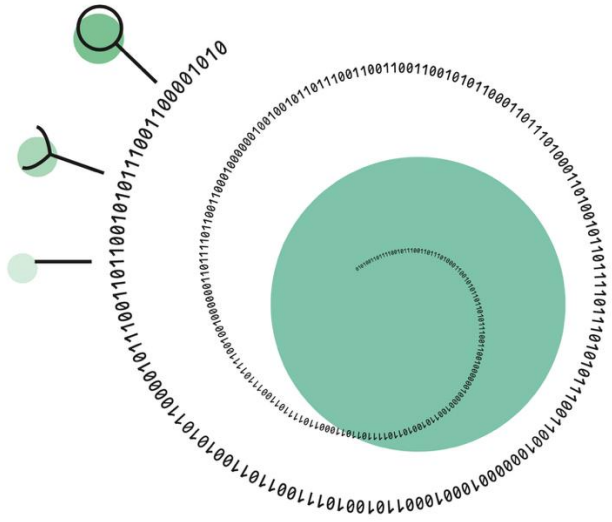
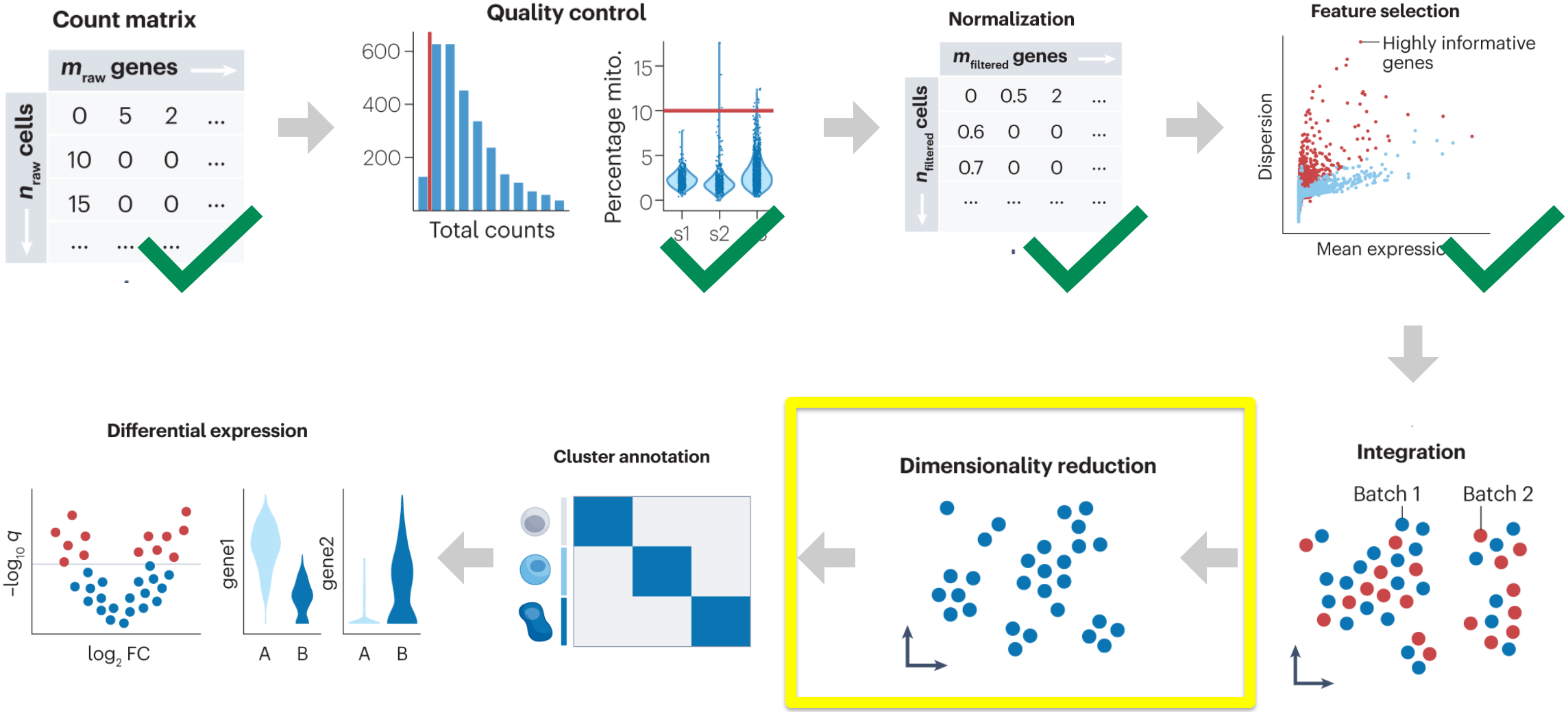


Single Dell Data Analysis Roundtable

Lecture 3 – Dimension reduction, clustering & visualization of high-dimensional data

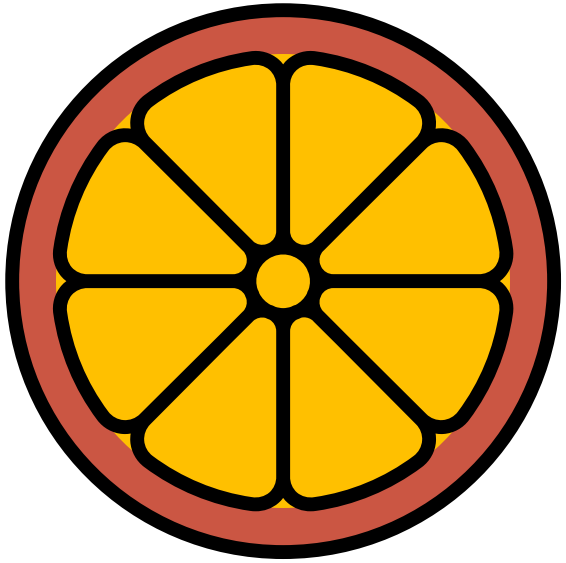
Lisa Buchauer
Professor of Systems Biology of Infectious Diseases
Department of Infectious Diseases and Intensive Care
Charité - Universitätsmedizin Berlin

Today



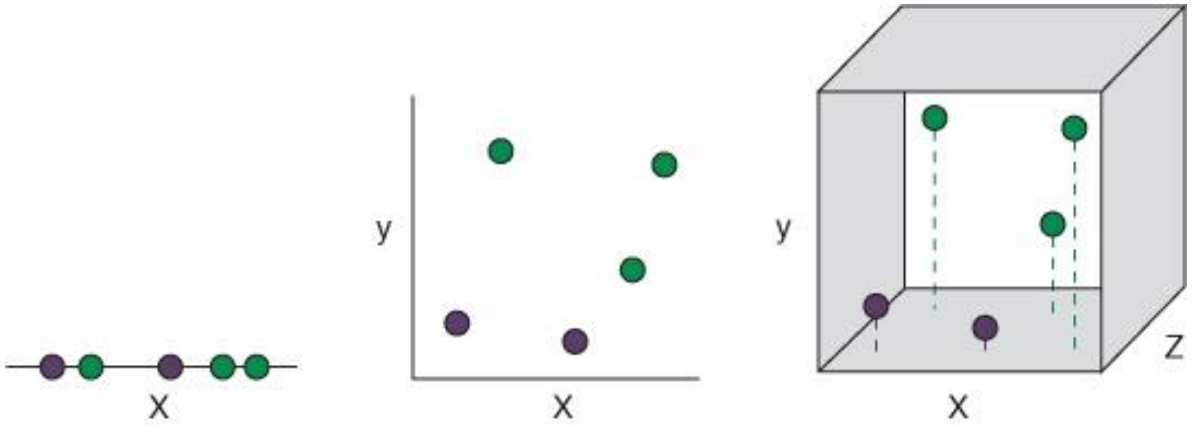
Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w

Losing intuition in high dimensions: a “fun fact”



In a high-dimensional orange, most of the mass is in the skin, not in the pulp

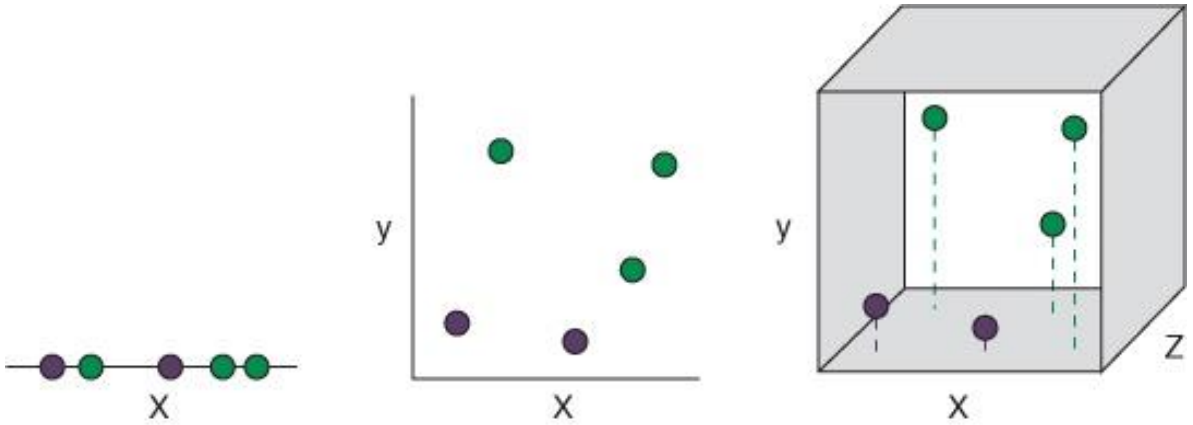
“The curse of dimensionality”



data sparsity
most of the high-dim space is empty

https://livebook.manning.com/concept/r/curse

“The curse of dimensionality”



data sparsity
most of the high-dim space is empty

distances may lose meaning

computational cost/ algorithmic efficiency

noise dimensions

visualization challenges

…

https://livebook.manning.com/concept/r/curse

Reducing dimensionality

Step 1: Feature selection

scanpy

```python
sc.pp.highly_variable_genes(adata, n_top_genes=2000, batch_key="sample")
```

Seurat

```r
pbmc <- FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)
```

https://cofactorgenomics.com/curse-of-dimensionality-wk-16/

From PCA to clusters and UMAP in basic tutorials

1

```
sc.tl.pca(adata)
```

2

```
sc.pp.neighbors(adata)
```

3

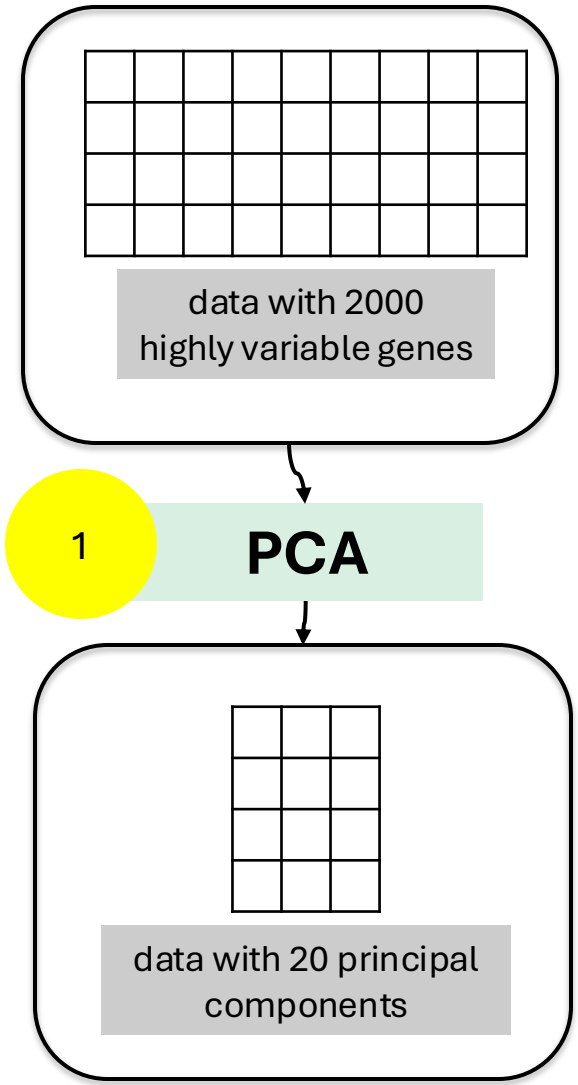
```
# Using the igraph implementation and a fixed number of iterations can be significantly fast
sc.tl.leiden(adata, flavor="igraph", n_iterations=2)
```

4

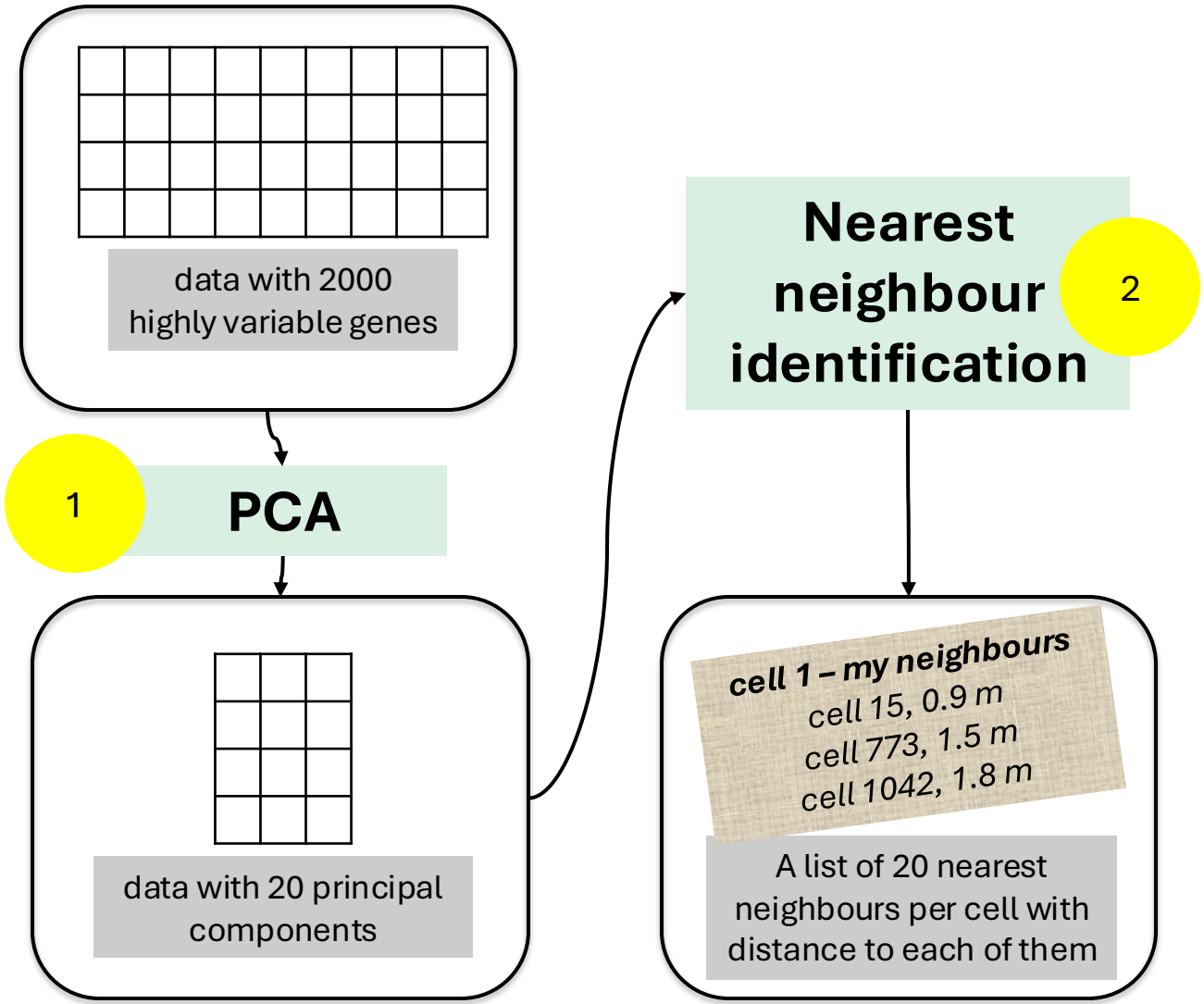
```
sc.tl.umap(adata)
```

https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html

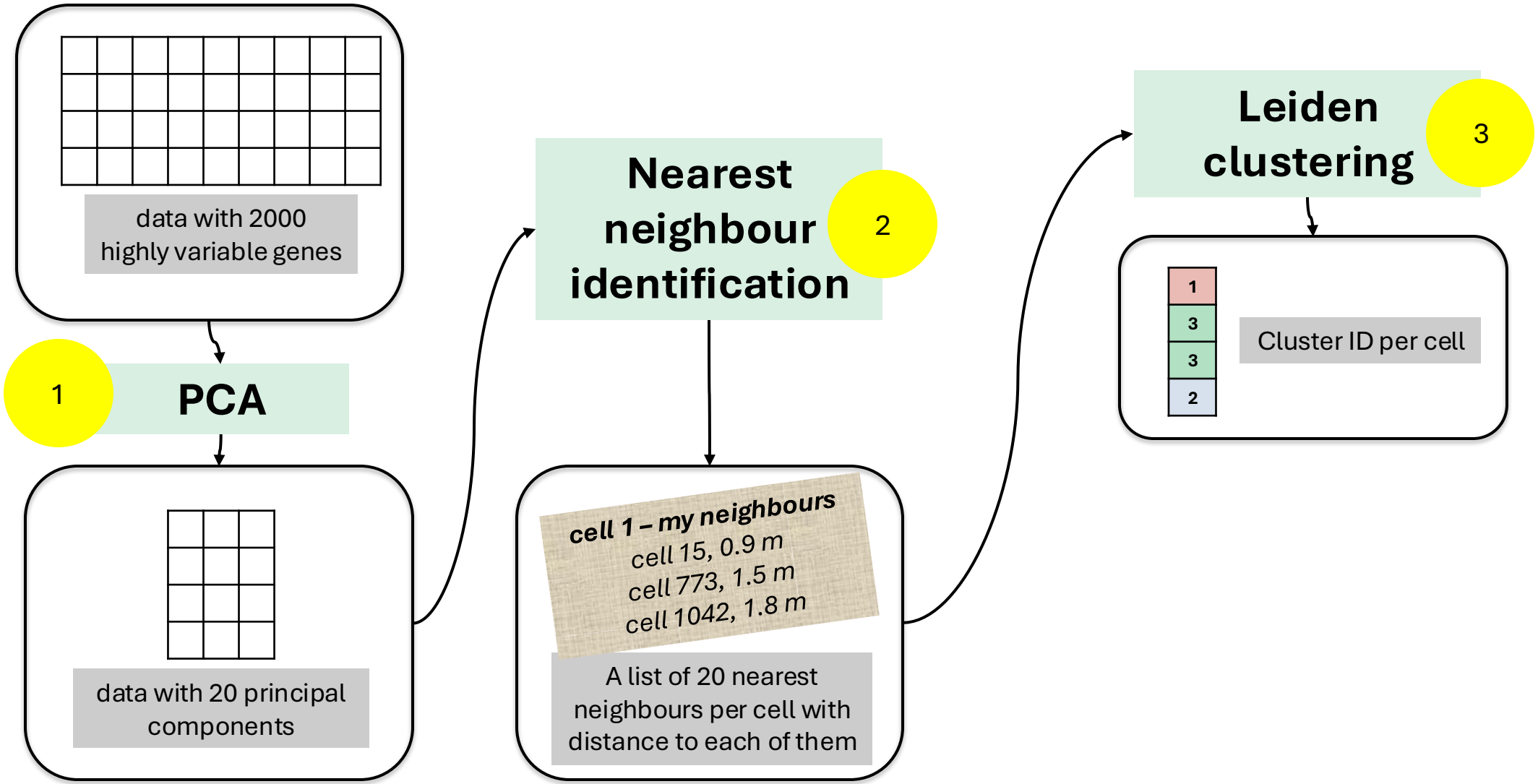
Data types along the processing path



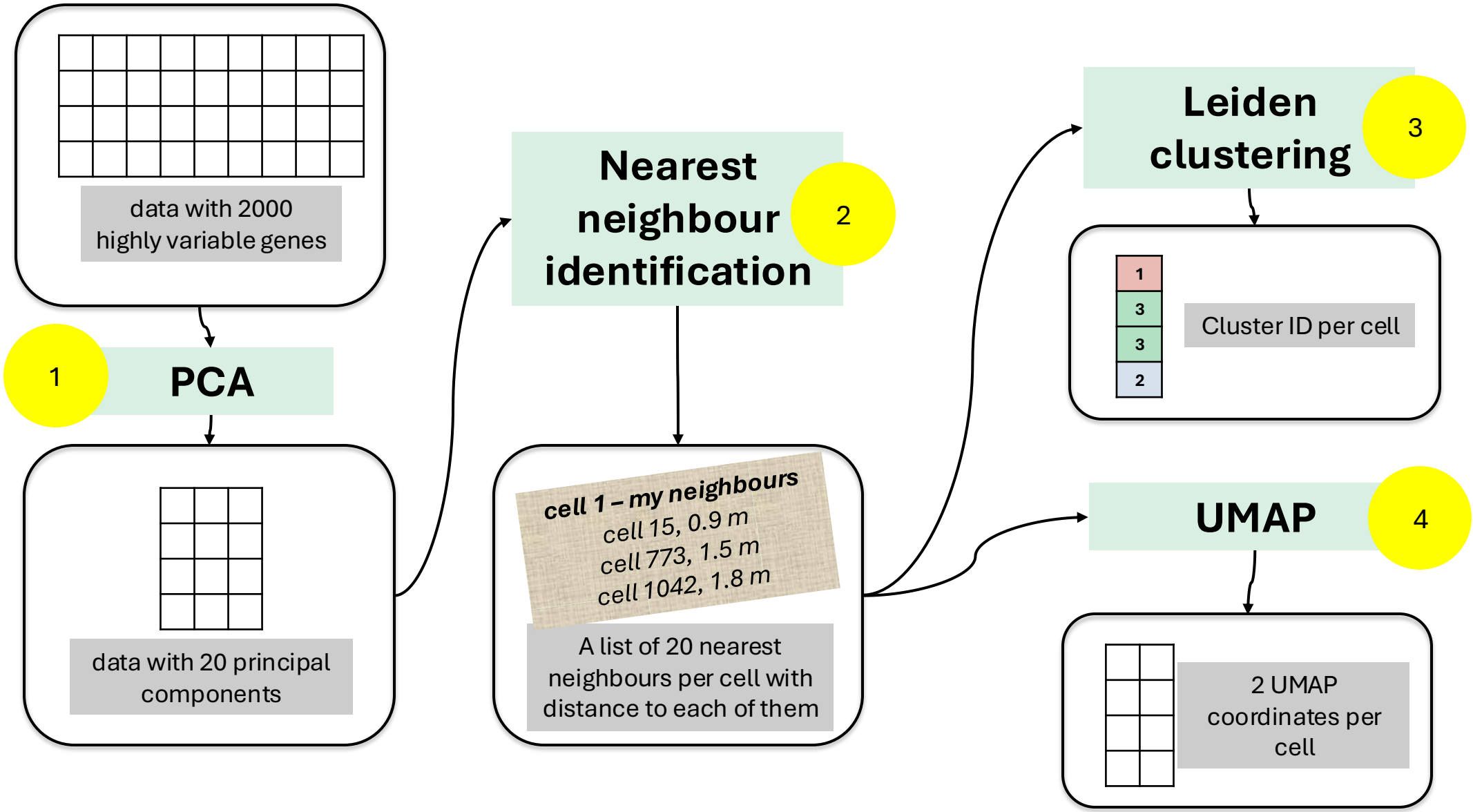
Data types along the processing path



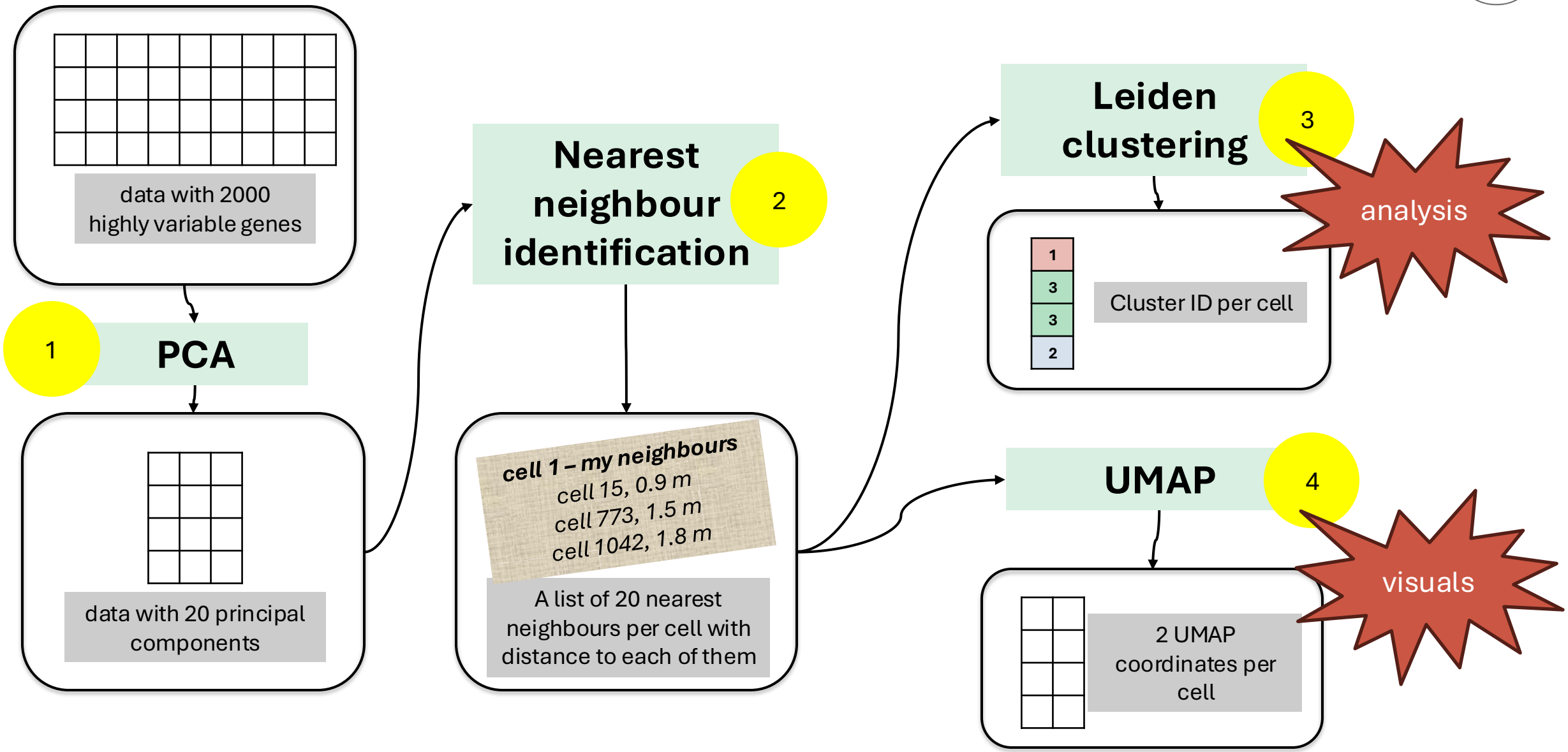
Data types along the processing path



Data types along the processing path



Data types along the processing path



1

Principal Component Analysis

python

```
sc.tl.pca(adata)
```

R

```
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
```

https://satijalab.org/seurat/articles/pbmc3k_tutorial
https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html

1

Principal Component Analysis

python

```
sc.tl.pca(adata)
```

R

```
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
```

Motivation

Less dimensions are better for downstream performance and interpretability

Many genes have correlated behaviour – not every original dimension adds value

https://satijalab.org/seurat/articles/pbmc3k_tutorial
https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html

1

Principal Component Analysis

python

```
sc.tl.pca(adata)
```

R

```
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
```

Motivation

Less dimensions are better for downstream performance and interpretability

Many genes have correlated behaviour – not every original dimension adds value

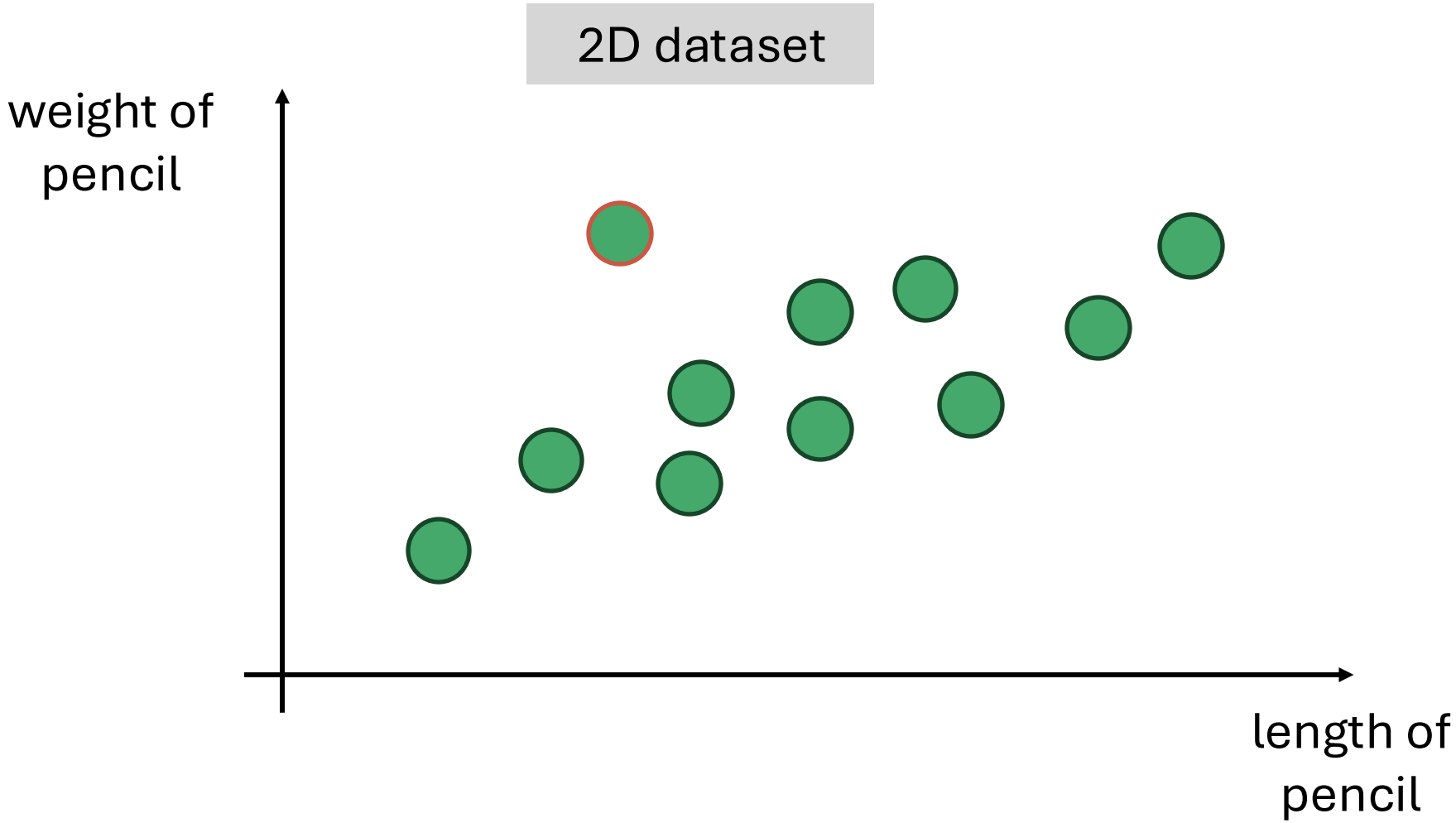
Goal

Move as much information as possible into as few dimensions as possible

https://satijalab.org/seurat/articles/pbmc3k_tutorial
https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html

1

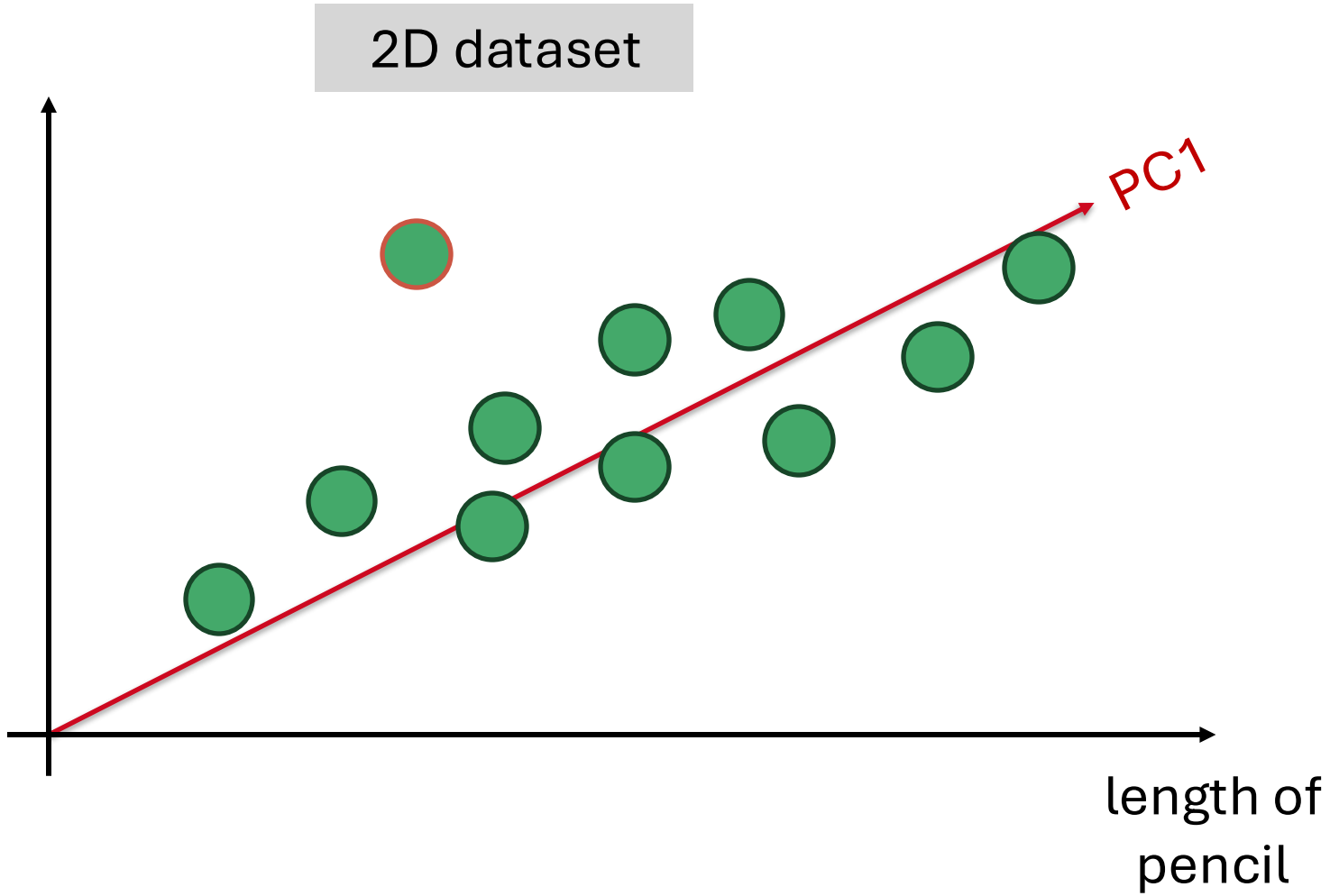
Principal Component Analysis



1

Principal Component Analysis

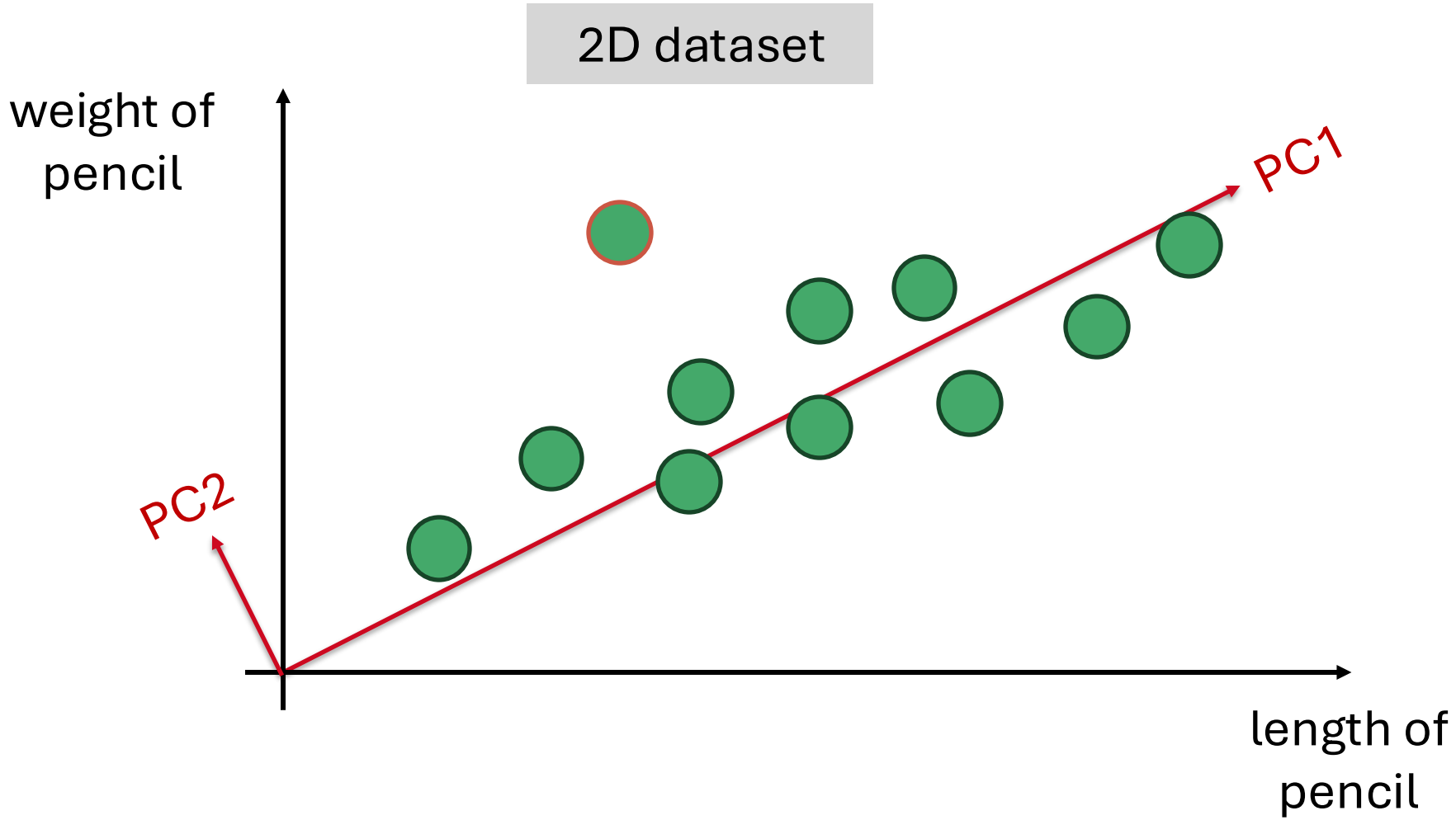
Find the dimension which captures the most variability



1

Principal Component Analysis

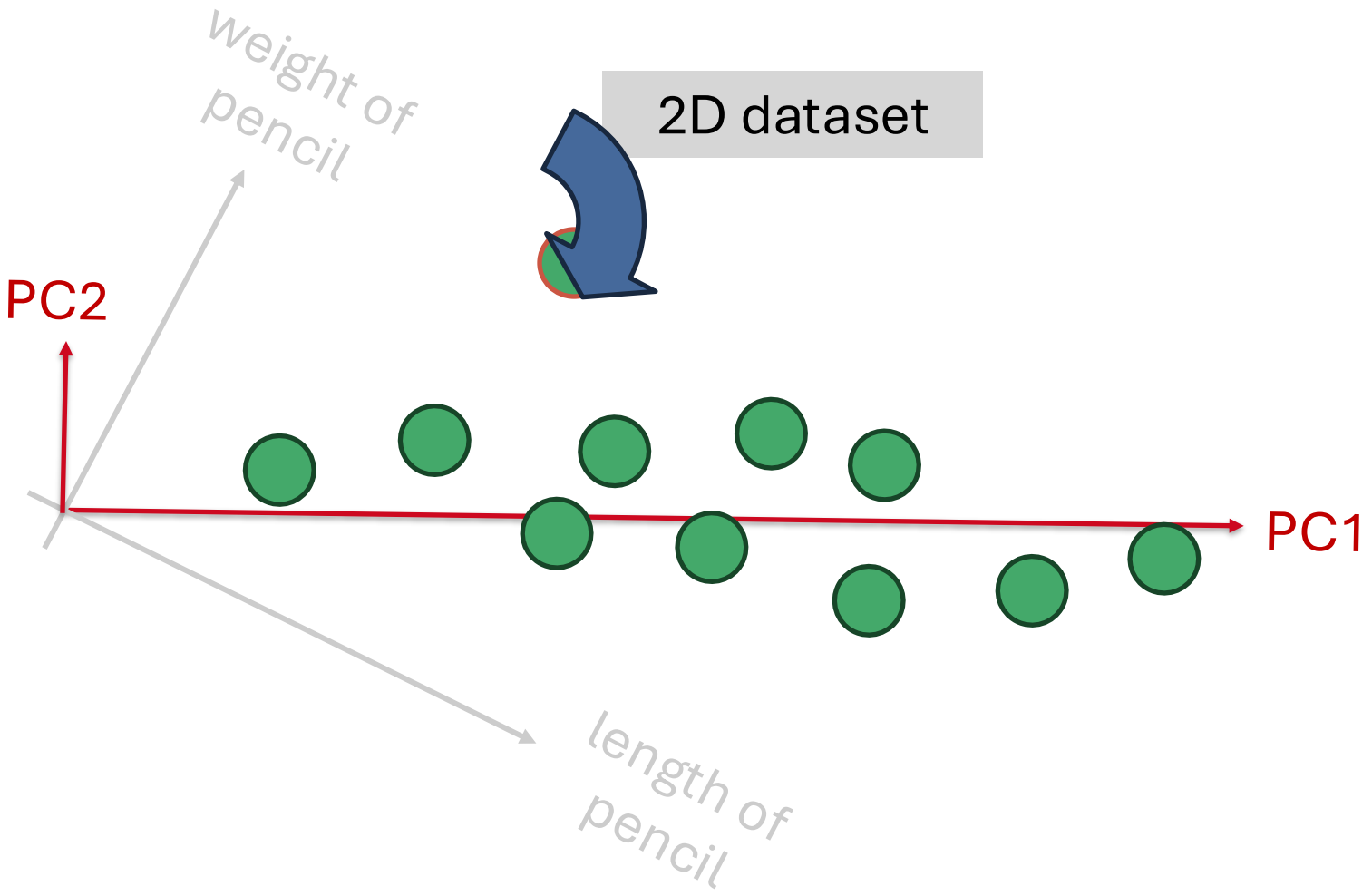
Find the dimension which captures the **2nd** most variability (etc)



1

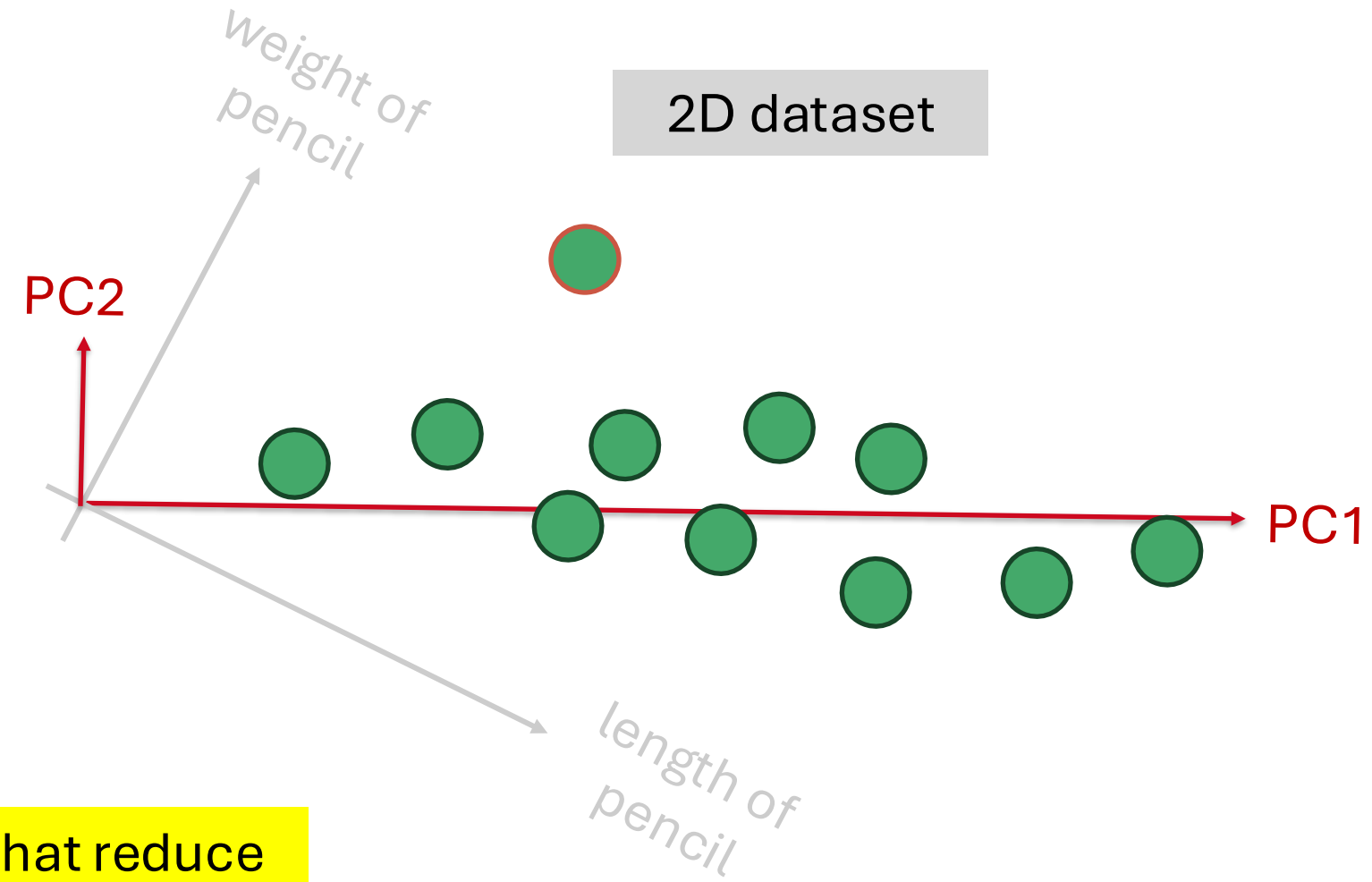
Principal Component Analysis

Describe each point in terms of the new rotated coordinates



1

Principal Component Analysis

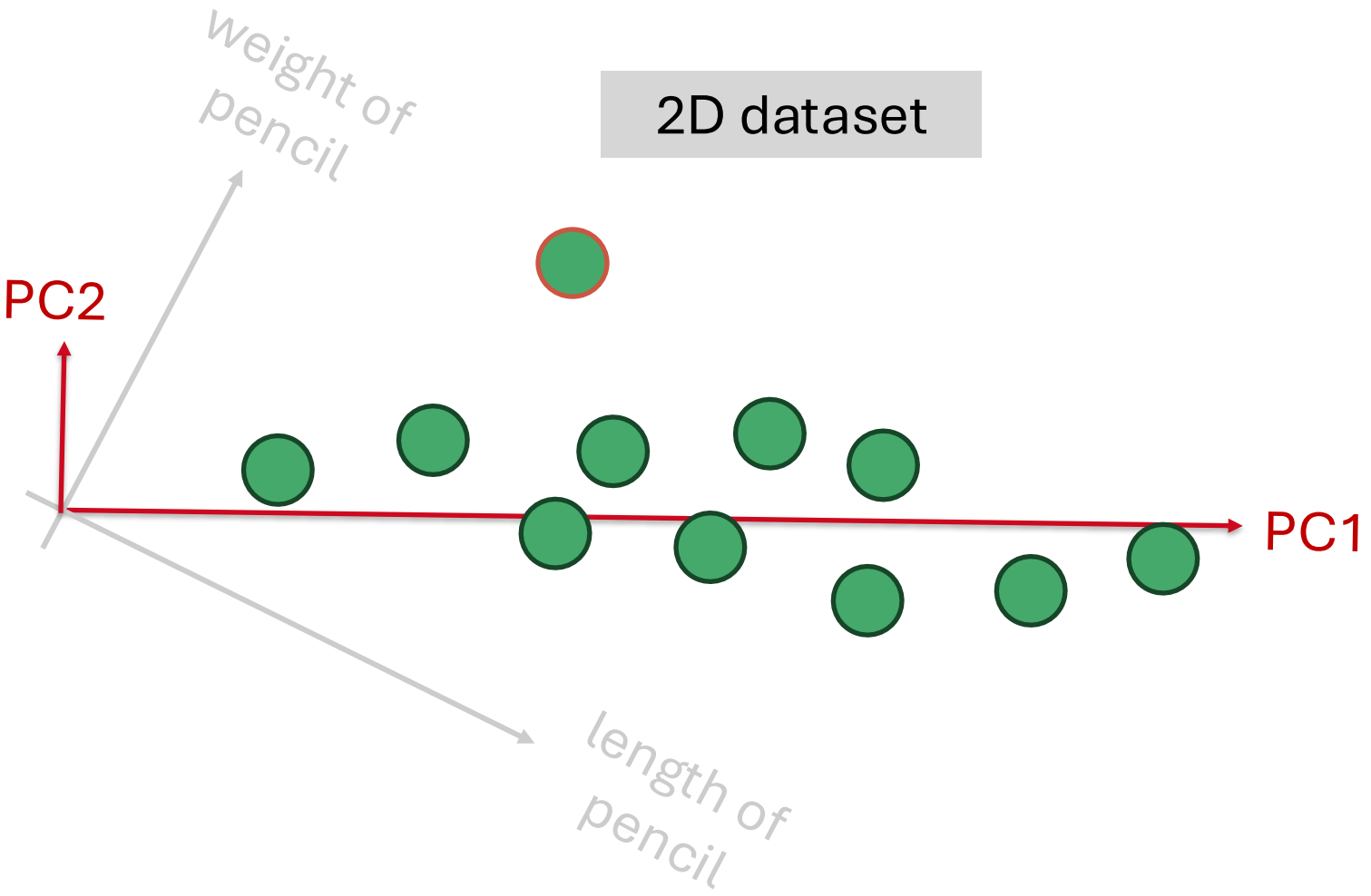


Describe each point in terms of the new rotated coordinates

?! How does that reduce dimensionality?

1

Principal Component Analysis

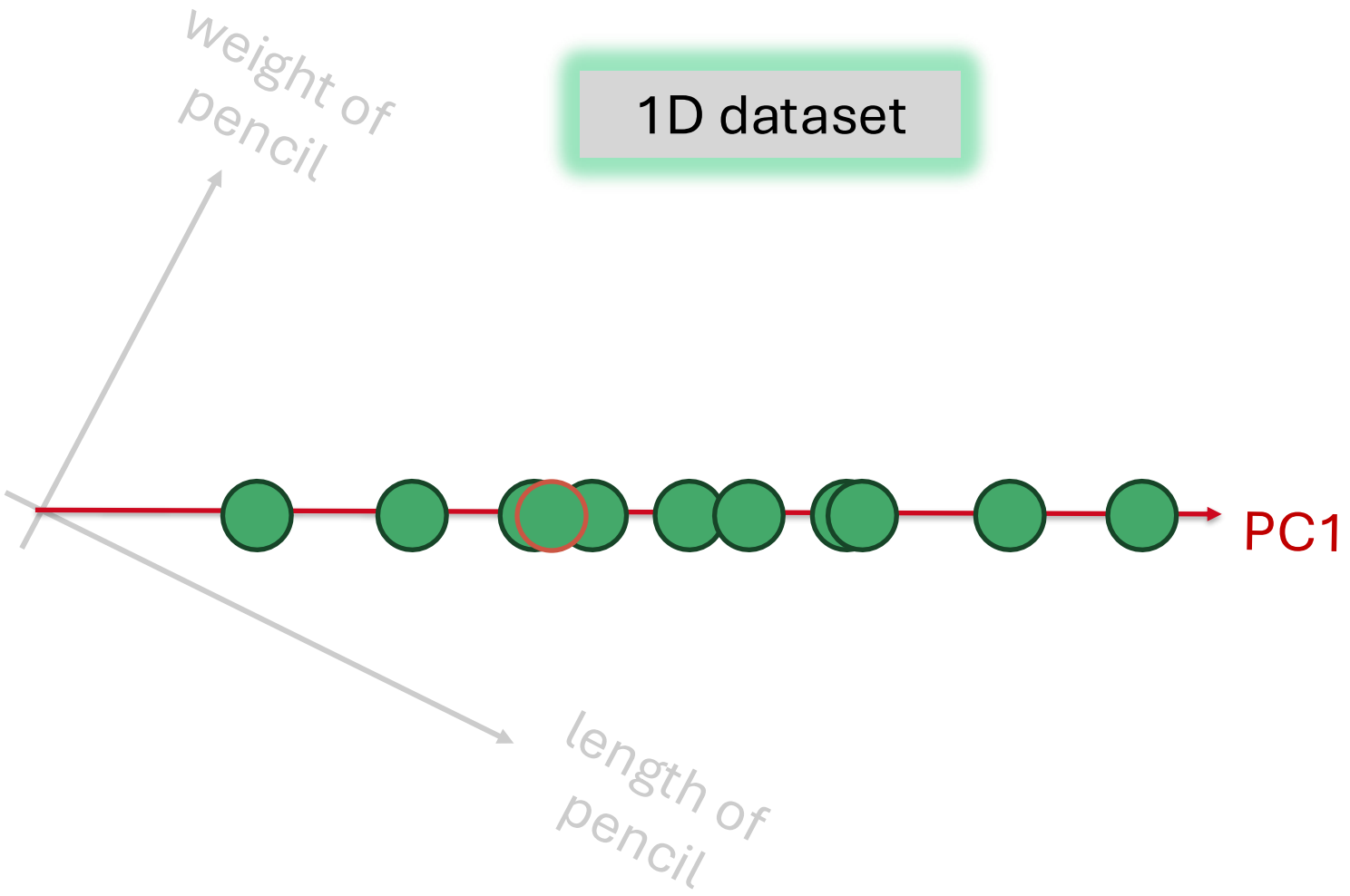


Inspect how much of the total variance is explained by each PC

	Explained variance
PC1	95%
PC2	5%

1

Principal Component Analysis



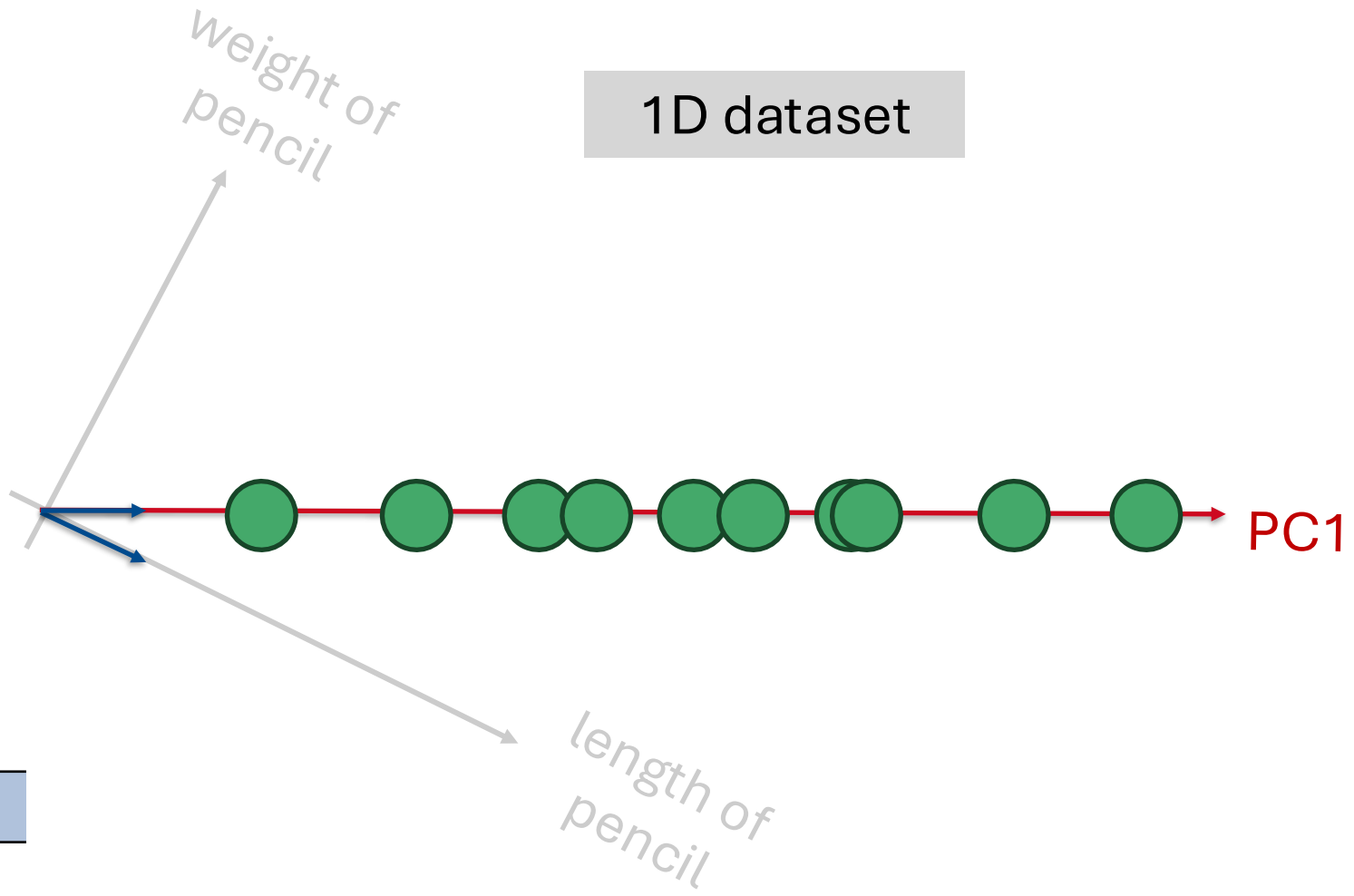
Discard lesser important dimensions

	Explained variance
PC1	95%

1

Principal Component Analysis

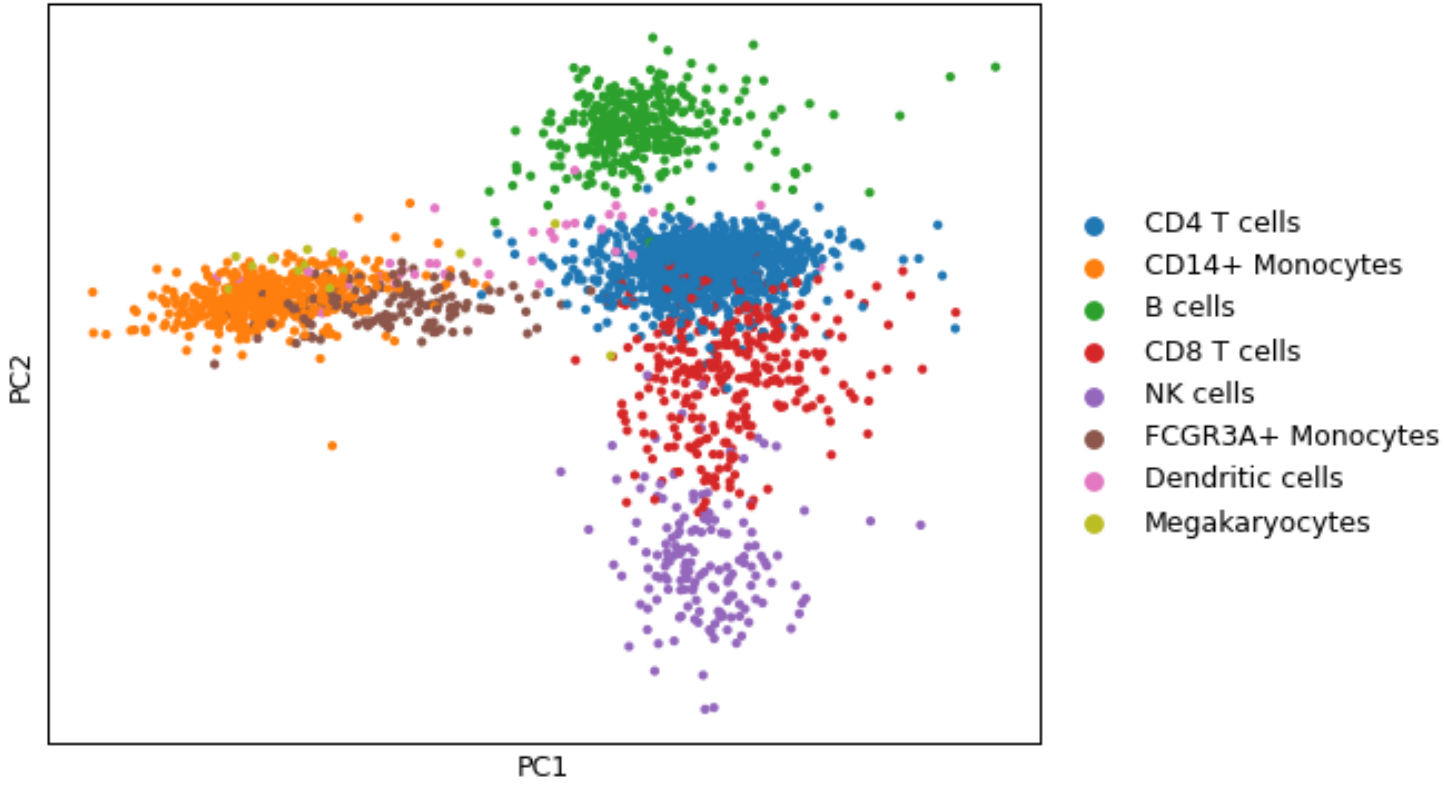
Loadings (projections of features onto PCs) help interpretability



	Loading on PC1
Length	0.9
Weight	0.2

1

Principal Component Analysis for single cell data

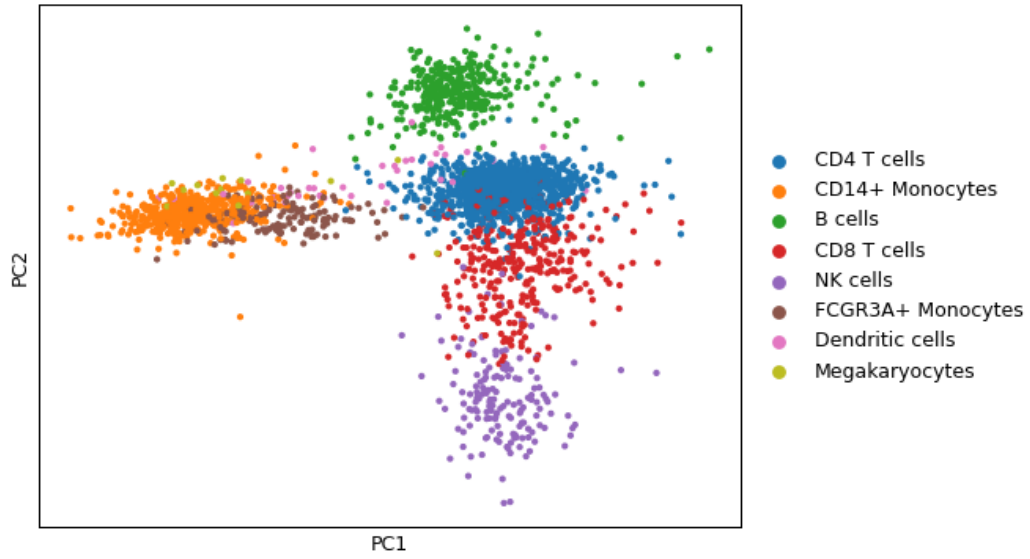


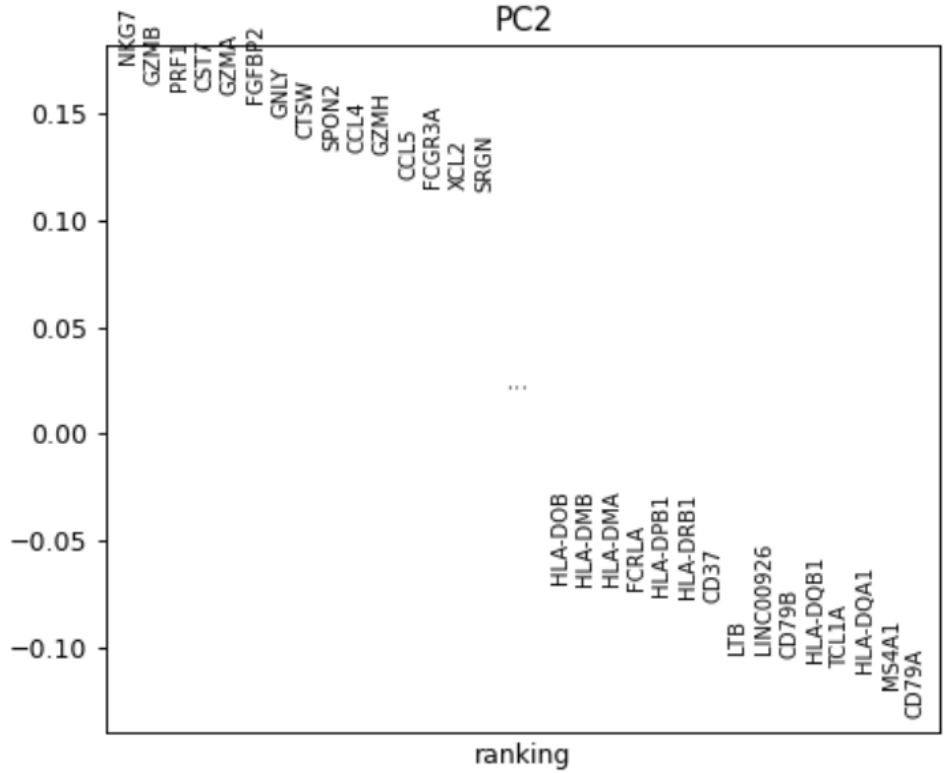
First 2 components are sufficient to separate B cells, T cells and monocytes!

https://scanpy.readthedocs.io/en/stable/api/generated/scanpy.pl.pca_overview.html

1

Principal Component Analysis for single cell data



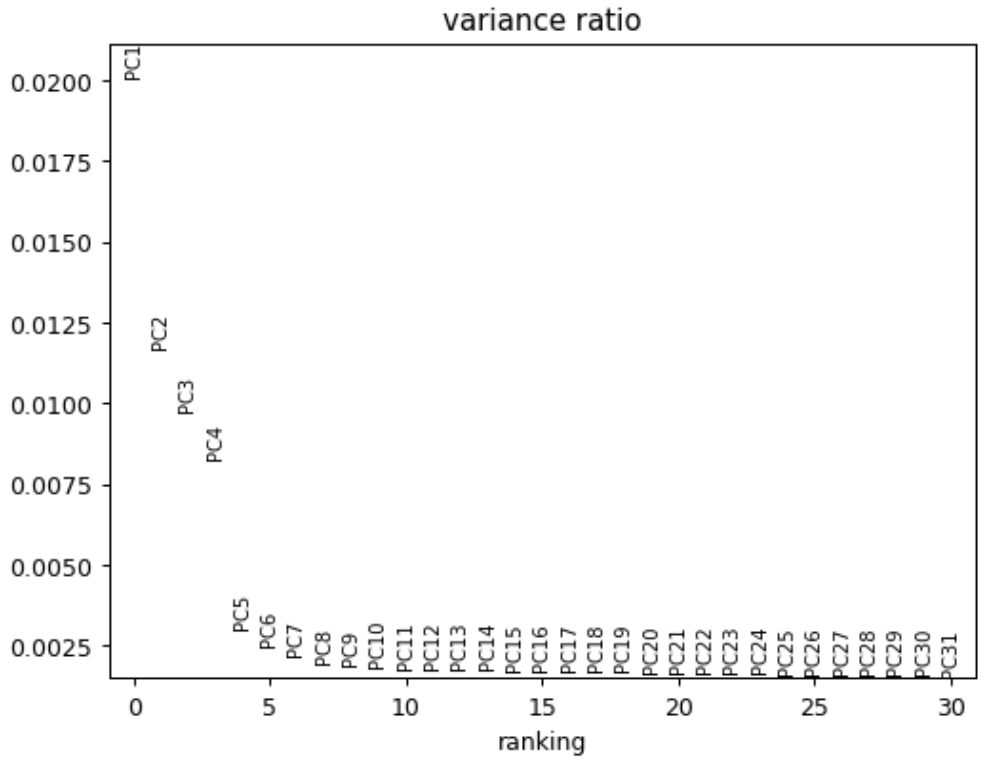


Important loadings on PC2 include NK cell marker NKG7 and B cell marker CD79a

https://scanpy.readthedocs.io/en/stable/api/generated/scanpy.pl.pca_overview.html

1

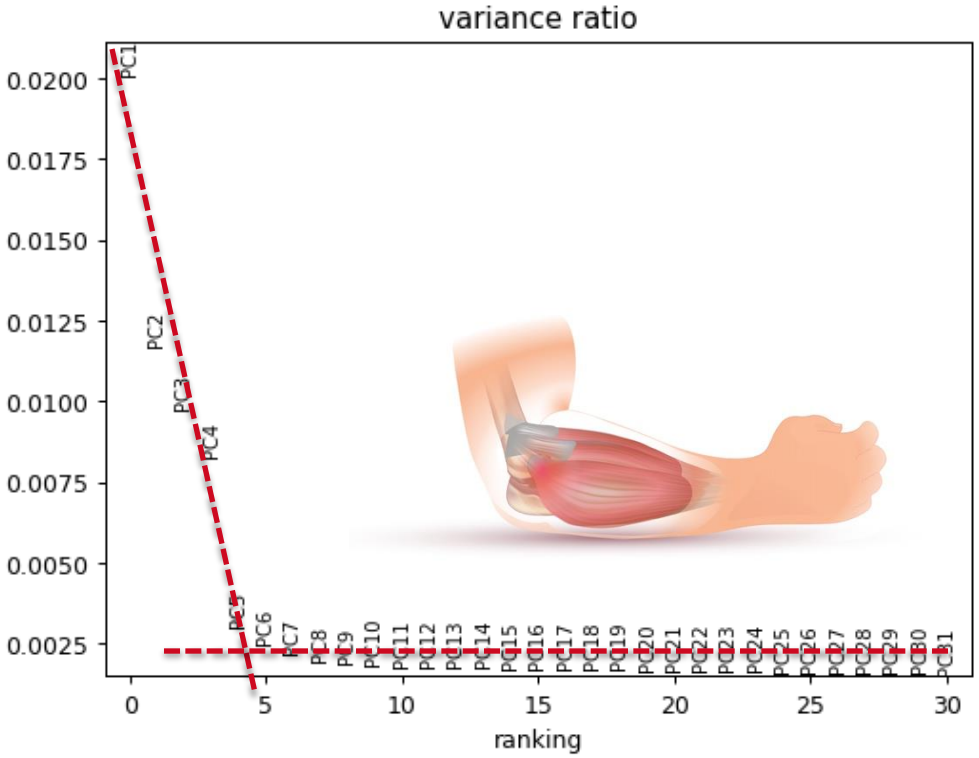
Principal Component Analysis for single cell data



https://scanpy.readthedocs.io/en/stable/api/generated/scanpy.pl.pca_overview.html

1

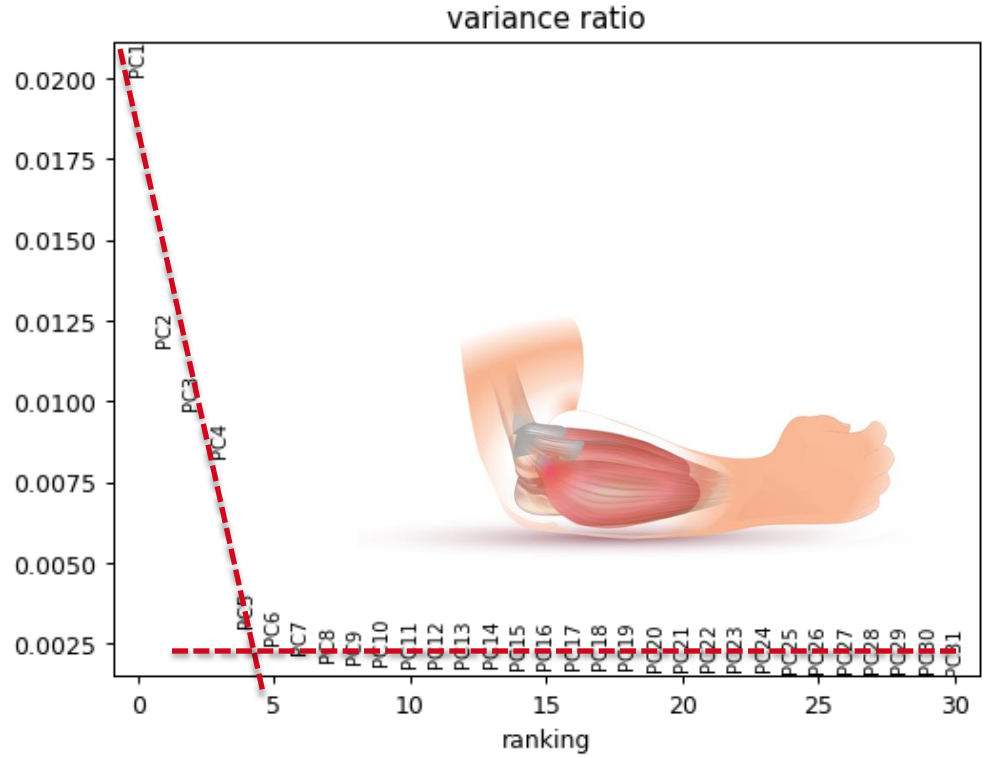
Principal Component Analysis for single cell data



https://scanpy.readthedocs.io/en/stable/api/generated/scanpy.pl.pca_overview.html

1

Principal Component Analysis for single cell data



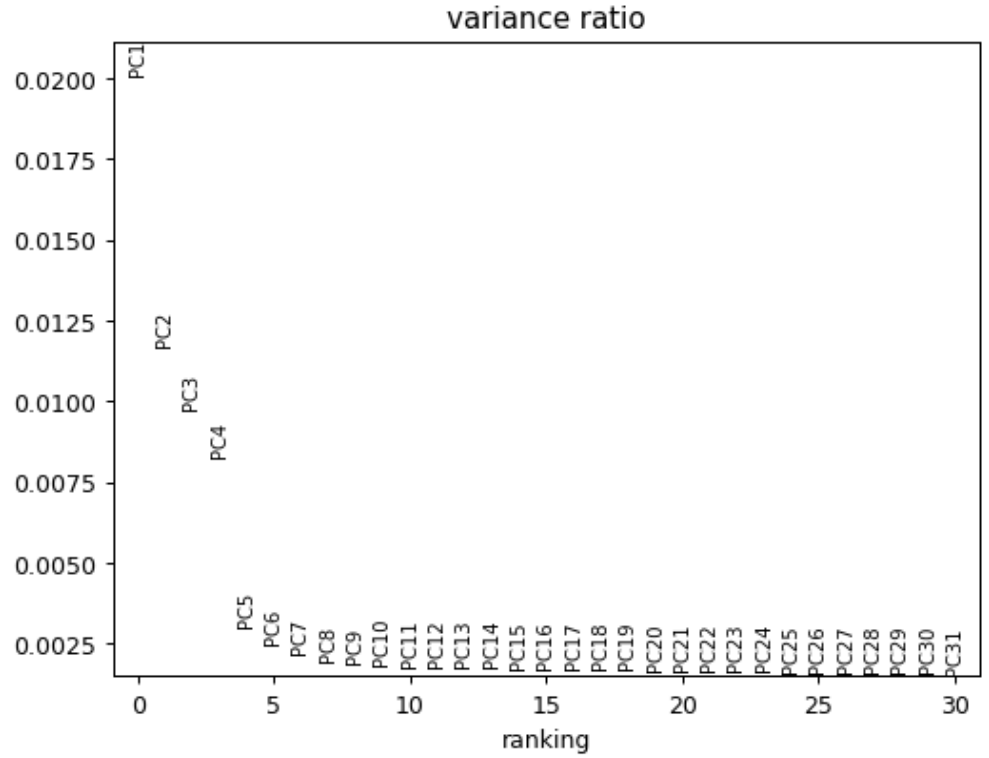
Elbow method can help to decide on the number of PCs to keep ...

- but may be too aggressive
- In this PBMC example, first 100 PCs explain 20% of variance

https://scanpy.readthedocs.io/en/stable/api/generated/scanpy.pl.pca_overview.html

1

Principal Component Analysis for single cell data



- Typically: keep 10-30 dimensions based on dataset complexity (e.g. number of cell types, conditions)
- If in doubt, keep more rather than removing too much info
- Repeat analysis with different choices of PCs

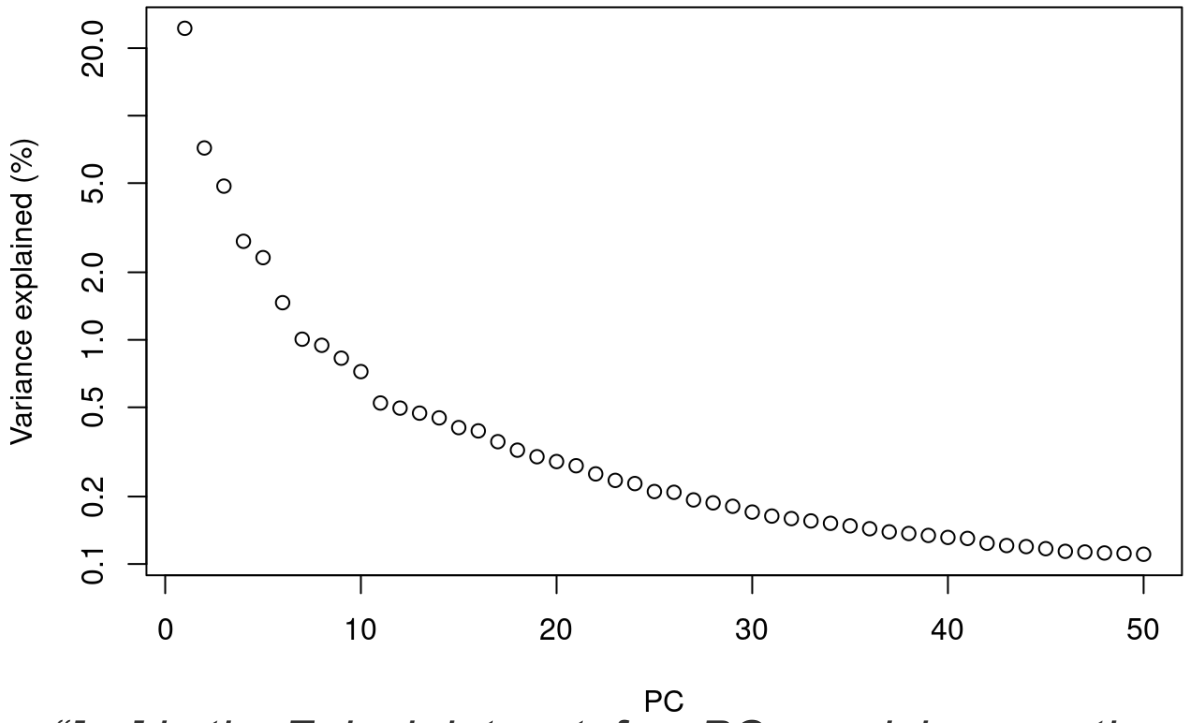
https://scanpy.readthedocs.io/en/stable/api/generated/scanpy.pl.pca_overview.html

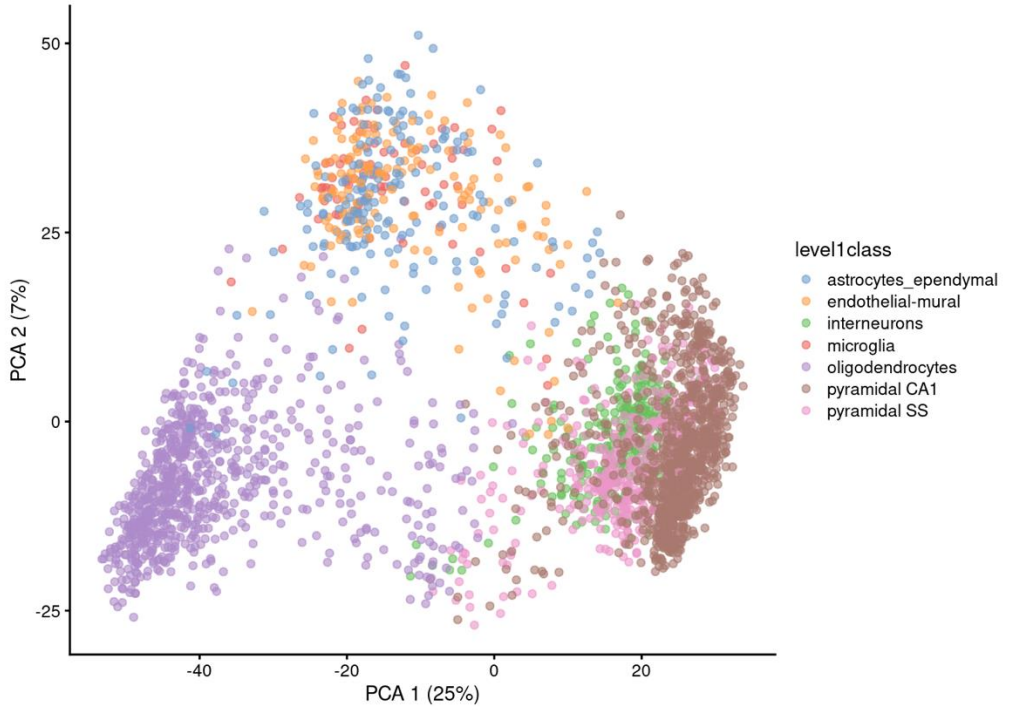
1

More diverse populations → first PCs capture more variability

Zeisel 2015 mouse brain dataset

(https://www.science.org/doi/10.1126/science.aaa1934, another classic dataset used in many tutorials)





"[…] in the Zeisel dataset, few PCs explain more than 1% of the variance in the entire dataset (Figure 4.1) and choosing between, say, 20 and 40 PCs would not even amount to four percentage points' worth of difference in variance."

https://bioconductor.org/books/3.13/OSCA.basic/dimensionality-reduction.html