GENE | X-t-SNE

Graph Enhanced Neighbor Embedding
Exponential and Student-t distributed Stochastic Neighbor Embedding

Visualizing High Dimensional Spaces that exhibit a Graph Structure

SAHIL LOOMBA

Visualizing High-D Data

- Equivalent to “dimensionality reduction” to 2 or 3 dimensions
- PCA: embeds data through a *linear* transformation while preserving *variance*
- Autoencoder: embeds data through a *nonlinear* transformation while preserving *information*[1]
- t-SNE: embeds data through a *nonlinear* transformation while preserving *local distances*[2]
- Manifold assumption: data lies on a smooth low-D manifold

[1]Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *science* 313.5786 (2006): 504-507.

[2]Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of Machine Learning Research* 9.Nov (2008): 2579-2605.

WYSS INSTITUTE

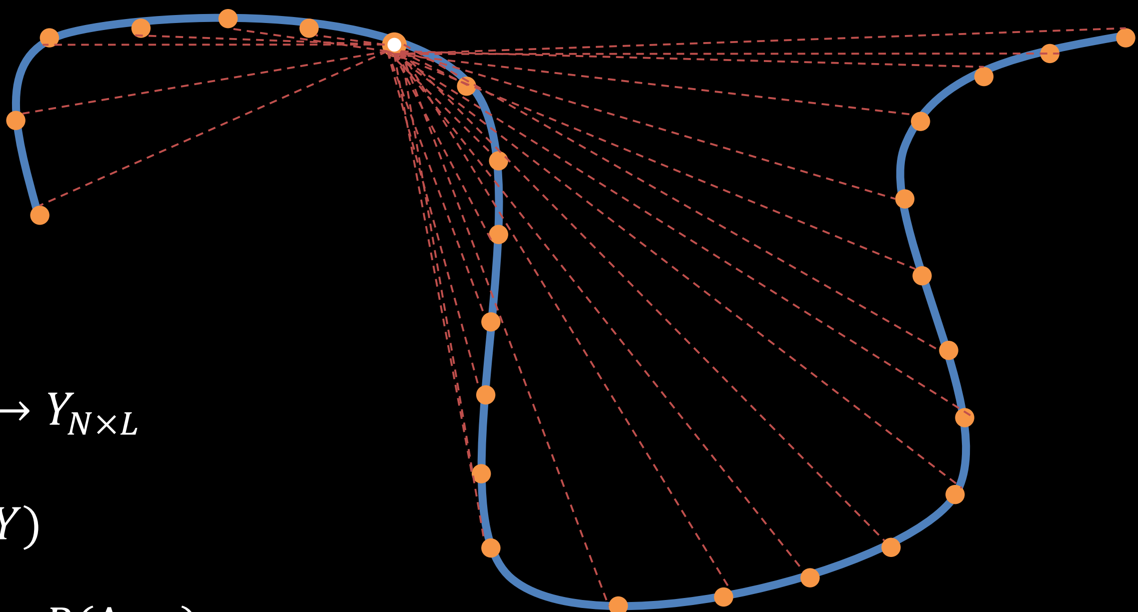
t-Stochastic Neighbor Embedding (t-SNE)

- Objective: preserve "local" distances of the high-D space in the low-D space

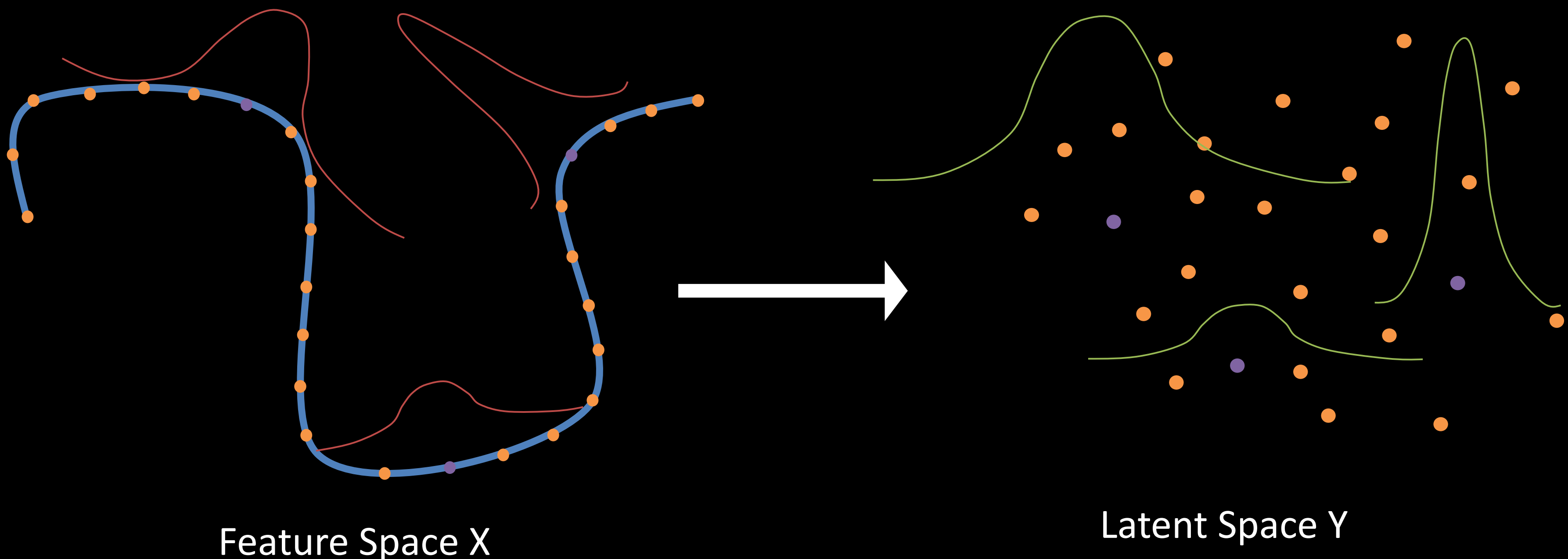
$f: X_{N \times P} \rightarrow Y_{N \times L}$

$P(X), P(Y)$

$P\left(\Delta x_{ij}\right) \approx P\left(\Delta y_{ij}\right)$

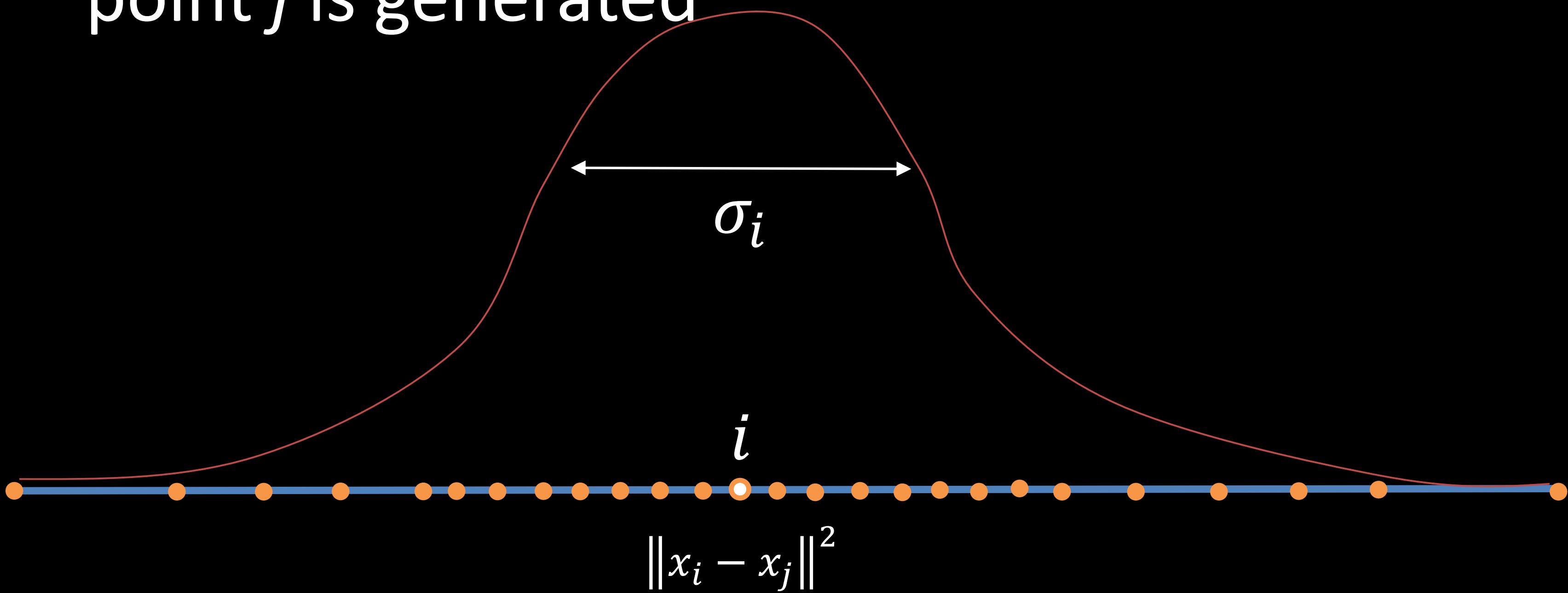


t-Stochastic Neighbor Embedding (t-SNE)



t-Stochastic Neighbor Embedding (t-SNE)

- For every point *i* in *X*, place an *isotropic* Gaussian around it from which every other point *j* is generated

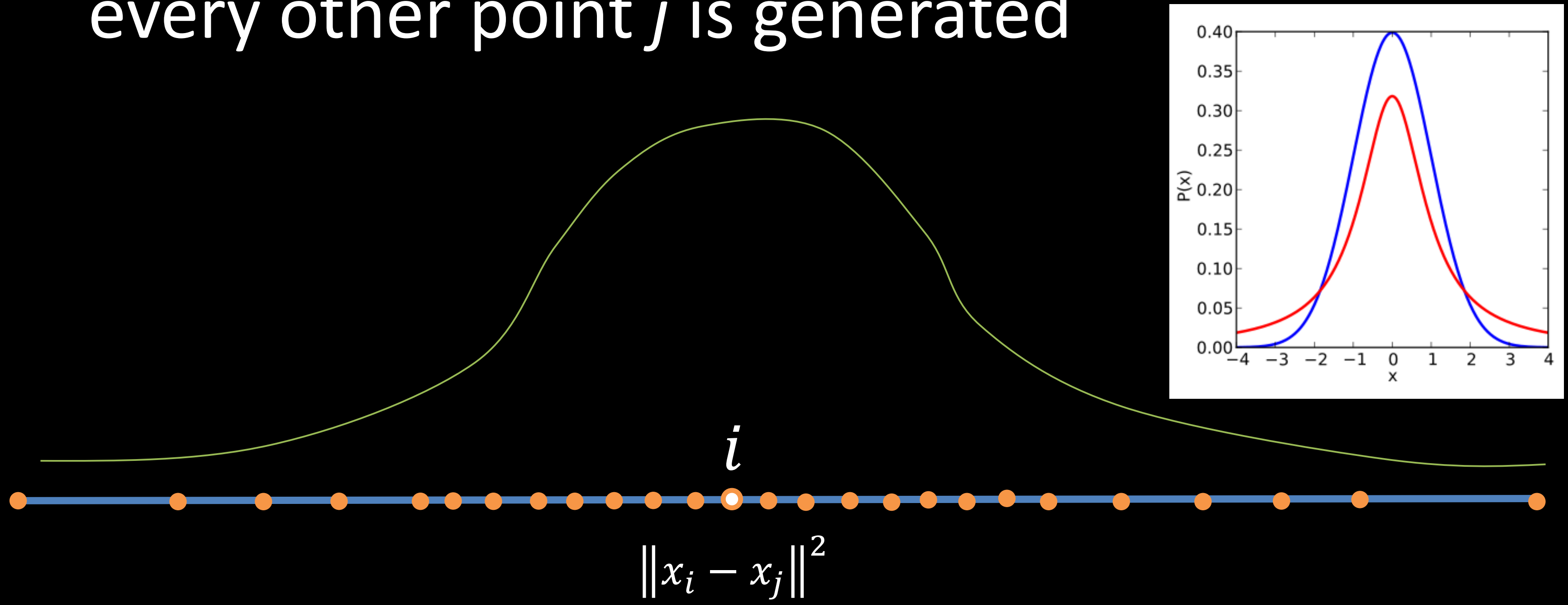


t-Stochastic Neighbor Embedding (t-SNE)

- Defining $P(X)$
  - $p_{j|i} = \frac{exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{j \neq i} exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}$
  - $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2}$
  - σ_i of kernel is found such that *perplexity* of conditional distribution is as per the user's requirement

t-Stochastic Neighbor Embedding (t-SNE)

- For every point *i* in *Y*, place a *heavy-tailed* distribution, such as the Student-t, from which every other point *j* is generated



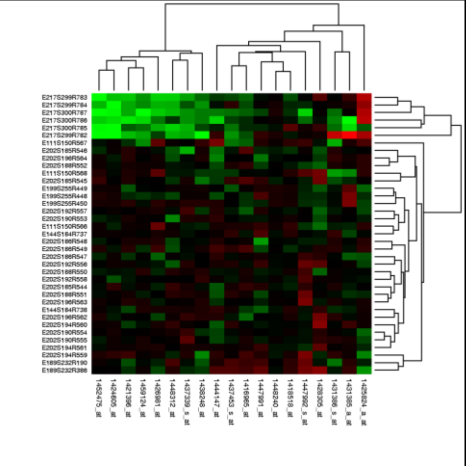
WYSS INSTITUTE

t-Stochastic Neighbor Embedding (t-SNE)

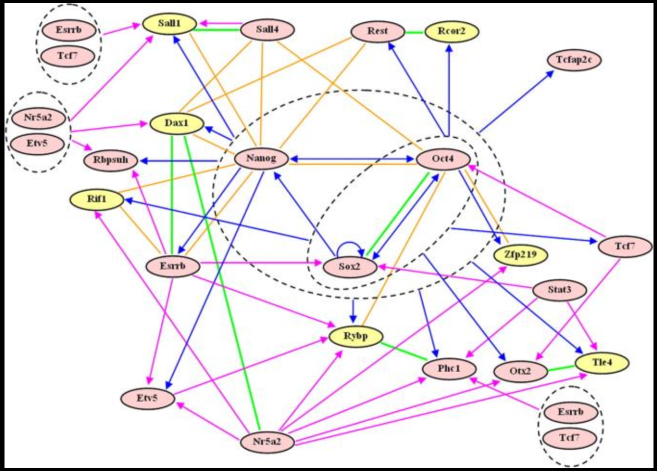
- Defining $P(Y)$
  - $q_{ij} = \frac{\left(1+\|x_i - x_j\|^2\right)^{-\nu}}{\sum_{j \neq i}\left(1+\|x_i - x_j\|^2\right)^{-\nu}}$
- Objective: $P\left(\Delta x_{ij}\right) \approx P(\Delta y_{ij})$
  - KL Divergence: $KL(P||Q) = \sum_{j \neq i} p_{ij} * log\left(\frac{p_{ij}}{q_{ij}}\right)$

Extending t-SNE to graph structured data

- Say we have a graph *G* associated with the data points in feature space *X*
- Can we use *G* as complementary information to *X* to embed the data points?

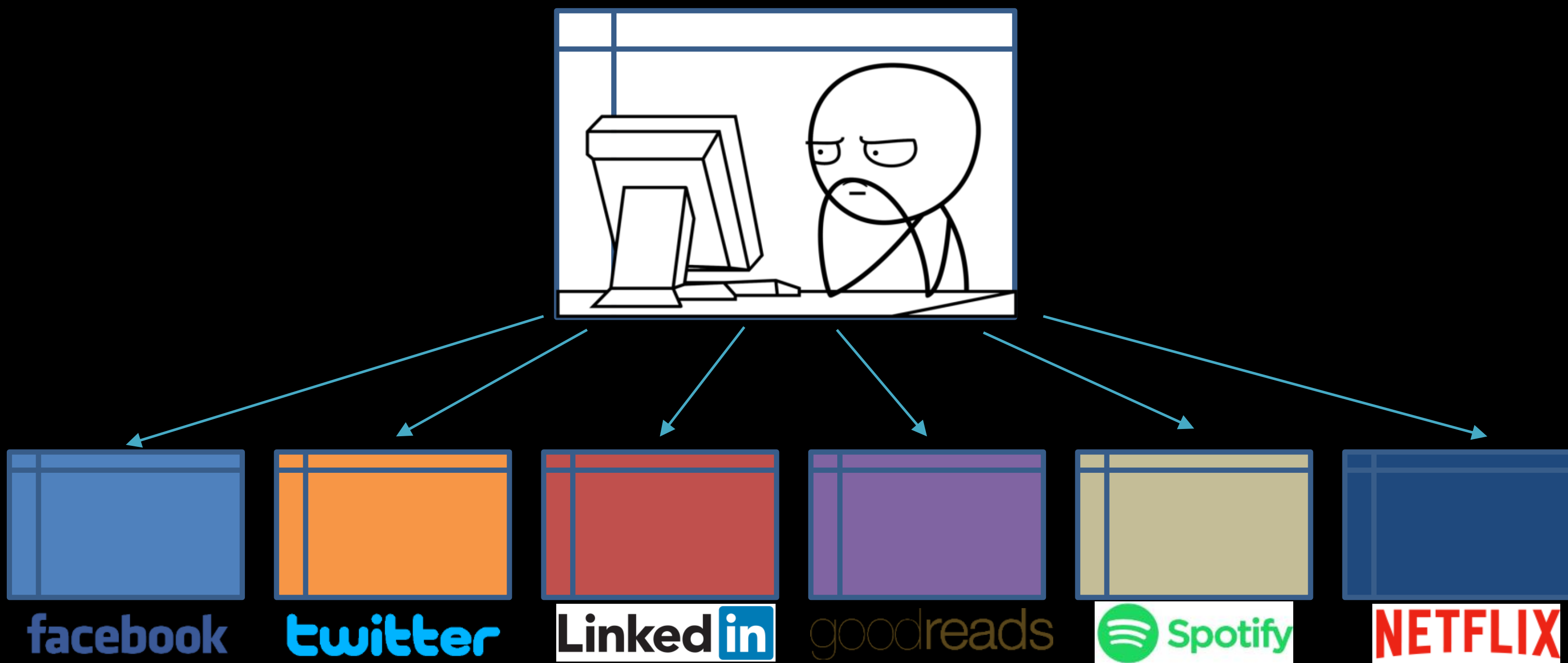


+



Hello, Internet Peeps

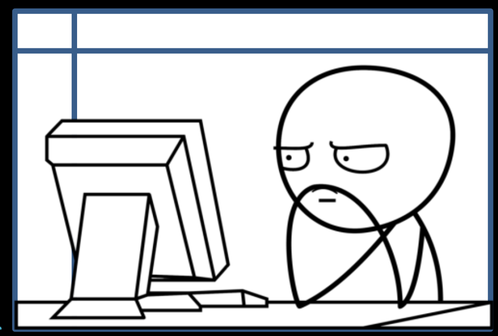
d dimensional latent ("hidden") space of people on the internet

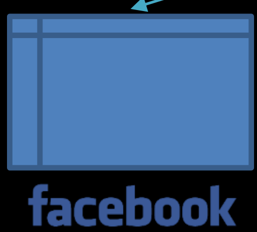


multiple "observed" modalities

Hello, Facebook Peeps

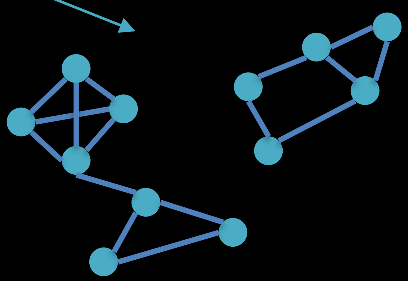
2 dimensional latent ("hidden") space of people on Facebook





interests, likes, preferences, …

N x P feature space



Graph of N nodes with E edges

Extending t-SNE to graph structured data

- $f: X_{N\times P} \times G_{N\times N} \rightarrow Y_{N\times L}$
- $f: X_{N\times P} \times G_{N\times N} \rightarrow Z_{N\times ?} \rightarrow Y_{N\times L}$
- Define $P(Z|X, G)$ and objective is

$$P\left(\Delta z_{ij}\right) \approx P(\Delta y_{ij})$$

- Let as assume Z distributes independently over X and G:

$$P(Z|X, G) = P(X) * P(G)$$

  - Problem 1: How do we define P(G)?
  - Problem 2: How do we define '?' ?
- Well, we need only care for conditional distribution of points!

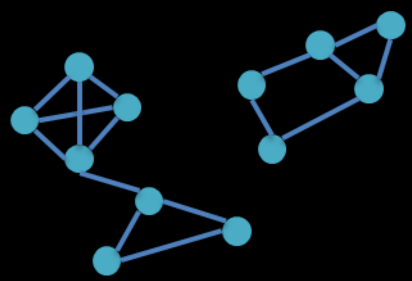
$$P\left(\Delta z_{j|i} | \Delta x_{j|i}, \Delta g_{j|i}\right) = P\left(\Delta x_{j|i}\right) * P(\Delta g_{j|i})$$

WYSS INSTITUTE

Conditional Distributions on Δ in G

- Define "distance between pairs of points *i* and *j*" as the "length of shortest path from *i* to *j*"
- From adjacency matrix A, calculate shortest path matrix Δ using Floyd–Warshall algorithm
- Set $\Delta g_{ij} = mean(\Delta g) + 1$

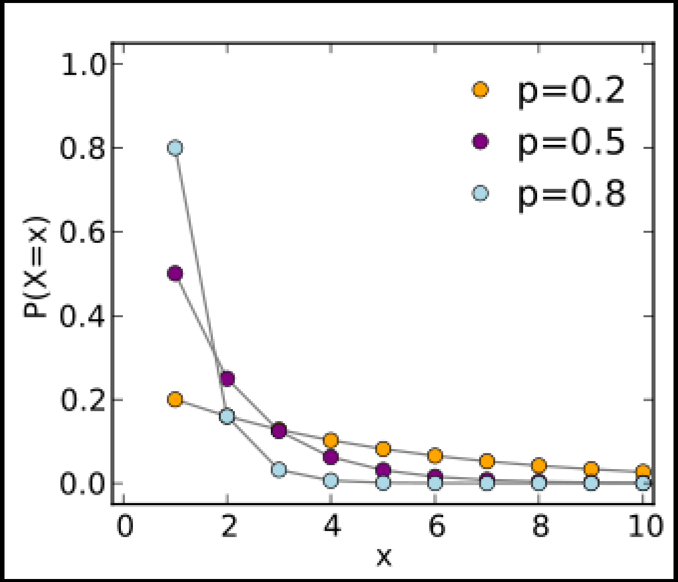
  if $\Delta g_{ij} > mean(\Delta g)$



(mean diameter of G; robust)

Conditional Distributions on Δ in G

- We use the geometric distribution for the kernel at point *i*
- $p_{j|i} = \frac{{\rho_i}^{\Delta_{ij}}}{\sum_{j \neq i} {\rho_i}^{\Delta_{ij}}}$
- ρ_i of kernel is found such that *perplexity* of conditional distribution is equal to degree of node *i* (number of immediate neighbors)



Total Joint Distributions on Δ in Z

- $P\left(\Delta z_{j|i} \middle| \Delta x_{j|i}, \Delta g_{j|i}\right) = P\left(\Delta x_{j|i}\right) * P(\Delta g_{j|i})$
- $p_{j|i} = \frac{exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right) * {\rho_i}^{\Delta_{ij}}}{\sum_{j \neq i} exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right) * \sum_{j \neq i} {\rho_i}^{\Delta_{ij}}}$
- $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2}$
- PROBLEM: the vanishing conditional

Extending t-SNE to graph structured data (attempt 2)

- Earlier, we had an "AND" space:

$$P(Z|X,G) = P(X) * P(G)$$

- Let's assume an "OR" space instead:

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

$$P(Z|X,G) = P(X) + P(G) - P(X) * P(G)$$

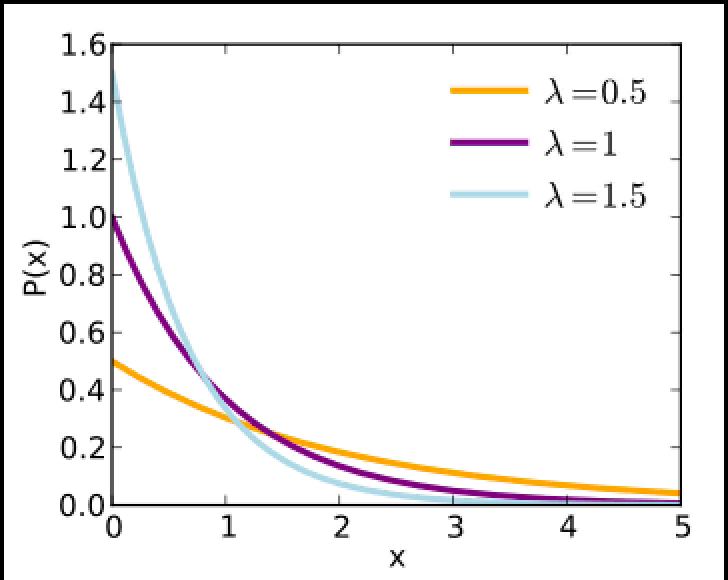
- Intuitively, we now extract "union" of information in the two spaces, rather than the "intersection"

Total Joint Distributions on Δ in Z

- $P\left(\Delta z_{j|i} | \Delta x_{j|i}, \Delta g_{j|i}\right) = P\left(\Delta x_{j|i}\right) + P\left(\Delta g_{j|i}\right) - P\left(\Delta x_{j|i}\right) * P\left(\Delta g_{j|i}\right)$
- $p_{j|i} = \frac{exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{j \neq i} exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)} + \frac{{\rho_i}^{\Delta_{ij}}}{\sum_{j \neq i} {\rho_i}^{\Delta_{ij}}} - \frac{exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right) * {\rho_i}^{\Delta_{ij}}}{\sum_{j \neq i} exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right) * \sum_{j \neq i} {\rho_i}^{\Delta_{ij}}}$
- $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2}$

What if Graph is weighted?

- Floyd–Warshall algorithm outputs a weighted shortest path matrix Ω
- Conditional distributions on Ω in G: we use the exponential distribution for the kernel at point *i*
  - $p_{j|i} = \frac{e^{-\lambda_i \omega_{ij}}}{\sum_{j \neq i} e^{-\lambda_i \omega_{ij}}}$
  - λ_i of kernel is found such that *perplexity* of conditional distribution is equal to degree of node *i* (number of immediate neighbors)



Generalized Extension of t-SNE

- All of these cases reduce to a generic eXponential family of conditional distributions in the high-D space: $\exp(-\eta x)$

Distribution	Variable (X)	Natural Parameter (η)
Gaussian: $e^{-\|x_i - x_j\|^2/2\sigma_i^2}$	$\|x_i - x_j\|^2$	$\eta_i = 1/2\sigma_i^2$
Geometric: ${\rho_i}^{\Delta_{ij}}$	Δ_{ij}	$\eta_i = \log(1/\rho_i)$
Exponential: $e^{-\lambda_i \omega_{ij}}$	ω_{ij}	$\eta_i = \lambda_i$

- Additionally, we can extend to *S* number of feature + graph spaces

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

Advantages of a Graph Structure

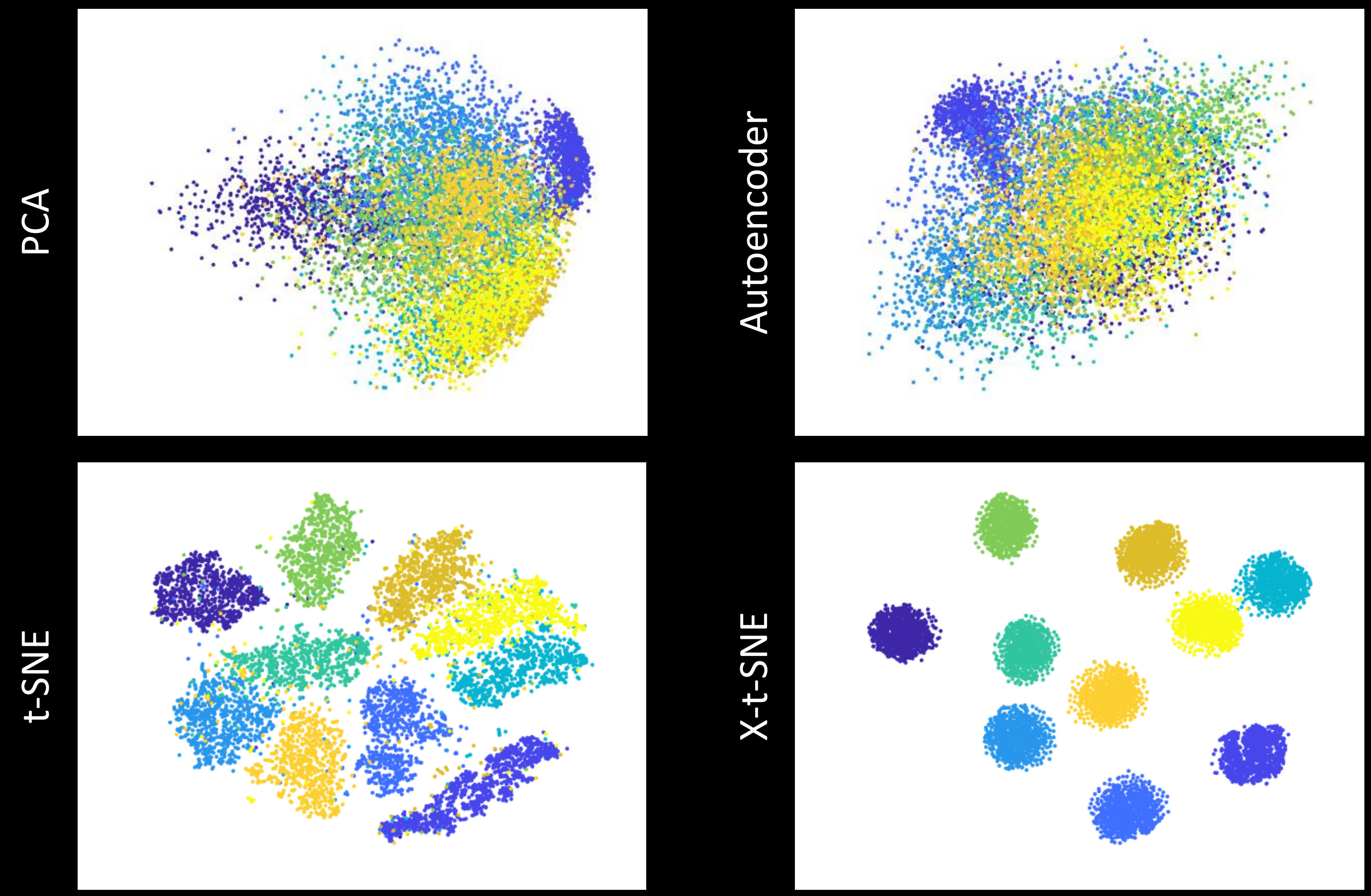
- Can encode arbitrarily complex relationships between data!
  - (Semi-)supervised learning: use clique graphs
  - Encode time-series: use chain/tree graphs
  - Encode multi-range correlations
- Combine disjointed feature spaces through graph bridges
  - "Graph-assisted" transfer learning

Experiments

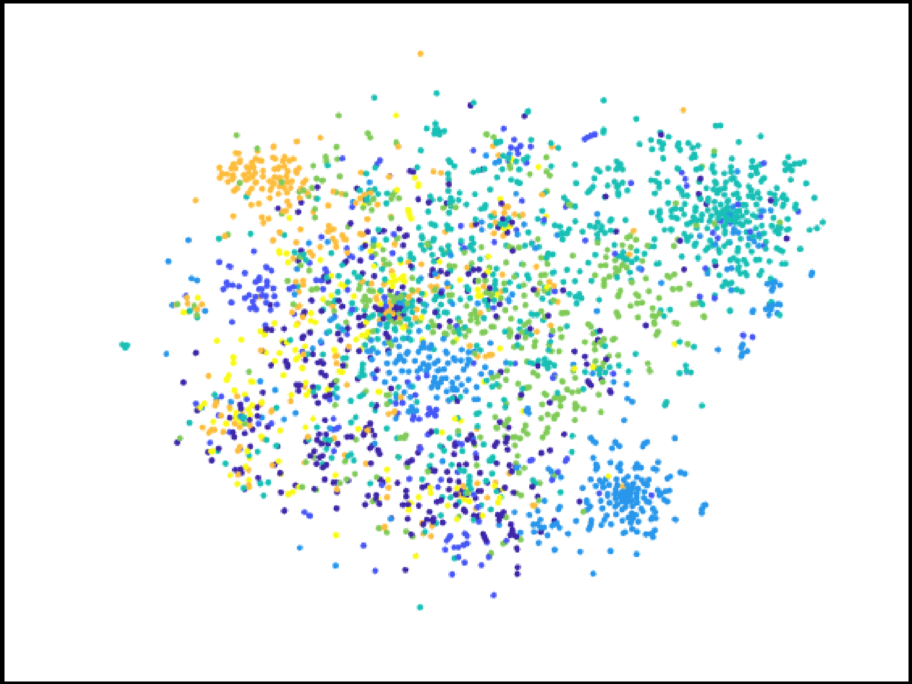
- MNIST Dataset with feature space of 10,000 28x28 b/w images of handwritten digits 0-9 and graph is a clique graph of supervision
- Citation Datasets with bag-of-words feature space of papers and citation networks as graphs:
  - Cora: 2708 papers, 7 paper types
  - Citeseer: 3312 papers, 6 paper types
- RECON2 “virtual” metabolic state Dataset[1]:
  - 2140 genes in a feature space of flux differences induced by single on/off perturbation across 7440 reactions
  - Gene co-participation graph, labeled by the most popular “subsystem” a gene participates in (91 subsystems)
  - Combos of 2s and 3s for random KOs (5000 each, total of 12140 “gene combos”)
- Lorenz attractor
  - Time is encoded as a chain-graph

[1]Thiele, Ines, et al. "A community-driven global reconstruction of human metabolism." *Nature biotechnology* 31.5 (2013): 419-425.

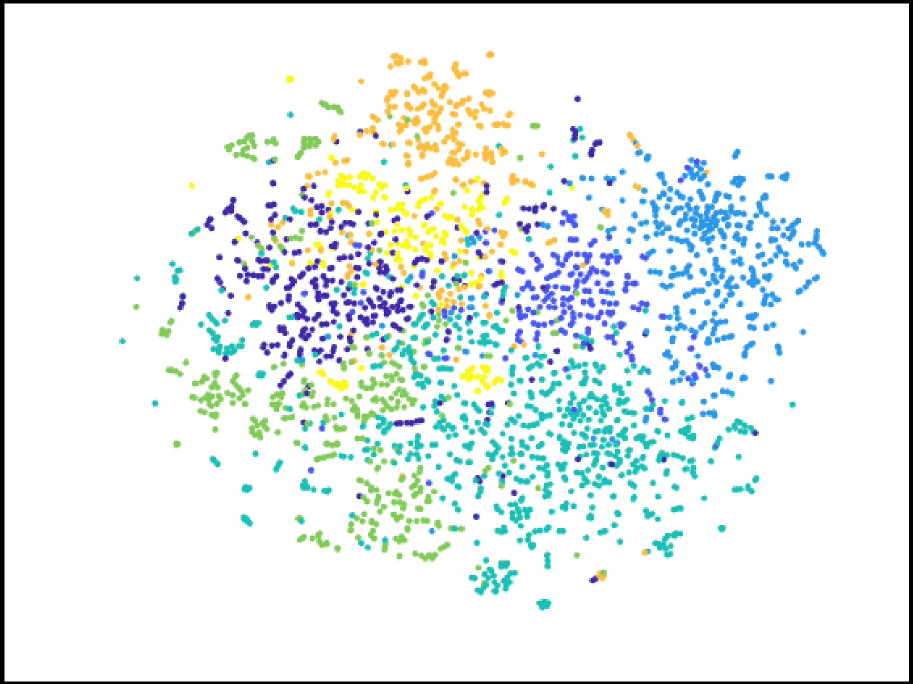
Experiment on MNIST Dataset



Experiment on Cora Dataset

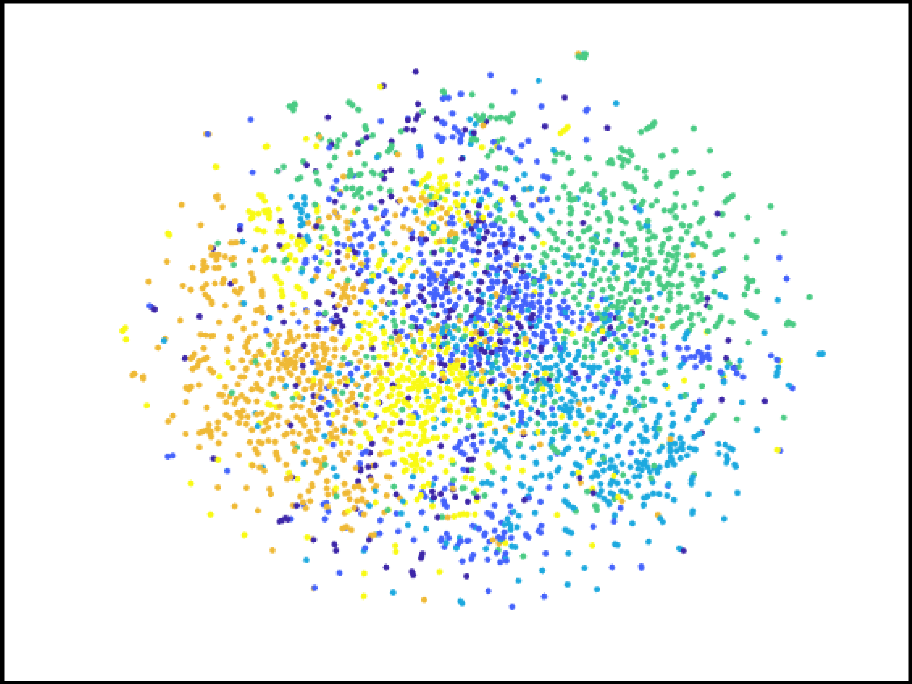


t-SNE

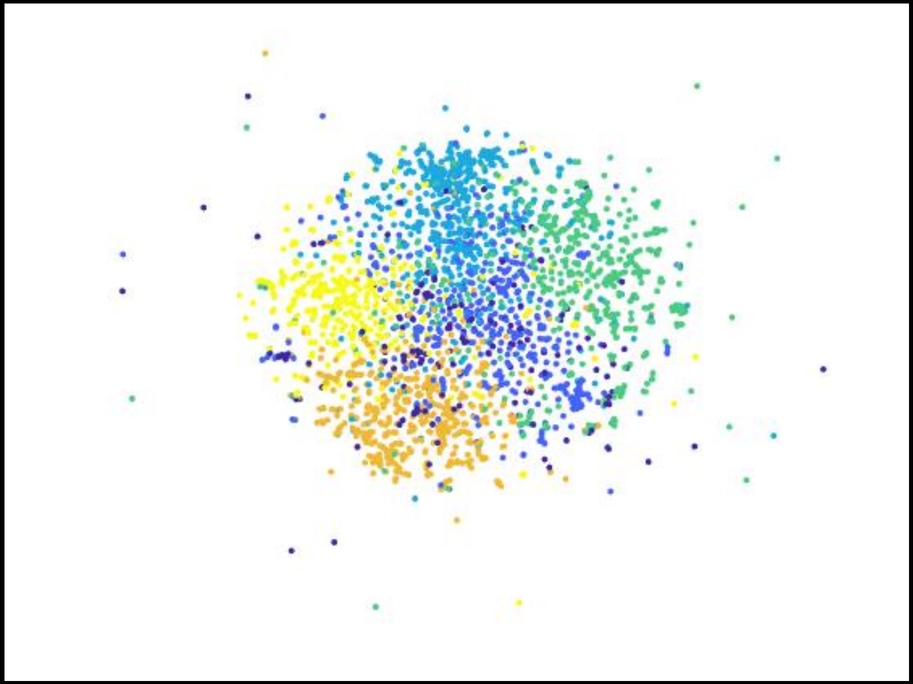


X-t-SNE

Experiment on Citeseer Dataset

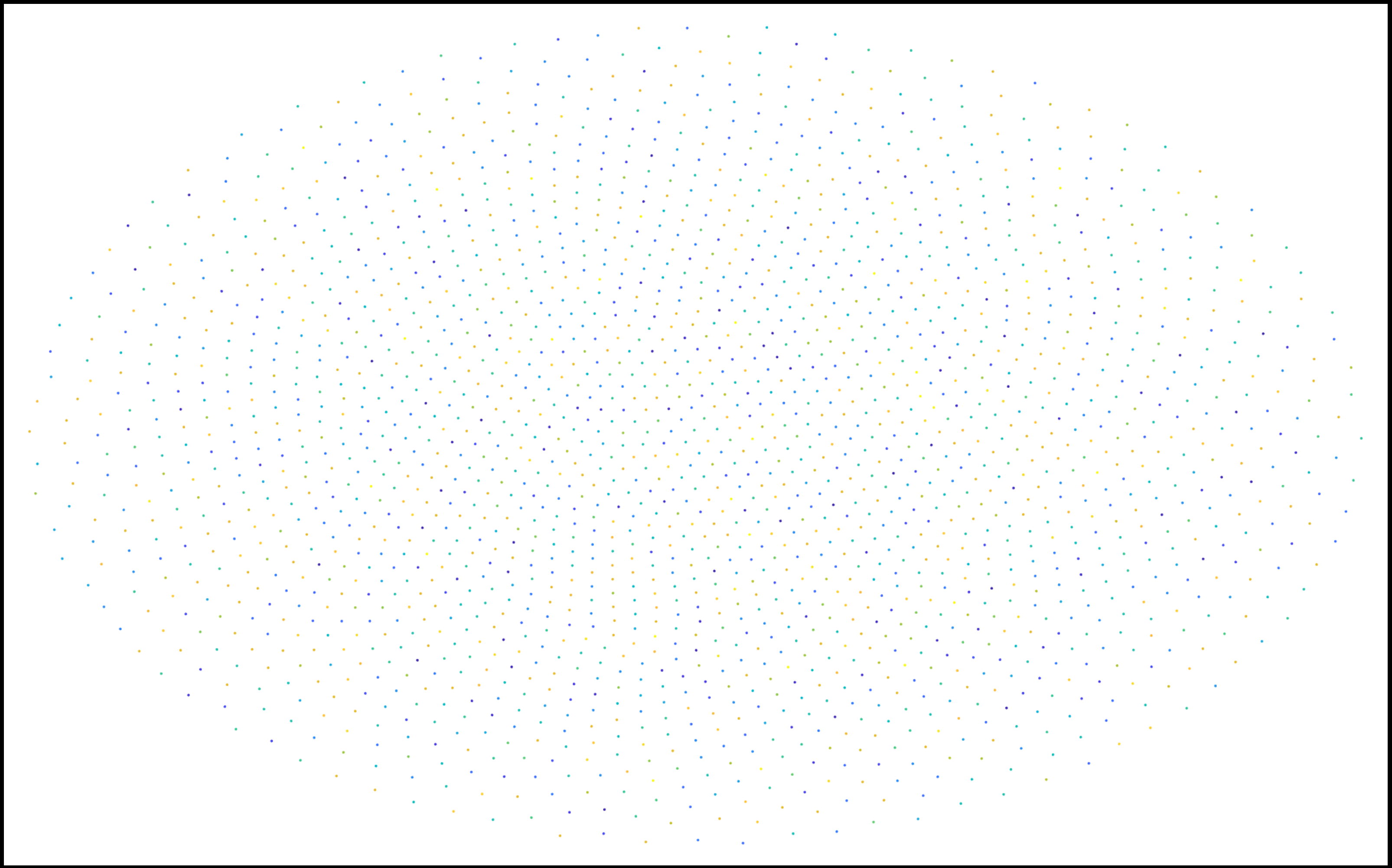


t-SNE



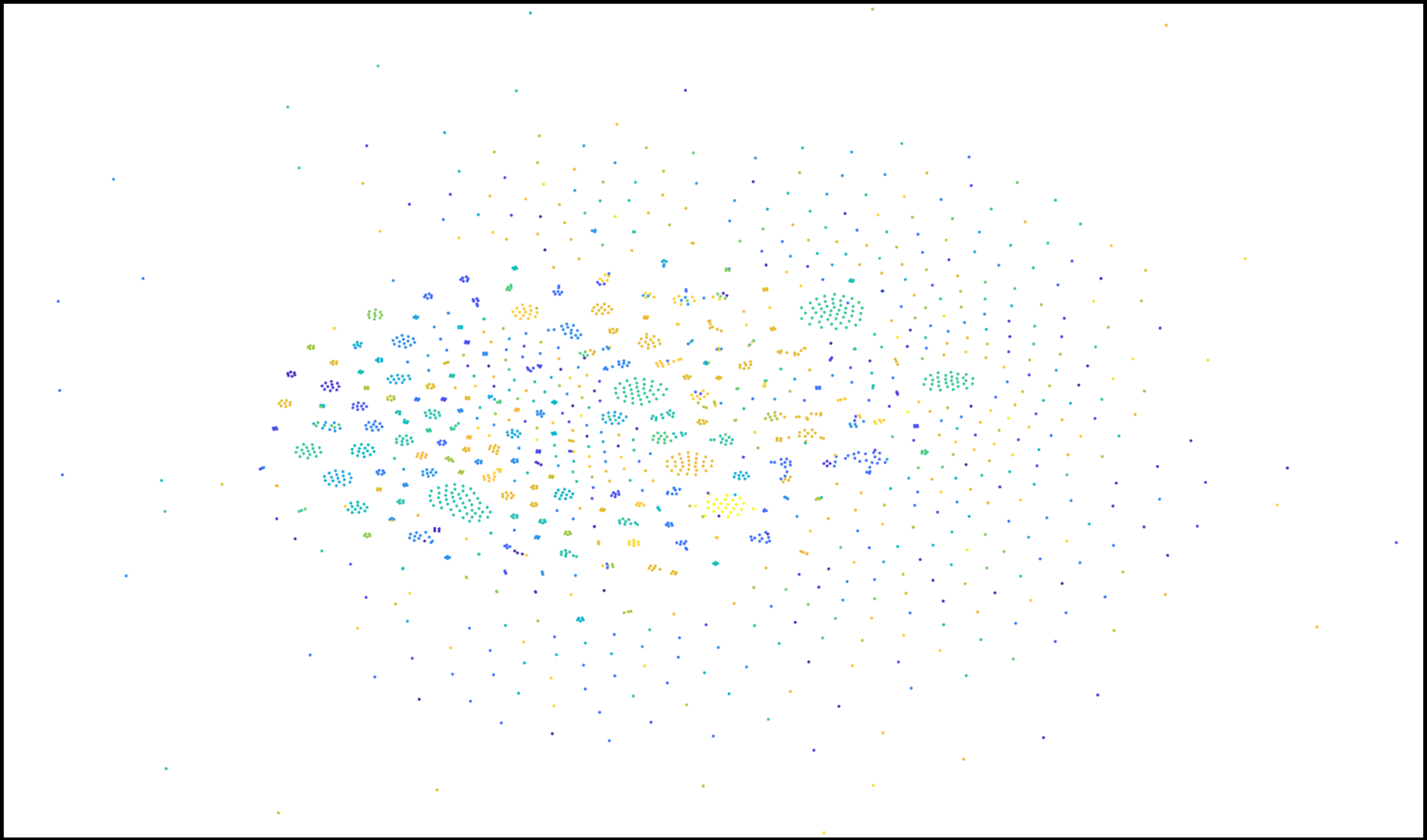
X-t-SNE

Experiment on RECON2



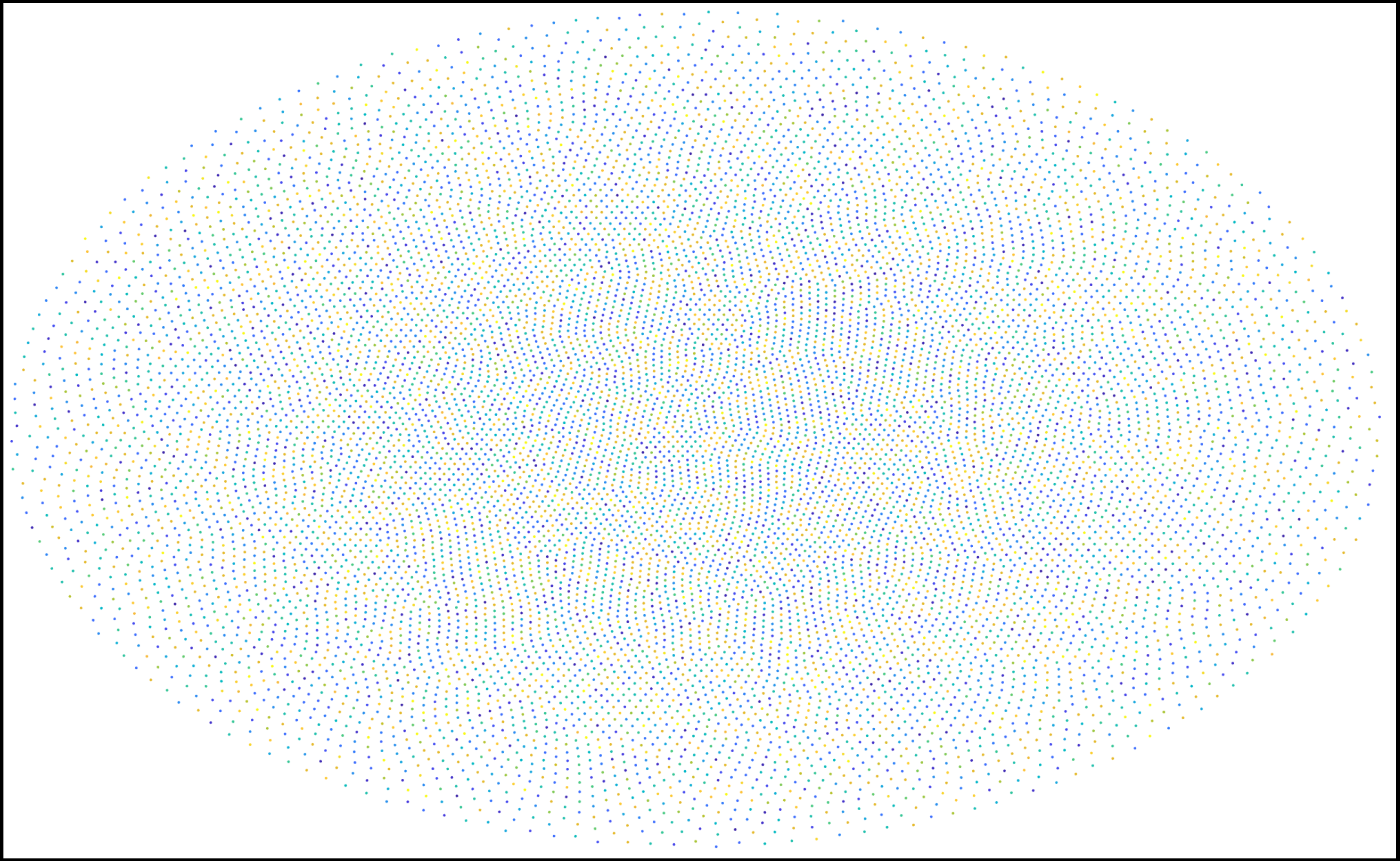
t-SNE

Experiment on RECON2



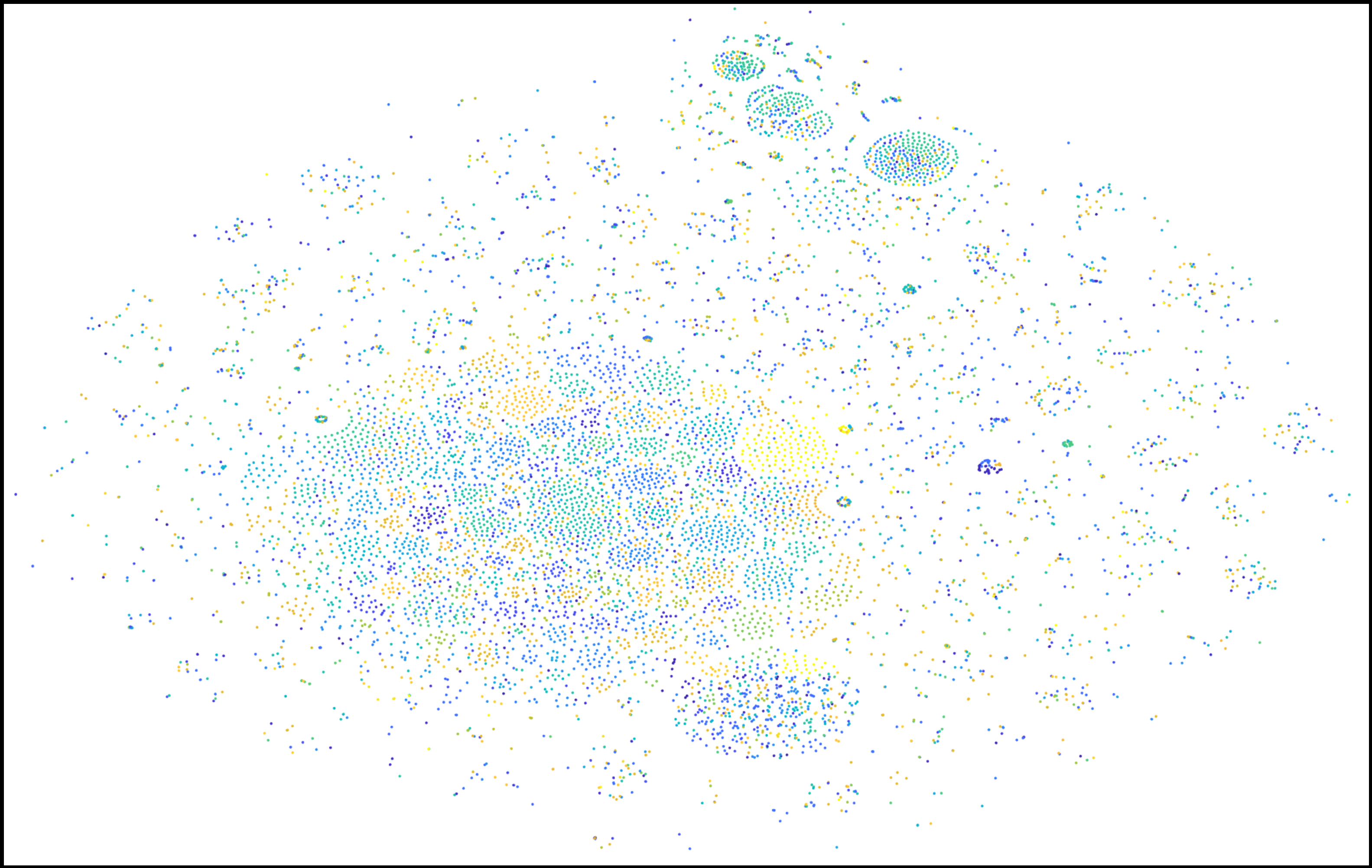
X-t-SNE

Experiment on RECON2 Combos



t-SNE

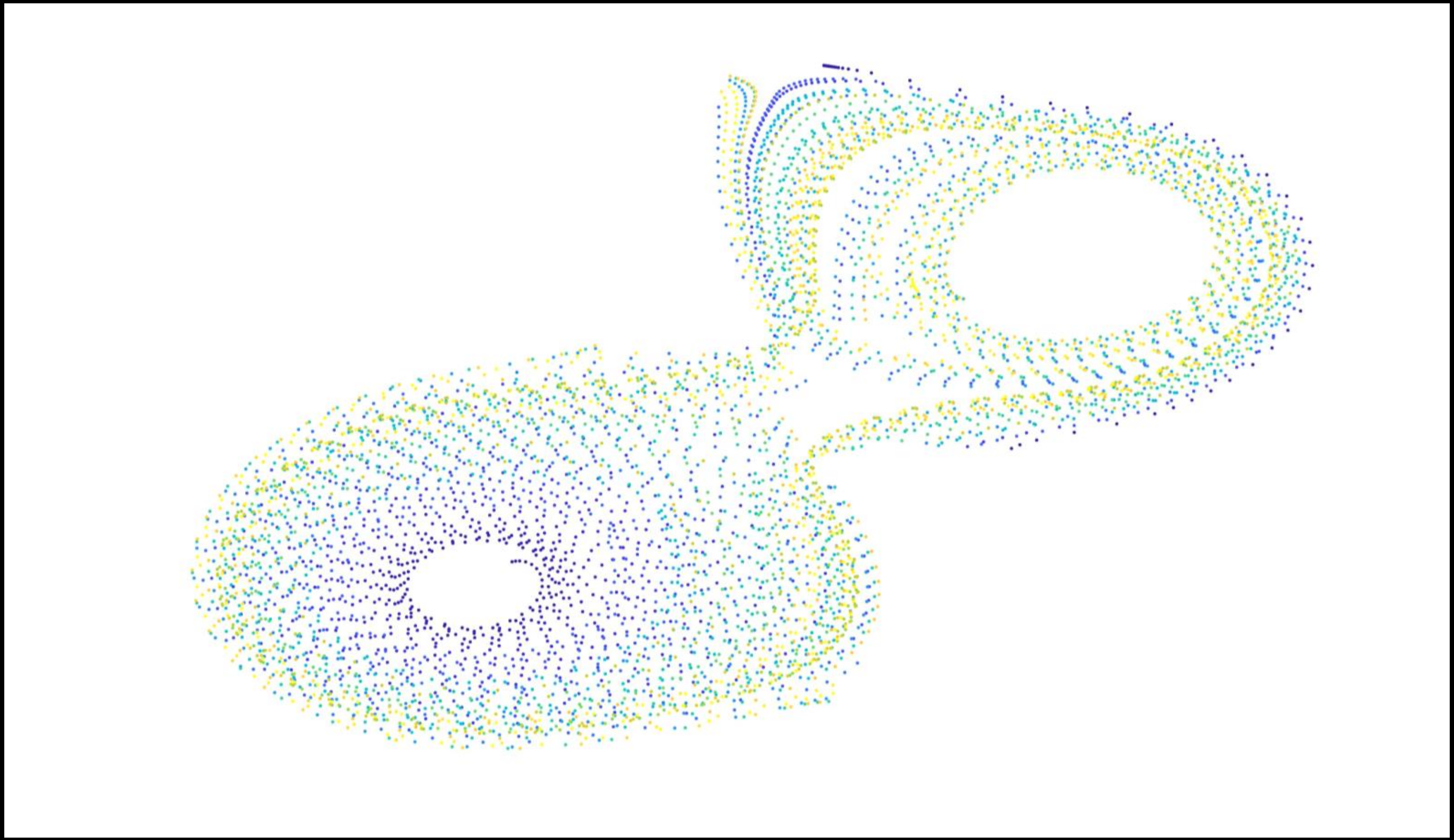
Experiment on RECON2 Combos



X-t-SNE

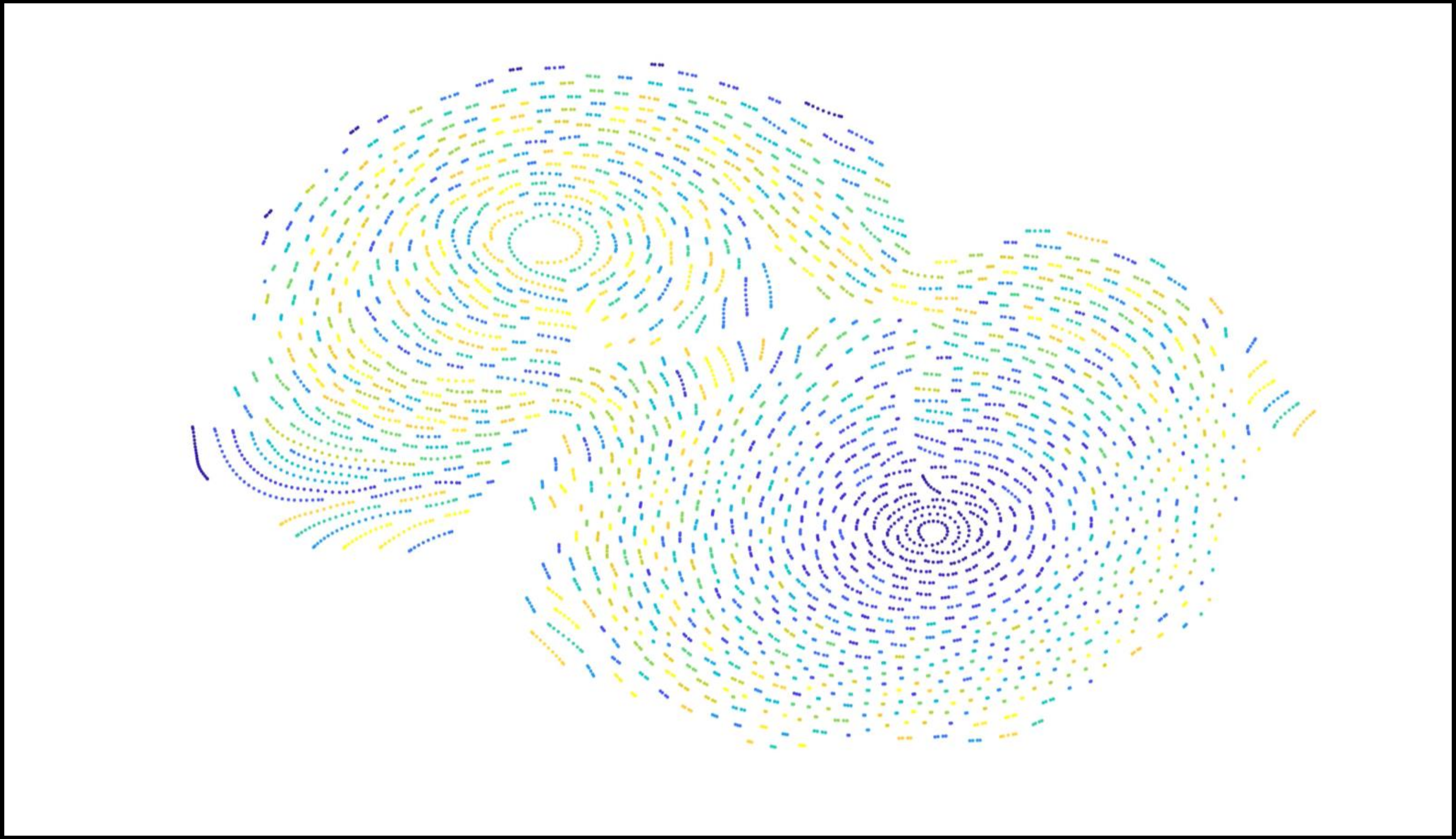
Wyss Institute

Lorenz Attractor



t-SNE

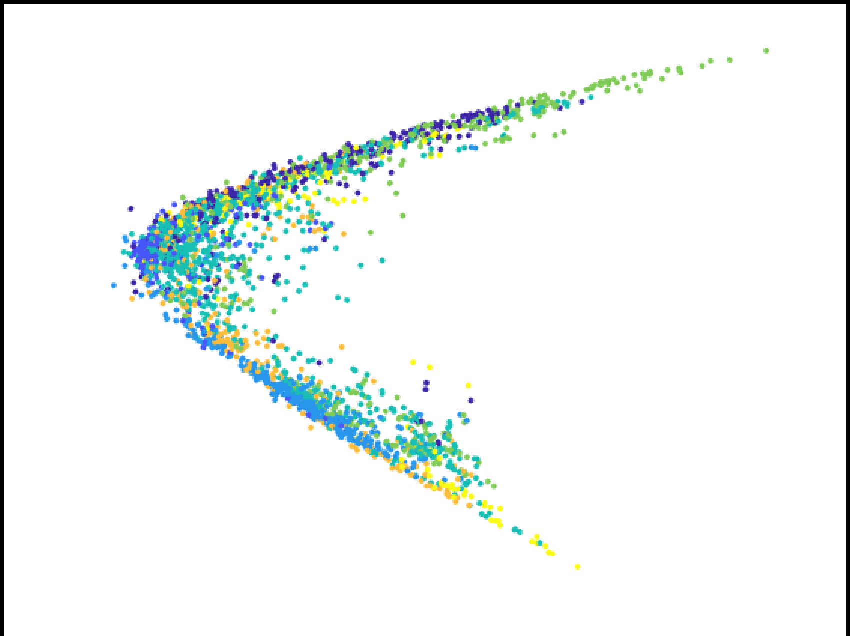
Lorenz Attractor



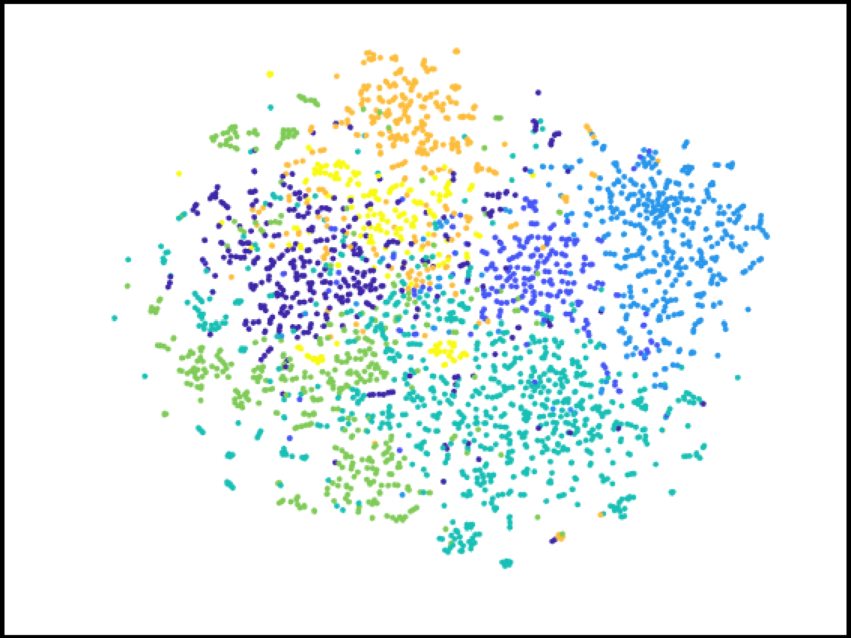
X-t-SNE

Wyss Institute

Comparative Results on Cora Dataset against state-of-the-art Algorithms



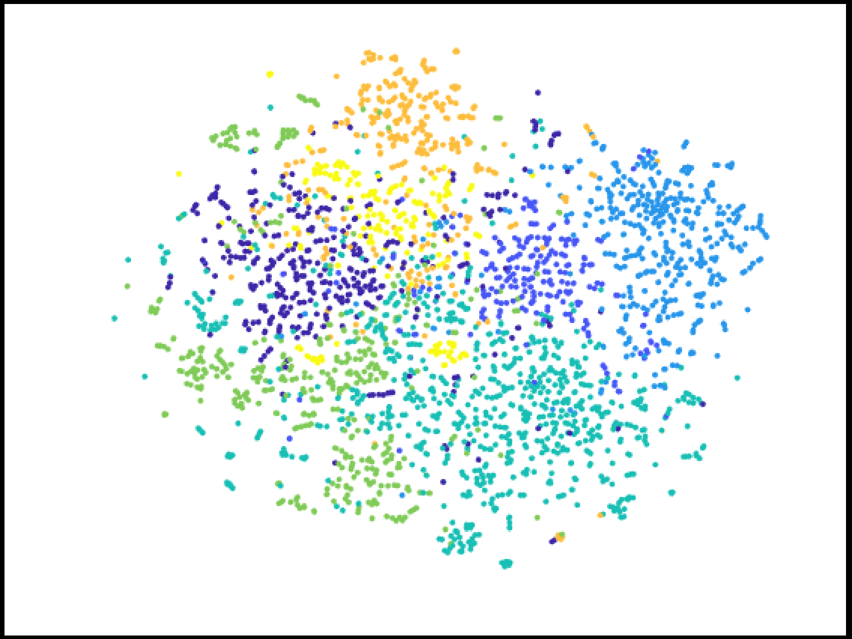
node2vec



X-t-SNE

Comparative Results on Cora Dataset against state-of-the-art Algorithms

Variational Graph Autoencoders



X-t-SNE

Scope

Biology

- Embed expression profiles in tissue/tumor/species specific GRN contexts (the Genotype-Tissue Expression (GTEx) Project)
- Multiomics with multigraph structures (layered X-t-SNEs)
- Track cell state evolution in an X-t-SNE landscape (preserving temporal neighborhoods)
- Graph enhanced computational drug discovery

(Meta) Machine Learning

- Opening the black box of deep learning: visualizing activation of hidden neurons in deep neural networks

Future Work

- More “perturbed” experiments on RECON2
- Compare to other graph embedding algorithms on quantifiable tasks such as link prediction:
  - Only graph structure embedders such as node2vec[1]
    - Principle: preserve neighborhood by simulating biased random walks on the graph
  - Graph + feature space embedders such as Variational Graph Autoencoders[2]
    - Principle: (1) preserve information + (2) latent space varies smoothly around a node’s neighborhood
- Generalized X-t-SNE with EM-style training
- Attach a decoder network to have a fully generative model (say predicting the high-D metabolic state for every gene-perturbation in the low-D space)

[1]Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

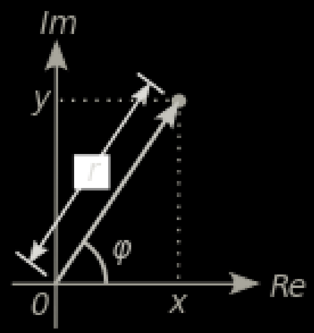
[2]Kipf, Thomas N., and Max Welling. "Variational Graph Auto-Encoders." *arXiv preprint arXiv:1611.07308* (2016).

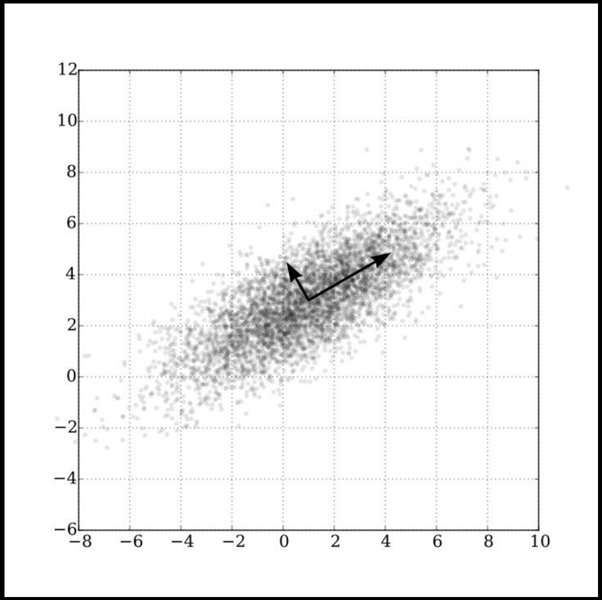
Conclusions

- Presented a new algorithm to overlay multiple layers of context to a feature space with an arbitrary level of complexity, that competes with state-of-the-art
- Achieved a low-D space that preserves local neighborhoods, which is good for visualization and more
  - evolutionary landscapes
  - creating similarity metrics
- This low-D space can be used as input to any other algorithm that cares for local neighborhoods (like clustering, or inducing maps between latent spaces)

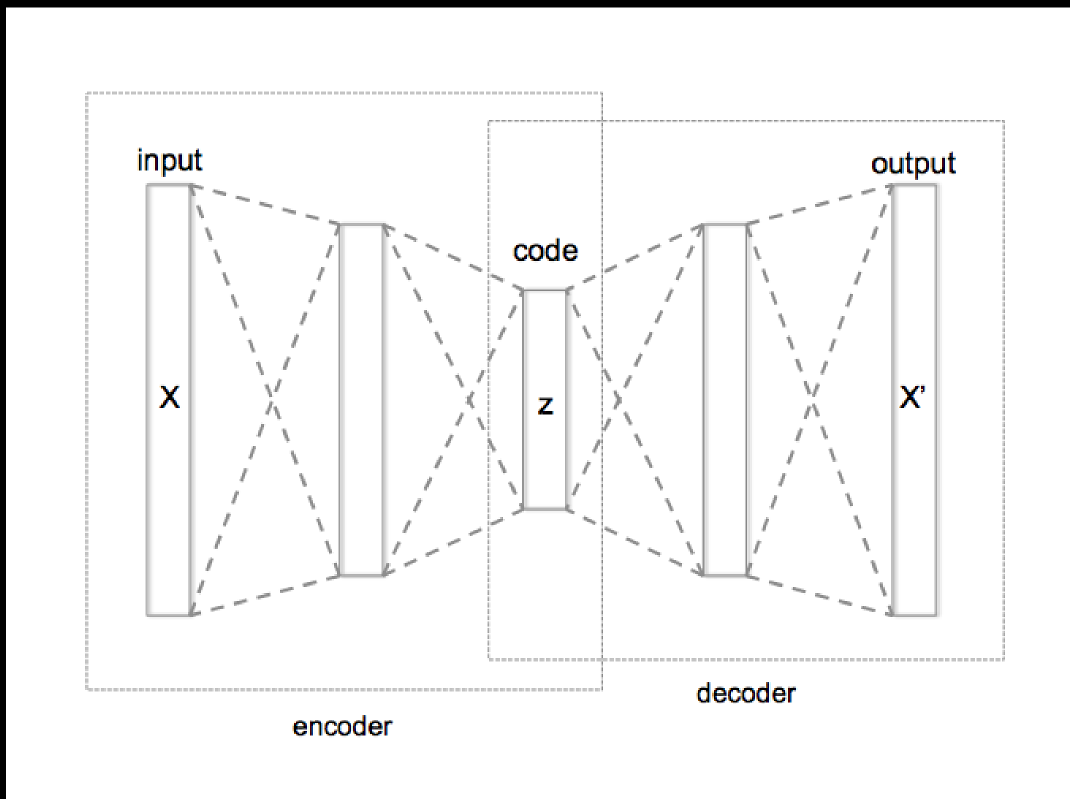
Aside: Principal Component Analysis (PCA)

- $X_{N\times P} = U_{N\times N}\Sigma_{N\times P}W^T_{P\times P} = U\Sigma W^T$
- Covariance Matrix: $X^T X = W\Sigma^2 W^T$
- $$\begin{aligned} Y_{N\times P} &= XW \\ &= U\Sigma \end{aligned}$$
- $$\begin{aligned} Y_{N\times L} &= XW_{P\times L} \\ Y_{N\times L} &= U_{N\times L}\Sigma_{L\times L} \end{aligned}$$

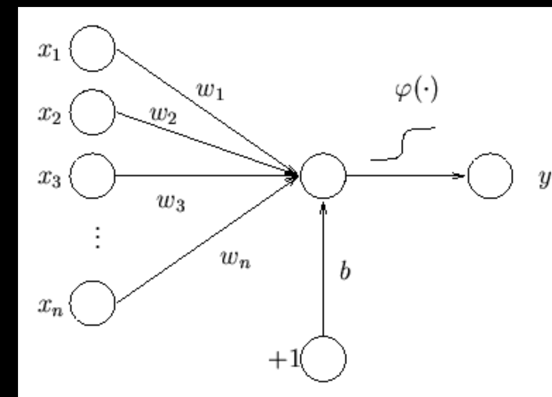




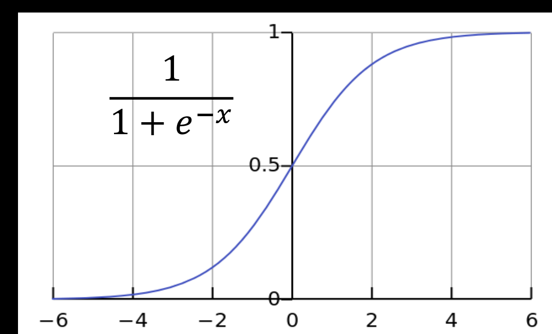
Aside: Autoencoder



Network Architecture



Artificial Neuron



Activation Function φ