T-DISTRIBUTED STOCHASTIC NEIGHBORHOOD EMBEDDING (T-SNE)

12.04.2019

Ravi Kiran Sarvadevabhatla

CVIT, IIIT Hyderabad

Visualization

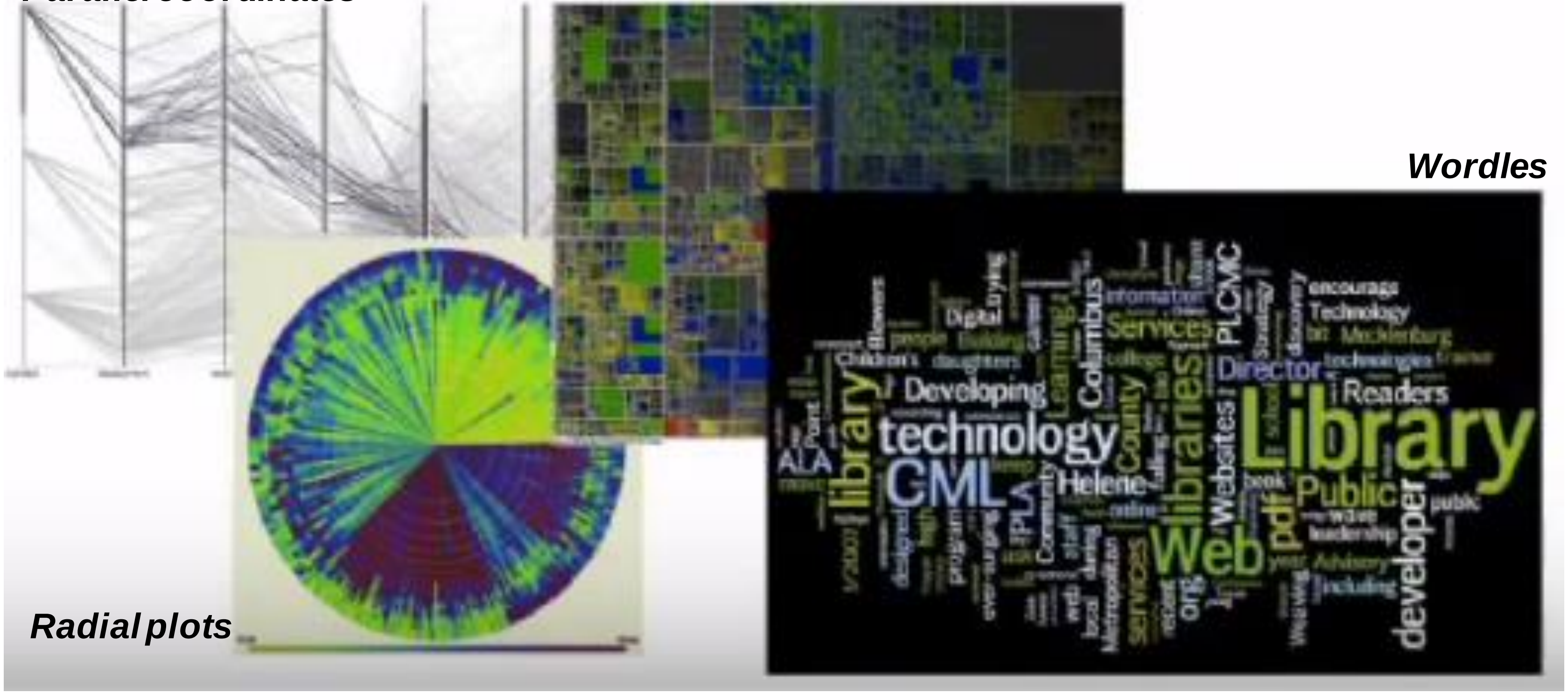
- Given a collection of N objects - x1,x2 … xN
- How can we get a feel for how these N objects are arranged in data space (d-dimensional)?
- Why go through all this in the first place?

Visualization

Parallel coordinates

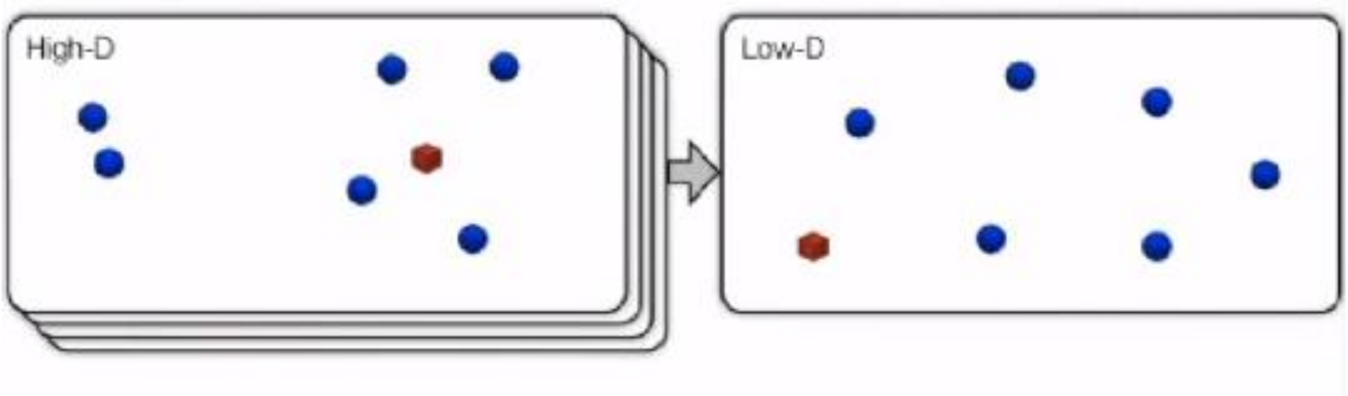
Treemaps

Wordles



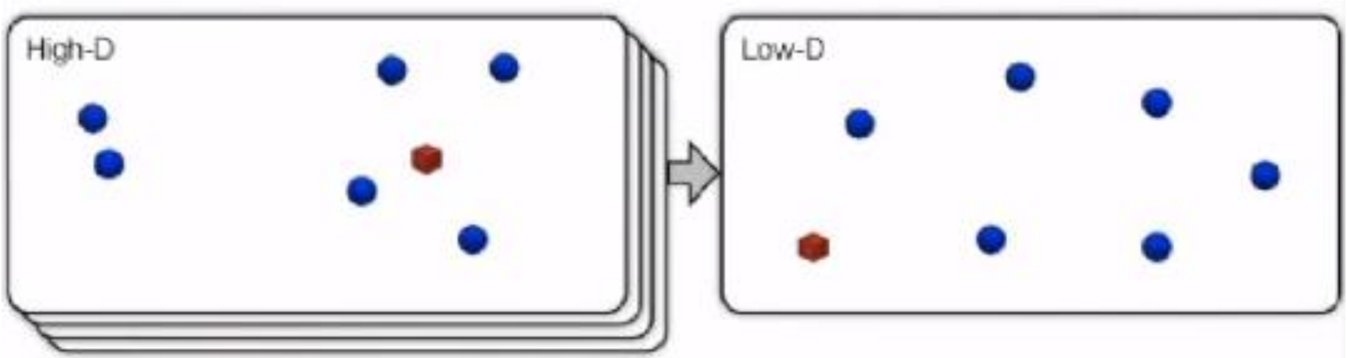
Radial plots


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 <u>similarities</u> in data

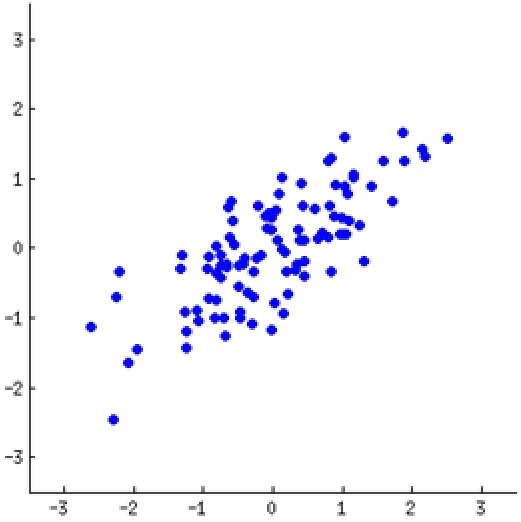
Formulation: Minimize some objective function which measures discrepancy between similarities in the data and similarities in the map.

PCA revisited ...

- What does PCA do ?

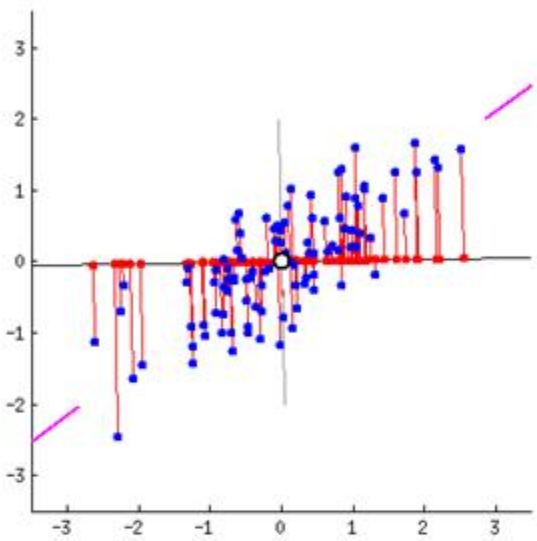
PCA revisited ...

- What does PCA do ?

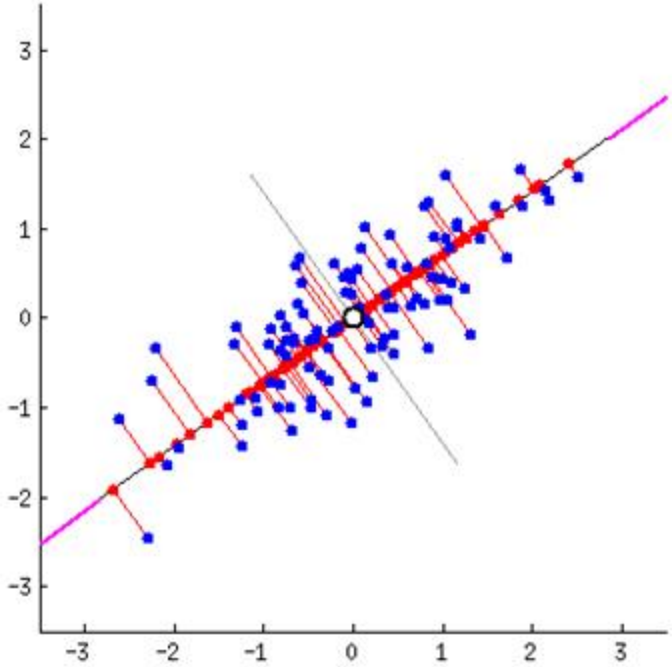


PCA revisited ...

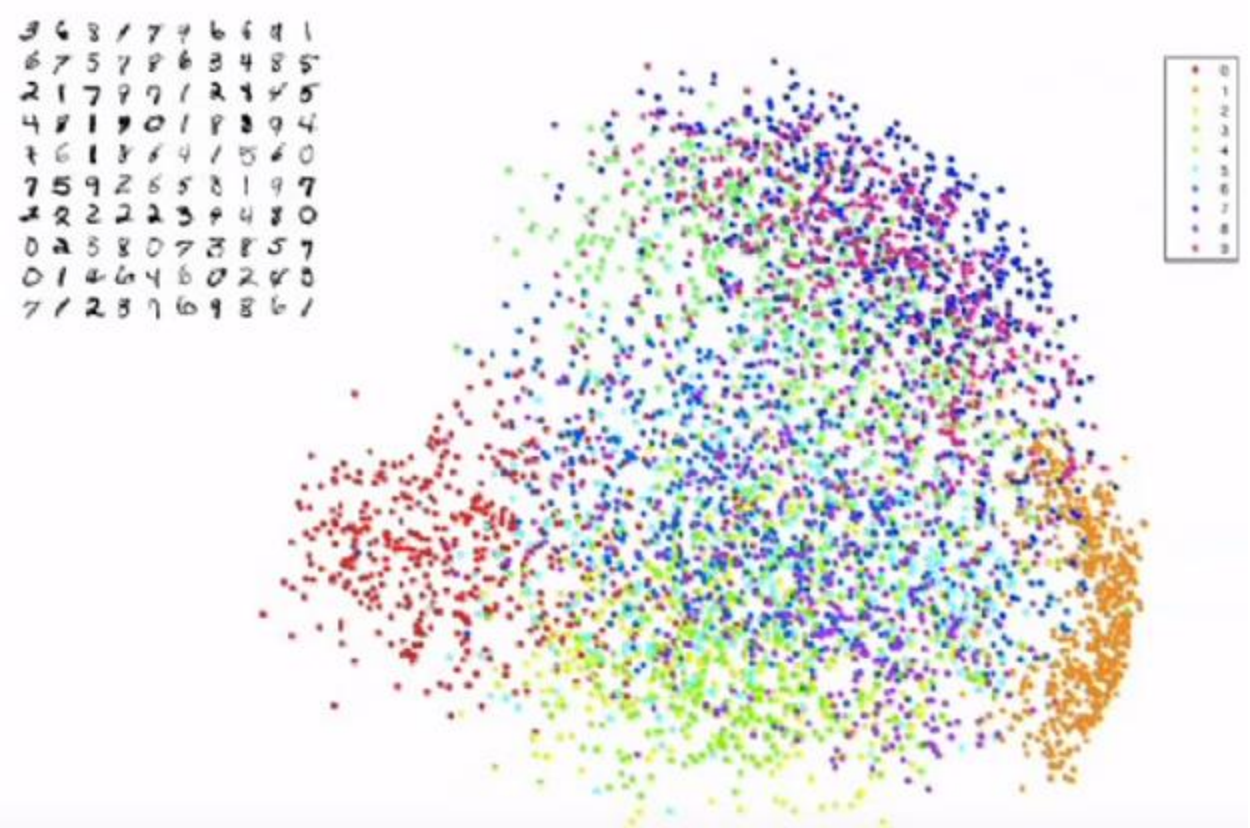
- What does PCA do ?



PCA revisited ...



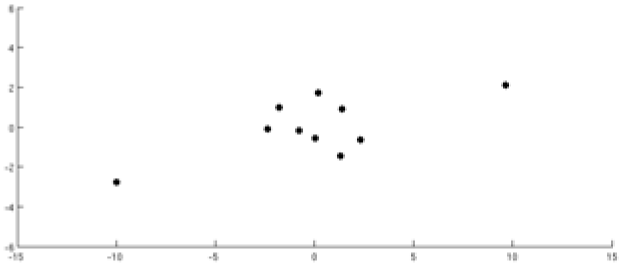
PCA revisited ...



PCA revisited ...

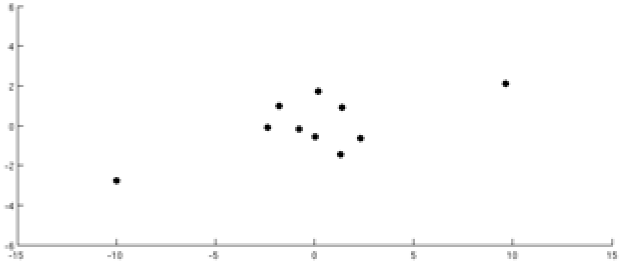
Is PCA minimizing the right objective function ?

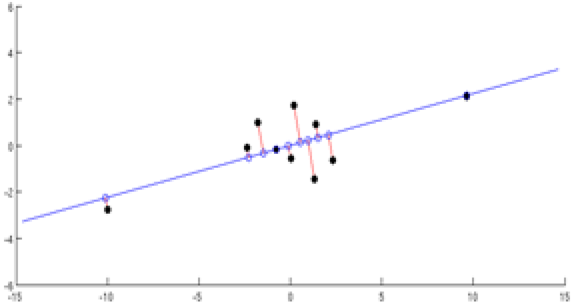
- PCA is mainly concerned with preserving **large** pairwise distances in the map



Is PCA minimizing the right objective function ?

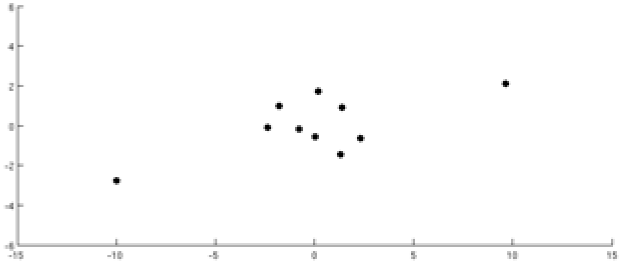
- PCA is mainly concerned with preserving **large** pairwise distances in the map

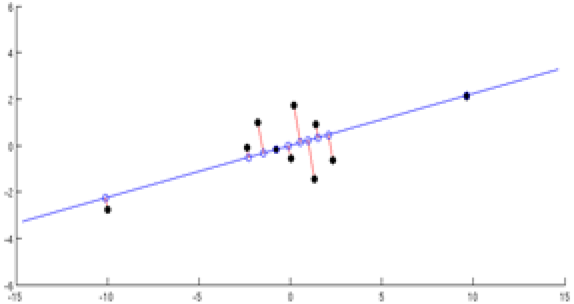




Is PCA minimizing the right objective function ?

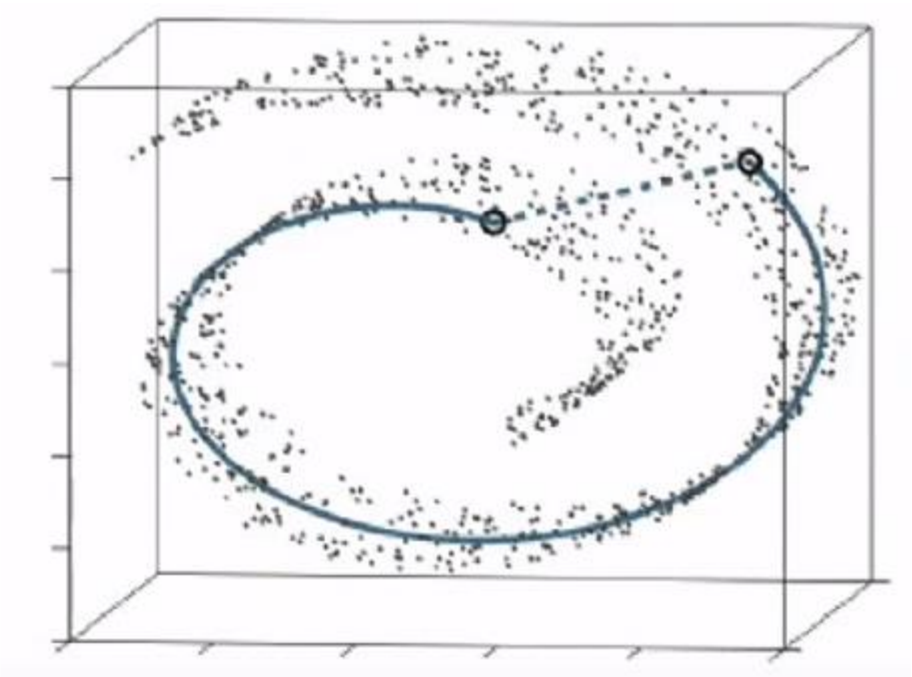
- PCA is mainly concerned with preserving **large** pairwise distances in the map





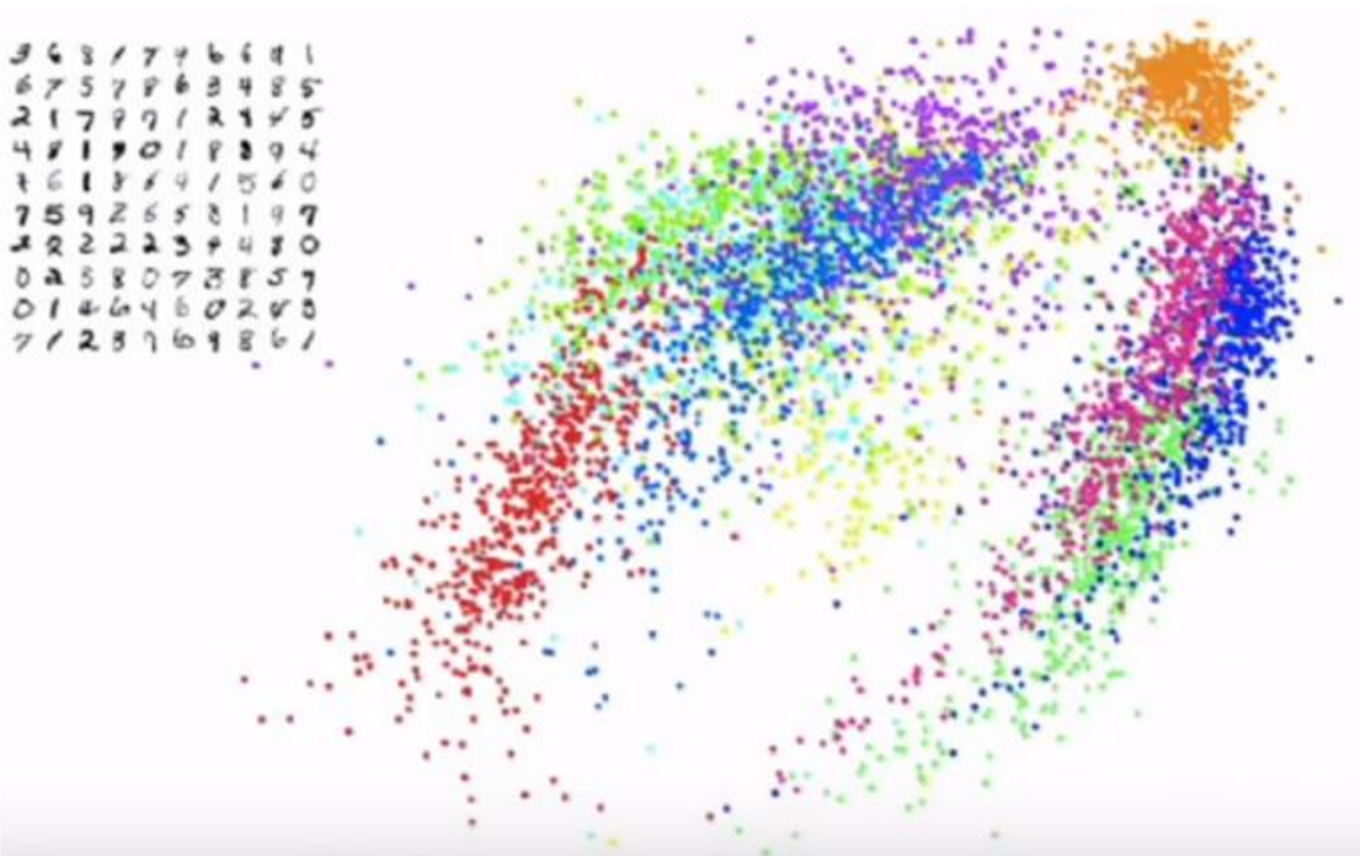
- “crowding” problem for smaller pairwise distances

Is PCA minimizing the right objective function ?

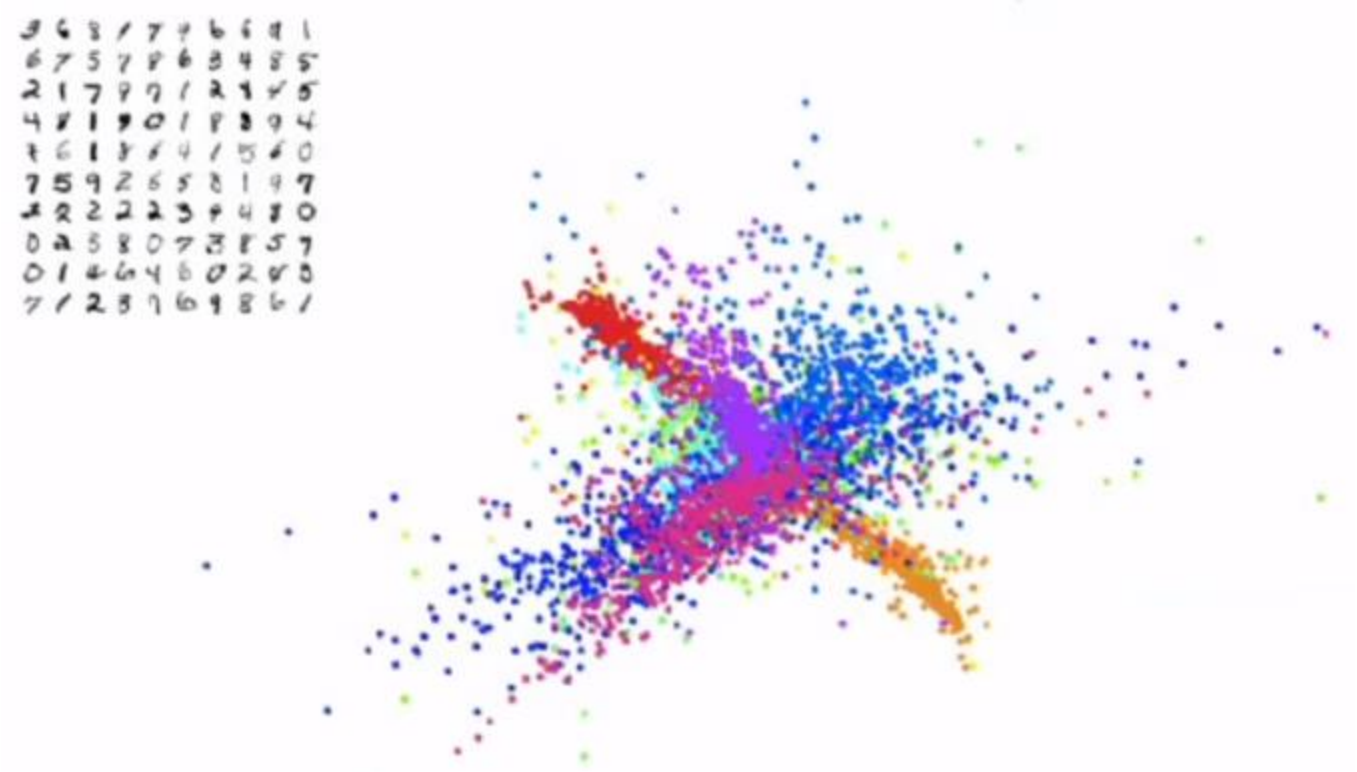


- Not so good at preserving local similarities
- Linearity is too restrictive an assumption !

Isomap

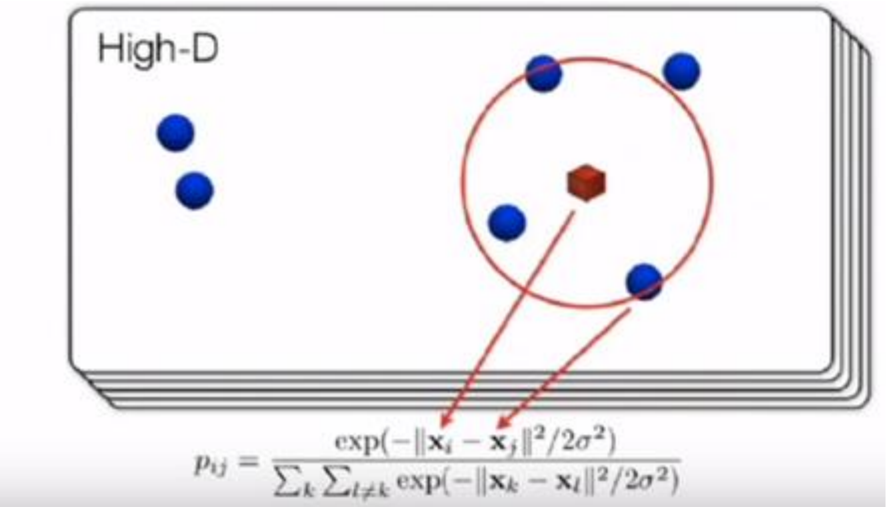


Locally linear embedding (LLE)



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=>Large p_ij, Far points => 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

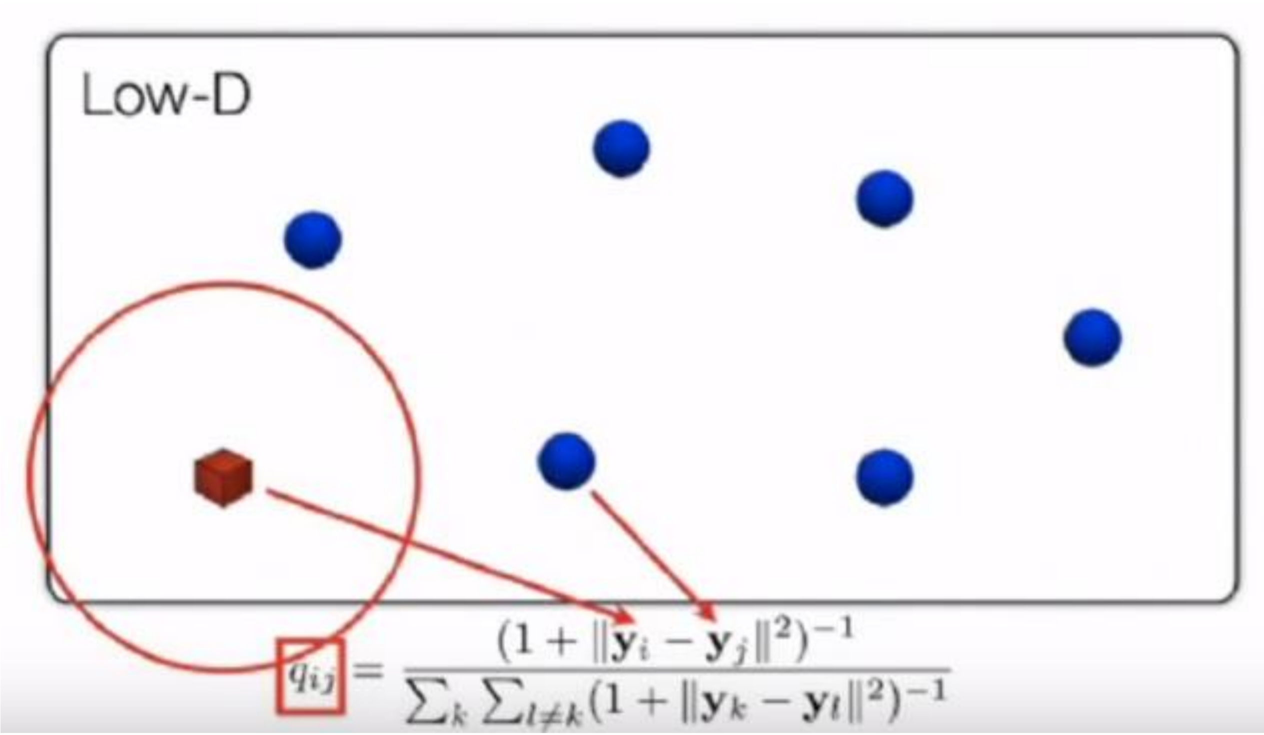
- We set the bandwidth σ_i such that the *conditional* has a *fixed perplexity*

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

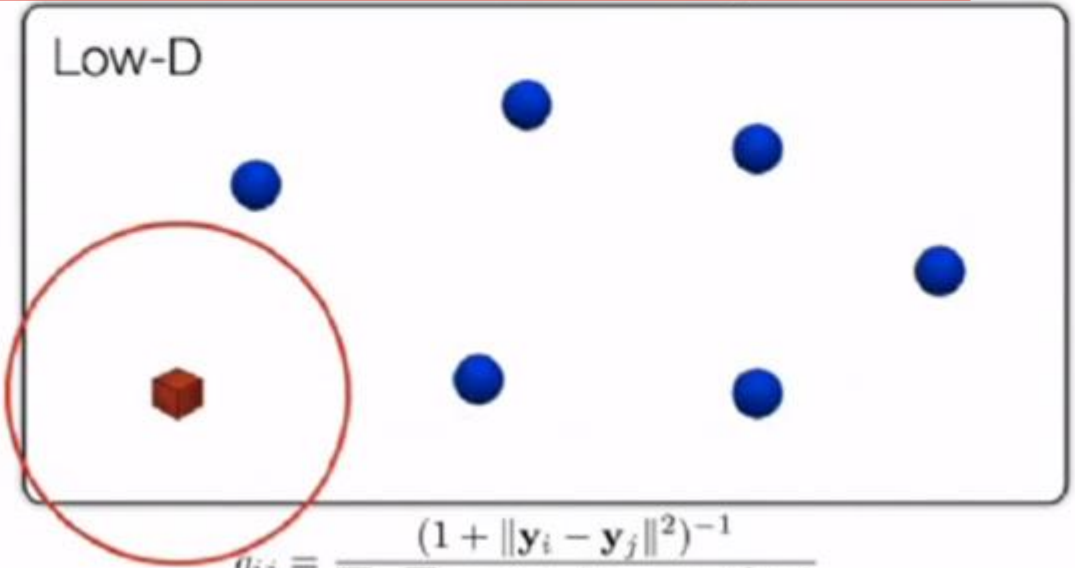
t-SNE : Modeling Similarities in Low-D



Lay out points in 2D/3D such that q_ij ~ p_ij

t-SNE : Modeling Similarities in Low-D

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

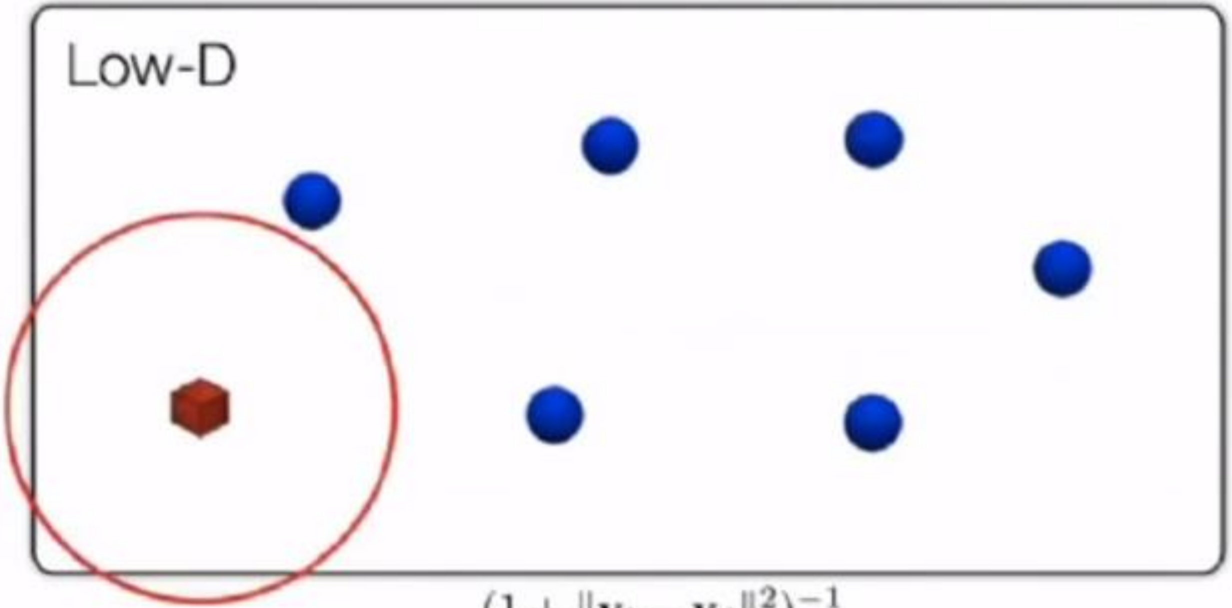


$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

Lay out points in 2D/3D such that q_ij ~ p_ij

t-SNE : Modeling Similarities in Low-D

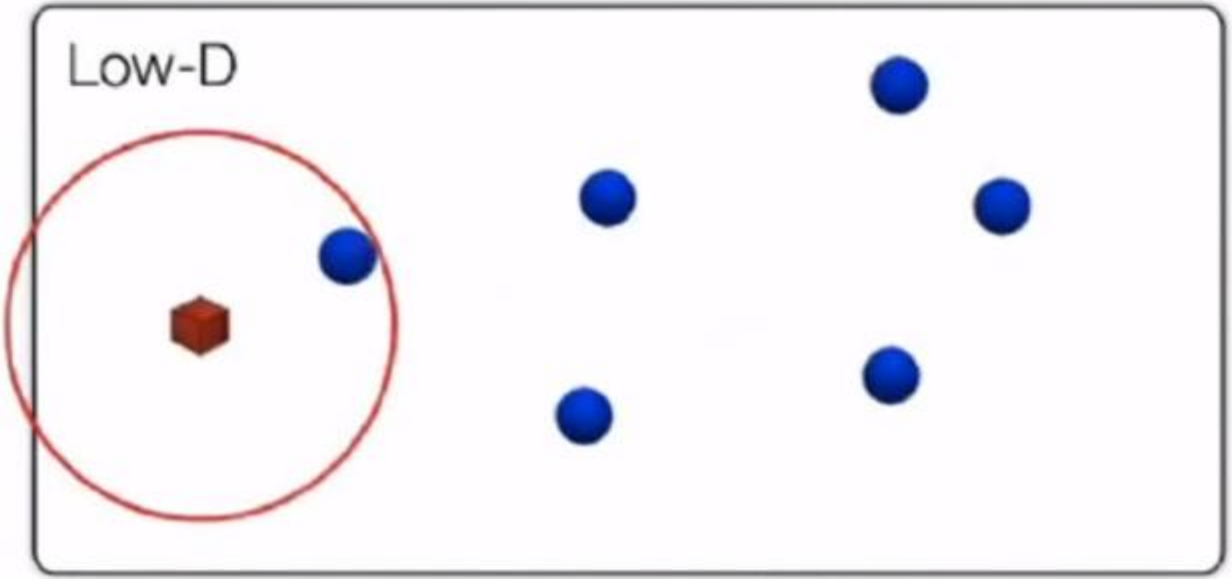
- Move points around to minimize: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$



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

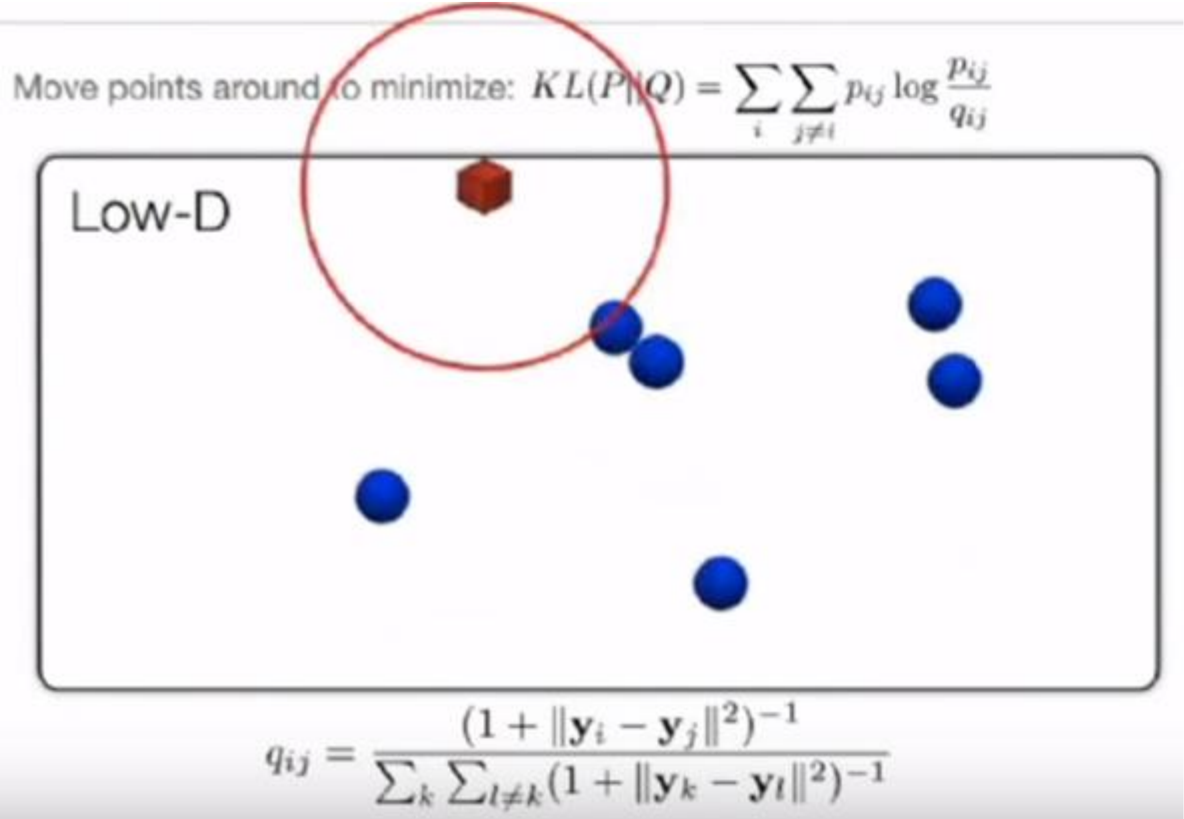
t-SNE : Modeling Similarities in Low-D

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



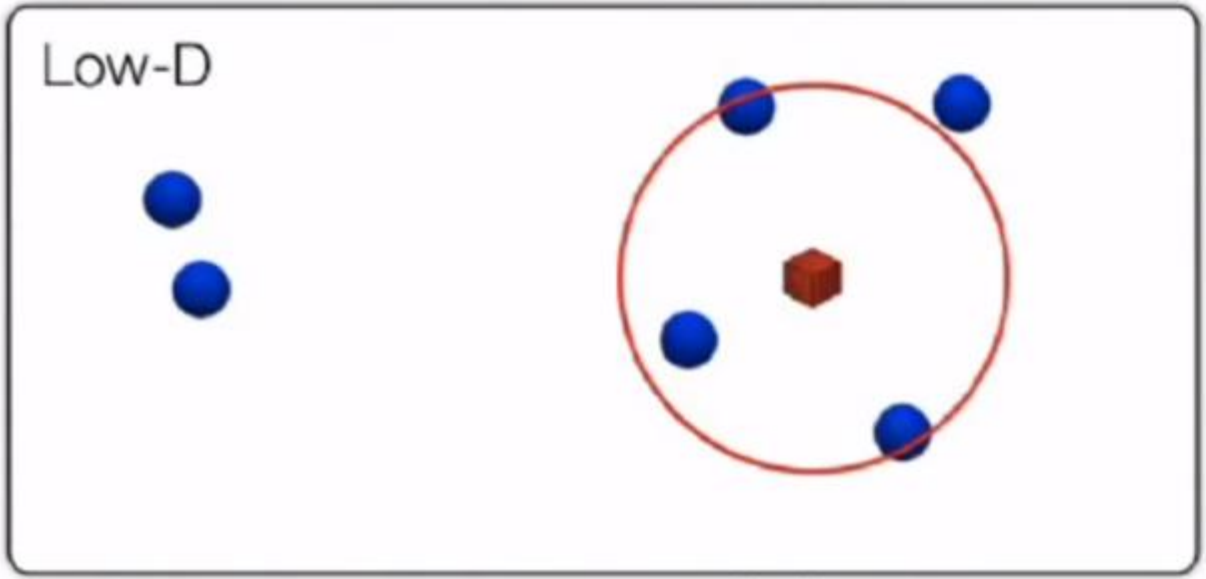
$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

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!

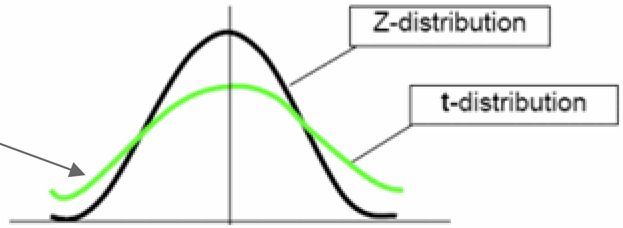
- Hence, t-SNE mainly preserves *local similarity structure* of the data

Motivation

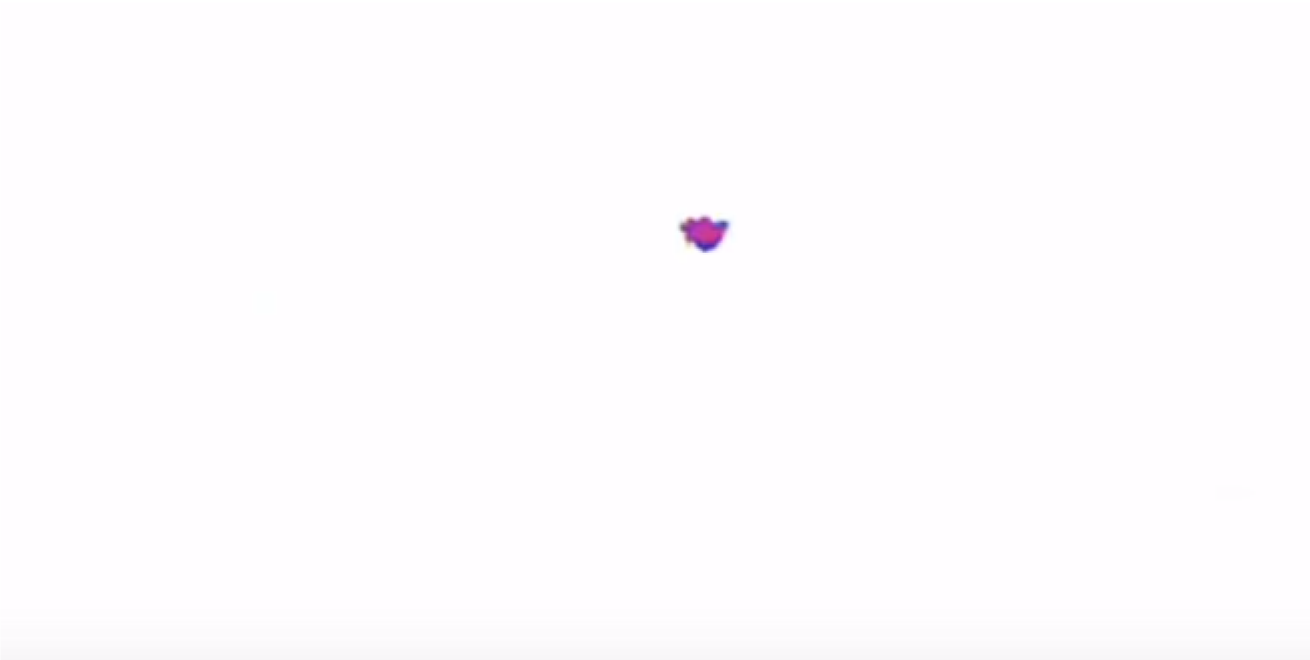
Why a Student-t distribution?

- 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!

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

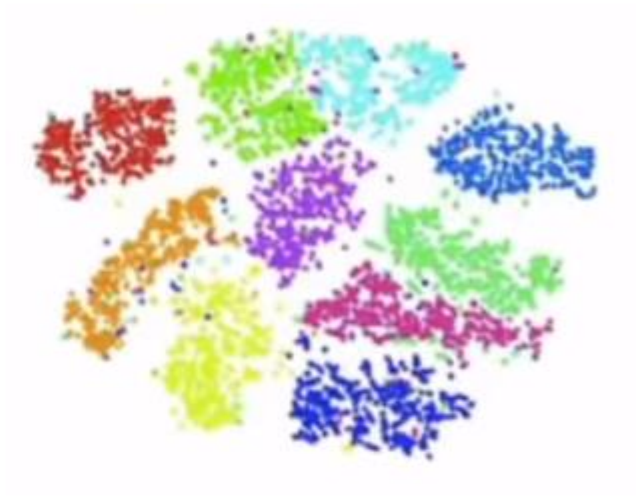


t-SNE in ACTION : Visualizing Handwritten DIgits

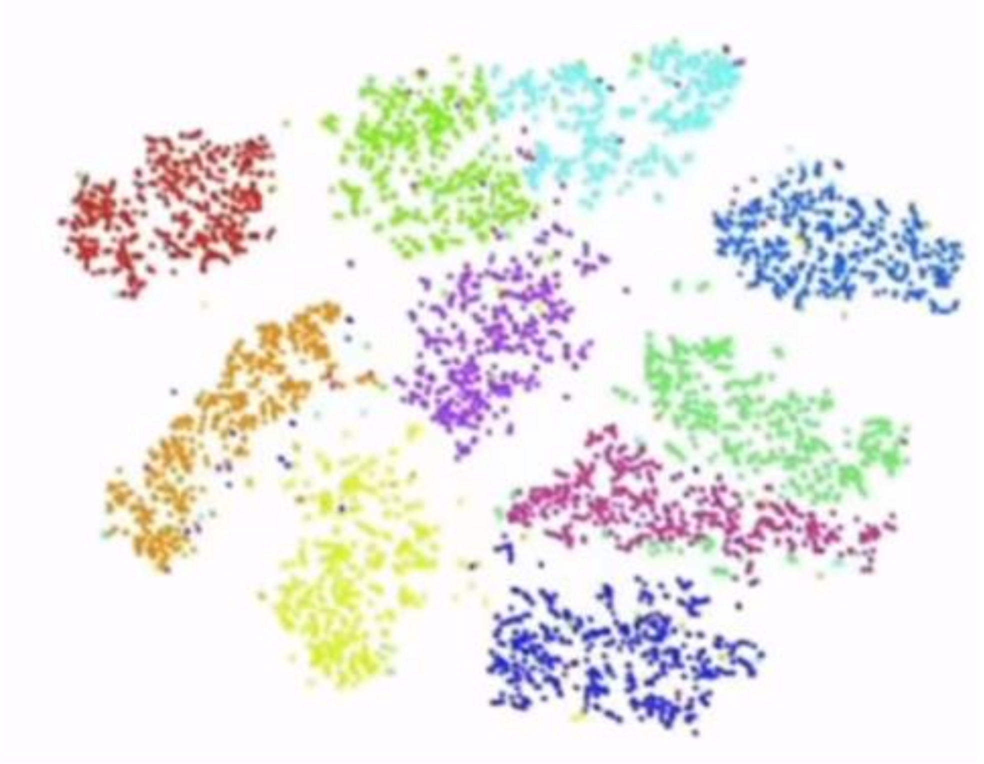


t-SNE in ACTION : Visualizing Handwritten DIgits

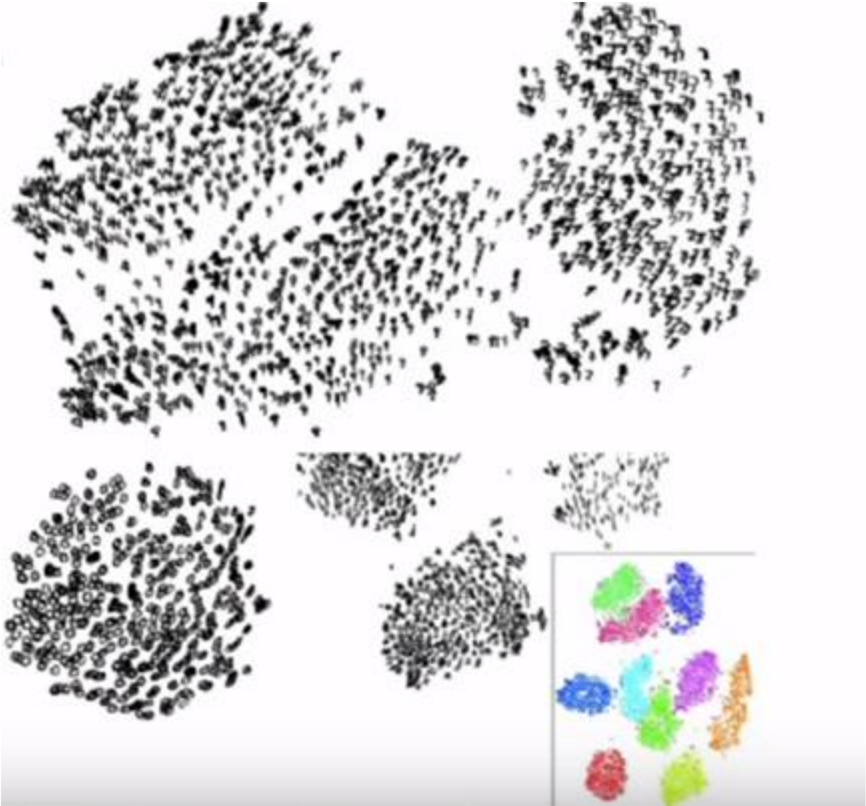
t-SNE in ACTION : Visualizing Handwritten DIgits



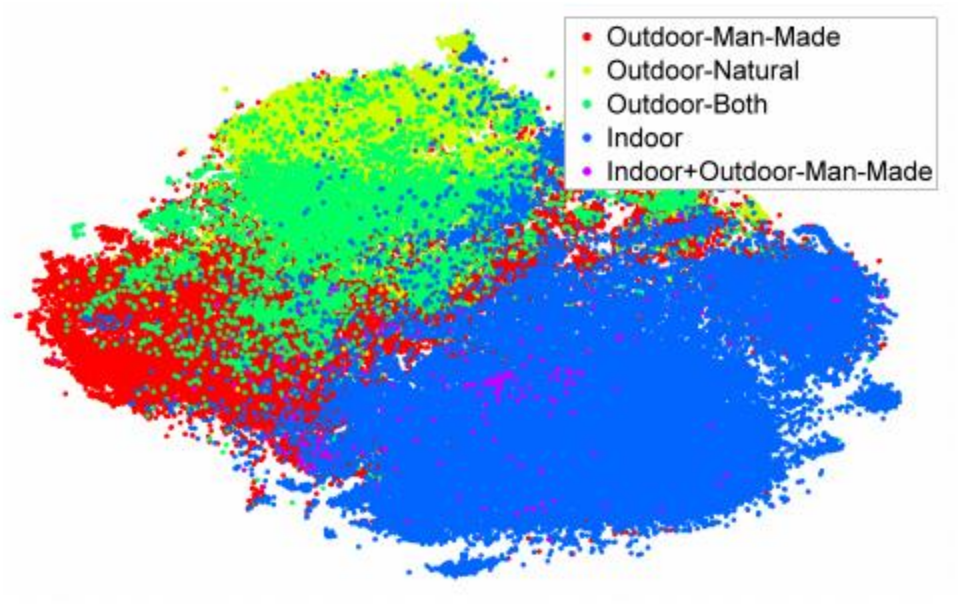
t-SNE in ACTION : Visualizing Handwritten DIgits



T-SNE : Preserving Local Substructure

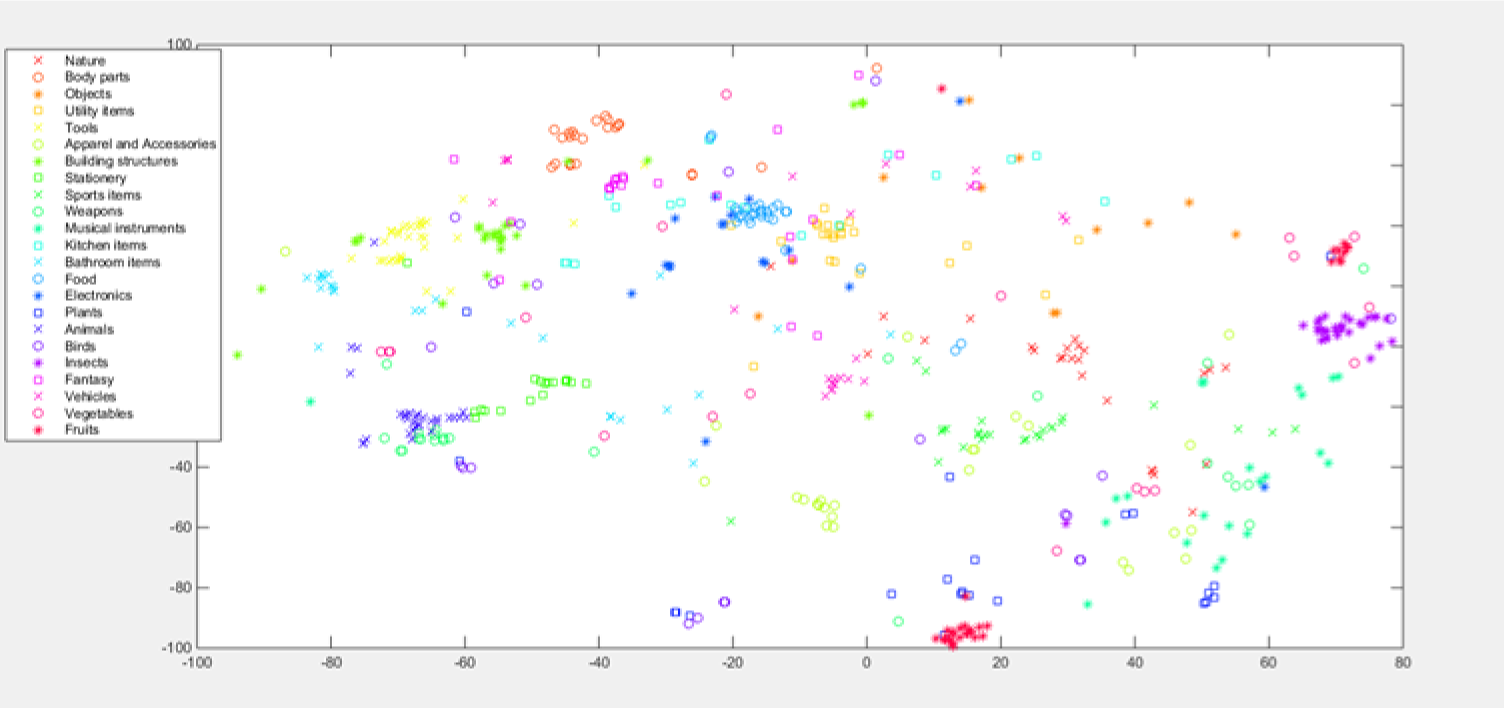


t-SNE : Deep Learning



Visualizing deep image features

t-SNE : Deep Learning



Visualizing sparsified sketch object features

t-SNE: Optimization

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

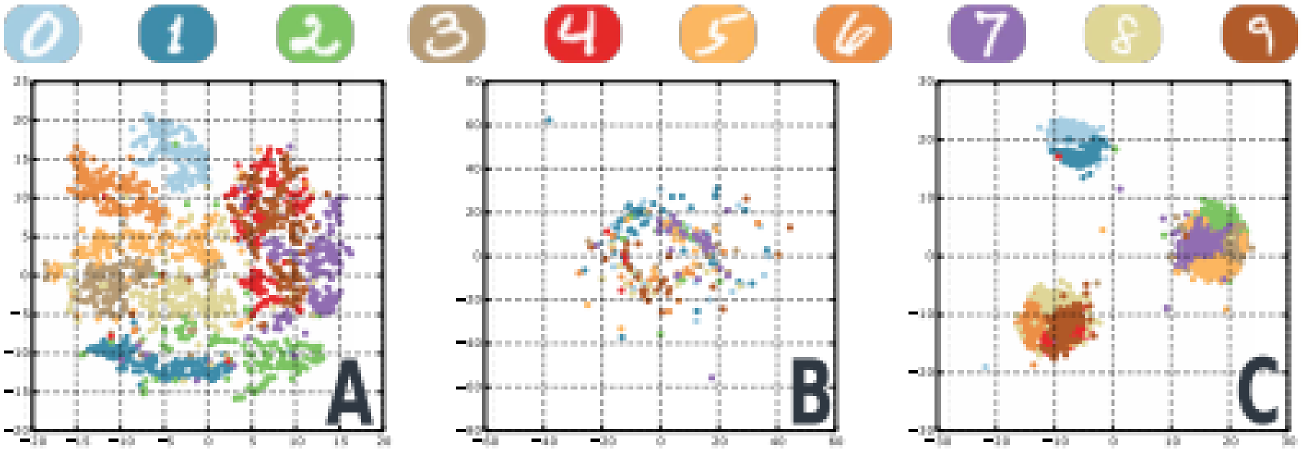
$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

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

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

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

t-SNE : Snack (ICCV 2015)

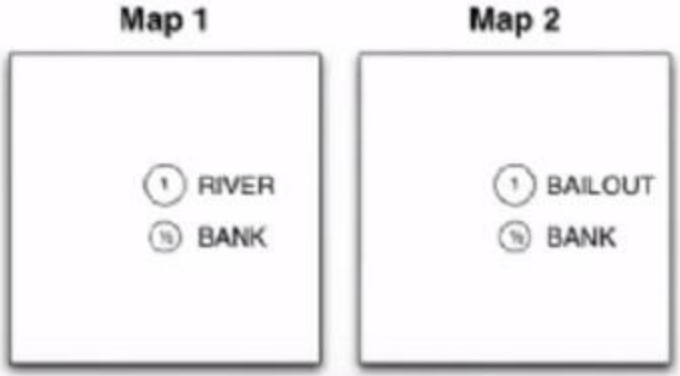


Learning Concept Embeddings With Combined Human-Machine Expertise, Wilber et al., ICCV 2015

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



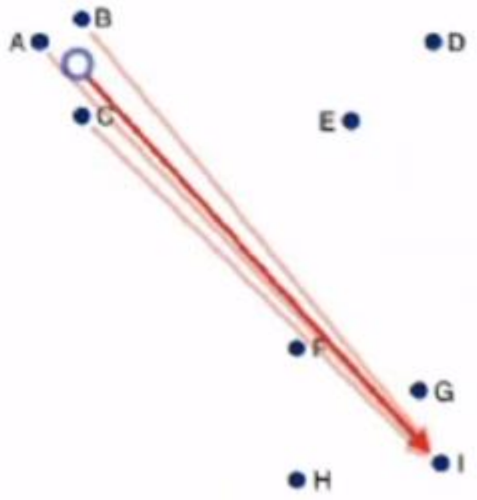
t-SNE : Multiple maps

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

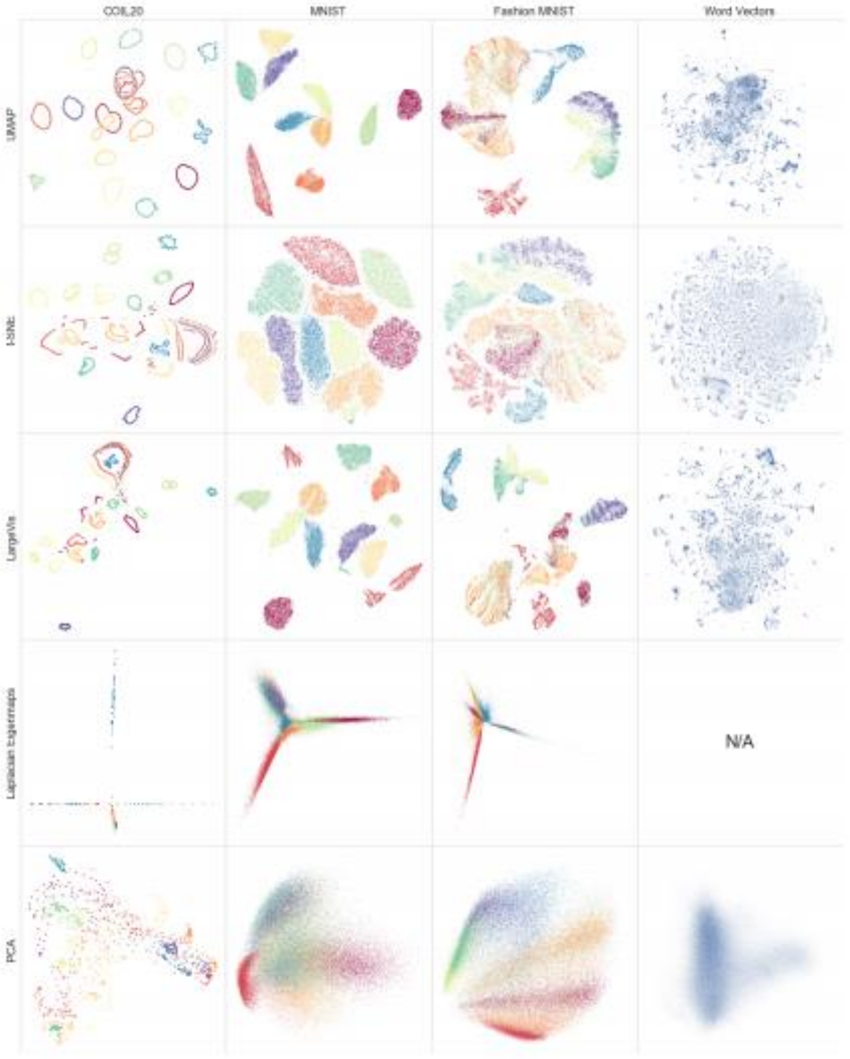
Important Caveat

- 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 ~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