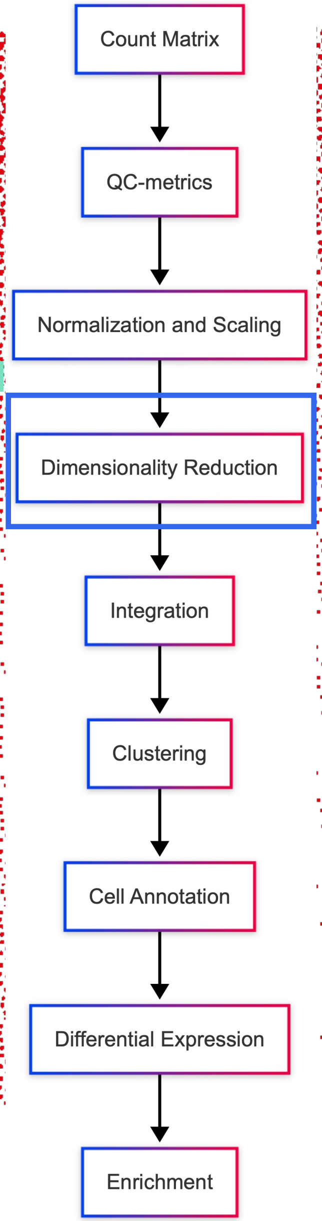
Swiss Institute of
Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Dimensionality reduction



Count Matrix

QC-metrics

Normalization and Scaling

Dimensionality Reduction

Integration

Clustering

Cell Annotation

Differential Expression

Enrichment

Deepak Tanwar

March 18-20, 2025

Adapted from previous year courses

Feedback from Geert van Geest

elixir SWITZERLAND

Learning objectives

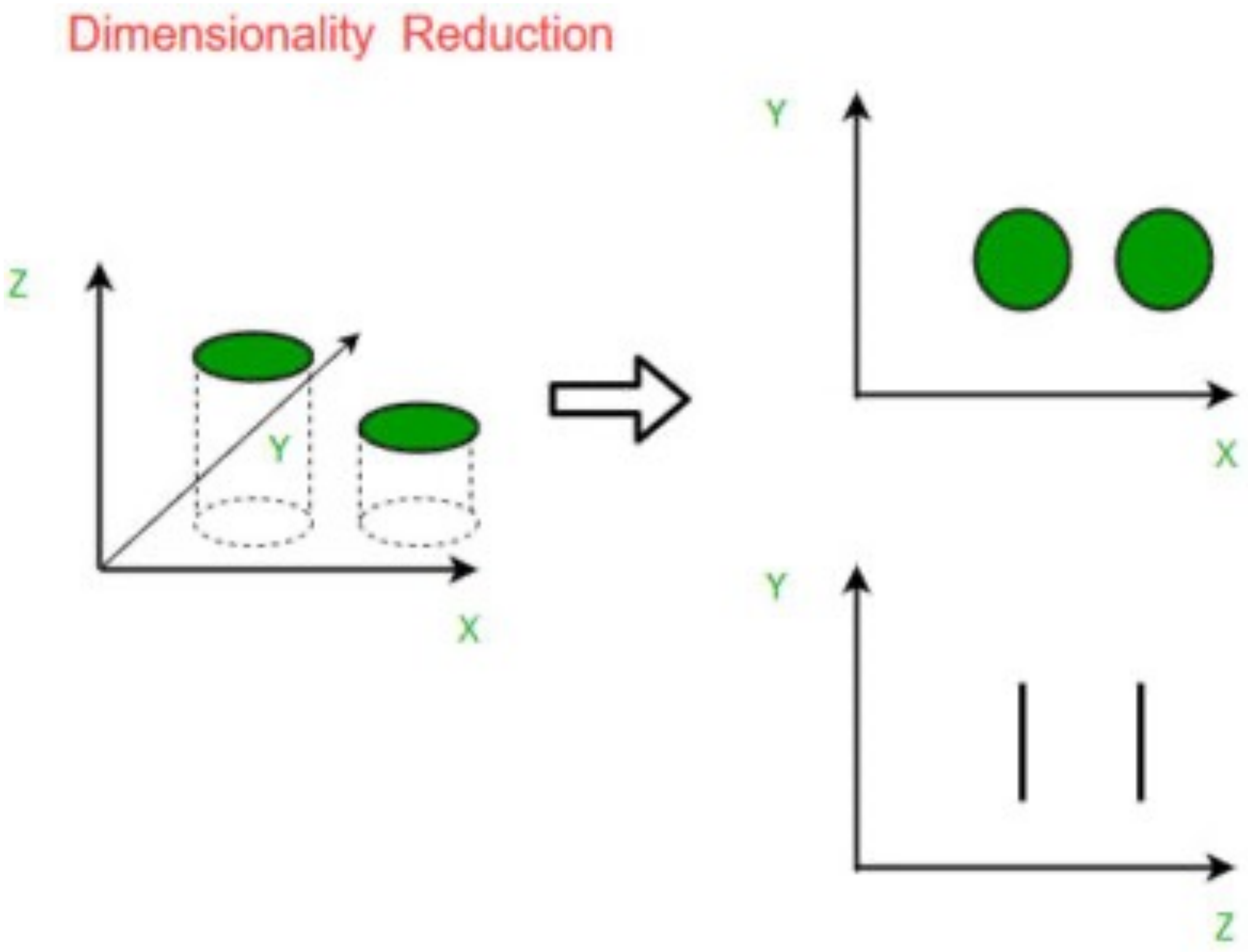
Understand the Curse of Dimensionality

Identify and Apply Dimensionality Reduction techniques

Understand PCA and UMAP

Determine appropriate scenarios for using UMAP in data analysis.

What is dimensionality reduction?



scRNA-seq is a high-throughput sequencing technology that produces datasets with high dimensions in the number of cells and genes. This immediately points to the fact that scRNA-seq data suffers from the ‘curse of dimensionality’.

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.

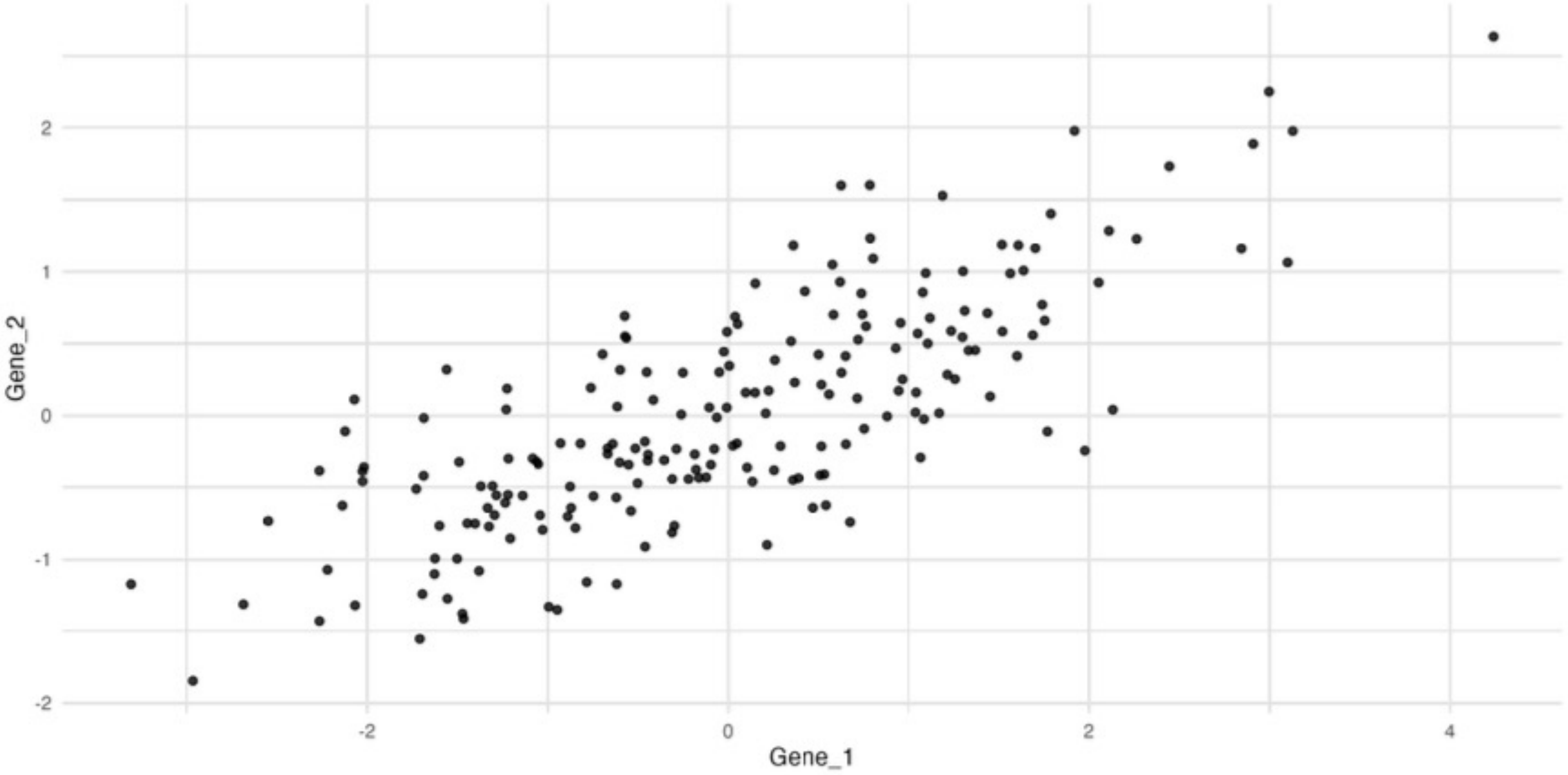
Need for dimensionality reduction in scRNA-seq data

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

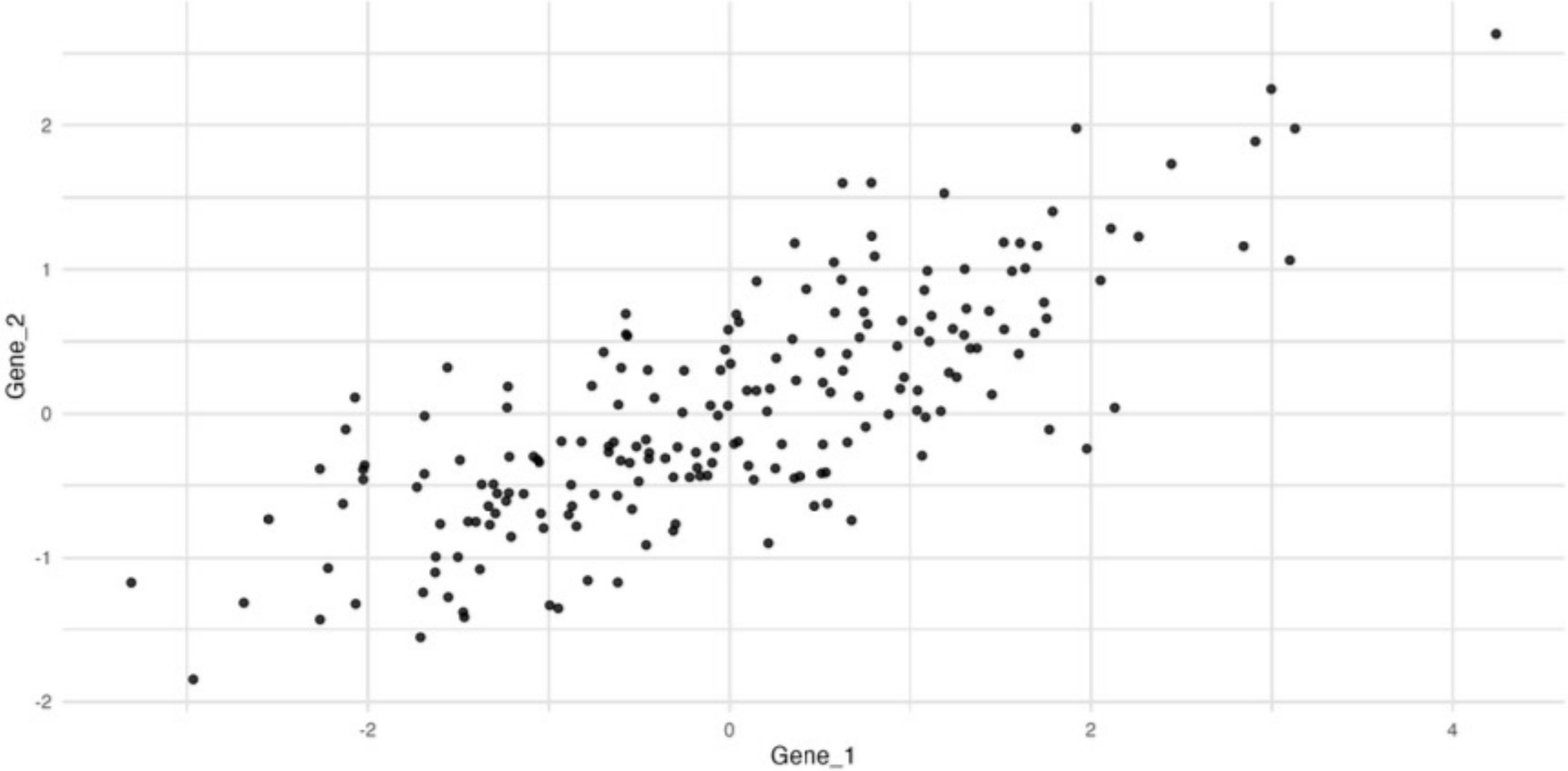
Principal Component Analysis

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



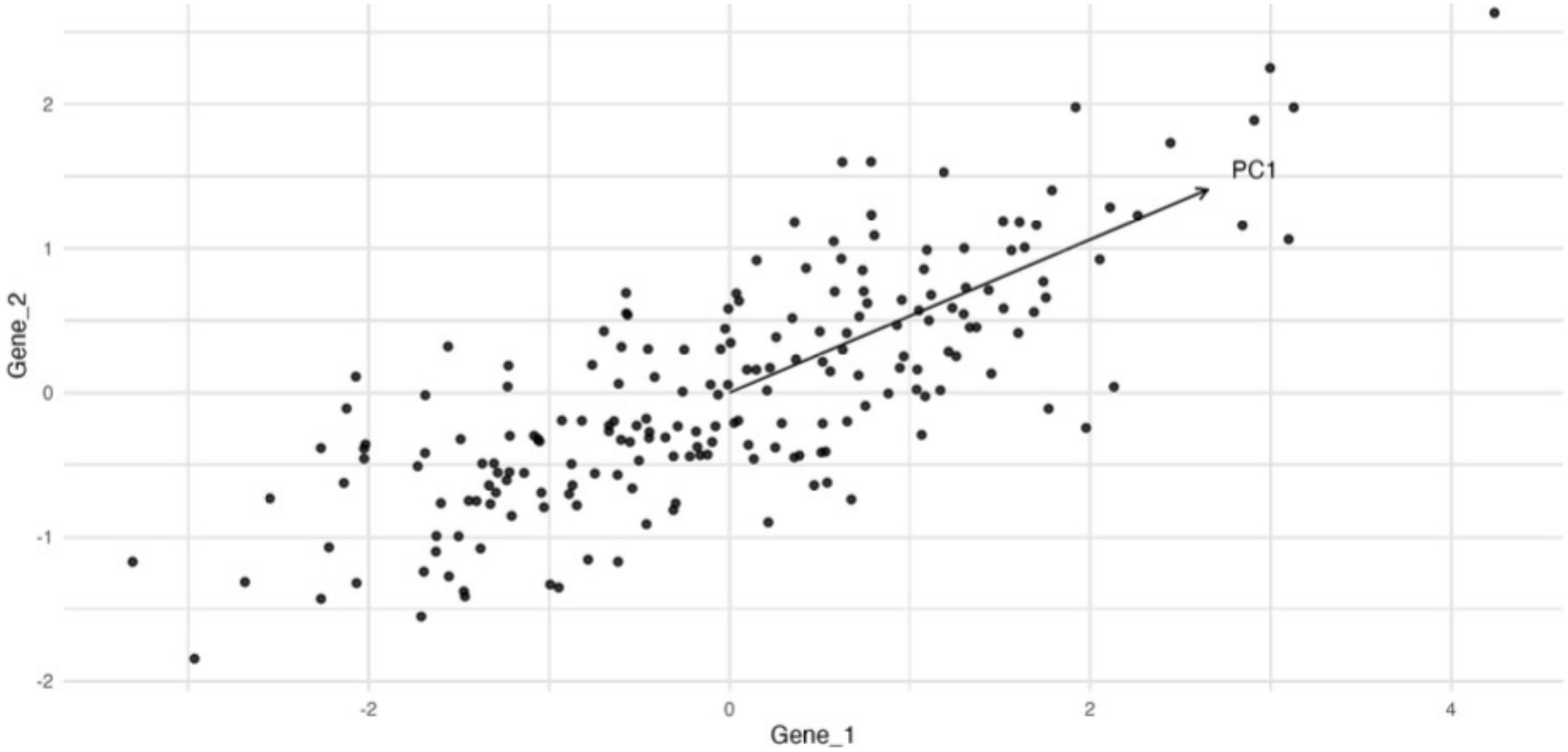
Principal Component Analysis

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



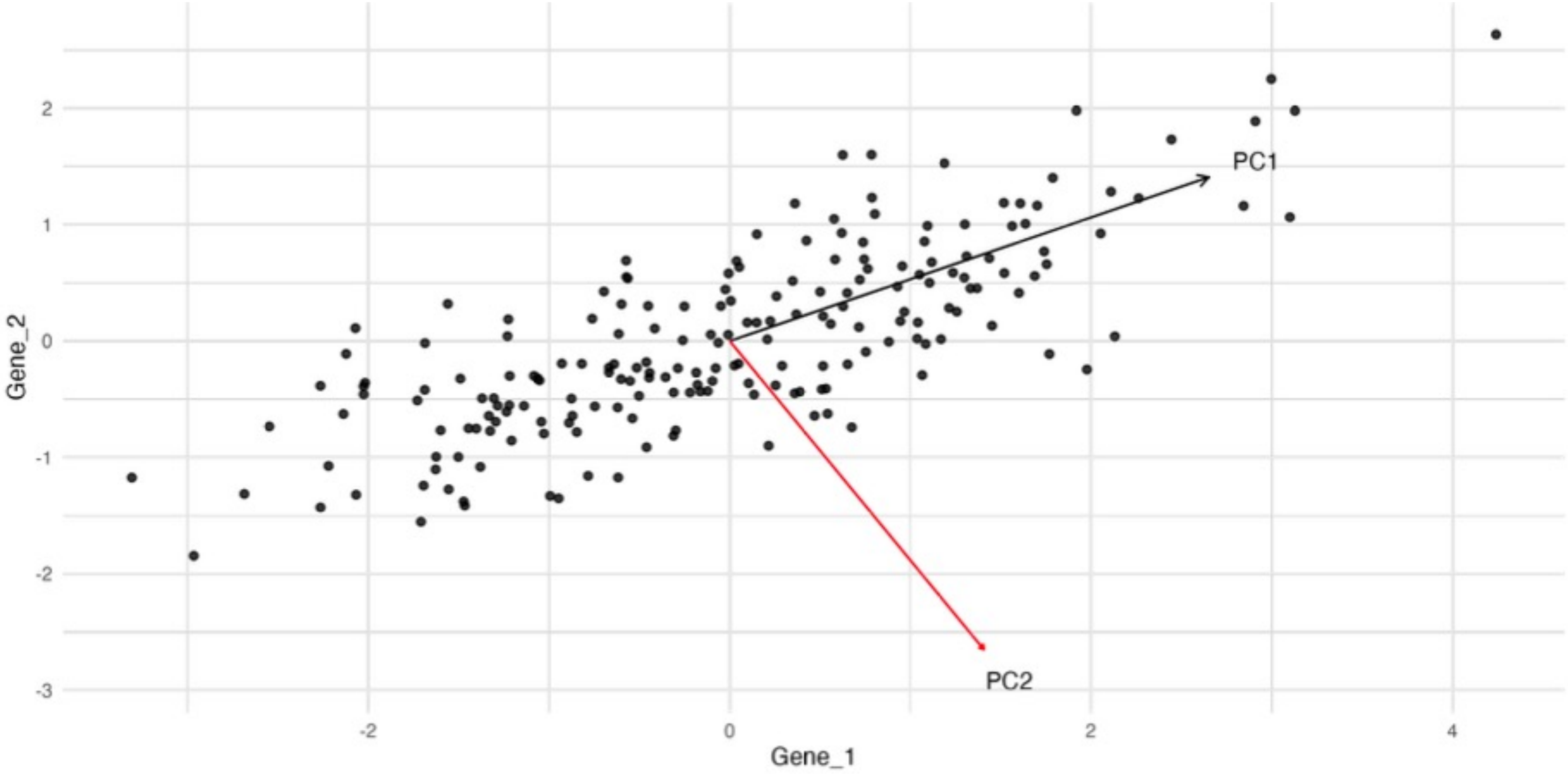
Principal Component Analysis

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



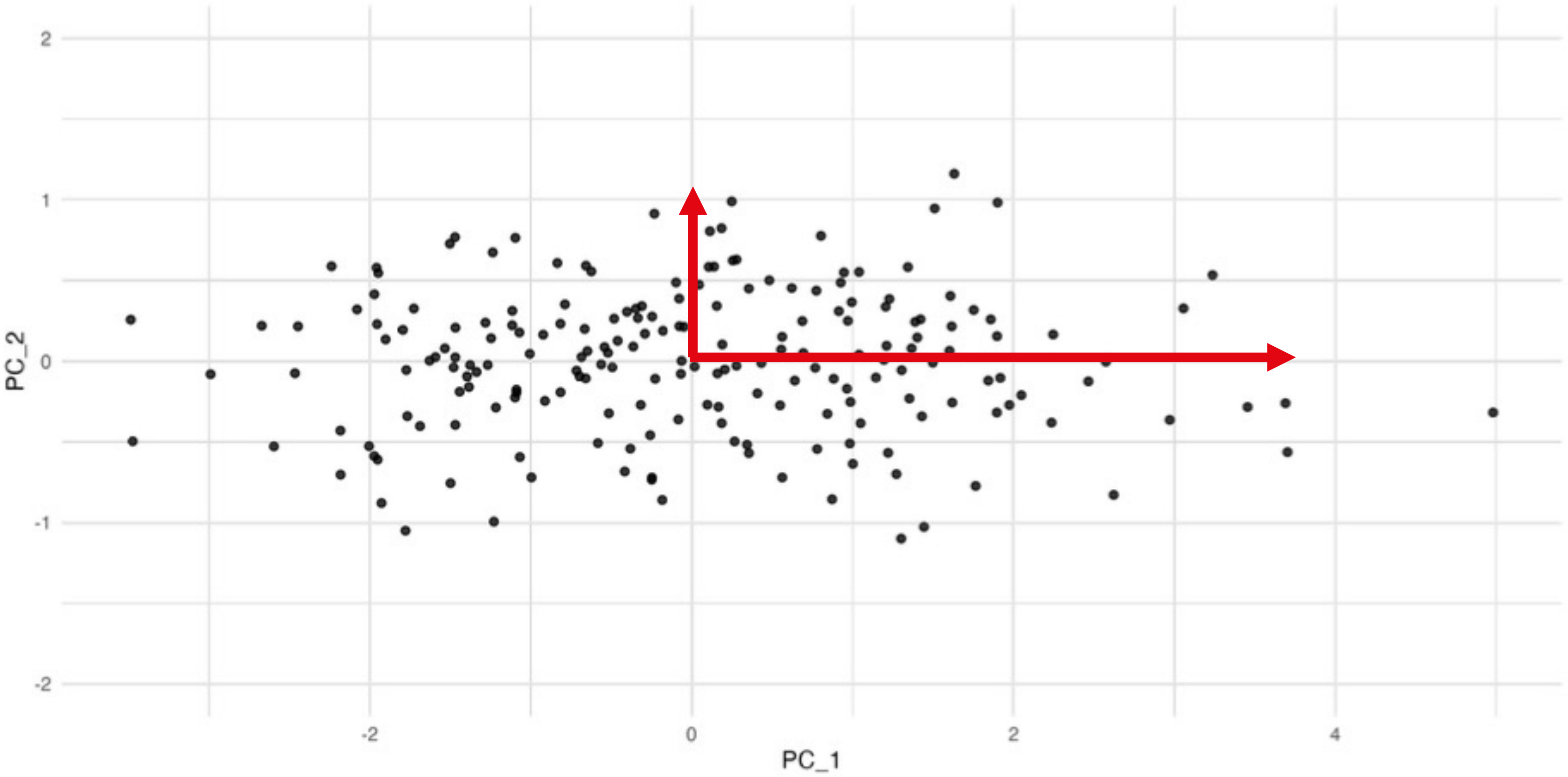
Principal Component Analysis

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



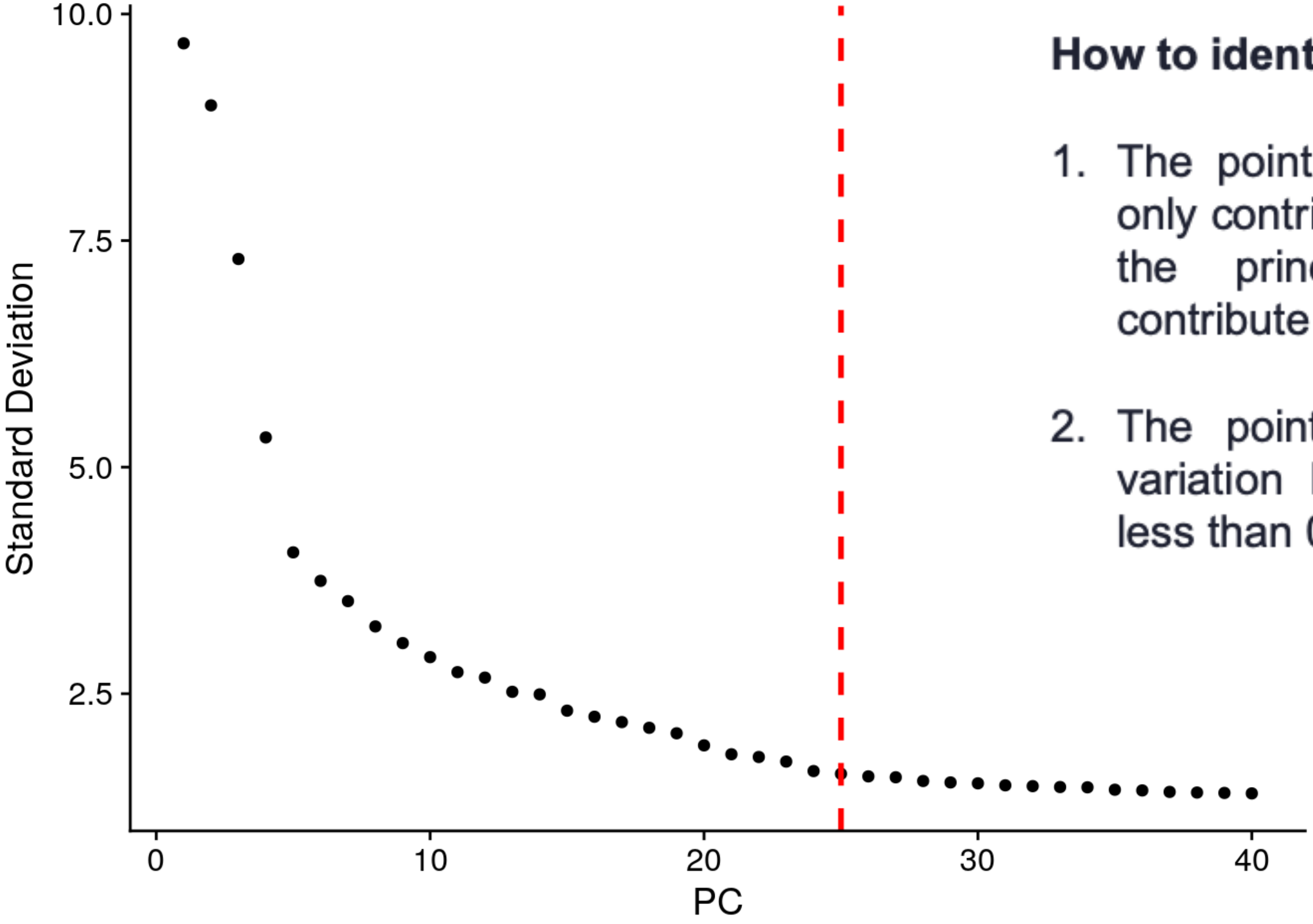
Principal Component Analysis

New axis that are linear combination of the original axes



Principal Component Analysis

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%.

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, another popular method, UMAP is faster and scales better with large datasets, making it ideal for single-cell datasets

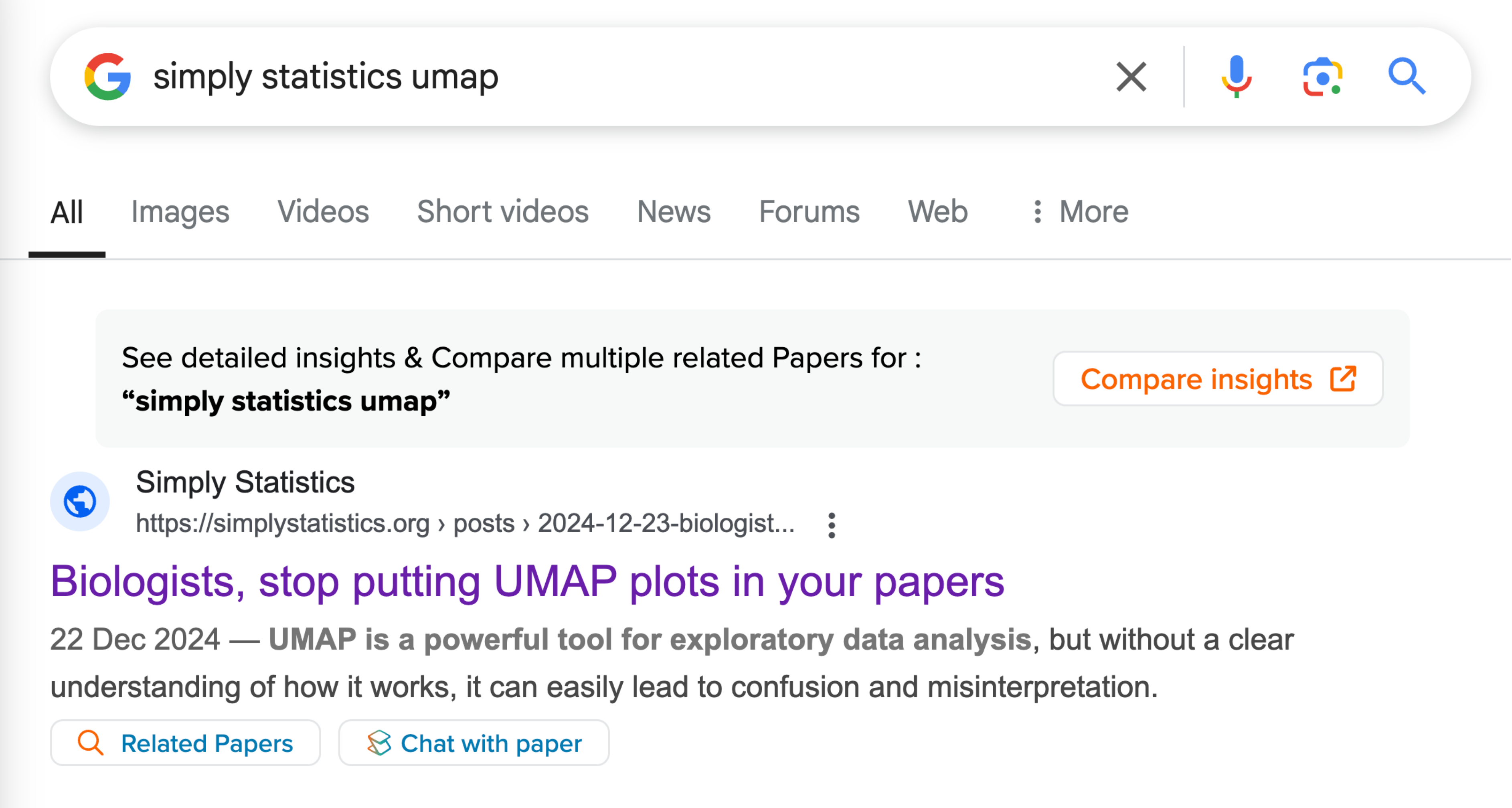
Why UMAP is performed after PCA?

PCA acts as a filter:

1. PCA helps **reduce noise** by capturing the most **informative features** (principal components)
2. UMAP struggles with **very high-dimensional data**
3. 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
4. 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**.*

Exercise (10 mins): when and when not to use UMAP?



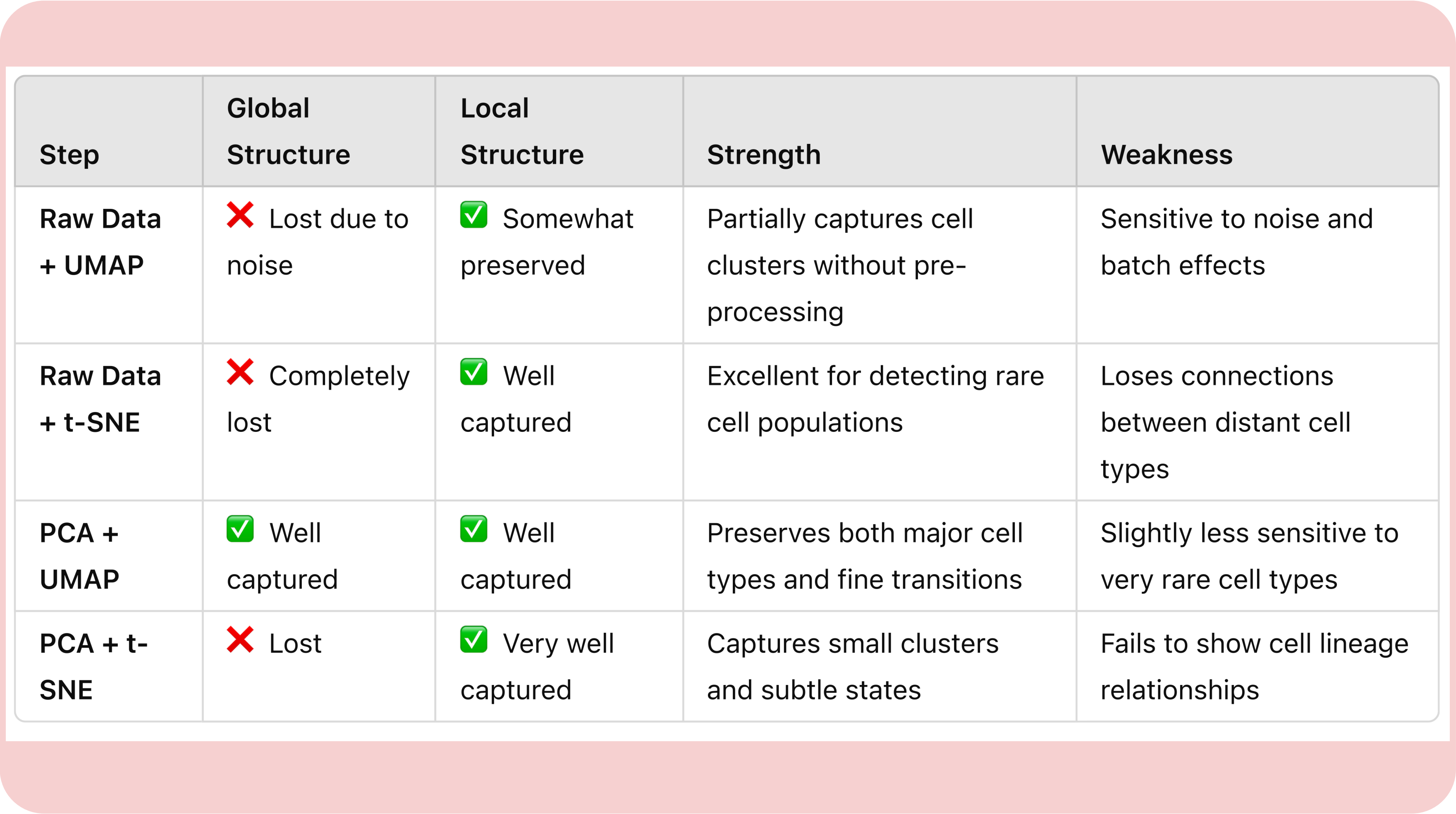
simply statistics umap

See detailed insights & Compare multiple related Papers for :
“simply statistics umap”

Simply Statistics
https://simplystatistics.org › posts › 2024-12-23-biologist...

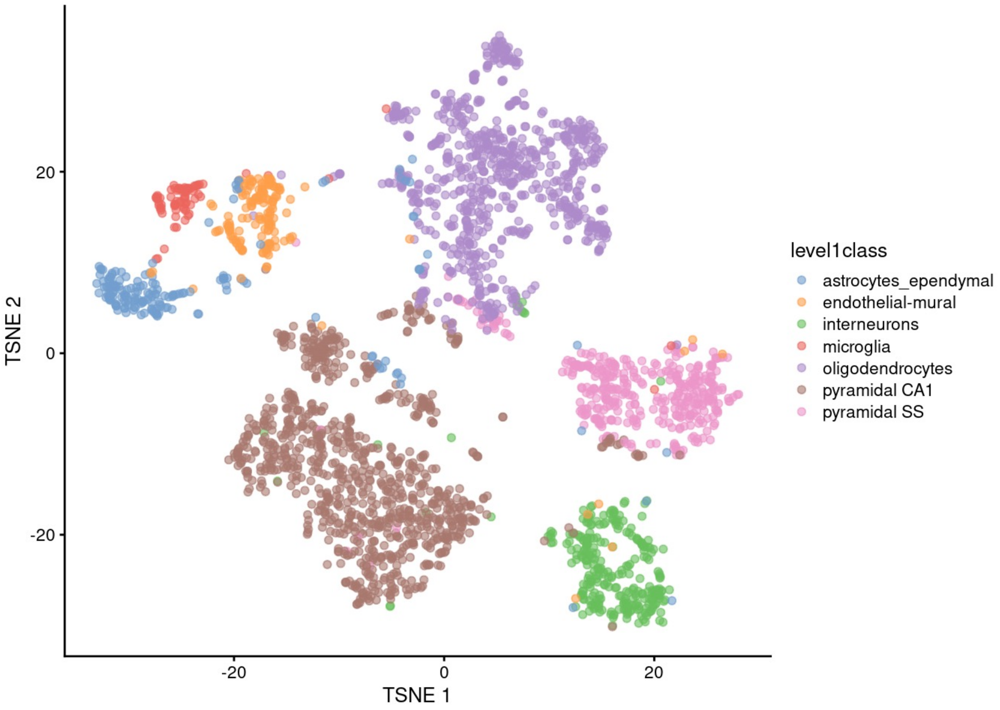
Biologists, stop putting UMAP plots in your papers

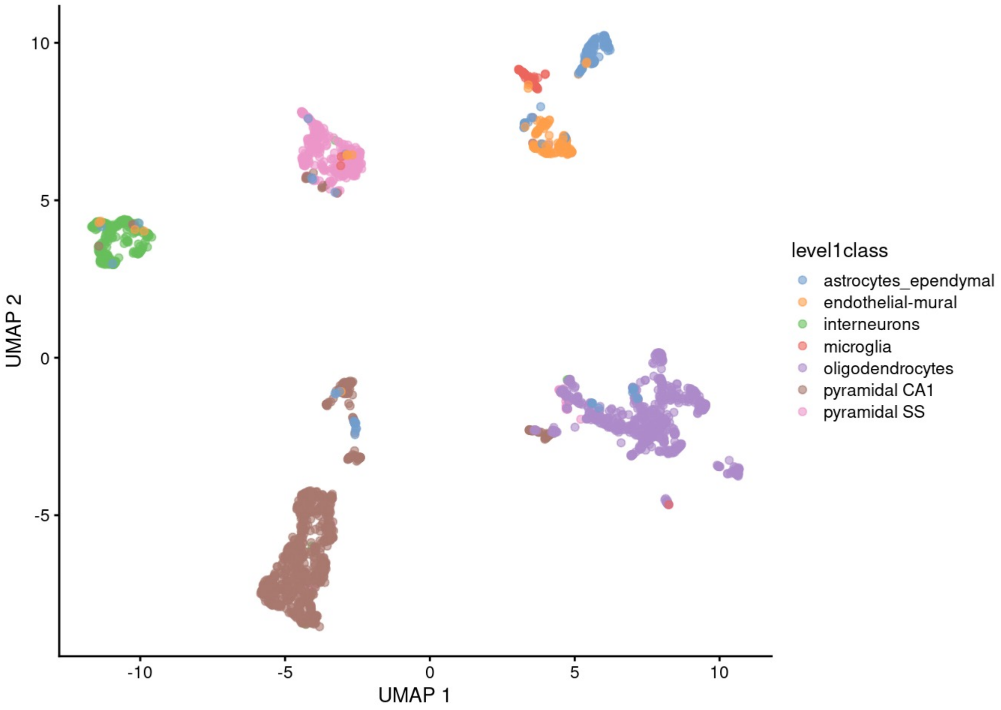
22 Dec 2024 — **UMAP is a powerful tool for exploratory data analysis**, but without a clear understanding of how it works, it can easily lead to confusion and misinterpretation.



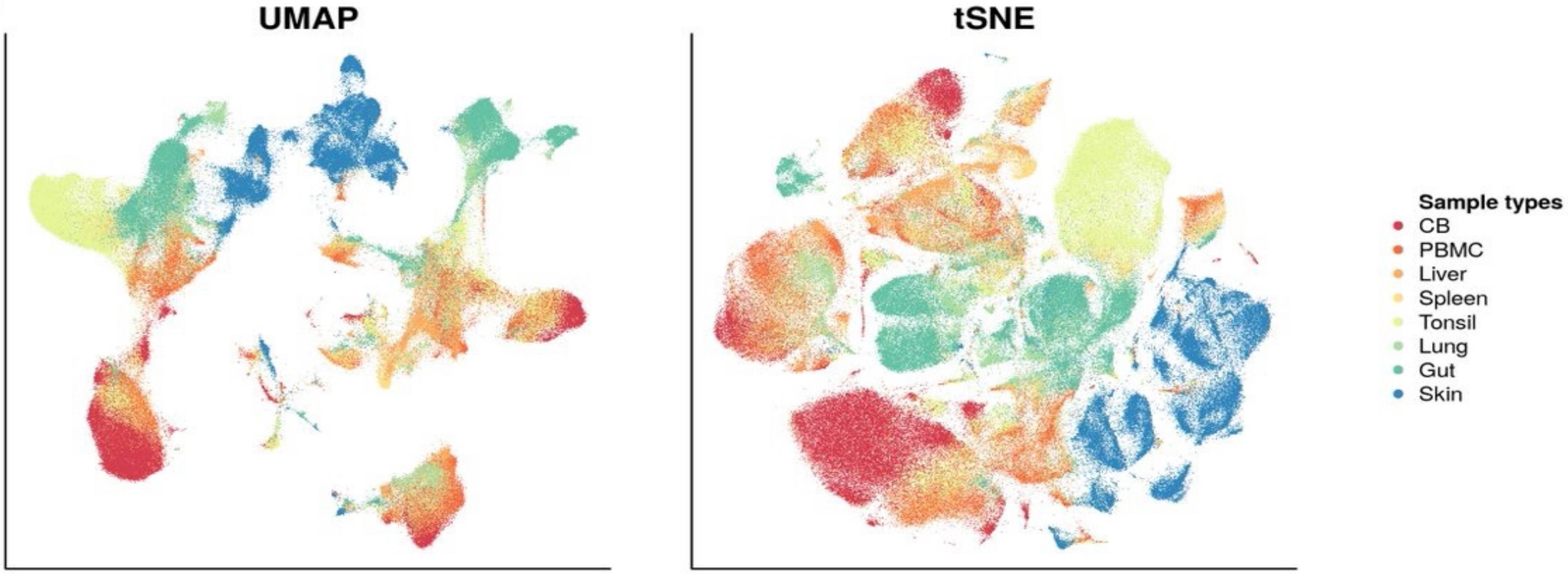
Step	Global Structure	Local Structure	Strength	Weakness
Raw Data + UMAP	❌ Lost due to noise	✅ Somewhat preserved	Partially captures cell clusters without pre-processing	Sensitive to noise and batch effects
Raw Data + t-SNE	❌ Completely lost	✅ Well captured	Excellent for detecting rare cell populations	Loses connections between distant cell types
PCA + UMAP	✅ Well captured	✅ Well captured	Preserves both major cell types and fine transitions	Slightly less sensitive to very rare cell types
PCA + t-SNE	❌ Lost	✅ Very well captured	Captures small clusters and subtle states	Fails to show cell lineage relationships

Global vs local structures





Global vs local structures



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.

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 dimensionality technique 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

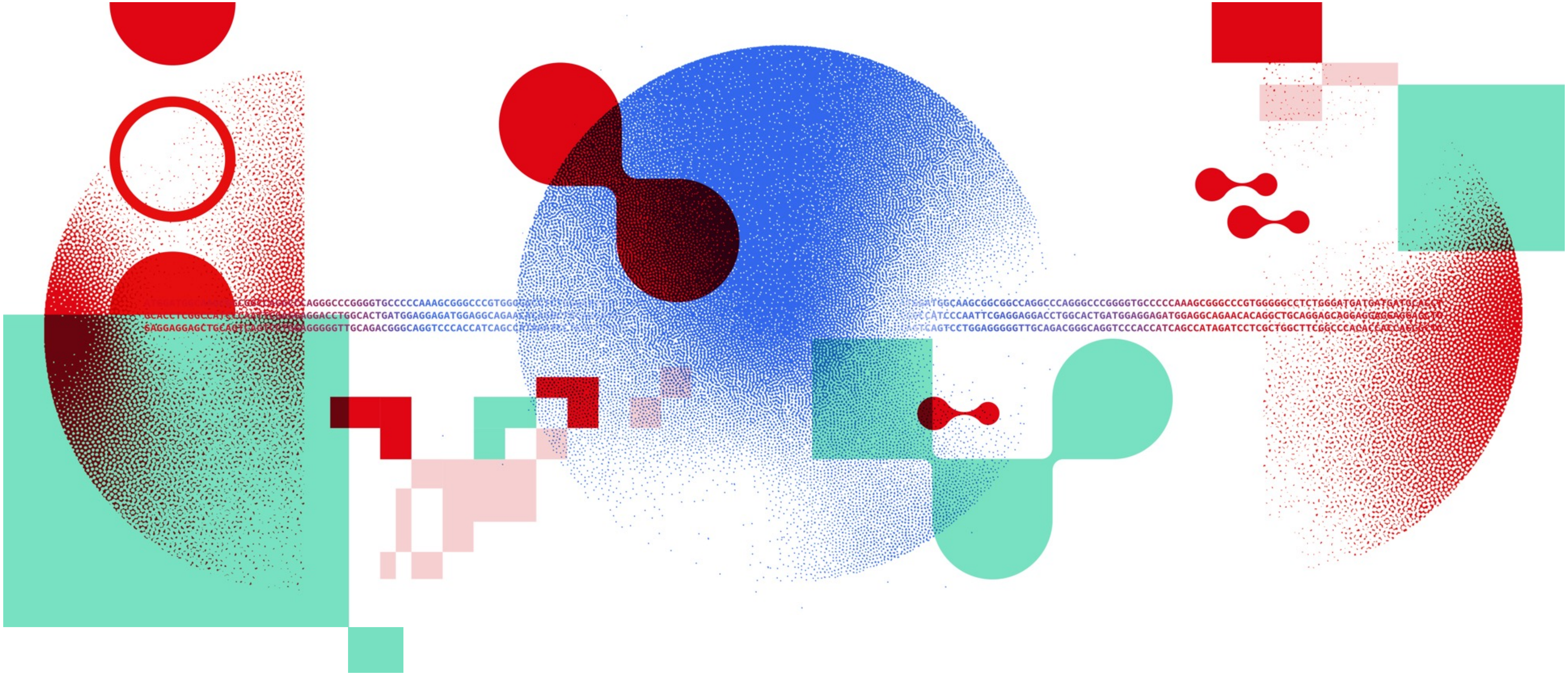
Principal Component Analysis (PCA): Identifies key directions in data, reduces dimensions

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



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss