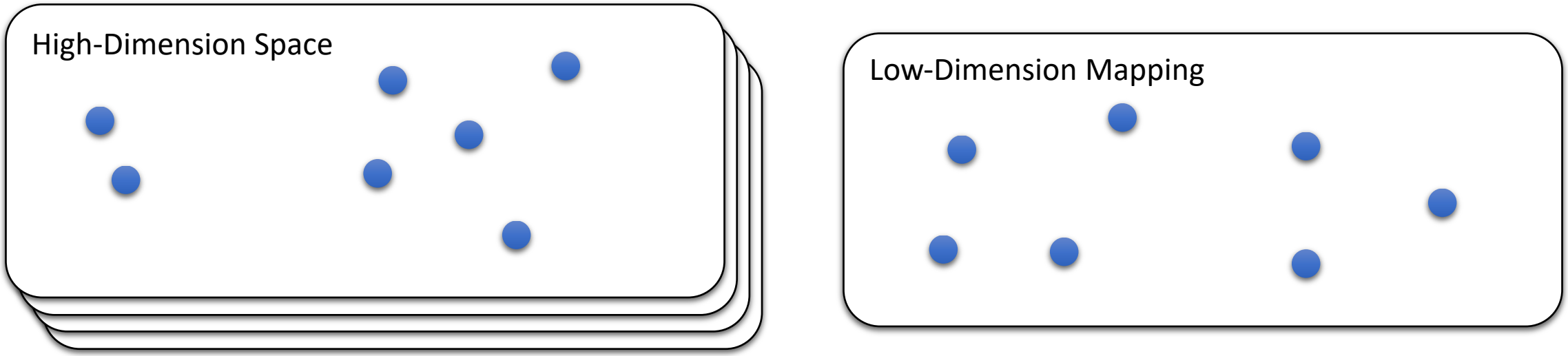
Dimension Reduction

From t-SNE to UMAP

Speaker: Jeff

Introduction

- **Concept:** Build map in which distances between points reflect similarities in the original data.

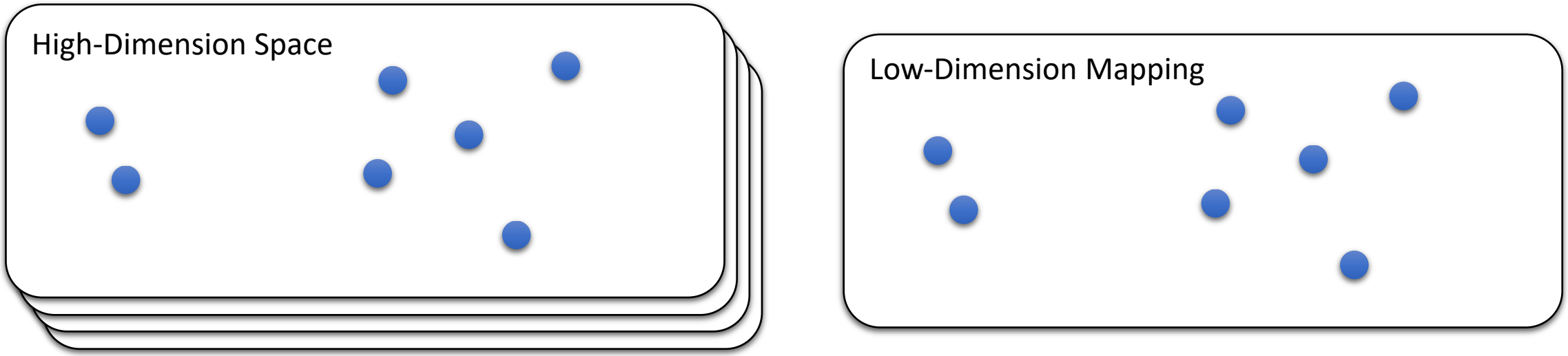


- Minimize some objective function that measures the discrepancy between similarities in the original data and similarities in the low dimensional map.

Introduction

- **Concept:** Build map in which distances between points reflect similarities in the original data.

Dimension Reduction / Embedding

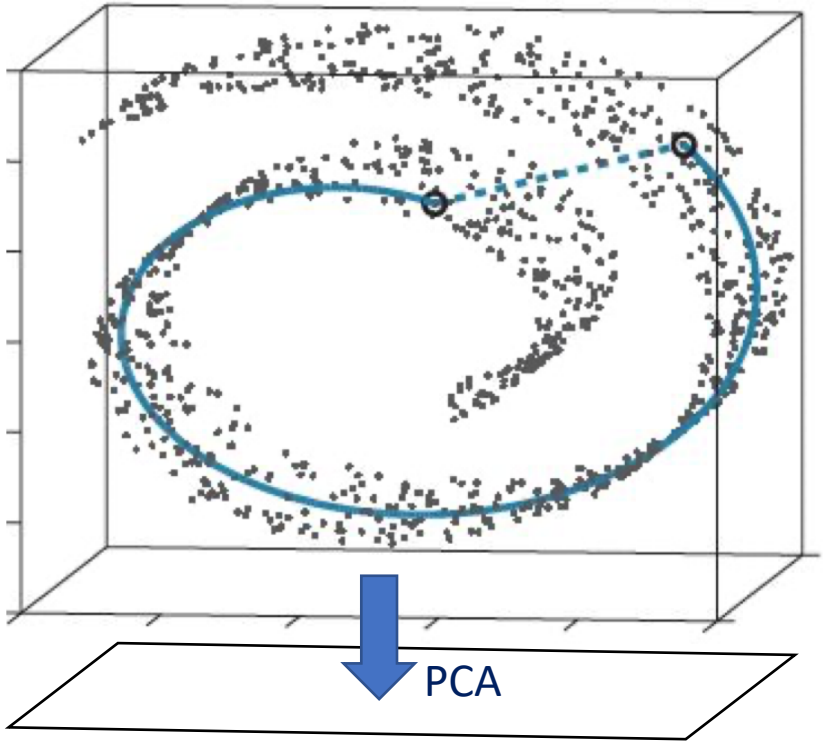


- Minimize some objective function that measures the discrepancy between similarities in the original data and similarities in the low dimensional map.

Introduction

- Matrix Factorization: Principal Component Analysis (PCA)
  - Preserving large pairwise distances in the low-D map.

  → Are such distances very reliable?



Distance under the 3-dim space. (Dash)

vs.

Distance over the manifold

- Similarity-based: SNE, t-SNE (van der Maaten, L. & Hinton, G. (2008)), UMAP (McInnes, L. et al.(2018))
- Autoencoder: scvis

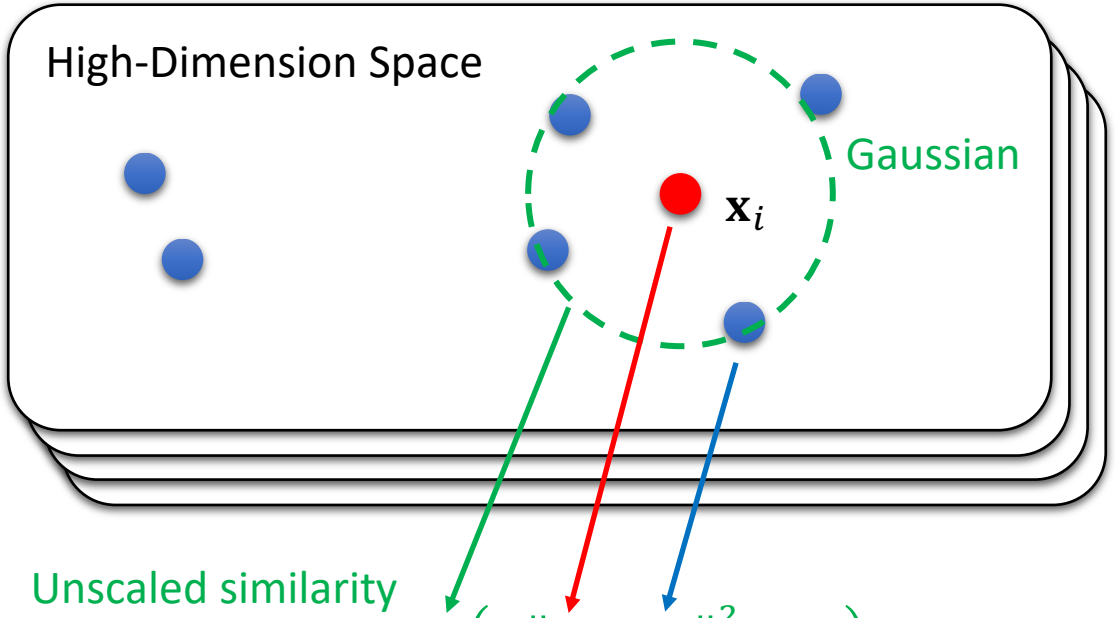
t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- t-SNE has already been successfully applied in a range of domains:
  - Bioinformatics, computer security, climate research, cancer research, etc.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Measure how close between two different high-dimensional objects.



Gaussian: $\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x_j-x_i)^2}{2\sigma^2}\right\}$

$$p_{ij} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2\right)}{\sum_k \sum_{l\neq k} \exp(-\|\mathbf{x}_k - \mathbf{x}_l\|^2/2\sigma^2)},$$ Scaled similarity

$$\sigma \; are \; all \; the \; same \; for \; each \; point$$

- Similar (closer) points in the high-D space → Large p_{ij}.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- In practice, we compute the input similarities slightly differently:

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_\mathrm{i} - \mathbf{x}_\mathrm{j}\|^2 / 2\sigma_i^2\right)}{\sum_{l \neq i} \exp\left(-\|\mathbf{x}_\mathrm{i} - \mathbf{x}_l\|^2 / 2\sigma_i^2\right)}$$ Scaled similarity

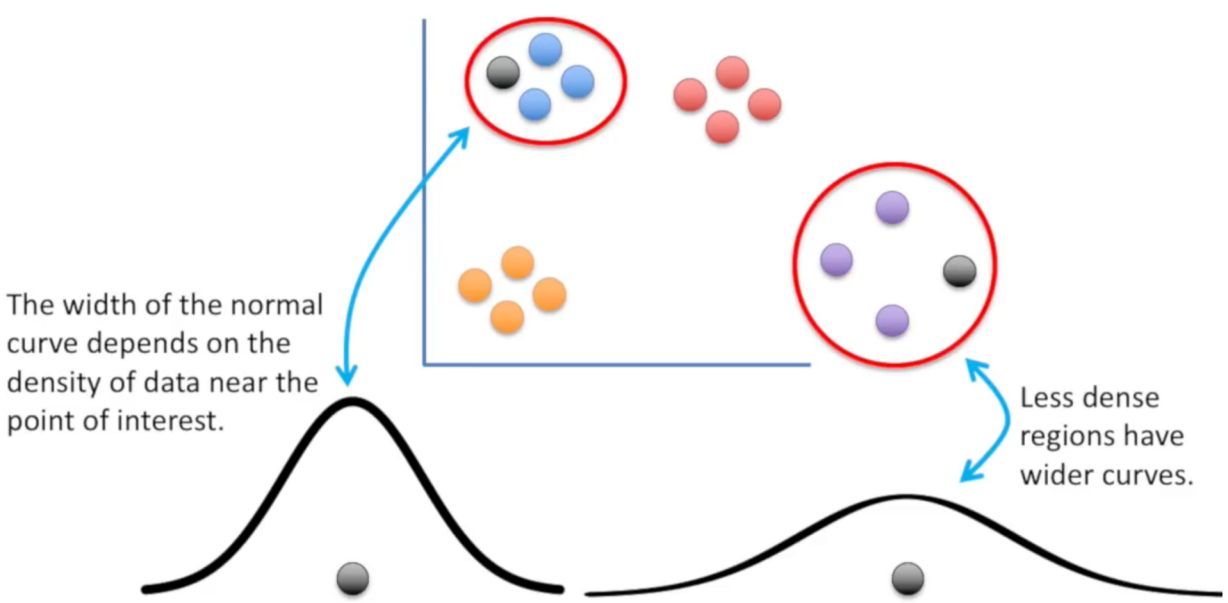
- $p_{j|i}$ only be comparable between different pairwise relationship after scaling.
- We set a different bandwidth σ_i for each point $\mathbf{x}_\mathrm{i}$.
  - Different part of the space may have different densities. (divide by σ_i → adapt to different densities.)
  - Each conditional dist. has a fixed *perplexity*.

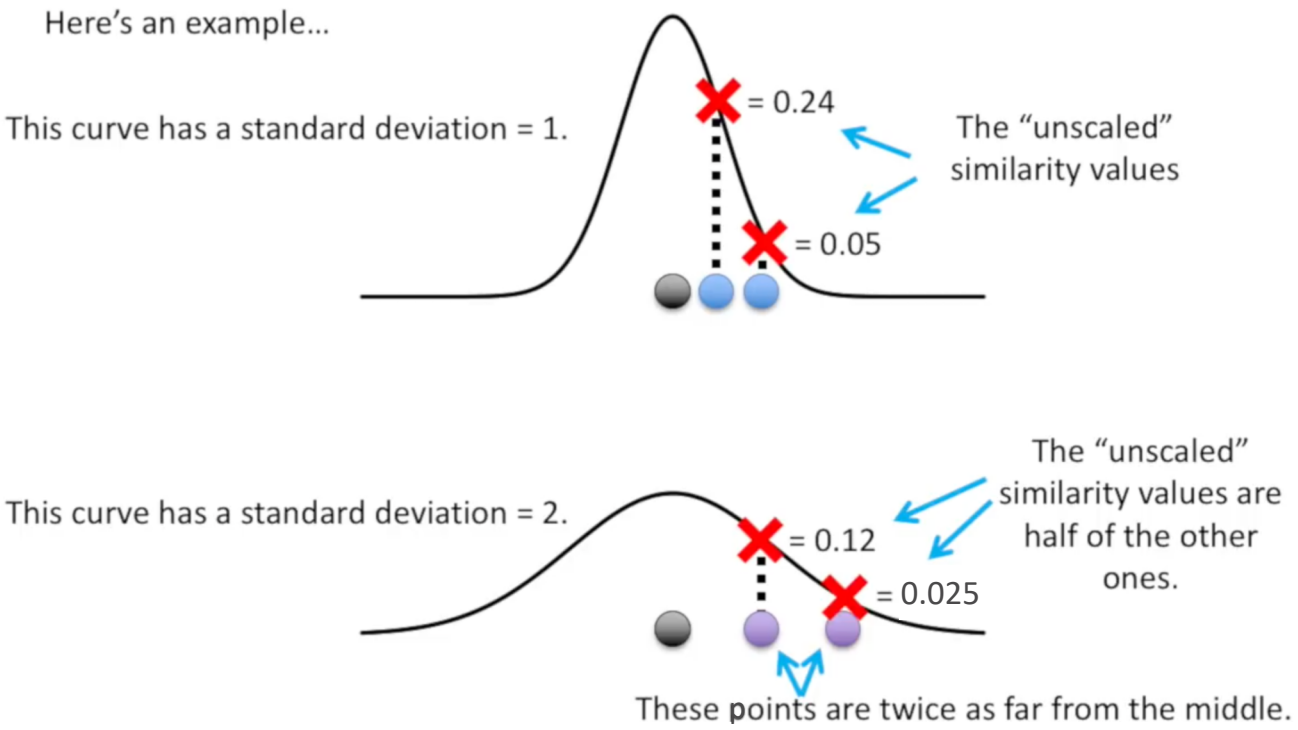
$$u = perp\left(p_{j|i}\right) = 2^{H\left(p_{j|i}\right)}, \qquad H\left(p_{j|i}\right) = -\sum_j p_{j|i} \log_2 p_{j|i} \rightarrow \ Shannon\ entropy$$

  - *perplexity* ↑ $\implies$ $Shannon\ entropy$ ↑ $\implies$ certainty ↓
    $\implies$ $p_{j|i}$ ↓, σ_i^2 ↑, Gaussian dist. → Uniform dist.
    $\implies$ t-SNE can cover more points. $\implies$ More global structure takes into account.
- Symmetrize the conditional dist.: $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$ → Similarities in High-D.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

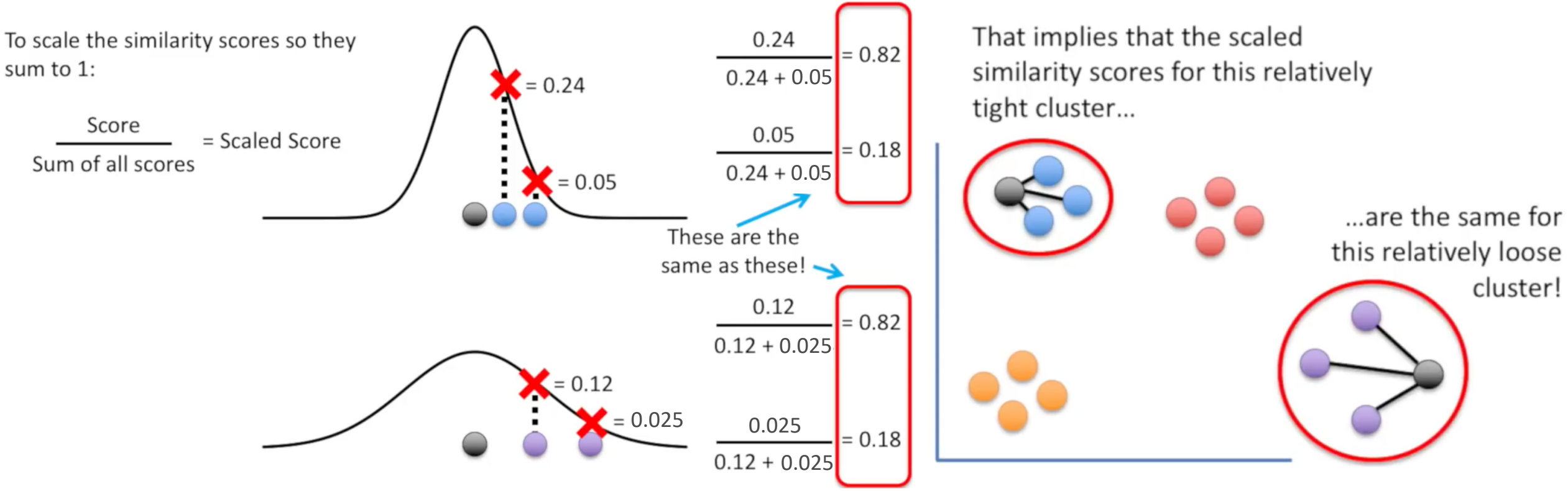
- The scaled similarity





t-Distributed Stochastic Neighbor Embedding (t-SNE)

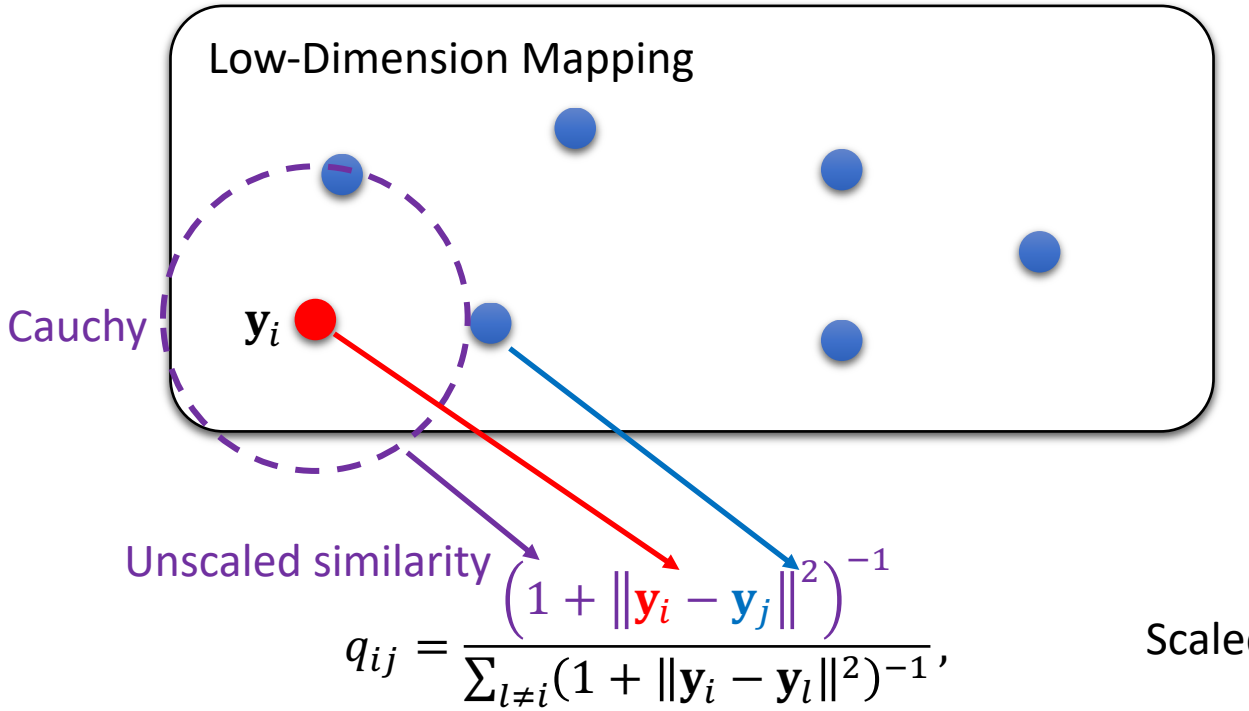
- The scaled similarity



The reality is a little more complicated, but only slightly.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Measure how close between two different low-dimensional map points.



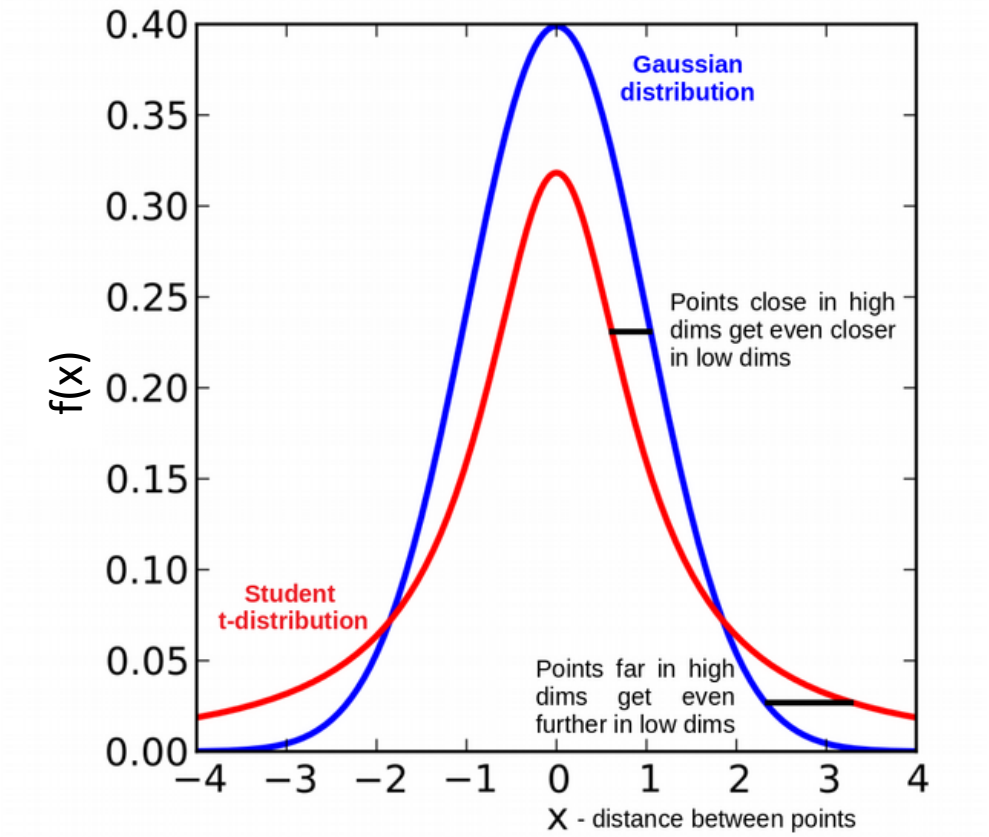
Cauchy: $\frac{1}{\pi\left(1+\left(\mathbf{y_j}-\mathbf{y_i}\right)^2\right)}$,

t-dist. , df = 1

- We want these q_{ij} to reflect p_{ij} as well as possible.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Why does t-SNE define map similarities as $q_{ij} \propto (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}$?
  - The denominator of p_{ij} and q_{ij} can be viewed as the normalization constant in the Gaussian and Student t-distribution.
  - Result: Dissimilar points have to be modeled as too far apart in the low-D map.

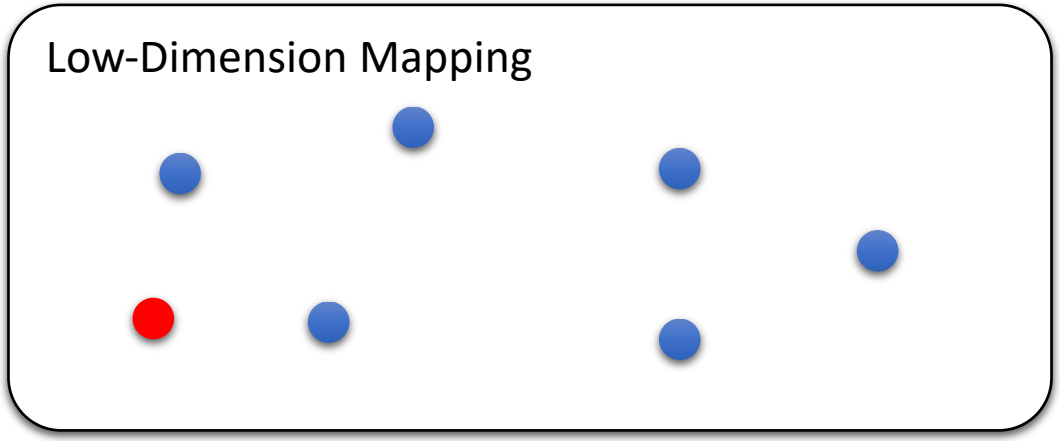


$\Longrightarrow$ Alleviate crowding problem.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Cost function: Measure the difference between these p_{ij}s and q_{ij}s.

$$KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

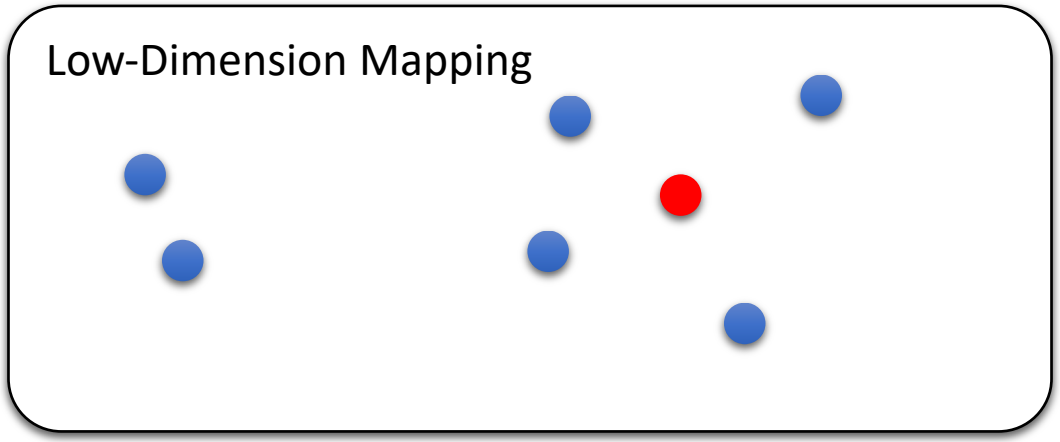


- Move points around to minimize $KL(P||Q)$.
  - Momentum gradient descent

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Cost function: Measure the difference between these p_{ij}s and q_{ij}s.

$$KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

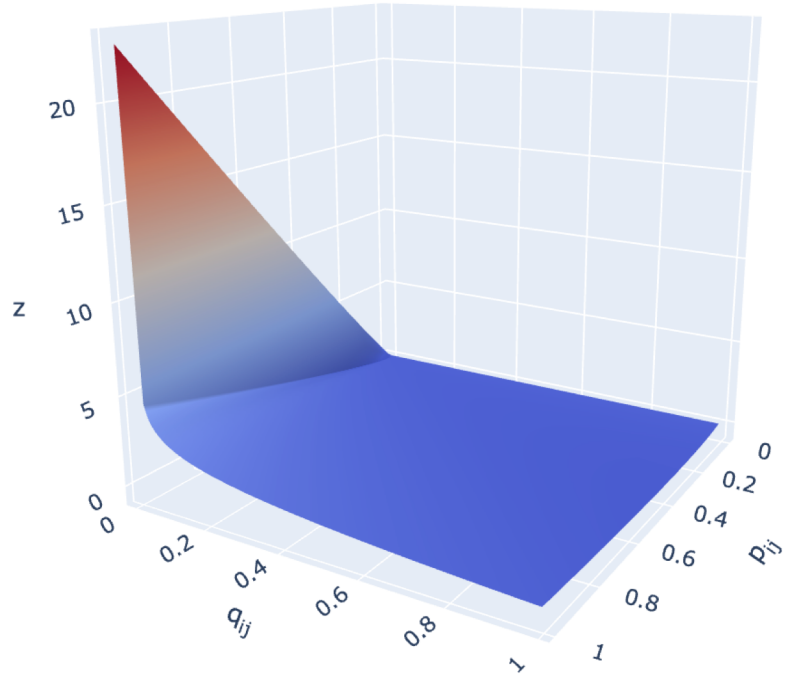


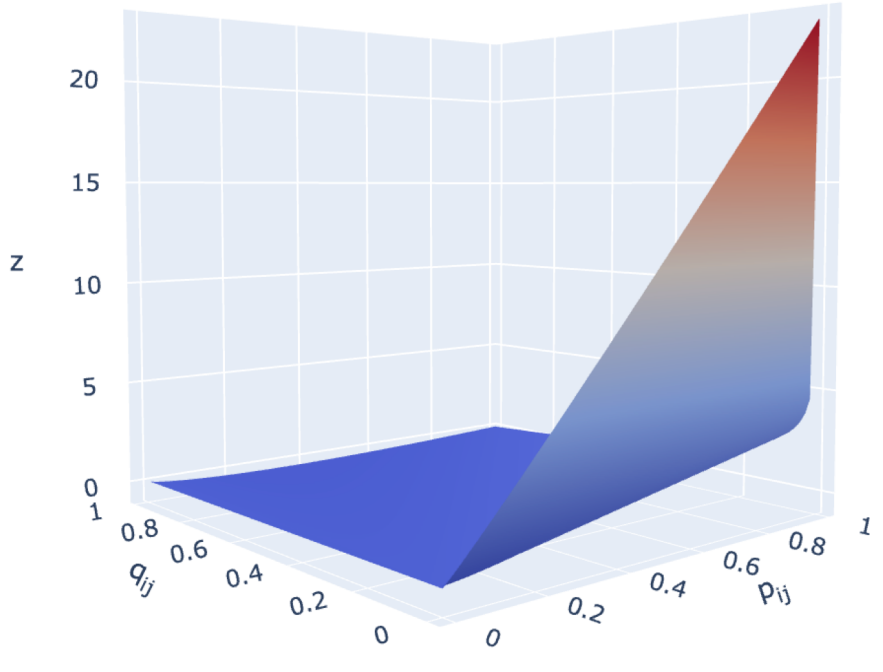
- Move points around to minimize $KL(P||Q)$.
  - Momentum gradient descent

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Kullback-Leibler divergences: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$
  - Similar in high-D, Dissimilar in low-D $\Longrightarrow$ Large p_{ij}, Small q_{ij} $\Longrightarrow$ Large cost (penalty)!
  - Dissimilar in high-D, similar in low-D $\Longrightarrow$ Small p_{ij}, Large q_{ij} $\Longrightarrow$ Small cost (penalty)!

  $\rightarrow$ KLD mainly preserves local similarity structure of the data.





t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Gradient of Kullback-Leibler divergences: $C = KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$
  - Let $d_{ij} = \|y_i - y_j\|,\ Z_i = \sum_{l \neq i} (1 + d_{il}^2)^{-1}$
  - $\frac{\partial C}{\partial y_i} = \sum_j \left(\frac{\partial C}{\partial d_{ij}} \cdot \frac{\partial d_{ij}}{\partial y_i} + \frac{\partial C}{\partial d_{ji}} \cdot \frac{\partial d_{ji}}{\partial y_i} \right) = \sum_j \left(\frac{\partial C}{\partial d_{ij}} + \frac{\partial C}{\partial d_{ji}} \right) \cdot \frac{y_i - y_j}{d_{ij}} = 2 \sum_j \frac{\partial C}{\partial d_{ij}} \cdot \frac{y_i - y_j}{d_{ij}}$
  - $\frac{\partial C}{\partial d_{ij}} = -\sum_{k \neq i} p_{ik} \frac{\partial}{\partial d_{ij}} log(q_{ik}) = -\sum_{k \neq i} p_{ik} \frac{\partial}{\partial d_{ij}} (log(q_{ik} Z_i) - log Z_i)$

    $= -\sum_{k \neq i} p_{ik} \left(\frac{1}{q_{ik} Z_i} \frac{\partial}{\partial d_{ij}} (1 + d_{ik}^2)^{-1} - \frac{1}{Z_i} \frac{\partial Z_i}{\partial d_{ij}} \right)$, $\frac{\partial}{\partial d_{ij}} (1 + d_{ik}^2)^{-1}, \frac{\partial}{\partial d_{ij}} Z_i$ only have value when $k, l = j$

    $= p_{ij} \left(\frac{1}{q_{ij} Z_i} 2 d_{ij} (1 + d_{ij}^2)^{-2} \right) + \sum_{k \neq i} \frac{-p_{ik}}{Z_i} \left((1 + d_{ij}^2)^{-2} \right) 2 d_{ij}$

    $= 2 d_{ij} \left(p_{ij} \frac{(1 + d_{ij}^2)^{-2}}{q_{ij} Z_i} - \frac{(1 + d_{ij}^2)^{-2}}{Z_i} \sum_{k \neq i} p_{ik} \right) = 2 d_{ij} \left(p_{ij} (1 + d_{ij}^2)^{-1} - q_{ij} (1 + d_{ij}^2)^{-1} \sum_{k \neq i} p_{ik} \right)$

    $= 2 d_{ij} \left(p_{ij} (1 + d_{ij}^2)^{-1} - q_{ij} (1 + d_{ij}^2)^{-1} \right) = 2 d_{ij} (p_{ij} - q_{ij}) (1 + d_{ij}^2)^{-1}$
  - $\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij}) \cdot \left(1 + \|y_i - y_j\|^2 \right)^{-1} \cdot (y_i - y_j)$

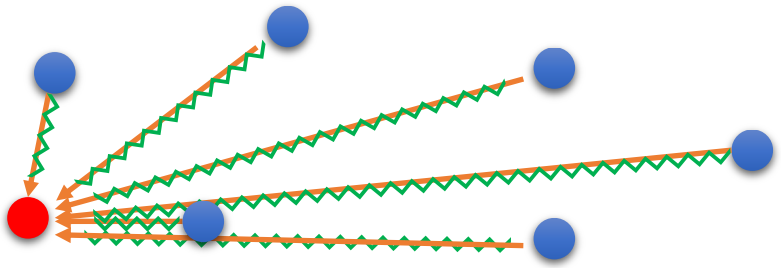
t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Gradient of Kullback-Leibler divergences
  - $\frac{\partial C}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij}) \cdot \left(1 + \|y_i - y_j\|^2\right)^{-1} \cdot \|y_i - y_j\| \cdot u\,, \quad u = \frac{y_i - y_j}{\|y_i - y_j\|}$ is an unit vector.

    Resultant force on a point

    Amount of the exertion/compression

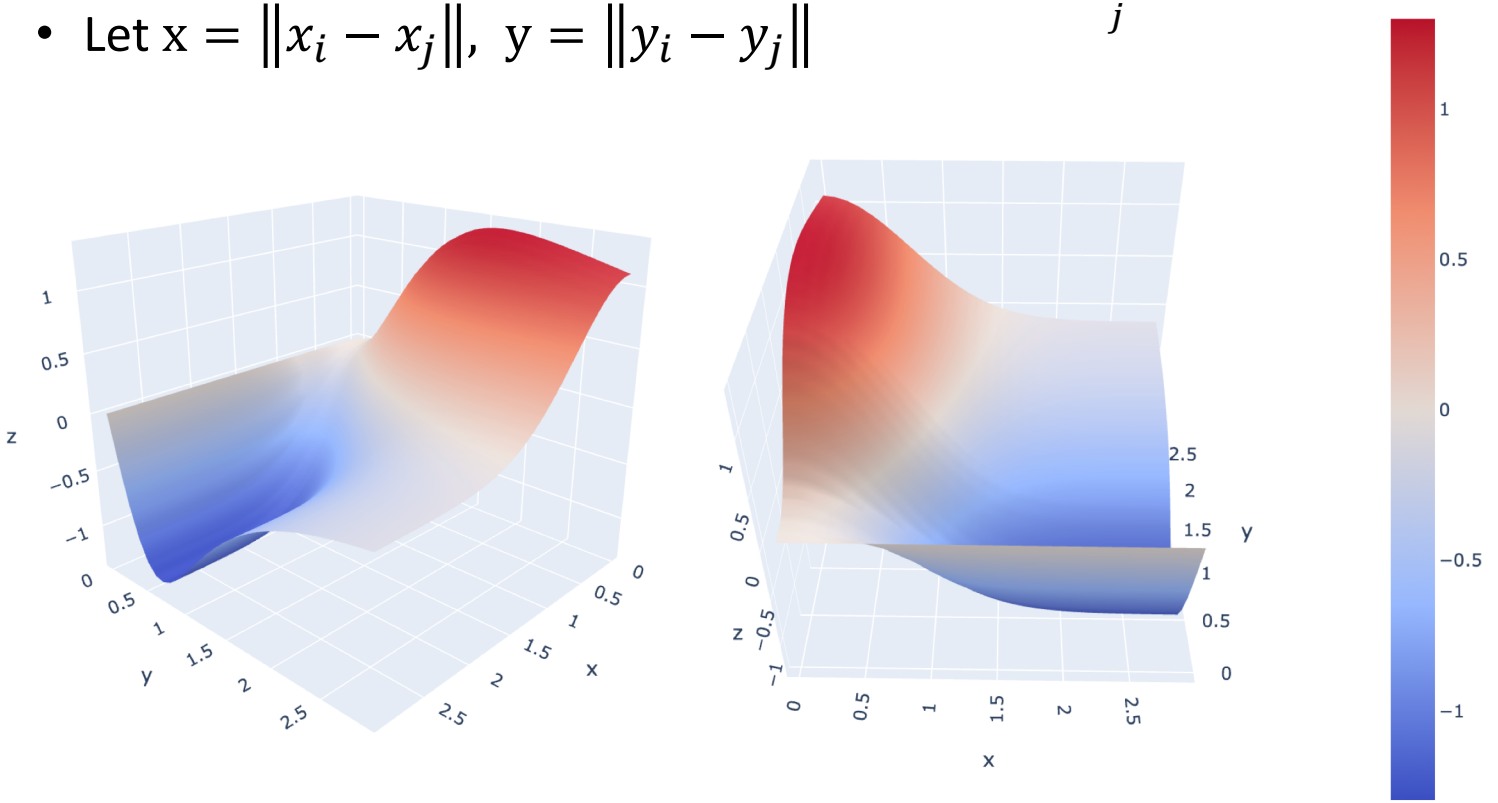
    Spring
  - Assume that the points in the low-D map are all connected with springs.

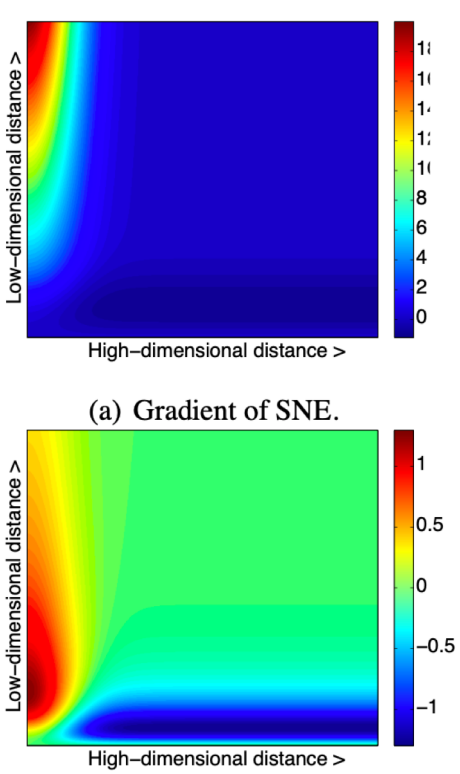


- If the q_{ij} is perfectly modeled the p_{ij}, then the amount of the exertion/compression would be zero.
  $\Longrightarrow$ No force in the spring.
- All pairwise between points (n^2 interactions between points) $\Longrightarrow$ Time consuming ($O(N^2)$).

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Gradient of Kullback-Leibler divergences $4\sum_j (p_{ij} - q_{ij}) \cdot \left(1 + \|y_i - y_j\|^2\right)^{-1} \cdot \|y_i - y_j\| \cdot u$
  - Let $\mathrm{x} = \|x_i - x_j\|,\ \mathrm{y} = \|y_i - y_j\|$





(a) Gradient of SNE.

(c) Gradient of t-SNE.

- Positive gradient $\Longrightarrow$ Attraction, Negative gradient $\Longrightarrow$ Repulsion
- (After considering q_{ij} in gradient completely) Dissimilar in high-D, similar in low-D $\Longrightarrow$ Repulsion.
- Modeling dissimilar (similar) data points by means of large (small) pairwise distances.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Barnes-Hut Approximation (van der Maaten, L. (2014))
  - Firstly, for the high-D data, simplify the computation of dissimilarities without negatively affecting the quality of the final embeddings.
  - Define a sparse approximation:
    - Let $\mathcal{N}_i$ represent the set of the $\lfloor 3u \rfloor$ nearest neighbors of $\mathbf{x}_\mathrm{i}$.
    - The setting of the σ_i^2 remains the same.
    - Redefine the pairwise similarities between different data points:

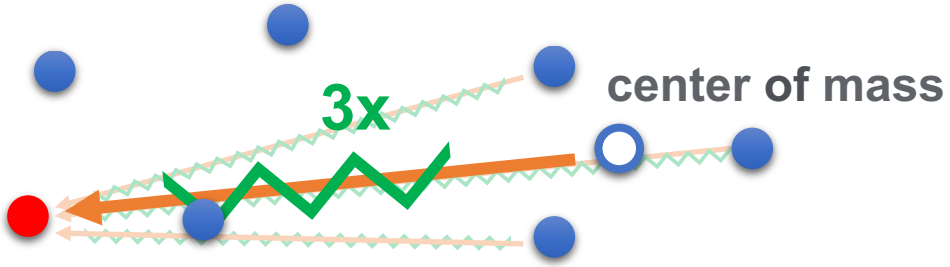
$$p_{j|i} = \begin{cases} \dfrac{\exp\left(-d(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j})^2 / 2\sigma_i^2\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(-d(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j})^2 / 2\sigma_i^2\right)}, & if\ j \in \mathcal{N}_i \\ 0, & otherwise \end{cases}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

  - The nearest-neighbor set $\mathcal{N}_i$ is constructed by using a depth-first search on the vantage-point tree that computes the distance of the objects stored in the nodes to the target object.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

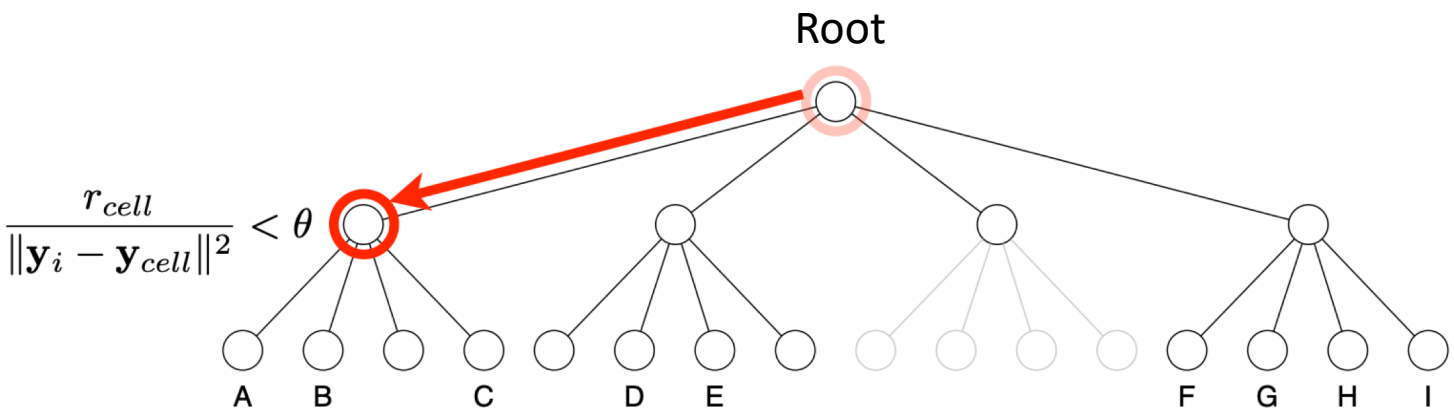
- Barnes-Hut Approximation (van der Maaten, L. (2014))
  - Then, for the low-D mapping, utilize Barnes-Hut approximation .
  - Many of the pairwise interactions between points are very similar.
  - Use the center of mass and combine the forces exerted by the points that are close together.
  - Use the new exertion/compression to approximate the original forces that are close together.

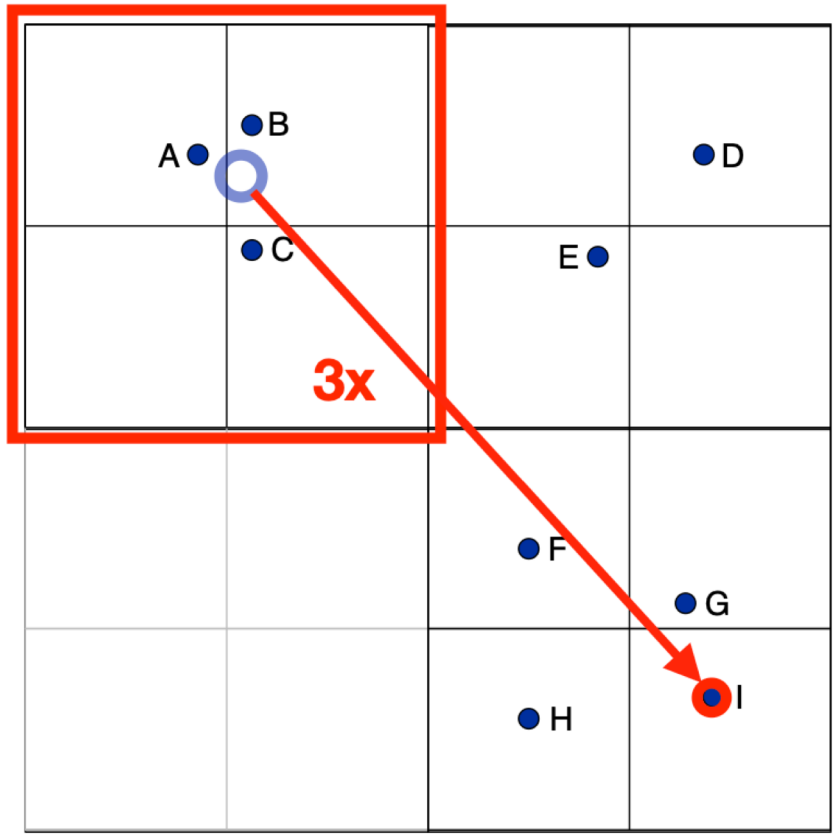


- Origin from the astronomy $\Longrightarrow$ Model the interactions between stars in the large galaxies of stars.
- Implement in practice: 2D: Quadtree/3D: Octtree
- O(N log N)

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Quadtree
  - A tree in which each node represents a rectangular cell.
  - Non-leaf node: Cell is split up into four smaller cells.
  - Leaf node: Cell contains at most one point of the embedding.
  - Root node: Cell contains the complete embedding.



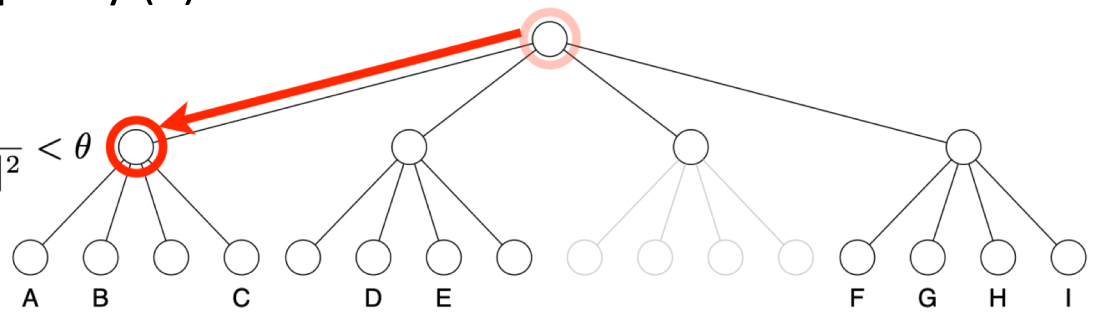


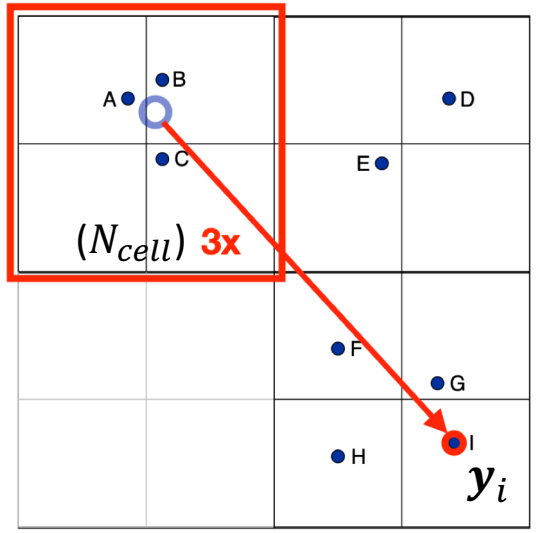
t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Quadtree
  - $\boldsymbol{y}_{cell}$: The center-of-mass of the embedding points that are located inside the corresponding cell.
  - N_{cell}: The total number of points that lie inside the cell.
  - r_{cell}: The length of the diagonal of the cell under consideration.
  - Construct quadtree in O(N) time by inserting the points one-by-one.
  - Splitting a leaf node whenever a second point is inserted in its cell. (A depth-first search)
  - Updating y_{cell} and N_{cell} of all visited nodes.
  - At every node in the quadtree, deciding whether the corresponding cell can be used as a "summary" for all points in that cell by using the following inequality (*).

(*) If a cell is sufficiently small and sufficiently far away from point $\boldsymbol{y}_i$ ⇒ Summary

$$\frac{r_{cell}}{\|\mathbf{y}_i - \mathbf{y}_{cell}\|^2} < \theta$$

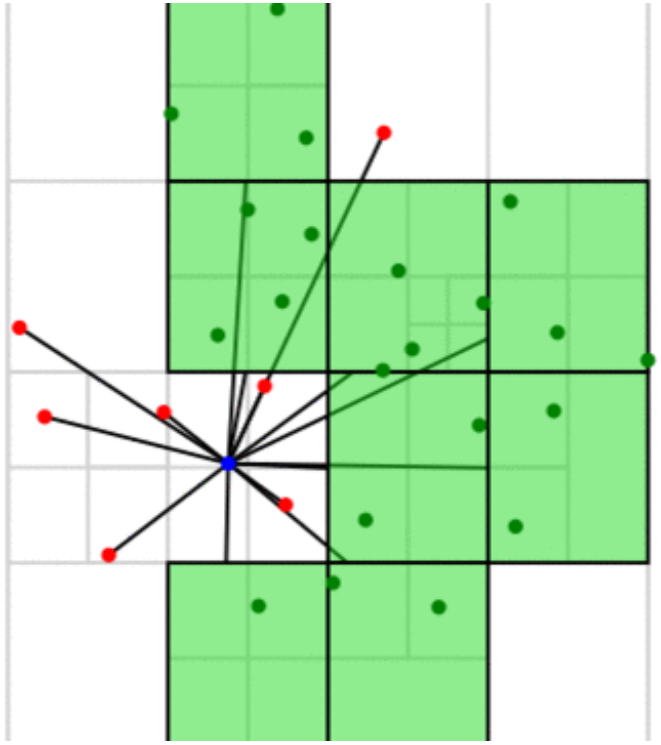


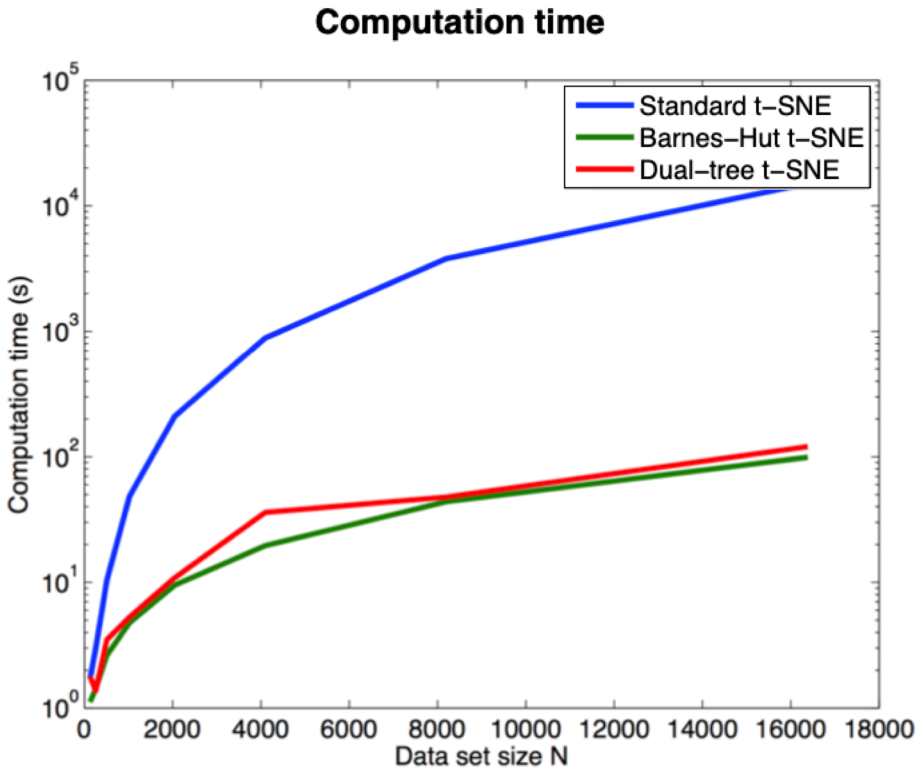


θ: speed v.s. accuracy. $\theta = 0$ ⇒ No speed (Standard t-SNE).

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Quadtree
  - A real example and the computation time.





t-Distributed Stochastic Neighbor Embedding (t-SNE)

Brief Summary

- t-SNE has been used in Bioinformatics, computer security, climate research, cancer research, etc.
- t-SNE use the Gaussian density to calculate a similarity measure between pairs of instances (points) in the high-D data.
- After that, due to the heavy tail property of the Student t-distribution, t-SNE use the Cauchy density to calculate a similarity measure between pairs of instances (points) in the low-D data.
- Optimize the low-D mapping layout by minimizing Kullback-Leibler divergence between the two similarities.
- Barnes-Hut Approximation was introduced with tree-based algorithm (Quadtree, Octtree) to accelerate the original t-SNE.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

Brief Summary

- There is two mainly important parameters that can be adjusted by user in t-SNE: *perplexity*, θ.

- In practice, by the current experience (Google Brain), small *perplexity* (e.g. 2) tends to cluster into many small clumps. In addition, parameter suggested by the author (5 ~ 50) is not always the optimal.
- By the experiment, *perplexity* value that exceeds the sample points would lead to unexpected behavior of the result.
- From the result of Google Brain, t-SNE indeed has a better performance than other older dimensionality-reduction algorithms.
- By studying how t-SNE behaves in simple cases (simple general case, different within- & between- clusters distance, noise recognition, topology pattern), it's possible to develop an intuition for what's going on.

- θ is a tradeoff between speed and accuracy. The lower θ is, the slower the speed is with higher accuracy.

Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP)

- Dimension reduction:
  - Preserve the data structure as much as possible.
  - The structure of the data → The interrelationships between different data points → Topology

- Except for the concept of building weighted neighborhood-graphs, the concept for UMAP is similar to t-SNE roughly.

- Topology: It is concerned with the properties of a geometric object that are preserved under continuous deformations, such as stretching, twisting, crumpling and bending, but not tearing or gluing.

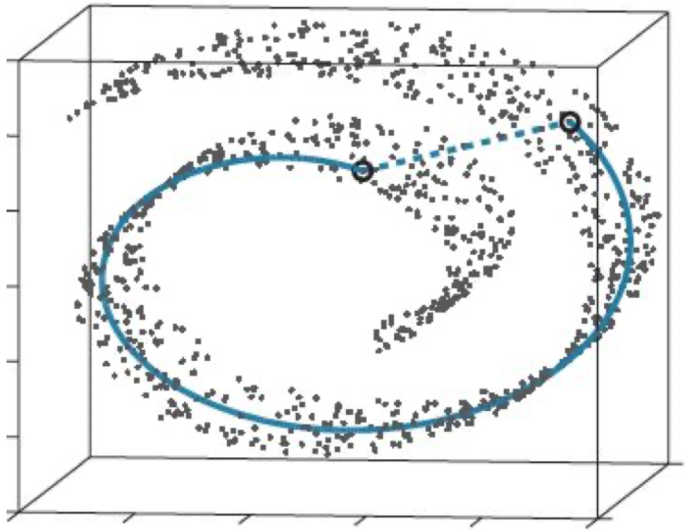
- Metric space: Simply to say, it is a space where we can measure distance.

- Topological space: It is defined as a set of points, along with a set of neighborhoods for each point, satisfying a set of axioms relating points and neighborhoods.

Uniform Manifold Approximation and Projection (UMAP)

- Homotopic

- Manifold is a kind of topological space that locally resembles Euclidean space near each point.



Uniform Manifold Approximation and Projection (UMAP)

- Fuzzy https://www.youtube.com/watch?time_continue=2&v=81HKNqruavc&feature=emb_logo

Uniform Manifold Approximation and Projection (UMAP)

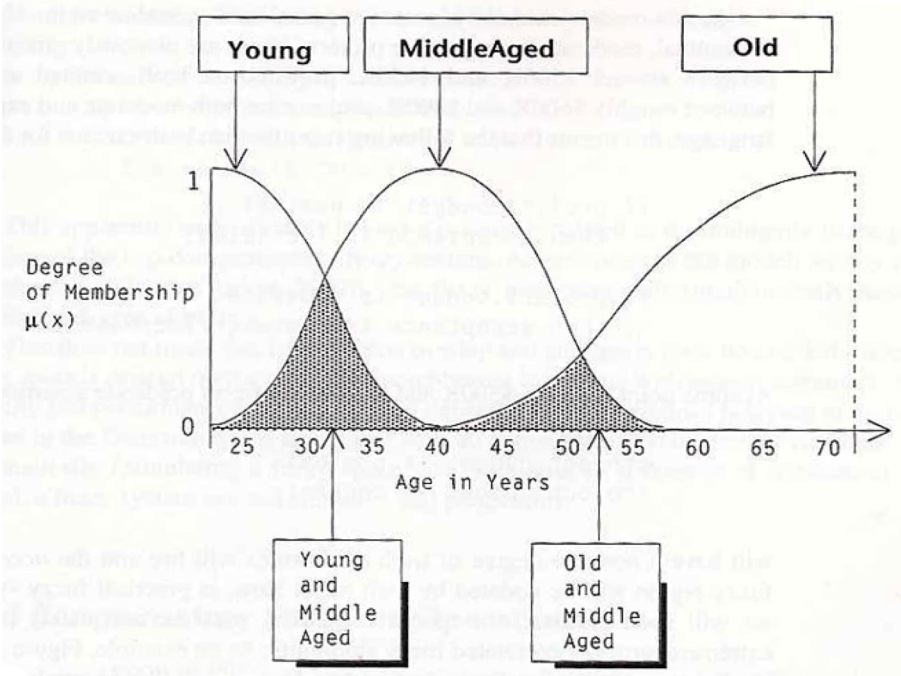
- Fuzzy
  - Another way to illustrate the uncertainty. (Different from the probability.)
  - Example: Having a date with a heterosexual stranger.
    - Before seeing the appearance of the stranger, we are not sure whether she (he) is beautiful (handsome)

      → Probability.
    - After seeing the appearance of the stranger, sometimes we still cannot conclude she (he) is beautiful (handsome). Furthermore, we might judge this by the "degree" of beautiful (handsome).

      → Fuzzy.
- Fuzzy set
  - It is somewhat like sets whose elements have degrees of membership.
- Membership function $\mu_A(x)$
  - X is a set, x is a point, and A is an arbitrary subset of X. $x \in X,\ A \subset X$.
  - Define $\mu_A(x)$: The degree of x belongs to A. $\mu_A(x) \in [0, 1]$ (formal term: The membership strength of x to the set A.)
  - It quantifies the grade of membership of the element in X to the fuzzy set A.

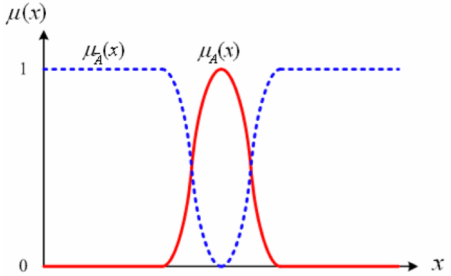
Uniform Manifold Approximation and Projection (UMAP)

- Membership function $\mu_A(x)$
  - Example
    - X: Age, A_1: Young, A_2: Middle-Aged, A_3: Old
    - $\mu_{A_1}(x) = \begin{cases} 1, & 0 \le x \le 25 \\ \frac{1}{1+\left(\frac{x-25}{5}\right)^2}, & 25 \le x \le 100 \end{cases}$

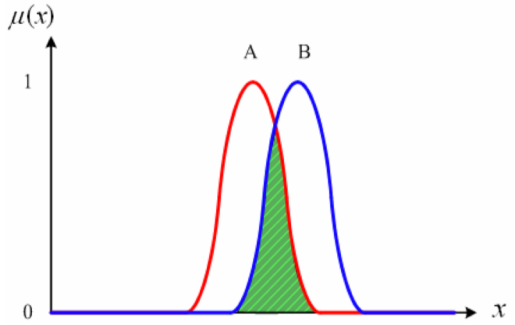


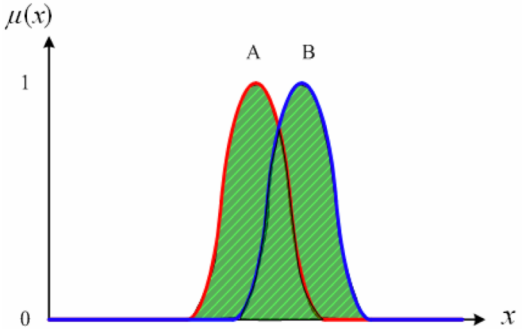
- Fuzzy operators

  Define A and B are fuzzy sets.



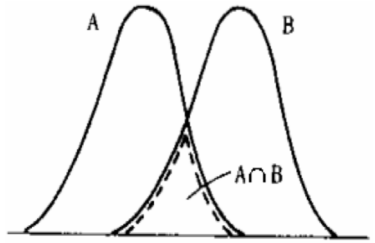
  - Complement $\bar{A}$:
    - $\mu_{\bar{A}}(x) = 1 - \mu_A(x)$
  - Intersection $A \cap B$ (T-norms):
    - $\mu_{A \cap B}(x) = \min[\mu_A, \mu_B] \equiv \mu_A(x) \wedge \mu_B(x)$
  - Union $A \cup B$ (T-conorms, S-norms):
    - $\mu_{A \cup B}(x) = \max[\mu_A, \mu_B] \equiv \mu_A(x) \vee \mu_B(x)$

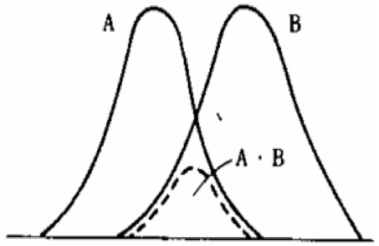


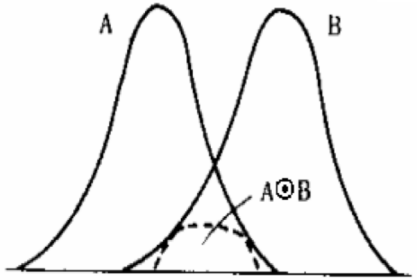


Uniform Manifold Approximation and Projection (UMAP)

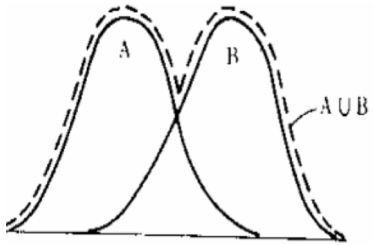
- Common operators for fuzzy operators
  - Intersection $A \cap B$:
    - Logical Product, Standard Intersection:
      - $t(\mu_A(x), \mu_B(x)) = \min[\mu_A, \mu_B] = \mu_{A \cap B}(x)$
    - Algebraic Product:
      - $t(\mu_A(x), \mu_B(x)) = \mu_A(x) \cdot \mu_B(x) = \mu_{A \cdot B}(x)$
    - Bounded Product:
      - $t(\mu_A(x), \mu_B(x)) = \max\{0,\ \mu_A(x) \cdot \mu_B(x) - 1\} = \mu_{A \odot B}(x)$

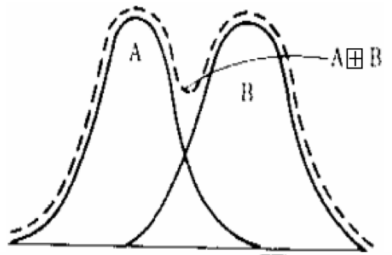


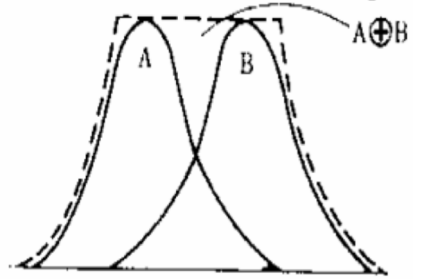




  - Union $A \cup B$:
    - Logical Sum, Standard Union:
      - $s(\mu_A(x), \mu_B(x)) = max[\mu_A, \mu_B] = \mu_{A \cup B}(x)$
    - Algebraic Sum:
      - $s(\mu_A(x), \mu_B(x)) = \mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x) = \mu_{A \boxplus B}(x)\ or\ \mu_{A+B}(x)$
    - Bounded Sum:
      - $s(\mu_A(x), \mu_B(x)) = min\{1,\ \mu_A(x) + \mu_B(x)\} = \mu_{A \oplus B}(x)$







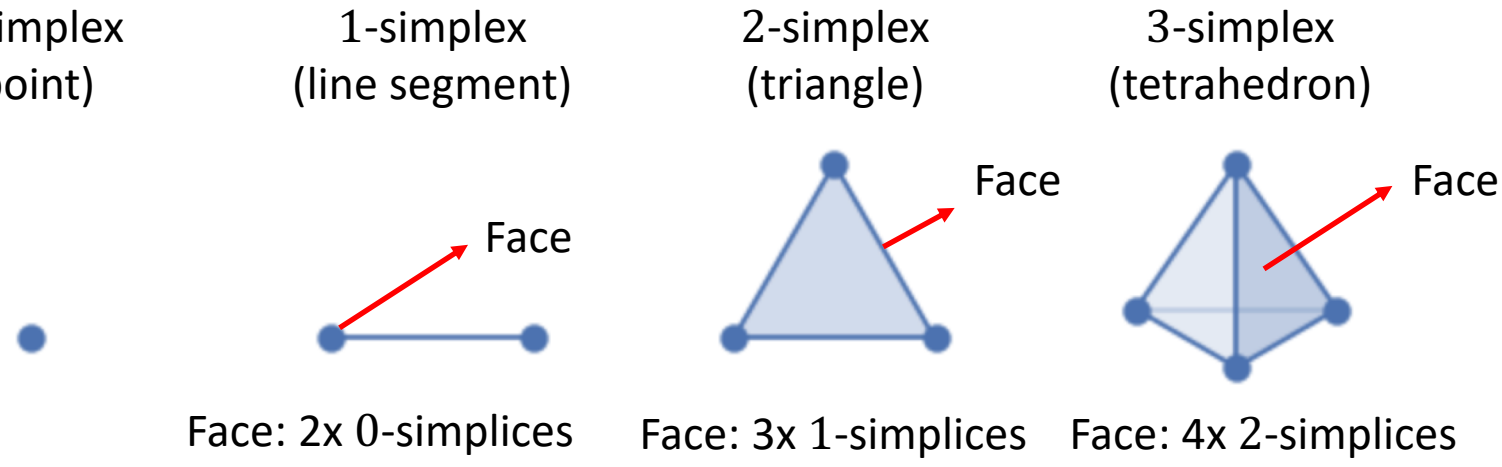
Uniform Manifold Approximation and Projection (UMAP)

- Measurement of the information for fuzzy
  - Herein, because there is no probabilistic concept, we need to redefine all the measurement.
  - Each measurement has many definitions, so the choice depends on the favor.
  - Shannon entropy (Li and Liu):
    - $H(A) = -\sum_{i=1}^{n}\{\mu_A(x_i)\log(\mu_A(x_i)) + [1-\mu_A(x_i)]\log(1-\mu_A(x_i))\}$
  - Cross entropy:
    - $H(A,B) = \sum_{i=1}^{n}\left\{\mu_A(x_i)\log\left(\frac{\mu_A(x_i)}{\mu_B(x_i)}\right) + [1-\mu_A(x_i)]\log\left(\frac{1-\mu_A(x_i)}{1-\mu_B(x_i)}\right)\right\}$
    - From https://link.springer.com/article/10.1186/s40467-015-0029-5
  - Jensen-Shannon divergence:
    - https://www.sciencedirect.com/science/article/abs/pii/S1568494619301693

Uniform Manifold Approximation and Projection (UMAP)

A Topology Primer

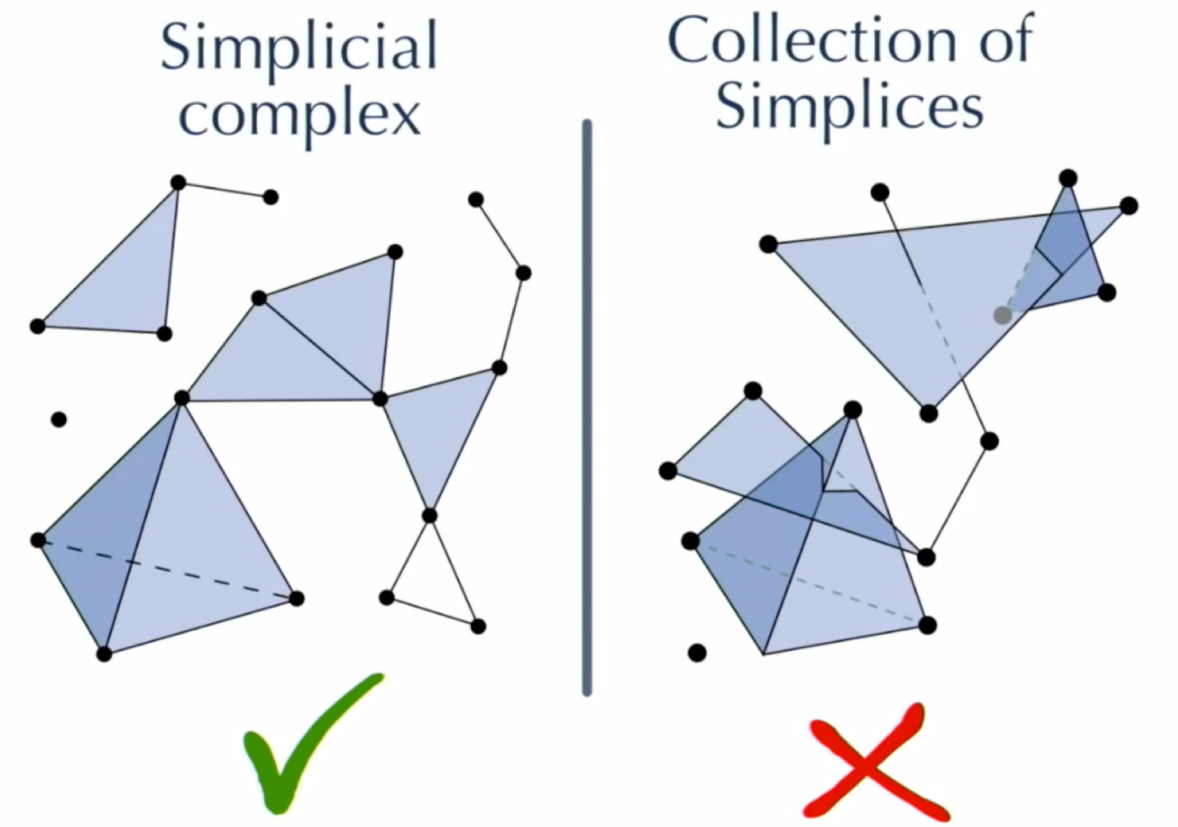
- Simplices and some definitions
  - A foundation to build a k-dimensional object geometrically. → building blocks.
  - k-simplex: The convex hull of $k + 1$ independent points.
  - Face: The convex hull of any nonempty subset of the $k + 1$ points that define a k-simplex is called a face of the simplex.
  - Example:



  - It can be used to build the easy combinatorial representation of topological space.

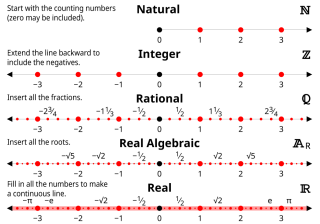
Uniform Manifold Approximation and Projection (UMAP)

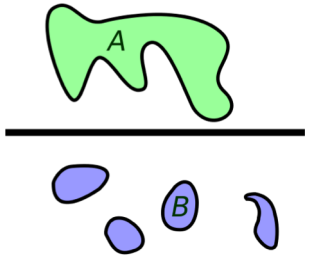
- Simplicial Complex
  - Glue simplices along their faces to form simplicial complexes.
  - It can build in practice for almost any topological space.
  - Complicated and continuous topology ⇒ Finite, simple and combinatorial objects.



Uniform Manifold Approximation and Projection (UMAP)

- Connected space
  - Disconnected space: A space X is disconnected if exists non-empty open sets U and V of X such that $U \cap V = \emptyset$ and $X = U \cup V$.
  - Connected space: A space X is called connected if it is not disconnected.
- Locally connected space
  - Locally connected: A space X is locally connected at x, if every neighborhood of x contains a connected open neighborhood.
  - Locally connected space: A space X is called locally connected if it is locally connected at every point $x \in X$.
  - Example:
    - The space $\mathbb{Q}$ (rational numbers) is neither connected nor locally connected.
    - The subspace $[0, 1] \cup [2, 3]$ of the real line $\mathbb{R}^1$ is locally connected but not connected.
    - Manifold is locally connected.

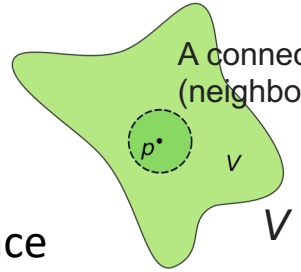




Connected space

Disconnected space

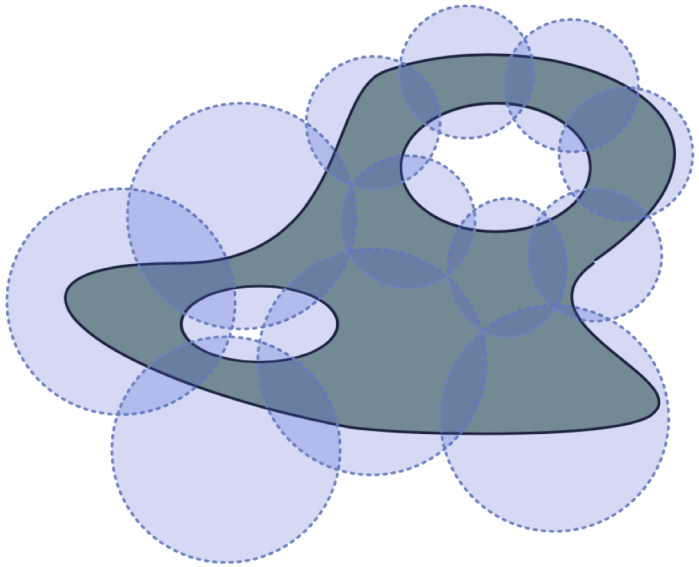
Locally connected space



V is a neighborhood of p

Uniform Manifold Approximation and Projection (UMAP)

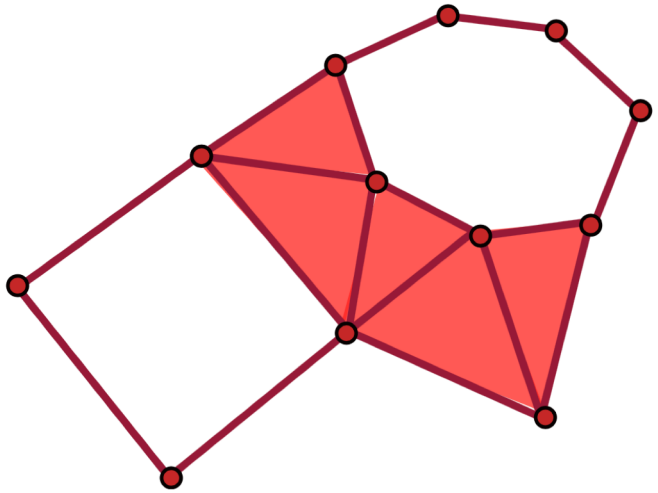
- Open cover and Čech complex
  - Open cover: A family of sets whose union is the whole space.
  - Čech complex: A combinatorial way to covert the open cover into a simplicial complex.
  - Nerve theorem: It ensures the homotopical equivalence (No information loss).



Open cover
(complicated and continuous topology)

Homotopically equivalent

Capture most information
we need to know.
By the Nerve Thm.

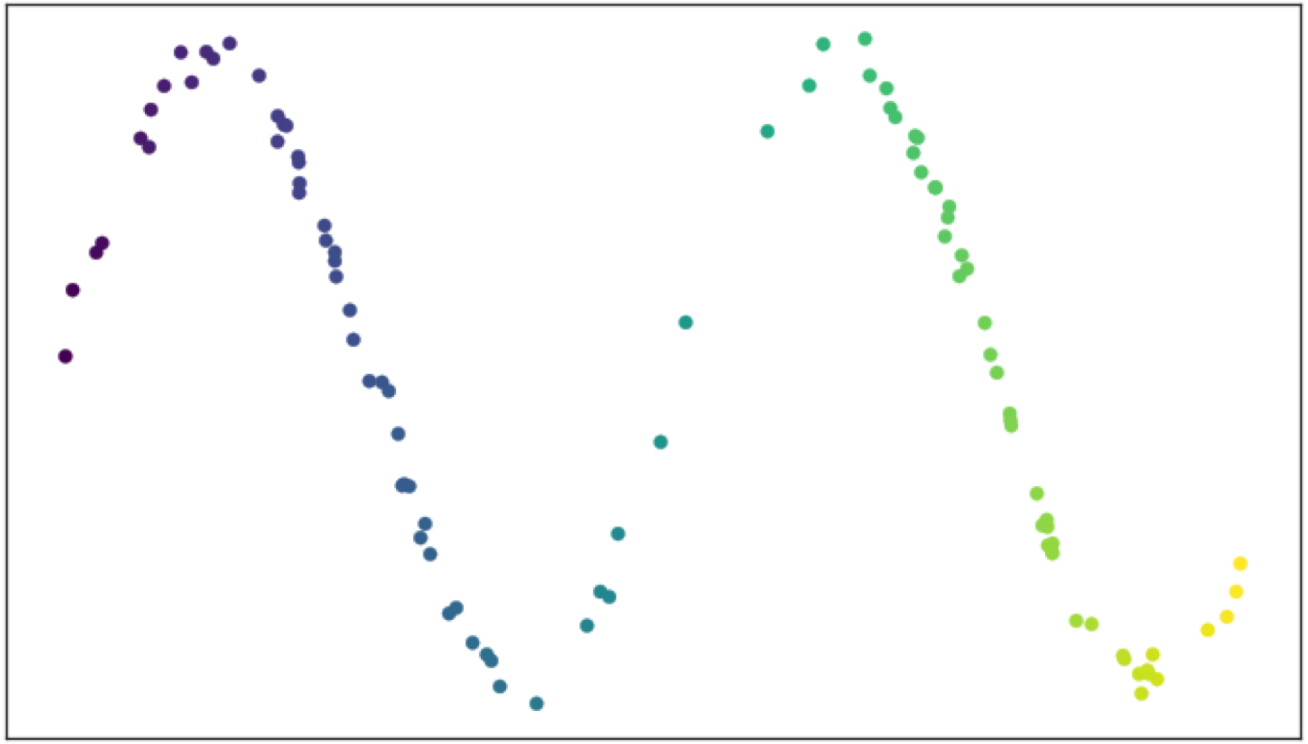


Simplicial complex, Čech complex
(Simple and combinatorial objects)

Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

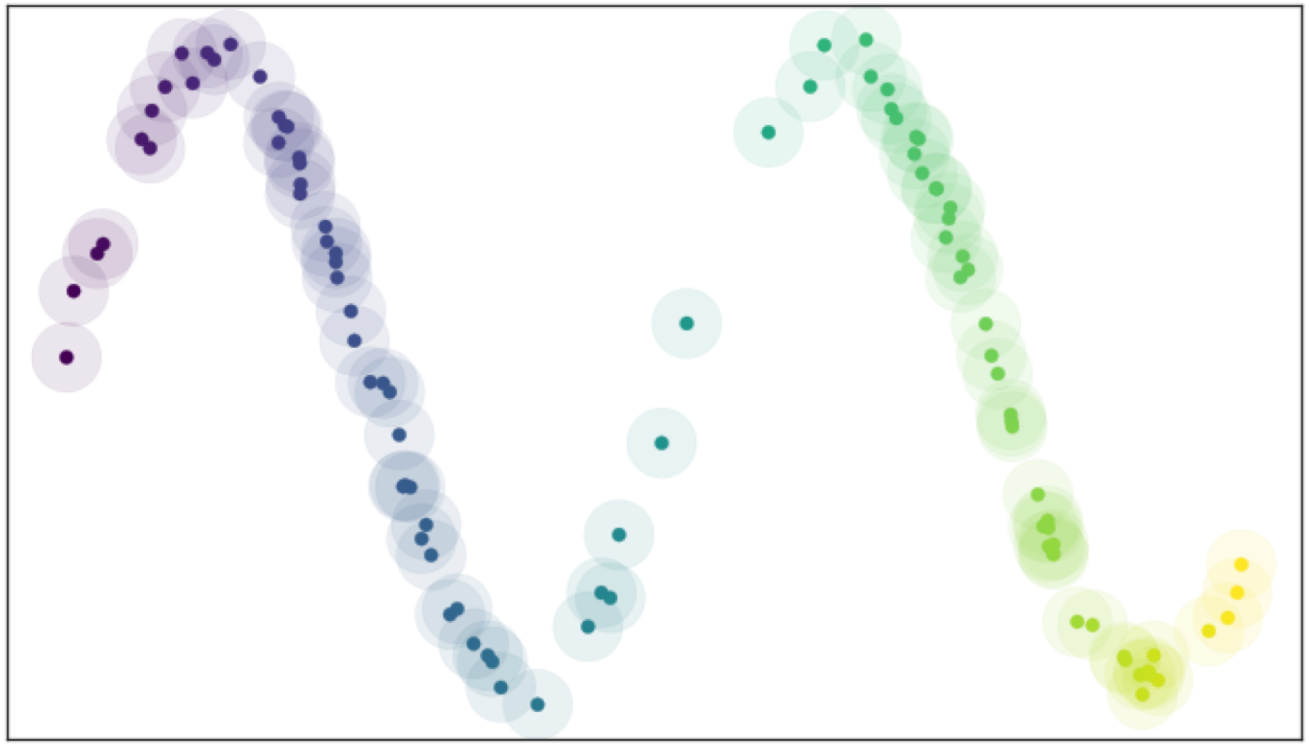
- Data: A noisy sine wave



Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

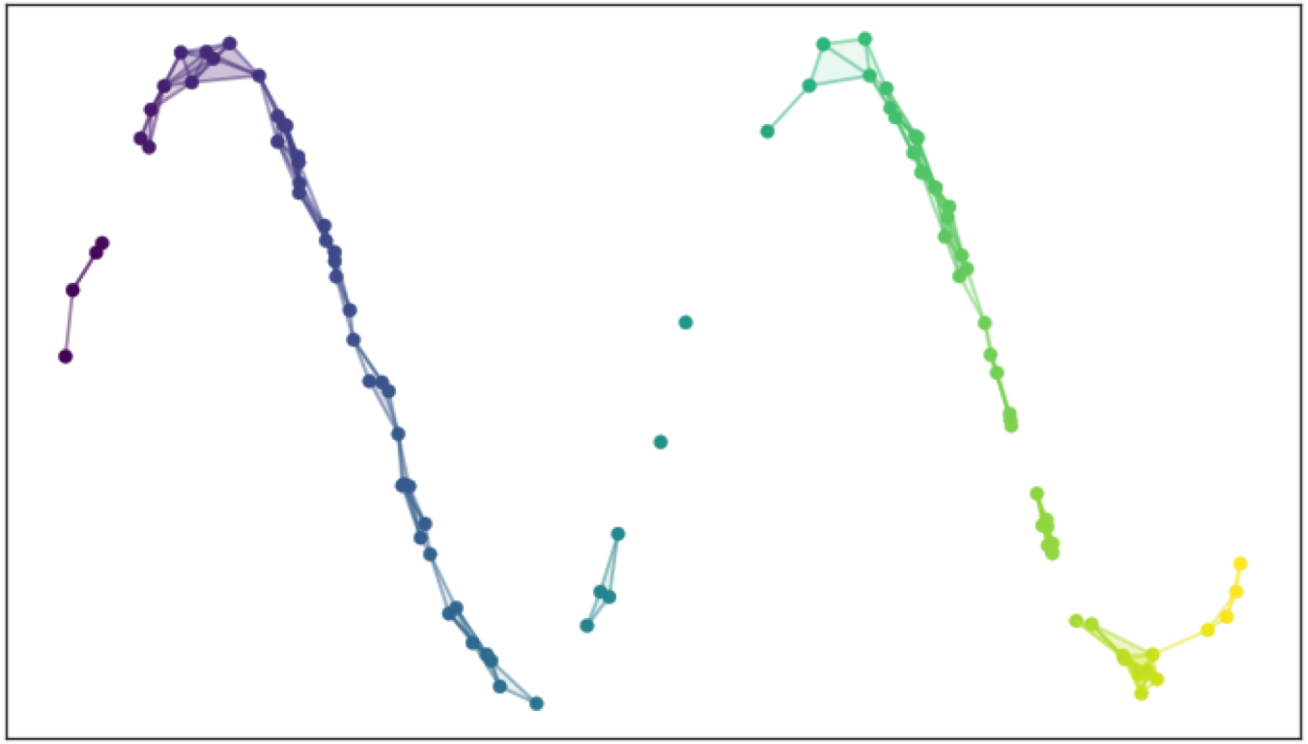
- Building open covers that can cover the manifold. (In this case, draw open balls)



Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

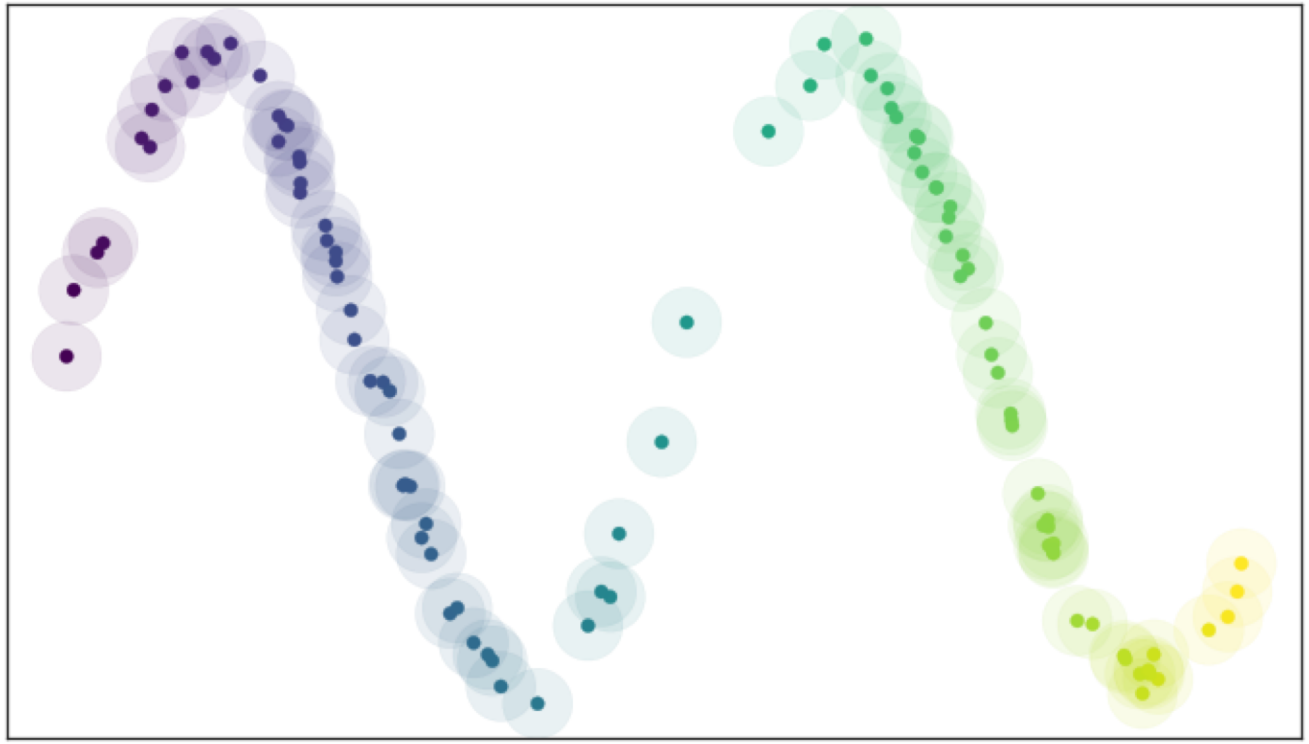
- Take the nerve of the cover to get the simplicial complex.
- A 1-simplex is formed by a pair of open sets that overlap.
- A 2-simplex is formed by triple open sets that overlap.
- ……



Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

- Assume we are in a metric space temporarily, the choice of radius for each open ball is important.
  - Too small → Many connected components
  - Too large → High dimensional simplices → Computation cost



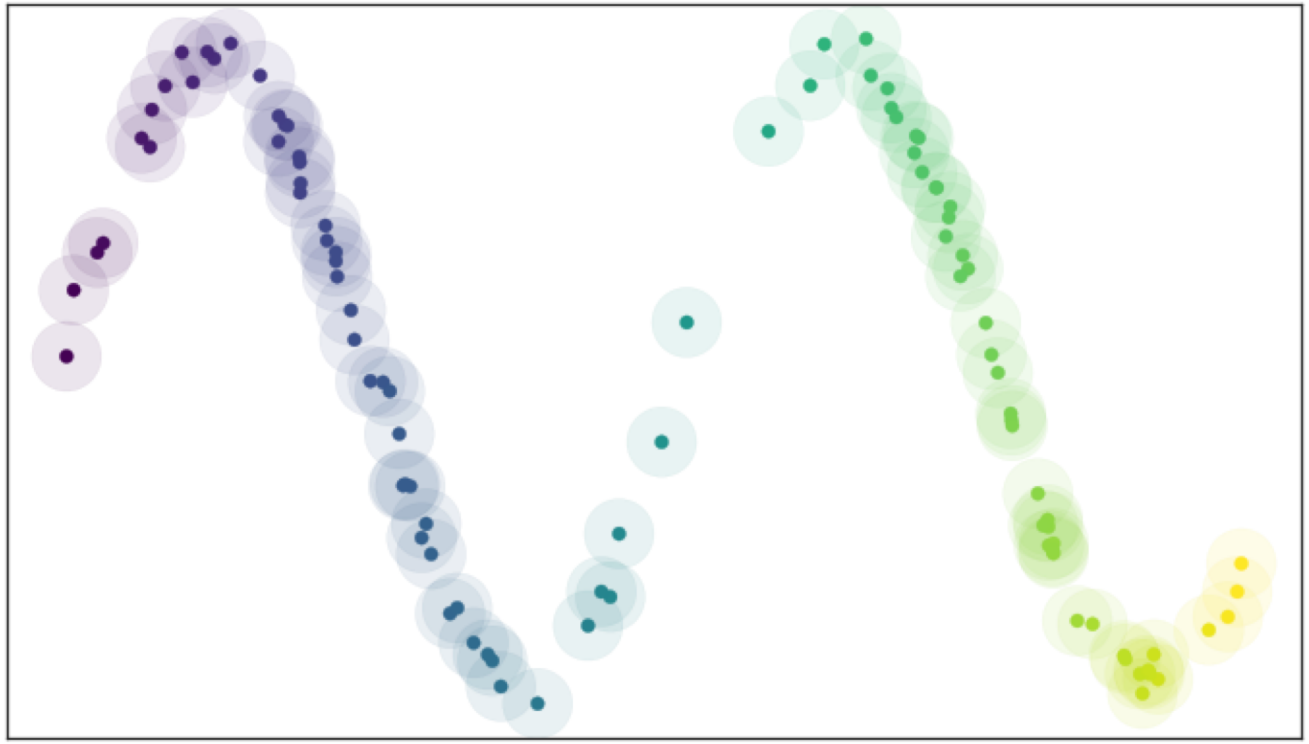
Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

- Attention: By the Nerve theorem, open covers should cover the whole space (manifold).
- However, there are gaps between the open covers.

  → This topology is not the topology we want to capture. (The simplicial complex cannot be used.)

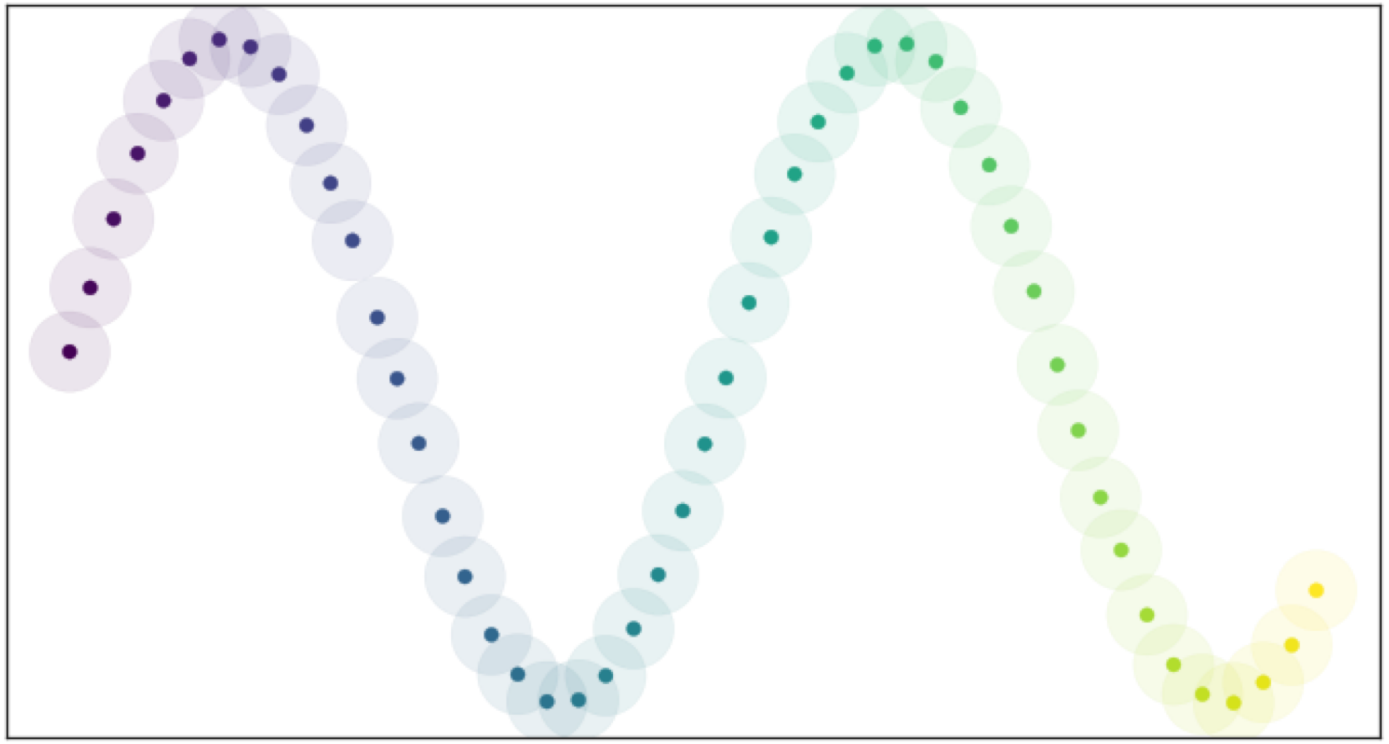
  → Some adjustments or assumptions are required to make.



Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

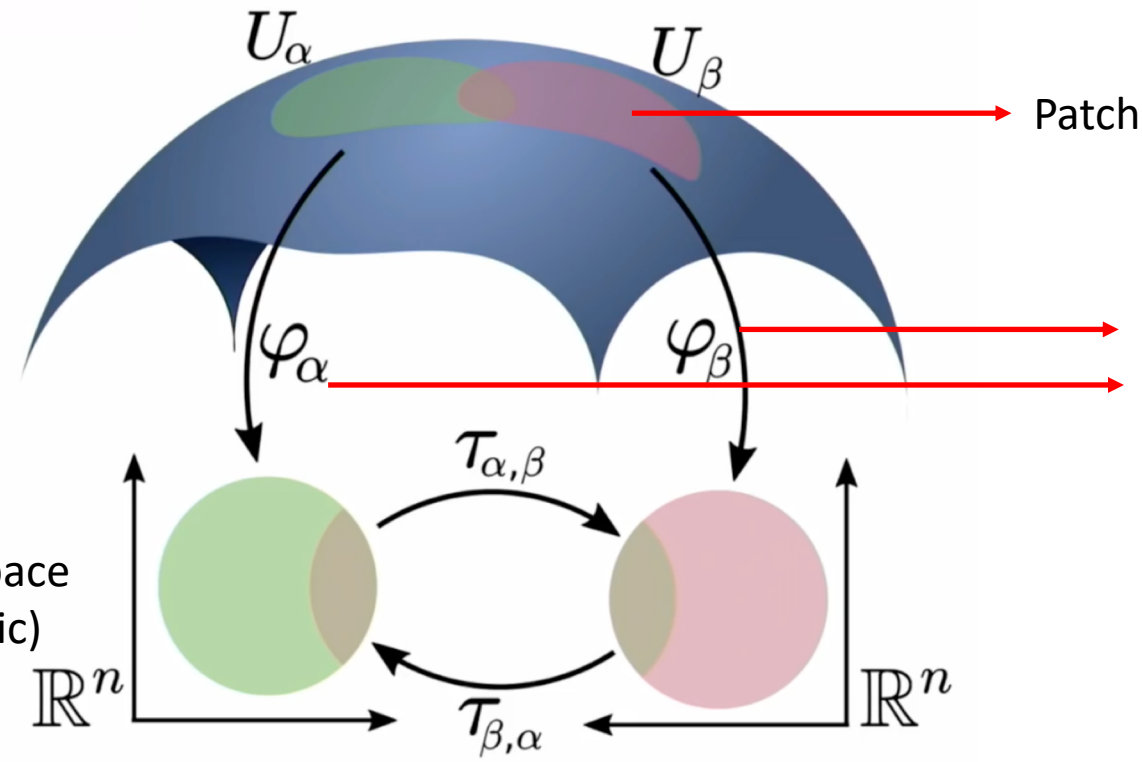
- If the data is uniformly distributed on the manifold, then the cover will be good.
  → Make assumption
- Assumption: The data is uniformly distributed on a manifold.



Uniform Manifold Approximation and Projection (UMAP)

- Some assumptions and definitions
  - Assumption: The data is uniformly distributed on a manifold.
    ⟶ To deal with real data, we have to define a Riemannian metric on the manifold.
  - Adapt uniform manifold to the real data under the Euclidean space with the following concept:

Manifold (Riemannian metric)



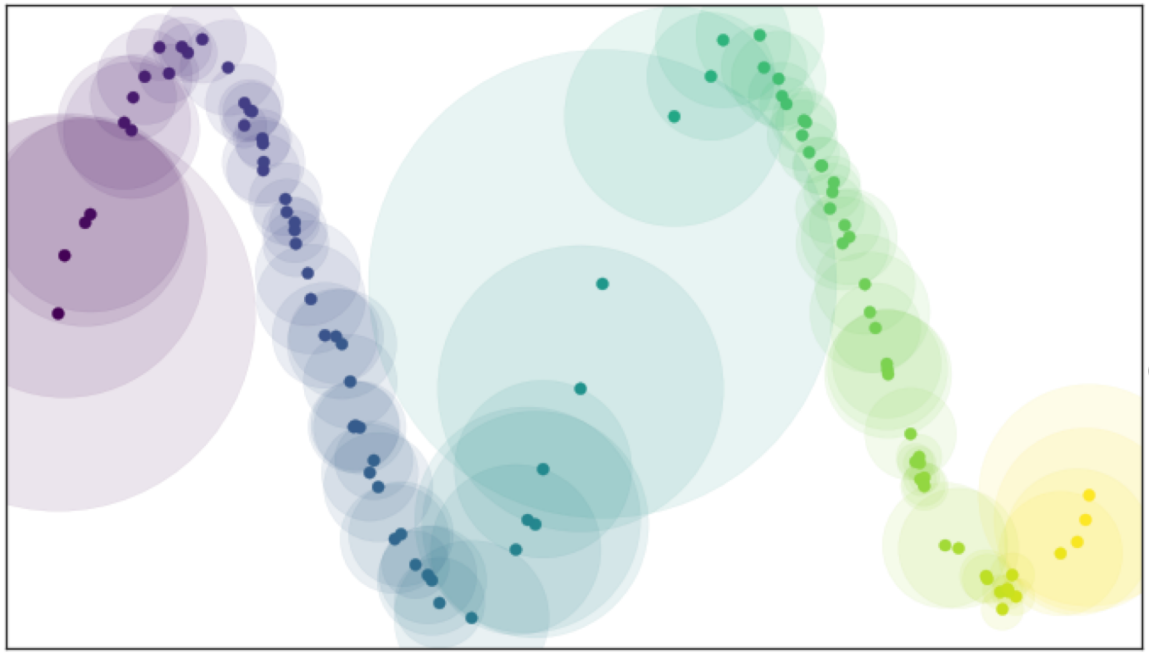
Map to the Euclidean space in different scales (depending on which patch you are in).

High-D Euclidean space (Riemannian metric)

Uniform Manifold Approximation and Projection (UMAP)

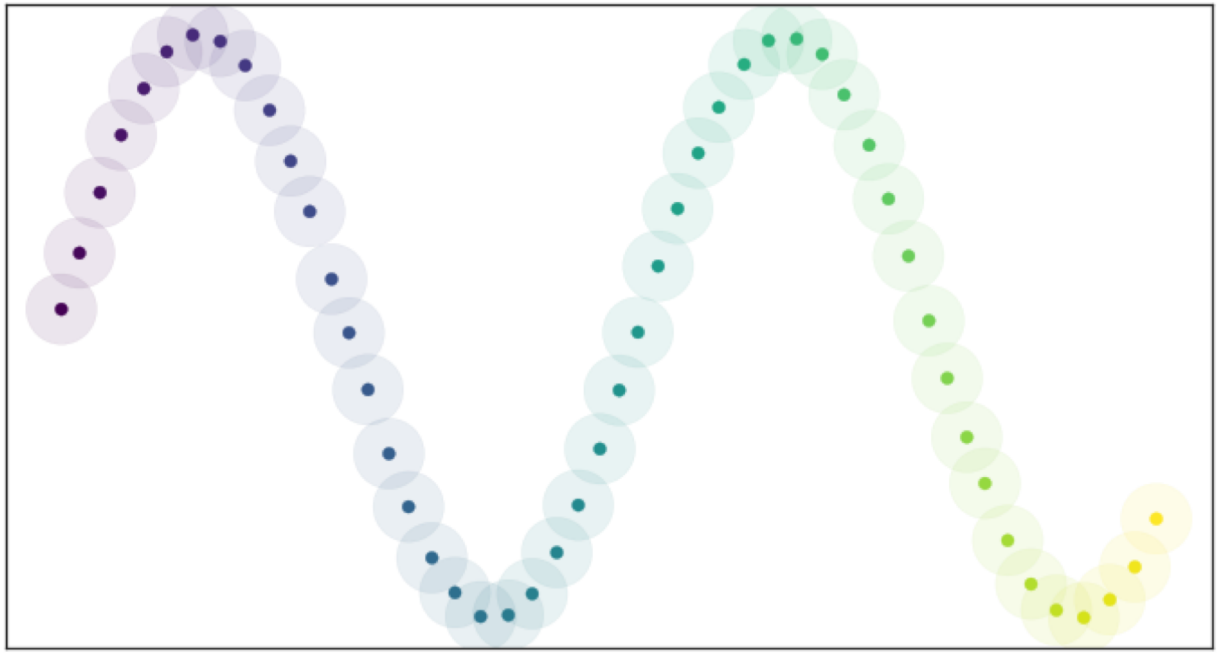
A toy example to construct the simplicial complex.

- The open balls are all the same size. (single unit ball)



$\Leftrightarrow$

equiv.



In the Euclidean space

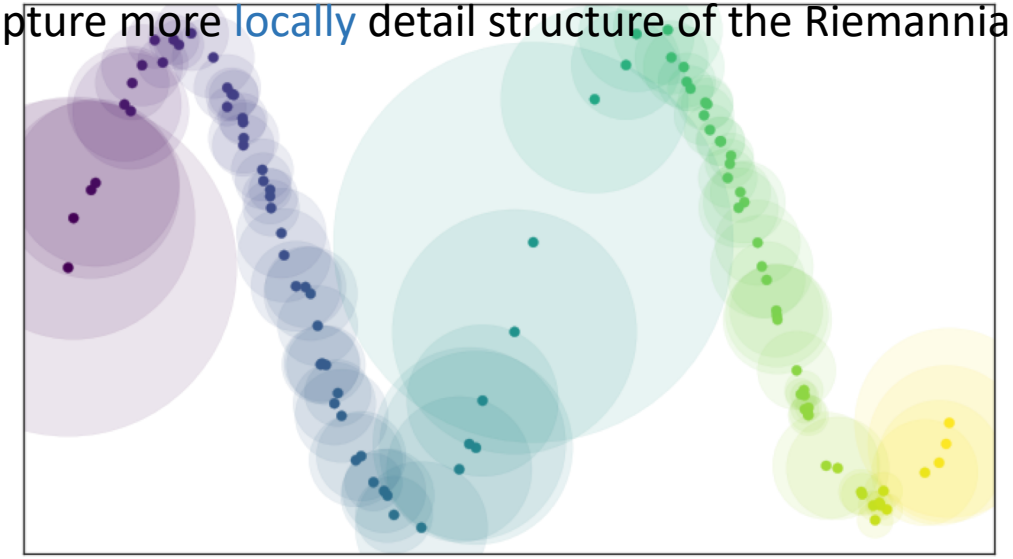
In the Manifold

radius one with a locally varying metric

Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

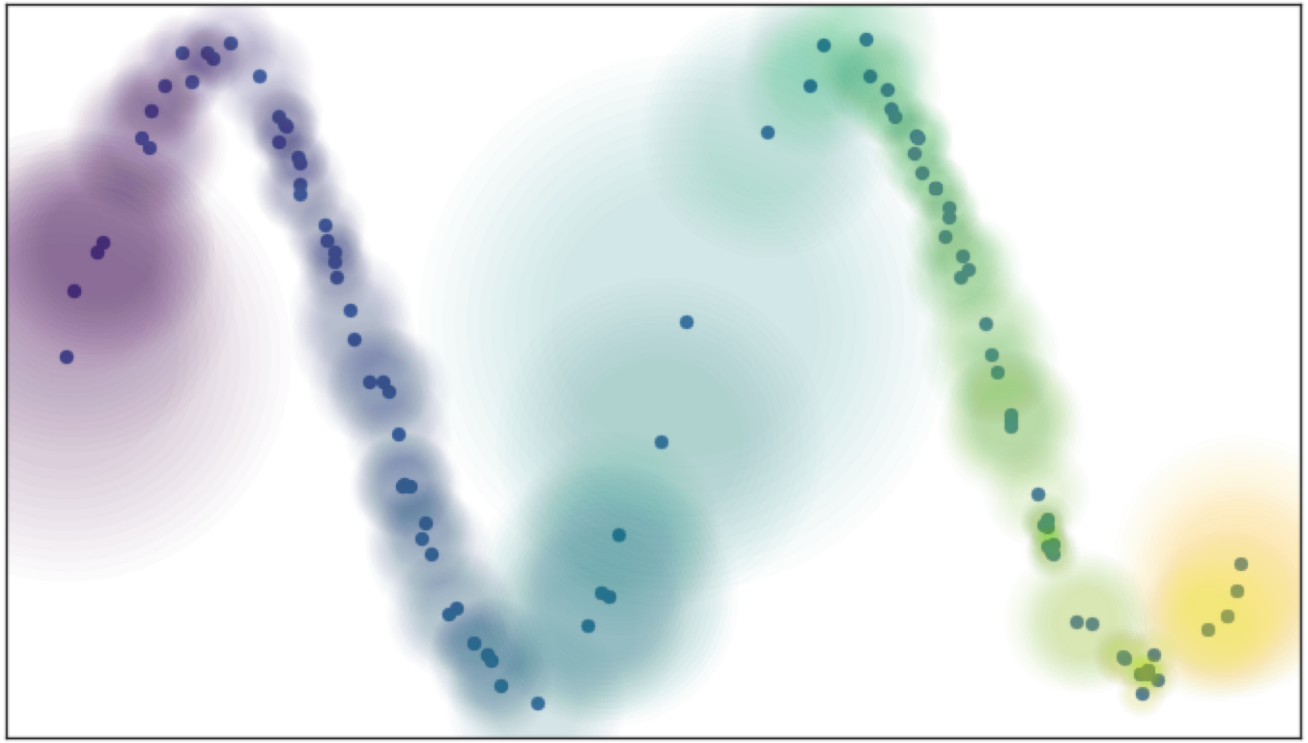
- UMAP provide a parameter *k*-neighbor to approximate the local sense of distance. (i.e. How locally we wish to estimate the Riemannian metric)
- Herein, in some sense, we use the neighboring distance to measure distance between points due to the “uniformly” distributed on the manifold. (Distances between points on the manifold are similar.)
  - k is a kind of sense to define φ_α & φ_β. (U_α and U_β are remain the same (single unit ball).)
  - If k is large → Capture more broadly accurate across the manifold as a whole.
  - If k is small → Capture more locally detail structure of the Riemannian metric.



Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

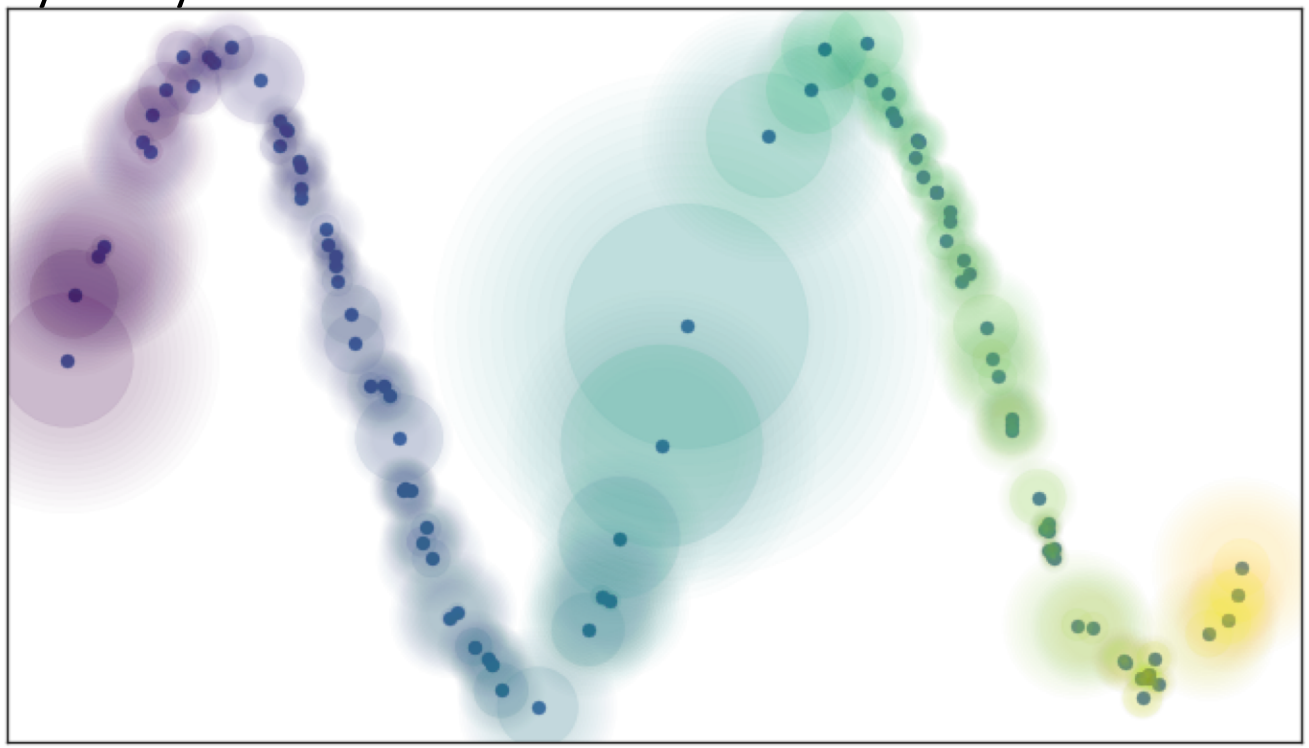
- Why choose a fixed radius?
  - Fixed radius: We cannot know how much a point in the open set of the point is.
- Various radius in our confidence: Fuzzy cover.



Faded confidence

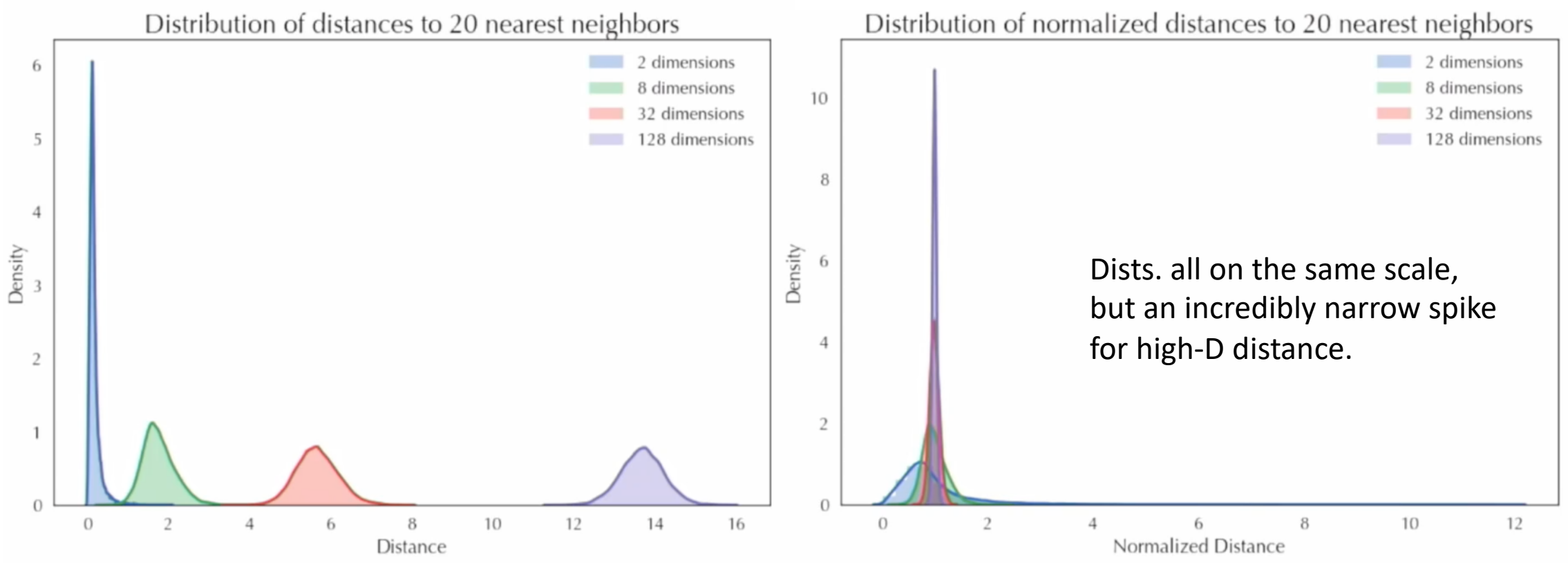
Uniform Manifold Approximation and Projection (UMAP)

- Some assumptions and definitions
  - Assumptions: The manifold is locally connected.
    → Isolated point is not allowed in the manifold.
    → Be completely confident in their distance out to the first nearest neighbor, and get fuzzy decay from there.



Uniform Manifold Approximation and Projection (UMAP)

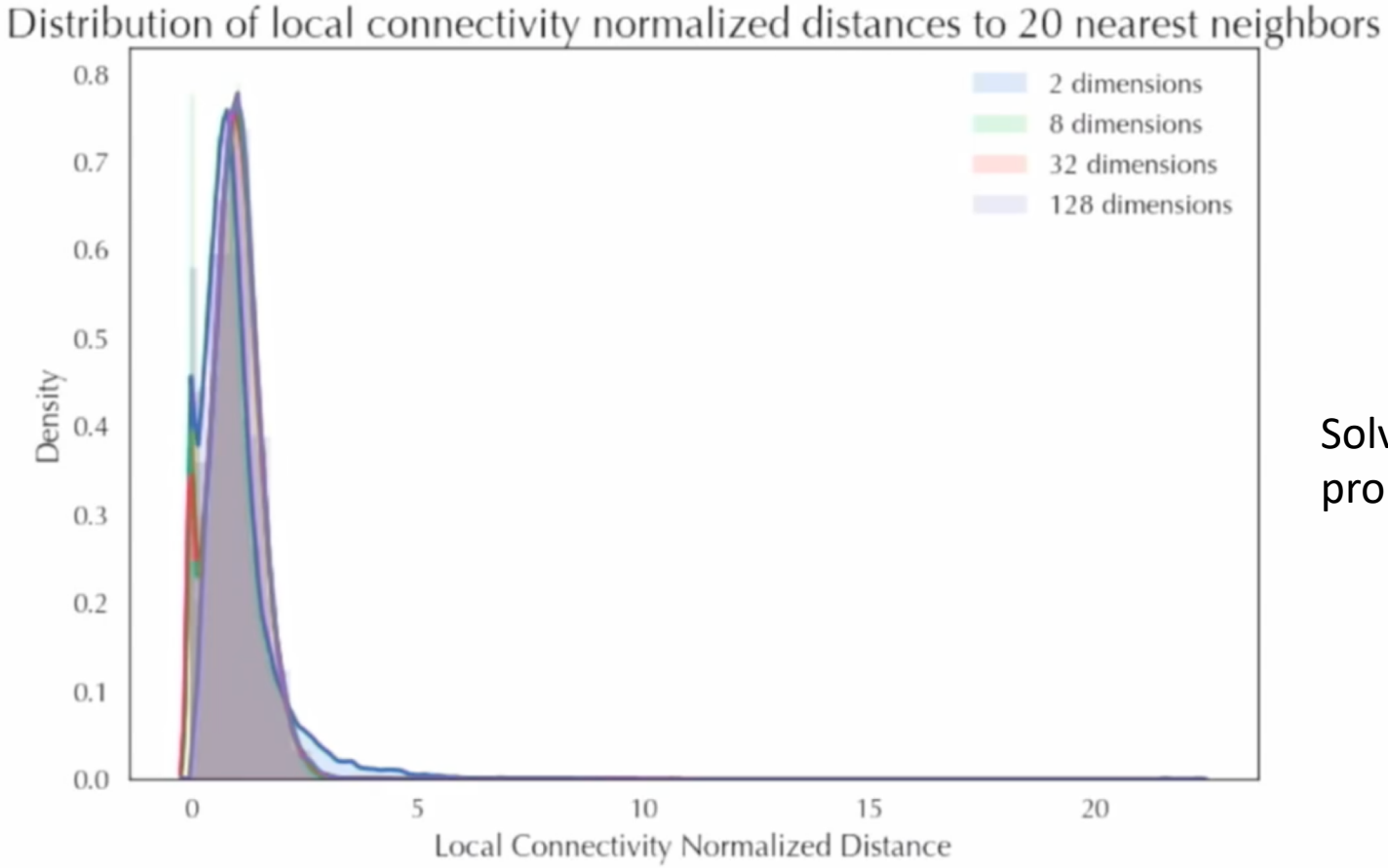
- The curse of the dimensionality (Before the local connectivity assumption)
  - Data points are sampled randomly on a normal distribution in varying dimension spaces.
  - Ideally, the dist. of distance should be similar after normalization (divided by the largest distance).



  - In high-D data, distances tend to be larger, but also more similar to one another. → tighter dist.

Uniform Manifold Approximation and Projection (UMAP)

- The curse of the dimensionality (After the local connectivity assumption)
  - Data is normalized after using local connectivity.
  - Distributions for different kinds of dimension are similar.

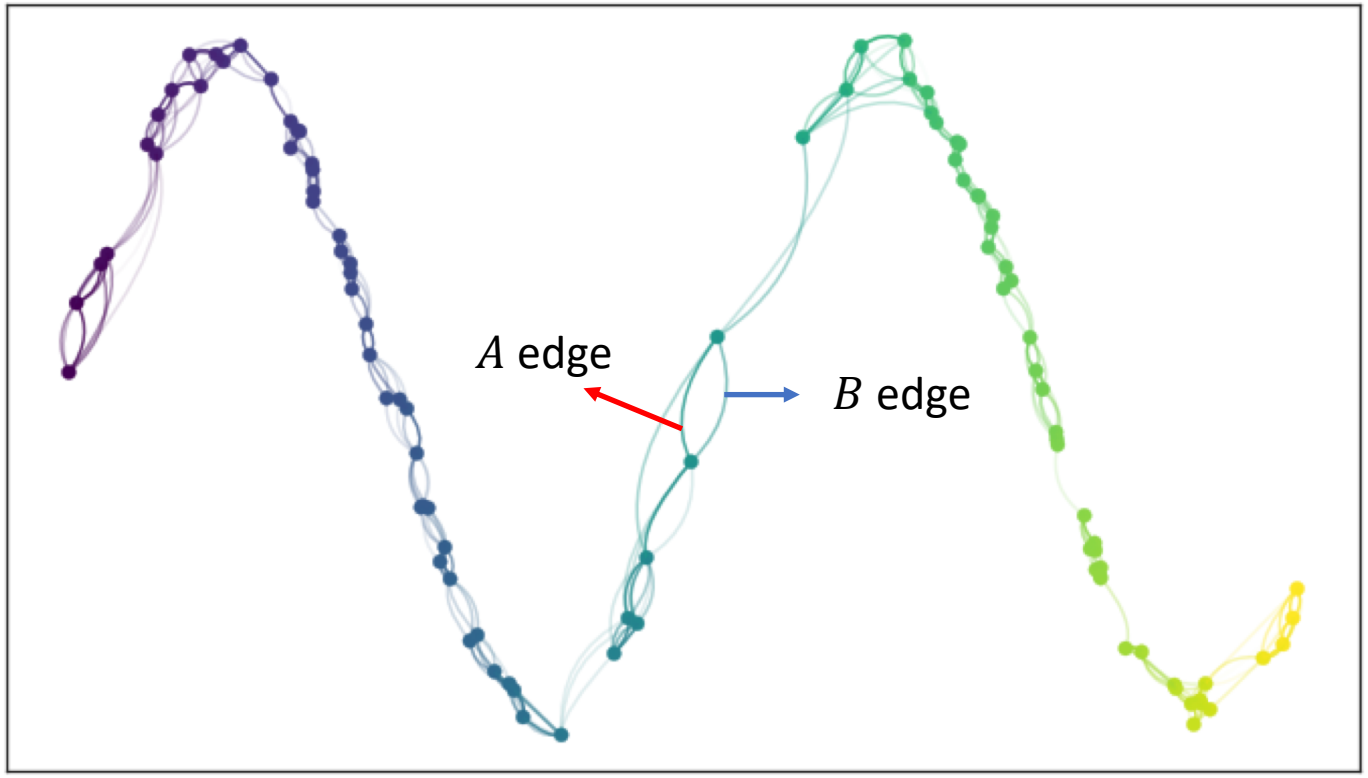


Solve the dimensionality problem.

Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

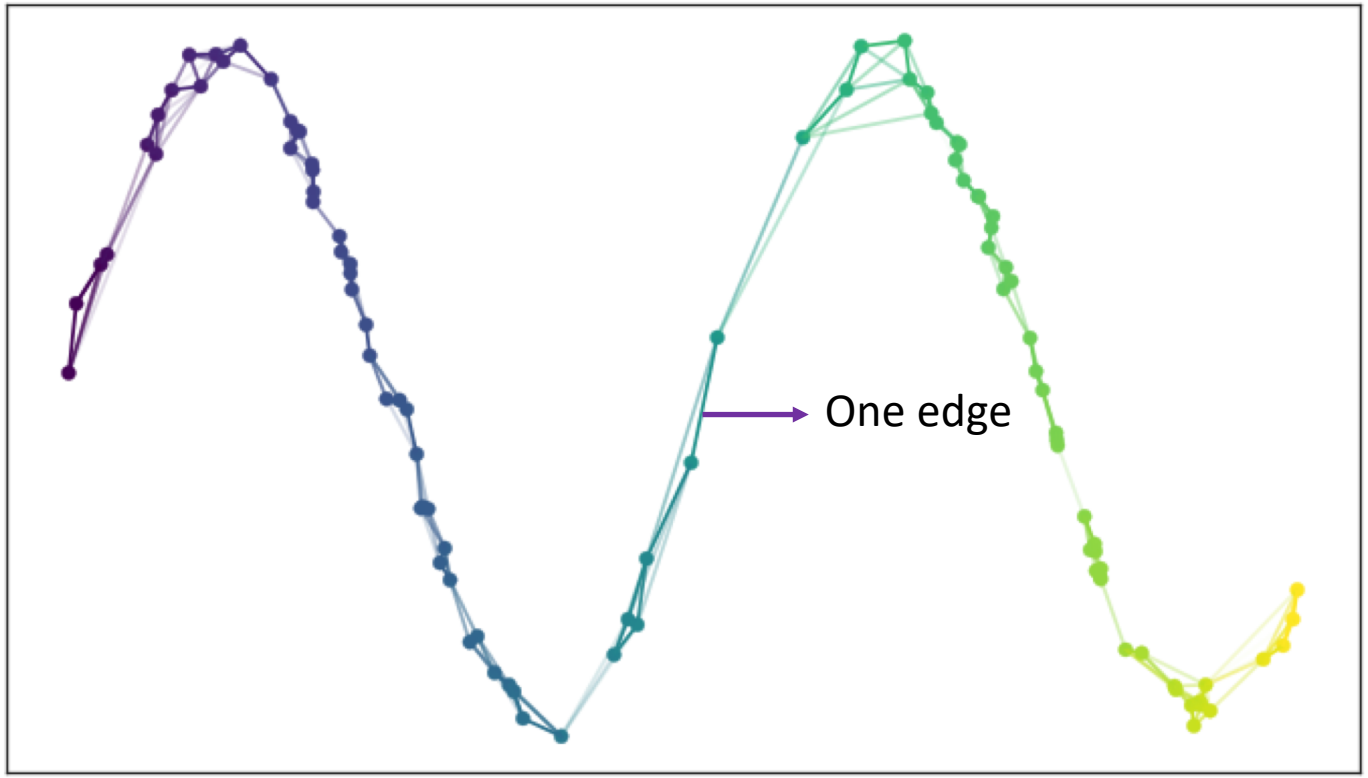
- Compatible local metrics
  - Incompatible local metrics ($\tau_{\alpha,\beta}$, $\tau_{\beta,\alpha}$): Multiple edges between any pair of points with different weights on them (e.g. A edge and B edge).
  - Weights: A similar concept for measuring the probability that the edge exists.



Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

- Compatible local metrics
  - Taking a fuzzy union: $\mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x)$
  - Combined weights: A similar concept for measuring the probability that either one of the edges exists. (either A edge exists or B edge exists.)



Uniform Manifold Approximation and Projection (UMAP)

- In low-Dimension mapping
  - Apply the same process to get a fuzzy graph.
  - Low-Dimension mapping manifold: e.g. $\mathbb{R}^2$ = Euclidean space
    → parameter: *d*. (The target embedding dimension.)
  - We cannot know the "correct" nearest neighbor distance for the local connectivity.
    → parameter: *min-dist*. (Aesthetic parameter.) → In some sense to define φ_α & φ_β for low-D mapping.

- Measure the difference between low-D and manifold:
  - $\mu(a)$: Weights (degree of at least an edge exists) in high dimension (manifold).
  - $v(a)$: Weights (degree of at least an edge exists) in low dimension.
  - Cross Entropy (definition)

$$\sum_{a \in A} \mu(a) \log\left(\frac{\mu(a)}{v(a)}\right) + \big(1-\mu(a)\big) \log\left(\frac{1-\mu(a)}{1-v(a)}\right)$$

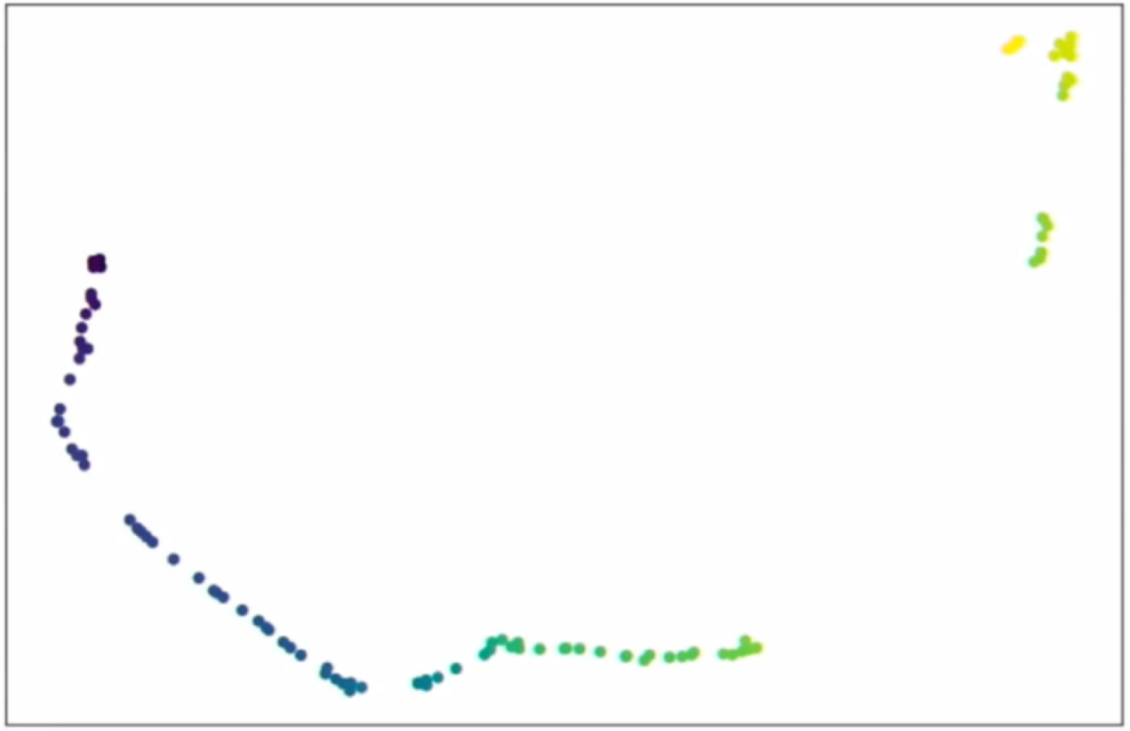
Get the clumps (Attraction)

Get the gaps (Repulsion)

Uniform Manifold Approximation and Projection (UMAP)

A toy example to construct the simplicial complex.

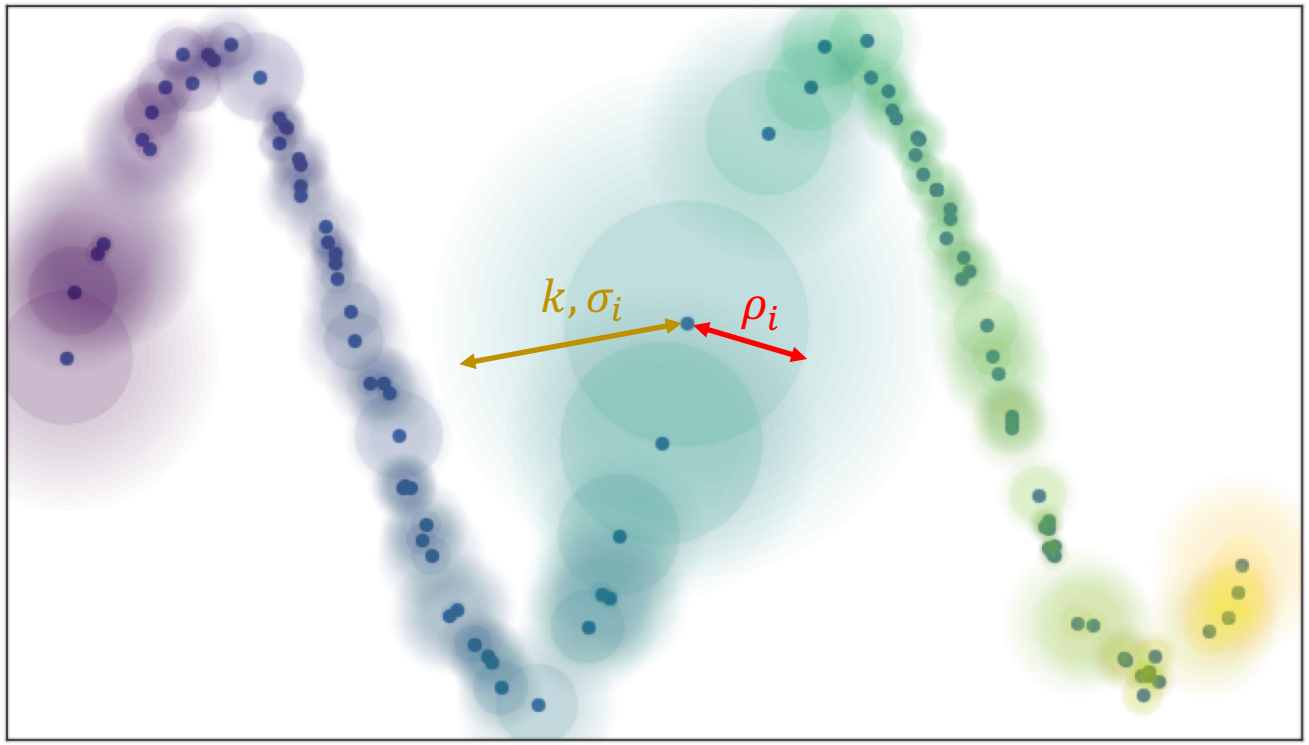
- Suppose we embed data into $\mathbb{R}^2$. i.e. From $\mathbb{R}^2 \rightarrow \mathbb{R}^2$.
- Final result:
  - There is a big gap in this embedded data.
  - More data in high-D can cover more original manifold space (more completed simplex representation) → Solve the gap problem.



Uniform Manifold Approximation and Projection (UMAP)

A computation view of UMAP

- In the High-D space (manifold):

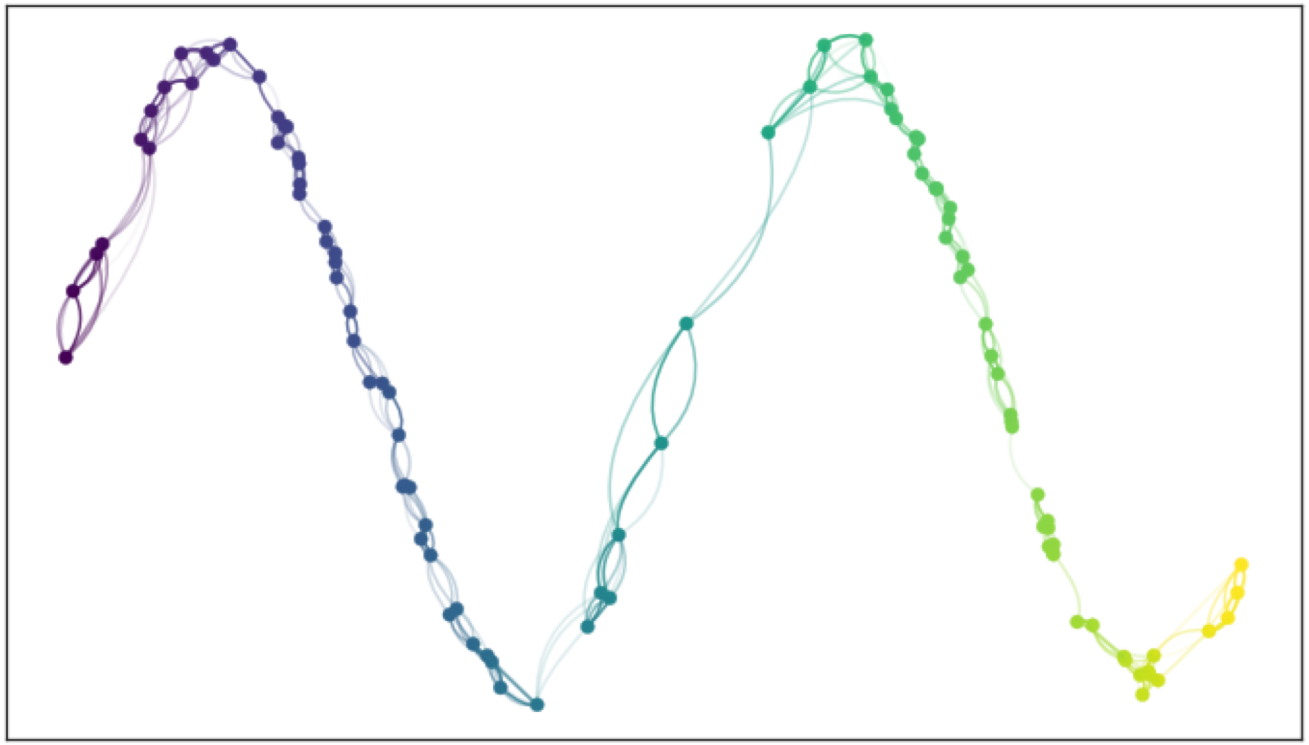


- Let $X = \{x_1, \dots, x_N\}$, be the input dataset.
- Let d be a metric, $d: X \times X \to \mathbb{R}_{\geq 0}$.
- Let the set $\{x_{i_1}, \dots, x_{i_k}\}$ be the k nearest neighbors of x_i under the metric d.
- $\rho_i = \min\left\{d\left(x_i, x_{i_j}\right) | 1 \leq j \leq k, d\left(x_i, x_{i_j}\right) > 0\right\}$
- Set σ_i satisfy $\sum_{j=1}^{k} \exp\left\{-\frac{\max\left(0, d\left(x_i, x_{i_j}\right) - \rho_i\right)}{\sigma_i}\right\} = \log_2(k)$, k is defined by user.

Uniform Manifold Approximation and Projection (UMAP)

A computation view of UMAP

- In the High-D space (manifold):

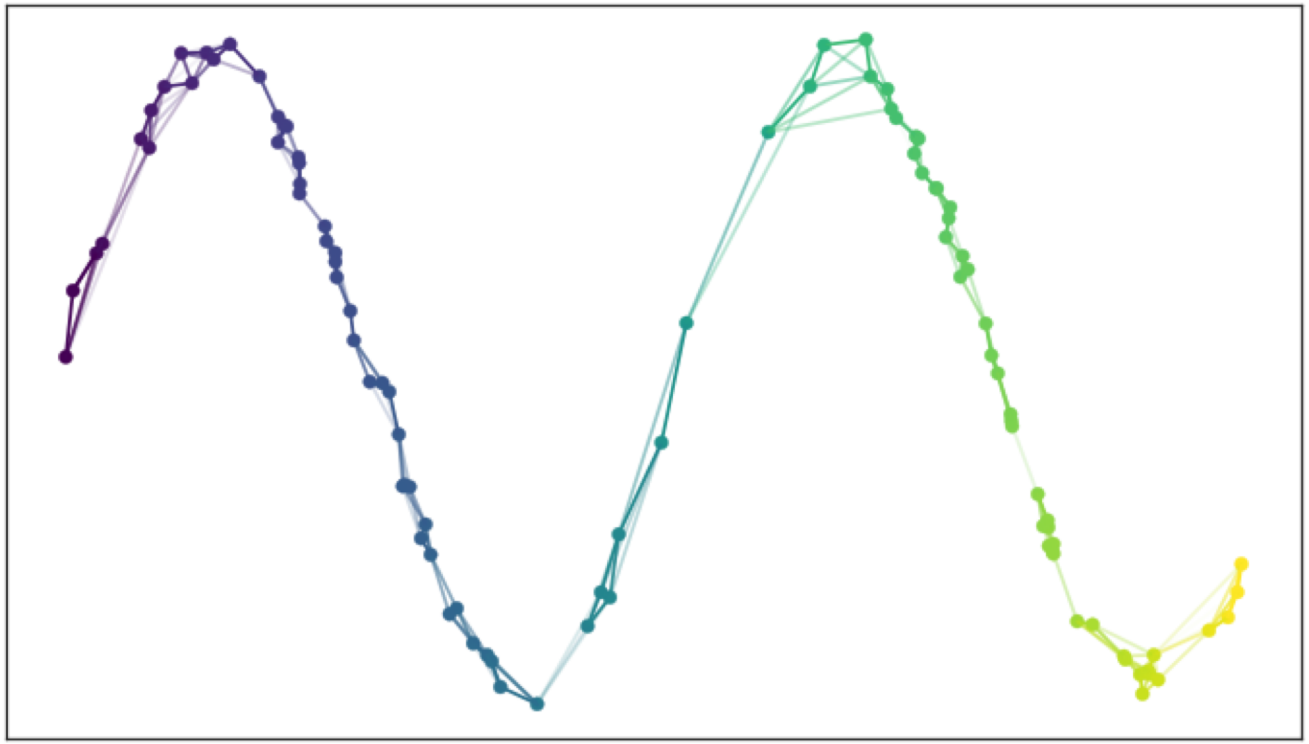


- Let $E = \left\{\left(x_i, x_{i_j}\right) | 1 \le j \le k, 1 \le i \le N\right\}$ be the set of directed edges.
- Define the weight function $w\left(\left(x_i, x_{i_j}\right)\right) = \exp\left\{-\frac{\max\left(0, d\left(x_i, x_{i_j}\right) - \rho_i\right)}{\sigma_i}\right\}$ (p_{ij} in the t-SNE).
- Define the weight matrix $A = [A_{ls}],\ where\ A_{ls} = \begin{cases} w\left(\left(x_i, x_{i_j}\right)\right), l = i, s = i_j \\ 0 \quad\quad , otherwise \end{cases}$

Uniform Manifold Approximation and Projection (UMAP)

A computation view of UMAP

- In the High-D space (manifold):



- Consider the symmetric matrix $B = A + A^T - A \circ A^T$, where $\circ$ is the Hadamard (or pointwise) product.
- In a meaningful way, the element in B represents $\mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x)$
- B is also the undirected weighted matrix.

Uniform Manifold Approximation and Projection (UMAP)

A computation view of UMAP

- In the Low-D mapping (Euclidean space):
  - The family of curves of the form (Student t-family):

$$v_{ij} = \frac{1}{1 + a\left\|y_i - y_j\right\|_2^{2b}}$$

  - When $a = 1, b = 1 \rightarrow$ Student t-kernel.
  - a and b can be specified by user or estimated by non-linear least squares fitting with *min-dist.*
  - The same concept of q_{ij} in the t-SNE.
  - Property of function in this family: Nicely differentiable functions.

- Cost function: Cross entropy for fuzzy set.
  - Define $w_{ij} = w\left(\left(x_i, x_{i_j}\right)\right)$
  - Consider the membership strength to be the probability (parameter: w_{ij}, statistics: v_{ij}) used in the parameter of a Bernoulli dist. → Find MLE ⇔ Minimize the cross entropy.

$$C = \sum_i \sum_{i \neq j} w_{ij} \log\left(\frac{w_{ij}}{v_{ij}}\right) + \left(1 - w_{ij}\right) \log\left(\frac{1 - w_{ij}}{1 - v_{ij}}\right)$$

Uniform Manifold Approximation and Projection (UMAP)

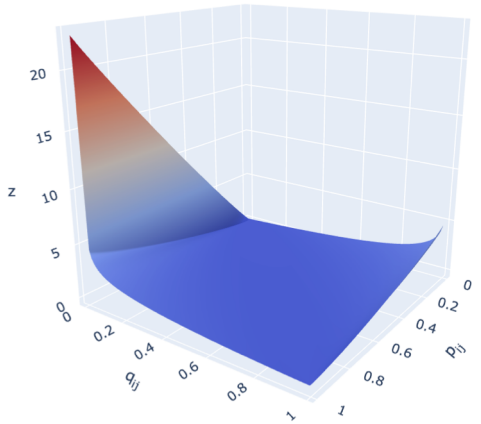
- Gradient of Cross entropy
  - Let $d_{ij} = \|y_i - y_j\|_2$

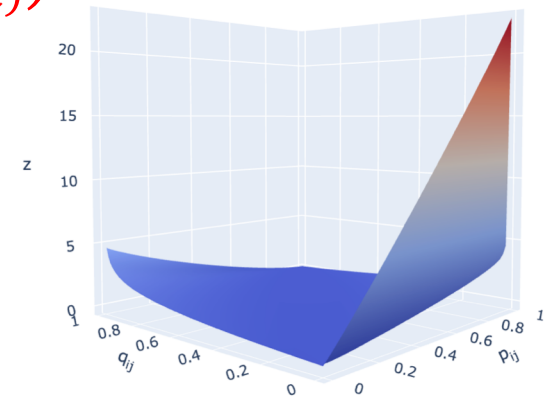
$$v_{ij} = \frac{1}{\left(1 + ad_{ij}^{2b}\right)}, \qquad 1 - v_{ij} = \frac{ad_{ij}^{2b}}{\left(1 + ad_{ij}^{2b}\right)}, \qquad \frac{\partial v_{ij}}{\partial d_{ij}} = -\frac{2abd_{ij}^{2b-1}}{\left(1 + ad_{ij}^{2b}\right)^2}$$

$$\begin{aligned}
\frac{\partial C}{\partial y_i} &= \sum_j \left[-\frac{w_{ij}}{v_{ij}} \cdot \frac{\partial v_{ij}}{\partial d_{ij}} + \frac{1 - w_{ij}}{1 - v_{ij}} \cdot \frac{\partial v_{ij}}{\partial d_{ij}} \right] \frac{\partial d_{ij}}{\partial y_i} \\
&= \sum_j \left[\left(-w_{ij}\left(1 + ad_{ij}^{2b}\right) + \frac{\left(1 - w_{ij}\right)\left(1 + ad_{ij}^{2b}\right)}{ad_{ij}^{2b}} \right) \cdot \frac{\partial v_{ij}}{\partial d_{ij}} \right] \frac{\partial d_{ij}}{\partial y_i} \\
&= \sum_j \frac{2abd_{ij}^{2(b-1)}}{1 + ad_{ij}^{2b}} \cdot w_{ij} \cdot d_{ij} \cdot \frac{\left(y_i - y_j\right)}{d_{ij}} \qquad \text{Attractive force} \\
&\quad - \frac{2b}{d_{ij}^2\left(1 + ad_{ij}^{2b}\right)} \cdot \left(1 - w_{ij}\right) \cdot d_{ij} \cdot \frac{\left(y_i - y_j\right)}{d_{ij}} \qquad \text{Repulsive force}
\end{aligned}$$

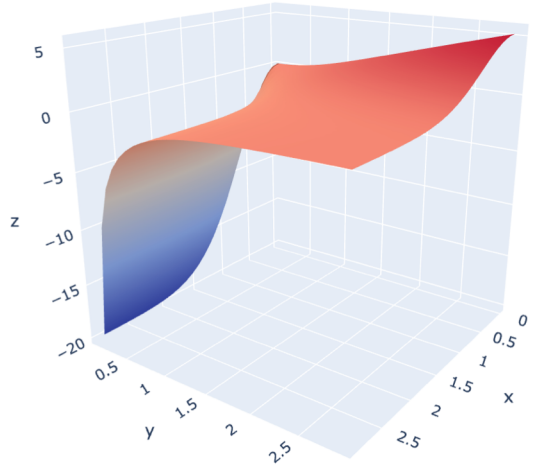
Uniform Manifold Approximation and Projection (UMAP)

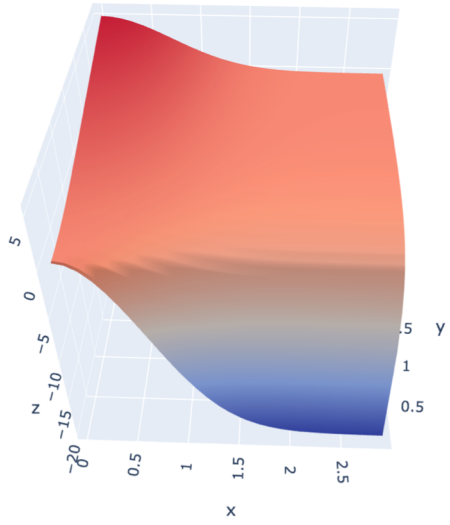
- Cost function: cross entropy $\sum_{a\in A} \mu(a) \log\left(\frac{\mu(a)}{v(a)}\right) + \left(1-\mu(a)\right) \log\left(\frac{1-\mu(a)}{1-v(a)}\right)$





- Gradient of cross entropy (After replace with exponential term and Student t-kernel.)





Uniform Manifold Approximation and Projection (UMAP)

- Implementation
  - Find the nearest neighbors efficiently → Random Projection trees + Nearest-Neighbor-Descent
  - Optimization improvement → Stochastic Gradient Descent + Negative Sampling, Probabilistic Edge Sampling

    ⟹ parameter: *n-epochs.* (The number of training epochs to use.)
  - Initialization for the low-D: Eigenvectors of the normalized Laplacian, Spectral layout.

    ⟹ faster convergence & greater stability

- Performance
  - High performance
  - Clean code
  - Custom distance metrics

Uniform Manifold Approximation and Projection (UMAP)

- Weaknesses
  - Interpretability.
    - Unlike PCA: a. Has factor loading. b. Embedding space has special meaning (greatest variance).
  - Small sample size may not work well.
    - Tend to find the spurious manifold structure from the noise of data.
    - Unstable due to the use of approximation techniques such as nearest neighbor algorithms and negative sampling. (occurred when sample size < 500)
  - Unsuitable if the global structure is of primary interest.
    - Similar to t-SNE and LargeVis, UMAP is derived from the axiom that local distance is more important than long range distance.

- Generalization
  - Modeled on the training data, use the model on the test data.
  - Be used as a generative model for original high dimensional space from embedding space.
  - (Semi-)Supervised version.
  - Combine different spaces with different metrics together and embedded them into one vector space. (Different metric for different datatype, Categorical: Jaccard or Dice, Ordinal: Manhattan.)

Uniform Manifold Approximation and Projection (UMAP)

Brief Summary

- UMAP has been used in bioinformatics, materials science, machine learning, etc.

- UMAP uses local manifold approximations and patches together their local fuzzy simplicial set representation to construct a topological representation in high-D data.

- A similar graph construction process can be used to construct an equivalent topological representation for low-D data.

- Optimize the layout of the data representation in the low dimensional space by minimizing the cross-entropy between the two topological representations.

Uniform Manifold Approximation and Projection (UMAP)

Brief Summary

- There are four mainly important parameters that can be adjusted by user: *k*-neighbor, *min-dist,* embedding dimension *d, n-epochs.*

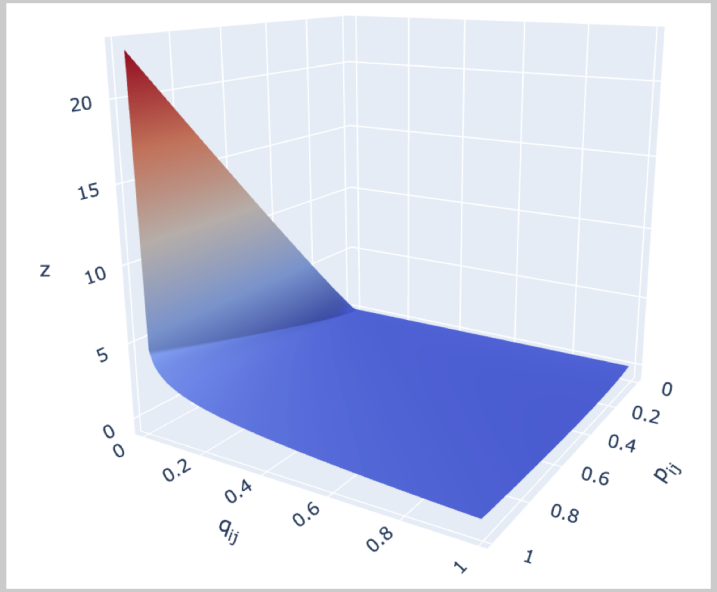
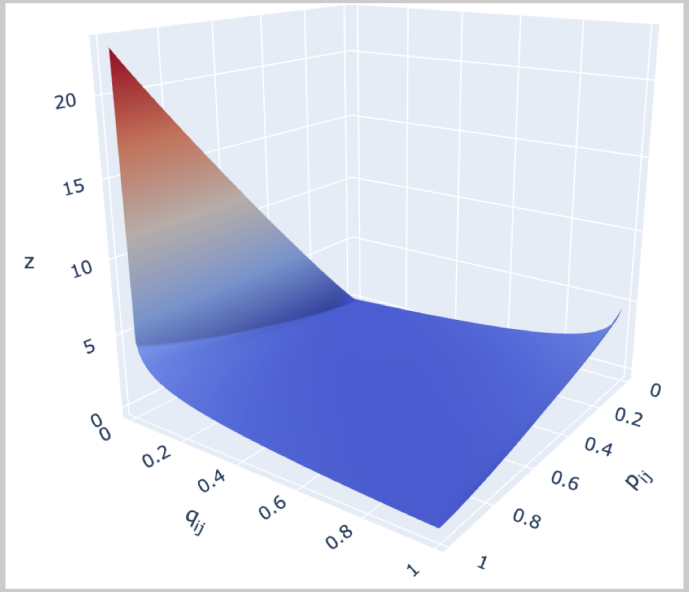
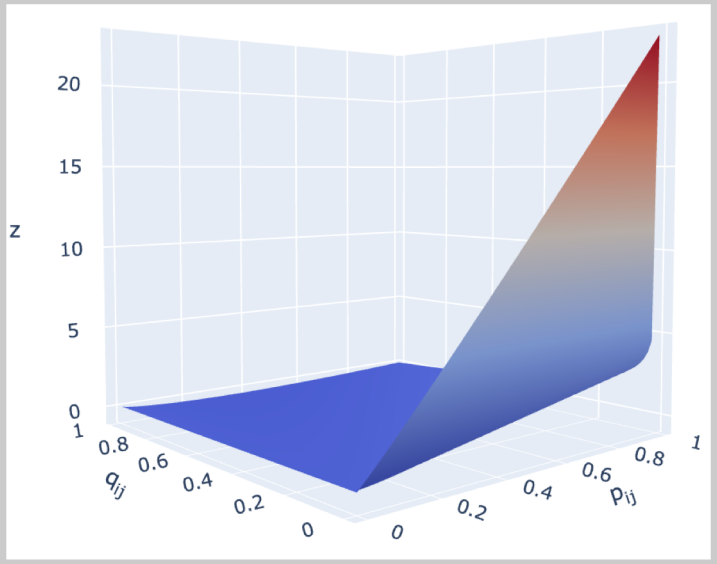
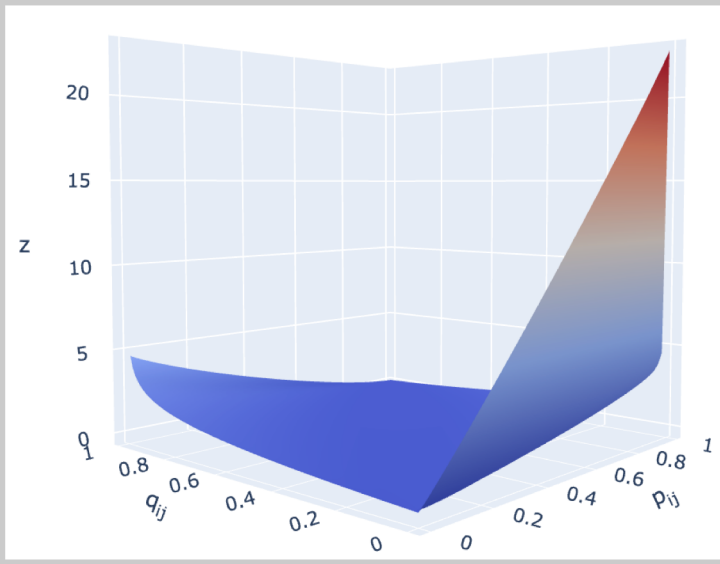
- A comprehensible concept for *k*-neighbor: Network, Polymers (Thermoplastic & Thermosetting)
- Large *k*-neighbor value captures large scale manifold structures, but loss detailed structure which will get averaged out in the local approximation.
- Smaller *k*-neighbor value captures detailed manifold structure, but tends to be broken into many small connected components.

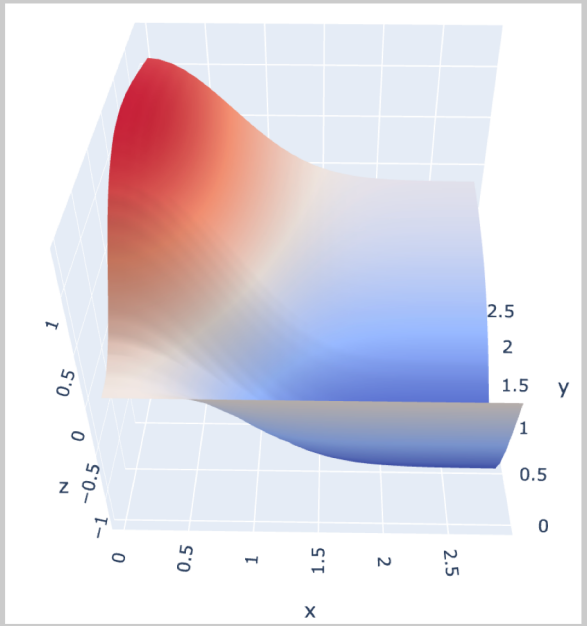
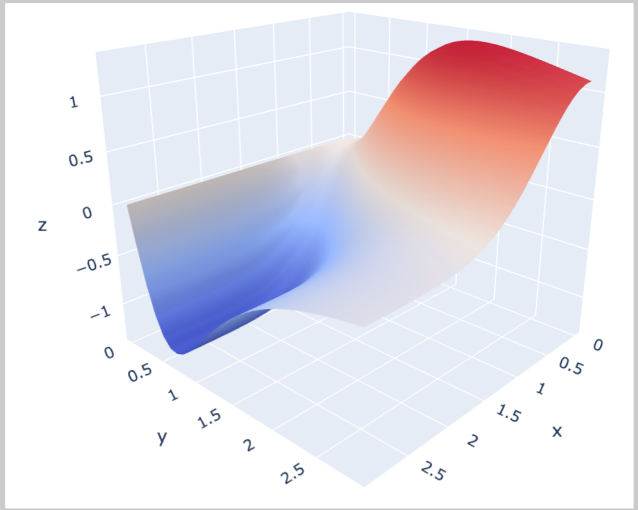
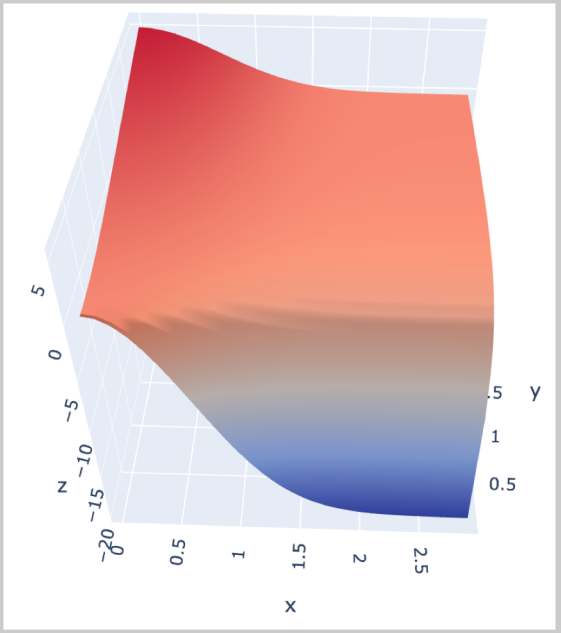
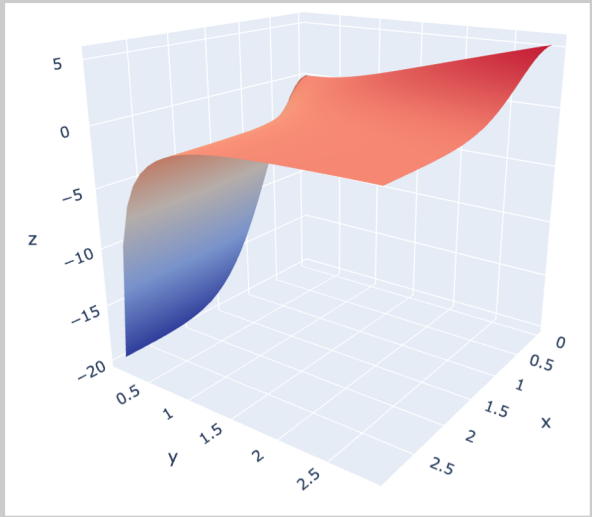
- Larger *min-dist* forces the embedding to spread points out more, and assist visualization. (Avoid potential overplotting issues)
- Smaller *min-dist* results in potentially densely packed regions, but will likely more faithfully represent the manifold structure.

Comparison and Evaluation

Comparison

Items	t-SNE	UMAP
Standpoint	High-dimensional Euclidean space	Topological Riemannian space (weighted graph)
Measurement for the distance between data points	Similarity, Probability	Membership strength (Fuzzy)
Kernel used for the original data	$\exp(\cdot)$ (Gaussian)	$\exp(\cdot)$
Kernel used for the reconstructed data	Student t-kernel	Student t-“family” kernel
Cost function	KL Divergence	Cross Entropy
Initialization for the optimization	Random	Graph Laplacian (Eigenvectors of Laplacian)
Optimization	Momentum gradient descent	Stochastic gradient descent

Items	t-SNE	UMAP
Cost function		
		

Items	t-SNE	UMAP
Gradient of the cost function	 	 

Evaluation

- https://www.youtube.com/watch?v=4NlvatkpV3s&feature=emb_logo

Evaluation

- From the author

	t-SNE	UMAP
COIL20	20 seconds	7 seconds
MNIST	22 minutes	98 seconds
Fashion MNIST	15 minutes	78 seconds
GoogleNews	4.5 hours	14 minutes

	UMAP speed up over t-SNE
COIL20	3x
MNIST	13x
Fashion MNIST	11x
GoogleNews	19x

- Google News dataset contains 200000 word vectors.
- UMAP preserves more global structure than t-SNE.
- Compare to t-SNE, UMAP has superior run time performance (on larger sample size, higher dimension of original dataset and higher dimension of embedding space.)

Evaluation

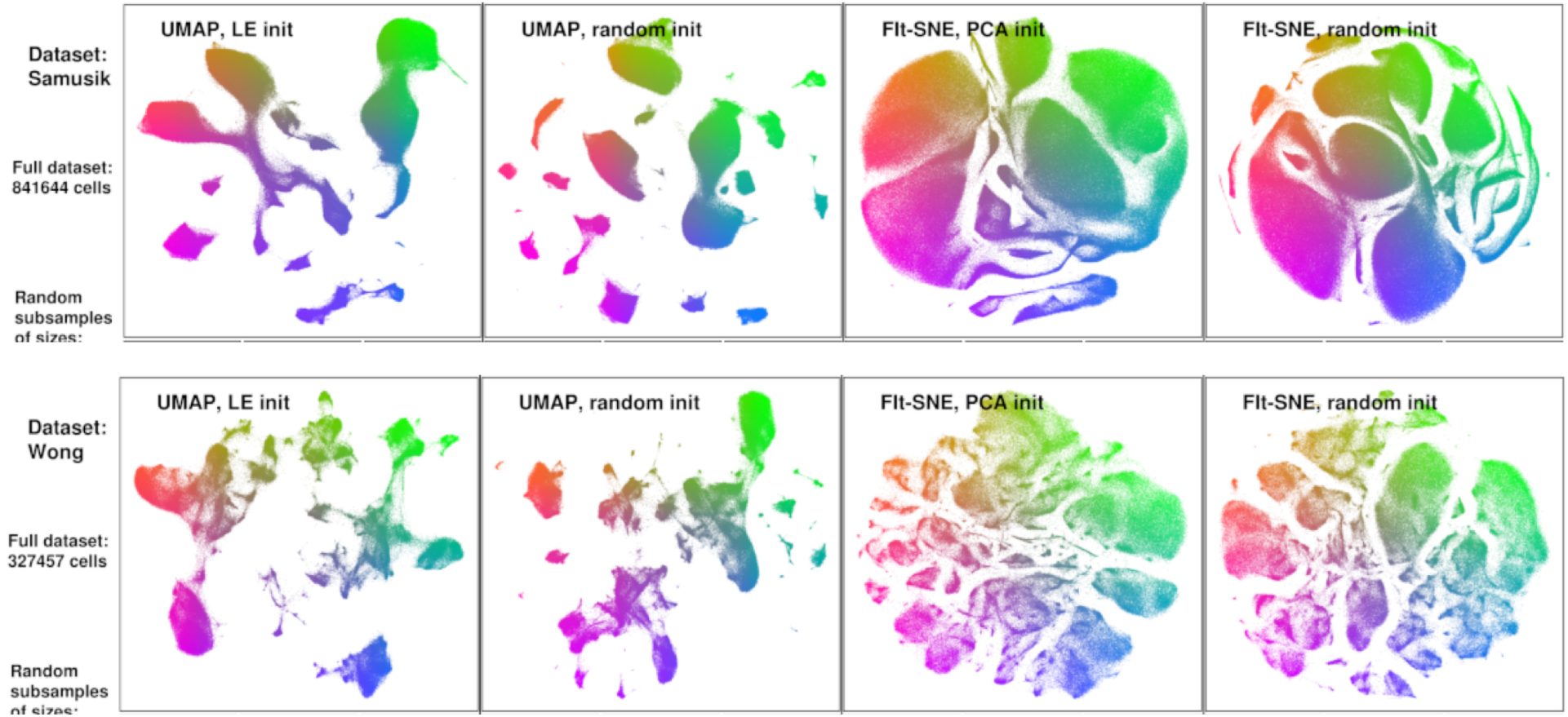
- Google AI team:

  https://pair-code.github.io/understanding-umap/?fbclid=IwAR2e4W-VnEVhjnmLi2HFHolEt-UFx_z-Mmkox1cKIu4us_SrN4dvURucmXo

  - A real world embedded example: Mammoth fossil example → UMAP indeed provide a better global structure preservation especially when the *perplexity* and *k*-neighbor are low. Furthermore, t-SNE can actually preserve global distances at large perplexity.
  - From the example of "clusters linked with a chain of points", it can be notice that the choice of hyperparameters depends on data. → Run different settings to get a better results.
  - At low *k*-neighbor value, spurious clustering can be observed.
  - In the most case, the performances of t-SNE & UMAP are very similar. However, in the example of "Containment – a dense, tight cluster inside of a wide, sparse cluster", t-SNE performs better then UMAP.
  - Cluster size and distance between clusters is not be meaningful for t-SNE and might not be meaningful for UMAP. (Since distance is warped.)

Myth broken?

- UMAP does not preserve global structure any better than t-SNE when using the same initialization.



Kobak, D., & Linderman, G. C. (2019) bioRxiv

Defending Justice?

- It conclude that it could not confirm that the tSNE and UMAP dimension reduction are affected by different initialization scenarios (based on the scRNAseq data).
  - https://towardsdatascience.com/why-umap-is-superior-over-tsne-faa039c28e99

- Other resources
  - https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17
  - https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668

Thank You!

Reference

- [pdf] Visualizing Data using t-SNE
- [pdf] Accelerating t-SNE using Tree-Based Algorithms
- [pdf] UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
- [pdf] UMAP does not preserve global structure any better than t-SNE when using the same initialization.

- t-SNE
  - https://www.youtube.com/watch?v=RJVL80Gg3lA
  - https://zhuanlan.zhihu.com/p/64664346
  - https://www.oreilly.com/content/an-illustrated-introduction-to-the-t-sne-algorithm/
  - https://math.stackexchange.com/questions/2227144/derivative-of-euclidean-norm-of-matrix-vector-product/2227237
  - https://medium.com/@layog/i-dont-understand-t-sne-part-1-50f507acd4f9
  - https://medium.com/@layog/i-do-not-understand-t-sne-part-2-b2f997d177e3
  - https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1
  - https://distill.pub/2016/misread-tsne/
  - https://www.microsoft.com/en-us/research/blog/optimizing-barnes-hut-t-sne/

Reference

- UMAP
  - https://umap-learn.readthedocs.io/en/latest/how_umap_works.html
  - https://www.youtube.com/watch?v=81HKNqruavc
  - https://www.youtube.com/watch?v=nq6iPZVUxZU&t=932s
  - https://www.youtube.com/watch?v=7pAVPjwBppo
  - https://www.youtube.com/watch?v=4NIvatkpV3s
  - https://en.wikipedia.org/wiki/Simplex
  - https://en.wikipedia.org/wiki/Simplicial_complex
  - https://en.wikipedia.org/wiki/Čech_complex
  - https://en.wikipedia.org/wiki/Connected_space
  - https://en.wikipedia.org/wiki/Locally_connected_space
  - https://en.wikipedia.org/wiki/Fuzzy_logic
  - https://en.wikipedia.org/wiki/Topological_manifold
  - https://mathworld.wolfram.com/LocallyConnected.html
  - https://sauln.github.io/mapper-presentation/#1
  - http://debussy.im.nuu.edu.tw/sjchen/Project_Courses/ML/Fuzzy.pdf
  - http://rportal.lib.ntnu.edu.tw/bitstream/20.500.12235/95760/2/n069475001302.pdf
  - https://web.math.sinica.edu.tw/math_media/d181/18102.pdf

Appendix

- Functor: A function between domains of discourse. E.g. a functor can transform the continuous topological space into the finite combinatorial object in a rigorous way.
- Adjunction: A near equivalence between domains of discourse. (the concept of having a pair of functors that go back and forth between two different domains of discourse. i.e. having a way of translating back and forth from one to the other and get an equivalence between two different domains of discourse)
- Limit: A solution to a system of constraints.
- Colimit: Gluing together a system of objects. E.g. Gluing together complex systems of objects into a single coherent whole object.