

T-DISTRIBUTED STOCHASTIC NEIGHBORHOOD EMBEDDING (T-SNE)

12.04.2019

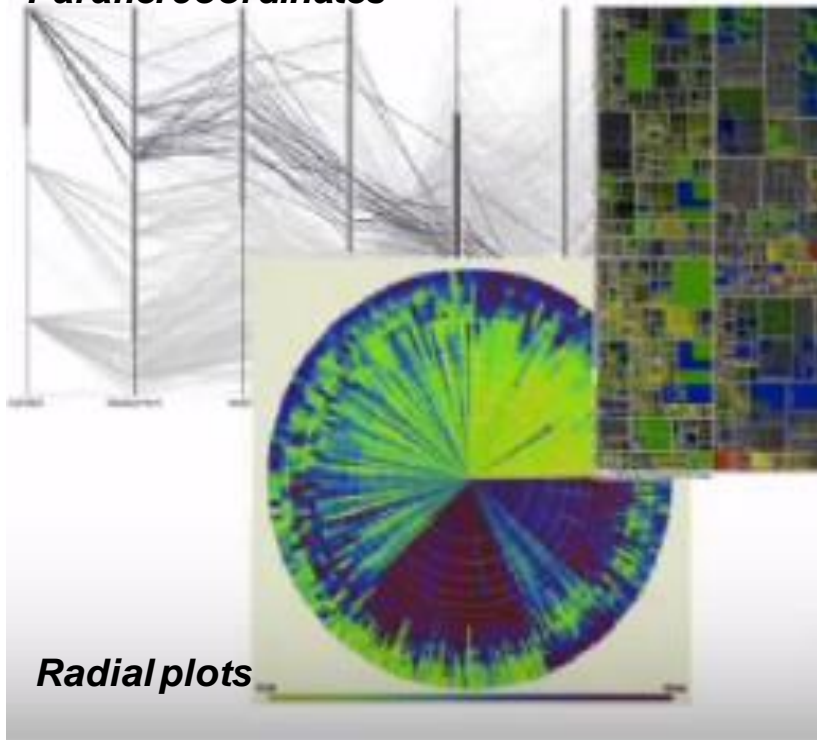
Ravi Kiran Sarvadevabhatla
CVIT, IIIT Hyderabad

VISUALIZATION

- Given a collection of N objects - $x_1, x_2 \dots x_N$
- How can we get a feel for how these N objects are arranged in data space (d -dimensional)?
- Why go through all this in the first place?

VISUALIZATION

Parallel coordinates



Treemaps

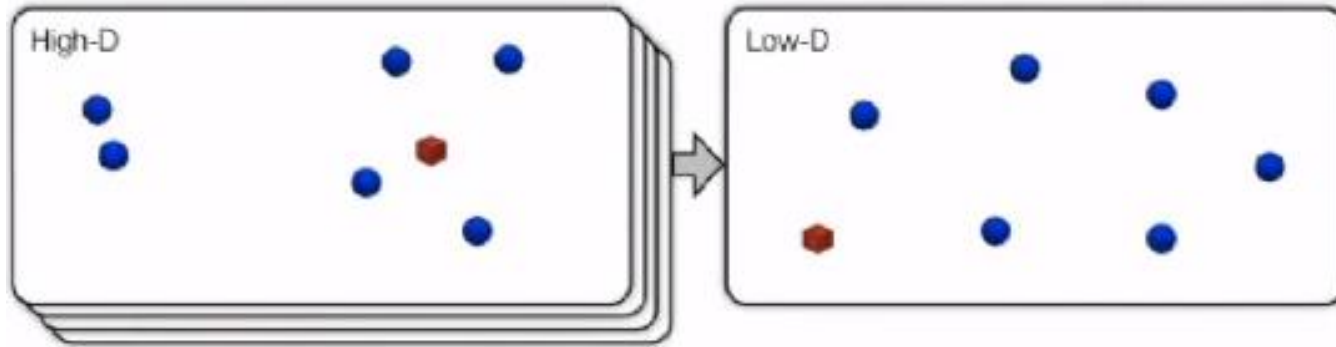


Wordles



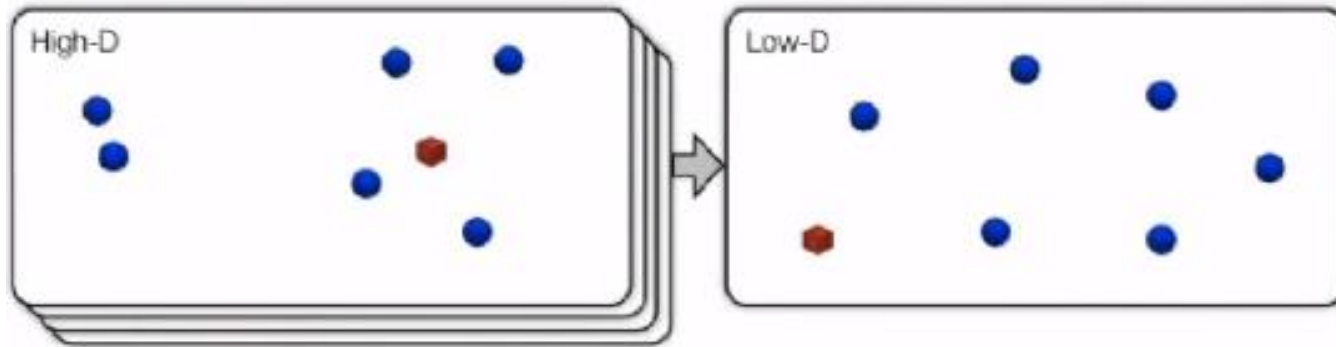
Radial plots

How can we visualize
high-dimensional data?



Build map in which distances between points reflect similarities in data

MATHEMATICALLY



Build map in which distances between points reflect similarities in data

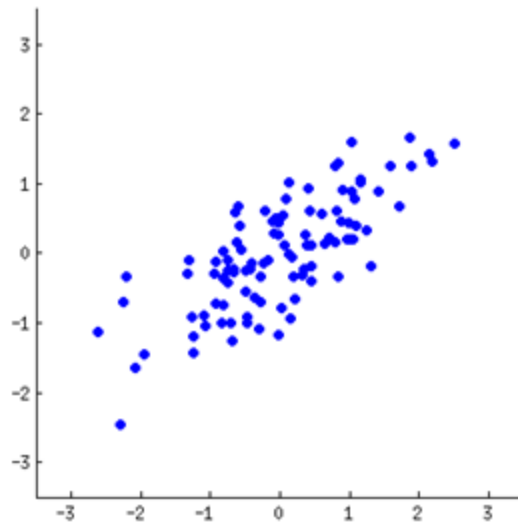
Formulation: Minimize some objective function which measures discrepancy between similarities in the data and similarities in the map.

PCA REVISITED ...

- What does PCA do ?

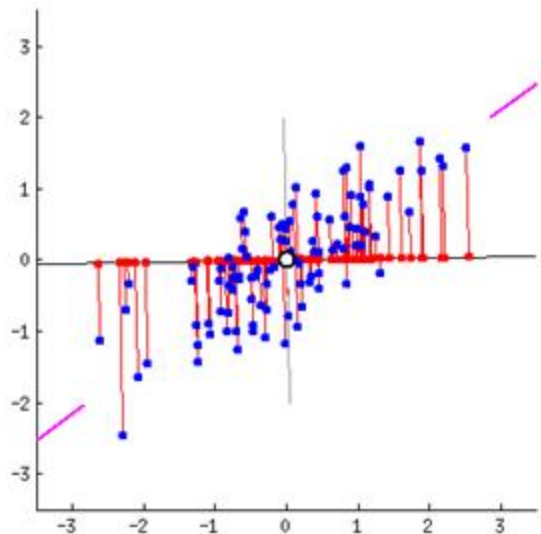
PCA REVISITED ...

- What does PCA do ?

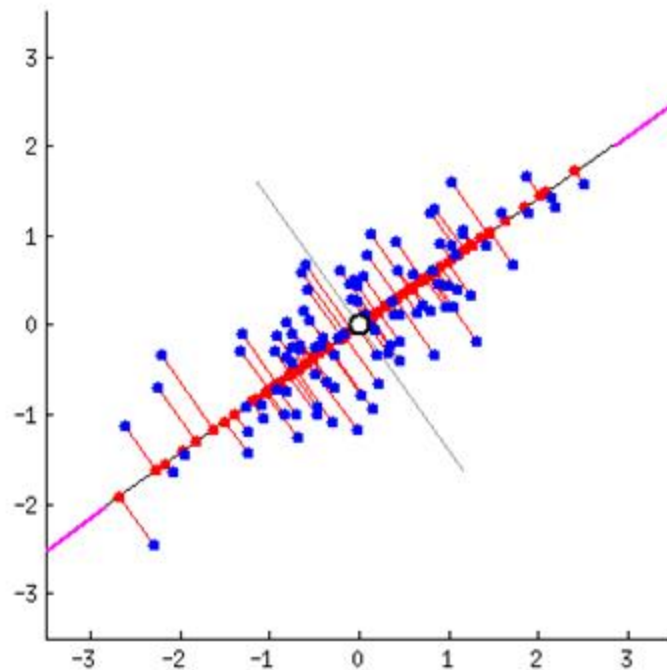


PCA REVISITED ...

- What does PCA do ?

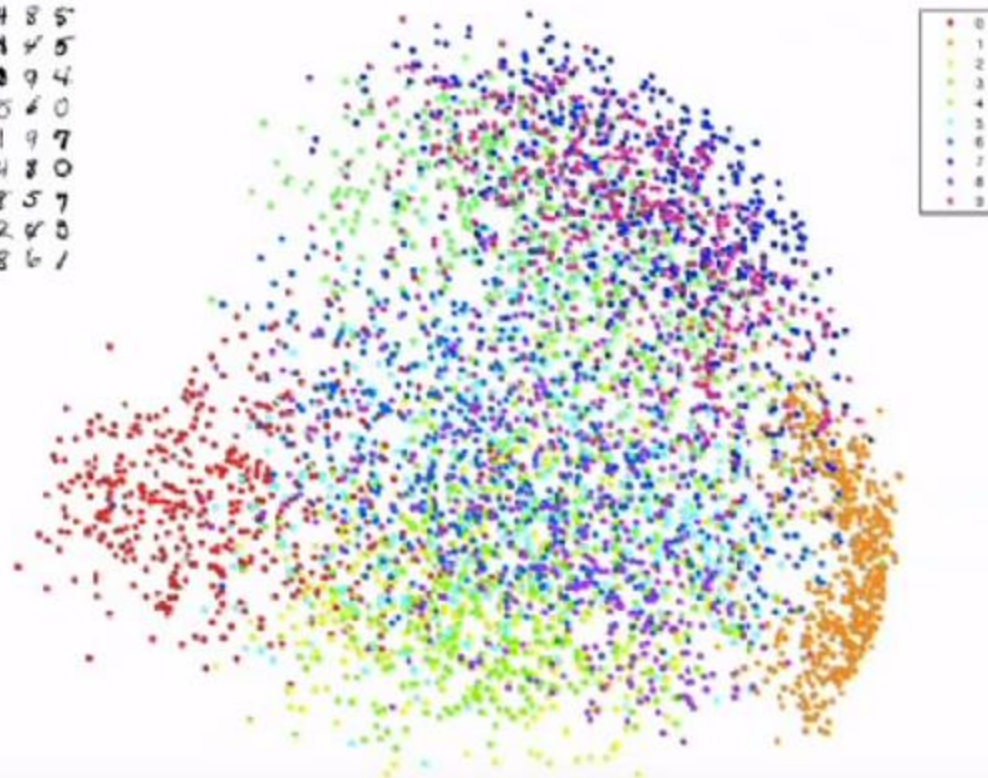


PCA REVISITED ...



PCA REVISITED ...

3 6 8 1 7 9 6 6 8 1
6 7 5 7 8 6 3 4 8 5
2 1 7 9 7 1 2 9 4 5
4 8 1 9 0 1 8 3 9 4
7 6 1 8 6 4 1 5 6 0
7 5 9 2 6 5 8 1 9 7
2 2 2 2 2 3 9 4 8 0
0 2 5 8 0 7 3 8 5 7
0 1 4 6 4 6 0 2 4 0
7 1 2 3 1 6 9 8 6 1



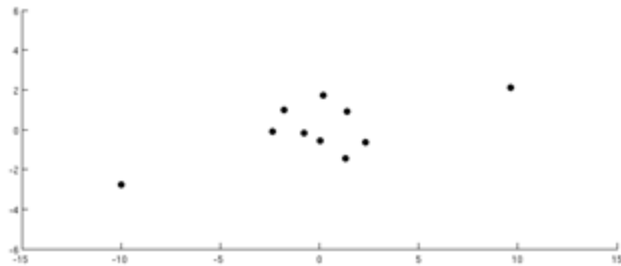
PCA REVISITED ...

3 6 3 1 7 9 6 4 1
6 7 5 7 8 6 3 4 8 5
2 1 7 9 7 1 2 9 4 5
4 8 1 9 0 1 8 3 9 4
3 6 1 8 6 4 1 5 6 0
7 5 9 2 6 5 8 1 9 7
2 2 2 2 2 3 9 4 8 0
0 2 5 8 0 7 3 8 5 7
0 1 4 6 4 6 0 2 4 8
7 1 2 3 1 6 9 8 6 1



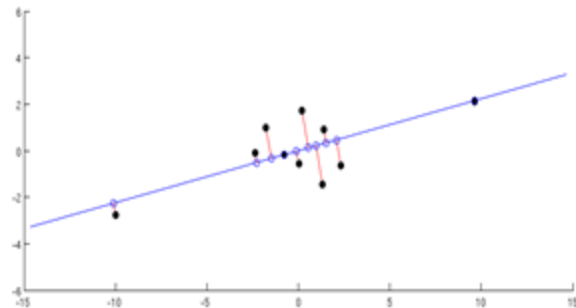
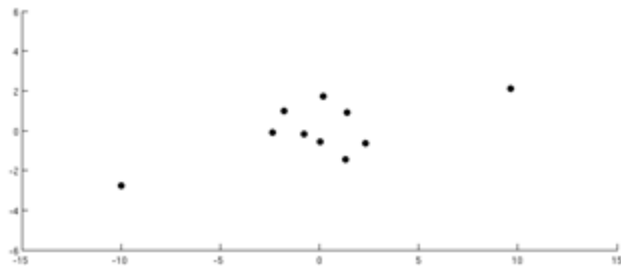
IS PCA MINIMIZING THE RIGHT OBJECTIVE FUNCTION ?

- PCA is mainly concerned with preserving **large** pairwise distances in the map



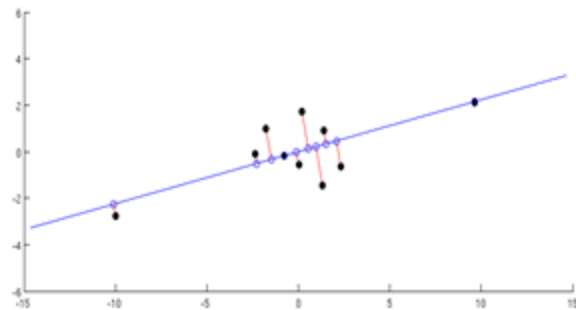
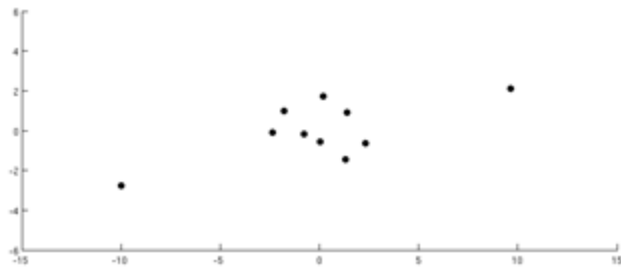
IS PCA MINIMIZING THE RIGHT OBJECTIVE FUNCTION ?

- PCA is mainly concerned with preserving **large** pairwise distances in the map



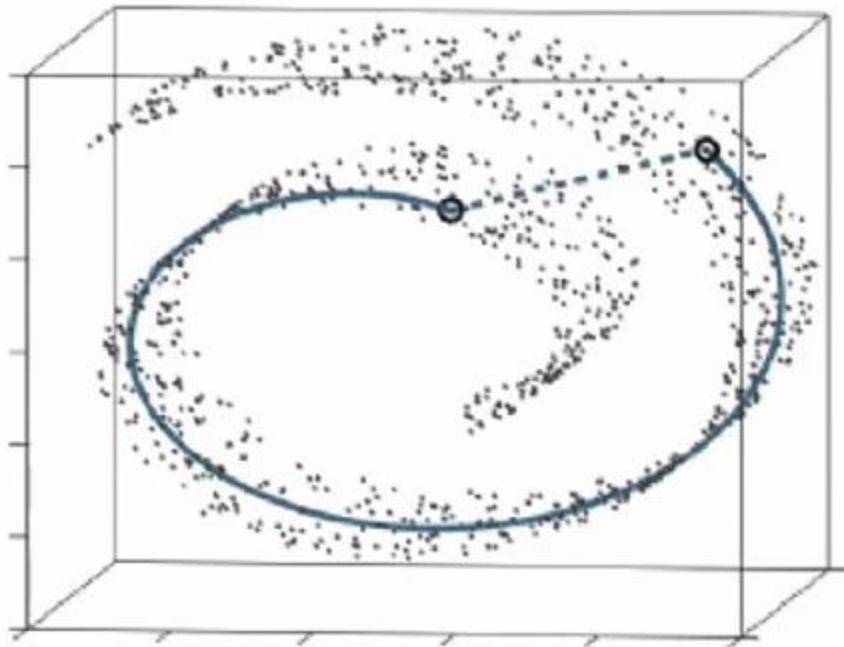
IS PCA MINIMIZING THE RIGHT OBJECTIVE FUNCTION ?

- PCA is mainly concerned with preserving **large** pairwise distances in the map



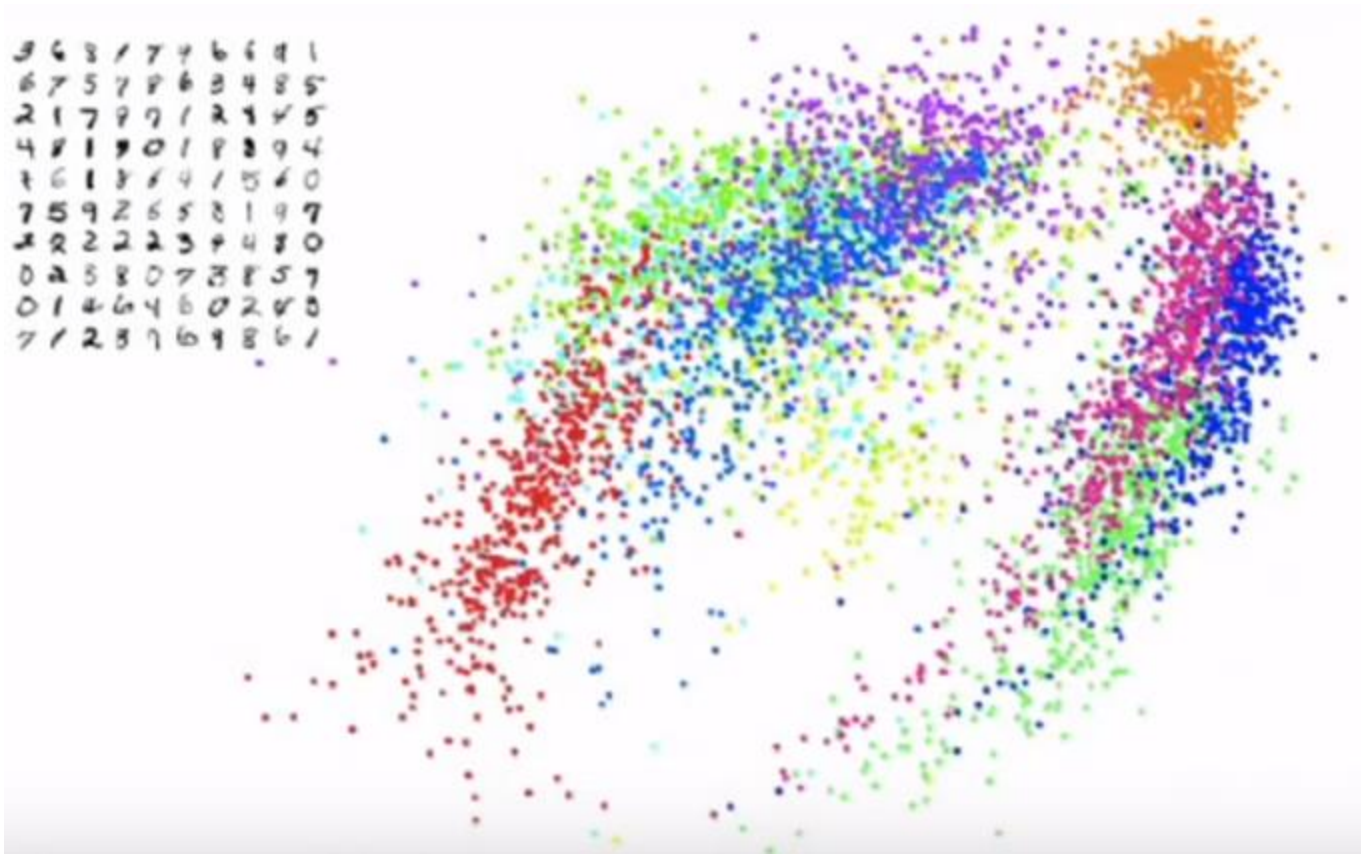
- “crowding” problem for smaller pairwise distances

IS PCA MINIMIZING THE RIGHT OBJECTIVE FUNCTION ?



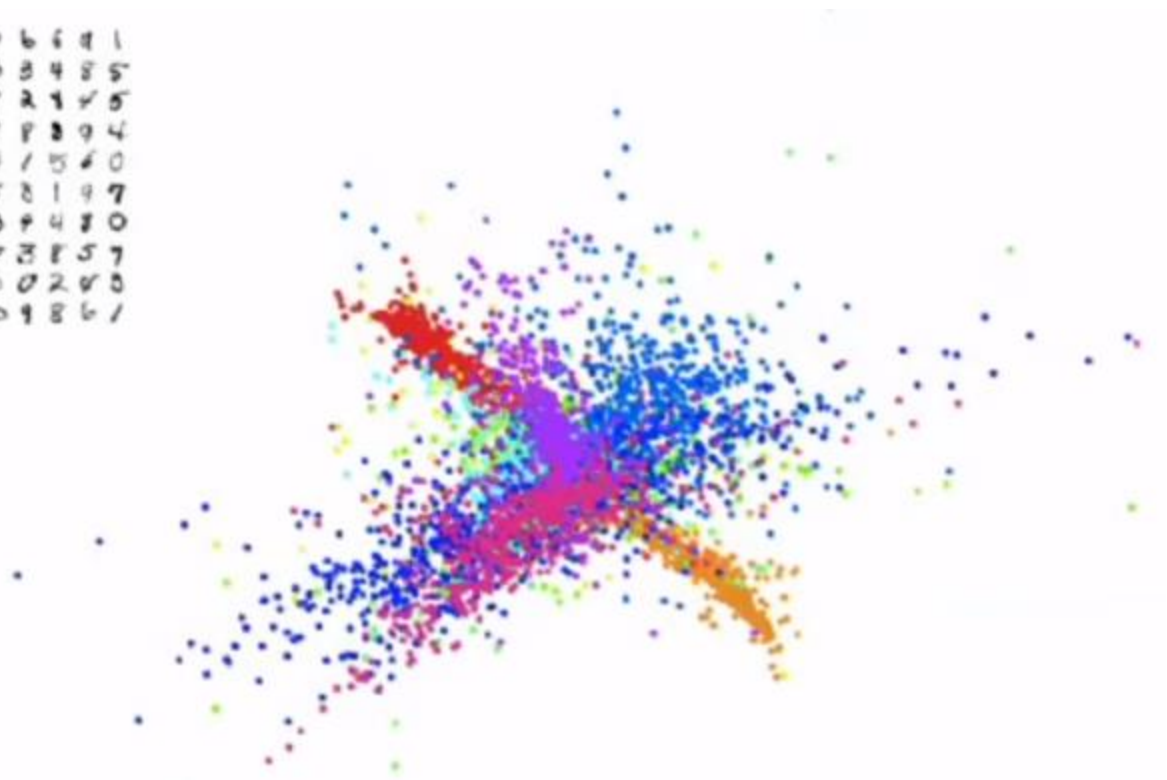
- Not so good at preserving local similarities
- Linearity is too restrictive an assumption !

ISOMAP



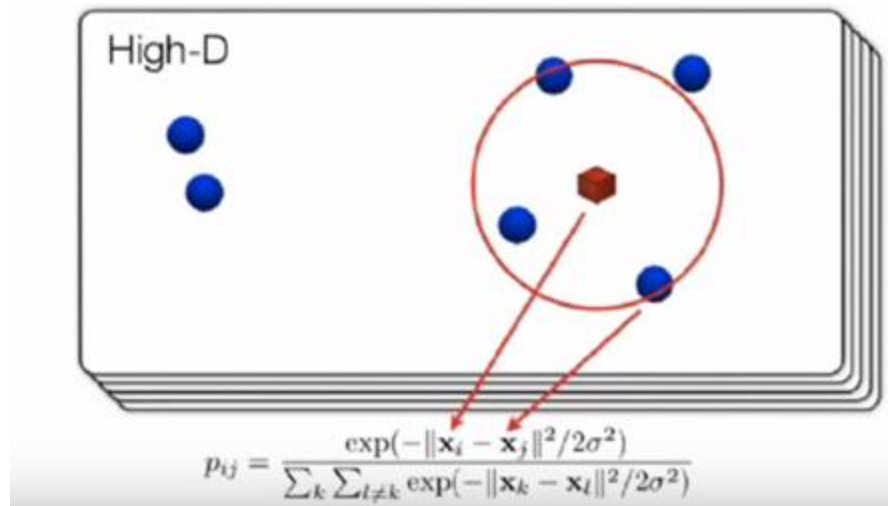
LOCALLY LINEAR EMBEDDING (LLE)

3 6 3 1 7 9 6 6 9 1
6 7 5 7 8 6 3 4 8 5
2 1 7 9 7 1 2 1 4 5
4 8 1 9 0 1 8 3 9 4
1 6 1 3 6 4 1 5 6 0
7 5 9 2 6 5 8 1 9 7
4 2 2 2 2 3 9 4 8 0
0 2 5 8 0 7 3 8 5 7
0 1 4 6 4 6 0 2 0 3
7 1 2 3 7 6 9 8 6 1



T-SNE : MODELING SIMILARITIES IN HIGH-D

Center a gaussian under each point i



Prob of picking point pair $(i,j) \propto p_{ij}$ [their similarity]

Nearby points \Rightarrow Large p_{ij} , Far points \Rightarrow Infinitesimal p_{ij}

T-SNE : MODELING SIMILARITIES IN HIGH-D

- In practice, we compute the input similarities slightly differently:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{j' \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_{j'}\|^2 / 2\sigma_i^2)}$$

*# of points
under each
Gaussian*

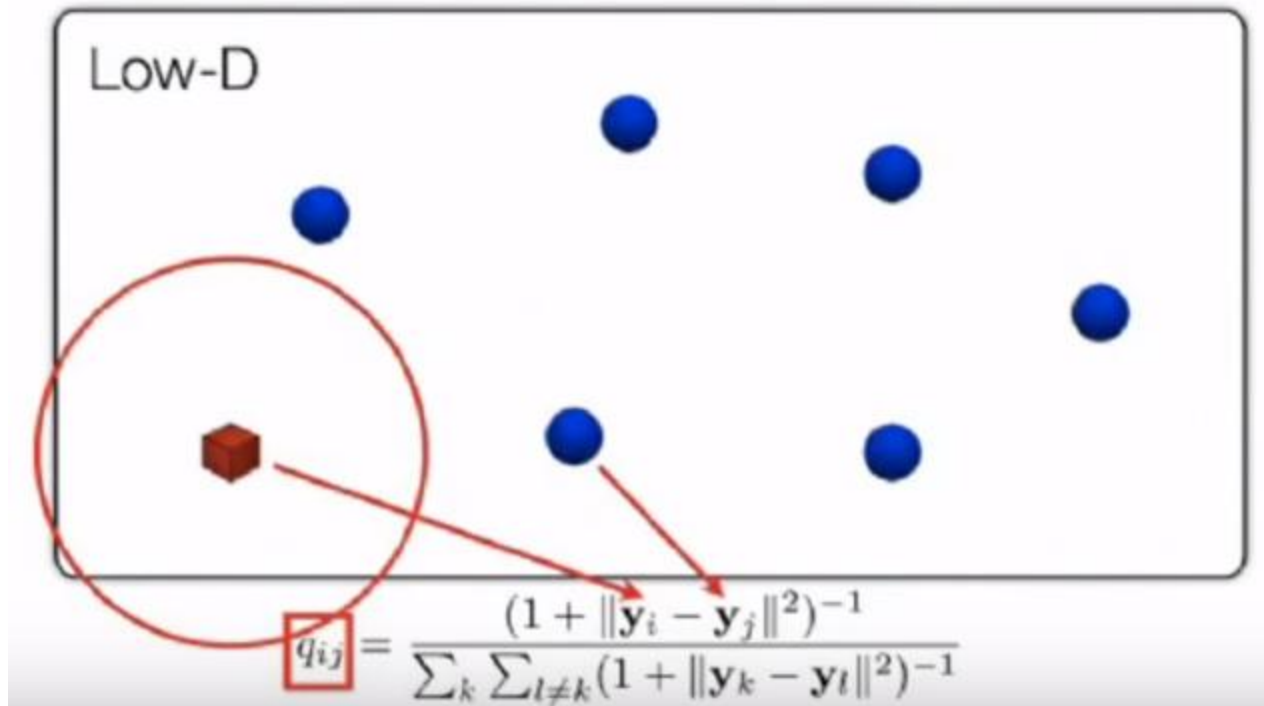
- We set the bandwidth σ_i such that the conditional has a fixed perplexity

allows for adapting to data distribution

- Finally, we symmetrize the conditionals: $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$

Provides robustness against outliers

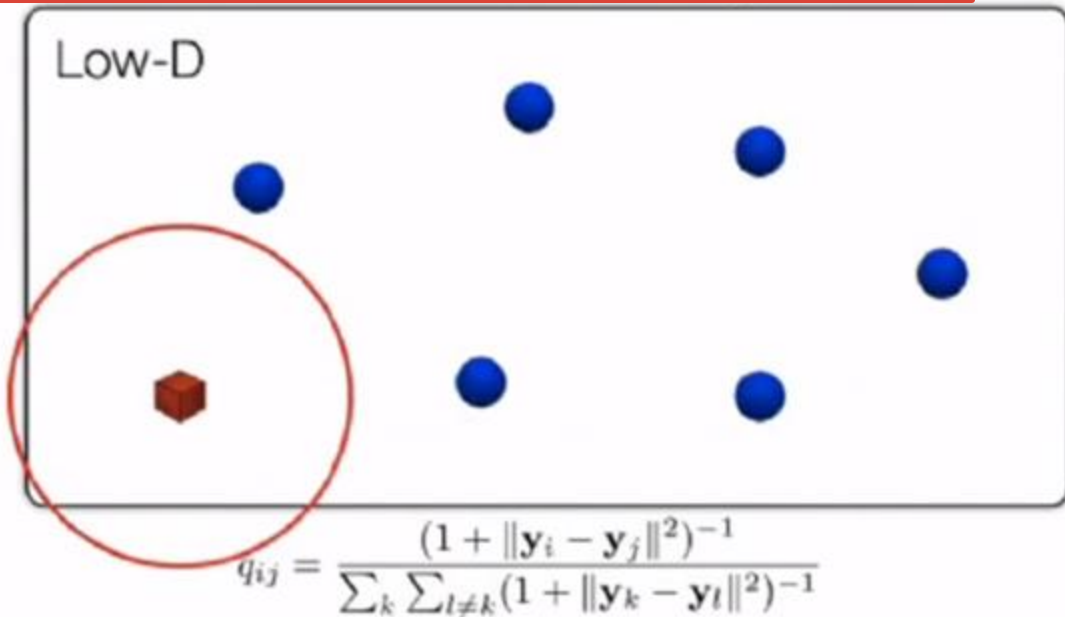
T-SNE : MODELING SIMILARITIES IN LOW-D



Lay out points in 2D/3D such that $q_{ij} \sim p_{ij}$

T-SNE : MODELING SIMILARITIES IN LOW-D

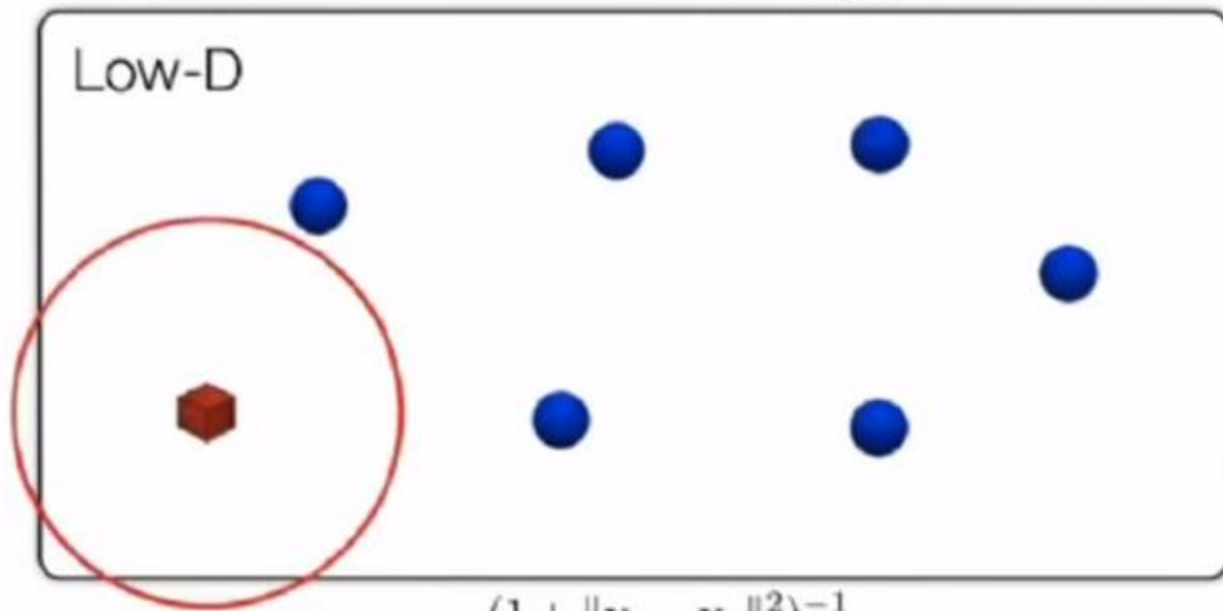
• Move points around to minimize: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$



Lay out points in 2D/3D such that $q_{ij} \sim p_{ij}$

T-SNE : MODELING SIMILARITIES IN LOW-D

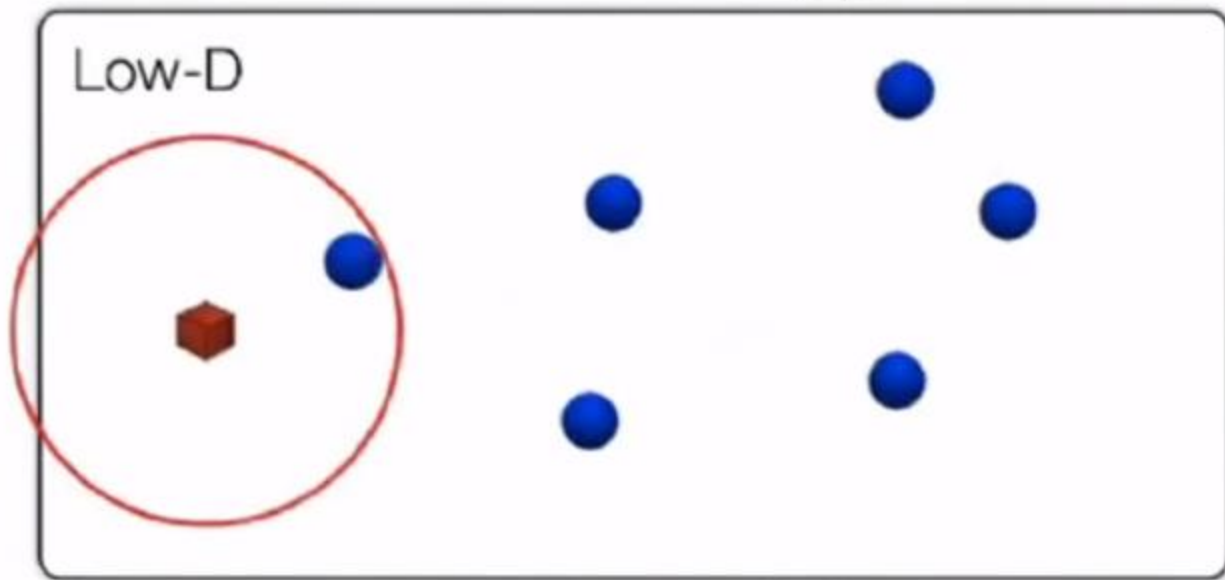
- Move points around to minimize: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$



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

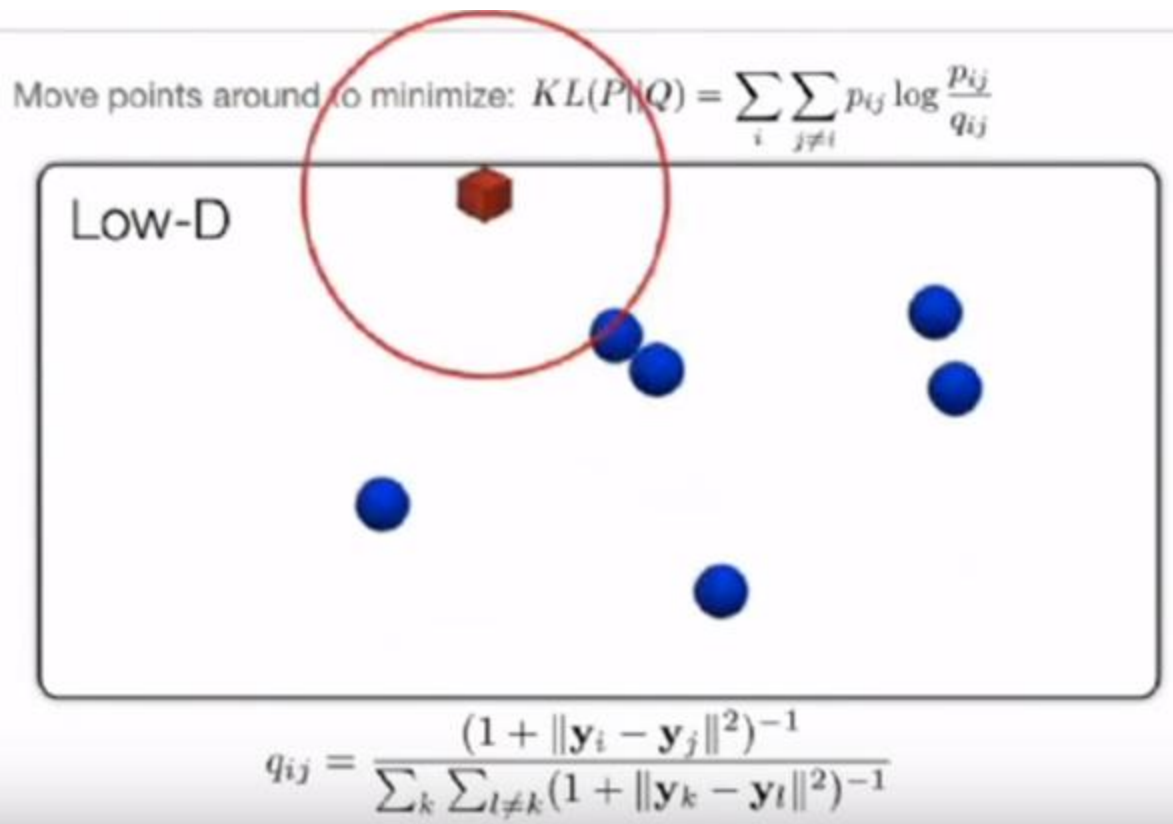
T-SNE : MODELING SIMILARITIES IN LOW-D

- Move points around to minimize: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$



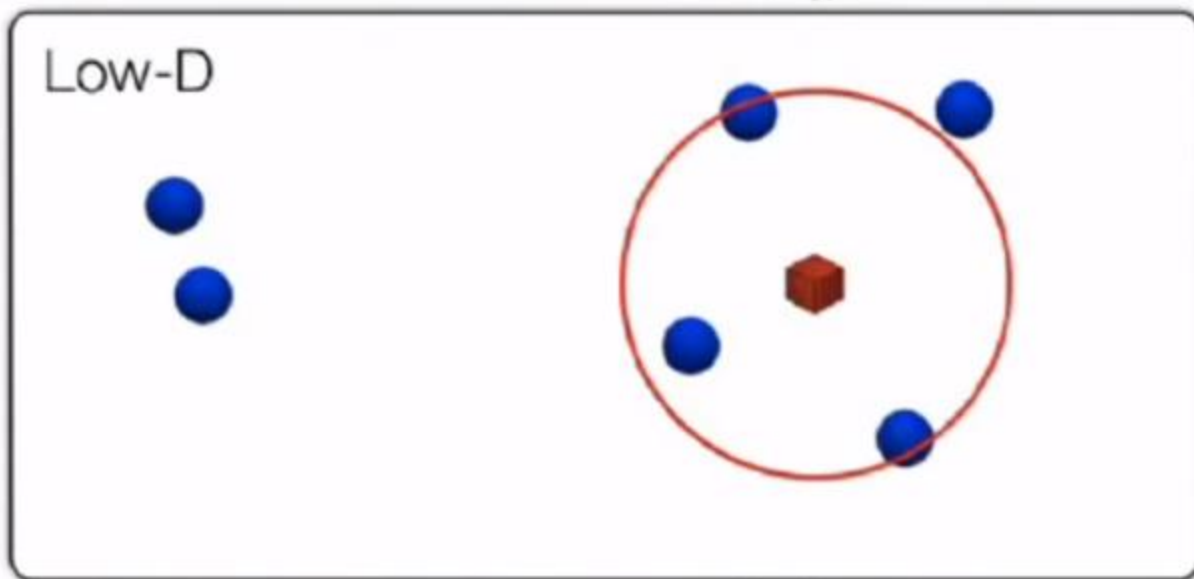
$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

T-SNE : MODELING SIMILARITIES IN LOW-D



T-SNE : MODELING SIMILARITIES IN LOW-D

Move points around to minimize: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$



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

T-SNE : SIGNIFICANCE OF KL-DIVERGENCE

- Kullback-Leibler divergence: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$

Not a symmetric distance. This is important!



- Large p_{ij} modeled by small q_{ij} ? Big penalty!

- Small p_{ij} modeled by large q_{ij} ? Small penalty!

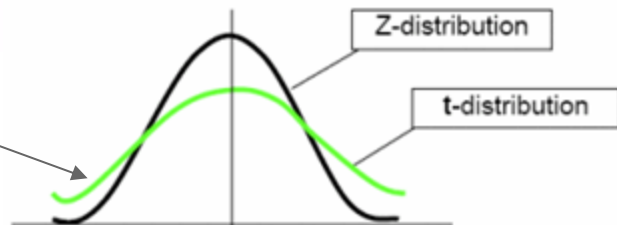
- Hence, t-SNE mainly preserves *local similarity structure* of the data

MOTIVATION

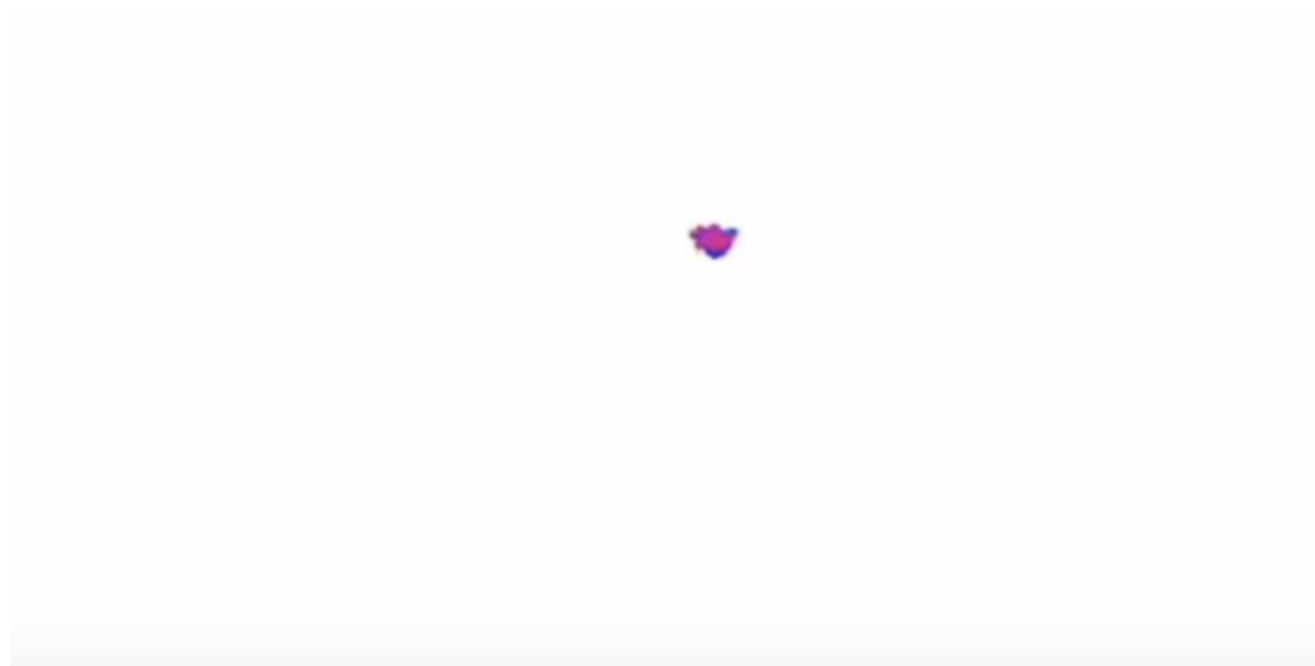
Why a Student-t distribution?

- Why do we define map similarities as $q_{ij} \propto (1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}$?
- Suppose data is *intrinsically* high-dimensional
- We try to model the *local structure* of this data in the map
- Result: dissimilar points have to be modeled as too far apart in the map!

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



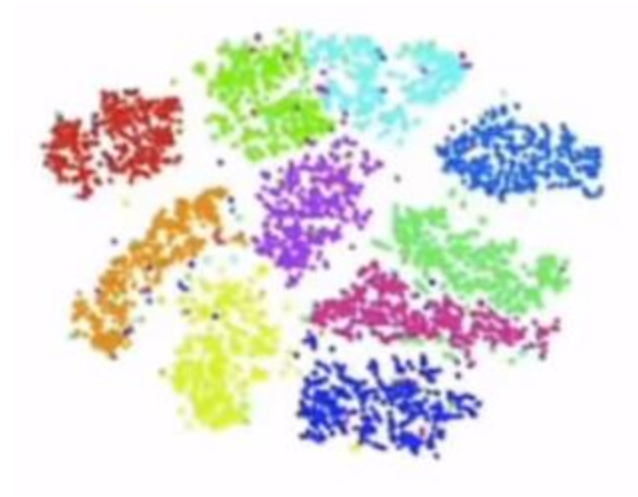
T-SNE IN ACTION : VISUALIZING HANDWRITTEN DIGITS



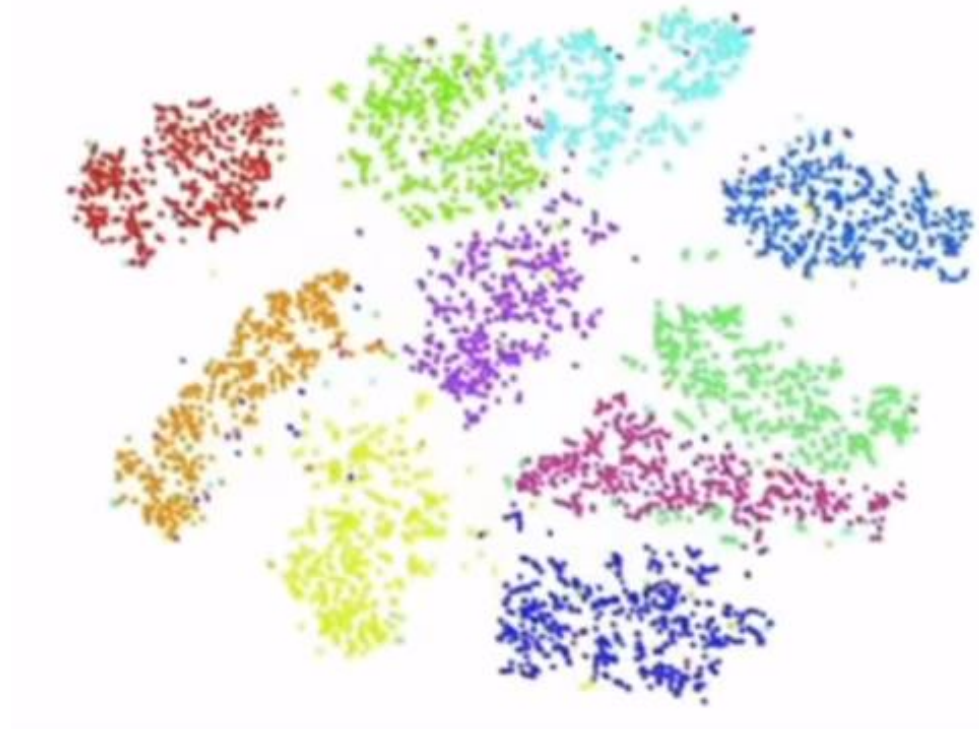
T-SNE IN ACTION : VISUALIZING HANDWRITTEN DIGITS



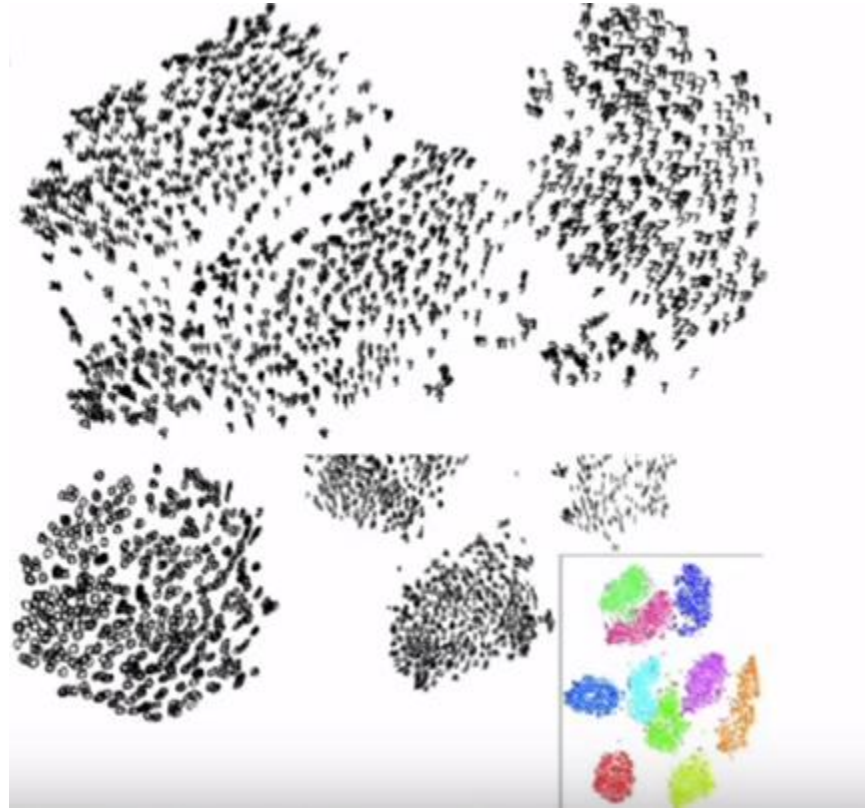
T-SNE IN ACTION : VISUALIZING HANDWRITTEN DIGITS



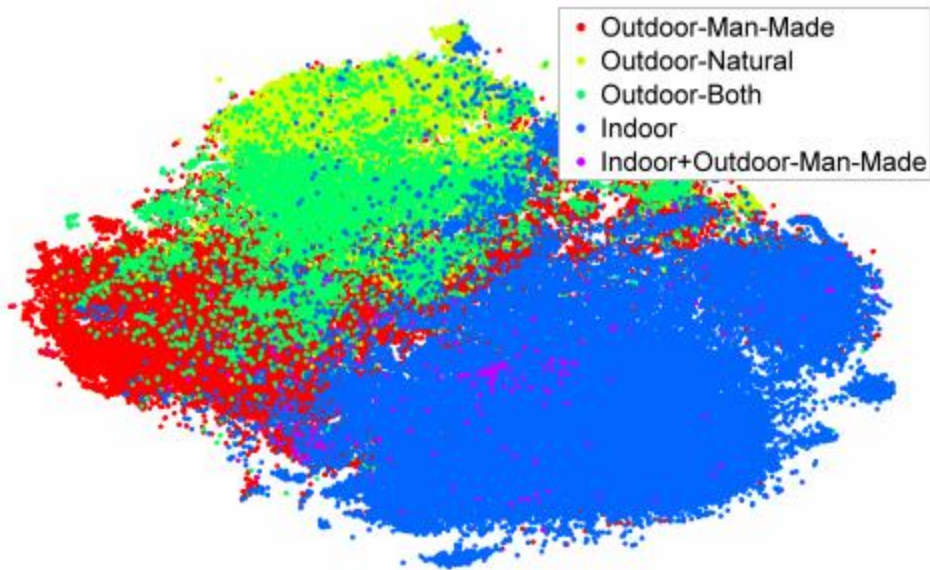
T-SNE IN ACTION : VISUALIZING HANDWRITTEN DIGITS



T-SNE : PRESERVING LOCAL SUBSTRUCTURE

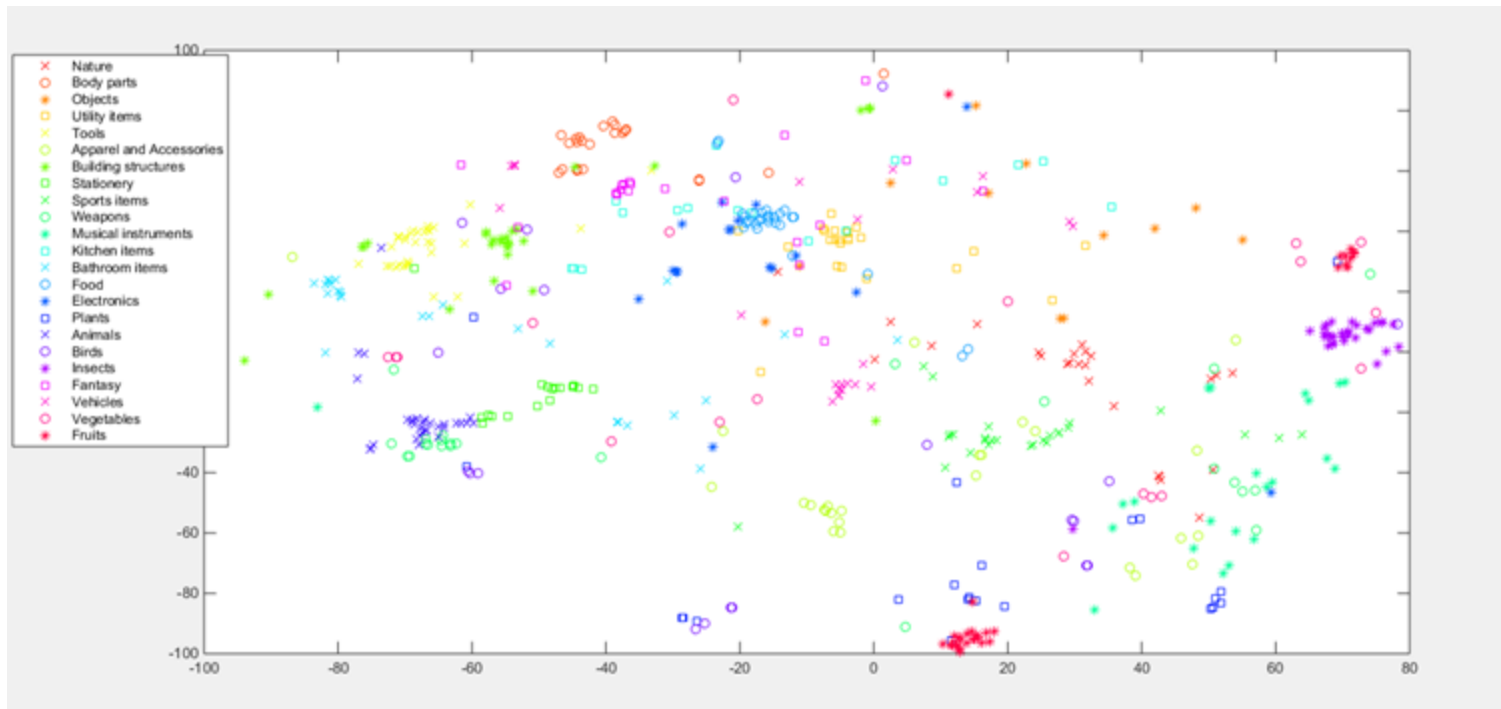


T-SNE : DEEP LEARNING



Visualizing deep image features

T-SNE : DEEP LEARNING



Visualizing sparsified sketch object features

T-SNE: OPTIMIZATION

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{j' \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_{j'}\|^2 / 2\sigma_i^2)}$$

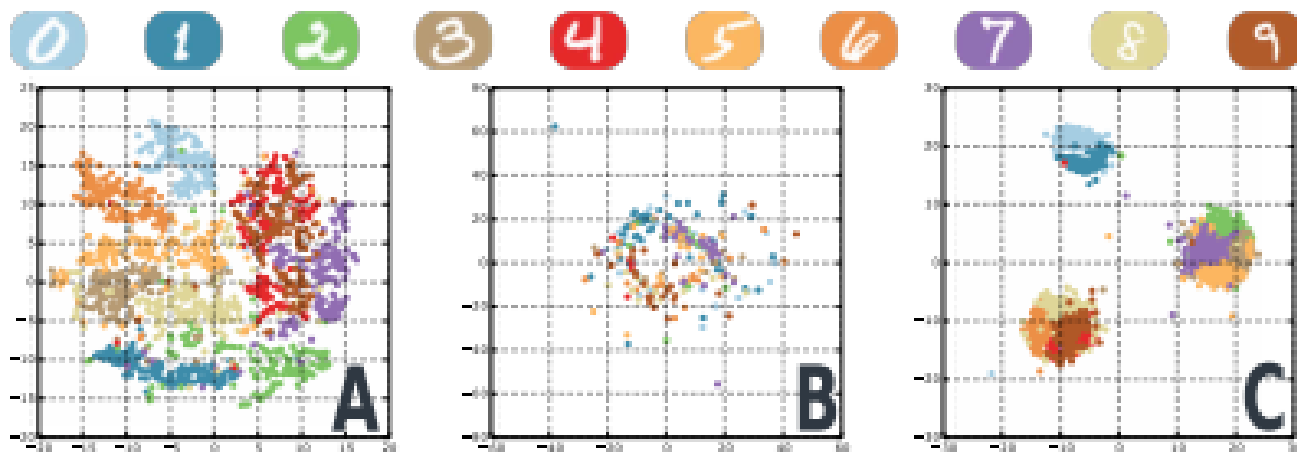
$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

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

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t) (\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)})$$

T-SNE : SNACK (ICCV 2015)



Learning Concept Embeddings With Combined Human-Machine Expertise, Wilber et al., ICCV 2015

T-SNE : MULTIPLE MAPS

Multiple maps t-SNE

- Construct multiple maps, and give each object a point in each map
- Assign an *importance weight* to each point
- Define the similarity between two points under the multiple maps model as a weighted sum over the similarities in the individual maps

Map 1



Map 2



T-SNE : MULTIPLE MAPS



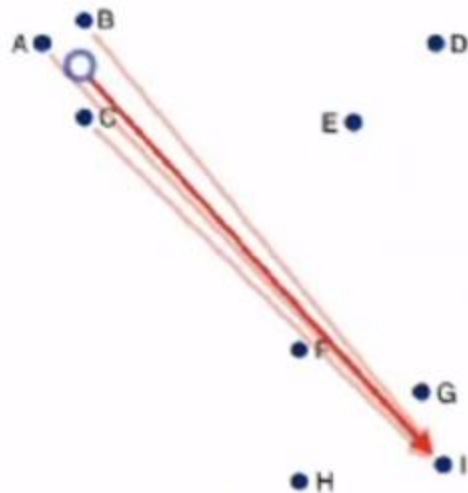
Monarchy



T-SNE: BARNES-HUT APPROXIMATION

Barnes-Hut approximation

- Approximate such similar interactions by a *single* interaction:



T-SNE : ADDING IT UP

- t-SNE = t-distributed stochastic neighborhood embedding
- 'N' : cares a lot about modelling local/nearby similarities well
- 'S' : gradient descent used to decide how to move points is stochastic
- t-distributed : Used to characterize similarities in low-d space
- Note: t-SNE is good for VISUALIZATION, not NECESSARILY for DIMENSIONALITY REDUCTION

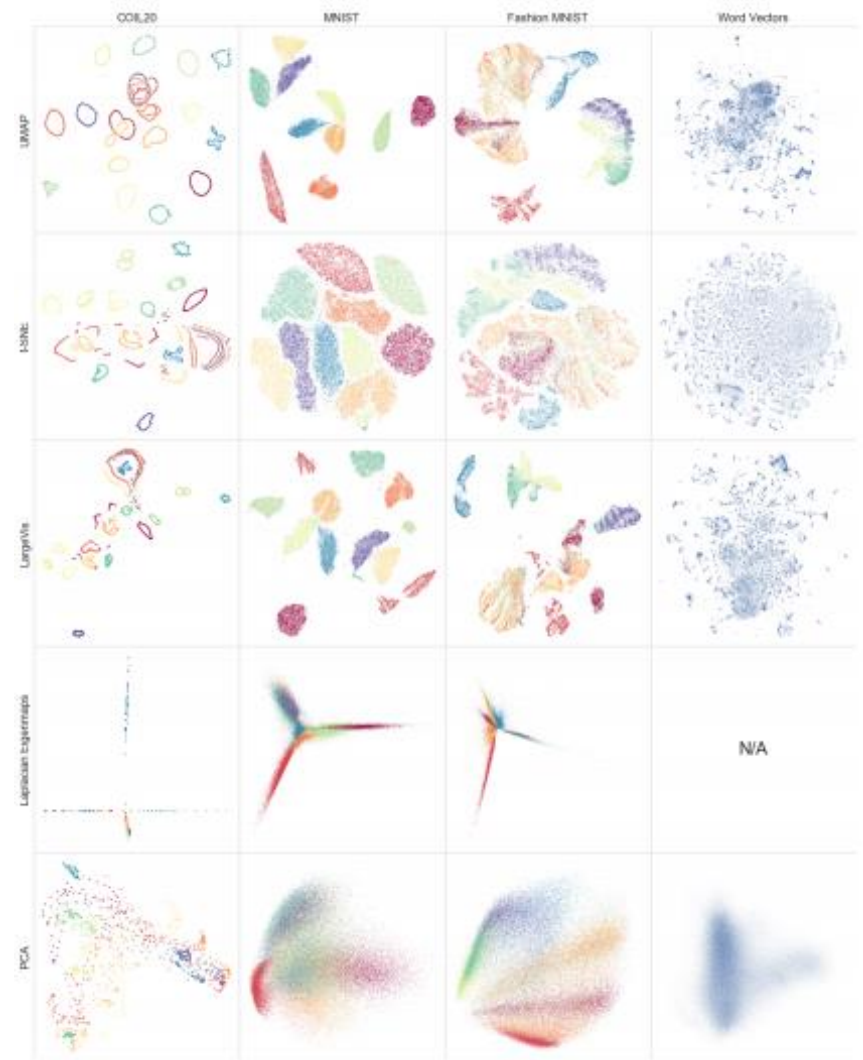
IMPORTANT CAVEAT

- t-SNE : Easy to “abuse”
- Will let you see what you “wish” to see :)
- How to interpret t-SNE : <http://distill.pub/2016/misread-tsne/>

ADDITIONALLY ...

1. t-SNE often fails to preserve global structure of the dataset;
2. t-SNE tends to suffer from "overcrowding" when N grows above $\sim 100k$;
3. Barnes-Hut runtime is too slow for large N .

UMAP (UNIVERSAL MANIFOLD APPROXIMATION)



REFERENCES

- Talk : <https://www.youtube.com/watch?v=RJVL80Gg3lA>
- Paper : <http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>
- <https://stats.stackexchange.com/questions/270391/should-dimensionality-reduction-for-visualization-be-considered-a-closed-probl/270414>