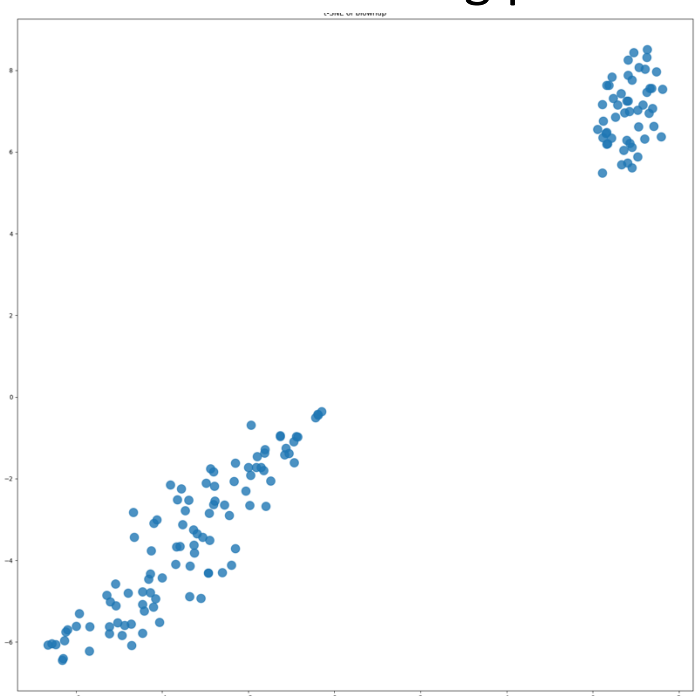
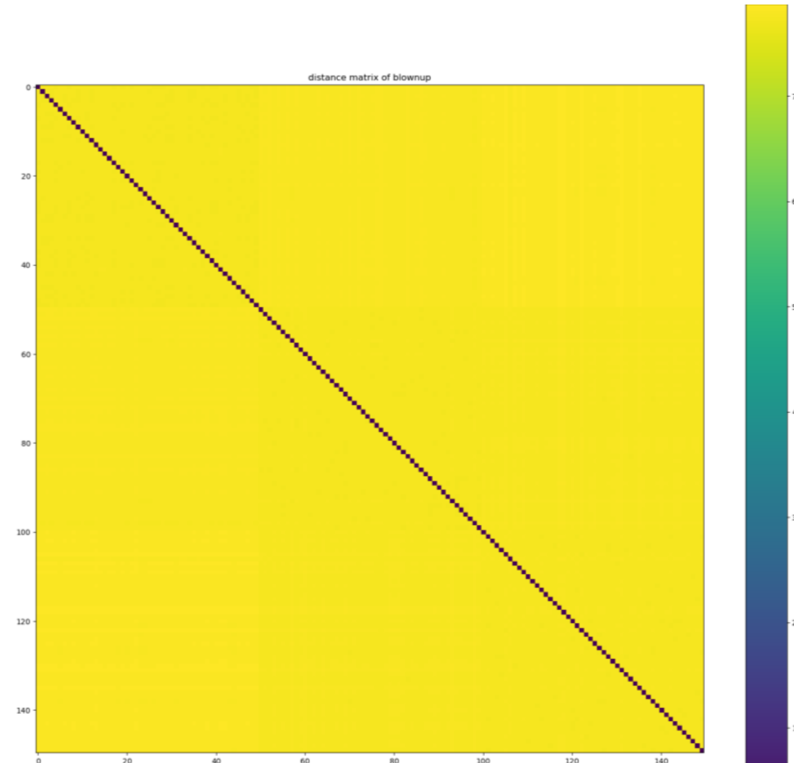
Perils of Using t-SNE (and friends)

Nakul Verma
Columbia University

Pop-Quiz!

Let's say you created a 2d t-SNE visualization of a dataset you collected and it produced the following plot.





Interpoint distance profile

Questions:

- What would you conclude about the clusters that may be present in your dataset?
- How confident are you about your conclusions?

Understanding the Visualizations

These critical questions require a white-box functional understanding of the visualization that was used (ie how exactly does t-SNE work).

let’s quickly review t-SNE and what is known
about its optimization criterion.

Stochastic Neighbor Embedding (SNE)

Goal: Find a low-dim. map that preserves the "local geometry"

local geometry = similarity between points in local neighborhoods

Idea:

Model the neighborhood structure/information as a probability distribution, then find a low-dimensional mapping that matches the same distribution!

Notation:

- $x_1,\ldots,x_n$ given high dim. data (given)
- $y_1,\ldots,y_n$ mapped low dim. Representation (to be learned)
- $p_{j|i}$ = probability of x_j being the neighbor of x_i (computed from data)
- $q_{j|i}$ = probability of y_j being the neighbor of y_i (to be matched to $p_{j|i}$)

Stochastic Neighbor Embedding

[Hinton and Roweis '03]

Stochastic Neighbor Embedding approach:

Probability model for high-dim input data

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\tau_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\tau_i^2)}$$

Meta parameter controlling the neighborhood size

Probability model for low-dim mapped data

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(-||y_i - y_k||^2)}$$

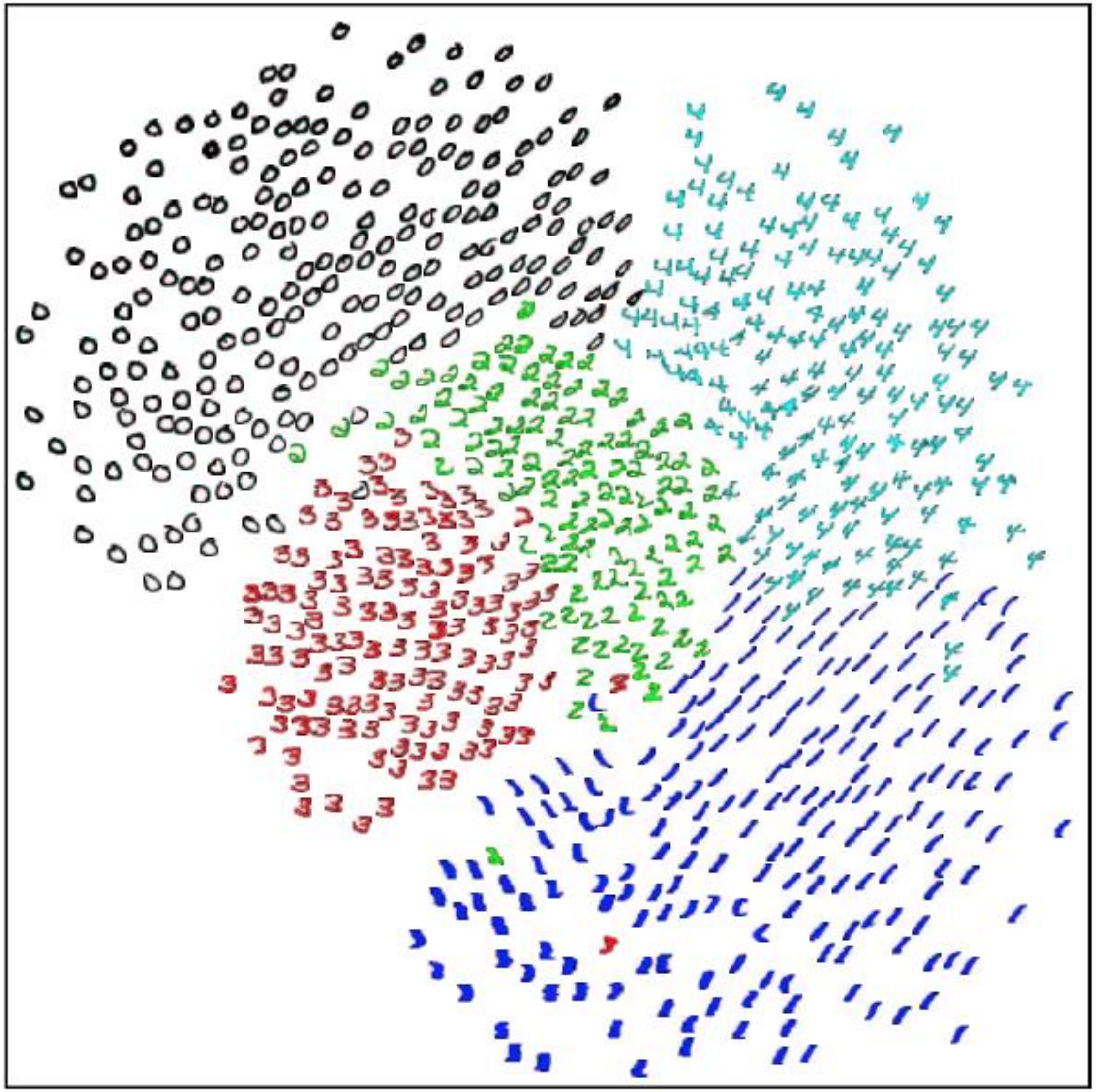
y's are the variables that need to be learned

Key optimization: Maximize the similarity between the distributions

minimize$_y$ $\sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$

Highly non-convex, just do gradient descent and settle with the local optimal solution

Stochastic Neighbor Embedding



The individual class clusters are well all together producing an effective visualization

But the clusters are NOT well separated

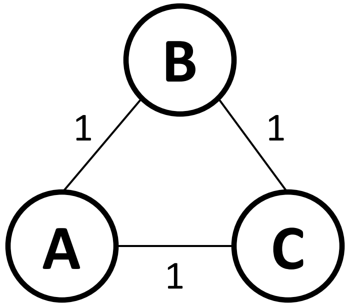
The issue: “crowding problem”

t-distributed Stochastic Neighbor Embedding

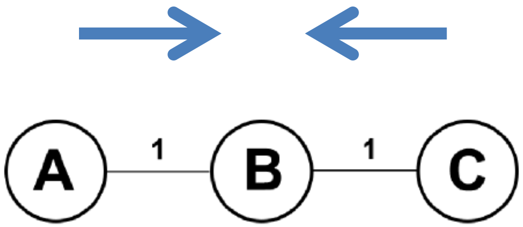
The crowding problem:

High dimensional data is being cramped into a low dimensional space. To match the probabilities, the clusters can "crowd" together

Consider three clusters A, B, C



Organization in high dimensions



Organization in low dimensions

Because of the gaussian-type neighborhood structure in low dimensions, large distance between A and C will be **penalized** a lot causing them to be mapped close (ie crowd) to each other

t-distributed Stochastic Neighbor Embedding

[Van der Maaten and Hinton '08]

Solution to the crowding problem

Idea: instead of using a thin-tailed Gaussian in the lower dimensions, we can use a heavier-tailed distribution, e.g. student's t-distribution!

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

Symmetrize the high dimensional neighborhood distribution

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

Use the heavier tailed student's t-distribution

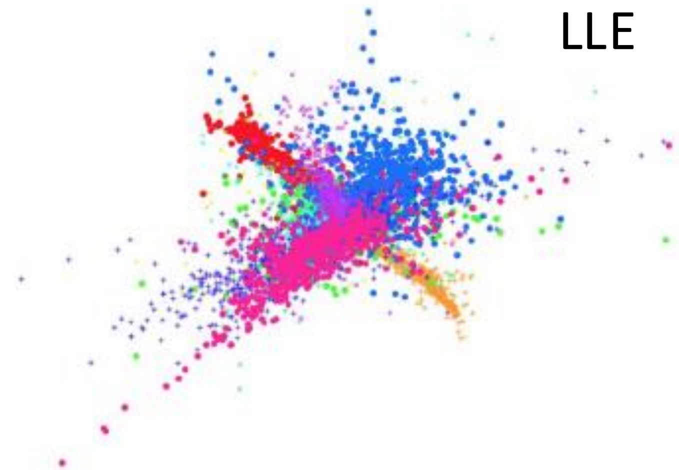
Final optimization:

$$\text{minimize}_y \quad \sum_i KL(P_i||Q_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

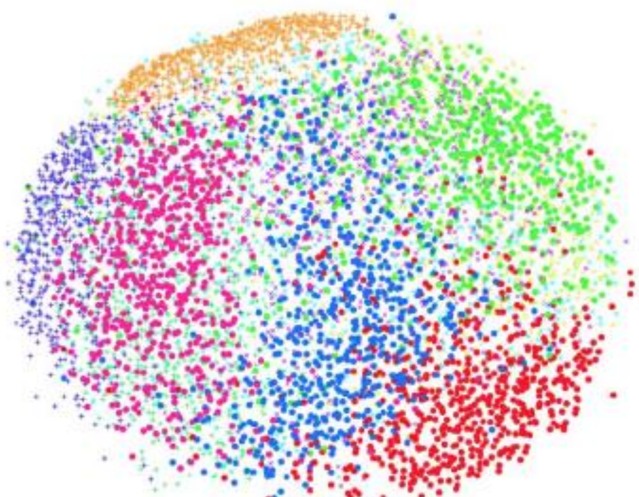
t-SNE on a Benchmark Dataset

PCA

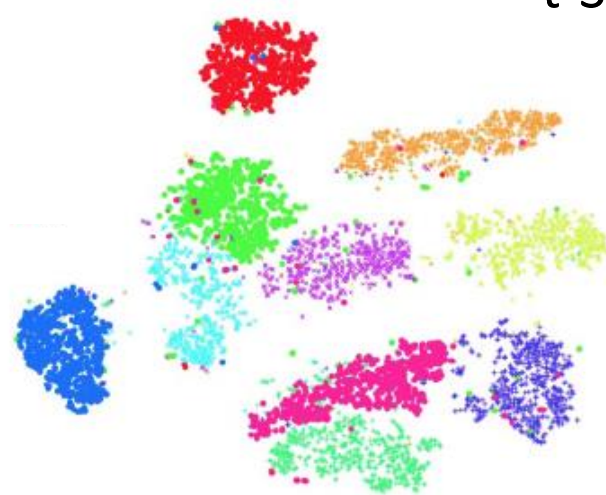
LLE



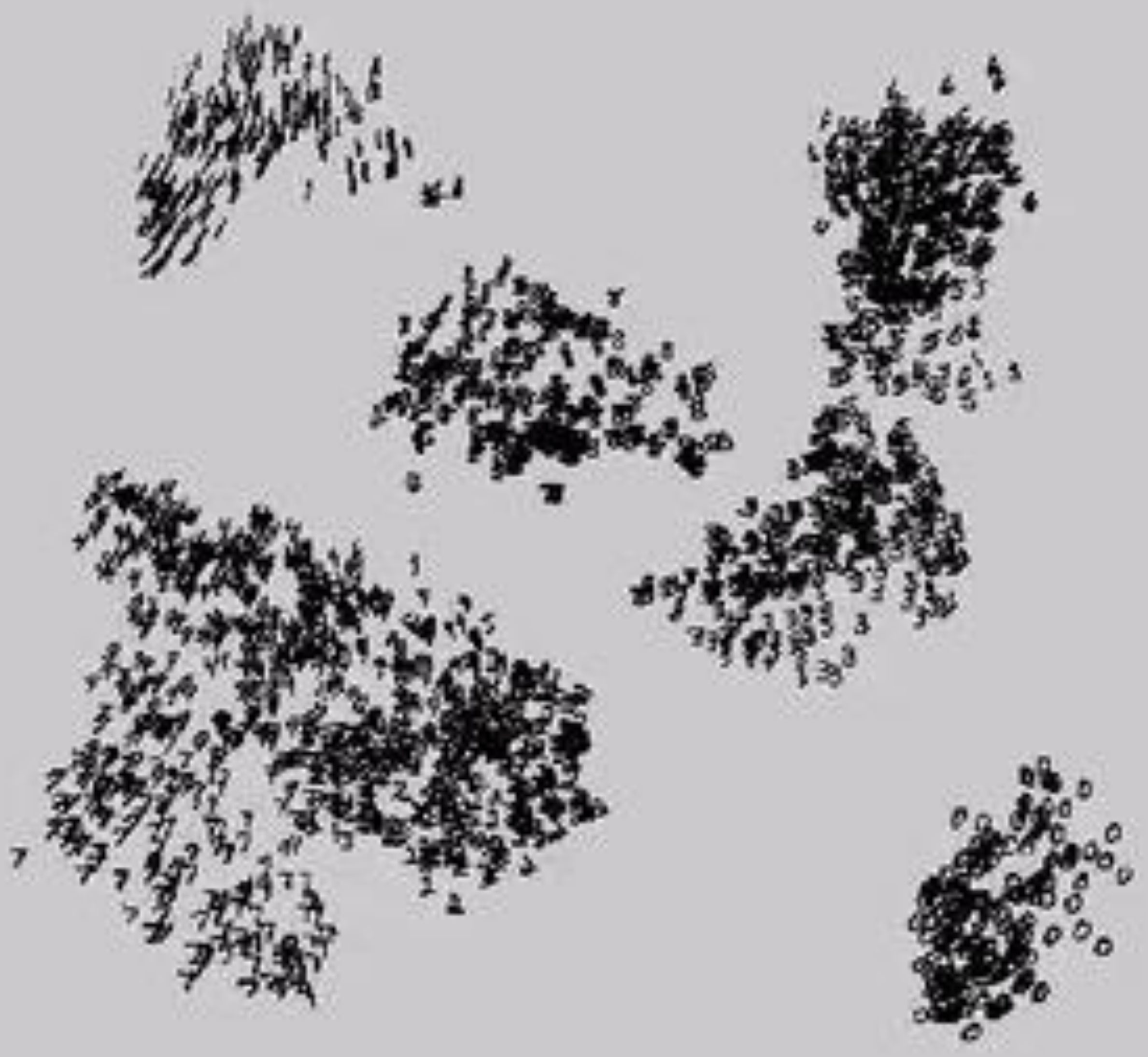
Sammon mapping



t-SNE



t-SNE



Question

So t-SNE visualization tends to unravel beautiful clear-cut clusters, and usually it "just works" in-practice straight out of the box.

Does it come with any sort of ***guarantees*** on the visualization it produces?

Results on True Positive discovery

Good News: If there are clear well-separated clusters in the high-dimensional input data, then 2D t-SNE visualization will be able unravel it.

Literature:

⇨ Global minima of t-SNE can reveal clusters for highly separated Gaussian-like clusters. [Shaham and Steinerberger '17]

- Very first theoretical result
- Cluster preservation defined in an odd unintuitive way
- Requires unrealistically large number of clusters to work

⇨ A local minima of t-SNE ran with exaggeration phase can potentially reveal well-separated clusters [Lindermann and Steinerberger '18]

- Analyzed by viewing the gradient update as a dynamical system
- The intra-cluster distances contract at a fast-enough rate

⇨ A local minima of t-SNE ran with exaggeration phase will reveal well-separated clusters [Arora, Hu, Khotari '19]

- Extends previous result and have an intuitive definition of 'reveal'
- Not only the clusters contract, but remain separated

Other Notable (Theoretical) Results

Some fundamental results are just being established…

- t-SNE gradient update acts a Markov chain and the visualization it produces is similar to doing spectral clustering on a specific kind of Laplacian matrix

  [Tony Cai and Ma ’22]

- t-SNE is consistent in the sense that embeddings generated by an i.i.d. sample from a fixed probability distribution converge in the limit

  [Auffinger and Fletcher ’23]

- t-SNE optimization provably has a minimizer (under mild assumptions)

  [Jeong and Wu ’24]

Negative (Theoretical) Results

All theoretical results (so far) are on “positive”, i.e. t-SNE works.

Are there any study on negative results?

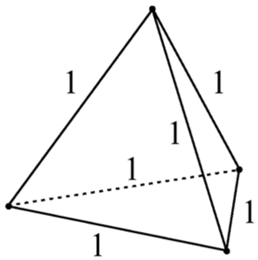
⇨ t-SNE always biases towards “clustering” an input dataset (even if there may not be any clusters in the input dataset) [Im, Verma, Branson ’18]

- can result in false cluster discovery
- provides a generalization to f-divergences to ameliorate this effect

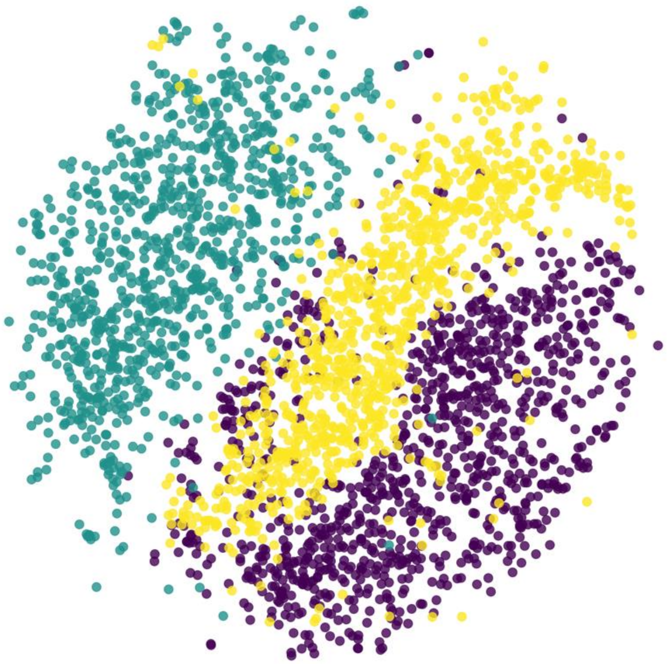
That’s it, exactly seven theoretical results exist on this topic.
(one negative result, all others positive :/)

(New) Results on False Positive Discovery

Claim: Any visualization that can be produced by t-SNE on a given dataset, can also be produced by a slight perturbation of a regular simplex!



QUIZ: One of these visualizations have been generated from MNIST dataset (3 digits), the other from slightly perturbing the simplex. Which one is which?



MNIST

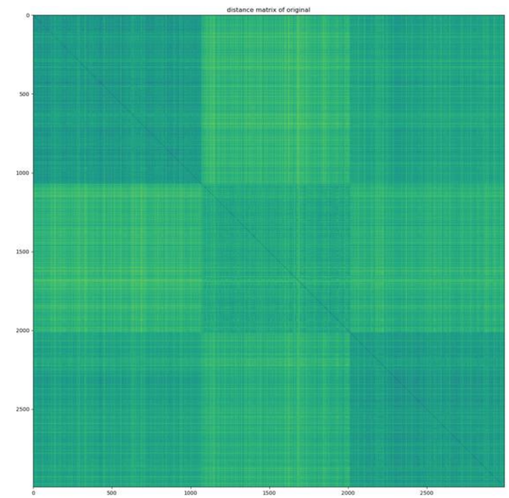
simplex

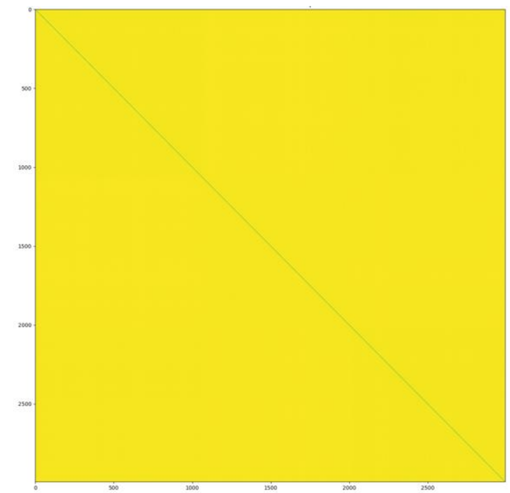
(New) Results on False Positive Discovery

MNIST

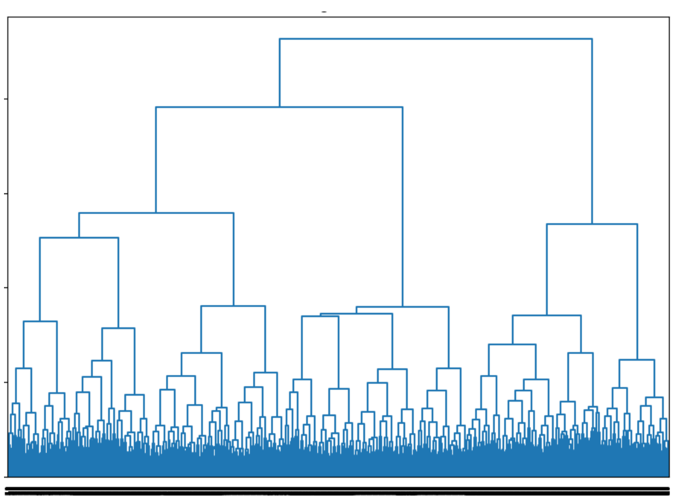
simplex

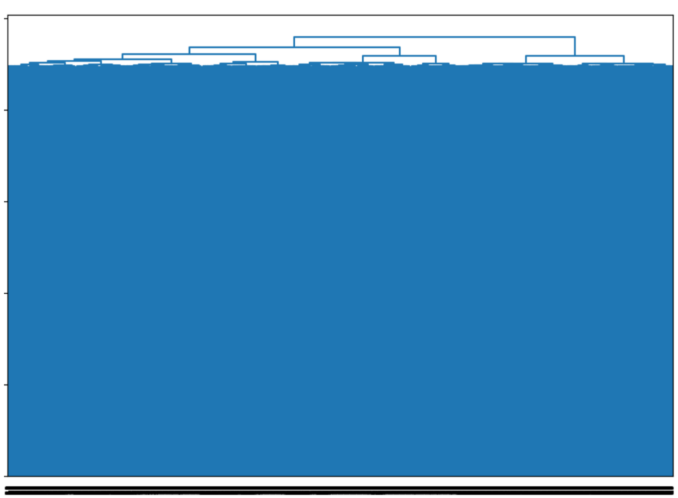
Interpoint distance profile





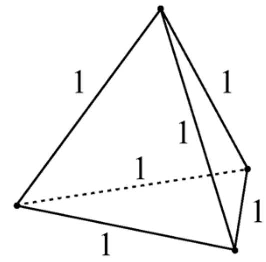
Hierarchical cluster profile





(New) Results on False Positive Discovery

Claim: Any visualization that can be produced by t-SNE on a given dataset, can also be produced by a slight perturbation of a regular simplex!



Proof Sketch:

The neighborhood probability matrix P induced by any input dataset can also be induced by a (perturbed) regular simplex.

How?

We show t-SNE's P matrix is both additive and multiplicative **invariant** to the pairwise interpoint distances.

Consider pairwise distances between three points:

	a-b	b-c	a-c
	5	10	10
(additive invariance)	5+10000	10+10000	10+10000
(multiplicative invariance)	(5+10000)/10000	(10+10000)/10000	(10+10000)/10000
(regular simplex!)	1.0005	1.0010	1.0010

(New) Results on False Positive Discovery

Try #2: One of these visualizations have been generated from IRIS dataset (3 clusters), the other from slightly perturbing the simplex. Which one is which?

IRIS dataset

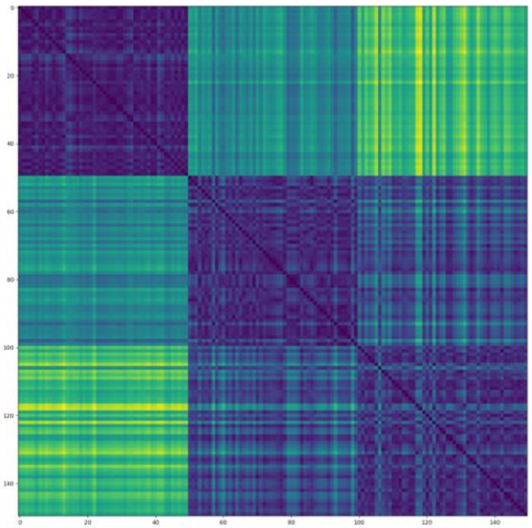
simplex

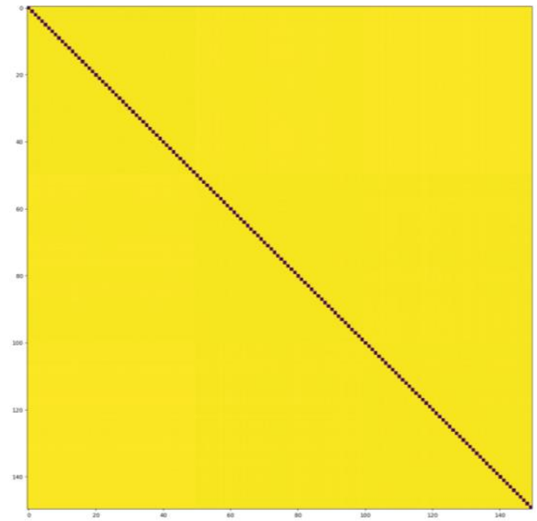
(New) Results on False Positive Discovery

IRIS

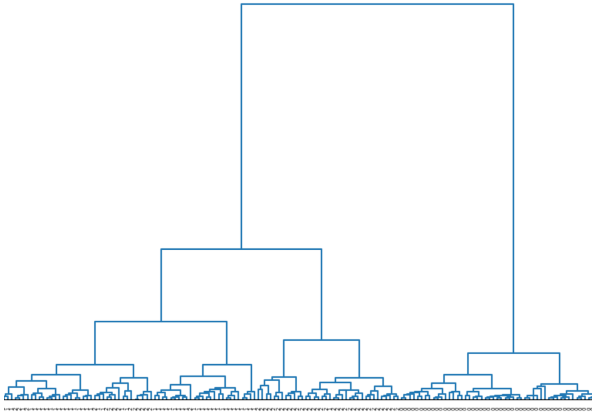
simplex

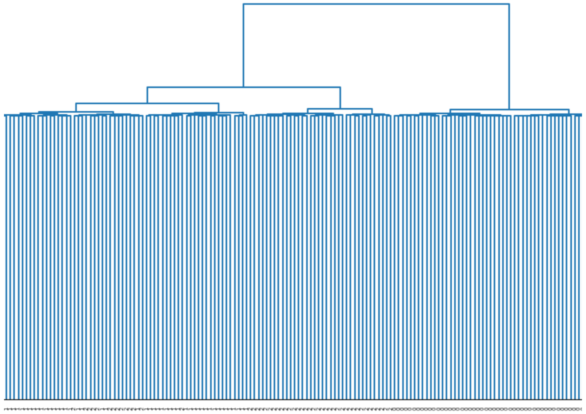
Interpoint distance profile





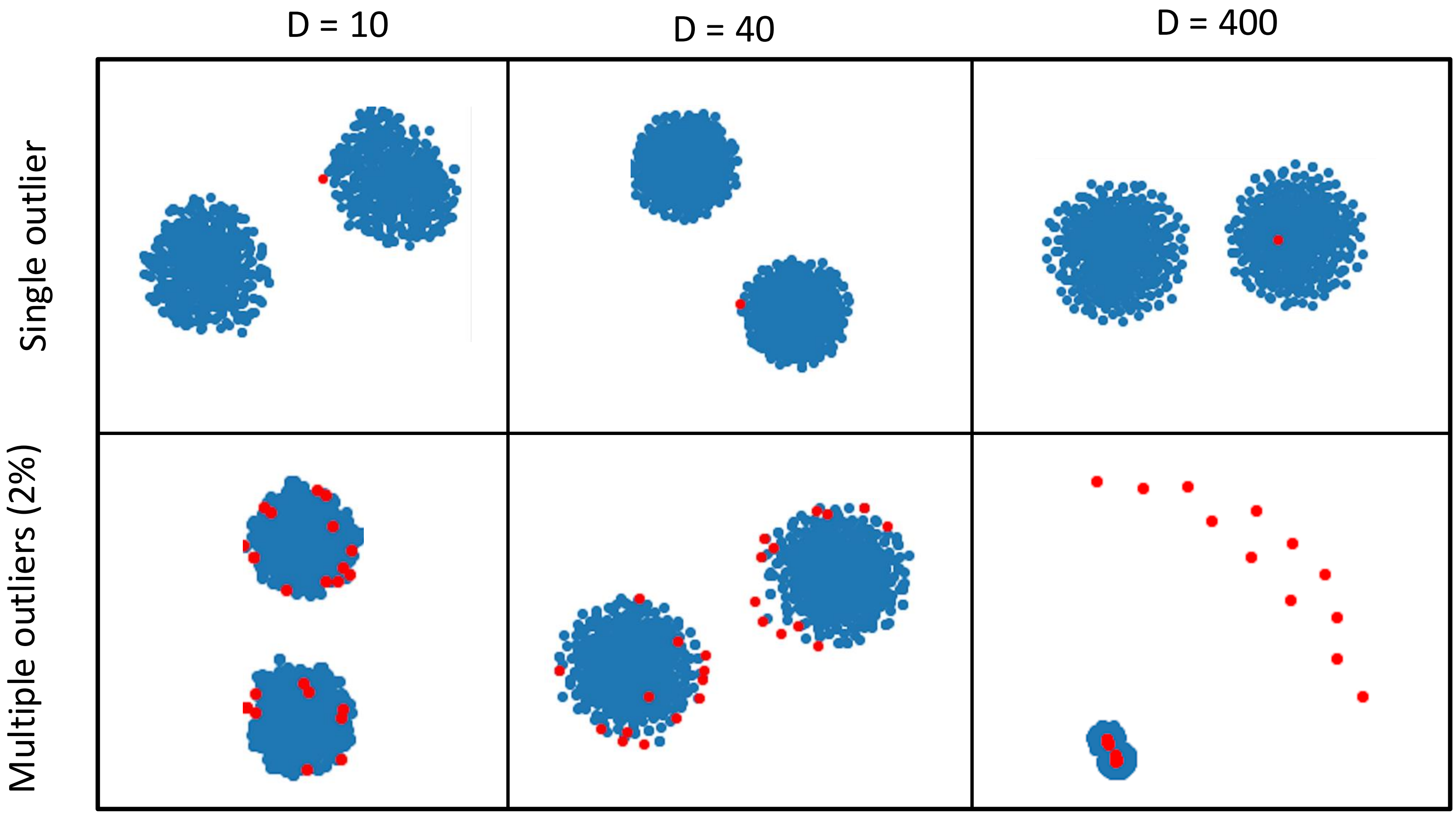
Hierarchical cluster profile





Effects in the Presence of Outliers

Claim: Extreme outliers in the input data cannot be shown as far away from the other (inlier) data points in any locally optimal t-SNE embedding

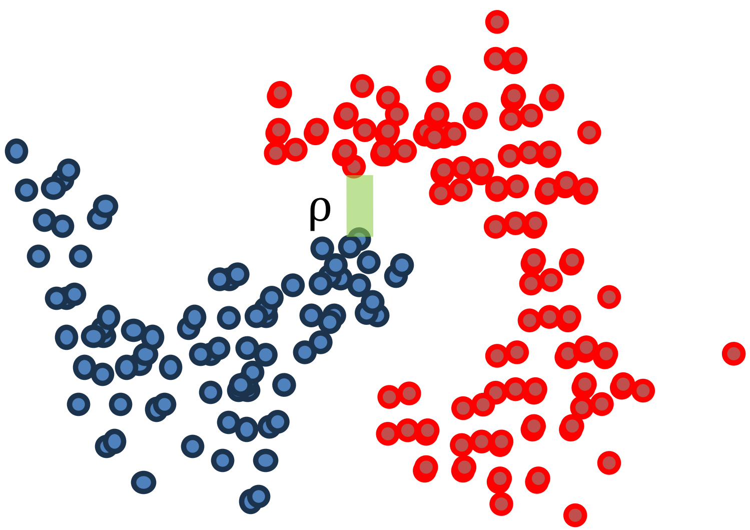


Universality Results

So we know tSNE can fail, is it possible to perhaps modify it or come up with an entirely new mapping (read U-MAPping) that works well?

How can we answer this question formally?

Given a dataset X of n points with a designated partition into *k* clusters, we say that a visualization (ie a map f: X → R^d) recovers the partition at resolution ρ if points from distinct clusters are mapped ρ away



Universality Results

Questions to ask:

- Can we design an *f* which recovers the partitions at some acceptable/tolerable resolution (say ρ = 1%, 0.1%, etc.) on input datasets with clear *k*-partitions?
- What restrictions, if any, such an *f* must have?

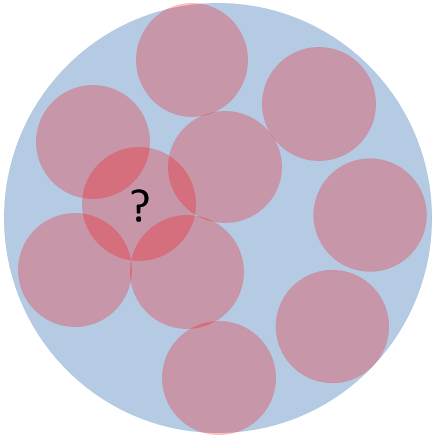
An interesting observation:

Any good *f* must obey the relationship

$$d \geq \log(k)/\log(1/\rho)$$

An elegant volume argument can be used.

Idea:



Embedding space – size $(1/\rho)^{nd}$

Each good cluster embedding – k^n total

Want: distinct cluster embeddings to not overlap, so $k^n \leq (1/\rho)^{nd}$

Universality Results

Implications:

Any visualization algorithm (tSNE, UMAP, autoencoder…) into 2-D **MUST fail*** on some dataset which has clear well-separated clusters with no outliers!

*fail means unable to recover/reveal/show the clusters

Alternatively, as a function of k (i.e. the number of clusters), any 2-D visualization **MUST suffer** the issues of the "crowding problem"

This result generalizes to any metric space (so the same bad news in spaces beyond Euclidean space, e.g. hyperbolic space, etc.)

Parting thoughts and future analysis

- t-SNE is a remarkably effective in visualizing cluster structure in data

  Arguably the best (along with UMAP) ultra low-dimensional technique that "just works"!

- t-SNE tends to cluster even when there may not be any clusters

  Can result in false cluster discovery!

  [Im, Verma, Branson '18]

  [Snoeck, Bergam, Verma '25?]

- t-SNE unfortunately doesn't behave well in the presence of outliers.

  Can result in false understanding of the dataset

  [Snoeck, Bergam, Verma '25?]

- Universal cluster-revealing visualizations are unfortunately not possible.

  [Snoeck, Bergam, Verma '25?]

Parting thoughts and future analysis

Other interesting avenues to explore…

- Hardness of the t-SNE objective
    - is it NP-hard?
    - does a good approximation to the objective exist?

- (theoretical) quality of the local minima

- Smart seeding/initialization

- There are absolutely no (theoretical) results on UMAP!!!

Questions/Discussion