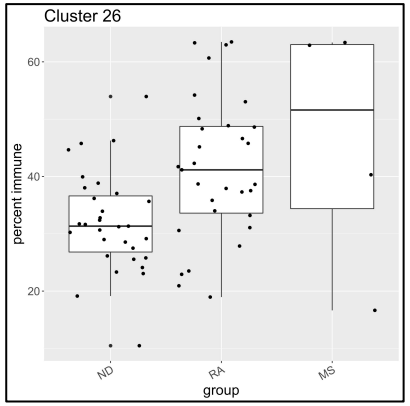
A comprehensive interrogation of the t-SNE algorithm for mass cytometry analysis

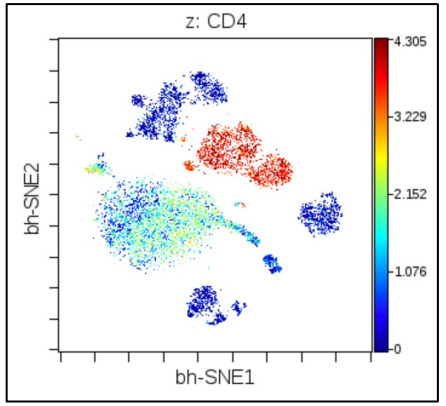
Tyler J Burns, PhD

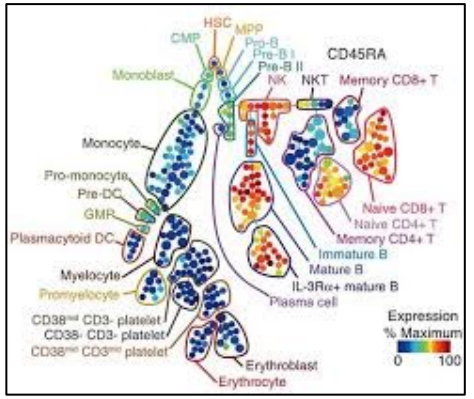
AG Mei

Deutsches Rheuma Forschung Zentrum

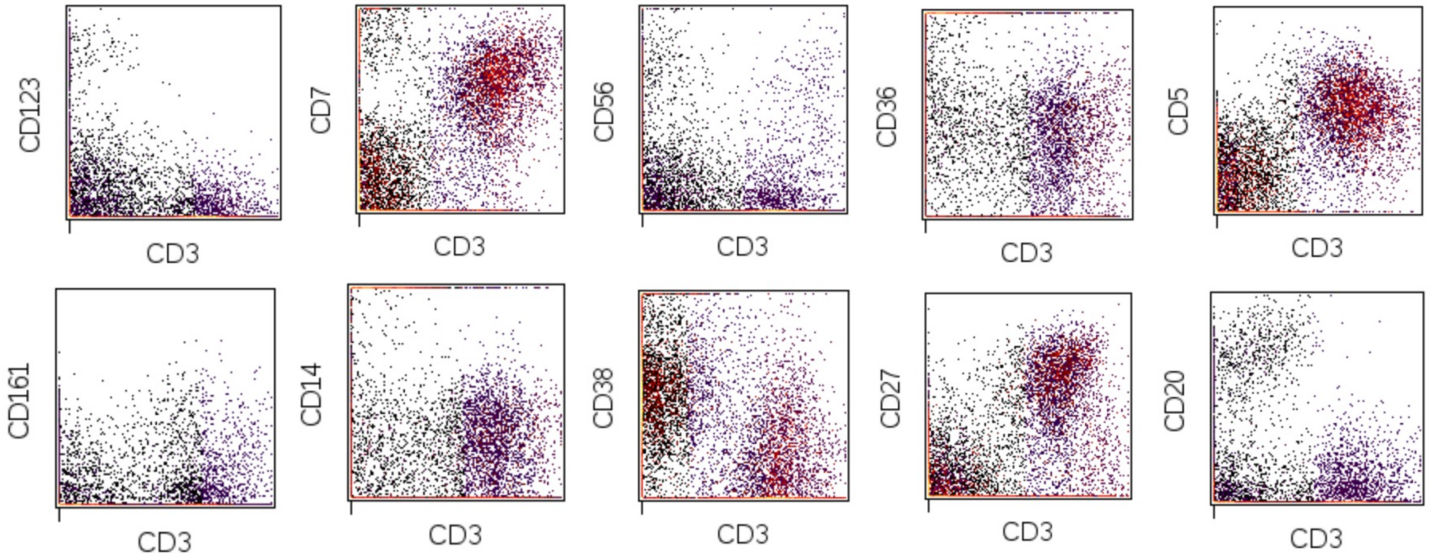
The big picture: from machine to human





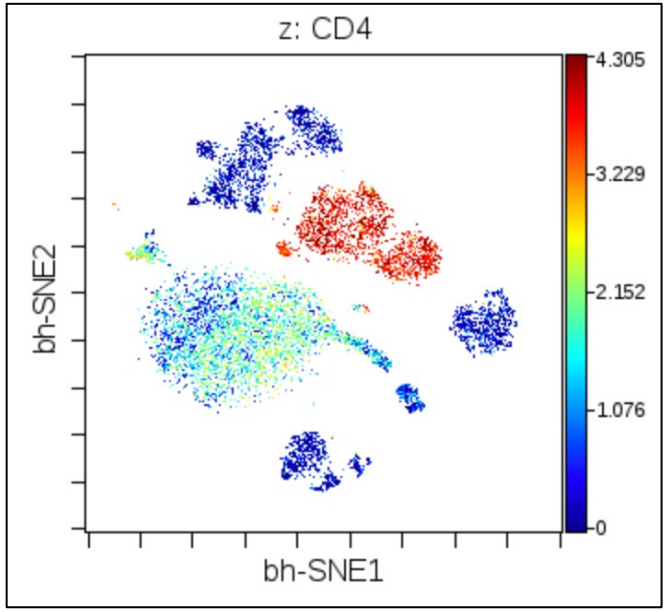


t-SNE makes large amounts of information human-palatable without too much human work




```
# 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

Background: the t-SNE algorithm as a dimension reducer

Visualizing Data using t-SNE

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

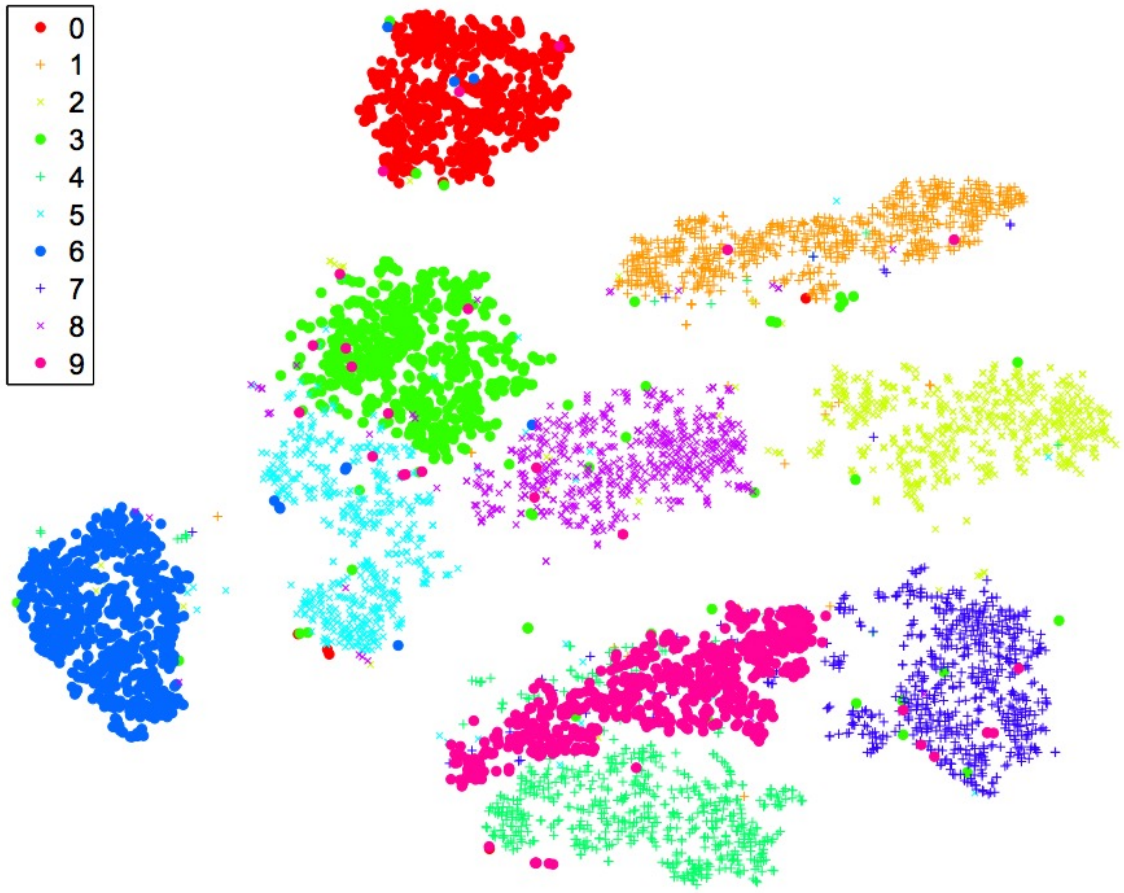
Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

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

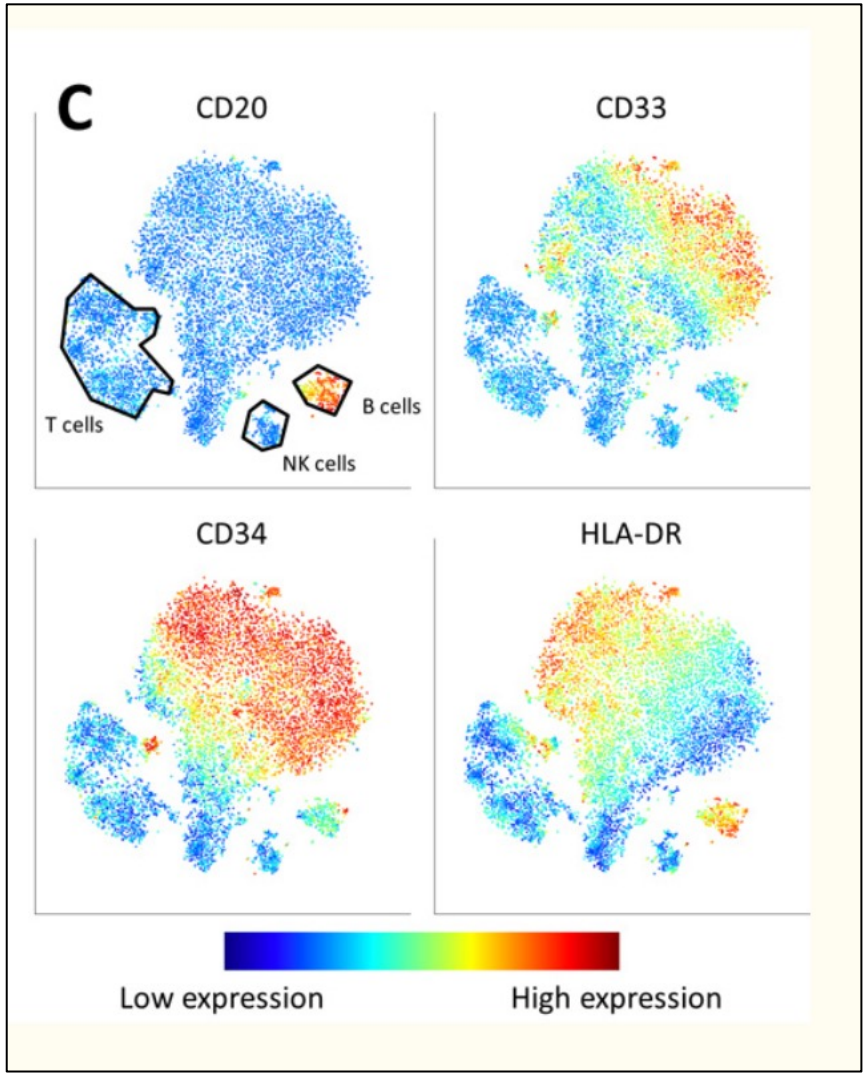
El-ad David Amir,[1] Kara L Davis,[2,3] Michelle D Tadmor,[1,3] Erin F Simonds,[2,3] Jacob H Levine,[1,3] Sean C Bendall,[2,3] Daniel K Shenfeld,[1,3] Smita Krishnaswamy,[1] Garry P Nolan,[2,4] and Dana Pe'er[1,4,*]

Author information ▶ Copyright and License information ▶ Disclaimer

Abstract

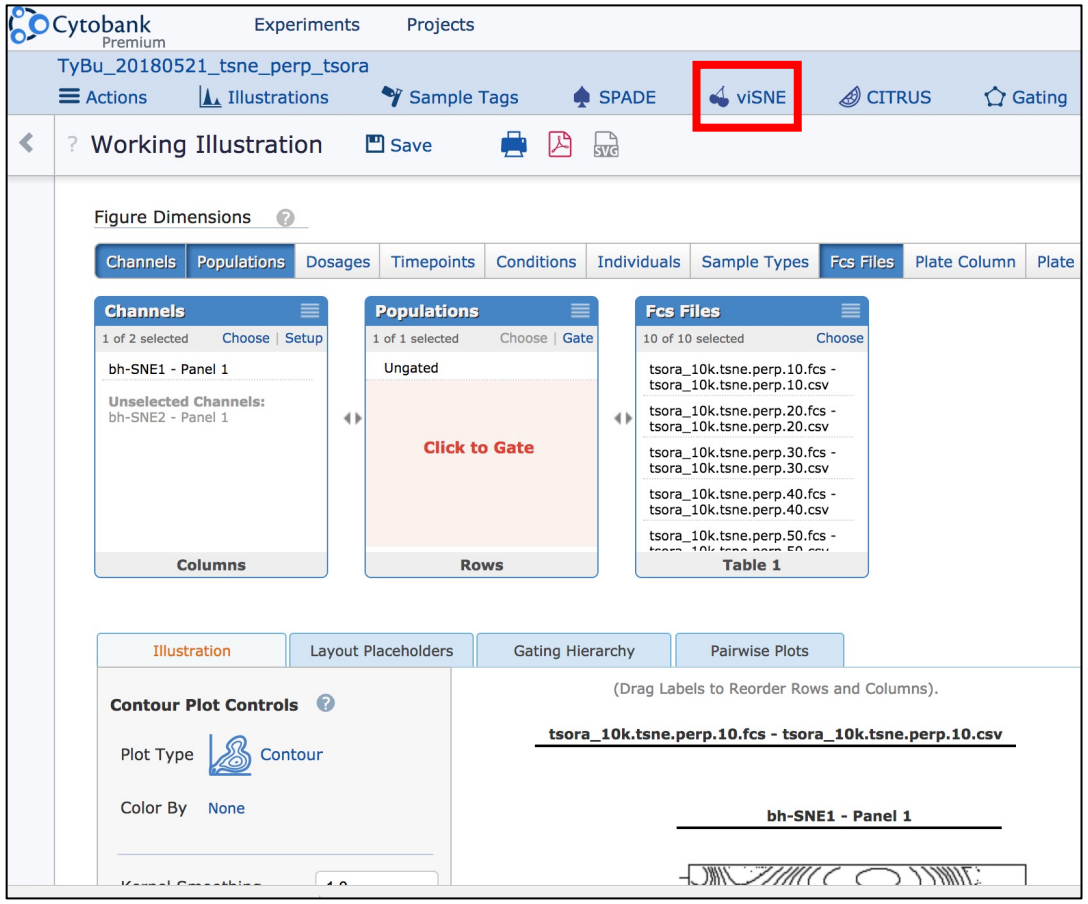
Go to:

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.

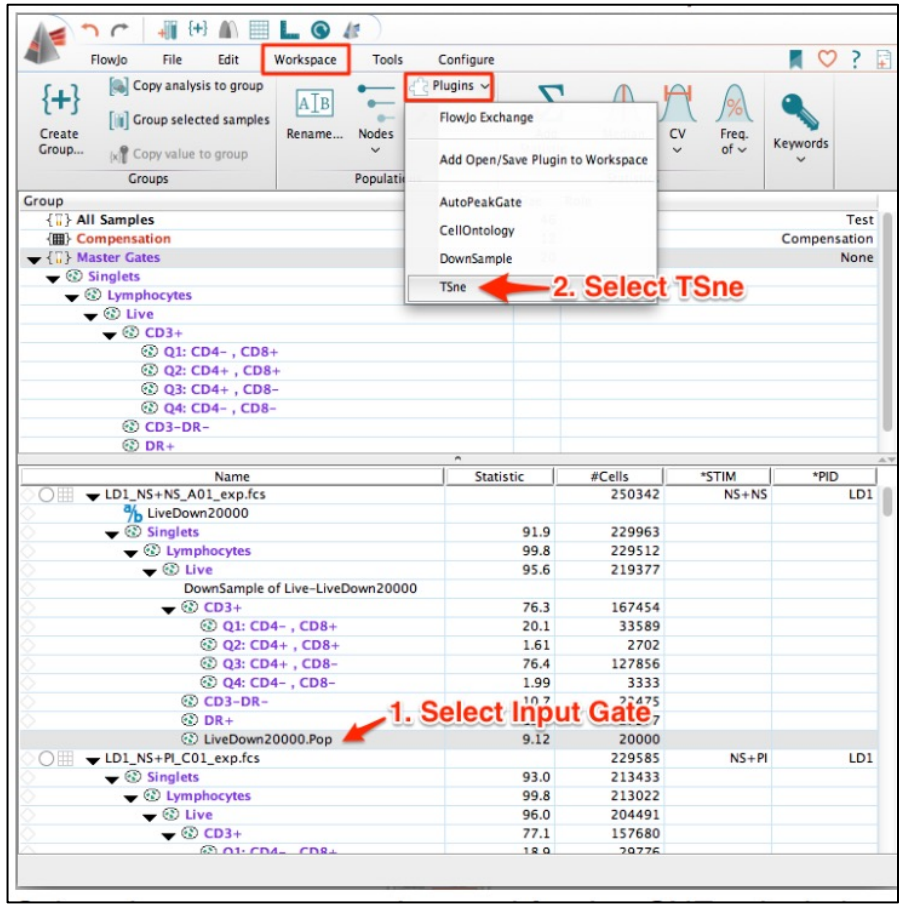


There are many other dimension reduction tools, but t-SNE is the most accessible

Cytobank

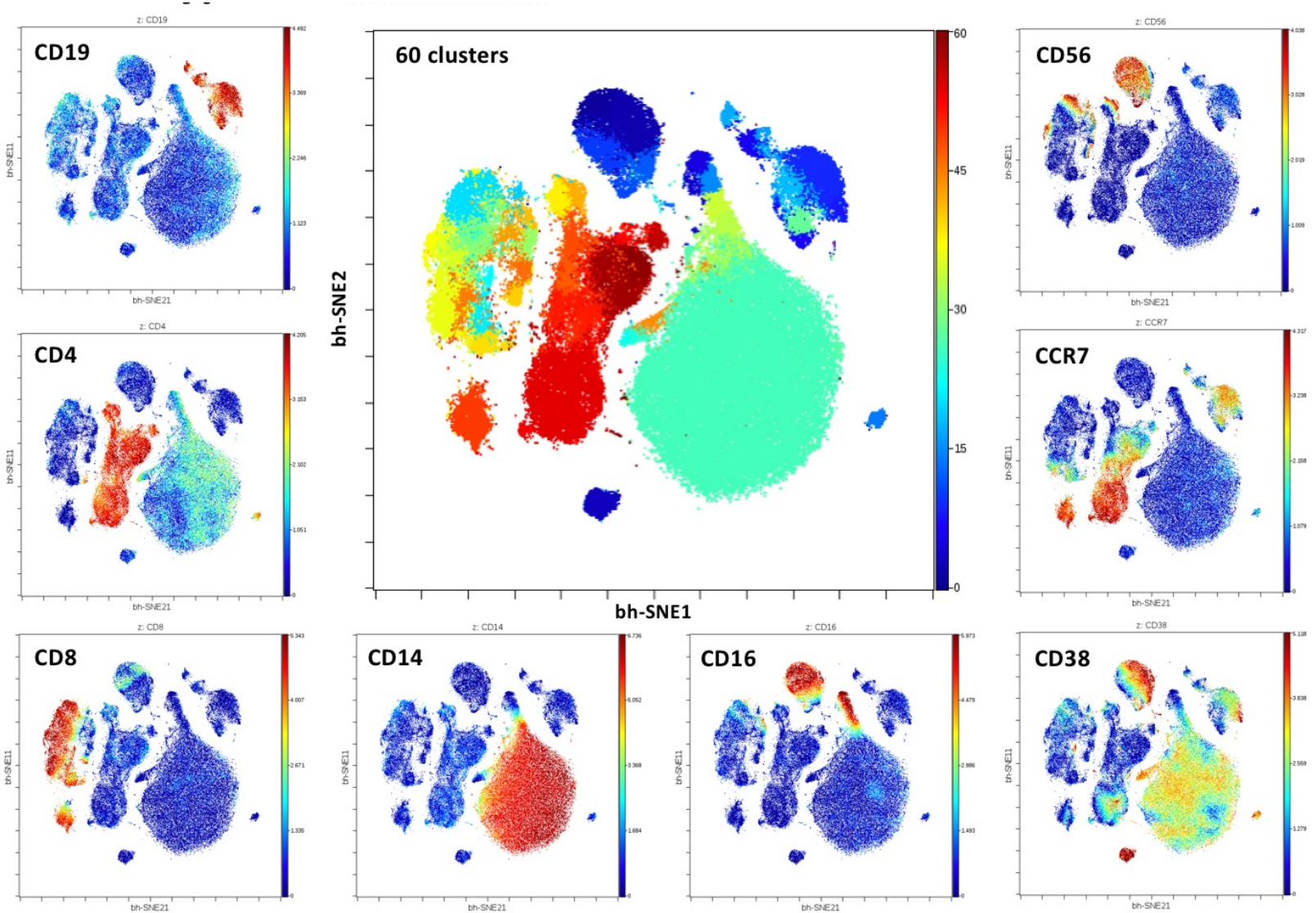


FlowJo



But what hidden information about t-SNE should we know for its proper use?

Why is t-SNE popular? In part because it looks nice and major subsets group together



Axel Schulz, Ph.D.

Credit for the following t-SNE visualization slides

StatQuest: t-SNE, Clearly Explained

17,938 views

489 10

StatQuest with Josh Starmer
Published on Sep 18, 2017

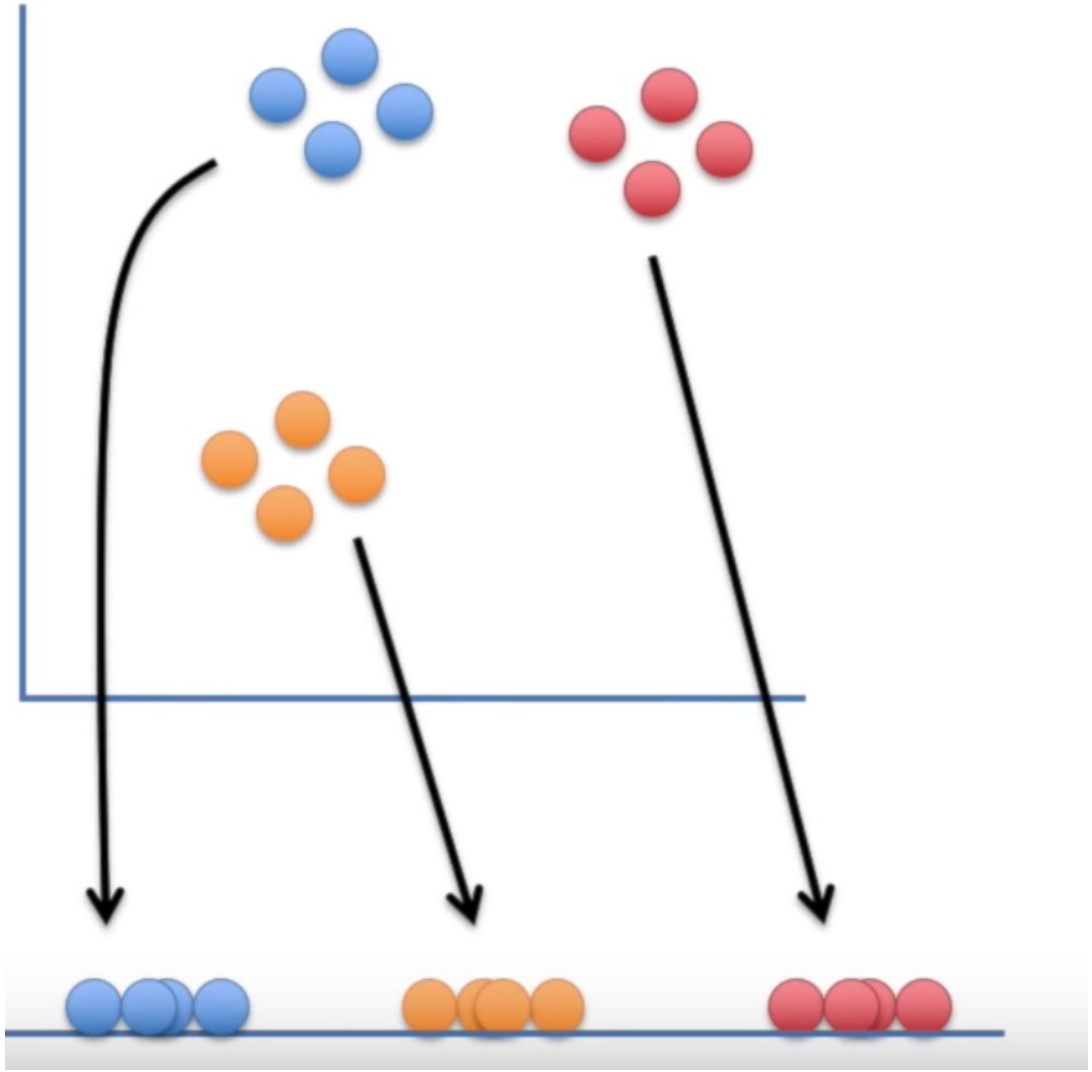
S

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:

SHOW MORE

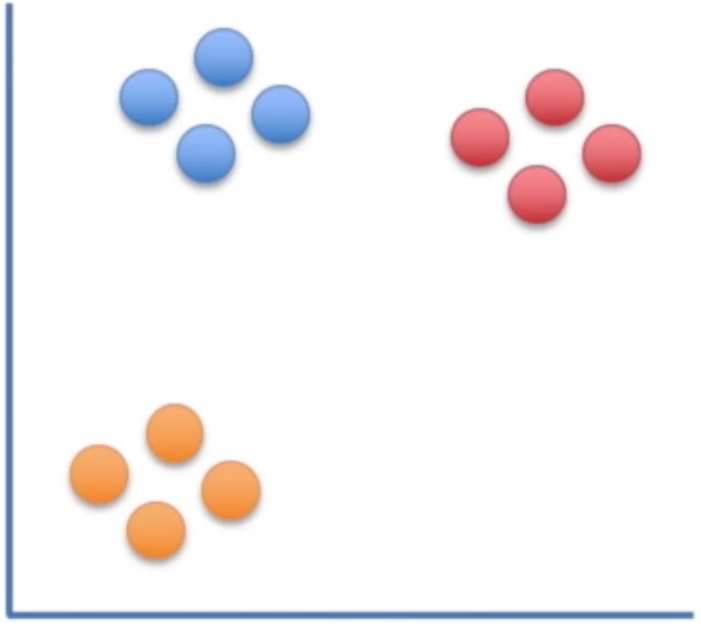
The goal of t-SNE is to reduce dimensions while preserving specific information

Higher dimensional space

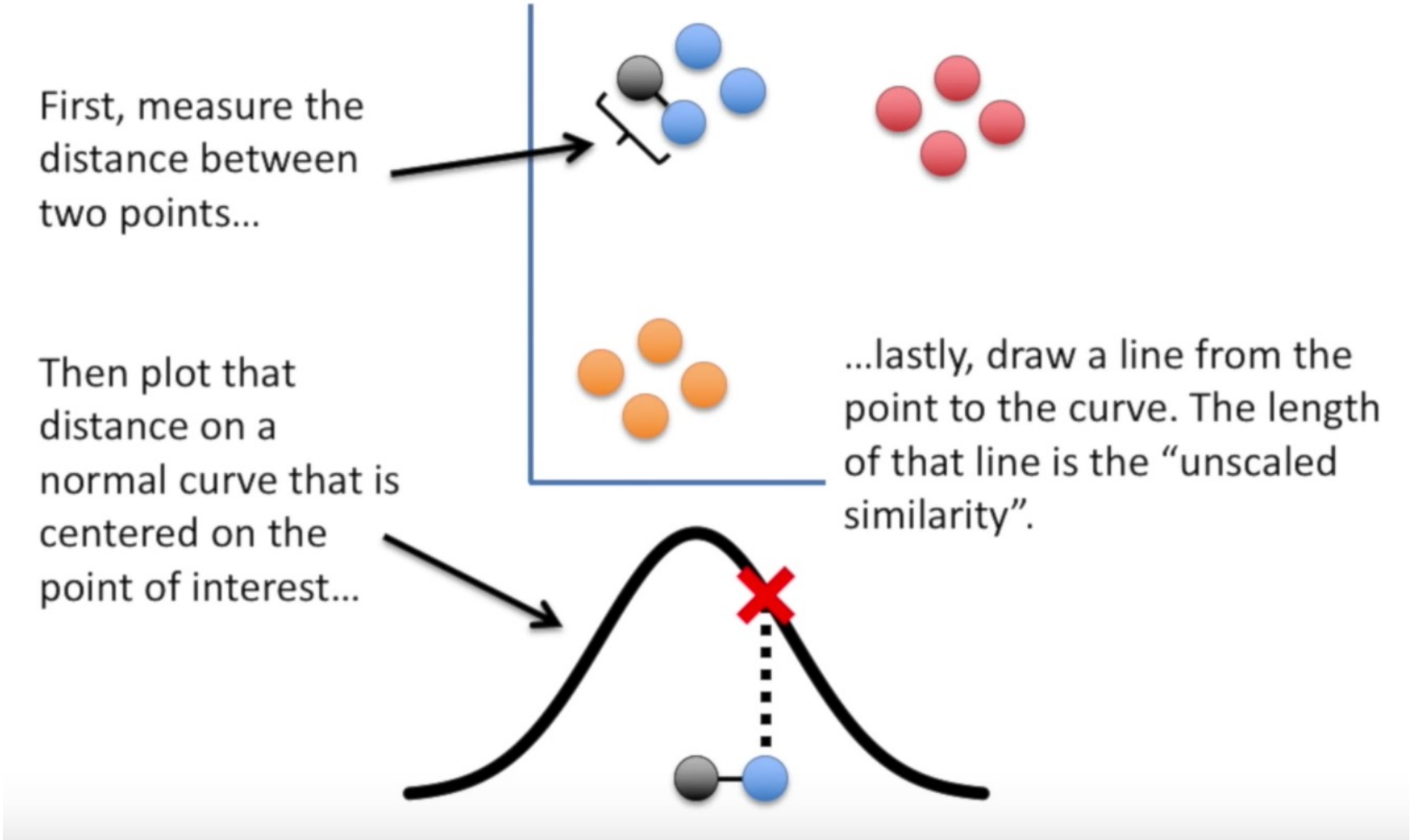


Low dimensional embedding

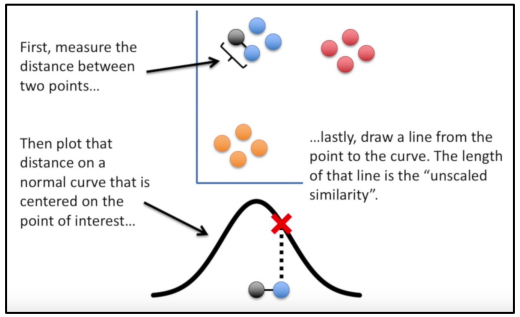
t-SNE starts with a low-d embedding of randomly placed points



t-SNE makes similarity scores…like distance but fitted to a distribution



These similarity scores go into similarity matrices



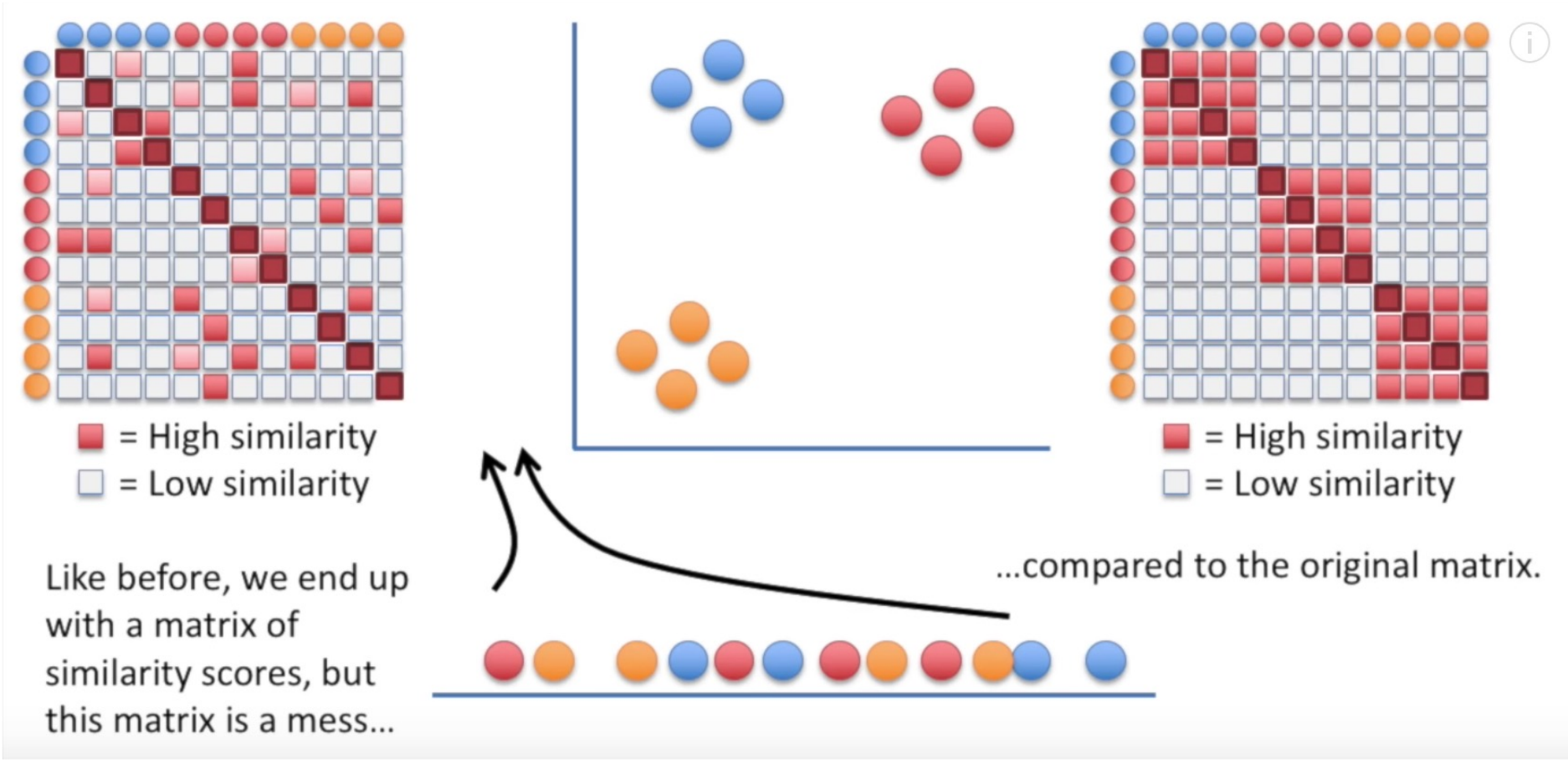
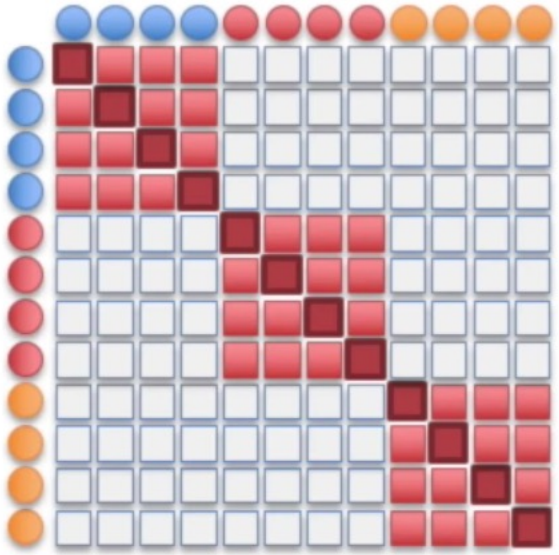
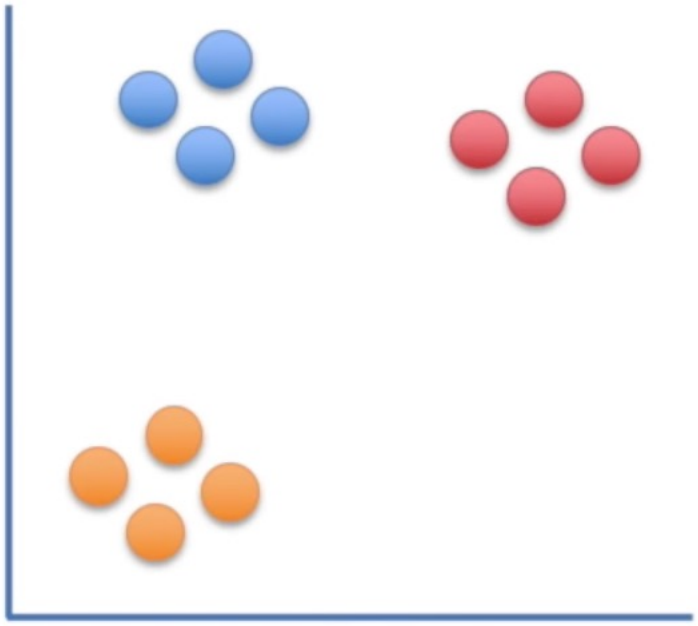
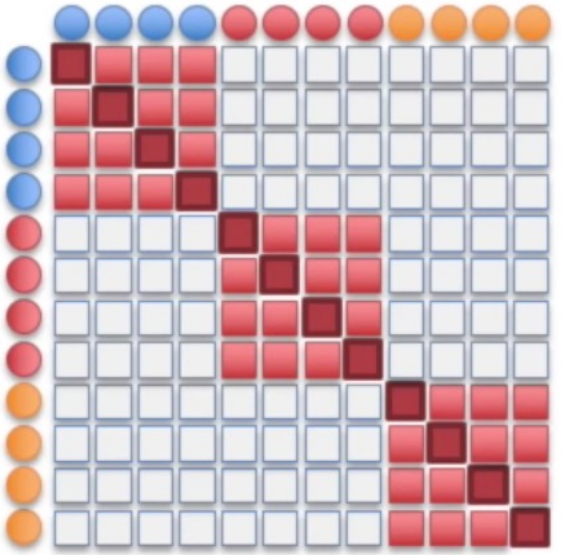


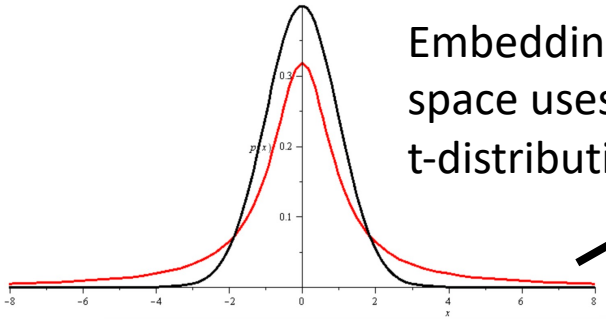
Image source: YouTube: StatQuest with Josh Starmer: *t-SNE, clearly explained*

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









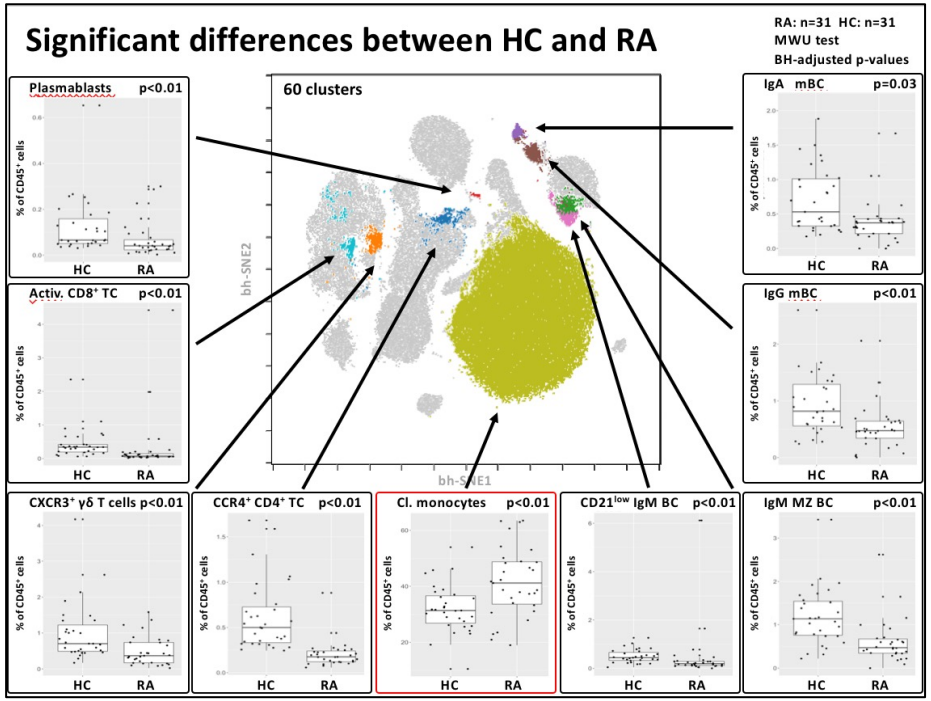
Embedding space uses t-distribution

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

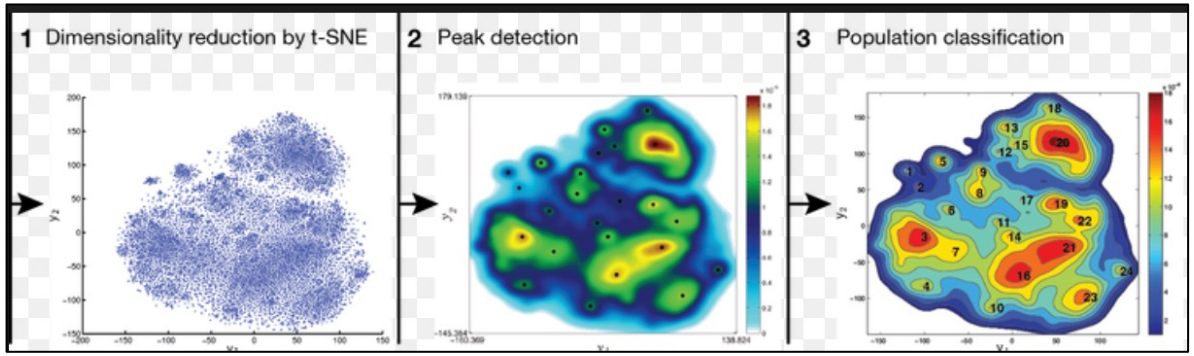
Two main uses of t-SNE

- Data visualization tool
  - Early phases: gain intuition about data
  - Late phases: summarize statistical output
- Part of a data analysis pipeline
  - Gating a t-SNE map
  - Clustering a t-SNE map

Example: Axel Schulz, AG Mei



Example: ACCENSE



The organization of my talk

- Part 1: Show how varying input affects t-SNE output (so you don’t have to)

- Part 2: Determine whether we can and/or should gate and cluster a t-SNE map

The organization of my talk

- **Part 1: Show how varying input affects t-SNE output (so you don't have to)**

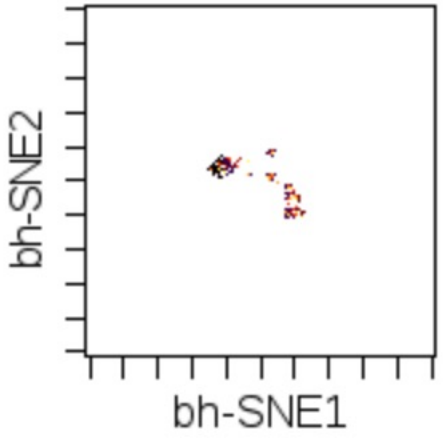
- Part 2: Determine whether we can and/or should gate and cluster a t-SNE map

What happens to t-SNE output when you vary the number of cells?

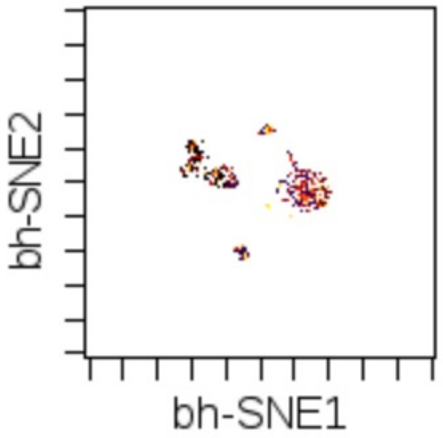
- t-SNE is typically viewed on a sub-sampled data due to run-time issues
- Data: healthy human PBMCs
- Procedure: run t-SNE with subsampled cells, ranging from 100 to 200,000.
- Visualize as a biaxial plot colored by the kernel density estimation
- Check to make sure the major subsets are still being compartmentalized

Altering the number of cells affects the amount of embedding space

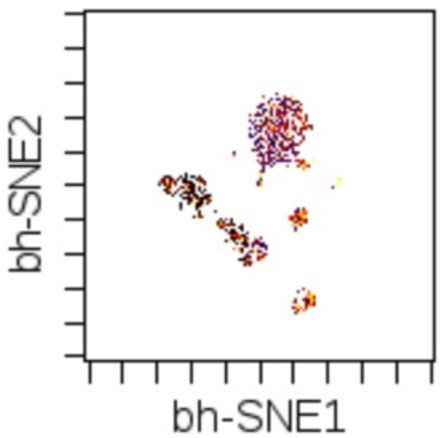
200 cells



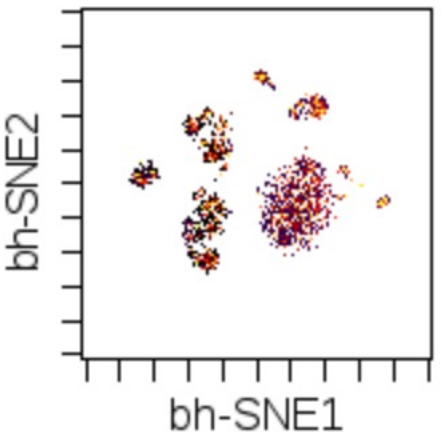
500 cells



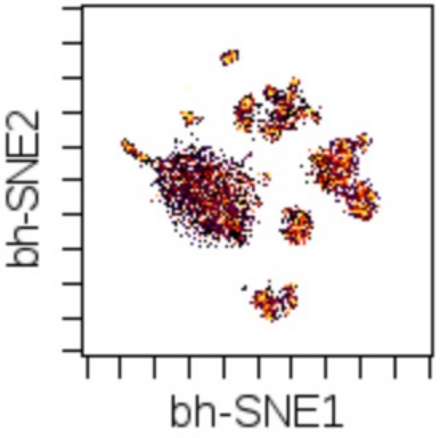
1000 cells



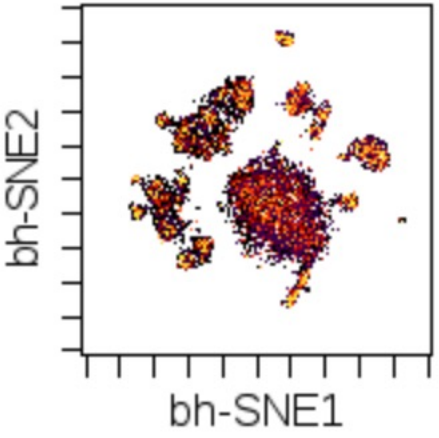
2000 cells



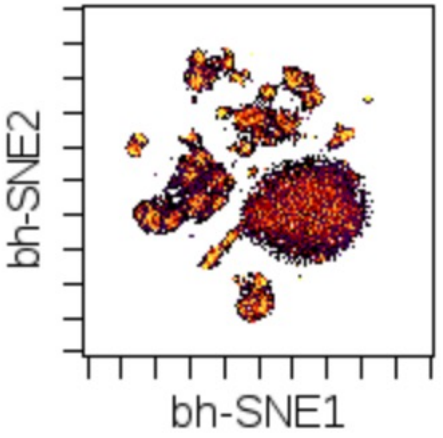
5000 cells



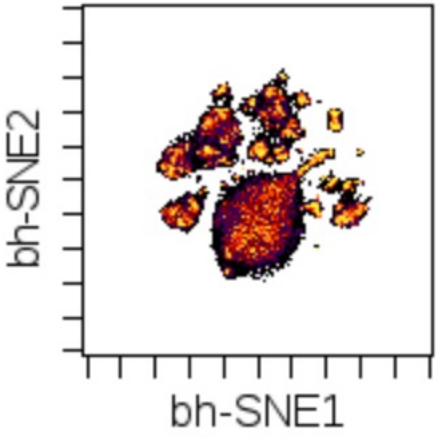
10000 cells



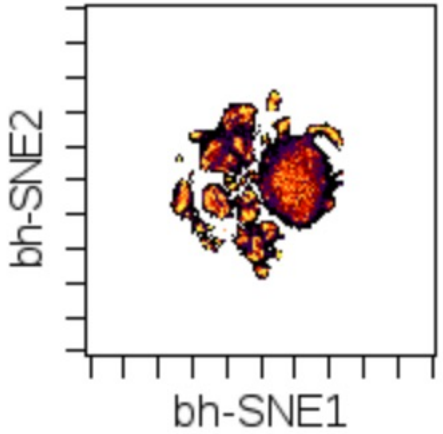
20000 cells



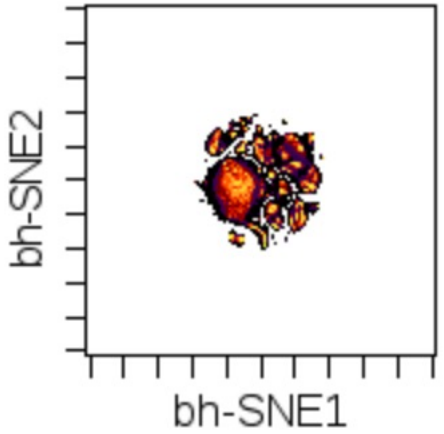
50000 cells



100000 cells

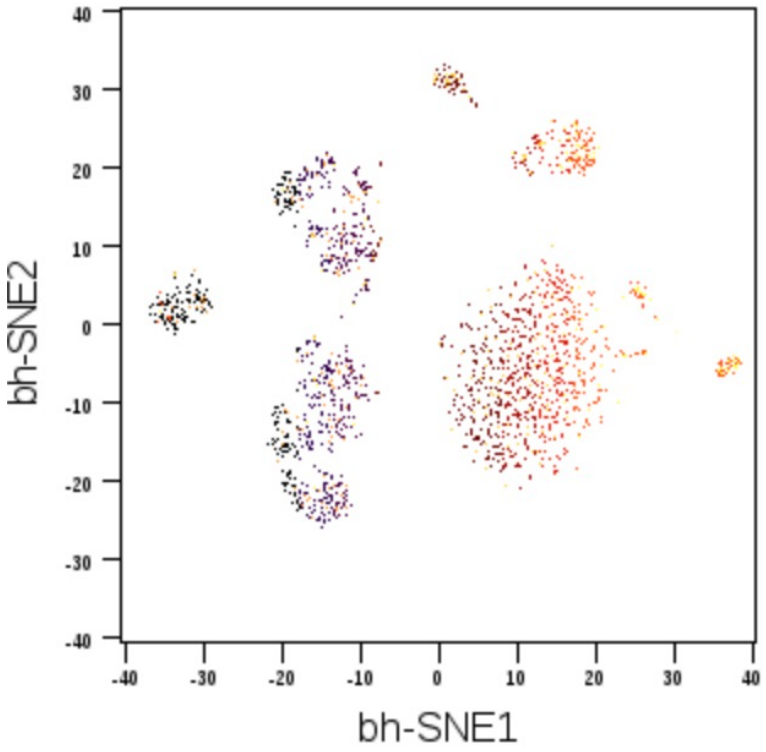


200000 cells

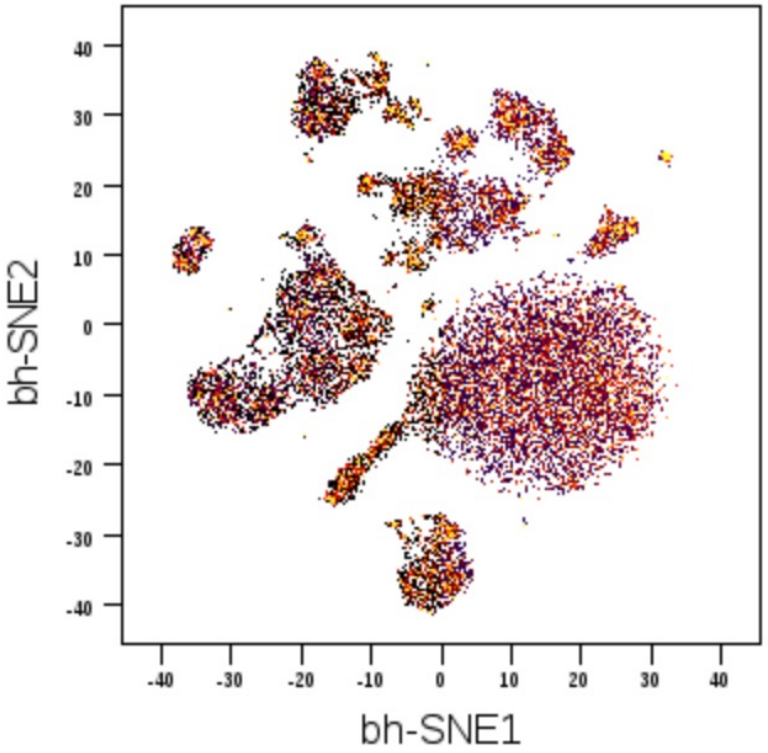


Global structure of t-SNE map doesn't appear to be affected by embedding space compression

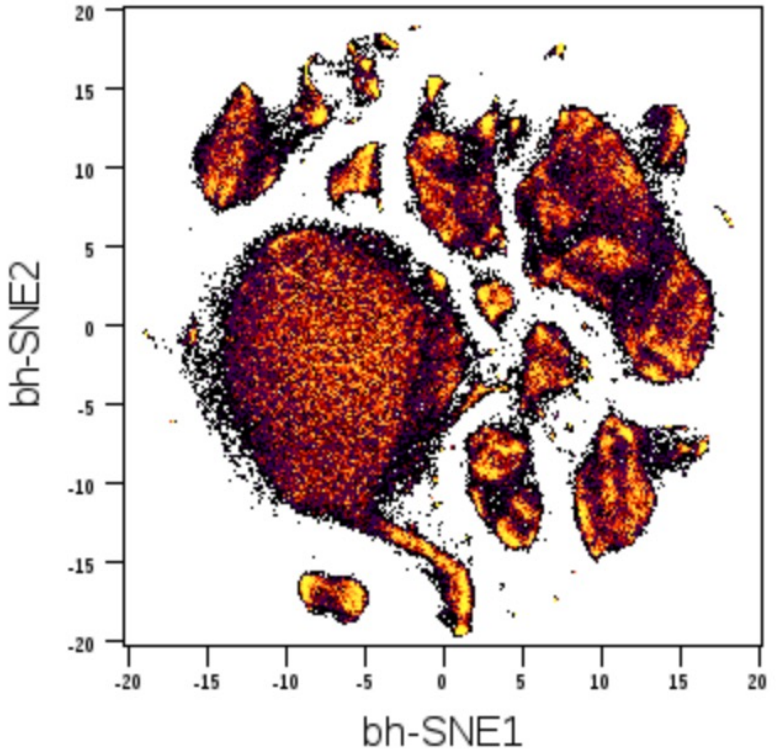
2k cells



20k cells



200k cells



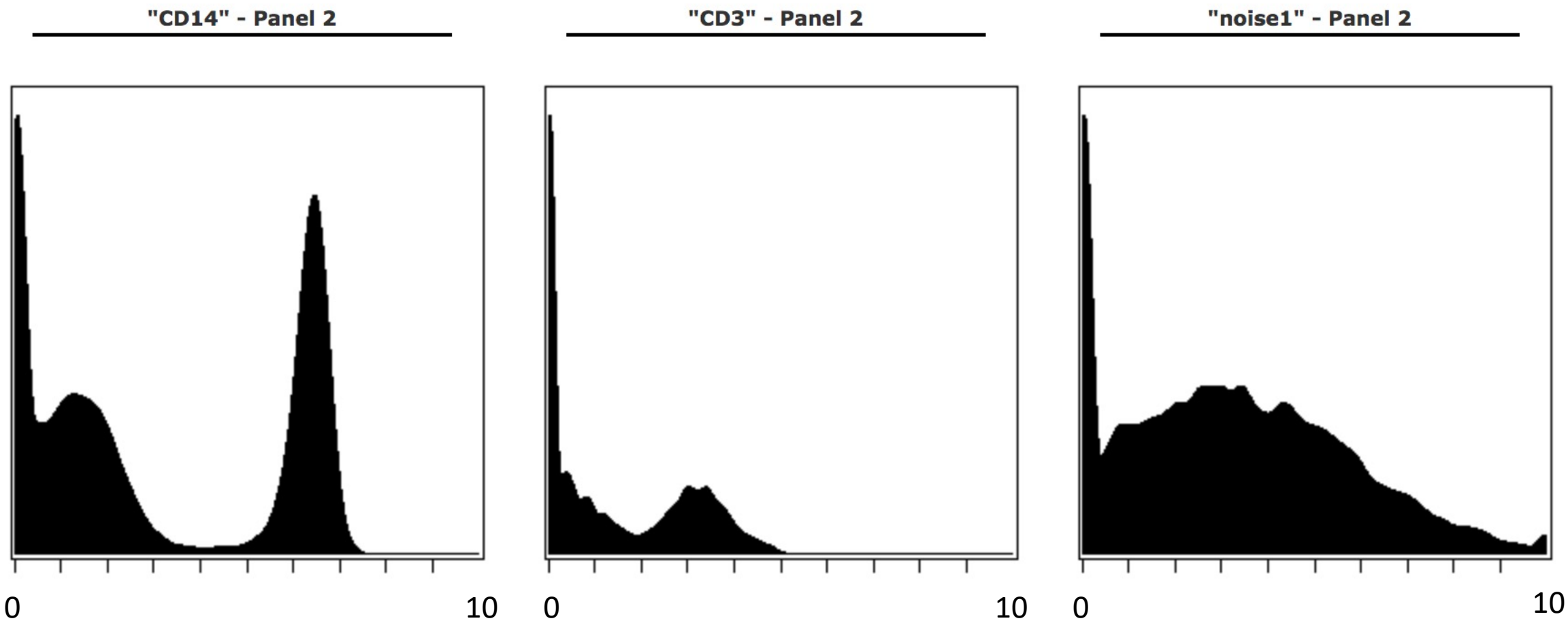
How robust is t-SNE visually to noise?

- Do “bad” or noisy markers mess up the output of the t-SNE map?
- Data: healthy human PBMCs (same as before). Simplified dataset with 6 markers.
- Procedure: Add random unimodal noise channels to the end of the dataset, and visualize the t-SNE output.
- Visualize as biaxial plot colored by Kernel Density Estimation

What adding noise to a dataset looks like

Real dimensions

Noise dimension



The structure of the data with noisy dimensions

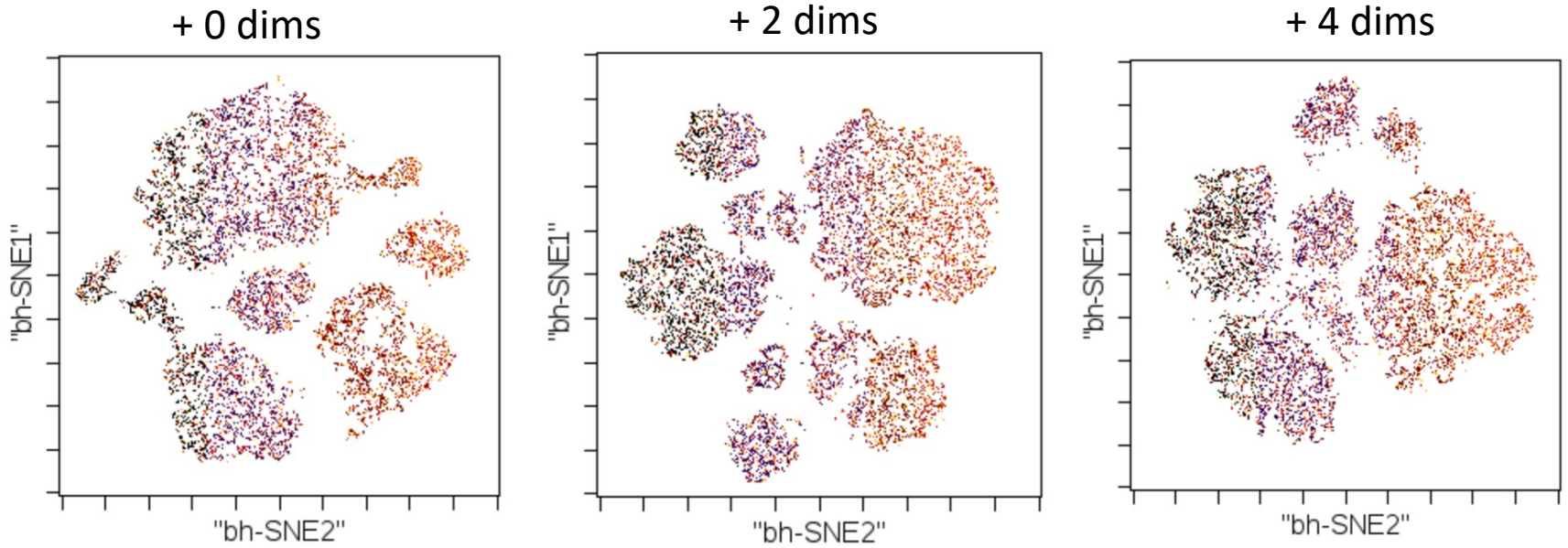
Run t-SNE using ALL OF THIS

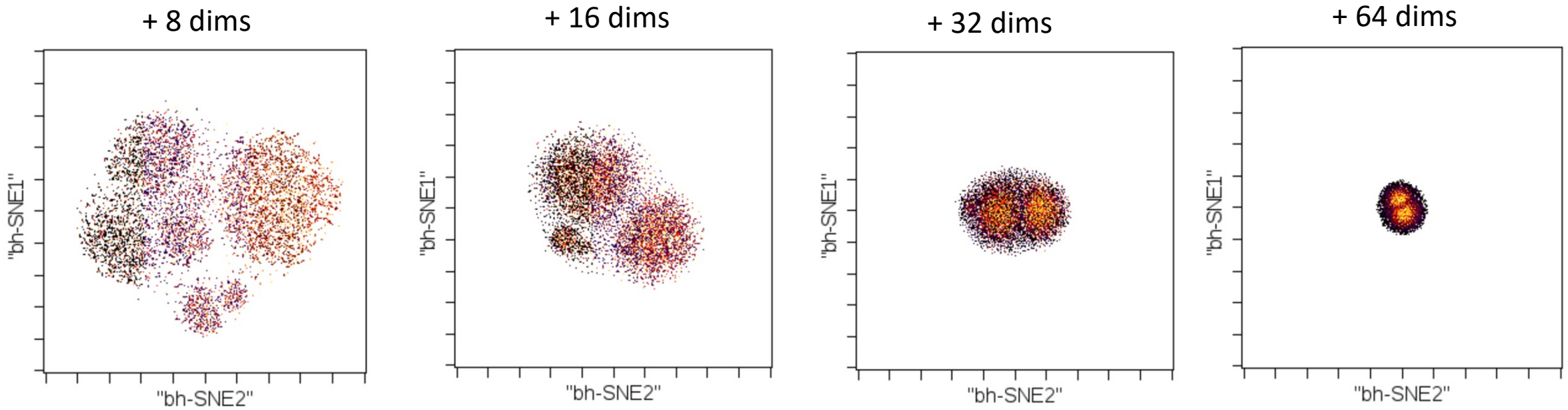
Real features

Noise features

Cells

Adding noise dimensions adversely affects t-SNE





Summary 1

- Adding more cells as input squishes the t-SNE output to the center
- Adding more cells as input maintains the shape of the islands, adds density details

- Adding noise dimensions adversely affects the topology of the t-SNE map. **So choose your panels carefully**.

The organization of my talk

- Part 1: Show how varying input affects t-SNE output (so you don’t have to)

- **Part 2: Determine whether we can and/or should gate and cluster a t-SNE map**

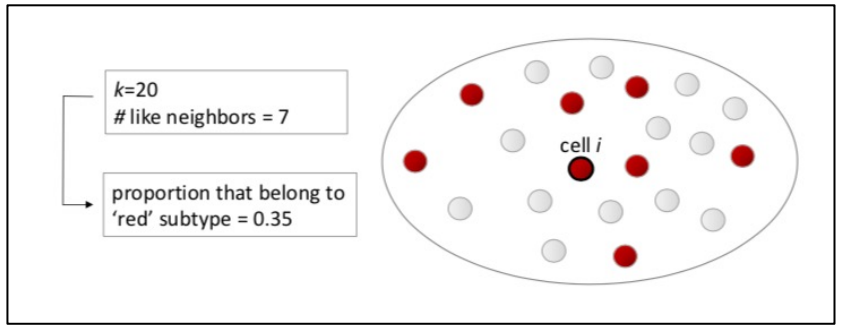
Low-dimension fidelity has been only recently addressed for single cell data

Comparative Analysis of Linear and Nonlinear Dimension Reduction Techniques on Mass Cytometry Data

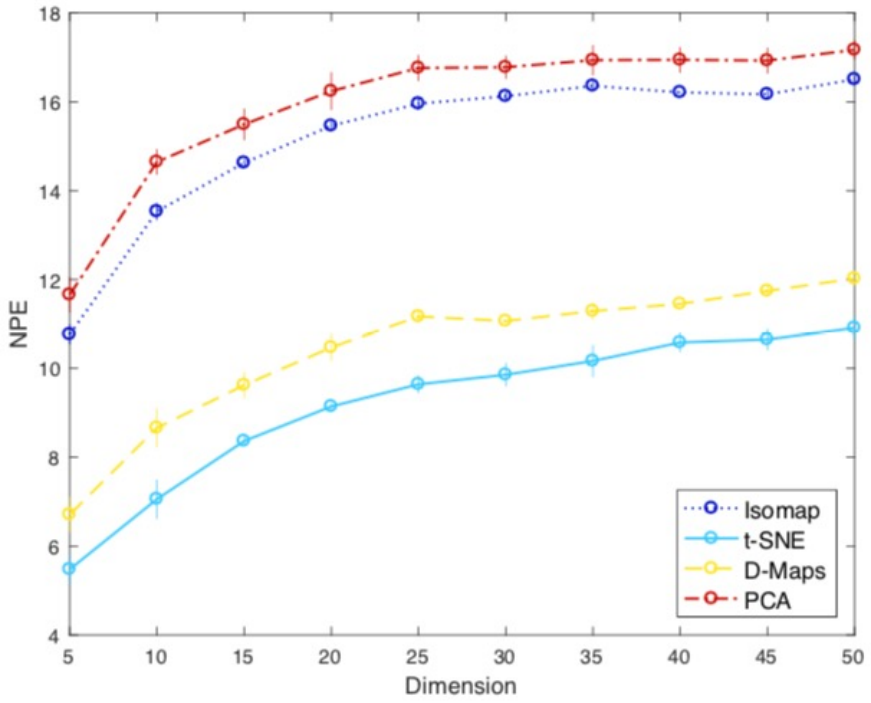
Anna Konstorum*[1], Nathan Jekel[2], Emily Vidal[3] and Reinhard Laubenbacher[1,4]

[1] *Center for Quantitative Medicine, UConn Health, Farmington, CT*
[2] *Department of Mathematics, Indiana University East, Richmond, IN*
[3] *Department of Mathematics, Angelo State University, San Angelo, TX*
[1,4] *Jackson Laboratory for Genomic Medicine, Farmington, CT*

This is based on manual gating, like the F1 Score for Clustering



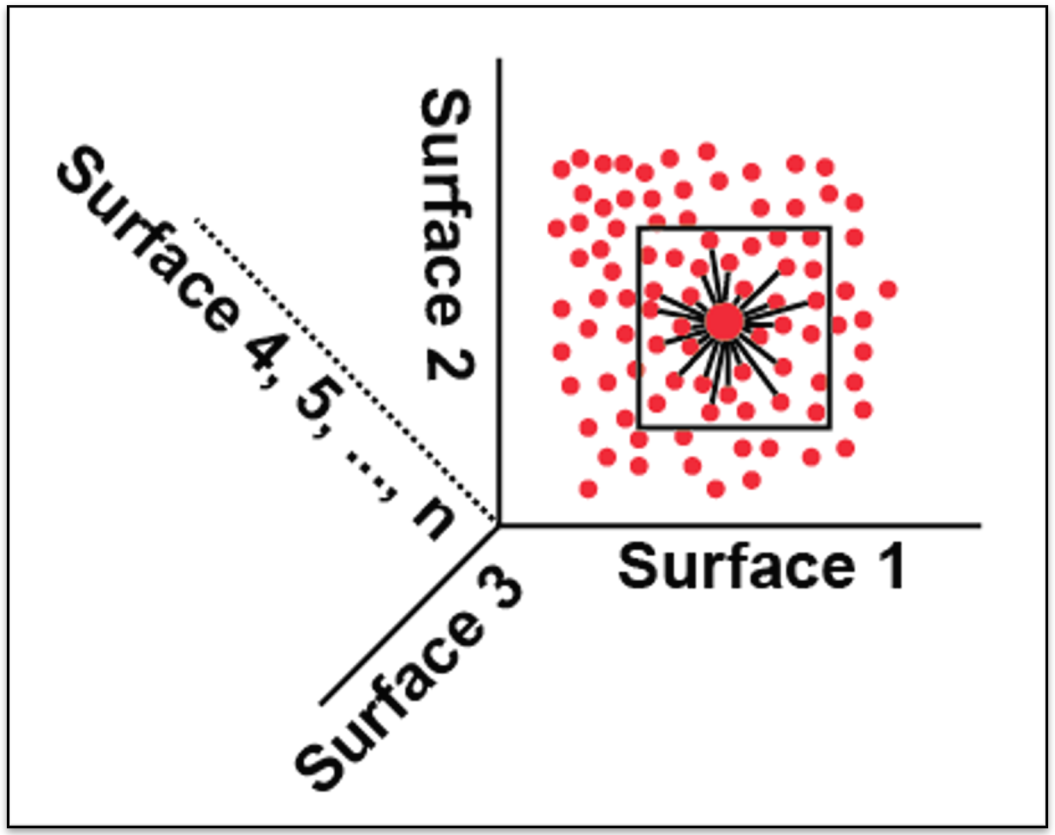
3.2 Neighborhood Proportion Error

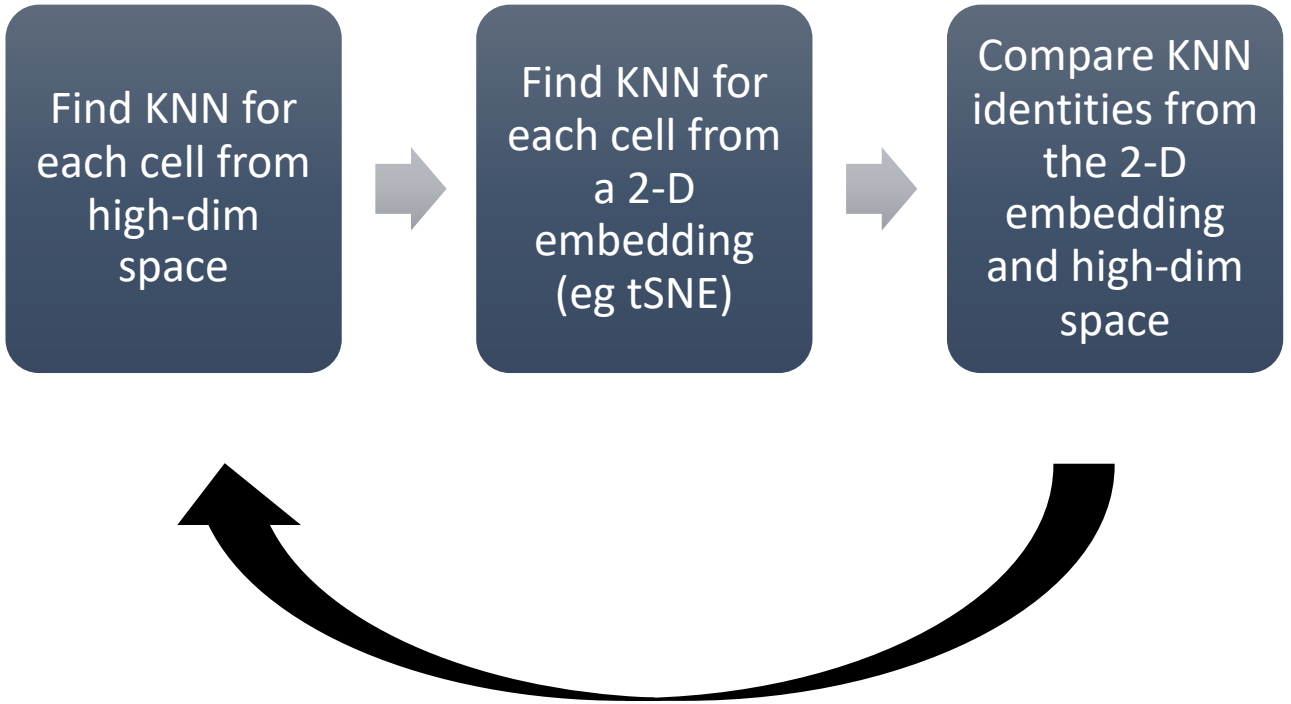


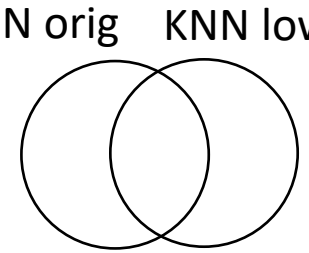
What is still needed from Low-D fidelity analysis

- Manual gating-free fidelity measure
- A way to assess LOCAL fidelity rather than global fidelity
- A deep-dive into a single algorithm rather than a high-level overview of multiple algorithms
- A software pipeline (eg. R package) that can incorporate new algorithms as they come out AFTER the paper is out.

KNN without manual gating to determine fidelity of lower dimensional embeddings







Bioconductor package: Sconify

Software for your KNN-based CyTOF needs

Sconify

platforms all | downloads available | posts 0 | in Bioc < 6 months | build ok

DOI: 10.18129/B9.bioc.Sconify

This is the **development** version of Sconify; for the stable release version, see Sconify.

A toolkit for performing KNN-based statistics for flow and mass cytometry data

Bioconductor version: Development (3.8)

This package does k-nearest neighbor based statistics and visualizations with flow and mass cytometery data. This gives tSNE maps"fold change" functionality and provides a data quality metric by assessing manifold overlap between fcs files expected to be the same. Other applications using this package include imputation, marker redundancy, and testing the relative information loss of lower dimension embeddings compared to the original manifold.

Author: Tyler J Burns

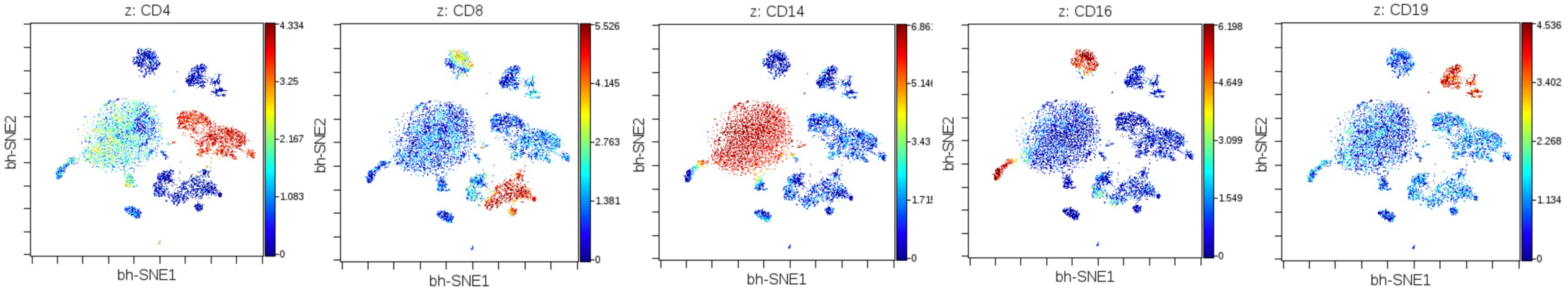
Maintainer: Tyler J Burns <burns.tyler at gmail.com>

Citation (from within R, enter `citation("Sconify")`):

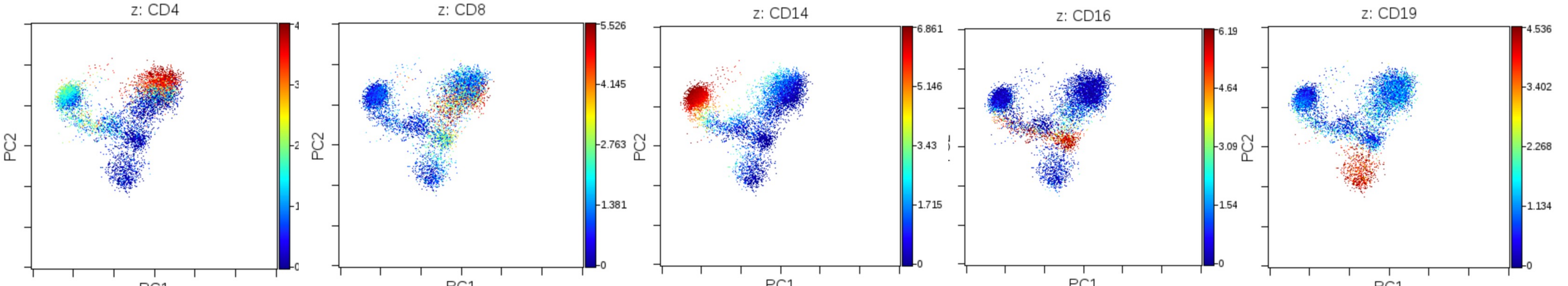
> Burns TJ (2018). *Sconify: A toolkit for performing KNN-based statistics for flow and mass cytometry data*. R package version 1.1.0.

A quick review: Principal Components Analysis (PCA) vs t-SNE

t-SNE (how most people do dim reduction for CyTOF)



PCA (the old or first-pass way of dim reduction)



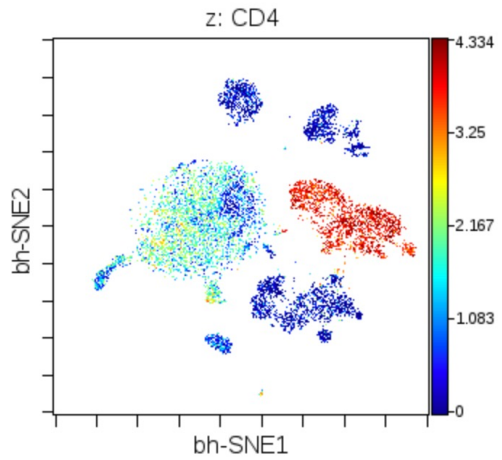
t-SNE preserves local structure at the expense of global structure

vs (KNN)

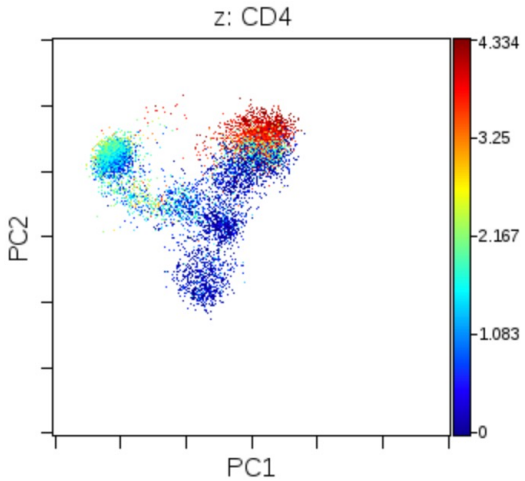
High-dimensional space

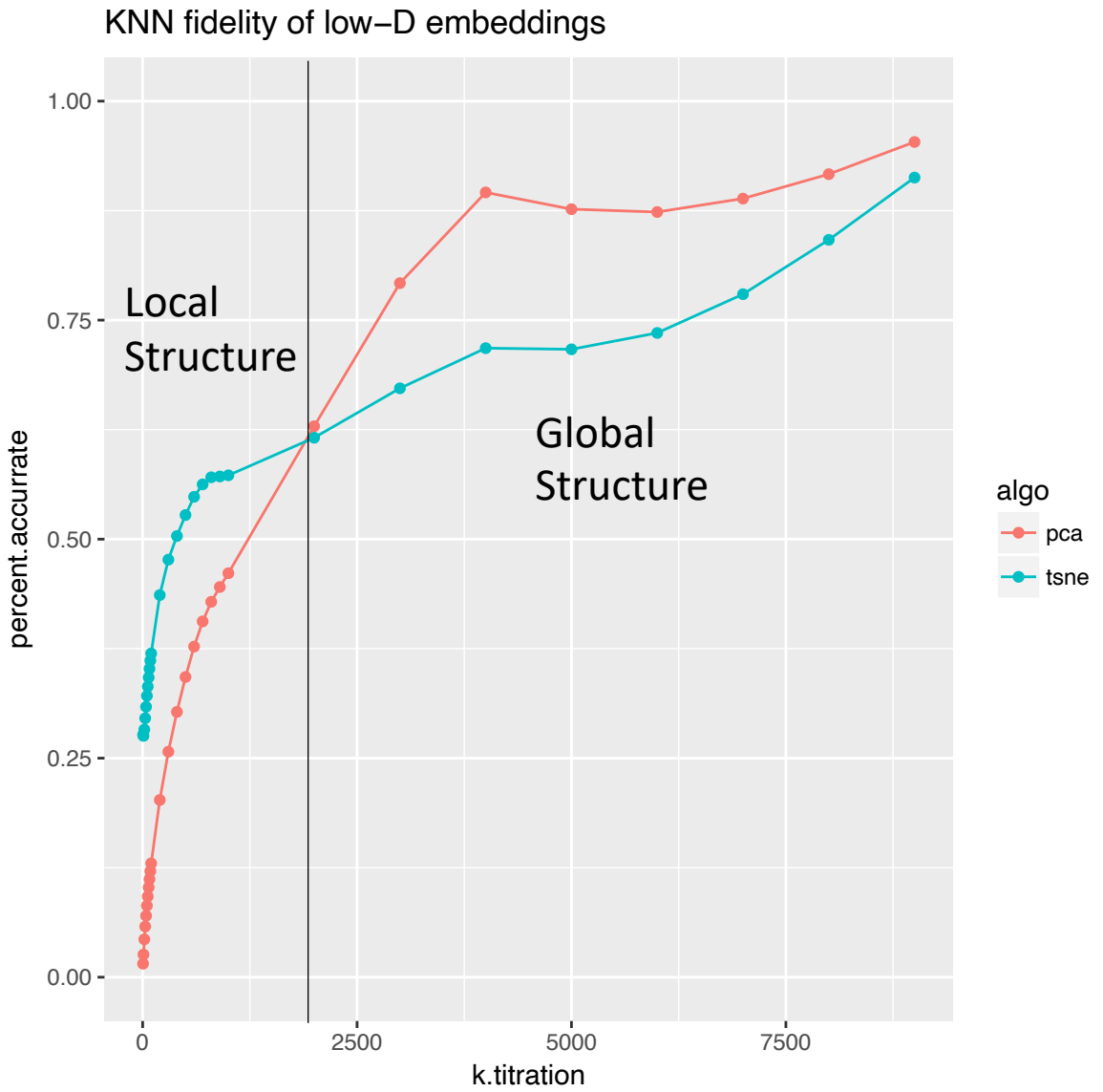
vs (KNN)

t-SNE



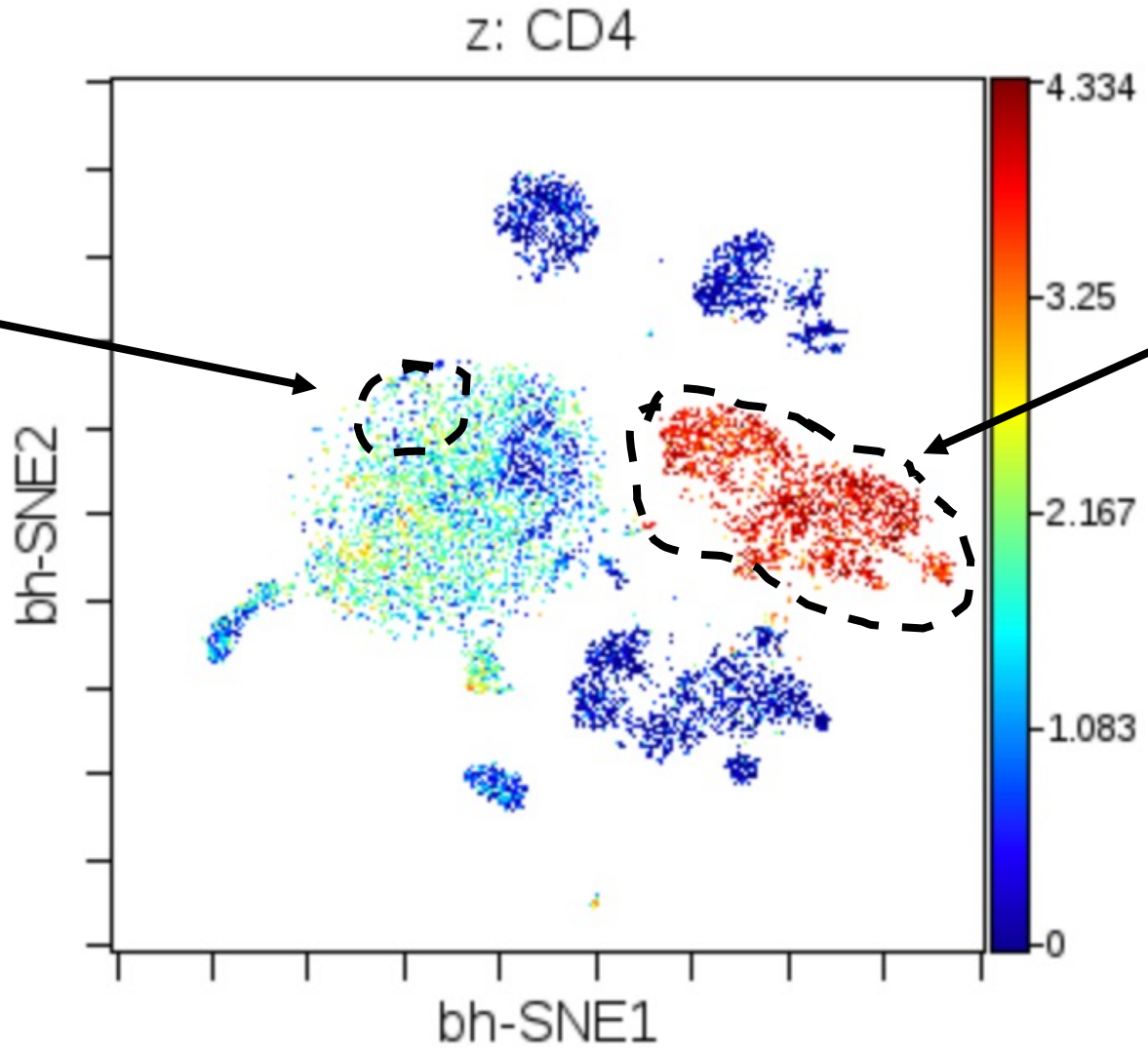
PCA





Does t-SNE preserve some regions better than others (should we gate the map?)

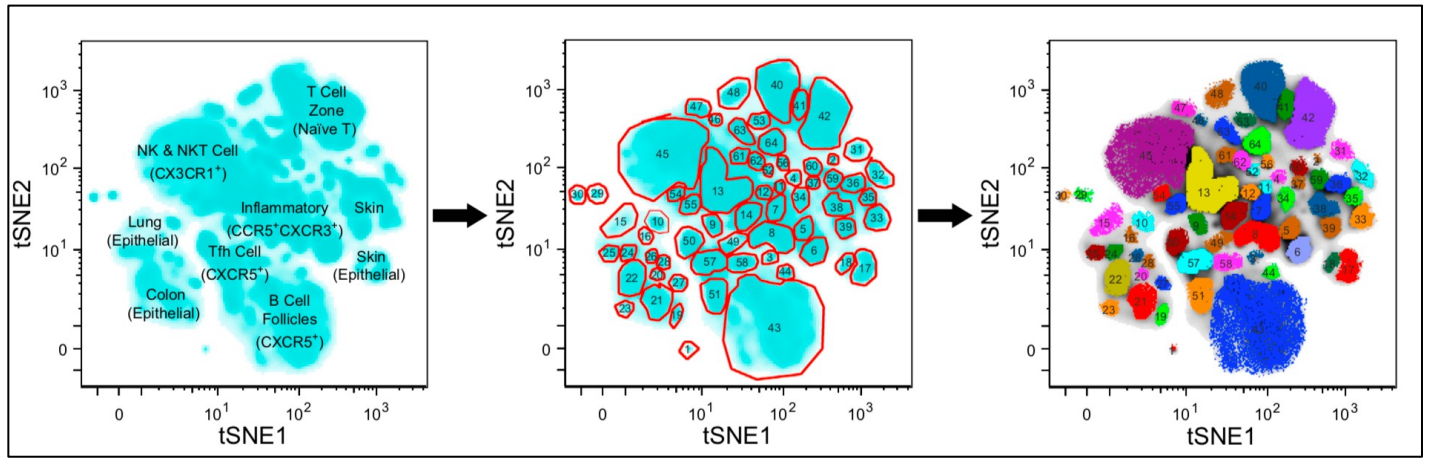
Is this a valid gate?



Is this a valid gate?

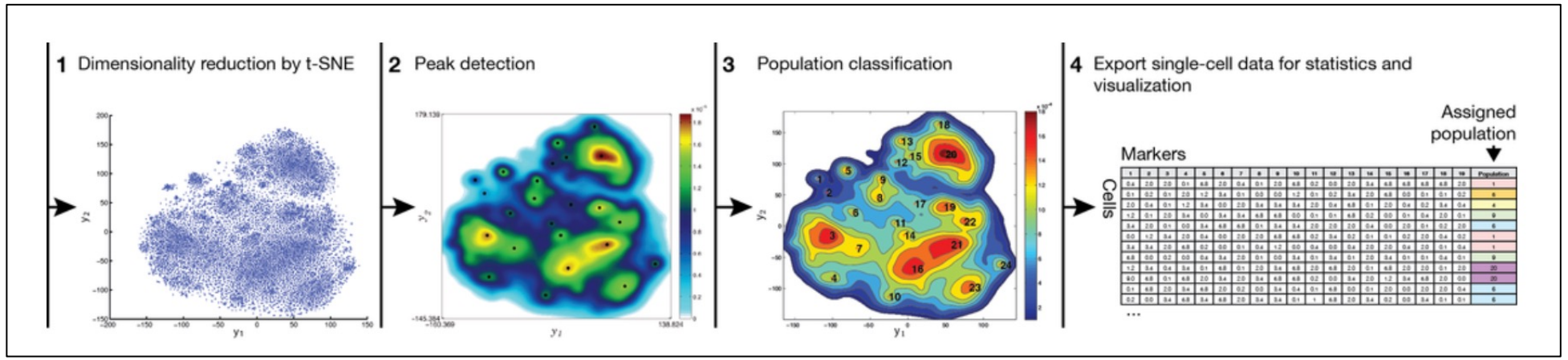
People are already gating and clustering t-SNE maps! Is this ok??

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



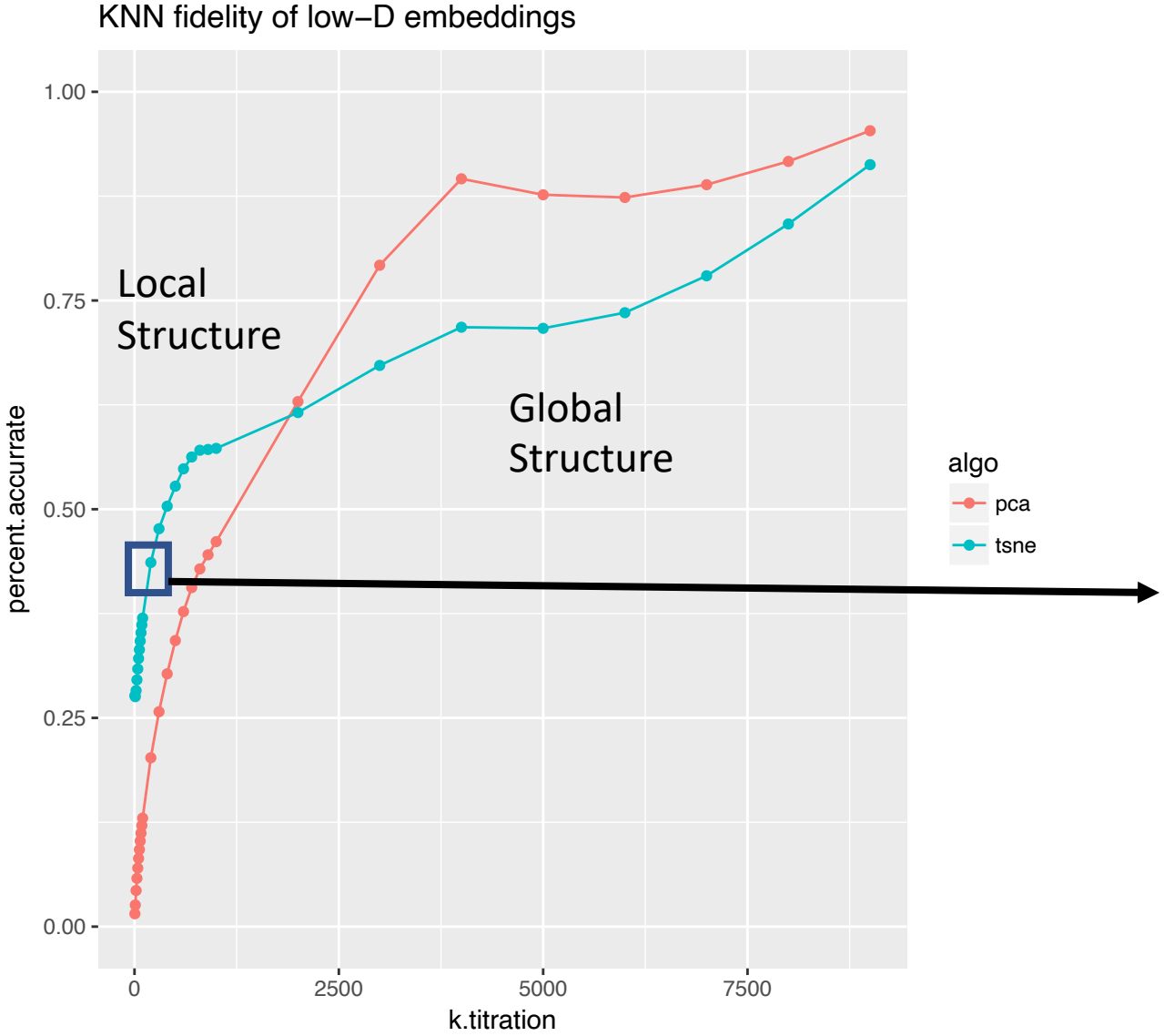
Wong *et al*, *Cell* 2016

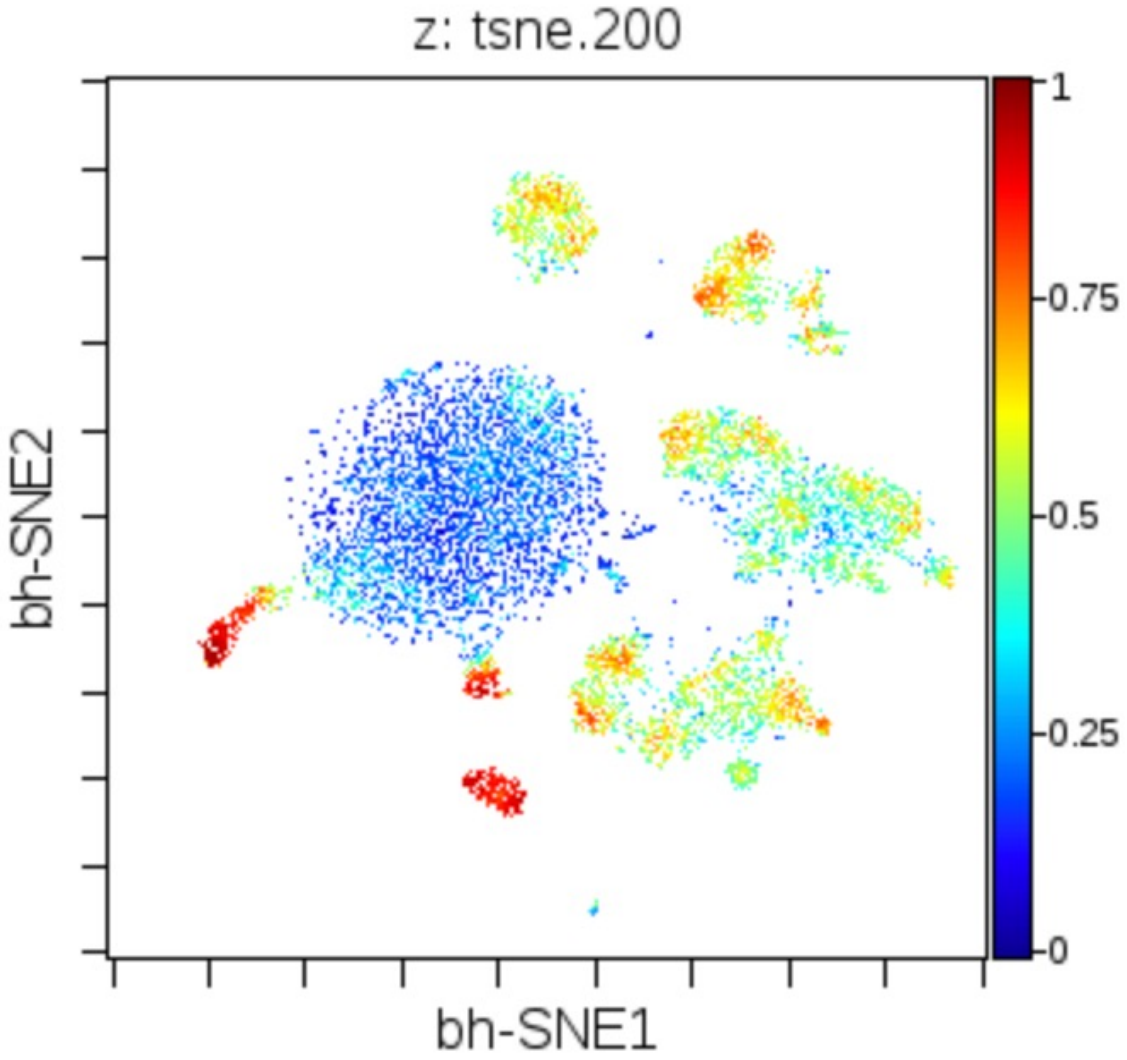
Accense (Petter Brodin): Clustering a t-SNE map



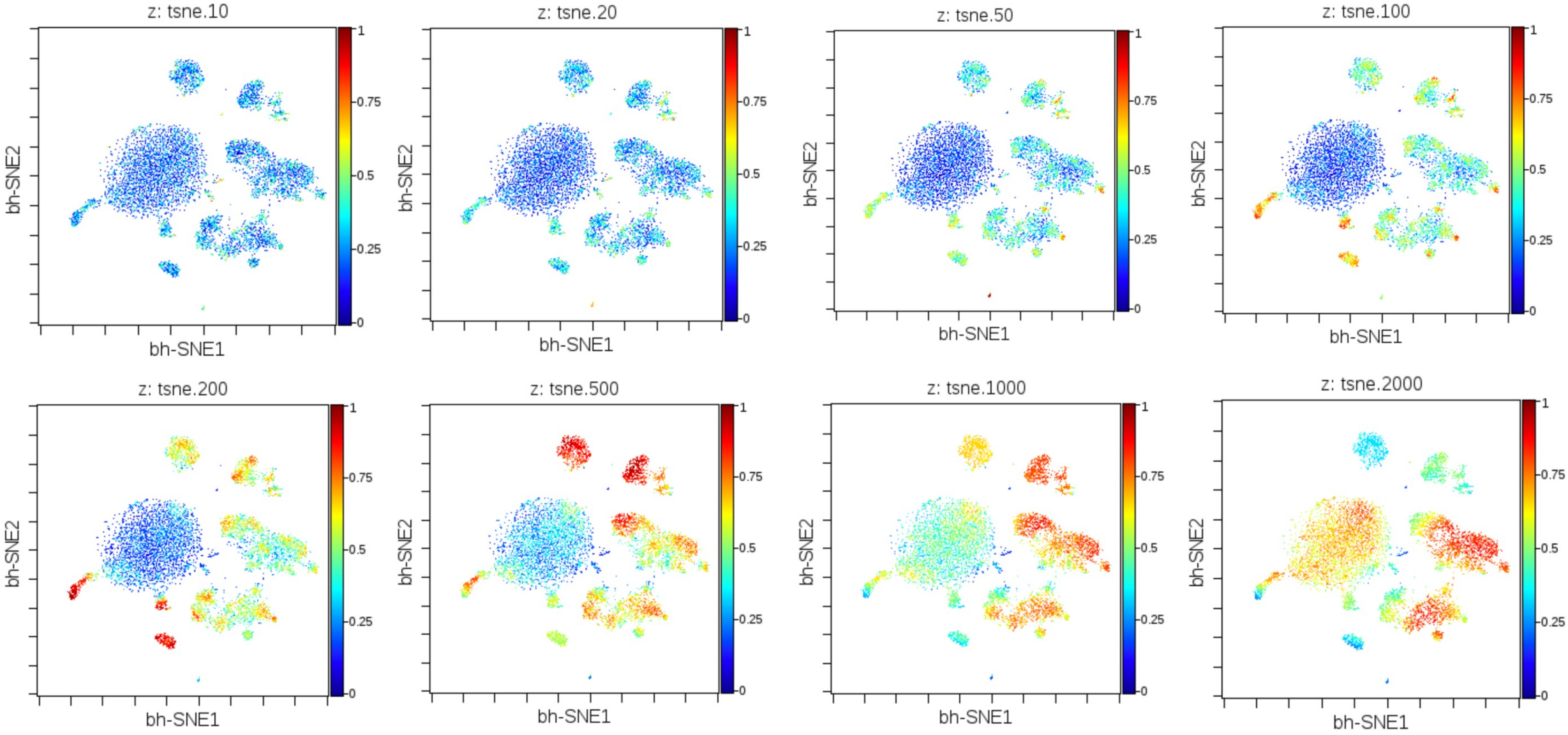
Shekar *et al*, *PNAS* 2014

Method: color t-SNE map by KNN fidelity for a given set of values K



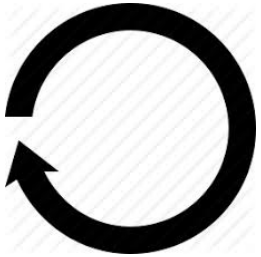


Local t-SNE fidelity sets guidelines for t-SNE clustering and gating

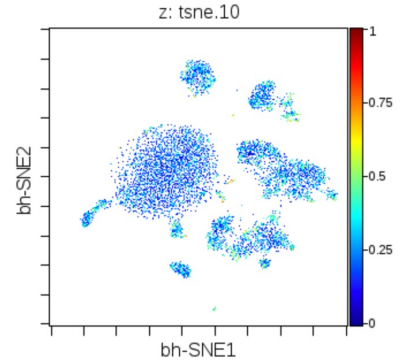


How consistent is one t-SNE run from another?

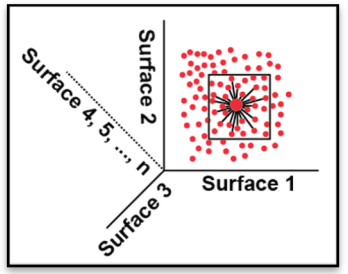
- Run t-SNE many times



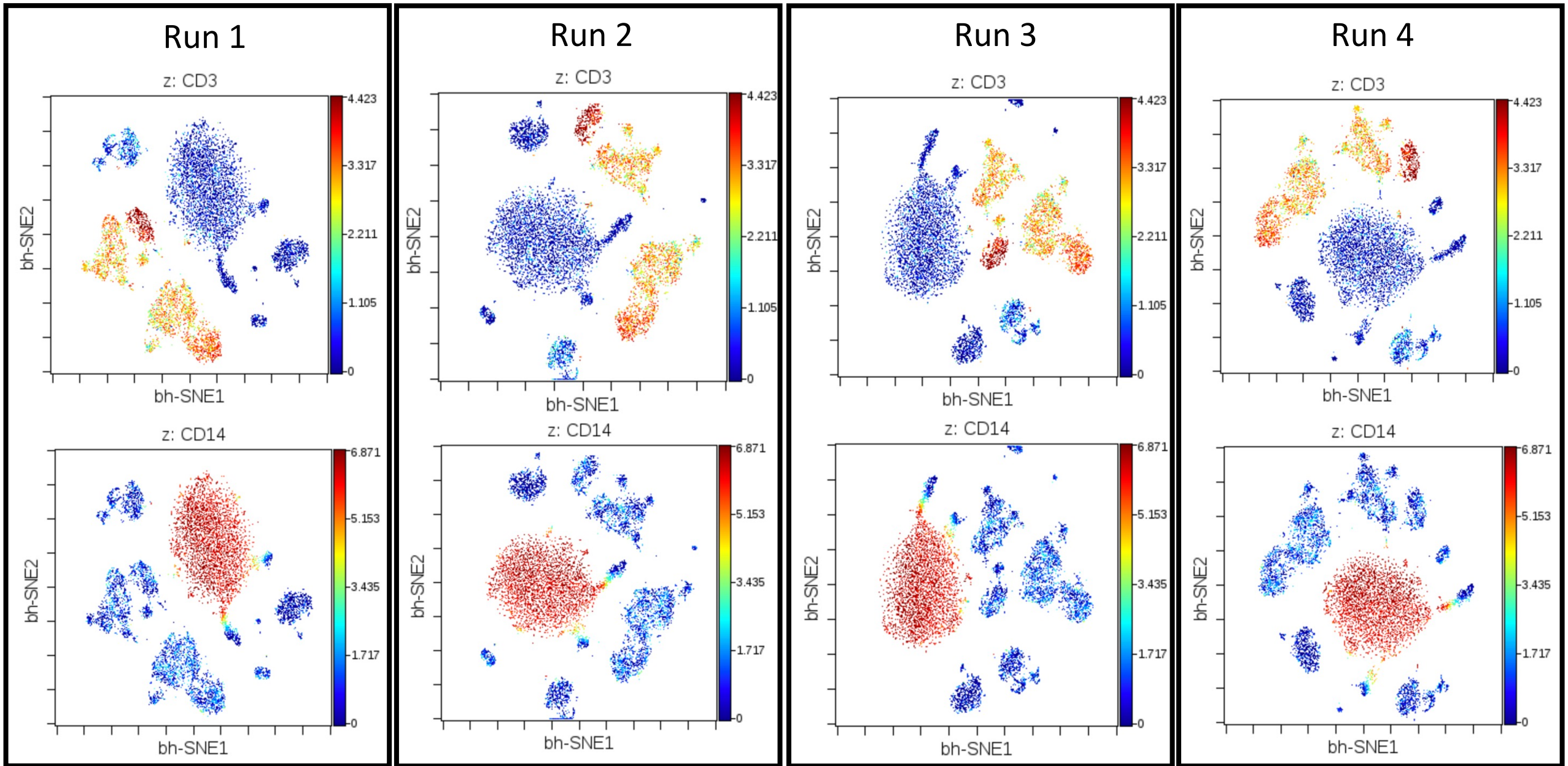
- Determine visual similarity of t-SNE maps



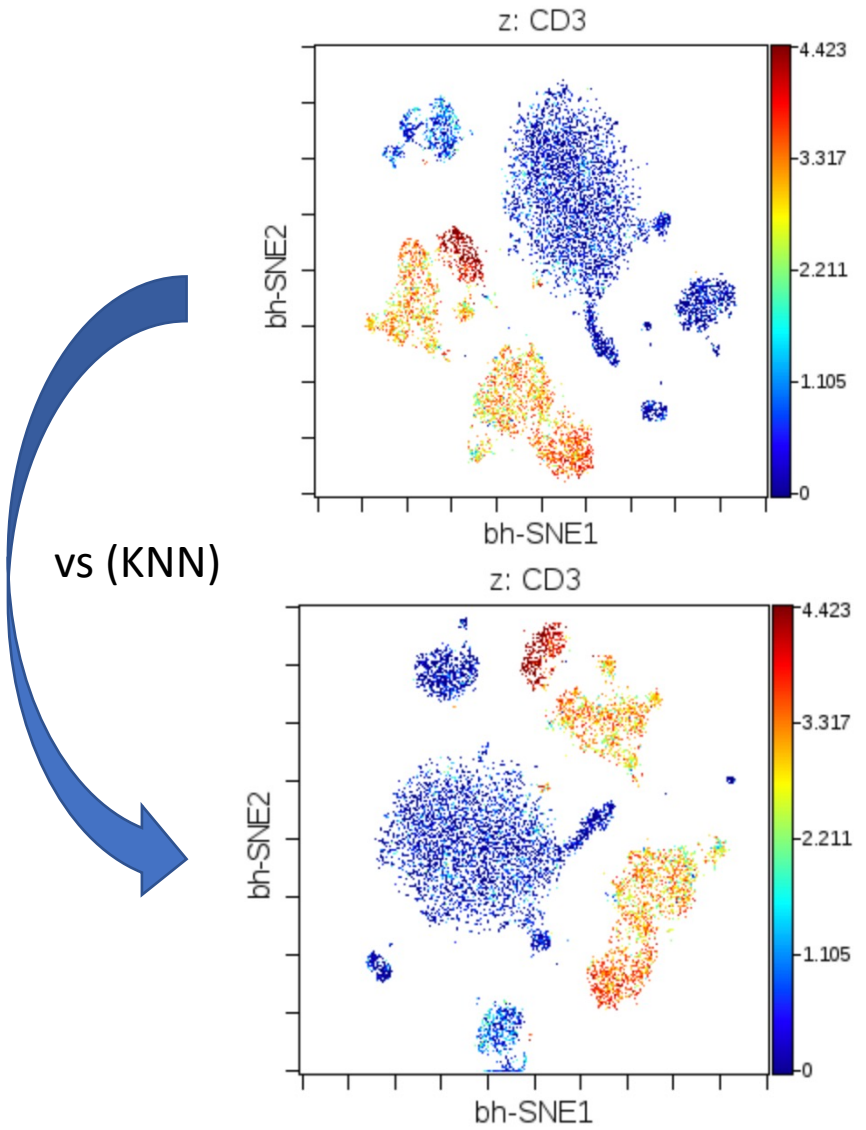
- Determine global KNN similarity of t-SNE maps
- Determine local KNN similarity of t-SNE maps

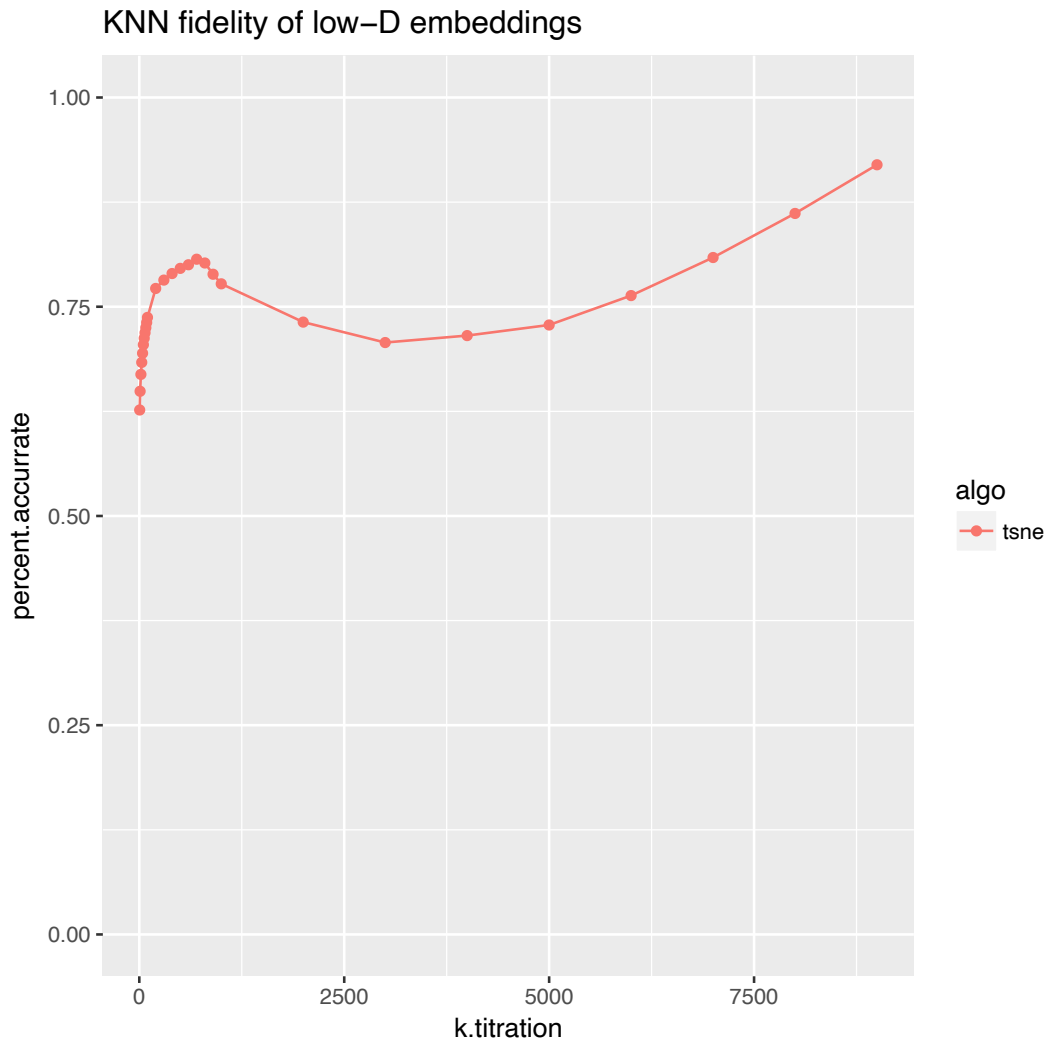


No two t-SNE maps are the same

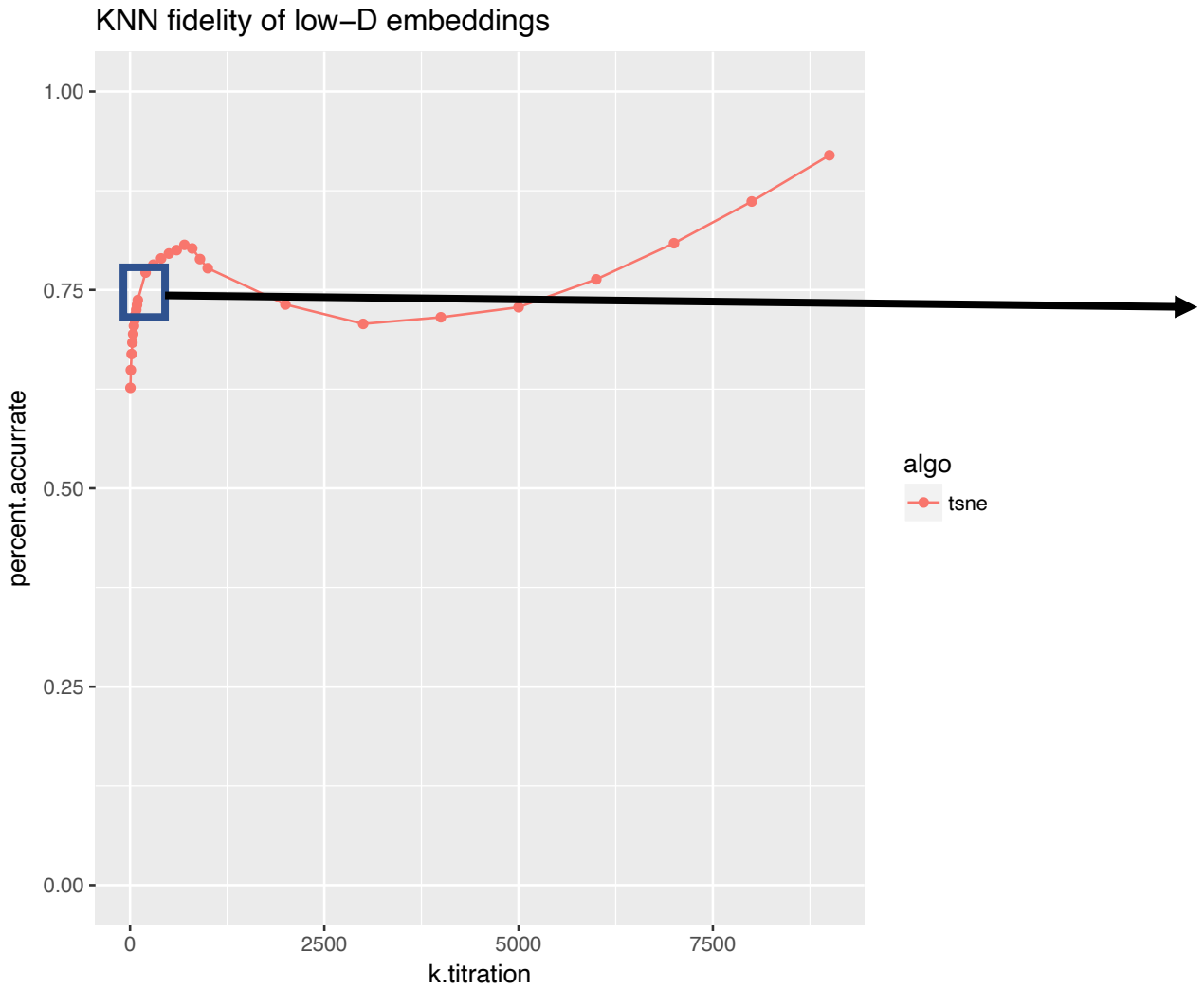


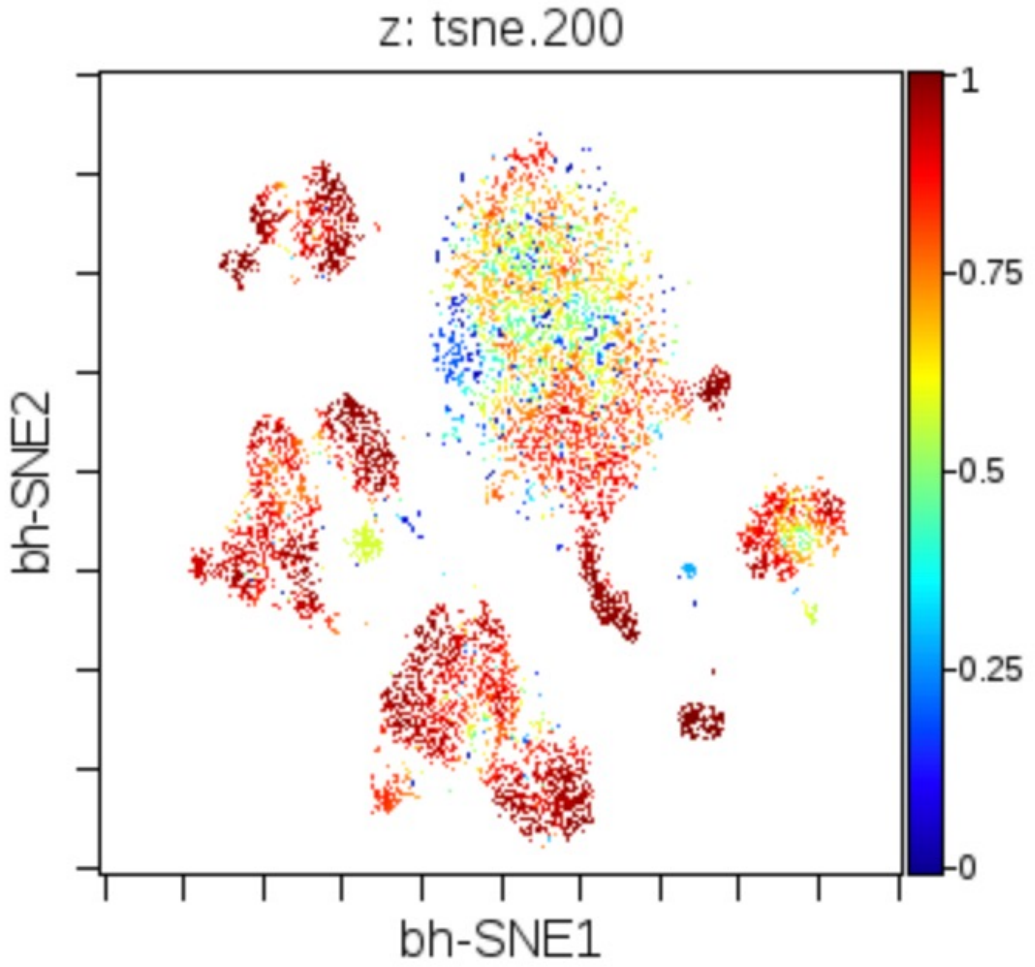
How consistent is one t-SNE run from another: KNN inspection



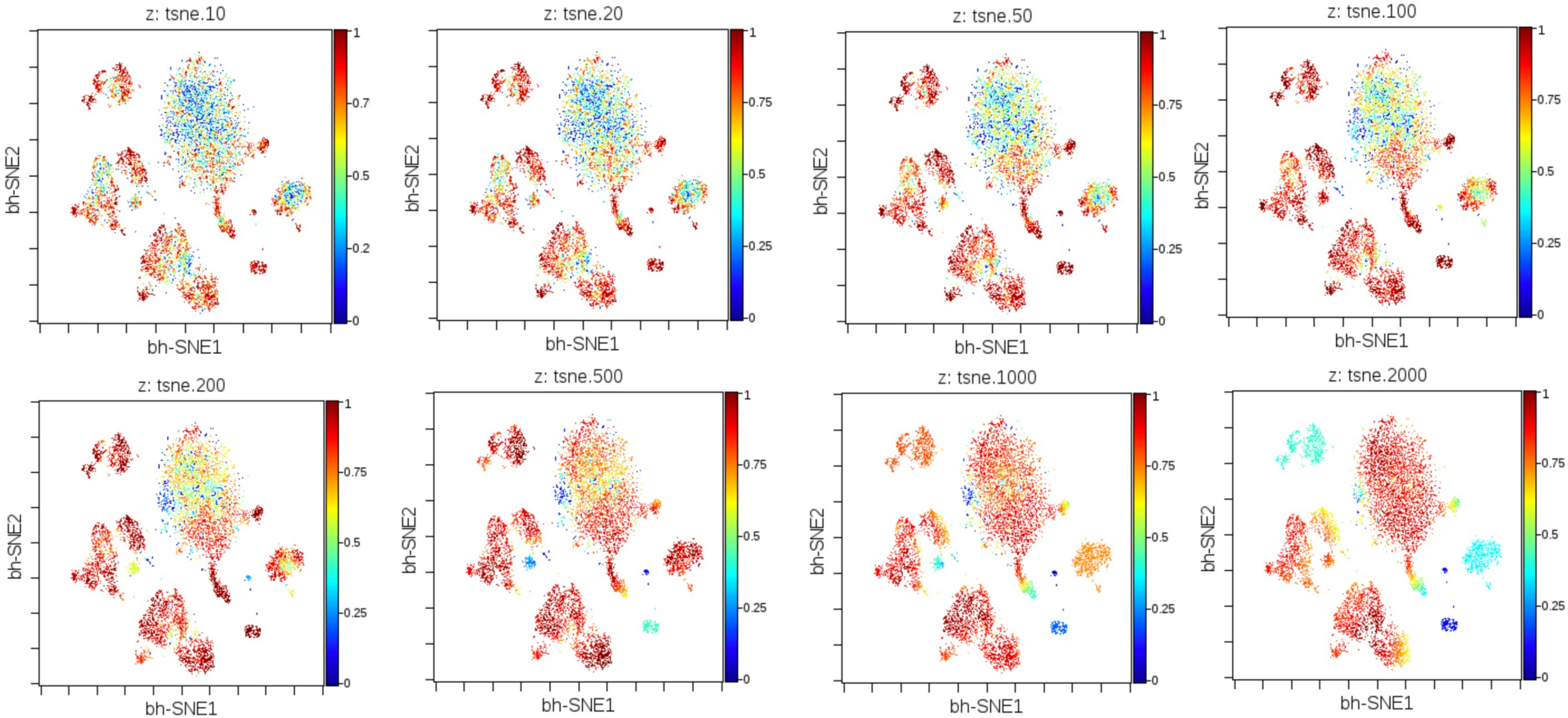


Method: color t-SNE map by KNN fidelity for a given set of values K





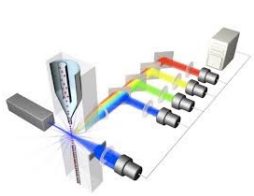
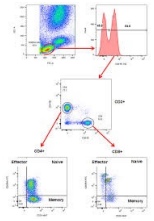
How consistent is one t-SNE run from another: KNN inspection

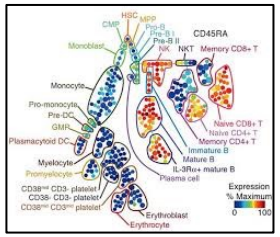


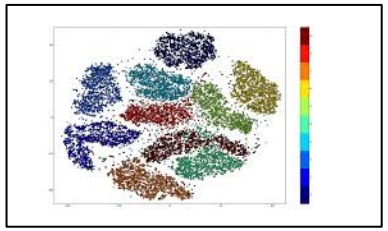
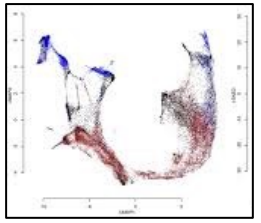
Summary 2

- t-SNE fidelity can be probed using KNN across a wide range of sizes
- t-SNE preserves local structure at the expense of global structure, with local
- t-SNE preserves particular regions more rigorously than others, and this can be used to guide any t-SNE based gating or clustering strategy
- t-SNE preserves local structure with roughly 60-80% consistency, while global island positions are jumbled across runs

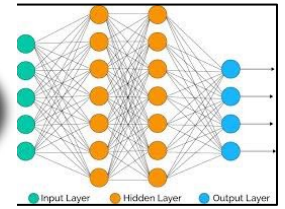
Conclusion: the structure of innovation





Tried and true

Bleeding edge

Yayy!

Acknowledgement

DRFZ BERLIN
Deutsches Rheuma-Forschungszentrum
Ein Institut der Leibniz-Gemeinschaft

Cytodiagnostics, Canada

Ben Pacheco

Miltenyi BioTec

Christian Dose, Susanne Krauthäuser

Henrik Mei Pawel Durek **Axel Schulz** Andreas Grützkau

Silke Stanislawiak **Sabine Baumgart** Marie Urbicht
Christina Schäfer Sarah Gillert
Heike Hirseland **Tyler Burns** Lisa Budzinski Edward Rullmann Julia Schulze

Stanford

Michael Leipold

Holden Maecker, Mark Davis, Garry Fathman

Scailyte

Manfred Claassen, Daniel Sonnleithner

Freie Universität Berlin

Prof. Dr. Susanne Hartmann
Institut für Immunologie
Dr. Svenja Steinfelder
Institut für Immunologie

IMMANUEL DIAKONIE

Prof. Dr. Andreas Krause
Innere Medizin, Rheumatologie und Klinische Immunologie
Prof. Dr. Andreas Michalsen
Naturheilkunde

DFG Deutsche Forschungsgemeinschaft

DGKL Deutsche Vereinte Gesellschaft für Klinische Chemie und Laboratoriumsmedizin e.V.
Nachwuchsakademie 2017

GerMaNet German Mass Cytometry Network

Be The Cure FOR RHEUMATOID ARTHRITIS

SYS INFLAME