Transformation invariant and outlier revealing dimensionality reduction using triplet embedding

Ehsan Amid
UC Santa Cruz
eamid@ucsc.edu

CMPS 242
December 7, 2017

Joint work with Manfred K. Warmuth

Overview

Outline

Dimensionality Reduction

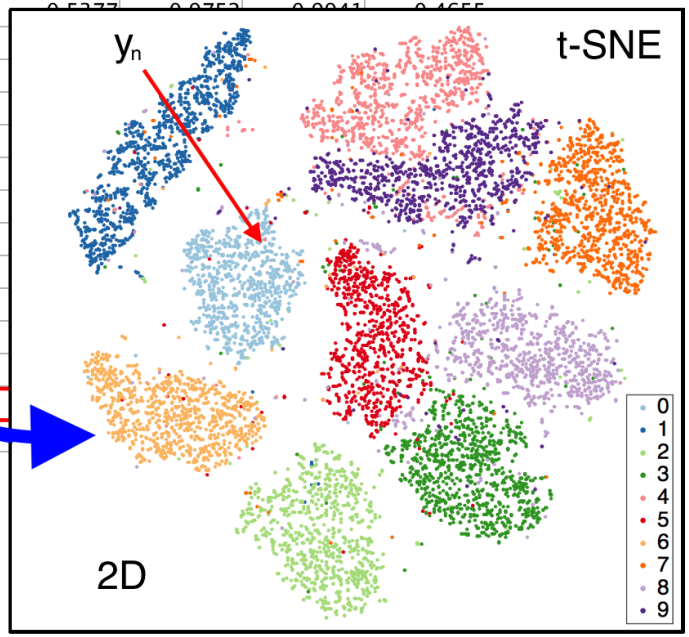
MNIST

	1	2	3	4	5	6	7	8	9	10
1	−0.3719	−1.5272	0.5759	−0.5991	0.9791	−0.5112	−0.2705	0.2778	0.3959	−1.1959
2	1.1788	0.6853	−0.4852	−0.3301	−0.1806	−0.7734	−0.4423	−0.3600	−0.6339	−0.0988
3	−1.8661	0.6259	1.2910	−1.2056	−1.7147	−1.3259	0.5882	0.4728	−0.1525	−0.1700
4	−0.5520	−1.5008	0.5909	0.2450	−0.7648	−0.6206	−0.6121	−1.1448	−0.0695	−0.6950
5	−0.5772	−2.0357	0.1287	1.5396	−0.3895	0.3938				
6	−0.7983	0.2106	0.8516	0.1278	1.1448	−0.8832				
7	−2.7710	0.9666	0.2987	0.3813	−0.0296	0.3866				
8	0.6923	−1.3235	1.0212	−0.2151	0.0815	−0.1155				
9	−3.6301	0.7912	−1.0352	−0.3903	−0.7005	0.1987				
10	0.6871	0.3195	0.3887	−1.8636	0.0575	0.0401				
11	−1.1250	−0.4799	−1.3353	−0.4977	2.2862	0.7745				
12	−1.4807	−0.1793	−2.1890	−0.3219	0.2592	−0.6512				
13	−1.0811	−1.8521	1.3248	−0.3865	−1.2451	0.1368				
14	−0.8135	1.6910	0.5258	−0.7722	0.157	1.0025				
15	0.7151	0.2077	−0.8843	0.3581	−0.4365	0.5765				
16	0.7337	−1.1998	−0.0254	0.3646	0.1651	−0.1967				
17	0.7133	−0.9934	−0.6849	1.3934	0.1603	−0562				
18	−0.3193	1.2468	−0.9969	1.5263	0.6650	−0.892				
19	0.3098	−0.5230	−0.7743	−1.5109	0.3518	0.7477				

196 features

x_n

N points



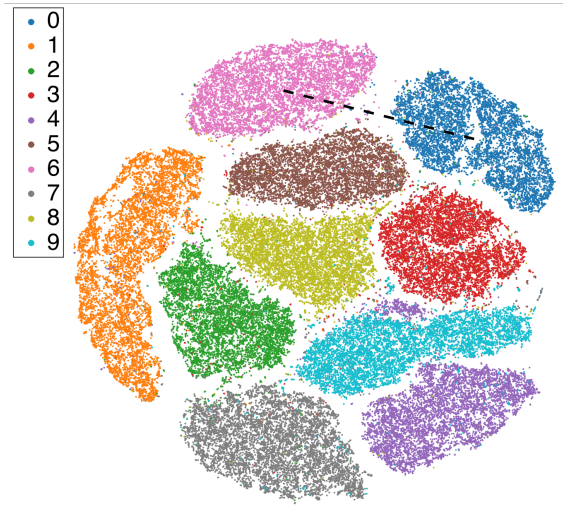
Scatter-plot of a high D data $\rightarrow$ 2D or 3D
t-SNE is the most commonly used method

How to test a DR method on a given natural data set?

- What invariances should be satisfied by a DR method?
- Can I trust a given DR method?
- Can I use it for cluster and outlier detection?

Invariances: subsets (even digits only)

t-SNE

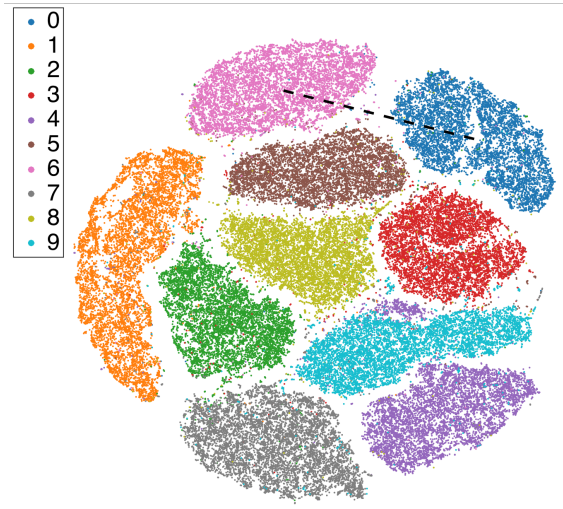


LargeVis

TriMap

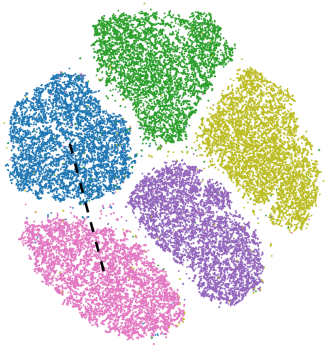
Invariances: subsets (even digits only)

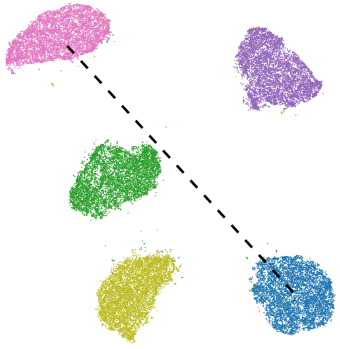
t-SNE

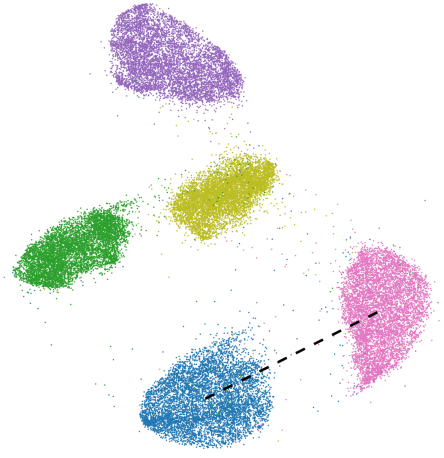


LargeVis

TriMap

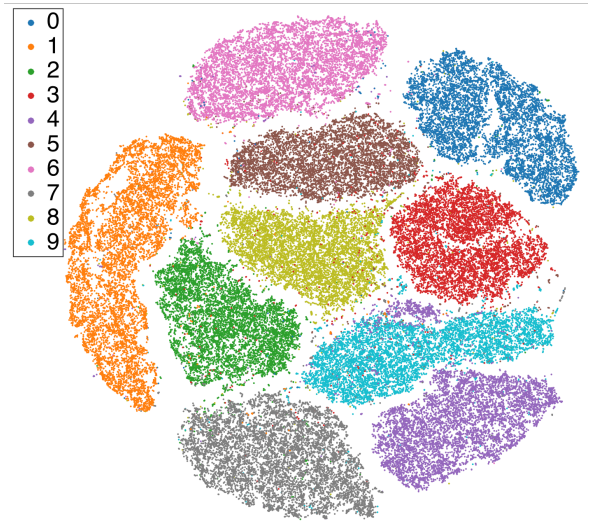






Invariances: sparseness (10% of the dataset)

t-SNE

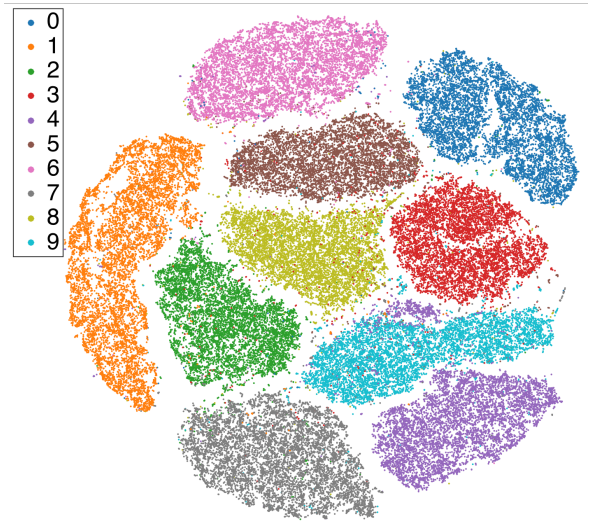


LargeVis

TriMap

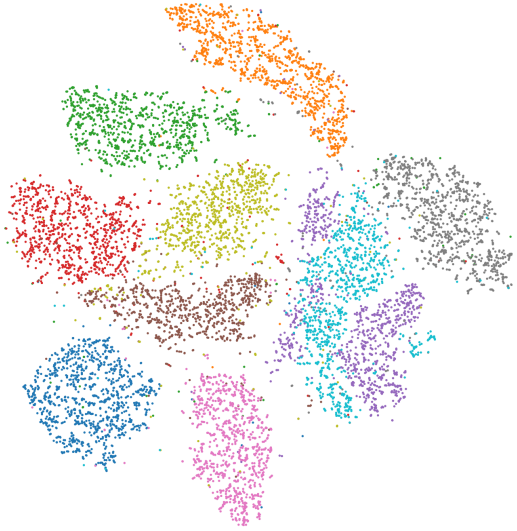
Invariances: sparseness (10% of the dataset)

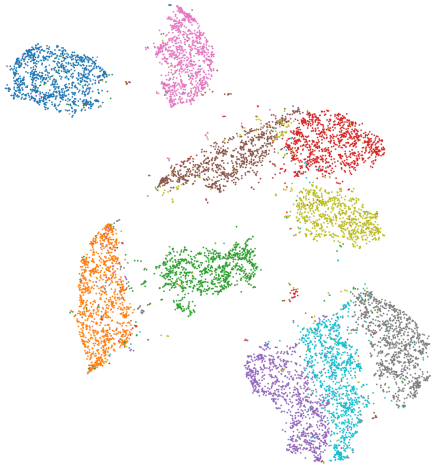
t-SNE



LargeVis

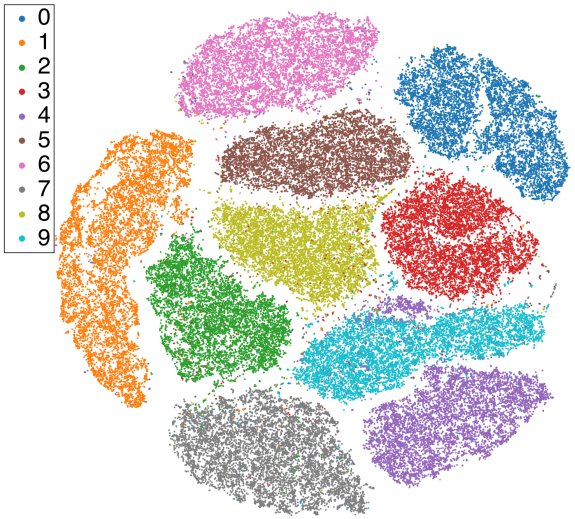
TriMap





Invariances: copies (2 copies shifted apart)

t-SNE

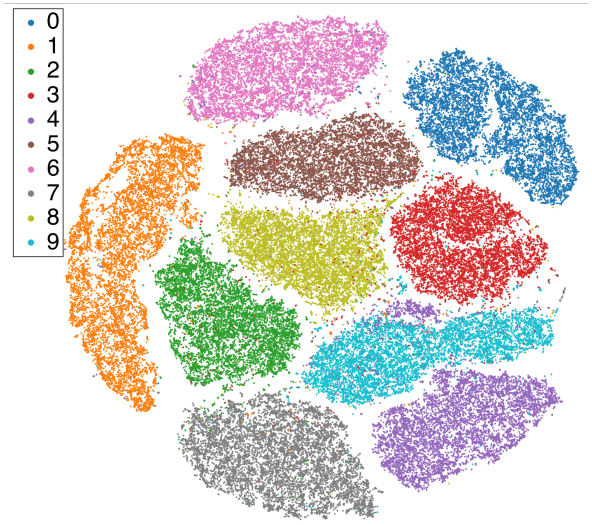


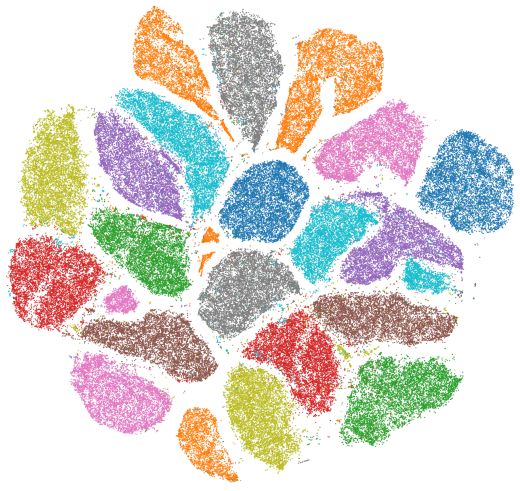
LargeVis

TriMap

Invariances: copies (2 copies shifted apart)

t-SNE



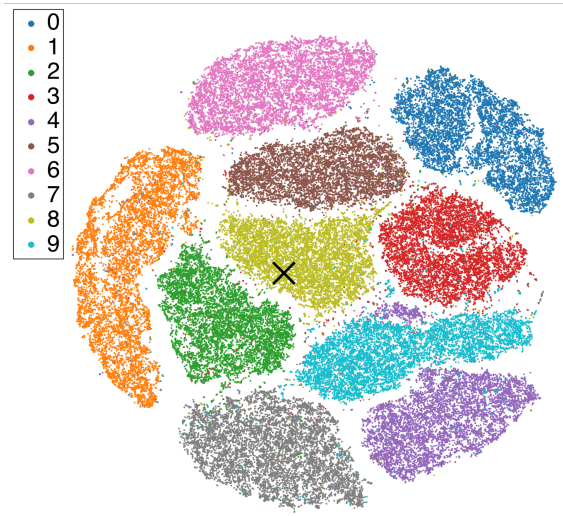


LargeVis

TriMap

Invariances: outliers (point X moved far away)

t-SNE

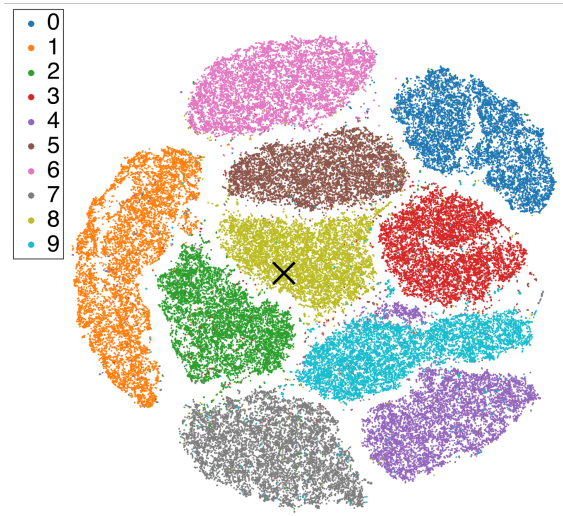


LargeVis

TriMap

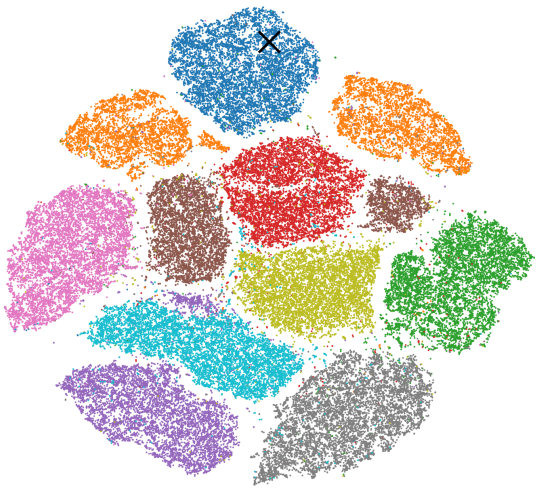
Invariances: outliers (point X moved far away)

t-SNE



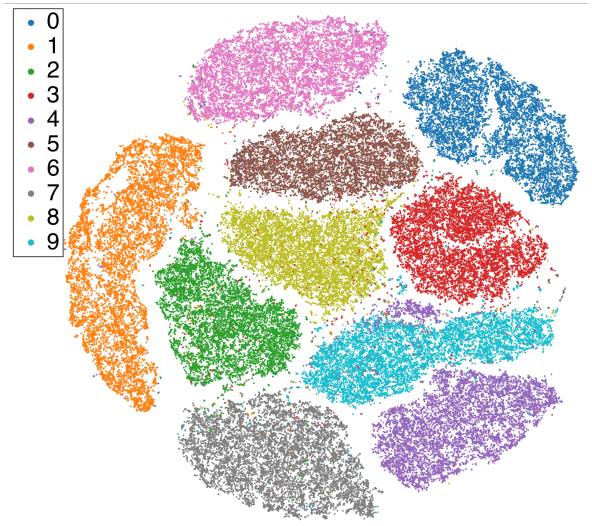
LargeVis

TriMap



Invariances: shifting (one cluster moved far away)

t-SNE

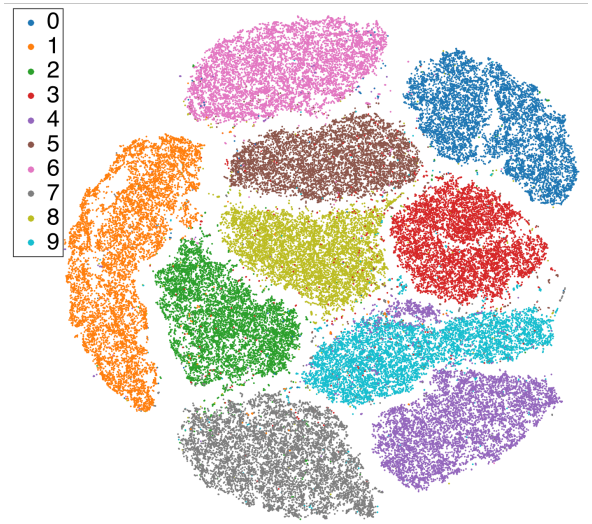


LargeVis

TriMap

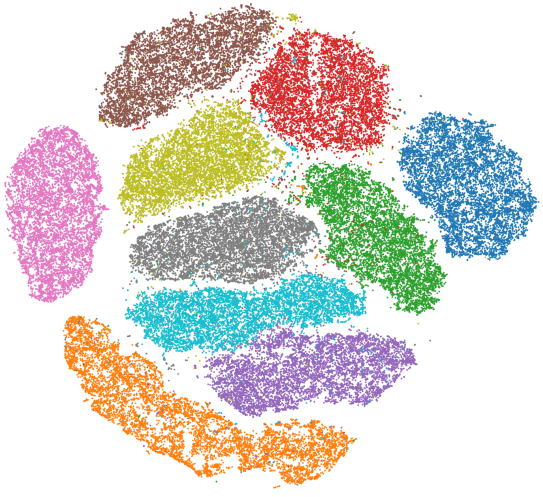
Invariances: shifting (one cluster moved far away)

t-SNE



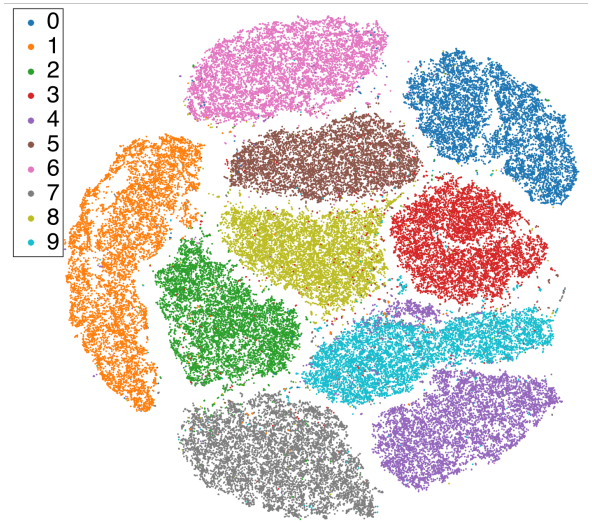
LargeVis

TriMap



Invariances: noise (+ Gaussian noise to 10% of points)

t-SNE

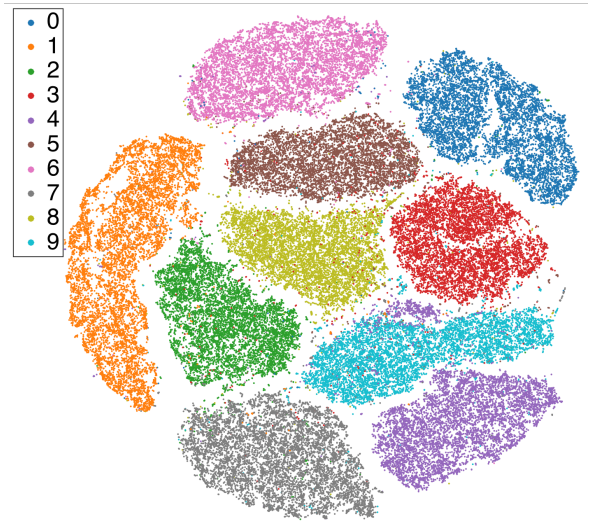


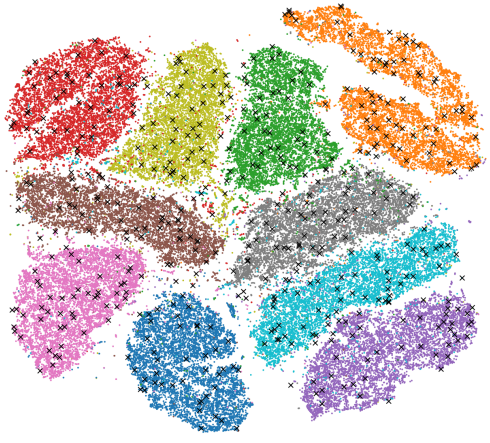
LargeVis

TriMap

Invariances: noise (+ Gaussian noise to 10% of points)

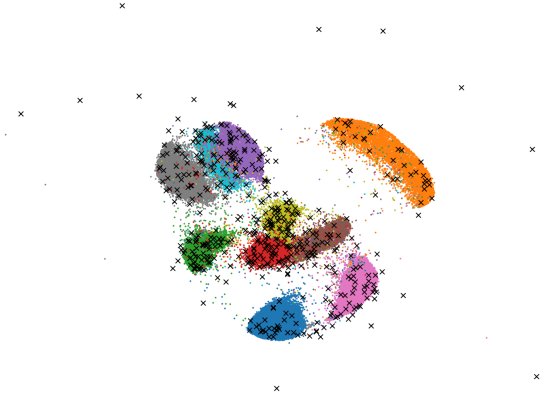
t-SNE





LargeVis

TriMap



Dimensionality Reduction?

Given a representation $\{\mathbf{x}_n\}$ for N items in a high-dimensional space, find a 2 dimensional representation $\{\mathbf{y}_n\}$ of the N items such that the *structure* is preserved as much as possible

Dimensionality Reduction?

Given a representation $\{\mathbf{x}_n\}$ for N items in a high-dimensional space, find a 2 dimensional representation $\{\mathbf{y}_n\}$ of the N items such that the *structure* is preserved as much as possible

Ie preserve

- clusters
- nearest neighbors
- pairwise distances
- low-dimensional manifolds

Dimensionality Reduction?

Given a representation $\{\mathbf{x}_n\}$ for N items in a high-dimensional space, find a 2 dimensional representation $\{\mathbf{y}_n\}$ of the N items such that the *structure* is preserved as much as possible

le preserve

- clusters
- nearest neighbors
- pairwise distances
- low-dimensional manifolds

Can't have everything

Dimensionality Reduction?

Given a representation $\{\mathbf{x}_n\}$ for N items in a high-dimensional space, find a 2 dimensional representation $\{\mathbf{y}_n\}$ of the N items such that the *structure* is preserved as much as possible

Ie preserve

- clusters
- nearest neighbors
- pairwise distances
- low-dimensional manifolds

Can't have everything

What is important?

Dimensionality Reduction?

Given a representation $\{\mathbf{x}_n\}$ for N items in a high-dimensional space, find a 2 dimensional representation $\{\mathbf{y}_n\}$ of the N items such that the *structure* is preserved as much as possible

Ie preserve

- clusters
- nearest neighbors
- pairwise distances
- low-dimensional manifolds

Can't have everything

What is important?

It only has to work for "natural" data sets

Outline

t-SNE Method

1) For each i, form neighborhood probabilities

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/{\sigma_i}^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/{\sigma_i}^2)} \quad \text{for } j \neq i$$

σ_i chosen based on perplexity parameter: $2^{\text{entropy}_i} \approx 30$

2) Symmetrize probabilities

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}, \quad p_{ii} := 0$$

t-SNE method

3) $\mathbf{x}_n$ mapped to $\mathbf{y}_n$ in 2D

Distribution in 2D space uses a t-distribution

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq \ell}(1 + \|\mathbf{y}_\ell - \mathbf{y}_k\|^2)^{-1}} \quad \text{for } j \neq i$$

- t-distribution has heavy tail
- works better than Gaussian

t-SNE method

3) $\mathbf{x}_n$ mapped to $\mathbf{y}_n$ in 2D

Distribution in 2D space uses a t-distribution

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq \ell}(1 + \|\mathbf{y}_\ell - \mathbf{y}_k\|^2)^{-1}} \quad \text{for } j \neq i$$

- t-distribution has heavy tail
- works better than Gaussian

4) Make two distributions similar by minimizing the KL divergence

$$\min_{\{\mathbf{y}_n\}} \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Via GD with momentum

- scale p_{ij} by 4 in first 100 rounds

t-SNE Method

- t-SNE forms nicely separated clusters
- Easy to apply on general datasets

t-SNE Method

- t-SNE forms nicely separated clusters
- Easy to apply on general datasets

- The complexity is $\mathcal{O}(N^2)$
  $\rightarrow$ can be reduced to $\mathcal{O}(N \log N)$
  via tree based approximation [barnshut]

t-SNE Method

- t-SNE forms nicely separated clusters
- Easy to apply on general datasets

- The complexity is $\mathcal{O}(N^2)$
  $\rightarrow$ can be reduced to $\mathcal{O}(N \log N)$
  via tree based approximation [barnshut]

- Everything is collapsed inside an orb
- Outliers become “inliers”
- Extremely misleading!

LargeVis Method

Preserves the **pairwise** similarities

1) For each i, find its k nearest-neighbors and form probabilities similar to t-SNE

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/{\sigma_i}^2)}{\sum_{k\in \text{knn}(i)} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/{\sigma_i}^2)} \quad \text{for } j \in \text{knn}(i)$$

$$\text{else} \quad p_{j|i} = 0$$

2) Define weights

$$w_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}, \quad p_{ii} := 0$$

LargeVis method

3) Maximize the following objective on similar pairs and randomly sampled dissimilar pairs

$$\sum_{(i,j) \text{ similar}} w_{ij} \log f(\|y_i - y_j\|) + \sum_{(i,k) \text{ not similar}} \gamma \log(1 - f(\|y_i - y_k\|))$$

where $f(\|y_i - y_j\|) = \frac{1}{1+a\|y_i - y_j\|^2}$ or $f(\|y_i - y_j\|) = \frac{1}{1+\exp(\|y_i - y_j\|^2)}$ is a similarity function.

Sampling distribution for dissimilar pairs $\propto d_j^{0.75}$ where d_j is the degree of point $j \rightarrow$ pulls the outliers back into the larger clusters!

LargeVis Method

- LargeVis also forms nicely separated clusters
- Scales linearly with n

LargeVis Method

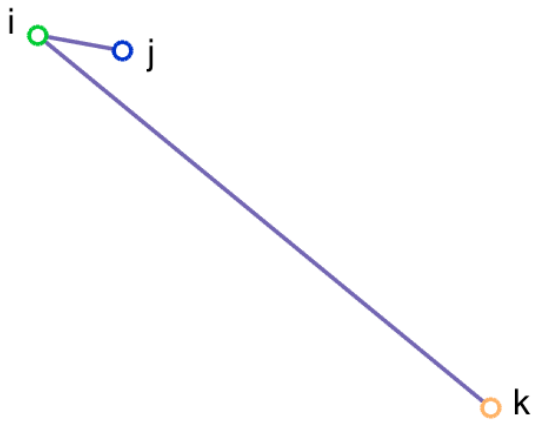
- LargeVis also forms nicely separated clusters
- Scales linearly with n

- Forms over-separated clusters
- Has no clue about the global scale of the data
- Outliers become “inliers” again
- Extremely misleading!

Outline

Solution: higher order of structure

Triplet (i, j, k)**:** "object i is closer to object j than to object k" (used in e.g. crowdsourcing)



Triplet Embedding: Given the set of triplets $\mathcal{T} = \{(i, j, k)\}$, represent $\{\mathbf{x}_n\}$ as $\{\mathbf{y}_n\}$ in 2D such that

$$\text{For each } (i, j, k) \in \mathcal{T} \quad \Longrightarrow \quad \|\mathbf{y}_i - \mathbf{y}_j\| < \|\mathbf{y}_i - \mathbf{y}_k\|, \text{ whp}$$

TriMap

$p_{ij} \geq 0$: pairwise similarity function between x_i and x_j in high-dimension

$q_{ij} \geq 0$: pairwise similarity function between y_i and y_j in low-dimension

$T = \{(i, j, k) : p_{ij} > p_{ik}\}$ set of all triplets

TriMap

$p_{ij} \geq 0$: pairwise similarity function between x_i and x_j in high-dimension

$q_{ij} \geq 0$: pairwise similarity function between y_i and y_j in low-dimension

$T = \{(i, j, k) : p_{ij} > p_{ik}\}$ set of all triplets

Mapping from high-dimension to high-dimension
(if p_{ij} and q_{ij} have identical form)

$$\frac{\left(\frac{p_{ij}}{p_{ik}}\right)}{\left(\frac{q_{ij}}{q_{ik}}\right)} = \frac{p_{ij}}{p_{ik}} \cdot \frac{q_{ik}}{q_{ij}} \to 1, \quad \forall (i, j, k) \in T$$

TriMap

$p_{ij} \geq 0$: pairwise similarity function between x_i and x_j in high-dimension

$q_{ij} \geq 0$: pairwise similarity function between y_i and y_j in low-dimension

$T = \{(i, j, k) : p_{ij} > p_{ik}\}$ set of all triplets

Mapping from high-dimension to high-dimension
(if p_{ij} and q_{ij} have identical form)

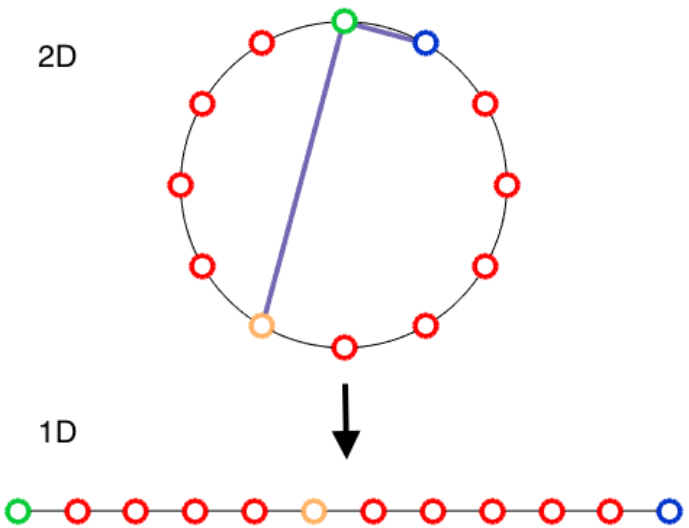
$$\frac{\left(\frac{p_{ij}}{p_{ik}}\right)}{\left(\frac{q_{ij}}{q_{ik}}\right)} = \frac{p_{ij}}{p_{ik}} \cdot \frac{q_{ik}}{q_{ij}} \to 1, \quad \forall (i, j, k) \in T$$

Mapping from high-dimension to low-dimension is different!

TriMap

In low-dimension:

- Less volume than high-dimension to place points
- Less degrees of freedom: not all the triplets can be satisfied



TriMap

To accommodate for lower volume:

- let the ratios go to 0 instead of 1!

$$\frac{\left(\frac{p_{ij}}{p_{ik}}\right)}{\left(\frac{q_{ij}}{q_{ik}}\right)} = \frac{p_{ij}}{p_{ik}} \cdot \frac{q_{ik}}{q_{ij}} \to 0, \quad \forall (i,j,k) \in T$$

TriMap

To accommodate for lower volume:

- let the ratios go to 0 instead of 1!

$$\frac{\left(\frac{p_{ij}}{p_{ik}}\right)}{\left(\frac{q_{ij}}{q_{ik}}\right)} = \frac{p_{ij}}{p_{ik}} \cdot \frac{q_{ik}}{q_{ij}} \to 0, \quad \forall (i,j,k) \in T$$

- use heavier-tail for q_{ij} (t-distribution) than p_{ij} (Gaussian)

TriMap

1) Set

$$p_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{\sigma_{ij}^2})$$

where $\sigma_{ij}^2 = \sigma_i \sigma_j$ and σ_i is average distance of i from its 10-th to 20-th NN

2) Define

$$q_{ij} = (1 + \frac{\|y_i - y_j\|^2}{\alpha})^{-\frac{1+\alpha}{2}}$$

α degrees of freedom

3) Minimize

$$L = \sum_{(i,j,k) \in T} \frac{p_{ij}}{p_{ik}} \cdot \frac{q_{ik}}{q_{ij}}$$

TriMap

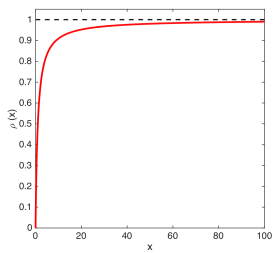
For unsatisfied triplets:

$$\left.\begin{array}{l} q_{ij} \text{ is small} \\ q_{ik} \text{ is large} \end{array}\right\} \quad \frac{q_{ik}}{q_{ij}} \text{ is large}$$

To accommodate for unsatisfied triplets:

- apply the robust transformation $\rho(\ell) = 1 - \frac{1}{1+\ell}$ on the ratios

notice that $\rho(\ell) \in [0, 1]$ for $\ell \geq 0$



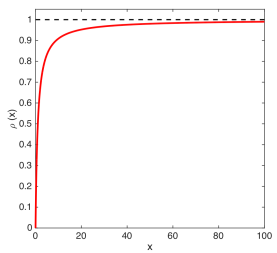
TriMap

For unsatisfied triplets:

$$\left.\begin{array}{l} q_{ij} \text{ is small} \\ q_{ik} \text{ is large} \end{array}\right\} \quad \frac{q_{ik}}{q_{ij}} \text{ is large}$$

To accommodate for unsatisfied triplets:

- apply the robust transformation $\rho(\ell) = 1 - \frac{1}{1+\ell}$ on the ratios



notice that $\rho(\ell) \in [0, 1]$ for $\ell \geq 0$

3) Minimize

$$L = \sum_{(i,j,k) \in T} \frac{p_{ij}}{p_{ik}} \cdot \rho\left(\frac{q_{ik}}{q_{ij}}\right)$$

TriMap

The method works well using all $\mathcal{O}(n^3)$ triplets for n points

Only feasible for toy datasets because $\mathcal{O}(n^3)$ triplets

Key observation: most triplets are redundant: e.g. (i, j, k) and (i, j, k') where k and k' are NN

Use a subset of triplets!

TriMap

The method works well using all $\mathcal{O}(n^3)$ triplets for n points

Only feasible for toy datasets because $\mathcal{O}(n^3)$ triplets

Key observation: most triplets are redundant: e.g. (i, j, k) and (i, j, k') where k and k' are NN

Use a subset of triplets!

Which triplets to use?

Use the triplets that have high p_{ij} value (i.e., KNN), thus a higher $\frac{p_{ij}}{p_{ik}}$ on average

Sampling process for triplets

$\rightarrow$ For each point i

$\rightarrow$ For each point j from m-nearest neighbors

$\rightarrow$ m' times: sample k randomly further away from i than j

Form triplet (i, j, k)

Sampling process for triplets

$\rightarrow$ For each point i

$\rightarrow$ For each point j from m-nearest neighbors

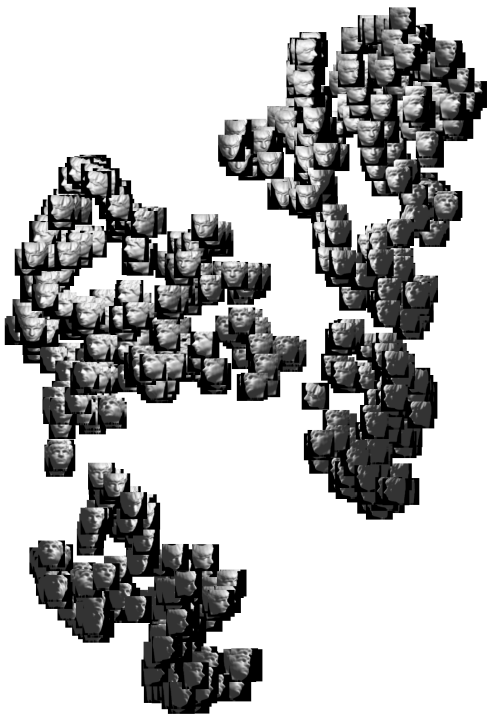
$\rightarrow$ m' times: sample k randomly further away from i than j

Form triplet (i, j, k)

- $(N\, m\, m')$ triplets in total
- If $m, m' \ll N \rightarrow$ then complexity is $\sim O(N)$
  We use $m = 50$, $m' = 10$
- Also add small number of randomly selected triplets

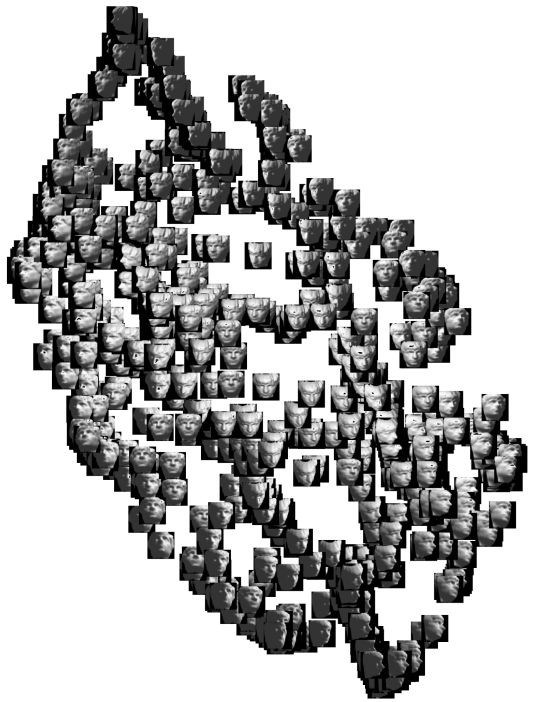
Some DR Results

t-SNE



LargeVis

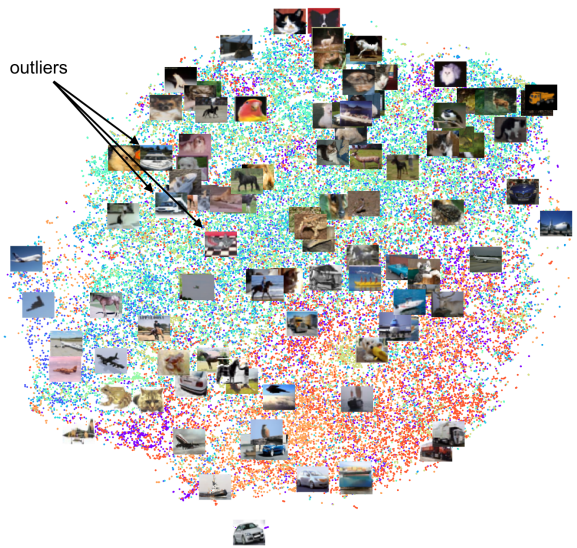
TriMap



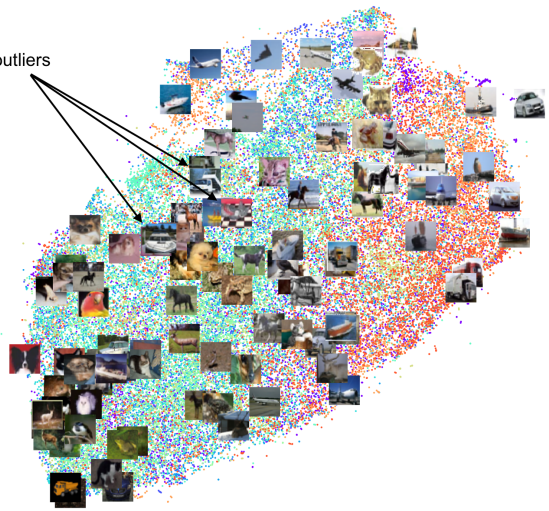
Faces dataset: the underlying manifold is recovered with TriMap ($n = 698$, $d = 4096$)

Some DR Results

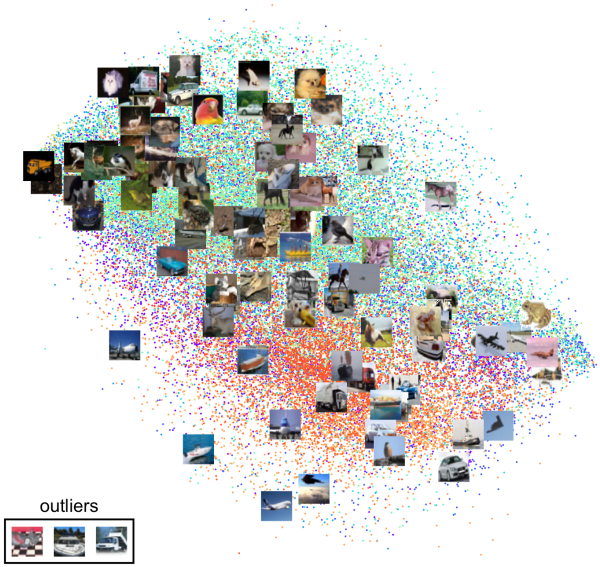
t-SNE



LargeVis



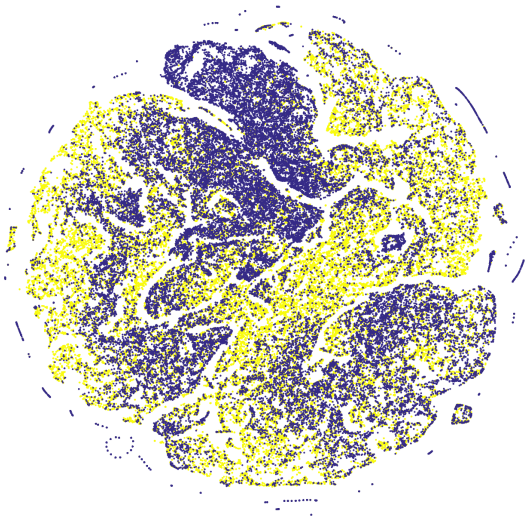
TriMap



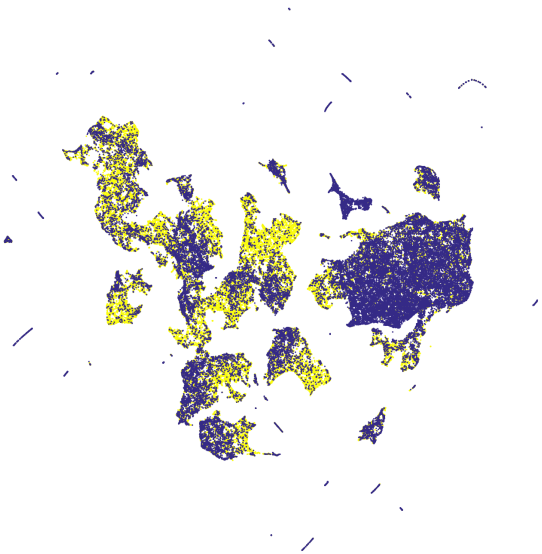
Cifar10 dataset: the ouliers detected by TriMap are shown on the bottom left corner of the figure (not the actual location). Both t-SNE and LargeVis place these outliers among other images (shown with arrows) ($n = 60{,}000$, $d = 1024 \times 3$)

Some DR Results

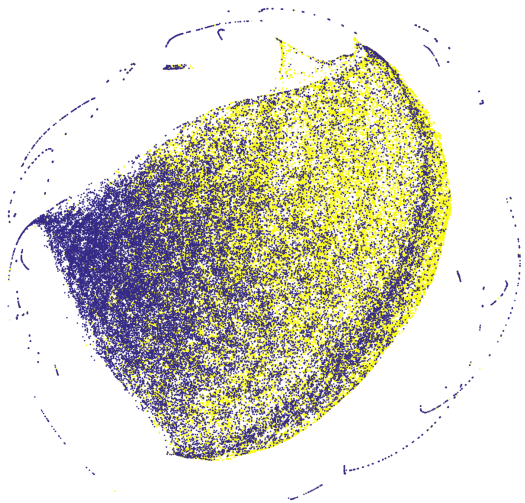
t-SNE



LargeVis



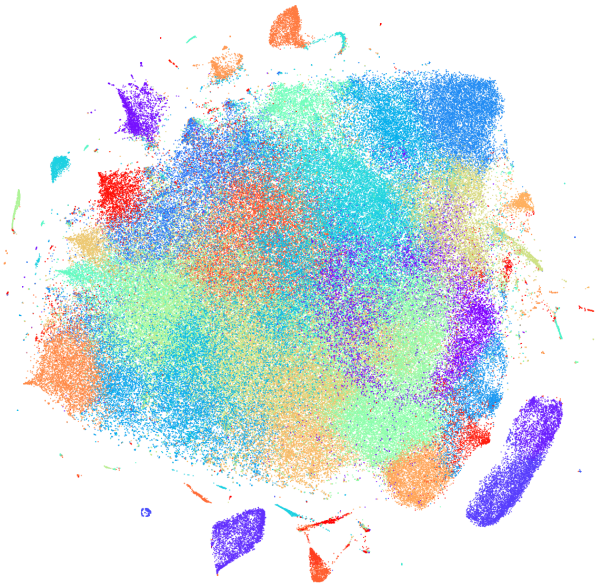
TriMap



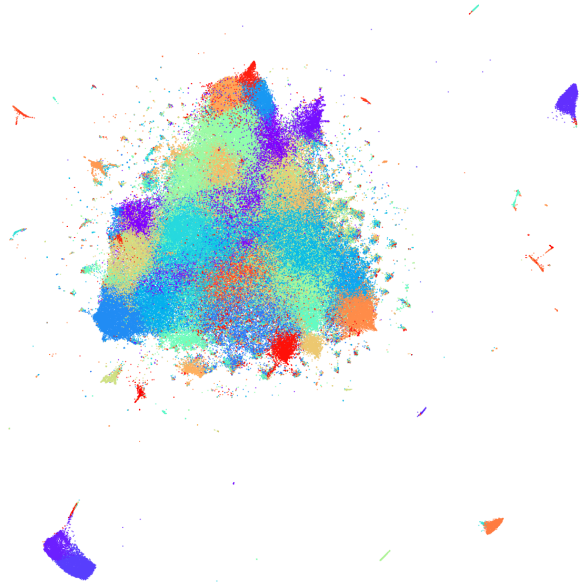
TV News: commercial (yellow), non-commercial (blue) broadcasts. TriMap finds a smooth map while the other two methods split the data into smaller patches ($n = 129{,}685$, $d = 50$)

Some DR Results

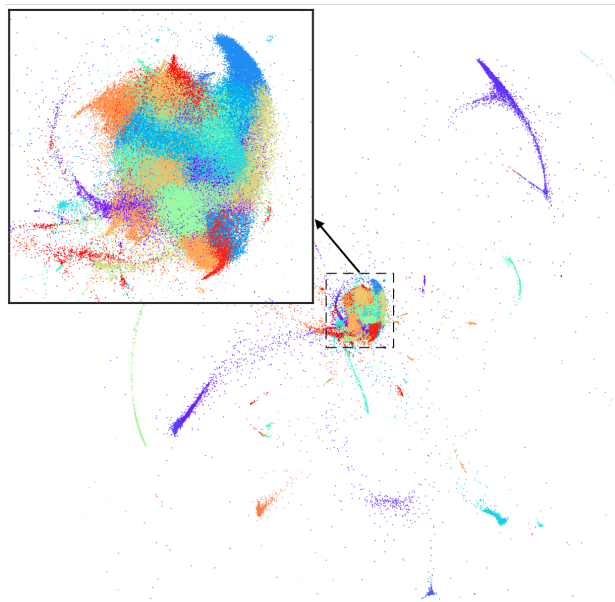
t-SNE



LargeVis



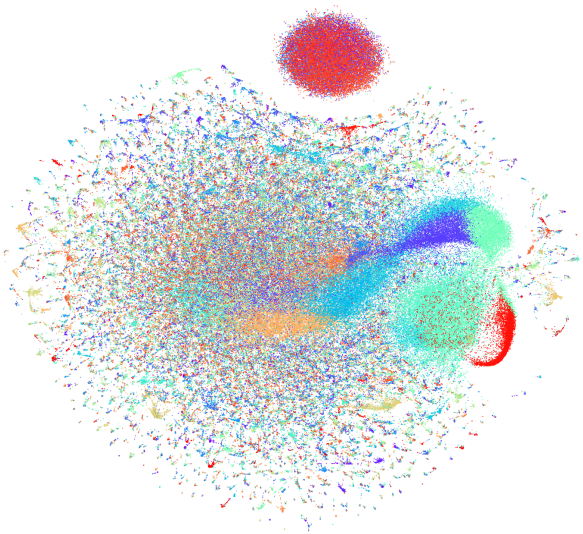
TriMap



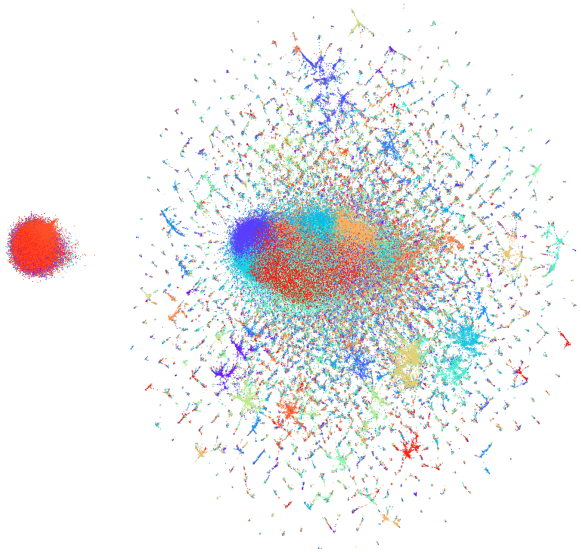
380k+ lyrics: for each lyric, we average over the representation of the words in that lyric. TriMap recovers multiple scales and outliers. The larger cluster found by both t-SNE and LargeVis is shown in the middle ($n = 266{,}557$, $d = 256$)

Some DR Results

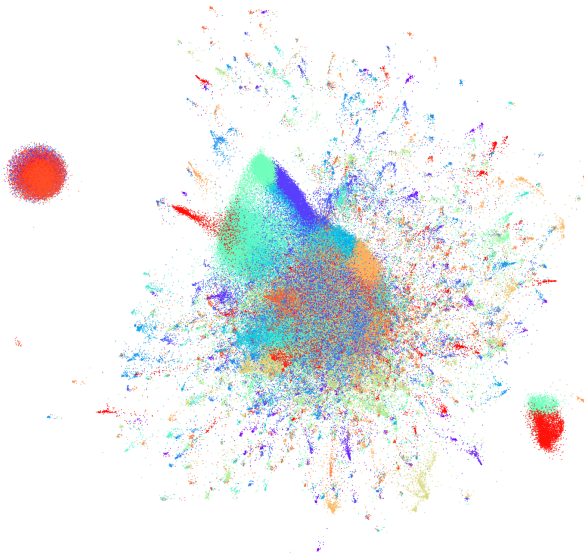
t-SNE



LargeVis



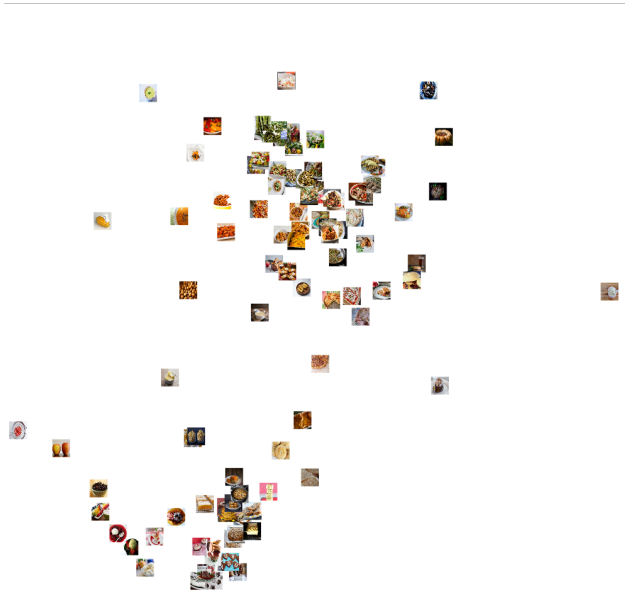
TriMap



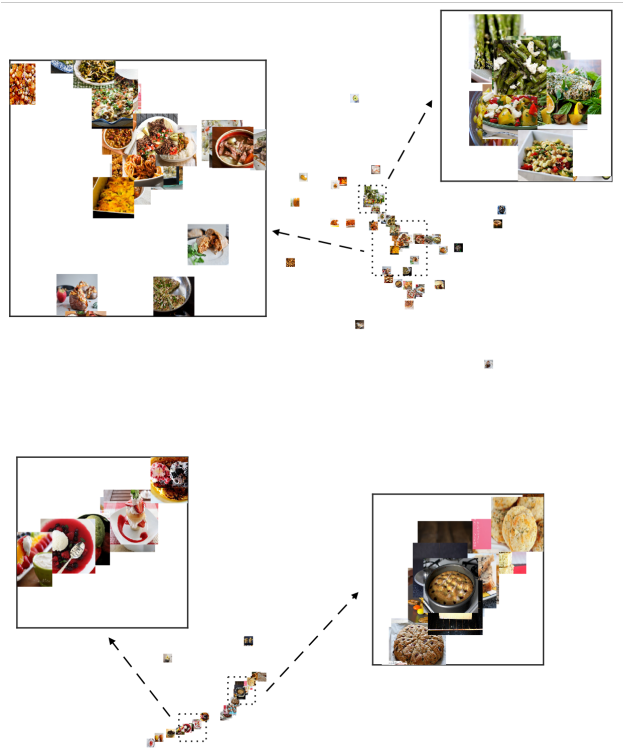
DBLP network: collaboration network of authors. Both LargeVis and TriMap produce compelling results. However, LargeVis tends to groups the outliers together into small clusters ($n = 317{,}080$, $d = 256$)

Triplet Embedding Results

t-STE

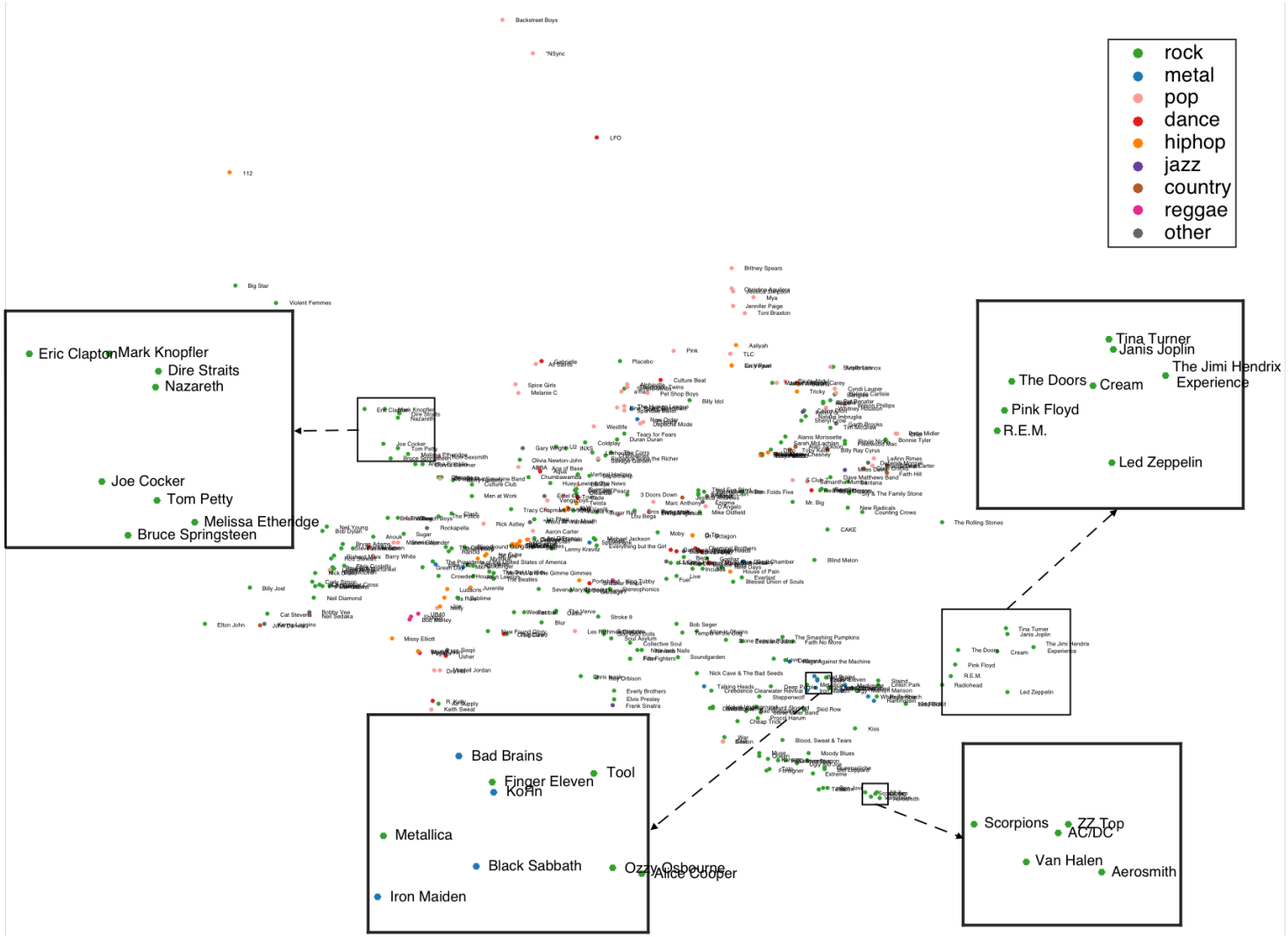


TriMap



Food dataset: no clear separation between the clusters in the t-STE. TriMap recovers three different clusters of food: “Vegetables and Meals” (top), “Ice creams and Deserts” (bottom left), and “Breads and Cookies” (bottom right)
($n = 100$, $|T| = 190{,}376$)

Triplet Embedding Results



Music dataset: similar artists are placed together ($n = 400$, $|T| = 9107$)

Conclusion

1. We can greatly improve DR methods based on triplets
2. & smart heuristic for sub sampling a linear number of triplets

3. What is the best natural set of invariances?
4. How can we create similar natural data sets for testing a DR method

What next?

1. Zooming
2. Movies
3. Speedup
4. Visualization