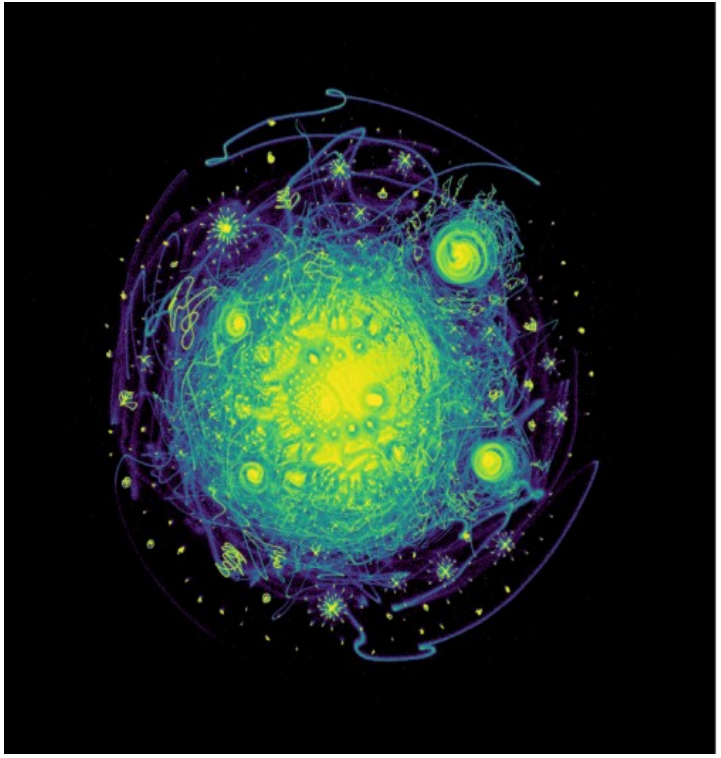
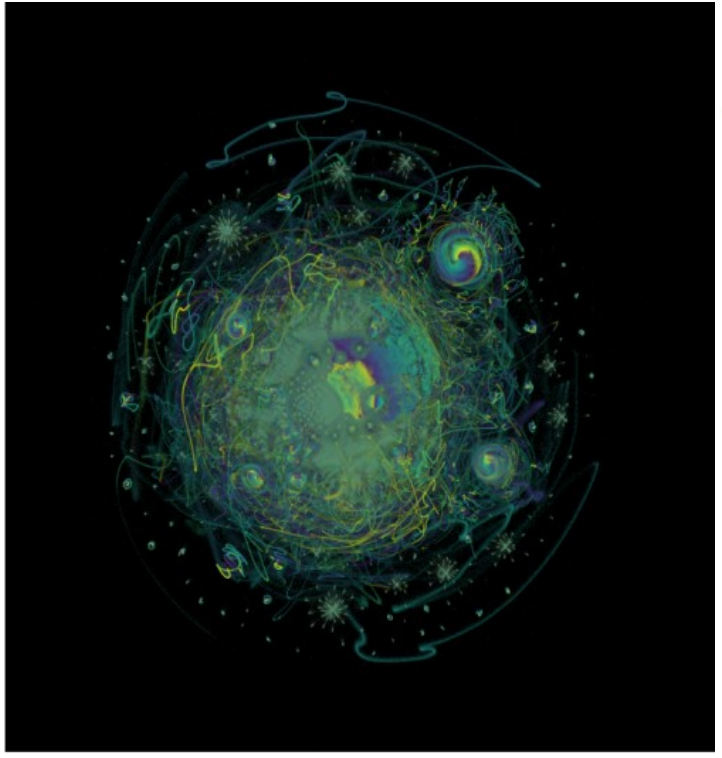
Dimension Reduction and UMAP for Omics Integration

Physalia course, online via zoom

Nikolay Oskolkov, MRG Group Leader, LIOS, Riga, Latvia





@NikolayOskolkov

@oskolkov.bsky.social

Personal homepage:
https://nikolay-oskolkov.com

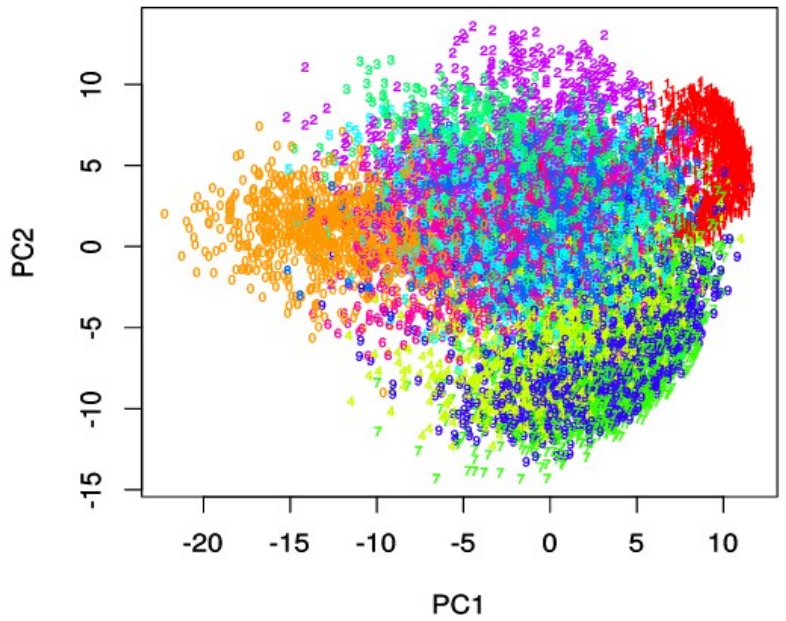
Image adapted from McInnes et al. 2018

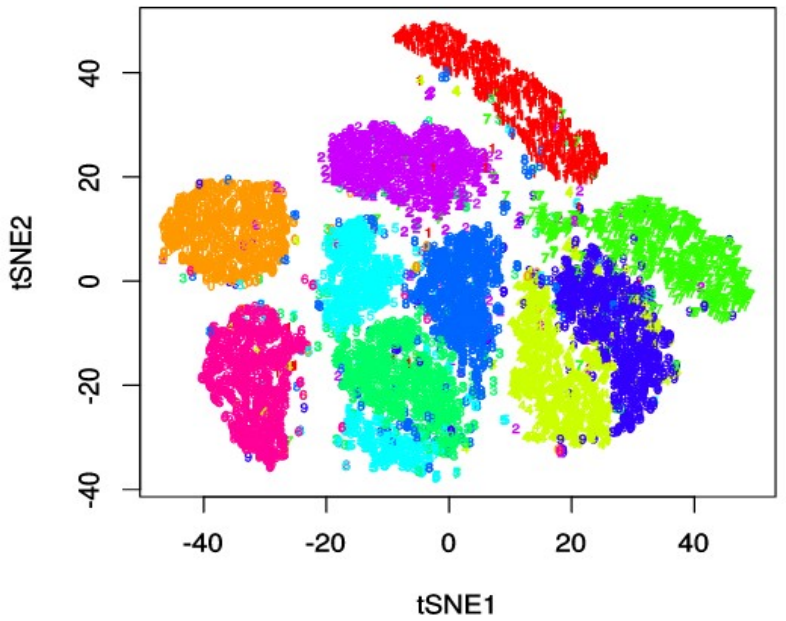
Dimensionality reduction
is also supposed to ... reduce dimensions

Dimension reduction: more than visualization

PCA PLOT WITH PRCOMP

tSNE MNIST

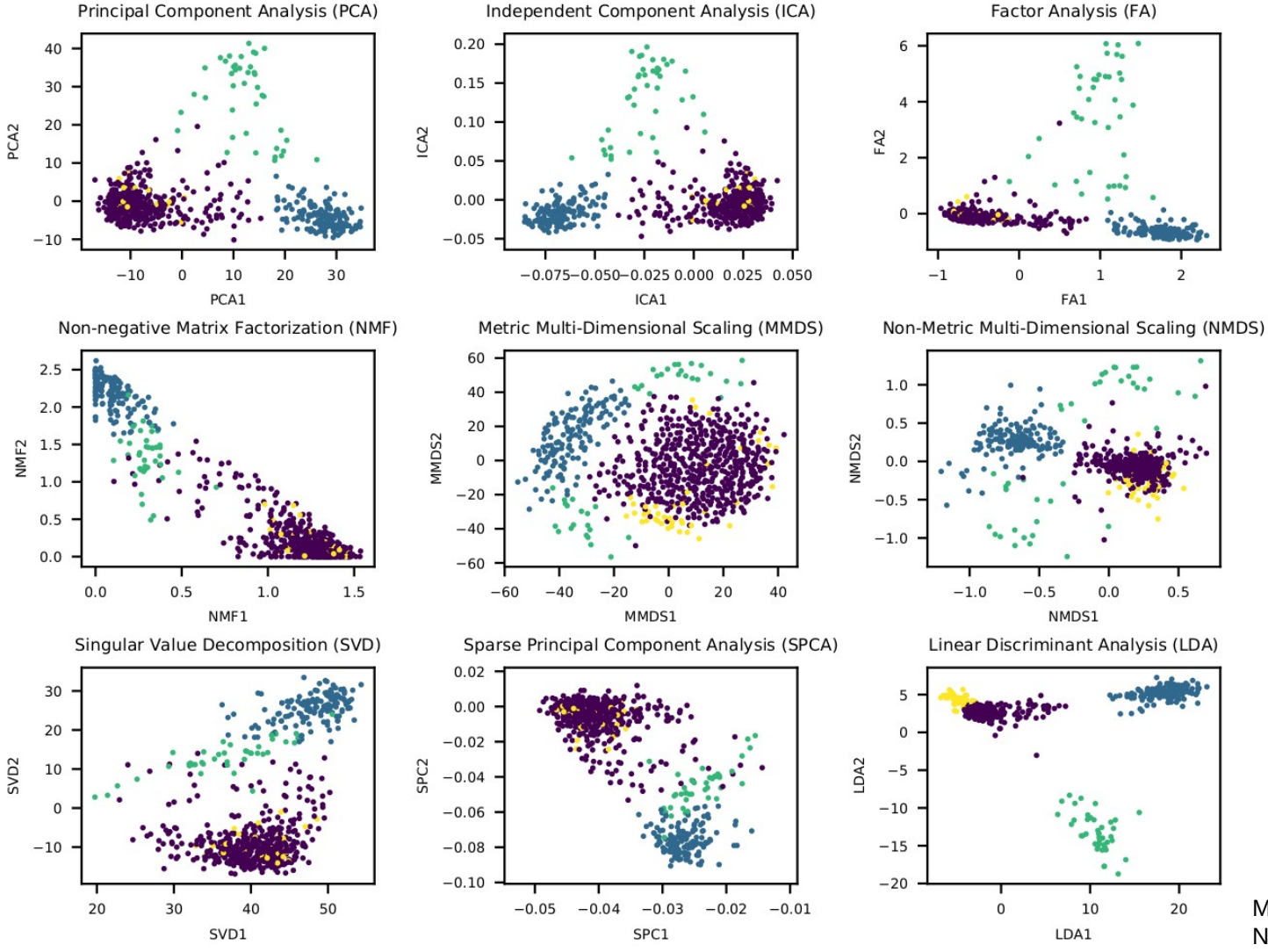




The goal of dimension reduction is not only visualization but also reducing dimensions

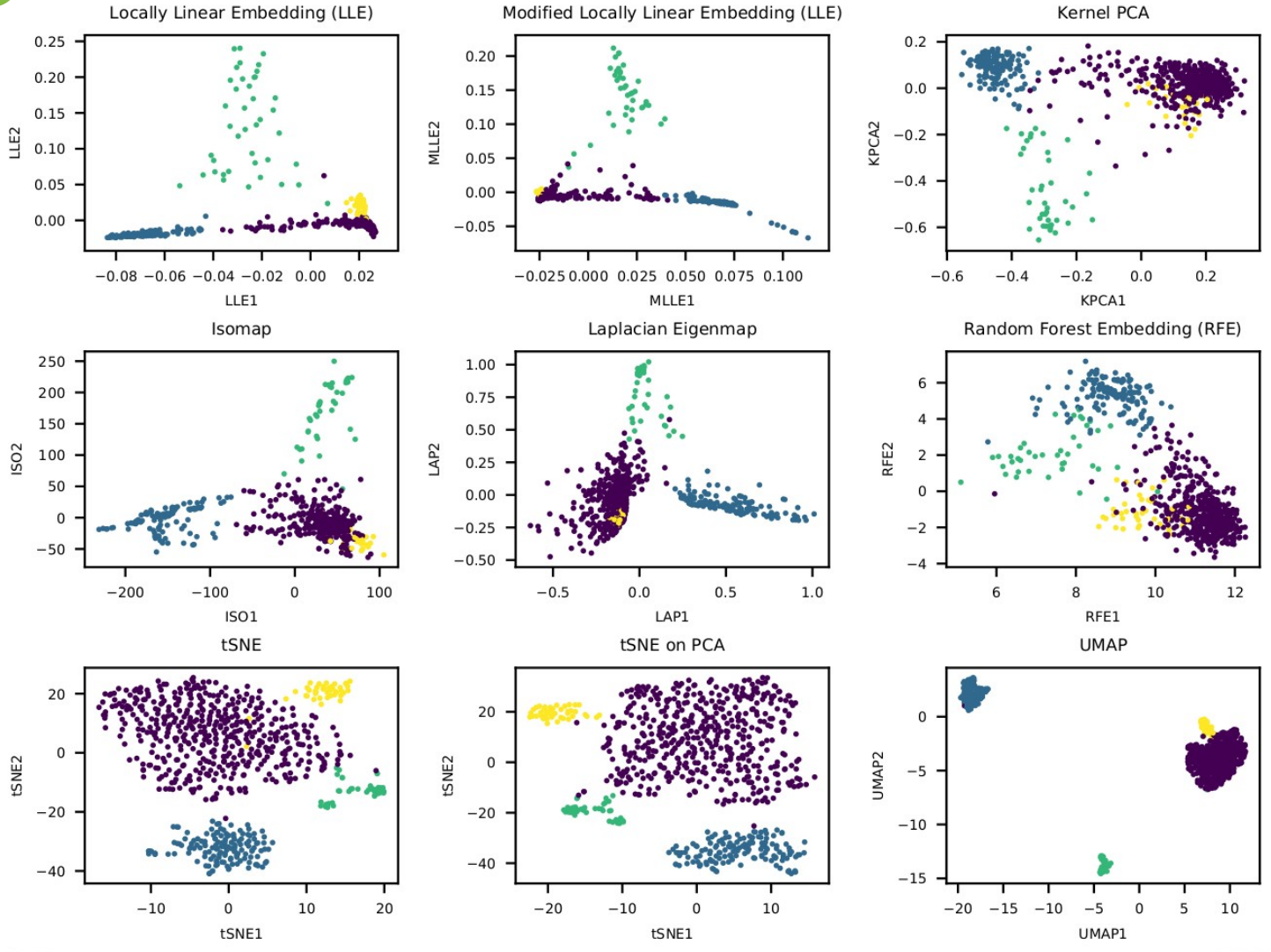
Dimension reduction techniques: linear vs. non-linear

Linear dimensionality reduction



M. Bartoschek, N. Oskolkov et al., Nature Communications 2018

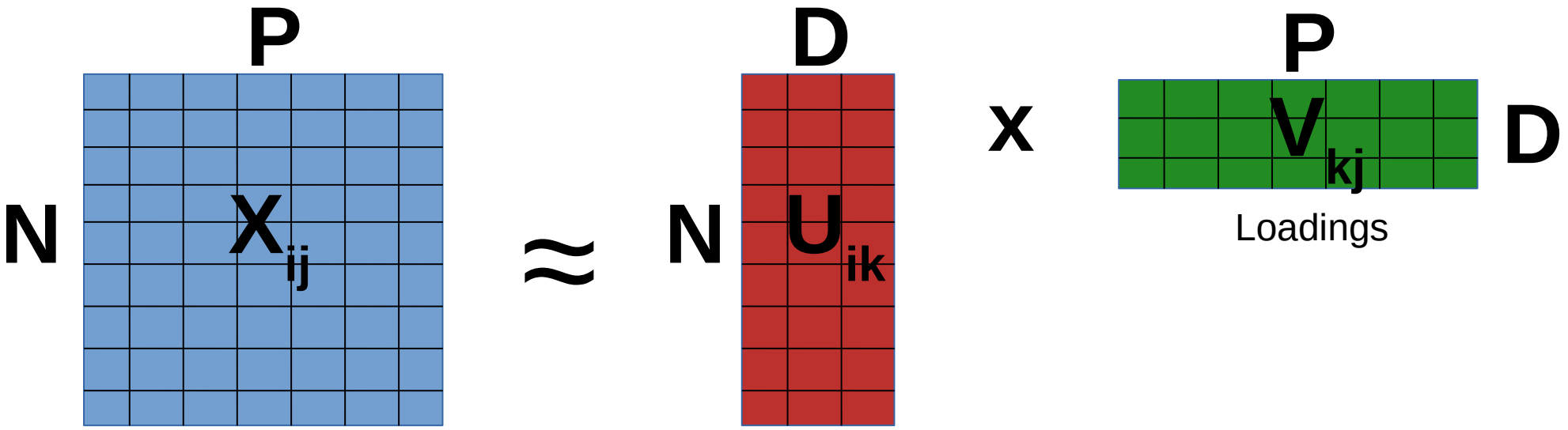
Non-linear dimensionality reduction



M. Bartoschek, N. Oskolkov et al., Nature Communications 2018

Linear dimension reduction: matrix factorization

$$\mathbf{X_{ij}} \approx \mathbf{U_{ik}}\mathbf{V_{kj}}$$



Data

Low-dimensional data representation (embeddings)

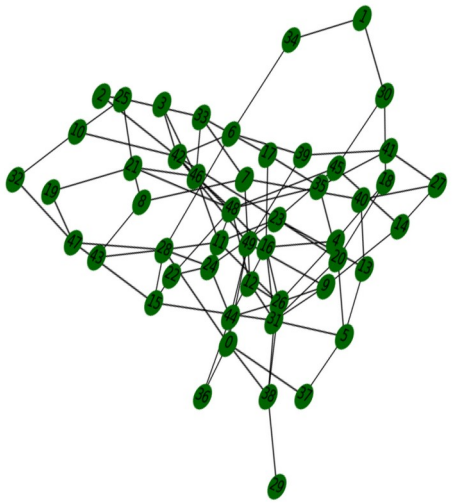
Loadings

$$\text{Loss} = \sum_{i=1}^{N}\sum_{j=1}^{P}(\mathbf{X_{ij}} - \mathbf{U_{ik}}\mathbf{V_{kj}})^2$$

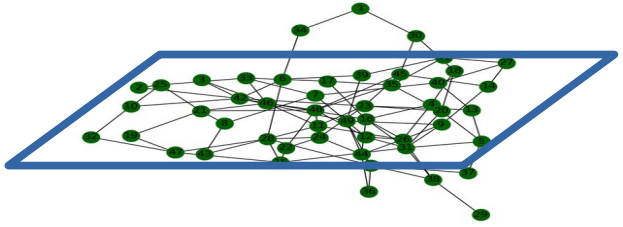
Non-linear dimension reduction: neighborhood graph

1) Construct high-dimensional graph

p_{ij}

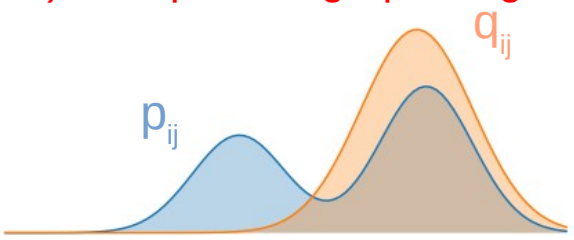


2) Construct low-dimensional graph



q_{ij}

3) Collapse the graphs together



Kullback-Leibler divergence

PCA dimension reduction algorithm

Coding in R:

```
data_centered <- scale(data, center = TRUE, scale = FALSE)

covariance <- t(data_centered) %*% data_centered

eig <- eigen(covariance)

plot(eig$vectors[,1:2]);
```

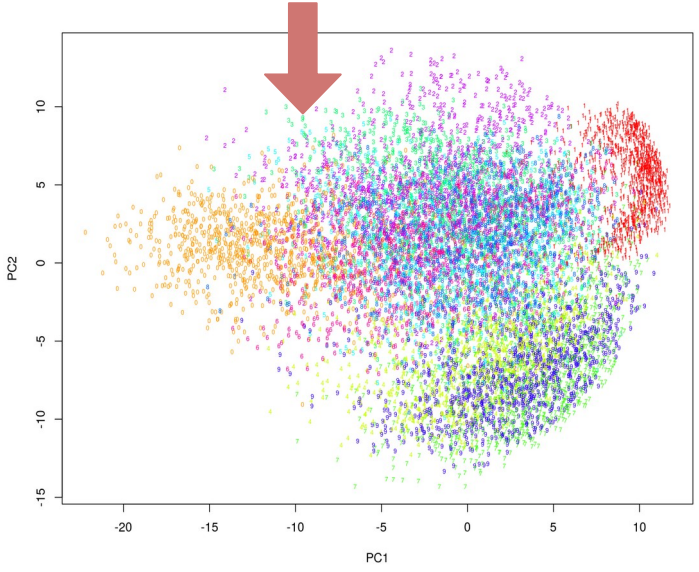
```
barplot(eig$values / sum(eig$values))
```

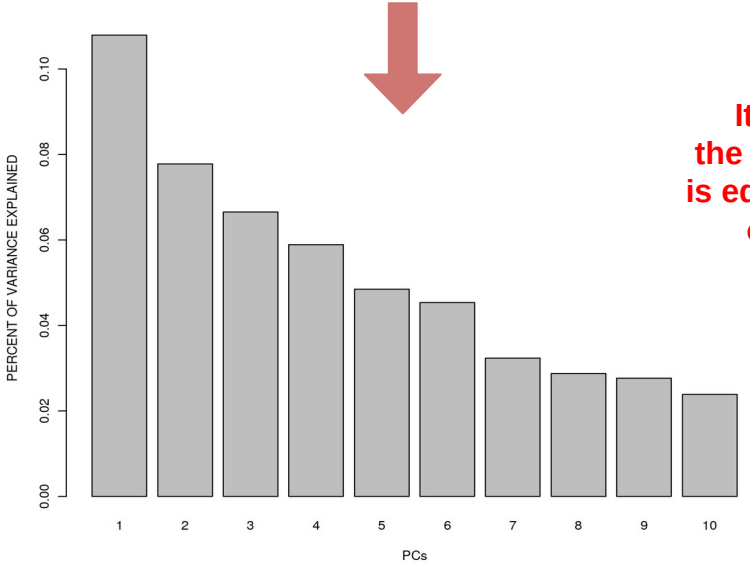
Mathematically:

$M_{ij} = X_{ij} - \mu_j$

$A = (1/N)*M^T*M$

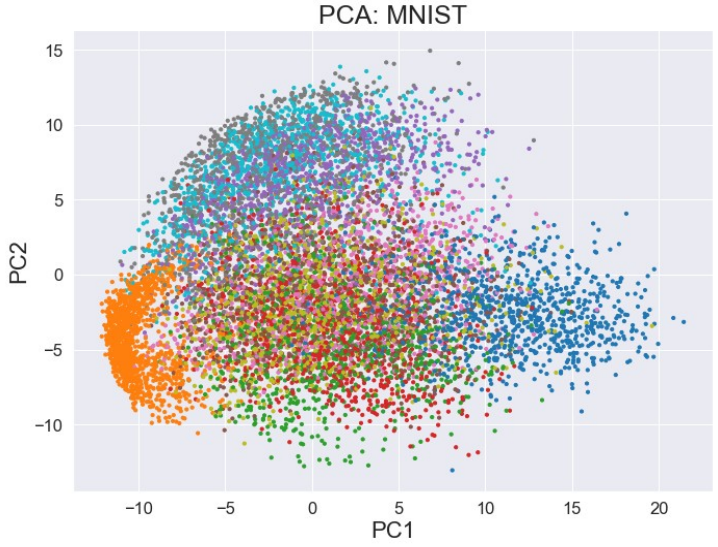
$A*u = \lambda*u$

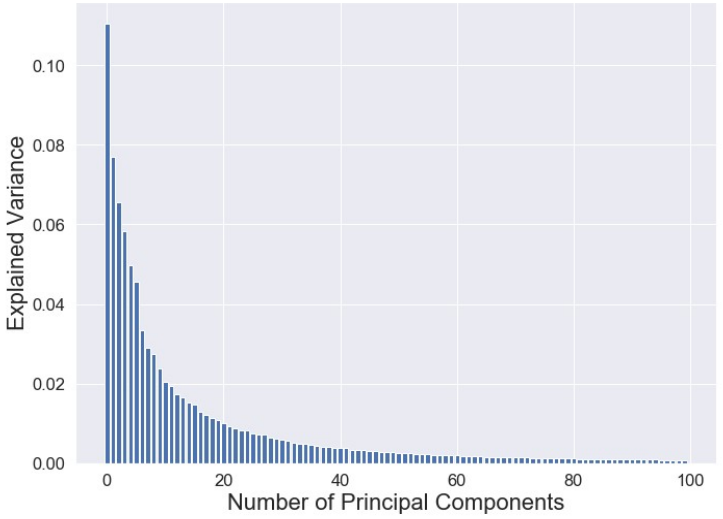


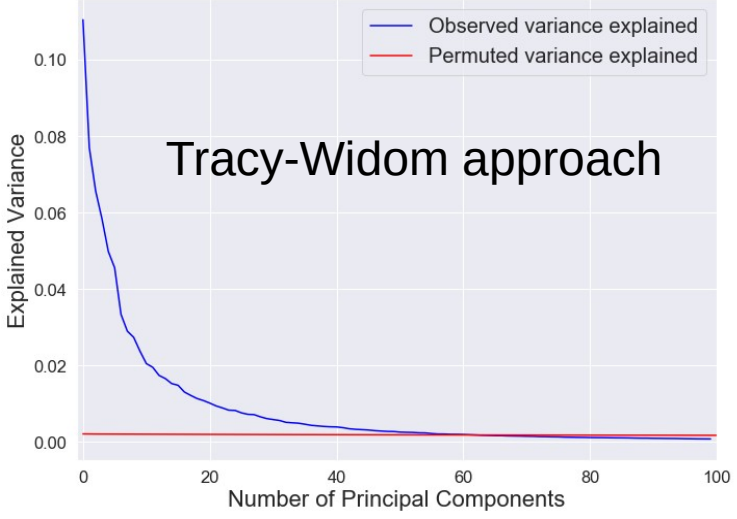


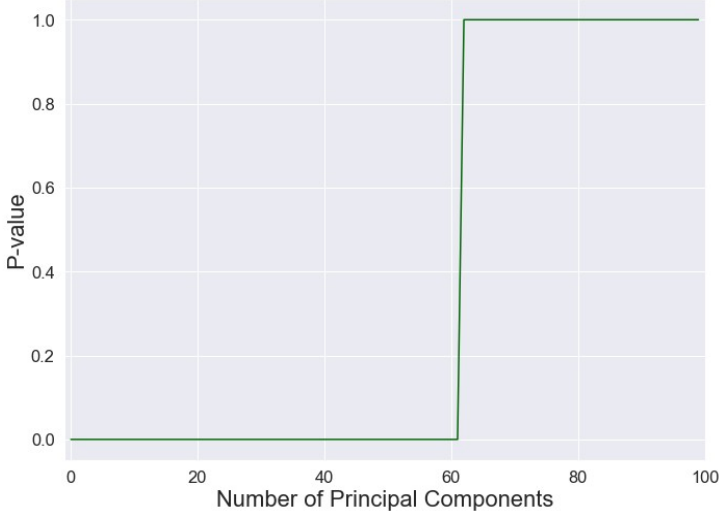
It can be analytically derived that the eigen value decomposition in PCA is equivalent to projecting data on axes of maximal variation in the data

Estimating the number of informative PCs





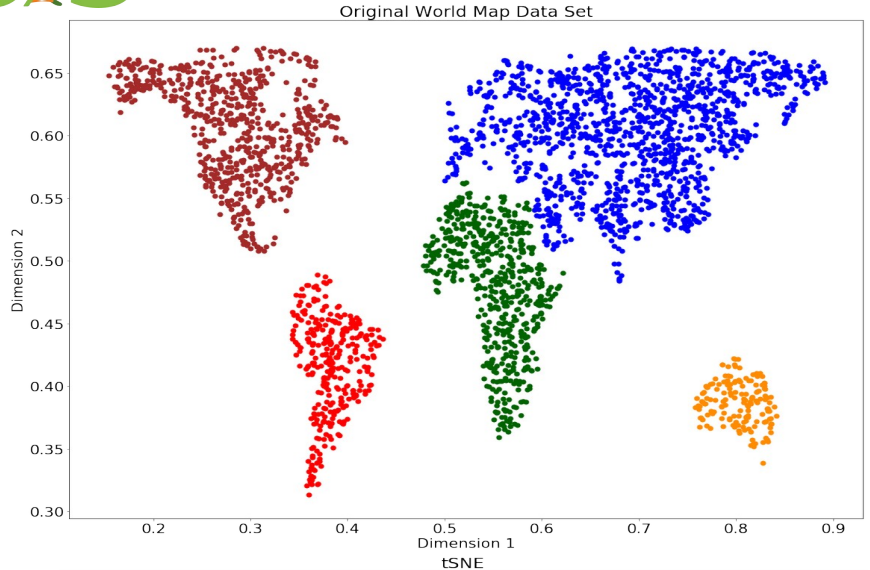


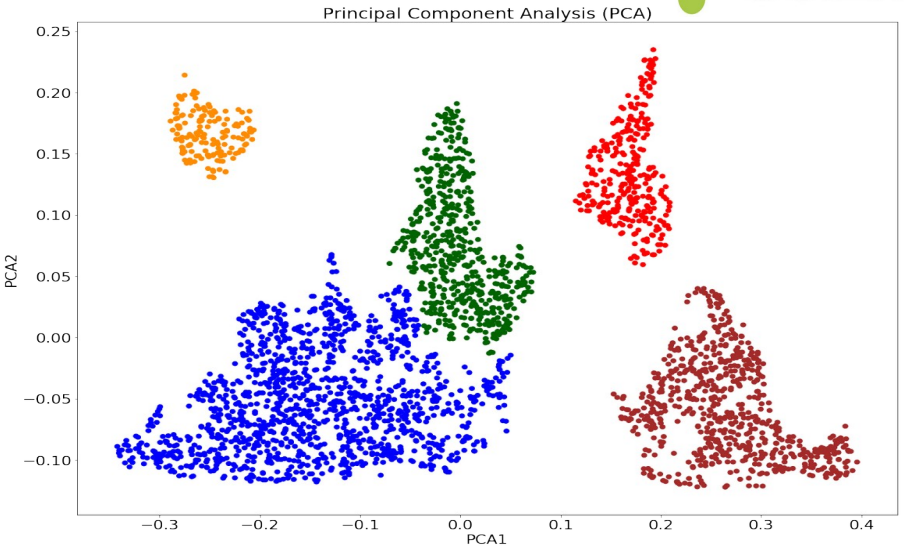


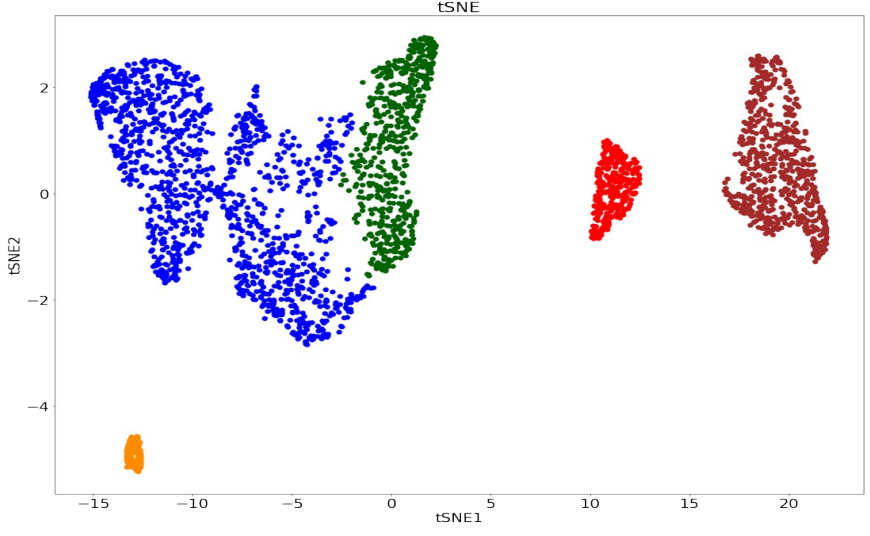
In Seurat: JackStraw

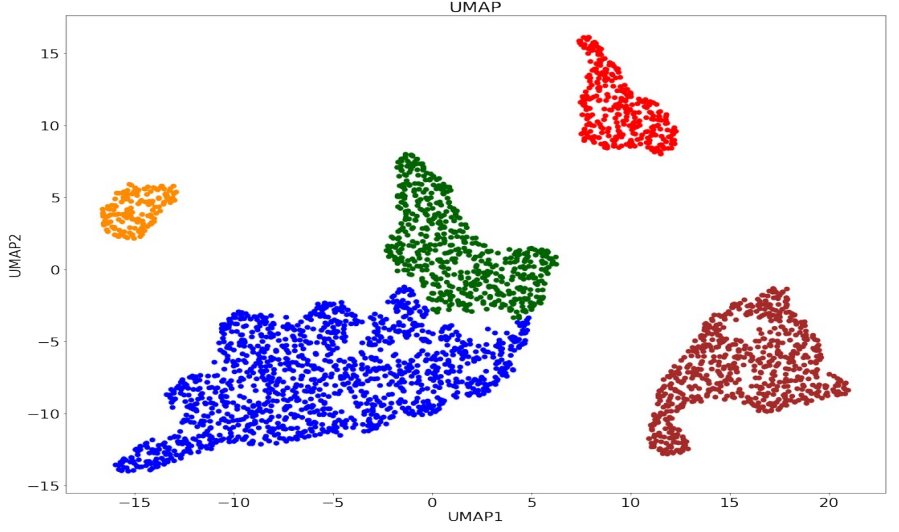
When linear dimension reduction is not good enough

PCA works fine on a linear manifold

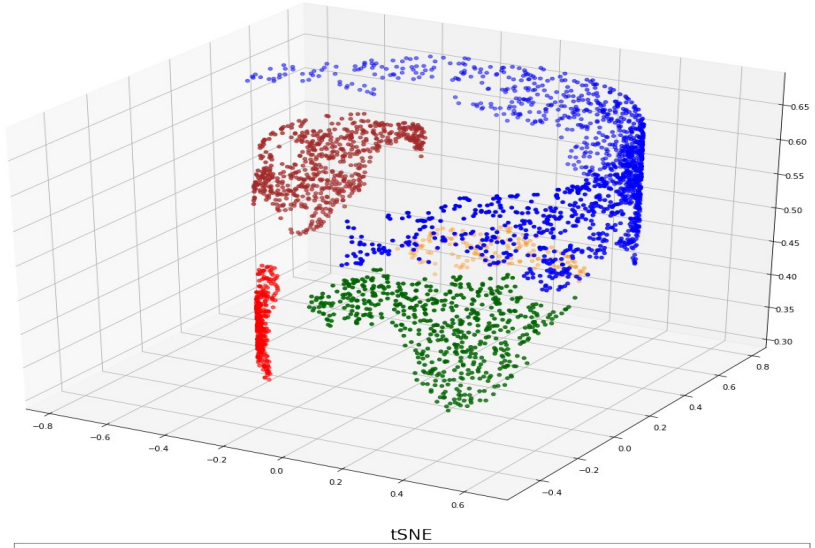


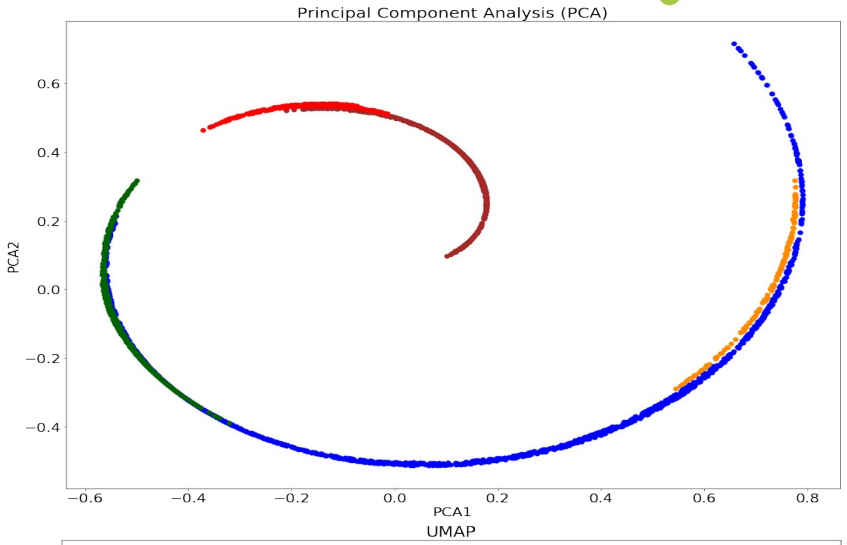




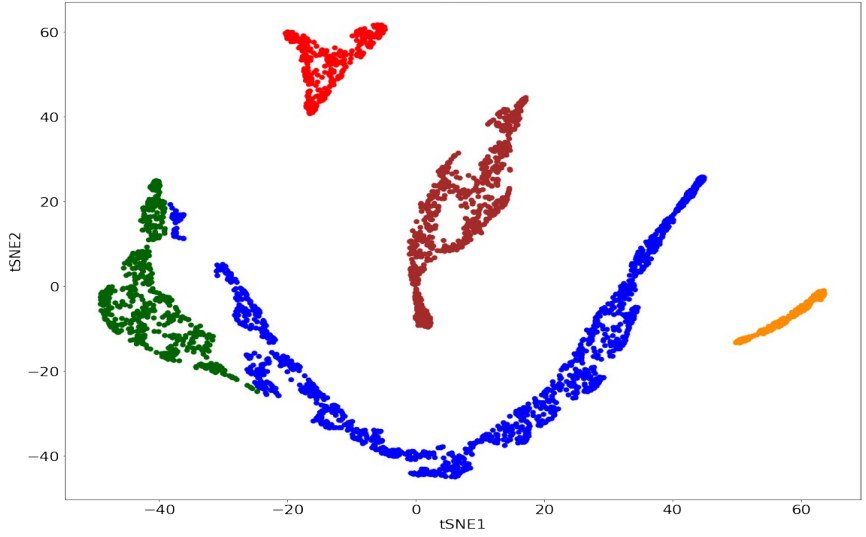


PCA vs. tSNE vs. UMAP on non-linear manifold

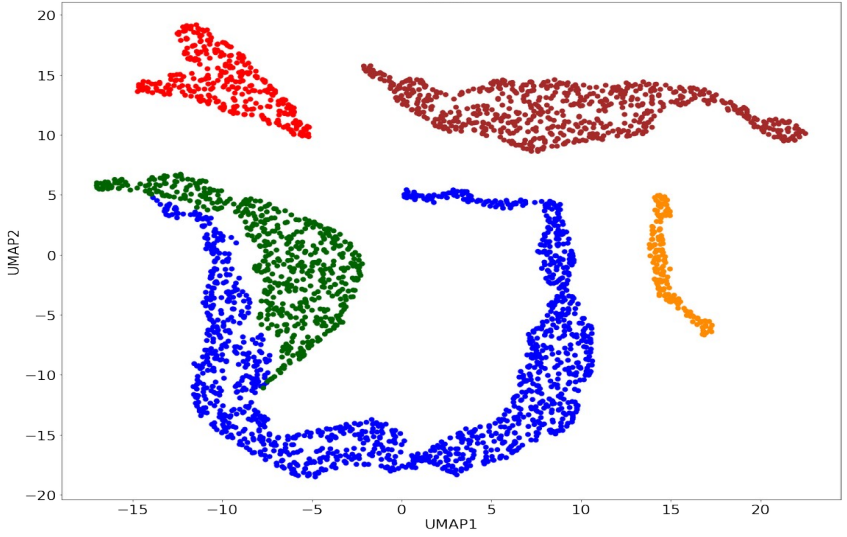




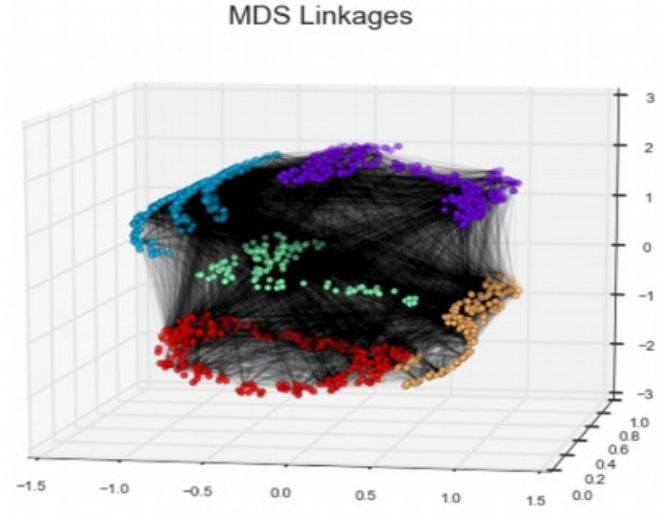
tSNE

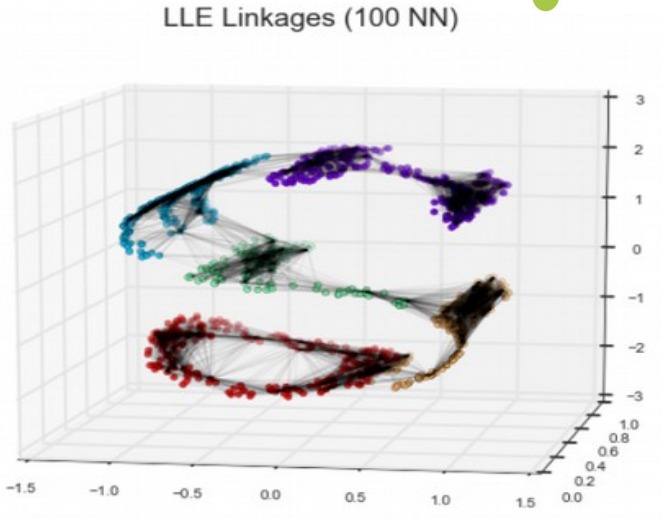


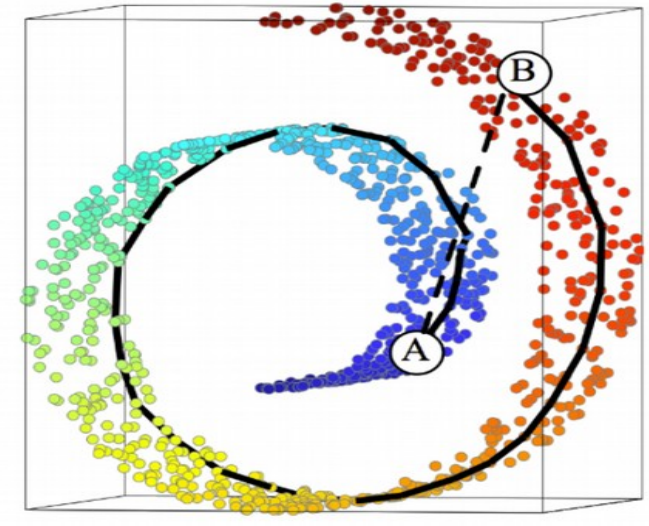
UMAP

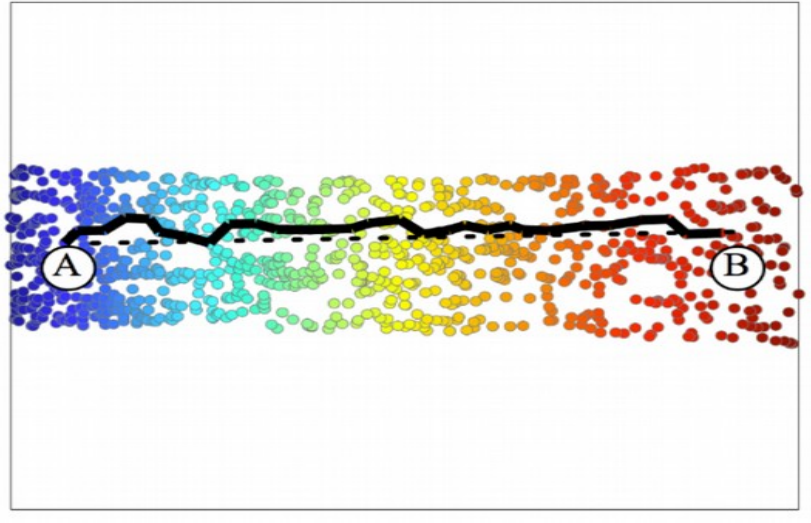


Why PCA can’t unwrap the Swiss Roll



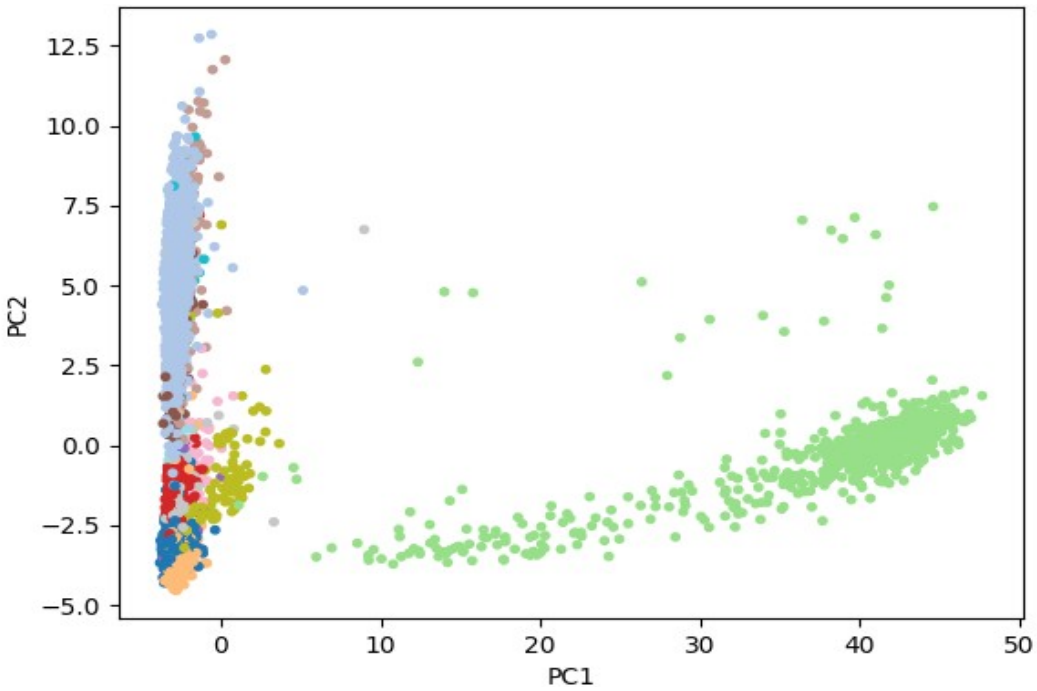


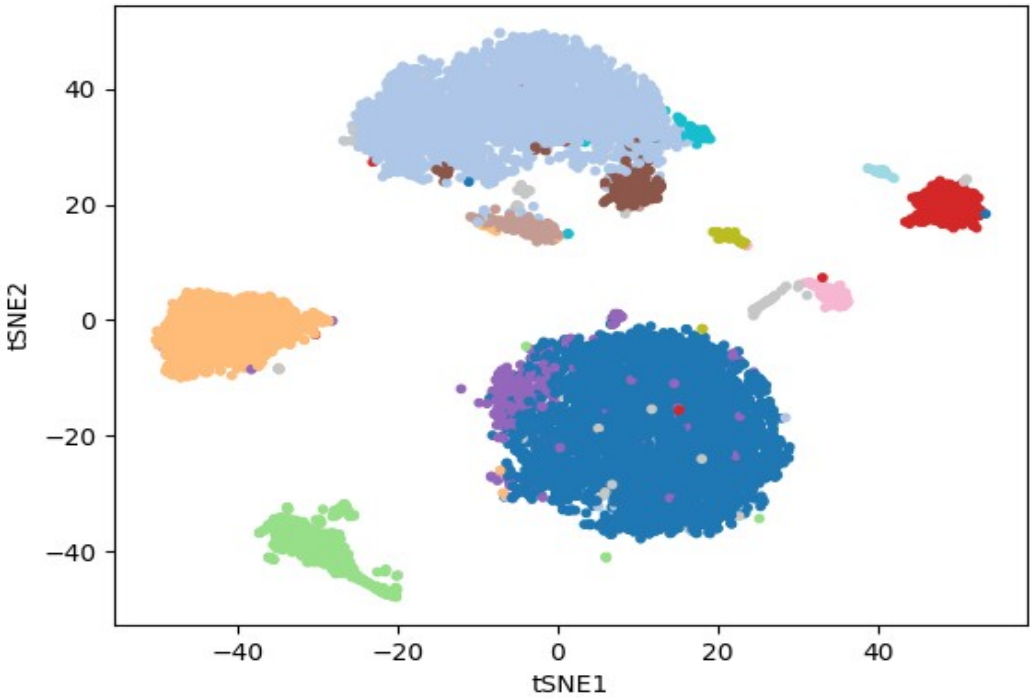




Why PCA is not enough in Single Cell

3k Peripheral Blood Mononuclear Cells (PBMC) available from 10X Genomics





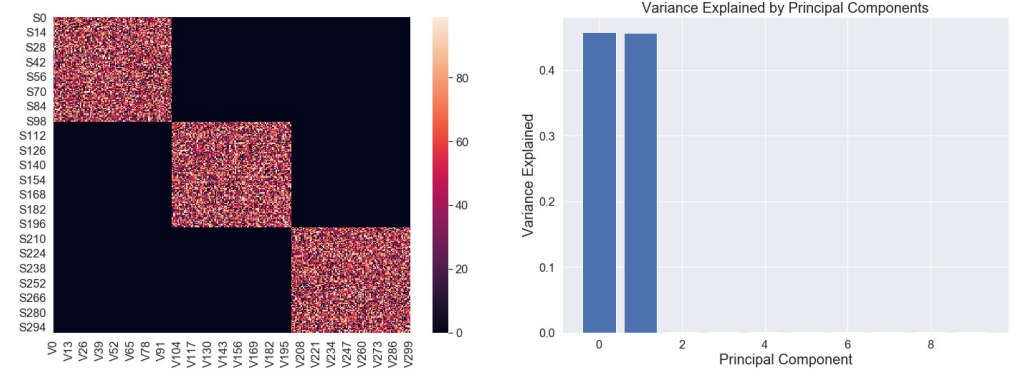
Two principal components (PCs) seem to be insufficient to fully reveal heterogeneity in single cell gene expression data.

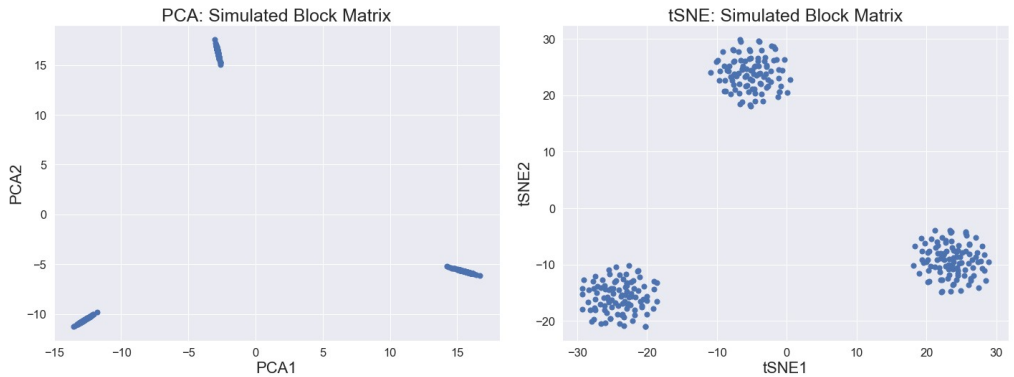
Solution: use more PCs or tSNE / UMAP

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

PCA vs. tSNE: when data complexity grows

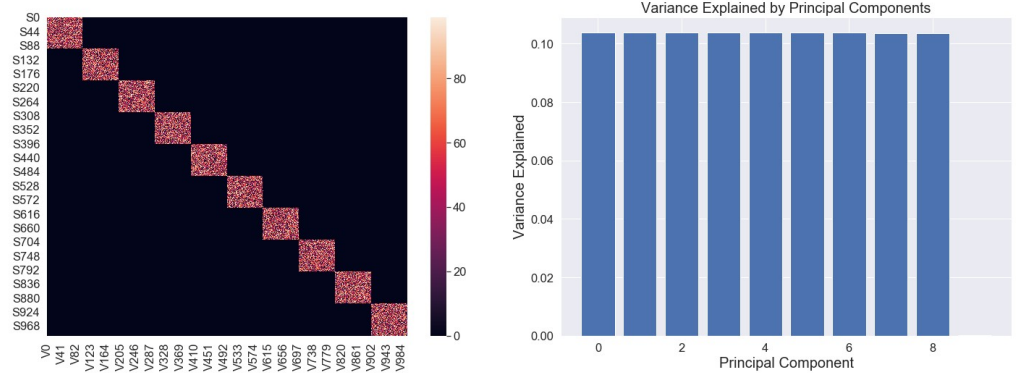
Three classes of data points

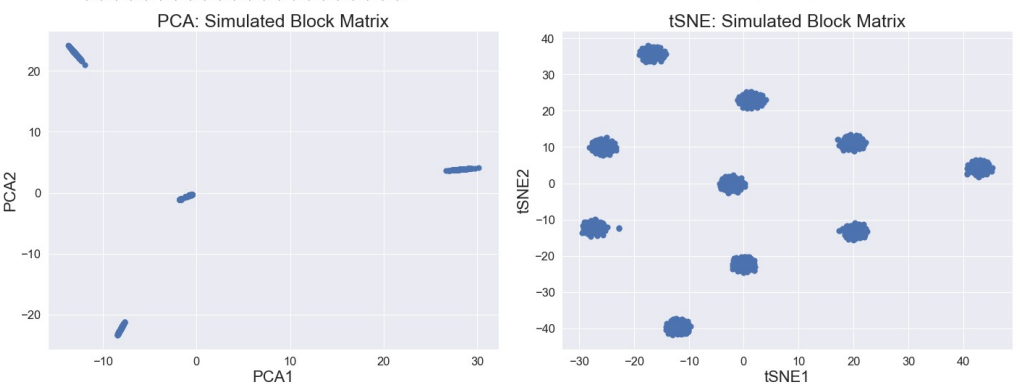




PCA and tSNE tell the same story

Ten classes of data points



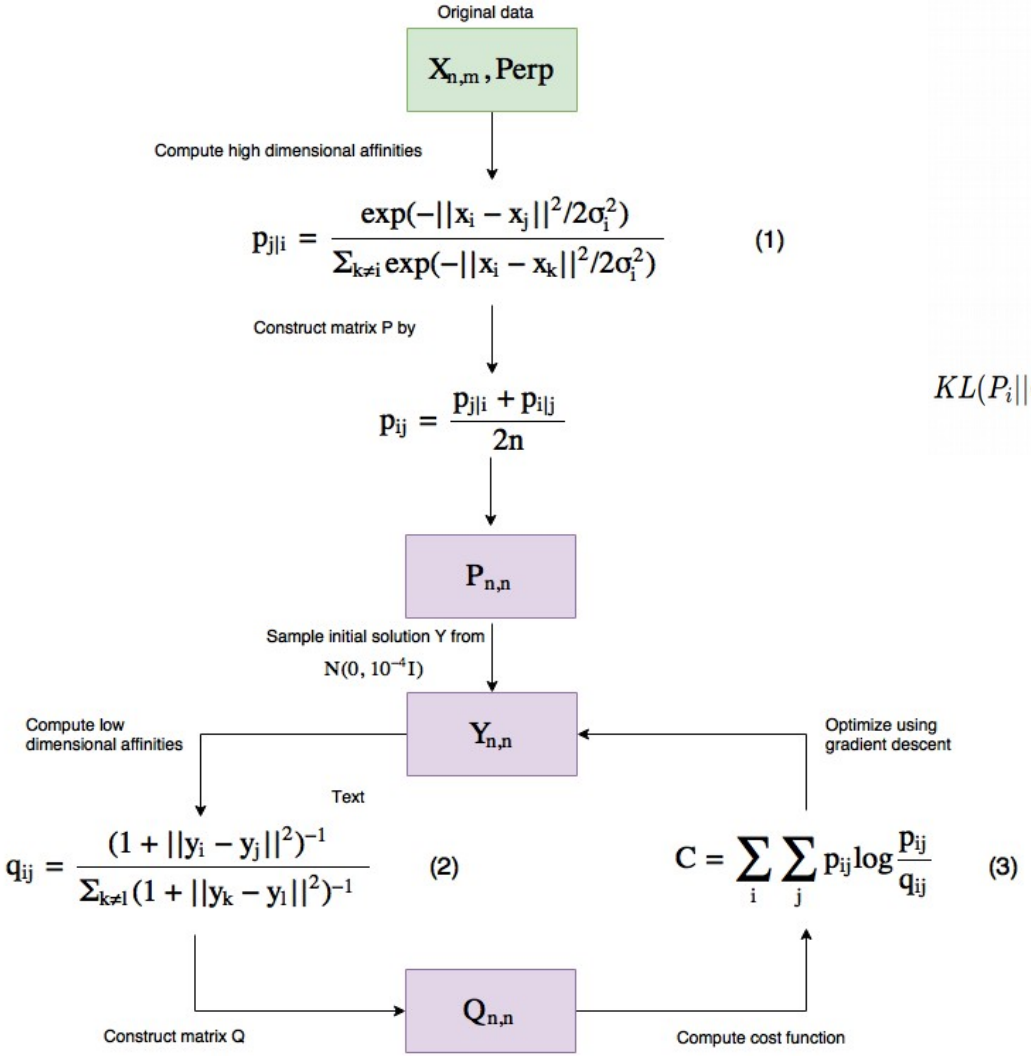


tSNE is more informative than PCA

Oskolkov et al., unpublished

Neighborhood graph dimension reduction: tSNE and UMAP

tSNE dimension reduction algorithm

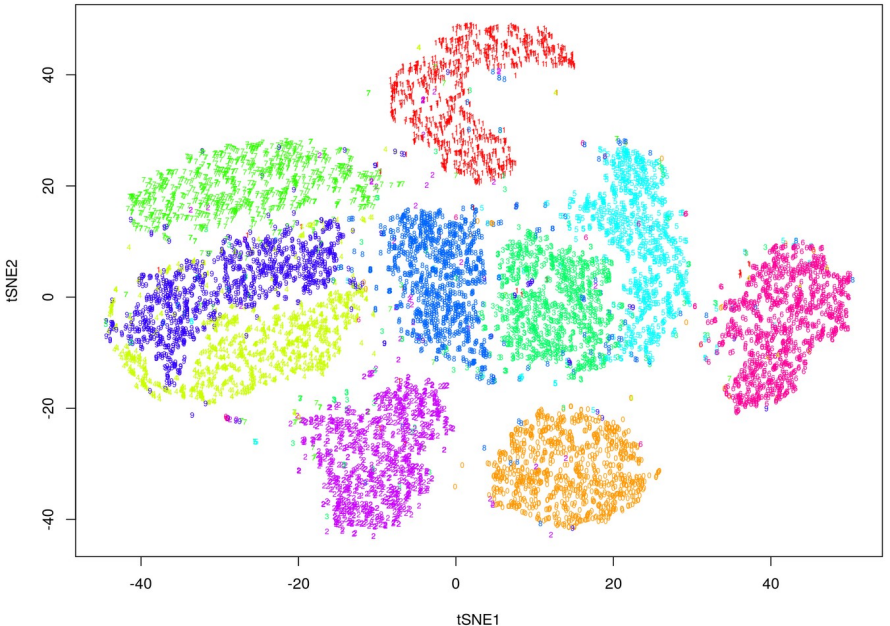


$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-||x_i - x_k||^2/2\sigma_i^2\right)}, \qquad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \tag{1}$$

$$\text{Perplexity} = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}} \tag{2}$$

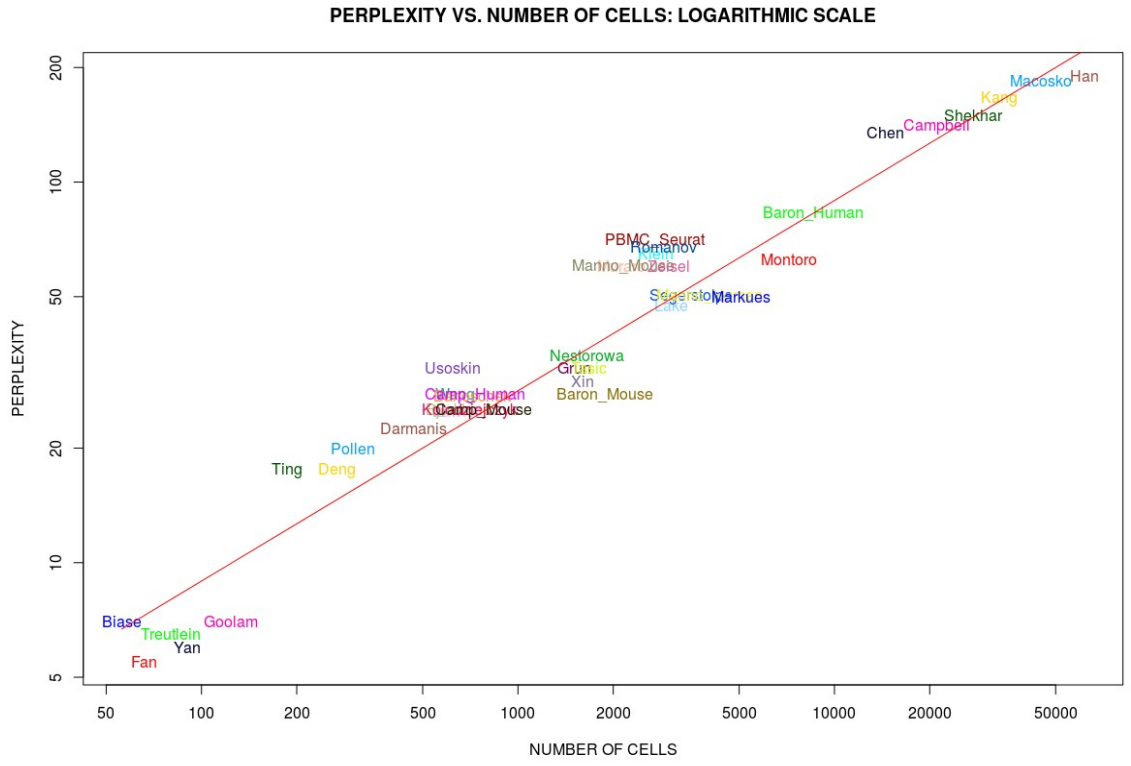
$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k\neq l}\left(1 + ||y_k - y_l||^2\right)^{-1}} \tag{3}$$

$$KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \qquad \frac{\partial KL}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + ||y_i - y_j||^2\right)^{-1} \tag{4}$$



How to select optimal perplexity

Van der Maaten: “Loosely speaking, one could say that a larger / denser dataset requires a larger perplexity.“



$$\log(Perp) = -0.179 + 0.51*\log(N)$$

$$Perp \sim N^{(1/2)}$$

Limitations of tSNE and promise of UMAP

tSNE does not scale for large data sets?

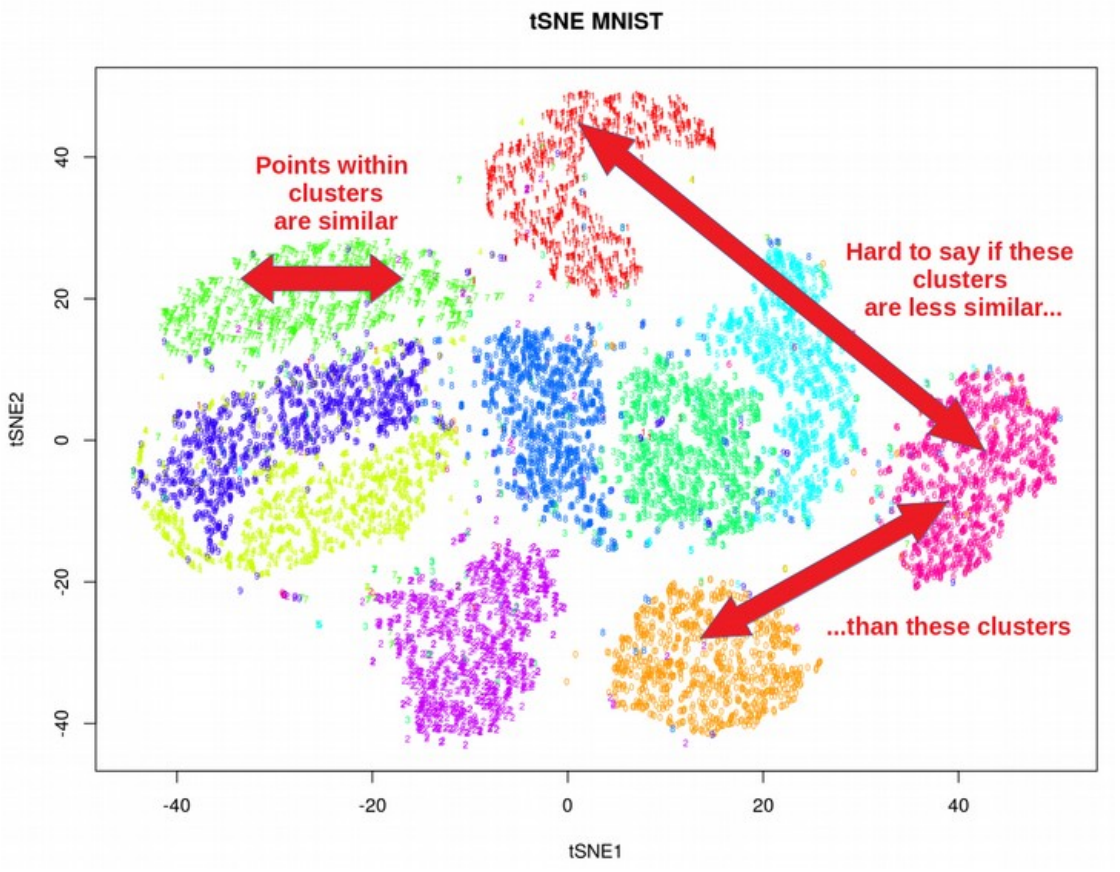
tSNE does not preserve global structure?

tSNE can only embed into 2-3 dims?

tSNE performs non-parametric mapping (no variance explained statistics)?

tSNE can not work with high-dimensional data directly (PCA needed)?

tSNE uses too much RAM at large perp?



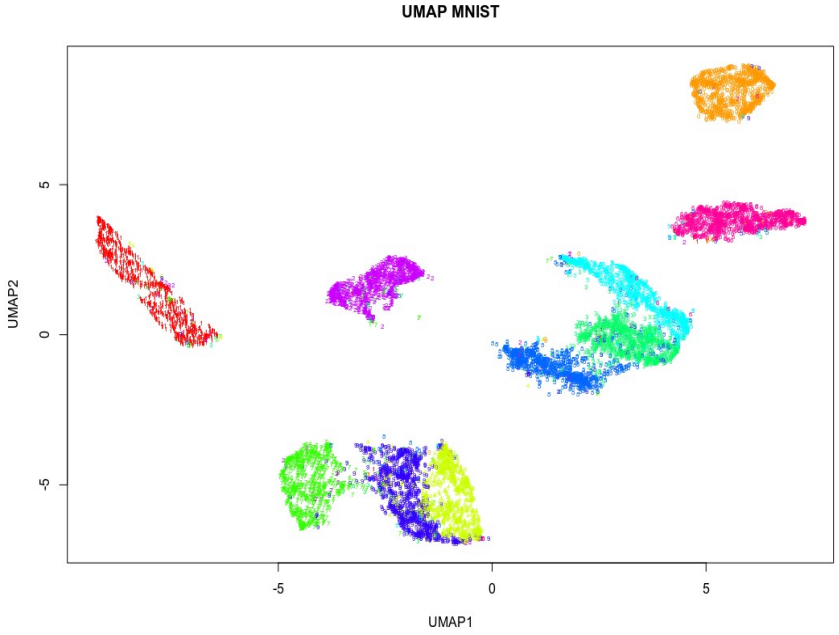
How is UMAP different from tSNE

UMAP uses local connectivity for high-dim probabilities

$$p_{i|j} = e^{-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}}$$

UMAP does not normalize probabilities (speed-up)

UMAP can deliver a number of components for clustering



UMAP uses Laplacian Eigenmap for initialization

UMAP uses Cross-Entropy (not KL) as cost function

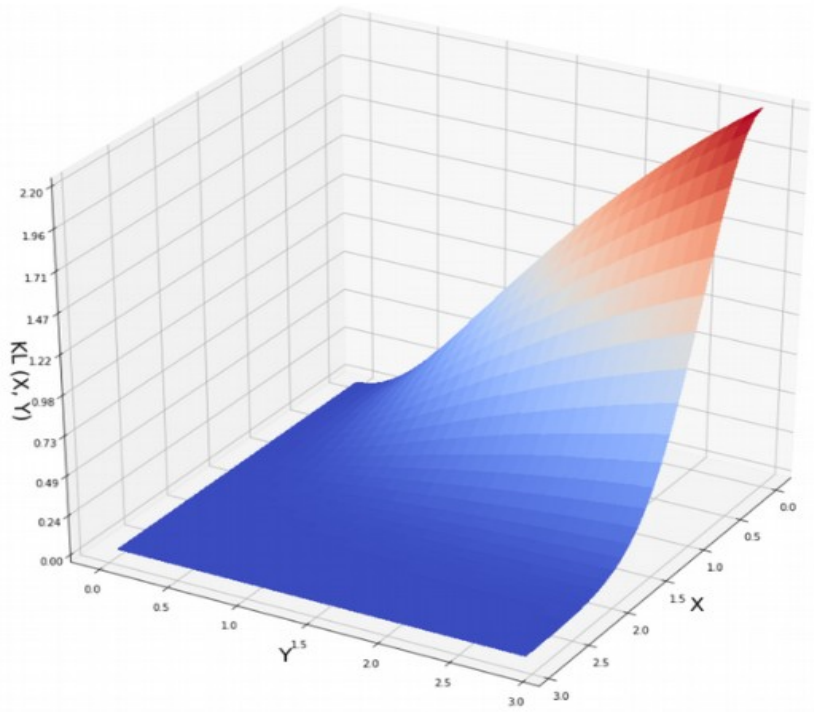
$$CE(X,Y) = \sum_i \sum_j \left[p_{ij}(X) \log\left(\frac{p_{ij}(X)}{q_{ij}(Y)}\right) + (1 - p_{ij}(X)) \log\left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)}\right) \right]$$

This is similar to tSNE cost function

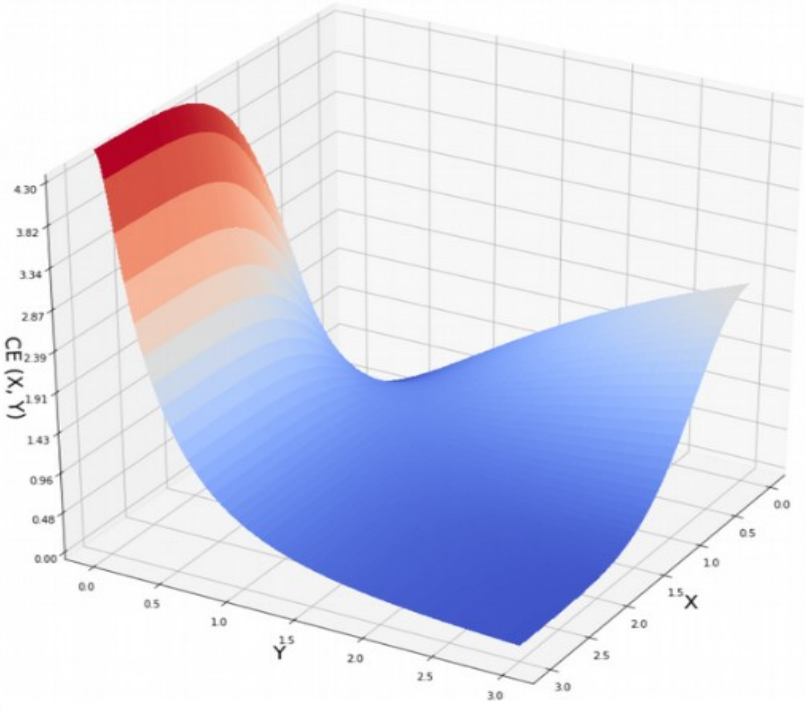
This term is UMAP specific

tSNE vs. UMAP:
global structure preservation

Cost function seems to make UMAP preserve more of global structure than tSNE

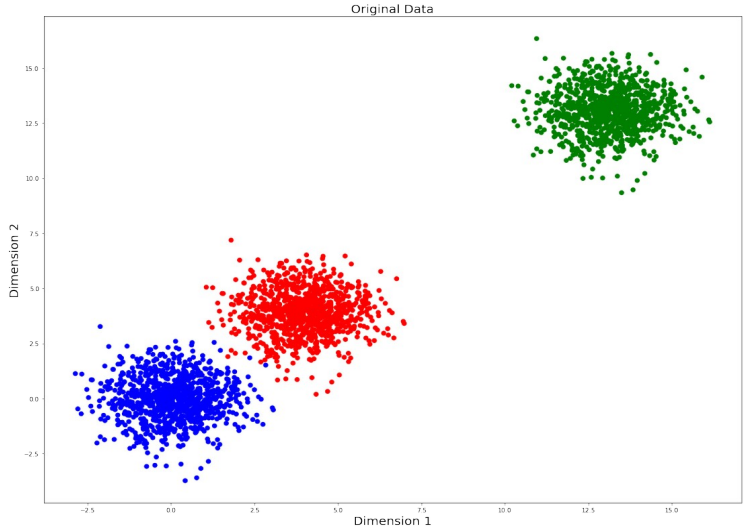


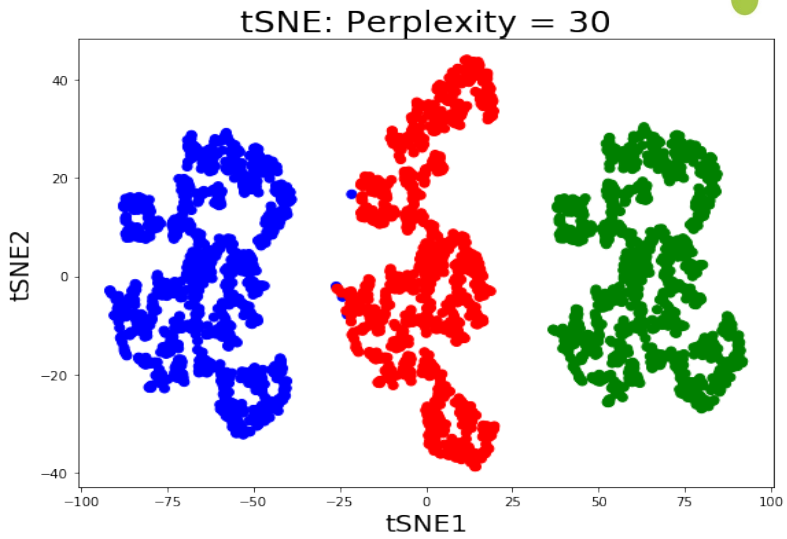
X → infinity, Y can be any

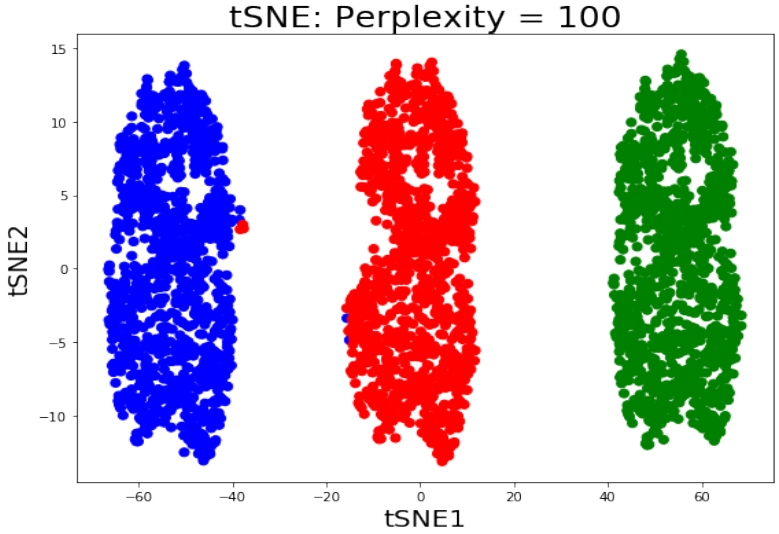


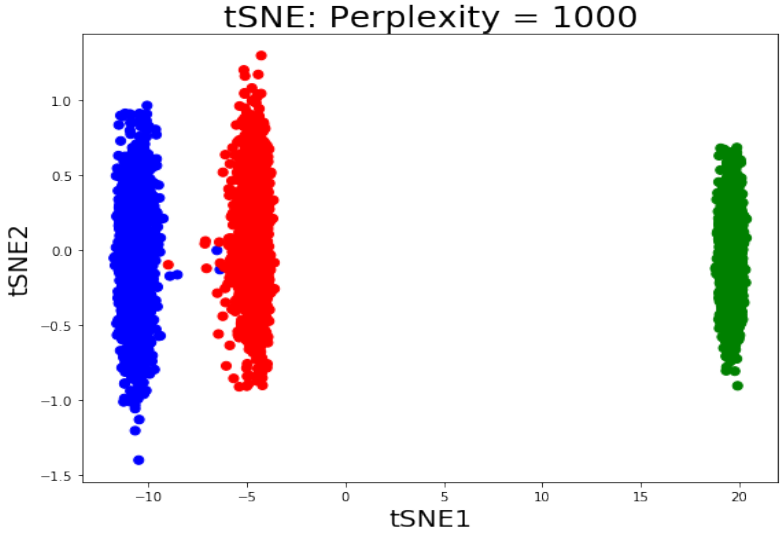
X → infinity, Y → infinity

Why preserving global structure is important



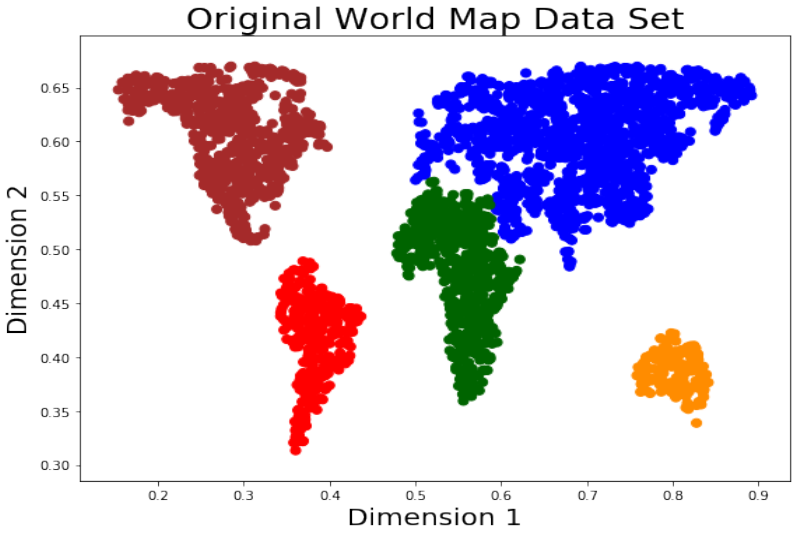


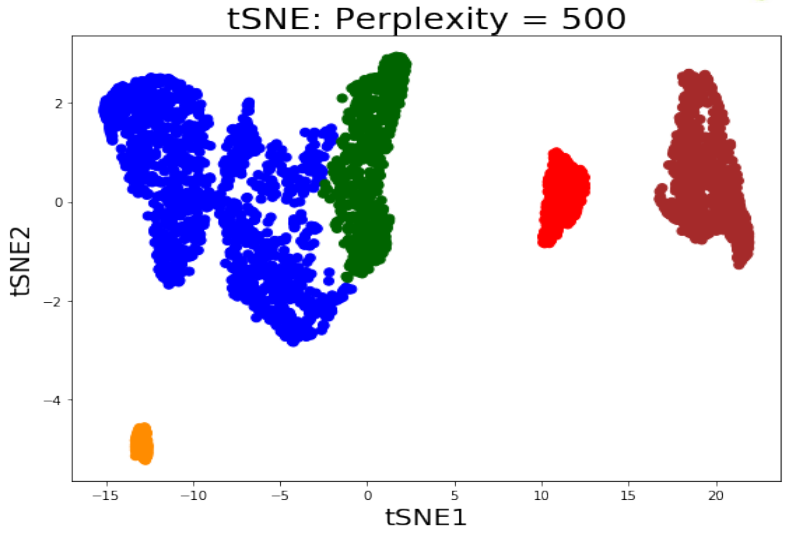


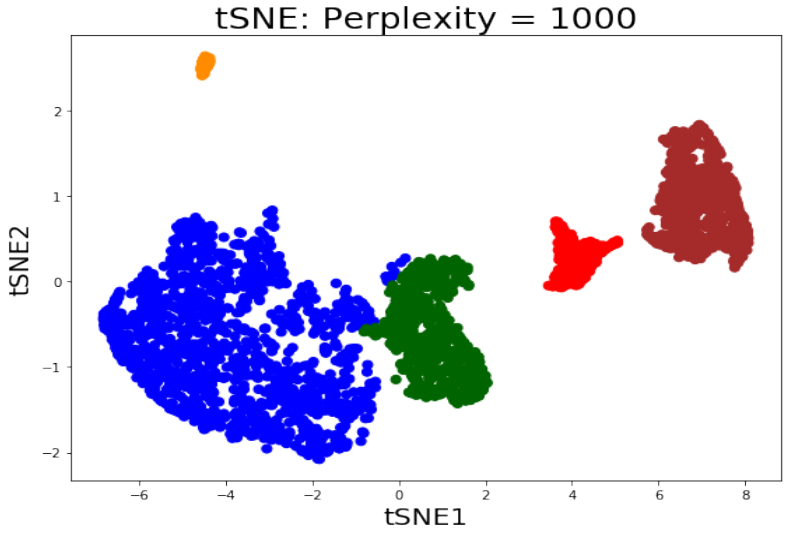


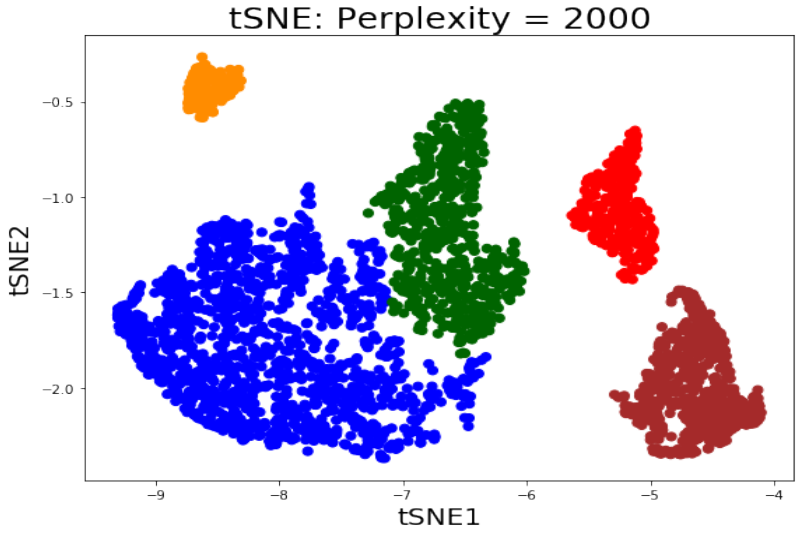
Can large perplexity solve the problem of global structure for tSNE?

Why preserving global structure is important



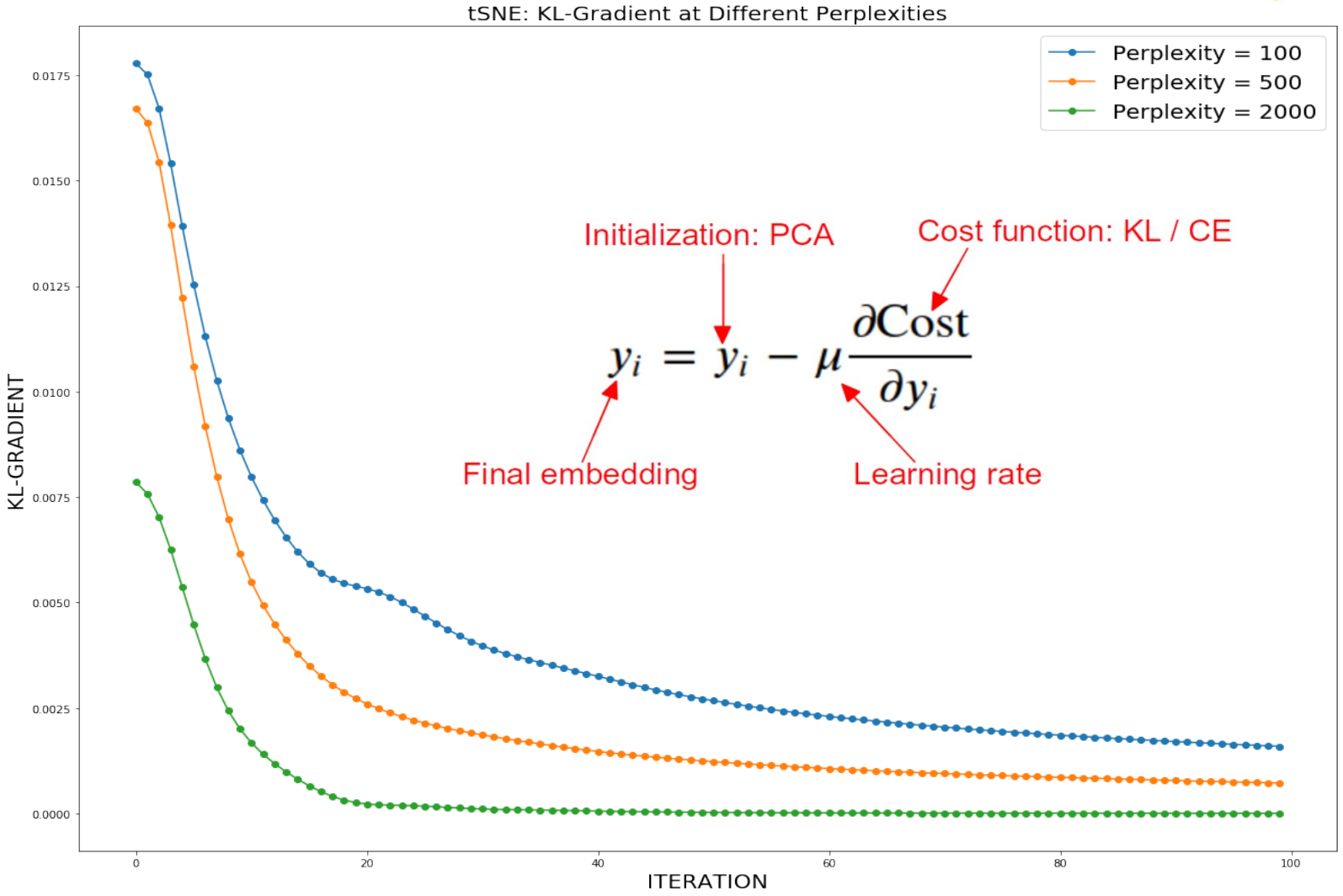




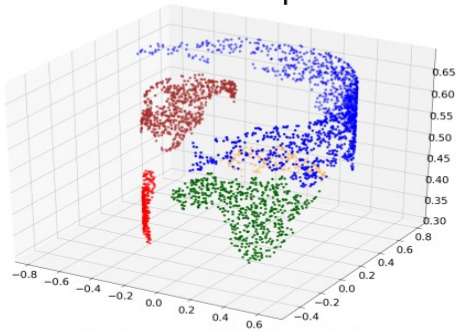
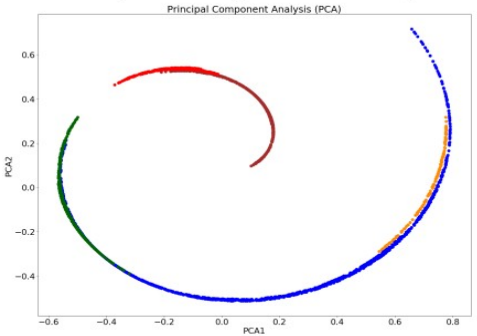
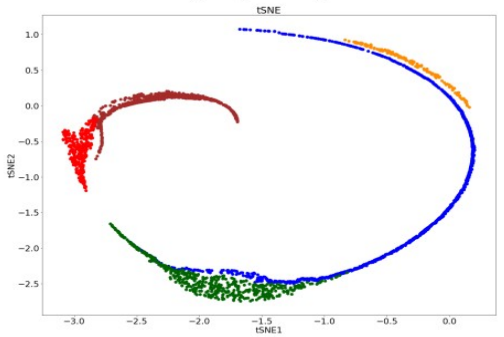
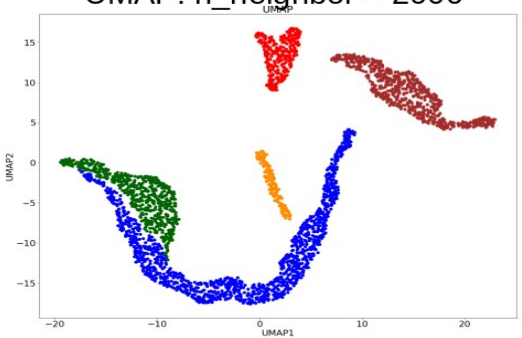
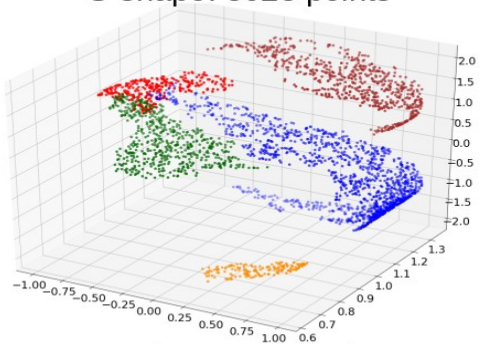
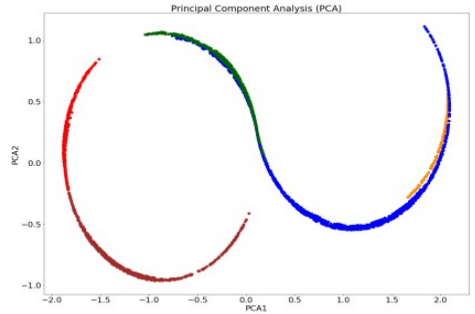
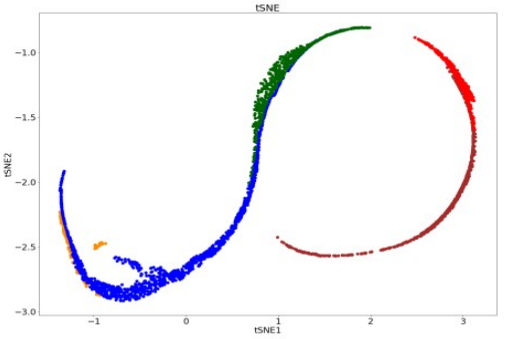
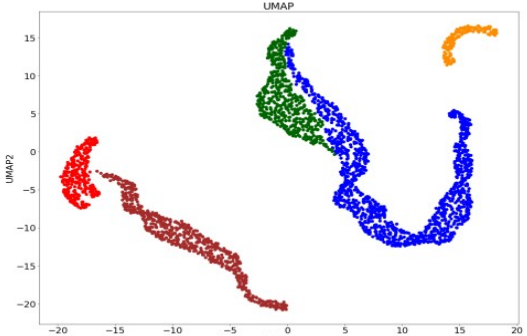
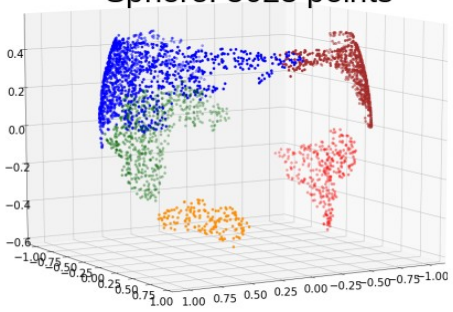
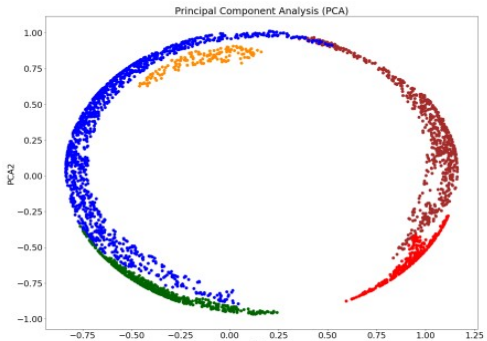
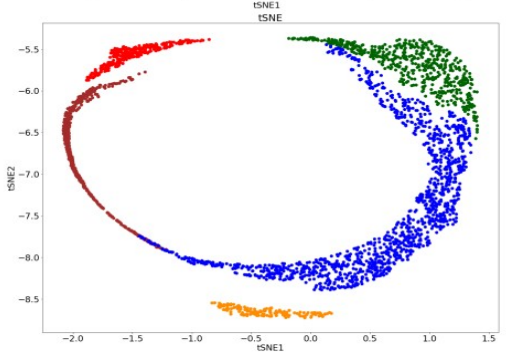
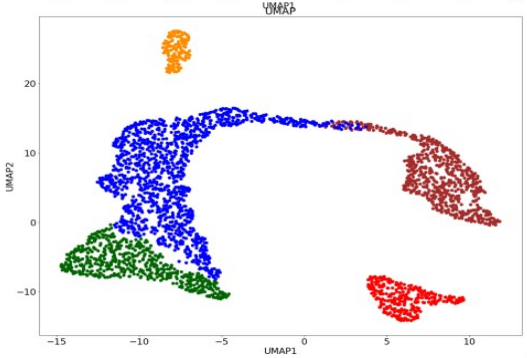


Can large perplexity solve the problem of global structure for tSNE?

KL-gradient goes to zero at large perplexity

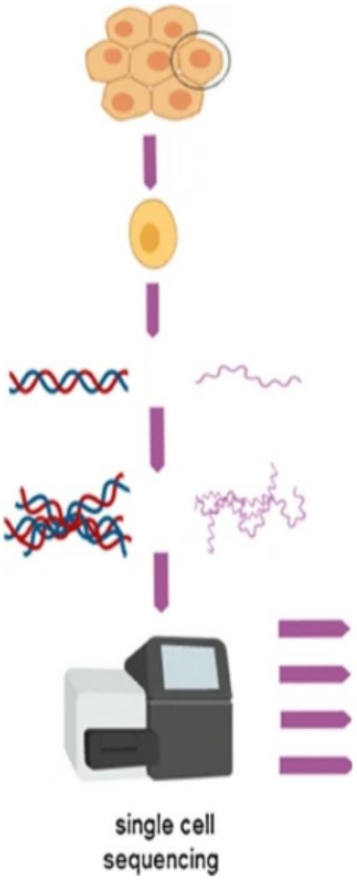


tSNE degrades to PCA on non-linear manifold at large perplexity

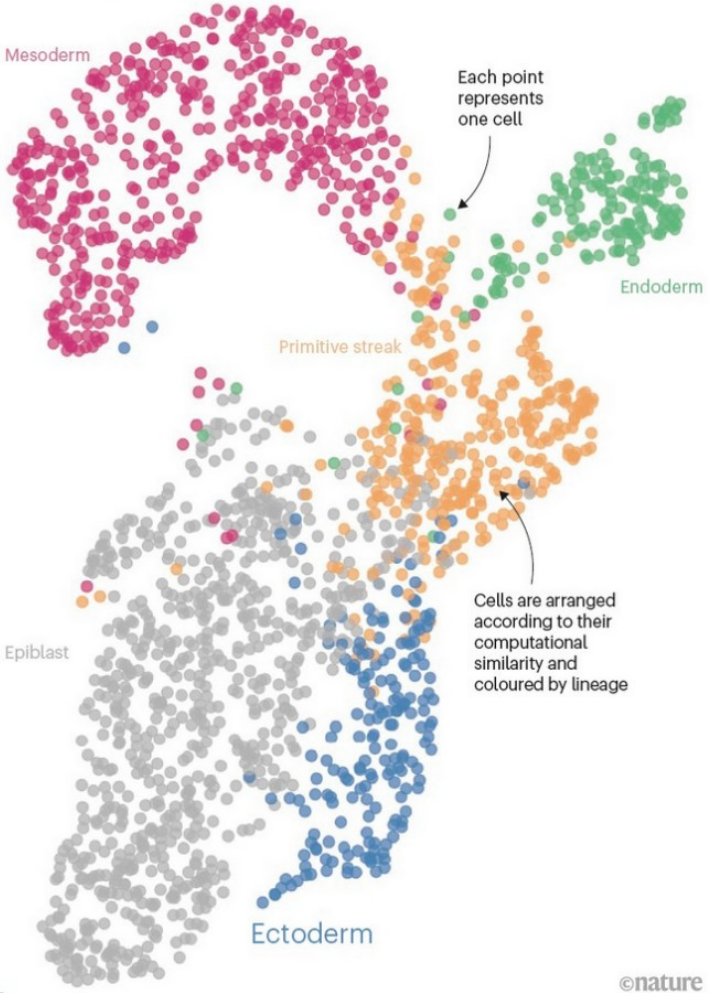
Swiss Roll: 3023 points	Principal Component Analysis	tSNE: perplexity = 2000	UMAP: n_neighbor = 2000
			
S-shape: 3023 points			
			
Sphere: 3023 points			
			

UMAP for Population Genomics applications: addressing some recent fair and unfair criticism

UMAP: Single Cell vs. PopGen

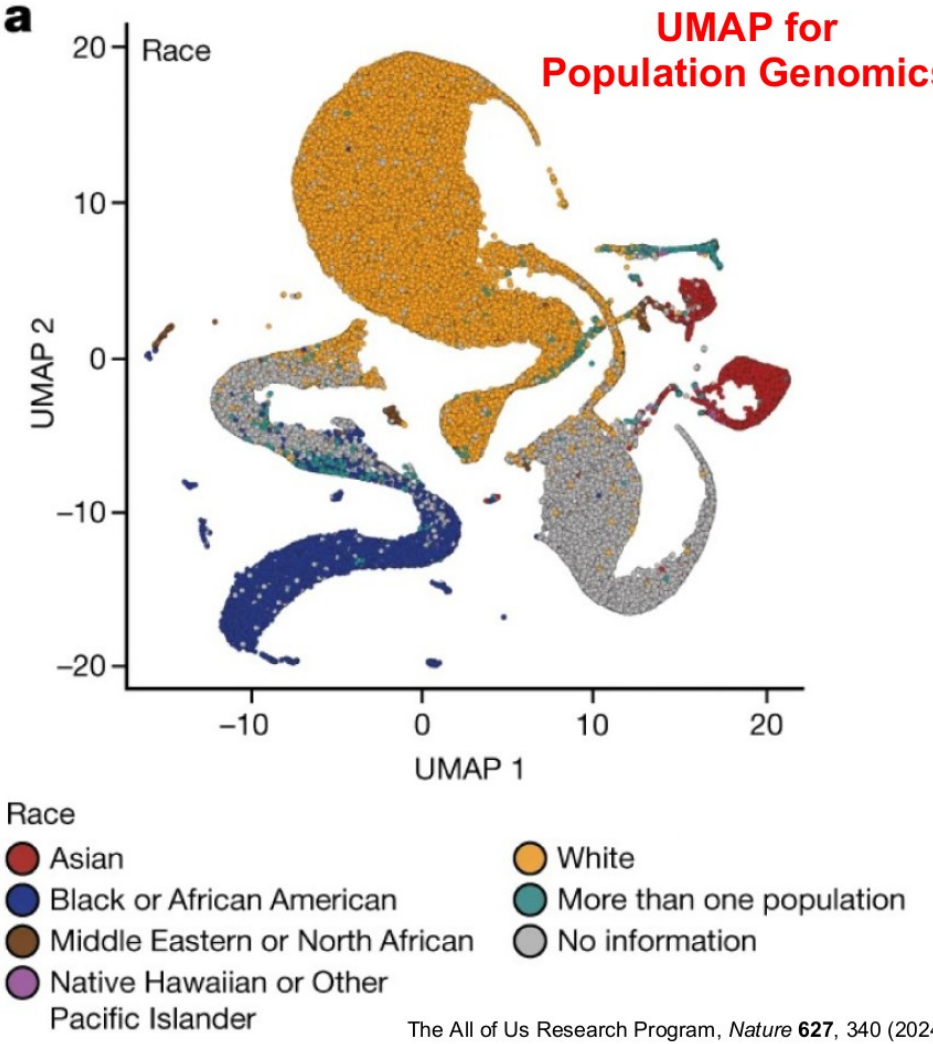


UMAP for Single Cell Biology



https://www.nature.com/articles/d41586-021-01994-w

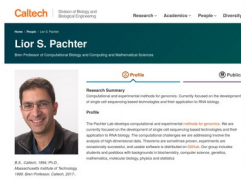
UMAP for Population Genomics



The All of Us Research Program, *Nature* **627**, 340 (2024)

UMAP (and Single Cell) Criticism

PLOS COMPUTATIONAL BIOLOGY



PERSPECTIVE

The specious art of single-cell genomics

Tara Chari[1], **Lior Pachter**[1,2]*

1 Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, United States of America, 2 Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, United States of America

* lpachter@caltech.edu

Abstract

Dimensionality reduction is standard practice for filtering noise and identifying relevant features in large-scale data analyses. In biology, single-cell genomics studies typically begin with reduction to 2 or 3 dimensions to produce "all-in-one" visuals of the data that are amenable to the human eye, and these are subsequently used for qualitative and quantitative exploratory analysis. However, there is little theoretical support for this practice, and we show that extreme dimension reduction, from hundreds or thousands of dimensions to 2, inevitably induces significant distortion of high-dimensional datasets. We therefore examine the practical implications of low-dimensional embedding of single-cell data and find that extensive distortions and inconsistent practices make such embeddings counter-productive for exploratory, biological analyses. In lieu of this, we discuss alternative approaches for conducting targeted embedding and feature exploration to enable hypothesis-driven biological discovery.

OPEN ACCESS

Citation: Chari T, Pachter L (2023) The specious art of single-cell genomics. PLoS Comput Biol 19(8): e1011288. https://doi.org/10.1371/journal.pcbi.1011288

Editor: Jason A. Papin, University of Virginia, UNITED STATES

Published: August 17, 2023

Data Availability Statement: Download links for the original data used to generate the figures and results in the paper are listed in Table A in S1 Text. Processed and normalized versions of the count matrices are available on CaltechData, with links provided in Table B in S1 Text. All analysis code used to generate the figures and results in the paper is available at https://github.com/pachterlab/CP_2023 and deposited at Zenodo (DOI https://doi.org/10.5281/zenodo.8087950). Code is provided in Colab notebooks which can be run for free on the Google cloud.

Funding: L.P. received the National Institutes of Health (nih.gov) award U19MH114830, administered by the National Institute of Mental Health (nimh.nih.gov). T.C. and L.P. were partially

Introduction

The high-dimensionality of "big data" genomics datasets has led to the ubiquitous application of dimensionality reduction to filter noise, enable tractable computation, and to facilitate exploratory data analysis (EDA). Ostensibly, the goal of this reduction is to preserve and extract local and/or global structures from the data for biological inference [1–3]. Trial and error application of common techniques has resulted in a currently popular workflow combining initial dimensionality reduction to a few dozen dimensions, often using principal component analysis (PCA), with further nonlinear reduction to 2 dimensions using t-SNE [4] or UMAP [1,2,5,6]. For single-cell genomics in particular, these embeddings are used extensively in qualitative and quantitative EDA tasks that fall into 4 main categories of applications (Fig 1, "Application"):

- Modality-mixing, integration, and reference mapping:

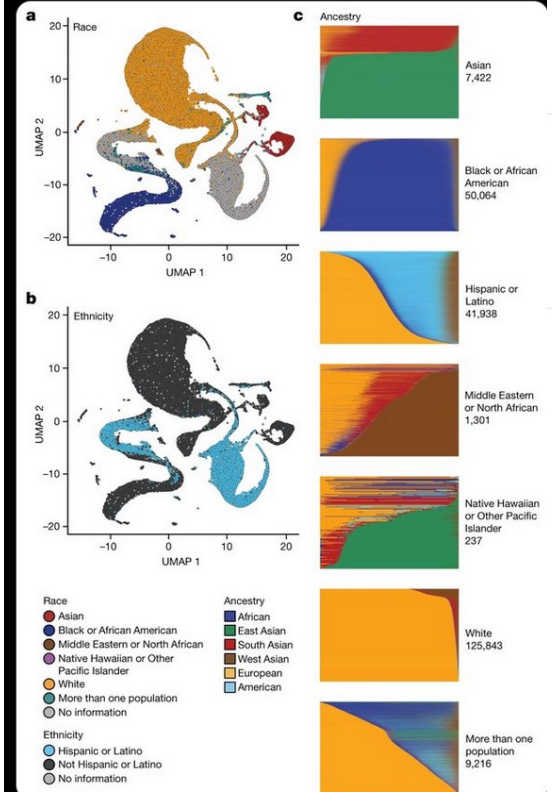
  Embeddings are used to visually assess the extent of integration, mixing, or similarities between cells from different batches [7–9] and to compare methods of integration/batch-correction [10]. For query dataset(s) mapped onto reference datasets/embeddings, visuals likewise provide an assessment of merged data similarities or differences [11,12].

- Cluster validation and relationships:

← **Post**

Jonathan Pritchard
@jkpritch

It's a pity that All of Us used UMAP to visualize ancestry variation in their new marker paper, out today in Nature.
The UMAP algorithm, by design, exaggerates the distinctiveness of the most frequent ancestries, a message that can be misinterpreted by the public.



Biologists, stop putting UMAP plots in your papers

UMAP is a powerful tool for exploratory data analysis, but without a clear understanding of how it works, it can easily lead to confusion and misinterpretation.

AUTHOR
Rafael Irizarry

PUBLISHED
Dec. 23, 2024

```
library(Matrix)
library(ggplot2)
library(dplyr)
library(umap)
set.seed(2024-6-21)
load("rda/pop_gen_sample.RData")
```

The UMAP craze in singe cell RNA-Seq

Single-cell RNA sequencing (scRNA-seq) has become one of the most widely used technologies in basic biology. With the rise of scRNA-seq, the use of **UMAP** has become ubiquitous in publications. While this dimensionality reduction technique is useful for exploratory data analysis, its overuse and misinterpretation have led to confusion and

Going through the post of Rafael Irizarry:

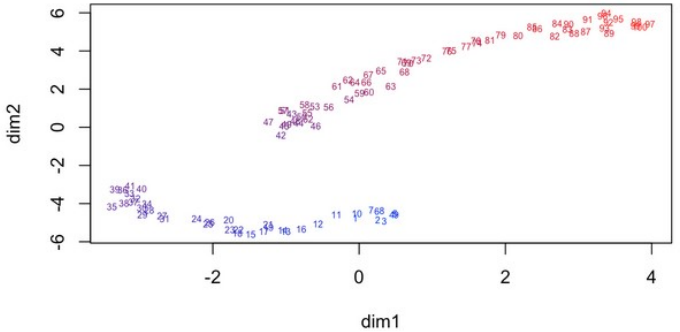
Point1: UMAP makes artificial clusters!

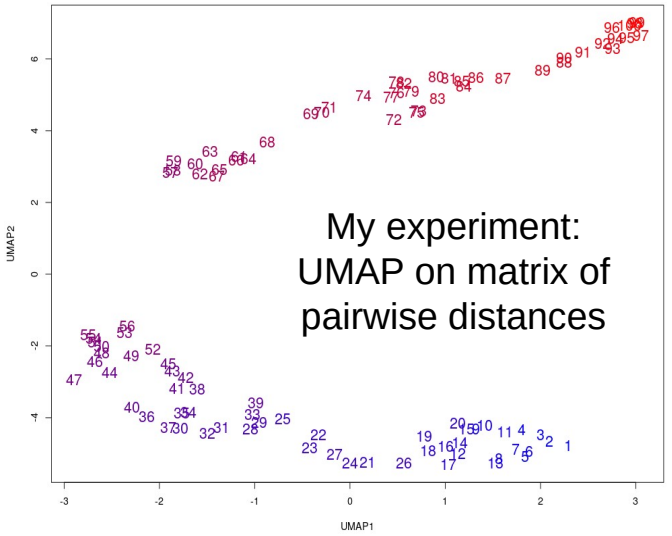
Are we looking at 3- or 100-dimensional data?

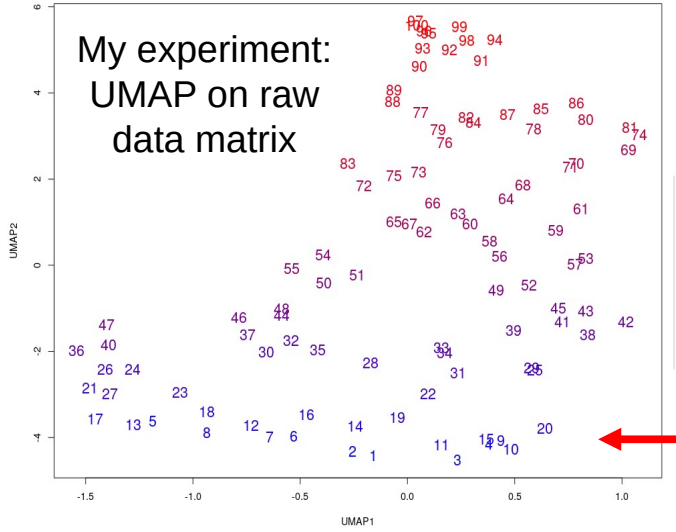
The issue becomes more significant when the underlying mathematics of UMAP is not fully understood. UMAP takes a p-dimensional vector of numeric values, such as gene expression in scRNA-Seq, and applies a mathematical transformation to produce two values, resulting in the two coordinates shown in the plot. But what exactly is this function? Do the authors who include these plots in papers fully understand the mathematics behind it? What genes are included in the calculation and how? How exactly does distance in the two dimensional summary relate to the actual distance in p-dimensional space? The actual summary function is rarely if ever explained, leaving readers uncertain about what the plot truly represents.

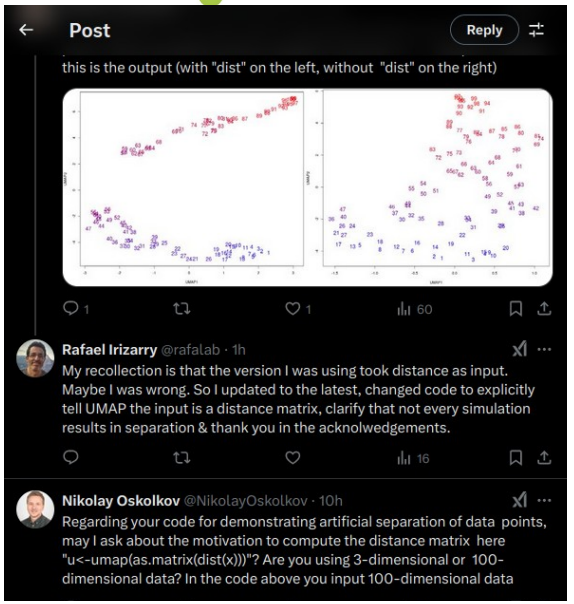
Additionally, UMAP is highly sensitive and can create separations in data that shouldn't necessarily exist. For example, consider applying UMAP to 100 randomly generated points from a multivariate normal distribution representing three correlated random variables:

```
Sigma <- matrix(.8, 3, 3); diag(Sigma) <- 1
x <- MASS::mvrnorm(100, rep(0,3), Sigma)
#x <- matrix(rnorm(100), ncol = 1)
u <- umap(as.matrix(dist(x)))
ranks <- rank(rowMeans(x))
colors <- colorRampPalette(c("blue", "red"))(nrow(x))
colormap <- colors[ranks]
plot(u$layout[,1], u$layout[,2], type = "n", xlab = "dim1", ylab = "dim2")
text(u$layout[,1], u$layout[,2], labels = ranks, col = colormap, cex = 0.5)
```



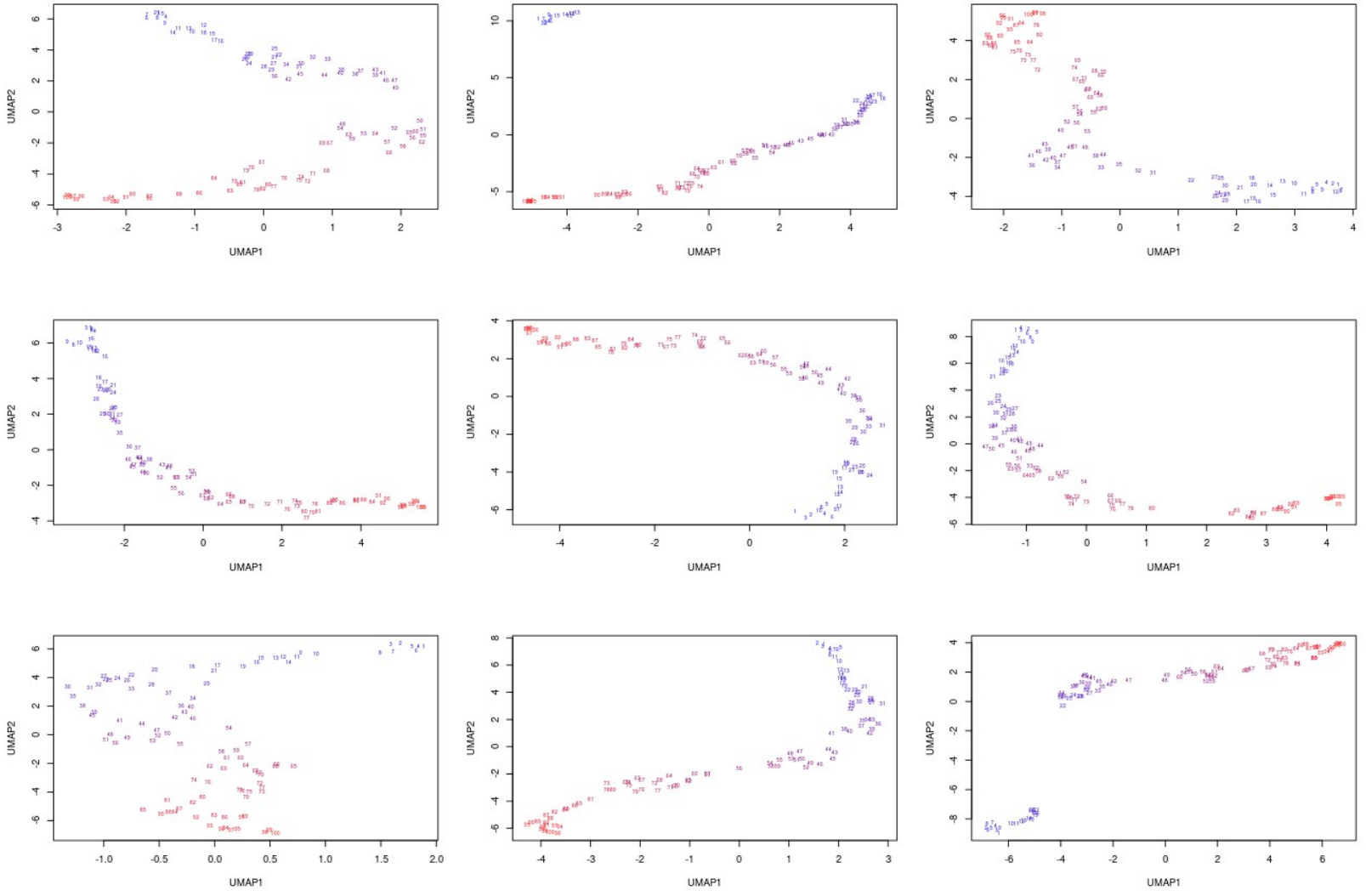




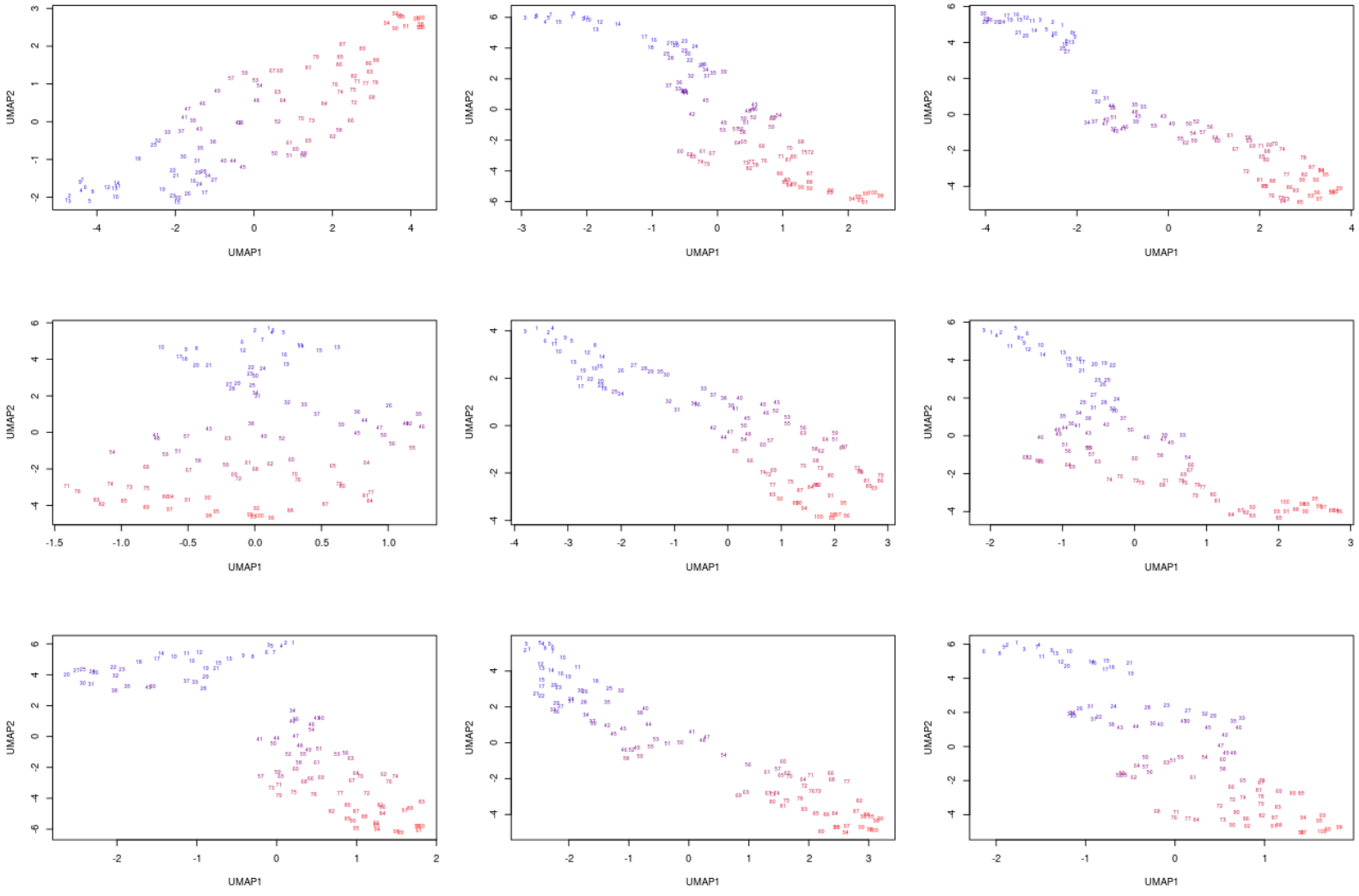



```
Sigma <- matrix(.8, 3, 3); diag(Sigma) <- 1
x <- MASS::mvrnorm(100, rep(0,3), Sigma)
custom.settings <- umap.defaults
custom.settings$input <- "dist"
u <- umap(as.matrix(dist(x)), config = custom.settings)
ranks <- rank(rowMeans(x))
colors <- colorRampPalette(c("blue", "red"))(nrow(x))
colormap <- colors[ranks]
plot(u$layout[,1], u$layout[,2], type = "n", xlab = "dim1", ylab = "dim2")
text(u$layout[,1], u$layout[,2], labels = ranks, col = colormap, cex = 0.5
```

UMAP on matrix of pairwise distances (N=100)

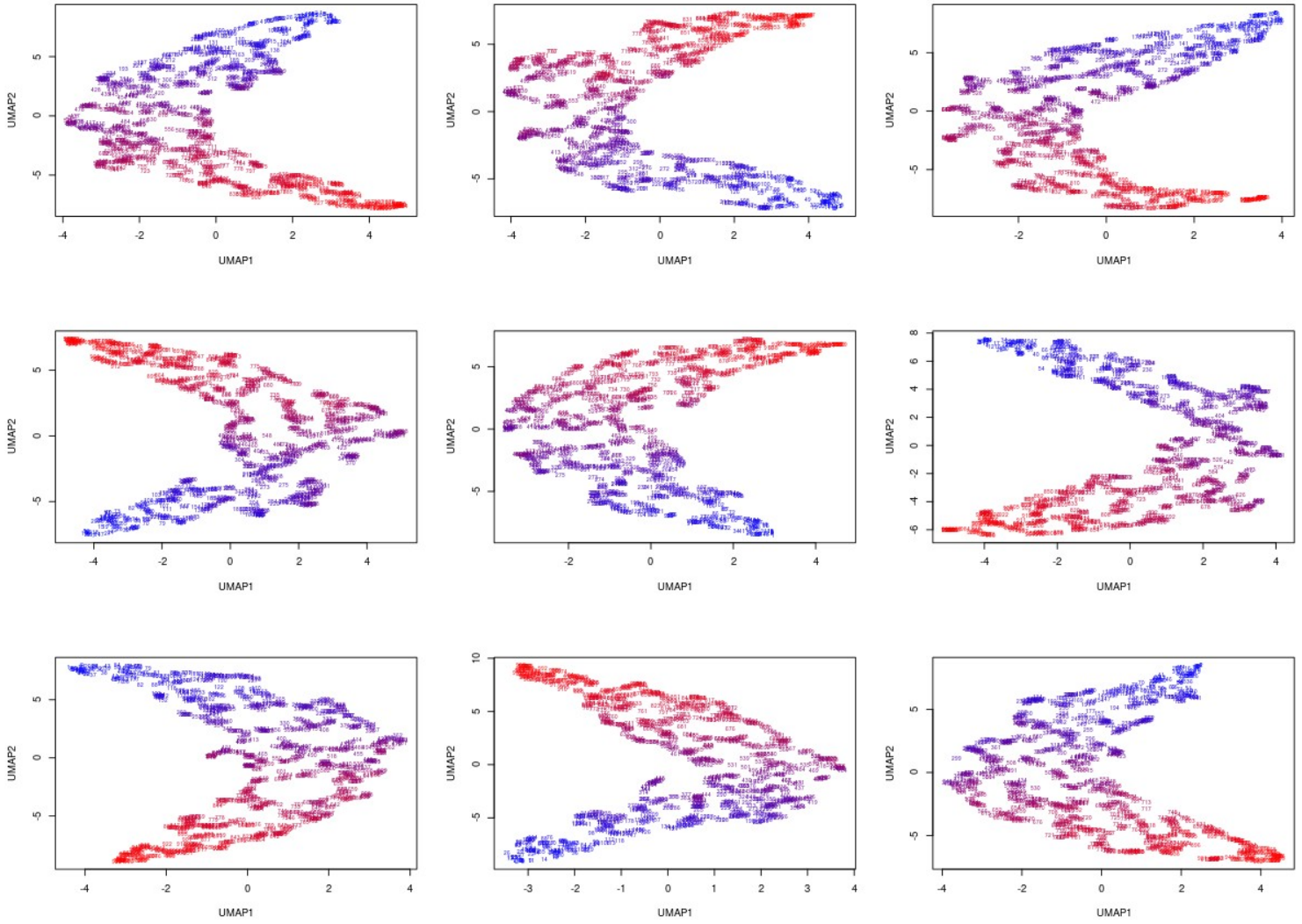


UMAP on raw data matrix (N=100)

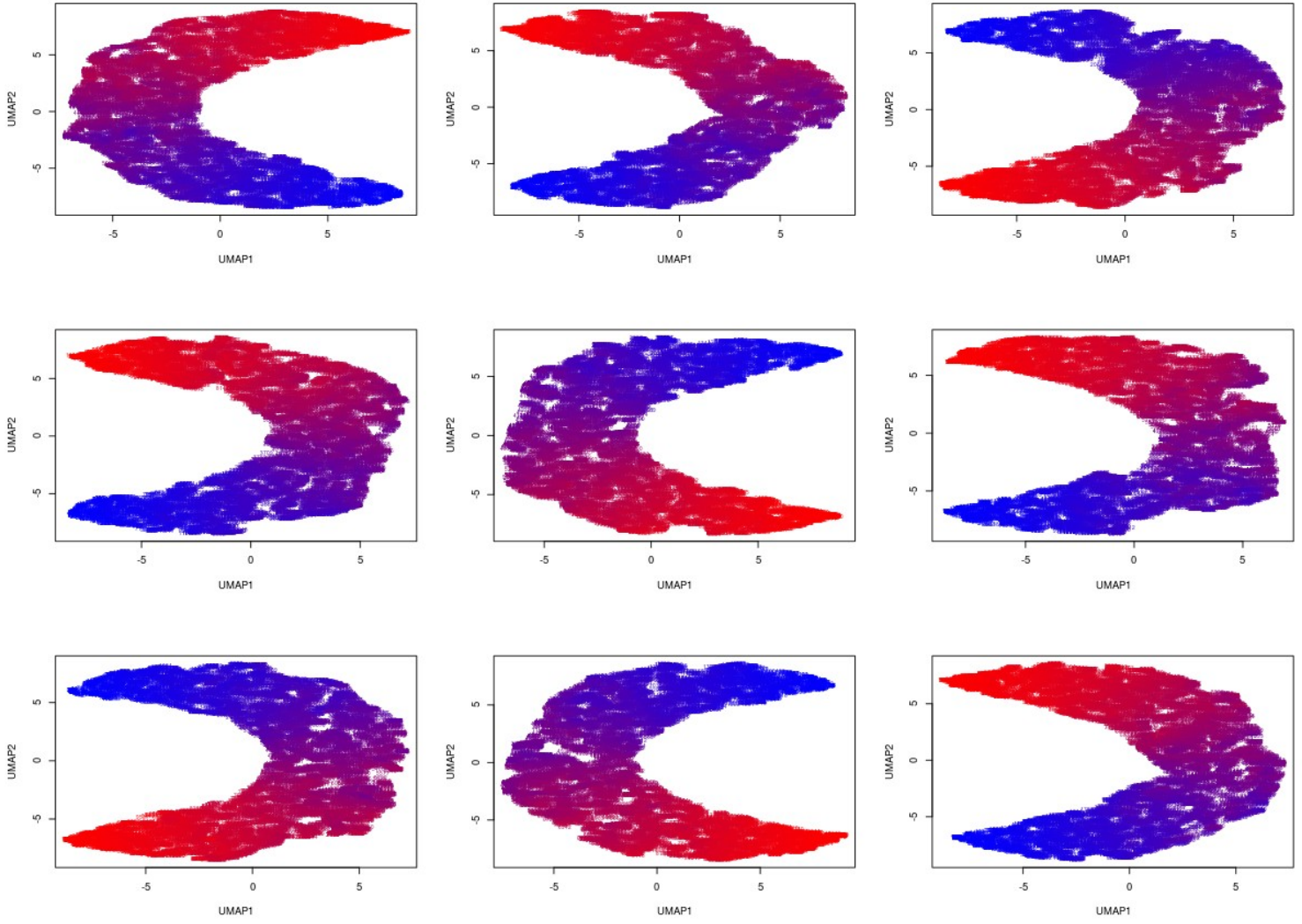


Is N=100 OK for statistics?

UMAP on raw data matrix (N=1000)



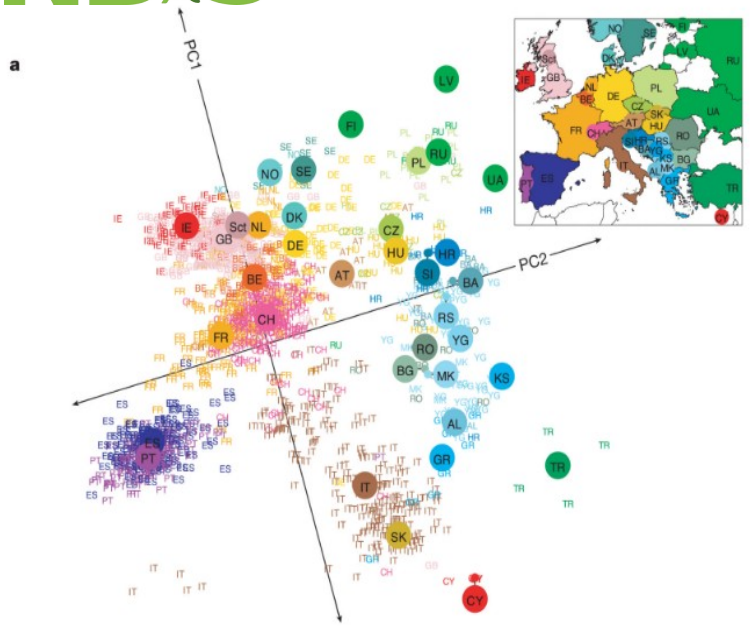
UMAP on raw data matrix (N=10 000)

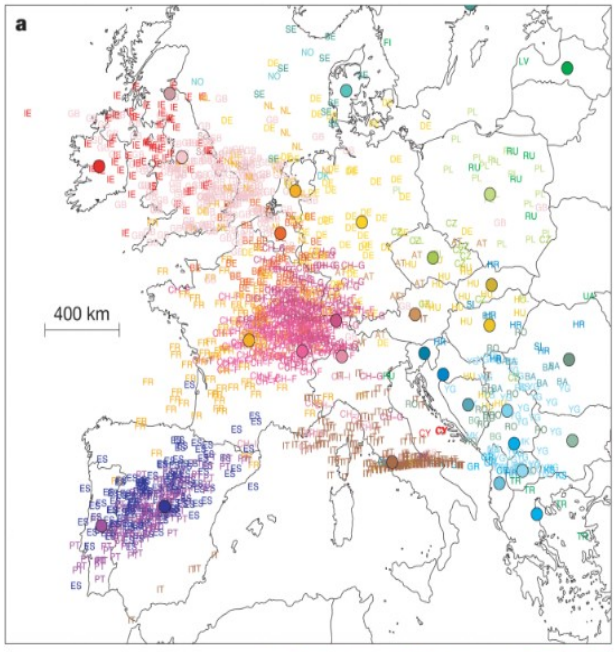


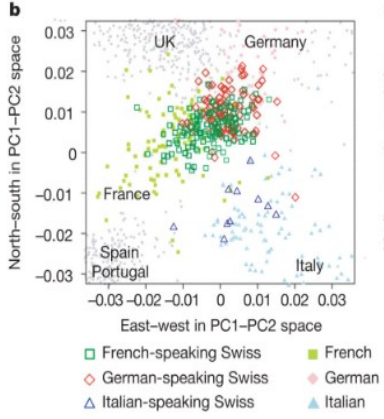
Going through the post of Rafael Irizarry:

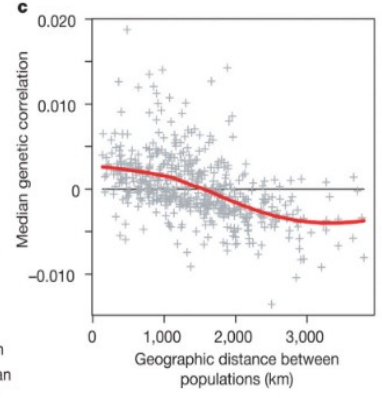
Point2: PCA better than UMAP for PopGen!

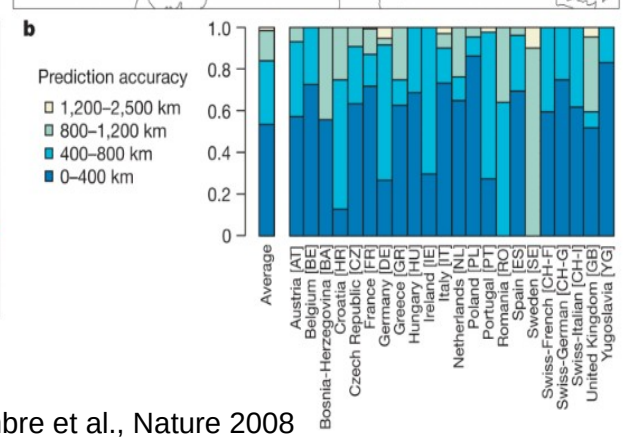
PCA gold standard in PopGen: still criticized











Novembre et al., Nature 2008

www.nature.com/scientificreports

scientific reports

OPEN

Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated

Eran Elhaik

Principal Component Analysis (PCA) is a multivariate analysis that reduces the complexity of datasets while preserving data covariance. The outcome can be visualized on colorful scatterplots, ideally with only a minimal loss of information. PCA applications, implemented in well-cited packages like EIGENSOFT and PLINK, are extensively used as the foremost analyses in population genetics and related fields (e.g., animal and plant or medical genetics). PCA outcomes are used to shape study design, identify, and characterize individuals and populations, and draw historical and ethnobiological conclusions on origins, evolution, dispersion, and relatedness. The replicability crisis in science has prompted us to evaluate whether PCA results are reliable, robust, and replicable. We analyzed twelve common test cases using an intuitive color-based model alongside human population data. We demonstrate that PCA results can be artifacts of the data and can be easily manipulated to generate desired outcomes. PCA adjustment also yielded unfavorable outcomes in association studies. PCA results may not be reliable, robust, or replicable as the field assumes. Our findings raise concerns about the validity of results reported in the population genetics literature and related fields that place a disproportionate reliance upon PCA outcomes and the insights derived from them. We conclude that PCA may have a biasing role in genetic investigations and that 32,000-216,000 genetic studies should be reevaluated. An alternative mixed-admixture population genetic model is discussed.

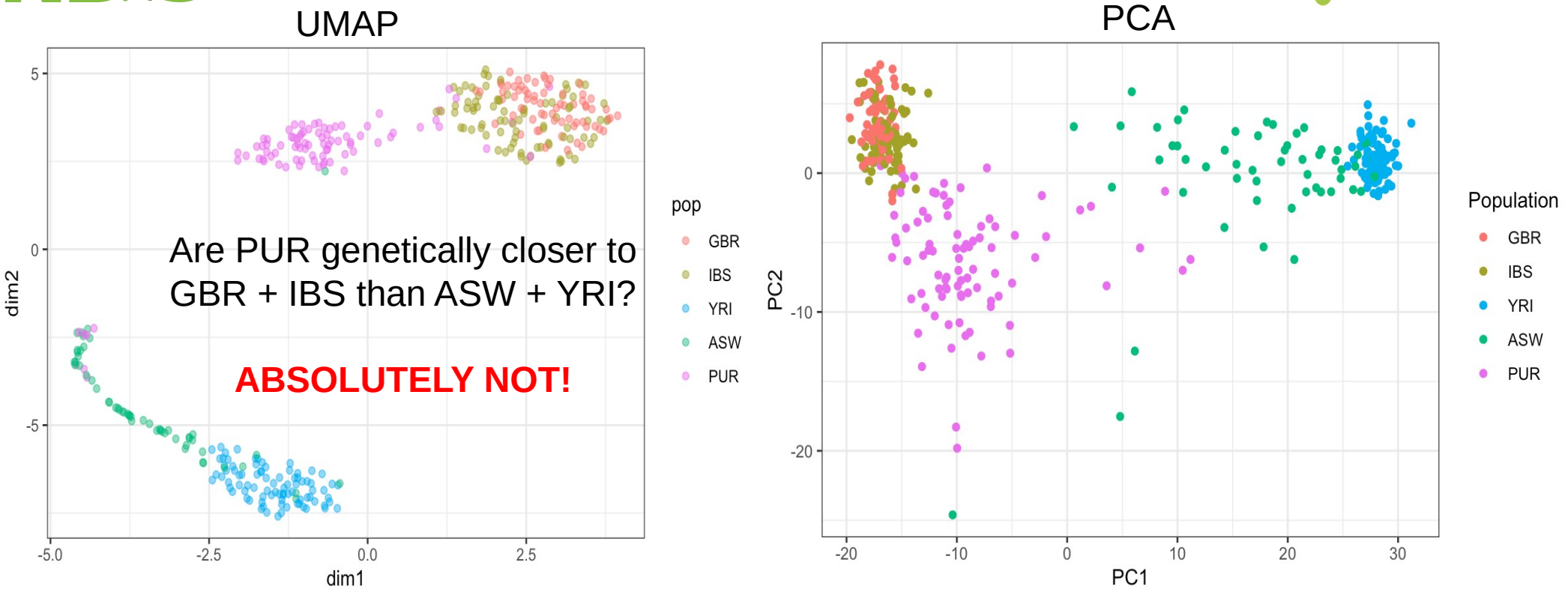
The ongoing reproducibility crisis, undermining the foundation of science[1], raises various concerns ranging from study design to statistical rigor[2,3]. Population genetics is confounded by its utilization of small sample sizes, ignorance of effect sizes, and adoption of questionable study designs. The field is relatively small and may involve financial interests[4–6] and ethical dilemmas[7,8]. Since biases in the field rapidly propagate to related disciplines like medical genetics, biogeography, association studies, forensics, and paleogenomics in humans and non-humans alike, it is imperative to ask whether and to what extent our most elementary tools satisfy risk criteria.

Principal Component Analysis (PCA) is a multivariate analysis that reduces the data's dimensionality while preserving their covariance. When applied to genotype bi-allelic data, typically encoded as AA, AB, and BB, PCA finds the eigenvalues and eigenvectors of the covariance matrix of allele frequencies. The data are reduced to a small number of dimensions termed principal components (PCs); each describes a decreased proportion of the genomic variation. Genotypes are then projected onto space spanned by the PC axes, which allows visualizing the samples and their distances from one another in a colorful scatter plot. In this visualization, sample overlap is considered evidence of identity, due to common origin or ancestry[9,10]. PCA's most attractive property for population geneticists is that the distances between clusters allegedly reflect the genetic and geographic distances between them. PCA also supports the projection of points onto the components calculated by a different dataset, presumably accounting for insufficient data in the projected dataset. Initially adapted for human genomic data in 1963[11], the popularity of PCA has slowly increased over time. It was not until the release of the SmartPCA tool (EIGENSOFT package)[10] that PCA was propelled to the front stage of population genetics.

PCA is used as the first analysis of data investigation and data description in most population genetic analyses, e.g., Refs.[12–15]. It has a wide range of applications. It is used to examine the population structure of a cohort or individuals to determine ancestry, analyze the demographic history and admixture, decide on the genetic similarity of samples and exclude outliers, decide how to model the populations in downstream analyses, describe the ancient and modern genetic relationships between the samples, infer kinship, identify ancestral clines in the data, e.g., Refs.[16–19], detect genomic signatures of natural selection, e.g., Ref.[20] and identify convergent evolution[21].

Department of Biology, Lund University, 22362 Lund, Sweden. email: eran.elhaik@biol.lu.se

HapMap3: UMAP vs PCA (Rafael Irizarry's post)

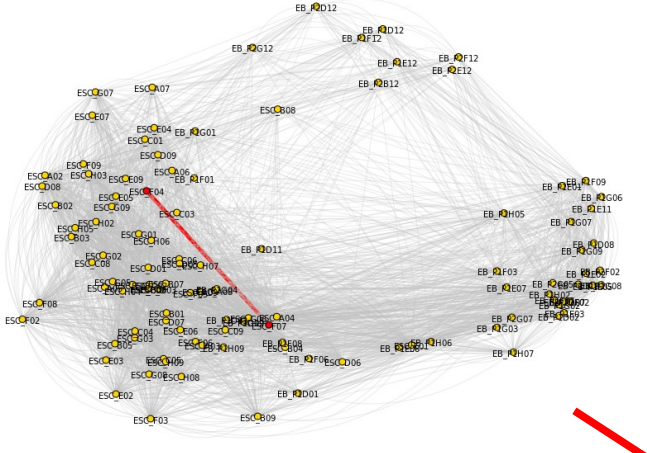


- Because of their meaningless inter-cluster distances tSNE / UMAP are less useful for population genomics than PCA
- The goal of tSNE / UMAP is to **discover clusters**, which is sufficient for Single Cell Biology but not for PopGen.
- In PopGen we generally do not discover clusters, we have an idea about e.g. human populations, and the aim is often to explore the **genetic relatedness** between the populations, a task UMAP can absolutely not solve!

UMAP for data integration

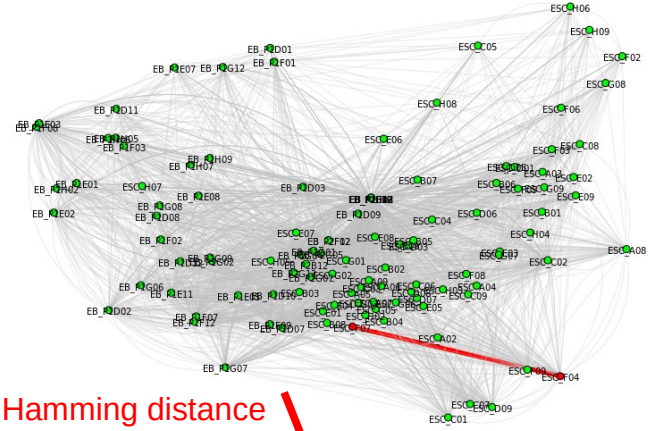
Graph Intersection Method

scRNAseq KNN Graph



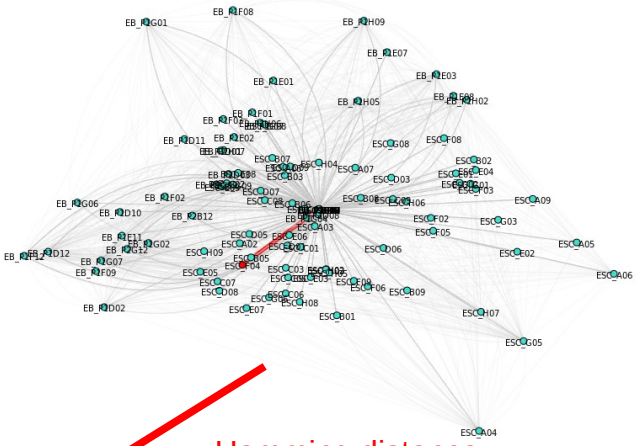
Euclidean distance

scBSseq KNN Graph



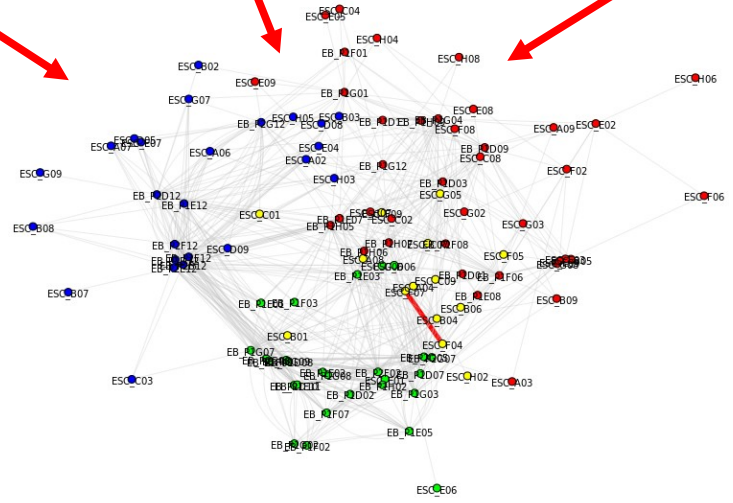
Hamming distance

scATACseq KNN Graph



Hamming distance

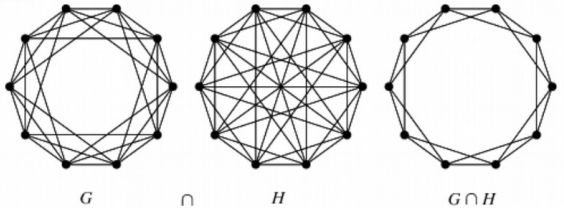
Consensus Graph



Keep edges consistently present across the Omics

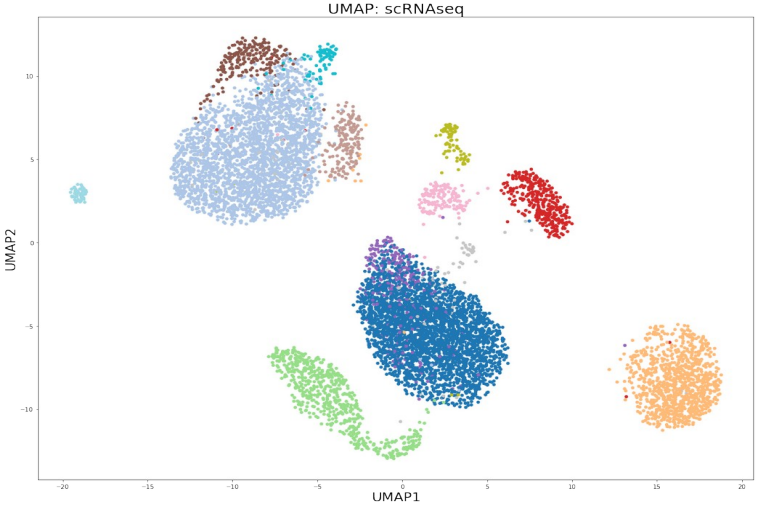
Graph Intersection

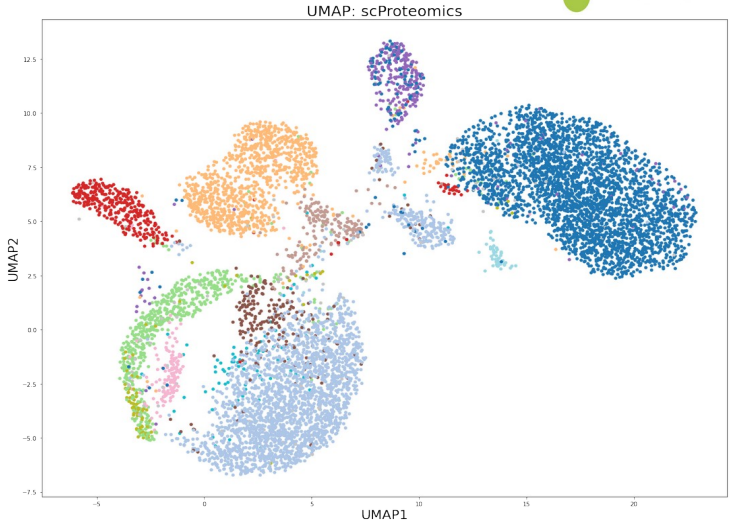
DOWNLOAD Wolfram Notebook

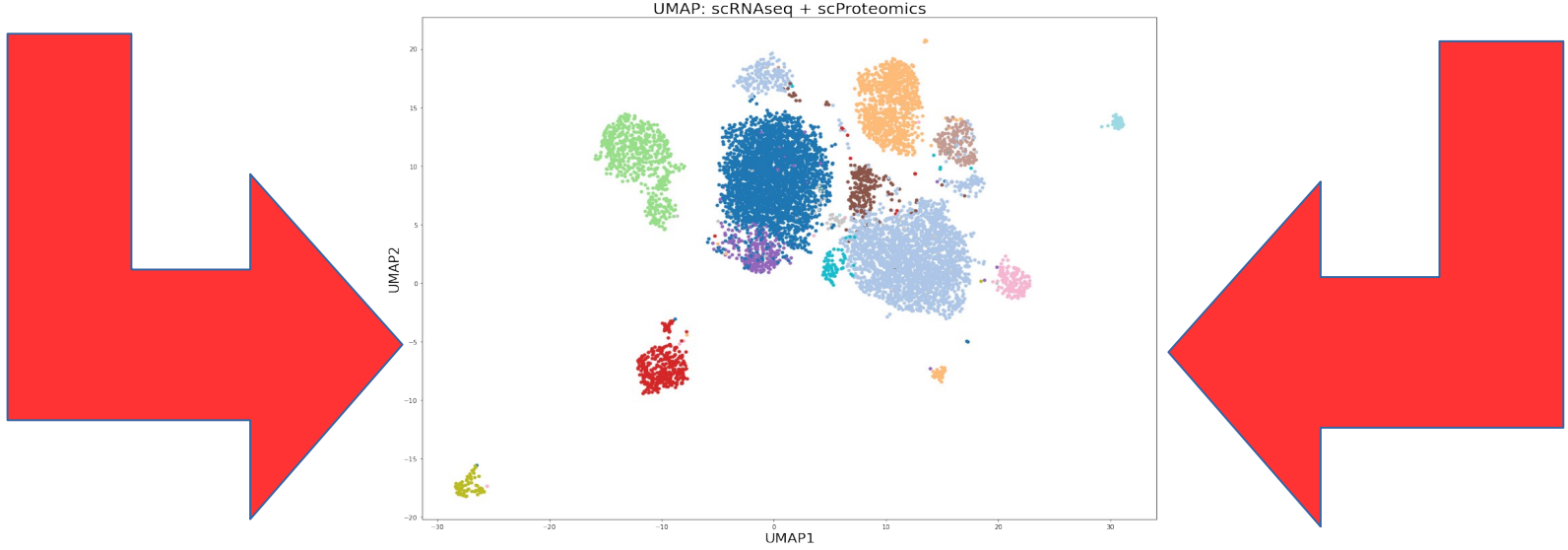


Let S be a set and $F = \{S_1, \ldots, S_p\}$ a nonempty family of distinct nonempty subsets of S whose union is $\bigcup_{i=1}^{p} S_i = S$. The intersection graph of F is denoted $\Omega(F)$ and defined by $V(\Omega(F)) = F$, with S_i and S_j adjacent whenever $i \neq j$ and $S_i \cap S_j \neq \emptyset$. Then a graph G is an intersection graph on S if there exists a family F of subsets for which G and $\Omega(F)$ are isomorphic graphs (Harary 1994, p. 19). Graph intersections can be computed in the Wolfram Language using `GraphIntersection[g, h]`.

UMAP for data integration: CITEseq







National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

LUNDS UNIVERSITET