

Machine Learning for Data Science (CS4786)

Lecture 4

Clustering

Course Webpage :

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

CLUSTERING CRITERION

- Minimize within cluster average dissimilarity

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

CLUSTERING CRITERION

- Minimize within cluster average dissimilarity

$$\begin{aligned} M_6 &= \sum_{j=1}^K \sum_{s \in C_j} \text{dissimilarity}(\mathbf{x}_s, C_j) \\ &= \sum_{j=1}^K \sum_{s \in C_j} \left(\frac{1}{|C_j|} \sum_{t \in C_j, t \neq s} \text{dissimilarity}(\mathbf{x}_s, \mathbf{x}_t) \right) \end{aligned}$$

CLUSTERING CRITERION

- Minimize within cluster average dissimilarity

$$\begin{aligned} M_6 &= \sum_{j=1}^K \sum_{s \in C_j} \text{dissimilarity}(\mathbf{x}_s, C_j) \\ &= \sum_{j=1}^K \sum_{s \in C_j} \left(\frac{1}{|C_j|} \sum_{t \in C_j, t \neq s} \text{dissimilarity}(\mathbf{x}_s, \mathbf{x}_t) \right) \\ &= \sum_{j=1}^K \frac{1}{|C_j|} \sum_{s \in C_j} \left(\sum_{t \in C_j, t \neq s} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 \right) \end{aligned}$$

CLUSTERING CRITERION

- Minimize within cluster average dissimilarity

$$\begin{aligned} M_6 &= \sum_{j=1}^K \sum_{s \in C_j} \text{dissimilarity}(\mathbf{x}_s, C_j) \\ &= \sum_{j=1}^K \sum_{s \in C_j} \left(\frac{1}{|C_j|} \sum_{t \in C_j, t \neq s} \text{dissimilarity}(\mathbf{x}_s, \mathbf{x}_t) \right) \\ &= \sum_{j=1}^K \frac{1}{|C_j|} \sum_{s \in C_j} \left(\sum_{t \in C_j, t \neq s} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 \right) \end{aligned}$$

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

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

CLUSTERING CRITERION

- minimizing $M_5 \equiv$ minimizing M_6

Lets build an Algorithm

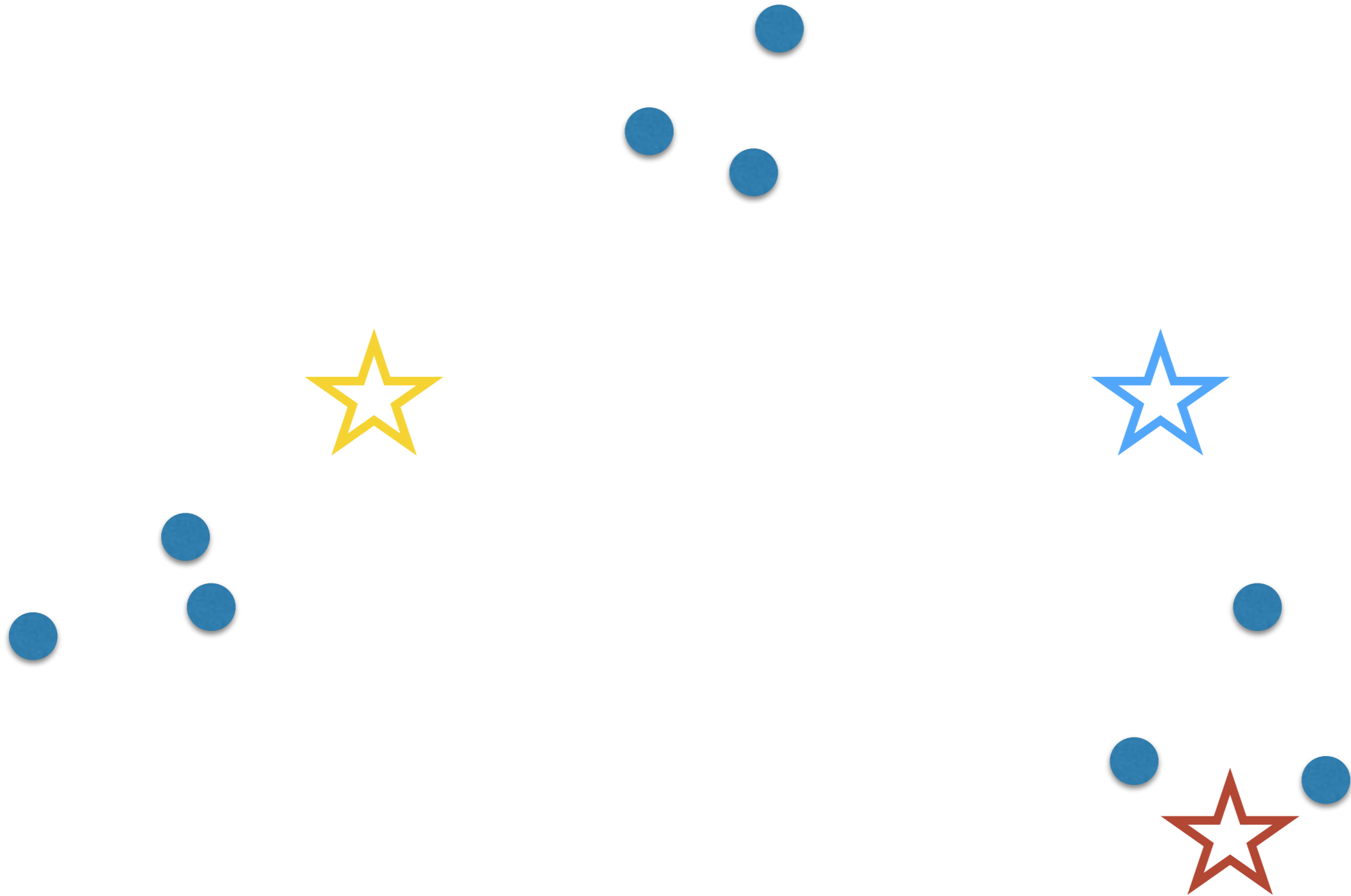
$$M_5 = \sum_{j=1}^K \sum_{t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2$$

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

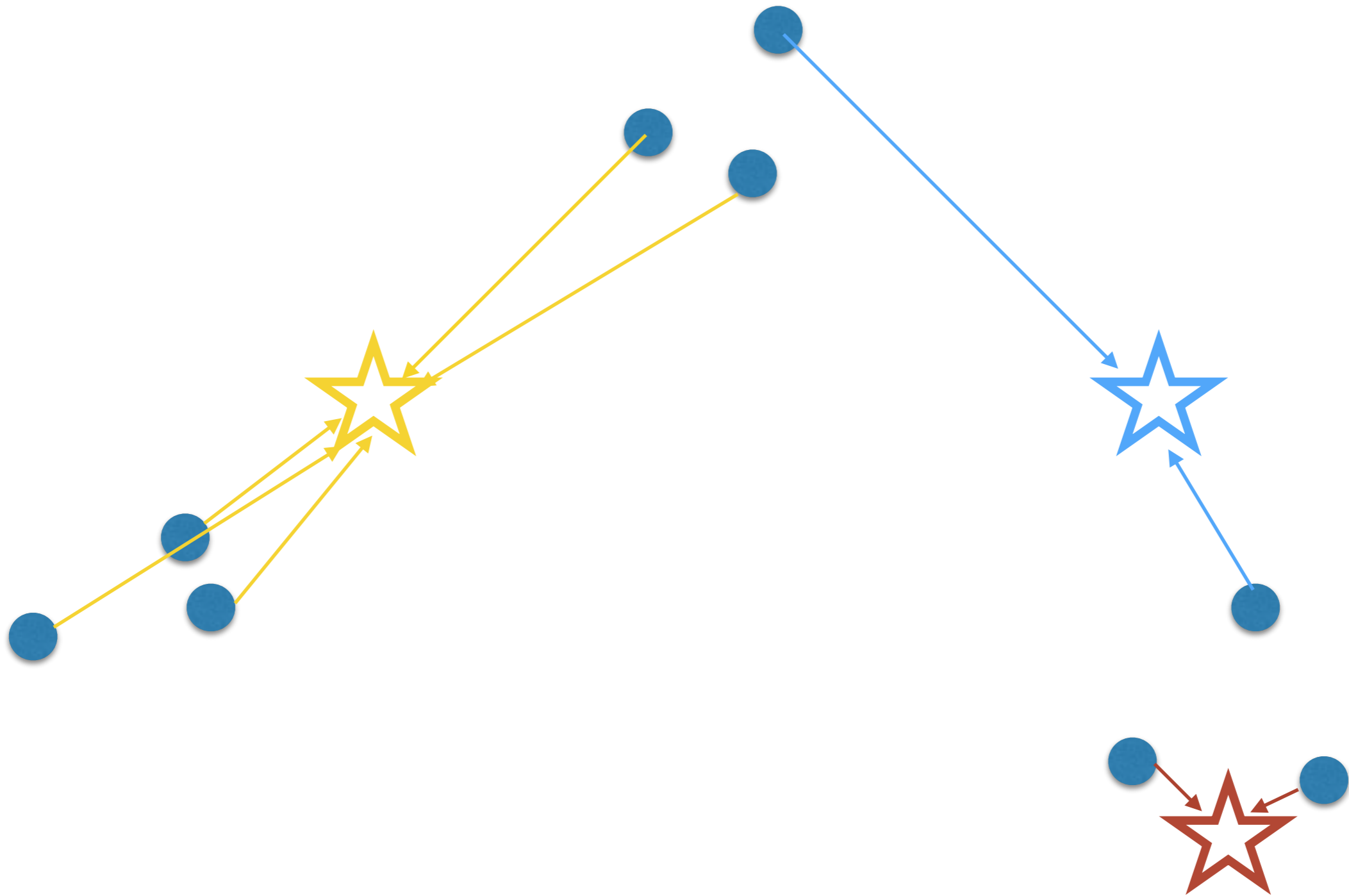
Demo



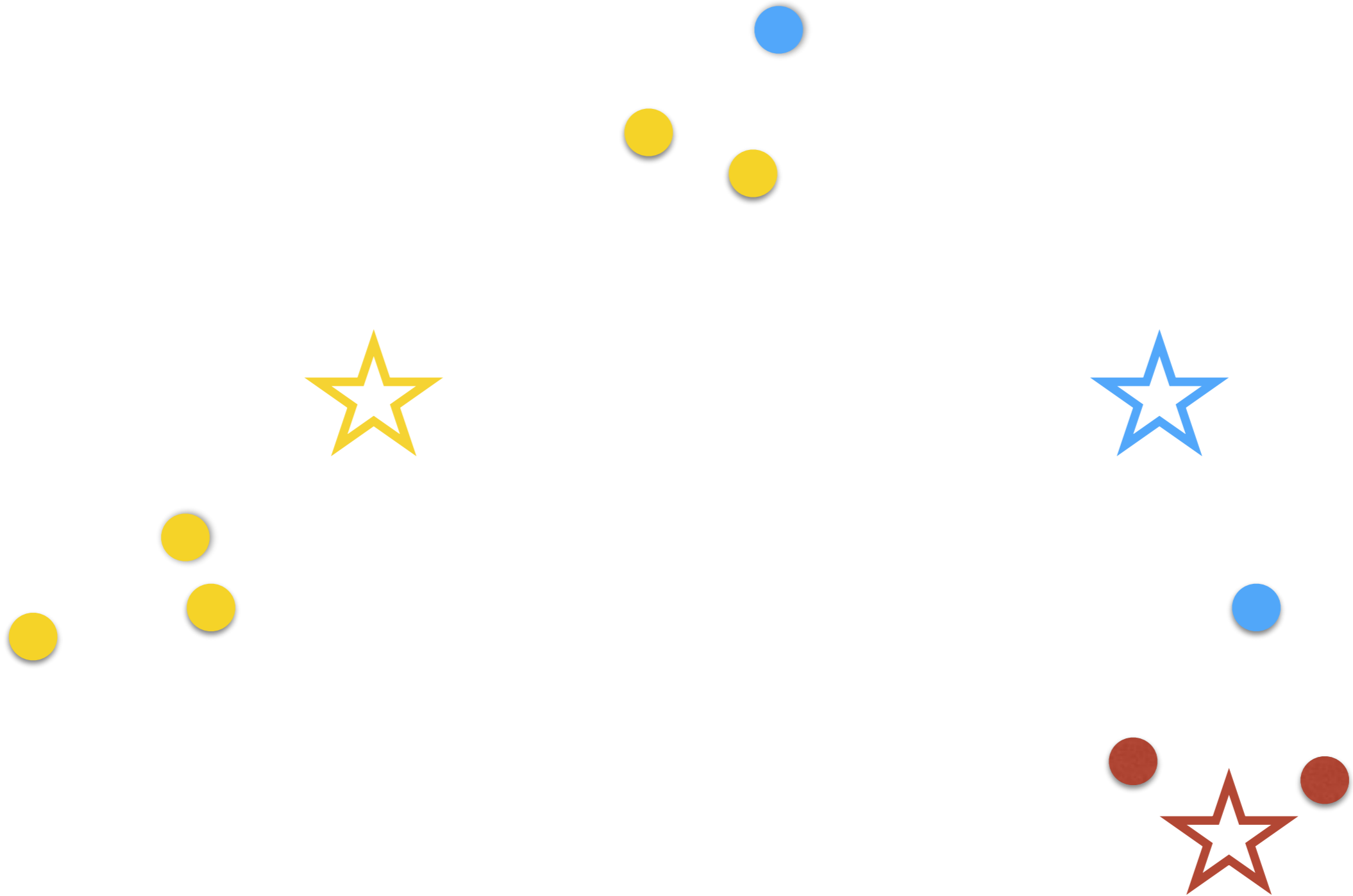
Demo



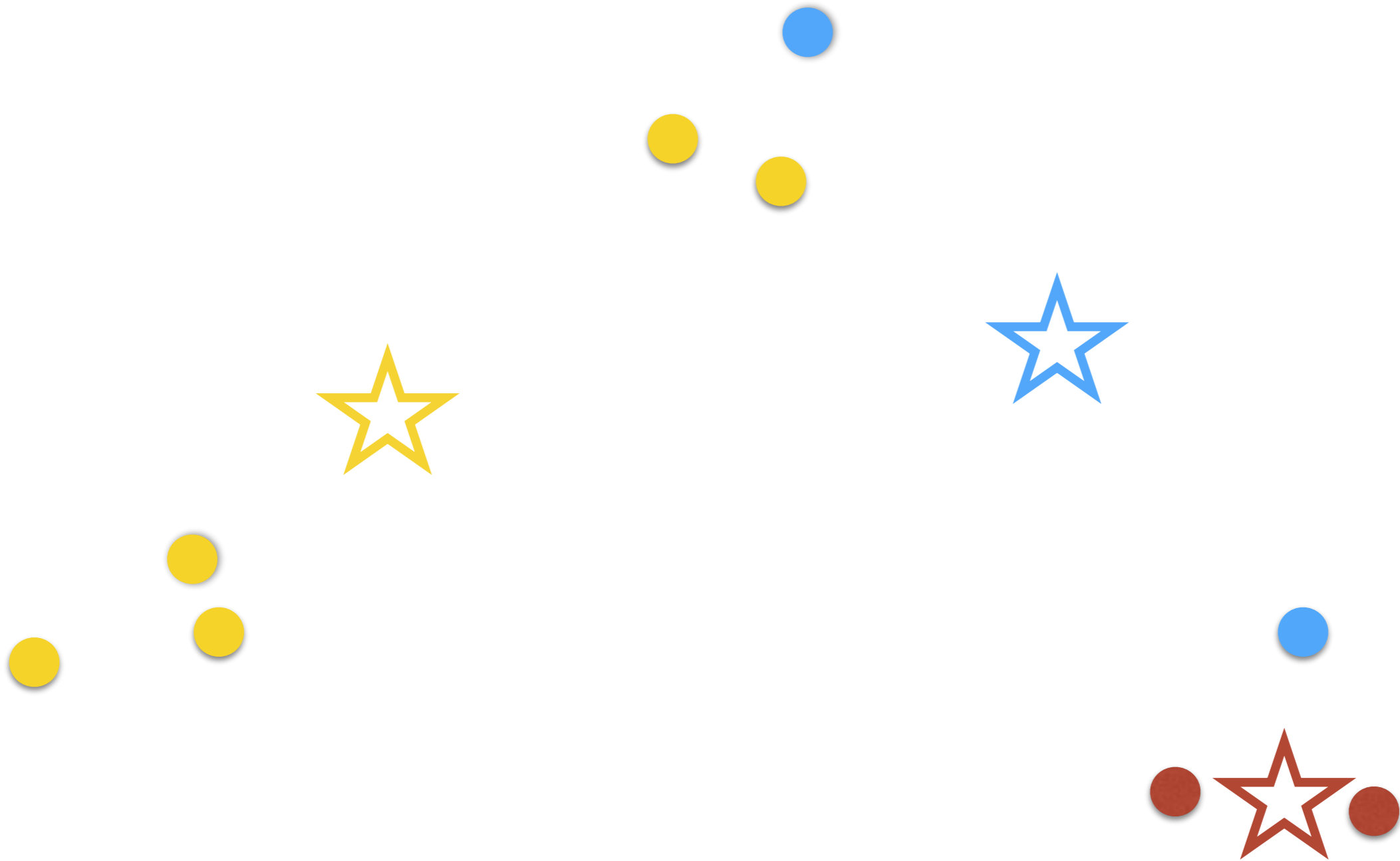
Demo



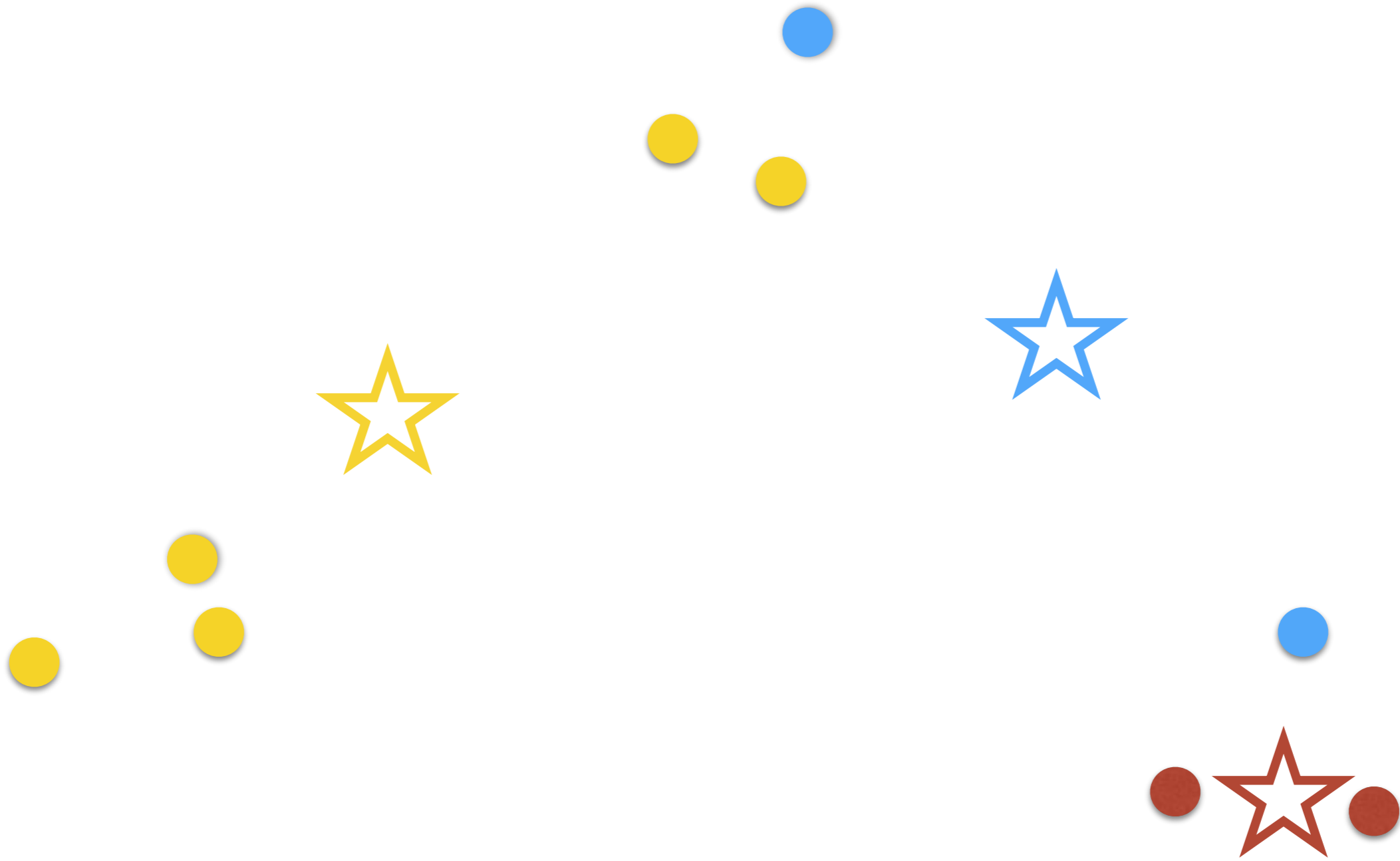
Demo



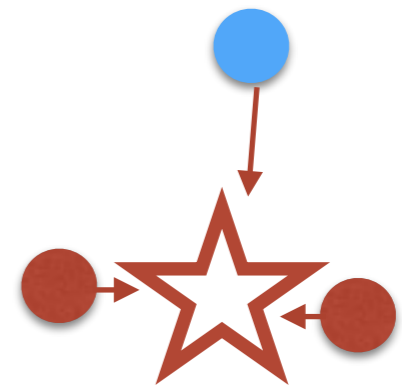
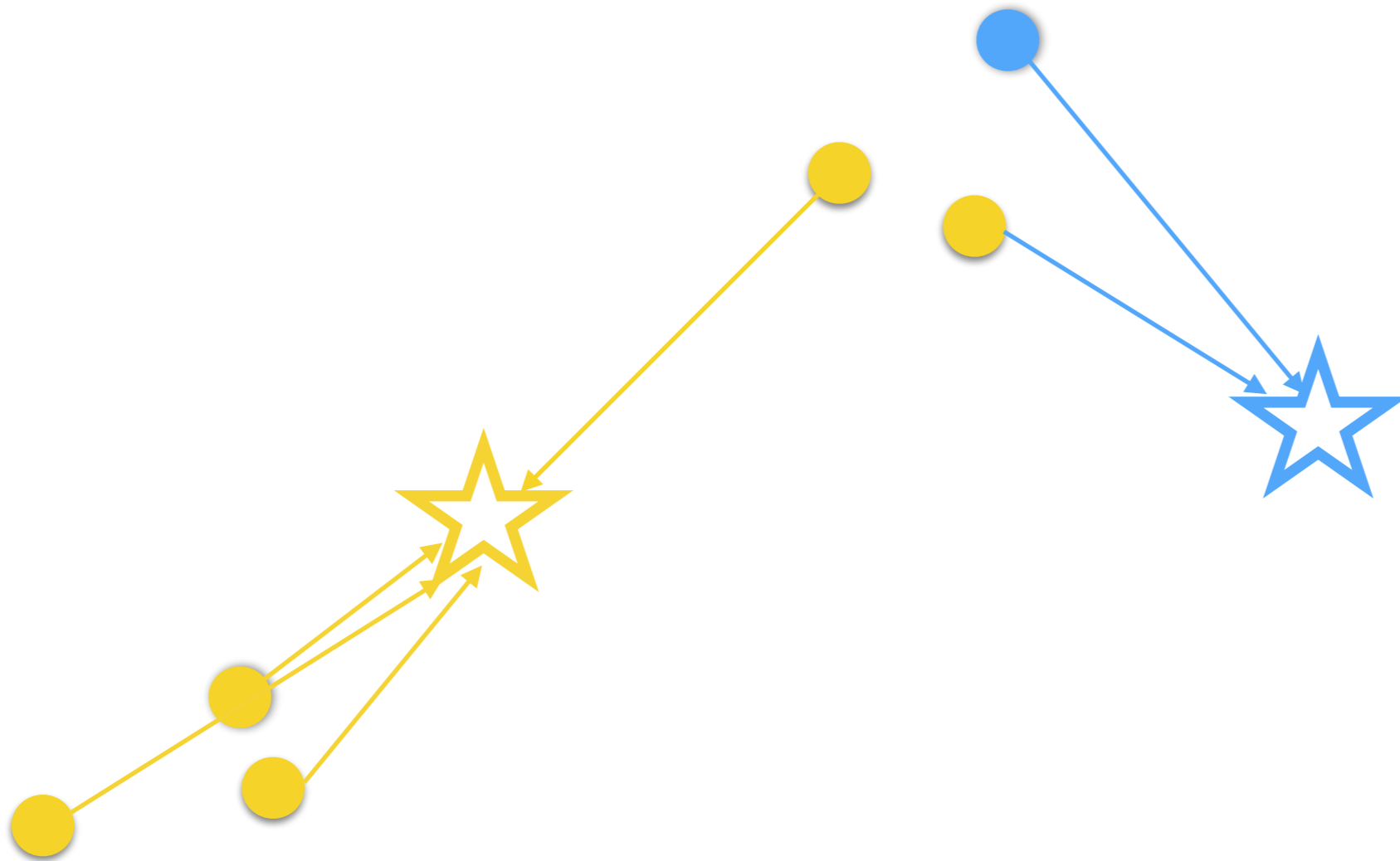
Demo



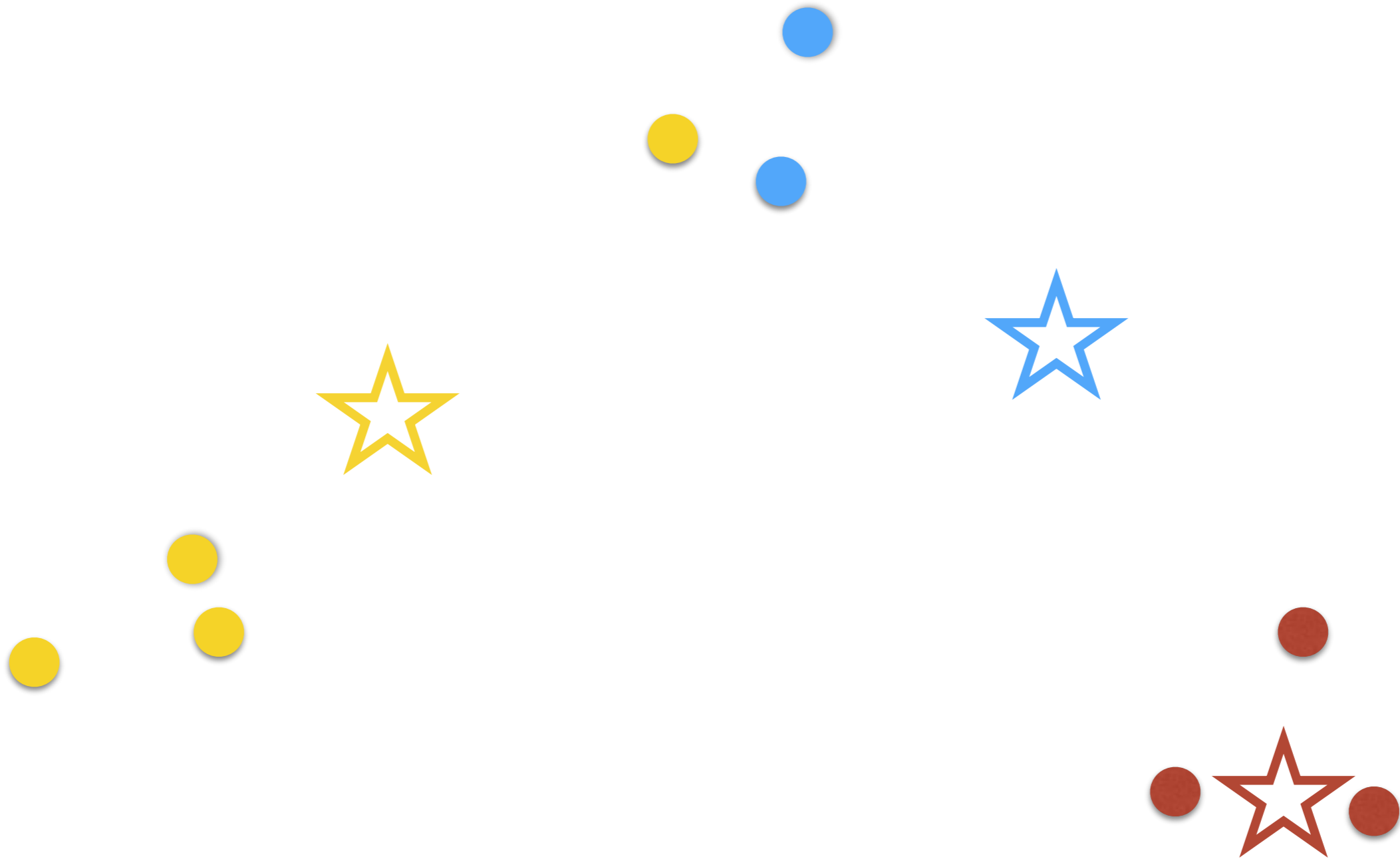
Demo



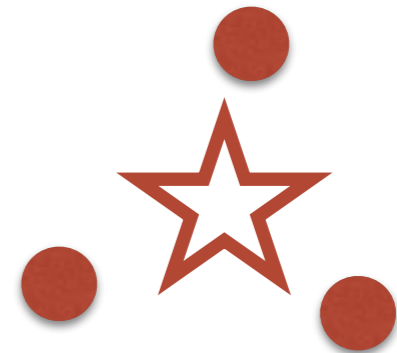
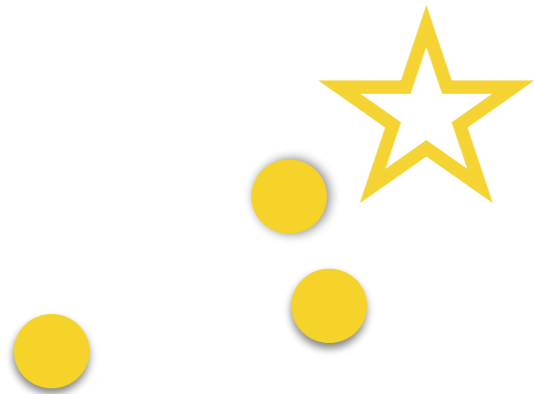
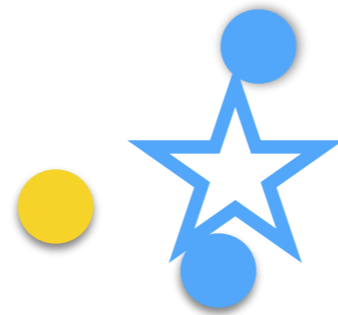
Demo



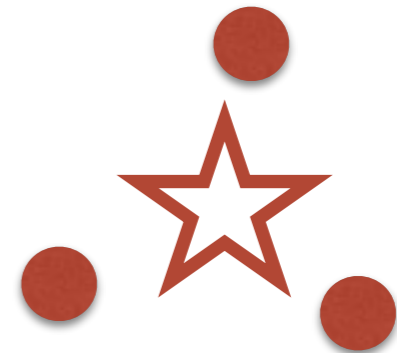
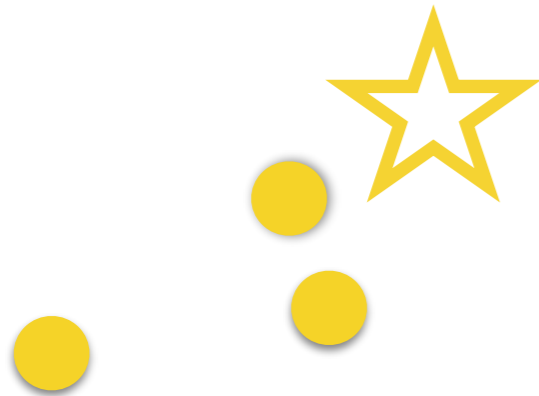
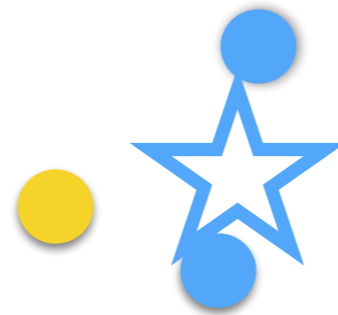
Demo



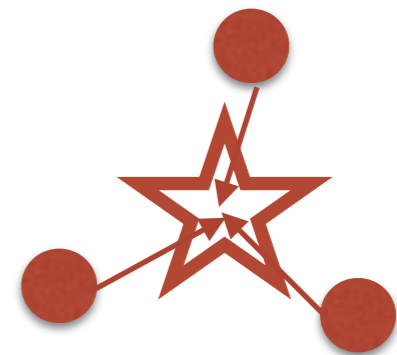
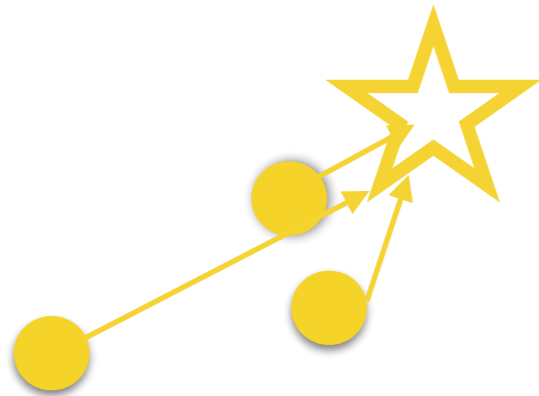
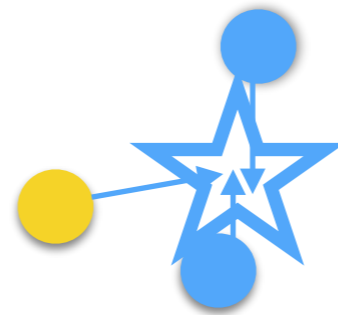
Demo



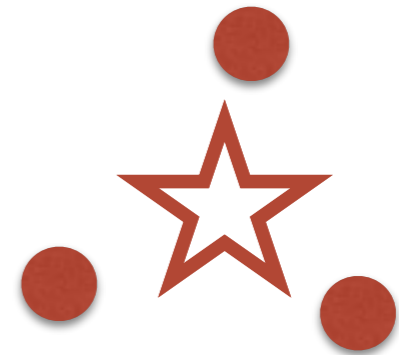
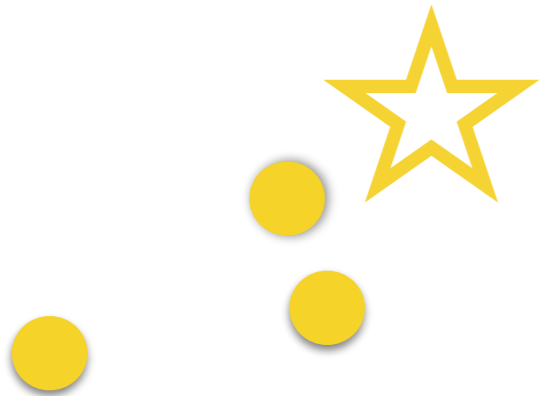
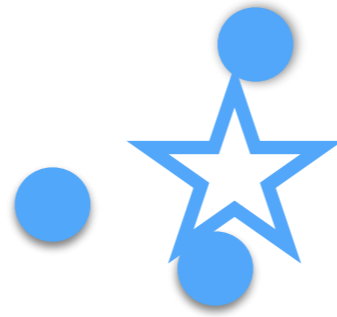
Demo



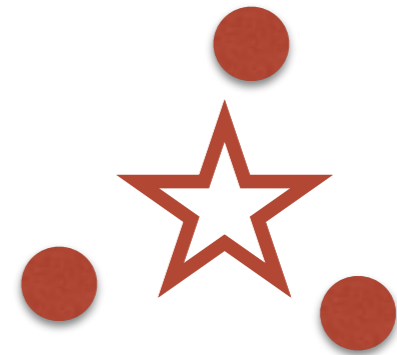
Demo



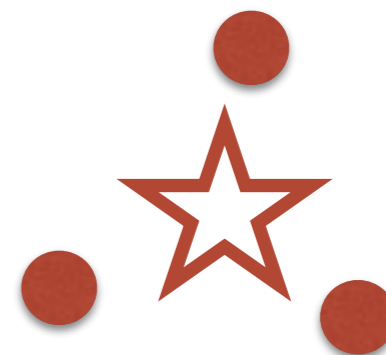
Demo



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 $t \in \{1, \dots, n\}$, set cluster identity of the point

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

- 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

K-means objective

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

K-means objective

$$\sum_{j=1}^K \sum_{t \in C_j} \left\| \mathbf{x}_t - \frac{1}{|C_j|} \sum_{s \in C_j} \mathbf{x}_s \right\|^2 = \min_{\mathbf{r}_1, \dots, \mathbf{r}_K} \sum_{j=1}^K \sum_{t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|^2$$
$$\parallel \parallel$$
$$M_5 = \min_{\mathbf{r}_1, \dots, \mathbf{r}_K} O(c; \mathbf{r}_1, \dots, \mathbf{r}_K)$$

K-means objective

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

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

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

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

Fact: Centroid is Minimizer

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

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

Proof

Proof

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

$$\mu_j = \frac{1}{|C_j|} \sum_{t \in C_j} \mathbf{x}_t$$

Proof

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

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

$$\mu_j = \frac{1}{|C_j|} \sum_{t \in C_j} \mathbf{x}_t$$

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

Proof

$$\begin{aligned} & \sum_{t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|^2 \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j + \mu_j - \mathbf{r}_j\|^2 \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2 + 2 \sum_{t \in C_j} (\mathbf{x}_t - \mu_j)^\top (\mu_j - \mathbf{r}_j) \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2 + 2 \left(\sum_{t \in C_j} \mathbf{x}_t - |C_j| \mu_j \right)^\top (\mu_j - \mathbf{r}_j) \end{aligned}$$
$$\mu_j = \frac{1}{|C_j|} \sum_{t \in C_j} \mathbf{x}_t$$

Proof

$$\begin{aligned} & \sum_{t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|^2 \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j + \mu_j - \mathbf{r}_j\|^2 \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2 + 2 \sum_{t \in C_j} (\mathbf{x}_t - \mu_j)^\top (\mu_j - \mathbf{r}_j) \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2 + 2 \left(\sum_{t \in C_j} \mathbf{x}_t - |C_j| \mu_j \right)^\top (\mu_j - \mathbf{r}_j) \end{aligned}$$
$$\mu_j = \frac{1}{|C_j|} \sum_{t \in C_j} \mathbf{x}_t$$

Proof

$$\begin{aligned} & \sum_{t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|^2 \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j + \mu_j - \mathbf{r}_j\|^2 \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2 + 2 \sum_{t \in C_j} (\mathbf{x}_t - \mu_j)^\top (\mu_j - \mathbf{r}_j) \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2 + 2 \left(\sum_{t \in C_j} \mathbf{x}_t - |C_j| \mu_j \right)^\top (\mu_j - \mathbf{r}_j) \\ &= \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 + \sum_{t \in C_j} \|\mu_j - \mathbf{r}_j\|^2 \\ &\geq \sum_{t \in C_j} \|\mathbf{x}_t - \mu_j\|^2 \end{aligned}$$
$$\mu_j = \frac{1}{|C_j|} \sum_{t \in C_j} \mathbf{x}_t$$

K-MEANS CONVERGENCE

- K-means algorithm converges to local minima of objective

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

- Proof:

K-MEANS CONVERGENCE

- K-means algorithm converges to local minima of objective

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

- Proof:

Clustering assignment improves objective:

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

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

K-MEANS CONVERGENCE

- K-means algorithm converges to local minima of objective

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

- Proof:

Clustering assignment improves objective:

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

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

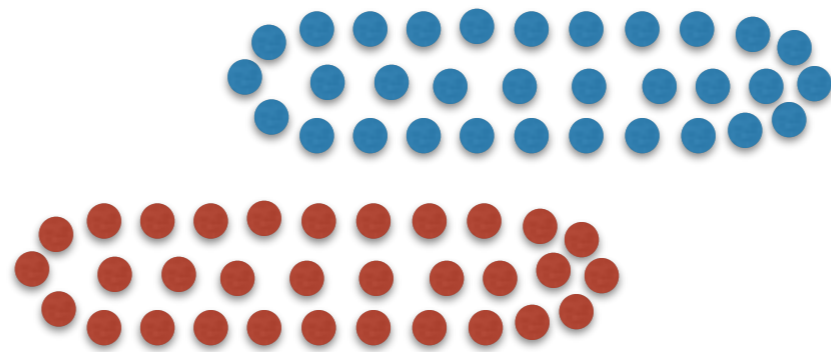
Computing centroids improves objective:

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

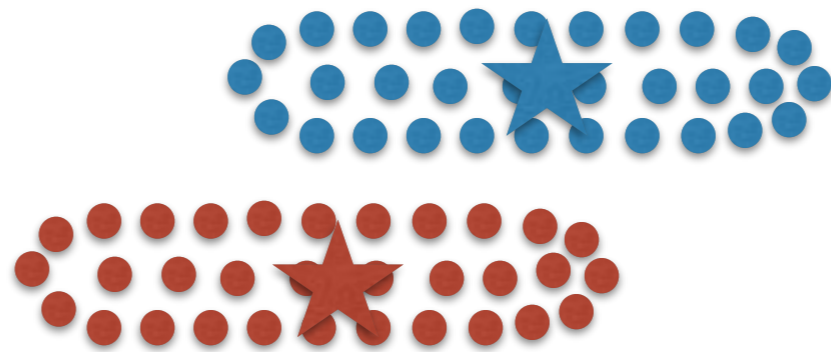
(By the fact about centroid)

Clustering demo

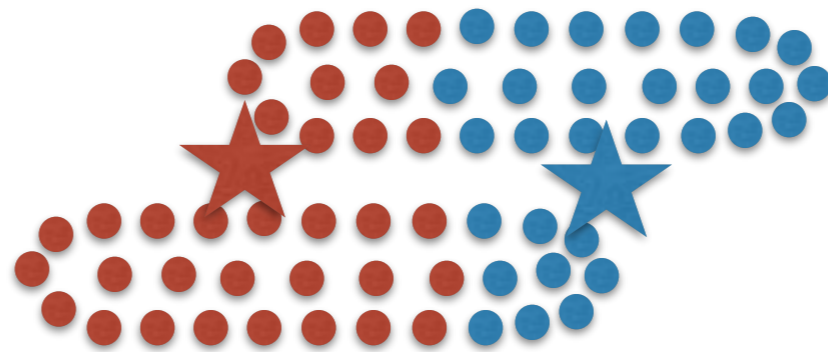
K-means: pitfalls



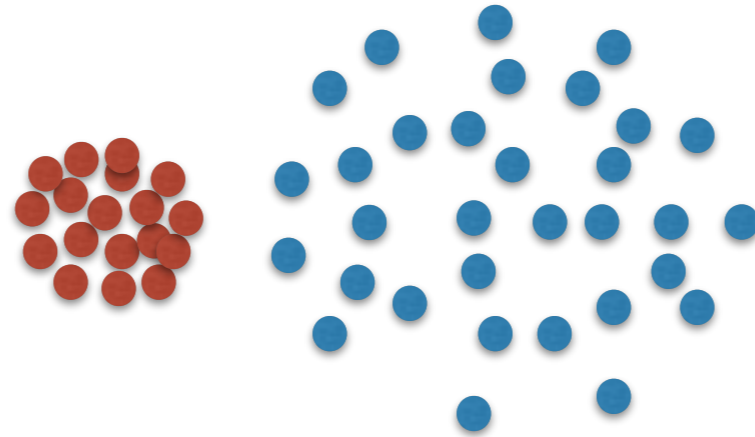
K-means: pitfalls



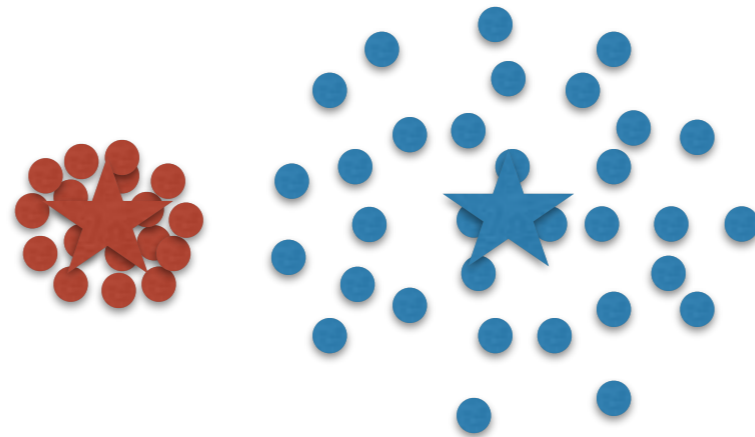
K-means: pitfalls



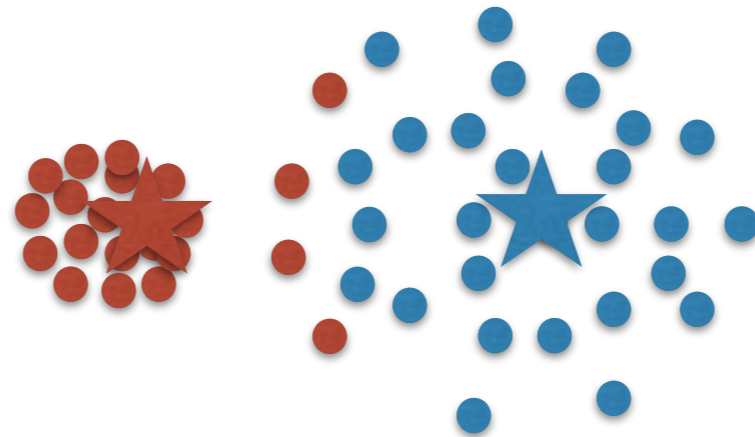
K-means: pitfalls



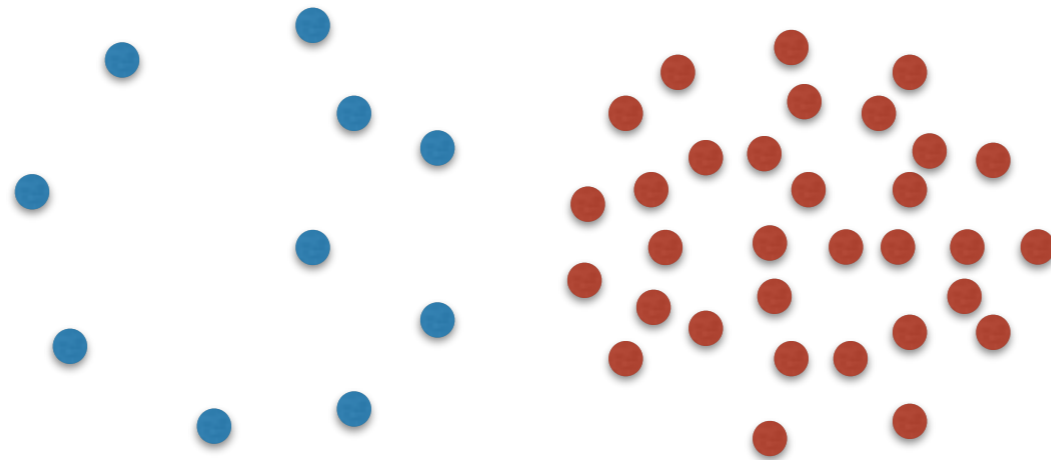
K-means: pitfalls



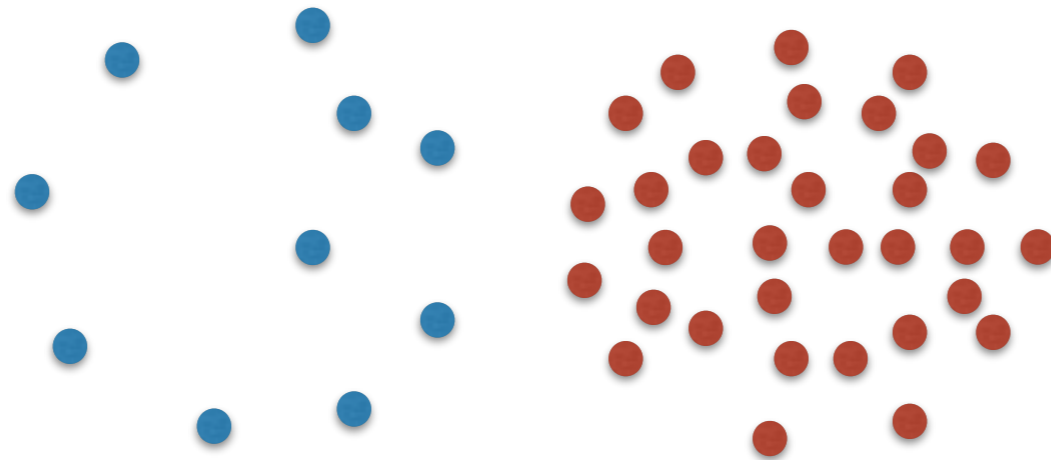
K-means: pitfalls



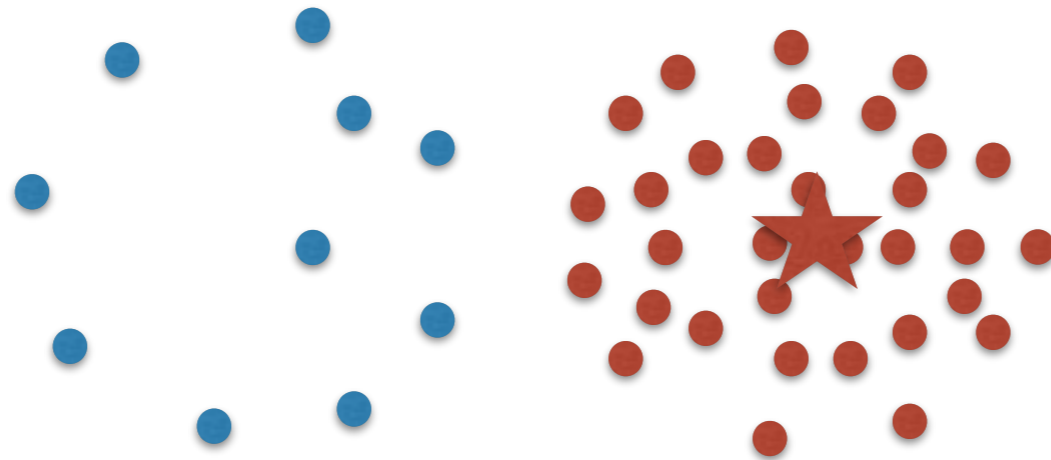
K-means: pitfalls



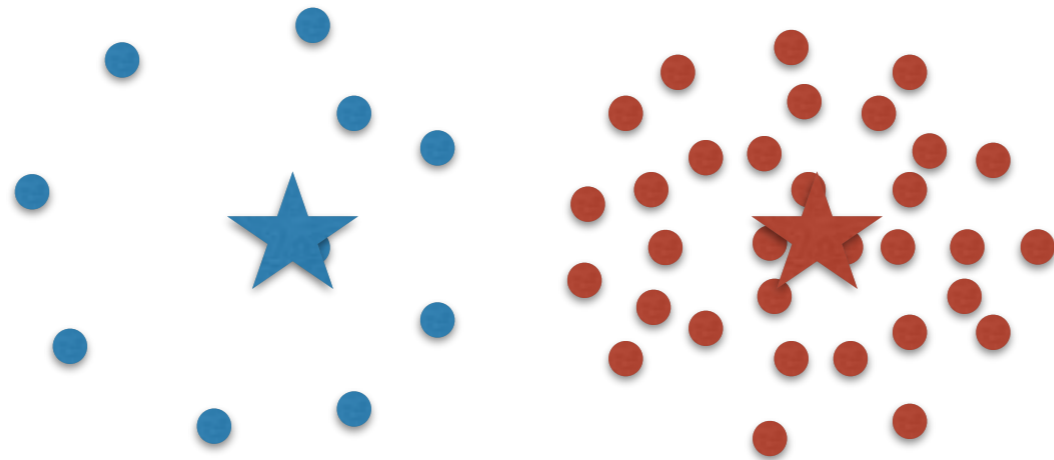
K-means: pitfalls



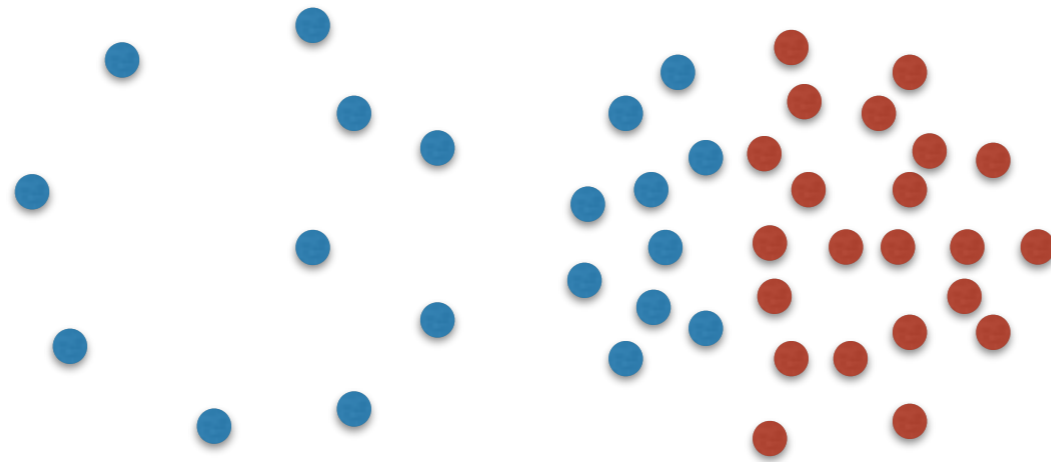
K-means: pitfalls



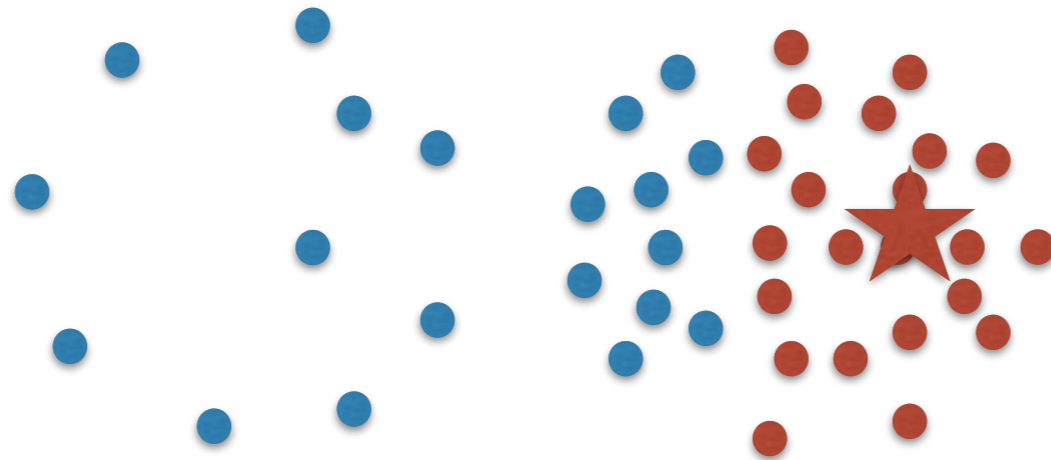
K-means: pitfalls



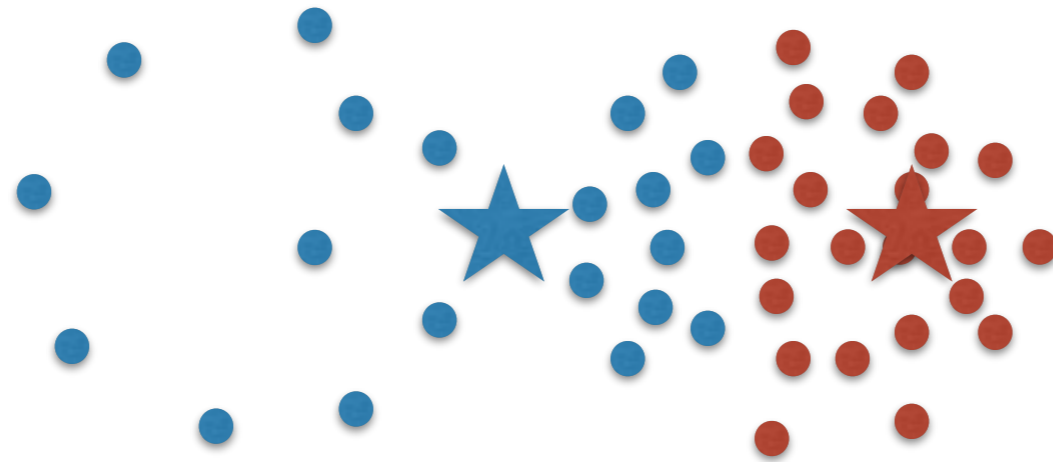
K-means: pitfalls



K-means: pitfalls



K-means: pitfalls



K-means: pitfalls

- Looks for spherical clusters
- Of same size
- And with roughly equal number of points

K-means: pitfalls

- Can we design algorithm that can address these shortcomings?