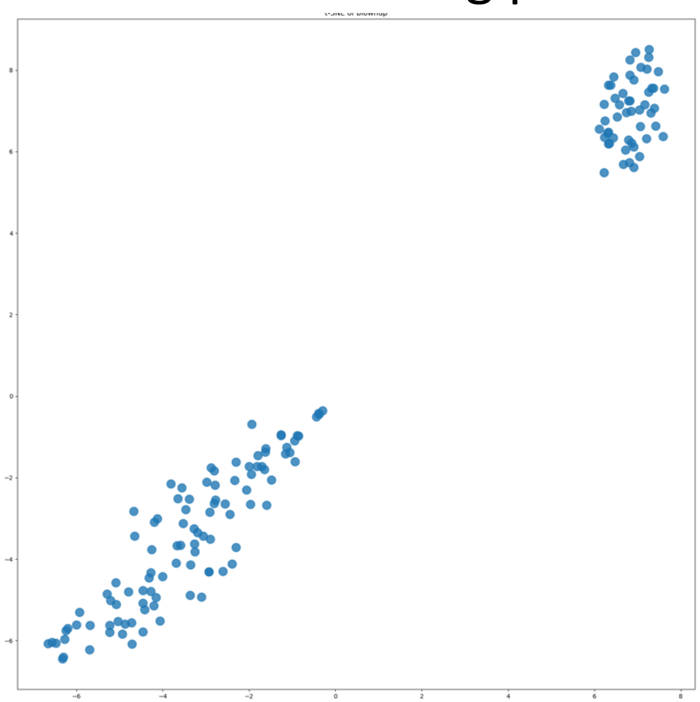
Perils of Using t-SNE (and friends)

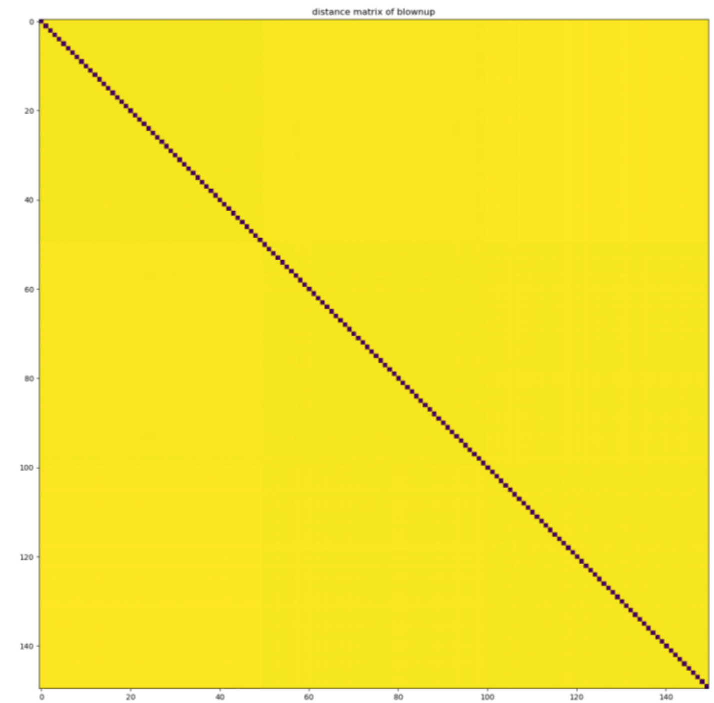
Nakul Verma

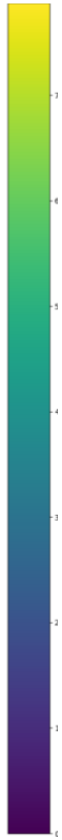
Columbia University

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

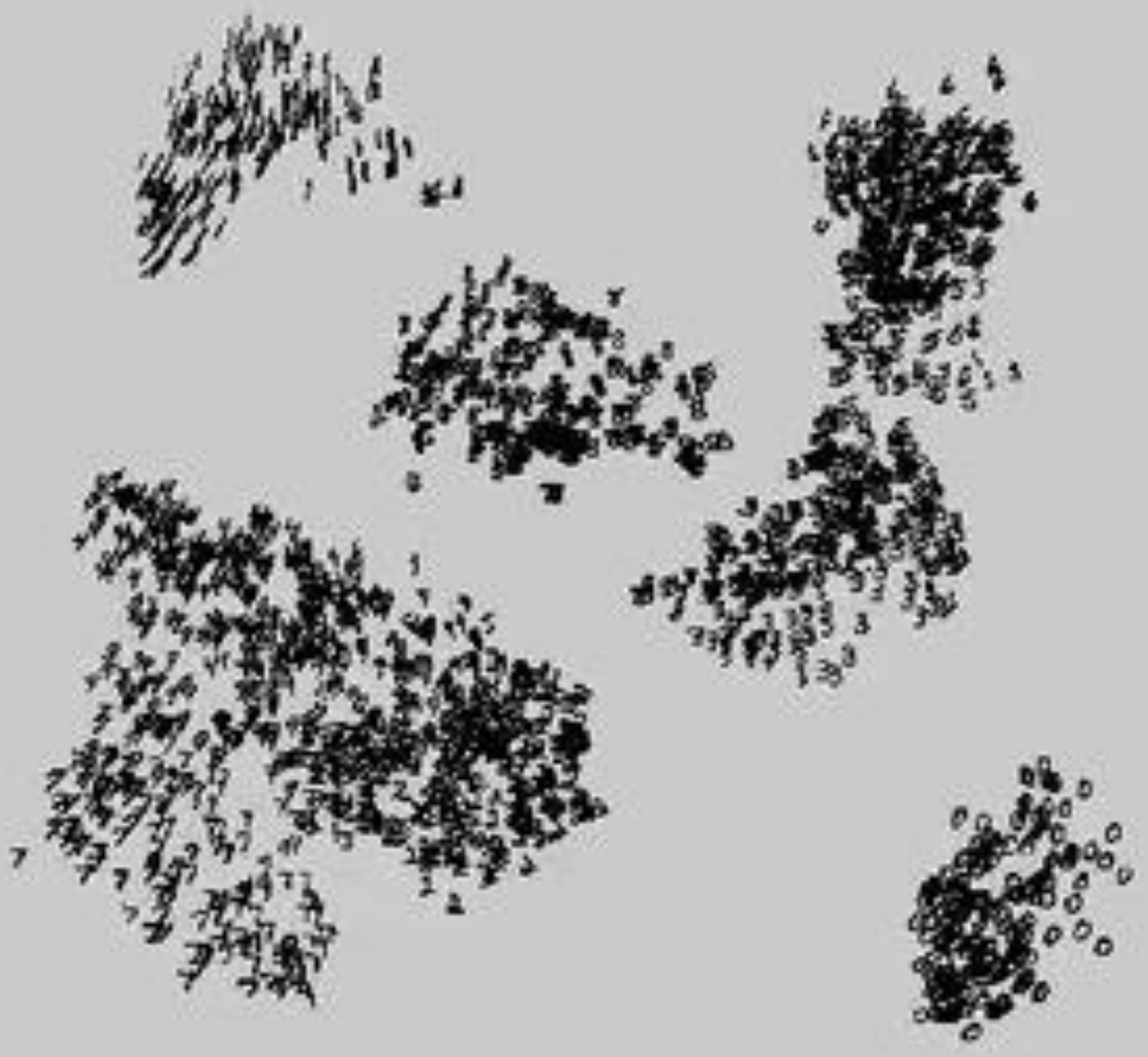
Idea:

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

(this, along with a lot of smart heuristics makes tSNE shine in practice)

The result…

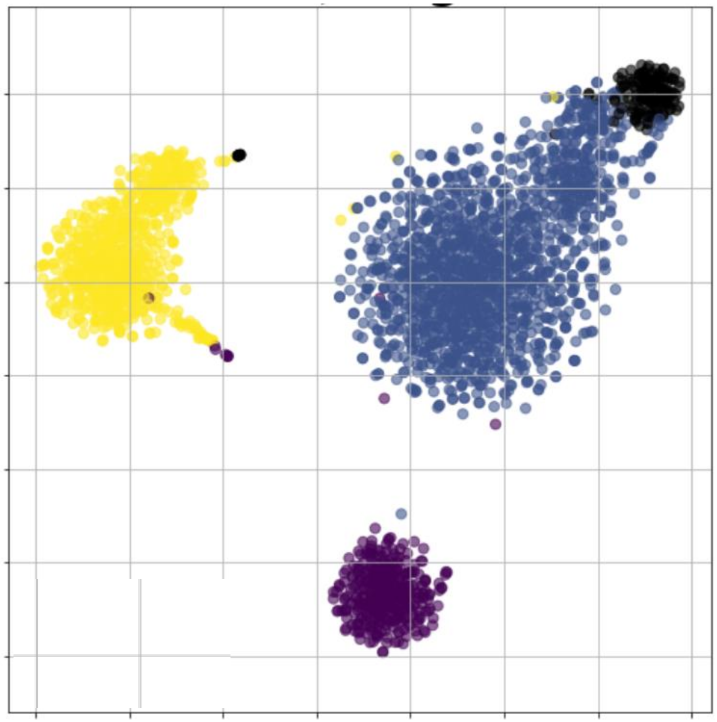
t-SNE – success story 1



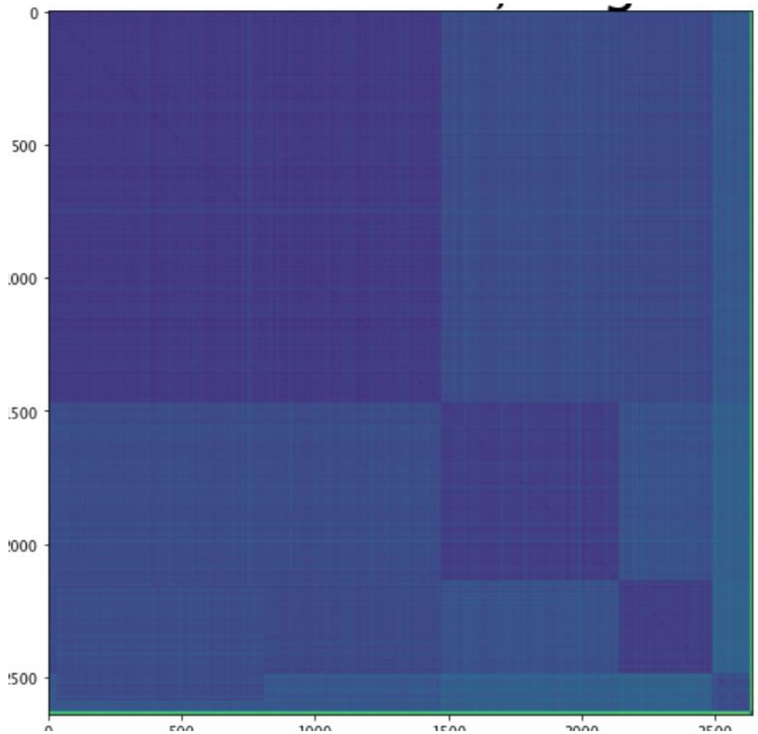
t-SNE – success story 2

PBMC3k Single Cell Data (2k points, 50 dimensions)

tSNE visualization



Original Pairwise distance structure



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

Negative (Theoretical) Results

Ok, so we have true positive results.

If data is truly clustered $\Rightarrow$ tSNE shows clusters

What about negative results? (false positives and false negatives)

NONE!?!

Why should we care?

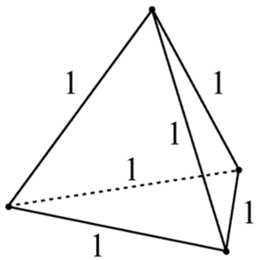
tSNE shows clusters $\not\Rightarrow$ data is truly clustered

For tSNE (or any data visualization method: PCA, UMAP, Isomap, LLE, …)

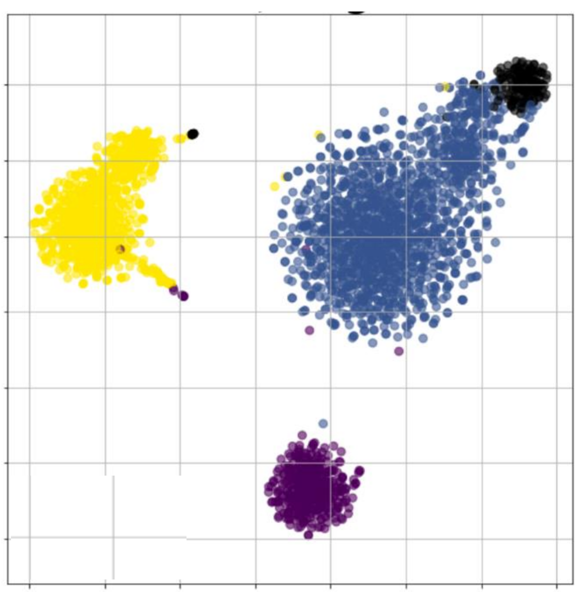
- Are false positives possible (shows clusters, but there aren't any)
- Are false negatives possible (shows no clusters, but there are)

(New) Results on False Positive Discovery

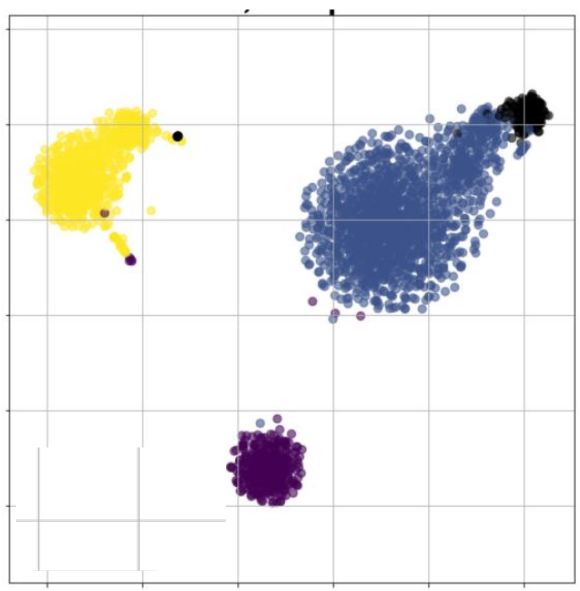
Claim: Any visualization that can be produced by t-SNE on a given dataset, can also be produced by a slight perturbation of a regular simplex!



QUIZ: One of these visualizations have been generated from PBMC3k dataset (~4 clusters), the other from slightly perturbing the simplex. Which one is which?



PBMC3k

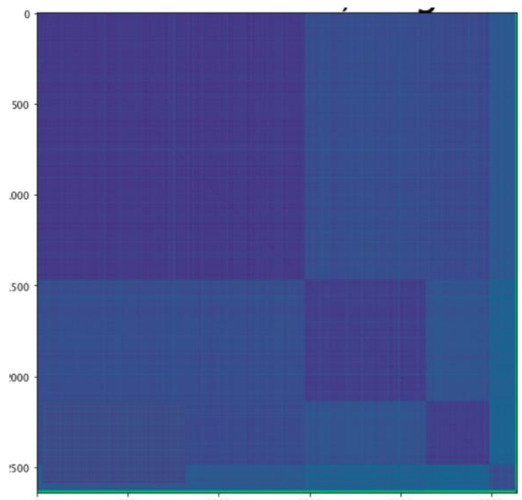


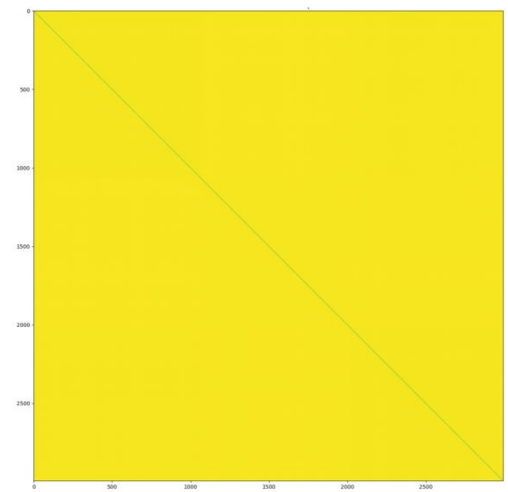
simplex

(New) Results on False Positive Discovery

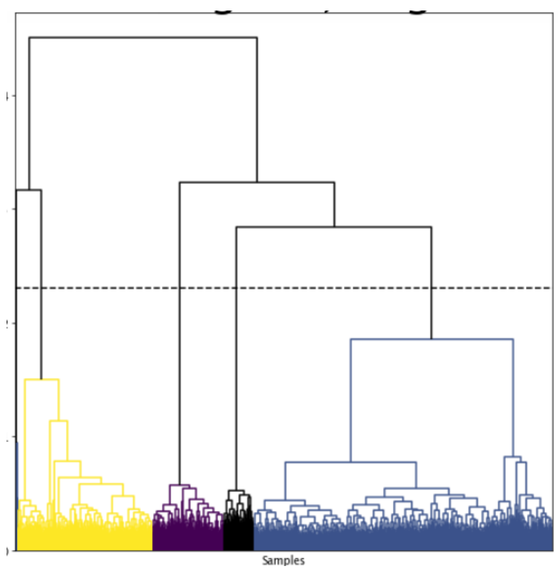
PBMC3k | simplex

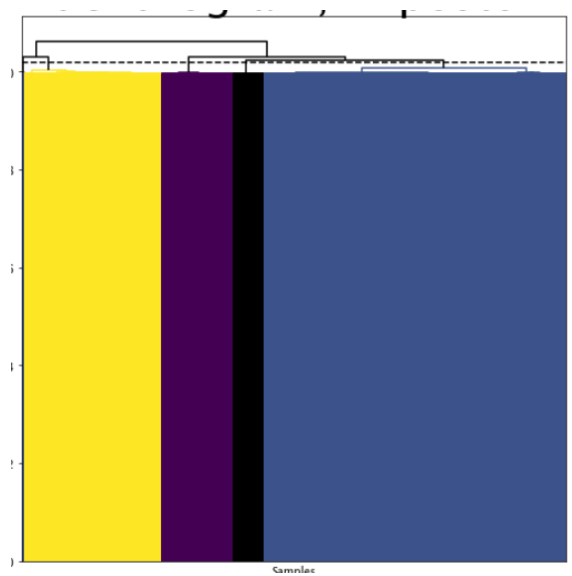
Interpoint distance profile





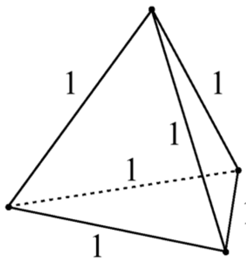
Hierarchical cluster profile





(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
(approx. 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

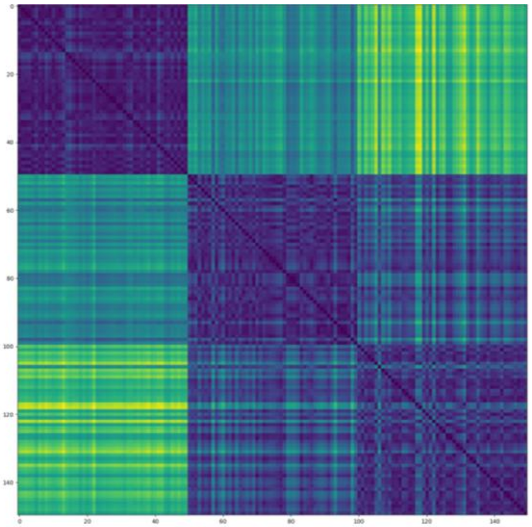
simplex

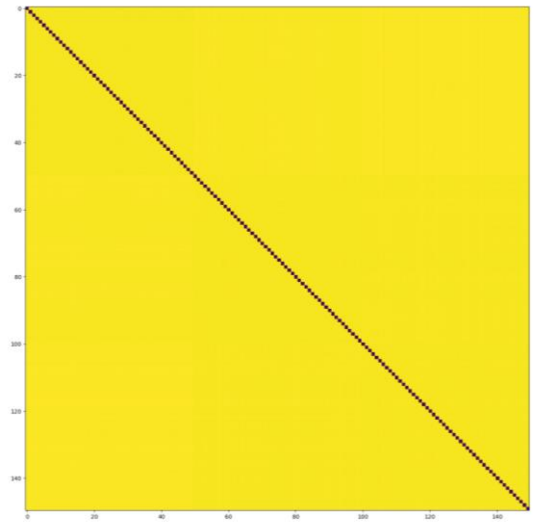
(New) Results on False Positive Discovery

IRIS

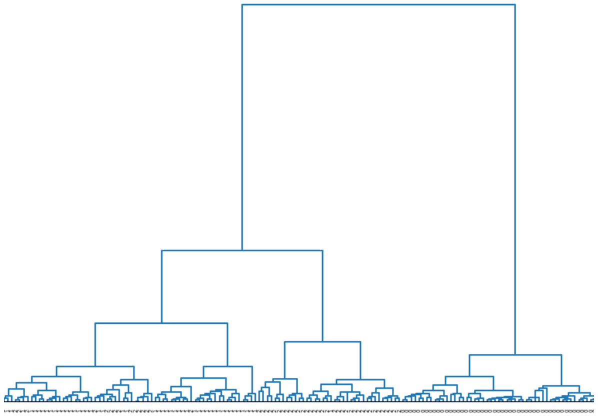
simplex

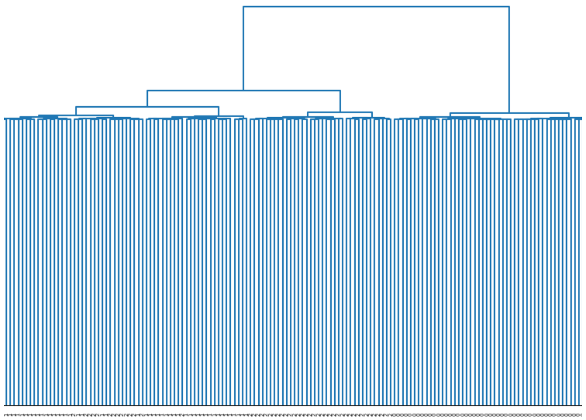
Interpoint distance profile





Hierarchical cluster profile

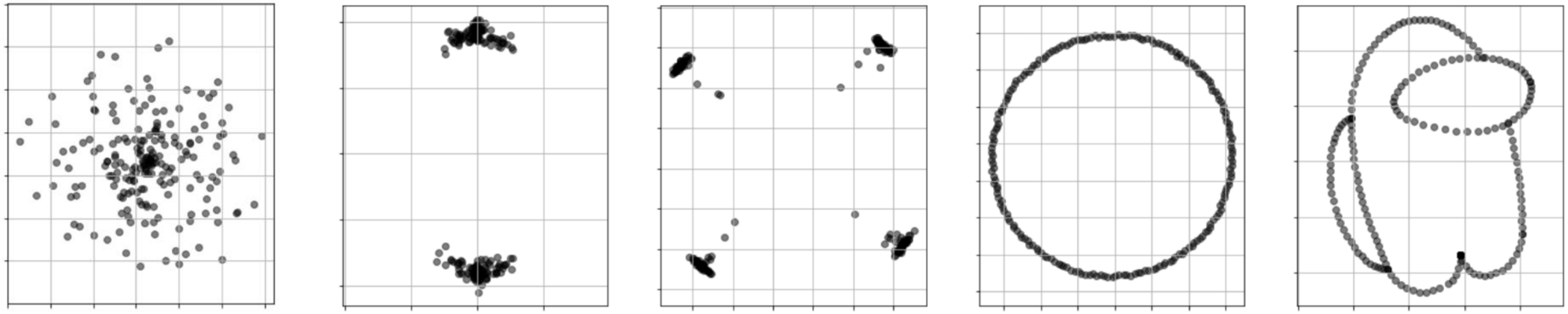




Implications

Corollary: A regular simplex (all interpoint distances equal) is an unstable input for tSNE algorithm.

small perturbations in the input can lead to dramatic changes in the output



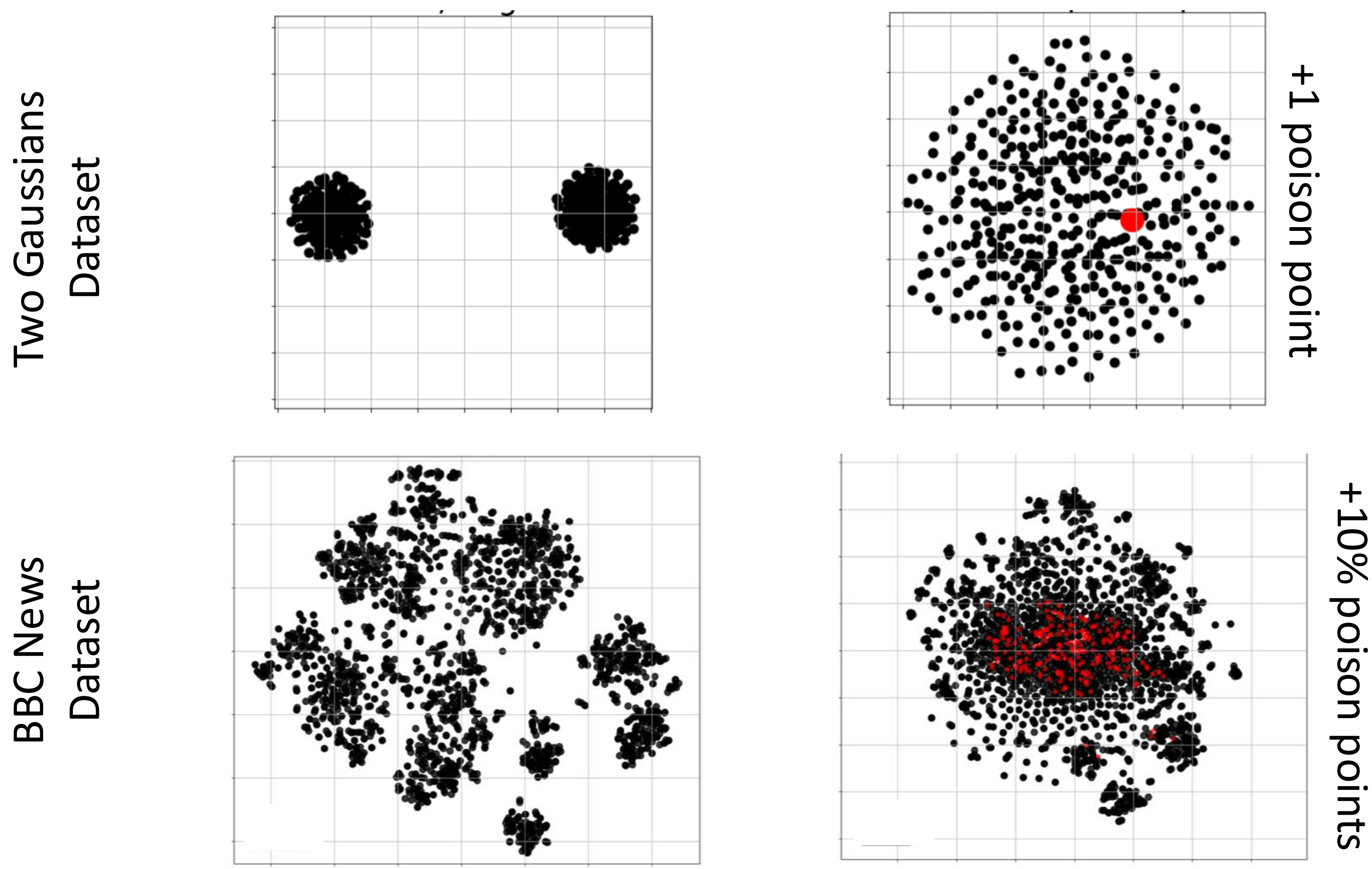
various tSNE outputs generated from very tiny perturbation of the same input (simplex)

Umm... ok so there is some pathological tSNE input, but my inputs are usually clustered and very unlike 'regular-simplex-like'

Well... high dimensional data is VERY simplex-like (provable and demonstrable)

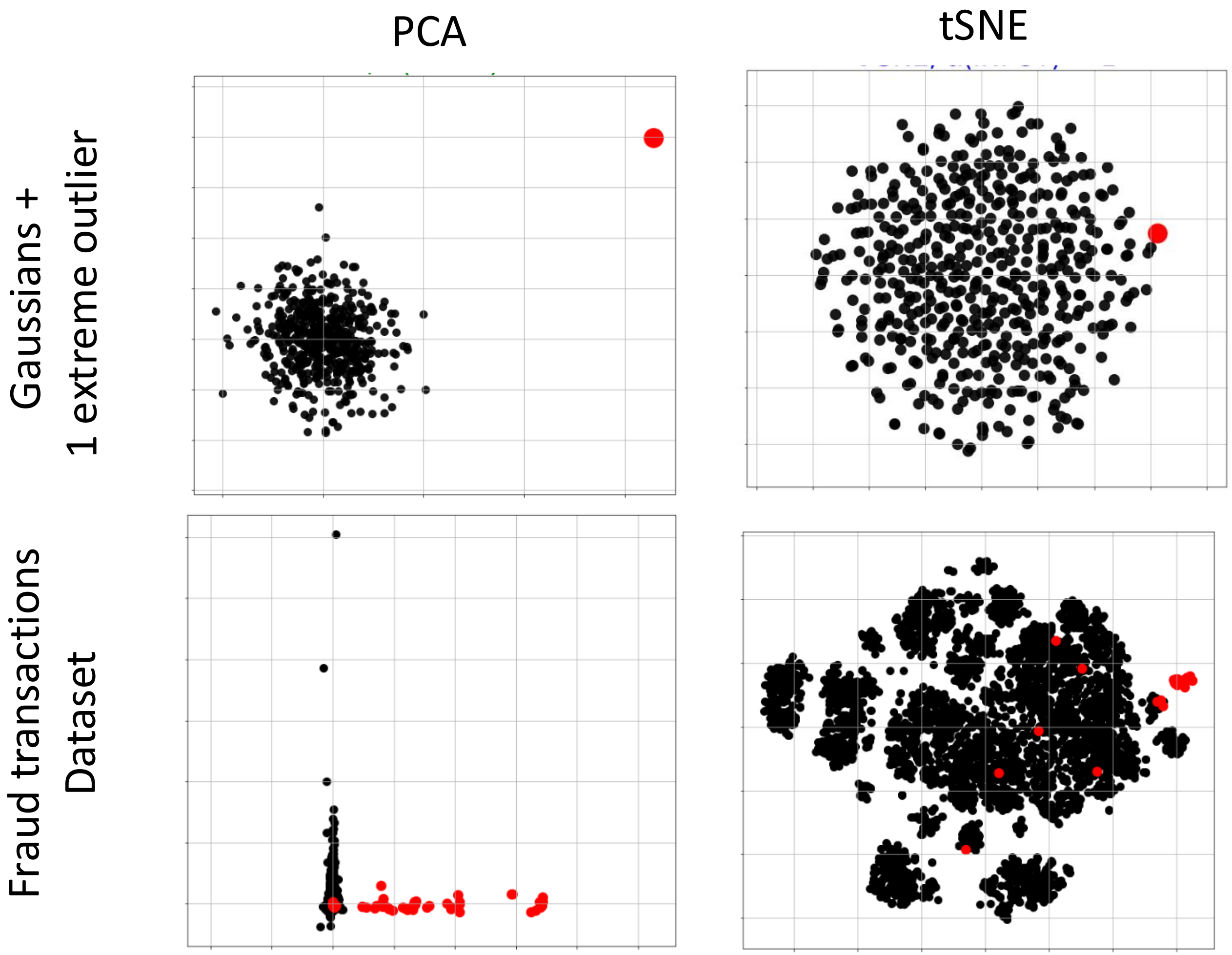
(New) Results on False Negatives

Claim: (some) Clustered input can result in unclustered tSNE visualization by simply introducing a few (sometimes even just one) strategically placed datapoints!



(New) Results on False Negatives

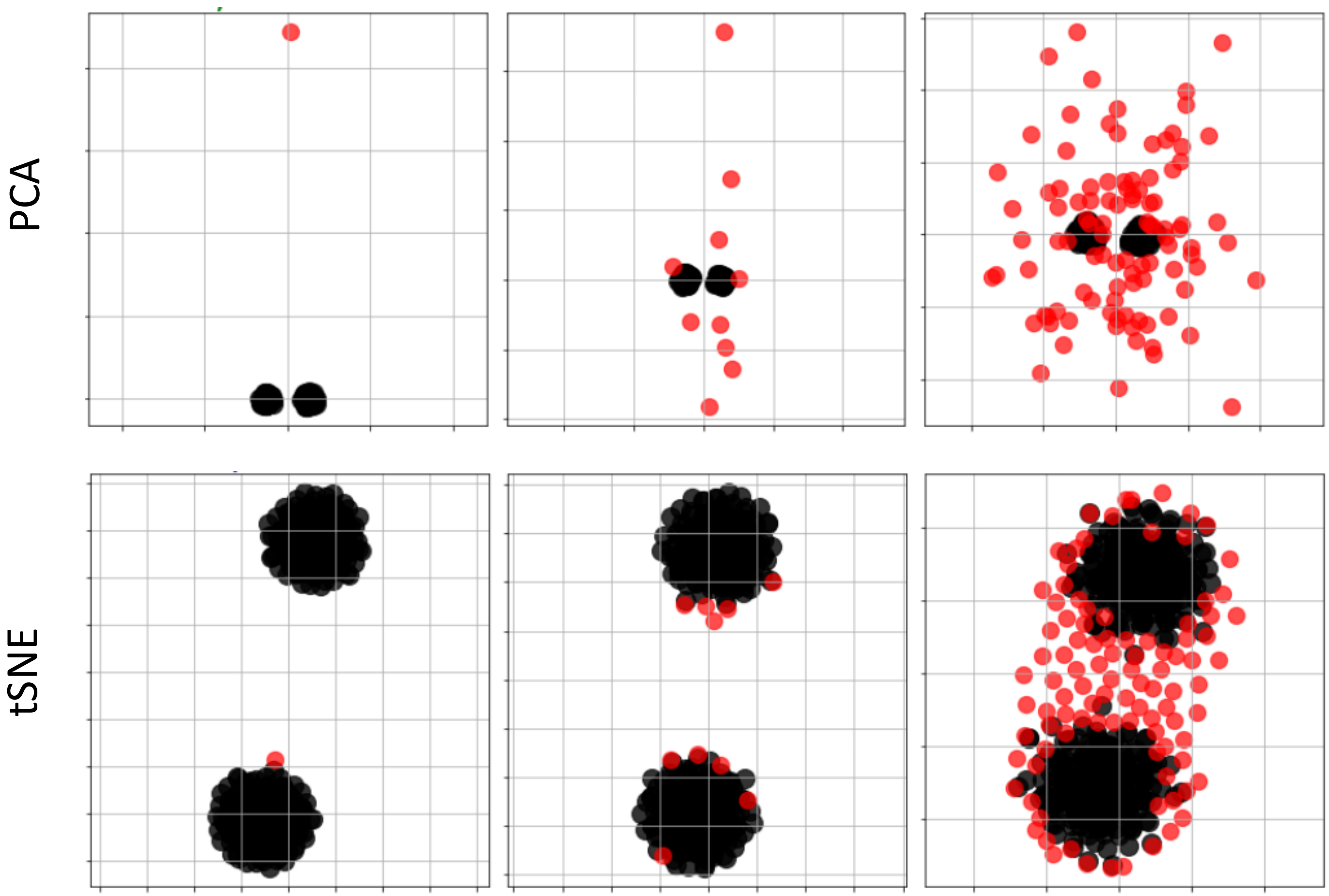
Claim: (extreme) outliers cannot be visualized by tSNE!



(New) Results on False Negatives

Claim: (extreme) outliers cannot be visualized by tSNE!

Two Gaussians + extreme outliers

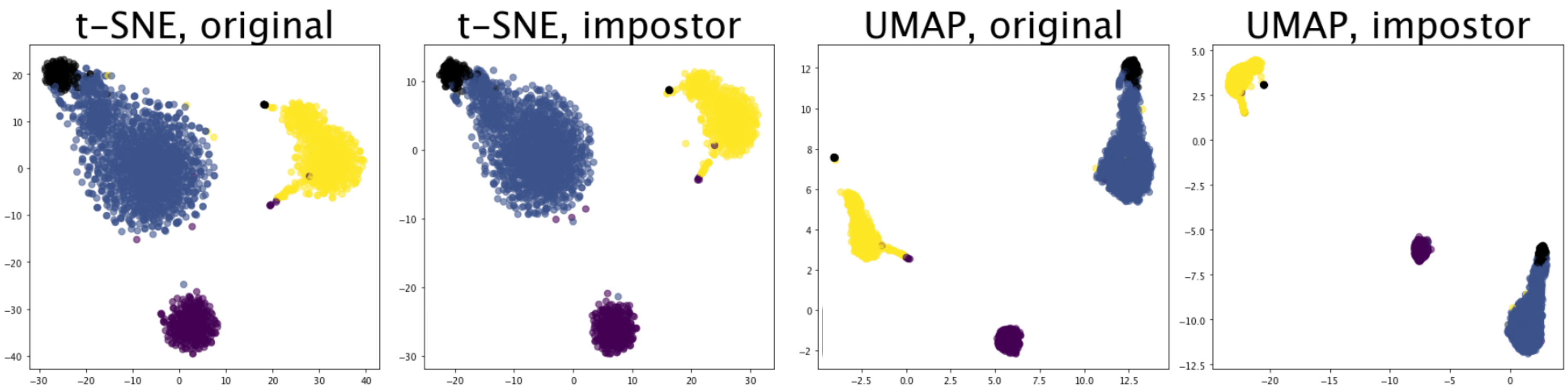


(New) What about UMAP?

HA… but I use UMAP!

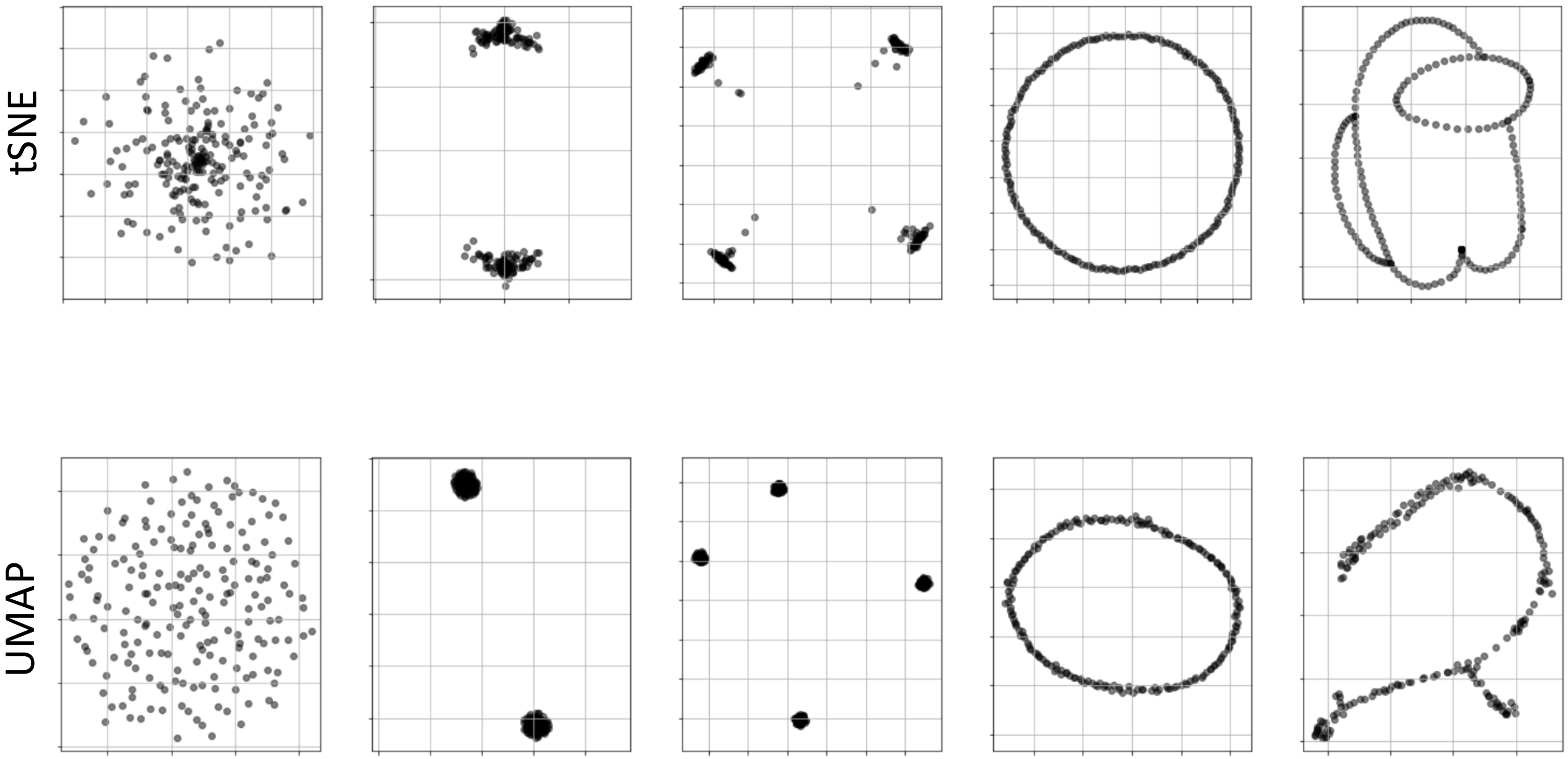
Unfortunately, same results

False Positives:



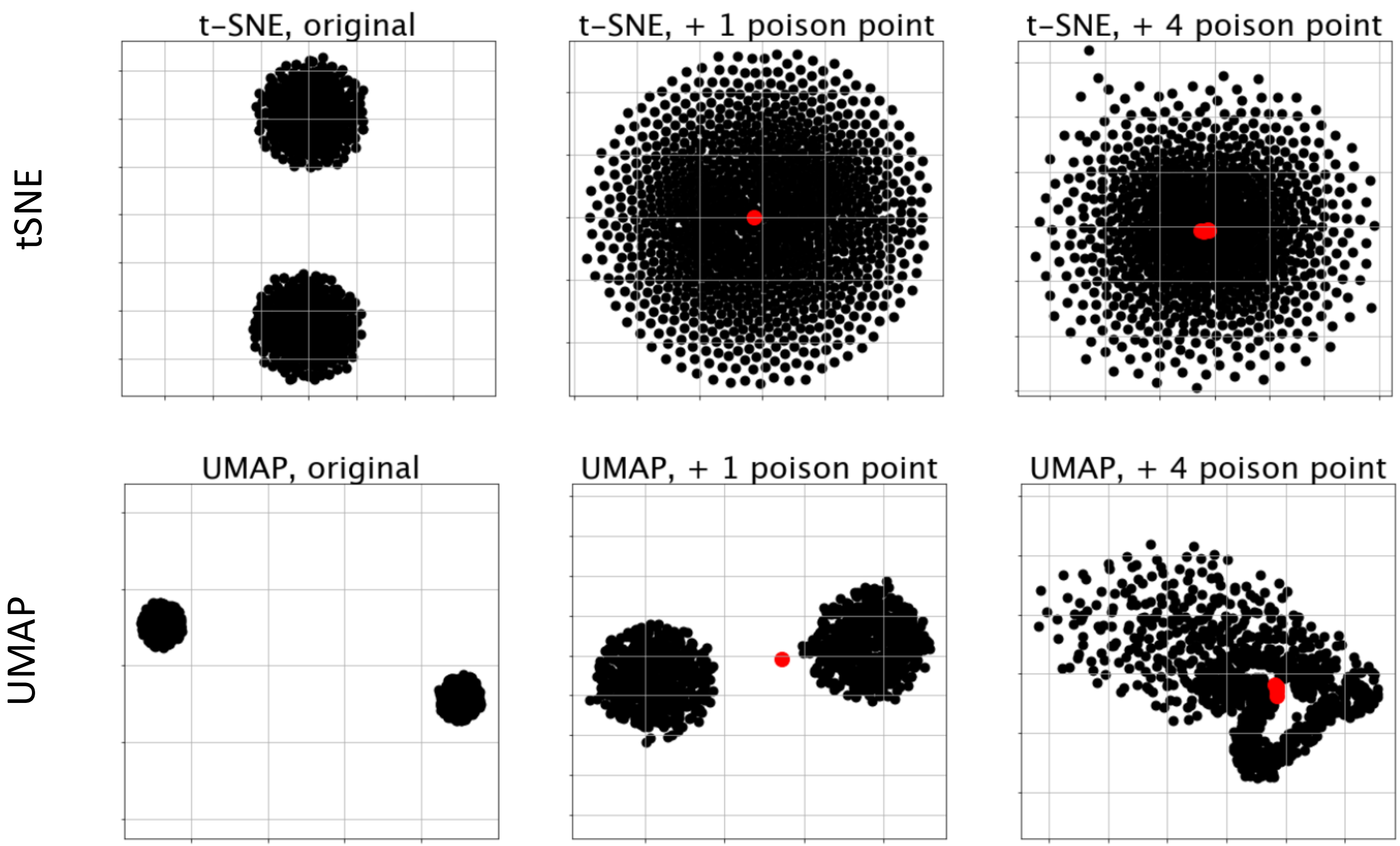
(new) UMAP Results

Output instability (minor perturbations of the input):



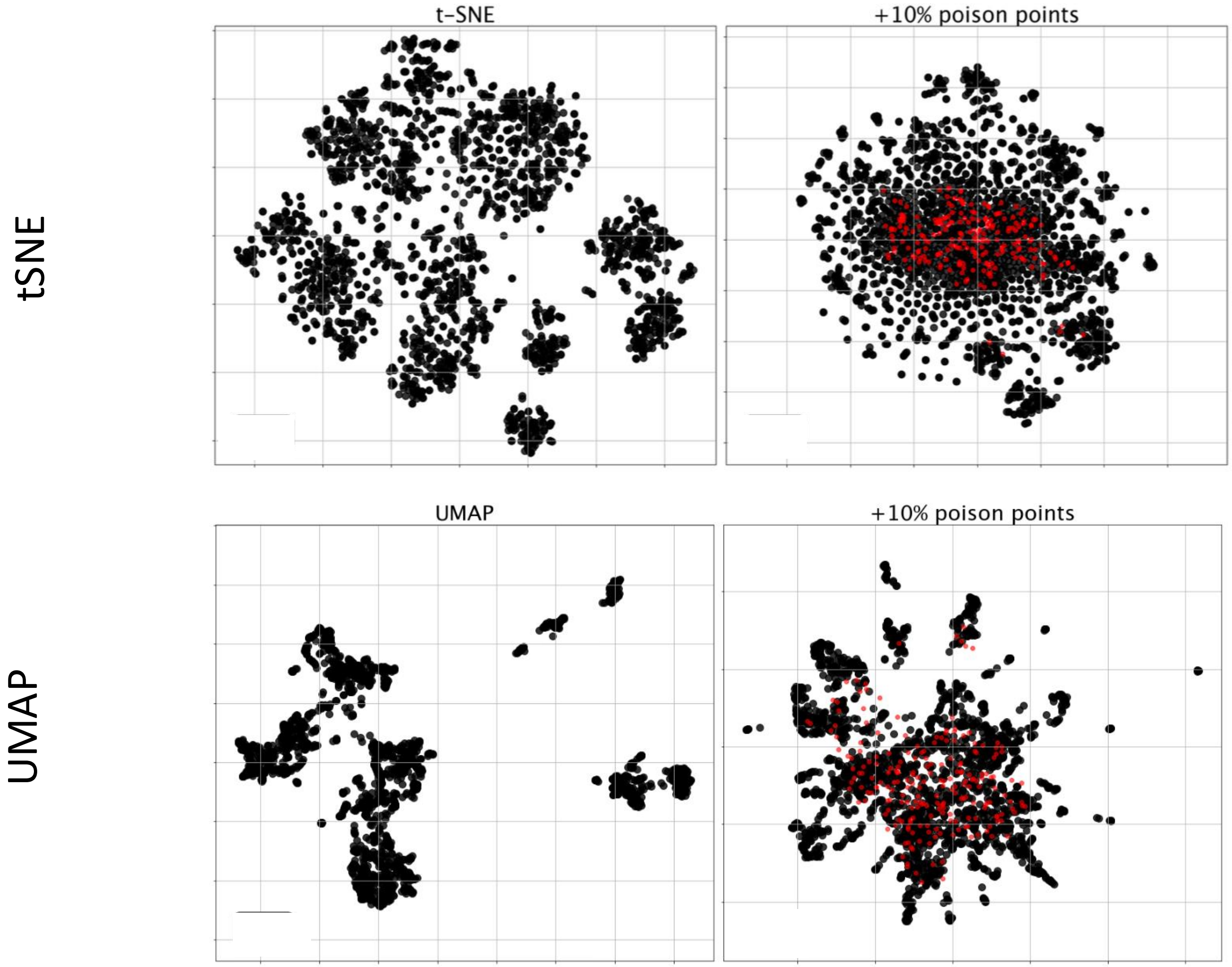
(new) UMAP Results

False Negatives (destroying clustering by adding poison points):



(new) UMAP Results

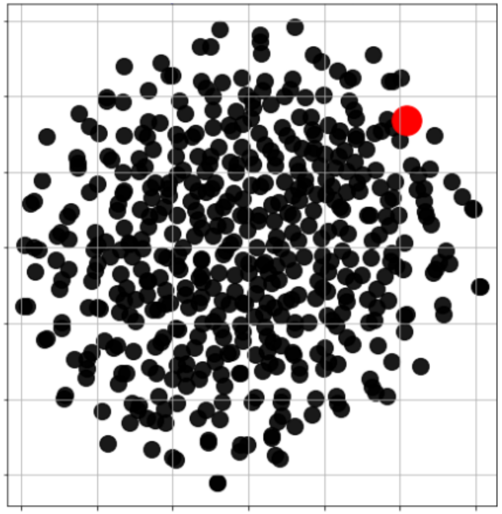
False Negatives (destroying clustering by adding poison points):



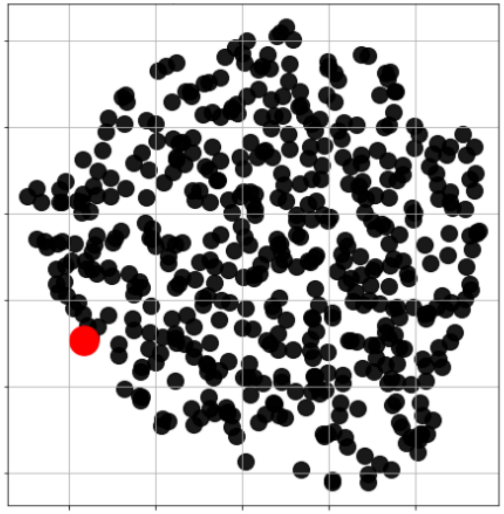
(new) UMAP Results

False Negatives (outlier extremity):

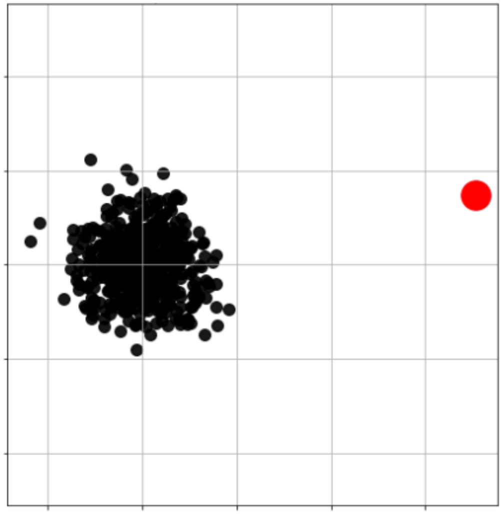
tSNE



UMAP

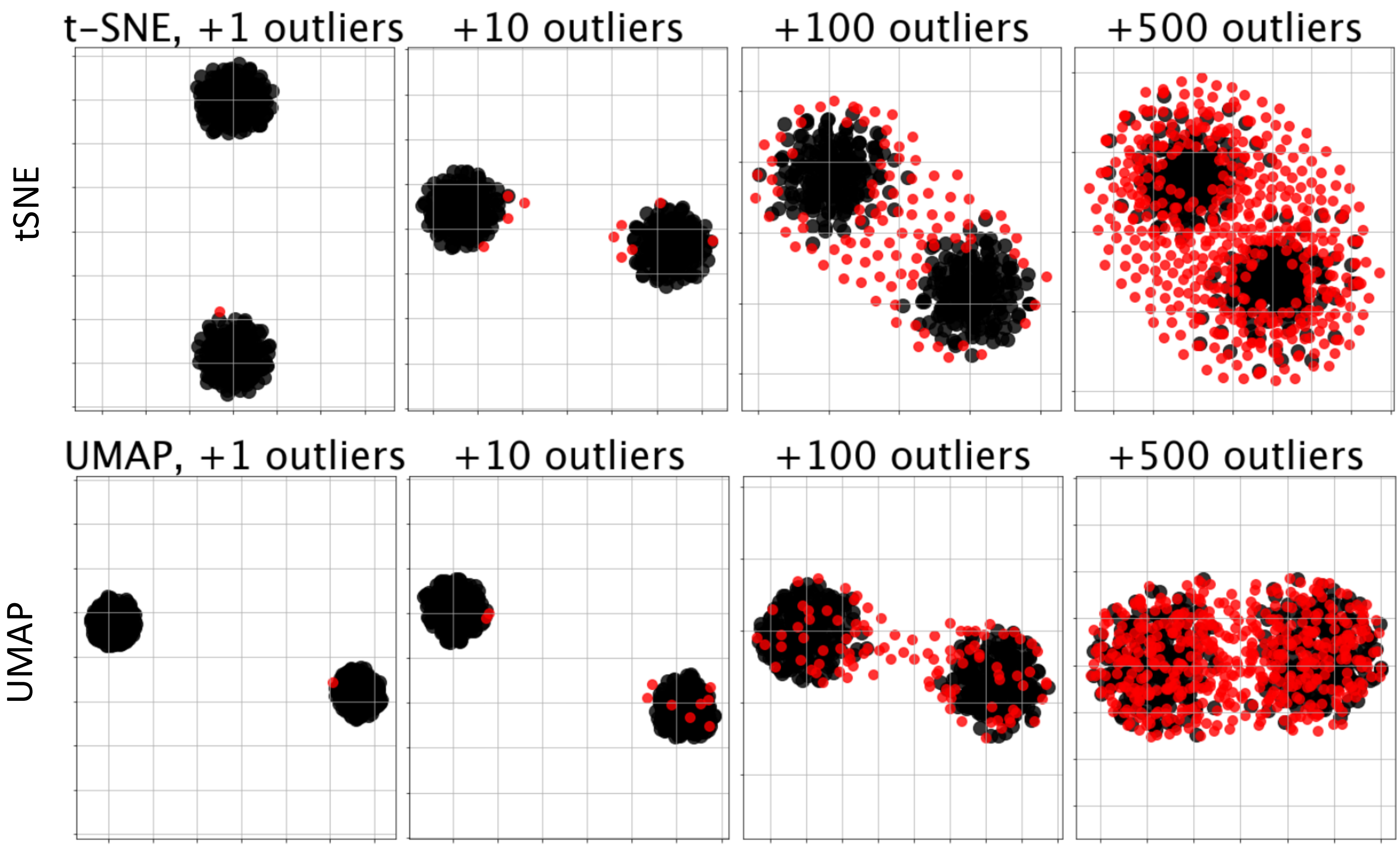


PCA



(new) UMAP Results

False Negatives (outlier extremity):



(new) Universality Results

So we know tSNE (and UMAP) can fail, is it possible to perhaps modify it or come up with an entirely new uber algorithm that works well?

In terms of neighborhood preservation, unfortunately… tl;dr - NO

Claim:

Any visualization algorithm (tSNE, UMAP, autoencoder…) into 2-D **MUST fail*** on some (in fact on almost all) datasets in terms of preserving local neighborhoods!

*fail means unable to properly maintain/show all local neighborhoods

So neighborhood distortion is almost imperative!

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]
  [Bergam, Snoeck, Verma '26]

- t-SNE unfortunately doesn't behave well in the presence of outliers.
  Can result in false understanding of the dataset

  [Bergam, Snoeck, Verma '26]

- Universal cluster-revealing visualizations are unfortunately not possible.

  [Snoeck, Bergam, Verma '26]

Parting thoughts and future analysis

Other interesting avenues to explore…

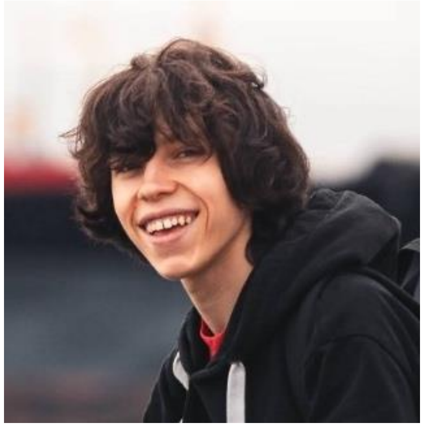
- Hardness of the t-SNE objective
  - is it NP-hard?
  - does a good approximation to the objective exist?
- (theoretical) quality of the local minima
- Smart seeding/initialization
- There are absolutely no (theoretical) results on UMAP!!!

Questions/Discussion

Thank You!

for patiently listening! ☺

Partners in Crime (aka Collaborators)



Szymon Snoeck

Noah Bergam

(2024-26+)

Daniel Im

Kristin Branson (2017)

(Honorary mention) Roian Egnor

(2015-17)