# STAT 847
# Reading Assignment 3

Chris Binoi Verghese
21092999

**Q1** (1 point): What are two methods used to assess the patterns in the data visually?

The two methods for reducing the complexity of a high-dimensional data space for assessing the patterns visually are:

1. t-SNE (t-distributed stochastic neighbor embedding)
2. UMAP (uniform manifold approximation and projection)

**Q2** (1 point): What's the difference between t-SNE and the more traditional dimension reduction method of PCA?

PCA is a linear approach to dimension reduction, relying on the addition of linear combinations of weighted raw feature values to find a lower-dimensional data space. In contrast, t-SNE utilizes a nonlinear combination of raw input features, providing a distinct treatment of data to uncover patterns in a lower-dimensional space.

**Q3** (1 point): What's an important difference between t-SNE and UMAP?

t-SNE is a probabilistic approach done by assuming that similarities equate to probabilities based on the overlaying of t-distributions in the lower-dimensional projection process.
On the other hand, UMAP avoids the notion of probabilistic similarities and is a graph-based approach that relies only on smoothed distance metrics.

**Q4** (1 point): How is the difference between the full and the reduced data measured when using t-SNE?

The dissimilarity between the original high-dimensional (full) data and the new reduced lower-dimensional representation is quantified using the Kullback–Leibler (KL) divergence. Low KL-divergence indicates highly similar distributions, while high values suggest divergence between the distributions.

**Q5** (2 point): Would running a t-SNE twice give exactly the same results twice? Why or why not?

No, because t-SNE involves a step where a new randomly distributed low-dimensional version of the dataset is created during each run, after which a new set of probabilities between all observations in this version is calculated for each iteration. The algorithm plots data randomly in a lower-dimensional setting, leading to different initializations and thus distinct results across multiple runs.

**Q6** (1 point): How does increasing the perplexity change the output of the t-SNE?

Increasing perplexity widens the distribution around each point, resulting in a more global solution. This wider distribution allows points potentially far apart in space to be assigned a higher probability of proximity to the distribution's center, fostering greater similarity between distant points.
Therefore, increasing the size of perplexity allows for a greater probability that distant points are treated as more similar.

**Q7** (2 points): In UMAP, what are the two searches used for learning the high-dimensional structure of the data?

There are two searches in the higher-dimensional space of the data:

1) between points i.e. local connection in a pointwise fashion across the full data space
2) around regions of points i.e. in the ambient space, achieved by taking the nerve of the now-learned higher-dimensional manifold.

**Q8** (1 point): How are the hyperparameters in t-SNE and in UMAP set to provide an ideal picture of the data.

There is no formal rule for selecting optimal hyperparameter values, however it is recommended adjust values until patterns stabilize. This improvement can be achieved through a grid search, providing several visualizations in a single plot to observe shifts across different hyperparameter values.