Machine Learning for Data Science (CS4786)
Lecture 3

Clustering

Course Webpage :

http://www.cs.cornell.edu/Courses/cs4786/2017fa/

Clustering Criterion

- Minimize total within-cluster dissimilarity

$$M_1 = \sum_{j=1}^{K} \sum_{s,t \in C_j} \text{dissimilarity}(x_t, x_s)$$

- Maximize between-cluster dissimilarity

$$M_2 = \sum_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \text{dissimilarity}(x_t, x_s)$$

- Maximize smallest between-cluster dissimilarity

$$M_3 = \min_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \text{dissimilarity}(x_t, x_s)$$

- Minimize largest within-cluster dissimilarity

$$M_4 = \max_{j \in [K]} \max_{s,t \in C_j} \text{dissimilarity}(x_t, x_s)$$

Clustering Criterion

- minimizing $M_1 \equiv$ maximizing M_2

Lets Build an Algorithm

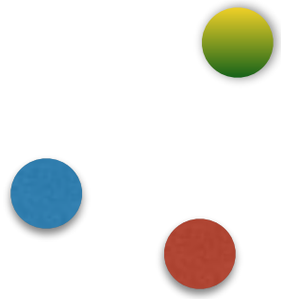
$$M_3 = \min_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \text{dissimilarity}(x_t, x_s)$$

Single Link Clustering

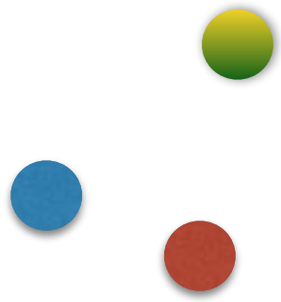
- Initialize n clusters with each point $\mathbf{x}_t$ to its own cluster
- Until there are only K clusters, do
  1. Find closest two clusters and merge them into one cluster

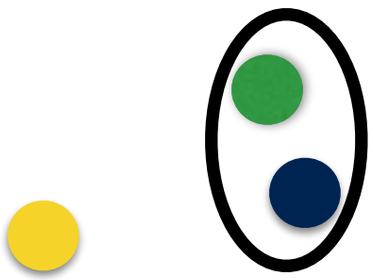
$$\text{dissimilarity}(C_i, C_j) = \min_{t \in C_i, s \in C_j} \text{dissimilarity}(\mathrm{x}_t, \mathrm{x}_s)$$

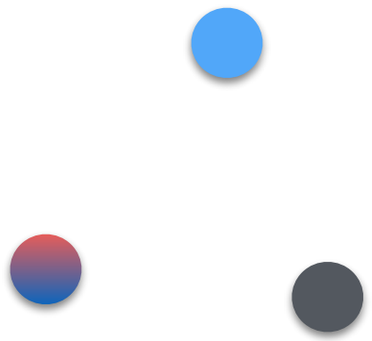
Demo



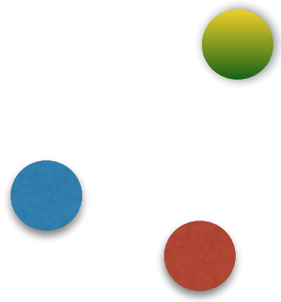
Demo

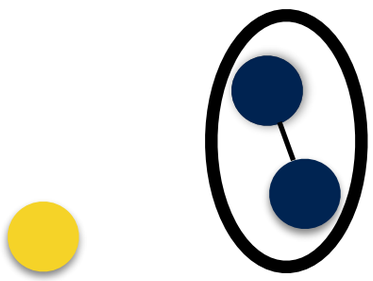


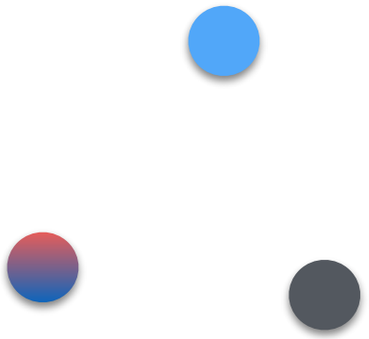




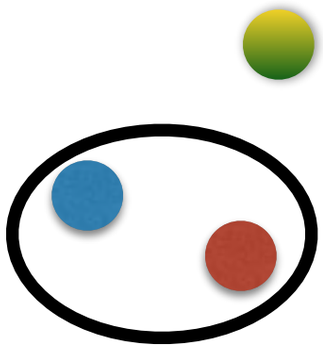
Demo

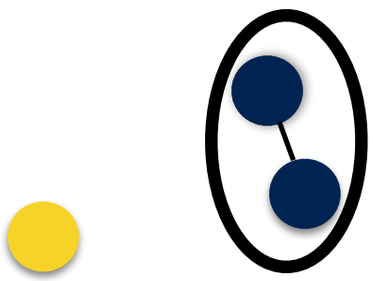


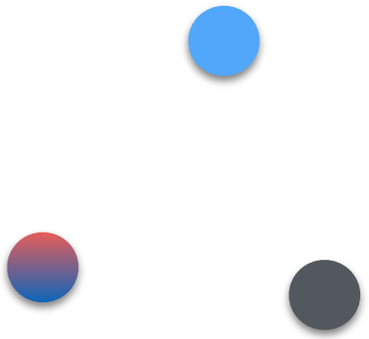




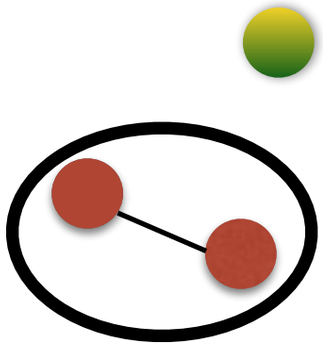
Demo

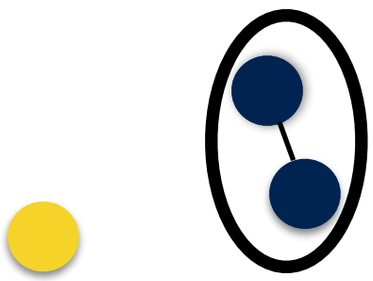


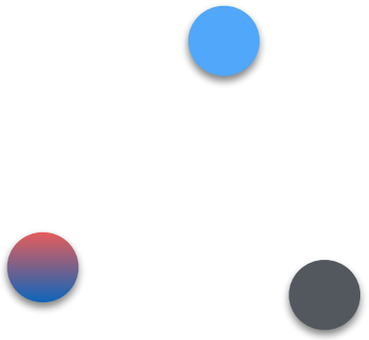




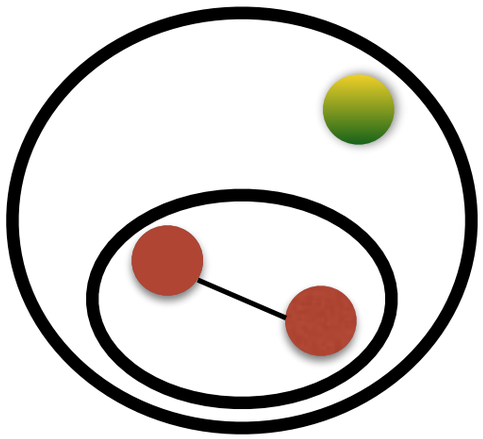
Demo

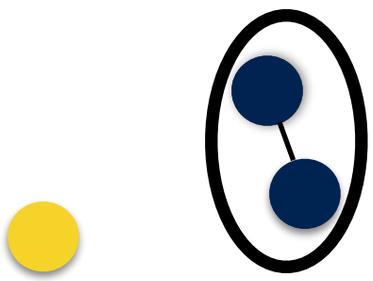


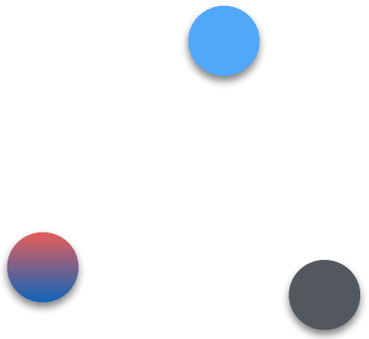




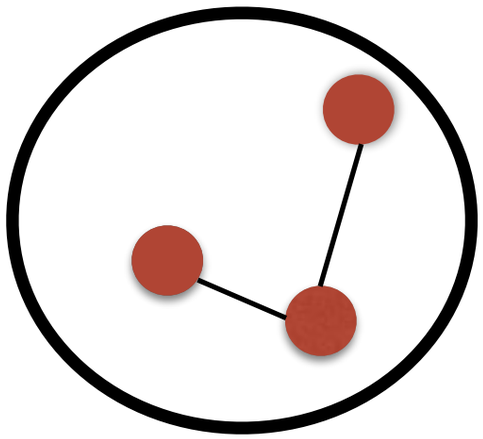
Demo

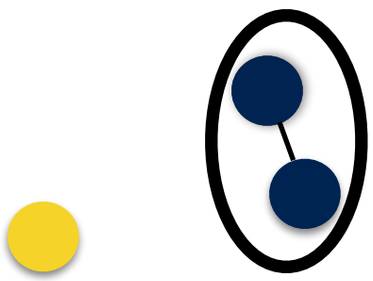


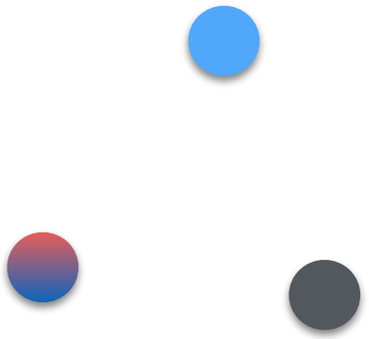




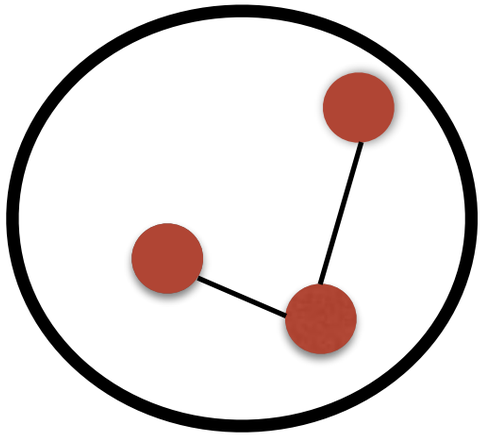
Demo

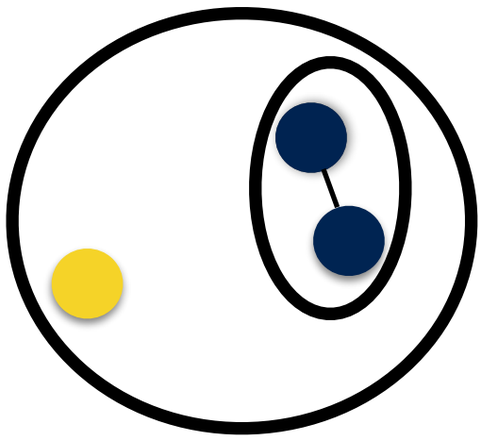


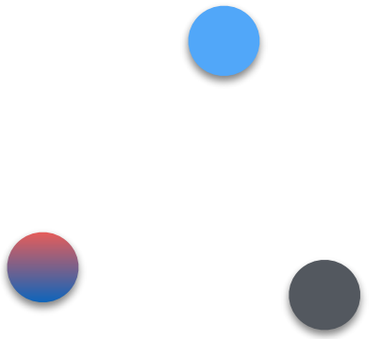




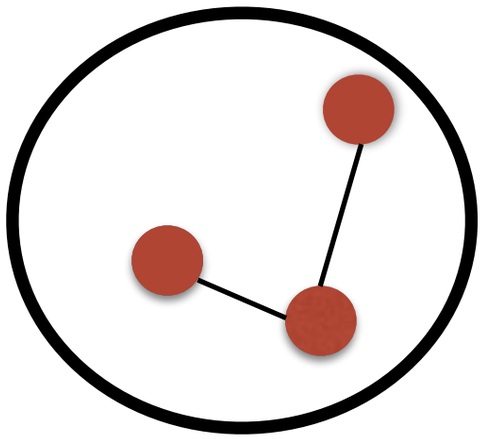
Demo

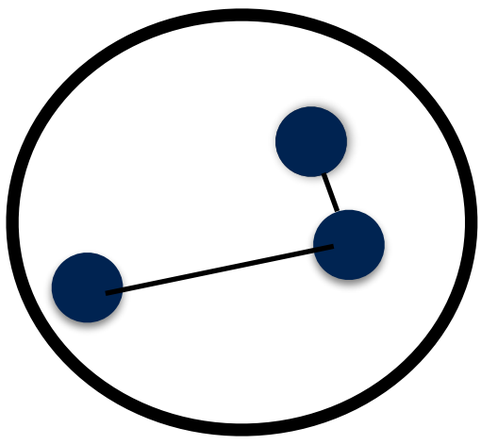


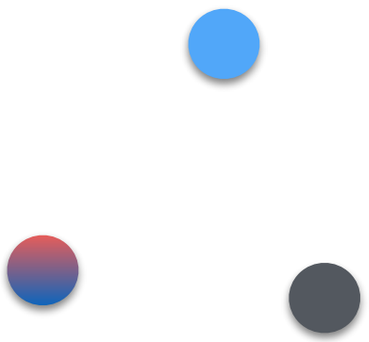




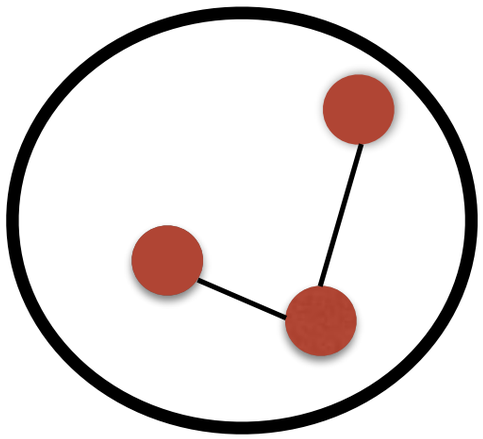
Demo

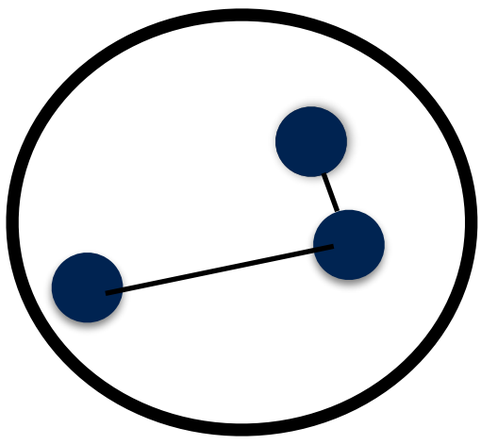


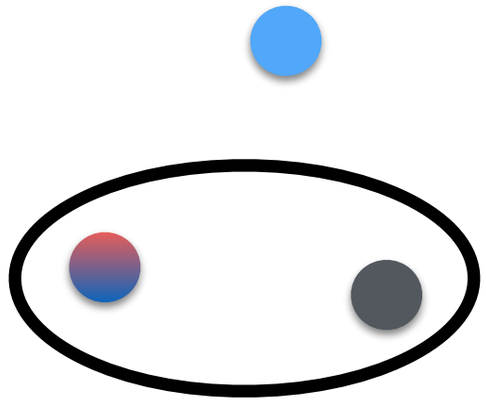




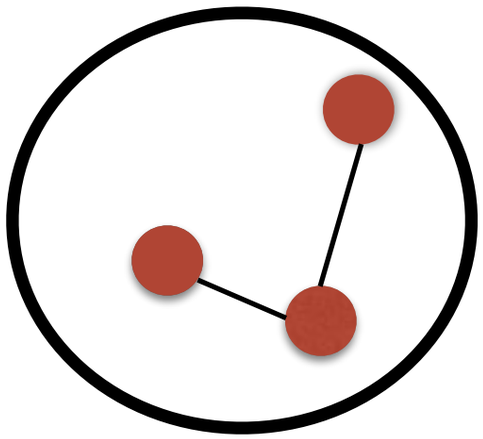
Demo

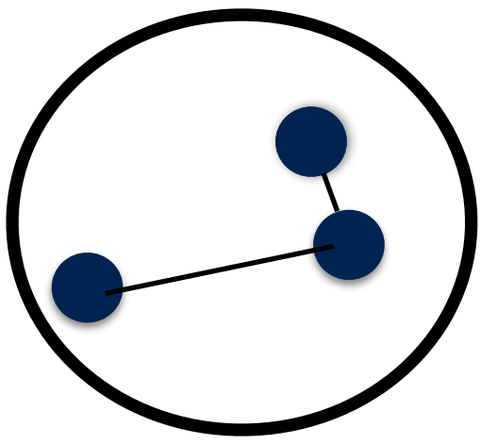


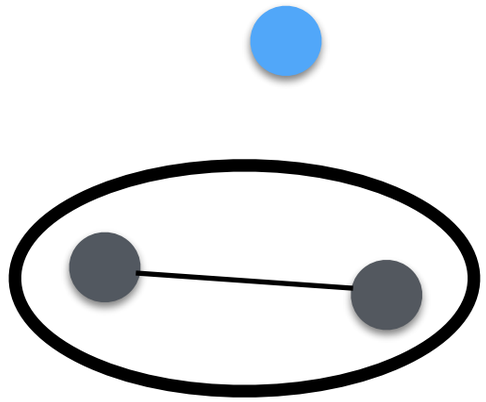




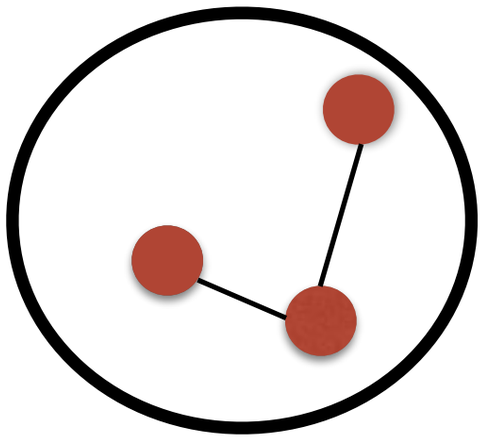
Demo

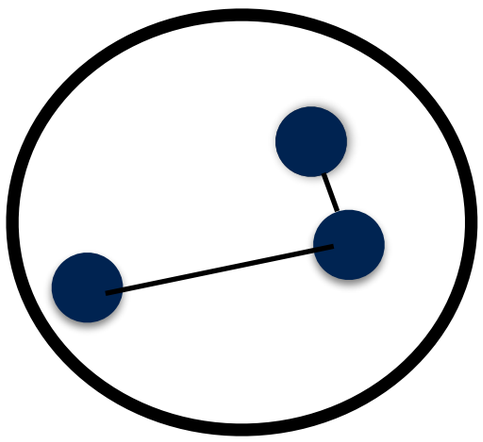


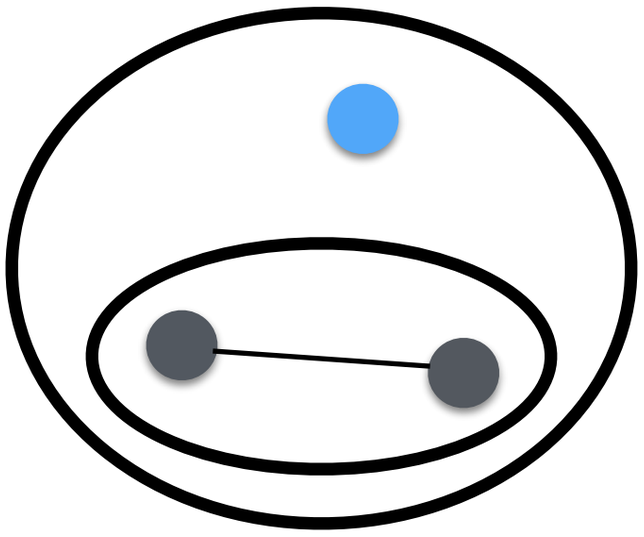




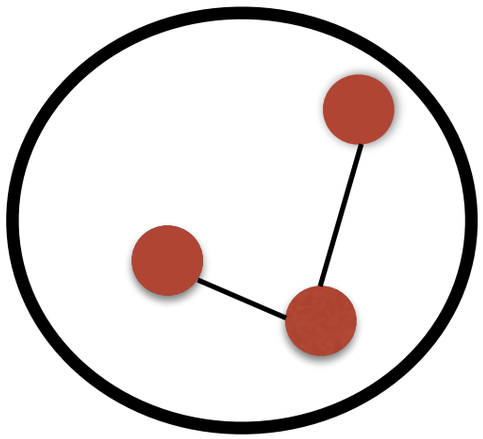
Demo

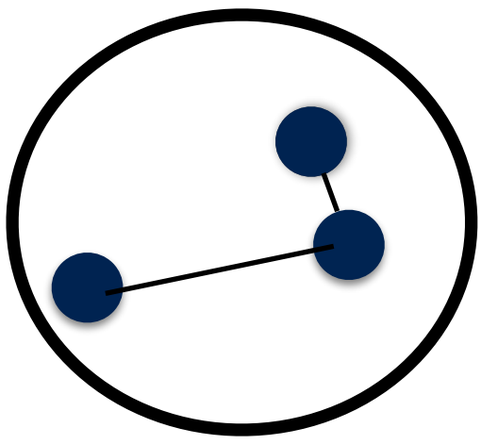


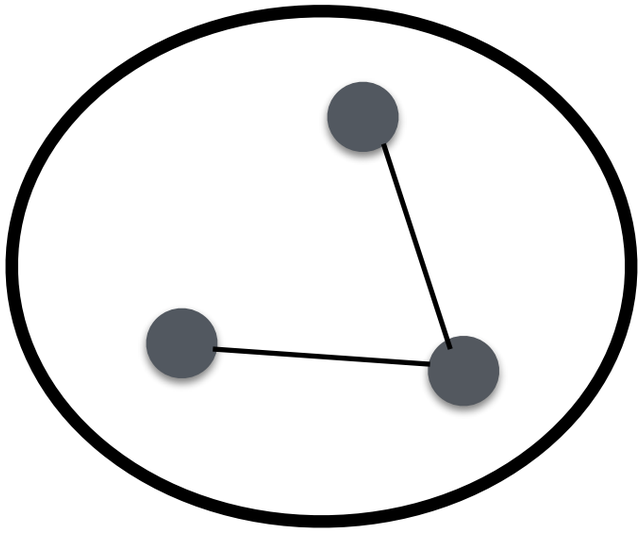




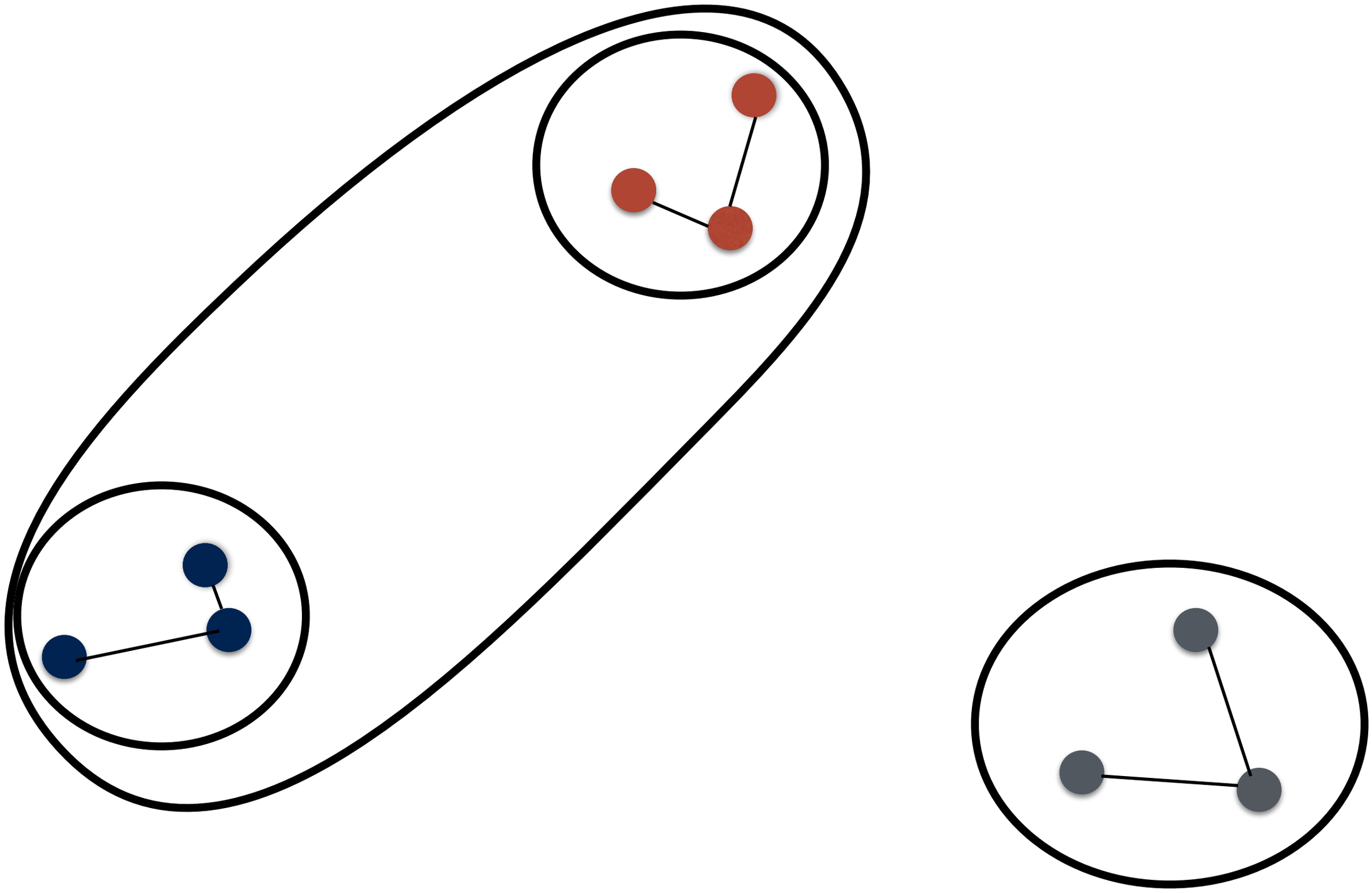
Demo



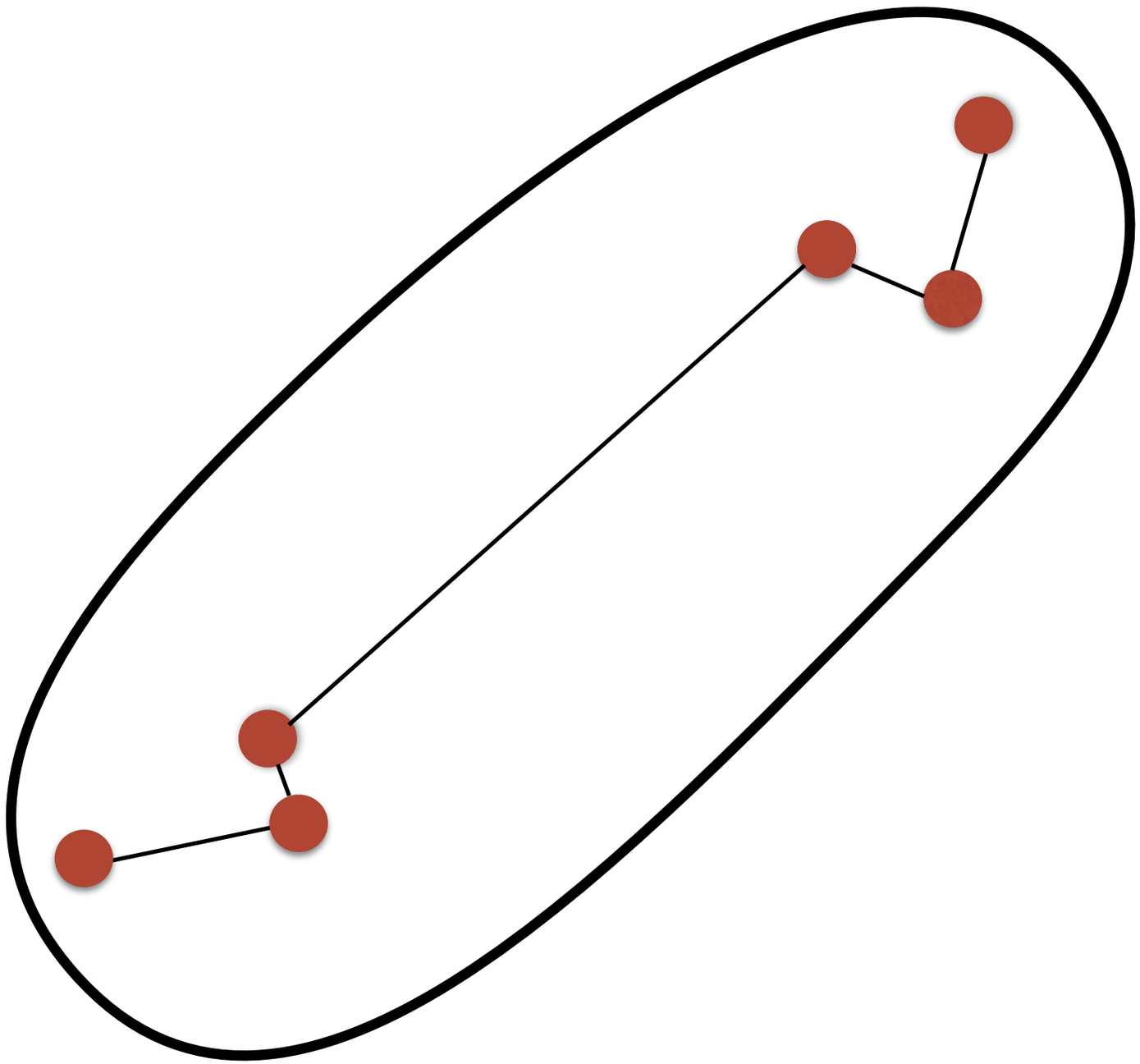


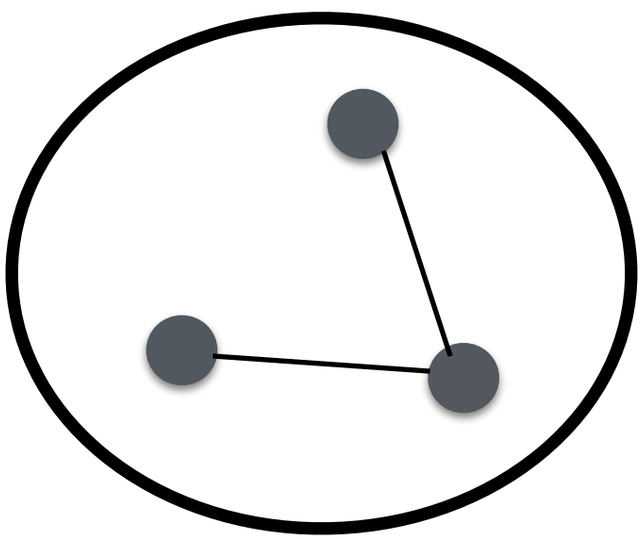


Demo

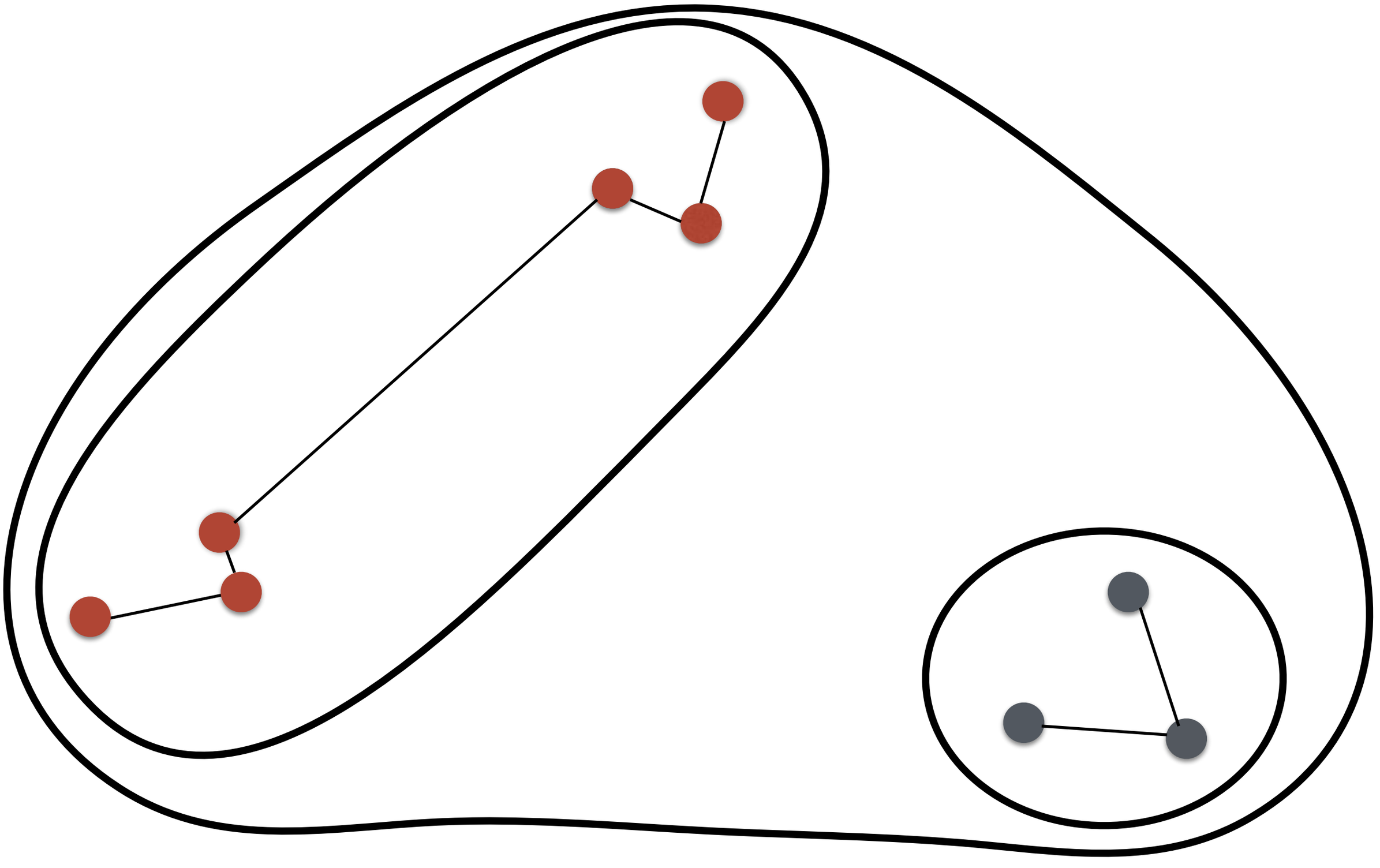


Demo

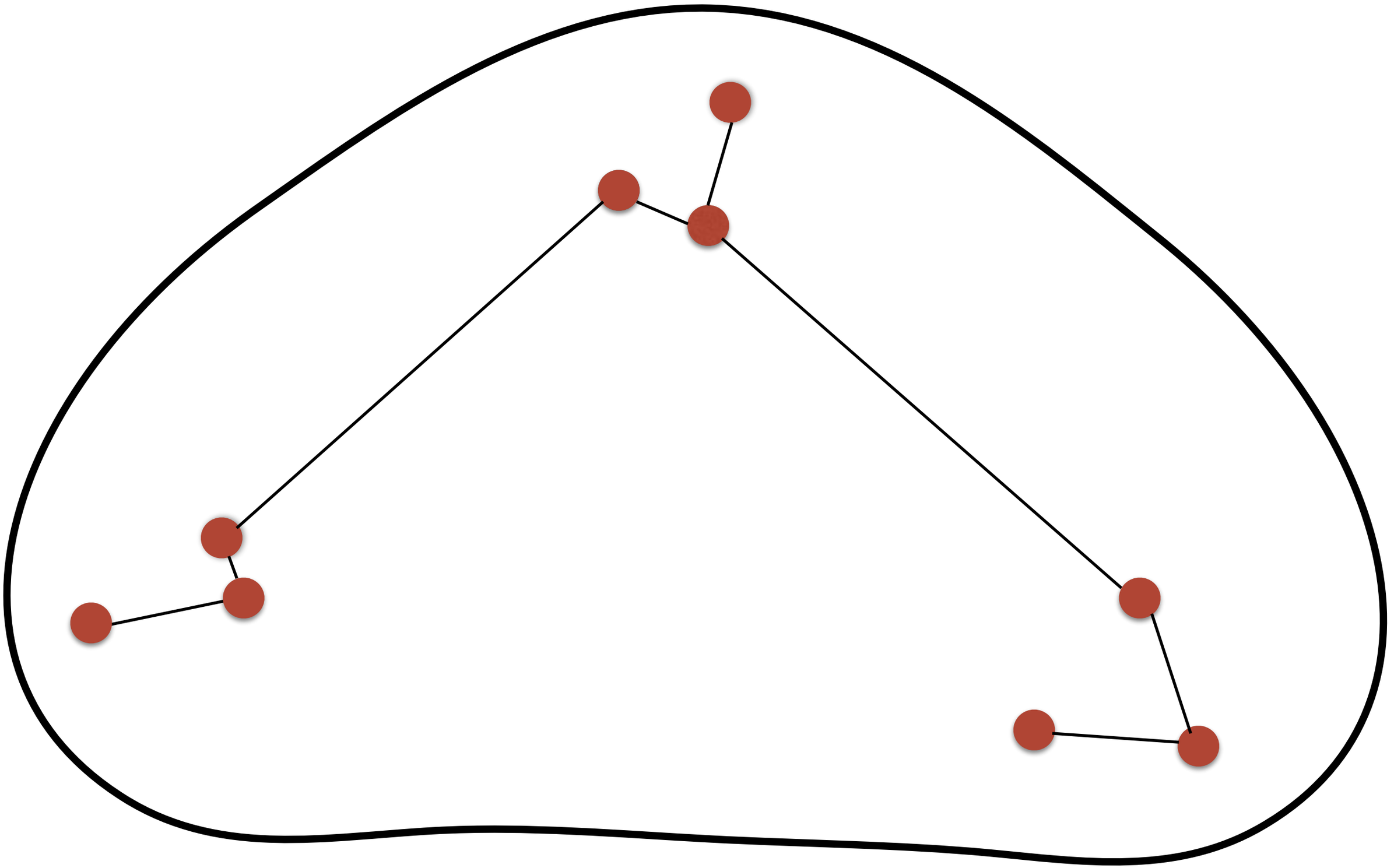




Demo



Demo



Single Link Objective

Objective for single-link:

$$M_3 = \min_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \text{dissimilarity}(x_t, x_s)$$

Single link clustering is optimal for above objective!

Single Link Objective

Proof:

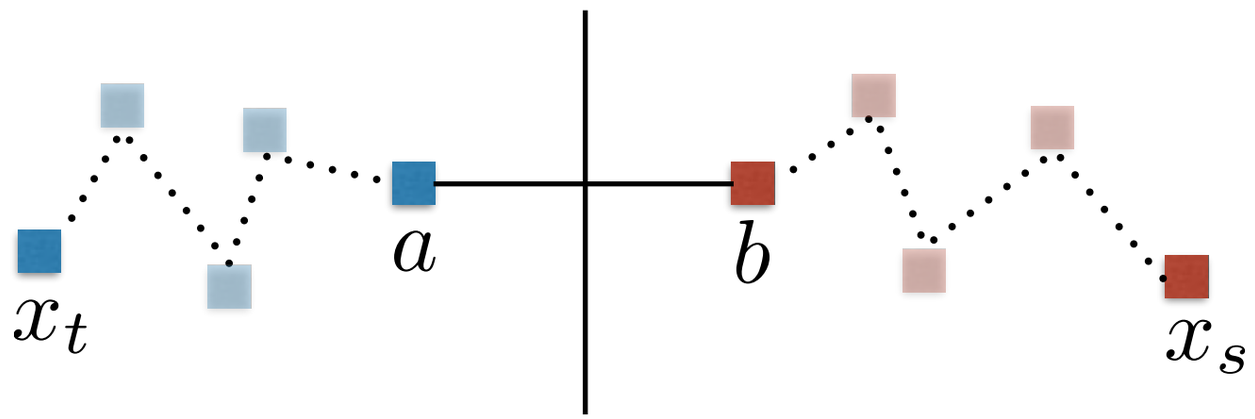
Say c is solution produced by single-link clustering

Key observation:

$$\min_{t,s:c(x_i)\neq c(x_j)} \text{dissimilarity}(x_i, x_j) > \text{Distance of points merged (edges on the tree)}$$

Say $c' \neq c$ then,

$$\exists\ t, s \text{ s.t. } c'(x_t) \neq c'(x_s) \text{ but } c(x_t) = c(x_s)$$



c' boundary

Clustering Criterion

- Minimize within cluster average dissimilarity

$$M_6 = \sum_{j=1}^{K} \sum_{s \in C_j} \text{dissimilarity}\left(\mathbf{x}_s, C_j\right)$$

$$= \sum_{j=1}^{K} \sum_{s \in C_j} \left(\frac{1}{|C_j|} \sum_{t \in C_j, t \neq s} \text{dissimilarity}\left(\mathbf{x}_s, \mathbf{x}_t\right) \right)$$

$$= \sum_{j=1}^{K} \frac{1}{|C_j|} \sum_{s \in C_j} \left(\sum_{t \in C_j, t \neq s} \left\| \mathbf{x}_s - \mathbf{x}_t \right\|_2^2 \right)$$

- Minimize within-cluster variance: $\mathbf{r}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$

$$M_5 = \sum_{j=1}^{K} \sum_{t \in C_j} \left\| \mathbf{x}_t - \mathbf{r}_j \right\|_2^2$$

Clustering Criterion

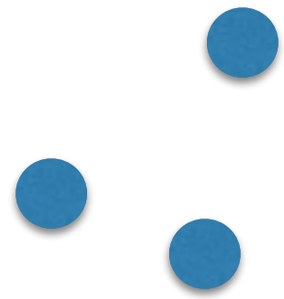
- minimizing $M_5 \equiv$ minimizing M_6

Lets build an Algorithm

$$M_5 = \sum_{j=1}^{K} \sum_{t \in C_j} \left\| \mathbf{x}_t - \mathbf{r}_j \right\|_2^2$$

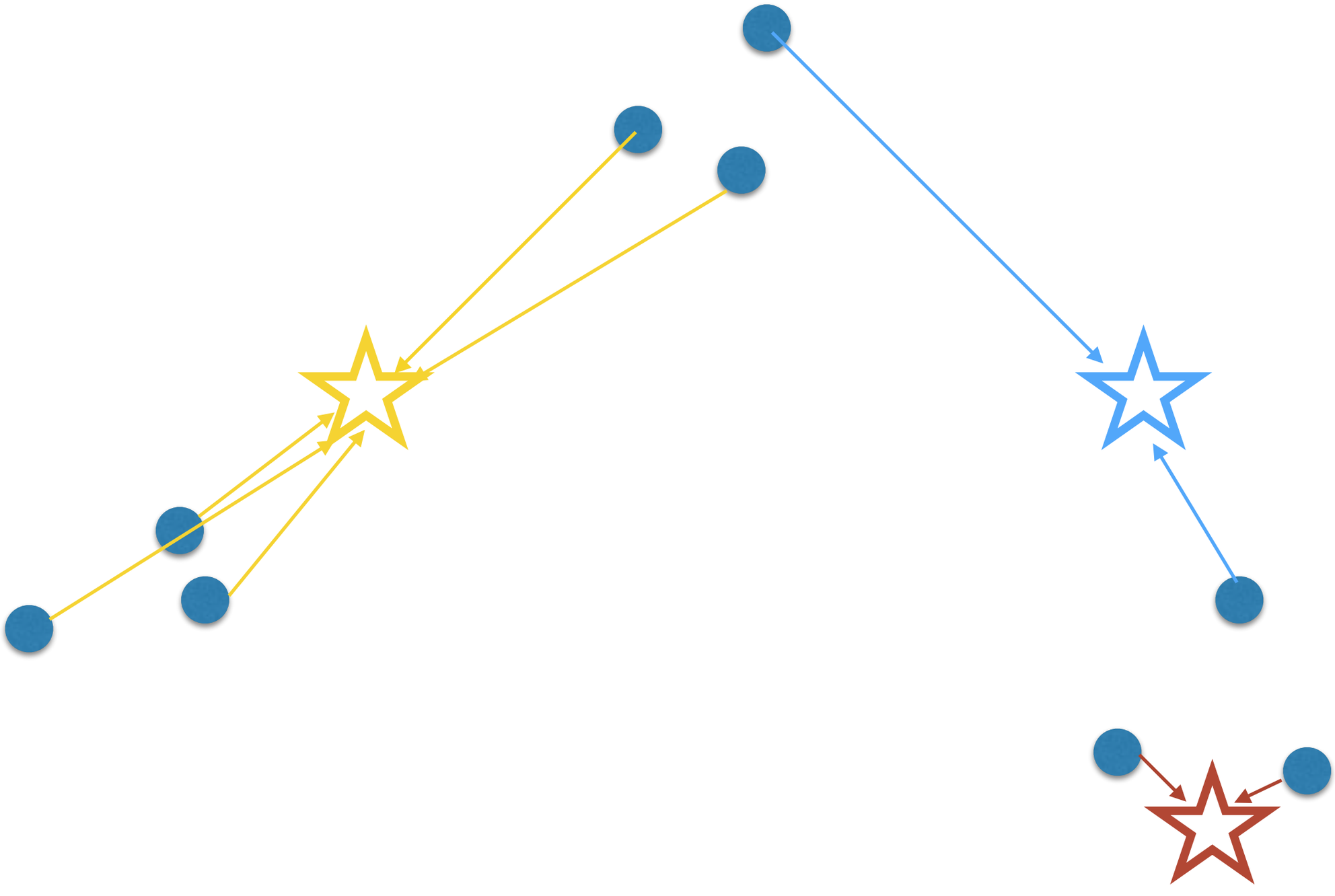
$$\text{where } \mathbf{r}_j = \frac{1}{|C_j|} \sum_{t \in C_j} \mathbf{x}_t$$

Demo



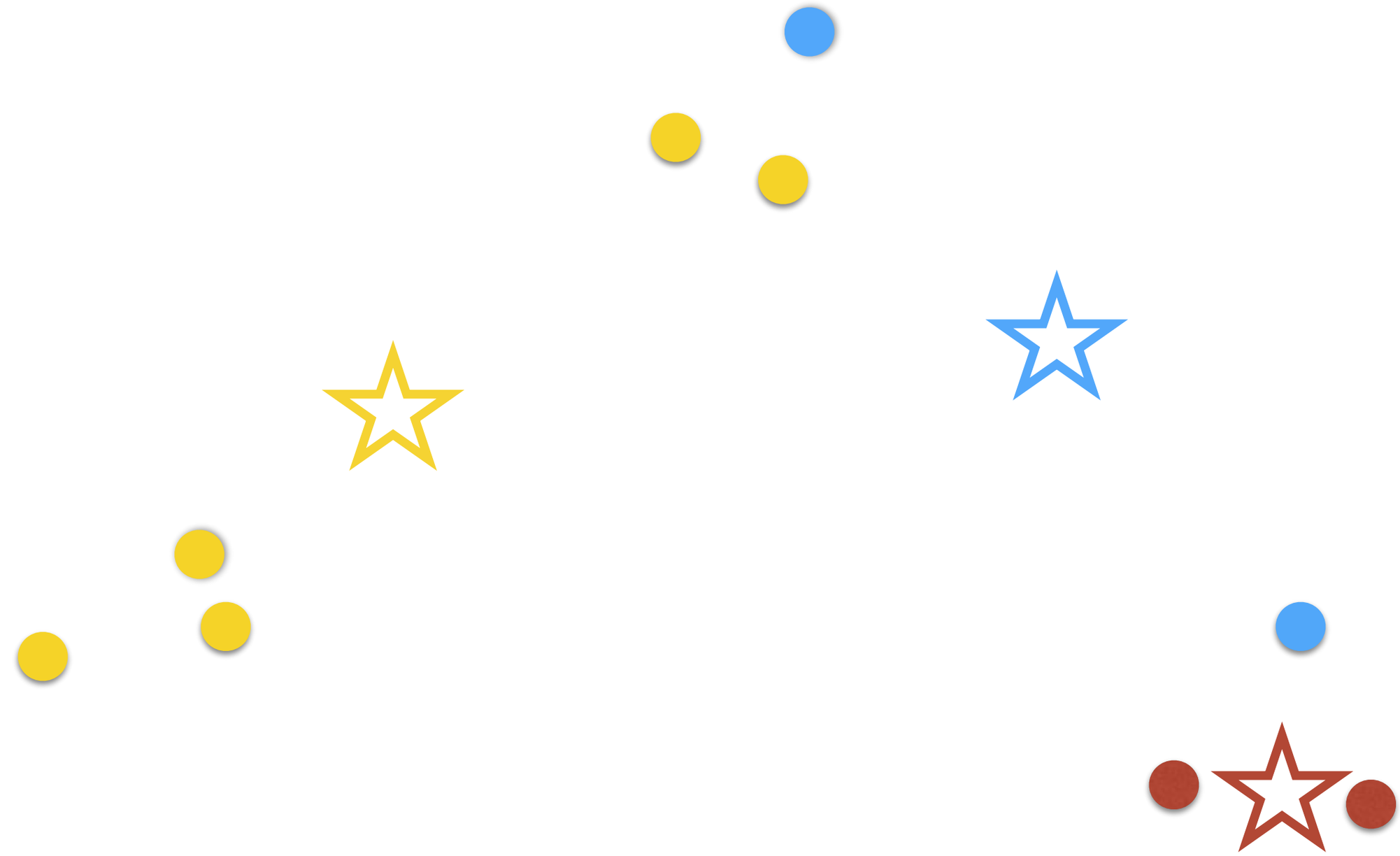
Demo

Demo

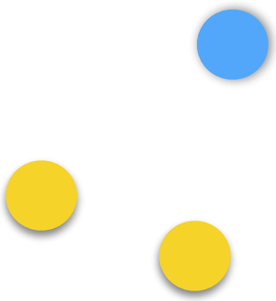


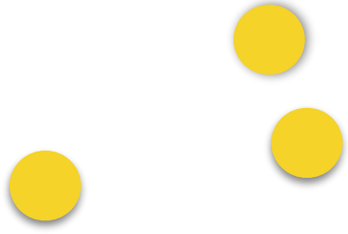
Demo

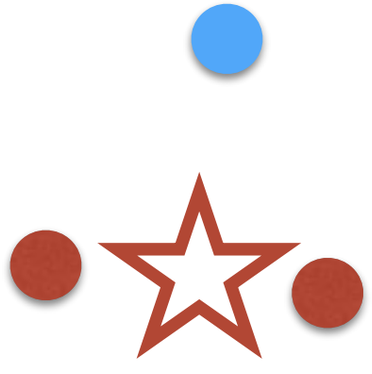
Demo



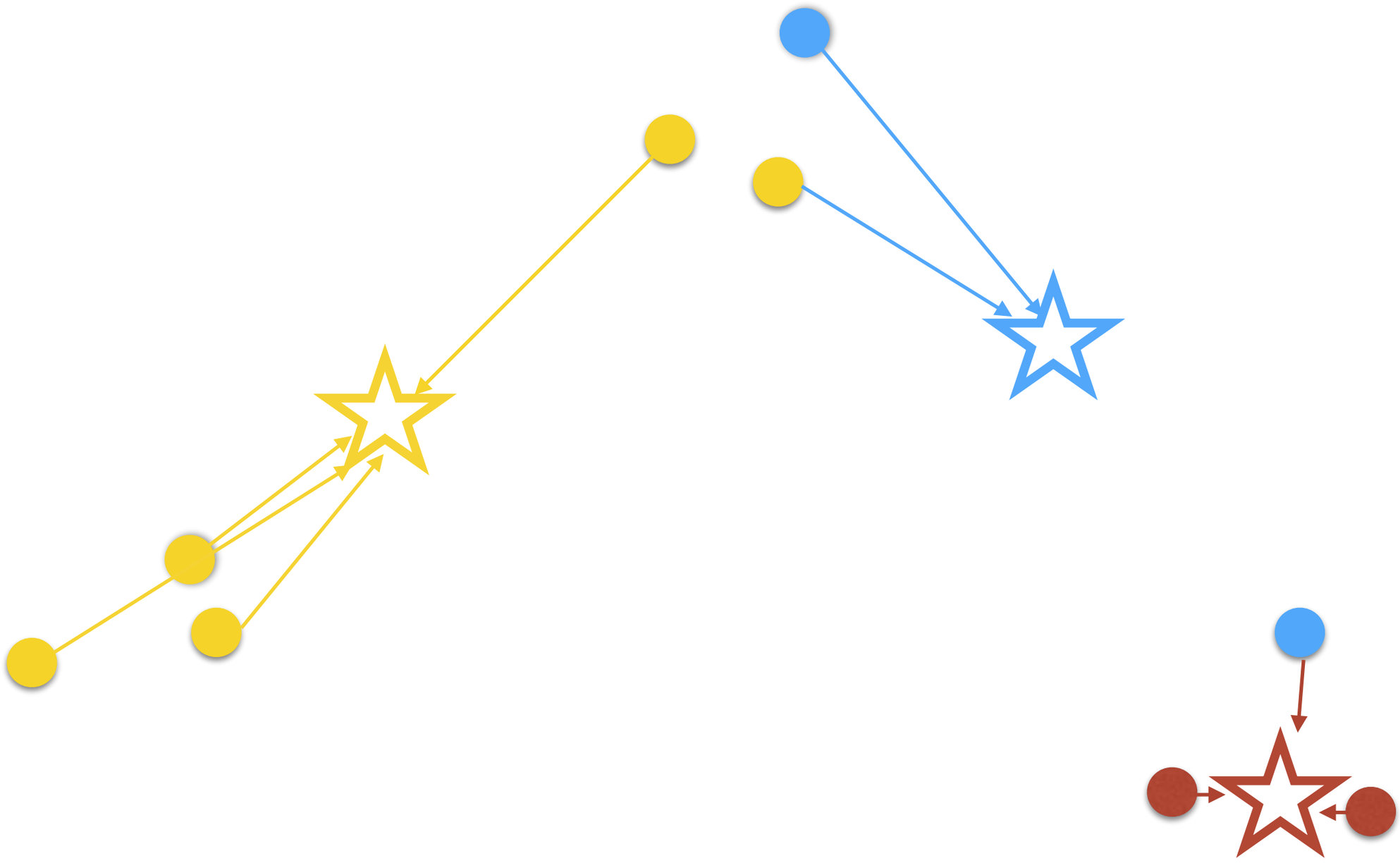
Demo



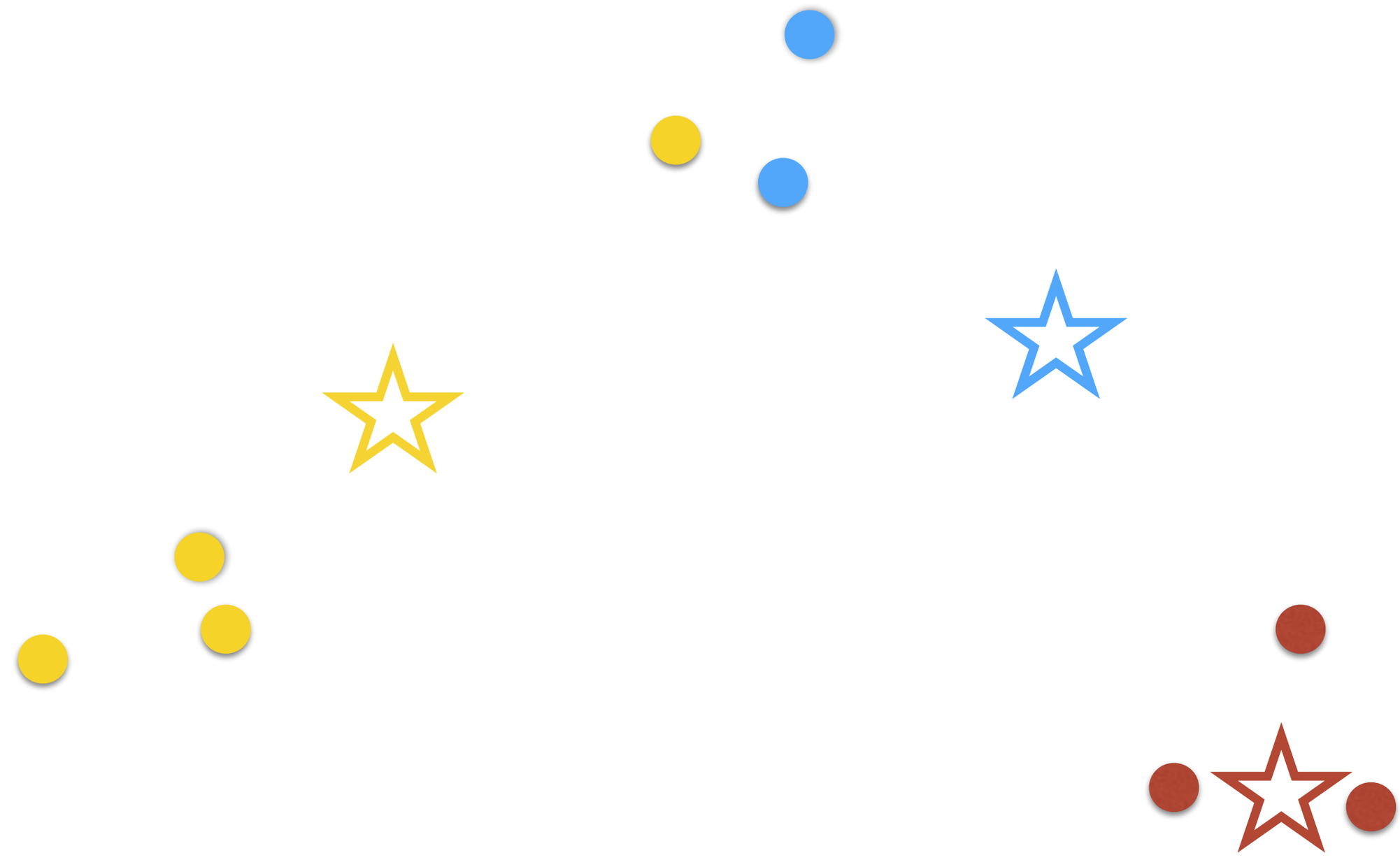




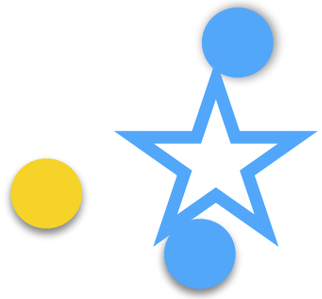
Demo

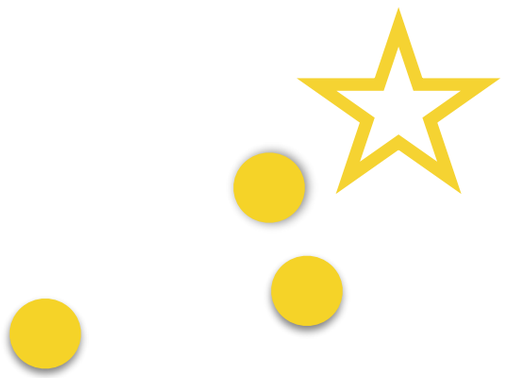


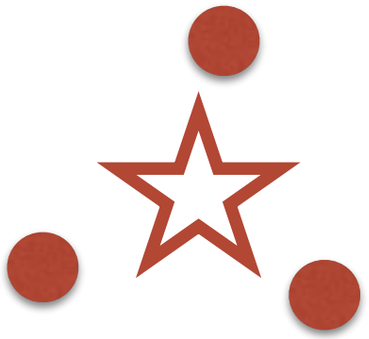
Demo



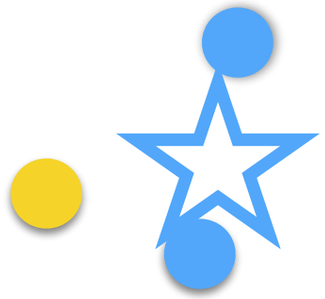
Demo

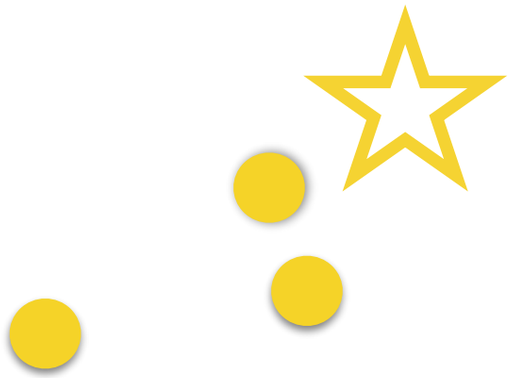


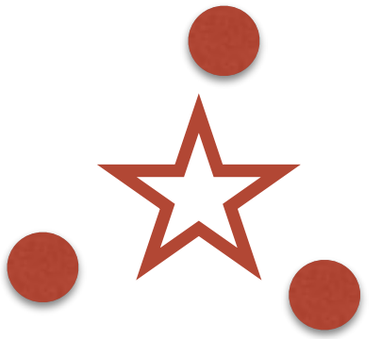




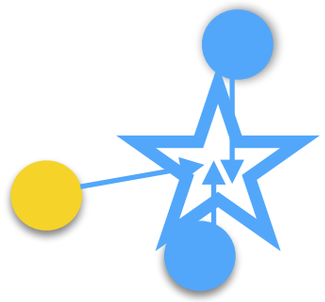
Demo

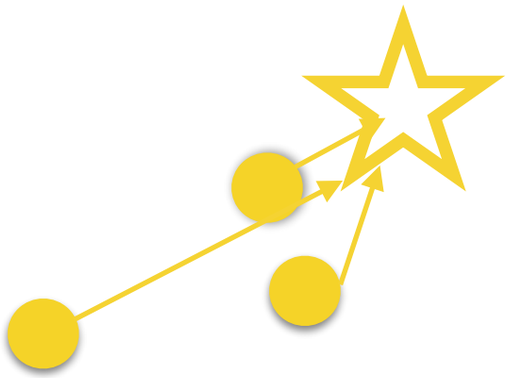


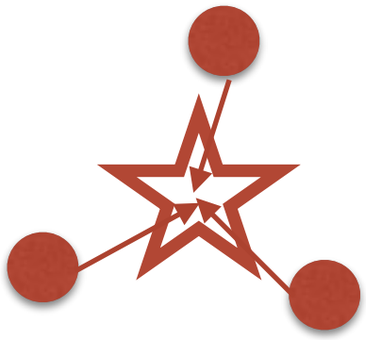




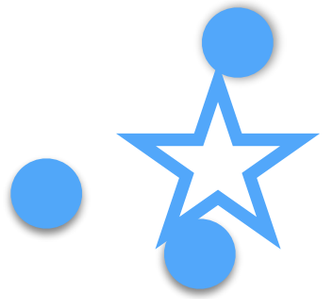
Demo

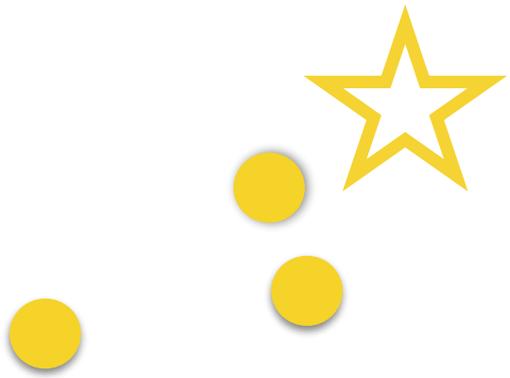


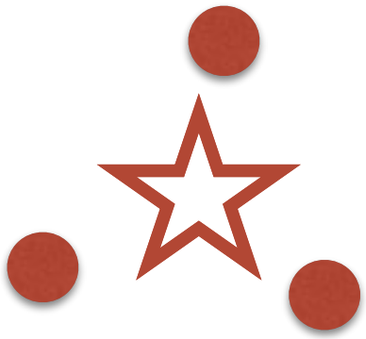




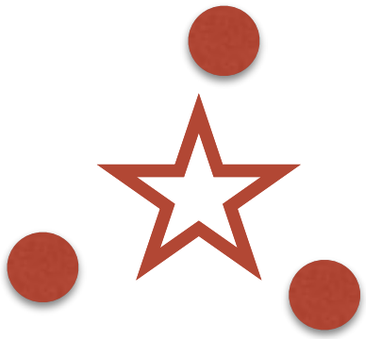
Demo



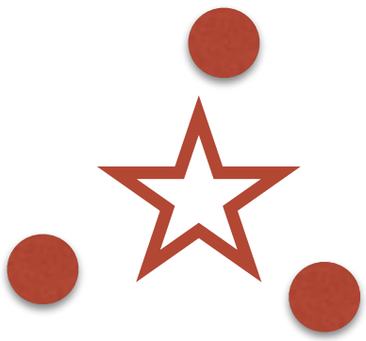




Demo



Demo



K-means Clustering

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \left\| \mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1} \right\|$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t$$

  3. $m \leftarrow m + 1$

K-means objective

$$\sum_{j=1}^{K} \sum_{t \in C_j} \left\| \mathbf{x}_t - \frac{1}{|C_j|} \sum_{s \in C_j} \mathbf{x}_s \right\|^2 = \min_{\mathbf{r}_1,\ldots,\mathbf{r}_K} \sum_{j=1}^{K} \sum_{t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|^2$$

$$\| \qquad\qquad\qquad\qquad \|$$

$$M_5 = \min_{\mathbf{r}_1,\ldots,\mathbf{r}_K} O(c; \mathbf{r}_1, \ldots, \mathbf{r}_K)$$

$$O\left(c; \mathbf{r}_1, \ldots, \mathbf{r}_K\right) = \sum_{j=1}^{K} \sum_{c(\mathbf{x}_t)=j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2$$

Minimize above objective over c and $\mathbf{r}_1, \ldots, \mathbf{r}_K$

Fact: Centroid is Minimizer

$$\forall \mathbf{r}_j, \quad \sum_{t \in C_j} \left\| \mathbf{x}_t - \frac{1}{|C_j|} \sum_{s \in C_j} \mathbf{x}_s \right\|^2 \leq \sum_{t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|^2$$

Proof

$$\sum_{t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|^2$$

$$= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j + \mu_j - \mathbf{r}_j\|^2 \qquad \mu_j = \frac{1}{|C_j|} \sum_{t \in C_j} \mathbf{x}_t$$

$$= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2 + 2 \sum_{t \in C_j} (\mathbf{x}_t - \mu_j)^\top (\mu_j - \mathbf{r}_j)$$

$$= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2 + 2 \left(\cancel{\sum_{t \in C_j} \mathbf{x}_t} - \cancel{|C_j| \mu_j} \right)^\top (\mu_j - \mathbf{r}_j)$$

$$= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2$$

$$\geq \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2$$

K-means Convergence

- K-means algorithm converges to local minima of objective

$$O\left(c; \mathbf{r}_1, \ldots, \mathbf{r}_K\right) = \sum_{j=1}^{K} \sum_{c(\mathbf{x}_t)=j} \left\|\mathbf{x}_t - \mathbf{r}_j\right\|_2^2$$

- Proof:

Clustering assignment improves objective:

$$O\left(\hat{c}^{m-1}; \mathbf{r}_1^{m-1}, \ldots, \mathbf{r}_K^{m-1}\right) \geq O\left(\hat{c}^{m}; \mathbf{r}_1^{m-1}, \ldots, \mathbf{r}_K^{m-1}\right)$$

(By definition of $\hat{c}^m(\mathbf{x}_t)$)

Computing centroids improves objective:

$$O\left(\hat{c}^{m}; \mathbf{r}_1^{m-1}, \ldots, \mathbf{r}_K^{m-1}\right) \geq O\left(\hat{c}^{m}; \mathbf{r}_1^{m}, \ldots, \mathbf{r}_K^{m}\right)$$

(By the fact about centroid)