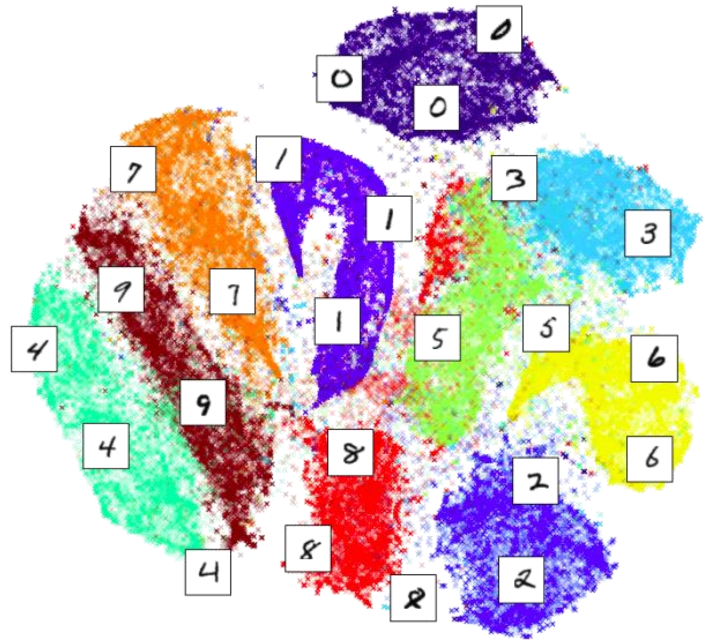


Dimensionality Reduction: t-SNE

Pilsung Kang
School of Industrial Management Engineering
Korea University

Stochastic Neighbor Embedding

Hinton and Roweis (2002)

- Stochastic Neighbor Embedding (SNE)
  - ✓ It is more important to get local distances right than non-local ones
  - ✓ SNE has a probabilistic way of deciding if a pairwise distance is local
  - ✓ Convert each high-dimensional similarity into the probability that one data point will pick the other data point as its neighbor
    - Probability of picking j given in high D
    - Probability of picking j given in low D

$$p_{j|i} = \frac{e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2}}}{\sum_{k \neq i} e^{-\frac{||\mathbf{x}_i - \mathbf{x}_k||^2}{2\sigma_i^2}}} \qquad q_{j|i} = \frac{e^{-||\mathbf{y}_i - \mathbf{y}_j||^2}}{\sum_{k \neq i} e^{-||\mathbf{y}_i - \mathbf{y}_k||^2}}$$

Stochastic Neighbor Embedding

- Picking the Radius of the Gaussian in p
  - ✓ We need to use different radii in different parts of the space so that we keep the effective number of neighbors about constant
  - ✓ A big radius leads to a high entropy for the distribution over neighbors of i, whereas a small radius leads to a low entropy
  - ✓ Decide what entropy you want and then find the radius that produces that entropy

$$Perplexity(P_i) = 2^{H(P_i)}$$

$$H(P_i) = \sum_j p_{j|i} log_2 p_{j|i}$$

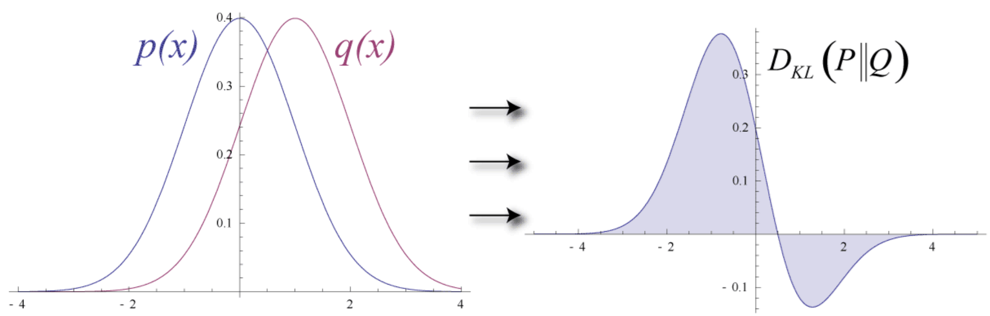
$$p_{j|i} = \frac{e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2}}}{\sum_{k \neq i} e^{-\frac{||\mathbf{x}_i - \mathbf{x}_k||^2}{2\sigma_i^2}}}$$

  - ✓ The performance of SNE is fairly robust to changes in the perplexity (5~50)

Stochastic Neighbor Embedding

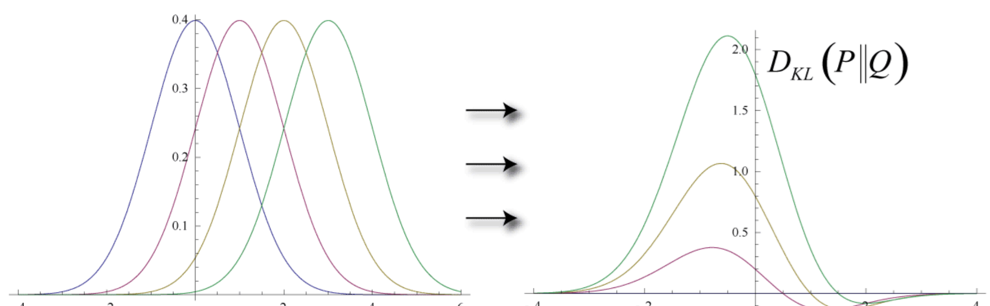
- Cost Function for a Low-dimensional Representation
  - ✓ Kullback-Leibler divergence
    - A non-symmetric measure of the difference between two probability distribution P and Q

$$Cost = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$



Original Gaussian PDF's

KL Area to be Integrated



Stochastic Neighbor Embedding

- Cost Function for a Low-dimensional Representation
  - ✓ Kullback-Leibler divergence
    - A non-symmetric measure of the difference between two probability distribution P and Q

$$Cost = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$

  - ✓ Gradient
    - Differencing Cost is tedious because $\mathbf{y}_k$ affect q_{ij} via the normalized term in Eq. (3), but the result is simple Hinton and Roweis (2002)
    - The gradient has a surprisingly simple form Maaten and Hinton (2008)

$$\frac{\partial C}{\partial \mathbf{y}_i} = 2 \sum_j (\mathbf{y}_j - \mathbf{y}_i)(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$$

Stochastic Neighbor Embedding

- Cost Function for a Low-dimensional Representation
  - ✓ Gradient
    - Differencing Cost is tedious because $\mathbf{y}_k$ affect q_{ij} via the normalized term in Eq. (3), but the result is simple Hinton and Roweis (2002)
    - The gradient has a surprisingly simple form Maaten and Hinton (2008)

$$Cost = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$

지금부터 내가

자세히 설명한다.

Stochastic Neighbor Embedding

- Gradient of the cost function

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

$$C = \sum_i \sum_j p_{j|i} log p_{j|i} - \sum_i \sum_j p_{j|i} log q_{j|i}$$

$$C' = - \sum_i \sum_j p_{j|i} log q_{j|i} \qquad \left(\frac{\partial C}{\partial y_t} = \frac{\partial C'}{\partial y_t}\right)$$

$$C' = \underbrace{- \sum_i p_{t|i} log q_{t|i}}_{①} \underbrace{- \sum_j p_{j|t} log q_{j|t}}_{②} \underbrace{- \sum_{i \neq t} \sum_{j \neq t} p_{i|j} log q_{i|j}}_{③}$$

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Stochastic Neighbor Embedding

- Gradient of the cost function

$$d_{ti} = exp\big(- ||\mathbf{y}_t - \mathbf{y}_i||^2 \big) = d_{it}$$

$$\frac{\partial d_{ti}}{\partial \mathbf{y}_t} = d'_{ti} = -2(\mathbf{y}_t - \mathbf{y}_i) exp\big(- ||\mathbf{y}_t - \mathbf{y}_i||^2 \big) = -2(\mathbf{y}_t - \mathbf{y}_i) d_{ti}$$

$$q_{t|i} = \frac{exp\big(- ||\mathbf{y}_i - \mathbf{y}_t||^2 \big)}{\sum_{k \neq i} exp\big(- ||\mathbf{y}_i - \mathbf{y}_k||^2 \big)} = \frac{d_{it}}{\sum_{k \neq i} d_{ik}}$$

$$q_{j|t} = \frac{exp\big(- ||\mathbf{y}_t - \mathbf{y}_j||^2 \big)}{\sum_{k \neq t} exp\big(- ||\mathbf{y}_t - \mathbf{y}_k||^2 \big)} = \frac{d_{tj}}{\sum_{k \neq t} d_{tk}}$$

$$q_{i|j} = \frac{exp\big(- ||\mathbf{y}_j - \mathbf{y}_i||^2 \big)}{\sum_{k \neq j} exp\big(- ||\mathbf{y}_j - \mathbf{y}_k||^2 \big)} = \frac{d_{ji}}{\sum_{k \neq j} d_{jk}}$$

Stochastic Neighbor Embedding

- Gradient of the cost function ①

$$\frac{\partial}{\partial y_t}\Big(-\sum_i p_{t|i} log q_{t|i}\Big) = -\sum_i p_{t|i} \cdot \frac{1}{q_{t|i}} \cdot \frac{\partial q_{t|i}}{\partial y_t}$$

$$= -\sum_i p_{t|i} \cdot \frac{1}{q_{t|i}} \cdot \frac{d'_{it} \cdot (\sum_{k \neq i} d_{ik}) - d_{it} \cdot d'_{it}}{(\sum_{k \neq i} d_{ik})^2}$$

$$= -\sum_i p_{t|i} \cdot \frac{1}{q_{t|i}} \cdot \frac{-2(\mathbf{y}_t - \mathbf{y}_i) \cdot d_{it} \cdot (\sum_{k \neq i} d_{ik}) + 2(\mathbf{y}_t - \mathbf{y}_i) \cdot d_{it}^2}{(\sum_{k \neq i} d_{ik})^2}$$

$$= -\sum_i p_{t|i} \cdot \frac{1}{q_{t|i}} \cdot \Big(-2(\mathbf{y}_t - \mathbf{y}_i) \cdot q_{t|i} + 2(\mathbf{y}_t - \mathbf{y}_i) \cdot q_{t|i}^2\Big)$$

$$= \sum_i p_{t|i} \cdot 2(\mathbf{y}_t - \mathbf{y}_i)(1 - q_{t|i})$$

Stochastic Neighbor Embedding

- Gradient of the cost function ②

$$\frac{\partial}{\partial y_t}\Big(-\sum_j p_{j|t} log q_{j|t}\Big) = -\sum_j p_{j|t} \cdot \frac{1}{q_{j|t}} \cdot \frac{\partial q_{j|t}}{\partial y_t}$$

$$= -\sum_j p_{j|t} \cdot \frac{1}{q_{j|t}} \cdot \frac{d'_{tj} \cdot (\sum_{k\neq t} d_{tk}) - d_{tj} \cdot (\sum_{k\neq t} d'_{tk})}{(\sum_{k\neq t} d_{tk})^2}$$

$$= -\sum_j p_{j|t} \cdot \frac{1}{q_{j|t}} \cdot \frac{-2(\mathbf{y}_t - \mathbf{y}_j) \cdot d_{tj} \cdot (\sum_{k\neq t} d_{tk}) - d_{tj} \cdot (\sum_{k\neq t} d'_{tk})}{(\sum_{k\neq t} d_{tk})^2}$$

$$= 2\sum_j p_{j|t} \cdot (\mathbf{y}_t - \mathbf{y}_j) + \sum_j p_{j|t} \cdot \frac{\sum_{k\neq t} d'_{tk}}{\sum_{k\neq t} d_{tk}} \qquad (d'_{tt} = 0, \sum_j p_{j|t} = 1)$$

$$= 2\sum_j p_{j|t} \cdot (\mathbf{y}_t - \mathbf{y}_j) + \sum_j \cdot \frac{d'_{tj}}{\sum_{k\neq t} d_{tk}}$$

Stochastic Neighbor Embedding

- Gradient of the cost function ②

$$= 2\sum_j p_{j|t} \cdot (\mathbf{y}_t - \mathbf{y}_j) + \sum_j \cdot \frac{d'_{tj}}{\sum_{k\neq t} d_{tk}}$$

$$= 2\sum_j p_{j|t} \cdot (\mathbf{y}_t - \mathbf{y}_j) - 2\sum_j (\mathbf{y}_t - \mathbf{y}_j) \cdot \frac{d_{tj}}{\sum_{k\neq t} d_{tk}}$$

$$= 2\sum_j p_{j|t} \cdot (\mathbf{y}_t - \mathbf{y}_j) - 2\sum_j (\mathbf{y}_t - \mathbf{y}_j) \cdot q_{j|t}$$

$$= 2\sum_j (\mathbf{y}_t - \mathbf{y}_j)(p_{j|t} - q_{j|t})$$

Stochastic Neighbor Embedding

- Gradient of the cost function ③

$$\frac{\partial}{\partial y_t}\Big(-\sum_{i\neq t}\sum_{j\neq t} p_{i|j} log q_{i|j}\Big) = -\sum_{i\neq t}\sum_{j\neq t} p_{i|j}\cdot\frac{1}{q_{i|j}}\cdot\frac{\partial q_{i|j}}{\partial y_t}$$

$$= -\sum_{i\neq t}\sum_{j\neq t} p_{i|j}\cdot\frac{1}{q_{i|j}}\cdot\frac{d'_{ji}\cdot\sum_{k\neq j} d_{jk} - d_{ji}\cdot d'_{jt}}{(\sum_{k\neq j} d_{jk})^2} \qquad (d'_{ji}=0)$$

$$= -\sum_{i\neq t}\sum_{j\neq t} p_{i|j}\cdot\frac{1}{q_{i|j}}\cdot\frac{2(\mathbf{y}_t-\mathbf{y}_j)\cdot d_{ji}\cdot d_{jt}}{(\sum_{k\neq j} d_{jk})^2}$$

$$= -\sum_{i\neq t}\sum_{j\neq t} p_{i|j}\cdot\frac{1}{q_{i|j}}\cdot 2(\mathbf{y}_t-\mathbf{y}_j)\cdot q_{i|j}\cdot q_{t|j}$$

$$= -\sum_{i\neq t}\sum_{j\neq t} 2(\mathbf{y}_t-\mathbf{y}_j)\cdot p_{i|j}\cdot q_{t|j}$$

Stochastic Neighbor Embedding

- Gradient of the cost function ① + ③

$$\underbrace{\sum_i p_{t|i} \cdot 2(\mathbf{y}_t - \mathbf{y}_i)(1 - q_{t|i})}_{\text{Replace the subscript } i \text{ with } j} - \sum_{i \neq t} \sum_{j \neq t} 2(\mathbf{y}_t - \mathbf{y}_j) \cdot p_{i|j} \cdot q_{t|j}$$

$$= 2\sum_j (\mathbf{y}_t - \mathbf{y}_j) \cdot p_{t|j} - 2\sum_j (\mathbf{y}_t - \mathbf{y}_j) \cdot p_{t|j} \cdot q_{t|j} - 2\sum_{i \neq t} \sum_{j \neq t} (\mathbf{y}_t - \mathbf{y}_j) \cdot p_{i|j} \cdot q_{t|j}$$

$$= 2\sum_j (\mathbf{y}_t - \mathbf{y}_j) \cdot p_{t|j} - 2\sum_i \sum_j (\mathbf{y}_t - \mathbf{y}_j) \cdot p_{i|j} \cdot q_{t|j}$$

$$= 2\sum_j (\mathbf{y}_t - \mathbf{y}_j) \cdot p_{t|j} - 2\sum_j \sum_i p_{i|j} \cdot (\mathbf{y}_t - \mathbf{y}_j) \cdot q_{t|j} \qquad (\sum_i p_{i|j} = 1)$$

$$= 2\sum_j (\mathbf{y}_t - \mathbf{y}_j) \cdot p_{t|j} - 2\sum_j (\mathbf{y}_t - \mathbf{y}_j) \cdot q_{t|j} = 2\sum_j (\mathbf{y}_t - \mathbf{y}_j)(p_{t|j} - q_{t|j})$$

Stochastic Neighbor Embedding

- Gradient of the cost function ① + ② + ③

$$2\sum_j (\mathbf{y}_t - \mathbf{y}_j)(p_{j|t} - q_{j|t}) + 2\sum_j (\mathbf{y}_t - \mathbf{y}_j)(p_{t|j} - q_{t|j})$$

$$= 2\sum_j (\mathbf{y}_t - \mathbf{y}_j)(p_{t|j} - q_{t|j} + p_{j|t} - q_{j|t})$$

- Update the coordinate in the lower dimension to minimize the cost function
  - ✓ Gradient update with a momentum term

$$\mathcal{Y}^{(t+1)} = \mathcal{Y}^{(t)} + \eta \frac{\partial C}{\partial \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t)} - \mathcal{Y}^{(t-1)}\right)$$

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

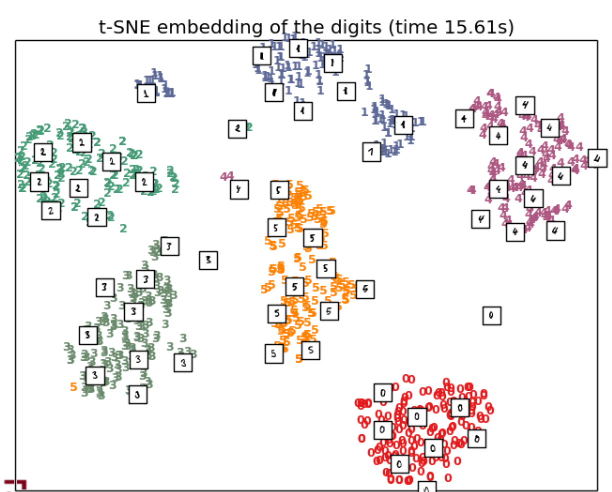
Stochastic Neighbor Embedding

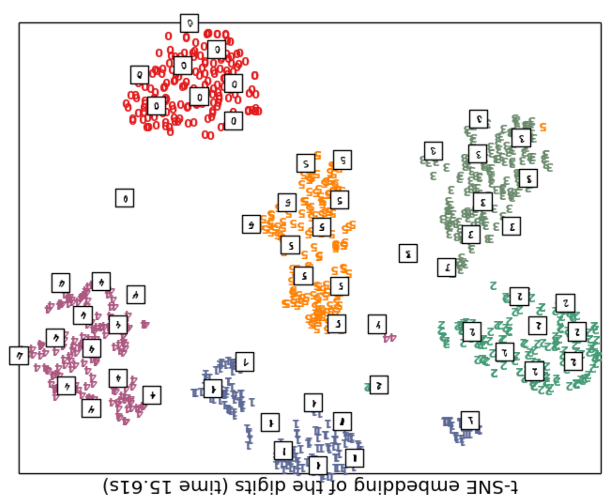
- From the original paper

$$\frac{\partial C}{\partial \mathbf{y}_i} = 2\sum_j (\mathbf{y}_j - \mathbf{y}_i)(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$$

- In this lecture note

$$\frac{\partial C}{\partial \mathbf{y}_t} = 2\sum_j (\mathbf{y}_t - \mathbf{y}_j)(p_{t|j} - q_{t|j} + p_{j|t} - q_{j|t})$$





Symmetric SNE

- Turning conditional probabilities into pairwise probabilities

$$p_{ij} = \frac{e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2}}}{\sum_{k \neq l} e^{-\frac{||\mathbf{x}_k - \mathbf{x}_l||^2}{2\sigma_i^2}}} \qquad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \qquad \sum_j p_{ij} > \frac{1}{2n}$$

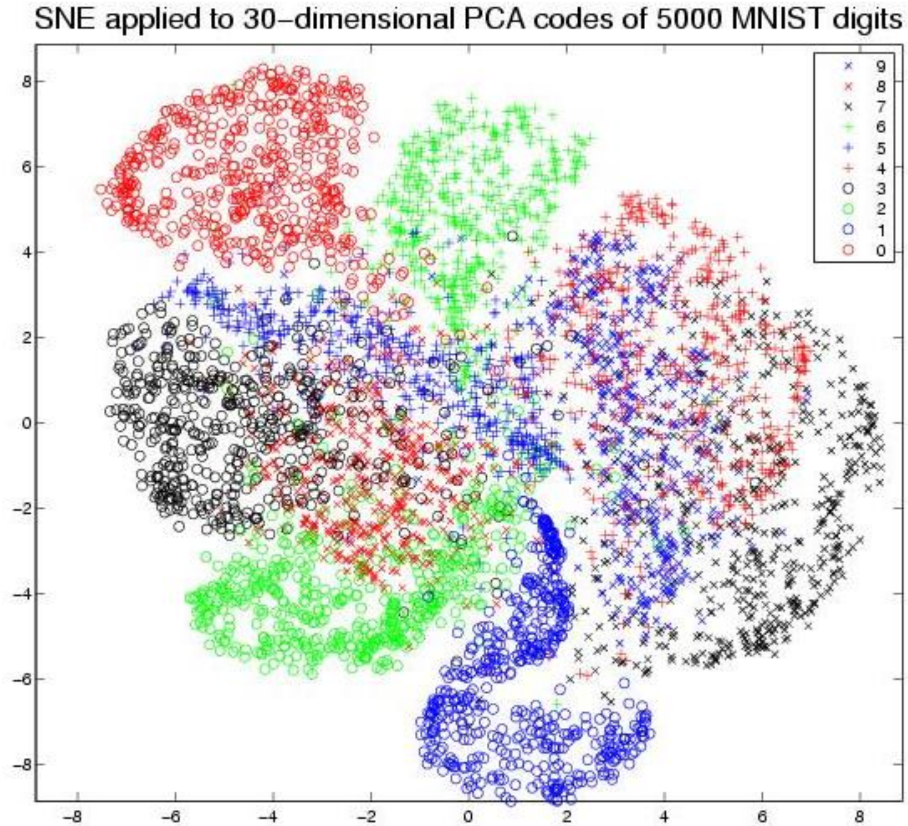
✓ Cost function and gradient

$$Cost = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_j (\mathbf{y}_j - \mathbf{y}_i)(p_{ij} - q_{ij})$$

Symmetric SNE

- Crowding problem
  - ✓ The area accommodating moderately distant data points is not large enough compared with the area accommodating nearby data points



t-SNE

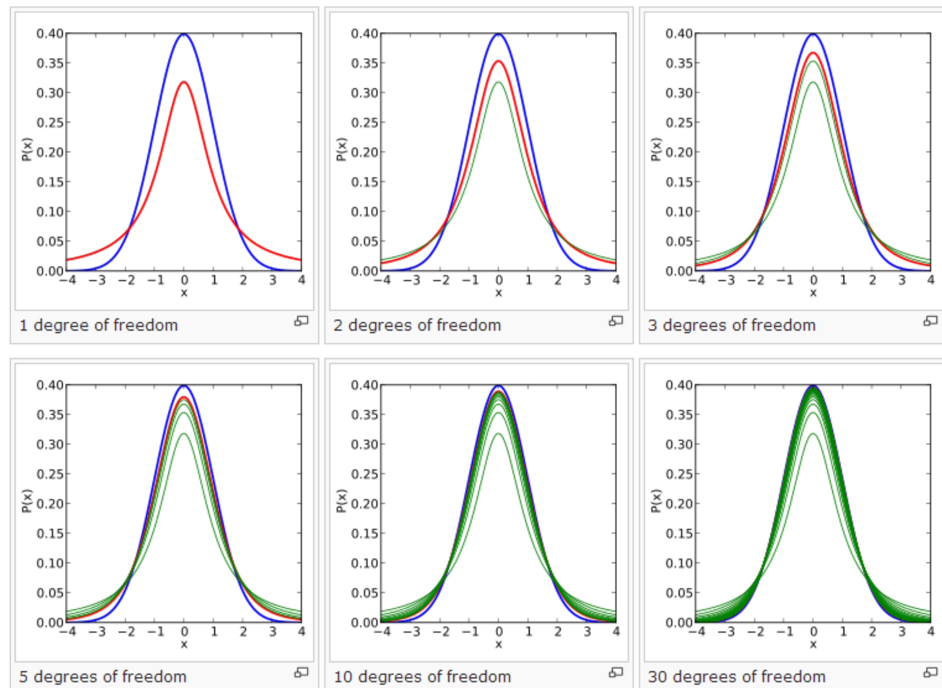
Maaten and Hinton (2008)

- Resolution to the Crowding Problem
  - ✓ Use a probability distribution that has much heavier tails than a Gaussian to convert distances into probabilities in the low-dimensional map
  - ✓ Student's t-distribution with one degree of freedom

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

$$\Gamma(n) = (n-1)!$$

$$q_{j|i} = \frac{(1+||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k\neq l}(1+||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}}$$



1 degree of freedom | 2 degrees of freedom | 3 degrees of freedom | 5 degrees of freedom | 10 degrees of freedom | 30 degrees of freedom

t-SNE

- Optimization of t-SNE

$$p_{ij} = \frac{e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2}}}{\sum_{k \neq l} e^{-\frac{||\mathbf{x}_k - \mathbf{x}_l||^2}{2\sigma_i^2}}} \qquad q_{ji} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}}$$

✓ Gradient:

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_j (\mathbf{y}_j - \mathbf{y}_i)(p_{ij} - q_{ij})(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}$$

t-SNE

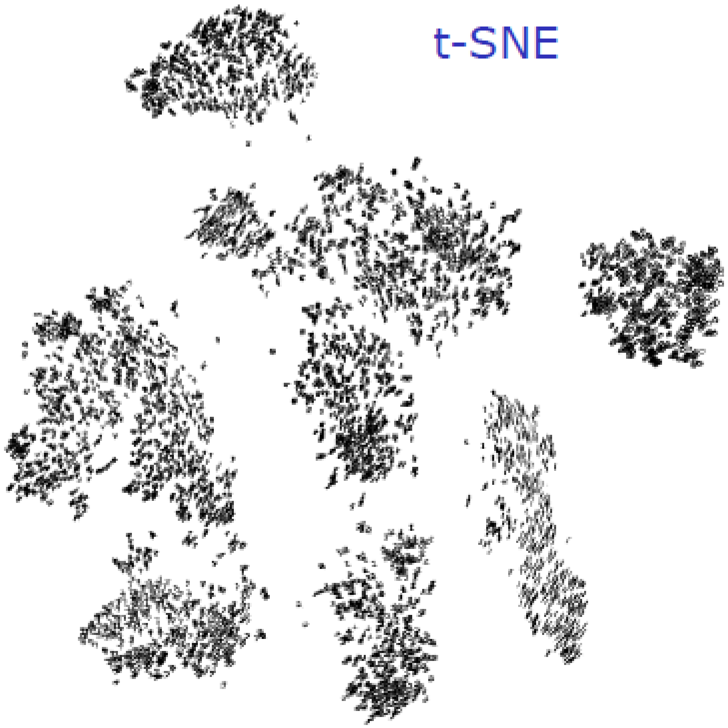
- t-SNE algorithm

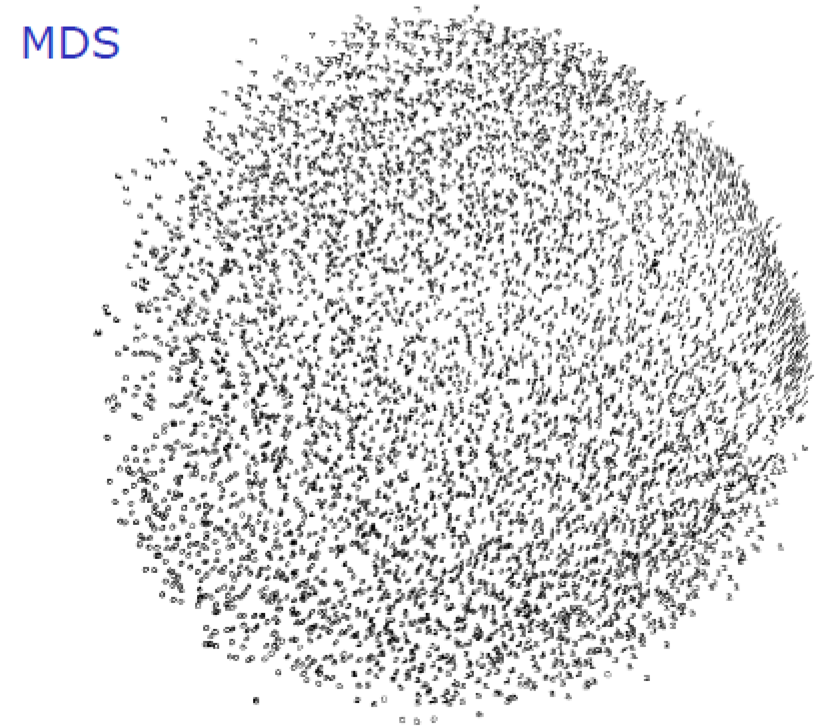
Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: data set $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$,
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, \dots, y_n\}$.
begin
  compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
  set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
  sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, \dots, y_n\}$ from $\mathcal{N}(0, 10^{-4} I)$
  for *t=1* **to** T **do**
    compute low-dimensional affinities q_{ij} (using Equation 4)
    compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
    set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)} \right)$
  end
end

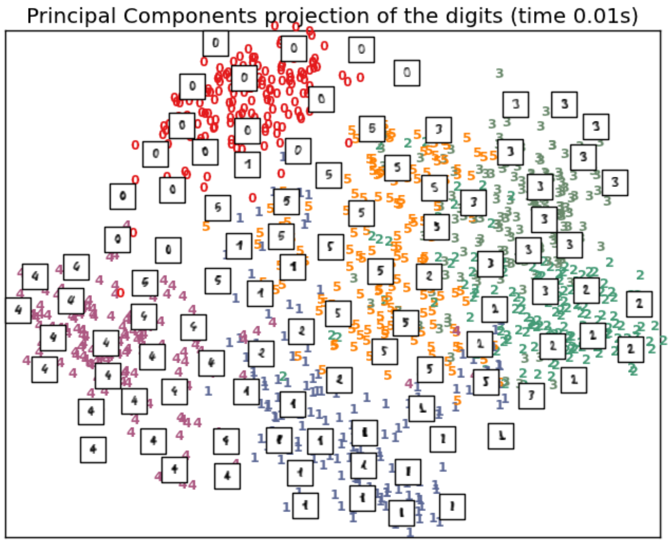
t-SNE vs. MDS

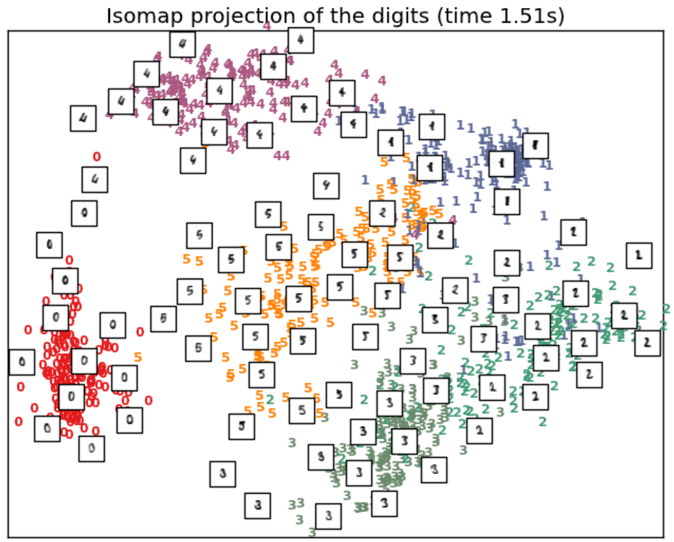
- MNIST dataset

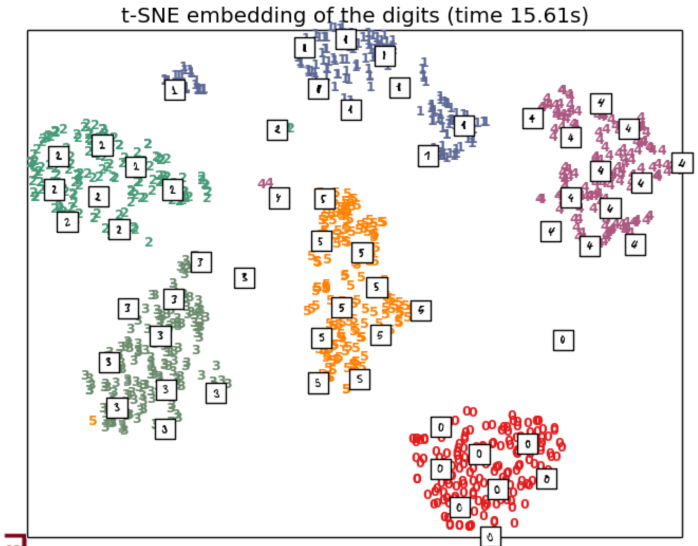




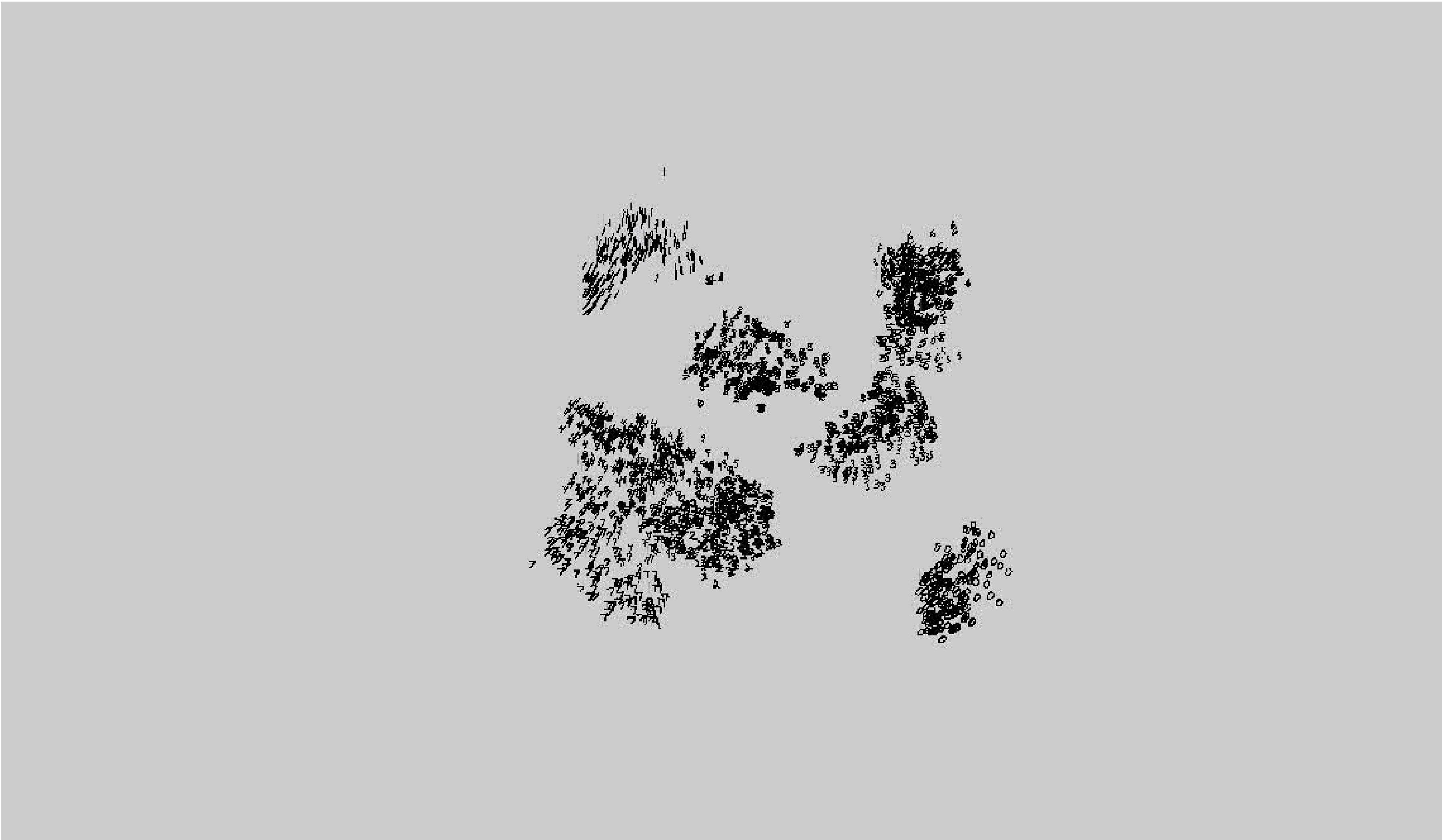
t-SNE Examples





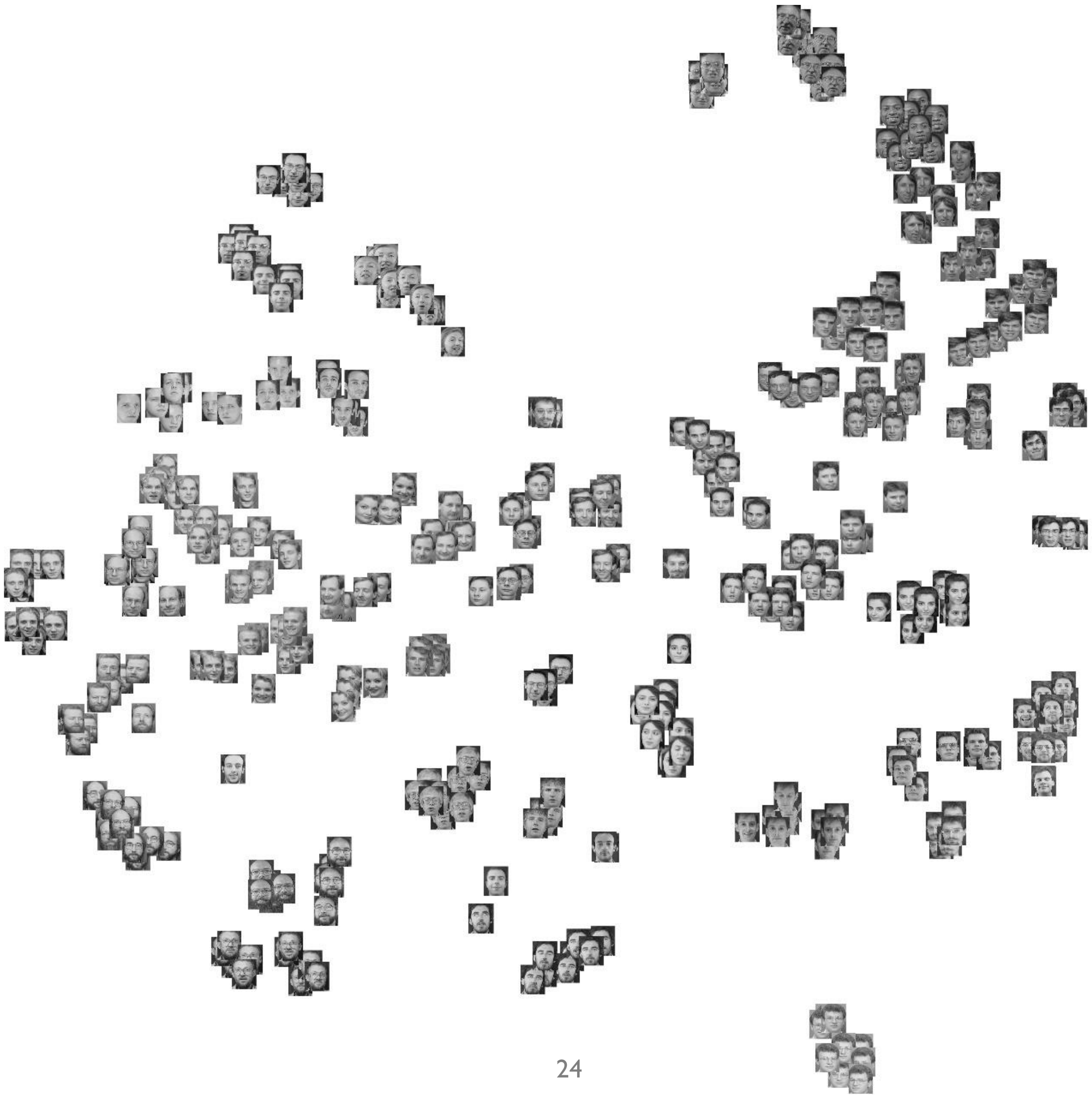


t-SNE Examples



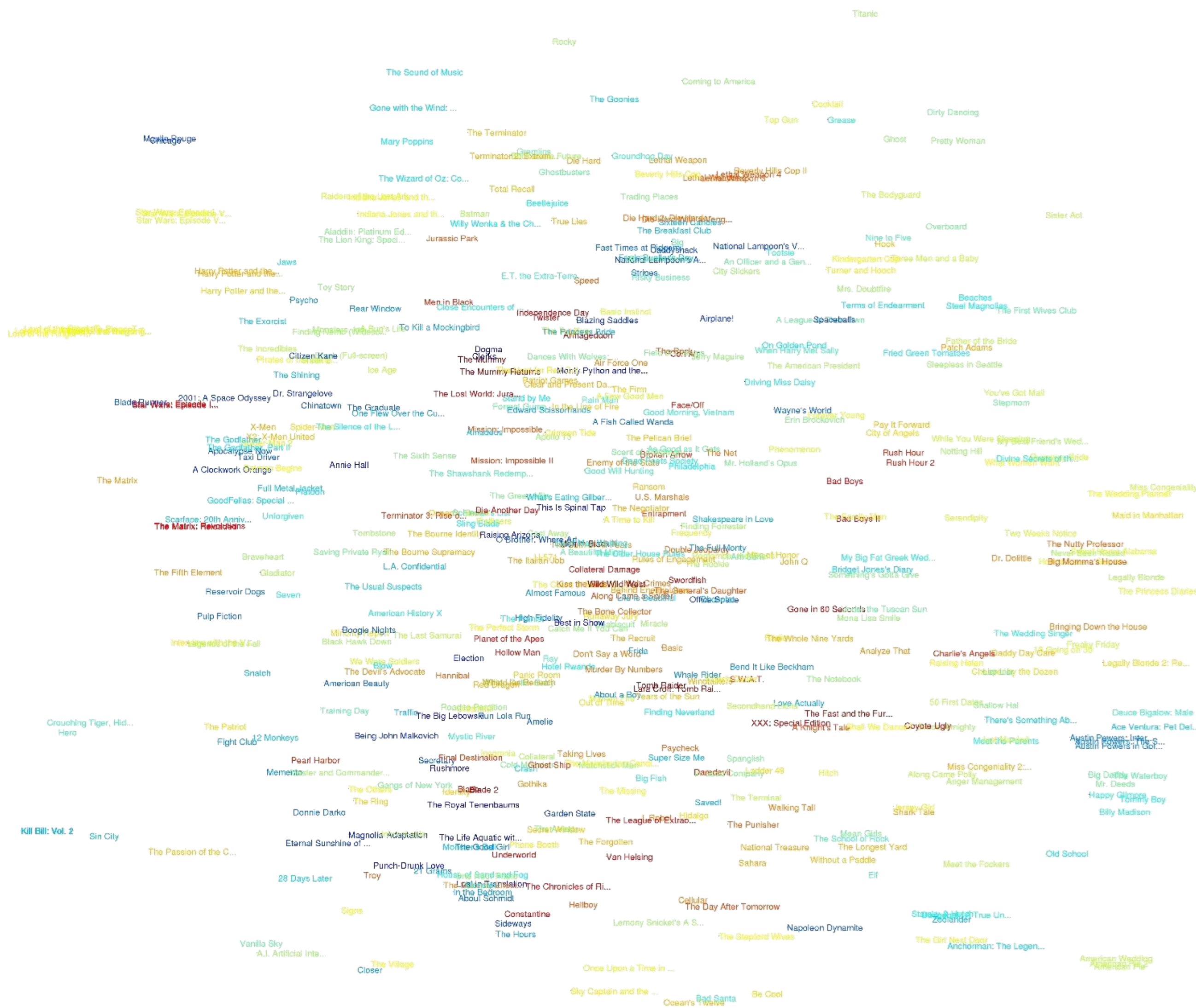
t-SNE Examples

- Olivetti faces datasets



t-SNE Examples

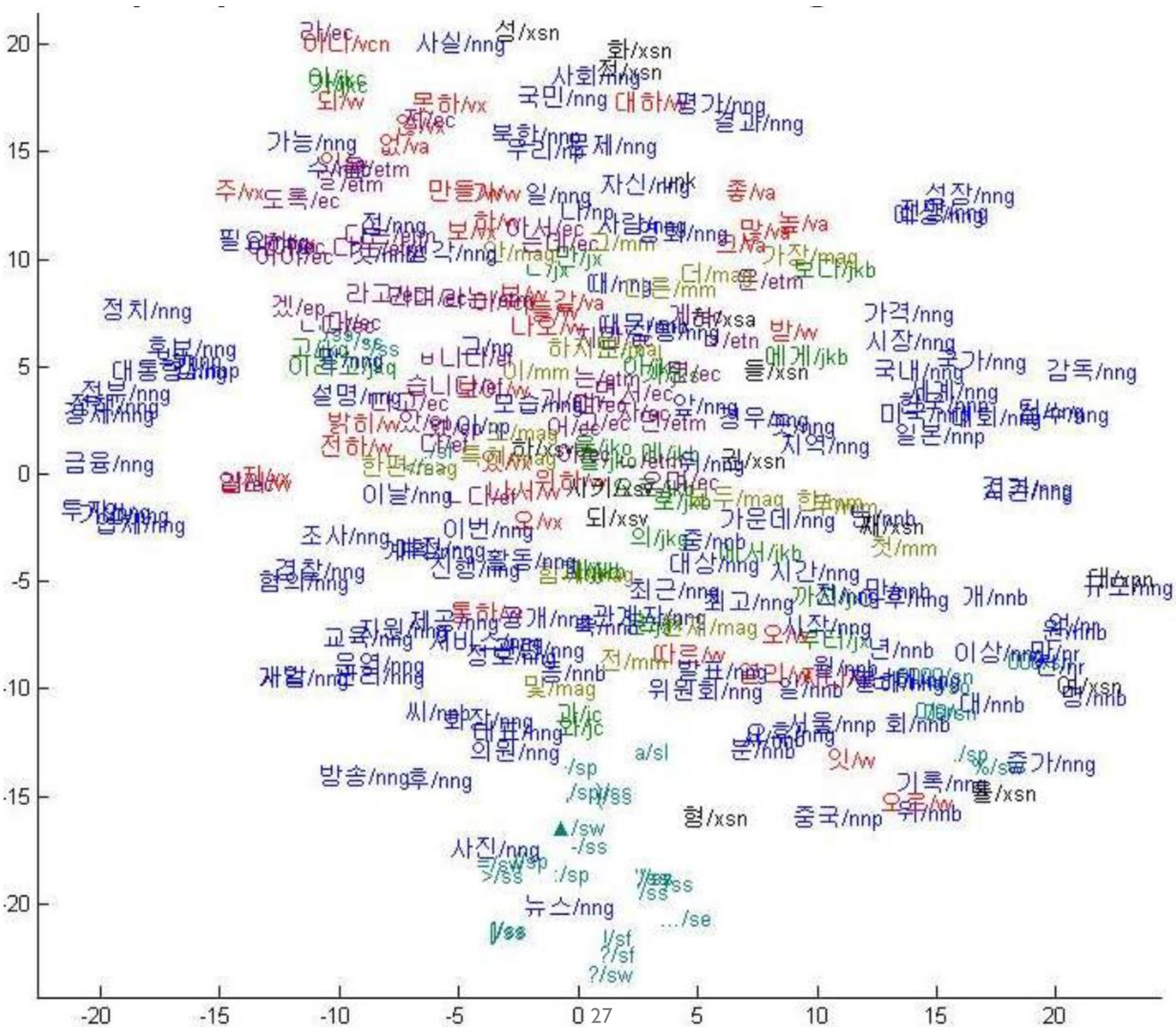
- Netflix dataset



t-SNE Examples

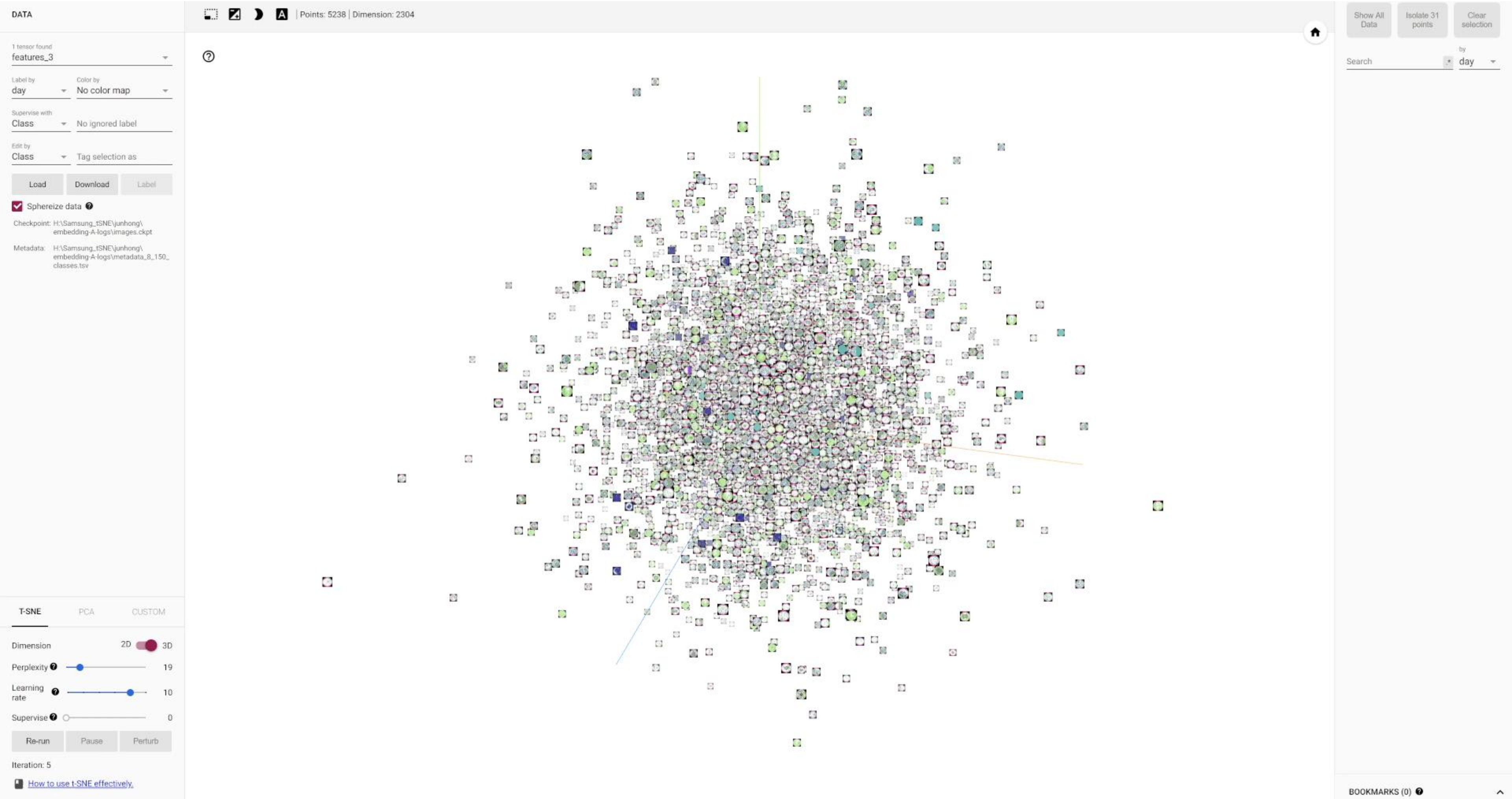
- CalTech-101

t-SNE Examples



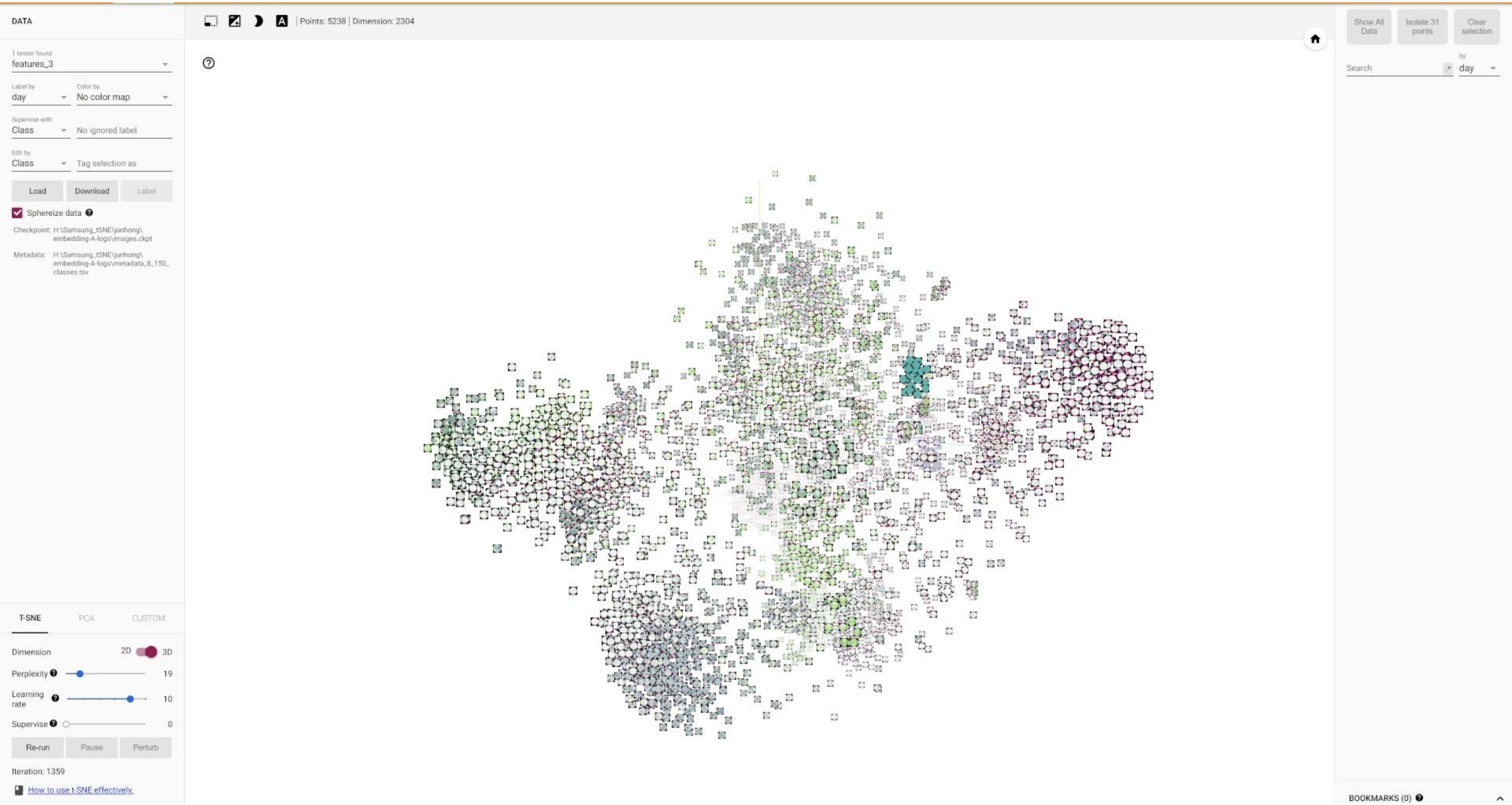
t-SNE Examples

- Wafer Bin Map (WBM) in Semiconductor Manufacturing



t-SNE Examples

- Wafer Bin Map (WBM) in Semiconductor Manufacturing




Q & A

References

Research Papers

- van der Maaten, L.J.P. and Hinton, G.E. (2008). Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research 9: 2579-2605.

Other materials

- Figure in the title page: https://nlml.github.io/in-raw-numpy/in-raw-numpy-t-sne/