T-SNE

(Dec 21, 2016)

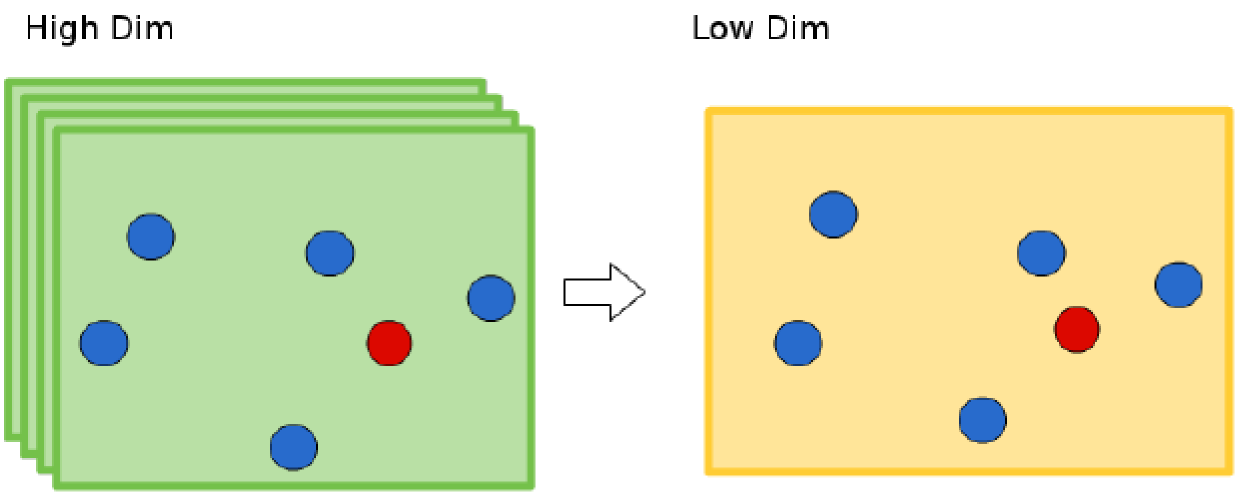
Cheng Jianlin

Outline

- I. Stochastic Neighbor Embedding
- II. t-Distributed Stochastic Neighbor Embedding
- III. Accelerating t-SNE using Tree-Based Algorithms

• I. Stochastic Neighbor Embedding

Preserve the neighborhood



• I. Stochastic Neighbor Embedding

·**First convert the high-dimensional Euclidean distances between datapoints into probabilities that represent similarities.**

·**Similarity of datapoints in High Dimension**

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

·**Similarity of datapoints in Low Dimension**

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

·**To evaluate a map:**

-- Compute the Kullback-Leibler divergence between the probabilities in the high-dimensional and low-dimensional spaces

• I. Stochastic Neighbor Embedding

- **Cost function:**

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

- **Perplexity:**

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

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

- **Gradient:**

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

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

• II. . t-Distributed SNE

- High-dimensional map:

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

- Low-dimensional map:

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

- Why a Student-t distribution:

--dissimilar points have to be modeled as too far apart in the map

• II. . t-Distributed SNE

Cost function:

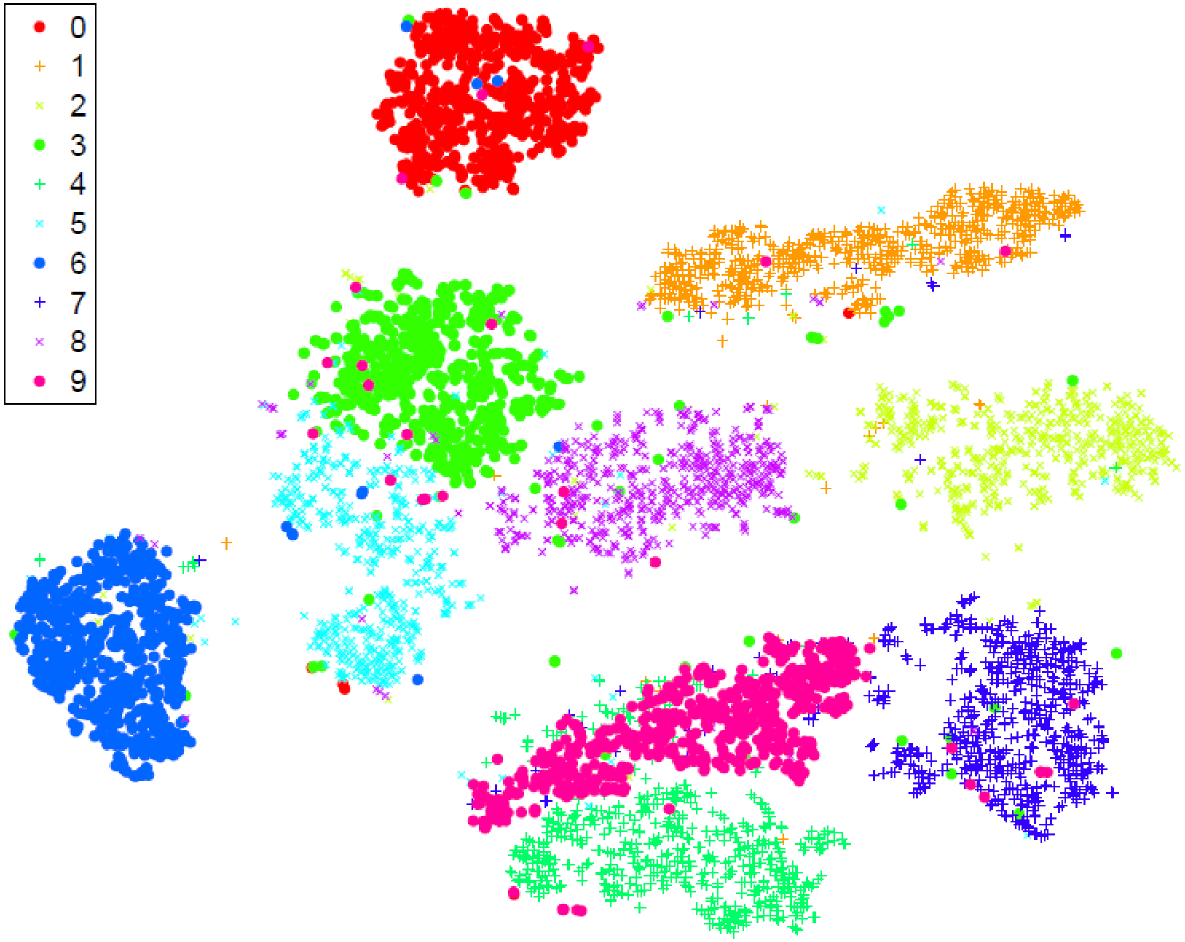
$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- Large p_{ij} modeled by small q_{ij}: Large penalty
- Small p_{ij} modeled by large q_{ij}: Small penalty
- t-SNE mainly preserves local similarity structure of the data

Gradient:

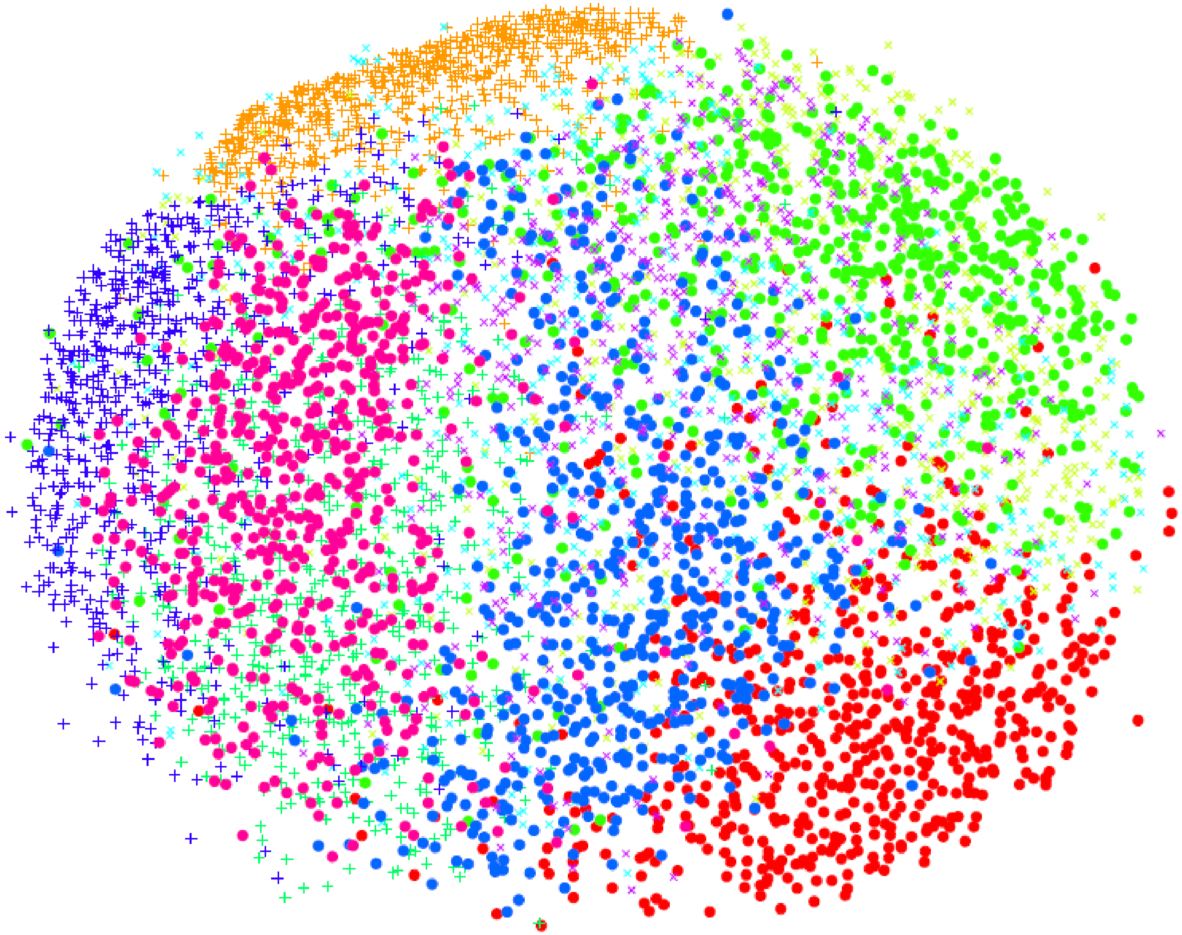
$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}$$

• II. . t-Distributed SNE



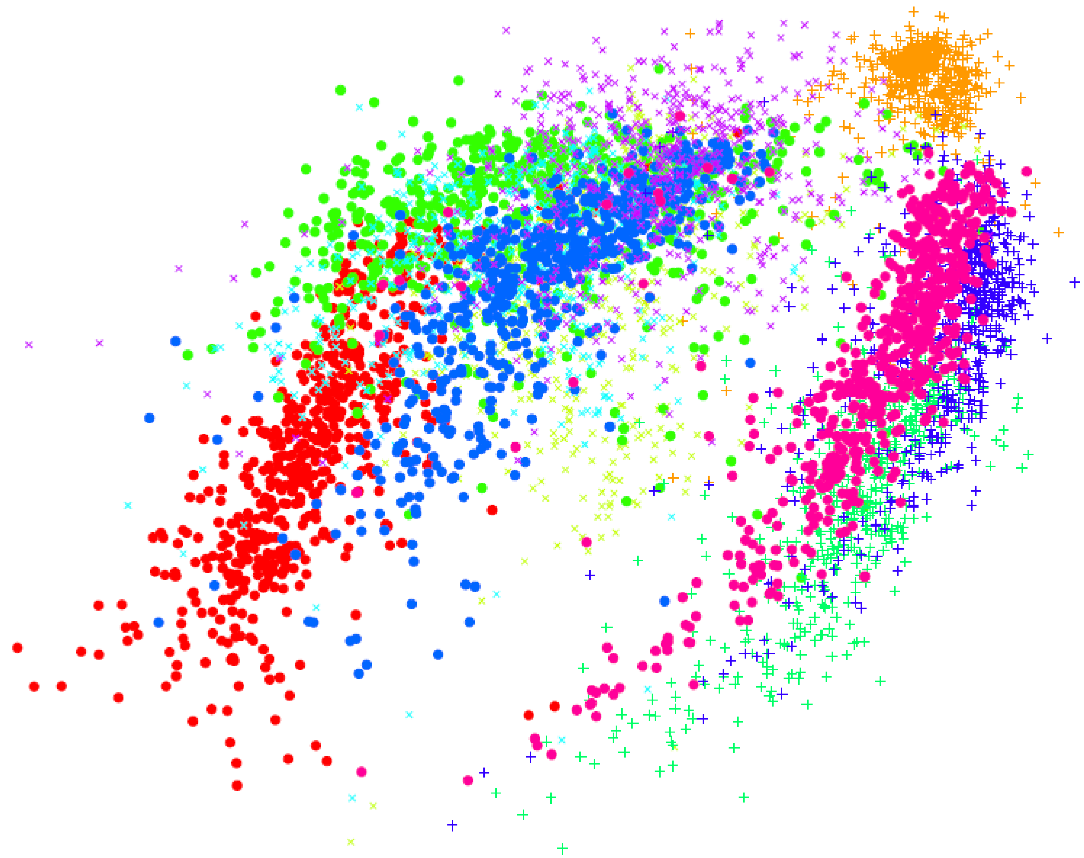
(a) Visualization by t-SNE.

- **II. . t-Distributed SNE**



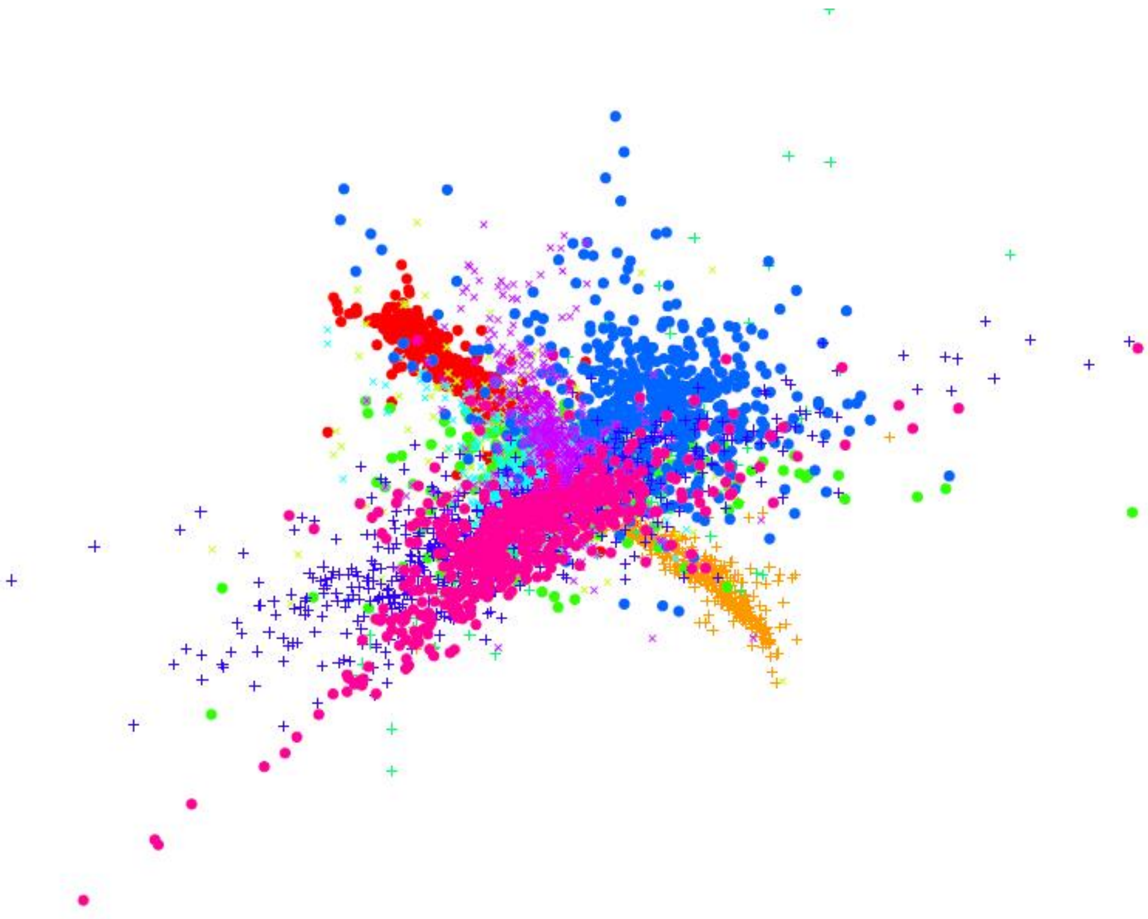
(b) Visualization by Sammon mapping.

• **II. . t-Distributed SNE**



(a) Visualization by Isomap.

- **II. . t-Distributed SNE**

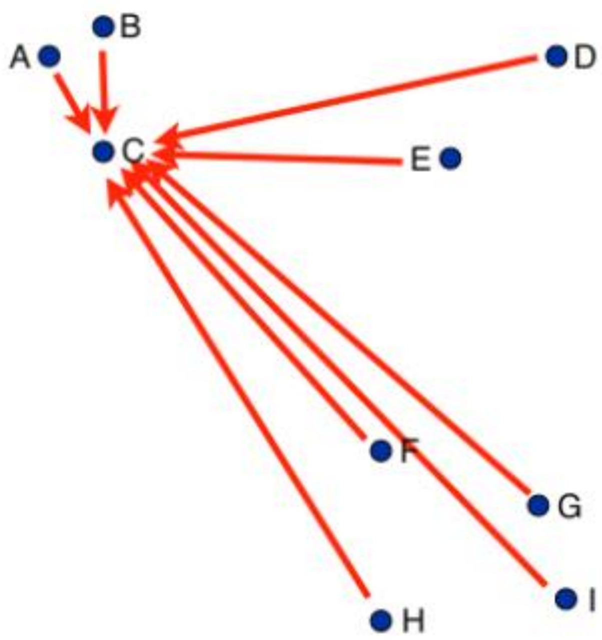


(b) Visualization by LLE.

• III. Accelerating t-SNE using Tree-Based Algorithms

•Gradient interpretation:

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

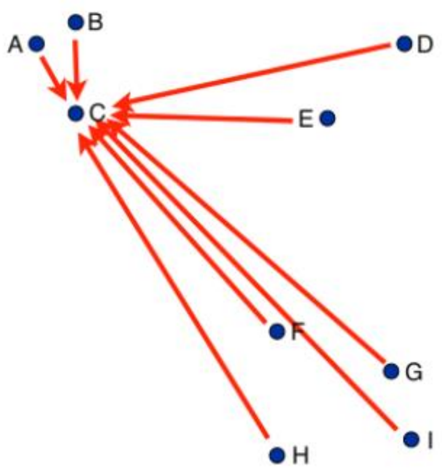


• III. Accelerating t-SNE using Tree-Based Algorithms

•Barnes-Hut Approximation

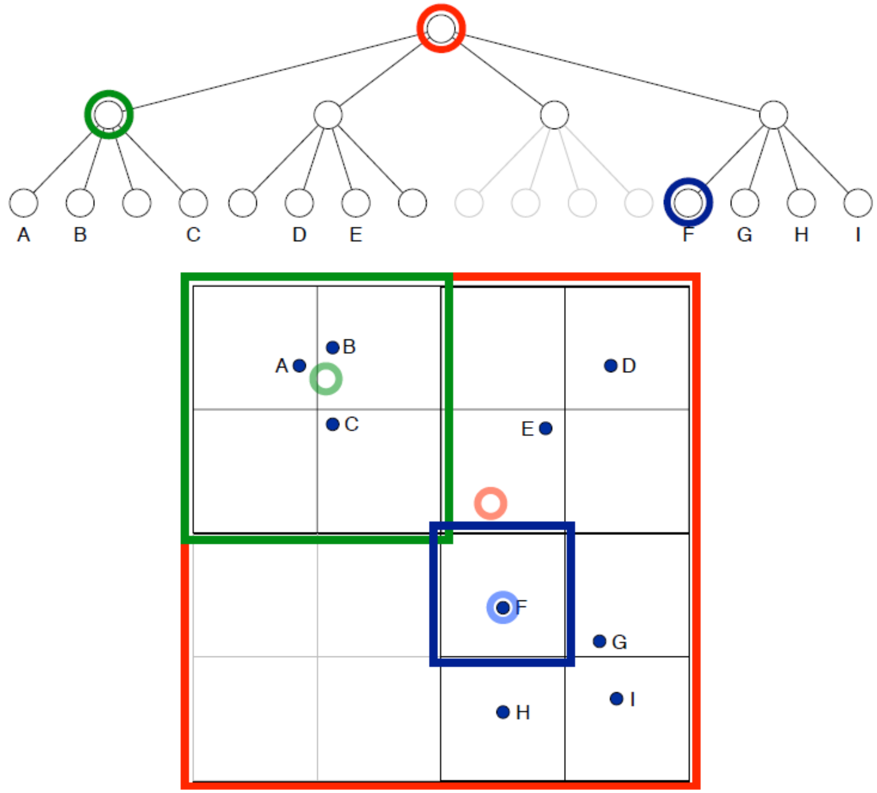
--Many of the pairwise interactions between points are very similar

Consider three points $\mathbf{y}_i$, $\mathbf{y}_j$, and $\mathbf{y}_k$ with $\|\mathbf{y}_i - \mathbf{y}_j\| \approx \|\mathbf{y}_i - \mathbf{y}_k\| \gg \|\mathbf{y}_j - \mathbf{y}_k\|$



• III. Accelerating t-SNE using Tree-Based Algorithms

•Barnes-Hut Approximation



--Summary

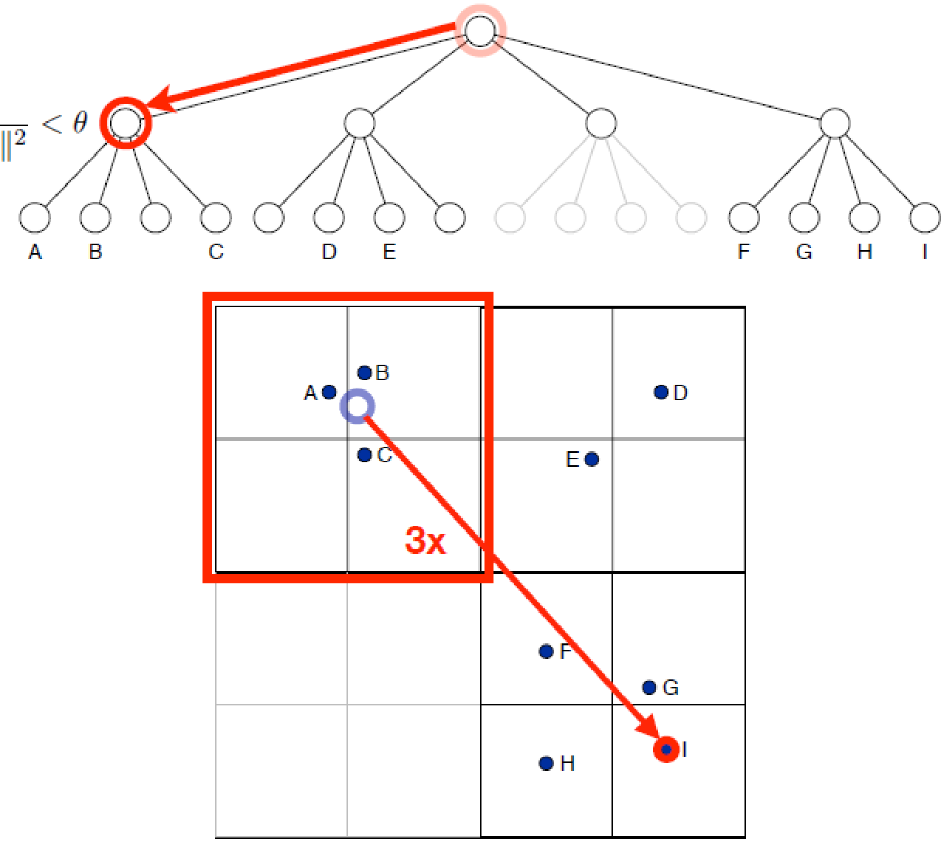
$$\frac{r_{cell}}{\|\mathbf{y}_i - \mathbf{y}_{cell}\|^2} < \theta$$

is a threshold that trades o speed and accuracy

• III. Accelerating t-SNE using Tree-Based Algorithms

•Dual-tree Approximation

$$\frac{r_{cell}}{\|\mathbf{y}_i - \mathbf{y}_{cell}\|^2} < \theta$$



--Summary

$$\frac{\max(r_{cell-A}, r_{cell-B})}{\|\mathbf{y}_{cell-A} - \mathbf{y}_{cell-B}\|^2} < \theta$$

Bibliography

- [1] Visualizing Data using t-SNE
- [2] Accelerating t-SNE using Tree-Based Algorithms

Thanks！