Data Analysis and Visualization

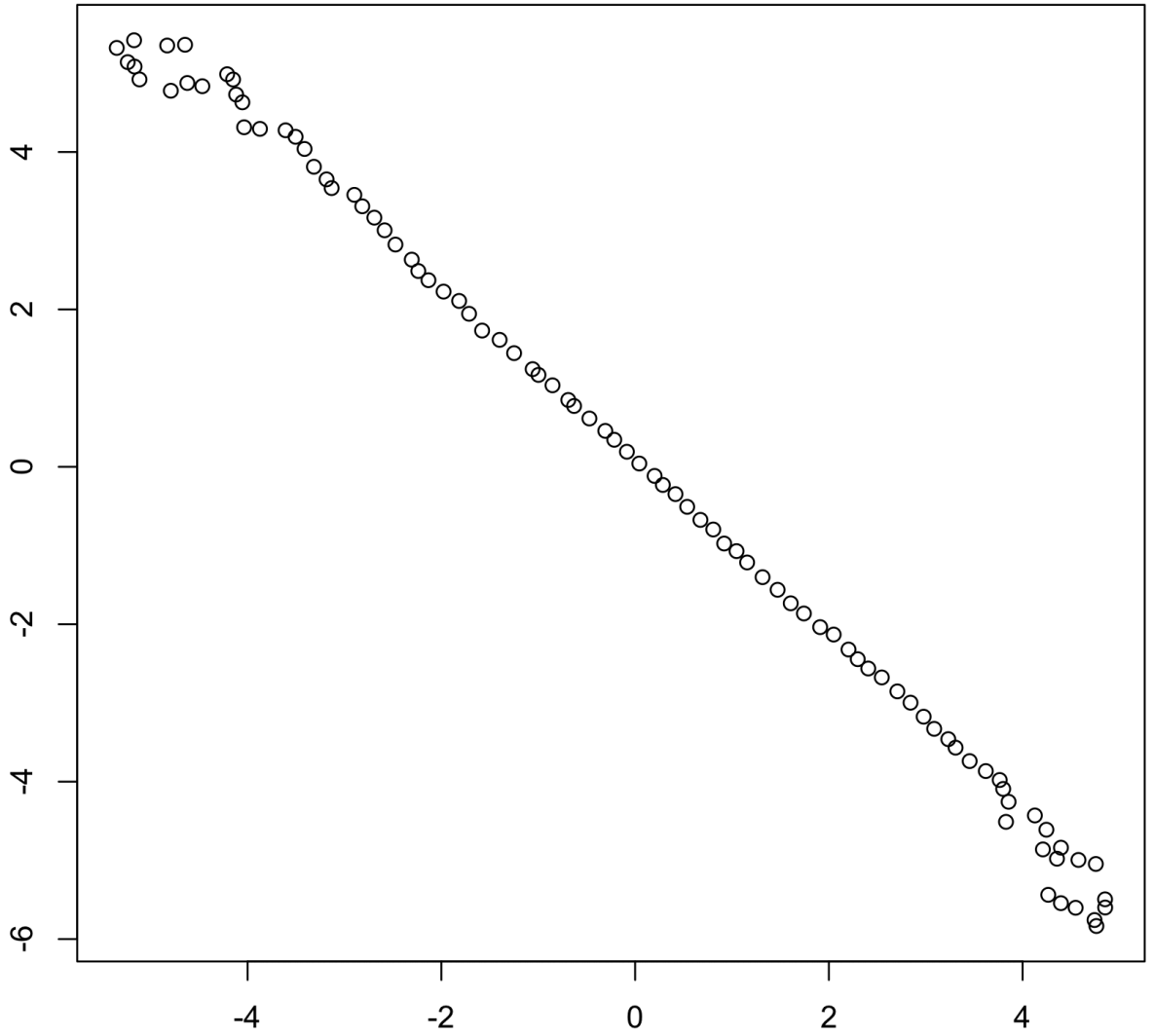
Universität Konstanz

Information Visualization

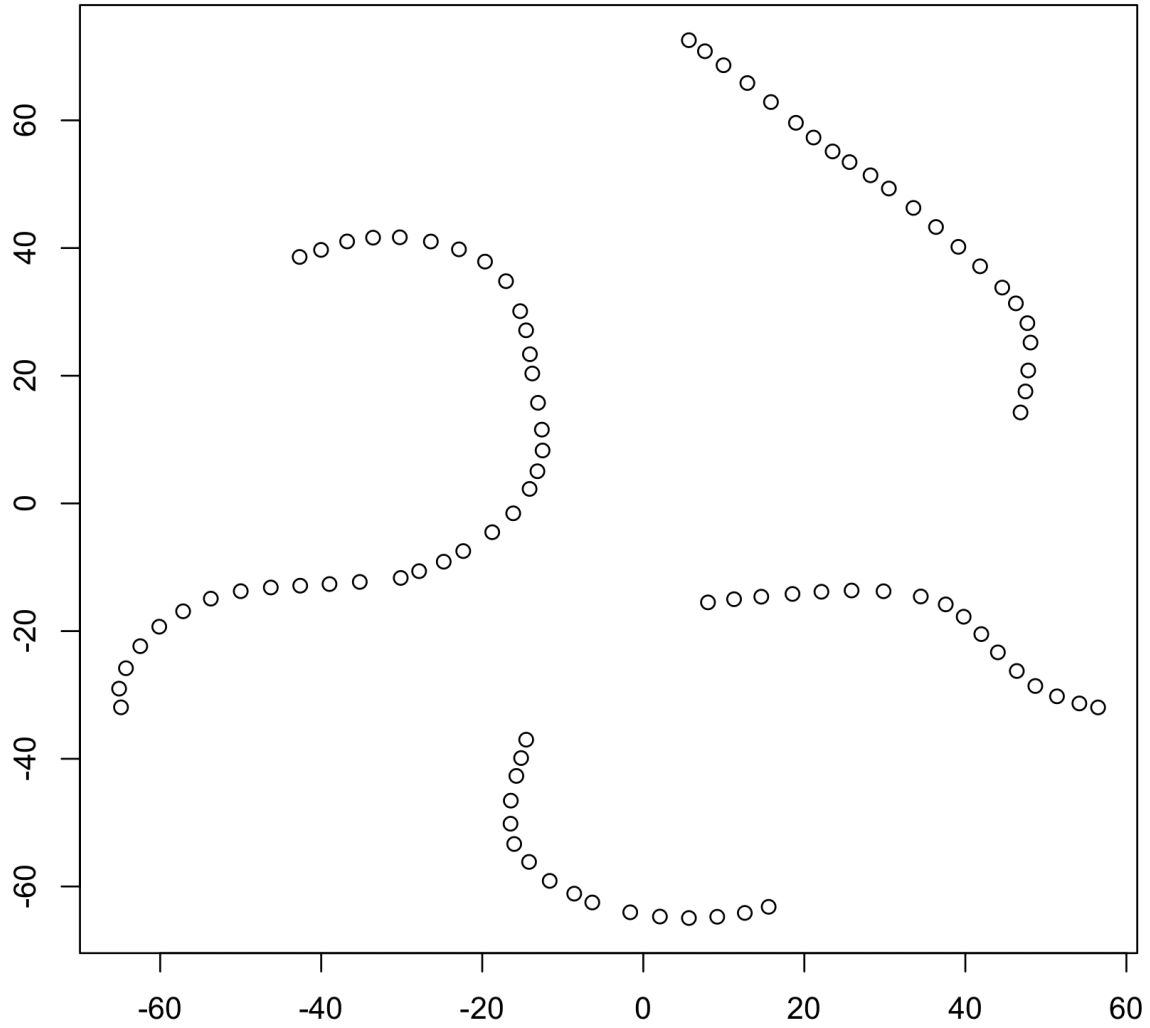
Projections



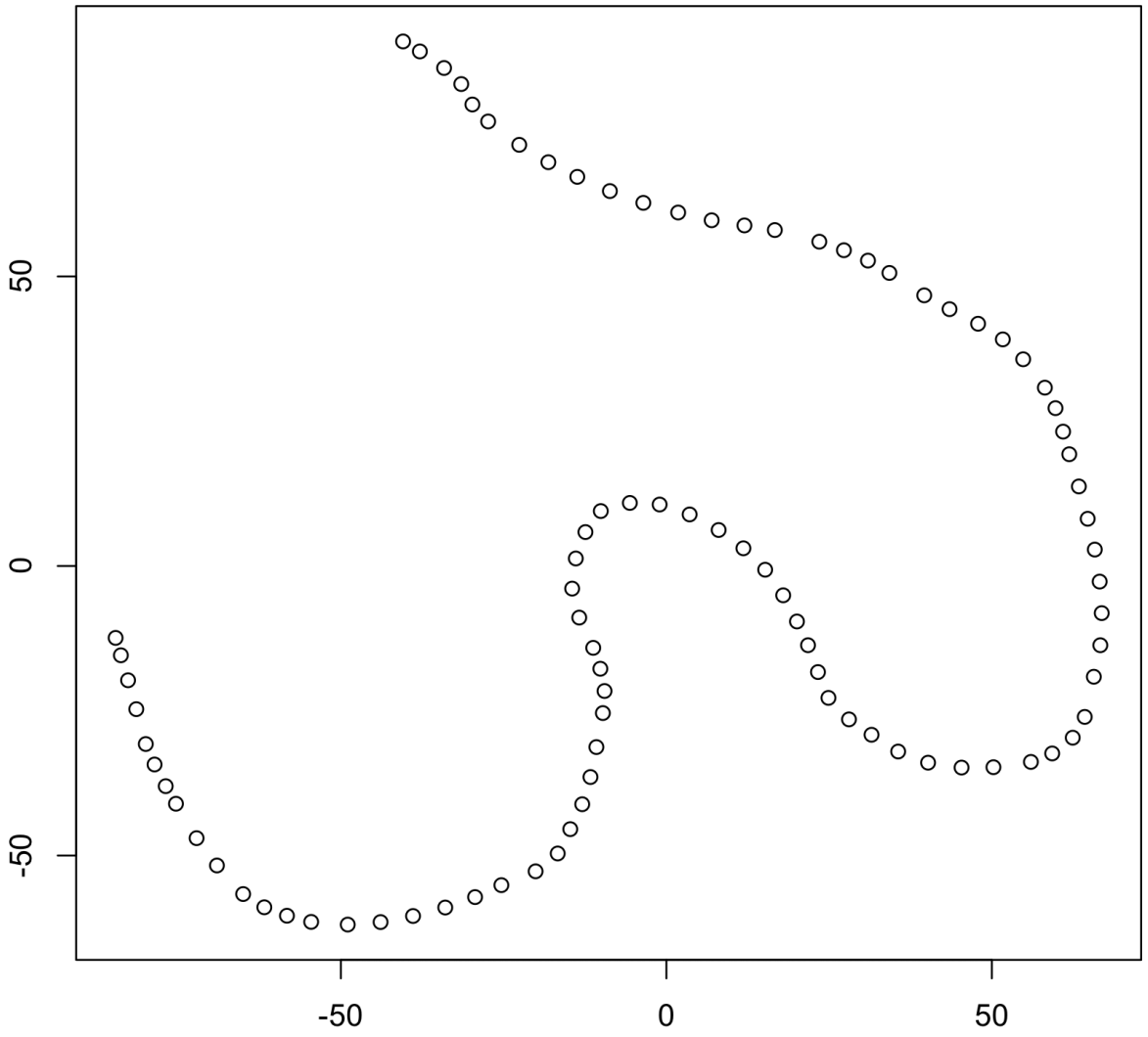
Dr. Johannes Fuchs



Projections



Projections



Projections

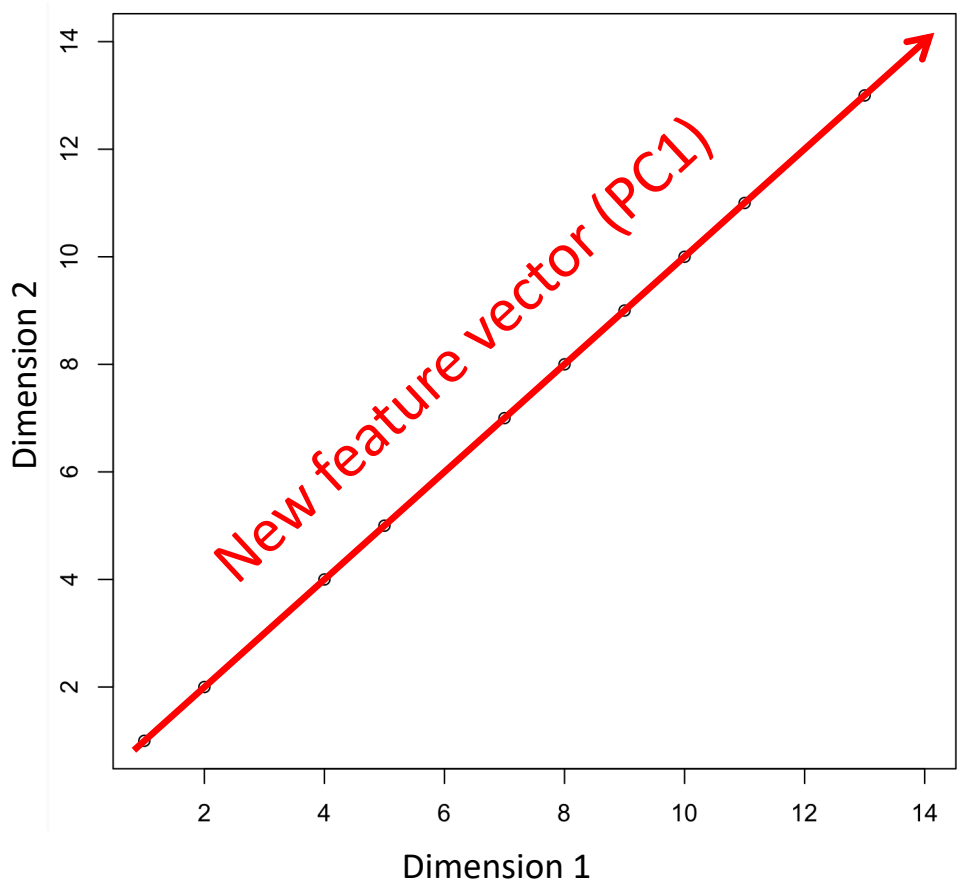
1. Principal Component Analysis (PCA)
2. Multi Dimensional Scaling (MDS)
3. t-Distributed Stochastic Neighboring Embedding (t-SNE)

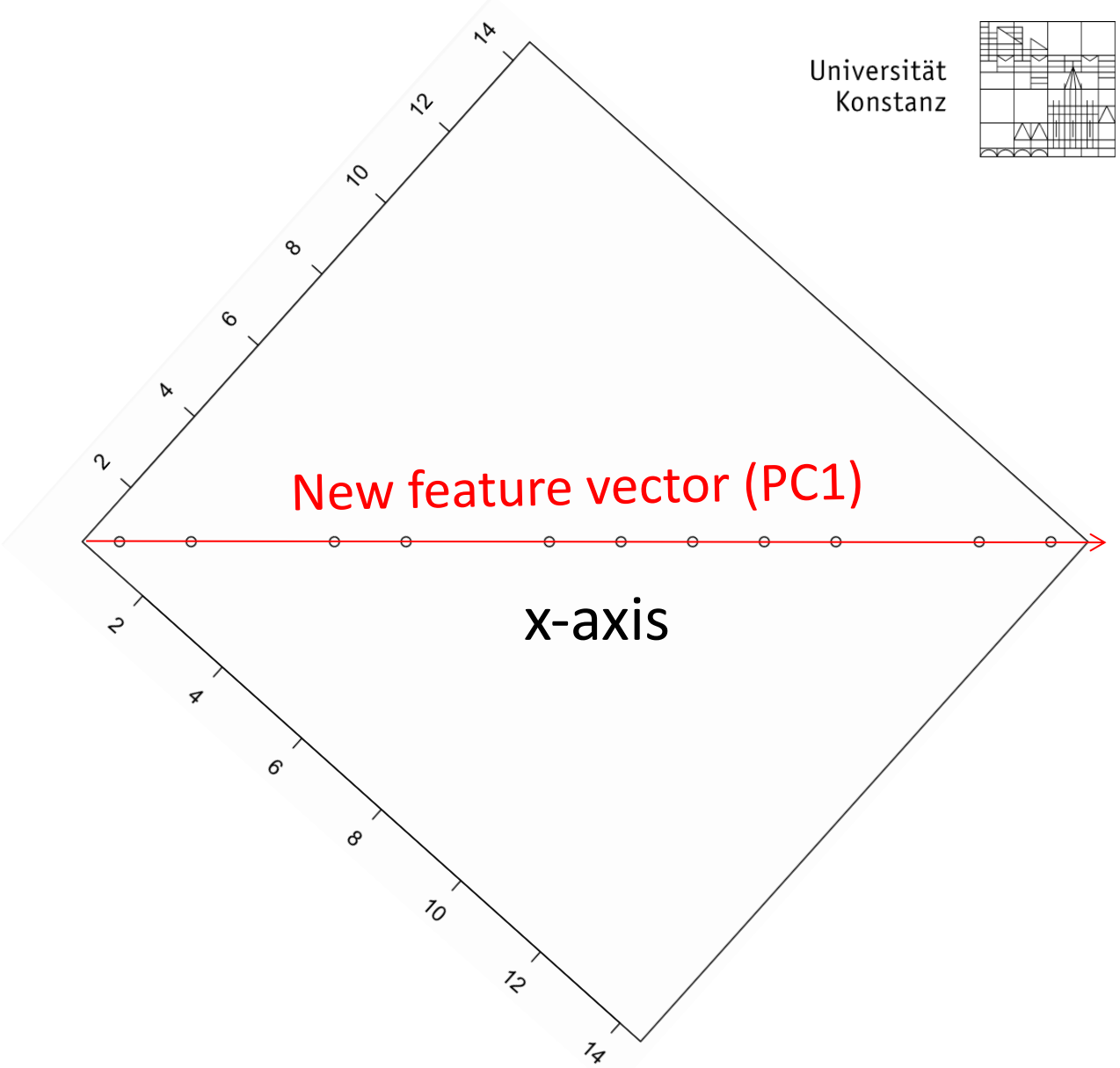
Projections - Agenda

Goal:

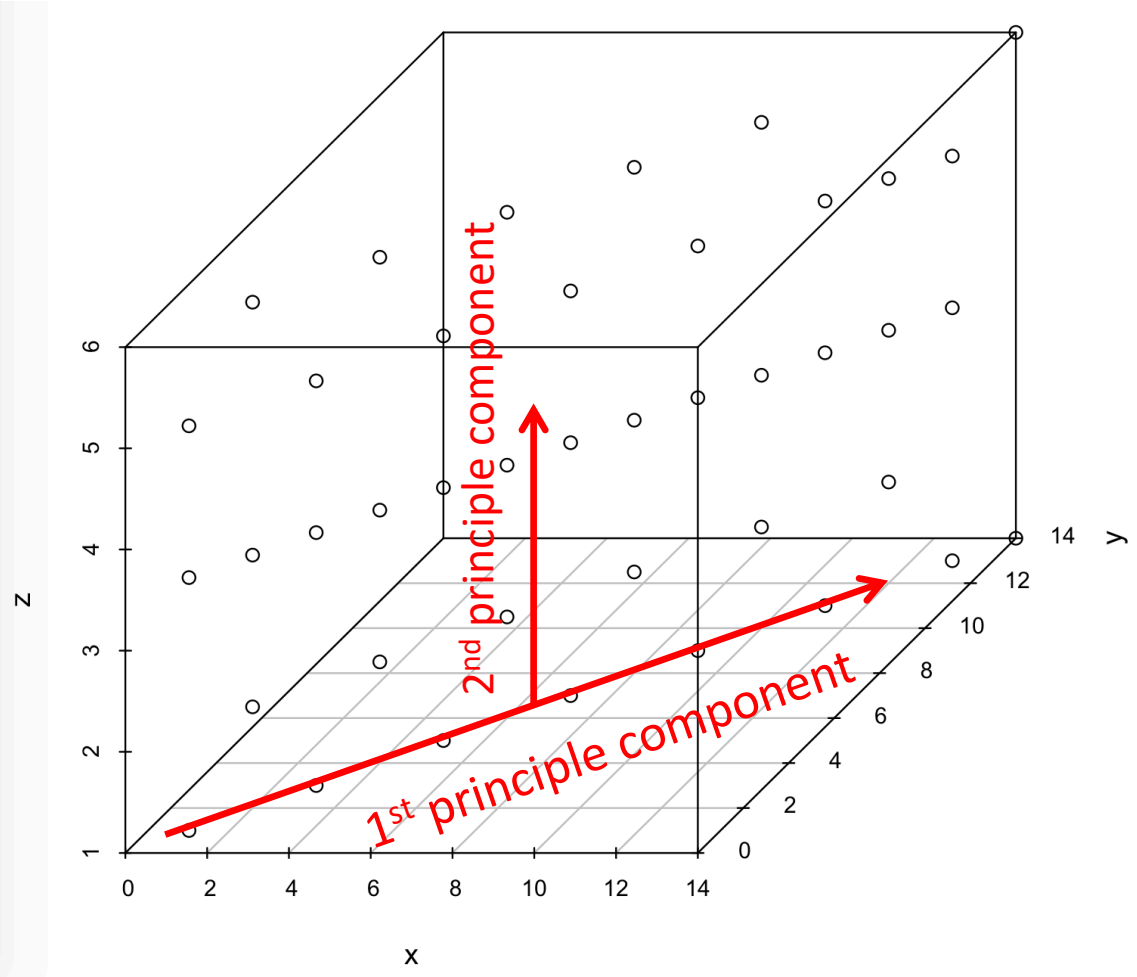
- find data representation that retains the maximum **nonredundant** and **uncorrelated** information

- explore statistical **correlations** among elements of the original patterns

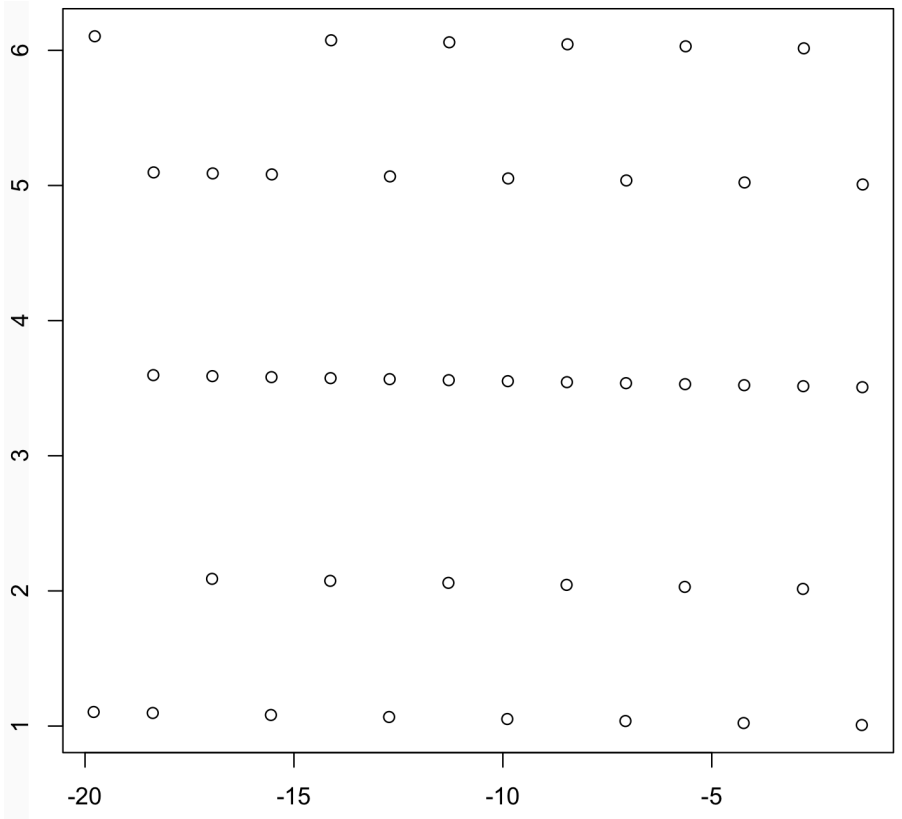




Projections - PCA



PCA



Projections - PCA

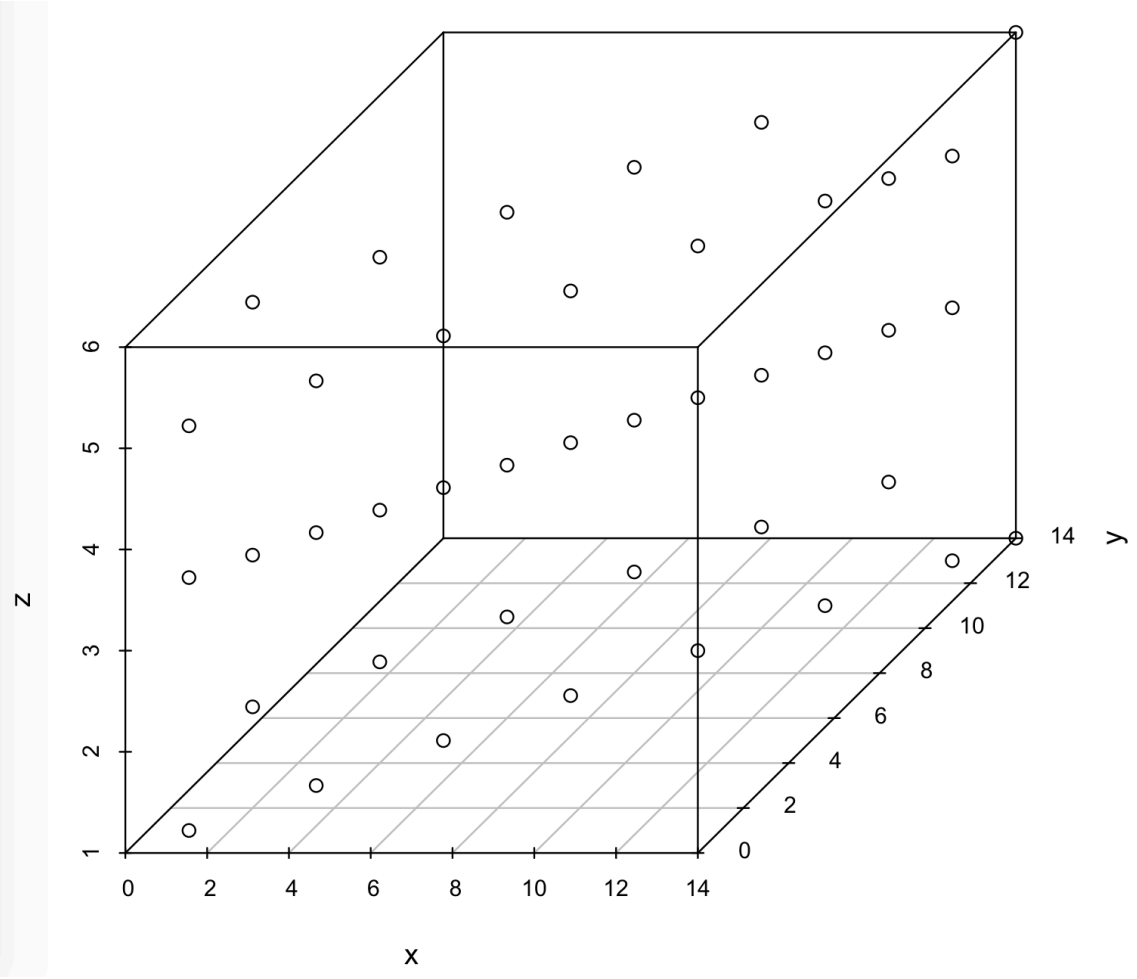
Summary:

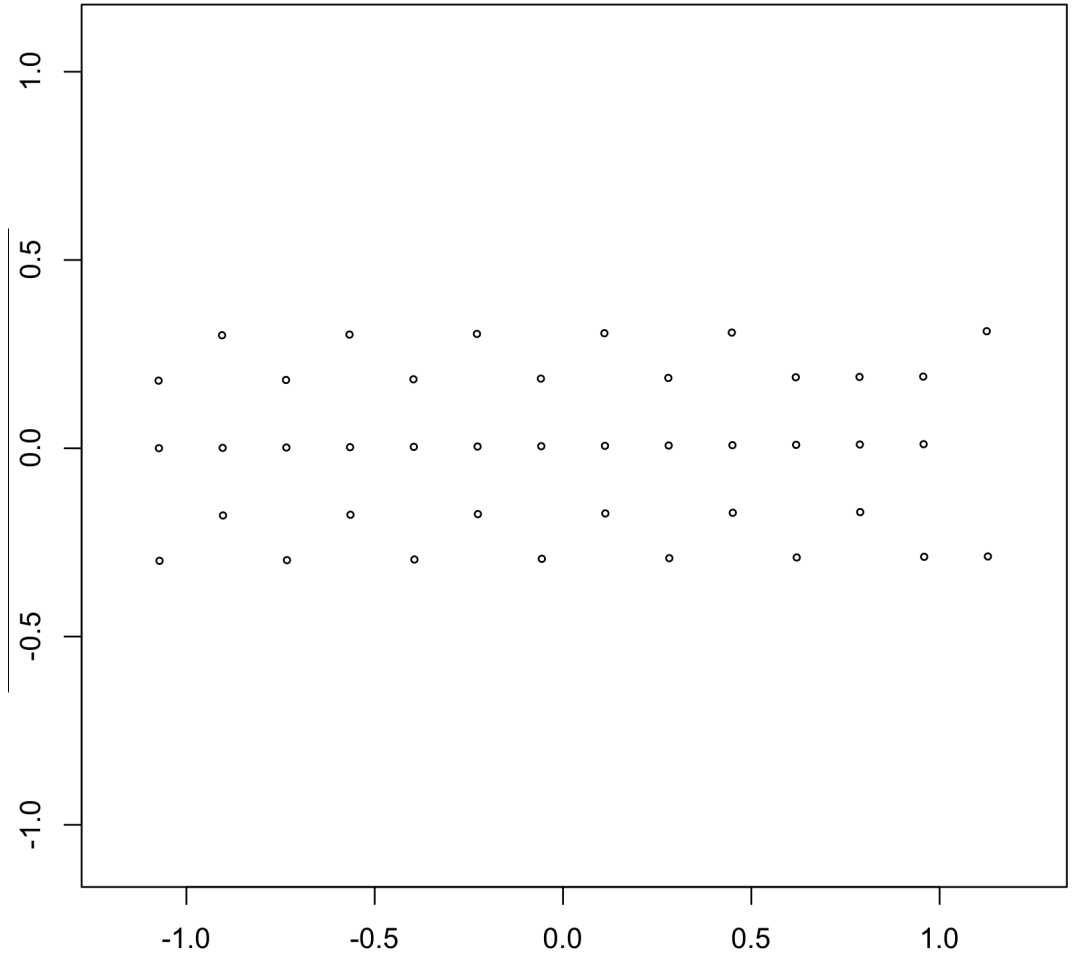
- number of dimensions is equal to the number of principal components
- is prone to oddities (e.g., outliers).
- tries to find hidden linear correlation.
- does not try to preserve the topology of the data.

Goal:

- place each data object in a lower dimensional space such that the **distances** between the data objects in the original space are preserved as well as possible.

- normally applied to project to two or three dimensions





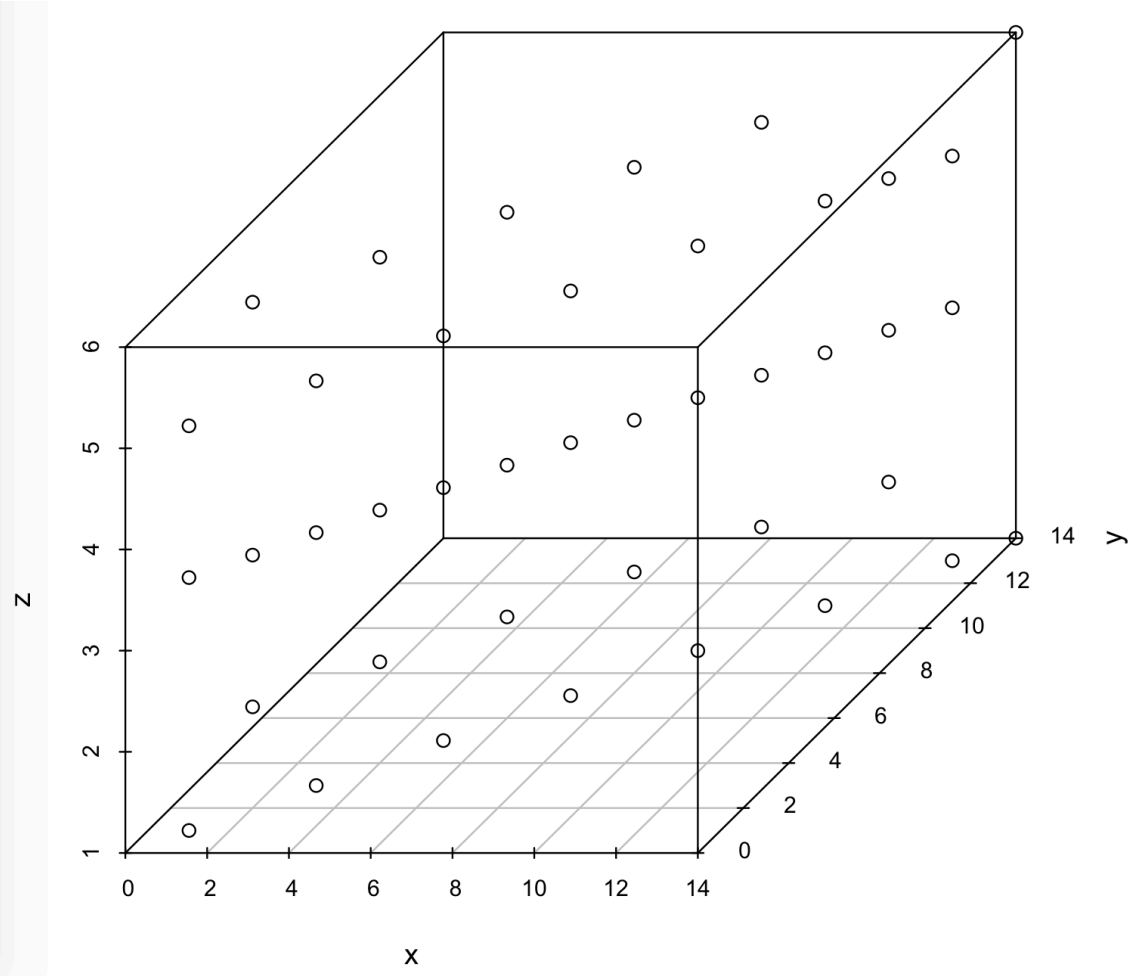
Projections - MDS

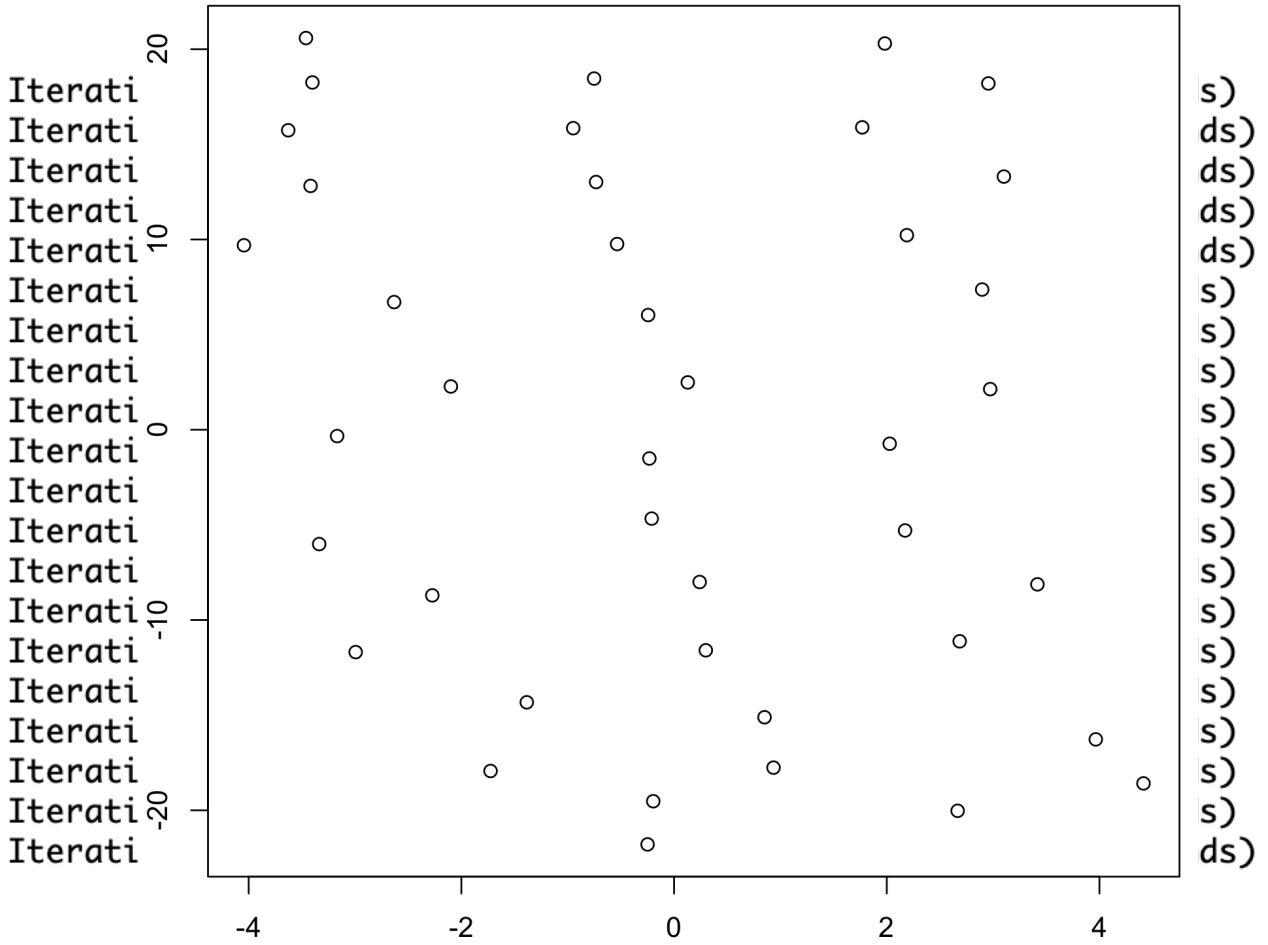
Summary:

- normally applied to project to two or three dimensions
- the accuracy can be judged by different (stress) measures
- new dimensions have no interpretation, just relative positions are important
- results are dependent on distance measures, stress measure, changes in the starting, termination and update conditions

Goal:

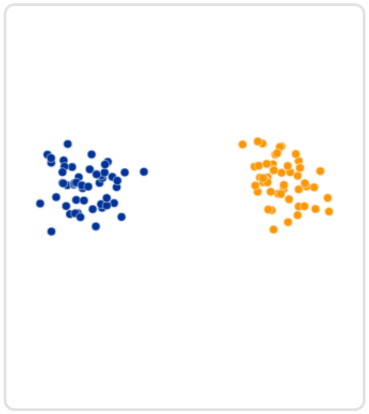
- Convert high-dimensional Euclidean distances into probabilities.
- Convert low-dimensional Euclidean distances into probabilities.
- Minimize the difference between the two.



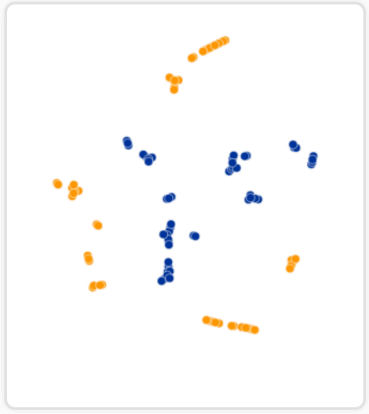


Projections – t-SNE

Maaten and Hinton, 2008



Original

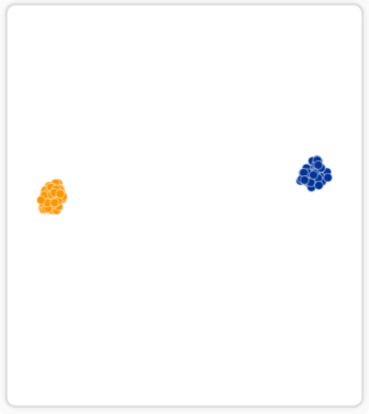


Perplexity: 2
Step: 5,000

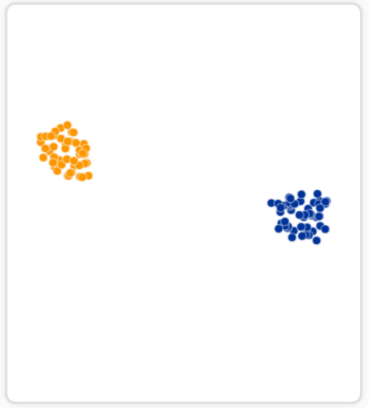


Perplexity: 5
Step: 5,000

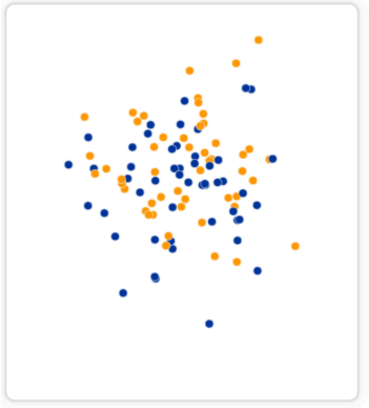
Strong influence of perplexity value on the projection



Perplexity: 30
Step: 5,000



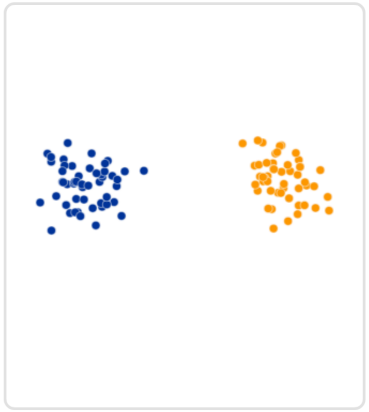
Perplexity: 50
Step: 5,000



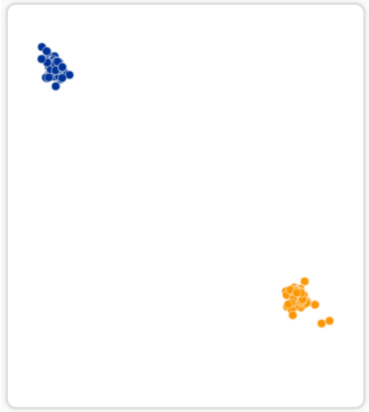
Perplexity: 100
Step: 5,000

Projections – t-SNE

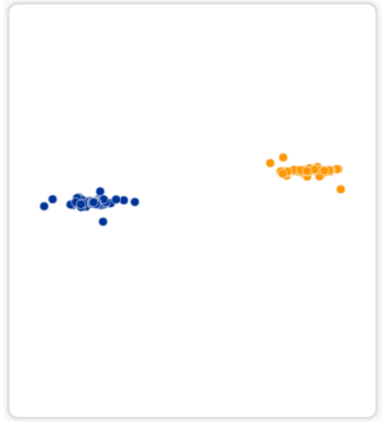
Wattenberg et al., 2016



Original

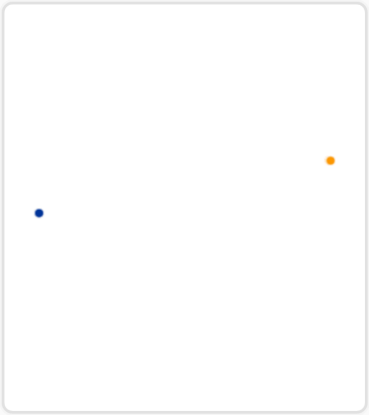


Perplexity: 30
Step: 10

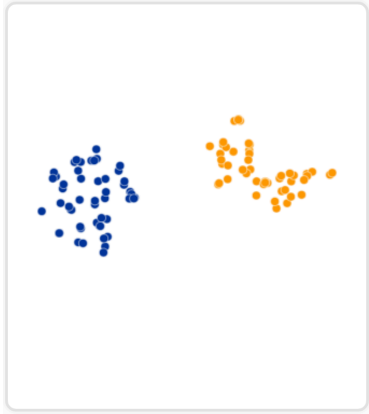


Perplexity: 30
Step: 20

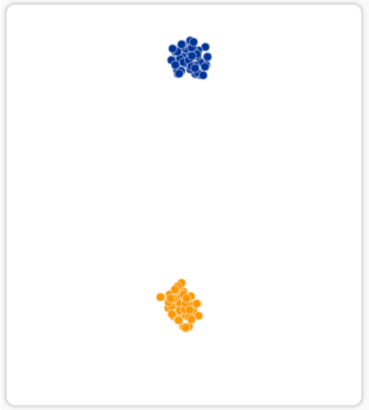
Number of optimal steps differ for different data sets.



Perplexity: 30
Step: 60



Perplexity: 30
Step: 120



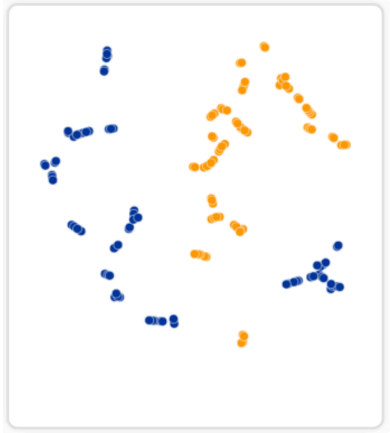
Perplexity: 30
Step: 1,000

Projections – t-SNE

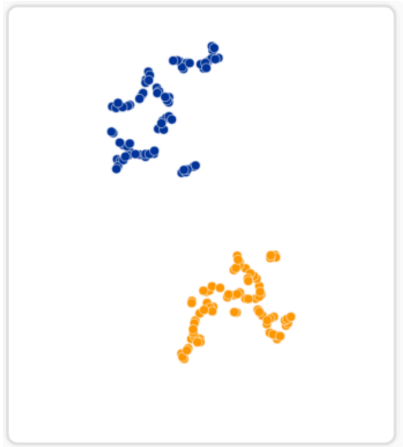
Wattenberg et al., 2016



Original

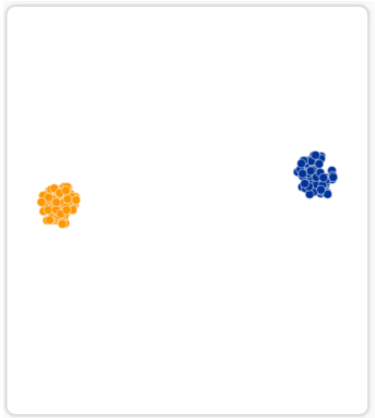


Perplexity: 2
Step: 5,000

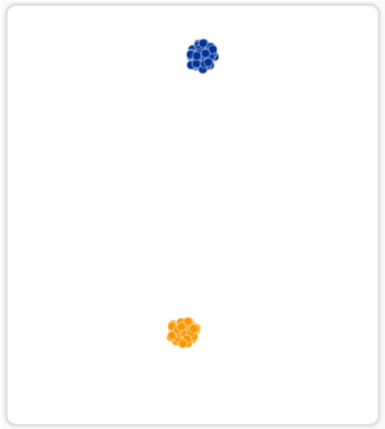


Perplexity: 5
Step: 5,000

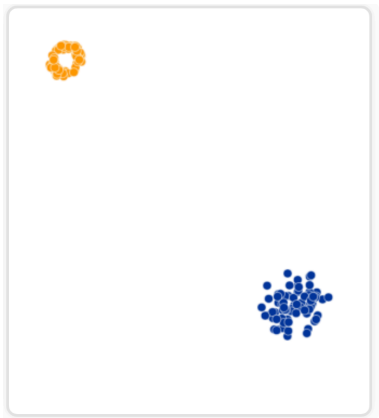
You cannot see relative sizes of clusters in t-SNE plots



Perplexity: 30
Step: 5,000



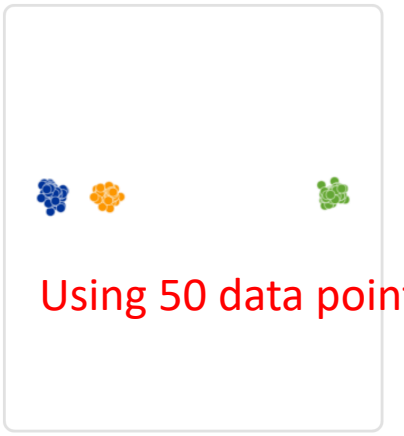
Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

Projections – t-SNE

Wattenberg et al., 2016



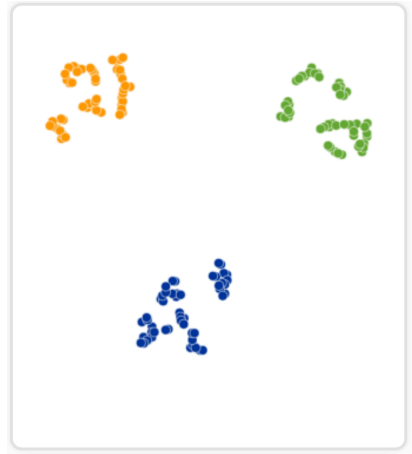
Original

Using 50 data points each



Perplexity: 2

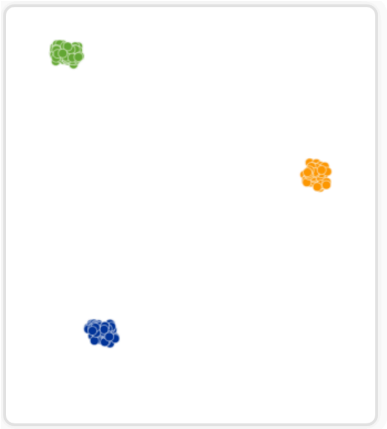
Step: 5,000



Perplexity: 5

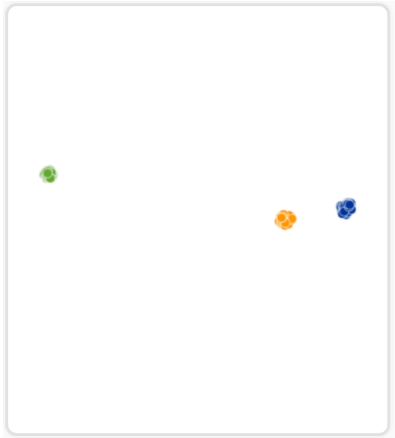
Step: 5,000

Distances between well-separated clusters may mean nothing



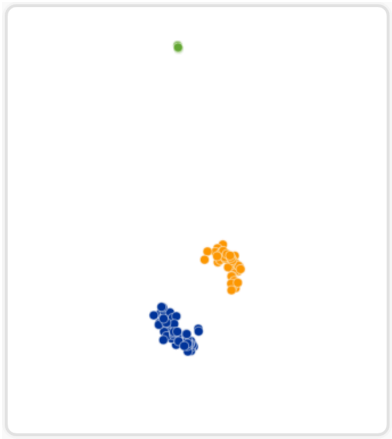
Perplexity: 30

Step: 5,000



Perplexity: 50

Step: 5,000

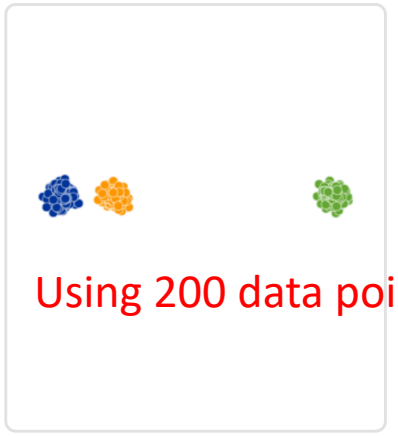


Perplexity: 100

Step: 5,000

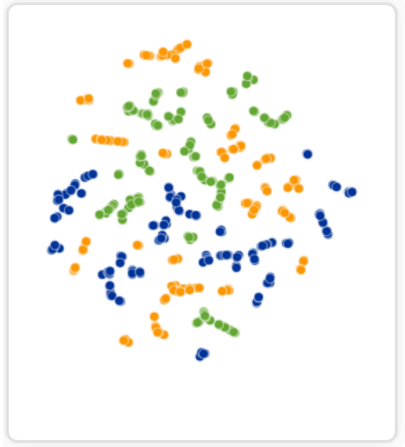
Projections – t-SNE

Wattenberg et al., 2016



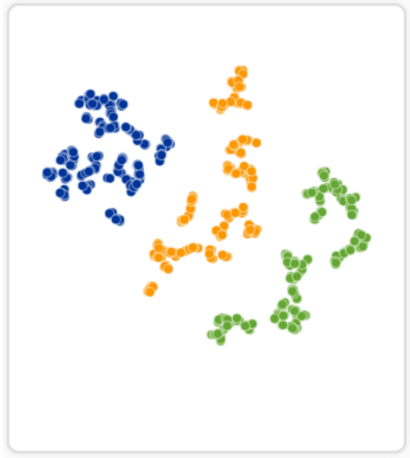
Original

Using 200 data points each



Perplexity: 2

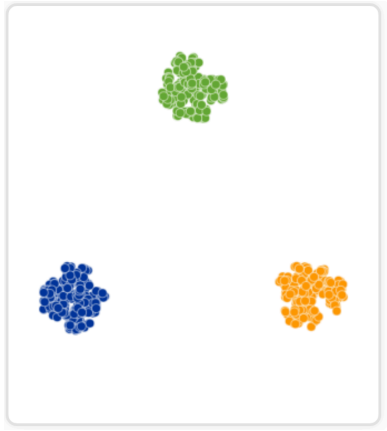
Step: 5,000



Perplexity: 5

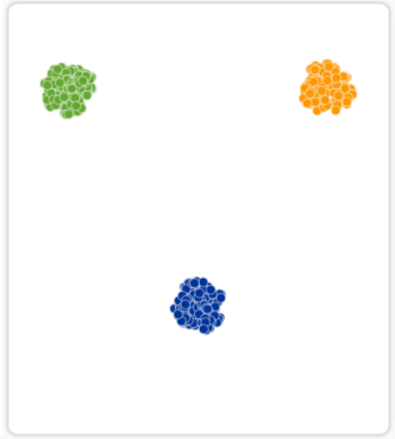
Step: 5,000

Distances between well-separated clusters may mean nothing



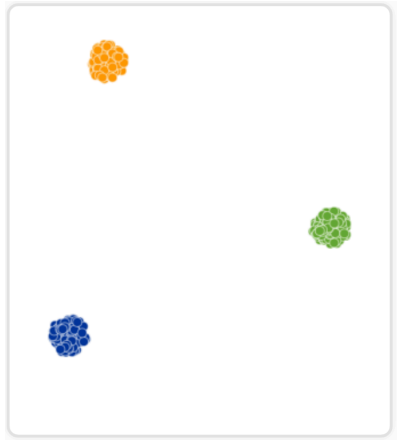
Perplexity: 30

Step: 5,000



Perplexity: 50

Step: 5,000

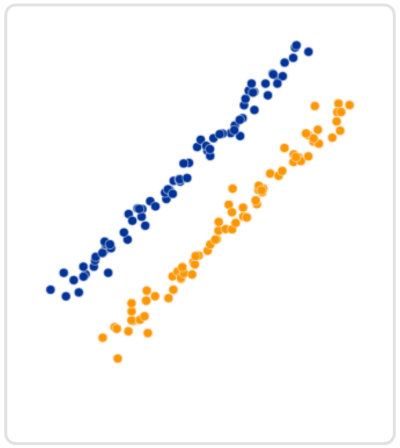


Perplexity: 100

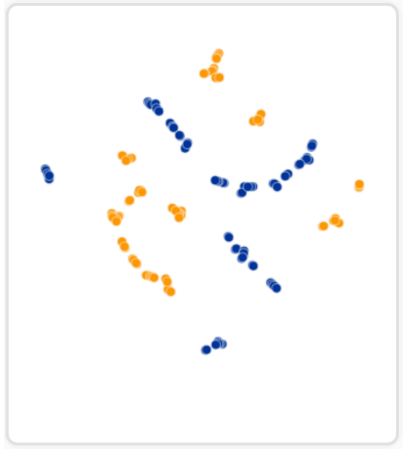
Step: 5,000

Projections – t-SNE

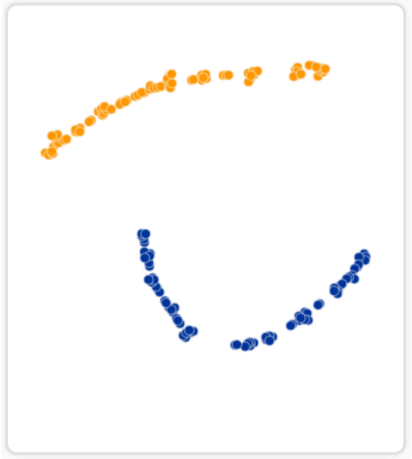
Wattenberg et al., 2016



Original

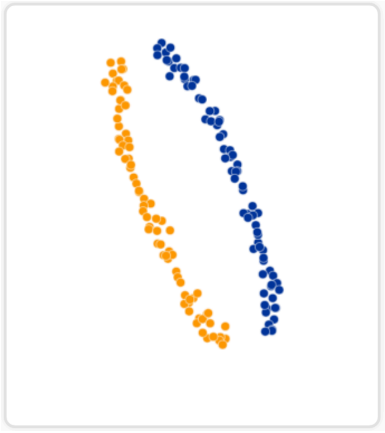


Perplexity: 2
Step: 5,000

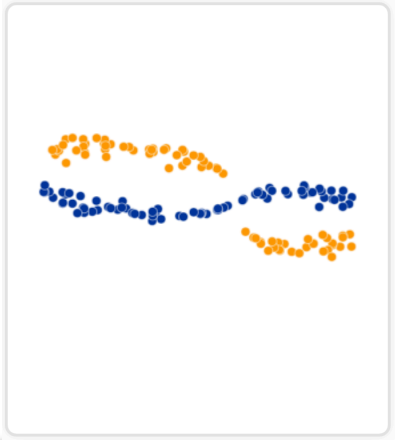


Perplexity: 5
Step: 5,000

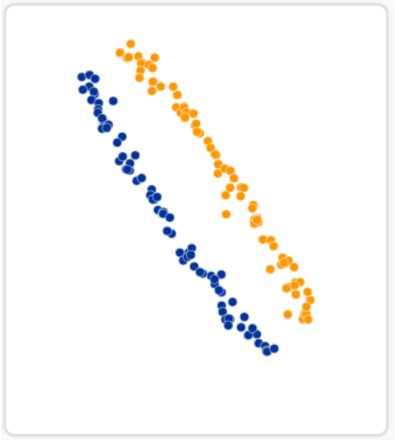
Be careful with linear relationships



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

Projections – t-SNE

Wattenberg et al., 2016

Summary:

- Because of the Gaussian kernel function local and global structures are preserved
- Really good for detecting clusters in the data

- Results depend on hyper-parameters:
  - Perplexity
  - Iterations
  - Learning rate

R-Code: 05_Projections.r

Simply apply the code snippets to your data set.

https://visxprojections.dbvis.de/client/index.html

Projections - Summary

- For detecting clusters, t-SNE is a suitable choice.
  - However, try multiple input parameters and look for consistent representations
  - Distances or densities have no meaning
- For detecting linear relationships, you can use a PCA projection.
- MDS can be used with different distance metrics. If the Euclidean distance is used, the result looks similar to a PCA projection.
  - Distances are preserved as good as possible.

Guidelines – Projections

References

Maaten and Hinton, 2008
Visualizing Data Using t-SNE
Maaten, Laurens van der and Hinton, Geoffrey: Journal of machine learning research, 2579—2605, 2008

Wattenberg et al., 2016
How to Use t–SNE Effectively
Wattenberg, Martin and Viégas, Fernanda and Johnson, Ian: Distill, 2016