Making low-dimensional embeddings more informative

Team: Linear Men

Team members:

Makhin Artem
Kuznetsov Mikhail
Mumladze Maximelian
Zelentsov Aleksei

Skoltech

Motivation

Low-dimensional embeddings can be extremely useful thing
(for example, for analysis)

One of the current state of the art methods (according to [1]) for making low-dimensional embeddings is tSNE.

But tSNE fails to model many of the important properties of the initial dataset, which can be useful for the researcher.

The purpose of our work is to **implement** and **investigate** improvement methods, **experimental proof** of the preservation of some properties of the dataset after low-dimensional transformation

Skoltech

[1] https://arxiv.org/pdf/2103.01828.pdf

Related works

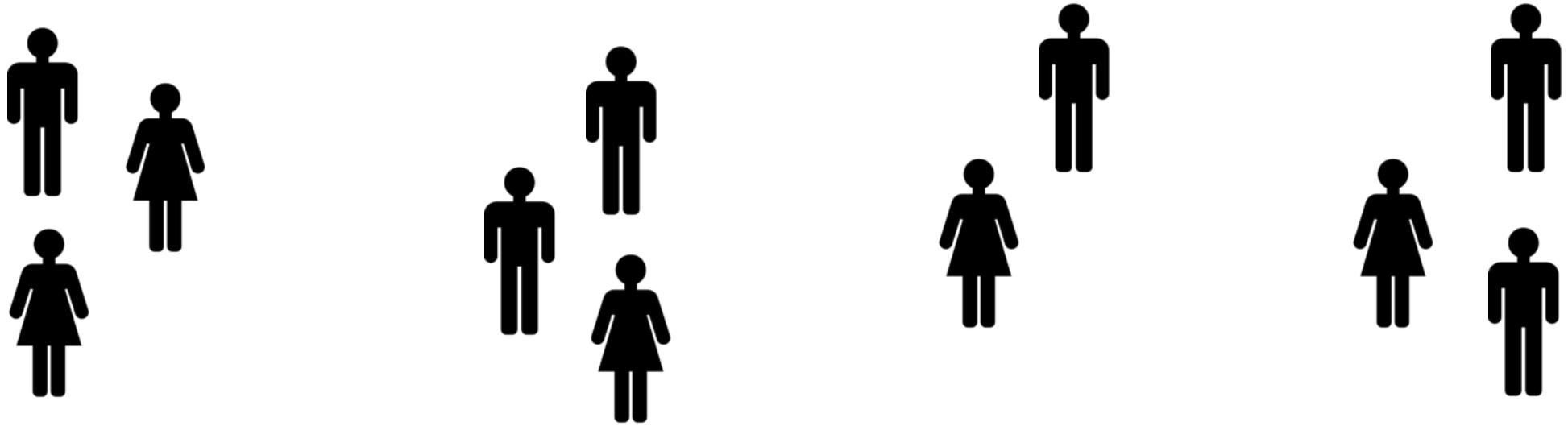
There are some works with modifications of t-SNE method by van der Maaten & Hinton (2008).

Names of this methods:

- JEDI and CONFETTI
- dtSNE
- ctSNE

Problem #1

If we have prior knowledge, we can exhibit meaningful low-lever structure.
This is implemented by JEDI

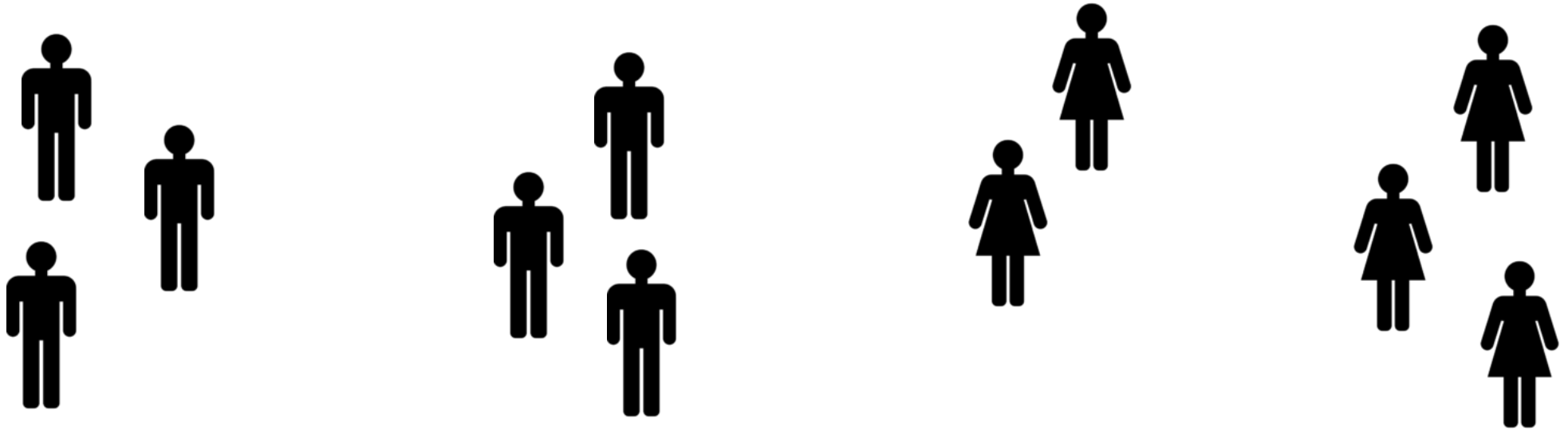


JEDI is useful both for unstructured and structured data analysis, provided we have prior knowledge

Skoltech

Problem #1

If we have prior knowledge, we can exhibit meaningful low-lever structure.
This is implemented by JEDI



JEDI is useful both for unstructured and structured data analysis, provided we have prior knowledge

Skoltech

JEDI

Vanilla t-SNE:

$$\arg\min_{Y} D_{\mathrm{KL}}(P \,||\, Q)$$

JEDI:

$$\arg\min_{Y} D_{\mathrm{KL}}(P \,||\, Q) - \mathrm{JS}^{\alpha}_{\beta}(P' \,||\, Q)$$

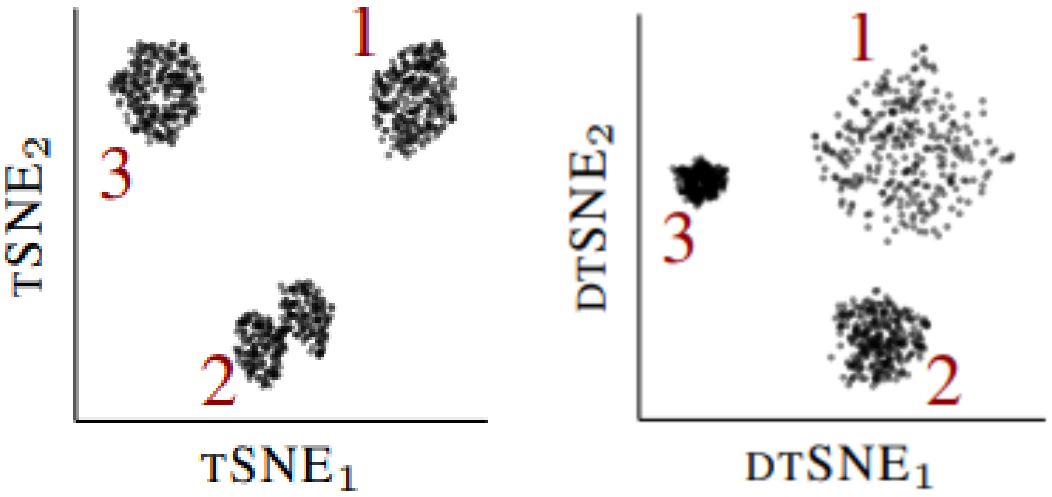
Parameterized Jensen Shannon Divergence:

$$\mathrm{JS}^{\alpha}_{\beta}(P' \,||\, Q) = \alpha D_{\mathrm{KL}}(P' \,||\, \beta Q + (1-\beta)P') + (1-\alpha) D_{\mathrm{KL}}(Q \,||\, \beta P' + (1-\beta)Q)$$

Implementation by Makhin Artem

Problem #2

Cluster sizes and densities in the embedding do not model those of high-dimensional data. **This can be solved by dtSNE**



dtSNE Preserving local densities in low-dimensional embeddings

dtSNE can help distinguish cell types while looking at relative differences in local densities in biological datasets

dtSNE

We want to map the distances of close neighbors of points in differently dense regions in **X** to regions in **Y** that show a similar relative difference in scale.
So, a scaling factor is defined:

$$\gamma_{ij} = \frac{\left((\sigma_i + \sigma_j)^2\right)^{-1}}{\max_{k,l}\left((\sigma_k + \sigma_l)^2\right)^{-1}} \qquad \sigma_{ij}^2 = \left(1/2(\sigma_i + \sigma_j)\right)^2$$

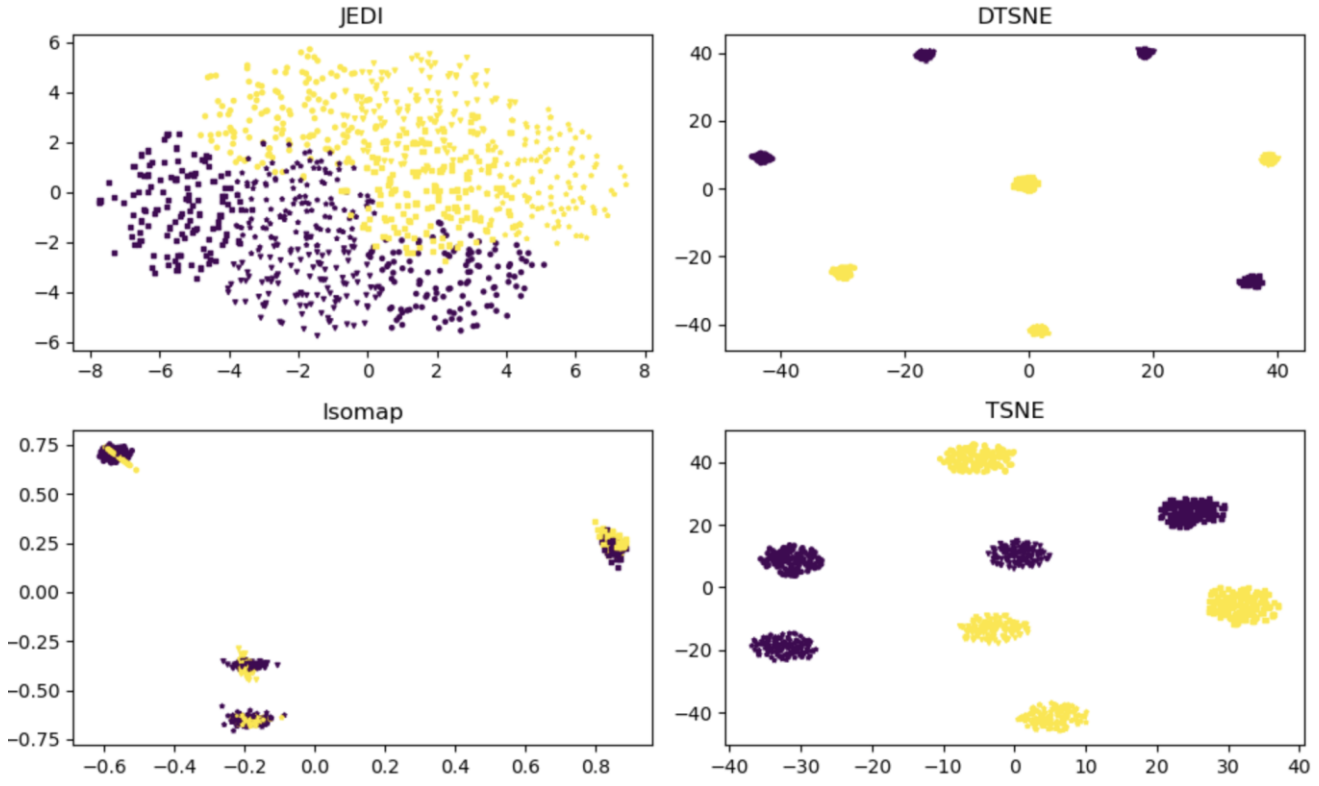
t-SNE	dt-SNE
$p_{j\|i} = \frac{\exp(-\|x_i - x_j\|_2^2 / (2\sigma_i^2))}{\sum_{k \neq i} \exp(-\|x_i - x_k\|_2^2 / (2\sigma_i^2))}$	$p_{j\|i} = \frac{\exp(-\|x_i - x_j\|_2^2 / (2\sigma_{ij}^2))}{\sum_{k \neq i} \exp(-\|x_i - x_k\|_2^2 / (2\sigma_{ik}^2))}$
$q_{ij} = \frac{(1 + \|x_i - x_j\|_2^2)^{-1}}{\sum_{k \neq l}(1 + \|x_k - x_l\|_2^2)^{-1}}$	$q_{ij} = \frac{(1 + \gamma_{ij}\|x_i - x_j\|_2^2)^{-1}}{\sum_{k \neq l}(1 + \gamma_{kl}\|x_k - x_l\|_2^2)^{-1}}$

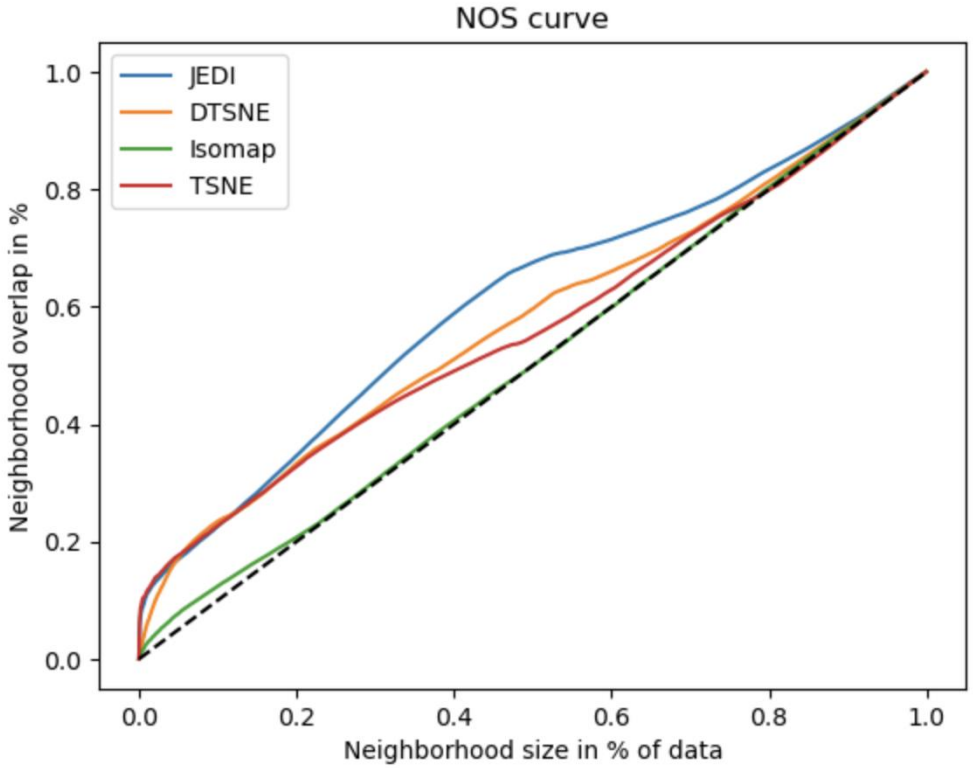
Skoltech

Implemented by Kuznetsov Mikhail

Experiments, JEDI

Dataset with 4 clusters in first four dimensions and 2 clusters in 5-6 dimensions.
Information from first dimensions is prior knowledge



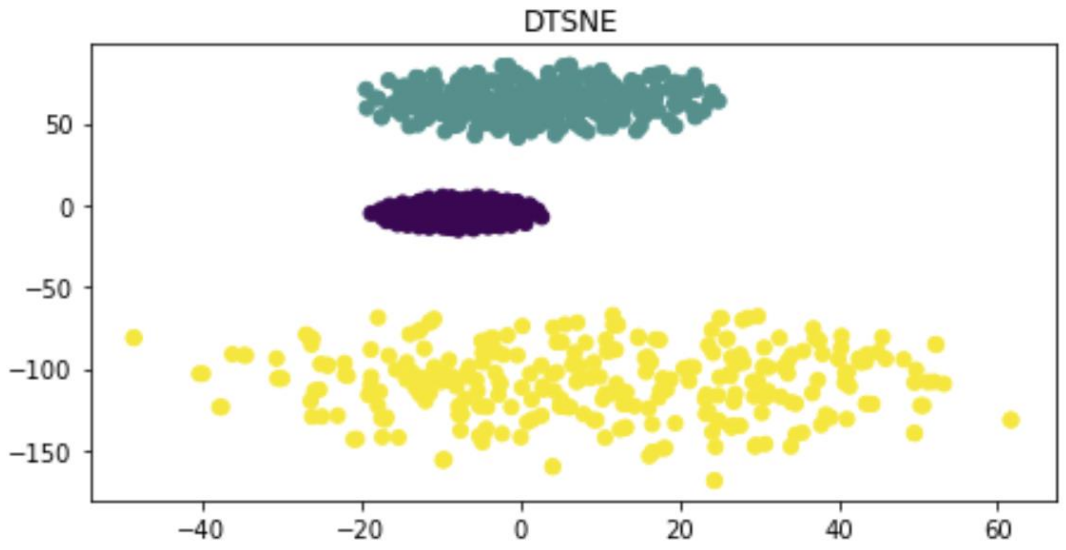


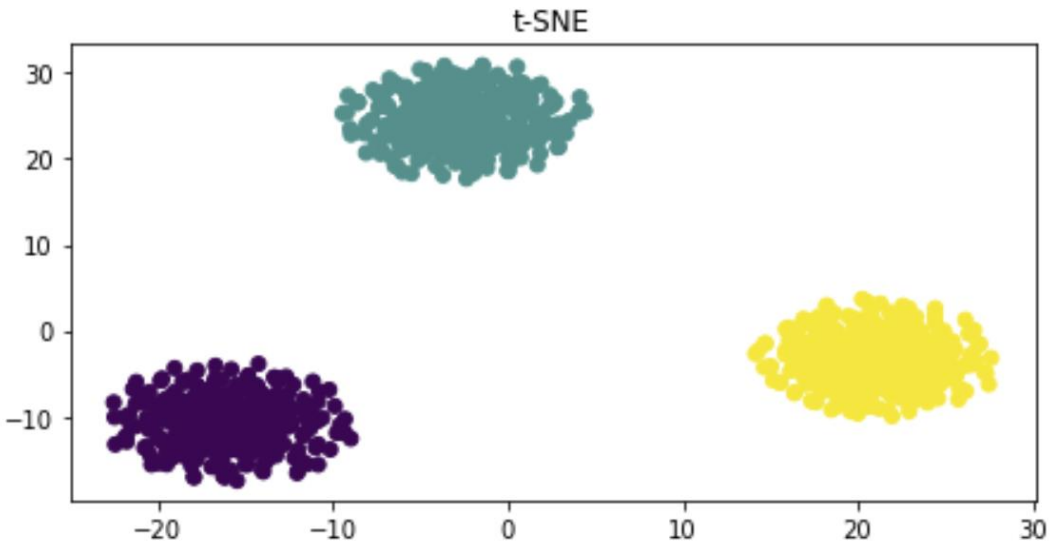
Skoltech

Done by Zelentsov Aleksei

Experiments, dtSNE, #1

50 dims, 3 Gaussian clusters, 3×300 points
The spread is scaled by 2, 4, 8 respectively



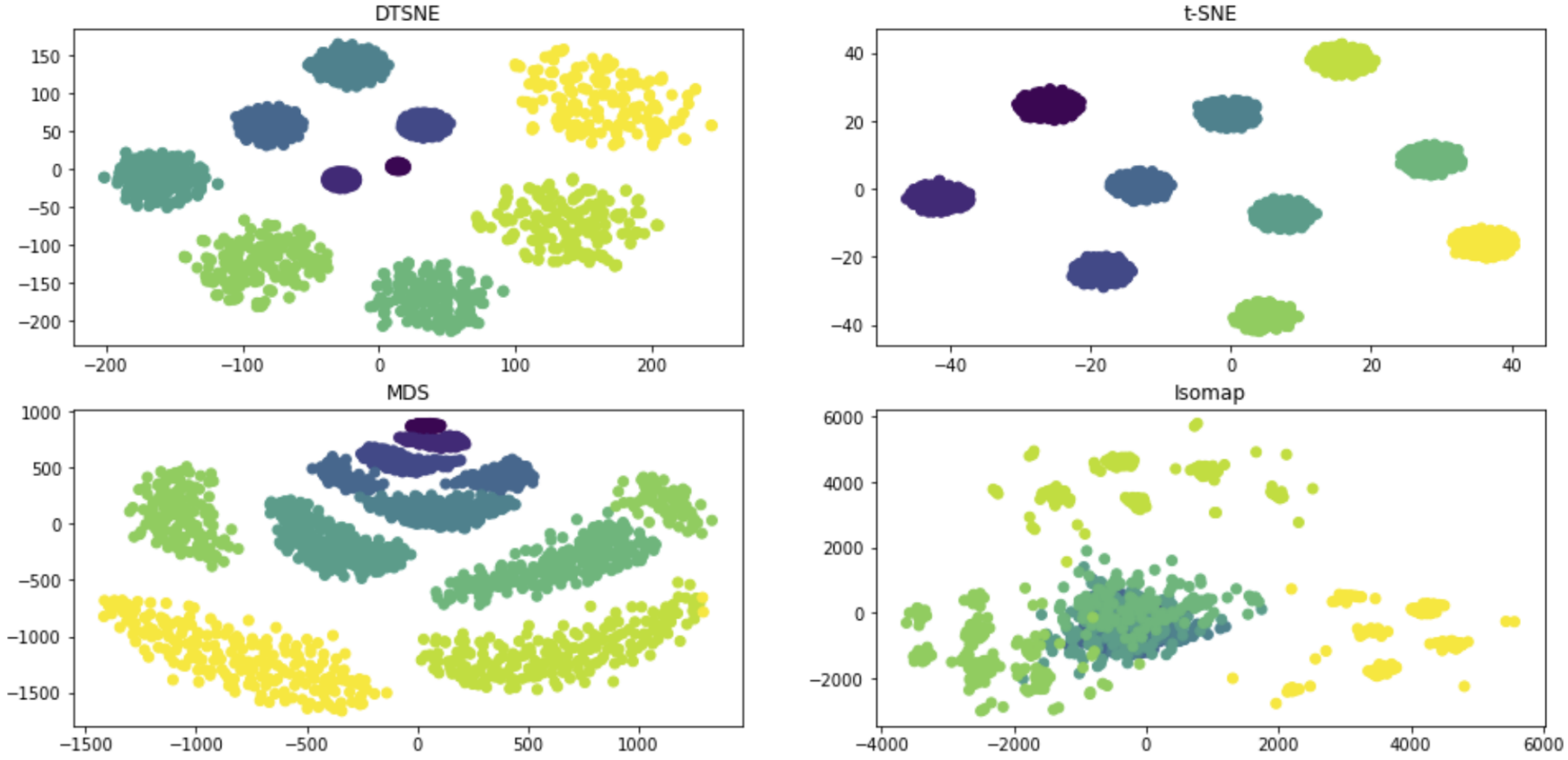


Skoltech

Done by Mumladze Maximelian

Experiments, dtSNE, #2

150 dims, 10 Gaussian clusters, 10×200 points
The spread is scaled by range(1, 11) respectively

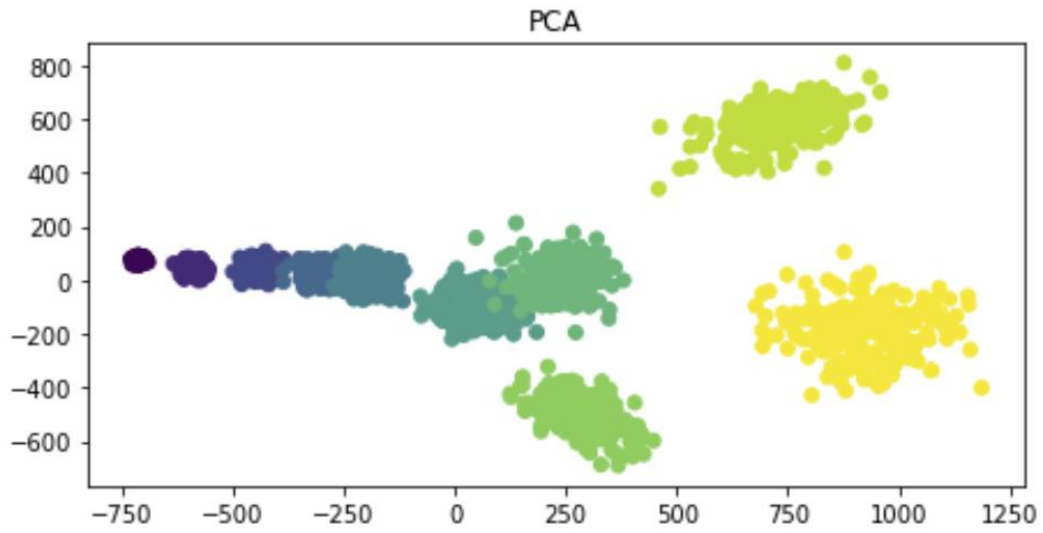


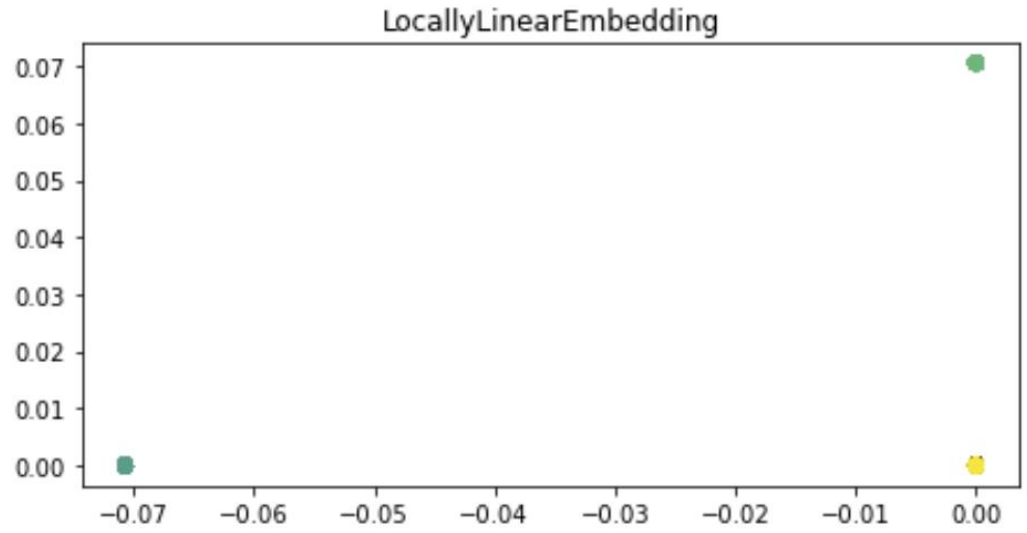
Skoltech

Done by Mumladze Maximelian

Experiments, dtSNE, #2

150 dims, 10 Gaussian clusters, 10×200 points
The spread is scaled by range(1, 11) respectively





Skoltech

Done by Mumladze Maximelian

Experiments, dtSNE, #2, Metrics

- ***global corr*** – spearman rank correlation between all high- and low-dimensional distances
- ***local corr*** – correlation between high- and low-dimensional distances of each point with its 100 closest neighbors
- ***rel reconstr*** - correlation between radii of balls enclosing the 100 neighbors of each point in high- and low-dimensional space

model	global corr	local corr	rel reconstr
DTSNE	0.590	**0.855**	**0.957**
t-SNE	0.530	0.101	0.038
MDS	**0.876**	0.825	0.956
Isomap	0.868	0.731	0.885
PCA	0.868	0.832	0.937
LocallyLinearEmbedding	-0.078	0.128	0.068

Skoltech

Done by Mumladze Maximelian

Conclusion

- We showed improved low-dimensional embeddings of high-dimensional data using two modifications **tSNE**: **JEDI** and **dtSNE**.

- Through our experiments, we demonstrated that **JEDI** is useful for structured and unstructured data analysis when prior knowledge is available, and **dtSNE** is effective in preserving cluster sizes and densities in the embedding.

- We compared our results with other well-known low-dimensional embedding techniques, such as **tSNE**, **MDS**, **ISOMAP**, **PCA**, and **LocallyLinearEmbedding**, and showed that **JEDI** and **dtSNE** outperform these techniques in certain metrics.