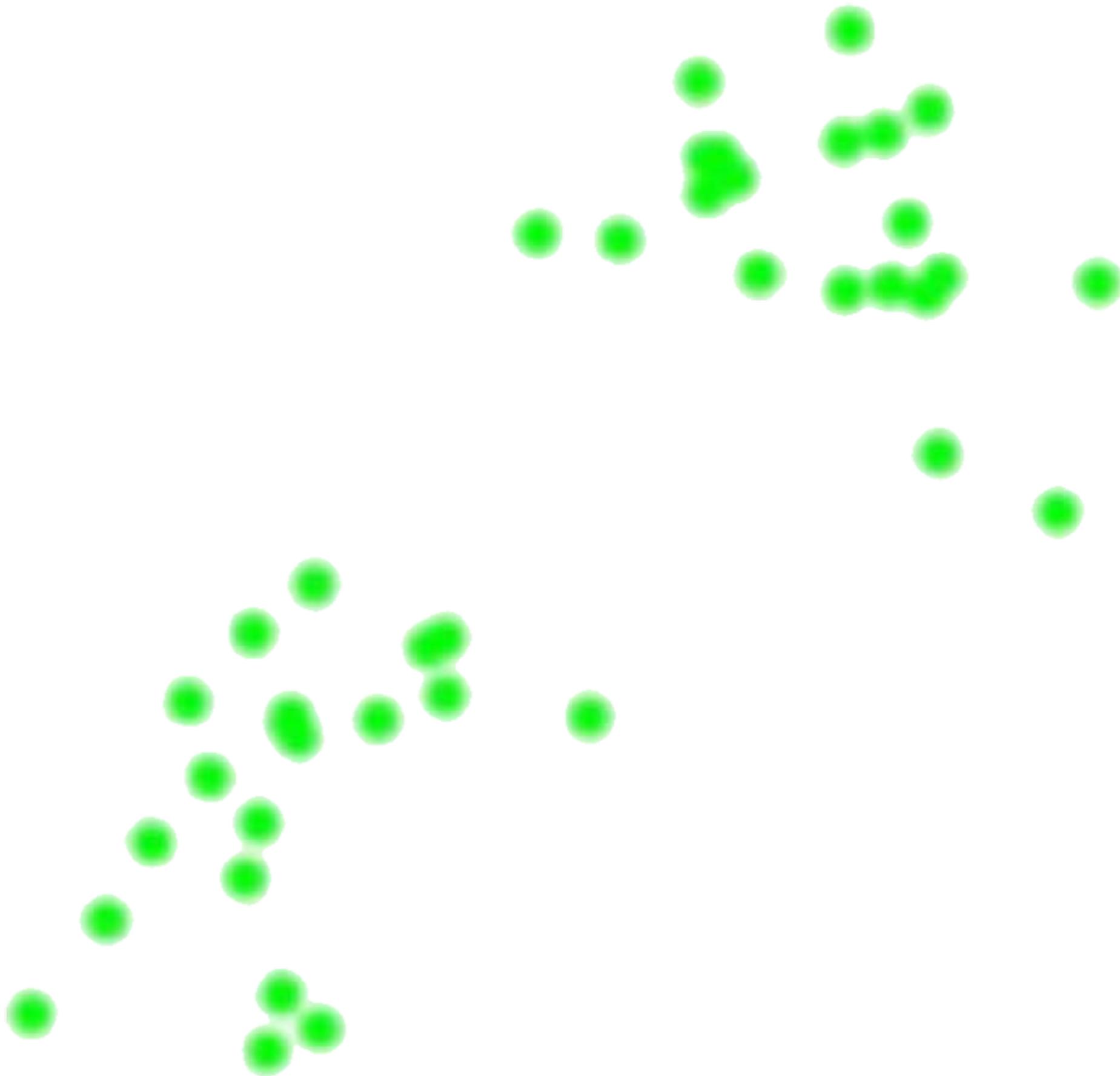


МЕТОД К СРЕДНИХ (K MEANS)

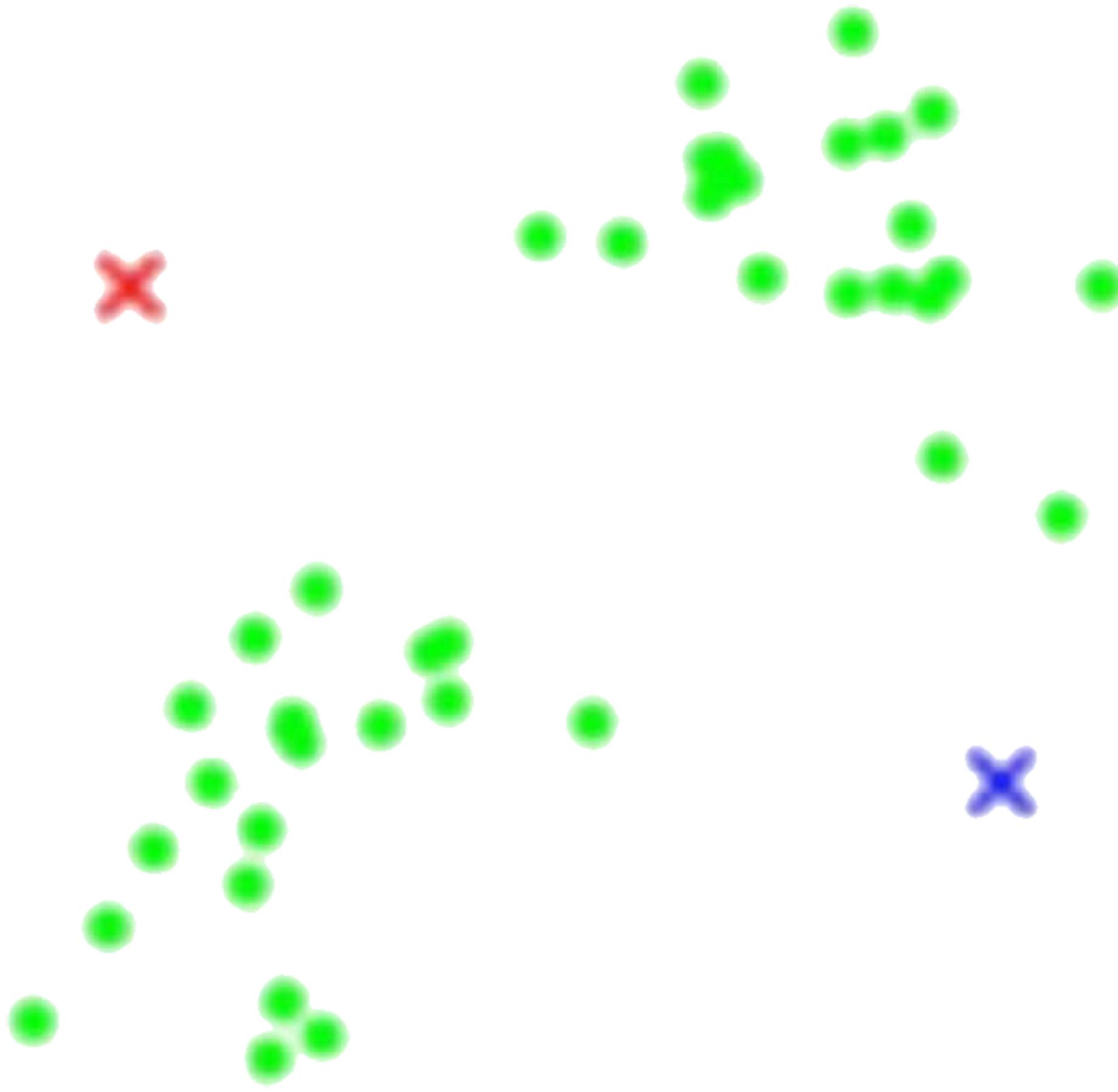
ПЛАН

1. Как работает K Means
2. Вариации K Means
3. Что делать, когда данных много: Mini Batch K-Means
4. Что делать, когда много признаков
5. Выбор начальных приближений: Kmeans++
6. Пример: уменьшение количества цветов в изображении
7. Работа K means с разными формами кластеров
8. Пример: мешок визуальных слов (bag of visual words)
9. Что оптимизирует K means

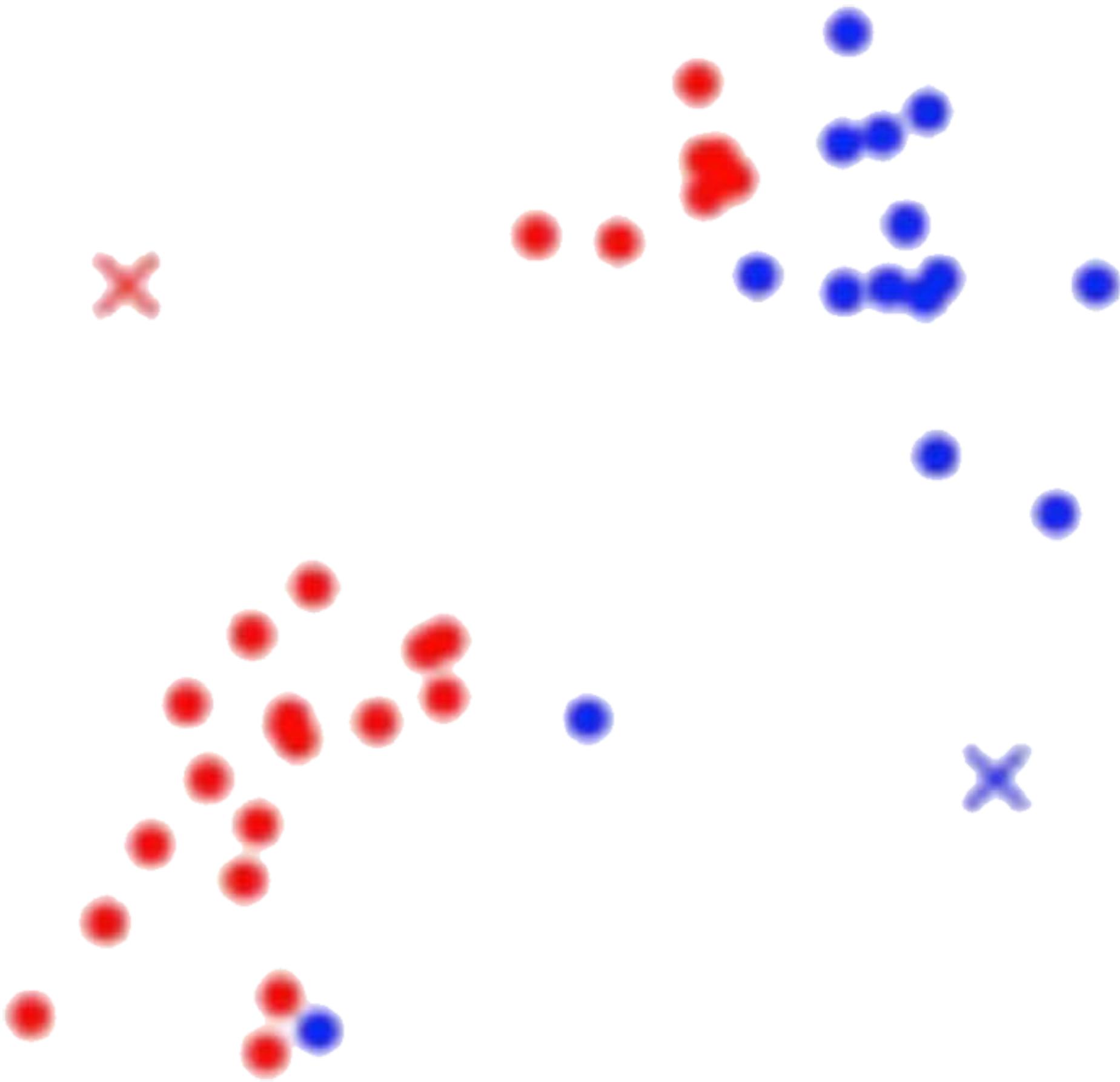
КАК РАБОТАЕТ K MEANS



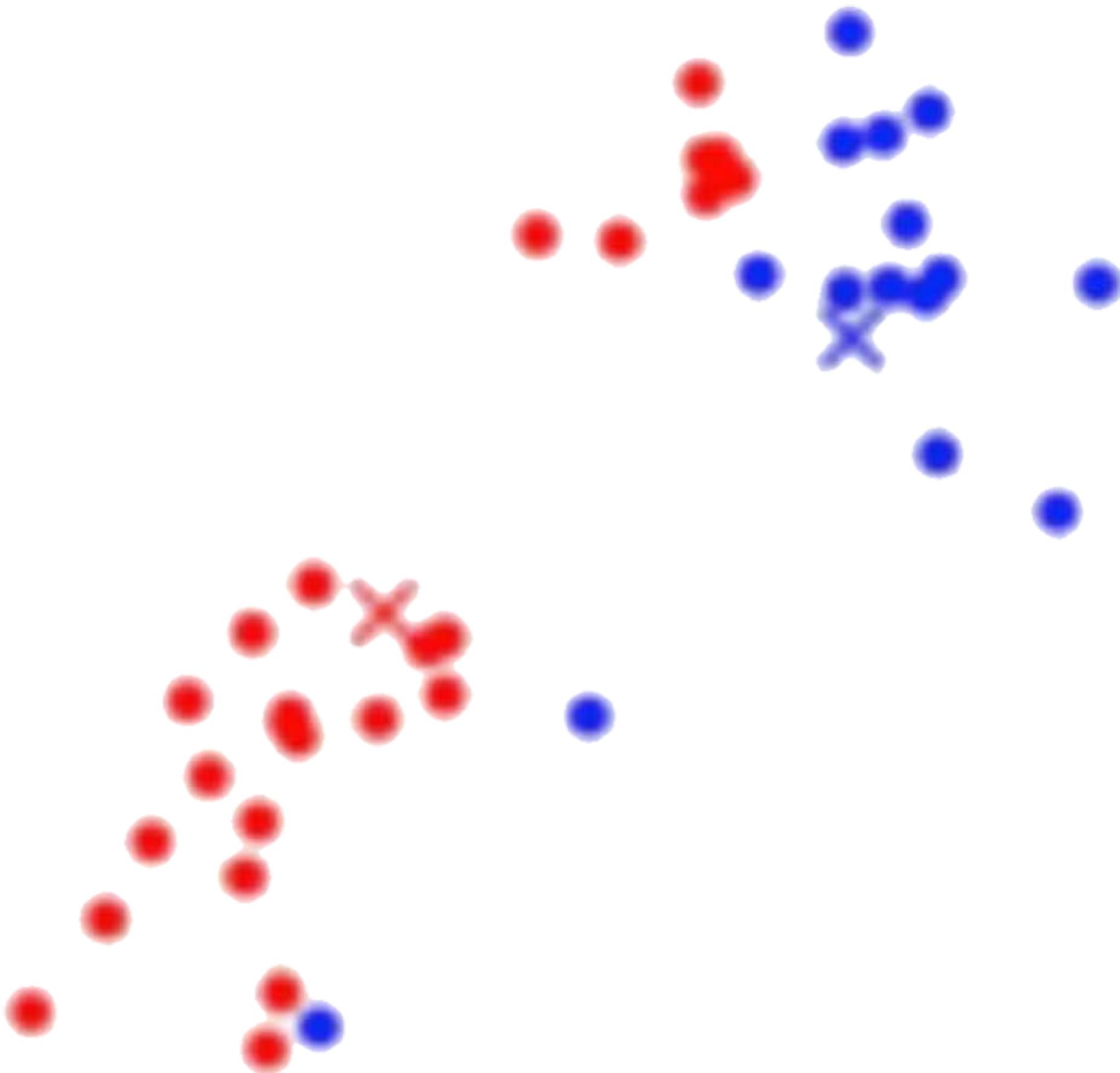
КАК РАБОТАЕТ K MEANS



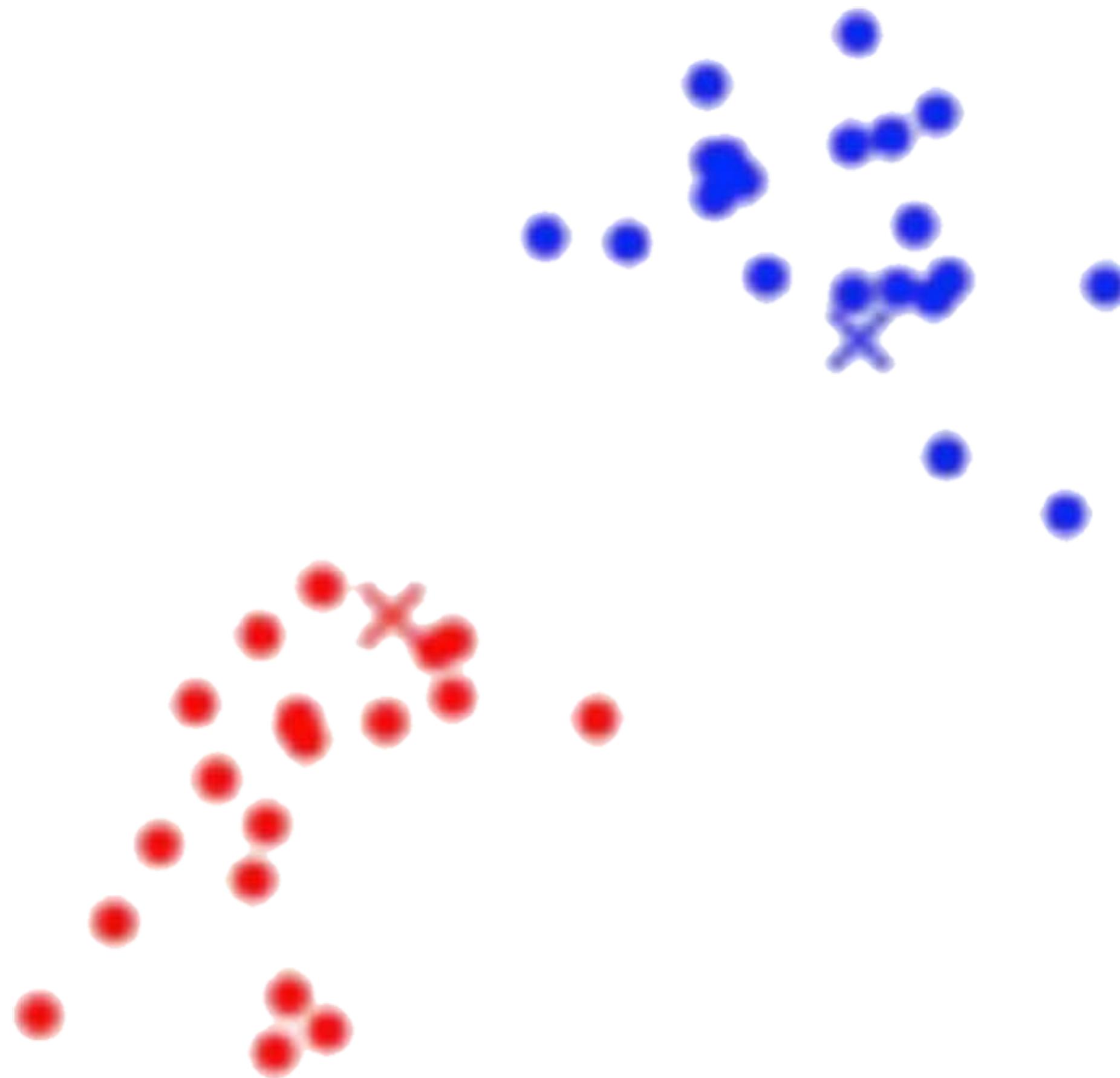
КАК РАБОТАЕТ K MEANS



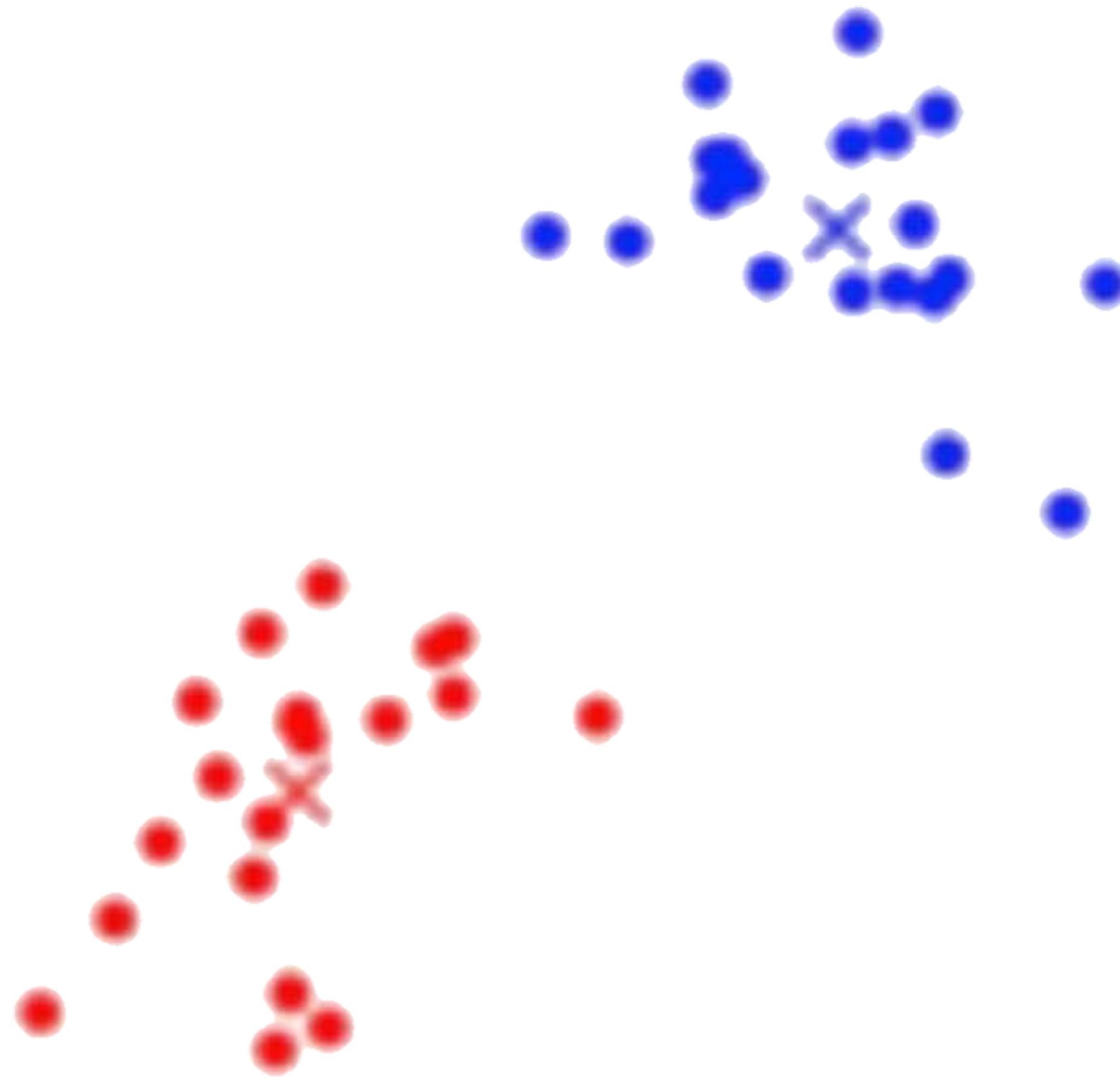
КАК РАБОТАЕТ K MEANS



КАК РАБОТАЕТ K MEANS



КАК РАБОТАЕТ K MEANS



ВАРИАЦИИ К MEANS

- В версии Болла Холла: уже рассказанный метод
- В версии Мак Кина: каждый раз, когда объект переходит из одного кластера в другой — центры кластеров пересчитываются

MINI-BATCH K MEANS

- Если данных много, относить объекты к кластерам и вычислять центры — достаточно долго
- Выход — на каждом шаге K Means работать со случайной подвыборкой из всех объектов
- В среднем все должно сходиться к тому же результату

ПОНИЖЕНИЕ РАЗМЕРНОСТИ ПРОСТРАНСТВА

- Каждое вычисление расстояния обычно требует $O(d)$ элементарных операций, где d — размерность пространства признаков
- Если признаков очень много, K Means начинает работать долго
- Решение — уменьшить число признаков
- Варианты: отбор признаков, метод главных компонент (PCA), сингулярное разложение (SVD) — об этом — далее в курсе

K MEANS++

- › В зависимости от начального приближения центров кластеров может потребоваться разное время для сходимости
- › Можно брать центры подальше друг от друга – для двух кластеров понятно, что это значит, а для K ?
- › Вариант выбора начальных приближений:
 - ▶ Первый центр выбираем случайно из равномерного распределения на выборке
 - ▶ Каждый следующий центр выбираем случайно из оставшихся точек так, чтобы вероятность выбрать каждую точку была пропорциональна квадрату расстояния от нее до ближайшего центра

ПРИМЕР: КВАНТИЗАЦИЯ ИЗОБРАЖЕНИЙ

Original image (96,615 colors)



ПРИМЕР: КВАНТИЗАЦИЯ ИЗОБРАЖЕНИЙ

Quantized image (64 colors, Random)

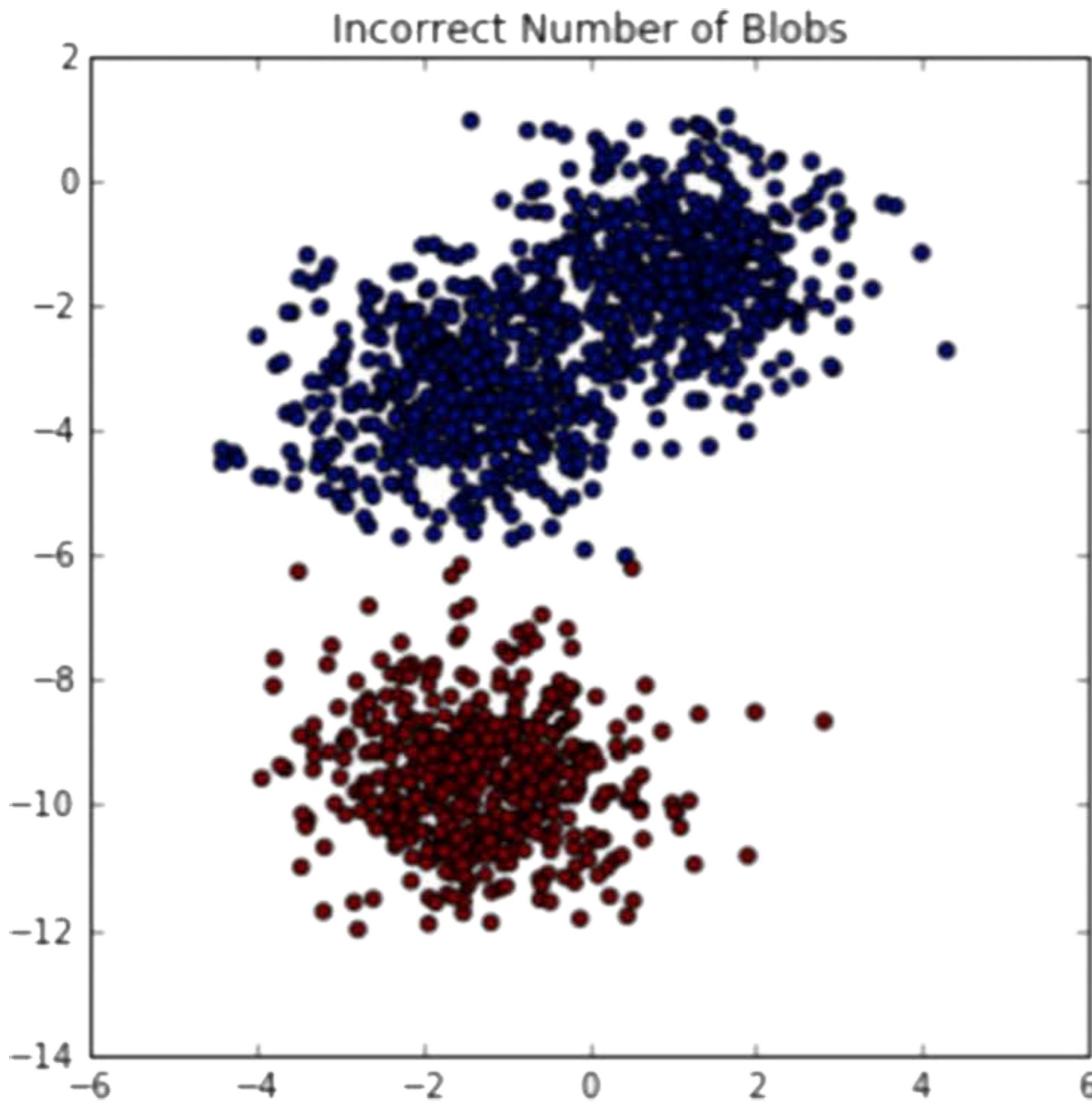


ПРИМЕР: КВАНТИЗАЦИЯ ИЗОБРАЖЕНИЙ

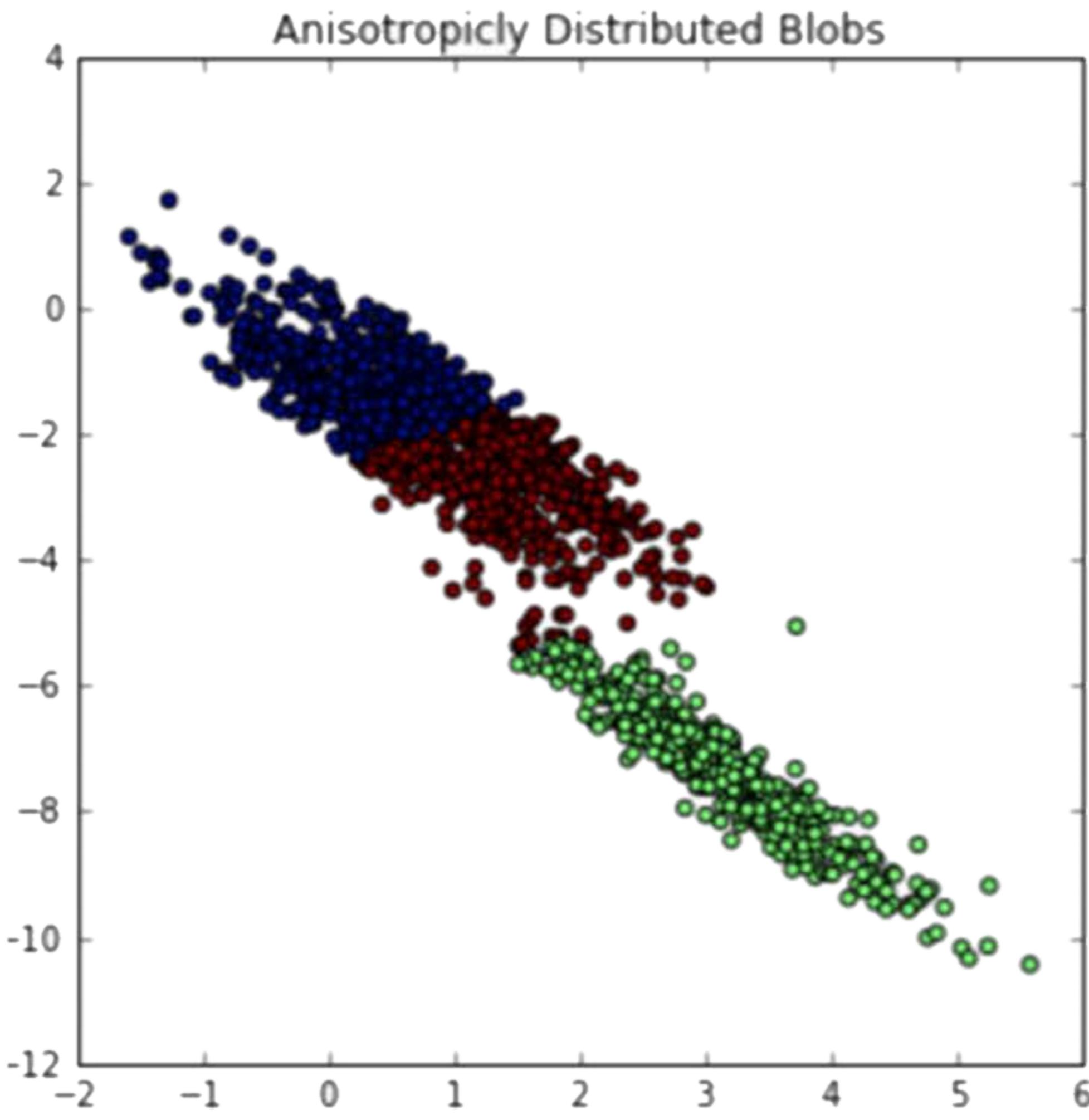
Quantized image (64 colors, K-Means)



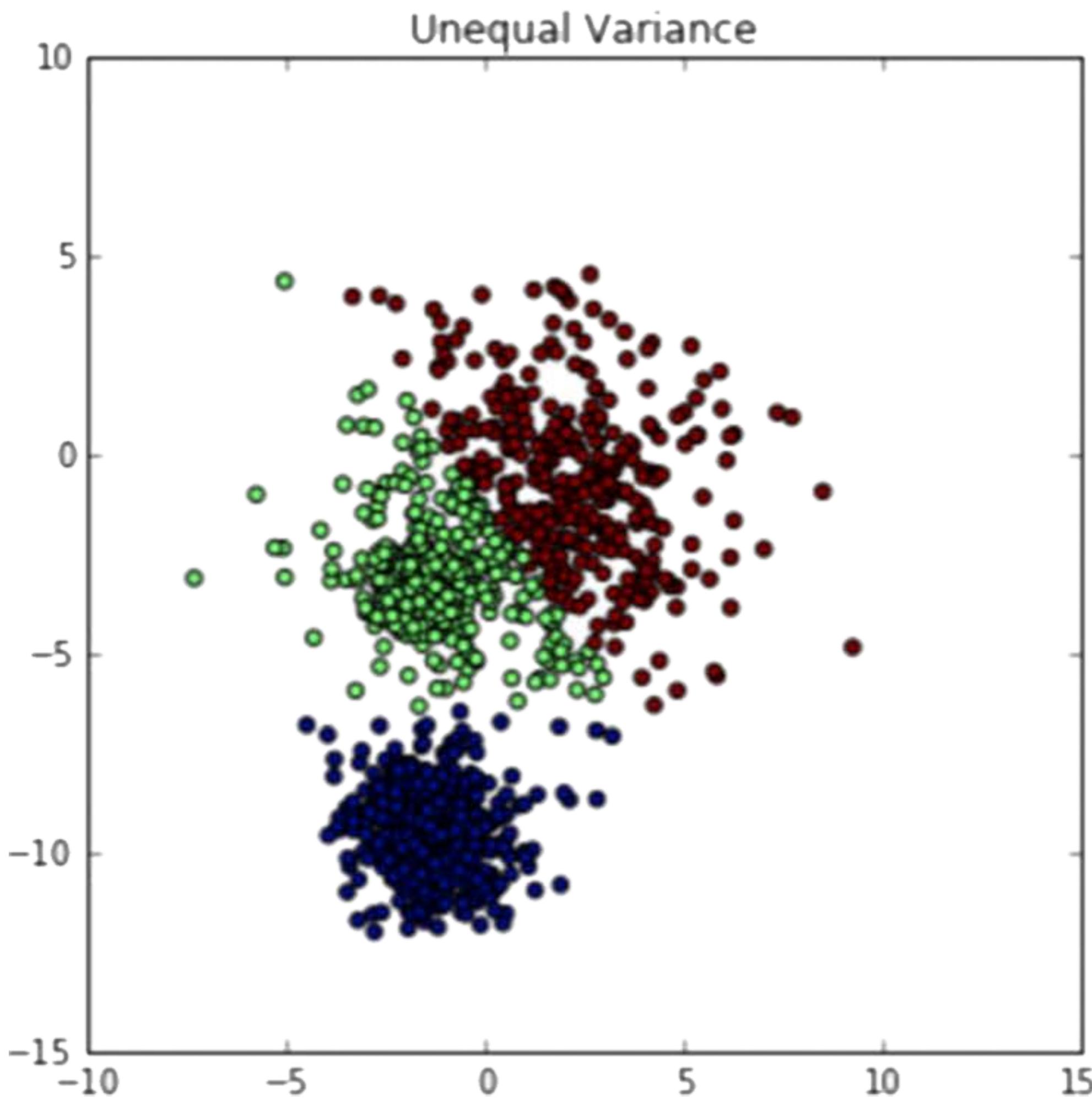
K-MEANS И РАЗНЫЕ ФОРМЫ КЛАСТЕРОВ



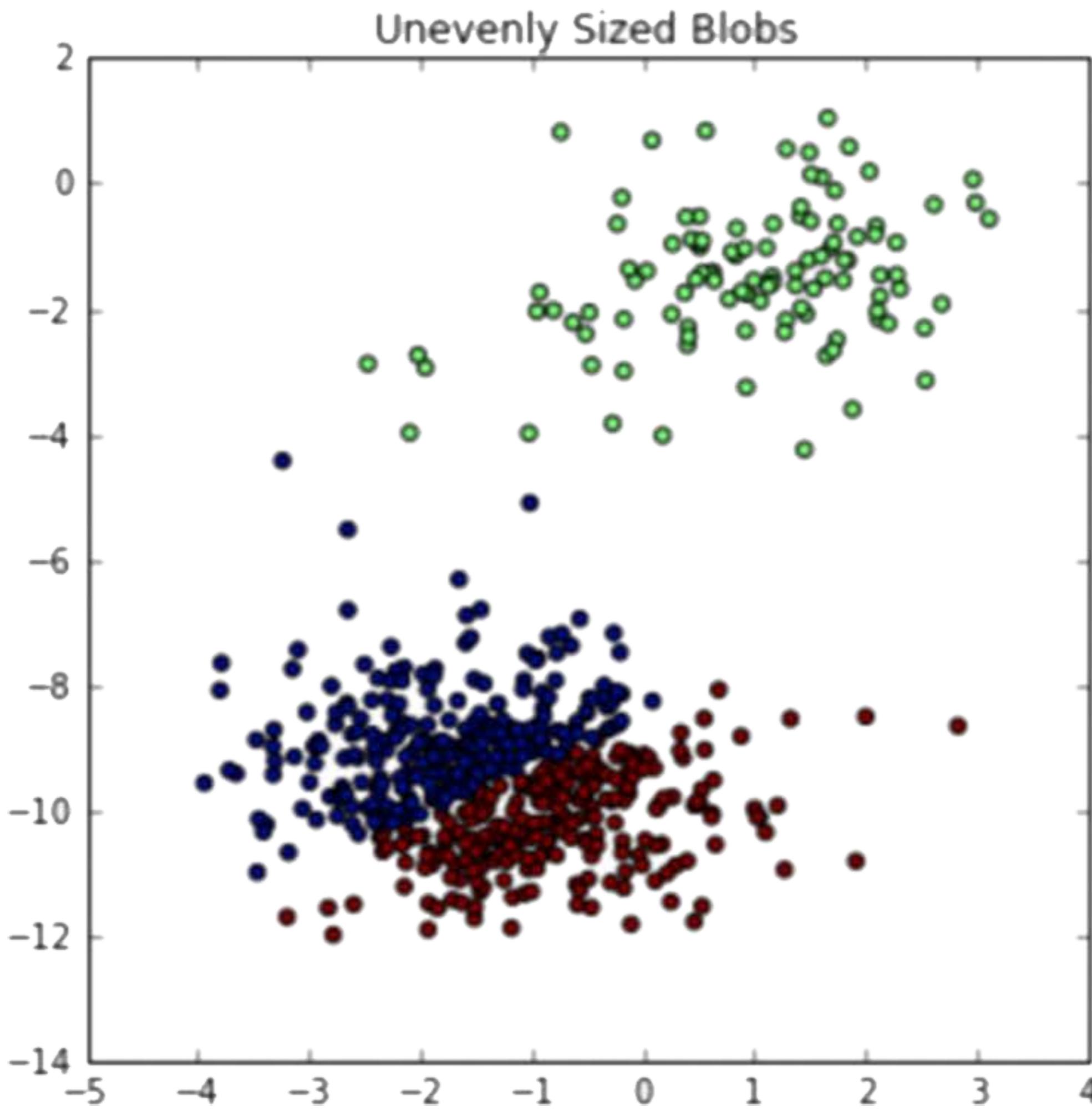
K MEANS И РАЗНЫЕ ФОРМЫ КЛАСТЕРОВ



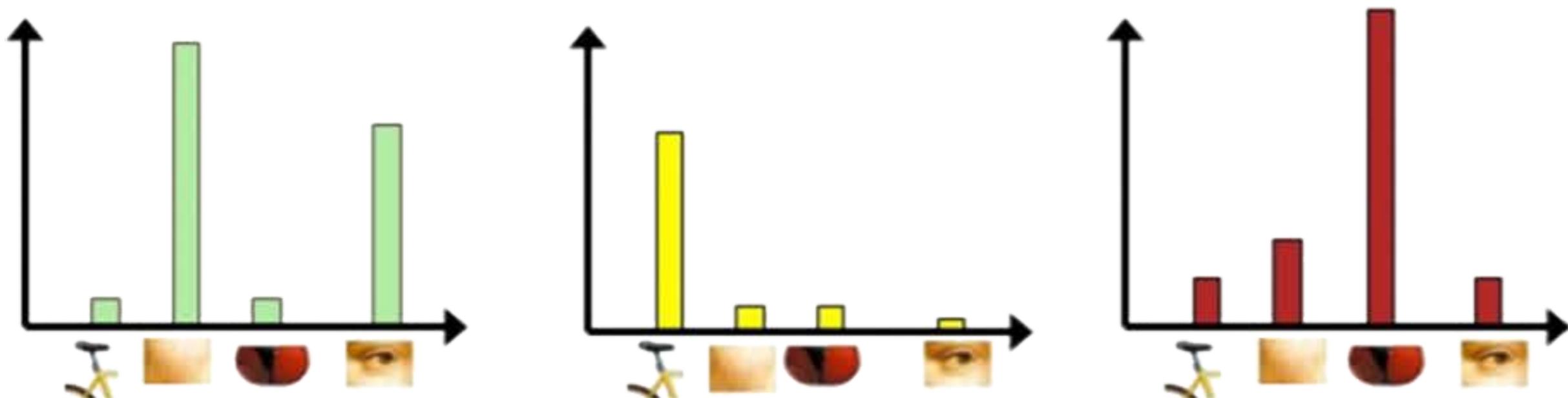
K MEANS И РАЗНЫЕ ФОРМЫ КЛАСТЕРОВ



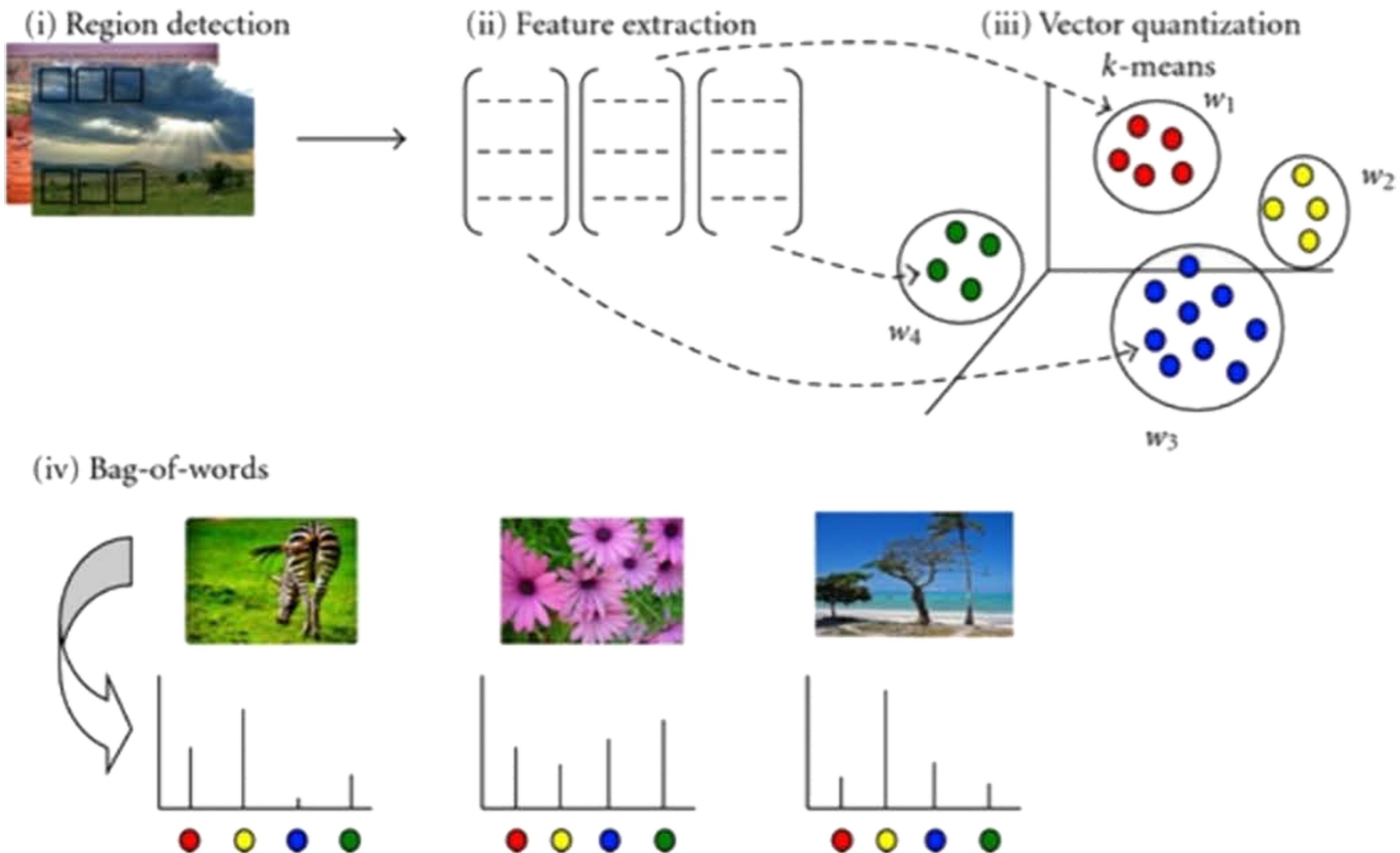
K-MEANS И РАЗНЫЕ ФОРМЫ КЛАСТЕРОВ



ПРИМЕР: МЕШОК ВИЗУАЛЬНЫХ СЛОВ



ПРИМЕР: МЕШОК ВИЗУАЛЬНЫХ СЛОВ



ЧТО ОПТИМИЗИРУЕТ K MEANS

- Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i < j} [y_i = y_j] \rho(x_i, x_j)}{\sum_{i < j} [y_i = y_j]} \rightarrow \min$$

ЧТО ОПТИМИЗИРУЕТ K MEANS

- Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i < j} [y_i = y_j] \rho(x_i, x_j)}{\sum_{i < j} [y_i = y_j]} \rightarrow \min$$

- Альтернативный вариант, если есть центры кластеров:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i: y_i = y} \rho^2(x_i, \mu_y) \rightarrow \min$$

ЧТО ОПТИМИЗИРУЕТ K MEANS

- В 1967 году Мак Кин показал, что для его версии K Means:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i: y_i = y} \rho^2(x_i, \mu_y) \rightarrow \min$$

ЧТО ОПТИМИЗИРУЕТ K MEANS

- K Means итеративно минимизирует среднее внутрикластерное расстояние:
 1. Объект присваивается к тому кластеру, центр которого ближе
 2. Центр кластера перемещается в среднее арифметическое векторов признаков объектов из него

ЧТО ОПТИМИЗИРУЕТ K MEANS

- K Means итеративно минимизирует среднее внутрикластерное расстояние:
 1. Объект присваивается к тому кластеру, центр которого ближе
 2. Центр кластера перемещается в среднее арифметическое векторов признаков объектов из него

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \operatorname{argmin}_{\mu} \frac{1}{N} \sum_{i=1}^N (\mu - x_i)^2$$

ЧТО ОПТИМИЗИРУЕТ K MEANS

› K Means итеративно минимизирует среднее внутрикластерное расстояние:

1. Объект присваивается к тому кластеру, центр которого ближе
2. Центр кластера перемещается в среднее арифметическое векторов признаков объектов из него

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \operatorname{argmin}_{\mu} \frac{1}{N} \sum_{i=1}^N (\mu - x_i)^2$$

$$\frac{d}{d\mu} \frac{1}{N} \sum_{i=1}^N (\mu - x_i)^2 = \frac{2}{N} \sum_{i=1}^N (\mu - x_i) = 0 \Rightarrow \mu = \frac{1}{N} \sum_{i=1}^N x_i$$

ИТОГИ

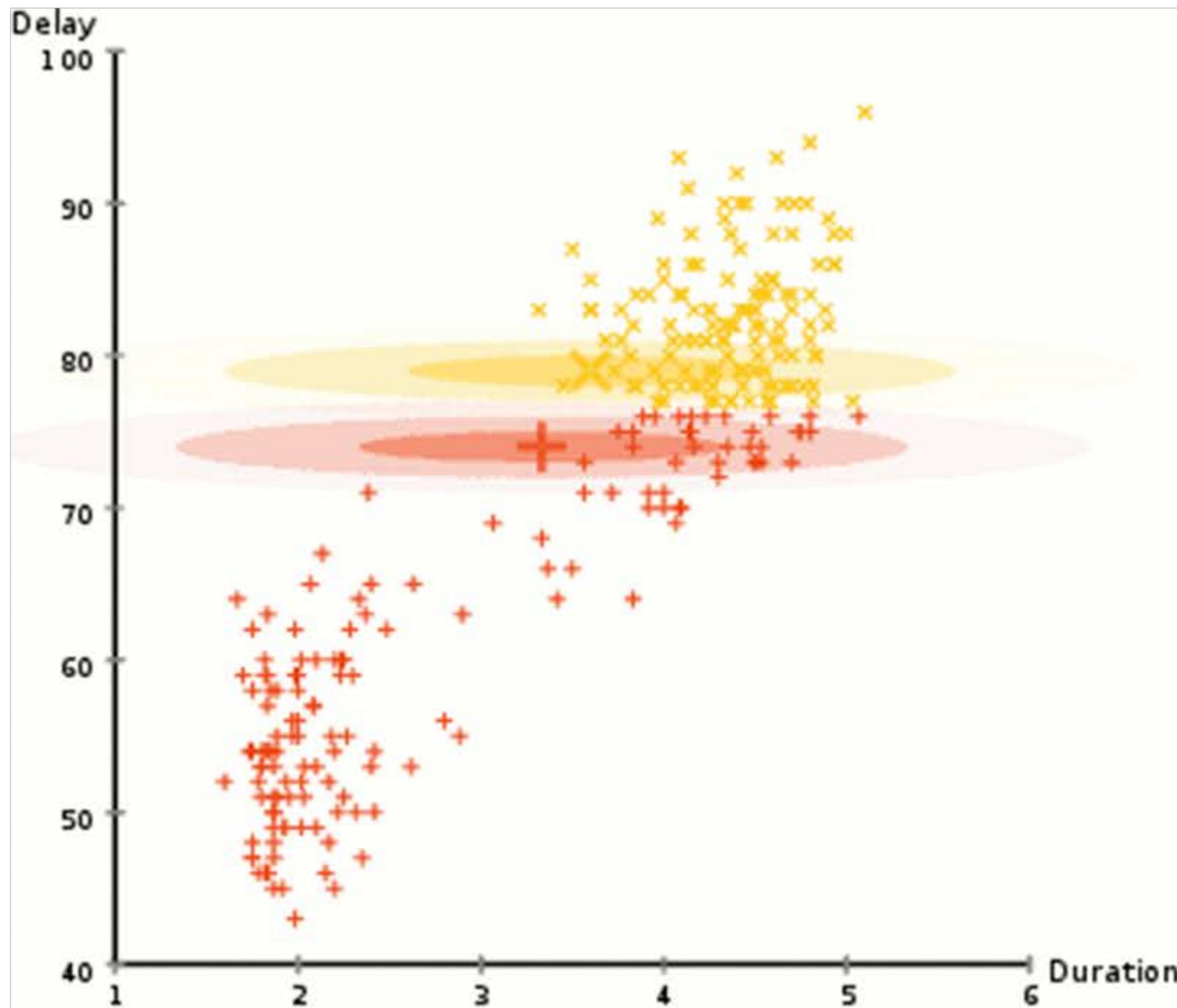
1. Как работает K Means
2. Вариации: K Means Болла-Холла и Мак Кина
3. Что делать, когда данных много: Mini Batch K-Means
4. Что делать, когда много признаков:
понижение размерности
5. Выбор начальных приближений: Kmeans++
6. Пример: квантизация изображений
7. Работа K means с разными формами
кластеров
8. Пример: мешок визуальных слов (bag of visual words)
9. Что оптимизирует K means

EXPECTATION MAXIMIZATION (ЕМ-АЛГОРИТМ)

ПЛАН

- › Как выглядит кластеризация с помощью ЕМ-алгоритма
- › Постановка задачи
- › Почему не решить «в лоб»
- › Описание ЕМ алгоритма
- › ЕМ-алгоритм в случае гауссовых распределений
- › Простое объяснение метода
- › Классическое объяснение метода
- › Для чего еще используют алгоритм

КАК ЭТО ВЫГЛЯДИТ



ПОСТАНОВКА ЗАДАЧИ

- Модель порождения данных:
 - ▶ Априорные вероятности кластеров —
 $\mathbf{w}_1, \dots, \mathbf{w}_K$
 - ▶ Плотности распределения кластеров —
 $p_1(x), \dots, p_K(x)$
 - ▶ Плотности распределения вектора признаков \mathbf{x} :

$$p(\mathbf{x}) = \sum_{j=1}^K \mathbf{w}_j p_j(\mathbf{x})$$

ПОСТАНОВКА ЗАДАЧИ

- Что будем делать:
По выборке оценим параметры модели:

$\mathbf{w}_1, \dots, \mathbf{w}_K$

$p_1(x), \dots, p_K(x)$

ПОСТАНОВКА ЗАДАЧИ

- Что будем делать:
По выборке оценим параметры модели:

$$\mathbf{w}_1, \dots, \mathbf{w}_K$$

$$p_1(x), \dots, p_K(x)$$

- Зачем:
Сможем оценивать вероятность
принадлежности к кластеру

ПОСТАНОВКА ЗАДАЧИ: РАЗДЕЛЕНИЕ СМЕСИ

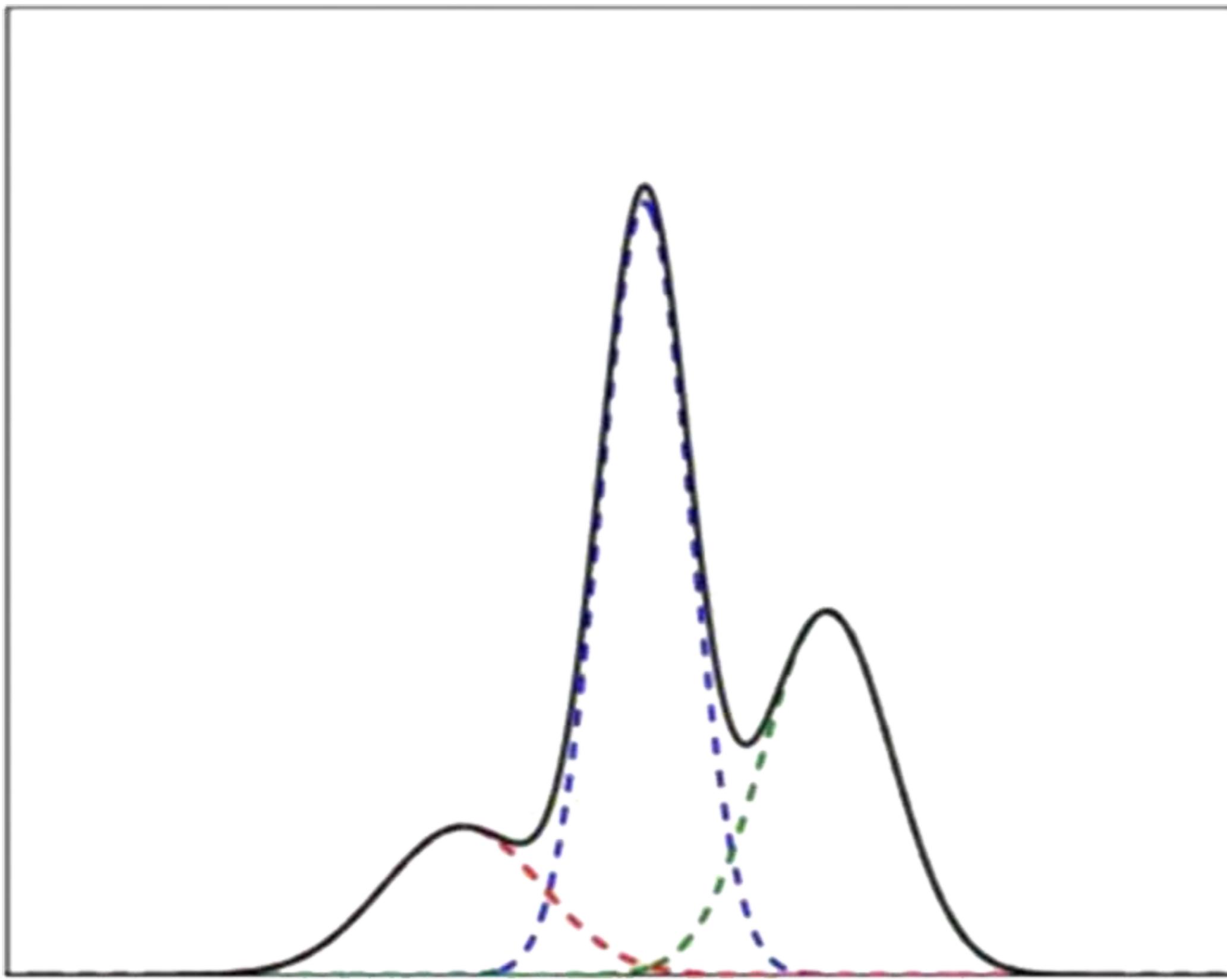
$$p(x) = \sum_{j=1}^K w_j p_j(x) \implies$$

\implies Оценить: w_1, \dots, w_K и $p_1(x), \dots, p_K(x)$

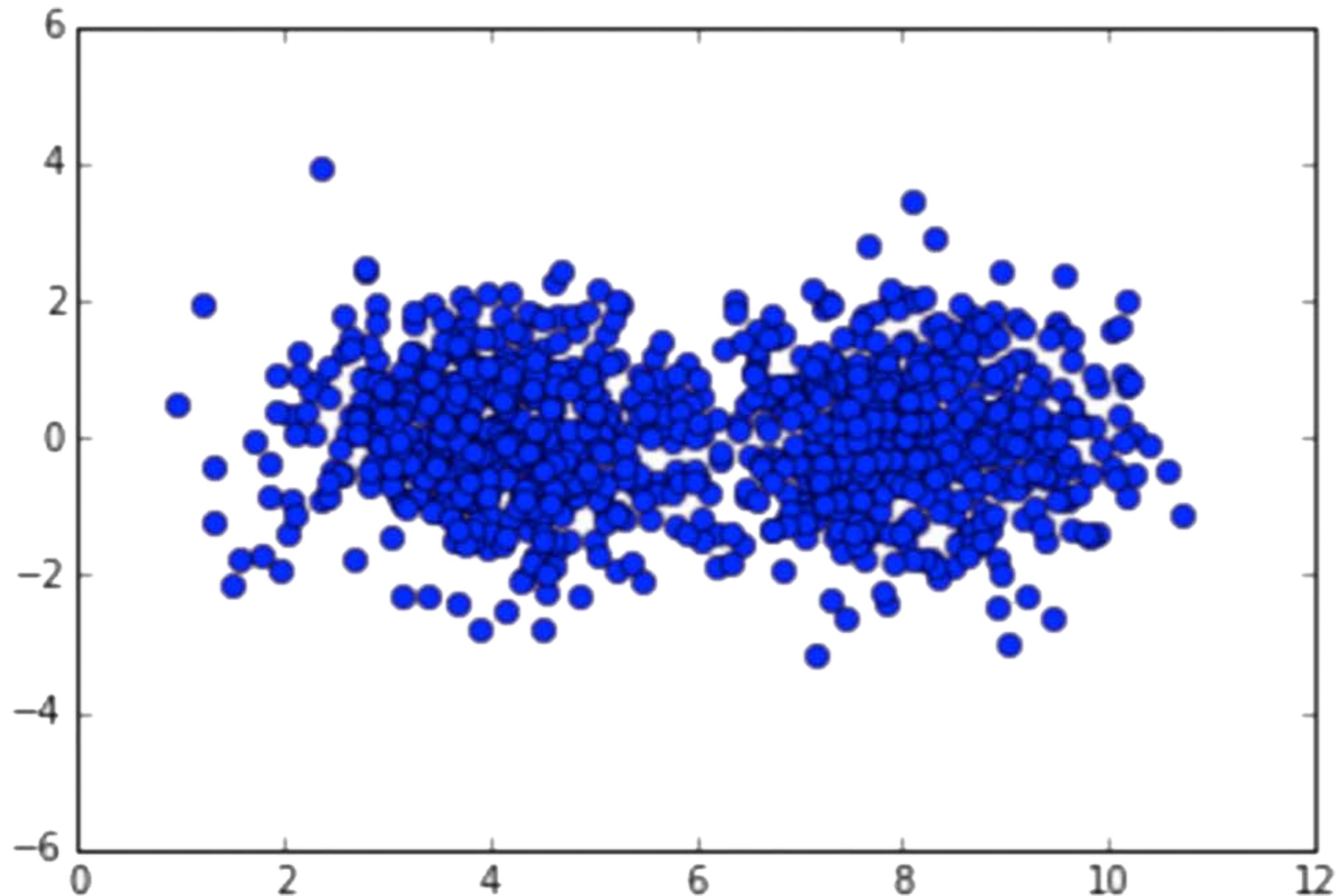
$$p_j(x) = \varphi(\theta_j; x)$$

Например, $p_j(x)$ — плотность нормального распределения (со своими параметрами для каждой компоненты)

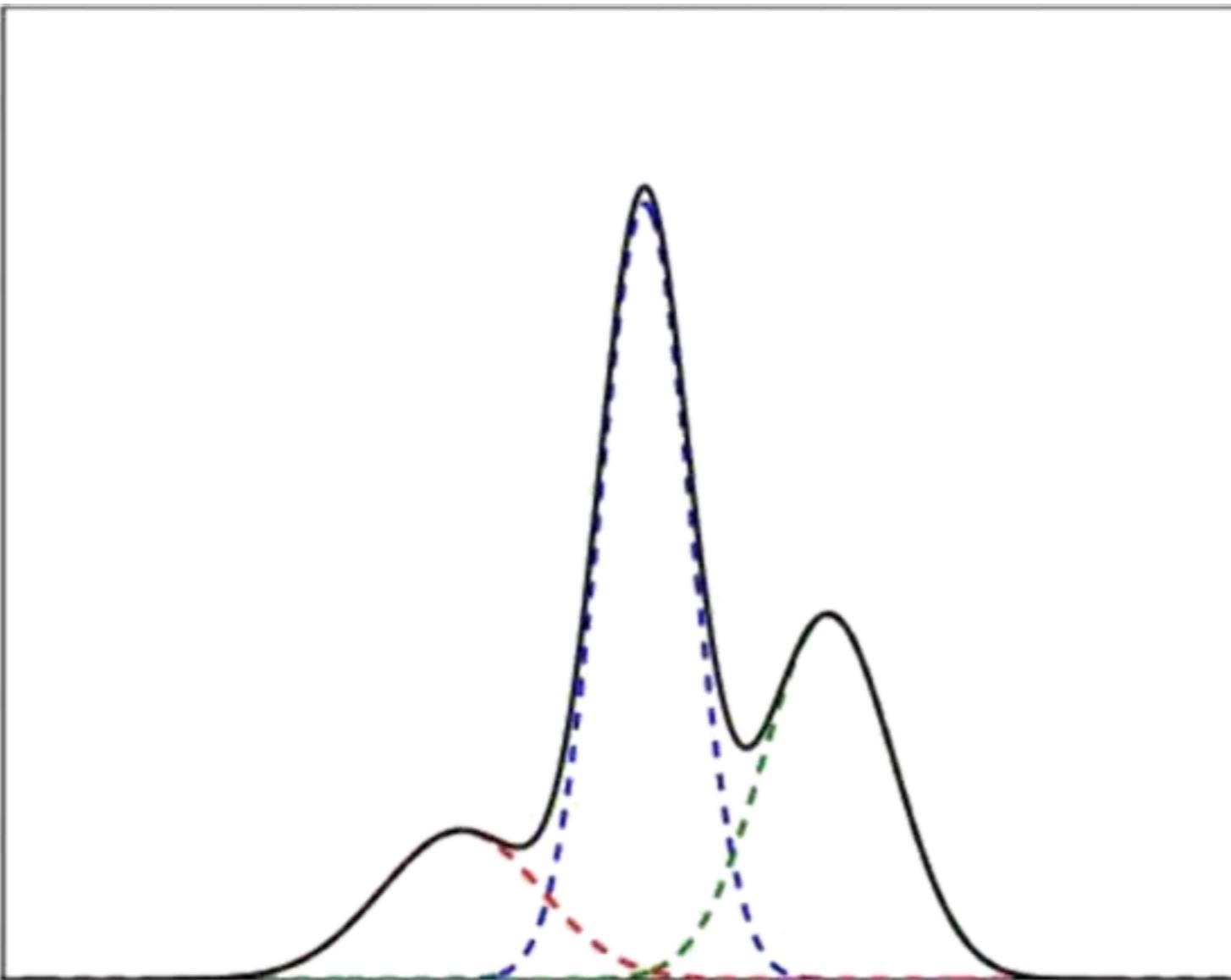
КАК ВЫГЛЯДИТ СМЕСЬ РАСПРЕДЕЛЕНИЙ



КАК ВЫГЛЯДИТ СМЕСЬ РАСПРЕДЕЛЕНИЙ



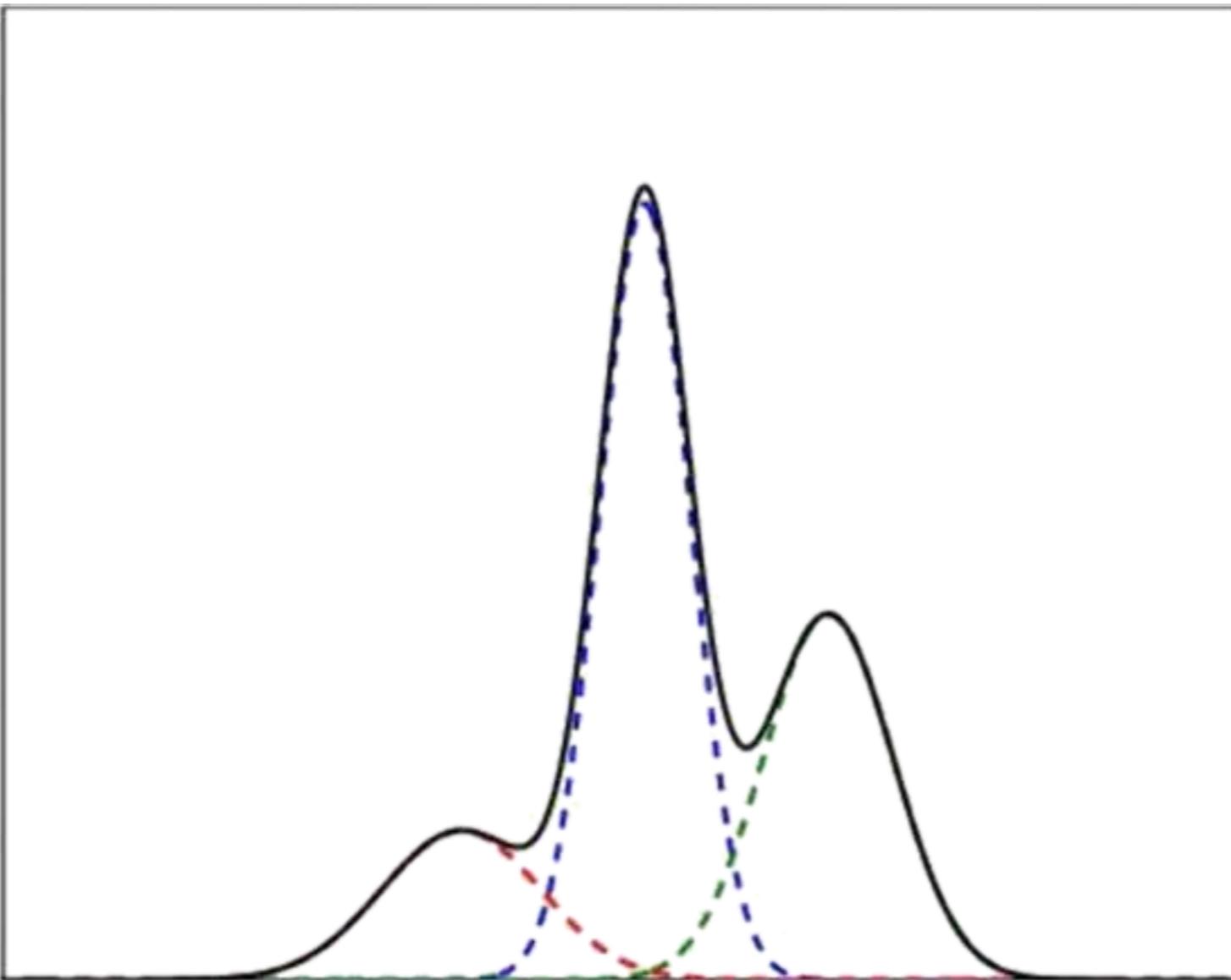
ПОЧЕМУ НЕ РЕШИТЬ ЗАДАЧУ «В ЛОБ»



$$p(x) = \sum_{j=1}^K w_j p_j(x), \quad p_j(x) = \varphi(\theta_j; x)$$

$$\mathbf{w}, \boldsymbol{\theta} = \operatorname{argmax}_{\boldsymbol{\theta}, \mathbf{w}} \sum_{j=1}^K \ln p(x_i)$$

ПОЧЕМУ НЕ РЕШИТЬ ЗАДАЧУ «В ЛОБ»



$$p(x) = \sum_{j=1}^K w_j p_j(x), \quad p_j(x) = \varphi(\theta_j; x)$$

$$\mathbf{w}, \boldsymbol{\theta} = \operatorname{argmax}_{\boldsymbol{\theta}, \mathbf{w}} \sum_{j=1}^K \ln p(x_i)$$

ЕМ-АЛГОРИТМ

$$p(x) = \sum_{j=1}^K w_j p_j(x) \quad p_j(x) = \varphi(\theta_j; x)$$

› Е-шаг:

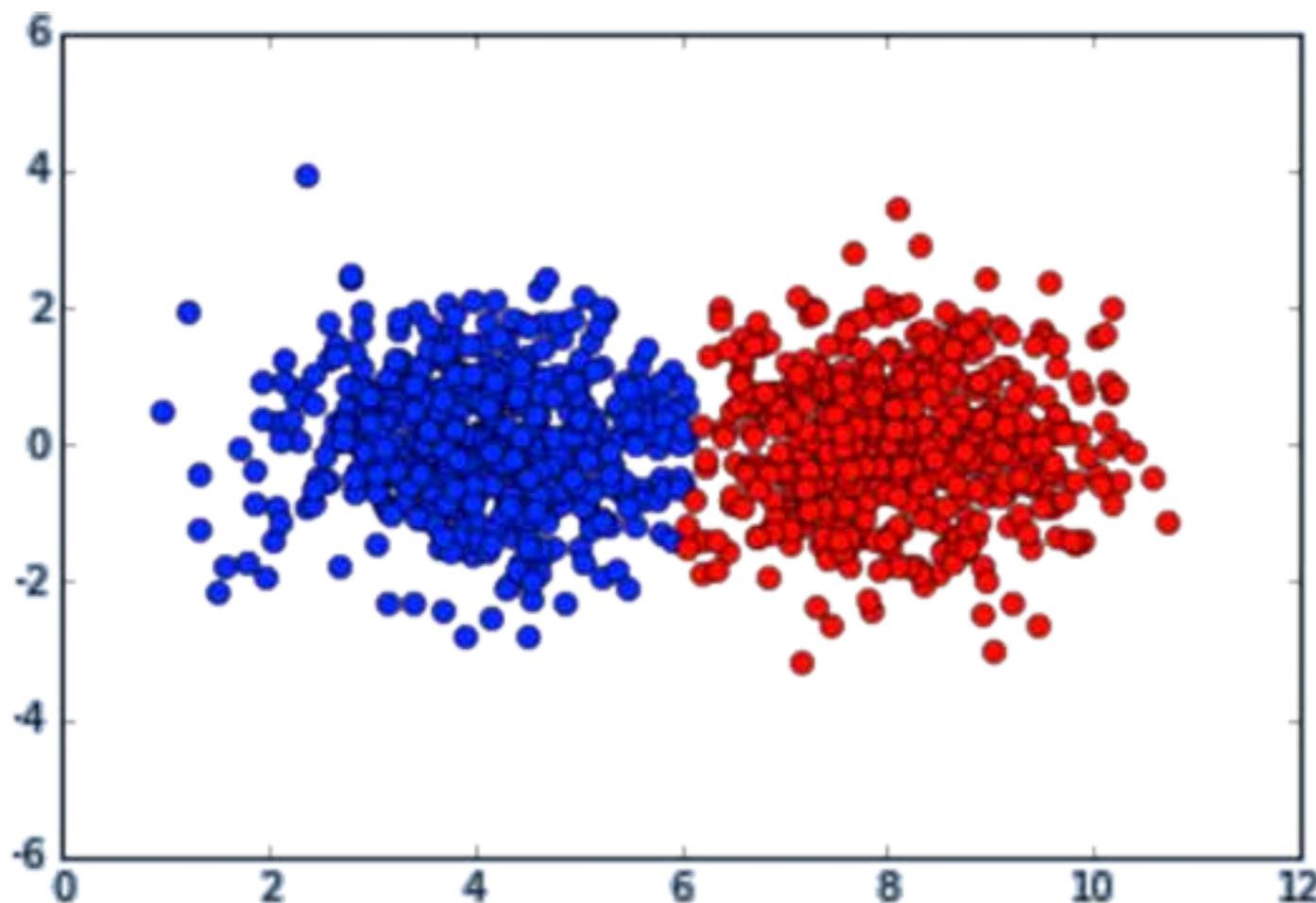
$$g_{ji} = p(j|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$$

› М-шаг:

$$w_j = \frac{1}{N} \sum_{i=1}^N g_{ji}$$

$$\theta_j = \operatorname{argmax}_{\theta} \sum_{i=1}^N g_{ji} \ln \varphi(\theta; x)$$

ПРИМЕР: 2 КЛАСТЕРА С ГАУССОВСКОЙ ПЛОТНОСТЬЮ



Относим x_i к кластеру j , для которого больше $p(j|x_i) = g_{ij}$

$$p(x) = w_1 p_1(x) + w_2 p_2(x)$$

» E-шаг: $g_{ji} = p(j|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$

» M-шаг: $w_j = \frac{1}{N} \sum_{i=1}^N g_{ji}, \quad \mu_j = \frac{1}{N w_j} \sum_{i=1}^N g_{ji} x_i$

$$\Sigma_j = \frac{1}{N w_j - 1} \sum_{i=1}^N g_{ji} (x_i - \mu_i)(x_i - \mu_j)^T$$

ПРОСТОЕ ОБЪЯСНЕНИЕ ЕМ-АЛГОРИТМА

- › Выбираем «скрытые переменные» таким образом, чтобы с ними было проще максимизировать правдоподобие
- › Е-шаг:
Оцениваем скрытые переменные
- › М-шаг:
Оцениваем w_1, \dots, w_K и $p_1(x), \dots, p_K(x)$, считая скрытые переменные зафиксированными

ПРОСТОЕ ОБЪЯСНЕНИЕ ЕМ-АЛГОРИТМА

» Е-шаг:

Для задачи разделения смеси подходят $P(j|x_i)$

Расписав по формуле Байеса, получаем:

$$P(j|x_i) = \frac{w_j p_j(x_i)}{\sum_{k=1}^K w_k p_k(x_i)}$$

» М-шаг:

Максимизируем правдоподобие по w_1, \dots, w_K и $p_1(x), \dots, p_K(x)$, считая $P(j|x_i)$ константами

Если выписать производные по параметрам и приравнять к нулю, получаем:

$$w_j = \frac{1}{N} \sum_{i=1}^N g_{ji}, \theta_j = \operatorname{argmax}_{\theta} \sum_{i=1}^N g_{ji} \ln \varphi(\theta; x)$$

 МФТИ.

КЛАССИЧЕСКОЕ ОБЪЯСНЕНИЕ ЕМ-АЛГОРИТМА

$$L(\theta; X, Z) = p(X, Z|\theta)$$

$$L(\theta; X) = p(X|\theta) = \sum_Z p(X, Z|\theta)$$

- » Е-шаг: $Q(\theta|\theta^{(t)}) = E_{Z|X,\theta^{(t)}} [\log L(\theta; X, Z)]$
- » М-шаг: $\theta^{(t+1)} = \operatorname{argmax}_{\theta} Q(\theta|\theta^{(t)})$

КАКИЕ ЕЩЁ ЗАДАЧИ РЕШАЮТ С ПОМОЩЬЮ ЕМ-АЛГОРИТМА

- Оценка параметров в других вероятностных моделях (не только в смеси распределений)
- Восстановление плотности распределения
- Классификация

РЕЗЮМЕ

- › Как выглядит кластеризация с помощью ЕМ-алгоритма
- › Постановка задачи
- › Почему не решить «в лоб»
- › Описание ЕМ алгоритма
- › ЕМ-алгоритм в случае гауссовых распределений
- › Простое объяснение метода
- › Классическое объяснение метода
- › Для чего еще используют алгоритм

АГЛОМЕРАТИВНАЯ ИЕРАХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

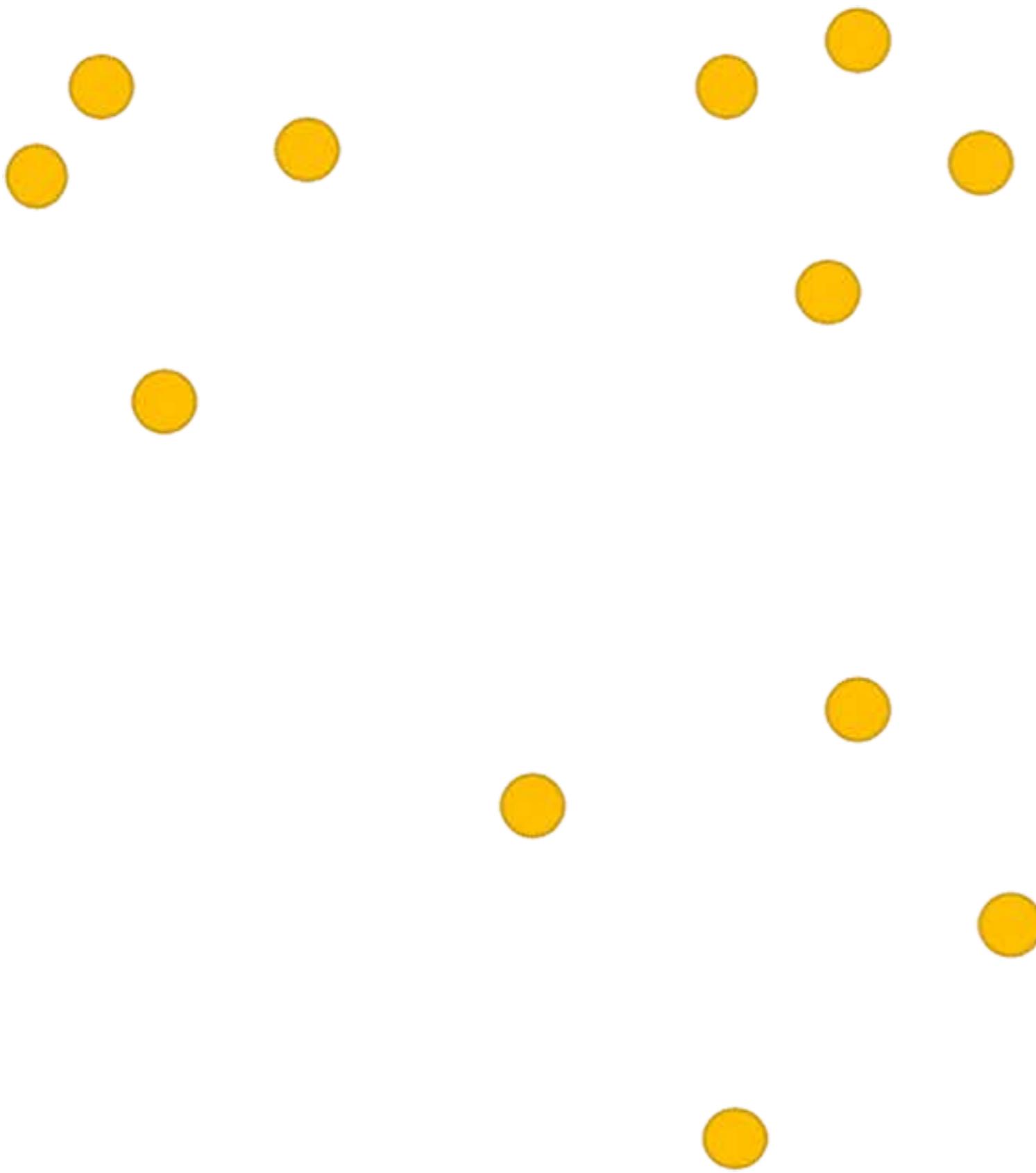
ПЛАН

- › Иерархическая кластеризация
- › Как устроена агломеративная кластеризация
- › Расстояние между кластерами
- › Формула Ланса-Уильямса
- › Дендрограммы
- › Примеры работы

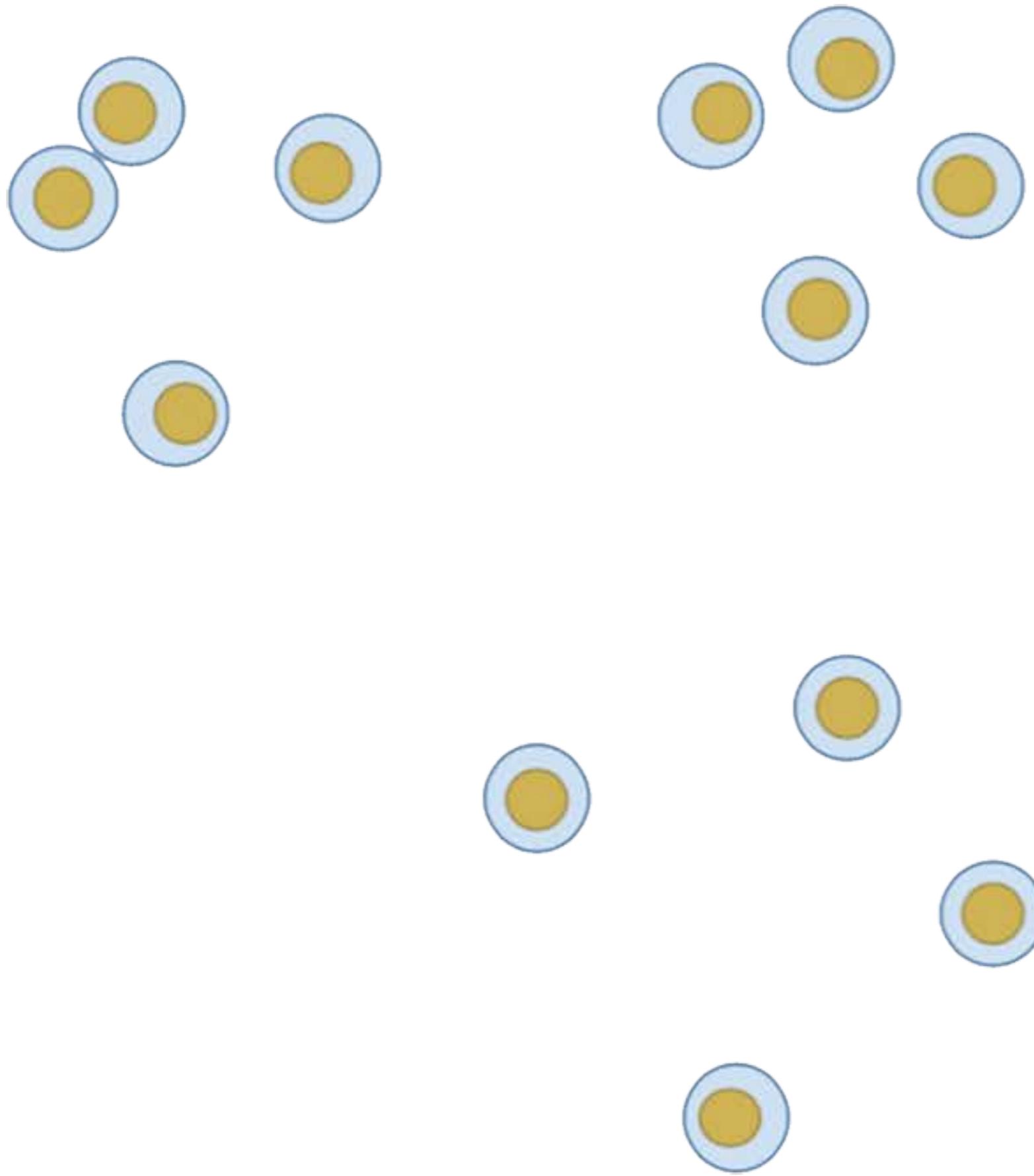
ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

- › Агломеративная
- › Дивизионная или дивизимная

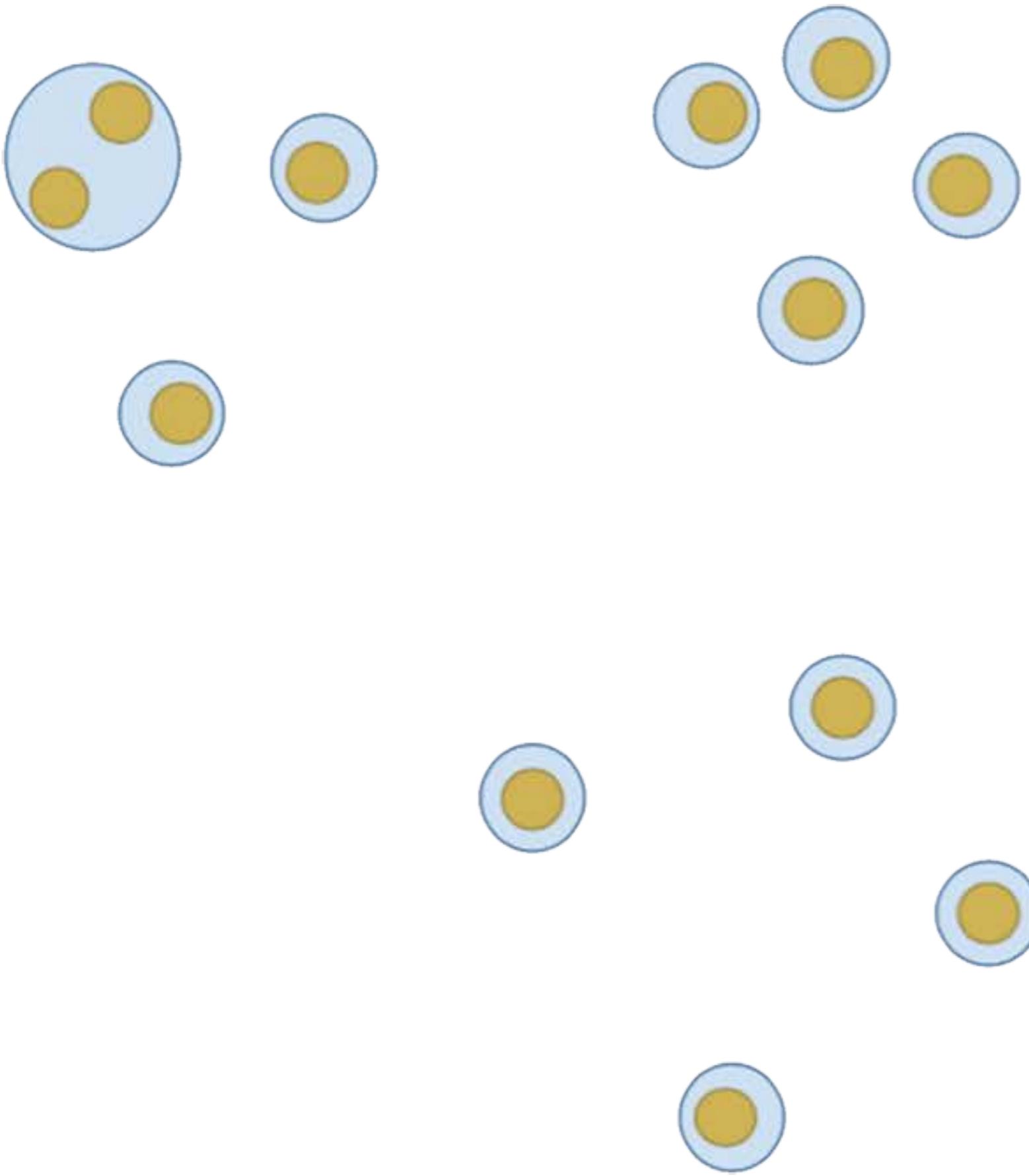
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



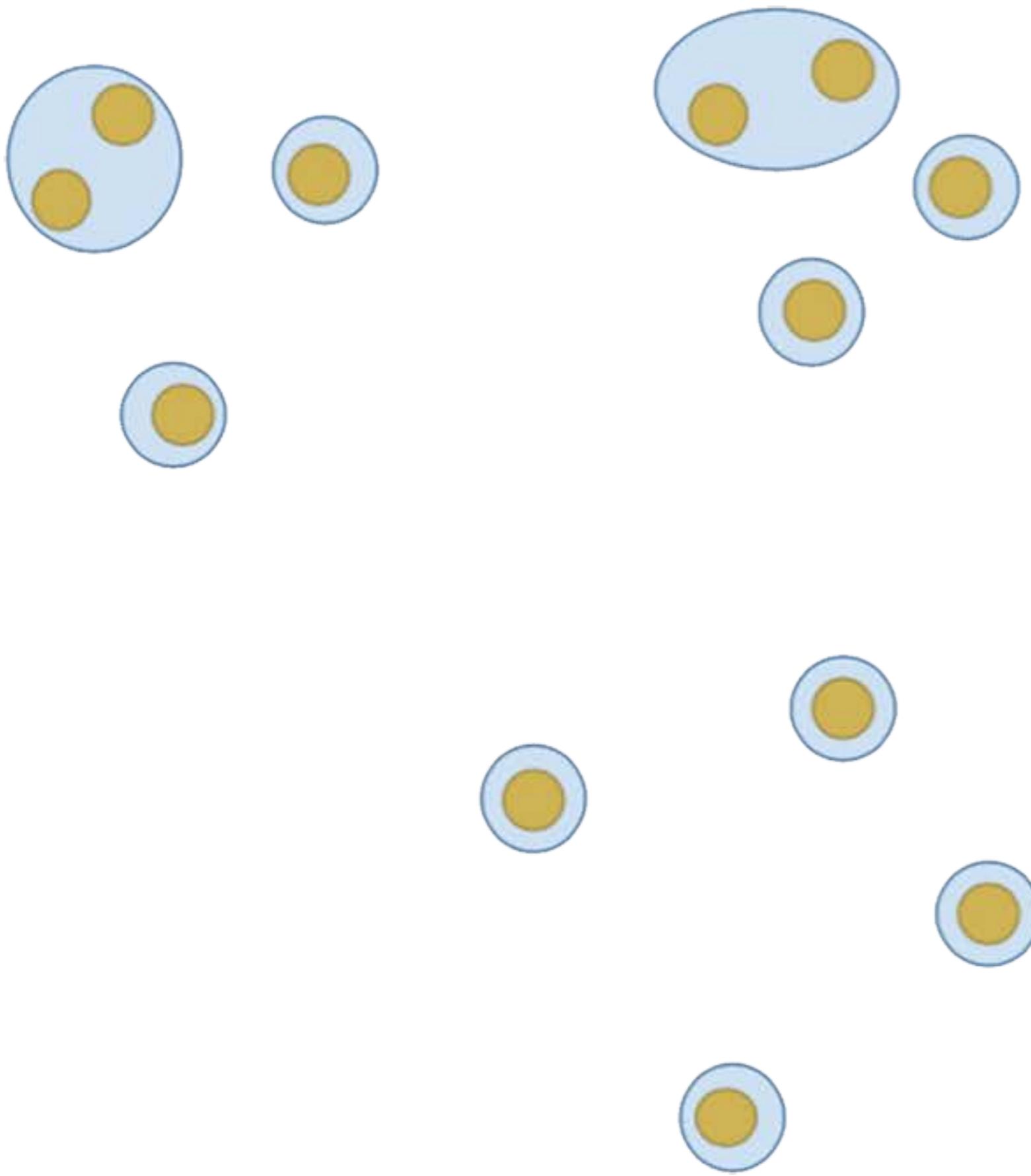
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



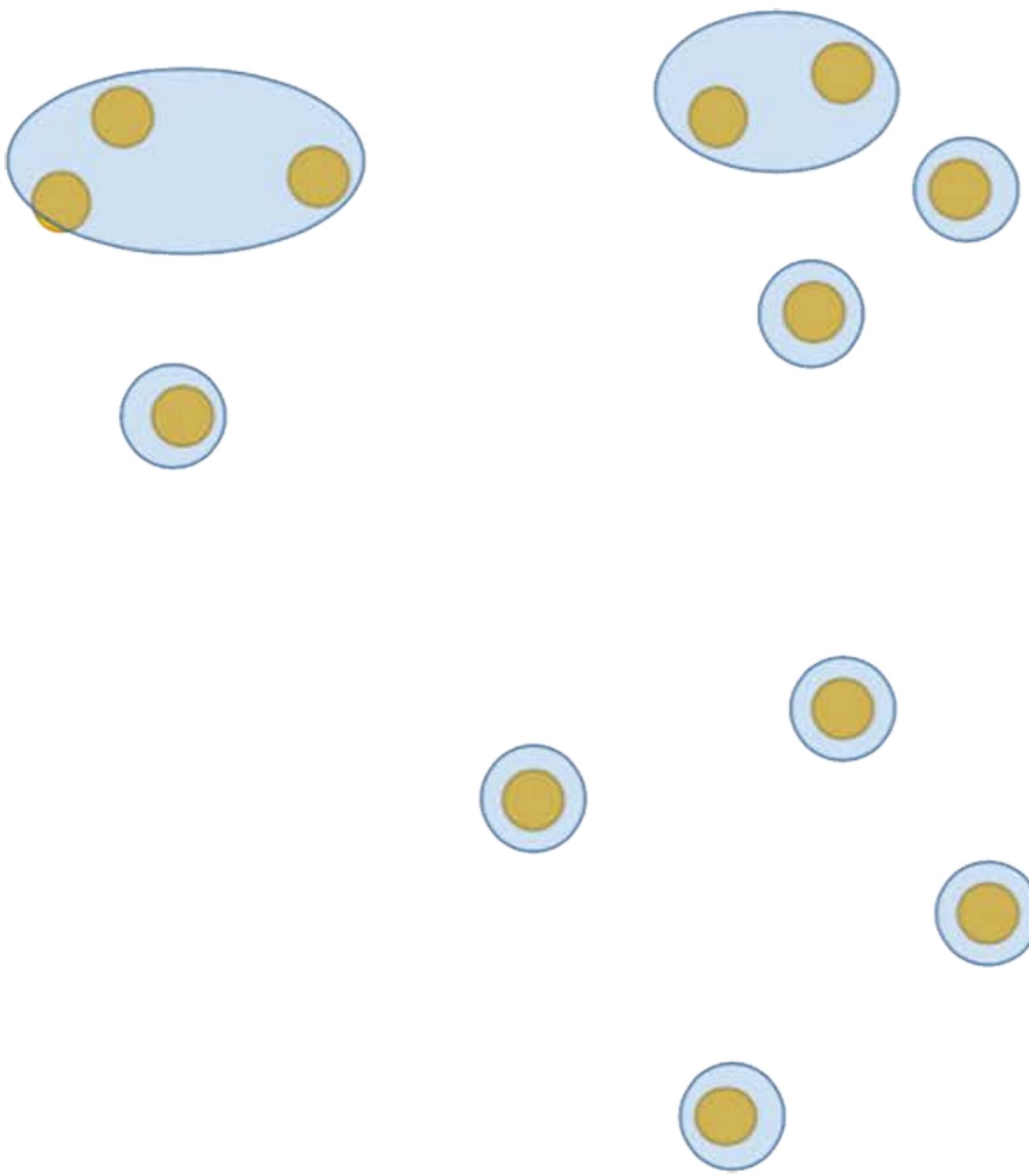
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



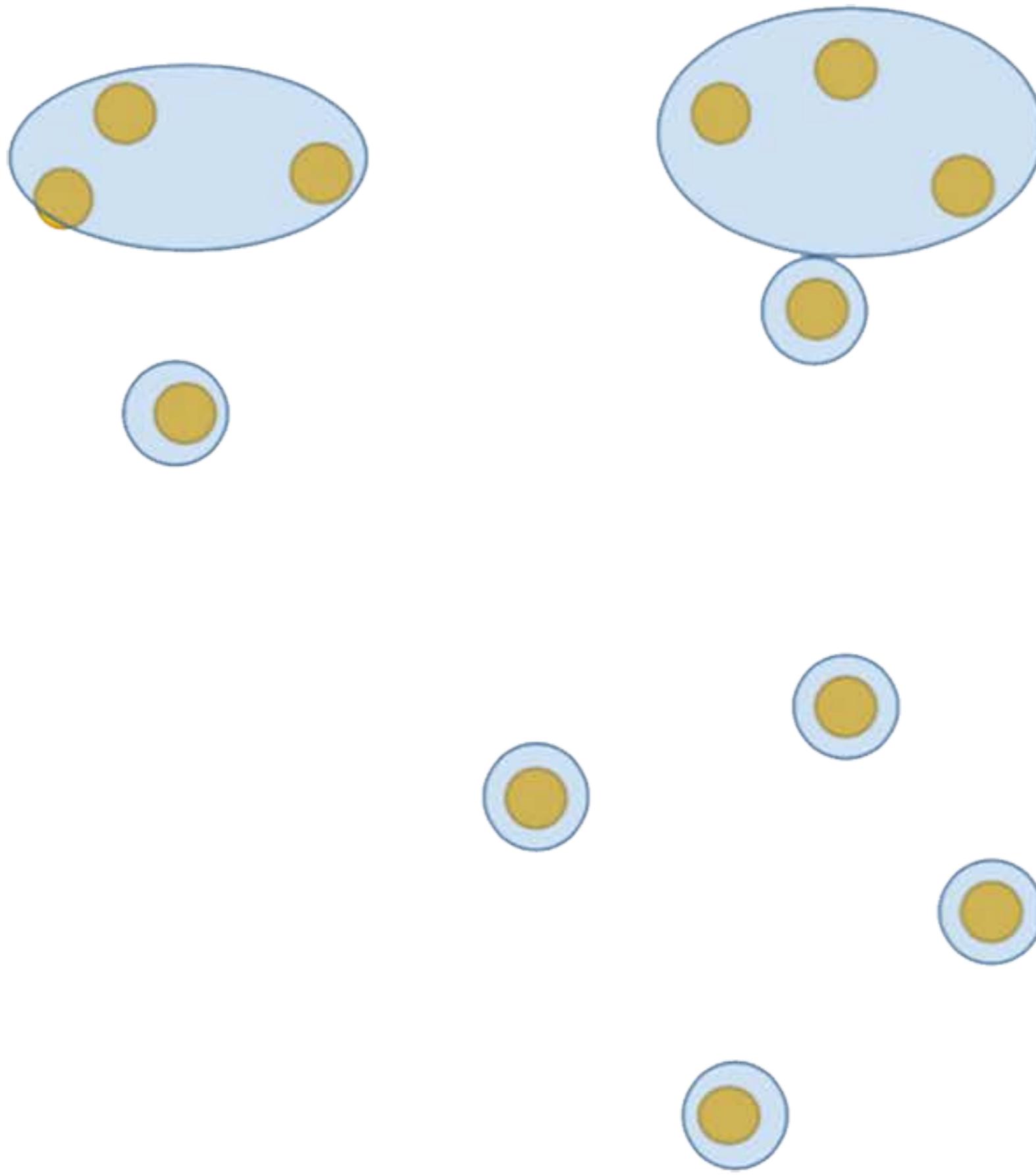
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



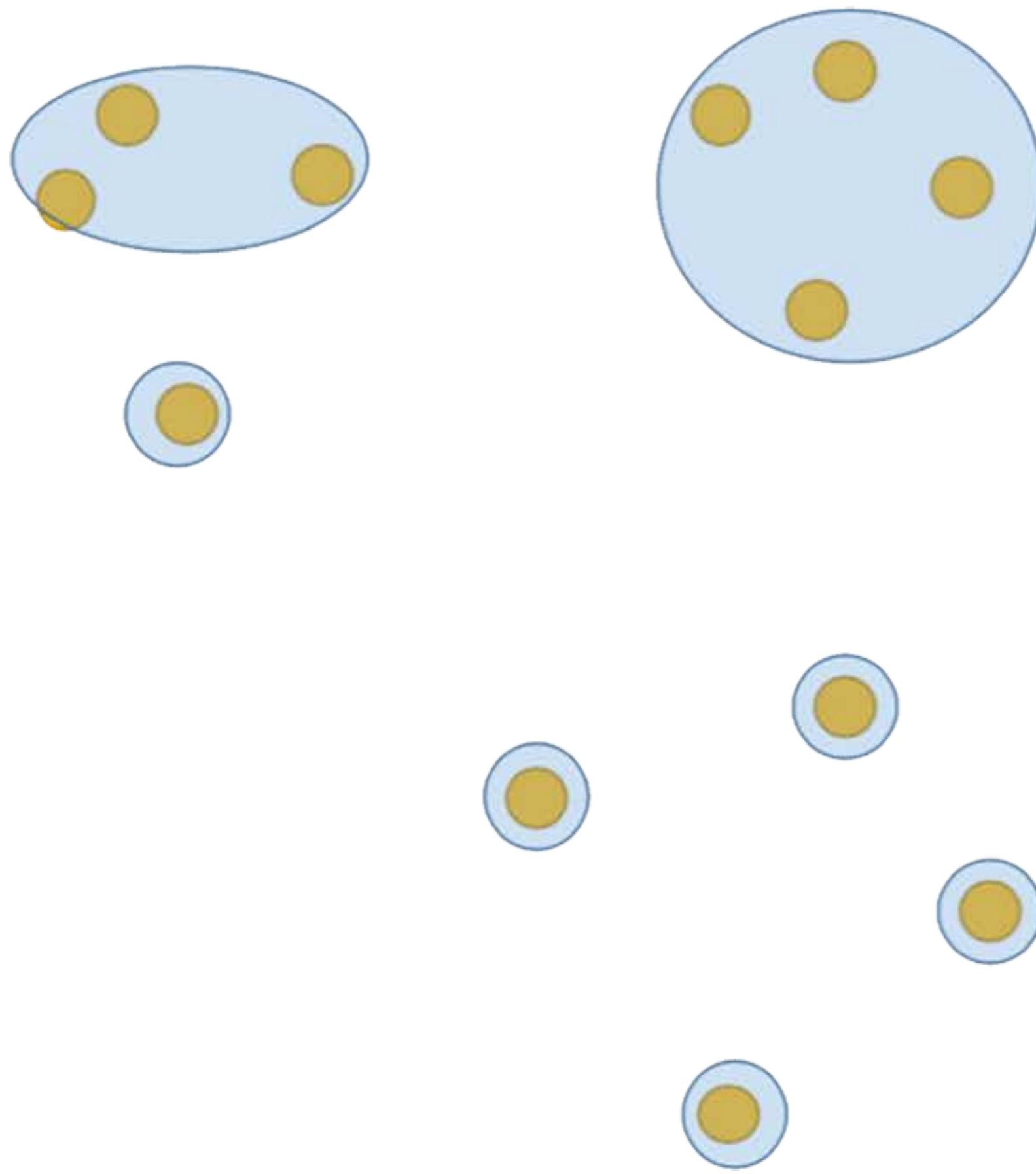
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



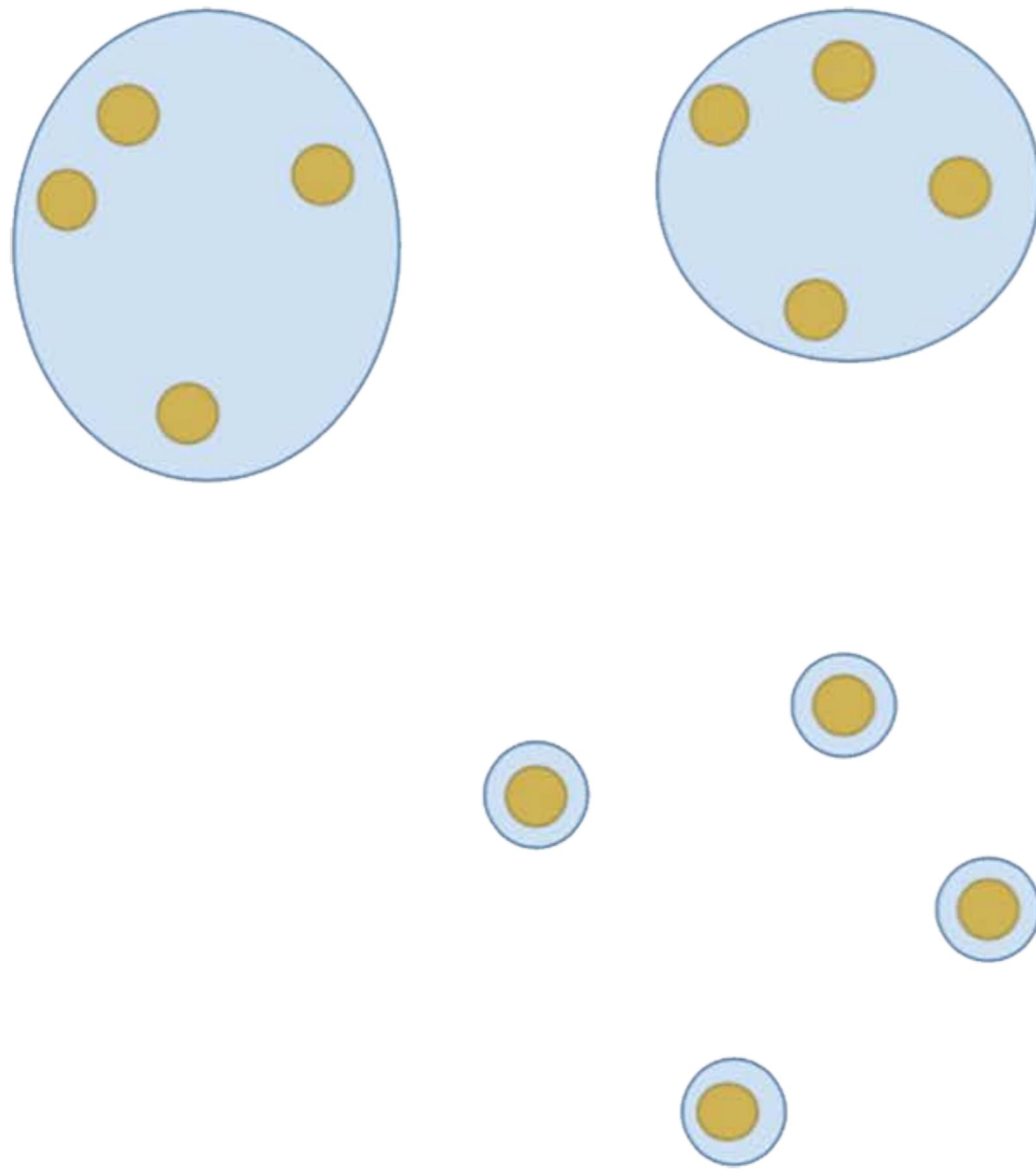
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



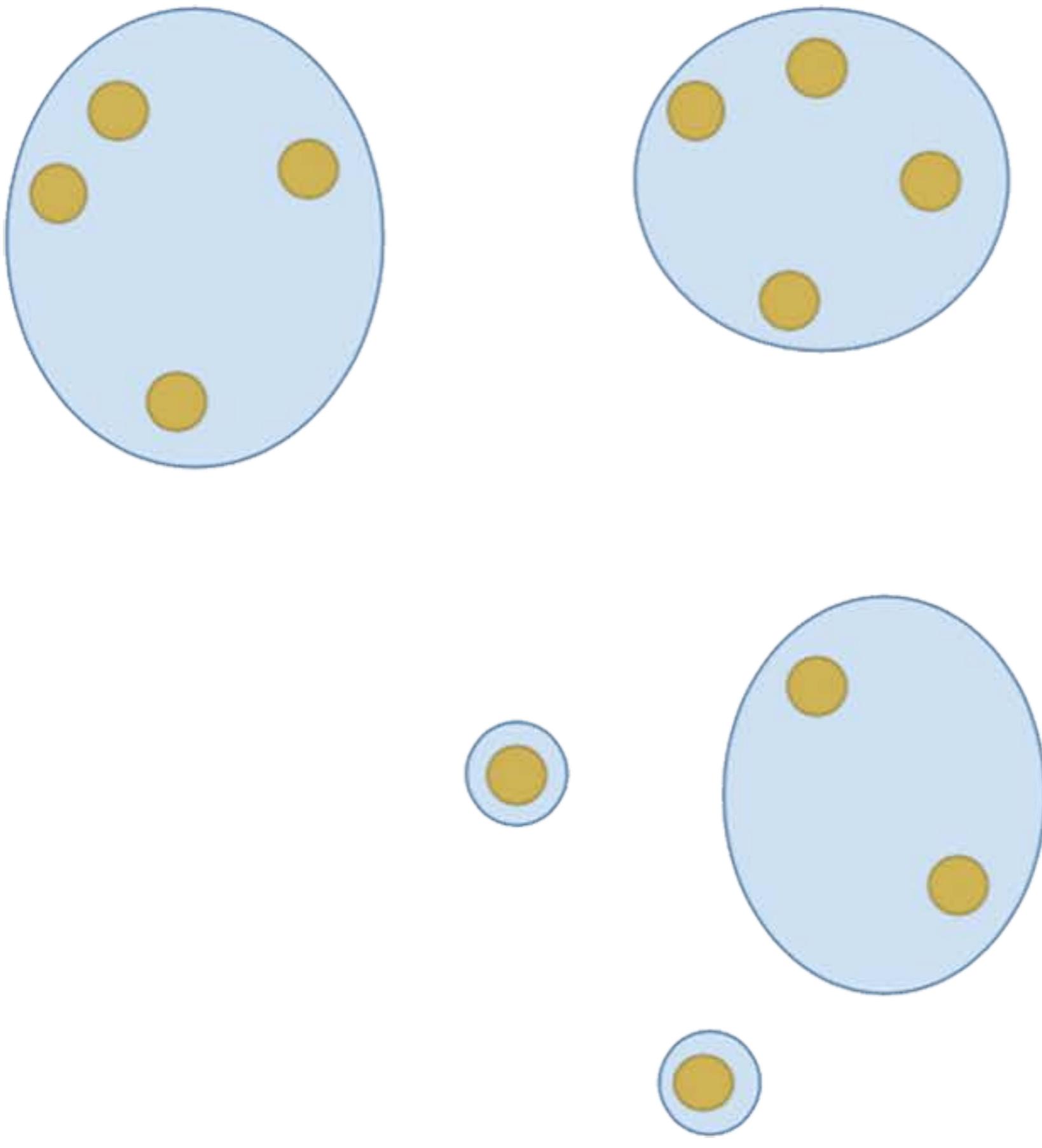
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



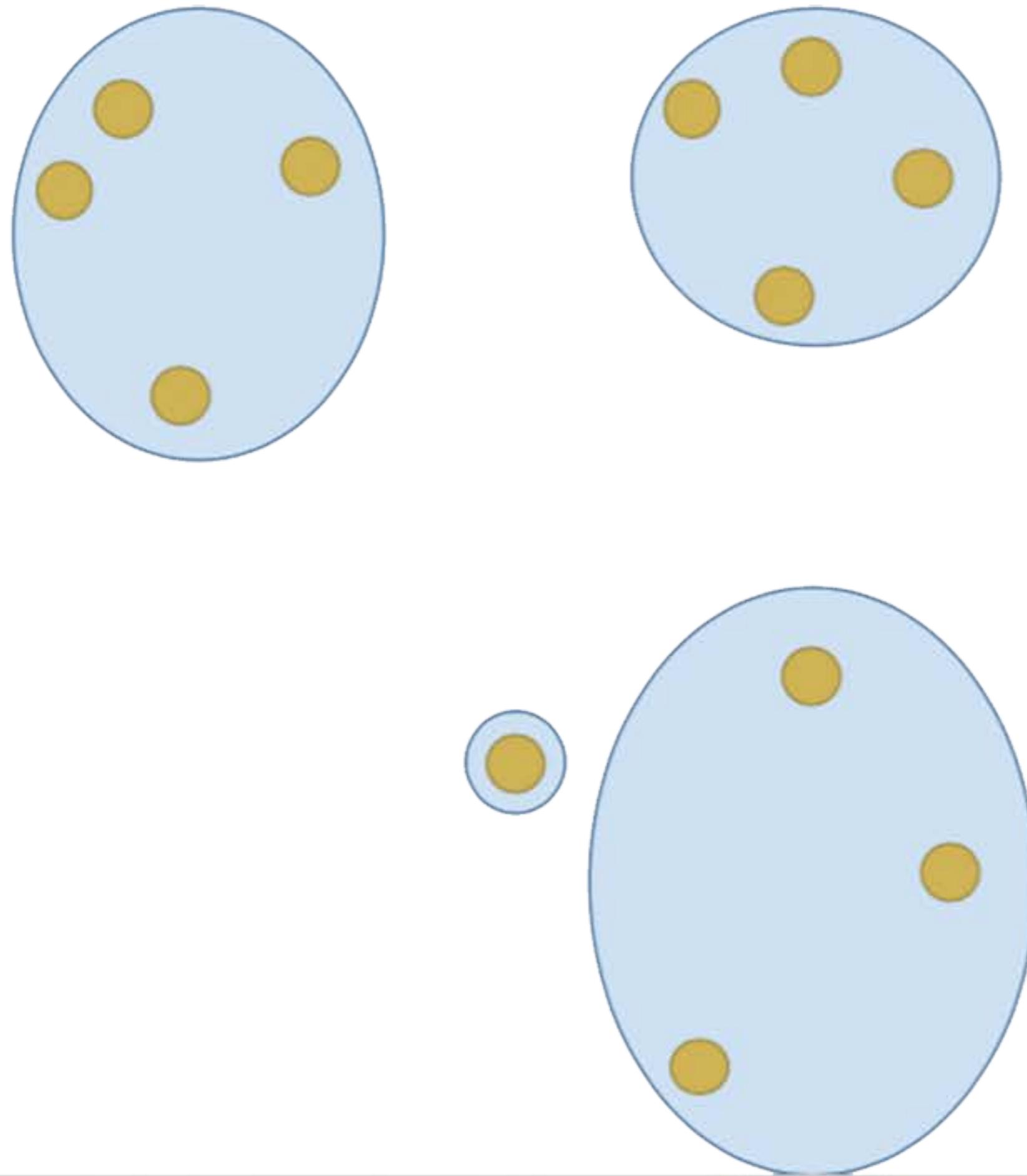
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



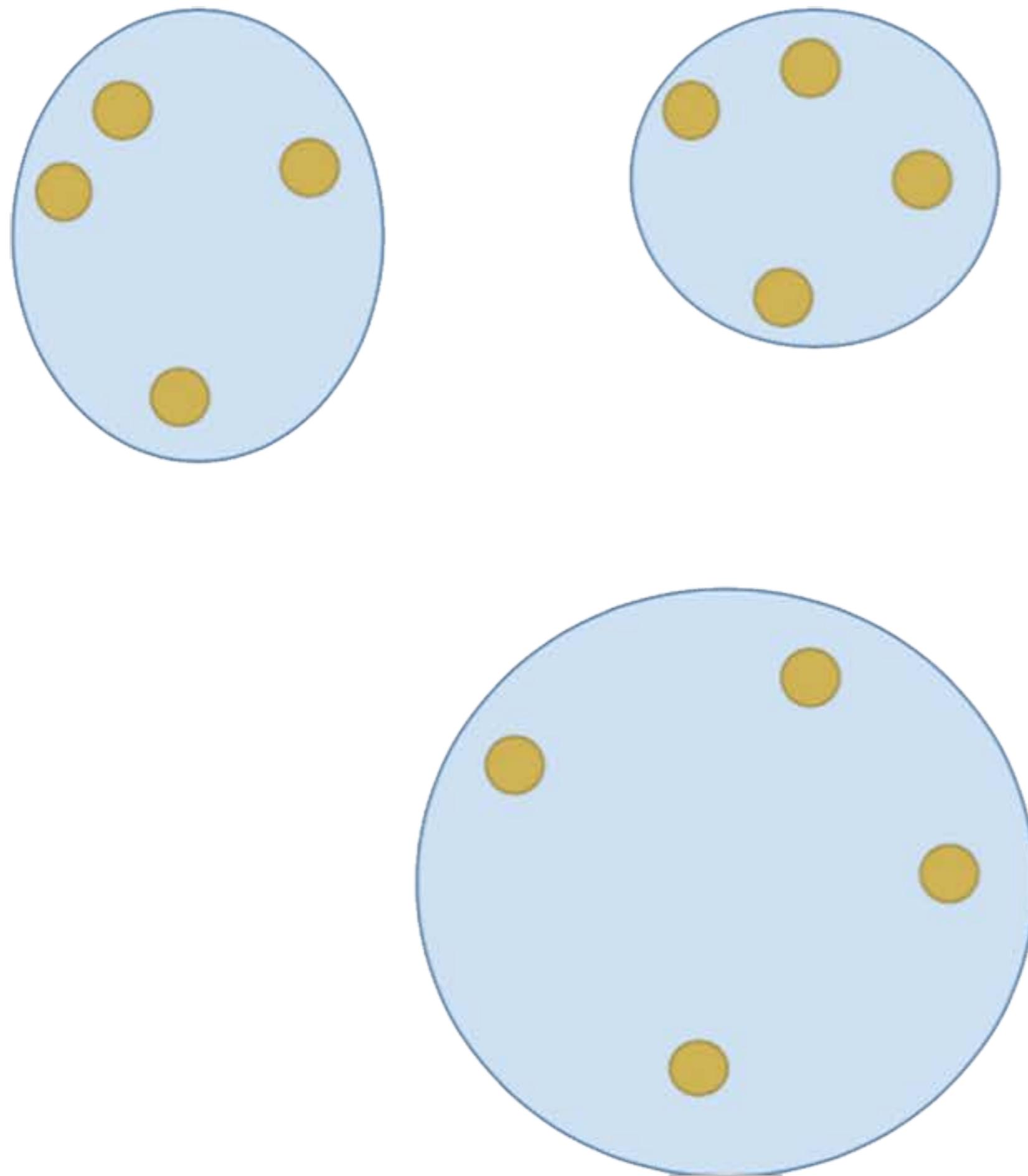
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



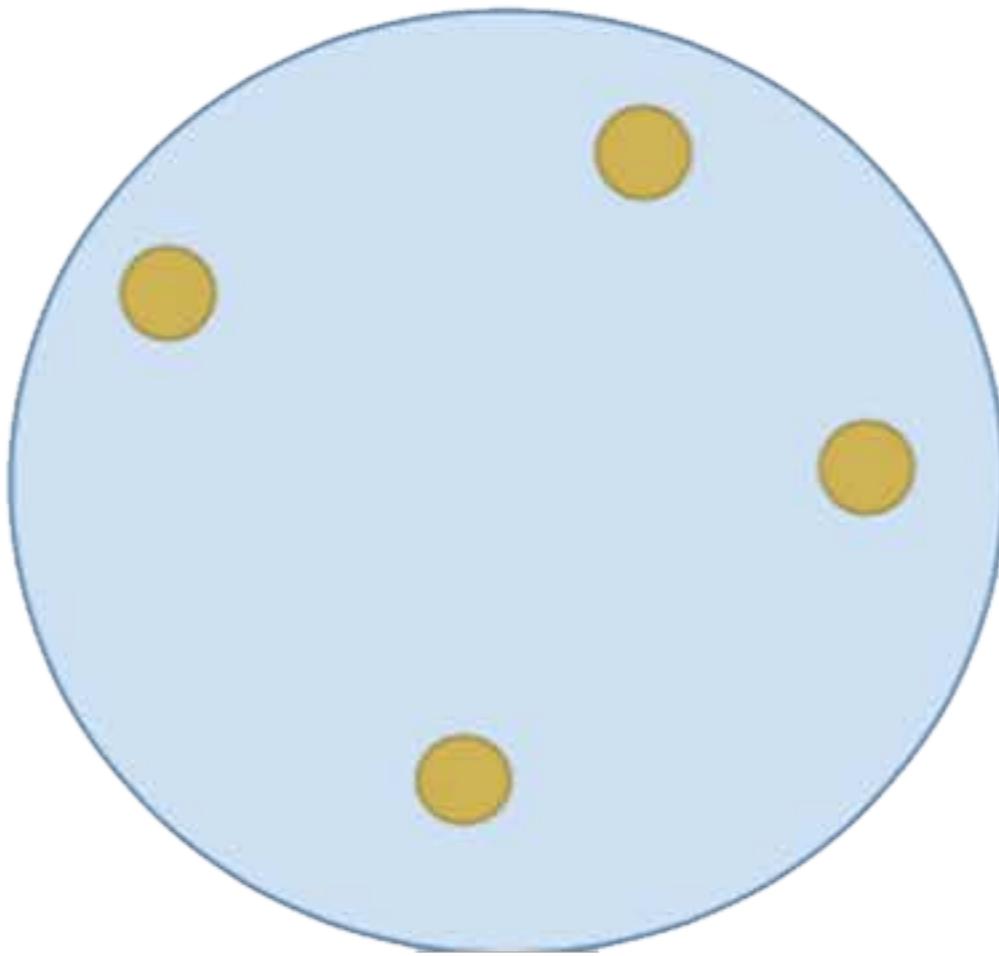
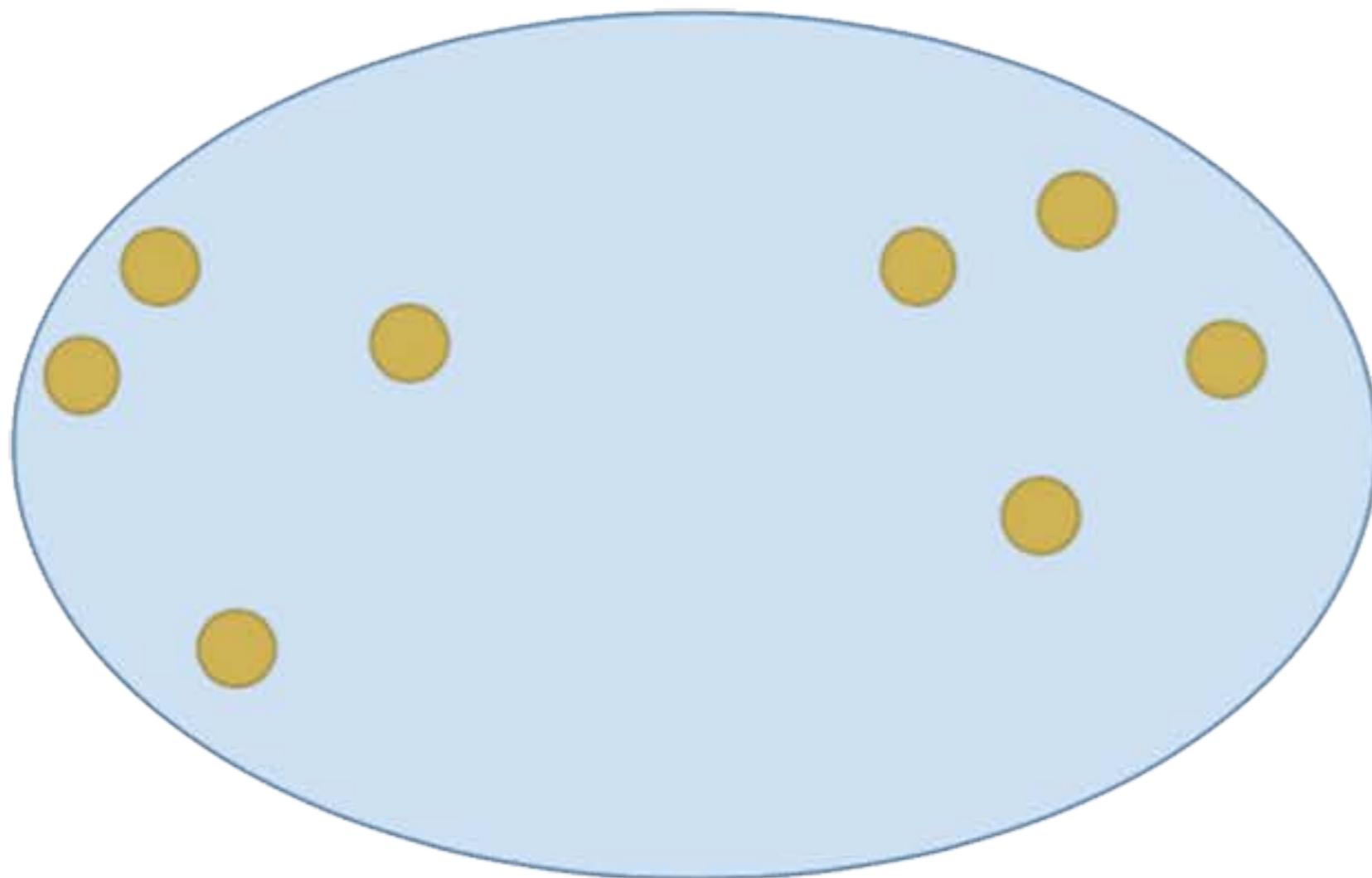
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



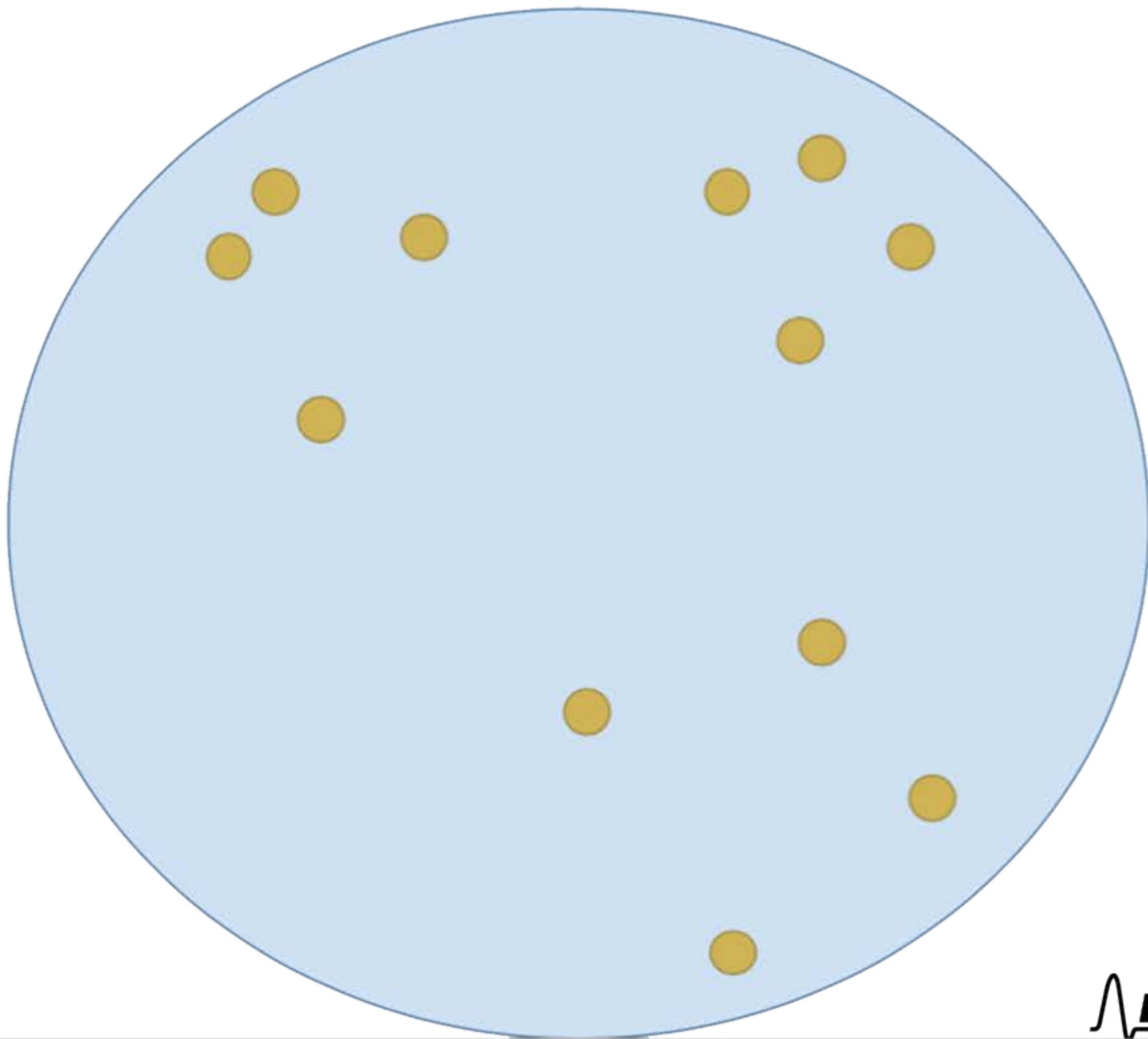
АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ



АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ

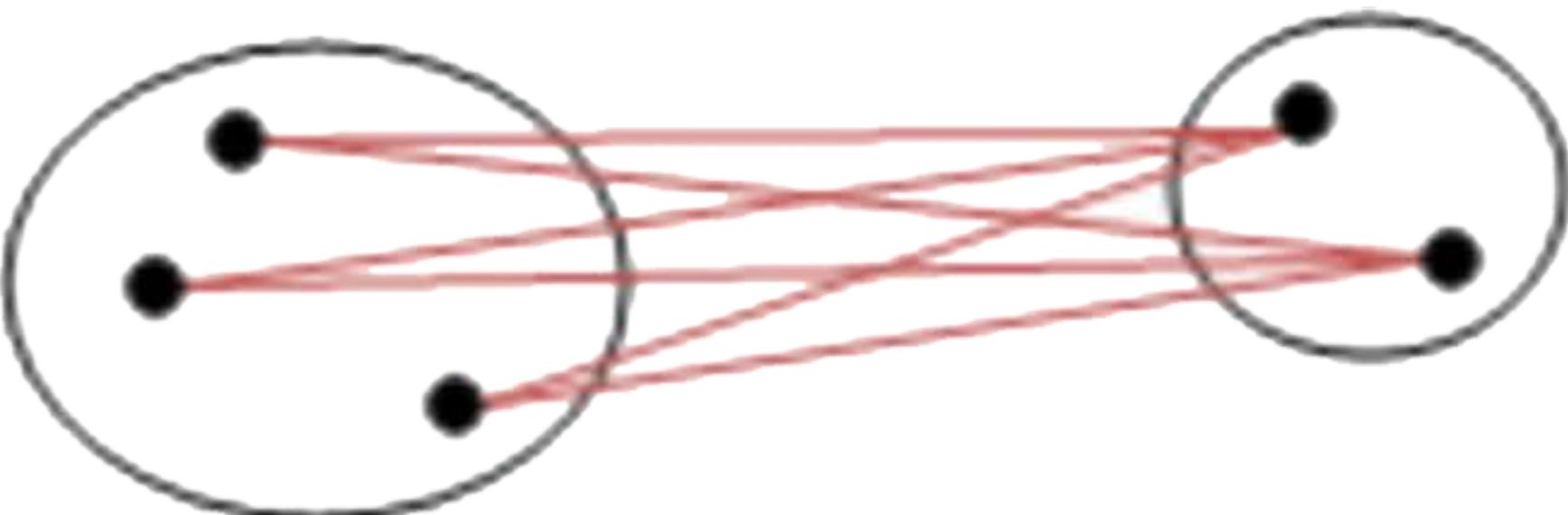


АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ

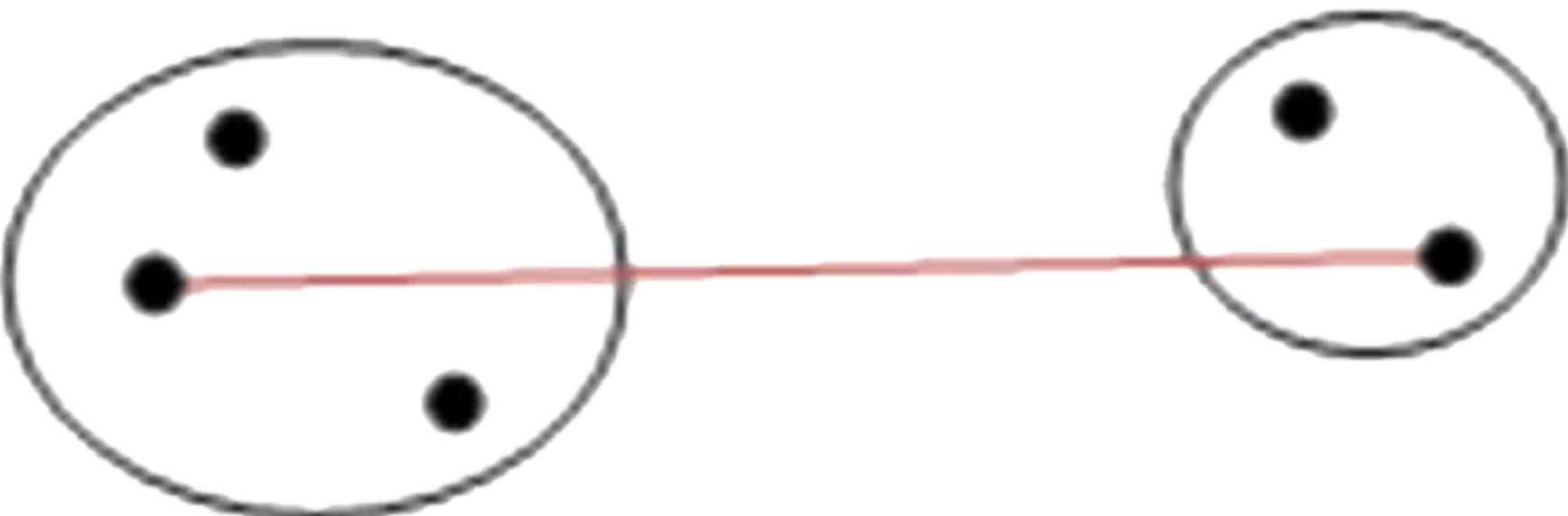


РАССТОЯНИЯ МЕЖДУ КЛАСТЕРАМИ

Average linkage



Complete linkage



Single linkage



ФОРМУЛА ЛАНСА-УИЛЬЯМСА

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \\ + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

ФОРМУЛА ЛАНСА-УИЛЬЯМСА

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \\ + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

» Расстояние ближнего соседа:

$$R^b(W, S) = \min_{w \in W, s \in S} \rho(w, s)$$

$$\alpha_U = \alpha_V = \frac{1}{2}$$

$$\beta = 0, \gamma = -\frac{1}{2}$$

» Расстояние дальнего соседа:

$$R^d(W, S) = \max_{w \in W, s \in S} \rho(w, s)$$

$$\alpha_U = \alpha_V = \frac{1}{2}$$

$$\beta = 0, \gamma = \frac{1}{2}$$

» Среднее расстояние:

$$R^c(W, S) = \frac{1}{|W||S|} \sum_{w \in W} \sum_{s \in S} \rho(w, s)$$

$$\alpha_U = \frac{|U|}{|W|}$$

$$\alpha_V = \frac{|V|}{|W|}$$

$$\beta = \gamma = 0$$

ФОРМУЛА ЛАНСА-УИЛЬЯМСА

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \\ + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

➤ Расстояние между центрами:

$$R^u(W, S) = \rho^2 \left(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|} \right)$$

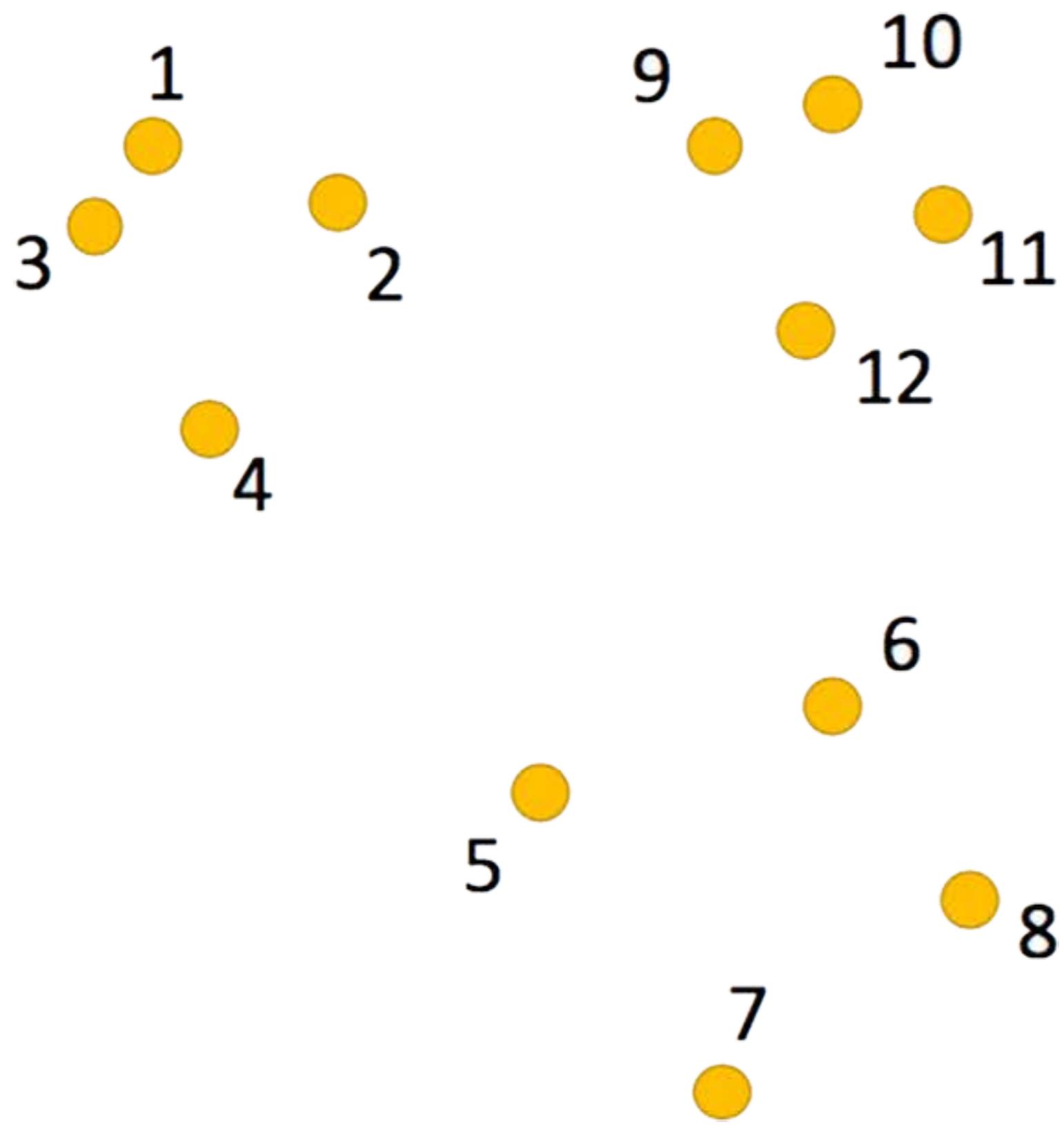
$$\alpha_U = \frac{|U|}{|W|}, \quad \alpha_V = \frac{|V|}{|W|}, \quad \beta = -\alpha_U \alpha_V, \quad \gamma = 0$$

➤ Расстояние Уорда:

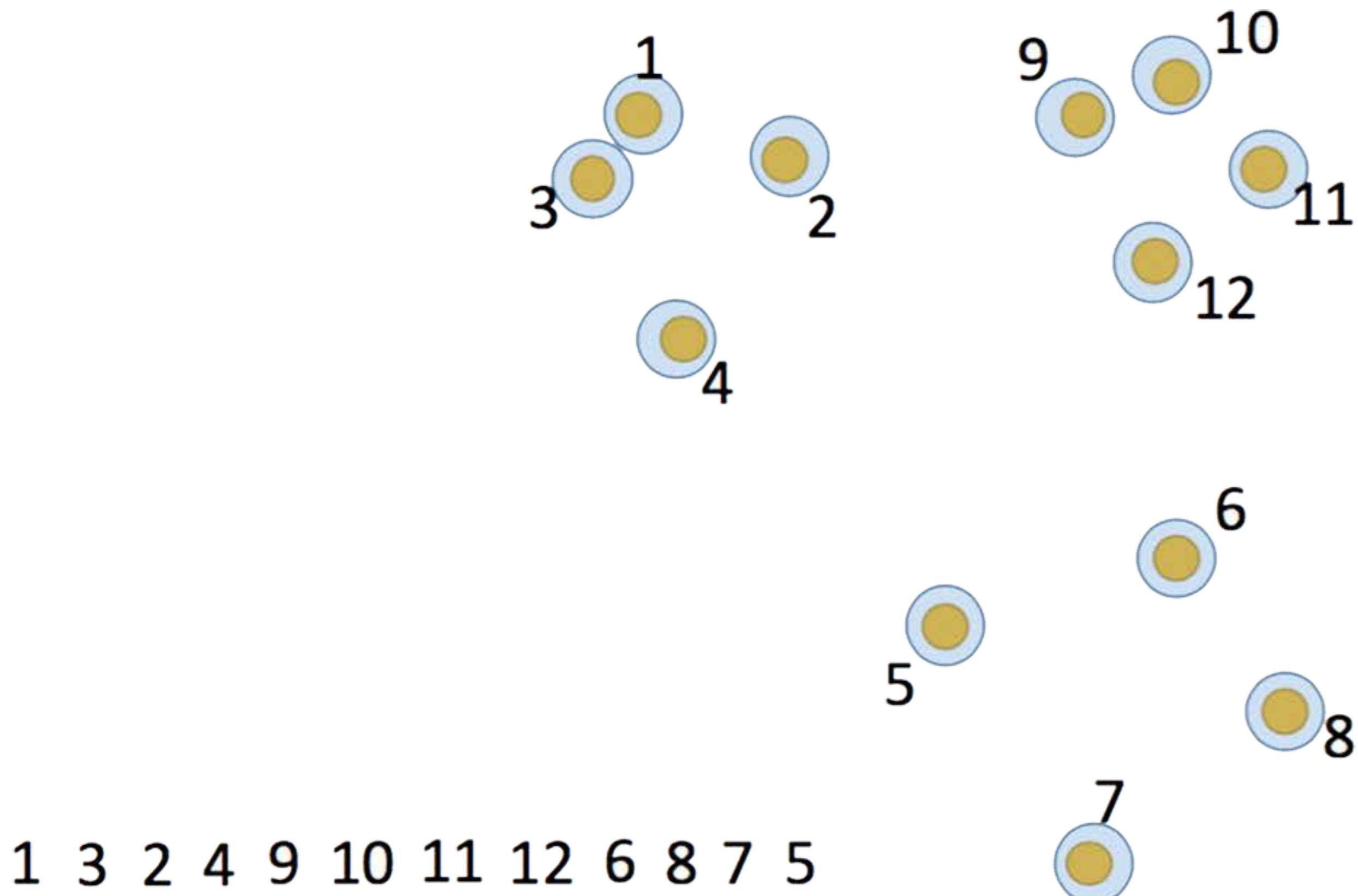
$$R^y(W, S) = \frac{|S||W|}{|S| + |W|} \rho^2 \left(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|} \right)$$

$$\alpha_U = \frac{|S| + |U|}{|S| + |W|}, \quad \alpha_V = \frac{|S| + |V|}{|S| + |W|}, \quad \beta = \frac{-|S|}{|S| + |W|}, \quad \gamma = 0$$

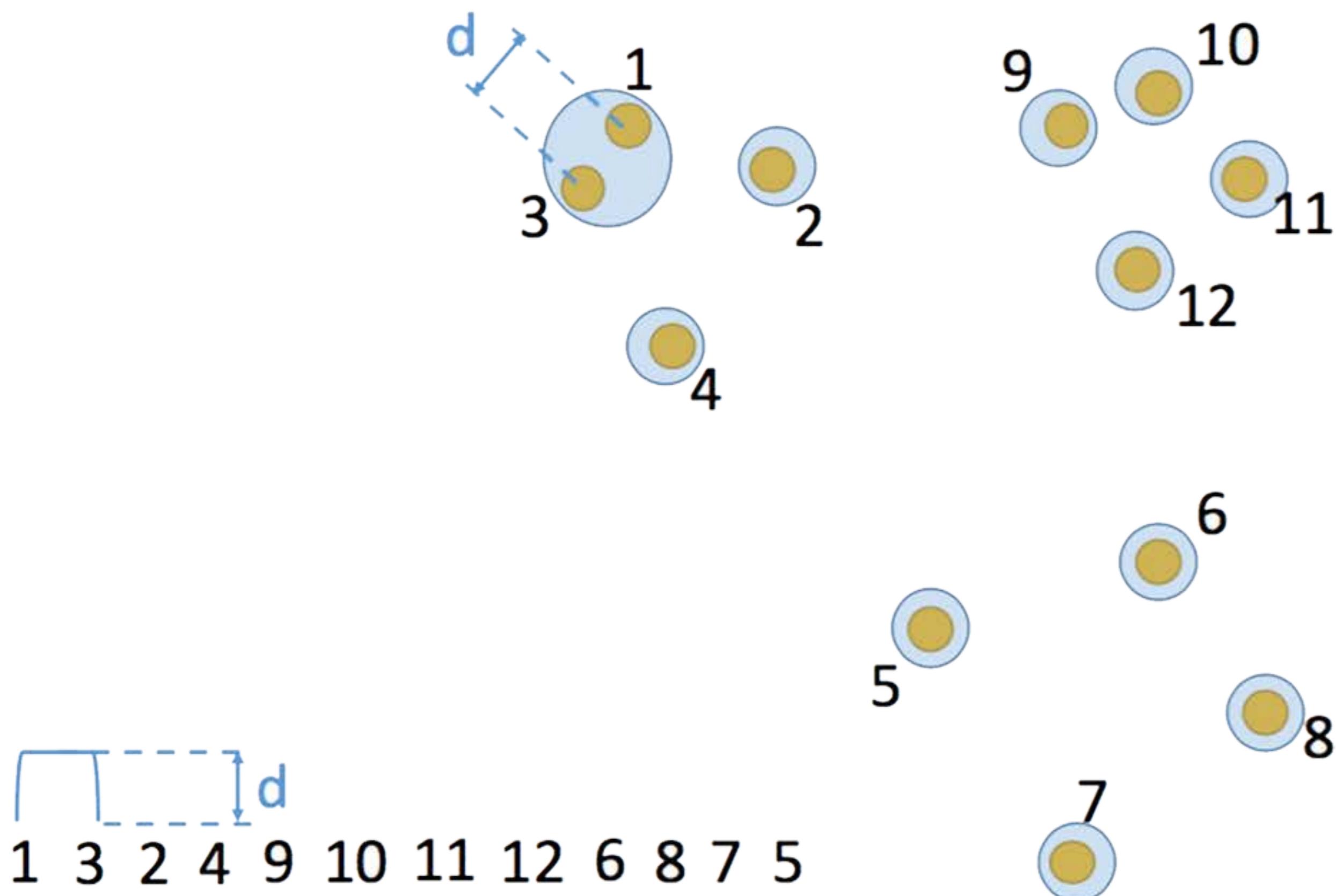
ДЕНДРОГРАММА



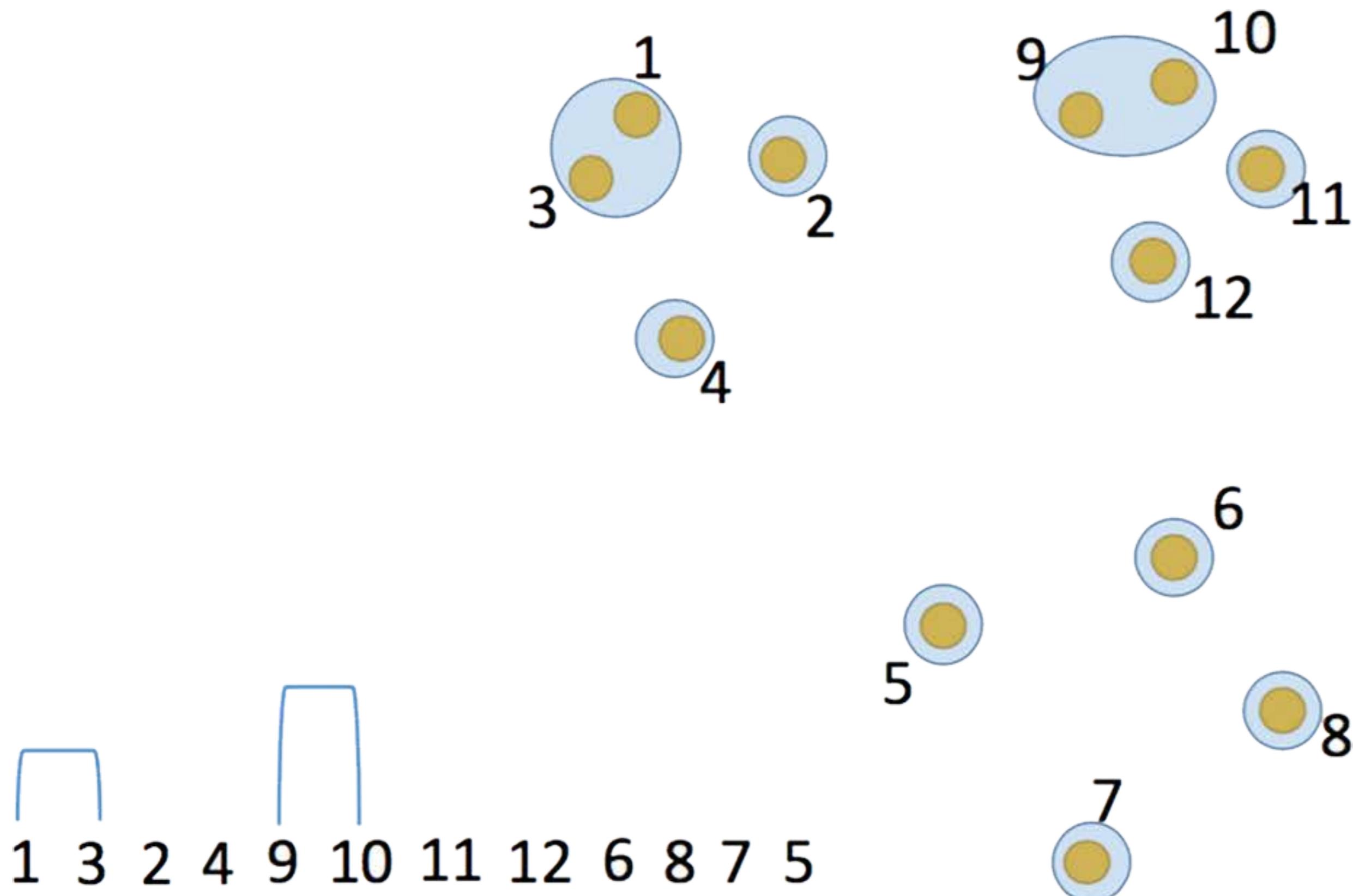
ДЕНДРОГРАММА



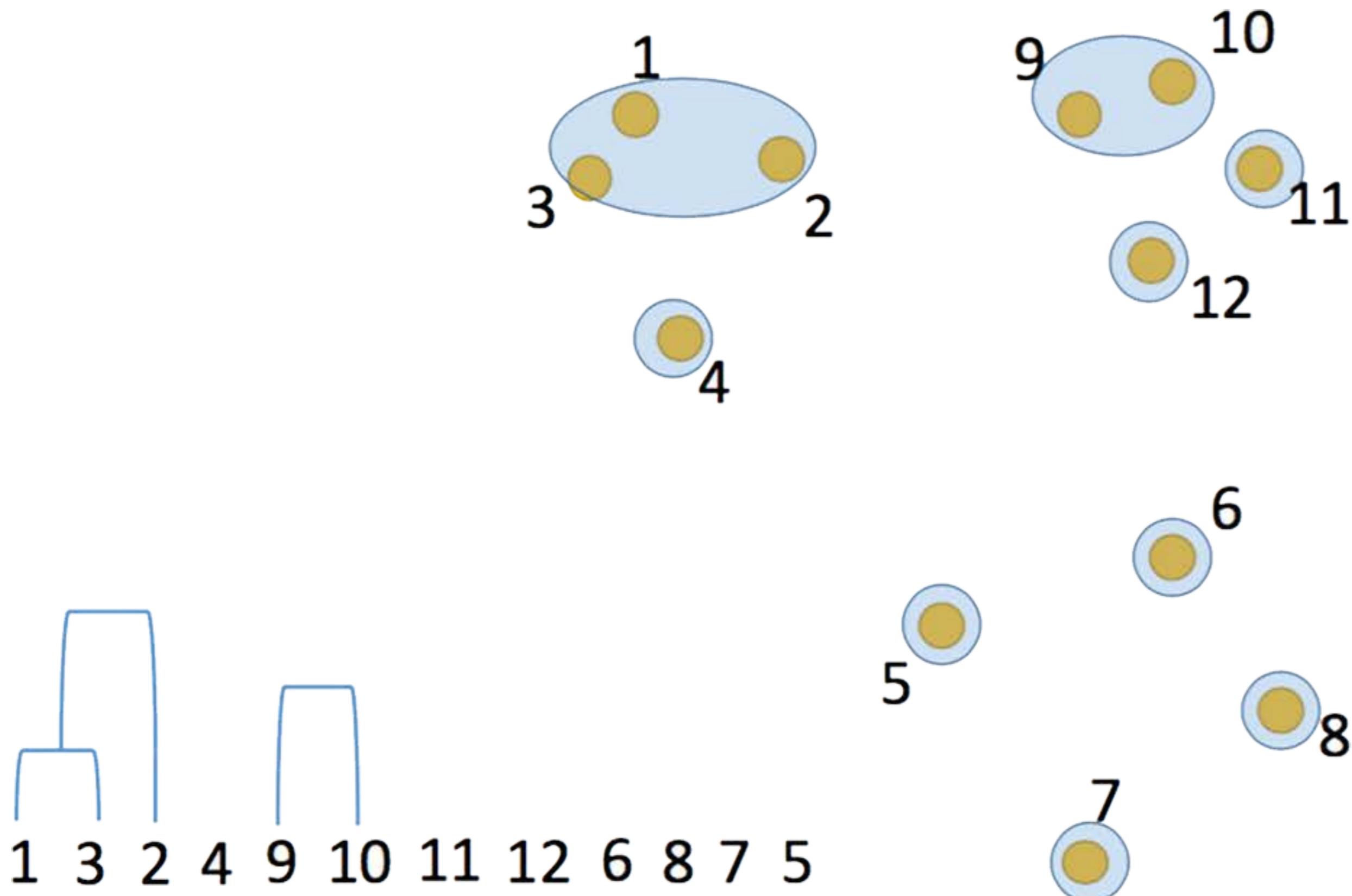
ДЕНДРОГРАММА



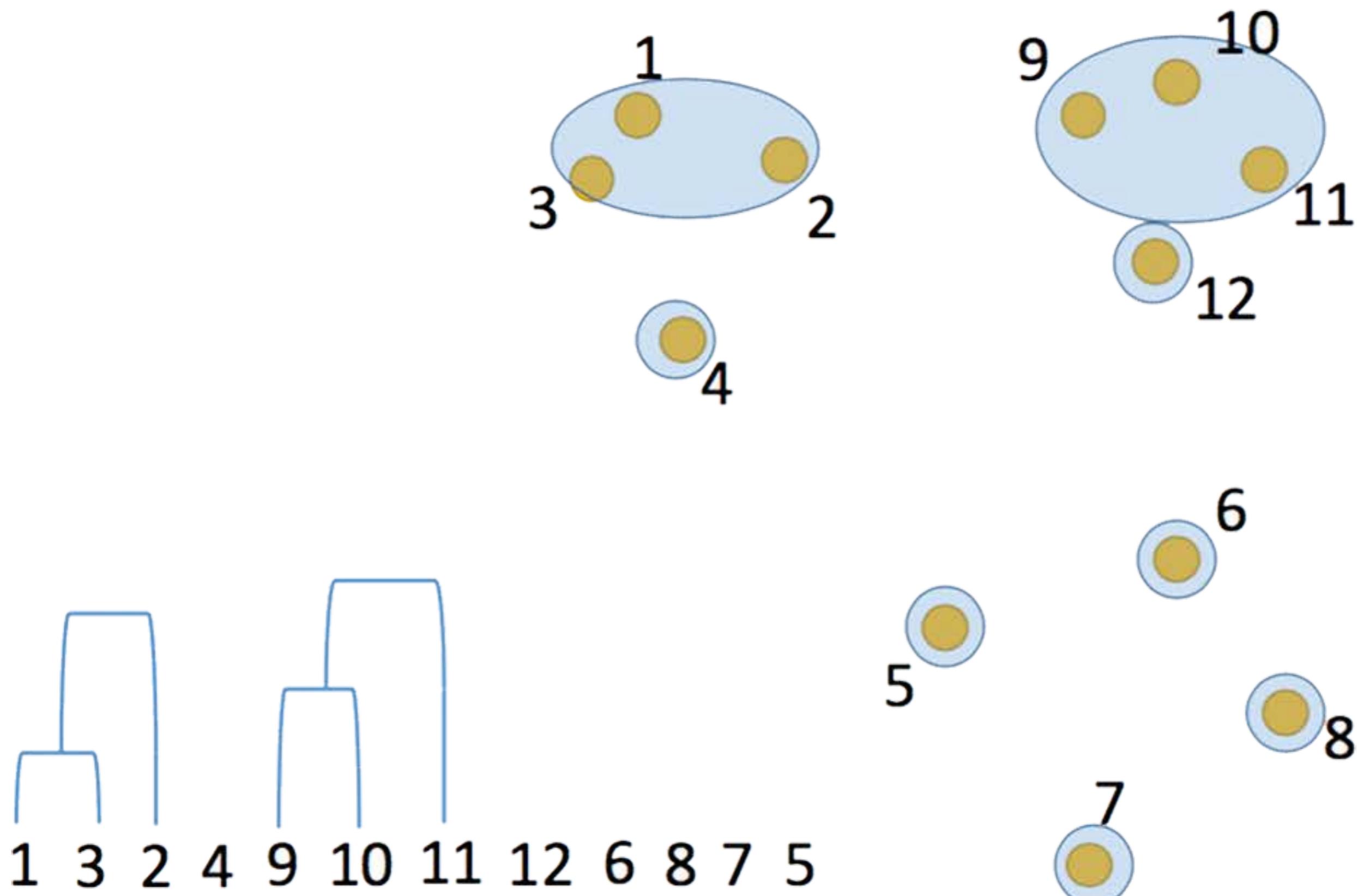
ДЕНДРОГРАММА



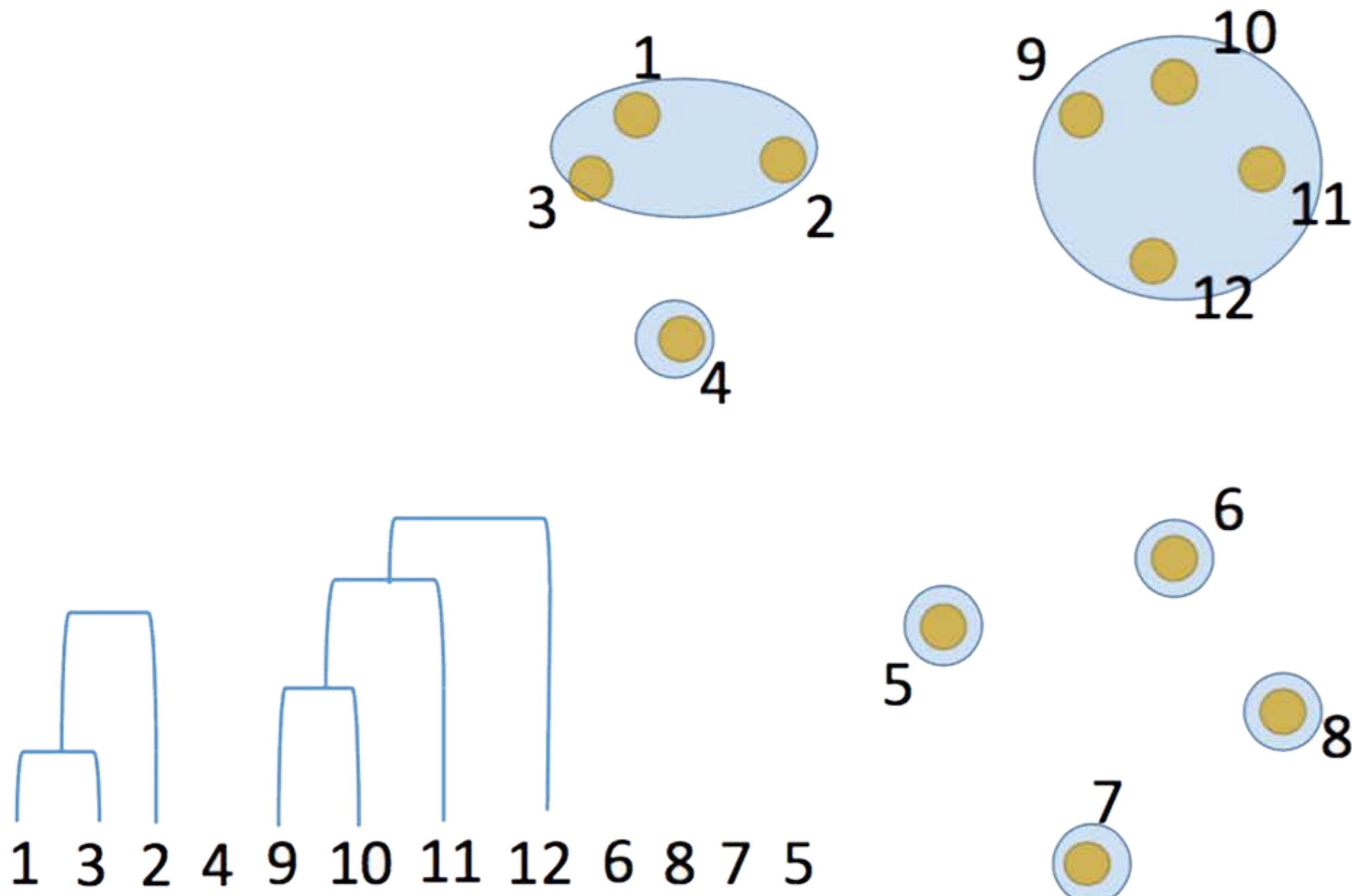
ДЕНДРОГРАММА



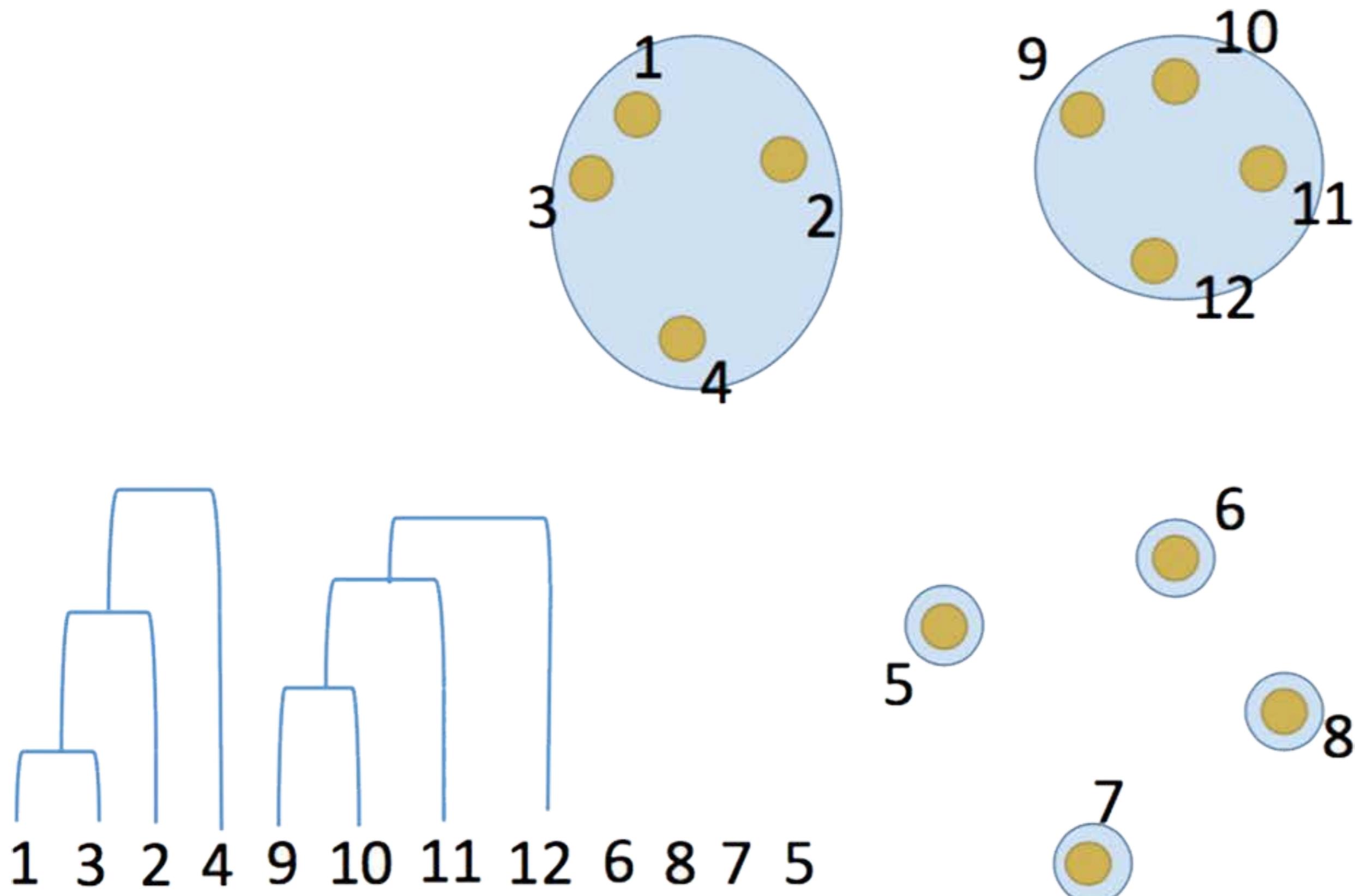
ДЕНДРОГРАММА



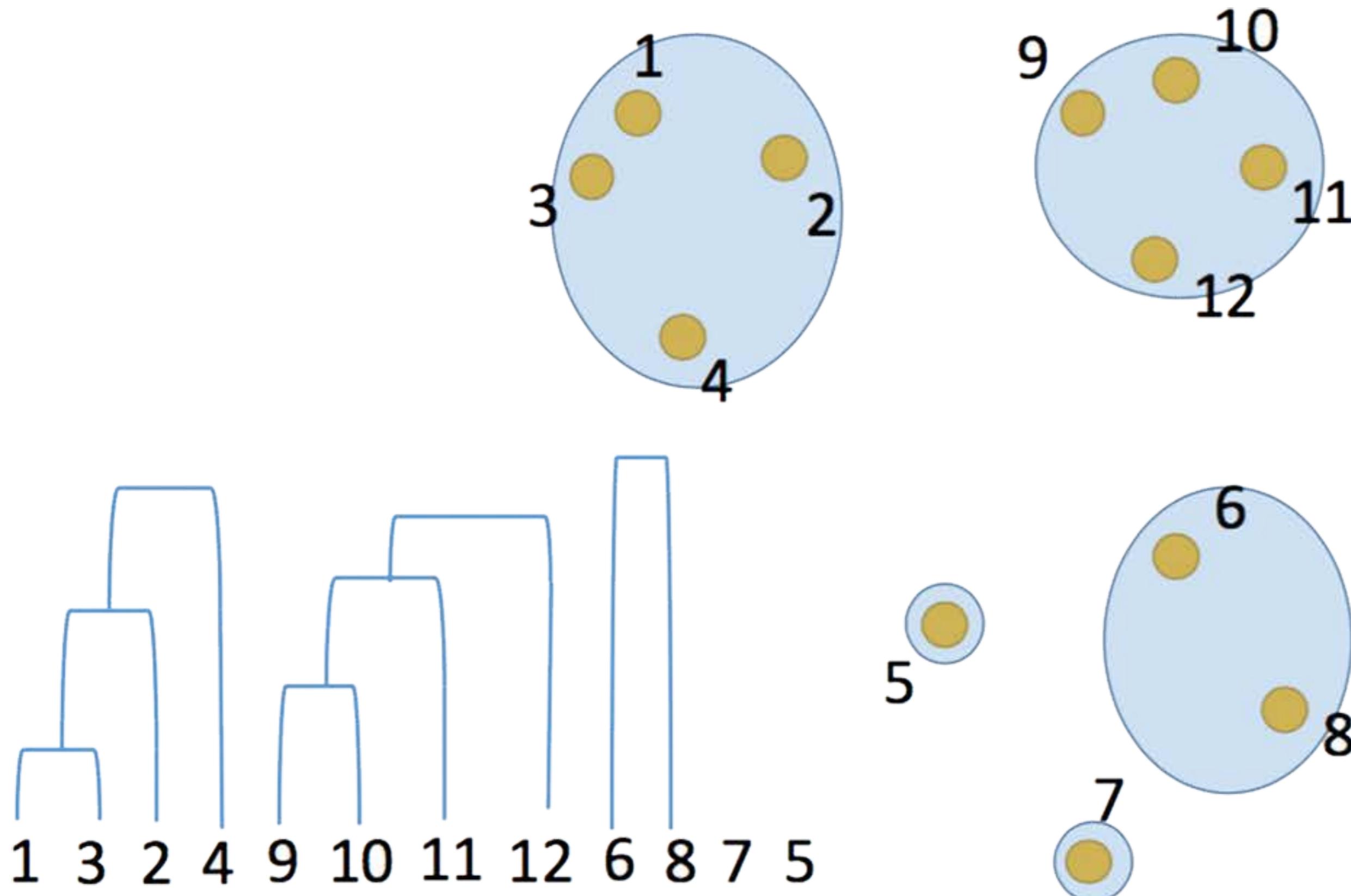
ДЕНДРОГРАММА



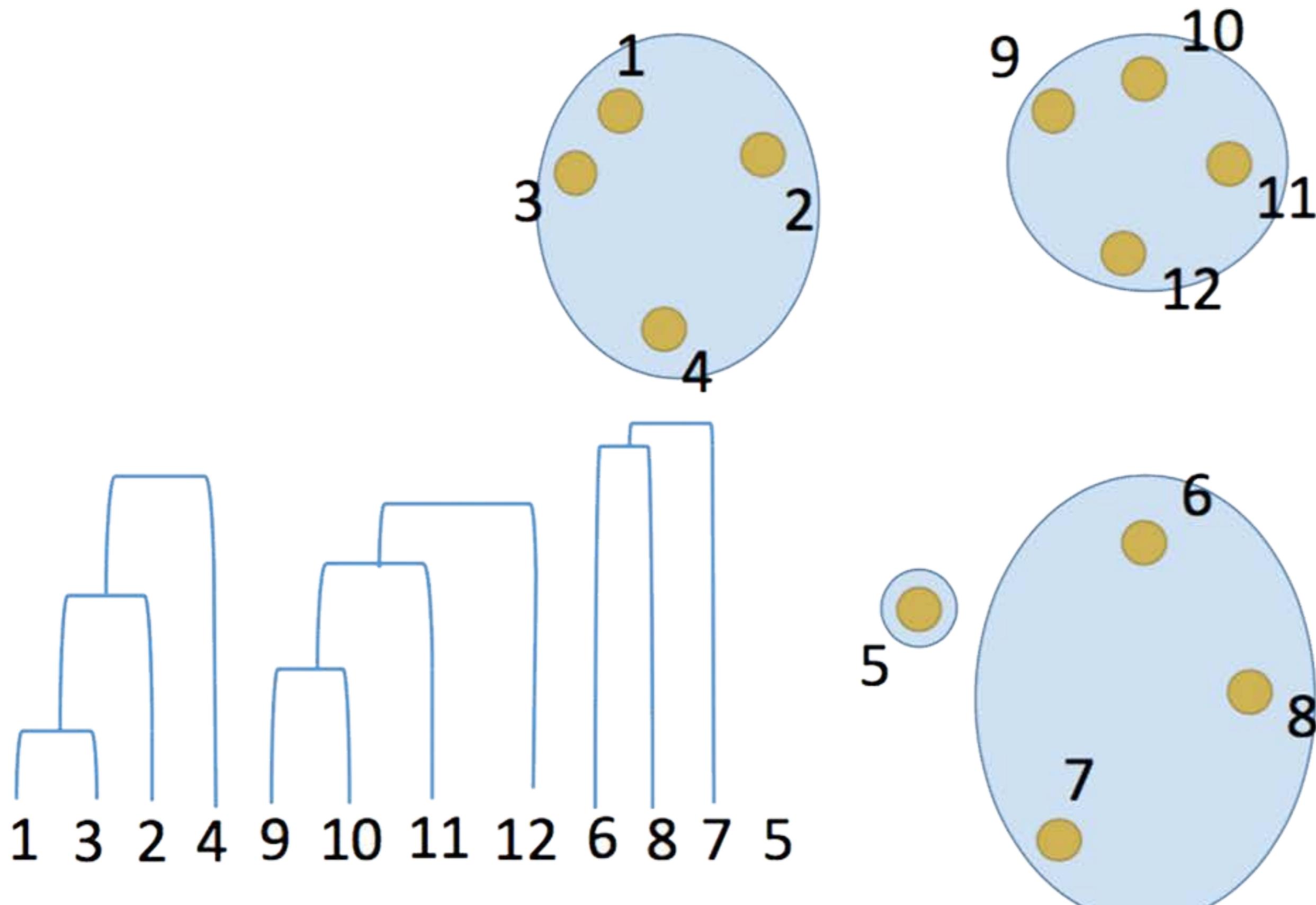
ДЕНДРОГРАММА



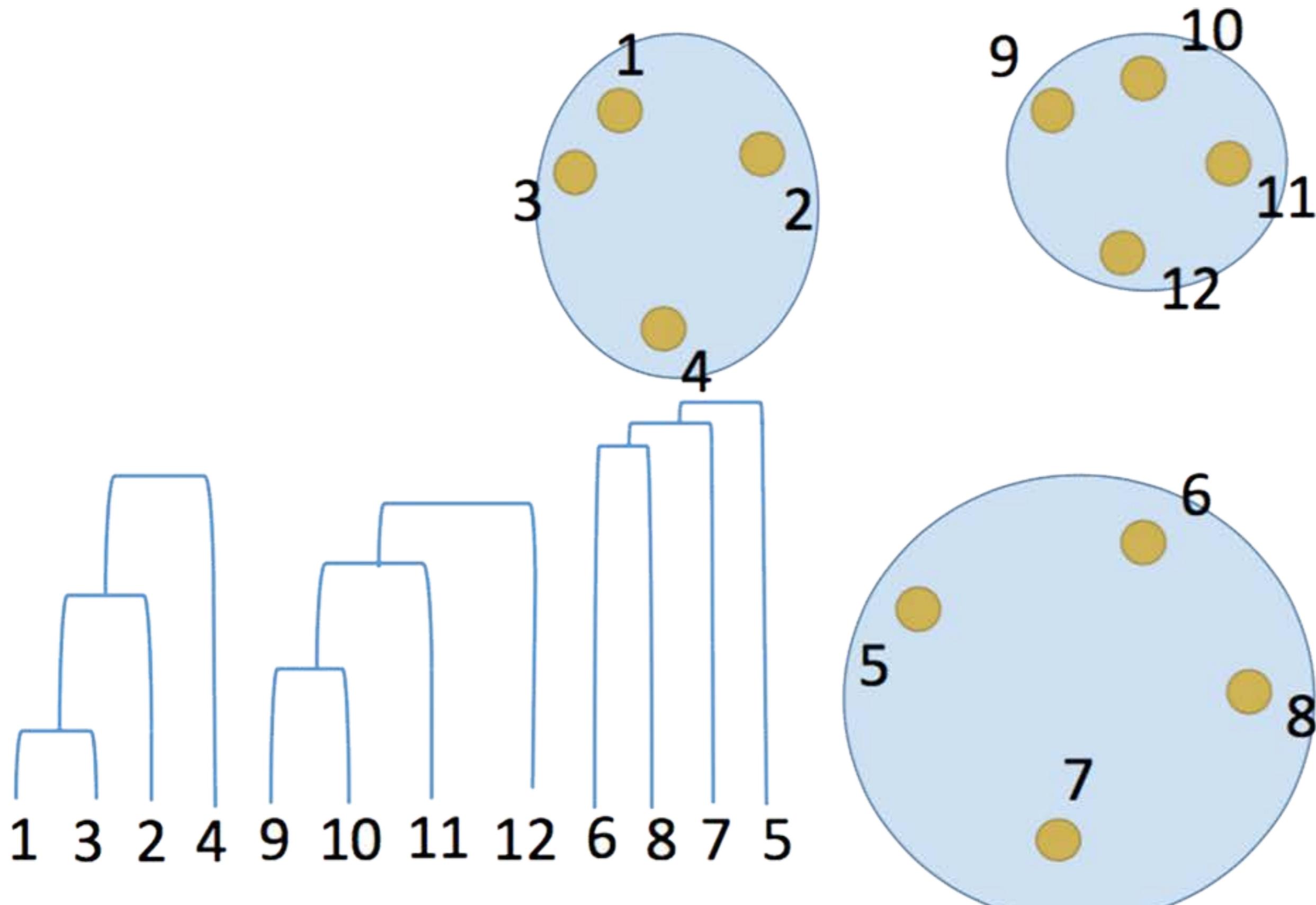
ДЕНДРОГРАММА



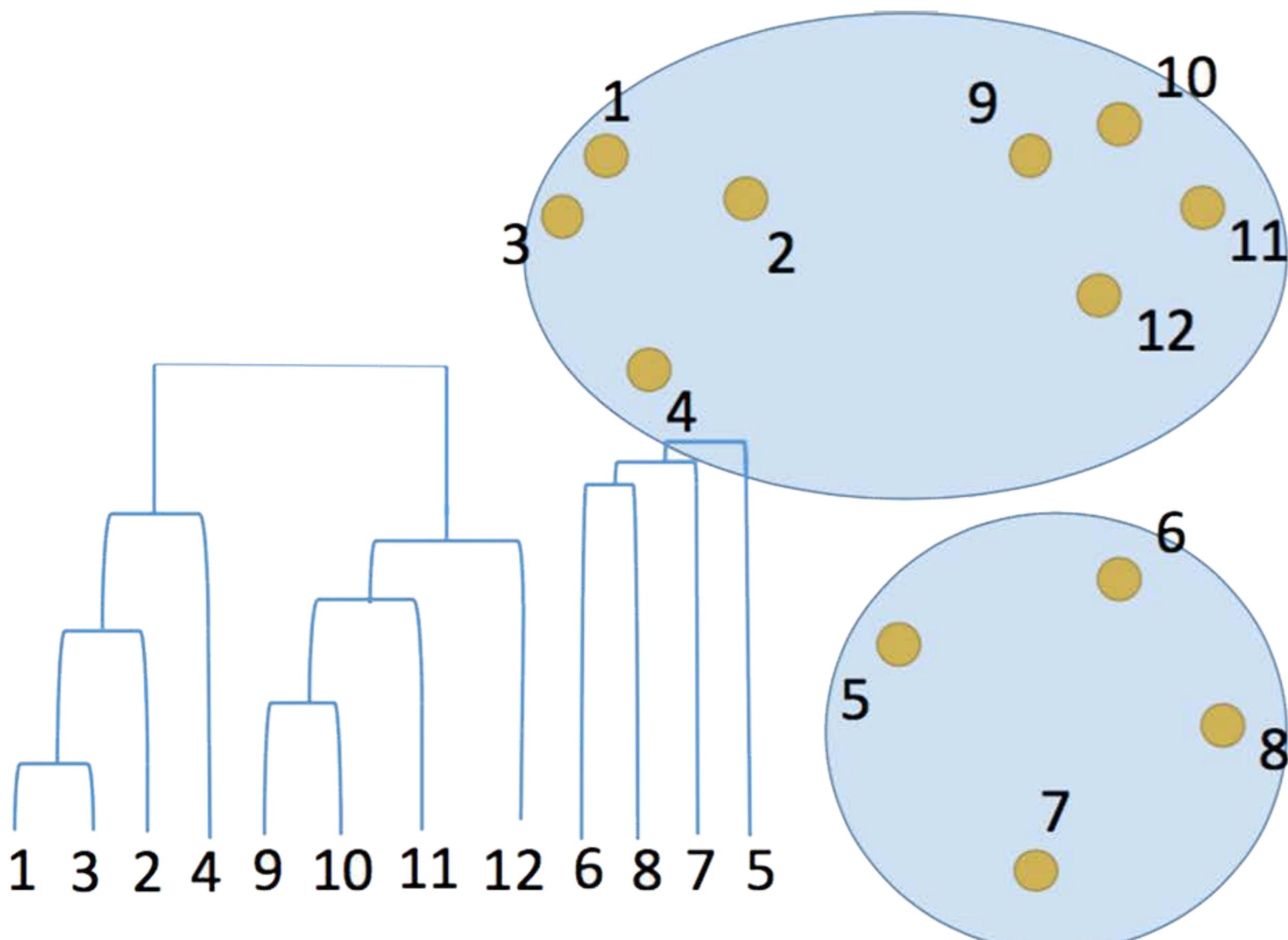
ДЕНДРОГРАММА



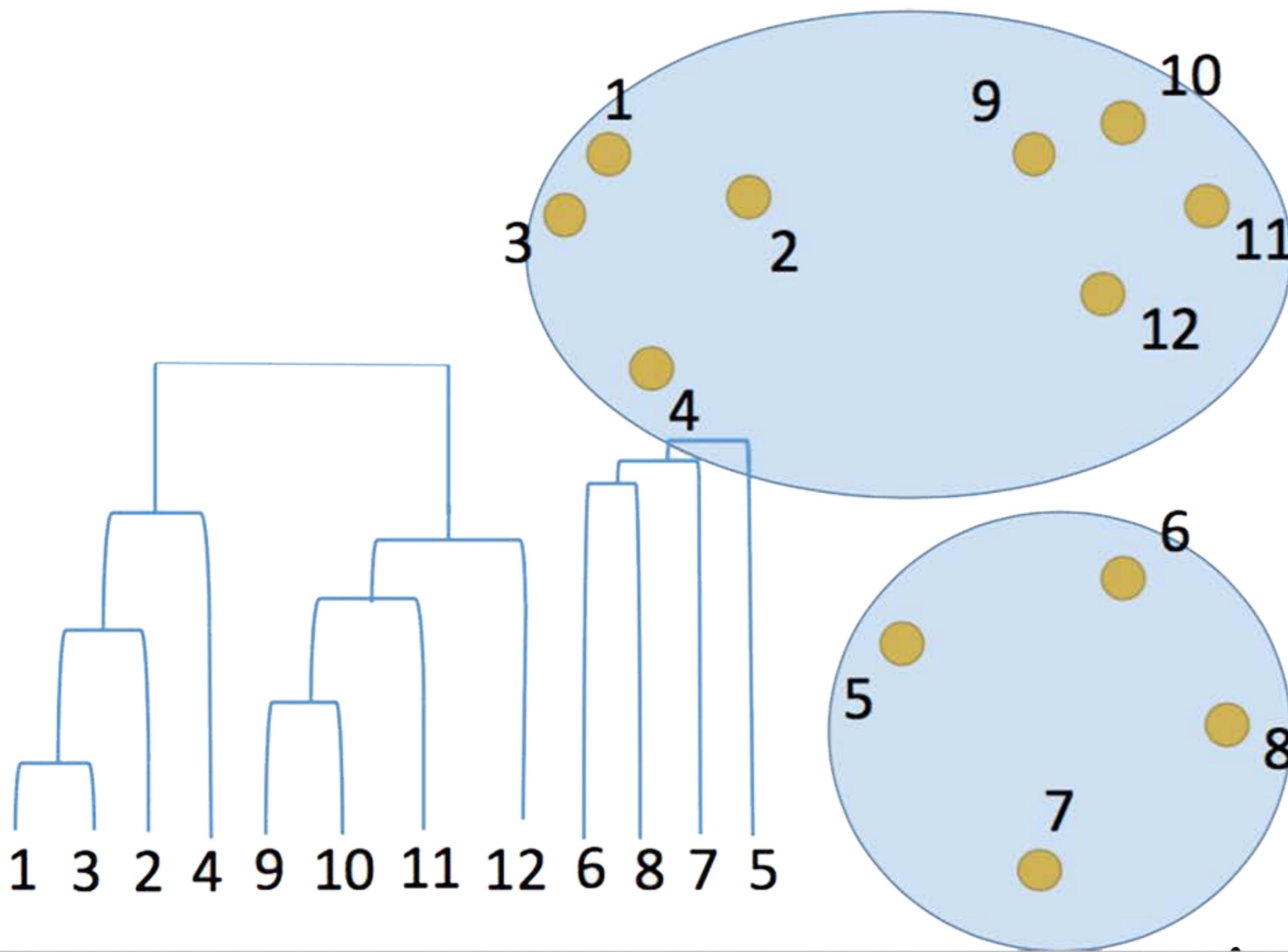
ДЕНДРОГРАММА



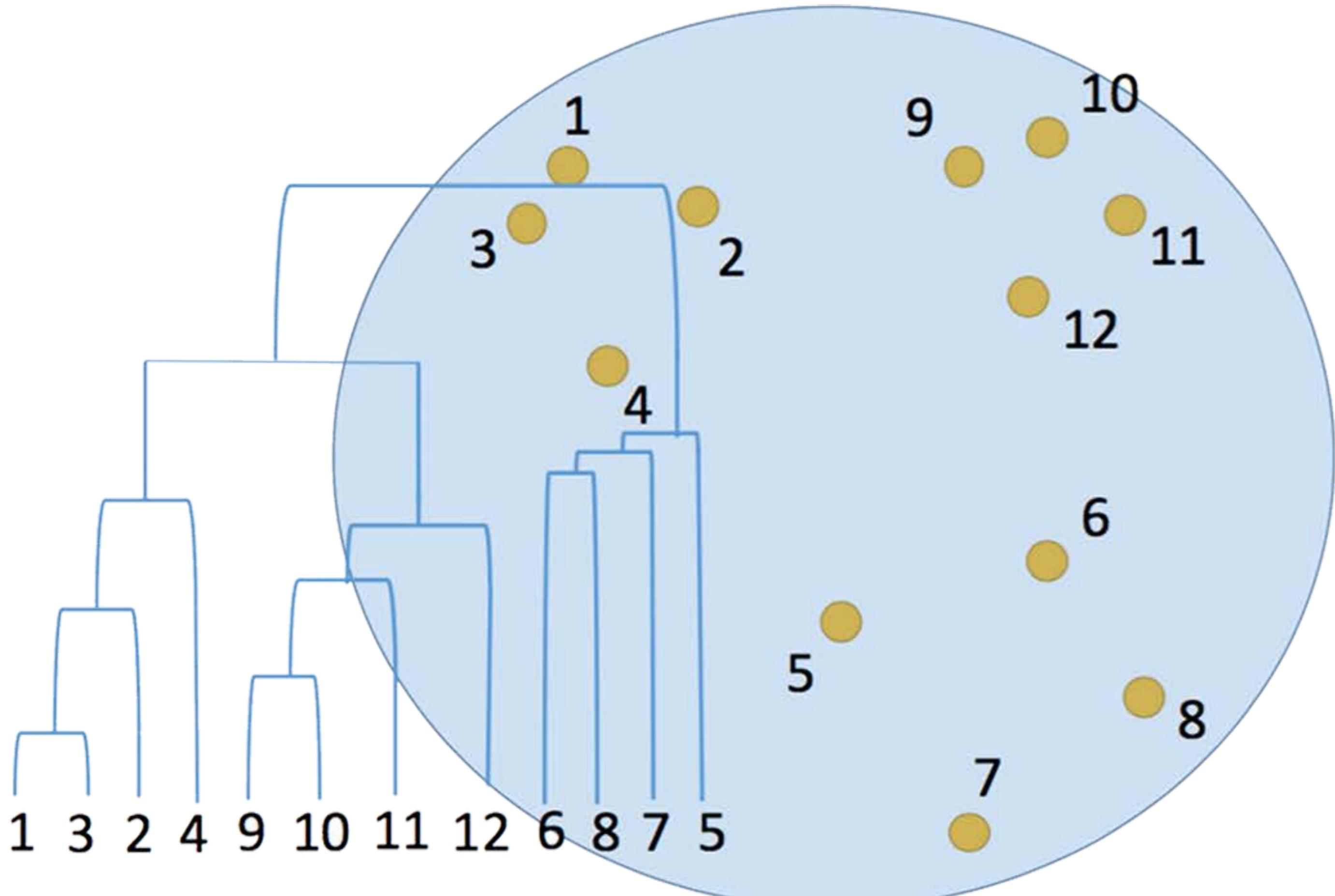
ДЕНДРОГРАММА



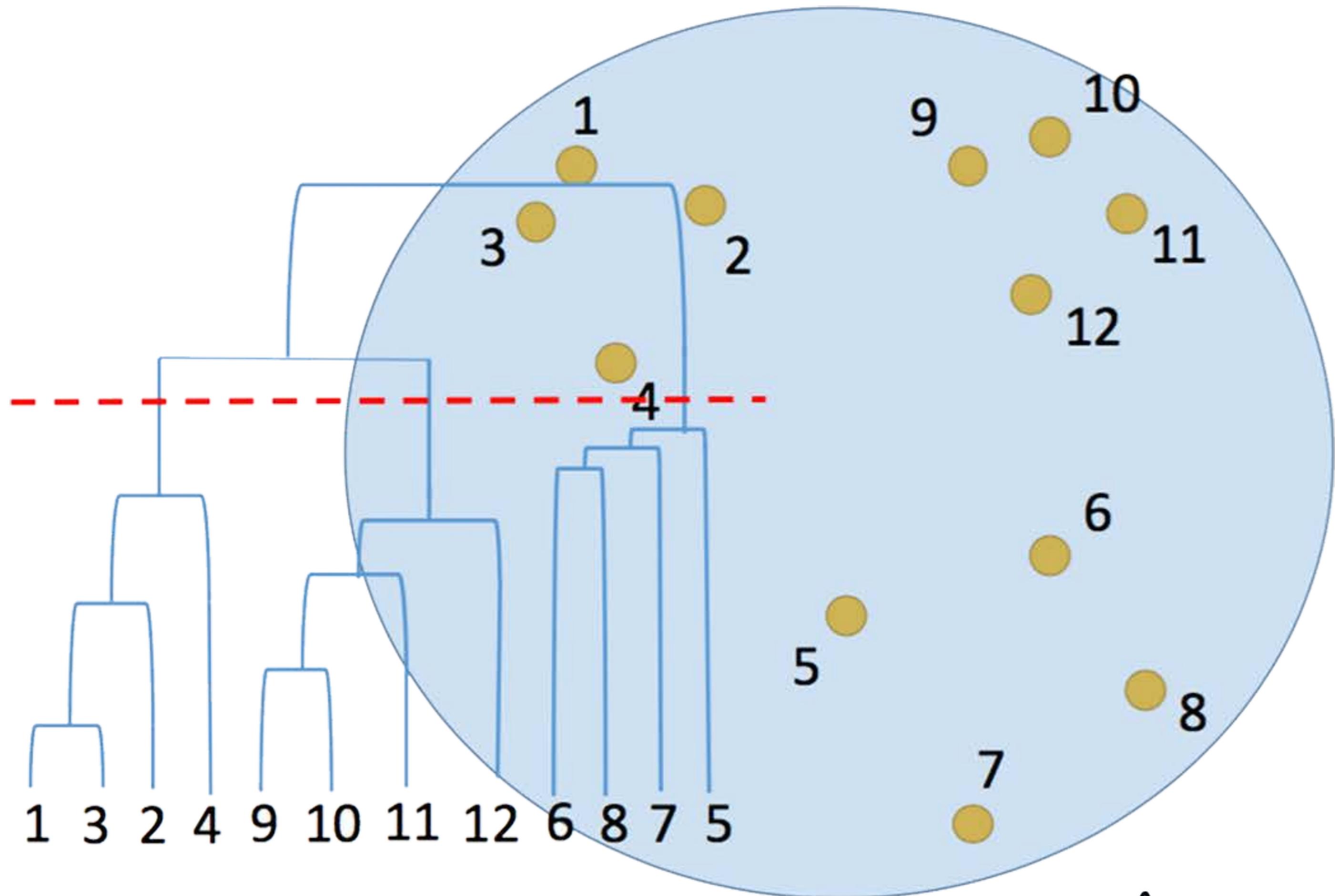
ДЕНДРОГРАММА



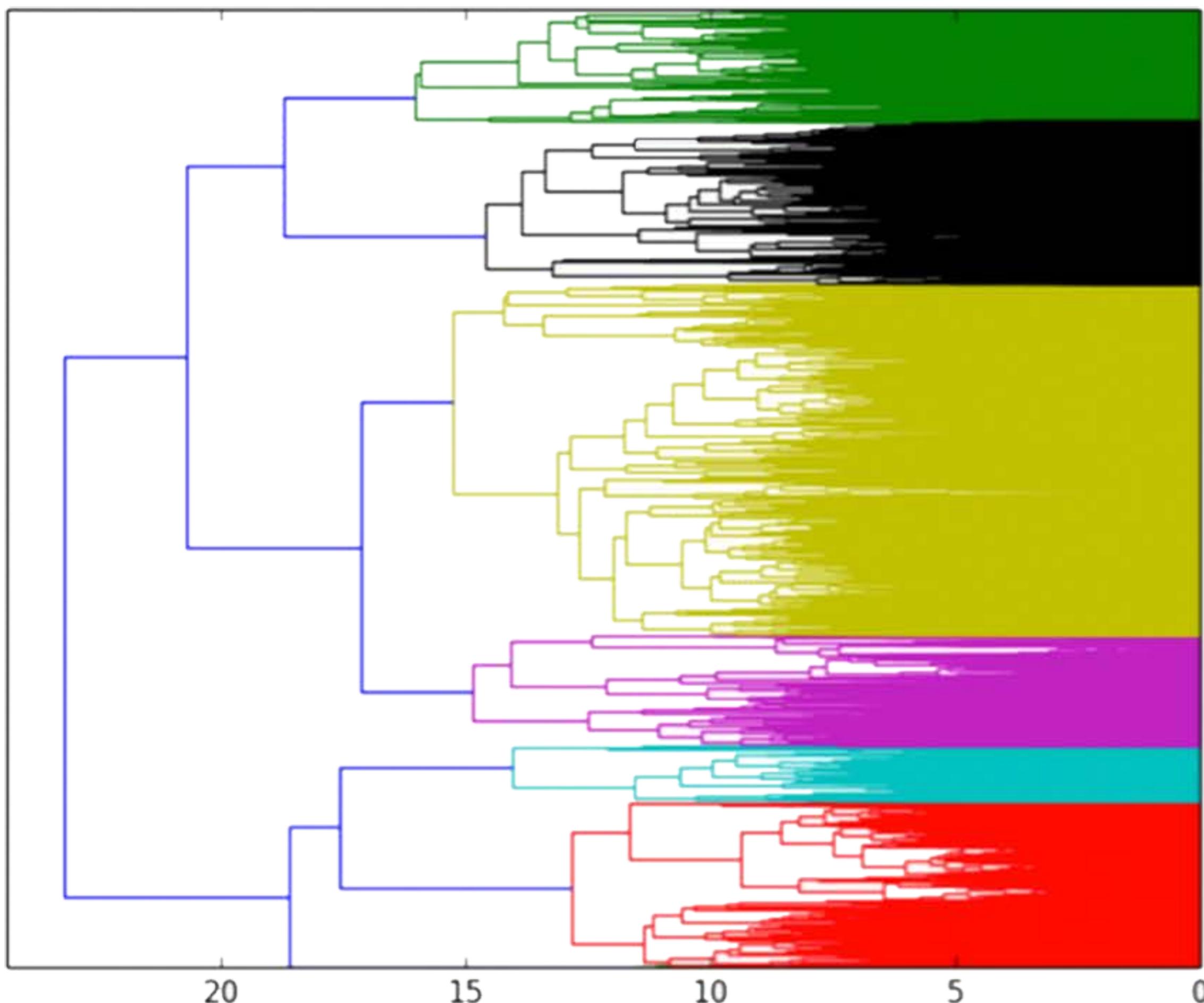
ДЕНДРОГРАММА



ДЕНДРОГРАММА

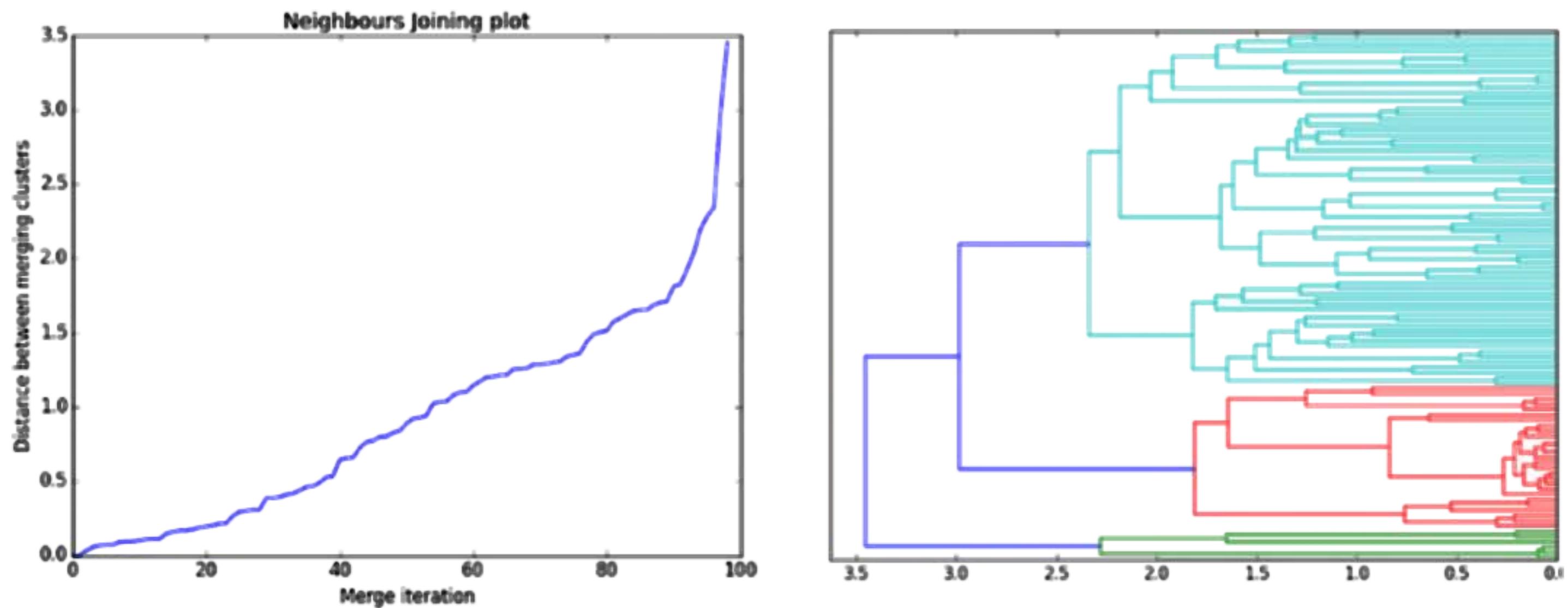


ПРИМЕР ДЕНДРОГРАММЫ



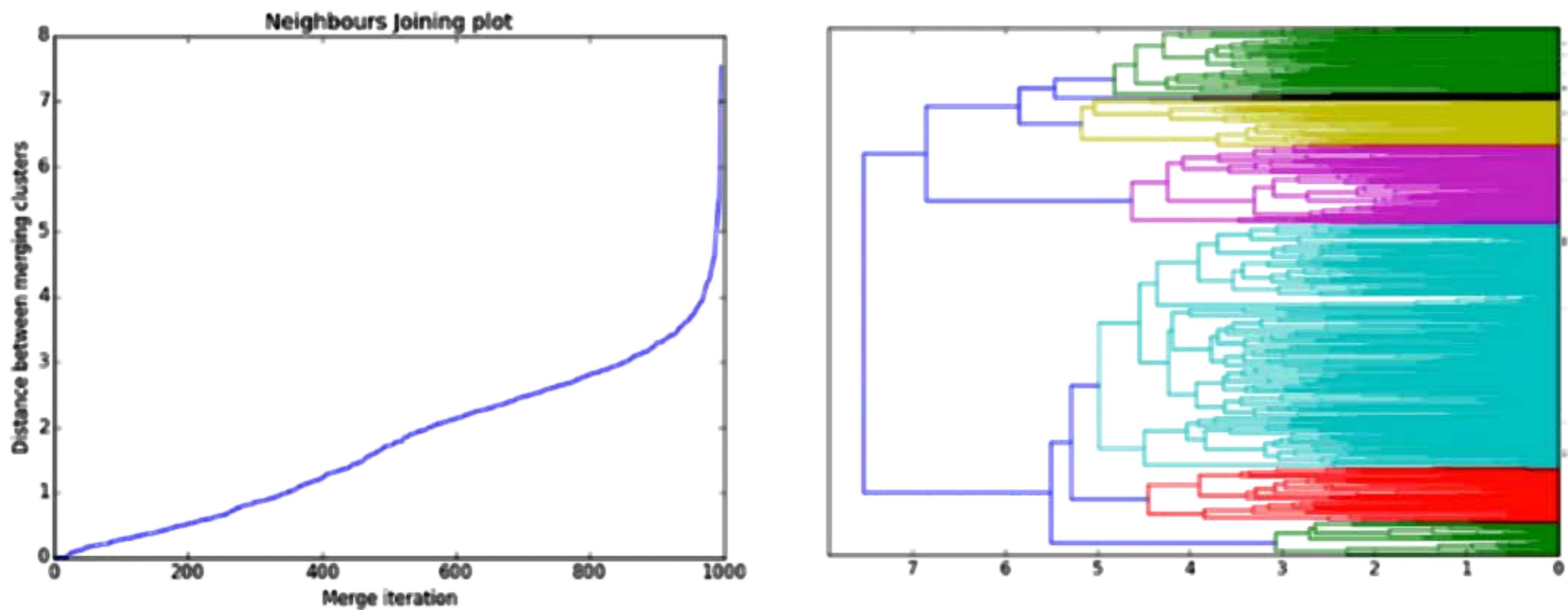
ПРИМЕР: РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

- На подвыборке из 100 писем



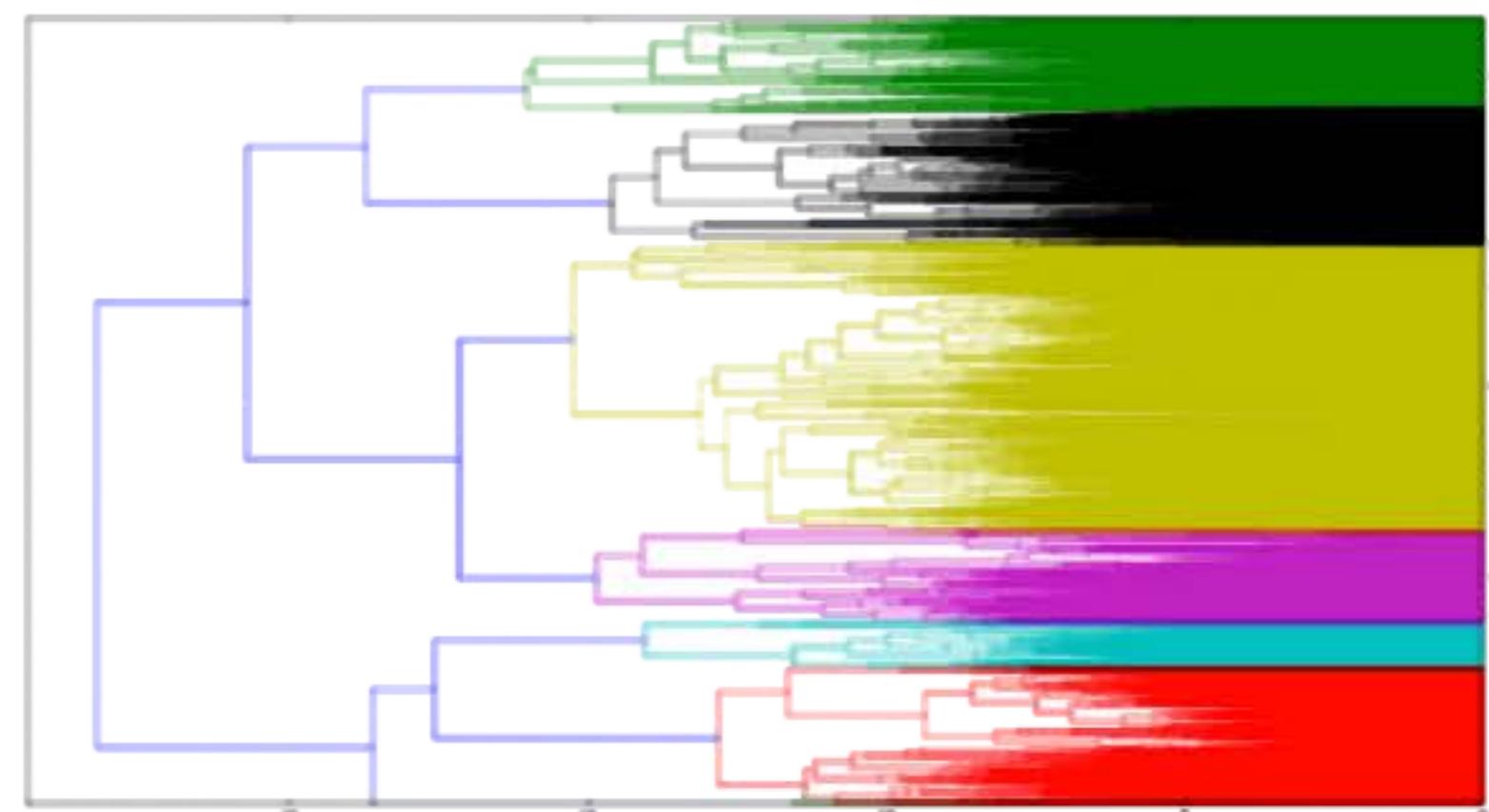
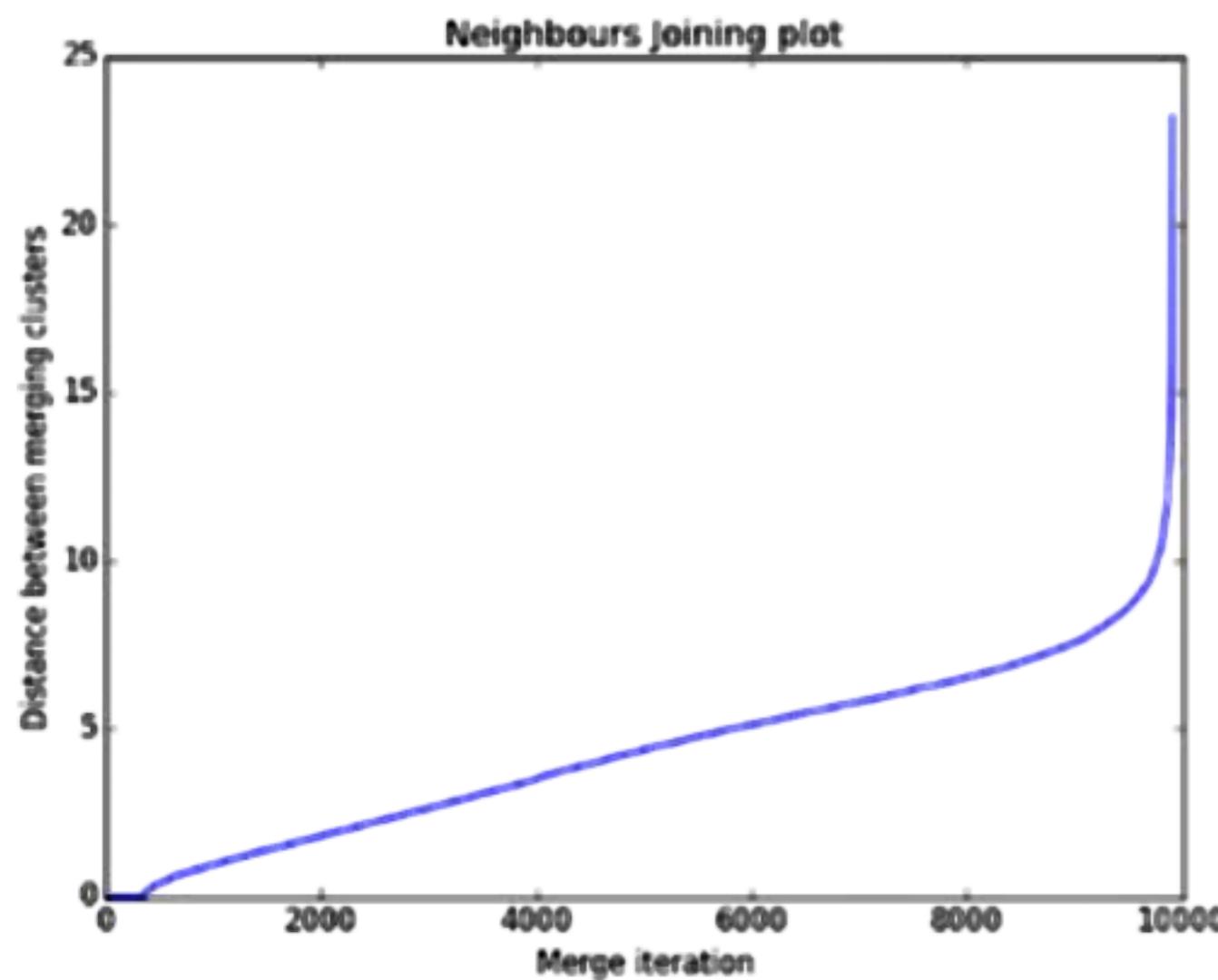
ПРИМЕР: РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

- На подвыборке из 1000 писем



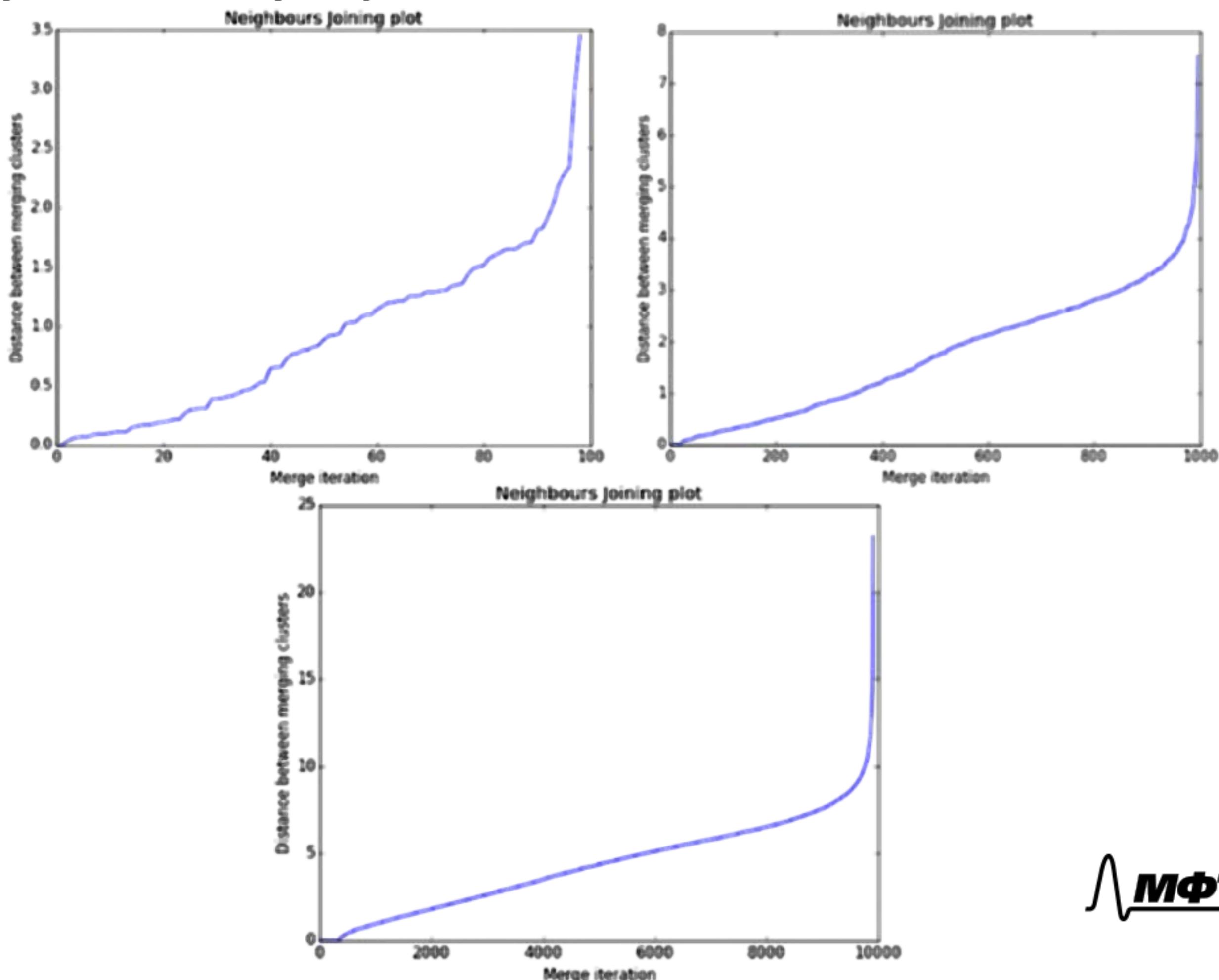
ПРИМЕР: РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

- На подвыборке из 10000 писем



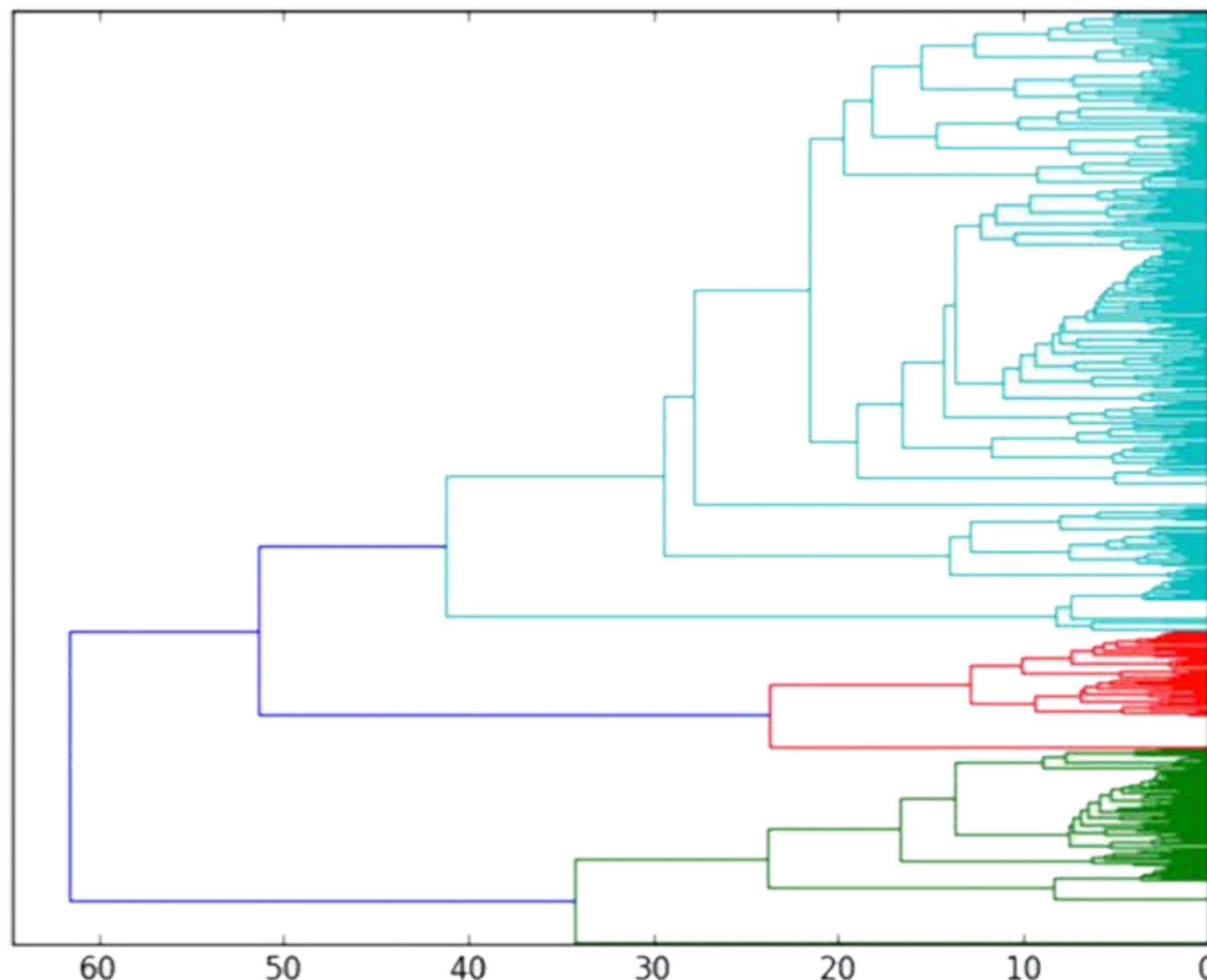
ПРИМЕР: РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

➤ Сравним графики: 100, 1000, 10000 писем

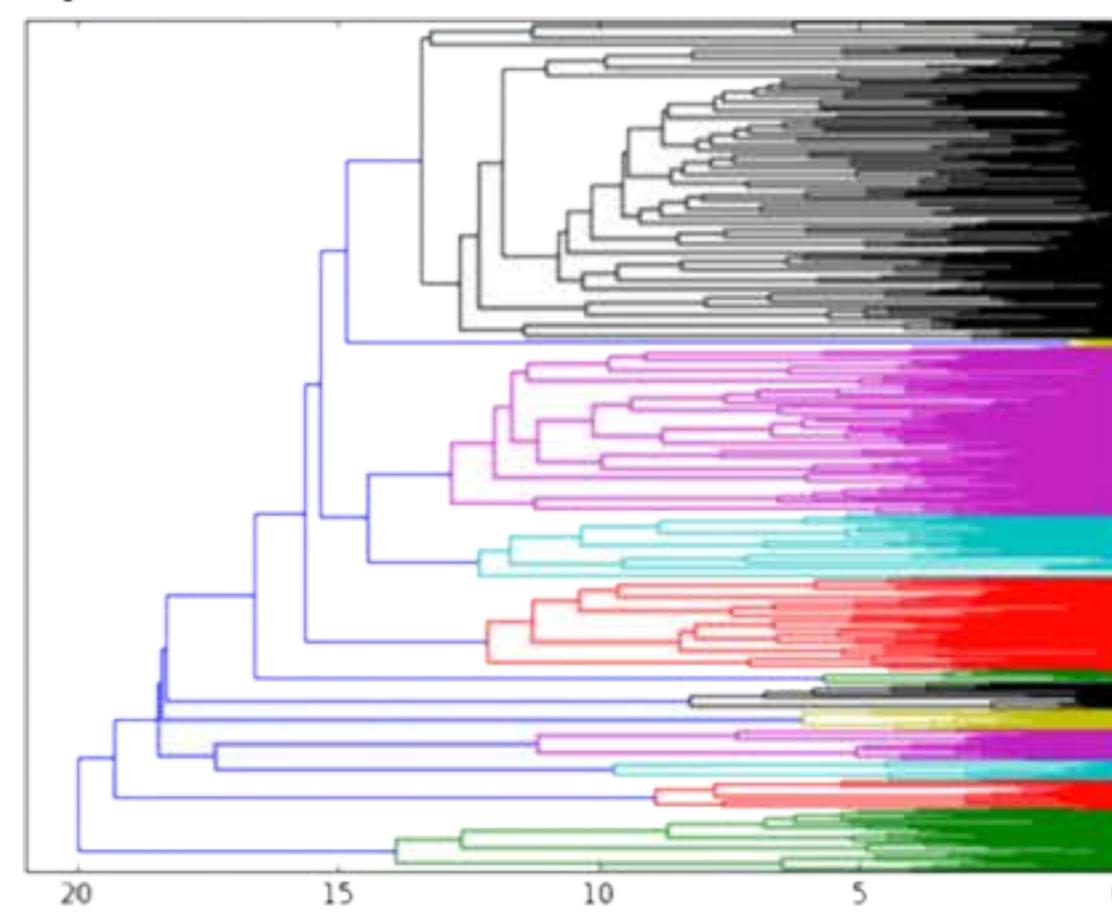
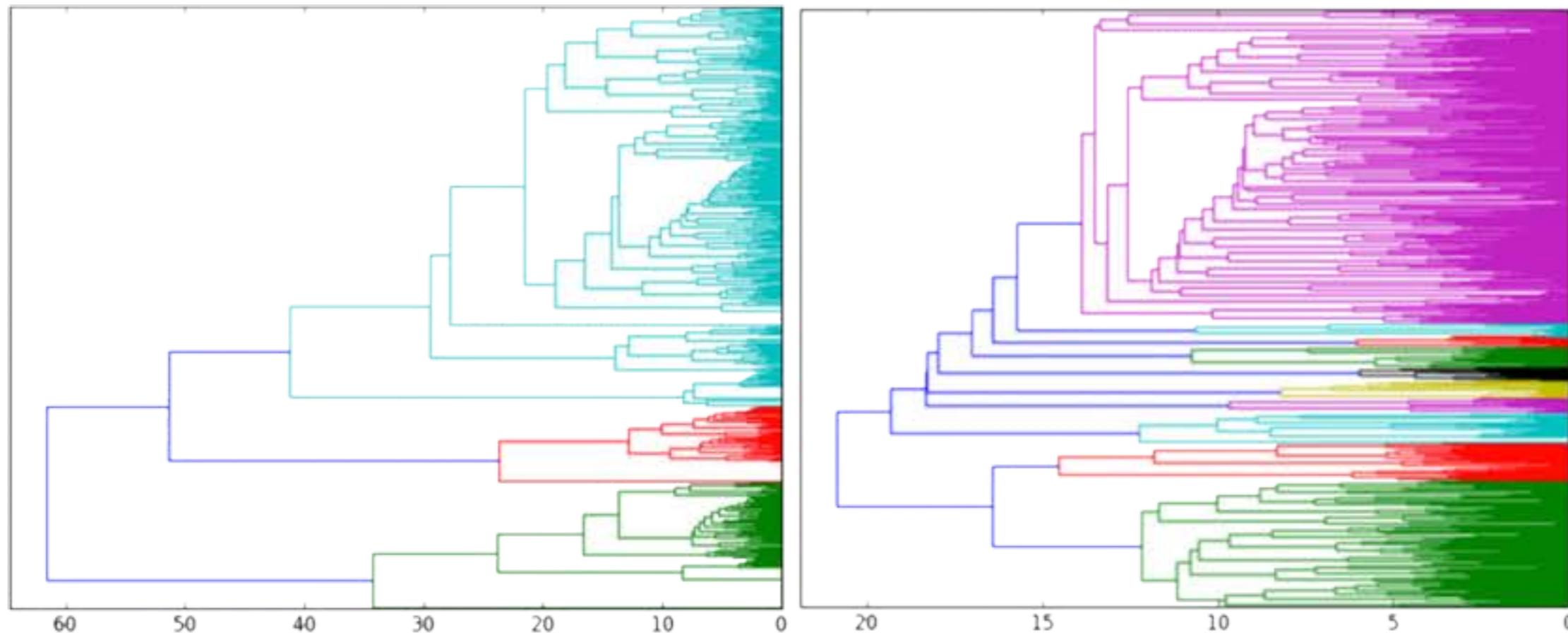


ПРИМЕР: ПЕРЕКОС В РАЗМЕРАХ КЛАСТЕРОВ

- › Дендрограмма, построенная для другой выборки текстов:

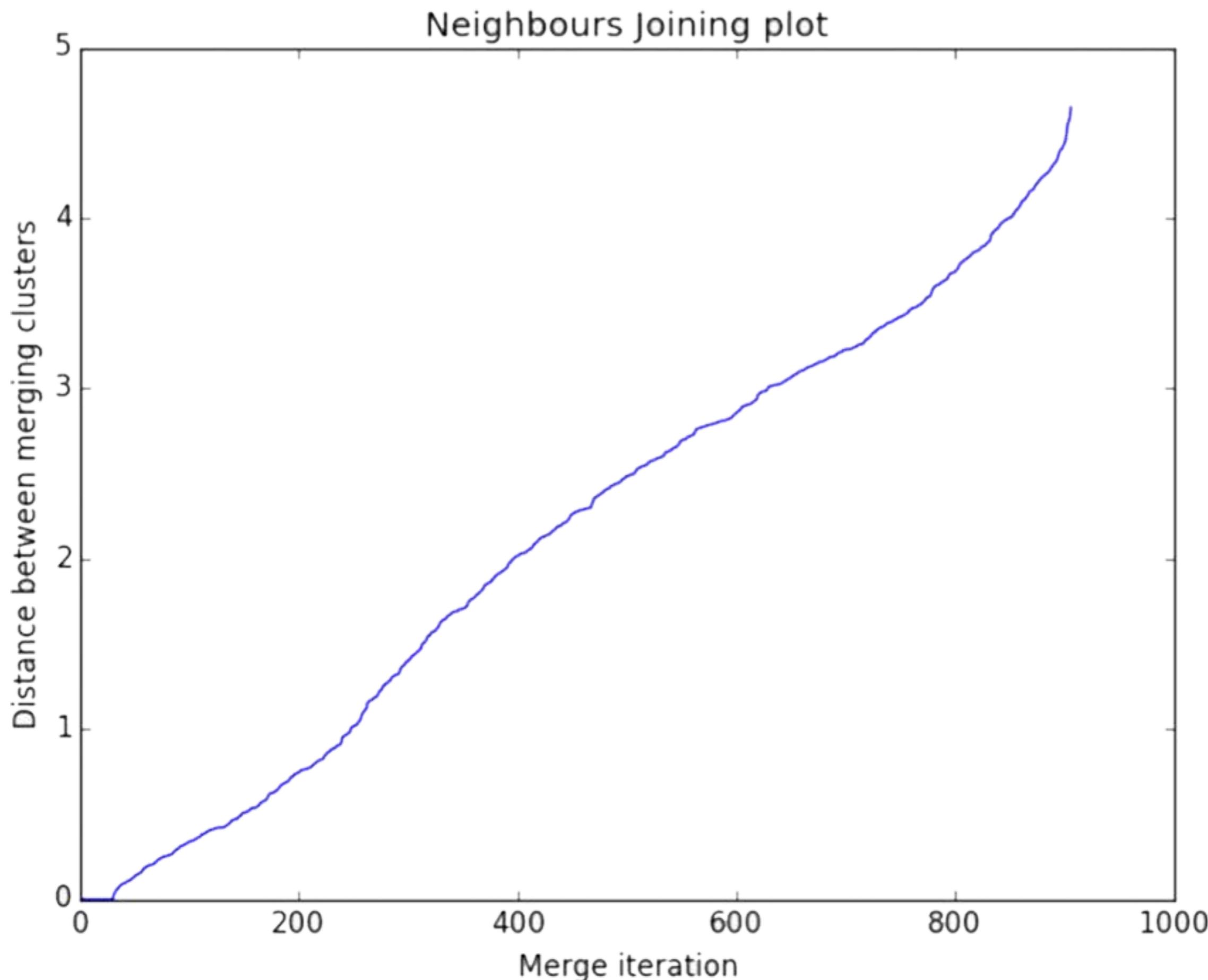


ПРИМЕР: ДОБАВЛЯЕМ SVD



SVD (еще меньше компонент)

ПРИМЕР: SVD И РАССТОЯНИЕ ПРИ СЛИЯНИИ



РЕЗЮМЕ

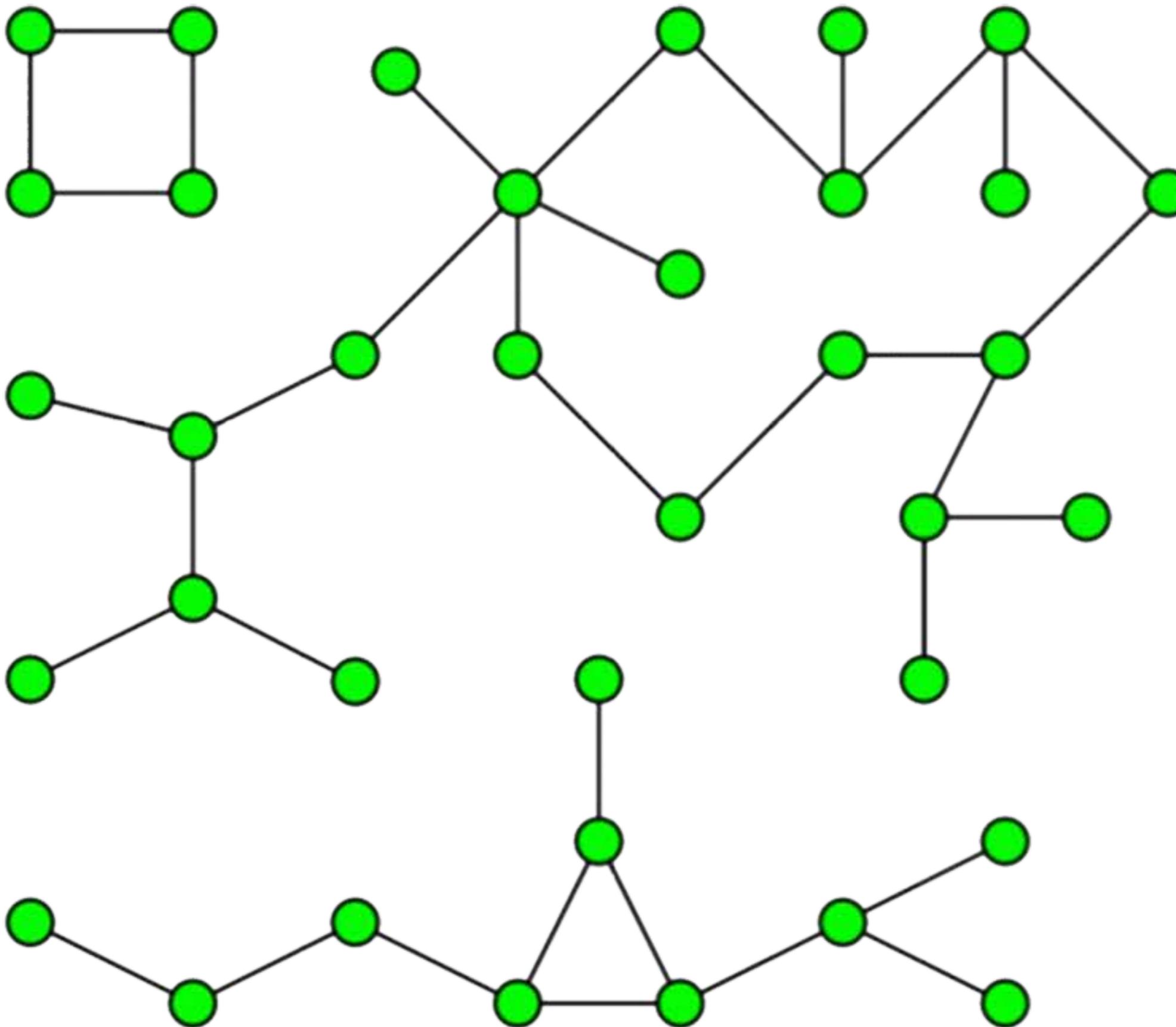
- › Иерархическая кластеризация
- › Как устроена агломеративная кластеризация
- › Расстояние между кластерами
- › Формула Ланса-Уильямса
- › Дендрограммы
- › Примеры работы

ПРОСТЫЕ ГРАФОВЫЕ МЕТОДЫ

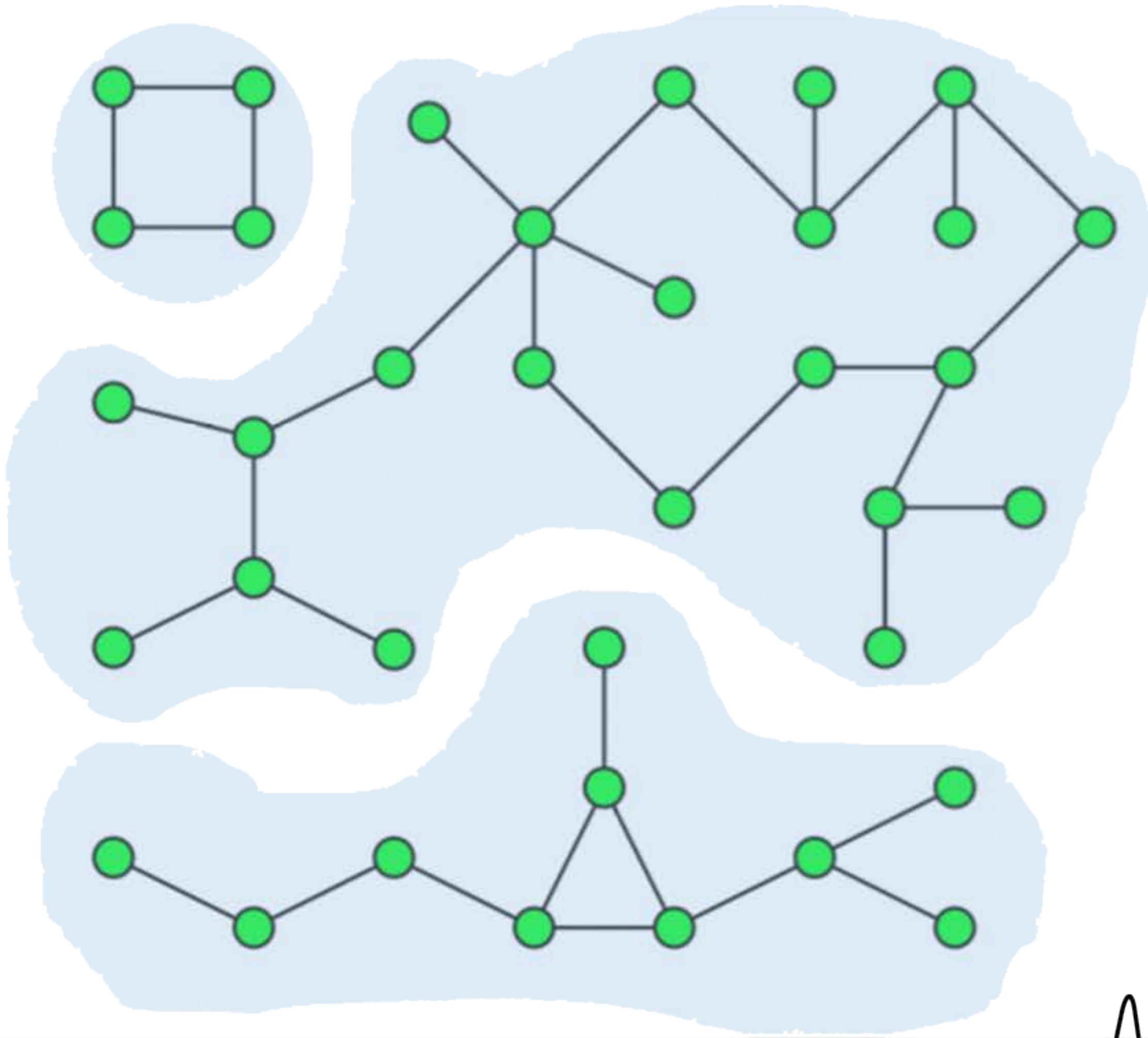
ПЛАН

- › Связные компоненты
- › Кластеризация с помощью выделения связных компонент
- › Минимальное оствовное дерево
- › Алгоритм Крускала
- › Кластеризация с помощью минимального оствового дерева

ВЫДЕЛЕНИЕ СВЯЗНЫХ КОМПОНЕНТ



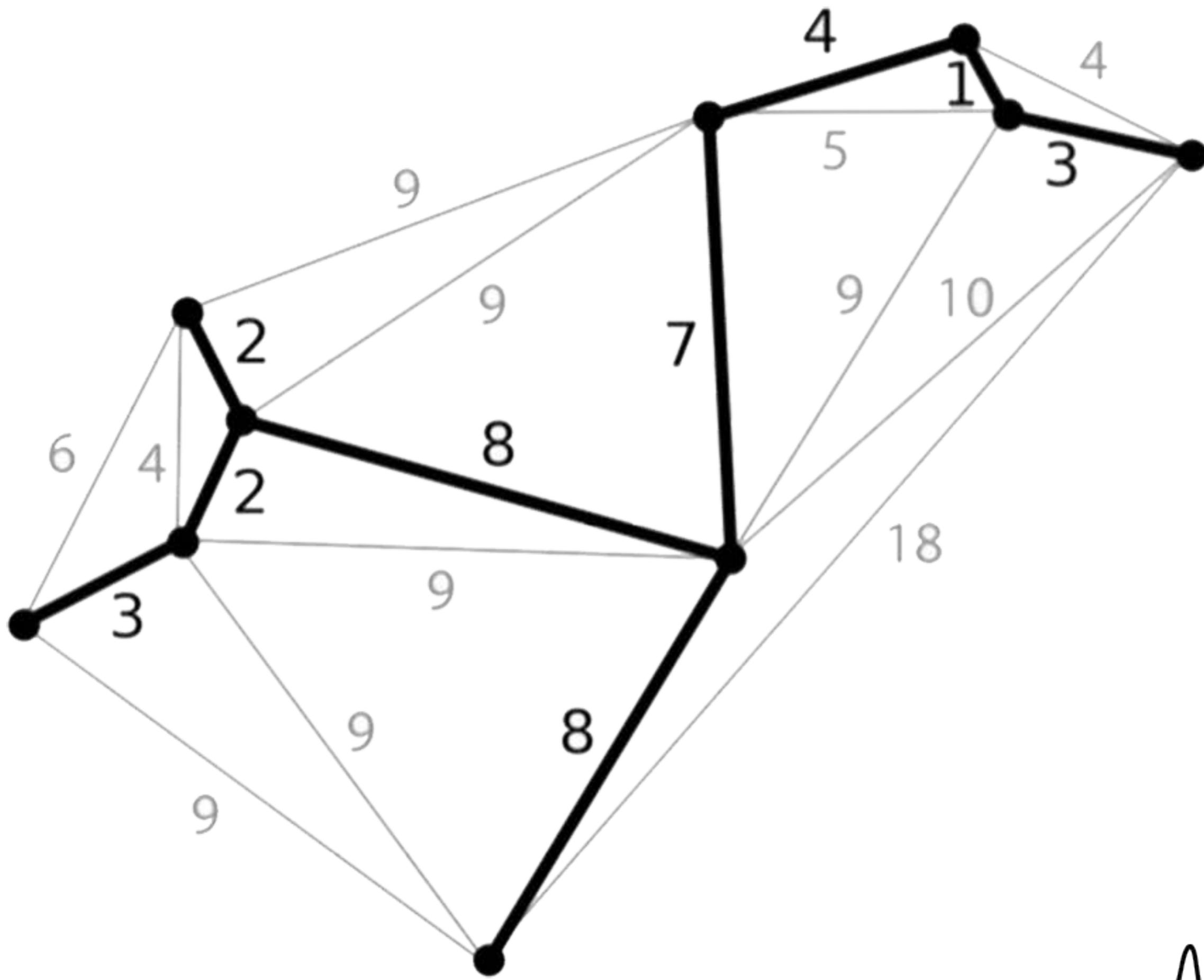
ВЫДЕЛЕНИЕ СВЯЗНЫХ КОМПОНЕНТ



КЛАСТЕРИЗАЦИЯ ПО КОМПОНЕНТАМ СВЯЗНОСТИ

- Соединяем ребром объекты, расстояние между которыми меньше R
- Выделяем компоненты связности
- Проблема: непонятно, как выбрать R , если нужно получить K кластеров

МИНИМАЛЬНОЕ ОСТОВНОЕ ДЕРЕВО



МИНИМАЛЬНОЕ ОСТОВНОЕ ДЕРЕВО

➤ Алгоритм Крускала (Kruskal):

1. Изначально множество уже найденных ребер пустое
2. На первом шаге добавляем ребро с минимальным весом
3. На каждом шаге добавляем ребро, одна из вершин которого лежит в множестве выбранных вершин, а другая – нет, при этом среди всех таких ребер выбираем ребро с наименьшим весом
4. В тот момент, когда задействованы все вершины графа – выбранные ребра образуют минимальное оставное дерево

КЛАСТЕРИЗАЦИЯ С ПОМОЩЬЮ МИНИМАЛЬНОГО ОСТОВНОГО ДЕРЕВА

- › Строим взвешенный граф, где веса рёбер — расстояния между объектами
- › Строим минимальное оствовное дерево для этого графа
- › Удаляем $K - 1$ ребро с максимальным весом
- › Получаем K компонент связности, которые интерпретируем как кластеры

РЕЗЮМЕ

- Связные компоненты
- Кластеризация с помощью выделения связных компонент
- Минимальное остовное дерево
- Алгоритм Крускала
- Кластеризация с помощью минимального остовного дерева

КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ ПЛОТНОСТИ ТОЧЕК (DENSITY BASED CLUSTERING)

ПЛАН

- › Идея методов на основе плотности точек
- › Пример основных, граничных и шумовых точек
- › DBSCAN
- › Пример работы DBSCAN
- › Определение числа кластеров
- › Настройка параметров DBSCAN

ИДЕЯ

DENSITY-BASED МЕТОДОВ

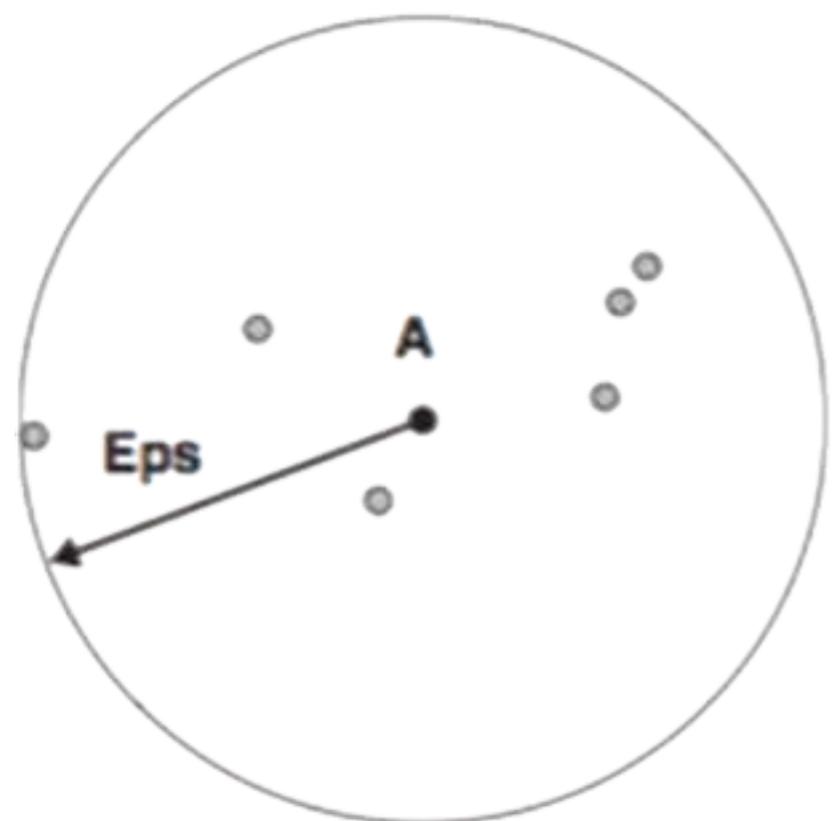


Figure 8.20. Center-based density.

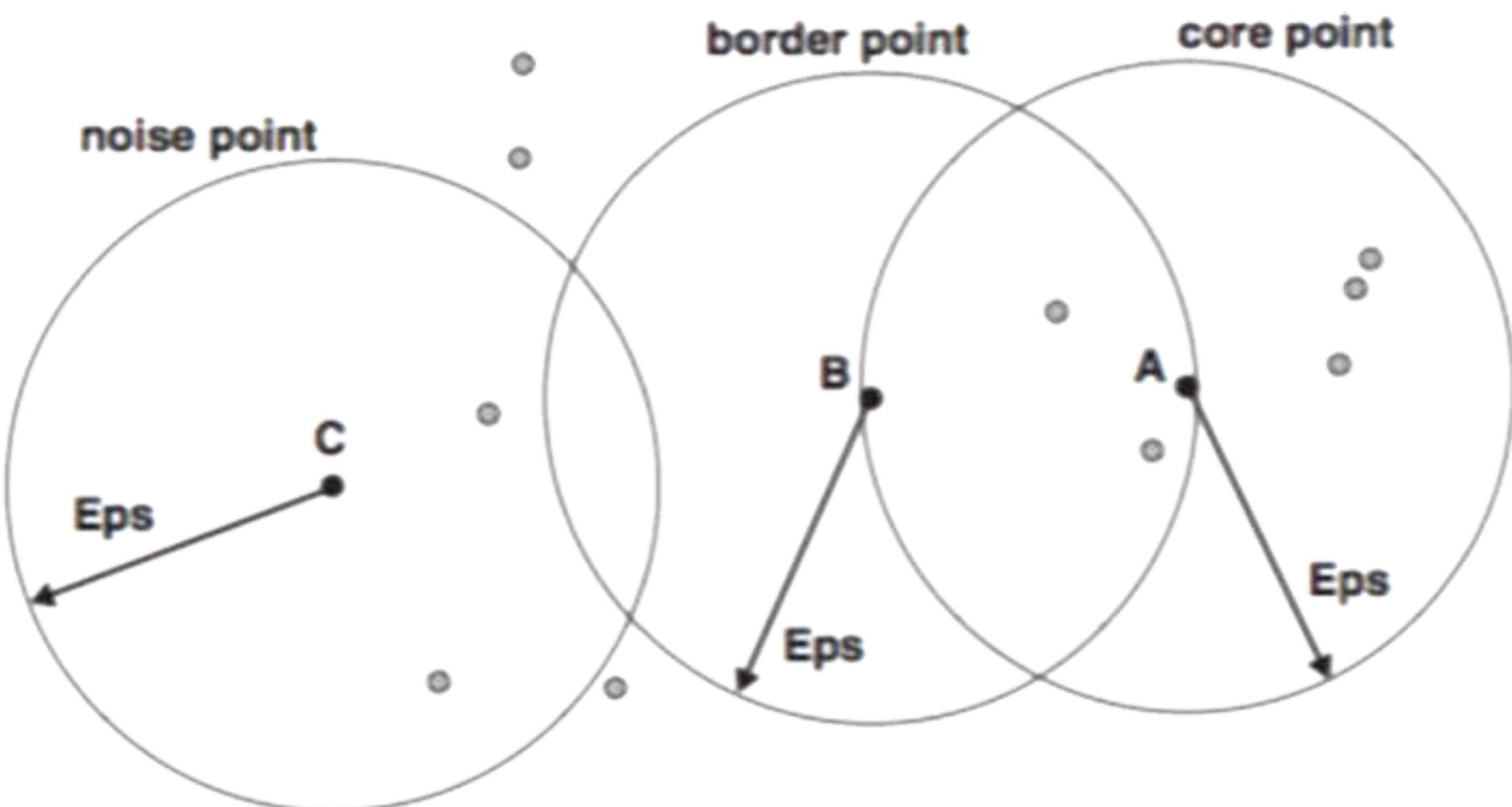
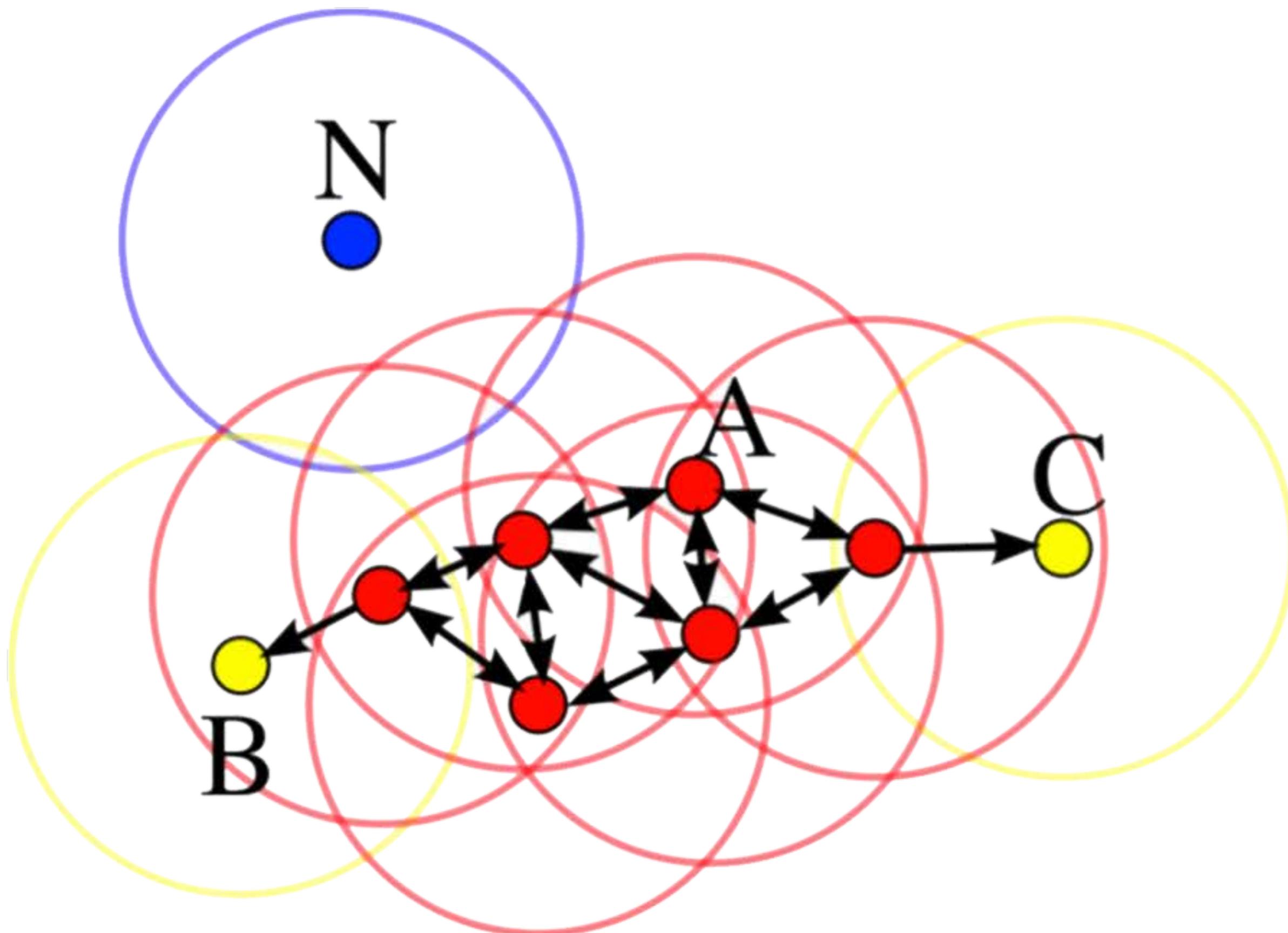


Figure 8.21. Core, border, and noise points.

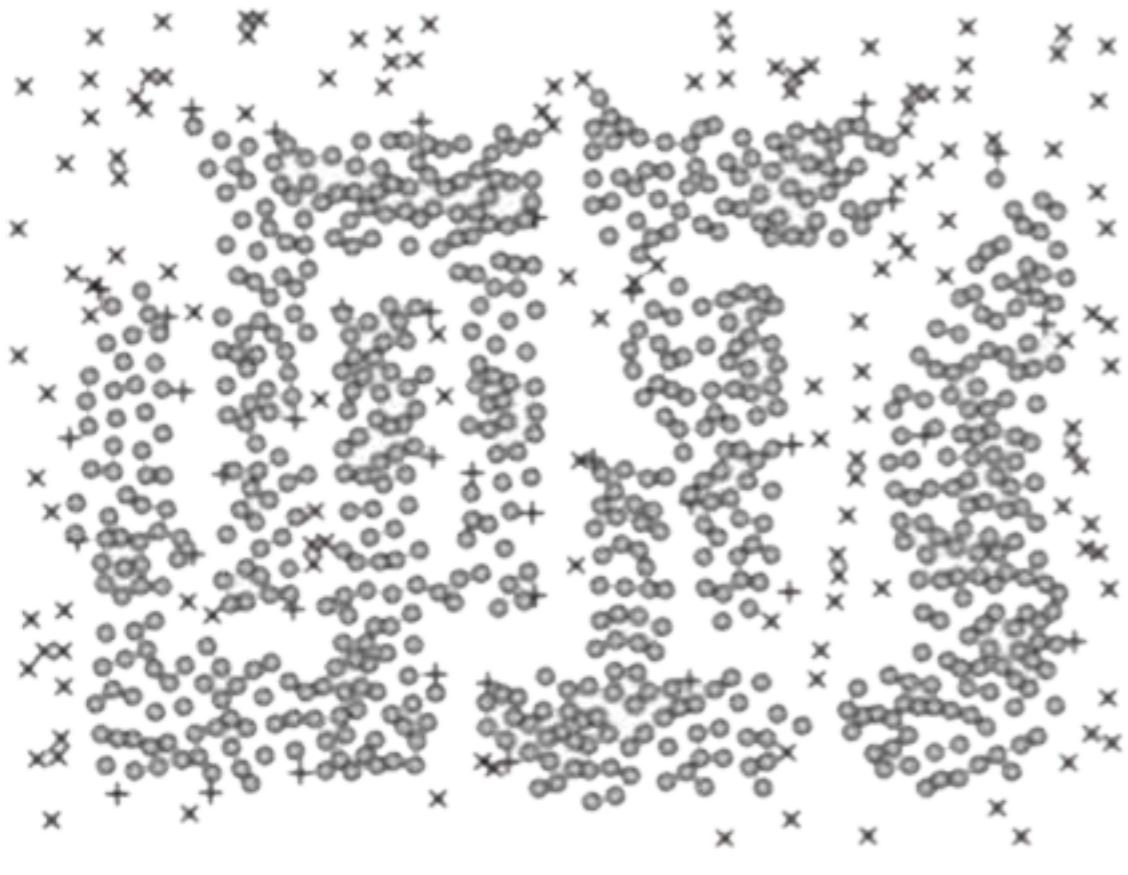
ОСНОВНЫЕ, ШУМОВЫЕ И ГРАНИЧНЫЕ ТОЧКИ



DBSCAN



(a) Clusters found by DBSCAN.



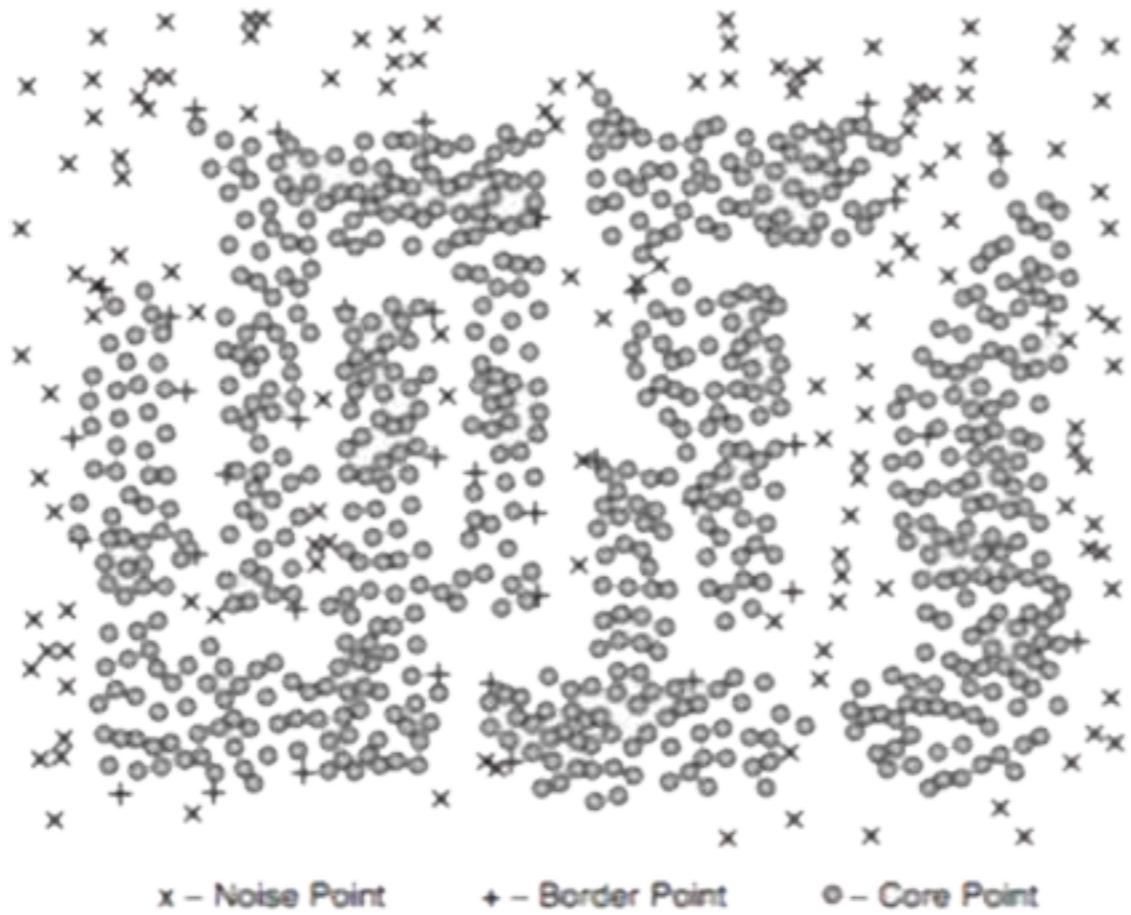
x – Noise Point + – Border Point ○ – Core Point

(b) Core, border, and noise points.

DBSCAN



(a) Clusters found by DBSCAN.



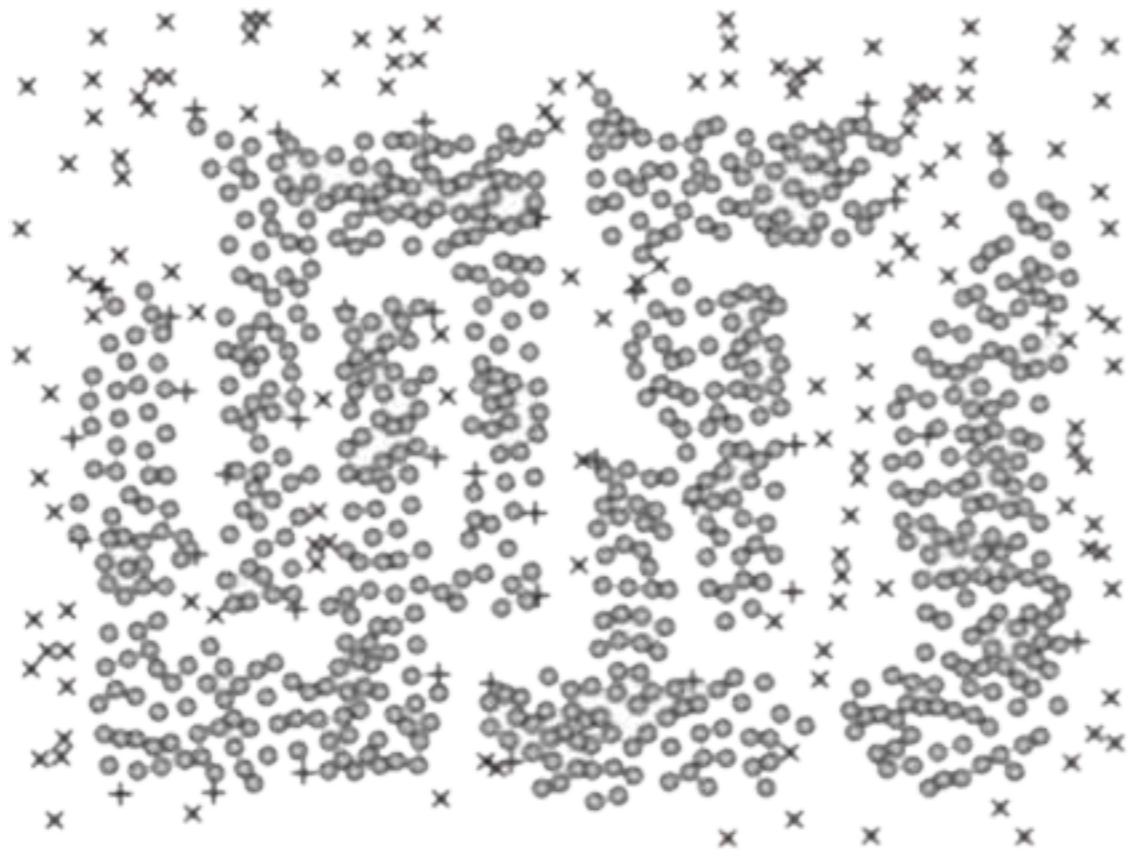
(b) Core, border, and noise points.

- 1: Пометить все точки, как основные, пограничные или шумовые.
- 2: Отбросить точки шума.

DBSCAN



(a) Clusters found by DBSCAN.



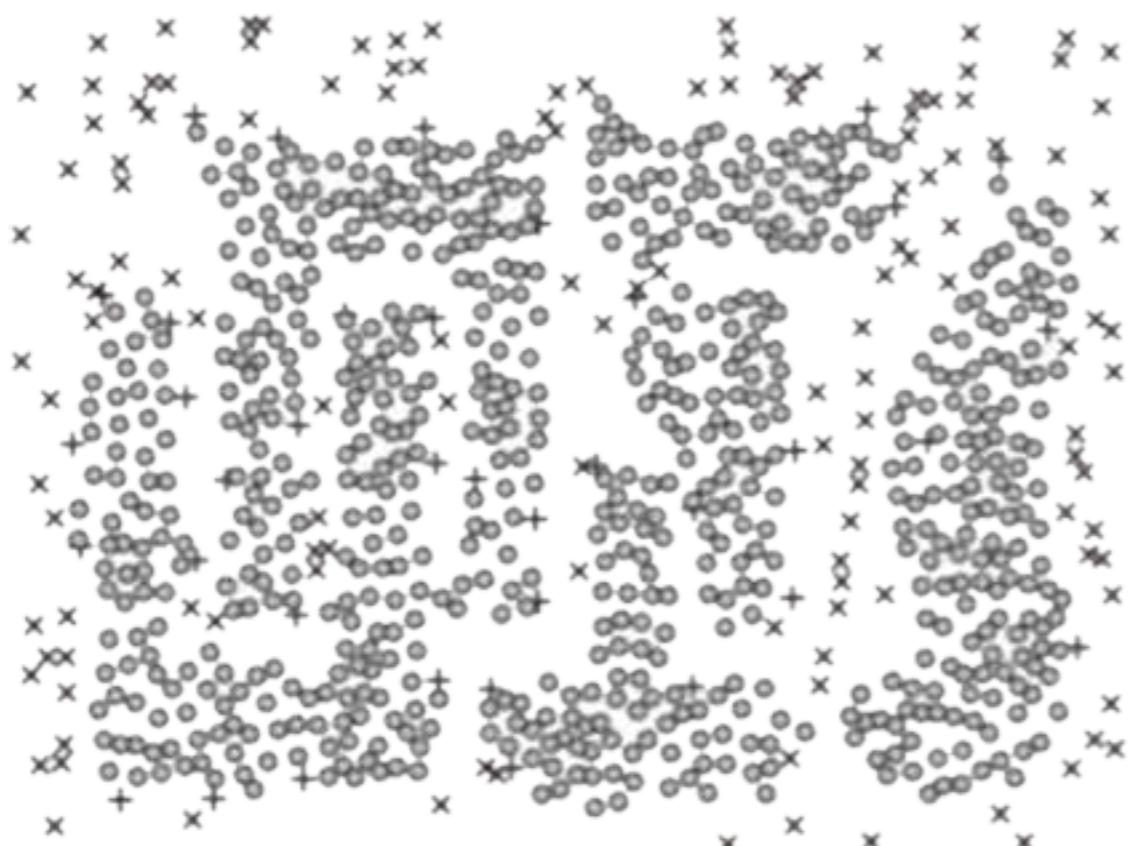
(b) Core, border, and noise points.

- 1: Пометить все точки, как основные, пограничные или шумовые.
- 2: Отбросить точки шума.
- 3: Соединить все основные точки, находящиеся на расстоянии ϵ радиуса одна от другой.

DBSCAN



(a) Clusters found by DBSCAN.



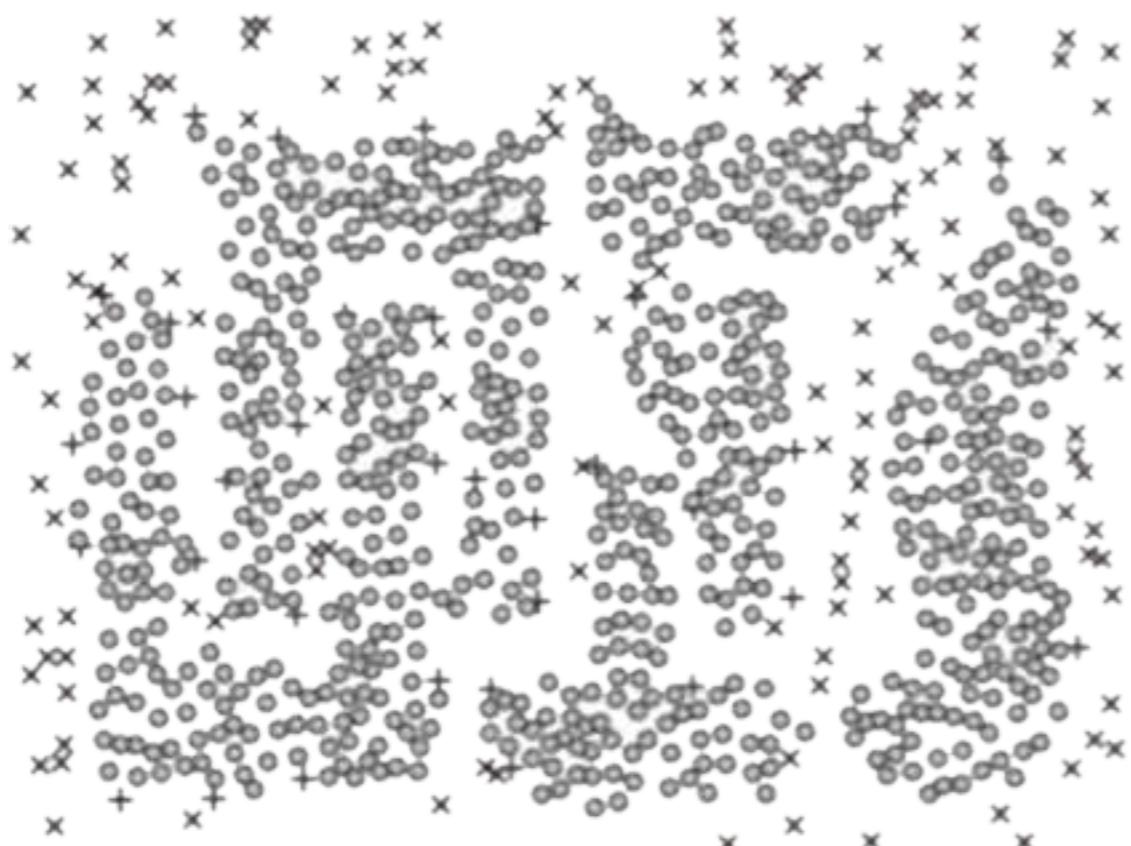
(b) Core, border, and noise points.

- 1: Пометить все точки, как основные, пограничные или шумовые.
- 2: Отбросить точки шума.
- 3: Соединить все основные точки, находящиеся на расстоянии ϵ радиуса одна от другой.
- 4: Объединить каждую группу соединённых основных точек в отдельный кластер.

DBSCAN



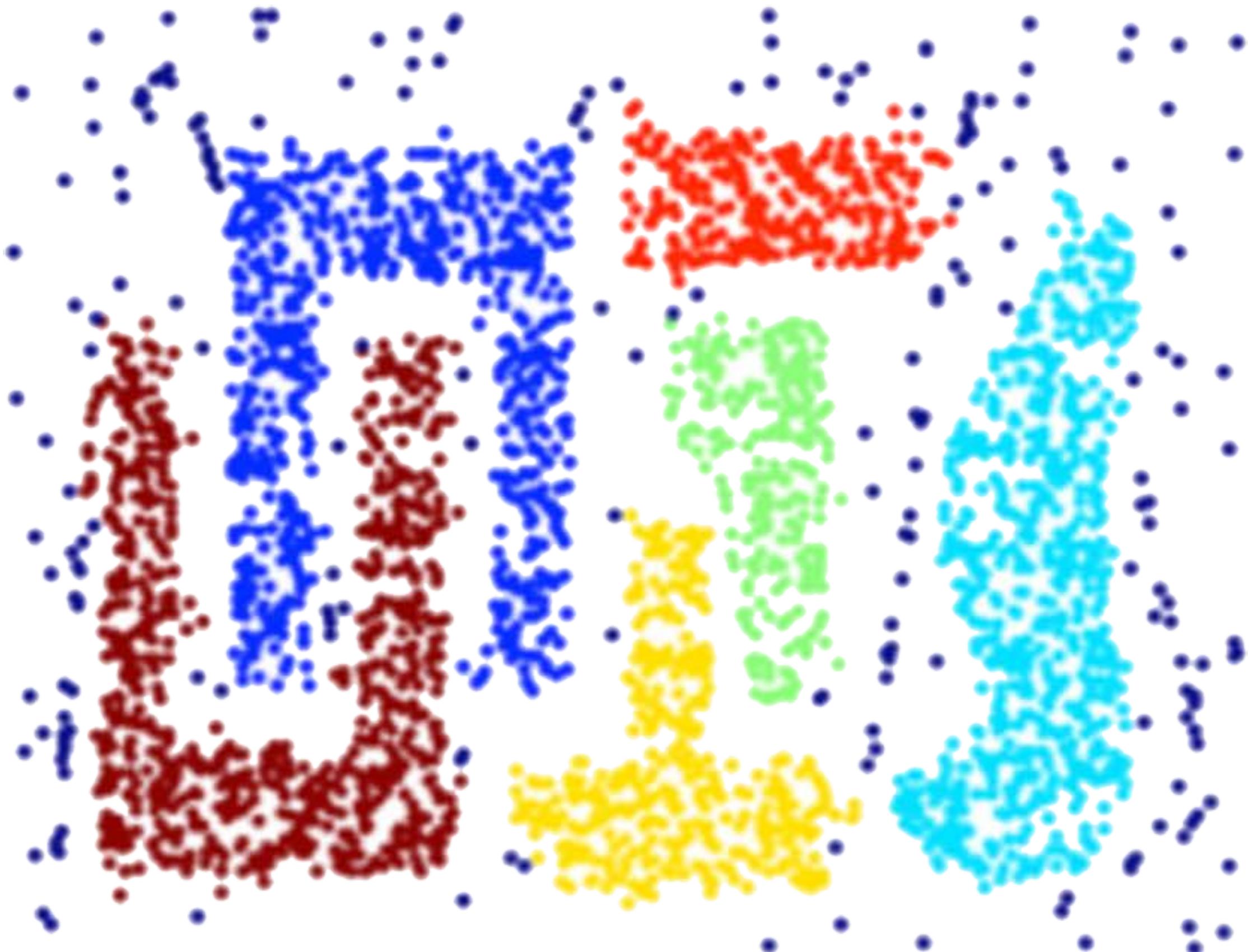
(a) Clusters found by DBSCAN.



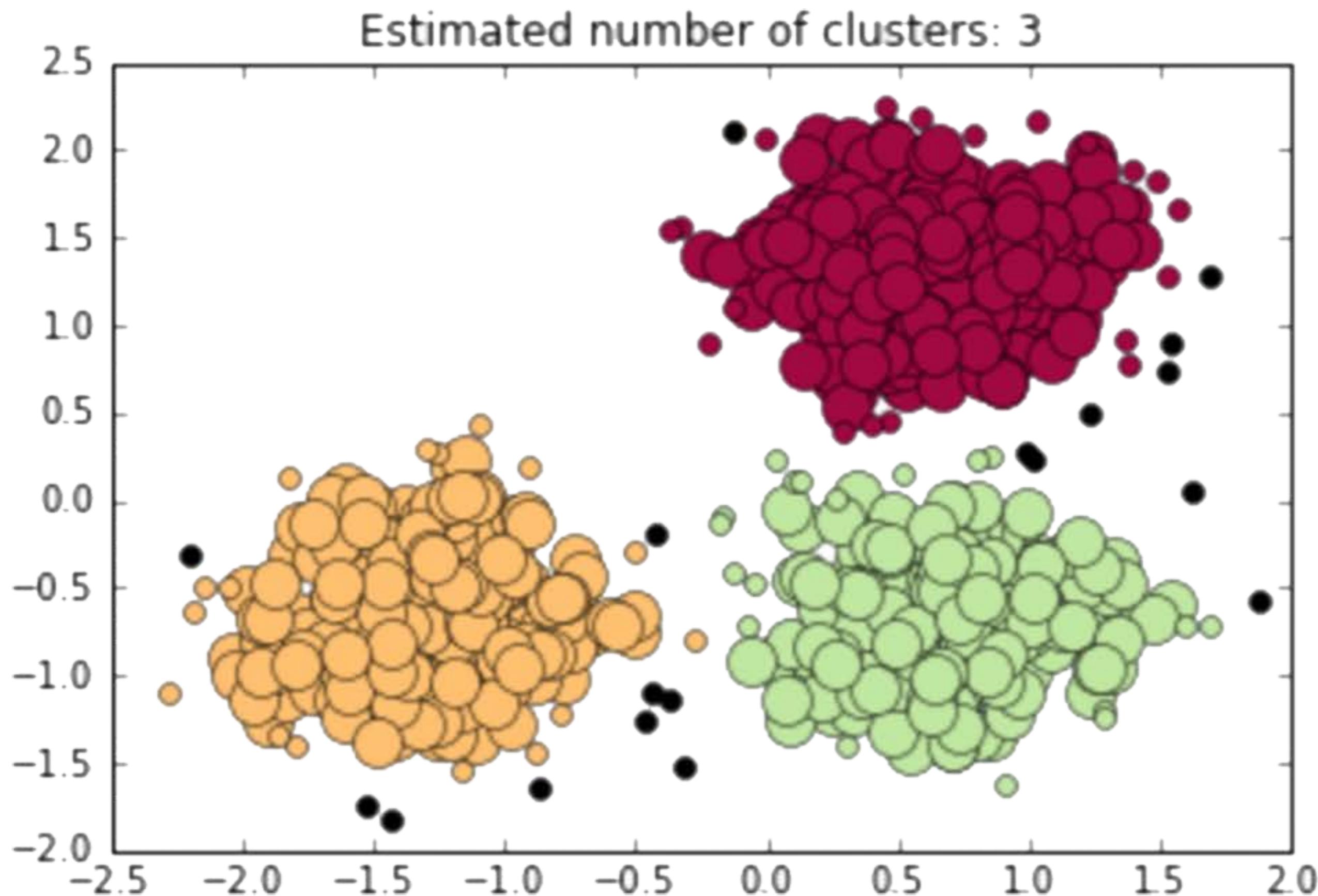
(b) Core, border, and noise points.

- 1: Пометить все точки, как основные, пограничные или шумовые.
- 2: Отбросить точки шума.
- 3: Соединить все основные точки, находящиеся на расстоянии ϵ радиуса одна от другой.
- 4: Объединить каждую группу соединённых основных точек в отдельный кластер.
- 5: Назначить каждую пограничную точку одному из кластеров, ассоциированных с ней основных точек.

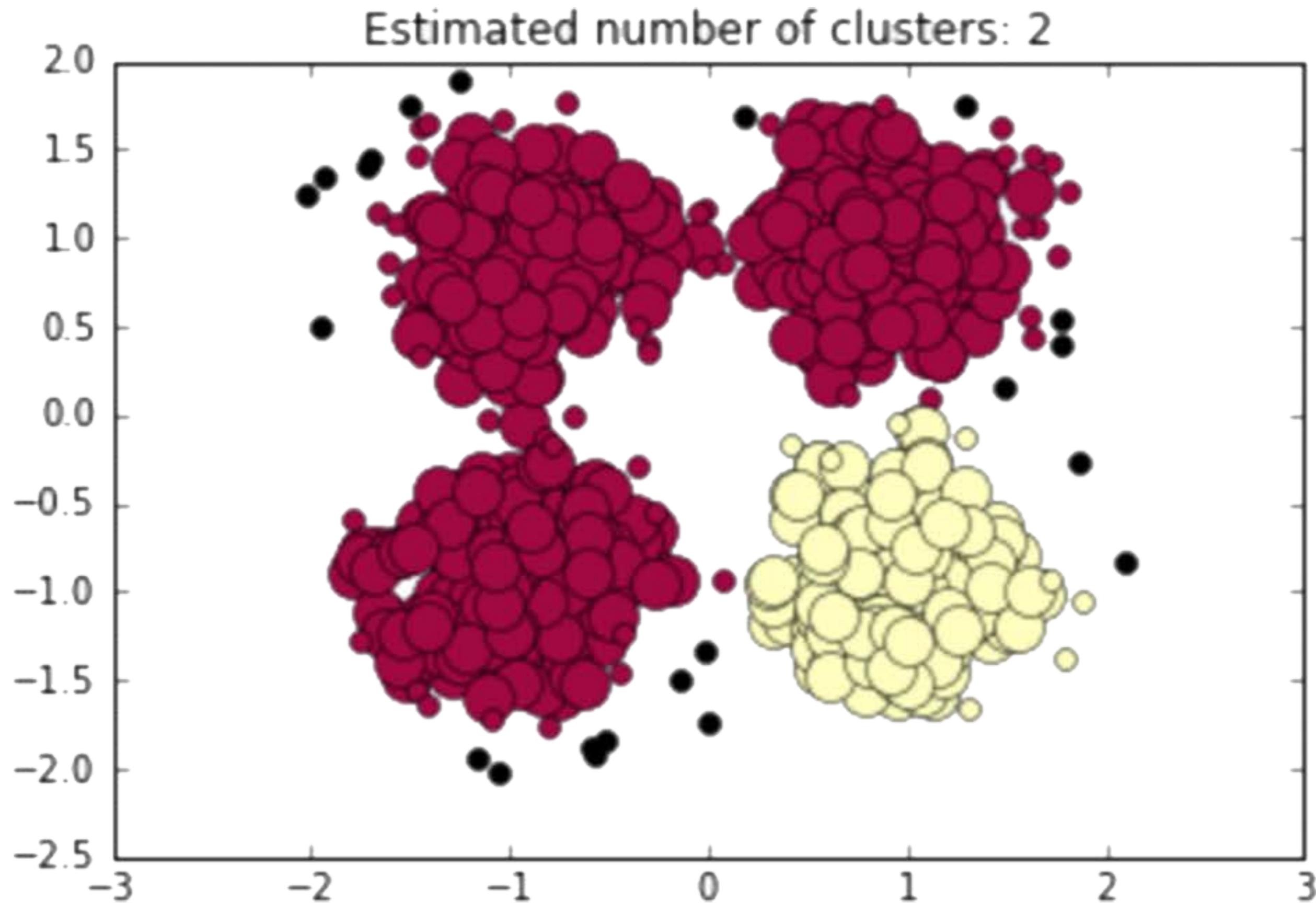
DBSCAN: РЕЗУЛЬТАТЫ РАБОТЫ



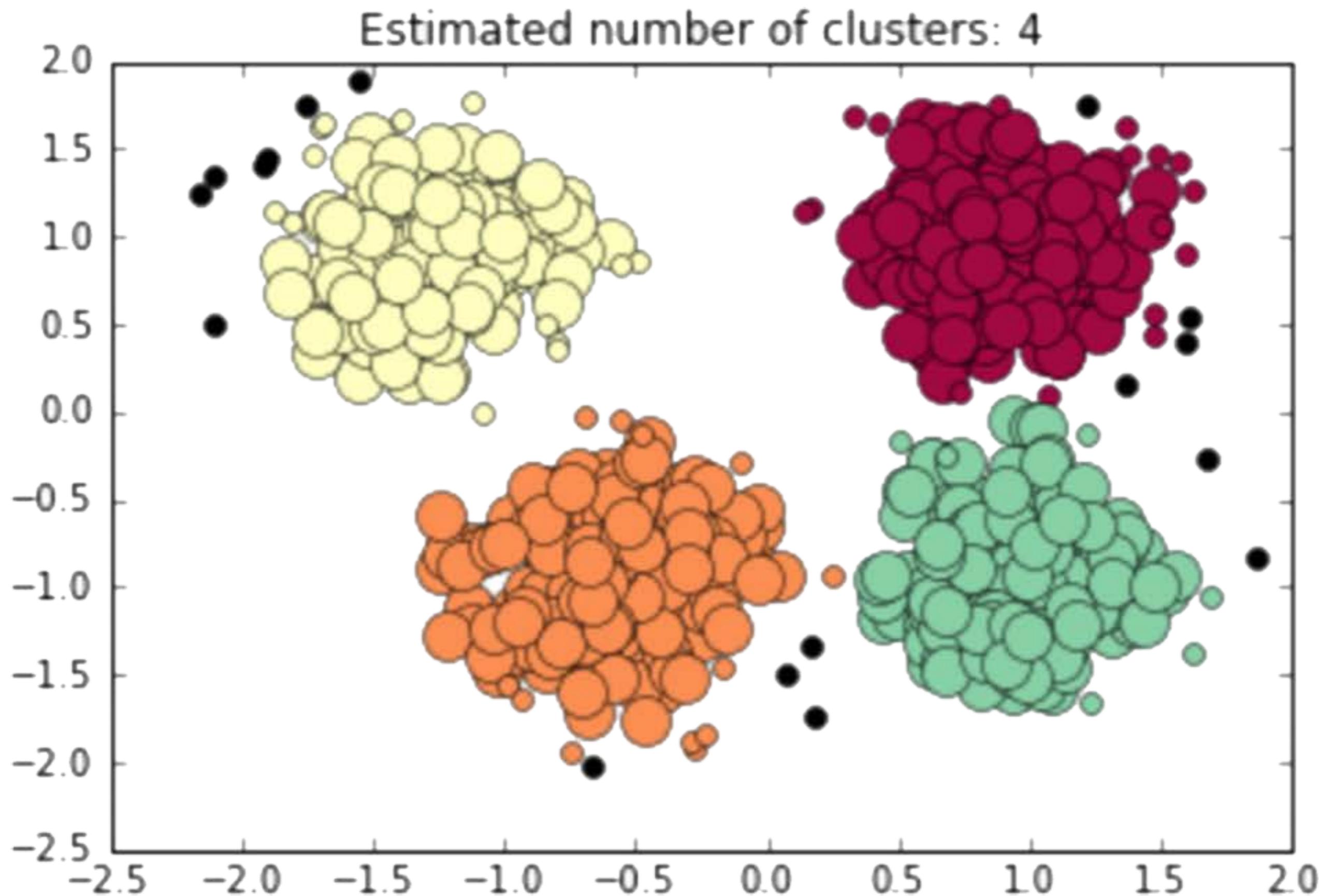
ОПРЕДЕЛЕНИЕ ЧИСЛА КЛАСТЕРОВ



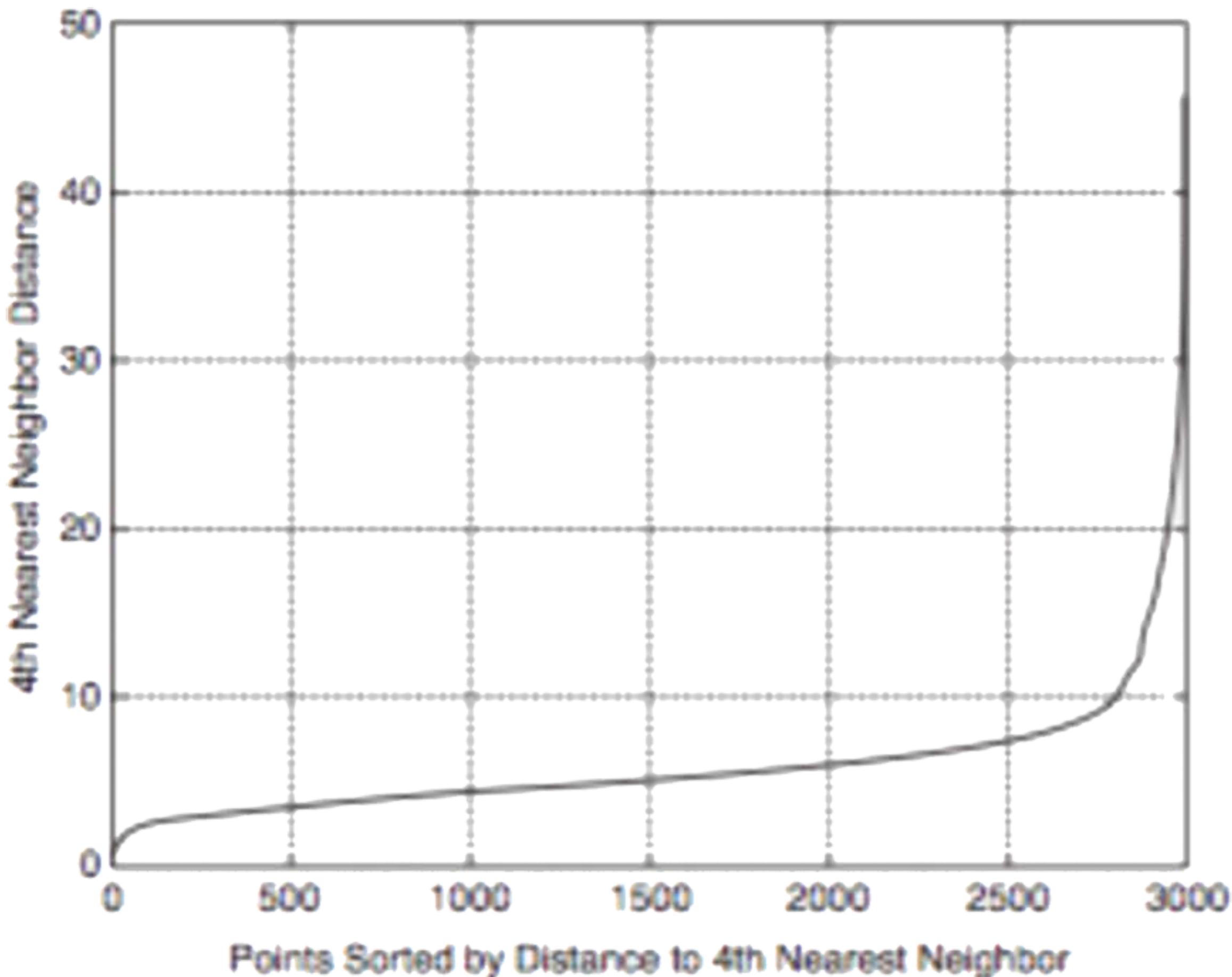
ОПРЕДЕЛЕНИЕ ЧИСЛА КЛАСТЕРОВ



ОПРЕДЕЛЕНИЕ ЧИСЛА КЛАСТЕРОВ



DBSCAN: ПОДБОР ПАРАМЕТРОВ



РЕЗЮМЕ

- › Идея методов на основе плотности точек
- › Пример основных, граничных и шумовых точек
- › DBSCAN
- › Пример работы DBSCAN
- › Определение числа кластеров
- › Настройка параметров DBSCAN

ОЦЕНКА КАЧЕСТВА И РЕКОМЕНДАЦИИ ПО РЕШЕНИЮ ЗАДАЧИ КЛАСТЕРИЗАЦИИ

ПЛАН

- › Среднее внутрикластерное и межкластерное расстояние
- › Силуэт (silhouette coefficient)
- › Подбор количества кластеров по силуэту
- › Проверка наличия кластерной структуры
- › Проблема выбора хороших признаков
- › Полнота и однородность (completeness & homogeneity)
- › Оценка качества с привлечением асессоров

СРЕДНЕЕ ВНУТРИКЛАСТЕРНОЕ РАССТОЯНИЕ

$$F_0 = \frac{\sum_{i < j} [y_i = y_j] \rho(x_i, x_j)}{\sum_{i < j} [y_i = y_j]} \rightarrow \min$$

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i: y_i = y} \rho^2(x_i, \mu_y) \rightarrow \min$$

СРЕДНЕЕ МЕЖКЛАСТЕРНОЕ РАССТОЯНИЕ

$$F_1 = \frac{\sum_{i < j} [y_i \neq y_j] \rho(x_i, x_j)}{\sum_{i < j} [y_i \neq y_j]} \rightarrow \max$$

$$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \rightarrow \max$$

КОМБИНИРУЕМ ФУНКЦИОНАЛЫ

$$F_0 = \frac{\sum_{i < j} [y_i = y_j] \rho(x_i, x_j)}{\sum_{i < j} [y_i = y_j]} \quad F_1 = \frac{\sum_{i < j} [y_i \neq y_j] \rho(x_i, x_j)}{\sum_{i < j} [y_i \neq y_j]}$$

$$F_0/F_1 \rightarrow \min$$

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i: y_i = y} \rho^2(x_i, \mu_y) \quad \Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu)$$

$$\Phi_0/\Phi_1 \rightarrow \min$$

КОЭФФИЦИЕНТ СИЛУЭТА

- » a : Среднее расстояние от данного объекта до всех других объектов из того же кластера
- » b : Среднее расстояние от данного объекта до всех объектов из ближайшего другого кластера

$$s = \frac{b - a}{\max(a, b)}$$

КОЭФФИЦИЕНТ СИЛУЭТА

- » a : Среднее расстояние от данного объекта до всех других объектов из того же кластера
- » b : Среднее расстояние от данного объекта до всех объектов из ближайшего другого кластера

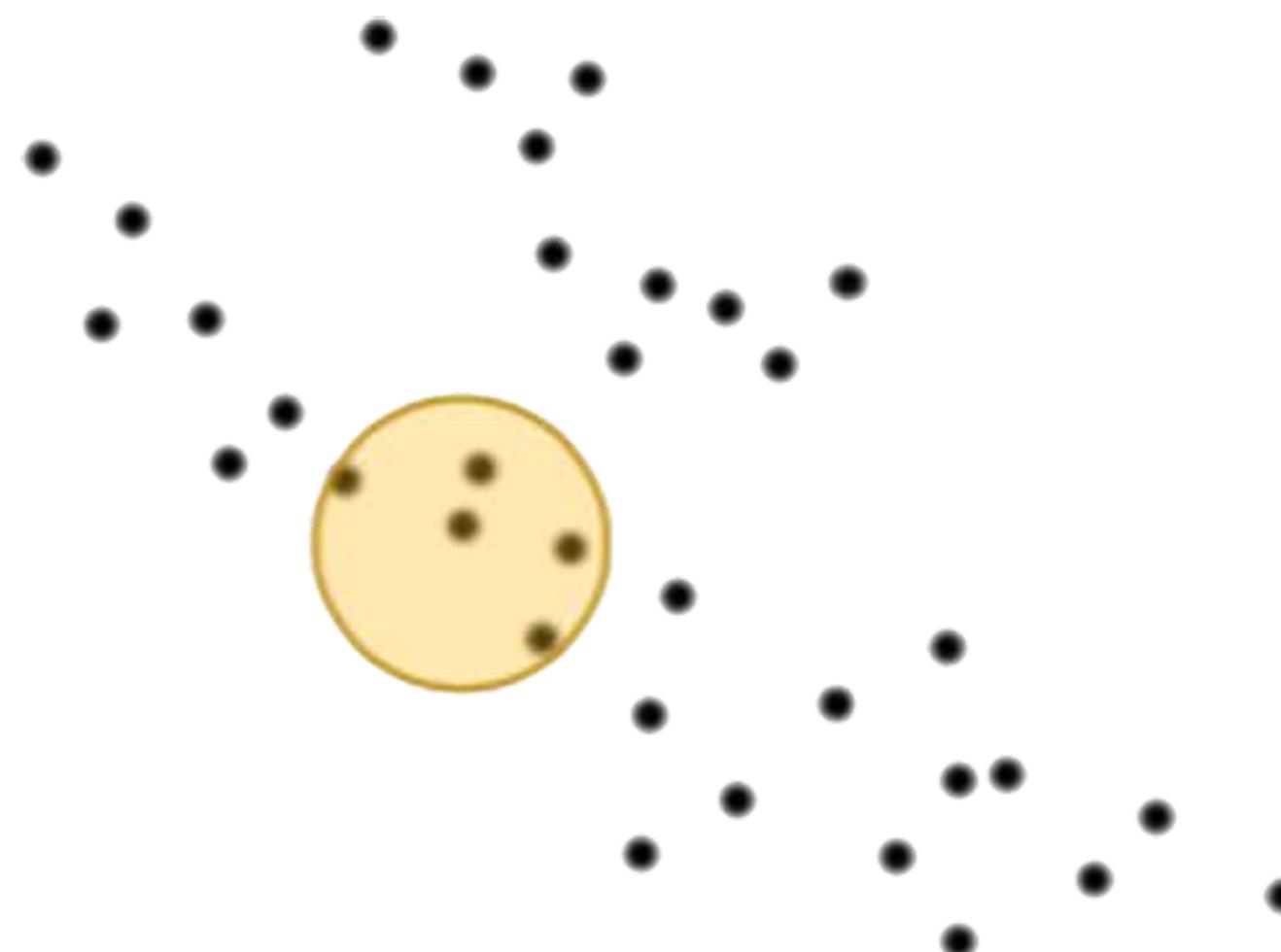
$$s = \frac{b - a}{\max(a, b)}$$



КОЭФФИЦИЕНТ СИЛУЭТА

- » a : Среднее расстояние от данного объекта до всех других объектов из того же кластера
- » b : Среднее расстояние от данного объекта до всех объектов из ближайшего другого кластера

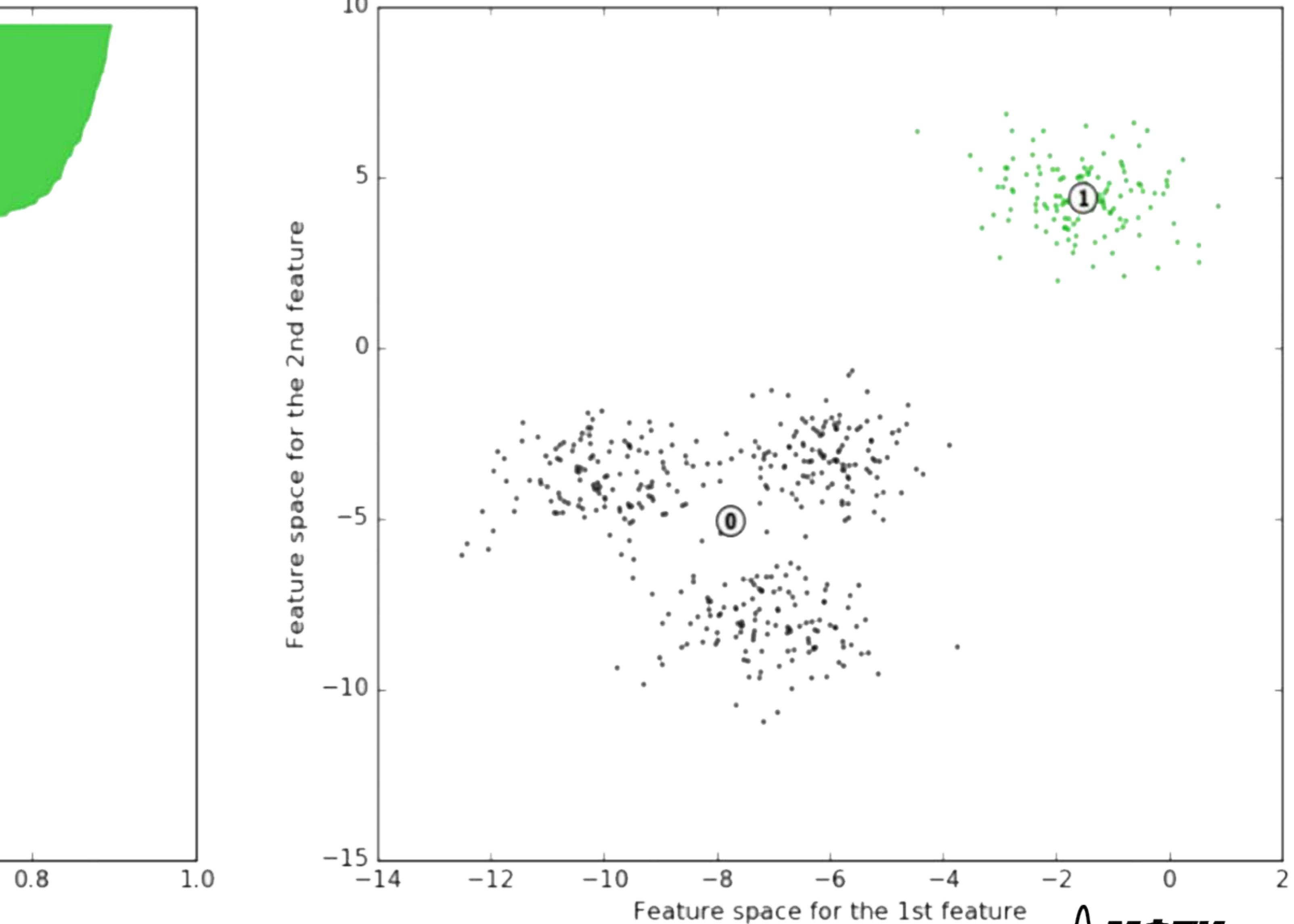
$$s = \frac{b - a}{\max(a, b)}$$



KMeans clustering on sample data with n_clusters = 2

ers.

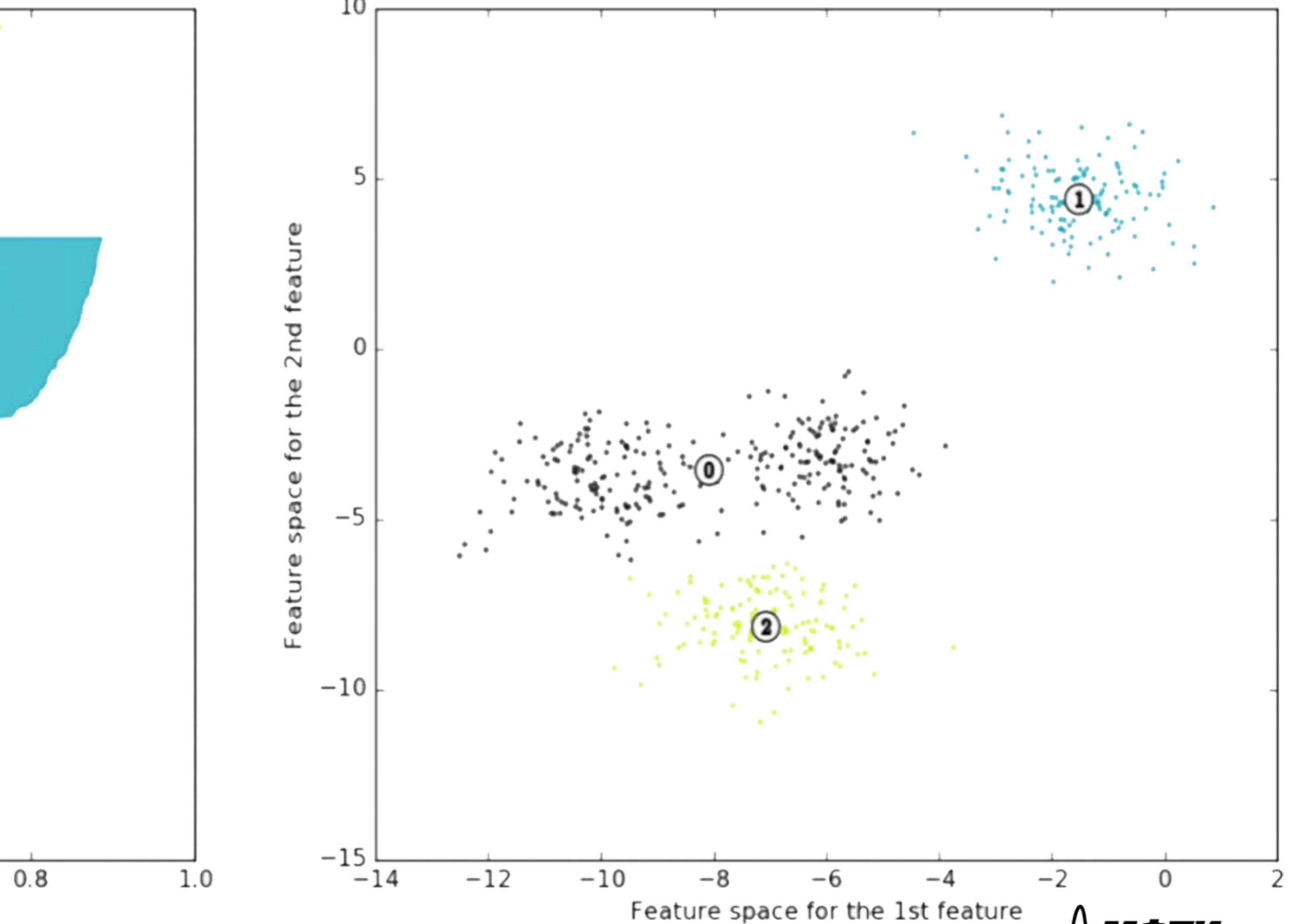
The visualization of the clustered data.



KMeans clustering on sample data with n_clusters = 3

ers.

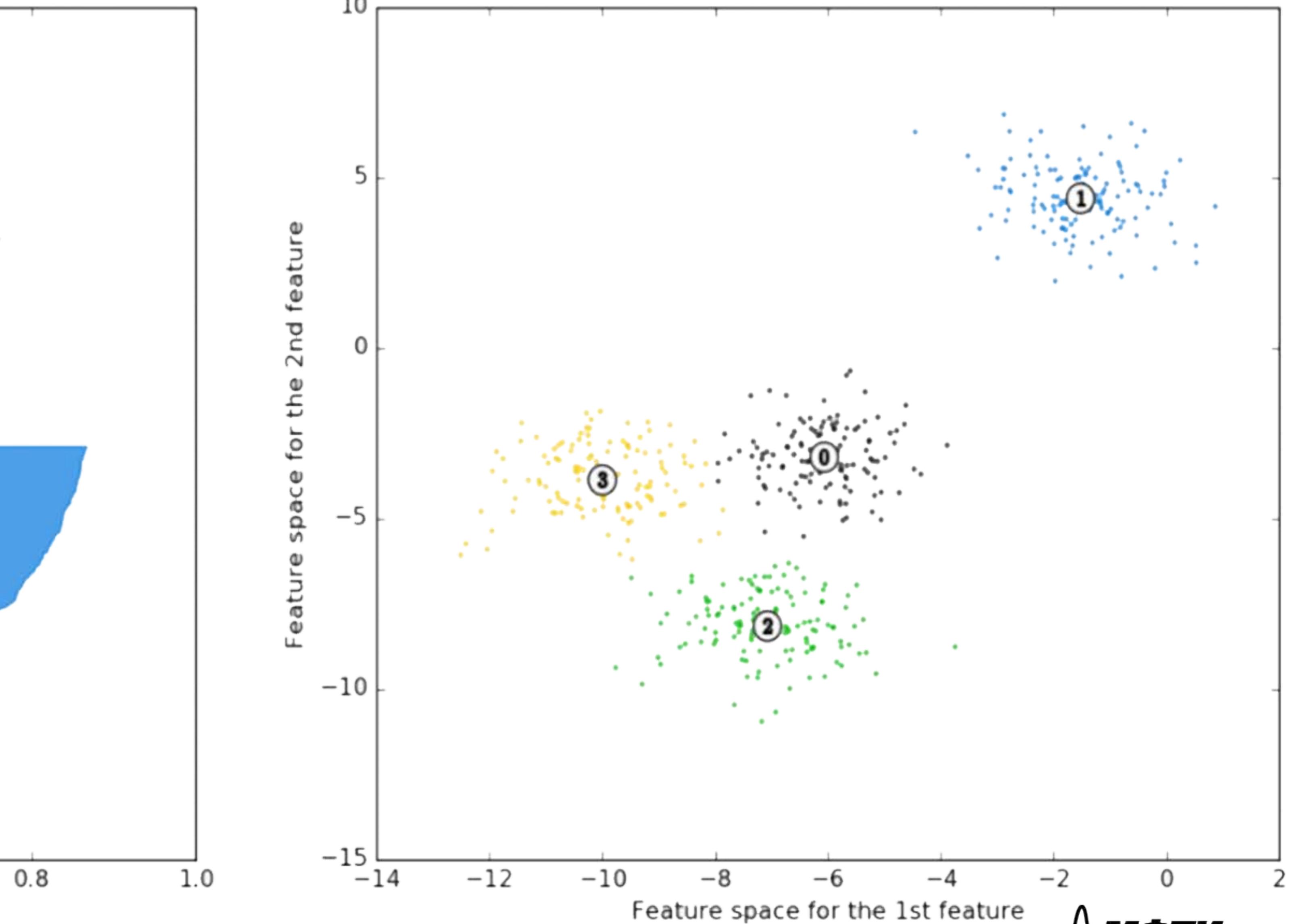
The visualization of the clustered data.



KMeans clustering on sample data with n_clusters = 4

ers.

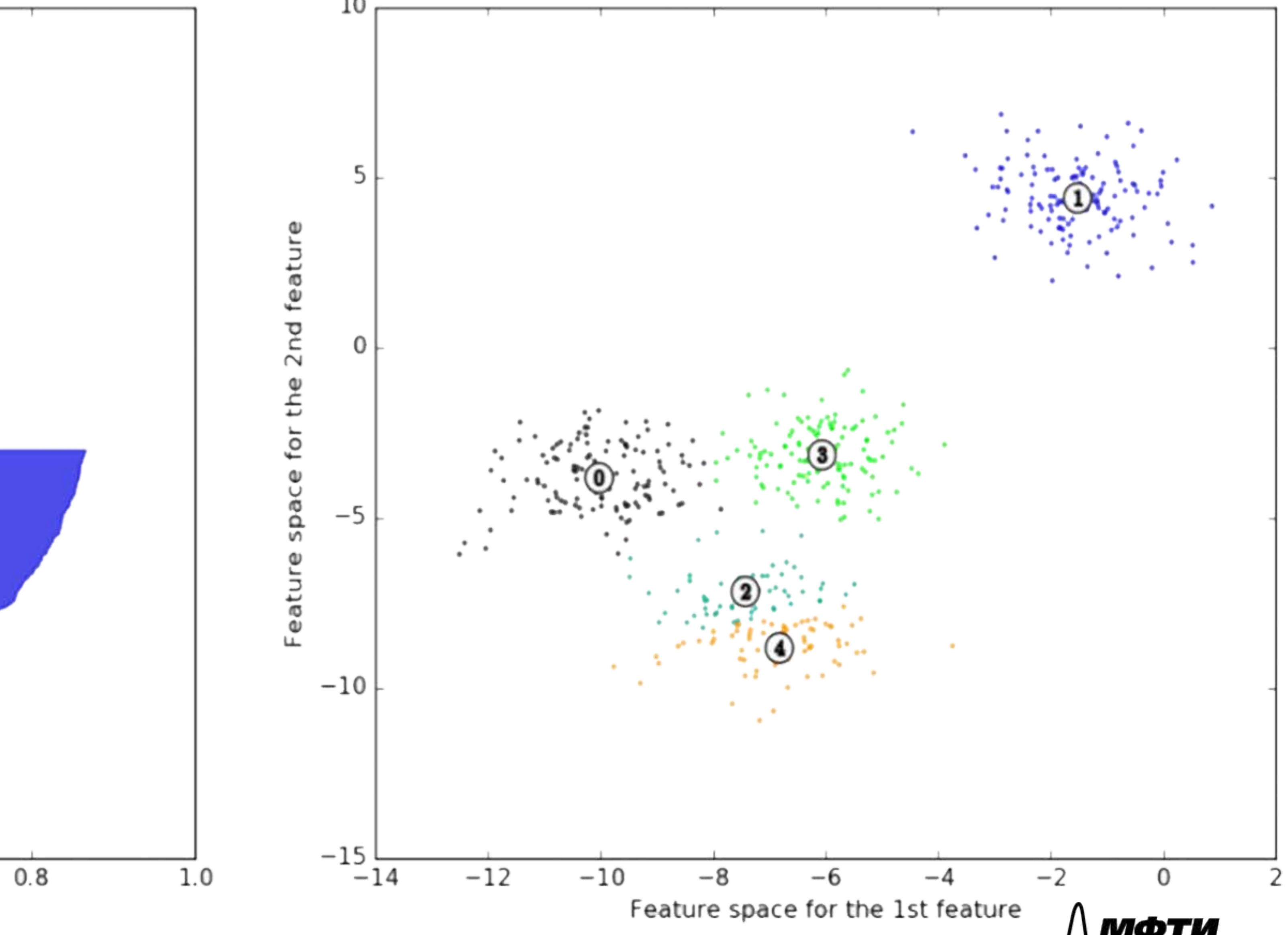
The visualization of the clustered data.



KMeans clustering on sample data with n_clusters = 5

ers.

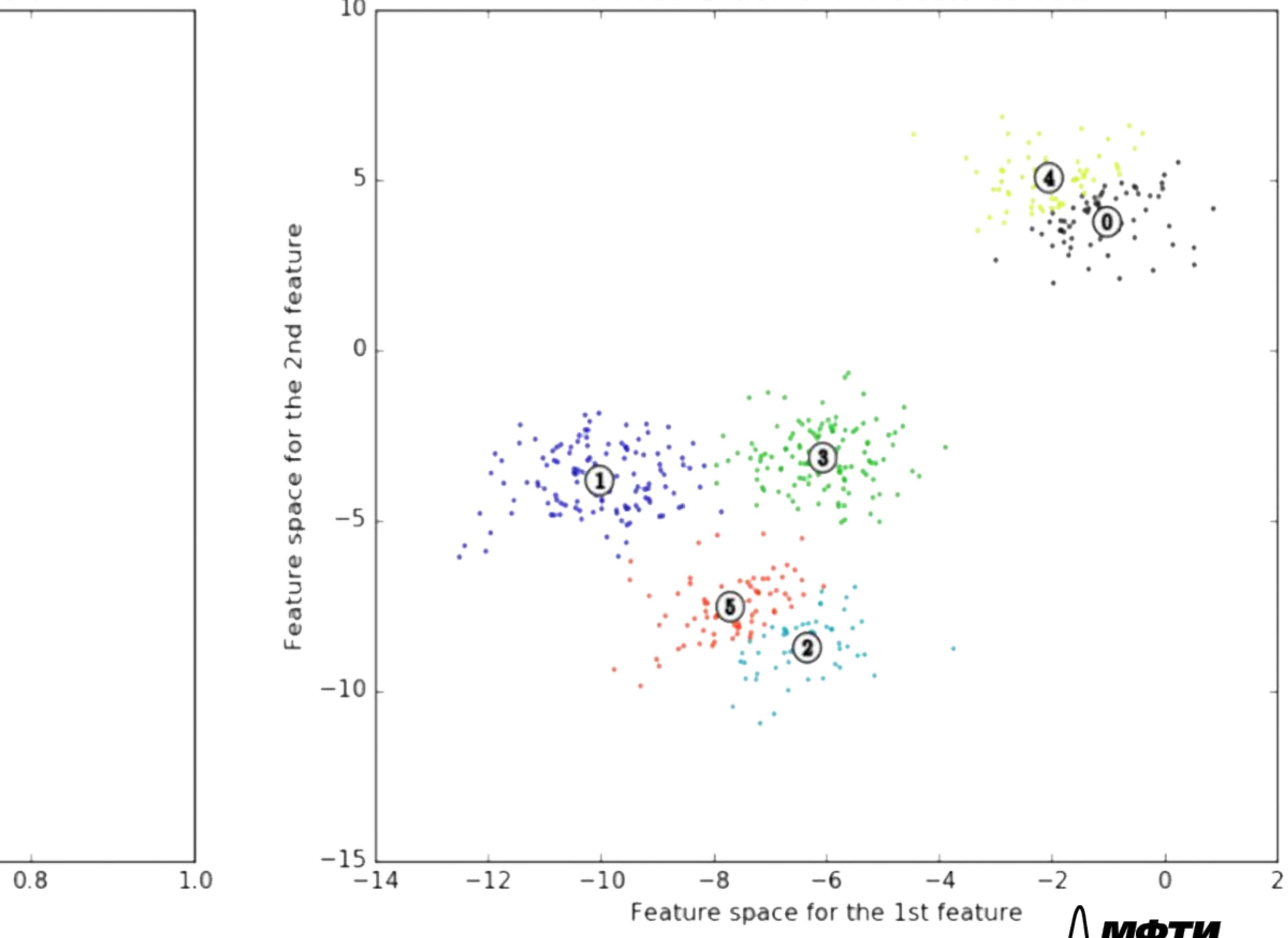
The visualization of the clustered data.



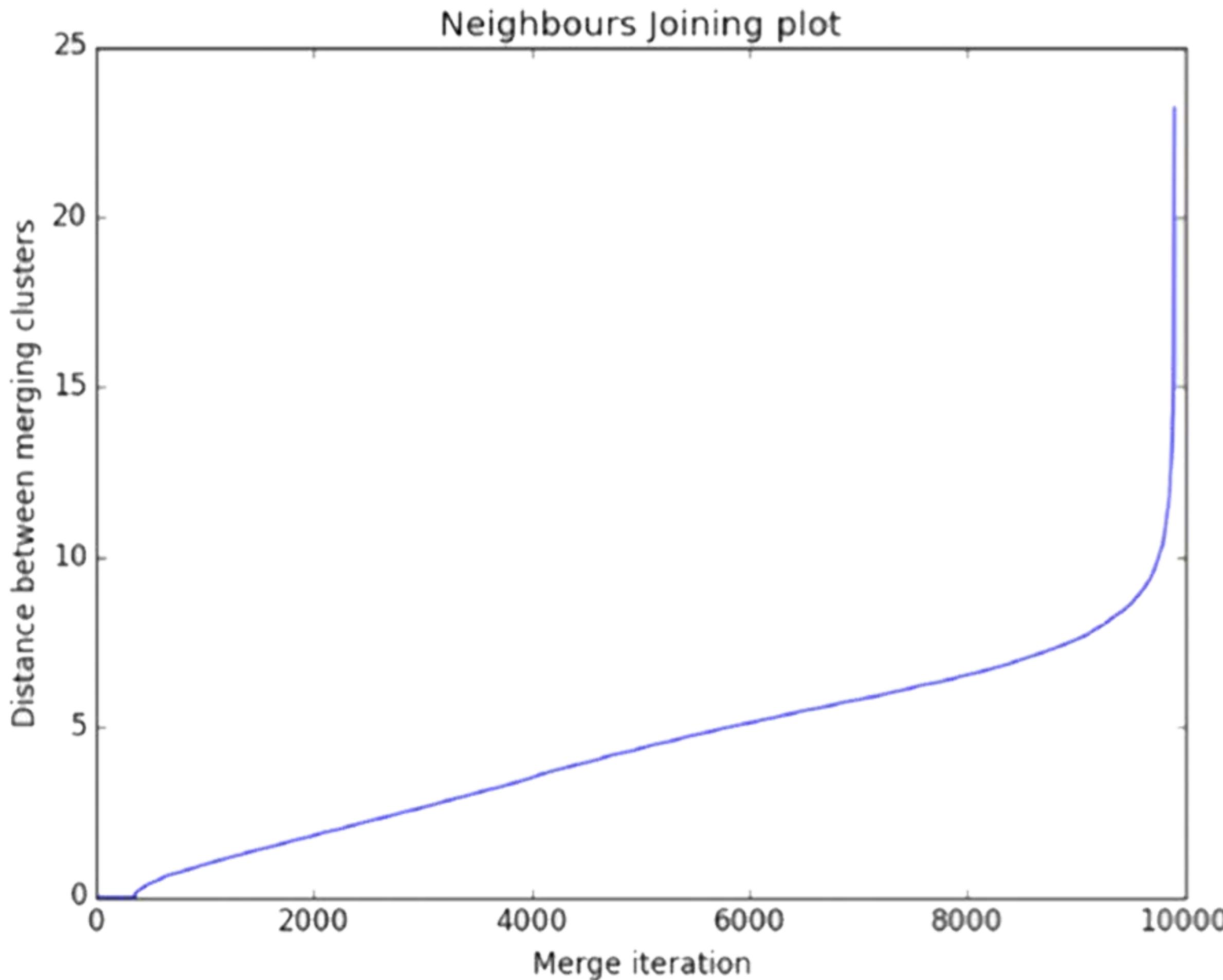
KMeans clustering on sample data with n_clusters = 6

ers.

The visualization of the clustered data.



ПРОВЕРКА НАЛИЧИЯ КЛАСТЕРНОЙ СТРУКТУРЫ



ПРОВЕРКА НАЛИЧИЯ КЛАСТЕРНОЙ СТРУКТУРЫ

- › Генерируем p случайных точек из равномерного распределения и p случайных из обучающей выборки
- › Вычисляем величину (статистика Хопкинса):

$$H = \frac{\sum_{i=1}^p w_i}{\sum_{i=1}^p u_i + \sum_{i=1}^p w_i}$$

ВЫБОР ПРИЗНАКОВ

- › Что хотим уметь делать:
Для разных признаков понимать, насколько
хорошо решена задача кластеризации
- › Зачем:
Тогда сможем выбирать наиболее
адекватные признаки
- › В чем проблема:
Текущие метрики зависят от признакового
пространства

ОДНОРОДНОСТЬ, ПОЛНОТА, V-МЕРА

- В каких случаях значения метрик максимальны:
 - ▶ Однородность: кластер состоит только из объектов одного класса
 - ▶ Полнота: все объекты из класса принадлежат к одному кластеру

ОДНОРОДНОСТЬ, ПОЛНОТА, V-МЕРА

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

ОДНОРОДНОСТЬ, ПОЛНОТА, V-МЕРА

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$H = - \sum_i p_i \ln p_i$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

ОДНОРОДНОСТЬ, ПОЛНОТА, V-МЕРА

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$H = - \sum_i p_i \ln p_i \quad H(C) = - \sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log \left(\frac{n_c}{n} \right)$$

ОДНОРОДНОСТЬ, ПОЛНОТА, V-МЕРА

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$H = - \sum_i p_i \ln p_i \quad H(C) = - \sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log \left(\frac{n_c}{n} \right)$$

$$P(c) = \frac{n_c}{n}$$

ОДНОРОДНОСТЬ, ПОЛНОТА, V-МЕРА

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$H = - \sum_i p_i \ln p_i \quad H(C) = - \sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log \left(\frac{n_c}{n} \right)$$

$$P(c) = \frac{n_c}{n}$$

$$H(C|K) = - \sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{n_{c,k}}{n} \cdot \log \left(\frac{n_{c,k}}{n} \right) \quad P(c|k) = \frac{n_{c,k}}{n_k}$$

ПРИВЛЕЧЕНИЕ АСЕССОРОВ ДЛЯ ОЦЕНКИ КАЧЕСТВА

- Если разметки нет, можно:
 - ▶ Использовать метрики без разметки
 - ▶ Создать разметку с помощью асессоров и использовать ее
 - ▶ Предложить асессорам отвечать на вопросы вида «допустимо ли эти объекты относить в один/в разные кластеры»

РЕЗЮМЕ

- › Среднее внутрикластерное и межкластерное расстояние
- › Силуэт (silhouette coefficient)
- › Подбор количества кластеров по силуэту
- › Проверка наличия кластерной структуры
- › Проблема выбора хороших признаков
- › Полнота и однородность (completeness & homogeneity)
- › Оценка качества с привлечением асессоров