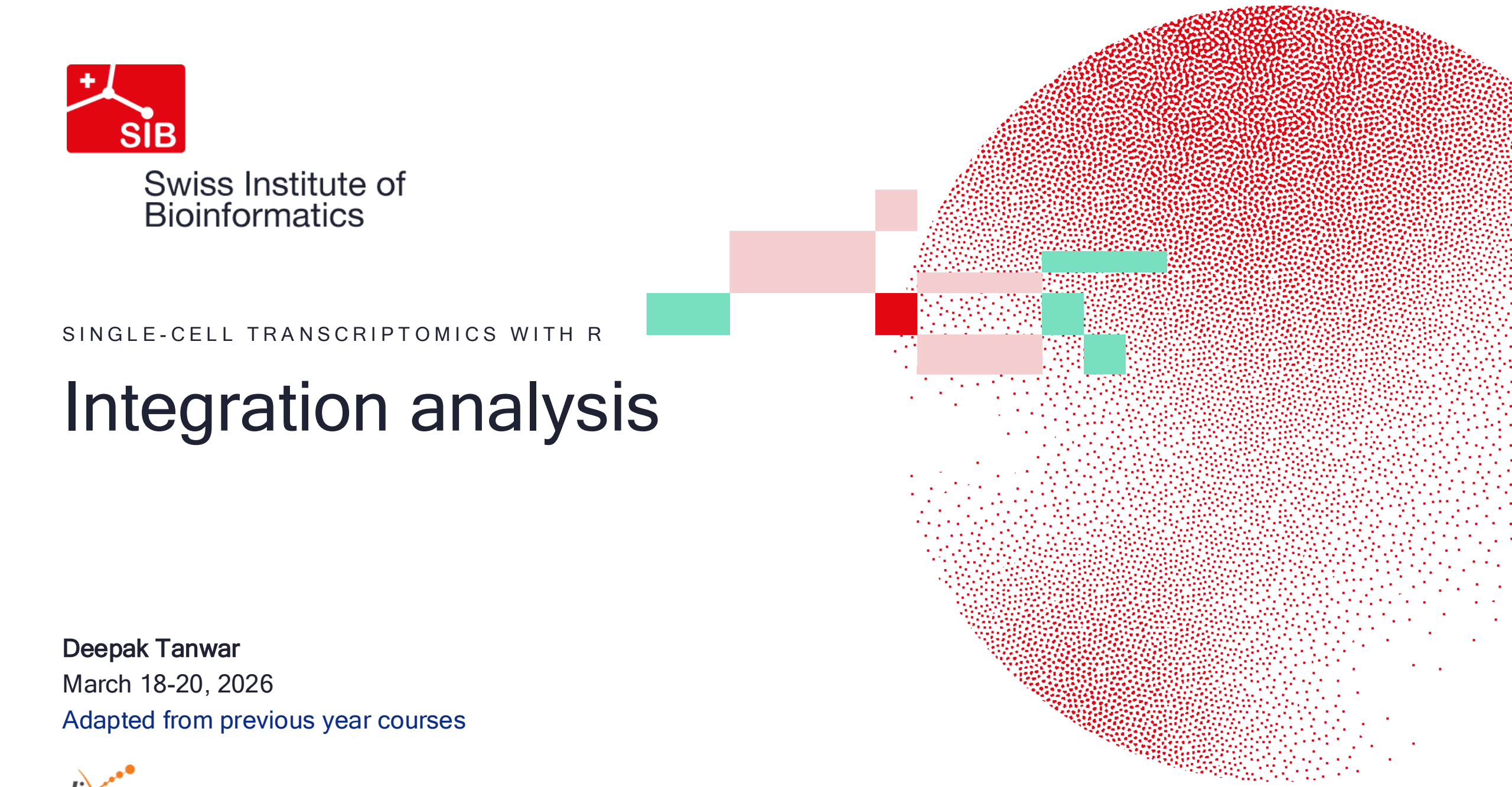


Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Integration analysis

Deepak Tanwar

March 18-20, 2026

Adapted from previous year courses

Learning objectives

Understand the importance of experimental design

Identify scenarios where integration is necessary for data analysis

Apply canonical correlation analysis (CCA) for integrating datasets

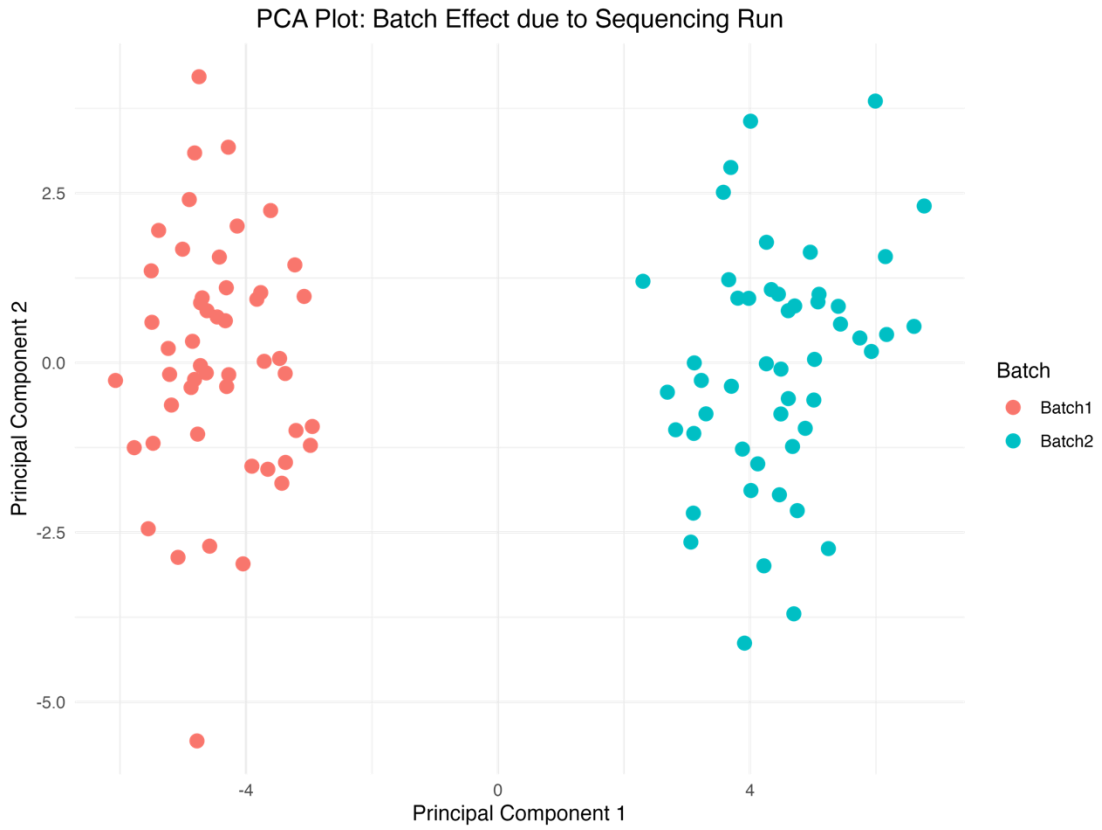
Unwanted Sources of Variation

Batch Effect is systematic techincal variations due to differences in:

a) cell isolation and handling protocols,

b) library preparation technology, and sequencing platforms

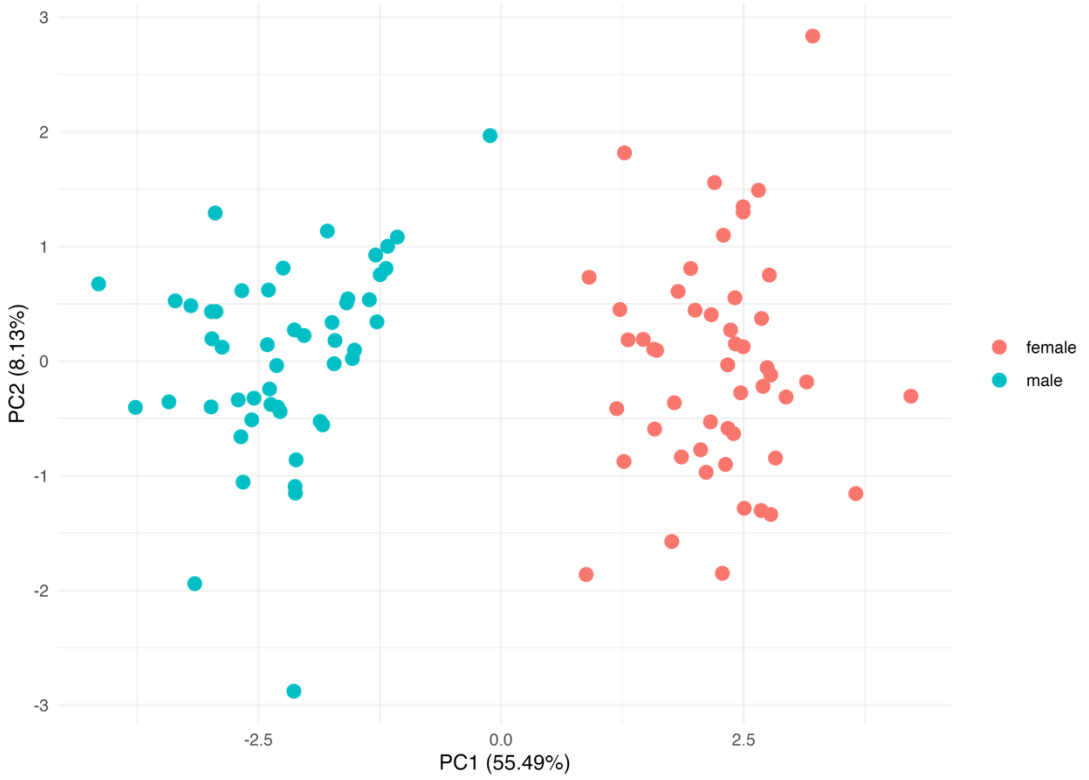
Batch effects can obscure true biological signals, making it difficult to compare datasets



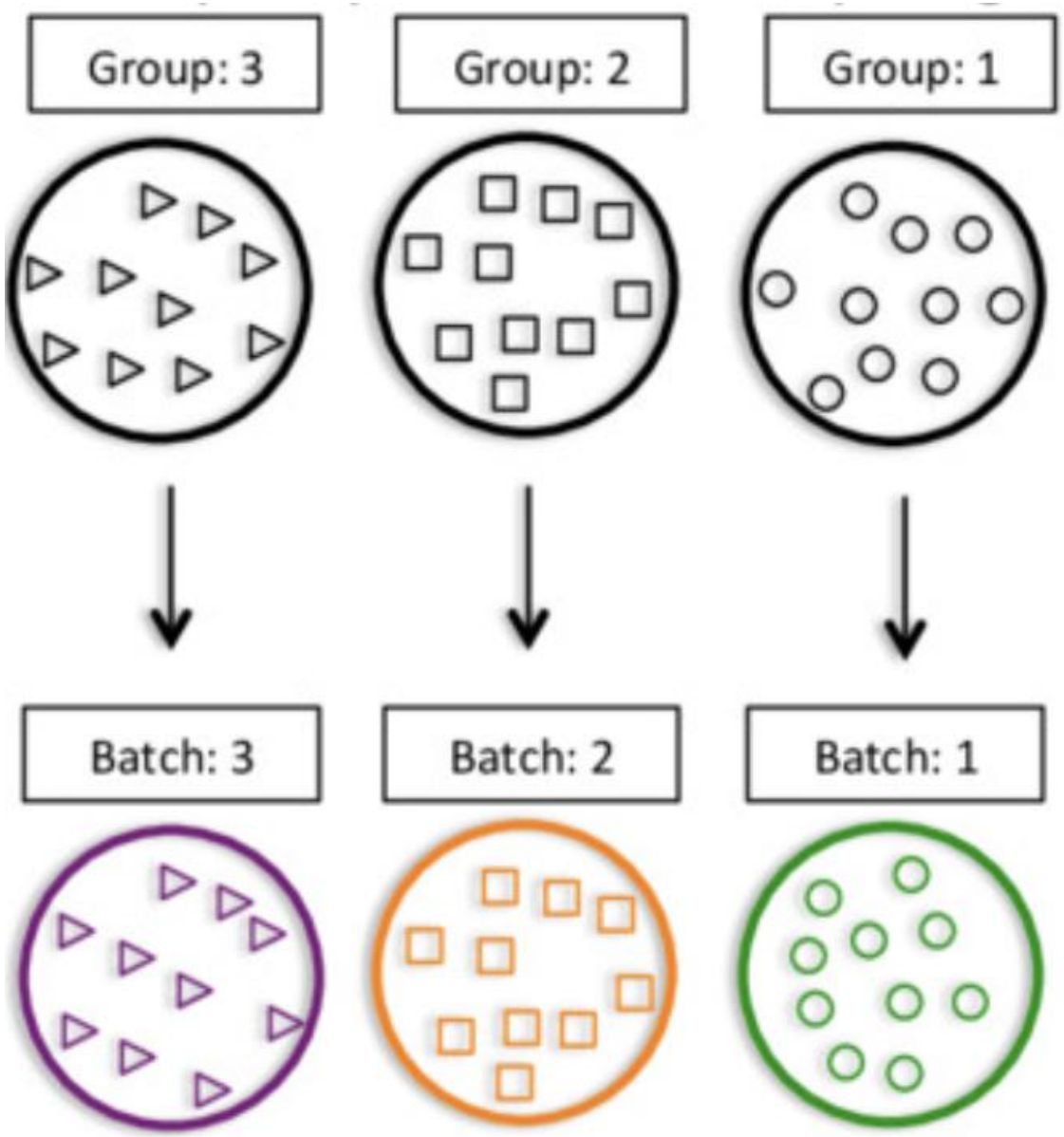
Unwanted Sources of Variation

Confounders are variables (e.g. Gender, Age) that could influnce gene expression

If they are not properly accounted for in the analysis they could potentially lead to misleading associations.



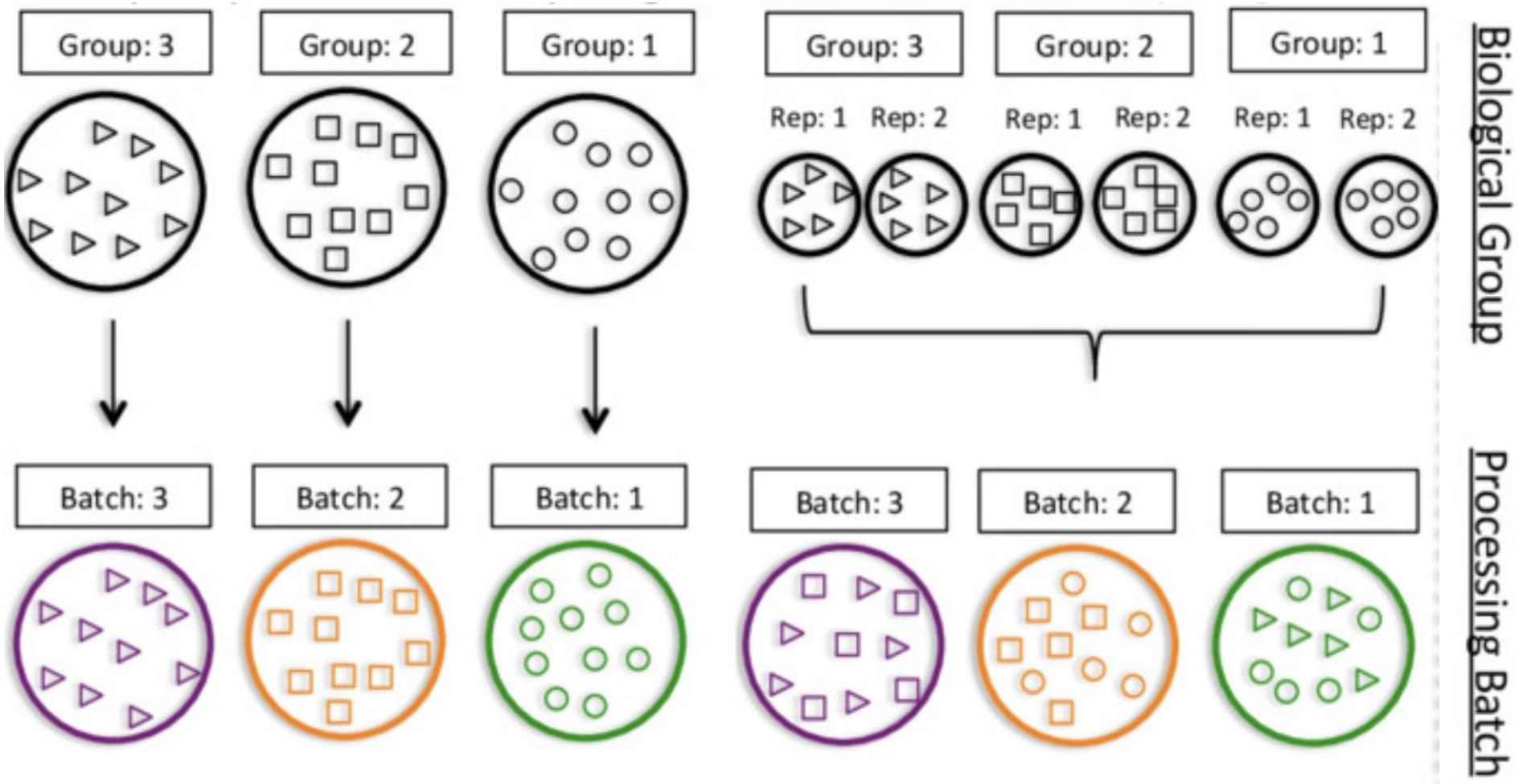
Experimental design matters



What changes would you make here to make the experimental design more optimal?

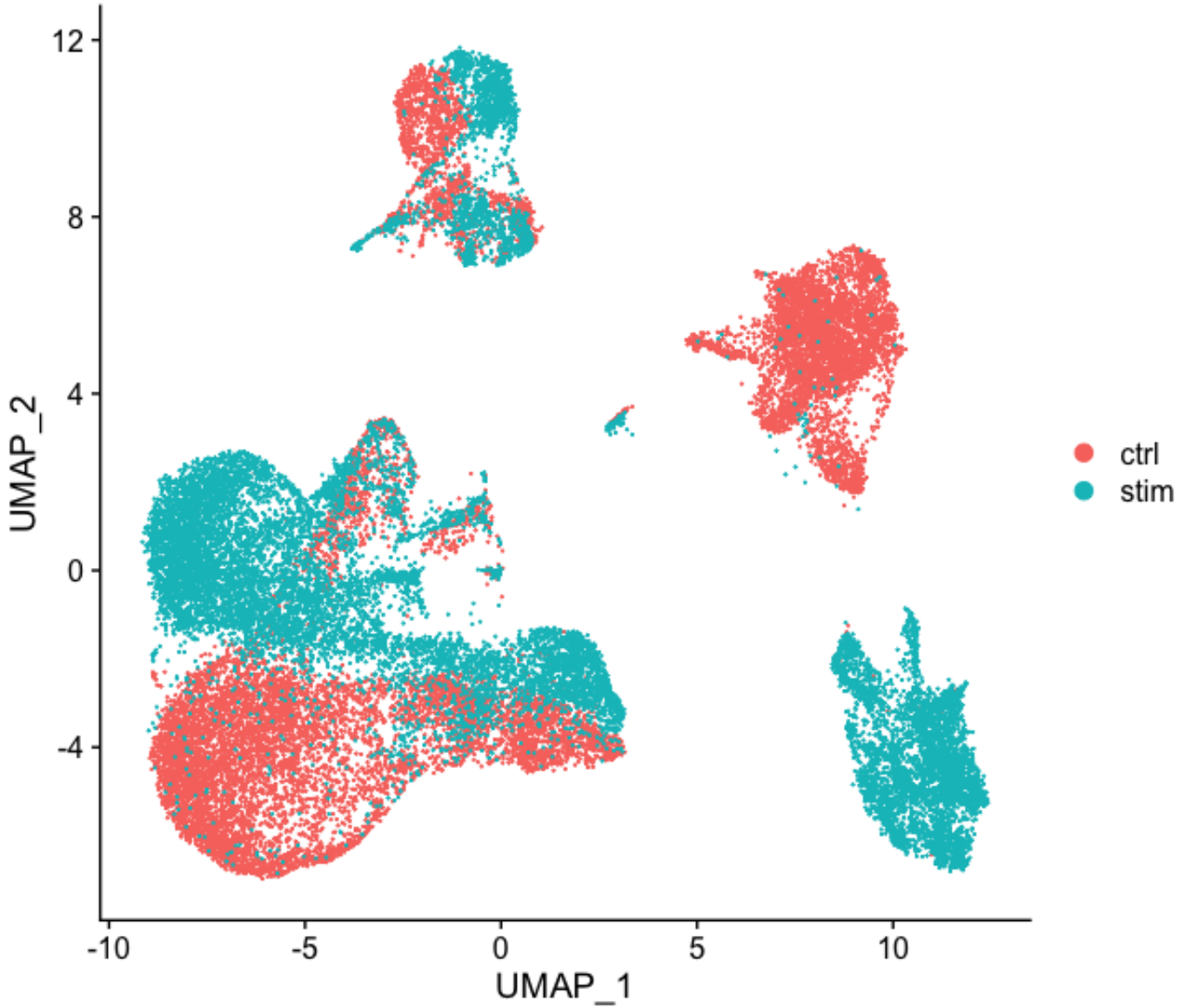
Hicks et al.

Experimental design matters

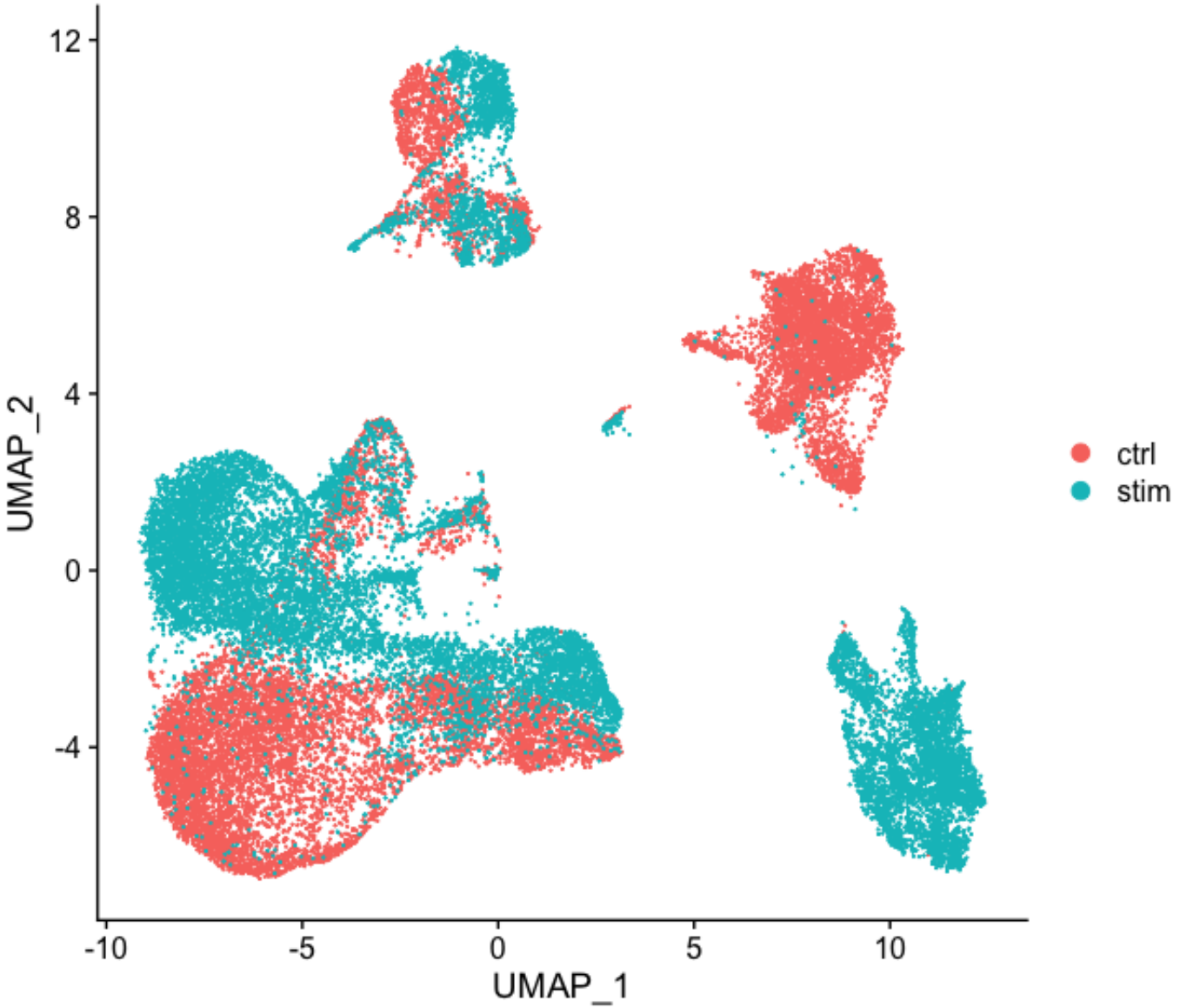


Hicks et al.

Exercise: Identify problem in this plot



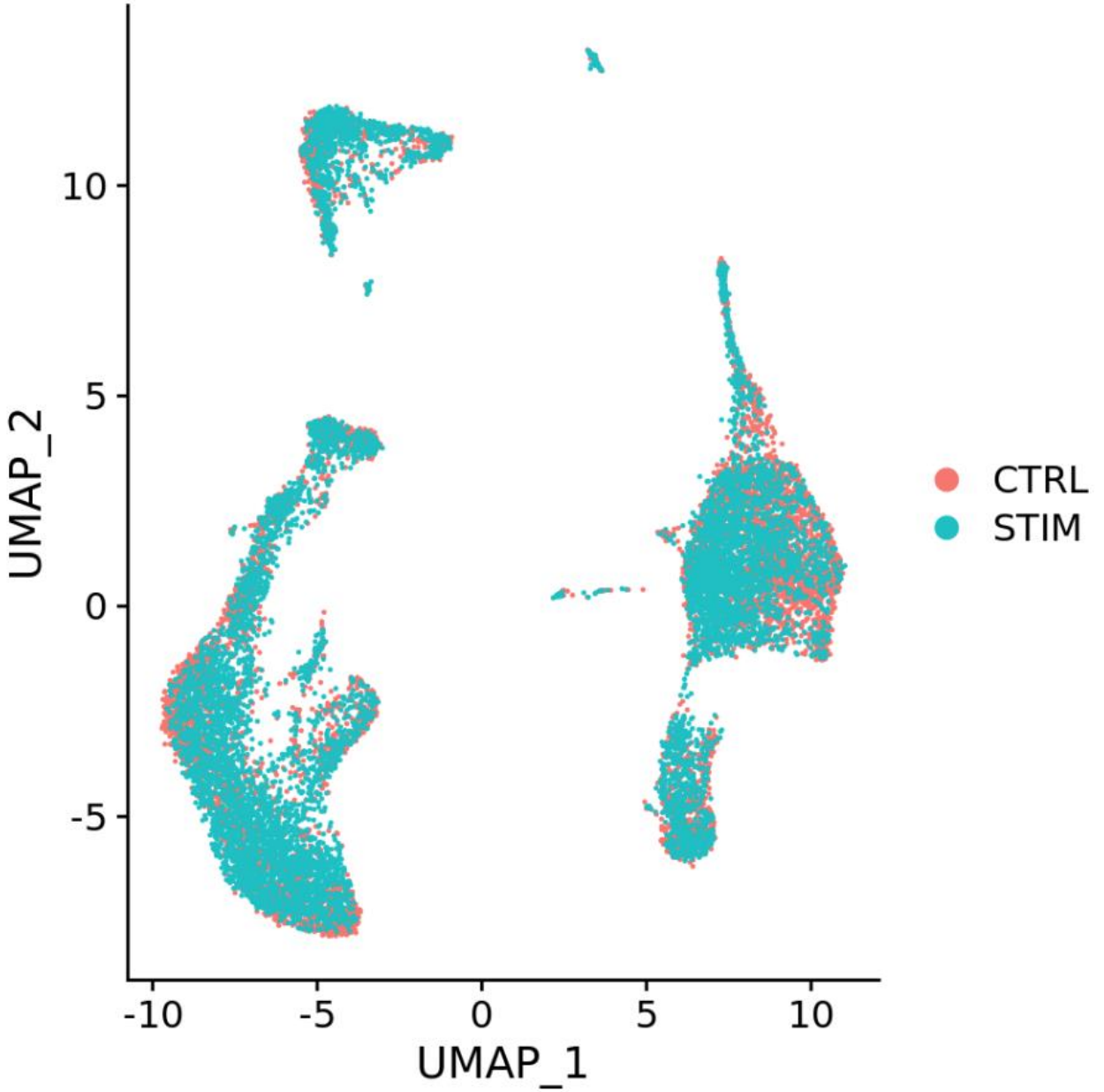
Exercise: Identify problem in this plot

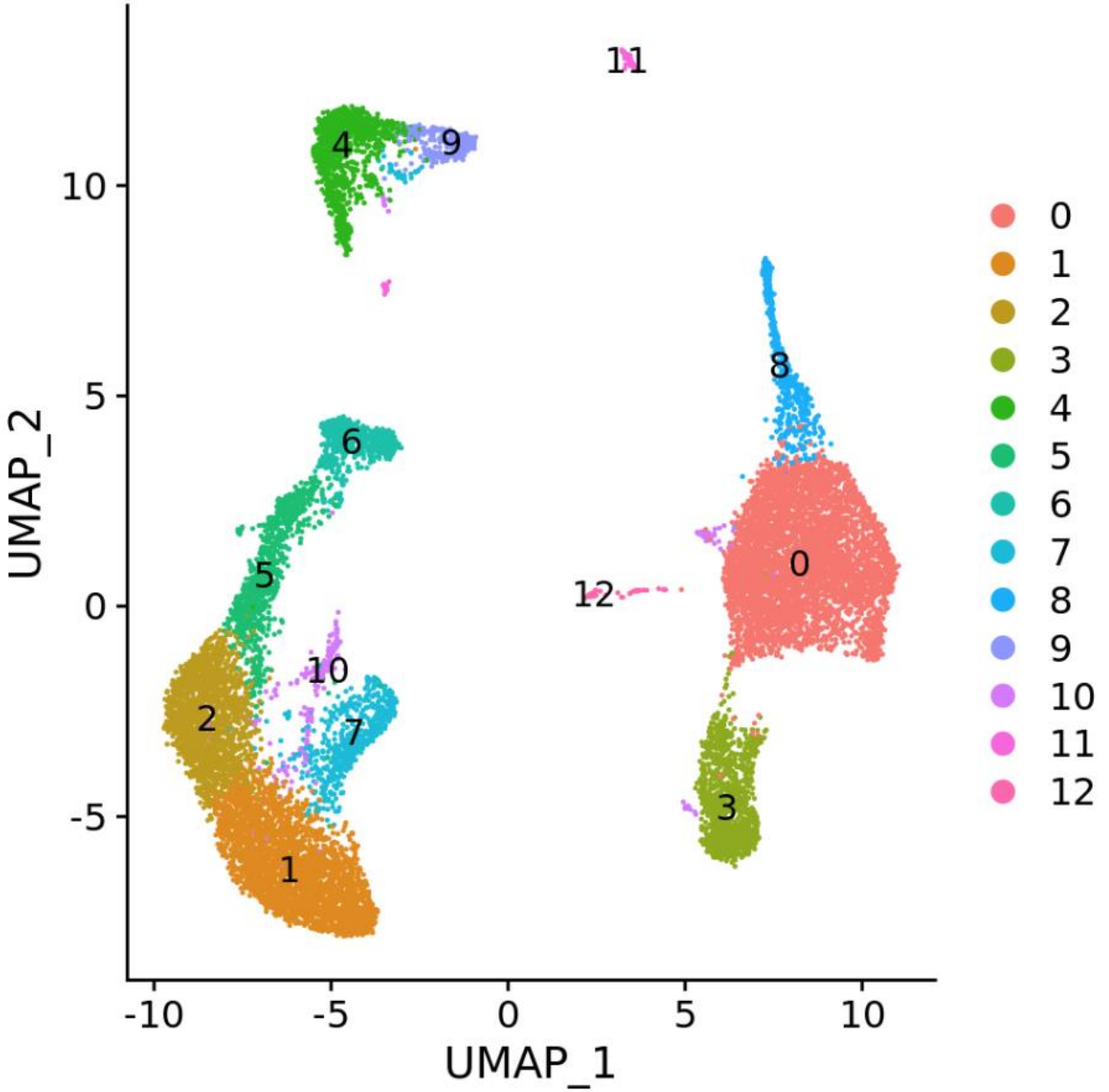


- **Explore the data:** do not just always perform integration because you think there might be differences

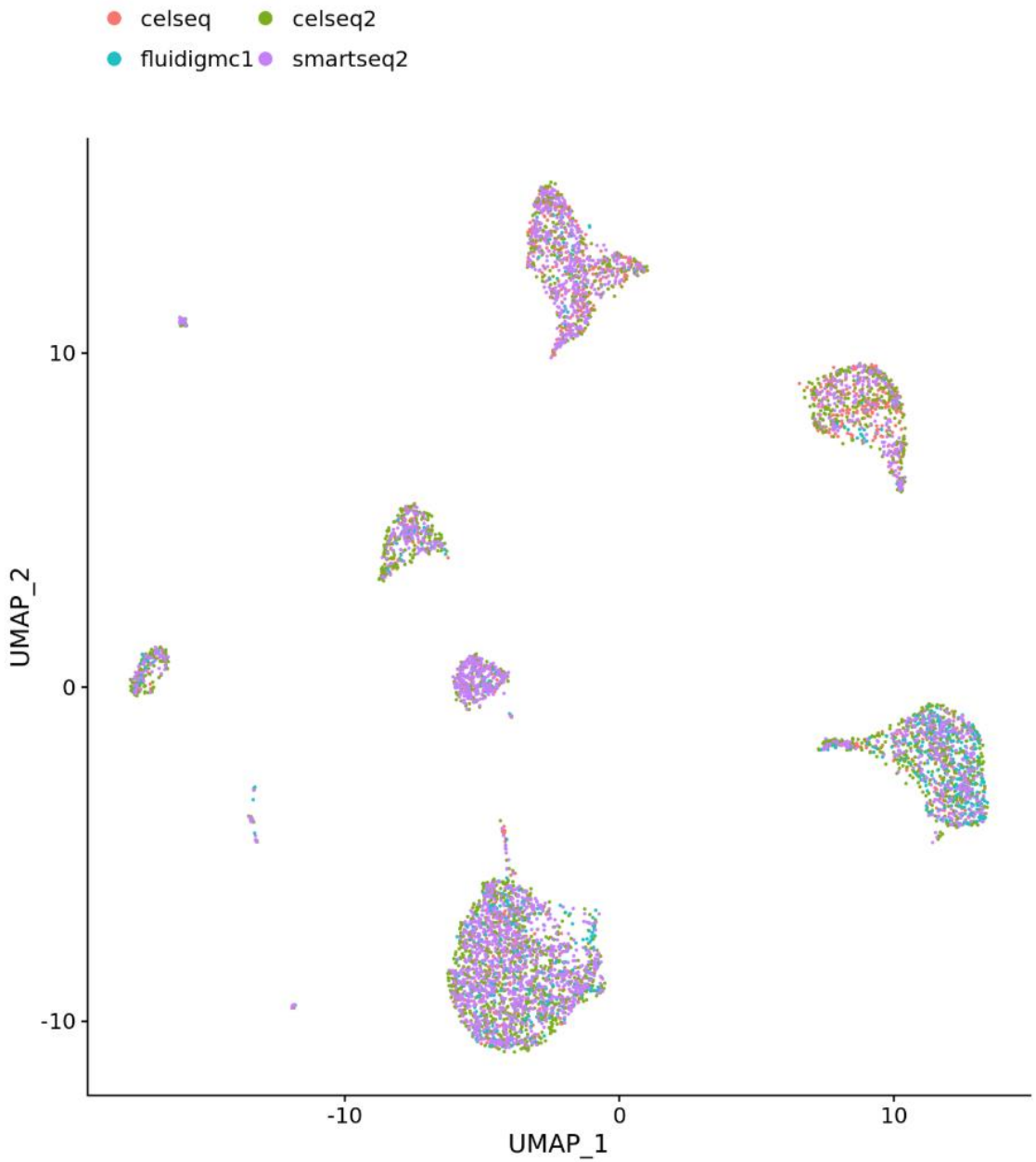
- If cells cluster by **sample, condition, batch, dataset, modality**, performing integration can help align cells across the groups to greatly improve the clustering and the downstream analyses.

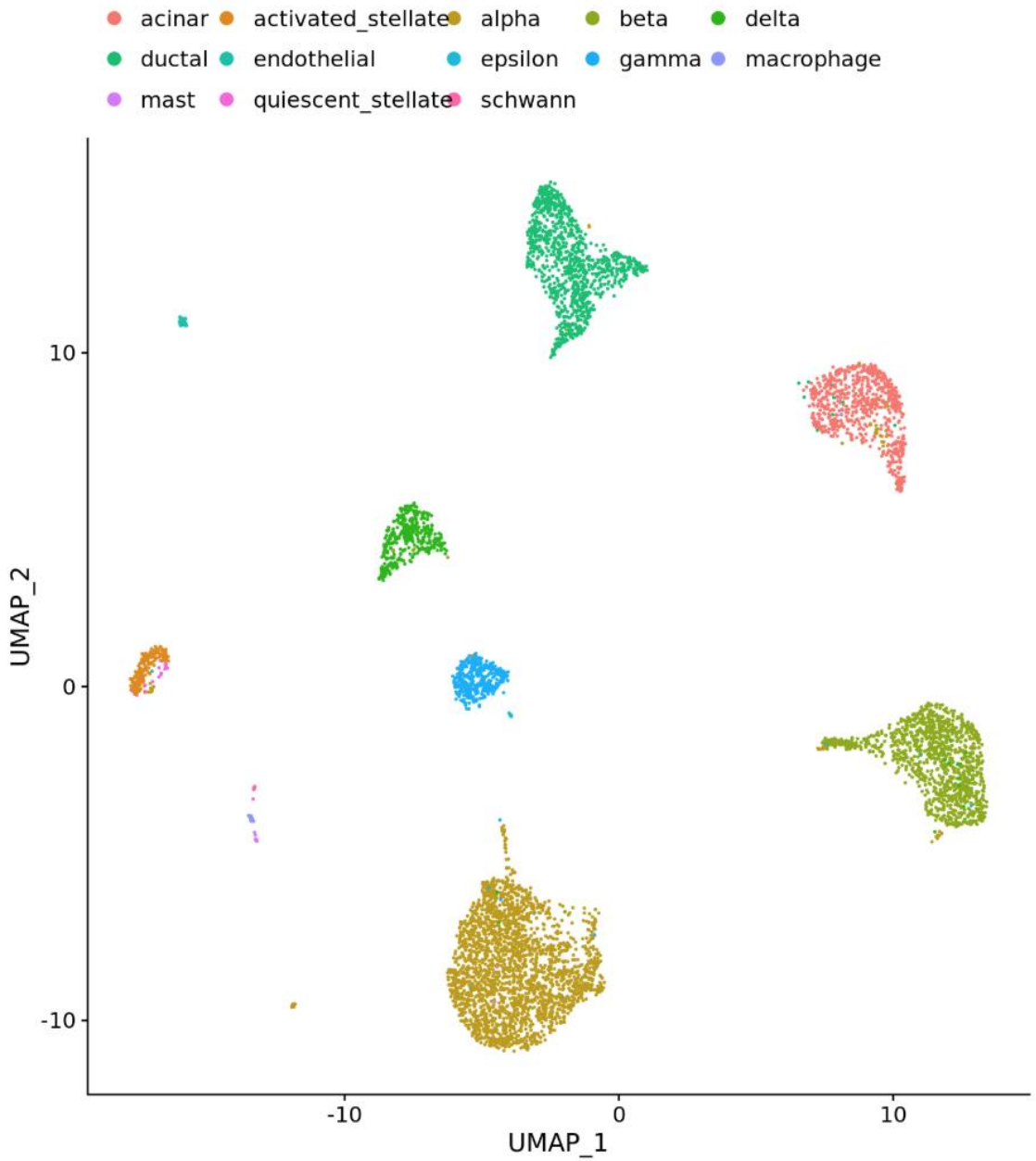
Example scenarios for integration: **conditions**



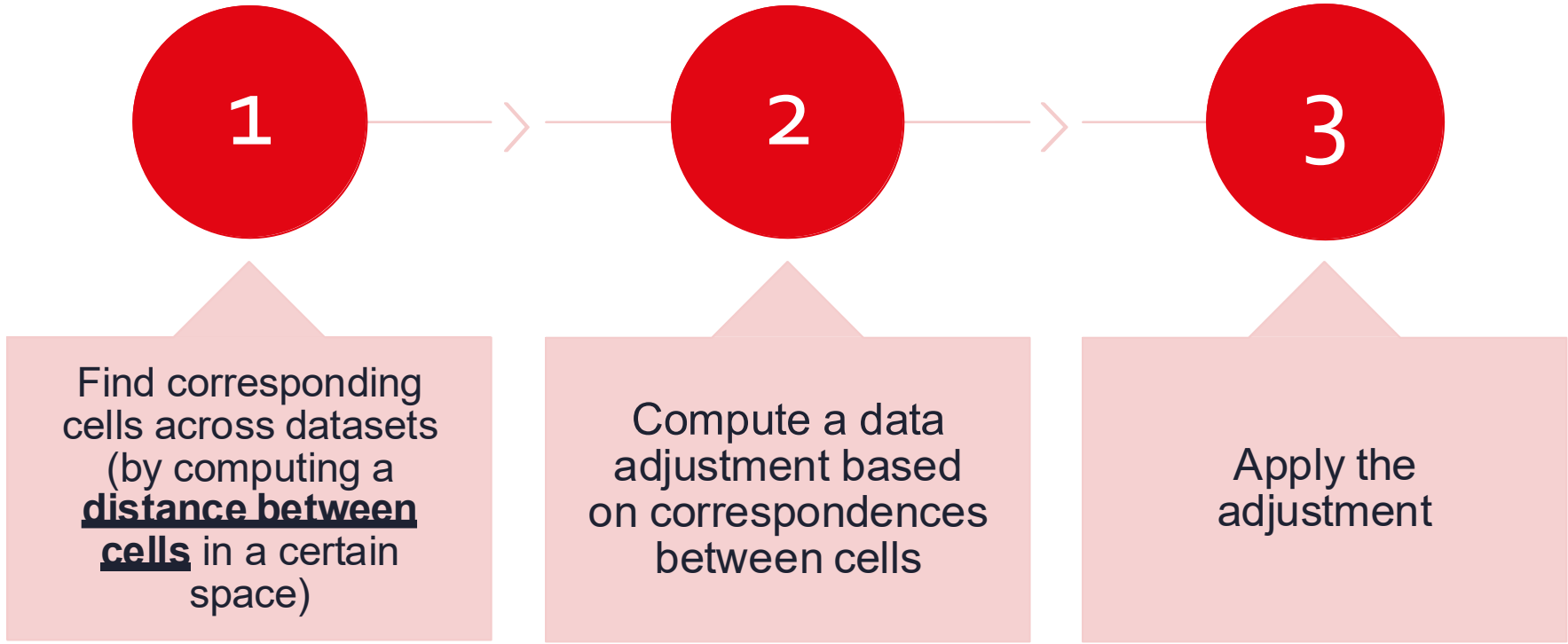


Example scenarios for integration: **datasets**



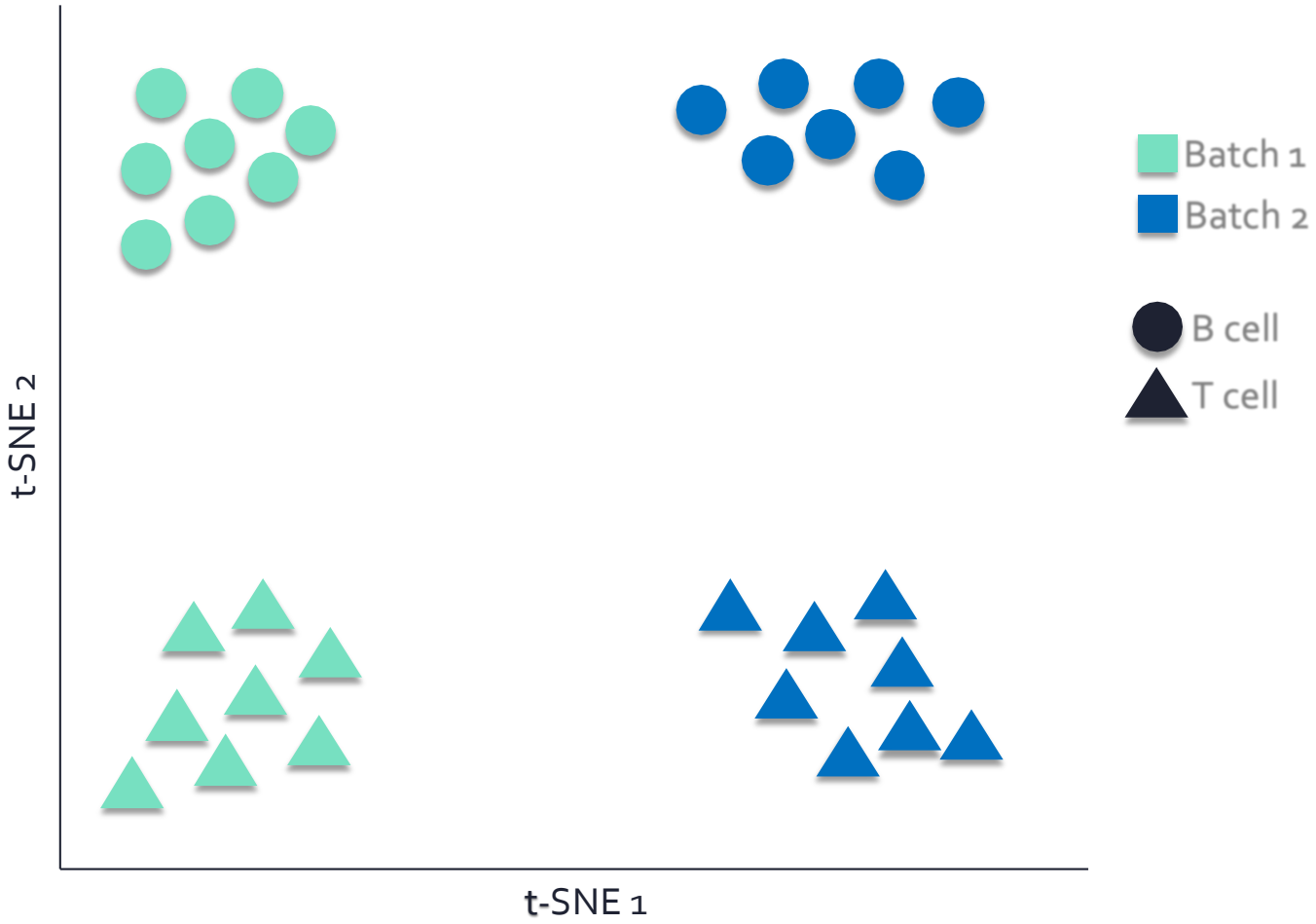


How to integrate the cells?

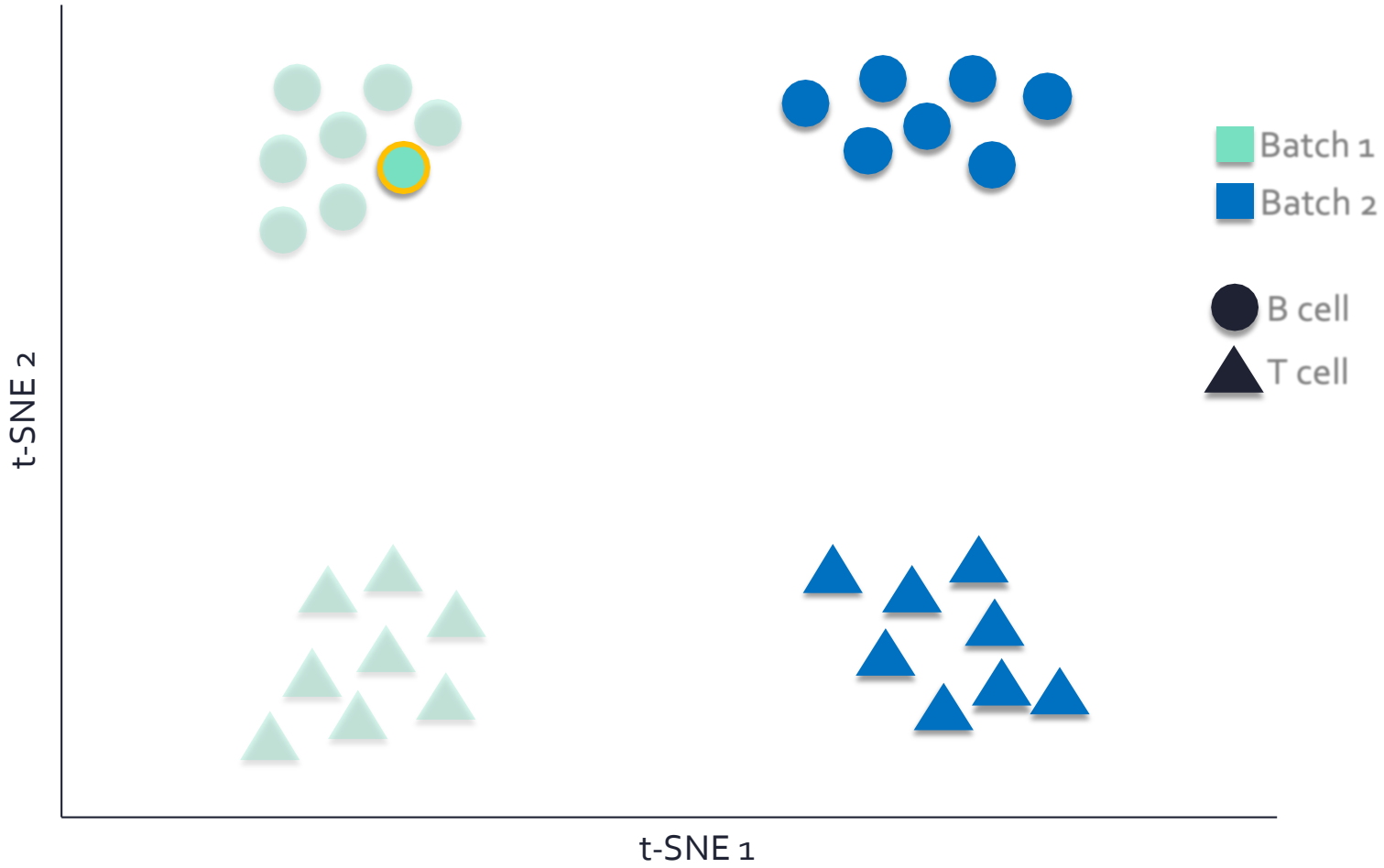


How to integrate the cells?

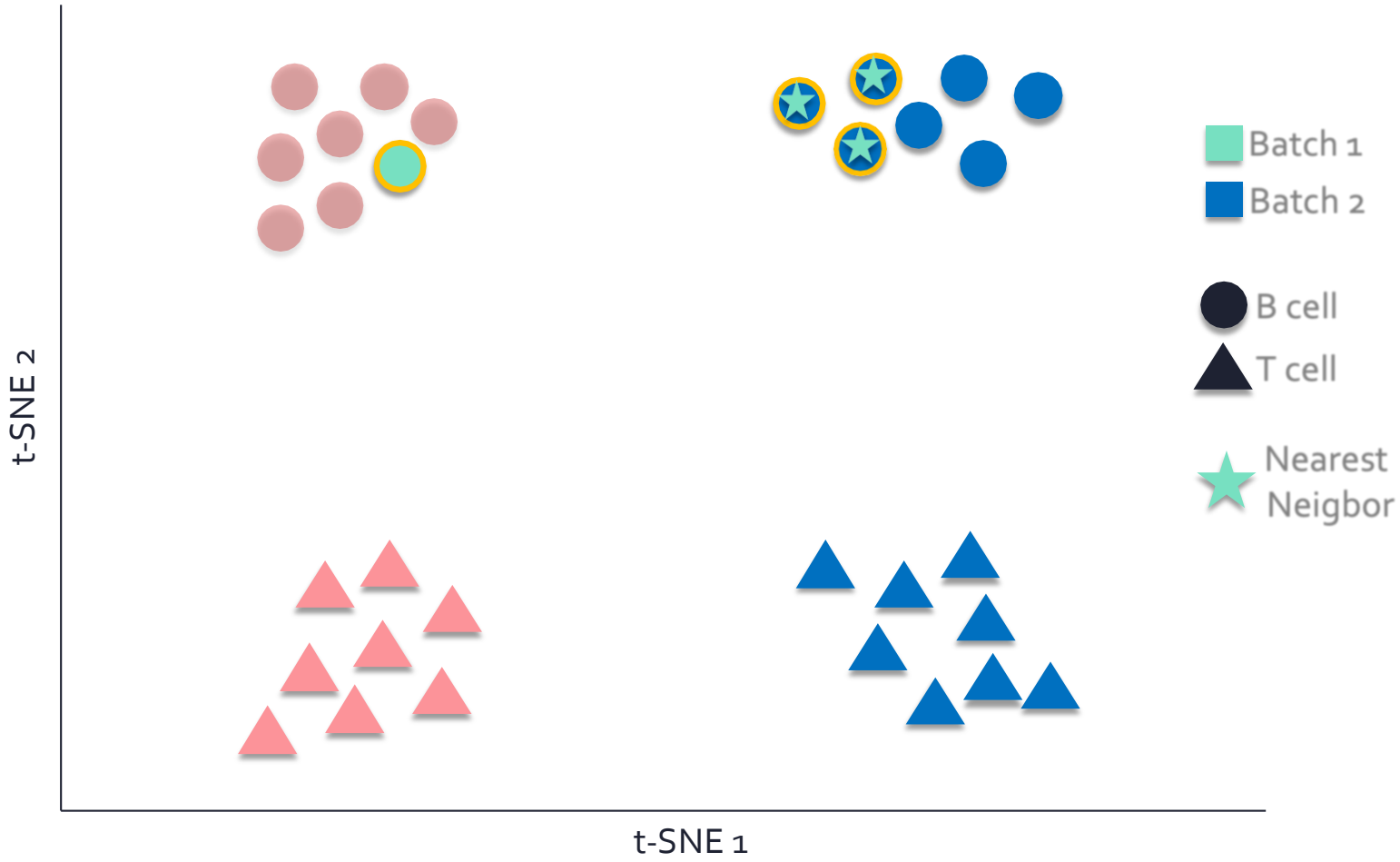
Example



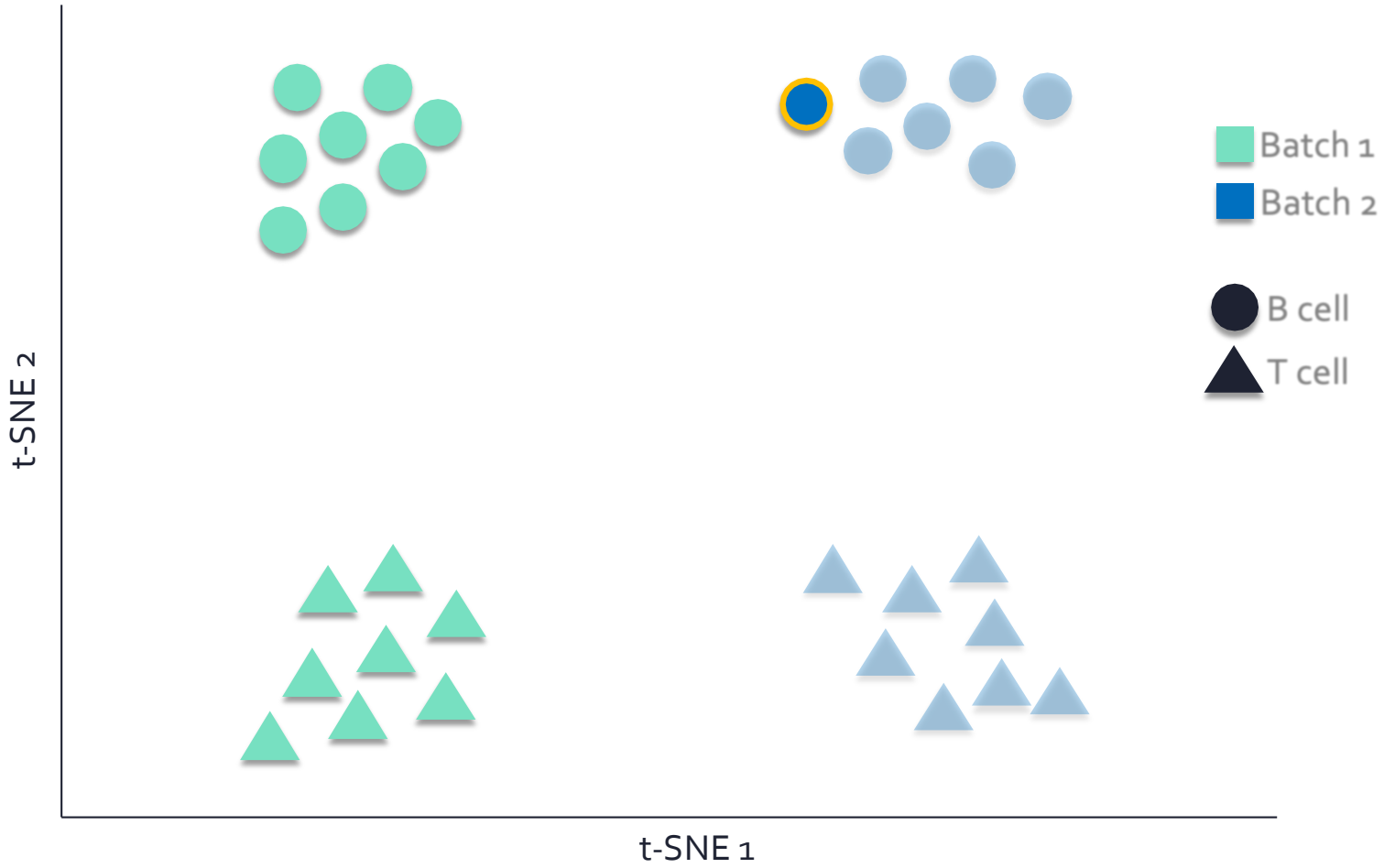
How to integrate the cells?



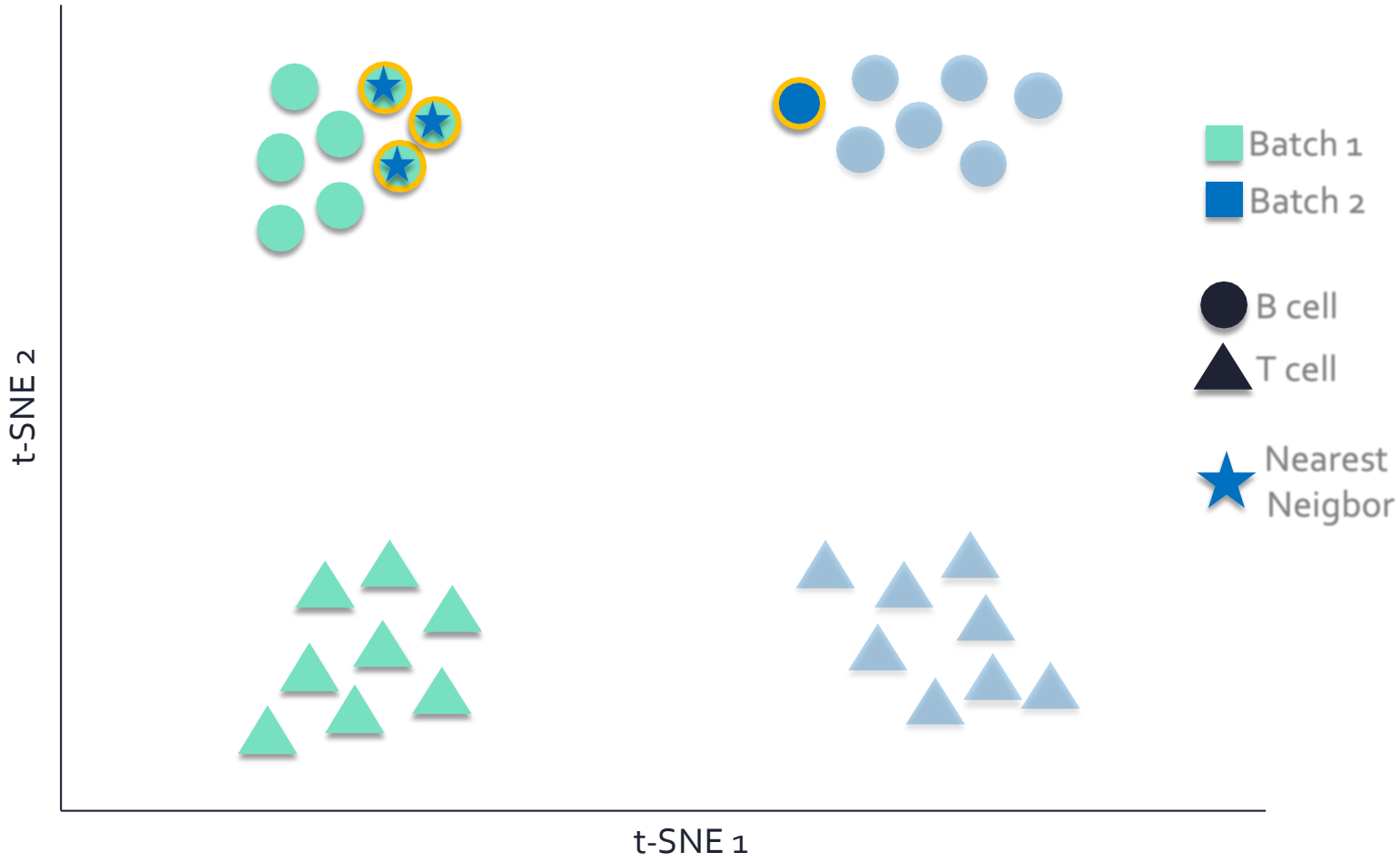
How to integrate the cells?



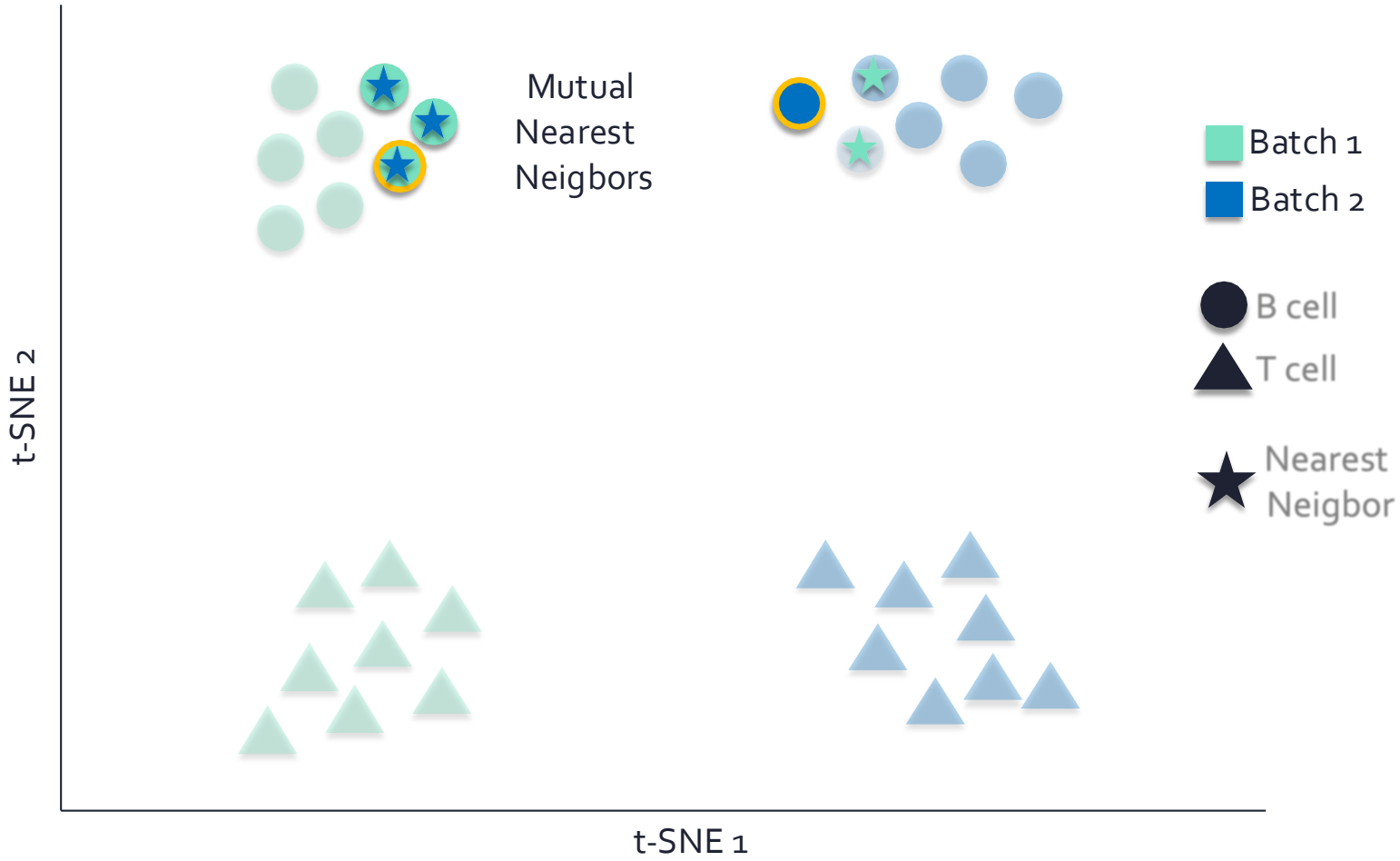
How to integrate the cells?



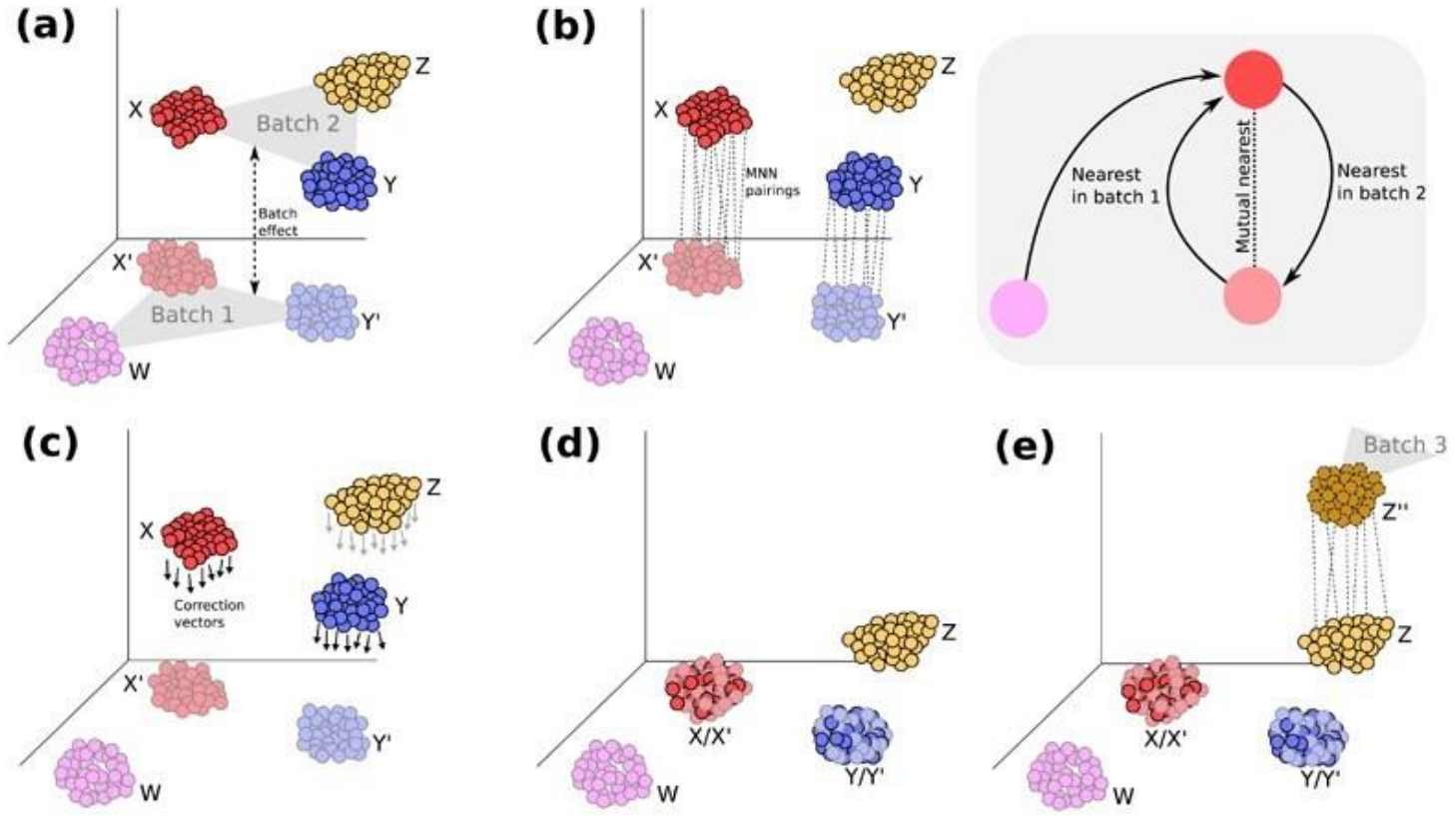
How to integrate the cells?



How to integrate the cells?

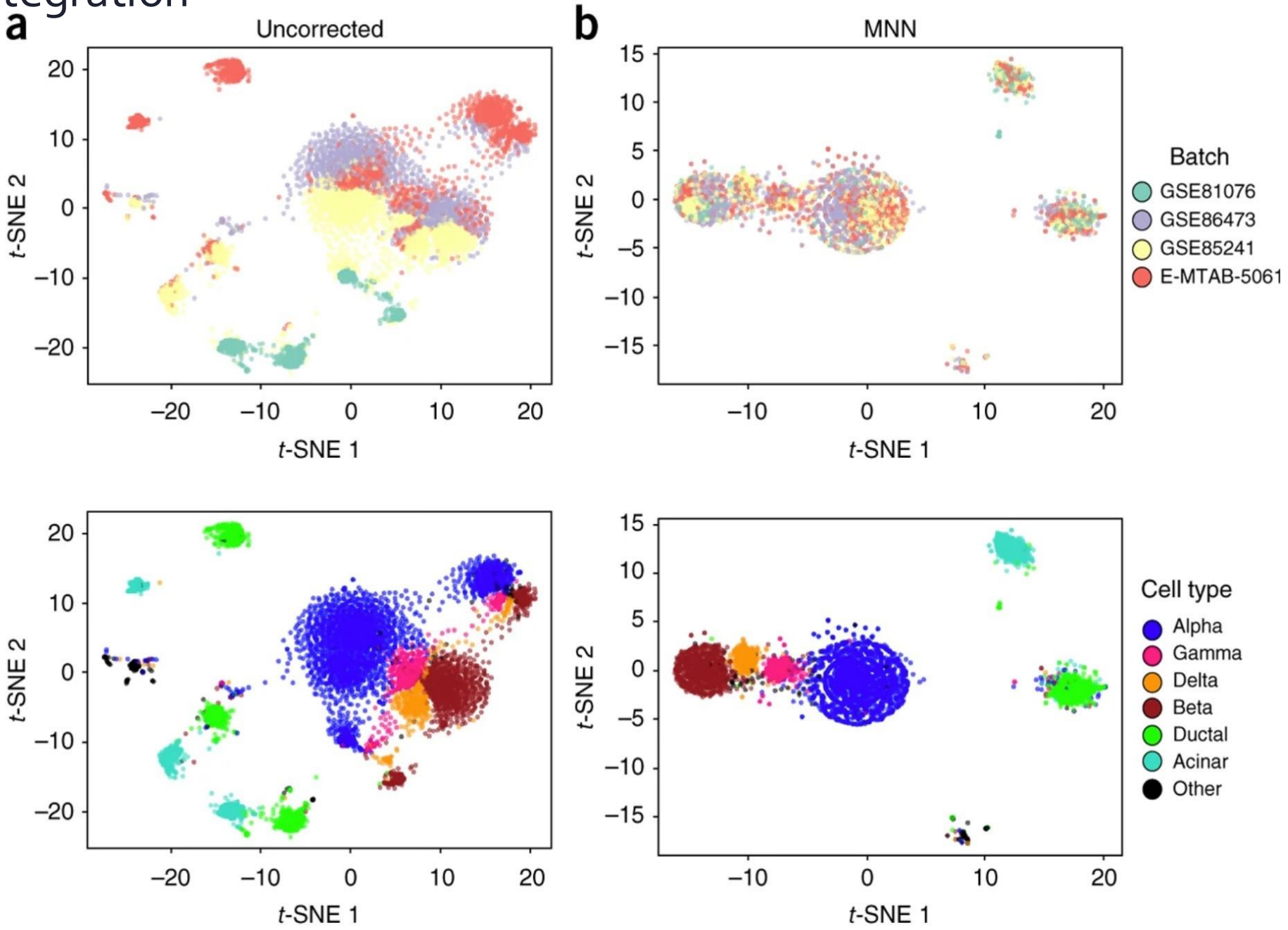


Mutual Nearest Neighbors



Adapted from: Haghverdi et al. Nat Biotechnol. 2018

Final Integration



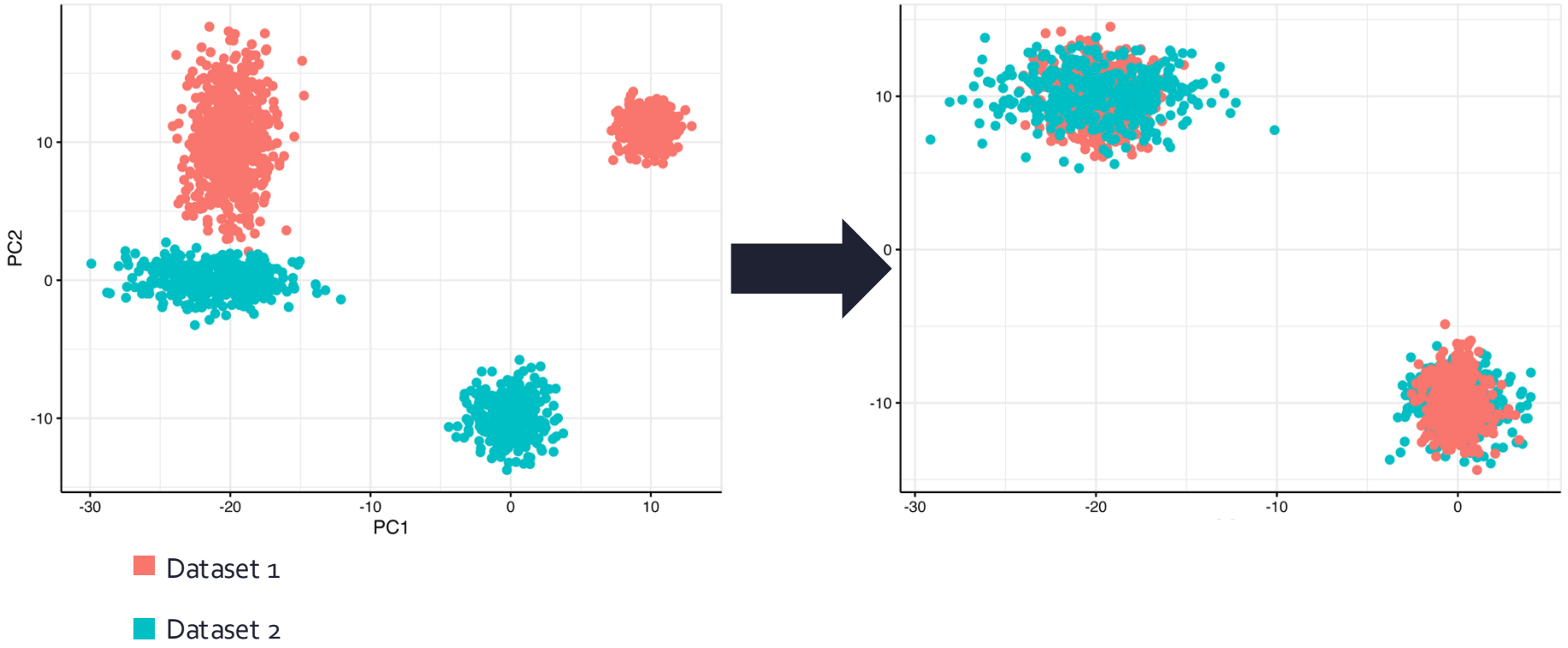
Canonical Correlation Analysis (CCA) + anchors

Find	Compute	Apply
Find corresponding cells across datasets (anchors) in L2-normalized CCA	Compute a data adjustment based on correspondences between cells	Apply the adjustment

Canonical Correlation Analysis (CCA) + anchors

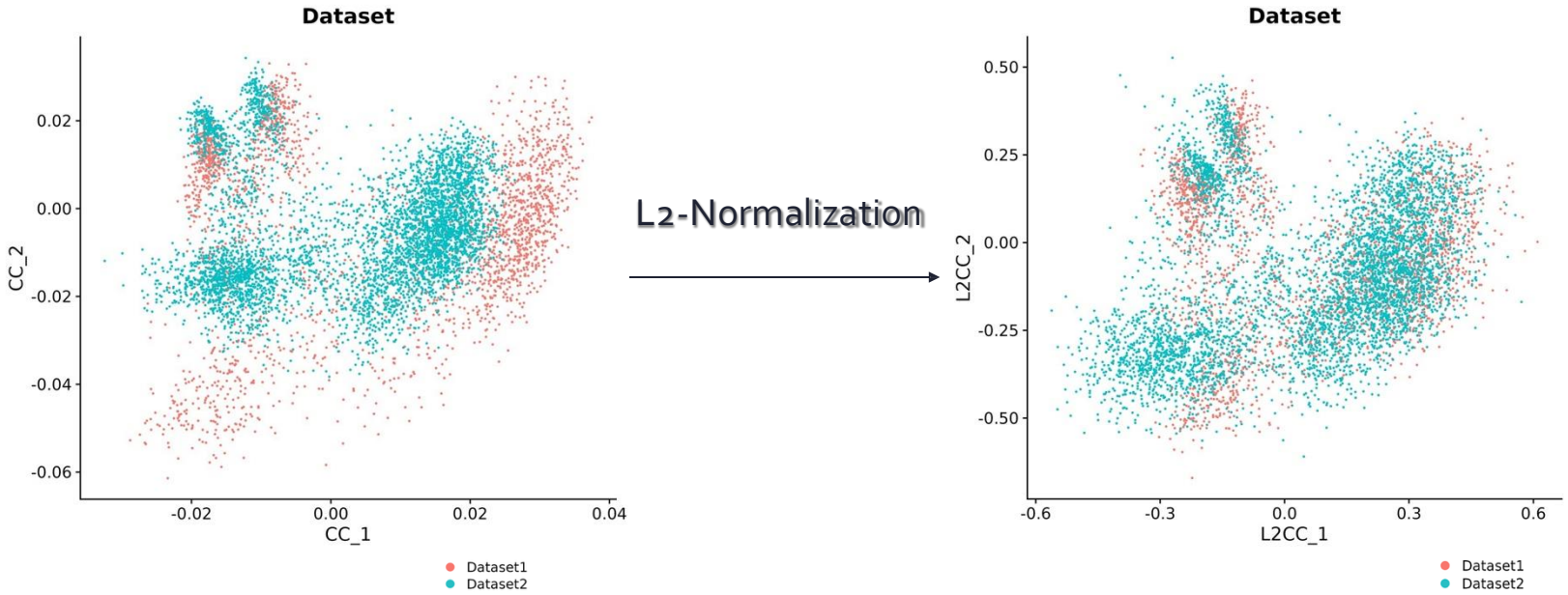
Step1

Find corresponding cells across datasets



Canonical Correlation Analysis (CCA) + anchors

L2-normalization of CCs



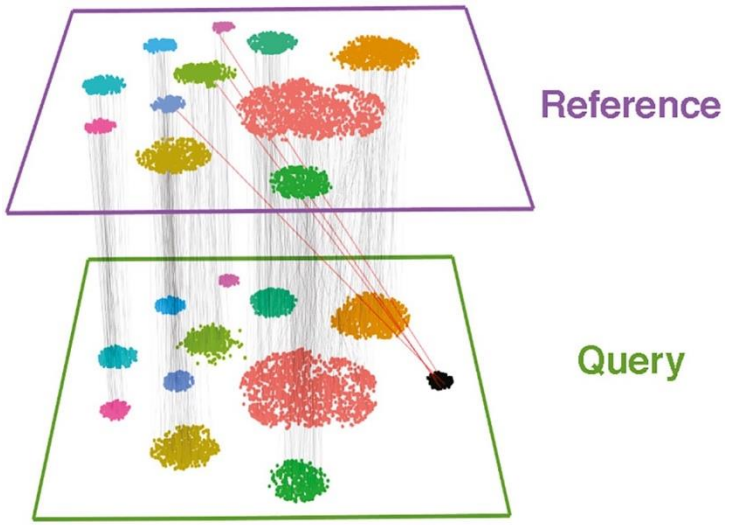
Imagine you have two datasets, A and B, each with a set of genes. CCA tries to find **linear combinations of genes** in A that correlate with corresponding combinations in B.

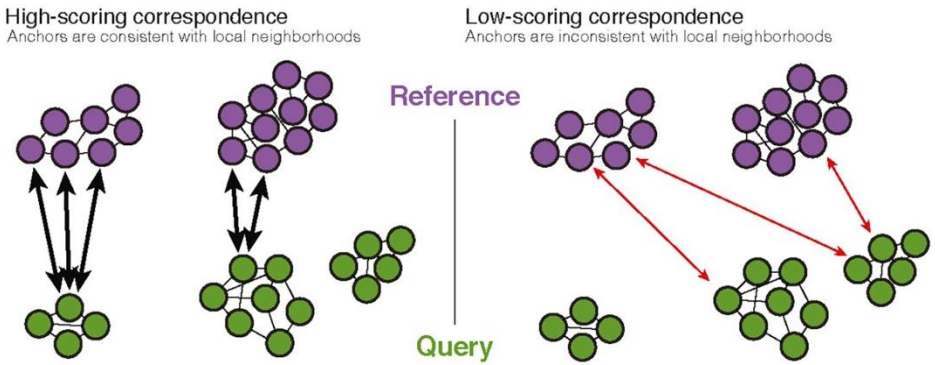
CCA looks for **pairs of "canonical" variables** (one from each dataset) that are maximally correlated with each other, capturing the shared structure

Canonical Correlation Analysis (CCA) + anchors

Anchors Identification

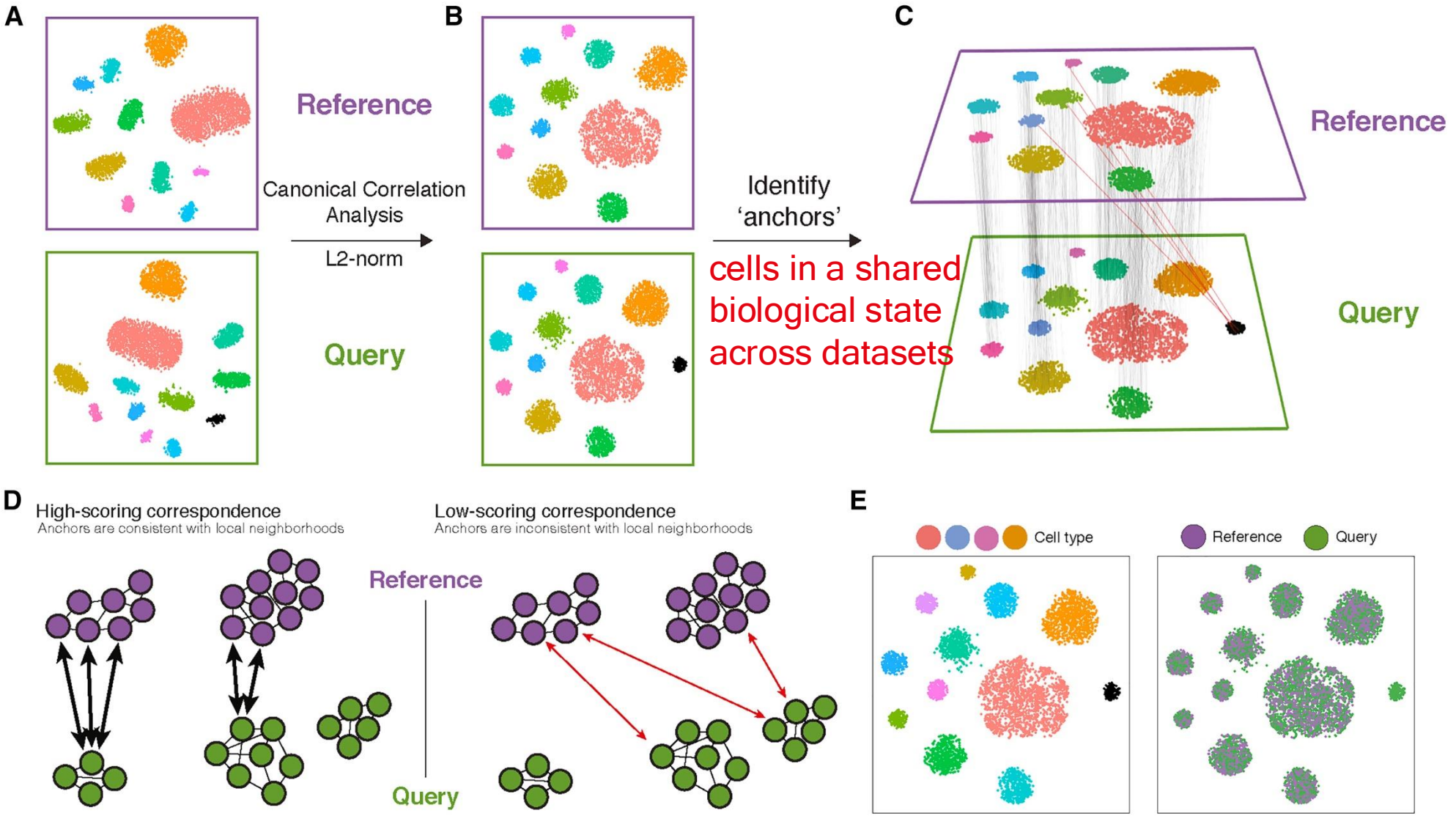
Seurat first performs **mutual nearest neighbours (MNN) matching** to identify cell pairs that are closest in gene expression space across datasets.





Integration using CCA: canonical correlation analysis

Datasets coming from 2 different platforms



scores to compute "correction" vectors for each query cell, transforming its expression so it can be jointly analyzed

Is CCA in Seurat really a CCA?

The "Seurat CCA" is taking the projection vector from the traditional CCA directly as cell embeddings. But in fact, **the classical definition of CCA would imply projecting genes into a common space rather than cells.**

Based on our understanding, the math behind the "Seurat CCA" algorithm is technically closer to a dual PCA formulation.

In the original paper, the assumption that the covariance matrix of gene expression is diagonal, is not necessary.

Furthermore, considering the formulation to preserve the most similarity (dual PCA), the low-dimensional cell embeddings should multiply the singular value, which is currently missing in the "Seurat CCA" algorithm.

And finally, there is an intrinsic connection between MNN and "Seurat CCA" (extended dual PCA).

https://xinmingtu.cn/blog/2022/CCA_dual_PCA/

Quiz

In which condition will you perform integration?

A) When cells cluster by sample
B) When cells cluster by condition
C) When cells cluster by batch
D) When cells cluster by dataset
E) None of the above
F) All of the above

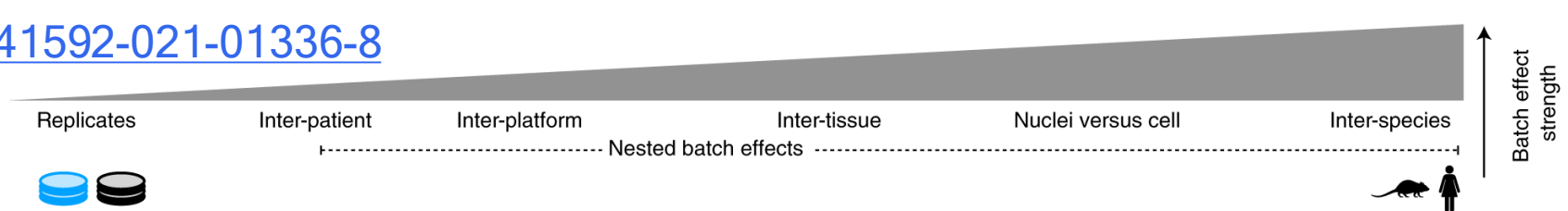
Guidelines to choose an integration method

a

	Considerations	scANVI	Scanorama embed	scVI	FastMNN embed	scGen	Harmony	FastMNN gene	Seurat v3 RPCA	BBKNN	Scanorama gene	ComBat	MNN	Seurat v3 CCA	trVAE	Conos	DESC	LIGER	SAUCIE embed	SAUCIE gene
Input	Programming language	Python	Python	Python	R	Python	R	R	R	Python	Python	Python/R	Python/R	R	Python	R	Python	R	Python	Python
	Method runs without additional information	✗				✗														
Scib results	Consistent top performer	✓	✓	✓		✓														
	Top method on small/ simple tasks		✓		✓	✓	✓													
	Top method on large/ complex tasks	✓	✓	✓		✓														
	Top method on ATAC data	–		–			✓											✓		
Task details	Integrates strong batch effects	✓	–	–		✓			–	–				–						
	Top method for recovery cell states or modules	✓	✓								✓	✓	✓							
	Confounding of bio and batch variance	✓	–			✓														
	Top method for trajectories	–	✓	–	✓	✓														
	Method deals with varying compositions											✗								
Speed	Fast method for quick results									✓		✓								
	Scales well to large datasets on CPU	✓	–	✓						✓	–								✓	✓
	Method has GPU support	✓		✓		✓									✓		✓		✓	✓
	Scales well to feature spaces beyond genes														✓	✓				
Output	Method shows corrected expression					✓		✓	✓		✓	✓	✓	✓						✓
	Method gives relative cell embeddings									✗						✗				

✓ Fulfills the criterion
– Partial fulfillment of criterion
✗ Does not fulfill criterion
Python
R

b



https://www.nature.com/articles/s41592-021-01336-8

Summary

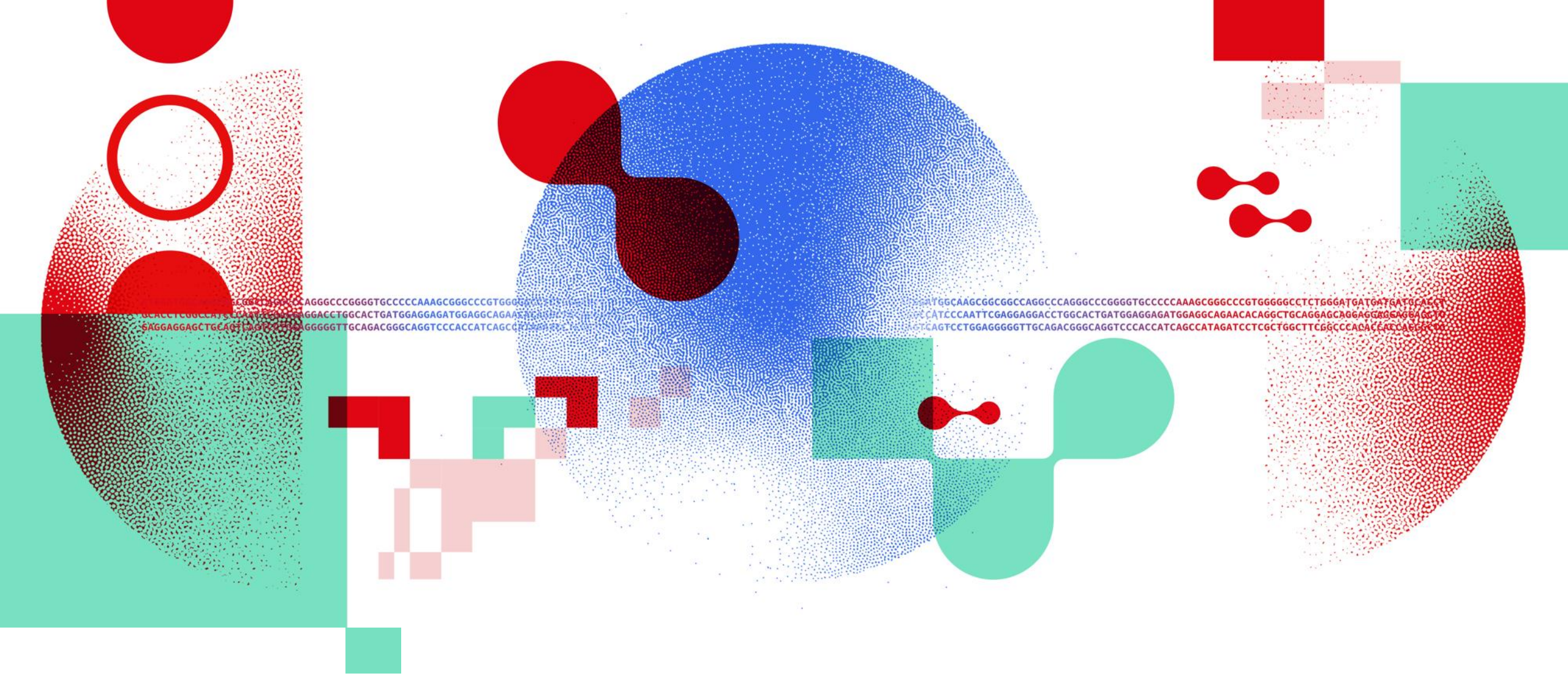
Experimental Design Matters: Optimize design to improve data quality and analysis

Integration Scenarios:

- Conditions: Compare different experimental conditions
- Datasets: Analyze data from different platforms together

Integration Using CCA:

- Align cells across groups to improve clustering and downstream analyses
- Compute correction vectors for each query cell to transform its expression for joint analysis



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss