TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

An introduction to Datascience

Anna Poetsch

Research Group „Biomedical Genomics“, Biotechnology Center TU Dresden, NCT Dresden, and CSBD

@apoetsch

Organisation

25.6.2024: Dimensionality reduction

2.7.2024: Machine Learning *

9.7.2024: Deep Learning, Large Language Models

16.7.2024: Summary, recap, Q&A *

*Melissa Sanabria & Pierre Joubert

TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

An introduction into Dimensionality reduction

Anna Poetsch

Research Group „Biomedical Genomics“, Biotechnology Center TU Dresden, NCT Dresden, and CSBD

@apoetsch

Organisation

25.6.2024: Dimensionality reduction

2.7.2024: Machine Learning

9.7.2024: Deep Learning, Large Language Models

16.7.2024: Summary, recap, Q&A

Principle 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.

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principle Component Analysis (PCA)

Principle 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 components in gene expression analysis

Collaboration AG Stange, UKD

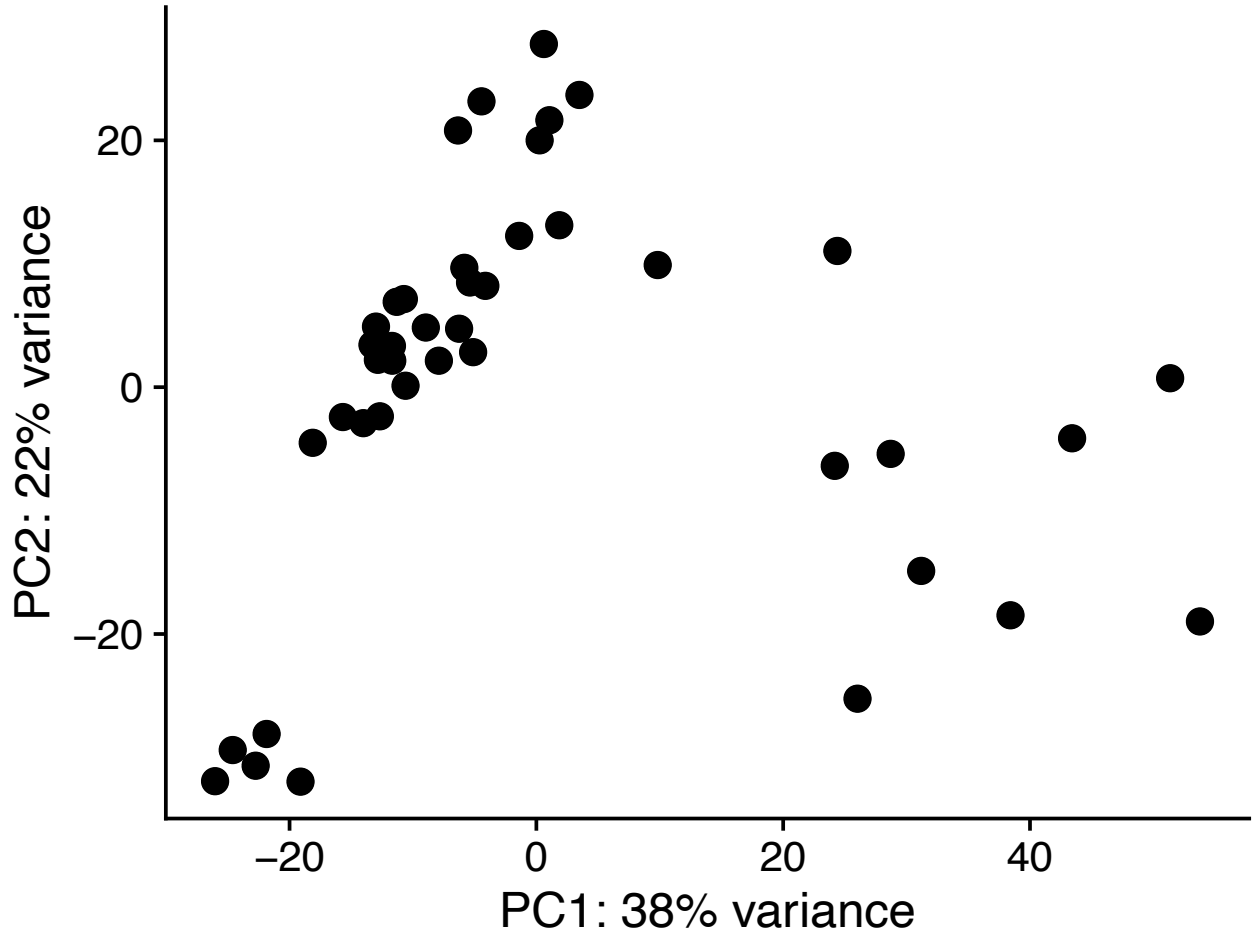
Principal components in gene expression analysis

Gene ID	Gene Name	Reference	Strand	Start	End	Coverage	FPKM	TPM
ENSG00000239906	-	1	-	139790	140339	0.222910	0.173640	0.358875
ENSG00000290825	DDX11L2	1	+	11869	14409	0.000000	0.000000	0.000000
ENSG00000223972	DDX11L1	1	+	12010	13670	0.000000	0.000000	0.000000
ENSG00000227232	WASH7P	1	-	14404	29570	0.569208	0.443395	0.916400
ENSG00000278267	MIR6859-1	1	-	17369	17436	4.279412	3.333523	6.889668
ENSG00000243485	MIR1302-2HG	1	+	29554	31109	0.000000	0.000000	0.000000
ENSG00000284332	MIR1302-2	1	+	30366	30503	0.000000	0.000000	0.000000
ENSG00000238009	-	1	-	89295	133723	0.000000	0.000000	0.000000
ENSG00000239945	-	1	-	89551	91105	0.000000	0.000000	0.000000
ENSG00000233750	CICP27	1	+	131025	134836	0.000000	0.000000	0.000000
ENSG00000268903	-	1	-	135141	135895	1.406623	1.095714	2.264602
ENSG00000269981	-	1	-	137682	137965	1.327465	1.034052	2.137161
ENSG00000241860	-	1	-	141474	173862	0.030956	0.080914	0.167231
ENSG00000222623	RNU6-1100P	1	-	157784	157887	0.000000	0.000000	0.000000
ENSG00000241599	-	1	+	160446	161525	0.000000	0.000000	0.000000
ENSG00000279928	DDX11L17	1	+	182696	184174	0.000000	0.000000	0.000000
ENSG00000279457	WASH9P	1	-	185217	195411	3.196135	2.489686	5.145639
ENSG00000273874	MIR6859-2	1	-	187891	187958	0.000000	0.000000	0.000000
ENSG00000228463	-	1	-	257864	359681	0.032588	0.037260	0.077007
ENSG00000286448	-	1	+	266855	268655	0.000000	0.000000	0.000000
ENSG00000236679	RPL23AP24	1	-	347982	348366	0.000000	0.000000	0.000000
ENSG00000236601	-	1	+	358857	366052	0.000000	0.000000	0.000000
ENSG00000290385	-	1	-	365389	522928	0.017129	0.501282	1.036041
ENSG00000269732	WBP1LP7	1	+	439870	440232	0.000000	0.000000	0.000000
ENSG00000284733	OR4F29	1	-	450740	451678	0.000000	0.000000	0.000000
ENSG00000237094	-	1	-	485026	485208	0.000000	0.000000	0.000000
ENSG00000233653	CICP7	1	+	487101	489906	0.000000	0.000000	0.000000
ENSG00000250575	-	1	-	491225	493241	0.058111	0.045267	0.093556
ENSG00000278757	U6	1	-	516376	516479	0.000000	0.000000	0.000000
ENSG00000272438	-	1	+	904834	915976	0.000000	0.000000	0.000000
ENSG00000230699	-	1	+	911435	914948	0.000000	0.000000	0.000000
ENSG00000241180	-	1	+	914171	914971	0.000000	0.000000	0.000000
ENSG00000288531	-	1	+	860226	868202	0.000000	0.000000	0.000000
ENSG00000230368	FAM41C	1	-	868071	876903	0.000000	0.000000	0.000000
ENSG00000234711	TUBB8P11	1	+	873292	874349	0.000000	0.000000	0.000000
ENSG00000283040	-	1	-	874529	877234	0.000000	0.000000	0.000000
ENSG00000223764	LINC02593	1	-	916865	921016	0.039258	0.030581	0.063204
ENSG00000187634	SAMD11	1	+	923923	944575	0.000000	0.000000	0.000000
ENSG00000188976	NOC2L	1	-	944203	959309	35.001949	55.589951	114.892357
ENSG00000187961	KLHL17	1	+	960584	965719	3.948853	5.744548	11.872734
ENSG00000230021	-	1	-	586071	827796	0.000000	0.000000	0.000000
ENSG00000235146	-	1	+	587629	594768	0.000000	0.000000	0.000000
ENSG00000225972	MTND1P23	1	+	629062	629433	3.741935	2.914846	6.024354
ENSG00000225630	MTND2P28	1	+	629640	630683	156.061310	121.566681	251.251953
ENSG00000237973	MTCO1P12	1	+	631074	632616	96.286453	75.004013	155.017029

Every gene represents a dimension

in every sample

A PCA for the colon data



Using PCA to look for batch effects

Before

After

Batch correction for direct use of counts only,
For differential analysis, a more sophisticated method is used

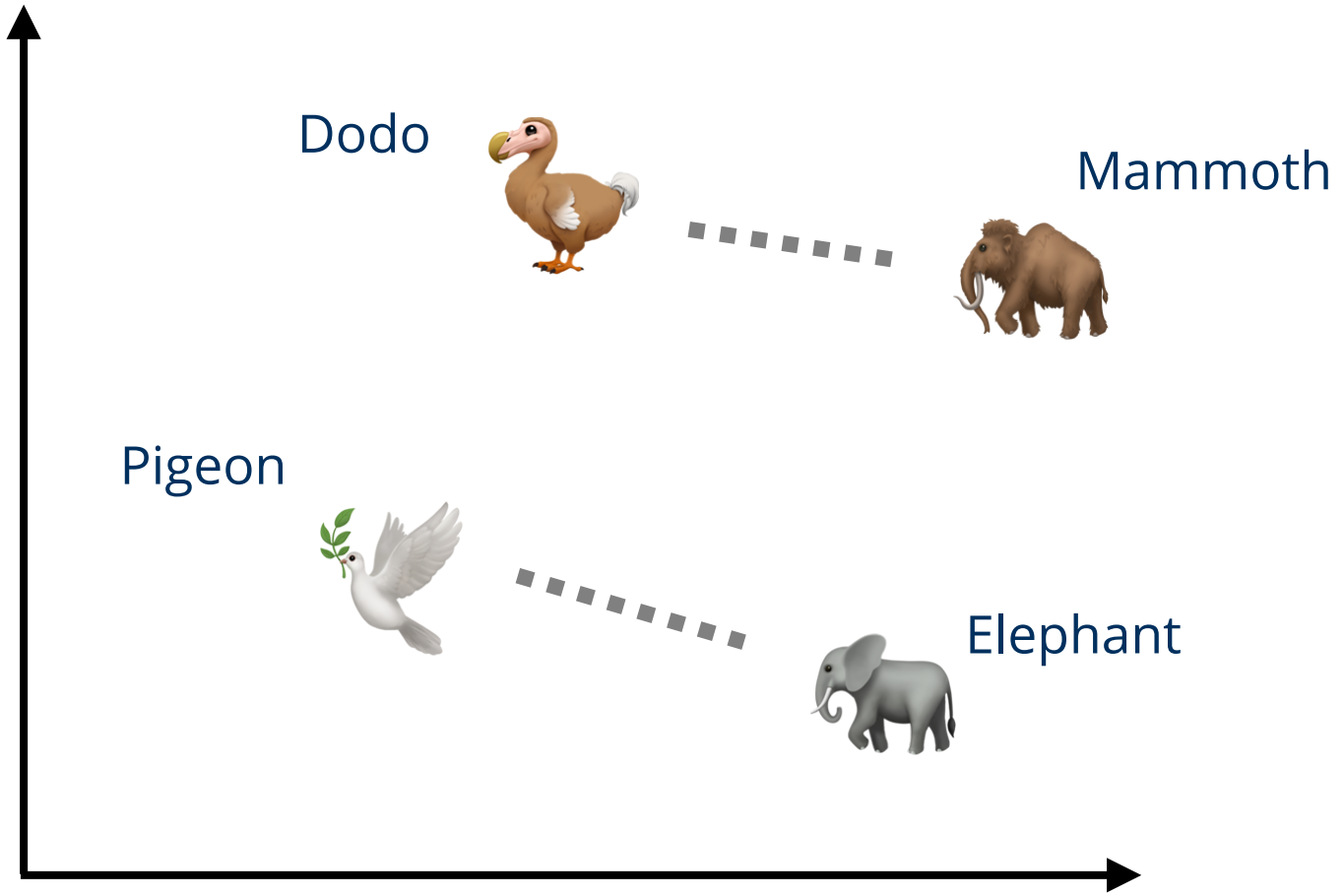
Is it the medium or is it the location?

Every medium activates its own transcriptional program

Uses of PCA

- get to know your multidimensional data
- explore batch effects
- evaluate results of batch correction
- explore grouping of samples
- explore relative relationships between groups

PCA in the embedding of language



Advantages of PCA and its downside

- well interpretable

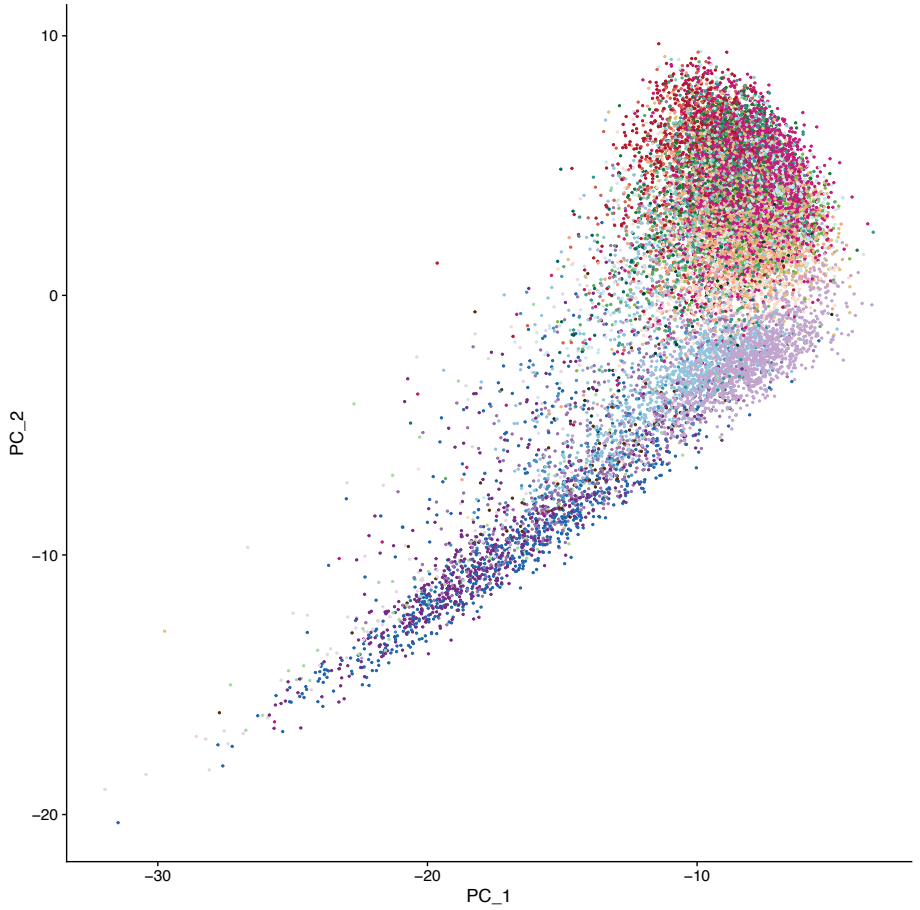
- linear

- some structure cannot be captured with a linear dimensionality reduction

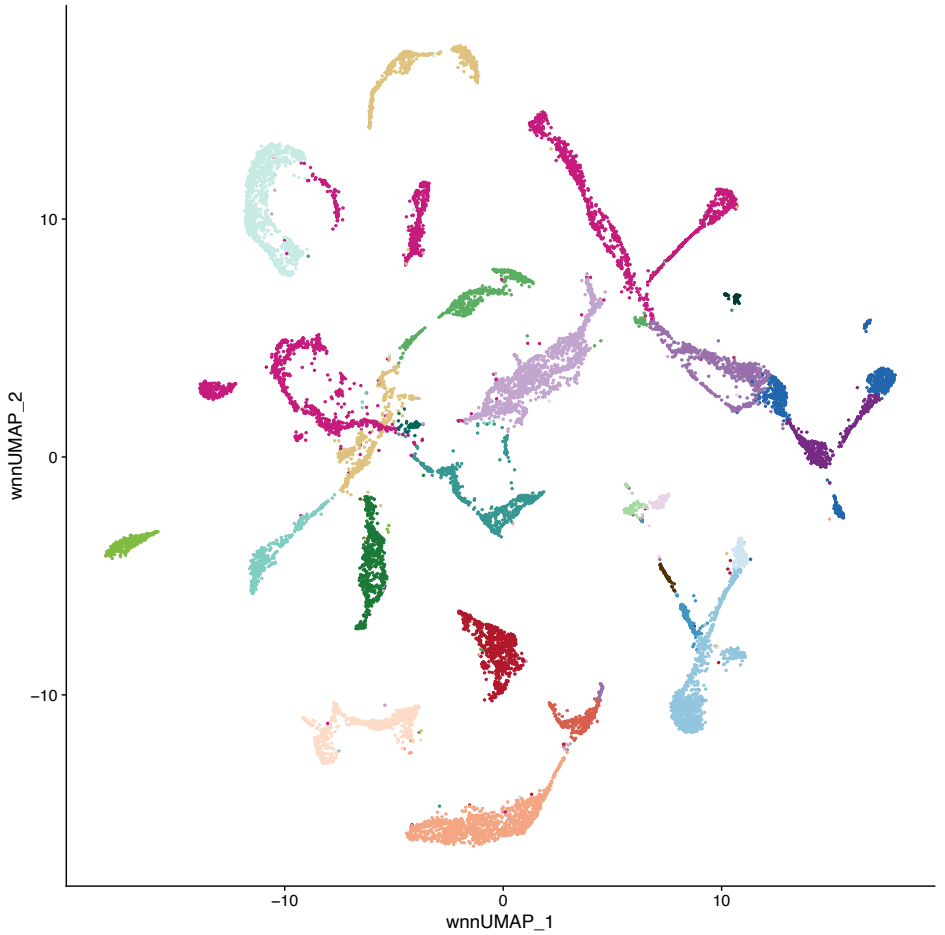
-> non-linear dimensionality reduction

Single cell gene expression data

PCA



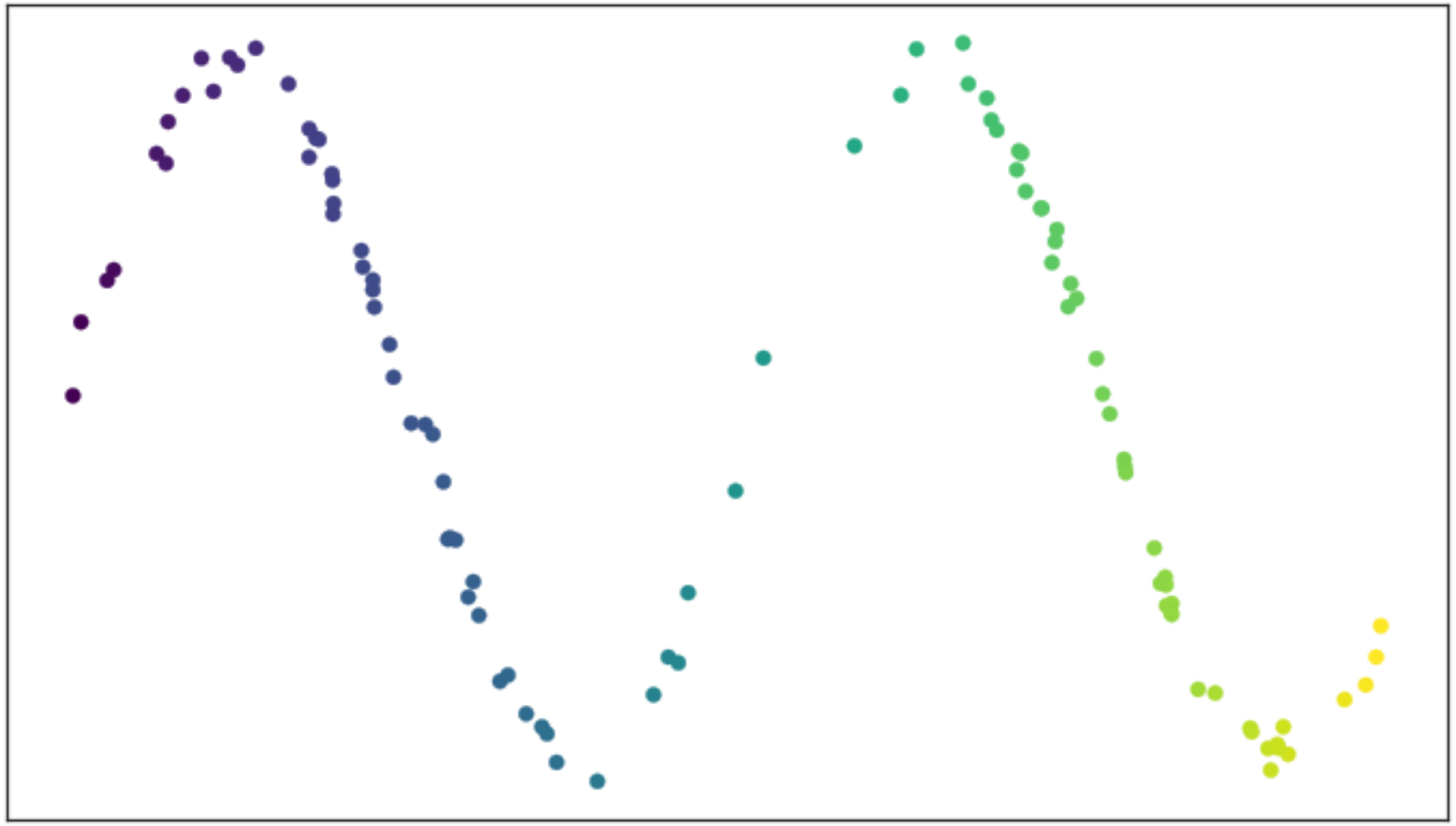
UMAP



Collaboration AG Becker, CRTD

UMAP with two dimensions

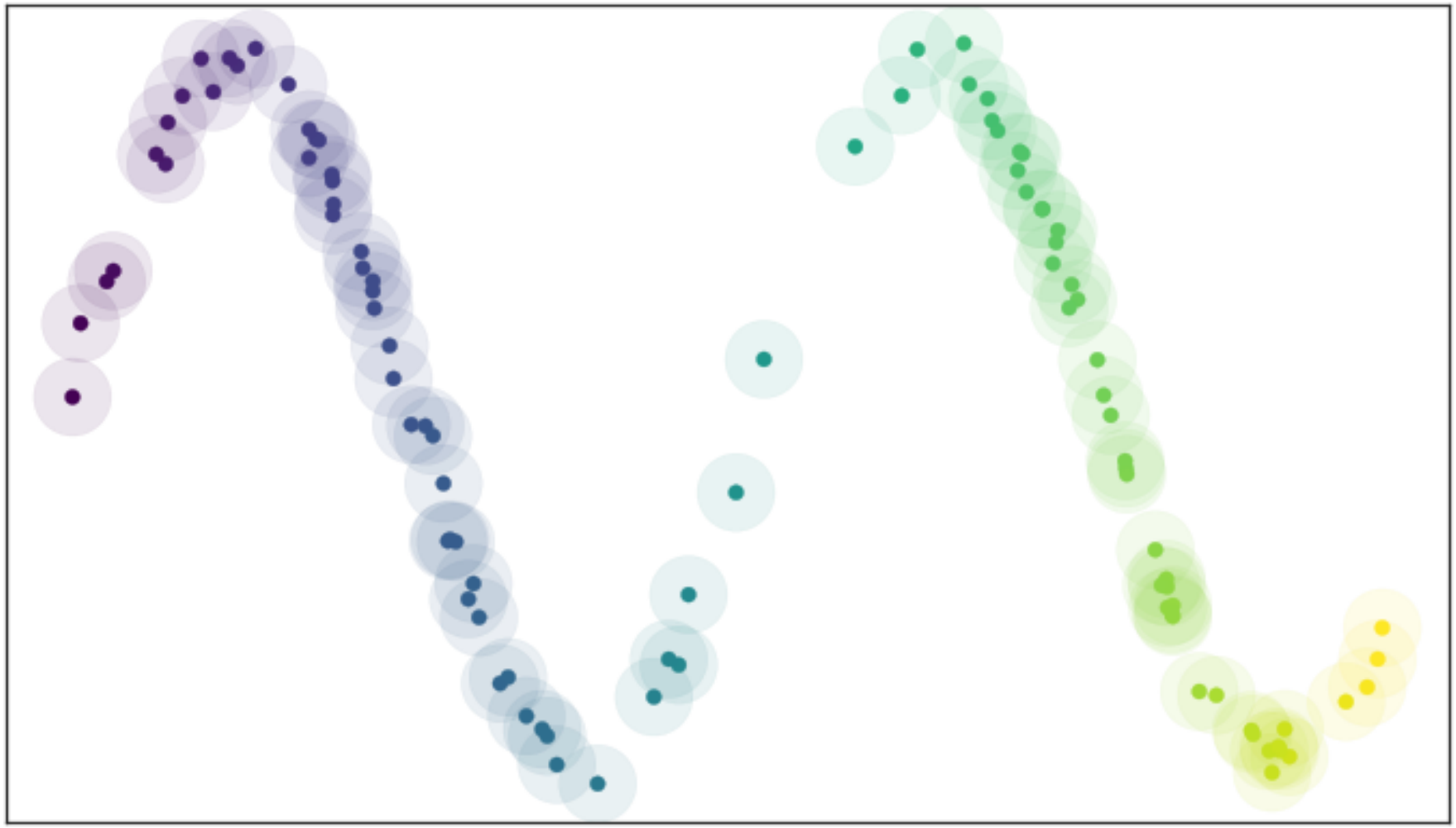
(UMAP=Uniform Manifold Approximation and Projection)



As example a sinus curve with some noise

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

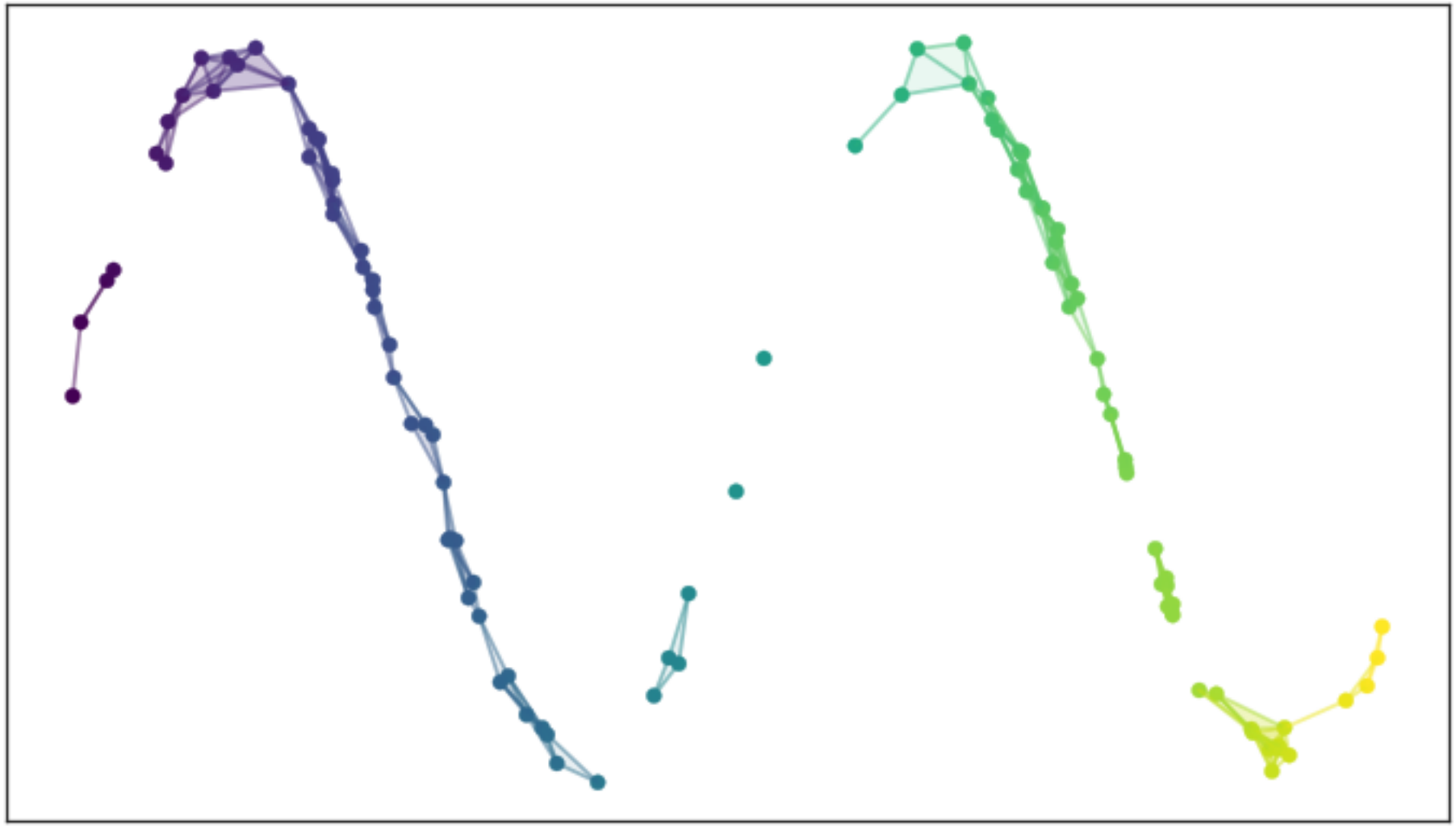
UMAP with two dimensions



A radius shows us whether there are neighbours

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

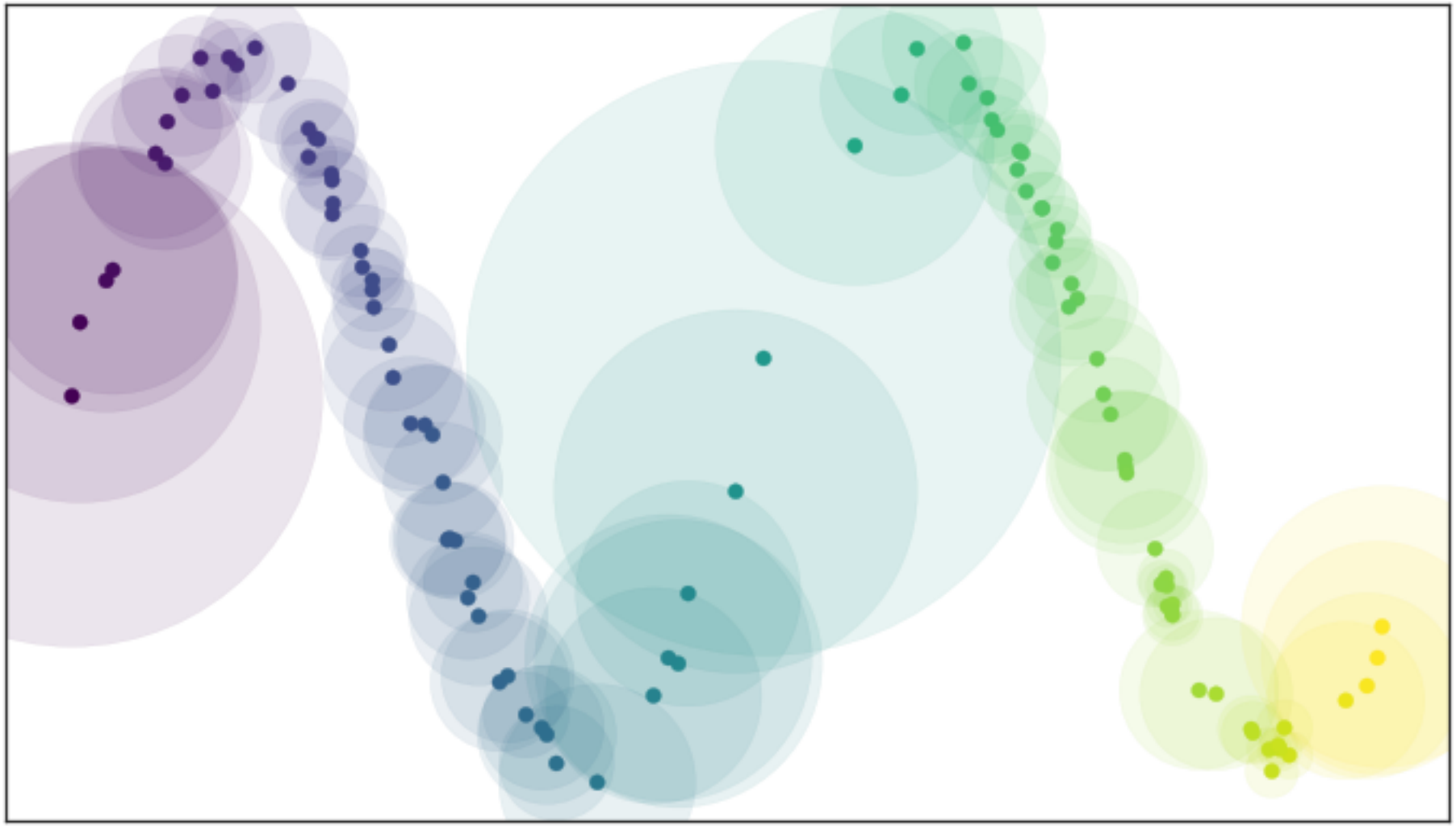
UMAP with two dimensions



The resulting yes-no answer is a bit unsatisfactory

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

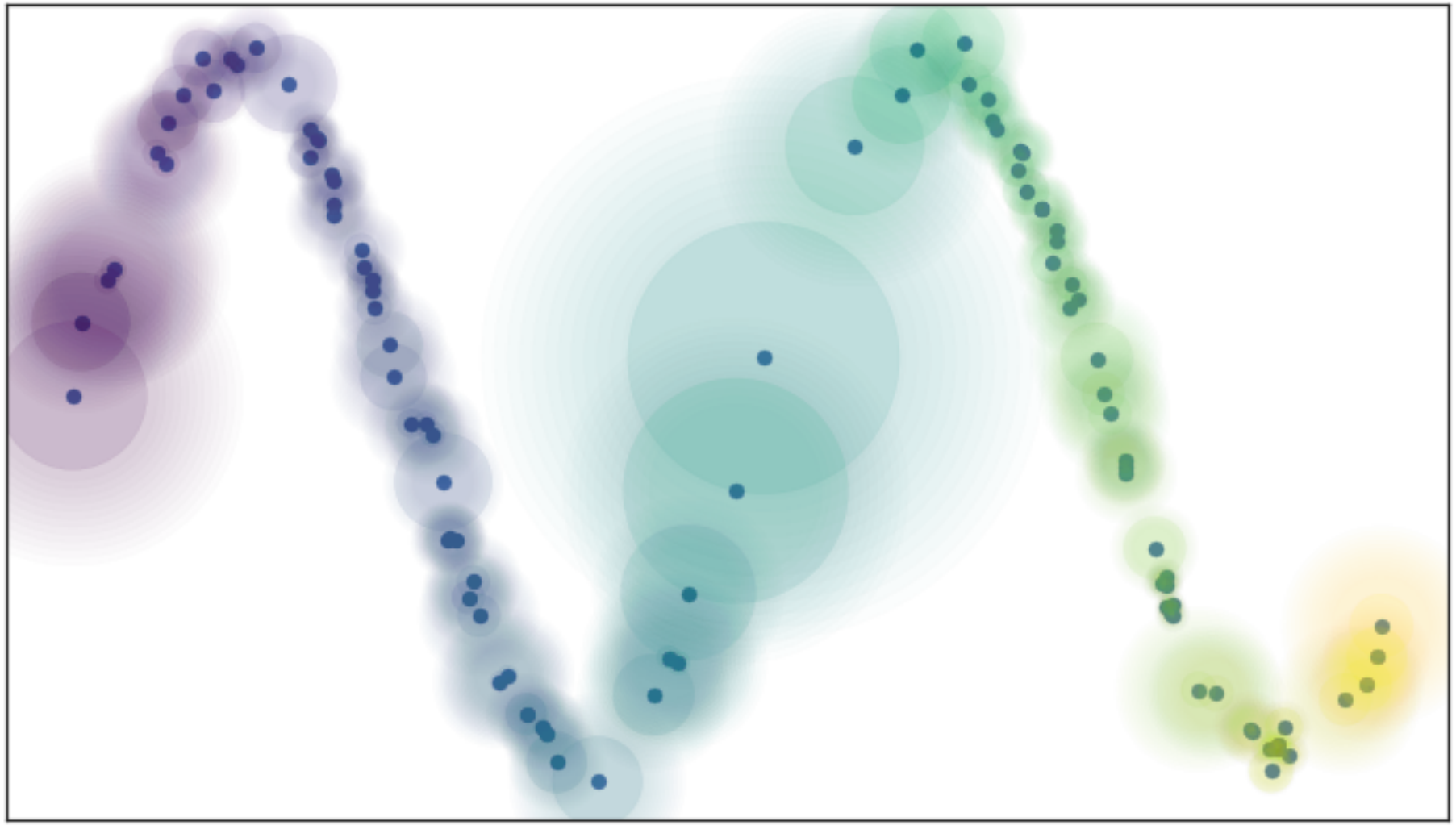
UMAP with two dimensions



Flexible radia also allow finding of isolated points

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

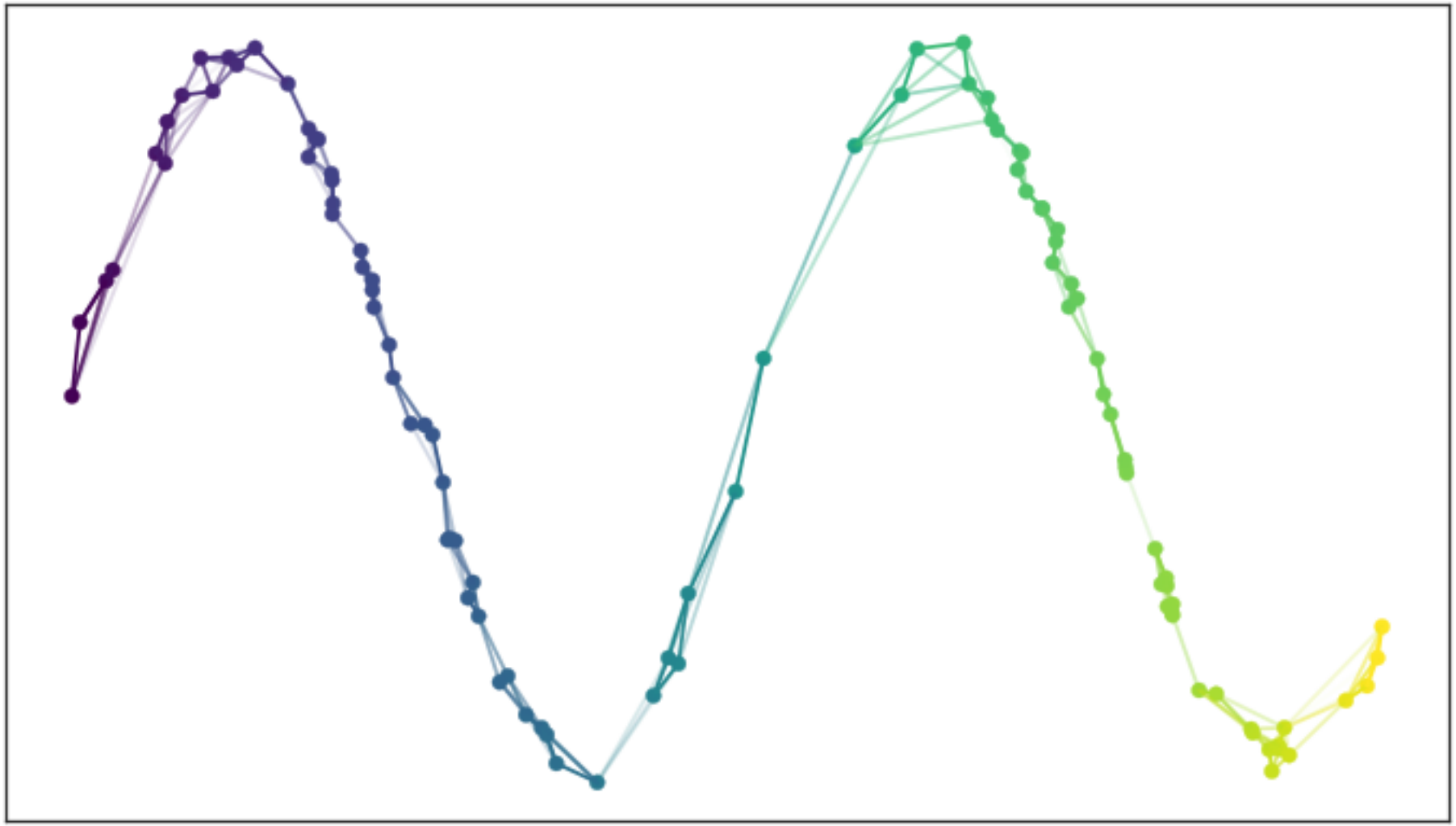
UMAP with two dimensions



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 with two dimensions



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

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

Dimensionality reduction with UMAP

- Optimisation algorithm for a “flexible” distance measurement to find a place in a low-dimensional space

- The result of the optimisation is dependent on the data and a random component

- UMAP can be tweaked with…
  …how distance is measured (metric)
  …how many neighbours are considered (n_neighbors)
  …how much points are allowed to overlay (min-dist)

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

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.

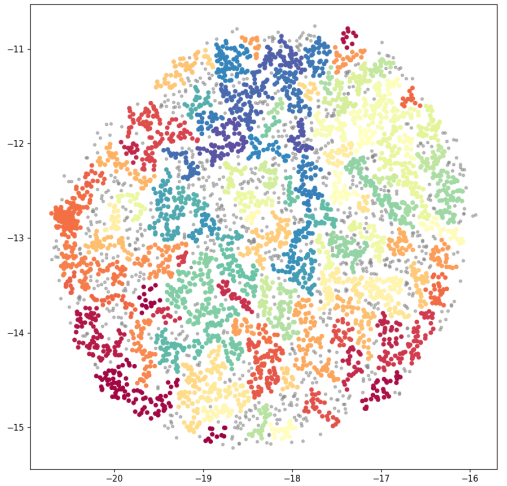
Distances between clusters might not mean anything

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

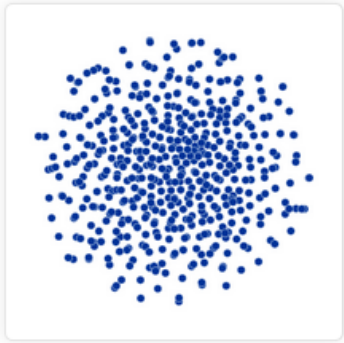
Since the UMAP algorithm is stochastic, different runs with the same hyperparameters can yield different results. Additionally, since the choice of hyperparameters is so important, it can be very useful to run the projection multiple times with various hyperparameters.

How to (mis)read UMAP



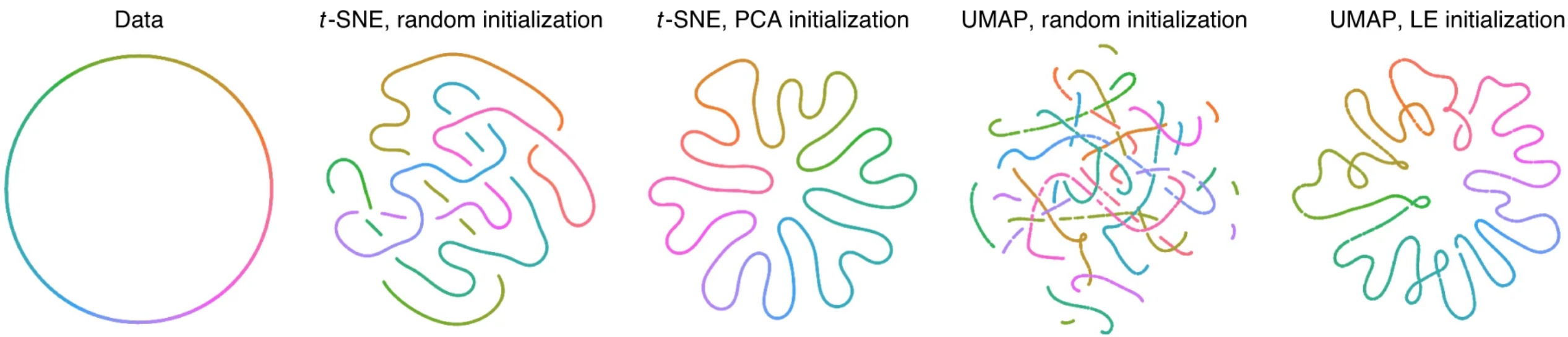
Random noise doesn’t always look random

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



Things to consider

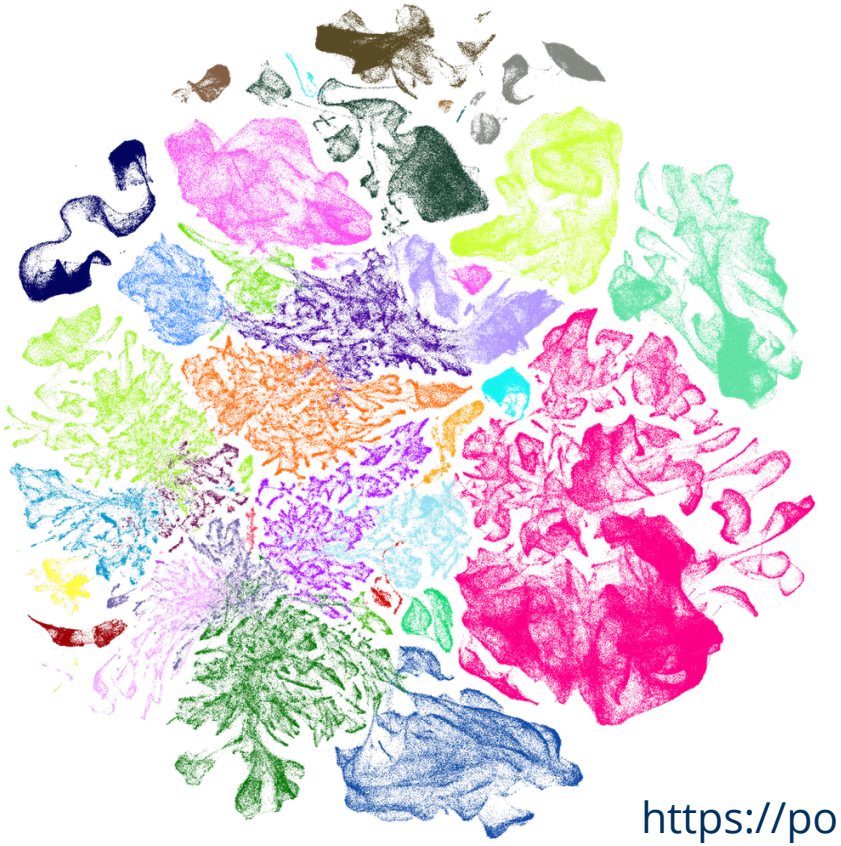
- Many parameters invite to “adjust” the data analysis
- Danger to over-interpret the visual “distance”
- How much data structure is preserved is still a matter of debate



https://www.nature.com/articles/s41587-020-00809-z

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

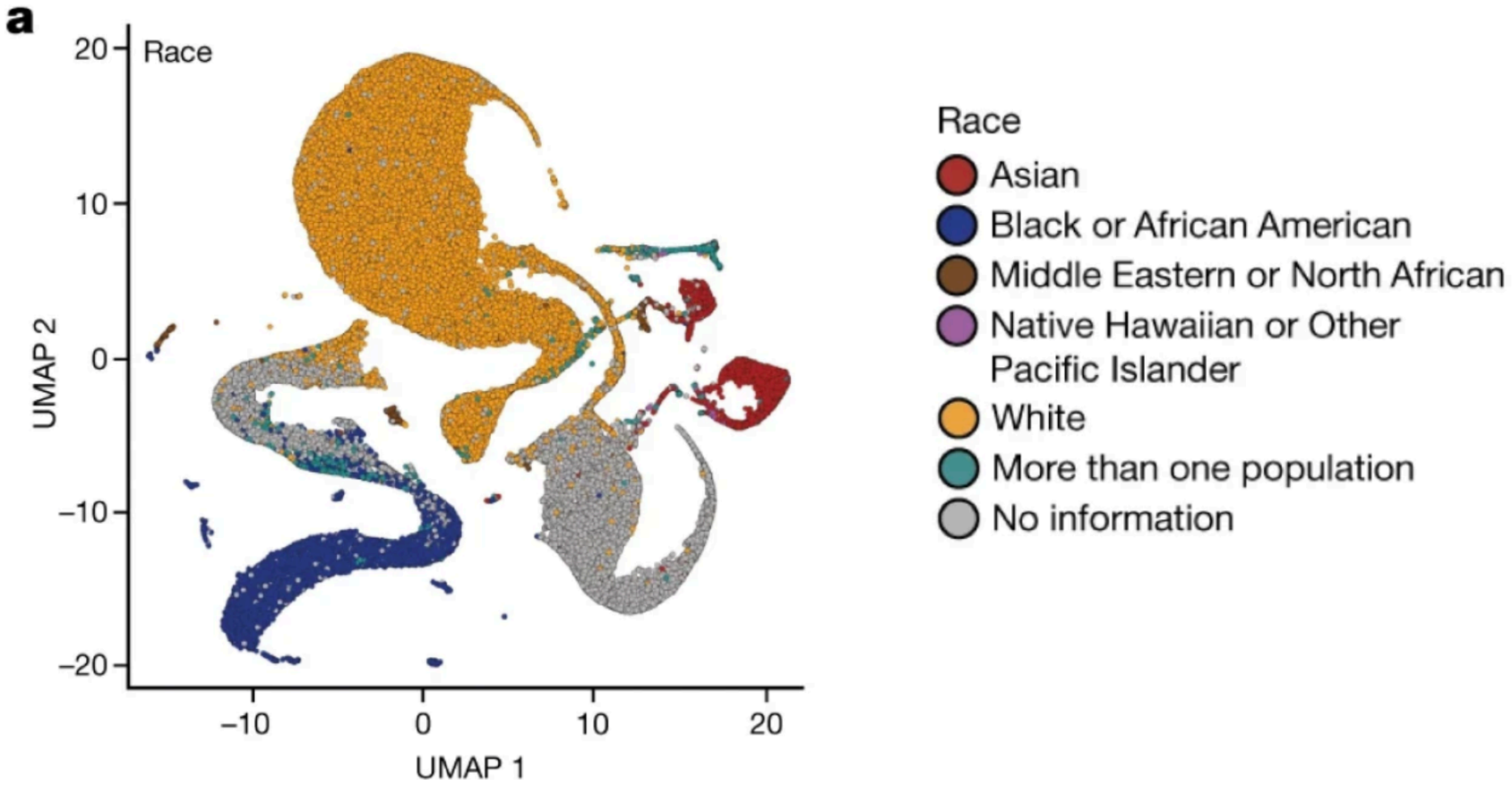


https://portal.brain-map.org/atlases-and-data/bkp/abc-atlas

‘All of Us’ genetics chart stirs unease over controversial depiction of race

Debate over figure connecting genes, race and ethnicity reignites concerns among geneticists about how to represent human diversity.

By Max Kozlov



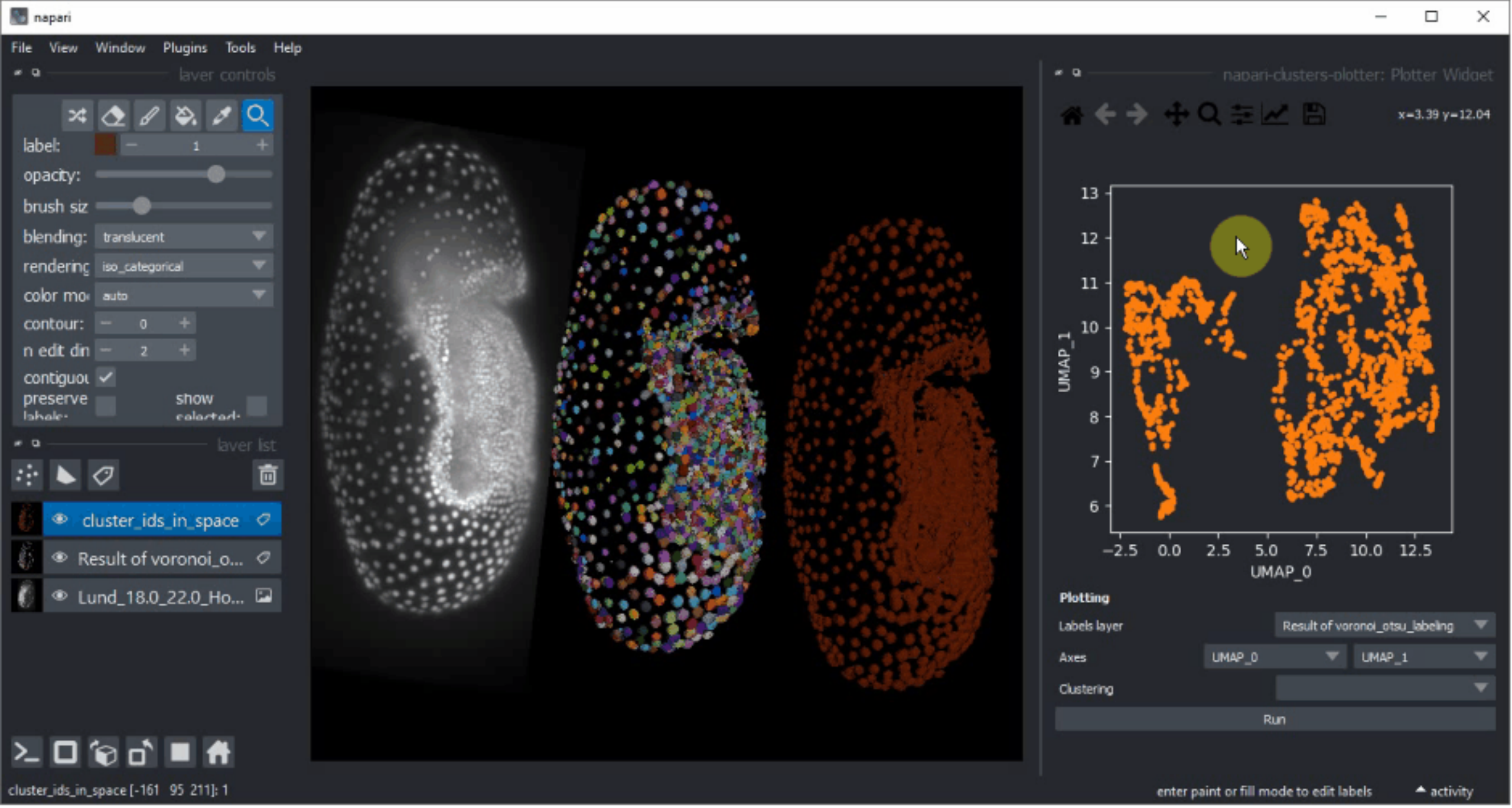
‘All of Us’ genetics chart stirs unease over controversial depiction of race

Debate over figure connecting genes, race and ethnicity reignites concerns among geneticists about how to represent human diversity.

By Max Kozlov

To a layperson, the chart shows several distinct colourful blobs that could be misinterpreted as supporting genetic essentialism — the pseudoscientific belief that racial or ethnic groups are distinct genetic categories, and that individuals of the same group are genetically similar, Birney says.

That is the opposite of what the data show, Bick says. “Our analysis reaffirms that race and ethnicity are social constructs that do not have a basis in genetics”.

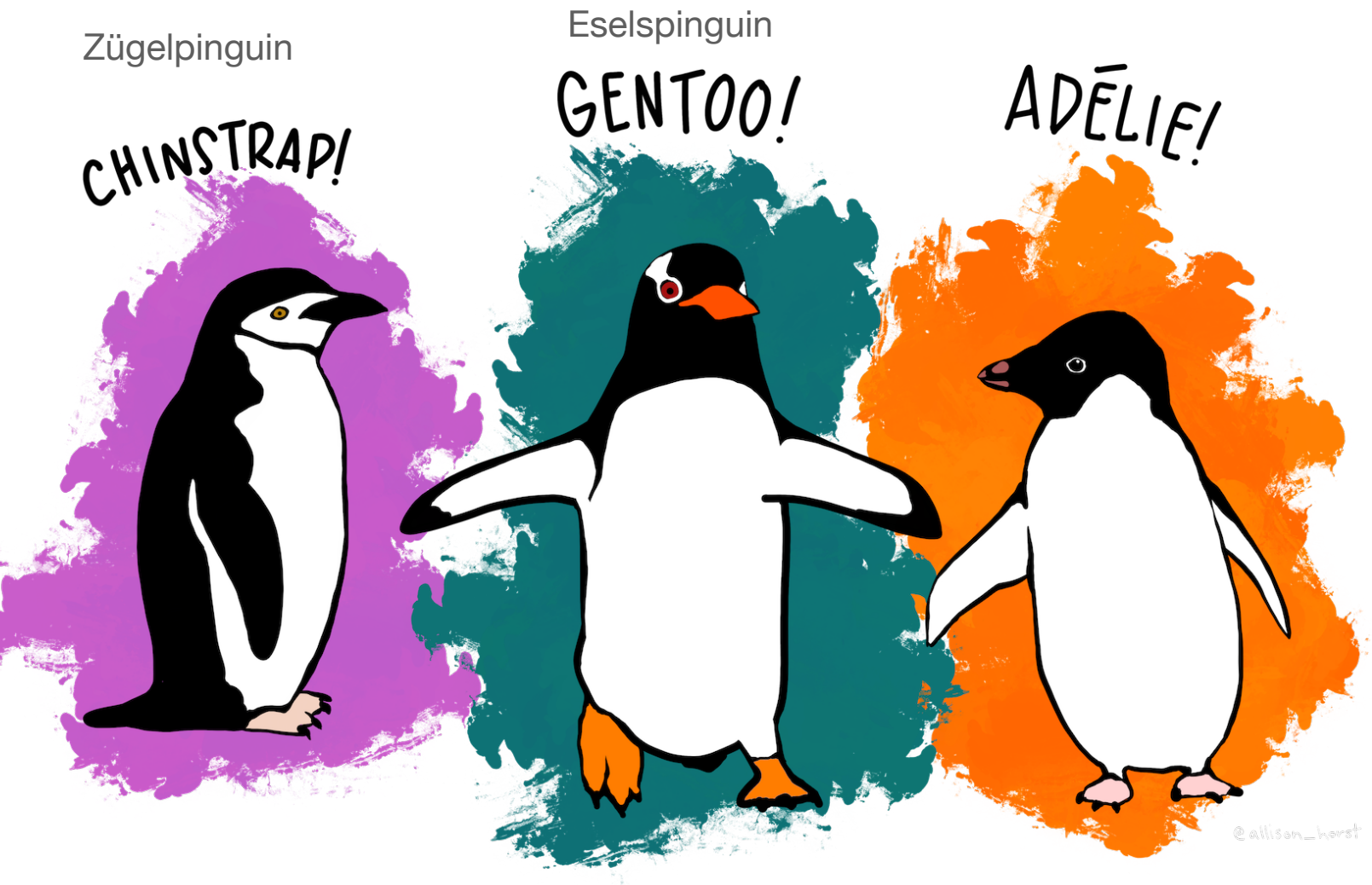

napari
File View Window Plugins Tools Help
layer controls
label:
opacity:
brush siz
blending: translucent
rendering iso_categorical
color mo auto
contour:
n edit dir
contiguou
preserve
show
layer list
cluster_ids_in_space
Result of voronoi_o...
Lund_18.0_22.0_Ho...
napari-clusters-plotter: Plotter Widget
x=3.39 y=12.04
UMAP_1
UMAP_0
Plotting
Labels layer
Result of voronoi_otsu_labeling
Axes
UMAP_0
UMAP_1
Clustering
Run
cluster_ids_in_space [-161 95 211]: 1
enter paint or fill mode to edit labels
activity

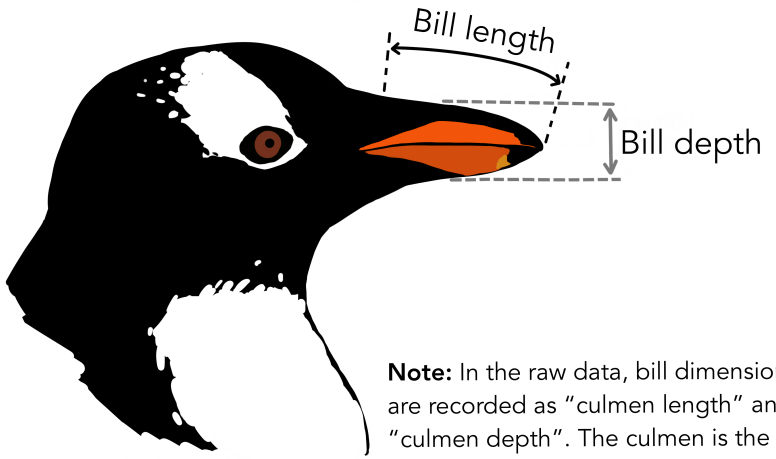
Sources and Material:

Tutorial: https://umap-learn.readthedocs.io/en/latest/
Paper: https://arxiv.org/abs/1802.03426
scRNA-Seq tutorial in Python: https://github.com/theislab/single-cell-tutorial
blood analysis in Python: https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html
blood analysis in R: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

PCA: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Exercise:





Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

https://colab.research.google.com/drive/1VOLoXAug76sXeL8PQiFUg3_diZrx2vXL?usp=sharing