Dimensionality Reduction and UMAP

Melissa Sanabria, TU Dresden
melissa.sanabria@tu-dresden.de

Principal Component Analysis (PCA)

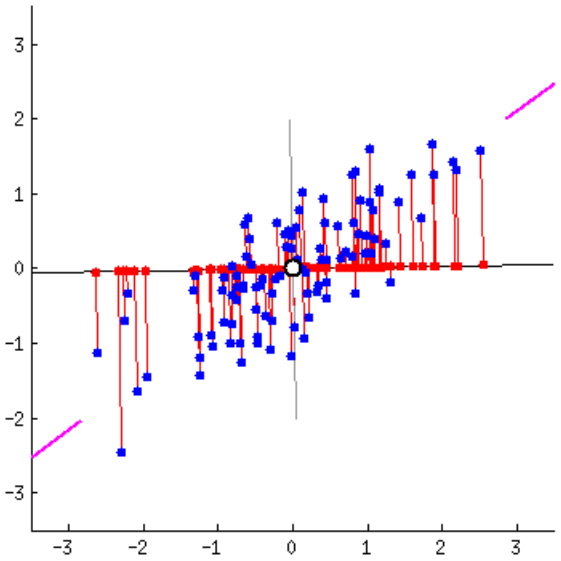
- PCA builds linear projections of data into a new coordinate system

- The coordinate system is chosen and ranked by the variance it explains in the data

- Usually the first principle components, the ones with the highest variance explained are shown

- How much variance they explain is indicative of how well the dimensionality reduction has worked

Principal Component Analysis (PCA)

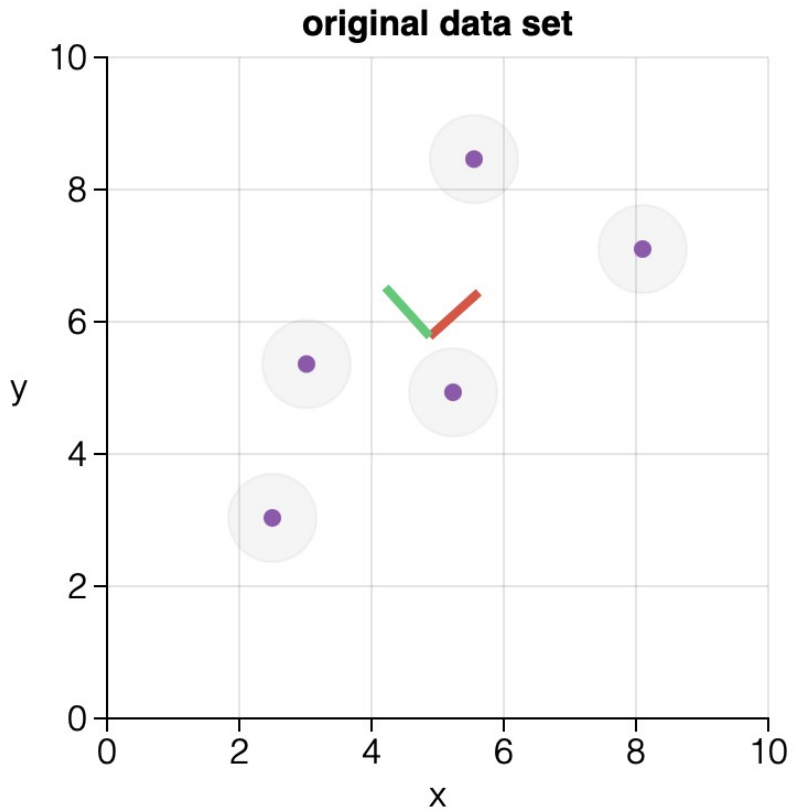


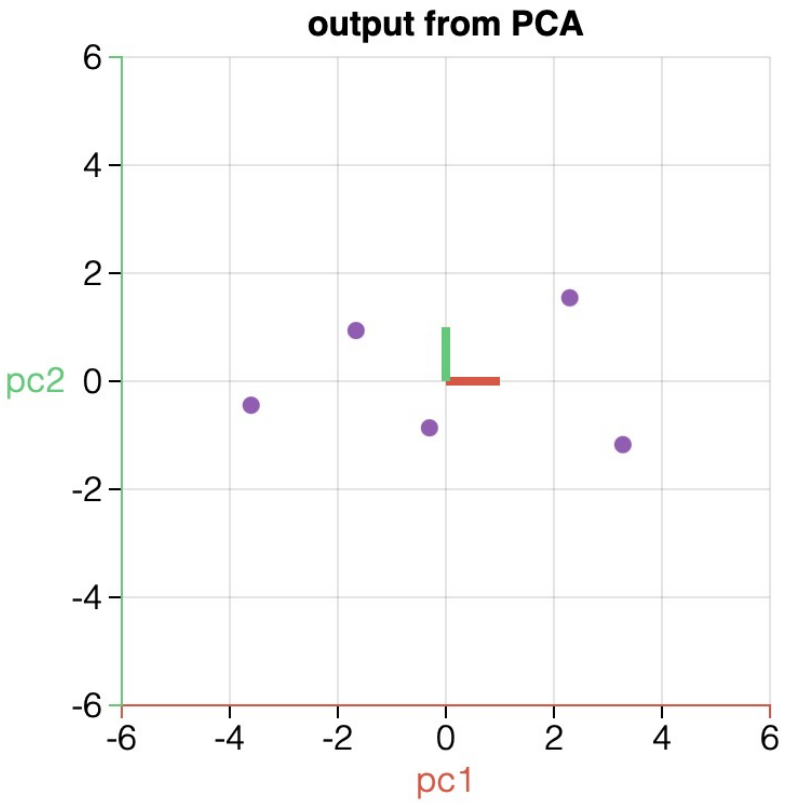
First PC accounts for the **largest possible variance** in the data set.

Line that matches the purple marks because it goes through the origin and it's the line in which the projection of the points (red dots) is the most spread out. Or mathematically speaking, it's the line that maximizes the variance (the average of the squared distances from the projected points (red dots) to the origin).

The second PC is calculated in the same way, with the condition that it is orthogonal to the first PC and that it accounts for the next highest variance.

Principal Component Analysis (PCA)





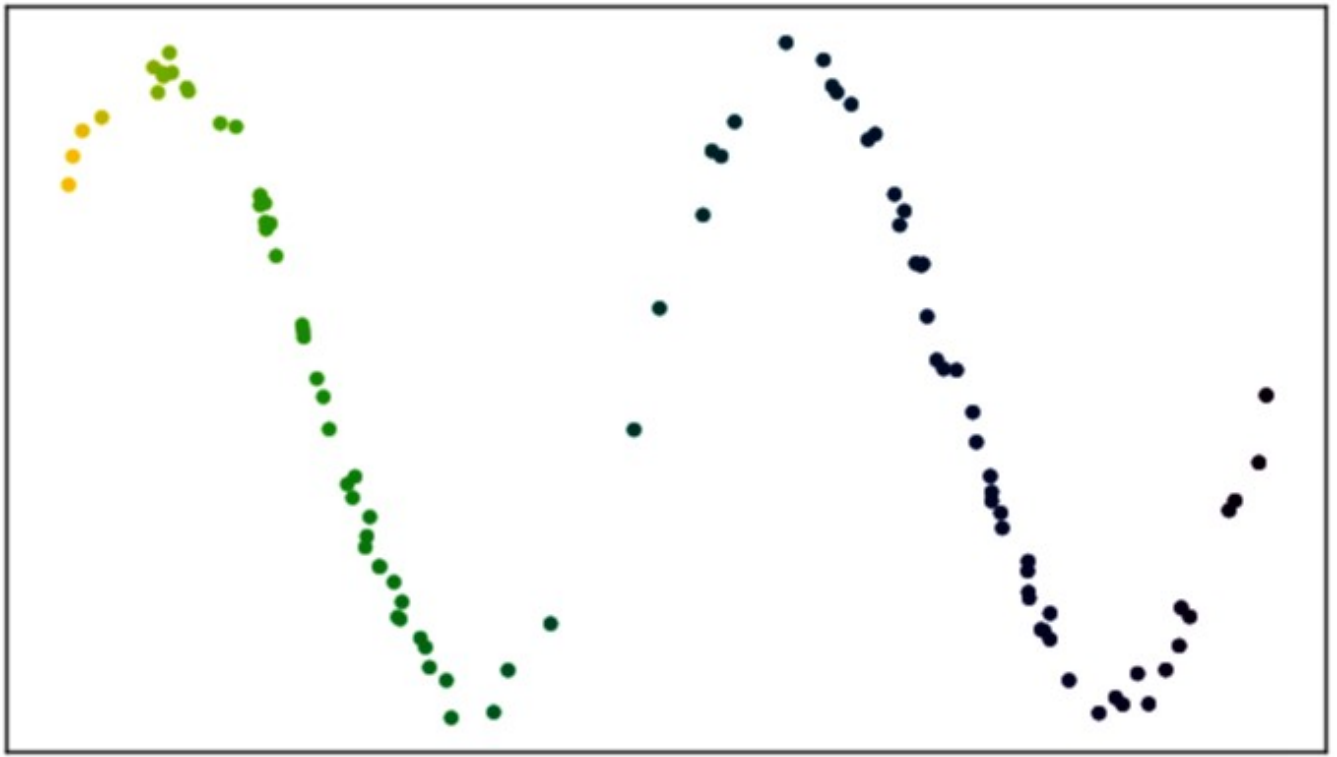
https://setosa.io/ev/principal-component-analysis/

Uniform Manifold Approximation and Projection (UMAP)

UMAP builds something called a **fuzzy simplicial complex**. This is really just a representation of a **weighted graph**, with edge weights representing the l**ikelihood that two points are connected**.

https://pair-code.github.io/understanding-umap/

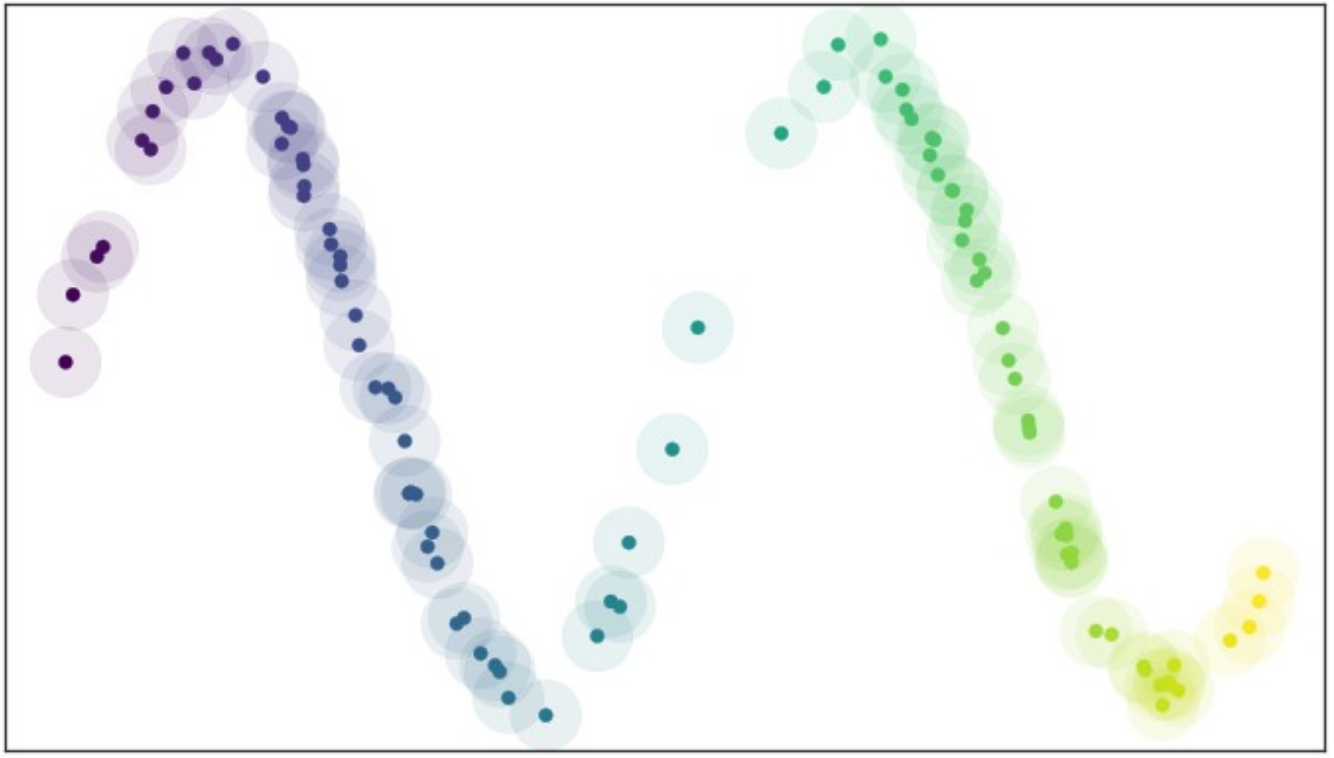
UMAP



As example a sinus curve with some noise

https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

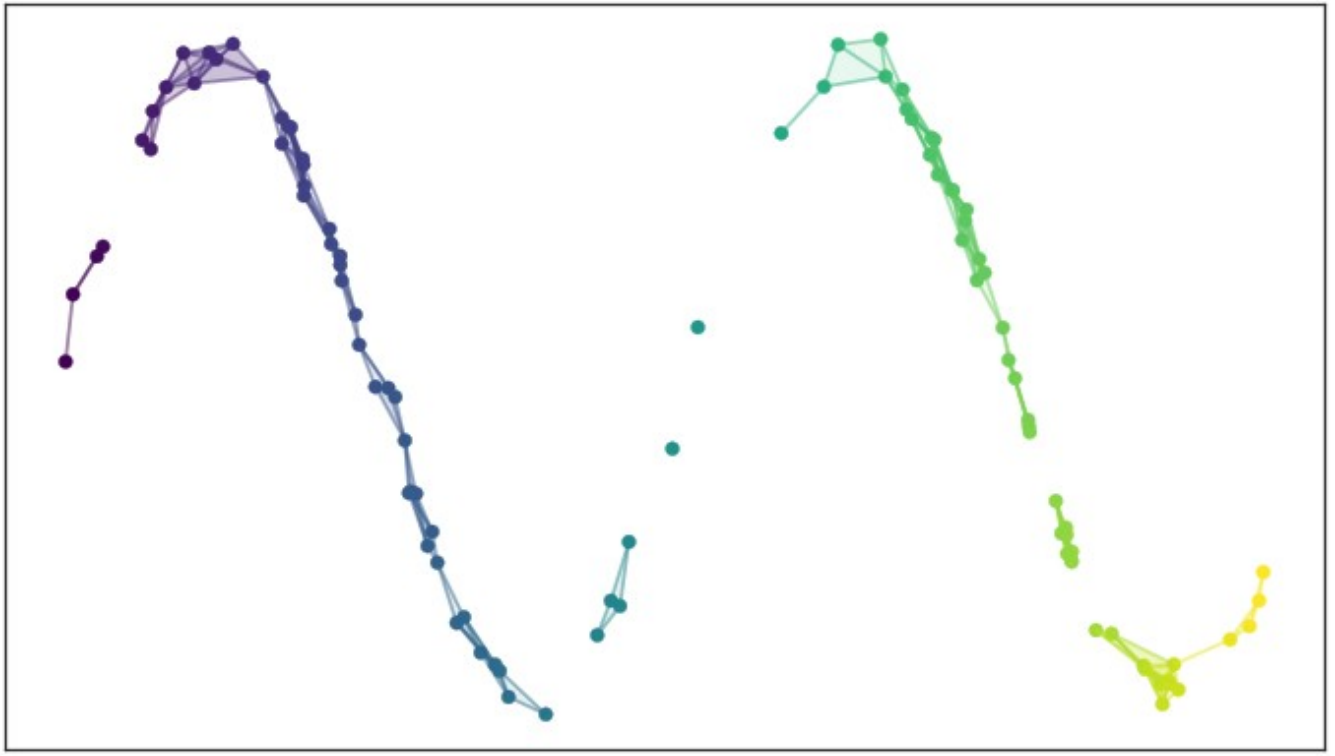
UMAP



A radius shows us whether there are neighbors

https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

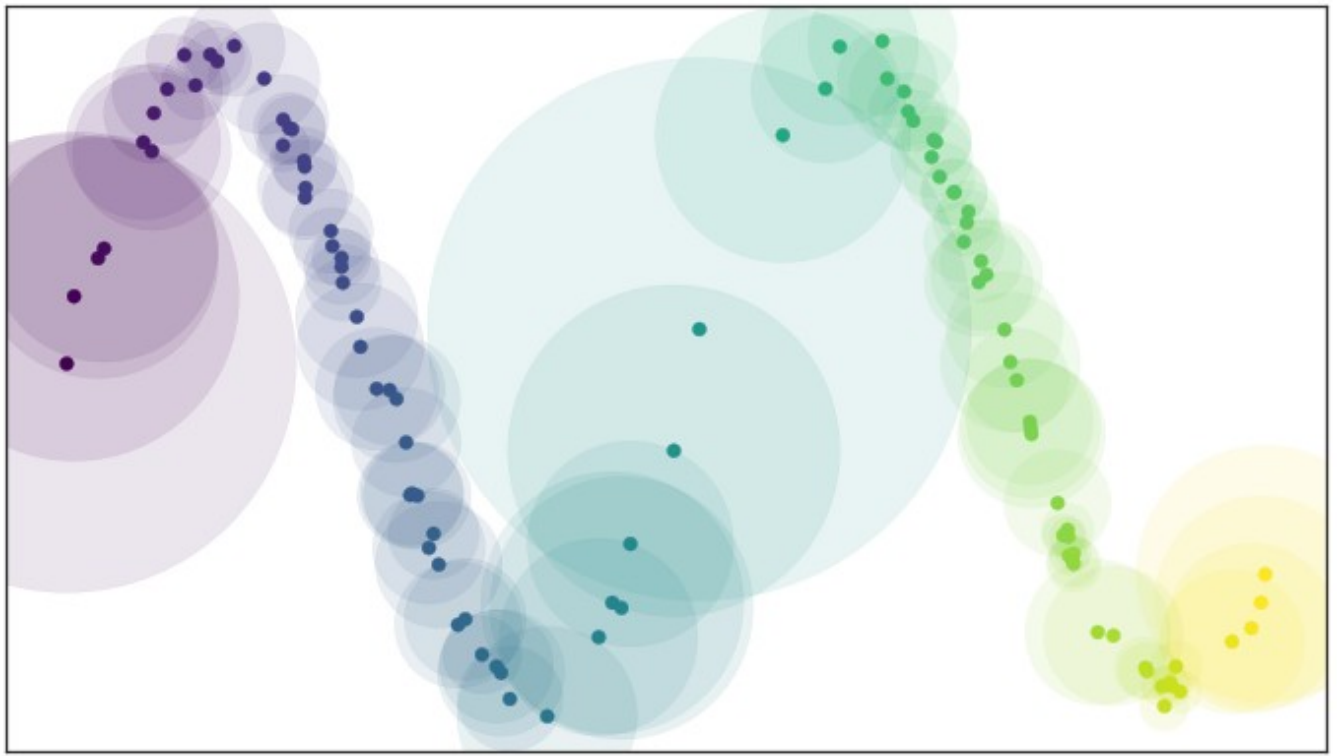
UMAP



The resulting yes-no answer is a bit unsatisfactory

https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

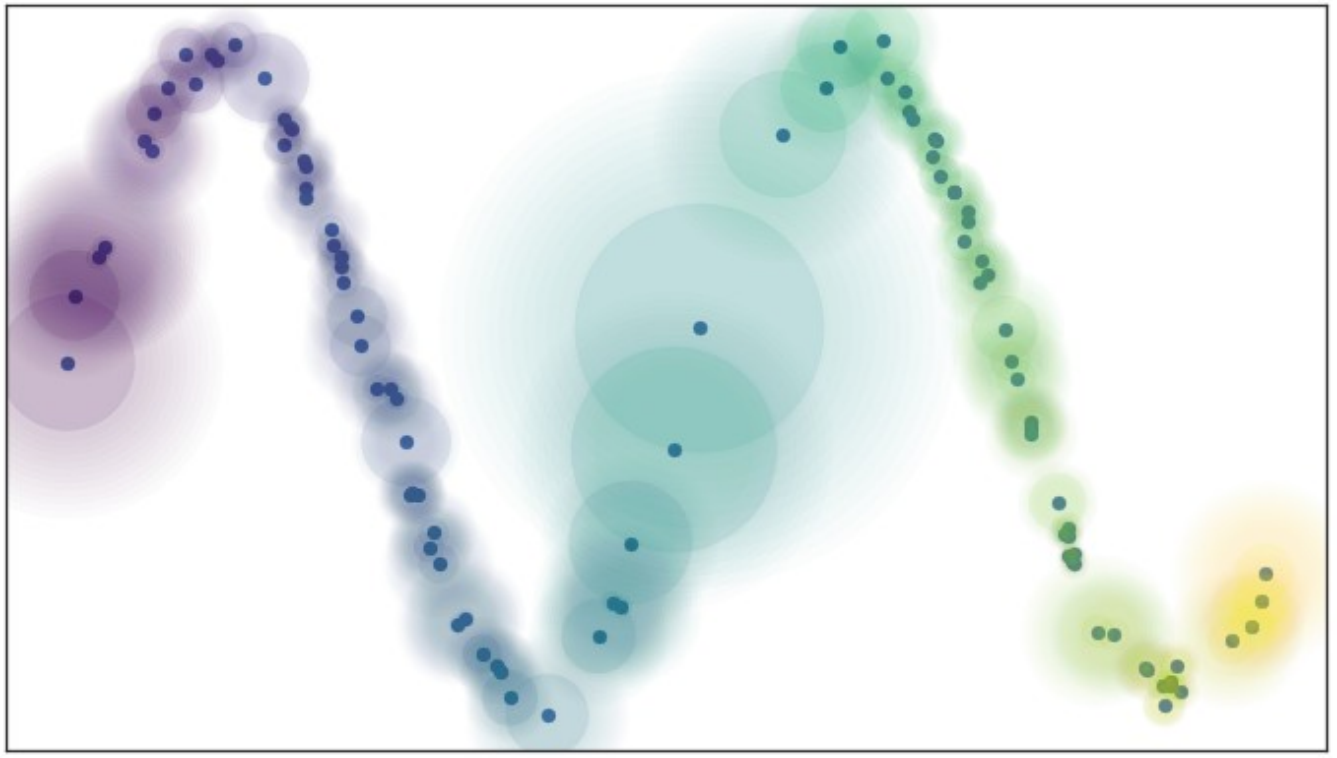
UMAP



Flexible radius also allow finding of isolated points

https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

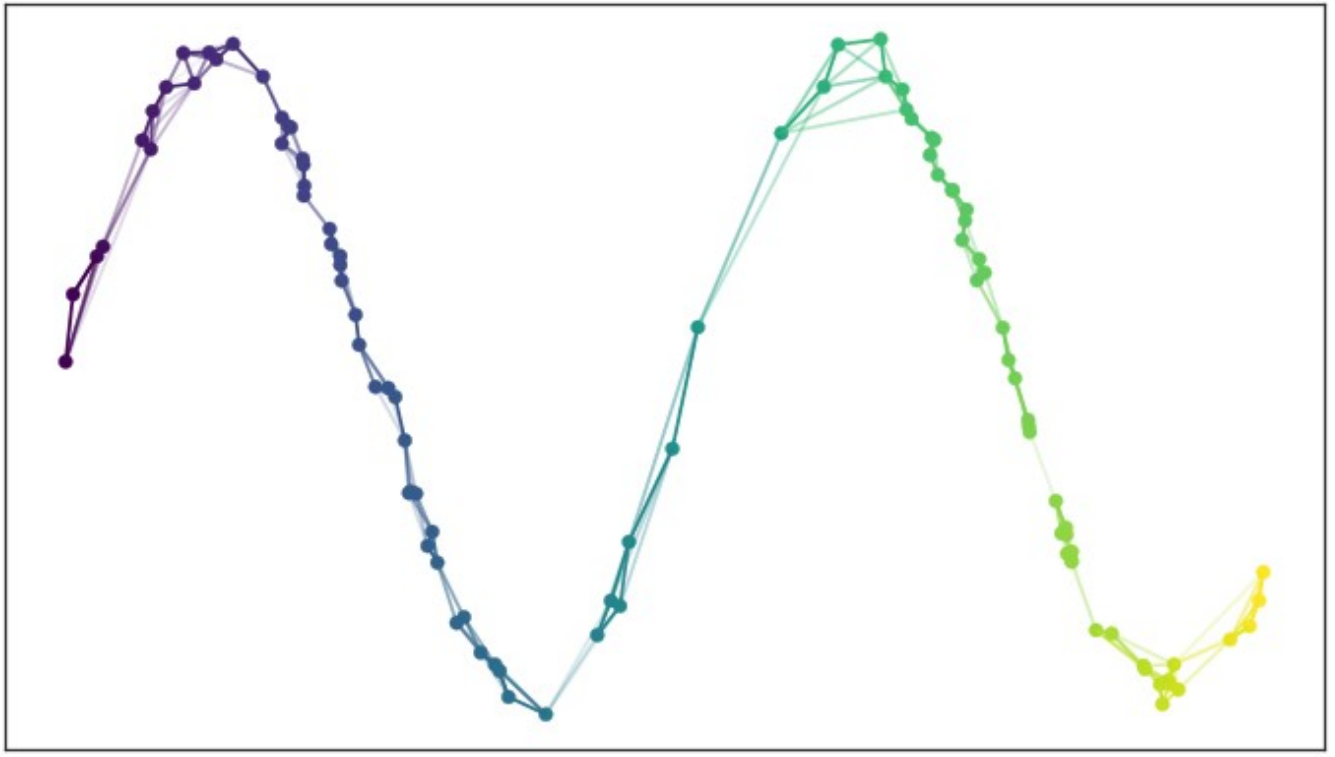
UMAP



A combination of a hard radius to the next neighbour and a flexible one beyond is more practical, because…

https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

UMAP



…it allows us to calculate probabilities, whether there are connections between the points.

https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

UMAP Parameters

…how distance is measured (**metric**)
…how many neighbours are considered (**n_neighbors**)
…how much points are allowed to overlay (**min-dist**)

How to (mis)read UMAP

Hyperparameters really matter

Choosing good values isn't easy, and depends on both the data and your goals. This is where UMAP's speed is a big advantage - By running UMAP multiple times with a variety of hyperparameters, you can get a better sense of how the projection is affected by its parameters.

Cluster sizes in a UMAP plot mean nothing

Just as in t-SNE, the size of clusters relative to each other is essentially meaningless. This is because UMAP uses local notions of distance to construct its high-dimensional graph representation.

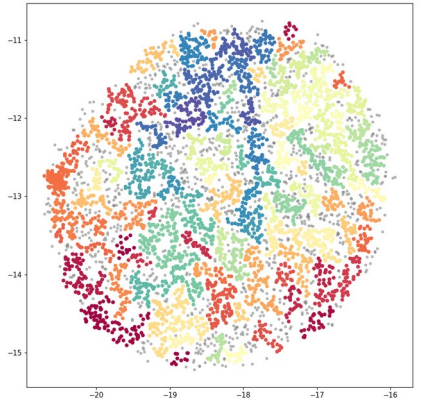
Distances between clusters might not mean anything

The distances between clusters is likely to be meaningless. While it's true that the global positions of clusters are better preserved in UMAP, the distances between them are not meaningful.

You may need more than one plot

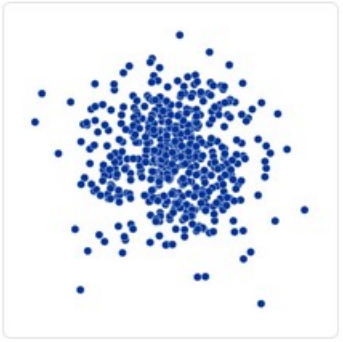
Since the UMAP algorithm is stochastic, different runs with the same hyperparameters can yield different results. Additionally, since the choice of hyperparameters is so important, it can be very useful to run the projection multiple times with various hyperparameters.

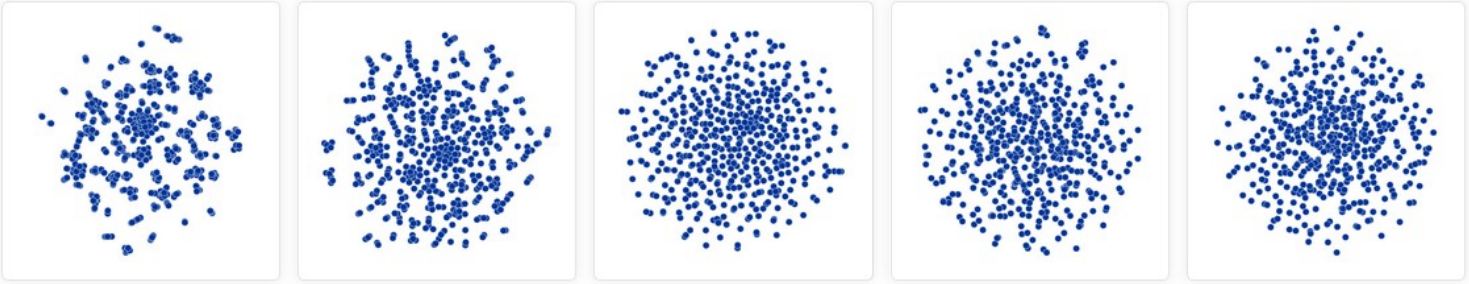
How to (mis)read UMAP

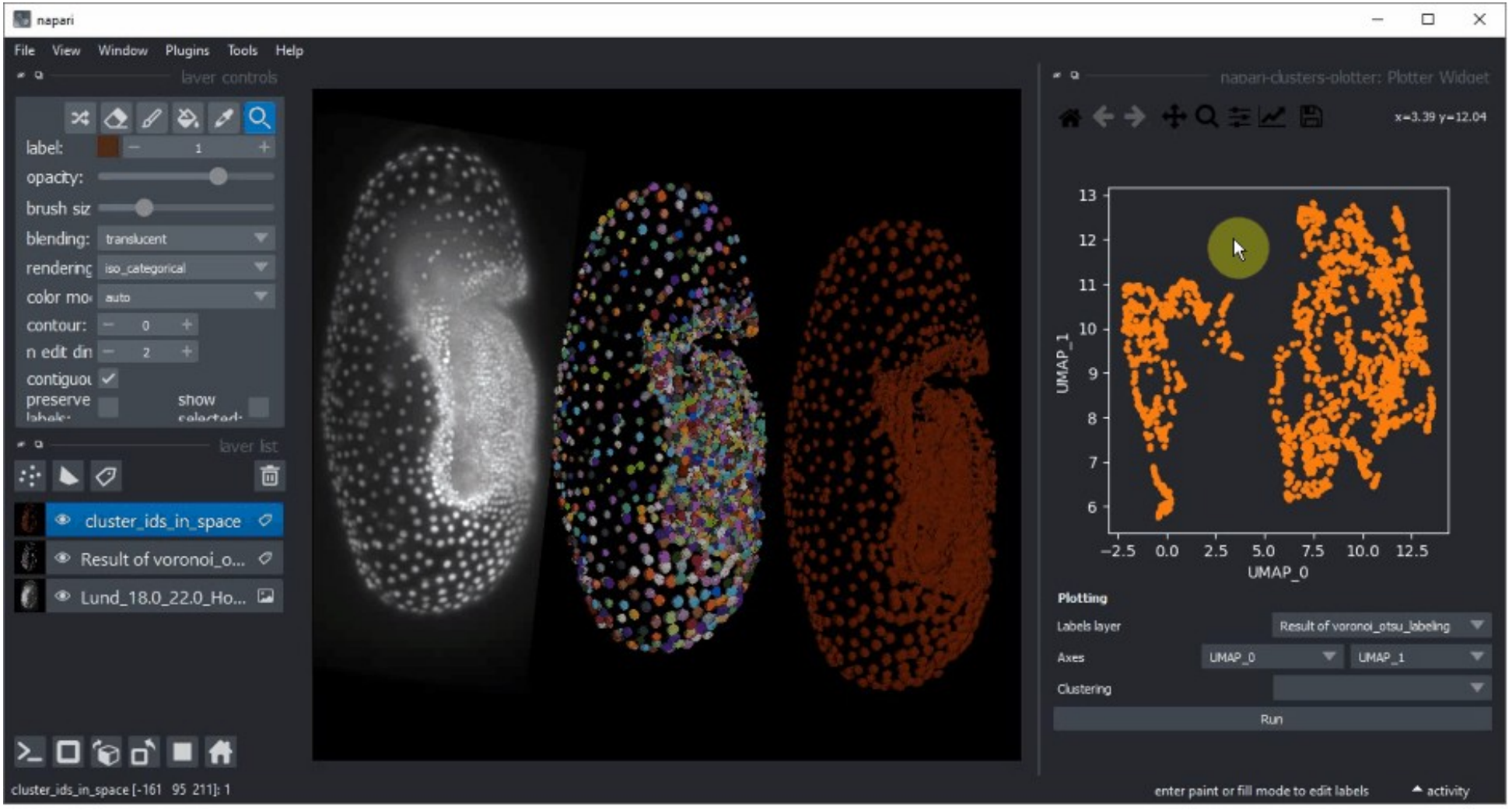


Random noise doesn’t always look random

Especially at low values of n_neighbors, spurious clustering can be observed.







@haesleinhuepf

https://www.napari-hub.org/plugins/napari-clusters-plotter

Applications of UMAP

Ready to go Tutorials:
scRNA-Seq tutorial in Python: https://github.com/theislab/single-cell-tutorial
scRNA-Seq blood analysis in Python: https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html
scRNA-Seq blood analysis in R: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html