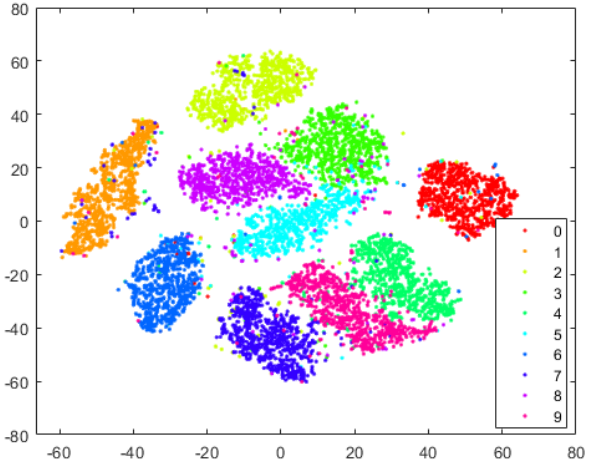
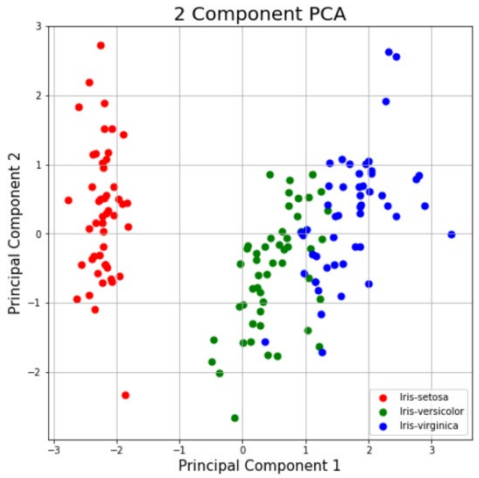
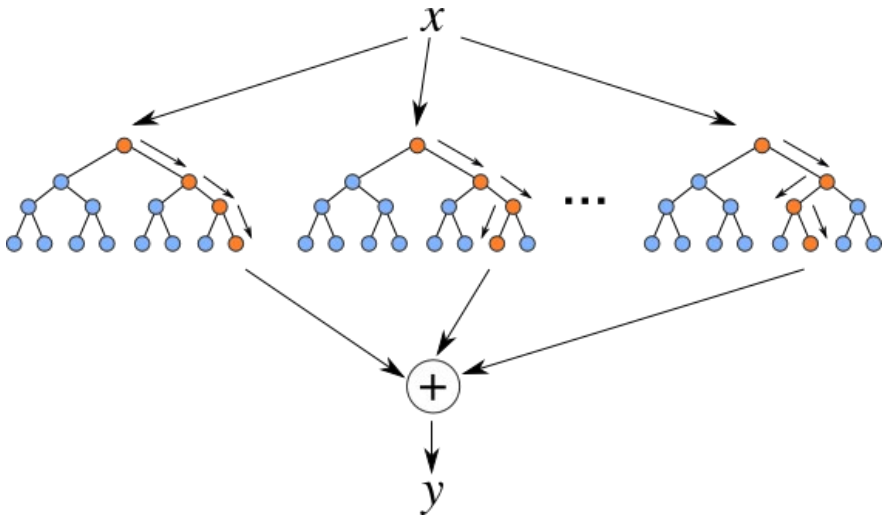
Less is more: An Intro to Dimensionality Reduction

Ron Rotkopf
LSCF, Weizmann Institute of Science







What is Dimensionality Reduction?

- Problem: many variables measured on the same sample.
- Some of these variables might be correlated.
- Per-variable tests might be redundant, or might miss important differences.
- Feature selection: Using a smaller set of variables to model the data.
- Feature extraction: This reduces the data in a high dimensional space to a lower dimension space.
- Find the “latent” features in your data.

Different Scenarios

- Few samples, defined groups/treatments (e.g. bulk mRNA-seq)
- Lots of samples (cells), exploratory analysis (single-cell)
- What are we asking?

Variable selection approaches

- Missing value ratio: remove features with lots of missing data.
- Low variance: remove features with low variance.
- Correlated variables: For a pair of highly correlated variables, one may be removed.
- Removal based on group test (keep only differing features).

Common Methods

- PCA (Principal Component Analysis)
- LDA (Linear Discriminant Analysis)
- GDA (Generalized Discriminant Analysis)
- t-SNE (t-Distributed Stochastic Neighbor Embedding)
- UMAP (Uniform Manifold Approximation and Projection)

PCA

- Original data – a set of (possibly) partially correlated variables
- Output – a new set of uncorrelated variables, each of which is a linear combination of the original variables.
- $PC1 = a_1x_1 + a_2x_2 + a_3x_3 + \ldots$
- $PC2 = b_1x_1 + b_2x_2 + b_3x_3 + \ldots$

PCA

- The first new variable (or principal component) accounts for as much of the variation in the original data amongst all linear combinations of the original variables.
- The next principal component accounts for as much as possible of the remaining variation, and so on.
- Factor loadings are the correlation coefficients between the variables and factors (components).

PCA

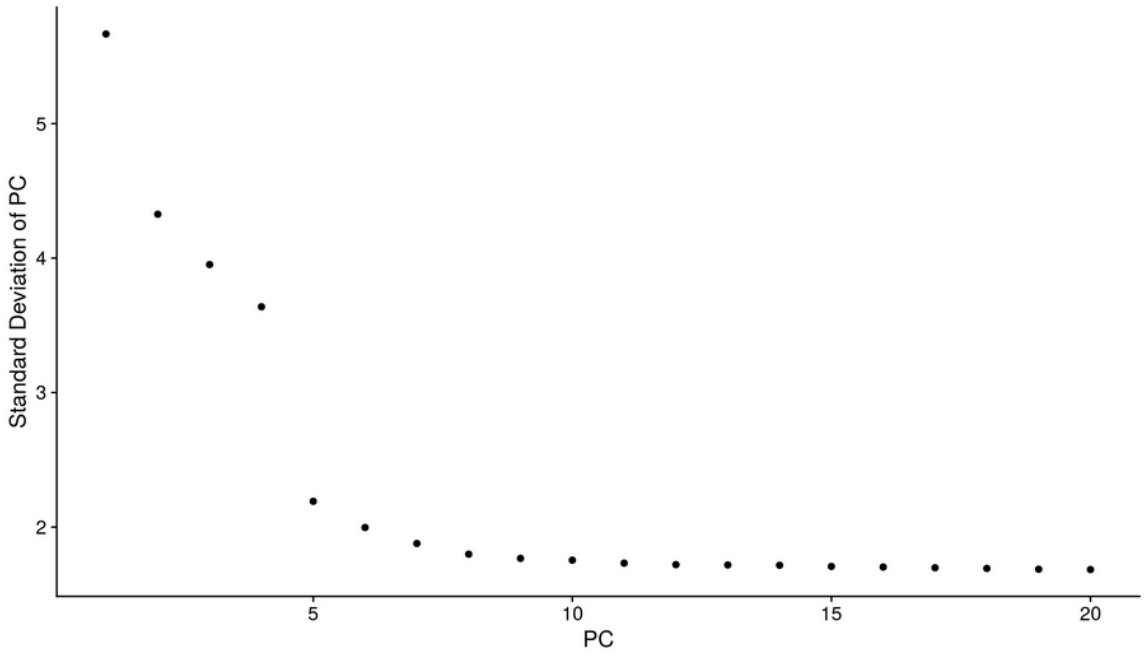
- Hopefully ,the first 2 or 3 PCs explain a large part of the variance.
- Always present the %variance on the PC axes.
- PCA doesn’t care about groups, and does not try to differentiate between treatments! It only cares about correlation and variance.

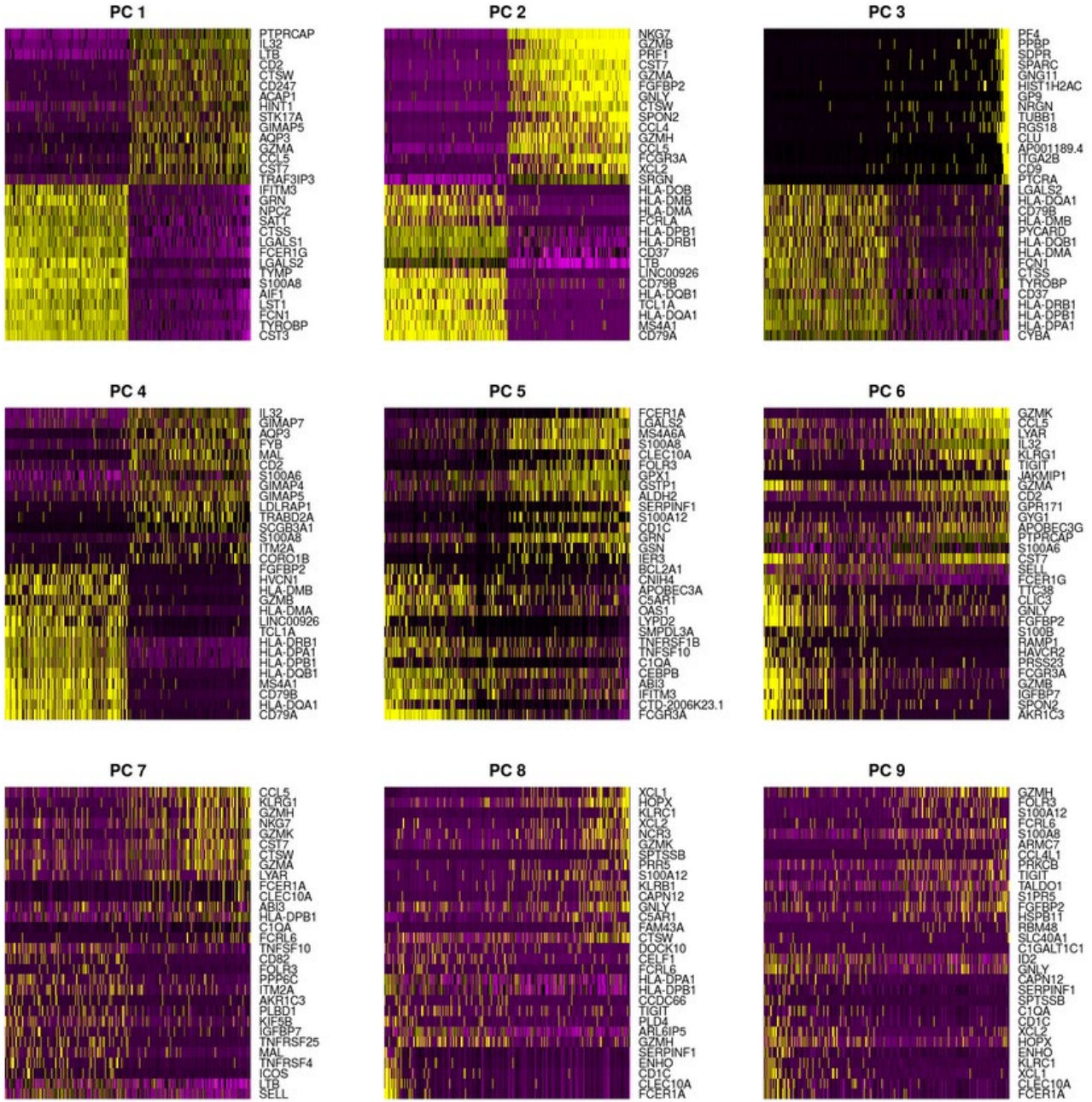
Example in R

PCA - Disadvantages

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which may not always exist.
- We may not know how many principal components to keep. In practice, some thumb rules are applied.
- The large distances between far-away points may distort our view.

Example from Seurat





LDA

- Similar to PCA, but supervised.

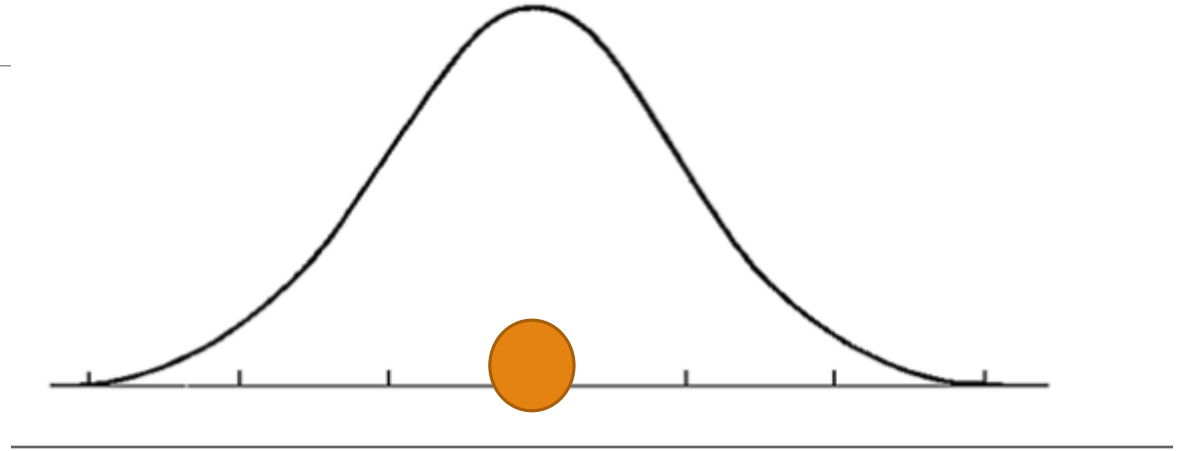
- Looks for combinations of variables that best explain the difference between groups.

t-SNE – t-Distributed Stochastic Neighbor Embedding

- Non-linear dimensionality reduction algorithm
- Based on probability distributions with random walk on neighborhood graphs to find the structure within the data.
- Local approaches seek to map nearby points on the manifold to nearby points in the low-dimensional representation.
- Global approaches attempt to preserve geometry at all scales, i.e mapping nearby points to nearby points and far away points to far away points.
- t-SNE tries to do both.

t-SNE

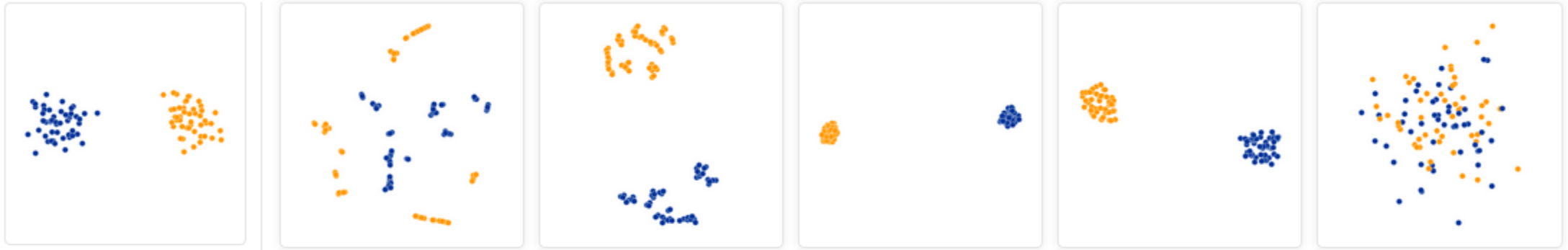
- Euclidean distances are converted to probabilities (similarities).
- Each point is considered as the center of a Gaussian distribution.
- t-SNE aims to get a 2D representation with similarities as close as possible to the high-dimensional data.



Connections between points can be envisioned as springs whose stiffness depends on the mismatch between the similarity of the two data points and the similarity of the two map points.

https://d3ansictanv2wj.cloudfront.net/rossant-f06184034ba66a0bd06a-001.html

Perplexity



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

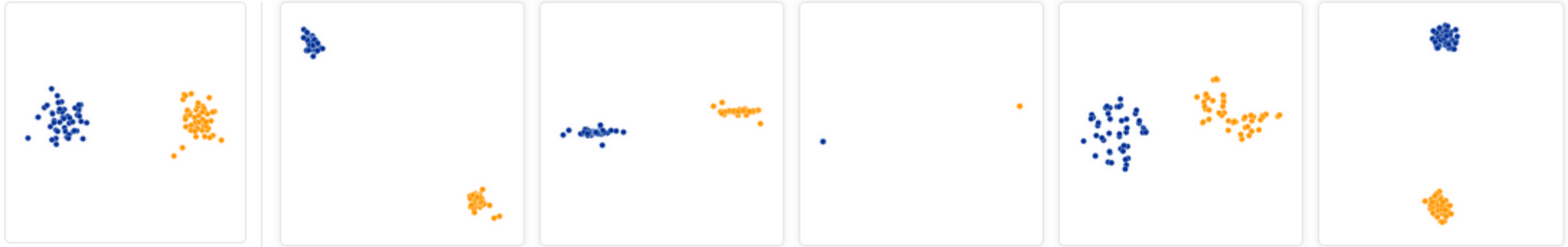
Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

https://distill.pub//enst-daersim/2016

Iterations



Original

Perplexity: 30
Step: 10

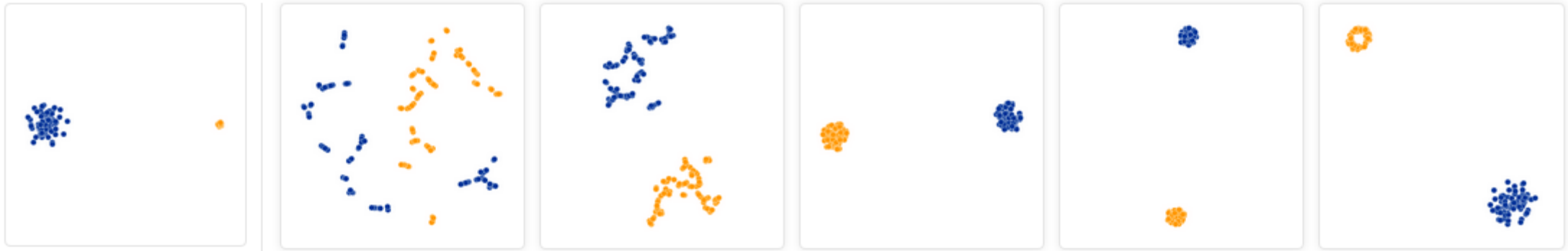
Perplexity: 30
Step: 20

Perplexity: 30
Step: 60

Perplexity: 30
Step: 120

Perplexity: 30
Step: 1,000

Cluster Sizes



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

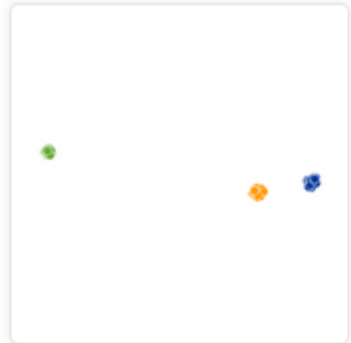
Distances Between Clusters

Original

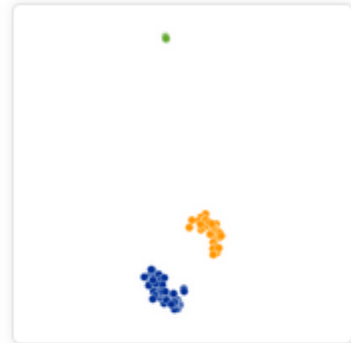
Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

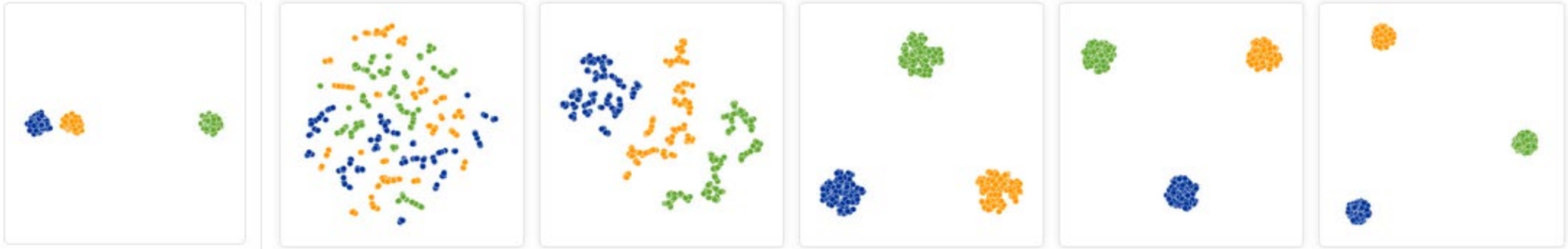


Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

Distances Between Clusters (more points)



Original

Perplexity: 2
Step: 5,000

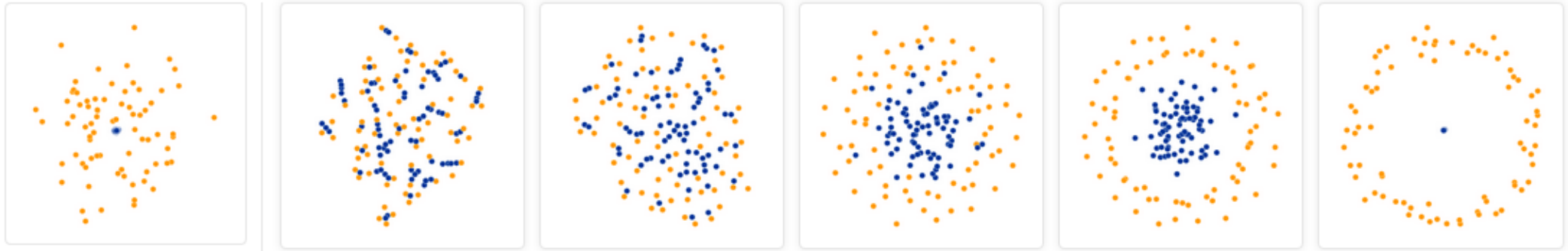
Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

Topology - Containment



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

UMAP - Uniform Manifold Approximation and Projection for Dimension Reduction

Based on ideas from topological data analysis.

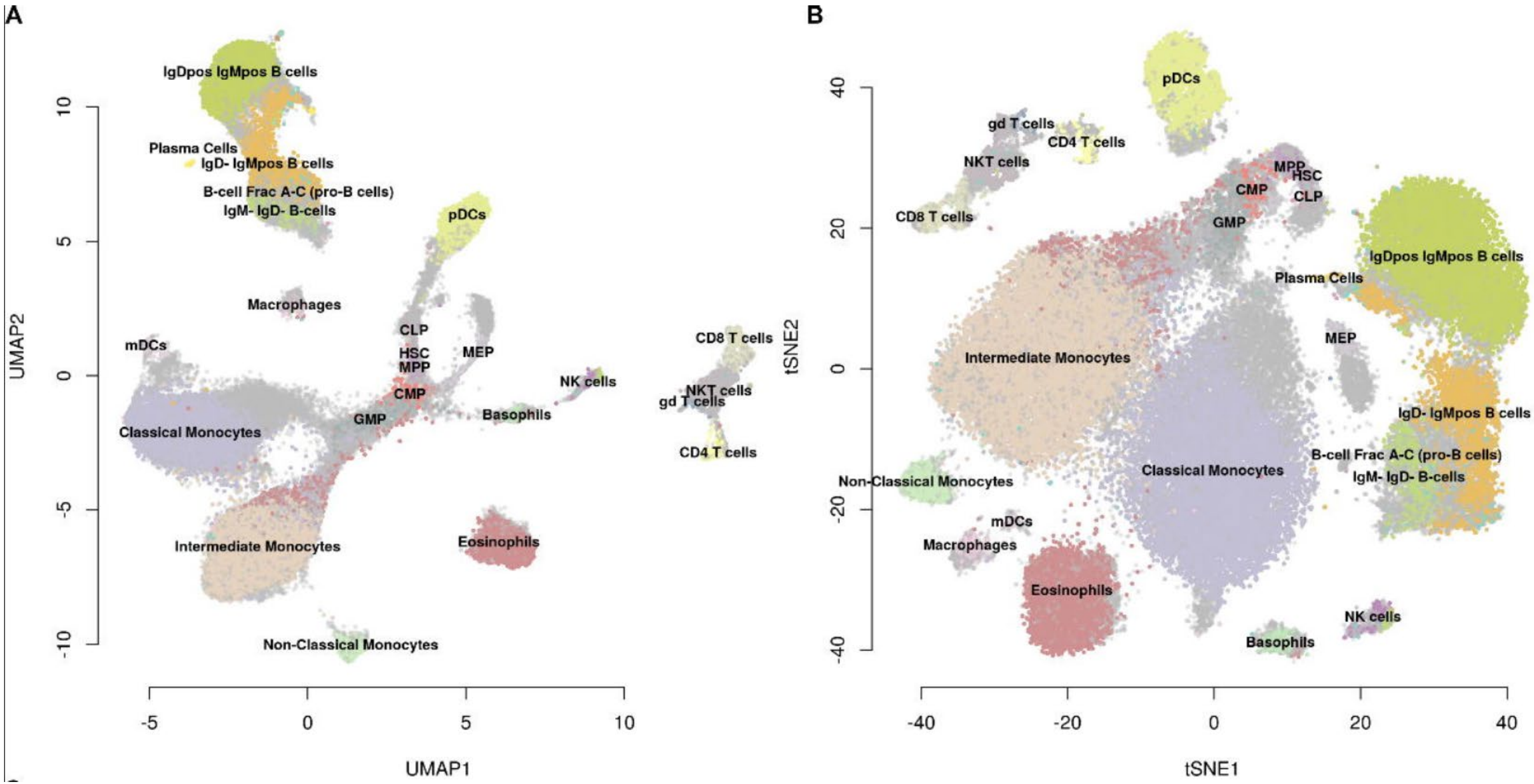
Compared to t-SNE:

- Preserves more of the global data structure (clusters are organized in a meaningful way).
- Shorter runtime.

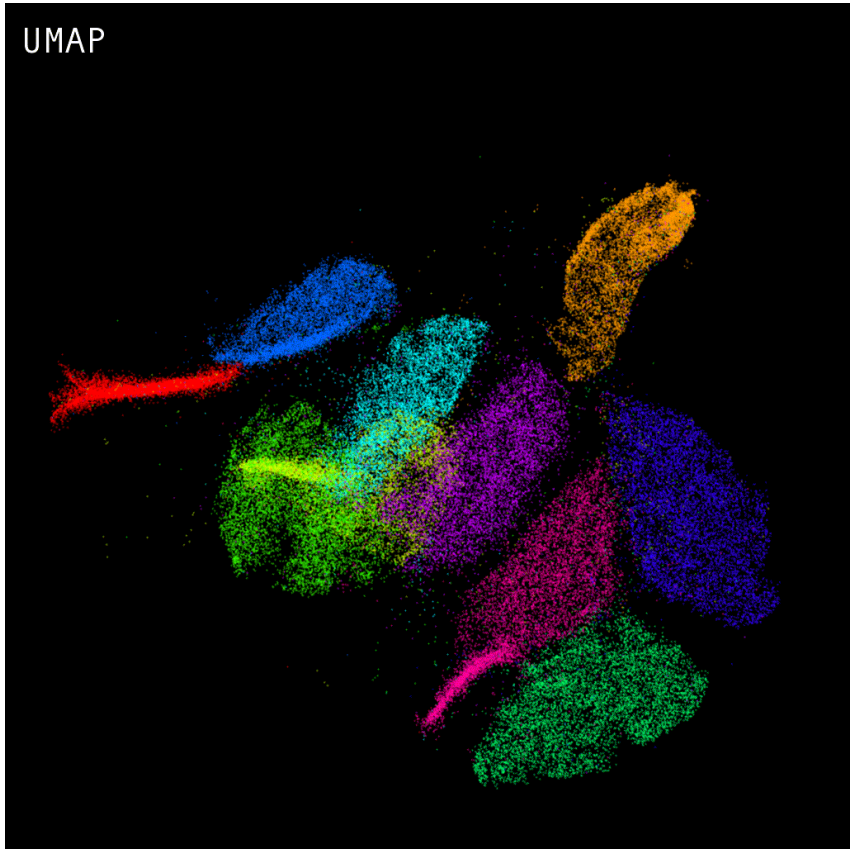
How UMAP works:
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html
Comparison with t-SNE: Becht et al. 2019
https://www.nature.com/articles/nbt.4314.pdf

Time Comparison

	t-SNE	UMAP
COIL20	20 seconds	7 seconds
MNIST	22 minutes	98 seconds
Fashion MNIST	15 minutes	78 seconds
GoogleNews	4.5 hours	14 minutes


A
IgDpos IgMpos B cells
Plasma Cells
IgD- IgMpos B cells
B-cell Frac A-C (pro-B cells)
IgM- IgD- B-cells
pDCs
Macrophages
CLP
HSC
MPP
MEP
mDCs
CMP
GMP
NK cells
Basophils
Classical Monocytes
CD8 T cells
NKT cells
gd T cells
CD4 T cells
Eosinophils
Intermediate Monocytes
Non-Classical Monocytes
UMAP2
UMAP1
10
5
0
-5
-10
-5
0
5
10
B
pDCs
gd T cells
CD4 T cells
NKT cells
MPP
HSC
CMP
CLP
CD8 T cells
GMP
IgDpos IgMpos B cells
Plasma Cells
MEP
Intermediate Monocytes
IgD- IgMpos B cells
B-cell Frac A-C (pro-B cells)
Classical Monocytes
IgM- IgD- B-cells
Non-Classical Monocytes
mDCs
Macrophages
Eosinophils
NK cells
Basophils
tSNE2
tSNE1
40
20
0
-20
-40
-40
-20
0
20
40

UMAP vs. t-SNE



Kylemcdonald.net