Кластеризация

Цели кластеризации

- Выделение закономерностей
- Построение иерархии множества объектов
- Упрощение дальнейшей обработки данных
- Сокращение объема данных
- Выделение нетипичных объектов
- Получение новых признаков

Кластеризация – обучение без учителя

- Нет точной постановки задачи
- Число кластеров?
- Критерий качества?

K-Means

Количество кластеров – задается.

μ_i – центр кластера C_i

Задача – минимизация:

$$\sum_{x_j} \min_{\mu_i} \left|\left|x_j - \mu_i\right|\right|_2^2$$

Алгоритм K-Means

1. Инициализируем центры кластеров (случайно или более хитрым образом).

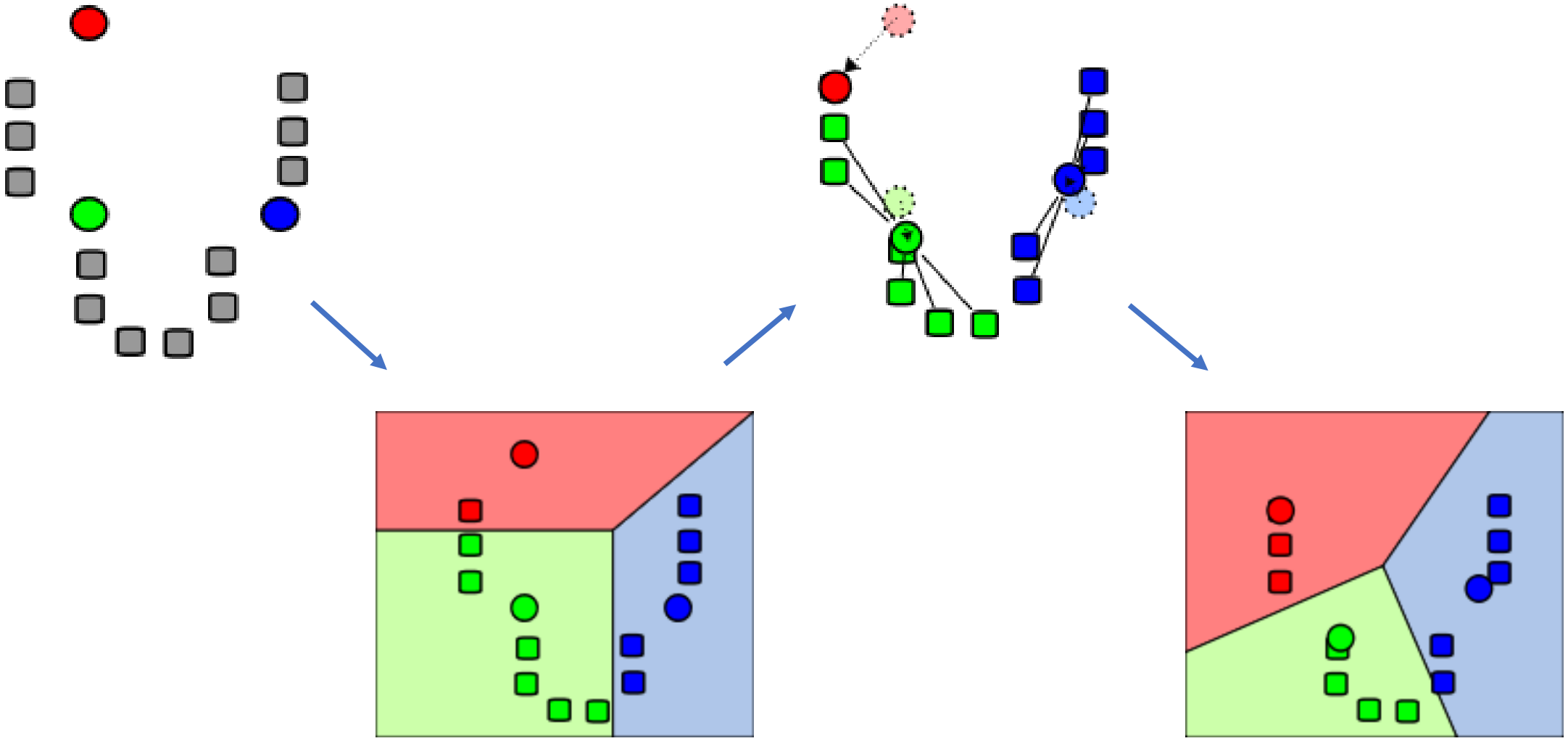
2. Припишем каждую точку к ближайшему центру.

3. Переместим центры кластеров в «центр масс» кластеров:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

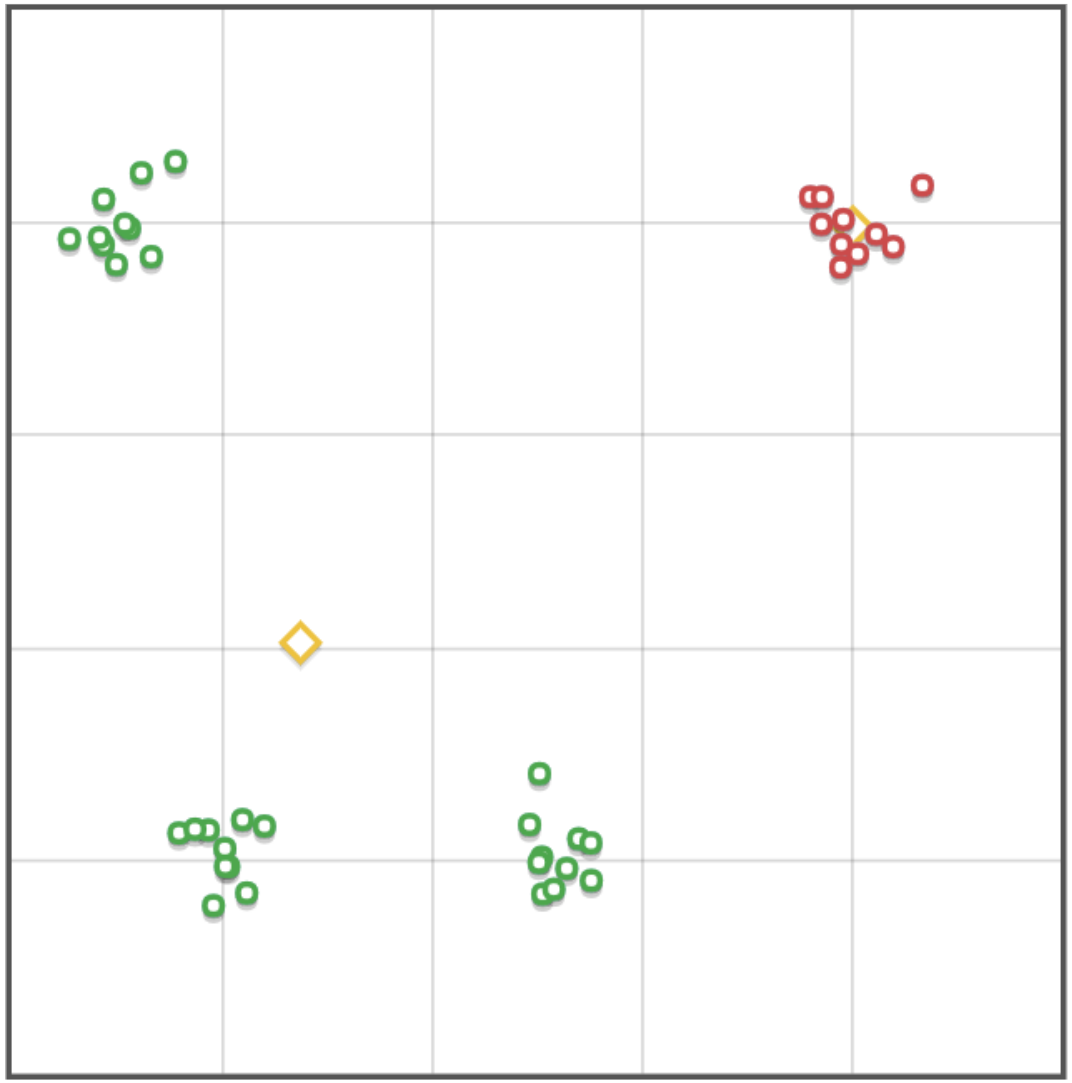
4. Повторяем шаги 2-3 до схождения.

K-Means



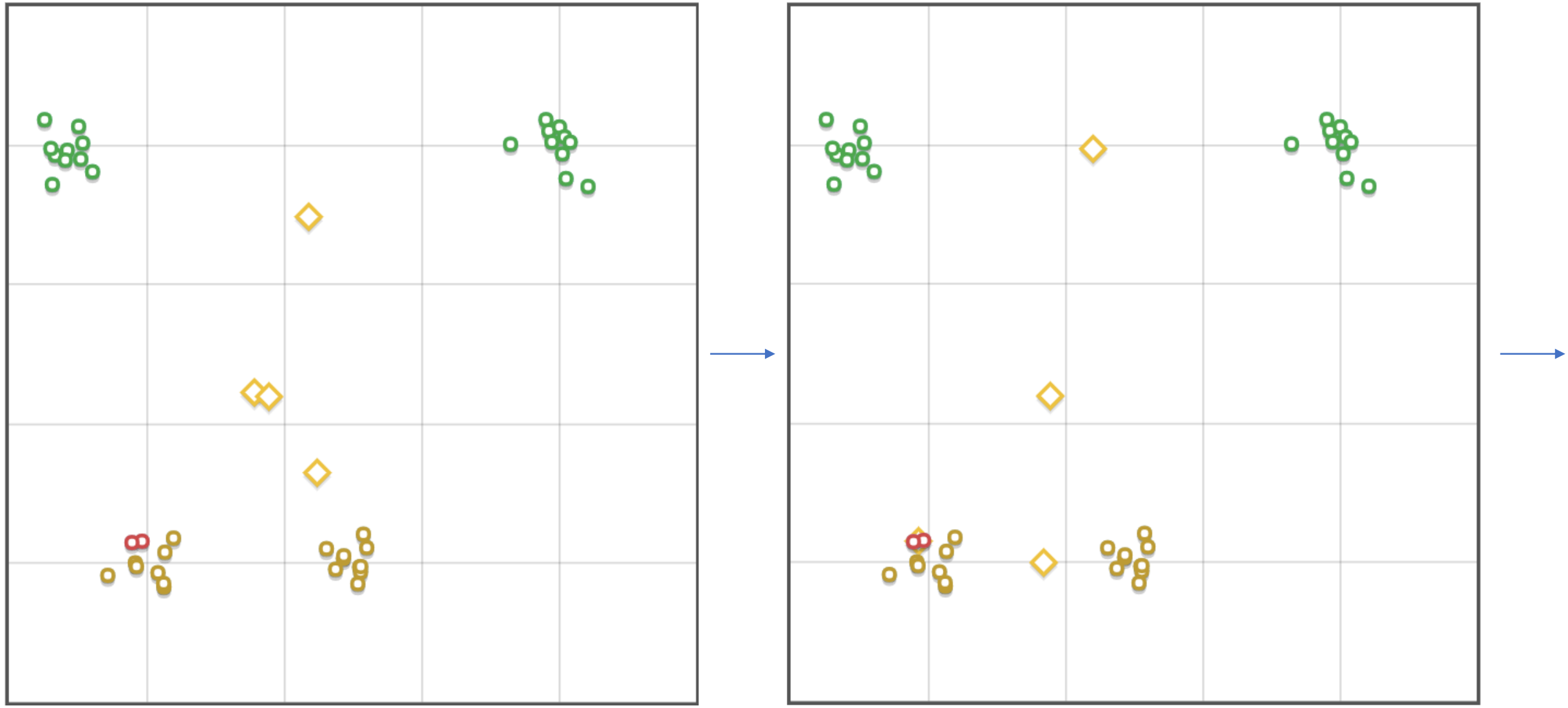
Проблемы K-Means

Выбор количества кластеров



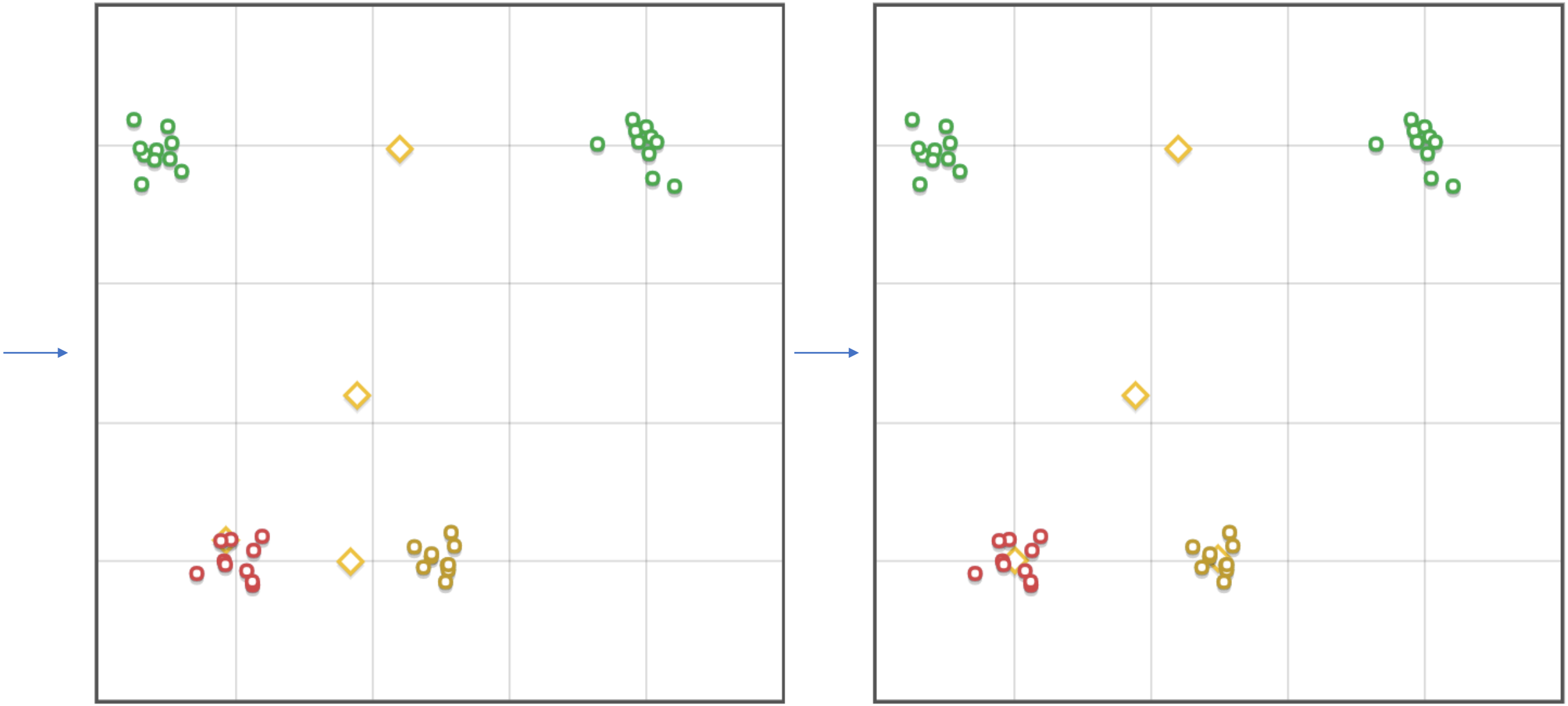
Проблемы K-Means

Выбор начального положения центров



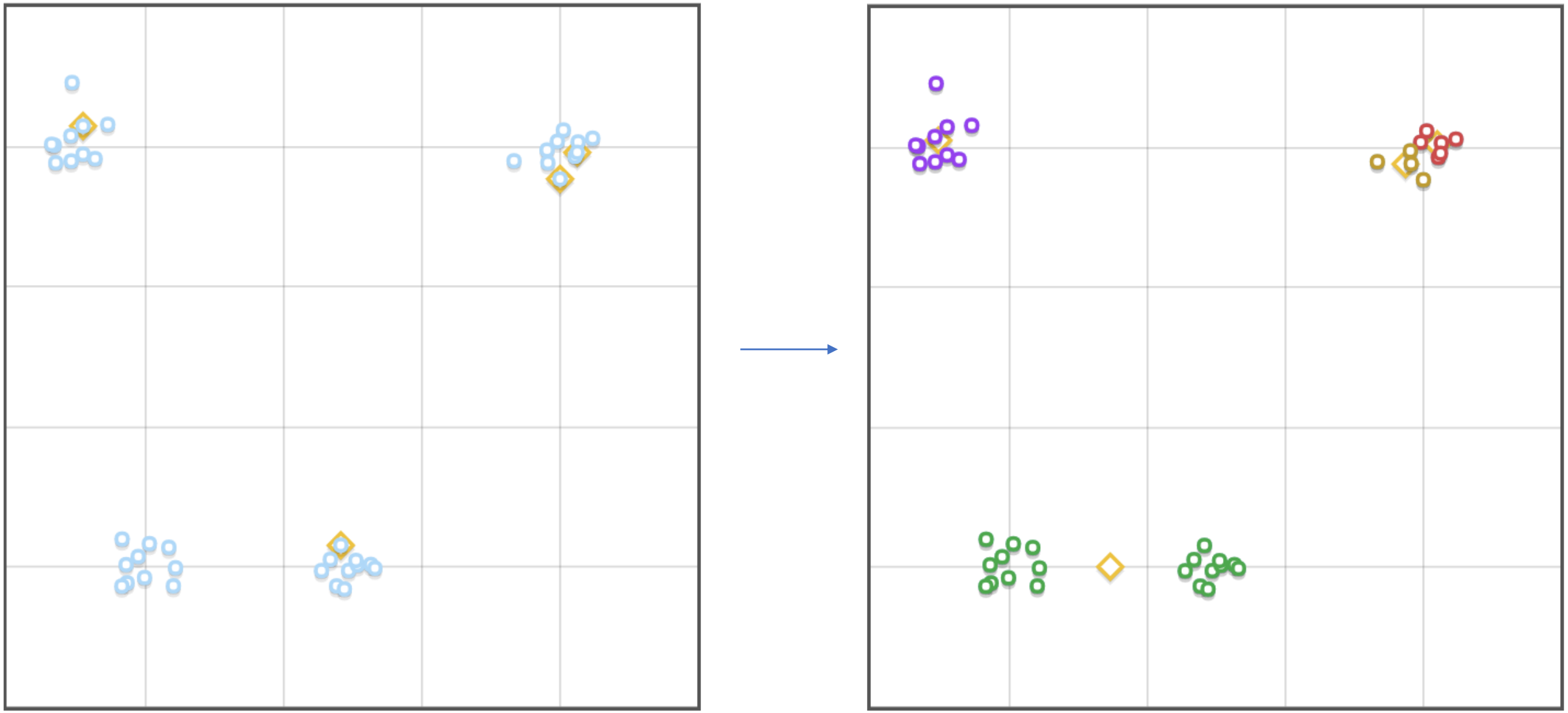
Проблемы K-Means

Выбор начального положения центров



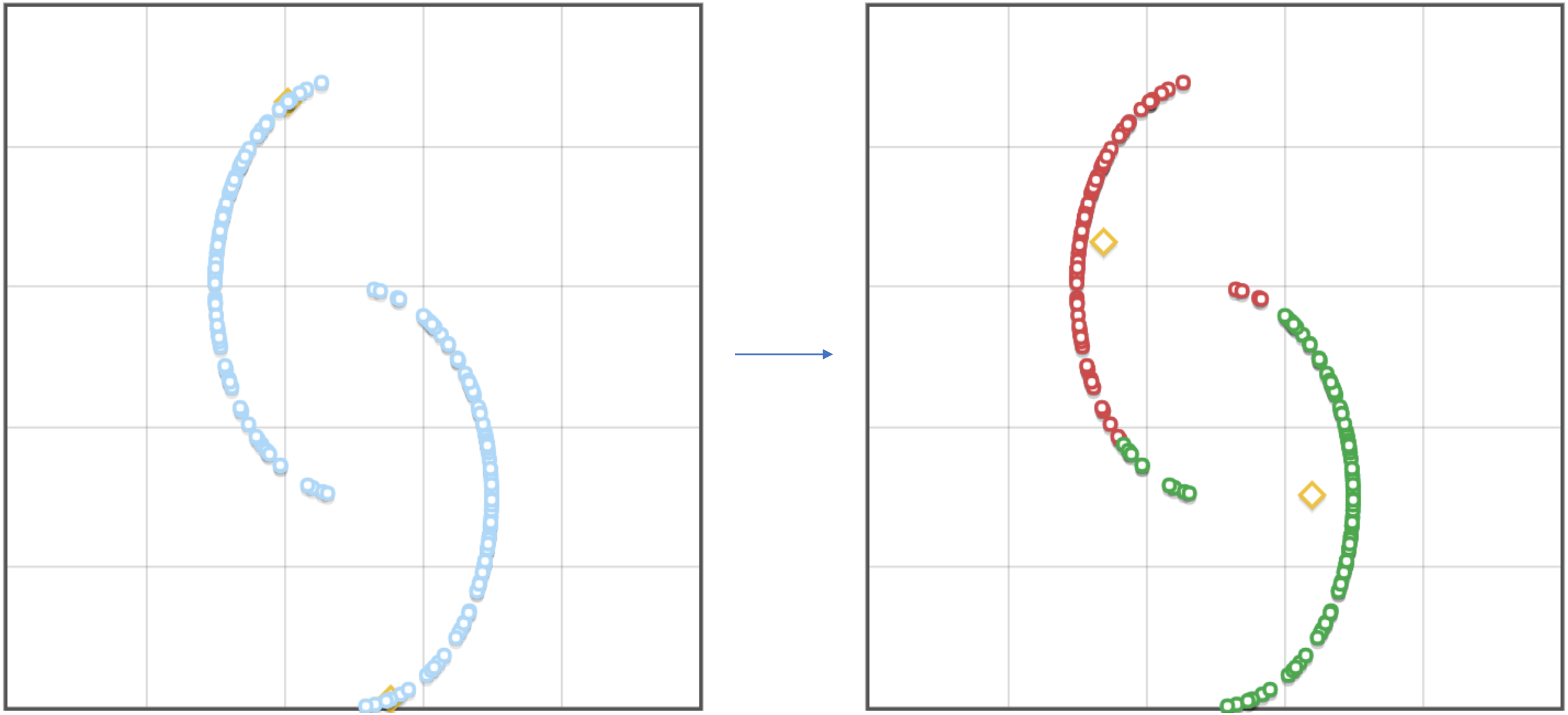
Проблемы K-Means

Выбор начального положения центров



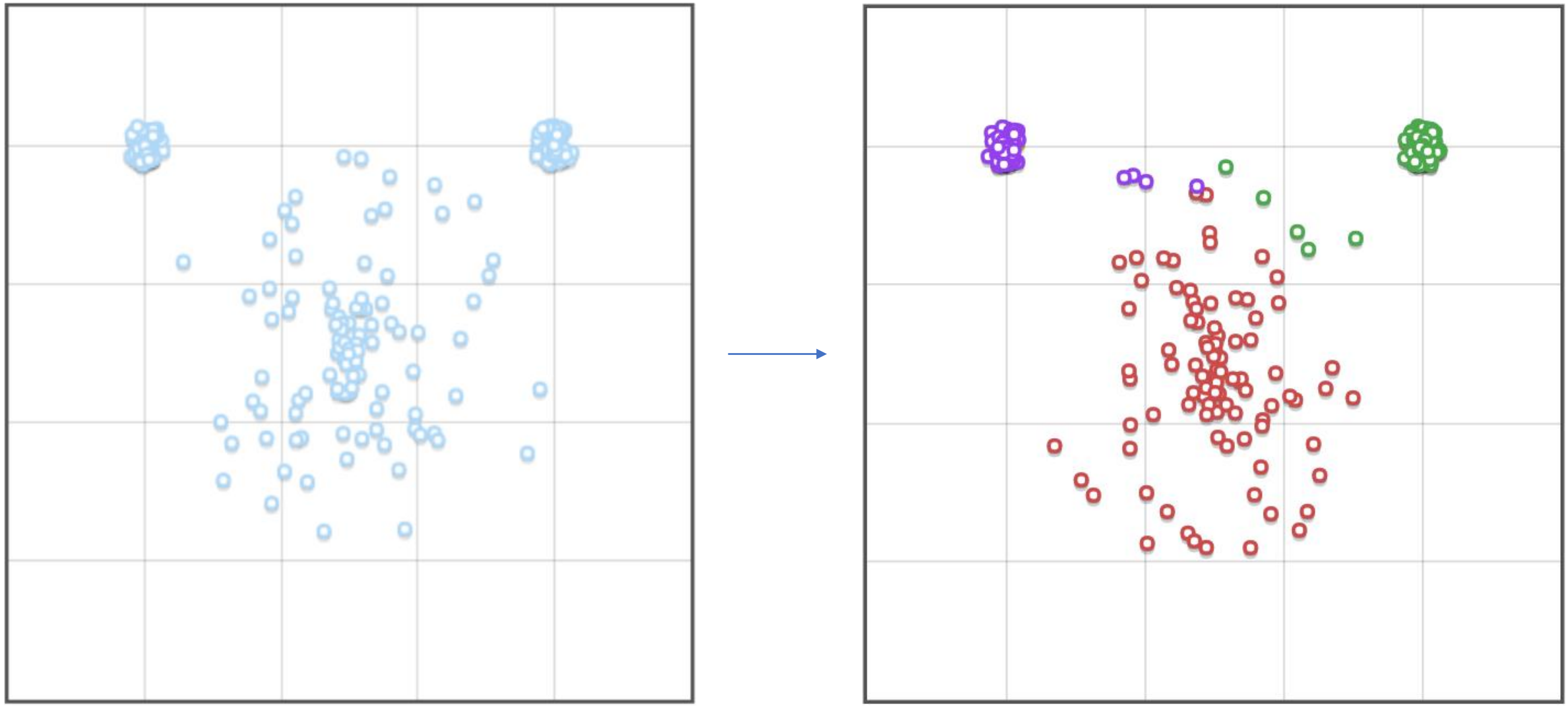
Проблемы K-Means

Несферические кластеры



Проблемы K-Means

Разноразмерные кластеры



Mean Shift

Задается не количество кластеров, а максимальный размер кластера.

Ищем плотные области, сдвигая центры.

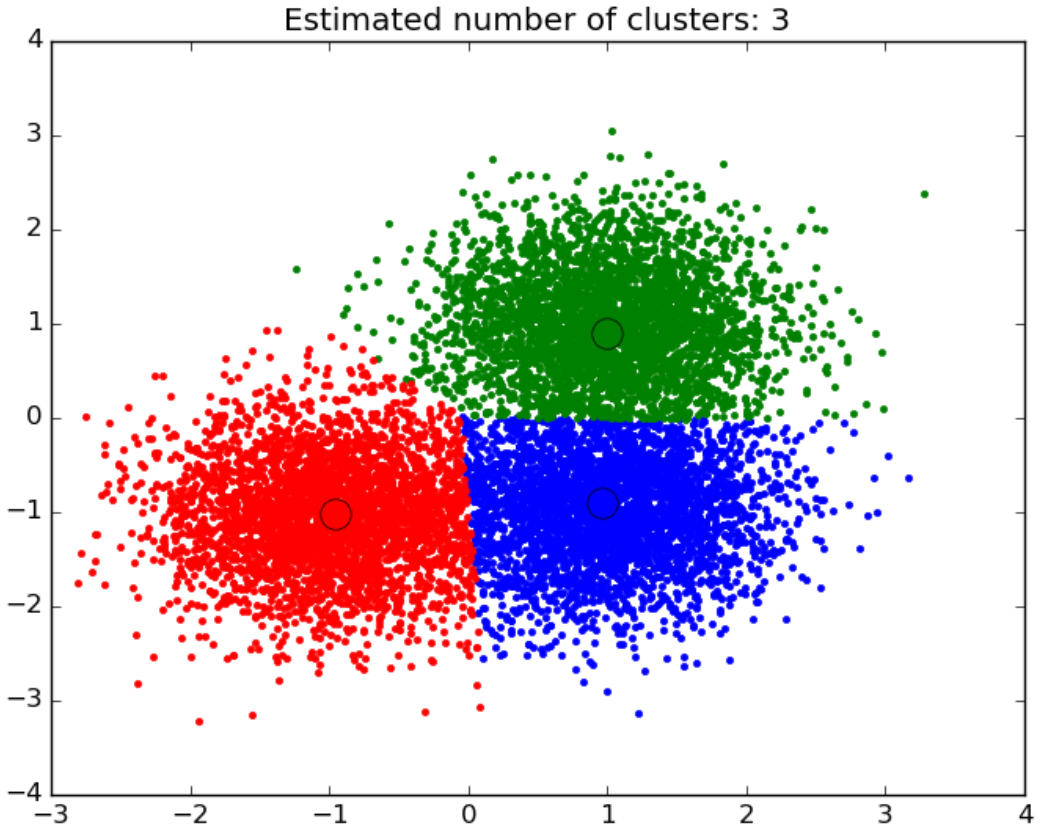
$$\mu_i(t+1) = \mathrm{m}\big(\mu_i(t)\big)$$

$$\mathrm{m}(\mu_i) = \frac{\sum_{\mathrm{x}_j \in N(\mu_i)} K\big(\mathrm{x}_j - \mu_i\big)\mathrm{x}_j}{\sum_{\mathrm{x}_j \in N(\mu_i)} K\big(\mathrm{x}_j - \mu_i\big)},$$

где $N(\mu_i)$ – окрестность точки μ_i, K – RBF (Radial Basis Function):

$$K\big(\mathrm{x}_j - \mu_i\big) = e^{-c\left|\left|\mathrm{x}_j - \mu_i\right|\right|_2^2}$$

Последний шаг – фильтруем центры, убираем дубликаты.

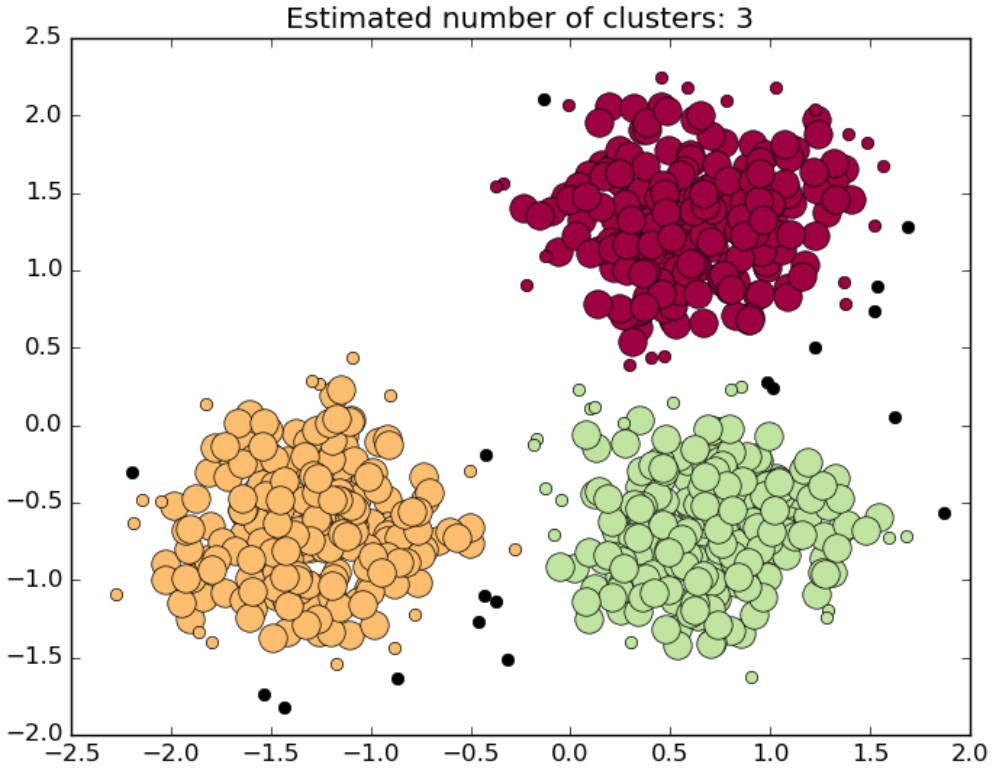


DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Количество кластеров – не задается.

Ищем объекты плотности (core samples) – такие объекты, в ϵ окрестности которых есть хотя бы m других объектов.

Объединяем объекты плотности на расстоянии ϵ и их окружение в кластеры.



Иерархическая кластеризация

Agglomerative Clustering

Иерархическая кластеризация снизу-вверх.

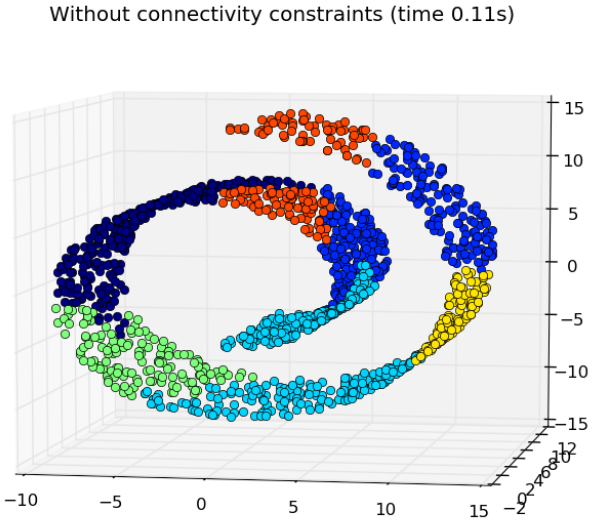
Начинаем с того, что каждая точка – отдельный кластер.

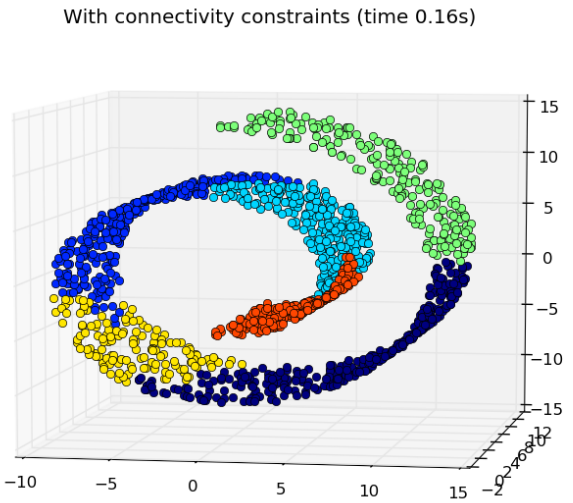
Три стратегии порядка слияния (linkage),
минимизирующие соответствующие величины:

- ward – дисперсия соединяемых кластеров
- average – среднее расстояние между объектами кластеров
- maximum (complete) – максимальное расстояние
между объектами кластеров

Сливаем пока не останется один кластер.

Можно добавить требование близости соединяемых кластеров (connectivity).





Affinity Propagation

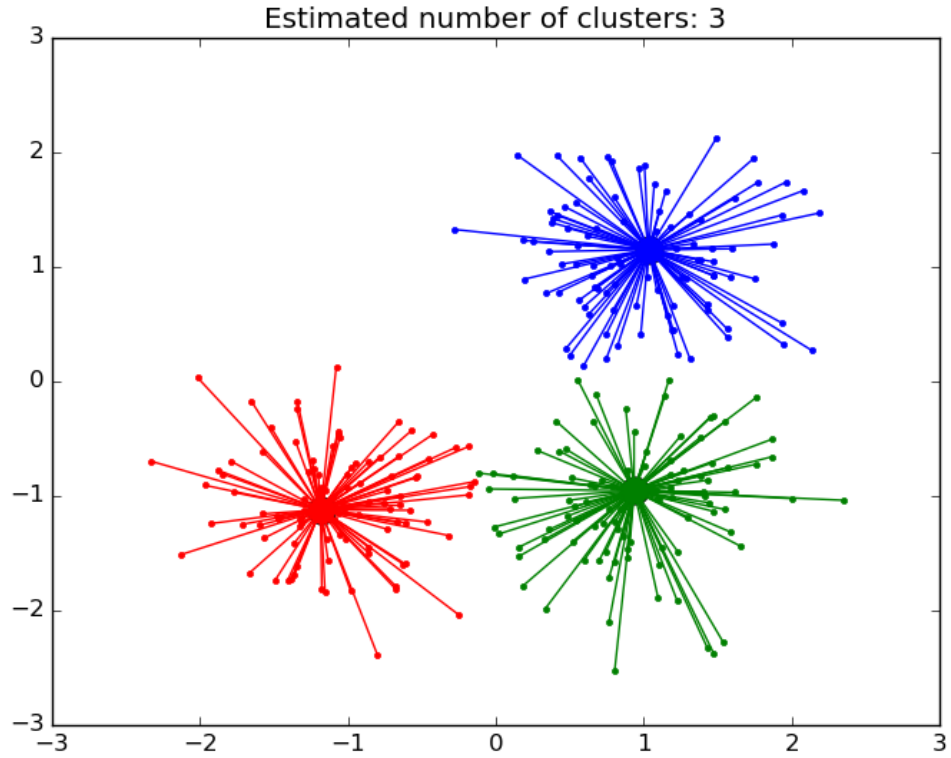
Affinity Propagation

Данные – близости объектов x_i и x_k - $s(i,k)$.

Распространяем до схождения 2 величины – *ответственность* $r(i,k)$ и *доступность* $a(i,k)$ (в начале - 0).

$$r(i,k) \leftarrow s(i,k) - \max_{m \neq k}\left(a(i,m) + s(i,m)\right)$$

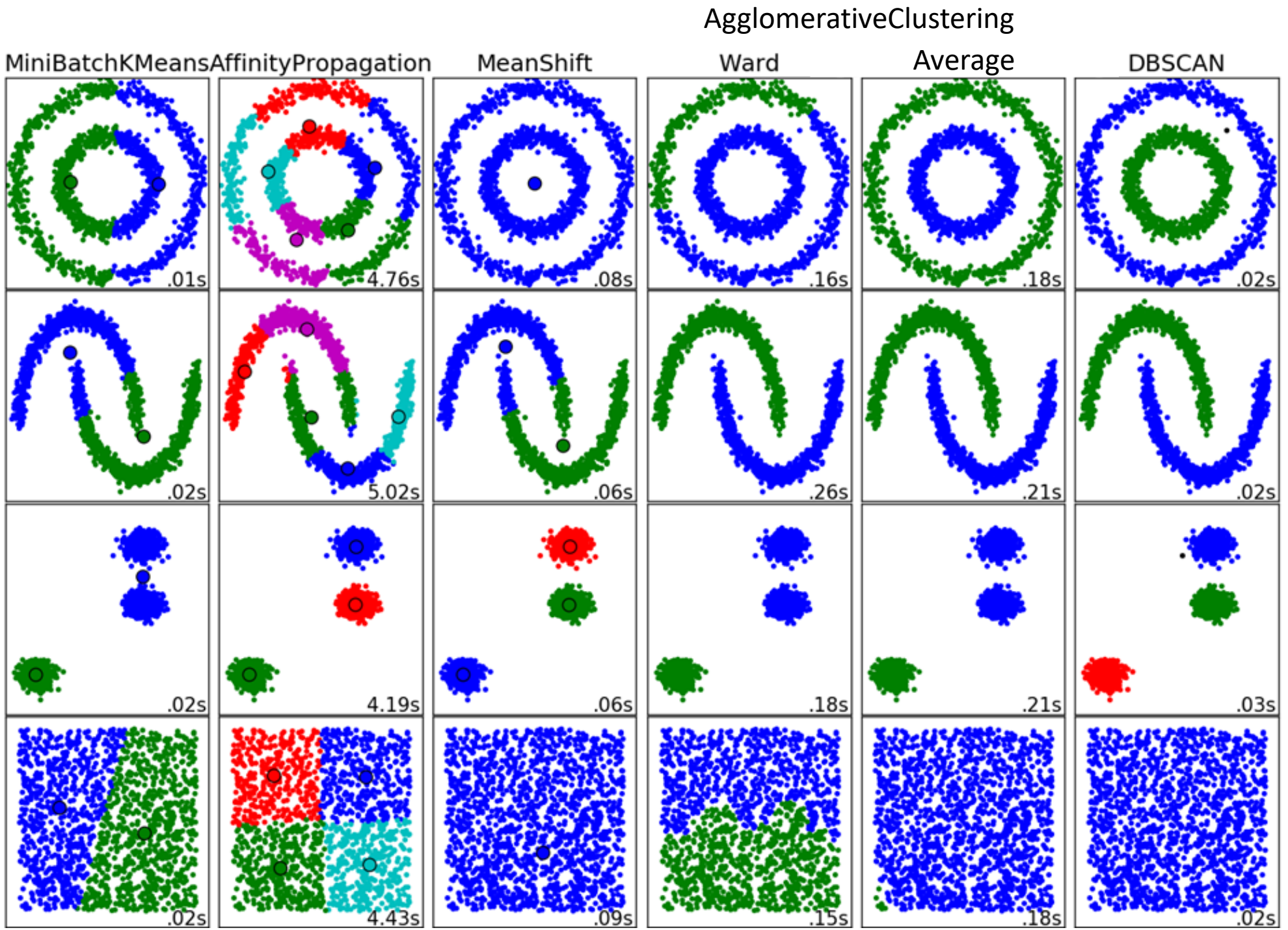
$$a(i,k) \leftarrow \begin{cases} \sum_{j \neq i} \max\left(0, r(j,k)\right), & \text{если } k = i \\ \min\left(0, r(k,k) + \sum_{j \neq i,k} \max\left(0, r(j,k)\right)\right), & \text{если } k \neq i \end{cases}$$



Выбираем самые «ответственные» и «доступные точки» точки:

$$\mu_i = \text{argmax}_k\left(a(i,k) + r(i,k)\right)$$

$s(i,i)$ - параметр, регулирующий количество кластеров


AgglomerativeClustering
MiniBatchKMeans
AffinityPropagation
MeanShift
Ward
Average
DBSCAN
.01s
4.76s
.08s
.16s
.18s
.02s
.02s
5.02s
.06s
.26s
.21s
.02s
.02s
4.19s
.06s
.18s
.21s
.03s
.02s
4.43s
.09s
.15s
.18s
.02s

Метрики кластеризации

Внутренние (internal):

- Davies-Bouldin index:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i} \left(\frac{\overline{\rho(\mu_i, \mathrm{x}^i)} + \overline{\rho(\mu_j, \mathrm{x}^j)}}{\rho(\mu_i, \mu_j)} \right)$$

- Dunn index:

$$D = \frac{\min\limits_{i \neq j} \rho(\mu_i, \mu_j)}{\max\limits_{x_i, x_j \in \mu} (x_i, x_j)}$$

Внешние (external):

- Purity:

Доля точек максимального класса в кластерах

- *метрики классификации

Active learning

Semi-supervised learning

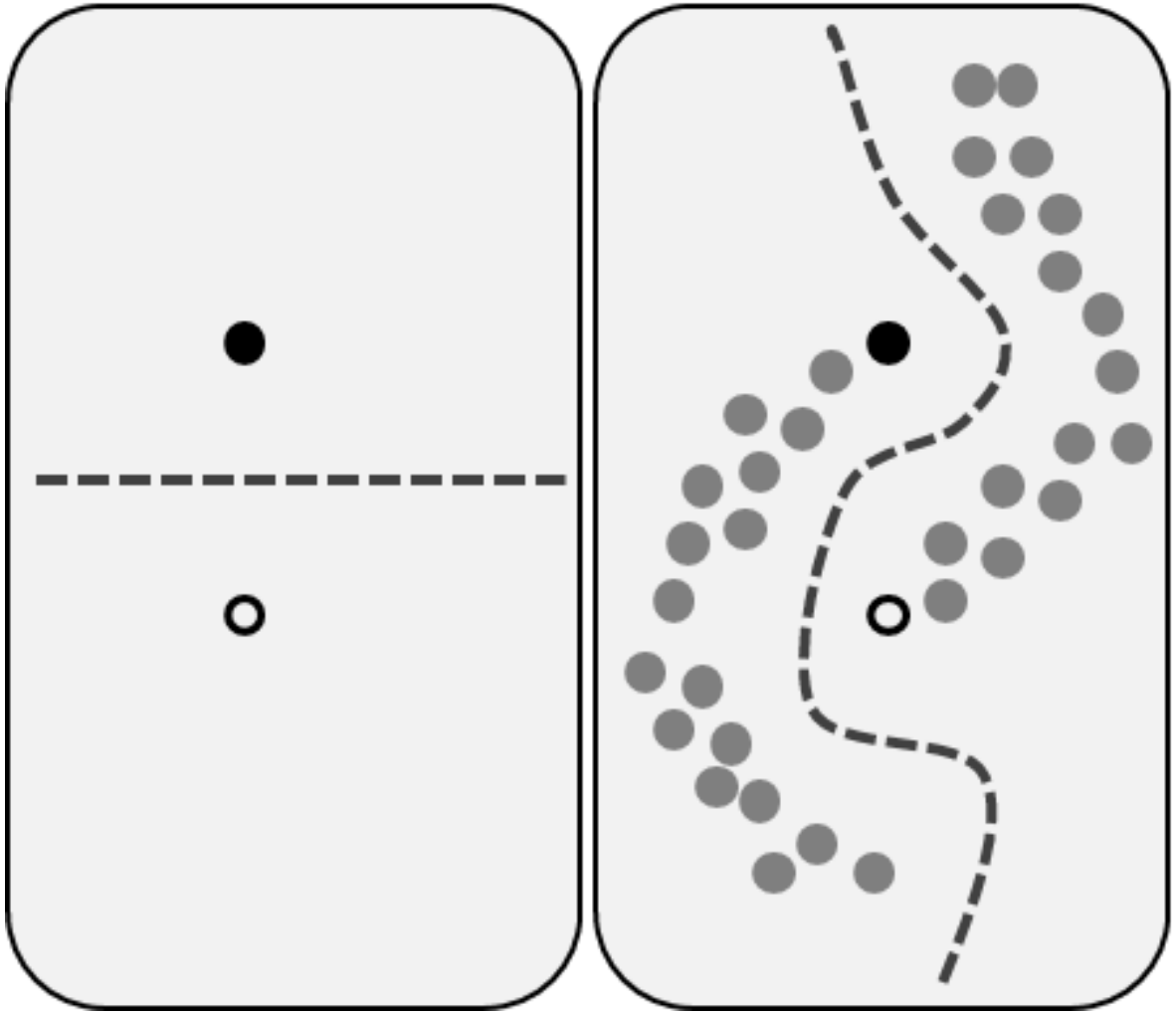
Зачастую получить данные (вектора признаков) довольно дешево, а разметить их довольно дорого, например из-за участия оценивающих людей.

Получить много данных можно, например:

- Веб-документы
- Речь
- Изображения и видео

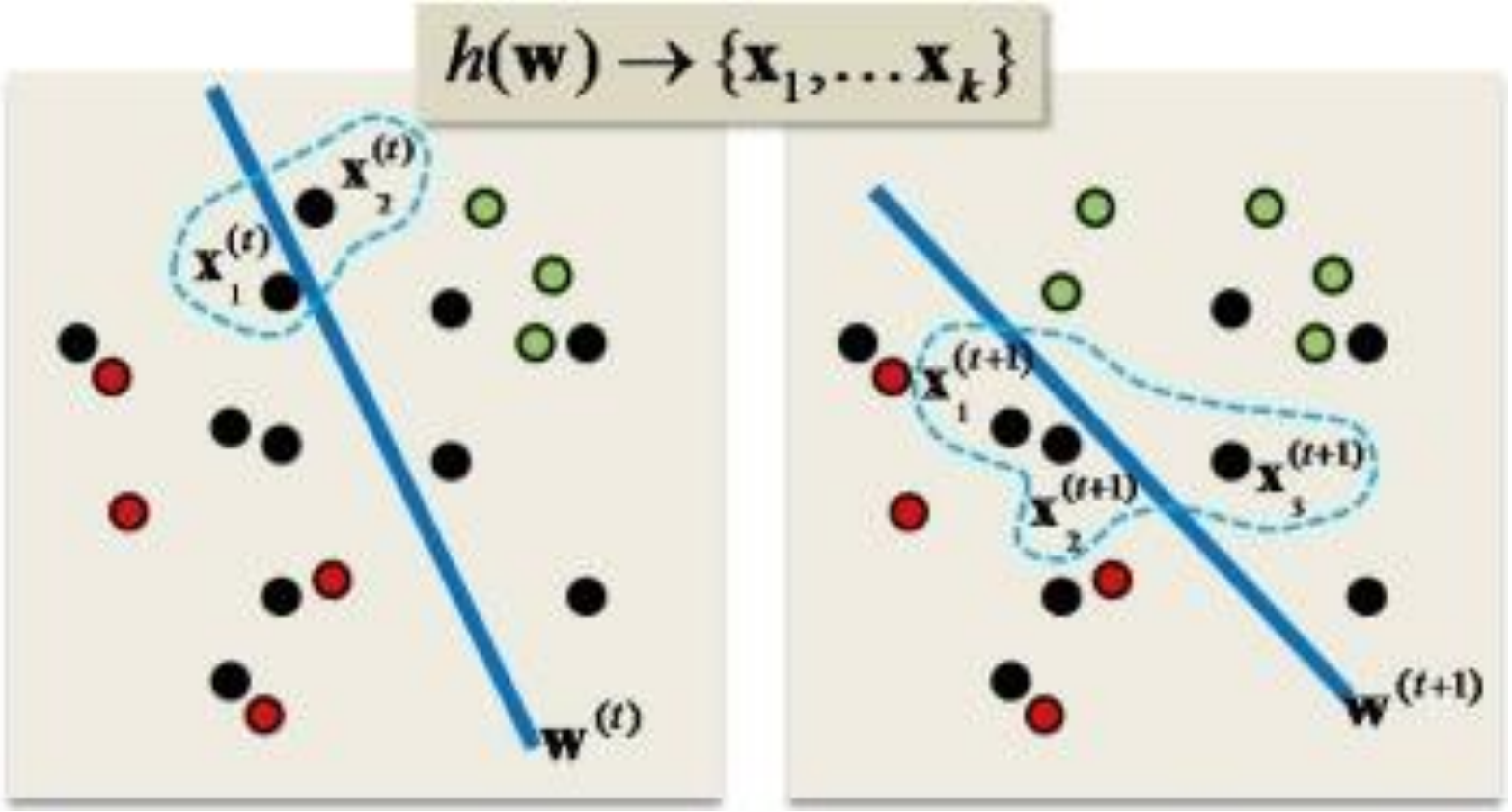
Тогда можно разметить не все вектора. Но информация о неразмеченных векторах все равно может пригодится.

Semi-supervised learning
(обучение с частичным привлечением учителя)

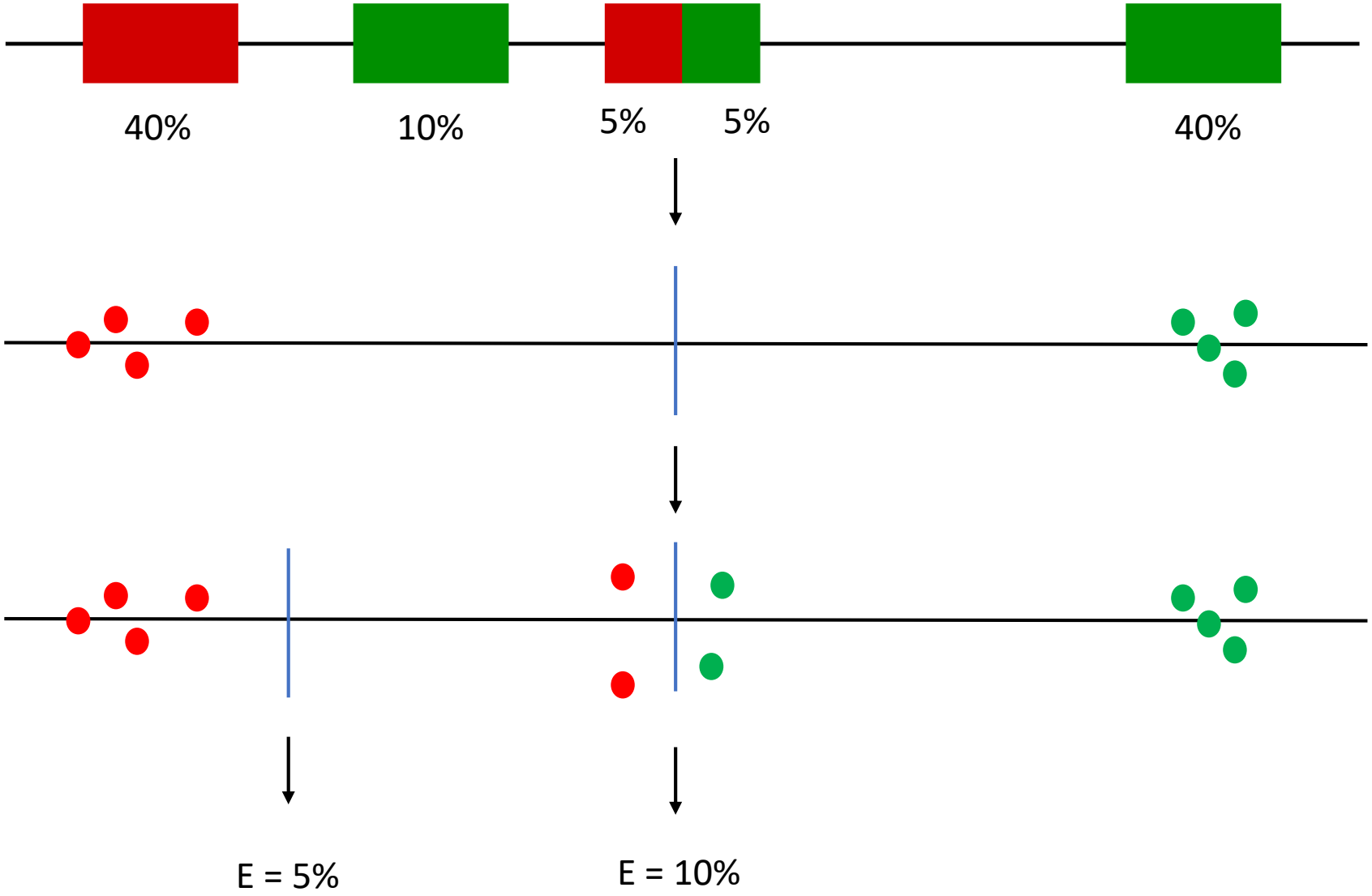


Active learning

Active learning – semi-supervised learning при котором мы сами выбираем вектора для оценки.

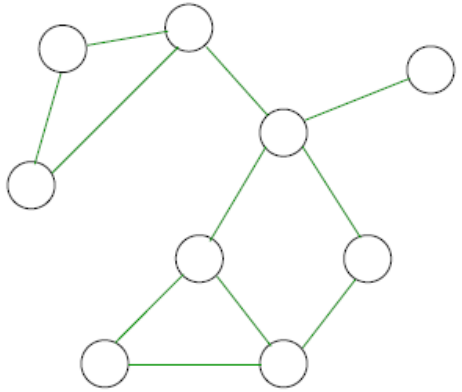


Sampling bias

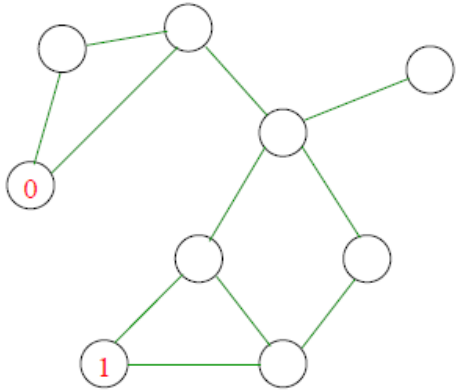


Распространение оценки
Label propagation

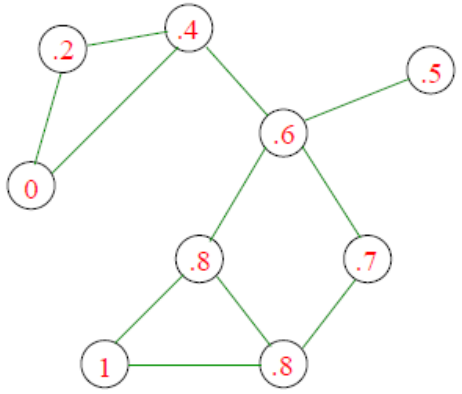
1) Построить граф близости



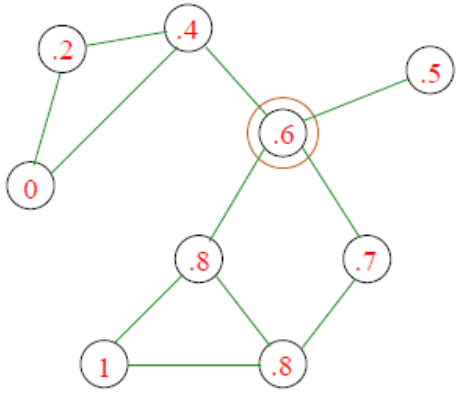
2) Оценить случайные точки



3) Распространить оценку

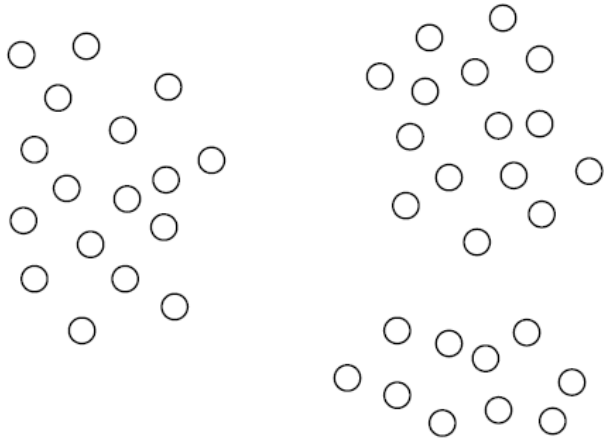


4) Оценить новую точку и вернуться в (3)

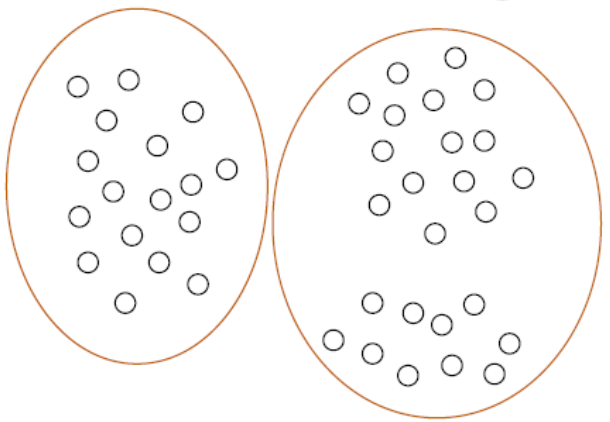


Переоценка кластеризации

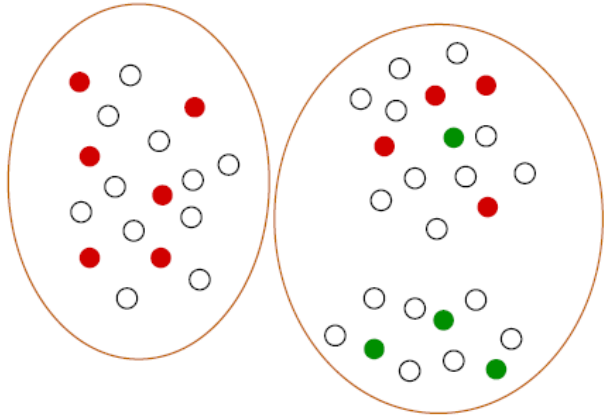
1) Неразмеченные данные



2) Найти кластеризацию



3) Разметить случайные точки в кластерах



4) Уточнить кластеризацию