Single-cell RNA-seq Analysis in Python

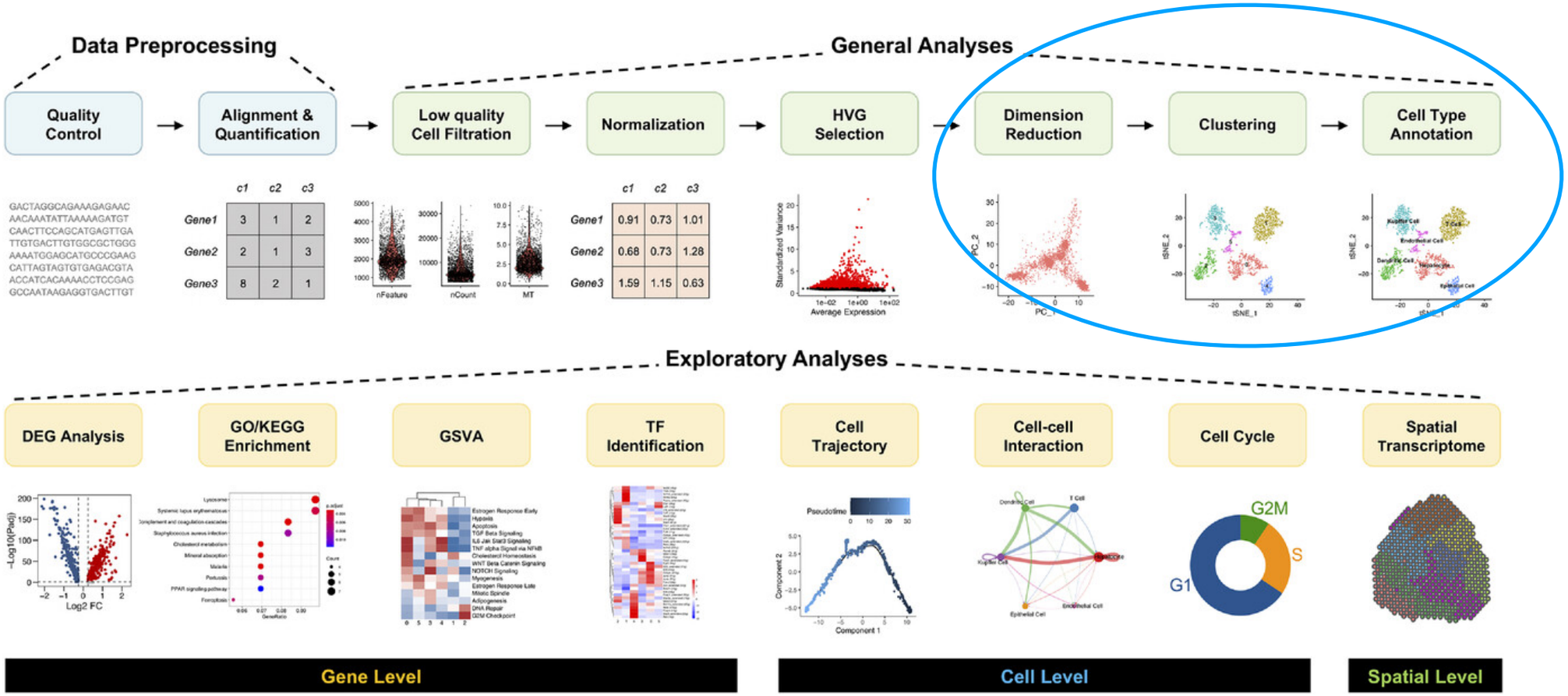
June 2024

Overview of Day #2

Day 2:

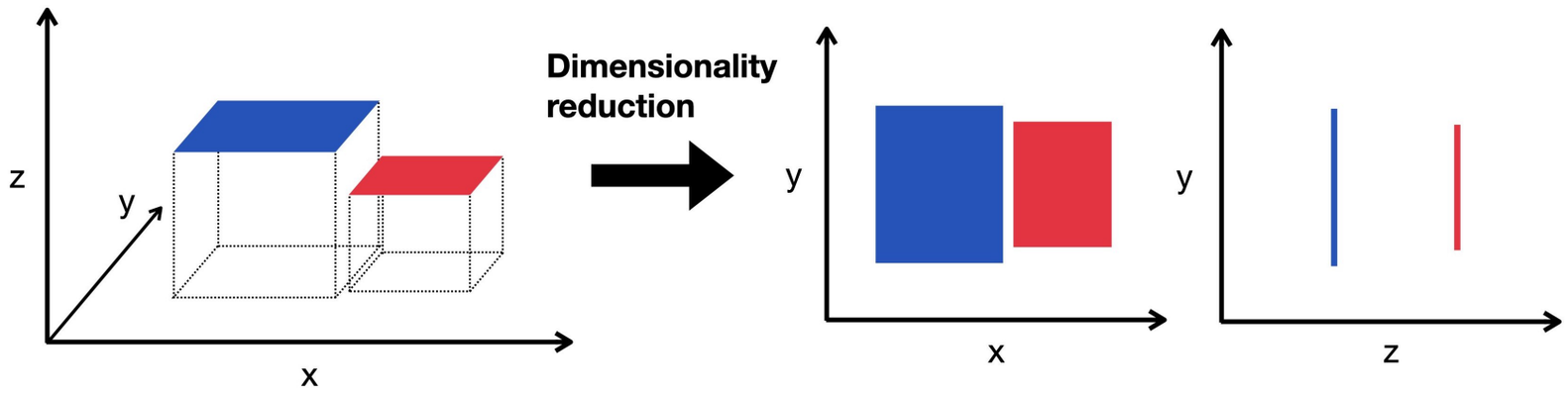
- Dimensionality reduction
- Clustering
- Cell type annotation
- Batch correction and data integration

Roadmap



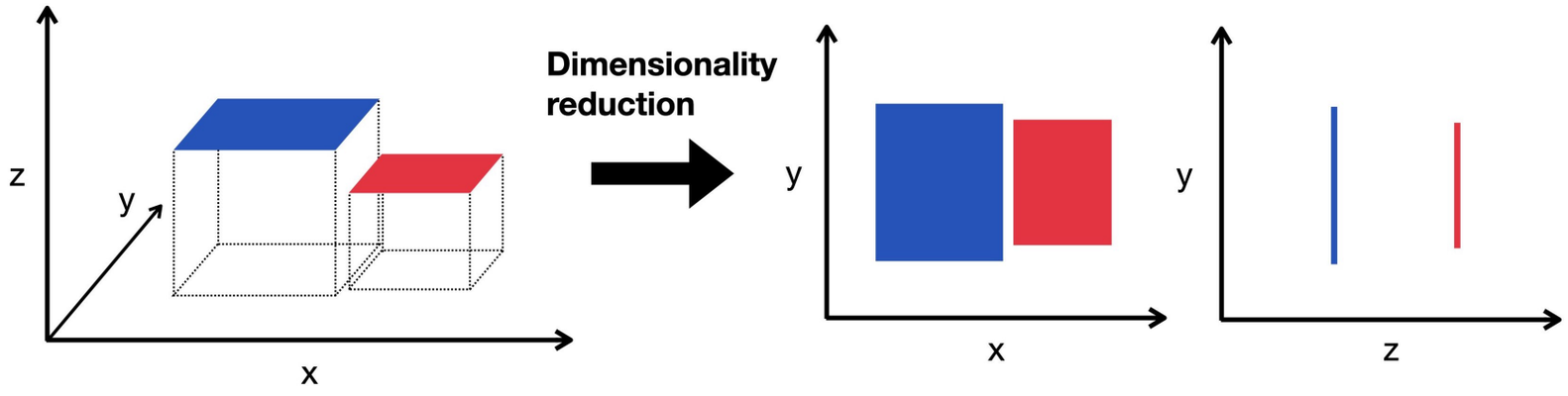
Jovic et al., 2022, Clinical and Translation Medicine

Dimensionality Reduction



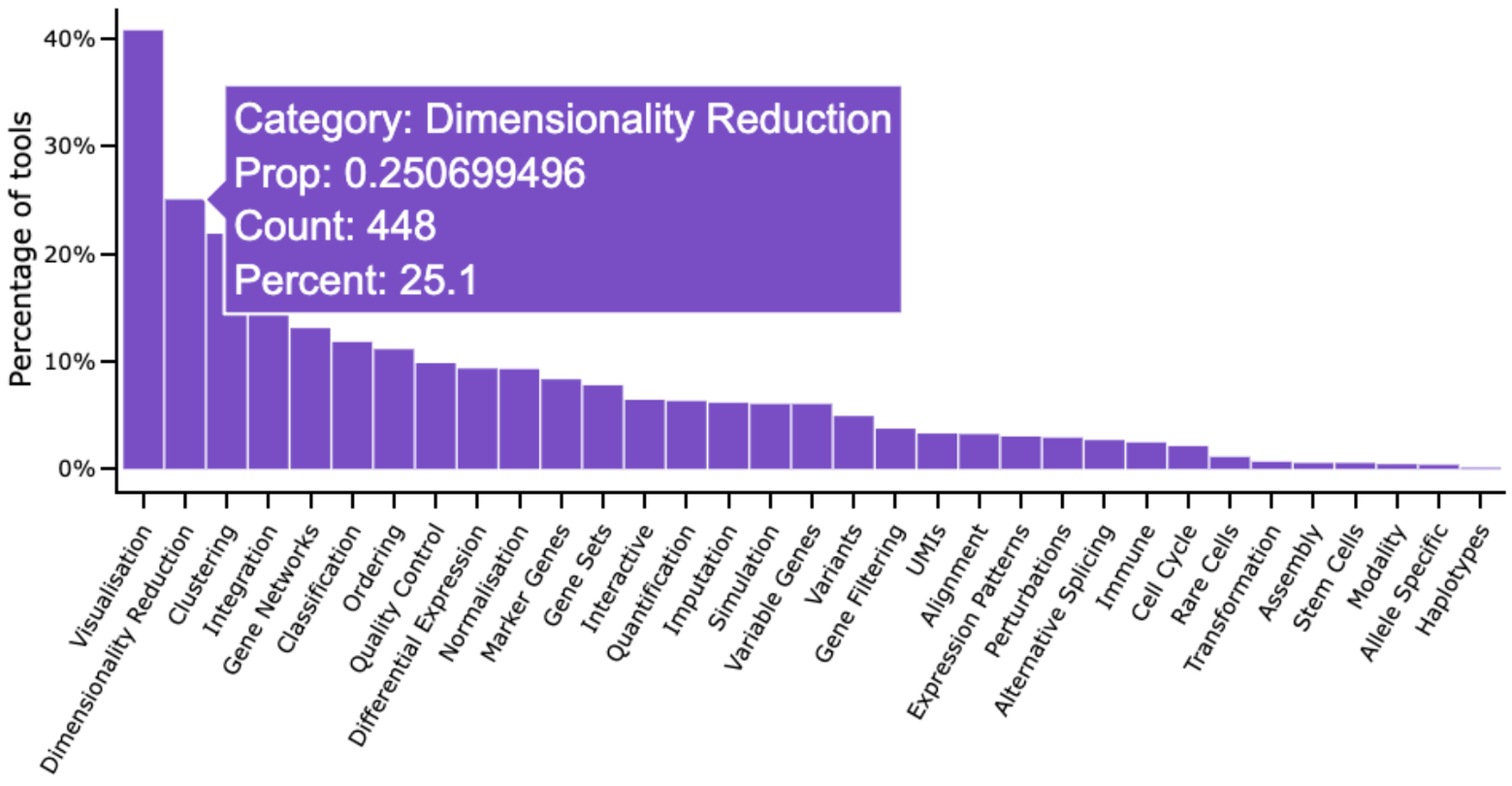
- Simplify complexity, so it becomes easier to work with.
- “Remove” redundancies in the data
- Identify the most relevant information (find and filter noise)
- Reduce computational time for downstream procedures
- Facilitate clustering, since some algorithms struggle with too many dimensions
- Data visualization

Dimensionality Reduction



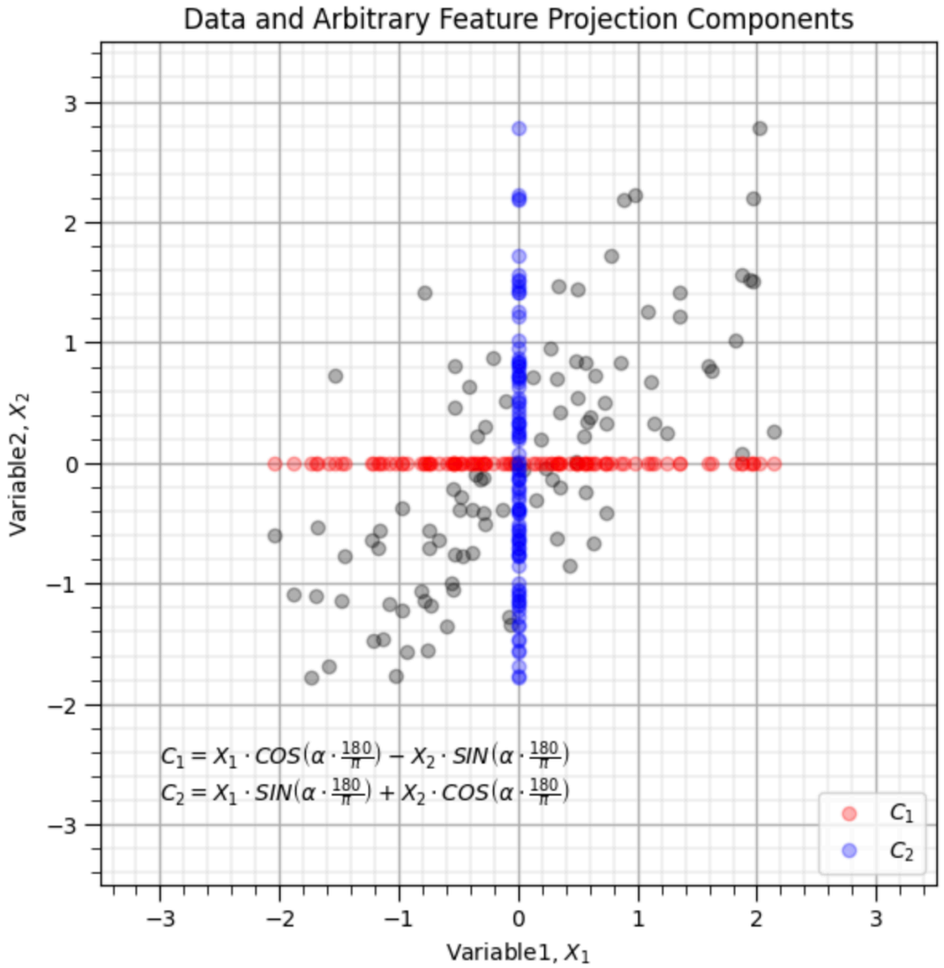
- They are not perfect representation of the high dimension
- One does loose information
- What is close in the projection might actually be far.
- What is far might actually be close
- Conclusions (specially biologically relevant conclusions) should NOT be drawn based on the dimensionality reduction.

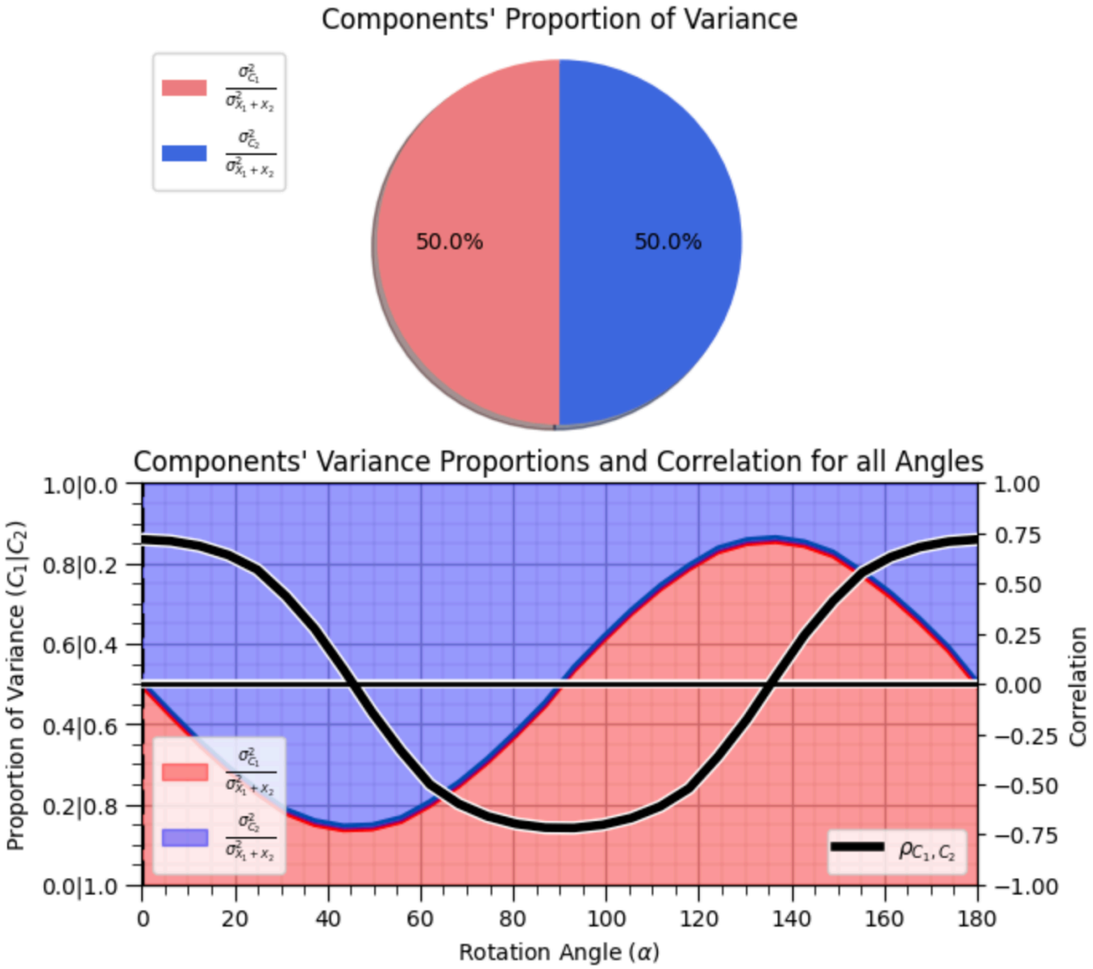
Dimensionality Reduction



https://www.scrna-tools.org/

PCA - Principal Component Analysis

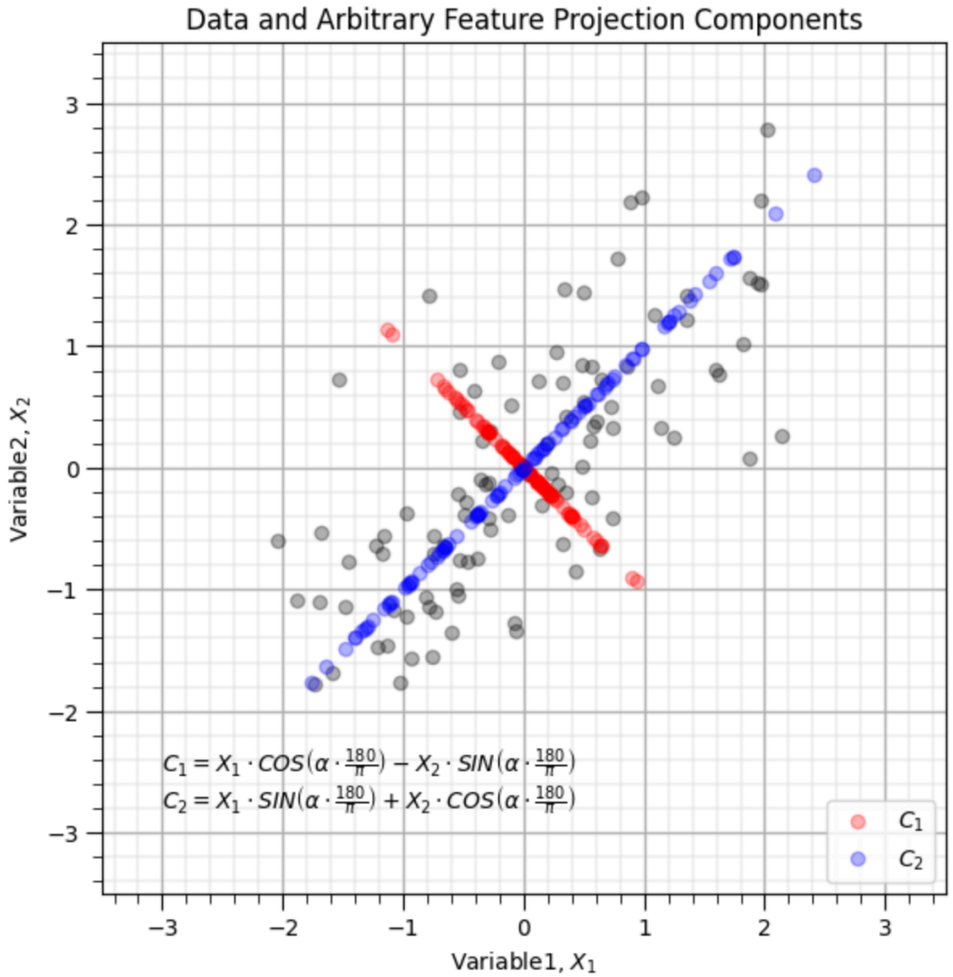


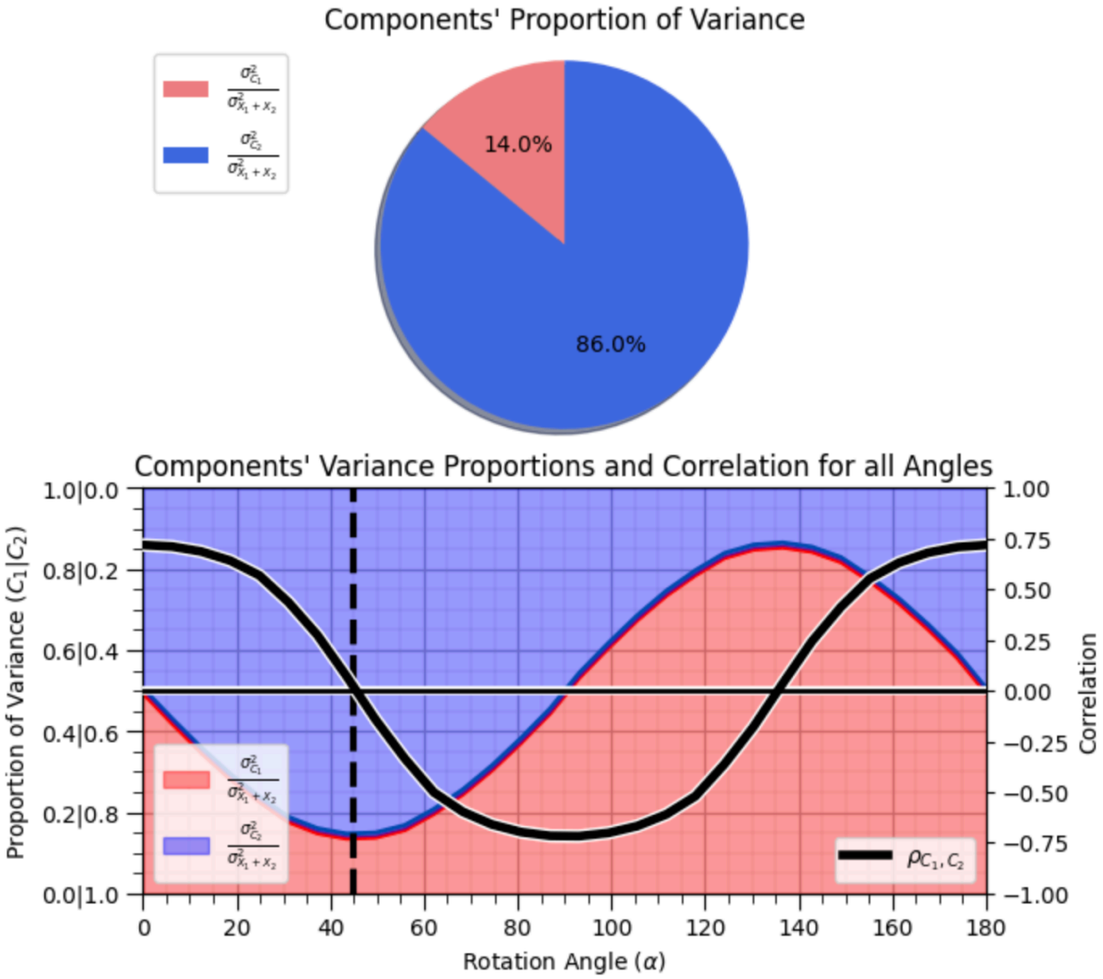


https://github.com/GeostatsGuy/DataScience_Interactive_Python/blob/main/Interactive_PCA_Rotation.ipynb

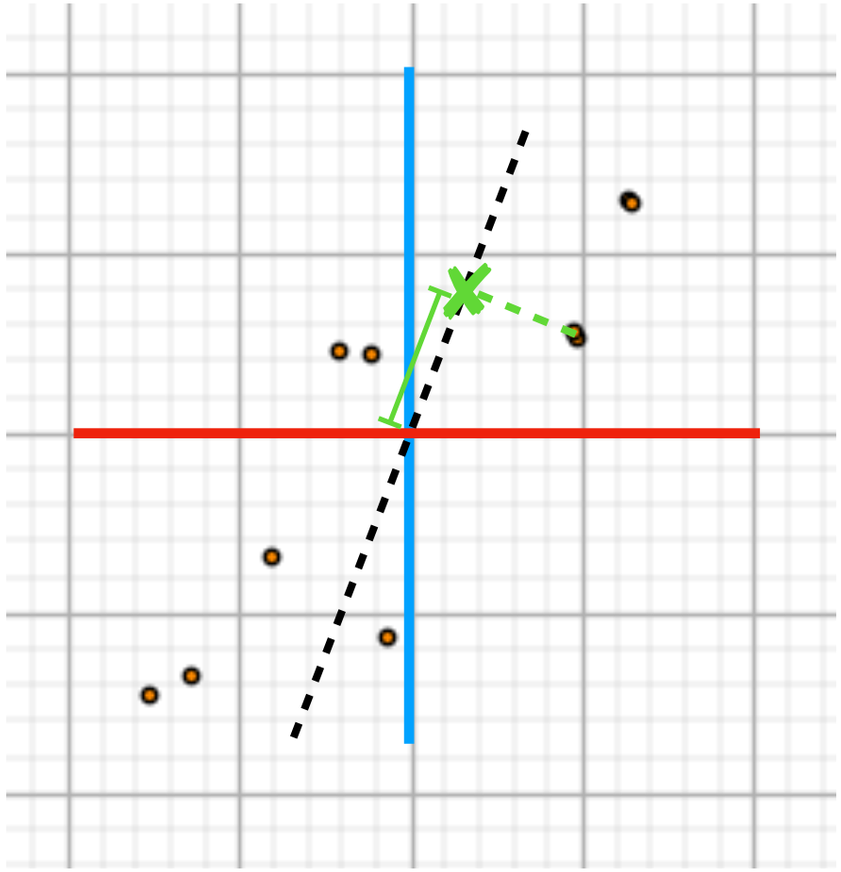
YouTube - StatQuest

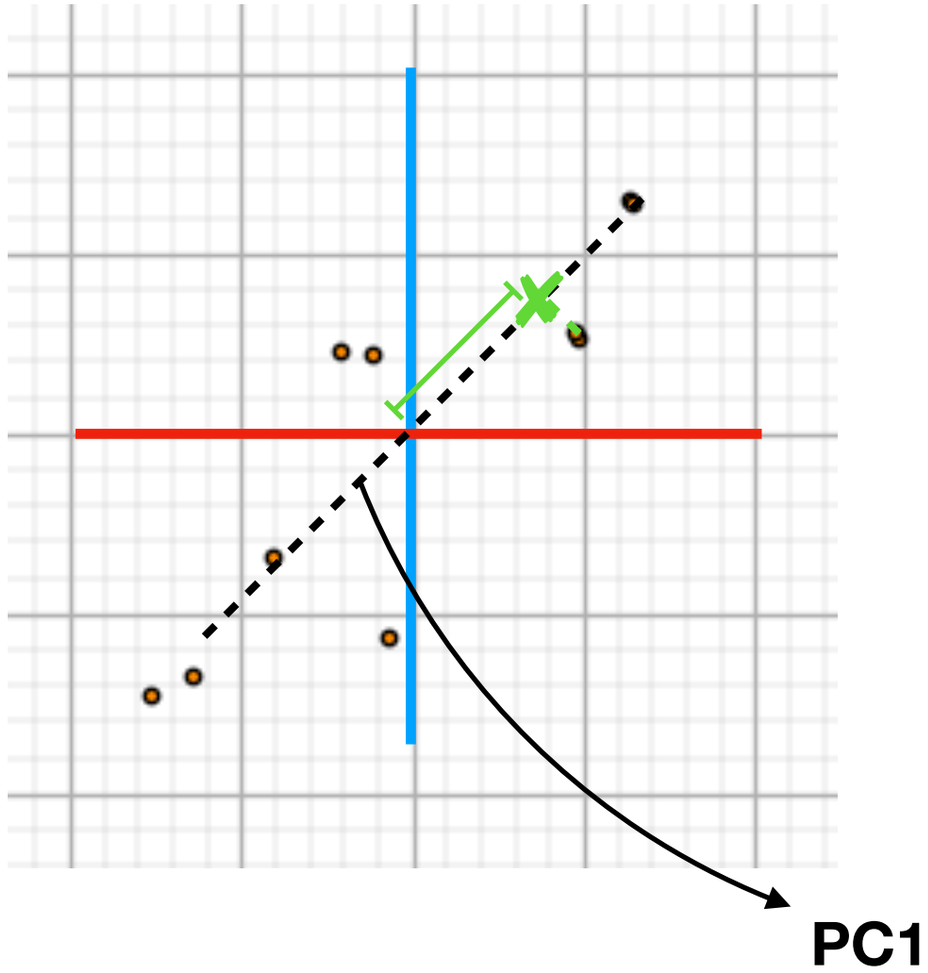
PCA - Principal Component Analysis



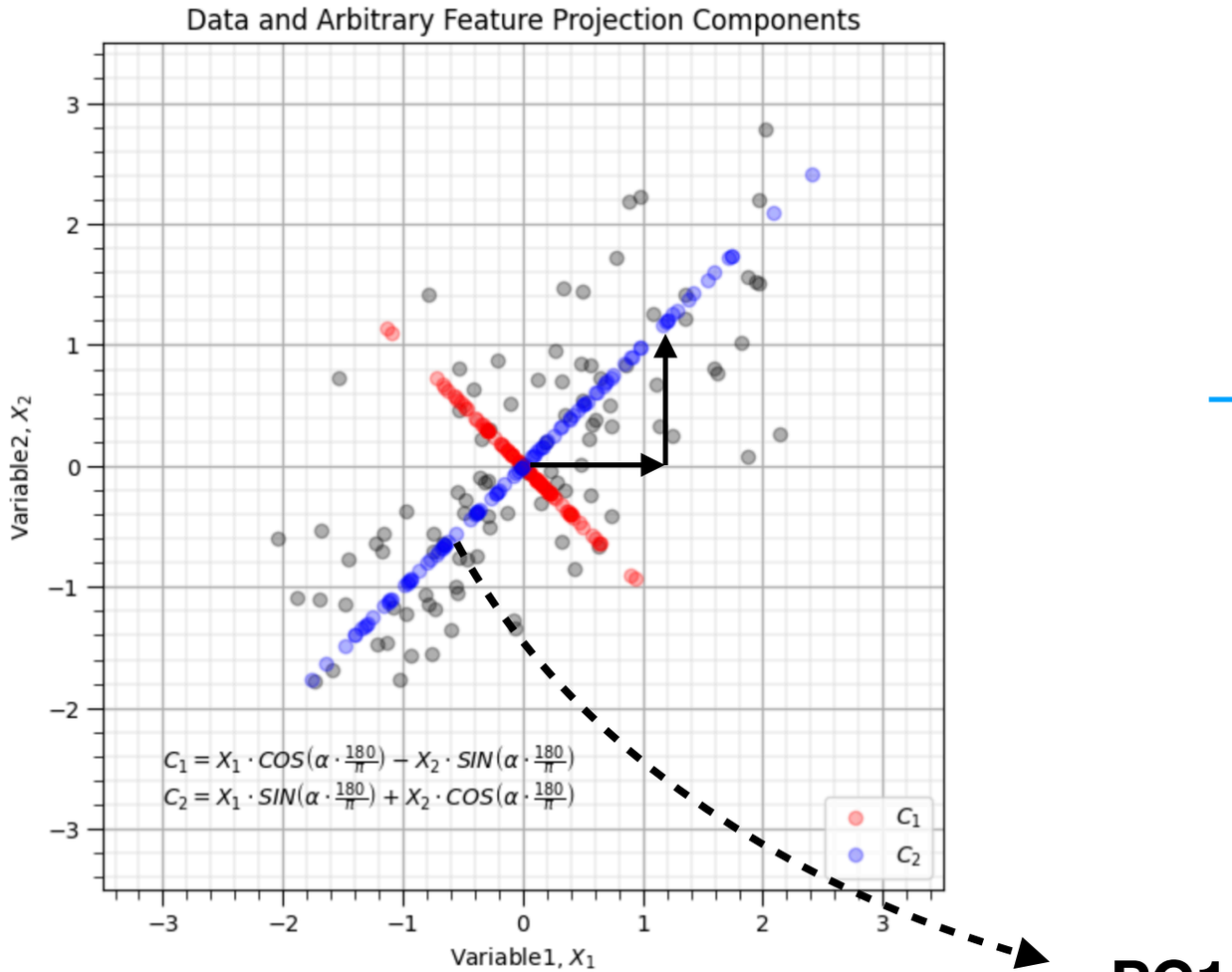


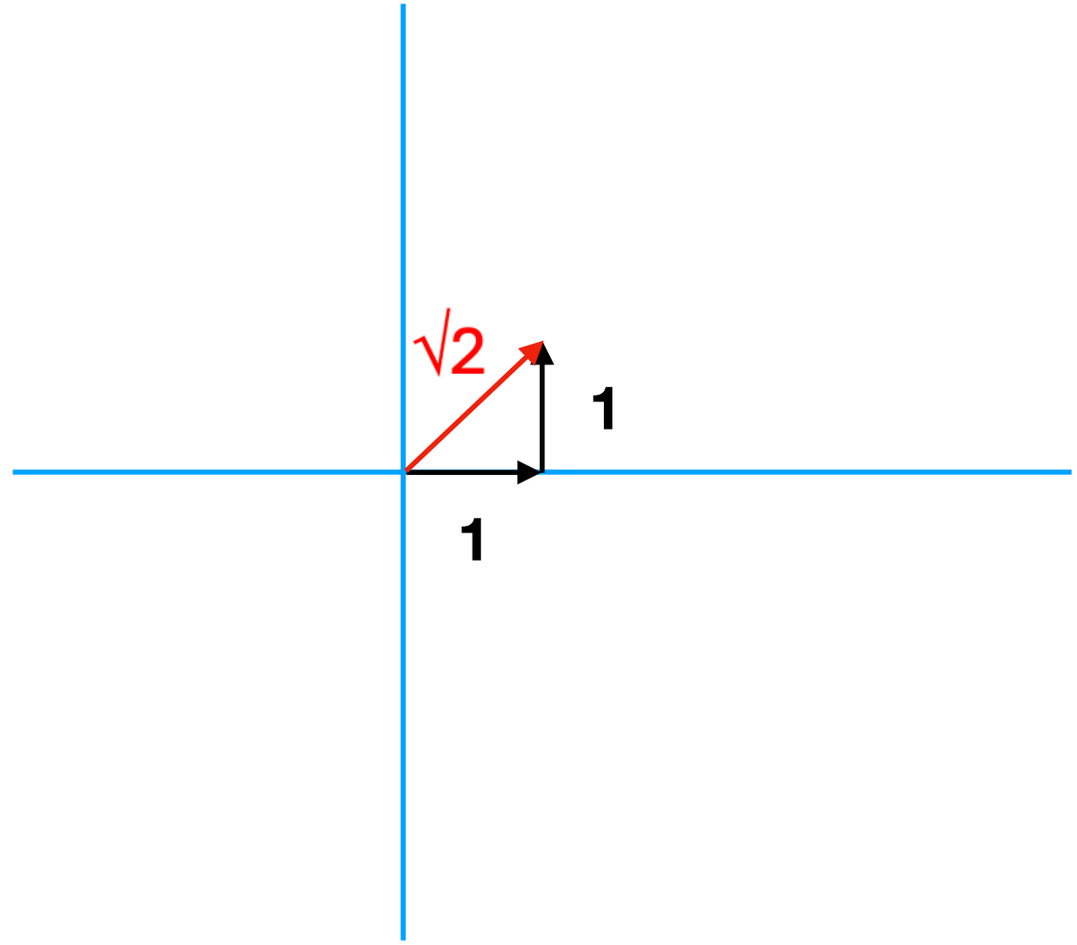
PCA - Principal Component Analysis





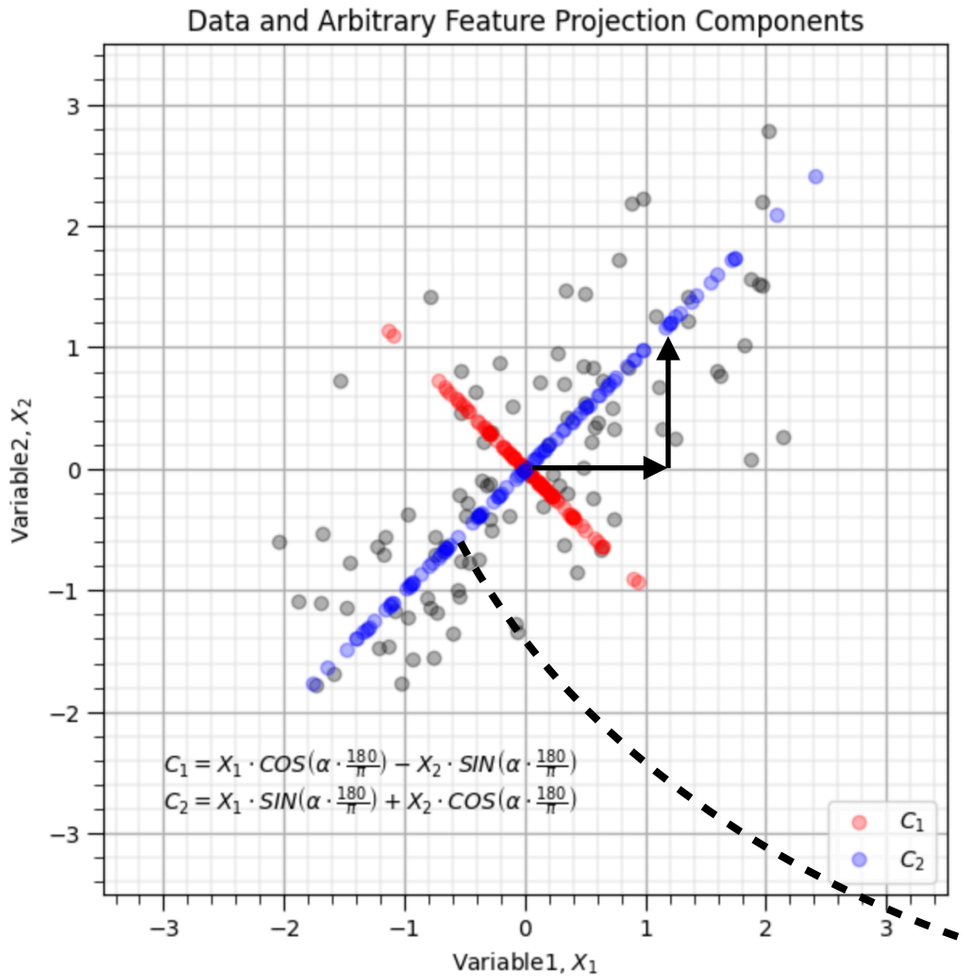
PCA - Principal Component Analysis

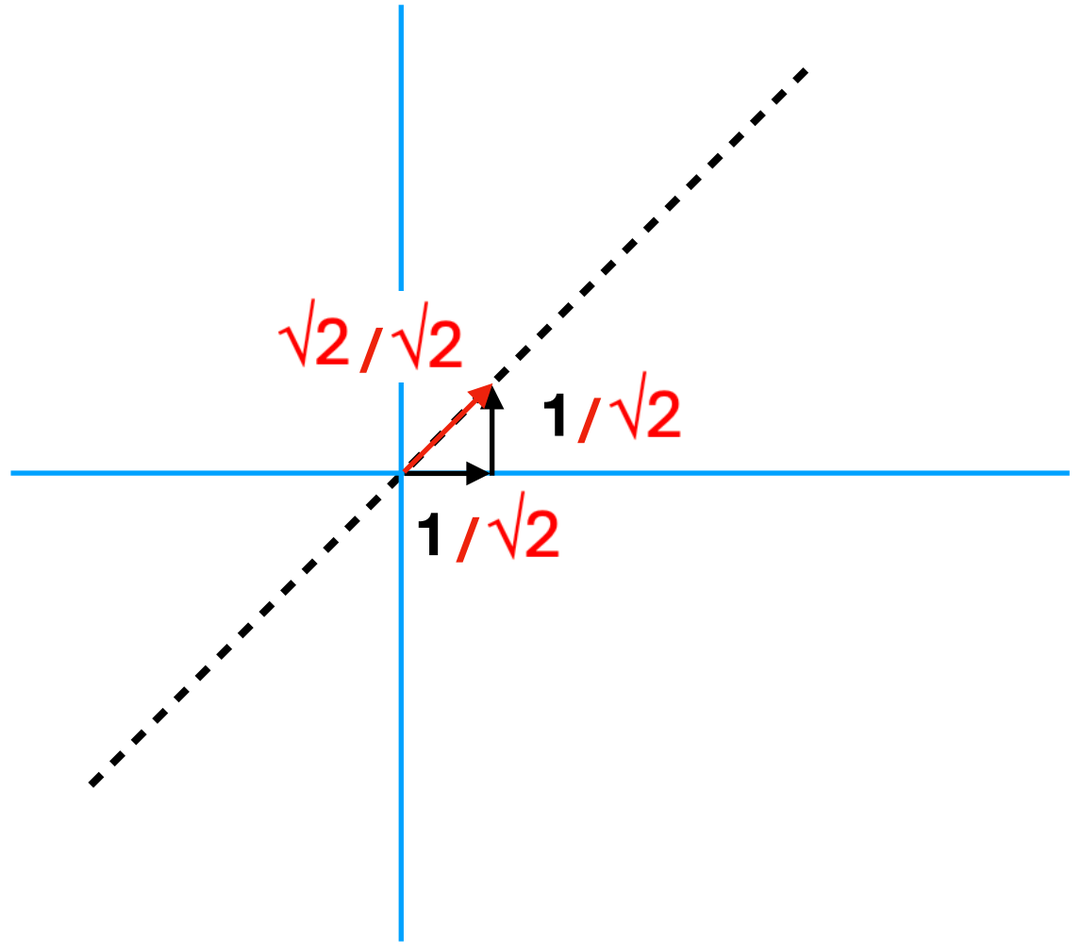




PC1

PCA - Principal Component Analysis

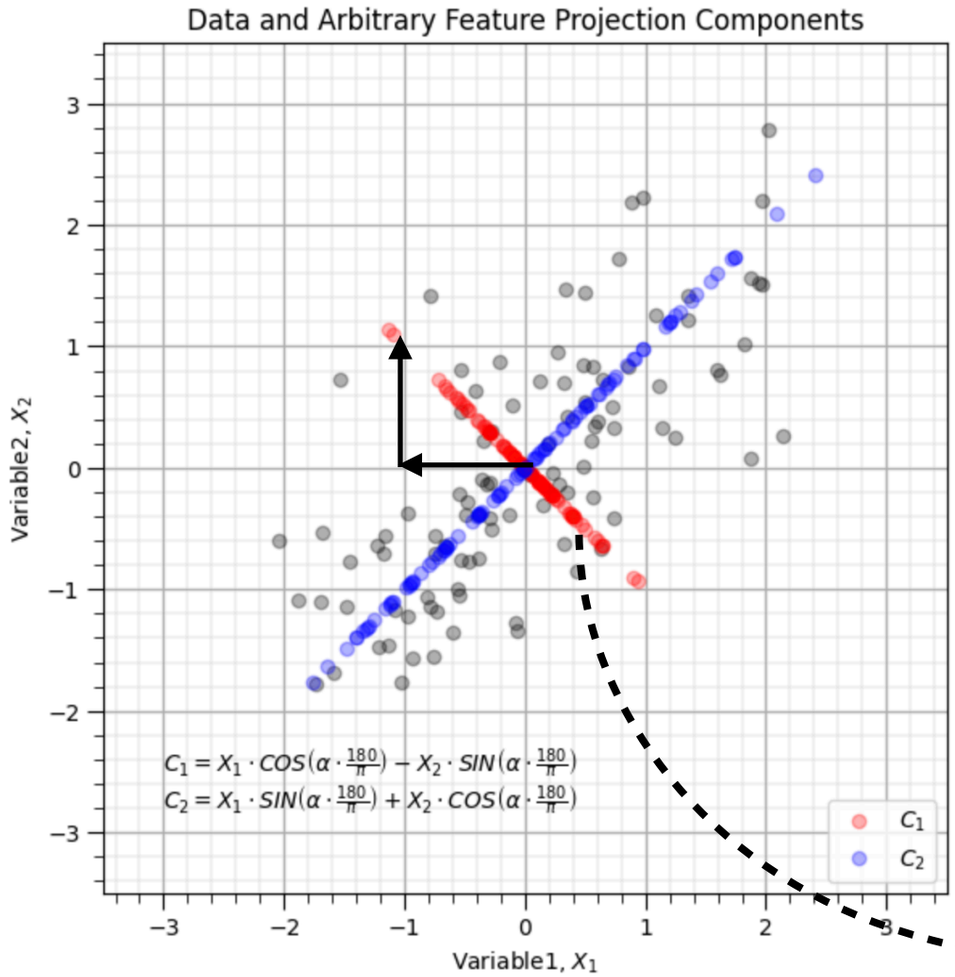


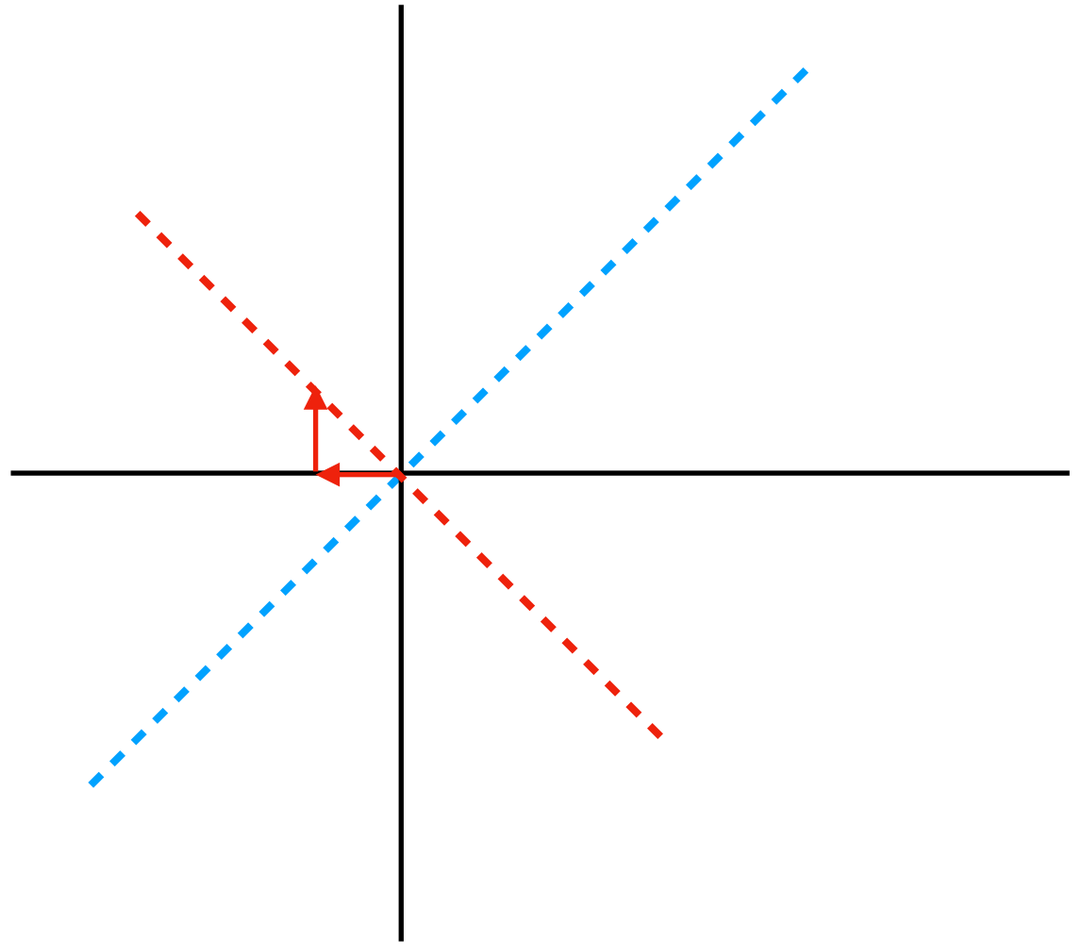


PC1

0.7 x Variable 1 + 0.7 x Variable 2

PCA - Principal Component Analysis





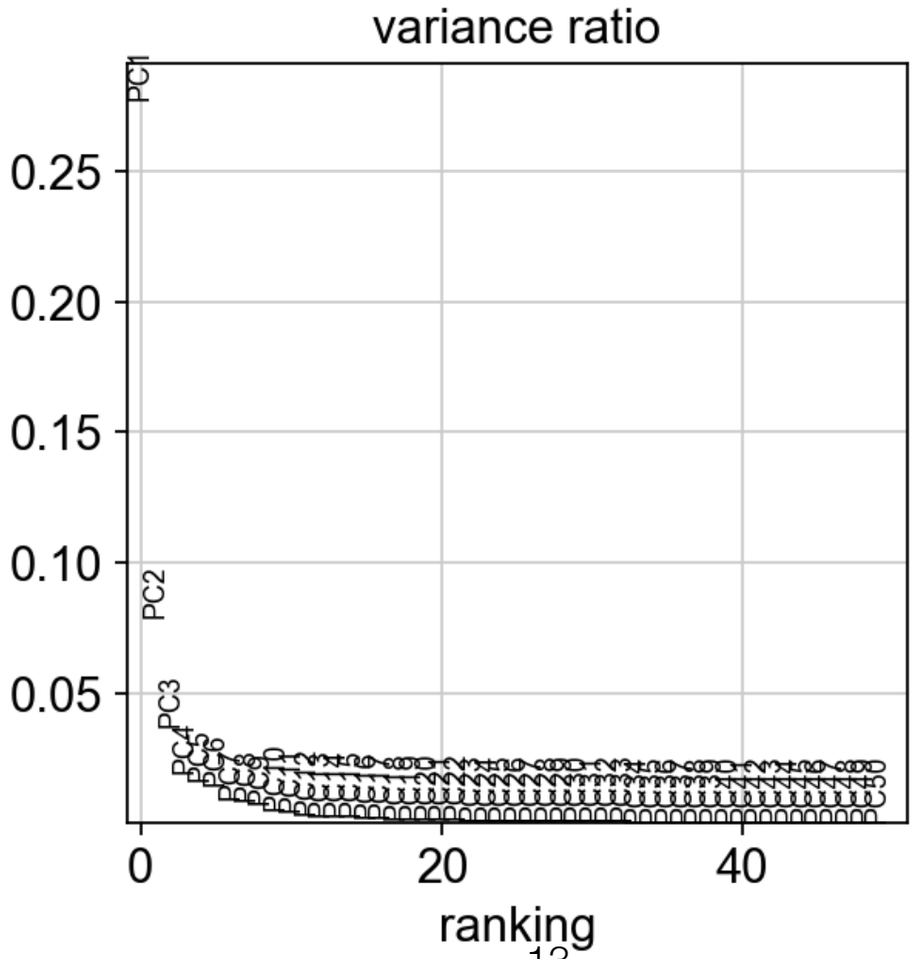
PC2

-0.7 x Variable 1 + 0.7 x Variable 2

PCA - Principal Component Analysis

How many PCs to include for the downstream analysis

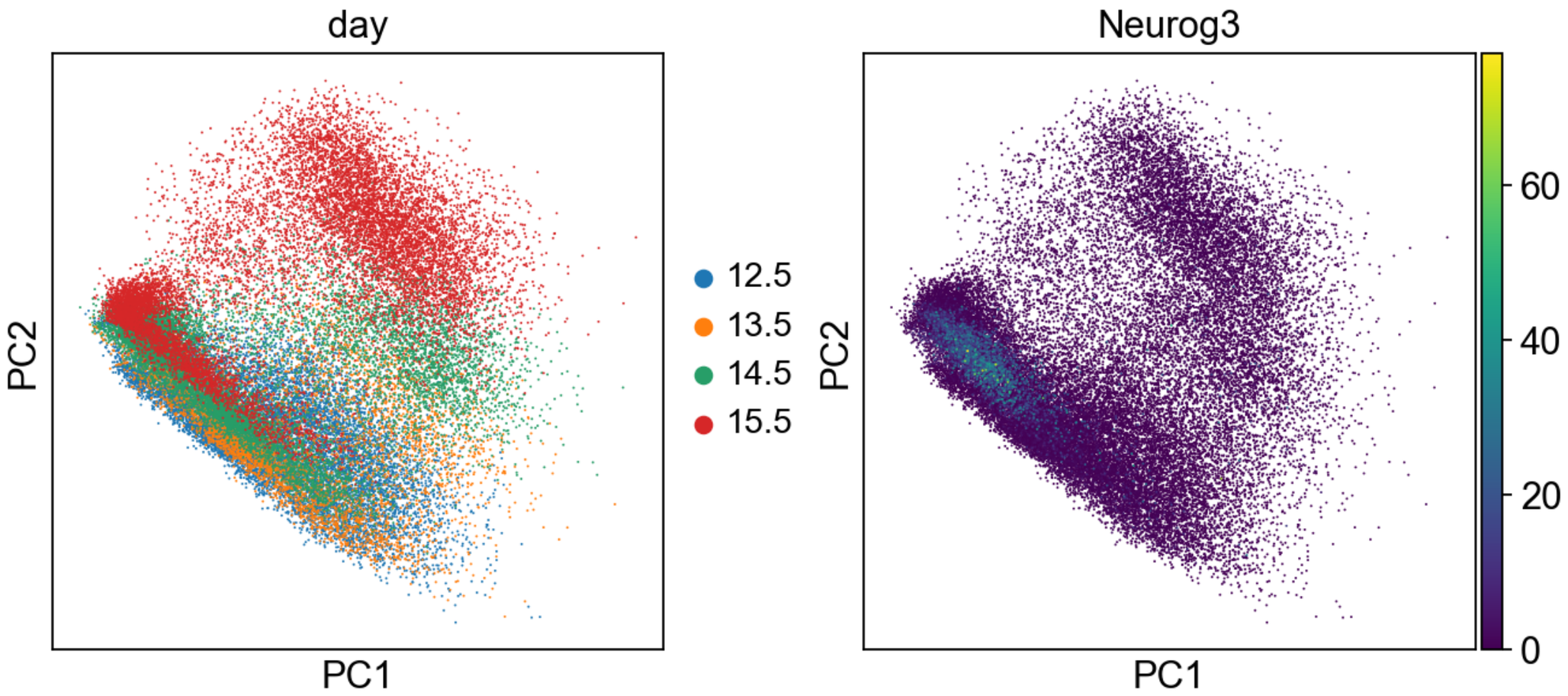
- PCs that explain at least 1% of variance
- Jackstraw of significant p-values
- The first 5-10 PCs
- Elbow plot / scree plot
- Scater library describes correlation between PCs and metadata, take PCs until the metadata information is covered.



PCA - Principal Component Analysis

- The PC are linear combination of the original axis.
- The estimated parameters of the linear combination is known and therefore we can know positively or negatively how much it goes into one direction or the other one.
- Indeed as the original axis are g1,g2,g3… and the new axis are a1g1 +a2g2…, one takes the ai that are the highest, positively and negatively and therefore knows which genes are mostly representing the axis you see.

PCA is not enough!



The genes that are associated with PC1 positively and negatively correspond to

Correlation between the PC and the gene’s expression

Are the genes with highest and lowest values calculated for the rotation matrix(loadings)

Are differentially expressed genes between PC1 and the others

Which dimensionality reduction method do you used/are you used to?

Dimensionality Reduction

- Xiang et al. compared the stability, accuracy and computing cost of 10 different dimensionality reduction methods

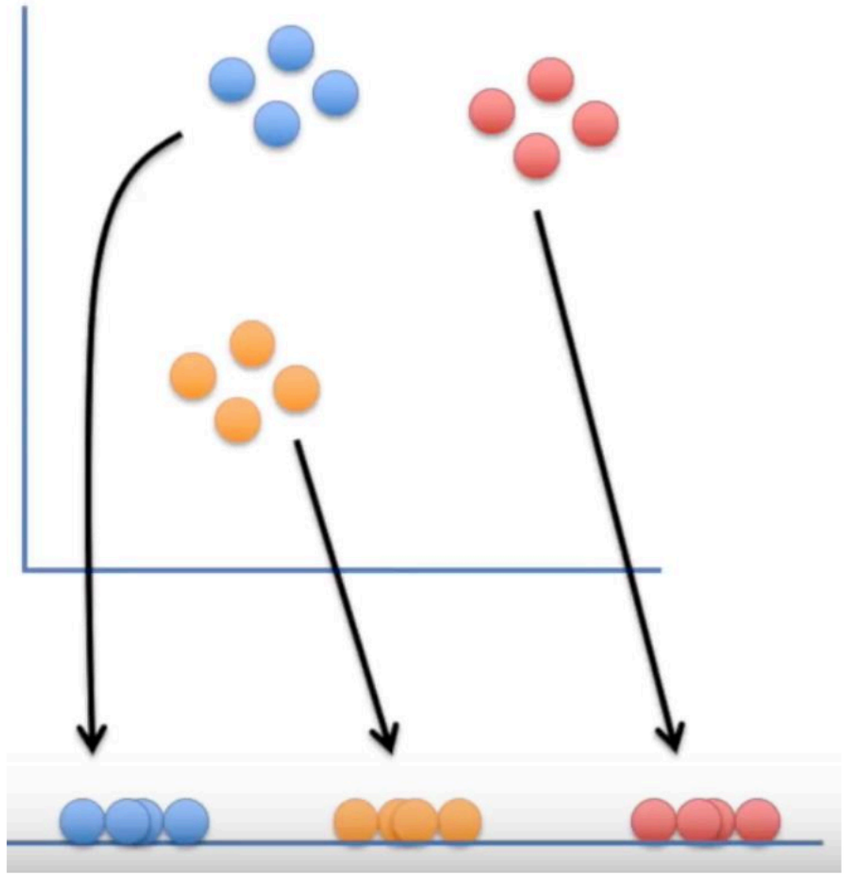
- t-distributed stochastic neighbor embedding (t-SNE) yielded the best overall performance

- Uniform manifold approximation and projection (UMAP) showed the highest stability and separates best the original cell populations

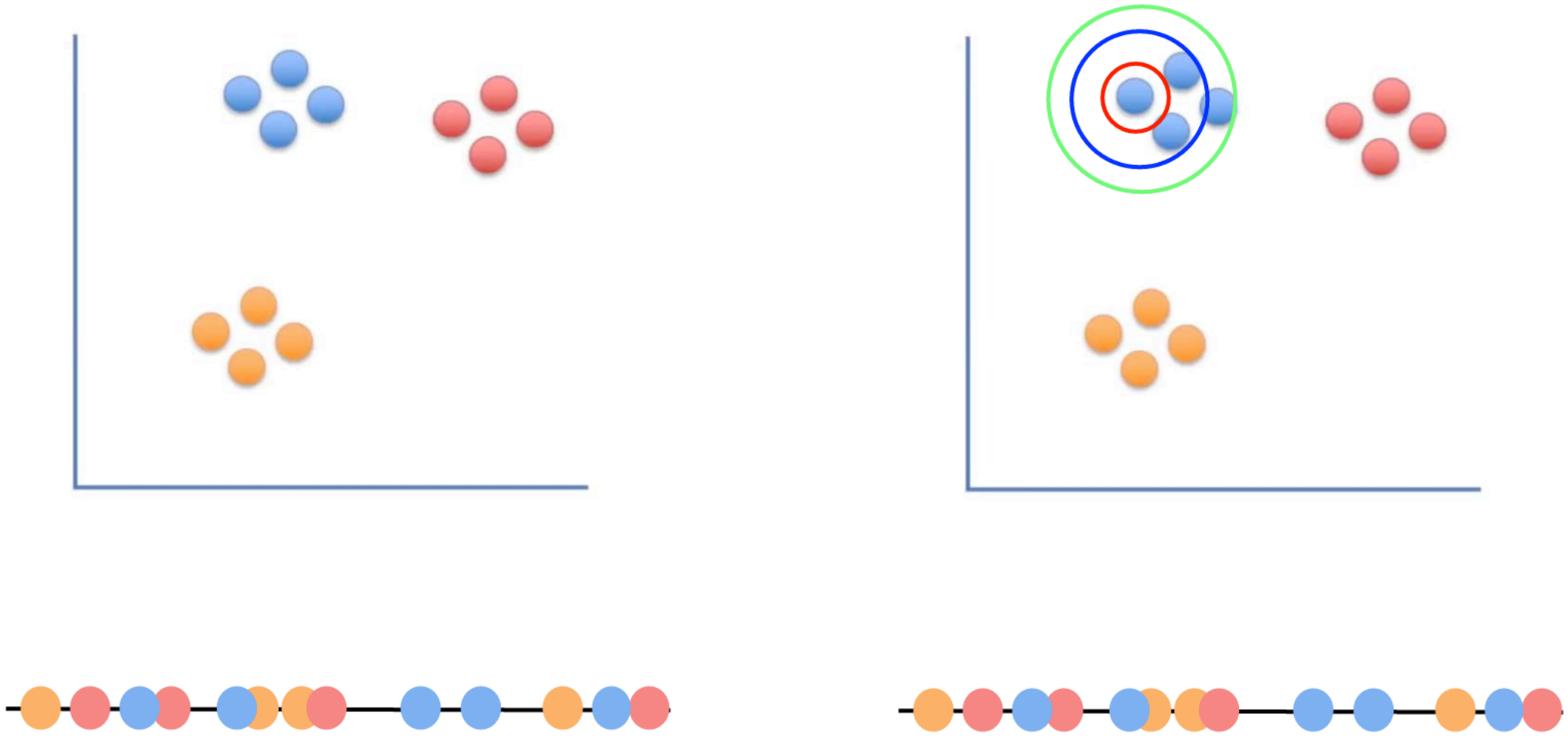
Xiang et al., 2021

t-SNE - t-distributed stochastic neighborhood embedding

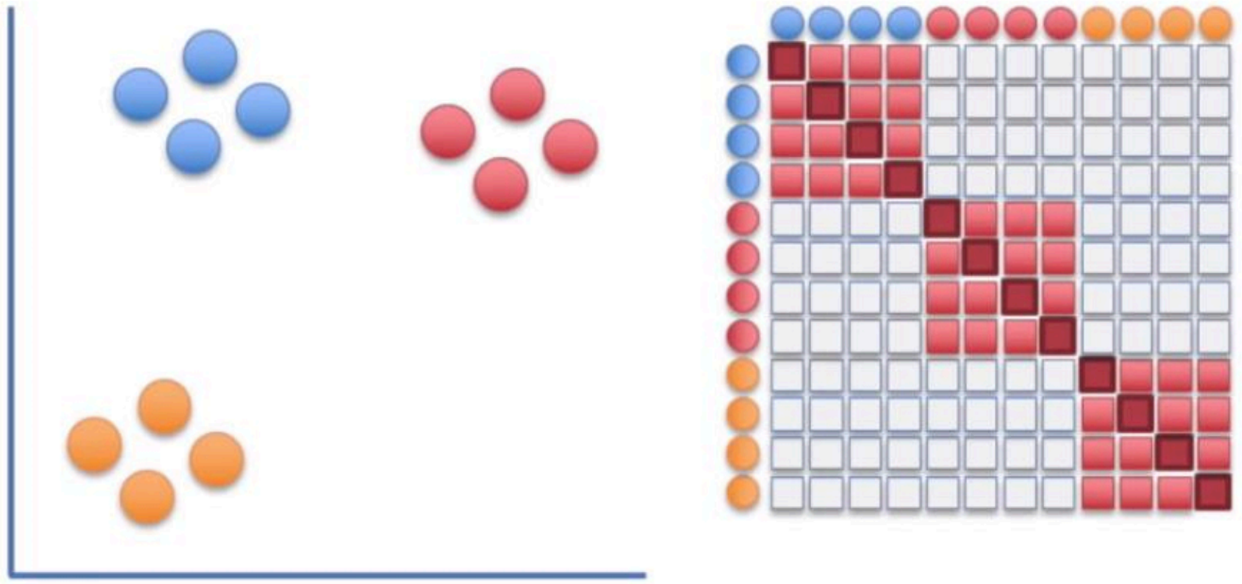
Find a right way to reduce dimension while keeping all the "clusters"



t-SNE

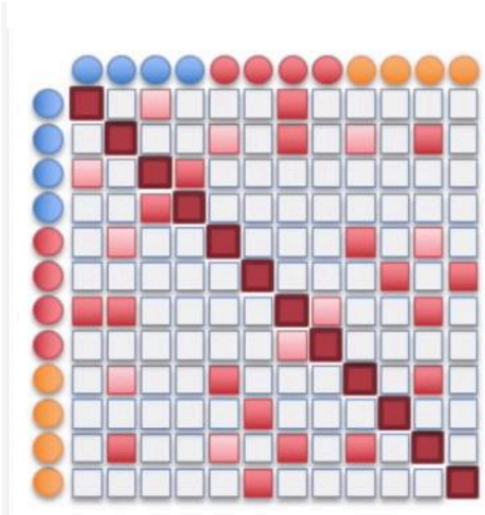


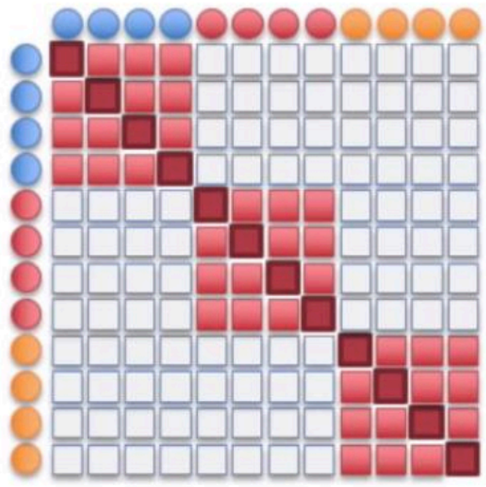
t-SNE

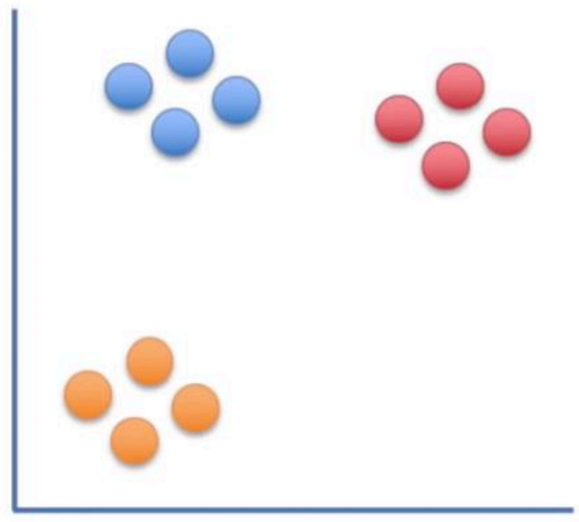


t-SNE

Move points little by little and redo calculation until you are « as close as possible » to the original similarity matrix or you reach a certain number of iteration (chosen by the user).





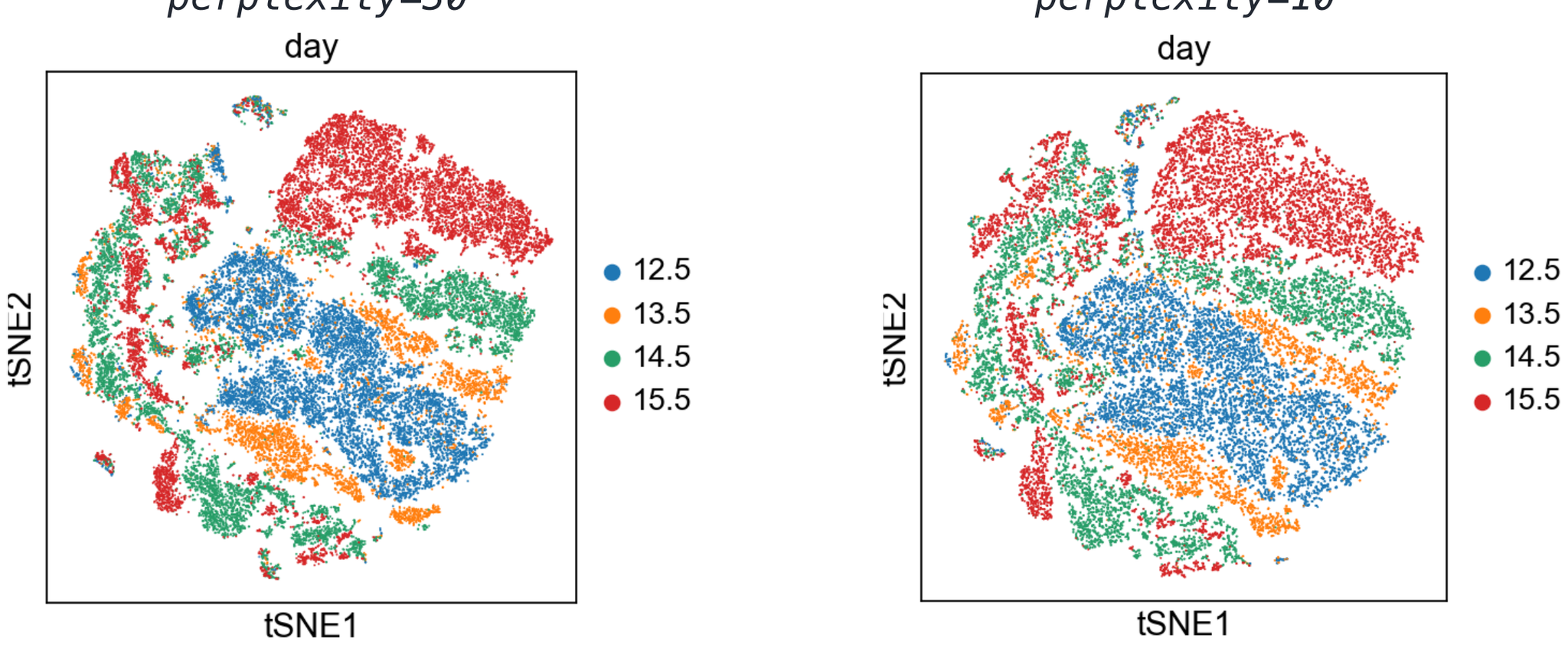


t-SNE

What parameters to consider

- Perplexity 30L => linked to parameter σ of all the points

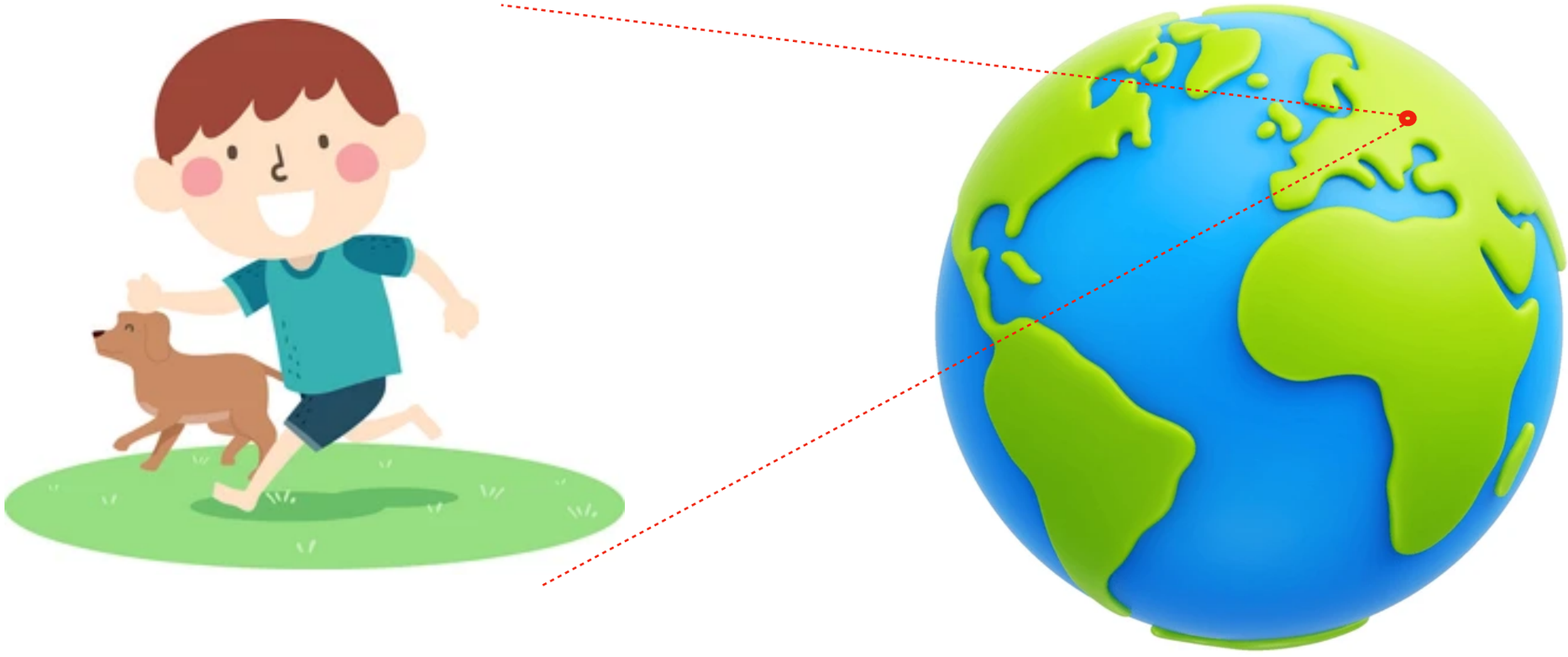
```
sc.tl.tsne(adata, perplexity=30, …)
```



UMAP - Uniform Manifold Approximation and Projection

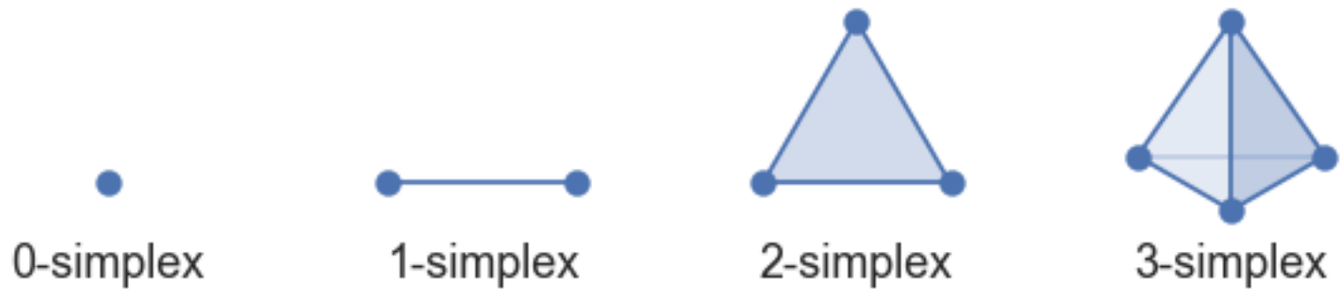
A non-linear graph-based method

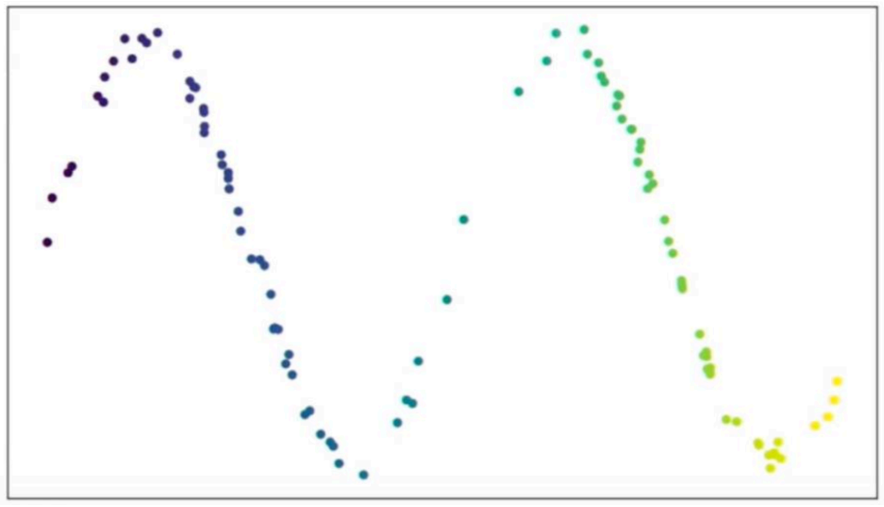
UMAP operates under the assumption that the data lie on a manifold.

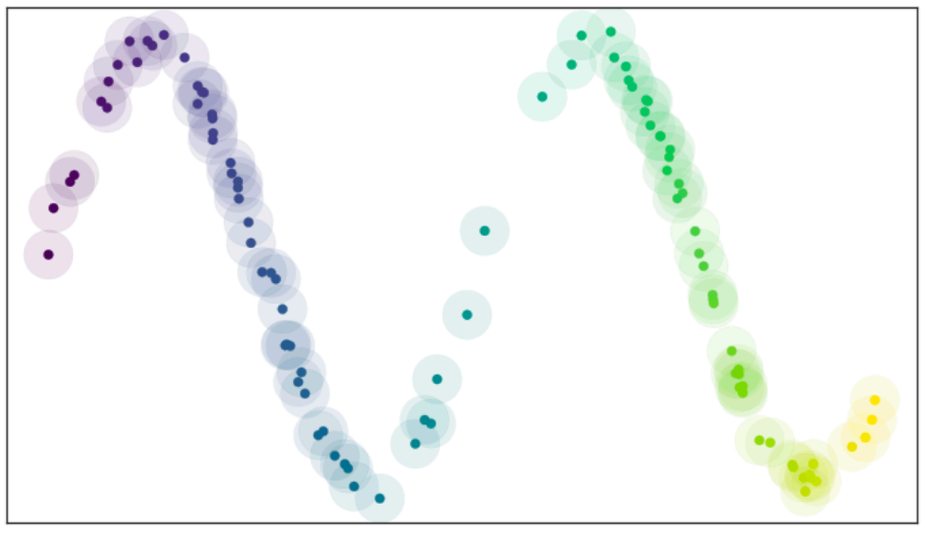


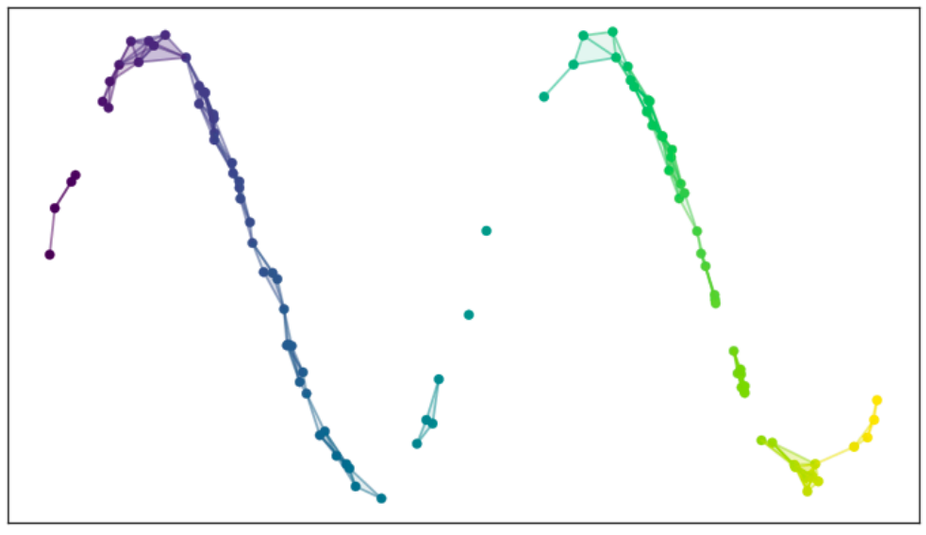
https://arxiv.org/abs/1802.03426

UMAP



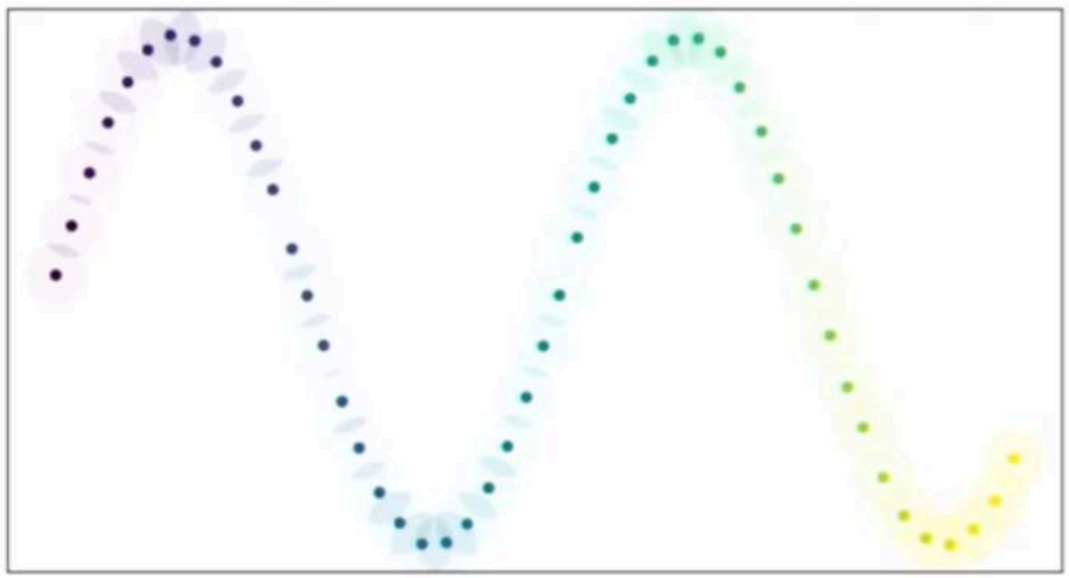






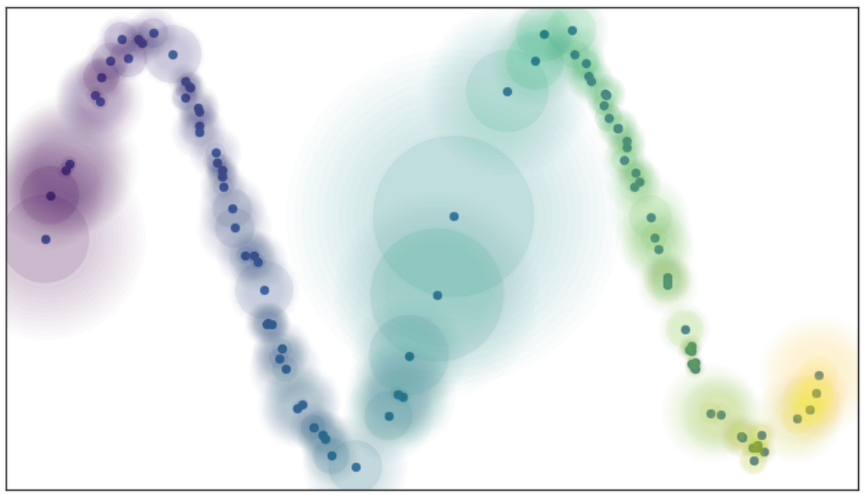
https://arxiv.org/abs/1802.03426

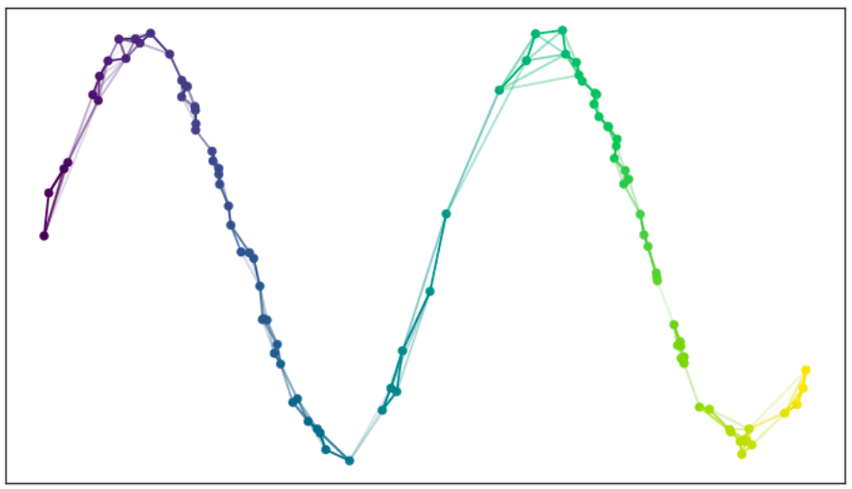
UMAP



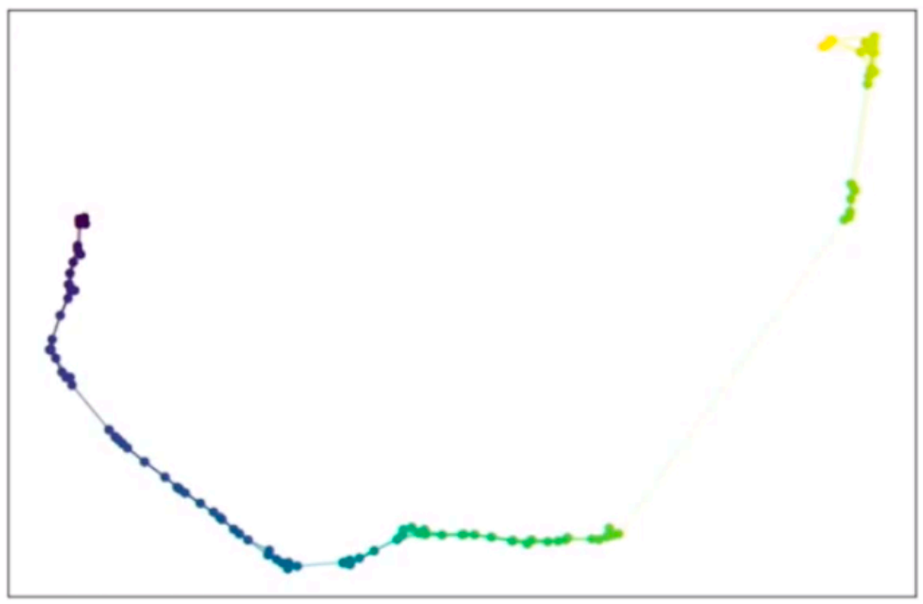
UMAP

Fuzzy topology





UMAP



UMAP

- The first phase is constructing a fuzzy topological representation (edges and weights)
- The second phase is optimizing the low dimensional representation

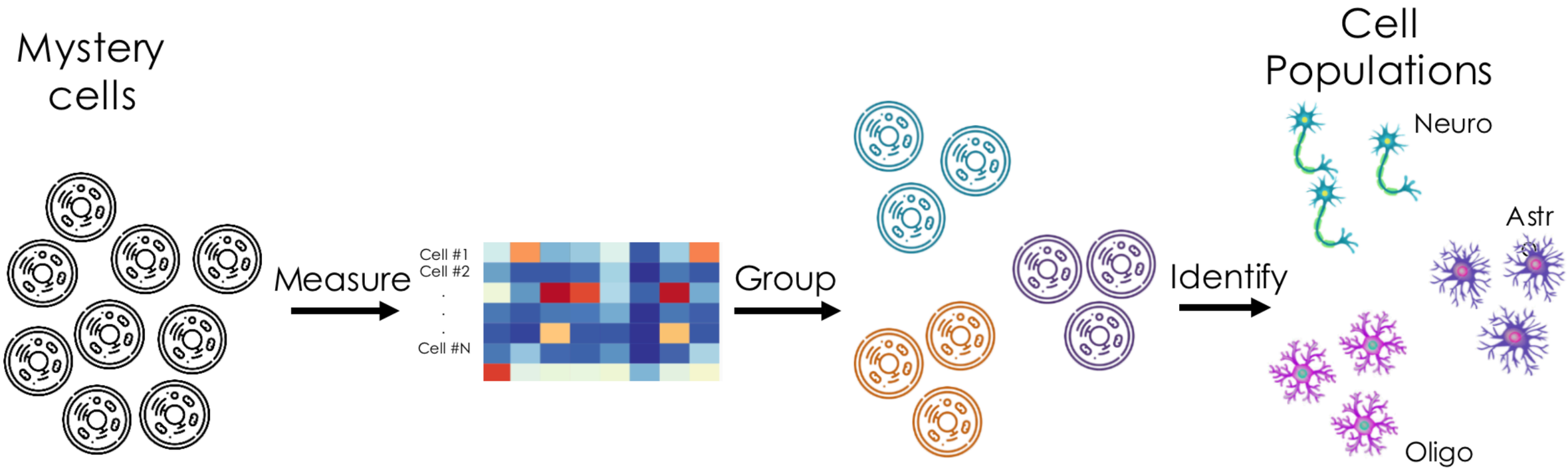
UMAP

what parameters to consider

```
sc.pp.neighbors(adata, n_neighbors, …)
```

```
sc.tl.umap(adata, min_dist, …)
```

Clustering



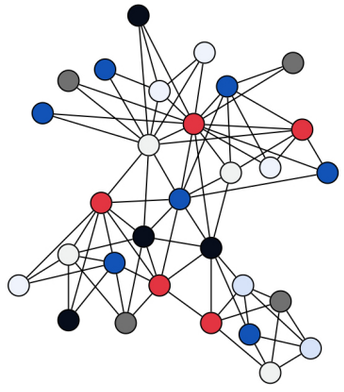
https://sib-swiss.github.io/

Clustering

Graph-based methods

First step) Generating a k-Nearest Neighbor (kNN) graph:
A graph in which two vertices p and q are connected by an edge, if the distance between p and q is among the k-th smallest distances from p to other objects from P.

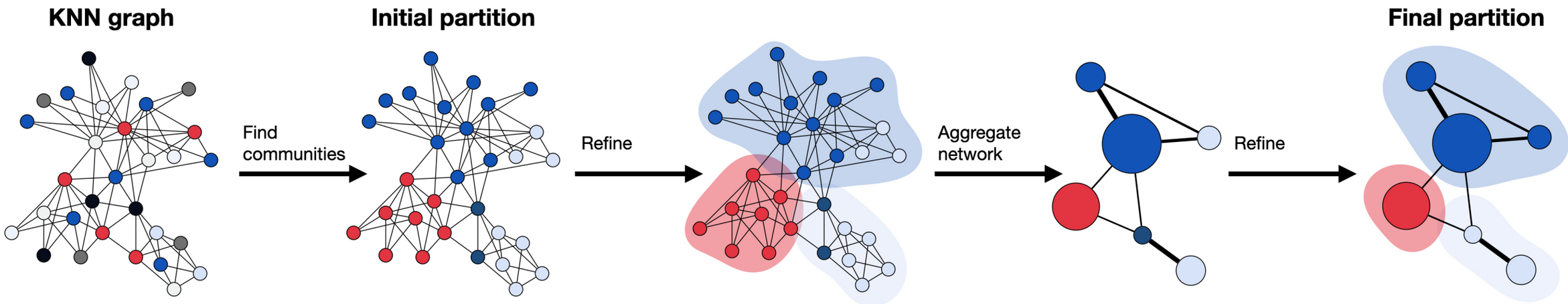
KNN graph



Clustering

Second step) Community detection (e.g. Louvain and Leiden)
Find a group of nodes with more edges inside the group than edges linking nodes of the group with the rest of the graph

The resolution parameter




```
scanpy.tl.leiden(adata, resolution=, …)
```

https://www.sc-best-practices.org/

Clustering Considerations

- The definition of cell type within your project
- The number of different cell types you expect
- The biological meaning of each cluster - check if any cluster is dominated by the low quality cells
- Clustering is subjective – No ground truth
- The stability of clusters

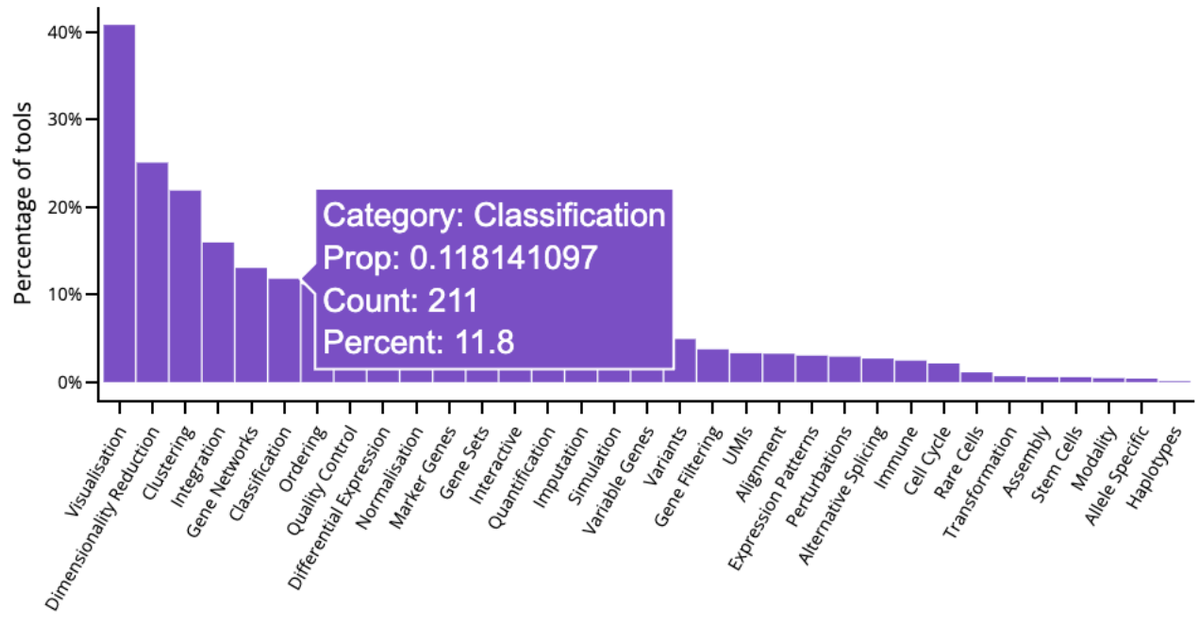
Cell-type Annotation

Manual annotation

- Based on a single or small set of marker genes
- Based on highly expressed genes in each cluster

Automated annotation

- Classifiers
- Reference mapping

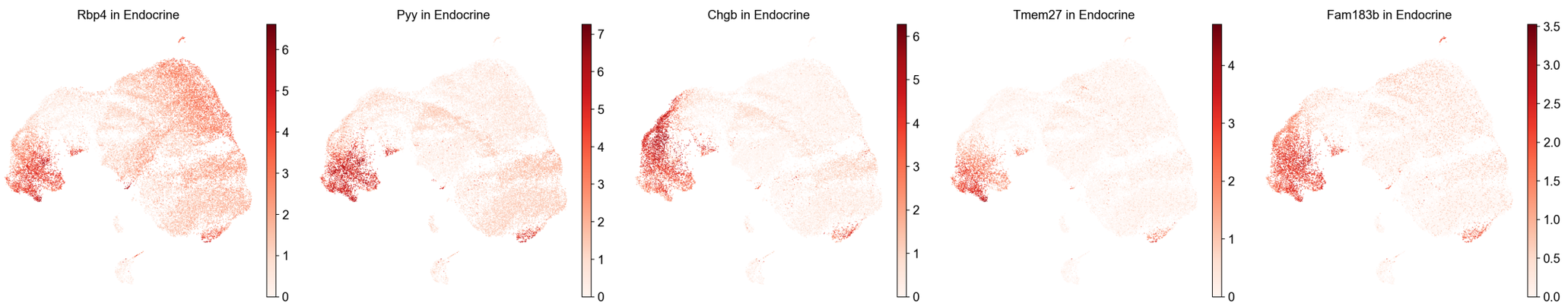


Cell-type Annotation

Manual annotation

- Based on a single or small set of marker genes

```
marker_genes = {
    "Multipotent": ["Dlk1", "Mdk"],
    "Tip": ["Vtn", "Myc", "Jam3"],
    "Trunk":["Notch2", "Cbx3"],
    "Acinar":["Cpa1", "Cel", "Rbpjl", "Reep5"],
"Ductal":["Sox9", "Anxa2", "Spp1"],
"EP":["Neurog3", "Hes6", "Btbd17", "Gadd45a"],
"Fev+":["Fev", "Cck", "Neurod1", "Vwa5b2", "Tox3"],
"Endocrine":["Rbp4", "Pyy", "Chgb", "Tmem27", "Fam183b"]}
```



Cell-type Annotation

Manual annotation

- Based on highly expressed genes in each cluster