Обучение без учителя

Лекция 7

Обучение без учителя (unsupervised learning)

Просто есть данные, без целевой переменной

Есть множество постановок задачи:

- Обучение ассоциативным правилам
- Кластеризация
- Визуализация
- Понижение размерности

Обучение ассоциативным правилам

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	shrimp	almonds	avocado	vegetables mix	green grapes	whole weat flour	yams	cottage cheese	energy drink	tomato juice	low fat yogurt	green tea	honey	salad
1	burgers	meatballs	eggs	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	chutney	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	turkey	avocado	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	mineral water	milk	energy bar	whole wheat rice	green tea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Параметры

1. $Support(A) = \frac{\text{Транзакции с } A}{\text{Все транзакции}}; Support(A \to B) = \frac{\text{Транзакции с } A \text{ и } B}{\text{Все транзакции}}$

2. $Confidence(A \to B) = \frac{\text{Транзакции с } A \text{ и } B}{\text{Транзакции с } A} = \frac{Support(A \to B)}{Support(A)}$

https://www.kaggle.com/code/evrenermis/association-rule-based-learning-explained

Обучение ассоциативным правилам

3. $Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)}$

$Lift(A \rightarrow B) = 1$ – товары независимы

$Lift(A \rightarrow B) > 1$ – товары зависимы положительно

$Lift(A \rightarrow B) < 1$ – товары зависимы негативно

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
6	(ground beef)	(mineral water)	0.098254	0.238368	0.040928	0.416554	1.747522
74	(ground beef)	(spaghetti)	0.098254	0.174110	0.039195	0.398915	2.291162
128	(ground beef)	(chocolate)	0.098254	0.163845	0.023064	0.234735	1.432669
175	(ground beef)	(milk)	0.098254	0.129583	0.021997	0.223881	1.727704

Кластеризация

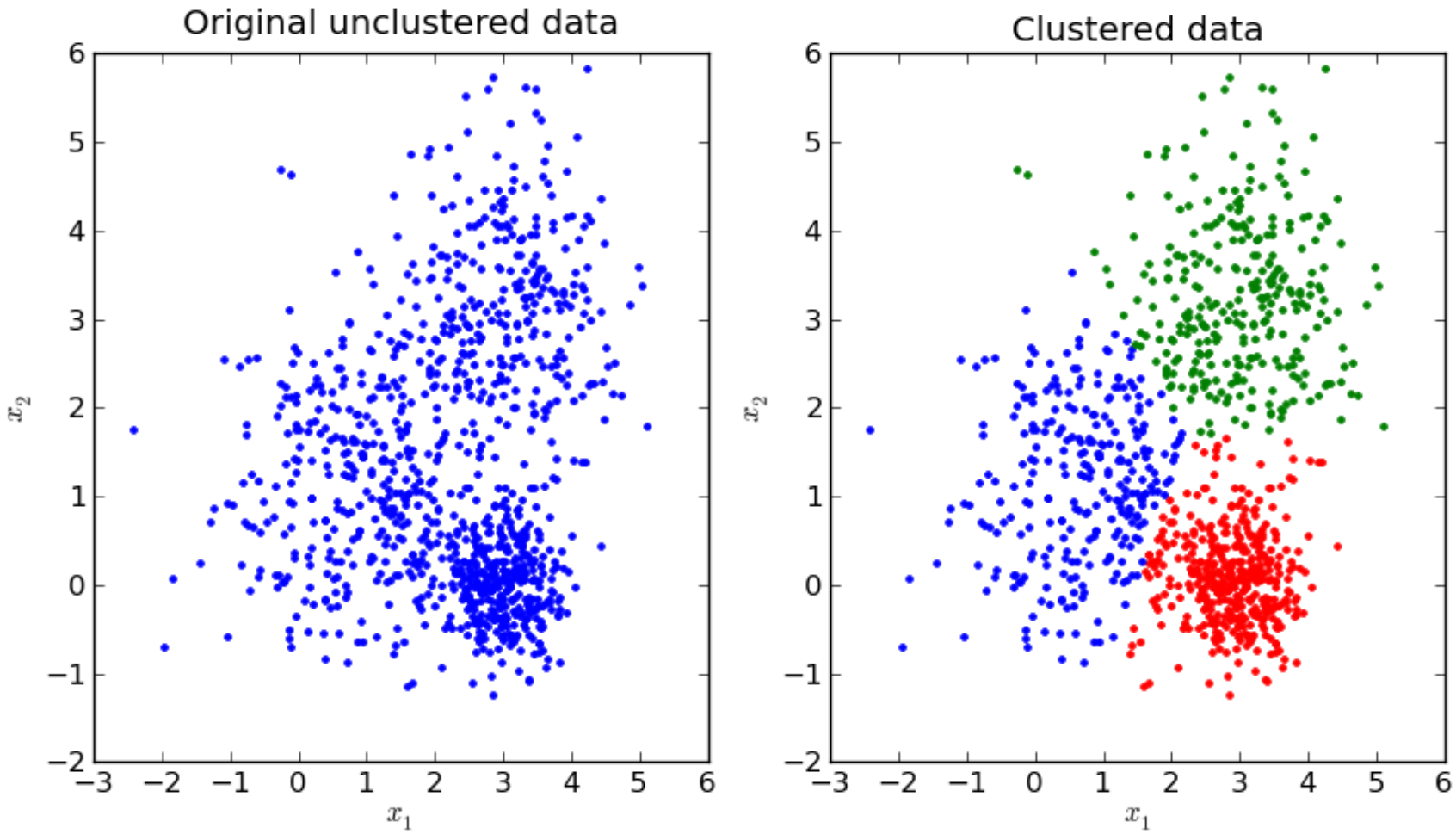
$X = (x_i)_{i=1}^{l}$

Необходимо построить такой алгоритм, который каждую точку соотнесет к одному из кластеров M: a: X $\rightarrow \{1, \ldots M\}$

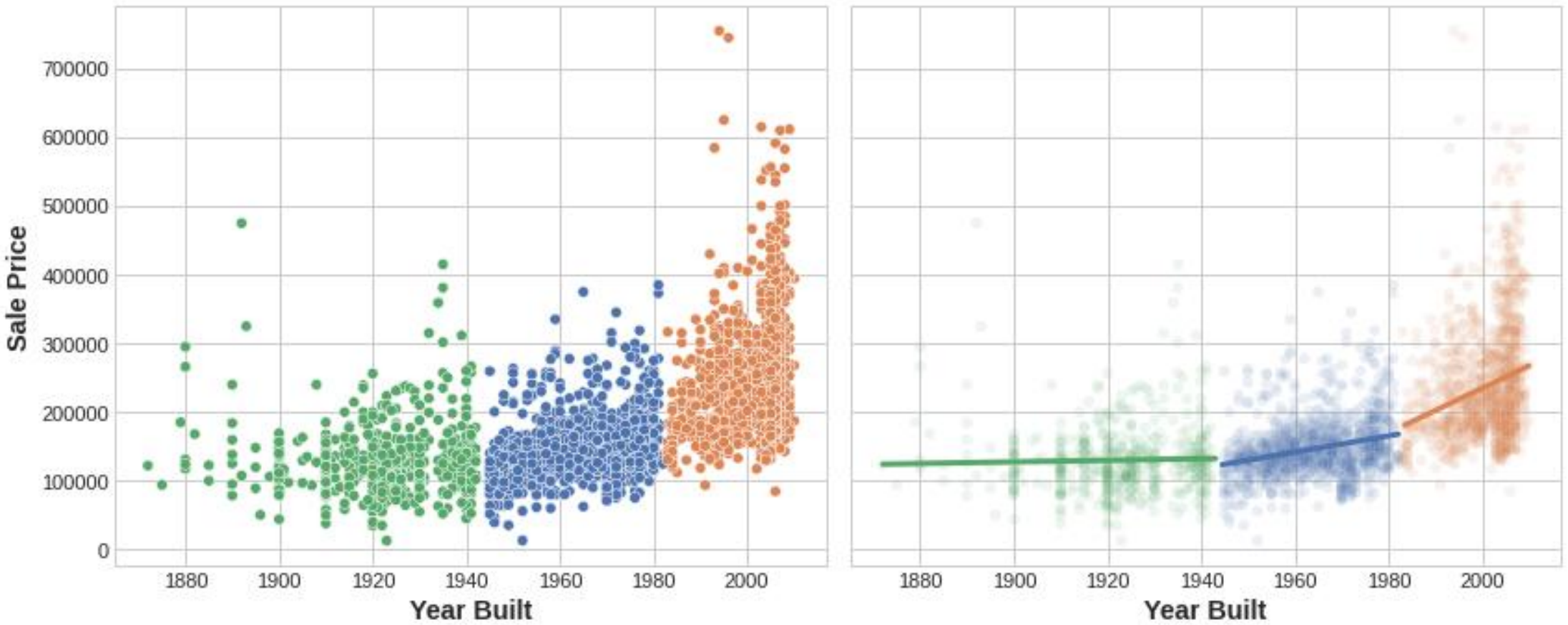
Примеры:

- Сегментация рынка
- Анализ социальных сетей
- Группировка результатов поиска
- Обнаружение аномалий

Пример кластеризации



Пример кластеризации



Кластеризация в общем

Выражение для учета расстояний между всеми парами точек:

$$T = \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} d_{ii'} = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\left(\sum_{C(i')=k} d_{ii'} + \sum_{C(i')\neq k} d_{ii'}\right)$$

$$T = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d_{ii'} + \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')\neq k} d_{ii'}$$

$$T = W(C) + B(C)$$

Кластеризация в общем

$$T = W(C) + B(C)$$

Внутрикластерное расстояние:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d_{ii'}$$

Межкластерное расстояние:

$$B(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')\neq k} d_{ii'}$$

Кластеризация. Решение

Можно решить оптимизационную задачу, пытаясь найти наилучшее разбиение N точек по K кластерам. Но для этого нужно перебрать все варианты, которых будет:

$$S(N,K) = \frac{1}{K!}\sum_{k=1}^{K}(-1)^{(K-k)}\binom{K}{k}k^N$$

Так например $S(10,4) = 34105$.

Но $S(19,4) \cong 10^{10}$

K-means

Будем считать расстояние как евклидово:

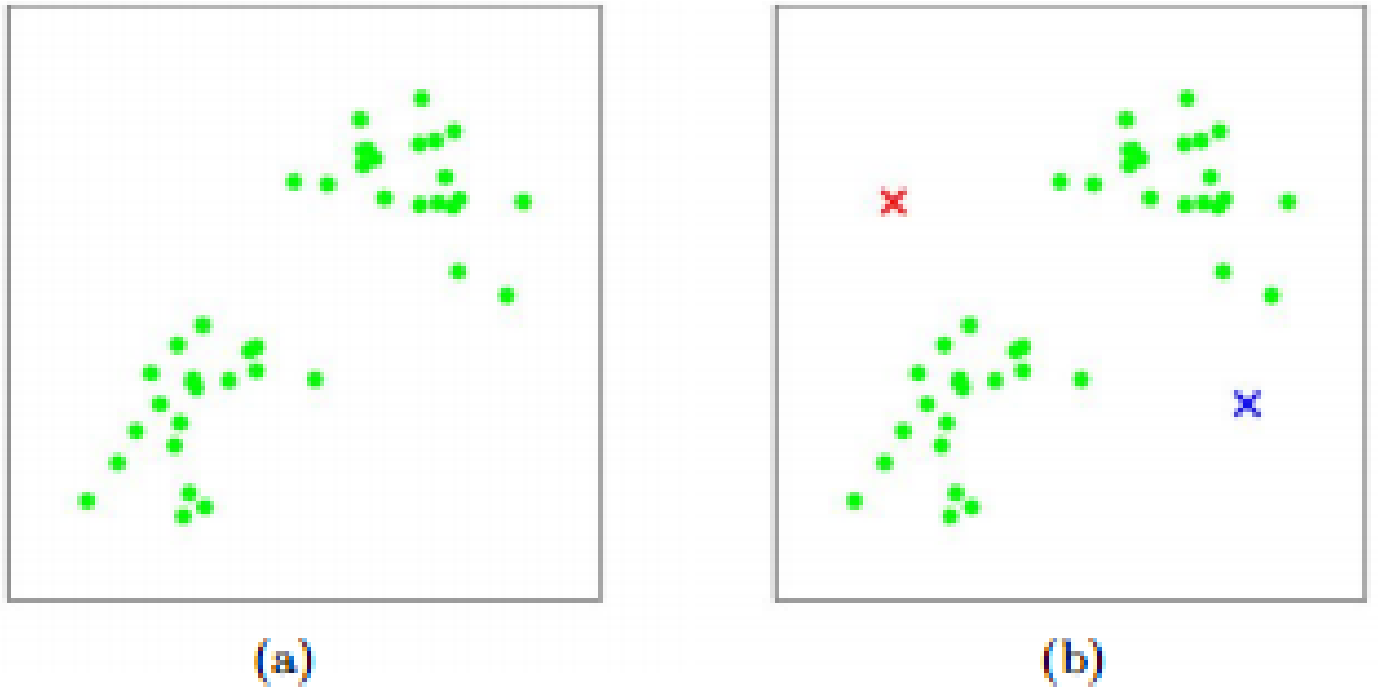
$$d(x_i, x_{i'}) = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 = \|x_i - x_{i'}\|^2$$

Распишем внутрикластерное расстояние как:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} \|x_i - x_{i'}\|^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2$$

K-means. Шаг 0

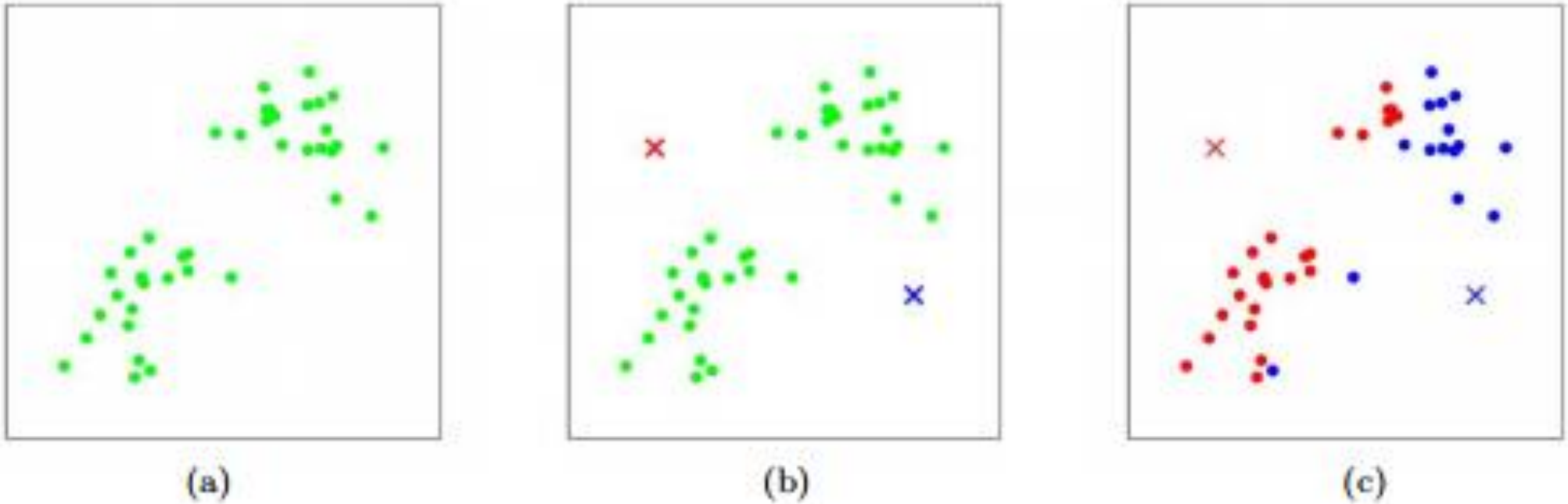
Случайно инициализируем центры кластеров $C_1, \ldots, C_k$.



K-means. Шаг 1

Обновляем принадлежности точек к кластерам

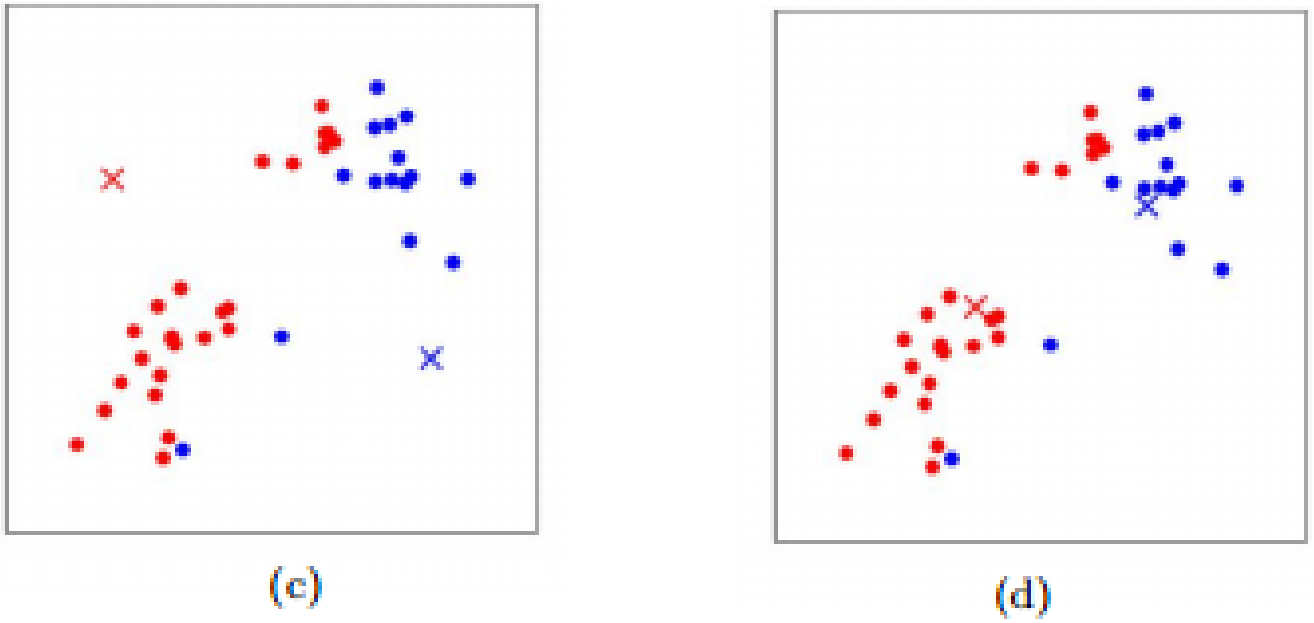
$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} \|x_i - m_k\|^2$$



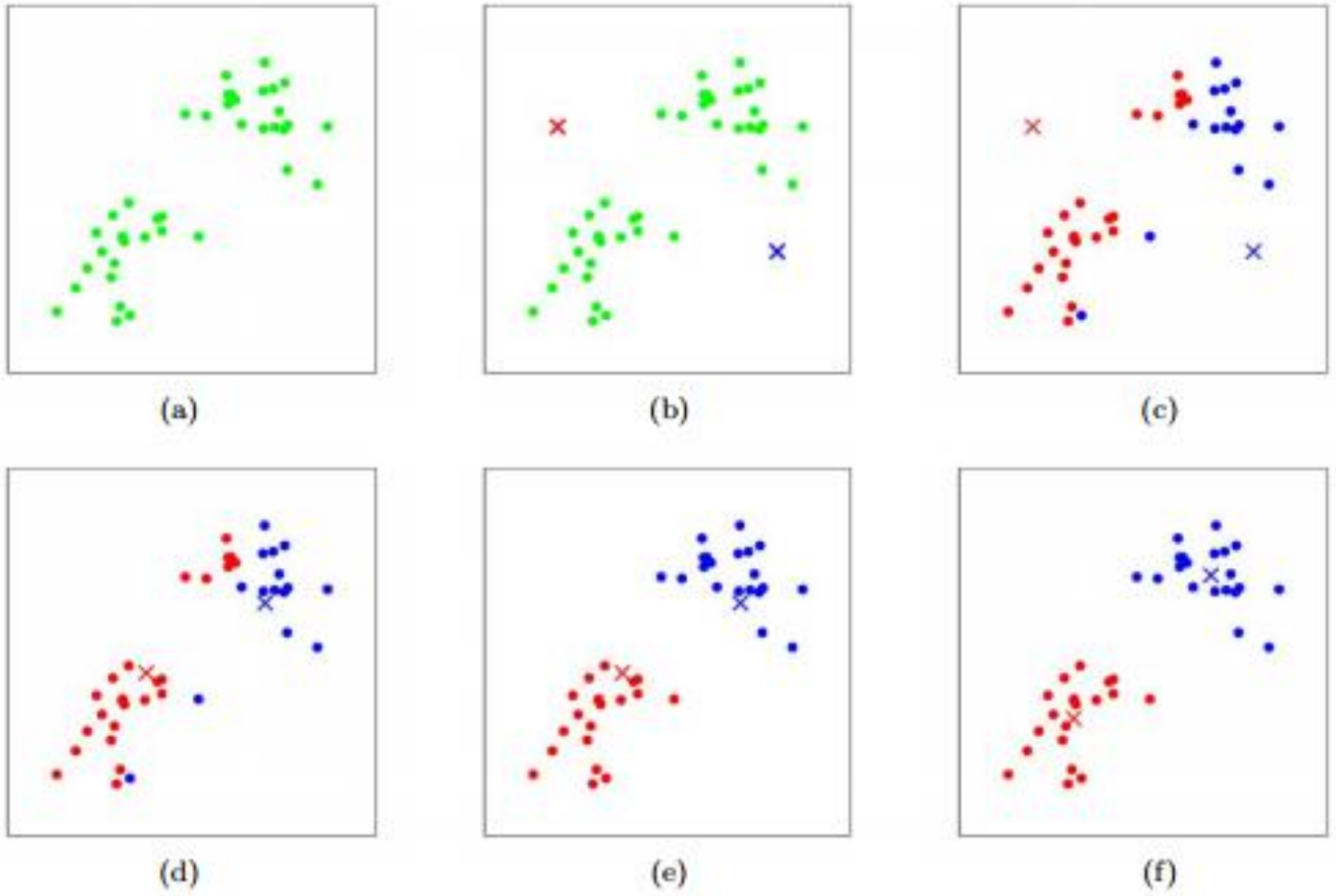
K-means. Шаг 2

Обновляем центры кластеров

$$\bar{x}_S = \underset{m}{\operatorname{argmin}} \sum_{i \in S} \|x_i - m_k\|^2$$



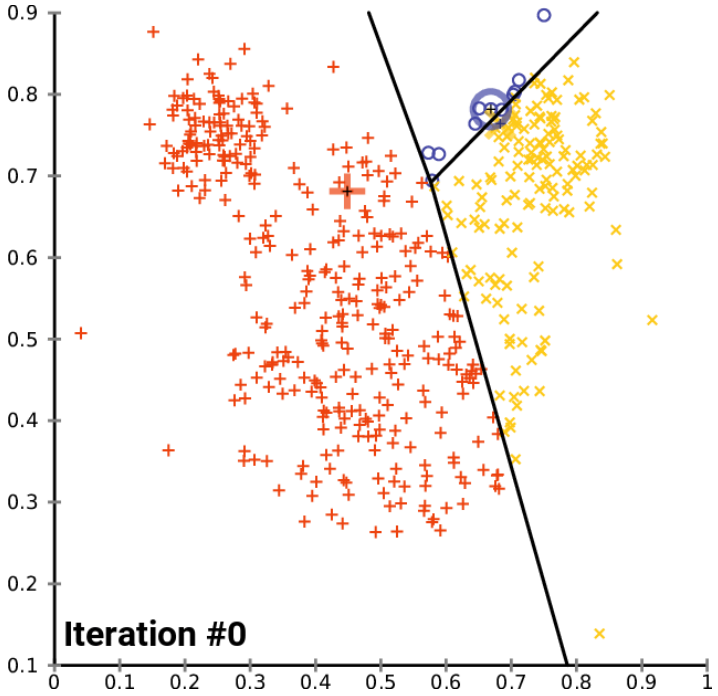
K-means. Повторяем шаги

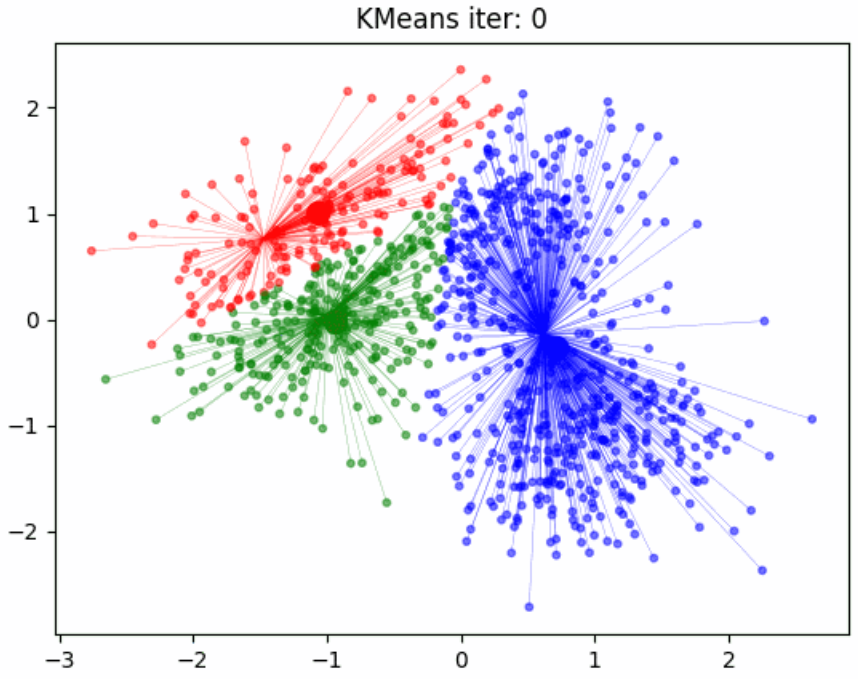


K-means. Повторяем шаги

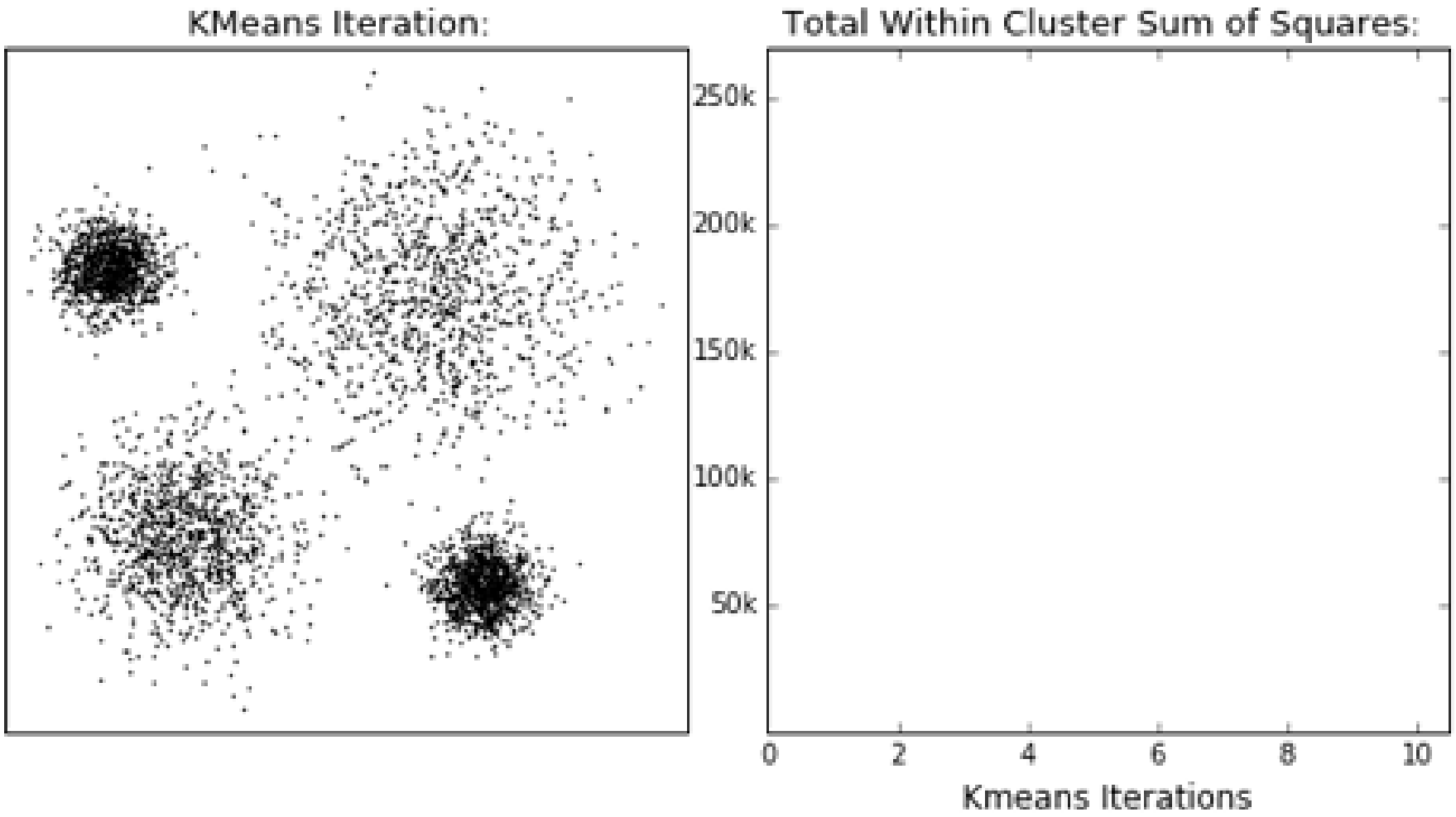
В общем решаем задачу:

$$C^* = \min_{C,\{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - m_k\|^2$$

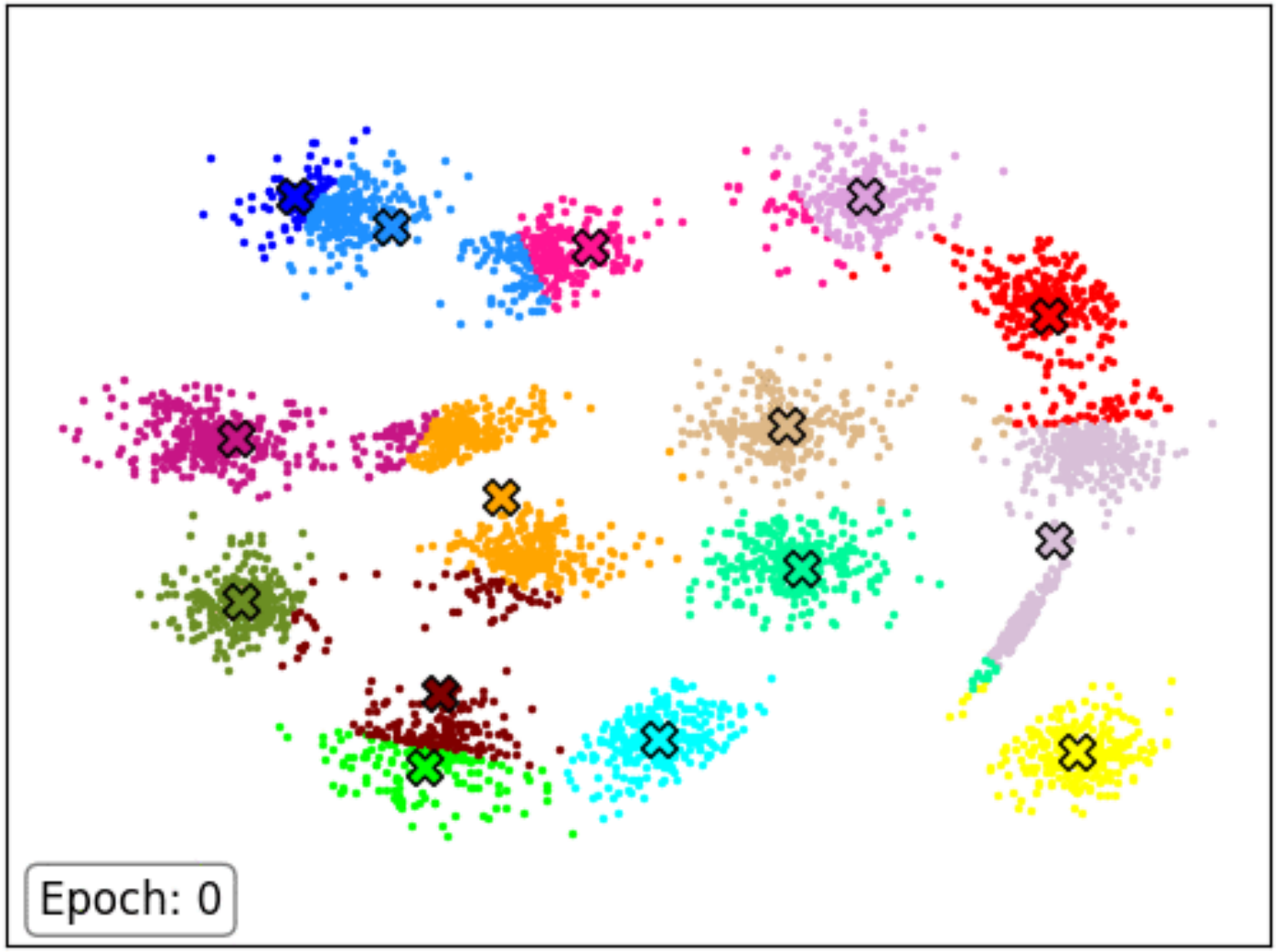




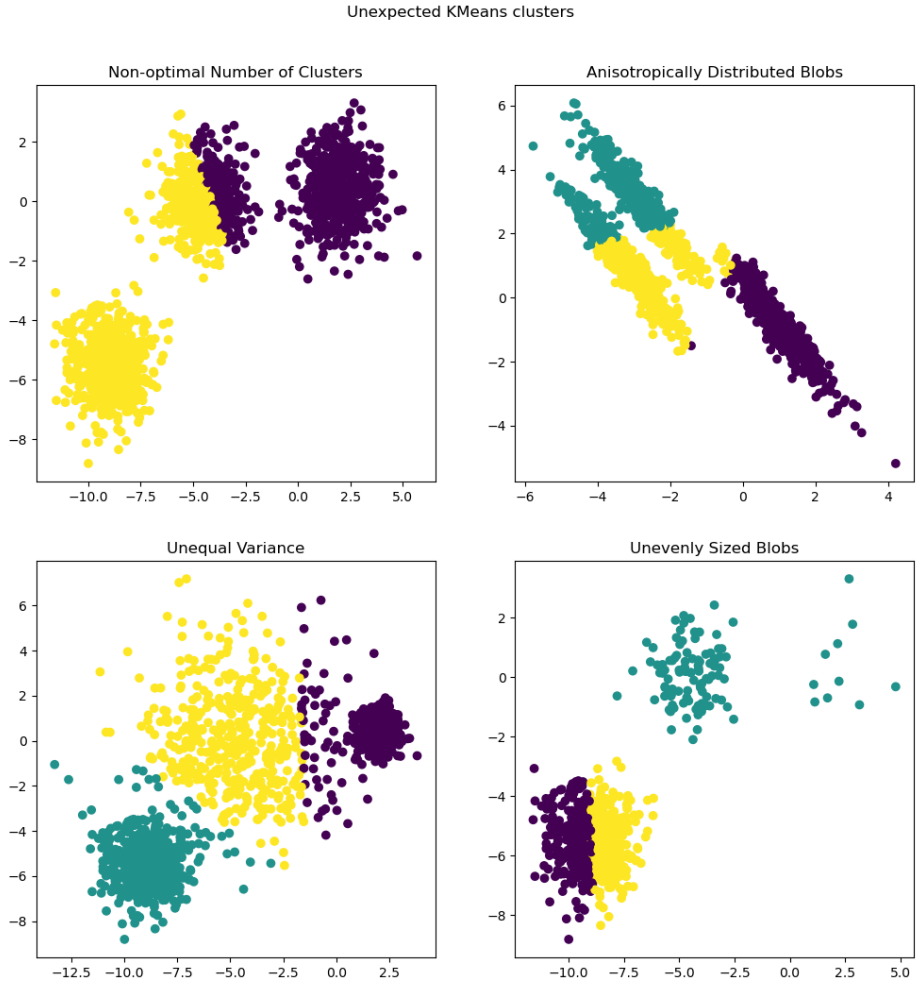
K-means. В картинках



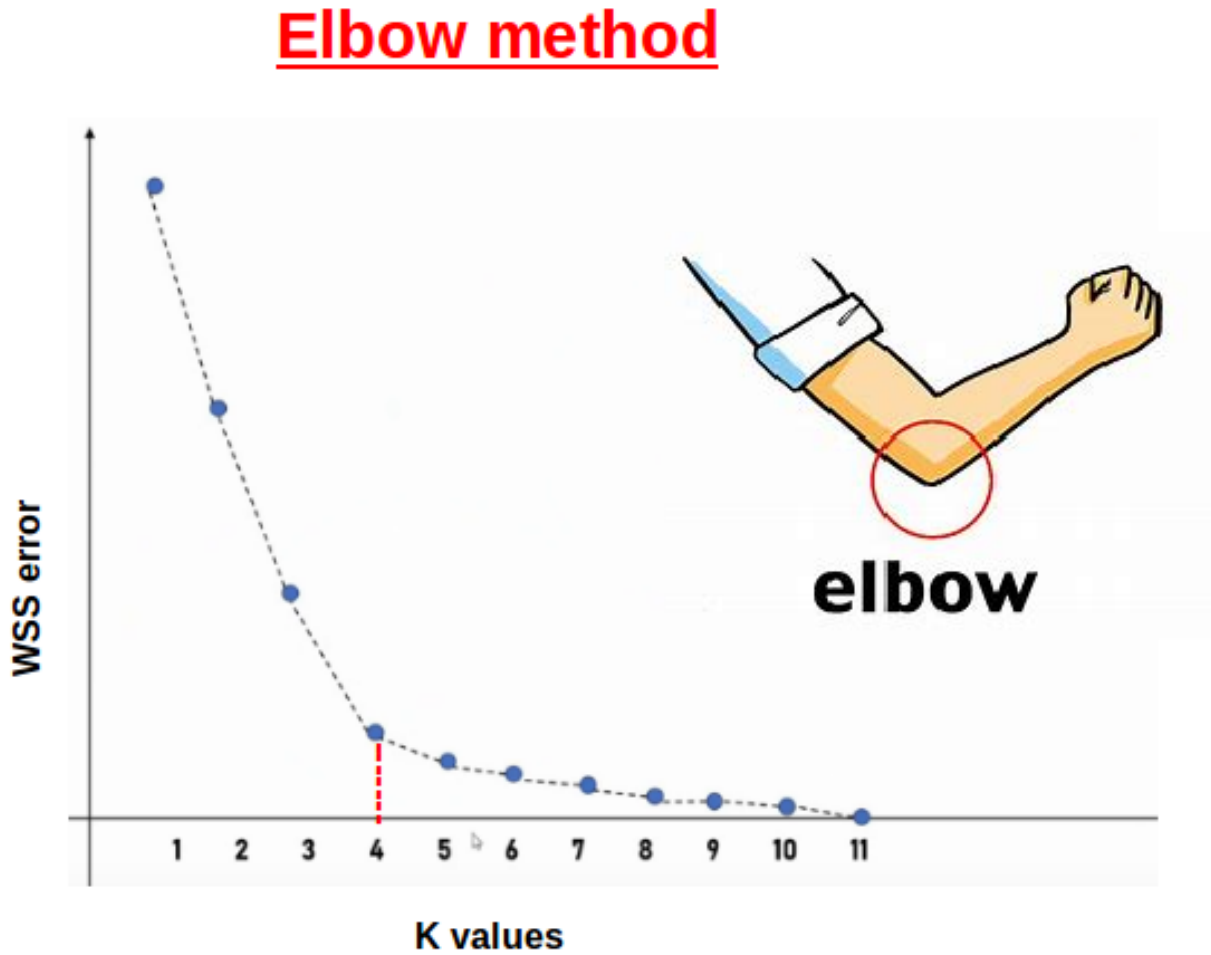
K-means++



K-means. Проблемы



K-means. Как найти K



Метрики в кластеризации

Внутрикластерное расстояние:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d_{ii'}$$

Межкластерное расстояние:

$$B(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')\neq k} d_{ii'}$$

Метрики в кластеризации

Коэффициент силуэта:

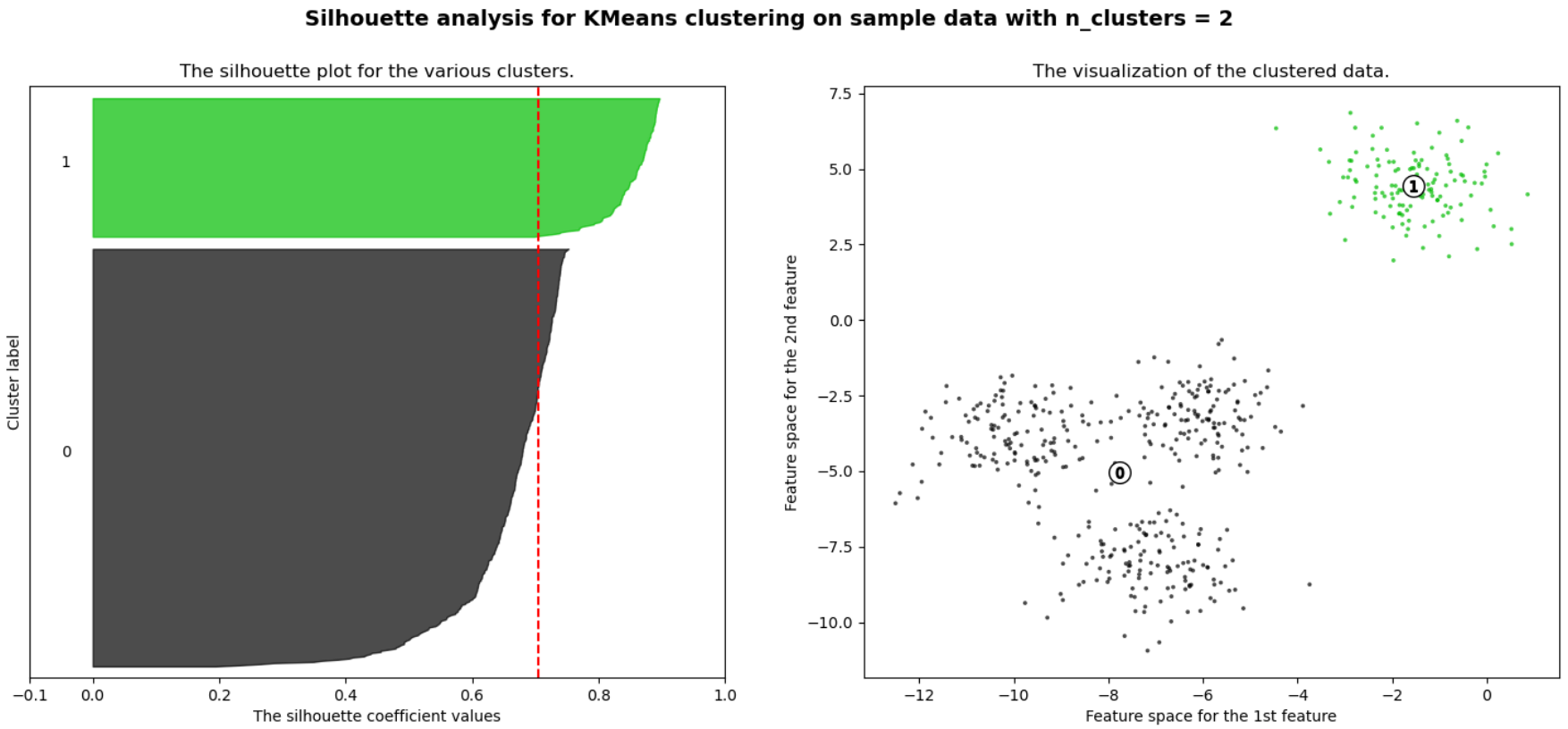
$$S(x_i) = \frac{B(x_i) - A(x_i)}{\max(B(x_i), A(x_i))},$$

где $A(x_i)$ – среднее расстояние между x_i и точками его кластера

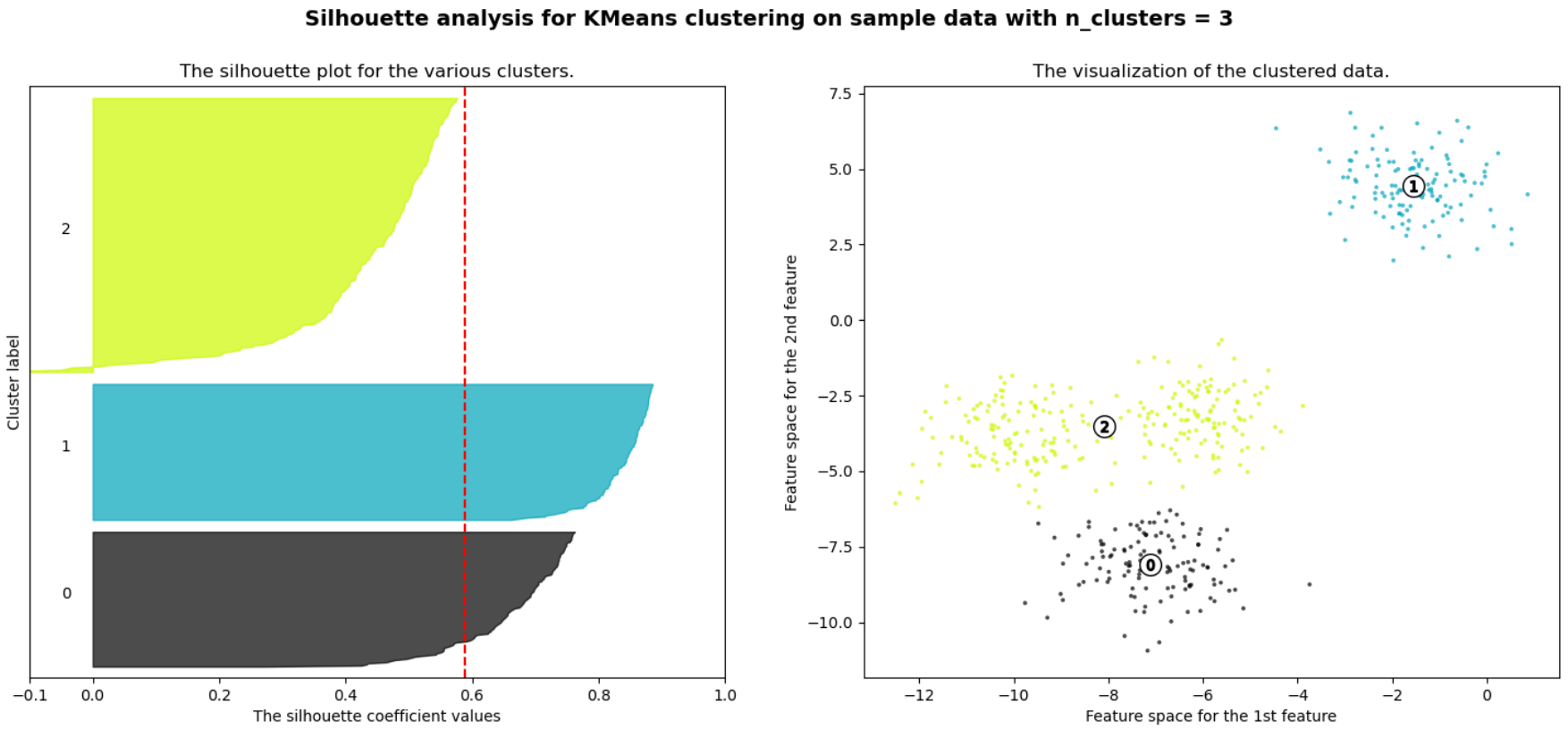
$B(x_i)$ – среднее расстояние между x_i и точками ближайшего кластера

Само значение $S(x_i) \in [-1; 1]$ и чем больше, тем лучше

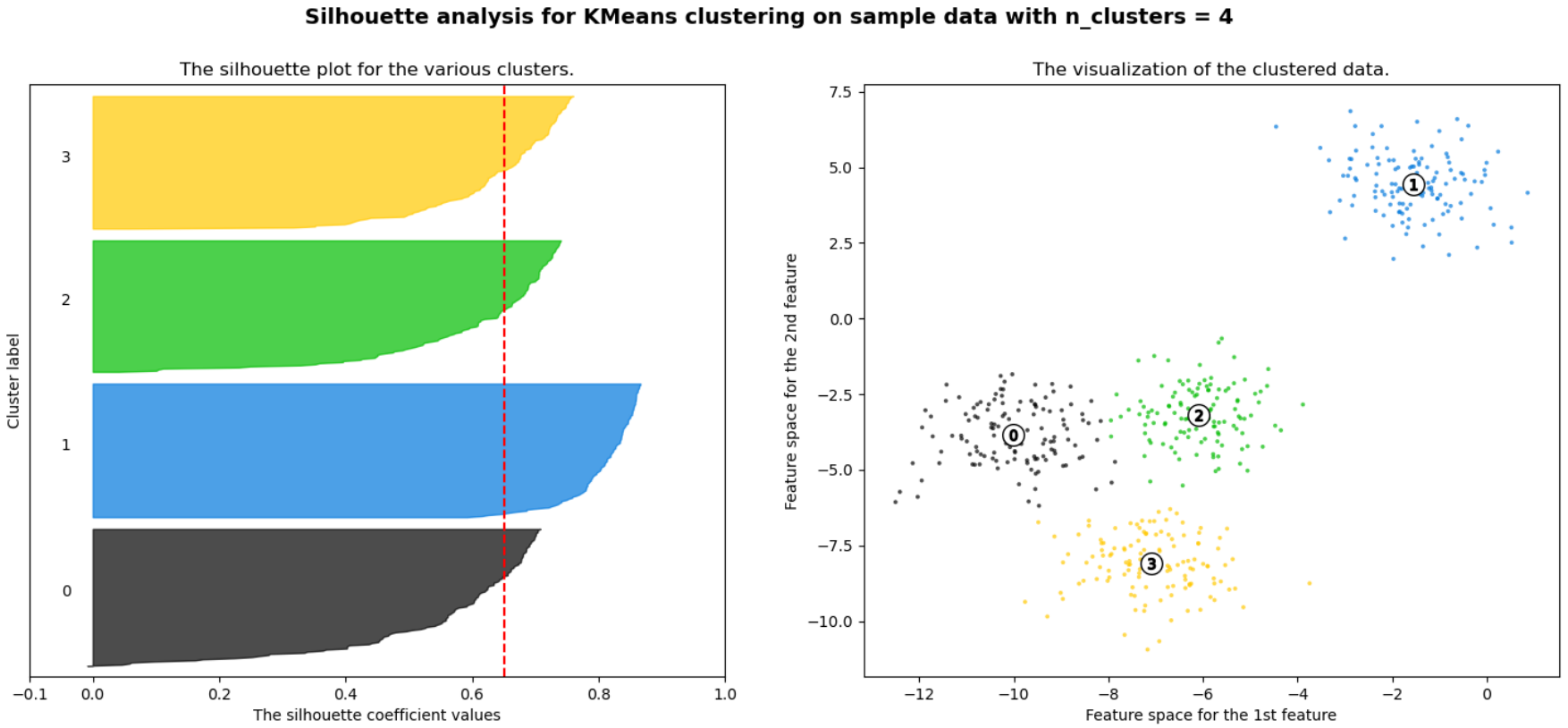
Коэффициент силуэта, 2 кластера



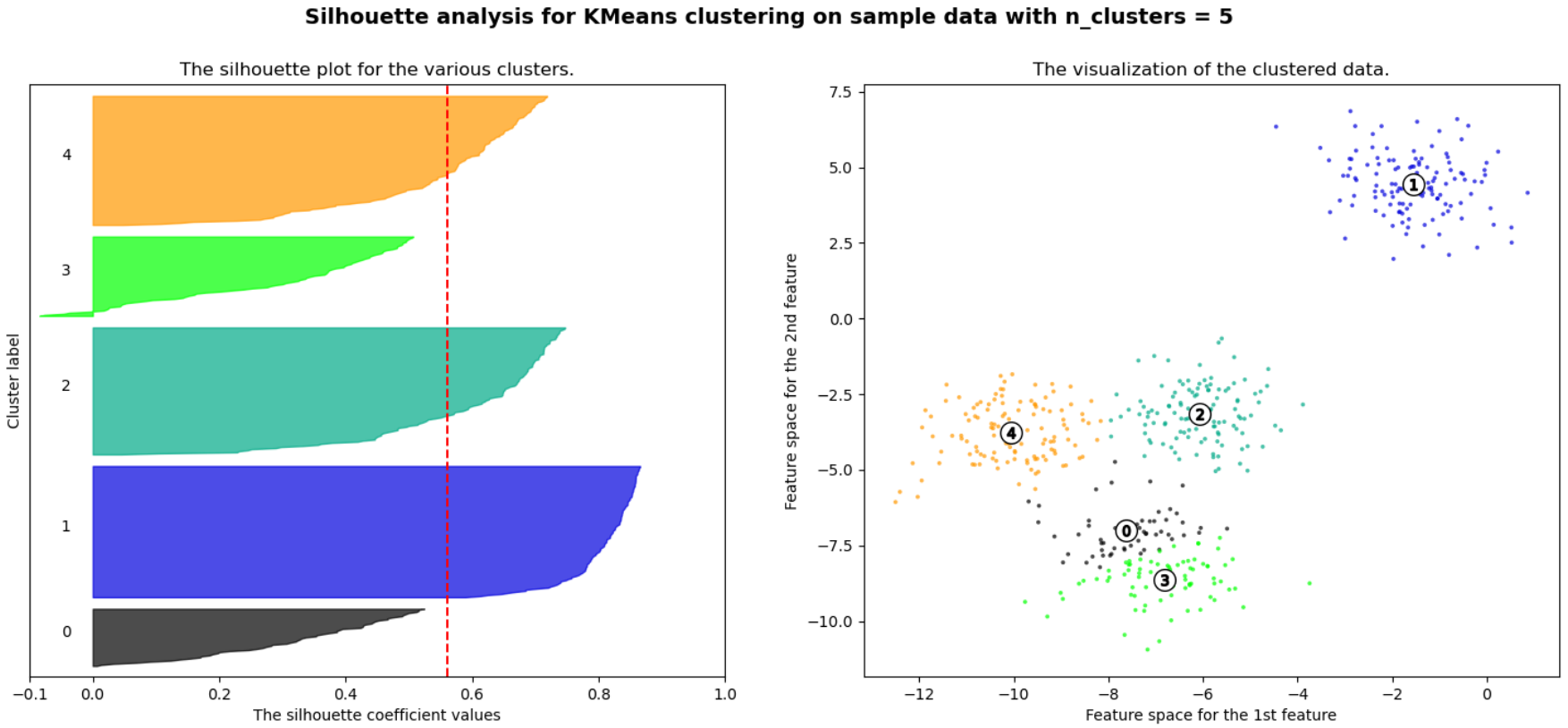
Коэффициент силуэта, 3 кластера



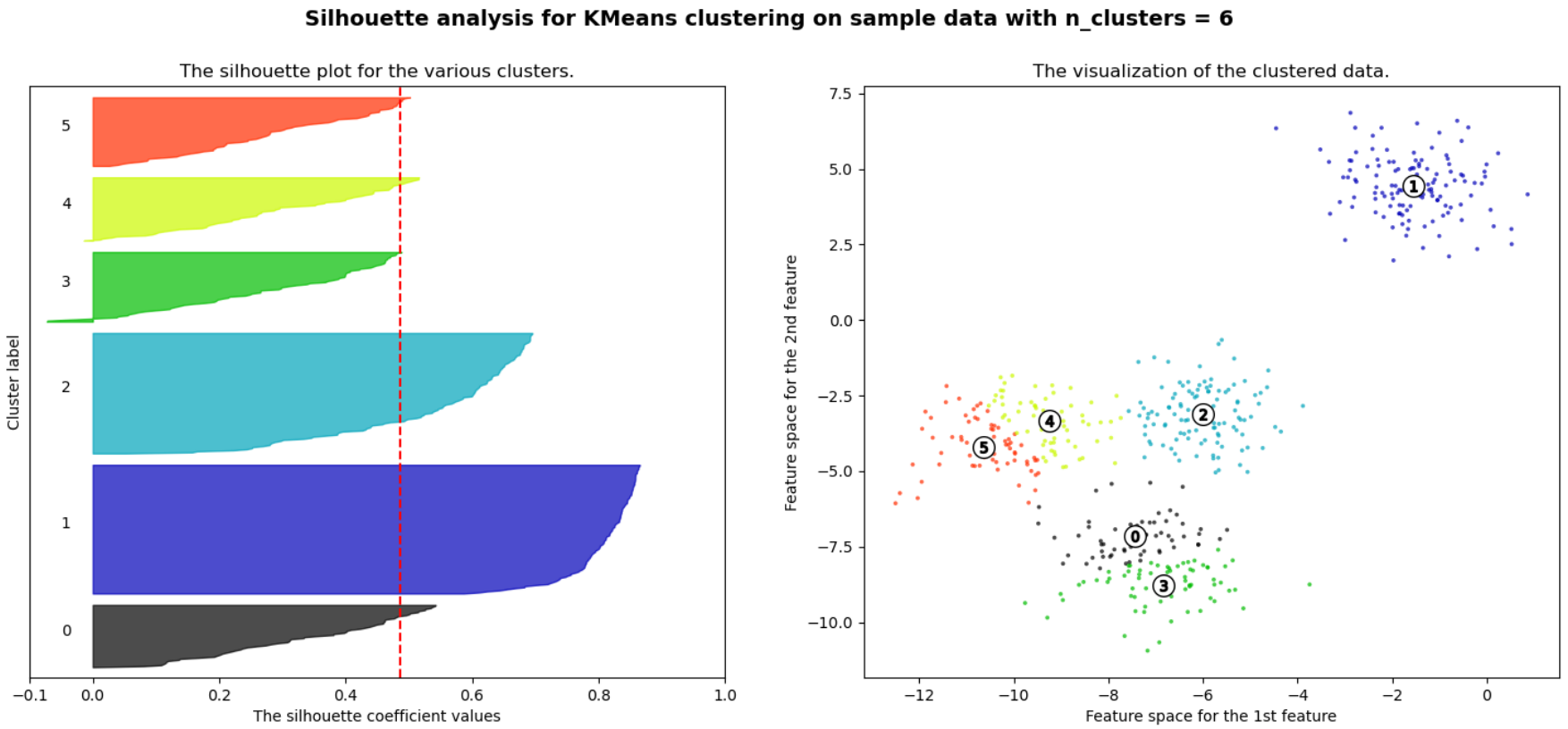
Коэффициент силуэта, 4 кластера



Коэффициент силуэта, 5 кластера



Коэффициент силуэта, 6 кластера



Метрики в кластеризации

Индекс Данна:

$$Dunn\ Index = \frac{\min\limits_{1 \le k < k' \le K} d(k, k')}{\max\limits_{1 \le k \le K} d(k)},$$

где $d(k, k')$ – межкластерное расстояние

$d(k)$ – внутрикластерное расстояние

Чем больше, тем лучше

Метрики в кластеризации. Есть разметка

Введем формулы энтропии:

$$H_{class} = -\sum_{c=1}^{C} \frac{m_c}{n} \log \frac{m_c}{n}$$

$$H_{clust} = -\sum_{k=1}^{K} \frac{n_k}{n} \log \frac{n_k}{n}$$

$$H_{class|clust} = -\sum_{c=1}^{C} \sum_{k=1}^{K} \frac{n_{ck}}{n_k} \log \frac{n_{ck}}{n_k}$$

Метрики в кластеризации. Есть разметка

Гомогенность:

$$Homogeneity = 1 - \frac{H_{class|clust}}{H_{class}}$$

Лучший вариант – каждый кластер содержит элементы только одного класса

Полнота:

$$Completeness = 1 - \frac{H_{clust|class}}{H_{clust}}$$

Лучший вариант – все объекты класса содержатся в одном кластере

Метрики в кластеризации. Есть разметка

V-мера:

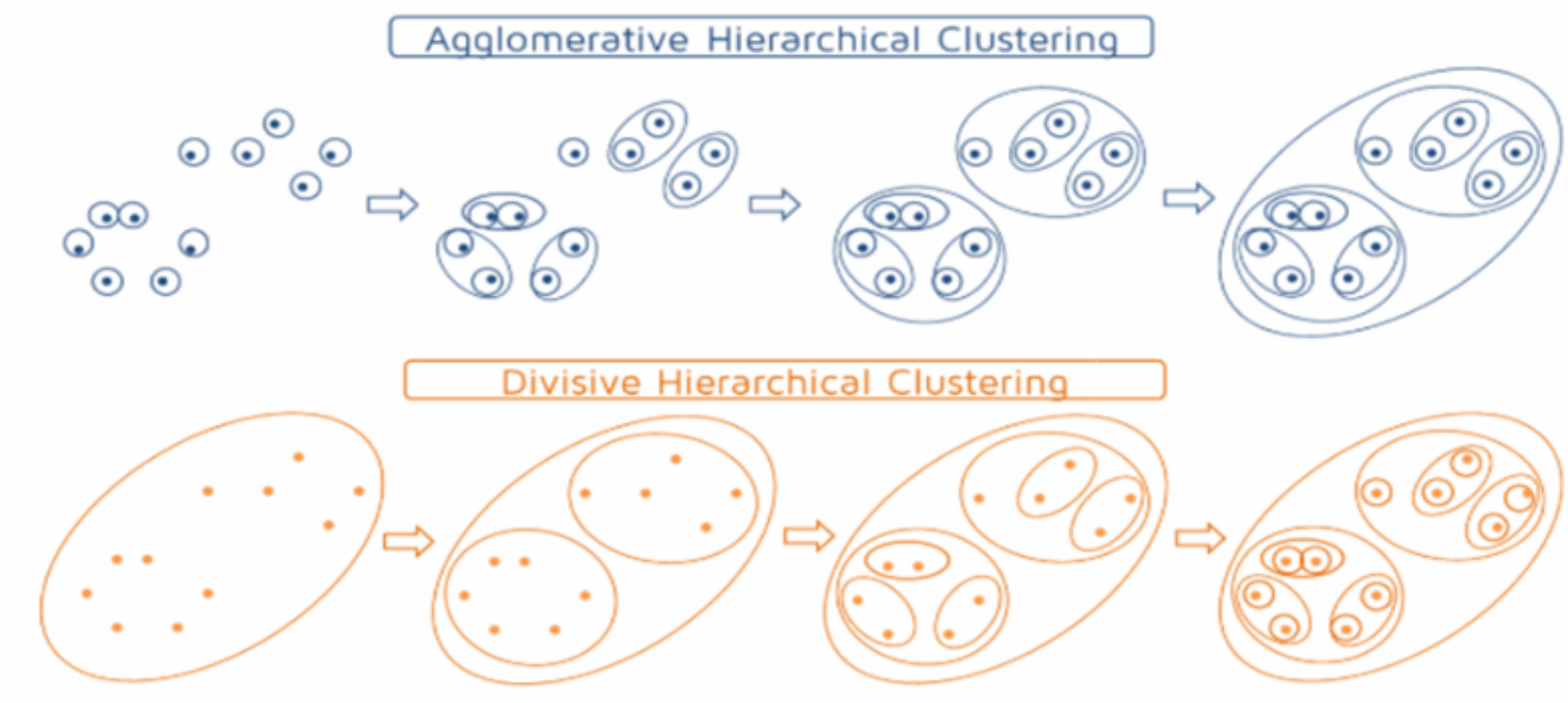
$$V_\beta = \frac{(1+\beta)\cdot Homogeneity \cdot Completeness}{\beta \cdot Homogeneity + Completeness}$$

$$V_1 = 2\frac{Homogeneity \cdot Completeness}{Homogeneity + Completeness}$$

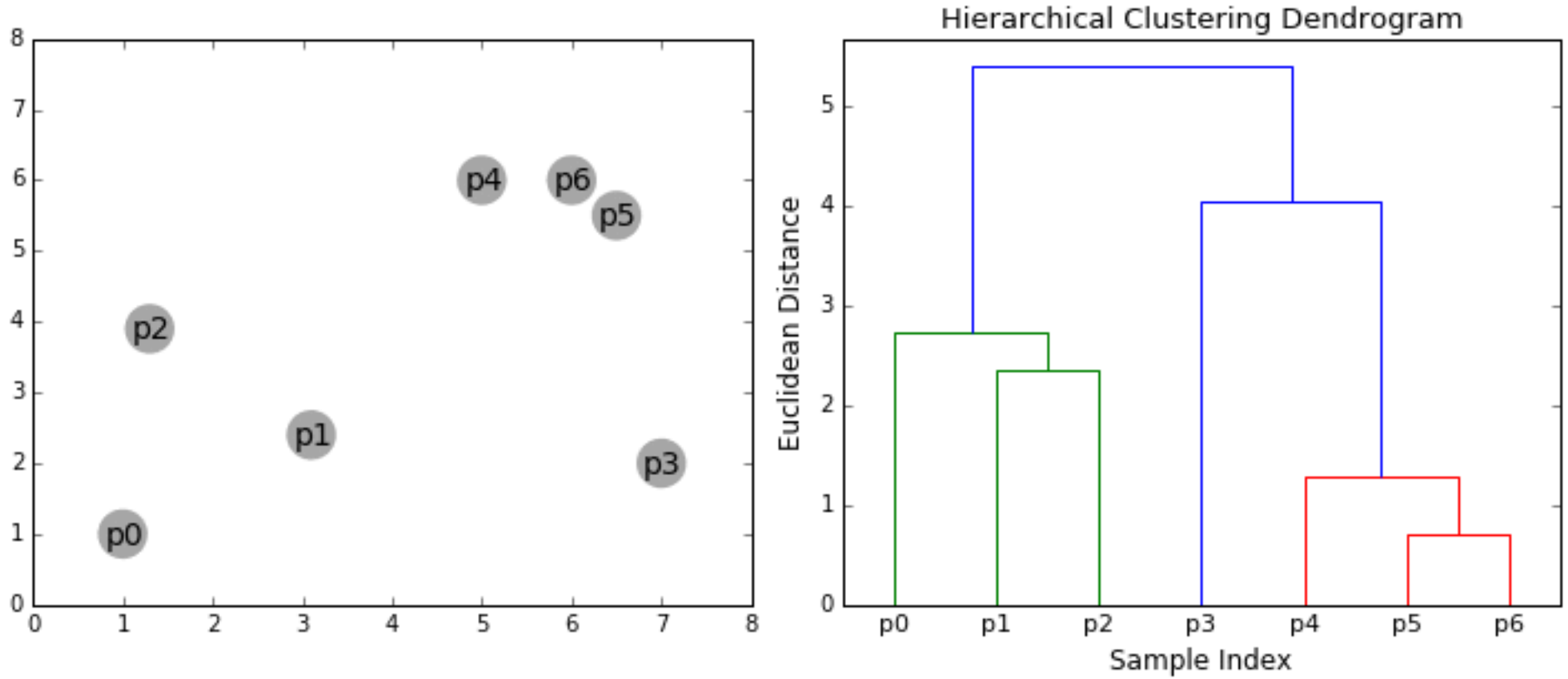
Дополнительно про метрики:

https://habr.com/ru/companies/yandex/articles/500742/

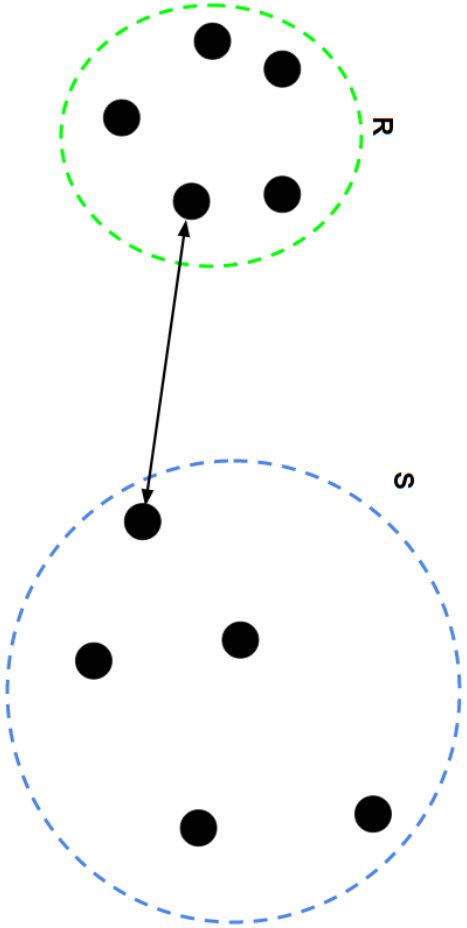
Иерархические алгоритмы кластеризации

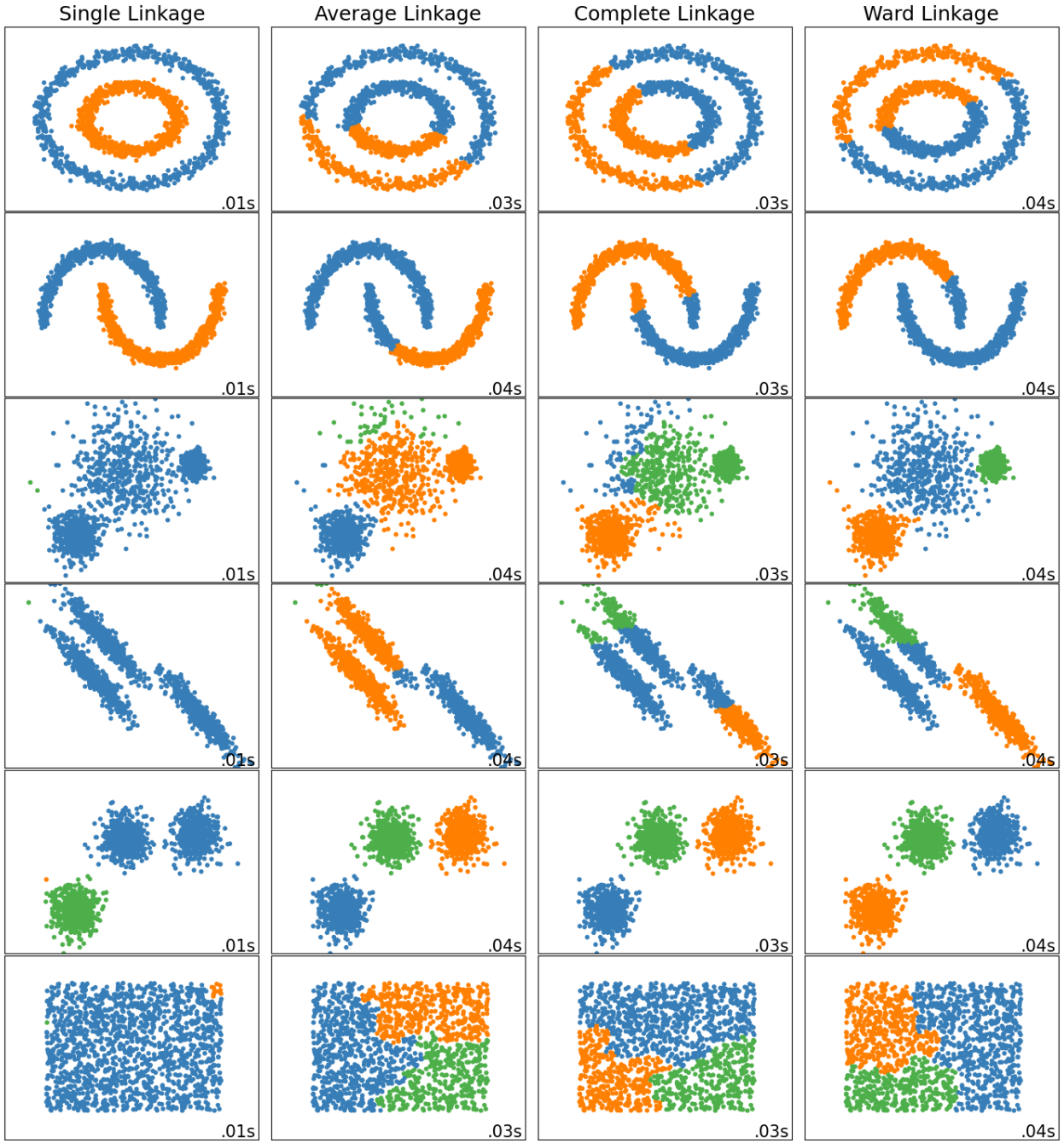


Дендрограммы

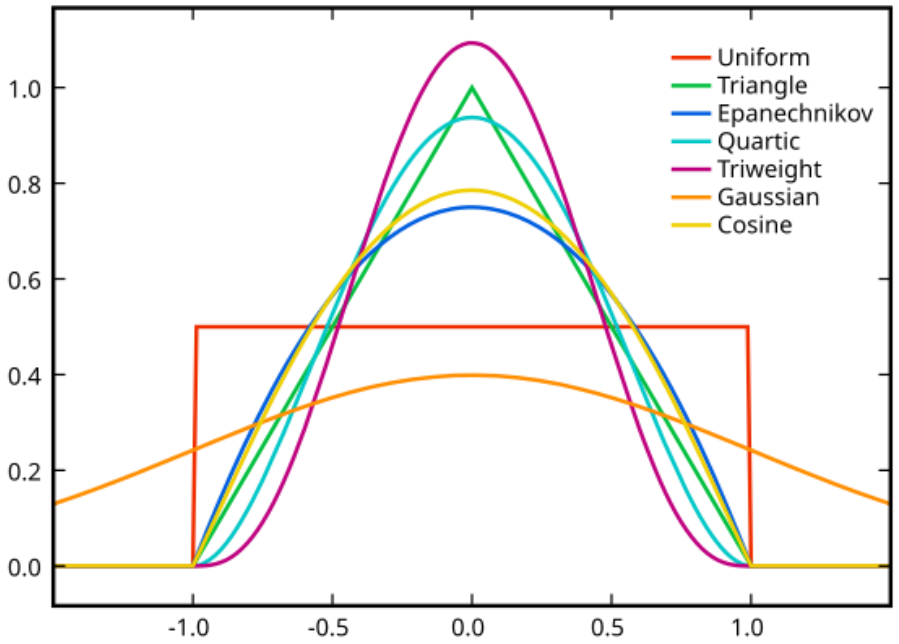


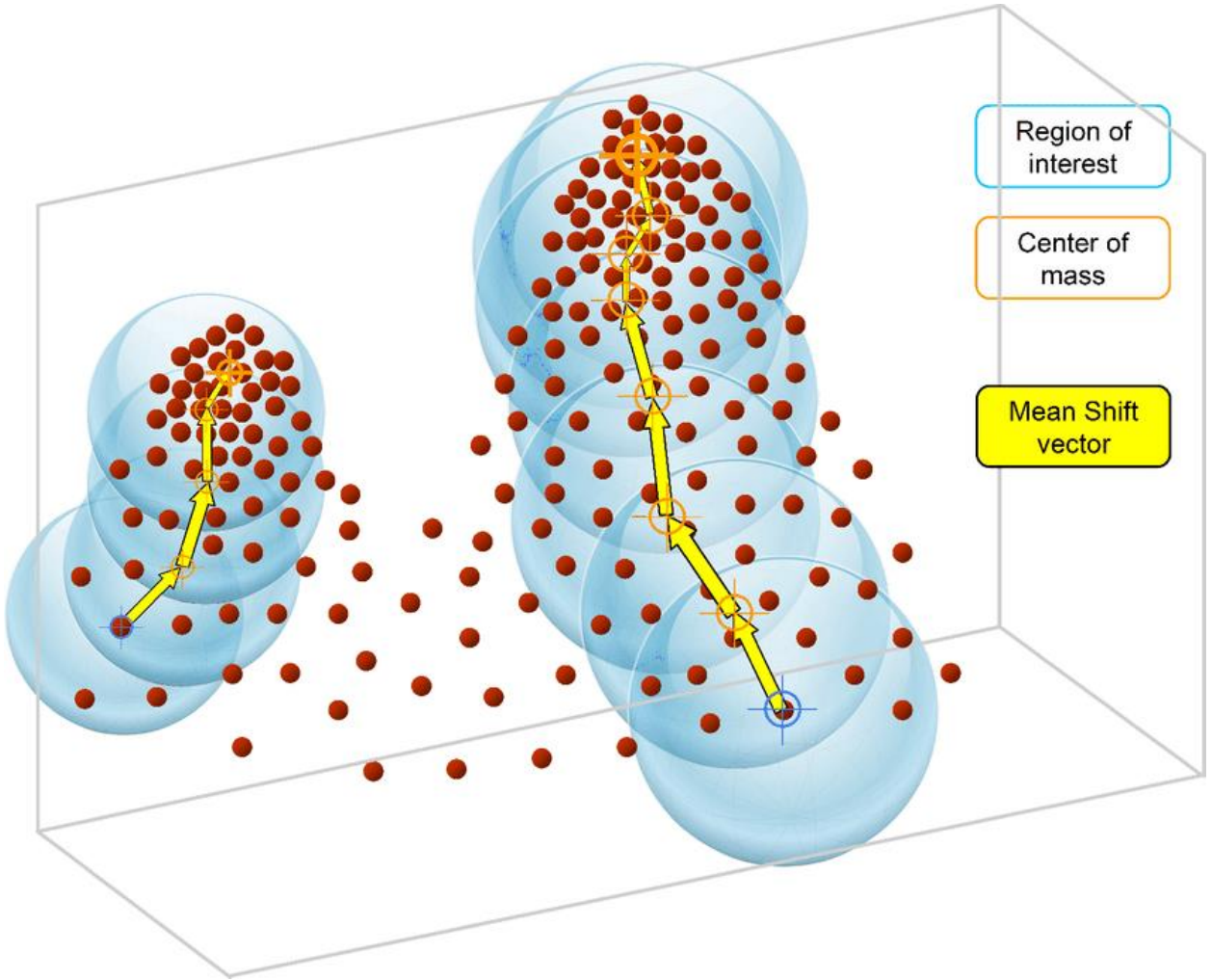
Можно по-разному считать межкластерные расстояния



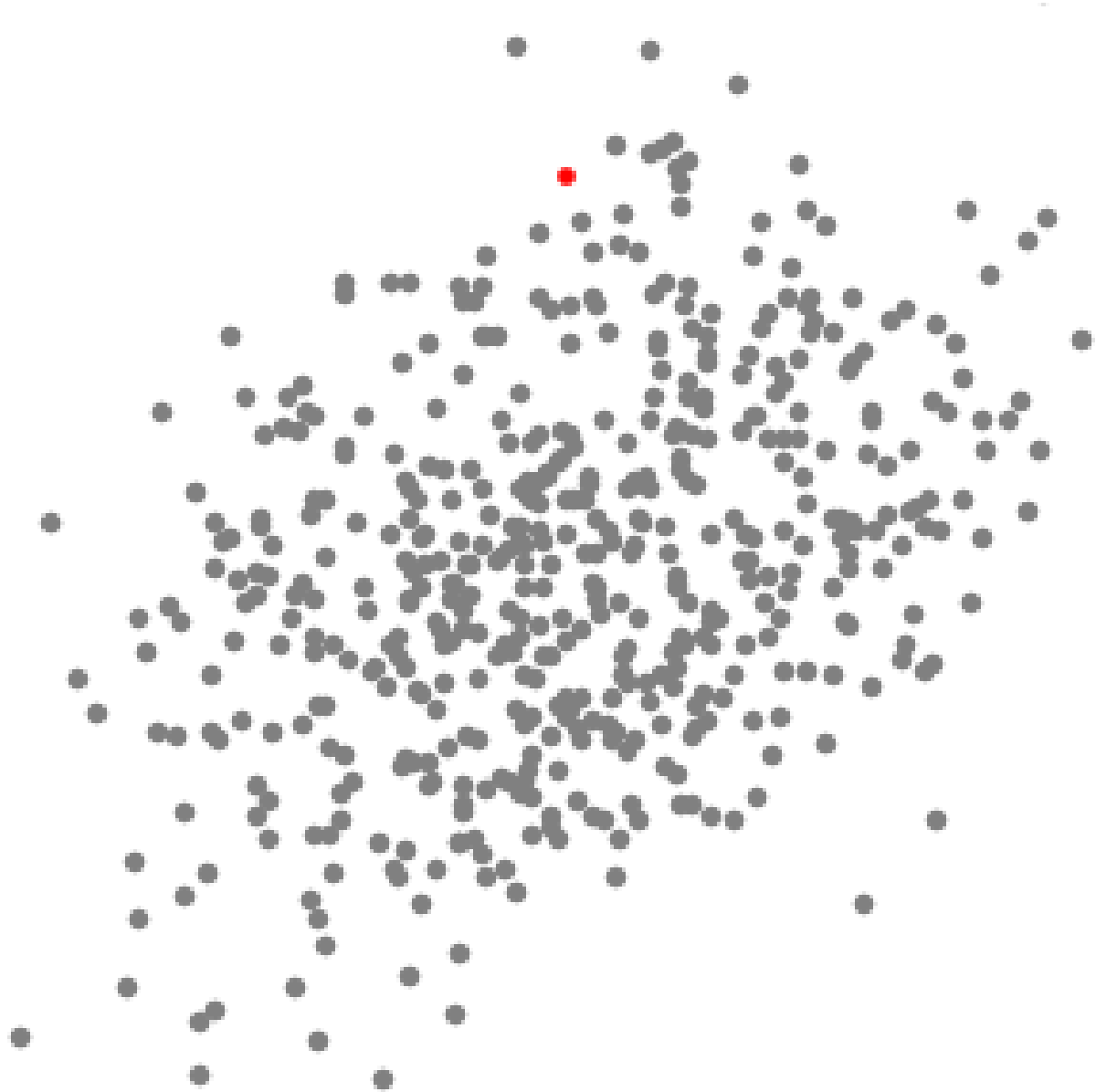


Mean Shift

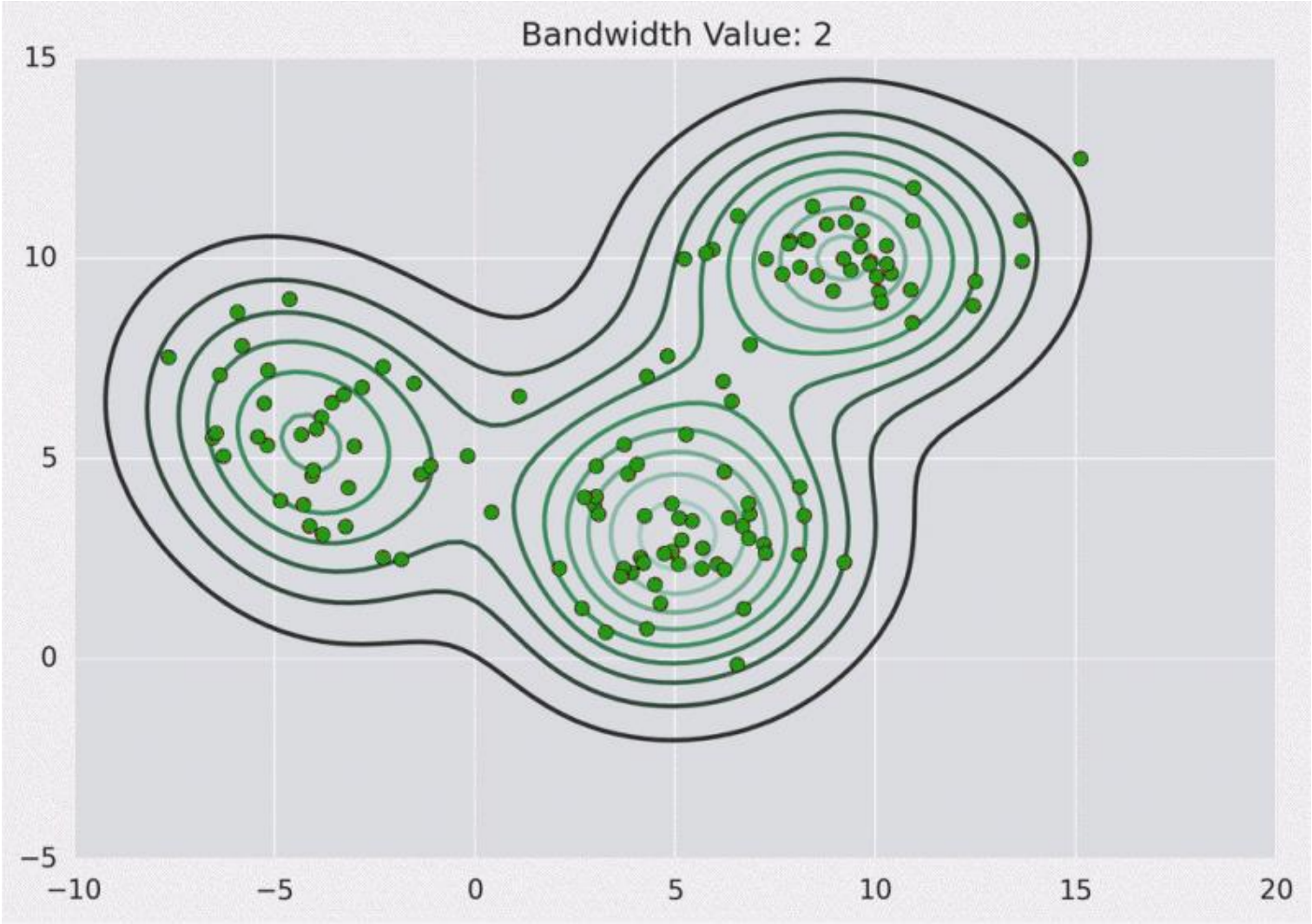

Uniform
Triangle
Epanechnikov
Quartic
Triweight
Gaussian
Cosine
1.0
0.8
0.6
0.4
0.2
0.0
-1.0
-0.5
0.0
0.5
1.0


Region of interest
Center of mass
Mean Shift vector

Mean Shift



Mean Shift

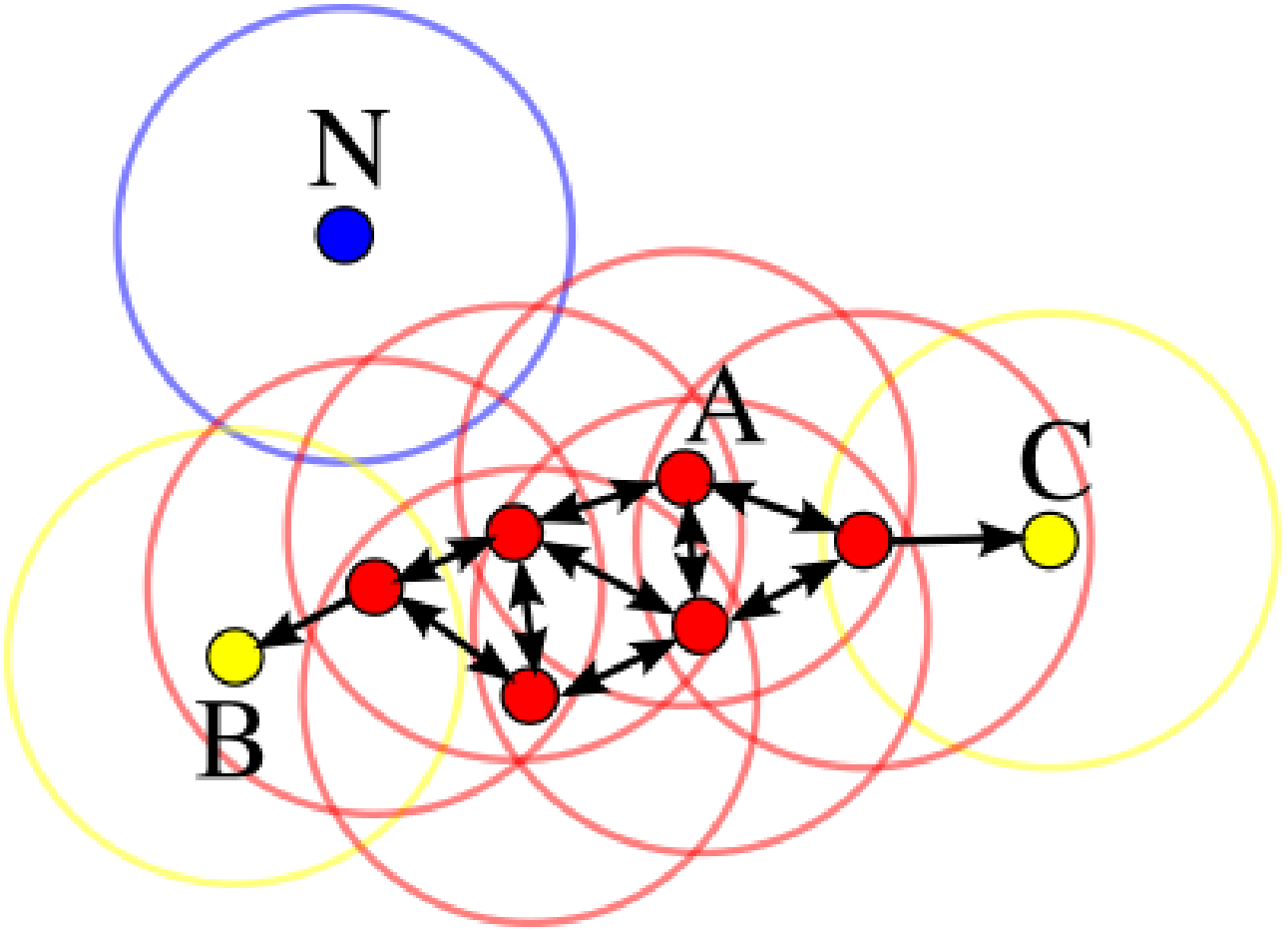


DBSCAN

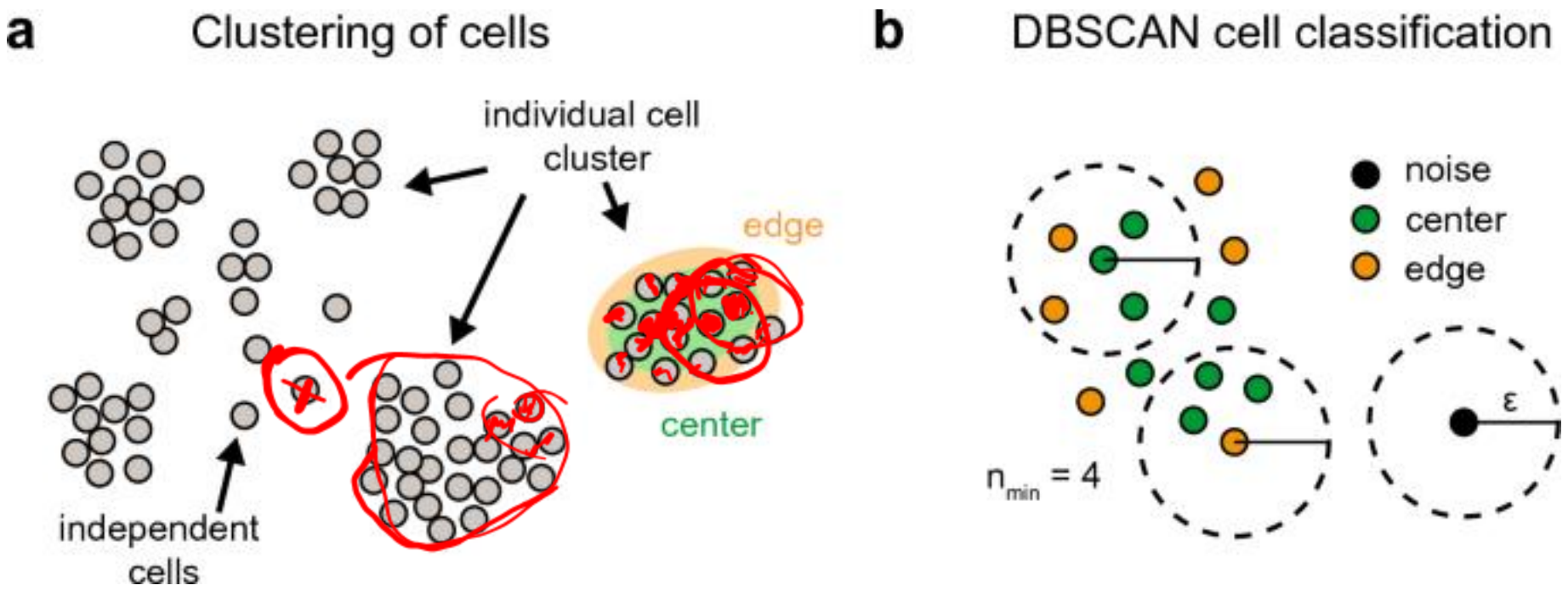
Типы точек:

1. **Основные** – в окрестности ε находится хотя бы N точек
2. **Граничные** – в окрестности ε кол-во **основных** точек $<$ N, но хотя бы 1
3. **Шумовые** – в окрестности ε нет **основных** точек и точек $< N$

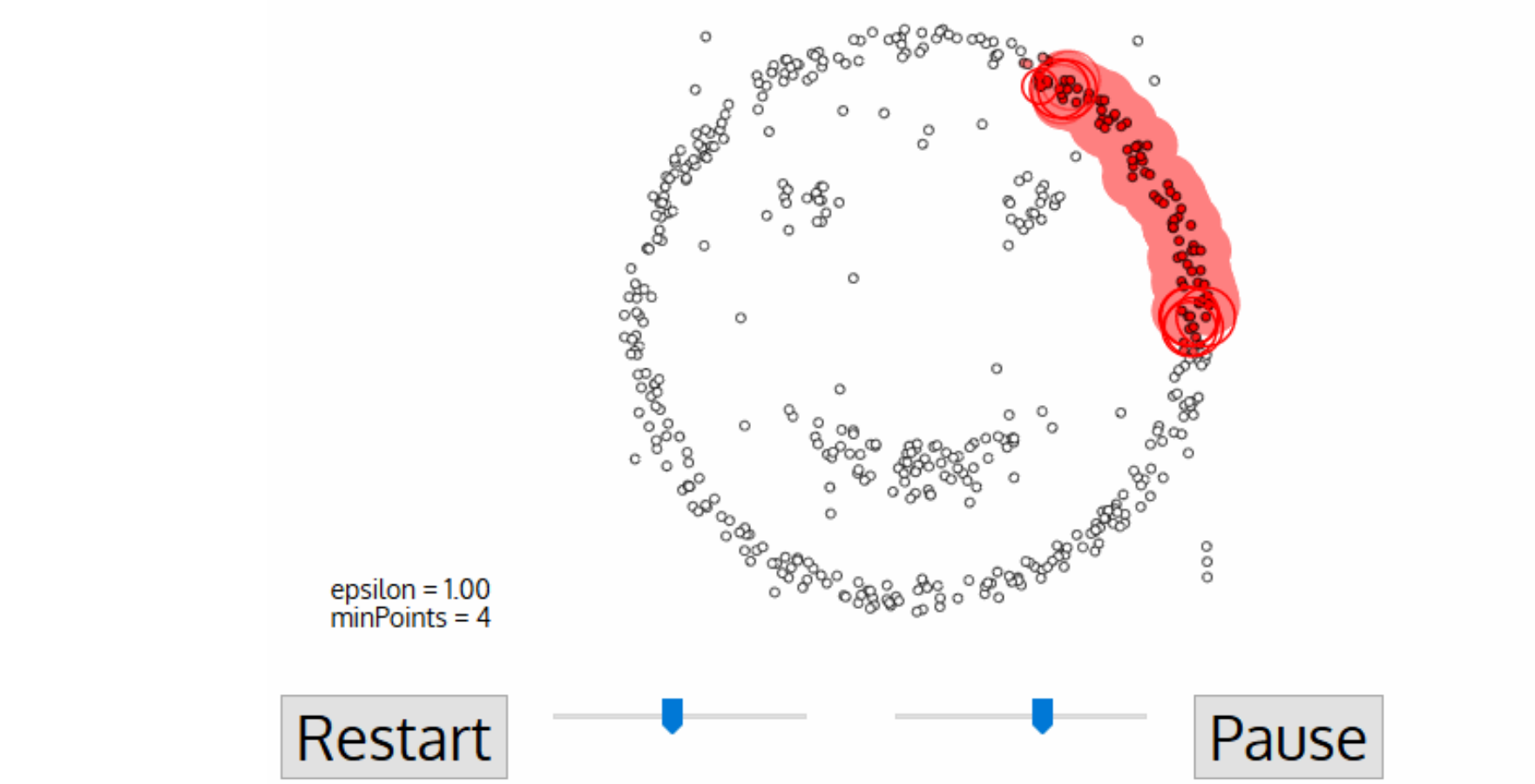
ε и N – гиперпараметры



DBSCAN. Пример



DBSCAN в действии

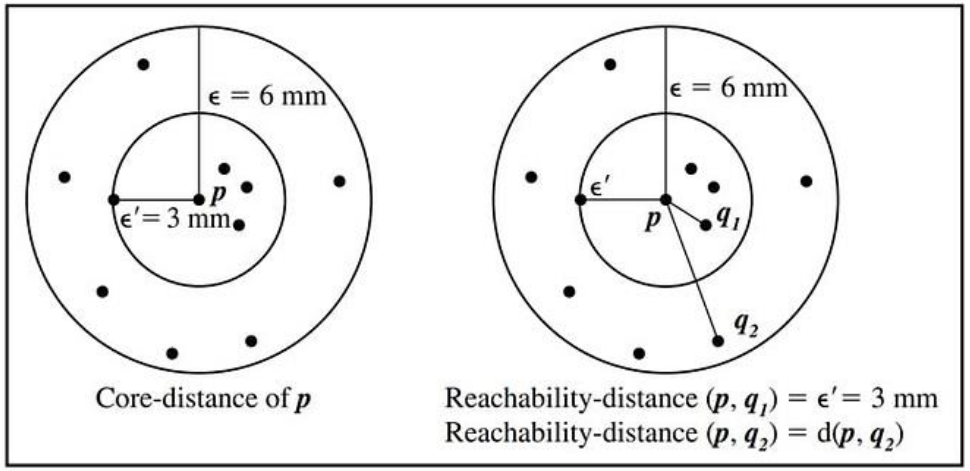


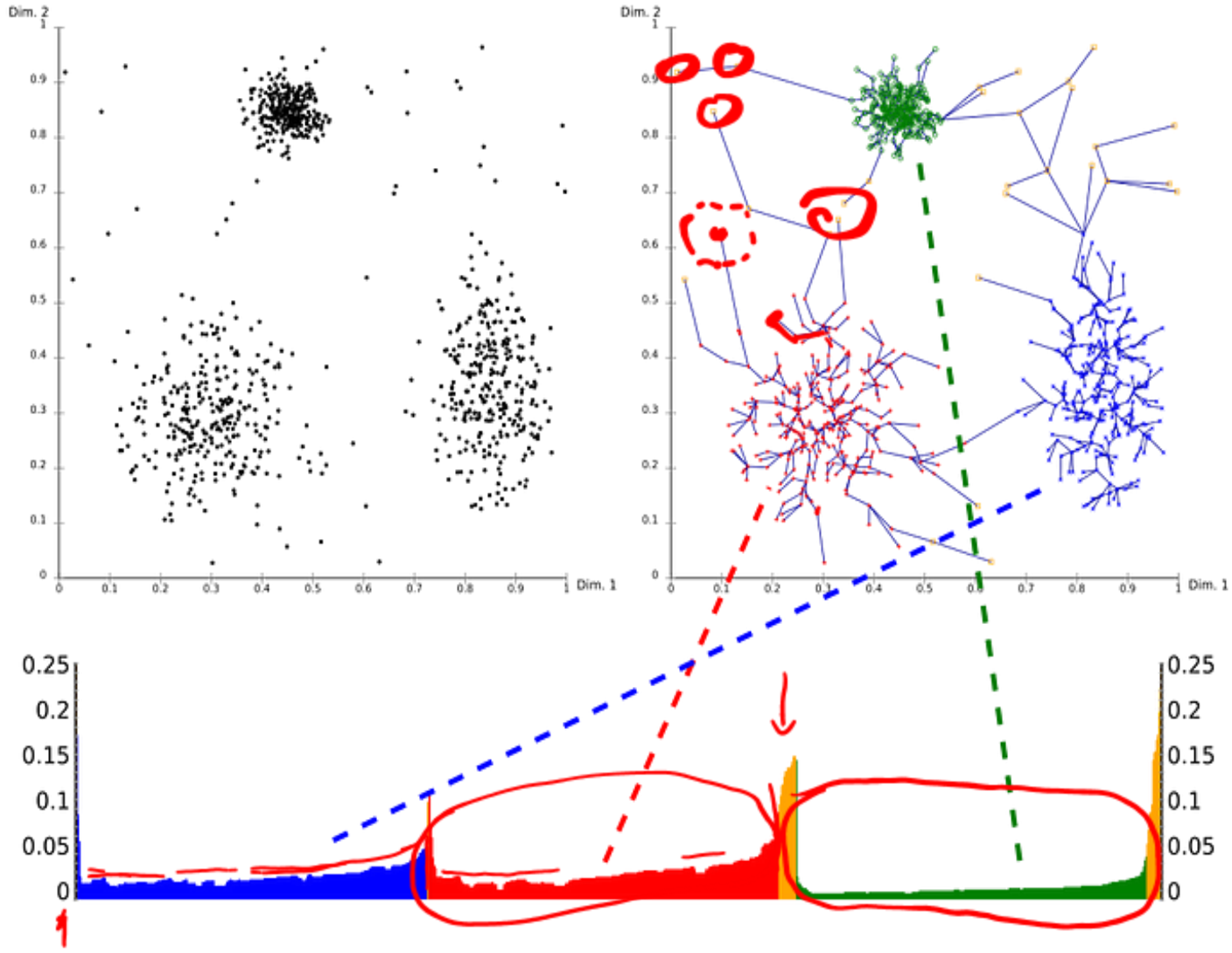
https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

OPTICS

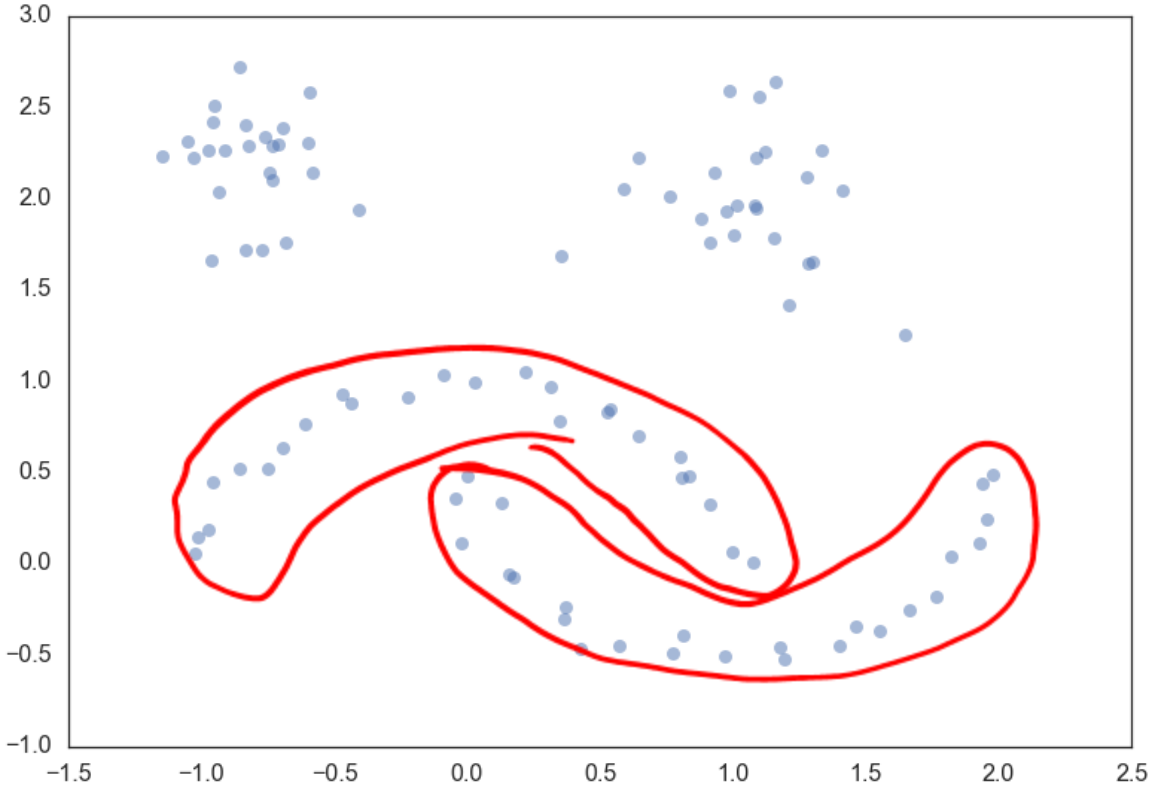
Изменения:

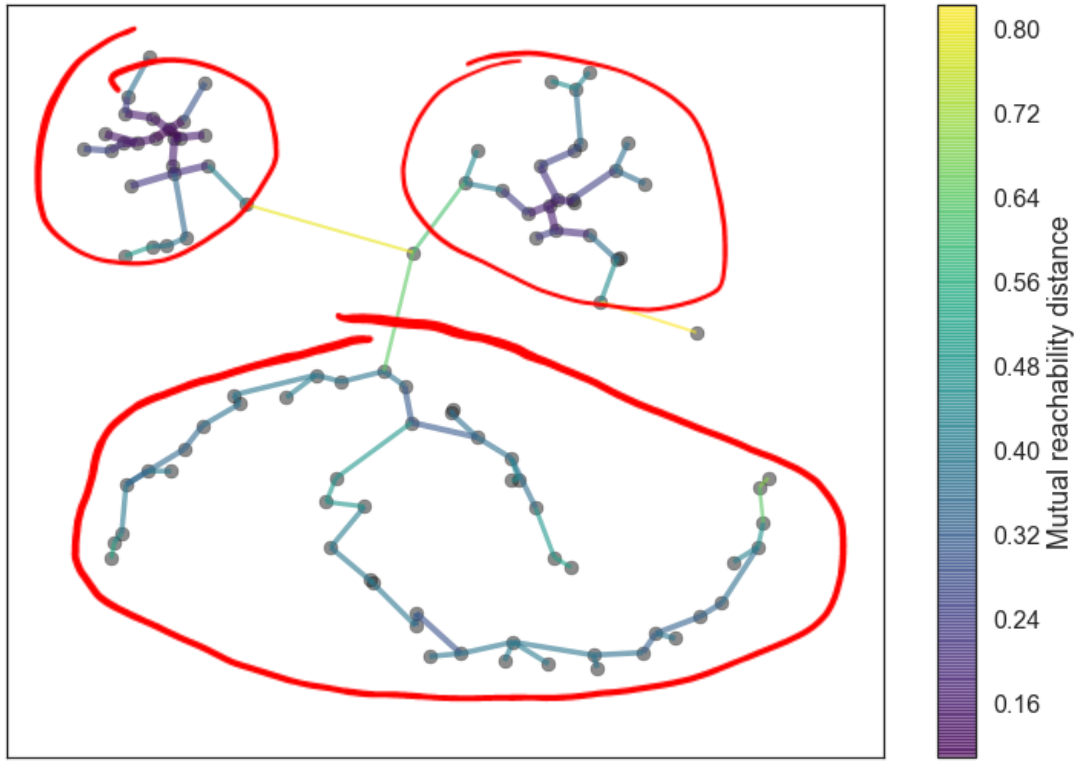
1. График доступности (или дендрограмма) – учитывает иерархию
2. Расстояние Core – ε'
3. Расстояние доступности – ε



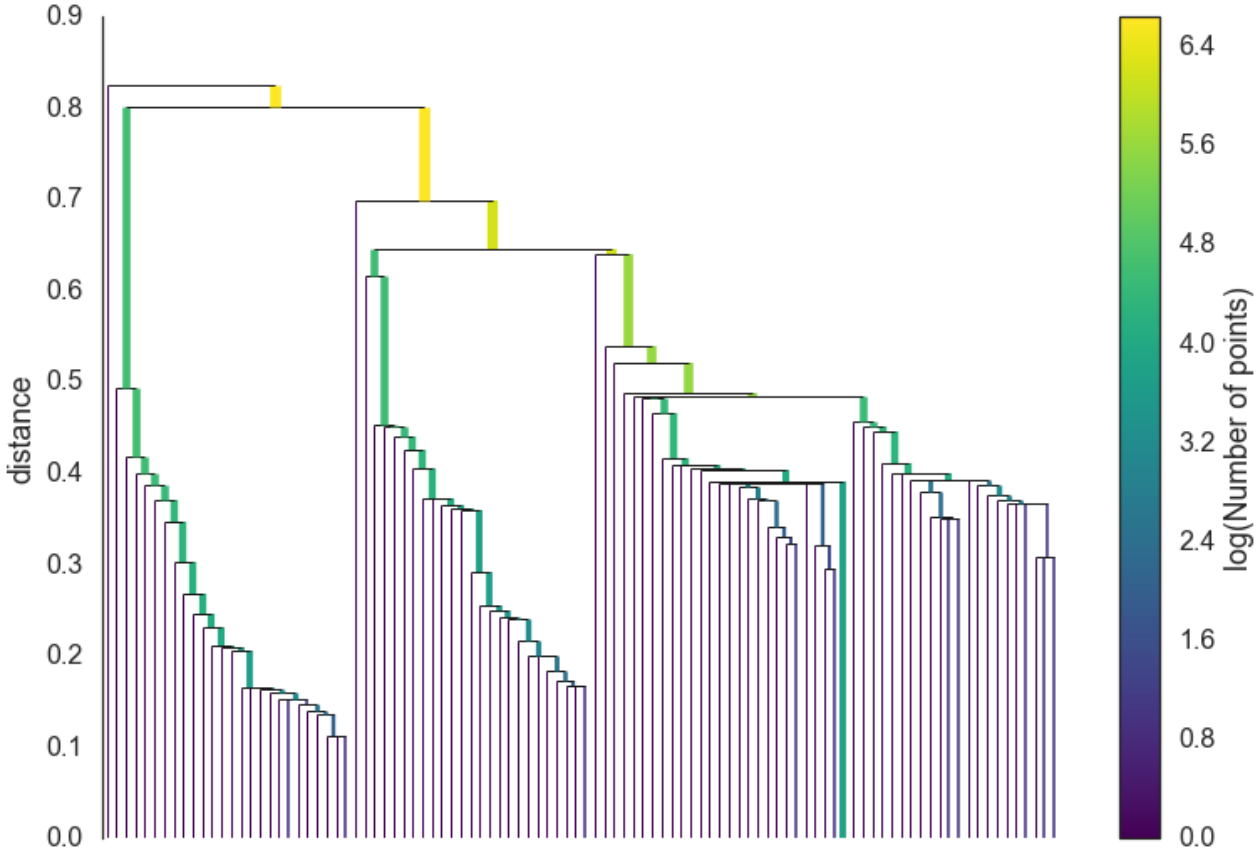


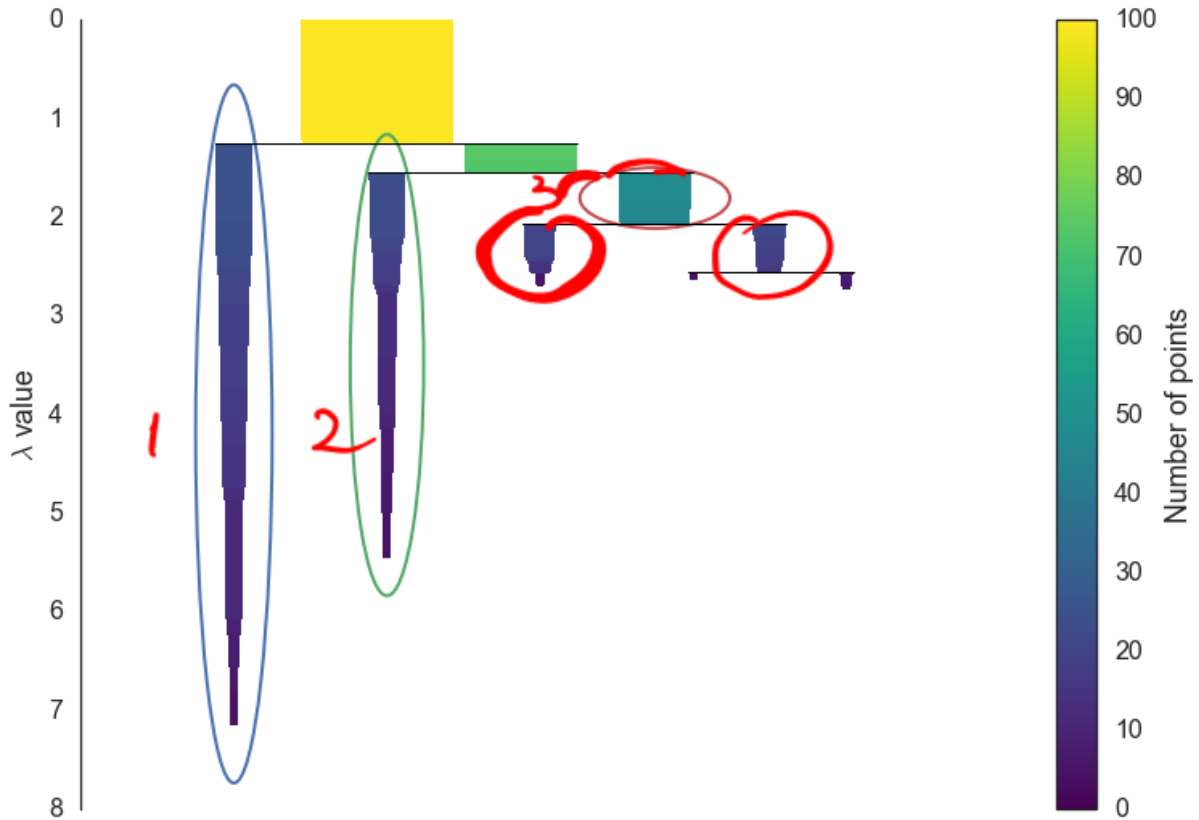
HDBSCAN

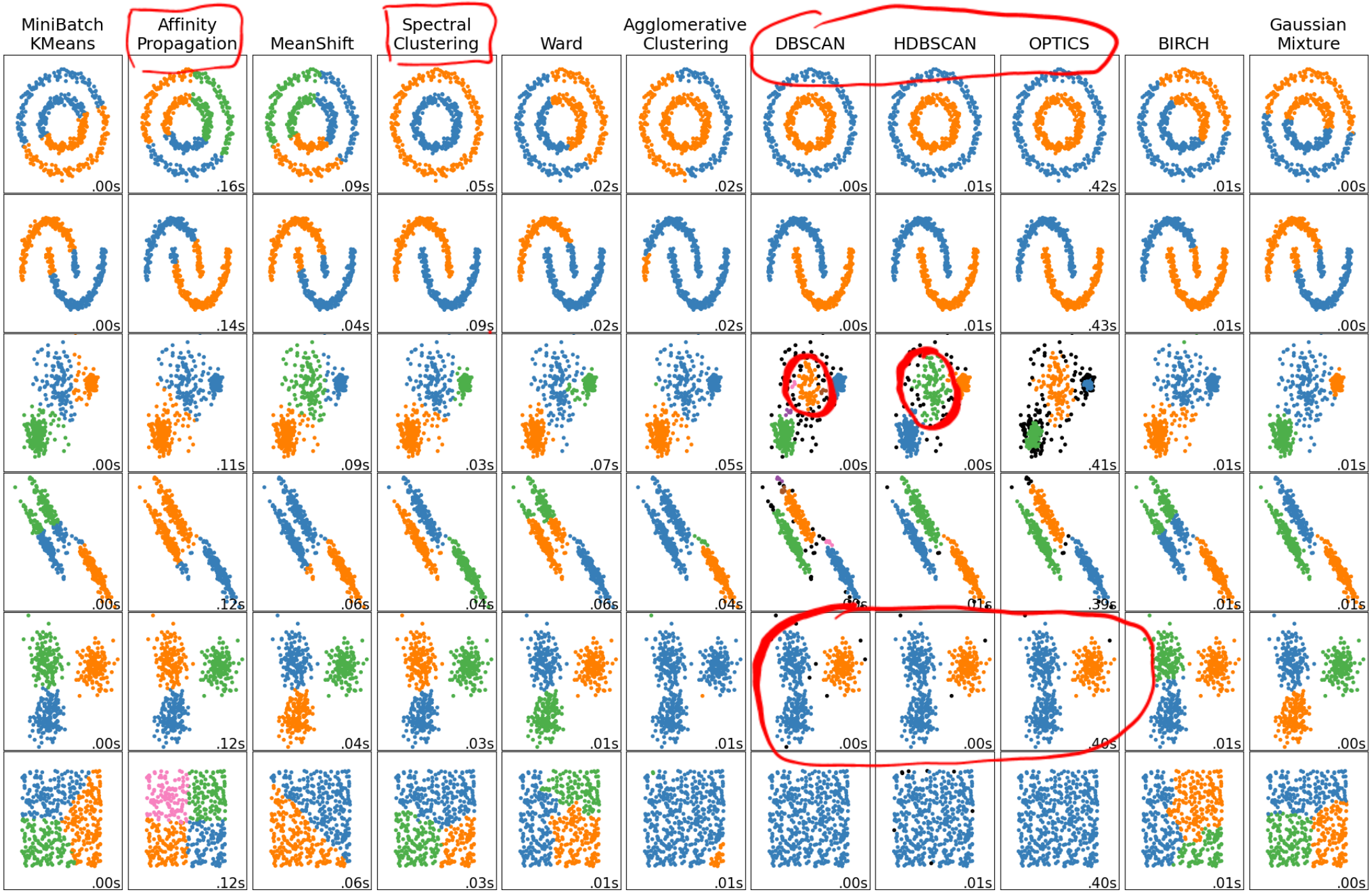

3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


0.80
0.72
0.64
0.56
0.48
0.40
0.32
0.24
0.16
Mutual reachability distance

HDBSCAN


distance
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
log(Number of points)
6.4
5.6
4.8
4.0
3.2
2.4
1.6
0.8
0.0


λ value
0
1
2
3
4
5
6
7
8
1
2
3
Number of points
100
90
80
70
60
50
40
30
20
10
0


MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
HDBSCAN
OPTICS
BIRCH
Gaussian Mixture
.00s .16s .09s .05s .02s .02s .00s .01s .42s .01s .00s
.00s .14s .04s .09s .02s .02s .00s .01s .43s .01s .00s
.00s .11s .09s .03s .07s .05s .00s .00s .41s .01s .01s
.00s .12s .06s .04s .06s .04s .01s .39s .01s .01s
.00s .12s .04s .03s .01s .01s .00s .00s .40s .01s .00s
.00s .12s .06s .03s .01s .01s .00s .01s .40s .01s .00s

Векторная квантизация

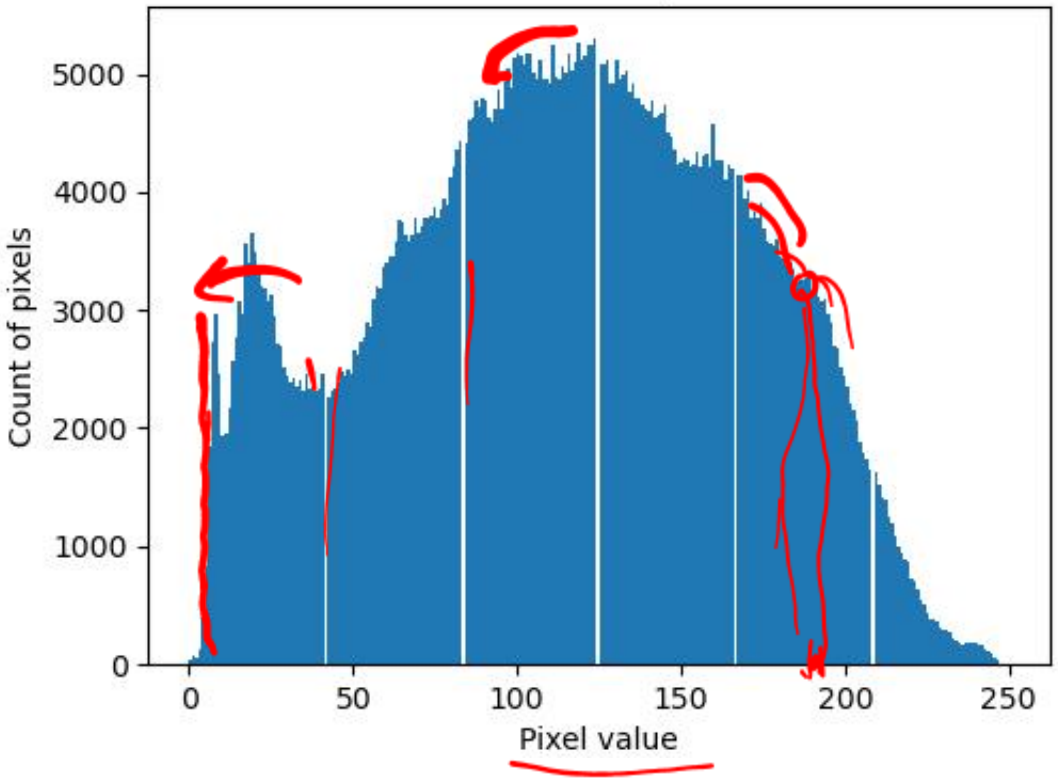
256 8 бит

3 бит — 8 ур. ур.

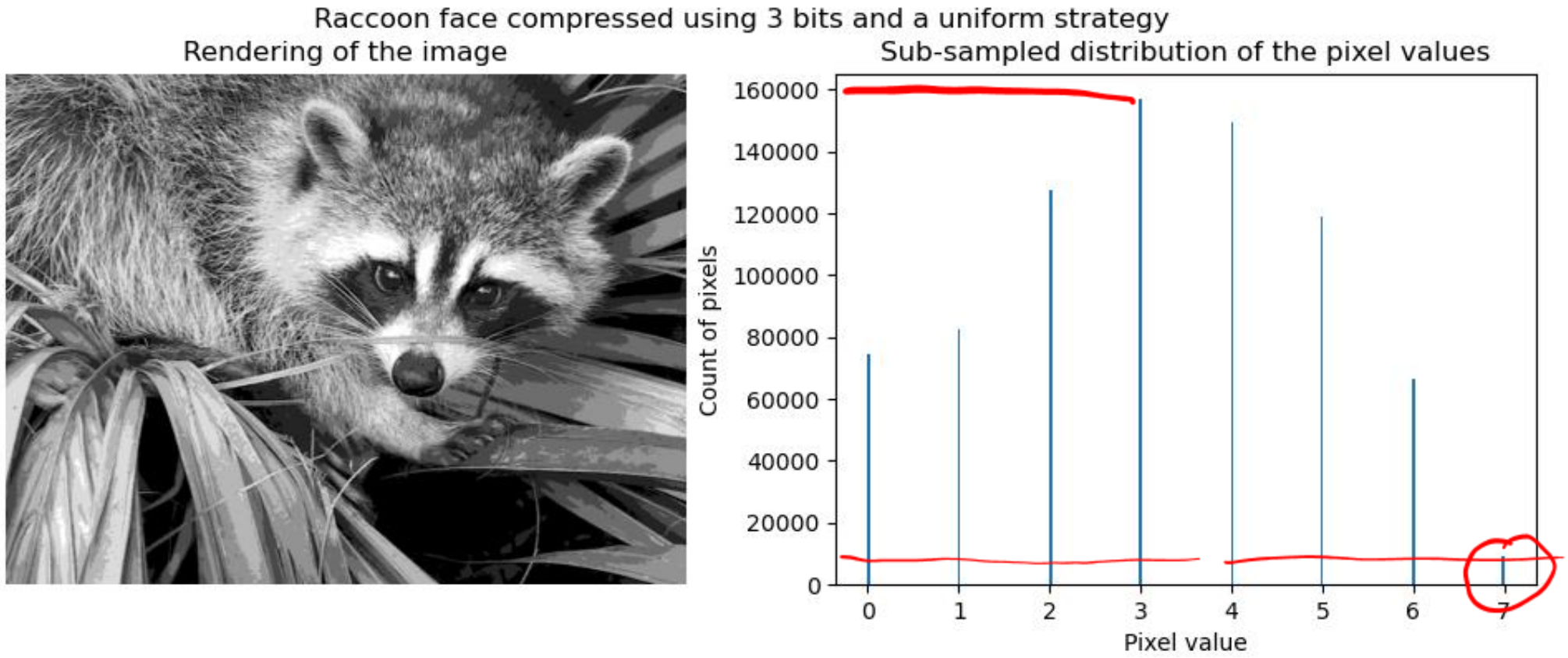
Original image of a raccoon face

Rendering of the image

Distribution of the pixel values



Векторная квантизация



Векторная квантизация