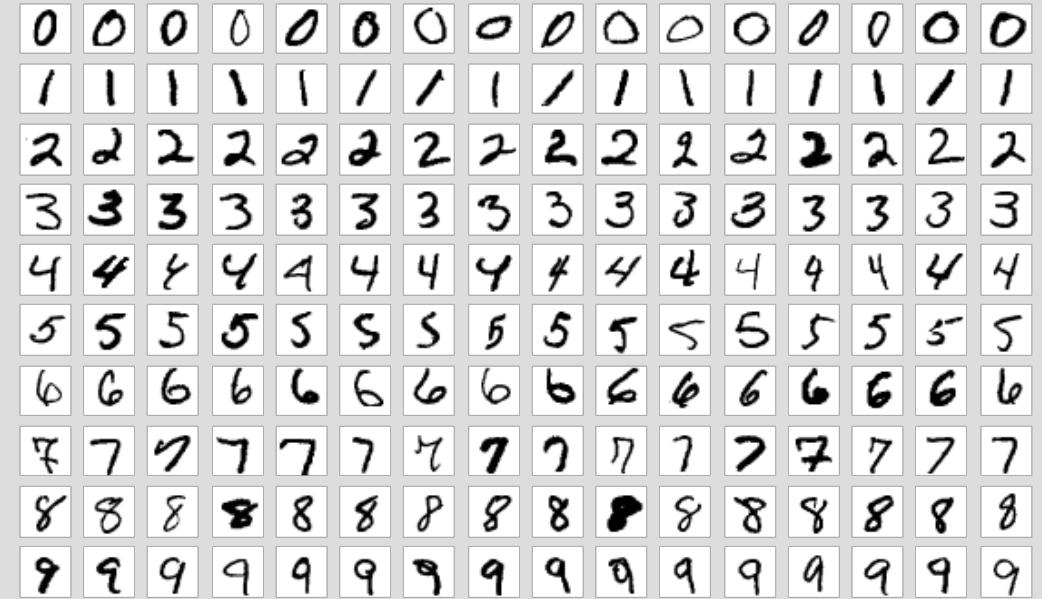


<https://lvdmaaten.github.io/tsne/>

## MNIST data

(<http://yann.lecun.com/exdb/mnist/>)

Contains 70,000  
handwritten digits from 0...10  
as vectors of length  $784 = 28 \times 28$



Left: 6000 digits analyzed with t-SNE

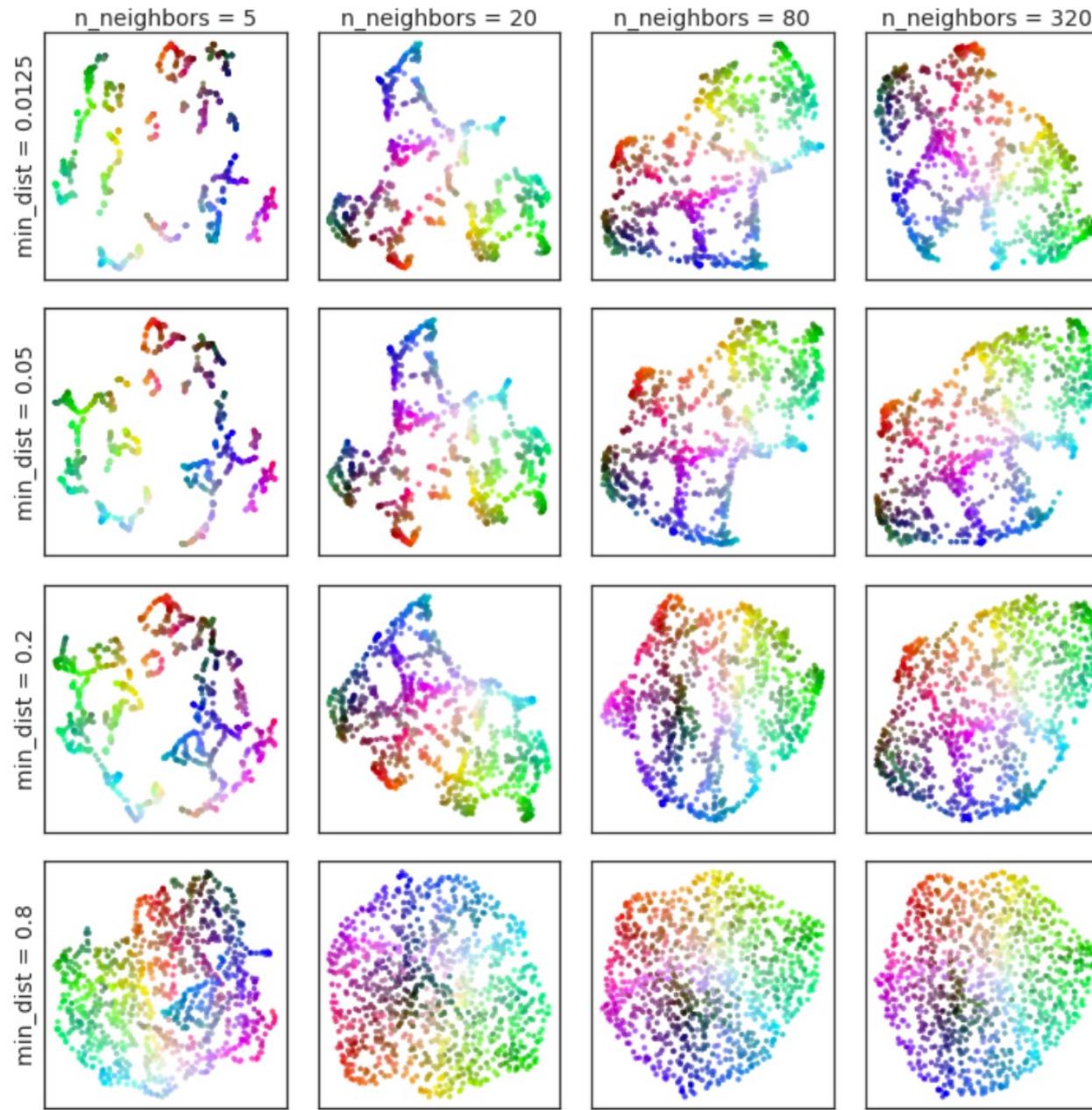
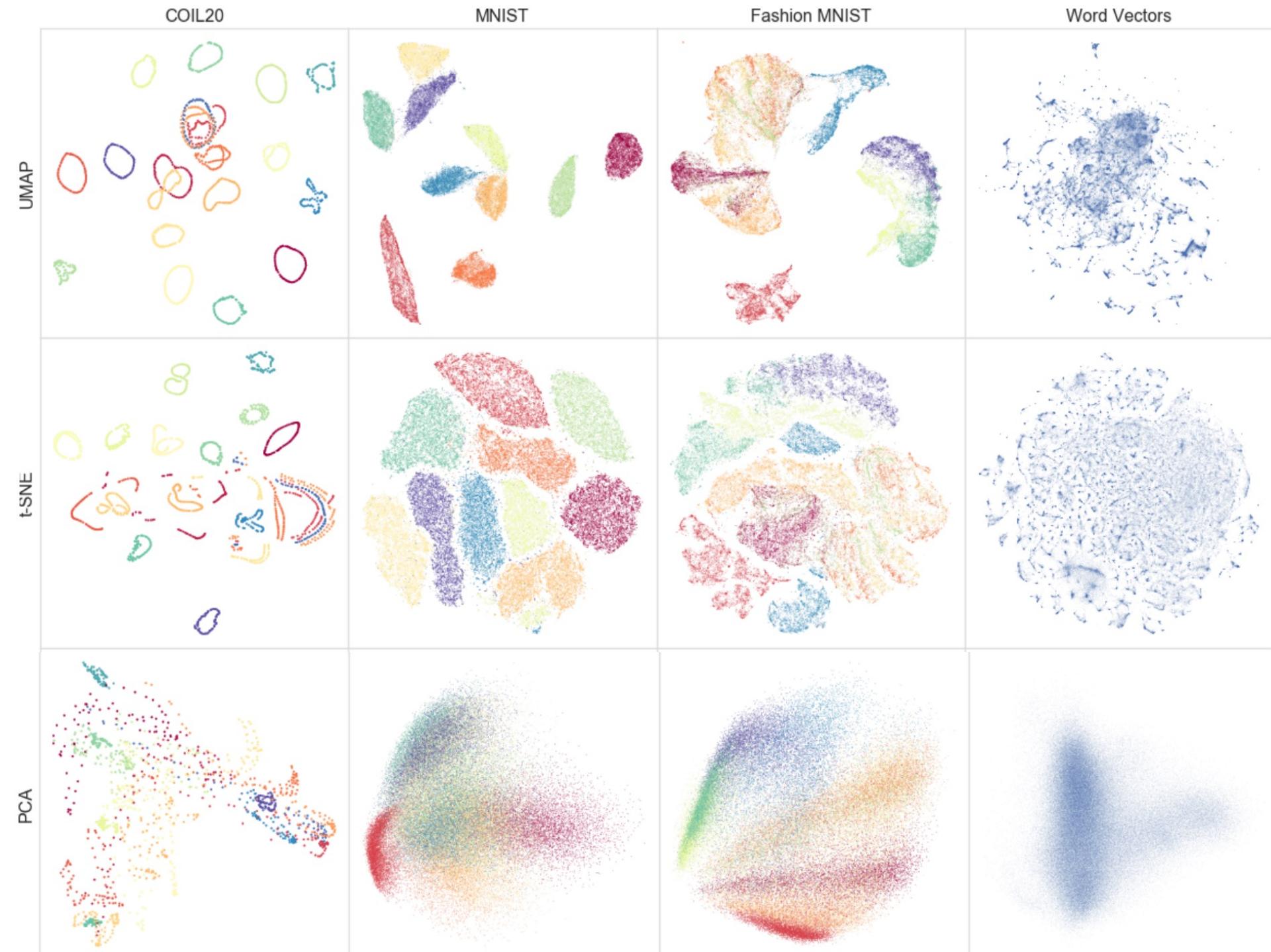


Figure 1: Variation of UMAP hyperparameters  $n$  and  $\text{min-dist}$  result in different embeddings. The data is uniform random samples from a 3-dimensional color-cube, allowing for easy visualization of the original 3-dimensional coordinates in the embedding space by using the corresponding RGB colour. Low values of  $n$  spuriously interpret structure from the random sampling noise – see Section 6 for further discussion of this phenomena.



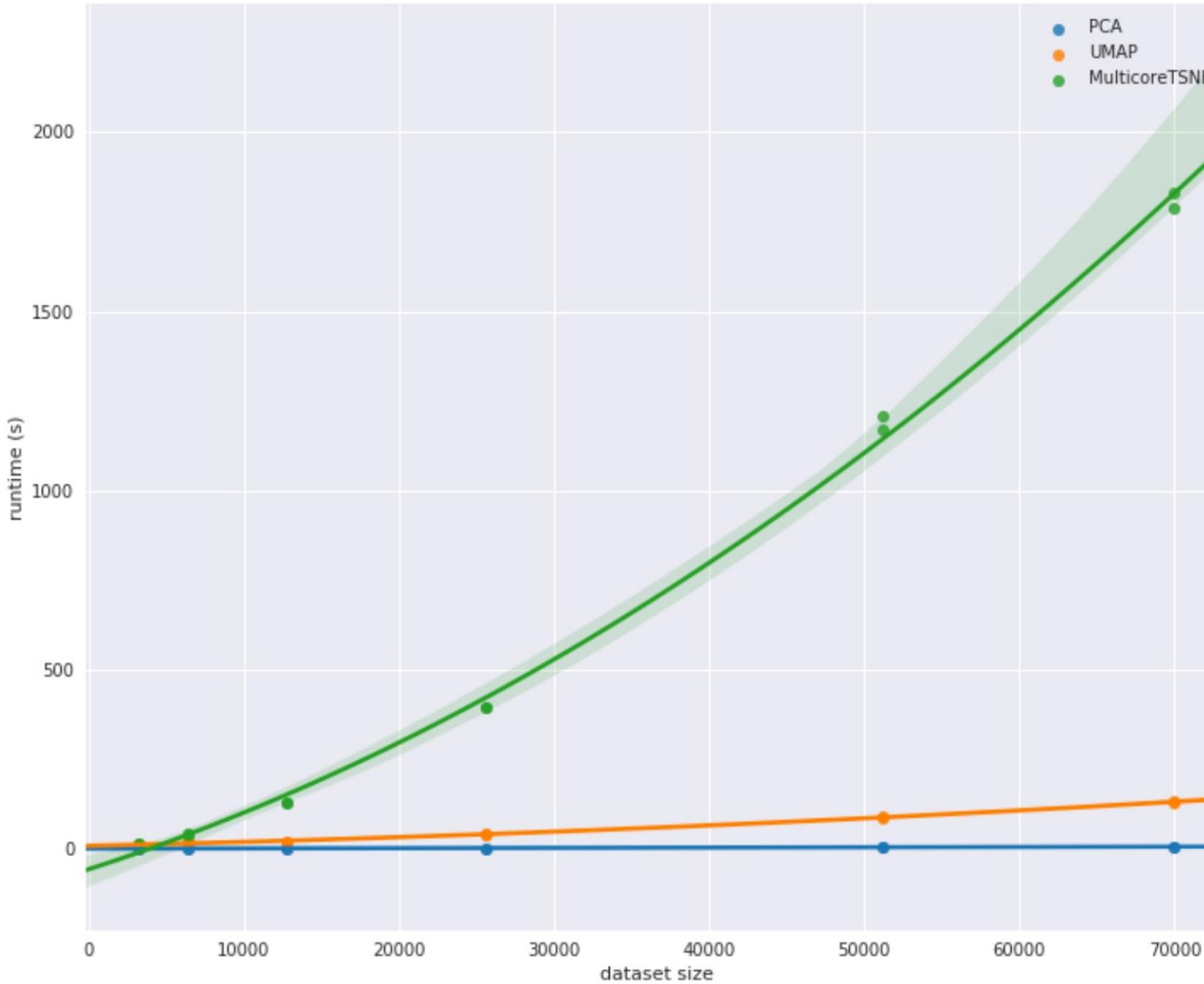
COIL 20: 1440 greyscale images of 20 objects under 72 different rotations spanning 360 degrees.

MNIST: 28x28 pixel grayscale images of handwritten digits 0...9. 70,000 images.

F-MNIST: 28x28 pixel grayscale images of fashion items (clothing, bags etc). 10 classes and 70,000 images.

Word Vectors: 3 million words and phrases from Google News documents and embedded into a 300 dimensional space via word2vec.

<https://arxiv.org/pdf/1802.03426.pdf>



MNIST dataset  
N = 70,000  
P = 784

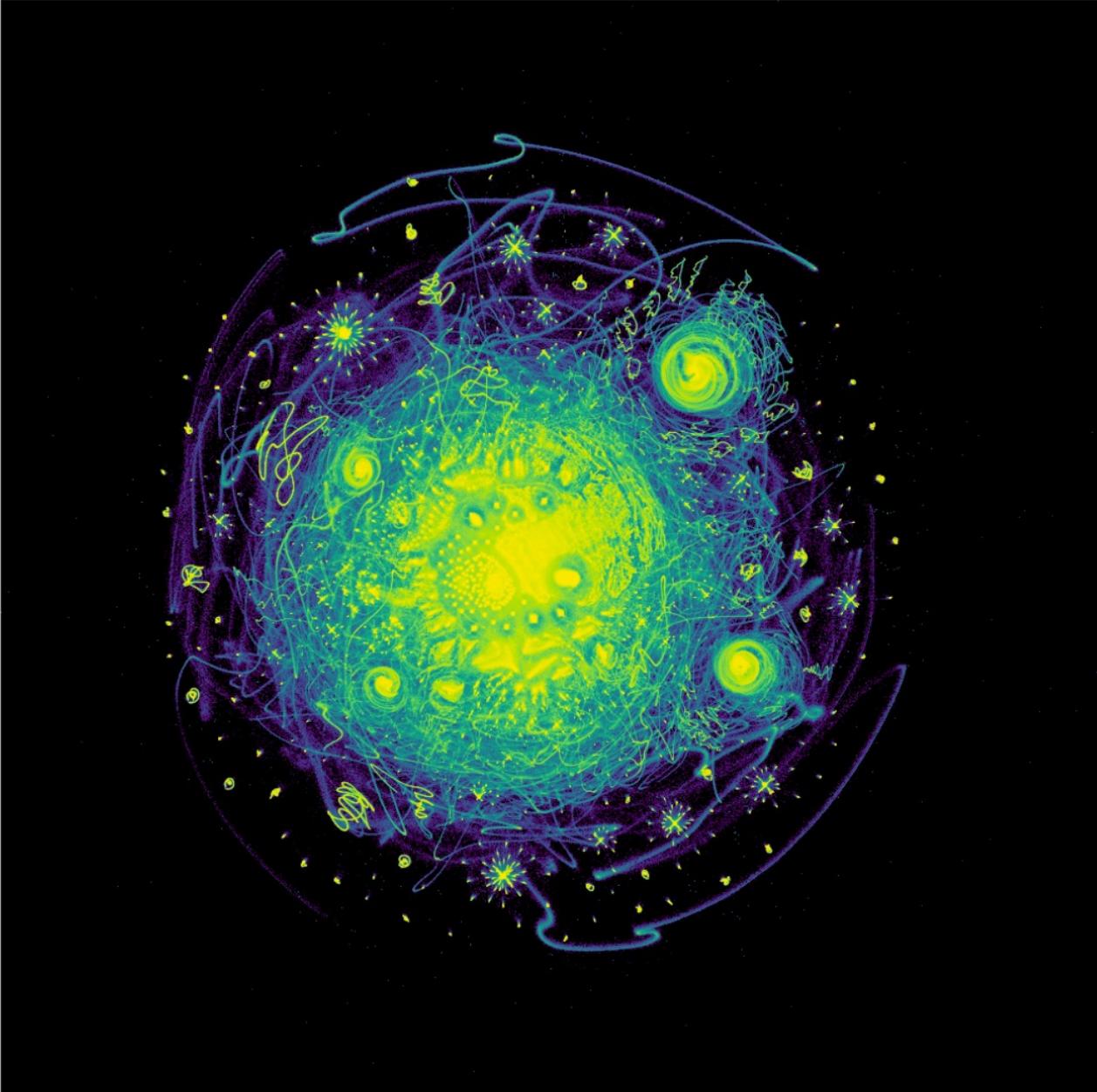


Figure 9: Visualization of 30,000,000 integers as represented by binary vectors of prime divisibility, colored by density of points.

By UMAP

<https://arxiv.org/pdf/1802.03426.pdf>