Visualizing High-Dimensional Datasets with Graph Structures

Sahil Loomba[1], Diogo M. Camacho[1], James J. Collins[1,2]

[1]Wyss Institute for Biologically Inspired Engineering at Harvard University, [2]Massachusetts Institute of Technology

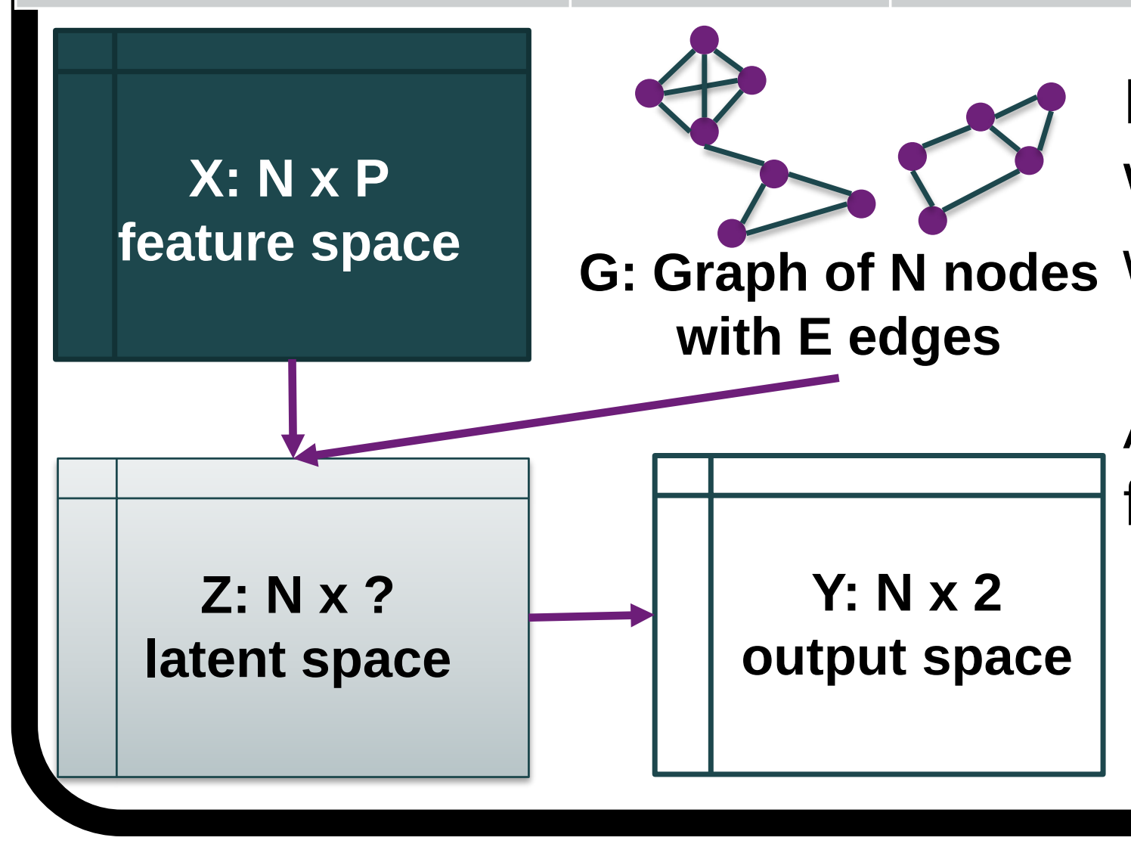
Abstract

High-dimensional datasets have become increasingly common in biology. While a litany of complex statistical and machine learning techniques can be applied to tease out patterns from these data, visualizing the data itself can offer key insights into the data distribution. Methods such as t-SNE (t-Stochastic Neighbor Embedding) provide a means to visualize high-dimensional data by reducing the data to a low dimensional (two or three dimensions) space. As biological datasets are associated with an underlying graph structure, incorporating network knowledge can not only better the visualization of these data, but also allow aberrations in the different biological organization spaces to be seen as perturbations to known biomolecular interactions. Additionally, graphs can encode any arbitrary knowledge such as label assignments as clique graphs, or time information as chain/tree graphs. For the visualization of such graph-based datasets, we extended t-SNE to a more generalized X-t-SNE (Exponential-family-t-Stochastic Neighbor Embedding). Our methodology uses the same Student's t-distribution in the low dimensional space, but generalized exponential family distributions in high dimensions. We principally and sequentially aggregate distributions in the high dimensional space while learning a mapping to the low dimensional space, which allows simultaneous visualization of multiple high dimensional feature spaces and graph structures. We apply our method to visualize abstract datasets such as the Lorenz attractor, popular ML datasets like MNIST handwritten digits, as well as biological datasets like human embryonic stem cell differentiation.

Method

We extend t-SNE into a generalized exponential-family-t-SNE or "X-t-SNE", wherein we impose one of the following exponential family conditional distributions in the feature space $x \sim \exp(-\eta x)$ (η decides the perplexity). But more often than not, a dataset would have a graph structure G *alongside* a continuous feature space X. To combine multiple output spaces, we define an intermediate latent space Z.

Distribution	Variable "x"	Suitable for	Interpretation of Variable "x"	Parameter "η"
Gaussian: $e^{-\|x_i-x_j\|^2/2\sigma_i^2}$	$\|x_i-x_j\|^2$	Continuous features	Euclidean distance b/w i & j in X	$\eta_i = 1/2\sigma_i^2$
Geometric: ${\rho_i}^{\Delta_{ij}}$	Δ_{ij}	Unweighted graphs	Shortest path length b/w i & j on G	$\eta_i = \log(1/\rho_i)$
Exponential: $e^{-\lambda_i \omega_{ij}}$	ω_{ij}	Weighted graphs	Shortest wtd path length b/w i & j on G	$\eta_i = \lambda_i$



Define $P(Z|X,G)$ and objective as $P(\Delta z_{ij}) \approx P(\Delta y_{ij})$

We use an "OR" formulation of Z so that intuitively we extract "union" of information in the two spaces:

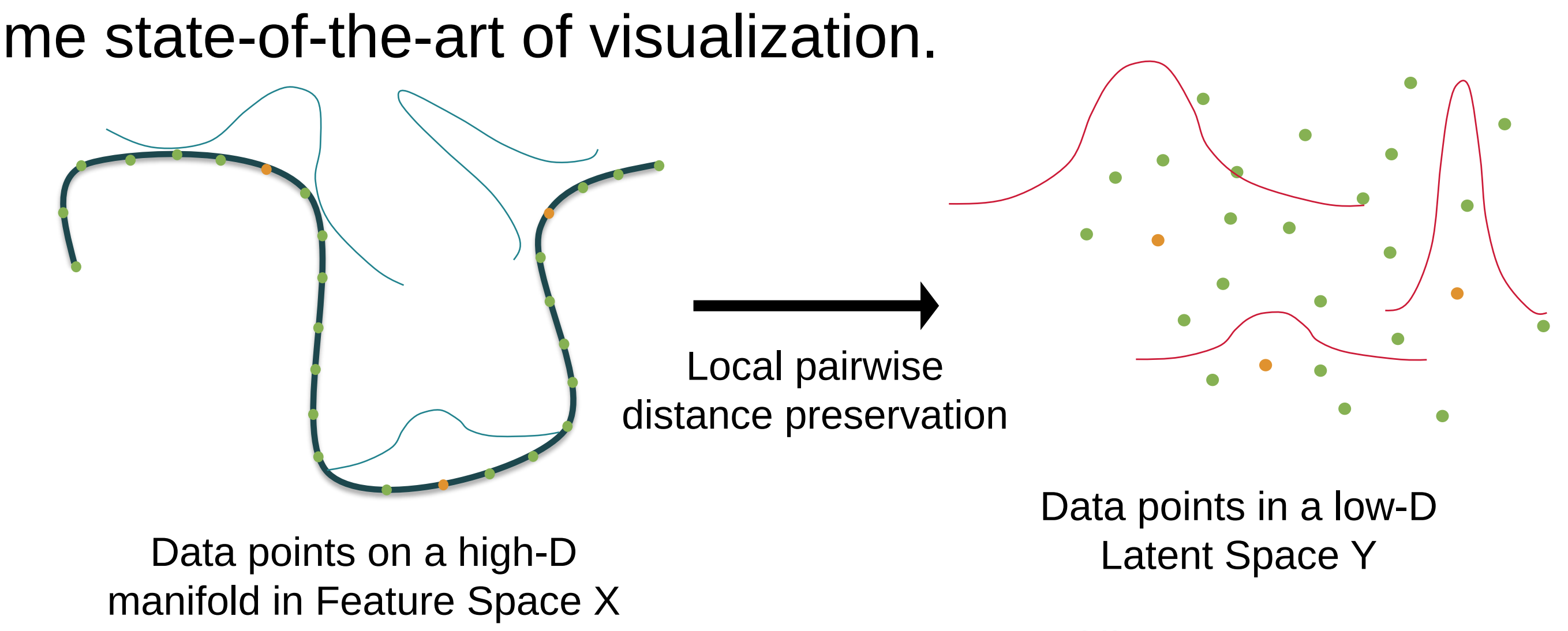
$$P(Z|X,G) = 1 - \left(1-P(X)\right)\left(1-P(G)\right)$$

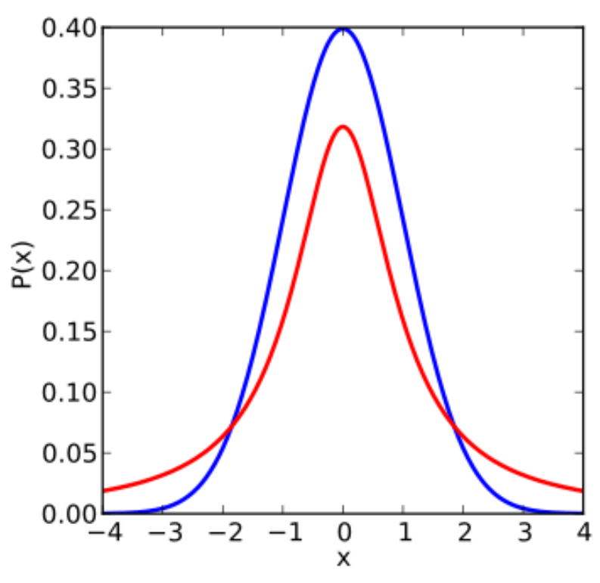
Additionally, we can extend this to S number of feature + graph spaces:

$$P(Z|X^1, X^2, \ldots, X^S) = 1 - \prod_{s=1}^{S}\left(1-P(X^s)\right)$$

Concept

The explosion of high-dimensional data in biology and life sciences has warranted the need of good data visualization algorithms, which can squish the "relevant" information onto just 2 or 3 dimensions. While popular dimensionality techniques such as PCA (which embeds data through a linear transformation while maximizing variance) and Auto-encoders (which do the same but through a non-linear one) can be used for this purpose, algorithms that preserve some notion of a "local pairwise distances" like t-SNE (Maaten & Hinton, 2008) have become state-of-the-art of visualization.


Local pairwise distance preservation
Data points on a high-D manifold in Feature Space X
Data points in a low-D Latent Space Y



Gaussian (blue) kernel in X, Student-t (red) kernel in Y

The idea being encoded is that we must preserve a neighborhood, i.e., points "probably" close in the high-D space must remain "probably" close in the low-D space. We define these distributions by:

- p_{ji}: For every point i in X, place an isotropic Gaussian around it from which every other point j is generated. σ_i of kernel is found such that *perplexity* of conditional distribution is as per the user's requirement.
- q_{ji}: For every point i in Y, place a heavy-tailed distribution, such as the Student-t, from which every other point j is generated.

$$p_{j|i} = \frac{exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{j\neq i} exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}$$

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

$$q_{ij} = \frac{\left(1 + \|x_i - x_j\|^2\right)^{-1}}{\sum_{j\neq i}\left(1 + \|x_i - x_j\|^2\right)^{-1}}$$

The mapping is "learnt" by enforcing that these two sets of distributions remain as identical as possible, by minimizing the distance (KL-divergence) between them.

Objective: $P(\Delta x_{ij}) \approx P(\Delta y_{ij})$

By minimizing KL Divergence:

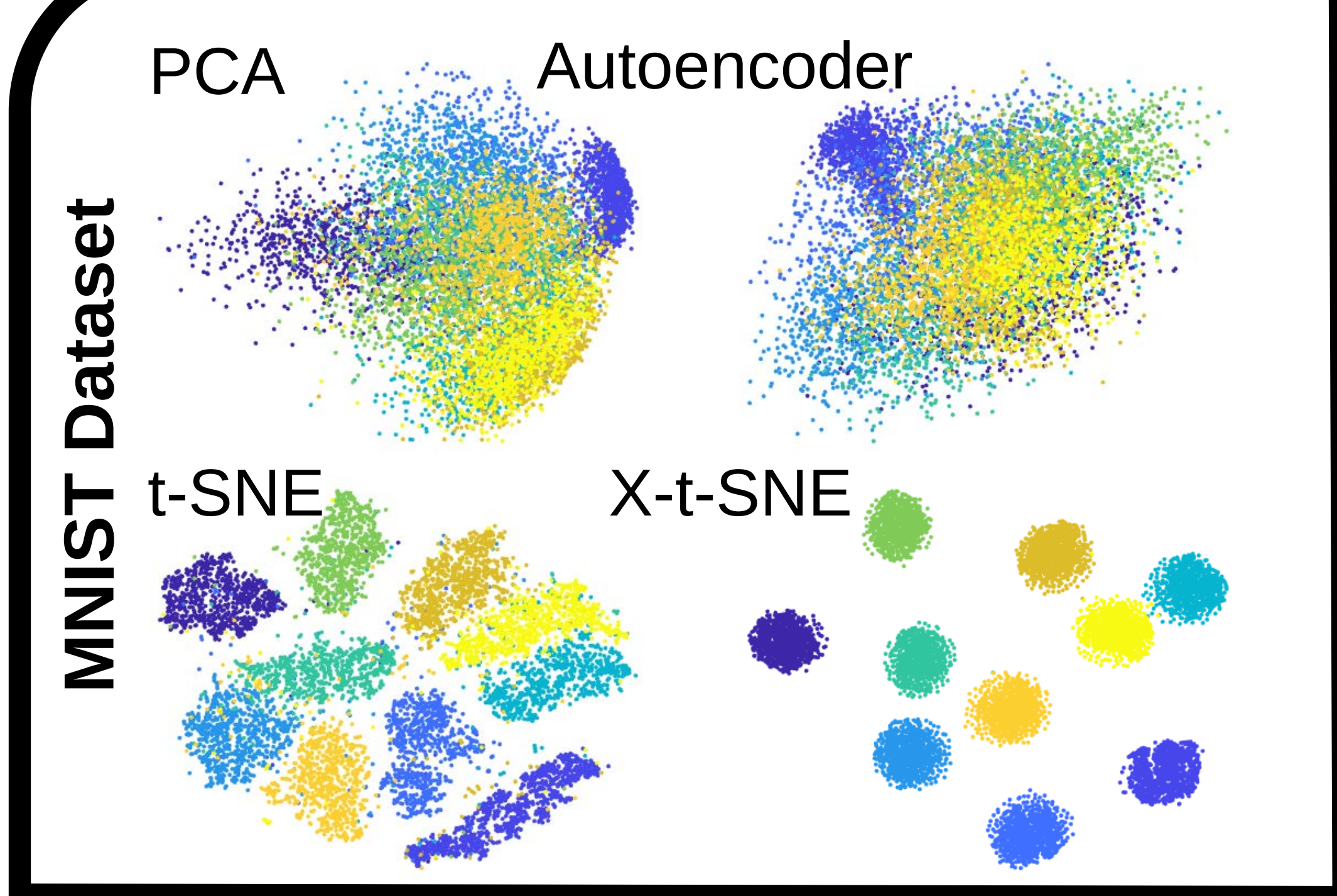
$$KL(P||Q) = \sum_{j\neq i} p_{ij} * log\left(\frac{p_{ij}}{q_{ij}}\right)$$

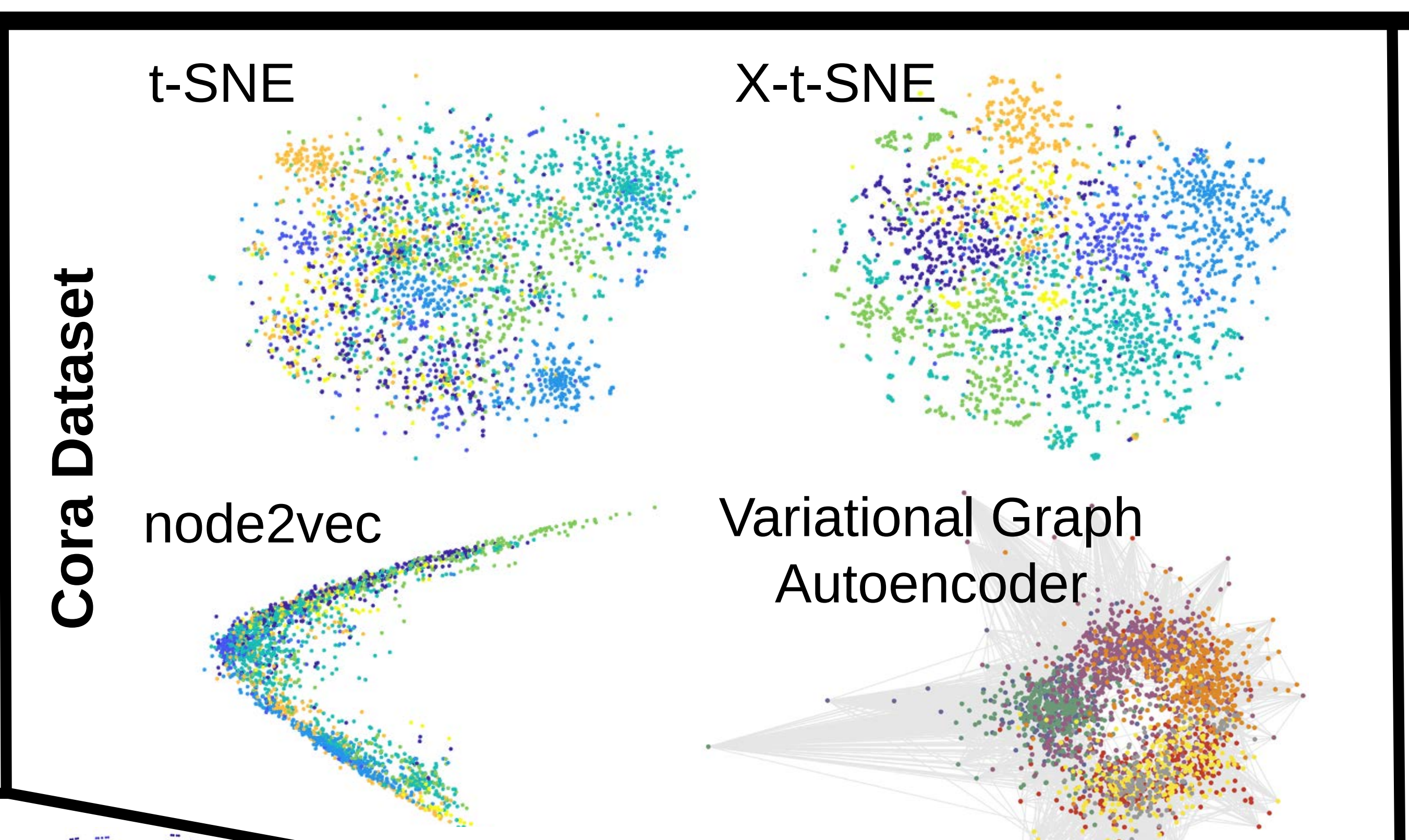
Datasets

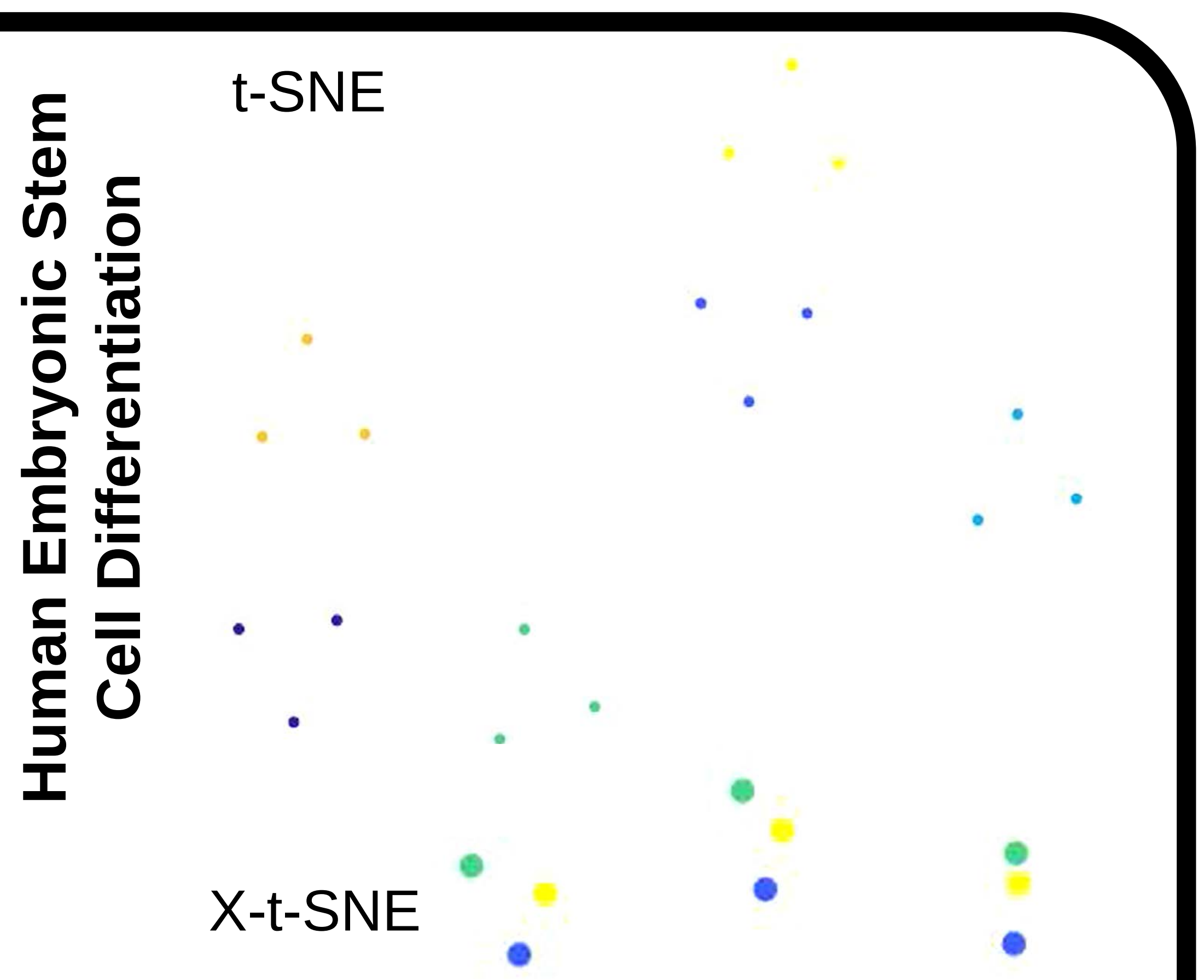
Graph structures can encode arbitrary relationships between data!

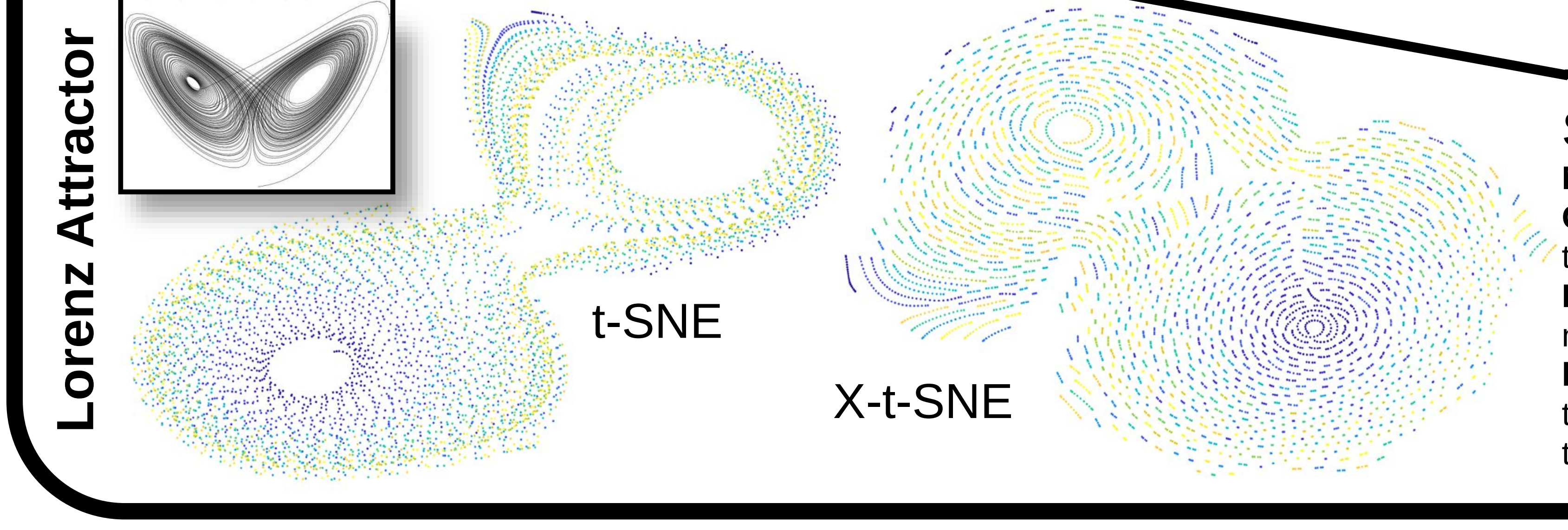
Dataset	# points	Feature Space X	Graph Space G	Labels
MNIST handwritten digits	10000	Black and white pixel values	A clique graph connecting points with same label	Digits 0-9
Cora paper citations	2708	Bag-of-words	Actual citation network	Paper topic type
Human embryonic stem cell differentiation	18	Transcriptomics	A claw graph of cell lineage in time	Cell type
3D Lorenz attractor	10000	Variable space of chaotic Lorenz system	A chain graph of changing variable space in time	Time point

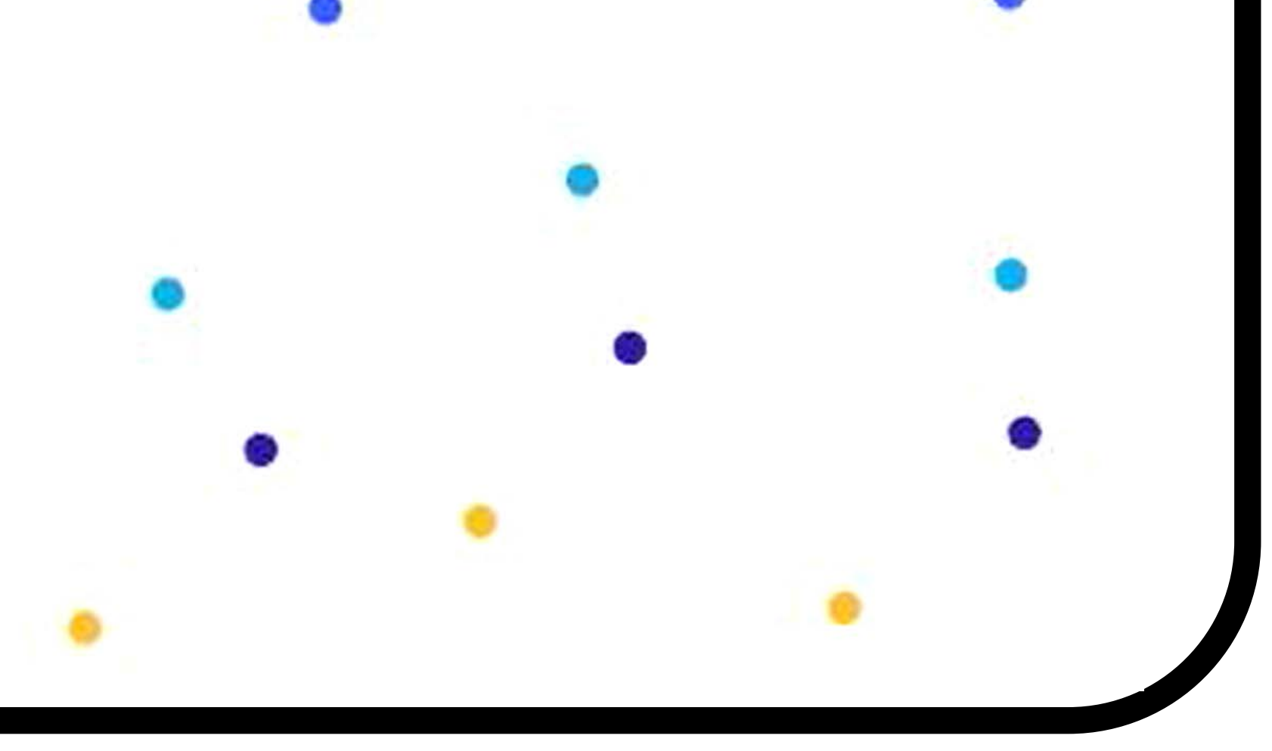
Results











Some interesting observations to note:

MNIST: X-t-SNE gives more clear clustering (semi-supervised)
Cora: X-t-SNE gives best visual clustering, better than state-of-the-art graph embedding algorithms like node2vec and VGAE
Lorenz: X-t-SNE is better able to maintain the butterfly manifold because the chain graph "smooths" the space out
hESC: While t-SNE only captures clusters of same cell type, X-t-SNE is also able to encode the notion of cell differentiation trajectories

Conclusions

We extended t-SNE into a generalized multi-output space method of visualizing data called X-t-SNE, that incorporates graphs to encode any complex relationship between the data being visualized. This has multiple applications in studying biological systems: (1) embedding expression profiles in tissue/tumor/species specific regulatory network contexts, (2) performing multiomics with multigraph structures (using layered X-t-SNEs), (3) tracking cell state evolution in an X-t-SNE landscape, etc. **Grant acknowledgements:** This work is supported by DARPA THoR 15-21.