

# Машинное обучение: обучение без учителя

Эмели Драль  
MADE академу, Москва 2020

# Алгоритмы машинного обучения

1. Обучение с учителем: линейные модели
2. Обучение с учителем: ансамбли моделей
3. Обучение с учителем: нейросетевые модели
- 4. Обучение без учителя: обзор методов**
5. Обучение с подкреплением
6. (optional) Рекомендательные системы

# Обучение без учителя: обзор методов

1. Задачи обучения без учителя
2. Алгоритмы кластеризации
3. Подходы к понижению размерности

# Базовые концепты

## Обучение с учителем

Объекты и признаки:

- $x$  – объект
- $y$  – ответ
- $(f_1, f_2 \dots f_n)$  – признаки, описывающие объекты
- $F^{(l,n)}$  – матрица объект-признак
- $X$  – пространство объектов
- $Y$  – пространство ответов

Модель:

- $a: X \rightarrow Y$
  - $a(x) = y$
  - $A$  – семейство моделей
- Оценка качества
- $Q(a, X)$  – ошибки модели  $a(x)$  на группе объектов  $X$

# Базовые концепты

# Обучение без учителя

Объекты и признаки:

- $x$  – объект
- $y$  – ответ
- $(f_1, f_2 \dots f_n)$  – признаки, описывающие объекты
- $X$  – пространство объектов
- $Y$  – множество кластеров

Функция расстояния:

- $\rho: X \times X \rightarrow [0; \infty)$  – семейство моделей

Алгоритм кластеризации:

- $a: X \rightarrow Y$

Оценка качества:

- Объекты одного кластера похожи
- Объекты разных кластеров существенно различаются

# Базовые концепты

## Частичное обучение

Объекты и признаки:

- $x$  – объект
- $(f_1, f_2 \dots f_n)$  – признаки, описывающие объекты
- $\{(x_1, y_1), \dots, (x_l, y_l)\}$  – данные с разметкой
- $\{x_{l+1}, \dots, x_k\}$  – данные без разметки
- $X$  – пространство объектов
- $Y$  – множество кластеров

Функция расстояния:

- $\rho: X \times X \rightarrow [0; \infty)$  – семейство моделей

Алгоритм кластеризации:

- $a: X \rightarrow Y$

Оценка качества:

- Объекты одного кластера похожи
- Объекты разных кластеров существенно различаются

Задачи обучения без учителя

# Задачи обучения без учителя

## 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

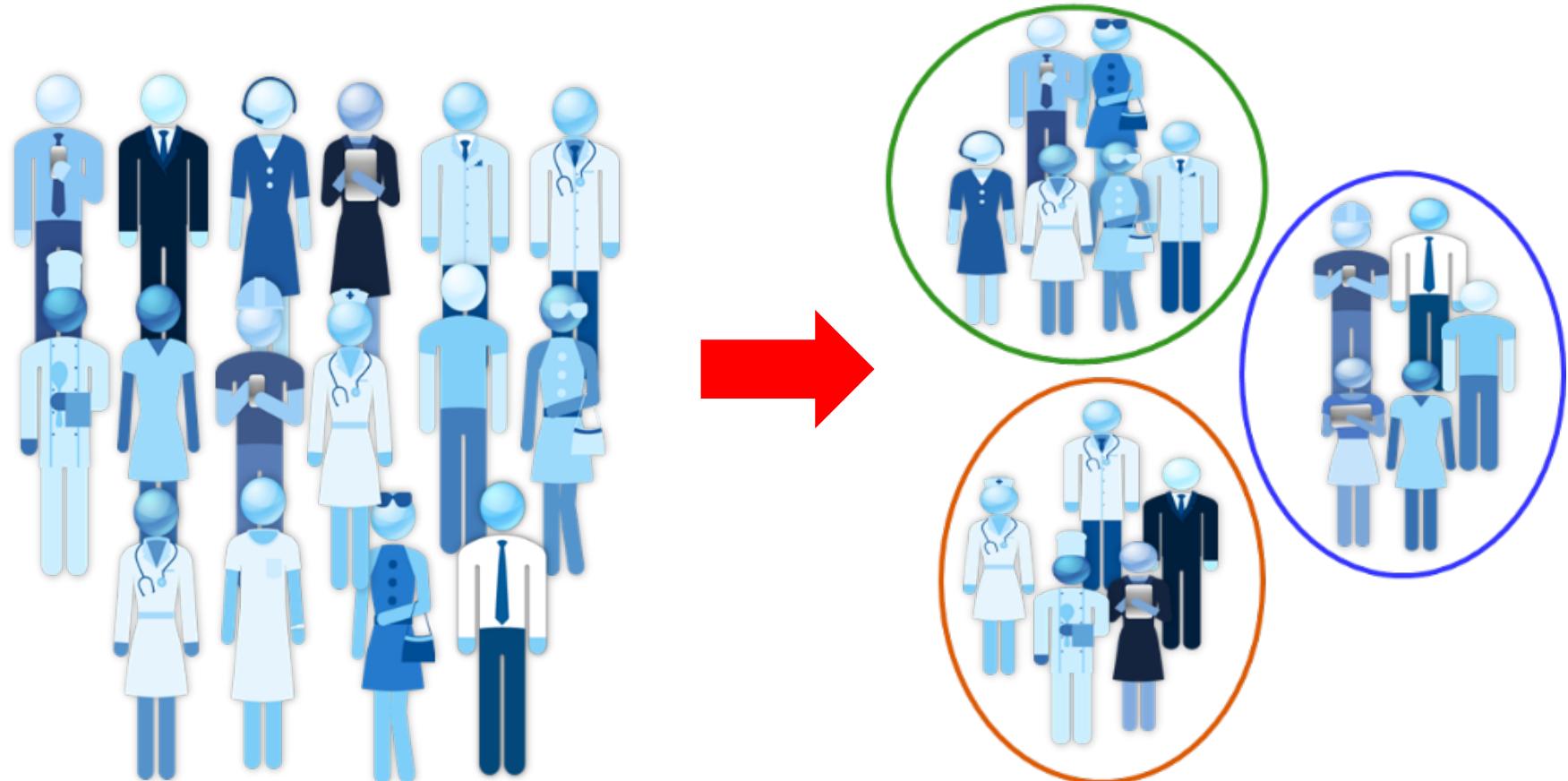
Задачи  
обучения без  
учителя

# Кластеризация



Задачи  
обучения без  
учителя

# Кластеризация



Задачи  
обучения без  
учителя

# Кластеризация



Clustering is subjective



Simpson's Family



School Employees



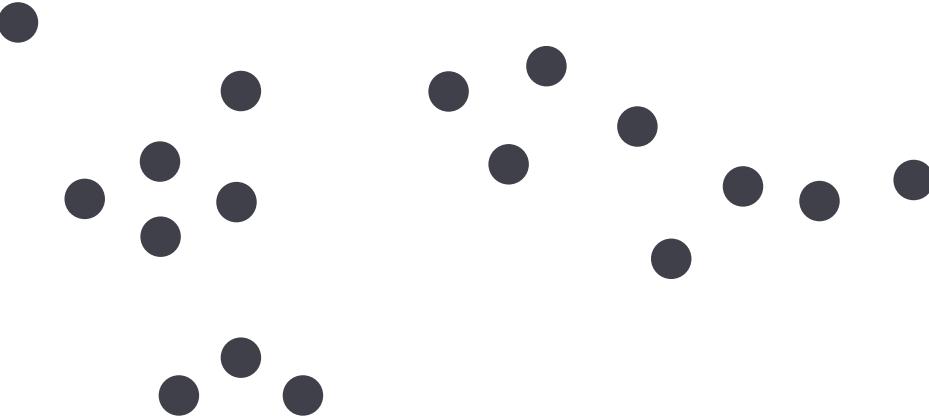
Females



Males

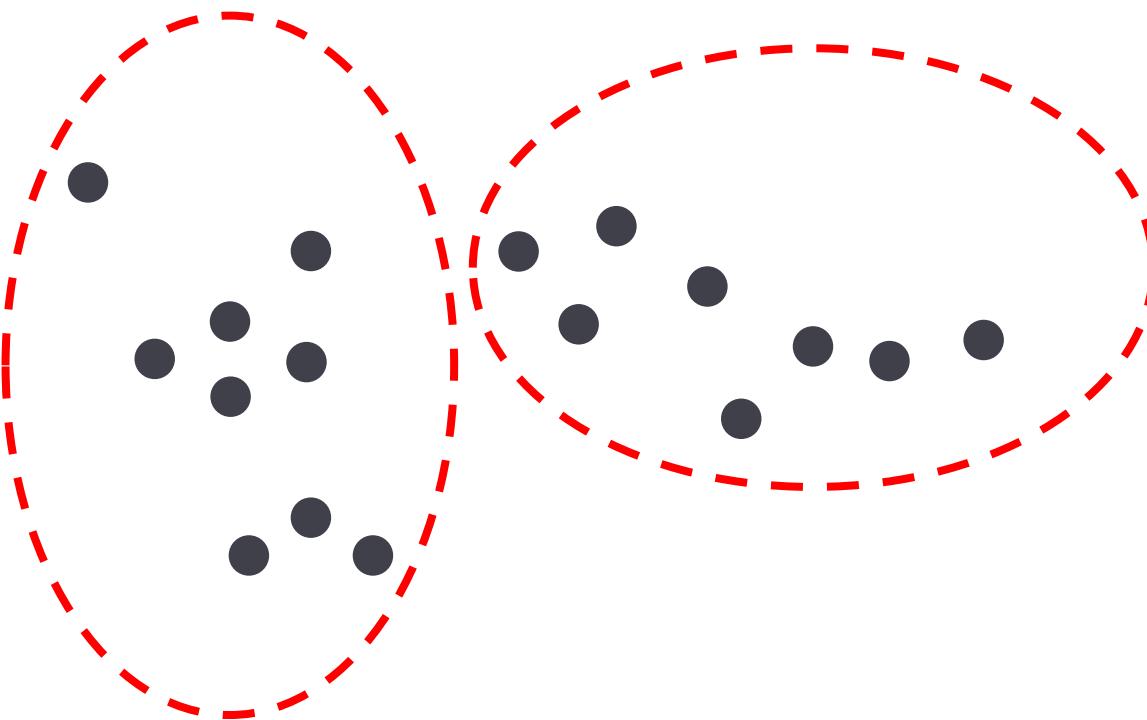
Задачи  
обучения без  
учителя

# Какова истинная структура?



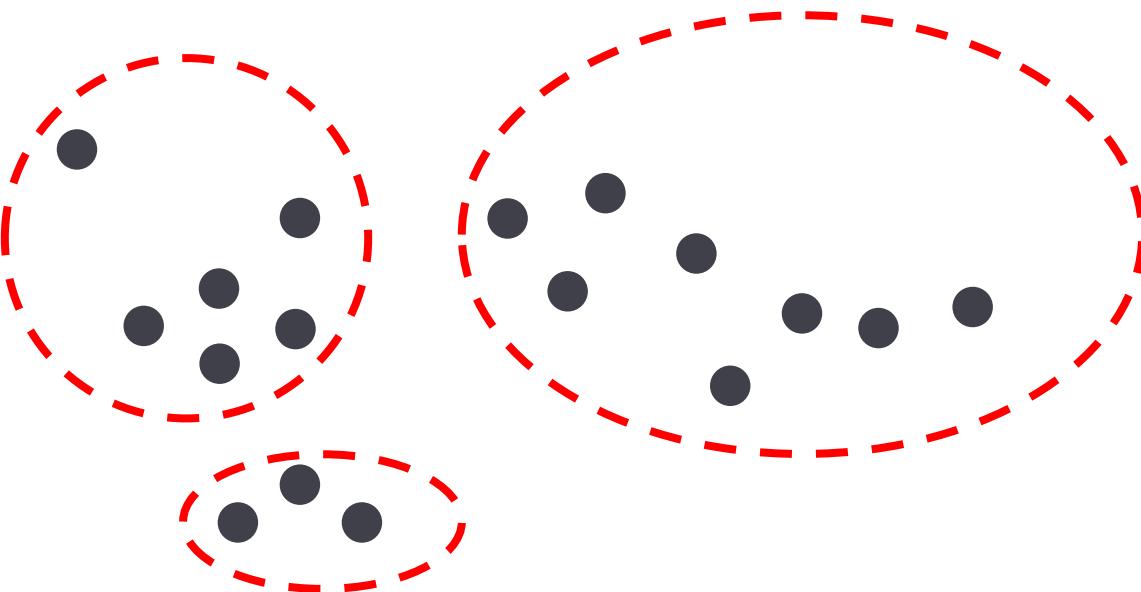
Задачи  
обучения без  
учителя

# Какова истинная структура?



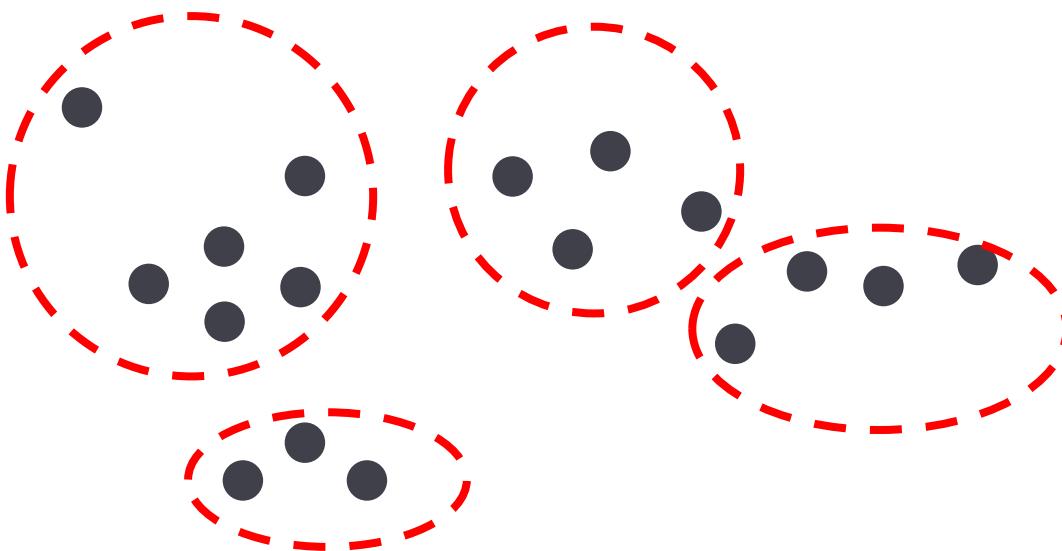
Задачи  
обучения без  
учителя

# Какова истинная структура?



Задачи  
обучения без  
учителя

# Какова истинная структура?



Задачи  
обучения без  
учителя

# Кластеризация



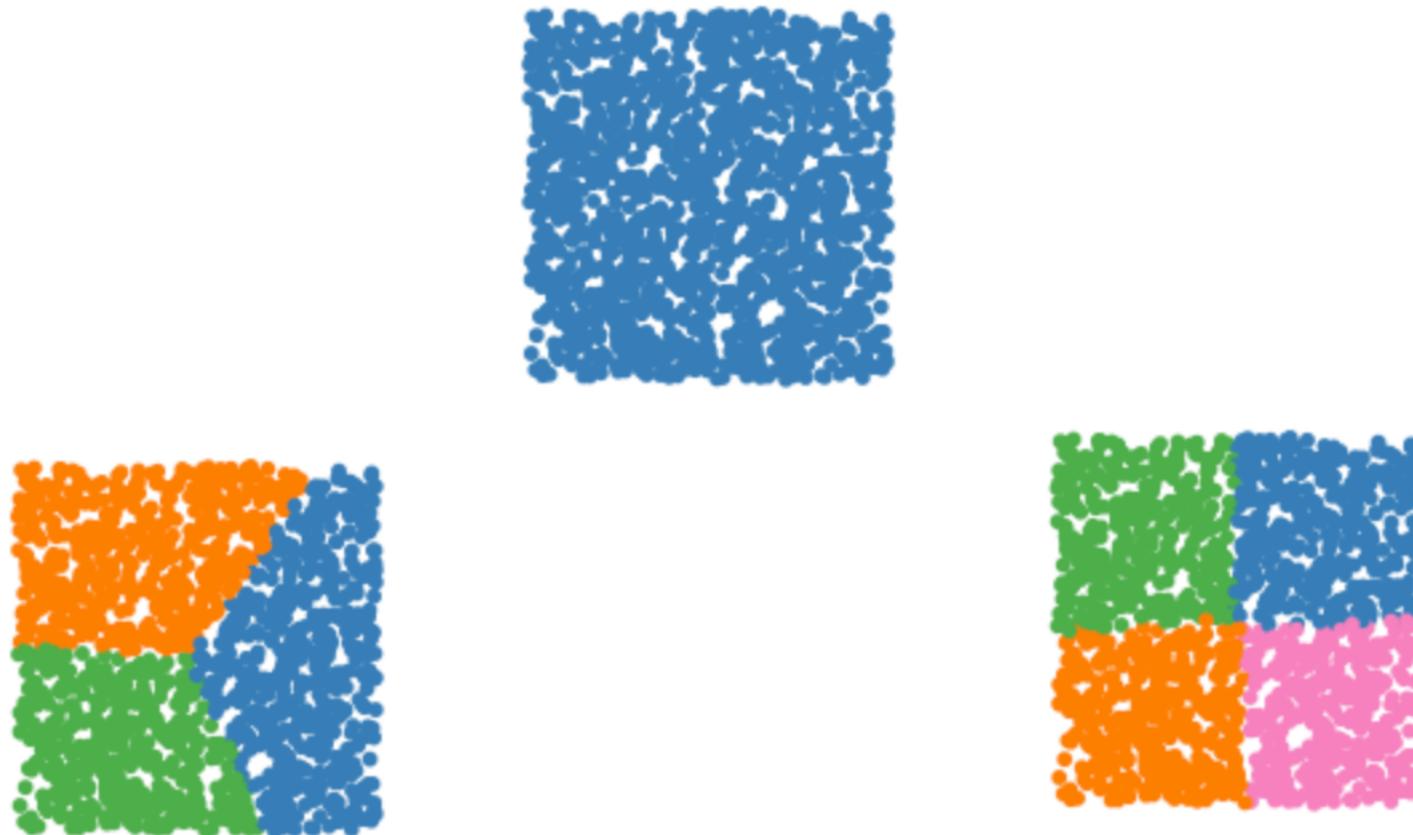
Задачи  
обучения без  
учителя

Не стоит ждать волшебного  
качества всегда



# Задачи обучения без учителя

А иногда стоит свести задачу к другой



# Задачи обучения без учителя

## Детектирование аномалий

- Детектирование аномалий (outlier detection, anomaly detection)
- Детектирование новизны (novelty detection)

Задачи  
обучения без  
учителя

# Детектирование аномалий



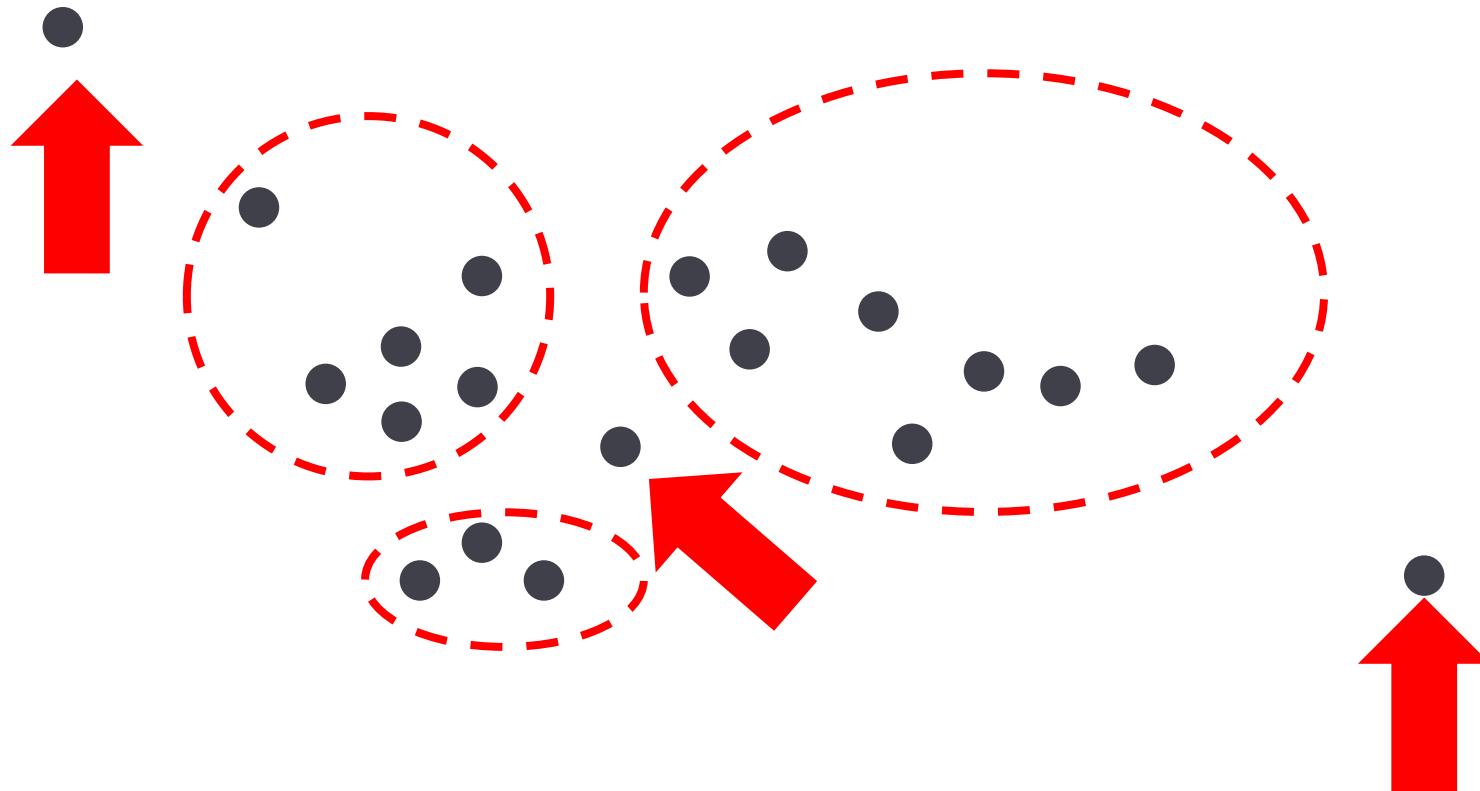
Задачи  
обучения без  
учителя

# Детектирование аномалий



Задачи  
обучения без  
учителя

# Детектирование аномалий



# Задачи обучения без учителя

## Понижение размерности данных

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

# Задачи обучения без учителя

## Понижение размерности данных

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

Зачем?

- исходные данные избыточны с точки зрения количества информации, необходимого для решения задачи
- оптимизация вычислительных затрат
- подготовка данных для дальнейшего анализа, например, визуального

# Задачи обучения без учителя

## Понижение размерности данных

	f1
x1	
x2	
x3	
x4	
x5	
x6	
x7	

Как изобразить выборку на  
плоскости?

# Задачи обучения без учителя

## Понижение размерности данных

	f1
x1	
x2	
x3	
x4	
x5	
x6	
x7	

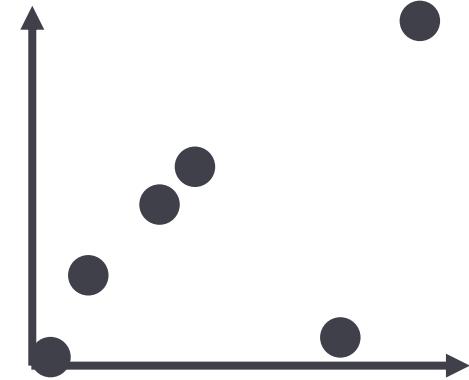


Как изобразить выборку на  
плоскости?

# Задачи обучения без учителя

## Понижение размерности данных

	f1	f2
x1		
x2		
x3		
x4		
x5		
x6		
x7		

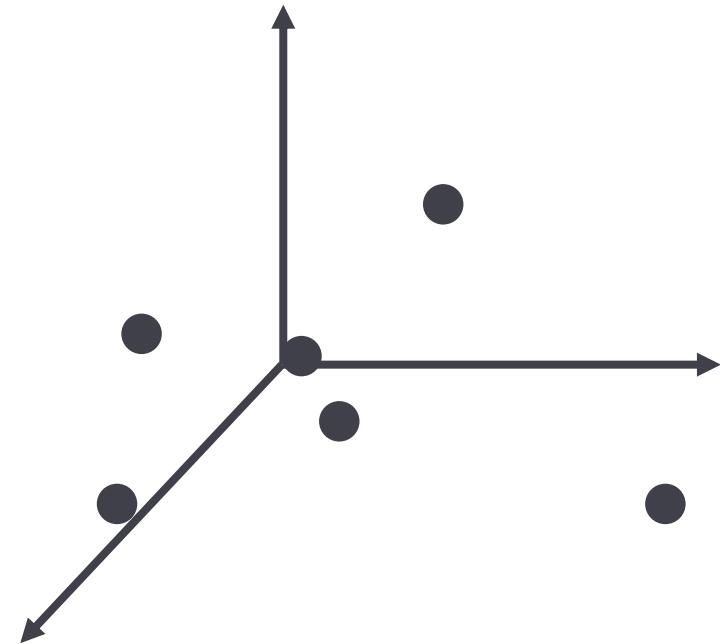


Как изобразить выборку на  
плоскости?

# Задачи обучения без учителя

## Понижение размерности данных

	f1	f2	f3
x1			
x2			
x3			
x4			
x5			
x6			
x7			



Как изобразить выборку на  
плоскости?

# Задачи обучения без учителя

## Понижение размерности данных

	f1	f2	f3	f4	f5	f6	f7	f8	f9
x1									
x2									
x3									
x4									
x5									
x6									
x7									

Как изобразить выборку на  
плоскости?

# Задачи обучения без учителя

## Понижение размерности данных

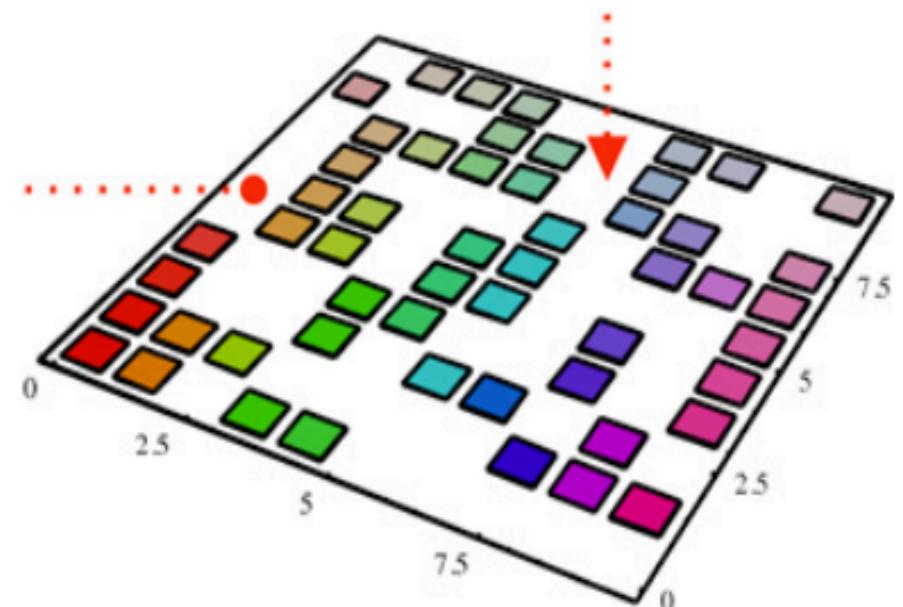
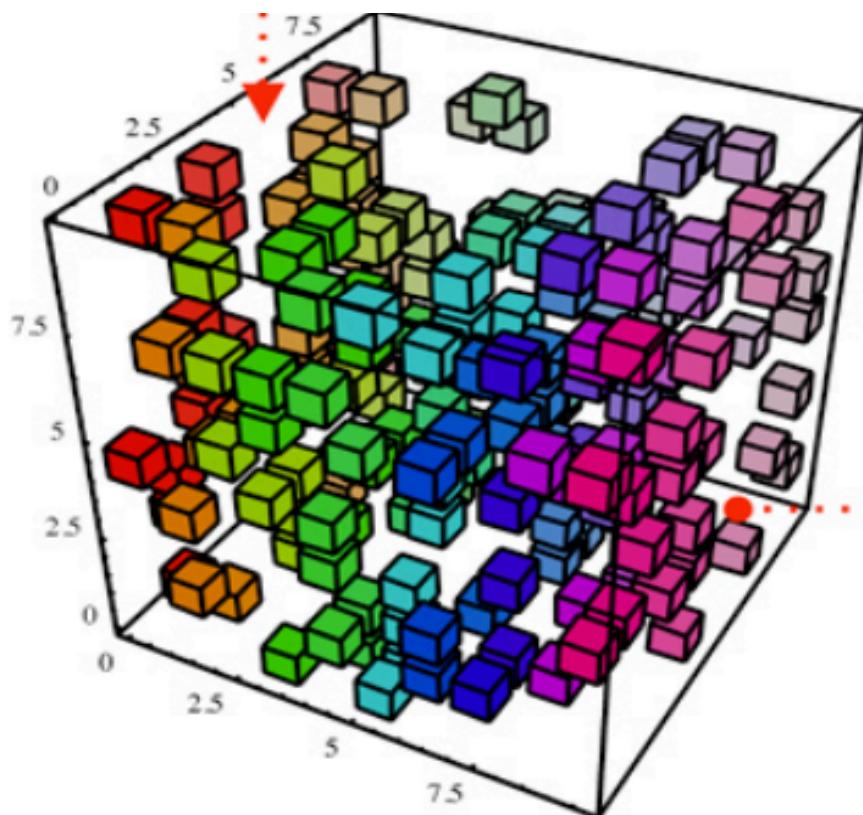
	f1	f2	f3	f4	f5	f6	f7	f8	f9
x1									
x2									
x3									
x4									
x5									
x6									
x7									

Как их изобразить выборку на  
плоскости?

1. Понизить размерность до такой, в  
которой мы можем визуализировать
2. Визуализировать

Задачи  
обучения без  
учителя

# Понижение размерности данных



# Задачи обучения без учителя

## Обучение без учителя

Чаще всего применяется для:

- Кластеризации данных
- Поиска аномалий
- Визуализации данных

# Алгоритмы кластеризации

## Алгоритмы 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

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

Алгоритмы  
кластеризации

# Среднее межклластерное расстояние

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

Алгоритмы  
кластеризации

## Алгоритмы кластеризации

# Качество кластеризации

- Как сильно объекты внутри одной группы **похожи** друг на друга –  $F_0$
- Как сильно объекты из разных групп **отличаются** друг от друга –  $F_1$
- Можно ли их объединить?

## Алгоритмы кластеризации

# Качество кластеризации

- Как сильно объекты внутри одной группы **похожи** друг на друга –  $F_0$
- Как сильно объекты из разных групп **отличаются** друг от друга –  $F_1$
- Можно ли их объединить?

## Алгоритмы кластеризации

# Качество кластеризации

- Как сильно объекты внутри одной группы **похожи** друг на друга –  $F_0$
  - Как сильно объекты из разных групп **отличаются** друг от друга –  $F_1$
- 
- Можно ли их объединить?
  - $F_0/F_1 \rightarrow \min$

## Алгоритмы кластеризации

# Качество кластеризации

Важный вопрос: **для чего** строится кластеризация?

Оценка качества помогает понять, способствуют ли результаты кластеризации решению конечной задачи?

- Эффективность маркетинговой кампании
- Качество модели и вклад признаков в модель
- Качество модели vs скорость обучения

# Алгоритмы кластеризации

## Алгоритмы

Мы обсудим три подхода, которые часто применяются на практике

- k-Means
- Графовые методы
- DBSCAN
- Агломеративная кластеризация

# Алгоритмы кластеризации

## Алгоритмы

Мы обсудим несколько подходов:

- k-Means
- Графовые методы
- DBSCAN
- Агломеративная кластеризация

# Алгоритмы кластеризации

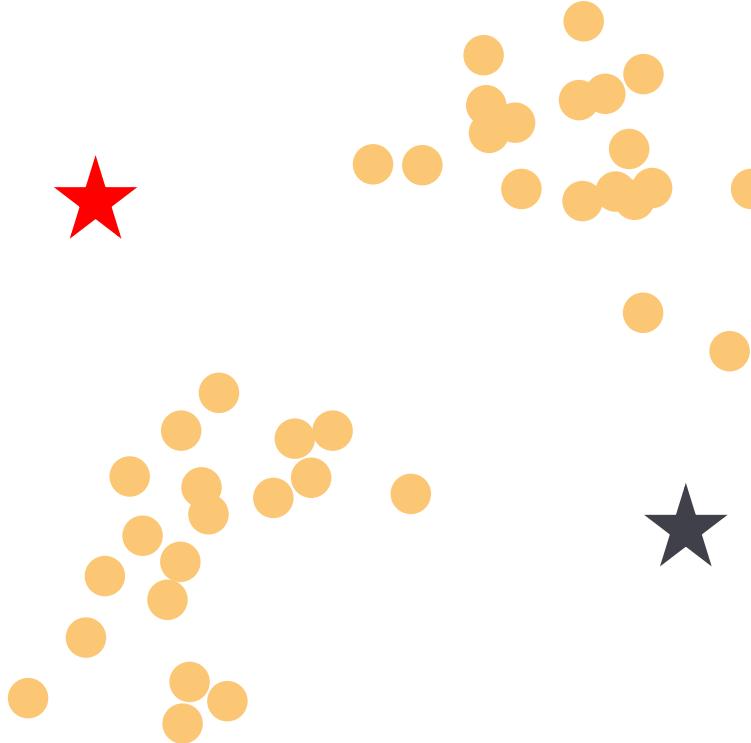
## Метод k-Means

Метод 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.

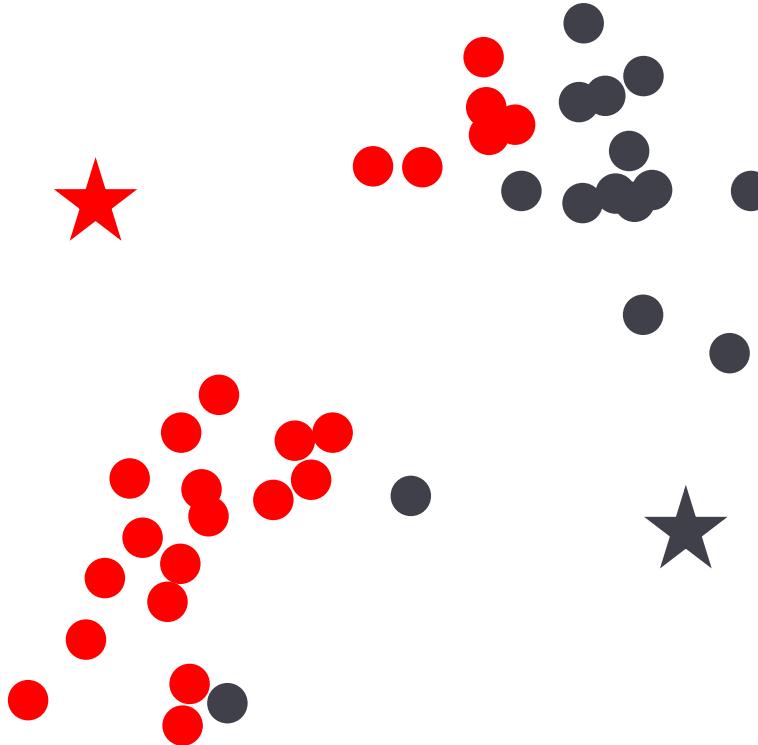
Алгоритмы  
кластеризации

# Метод k-Means



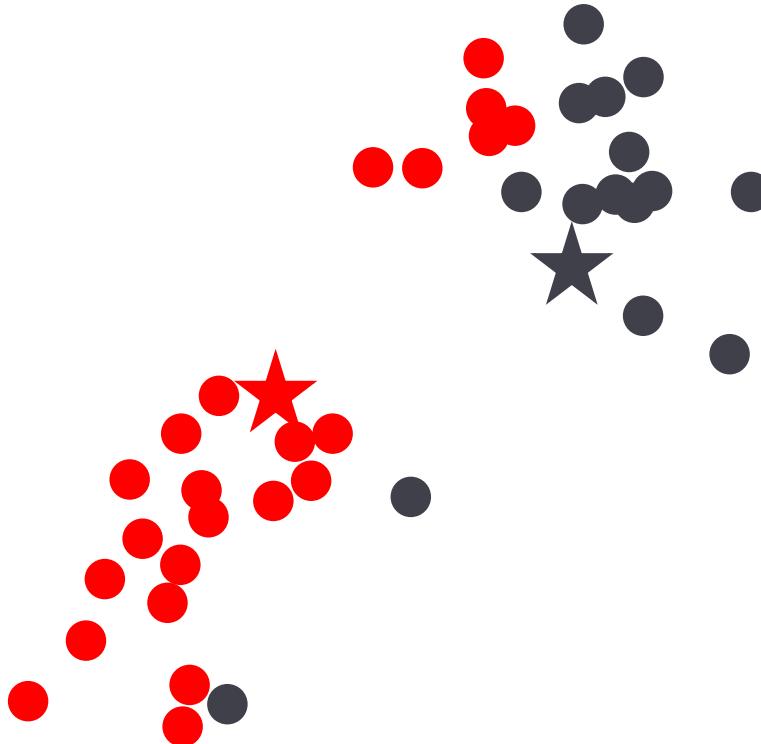
Алгоритмы  
кластеризации

# Метод k-Means



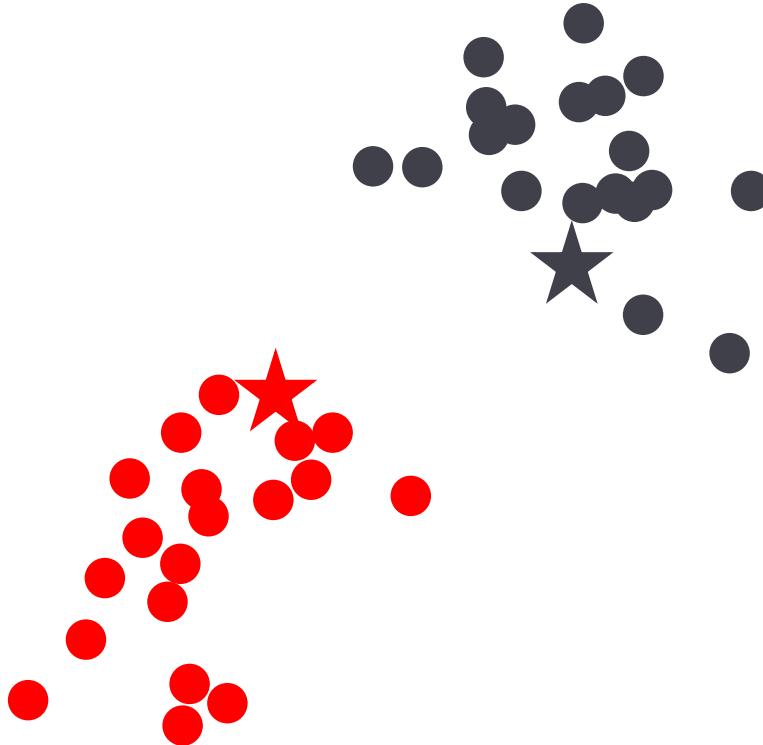
# Алгоритмы кластеризации

## Метод k-Means



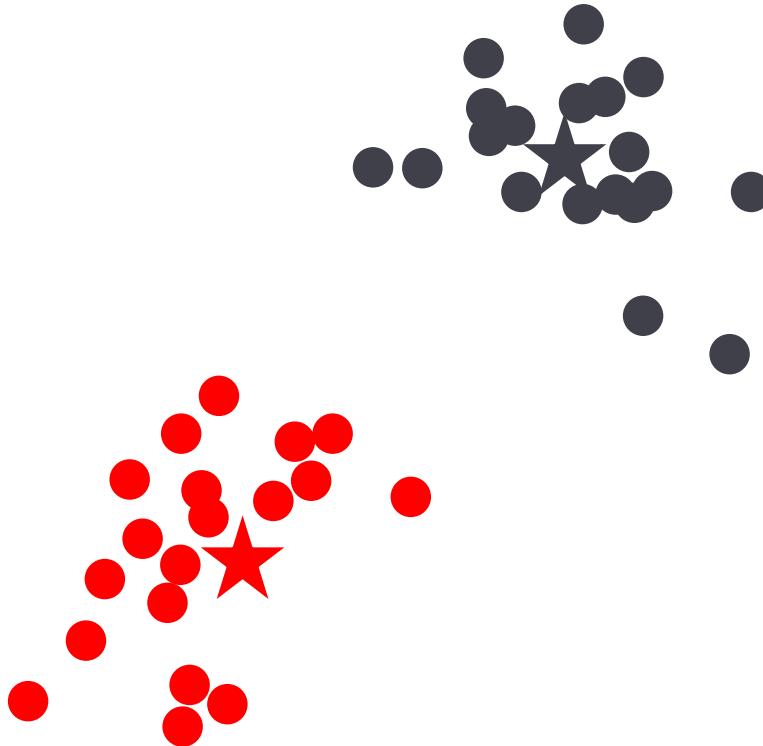
Алгоритмы  
кластеризации

# Метод k-Means



Алгоритмы  
кластеризации

# Метод k-Means



## Алгоритмы кластеризации

# Метод k-Means

Метод k-means итеративно минимизирует среднее внутрикластерное расстояние.

- Метод чувствителен к начальной инициализации центров кластеров. В случае неудачной инициализации результаты будут неудовлетворительными.
- Метод не оптимизирует форму кластеров
- Метод не оптимизирует количество кластеров
- Метод не устойчив к выбросам
- Метод применим в линейных пространствах (нам надо уметь считать среднее)

## Алгоритмы кластеризации

# Графовые методы

Графовые подходы основаны на представлении данных в виде графа, где множеством вершин являются объекты, а множеством рёбер – попарные расстояния между объектами.

Мы рассмотрим:

- Алгоритм на основе связных компонент
- Алгоритм на основе оствового дерева

# Алгоритмы кластеризации

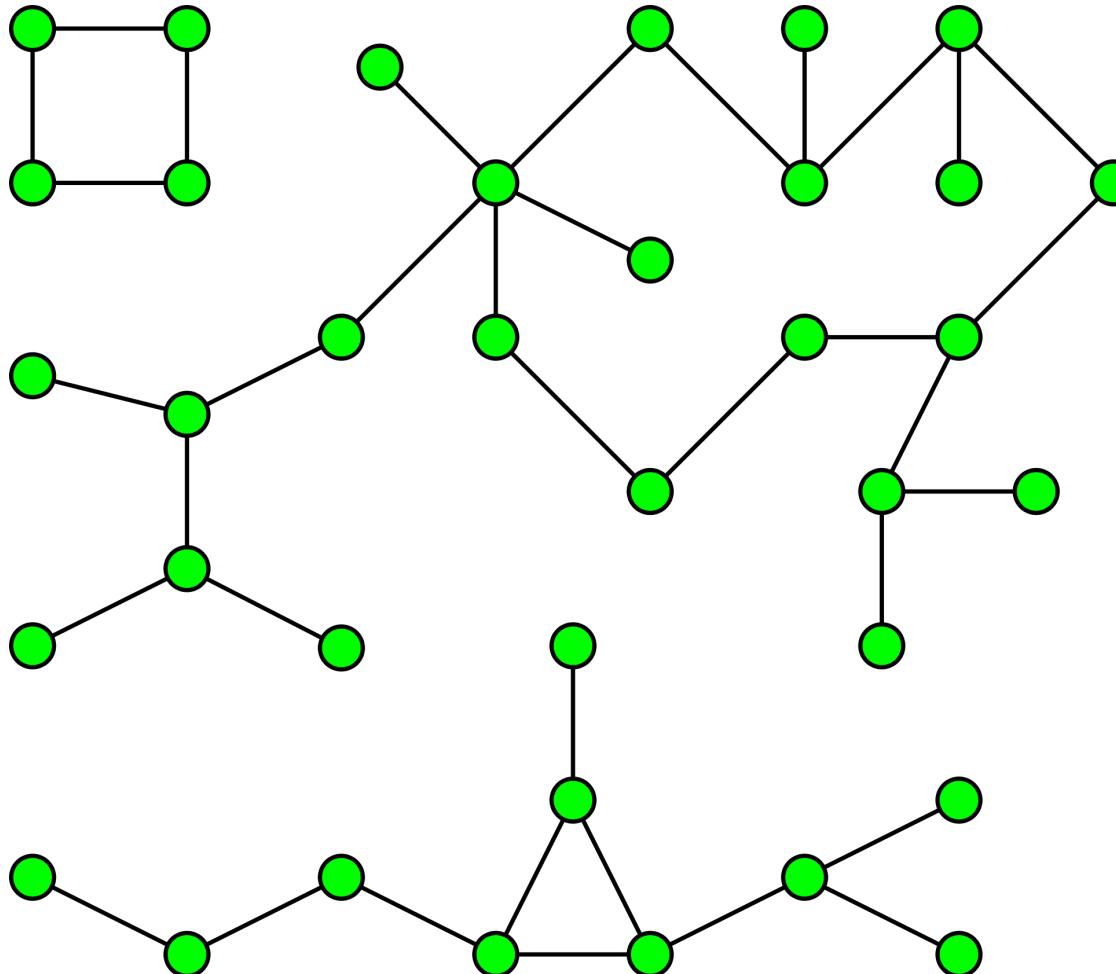
## Выделение связных компонент

Подход на основе компонент связности:

1. Соединяем ребром объекты, расстояние между которыми меньше  $R$
2. Выделяем компоненты связности

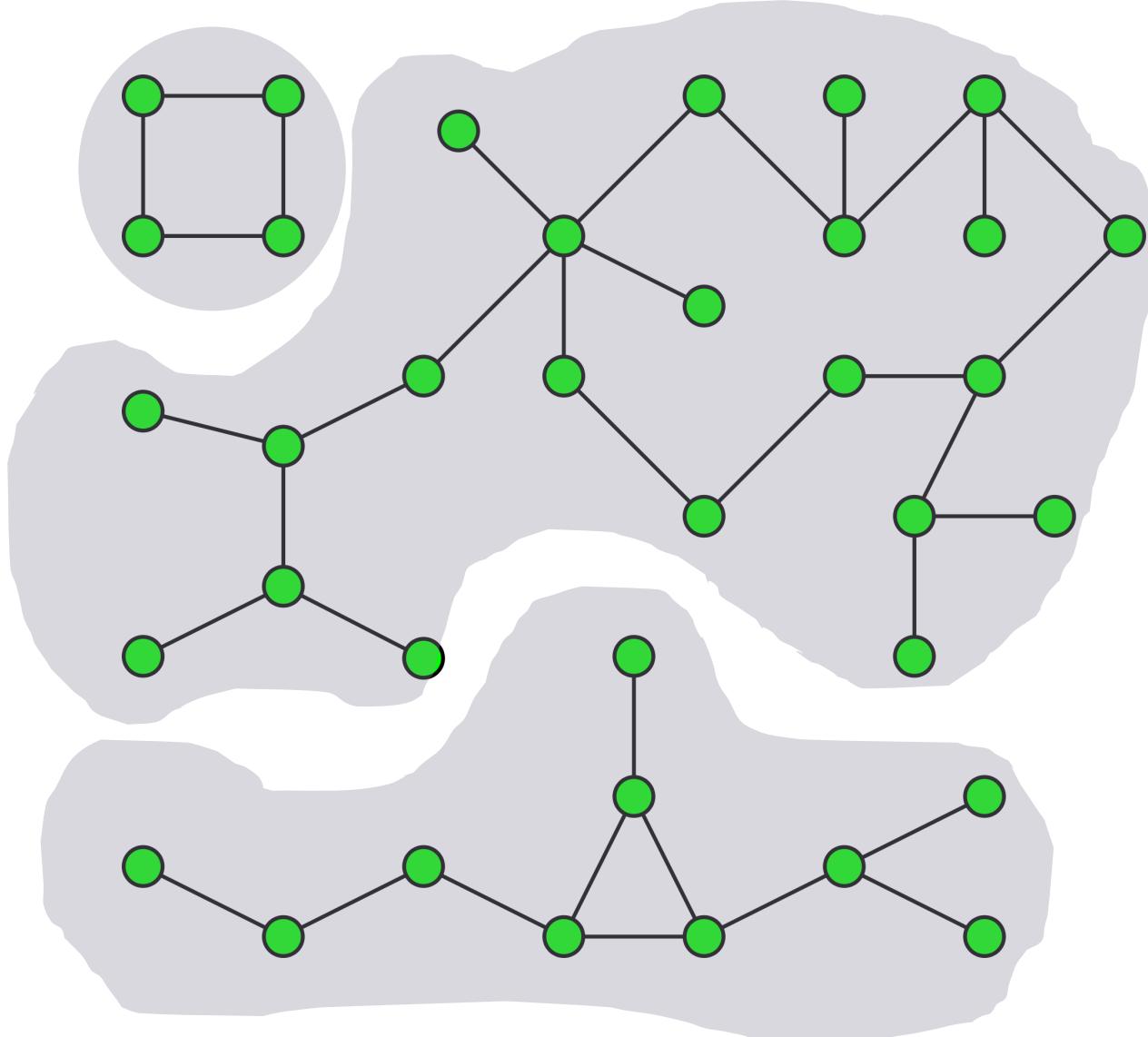
# Алгоритмы кластеризации

## Выделение связных компонент



## Алгоритмы кластеризации

# Выделение связных компонент



# Алгоритмы кластеризации

## Выделение связных компонент

Подход на основе компонент связности:

1. Соединяем ребром объекты, расстояние между которыми меньше  $R$
2. Выделяем компоненты связности

Проблема: непонятно, как выбрать  $R$ , если нужно получить  $K$  кластеров

# Алгоритмы кластеризации

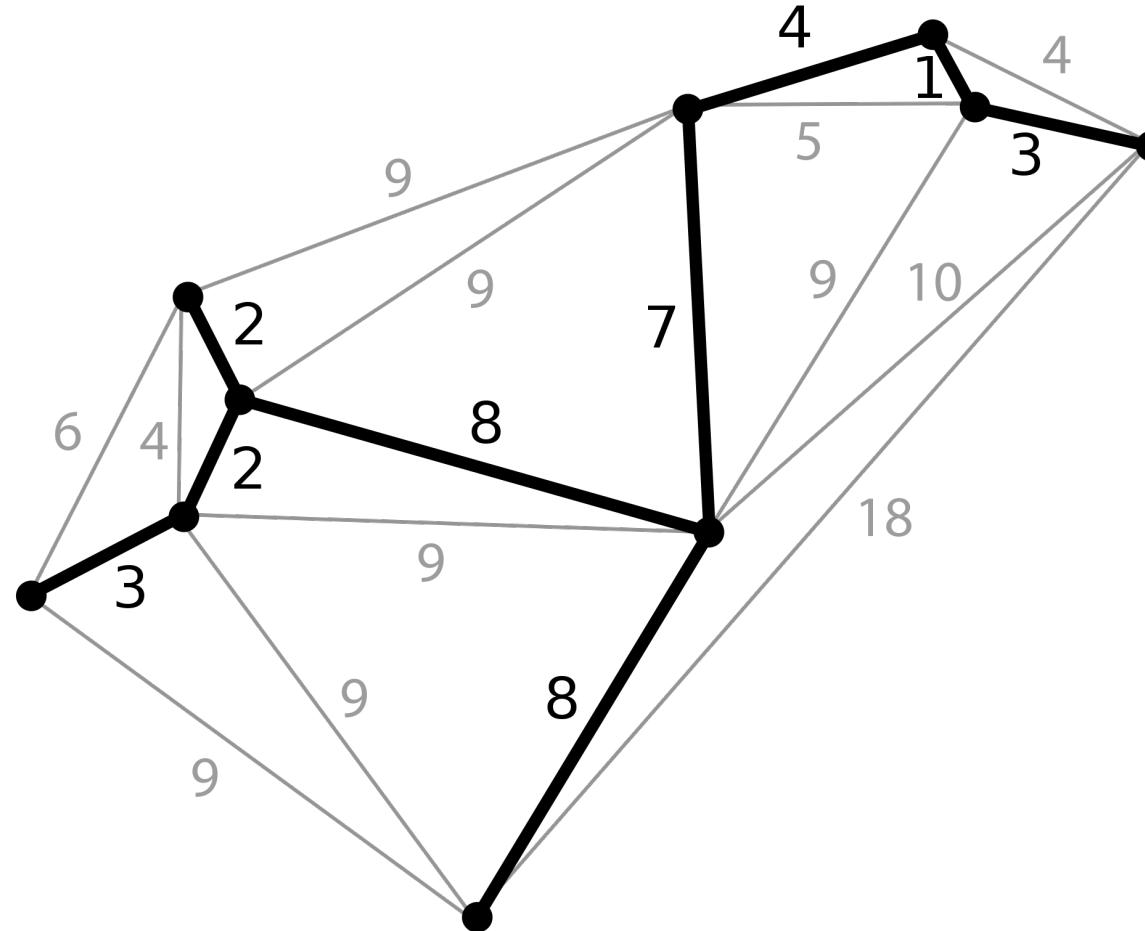
## Минимальное оставное дерево

Подходы на основе оставного дерева:

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

# Алгоритмы кластеризации

## Минимальное остовное дерево



# DBSCAN

Density-based spatial clustering of applications with noise

Метод кластеризации на основе плотности объектов в пространстве.

Алгоритмы  
кластеризации

# Алгоритмы кластеризации

## DBSCAN

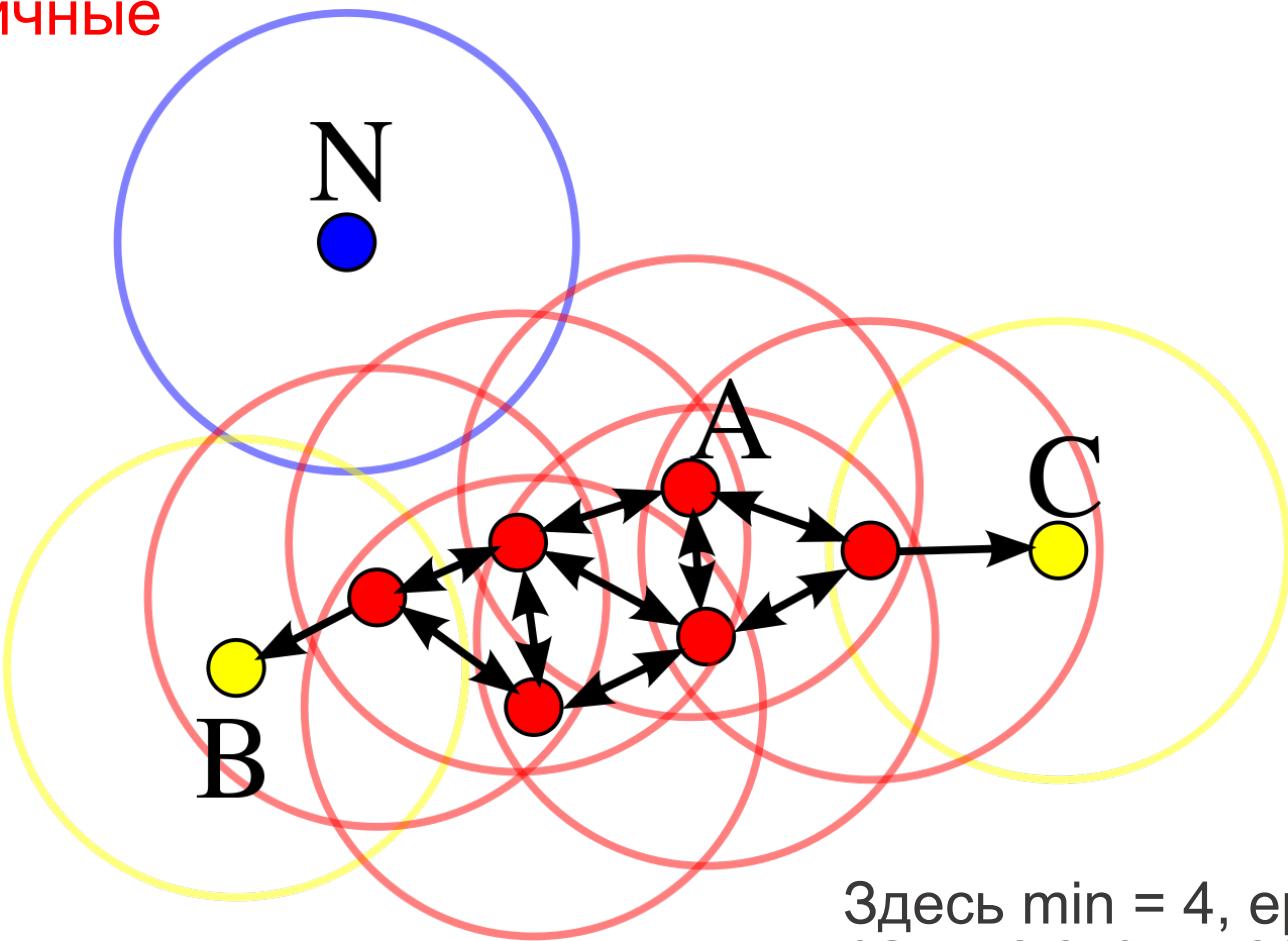
Метод делит объекты на 3 группы: **основные, шумовые и граничные**

- Точка  $p$  является основной, если по меньшей мере  $min$  точек находятся на расстоянии, не превосходящем  $\text{eps}$ , до неё
- Точка  $q$  прямо достижима из  $p$ , если точка  $q$  находится на расстоянии, не большем  $\text{eps}$  от  $p$
- Точка  $A$  достижима из  $p$ , если имеется путь  $p_1, \dots, p_n$ , где  $p_1 = p$  и  $p_n = A$ , а каждая  $p_{i+1}$  достижима прямо из  $p_i$  (все точки на пути должны быть основными, за исключением  $q$ )
- Все точки, не достижимые из основных точек, считаются выбросами
- Все точки, не являющиеся выбросами и основными точками – граничные

## Алгоритмы кластеризации

# DBSCAN

Метод делит объекты на 3 группы: **основные, шумовые и граничные**



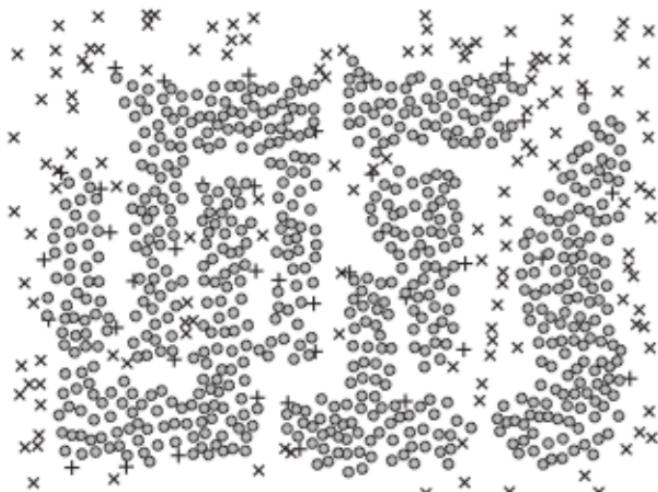
Здесь  $min = 4$ ,  $\text{eps}$  –  
радиус окружностей

# Алгоритмы кластеризации

## DBSCAN: алгоритм



(a) Clusters found by DBSCAN.

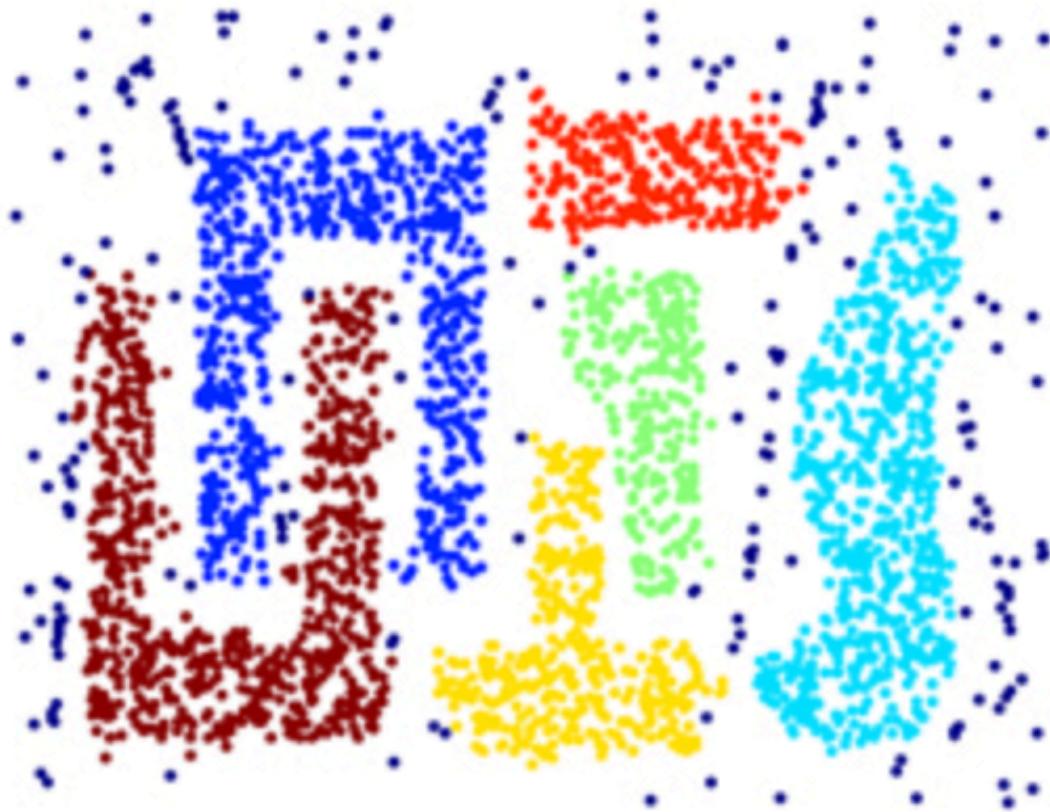


(b) Core, border, and noise points.

- 1: Пометить все точки, как основные, пограничные или шумовые.
- 2: Отбросить точки шума.
- 3: Соединить все основные точки, находящиеся на расстоянии  $Eps$  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.

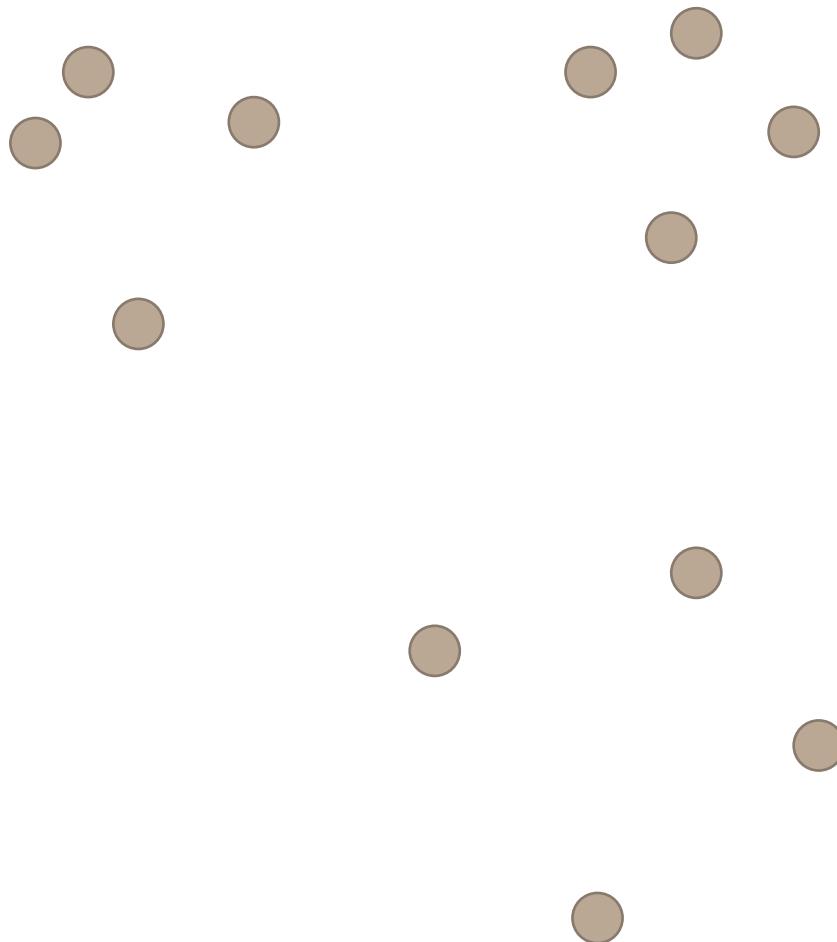
Алгоритмы  
кластеризации

# DBSCAN: алгоритм



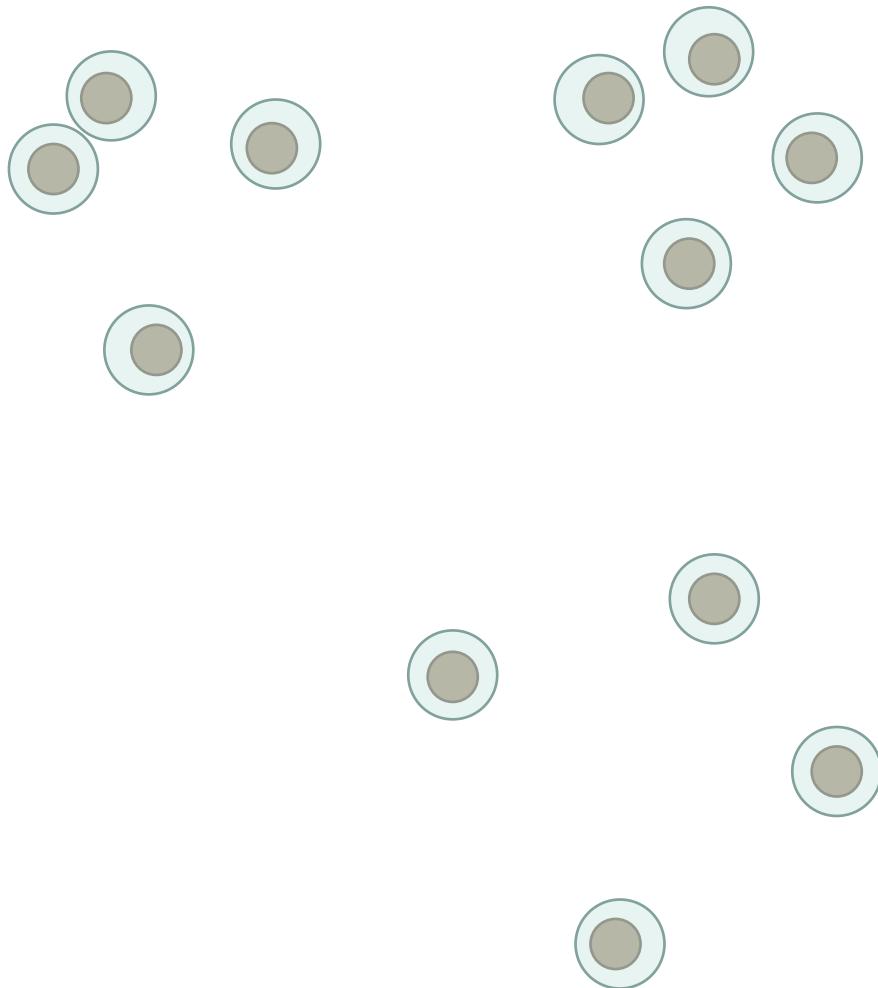
# Иерархическая агломеративная кластеризация

Алгоритмы  
кластеризации



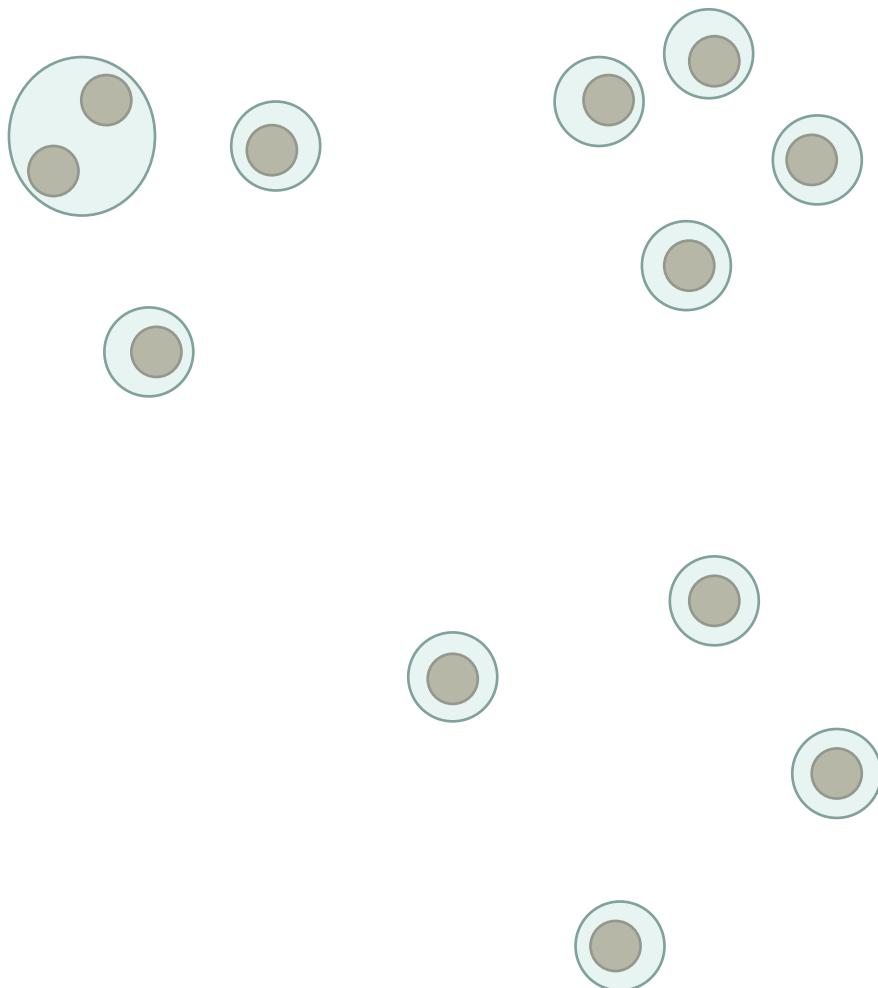
# Иерархическая агломеративная кластеризация

Алгоритмы  
кластеризации



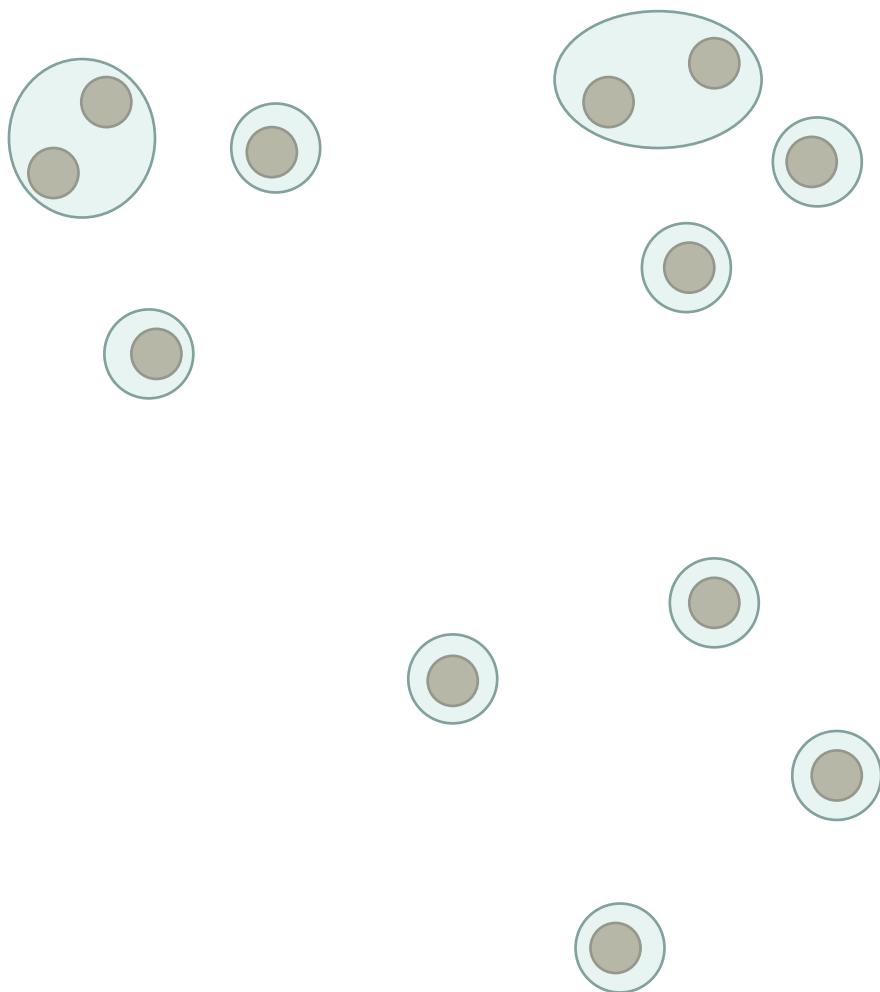
# Иерархическая агломеративная кластеризация

Алгоритмы  
кластеризации



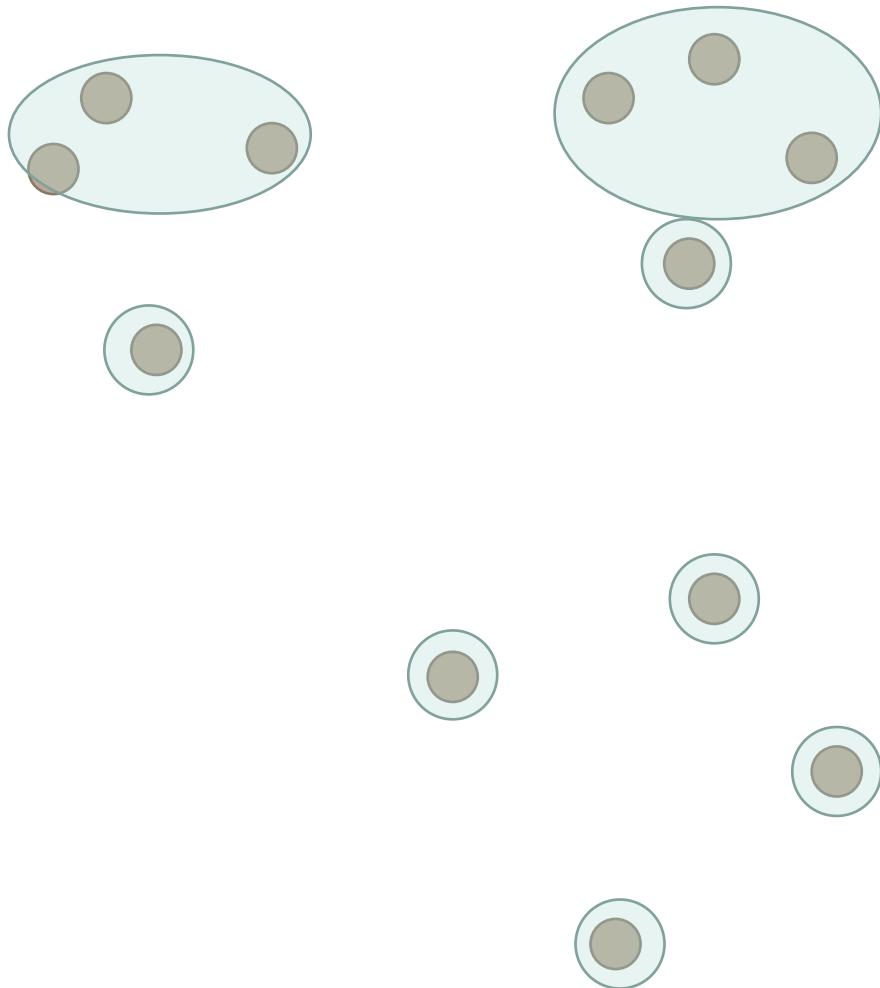
# Иерархическая агломеративная кластеризация

Алгоритмы  
кластеризации



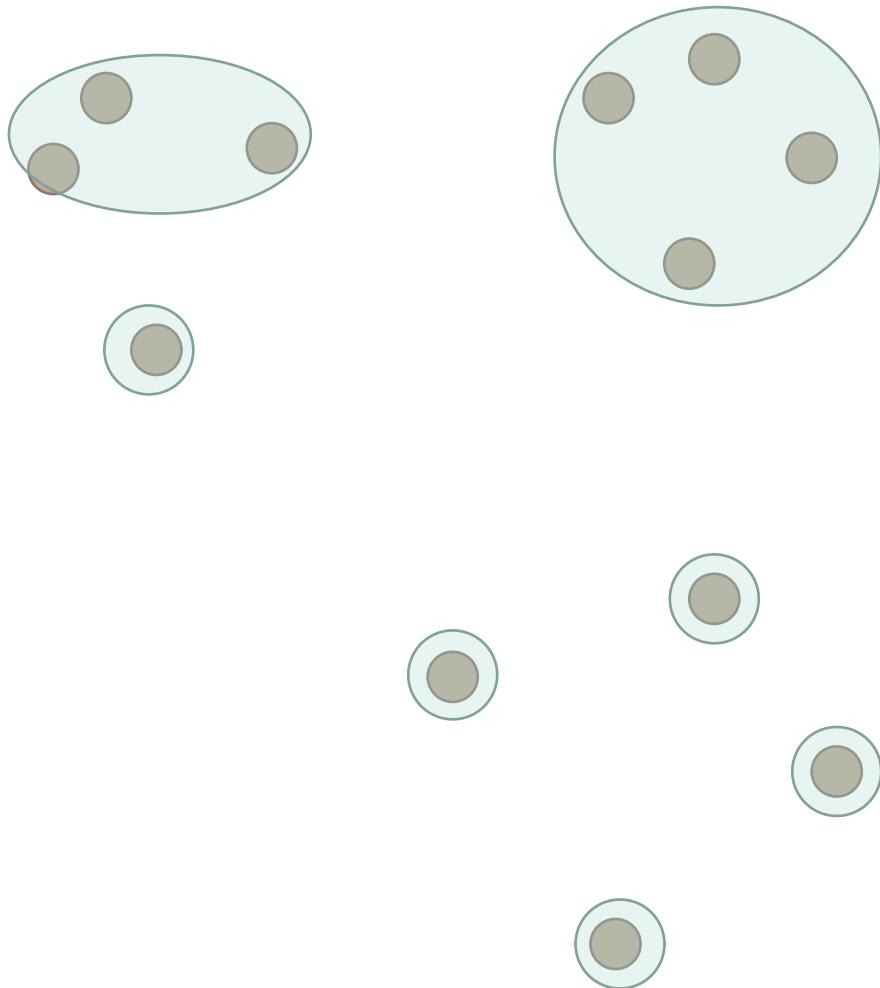
Алгоритмы  
кластеризации

# Иерархическая агломеративная кластеризация



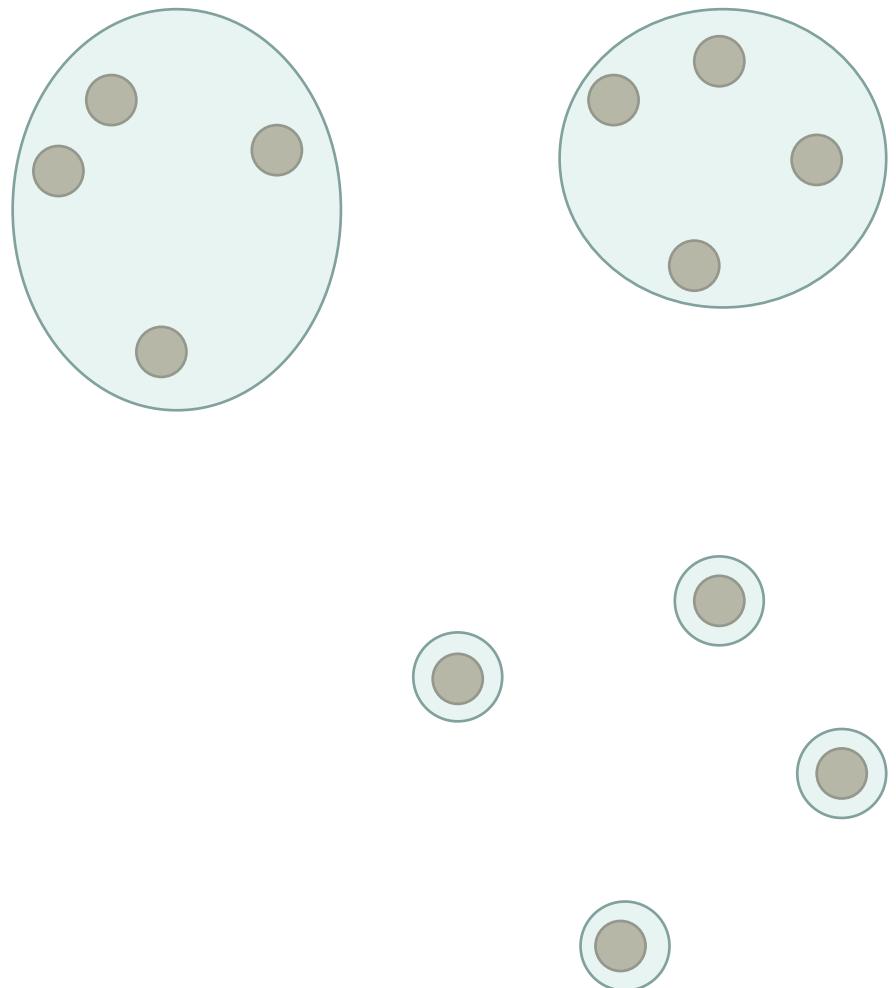
Алгоритмы  
кластеризации

# Иерархическая агломеративная кластеризация



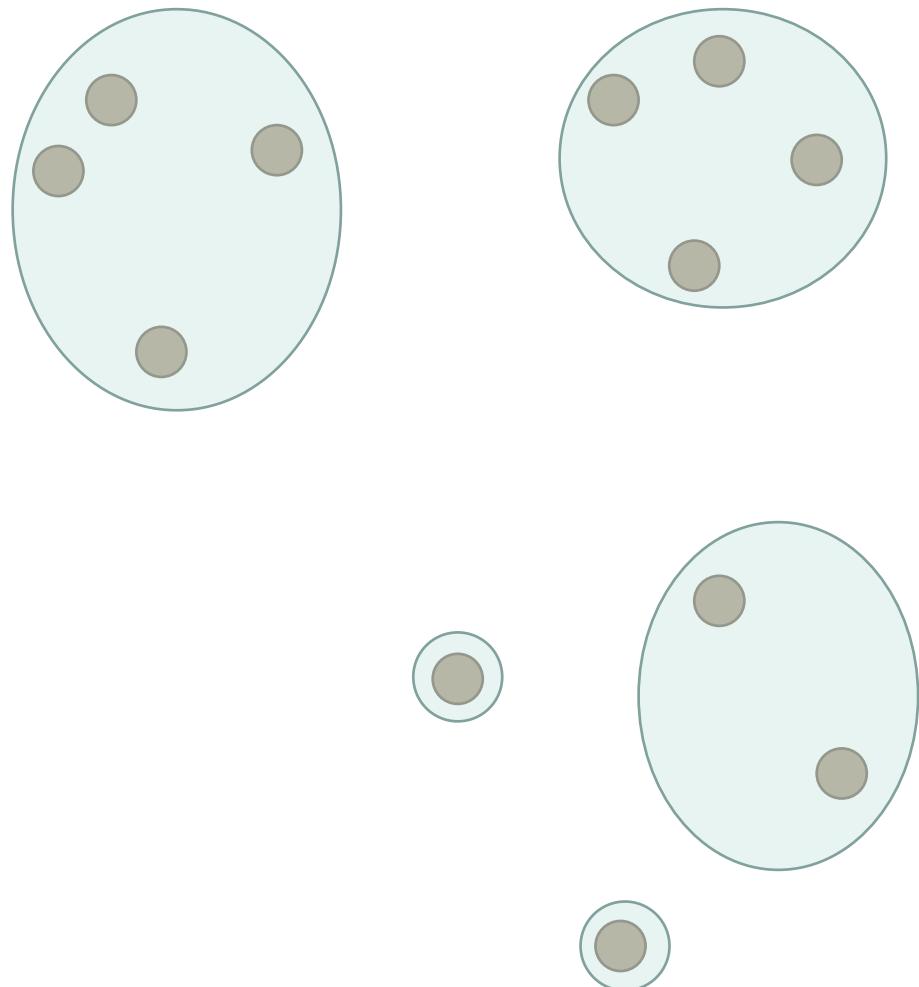
Алгоритмы  
кластеризации

# Иерархическая агломеративная кластеризация



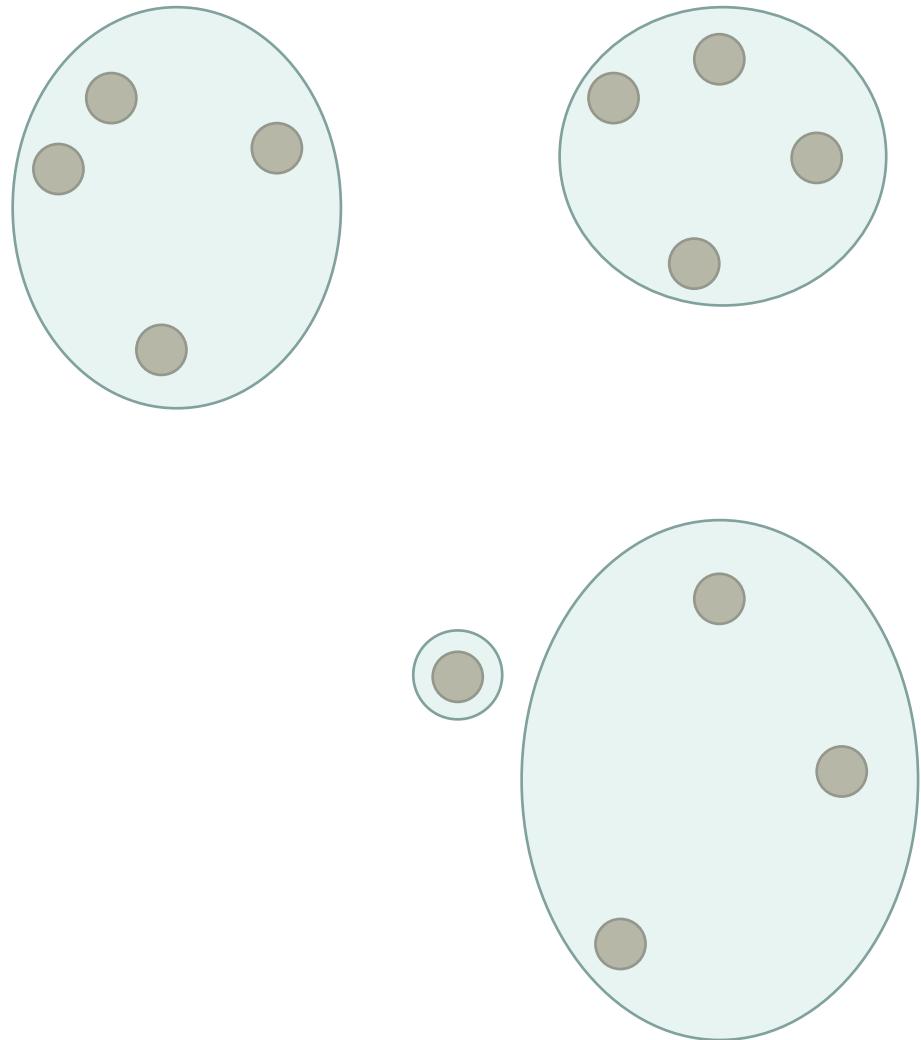
Алгоритмы  
кластеризации

# Иерархическая агломеративная кластеризация



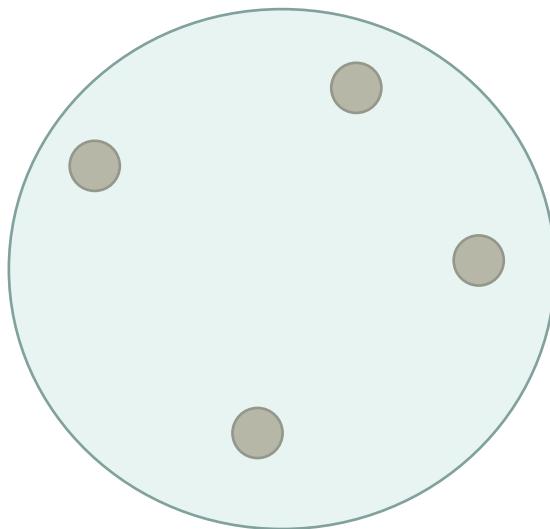
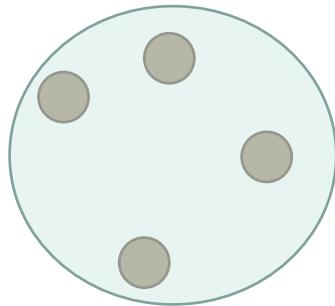
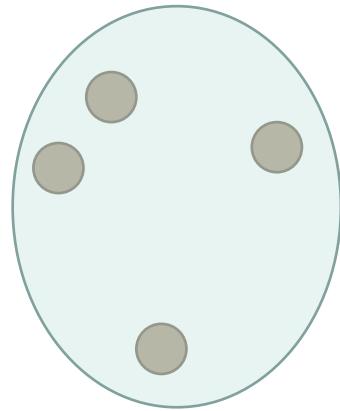
Алгоритмы  
кластеризации

# Иерархическая агломеративная кластеризация



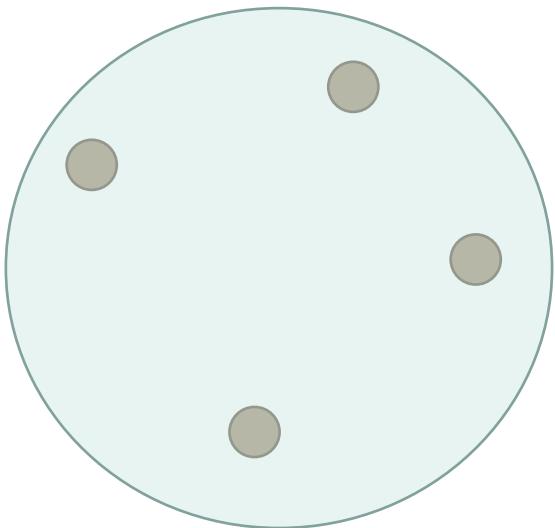
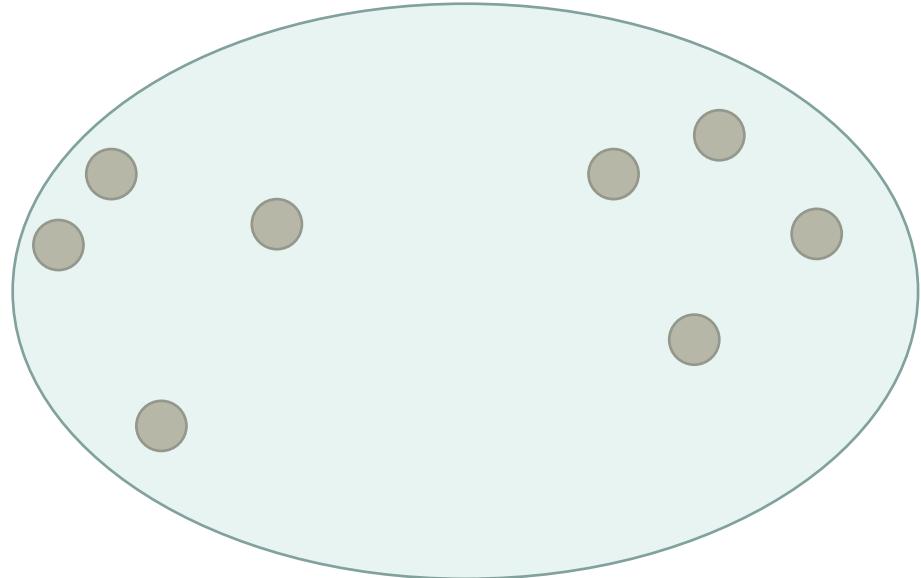
Алгоритмы  
кластеризации

# Иерархическая агломеративная кластеризация



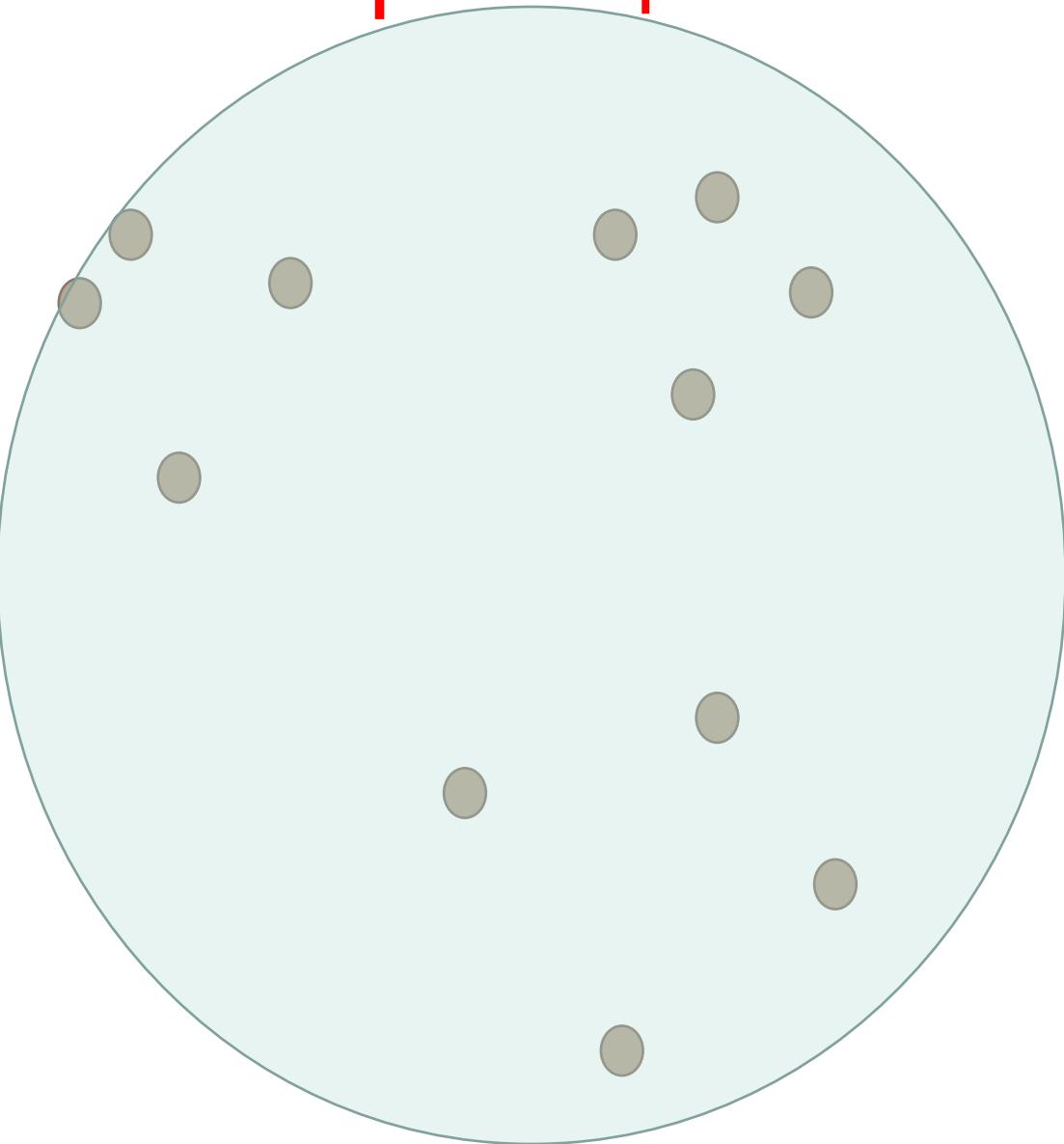
Алгоритмы  
кластеризации

# Иерархическая агломеративная кластеризация



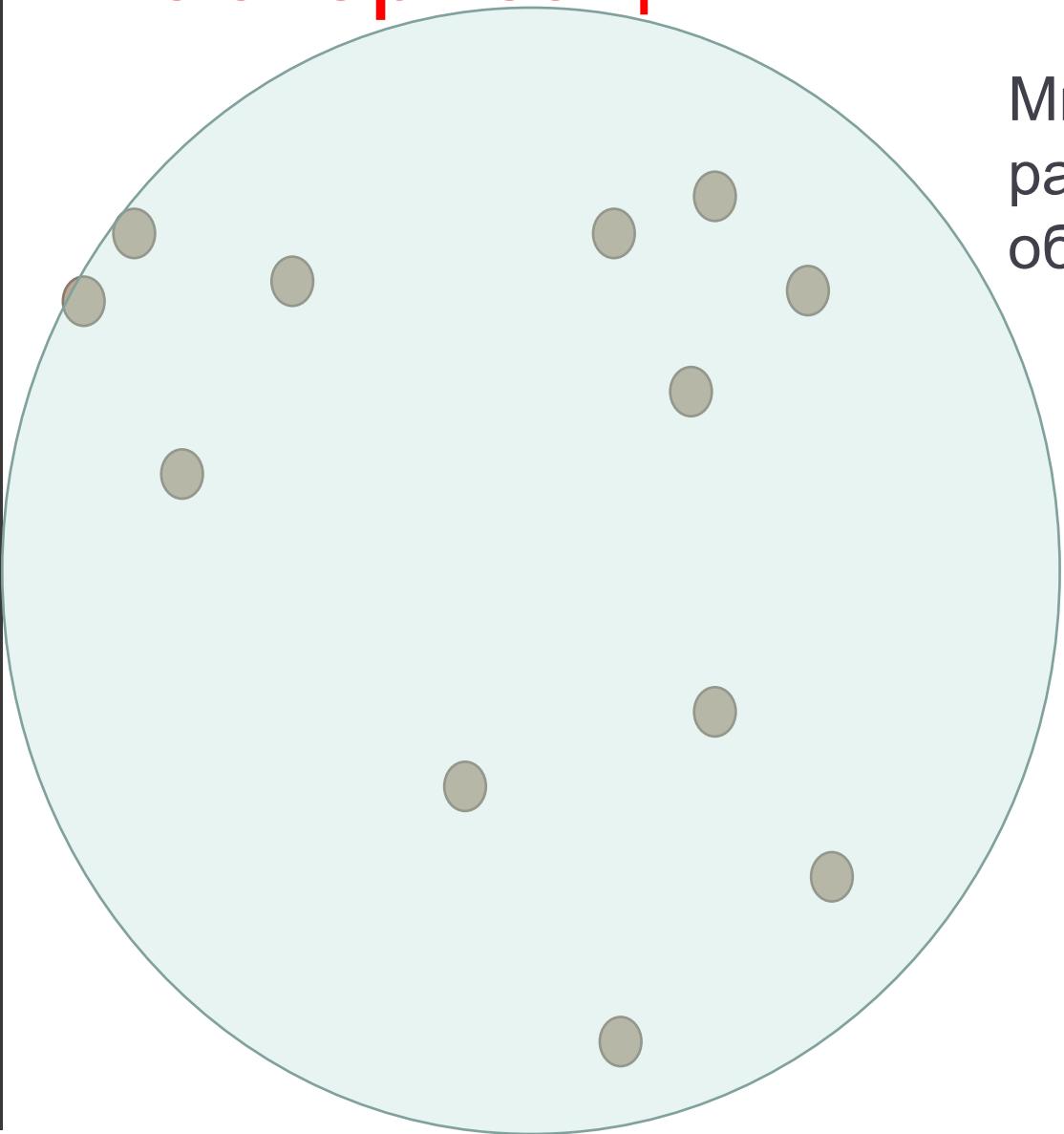
# Иерархическая агломеративная кластеризация

Алгоритмы  
кластеризации



# Иерархическая агломеративная кластеризация

Алгоритмы  
кластеризации



Мы умеем считать  
расстояние между  
объектами.

А как оценить  
расстояние между  
кластерами?

# Межкластерное расстояние

Формула Ланса-Уильямса:

$$\begin{aligned} R(U \cup V, S) \\ = \alpha_U R(U, S) + \alpha_V R(V, S) + \beta R(U, V) + \gamma |(R(U, S) - R(V, S)| \end{aligned}$$

Алгоритмы  
кластеризации

# Алгоритмы кластеризации

## Межклusterное расстояние

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:*

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

*Расстояние дальнего соседа:*

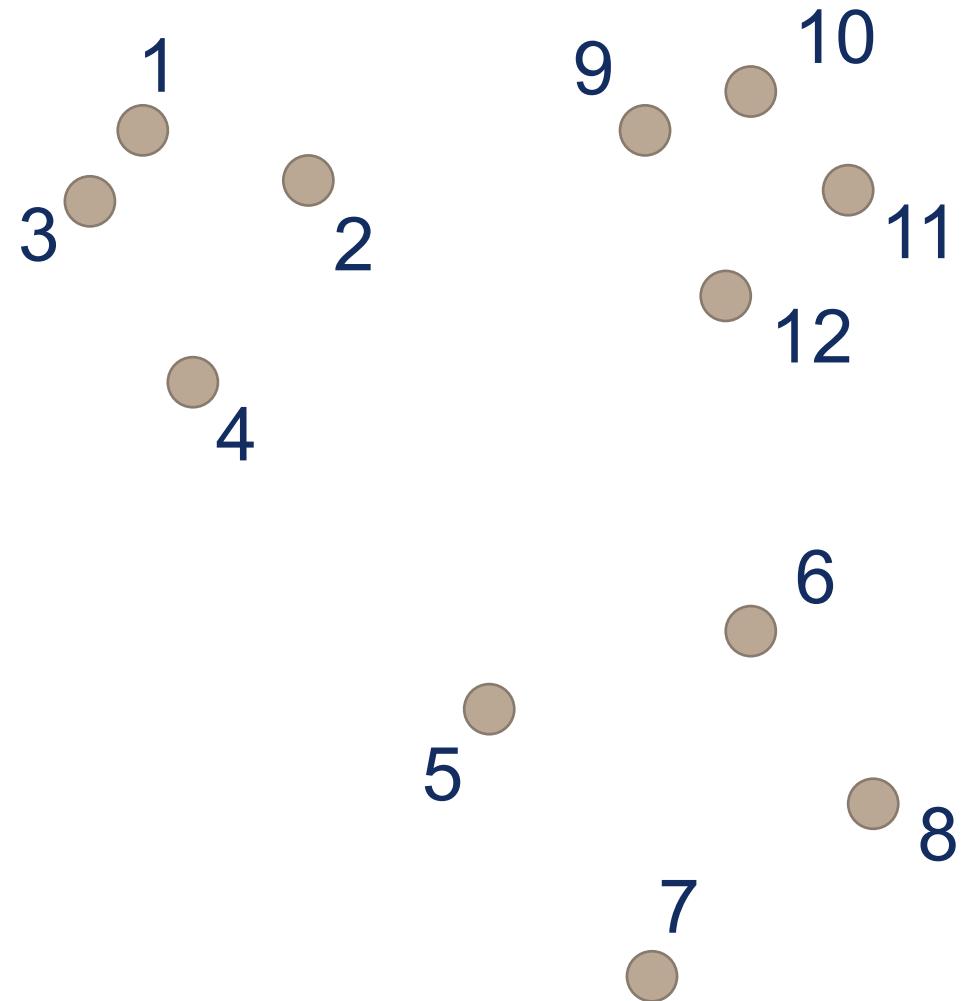
$$R^d(W, S) = \max_{w \in W, s \in S} \rho(w, s); \quad \alpha_U = \alpha_V = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}.$$

*Среднее расстояние:*

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

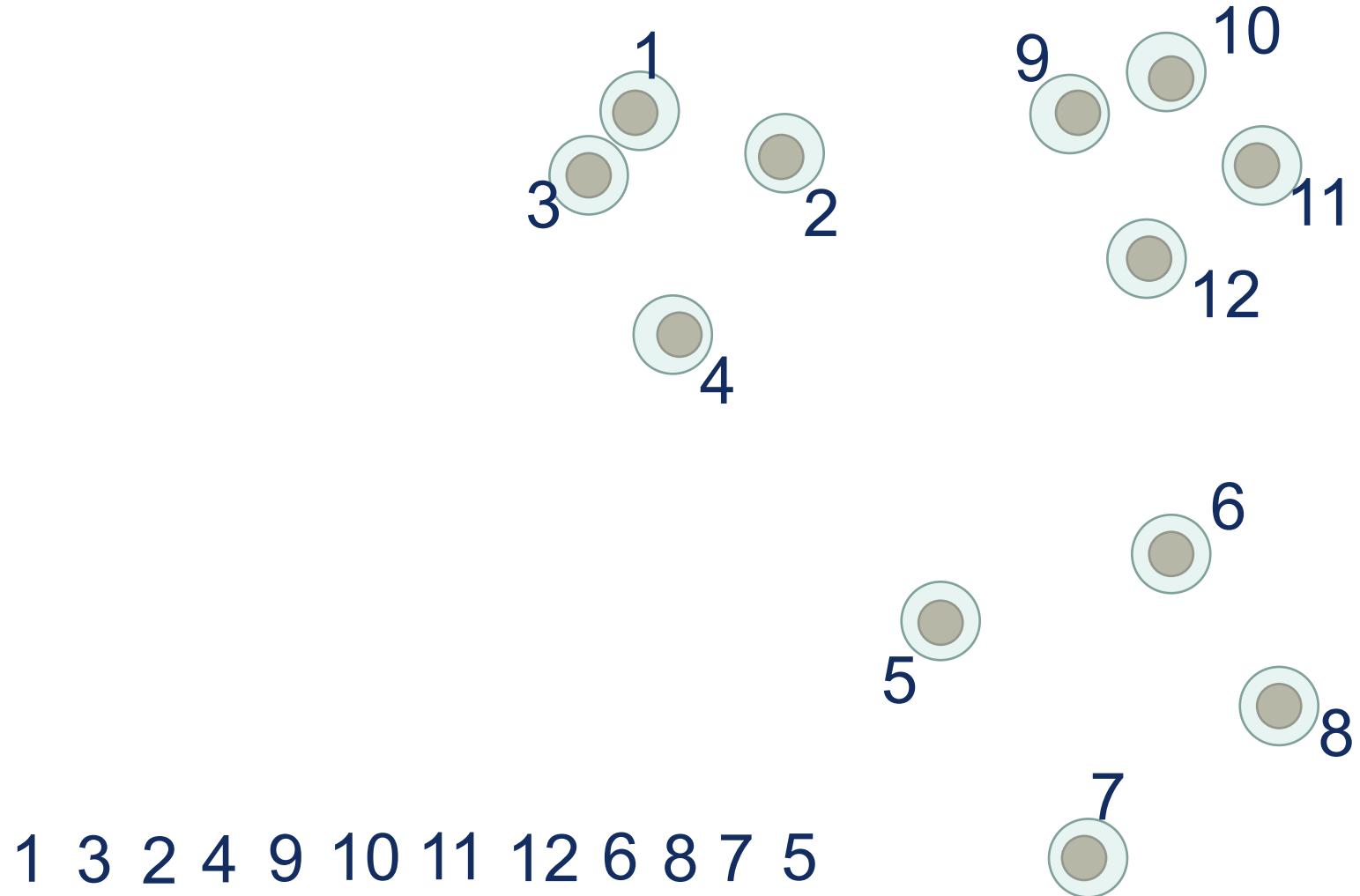
# Алгоритмы кластеризации

## Дендрограмма



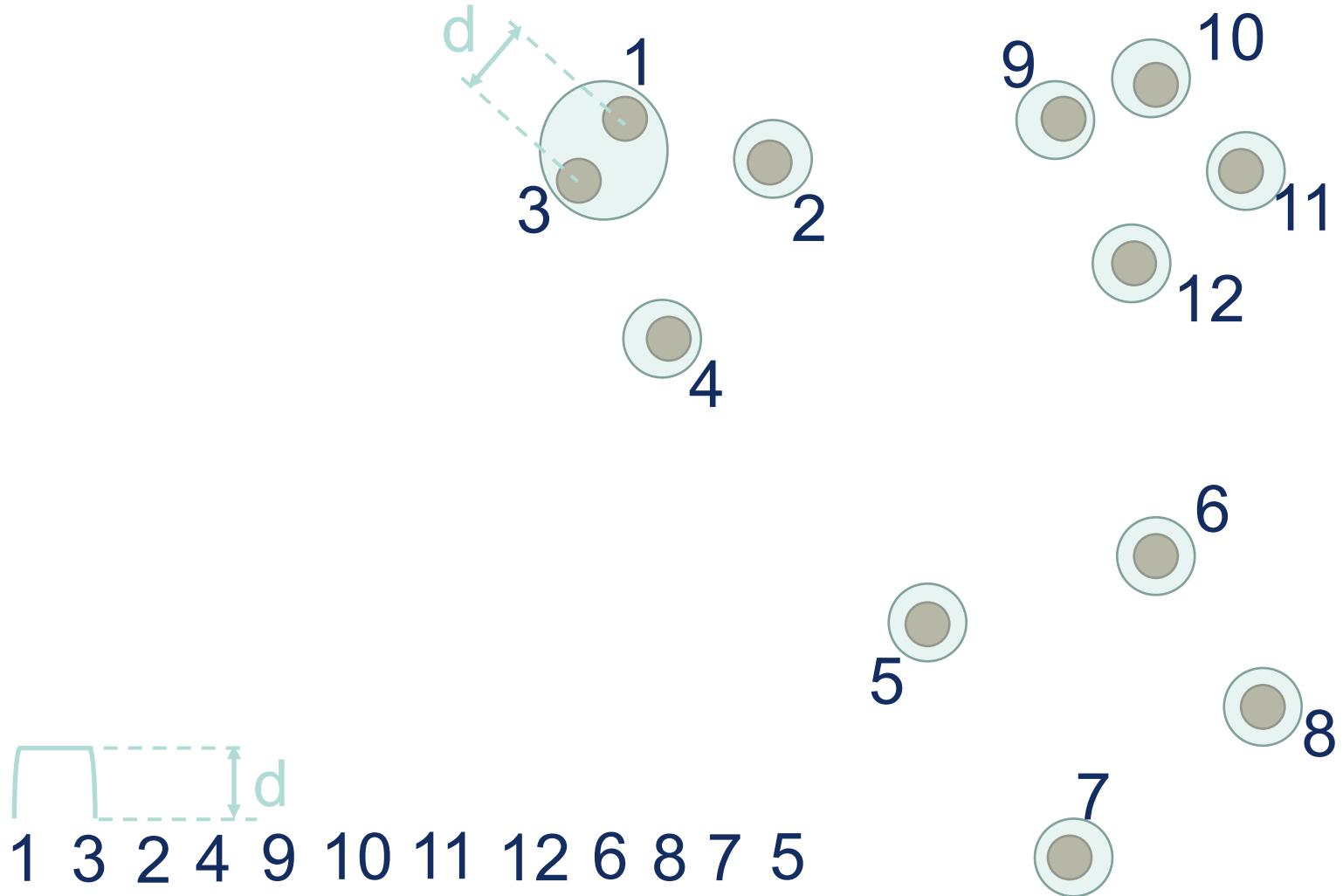
## Алгоритмы кластеризации

# Дендрограмма



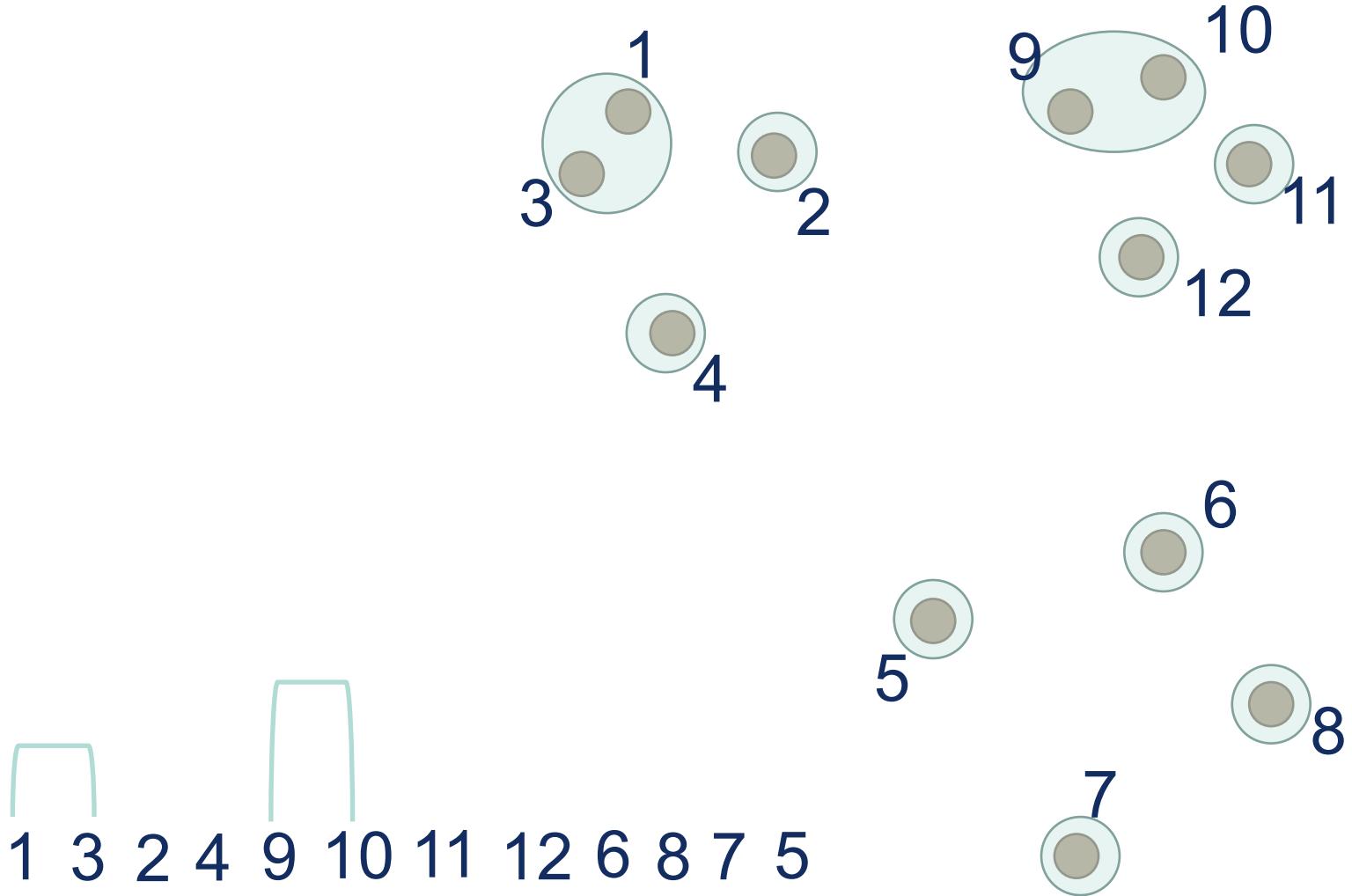
## Алгоритмы кластеризации

# Дендрограмма



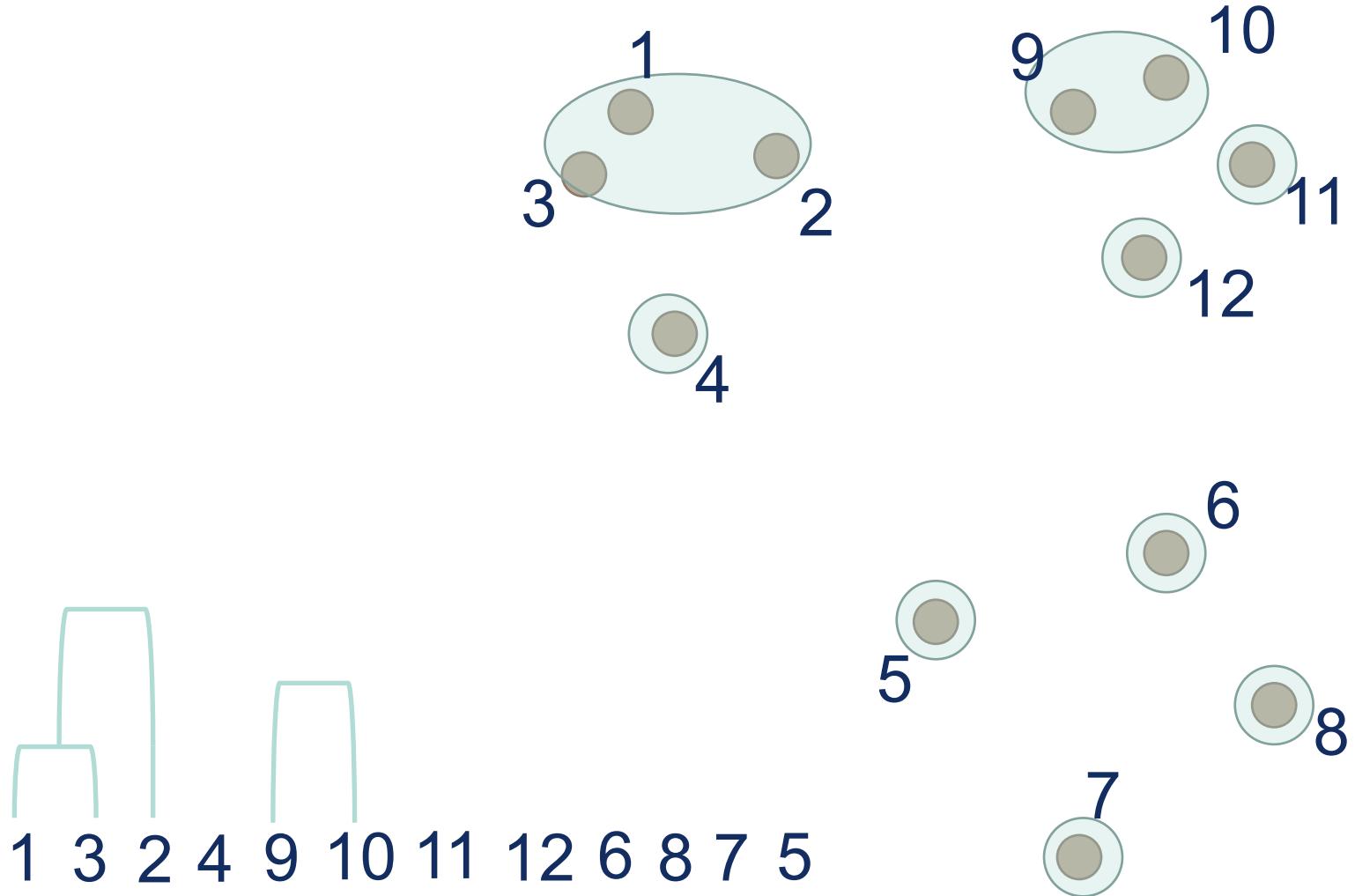
## Алгоритмы кластеризации

# Дендрограмма



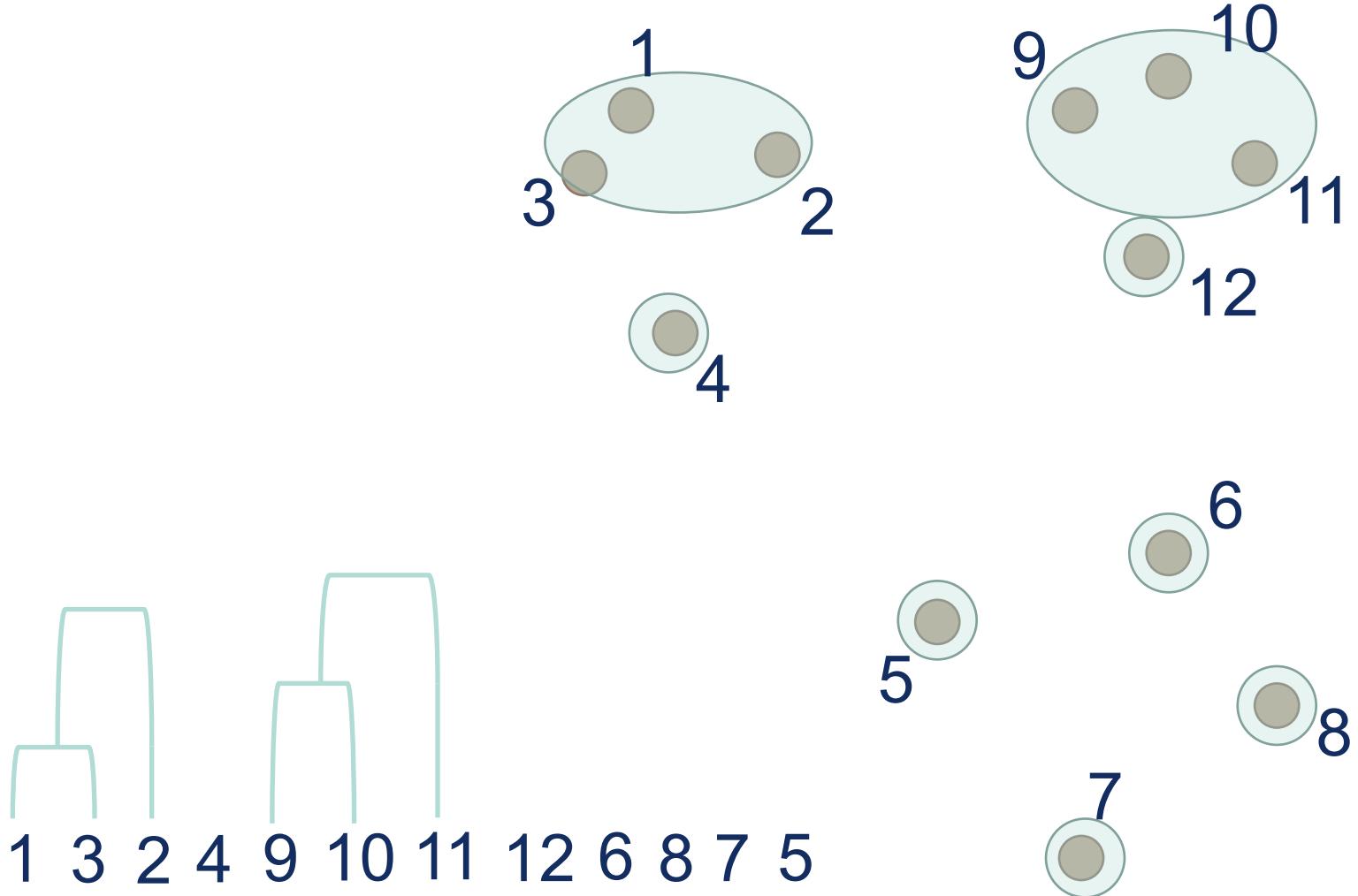
## Алгоритмы кластеризации

# Дендрограмма



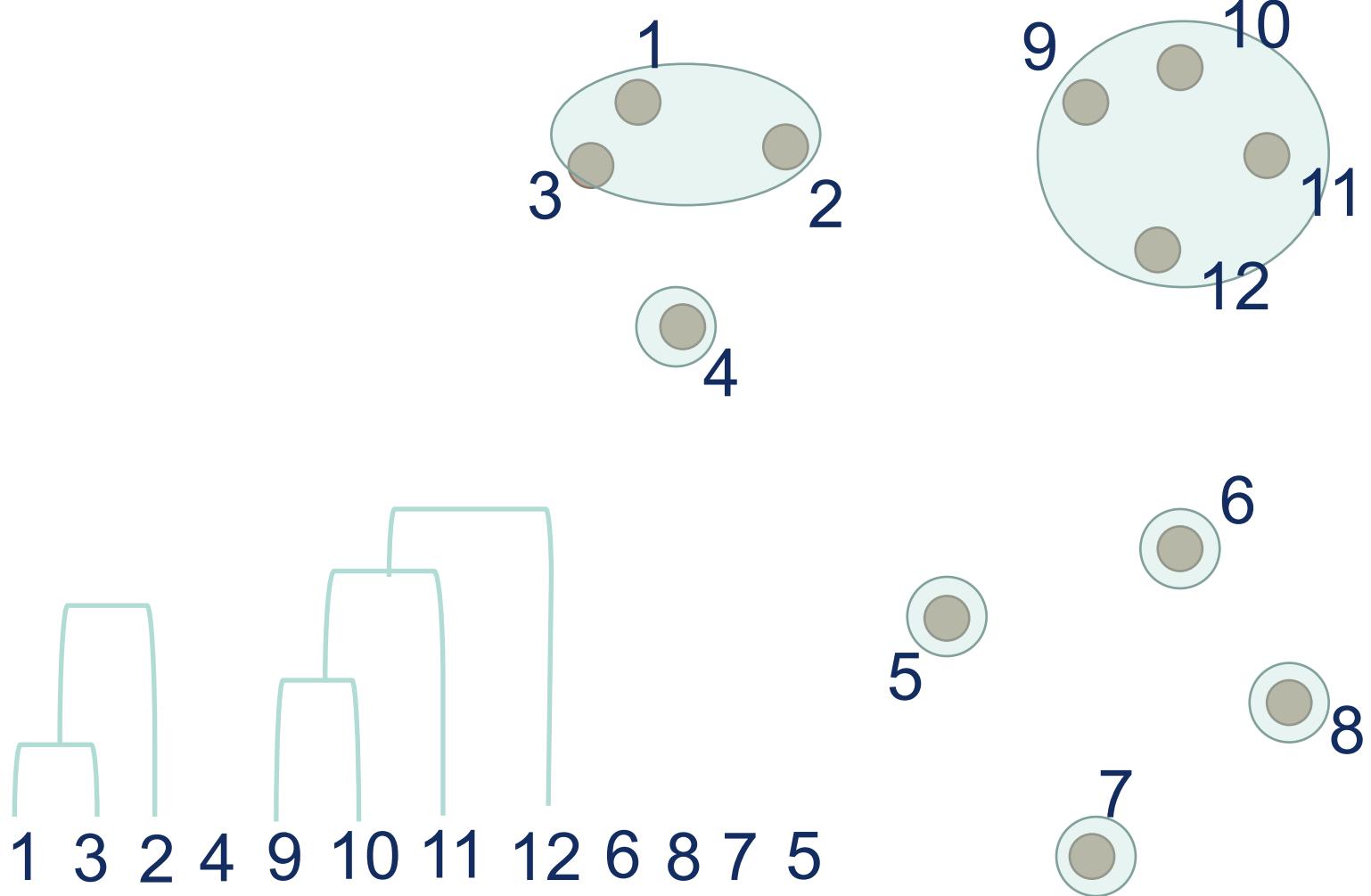
## Алгоритмы кластеризации

# Дендрограмма



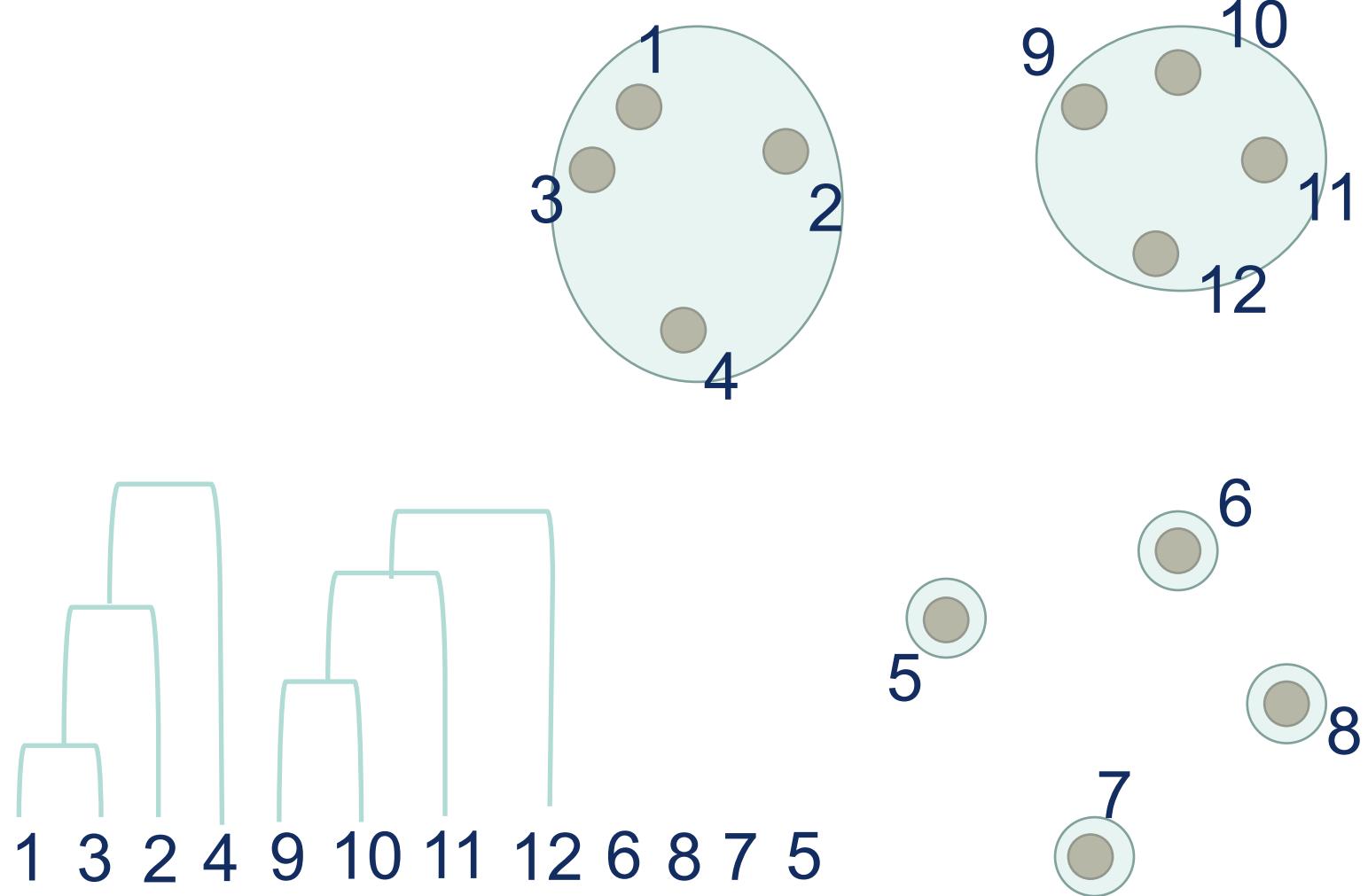
# Алгоритмы кластеризации

## Дендрограмма



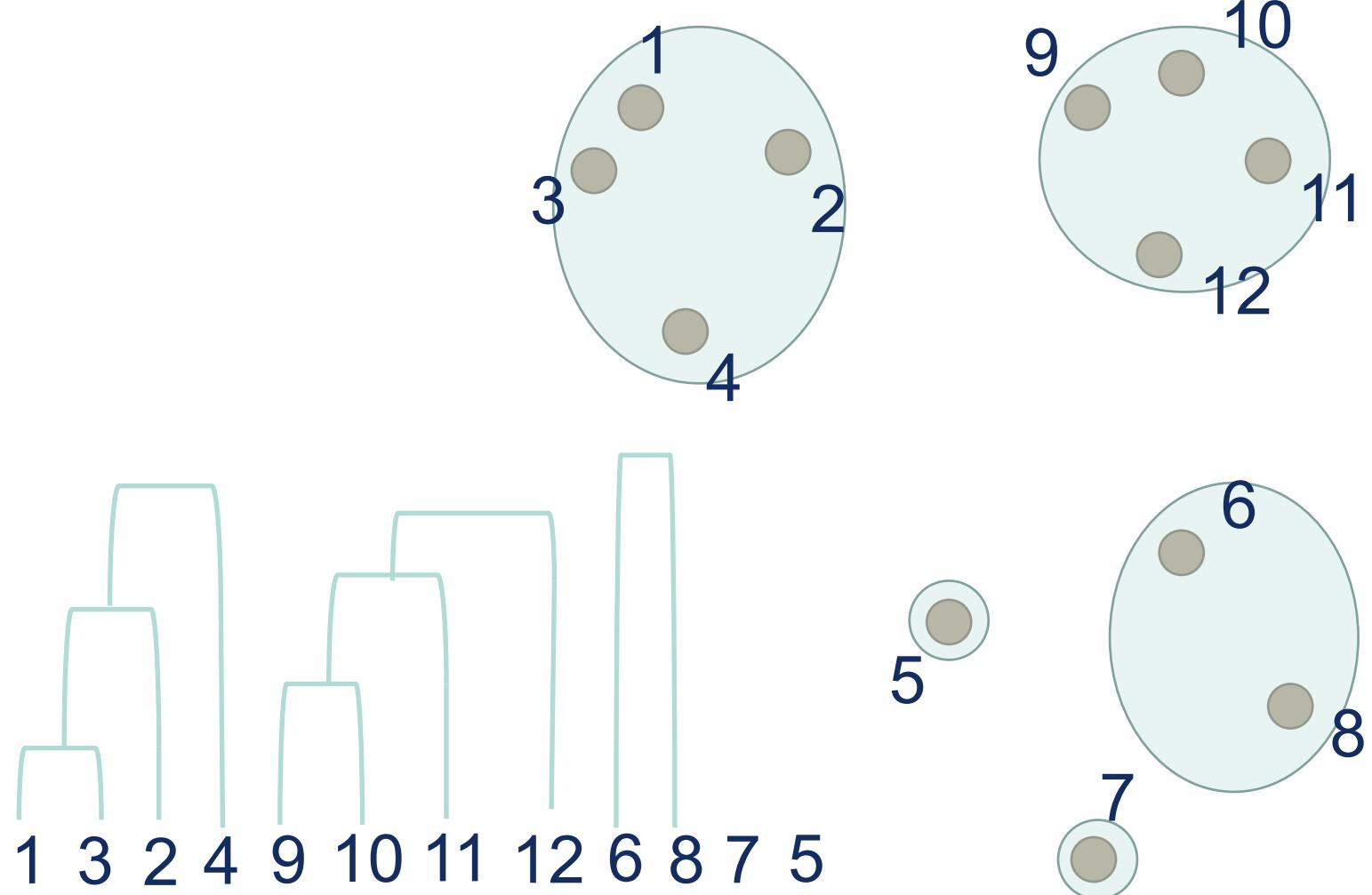
# Алгоритмы кластеризации

## Дендрограмма



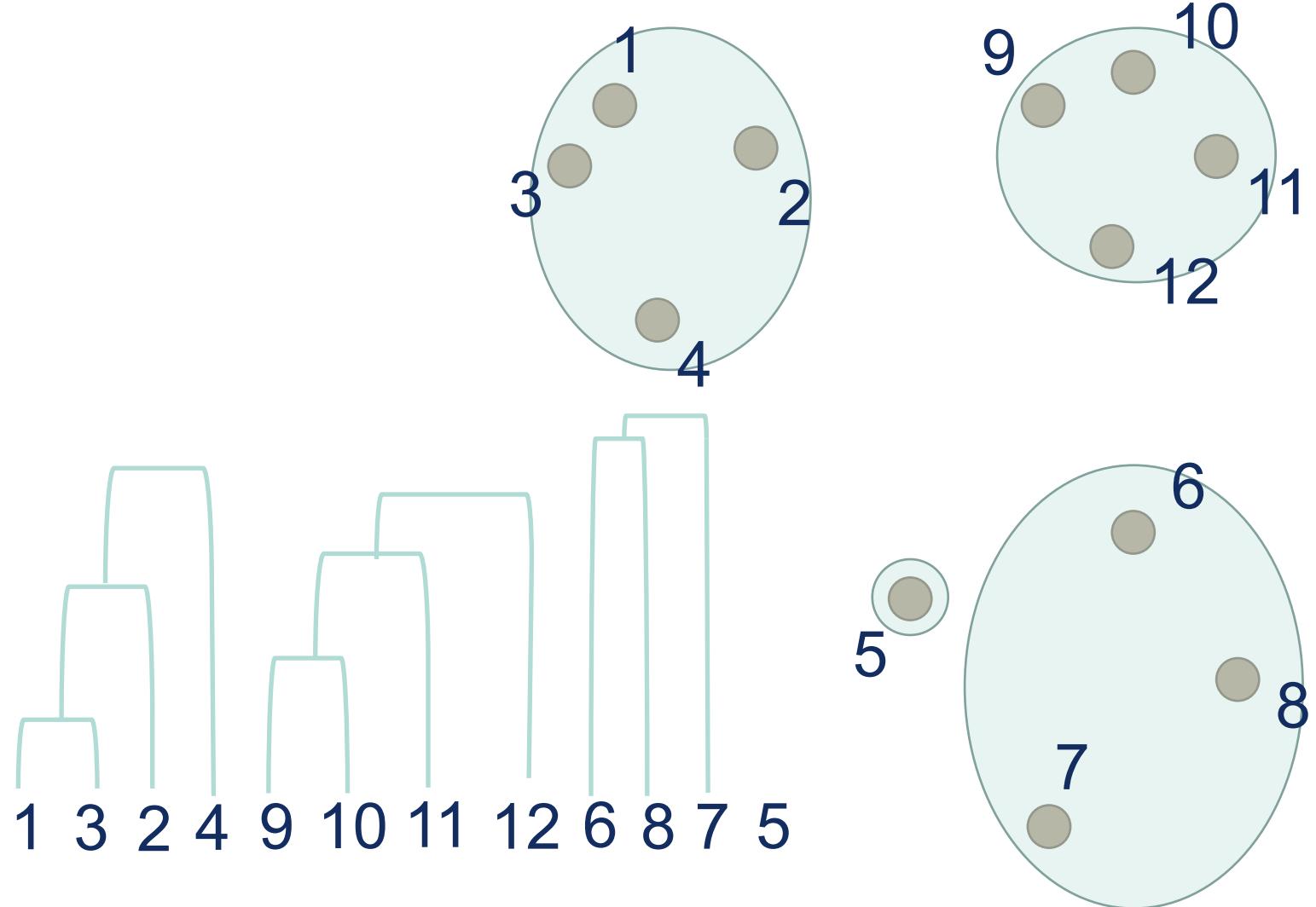
# Алгоритмы кластеризации

## Дендрограмма



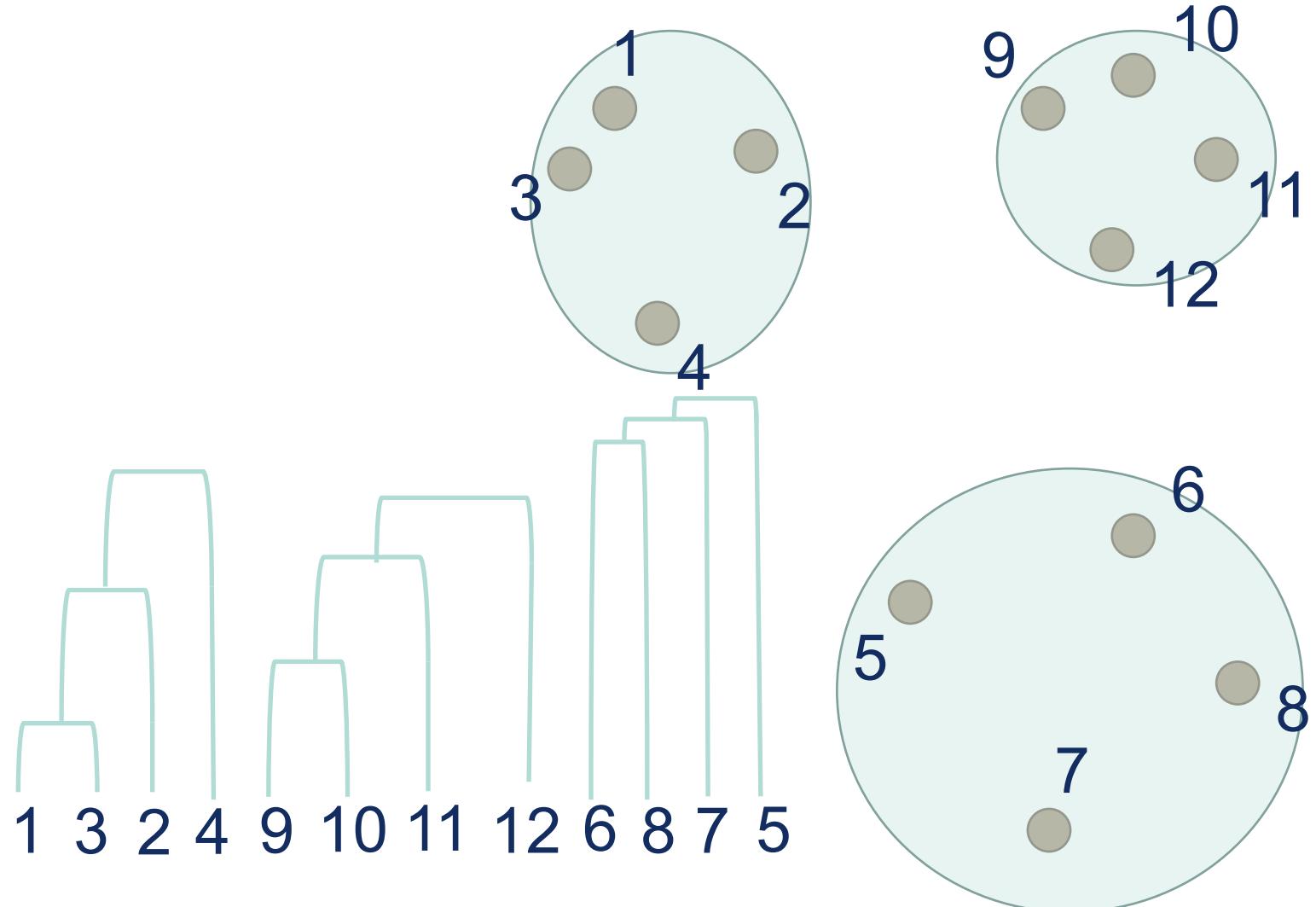
# Алгоритмы кластеризации

## Дендрограмма



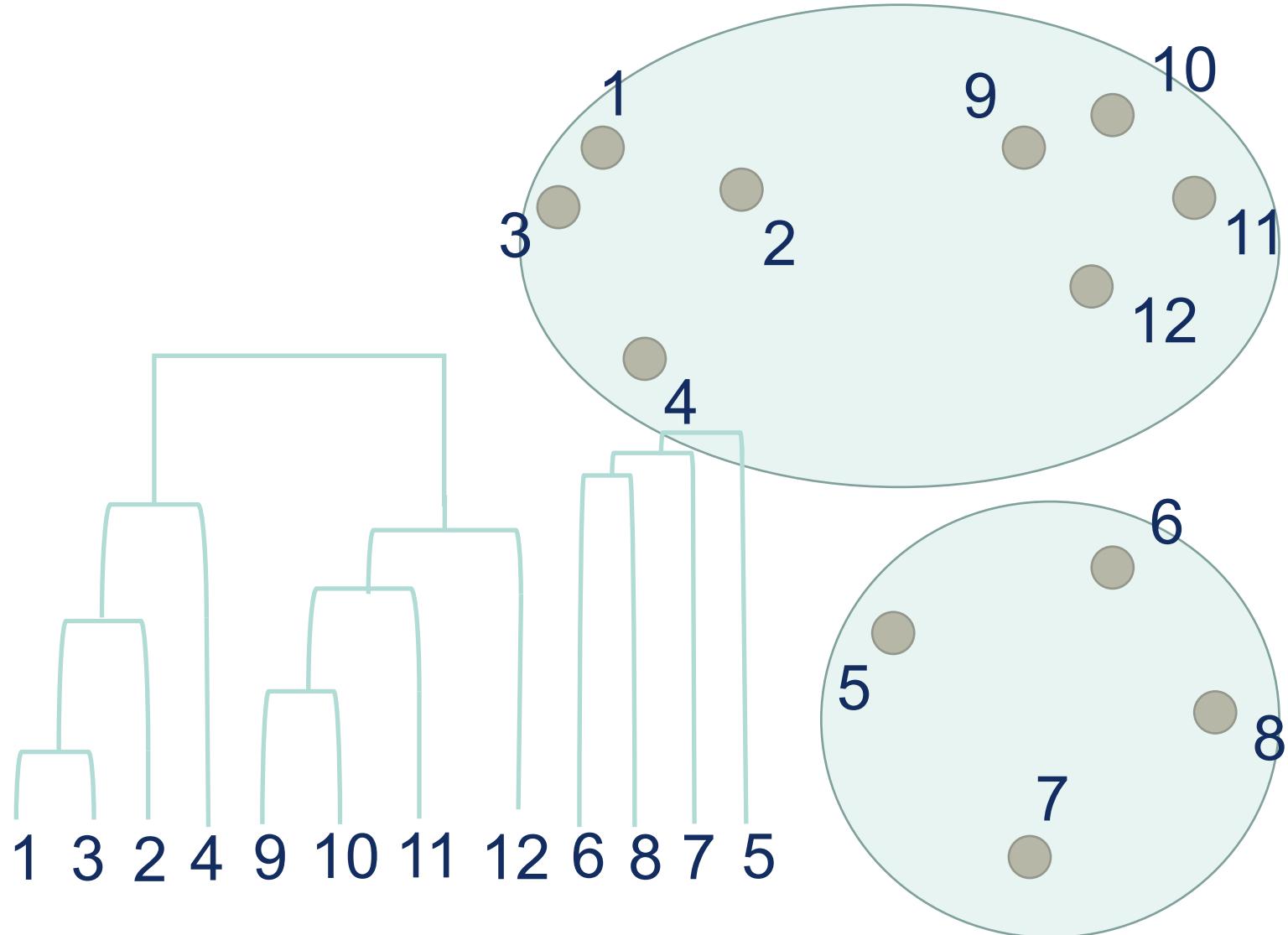
# Алгоритмы кластеризации

## Дендрограмма



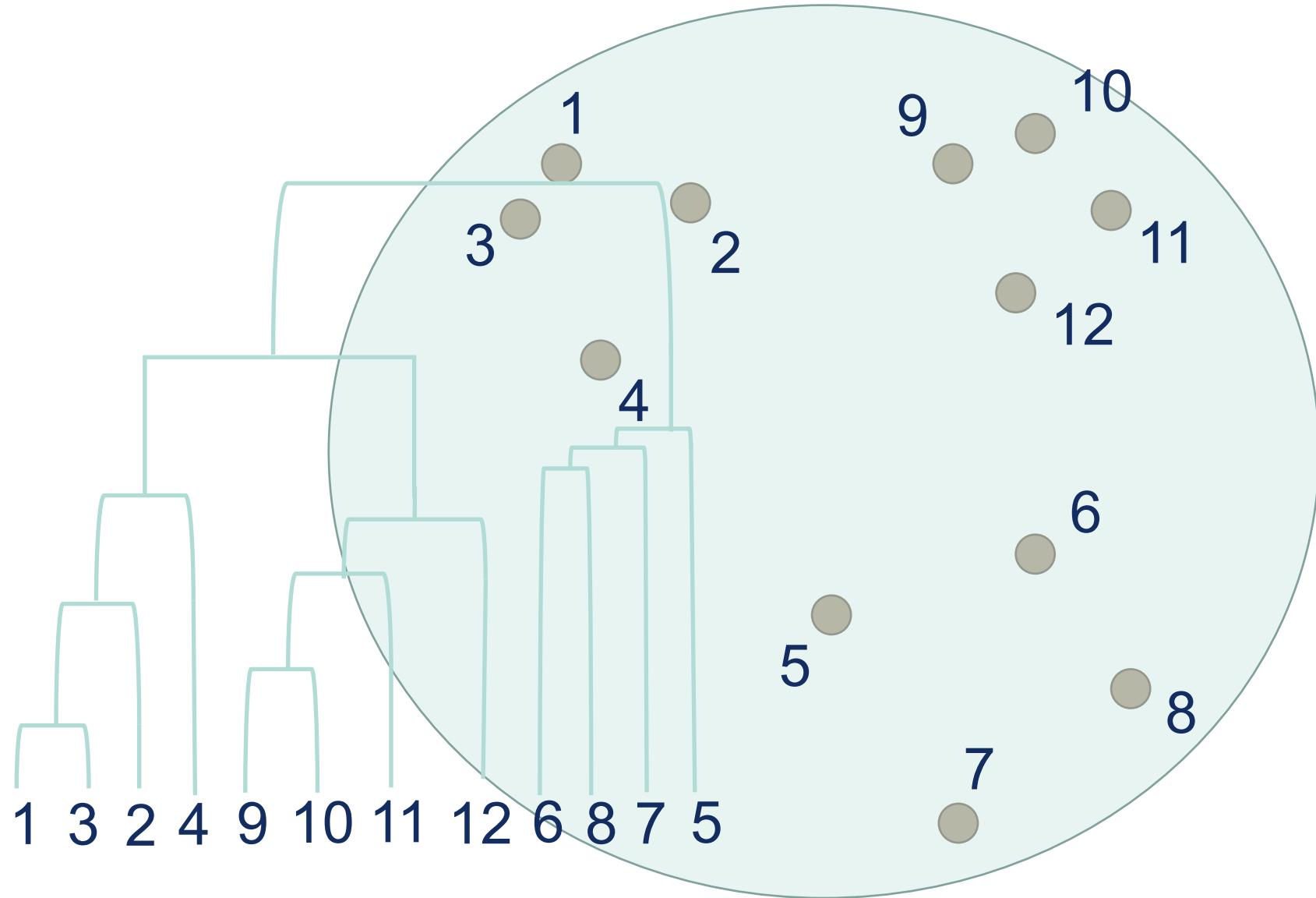
# Алгоритмы кластеризации

## Дендрограмма



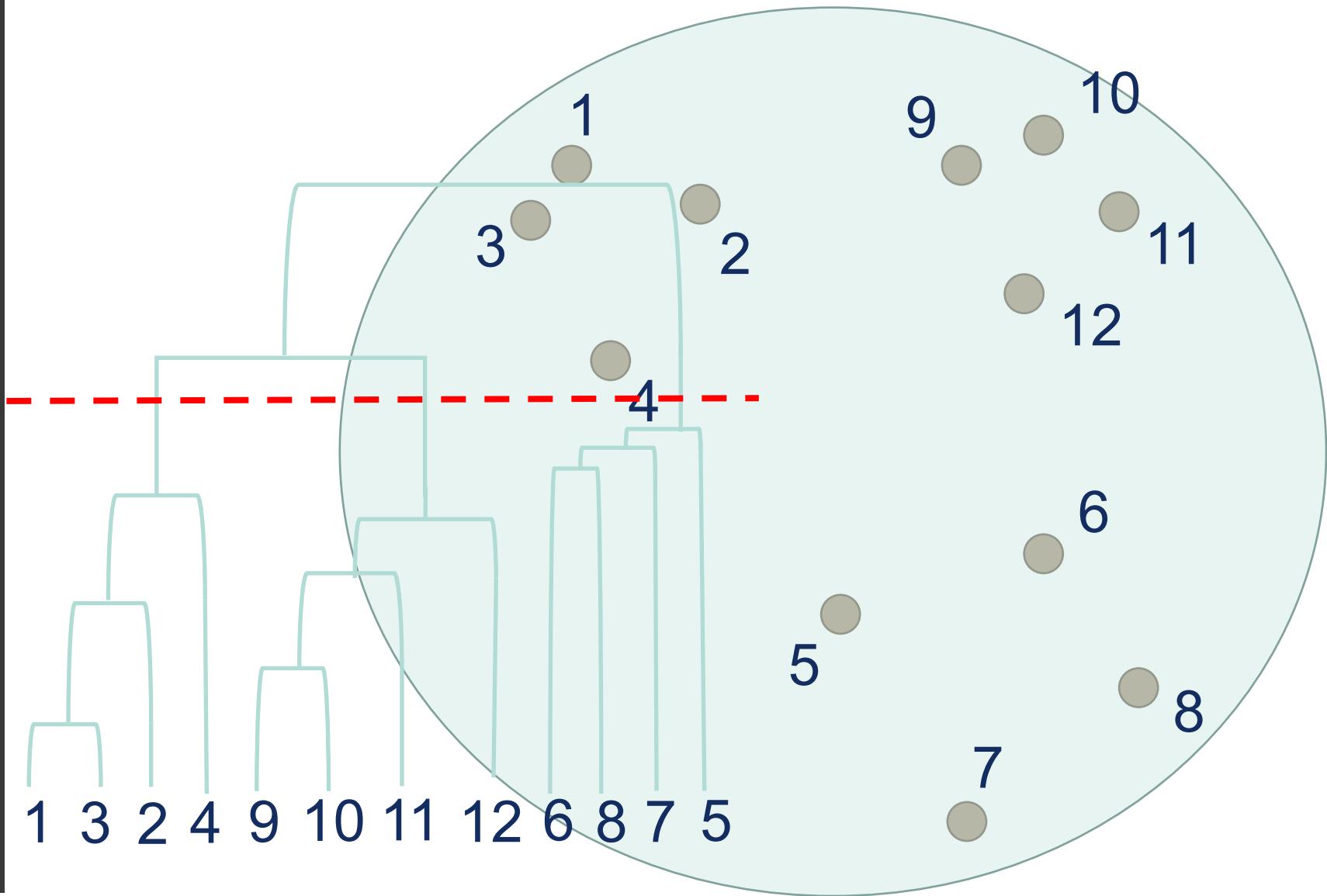
## Алгоритмы кластеризации

# Дендрограмма



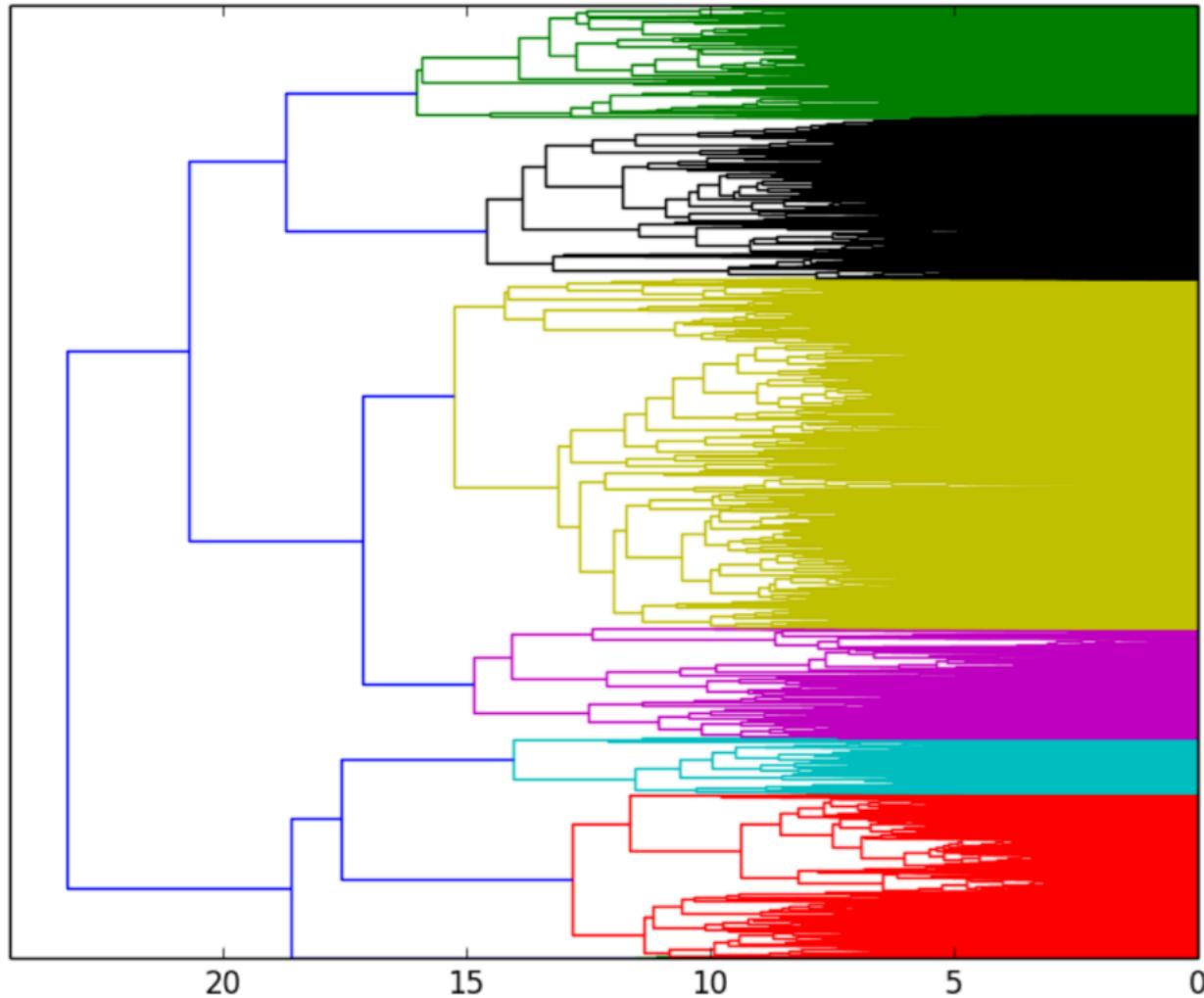
## Алгоритмы кластеризации

# Дендрограмма



# Алгоритмы кластеризации

## Пример: кластеризация писем



# Алгоритмы понижения размерности

# Понижение размерности

## Постановка задачи

Что требуется сделать?

1. Сгенерируем новые признаки, на основе исходных
2. Новых признаков должно быть меньше, чем исходных
3. Потеря информации должна быть минимальной: требуется сохранить закономерности в данных

# Понижение размерности

## Методы понижения размерности

Методы понижения размерности можно разделить на 2 группы:

1. Линейные: новые признаки представляют собой линейную комбинацию исходных
2. Нелинейные: новые признаки не являются линейной комбинацией исходных

# Понижение размерности

## Методы понижения размерности

Методы понижения размерности можно разделить на 2 группы:

1. Линейные: новые признаки представляют собой линейную комбинацию исходных
  - Random projections
  - PCA
2. Нелинейные: новые признаки не являются линейной комбинацией исходных
  - MDS (multi dimensional scaling)
  - SNE (stochastic neighbor embedding)
  - t-SNE (t-distributed stochastic neighbor embedding)

# Понижение размерности

## Методы понижения размерности

Методы понижения размерности можно разделить на 2 группы:

1. Линейные: новые признаки представляют собой линейную комбинацию исходных
  - Random projections
  - PCA
2. Нелинейные: новые признаки не являются линейной комбинацией исходных
  - MDS (multi dimensional scaling)
  - SNE (stochastic neighbor embedding)
  - t-SNE (t-distributed stochastic neighbor embedding)
3. Нейросетевой подход
  - Autoencoders

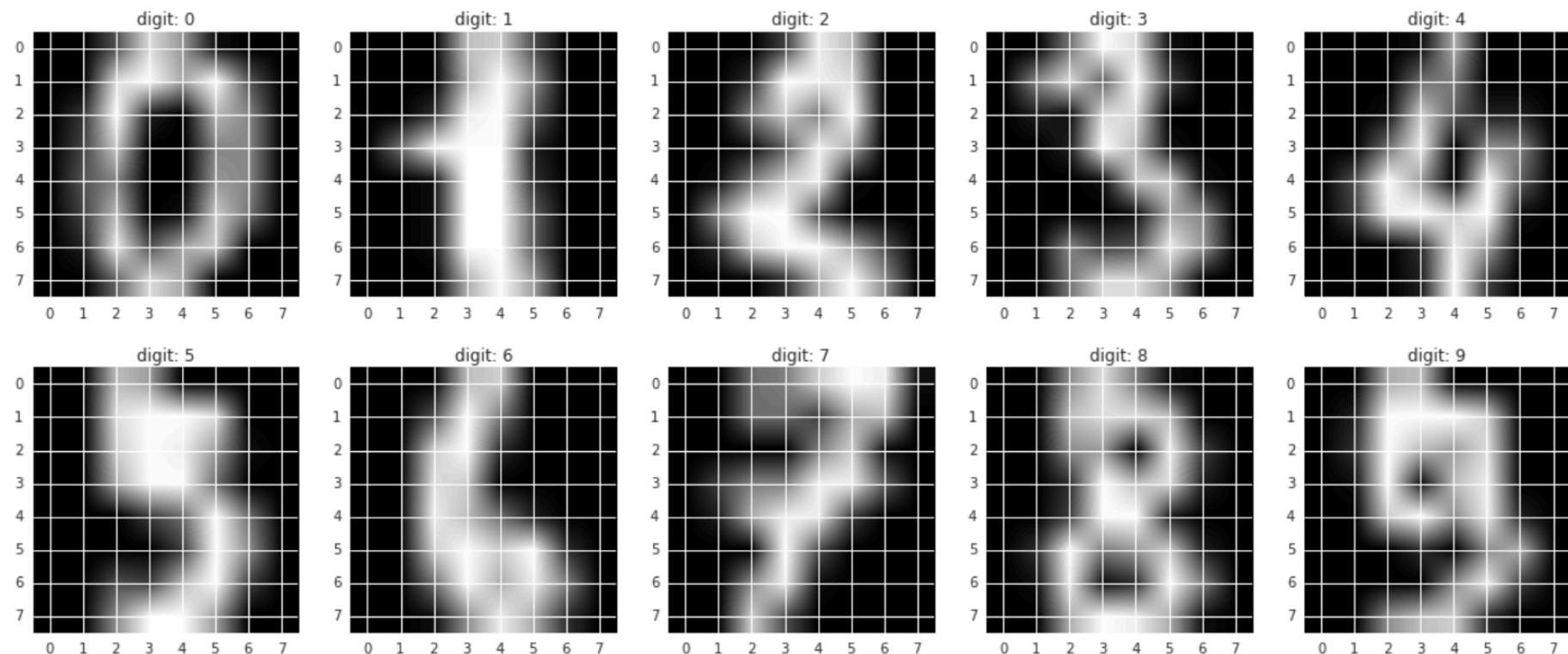
# Понижение размерности

## Пример: digits recognition

- 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

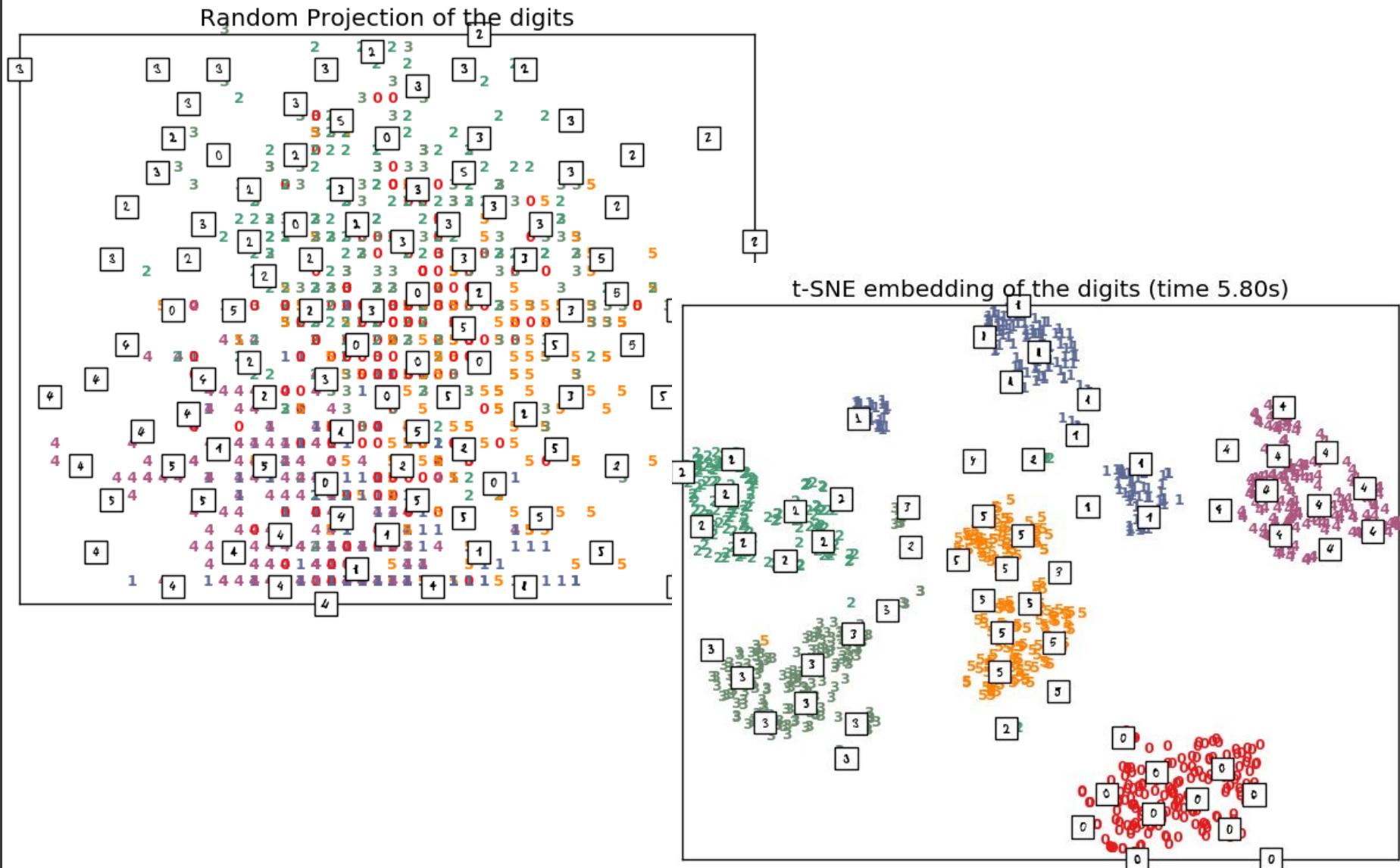
# Понижение размерности

## Пример: digits recognition



# Понижение размерности

## Пример: digits recognition



# Обучение без учителя

## To take away

- Задачи обучения без учителя: кластеризация, понижение размерности, поиск аномалий
- Кластеризация – некорректно поставленная задача. Для решения существует много алгоритмов и эвристик, однако результаты всегда субъективны
- DBSCAN – популярный алгоритм кластеризации
- Агломеративная кластеризация – популярный подход к иерархической кластеризации
- Методы кластеризации и классификации адаптируются к частичному обучению, методы кластеризации адаптируются более просто

# Машинное обучение: обучение без учителя

Спасибо!  
Эмели Драль