DLI Accelerated Data Science Teaching Kit

Lecture 15.5 - t-SNE

t-SNE

t-distributed Stochastic Neighbor Embedding

- Non-linear dimensionality reduction method developed for visualizing high-dimensional data in low-dimensional space (e.g., 2D, 3D) [1]
- Widely used in numerous fields and applications
- Main ideas
    - Models similarities between data points as joint probabilities
        - Similar (dissimilar) points assigned a higher (lower) probability
    - Represents each high-dimensional point by a low-dimensional version (e.g., 2D)
    - Minimizes KL divergence between the joint probabilities between originally high-dimensional data and low-dimensional representation

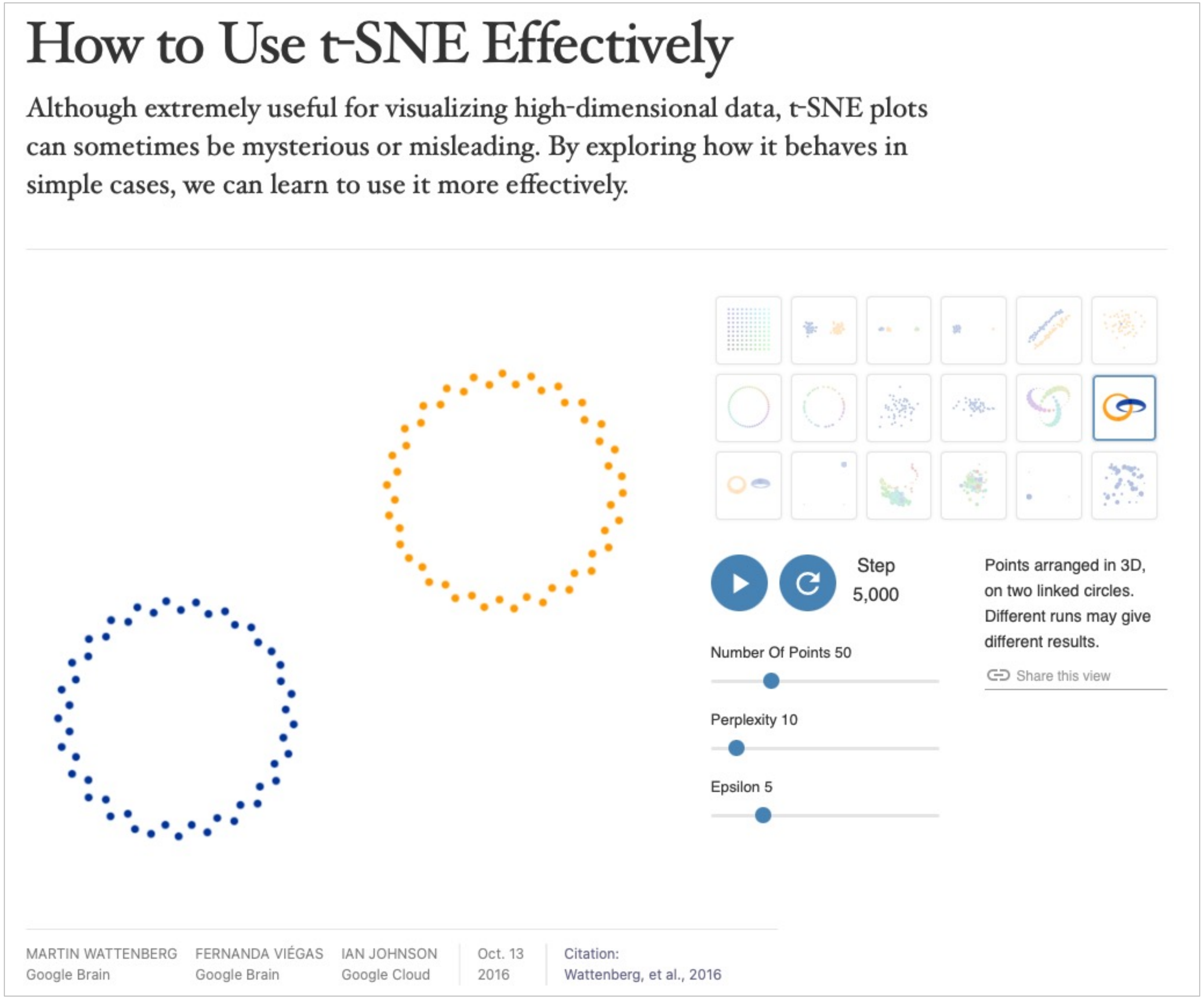
[1] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

t-SNE

t-distributed Stochastic Neighbor Embedding

- For super-high dimensional data, typically first apply another dimensionality reduction method (e.g., SVD)
- Can help discover patterns other techniques cannot (e.g., when linear assumptions violated for PCA)
- Extension over SNE
    - t-SNE uses heavy-tailed t-distribution (vs Gaussian)
    - Suitable for reducing to a very low dimensions (e.g., 2D)
- Used to be considered a “slow” method (solving n-body problem)
- Now has fast approximate algorithms, and GPU acceleration (available in RAPIDS)

[1] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.



How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.

MARTIN WATTENBERG Google Brain

FERNANDA VIÉGAS Google Brain

IAN JOHNSON Google Cloud

Oct. 13 2016

Citation: Wattenberg, et al., 2016

Important Considerations when Using t-SNE

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

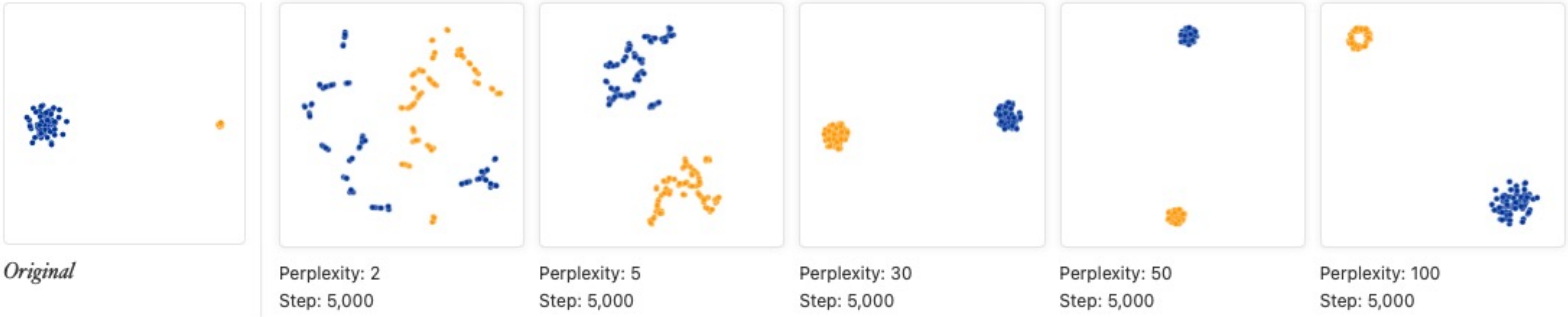
1. Hyperparameters really matter

Perplexity recommended to be 5-50. Should be smaller than number of data points.

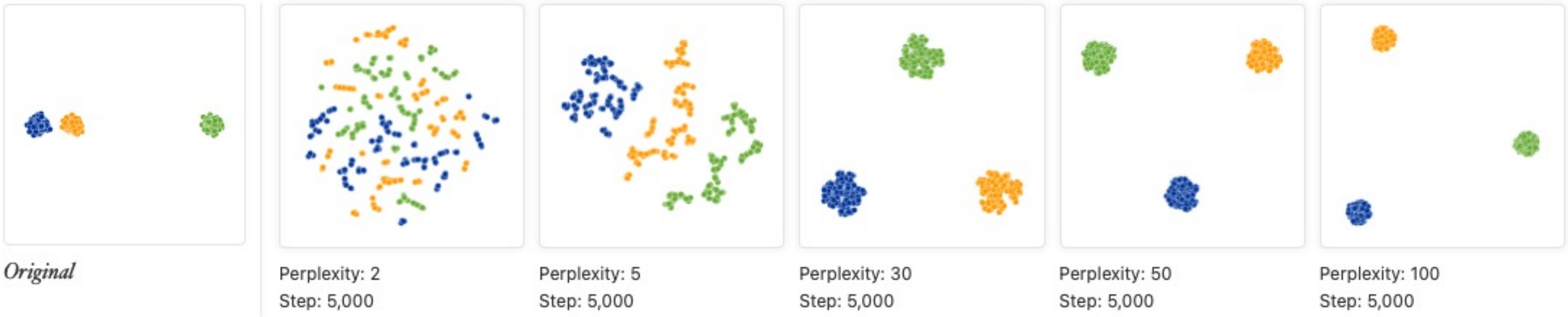
1. Hyperparameters really matter

“Pinched” may indicate stopping too early. Should wait until convergence.

2. Cluster sizes in a t-SNE plot mean nothing



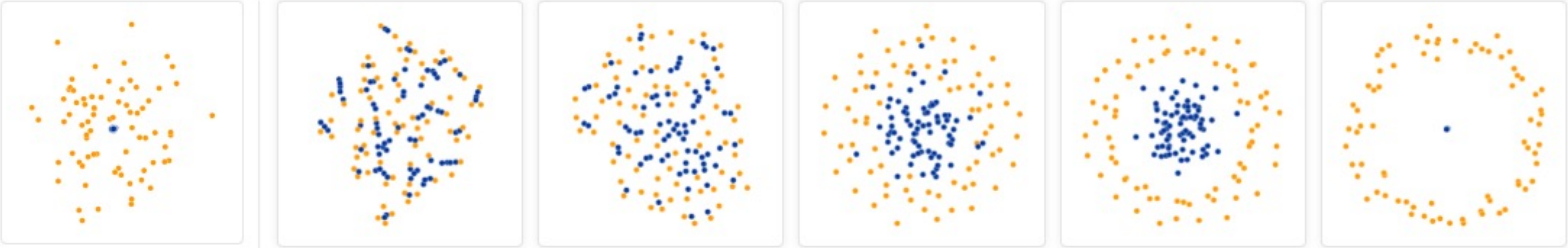
3. Distances between clusters might not mean anything



5. You can see some shapes, sometimes

See shapes as “right” perplexity. t-SNE tends to magnify dense regions.

6. For topology, you may need more than one plot



Original

Perplexity: 2
Step: 5,000

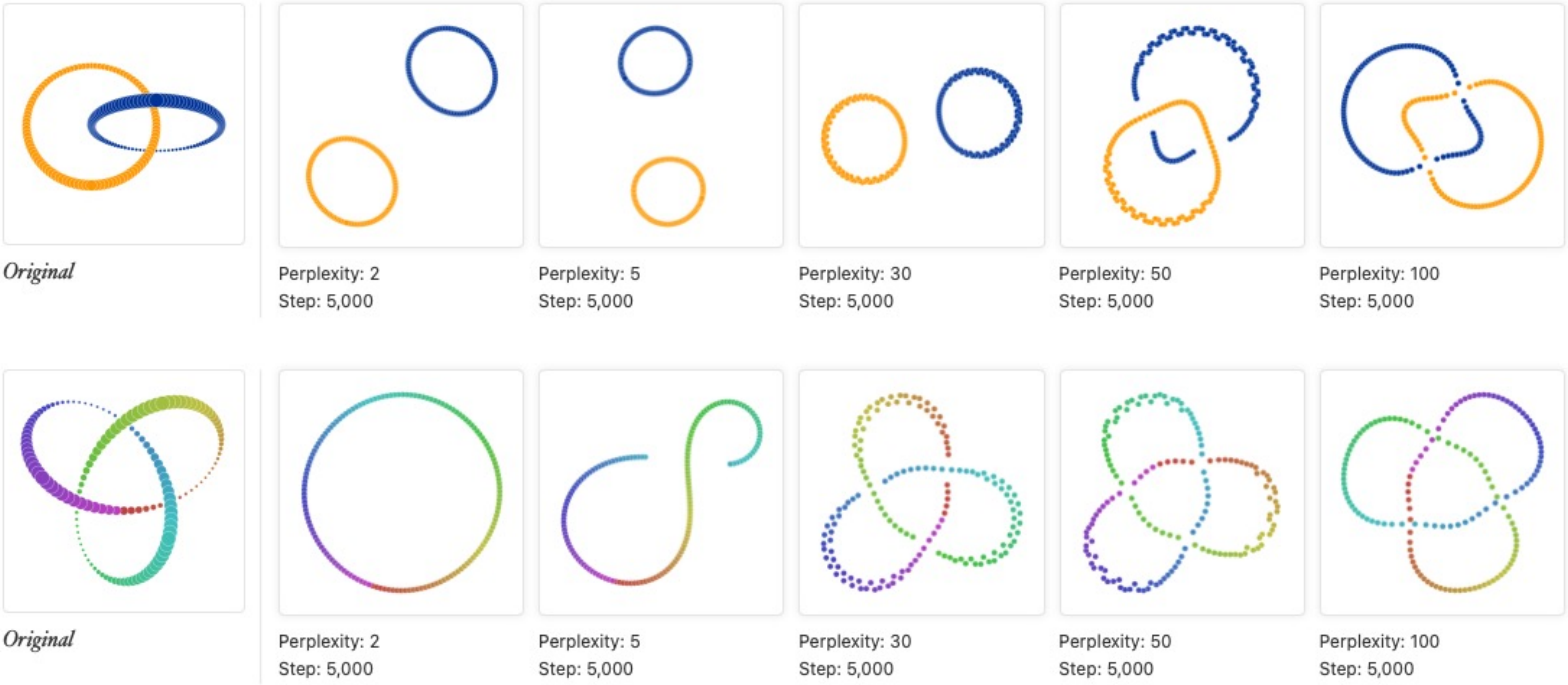
Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

6. For topology, you may need more than one plot



DLI Accelerated Data Science Teaching Kit

Thank You