t-SNE exaggerates clusters, provably

Noah Bergam

ICLR 2026
Joint work with

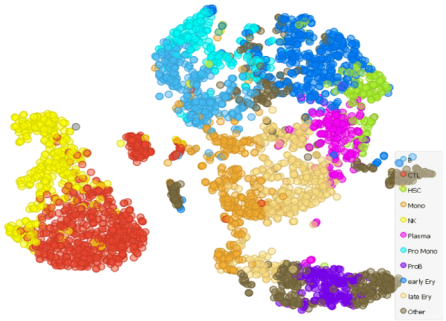
Szymon Snoeck

Nakul Verma

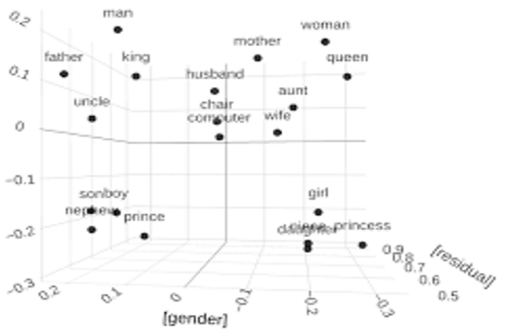
What is t-SNE?

t-SNE (t-distributed stochastic neighbor embedding) is a very popular data visualization algorithm.

cells:



words:



Wide practical adoption in research…poor basic understanding of how it works (likewise UMAP, TriMAP, etc.)

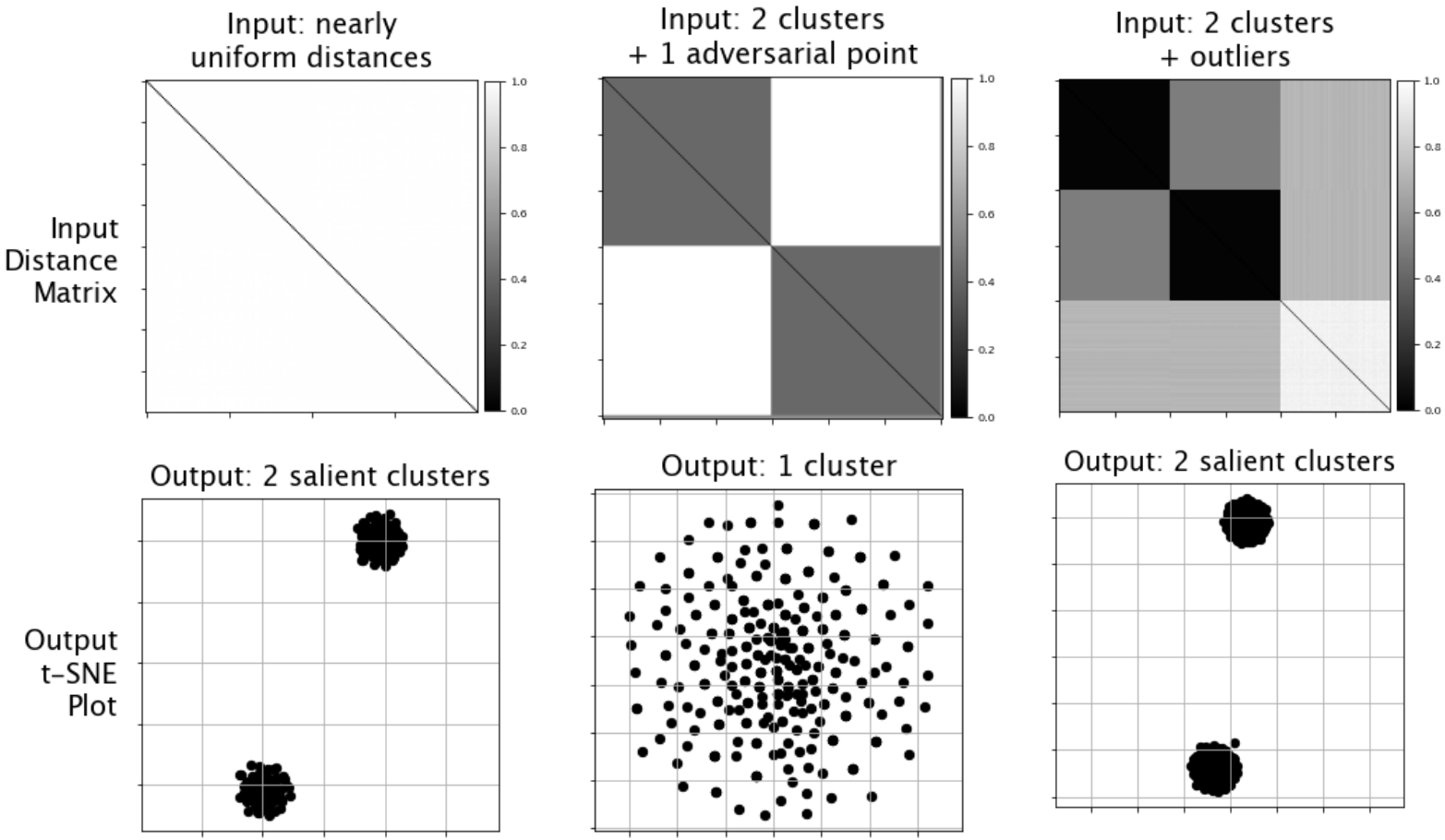
Theoretical Understanding of t-SNE

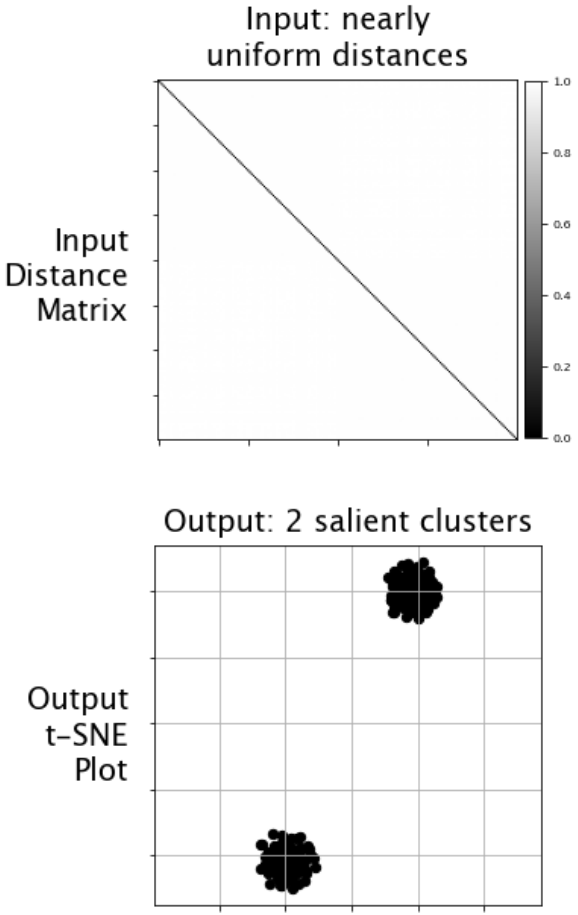
Previous Work [SS'17, LS'17, AHK'18, CM'20]:
If X is clustered, is t-SNE(X) clustered? Yes.
(true positive guarantee)

Our Work: Explaining + quantifying failure modes.
If t-SNE(X) is clustered, is X clustered? No.
(false positive risk)
If t-SNE(X) is un-clustered, is X un-clustered? No.
(false negative risk)

We initiate the theoretical study of t-SNE's failure modes.

(1) False Positive Clusters
(2) False Negative Clusters
(3) Suppression of outliers


Input: nearly uniform distances
Input: 2 clusters + 1 adversarial point
Input: 2 clusters + outliers
Input Distance Matrix
1.0
0.8
0.6
0.4
0.2
0.0
Output: 2 salient clusters
Output: 1 cluster
Output: 2 salient clusters
Output t-SNE Plot
(1) False Positive
(2) False Negative
(3) Outlier Suppression

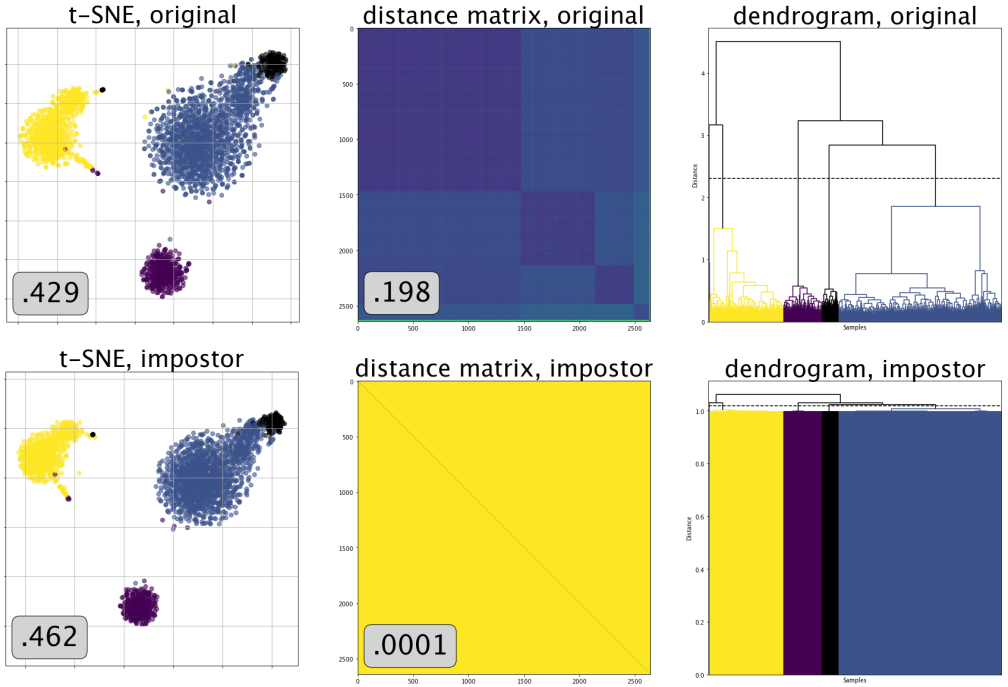


False Positive: Why?

Fact: given a Euclidean distance matrix, if you add a constant to all the distances, the resulting matrix is still a Euclidean distance matrix.

t-SNE is invariant to this operation [LV'11,'14].

Corollary 1: Any dataset has an arbitrarily “un-clustered” impostor with the same t-SNE output.



Corollary 2: all possible t-SNE outputs can be produced by arbitrarily small perturbations of a unit regular simplex.

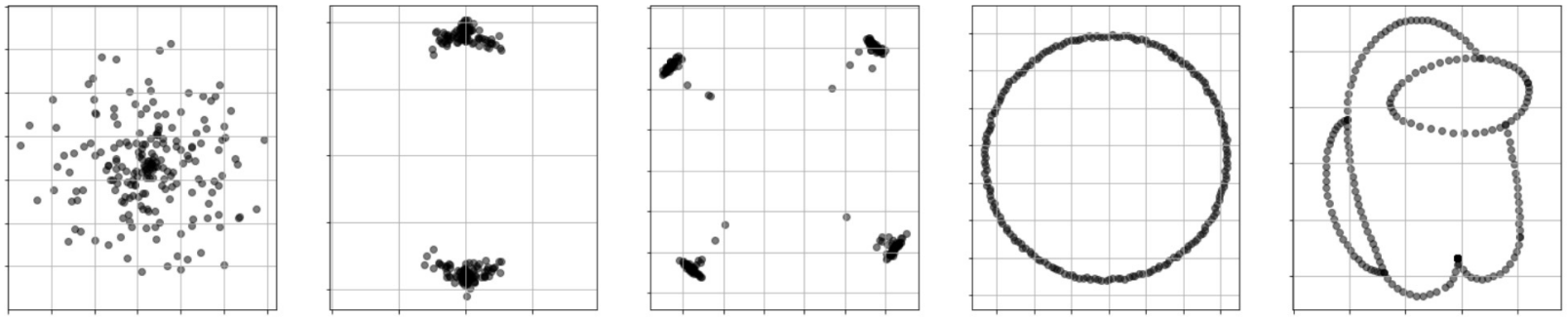
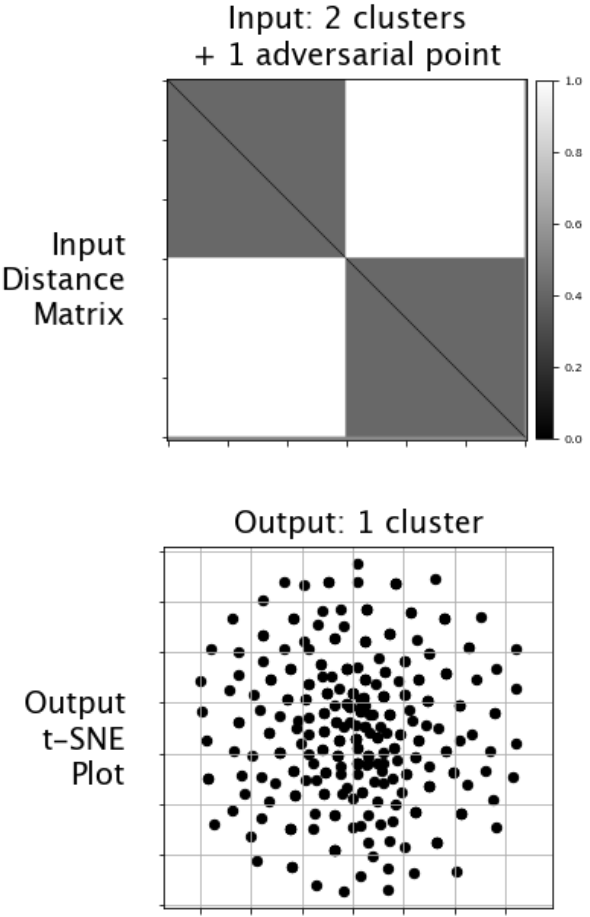


Figure 3: Various different 2D t-SNE visualizations produced by adversarial perturbations of a 200-point unit regular simplex. Each pair of perturbations satisfies the conditions of Theorem 5 for $\epsilon = 0.01$.



(2) False Negative

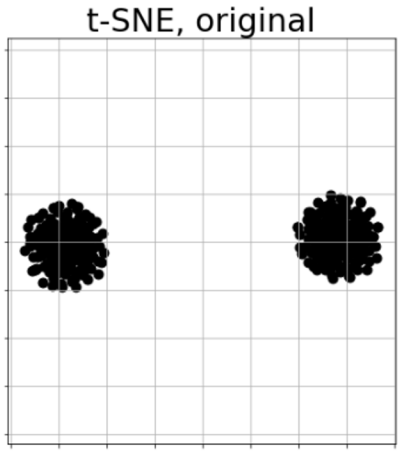
False Negative: Why?

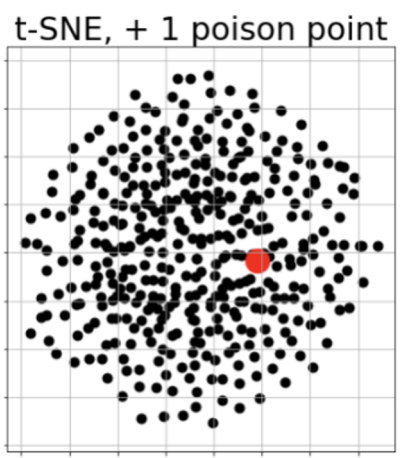
The adversarial point is placed such that it is the nearest neighbor to all other points (a “hub”).

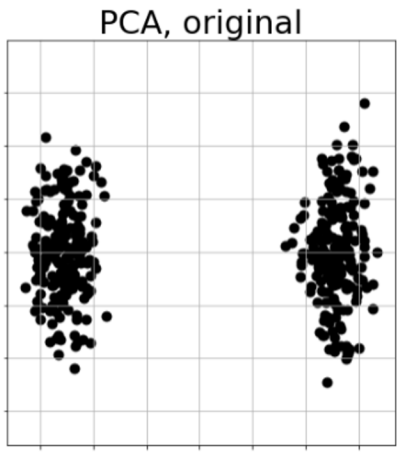
(Hubs only exists for intrinsically high-dimensional, i.e. near-simplex, point sets.)

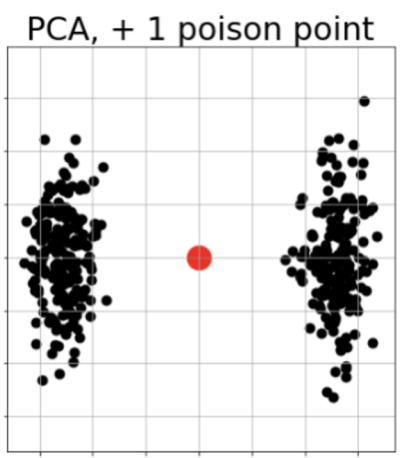
This hub “glues the clusters together.”
(We formalize this in the paper.)

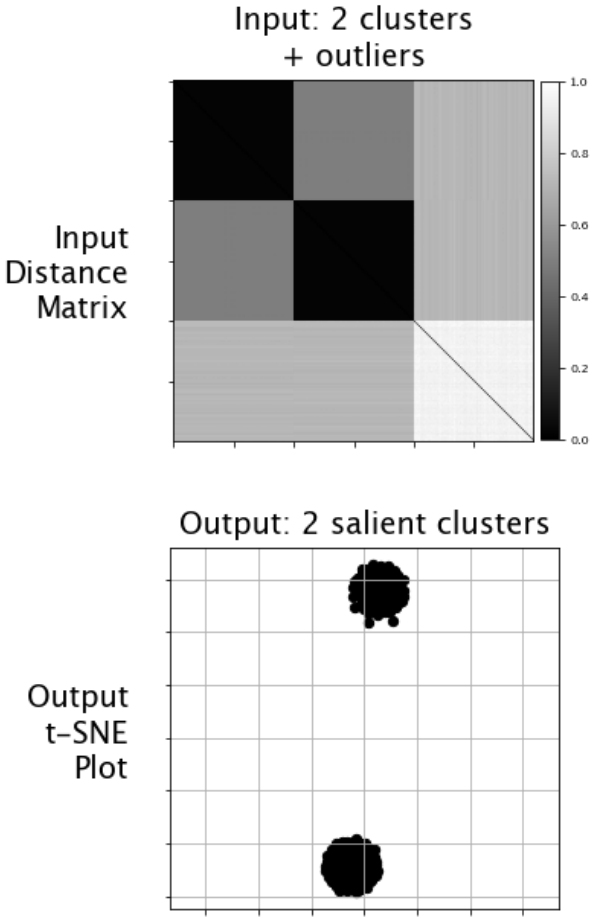
Input dataset: mixture of two well-separated Gaussians











(2) False Negative

Outlier Suppression: Why?

t-SNE does not work directly with the input and output points: it “views” the points through a kernel.

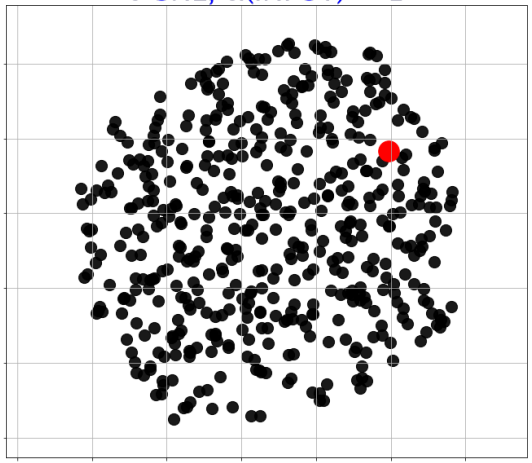
Input Kernel P ≈ nearest-neighbor graph
Output Kernel Q ≈ radius-neighbor graph

t-SNE wants to make Q look like P.
This distorts outliers!

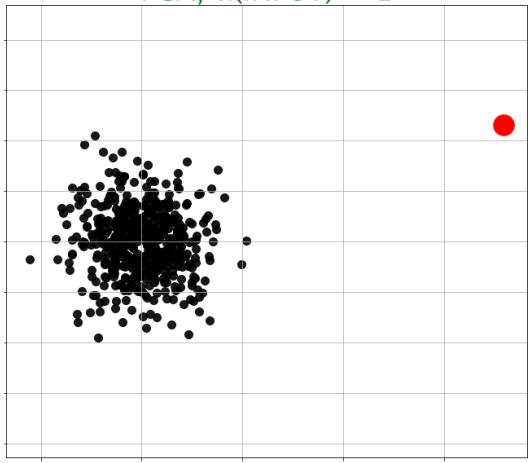
α = “how extreme is the worst outlier”

Synthetic Data: Gaussian plus outlier(s)

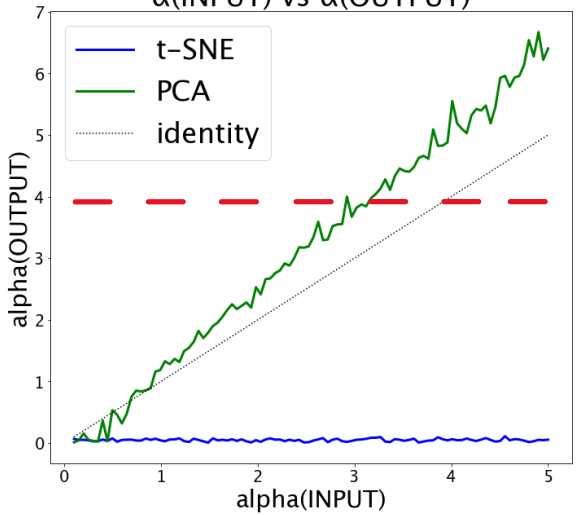
t-SNE, α(INPUT) = 1



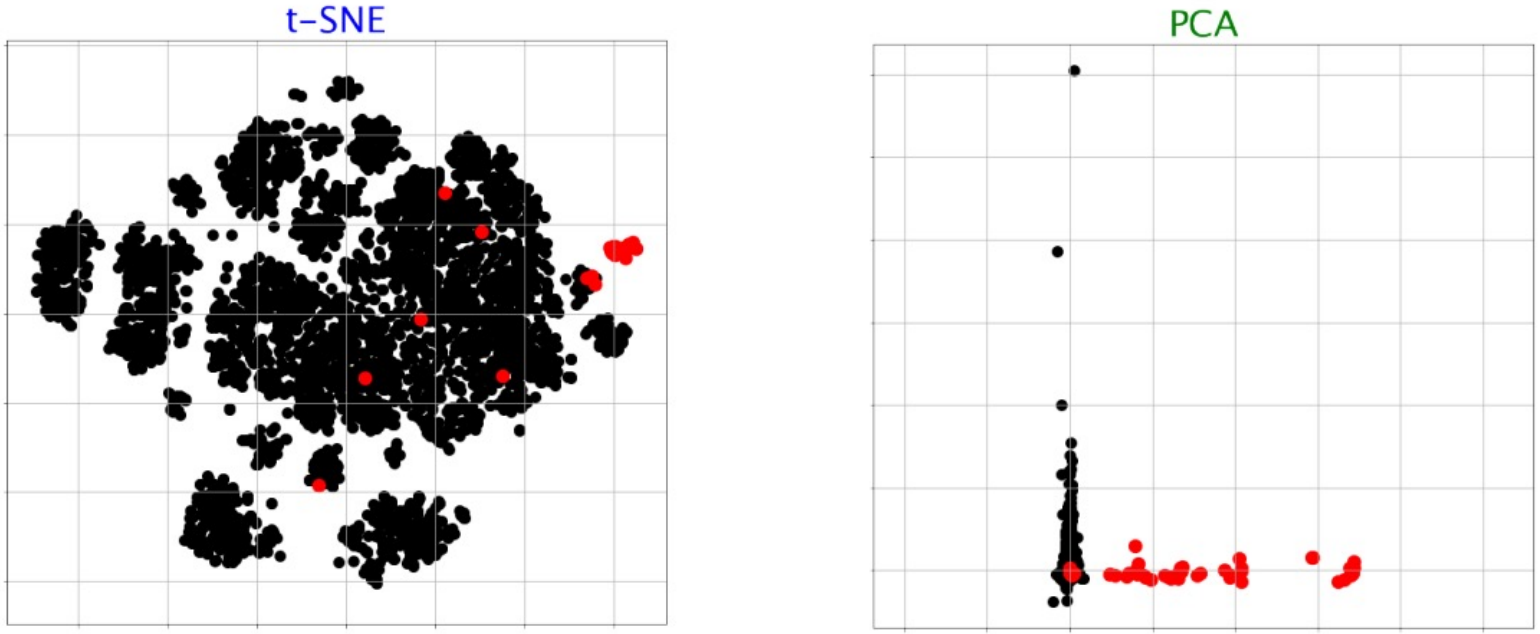
PCA, α(INPUT) = 1



α(INPUT) vs α(OUTPUT)



Theorem: α(any stationary t-SNE embedding) $\leq$ 4



Real-World Example: Credit card fraud dataset (red = fraud)

Discussion

“t-SNE bad? Guess I’ll use UMAP.”

We observe, empirically, that similar failure modes apply.

No free lunch: Data visualization always distorts information!

Important to identify what info is distorted and why.

Theory of data visualization is very underdeveloped!

Thank you!