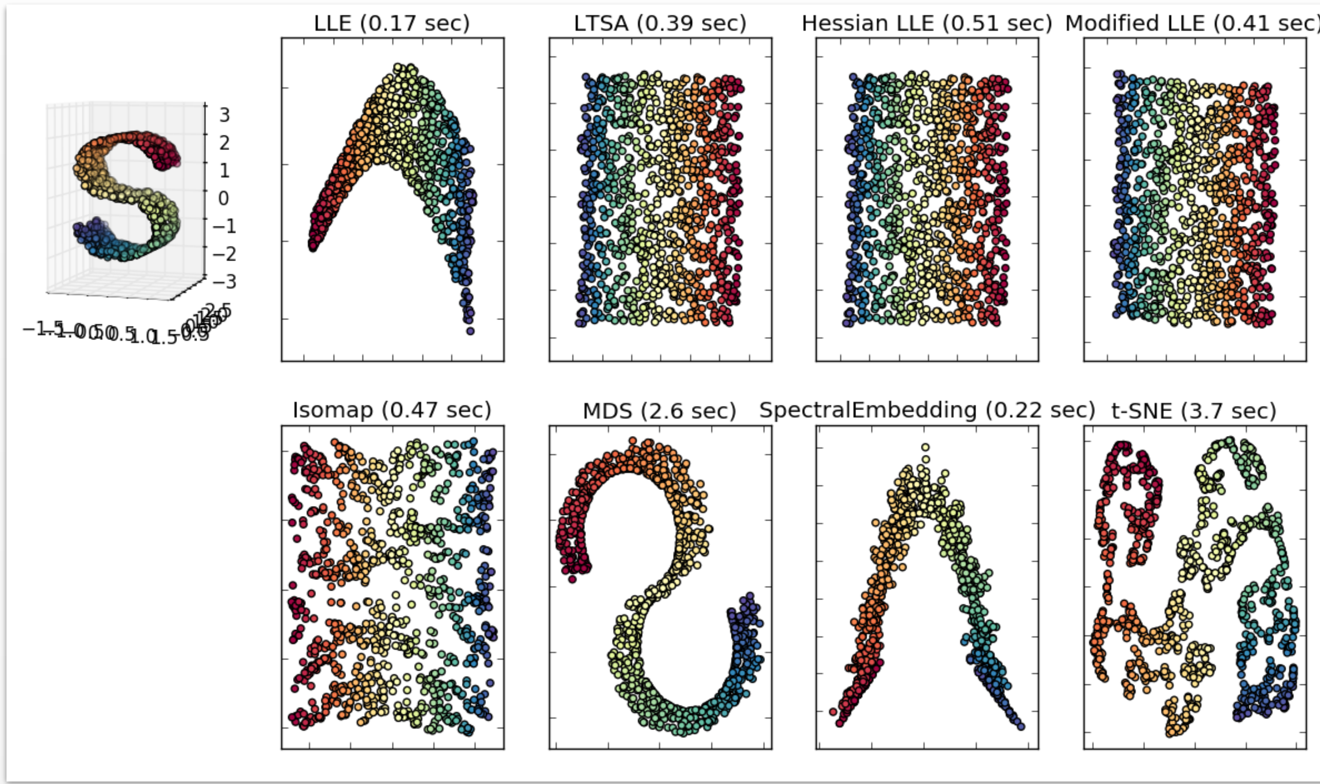
Dimensionality Reduction

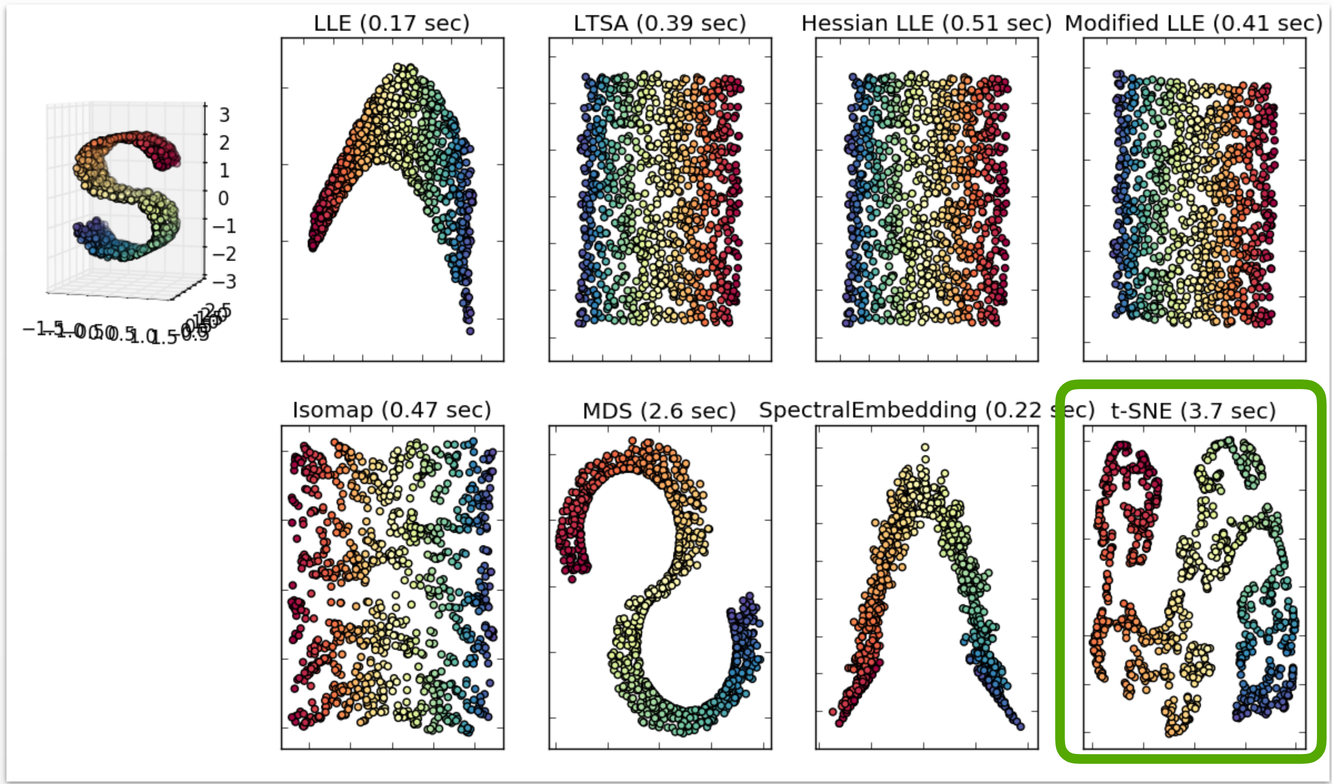
Shantanu Jain

Manifold Learning



Idea: Perform a *non-linear* dimensionality reduction in a manner that preserves proximity (but not distances)

Manifold Learning



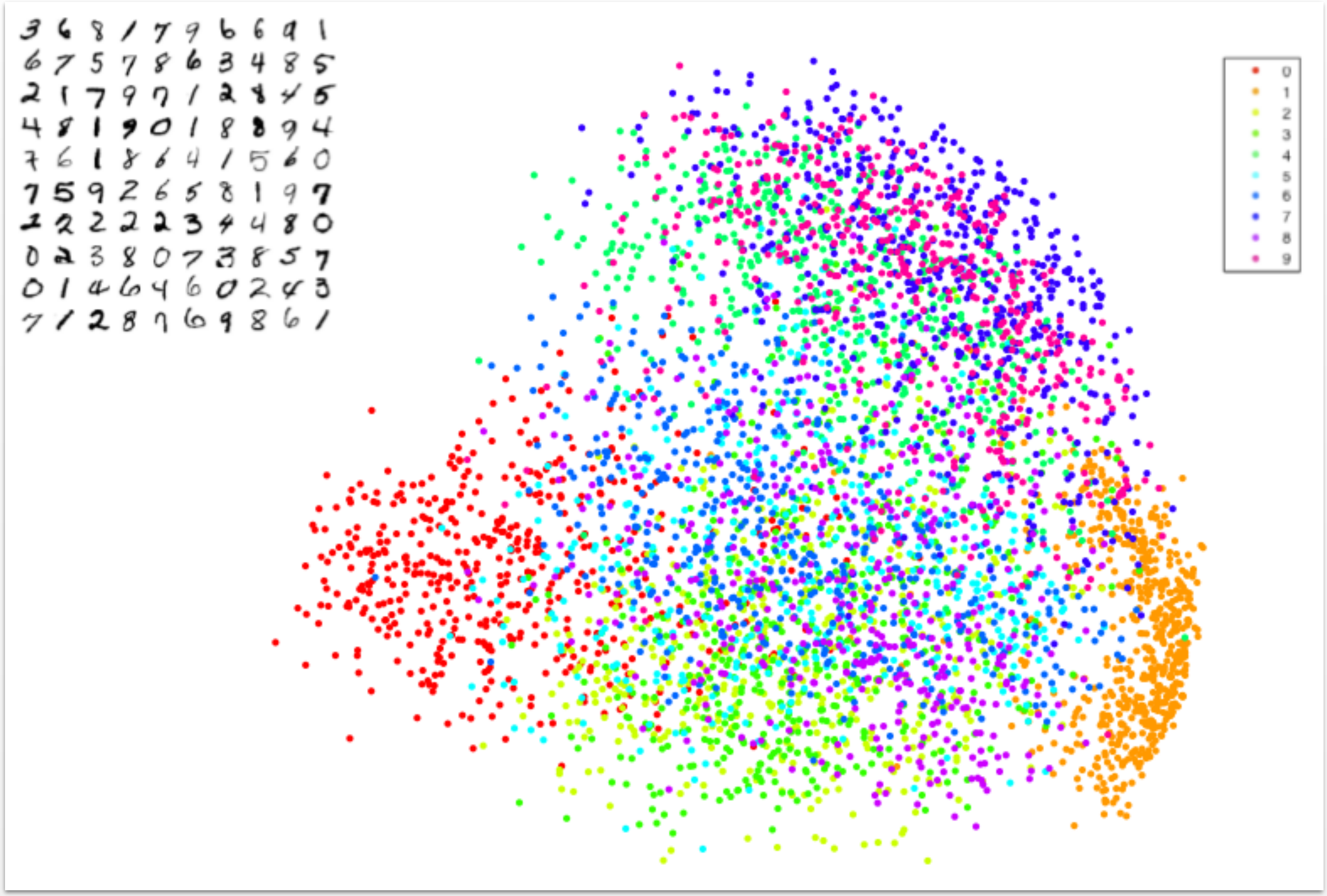
Visualizing data using **t-SNE** [PDF] jmlr.org

L Maaten, G Hinton - Journal of Machine Learning Research, 2008 - jmlr.org

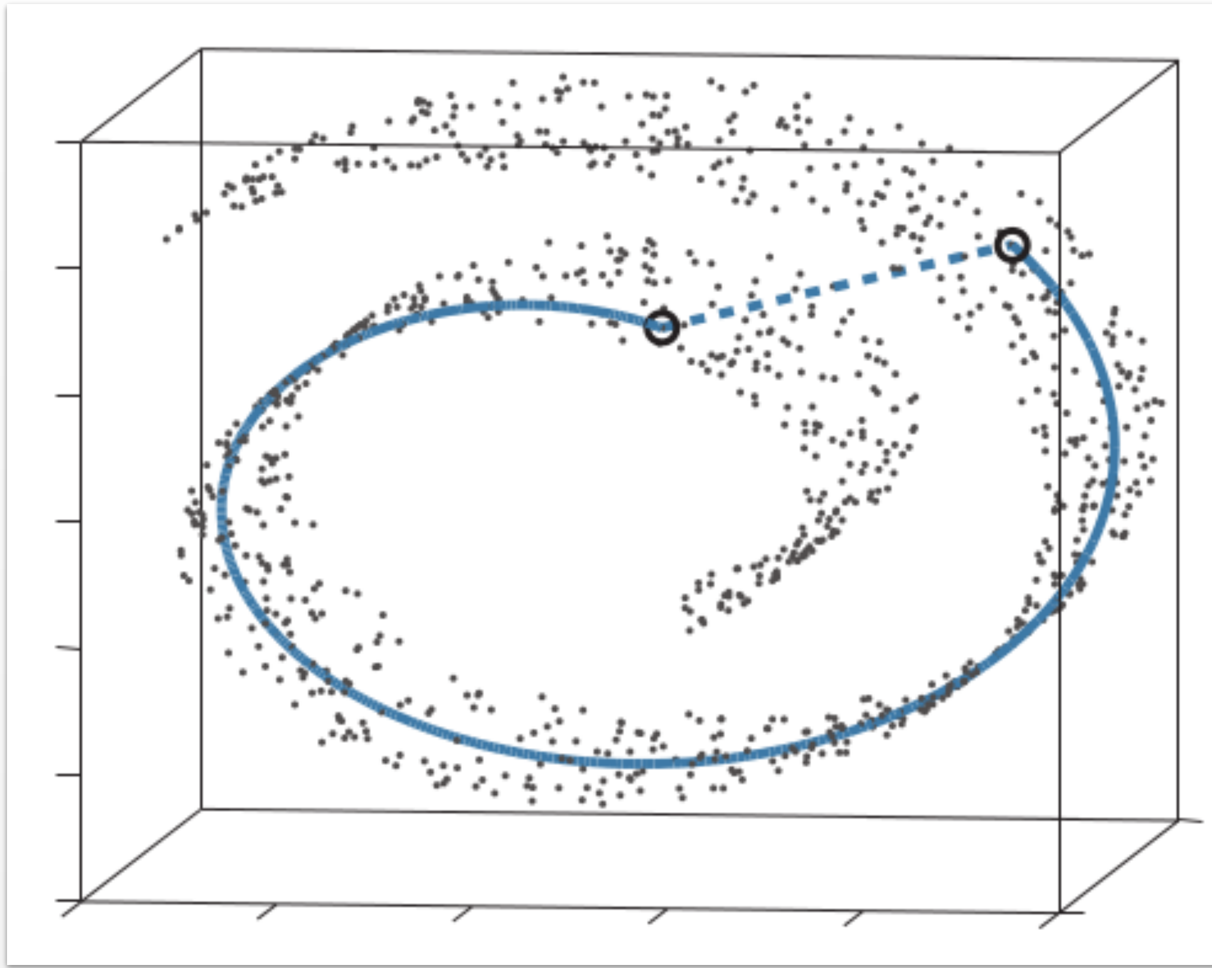
Abstract We present a new technique called" **t-SNE**" that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map. The technique is a variation of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) that is much **...**

Cited by 1771 Related articles All 35 versions Cite Save

PCA on MNIST Digits

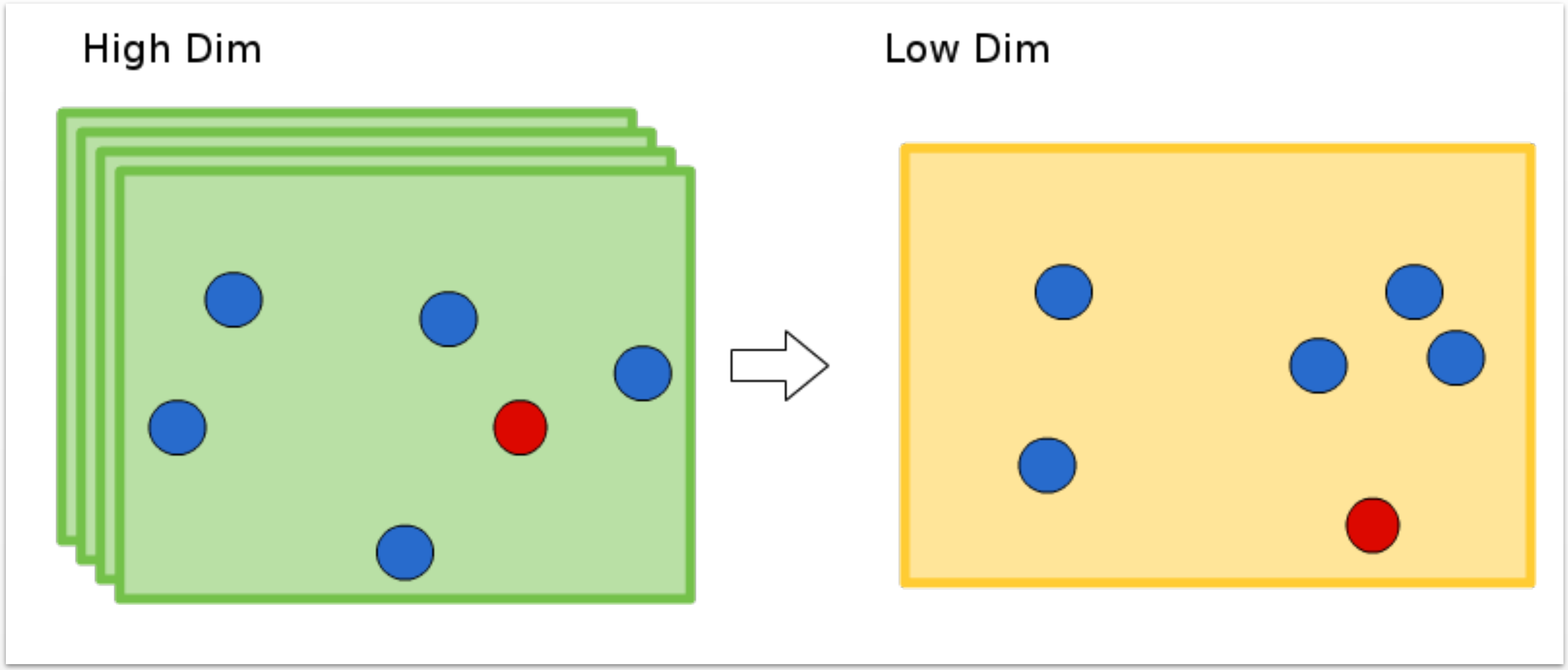


Swiss Roll



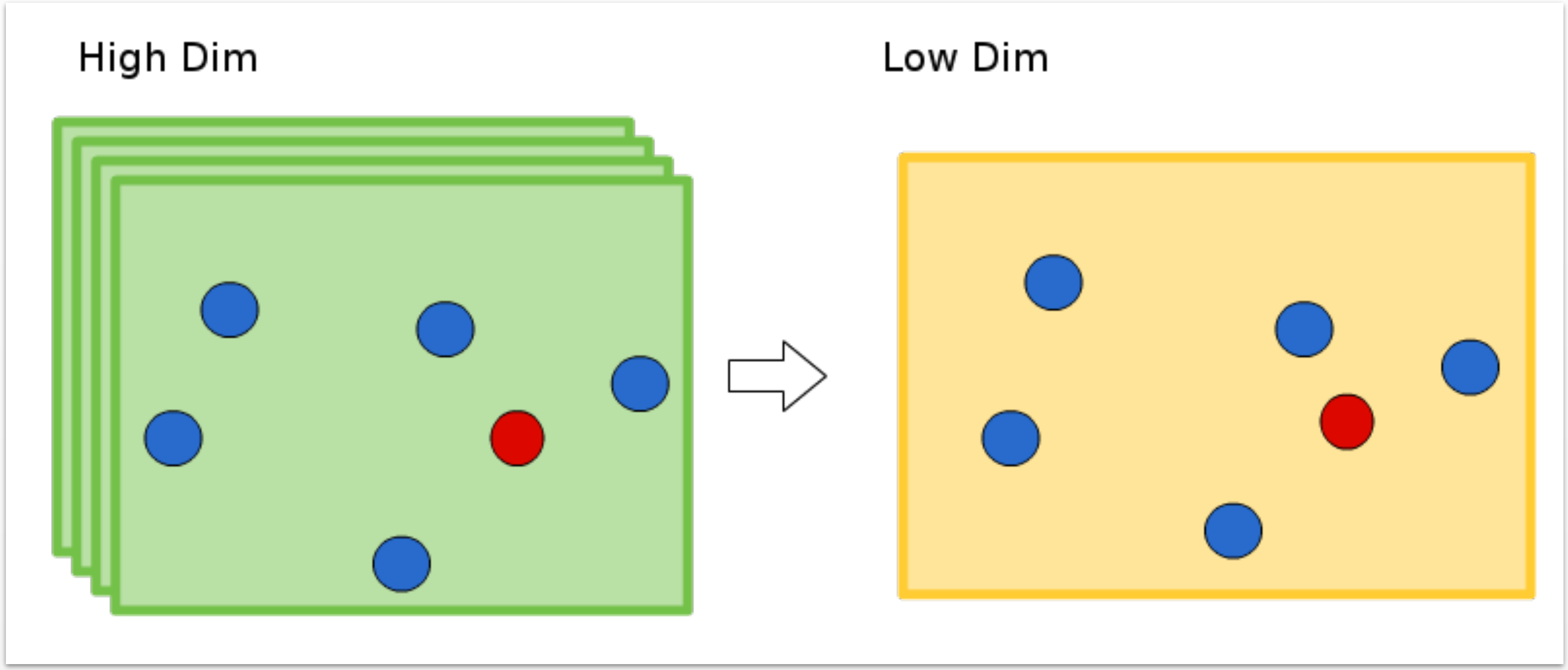
Euclidean distance is not always
a *good* notion of proximity

Non-linear Projection



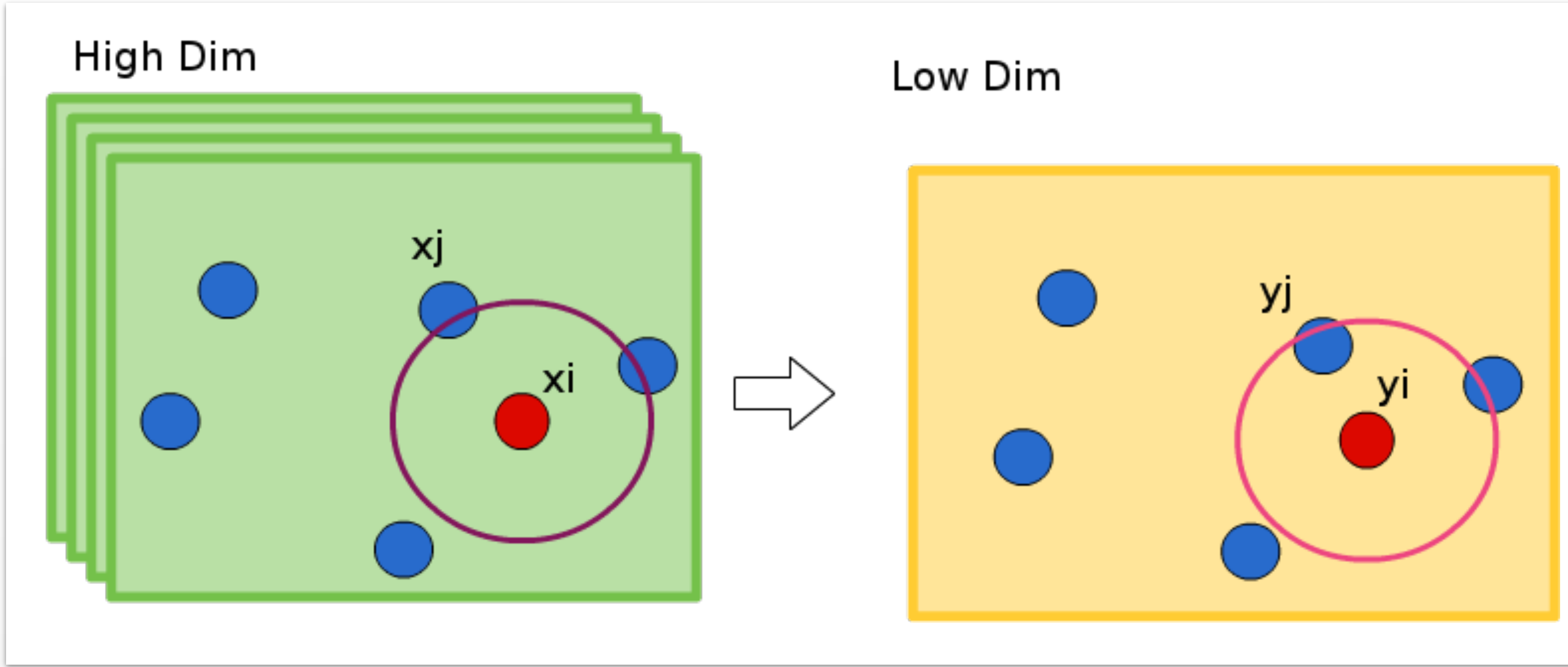
Bad projection: relative position to neighbors changes

Non-linear Projection



Intuition: Want to preserve *local* neighborhood

Stochastic Neighbor Embedding



Similarity in *high* dimension

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

Similarity in *low* dimension

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

Stochastic Neighbor Embedding

- Similarity of datapoints in High Dimension

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

- Similarity of datapoints in Low Dimension

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

- Cost function

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_{j\neq i} p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$

$P_i = [p_{j|i}]_{j\neq i}$

Vector with entries $p_{j|i}$ for all $j \neq i$

$Q_i = [q_{j|i}]_{j\neq i}$

Idea: Optimize y_i via gradient descent on C

Stochastic Neighbor Embedding

Gradient has a surprisingly simple form

$$\frac{\partial C}{\partial y_i} = \sum_{j \neq i} (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

The gradient update with momentum term is given by

$$Y^{(t)} = Y^{(t-1)} + \eta \frac{\partial C}{\partial y_i} + \beta(t)(Y^{(t-1)} - Y^{(t-2)})$$

$Y^{(t)}$ is a matrix containing the low-dimension representation of all the points at iteration t

Stochastic Neighbor Embedding

Gradient has a surprisingly simple form

$$\frac{\partial C}{\partial y_i} = \sum_{j \neq i} (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

The gradient update with momentum term is given by

$$Y^{(t)} = Y^{(t-1)} + \eta \frac{\partial C}{\partial y_i} + \beta(t)(Y^{(t-1)} - Y^{(t-2)})$$

Problem: $p_{j|i}$ is not equal to $p_{i|j}$

Symmetric SNE

- Minimize a single KL divergence between a joint probability distribution

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

Old cost function
$\sum_i \sum_{j \neq i} p_{j|i} \log \frac{q_{j|i}}{p_{j|i}}$

- The obvious way to redefine the pairwise similarities is

$$p_{ij} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq l} exp(-||x_l - x_k||^2/2\sigma^2)}$$

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

If the i^{th} point is an outlier all p_{ij}'s are small. Which means that the cost function C is insensitive to the positioning of the i^{th} point's representation in the lower dimensional space

Symmetric SNE

- Minimize a single KL divergence between a joint probability distribution

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

- Solution for weakly determined outlier points.

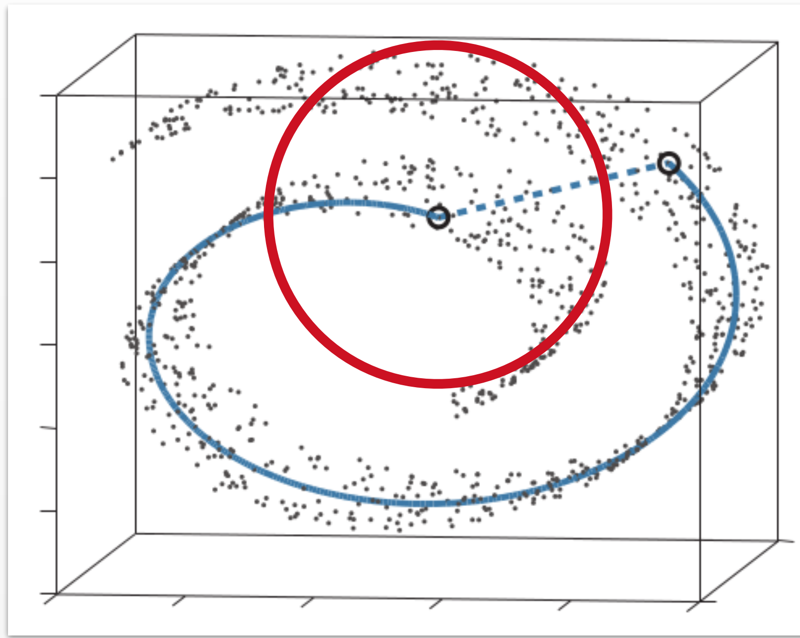
$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

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

The total probability of the i^{th} point is at least $\frac{1}{2N}$

$$\begin{aligned} p_i &= \textstyle\sum_{j \neq i} p_{ij} \\ &= \frac{\sum_{j \neq i} p_{j|i} + \sum_{j \neq i} p_{i|j}}{2N} \\ &= \frac{1 + \sum_{j \neq i} p_{i|j}}{2N} \\ &\geq \frac{1}{2N} \end{aligned}$$

Choosing the bandwidth

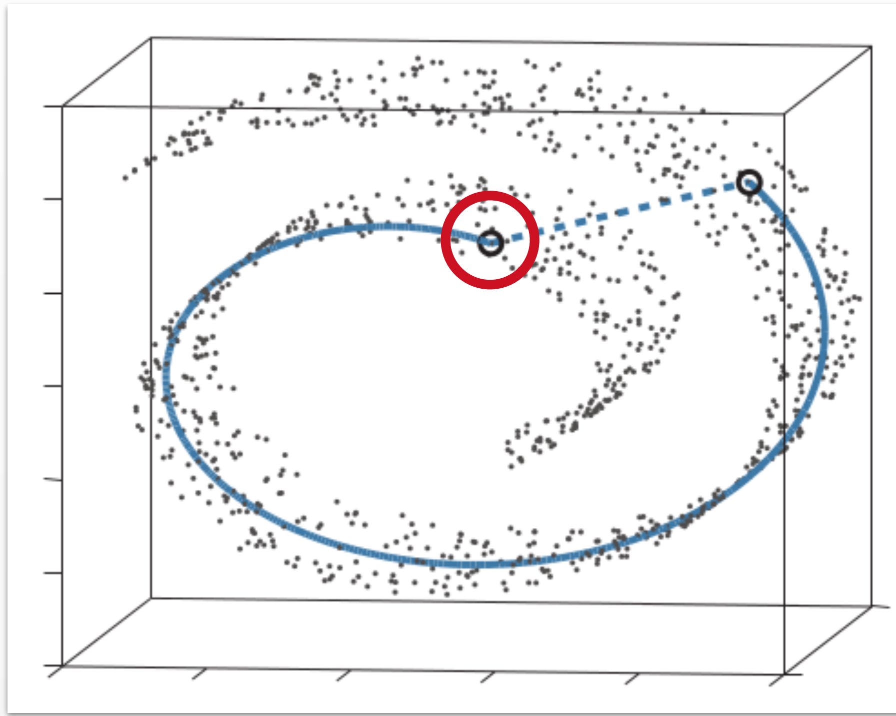


$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

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

Bad σ: Neighborhood is not local in manifold

Choosing the bandwidth

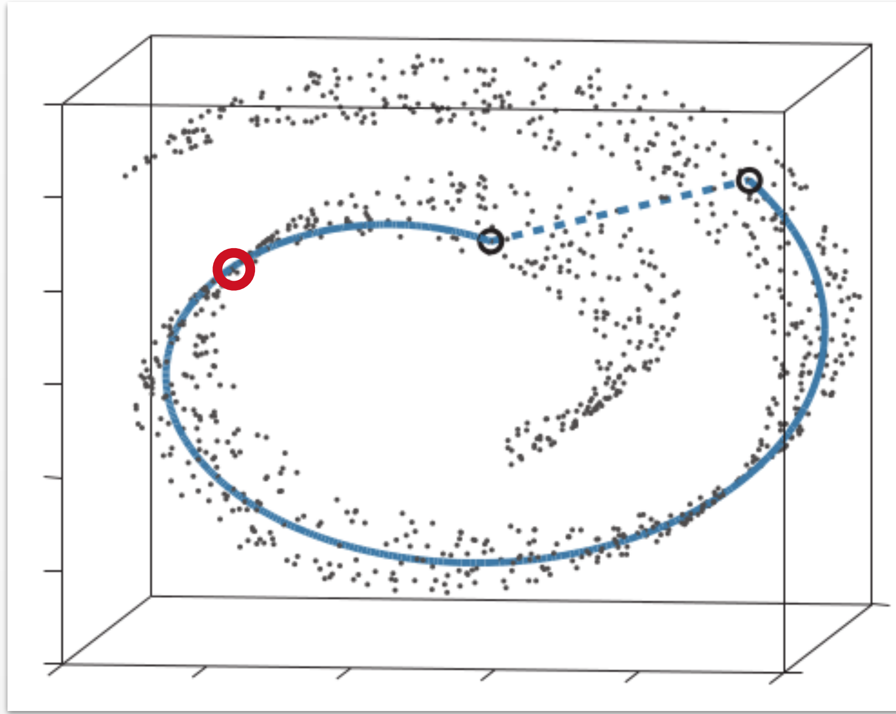


$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

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

Solution: Define σ_i per point. Good σ_i: Neighborhood contains 5-50 points

Choosing the bandwidth

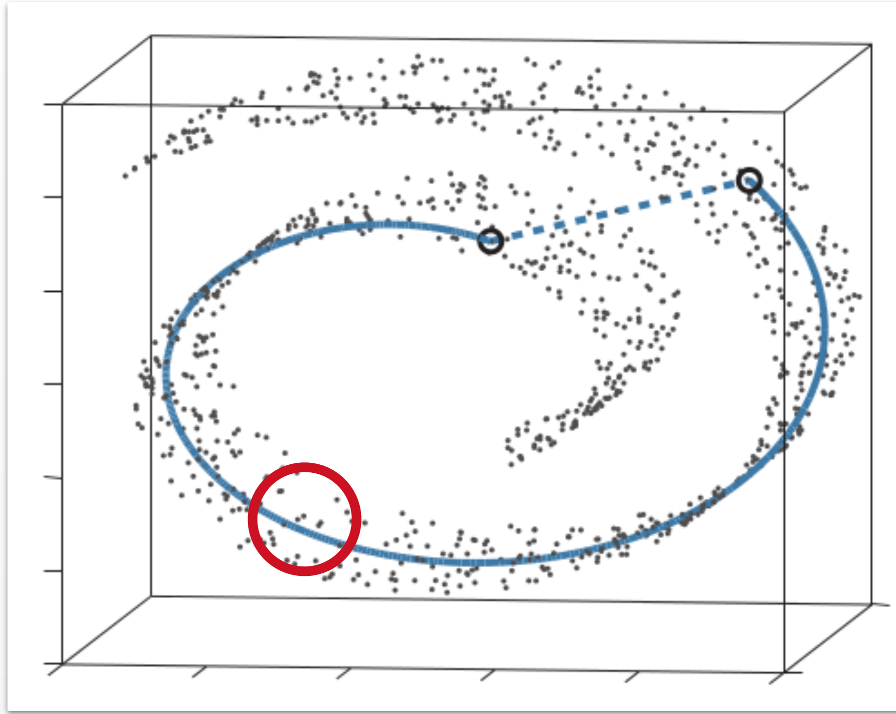


$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

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

Solution: Define σ_i per point.

Choosing the bandwidth

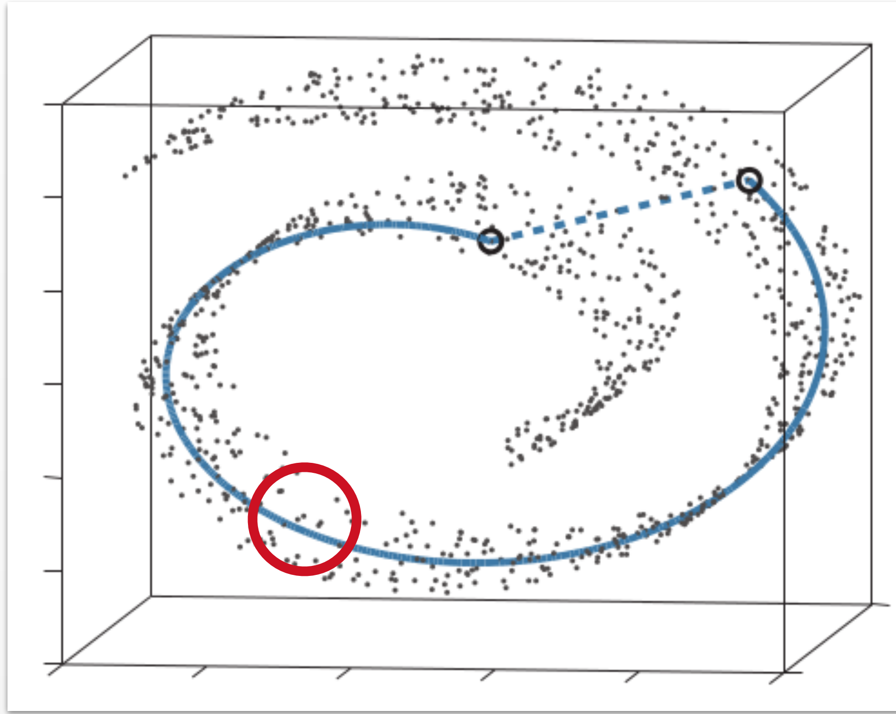


$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

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

Solution: Define σ_i per point.

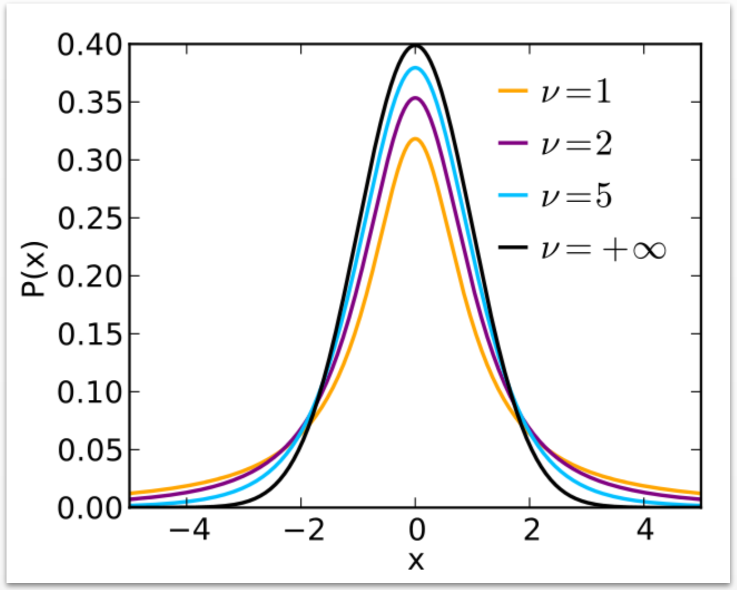
Choosing the bandwidth



$$\mathrm{Perp}(\mathbf{p}_{j|i}) = \exp H(\mathbf{p}_{j|i}) = \exp^{-\sum_j \mathbf{p}_{j|i} \log \mathbf{p}_{j|i}}$$

Set σ_i to ensure constant perplexity

t-SNE: SNE with a t-Distribution



$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Similarity in *High* Dimension

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

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

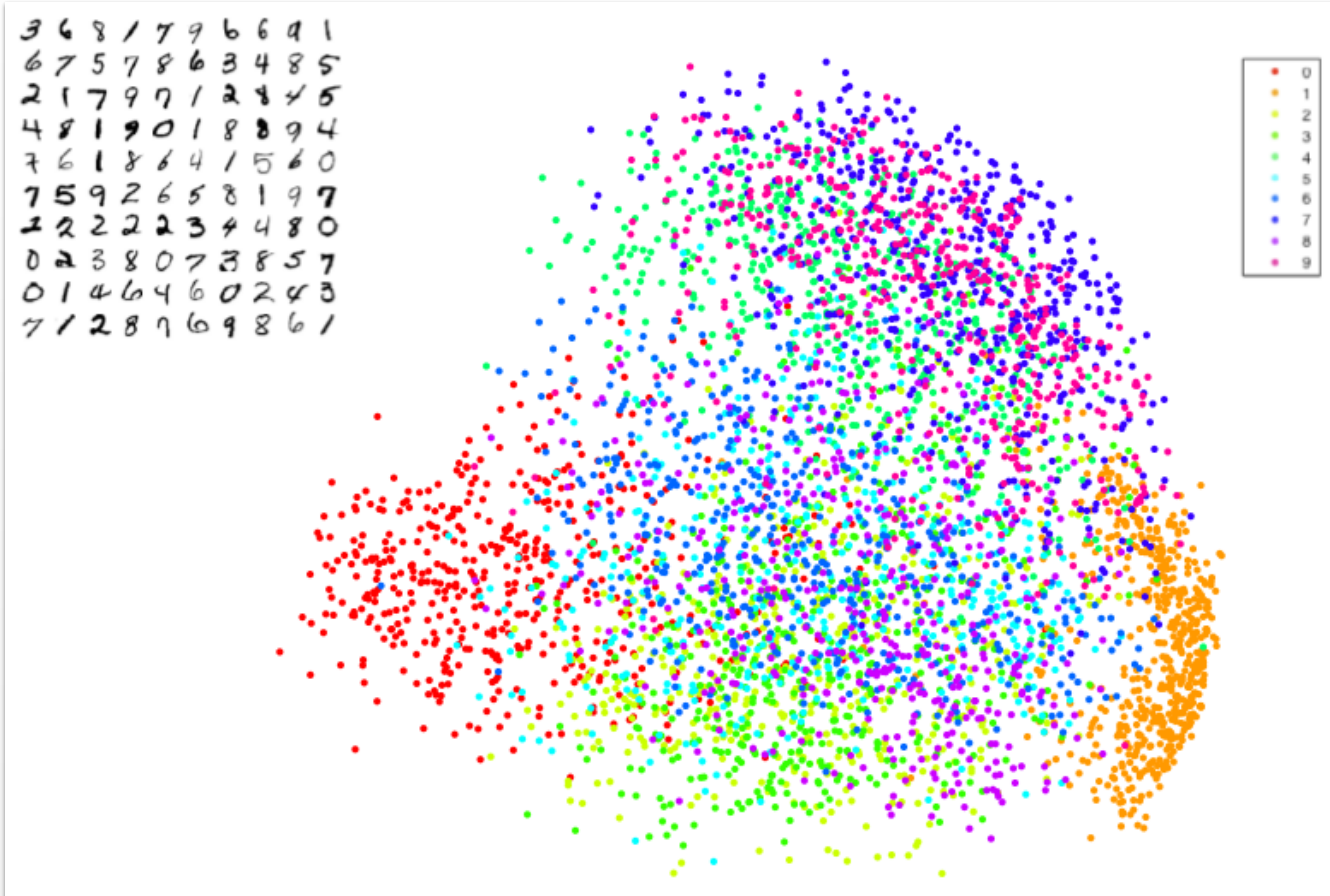
Similarity in *Low* Dimension

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

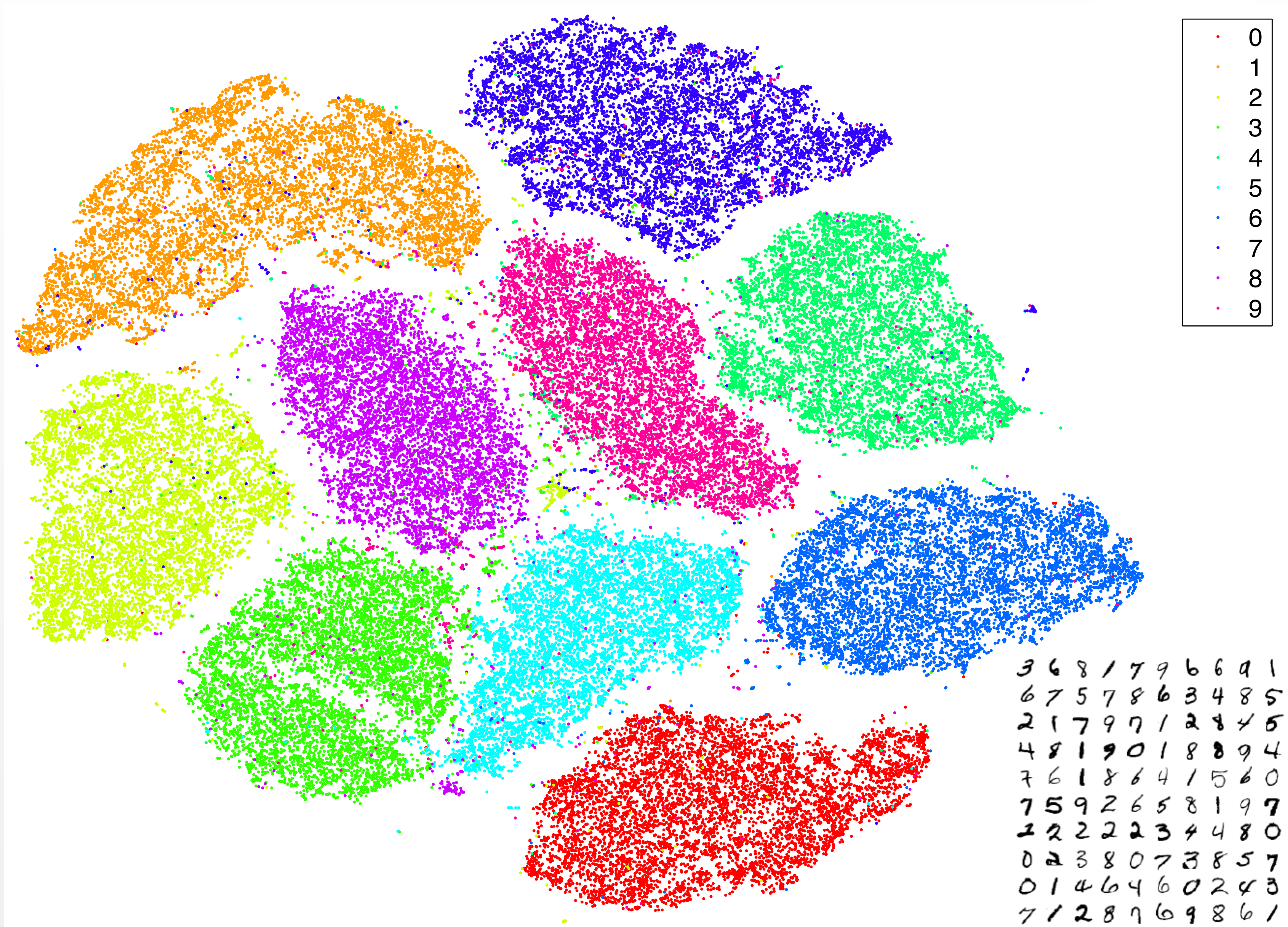
Gradient

$$\frac{\partial C}{\partial y_i} = 4\sum_{j\neq i}(p_{ij} - q_{ij})(1+||y_i - y_j||^2)^{-1}(y_i - y_j)$$

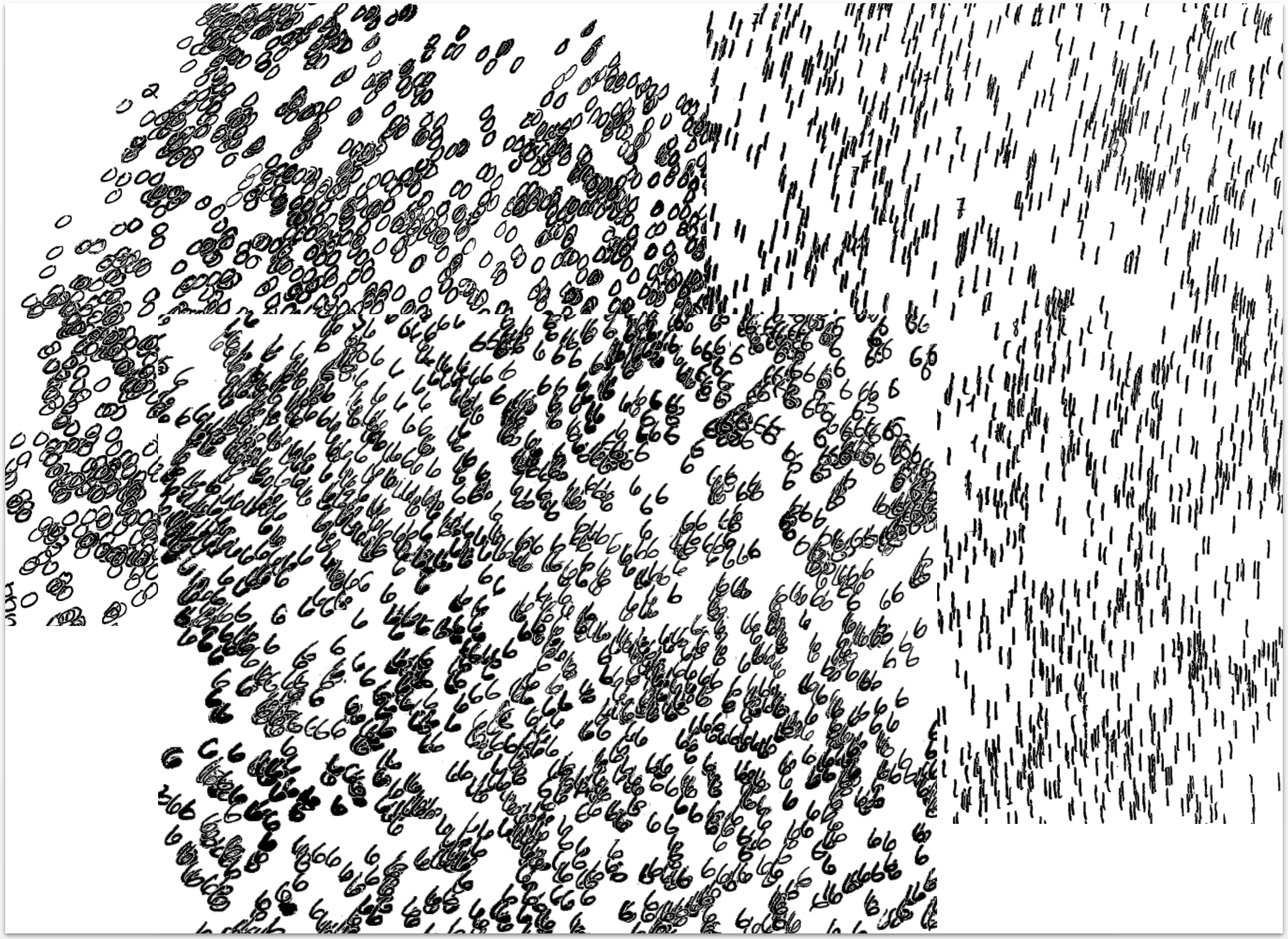
PCA on MNIST Digits



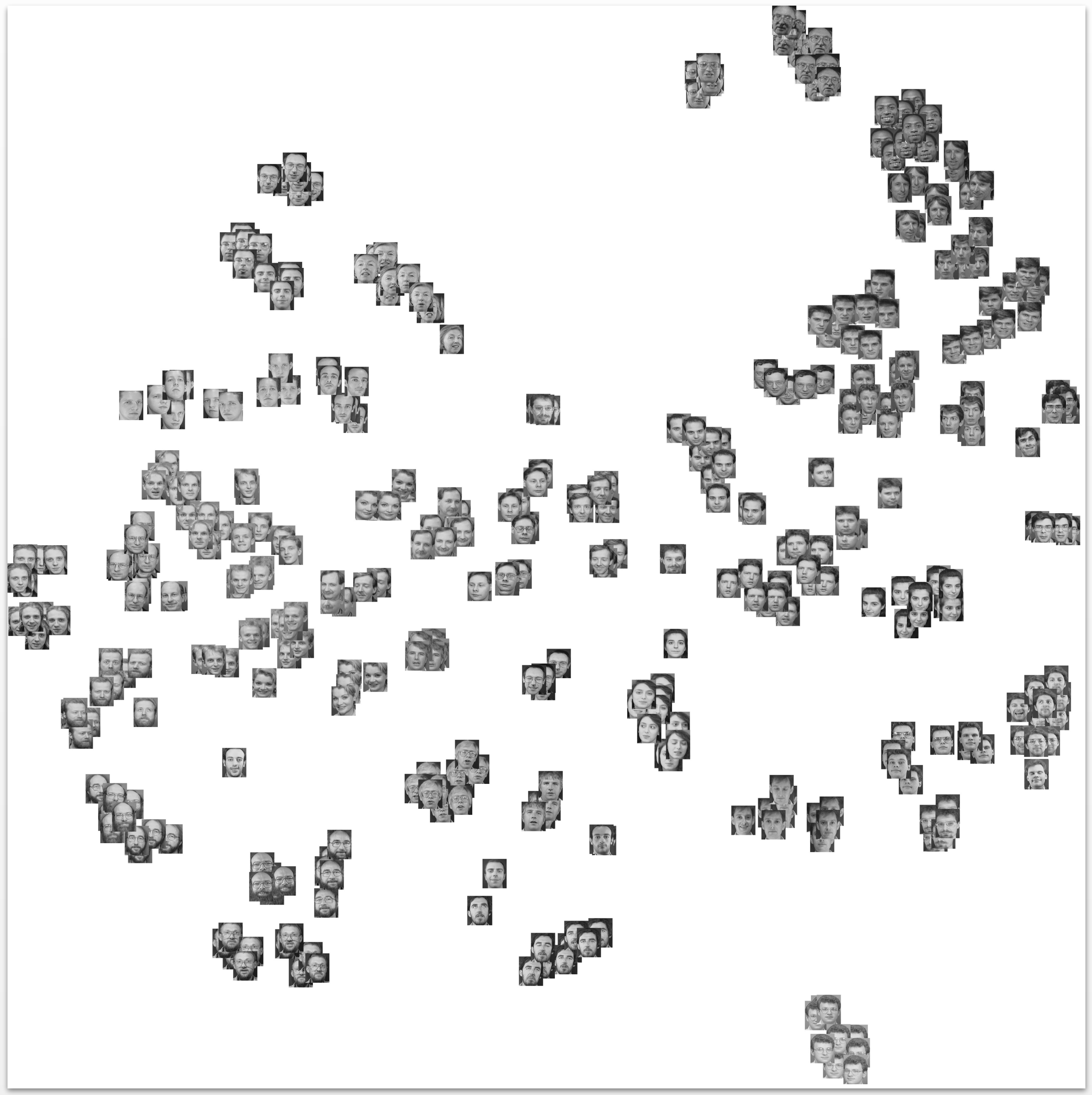
t-SNE on MNIST Digits



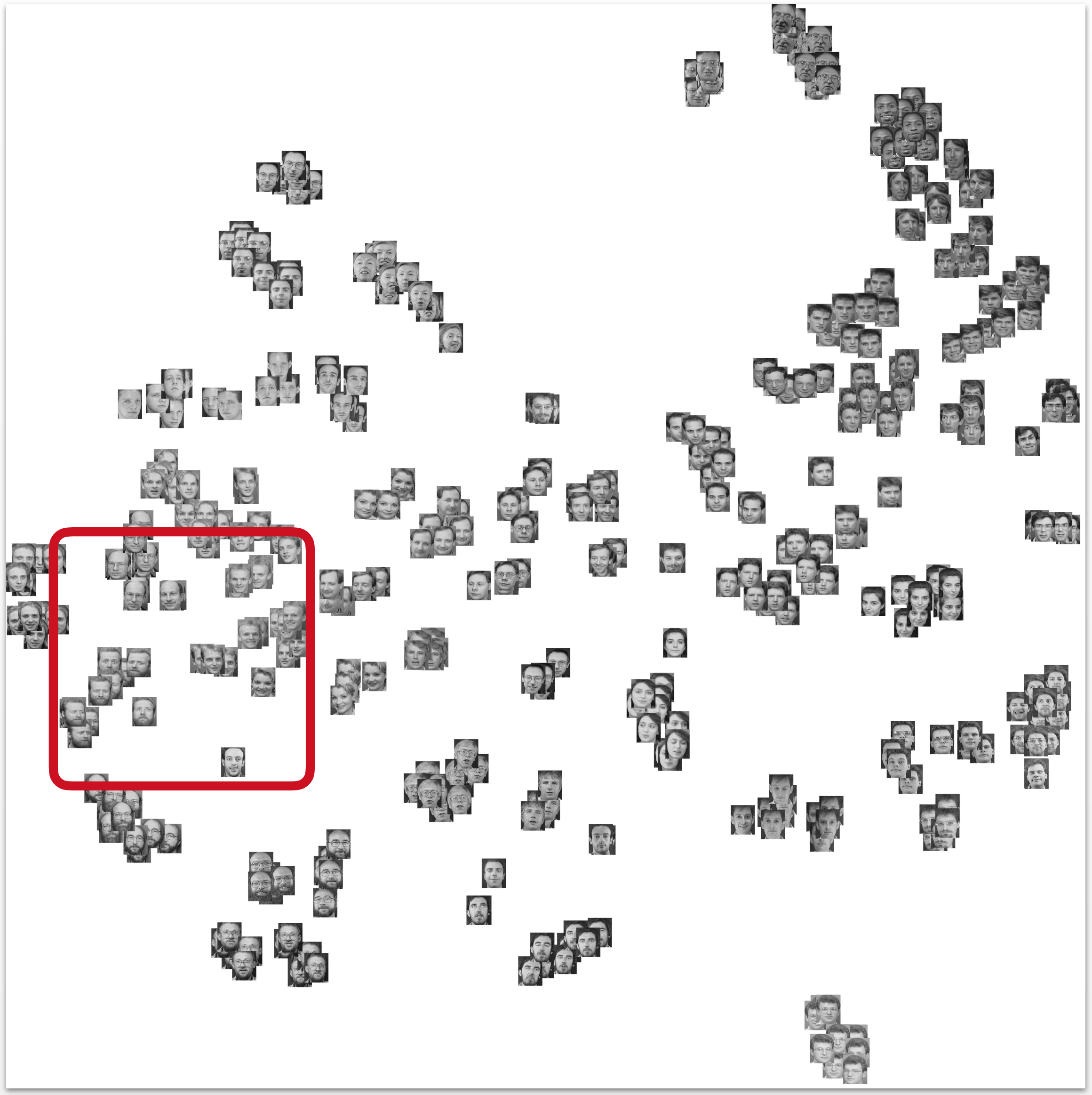
t-SNE on MNIST Digits



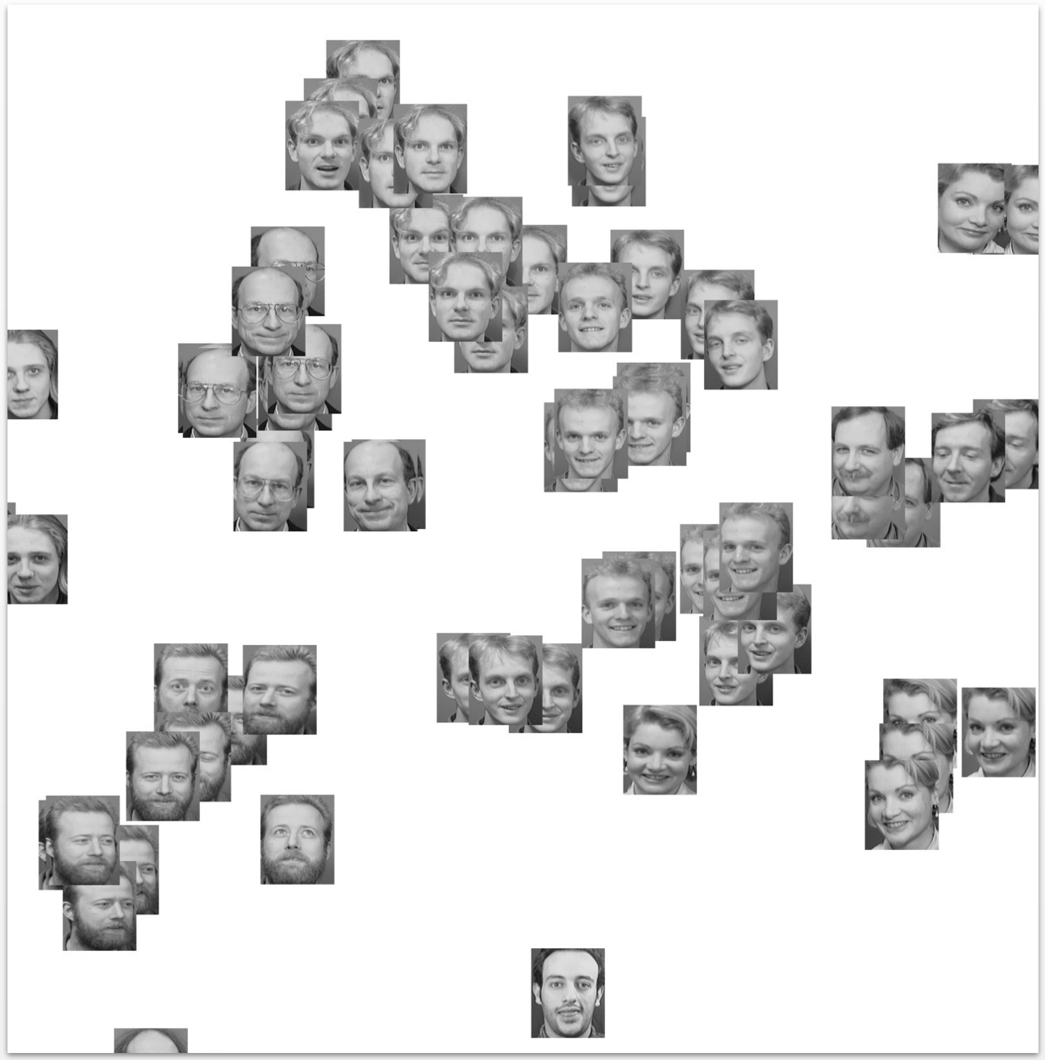
t-SNE on Olivetti Faces



t-SNE on Olivetti Faces



t-SNE on Olivetti Faces



Manifold Learning