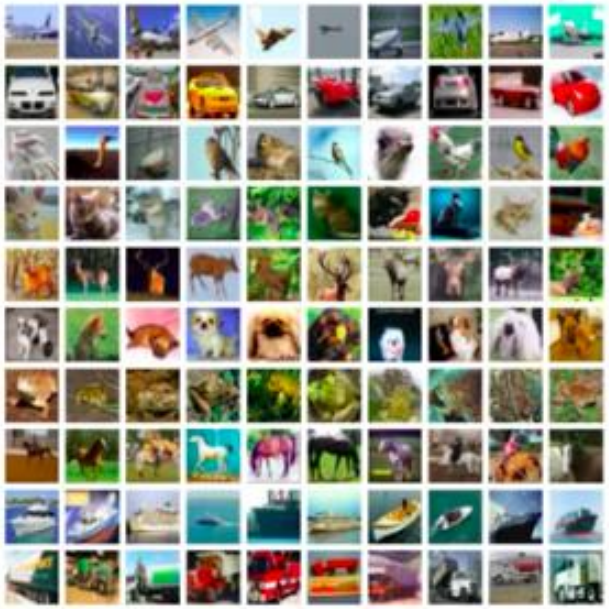


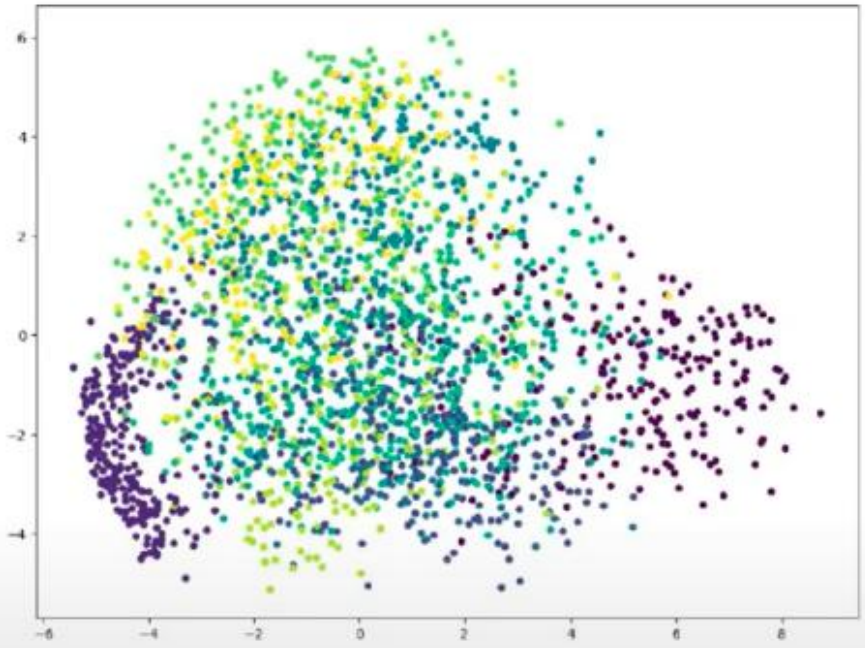
CIFAR-10

32*32 = 1024

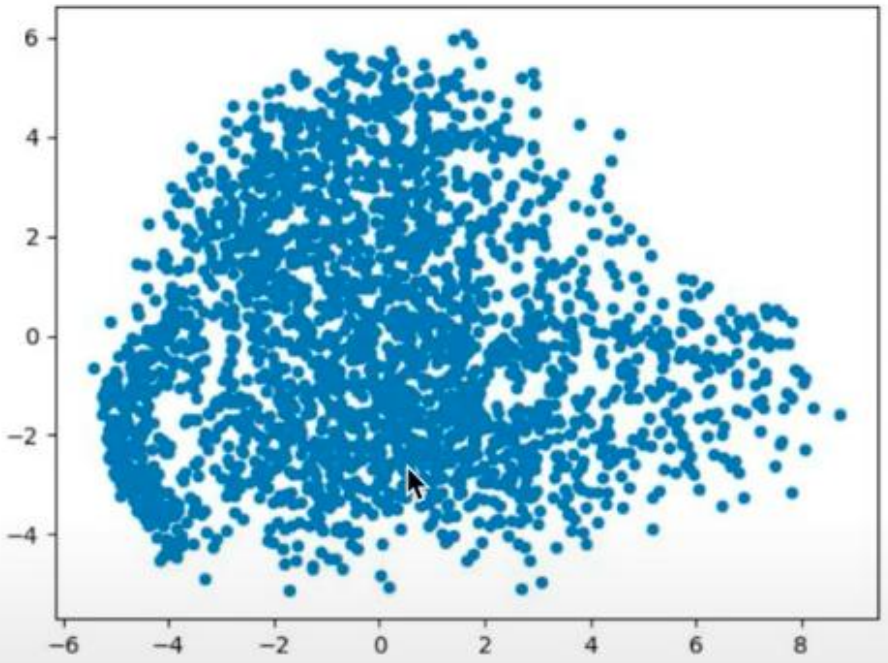
	FAM114A2	RNF20	NOL10	HAUS6	CPPED1	ACER3	ACER3
GSM115061	6.32987	7.96496	7.79692	4.40904	8.42627	7.18535	6.14987
GSM115062	6.70648	8.10416	7.62097	4.89053	8.79704	6.86707	5.60573
GSM115063	6.96342	8.20729	7.84963	5.01553	8.31353	7.48224	6.44482
GSM115064	6.70648	8.24981	7.69327	5.13721	8.64346	8.00825	6.96838
GSM115065	7.65958	8.92002	8.16916	5.96959	8.89742	8.1957	7.41619
GSM115066	7.57973	8.81423	8.09997	6.15961	8.81846	8.18607	7.43742
GSM115067	6.61016	8.02588	7.3402	4.71896	8.22403	8.06229	7.02589
GSM115068	6.97368	8.83162	7.63945	5.17843	8.4475	8.23283	7.65563
GSM115069	6.74944	8.58843	7.50808	4.82275	8.49804	8.11841	7.3771
GSM115070	6.29895	8.16635	7.87113	4.32141	7.7719	6.66789	5.29641
GSM115071	6.14449	8.01832	7.62978	3.92688	7.5319	7.0351	5.48401
GSM115072	6.43854	8.09891	8.0752	4.33687	7.91538	6.78031	5.69808
GSM115073	6.70648	8.09316	7.65406	4.9222	8.97477	6.96544	6.05961
GSM115074	6.50252	8.3742	7.46079	5.15525	8.80994	7.16648	6.30934
GSM115075	6.7622	8.64804	7.68013	4.85529	8.70325	6.64742	5.60776

Gene expressions

60,000+

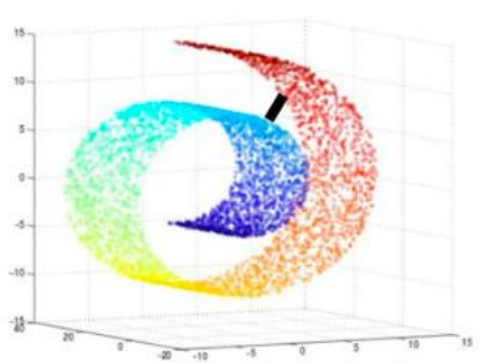


Visualization with labels

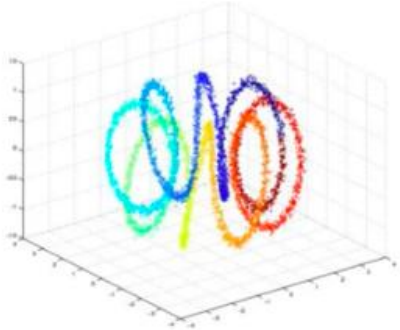


Visualization without labels

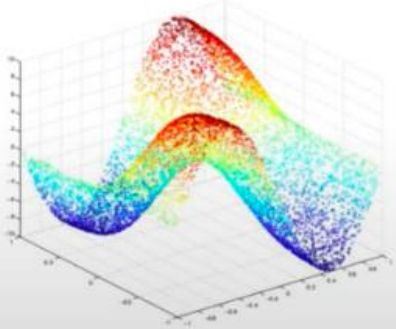
- PCA is good for dimension reduction
- But it is a linear algorithm
  - It cannot represent complex relationship between features



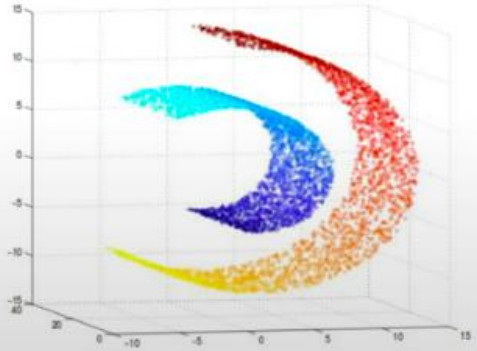
(a) Swiss roll dataset.



(b) Helix dataset.

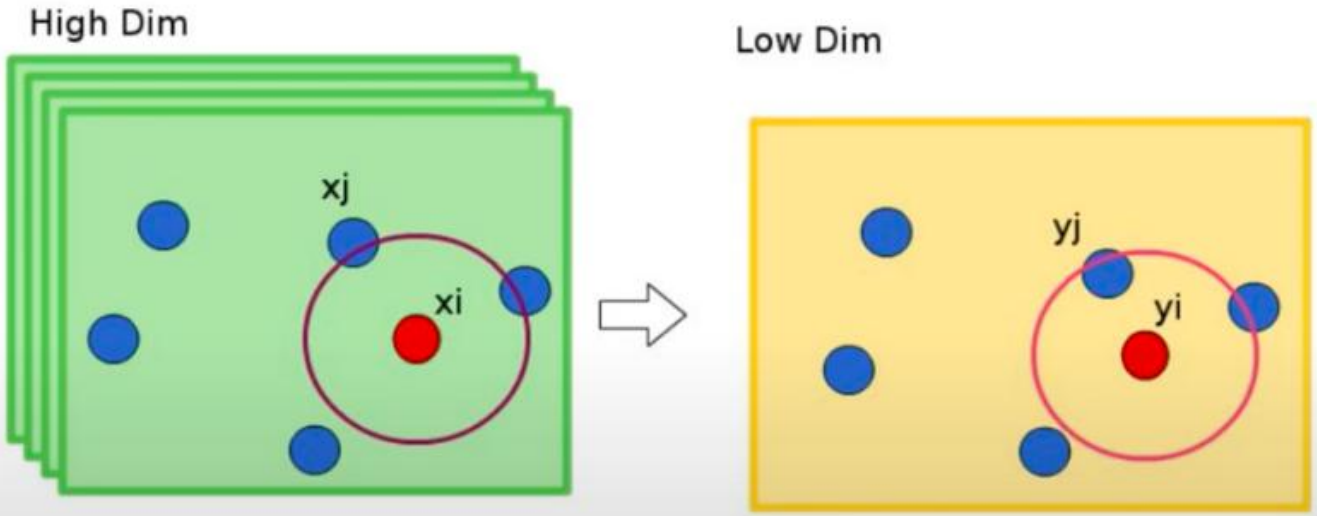


(c) Twinpeaks dataset.

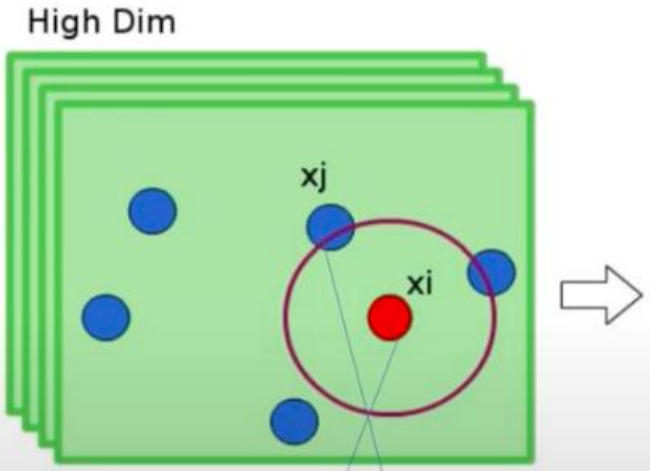


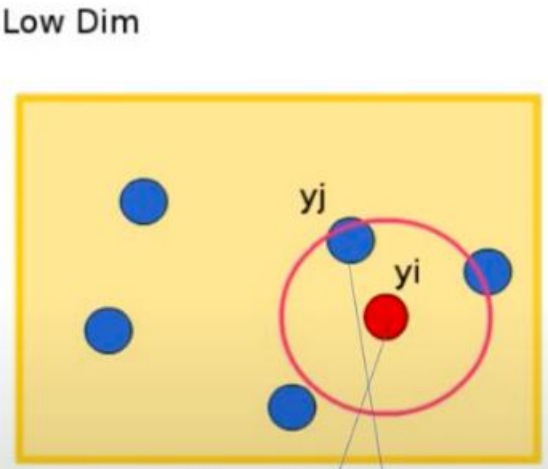
(d) Broken Swiss roll dataset.

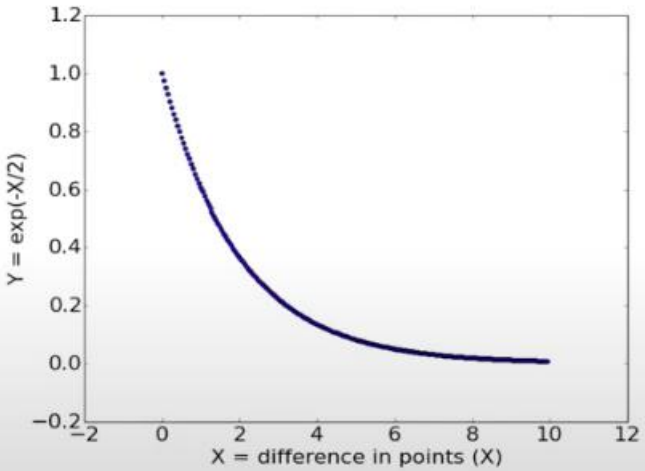
- A collection of N high-dimensional objects are given (x_1, x_2, ……,x_n)
- Build a map in which distances between points reflect similarities in the data



Converting the high-D Euclidean distance into conditional probabilities that represents similarities





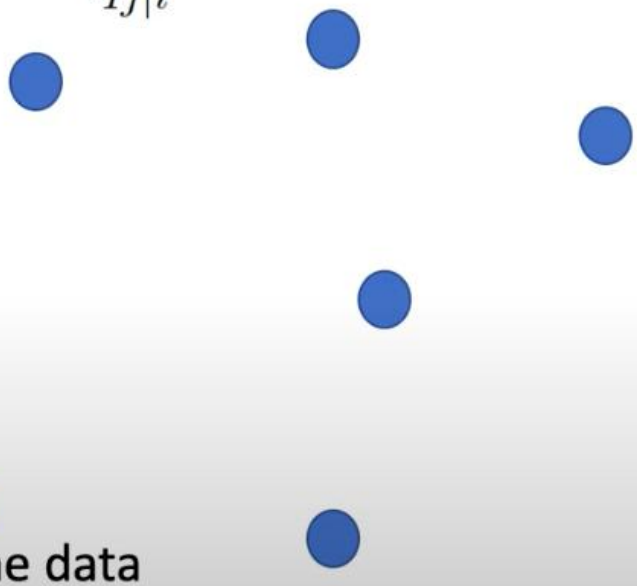


$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

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

Move points around to minimize the Kullback-Leibler divergence between ***P*** and ***Q***

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

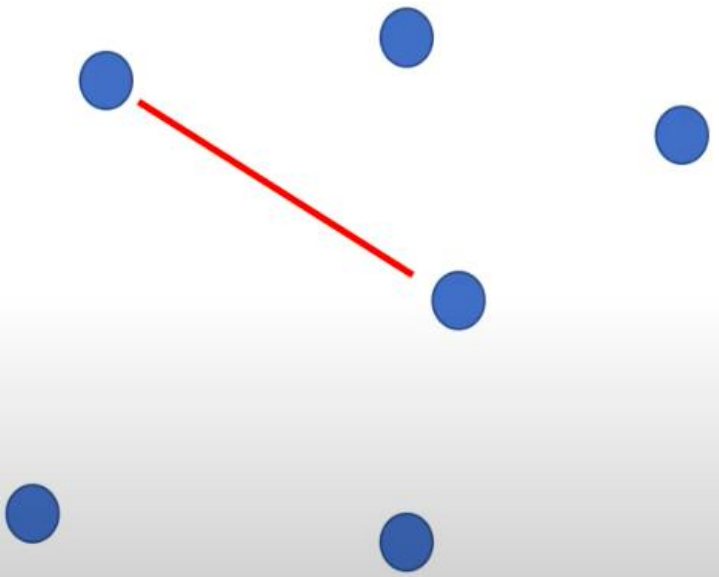


- if *p* = *q* then log(1) = 0
- Penalize when *p != q*
  - Large ***p*** modeled by small ***q***: Big penalty
  - Small ***p*** modeled by large ***q***: Small penalty
- Hence, SNE preserve the local similarity structure of the data

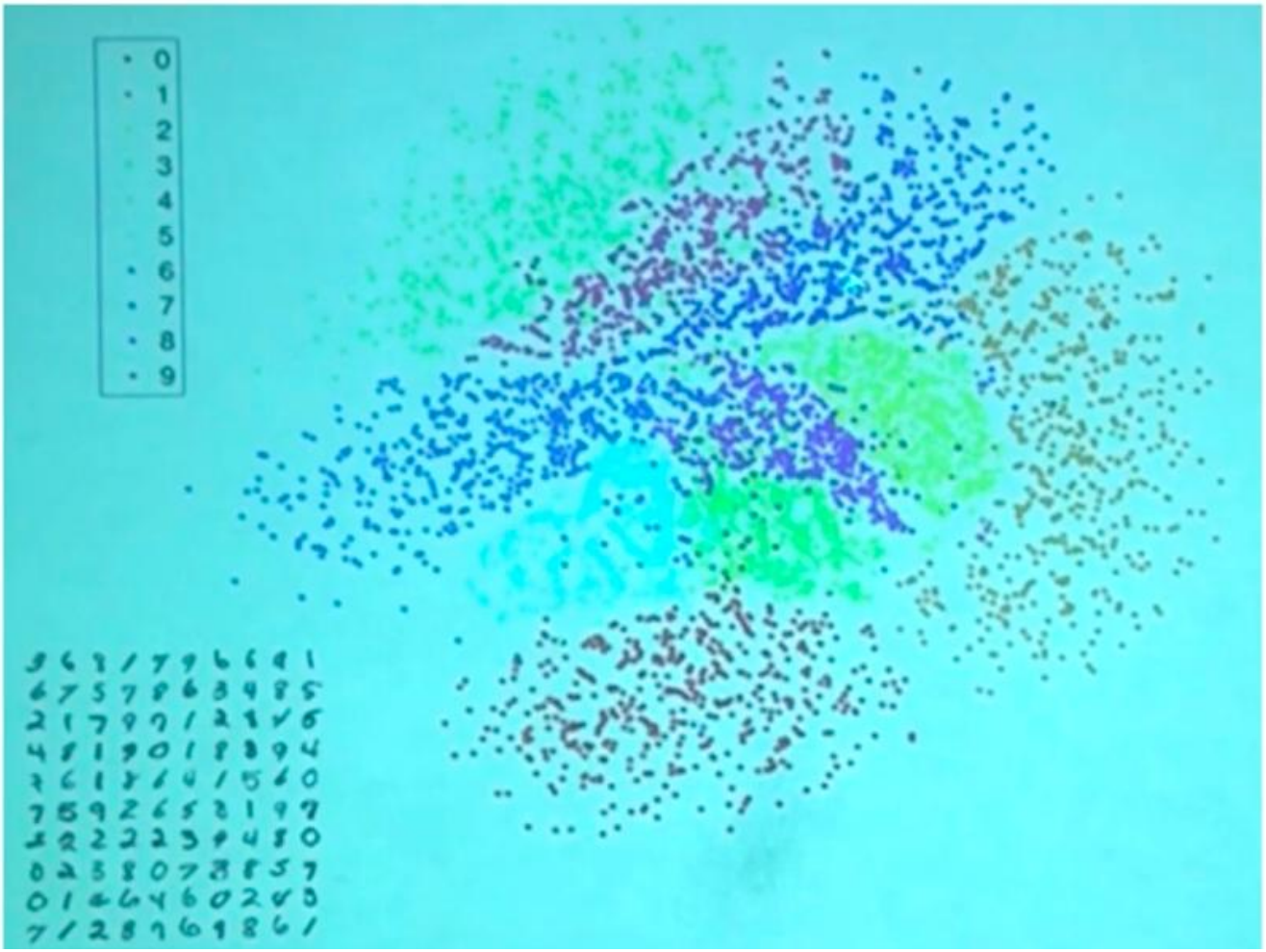
Move points around to minimize the Kullback-Leibler divergence between ***P*** and ***Q***

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

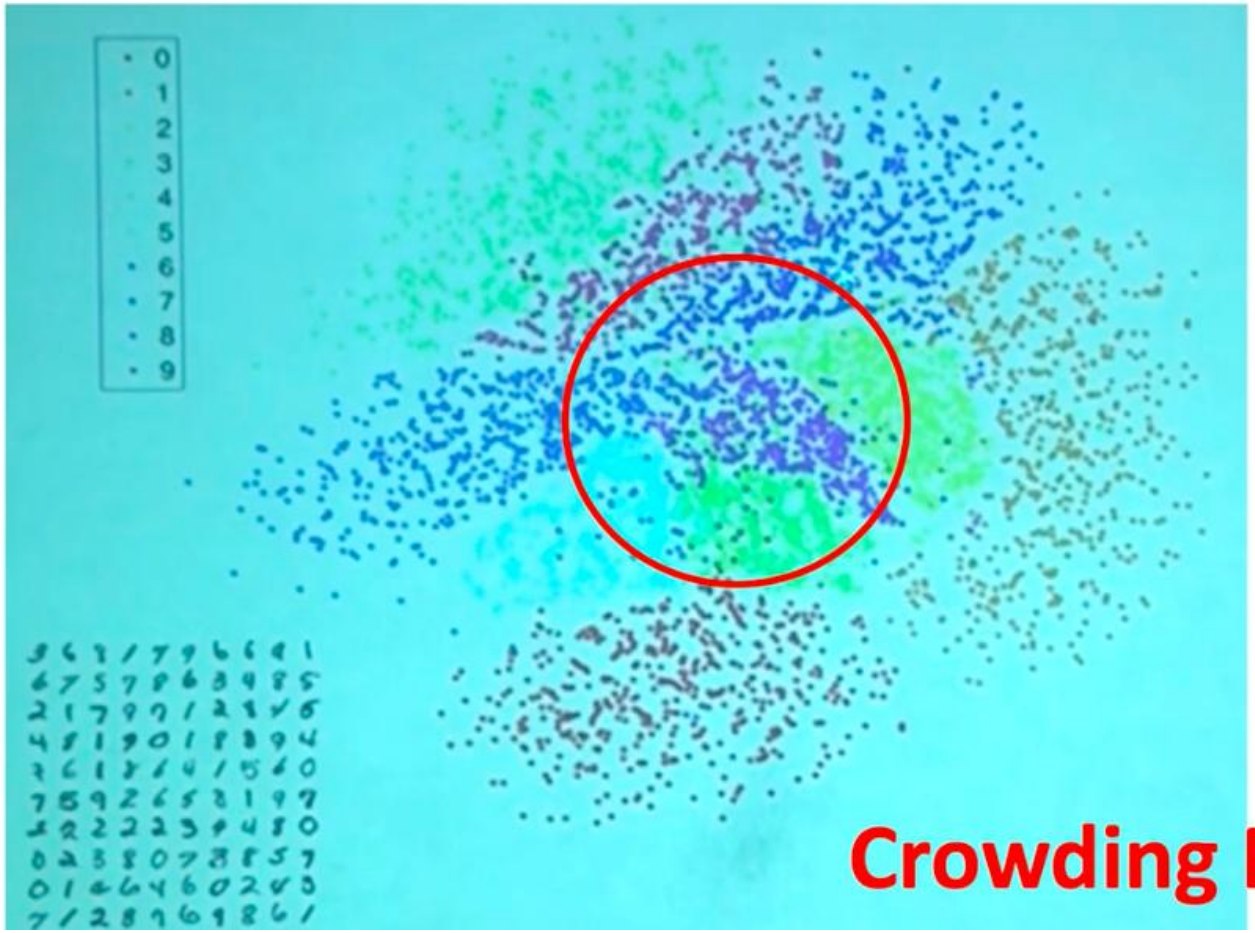
$$\frac{dC}{dy_i} = 2 \sum_j \underbrace{(y_i - y_j)}_{\text{Spring}} (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$$



Visualization of MNIST using SNE

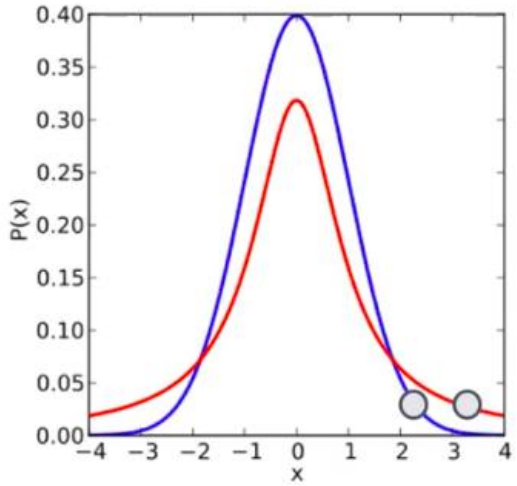


Visualization of MNIST using SNE

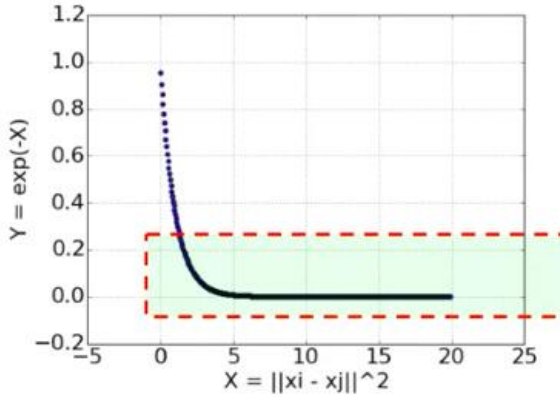


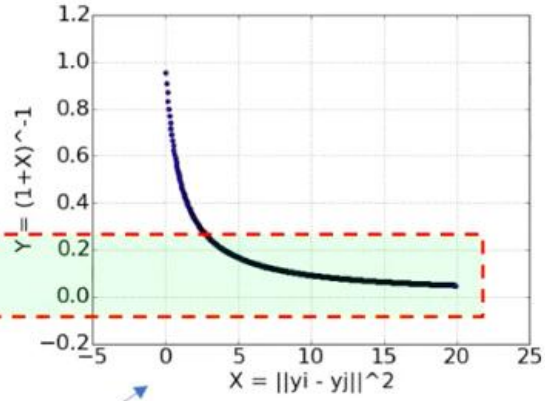
Crowding Problem

Student's t-distributed SNE



Blue = Gaussian
Red = Student's t





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

- As a result, dissimilar objects are allowed to be modeled far apart

Cost Function

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

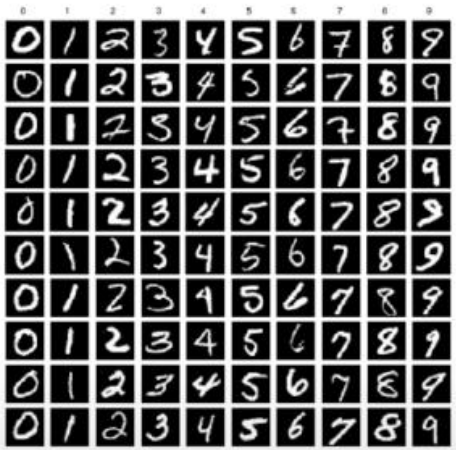
Gradient of *C* with respect to *yi*

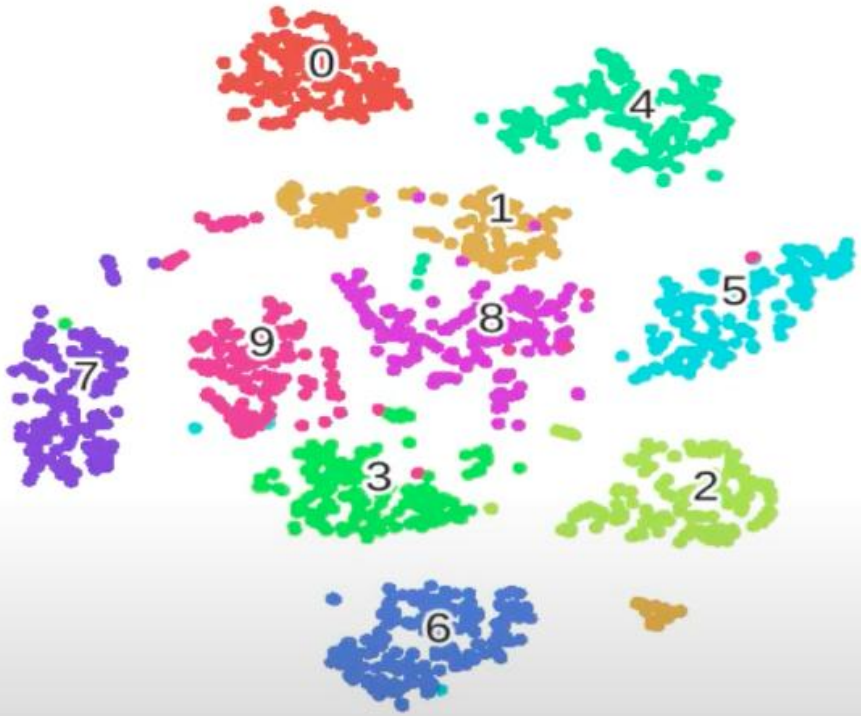
$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}$$

+ve/-ve

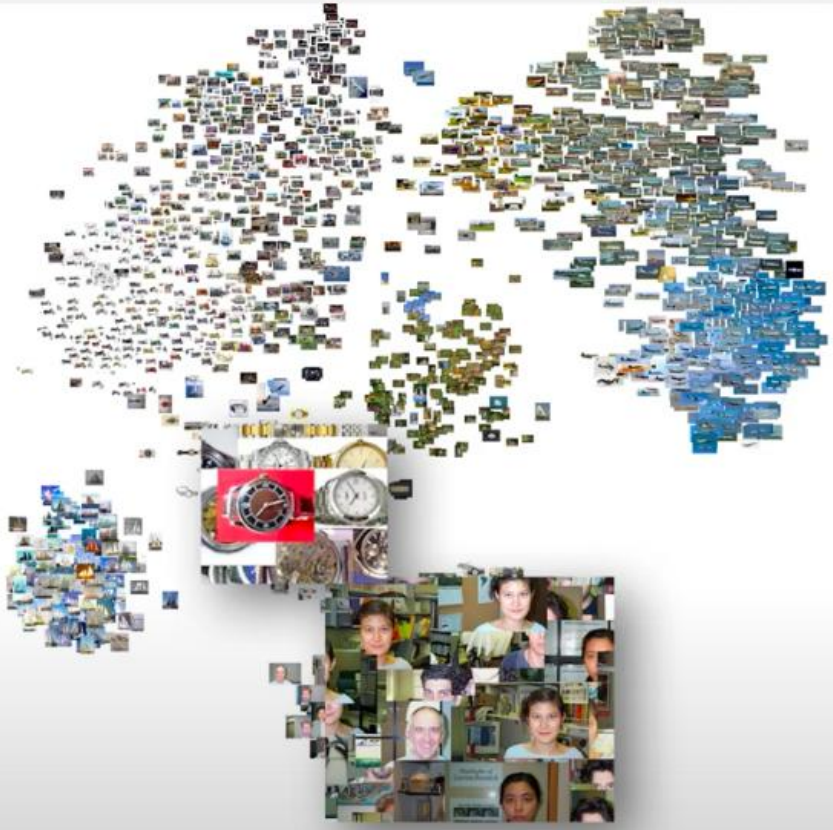
Always +ve (closer *yi,yj* get higher weight)

Visualization of MNIST Using t-SNE

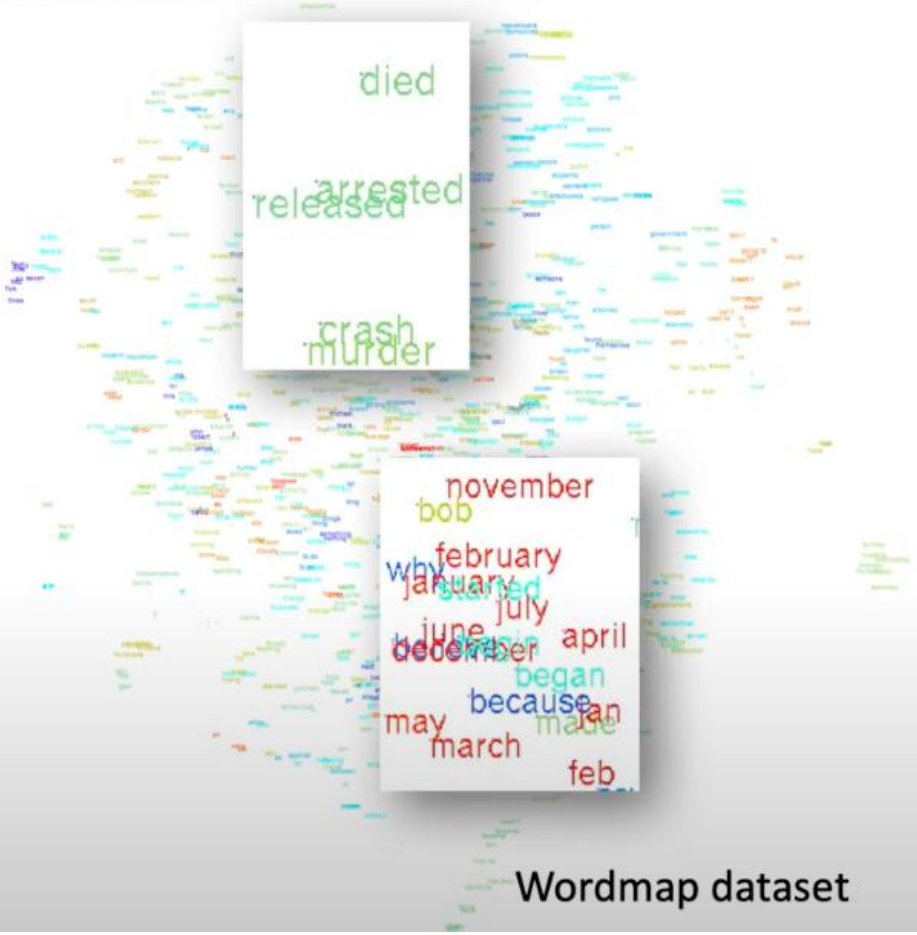




Example Datasets



CalTech-101 dataset



Wordmap dataset

Limitations!

- Not for dimensionality reduction, only for visualization!
- Less successful if data has very high intrinsic dimensionality
  - First reduce dimension using any non-linear model (e.g. autoencoder)
- Cost function is not convex
  - Carefully choose optimization parameters