A visual interrogation of dimension reduction tools for single-cell analysis

Tyler J Burns, PhD

AG Mei, DRFZ Berlin

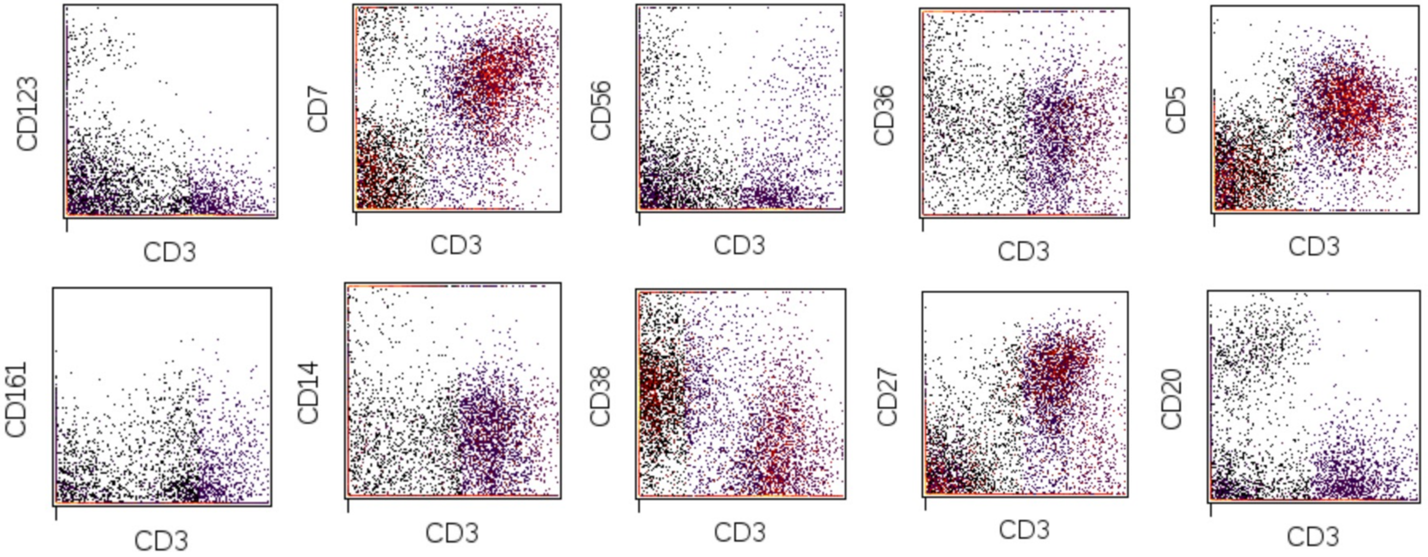
Outline

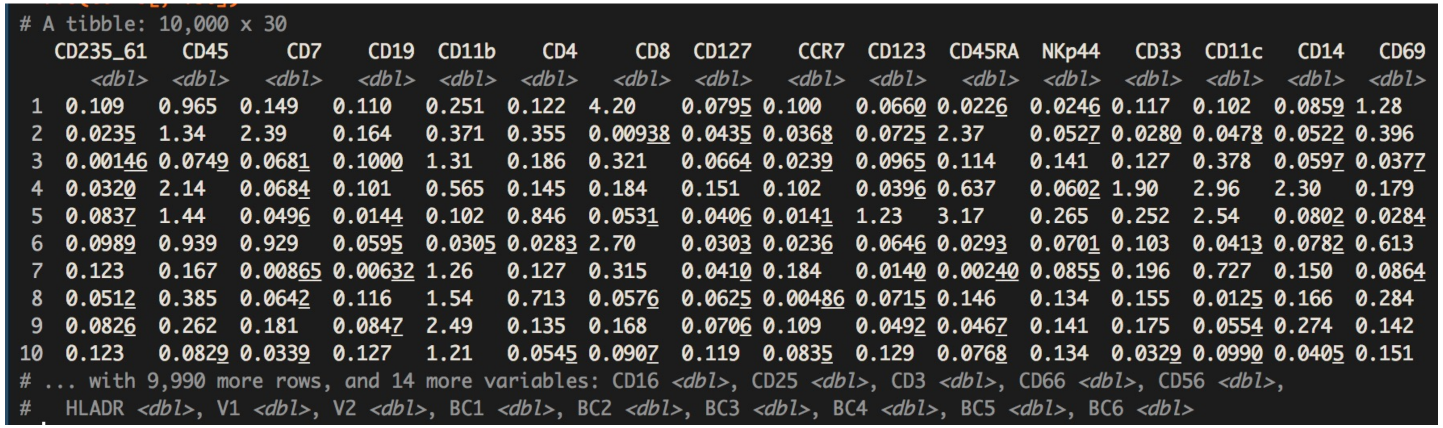
- Part 1: Introduction
- Part 2: Preservation of local structure
- Part 3: Preservation of global structure

Outline

- **Part 1: Introduction**
- Part 2: Preservation of local structure
- Part 3: Preservation of global structure

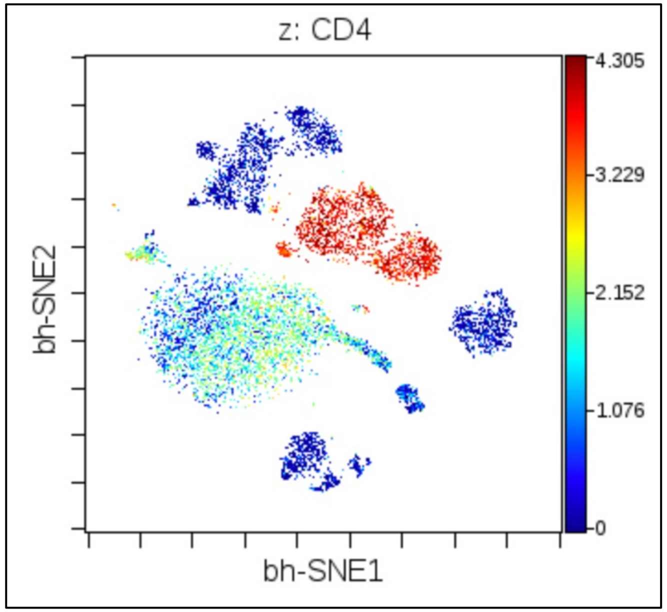
Dimension reduction makes large amounts of information human-readable without too much human work





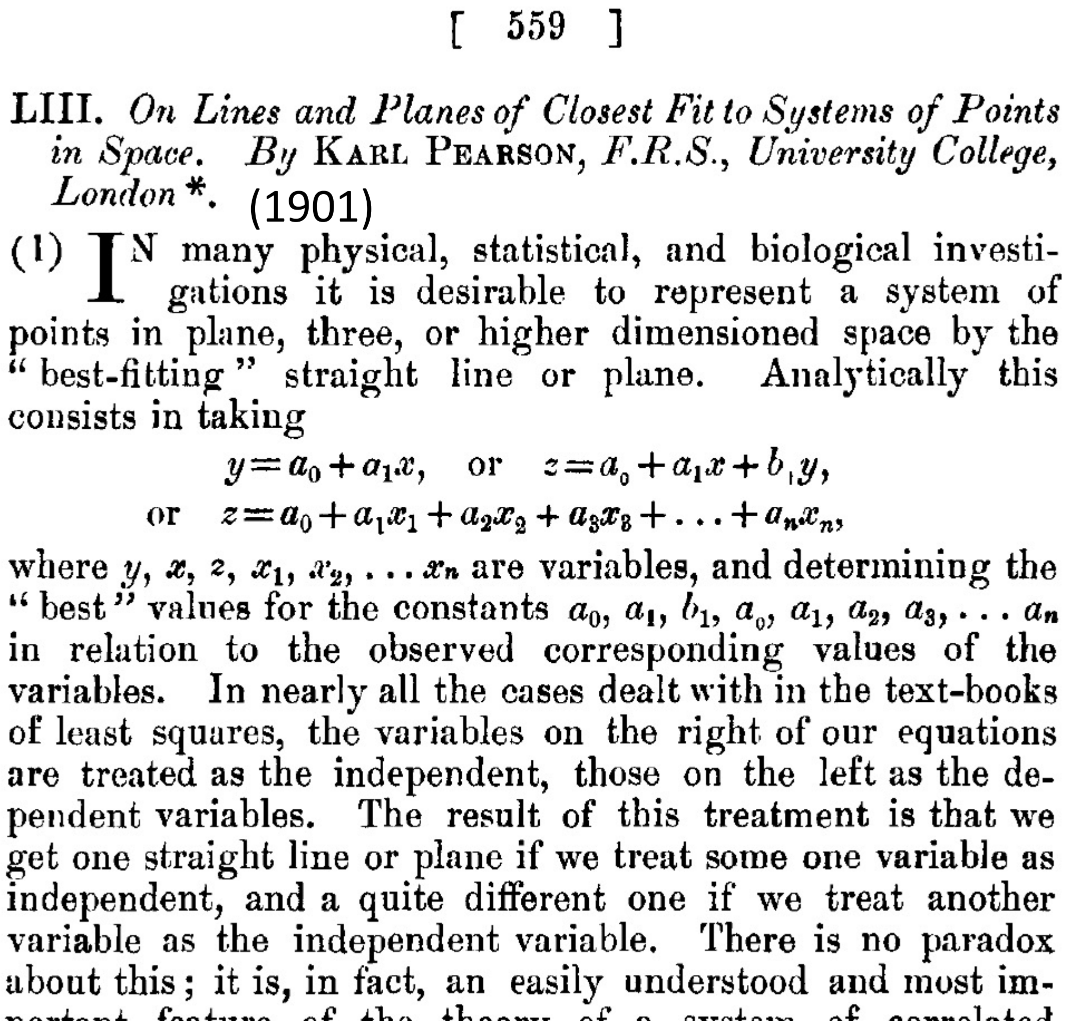
```
# A tibble: 10,000 x 30
   CD235_61   CD45     CD7    CD19  CD11b    CD4     CD8  CD127    CCR7  CD123  CD45RA  NKp44   CD33  CD11c   CD14   CD69
      <dbl>  <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
 1 0.109    0.965  0.149   0.110   0.251  0.122  4.20    0.0795 0.100   0.0660 0.0226  0.0246 0.117  0.102  0.0859 1.28
 2 0.0235   1.34   2.39    0.164   0.371  0.355  0.00938 0.0435 0.0368  0.0725 2.37    0.0527 0.0280 0.0478 0.0522 0.396
 3 0.00146  0.0749 0.0681  0.1000  1.31   0.186  0.321   0.0664 0.0239  0.0965 0.114   0.141  0.127  0.378  0.0597 0.0377
 4 0.0320   2.14   0.0684  0.101   0.565  0.145  0.184   0.151  0.102   0.0396 0.637   0.0602 1.90   2.96   2.30   0.179
 5 0.0837   1.44   0.0496  0.0144  0.102  0.846  0.0531  0.0406 0.0141  1.23   3.17    0.265  0.252  2.54   0.0802 0.0284
 6 0.0989   0.939  0.929   0.0595  0.0305 0.0283 2.70    0.0303 0.0236  0.0646 0.0293  0.0701 0.103  0.0413 0.0782 0.613
 7 0.123    0.167  0.00865 0.00632 1.26   0.127  0.315   0.0410 0.184   0.0140 0.00240 0.0855 0.196  0.727  0.150  0.0864
 8 0.0512   0.385  0.0642  0.116   1.54   0.713  0.0576  0.0625 0.00486 0.0715 0.146   0.134  0.155  0.0125 0.166  0.284
 9 0.0826   0.262  0.181   0.0847  2.49   0.135  0.168   0.0706 0.109   0.0492 0.0467  0.141  0.175  0.0554 0.274  0.142
10 0.123    0.0829 0.0339  0.127   1.21   0.0545 0.0907  0.119  0.0835  0.129  0.0768  0.134  0.0329 0.0990 0.0405 0.151
# ... with 9,990 more rows, and 14 more variables: CD16 <dbl>, CD25 <dbl>, CD3 <dbl>, CD66 <dbl>, CD56 <dbl>,
#   HLADR <dbl>, V1 <dbl>, V2 <dbl>, BC1 <dbl>, BC2 <dbl>, BC3 <dbl>, BC4 <dbl>, BC5 <dbl>, BC6 <dbl>
```

t-SNE(cells)



Amir et al, *Nat Biotechnology* 2013

Early dimension reduction tools: Principal Component Analysis (PCA)



[559]

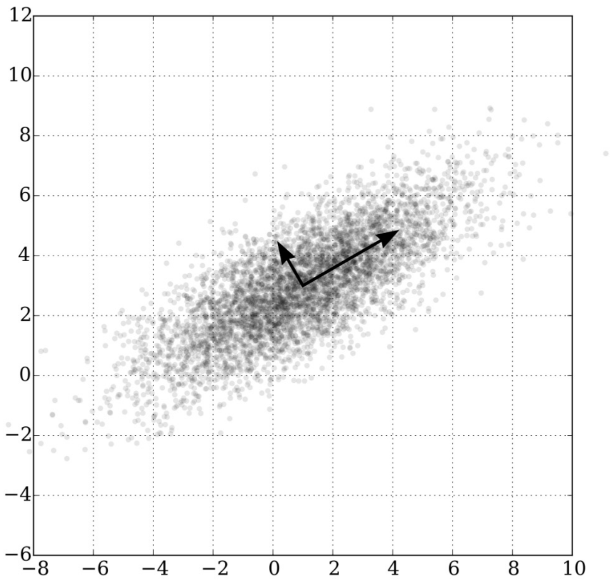
LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* *By* KARL PEARSON, *F.R.S., University College, London**. (1901)

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y=a_0+a_1x, \quad \text{or} \quad z=a_0+a_1x+b_1y,$$
$$\text{or} \quad z=a_0+a_1x_1+a_2x_2+a_3x_3+\dots+a_nx_n,$$

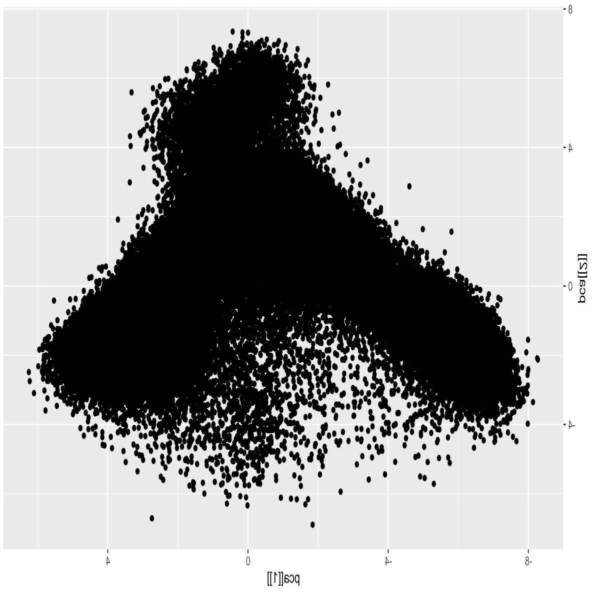
where $y, x, z, x_1, x_2, \dots x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \dots a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books of least squares, the variables on the right of our equations are treated as the independent, those on the left as the dependent variables. The result of this treatment is that we get one straight line or plane if we treat some one variable as independent, and a quite different one if we treat another variable as the independent variable. There is no paradox about this; it is, in fact, an easily understood and most important feature of the theory of a system of correlated

Axes span the direction with highest variance

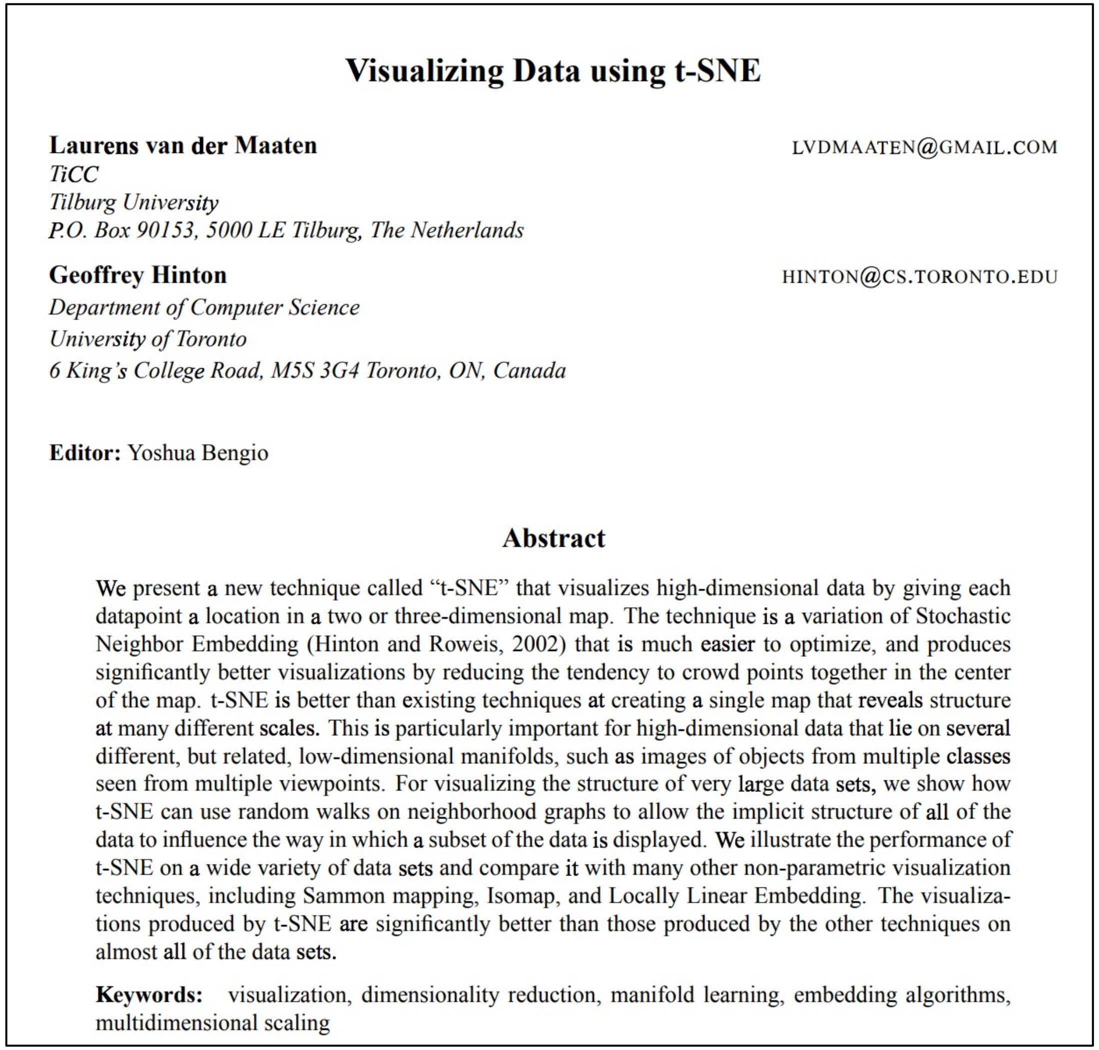


https://en.wikipedia.org/wiki/Principal_component_analysis

Samusik_01 bone marrow CyTOF dataset



t-SNE preserves local information, produces more well clustered maps



Visualizing Data using t-SNE

Laurens van der Maaten LVDMAATEN@GMAIL.COM
TiCC
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

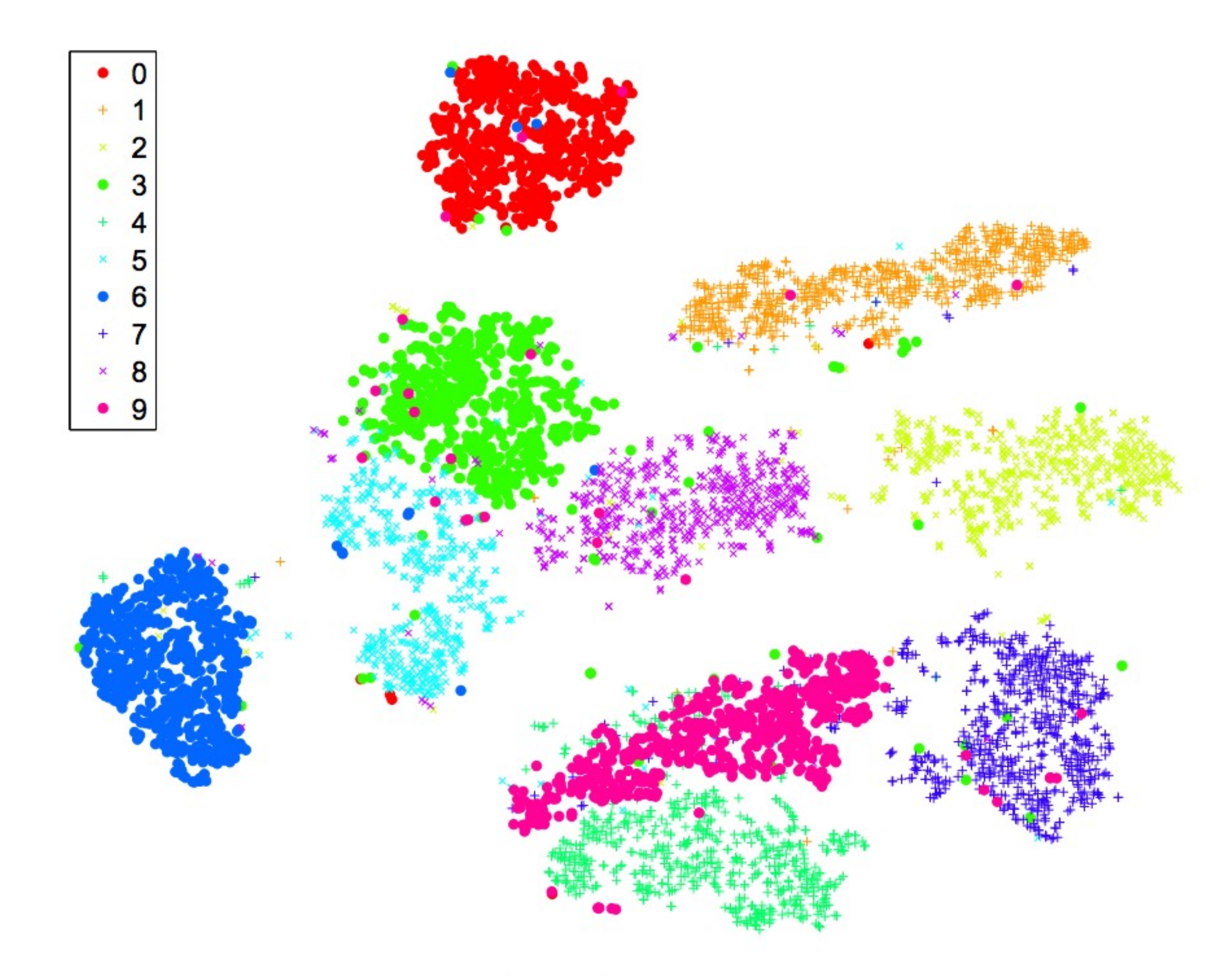
Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

Editor: Yoshua Bengio

Abstract

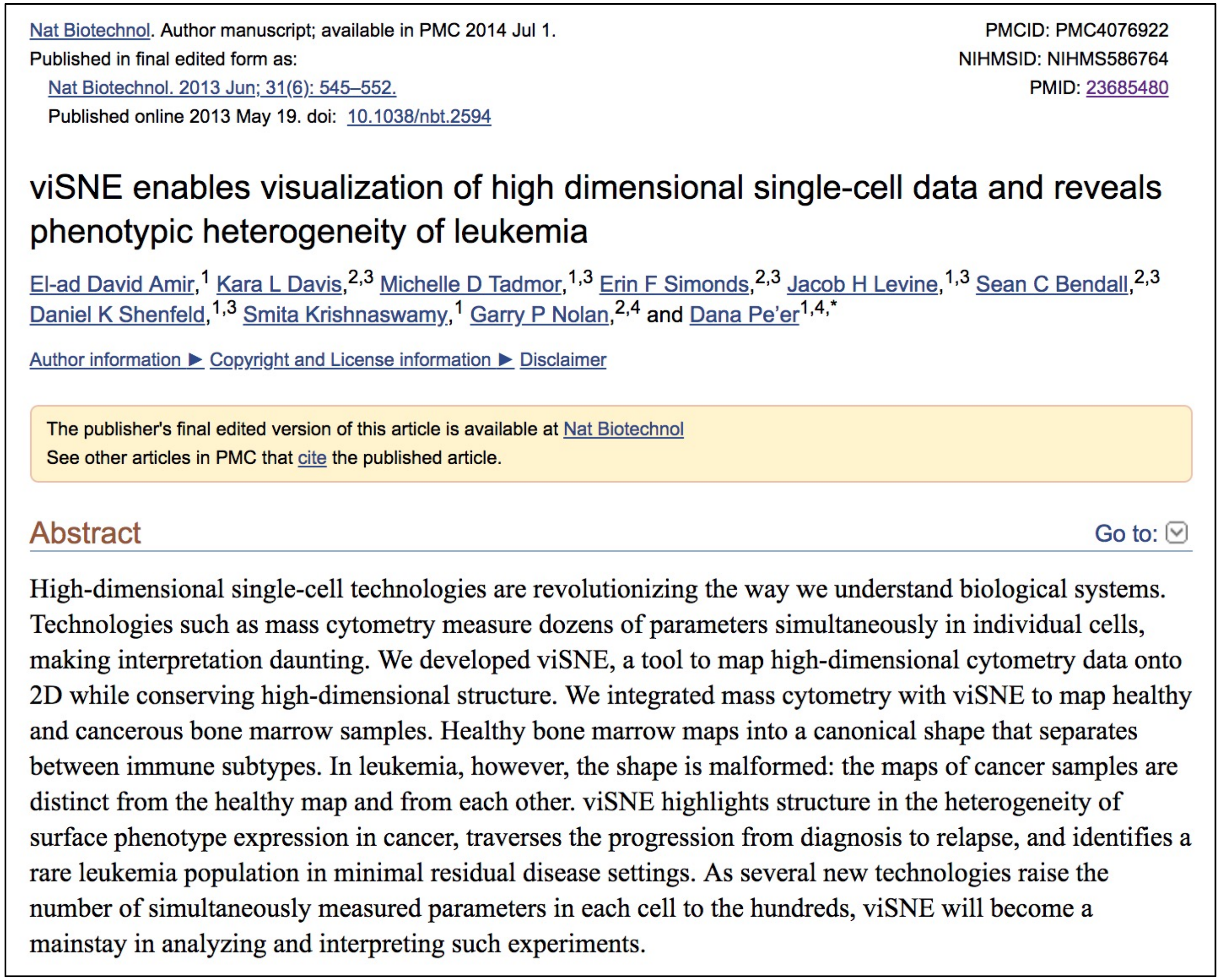
We present a new technique called "t-SNE" that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map. The technique is a variation of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) that is much easier to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map. t-SNE is better than existing techniques at creating a single map that reveals structure at many different scales. This is particularly important for high-dimensional data that lie on several different, but related, low-dimensional manifolds, such as images of objects from multiple classes seen from multiple viewpoints. For visualizing the structure of very large data sets, we show how t-SNE can use random walks on neighborhood graphs to allow the implicit structure of all of the data to influence the way in which a subset of the data is displayed. We illustrate the performance of t-SNE on a wide variety of data sets and compare it with many other non-parametric visualization techniques, including Sammon mapping, Isomap, and Locally Linear Embedding. The visualizations produced by t-SNE are significantly better than those produced by the other techniques on almost all of the data sets.

Keywords: visualization, dimensionality reduction, manifold learning, embedding algorithms, multidimensional scaling



(a) Visualization by t-SNE.

viSNE: the adaptation of t-SNE to CyTOF



Nat Biotechnol. Author manuscript; available in PMC 2014 Jul 1.
Published in final edited form as:
Nat Biotechnol. 2013 Jun; 31(6): 545–552.
Published online 2013 May 19. doi: 10.1038/nbt.2594

PMCID: PMC4076922
NIHMSID: NIHMS586764
PMID: 23685480

viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia

El-ad David Amir,[1] Kara L Davis,[2,3] Michelle D Tadmor,[1,3] Erin F Simonds,[2,3] Jacob H Levine,[1,3] Sean C Bendall,[2,3] Daniel K Shenfeld,[1,3] Smita Krishnaswamy,[1] Garry P Nolan,[2,4] and Dana Pe'er[1,4,*]

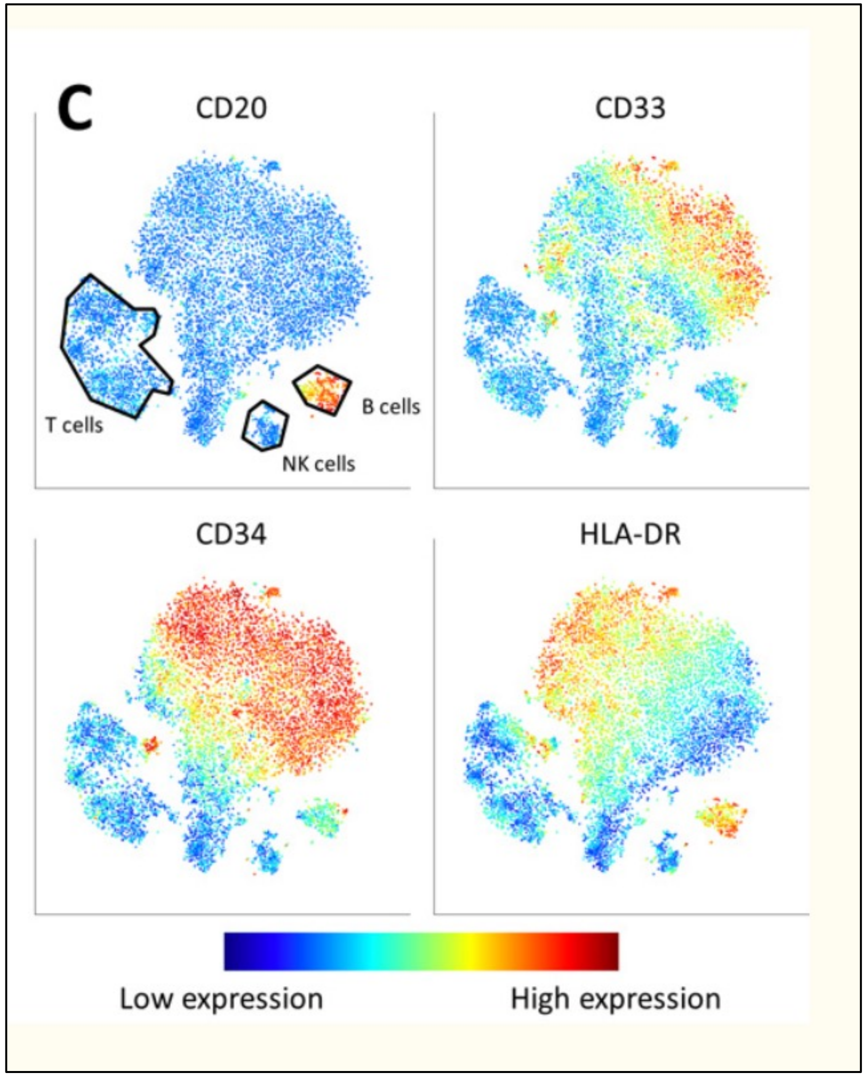
Author information ▶ Copyright and License information ▶ Disclaimer

The publisher's final edited version of this article is available at Nat Biotechnol
See other articles in PMC that cite the published article.

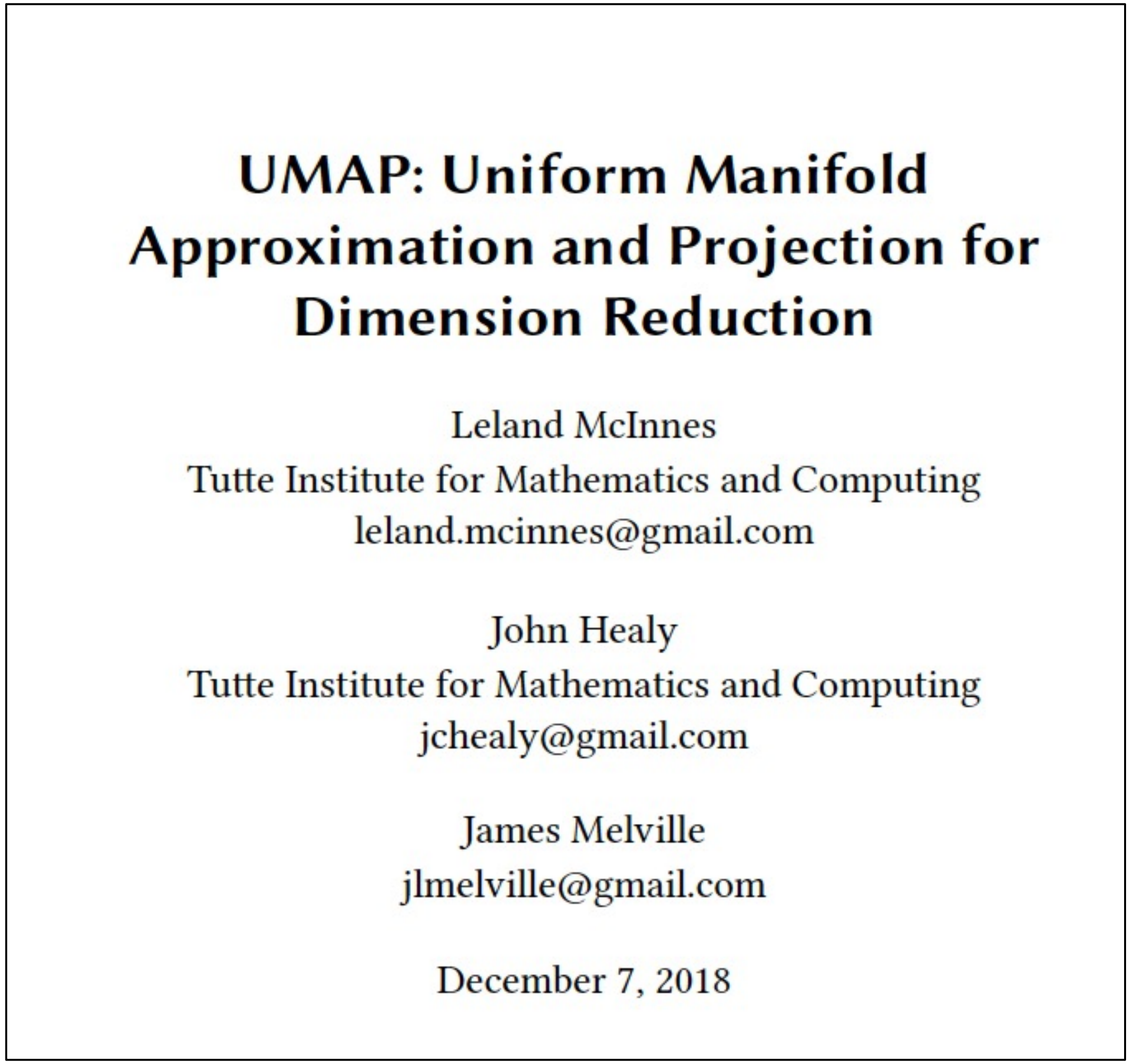
Abstract

Go to: ☐

High-dimensional single-cell technologies are revolutionizing the way we understand biological systems. Technologies such as mass cytometry measure dozens of parameters simultaneously in individual cells, making interpretation daunting. We developed viSNE, a tool to map high-dimensional cytometry data onto 2D while conserving high-dimensional structure. We integrated mass cytometry with viSNE to map healthy and cancerous bone marrow samples. Healthy bone marrow maps into a canonical shape that separates between immune subtypes. In leukemia, however, the shape is malformed: the maps of cancer samples are distinct from the healthy map and from each other. viSNE highlights structure in the heterogeneity of surface phenotype expression in cancer, traverses the progression from diagnosis to relapse, and identifies a rare leukemia population in minimal residual disease settings. As several new technologies raise the number of simultaneously measured parameters in each cell to the hundreds, viSNE will become a mainstay in analyzing and interpreting such experiments.



Emergence of UMAP as an alternative to t-SNE for single-cell analysis



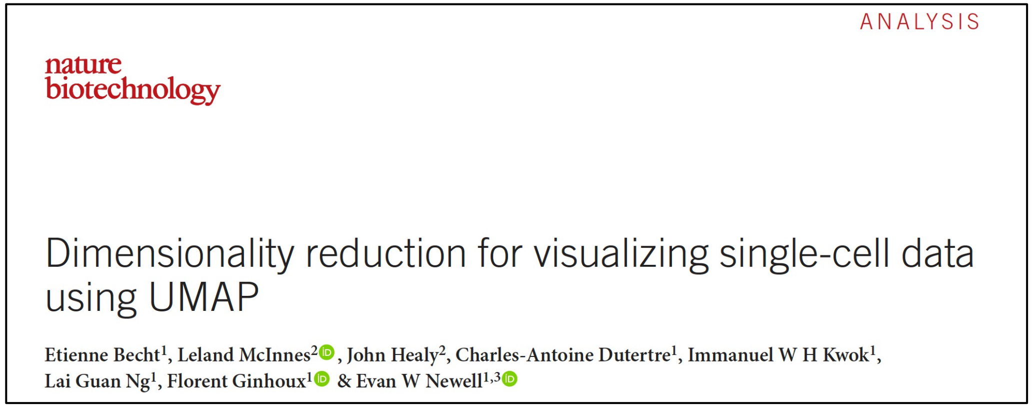
UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Leland McInnes
Tutte Institute for Mathematics and Computing
leland.mcinnes@gmail.com

John Healy
Tutte Institute for Mathematics and Computing
jchealy@gmail.com

James Melville
jlmelville@gmail.com

December 7, 2018

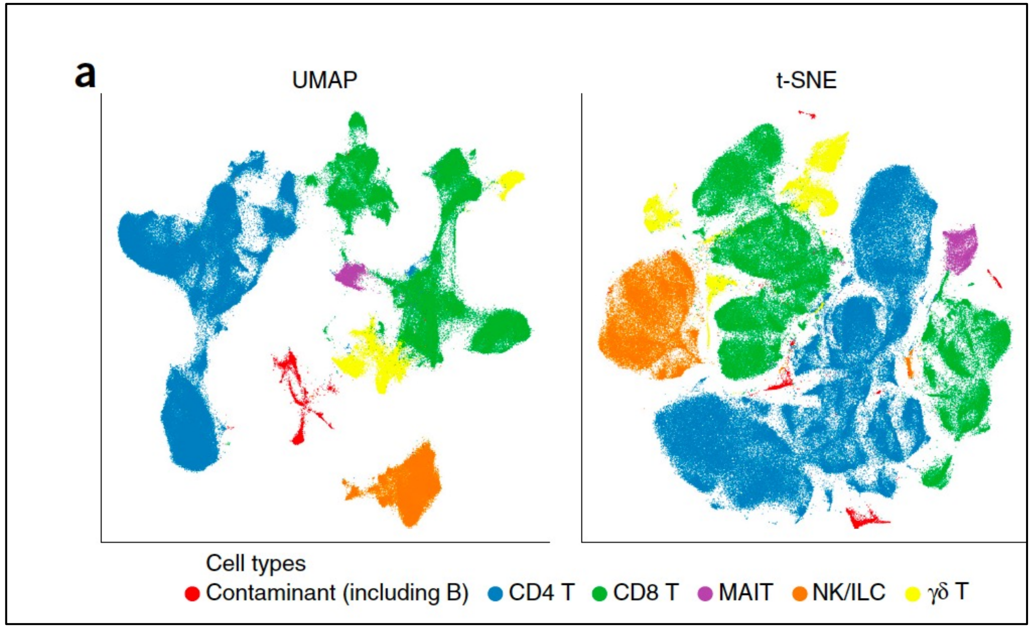


ANALYSIS

nature biotechnology

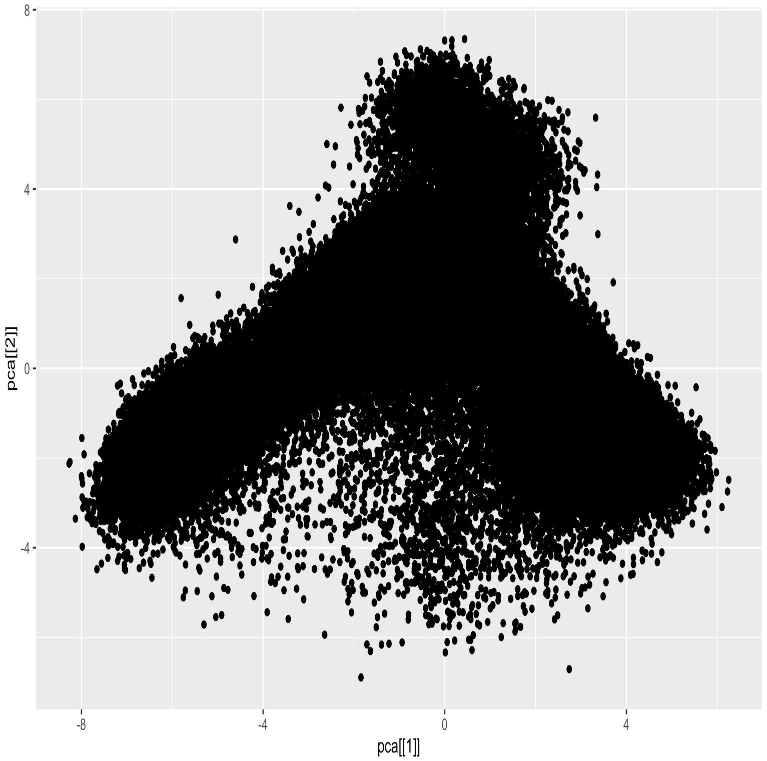
Dimensionality reduction for visualizing single-cell data using UMAP

Etienne Becht[1], Leland McInnes[2], John Healy[2], Charles-Antoine Dutertre[1], Immanuel W H Kwok[1], Lai Guan Ng[1], Florent Ginhoux[1] & Evan W Newell[1,3]

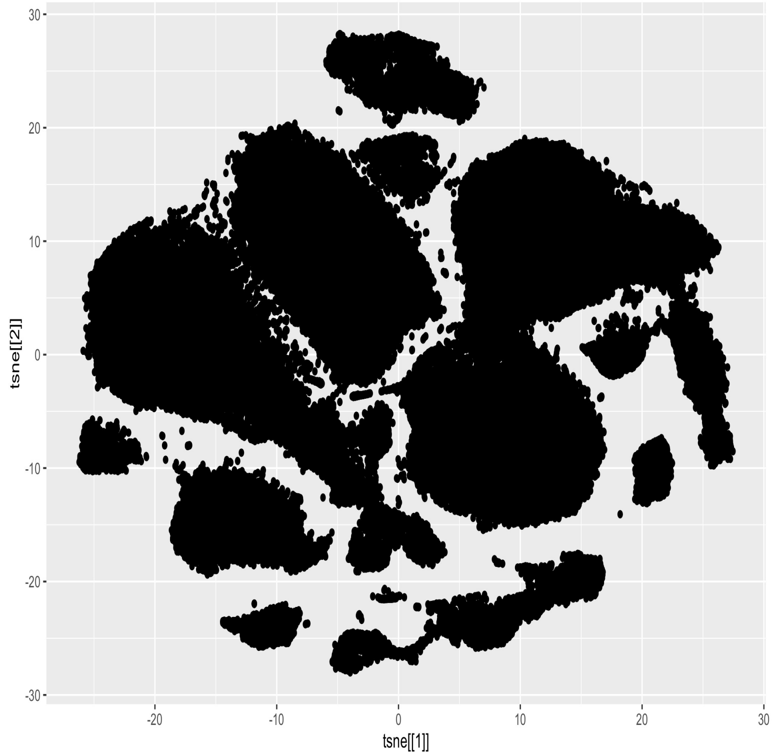


What PCA, t-SNE and UMAP look like on a bone marrow CyTOF dataset

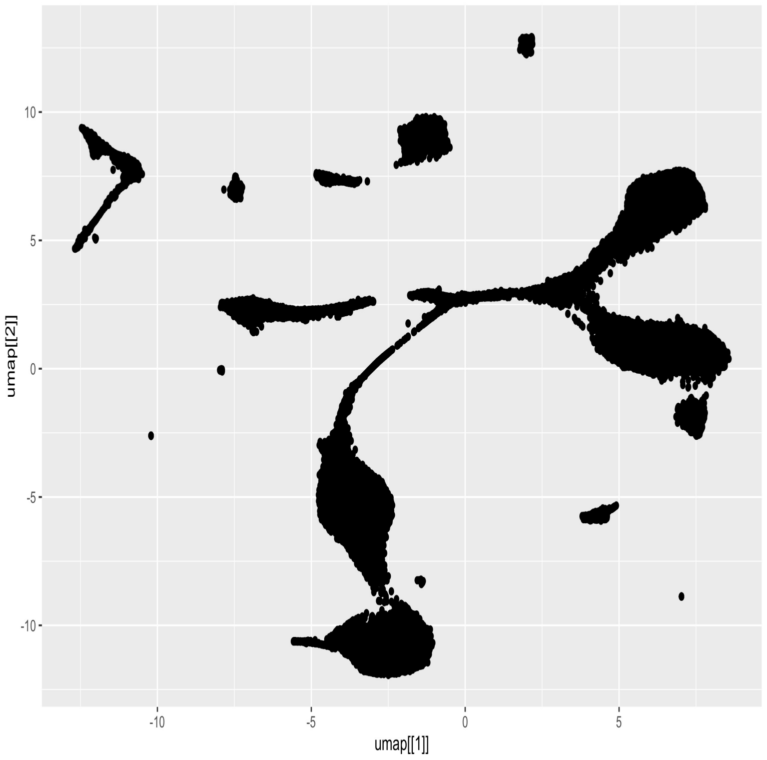
PCA



t-SNE



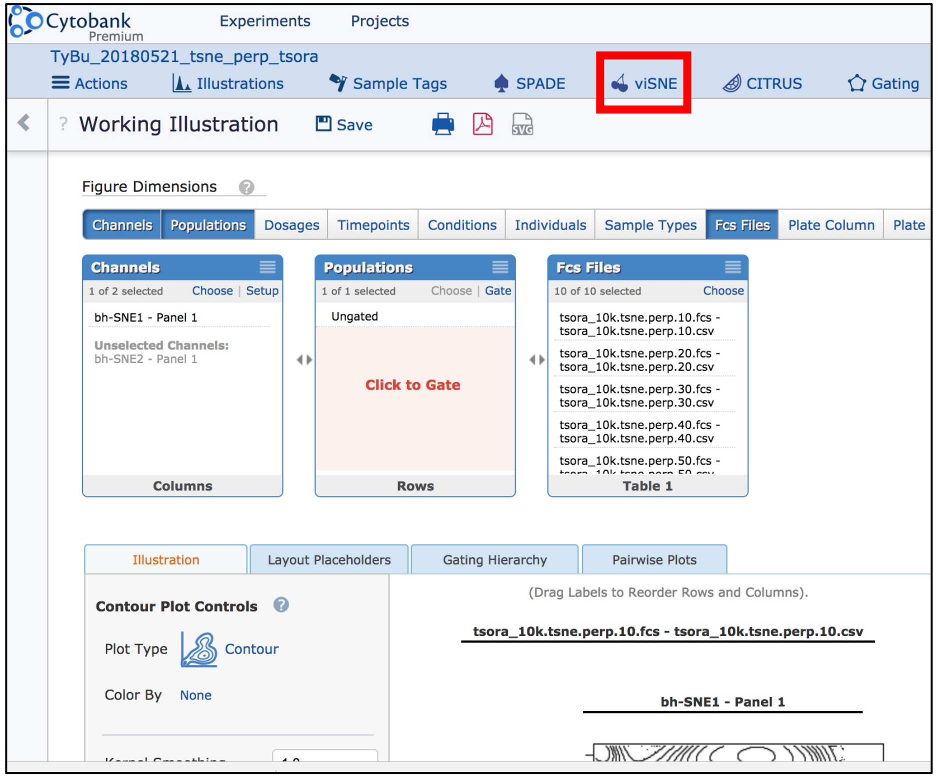
UMAP



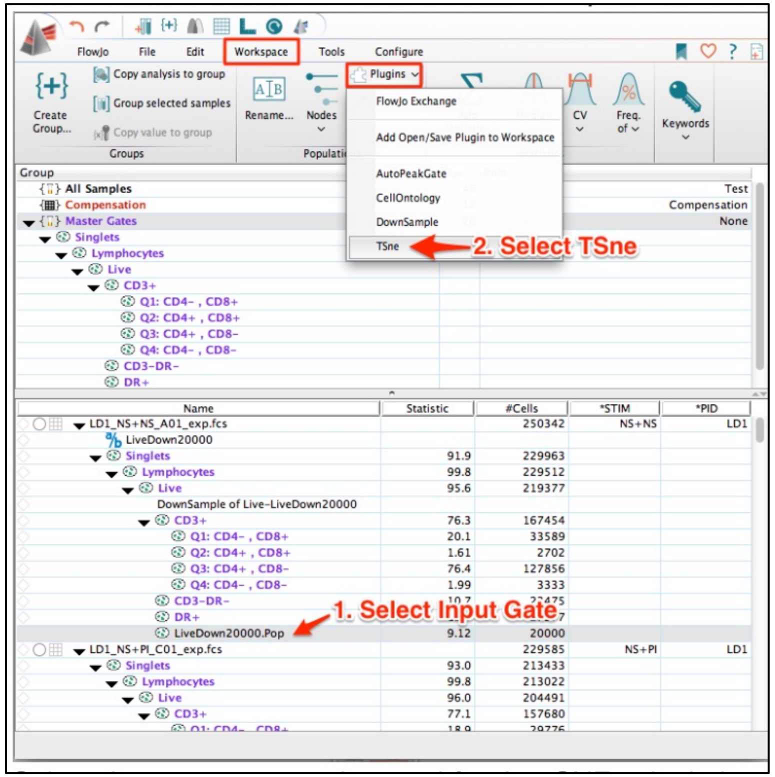
Dataset: Samusik *et al*, *Nature Methods* 2016

t-SNE and UMAP are accessible from single-cell analysis user interfaces

Cytobank



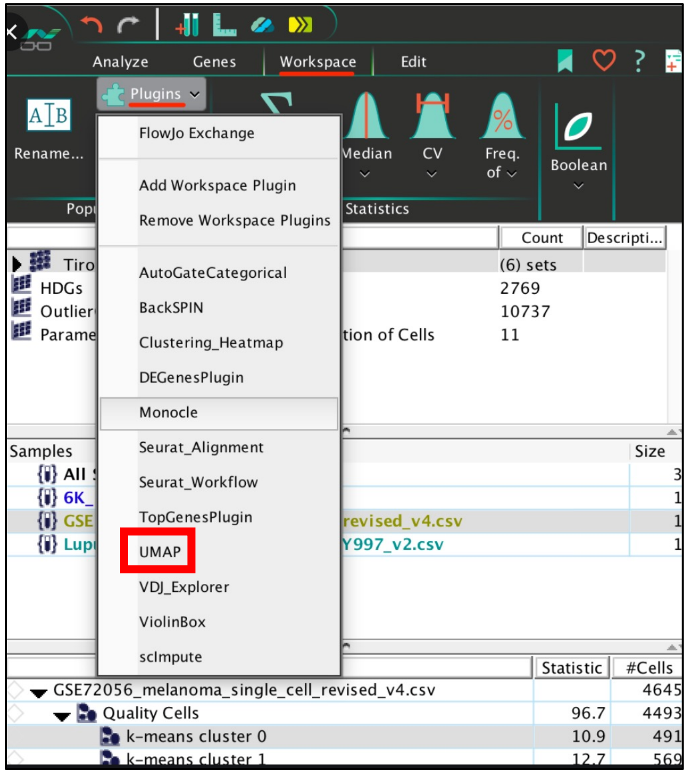
FlowJo



2. Select TSne

Name	Statistic	#Cells	*STIM	*PID
LD1_NS+NS_A01_exp.fcs		250342	NS+NS	LD1
LiveDown20000				
Singlets	91.9	229963		
Lymphocytes	99.8	229512		
Live	95.6	219377		
DownSample of Live-LiveDown20000				
CD3+	76.3	167454		
Q1: CD4- , CD8+	20.1	33589		
Q2: CD4+ , CD8+	1.61	2702		
Q3: CD4+ , CD8-	76.4	127856		
Q4: CD4- , CD8-	1.99	3333		
CD3-DR-	10.7	22475		
DR+	[illegible]	[illegible]		
LiveDown20000.Pop	9.12	20000		
LD1_NS+PI_C01_exp.fcs		229585	NS+PI	LD1
Singlets	93.0	213433		
Lymphocytes	99.8	213022		
Live	96.0	204491		
CD3+	77.1	157680		
Q1: CD4- , CD8+	18.9	29776		

1. Select Input Gate

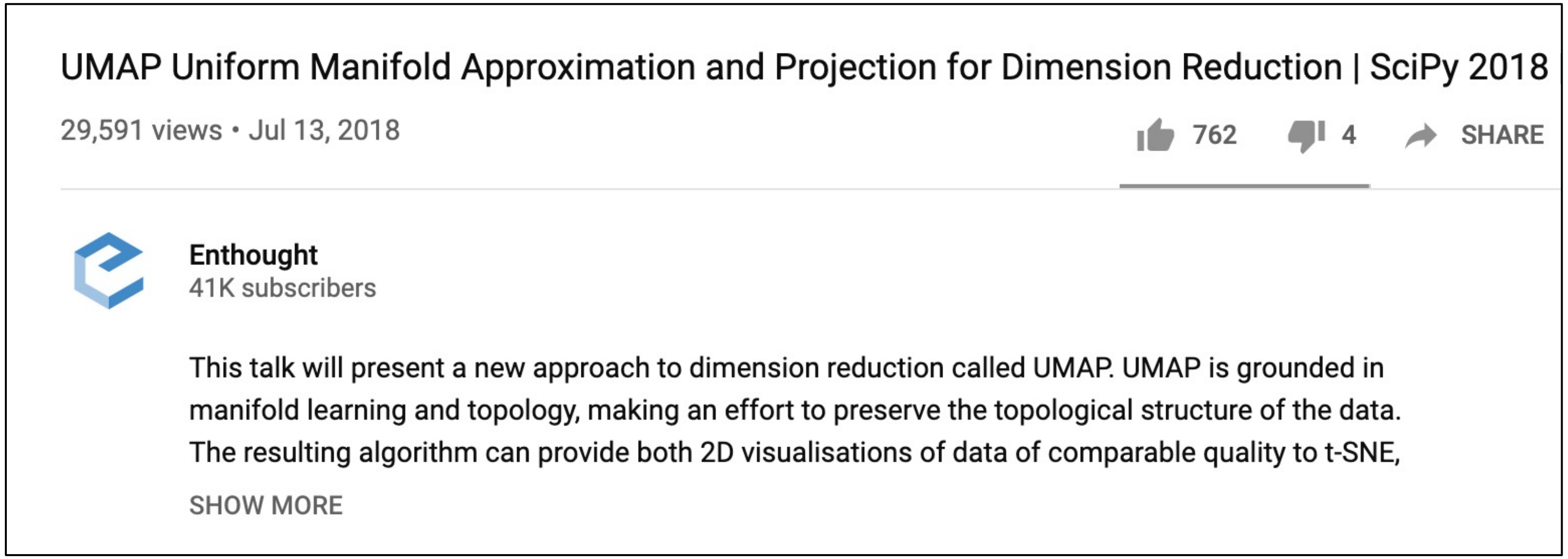


	Statistic	#Cells
GSE72056_melanoma_single_cell_revised_v4.csv		4645
Quality Cells	96.7	4493
k-means cluster 0	10.9	491
k-means cluster 1	12.7	569

Others: Astrolabe, OMIQ, Tercen

What additional information about dimension reduction maps should we know for their proper use?

Credit for the following t-SNE and UMAP explanations

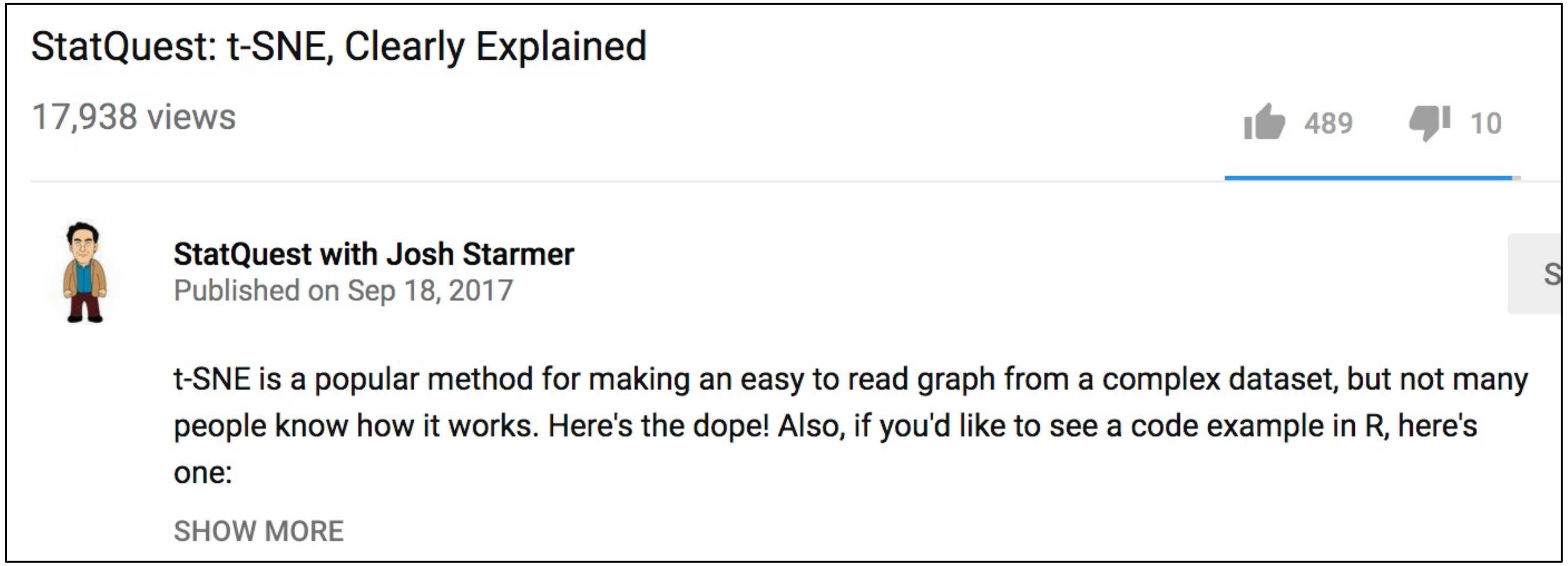


UMAP Uniform Manifold Approximation and Projection for Dimension Reduction | SciPy 2018

29,591 views • Jul 13, 2018

Enthought
41K subscribers

This talk will present a new approach to dimension reduction called UMAP. UMAP is grounded in manifold learning and topology, making an effort to preserve the topological structure of the data. The resulting algorithm can provide both 2D visualisations of data of comparable quality to t-SNE,



StatQuest: t-SNE, Clearly Explained

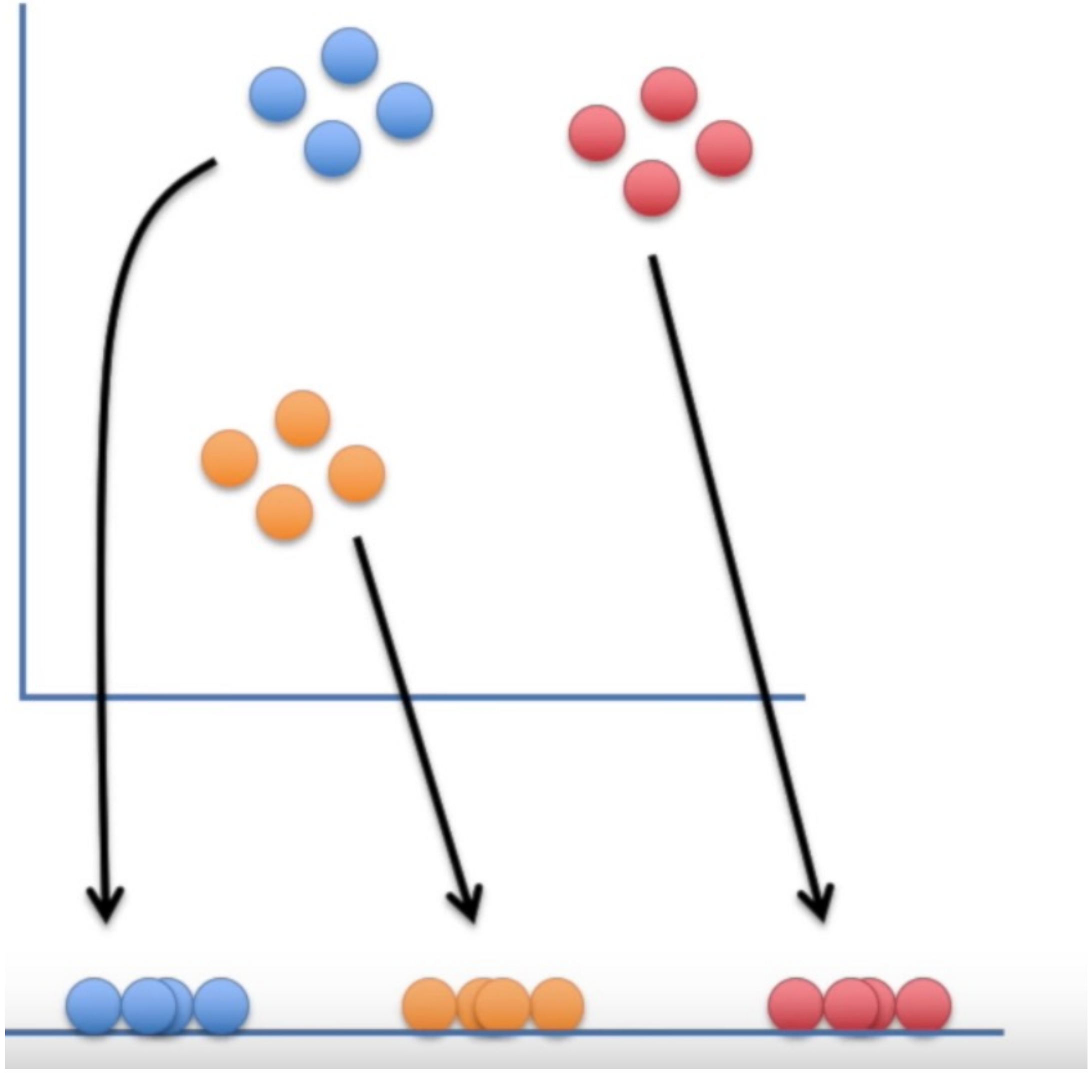
17,938 views

StatQuest with Josh Starmer
Published on Sep 18, 2017

t-SNE is a popular method for making an easy to read graph from a complex dataset, but not many people know how it works. Here's the dope! Also, if you'd like to see a code example in R, here's one:

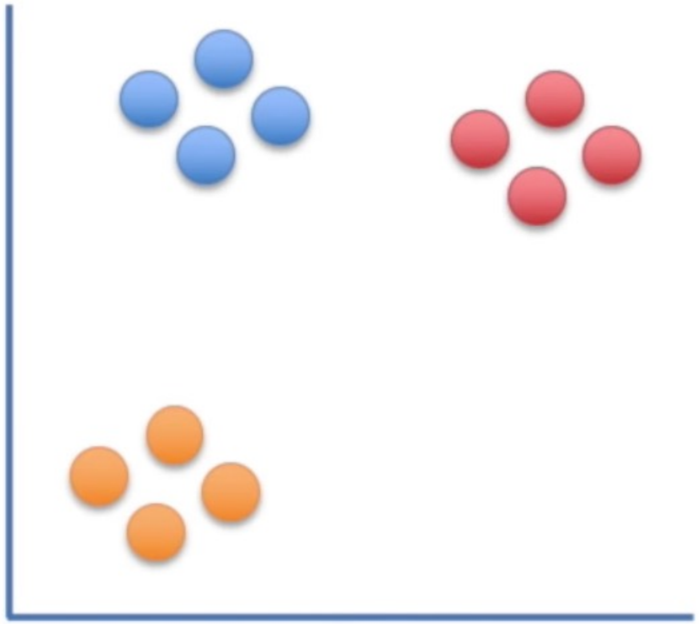
The goal of t-SNE and UMAP is to reduce dimensions while preserving specific information about each cell’s neighbors

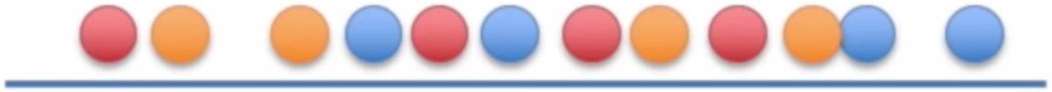
Higher dimensional space



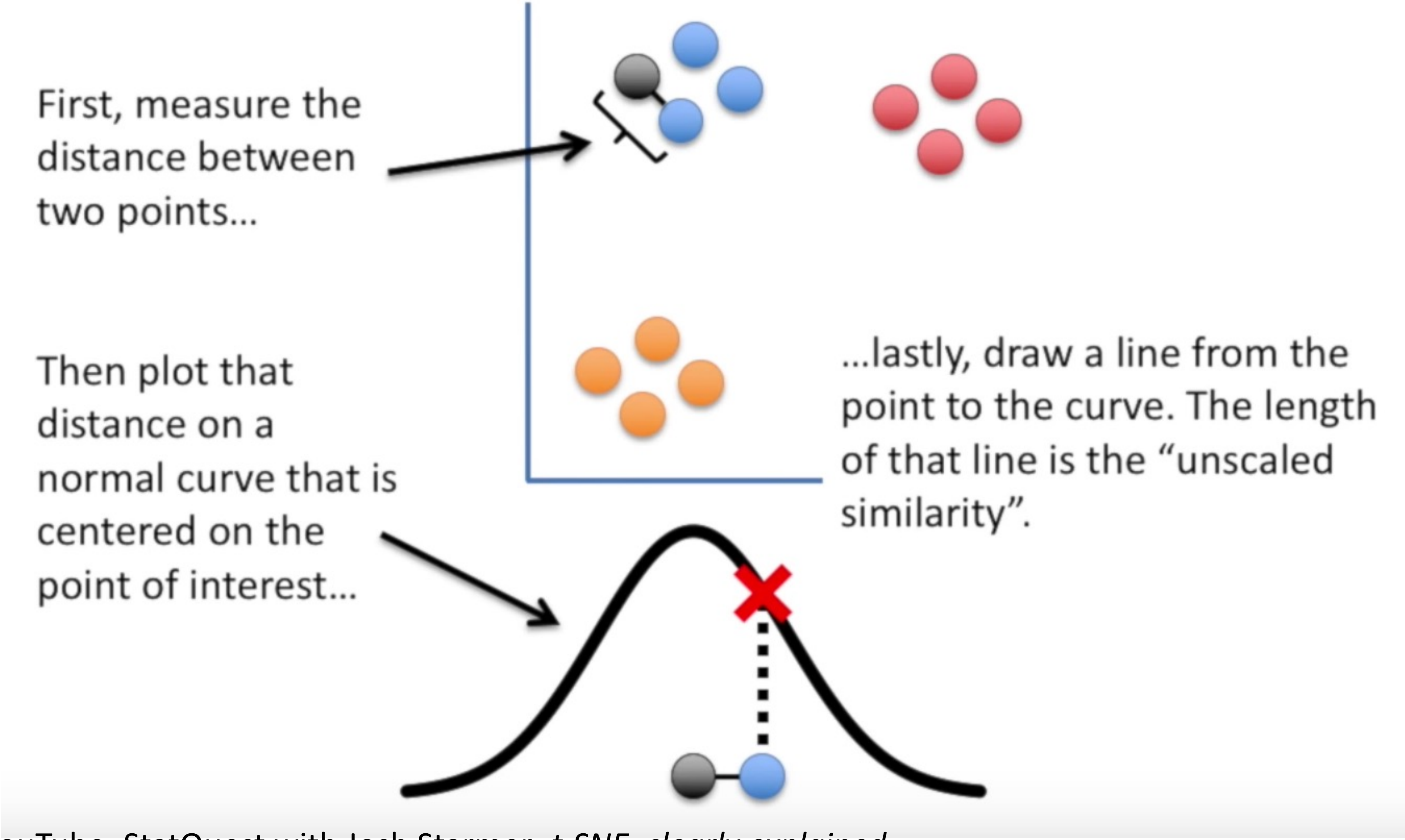
Low dimensional embedding

t-SNE and UMAP start with a low-dimensional embedding of randomly placed points





t-SNE weights its neighbors based on distance fitted to a distribution



First, measure the distance between two points…

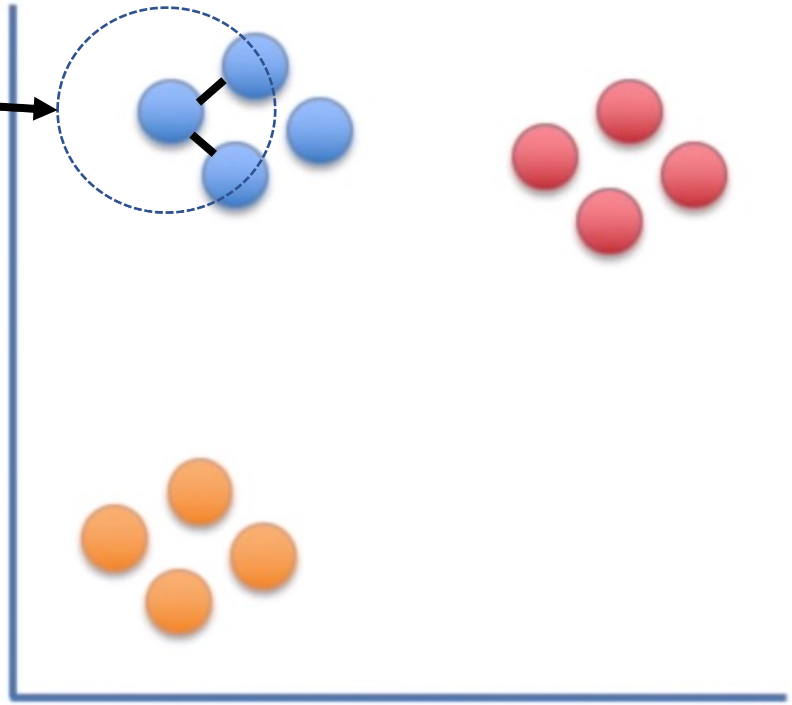
Then plot that distance on a normal curve that is centered on the point of interest…

…lastly, draw a line from the point to the curve. The length of that line is the “unscaled similarity”.

Image source: YouTube: StatQuest with Josh Starmer: *t-SNE, clearly explained*

UMAP weights its neighbors based on topology

Find the diameter
of a neighborhood.
Think of it like a ball.



Find the probability that
a 1-simplex exists between
two points in a neighborhood

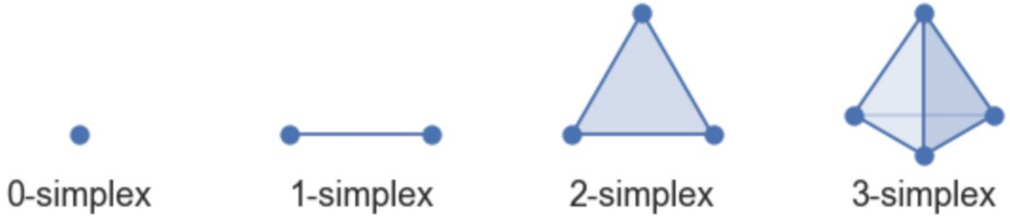


Image source: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

Resulting neighbor graph is a
bunch of simplexes glued together
(simplicial complex). Simpler structure
but preserves topological information.

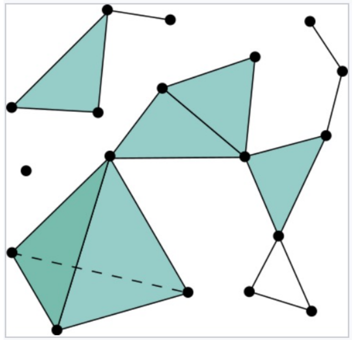
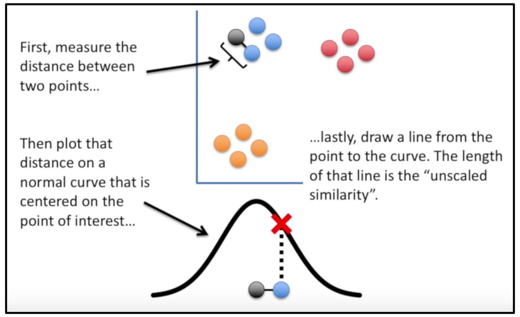


Image source: https://en.wikipedia.org/wiki/Simplicial_complex

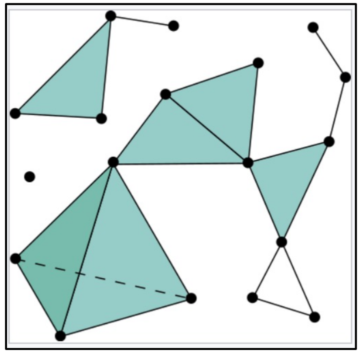
The weighted neighborhood graphs can be represented as similarity matrices

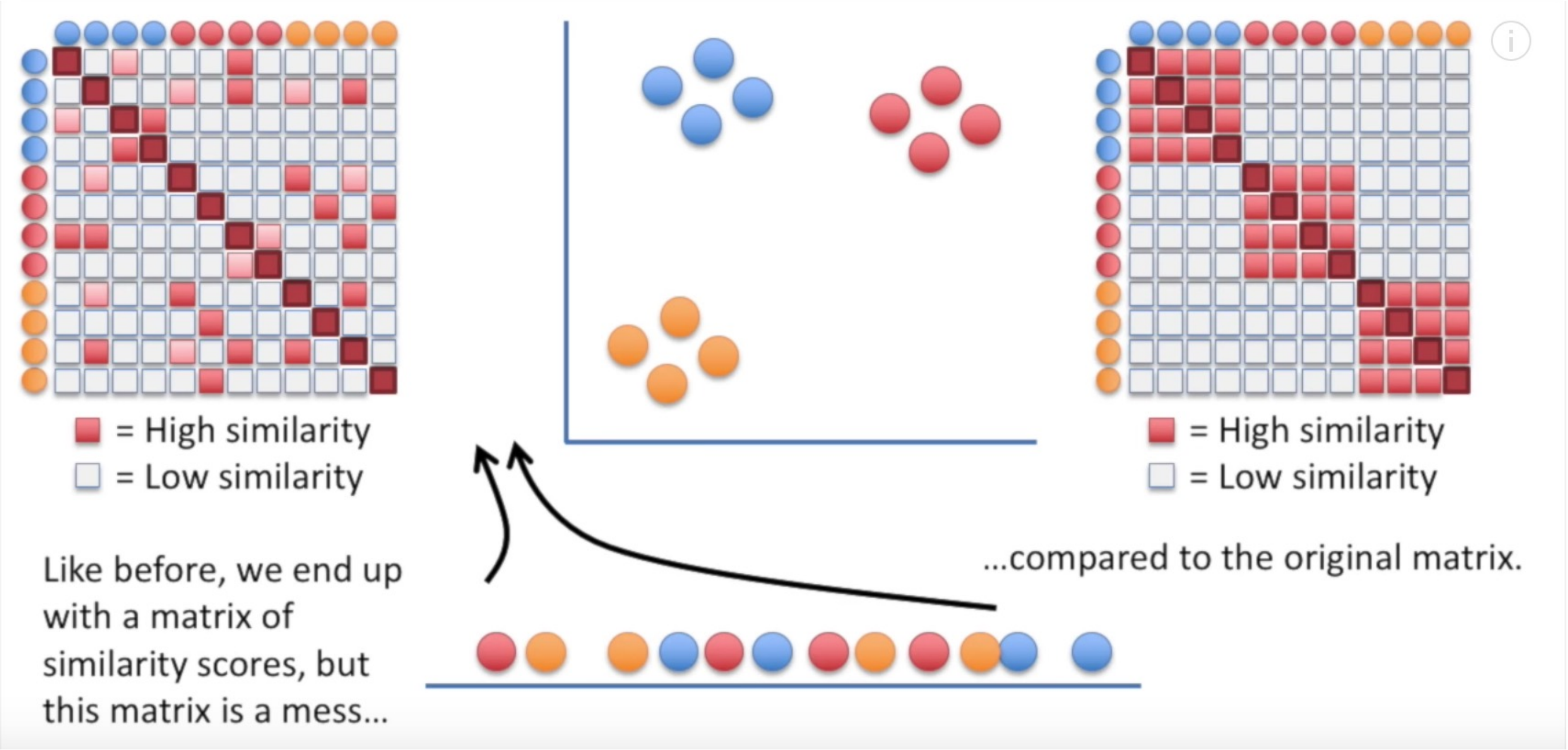


First, measure the distance between two points...

Then plot that distance on a normal curve that is centered on the point of interest...

...lastly, draw a line from the point to the curve. The length of that line is the "unscaled similarity".





🟥 = High similarity

□ = Low similarity

Like before, we end up with a matrix of similarity scores, but this matrix is a mess...

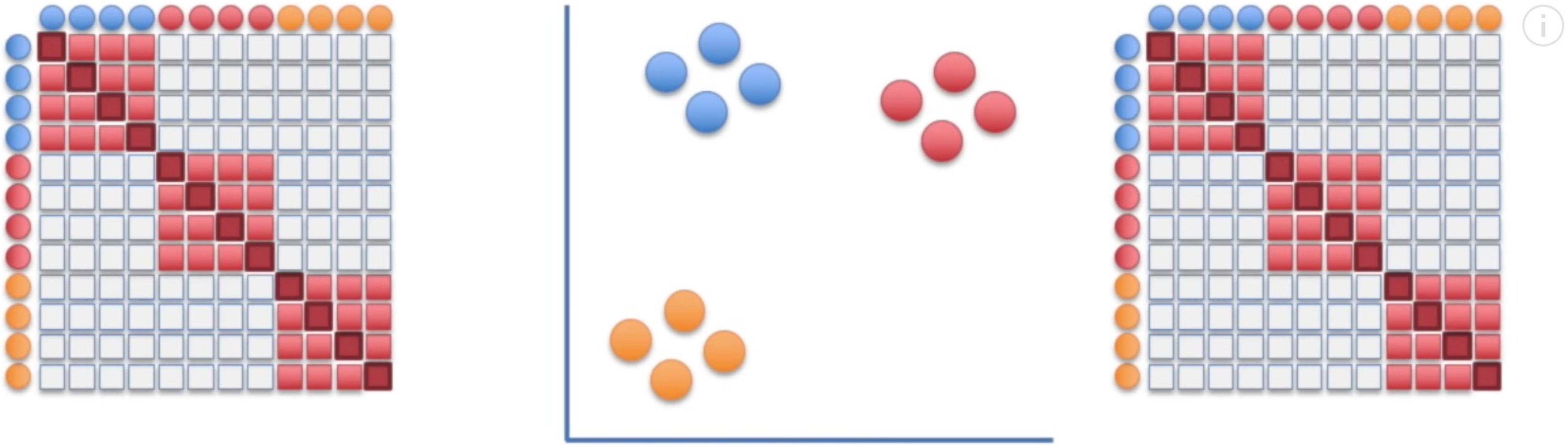
...compared to the original matrix.

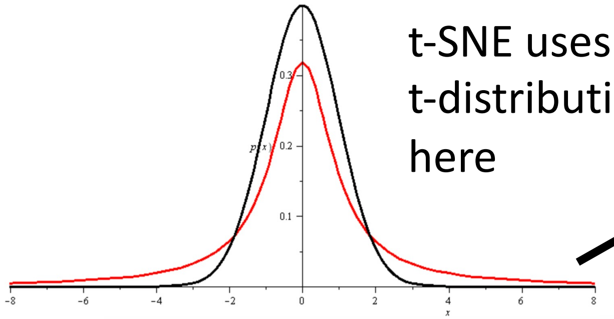
🟥 = High similarity

□ = Low similarity

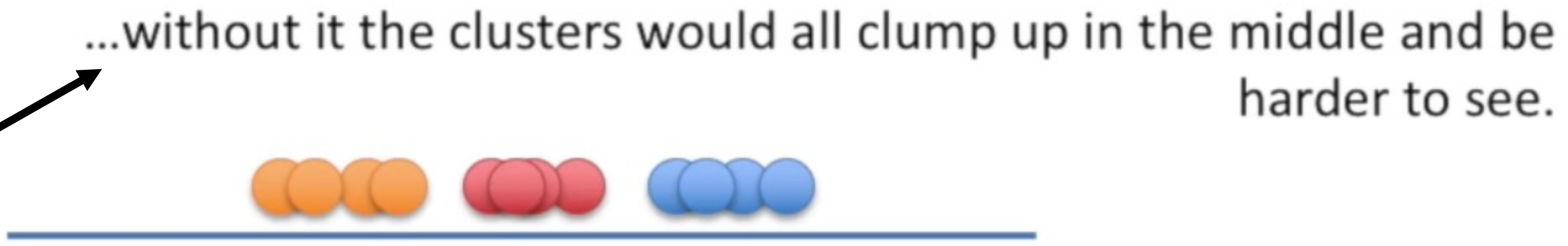
Image source: YouTube: StatQuest with Josh Starmer: *t-SNE, clearly explained*

Make these similarity matrices as similar to each other as possible, and then you’re done





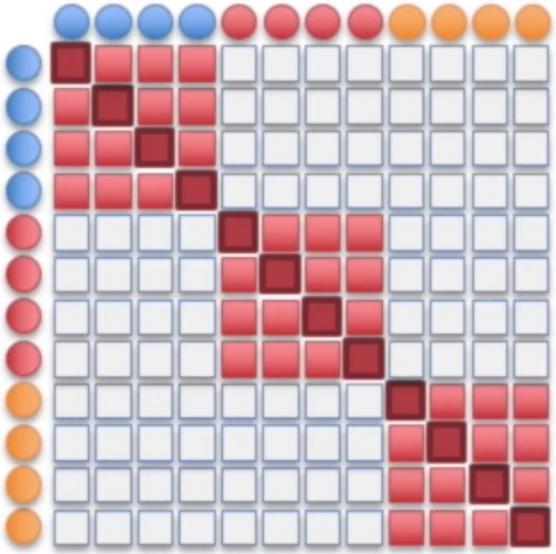
t-SNE uses a t-distribution here

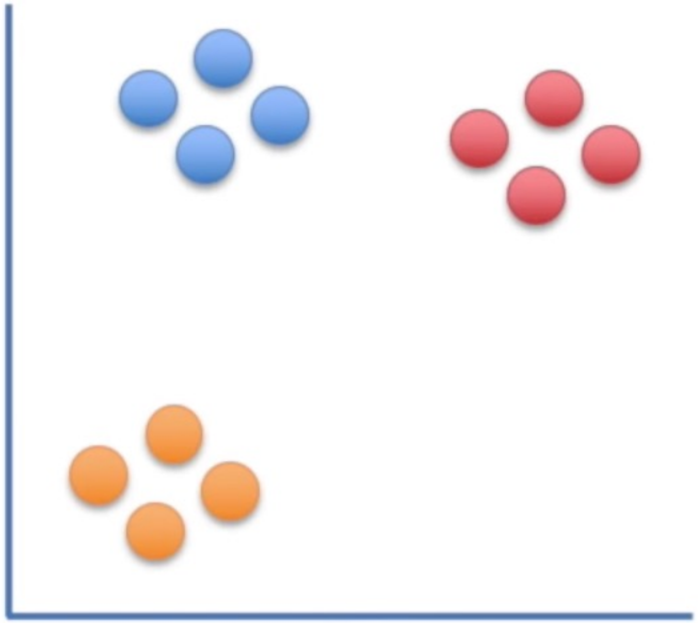


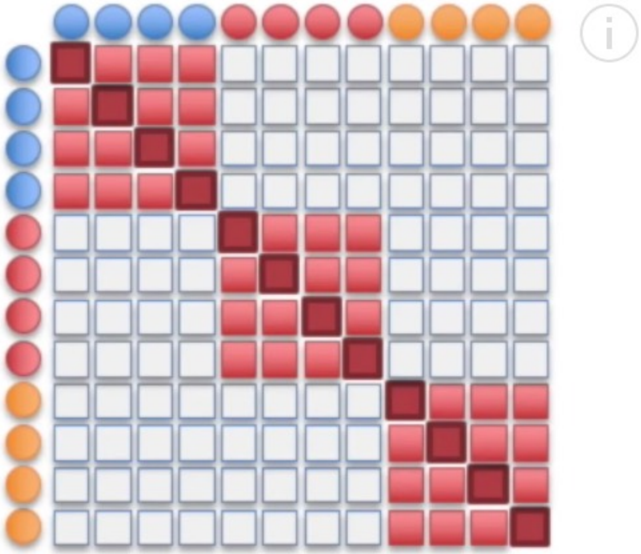
...without it the clusters would all clump up in the middle and be harder to see.

Image source: YouTube: StatQuest with Josh Starmer: *t-SNE, clearly explained*

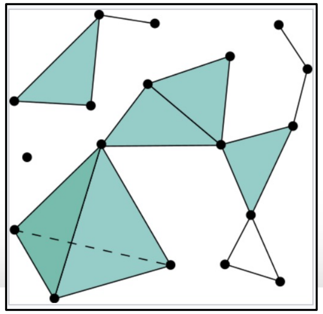
Make these similarity matrices as similar to each other as possible, and then you're done

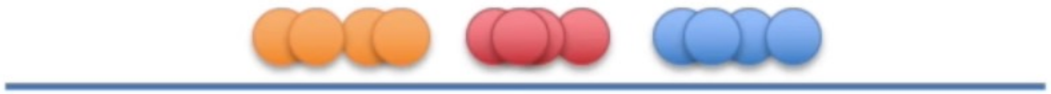






UMAP also makes a
2-D simplicial complex

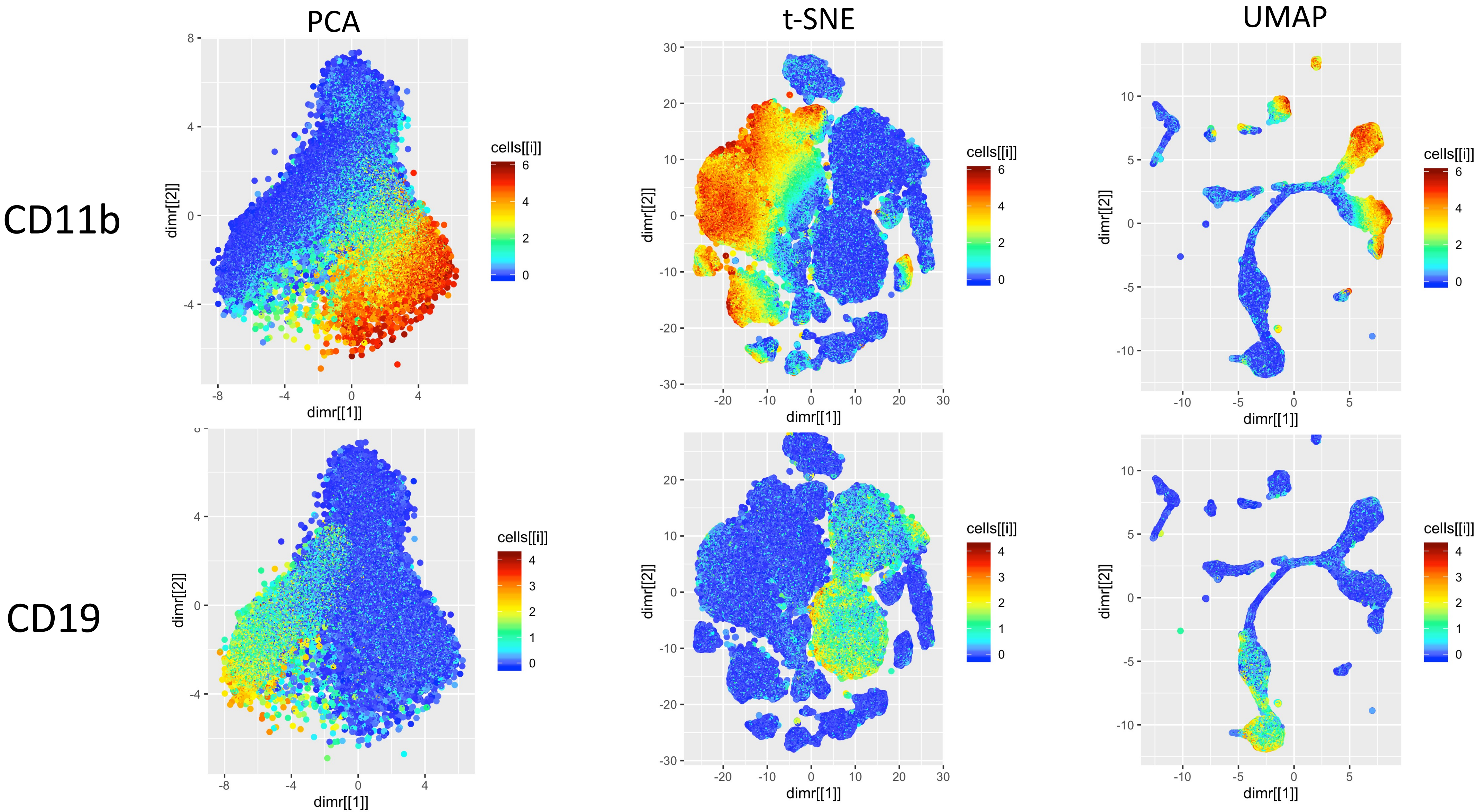




UMAP only moves one or a few cells at a time, rather than all of them as t-SNE does.

A cell will move toward one cell and away from another

Dimension reduction maps group similar cells near each other

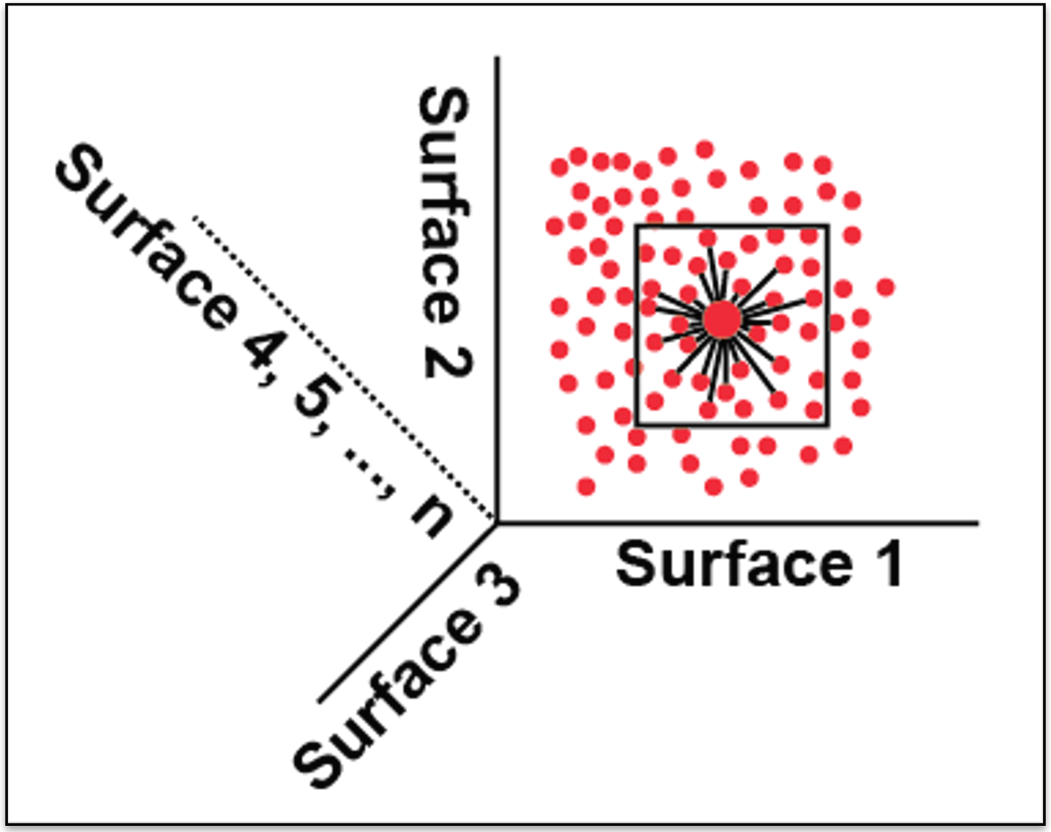


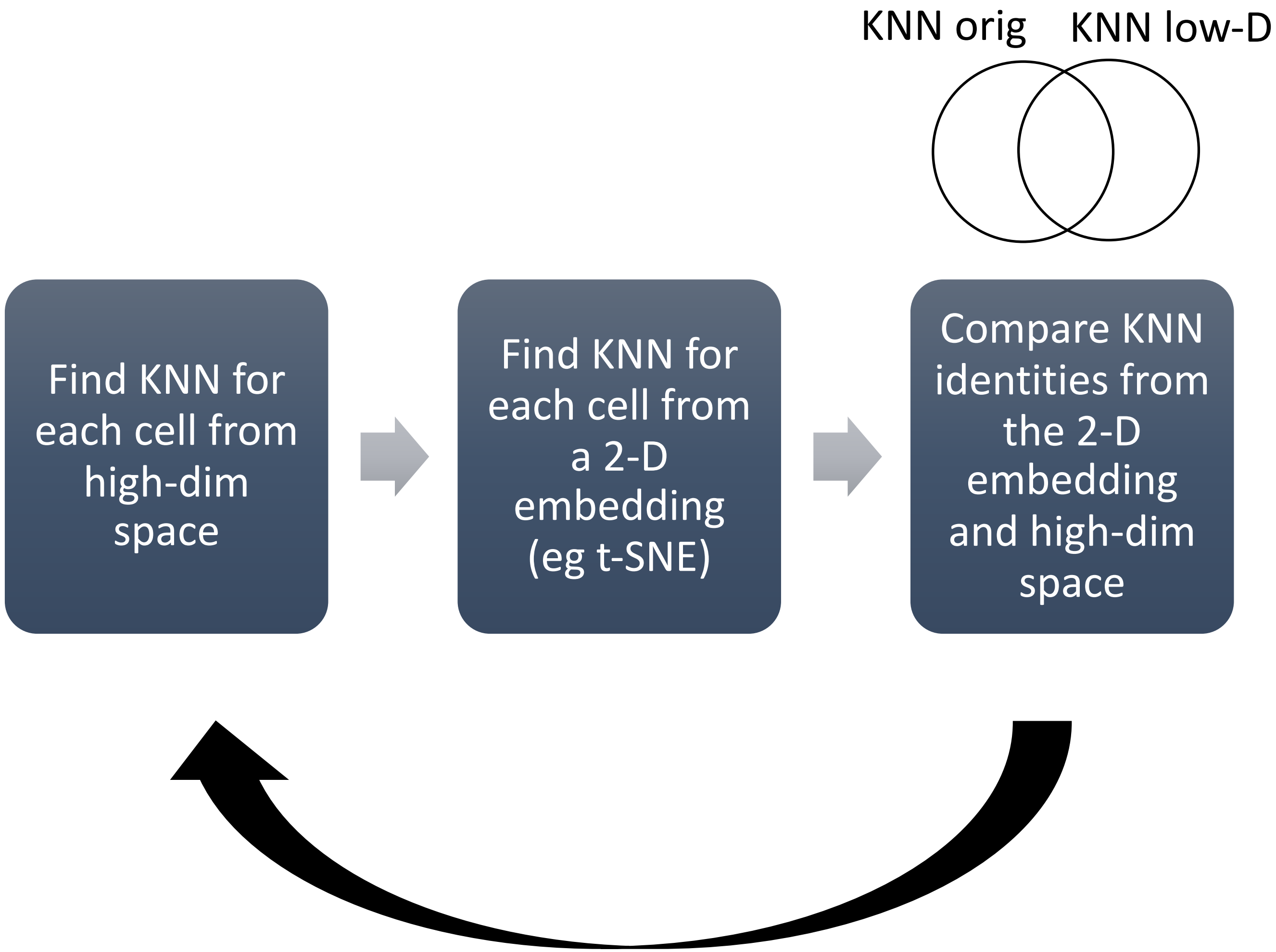
Samusik
Bone
marrow

Outline

- Part 1: Introduction
- **Part 2: Preservation of local structure**
- Part 3: Preservation of global structure

KNN to determine preservation of lower dimensional embeddings




KNN orig

KNN low-D

Find KNN for each cell from high-dim space

Find KNN for each cell from a 2-D embedding (eg t-SNE)

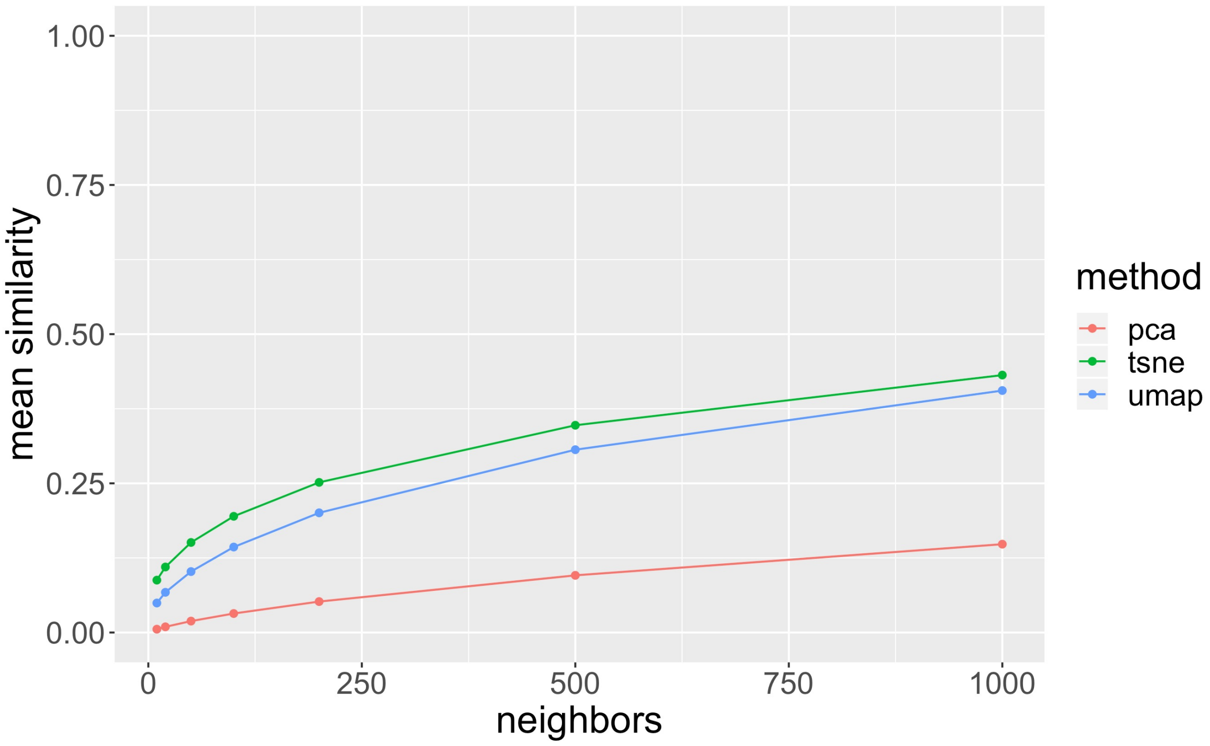
Compare KNN identities from the 2-D embedding and high-dim space

Repeat across a wide range of values for K

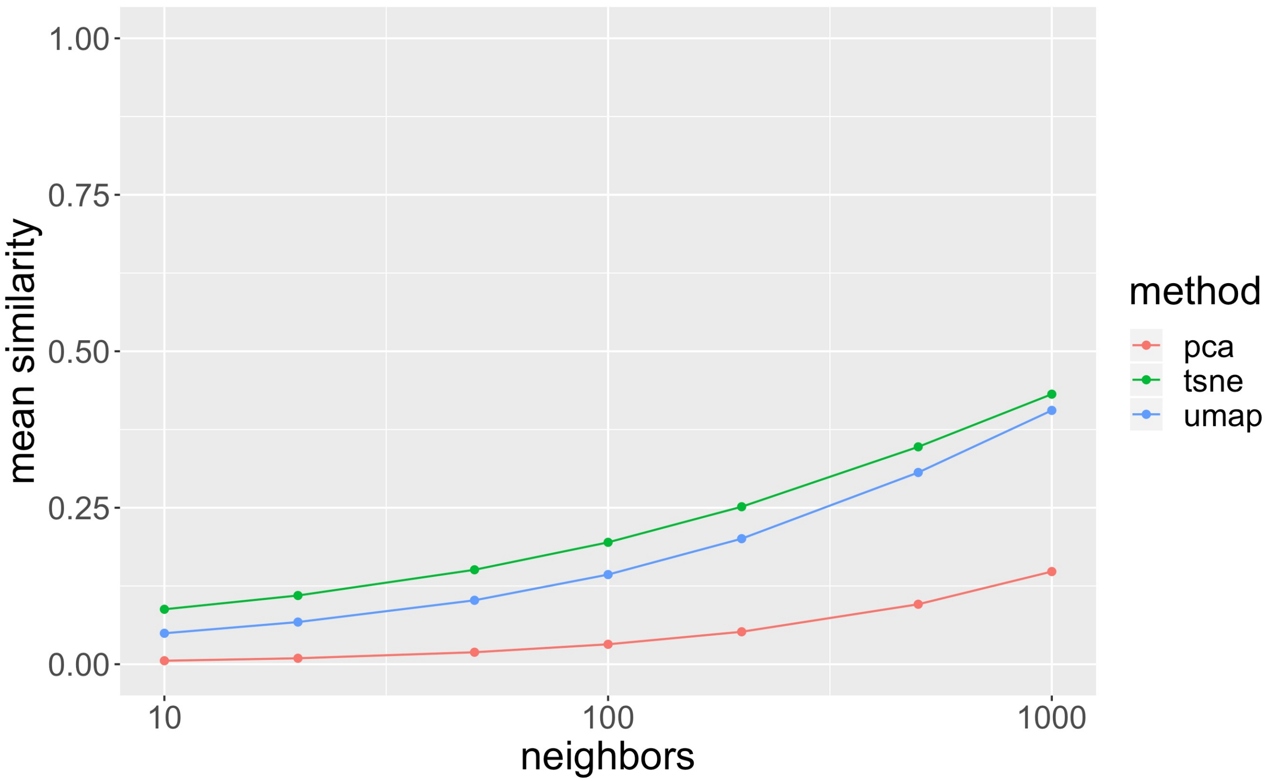
Bioconductor package: Sconify

Global KNN comparison between t-SNE, UMAP, and PCA

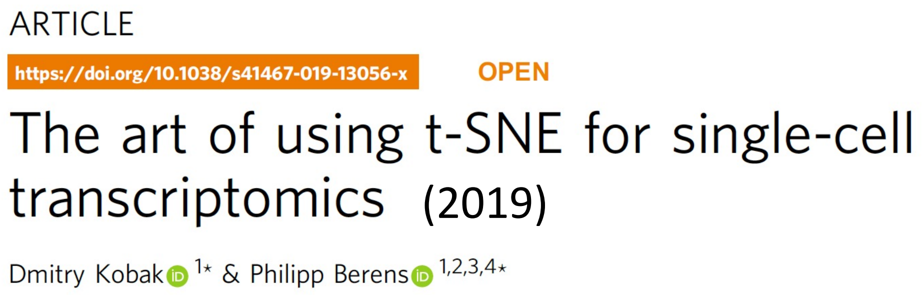
Dataset: Samusik Bone marrow (public)
Num. cells: 100k



X axis is on a log scale



t-SNE outperforms UMAP in KNN preservation, has been observed in scRNA seq data



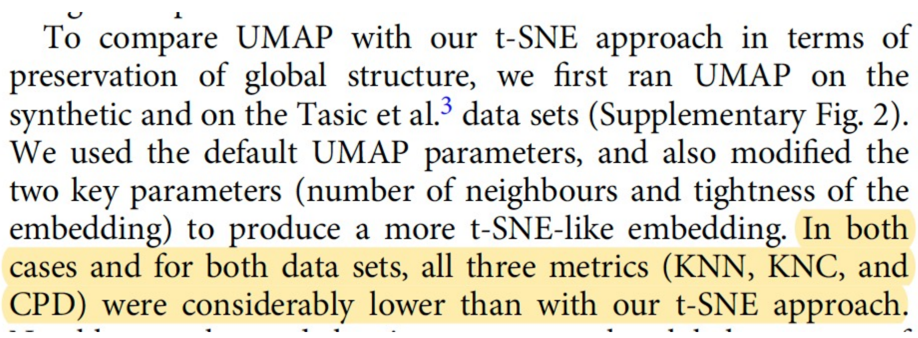
ARTICLE

https://doi.org/10.1038/s41467-019-13056-x OPEN

The art of using t-SNE for single-cell transcriptomics (2019)

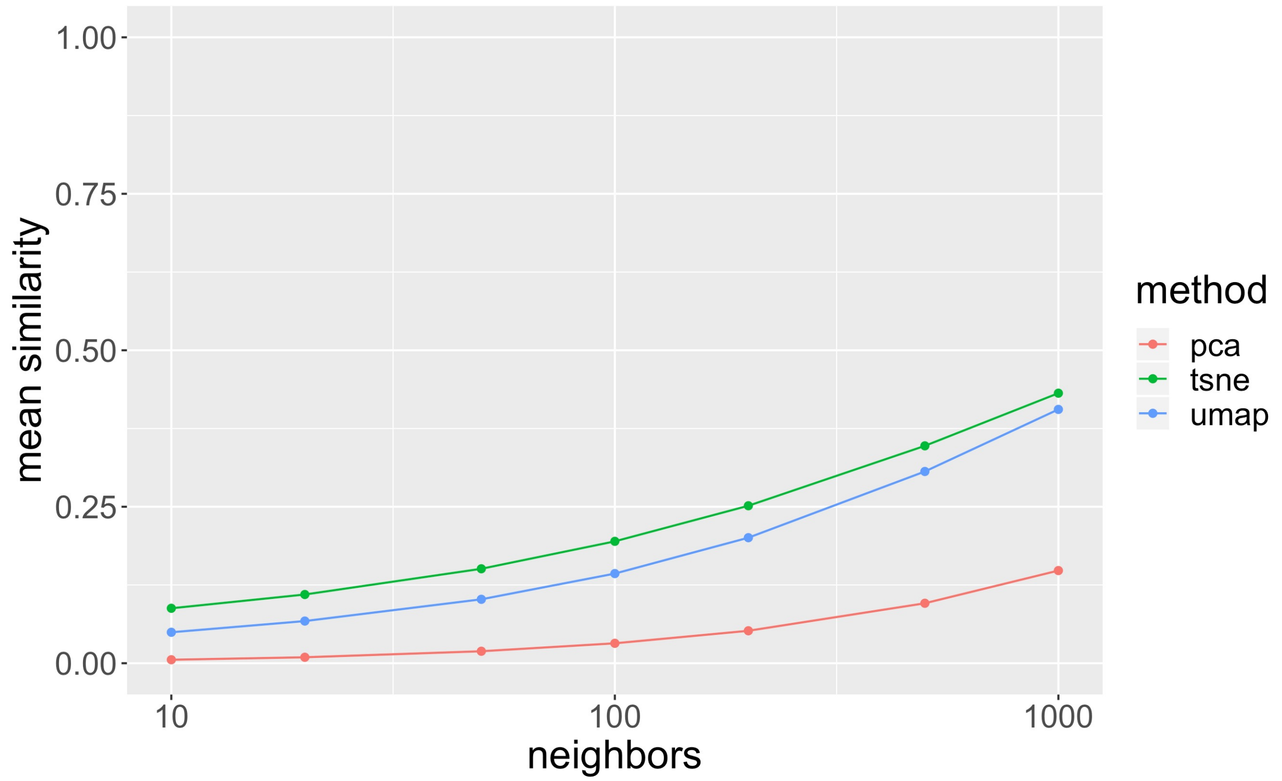
Dmitry Kobak[1*] & Philipp Berens[1,2,3,4*]

(Also did KNN preservation, K = 10 only)

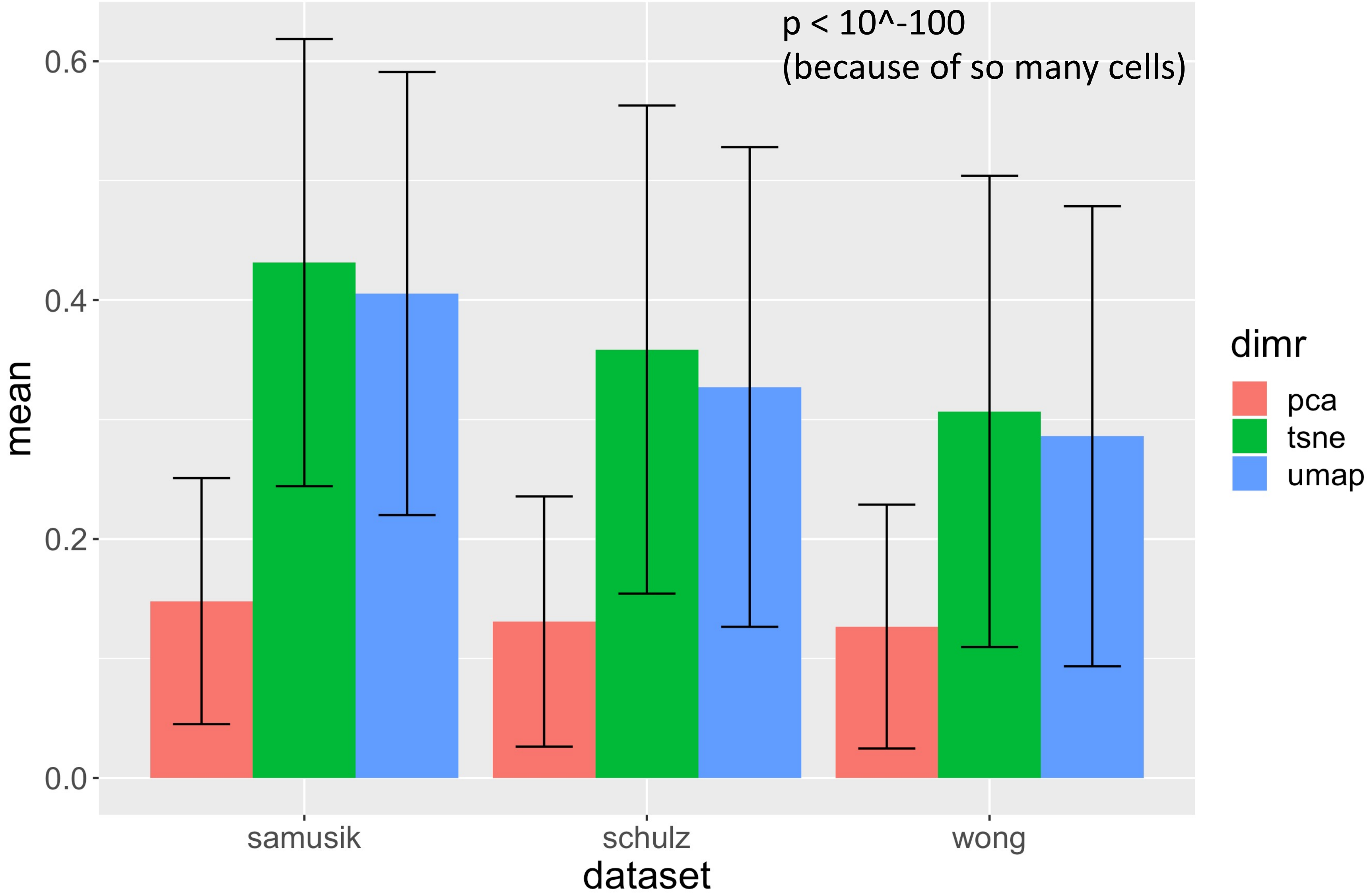


To compare UMAP with our t-SNE approach in terms of preservation of global structure, we first ran UMAP on the synthetic and on the Tasic et al.[3] data sets (Supplementary Fig. 2). We used the default UMAP parameters, and also modified the two key parameters (number of neighbours and tightness of the embedding) to produce a more t-SNE-like embedding. In both cases and for both data sets, all three metrics (KNN, KNC, and CPD) were considerably lower than with our t-SNE approach.

X axis is on a log scale



Results confirmed across 3 datasets, but with very large standard deviation

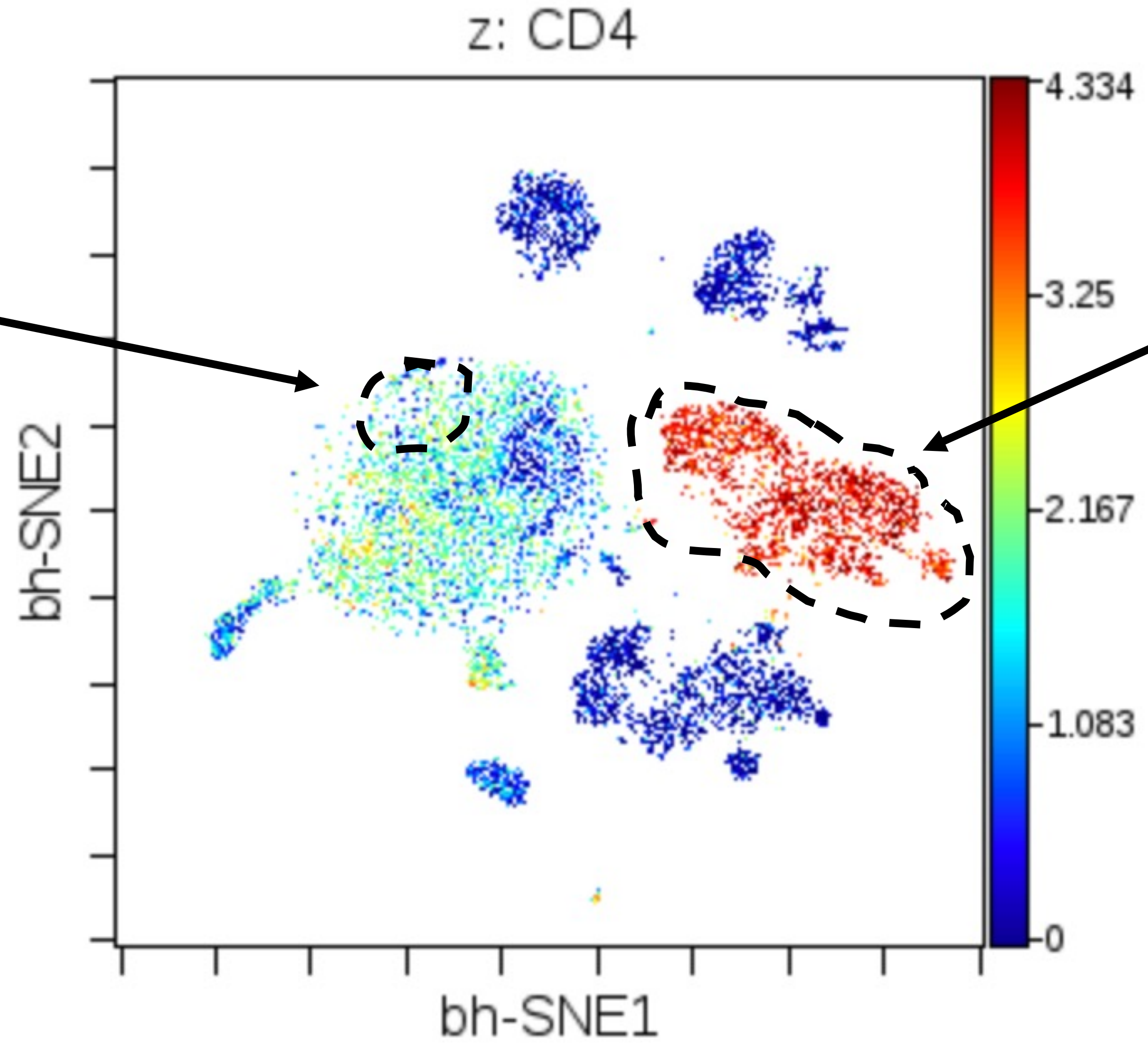


p < 10^-100
(because of so many cells)

Cells: 100k
k = 1000

Does dimension reduction maps preserve some regions better than others (should and/or how should we gate the map?)

Is this a valid gate?

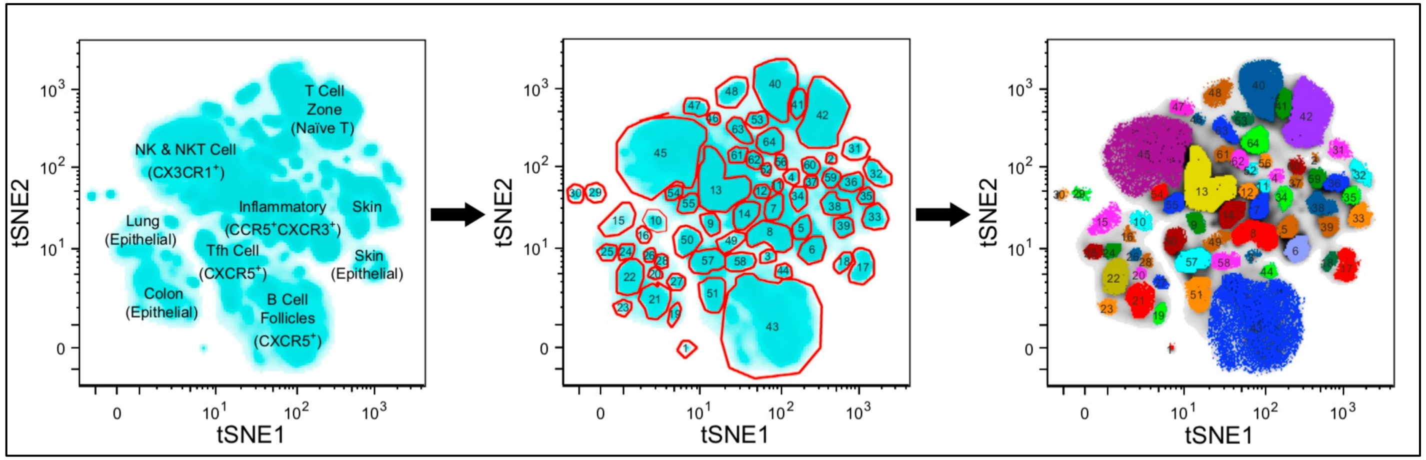


Is this a valid gate?

Data: Axel Schulz, PhD

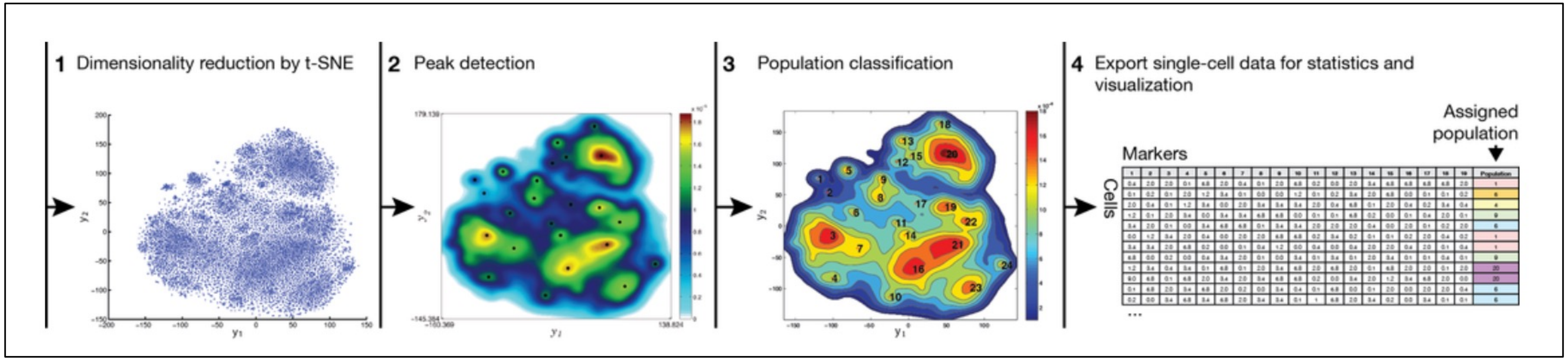
People are already gating and clustering dimension reduction maps. Guidelines are needed!

Michael Wong and Evan Newell: Manually gating a t-SNE map



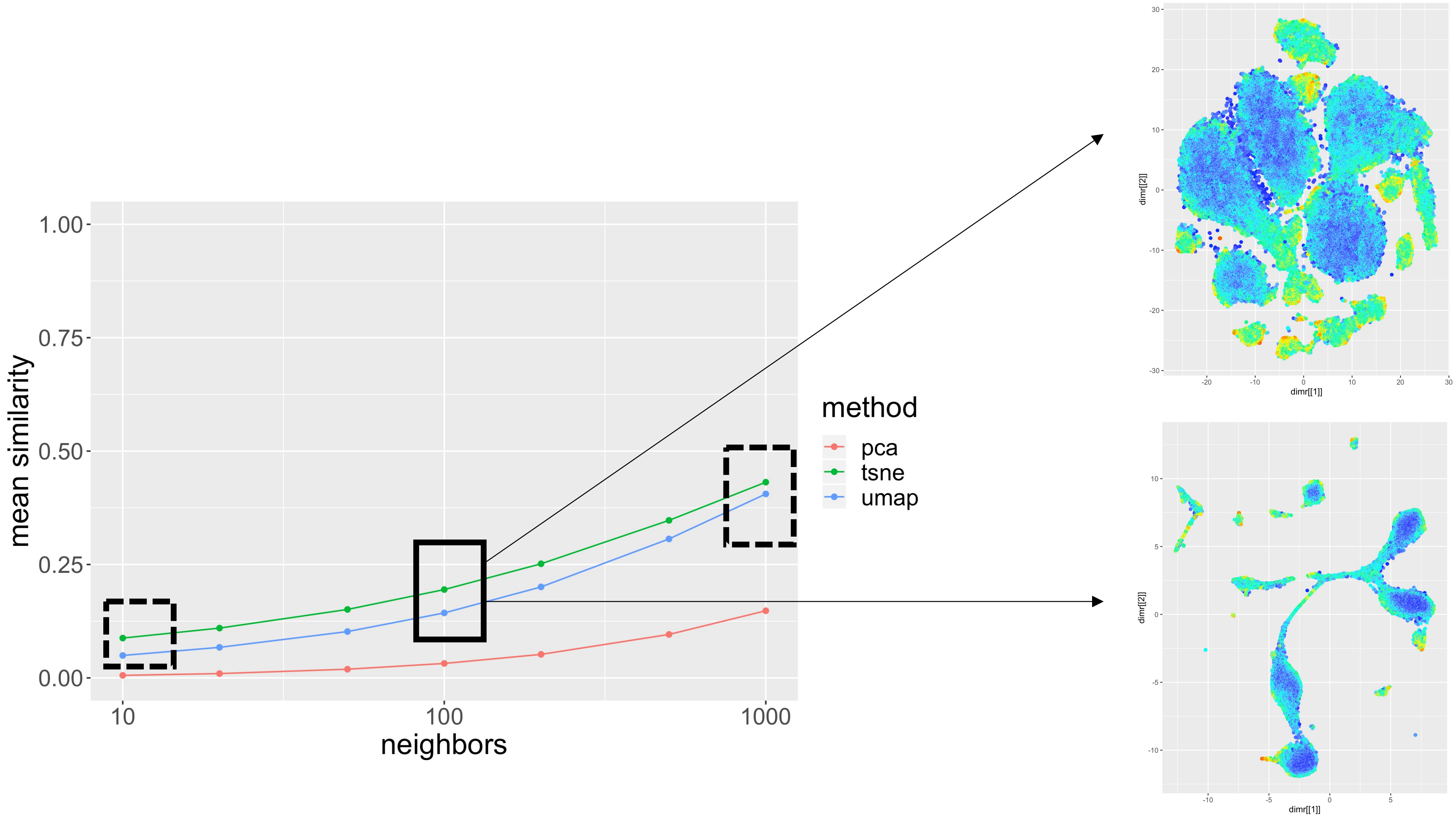
Wong *et al*, *Cell* 2016

Accense (Petter Brodin): Clustering a t-SNE map

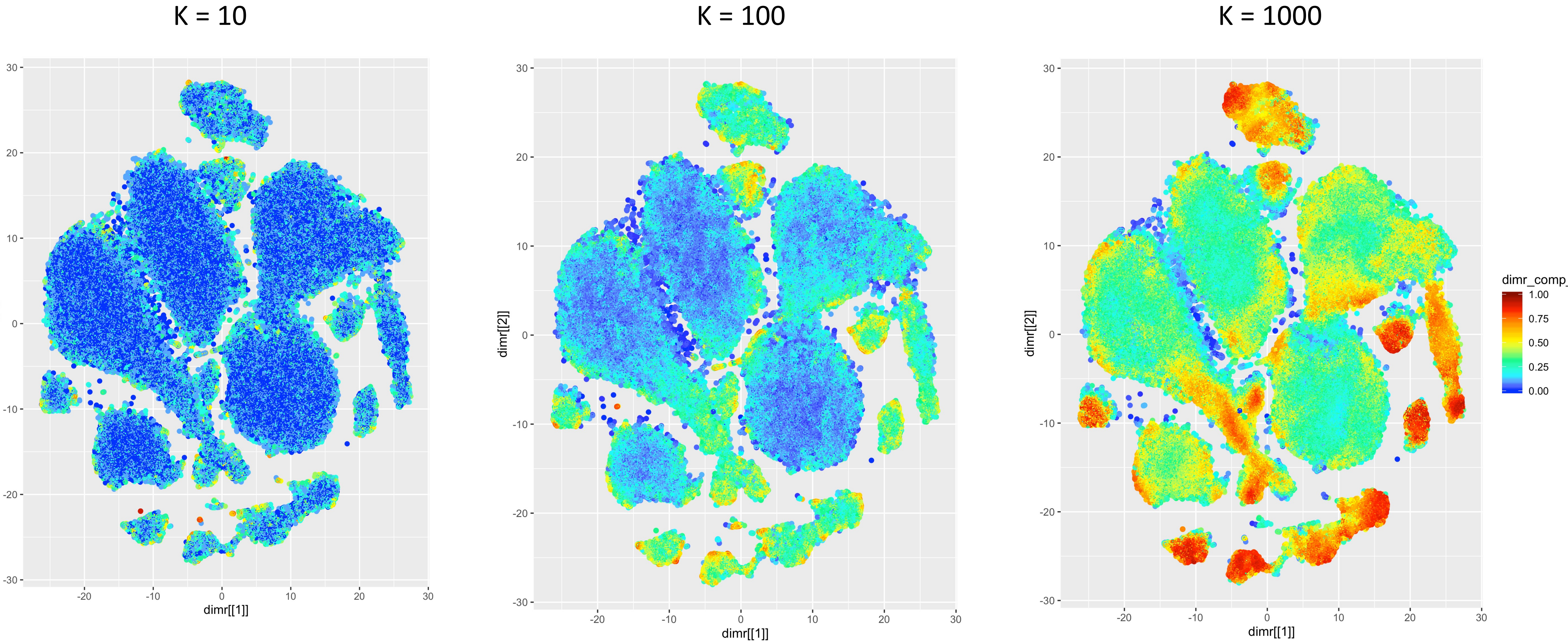

1 Dimensionality reduction by t-SNE
2 Peak detection
3 Population classification
4 Export single-cell data for statistics and visualization
Assigned population
Markers
Cells
y₁
y₂

Shekar *et al*, *PNAS* 2014

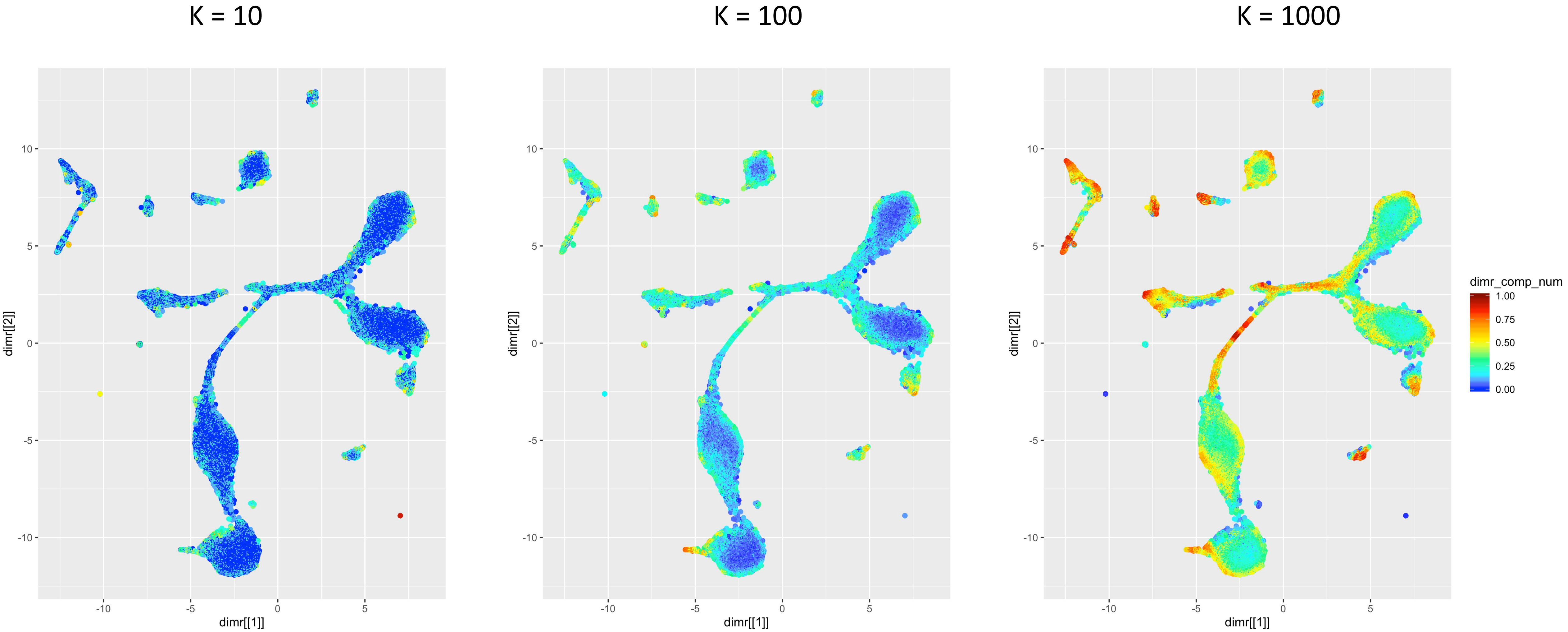
Color a dimension reduction map by it’s own neighborhood preservation, given k



Local comparison for t-SNE



Local comparison for UMAP



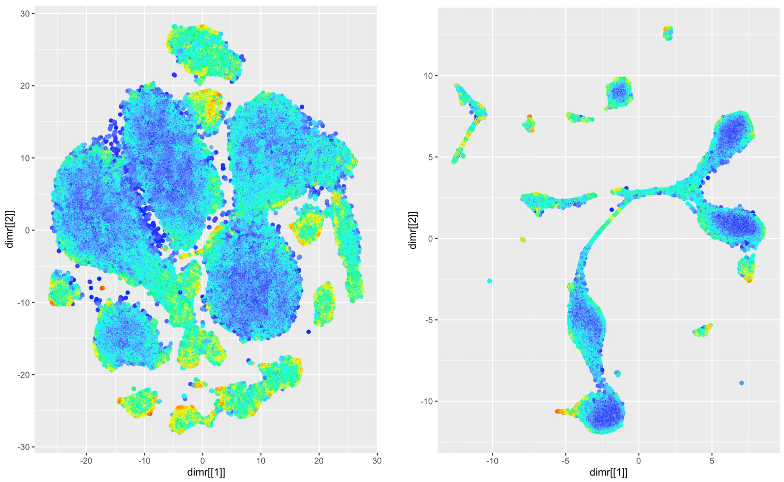
K = 10

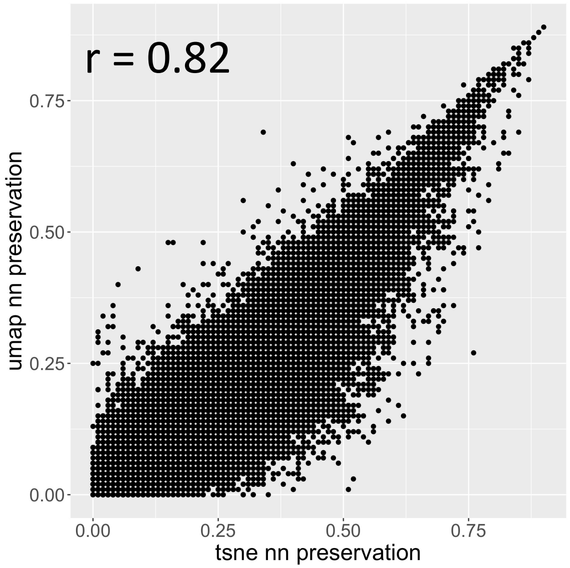
K = 100

K = 1000

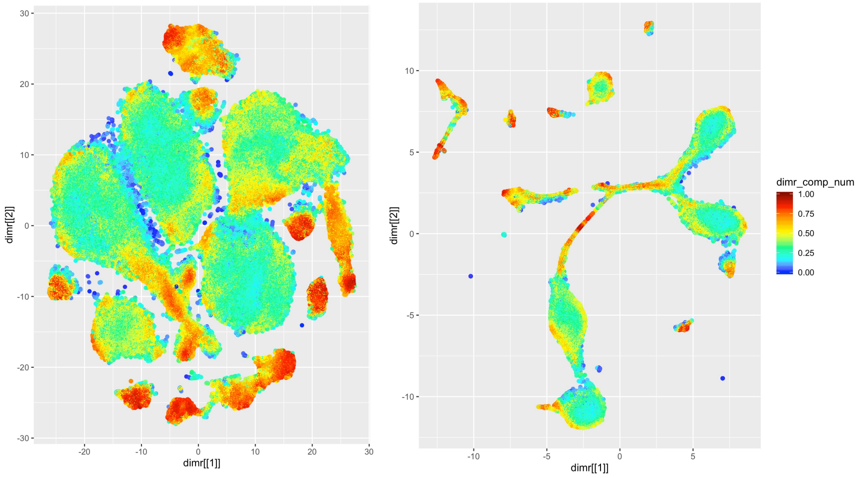
t-SNE and UMAP are preserving the data in a similar manner

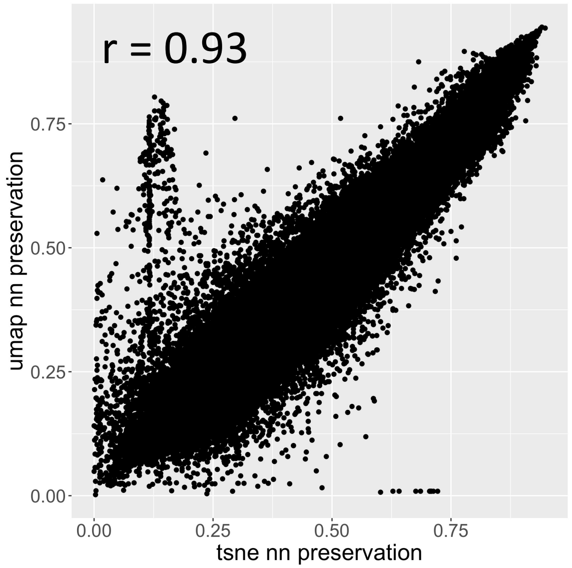
K = 100





K = 1000





Part 1 conclusions

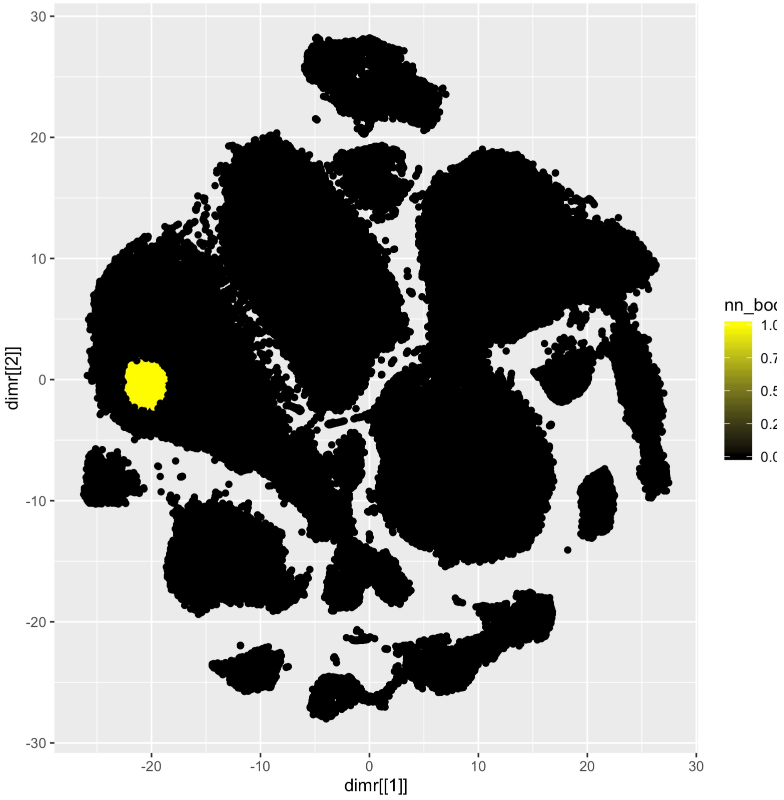
- t-SNE outperforms UMAP (though only slightly) in KNN preservation
- Both t-SNE and UMAP outperform PCA in KNN preservation
- KNN preservation performance varies in specific patterns across both t-SNE and UMAP
- t-SNE and UMAP have better KNN preservation in smaller islands/corridors in the data. Implications on how to gate the maps

Outline

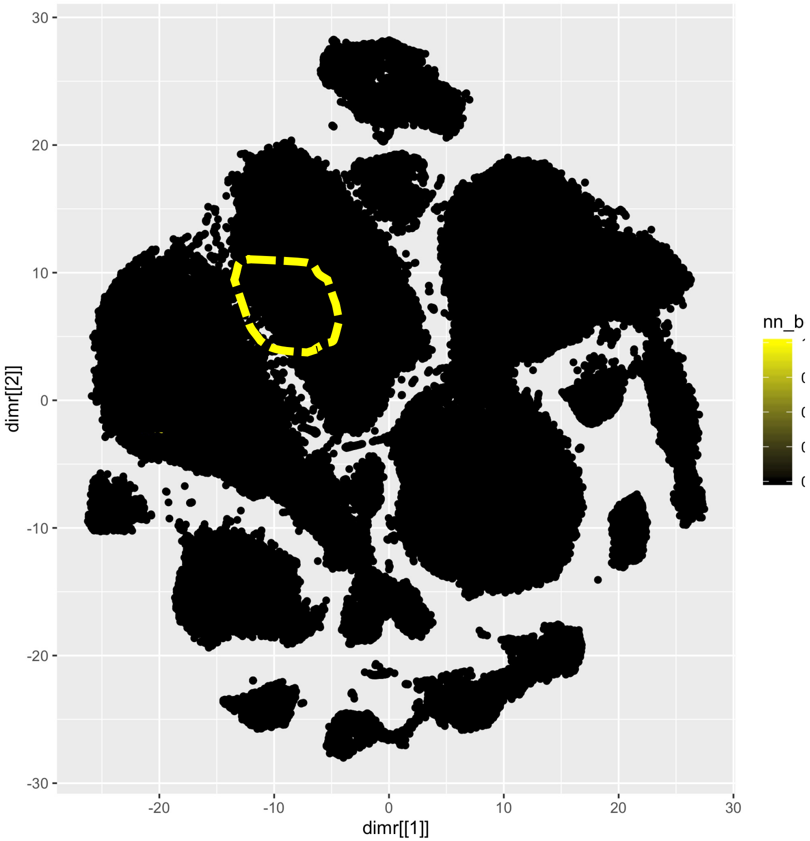
- Part 1: Introduction
- Part 2: Preservation of local structure
- **Part 3: Preservation of global structure**

If a “gate” on the map has 30% KNN preservation, where are the other cells?

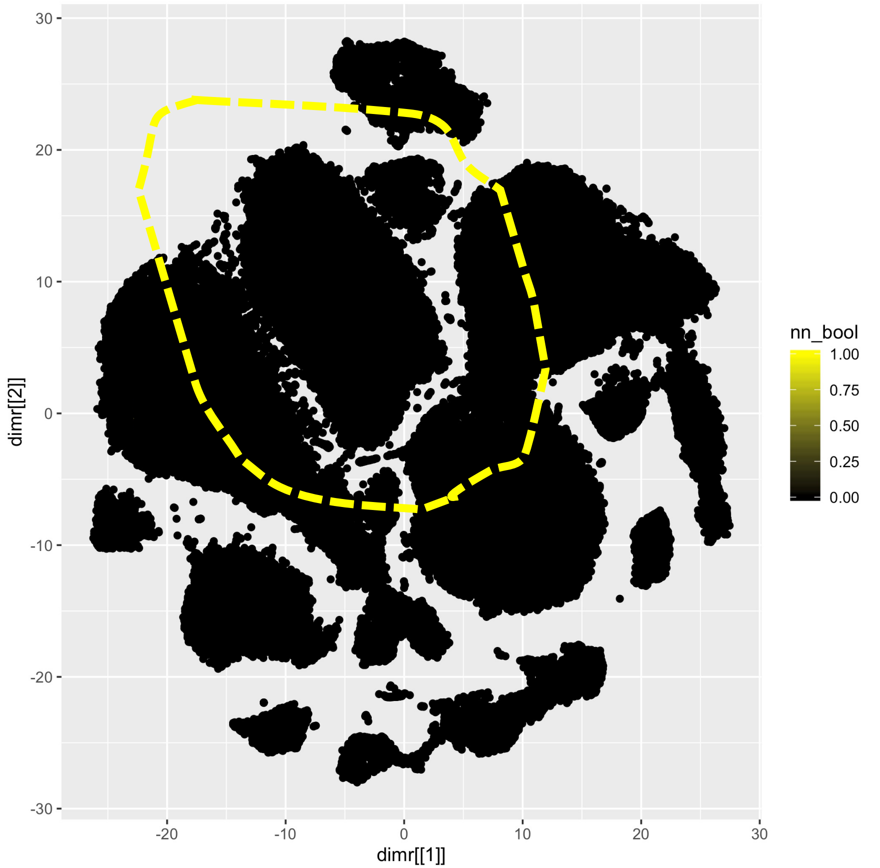
KNN ID on t-SNE map



KNN from hi-D, locations
Hypothesis 1

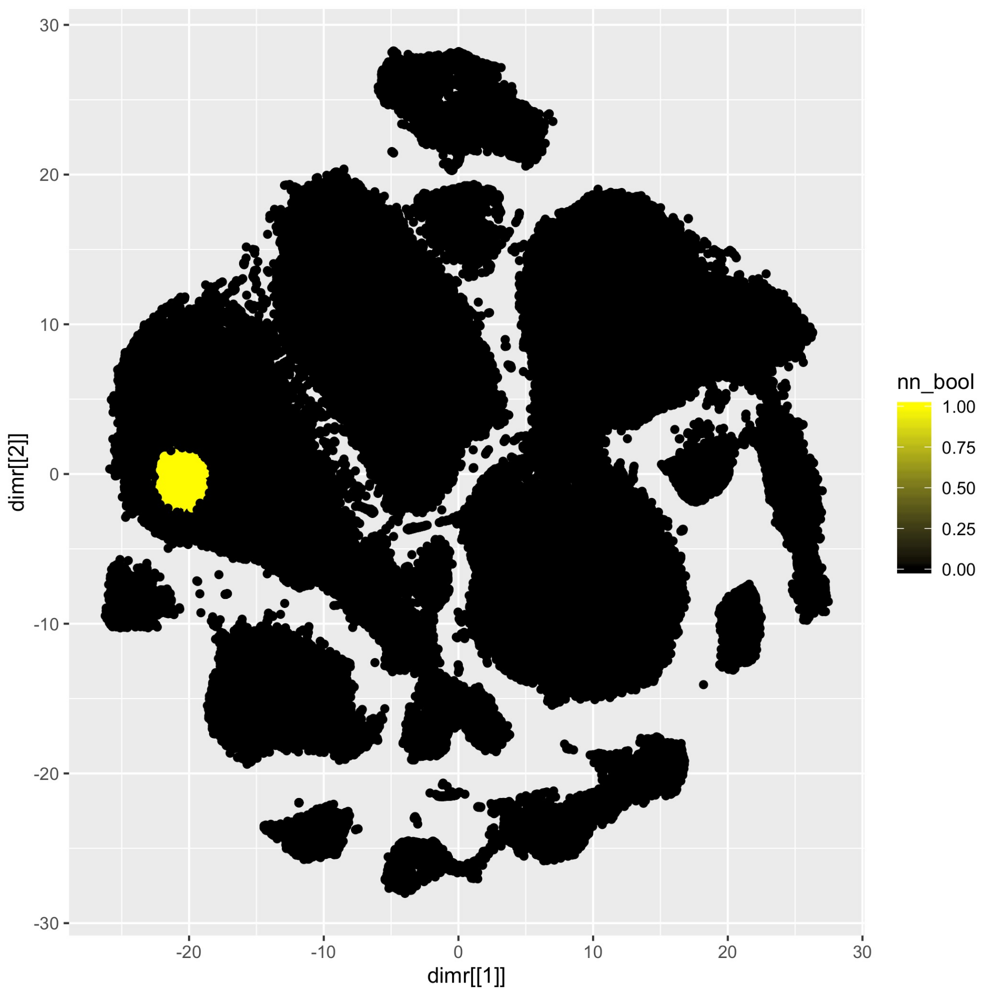


KNN from hi-D, locations
Hypothesis 2

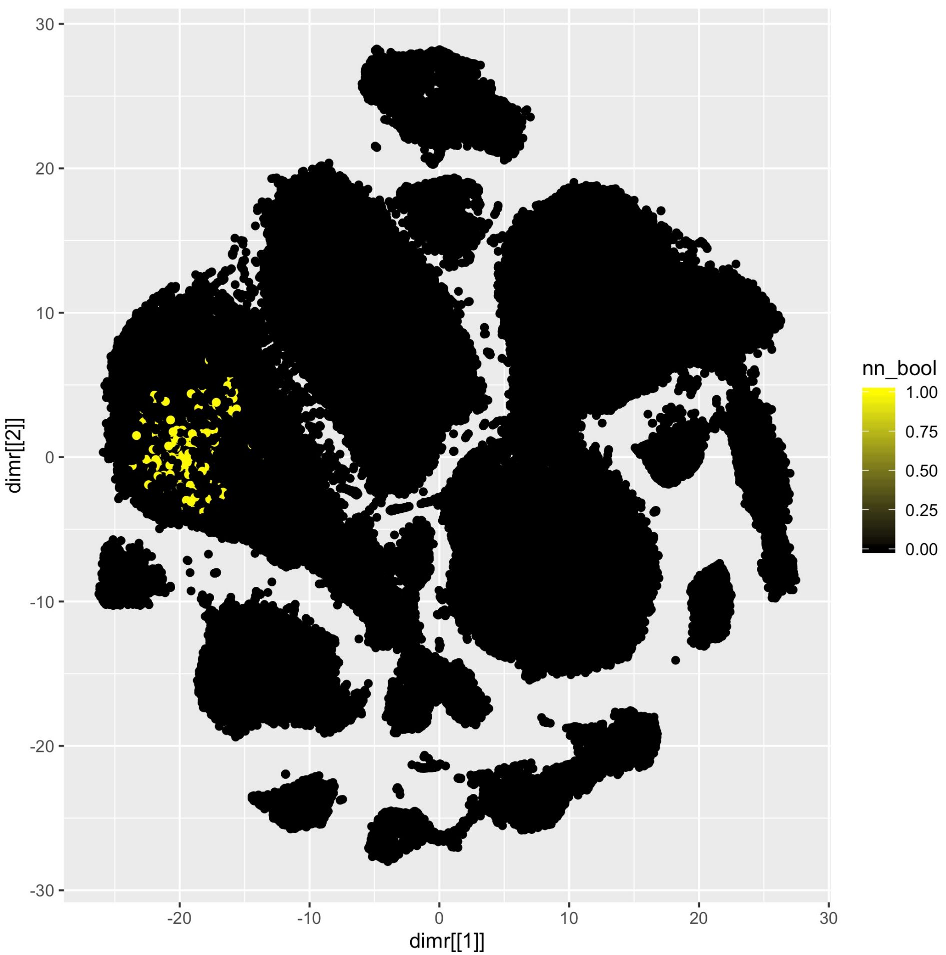


KNN identity for t-SNE, k = 1000, cell 1

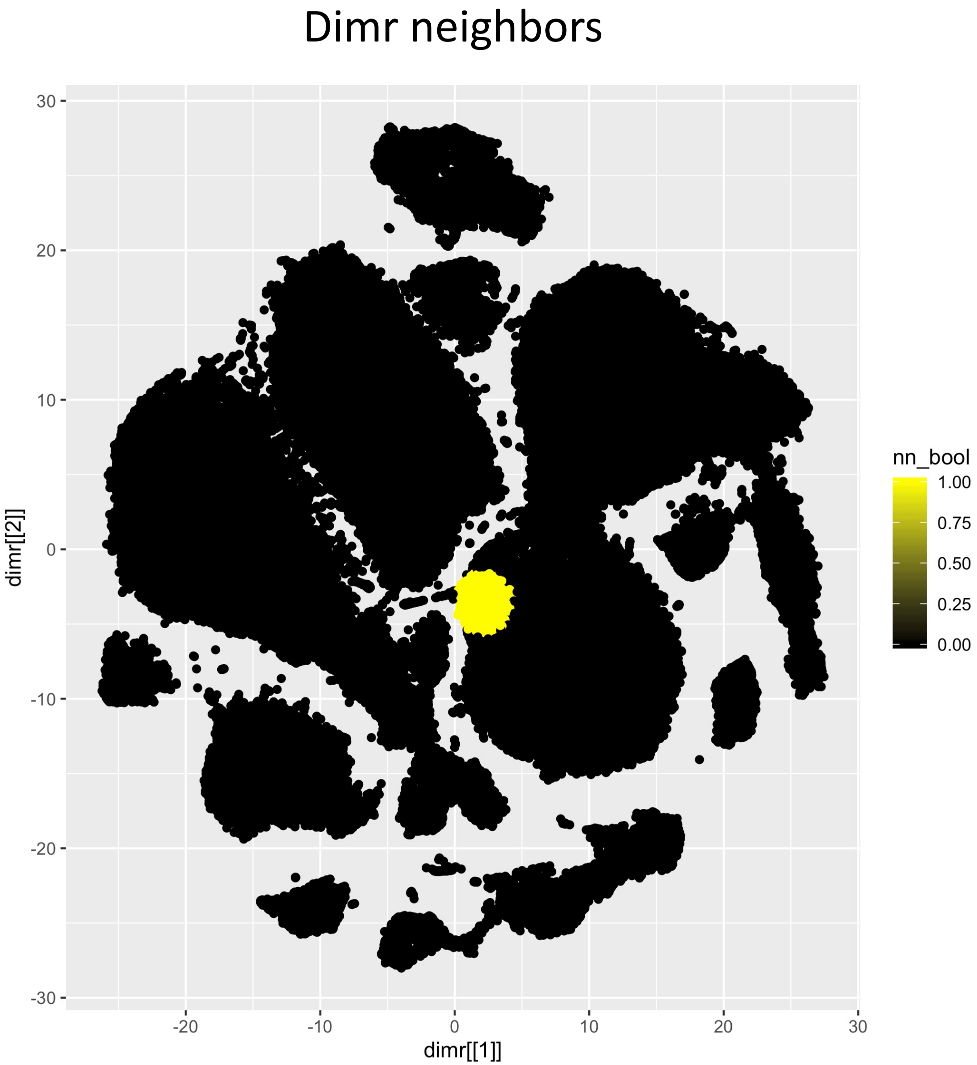
Dimr neighbors

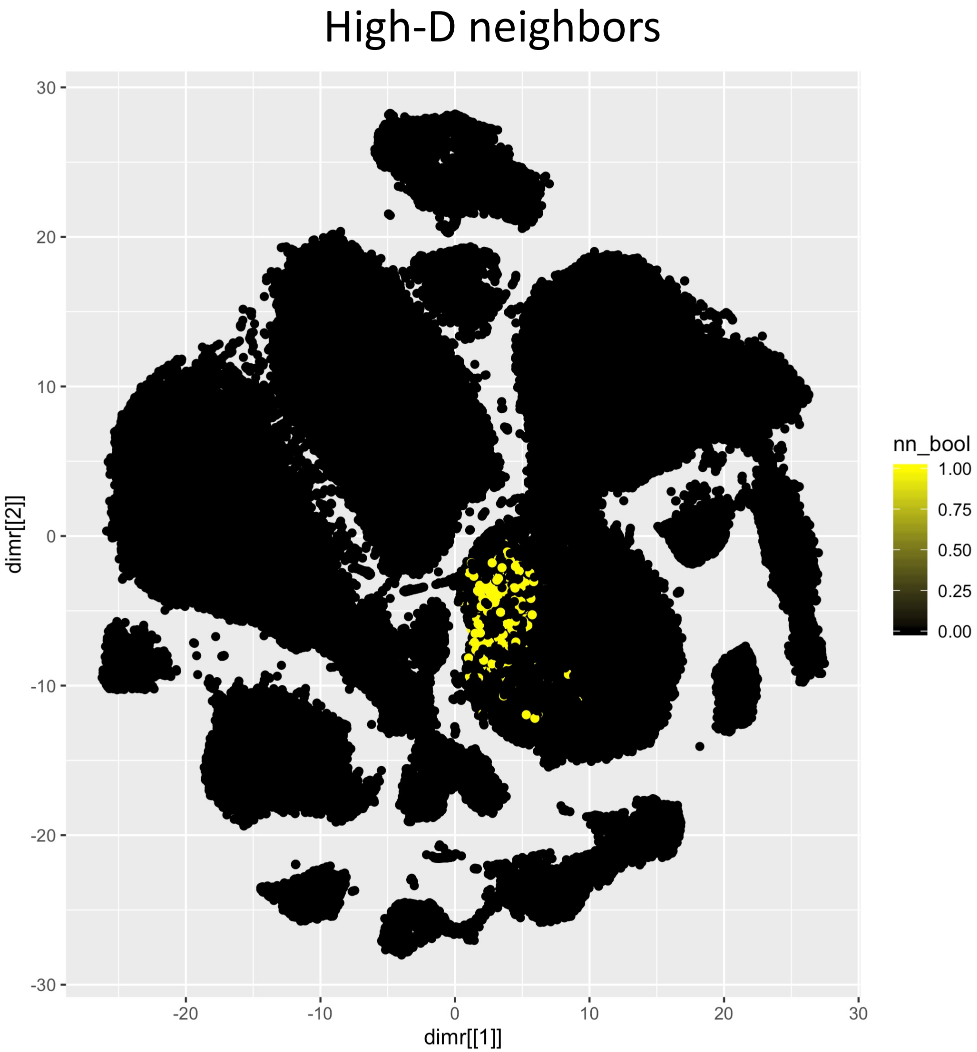


High-D neighbors



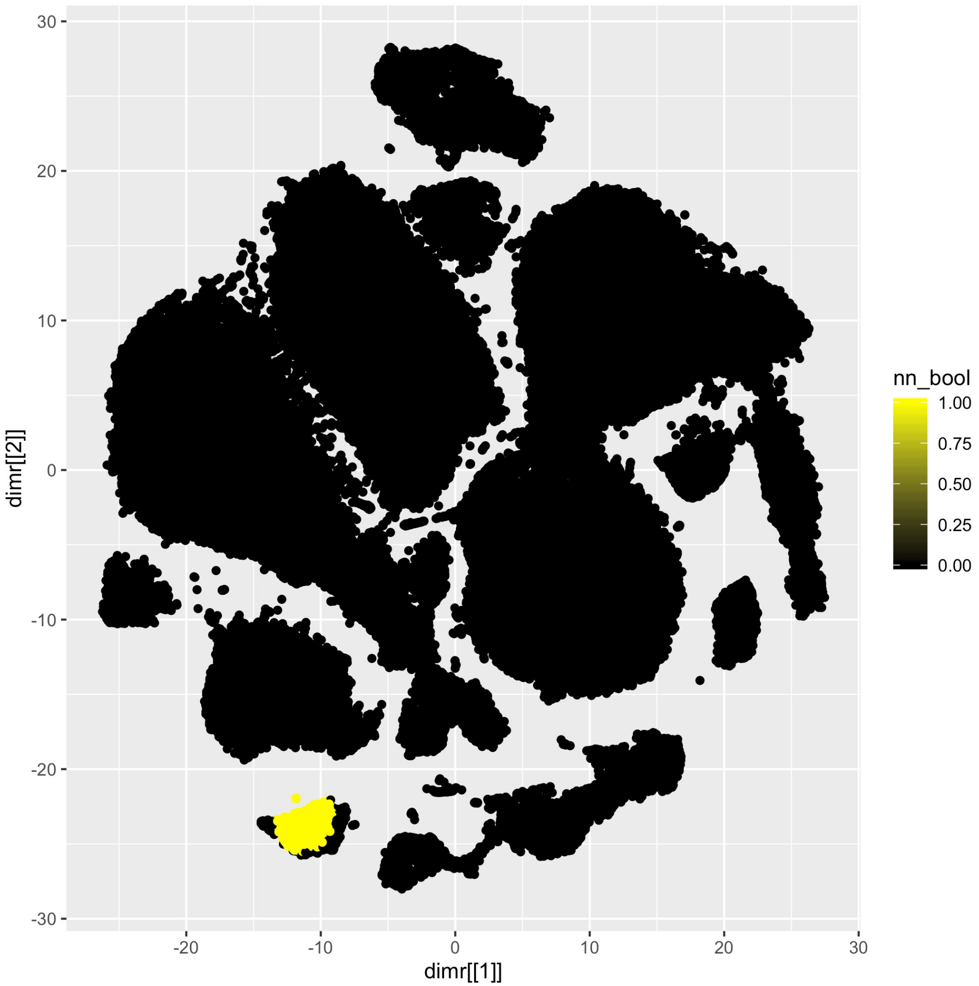
KNN identity for t-SNE, k = 1000, cell 4



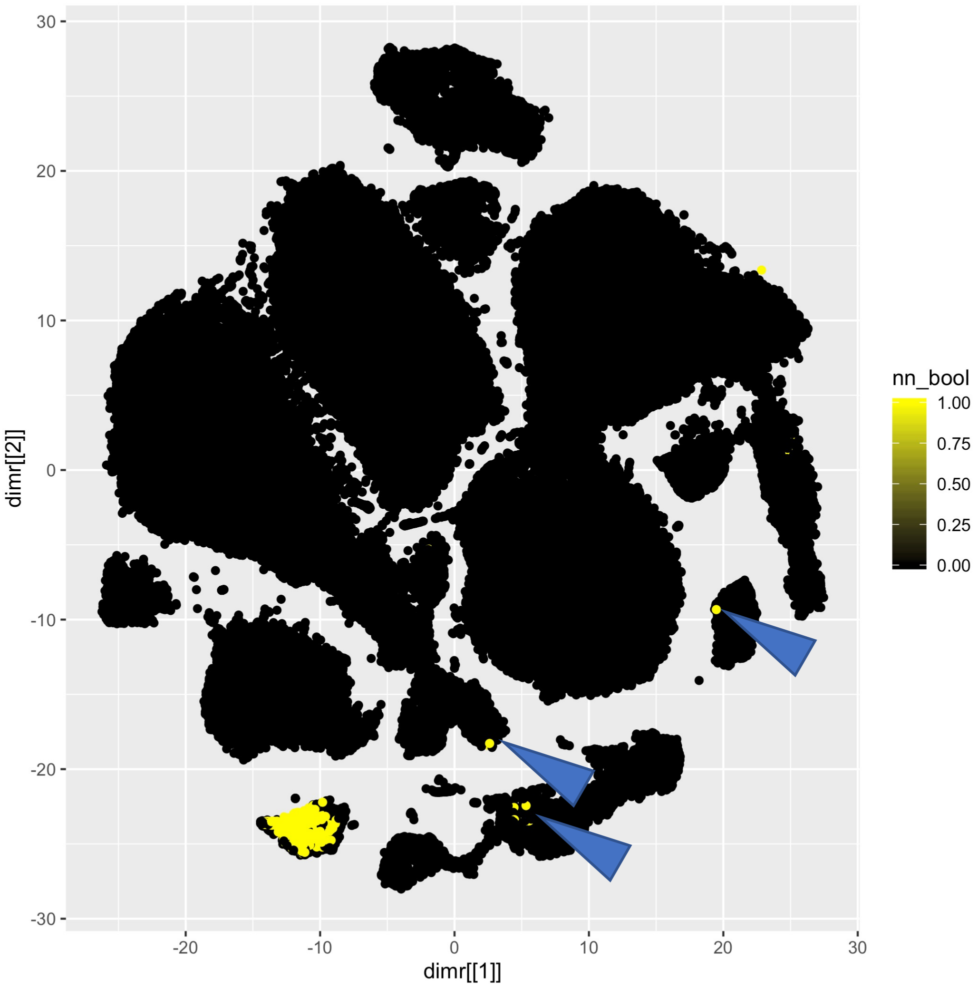


KNN identity for t-SNE, k = 1000, cell 6

Dimr neighbors

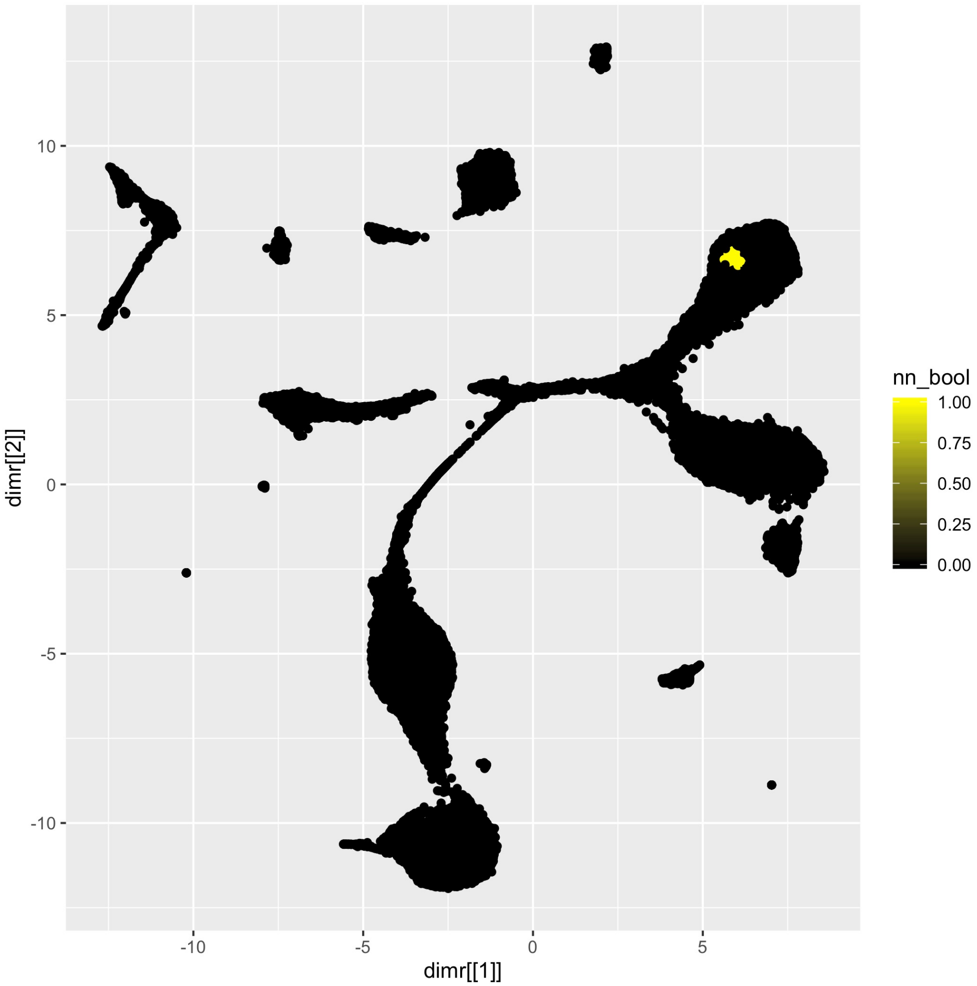


High-D neighbors

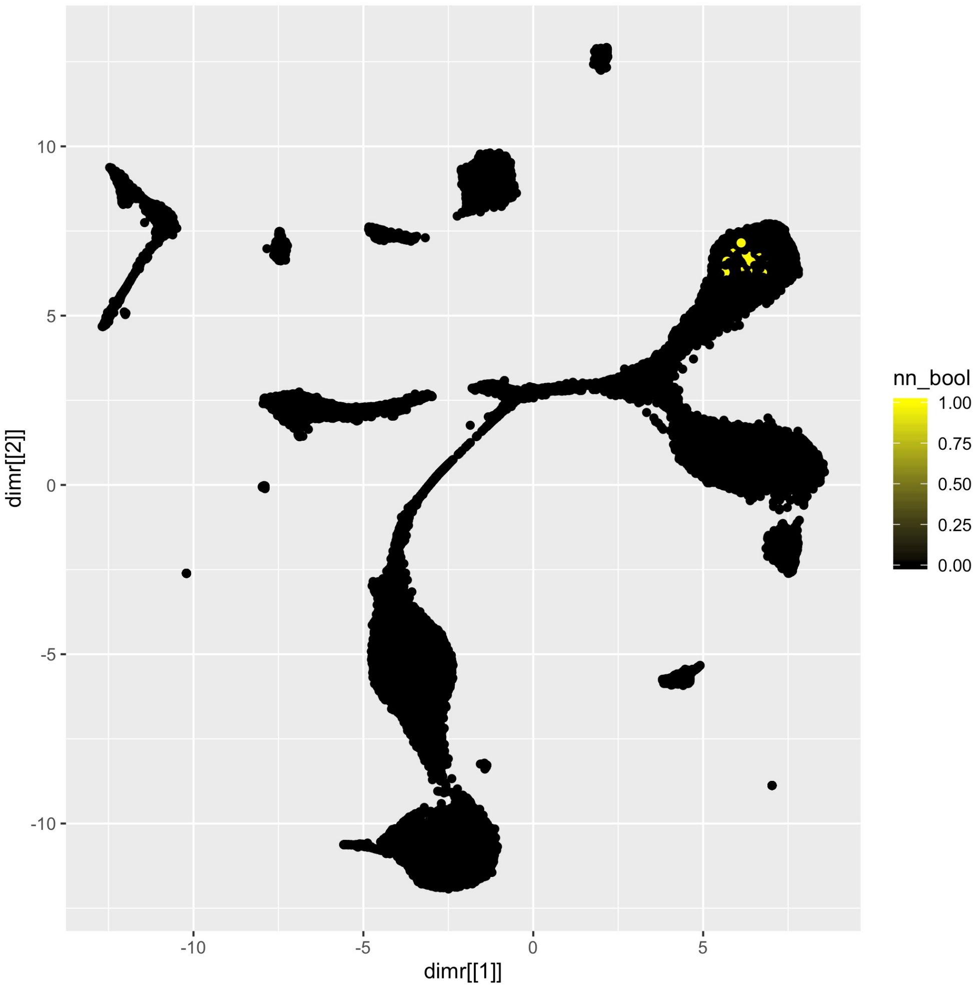


KNN identity for UMAP, k = 1000, cell 1

Dimr neighbors

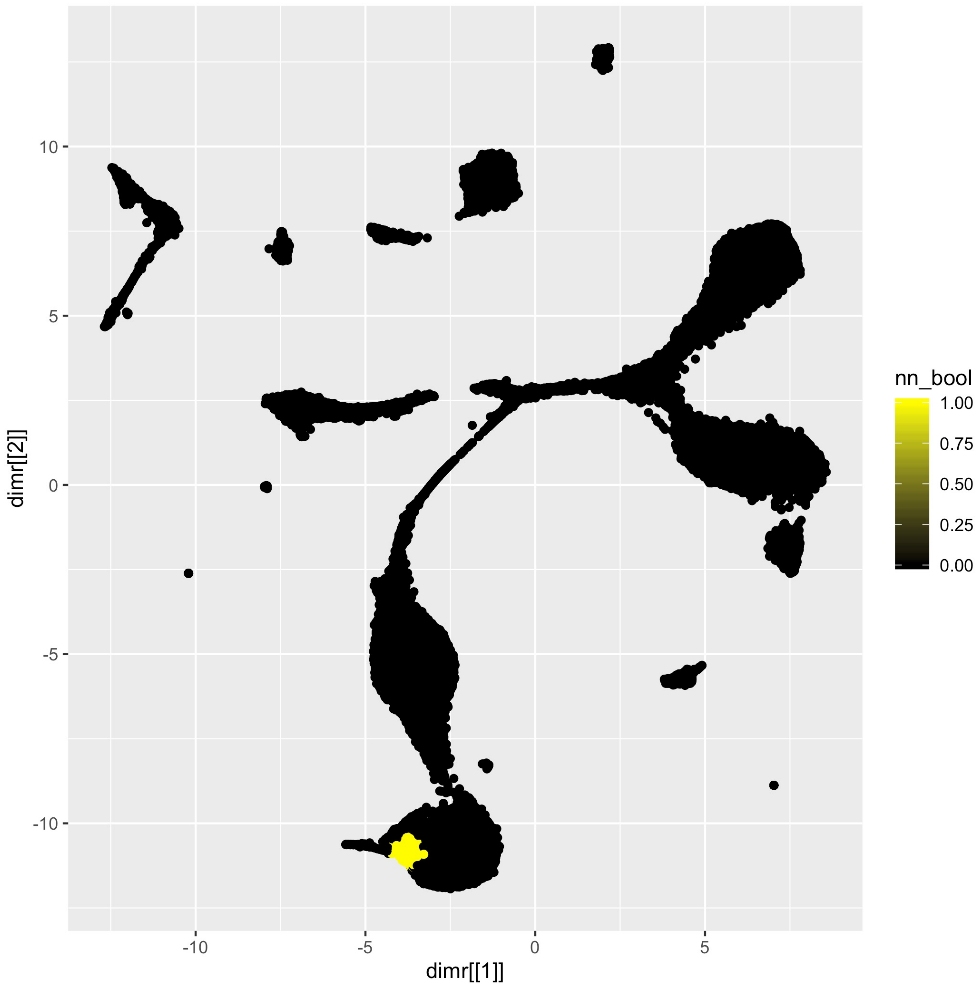


High-D neighbors

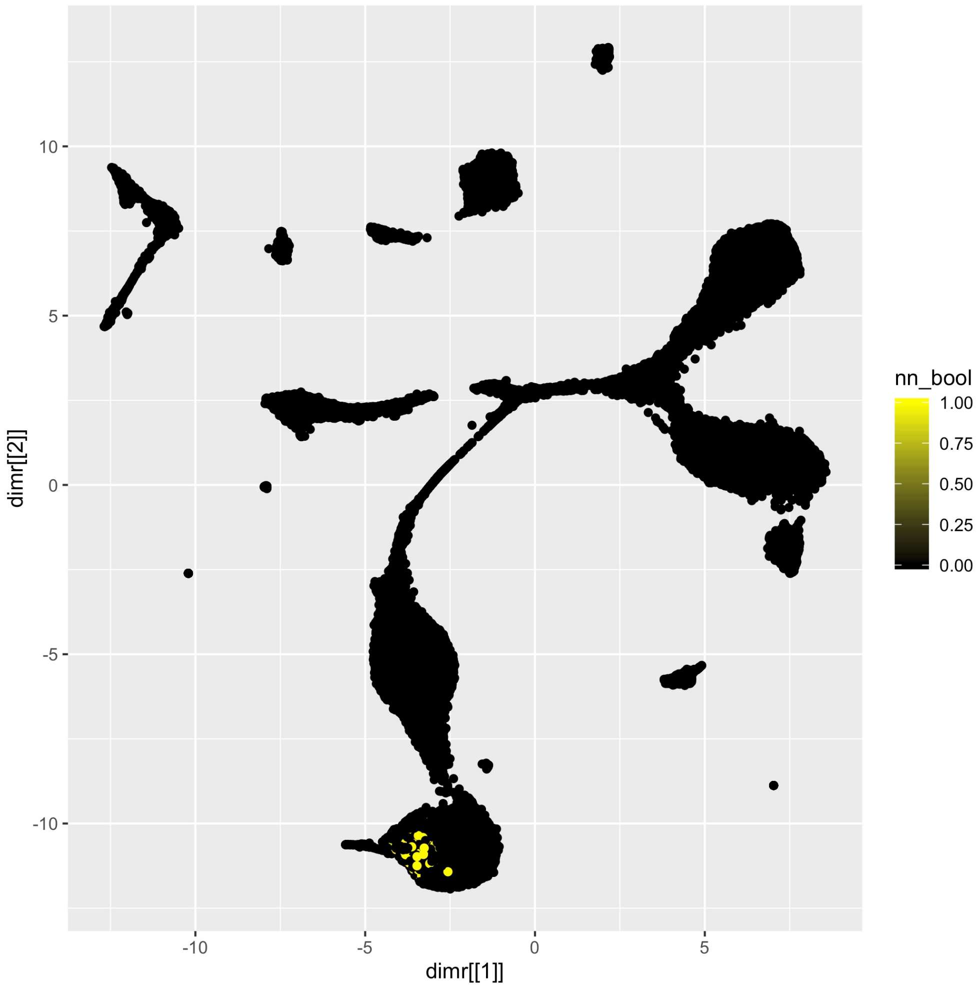


KNN identity for UMAP, k = 1000, cell 4

Dimr neighbors

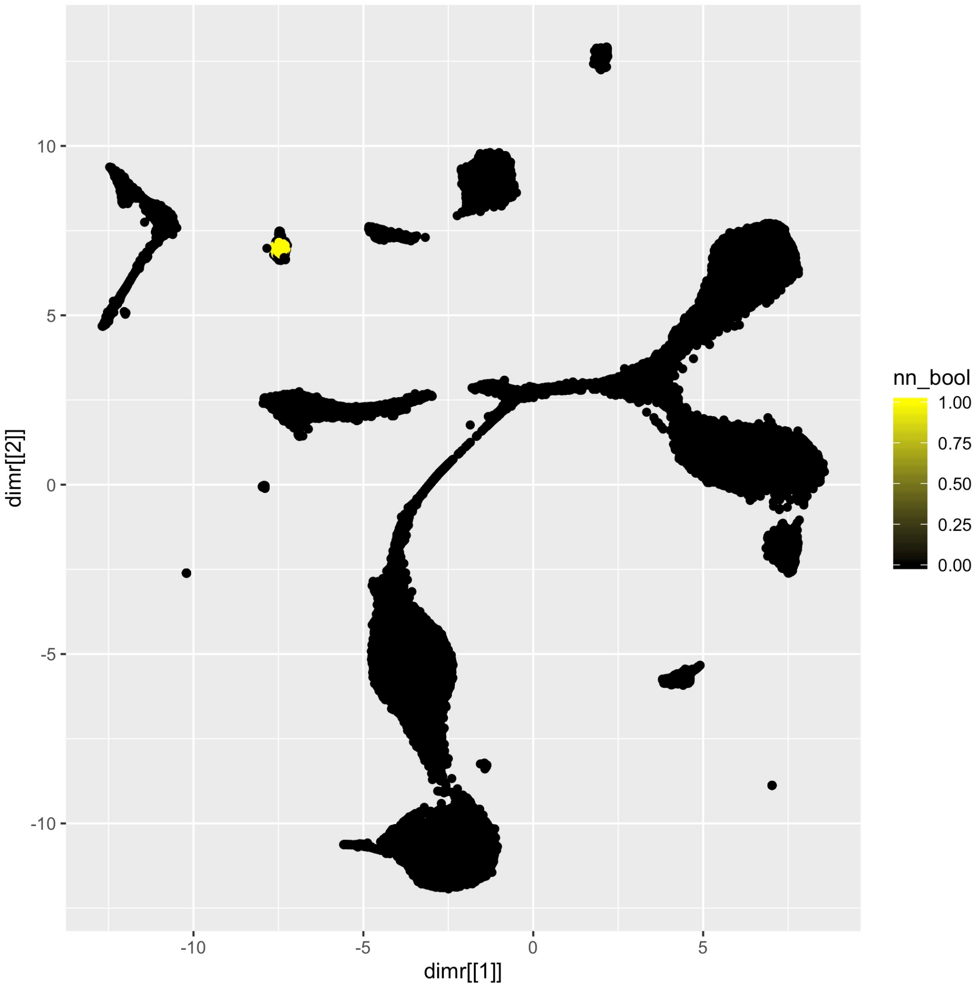


High-D neighbors

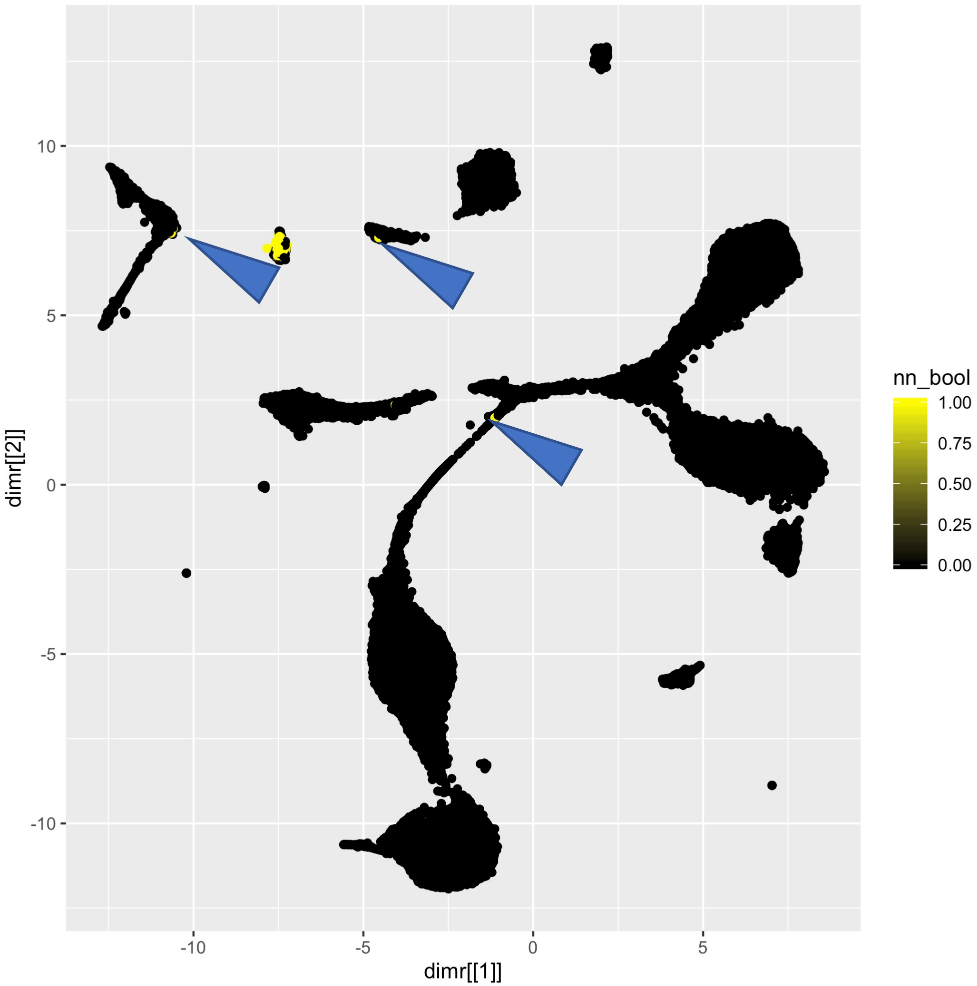


KNN identity for UMAP, k = 1000, cell 6

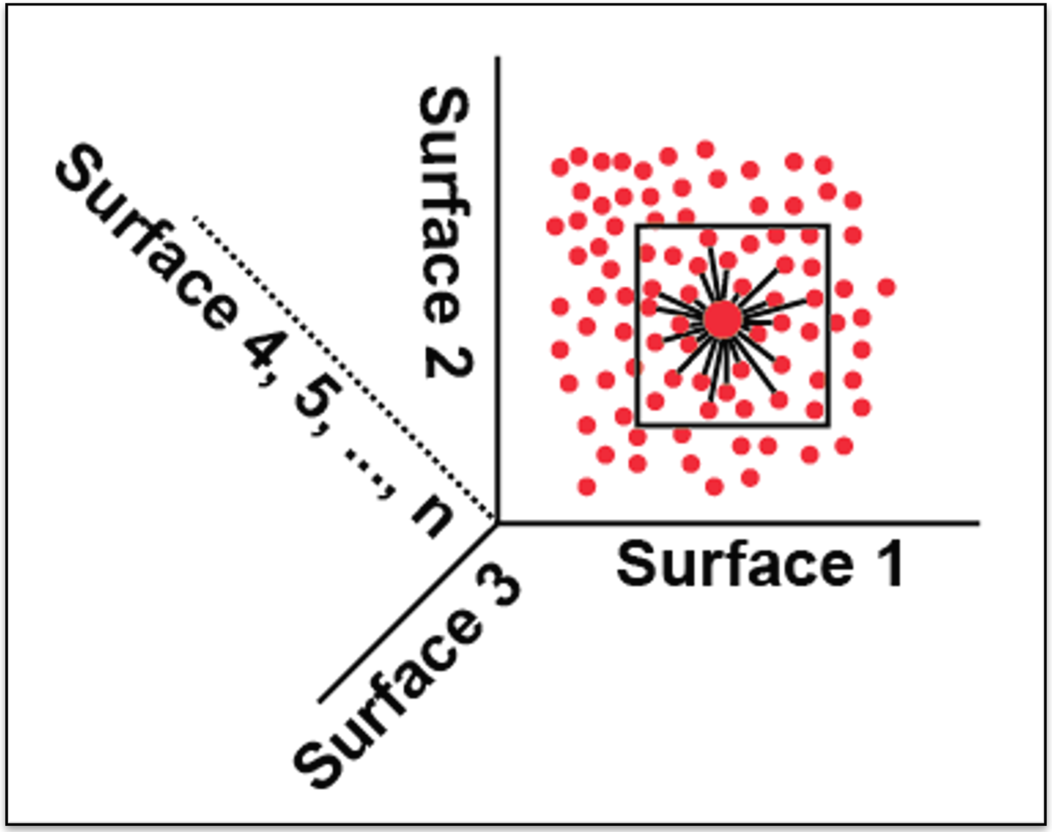
Dimr neighbors

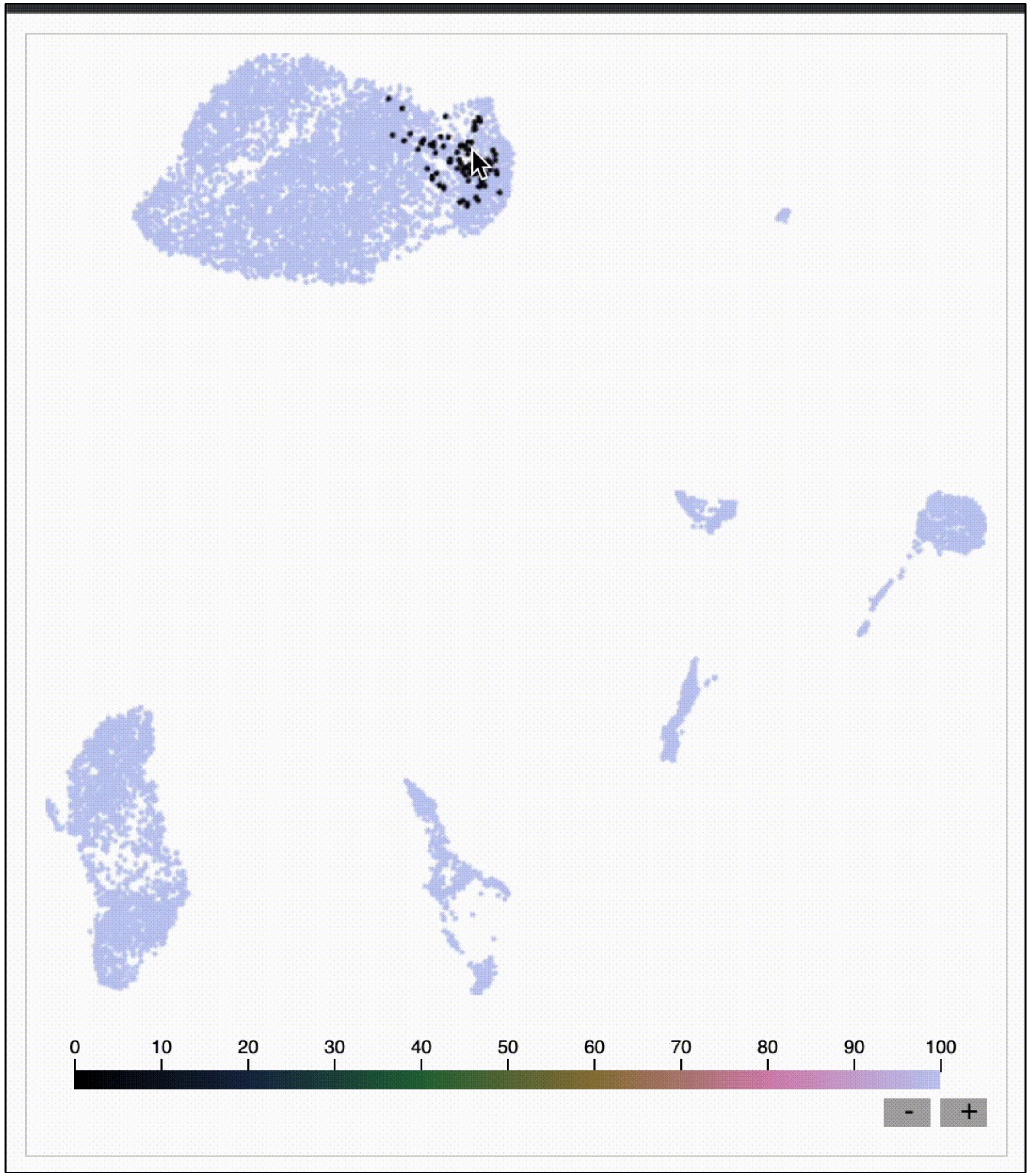


High-D neighbors

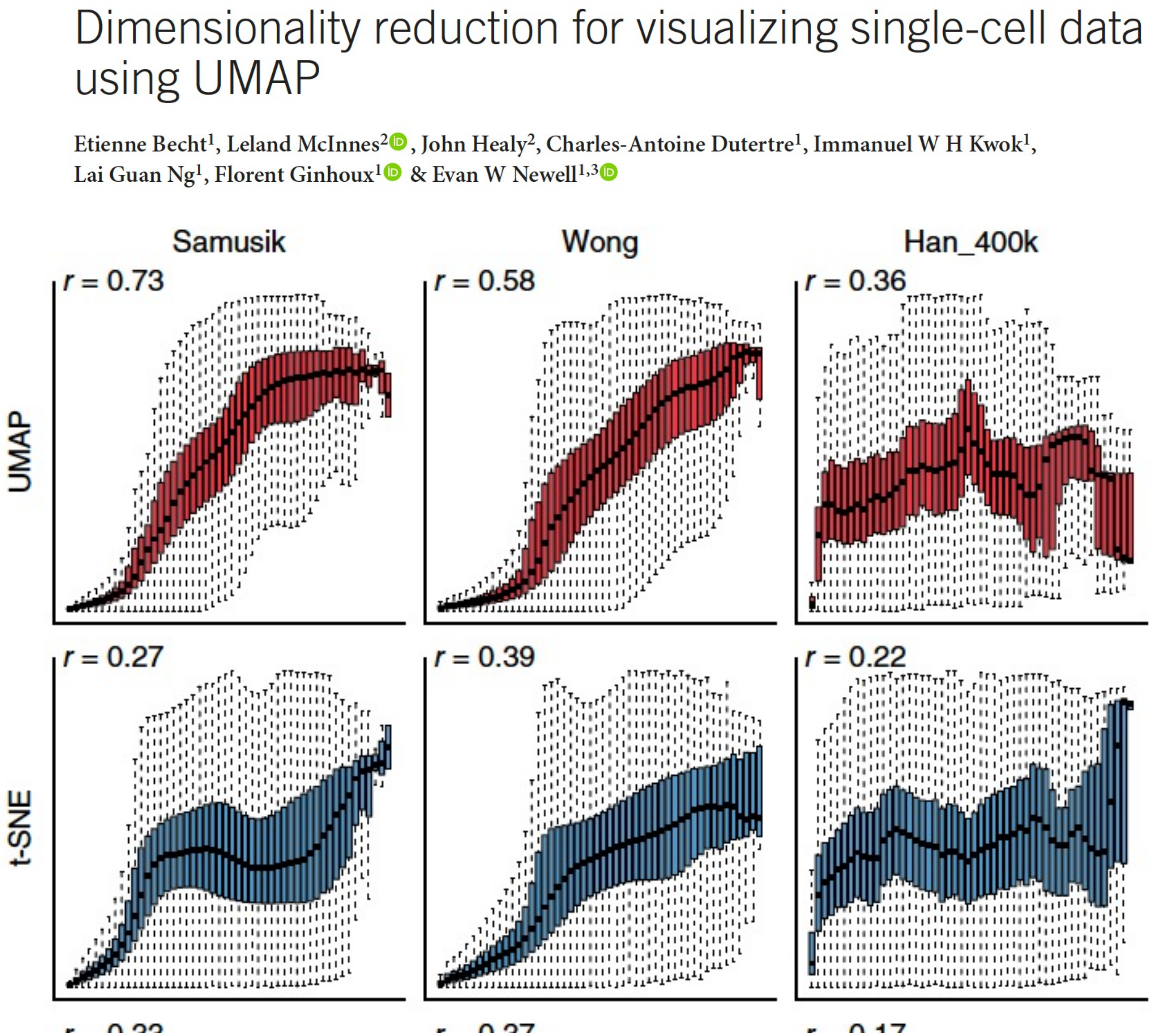


Use my tool knn_sleepwalk see the feature space KNN for your own data





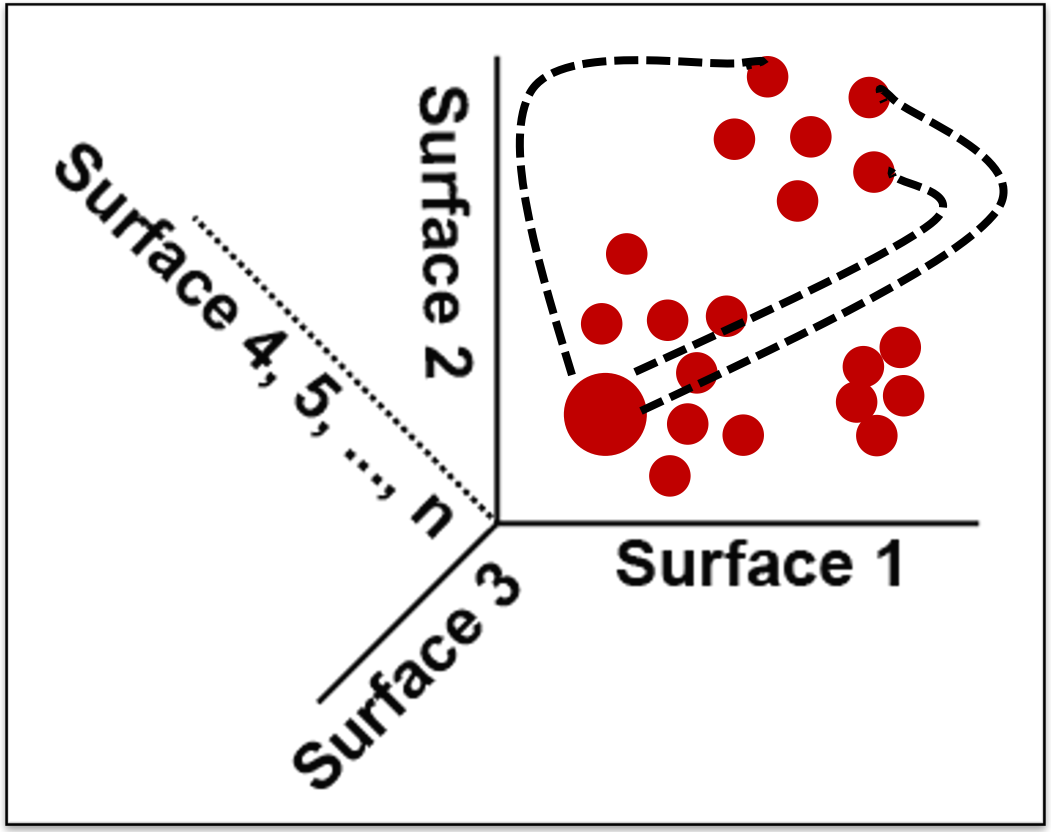
Global preservation, measured by pairwise distances

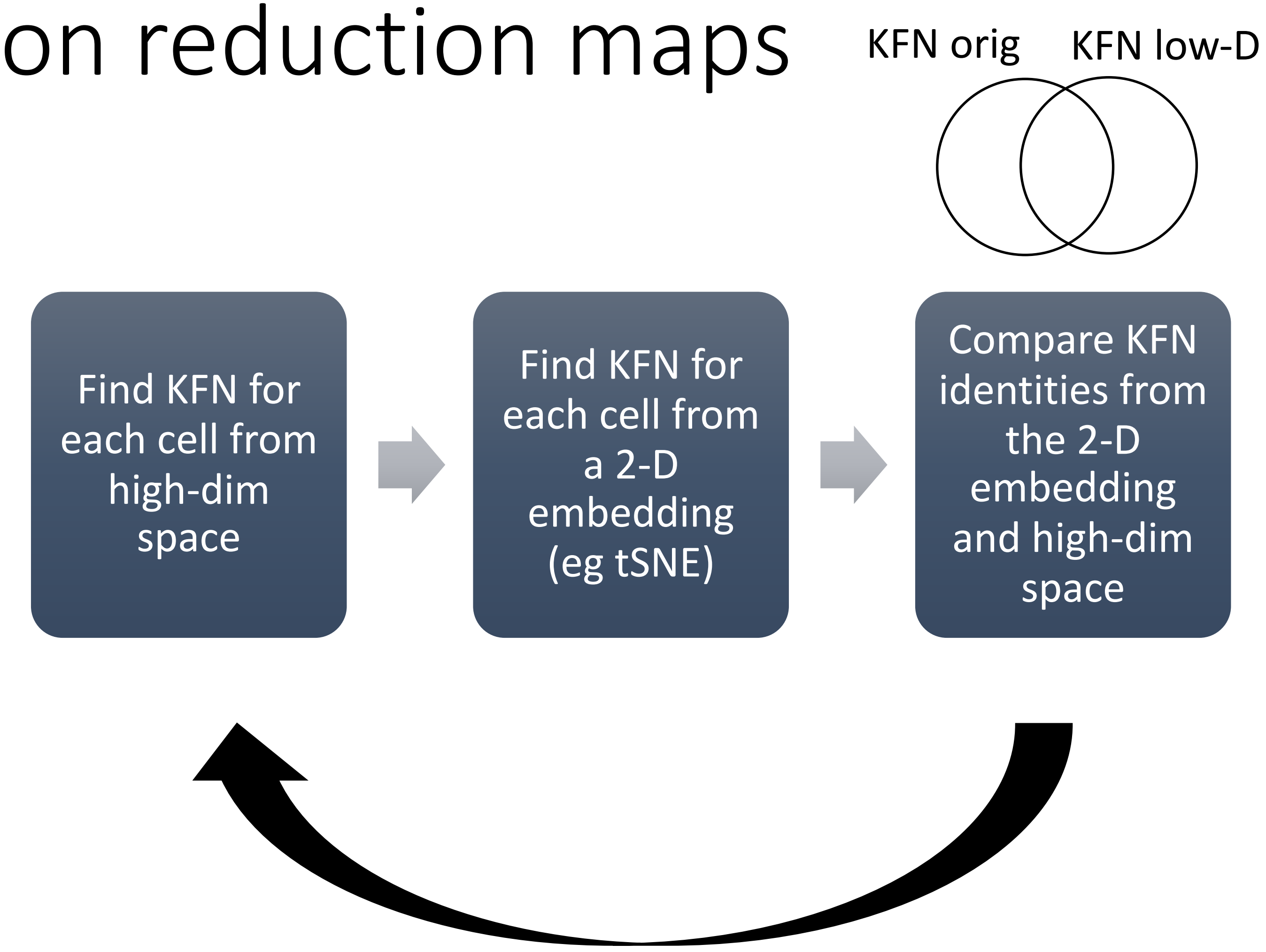


Dimensionality reduction for visualizing single-cell data using UMAP

Etienne Becht[1], Leland McInnes[2], John Healy[2], Charles-Antoine Dutertre[1], Immanuel W H Kwok[1], Lai Guan Ng[1], Florent Ginhoux[1] & Evan W Newell[1,3]

K-**farthest** neighbors (KFN) to determine global preservation of dimension reduction maps





Find KFN for each cell from high-dim space

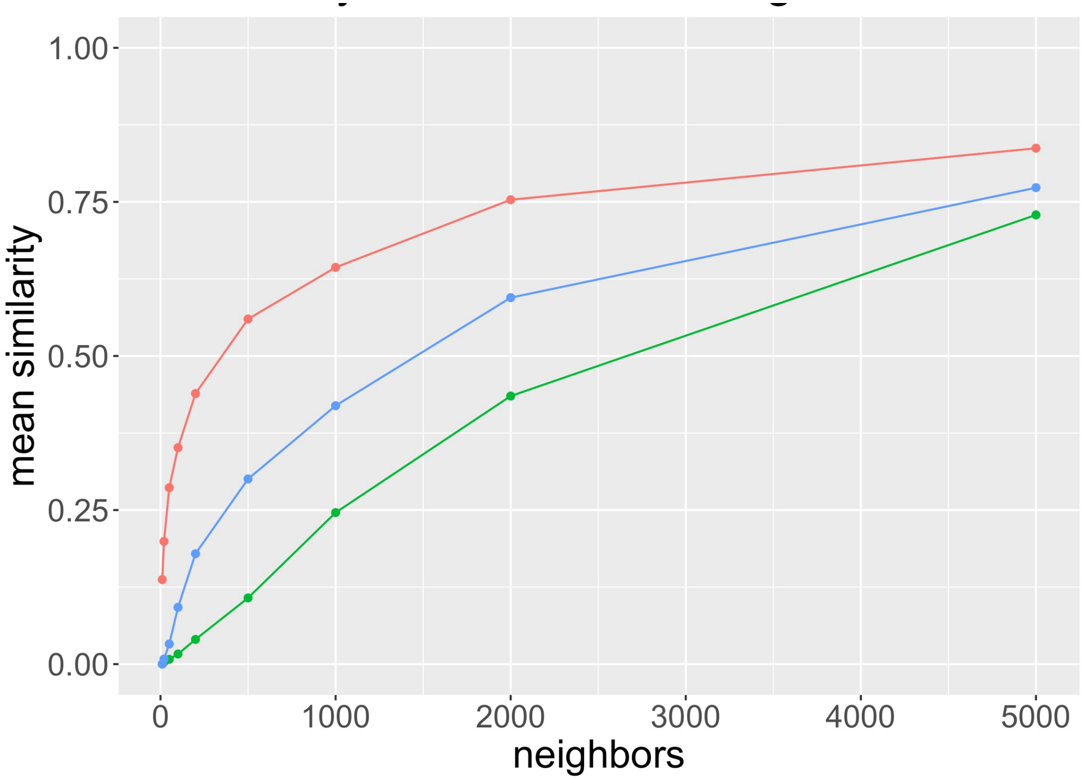
Find KFN for each cell from a 2-D embedding (eg tSNE)

Compare KFN identities from the 2-D embedding and high-dim space

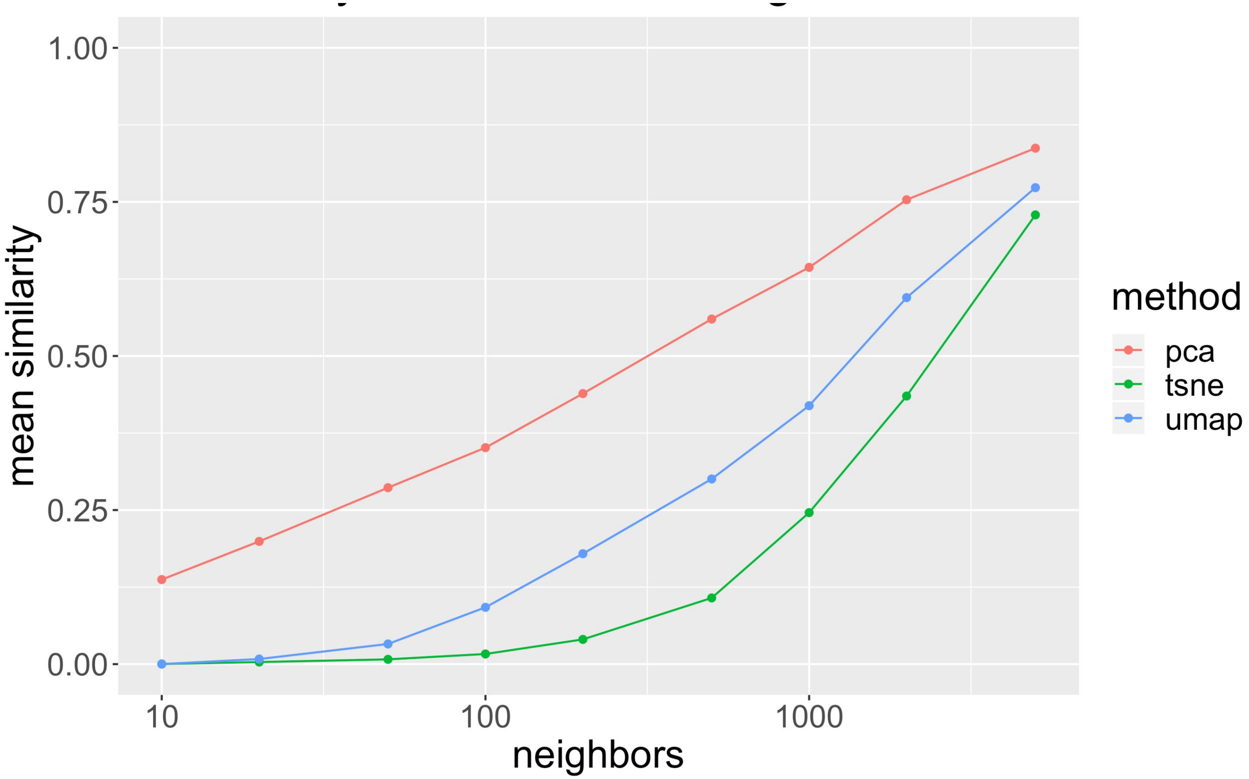
Repeat across a wide range of values for K

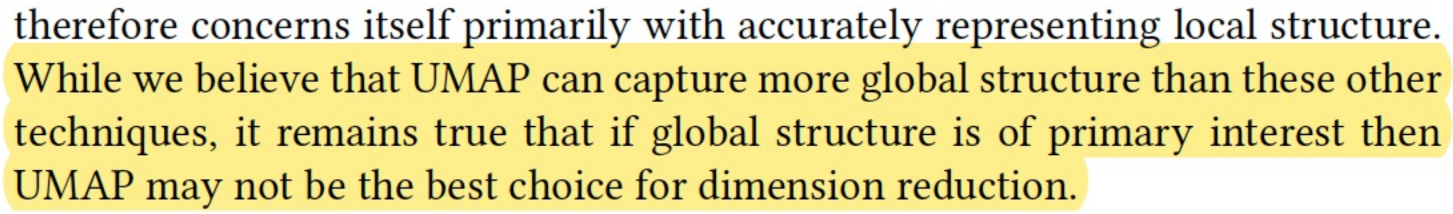
Bioconductor package: Sconify

Global KFN comparison between PCA, t-SNE and UMAP (10k cell subsample)



X axis is on a log scale

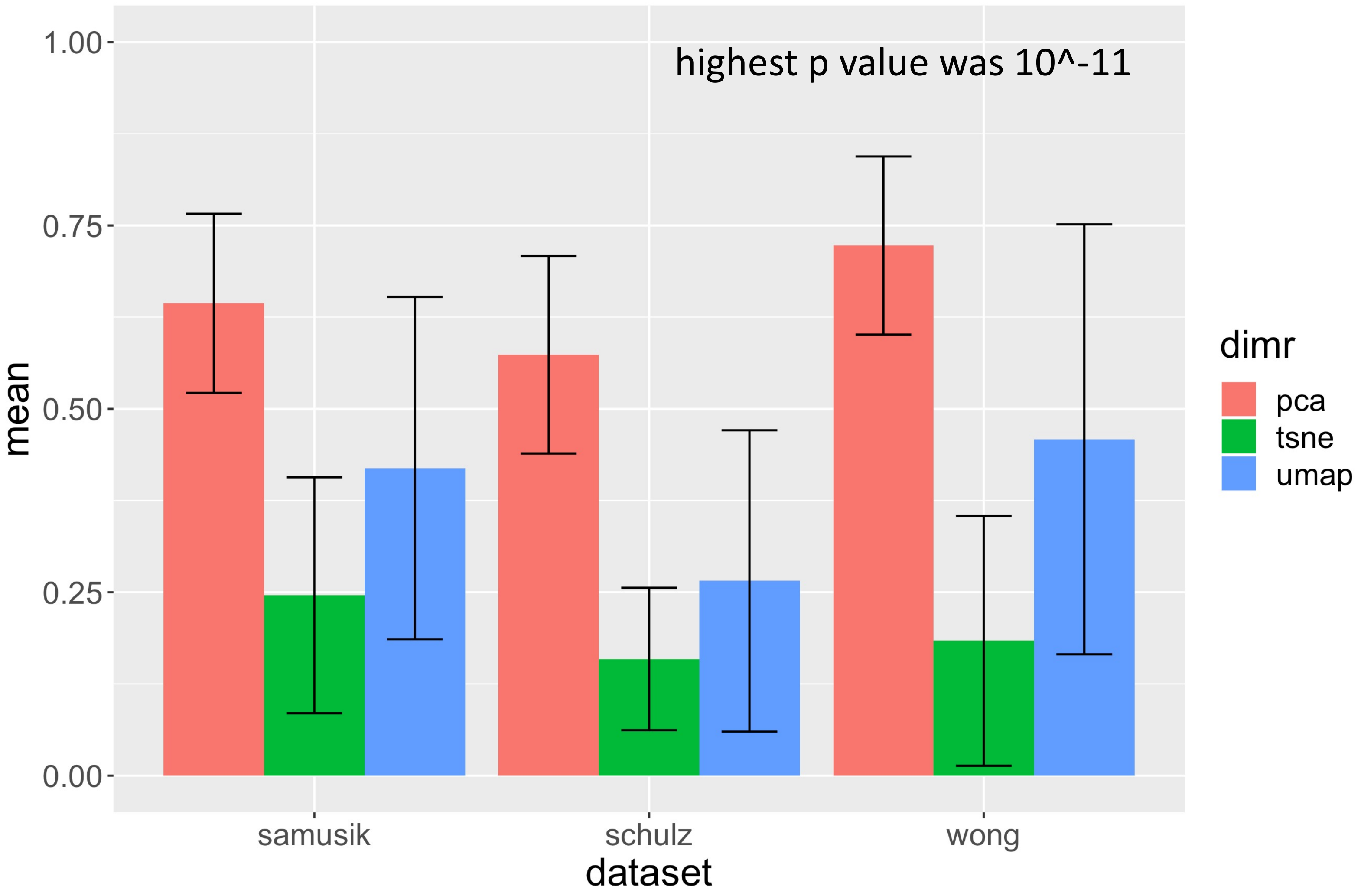




therefore concerns itself primarily with accurately representing local structure. While we believe that UMAP can capture more global structure than these other techniques, it remains true that if global structure is of primary interest then UMAP may not be the best choice for dimension reduction.

Mcinness *et al*, *Arxiv* 2018
(the UMAP paper)

Across 3 datasets, bar plots with error bars and p values



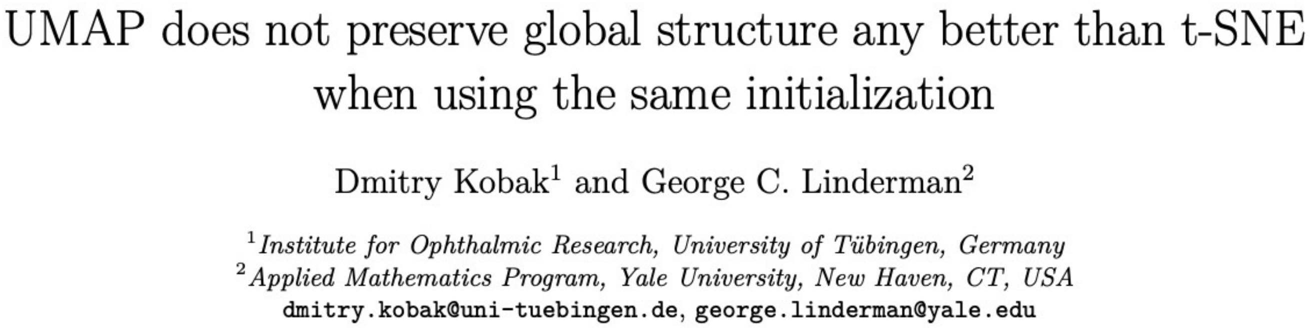
highest p value was 10^-11

Cells: 10k
k = 1000

Part 3 conclusions

- Nearest neighborhoods computed from high-D space and dimension reduction space occupy similar regions
- Positioning of the islands relative to each other could be arbitrary
- K-farthest neighborhood (KFN) preservation reveals global structure preservation: PCA > UMAP > t-SNE

Next steps: initialization matters

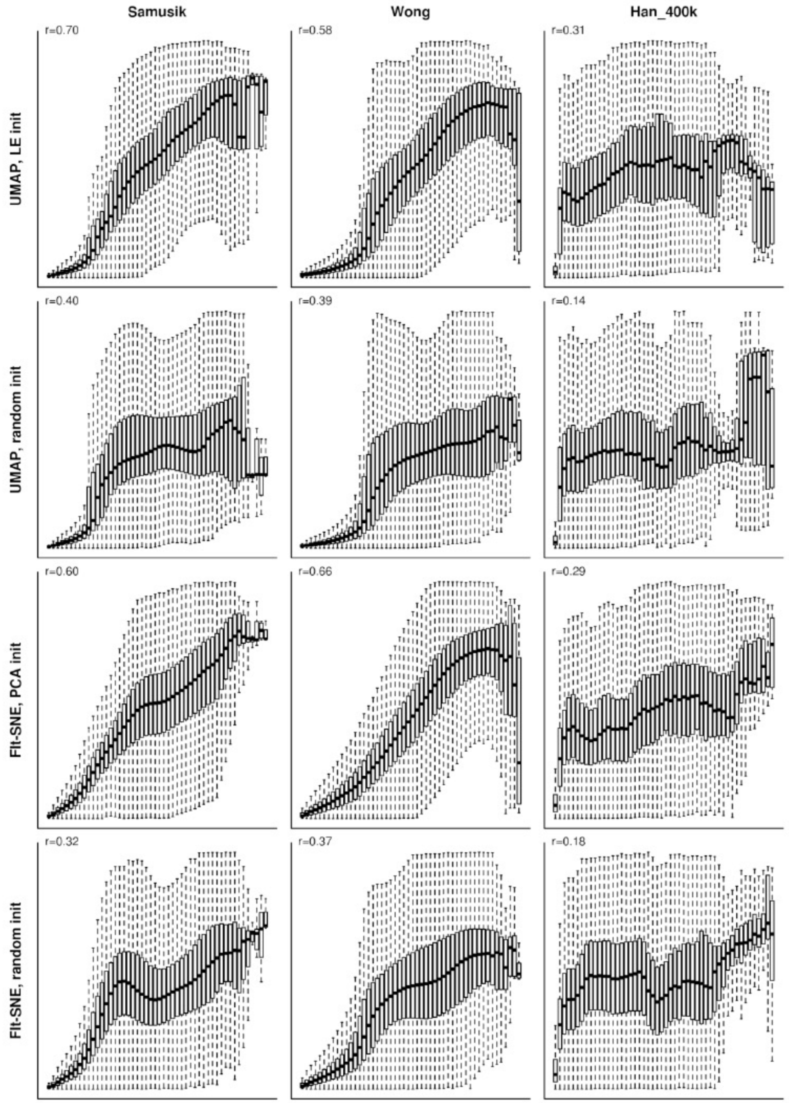


UMAP does not preserve global structure any better than t-SNE when using the same initialization

Dmitry Kobak[1] and George C. Linderman[2]

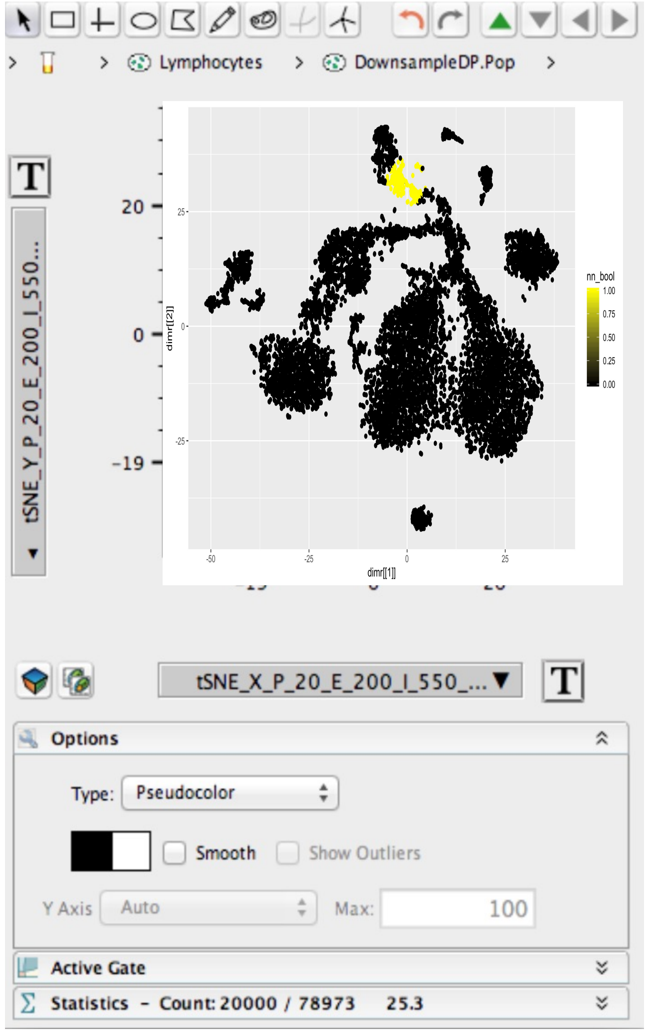
[1] *Institute for Ophthalmic Research, University of Tübingen, Germany*
[2] *Applied Mathematics Program, Yale University, New Haven, CT, USA*
dmitry.kobak@uni-tuebingen.de, george.linderman@yale.edu

(BioRxiv)



Toward a ”safe” manual gating interface for dimension reduction maps

Identity color (dimr)



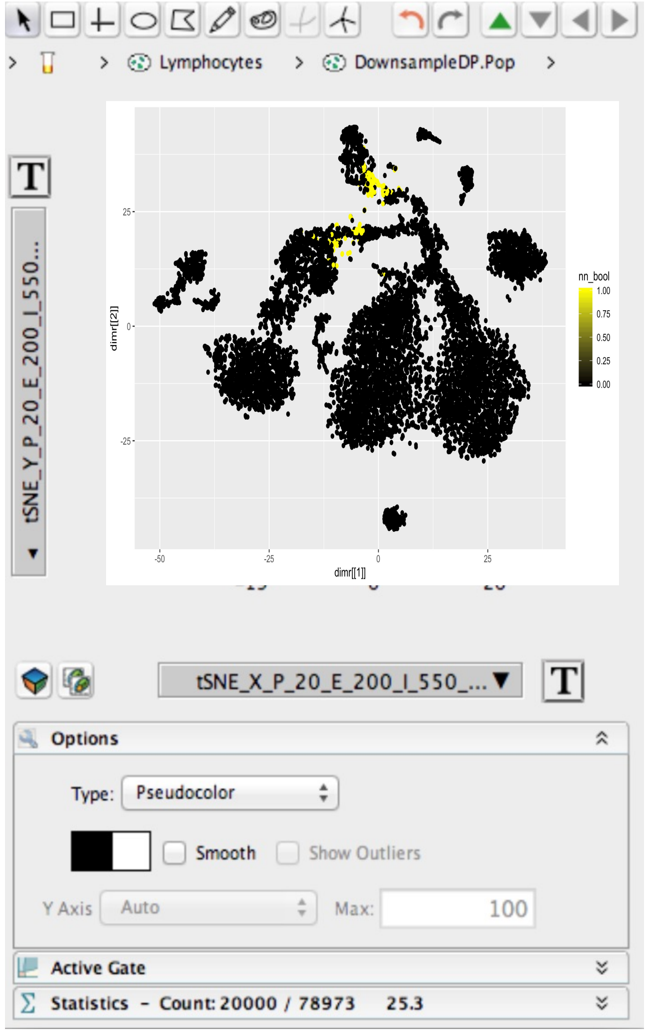
Type: Pseudocolor

□ Smooth □ Show Outliers

Y Axis Auto Max: 100

Statistics – Count: 20000 / 78973 25.3

Identity color (high-d)



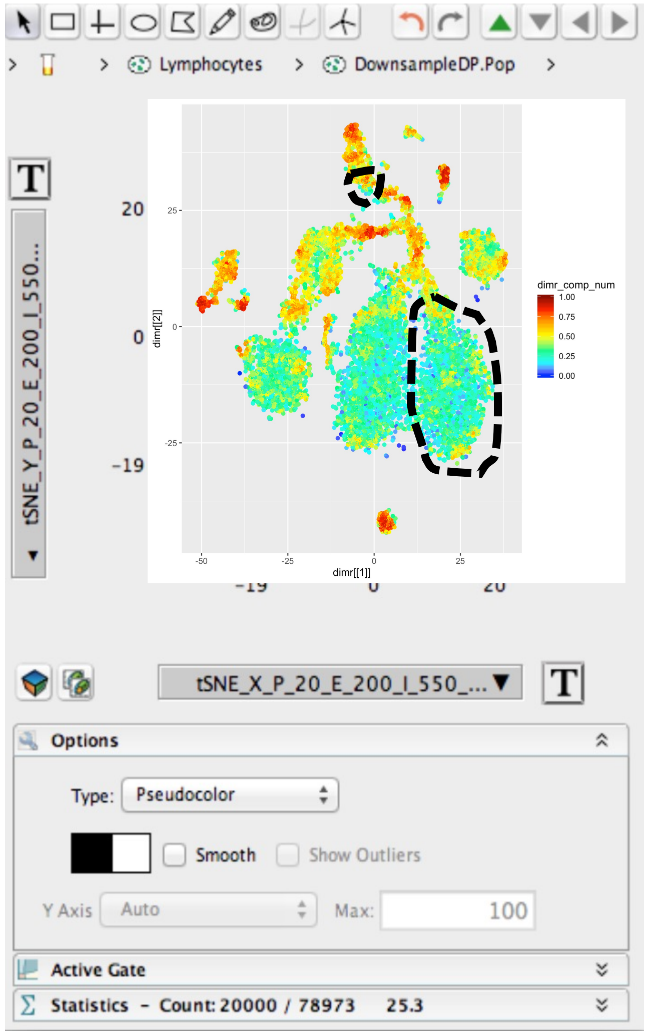
Type: Pseudocolor

□ Smooth □ Show Outliers

Y Axis Auto Max: 100

Statistics – Count: 20000 / 78973 25.3

Identity comparison



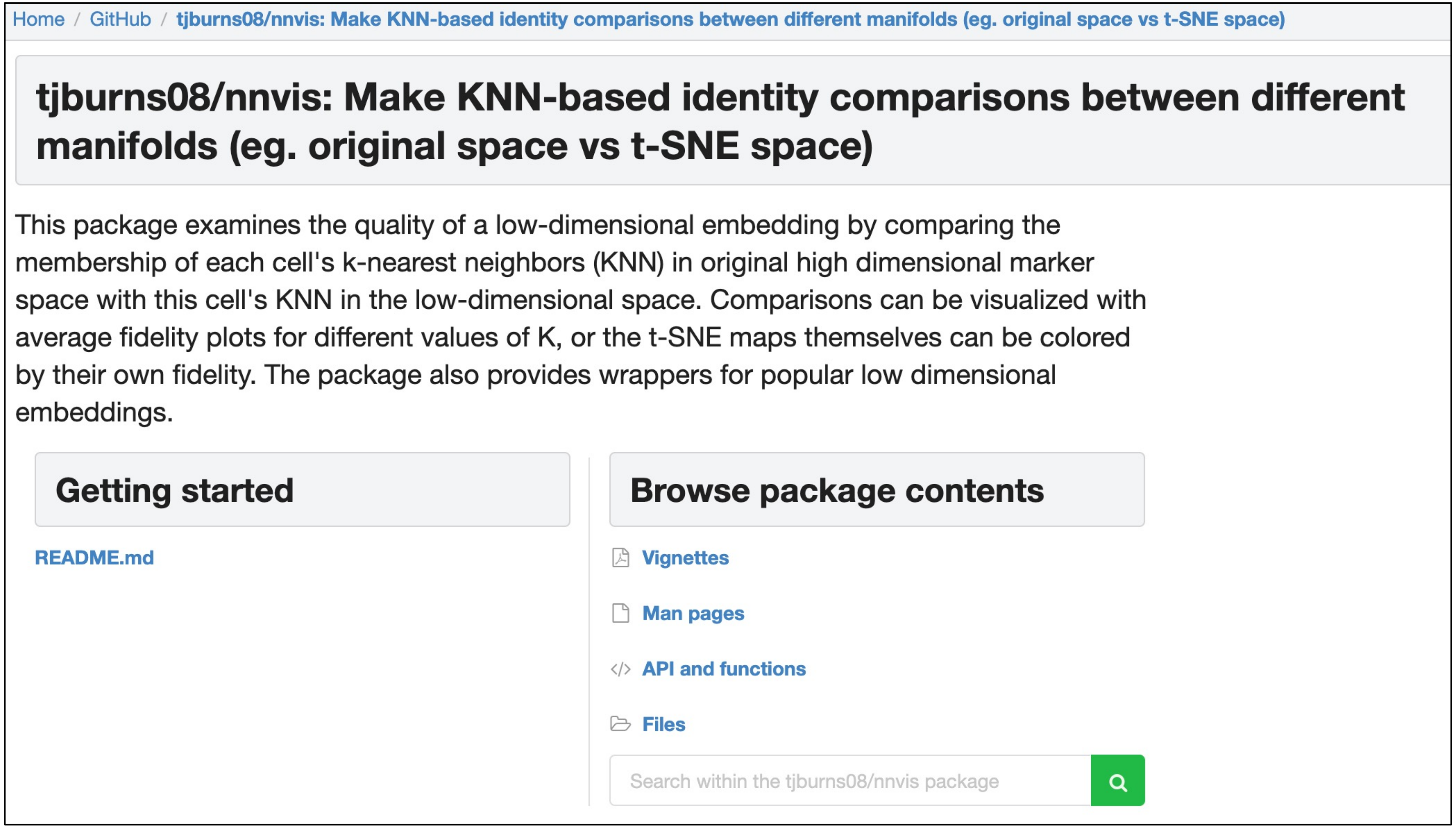
Type: Pseudocolor

□ Smooth □ Show Outliers

Y Axis Auto Max: 100

Statistics – Count: 20000 / 78973 25.3

nnvis: an R package to do neighbor-based preservation analysis on your dimension reductions



Home / GitHub / tjburns08/nnvis: Make KNN-based identity comparisons between different manifolds (eg. original space vs t-SNE space)

tjburns08/nnvis: Make KNN-based identity comparisons between different manifolds (eg. original space vs t-SNE space)

This package examines the quality of a low-dimensional embedding by comparing the membership of each cell's k-nearest neighbors (KNN) in original high dimensional marker space with this cell's KNN in the low-dimensional space. Comparisons can be visualized with average fidelity plots for different values of K, or the t-SNE maps themselves can be colored by their own fidelity. The package also provides wrappers for popular low dimensional embeddings.

Getting started

README.md

Browse package contents

Vignettes

Man pages

API and functions

Files

Search within the tjburns08/nnvis package

Acknowledgments

Heike Hirseland Antonia Niedobitek René Riedel

Sarah Gräßle Marie Urbicht Silke Stanislawiak

DRFZ BERLIN

Deutsches Rheuma-Forschungszentrum
Ein Institut der Leibniz-Gemeinschaft

Sabine Baumgart **Andreas Grützkau** Pawel Durek

Henrik Mei **Axel Schulz**