# 1 Intro - Clustering

**Clustering is the process of identifying 'similar' groups of data. The metric by which similarity is measured can vary, e.g. Euclidean distance, density, cosine similarity (to name a few.)**

## 1.1 K-Means Clustering

Groups $N$ data points into $K$ clusters by Euclidean distance (l2 norm).

- Initialise $\mu_k$ randomly for all $k \in K$
- For each $x_n$:
  - for $k = argmin_k ||x_n - \mu_k||$, $r_{nk} = 1$
- For each $k \in K$:
  - $\mu_k = \frac{\sum^N r_{nk} x_n}{\sum^N r_{nk}}$

# 2 Gaussian Mixture Models

## 2.1 Model definition

**Imagine now that points could partially belong to multiple clusters; a soft assignment. The probability of observing data point $x_n$ in cluster $k$ is proportional to its assignment weighting.** One way to do this is to assume a Gaussian distribution for the points within each cluster, and a multinomial distribution for the clusters themselves. Using a generative process:

1. First pick a cluster by rolling a dice (with parameter $\phi$)
2. Second generate a data point for this cluster based on its distribution (parameters $\mu_k$, $\Sigma_k$)

$$p(x_n, C_k) = \phi_k N(x_n; \mu_k, \Sigma_k)$$

- Because we do not know what cluster $x_n$ was actually generated from, we need to sum over the marginal probability to obtain the distribution. $C_k$ is, in effect, a 'latent' ('hidden') variable.
- for convention sake let us denote $z_n k = C_k$ for data point $x_n$
- So our probability distribution becomes:

$$p(x_n) = \sum^K \phi_k N(x_n; \mu_k, \Sigma_k)$$

- The log likelihood is:

$$ln(\mathcal{L}) = \sum^N ln\big(\sum^K \phi_k N(x_n; \mu_k, \Sigma_k)\big)$$

- (note the sum inside the log; it makes solving a bit tougher having a latent variable)

## 2.2 MLE Parameter Estimates

$$\mu_k = \frac{1}{N_k} \sum^k \gamma(z_{nk}) x_n \tag{1}$$

$$\Sigma_k = \frac{1}{N_k} \sum^k \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \tag{2}$$

$$\phi_k = \frac{N_k}{N} \tag{3}$$

where:

$$N_k = \sum^N \gamma(z_{nk}) \tag{4}$$

These all intuitively make sense for the MLE estimates; Considering $\gamma(z_{nk})$ as the portion of $x_n$ that belongs to cluster $k$. $\mu_k$ and $\Sigma_k$ estimates take this into account. $\phi_k$ is the relative fraction of each clusters mass.