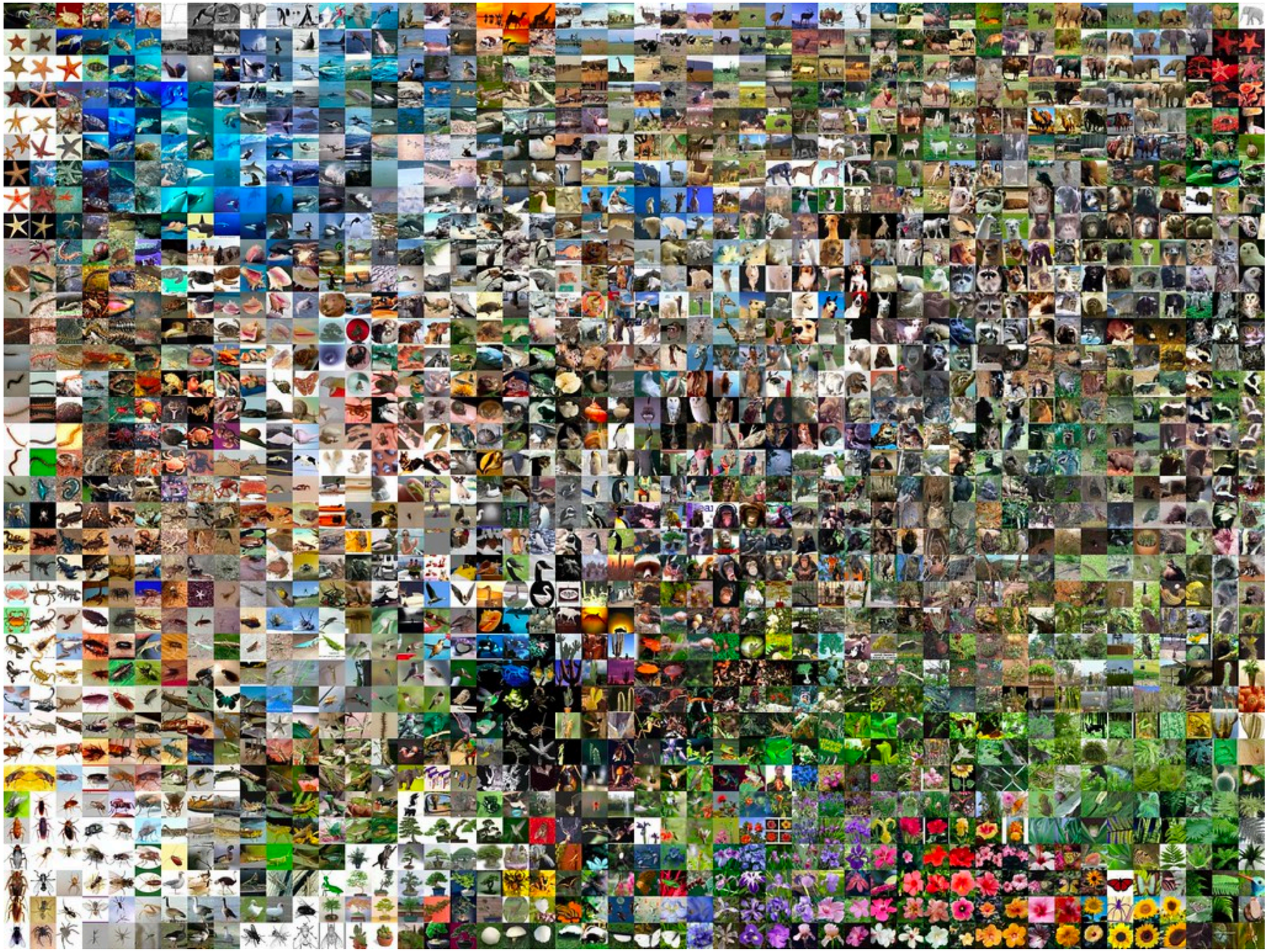
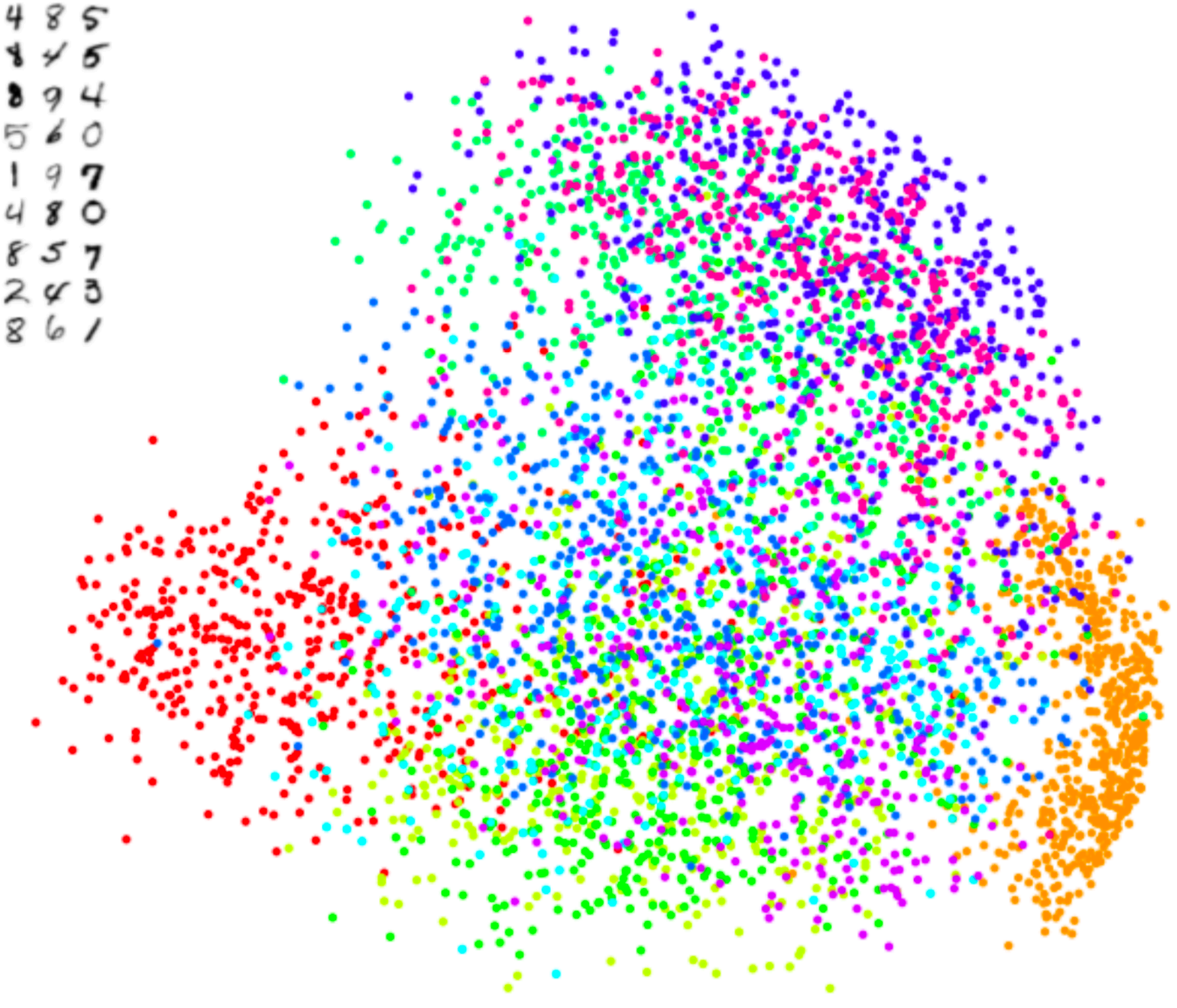
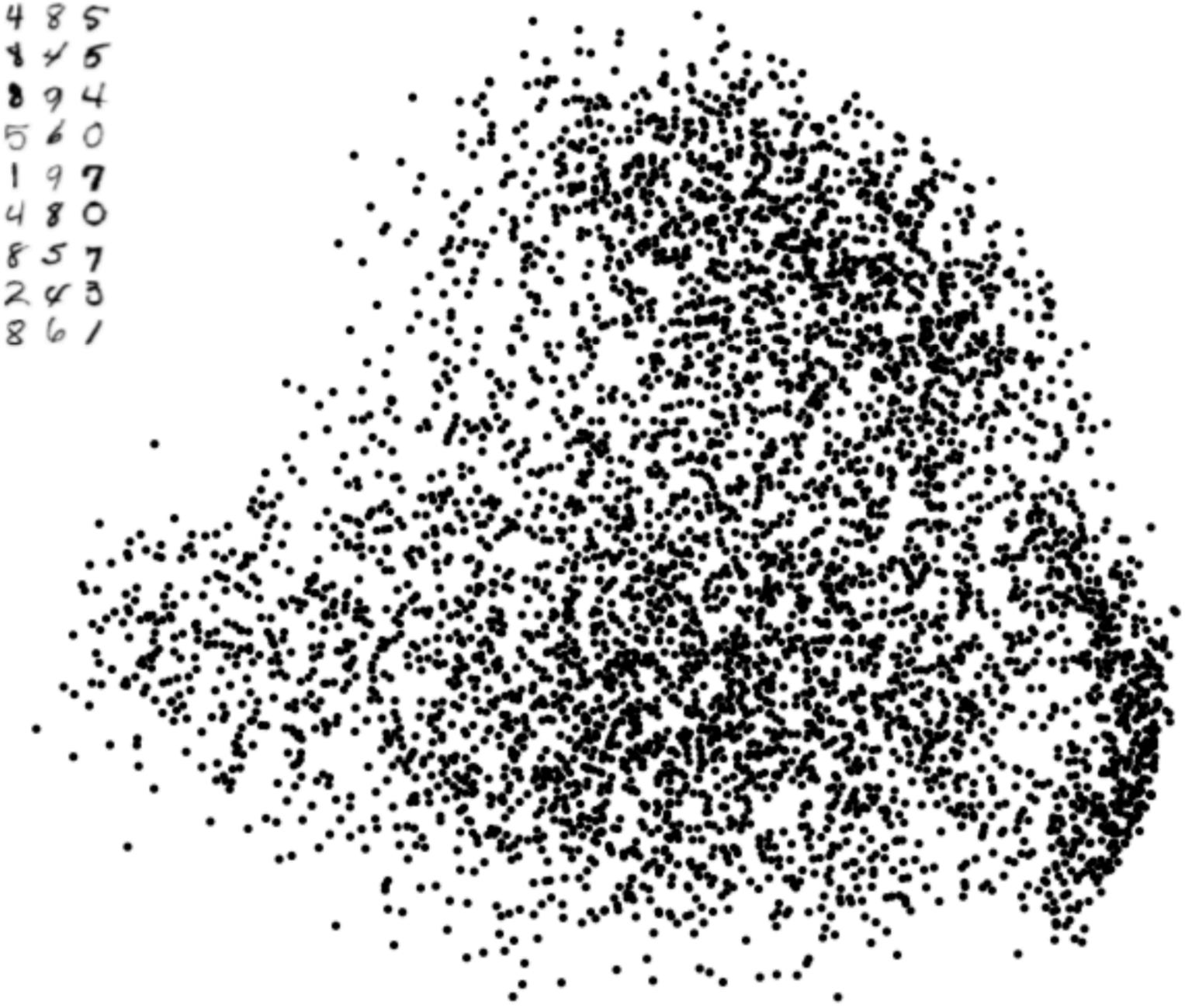


t-sne

mnist in pca

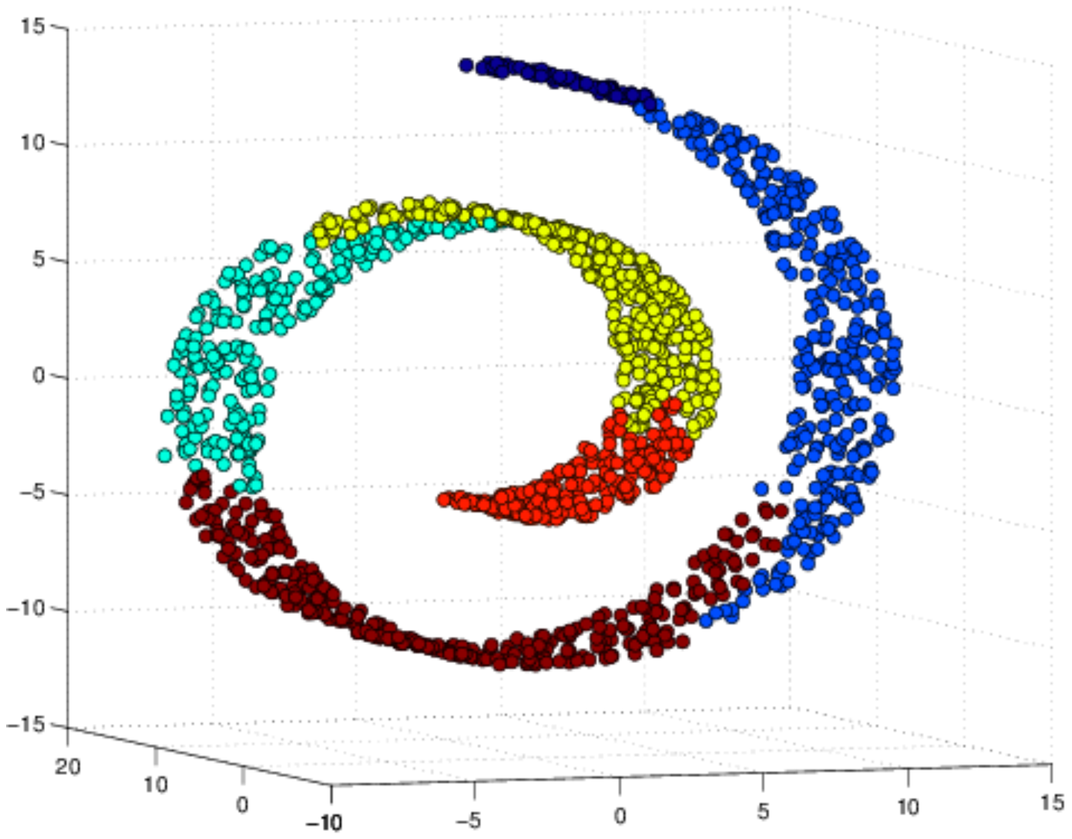


mnist in pca

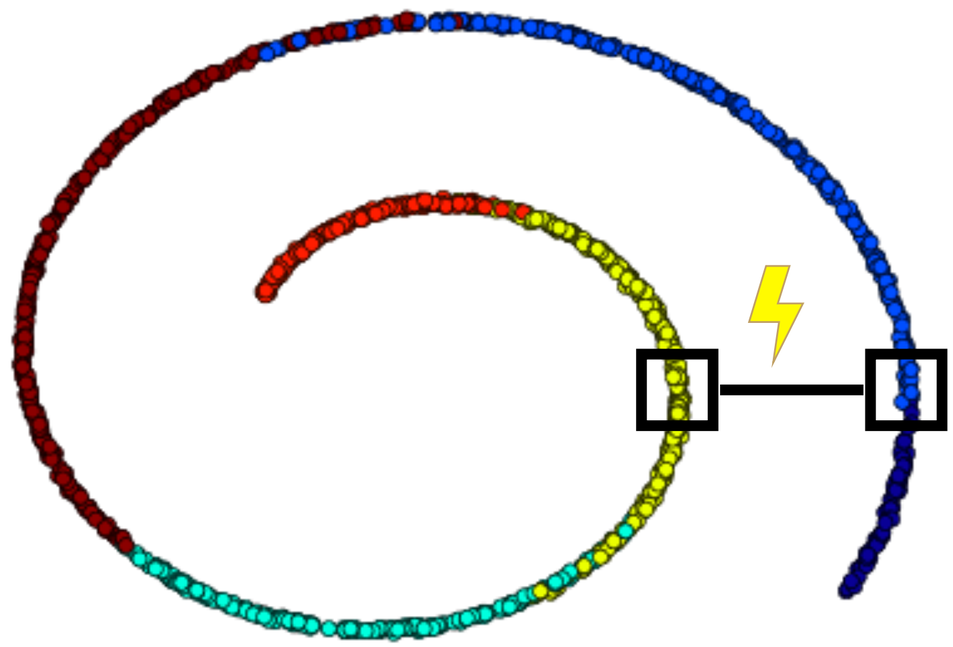


pca drawbacks

pca targets dimensionality, preserving large pairwise distances in the map,
but cannot catch the structure of the data



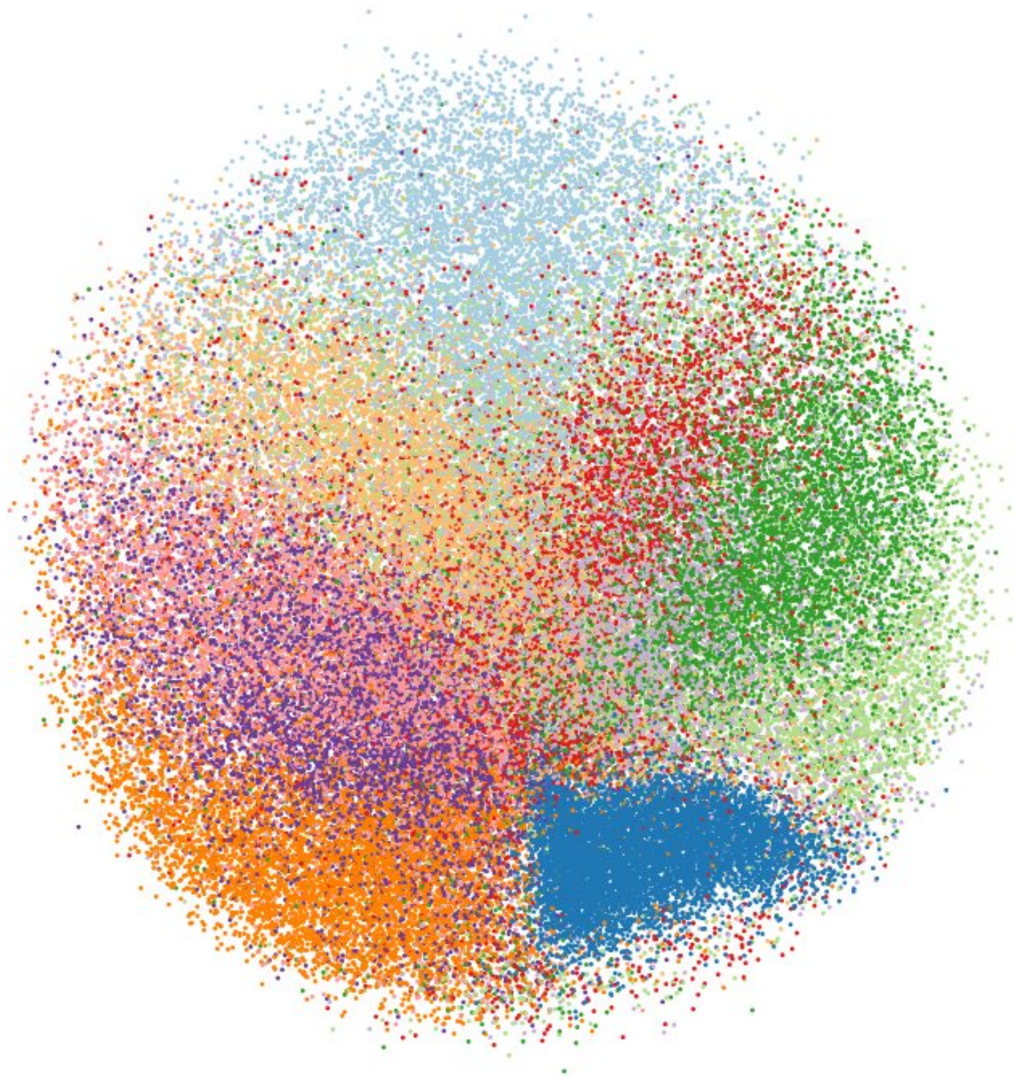
original
swiss roll data



pca
90% explained variance

mnist in mds

mnist in mds

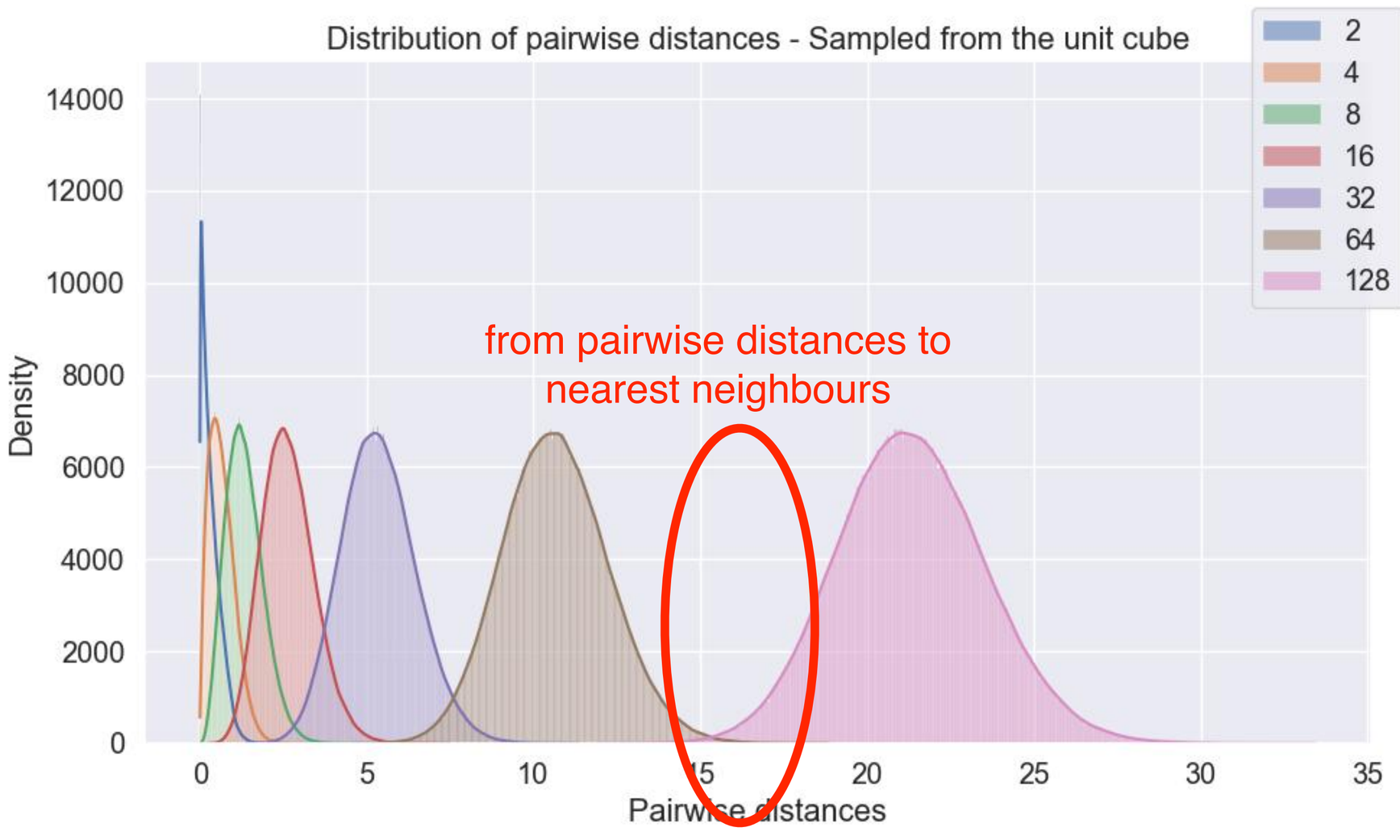


Lambert et al, 2022

mds drawbacks

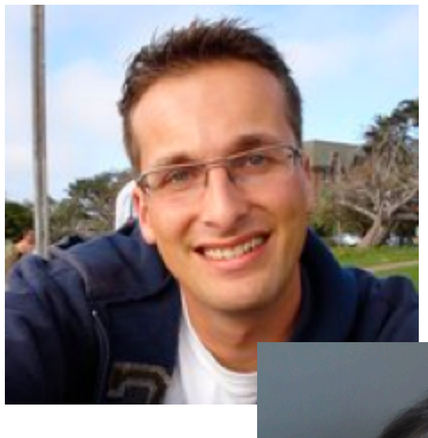
computational complexity

mds tries to preserve pairwise distances from high-dimensional space to low-dimensional space but this is not possible (curse of dimensionality)



t-sne

what is reliable are the very small euclidean distances between neighbouring points



t-distributed stochastic neighbor embedding

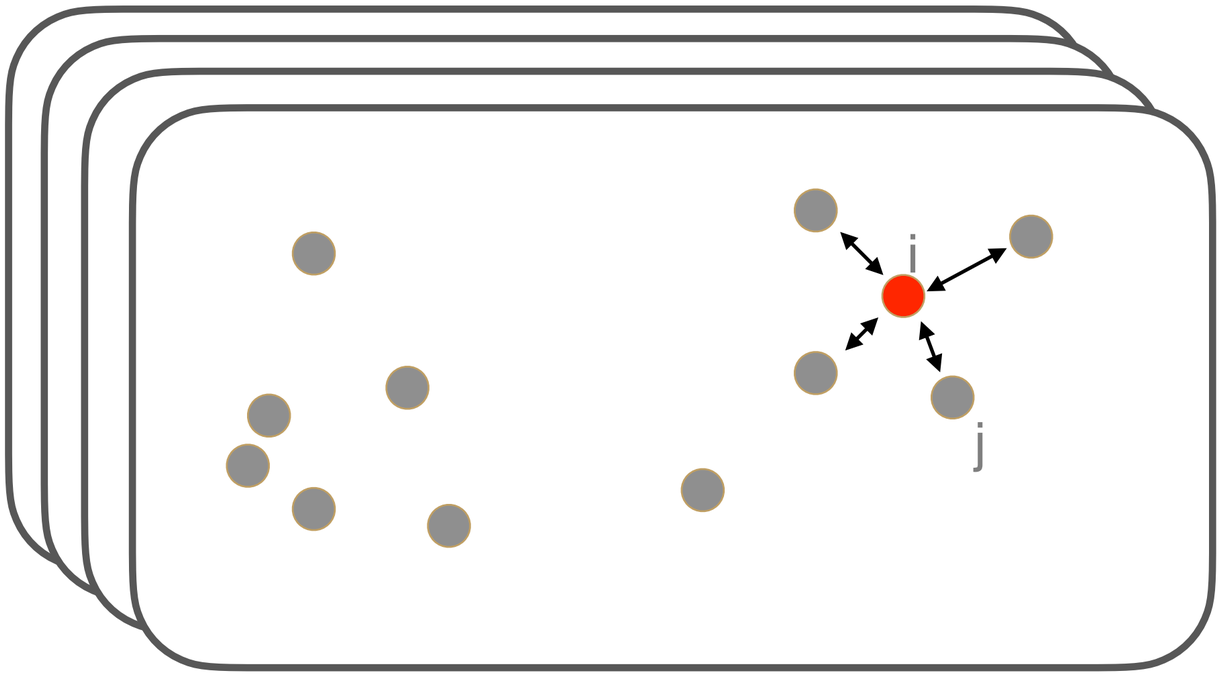
Journal of Machine Learning Research 9 (2008) 2579-2605 Submitted 5/08; Revised 9/08; Published 11/08

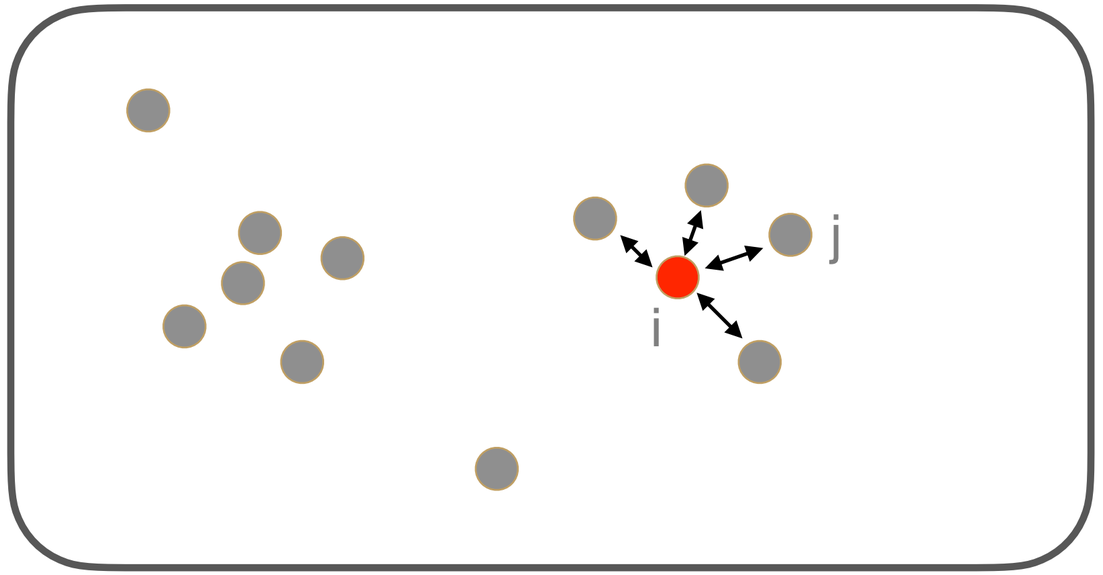
Visualizing Data using t-SNE

Laurens van der Maaten LVDMAATEN@GMAIL.COM
TiCC
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

focus on neighbours





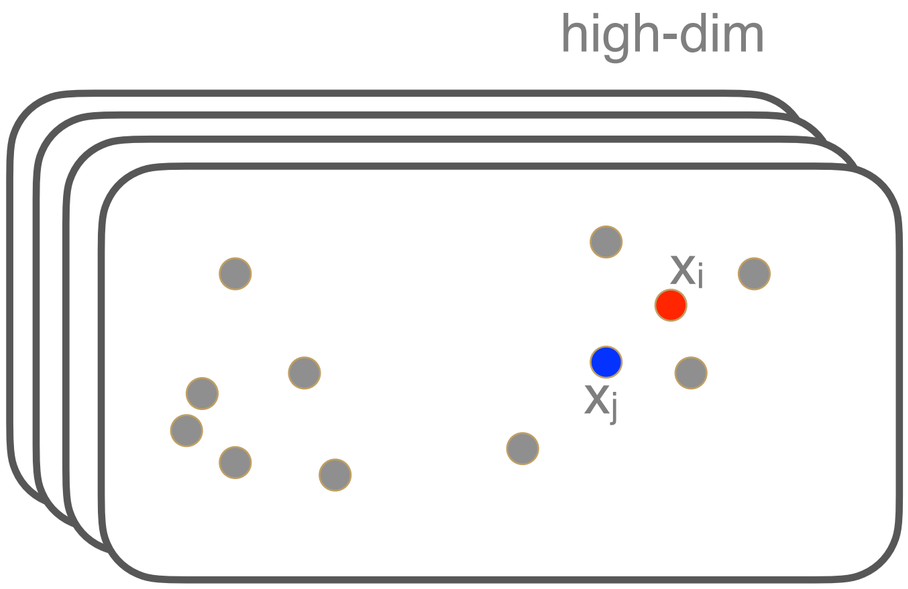
$$L = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

similarities* (affinities) in high-dim

similarities* (affinities) in low-dim

* similarities sum to 1

t-sne

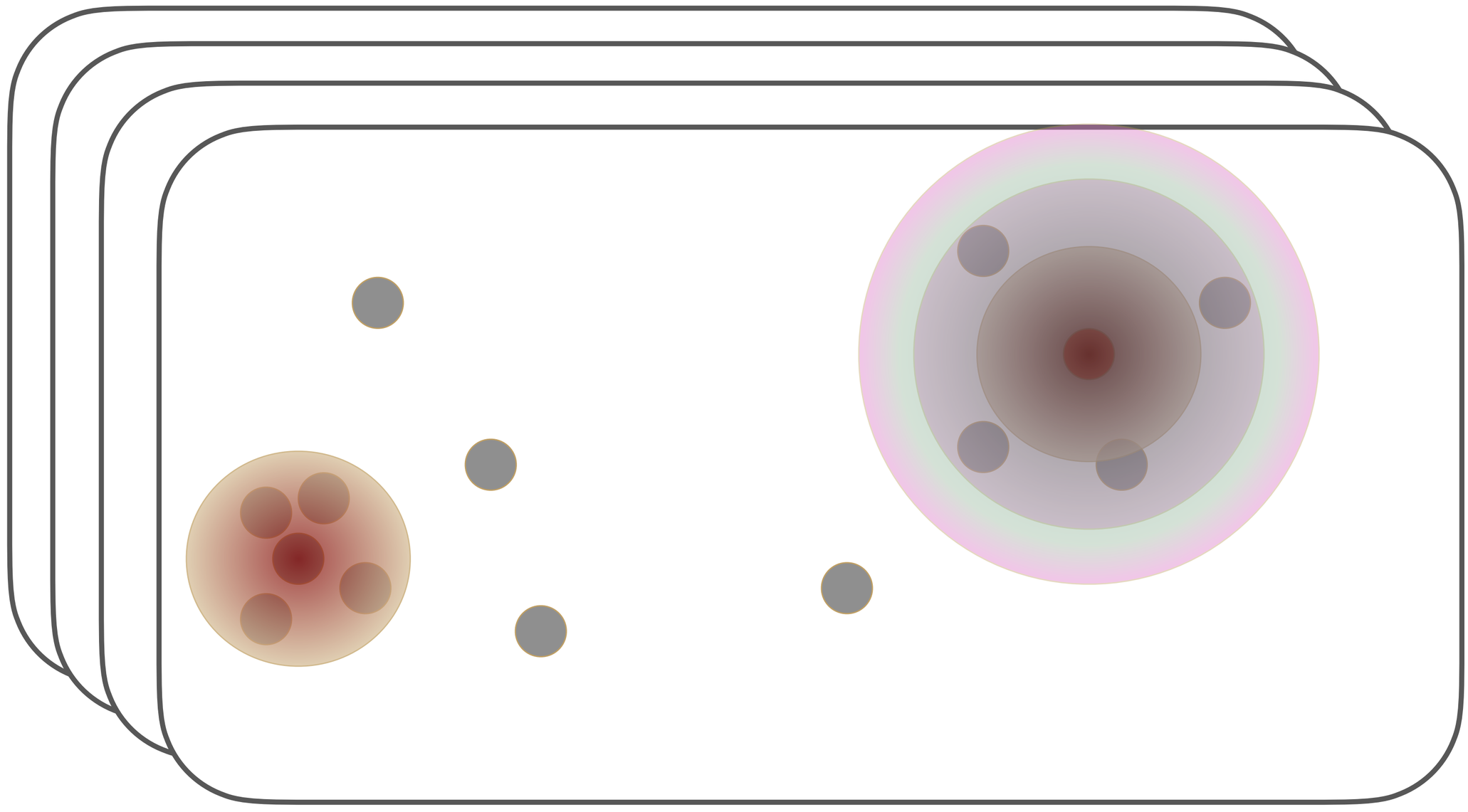


$$p_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq l} \exp\left(-\frac{\|x_k - x_l\|^2}{2\sigma_i^2}\right)}$$

gaussian

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

focus on neighbours



use a gaussian function to weight distances

perplexity

the only parameter to set is the variance σ_i for the high-dim gaussian of p_{ij}

no single value of σ_i can be optimal for all data points in the data set because the density of the data is likely to vary

in dense regions, a smaller value of σ_i is usually more appropriate than in sparser regions

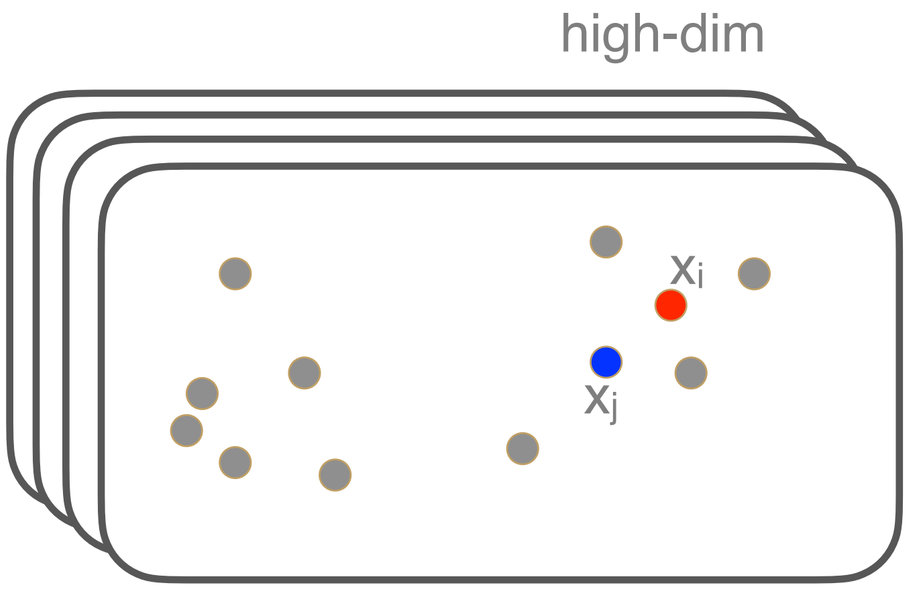
any particular value of σ_i induces a probability distribution P_i that has an entropy which increases as σ_i increases

tsne performs a binary search for the value of σ_i that produces a P_i with a fixed perplexity that is specified by the user

$$\text{Perp}(P_i) = 2^{H(P_i)} = 2^{-\sum_j p_{ij} \log_2 p_{ij}}$$

perplexity can be interpreted as a smooth measure of the effective number of neighbors, typical values are between 5 and 50

t-sne

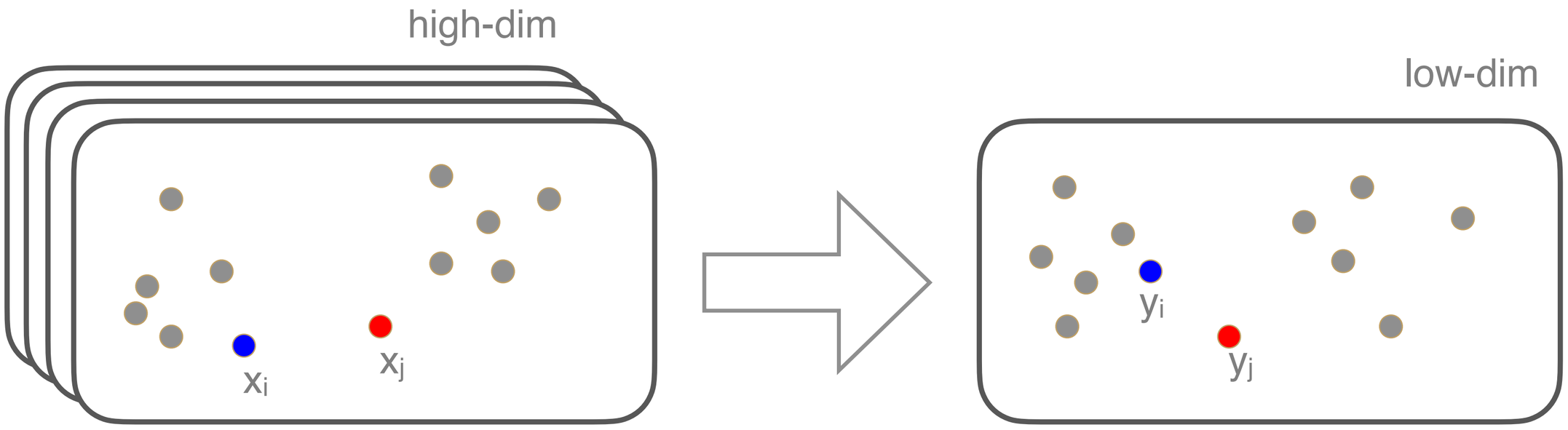


$$p_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq l} \exp\left(-\frac{\|x_k - x_l\|^2}{2\sigma_i^2}\right)}$$

gaussian

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

tsne



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

gaussian

$$q_{ij} = \frac{w_{ij}}{Z}, \quad w_{ij} = k(\|y_i - y_j\|), \quad Z = \sum_{k \neq l} w_{kl}$$

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

gaussian

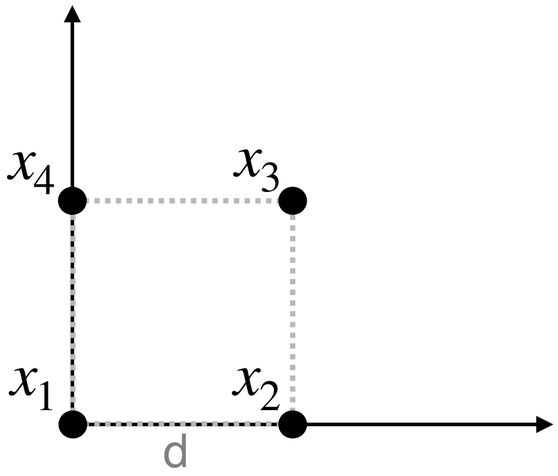
$$k(d) = 1/(1 + d^2) \qquad q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

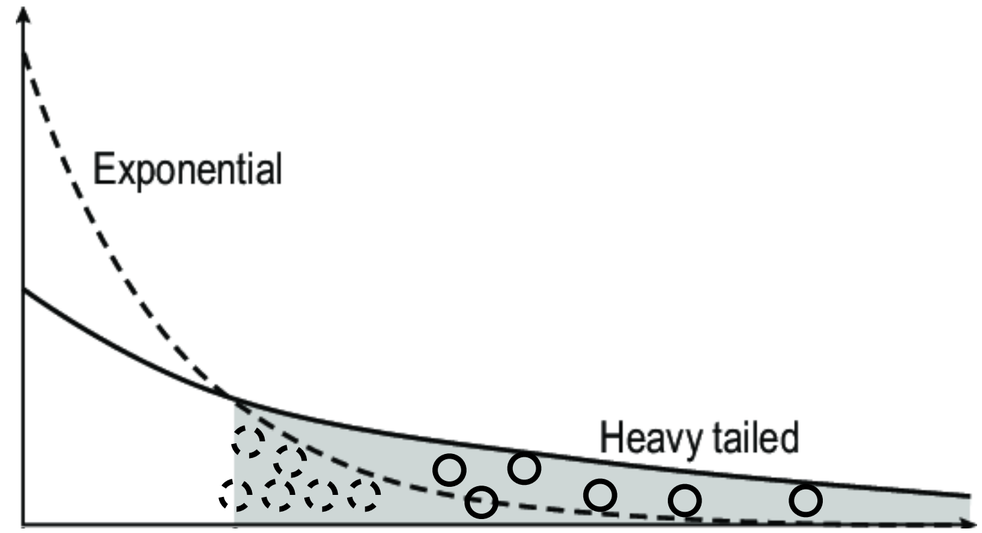
t-student

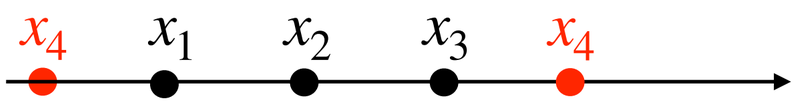
t-sne

crowding problem

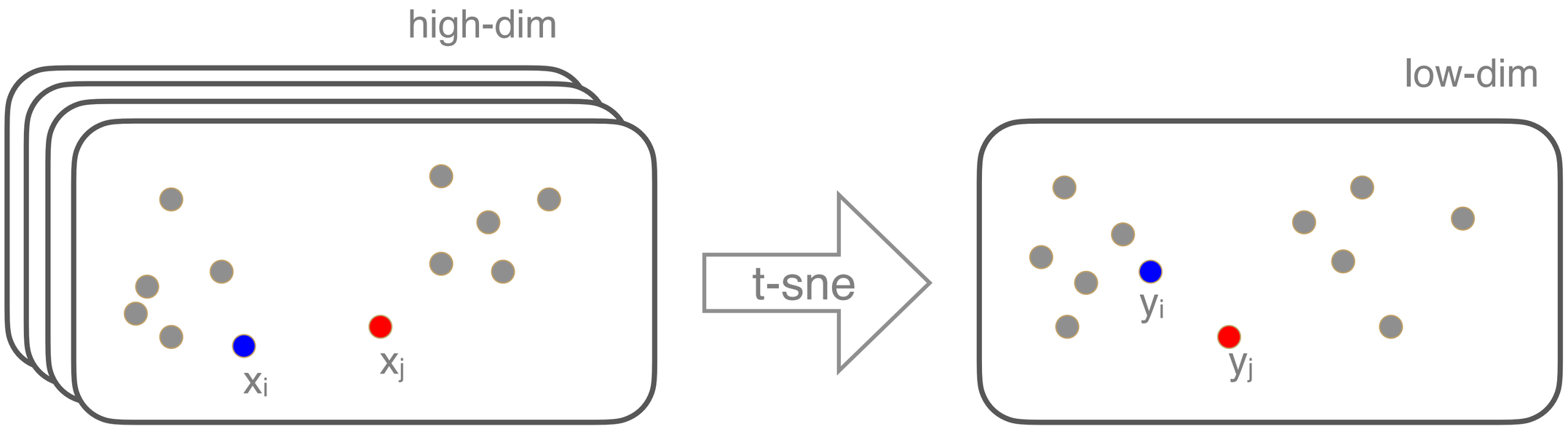
the volume of a sphere centered on datapoint i scales as r^m, where r is the radius and m the dimensionality of the sphere; if we want to model the small distances accurately in the map, most of the points that are at a moderate distance from datapoint i will have to be placed much too far away in the two-dimensional map.







t-sne



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

gaussian

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

t-student

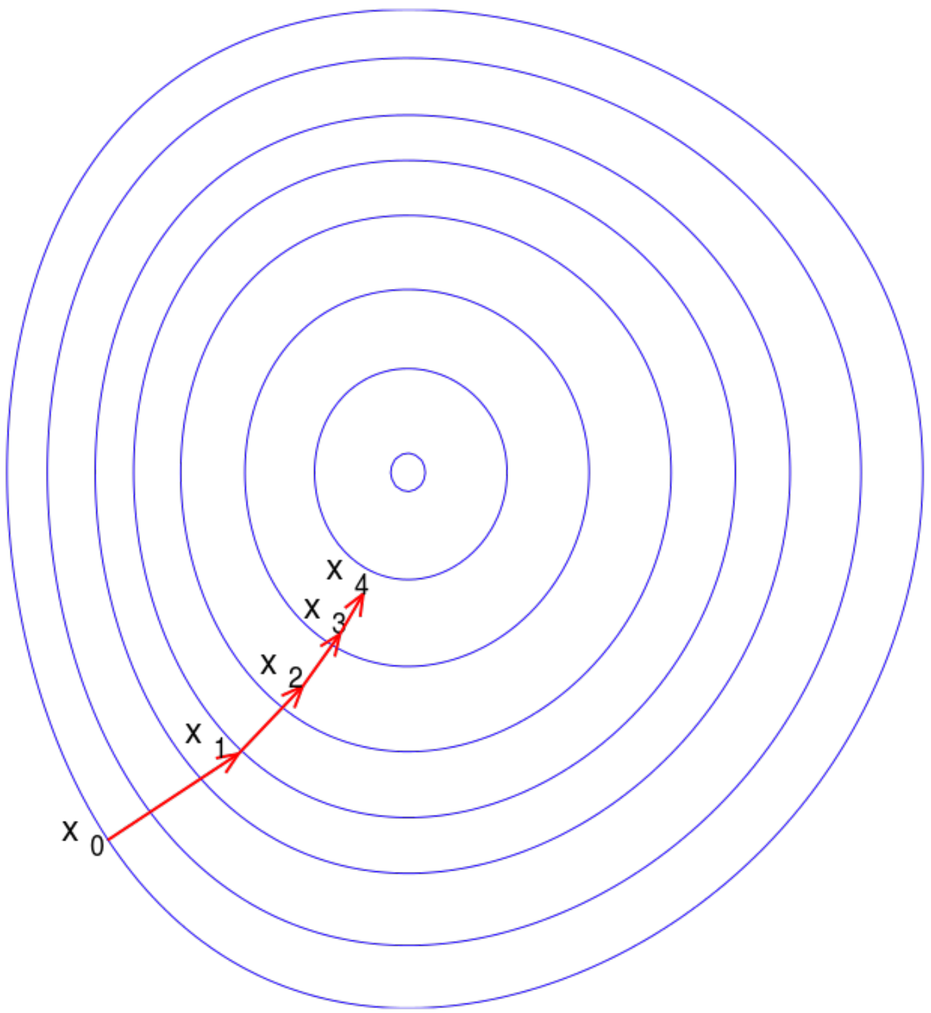
objective function: minimizing $C = \mathrm{KL}(P\|Q) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_{ij}}$

kullback-leibler divergence

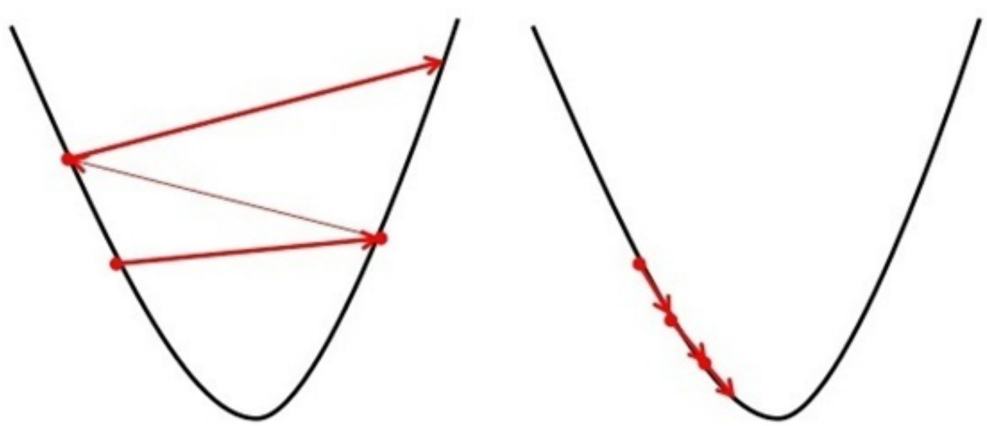
gradient descent

$f:\mathbf{R}^p \to \mathbf{R}$, defined and differentiable in a neighborhood of a point a,
decreases *fastest* if one goes from a in the direction of the negative gradient
of f at a, $-\nabla f(a)=-(\partial f/\partial x_1,\ldots,\partial f/\partial x_p)|_a$

if $a_{n+1}=a_n-\gamma\nabla f(a_n)$ for a small positive γ, then $f(a_{n+1})\leqq f(a_n)$



start with a guess x_0 of a local minimum for f
and consider the sequence $x_0,x_1,x_2,\ldots$
with $x_{n+1}=x_n-\gamma_n\nabla f(x_n)$:
then we have a monotonic sequence $f(x_0)\geqq f(x_1)\geqq f(x_2)\geqq\ldots$ so that the sequence
$x_0,x_1,x_2,\ldots$ can converge to the minimum of f



t-sne optimization

$$C = \mathrm{KL}(P\|Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}} = \sum_{i,j} p_{ij} \log p_{ij} - \sum_{i,j} p_{ij} \log q_{ij} \overset{\text{independent from q}}{=} - \sum_{i,j} p_{ij} \log q_{ij} = - \sum_{i,j} p_{ij} \log \frac{w_{ij}}{Z}$$

$$\sum_{i,j} p_{ij} = 1$$

$$= - \sum_{i,j} p_{ij} \log w_{ij} + \sum_{i,j} p_{ij} \log Z = - \sum_{i,j} p_{ij} \log w_{ij} + \log Z = - \sum_{i,j} p_{ij} \log w_{ij} + \sum_{i,j} \log w_{ij} =$$

$$C = - \sum_{i,j} p_{ij} \log w_{ij} + \sum_{i,j} \log w_{ij}$$

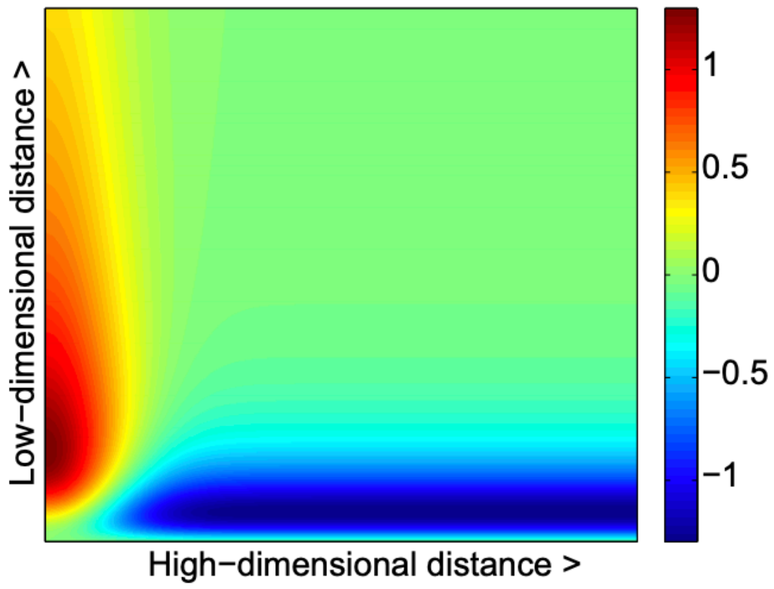
Attractive force:
close neighbours
attract each other

Repulsive force:
all points repulse
each other

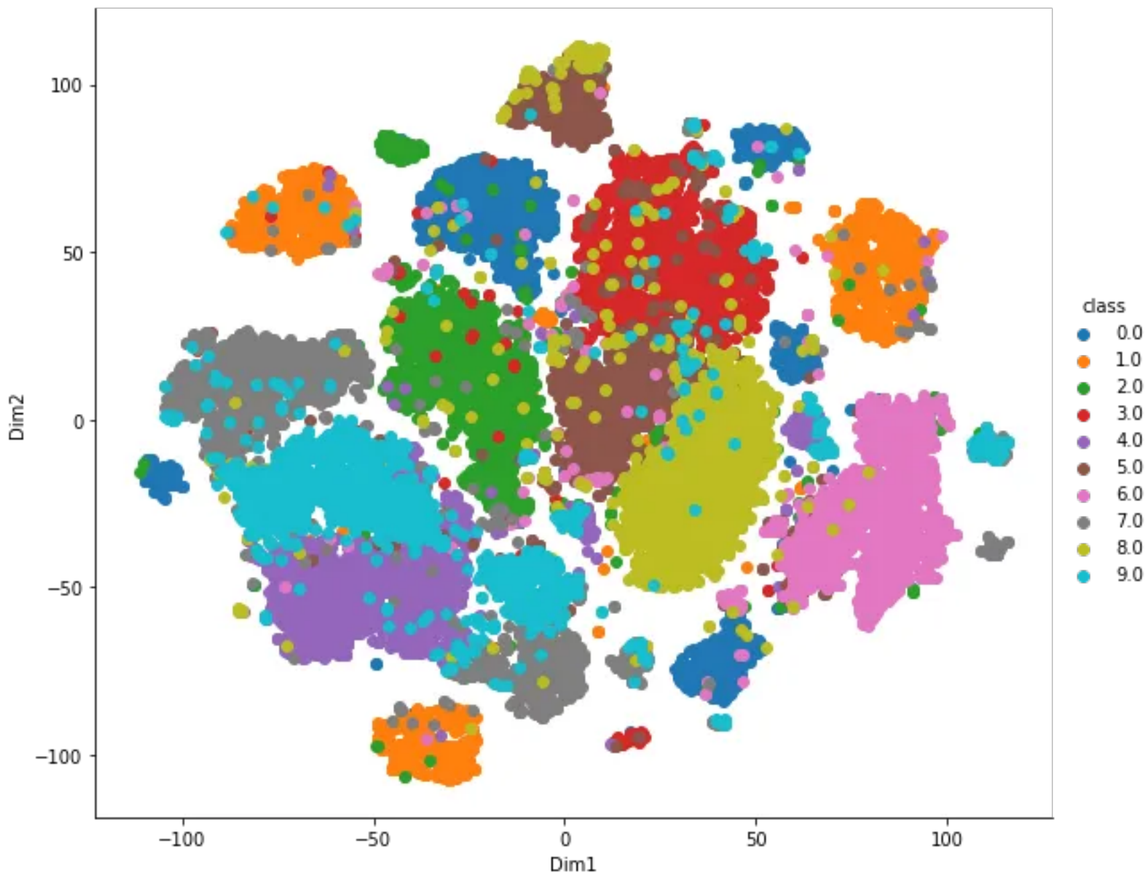
$$\frac{\delta C}{\delta y_i} = -2 \sum_j p_{ij} \frac{1}{w_{ij}} \frac{\delta w_{ij}}{\delta y_i} + 2 \frac{1}{Z} \sum_j \frac{\delta w_{ij}}{\delta y_i} \sim \sum_j p_{ij} w_{ij} \left(y_i - y_j \right) - \frac{1}{Z} \sum_j w_{ij}^2 \left(y_i - y_j \right)$$

tsne optimization

$$\frac{\delta C}{\delta y_i} \sim \sum_j p_{ij} w_{ij} \left(y_i - y_j \right) - \frac{1}{Z} \sum_j w_{ij}^2 \left(y_i - y_j \right)$$

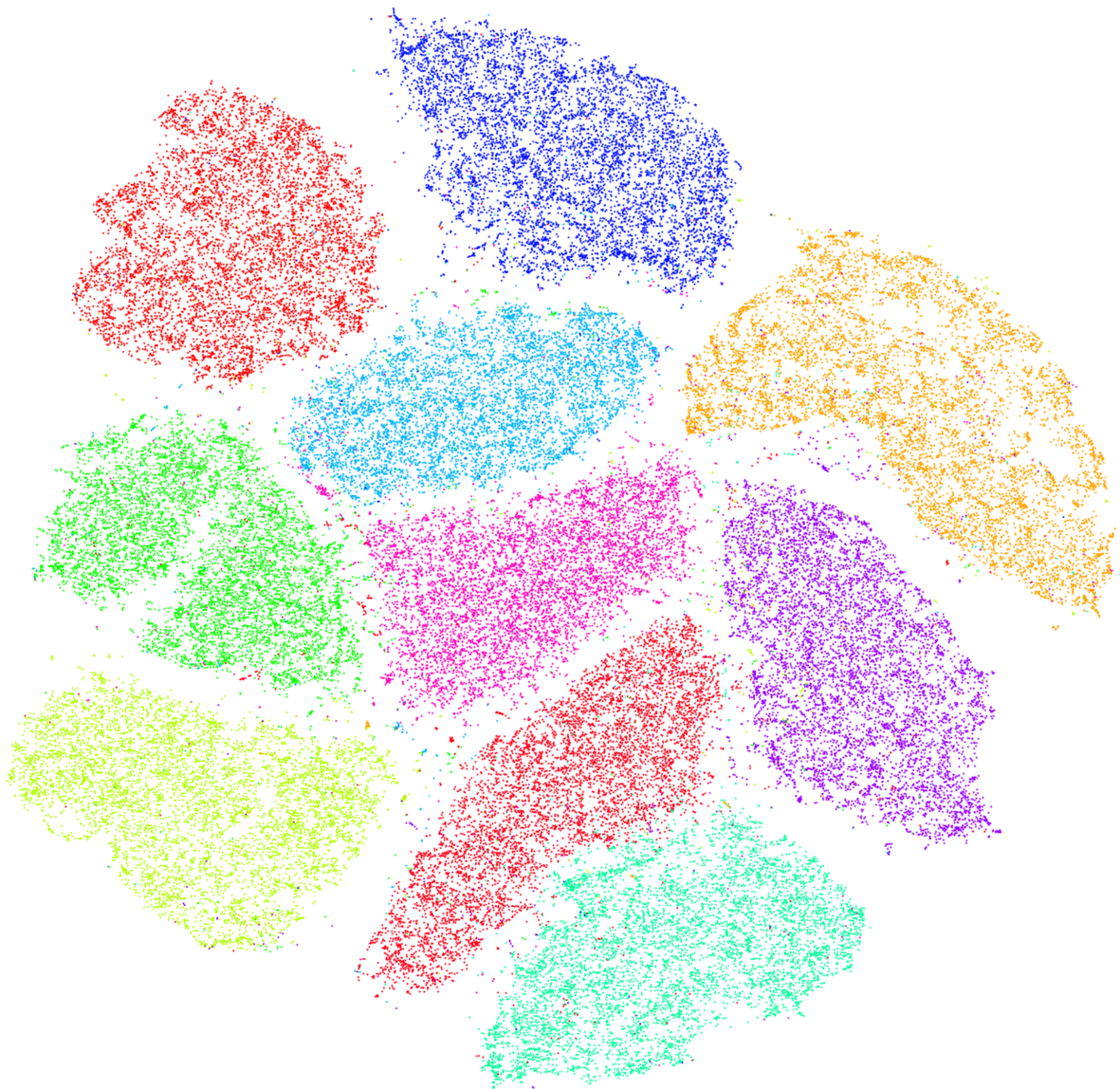


early compression:
force the map points to stay close together at the start of the optimization, implemented by adding an additional l2-penalty to the cost function that is proportional to the sum of square distances of the map points from the origin



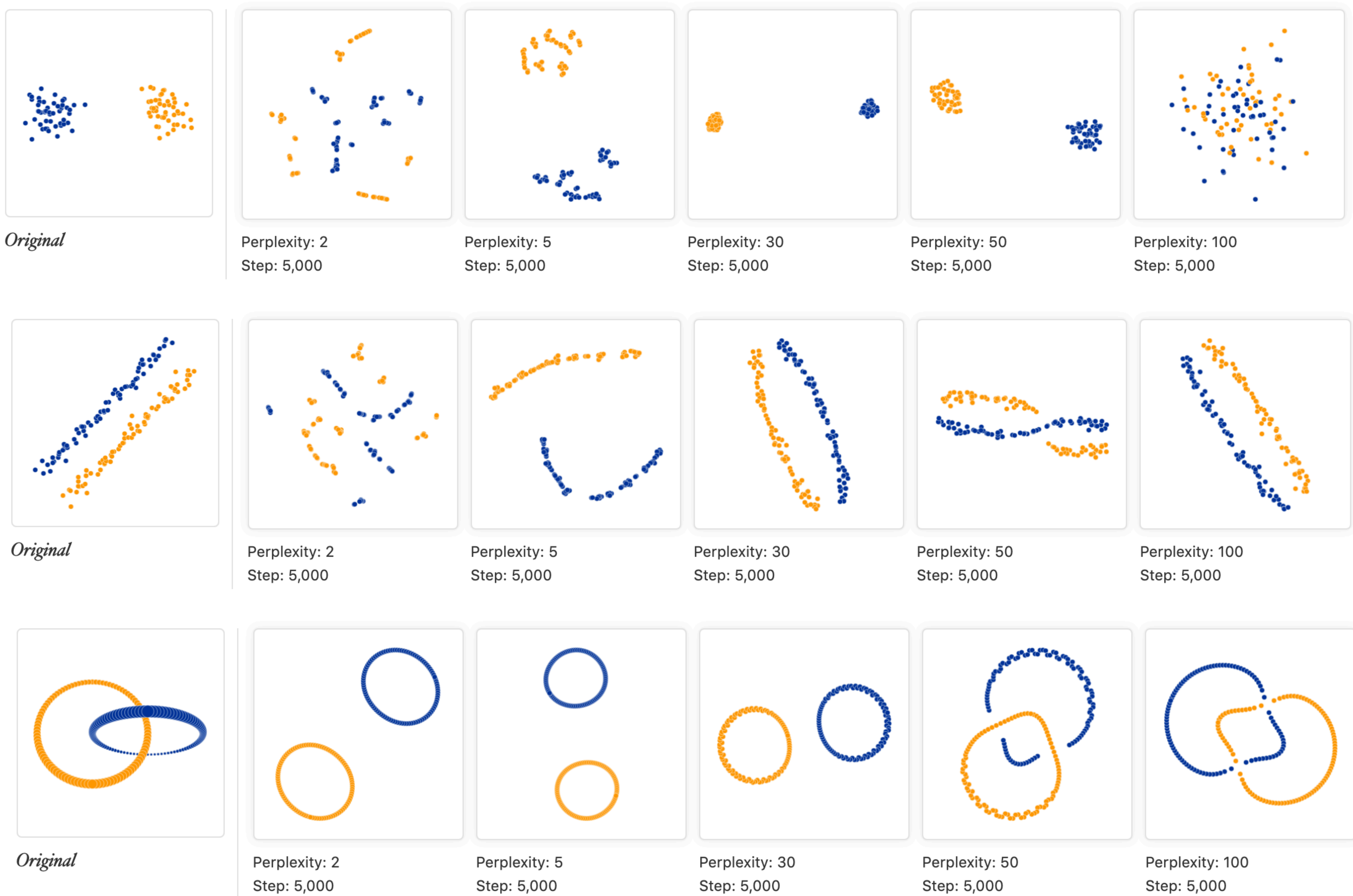
early exaggeration:
multiply all of the p_{ij}'s by, for example, 4, in the initial stages of the optimization, modeling the large p_{ij}'s by fairly large q_{ij}'s; natural clusters in the data tend to form tight widely separated clusters in the map

mnist in t-sne



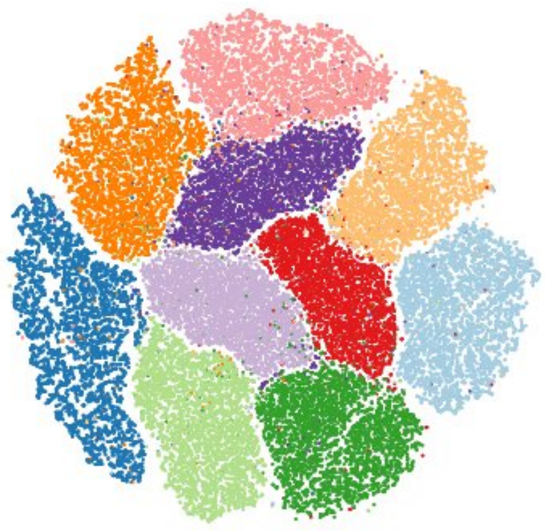
perplexity

[https://distill.pub/2016/misread-tsne/]



perplexity

Perplexity=9



Perplexity=30

Perplexity=90

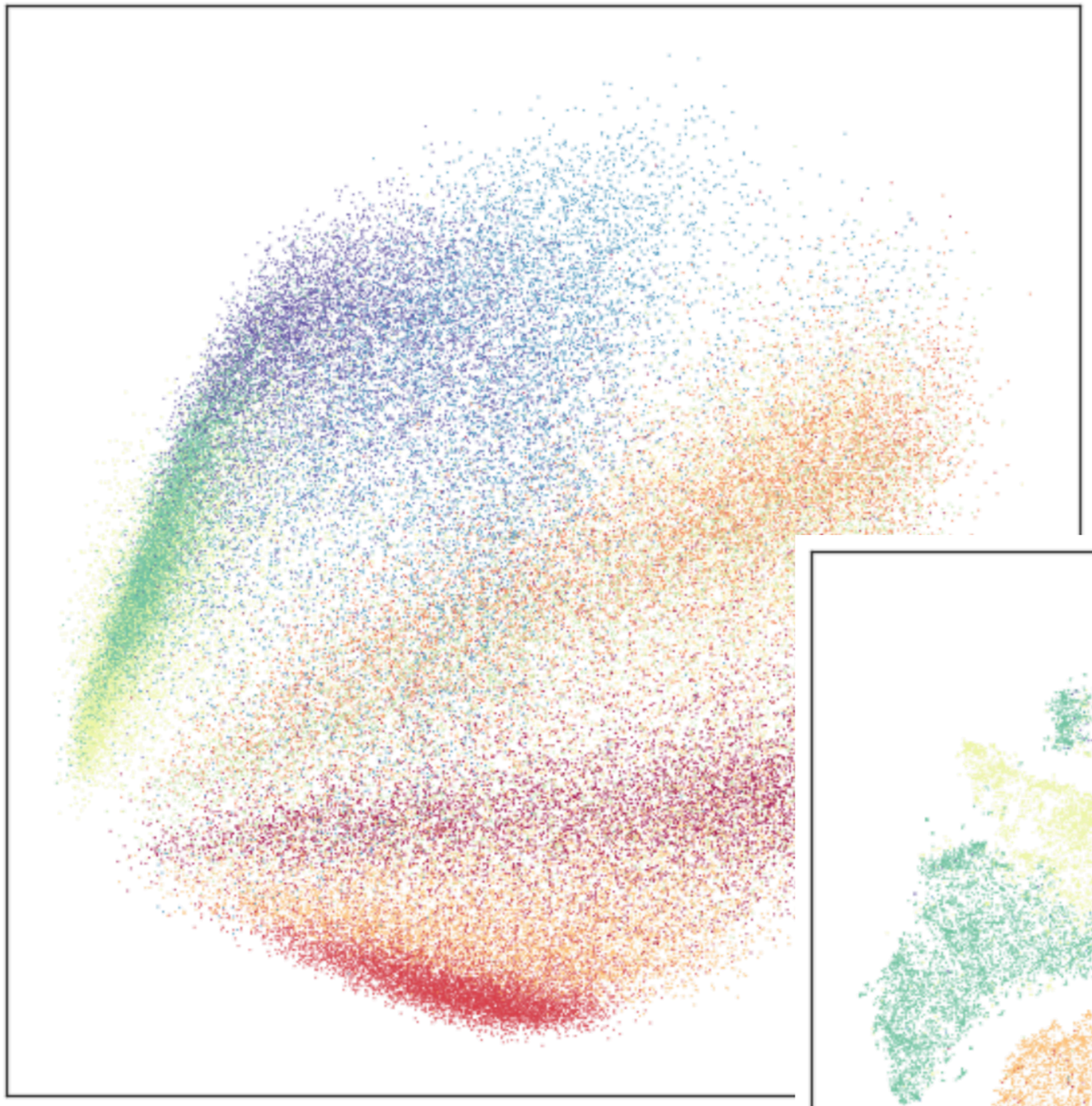
Perplexity=300

examples

fashion mnist

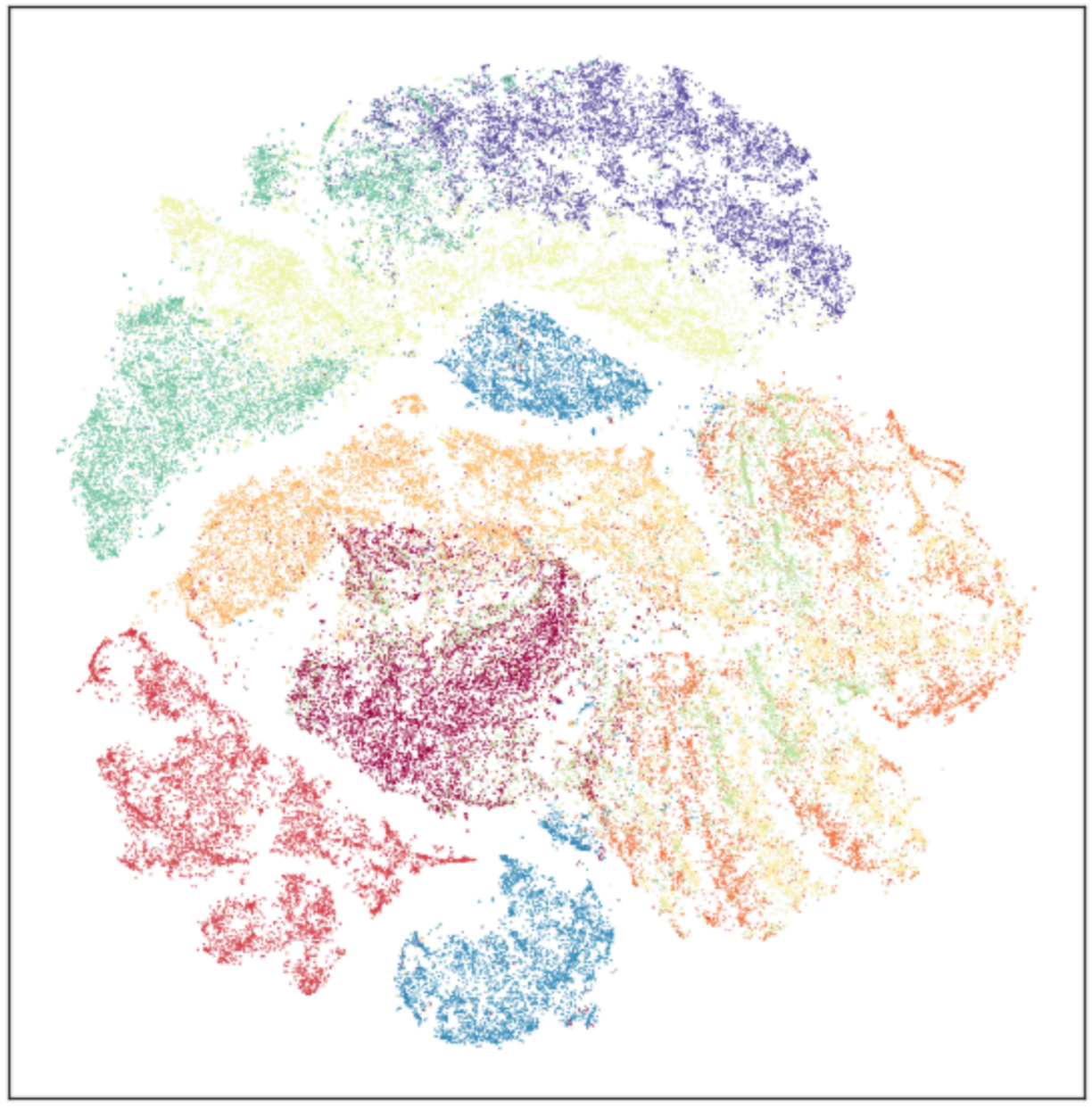
examples

fashion mnist



pca

tsne

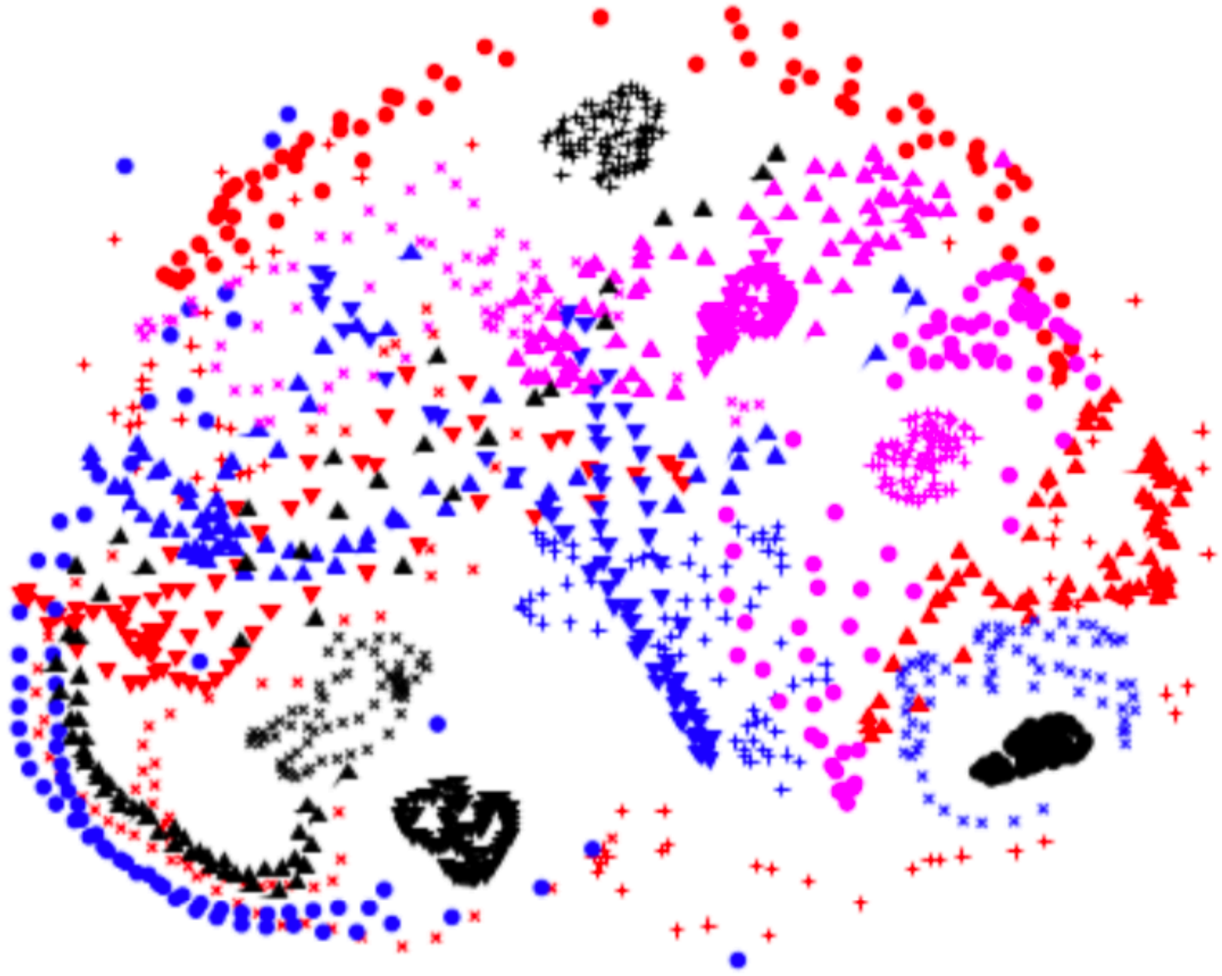


examples

coil20

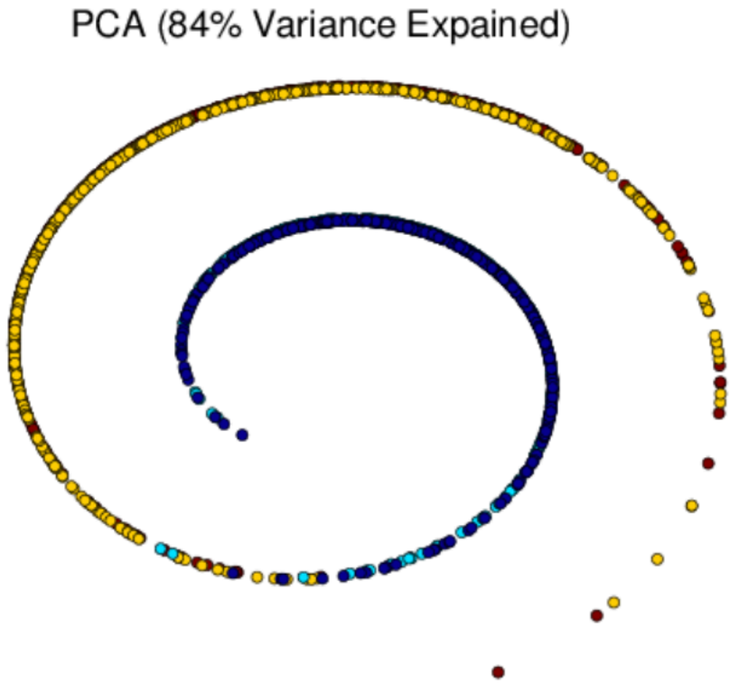
tsne

sammon

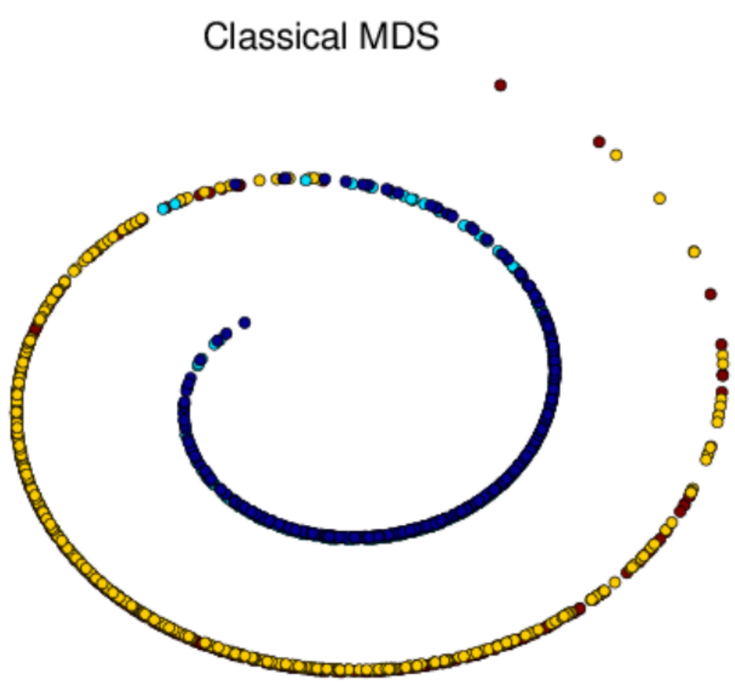


examples

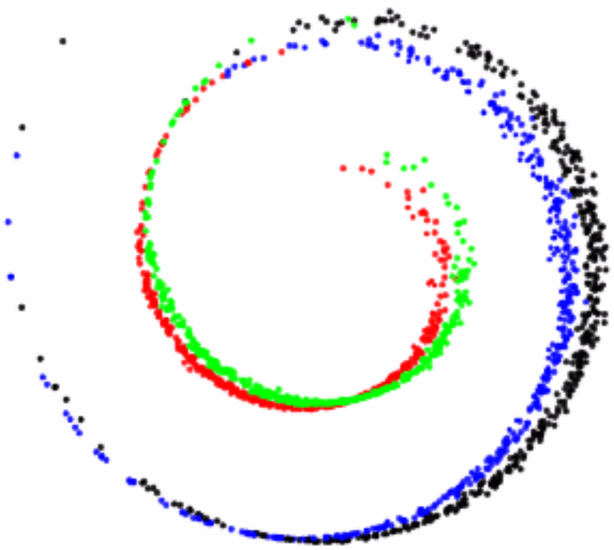
(double) swiss roll

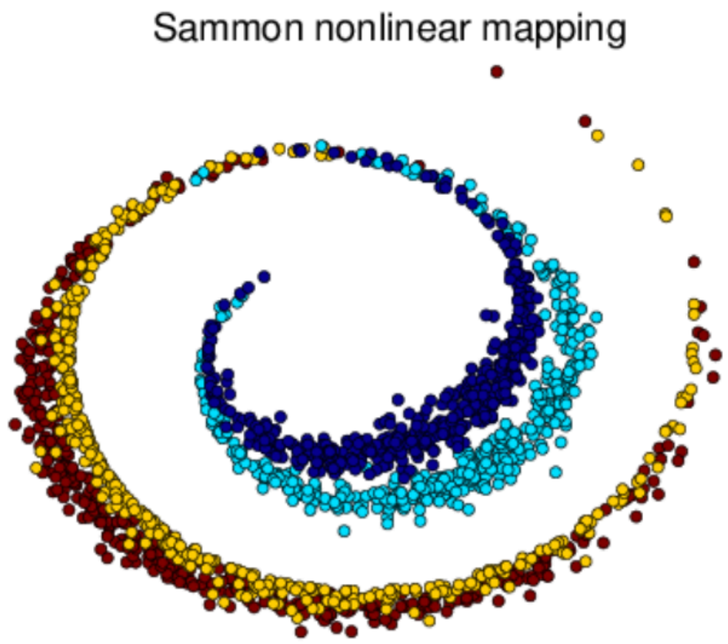


(a) PCA (PC1 versus PC2)

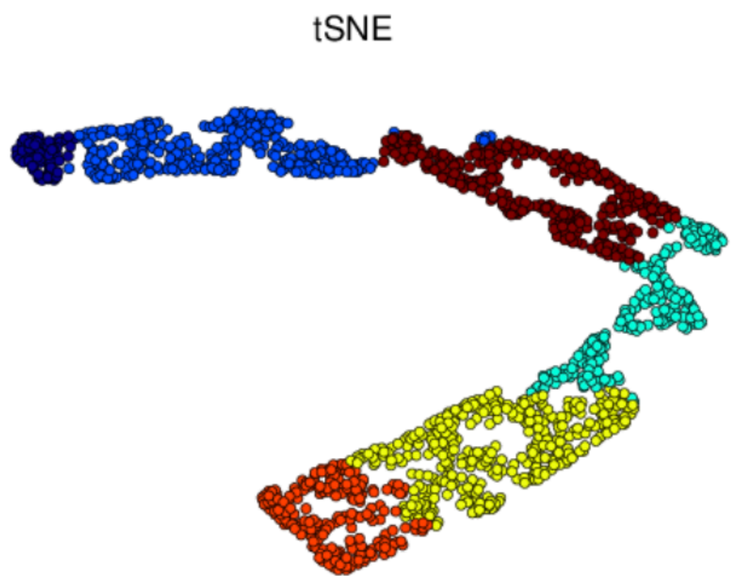


(b) Classical MDS





(c) Sammon



(d) tSNE

t-sne - summary

Non linear method

Preserve local structure - Global structure not preserved

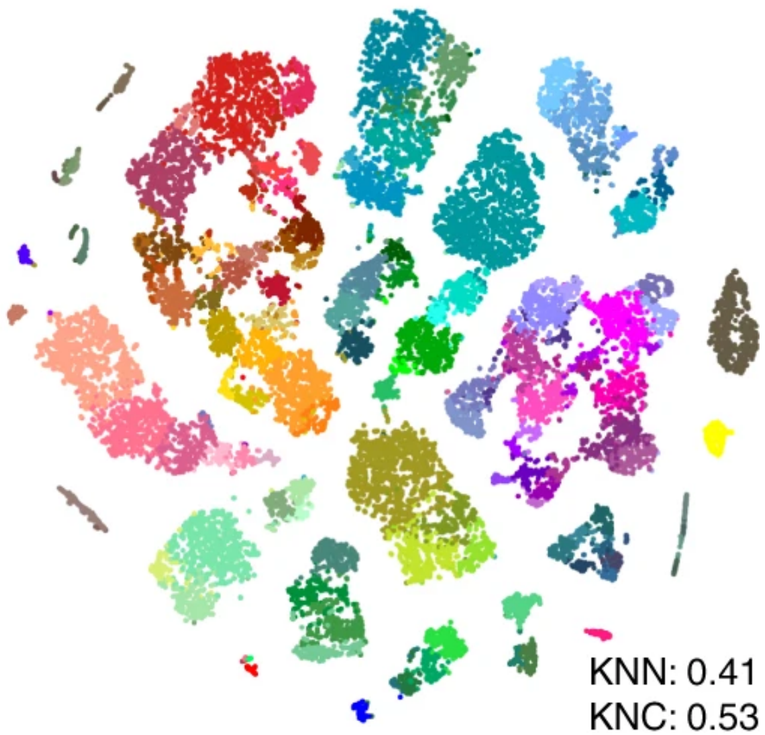
cluster size and inter-cluster distance not meaningful

hyperparameters choice is critical

stochastic (use informative initialisation, if possible)

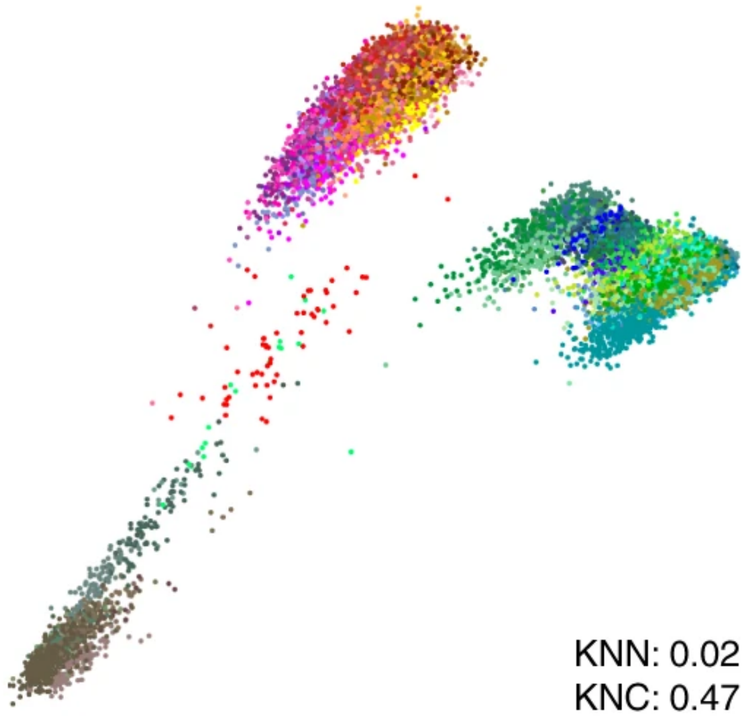
t-sne - initialisation

Default t-SNE
Random initialisation



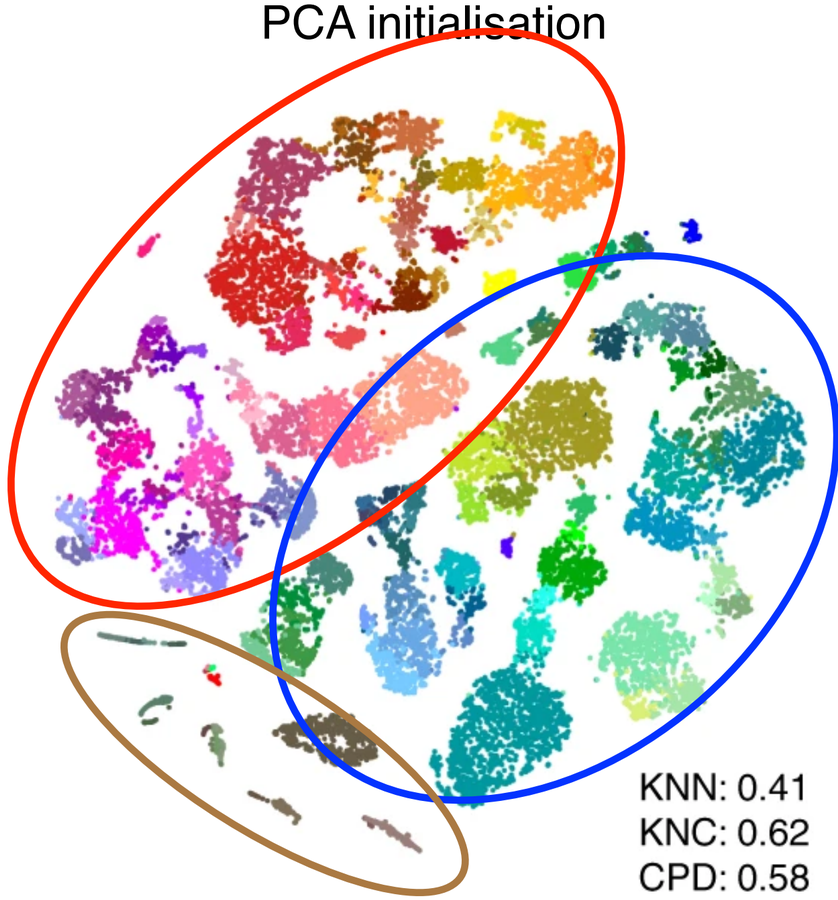
KNN: 0.41
KNC: 0.53
CPD: 0.24

PCA



KNN: 0.02
KNC: 0.47
CPD: 0.91

t-SNE
PCA initialisation



KNN: 0.41
KNC: 0.62
CPD: 0.58

Kobak et al, 2019

t-sne - summary

Non linear method

Preserve local structure - Global structure not preserved

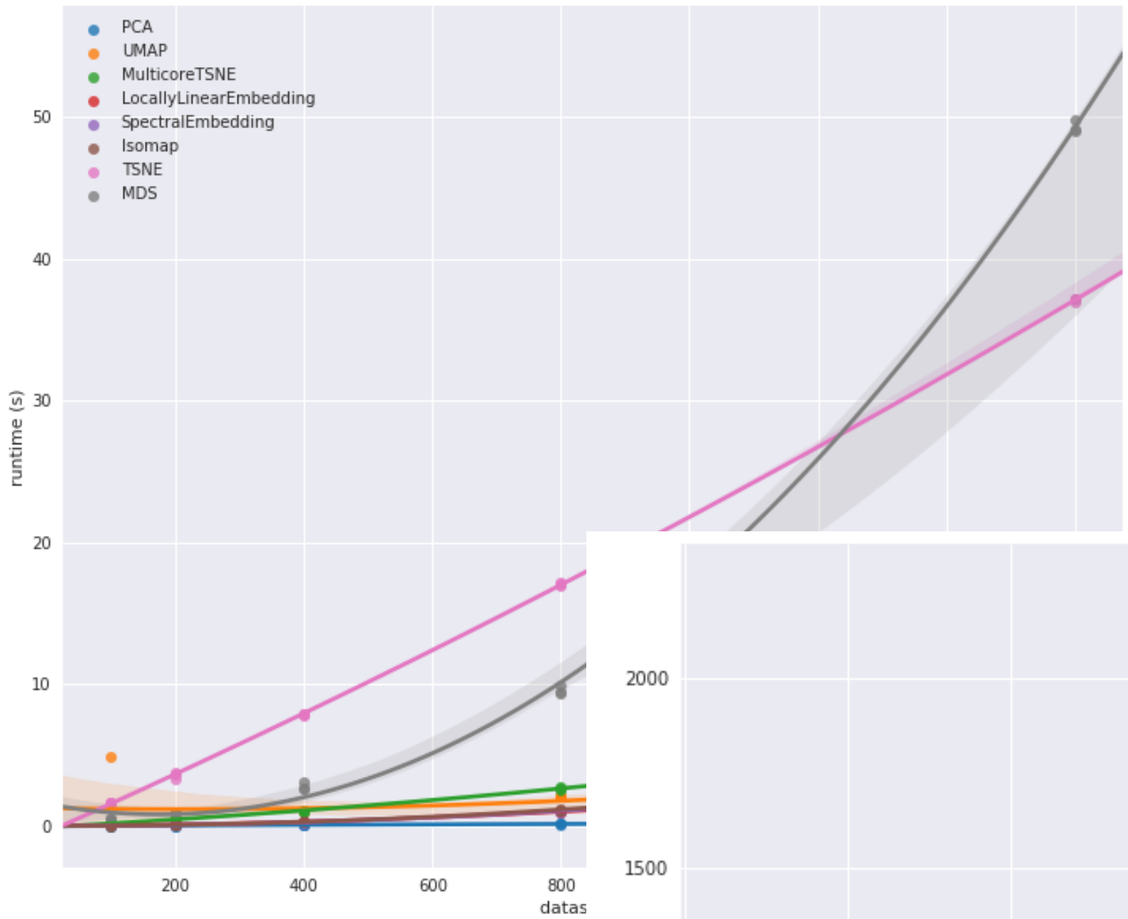
cluster size and inter-cluster distance not meaningful

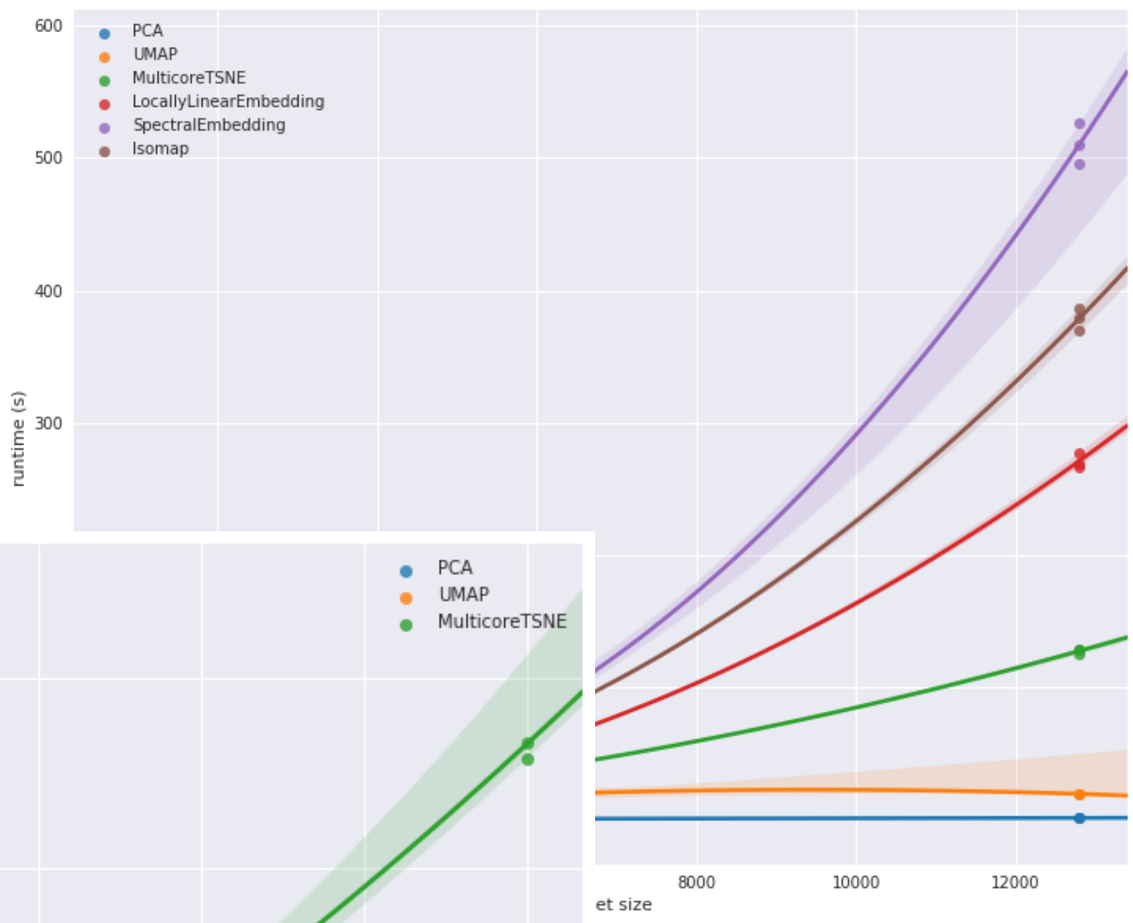
hyperparameters choice is critical

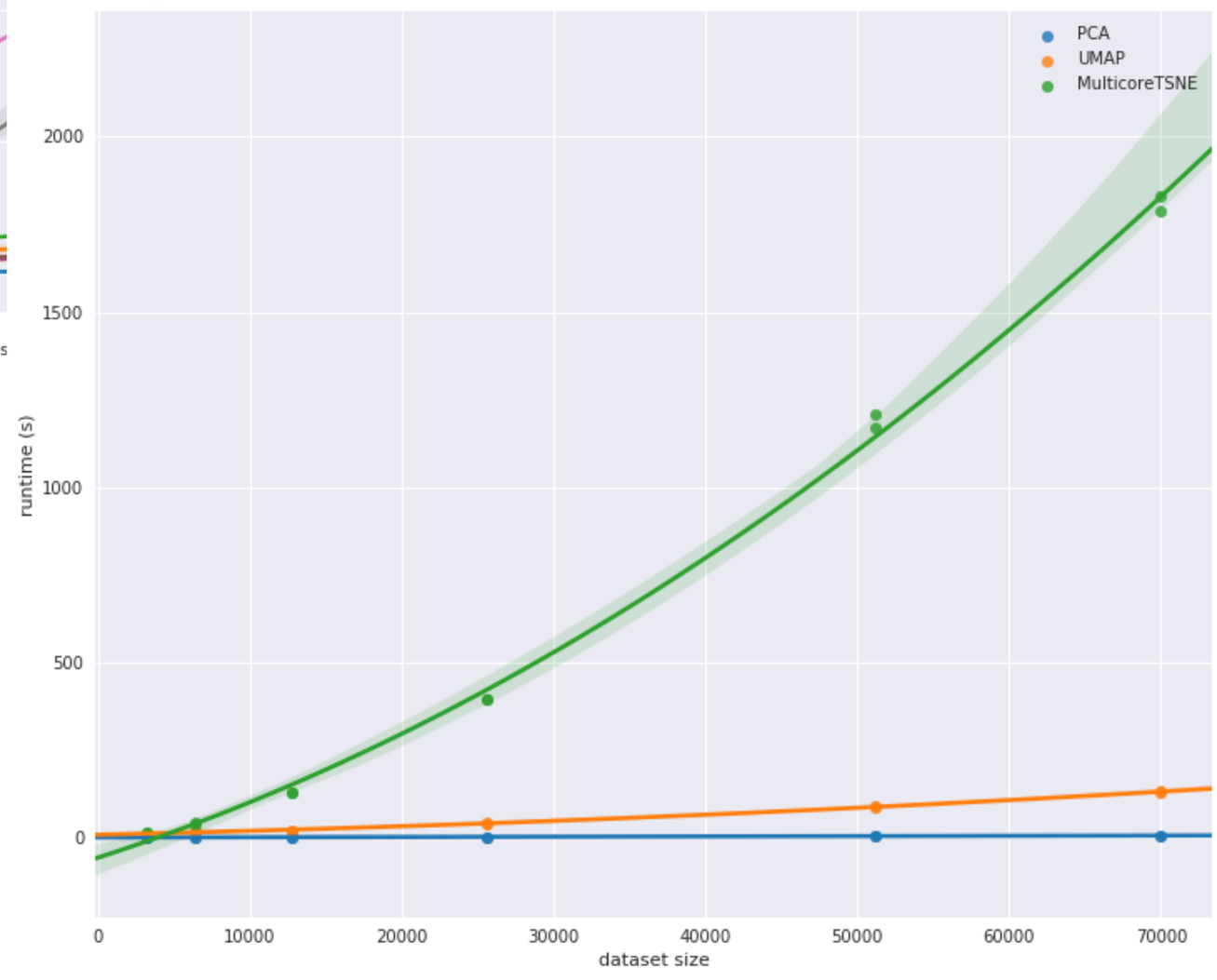
stochastic (use informative initialisation, if possible)

slow on large datasets

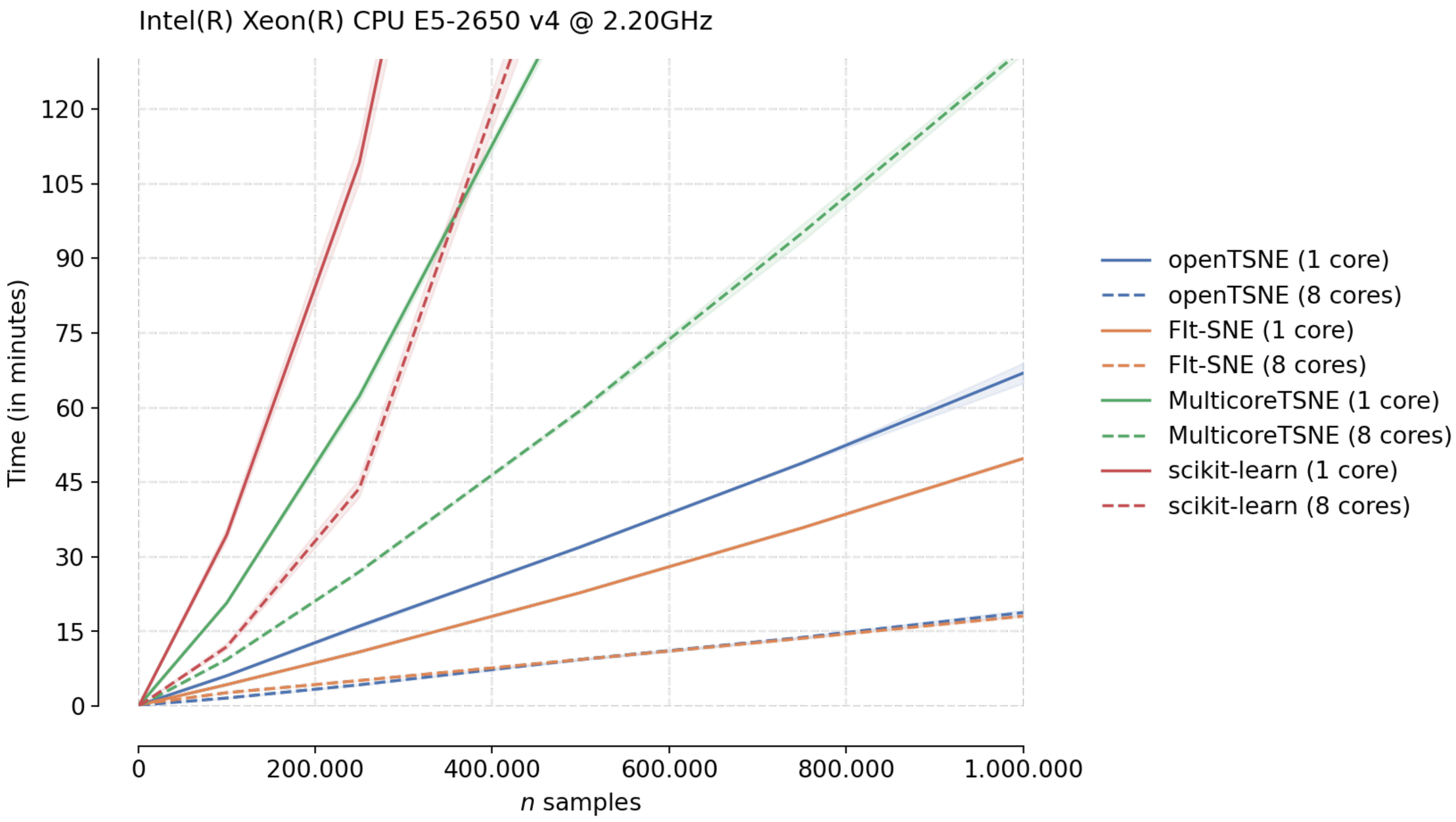
t-sne - running time







t-sne - running time



https://opentsne.readthedocs.io/en/stable/benchmarks.html

T-sne - summary

Non linear method

Preserve local structure - Global structure not preserved

cluster size and inter-cluster distance not meaningful

hyperparameters choice is critical

stochastic (use informative initialisation, if possible)

~~slow on large datasets~~