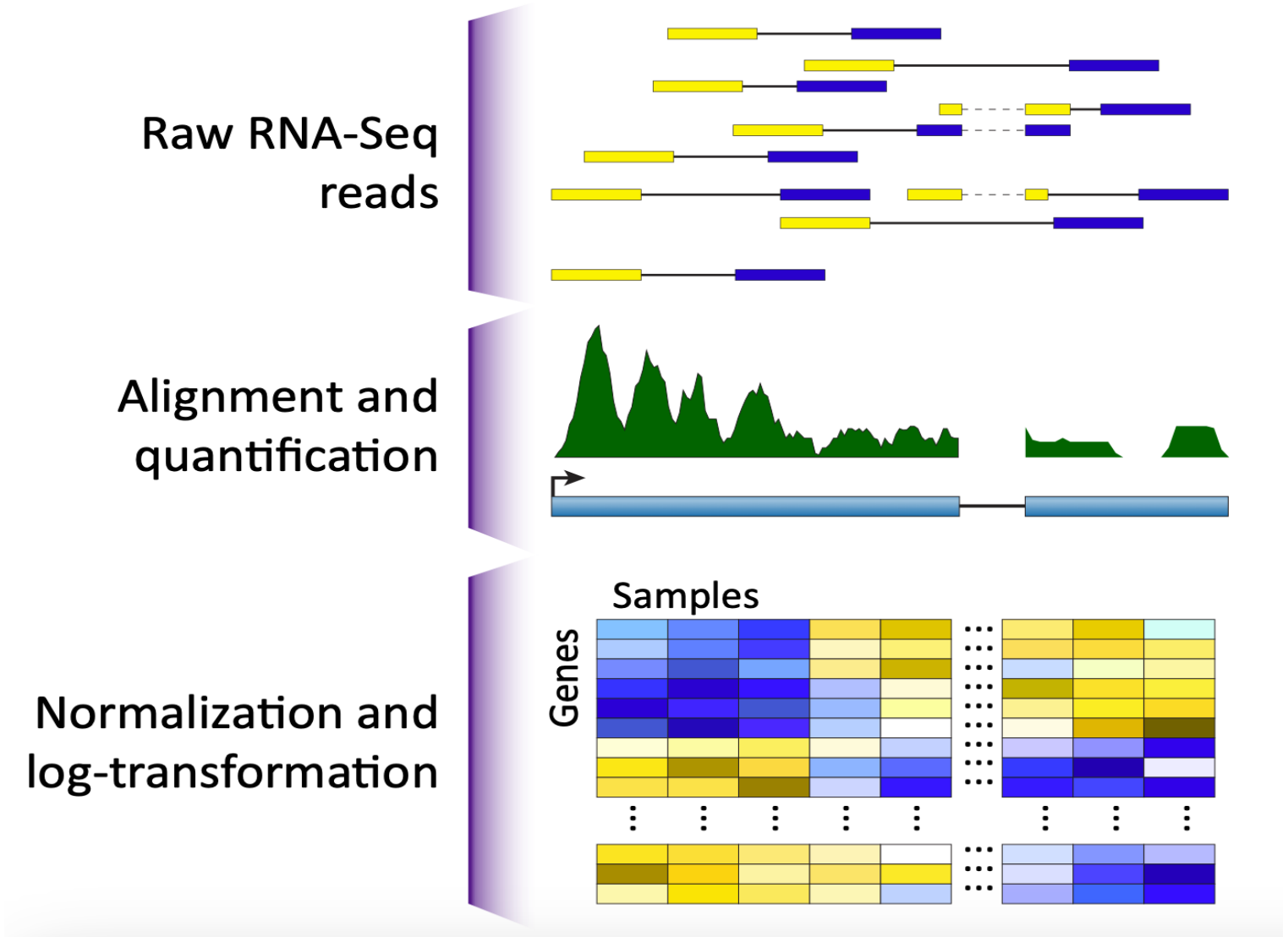
1. Removing redundancies in the data
2. Identify most relevant information
3. Reduce computational time for downstream analysis
4. Exploratory data analysis

What is dimensionality reduction?

What is dimensionality reduction?

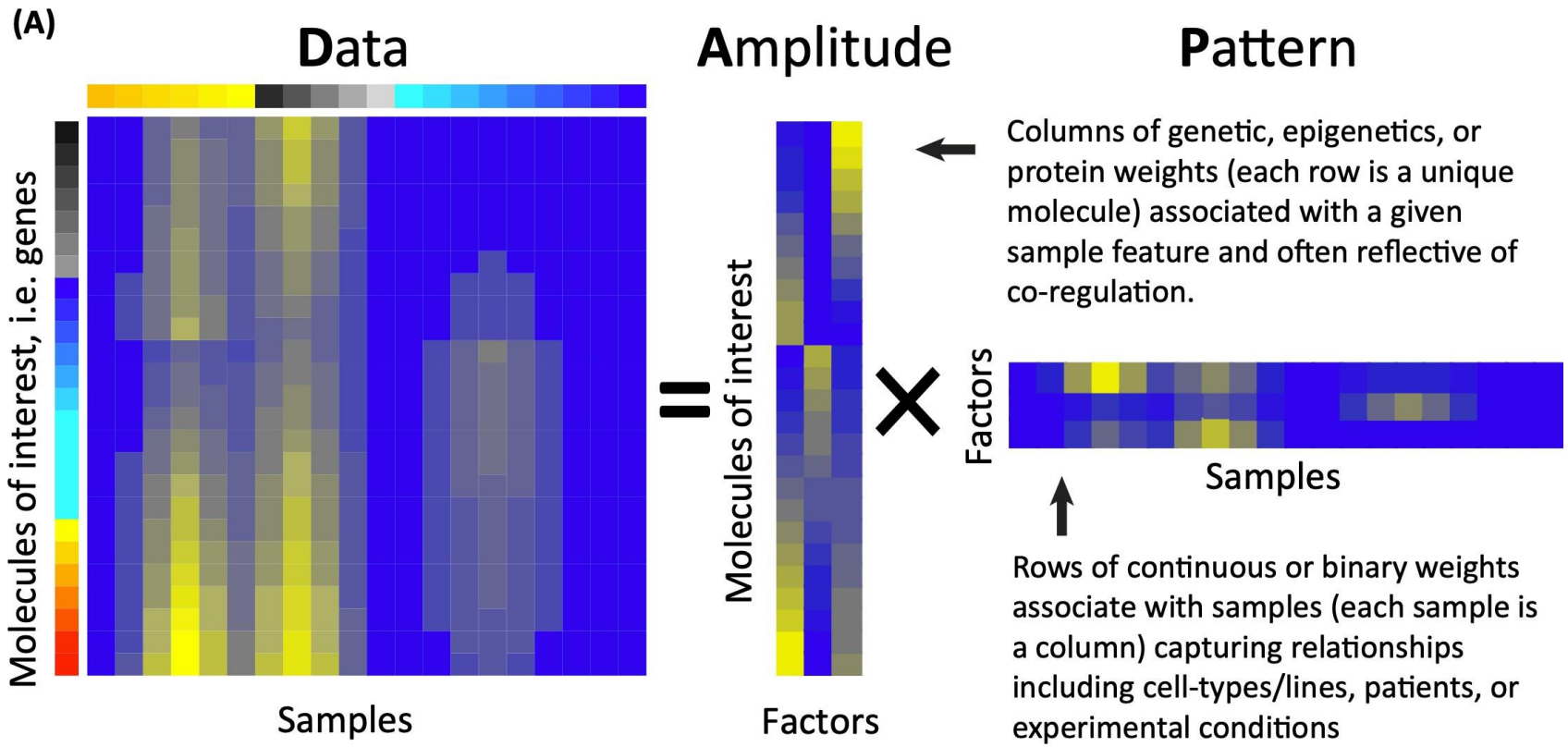


It is all about matrix



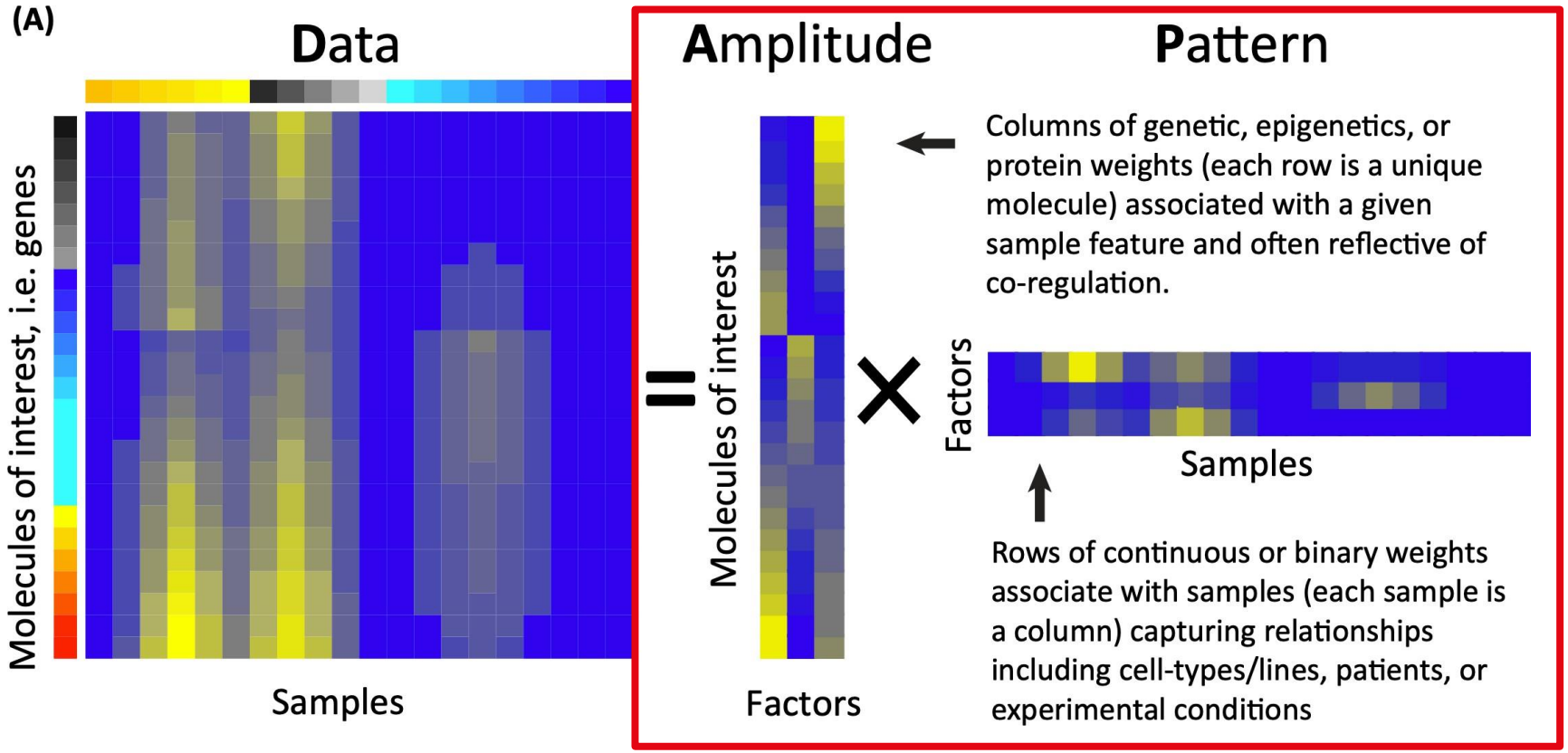
Adapted from : Stein-O'Brien, et al. Trends in Genetics (2018)

Matrix Factorization



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

Matrix Factorization



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

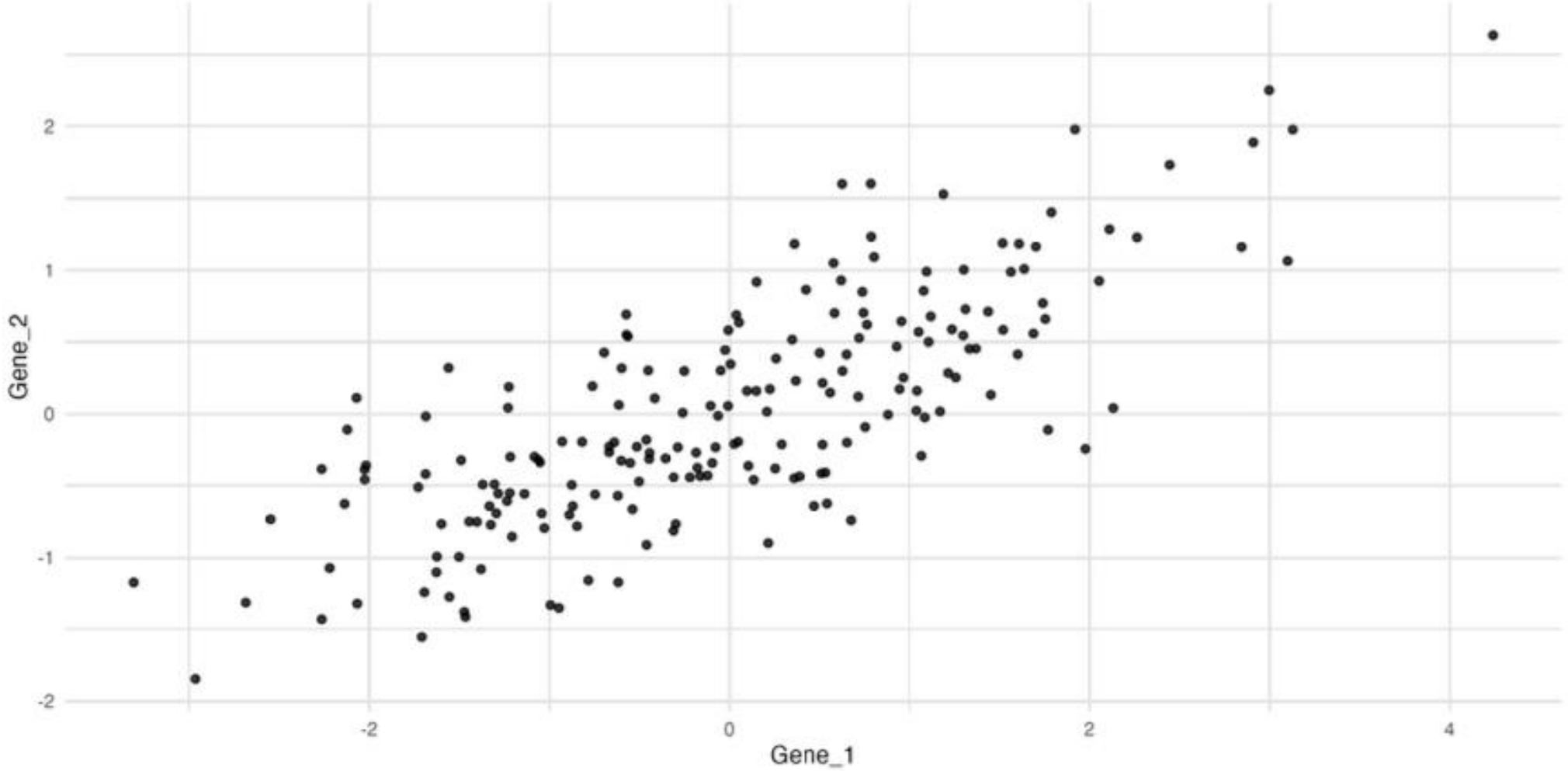
- **Eigenvectors** (Principal Components, PCs): Represent new axes (or directions) in the data space along which the variation is maximized (aka gene weights)
- **Eigenvalues**: Indicate the amount of variance explained by each PC

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.

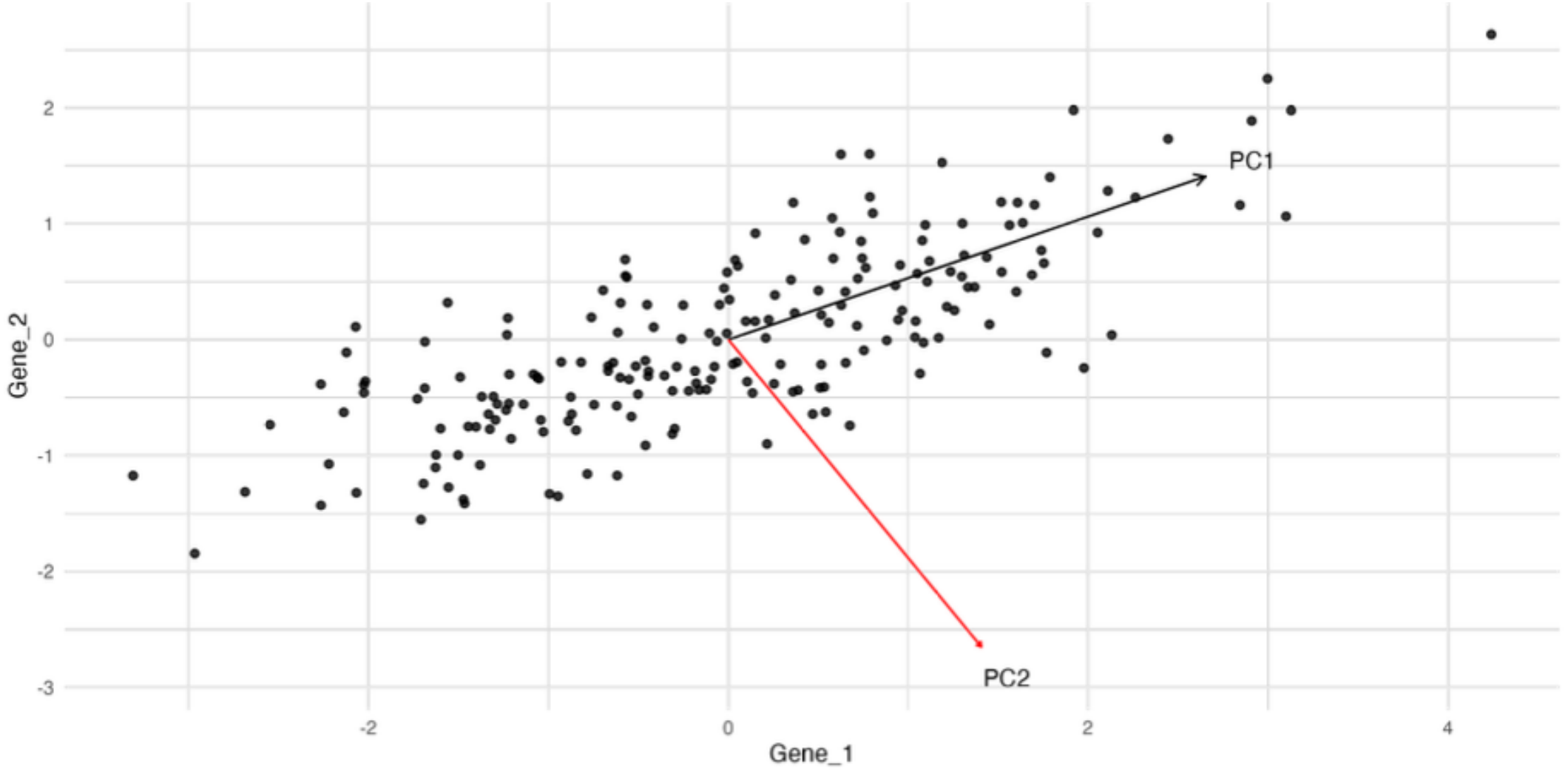
Principal Component Analysis

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



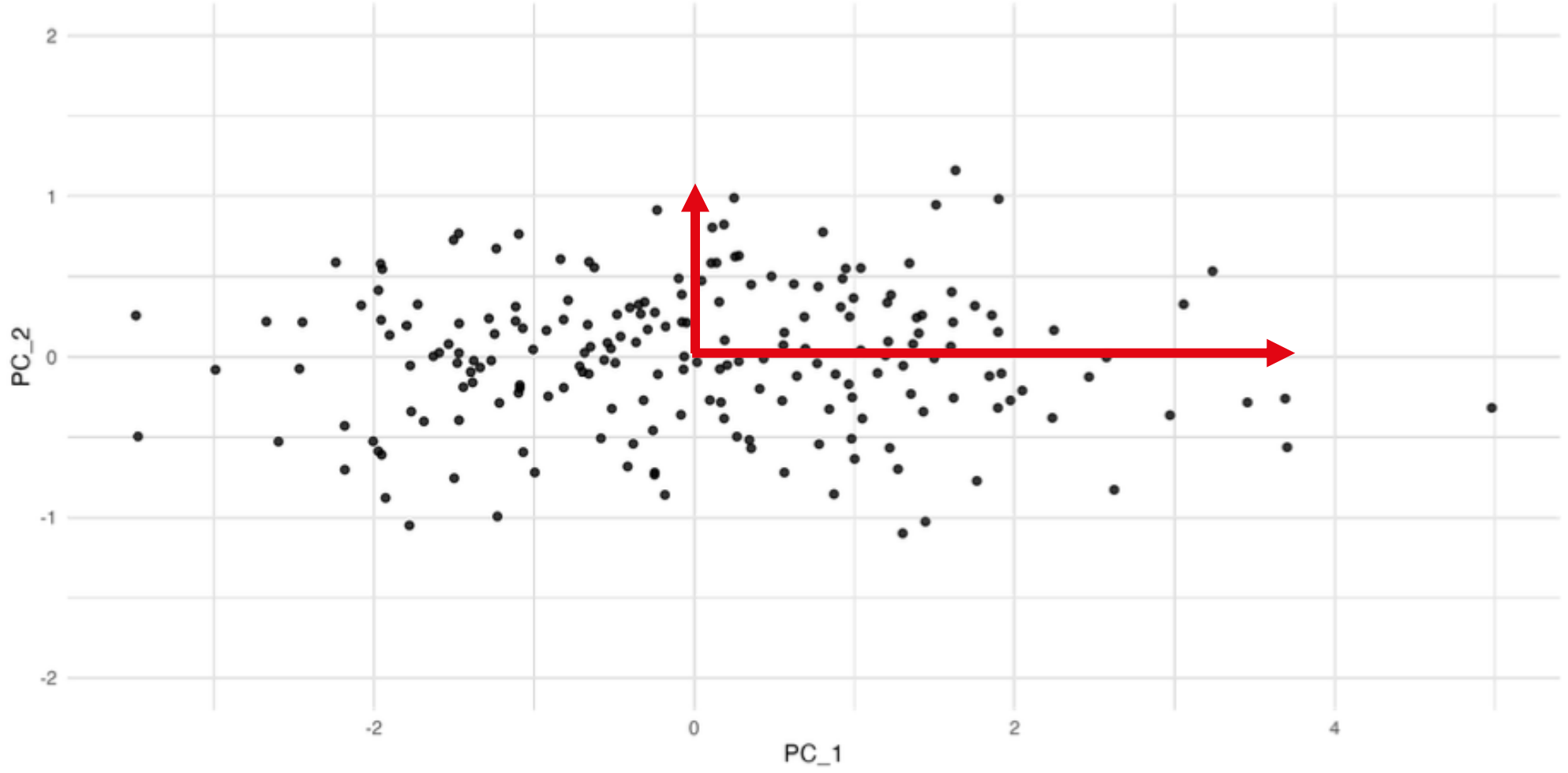
Principal Component Analysis

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



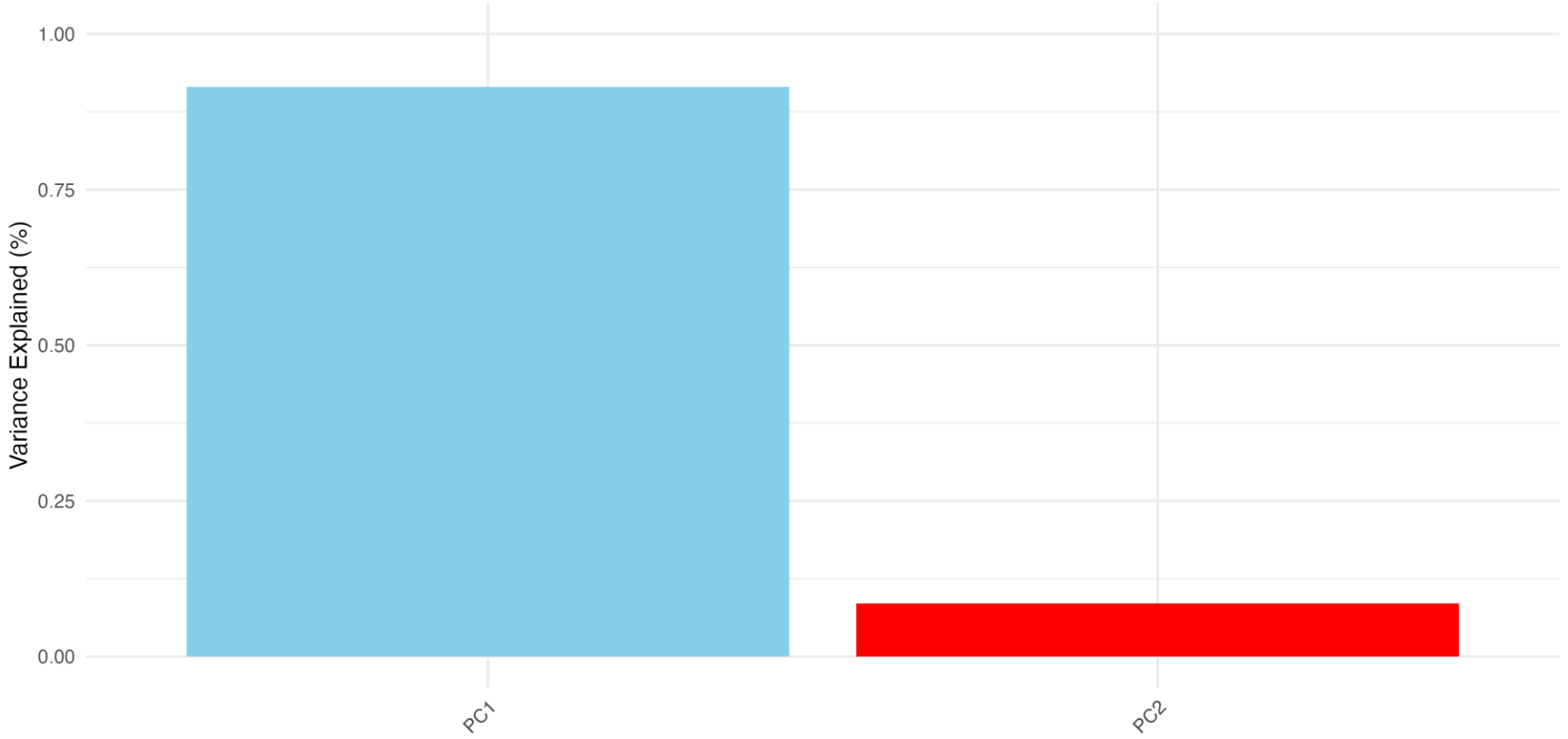
Principal Component Analysis

New axis that are linear combination of the original axes



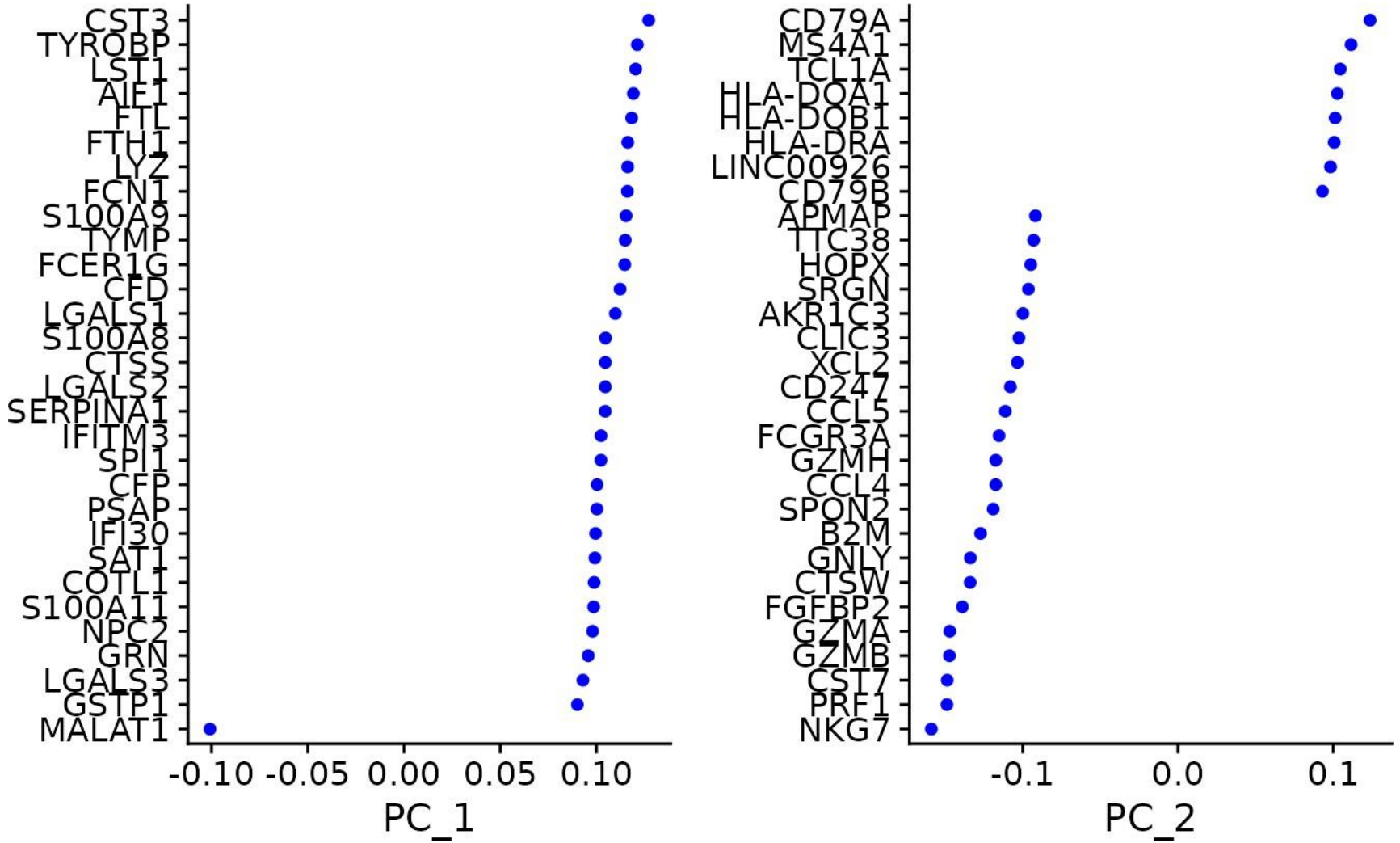
Principal Component Analysis

New axis that are linear combination of the original axes



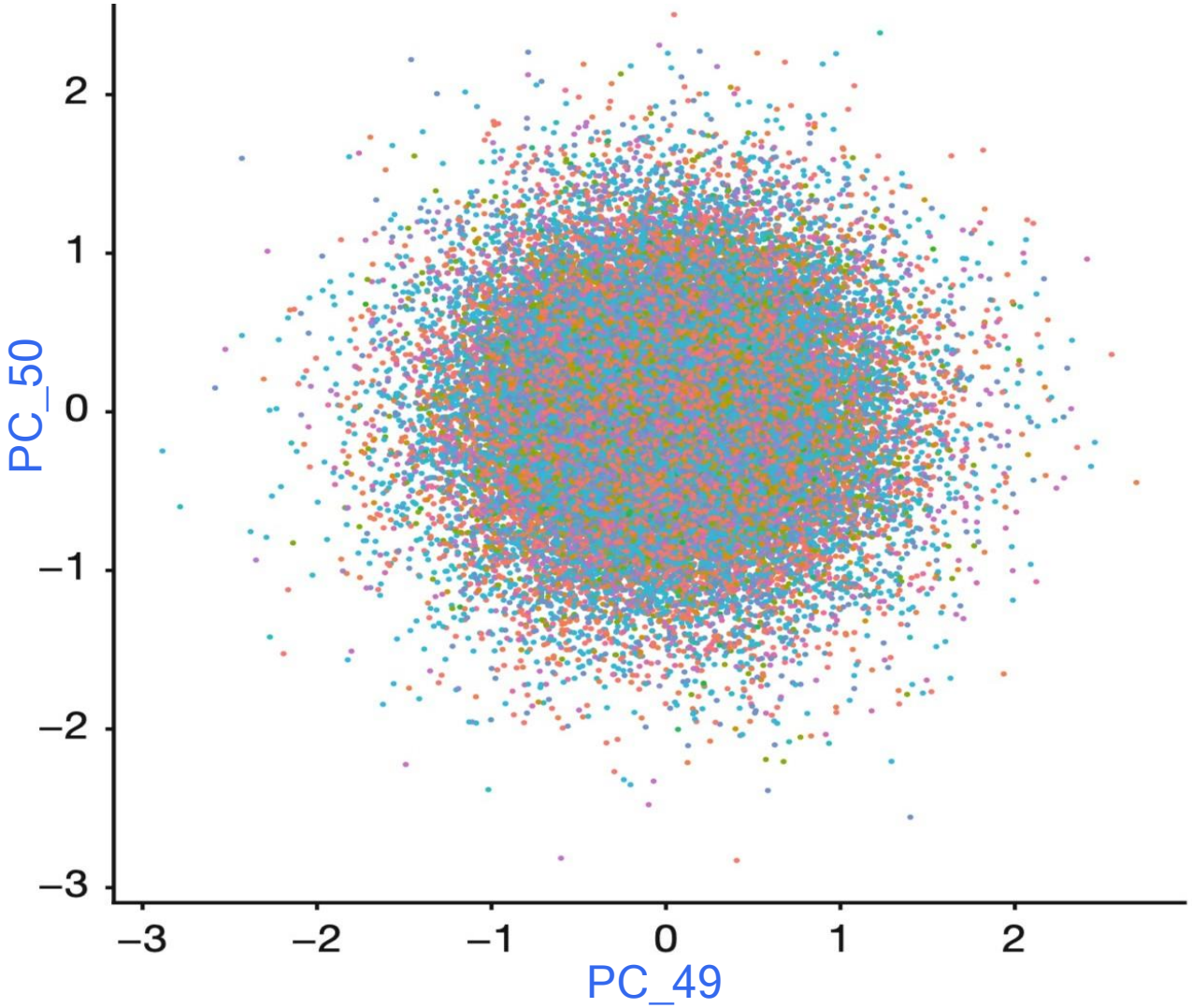
Choosing the number of Principal Components

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 Principal Components

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 Principal Components

Choosing the number of Principal Components

We could use some heuristic approaches:

- PCs that explain at least 1% of variance

Choosing the number of Principal Components

We could use some heuristic approaches:

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

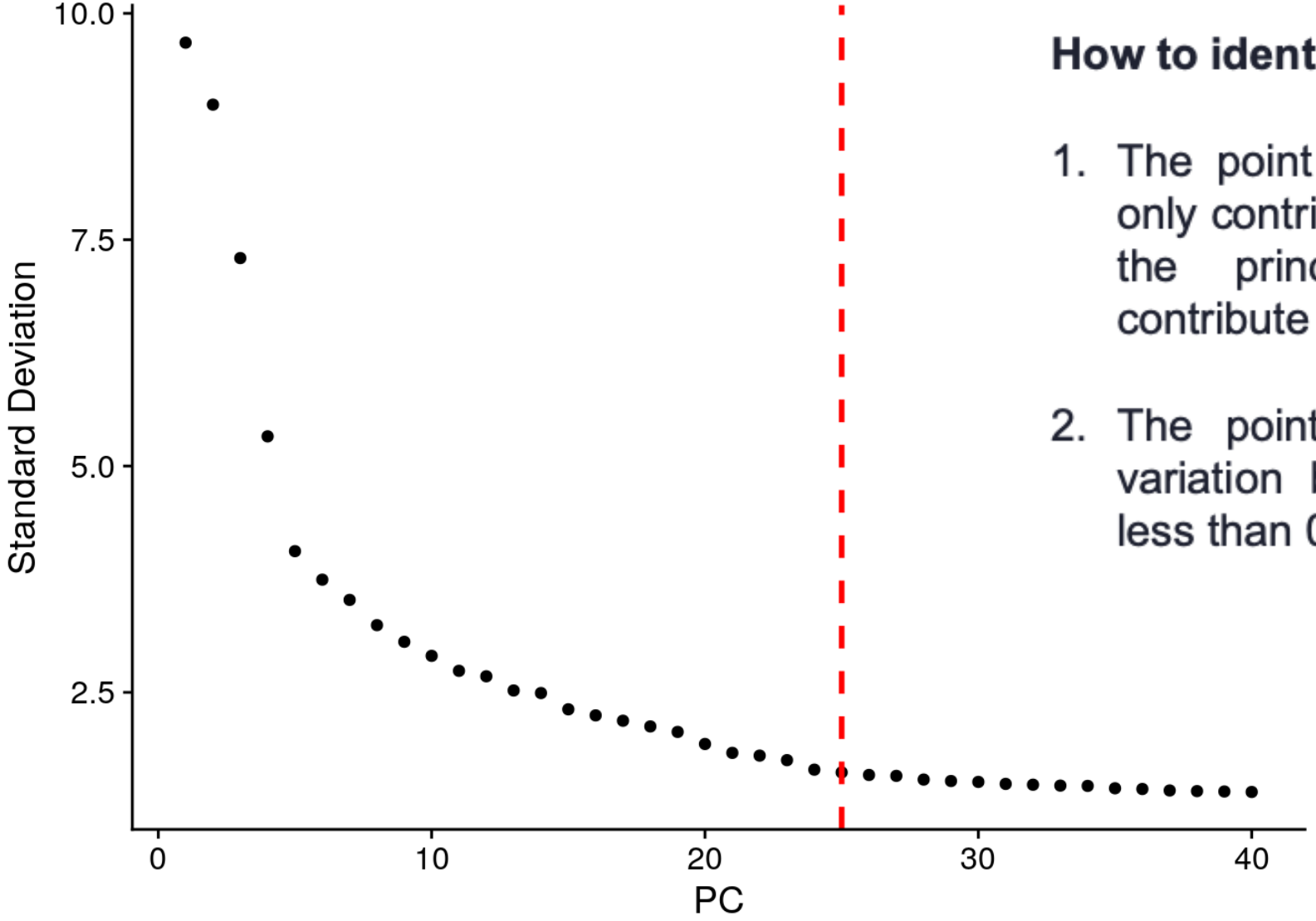
Choosing the number of Principal Components

We could use some heuristic approaches:

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

Principal Component Analysis

Choosing the number of PCs (elbow point)

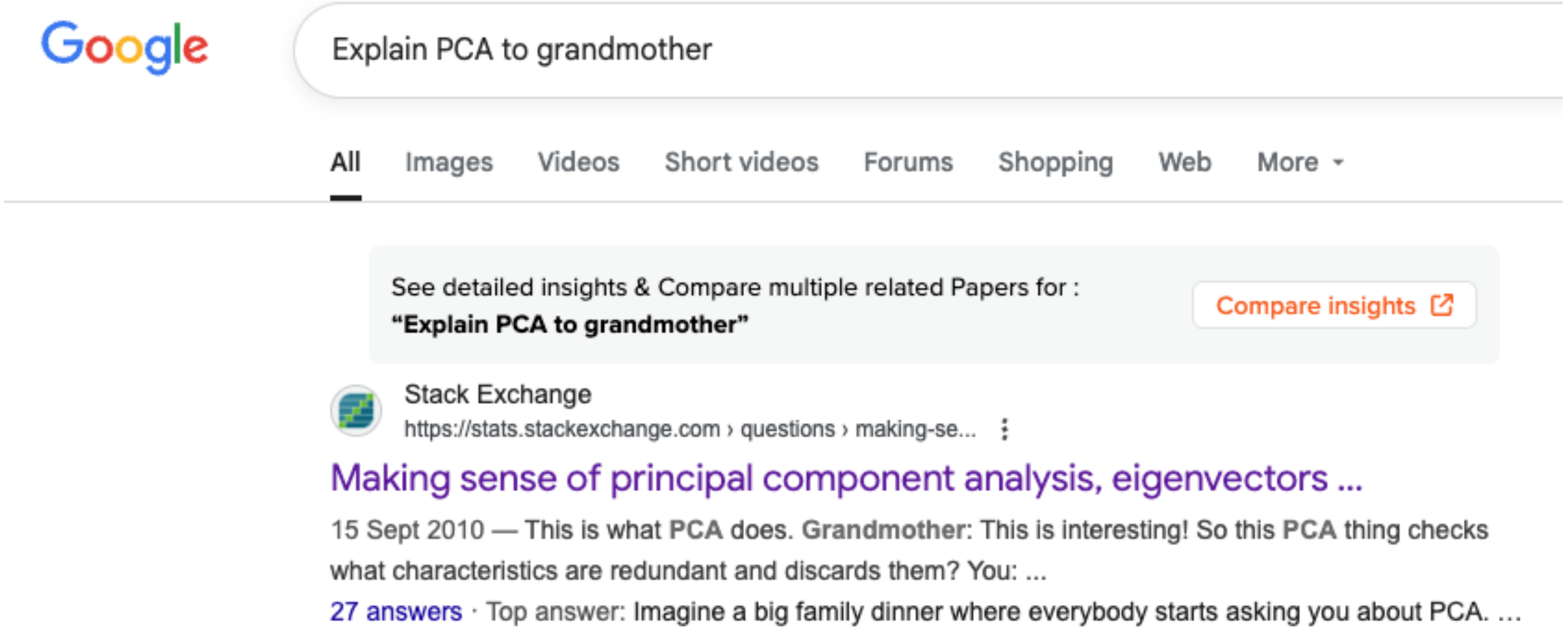


How to identify the Elbow point:

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%.

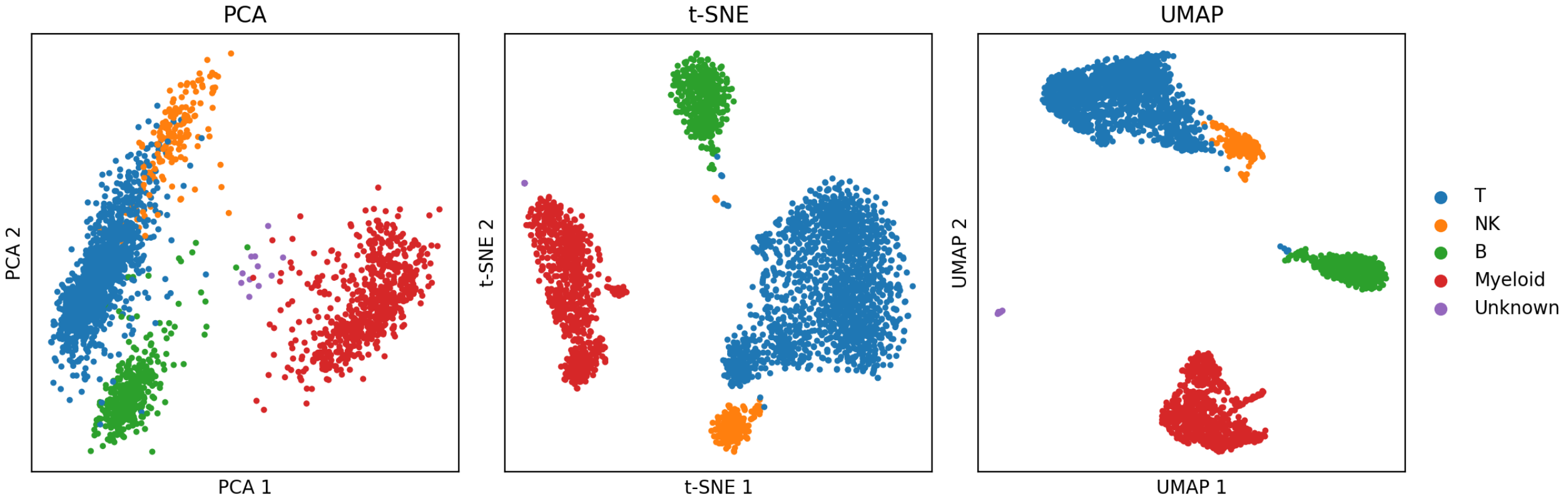
Diving into Principal Component Analysis



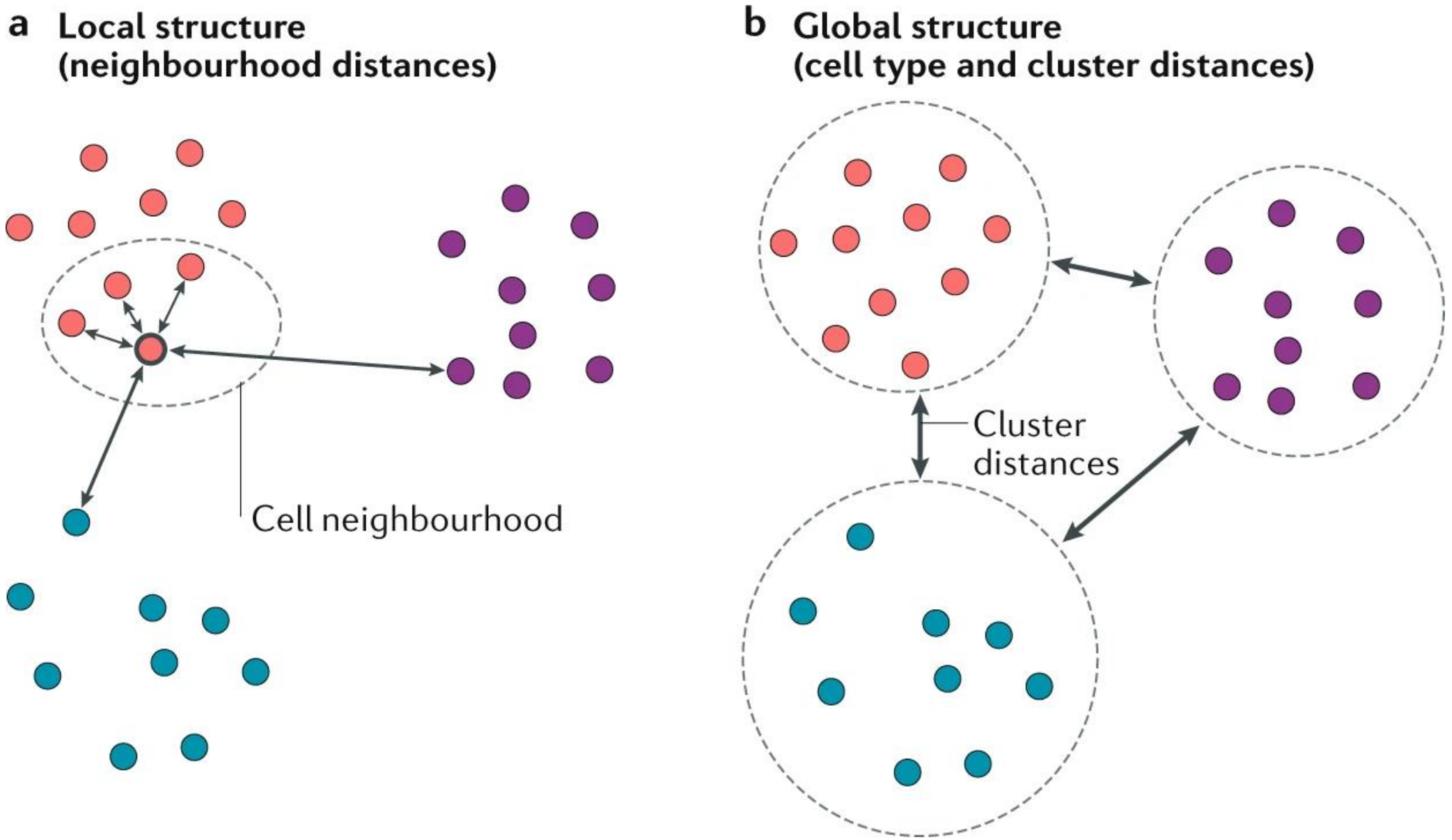
Practical considerations

- 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..

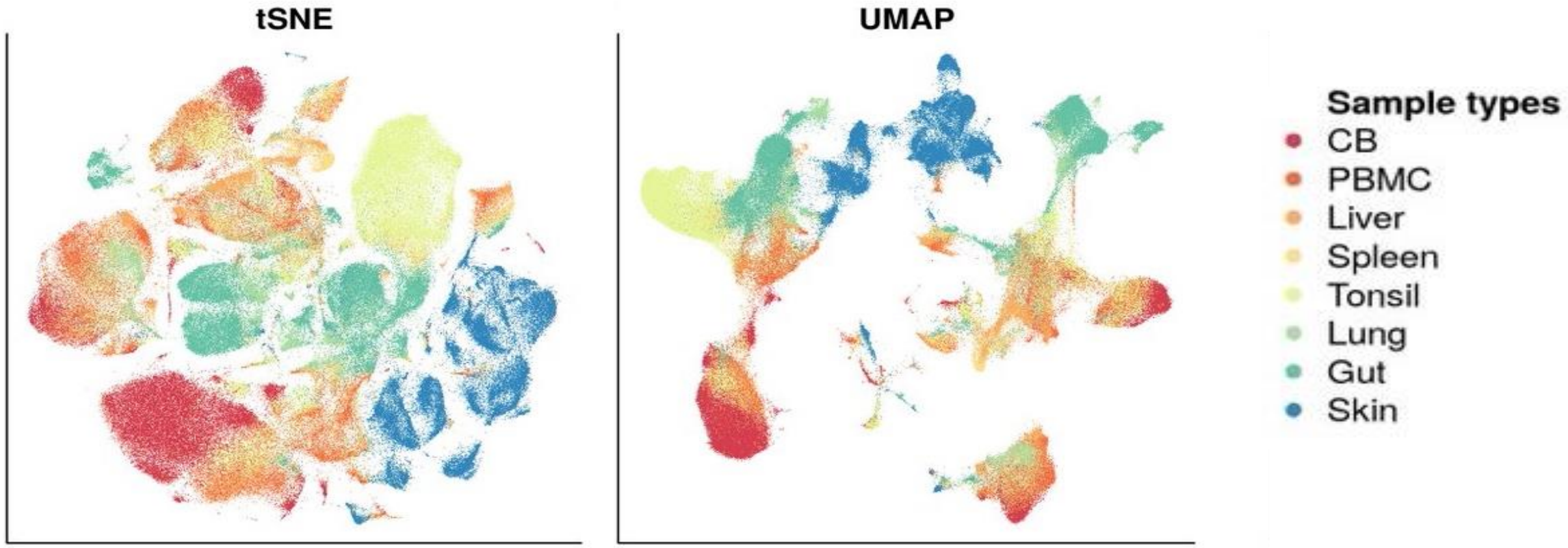
Data visualization techniques



Structures in data



Global vs local structures



UMAP (Uniform Manifold Approximation and Projection)

UMAP helps visualize high-dimensional data in a low-dimensional space:

1. UMAP preserves both the local and global structure of the data, allowing researchers to identify cell clusters and relationships between different cell types
2. UMAP allows for easy visualization of complex cellular heterogeneity and developmental trajectories
3. Compared to t-SNE, UMAP is faster and scales better with large datasets, making it ideal for single-cell datasets
4. Works on original data, but best on PCA reduced dimension (default in Seurat)

Why UMAP is performed after PCA?

Why UMAP is performed after PCA?

PCA acts as a filter:

1. PCA helps **reduce noise** by capturing the most **informative features** (principal components)
2. UMAP is computationally **faster and more accurate** when working on a smaller number of dimensions (like 20 PCs) instead of the original thousands of genes
3. PCA removes **redundant and highly correlated features**, preventing UMAP from overfitting to technical noise or batch effects

Why UMAP is performed after PCA?

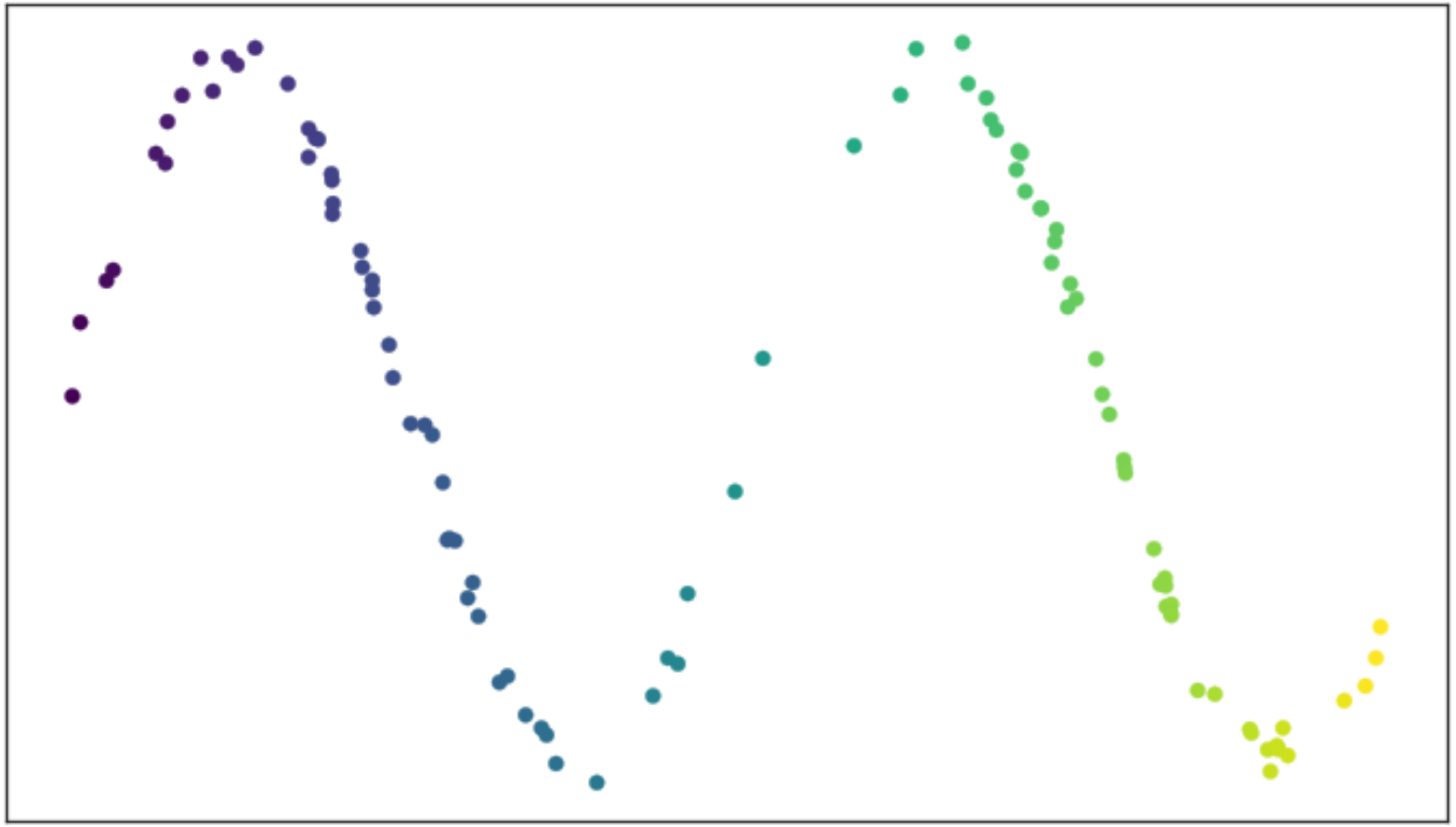
PCA acts as a filter:

1. PCA helps **reduce noise** by capturing the most **informative features** (principal components)
2. UMAP is computationally **faster and more accurate** when working on a smaller number of dimensions (like 20 PCs) instead of the original thousands of genes
3. PCA removes **redundant and highly correlated features**, preventing UMAP from overfitting to technical noise or batch effects

*Reduces the data from **~20,000 genes to ~50 principal components**, which is still enough for UMAP to **capture both global and local structure**.*

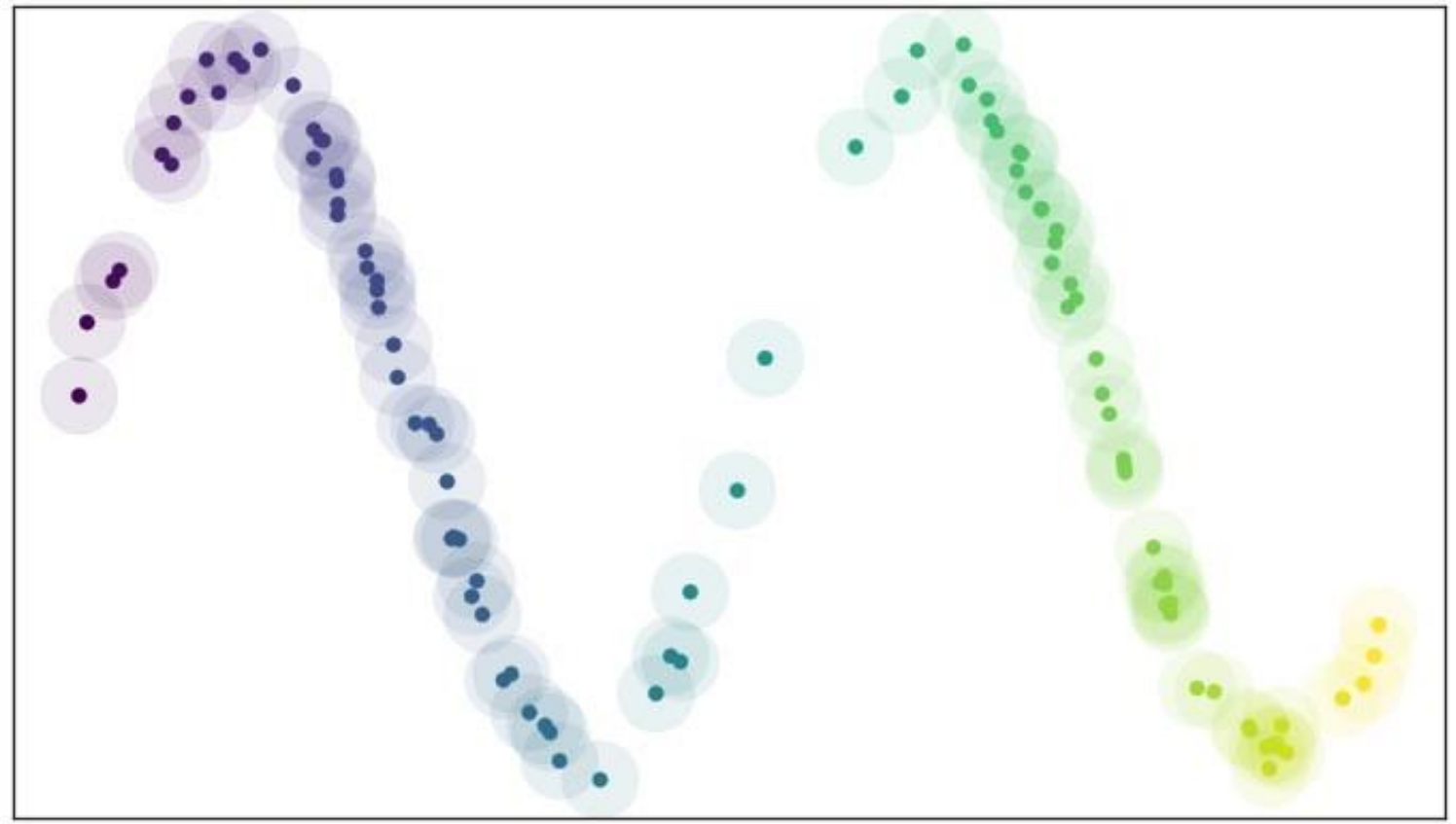
UMAP Theory

UMAP extends a radius outwards from each point



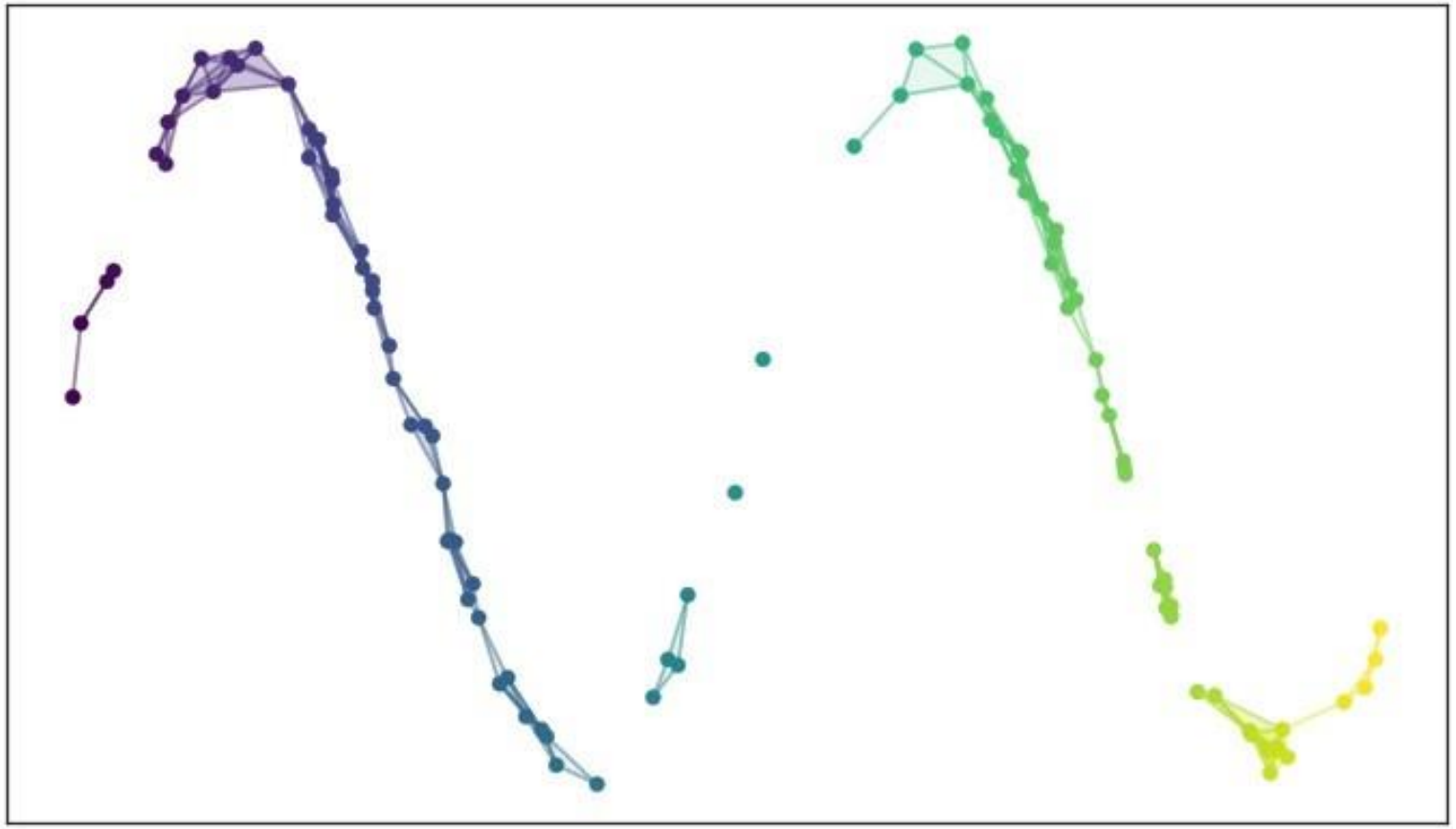
UMAP Theory

UMAP extends a radius outwards from each point



UMAP Theory

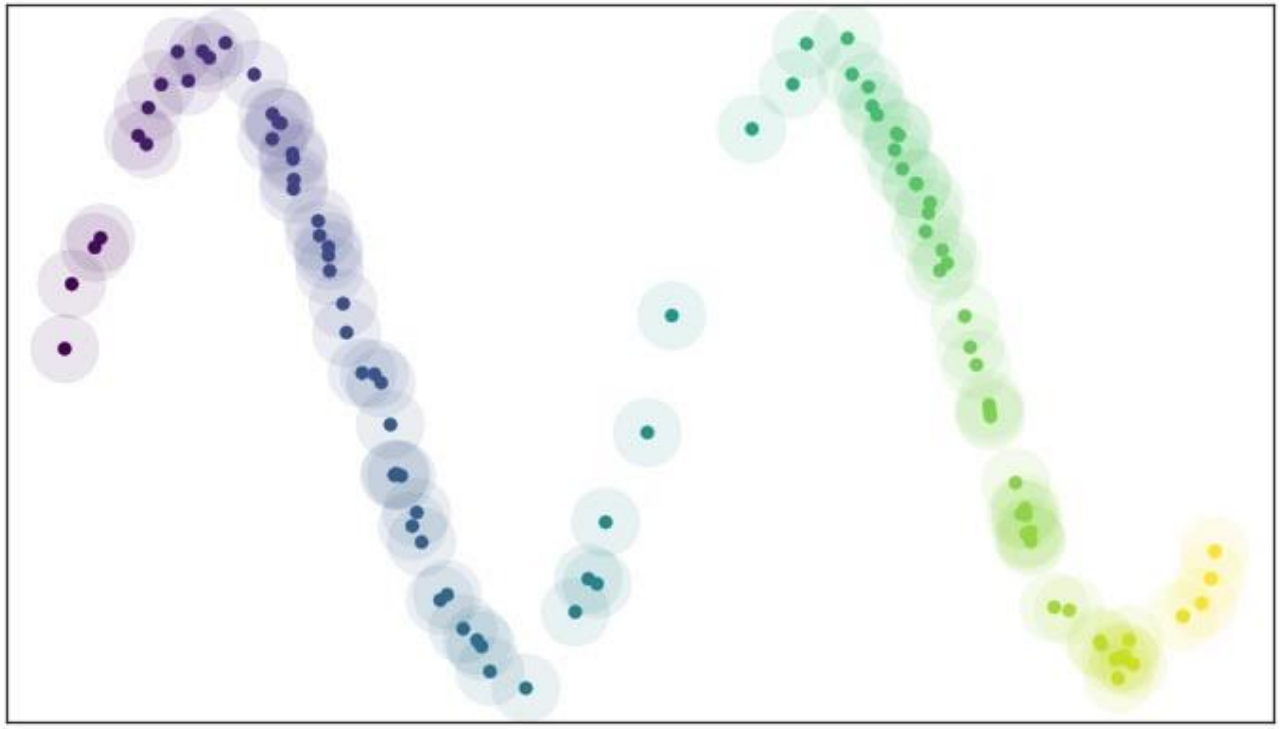
Connect points when those radii overlap



UMAP Theory

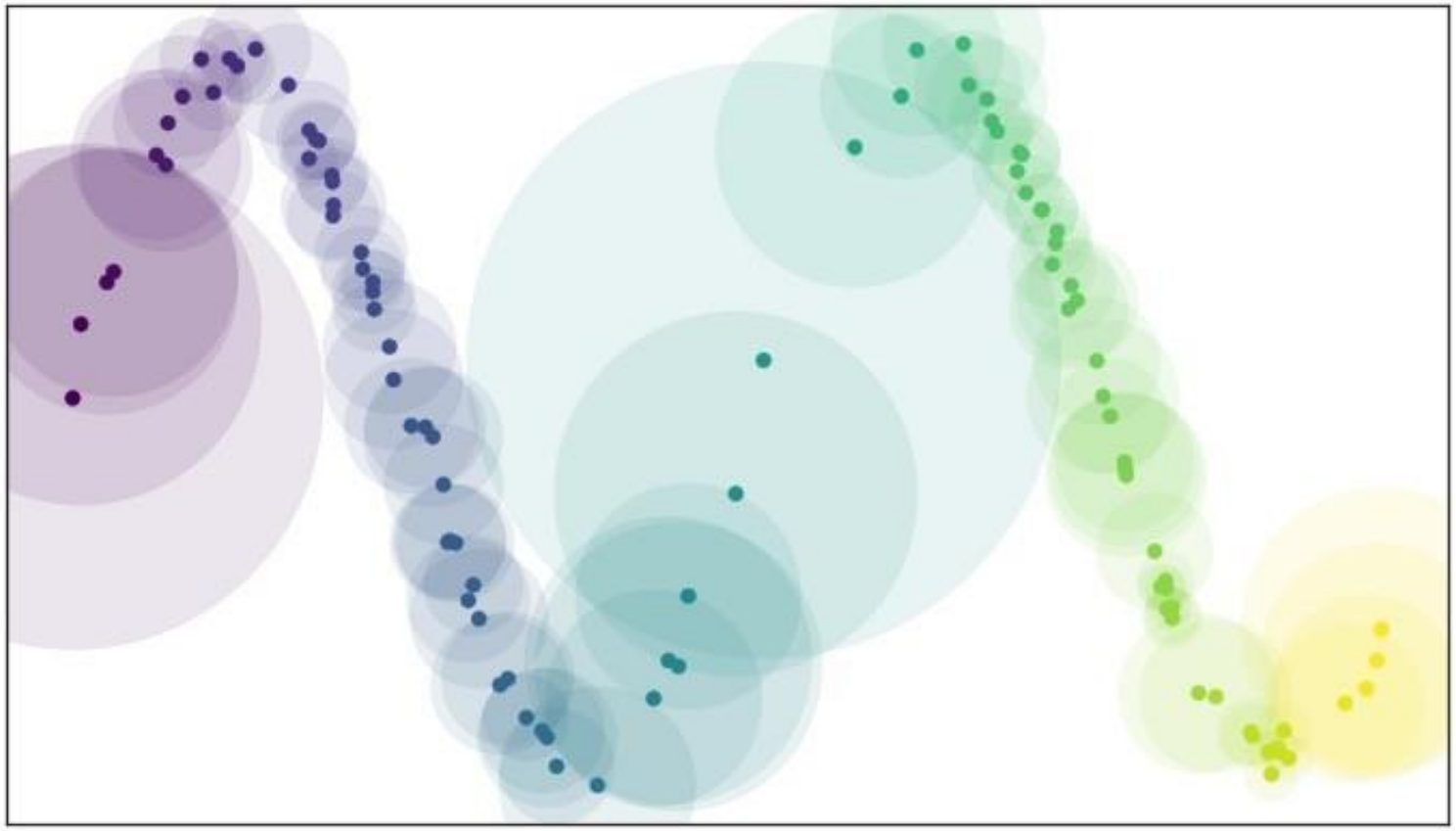
Choosing this radius is critical:

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



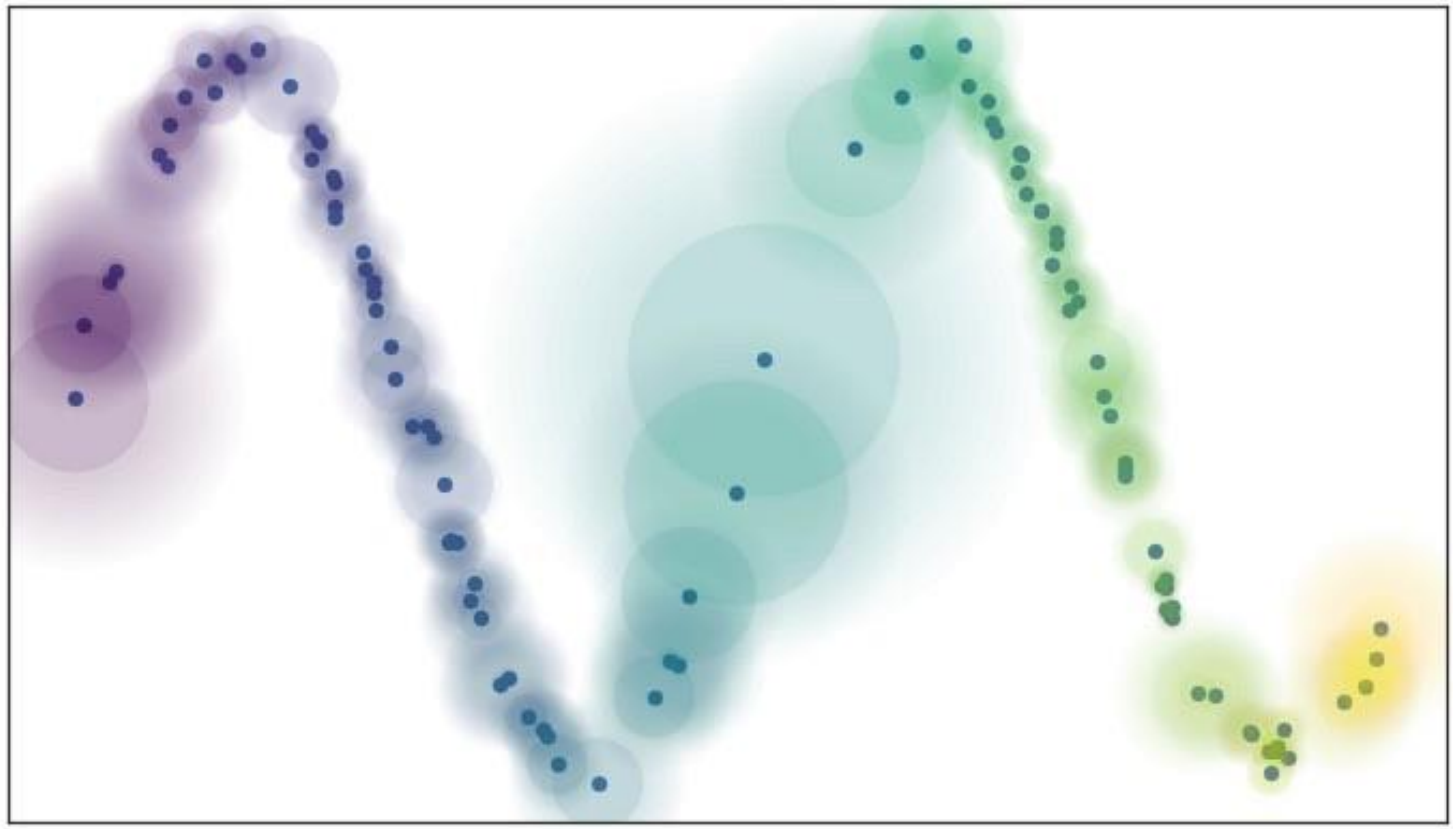
UMAP Theory

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**.



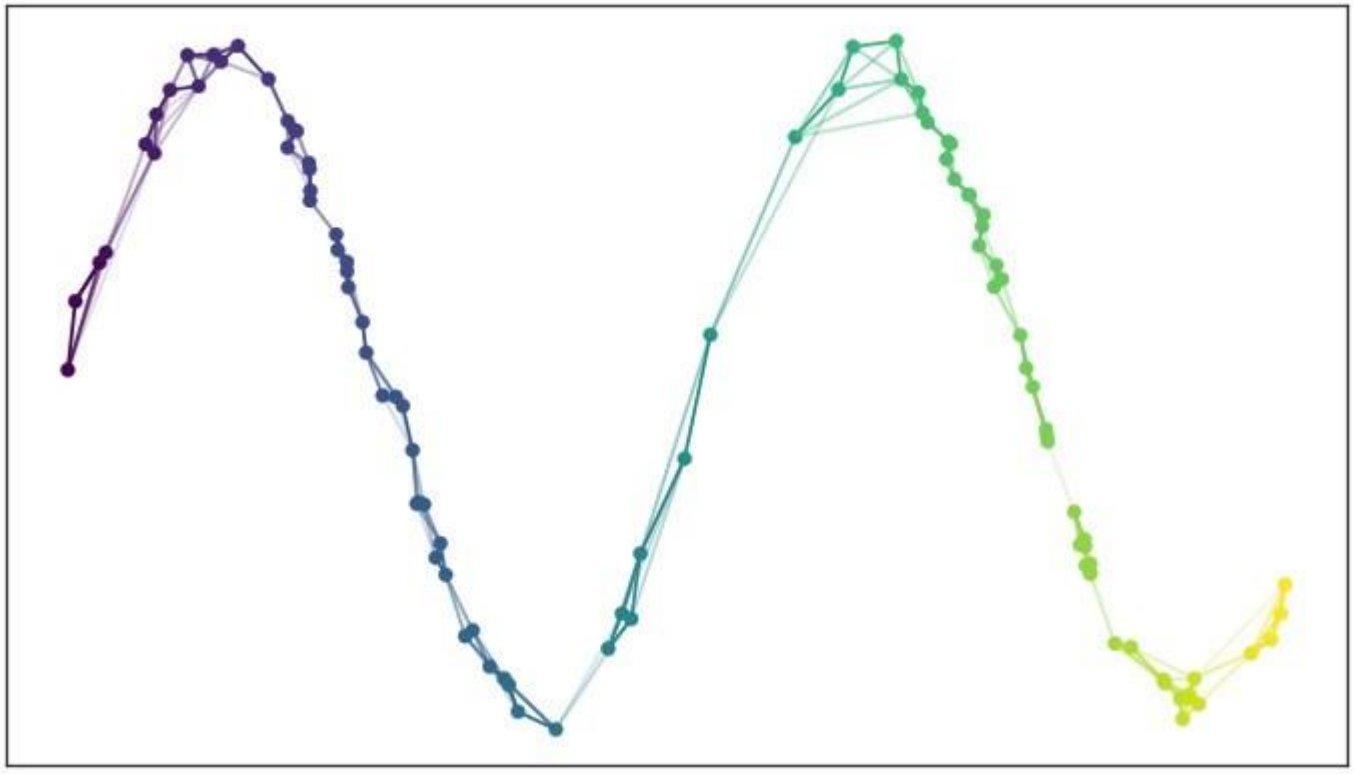
UMAP Theory

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



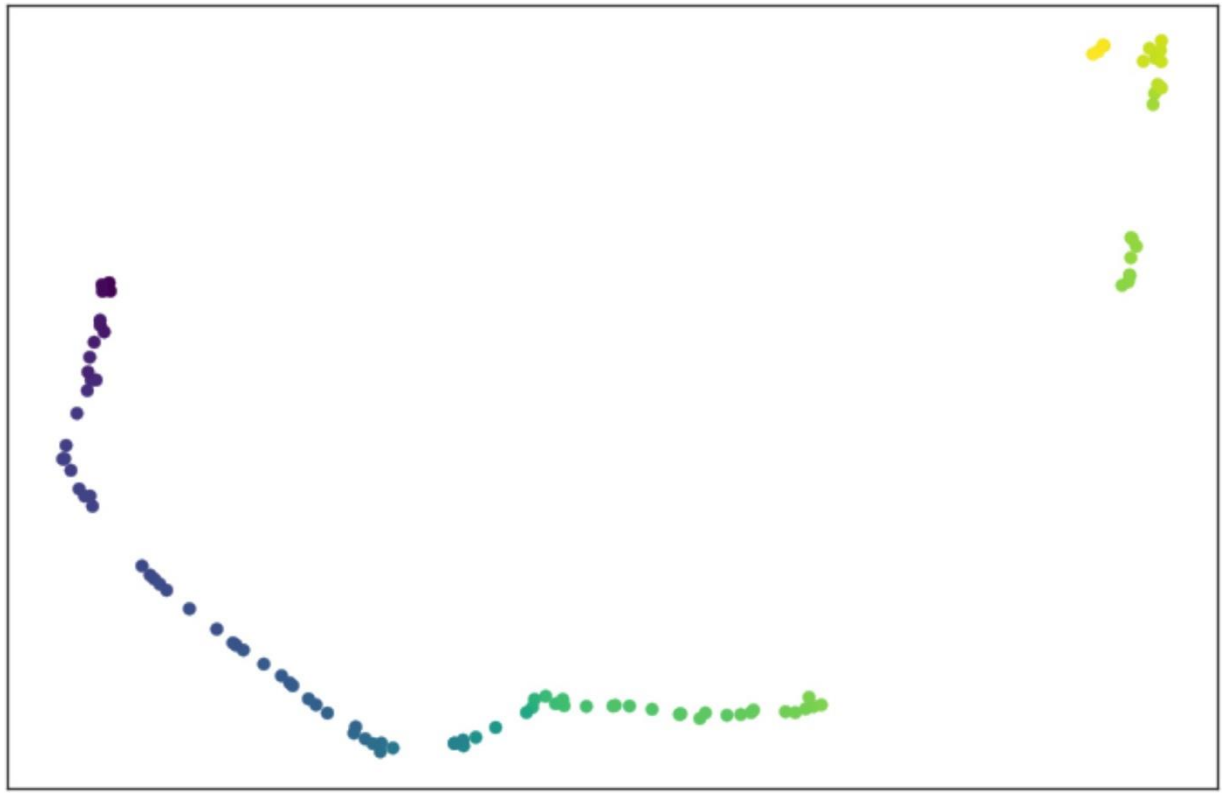
UMAP Theory

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.



UMAP Theory: Final Step

UMAP first builds a graph of cell relationships, then arranges the cells in 2D/3D so that neighbors stay close and others stay apart.



UMAP: Key hyperparameters

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.**

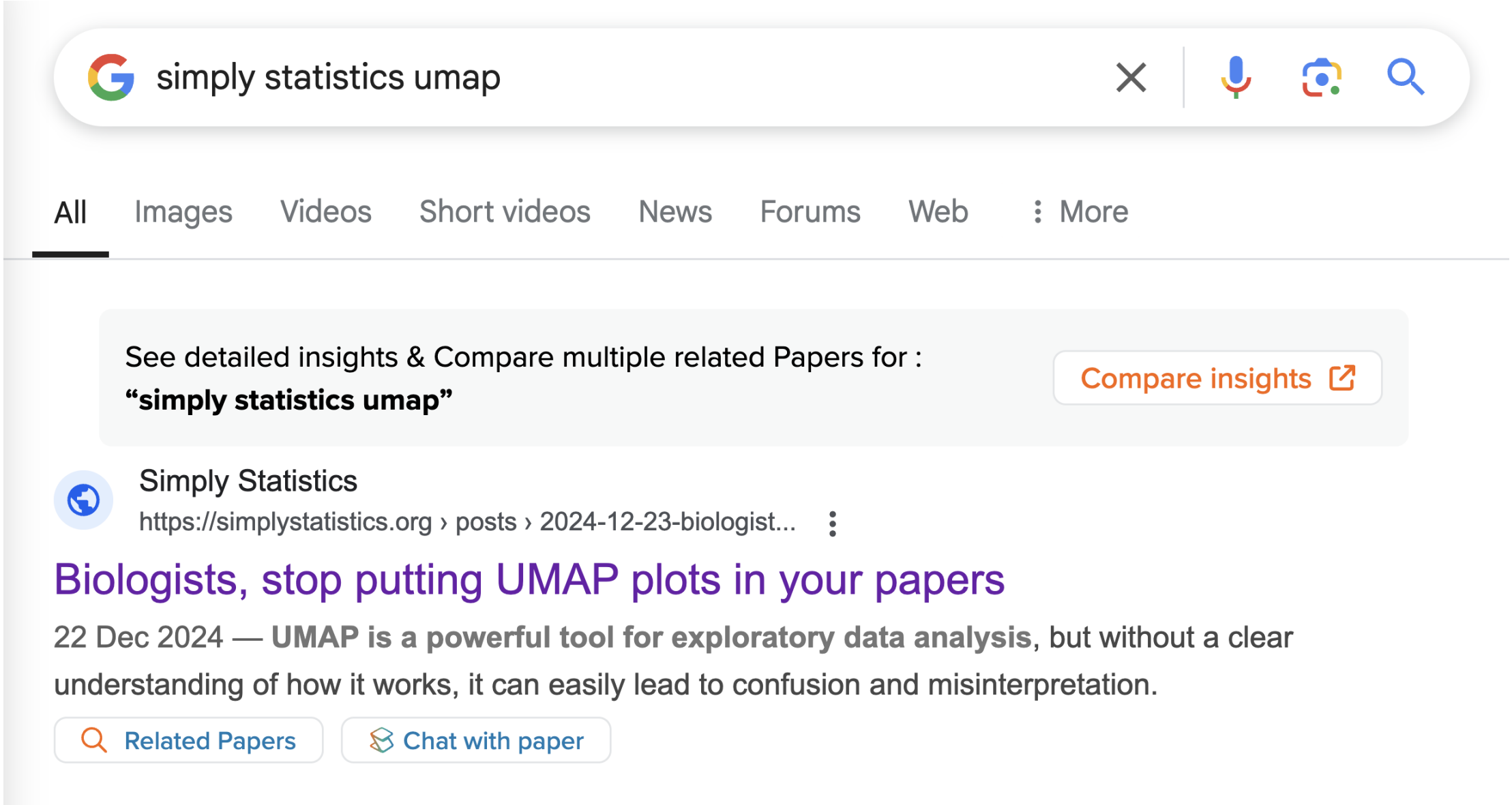
     Lower values (e.g., 0.001) will result in more compact clusters, making it easier to identify tight groupings.

     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).

When and when not to use UMAP?



Skepticism about UMAP

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.

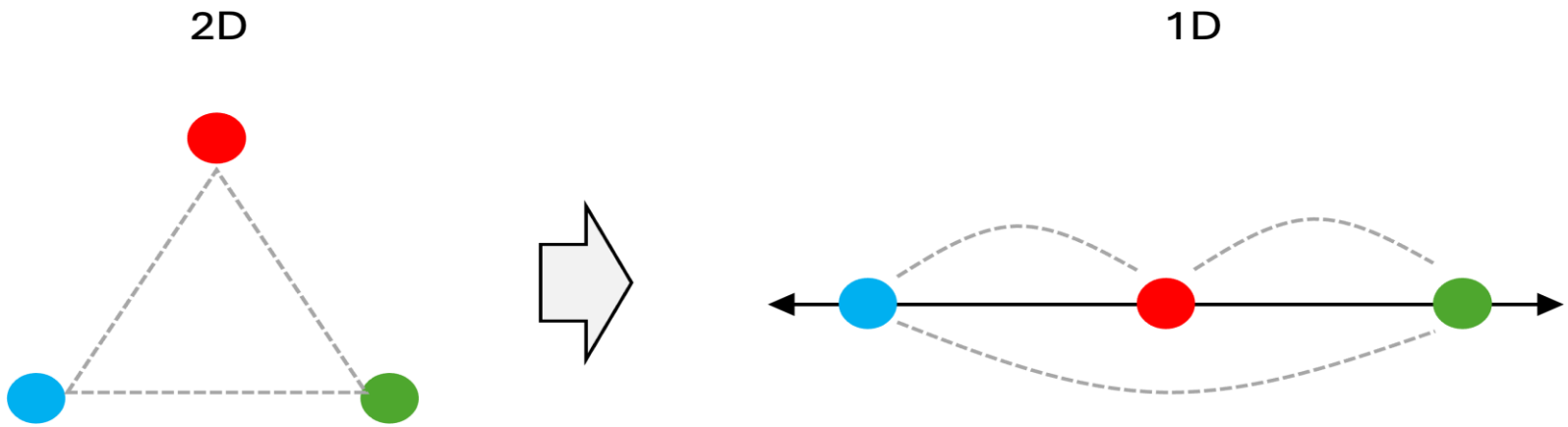
Groups or clusters that appear well-separated in the 2D plot might not actually be as distinct in the high-dimensional space

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)

Considerations

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 omission, 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

Quiz

1. How to determine the number of PCs after PCA analysis?

A) Select the PCs with the highest eigenvalues.
B) Use the scree plot and select the point where the "elbow" occurs.
C) Retain all components to avoid loss of information.
D) Select the PCs that capture at least 50% of the variance.

2. Which technique aims to preserves both global and local structure of the scRNA-seq data?

A) PCA

B) UMAP

C) PCA + UMAP

D) None

Summary

Curse of Dimensionality: High-dimensional data often contains noise and redundancy

Need for Dimensionality Reduction: Essential for efficient and effective data analysis

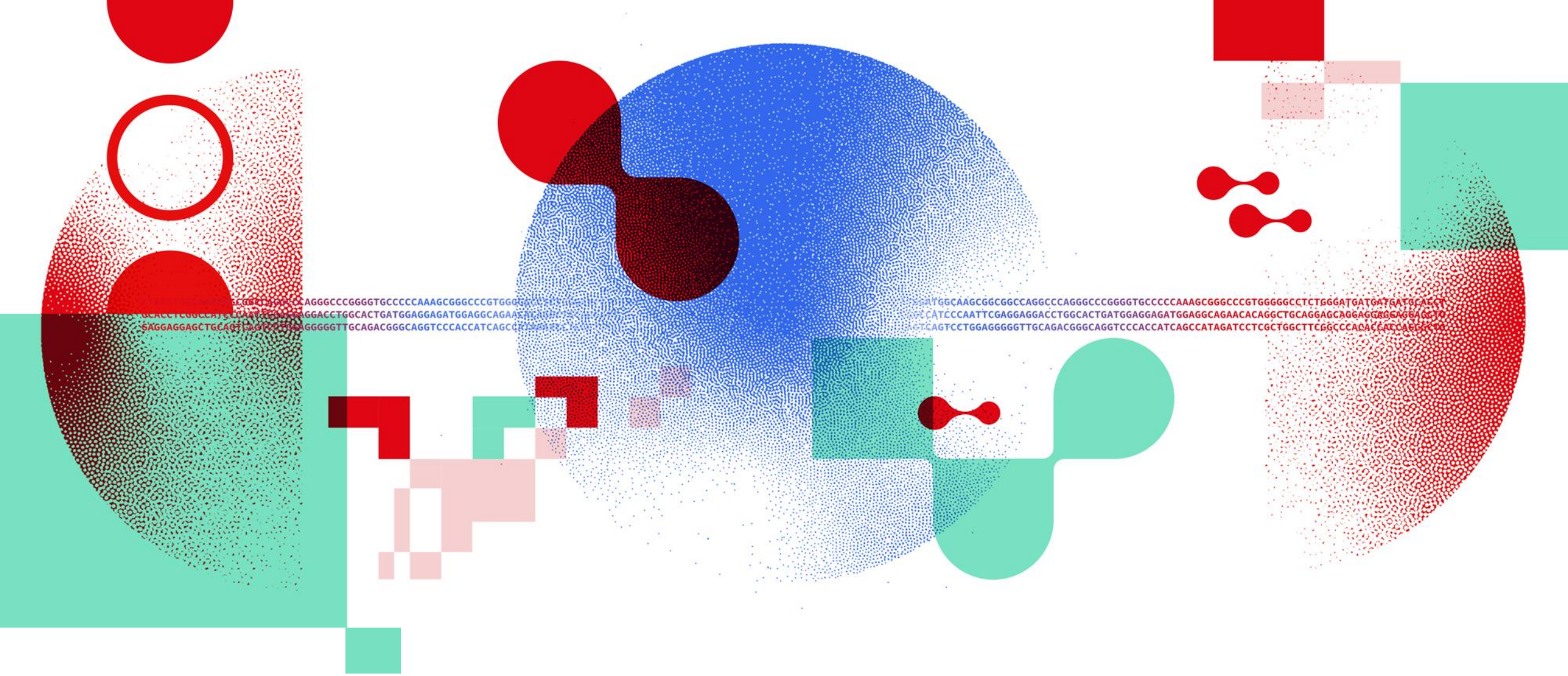
Principal Component Analysis (PCA): Identifies key directions in data, reduces dimensions

UMAP: Visualizes data, preserves structure, faster and scalable

PCA + UMAP Workflow: PCA reduces noise, UMAP visualizes reduced dimensions

References

1. R. Bellman, R.E. Bellman, and Rand Corporation. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957. URL: https://books.google.de/books?id=rZW4ugAACAAJ.
2. https://www.biostars.org/p/381993/
3. UMAP: https://www.youtube.com/watch?v=eN0wFzBA4Sc
4. https://www.mdpi.com/2079-7737/13/7/512
5. https://simplystatistics.org/posts/2024-12-23-biologists-stop-including-umap-plots-in-your-papers/
6. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011288
7. https://sites.gatech.edu/omscs7641/2024/03/10/no-straight-lines-here-the-wacky-world-of-non-linear-manifold-learning/
8. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012403
9. https://pair-code.github.io/understanding-umap/



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss