Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: soumyad@cse.iitk.ac.in

Acknowledgements

- Some of the following slides are adapted from the excellent course materials and tutorials made available by:
  - Prof. Klaus Mueller (State University of New York at Stony Brook)
  - Prof. Tamara Munzner (University of British Columbia)
  - Zaur Fataliyev (Research Scientist – Meta)

Study Materials for Lecture 12

- Visualizing High-Dimensional Data: Advances in the Past Decade; S. Liu et al., TVCG2016
- t-SNE: https://distill.pub/2016/misread-tsne/
- UMAP: https://pair-code.github.io/understanding-umap/
- Footnotes in slides

Assignment 1: Due Feb 16, 2024

- Assignment 1 is due on Feb 16th 11:59pm
  - Submit via HelloIITK
  - Follow submission instructions otherwise points will be deducted
  - If you miss deadline but still want partial credit, follow the late submission policy
    - 2 late days with penalty

Mid Sem Exam

- Saturday 22nd Feb 8-10am
- Location: TBD
- **Please bring your institute id card (mandatory)**
- No classes during Mid Sem week

Mid Sem Exam

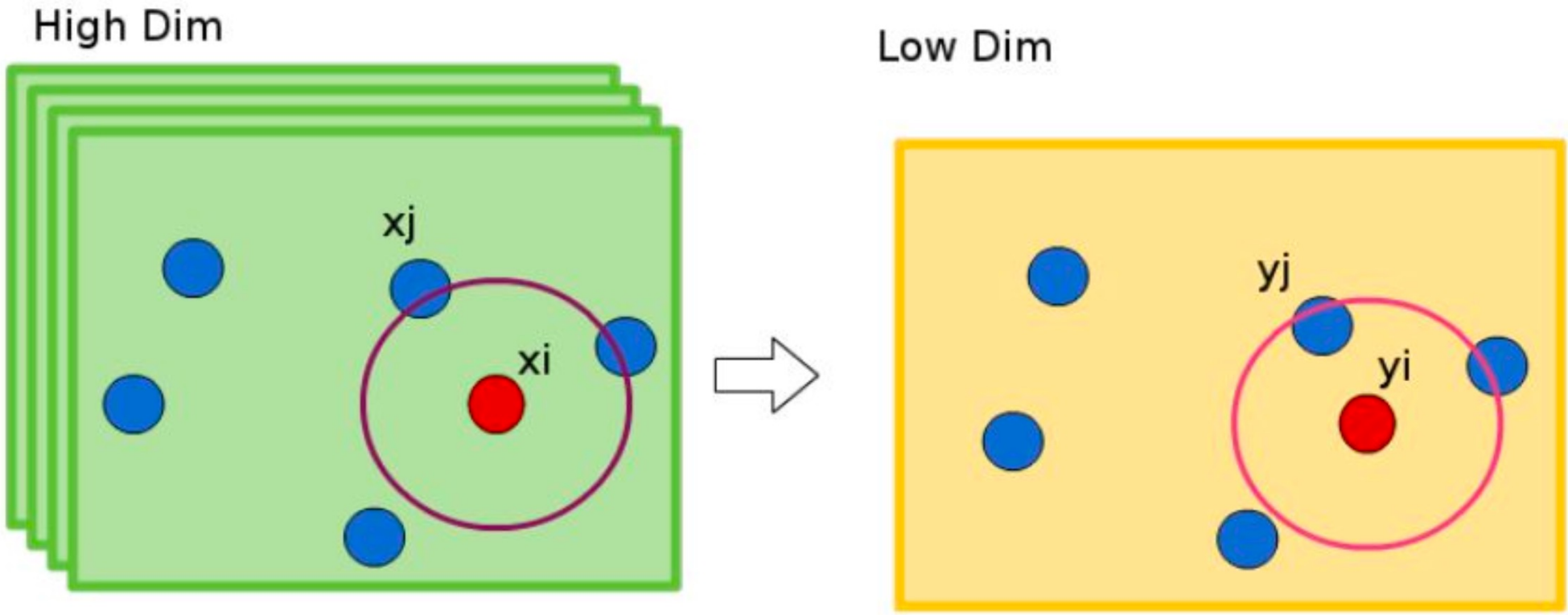
- All topics covered until 19th Feb are part of the mid semester examination
- All the libraries we have used/seen so far can be excluded, no coding questions
- Test will be primarily on your understanding of the concepts of the topics discussed in the class

t-Distributed Stochastic Neighbor Embedding (t-SNE)

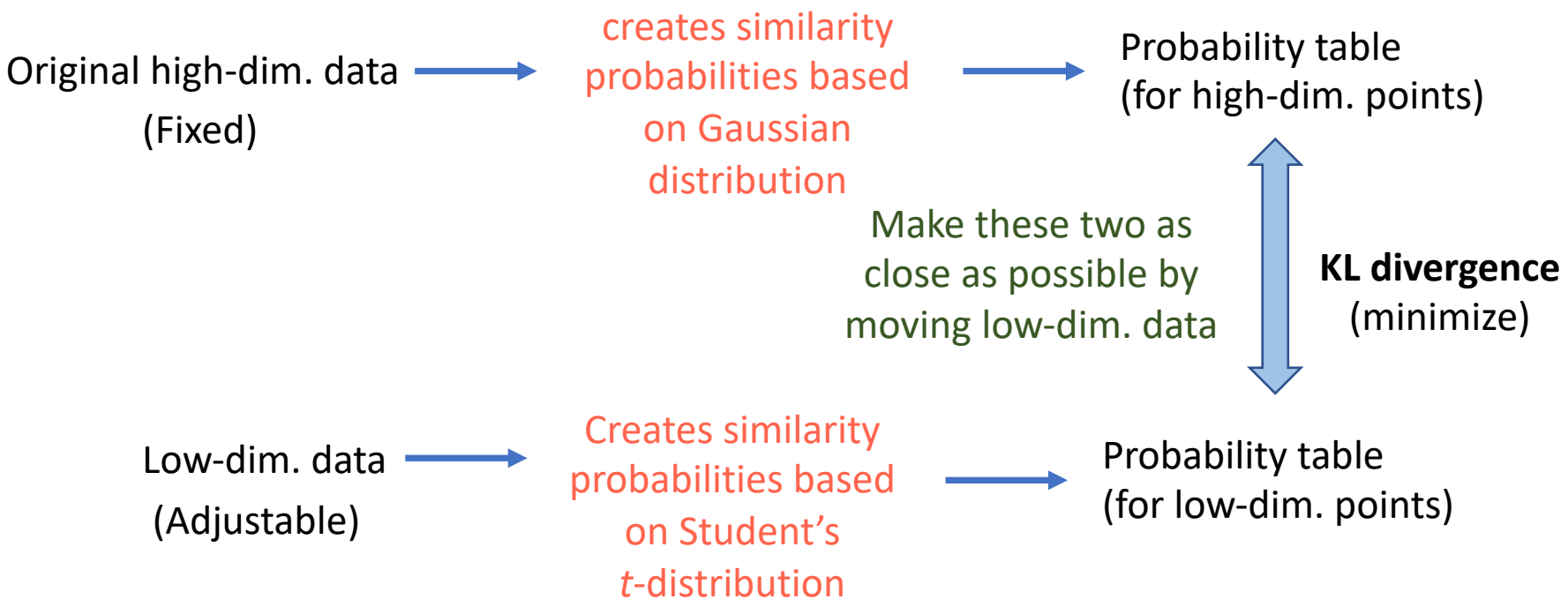
- PCA is not always effective in finding patterns in low dimensional visualization space
  - It is a linear algorithm, meaning that it cannot represent complex nonlinear relationship between features
- t-Distributed Stochastic Neighbor Embedding: A nonlinear dimensionality reduction technique
- t-SNE is specifically designed for high dimensional data visualization purposes
  - **Paper:** https://jmlr.org/papers/v9/vandermaaten08a.html
  - ~51,000 citations!!

t-SNE : Underlying Idea

Measure pairwise distances between high dimensional and low dimensional points

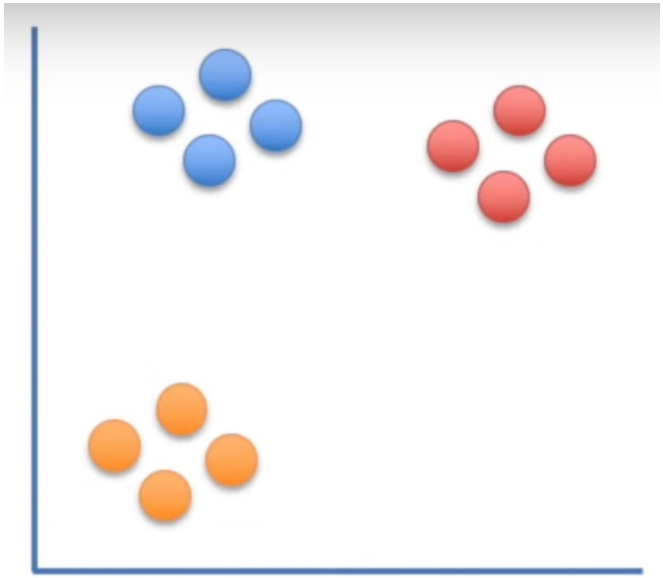


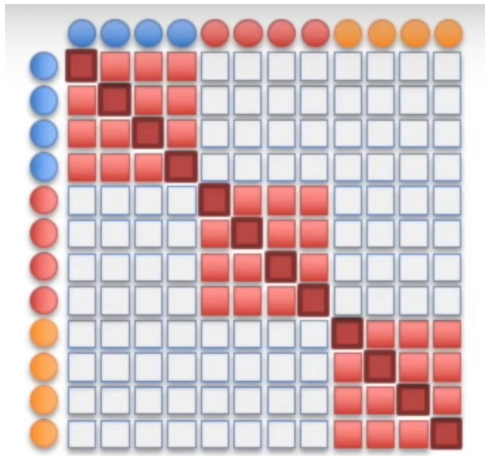
t-SNE : Underlying Idea



t-SNE : Underlying Idea

High Dim data

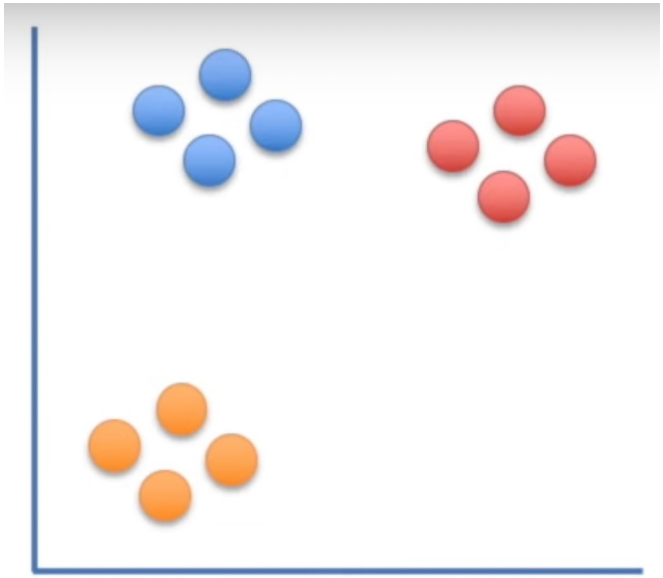


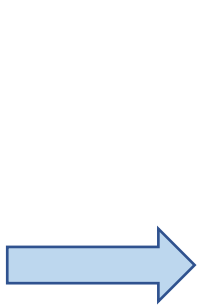


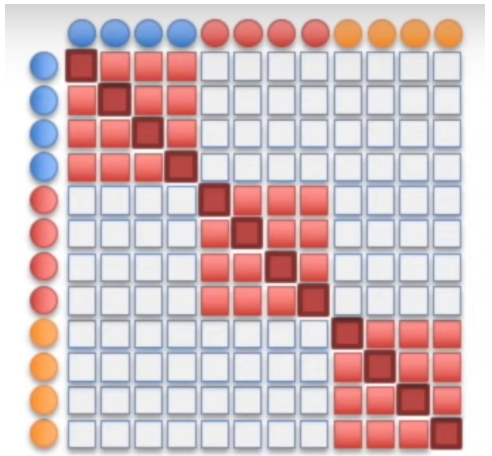
Similarity Probability table in high dimension

t-SNE : Underlying Idea

High Dim data

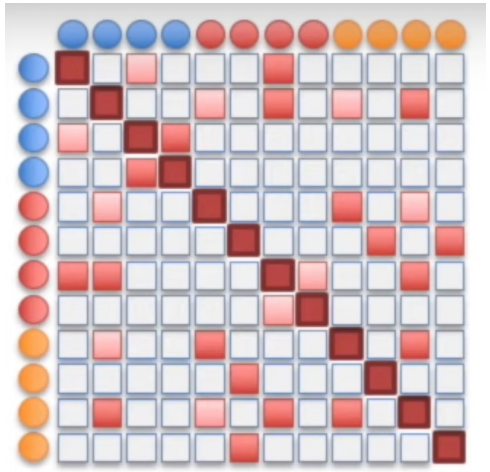






Similarity Probability table in high dimension

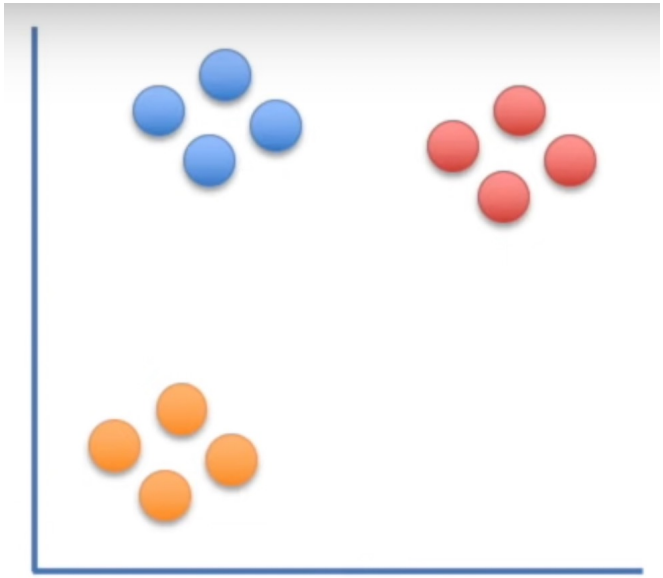
Low Dim data

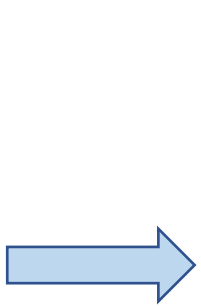


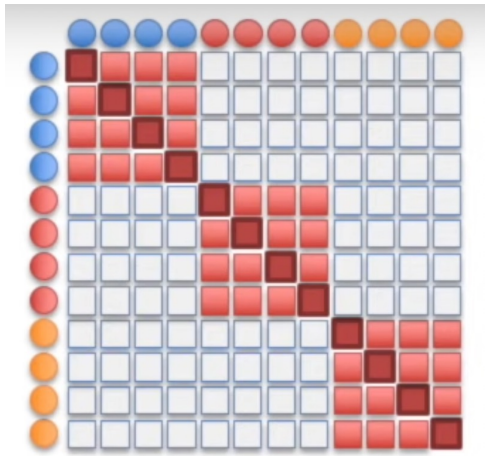
Initial Similarity Probability table in low dimension

t-SNE : Underlying Idea

High Dim data

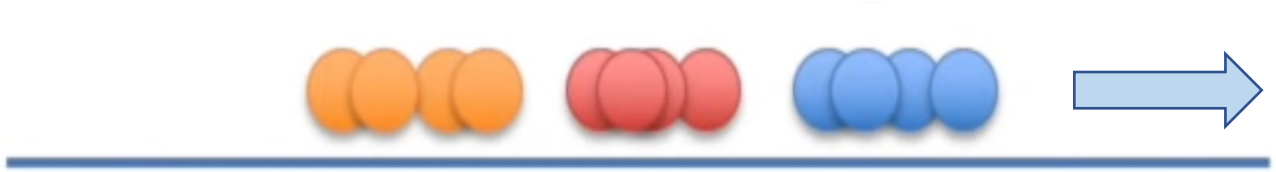


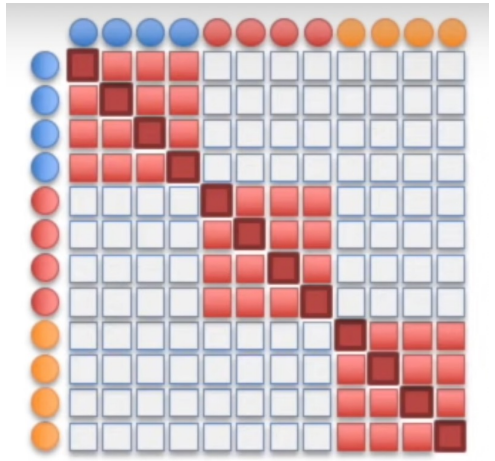




Similarity Probability table in high dimension

Low Dim data





Final Similarity Probability table in low dimension

t-SNE: Measure Distances and Optimize

- Similarity of datapoints in high dimensional space

$$p_{ij} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq l} exp(-||x_l - x_k||^2/2\sigma^2)}$$

- Similarity of datapoints in low dimensional space

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

- Cost function: Minimize KL Divergence

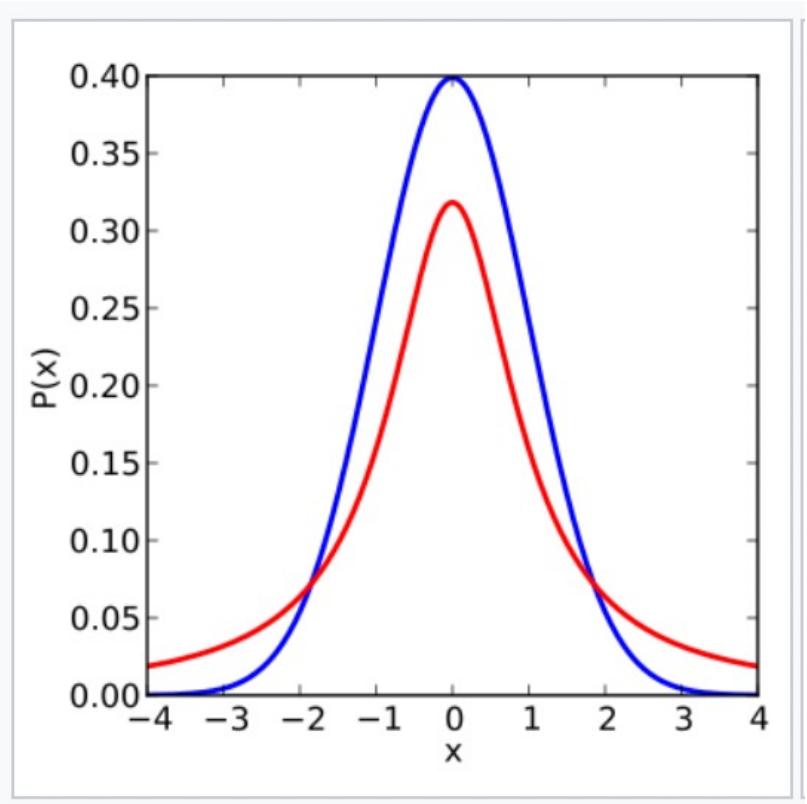
$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

KL Divergence: Kullback–Leibler divergence is a measure of how one probability distribution P is different from a second, reference probability distribution Q

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

What is the t in the t-SNE ?

- To measure distance between points in the low dimensional space, a student's **t**-distribution is used instead of a Gaussian distribution

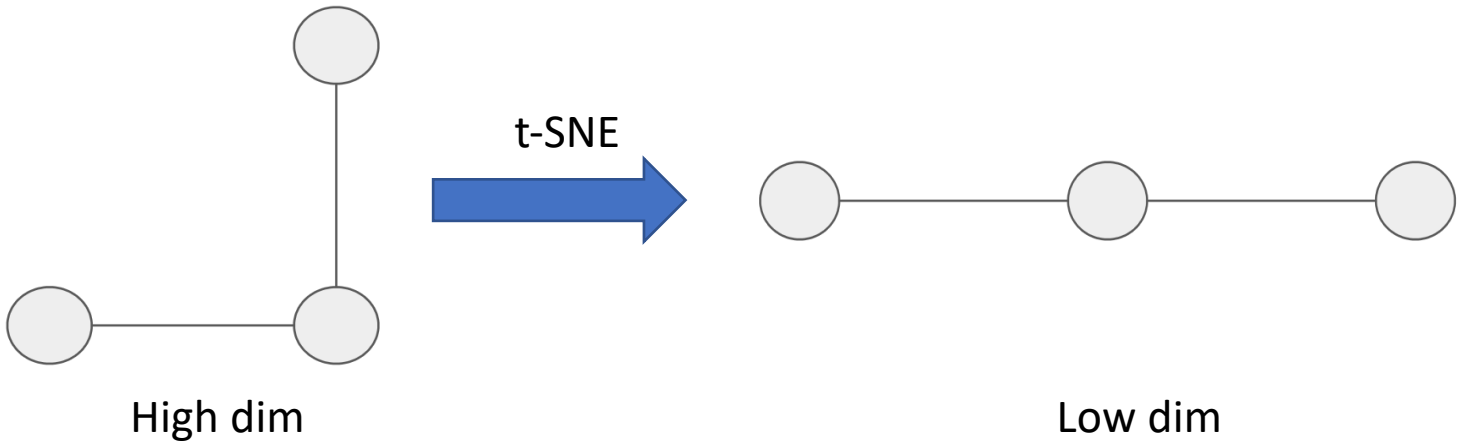


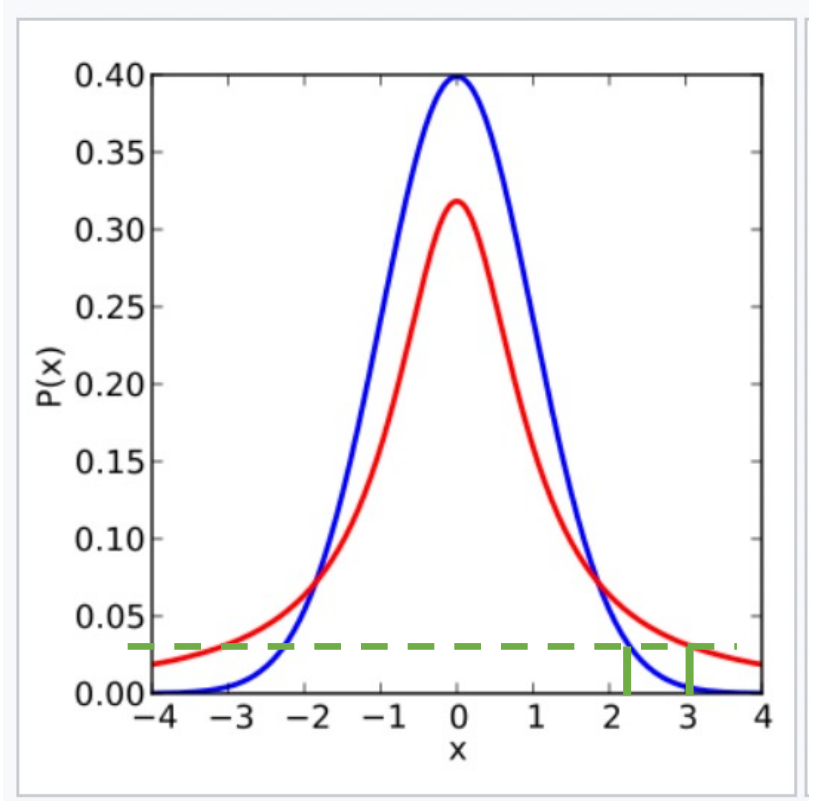
Blue: Gaussian distribution
Red: t-Distribution

- The student's t-distribution has a heavier tail compared to the Gaussian distribution

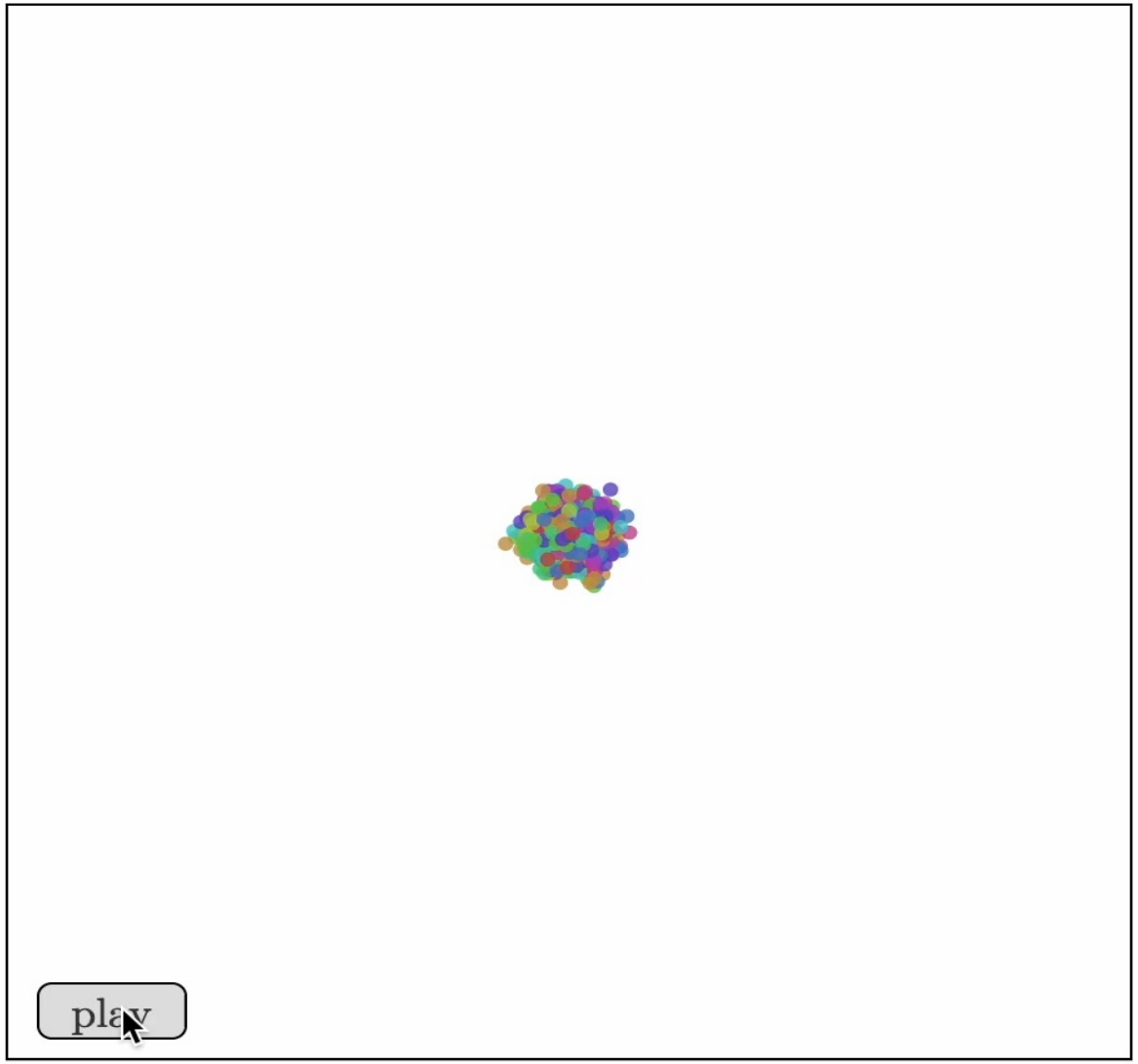
Why t-distribution? Crowding Problem

- Goal is to preserve local structure in low dimensional embedding
- Points which are far apart in high dimensional space should be far apart after projection
- The heavy tailed t-distribution helps to achieve it



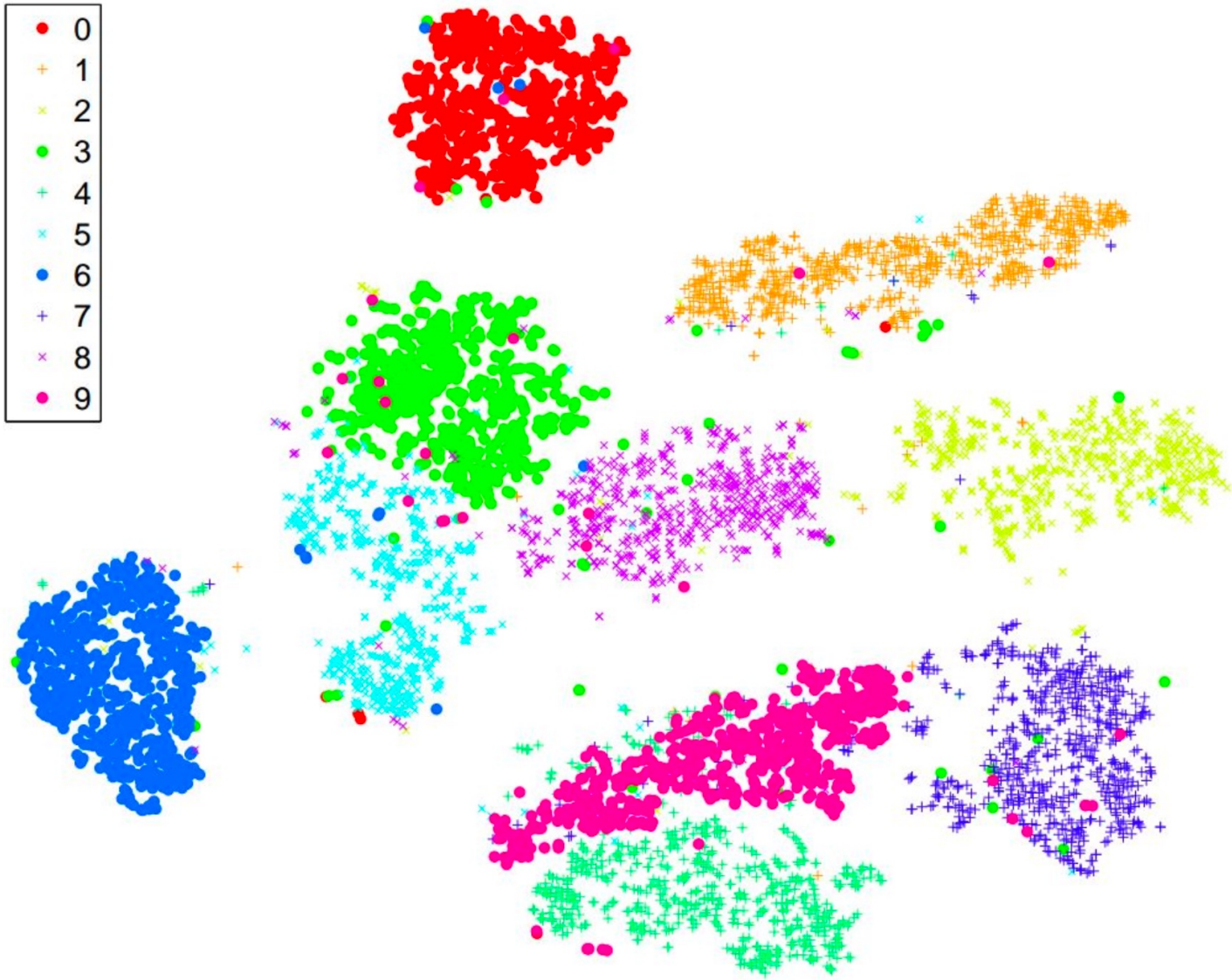


Execution of t-SNE on MNIST Data



Visualizing MNIST with t-SNE

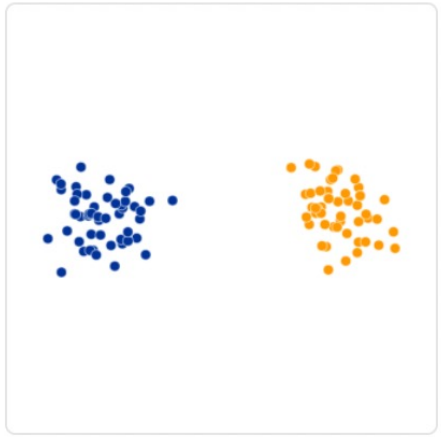
t-SNE on MNIST Data



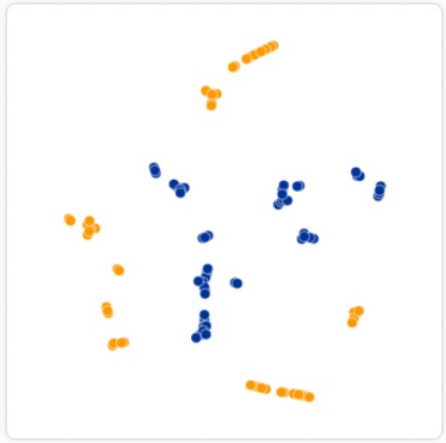
How to use t-SNE Effectively

- A key hyperparameter is perplexity
- It is a parameter that determines how to balance attention between local and global structures
  - Intuition: A guess about the number of close neighbors each point has
  - Changing this parameter has significant impact on final layout

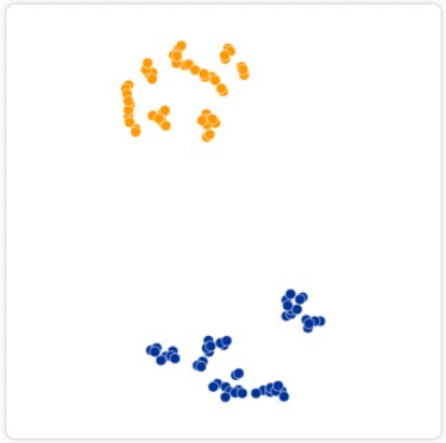
Impact of Perplexity in t-SNE



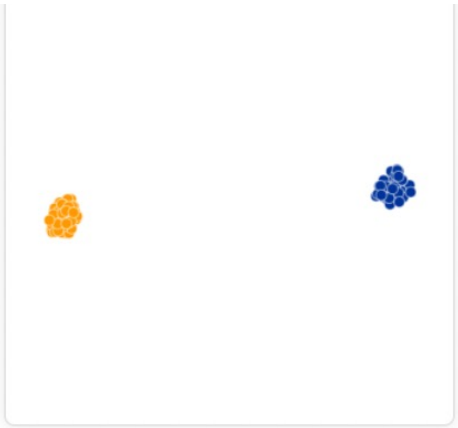
Original



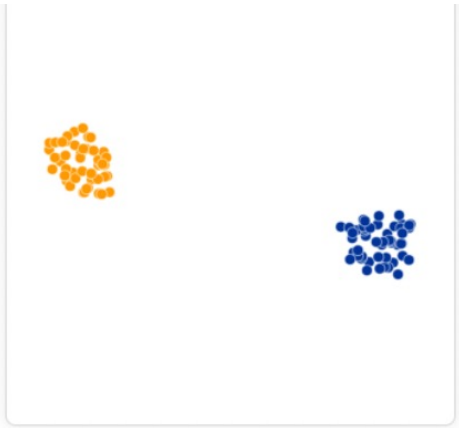
Perplexity: 2
Step: 5,000



Perplexity: 5
Step: 5,000



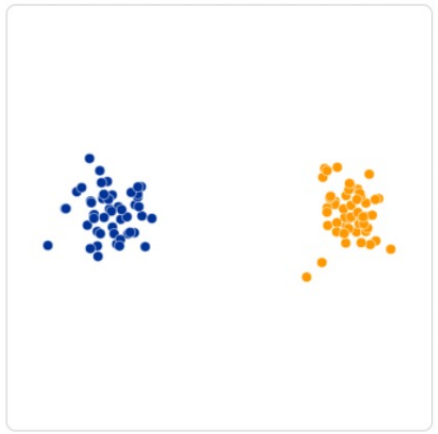
Perplexity: 30
Step: 5,000



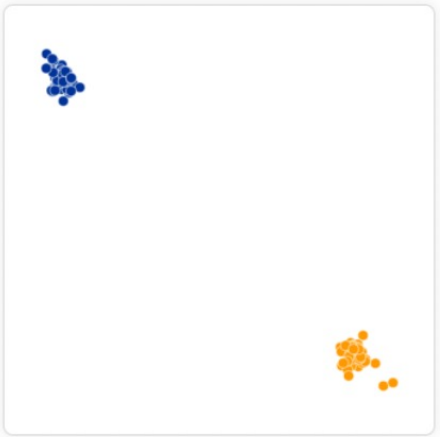
Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

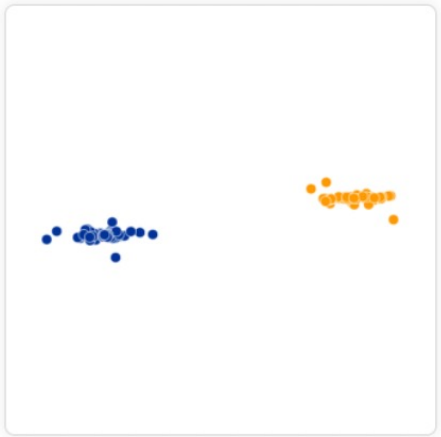
Impact of Epsilon (#iterations) in t-SNE



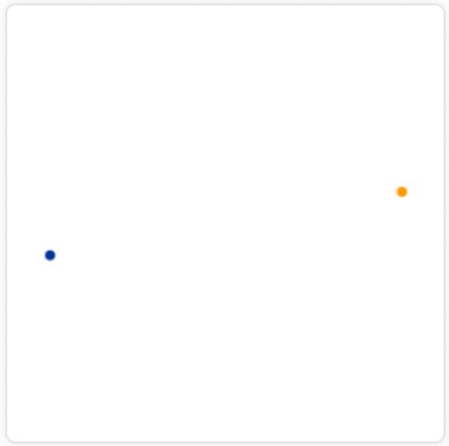
Original



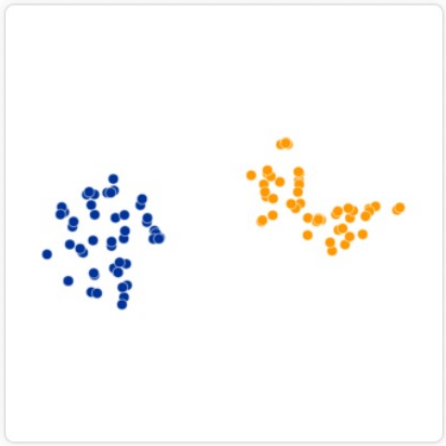
Perplexity: 30
Step: 10



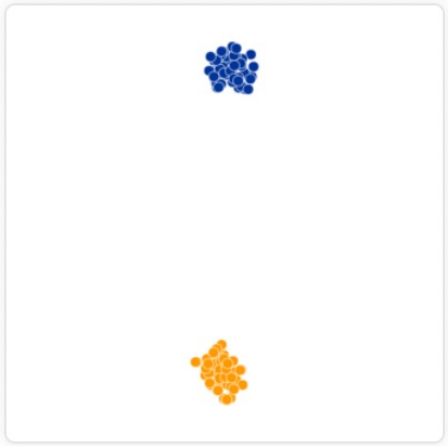
Perplexity: 30
Step: 20



Perplexity: 30
Step: 60



Perplexity: 30
Step: 120



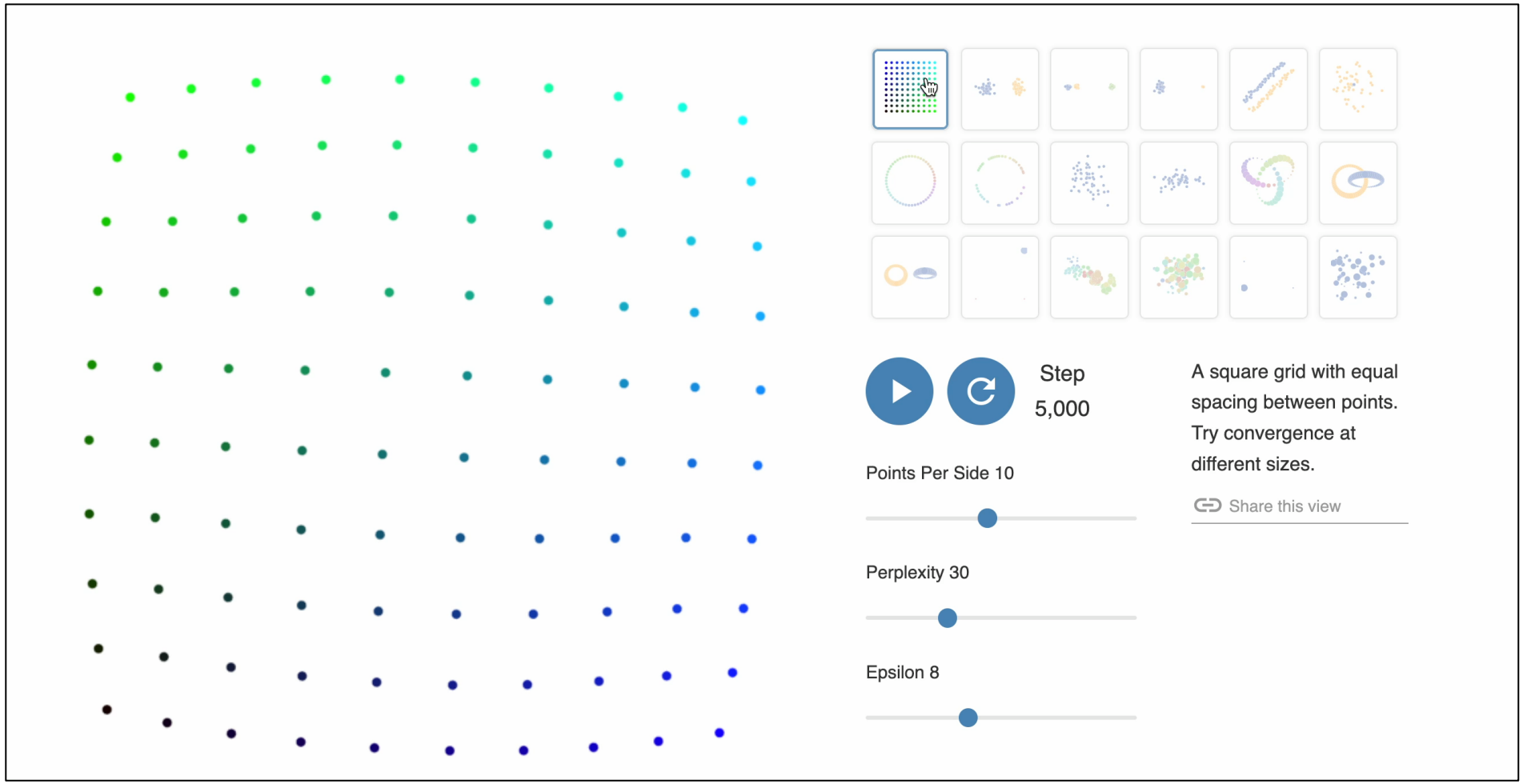
Perplexity: 30
Step: 1,000

Other Key Aspects of t-SNE

- Cluster sizes in a t-SNE plot may mean nothing
- Distances between clusters might not mean anything
- Random noise doesn’t always look random
- For topology, you may need more than one plot at different perplexity values
- Try with different number of iterations to ensure the algorithm has converged

Excellent Online Resource for Learning t-SNE

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

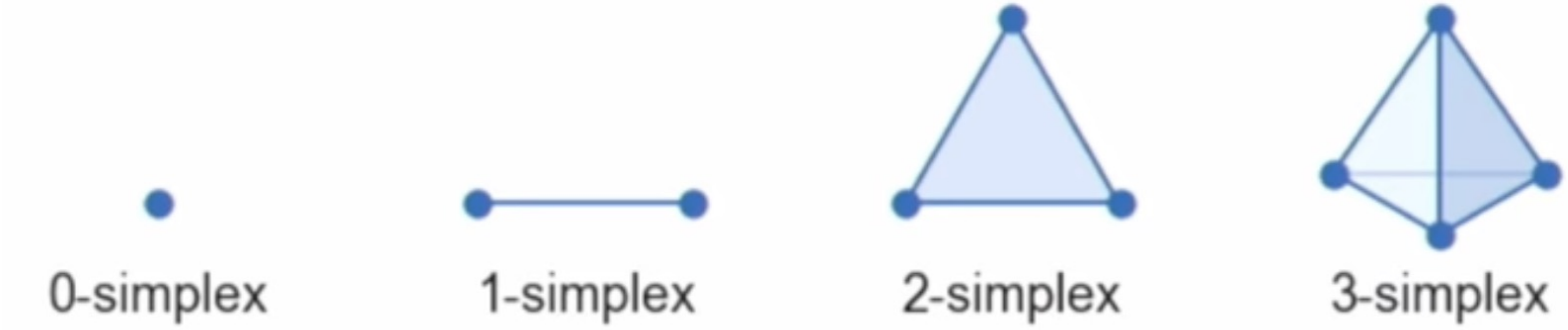


Uniform Manifold Approximation and Projection (UMAP)

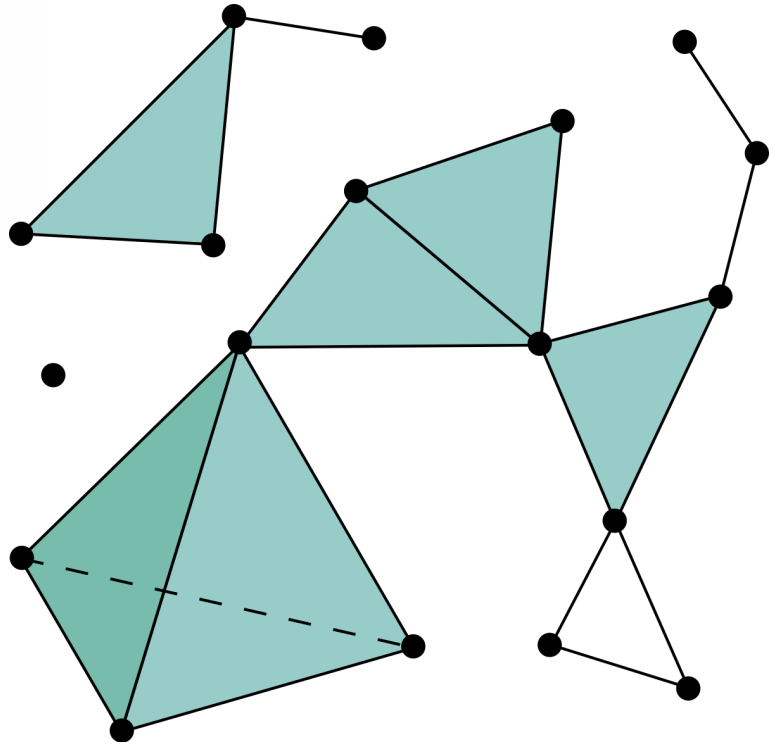
- A dimensionality reduction technique that assumes the available data samples are evenly (**uniformly**) distributed across a topological space (**manifold**), which can be **approximated** from these finite data samples and mapped (**projected**) to a lower-dimensional space.
- Key steps:
  - Learning the manifold structure in the high-dimensional space
  - Finding a low-dimensional representation of the high dim. manifold
- Given the high dimensional graph structure, UMAP projects the data into lower dimensions via a **force-directed graph layout algorithm**
- **Paper:** https://arxiv.org/abs/1802.03426
  - ~15,000 citations!!

Simplex and Simplicial Complex

- A simplex is a generalization of the notion of a triangle or tetrahedron to arbitrary dimensions



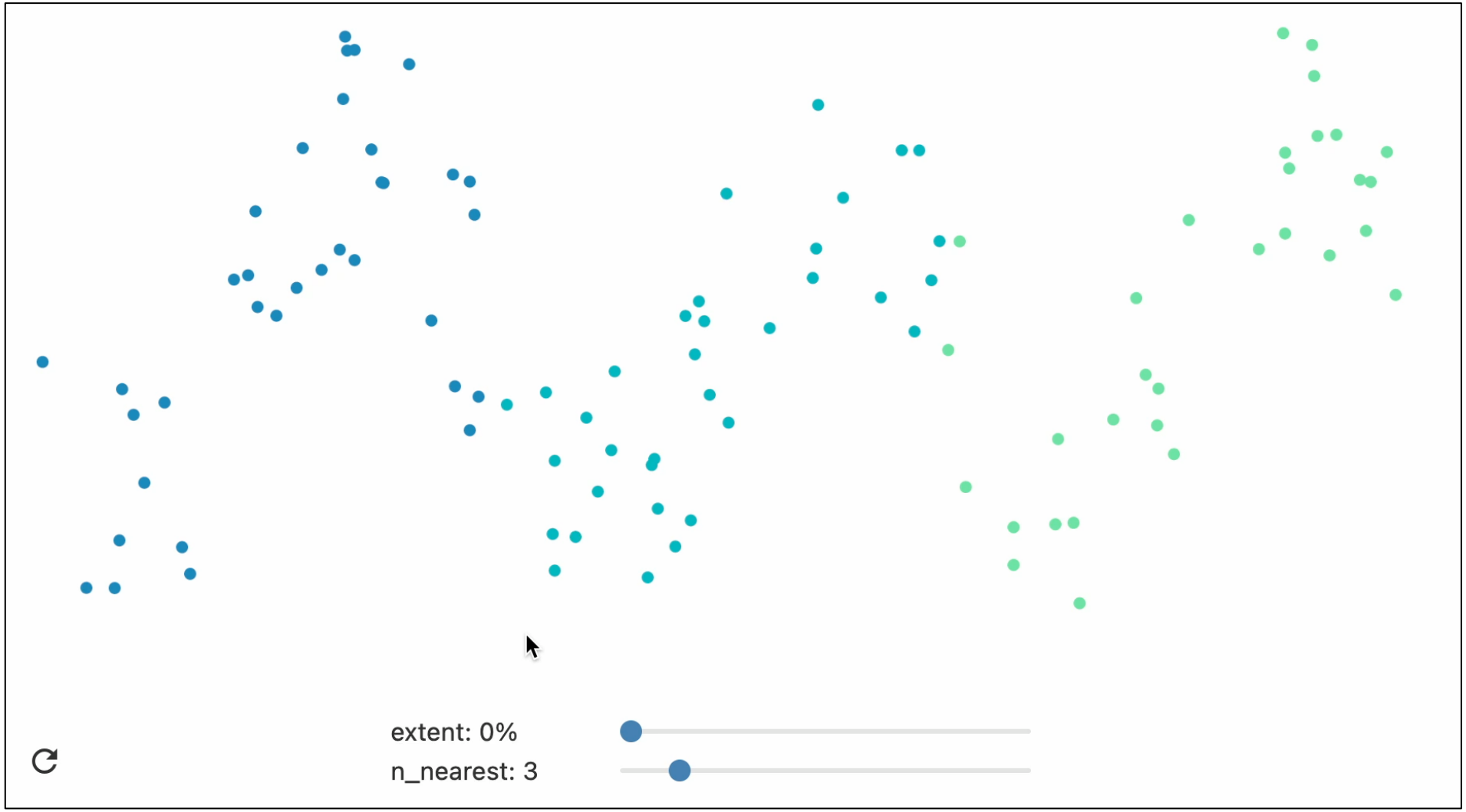
- A simplicial complex is a set composed of points, line segments, triangles, and their n-dimensional counterparts



UMAP

- Conceptually very similar to t-SNE
  - Construct a high dimensional graph representation of the data
  - Optimizes a low-dimensional graph to be as structurally similar as possible
- Idea behind constructing the high dimensional graph
  - Build a ‘fuzzy simplicial complex’
    - A weighted graph where edge weights represent likelihood that two points are connected
  - Connectedness: Grow a radius outward from each point and when it overlaps with a neighbor, connect the points
  - Then make the graph "fuzzy" by decreasing the likelihood of connection as the radius grows outward
  - Each point must be connected to at least its closest neighbor to capture local structure

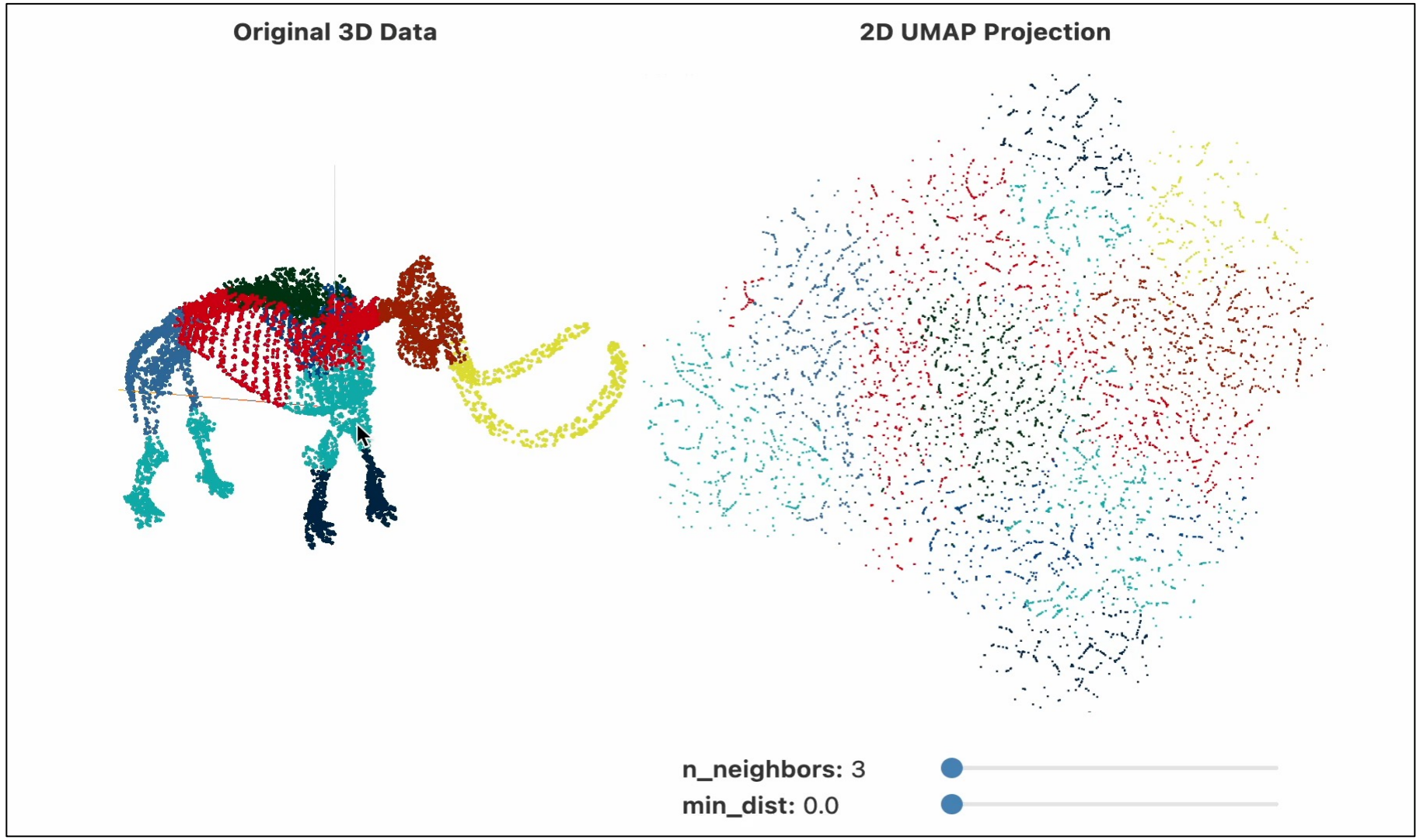
UMAP: High Dimensional Graph Construction



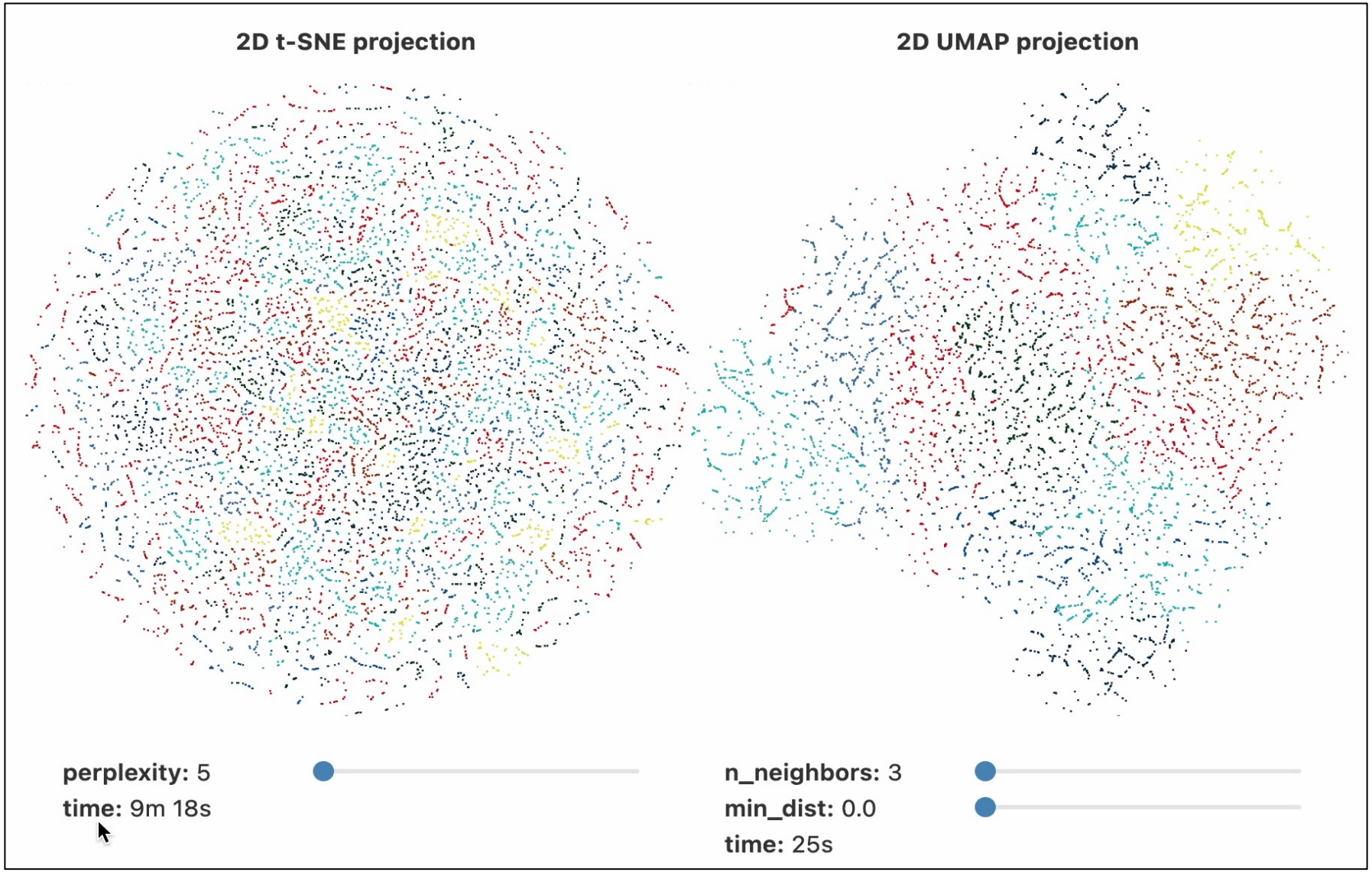
Two Important Parameters of UMAP

- Number of nearest neighbors
  - the number of approximate nearest neighbors used to construct the initial high-dimensional graph
- Minimum distance
  - the minimum distance between points in low-dimensional space controls how tightly UMAP clumps points together
    - Low values leading to more tightly packed embeddings
    - Larger values will make UMAP pack points together more loosely, focusing instead on the preservation of the broad topological structure

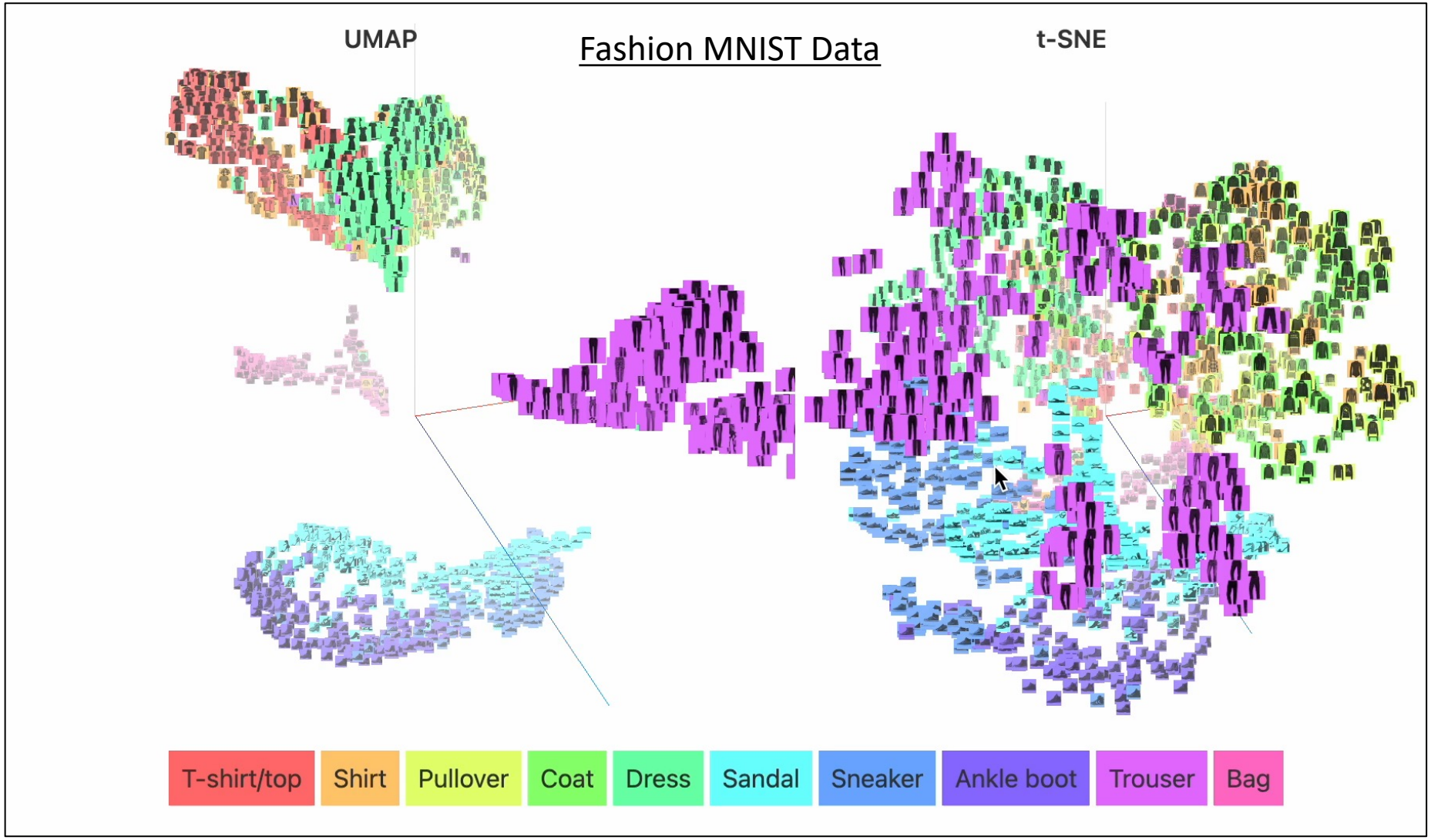
Two Important Parameters of UMAP



UMAP vs t-SNE


2D t-SNE projection
2D UMAP projection
perplexity: 5
time: 9m 18s
n_neighbors: 3
min_dist: 0.0
time: 25s

UMAP vs t-SNE


UMAP
Fashion MNIST Data
t-SNE
T-shirt/top
Shirt
Pullover
Coat
Dress
Sandal
Sneaker
Ankle boot
Trouser
Bag

Comments to both t-SNE and UMAP Methods

- Hyperparameters really matter!
- Cluster sizes may mean nothing
- Distances between clusters might not mean anything
- You may need more than one plot
- Random noise doesn’t always look random

A Comparative Study & Performance

MNIST dataset (downsampled to 2000 points)

PCA: 0.82 sec

LLE: 260 sec

Isomap: 280 sec

t-SNE: 250 sec

UMAP: 44 sec