A New Technique for Dimension Reduction for Data Visualization

Cynthia Rudin

Duke University

Joint work with Yingfan Wang, Haiyang Huang, and Yaron Shaposhnik

[PDF] Visualizing data using **t-SNE**.

L Van der Maaten, G Hinton - Journal of machine learning research, 2008 - jmlr.org

… **t-SNE** is better than existing techniques at creating a single … large data sets, we show how **t-SNE** can use random walks on … We illustrate the performance of **t-SNE** on a wide variety of …

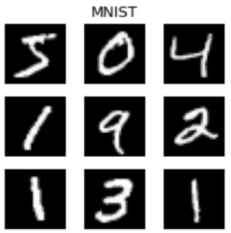
☆ Save 99 Cite Cited by 25211 Related articles All 47 versions »

t-SNE is a dimension reduction algorithm.

Input: high-dimensional data

Output: low-dimensional data that preserves…

- the graph structure?
- local neighborhoods?
- global structure?



t-SNE on MNIST

[PDF] Visualizing data using **t-SNE**.

L Van der Maaten, G Hinton - Journal of machine learning research, 2008 - jmlr.org

… **t-SNE** is better than existing techniques at creating a single … large data sets, we show how **t-SNE** can use random walks on … We illustrate the performance of **t-SNE** on a wide variety of …

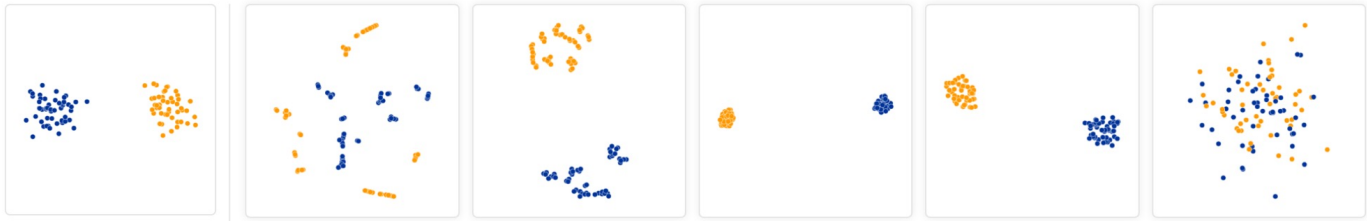
☆ Save 99 Cite Cited by 25211 Related articles All 47 versions »

How to Use t-SNE Effectively

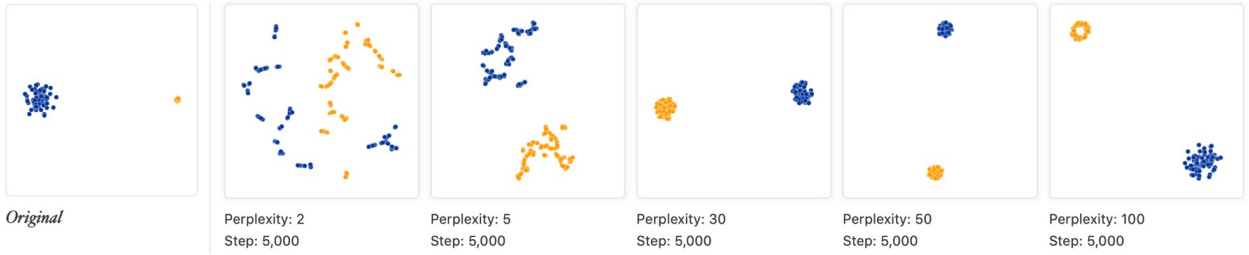
MARTIN WATTENBERG, Google Brain
FERNANDA VIÉGAS, Google Brain
IAN JOHNSON, Google Cloud

Oct. 13, 2016

1. Those hyperparameters really matter

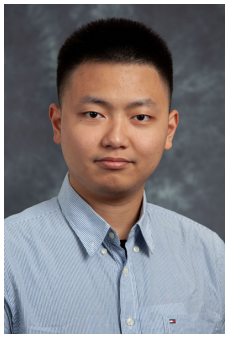


2. Cluster sizes in a t-SNE plot mean nothing



SAMSI Interpretable Deep Learning Working Group

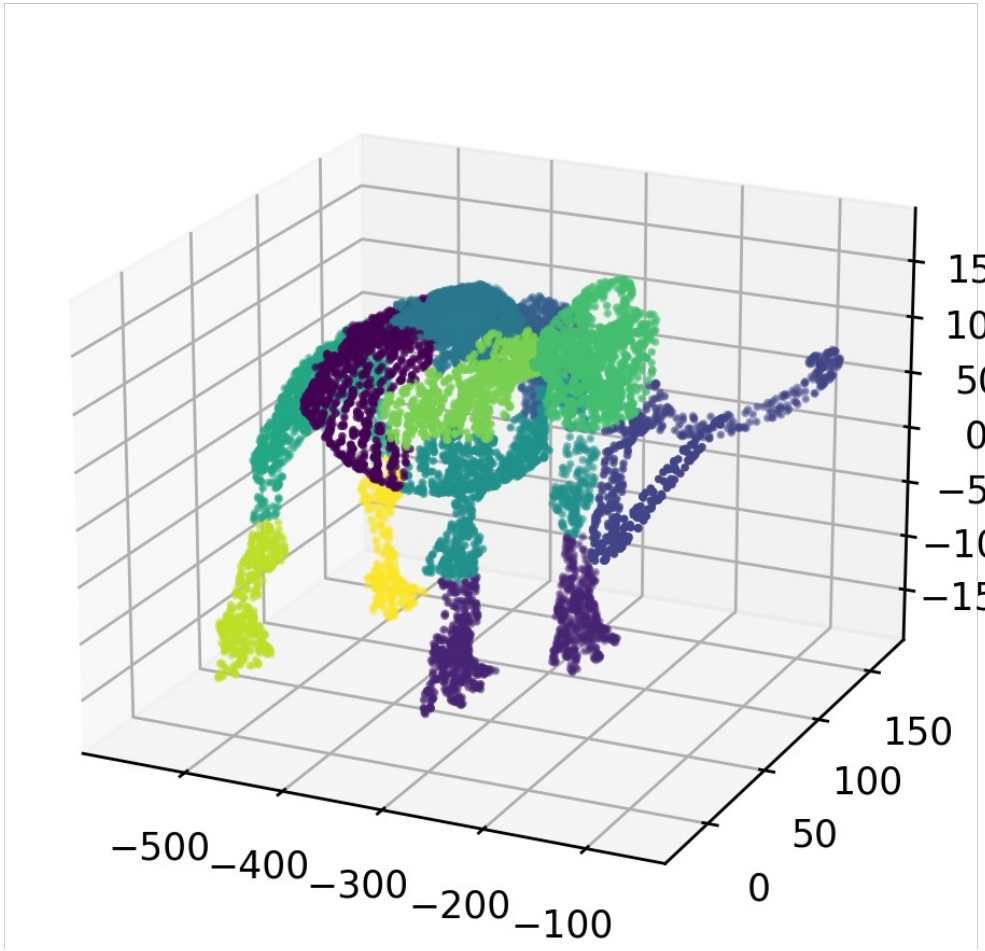
Yingfan Wang
PhD student, Duke



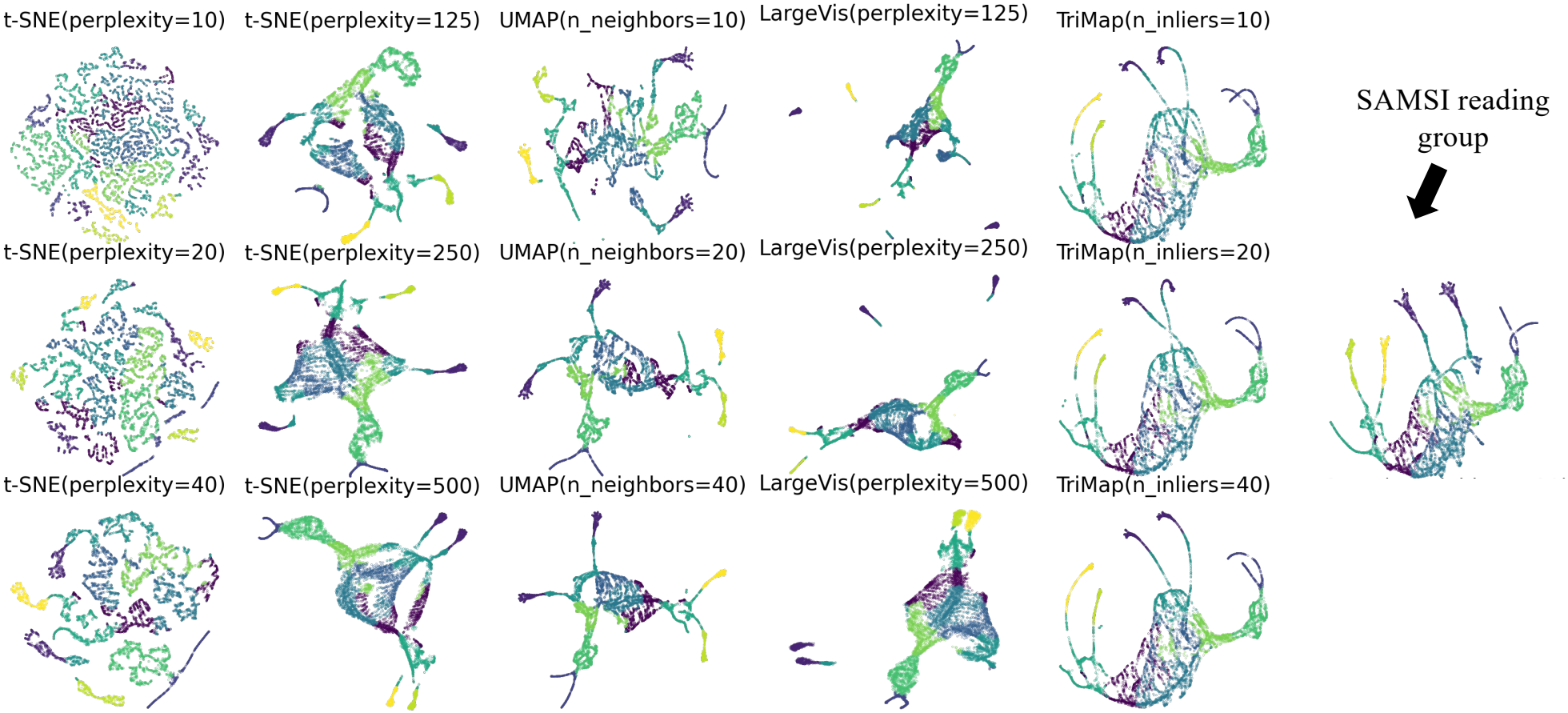
Haiyang Huang
PhD student, Duke

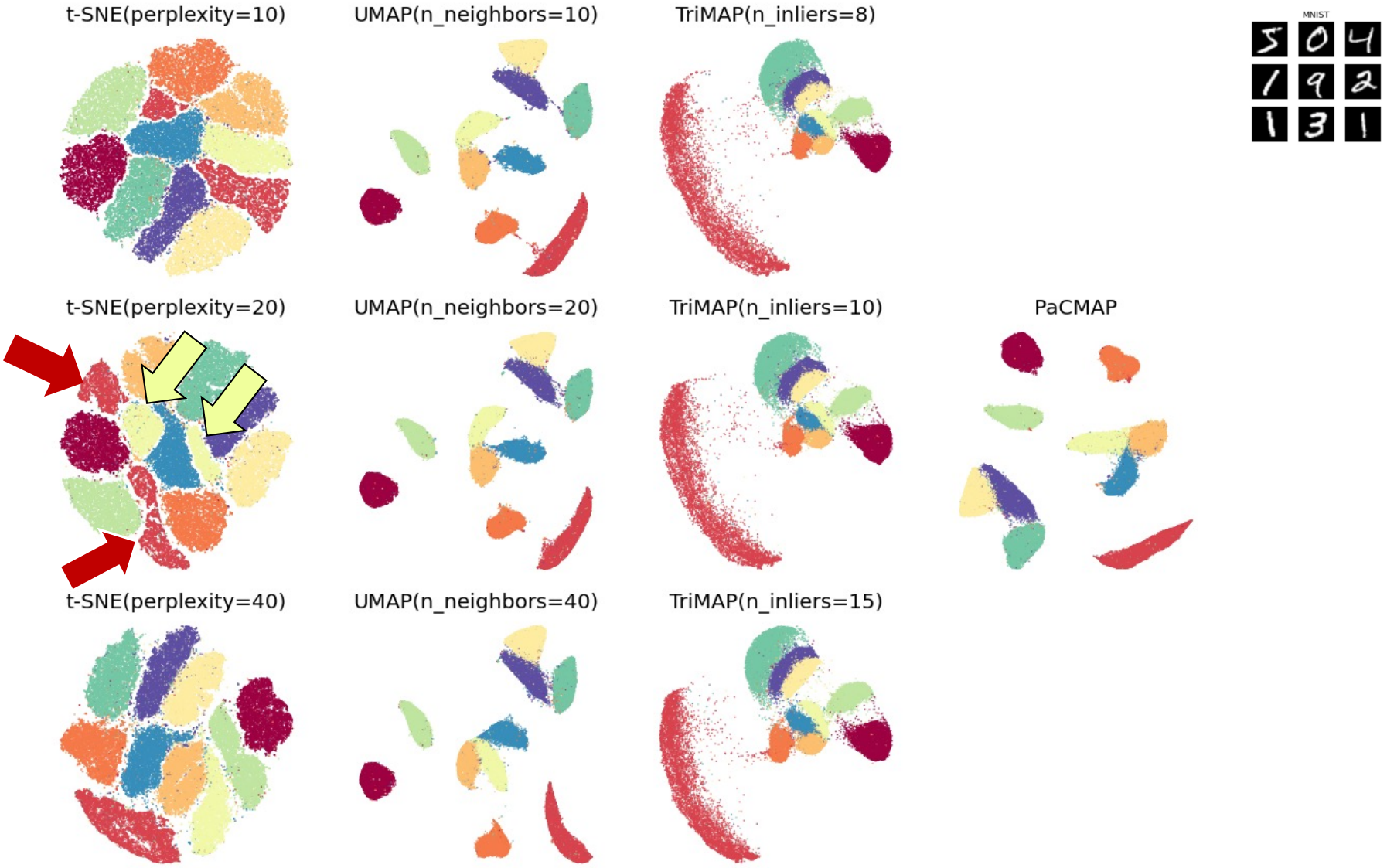
Yaron Shaposhnik
Asst. Prof., U Rochester

Original Mammoth



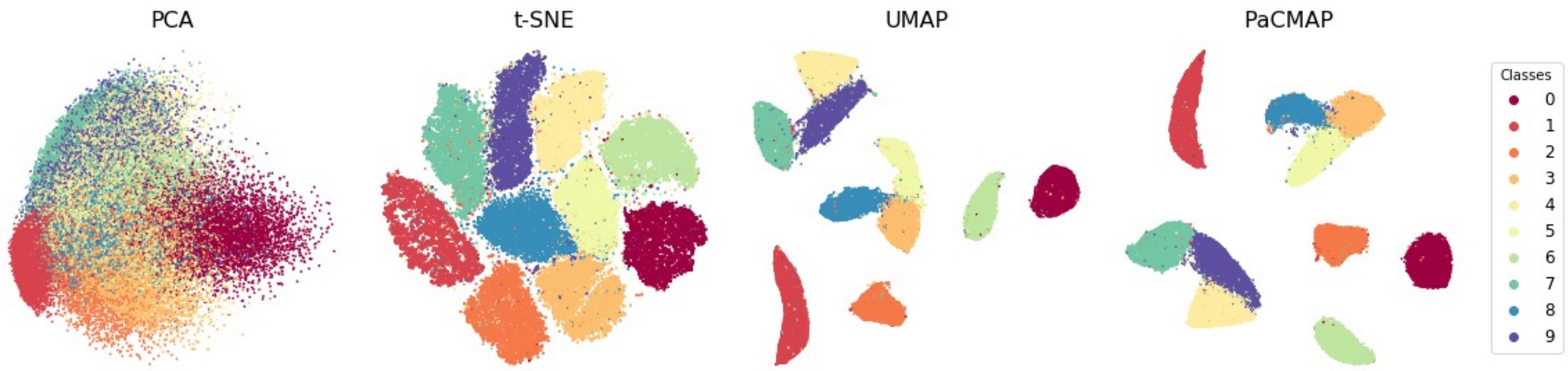
Task: 3d to 2d.


t-SNE(perplexity=10)
t-SNE(perplexity=125)
UMAP(n_neighbors=10)
LargeVis(perplexity=125)
TriMap(n_inliers=10)
SAMSI reading group
t-SNE(perplexity=20)
t-SNE(perplexity=250)
UMAP(n_neighbors=20)
LargeVis(perplexity=250)
TriMap(n_inliers=20)
t-SNE(perplexity=40)
t-SNE(perplexity=500)
UMAP(n_neighbors=40)
LargeVis(perplexity=500)
TriMap(n_inliers=40)


t-SNE(perplexity=10)
UMAP(n_neighbors=10)
TriMAP(n_inliers=8)
MNIST
t-SNE(perplexity=20)
UMAP(n_neighbors=20)
TriMAP(n_inliers=10)
PaCMAP
t-SNE(perplexity=40)
UMAP(n_neighbors=40)
TriMAP(n_inliers=15)

Local vs Global

- Local structure: local neighborhood graph, nearest neighbors
- Global structure: relationships between clusters, respect relative distances between points in high-dimensional space.



(mainly global) (mainly local) (mainly local) (both, actually)

Global Methods

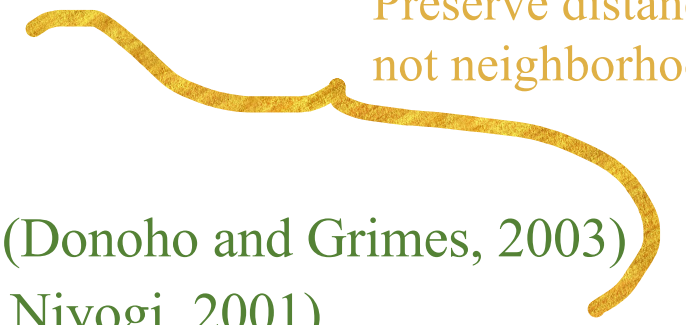
- PCA (Pearson, 1901)
- MDS (Torgerson, 1952)

:

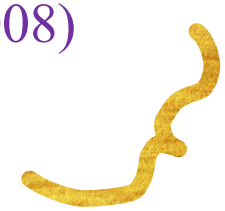
Local Methods

- LLE (Roweis and Saul, 2000),
- Isomap (Tenenbaum et al., 2000)
- Hessian Local Linear Embedding (Donoho and Grimes, 2003)
- Laplacian Eigenmaps (Belkin and Niyogi, 2001)
- Stochastic Neighborhood Embedding (SNE) (Hinton and Roweis, 2003)
- t-SNE (van der Maaten and Hinton, 2008)
- LargeVis (Tang et al., 2016)
- UMAP (McInnes et al., 2018)

:



Crowding problem



Preserve neighborhoods

Global Methods

- PCA (Pearson, 1901)
- MDS (Torgerson, 1952)

:

Local Methods

- LLE (Roweis and Saul, 2000),
- Isomap (Tenenbaum et al., 2000)
- Hessian Local Linea 2003)
- Laplacian Eigenmap
- Stochastic Neighborl d Roweis, 2003)
- t-SNE (van der Maat
- LargeVis (Tang et al., 2016)
- UMAP (McInnes et al., 2018

:

Article | Open Access | Published: 28 November 2019

The art of using t-SNE for single-cell transcriptomics

Dmitry Kobak & Philipp Berens

Nature Communications **10**, Article number: 5416 (2019) | Cite this article

How to Use t-SNE Effectively

MARTIN WATTENBERG, Google Brain
FERNANDA VIÉGAS, Google Brain
IAN JOHNSON, Google Cloud

Oct. 13 2016

Computer Science > Artificial Intelligence

[Submitted on 10 Aug 2017]

Automatic Selection of t-SNE Perplexity

Yanshuai Cao, Luyu Wang

t-Distributed Stoch
dimensionality redu
hyperparameter tha

Automated optimal parameters for T-distributed stochastic neighbor embedding improve visualization and allow analysis of large datasets

October 2018
DOI: 10.1101/451690
Project: Automated Analysis of Flow Cytometry Multidimensional Datasets

Authors:

Anna C Belkina
Boston University

Christopher O. Ciccolella

Rina Anno

Josef Spidlen

Algorithm	Graph components and Loss function
t-SNE (van der Maaten & Hinton, 2008)	Graph components: Edges (i, j) $\text{Loss}_{i,j}^{\text{t-SNE}} = p_{ij} \log \frac{p_{ij}}{q_{ij}}$, where $q_{ij} = \frac{\left(1+\|\mathbf{y}_i-\mathbf{y}_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\|\mathbf{y}_k-\mathbf{y}_l\|^2\right)^{-1}}$ where p_{ij} is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and other $\mathbf{x}_\ell$'s.
UMAP (McInnes et al., 2018)	Graph components: Edges (i, j) $\text{Loss}_{i,j}^{\text{UMAP}} = \begin{cases} \bar{w}_{i,j} \log\left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1} & i, j \text{ neighbors} \\ (1 - \bar{w}_{i,j}) \log\left(1 - \left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1}\right) & \text{otherwise,} \end{cases}$ where $\bar{w}_{i,j}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and nearby $\mathbf{x}_\ell$'s.
TriMAP (Amid & Warmuth, 2019)	Graph components: Triplets (i, j, k) where $\text{Distance}_{i,j} \leq \text{Distance}_{i,k}$ $\text{Loss}_{i,j,k}^{\text{TM}} = \omega_{i,j,k} \frac{s(\mathbf{y}_i,\mathbf{y}_k)}{s(\mathbf{y}_i,\mathbf{y}_j)+s(\mathbf{y}_i,\mathbf{y}_k)}$, where $s(\mathbf{y}_i, \mathbf{y}_j) = \left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}$ and $\omega_{i,j,k}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$ and nearby points.

Hard to understand what's important here…

Start from the obvious:

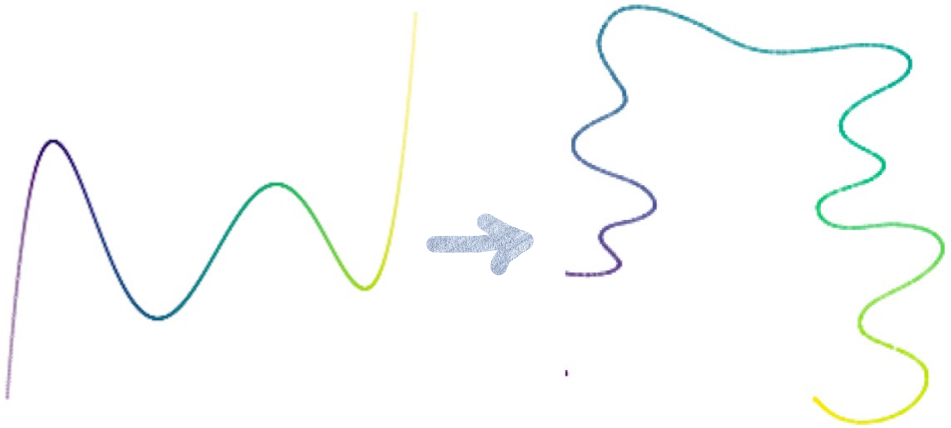
- Attraction: high-dimensional neighbors should be attracted.
- Repulsion: points far in original space should be far in low-dim space.

But that's not enough…

$$\sum_{(i,j)\in\mathcal{T}_{\text{neighbors}}} l^{\text{attract}}(i,j) + \sum_{(i,k)\in\mathcal{T}_{\text{further}}} l^{\text{repulse}}(i,k)$$

Attract neighbors

Repulse far points



Start from the obvious:

- Attraction: high-dimensional neighbors should be attracted.
- Repulsion: points far in original space should be far in low-dim space.

After a huge amount of experimentation, we found that:

- Certain specific properties of the loss function are important for local structure.
- The choice of which graph components to exert forces on is important for global structure.

$$\sum_{\text{Subset of graph components } \{i\}} \text{Weight}^{\mathbf{X}}(\mathcal{C}_i^H) \cdot \text{Loss}^{\mathbf{Y}}(\mathcal{C}_i^L)$$

After a huge amount of experimentation, we found that:

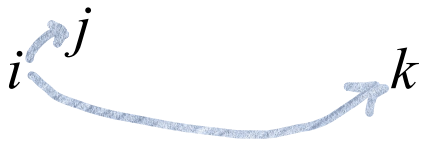
- Certain specific properties of the loss function are important for local structure.
- The choice of which graph components to exert forces on is important for global structure.

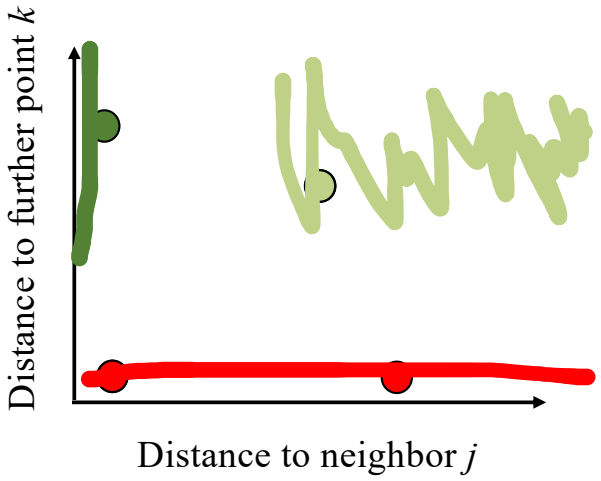
Algorithm	Graph components and Loss function
t-SNE (van der Maaten & Hinton, 2008)	Graph components: Edges (i, j) $\text{Loss}_{i,j}^{\text{t-SNE}} = p_{ij} \log \frac{p_{ij}}{q_{ij}}$, where $q_{ij} = \frac{\left(1+\|\mathbf{y}_i-\mathbf{y}_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\|\mathbf{y}_k-\mathbf{y}_l\|^2\right)^{-1}}$ where p_{ij} is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and other $\mathbf{x}_\ell$'s.
UMAP (McInnes et al., 2018)	Graph components: Edges (i, j) $\text{Loss}_{i,j}^{\text{UMAP}} = \begin{cases} \bar{w}_{i,j} \log\left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1} & i, j \text{ neighbors} \\ (1 - \bar{w}_{i,j}) \log\left(1 - \left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1}\right) & \text{otherwise,} \end{cases}$ where $\bar{w}_{i,j}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and nearby $\mathbf{x}_\ell$'s.
TriMAP (Amid & Warmuth, 2019)	Graph components: Triplets (i, j, k) where $\text{Distance}_{i,j} \leq \text{Distance}_{i,k}$ $\text{Loss}_{i,j,k}^{\text{TM}} = \omega_{i,j,k} \frac{s(\mathbf{y}_i,\mathbf{y}_k)}{s(\mathbf{y}_i,\mathbf{y}_j)+s(\mathbf{y}_i,\mathbf{y}_k)}$, where $s(\mathbf{y}_i, \mathbf{y}_j) = \left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}$ and $\omega_{i,j,k}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$ and nearby points.

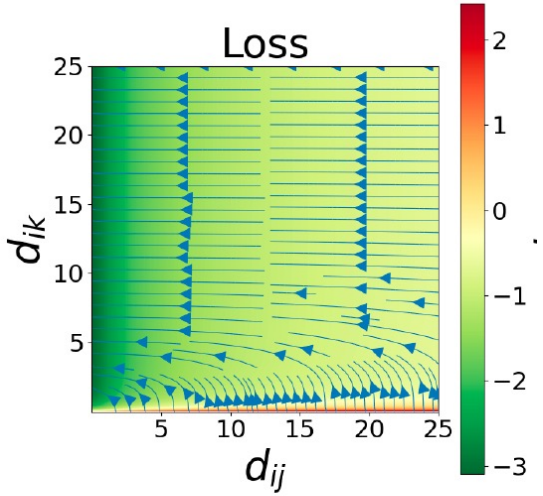
Hard to understand what's important here…

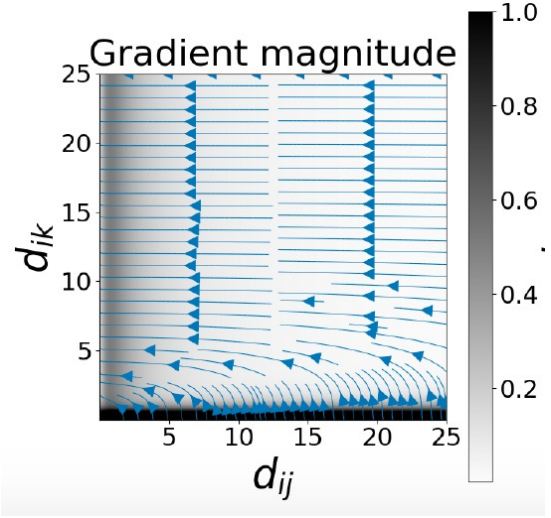
The “rainbow” plot

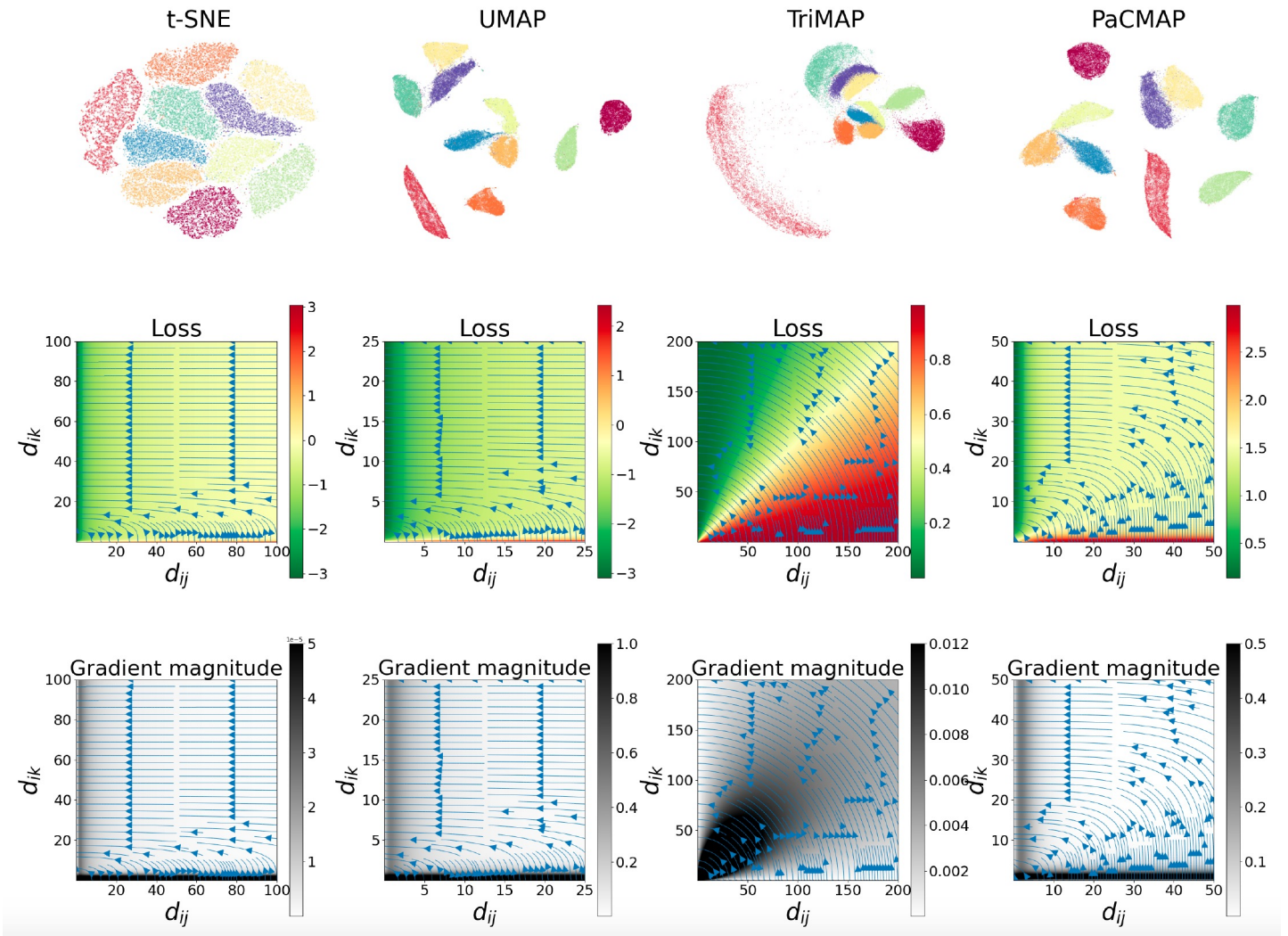
Triple i, j (neighbor), k (further)



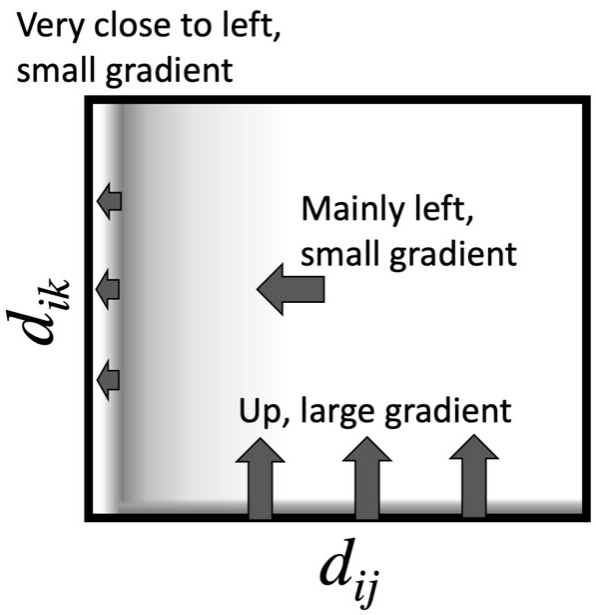






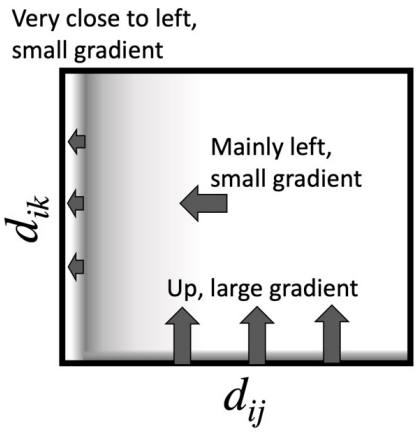

t-SNE
UMAP
TriMAP
PaCMAP
Loss
Gradient magnitude
d_{ik}
d_{ij}

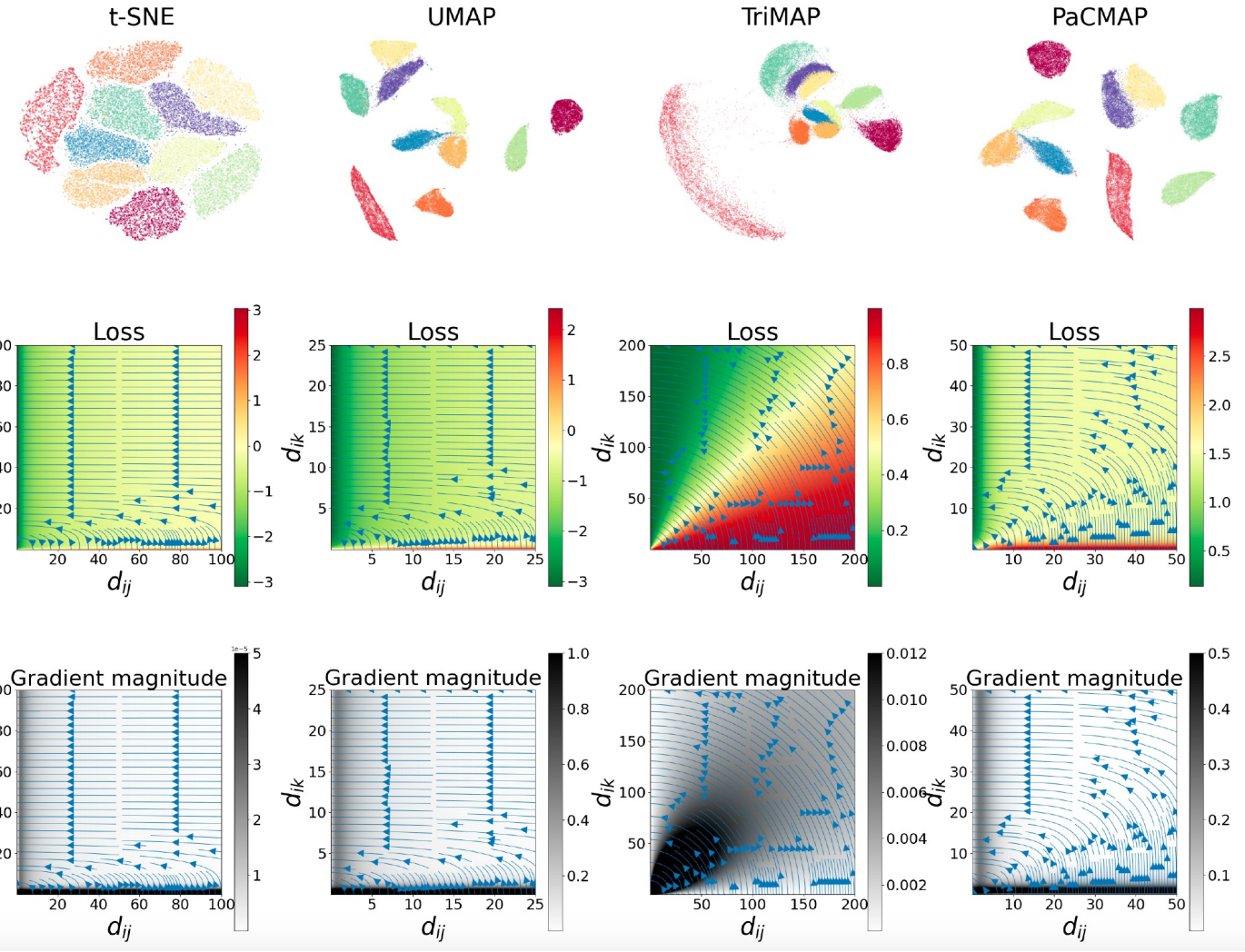
Our principles for a good loss function



1. Monotonicity
2. Except at bottom, gradient should go mainly to the left.
(if further point is sufficiently far, should focus on pulling neighbor closer.)
3. At bottom, gradient goes up.
(push further points away really hard)
4. At left, gradient has small magnitude.
(don't crowd, relax when close enough)
5. At bottom, gradient has large magnitude.
(push farther point away)
6. Gradient fades as neighbor gets farther away.
(give up on neighbors when they are too far)

Our principles for a good loss function





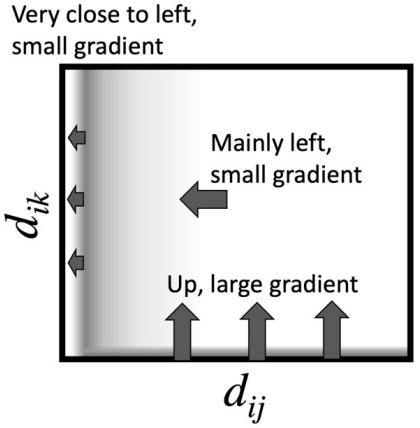
Bad loss functions

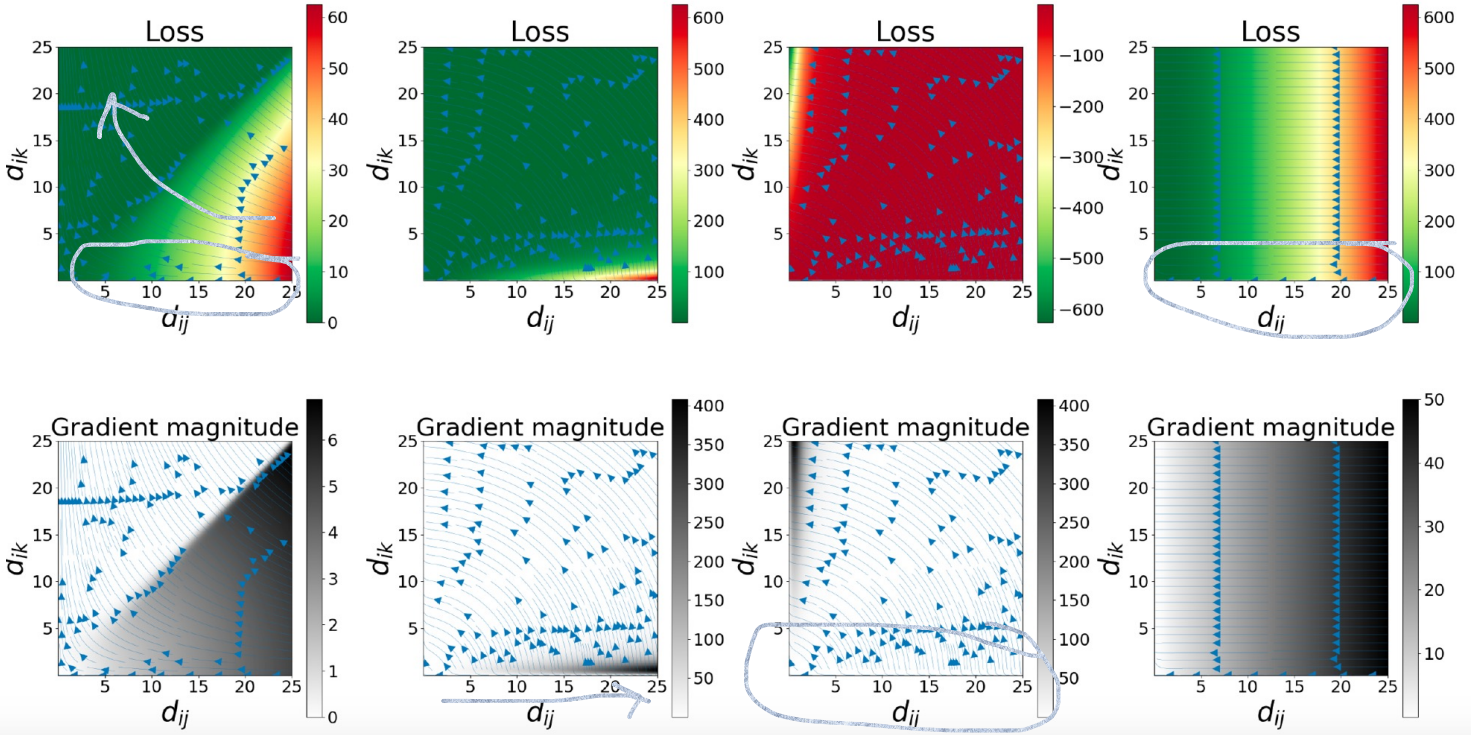
$Loss = log\left(1 + exp\left(\frac{d_{ij}^2 - d_{ik}^2}{10}\right)\right)$

$Loss = \frac{d_{ij}^2 + 1}{d_{ik}^2 + 1}$

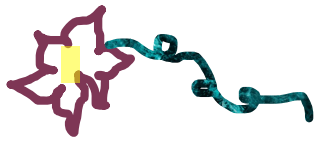
$Loss = -\frac{d_{ik}^2 + 1}{d_{ij}^2 + 1}$

$Loss = log(1 + exp(d_{ij}^2) + exp(-d_{ik}^2))$

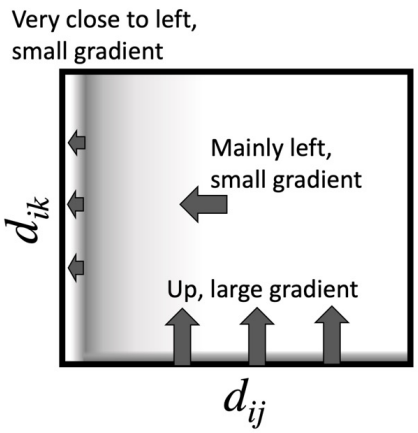




supposed to give up on far neighbors…



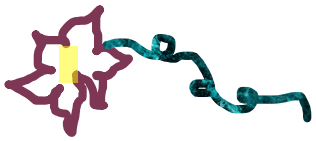
Our principles for a good loss function



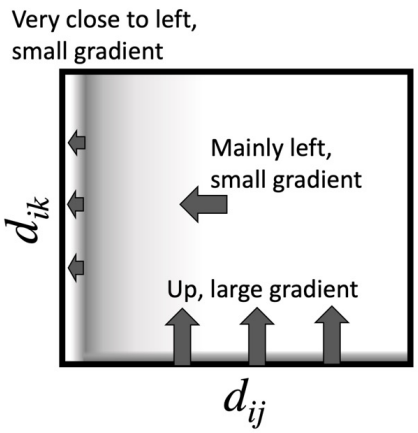
VS.

Algorithm	Graph components and Loss function
t-SNE	Graph components: Edges (i, j) $\text{Loss}_{i,j}^{\text{t-SNE}} = p_{ij} \log \frac{p_{ij}}{q_{ij}}$, where $q_{ij} = \frac{\left(1+\|\mathbf{y}_i-\mathbf{y}_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\|\mathbf{y}_k-\mathbf{y}_l\|^2\right)^{-1}}$ where p_{ij} is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and other $\mathbf{x}_\ell$'s.
UMAP	Graph components: Edges (i, j) $\text{Loss}_{i,j}^{\text{UMAP}} = \begin{cases} \bar{w}_{i,j} \log\left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1} & i, j \text{ neighbors} \\ (1 - \bar{w}_{i,j}) \log\left(1 - \left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1}\right) & \text{otherwise,} \end{cases}$ where $\bar{w}_{i,j}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and nearby $\mathbf{x}_\ell$'s.
TriMAP	Graph components: Triplets (i, j, k) where $\text{Distance}_{i,j} \leq \text{Distance}_{i,k}$ $\text{Loss}_{i,j,k}^{\text{TM}} = \omega_{i,j,k} \frac{s(\mathbf{y}_i,\mathbf{y}_k)}{s(\mathbf{y}_i,\mathbf{y}_j)+s(\mathbf{y}_i,\mathbf{y}_k)}$, where $s(\mathbf{y}_i, \mathbf{y}_j) = \left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}$ and $\omega_{i,j,k}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$ and nearby points.

PaCMAP's loss is simpler.



Our principles for a good loss function



VS.

Algorithm	Graph components and Loss function
t-SNE	Graph components: Edges (i,j) $\text{Loss}_{i,j}^{\text{t-SNE}} = p_{ij} \log \frac{p_{ij}}{q_{ij}}$, where $q_{ij} = \frac{(1+\|\mathbf{y}_i-\mathbf{y}_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|\mathbf{y}_k-\mathbf{y}_l\|^2)^{-1}}$ where p_{ij} is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and other $\mathbf{x}_\ell$'s.
UMAP	Graph components: Edges (i,j) $\text{Loss}_{i,j}^{\text{UMAP}} = \begin{cases} \bar{w}_{i,j} \log\left(1+a\left(\|\mathbf{y}_i-\mathbf{y}_j\|_2^2\right)^b\right)^{-1} & i,j \text{ neighbors} \\ (1-\bar{w}_{i,j}) \log\left(1-\left(1+a\left(\|\mathbf{y}_i-\mathbf{y}_j\|_2^2\right)^b\right)^{-1}\right) & \text{otherwise,} \end{cases}$ where $\bar{w}_{i,j}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and nearby $\mathbf{x}_\ell$'s.
TriMAP	Graph components: Triplets (i,j,k) where $\text{Distance}_{i,j} \leq \text{Distance}_{i,k}$ $\text{Loss}_{i,j,k}^{\text{TM}} = \omega_{i,j,k} \frac{s(\mathbf{y}_i,\mathbf{y}_k)}{s(\mathbf{y}_i,\mathbf{y}_j)+s(\mathbf{y}_i,\mathbf{y}_k)}$, where $s(\mathbf{y}_i,\mathbf{y}_j) = \left(1+\|\mathbf{y}_i-\mathbf{y}_j\|^2\right)^{-1}$ and $\omega_{i,j,k}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$ and nearby points.

PaCMAP's loss is simpler.

$$\text{distance}\,(i,j) := \|\boldsymbol{y_i} - \boldsymbol{y_j}\|^2 + 1$$

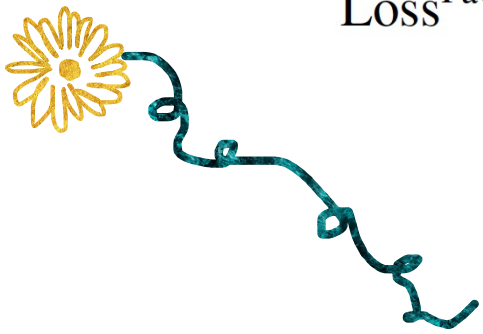
$$\text{Loss}^{\text{PaCMAP}} = w_{\text{neighbors}}\text{Loss}_{\text{neighbors}} + \quad ??? \quad + w_{FP}\text{Loss}_{FP}$$

$$\text{Loss}_{\text{neighbors}} = \frac{\text{distance}\,(i,j)}{\text{distance}\,(i,j)+10}$$

attractive

$$\text{Loss}_{FP} = \frac{1}{\text{distance}\,(i,l)+1}$$

repulsive



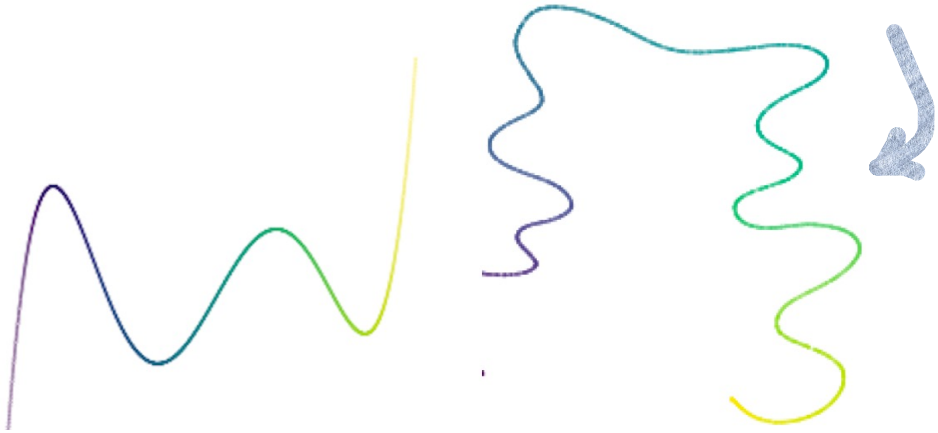
- Certain specific properties of the loss function are important for local structure.
- The choice of which graph components to exert forces on is important for global structure.

(UMAP)

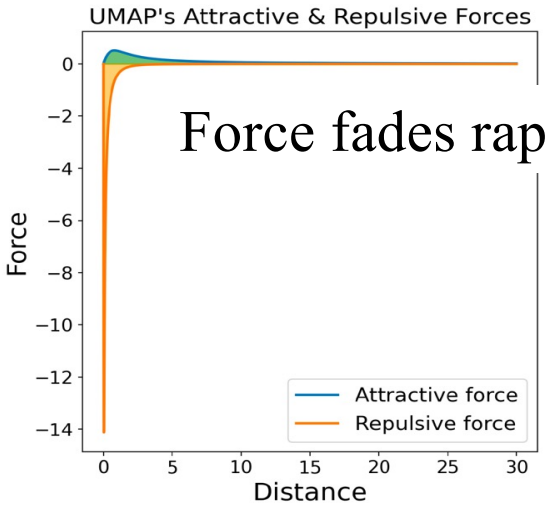
$$\sum_{(i,j)\in\mathcal{T}_{\text{neighbors}}} l^{\text{attract}}(i,j) + \sum_{(i,k)\in\mathcal{T}_{\text{further}}} l^{\text{repulse}}(i,k)$$

Attract neighbors

Repulse far points



Not enough forces!



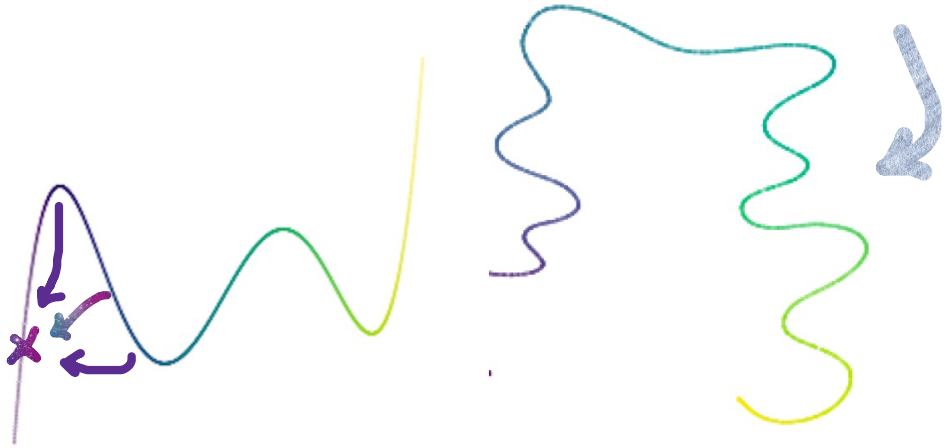
Force fades rapidly!

(UMAP)

$$\sum_{(i,j)\in\mathcal{T}_{\text{neighbors}}} l^{\text{attract}}(i,j) + \sum_{(i,k)\in\mathcal{T}_{\text{further}}} l^{\text{repulse}}(i,k)$$

Attract neighbors

Repulse far points



Not enough…

Let's attract "mid-near" points
to maintain global structure

PaCMAP's Loss

$$\text{Loss}^{\text{PaCMAP}} = w_{\text{neighbors}}\text{Loss}_{\text{neighbors}} + w_{MN}\text{Loss}_{MN} + w_{FP}\text{Loss}_{FP}$$

$$\text{distance}\,(i,j) := \|\boldsymbol{y_i} - \boldsymbol{y_j}\|^2 + 1$$

$$\text{Loss}_{\text{neighbors}} = \frac{\text{distance}\,(i,j)}{\text{distance}\,(i,j)+10}$$

Neighbors:
attractive

$$\text{Loss}_{MN} = \frac{\text{distance}\,(i,l)}{\text{distance}\,(i,l)+10000}$$

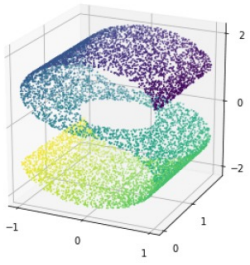
Mid-near pairs:
mild attractive

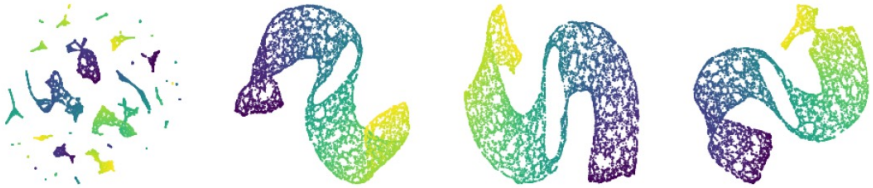
$$\text{Loss}_{FP} = \frac{1}{\text{distance}\,(i,l)+1}$$

Further points:
repulsive

Mid near pair for i : sample 6 observations, choose the second closest of the 6, pair it with i.

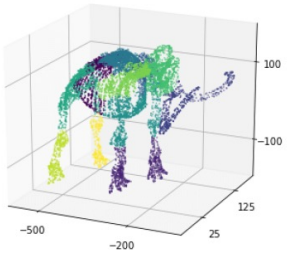
Mid-near pairs have an effect!

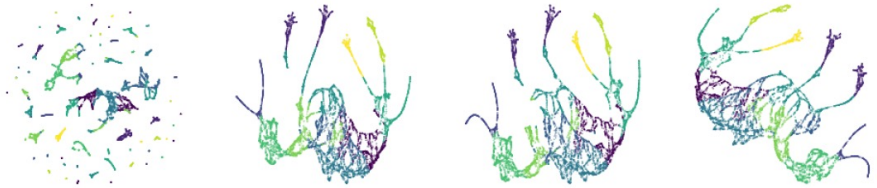




0 2 5 8

mid-near pairs





0 2 5 8

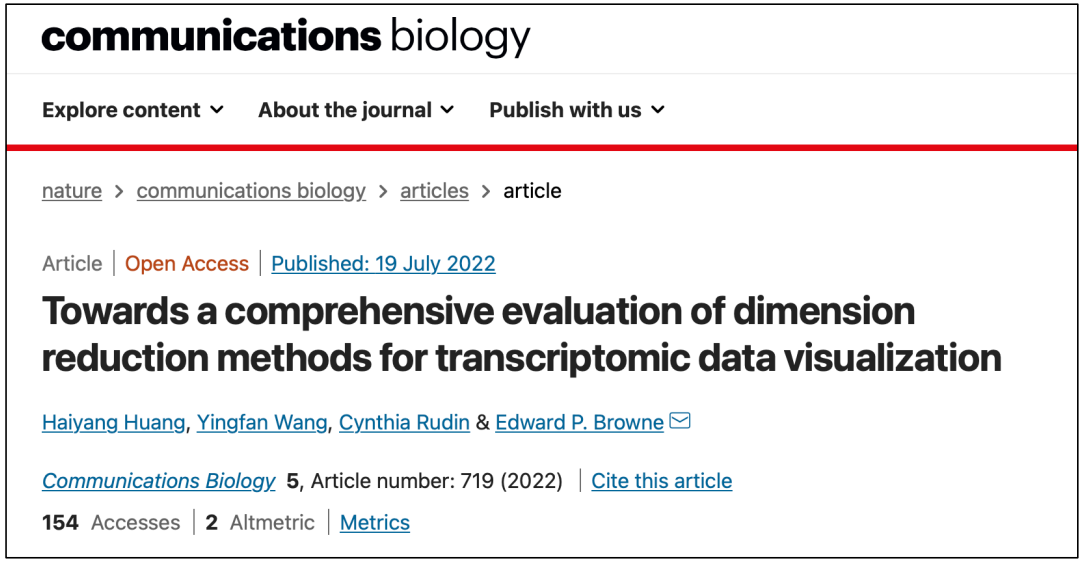
mid-near pairs

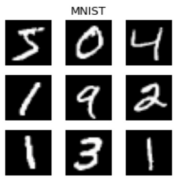
- Certain specific properties of the loss function are important for local structure.
- The choice of which graph components to exert forces on is important for global structure.

How to evaluate DR?

How to Evaluate DR Algorithms?

- Local structure – DR then supervised classification in 2D
- Global structure – triplet loss
- Sensitivity to parameter choices
- Sensitivity to preprocessing choices
- Computational efficiency





t-SNE(perplexity=10)

UMAP(n_neighbors=10)

TriMAP(n_inliers=8)

t-SNE(perplexity=20)

UMAP(n_neighbors=20)

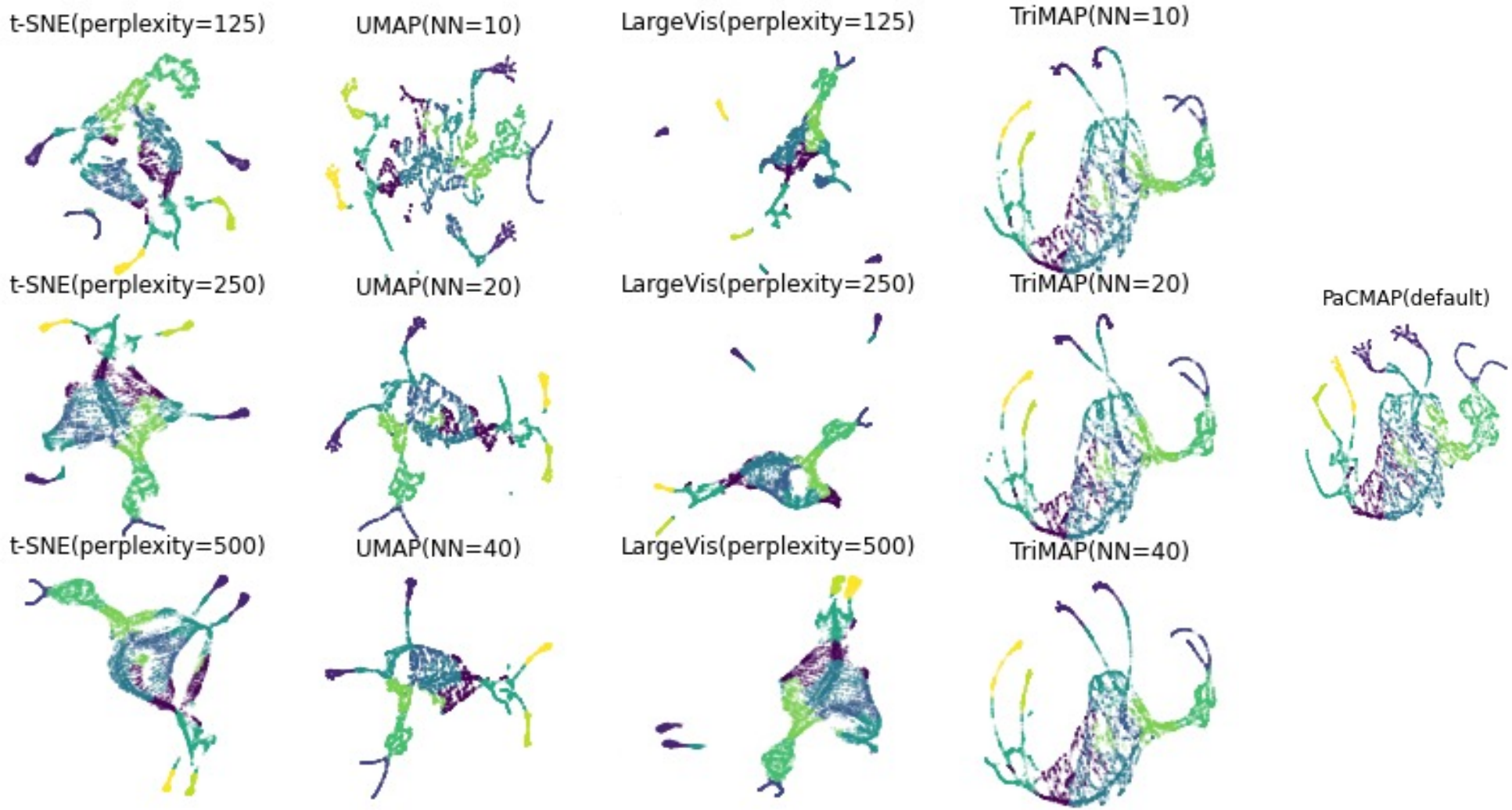
TriMAP(n_inliers=10)

PaCMAP

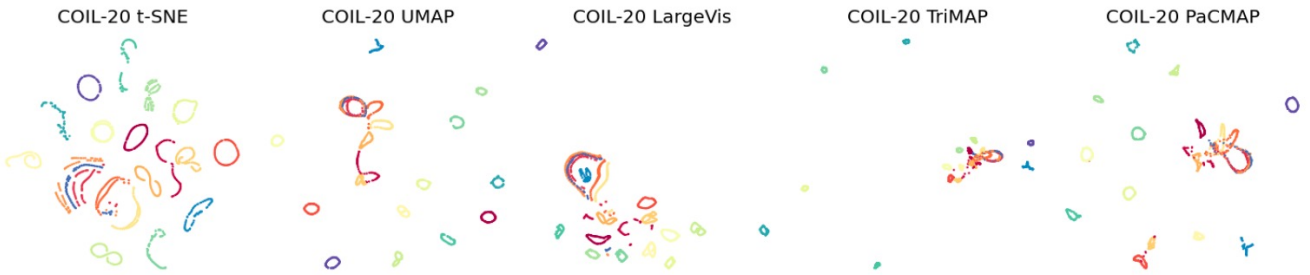
t-SNE(perplexity=40)

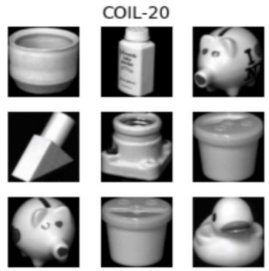
UMAP(n_neighbors=40)

TriMAP(n_inliers=15)

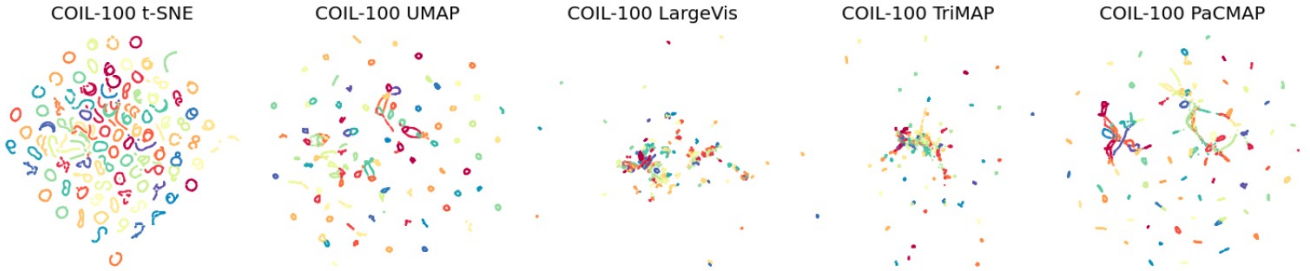

t-SNE(perplexity=125)
UMAP(NN=10)
LargeVis(perplexity=125)
TriMAP(NN=10)
t-SNE(perplexity=250)
UMAP(NN=20)
LargeVis(perplexity=250)
TriMAP(NN=20)
PaCMAP(default)
t-SNE(perplexity=500)
UMAP(NN=40)
LargeVis(perplexity=500)
TriMAP(NN=40)

COIL-20 Data

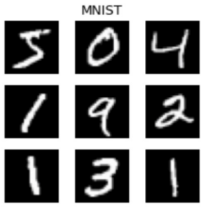




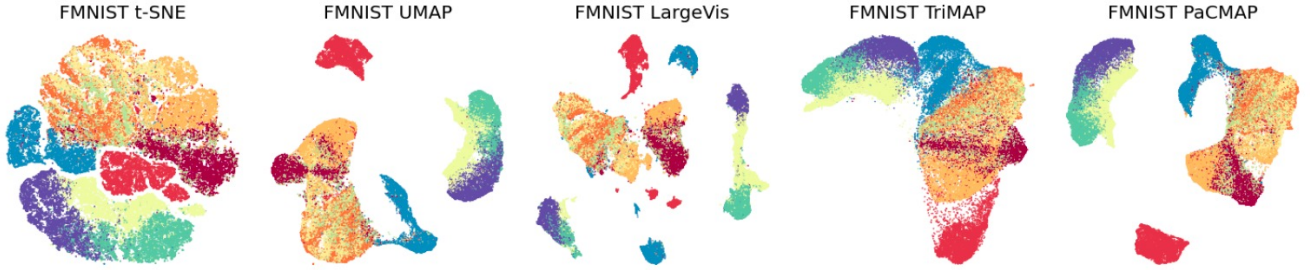
COIL-100 Data

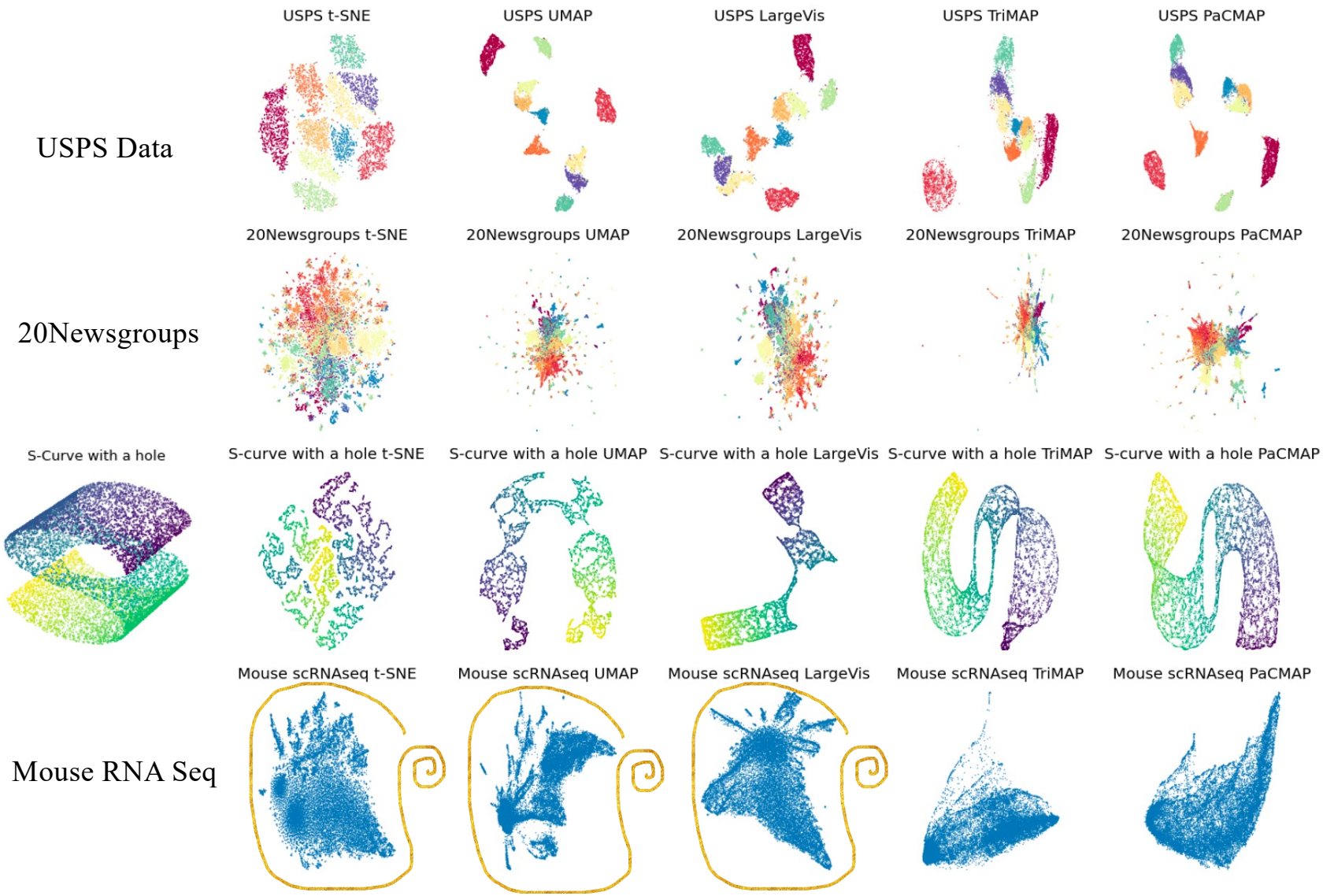


MNIST Data



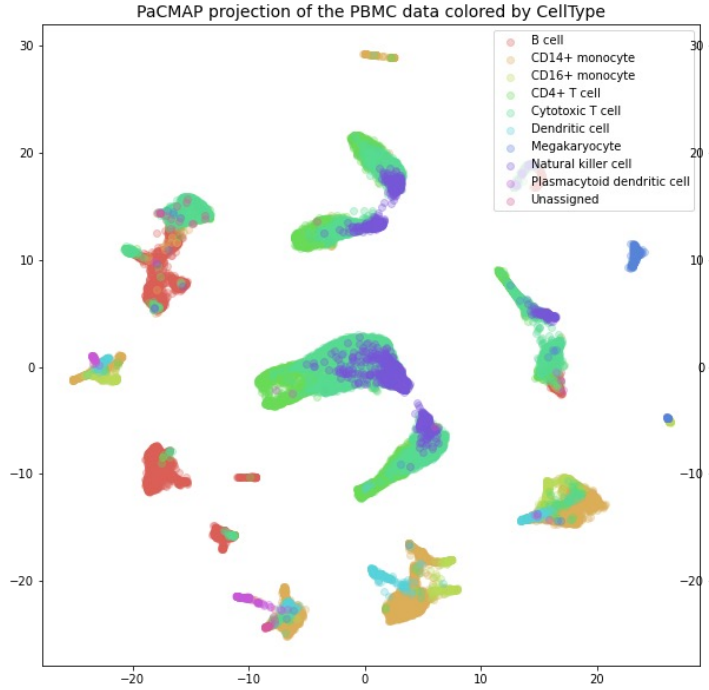
Fashion MNIST

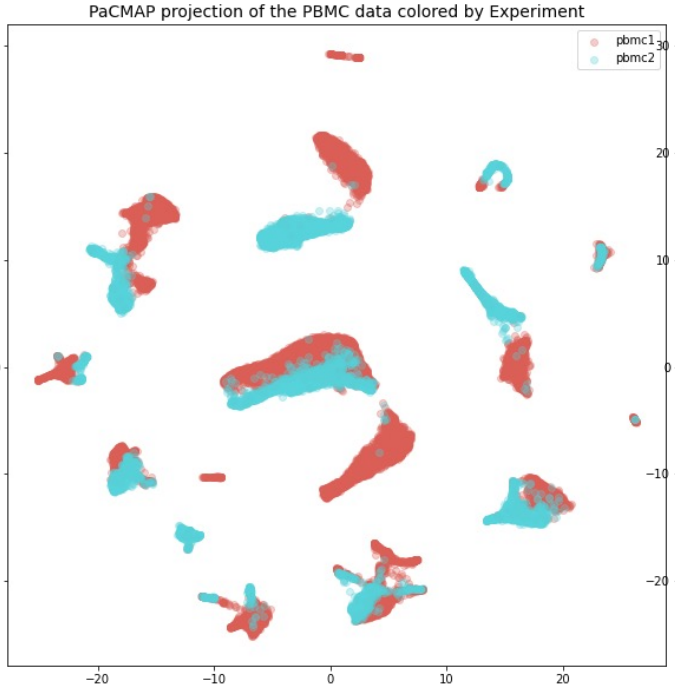


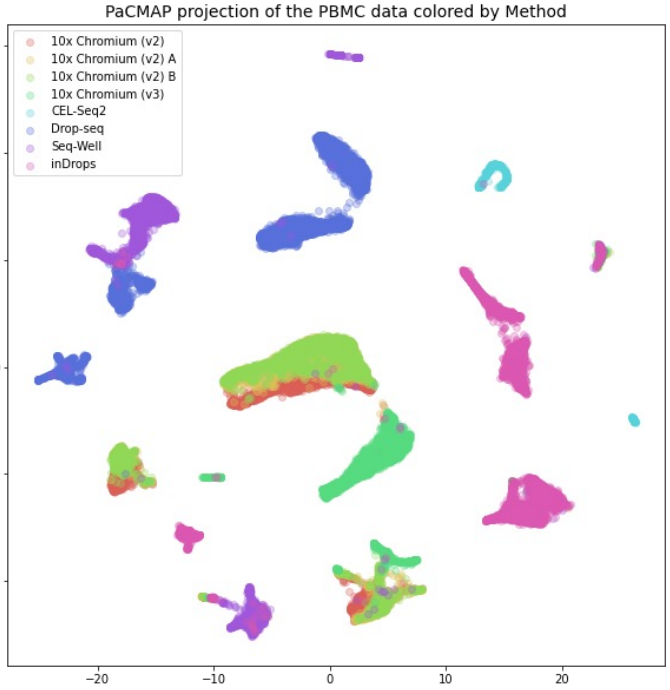

USPS Data
USPS t-SNE
USPS UMAP
USPS LargeVis
USPS TriMAP
USPS PaCMAP
20Newsgroups
20Newsgroups t-SNE
20Newsgroups UMAP
20Newsgroups LargeVis
20Newsgroups TriMAP
20Newsgroups PaCMAP
S-Curve with a hole
S-curve with a hole t-SNE
S-curve with a hole UMAP
S-curve with a hole LargeVis
S-curve with a hole TriMAP
S-curve with a hole PaCMAP
Mouse RNA Seq
Mouse scRNAseq t-SNE
Mouse scRNAseq UMAP
Mouse scRNAseq LargeVis
Mouse scRNAseq TriMAP
Mouse scRNAseq PaCMAP

PBMC Data, PaCMAP projection, Colored 3 Ways

(PBMC is Peripheral Blood Mononuclear Cell)







Colored by Cell Type

Colored by Experiment

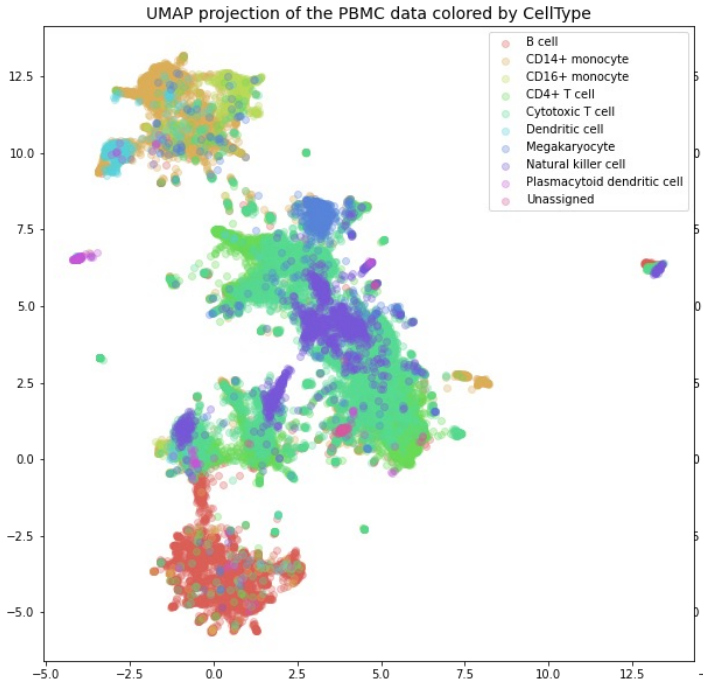
Colored by Method

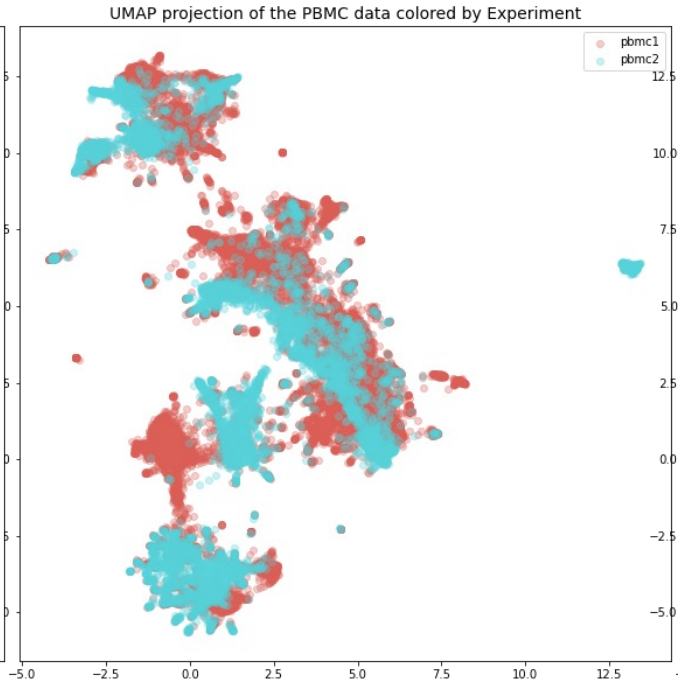
PaCMAP: 23.90 seconds

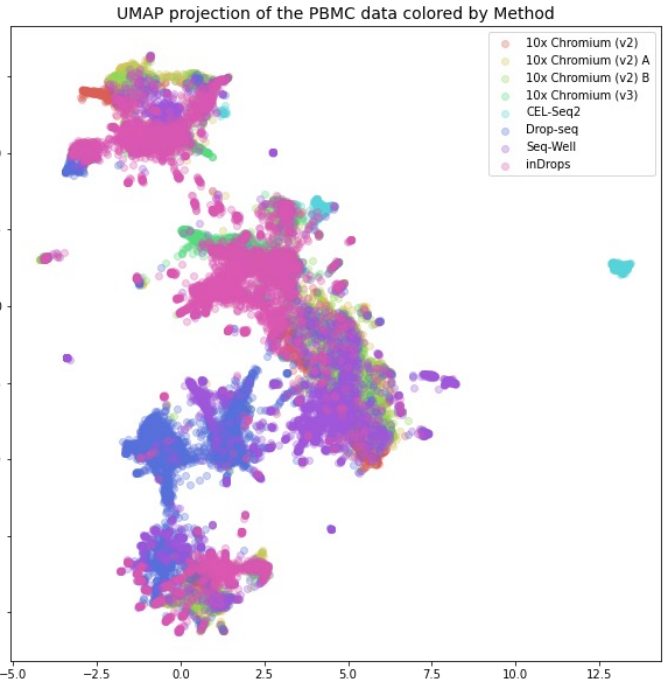
figure credit: Carla Moelbert

Data from Ding J, Adiconis X, Simmons SK, Kowalczyk MS et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol 2020 Jun;38(6):737-746. PMID: 32341560

PBMC Data, UMAP projection, Colored 3 Ways







Colored by Cell Type

Colored by Experiment

Colored by Method

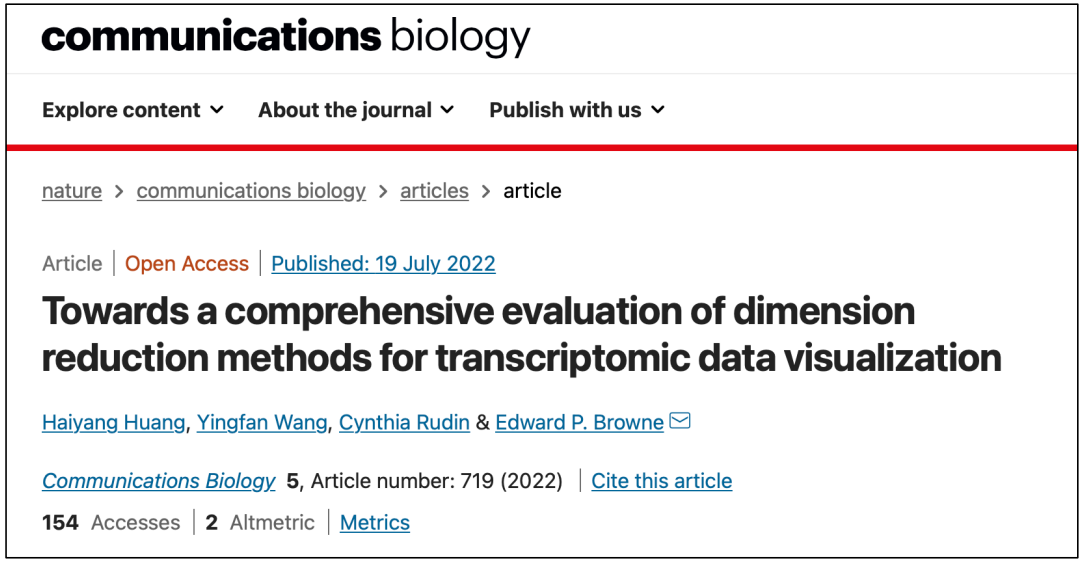
Data from Ding J, Adiconis X, Simmons SK, Kowalczyk MS et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol 2020 Jun;38(6):737-746. PMID: 32341560

UMAP: 75.82 seconds

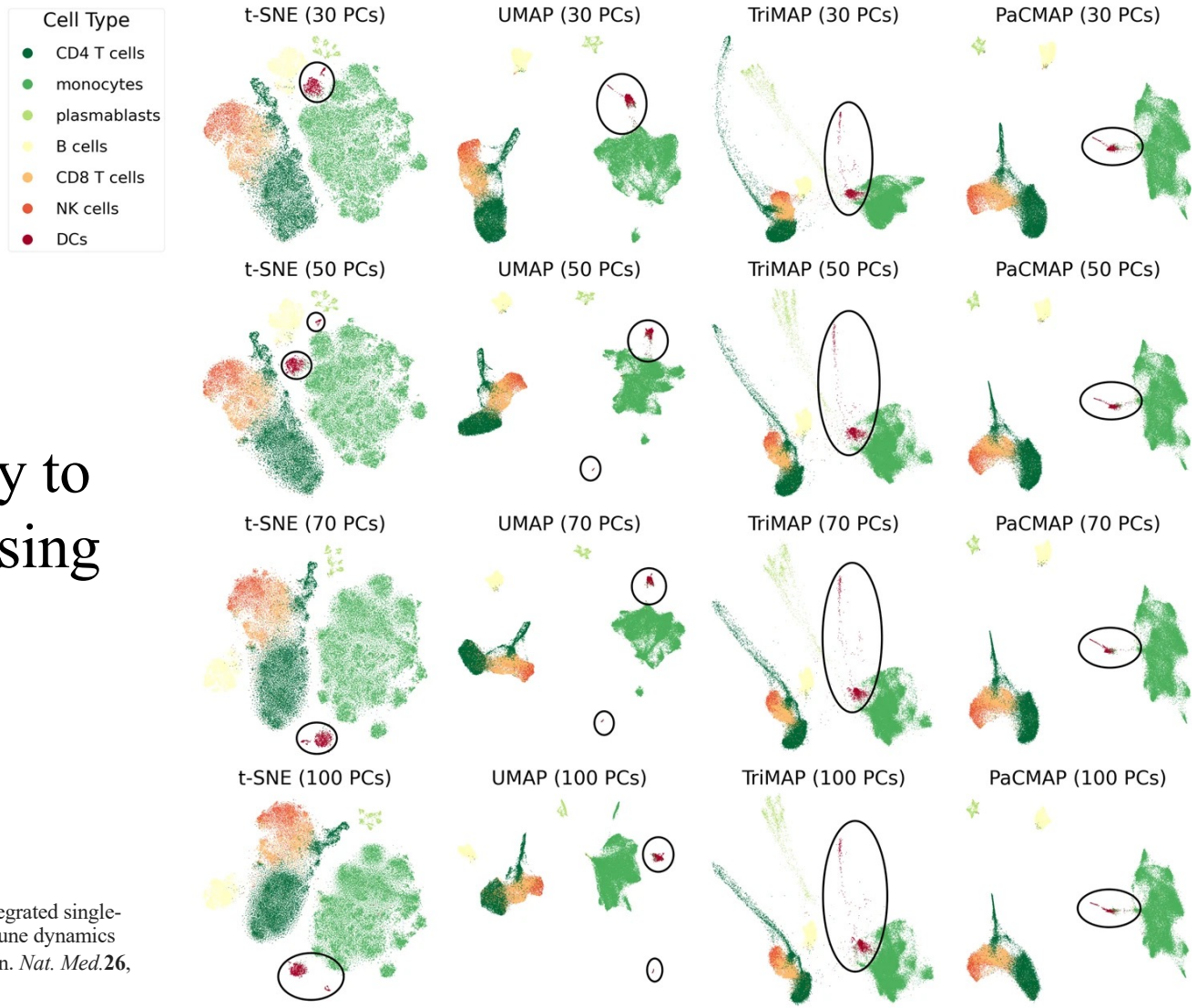
figure credit: Carla Moelbert

How to Evaluate DR Algorithms?

- Local structure – DR then supervised classification in 2D
- Global structure – triplet loss
- Sensitivity to parameter choices
- Sensitivity to preprocessing choices
- Computational efficiency



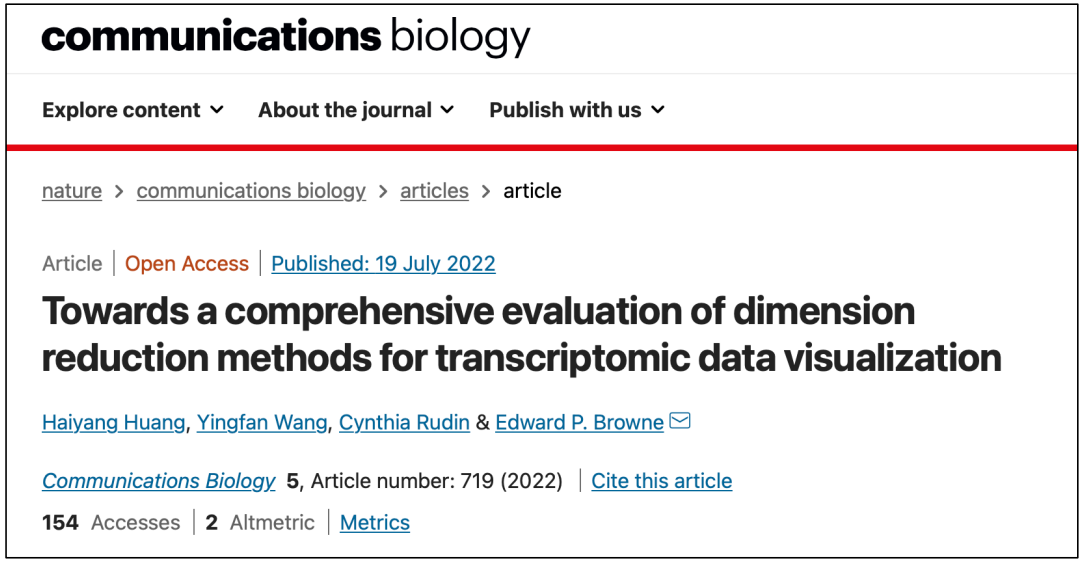
Sensitivity to Preprocessing



Data from: Kazer, S. W. et al. Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection. *Nat. Med.***26**, 511–518 (2020).

How to Evaluate DR Algorithms?

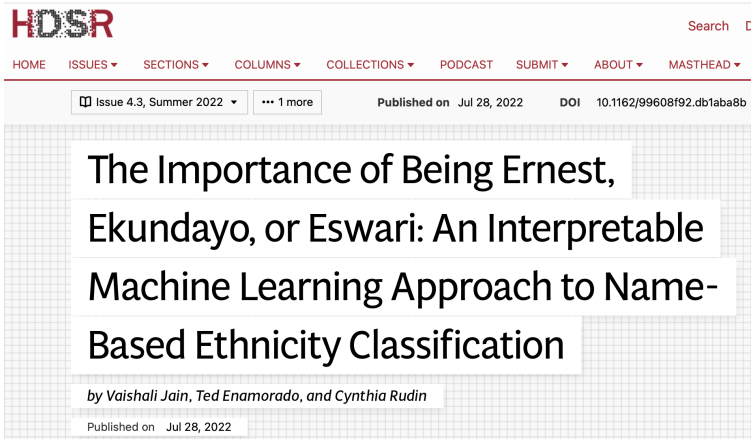
- Local structure – DR then supervised classification in 2D
- Global structure – triplet loss
- Sensitivity to parameter choices
- Sensitivity to preprocessing choices
- Computational efficiency

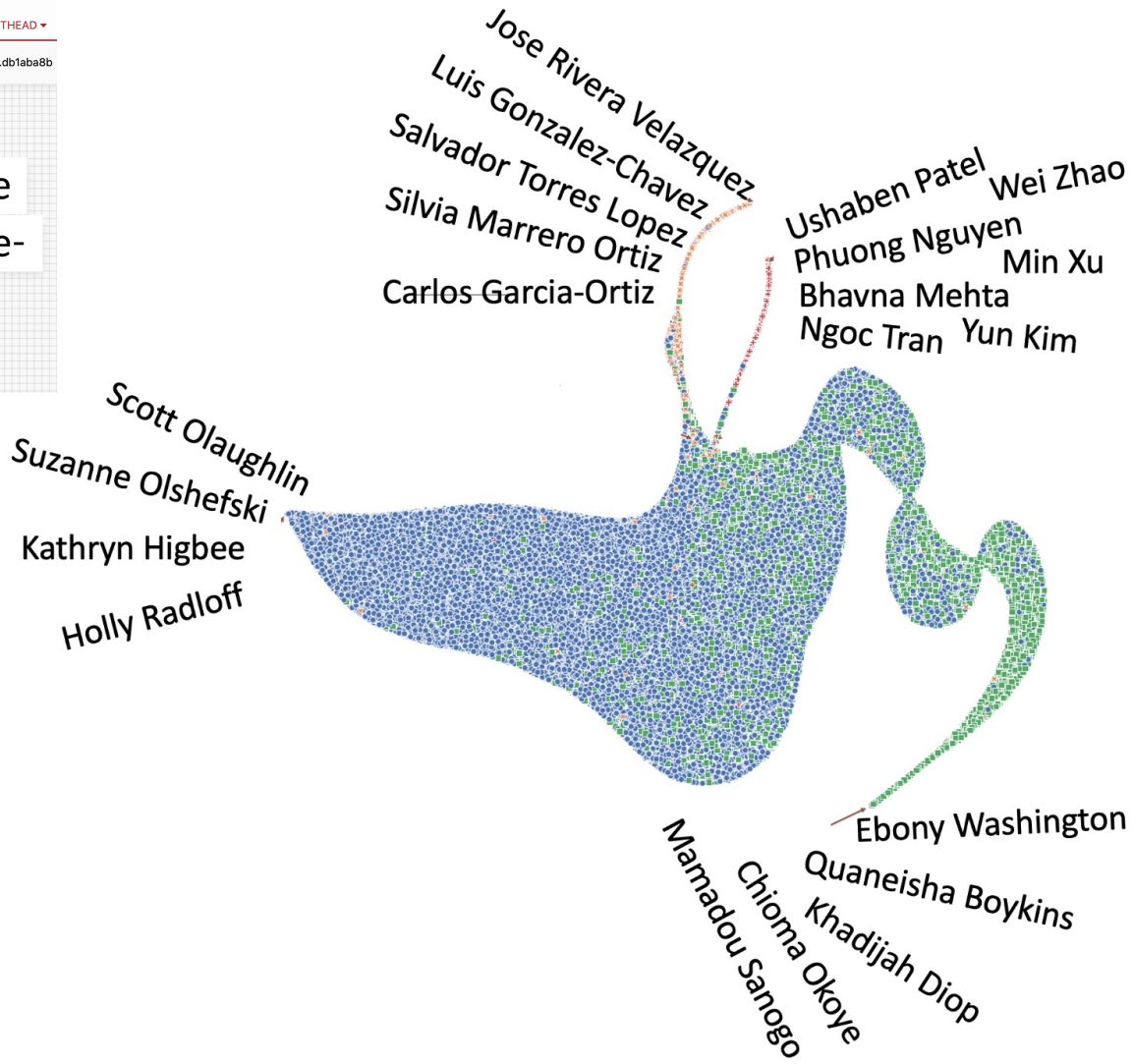


Run time

Dataset (Size)	t-SNE	LargeVis	UMAP	TriMAP	PaCMAP
Olivetti Faces (0.4K)	00:00:04	00:08:13	00:00:02	**00:00:01**	**00:00:01**
COIL-20 (1.4K)	00:00:08	00:10:18	00:00:05	00:00:02	**00:00:01**
COIL-100 (7.2K)	00:00:49	00:09:53	00:00:10	00:00:06	**00:00:03**
S-Curve with Hole (9.5K)	00:01:17	00:10:09	00:00:15	00:00:08	**00:00:05**
USPS (9.5K)	00:01:14	00:10:15	00:00:15	00:00:07	**00:00:05**
Mammoth (10K)	00:00:58	00:10:36	00:00:16	00:00:08	**00:00:05**
20Newsgroups (18K)	00:03:29	00:11:40	00:00:19	00:00:18	**00:00:12**
Mouse scRNA-seq (20K)	00:04:43	00:12:52	00:00:24	00:00:20	**00:00:13**
MNIST (70K)	00:14:02	00:20:19	00:01:09	00:01:14	**00:00:52**
F-MNIST (70K)	00:12:43	00:17:11	00:00:59	00:01:13	**00:00:47**
Flow Cytometry (3M)	-	-	-	02:10:27	**00:58:28**
KDD Cup99 (4M)	-	-	-	03:34:57	**02:05:19**

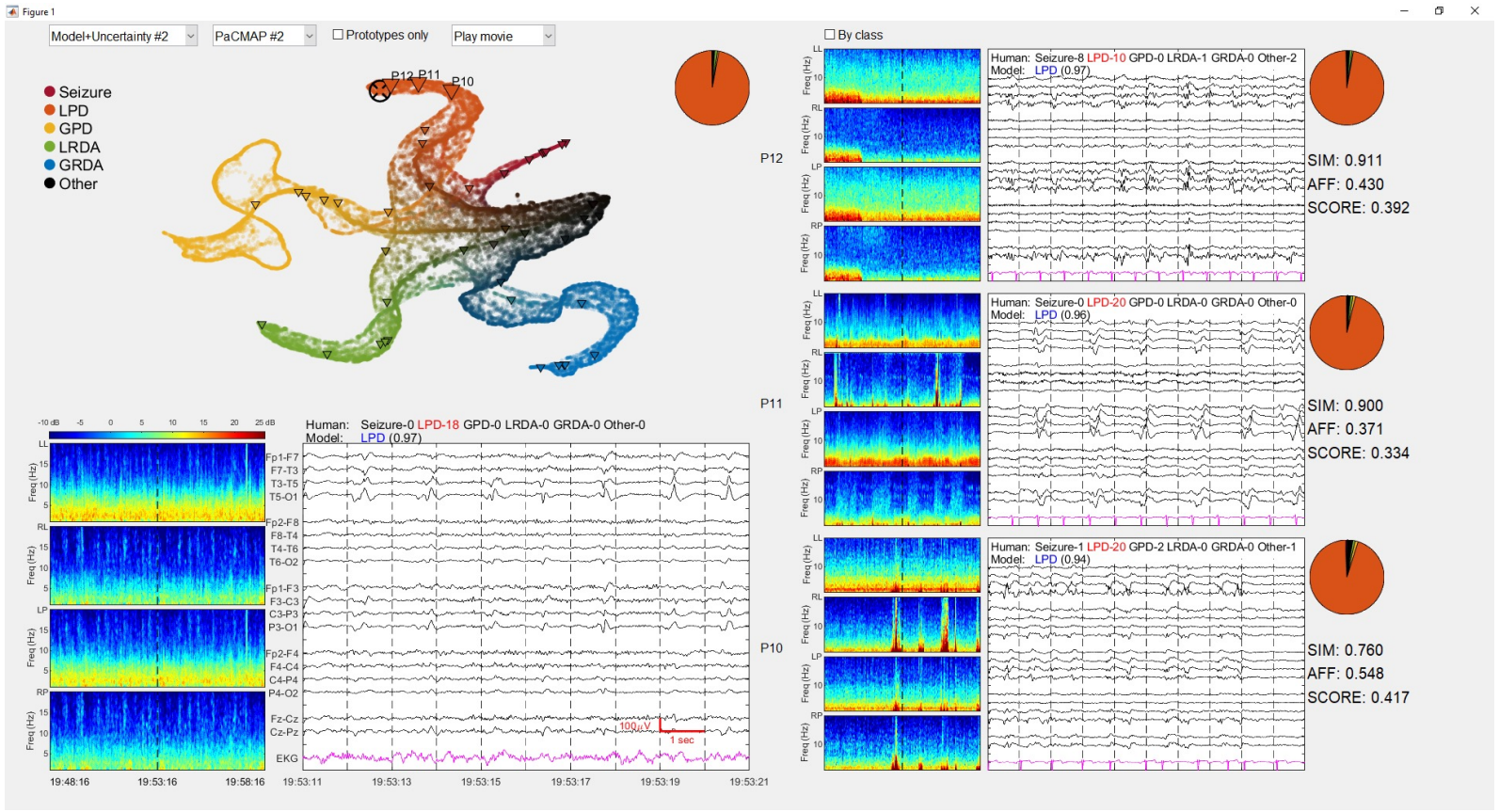
Name-Ethnicity Classification
(helpful for assessing fairness)


HDSR
Search
HOME
ISSUES
SECTIONS
COLUMNS
COLLECTIONS
PODCAST
SUBMIT
ABOUT
MASTHEAD
Issue 4.3, Summer 2022
1 more
Published on Jul 28, 2022
DOI 10.1162/99608f92.db1aba8b
The Importance of Being Ernest, Ekundayo, or Eswari: An Interpretable Machine Learning Approach to Name-Based Ethnicity Classification
by Vaishali Jain, Ted Enamorado, and Cynthia Rudin
Published on Jul 28, 2022

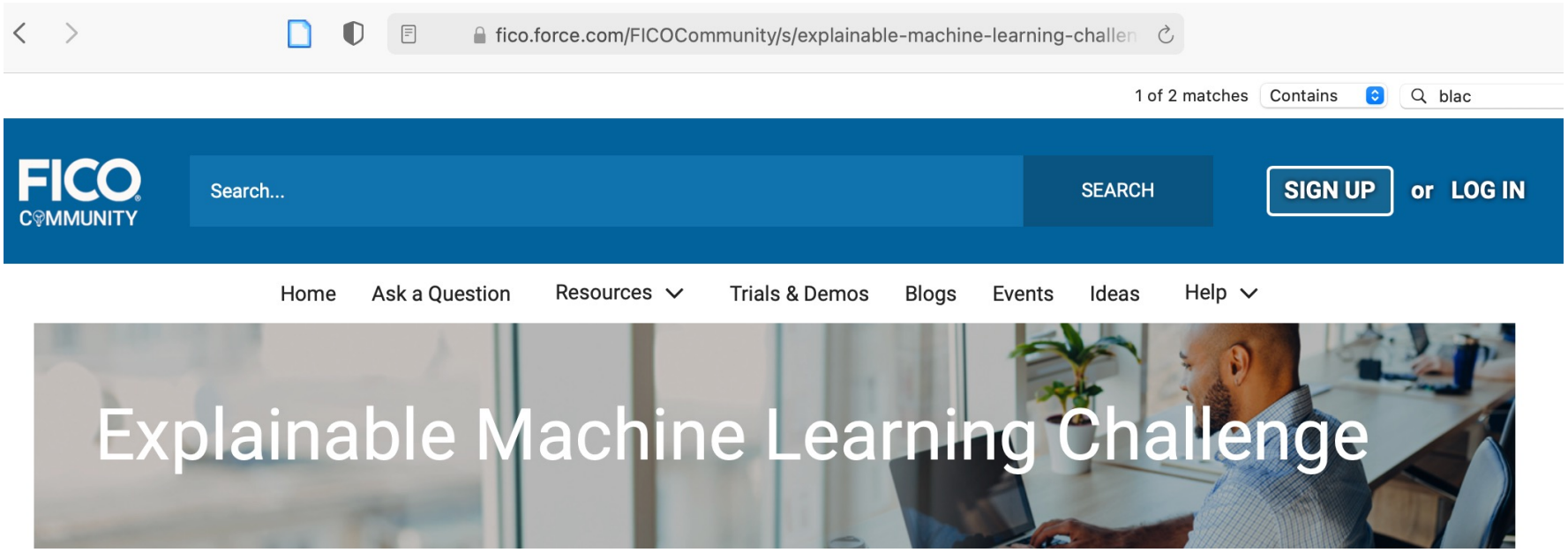

Jose Rivera Velazquez
Luis Gonzalez-Chavez
Salvador Torres Lopez
Silvia Marrero Ortiz
Carlos Garcia-Ortiz
Ushaben Patel
Wei Zhao
Phuong Nguyen
Min Xu
Bhavna Mehta
Ngoc Tran
Yun Kim
Scott Olaughlin
Suzanne Olshefski
Kathryn Higbee
Holly Radloff
Ebony Washington
Quaneisha Boykins
Khadijah Diop
Chioma Okoye
Mamadou Sanogo

EthnicIA

EEG Monitoring



joint work with Alina Barnett, Zhicheng Guo, Jing Jing and Brandon Westover

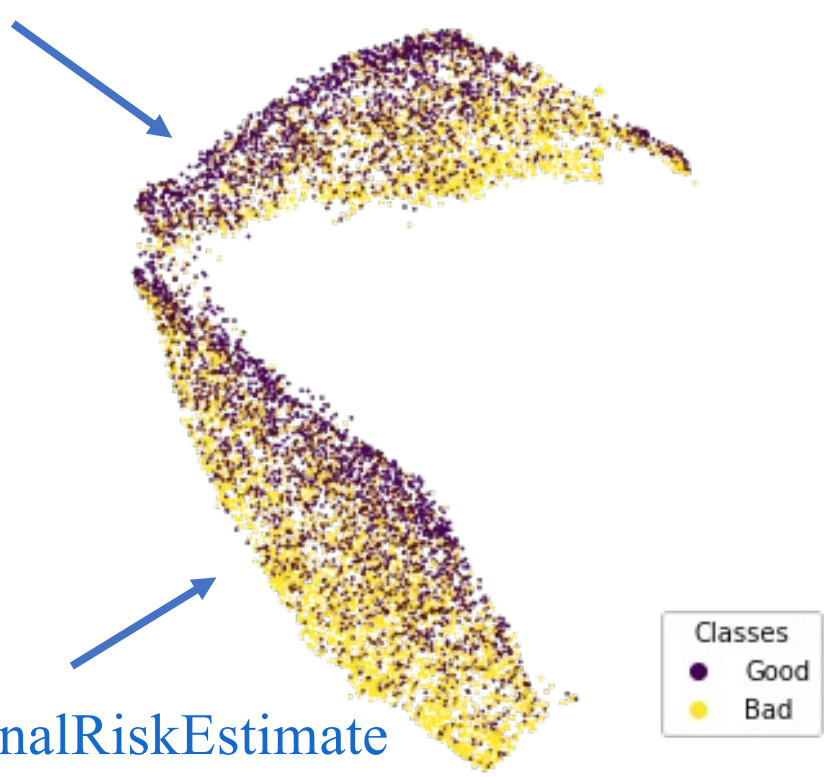


Home Equity Line of Credit (HELOC) Dataset

This competition focuses on an anonymized dataset of Home Equity Line of Credit (HELOC) applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price). The customers in this dataset have requested a credit line in the range of $5,000 - $150,000. The fundamental task is to use the information about the applicant in their credit report to predict whether they will repay their HELOC account within 2 years. This prediction is then used to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.

PaCMAP result on FICO

higher PercentTradesWBalance,
NetFractionRevolvingBurden and
NetFractionInstallBurden

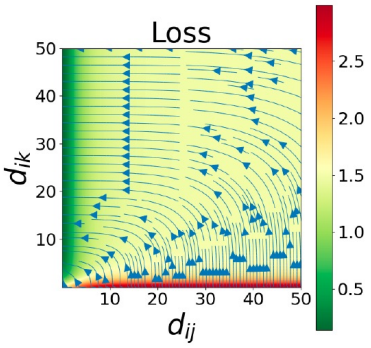


has missing data

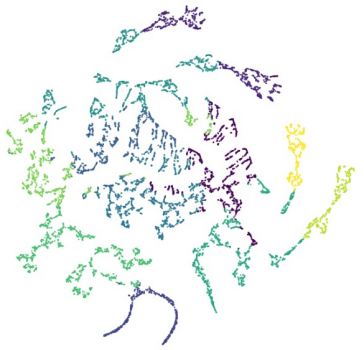
higher ExternalRiskEstimate

credit: Yingfan Wang

Summary

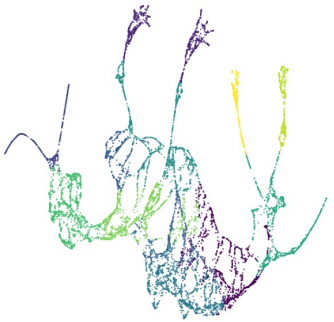


t-SNE (Mammoth)



- What makes DR algorithms succeed/fail?
  - Local structure: A good loss function.
    - Should obey 6 principles.
    - Rainbow plot allows to compare across algorithms.
  - Global structure:
    - We suggest forces on non-neighbors. Mid-near pairs.
- PaCMAP
  - A simpler loss function involving 3 terms.
  - Preserves local and global structure, fast run time.

PaCMAP (Mammoth)



JMLR

Home Page

Papers

Submissions

News

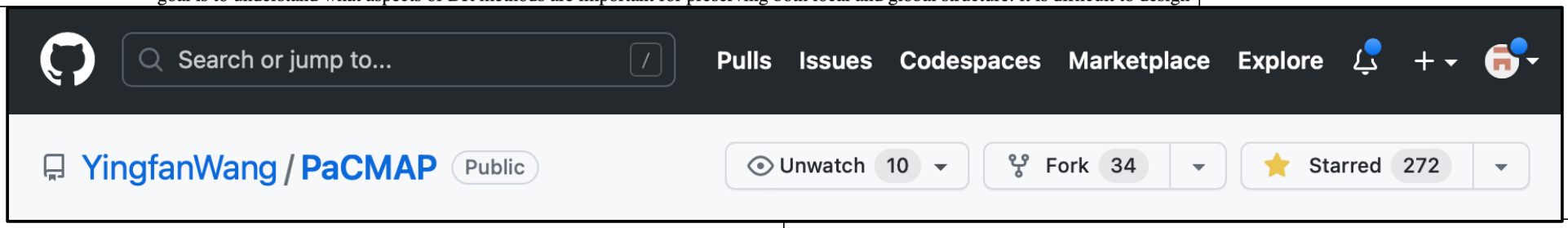
Editorial Board

Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization

Yingfan Wang, Haiyang Huang, Cynthia Rudin, Yaron Shaposhnik; 22(201):1–73, 2021.

Abstract

Dimension reduction (DR) techniques such as t-SNE, UMAP, and TriMap have demonstrated impressive visualization performance on many real-world datasets. One tension that has always faced these methods is the trade-off between preservation of global structure and preservation of local structure: these methods can either handle one or the other, but not both. In this work, our main goal is to understand what aspects of DR methods are important for preserving both local and global structure: it is difficult to design



*Winner of the **2023 John M. Chambers Statistical Software Award** from the American Statistical Association

PaCMAP: https://github.com/YingfanWang/PaCMAP
via pip: pip install pacmap

communications biology

Explore content | About the journal | Publish with us

nature > communications biology > articles > article

Article | Open Access | Published: 19 July 2022

Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization

Haiyang Huang, Yingfan Wang, Cynthia Rudin & Edward P. Browne

Communications Biology **5**, Article number: 719 (2022) | Cite this article

154 Accesses | **2** Altmetric | Metrics