- Goal: Structure documents by topic

--- Discover groups (clusters) of related articles
--- No labels provided uncover cluster structure from input alone
Input: docs as vectors x_i
Output: cluster labels z_i.

- What defines a cluster?

--- Cluster defined by center & shape/spread.
--- Assign observation x_i(doc) to cluster k (topic label) if: 1) Score under cluster k is higher than under others; 2) For simplicity, often define score as distance to cluster center (ignoring shape).

- K-means algorithm

0) Initialize cluster centers: $\mu_1, \mu_{2,} \ldots, \mu_k$.
1) Assign observations to closest cluster center: $z_i \leftarrow argmin_j \left\|\mu_j - x_i\right\|_2^2$
2) Revise cluster c enters as mean of assigned observations: $\mu_j = \frac{1}{n_j}\sum_{i;\, z_i=j} x_i$
3) Repeat 1+2, until convergence.

- A coordinate descent algorithm

1) Assign observations to closest cluster center:

$$z_i \leftarrow \arg min_j \left\|\mu_j - x_i\right\|_2^2$$

2) Revise cluster centers as mean of assigned observations:

$$\mu_j = \frac{1}{n_j}\sum_{i:\, z_i=j} x_i$$

$$\mu_j \leftarrow \arg min_\mu \sum_{i:\, z_i=j} \|\mu - x_i\|_2^2$$

- Converges to local optimum

--- k-means is very sensitive to initialization and you can actually get very crazy solutions.

- K-means++ overview:

--- Initialization of k-means algorithm is critical to quality of local optima found.
--- Smart initialization:
1) Choose first cluster center uniformly at random from data points;
2) For each observation x_i, compute distance d(x) to nearest cluster center;
3) Choose new cluster center from amongst data points, with probability of x being chosen proportional to $d(x)^2$;
4) Repeat Steps 2 and 3, until k centers have been chosen.

- K-means++ pros/cons

--- Computationally costly relative to random initialization, but the subsequent k-means often converges more rapidly.
--- Tends to improve quality of local optimum and lower runtime.

- K-means objective

--- k-means is trying to minimize the sum of squared distances:

$$\sum_{j=1}^{k} \sum_{i:\, z_i=j} \left\| \mu_j - x_i \right\|_2^2$$

- What happens as k increases?

--- Can refine clusters more and more to the data – Overfitting!