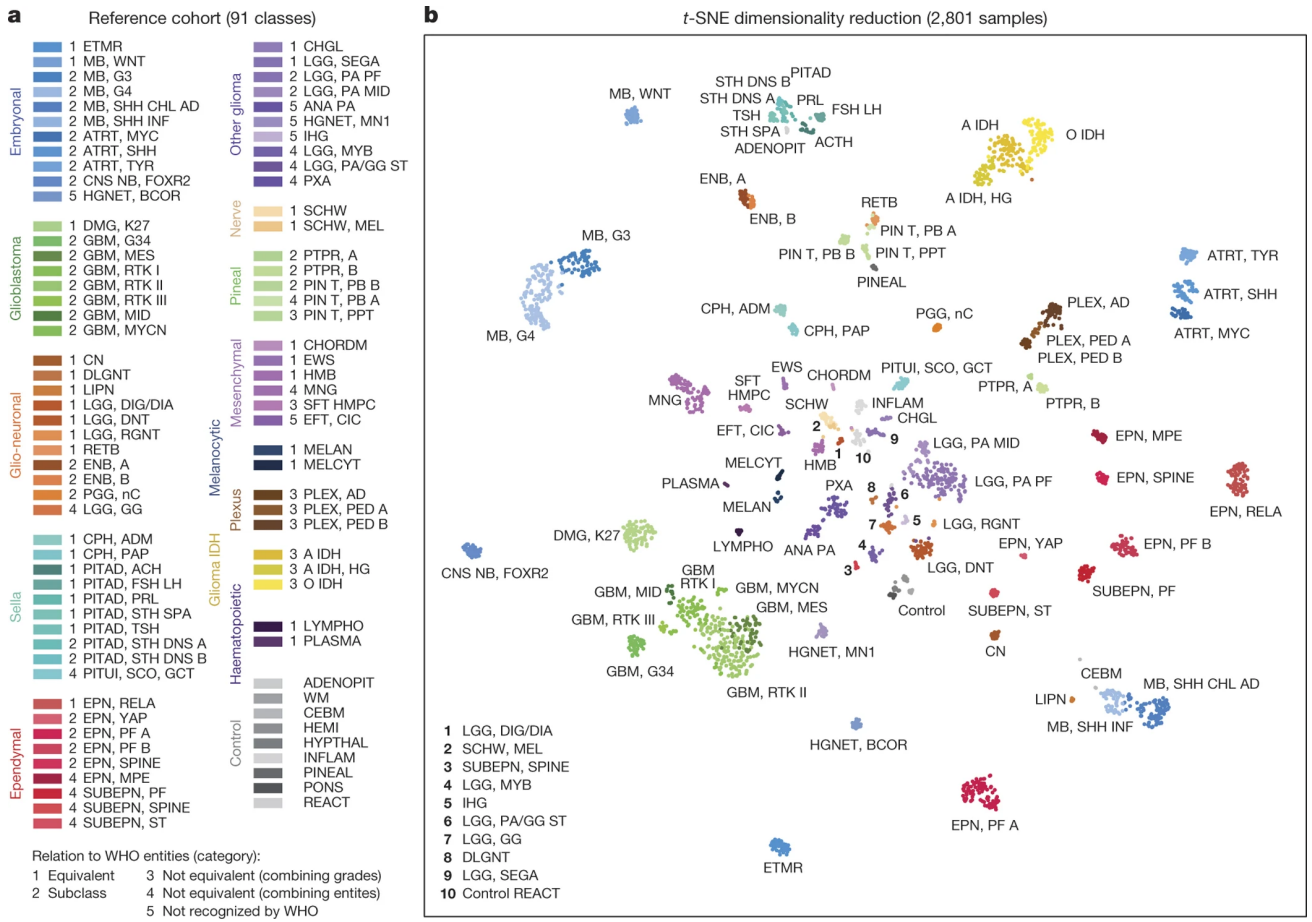
Dimension Reduction

Jun Kang

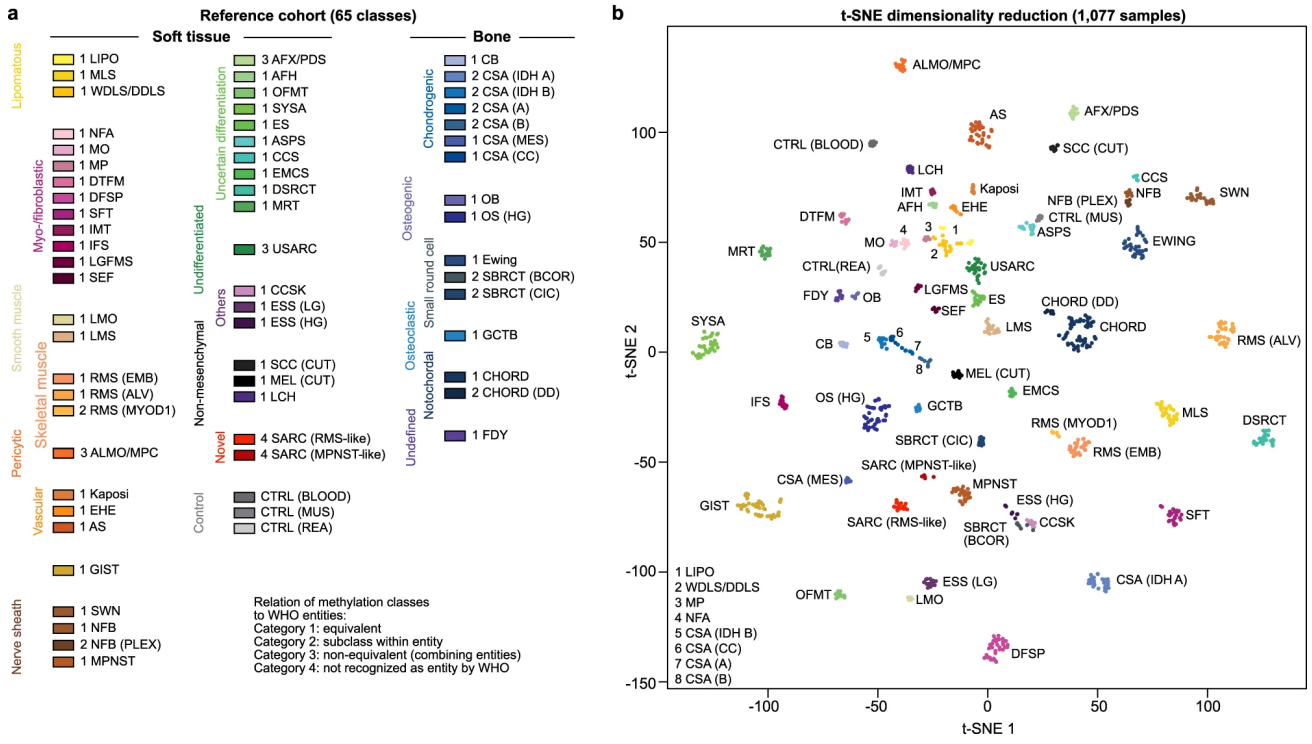
2022-04-29

DNA methylation-based classification of central nervous system tumours

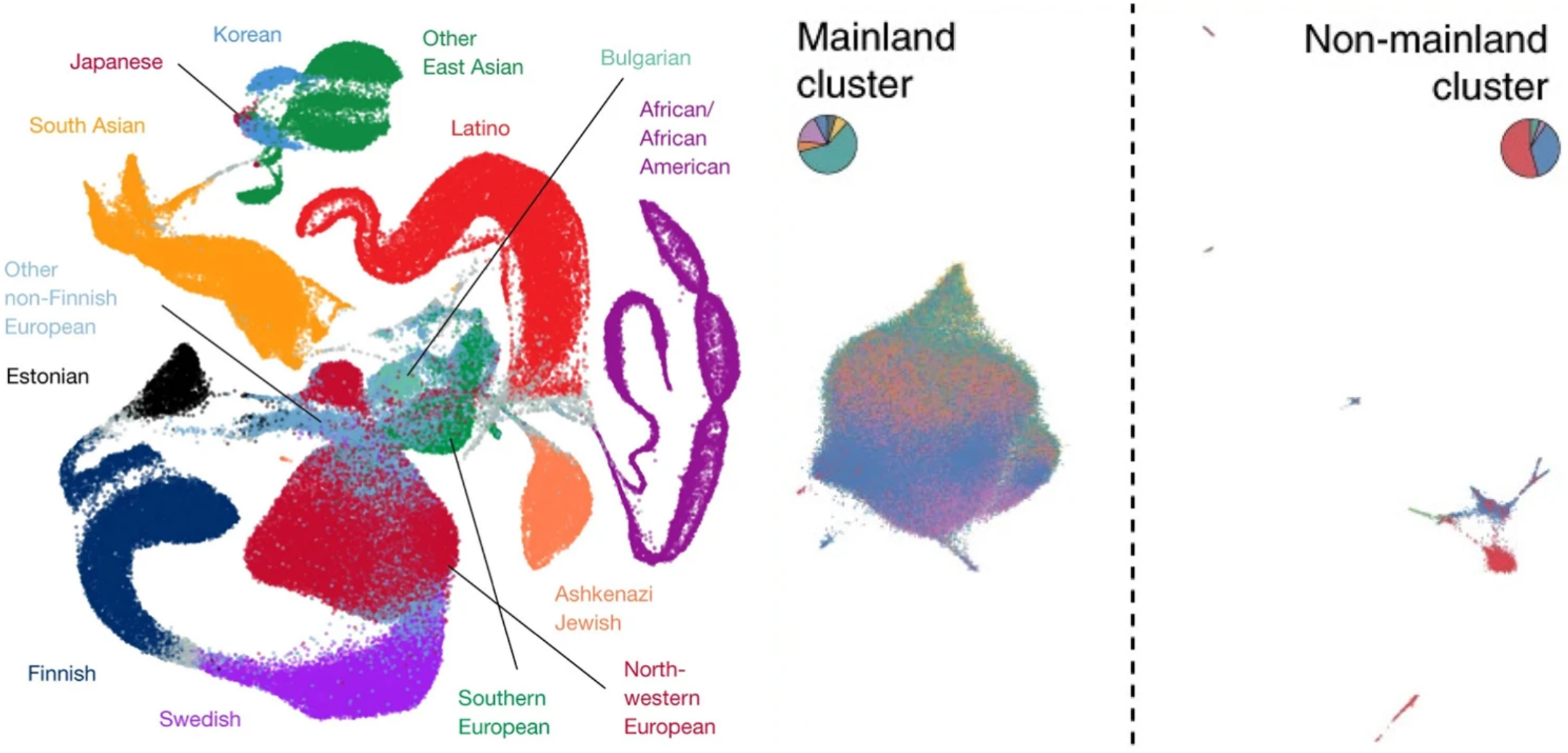


Nature 555, 469–474 (2018). https://doi.org/10.1038/nature26000

Sarcoma classification by DNA methylation profiling



Nat Commun 12, 498 (2021). https://doi.org/10.1038/s41467-020-20603-4



(a) gnomADv3 data visualized using UMAP.

(b) BBJ data visualized using UMAP.

J Hum Genet 66, 85–91 (2021). https://doi.org/10.1038/s10038-020-00851-4

High dimensional data

- Population genetics
- Single cell sequencing
- Spatial transcriptomics

Dimension reduction technique

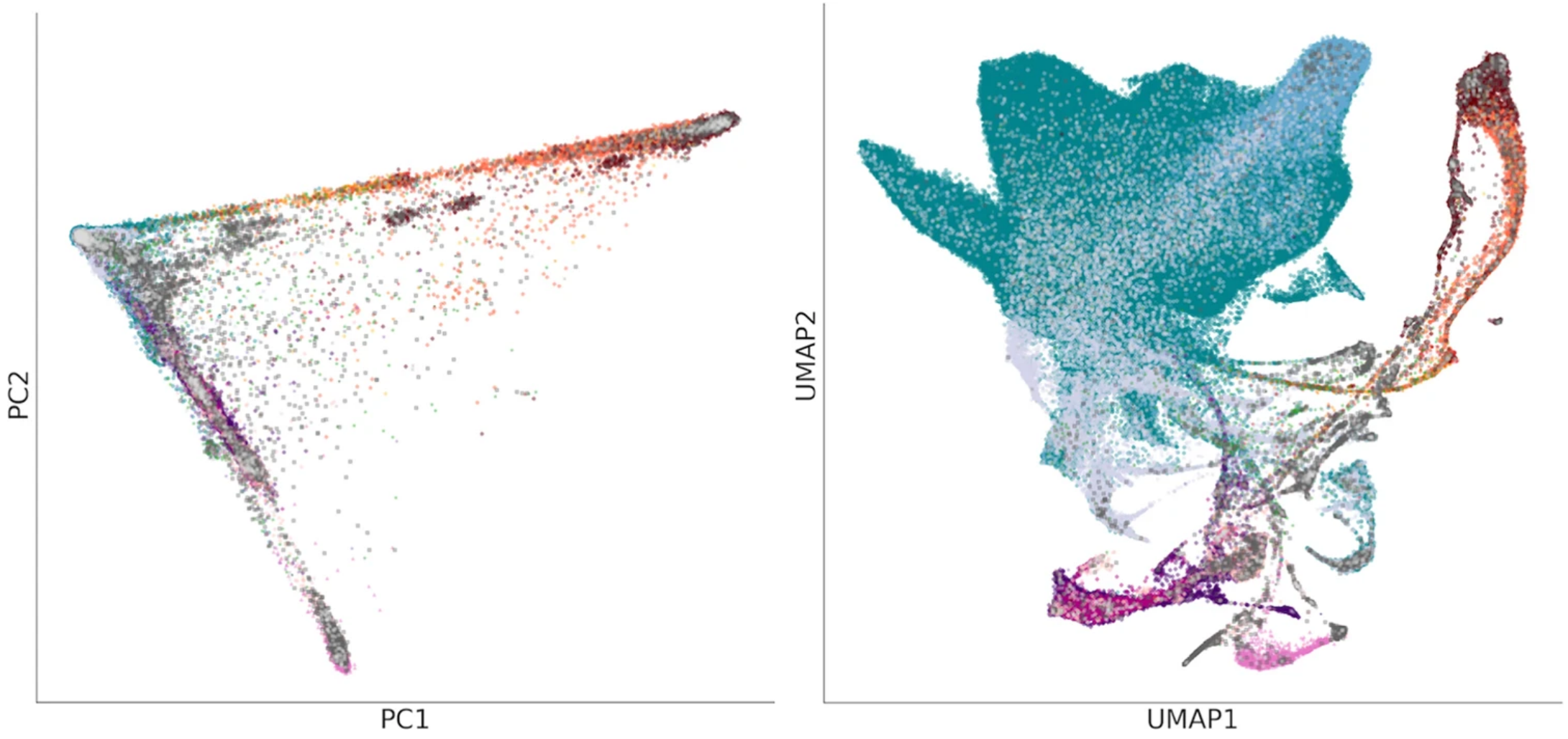
- Principal component analysis (PCA)
- t-Distributed Stochastic Neighbor Embedding (t-SNE)
- Uniform Manifold Approximation and Projection (UMAP)

Dimension reduction

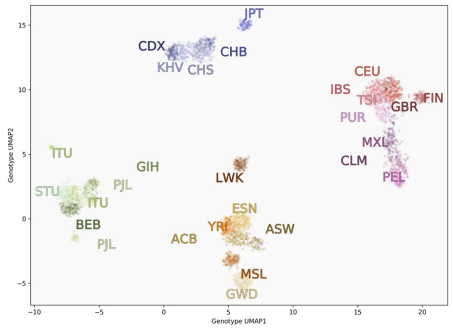
- Change and select basis to clustering
- Visualization in 2 dimensional space

Concepts

- Dimension
- Basis or Latent features
- Graph
- Projection or Embedding
- Linearity

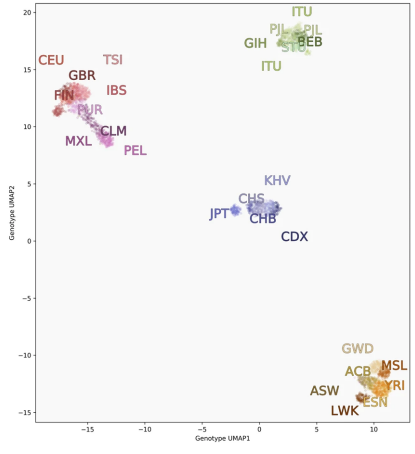


J Hum Genet 66, 85–91 (2021). https://doi.org/10.1038/s10038-020-00851-4

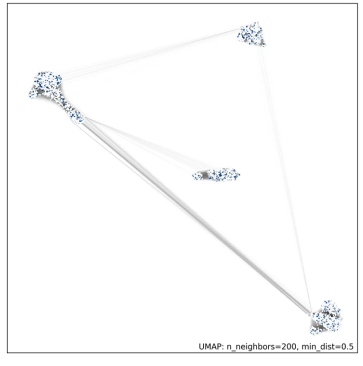


(a) UMAP with 15 neighbours.

(b) Connectivity map of 15 neighbours.



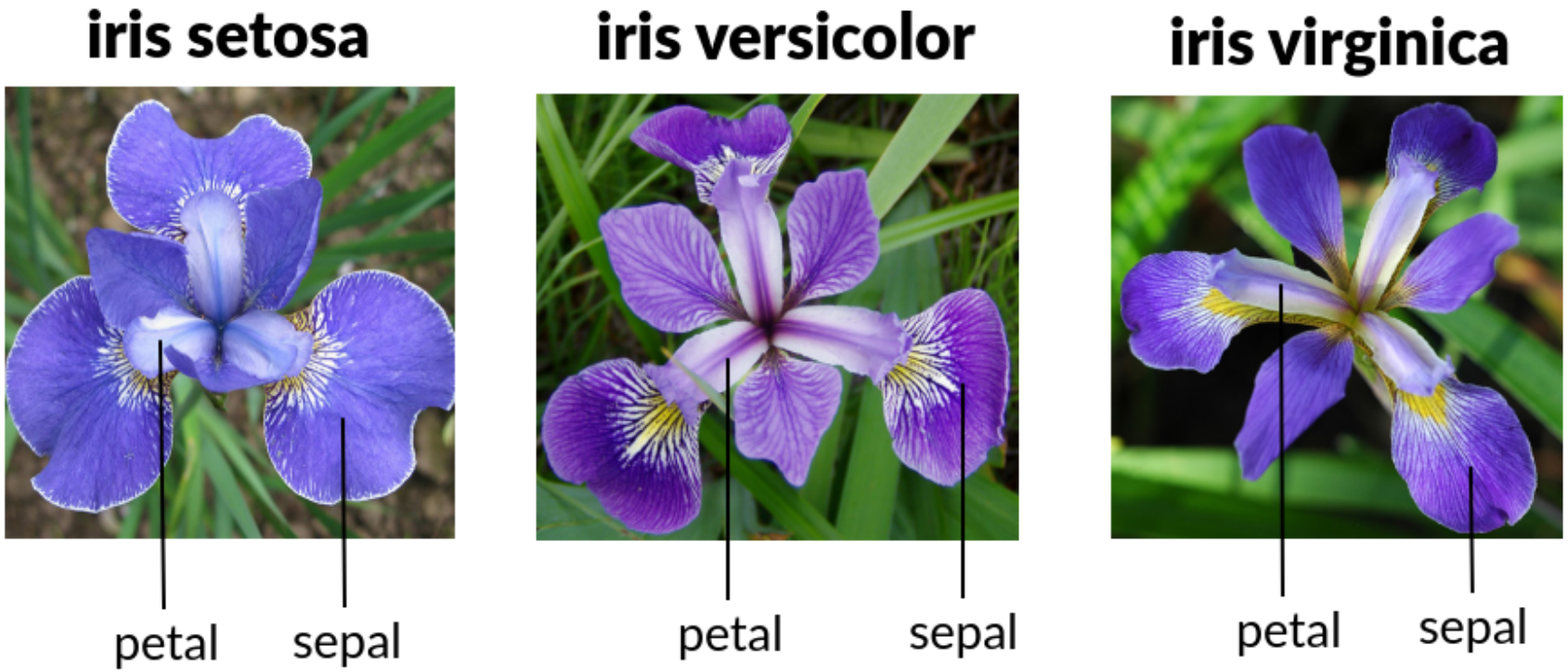
(c) UMAP with 200 neighbours.



(d) Connectivity map of 200 neighbours.

J Hum Genet 66, 85–91 (2021). https://doi.org/10.1038/s10038-020-00851-4

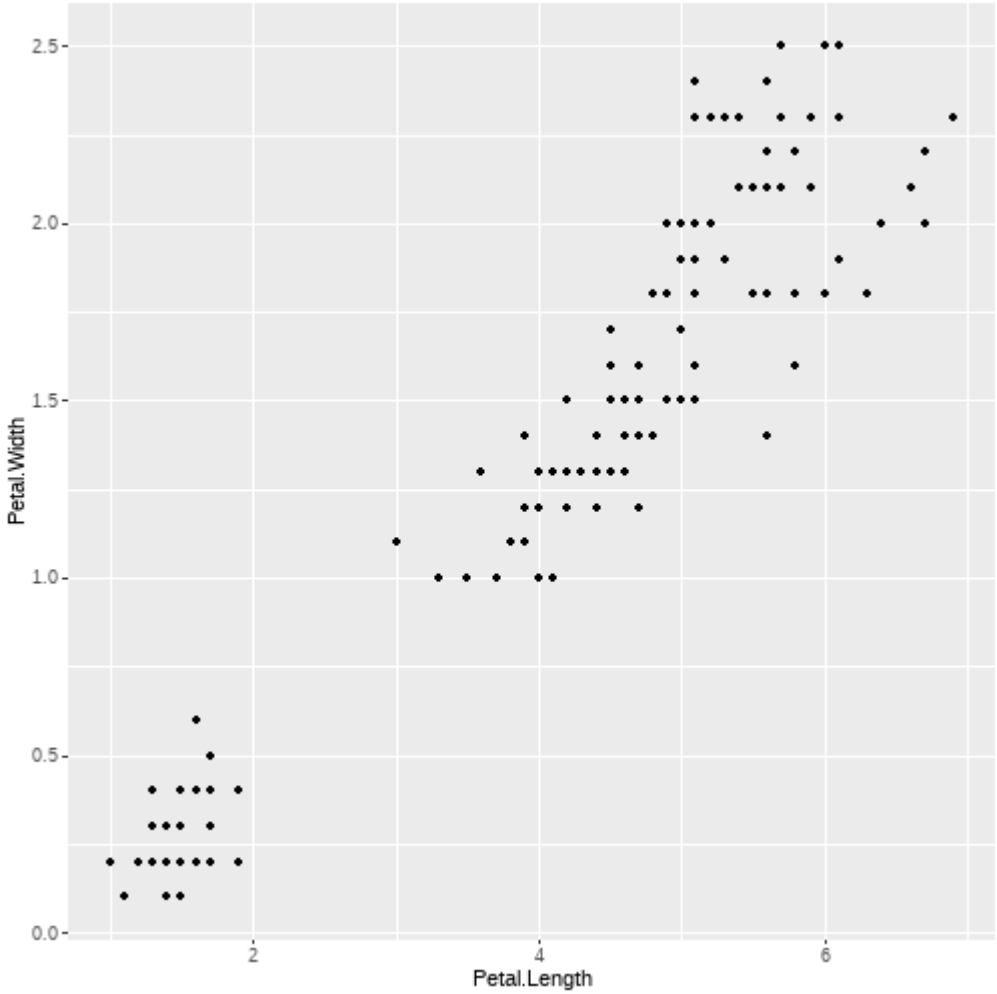
Iris



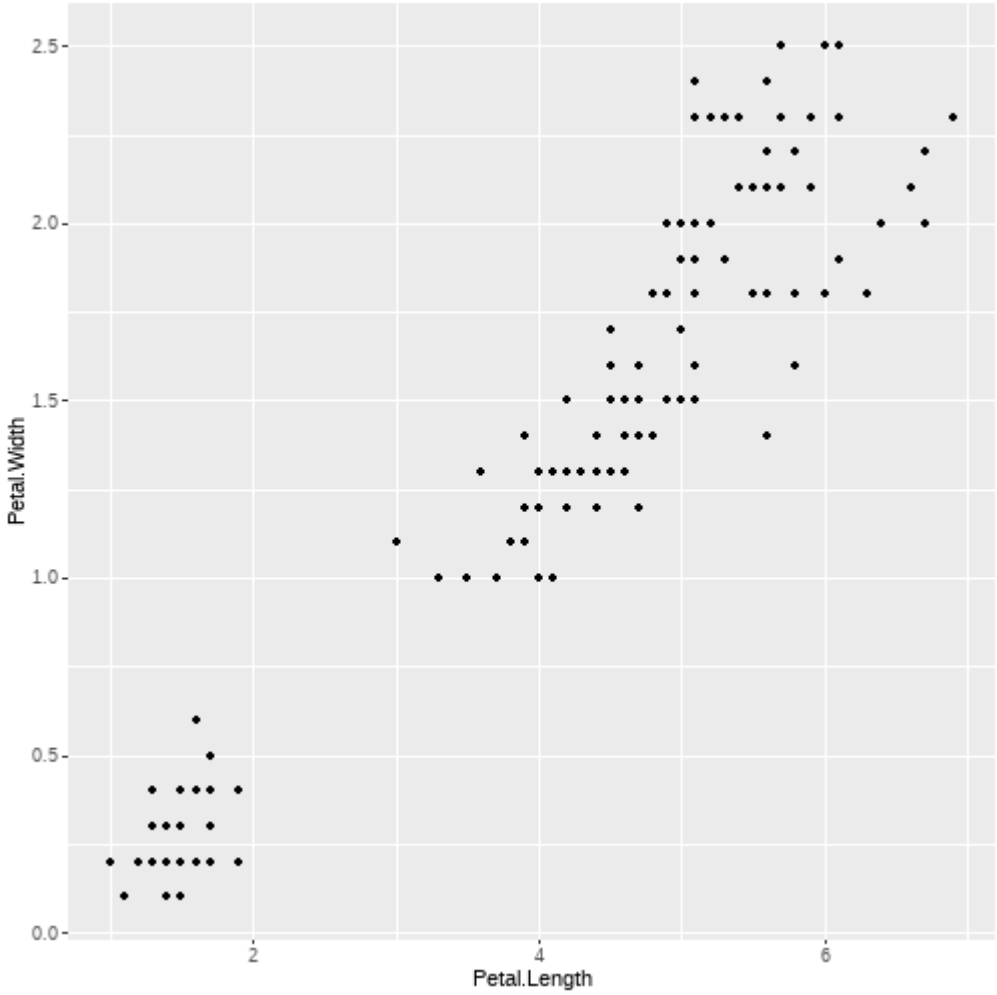
/ˈpedl/, /ˈsēpəl/

Iris data

Dimension reducsion by human selection



Clustering? and select one dimension



Principal component analysis (PCA)

Linearity

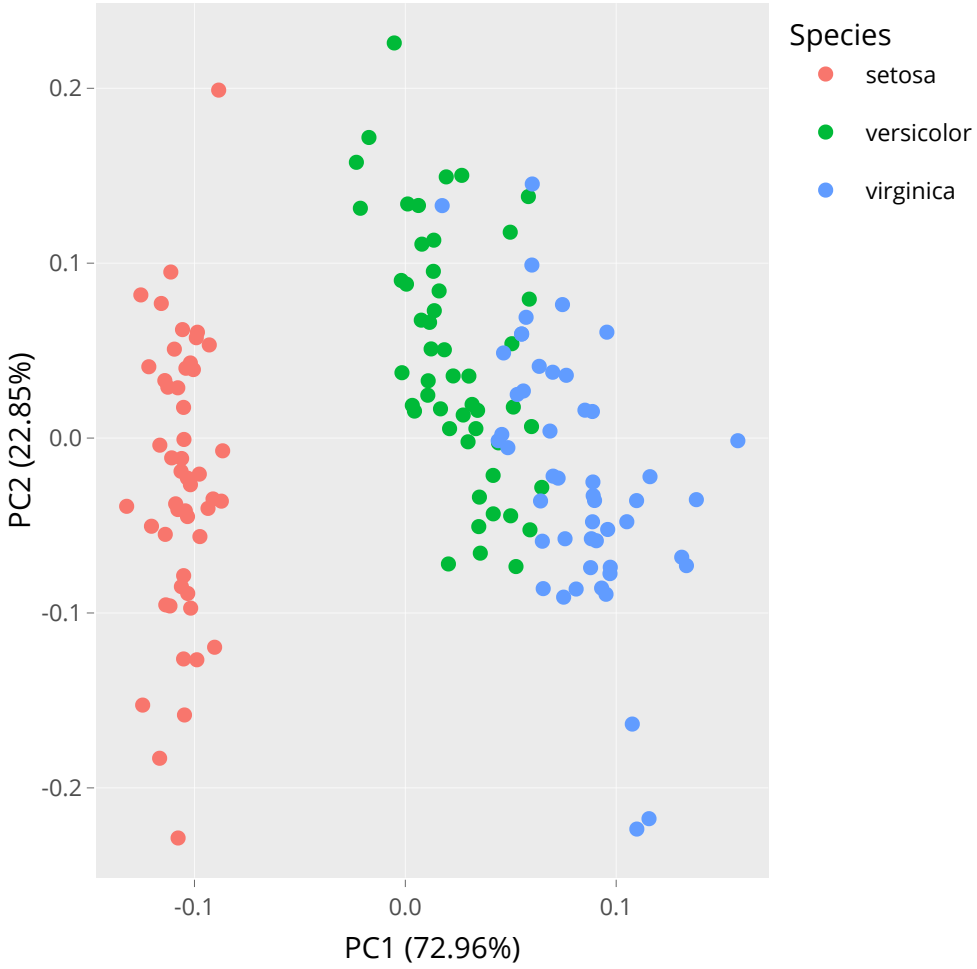
$$PC1 = \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 \ldots + \alpha_n v_n$$
$$PC2 = \beta_2 v_1 + \beta_2 v_2 + \beta_3 v_3 \ldots + \beta_n v_n$$

Projection

Orthogonality

- PCs are orthogonal each other

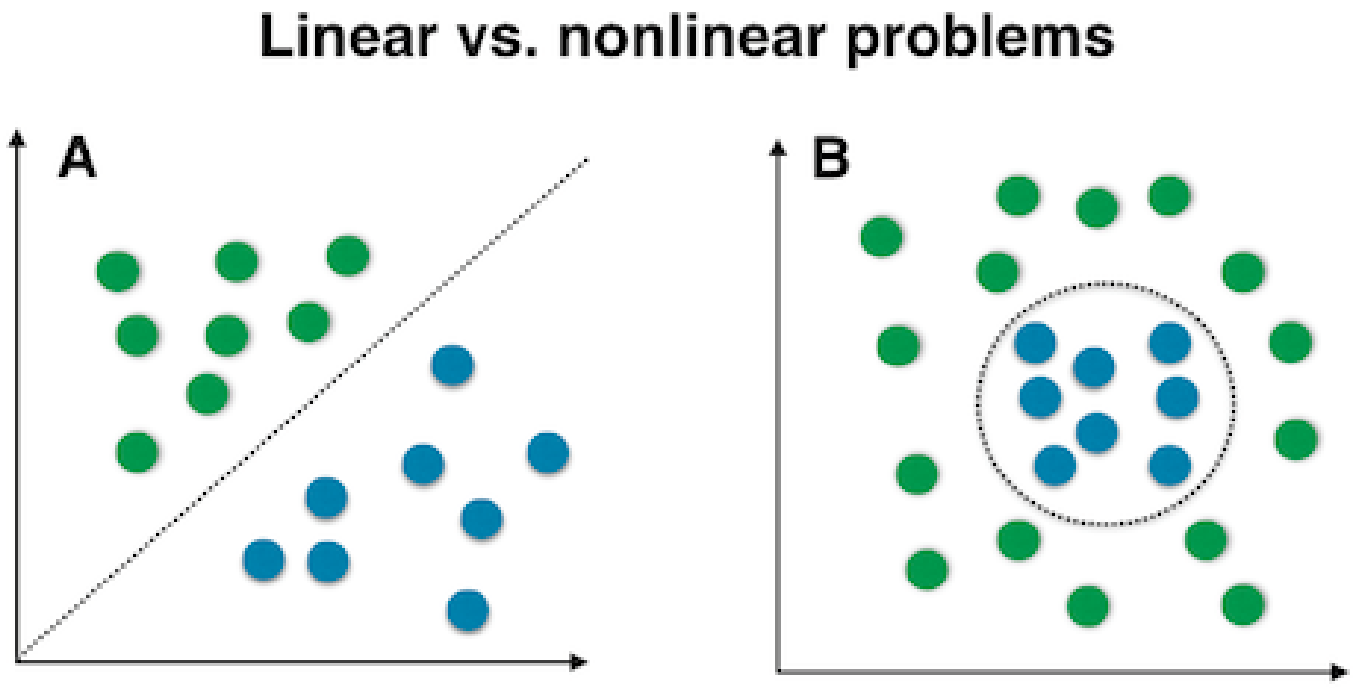
Principal component analysis (PCA)



Latent variables (features)

- Linear
- Non-linear

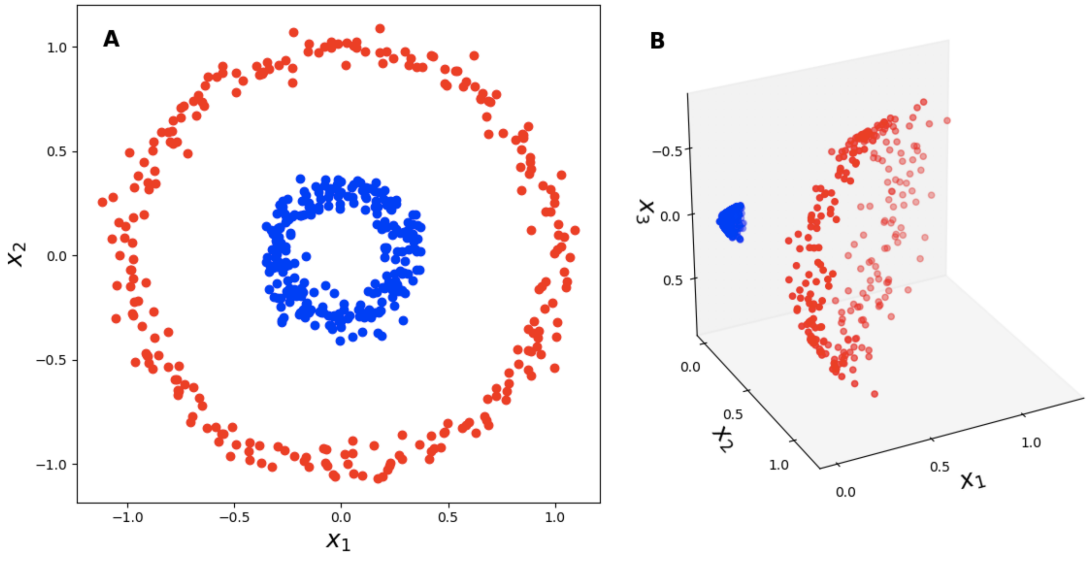
Non-linear



http://rasbt.github.io/mlxtend/user_guide/feature_extraction/RBFKernelPCA/

Non-linear

- Feature expansion: Lower dimension to higher dimension
- Gaussian Kernel (Radial basis function kernel)

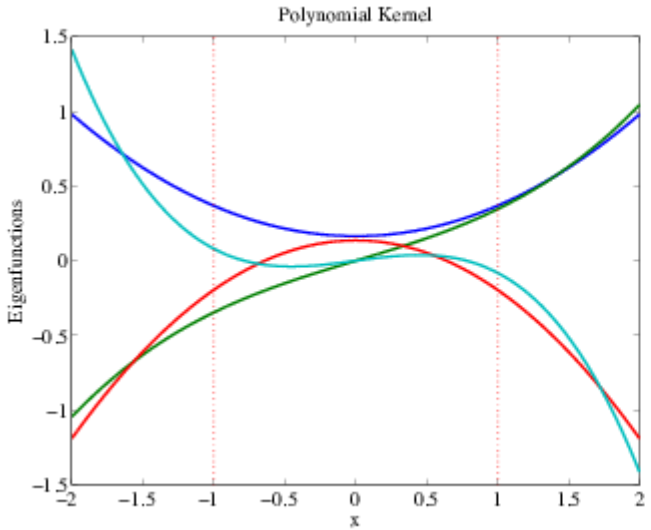


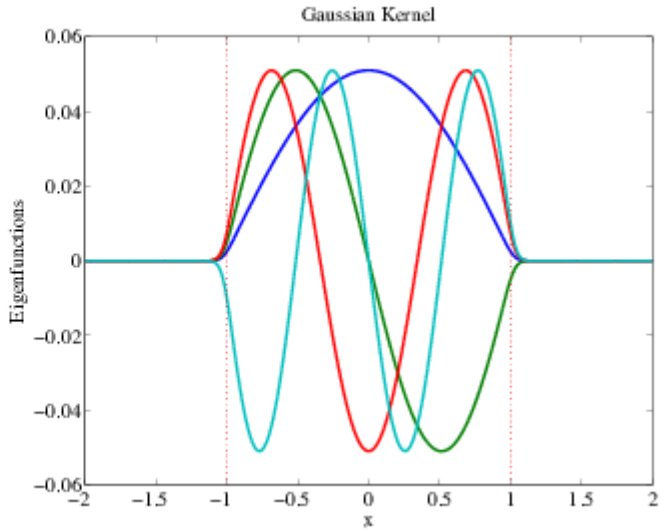
The "lifting trick". (a) A binary classification problem that is not linearly separable in $\mathbb{R}^2$ (b) A lifting of the data into $\mathbb{R}^3$ using a polynomial kernel, $\varphi([x_1 \;\; x_2]) = [x_1^2 \;\; x_2^2 \;\; \sqrt{2}x_1x_2]$

Polynomial kernal $(x_1 + x_2)^2 = x_1^2 + x_2^2 + 2x_1x_2$

https://gregorygundersen.com/blog/2019/12/10/kernel-trick/

Kernel basis functions (Eigenfunctions)





Electronic Journal of Statistics. 10. 423-463. 10.1214/16-EJS1112.

Manifold Based Dimensionality Reduction

- Key Assumption: Points live on a low dimensional manifold
- Manifold: subspace that looks locally Euclidean
- Given data, can we uncover this manifold?

Can we unfold this?

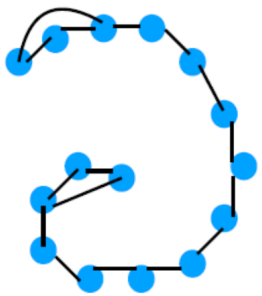
https://www.cs.cornell.edu/courses/cs4786/2020sp/lectures/lec08.pdf

Method I: Isomap

1. For every point, find its (k-) Nearest Neighbors

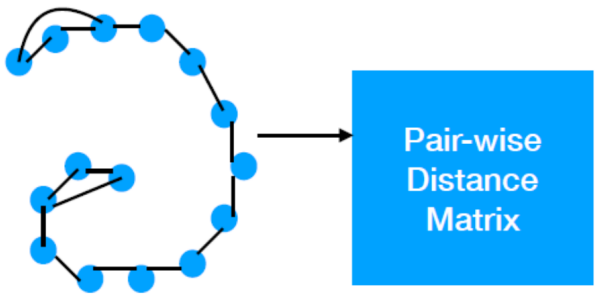
Method I: Isomap

1. For every point, find its (k-) Nearest Neighbors
2. Form the Nearest Neighbor graph



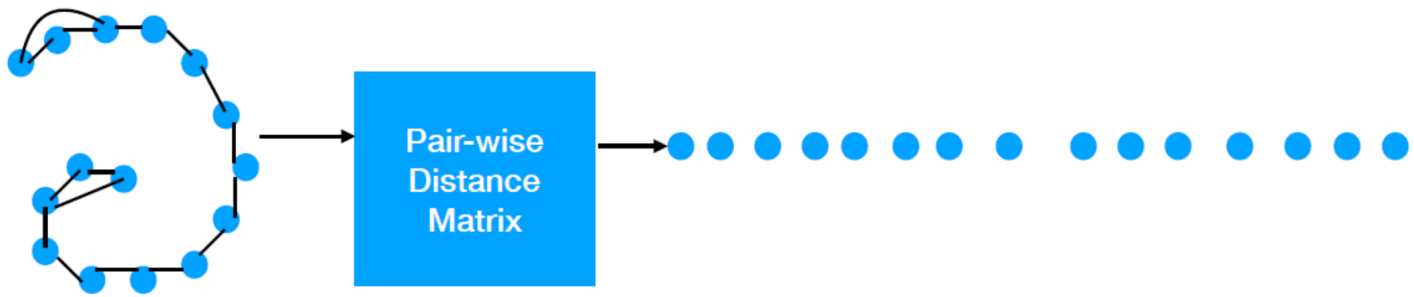
Method I: Isomap

1. For every point, find its (k-) Nearest Neighbors
2. Form the Nearest Neighbor graph
3. For every pair of points A and B, distance between point A to B is shortest distance between A and B on graph



Method I: Isomap

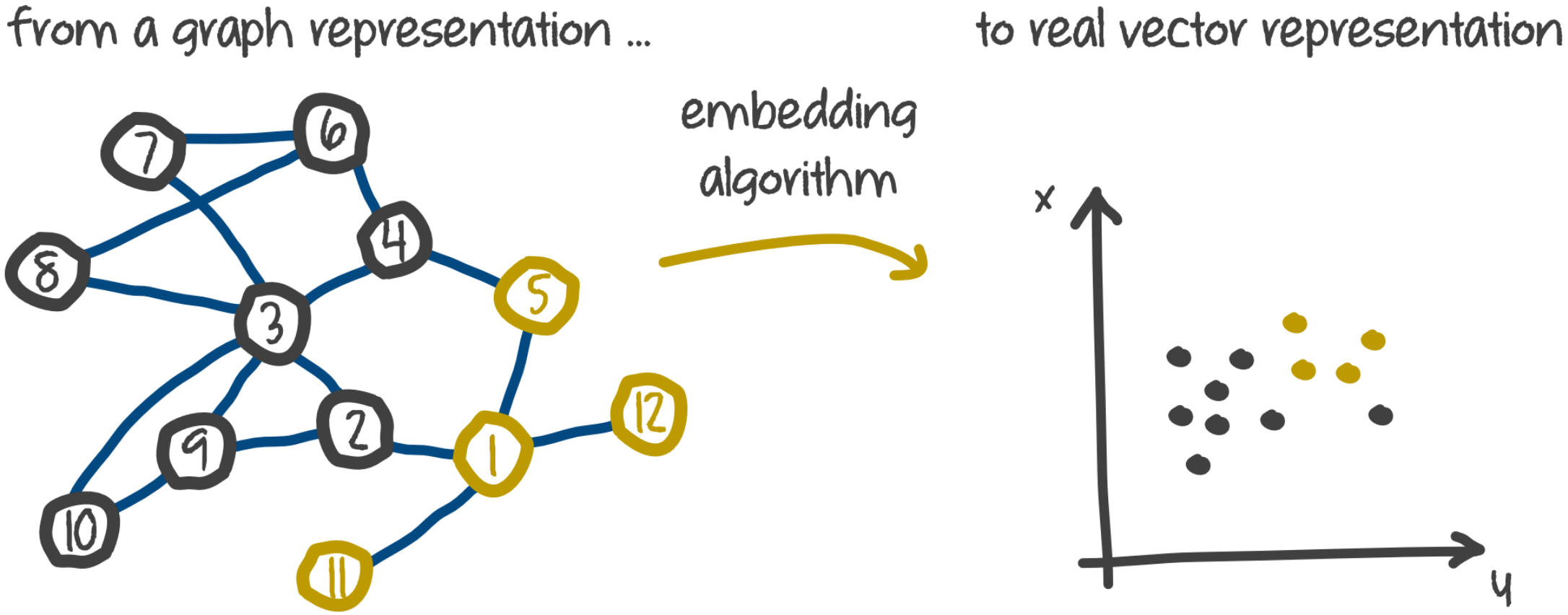
1. For every point, find its (k-) Nearest Neighbors
2. Form the Nearest Neighbor graph
3. For every pair of points A and B, distance between point A to B is shortest distance between A and B on graph
4. Find points in low dimensional space such that distances between points in this space is equal to distance on graph.



t-Distributed Stochastic Neighbor Embedding (t-SNE)

https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

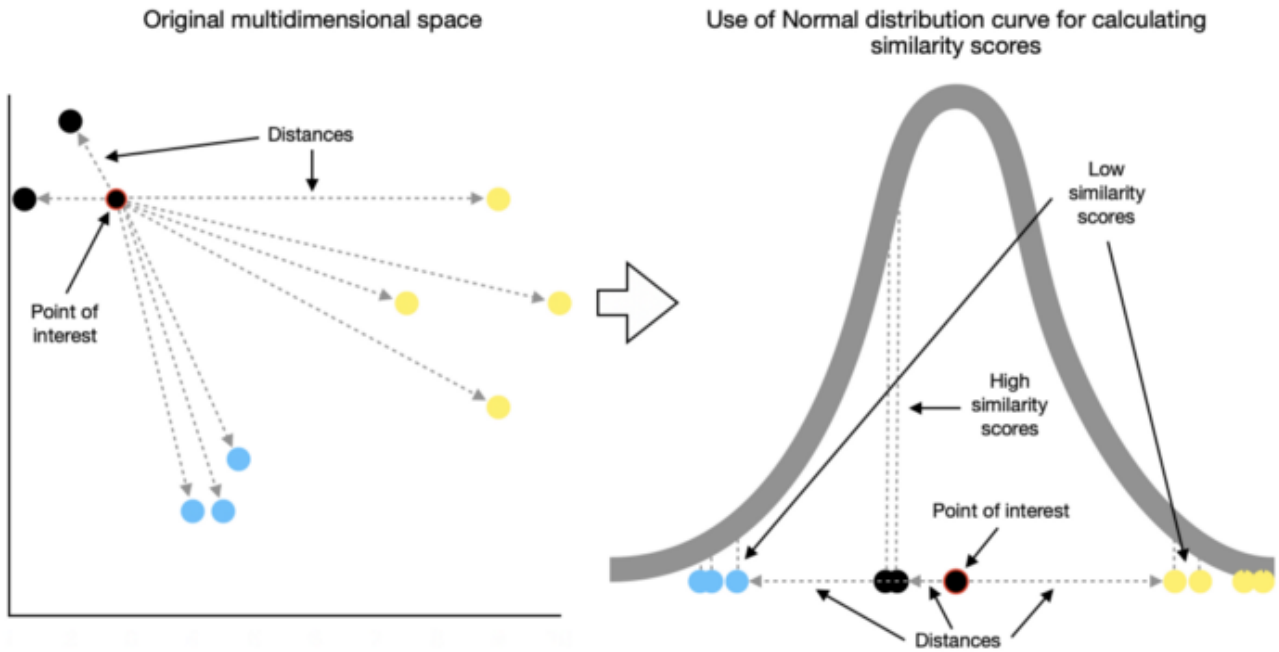
Graph (Neighbor) Embedding



https://gearons.org/blog/2016/MAGE/

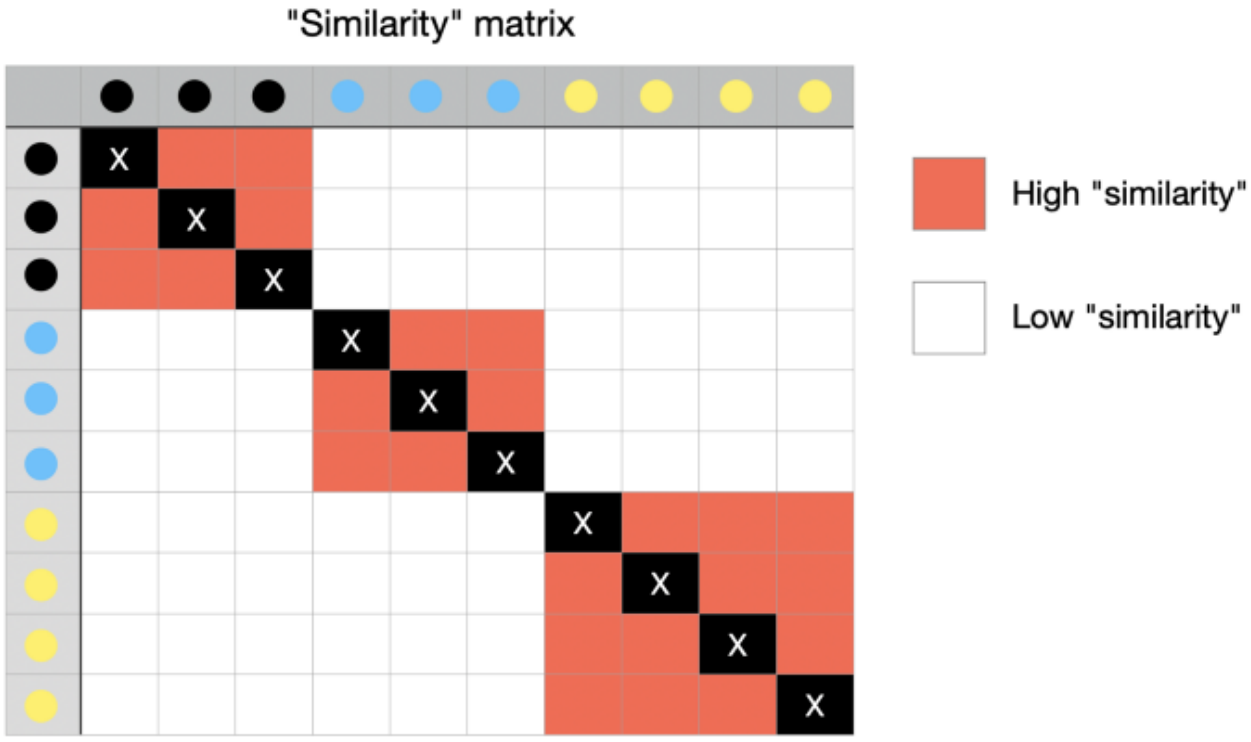
Probablistic/Stochastic Neighbor Embedding (SNE)

Similarity scores (probablistic/stochastic)



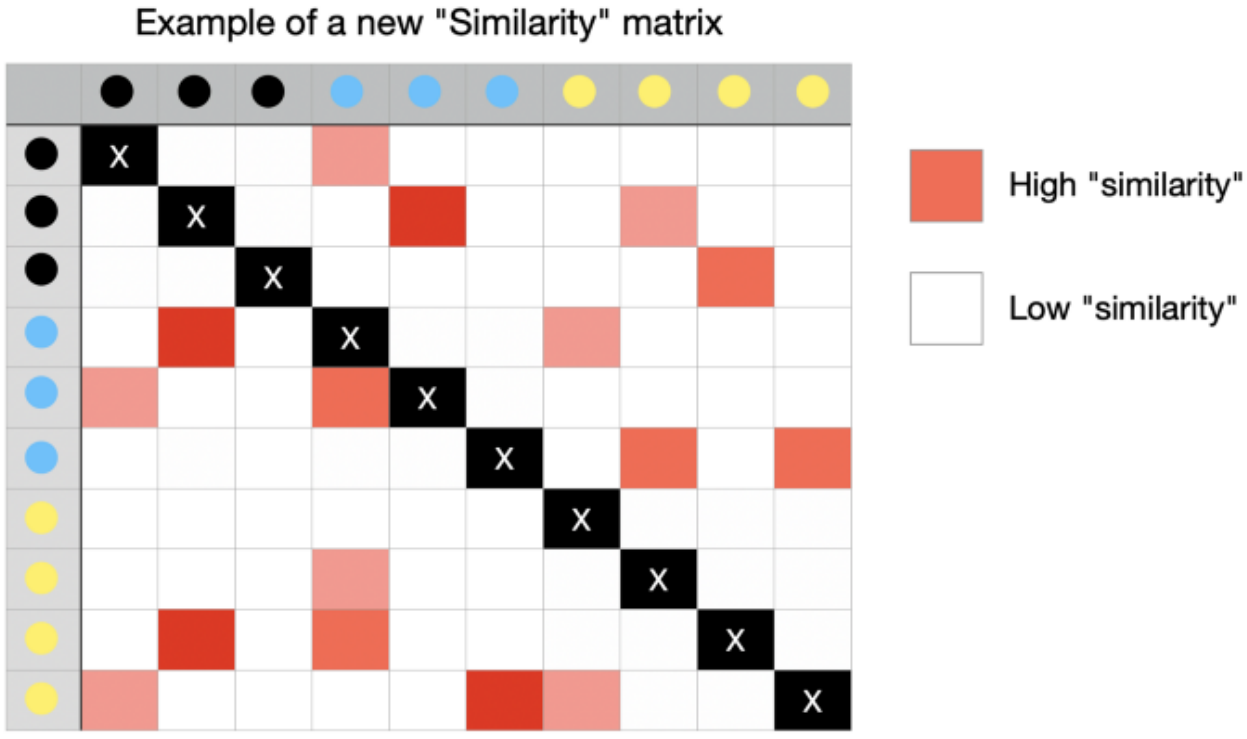
https://towardsdatascience.com/t-sne-machine-learning-algorithm-a-great-tool-for-dimensionality-reduction-in-python-ec01552f1a1e

Similarity matrix (High dimension)



https://towardsdatascience.com/t-sne-machine-learning-algorithm-a-great-tool-for-dimensionality-reduction-in-python-ec01552f1a1e

Similarity matrix (low dimension initial)

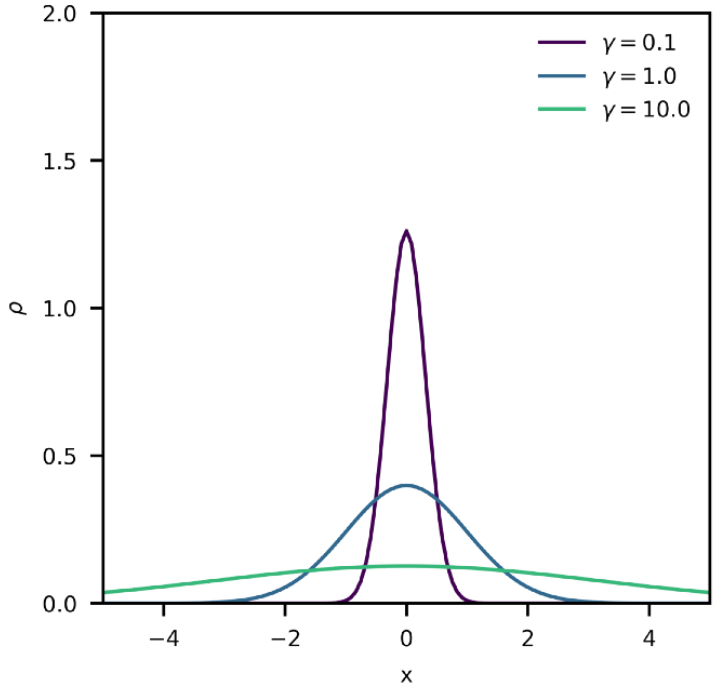


https://towardsdatascience.com/t-sne-machine-learning-algorithm-a-great-tool-for-dimensionality-reduction-in-python-ec01552f1a1e

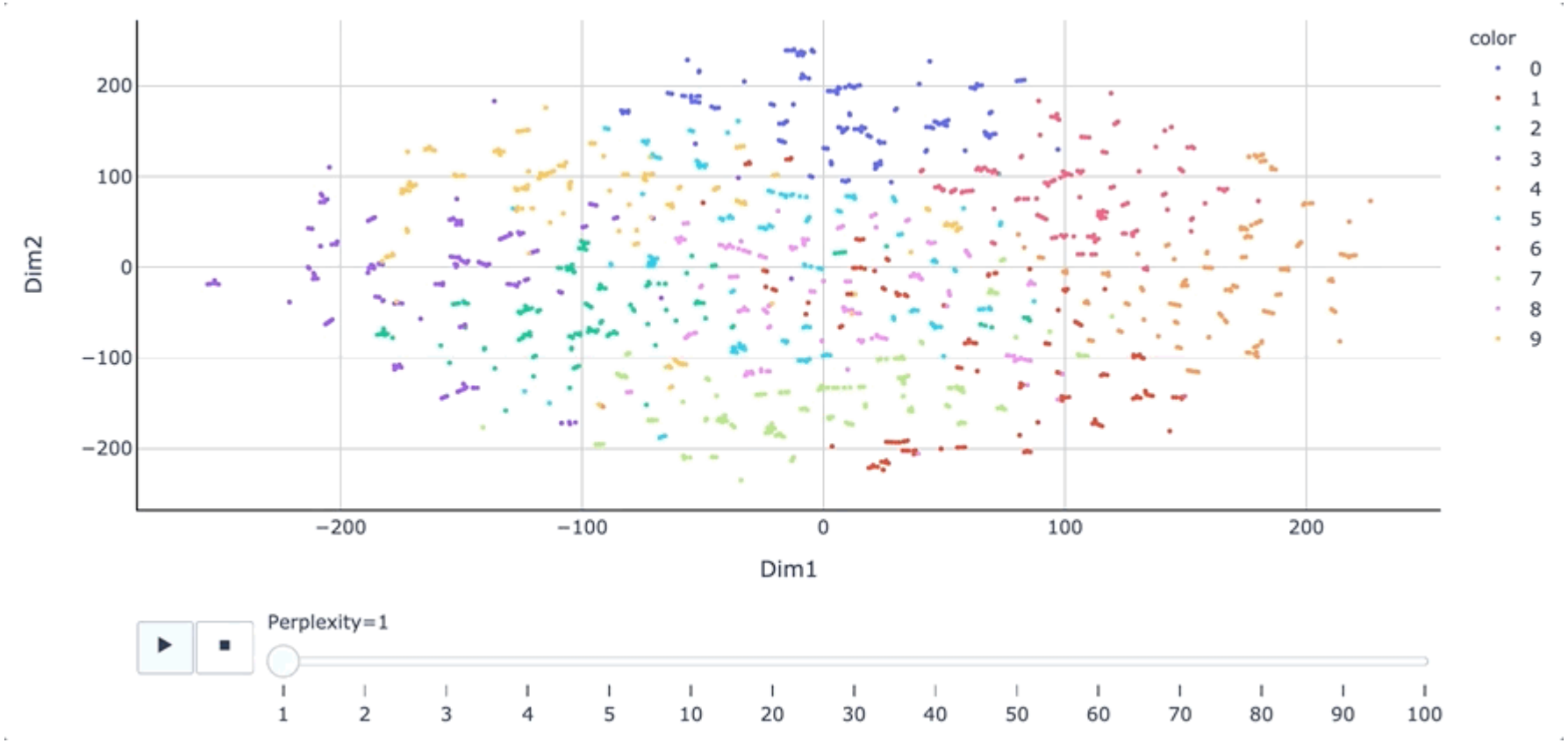
Minimize the Kullback–Leibler divergence (KL divergence) through gradient descent.

- Learning rate
- Iteration number

Perplexity



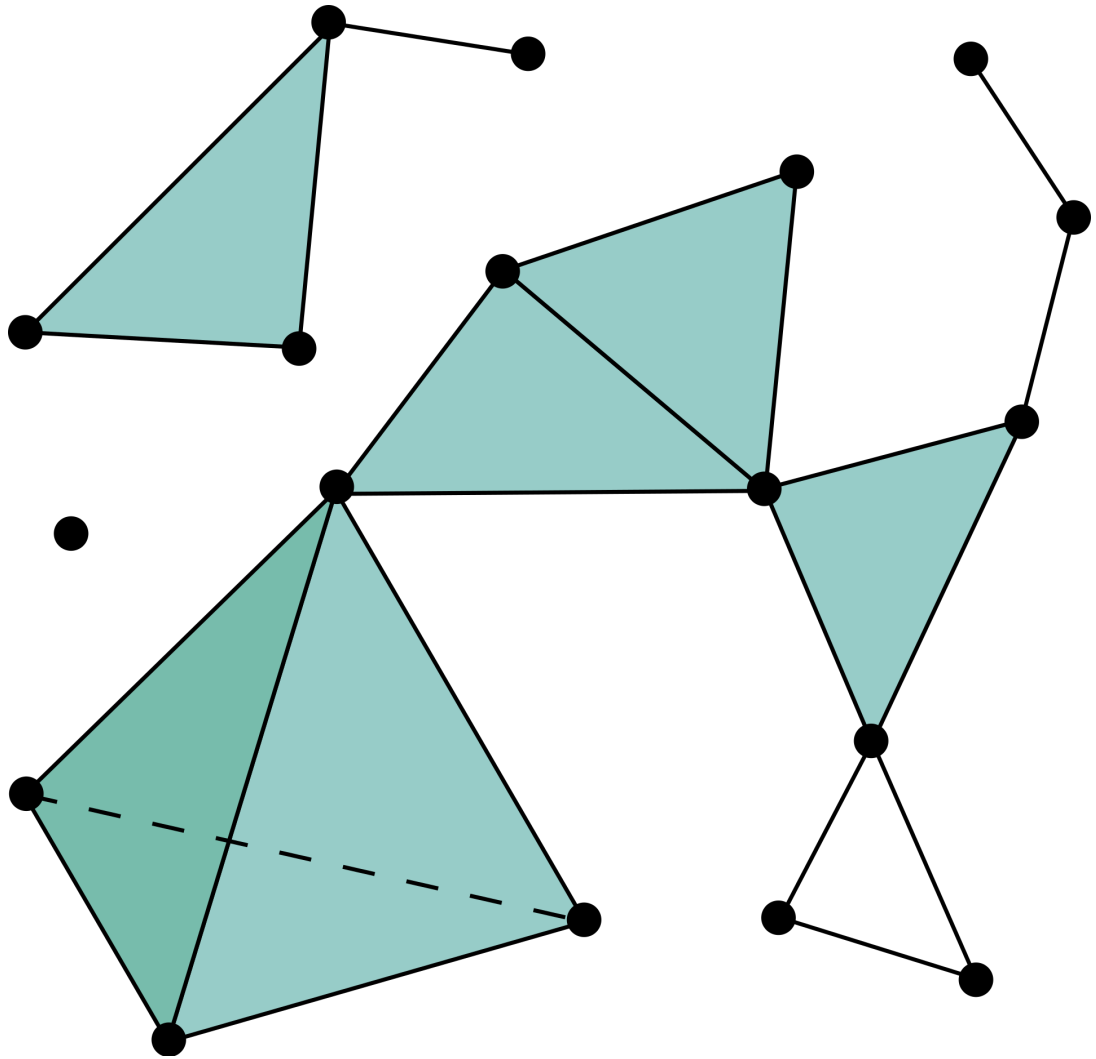
Perplexity



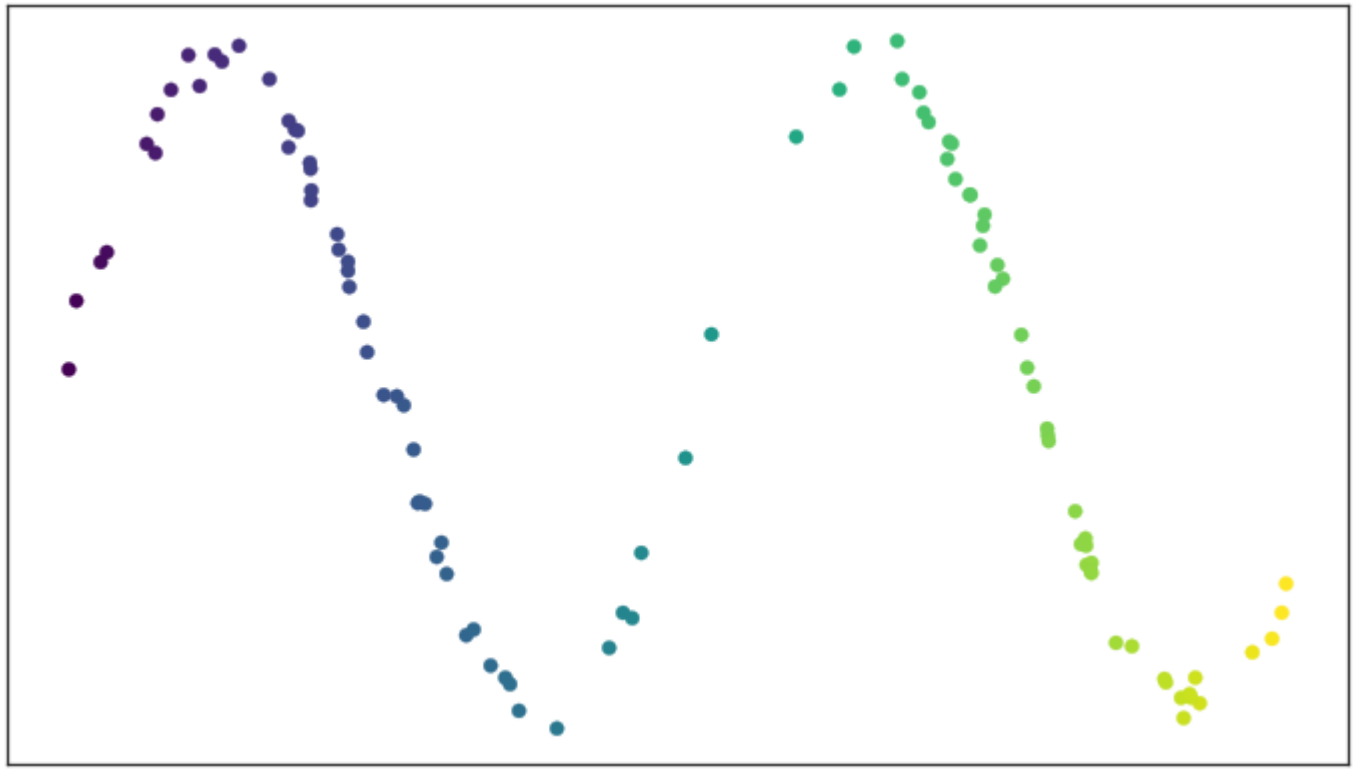
https://towardsdatascience.com/t-sne-machine-learning-algorithm-a-great-tool-for-dimensionality-reduction-in-python-ec01552f1a1e

https://distill.pub/2016/misread-tsne/

Simplicial complex

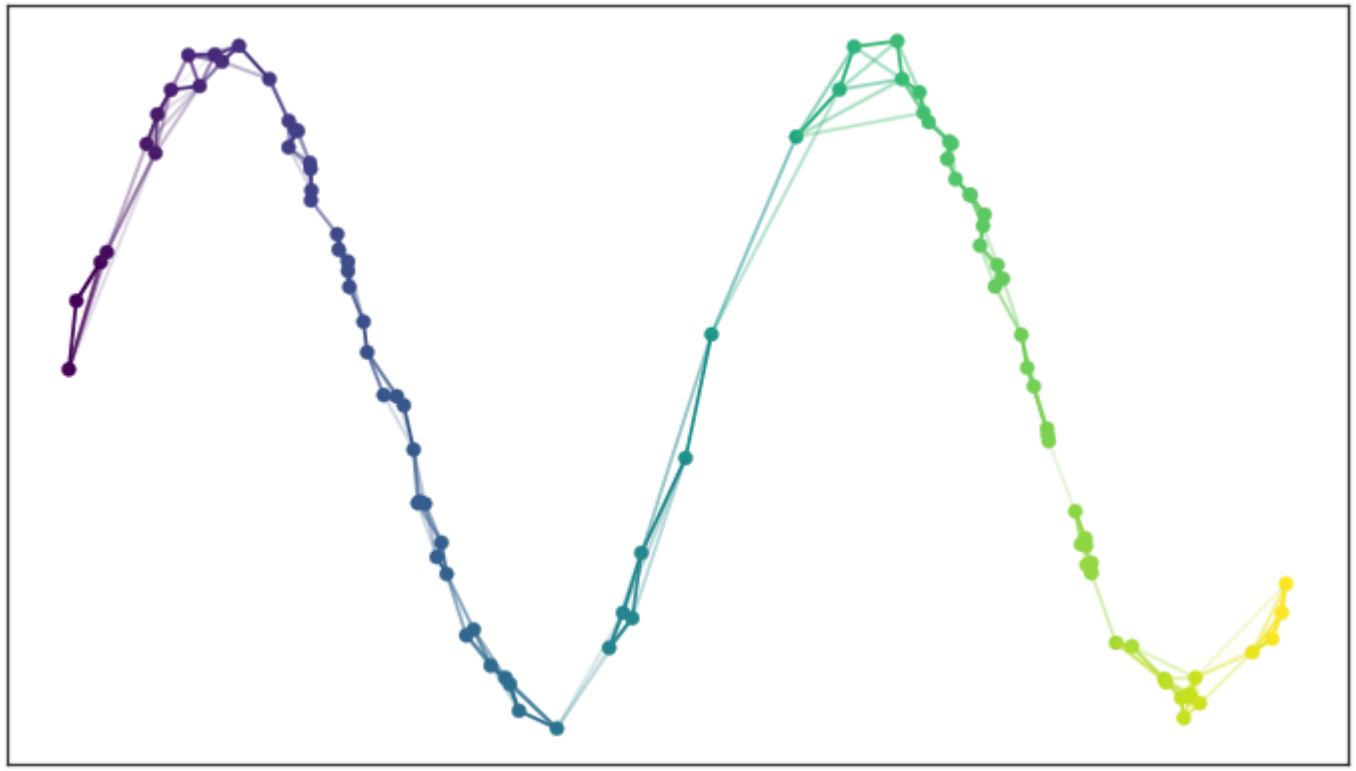


Test data set of a noisy sine wave



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

Graph with combined edge weights



Minimize cross entropy

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)$$

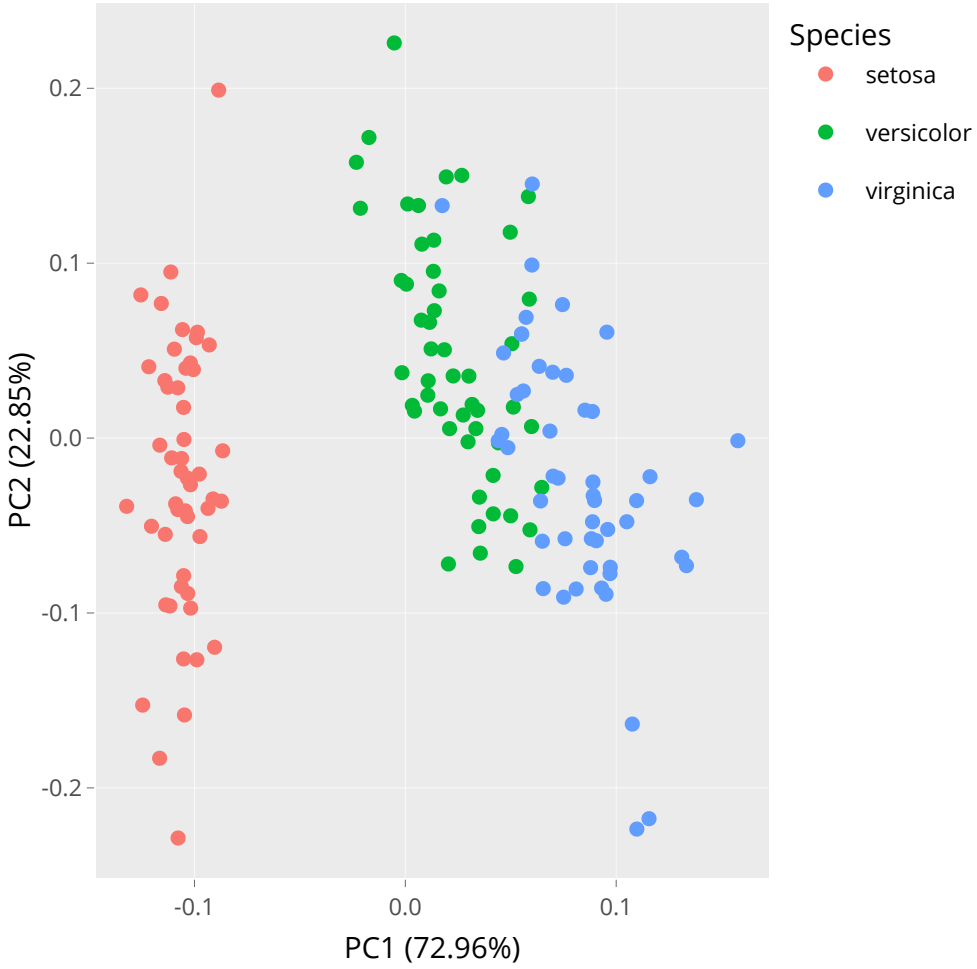
Similar and different with t-SNE

- Similarity
  - Exponential probability distribution
- Minimize cross entropy
  - Gradient descent

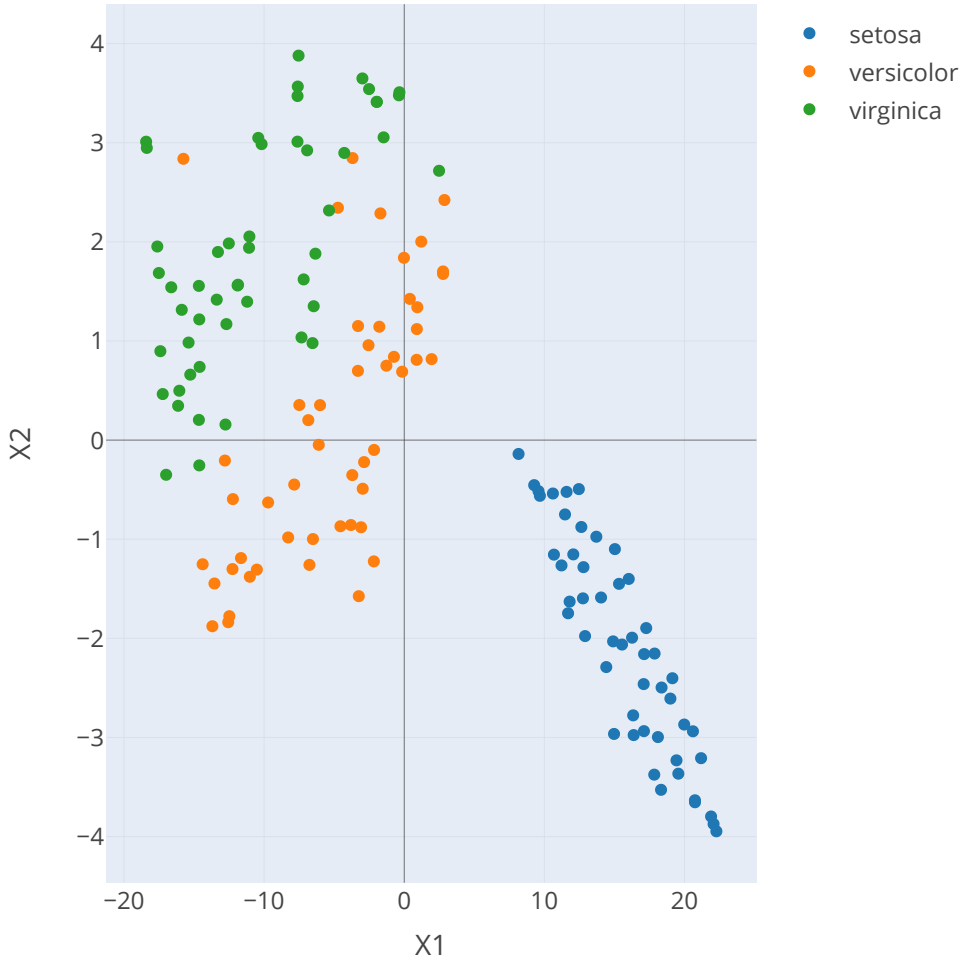
Parameters of UMAP

- Number of neighbors
- Minimal distance
- Learning rate
- Number of epoch

Principal component analysis (PCA)



t-Distributed Stochastic Neighbor Embedding (t-SNE)



Uniform Manifold Approximation and Projection (UMAP)