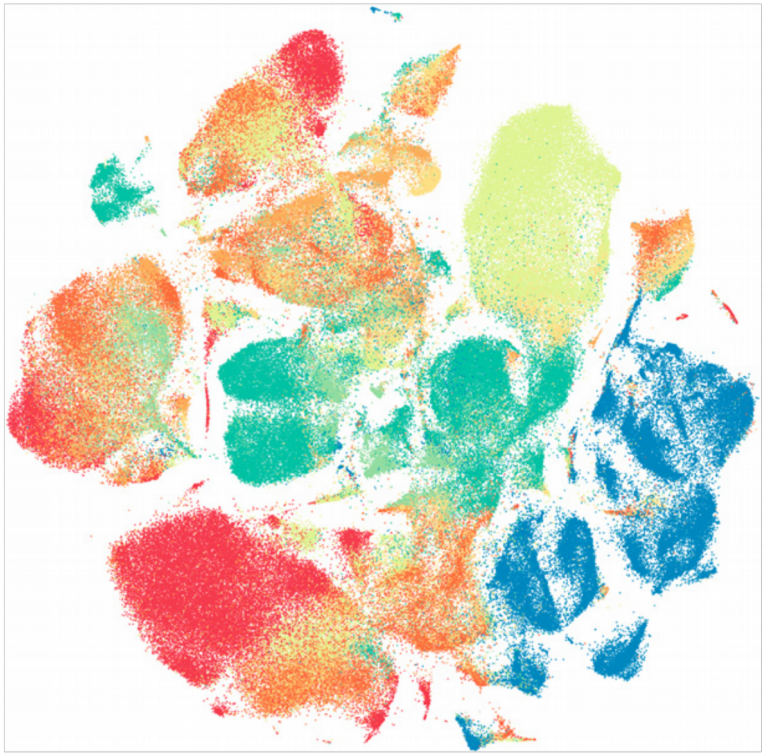
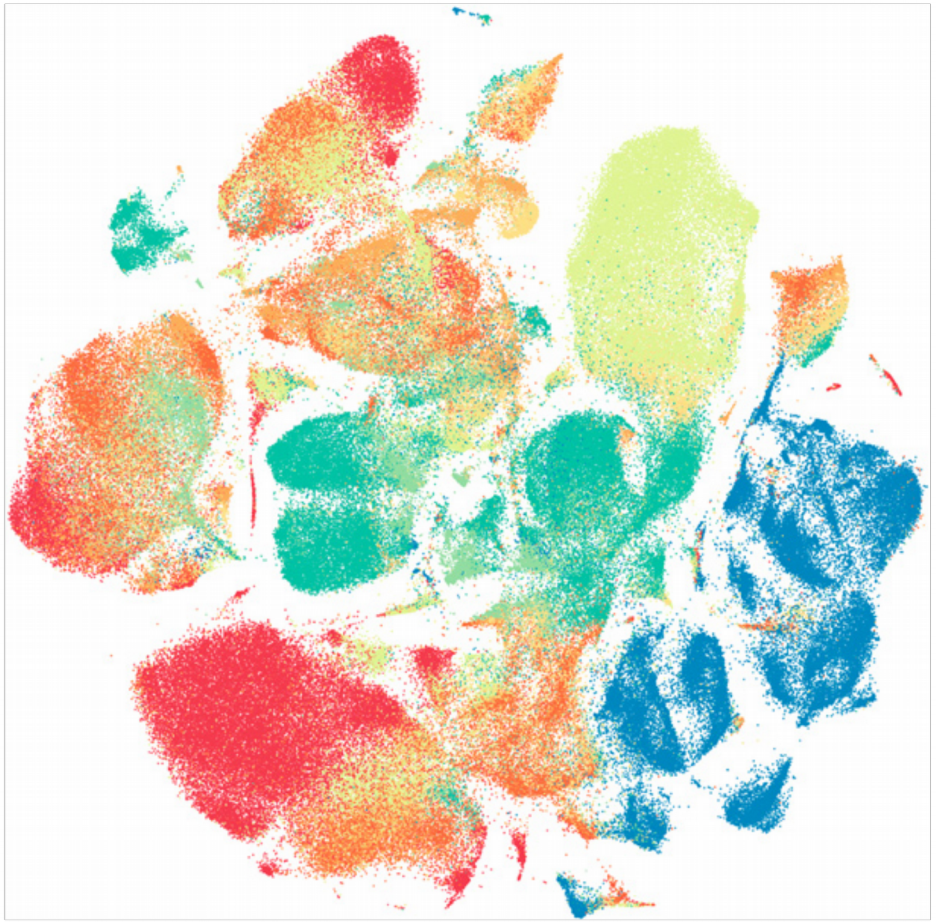


14: The truth about t-SNE

https://github.com/matthiaskoenig/itbtechtalks

Dr Matthias König
Humboldt University Berlin,
Institute for Theoretical Biology



"**t-SNE** has become widespread in the field of machine learning, since it **has an almost magical ability to create compelling two-dimensional "maps"** from data with hundreds or even thousands of dimensions.

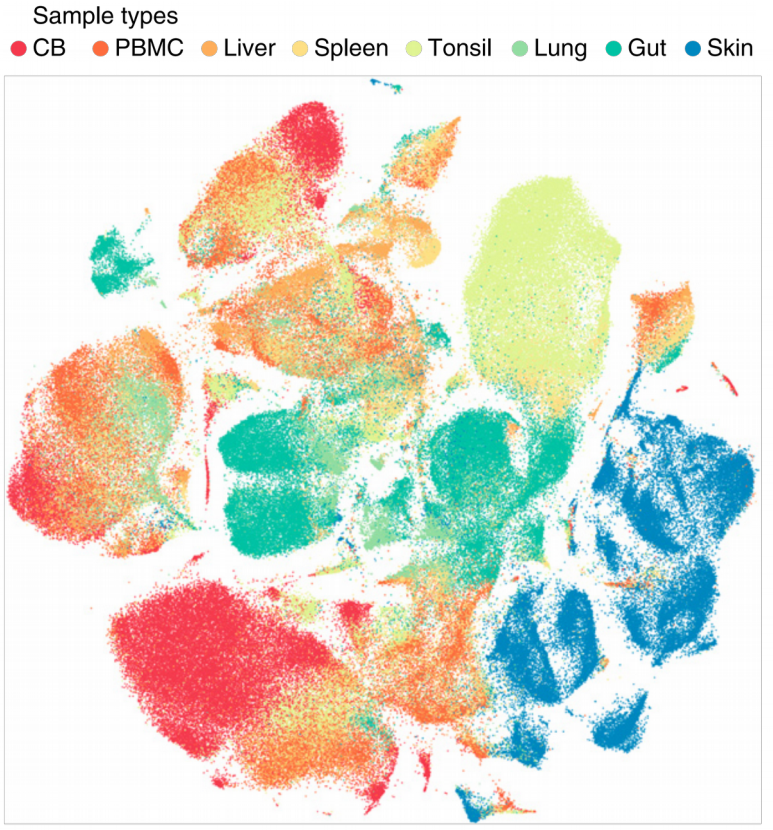
Although impressive, **these images can be tempting to misread.**"

https://distill.pub/2016/misread-tsne/
Dimensionality reduction for visualizing single-cell data using UMAP; Becht 2019; Nat. Biotechnology

What is t-SNE?

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets.

- 2D projection of high-dimensional data which allows to visualize cluster (using local similarities)
- extremely popular in the field of single-cell and comparative omics
- models each high-dimensional object by a two- or three-dimensional point in such a way that **similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability**
- The **algorithm is non-linear and adapts to the underlying data**, performing **different transformations on different regions**.
- implementations in **Matlab**, **C++**, **CUDA**, **Python**, **Torch**, **R**, **Julia**, and **JavaScript**
- **Although extremely useful** for visualizing high-dimensional data, t-SNE plots **can sometimes be mysterious or misleading**.



https://distill.pub/2016/misread-tsne/
Dimensionality reduction for visualizing single-cell data using UMAP; Becht 2019; Nat. Biotechnology

1.1Those hyperparameters really matter

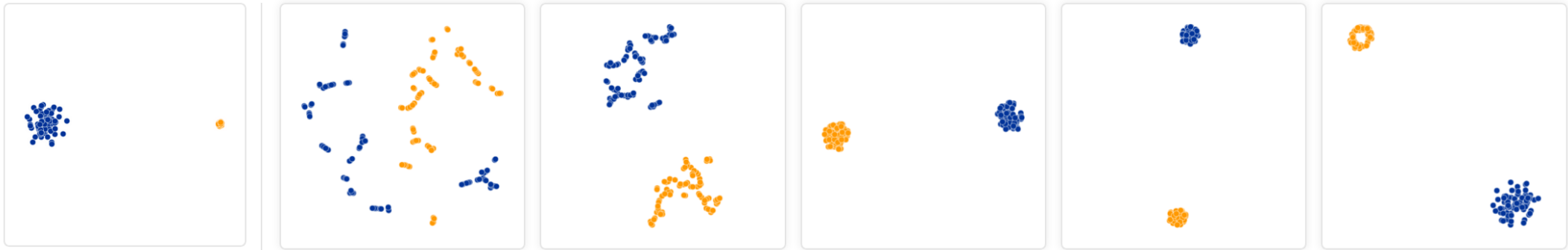
- **hello world of t-SNE**: a data set of two widely separated clusters
- with suggested **perplexity values (5-50)** the diagram shows **clusters**, although **with very different shapes**
- with low perplexity 2, local variations dominate
- perplexity 100 pitfall: for algorithm to operate properly perplexity < # data points

https://distill.pub/2016/misread-tsne/

1.2Those hyperparameters really matter

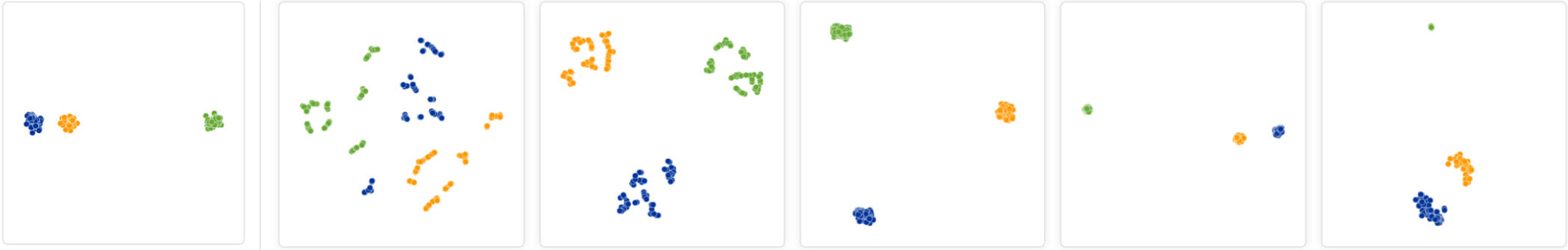
- Each of the plots before was made with 5,000 iterations with a **learning rate** (often called “epsilon”) of 10, and had reached a point of **stability** by step 5,000.
- If you see a t-SNE plot with **strange “pinched” shapes**, chances are the process was stopped too early. Unfortunately, there’s no fixed number of steps that yields a stable result. **Different data sets** can **require different numbers of iterations** to converge.

2 Cluster sizes in a t-SNE plot mean nothing



- Surprisingly, the two clusters look about same size in the t-SNE plots. What's going on?
- The t-SNE algorithm **adapts its notion of "distance" to regional density variations** in the data set. As a result, it naturally **expands dense clusters**, and **contracts sparse ones**, evening out cluster sizes.

3 Distances between clusters might not mean anything



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

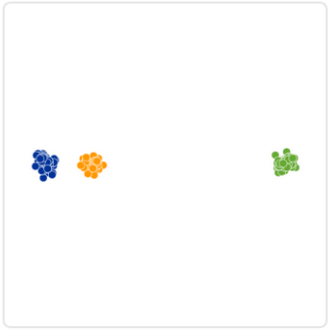
Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

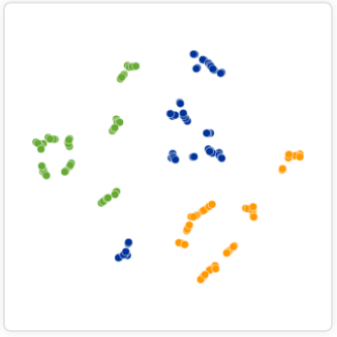
Perplexity: 100
Step: 5,000

- At **perplexity 50**, the diagram gives a **good sense of the global geometry**.
- For lower perplexity values the clusters look equidistant. When the perplexity is 100, we see the global geometry fine, but one of the cluster appears, falsely, much smaller than the others.
- So is perplexity 50 a good value for global geometry?

3 Distances between clusters might not mean anything

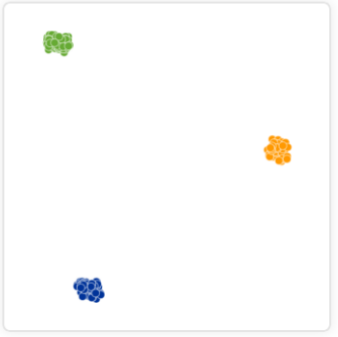


Original



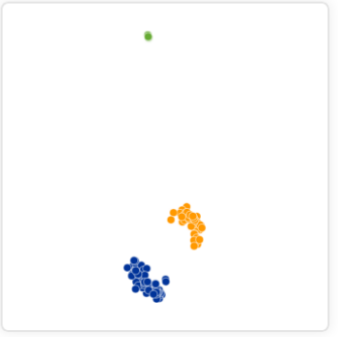
Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000



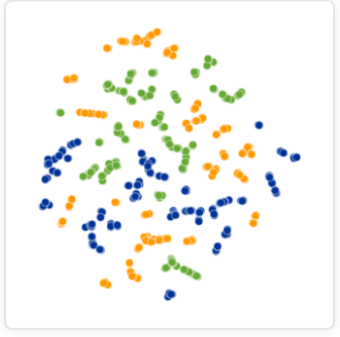
Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

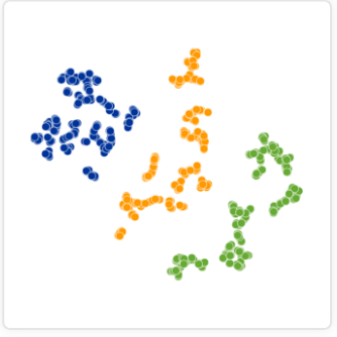


Perplexity: 100
Step: 5,000

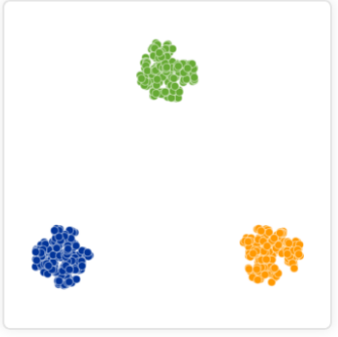
Original



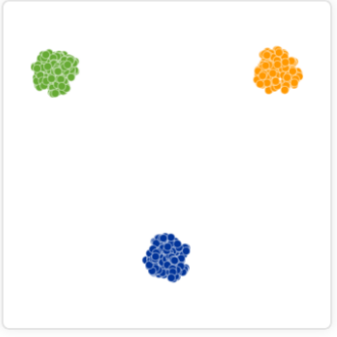
Perplexity: 2
Step: 5,000



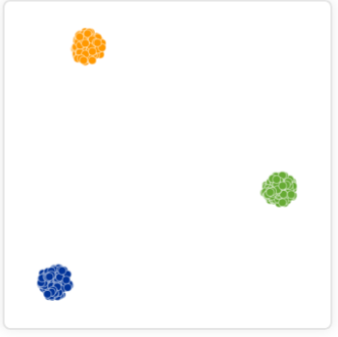
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000

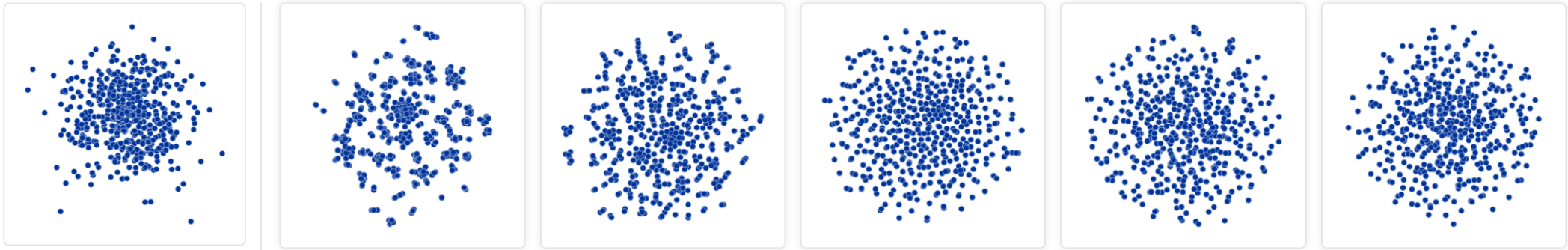


Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

4 Random noise doesn't always look random



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

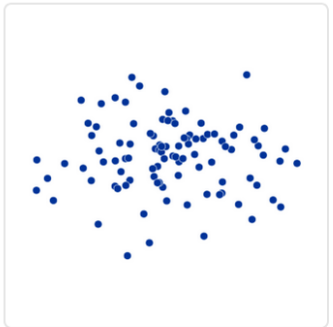
Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

- 500 points drawn from a unit Gaussian distribution in 100 dimensions
- A classic pitfall is thinking you see patterns in what is really just random data. Recognizing noise when you see it is a critical skill, but it takes time to build up the right intuitions. A tricky thing about t-SNE is that it throws a lot of existing intuition out the window.
- **those "clumps" aren't meaningful**. If you look back at previous examples, low perplexity values often lead to this kind of distribution. **Recognizing** these **clumps as random noise** is an important part of reading t-SNE plots.

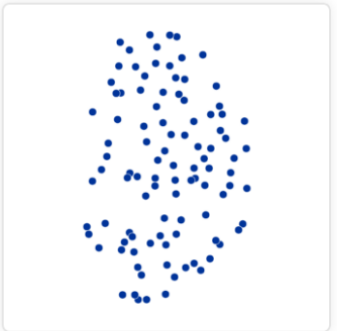
5 You can see some shapes, sometimes



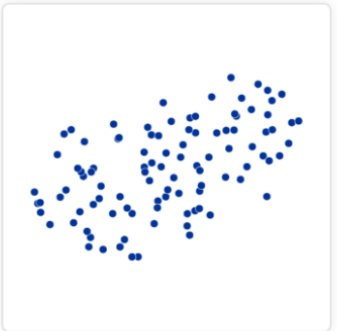
Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

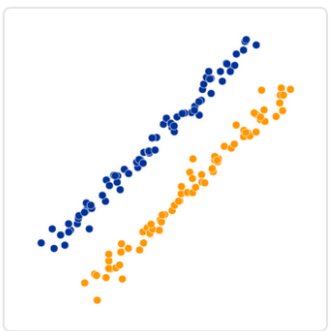


Perplexity: 30
Step: 5,000

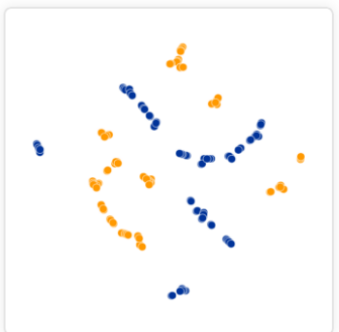


Perplexity: 50
Step: 5,000

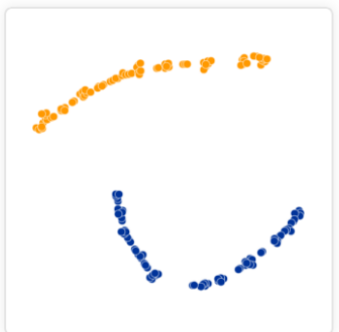
Perplexity: 100
Step: 5,000



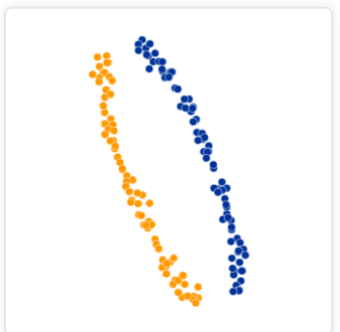
Original



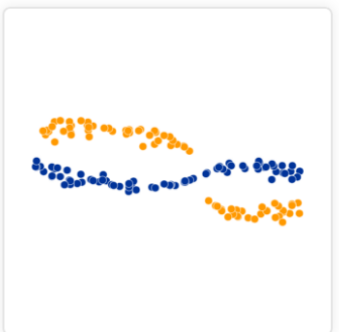
Perplexity: 2
Step: 5,000



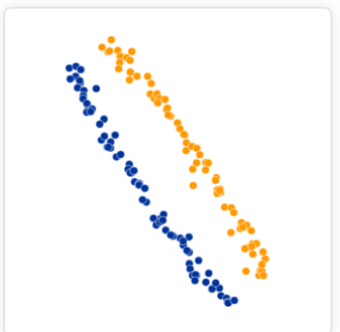
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



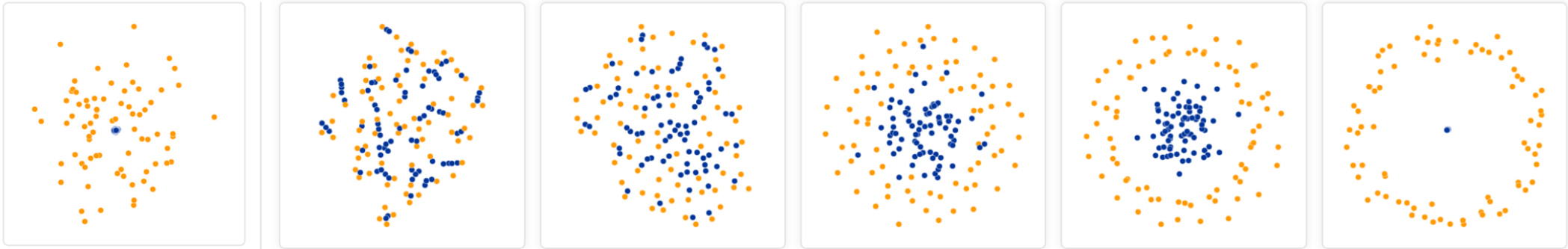
Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

- **at low perplexity, local effects and meaningless “clumping” take center stage**. More extreme shapes also come through, but again only at the right perplexity.

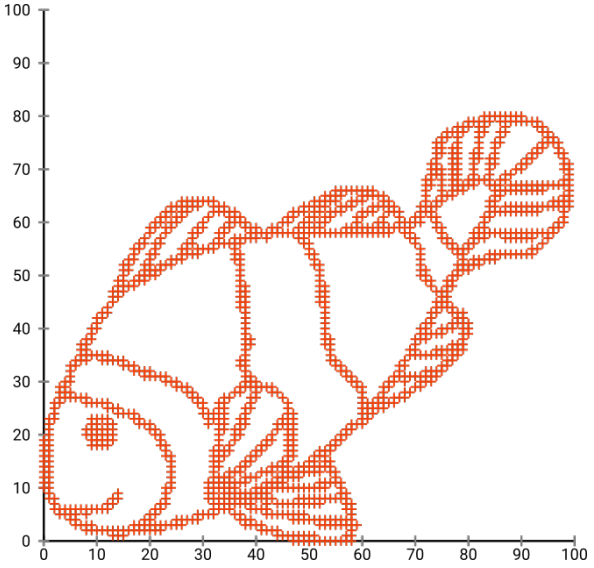
6 For topology, you may need more than one plot

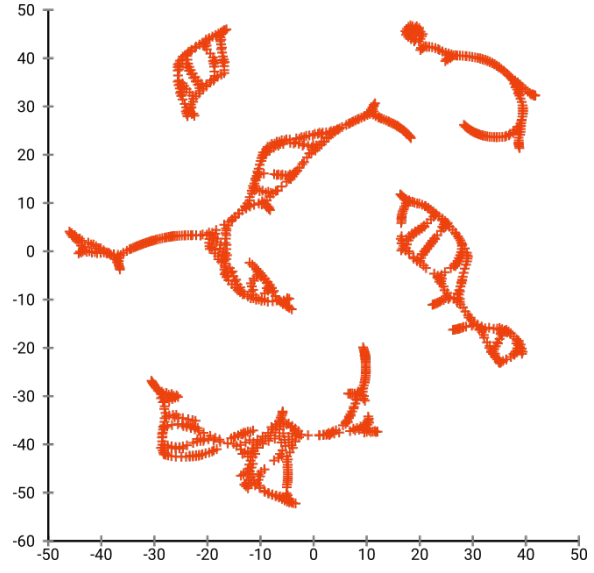


- two groups of 75 points in 50 dimensional space (blue contained in yellow)
- **perplexity 30** view **shows the basic topology correctly**, but again t-SNE greatly exaggerates the size of the smaller group of points.
- At **perplexity 50**, there's a new phenomenon: the outer group becomes a circle, as the plot tries to depict the fact that all its points are about the same distance from the inner group.

The truth about t-SNE

- **t-SNE is incredibly flexible**, and can often find structure where other dimensionality-reduction algorithms cannot. Unfortunately, that very **makes it tricky to interpret**.
- t-SNE **does not preserve distances nor density**. It **only to some extent preserves nearest-neighbors**. The difference is subtle, but **affects any density- or distance based algorithm**.
- The visual clusters can be **strongly influenced by** the chosen **parameterization** and therefore a good understanding of the parameters for t-SNE is necessary. Interactive exploration may thus be necessary to choose parameters and validate results.





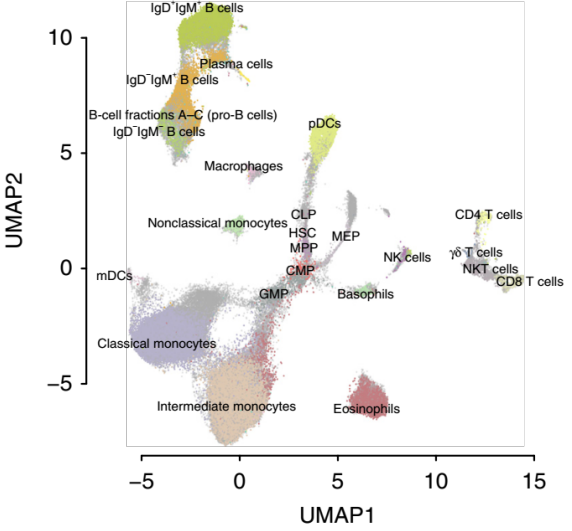
t-SNE uses the t-distribution in the projected space. In contrast to the Gaussian distribution used by regular SNE, this means **most points will repel each other**, because they have 0 affinity in the input domain (Gaussian gets zero quickly), but >0 affinity in the output domain. Sometimes (as in MNIST) this makes nicer visualization. In particular, **it can help "splitting" a data set a bit more** than in the input domain.

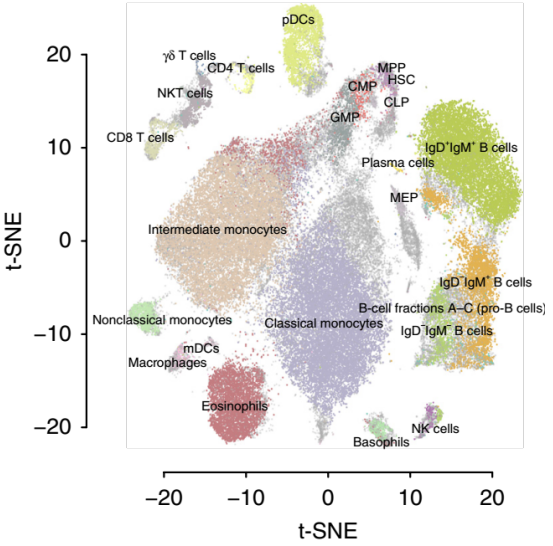
https://stats.stackexchange.com/questions/263539/clustering-on-the-output-of-t-sne/264647#264647

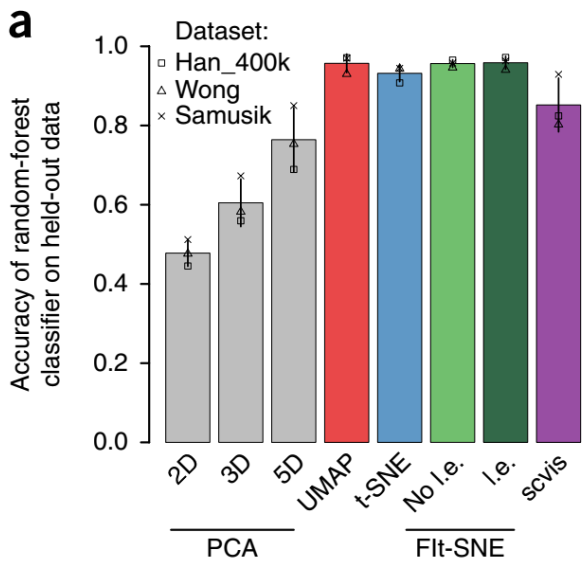
UMAP

(uniform manifold approximation and projection)

Comparing performance of UMAP with five other tools: **fastest run times, highest reproducibility** and the **most meaningful organization of cell clusters.**







Dimensionality reduction for visualizing single-cell data using UMAP; Becht 2019; Nat. Biotechnology

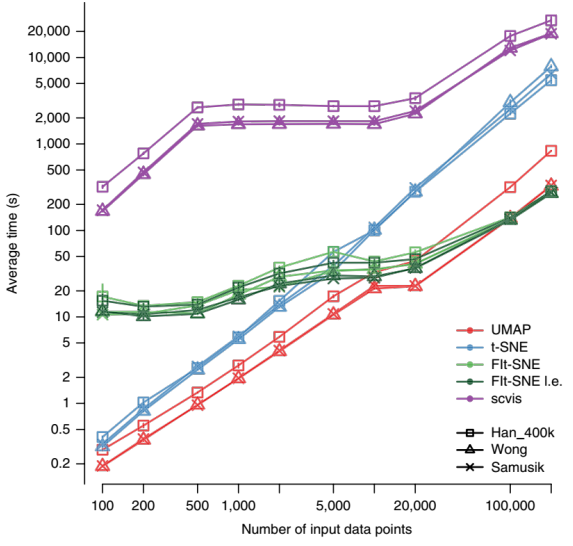


Figure 3 Run times of five dimensionality reduction methods for inputs of varying sizes. The average run time of three random subsamples is represented, with vertical bars representing s.d. after log-transforming the run times.

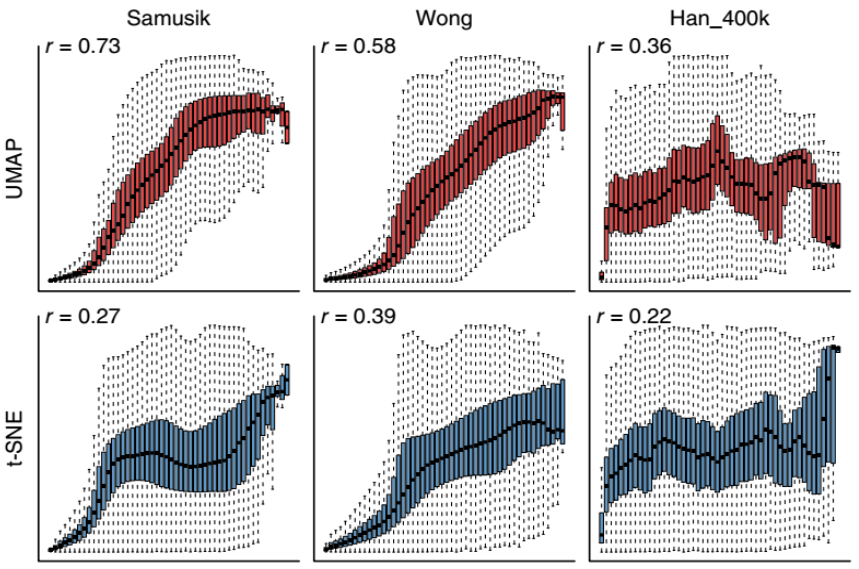


Figure 5 Preservation of pairwise distances in embeddings.

References

- Interactive exploration of t-SNE (basis of most of the talk)
  https://distill.pub/2016/misread-tsne/
- Visualizing data using t-SNE (google talk, indroducing t-SNE method and applications)
  https://www.youtube.com/watch?v=RJVL80Gg3lA&list=UUtXKDgv1AVoG88PLl8nGXmw
- Developer of t-SNE
  https://lvdmaaten.github.io/tsne/
- Wikipedia article (math background)
  https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding
- t-SNE and clustering discussion
  https://stats.stackexchange.com/questions/263539/clustering-on-the-output-of-t-sne/264647#264647
- Dimensionality reduction for visualizing single-cell data using UMAP; Becht 2019; Nat. Biotechnology
  https://www.nature.com/articles/nbt.4314