

Data Mining in Action

Лекция 3
Обучение без учителя

Что такое обучения без учителя или
unsupervised learning?

Обучение без учителя

- Известны только описания объектов
- Требуется обнаружить внутренние взаимосвязи, зависимости, закономерности

Задачи обучения без учителя

- Кластеризация
- Детектирование аномалий
- Понижение размерности
- Визуализация данных

На этой лекции

- Вы научитесь различать задачи
 - кластеризации
 - детектирования аномалий
 - понижения размерности
 - визуализации данных
- Узнаете подходы к оценке качества кластеризации
- Узнаете основные алгоритмы кластеризации

Задачи обучения без учителя

Кластеризация

- Требуется разбить объекты на группы

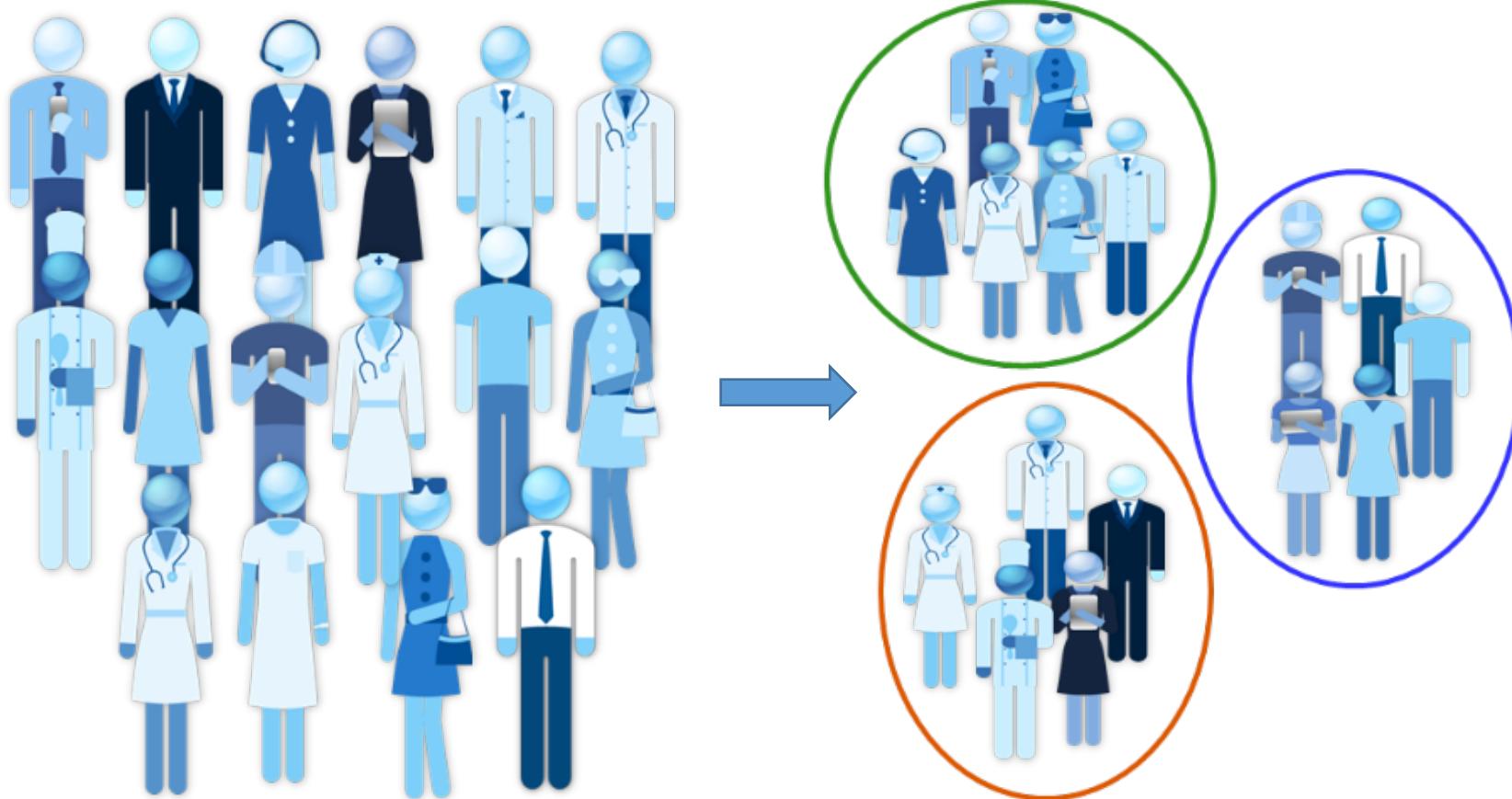
Кластеризация

- Требуется разбить объекты на группы таким образом, чтобы:
 - группы отражали **структуру** исходных данных
 - объекты внутри одной группы были **похожи** друг на друга
 - объекты из разных групп **отличались** друг от друга

Кластеризация



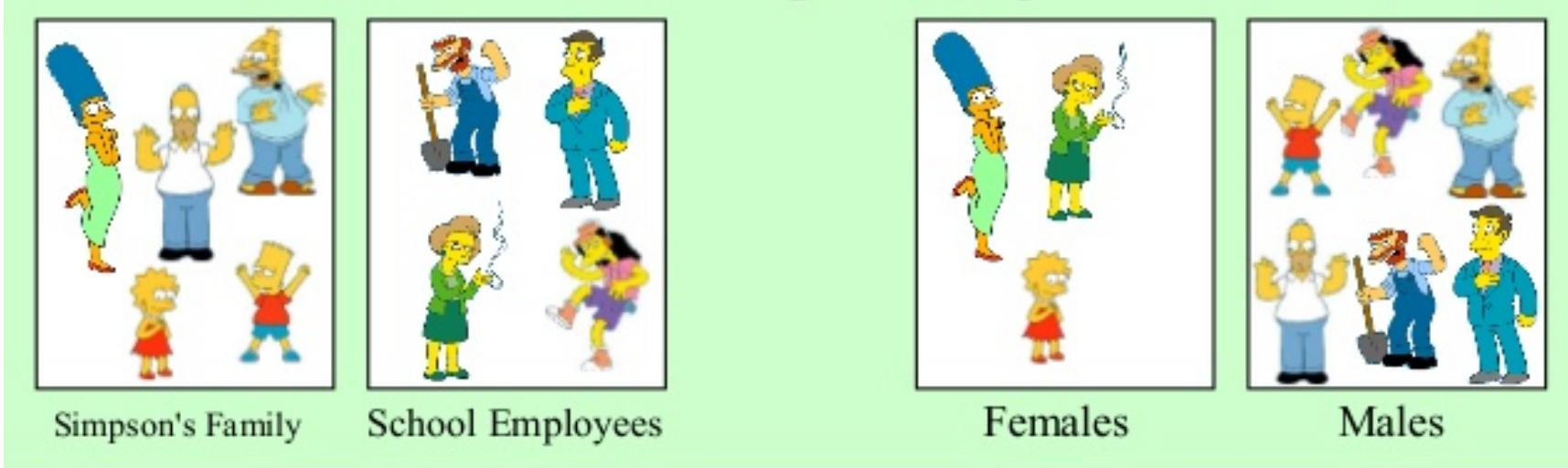
Кластеризация



Какова правильная групповая структура?



Кластеризация субъективна



Классификация vs Кластеризация

Обучающая выборка:

x_1, \dots, x_l - объекты

y_1, \dots, y_l - ответы

Тестовая выборка:

x_{l+1}, \dots, x_{l+u}

Задача: построить $a(x) \approx y$

Обучающая выборка:

x_1, \dots, x_l - объекты

Тестовая выборка:

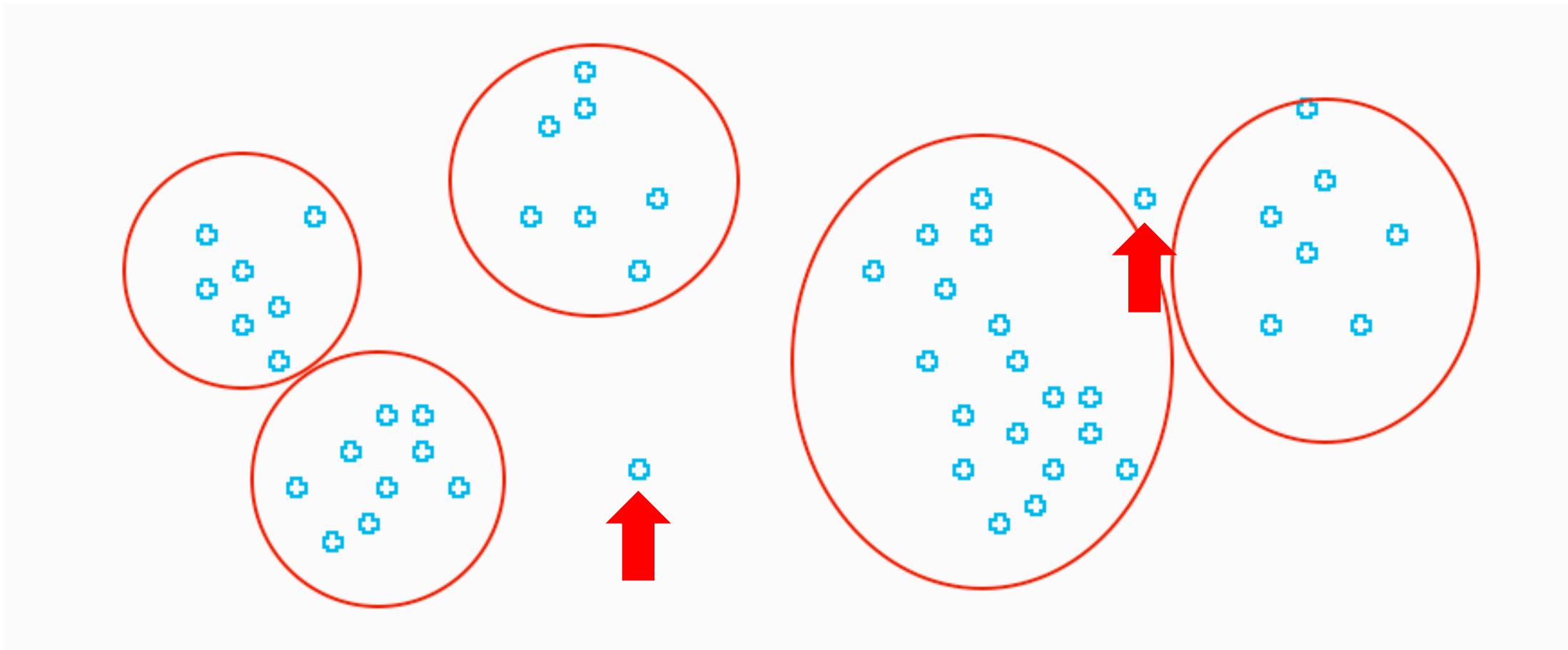
x_1, \dots, x_l - объекты

Задача: разбить объекты на
группы

Детектирование аномалий

- Детектирование выбросов или outlier detection
- Детектирование новизны или novelty detection

Детектирование аномалий



Детектирование аномалий



Детектирование аномалий



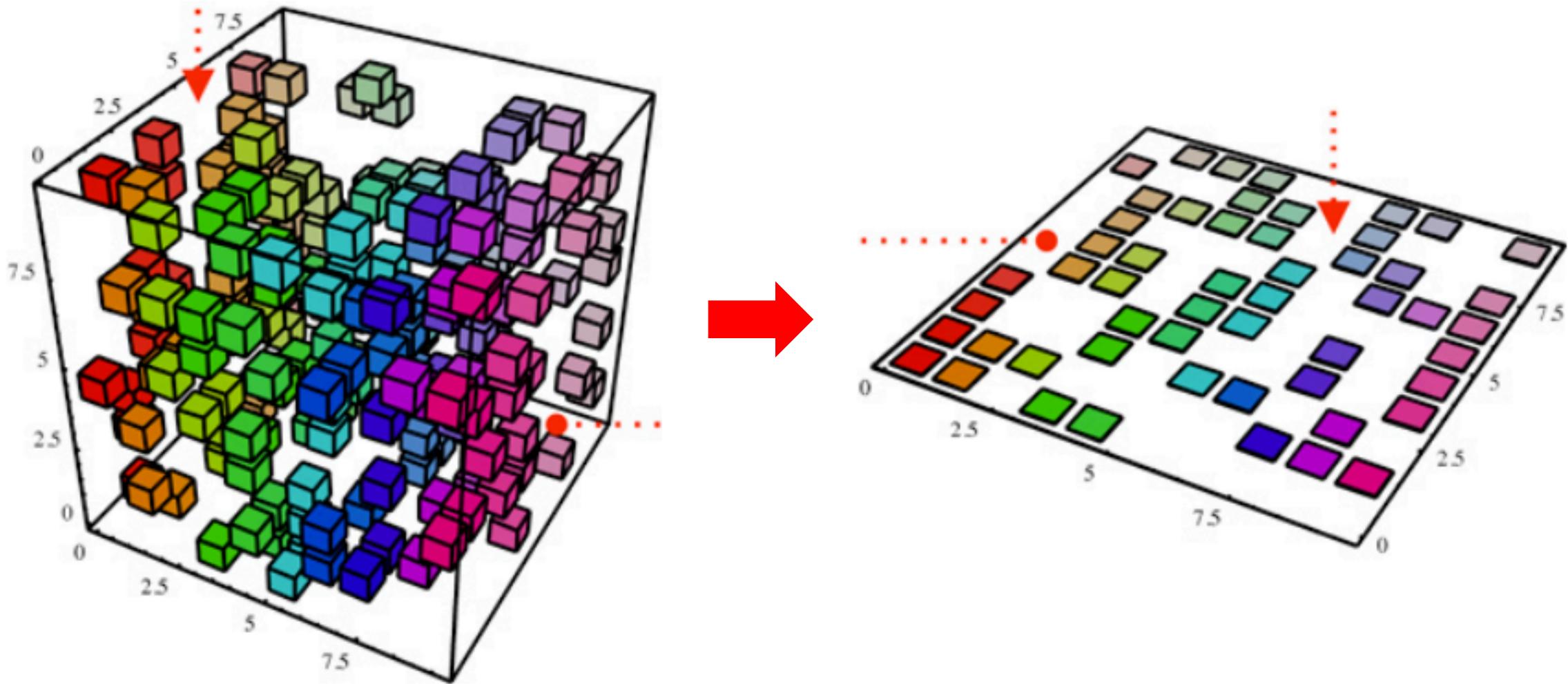
Понижение размерности данных

- Процесс уменьшения размерности анализируемого множества данных до размера, оптимального с точки зрения решаемой задачи

Зачем снижать размерность данных?

- Исходные данные избыточны с точки зрения количества информации, необходимого для решения задачи
- Оптимизация вычислительных затрат
- Подготовка данных для дальнейшего анализа

Понижение размерности



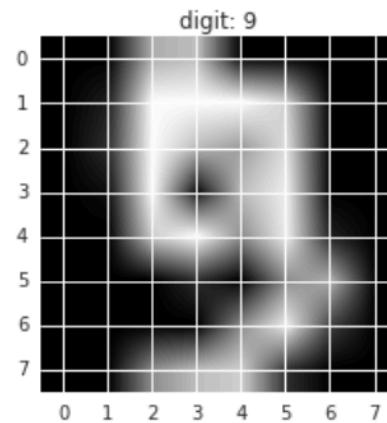
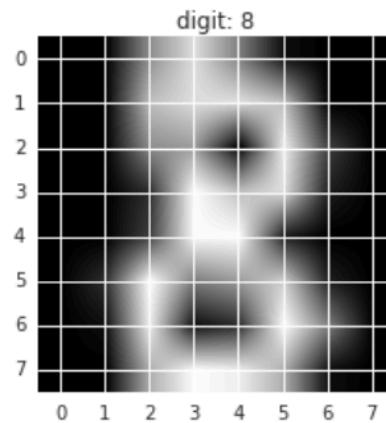
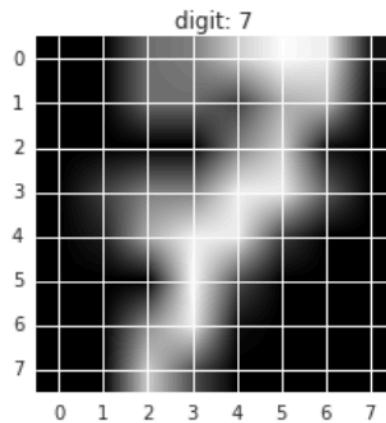
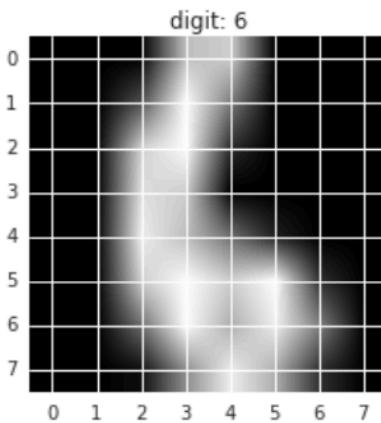
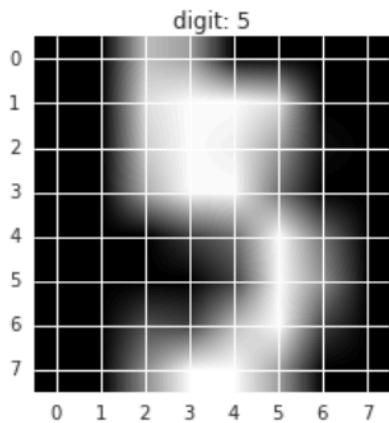
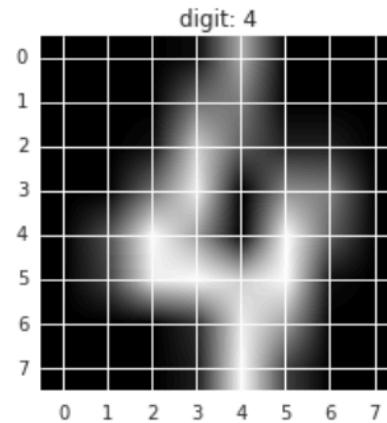
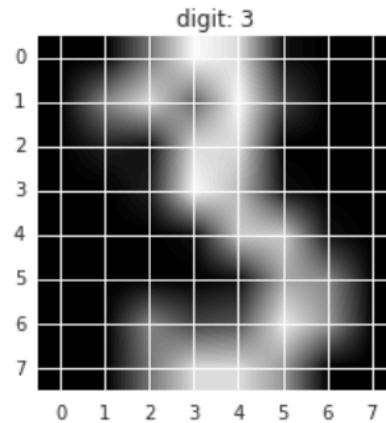
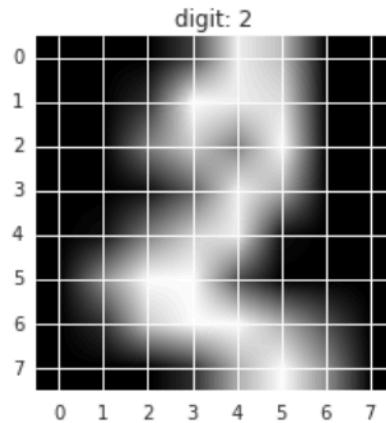
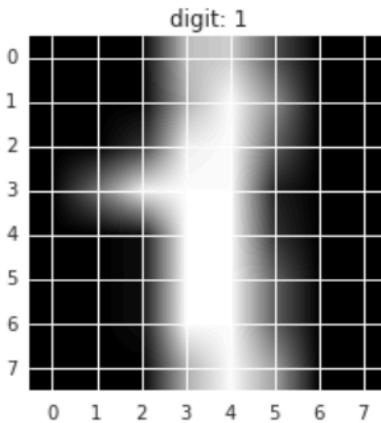
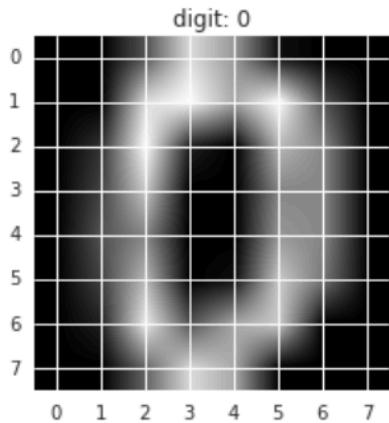
Визуализация данных

- Наглядное представление данных для эффективного восприятия и анализа

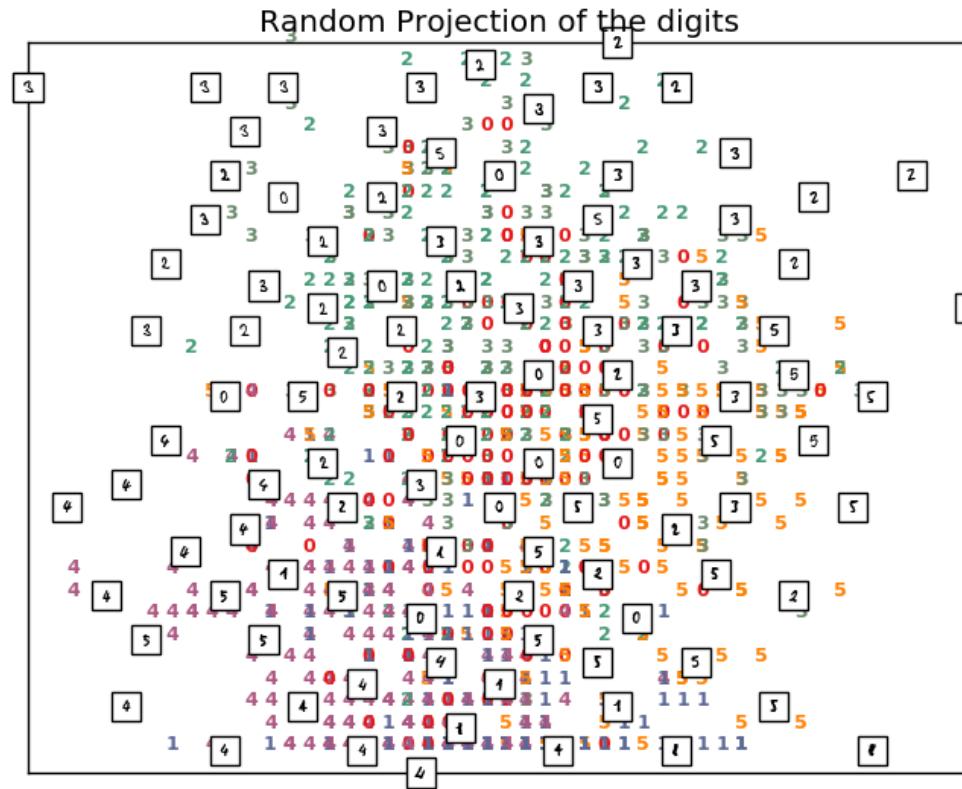
Распознавание рукописных цифр

- X: каждая картинка описывается набором пикселей, где яркость пикселя соответствует силе нажатия ручки на бумагу
- Y: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} – исходная цифра, которую написали от руки
- Задача: распознать исходную цифру по её изображению

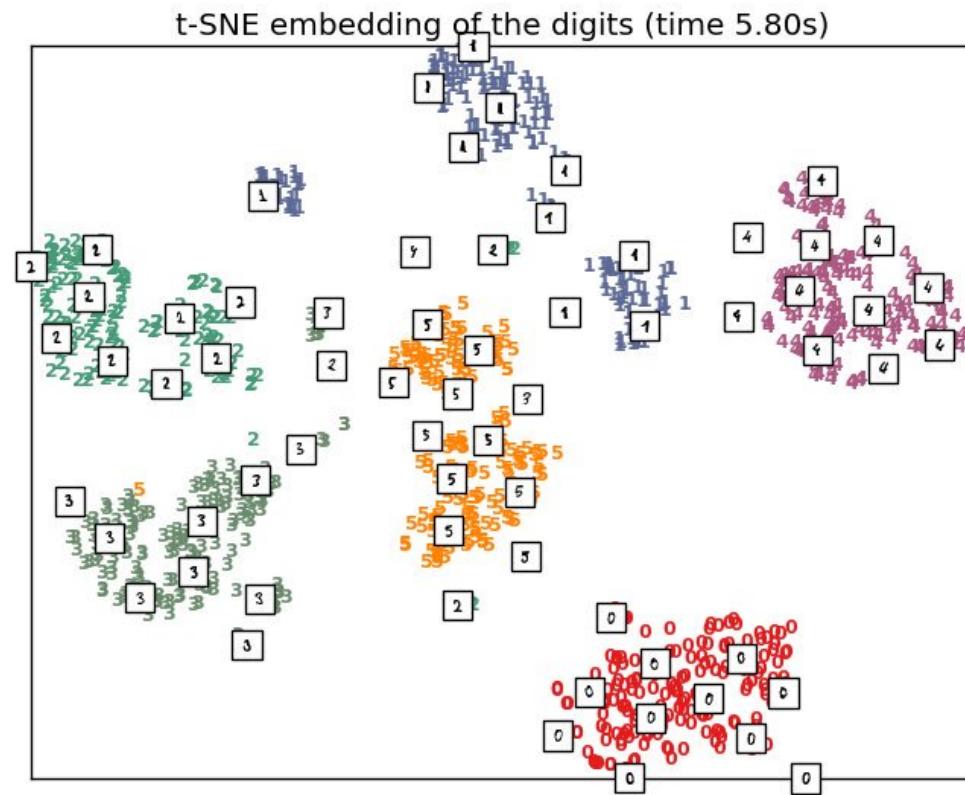
Насколько эта задача сложна?



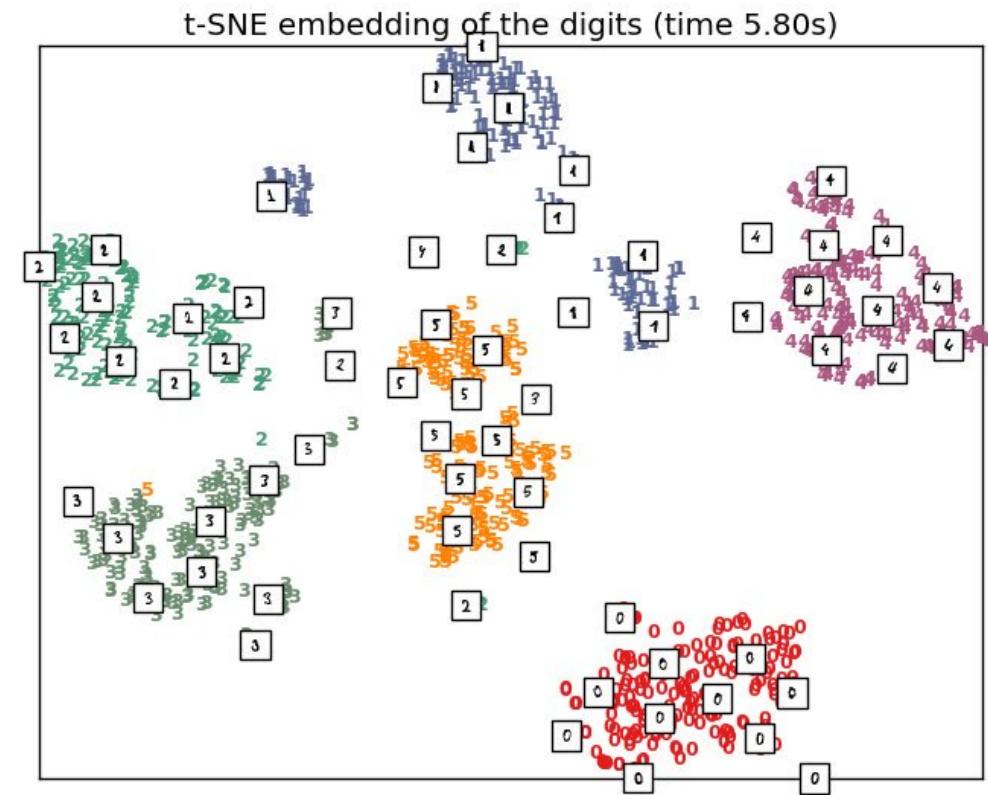
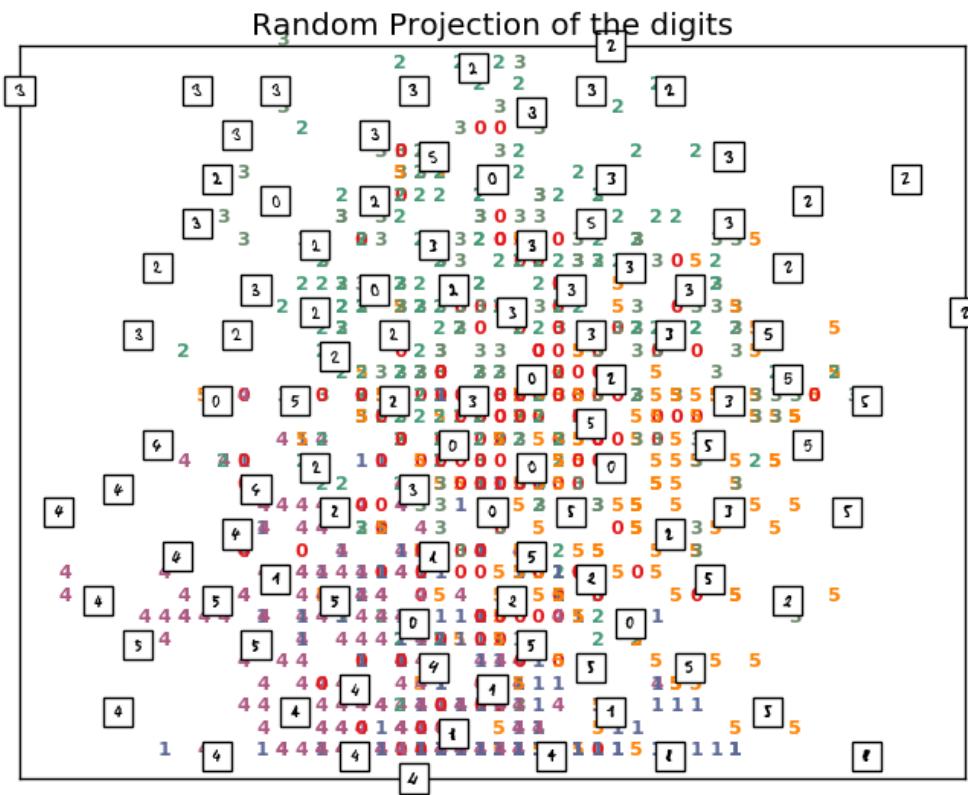
Визуализация + уменьшение размерности



Визуализация + уменьшение размерности



Важно, чтобы было «красиво»



Как оценить качество кластеризации?

Качество кластеризации

- Как сильно объекты внутри одной группы **похожи** друг на друга?
- Как сильно объекты из разных групп **отличаются** друг от друга?
- Насколько группы отражают **структуру** исходных данных?

Среднее внутрикластерное расстояние

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

Среднее межклusterное расстояние

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

Как объединить критерии F_0 и F_1 ?

Как объединить критерии F_0 и F_1 ?

- $F_0/F_1 \rightarrow \min$

Оценка качества решения конечной задачи

- Зачем делается кластеризация?
- Помогают ли её результаты решить конечную задачу?

Оценка качества решения конечной задачи

- Зачем делается кластеризация?
- Помогают ли её результаты решить конечную задачу?
- Эффективность маркетинговой кампании
- Качество модели и вклад признаков в модель
- Качество модели vs скорость обучения

Основные алгоритмы кластеризации

Алгоритмы кластеризации

- Для разных задач подходят разные алгоритмы
- Важно учитывать:
 - Особенности данных
 - Желаемую форму кластеров
 - Количество и размер кластеров
 - Вложенность кластеров
 - Мягкость/жесткость кластеризации

Алгоритмы кластеризации

Мы рассмотрим:

- K-means
- Графовые алгоритмы
- Алгоритмы на основе плотности точек
- Иерархическую кластеризацию

Алгоритм k-means (k средних)

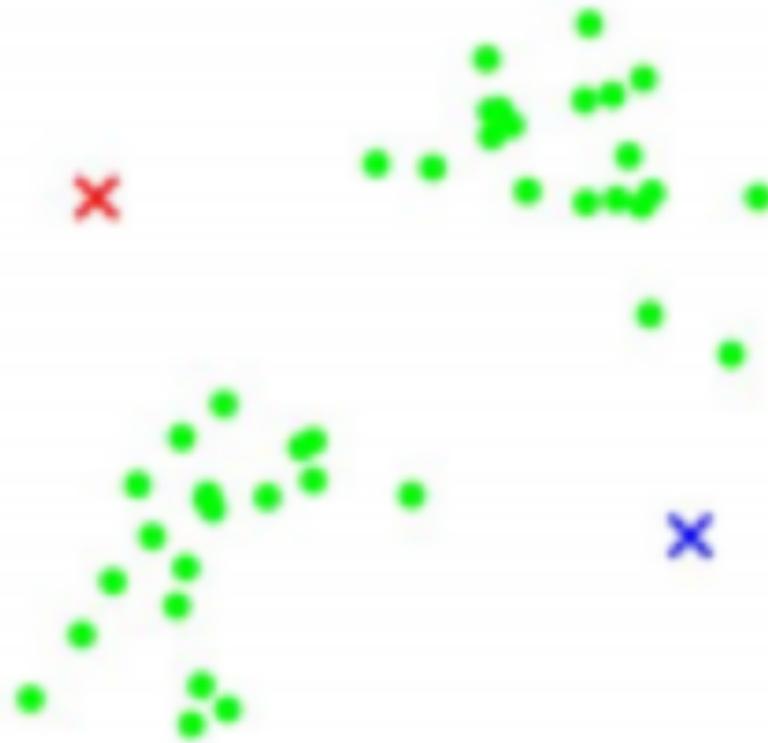
k-means итеративно минимизирует среднее внутрикластерное расстояние:

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

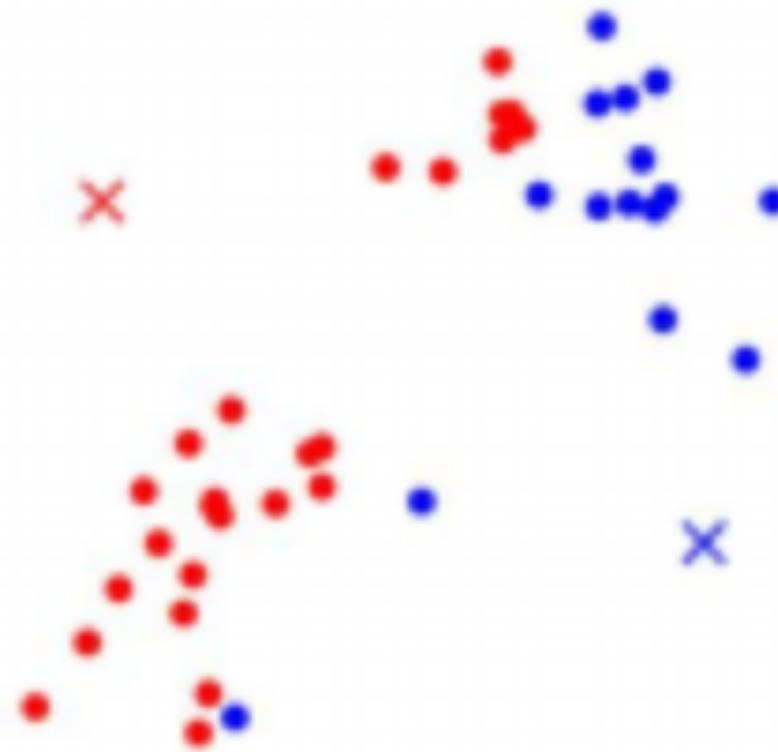
Как работает k-means?



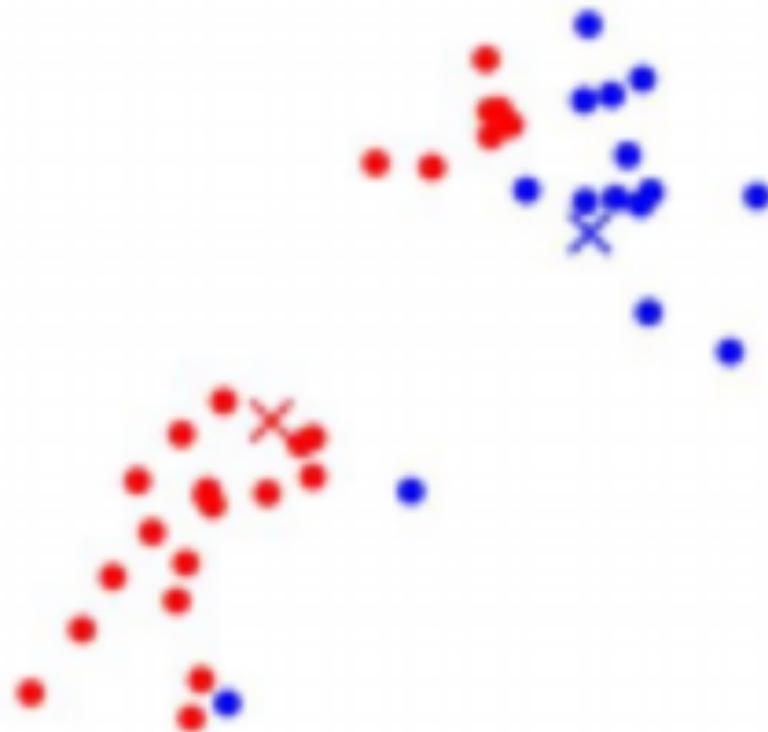
Как работает k-means?



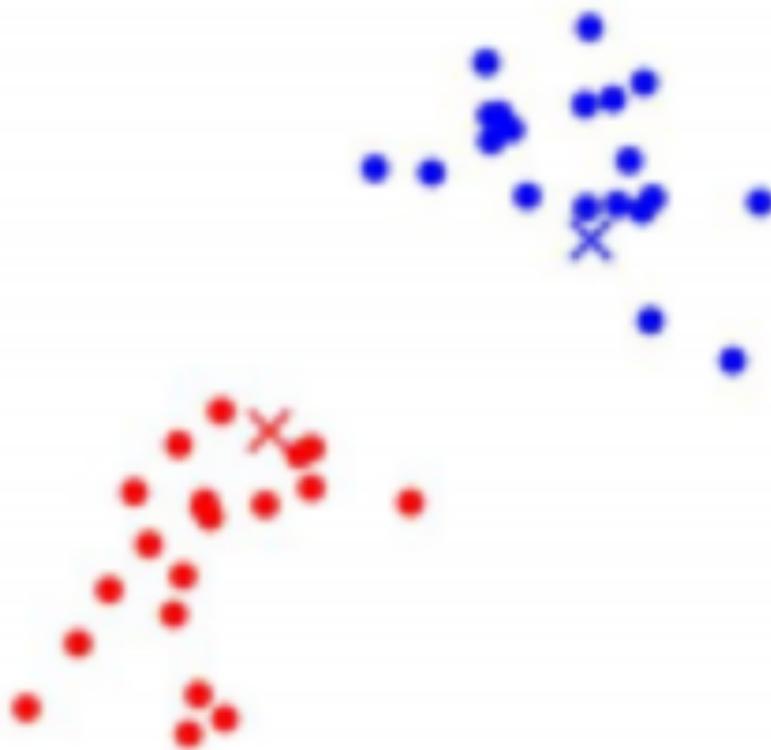
Как работает k-means?



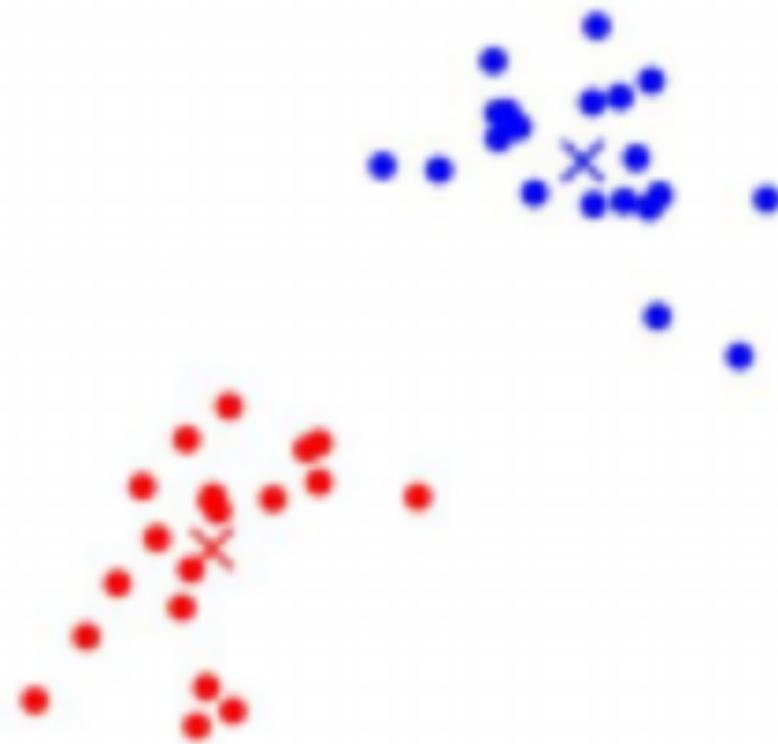
Как работает k-means?



Как работает k-means?



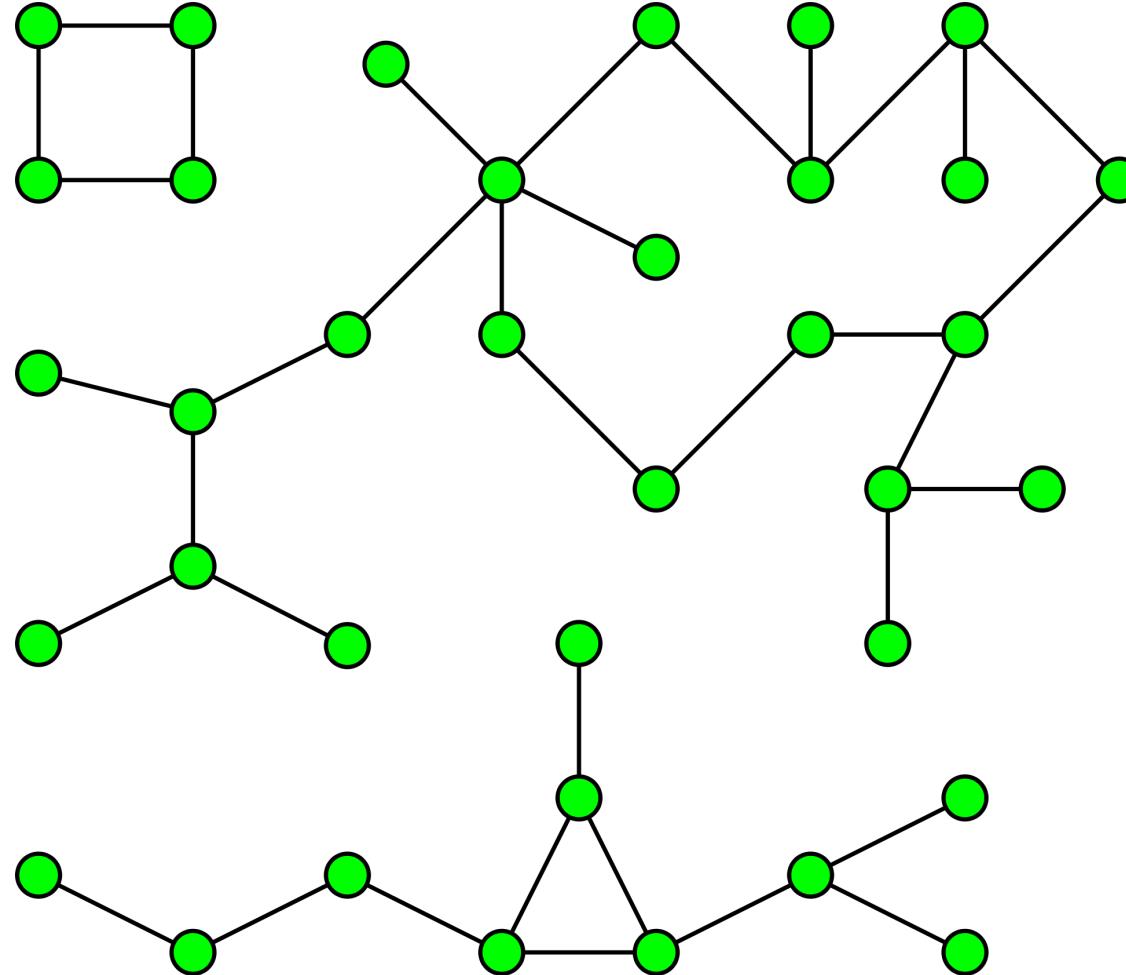
Как работает k-means?



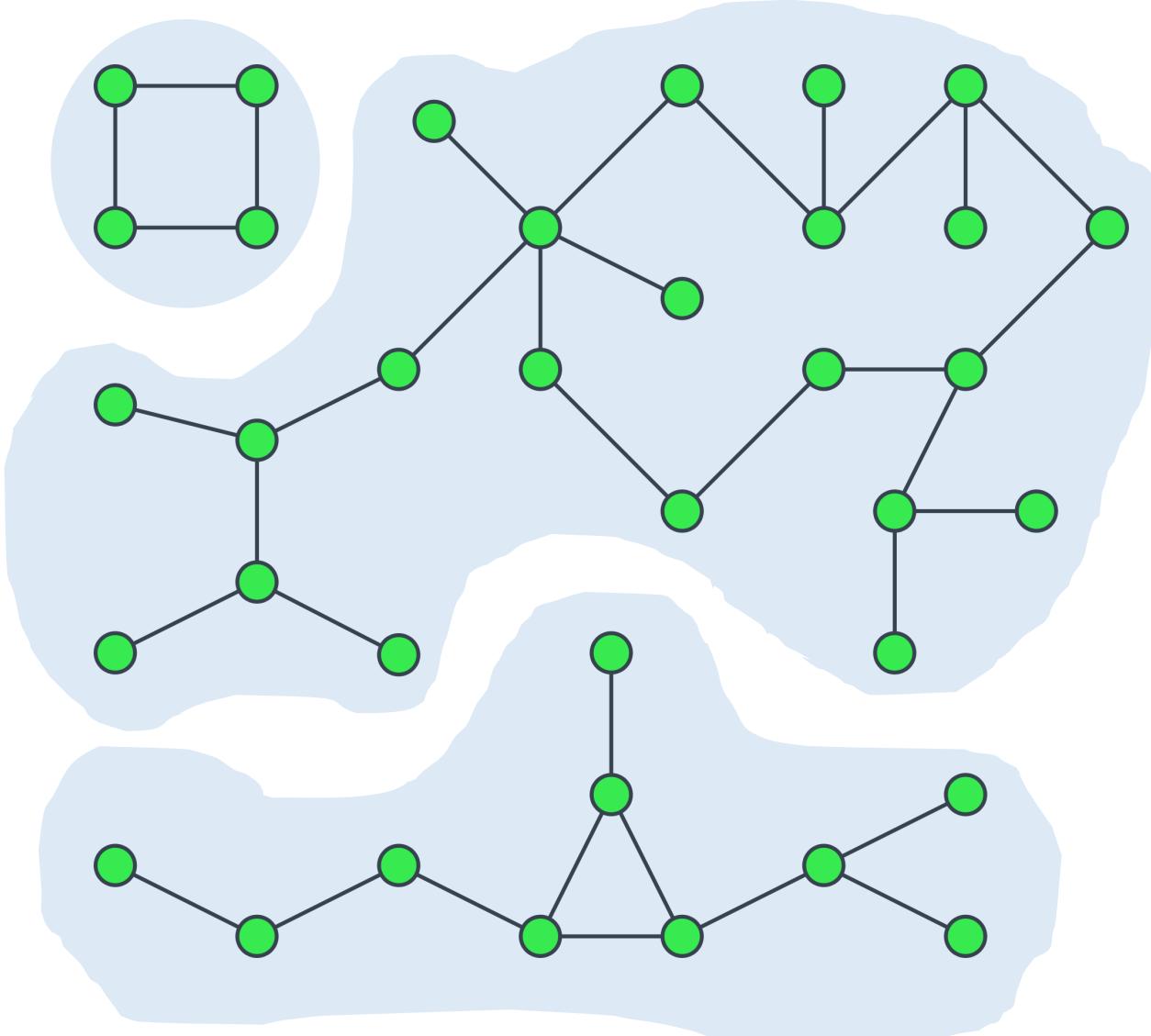
Графовые алгоритмы

- Алгоритмы на основе связных компонент
- Алгоритм на основе оставного дерева

Выделение связных компонент



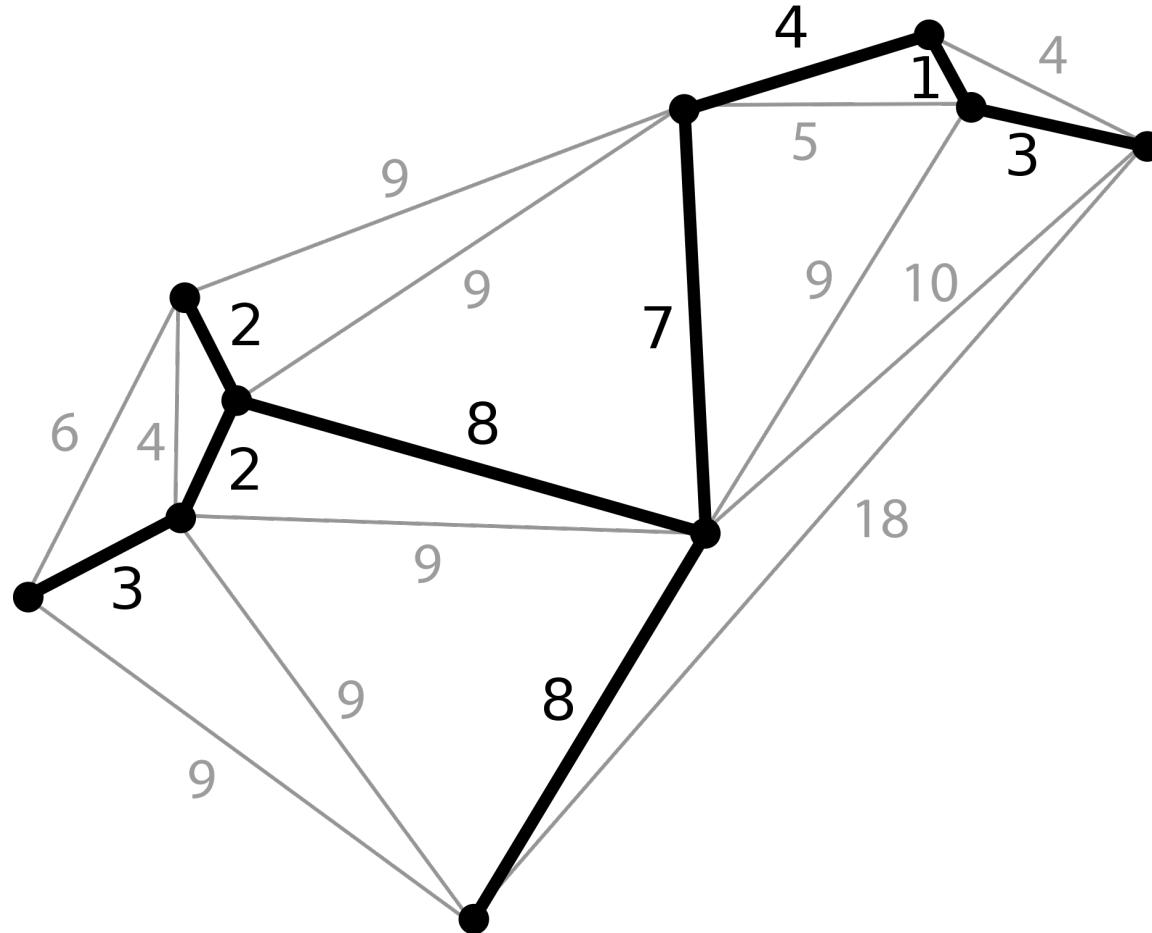
Выделение связных компонент



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 оставное 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 вершина 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

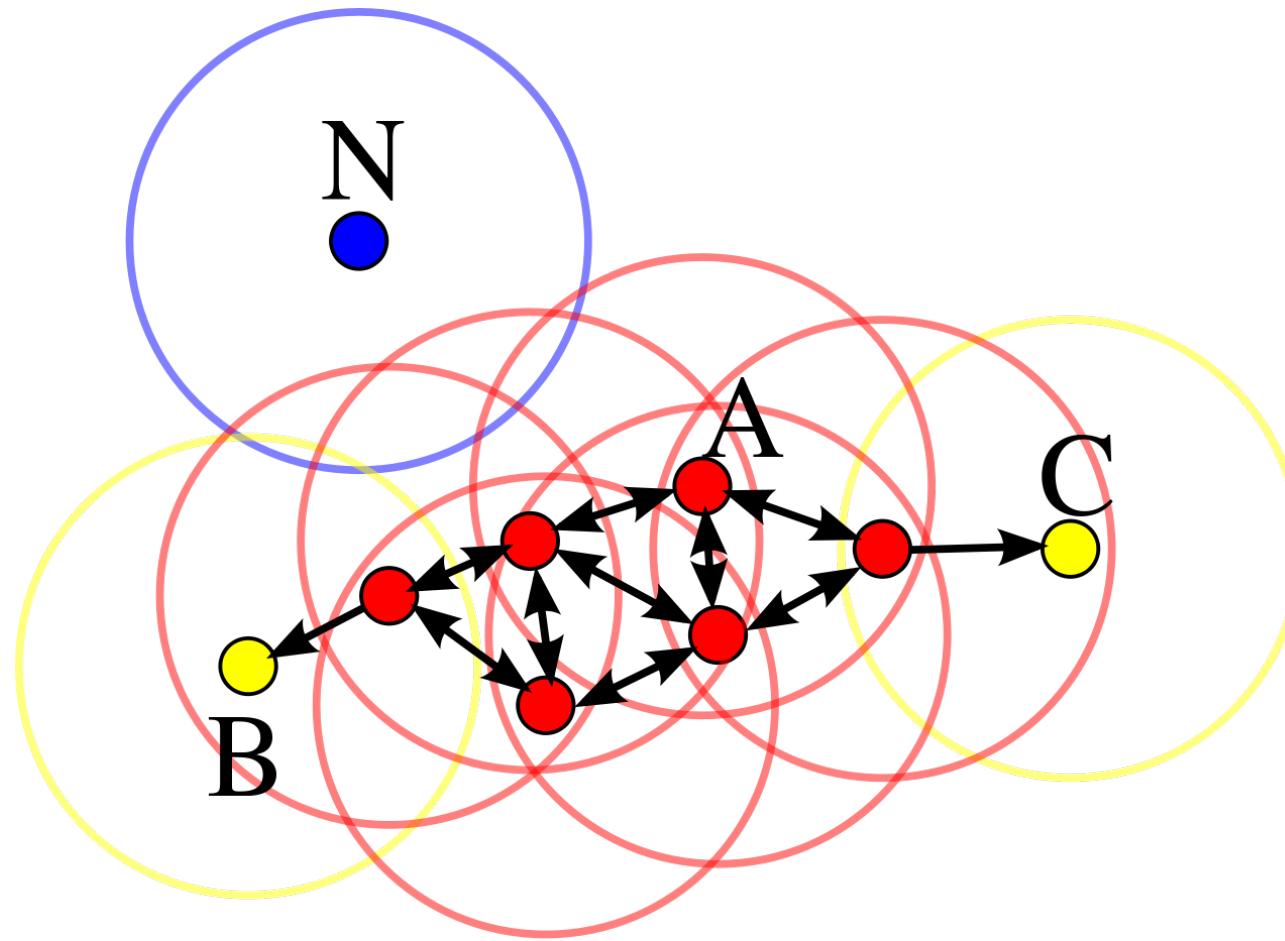
Кластеризация с помощью минимального оствового дерева

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

Density-based методы

- Основываются на плотности точек
- DBSCAN - Density-based spatial clustering of applications with noise

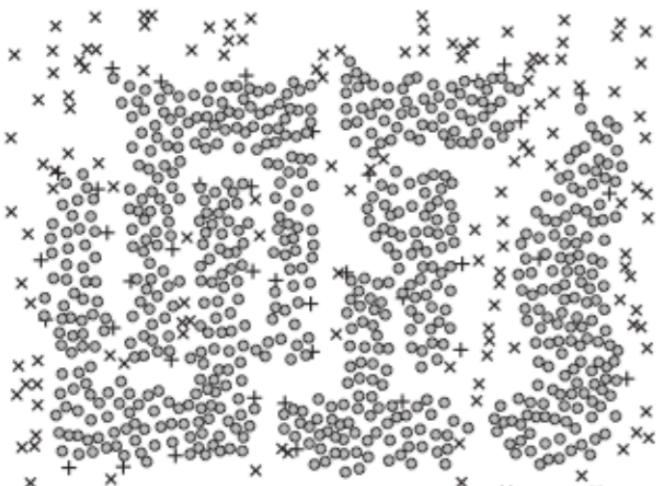
Основные, шумовые и граничные точки



DBSCAN



(a) Clusters found by DBSCAN.

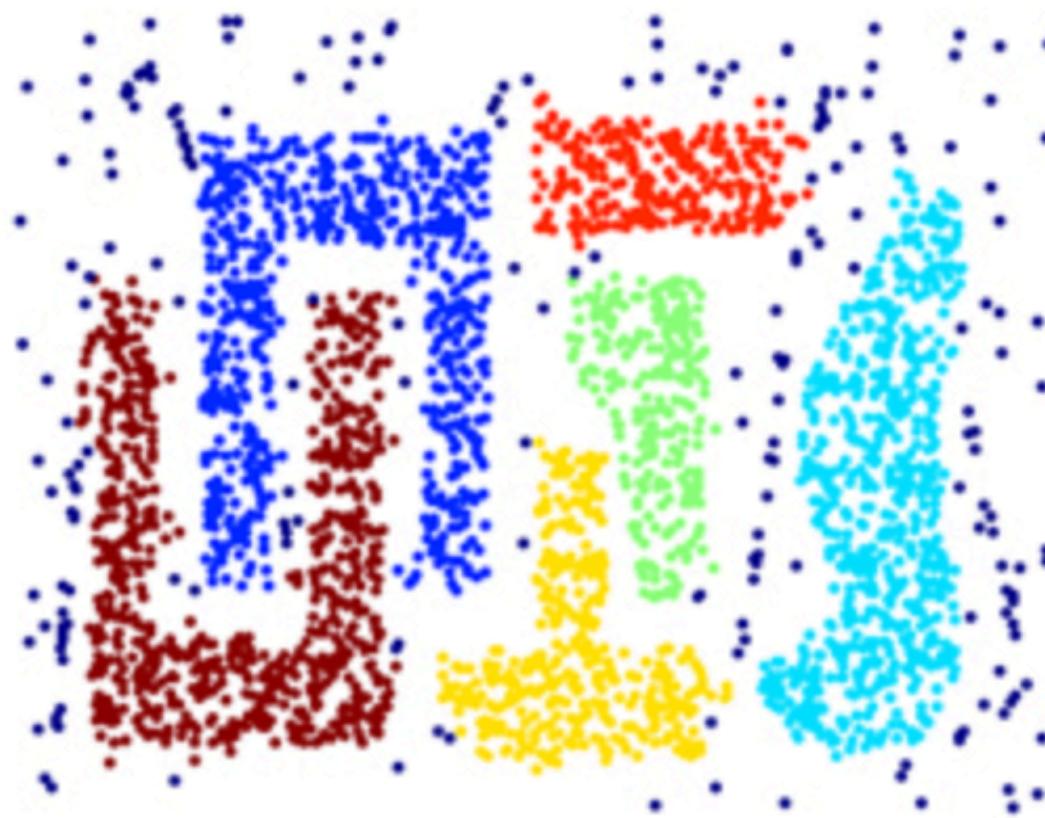


x – Noise Point + – Border Point o – Core Point

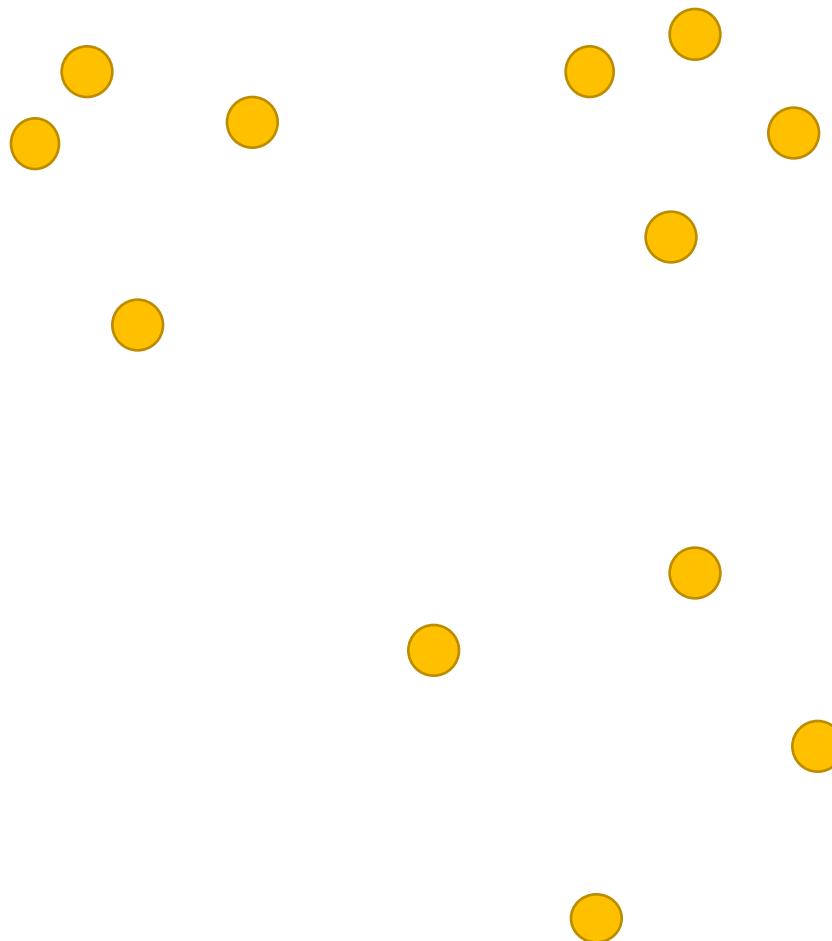
(b) Core, border, and noise points.

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

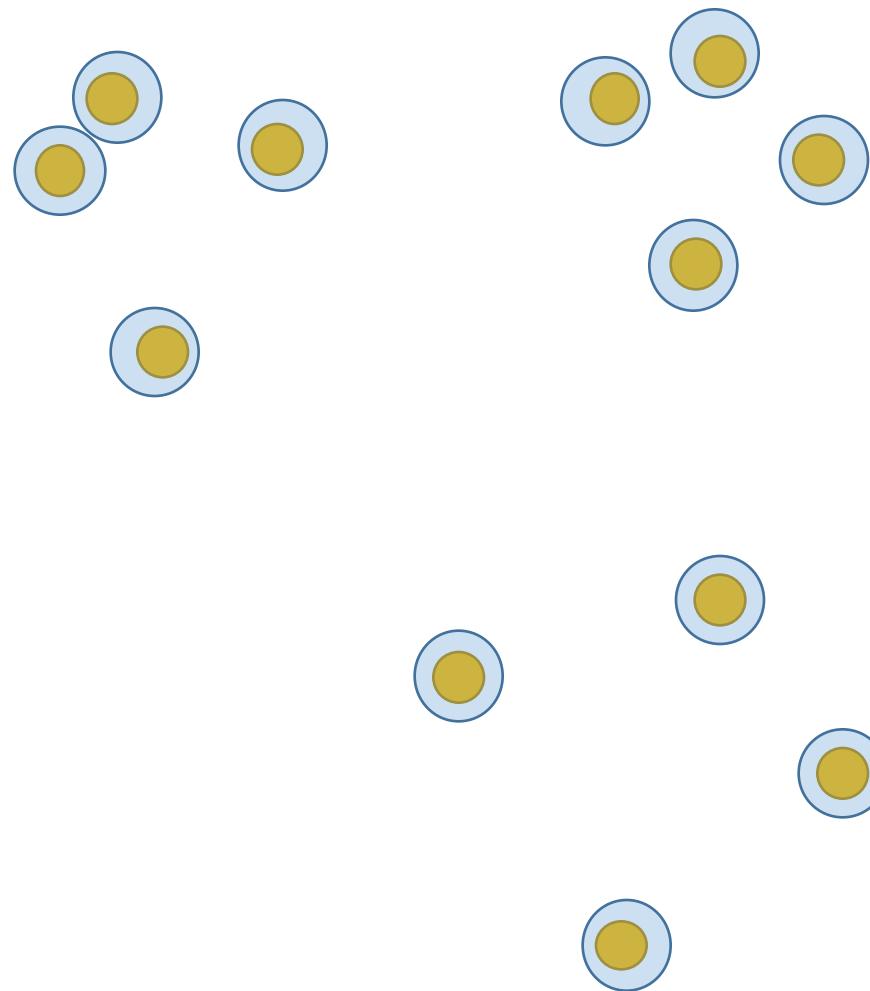
DBSCAN: результаты работы



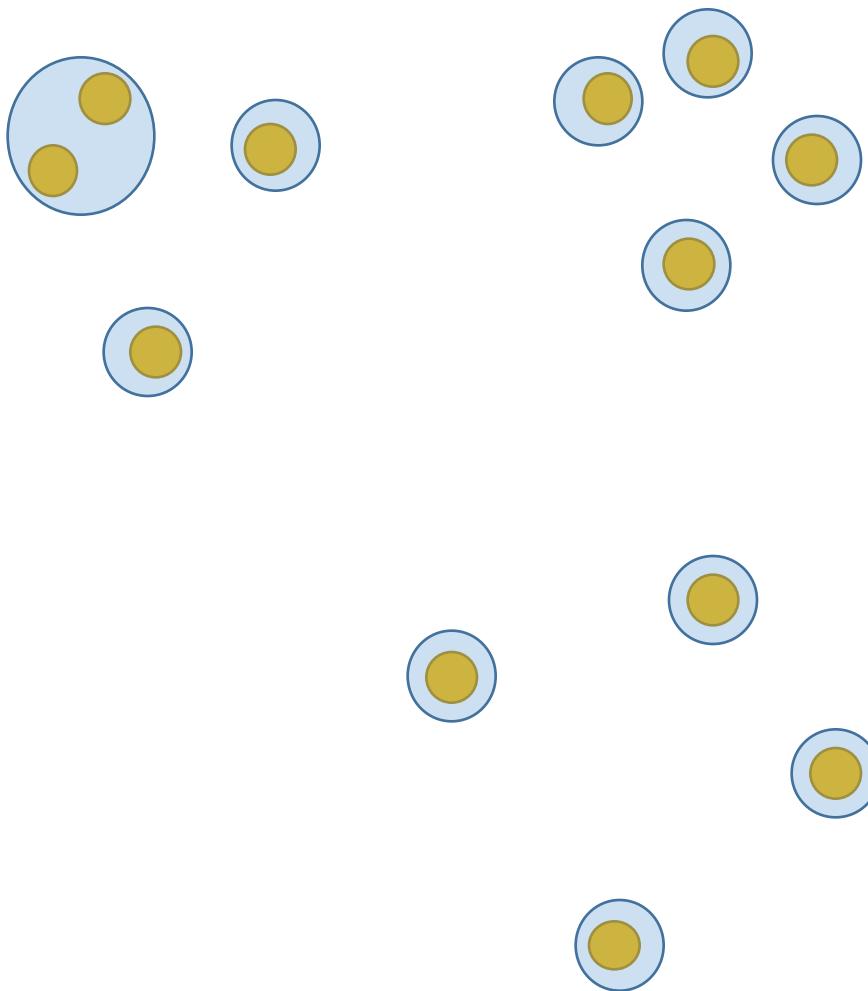
Иерархическая агломеративная кластеризация



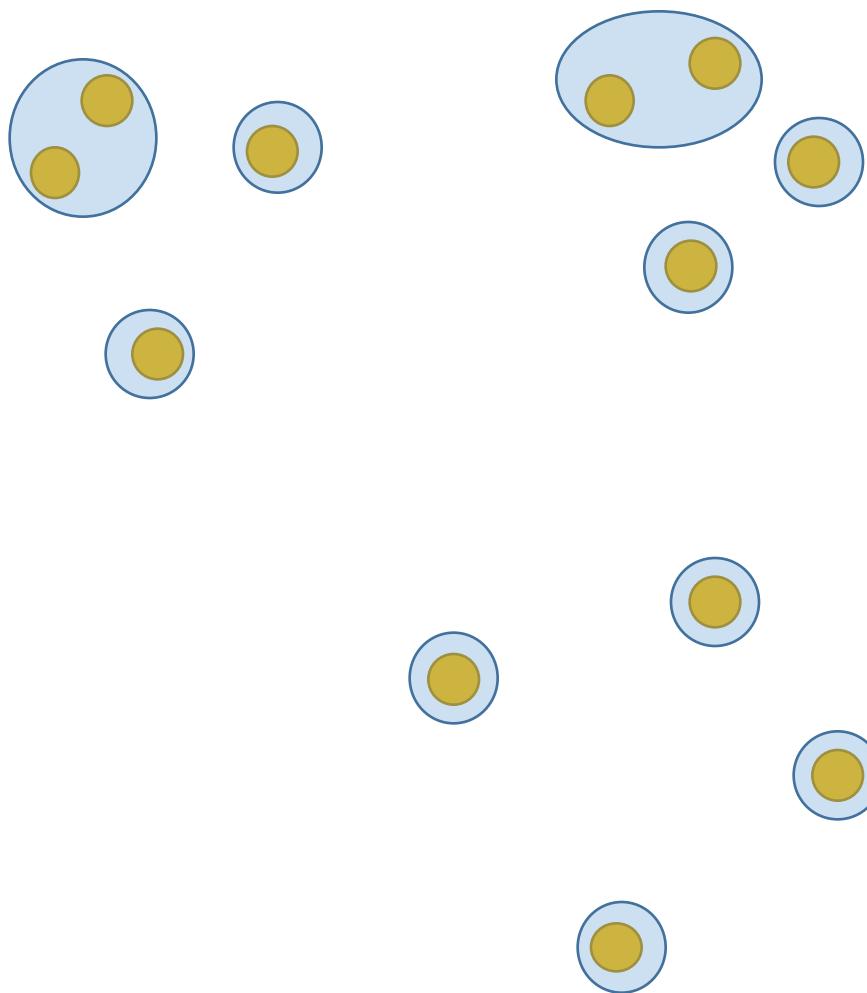
Иерархическая агломеративная кластеризация



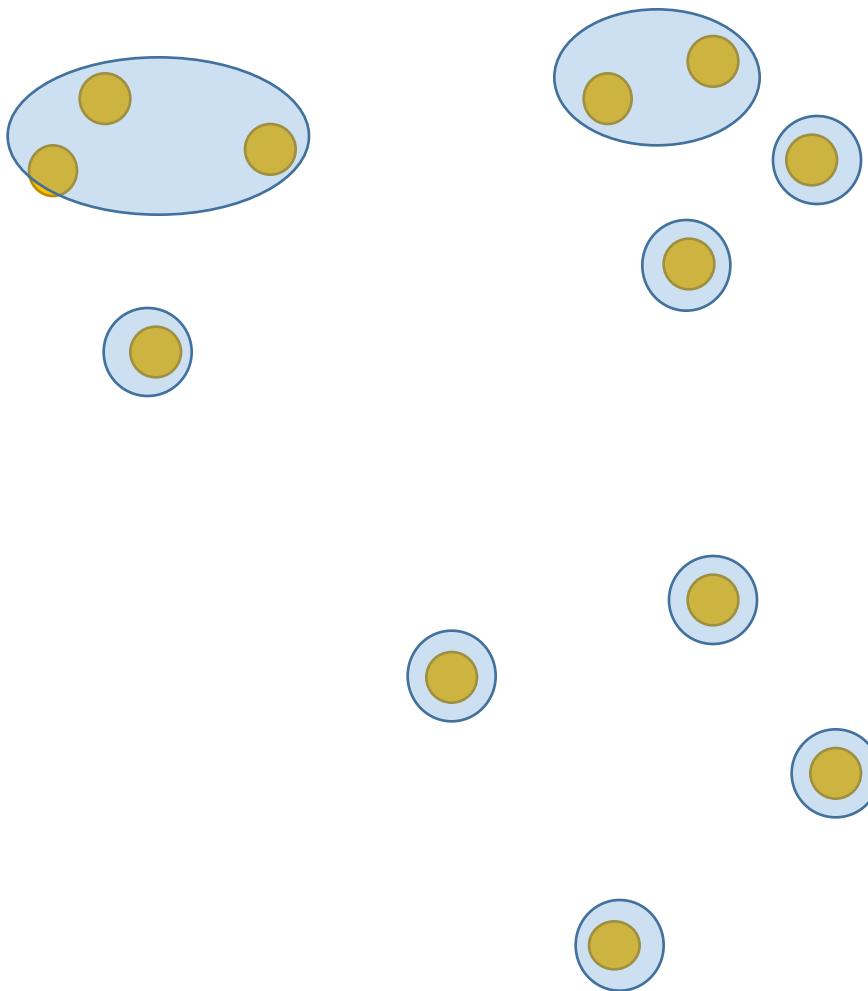
Иерархическая агломеративная кластеризация



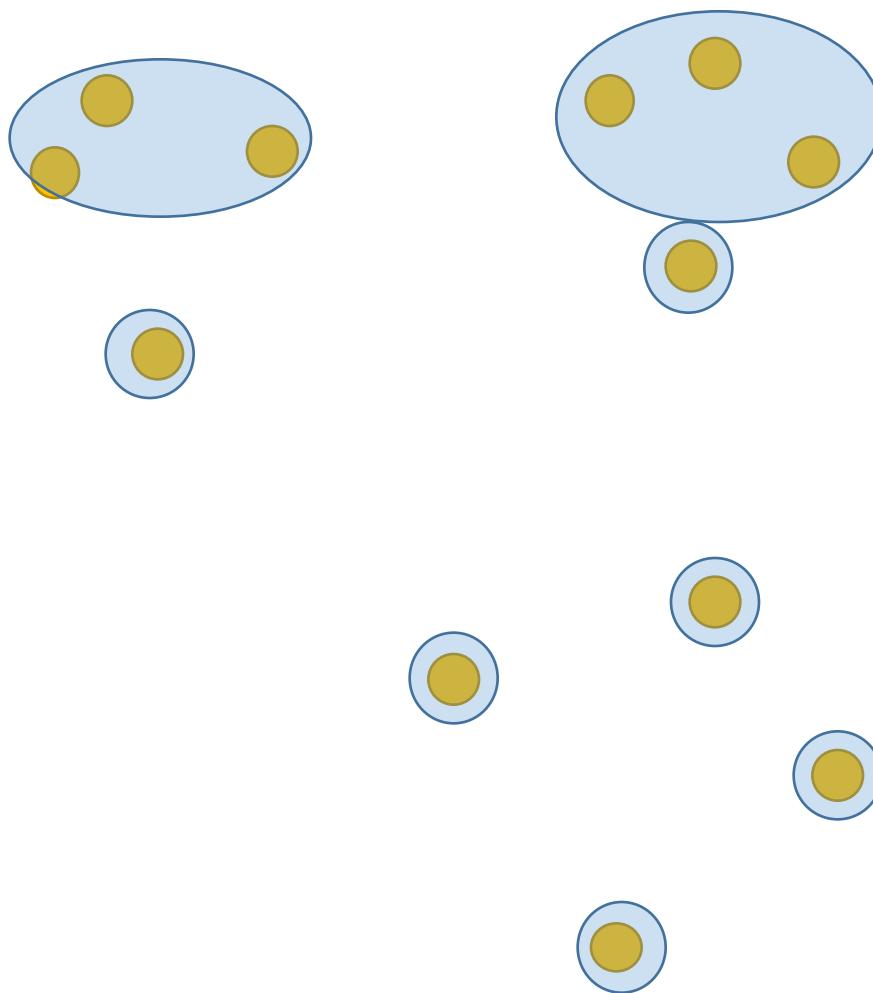
Иерархическая агломеративная кластеризация



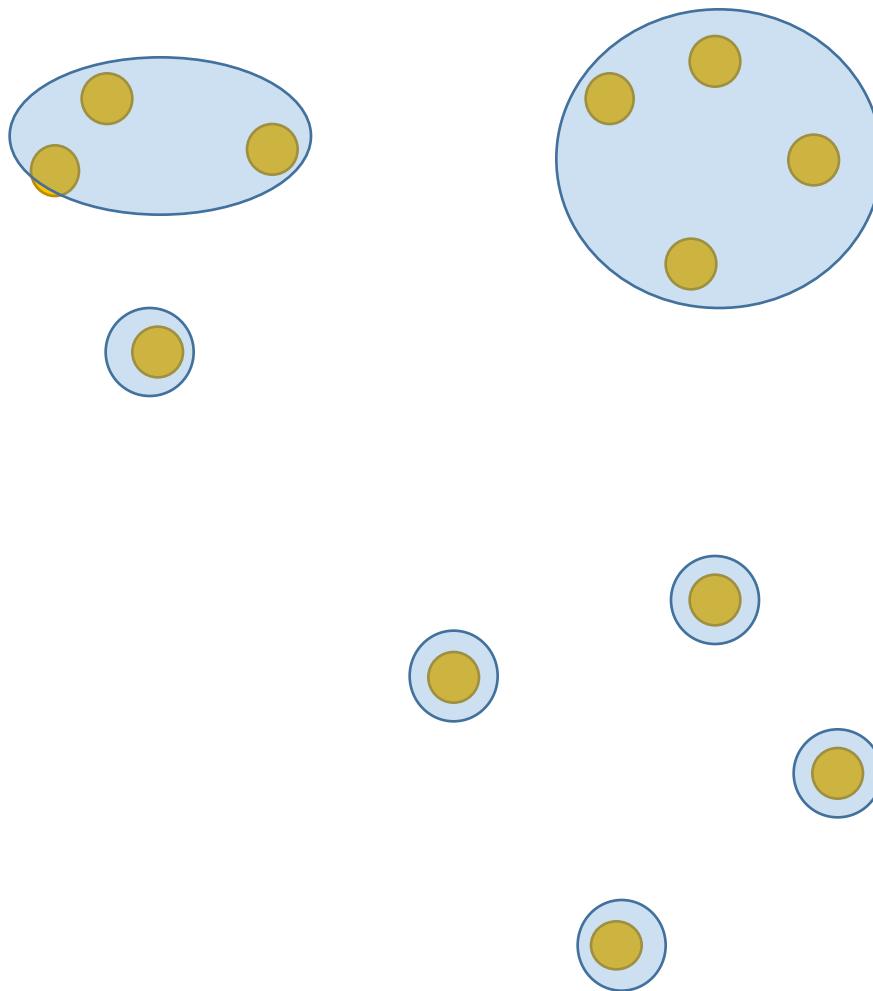
Иерархическая агломеративная кластеризация



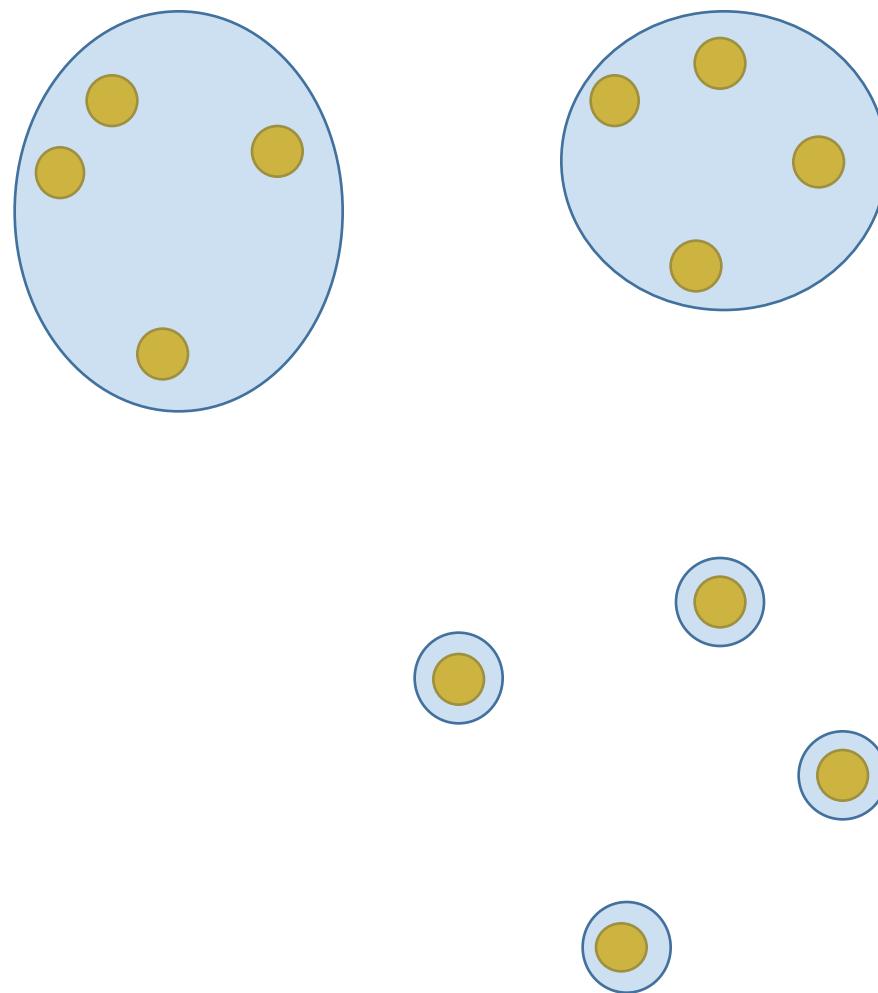
Иерархическая агломеративная кластеризация



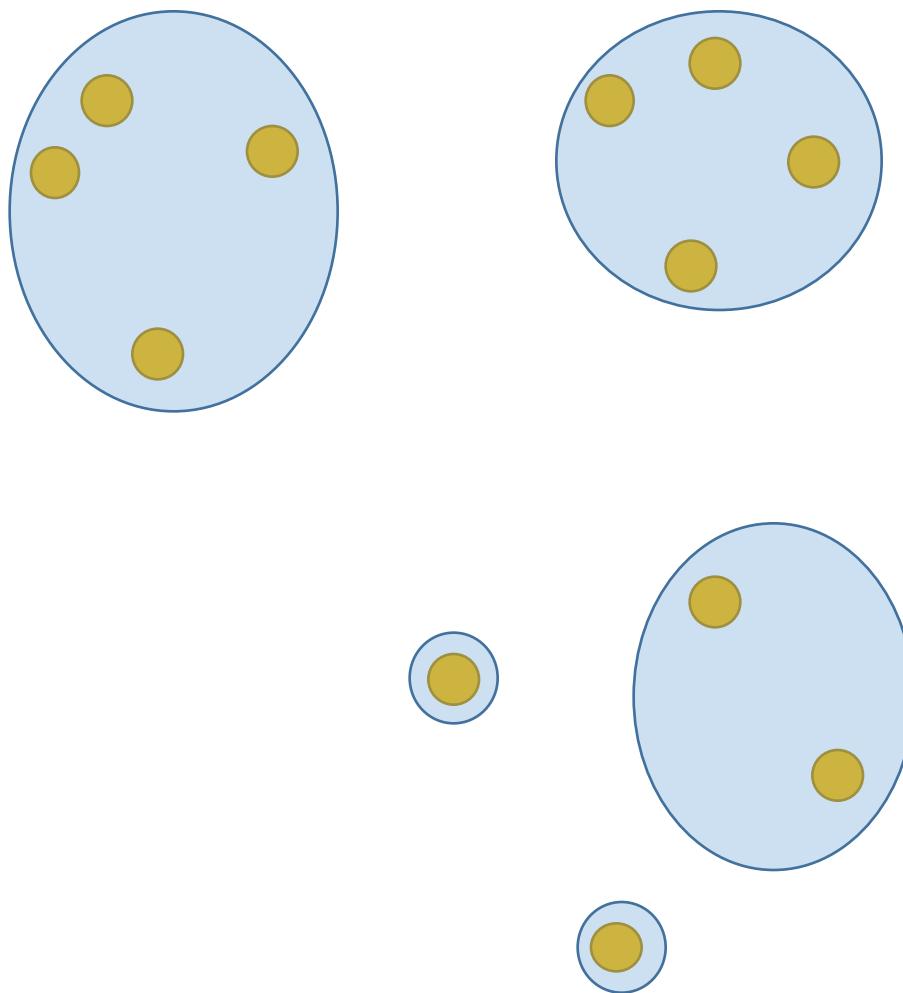
Иерархическая агломеративная кластеризация



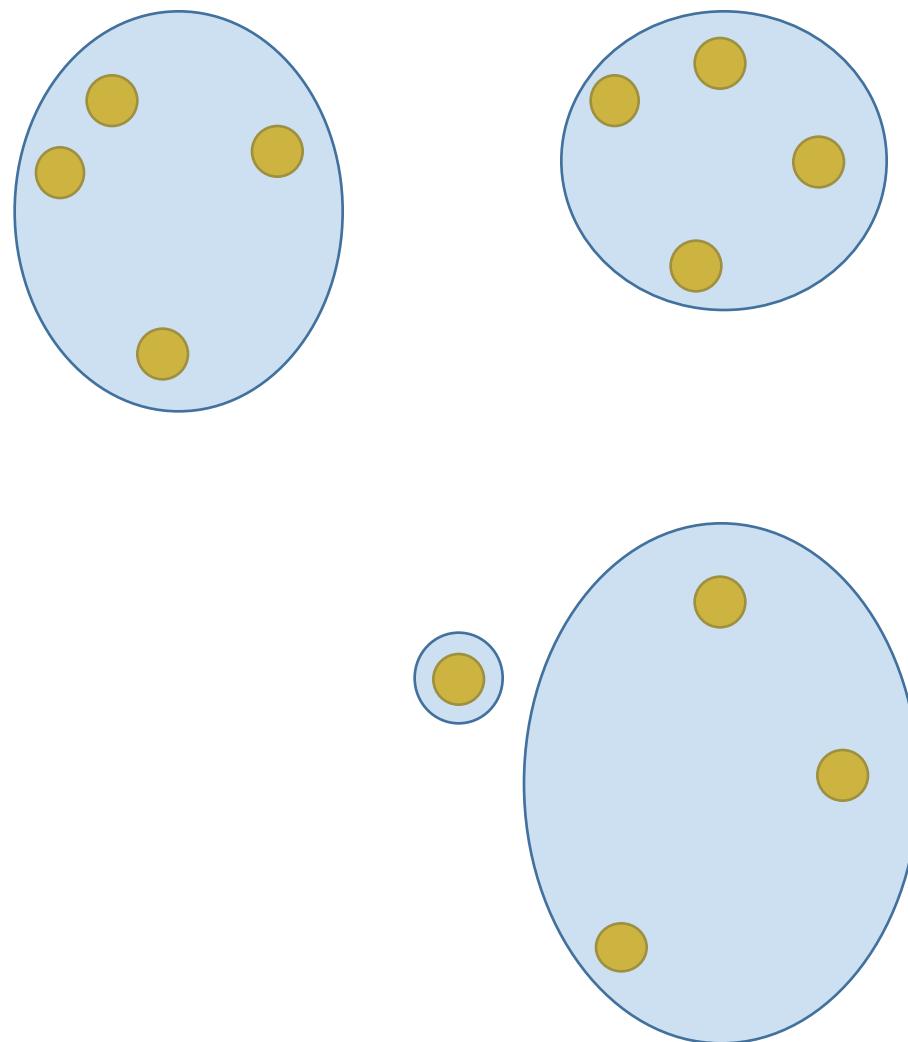
Иерархическая агломеративная кластеризация



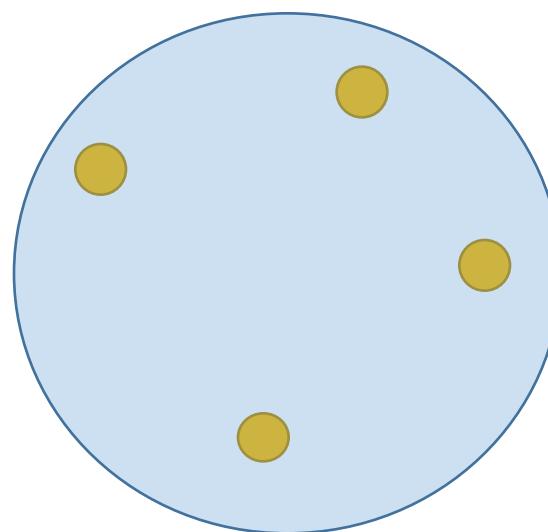
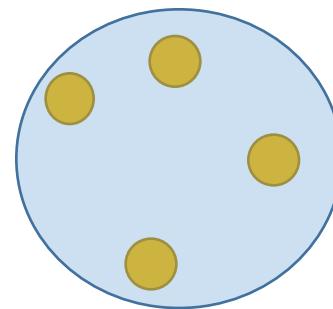
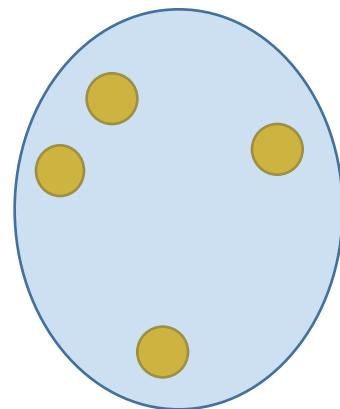
Иерархическая агломеративная кластеризация



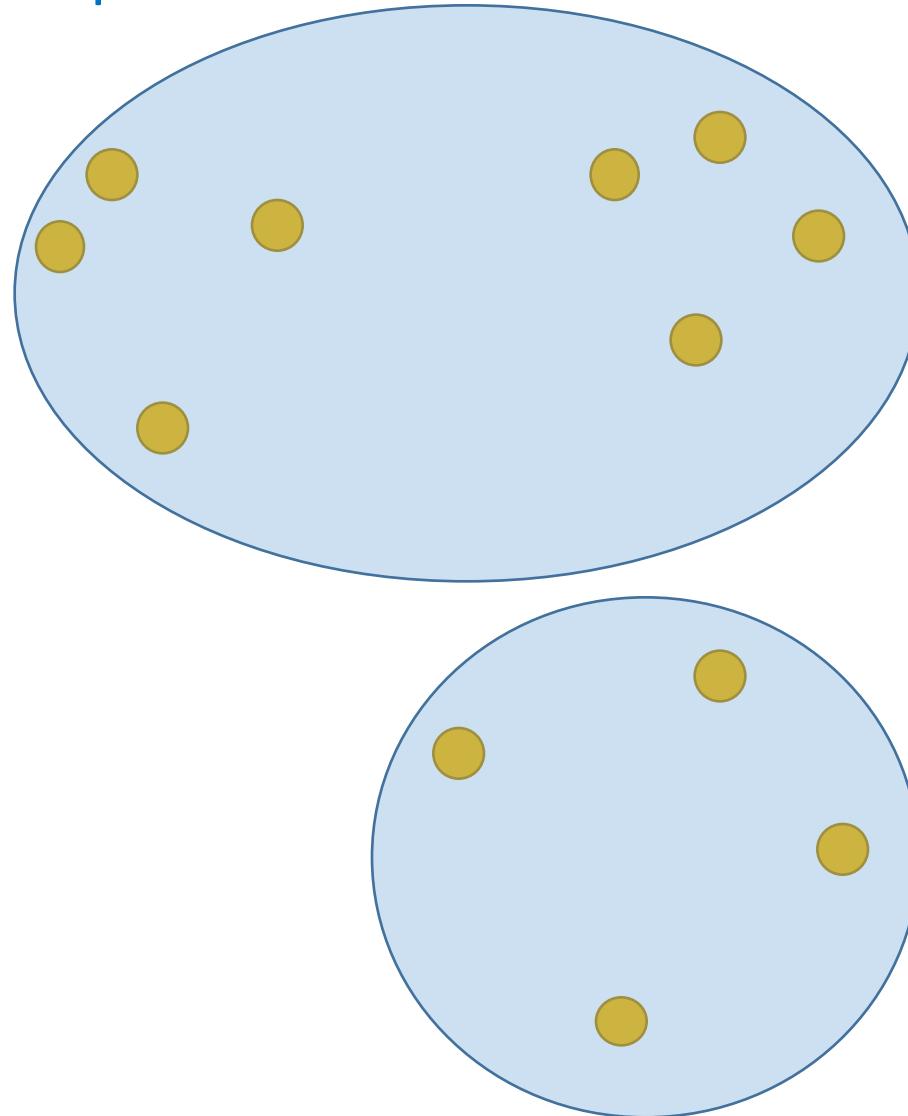
Иерархическая агломеративная кластеризация



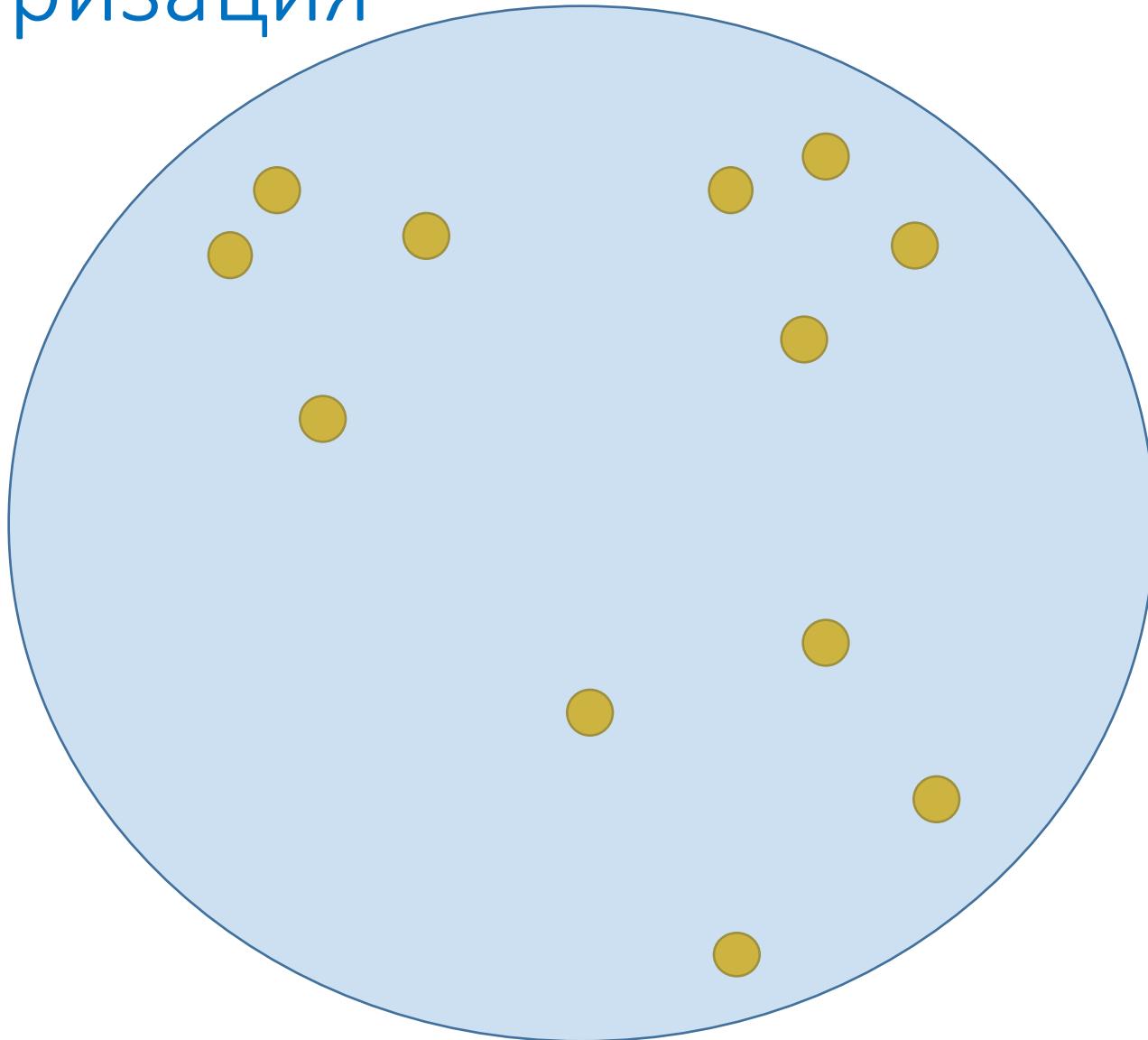
Иерархическая агломеративная кластеризация



Иерархическая агломеративная кластеризация



Иерархическая агломеративная кластеризация



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:

$$R^6(W, S) = \min_{w \in W, s \in S} \rho(w, s); \quad \alpha_U = \alpha_V = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}.$$

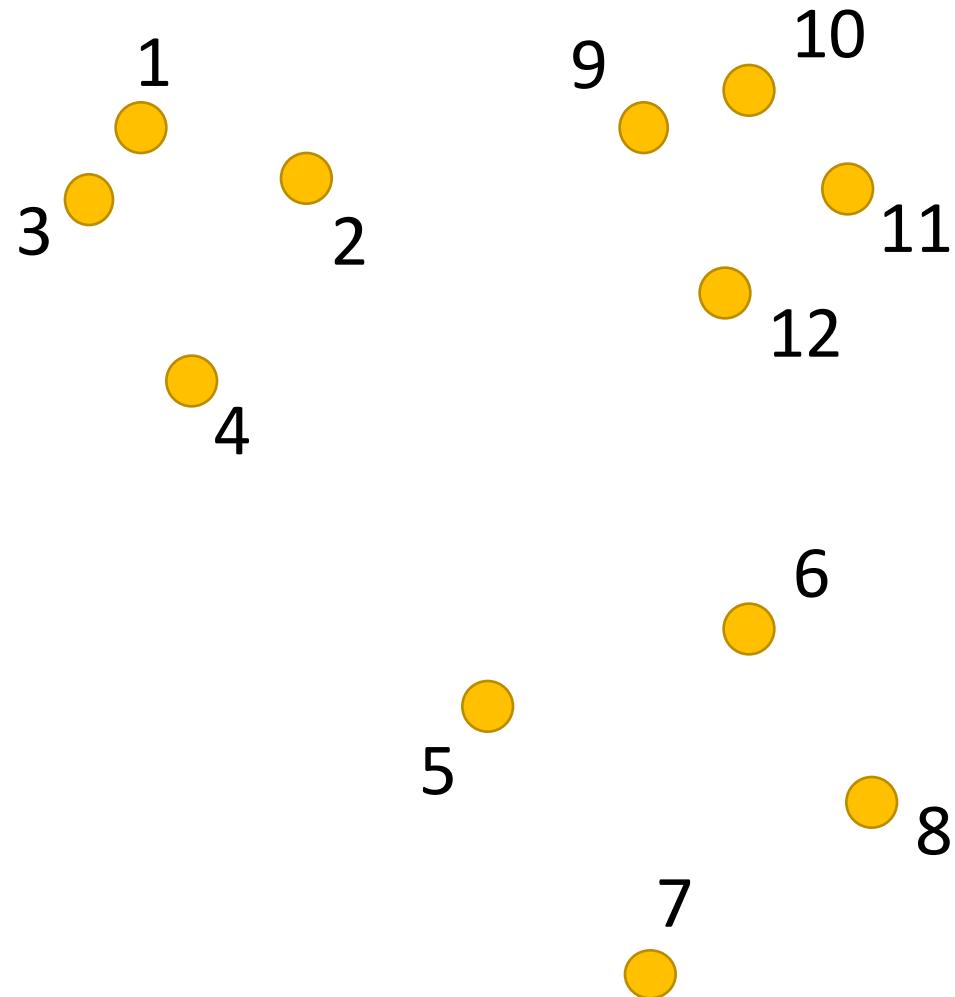
Расстояние дальнего соседа:

$$R^\Delta(W, S) = \max_{w \in W, s \in S} \rho(w, s); \quad \alpha_U = \alpha_V = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}.$$

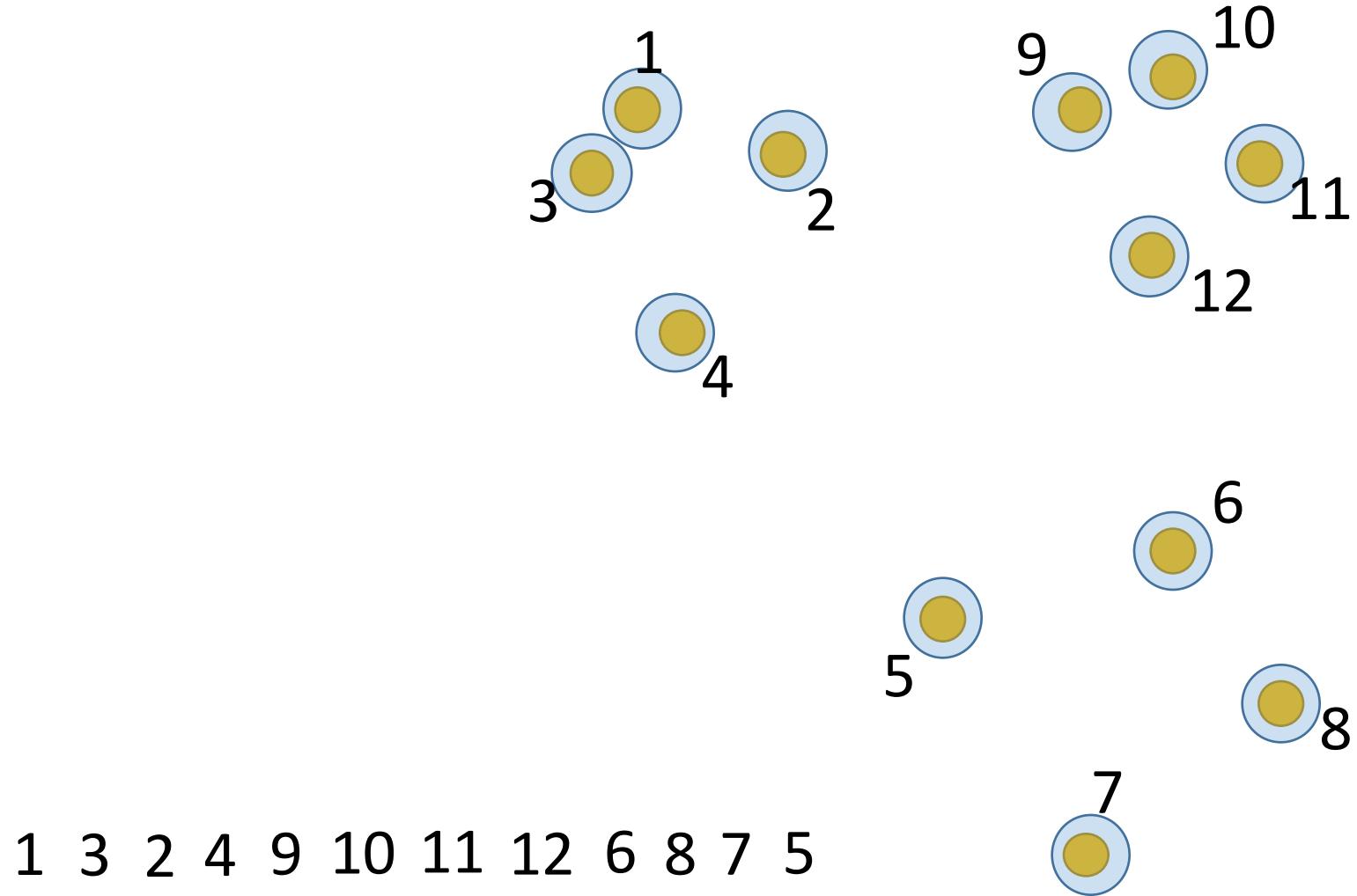
Среднее расстояние:

$$R^c(W, S) = \frac{1}{|W||S|} \sum_{w \in W} \sum_{s \in S} \rho(w, s); \quad \alpha_U = \frac{|U|}{|W|}, \alpha_V = \frac{|V|}{|W|}, \beta = \gamma = 0.$$

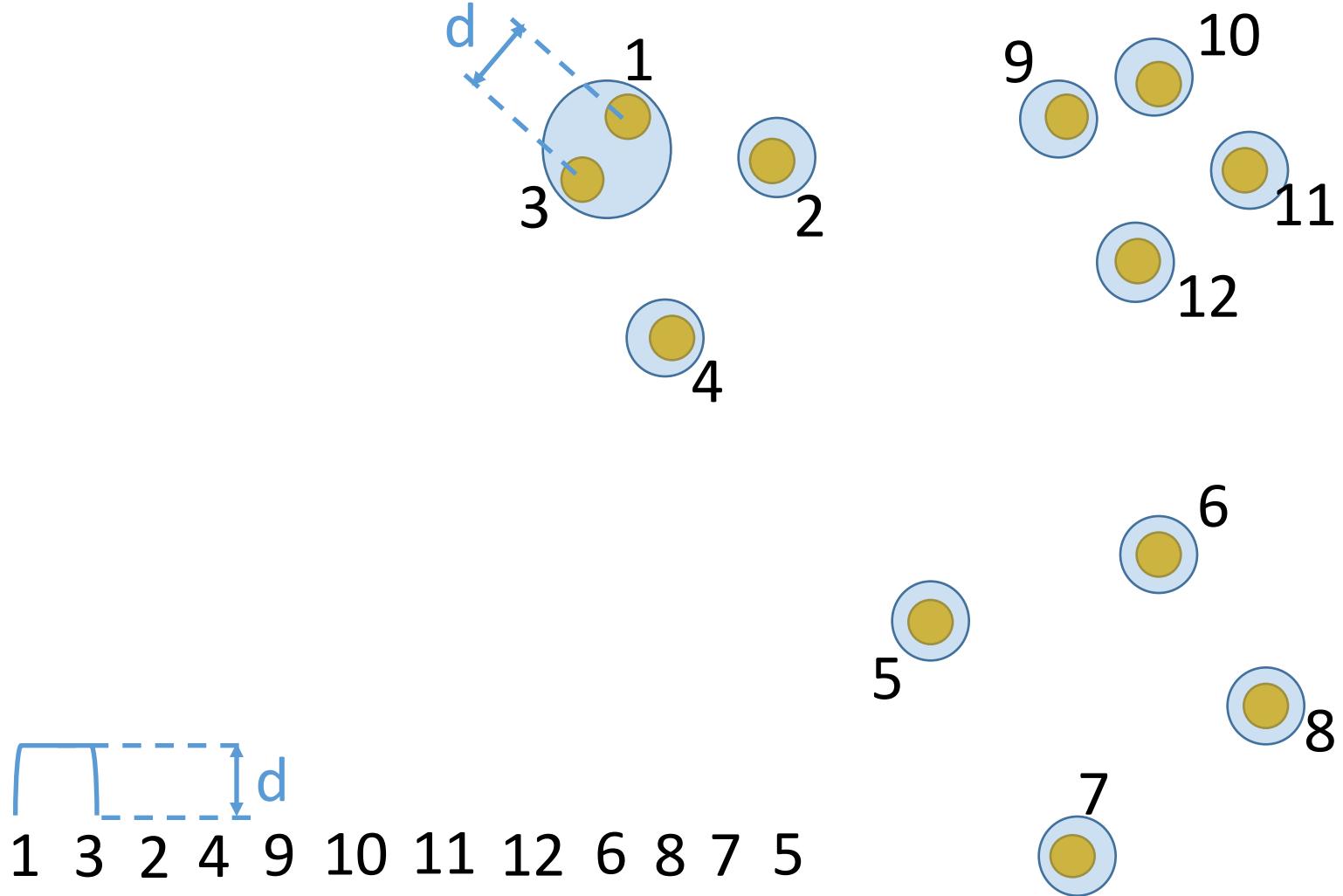
Дендрограмма



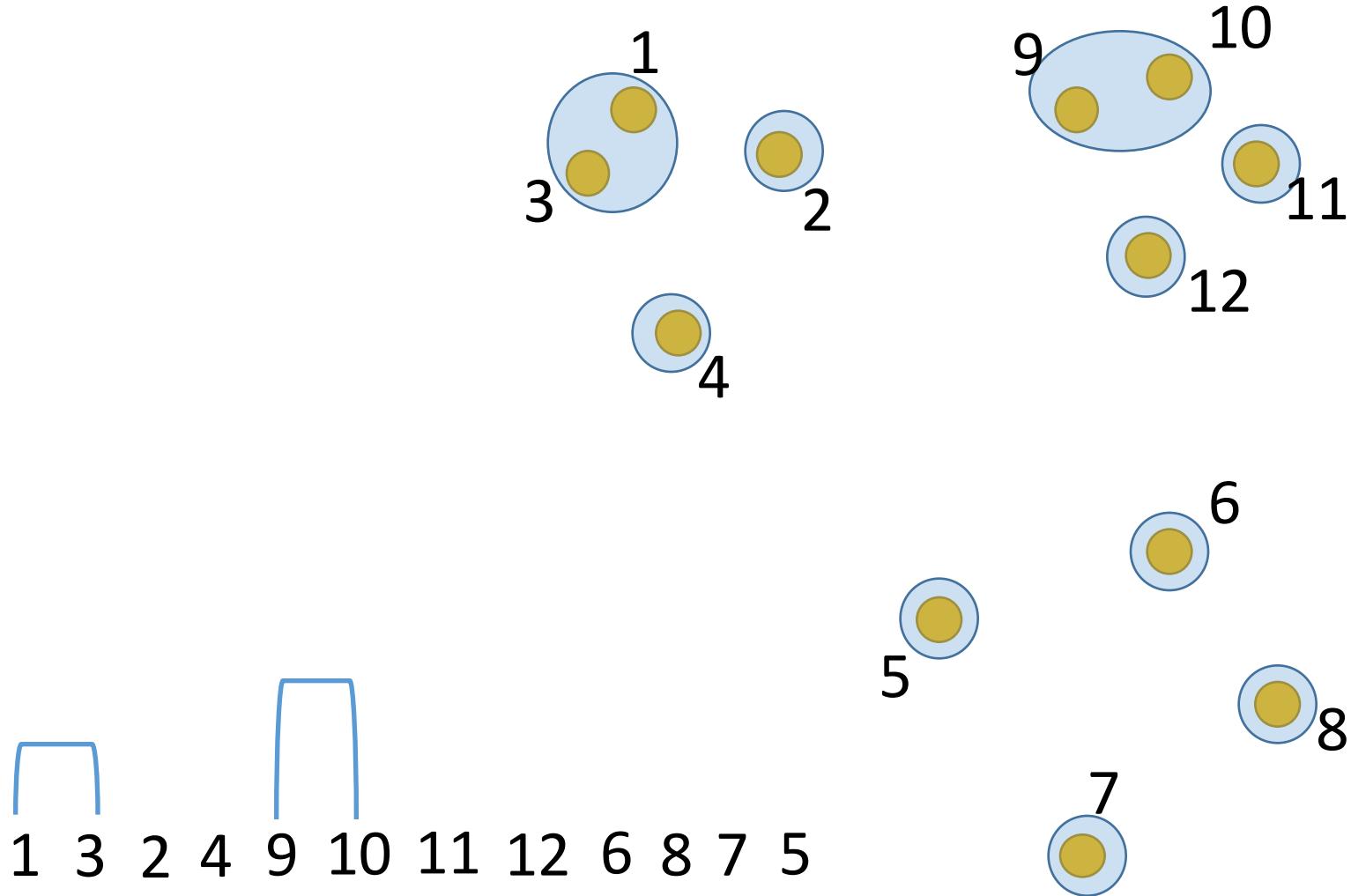
Дендрограмма



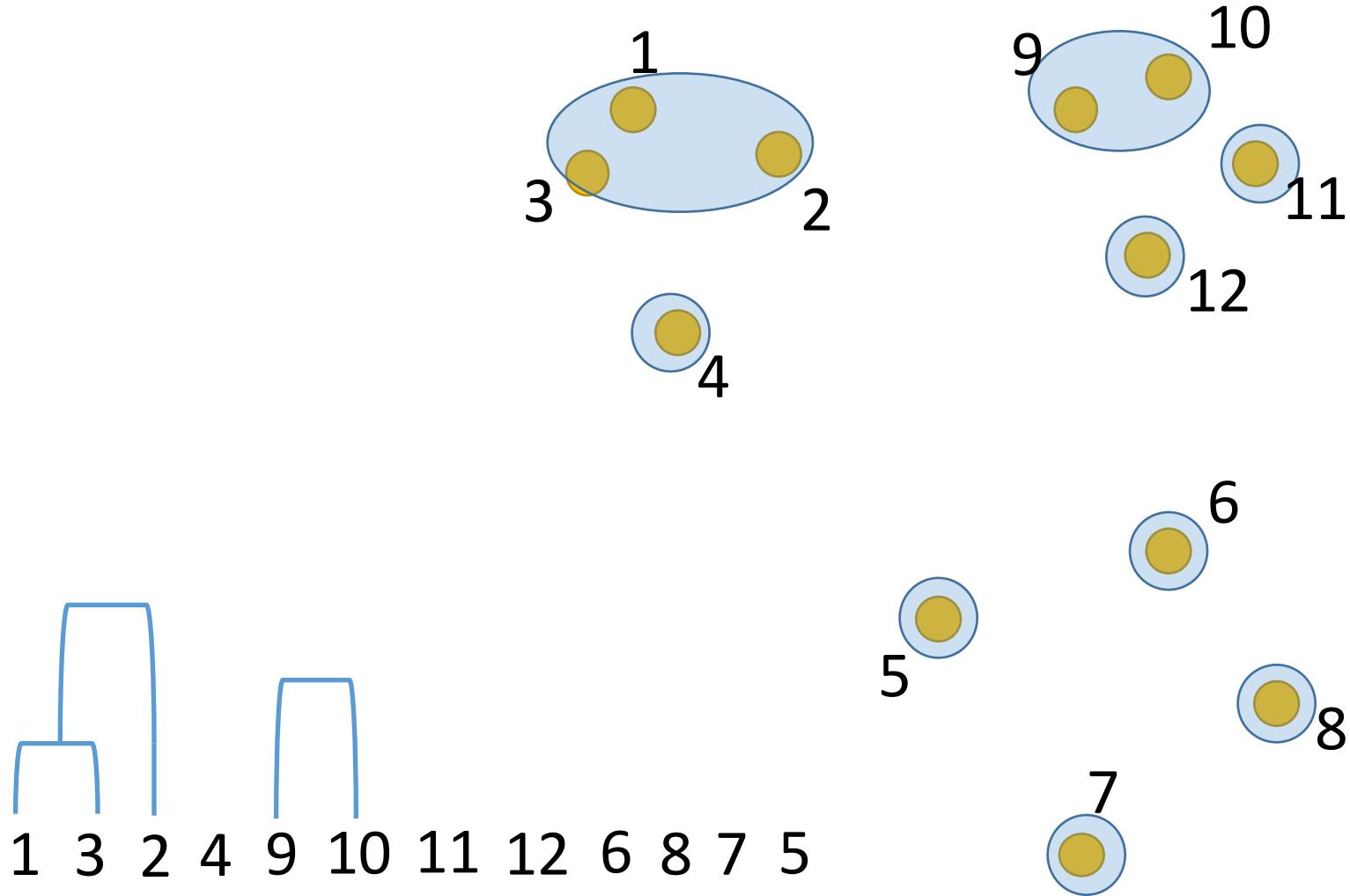
Дендрограмма



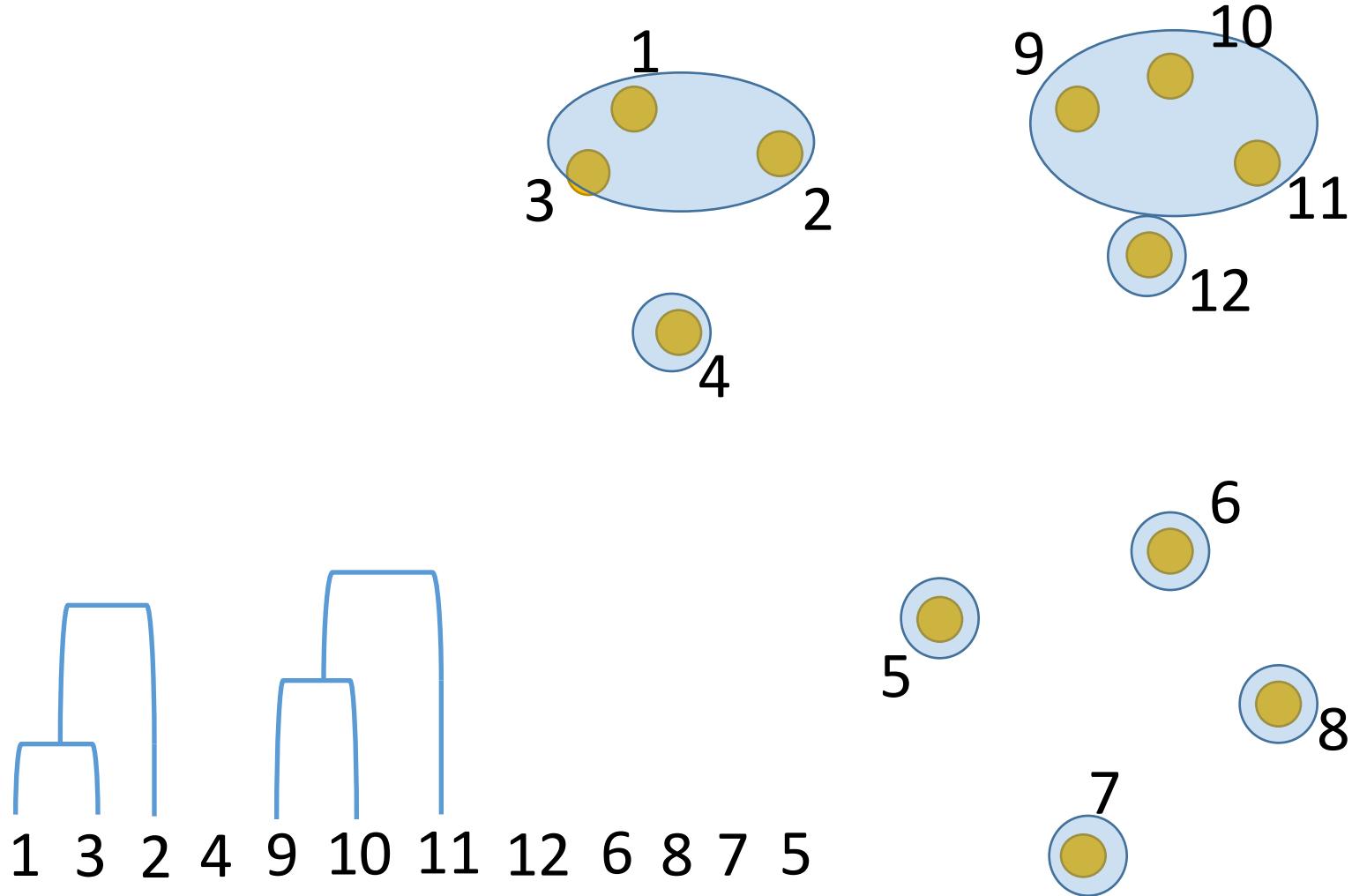
Дендрограмма



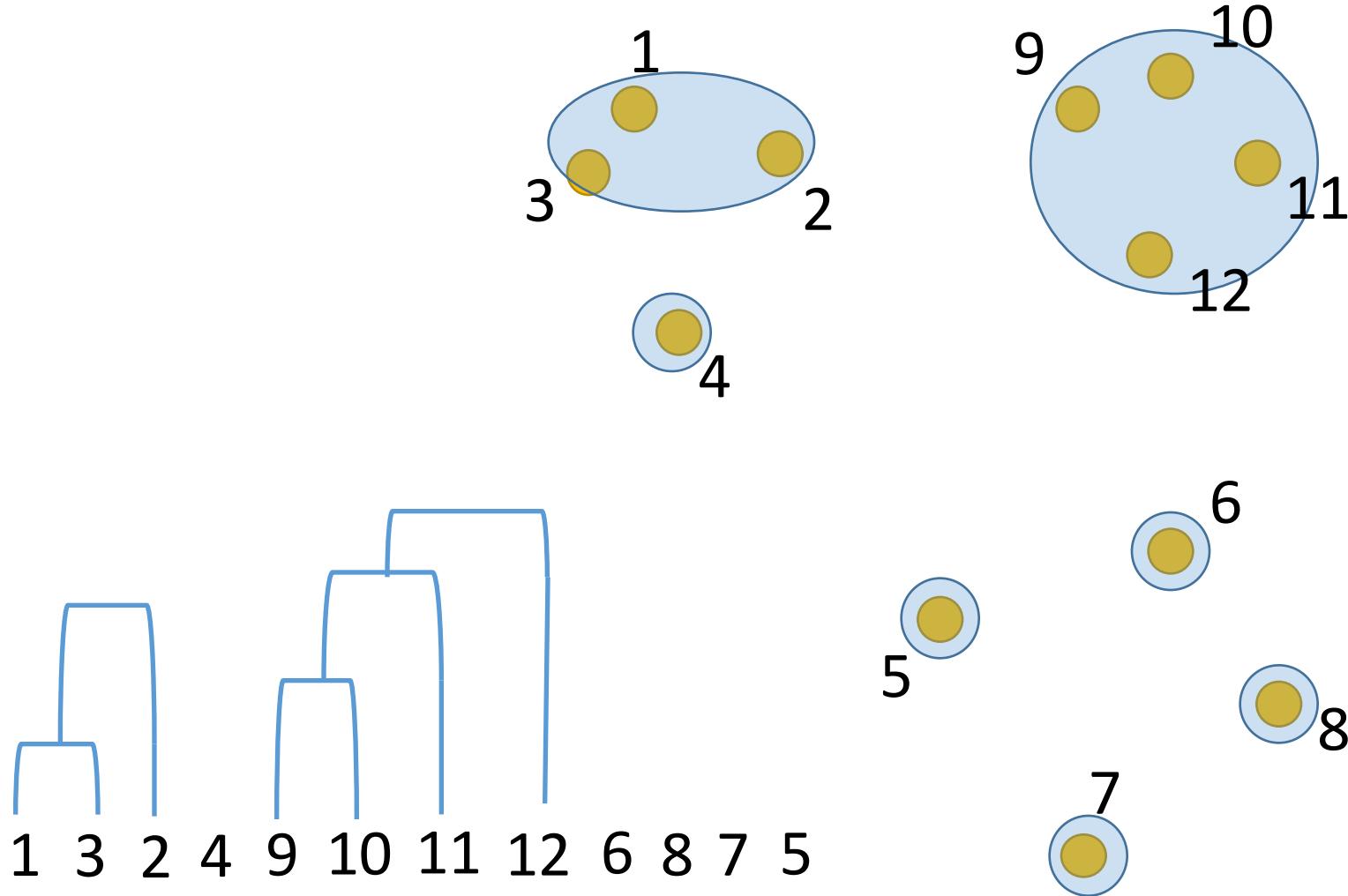
Дендрограмма



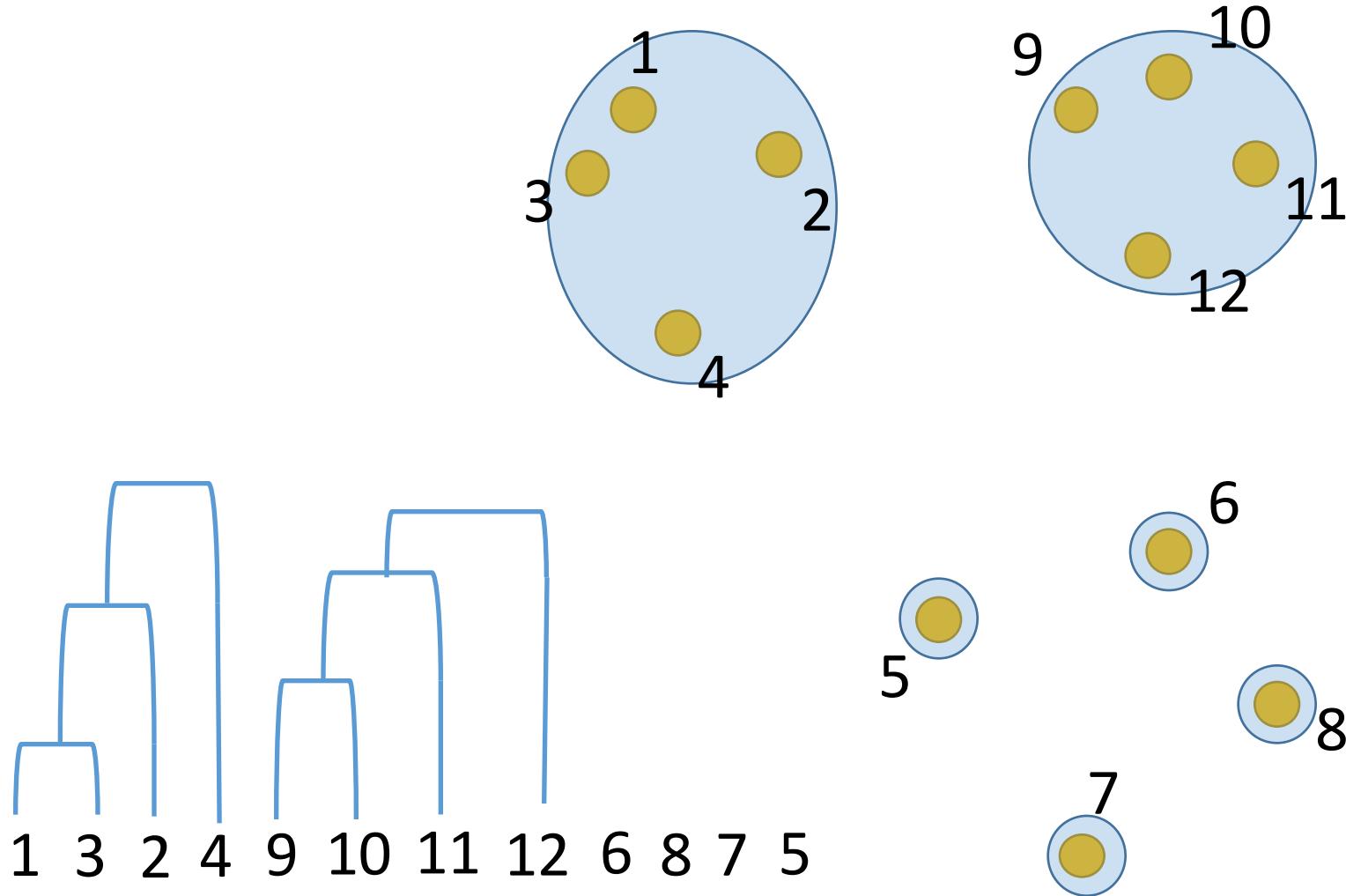
Дендрограмма



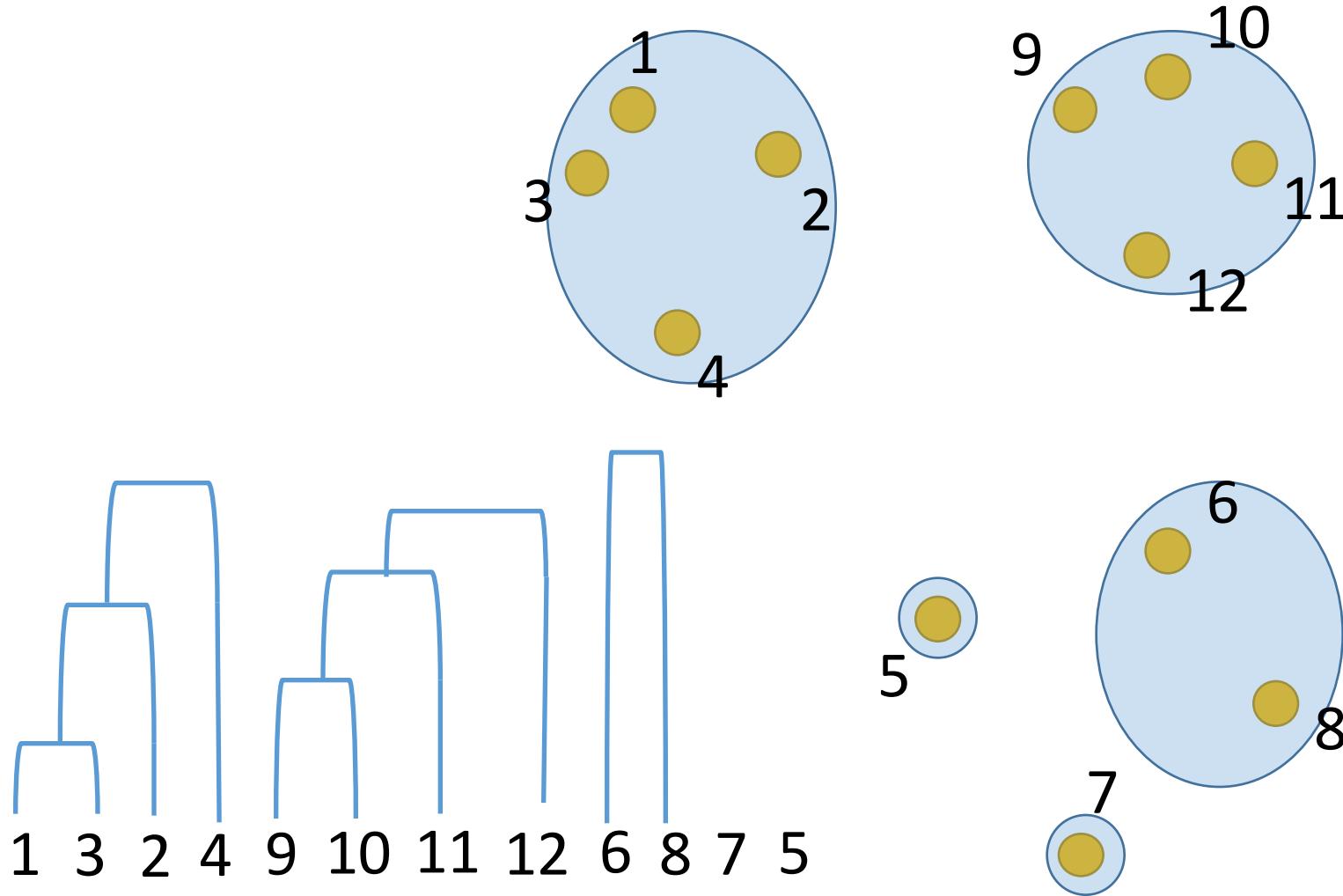
Дендрограмма



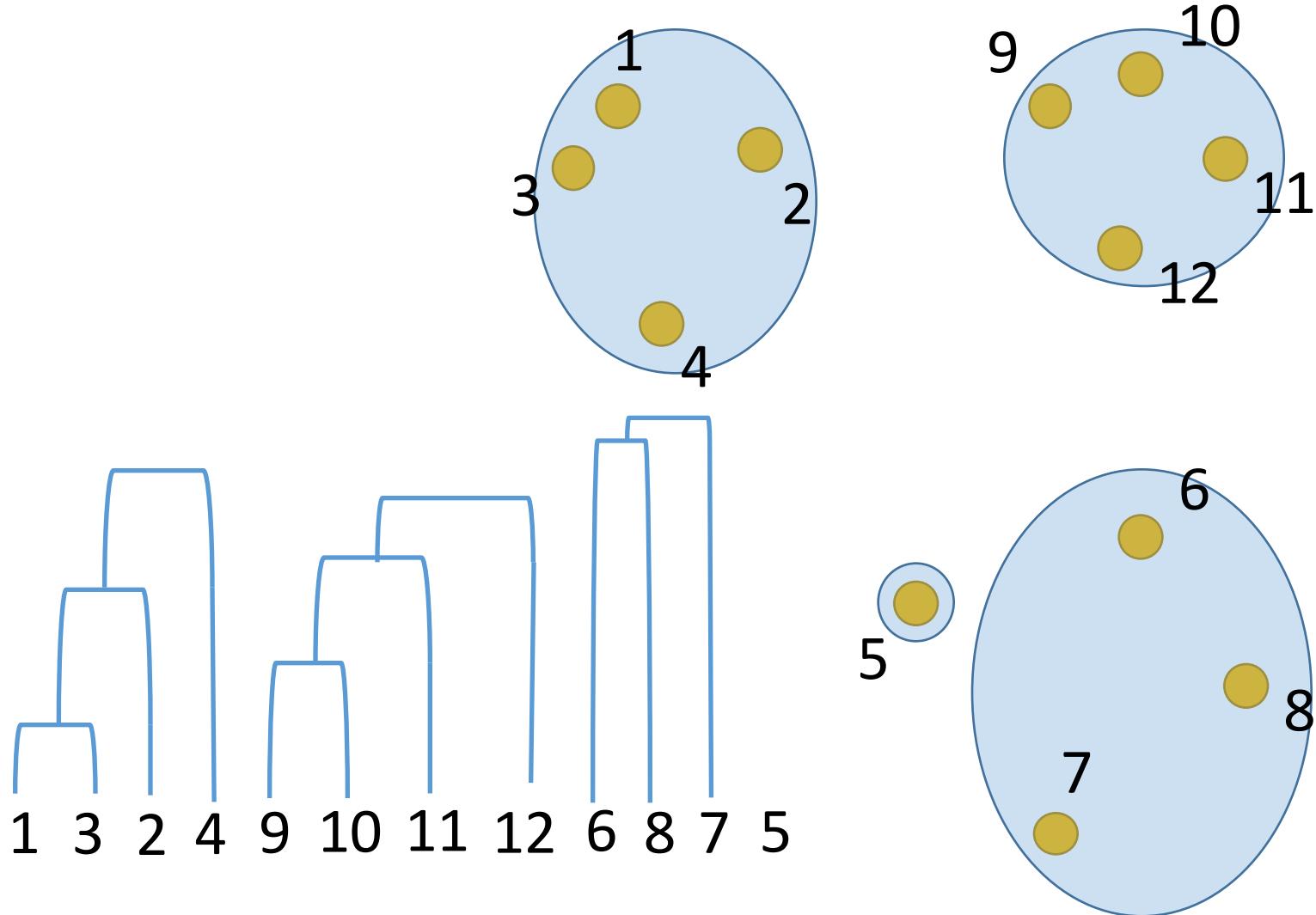
Дендрограмма



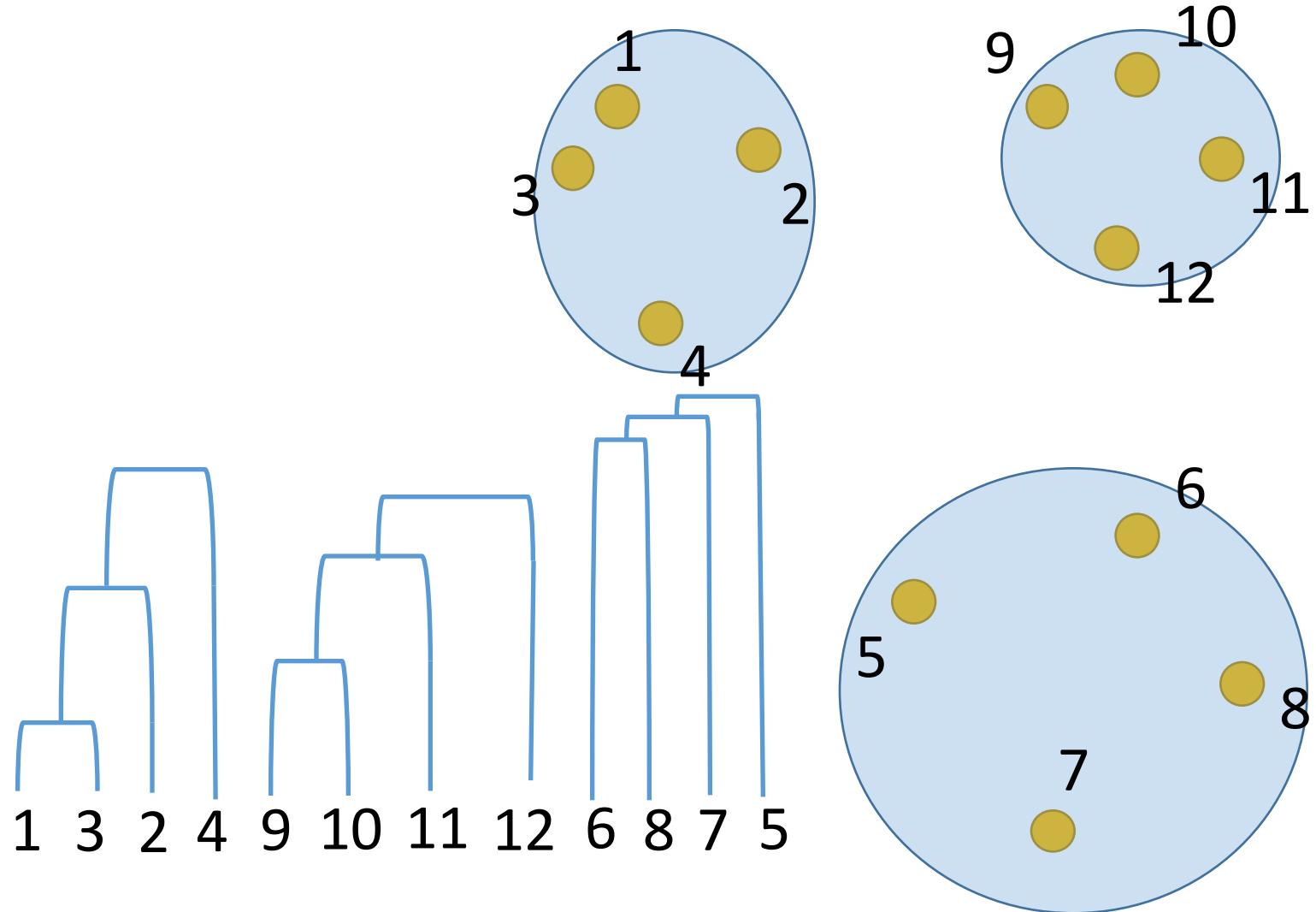
Дендрограмма



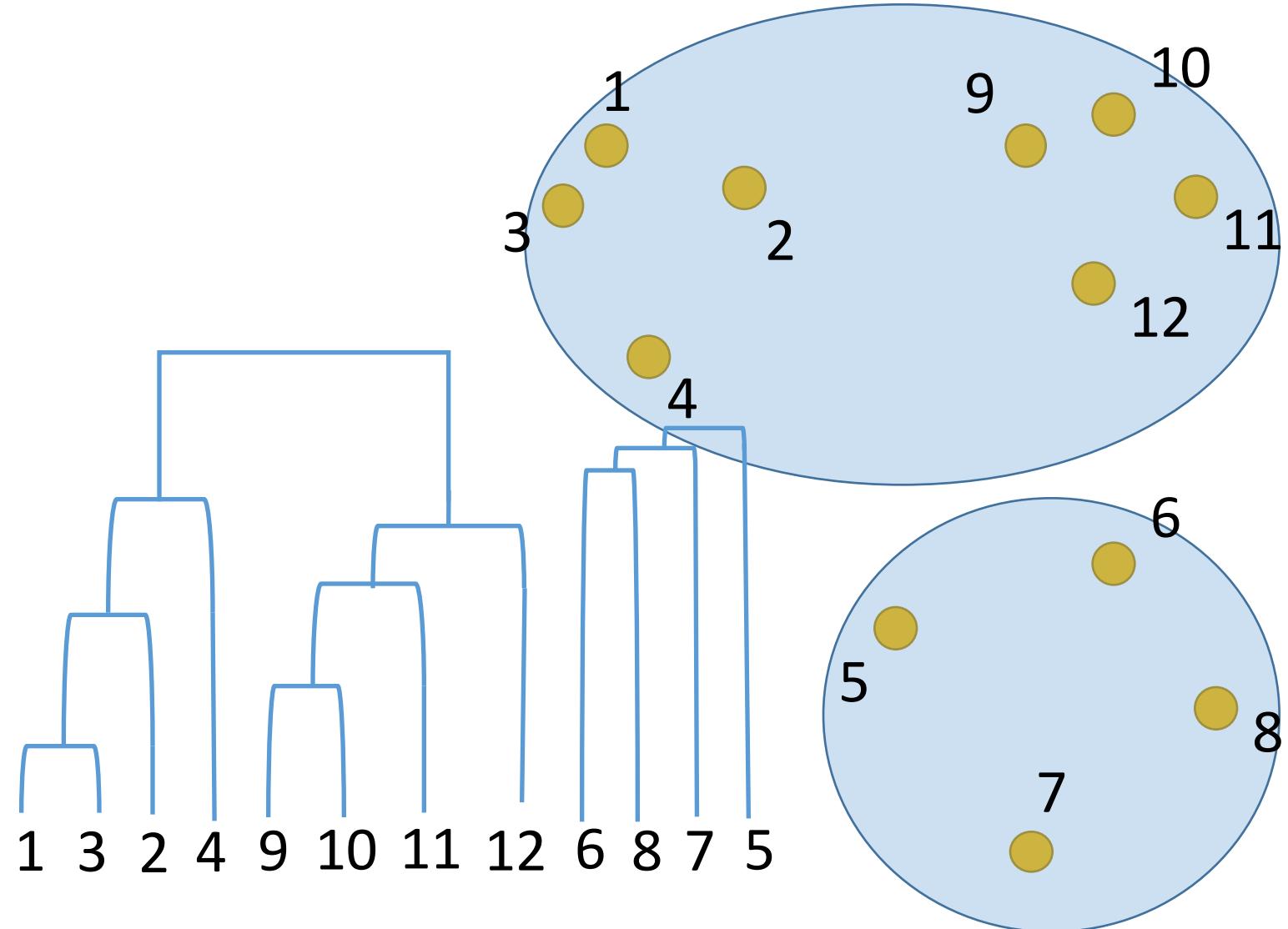
Дендрограмма



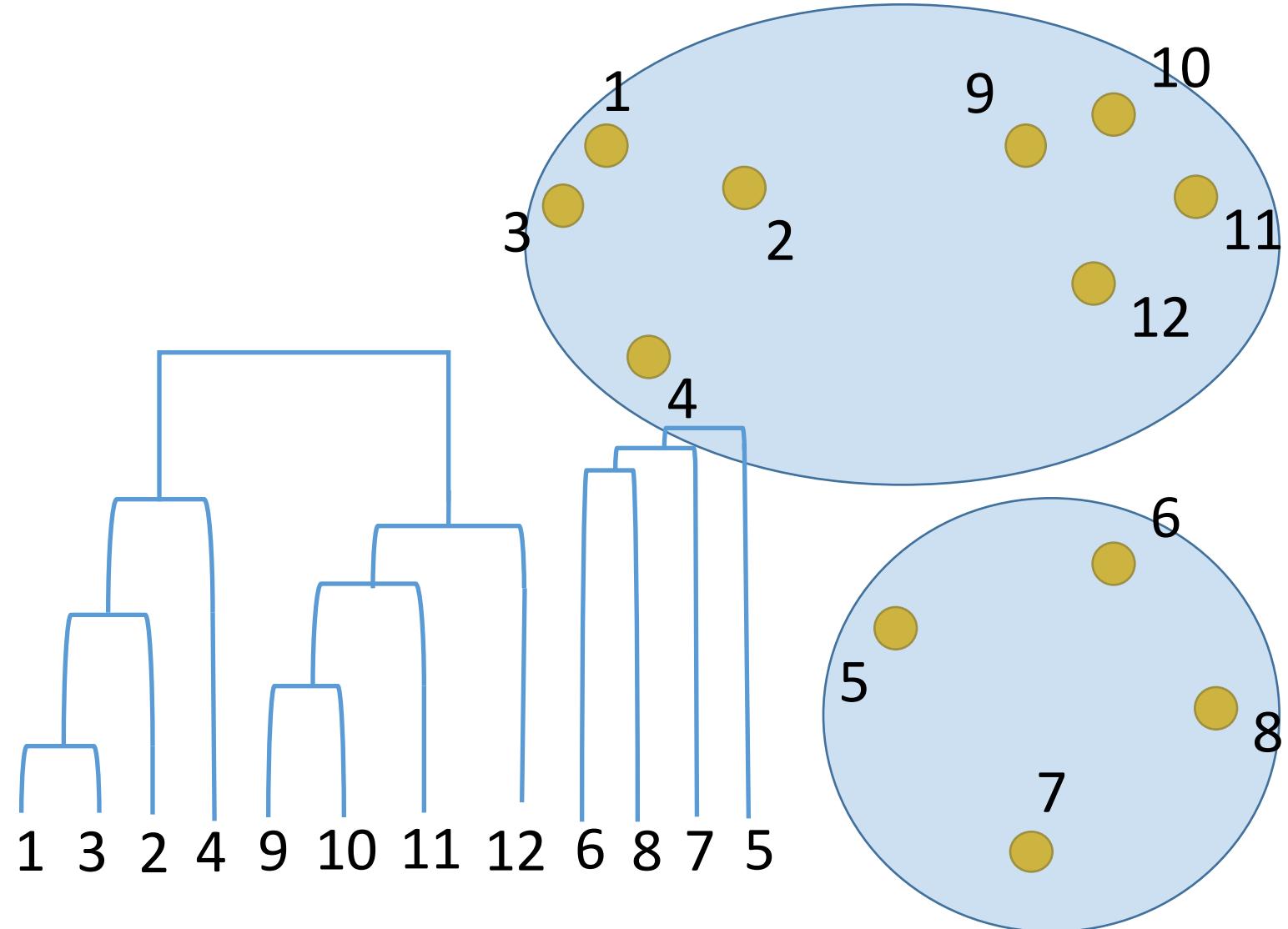
Дендрограмма



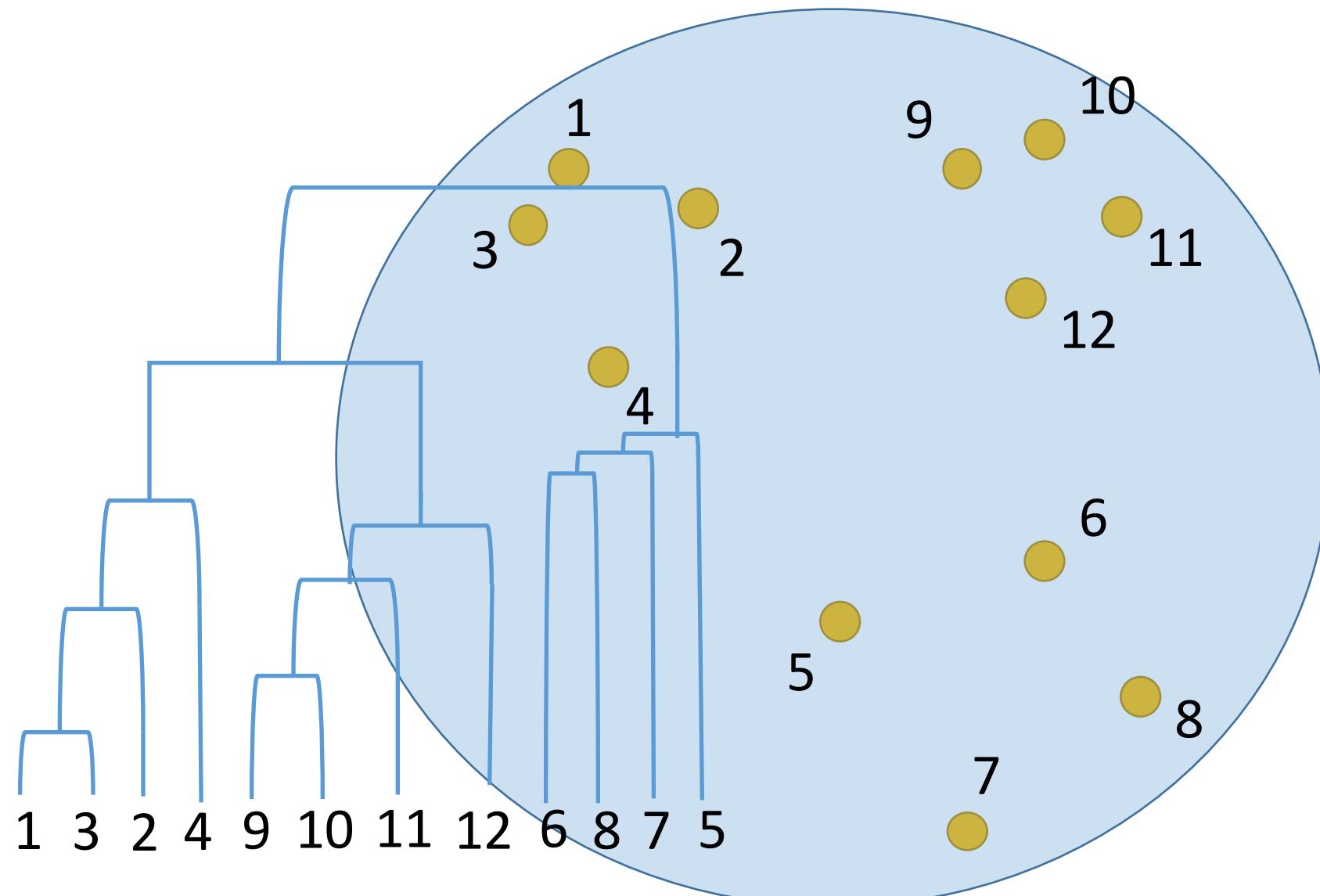
Дендрограмма



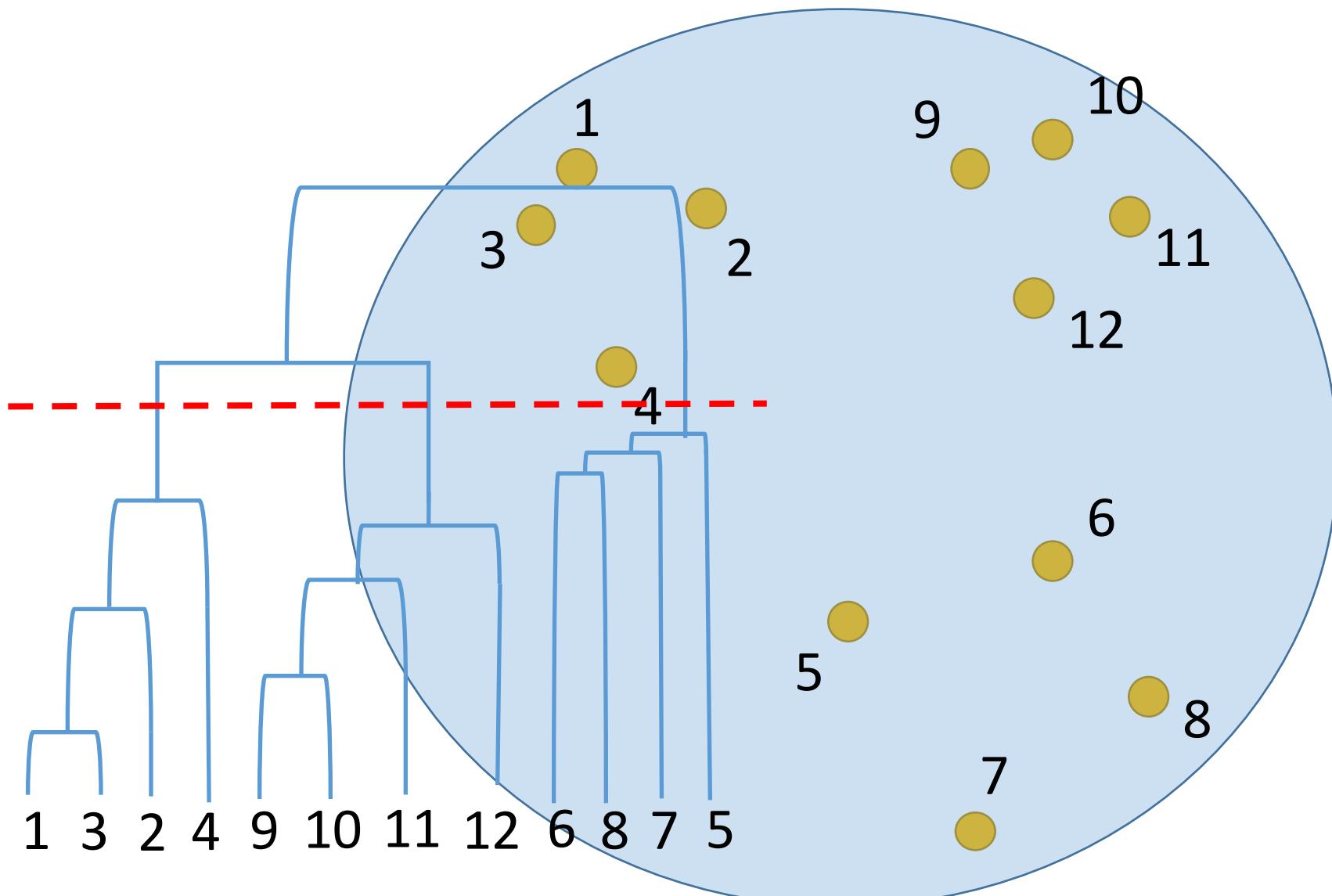
Дендрограмма



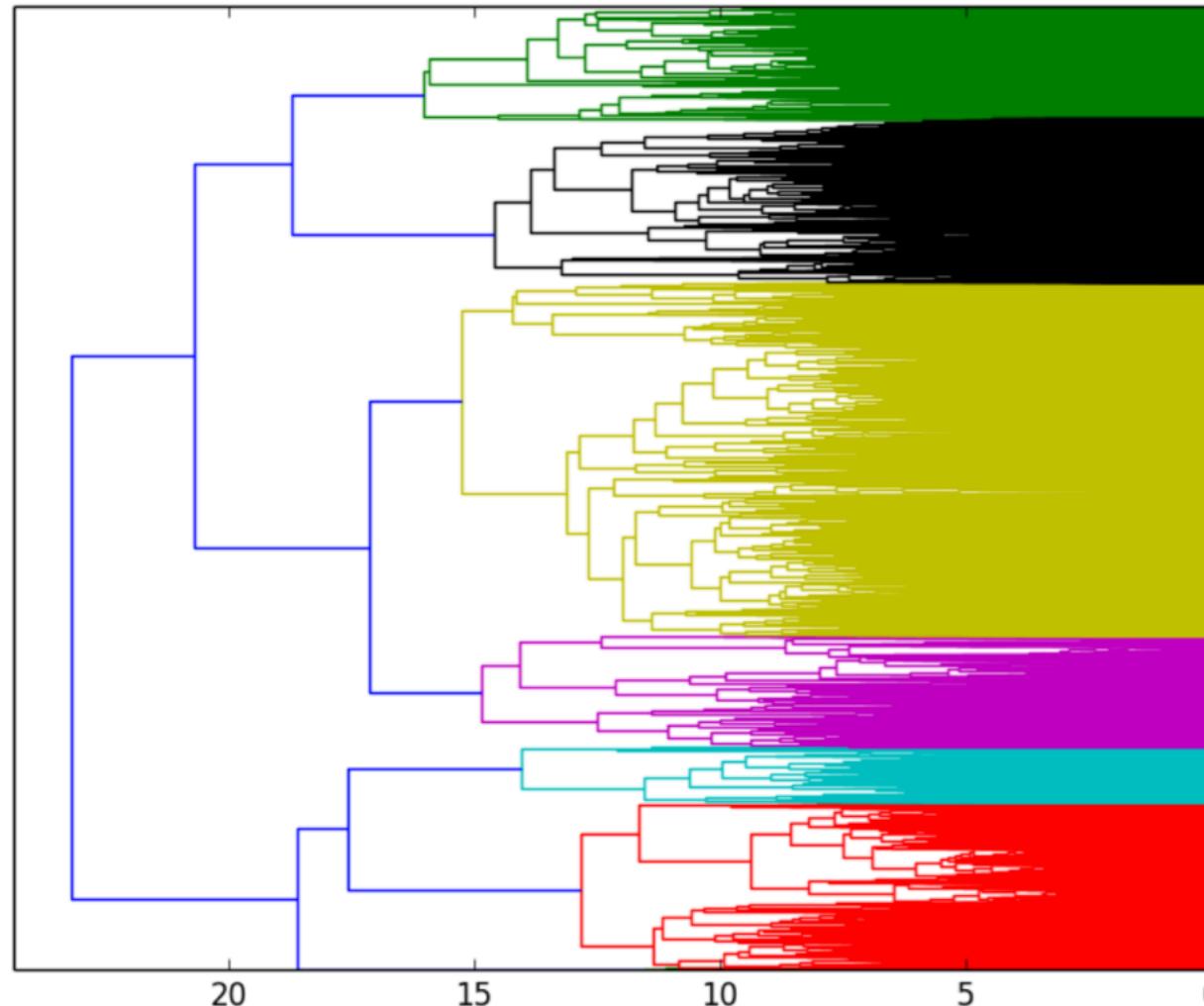
Дендрограмма



Дендрограмма



Пример: кластеризация писем



Резюме

- Обучение без учителя: кластеризация, понижение размерности, визуализация данных и пр.
- Мы рассмотрели некоторые алгоритмы кластеризации: k-means, DBSCAN, графовые методы, агломеративную кластеризацию
- Существует множество алгоритмов понижения размерности
- А также множество алгоритмов обнаружения аномалий