Dimensionality Reduction

Chongzhi Zang

Acknowledgements: Some materials in the slides are borrowed from
Caltech Bi/BE/CS 183 course taught by Lior Pachter (under CC BY 4.0 license).
Video explanations are from StatQuest youtube channel by Josh Starmer.
Copyright of images from internet belongs to their respective owners.

Outline

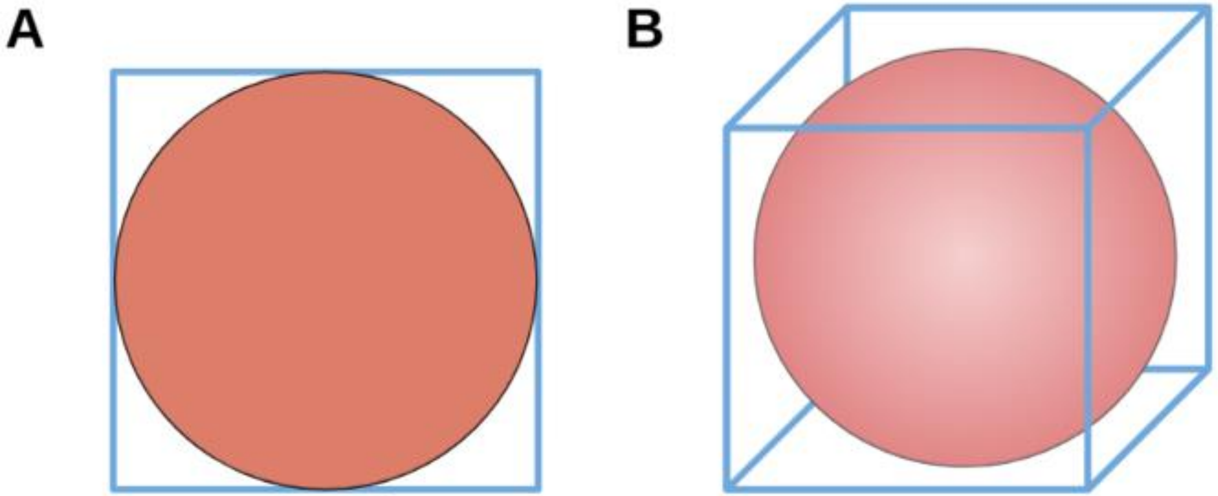
- Motivation
- Principal component analysis (PCA)
- t-distributed stochastic neighbor embedding (t-SNE)
- Uniform manifold approximation and projection (UMAP)
- Autoencoder

Motivation

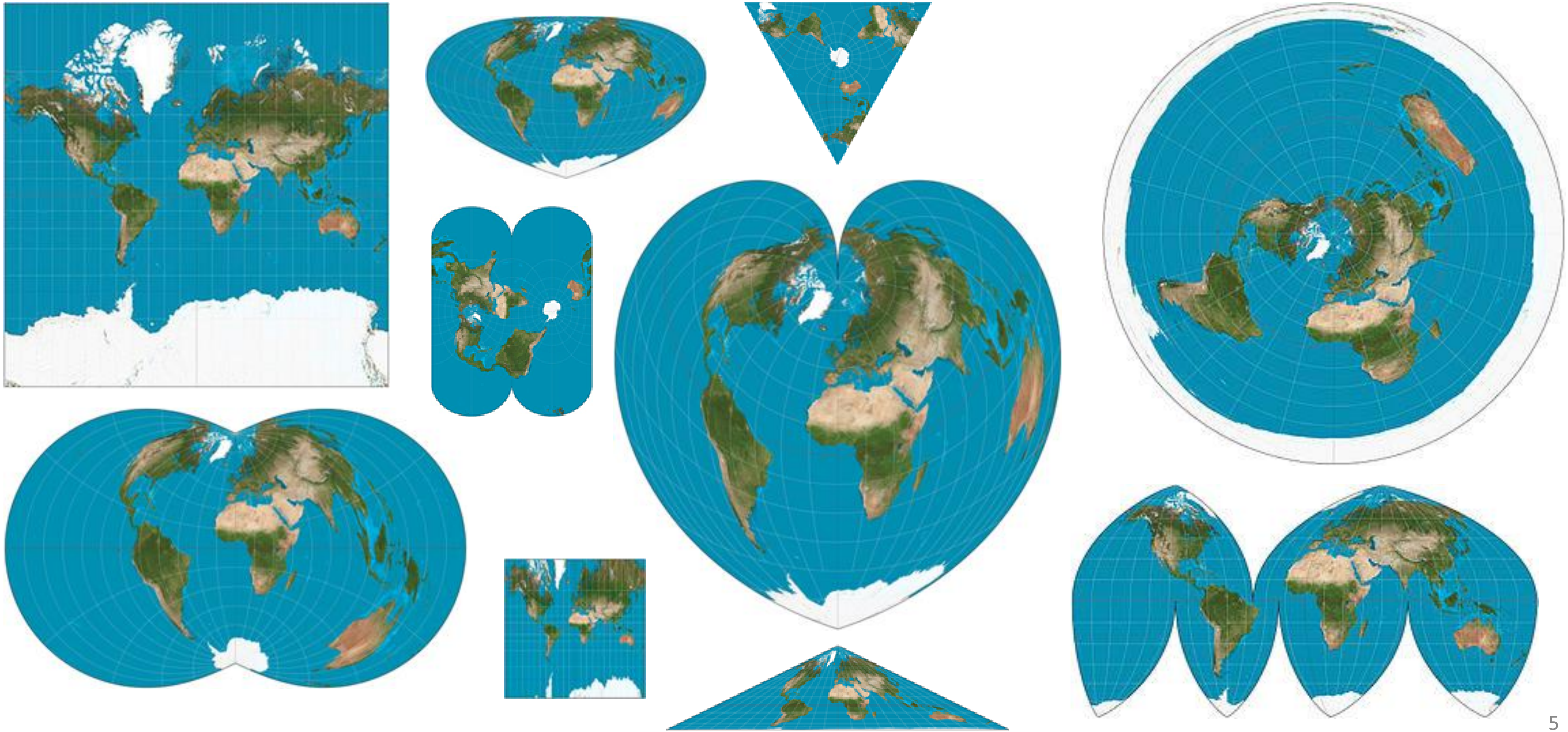
- Curse of dimensionality
- High-throughput biological data can be high-dimensional
- Data analysis and data visualization

Curse of dimensionality

- Higher dimension: more space
- A random point in the space tends to be far from the center and close to the border
- Sparsity: average distance between two random points:
  - in a unit square is ~0.52
  - in a unit 3D cube is ~0.66
  - in a unit 10^6D hypercube is ~408.25
- Sampling effort increases exponentially



Map projection is 3D→2D dimensionality reduction

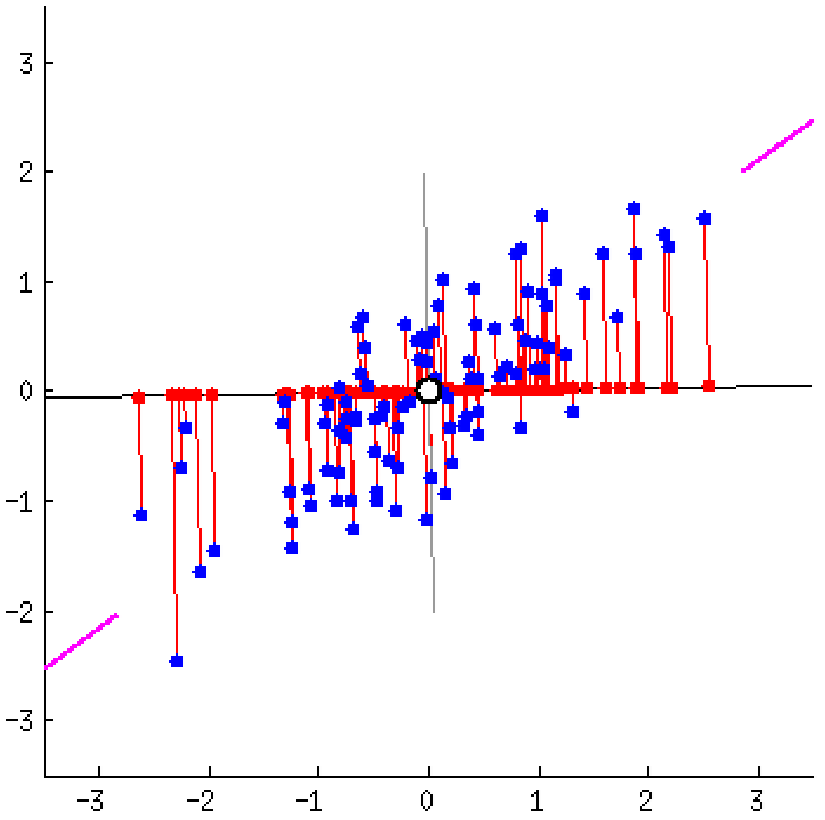


Principal component analysis (PCA)

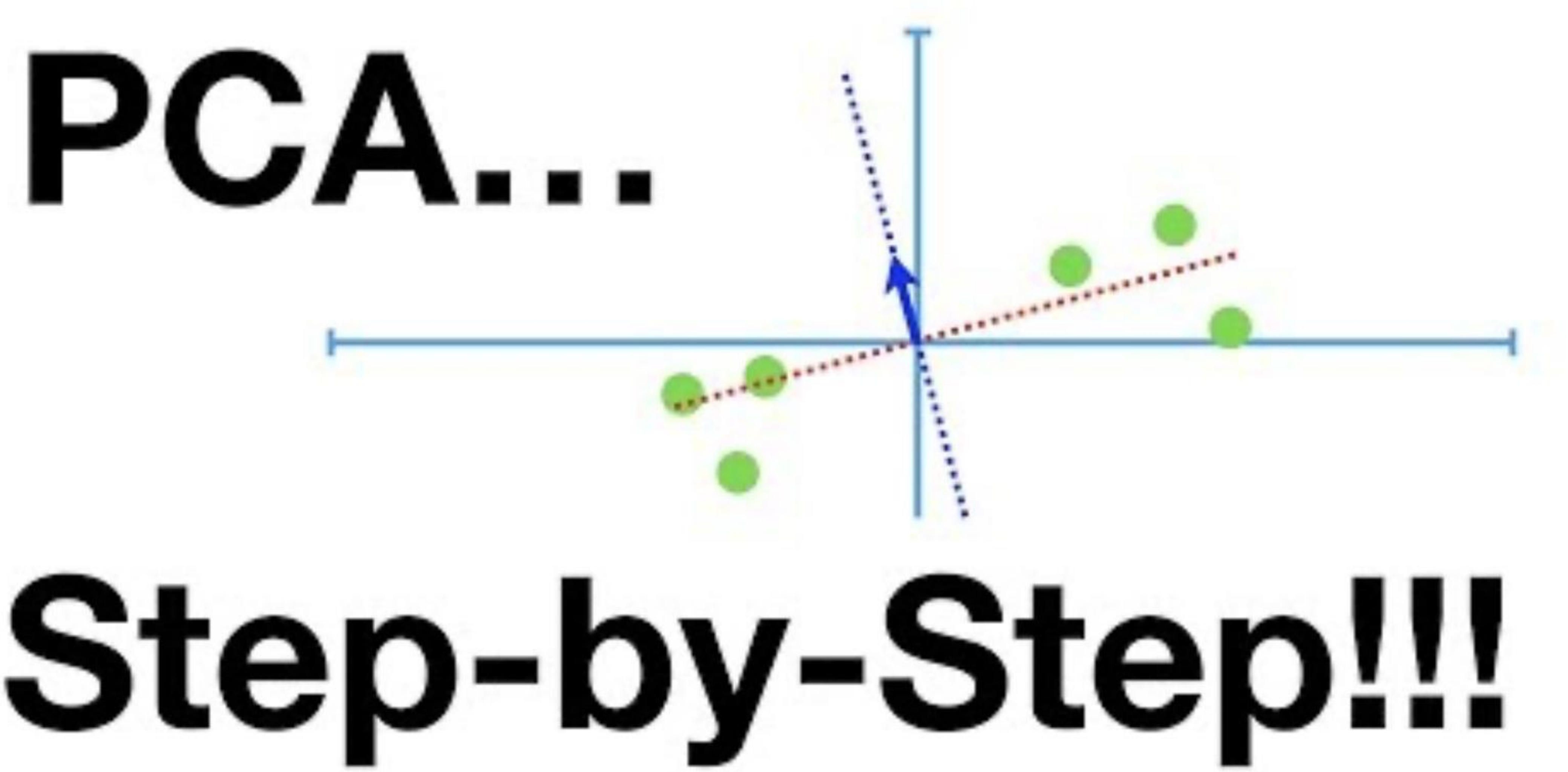
- PCs: axis where the variance of the projected data points is maximized

- Find the line with the property that the average squared distance of the points to that line is minimized

- Calculation: based on Singular Vector Decomposition (SVD)



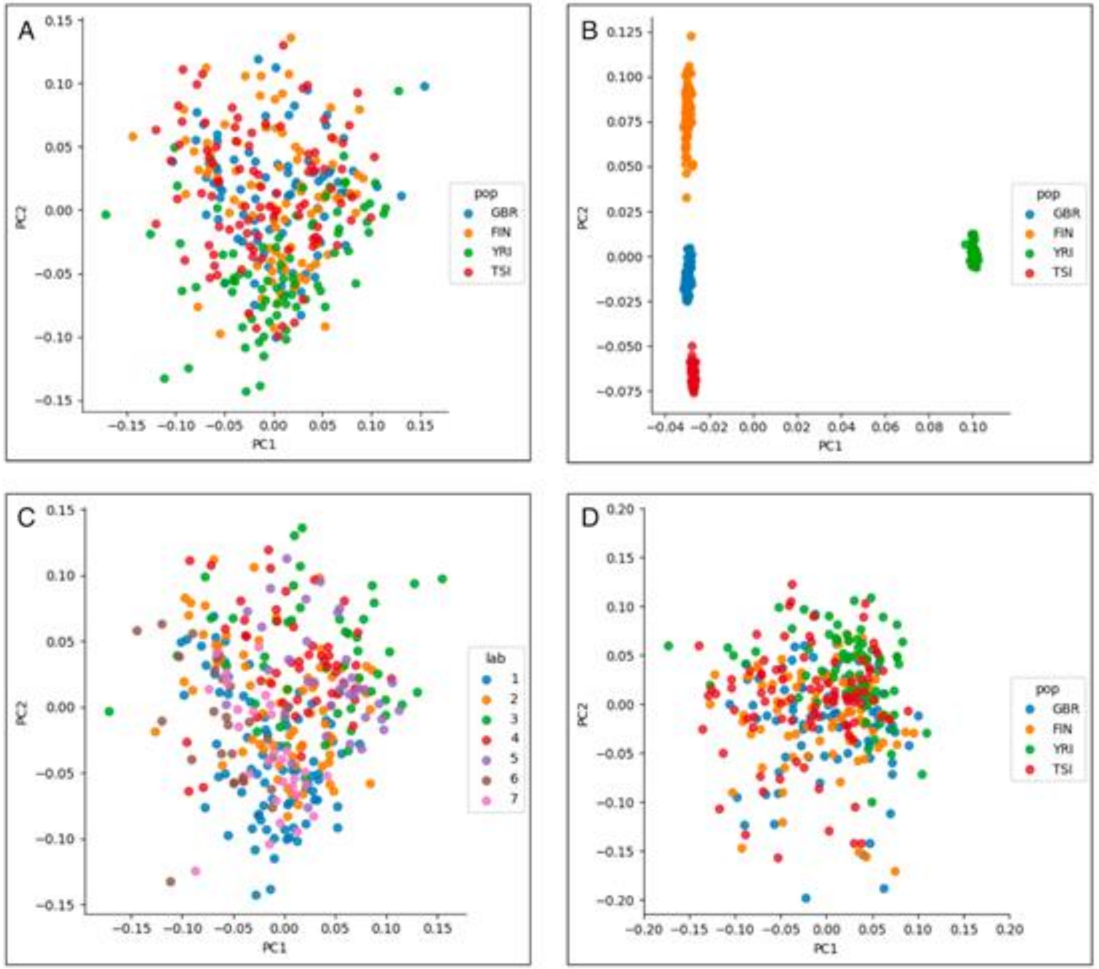
Principal Component Analysis (PCA)



https://www.youtube.com/watch?v=FgakZw6K1QQ

PCA finds projections that maximize retained **variance…**
…not necessarily retained **biology**

- An analysis of data from (Lappalainen et al., 2013), which included (bulk) gene expression measurements paired with genotypes for individuals from 5 populations (4 were studied) showed little clustering by population in the PCA for expression data (A) and separated cluster for genotypes (B).
- The structure in the gene expression data could be partially explained by batch effect (samples were assayed at different centers), a signal that dominated others (C).
- While removal of batch effect revealed more population structure, it was still not as strong as with genotype data.



Brown et al., 2018

The goal of dimensionality reduction

- The goal of dimension reduction depends on the application. It may be to:
    - remove “extraneous”, noisy, dimensions in the data.
    - produce smaller matrices that are smaller to store and computationally more tractable to analyze.
    - visualize “structure” in data. This can require methods that preserve various attributes of the data such as distances between points, or local neighborhoods of points.
- PCA is a *linear* dimension reduction method; it is a *projection* of the data to lower dimension. There are other linear dimension reduction methods, and there are non-linear dimension reduction methods.

Non-linear dimension reduction by t-SNE

- t-distributed stochastic neighbor embedding (t-SNE)
- Introduced in (van der Maaten and Hinton, 2008)
- A non-linear dimensionality reduction approach that attempts to map a distribution of pairwise distances among *n* high-dimensional samples from their high dimension to a distribution of pairwise distances of the *n* samples in a low dimension.

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \ln \left(\frac{p_{ij}}{q_{ij}} \right)$$

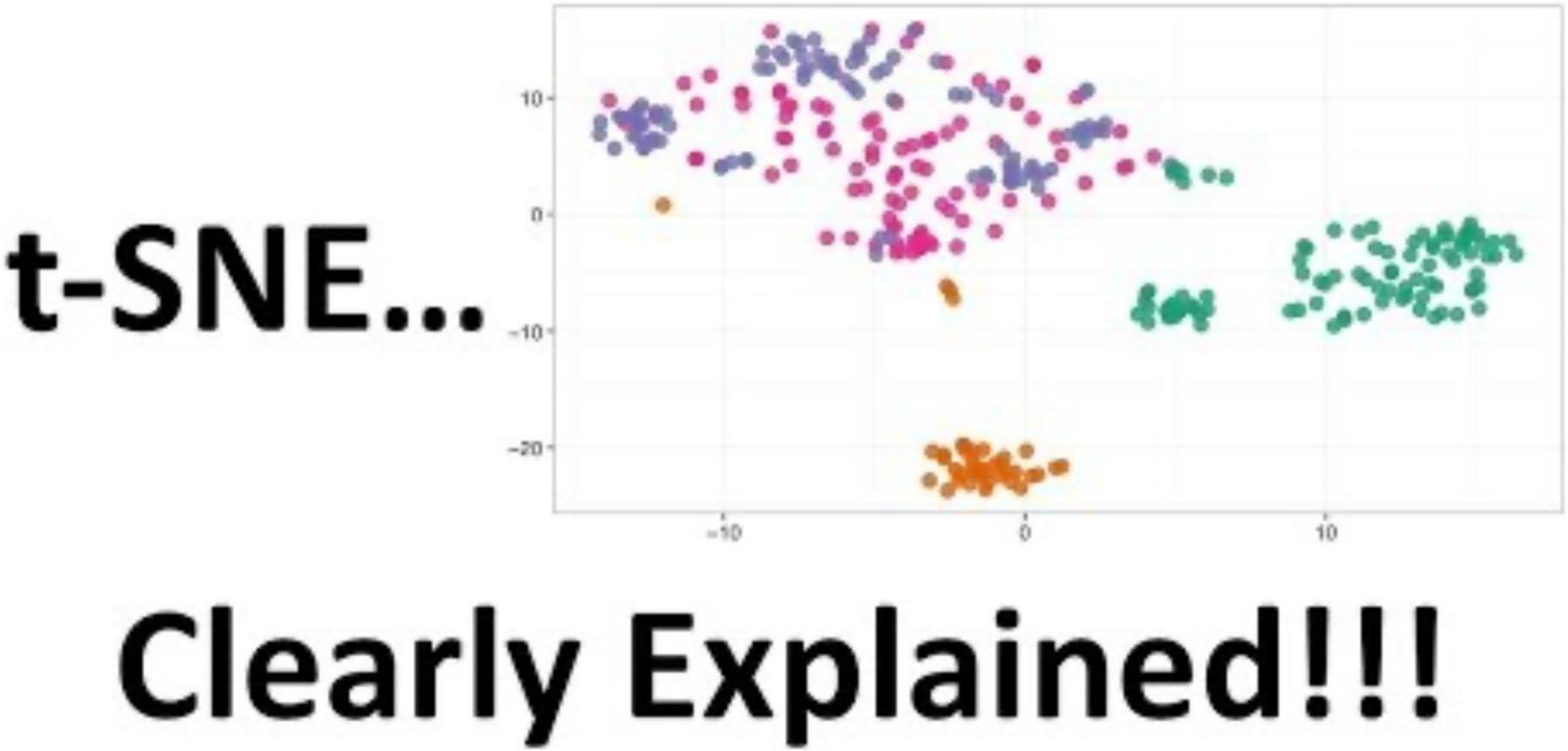
$$p_{i|j} = \frac{e^{\frac{-||x_i - x_j||^2}{2\sigma_i^2}}}{\sum_{k \neq i} e^{\frac{-||x_i - x_k||^2}{2\sigma_i^2}}}$$

- Minimizes the Kullback-Leibler divergence between a Gaussian distribution used to model distances in the ambient space, and a Student t-distribution modeling distances in low-dimension (2d or 3d).

$$q_{i|j} = \frac{(1+||y_i - y_j||)^{-1}}{\sum_{k \neq l}(1+||x_i - x_k||)^{-1}}$$

- Theorem: (Linderman and Steinerberger, 2019): There are parameters for this algorithm that ensure rapid convergence. The algorithm behaves like spectral clustering (under some assumptions).

t-SNE

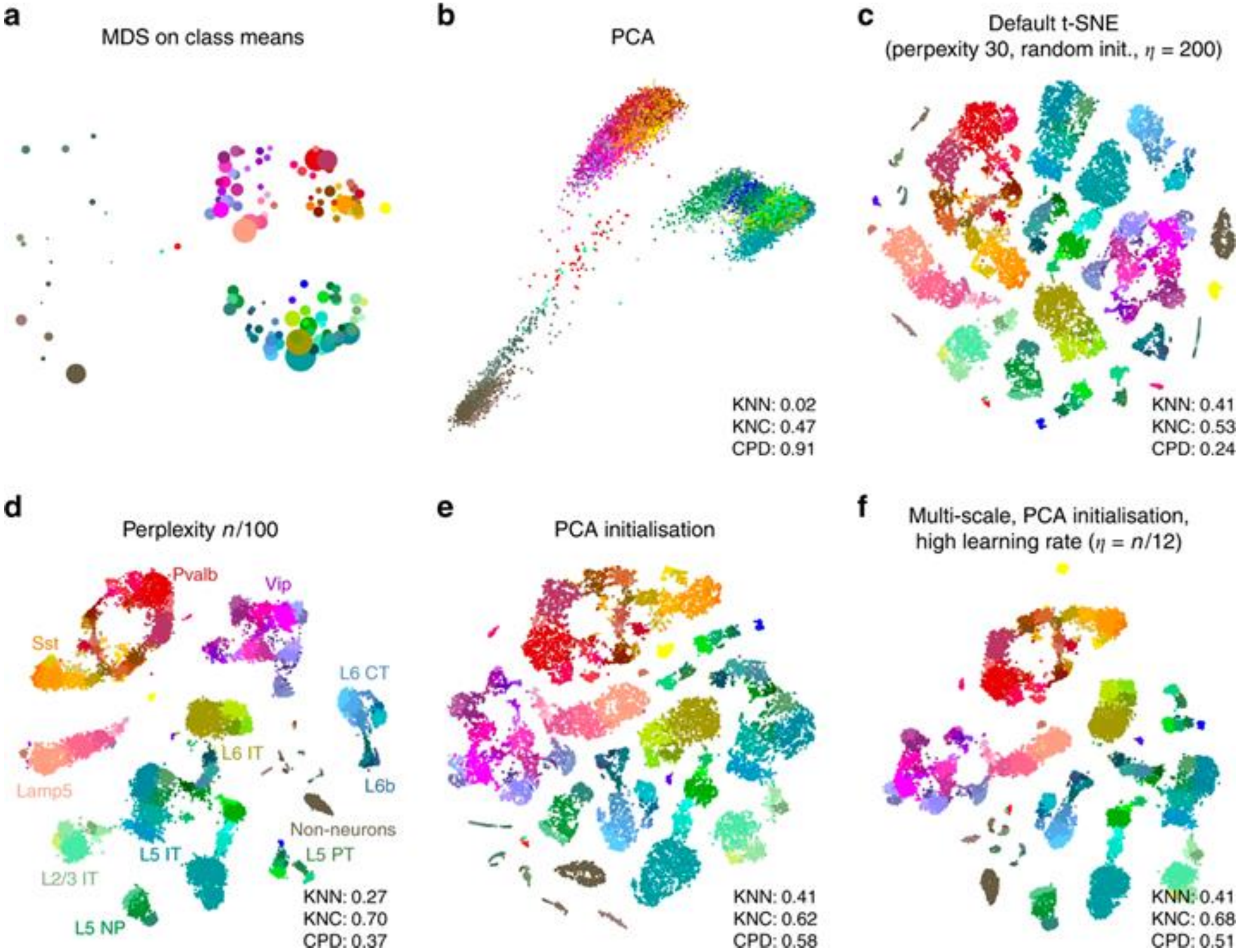


https://www.youtube.com/watch?v=NEaUSP4YerM

The art of using t-SNE (Kobak and Berens, 2019)

- t-SNE appears to produce “prettier” images if the data matrix is first reduced in dimension with PCA.

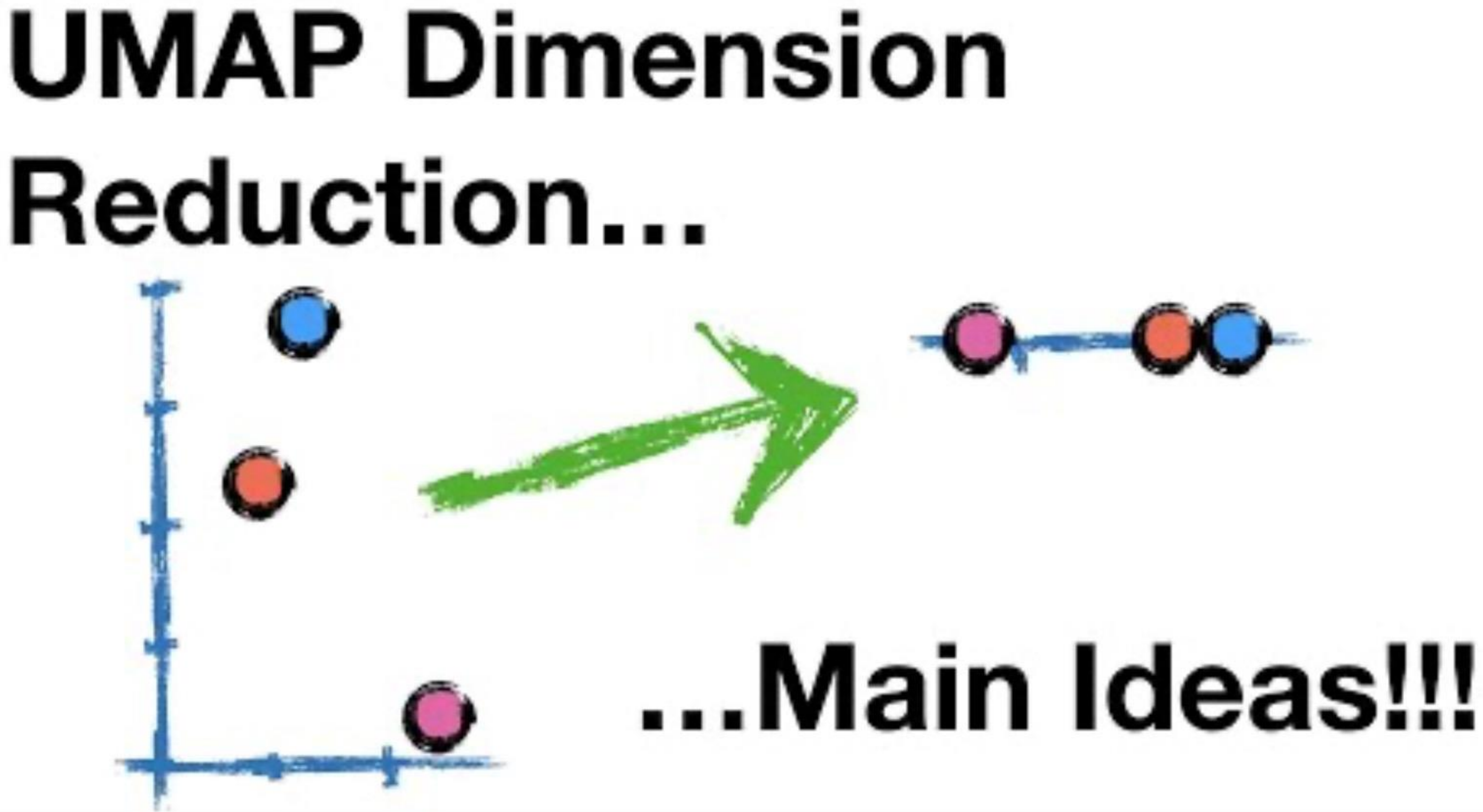
- Results are strongly dependent on parameters used (“one may worry that this gives a researcher too many knobs to tune”).



Non-linear dimension reduction by UMAP

- Uniform manifold approximation and projection (UMAP)
- Introduced in (McInnes et al., 2018).
- Based on intuition gleaned from geometry to construct a weighted graph on a point set that is then embedded in low (usually two) dimension: “It turns out that we can actually formalize all of this by stealing the singular set and geometric realization functors from algebraic topology and then adapting them to apply to metric spaces and fuzzy simplicial sets.” - from the UMAP documentation.
- Widely used in single-cell RNA-seq analysis.

UMAP



https://www.youtube.com/watch?v=eN0wFzBA4Sc

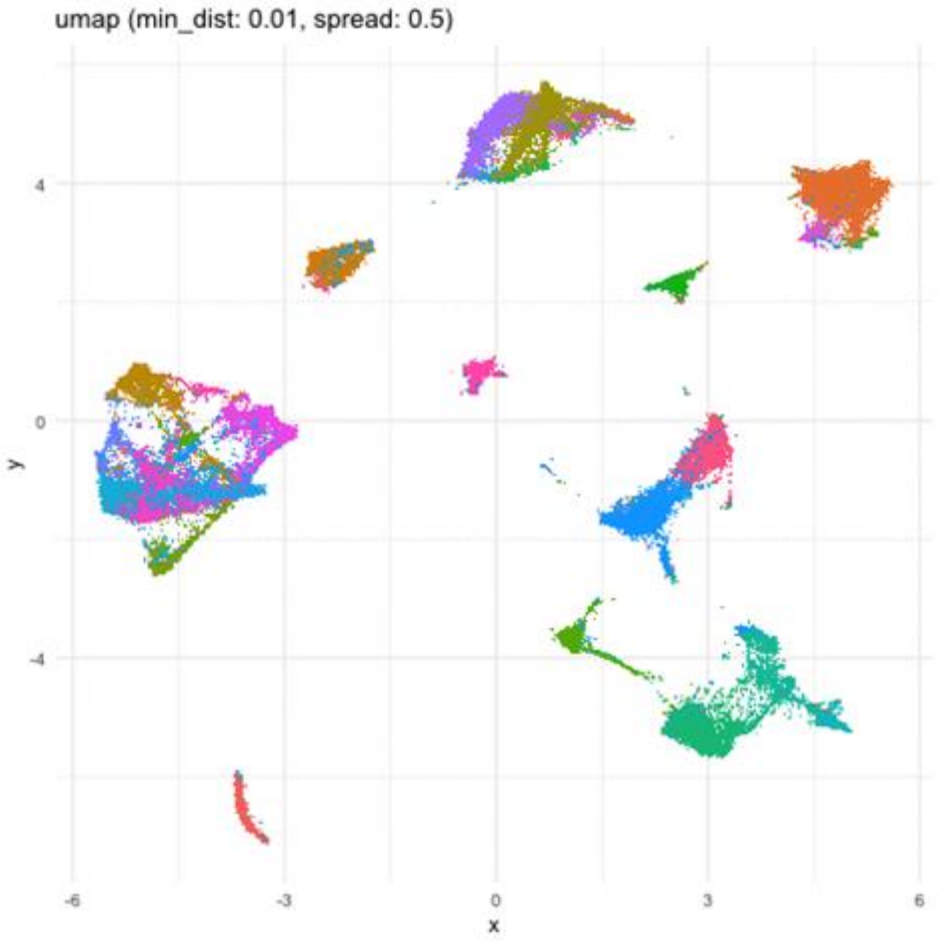
UMAP results also depend on numerous parameters

- The example on the right is from (Fan, 2022).

- UMAP is used to “validate” clustering of cells. Cells are painted by cluster and separation in UMAP is taken to be confirmation that the clustering was effective and biologically meaningful.

- Parameters are frequently “optimized” to produce images that are confirmatory. Confirmation bias is (hopefully) avoided by following up with experiments to confirm hypothesis based on the visualizations.

- UMAP embeddings are also used as the basis for further quantitative analyses.

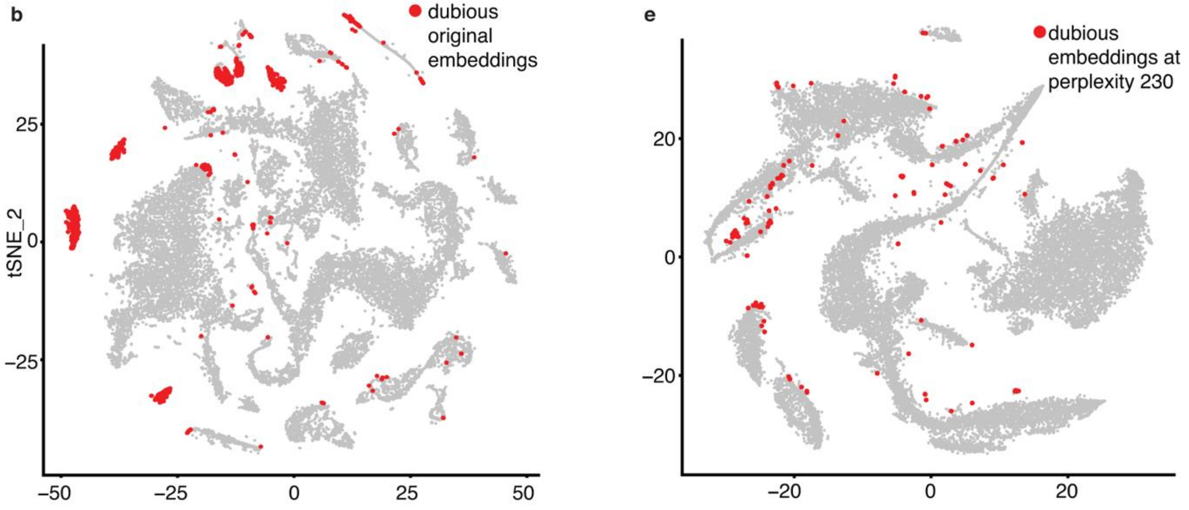


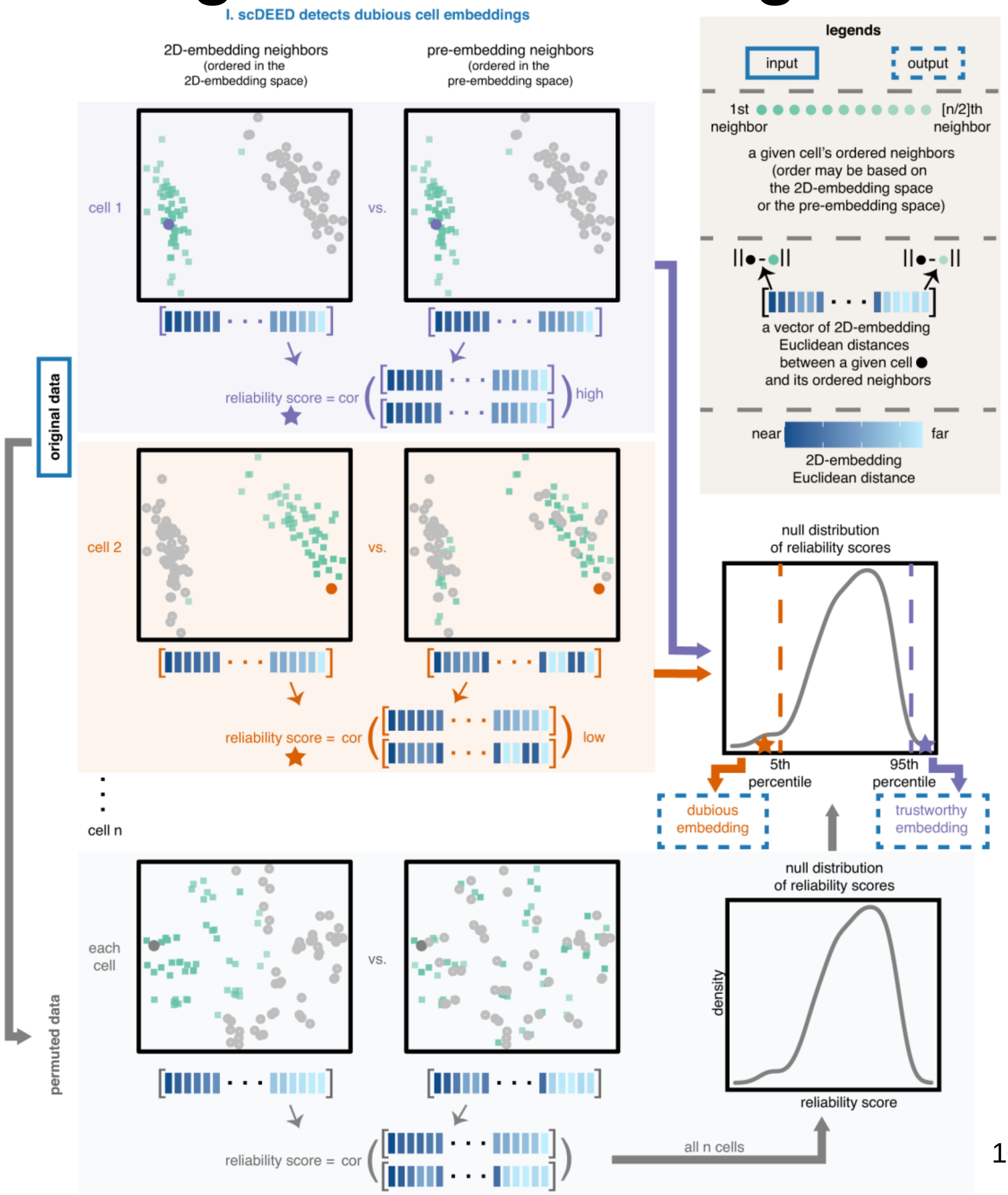
scDEED: a statistical method for detecting dubious 2D single-cell embeddings (Xia et al. 2024)

- Calculate the correlation of neighbor distance between original space and reduced dimension space.

- Identify dubious cell embeddings

- Guide parameter settings





I. scDEED detects dubious cell embeddings
2D-embedding neighbors (ordered in the 2D-embedding space)
pre-embedding neighbors (ordered in the pre-embedding space)
legends
input
output
1st neighbor
[n/2]th neighbor
a given cell's ordered neighbors (order may be based on the 2D-embedding space or the pre-embedding space)
a vector of 2D-embedding Euclidean distances between a given cell ● and its ordered neighbors
near
far
2D-embedding Euclidean distance
original data
cell 1
vs.
reliability score = cor
high
cell 2
vs.
reliability score = cor
low
cell n
null distribution of reliability scores
5th percentile
95th percentile
dubious embedding
trustworthy embedding
permuted data
each cell
vs.
reliability score = cor
all n cells
null distribution of reliability scores
density
reliability score

Autoencoder

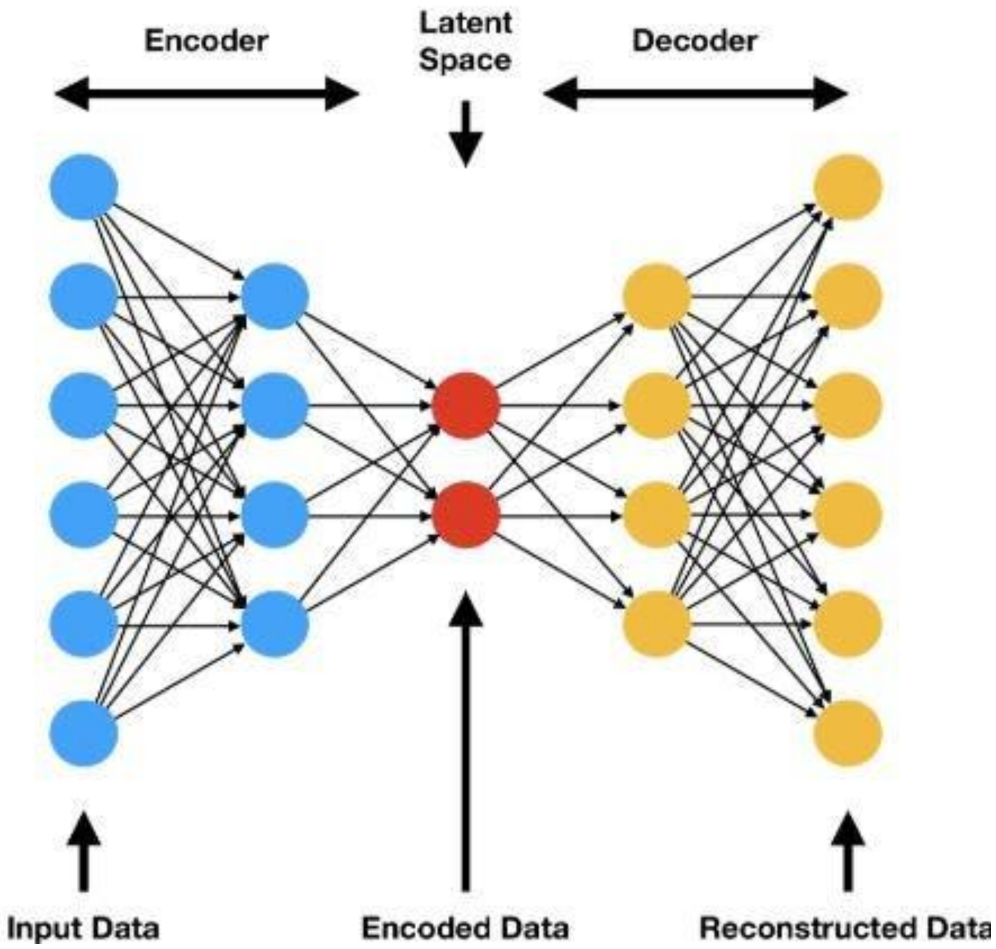
- An artificial neural network for dimensionality reduction

- Encoder & Decoder

- Lower-dimensional embedding in Latent Space

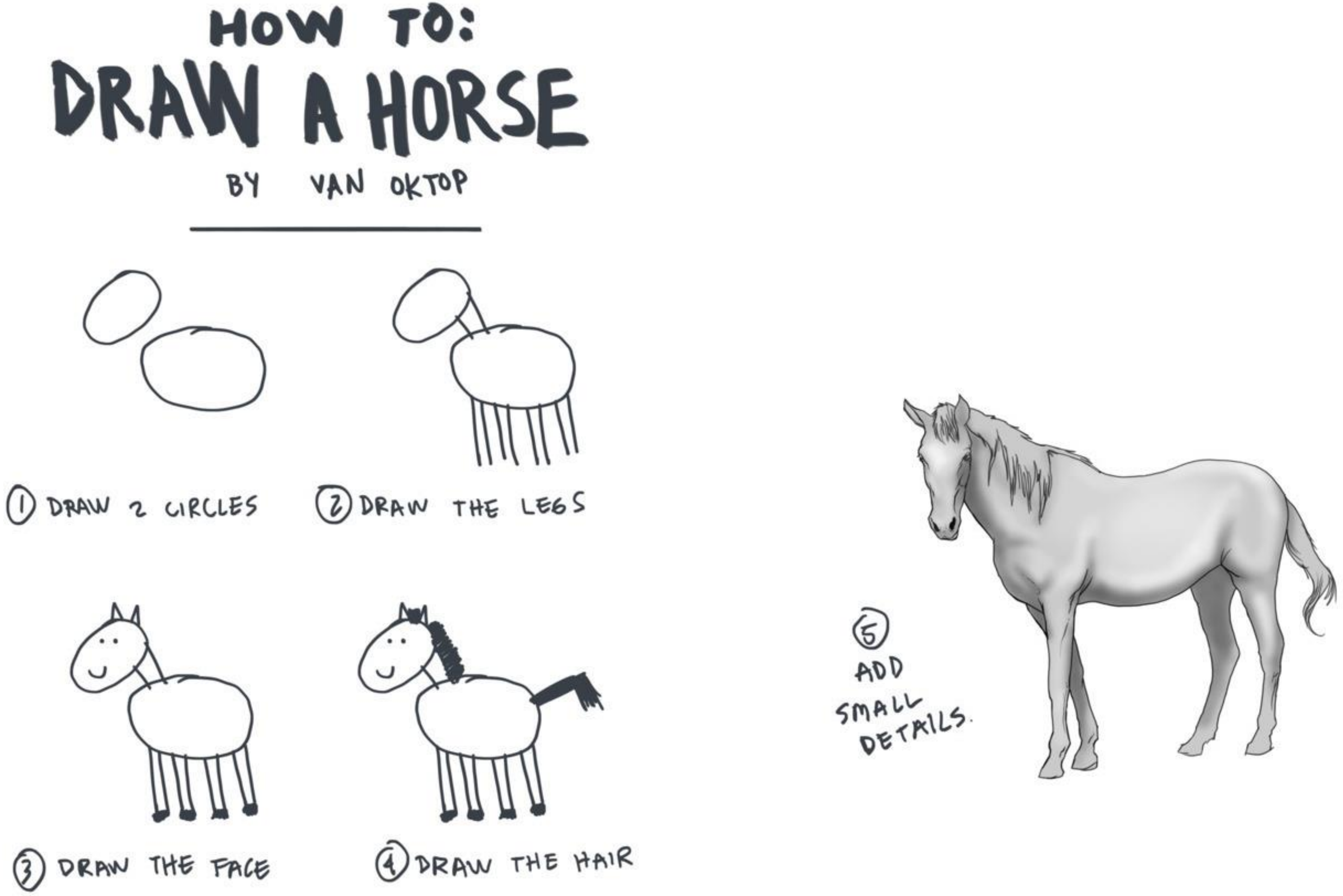
- Non-linear

- Model interpretability



Summary

- Curse of dimensionality
- Linear dimensionality reduction: PCA
- Non-linear dimensionality reduction: t-SNE, UMAP, autoencoder
- t-SNE and UMAP are good for data visualization, but not necessarily appropriate for scientific discovery.


HOW TO:
DRAW A HORSE
BY VAN OKTOP
1 DRAW 2 CIRCLES
2 DRAW THE LEGS
3 DRAW THE FACE
4 DRAW THE HAIR
5 ADD SMALL DETAILS.