

Кластеризация

- Дано: матрица «объекты-признаки» X
- Найти:
 1. Множество кластеров Y
 2. Алгоритм кластеризации $a(x)$, который приписывает каждый объект к одному из кластеров
- Каждый кластер состоит из похожих объектов
- Объекты из разных кластеров существенно отличаются

Отличия

Обучение с учителем

- Цель: минимизация функционала ошибки
- Множество ответов известно заранее
- Конкретные способы измерения качества

Кластеризация

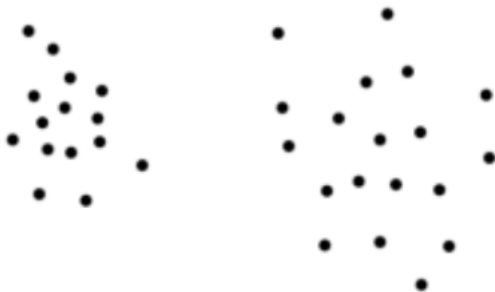
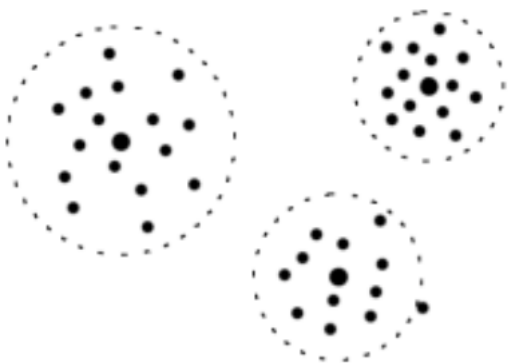
- Нет строгой постановки
- Множество кластеров неизвестно
- Правильные ответы отсутствуют — нельзя измерить качество

Зачем кластеризовать?

- Маркетинг: искать похожих клиентов
 - Модерация: проверять только одно сообщение из кластера
 - Соц. опросы: выделять группы схожих анкет
 - Соц. сети: искать сообщества
-
- Выявлять типы людей и формировать поведенческие паттерны для каждого типа

Виды кластеризации

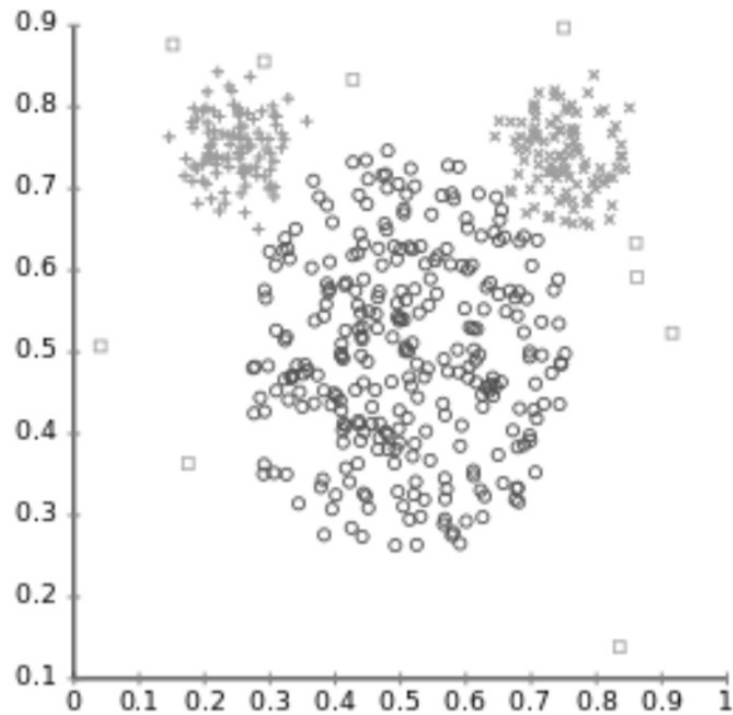
Форма кластеров



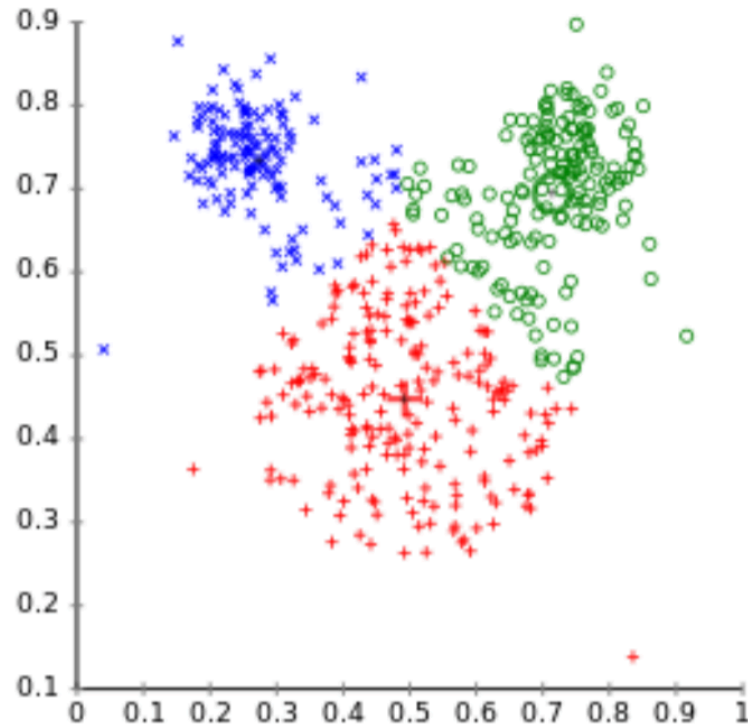
Форма кластеров



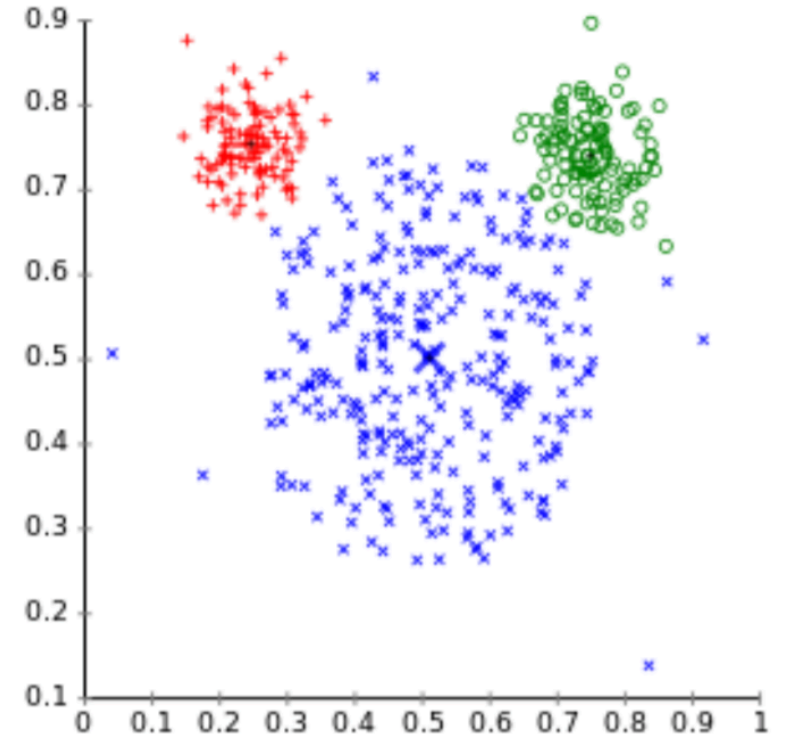
Различия в результатах работы



Исходная выборка
("Mouse" dataset)

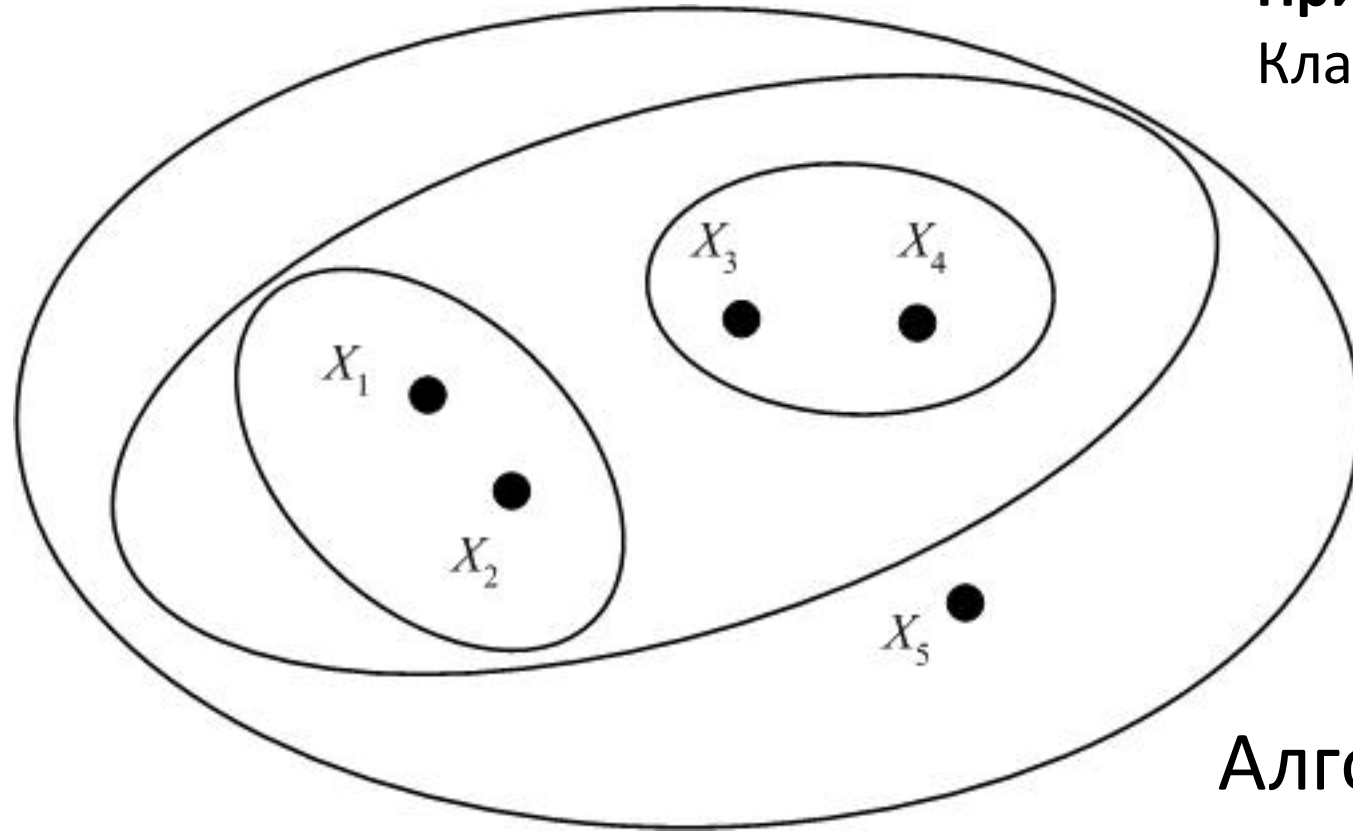


Метод 1



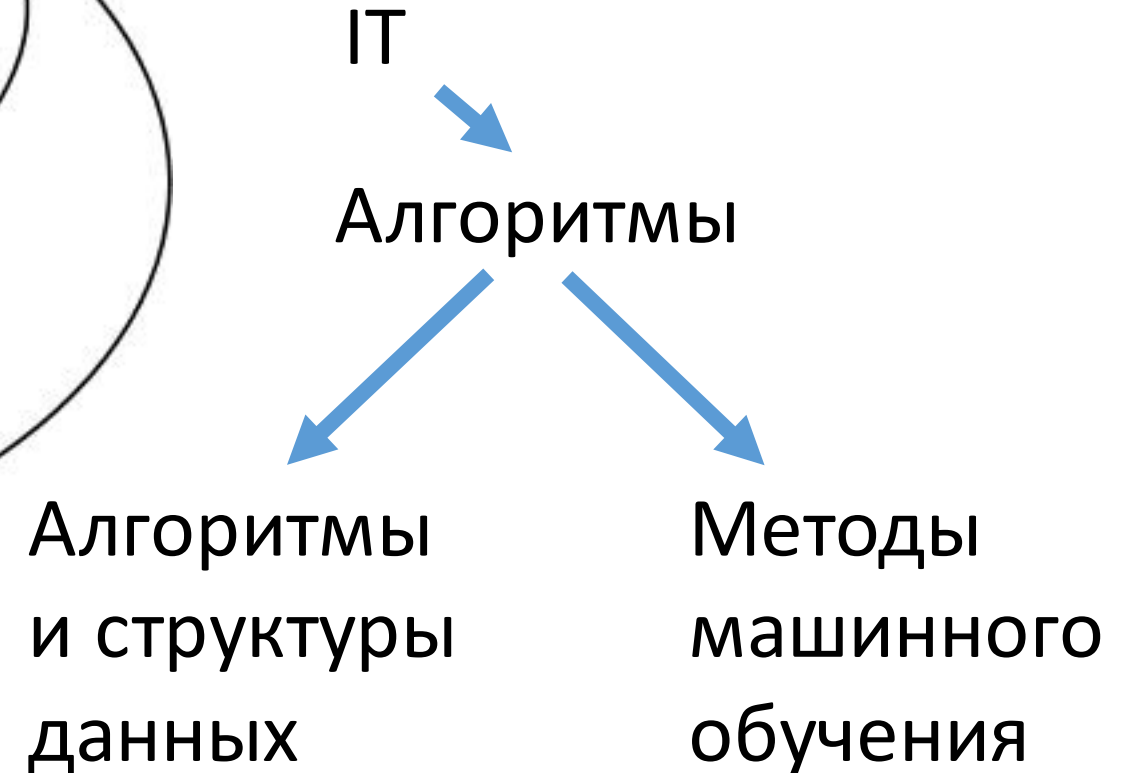
Метод 2

Иерархическая кластеризация



Пример:

Кластеризация статей с Хабрахабра



Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 1: в один кластер должны попадать новости на одну тему



Батыршин сыграет вместо Хабарова у «Магнитки» в матче с «Салаватом»

Место в третьей паре защиты «Магнитки» на третью встречу плей-офф Кубка Гагарина с «Салаватом Юлаевым» занял защитник Рафаэль Батыршин, сообщает из Уфы корреспондент «Чемпионата» Павел Панышев. Травмированный Ярослав Хабаров выбыл на неопределённый срок. Для форварда Оскара Осалы сезон закончен.



Футболисты ЦСКА проиграли «Долгопрудному» в товарищеском матче

Футболисты московского ЦСКА со счетом 2:3 проиграли клубу второго дивизиона "Долгопрудный" в товарищеском матче, который состоялся в Москве на стадионе "Октябрь". У армейцев забитыми мячами отличились Александр Цауня (15-я минута) и Сергей Ткачев (54).

Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 2: в один кластер должны попадать новости об одном «большом» событии



Керлингистки сборной РФ сделали правильные выводы после ОИ - Сидорова
10:38 26.03.2014



Путин призвал МВД использовать в Крыму опыт работы на Олимпиаде
14:13 21.03.2014



Два "олимпийских" спецавтопарка останутся в Сочи как наследие Игр
11:50 26.03.2014

Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 3: в один кластер должны попадать тексты об одной и той же новости

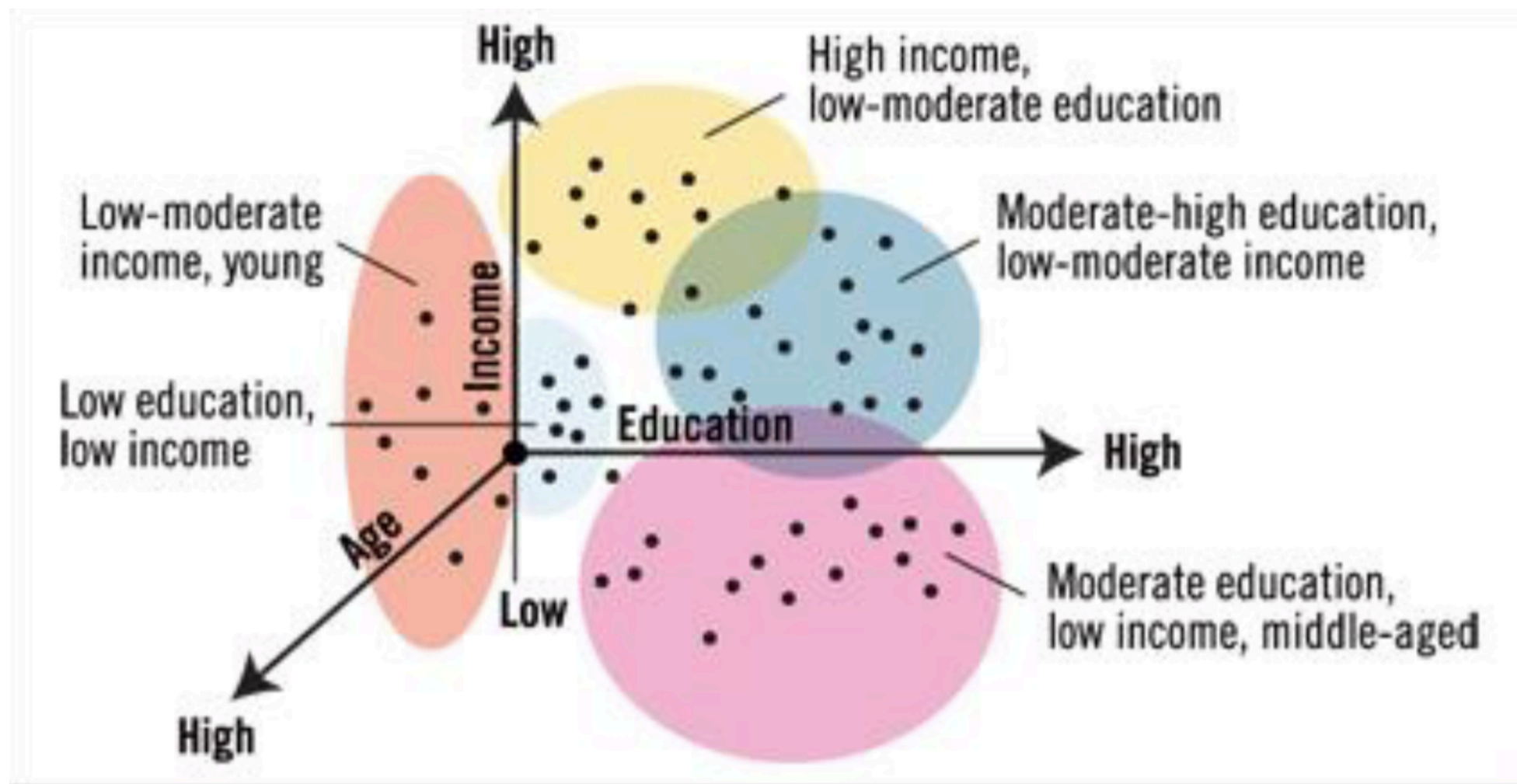
11:41, 08 ФЕВРАЛЯ 2014

Открытие Олимпиады в Сочи
посмотрели несколько миллиардов
человек

Олимпиада в Сочи открыта

**Церемония открытия Олимпиады в
Сочи. Онлайн-репортаж**

Вспомогательные задачи



Вспомогательные задачи

Цель: улучшение распознавания

5

5

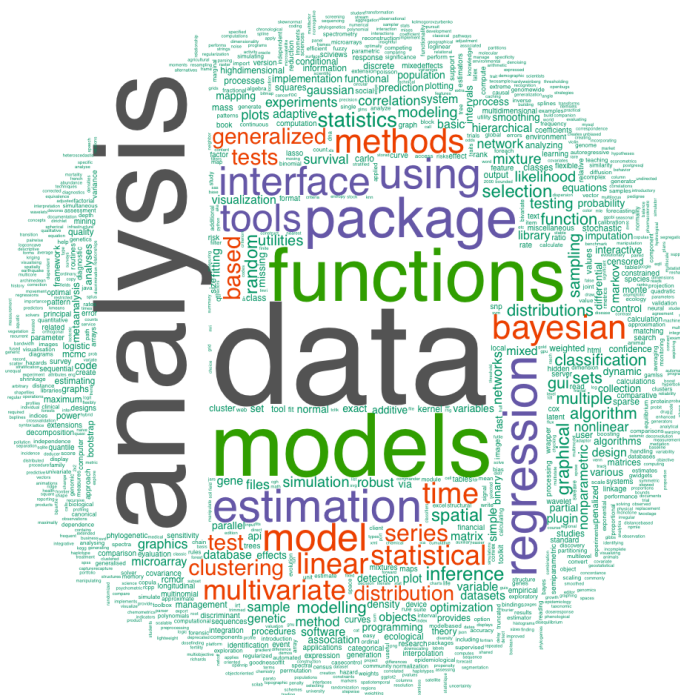
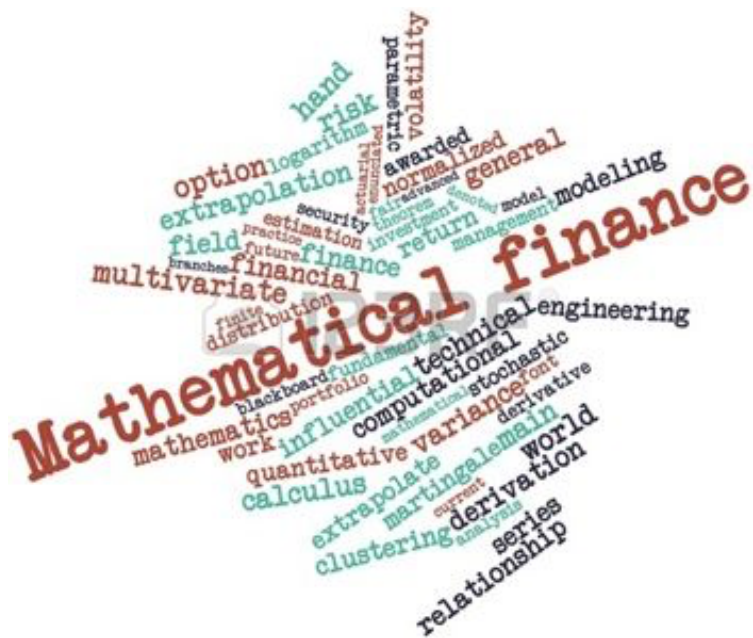
5

5

5

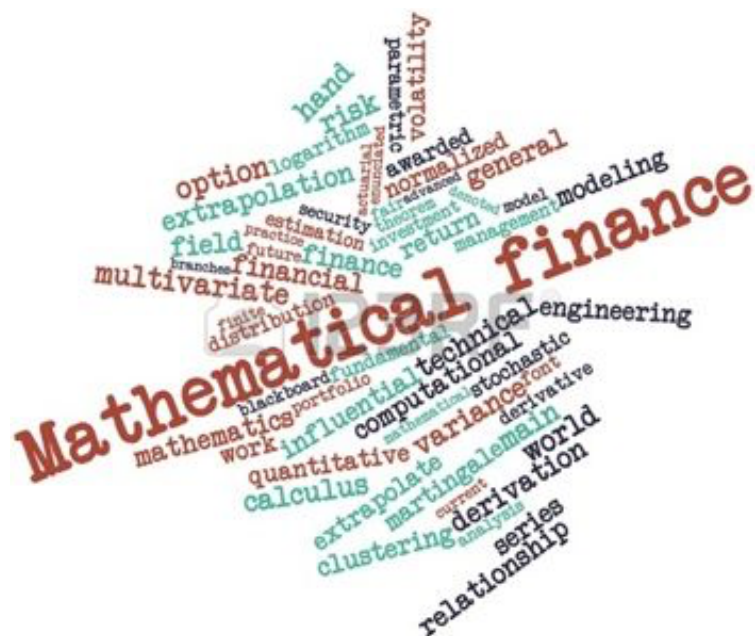
«Жёсткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»

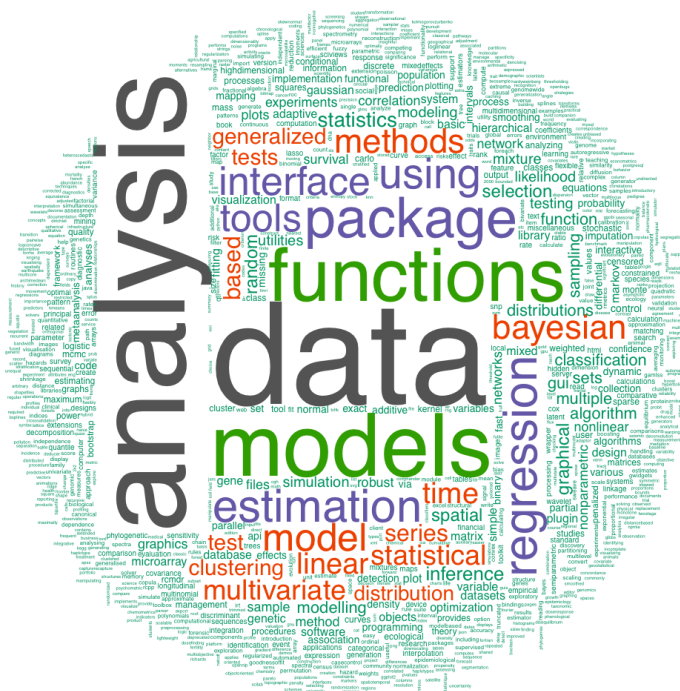


«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»



0.2



0.3



0.5

Типы задач кластеризации

- Форма кластеров, которые нужно выделять
- Плоская или древовидная структура
- Размер кластеров
- Конечная задача или вспомогательная
- Жесткая или мягкая кластеризация

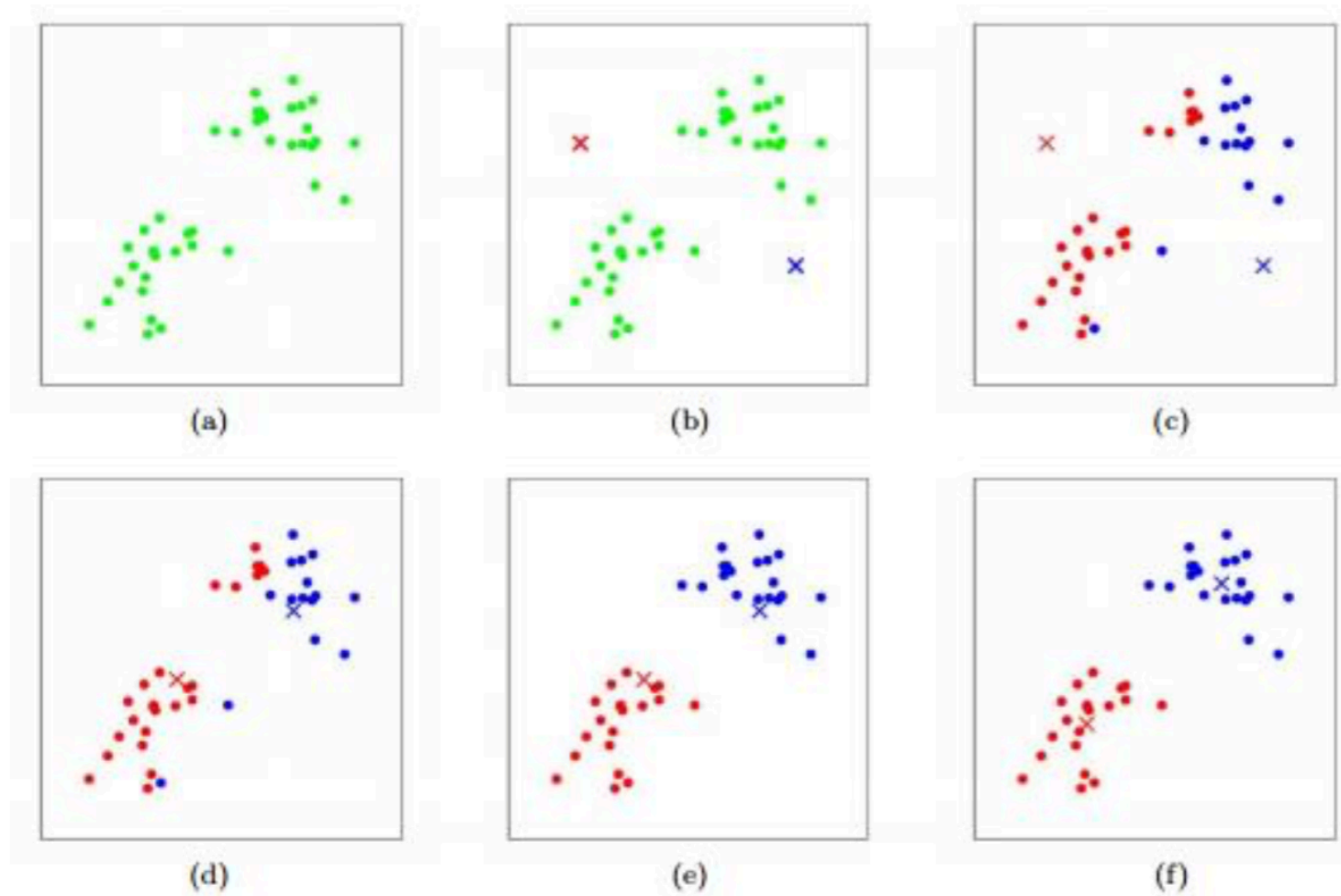
K-Means

K-Means

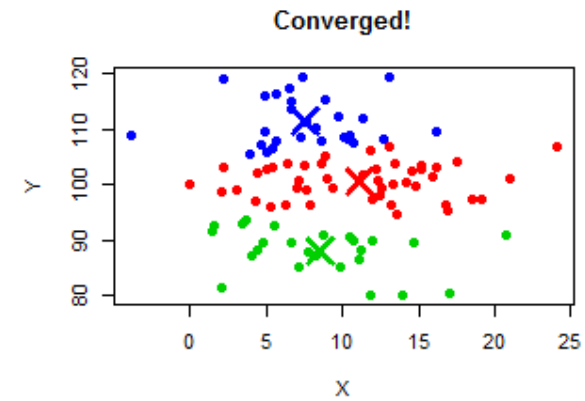
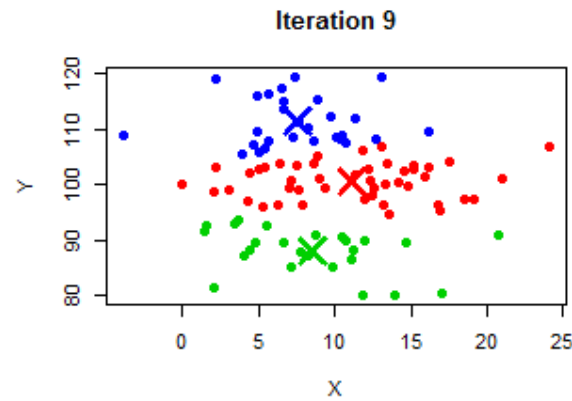
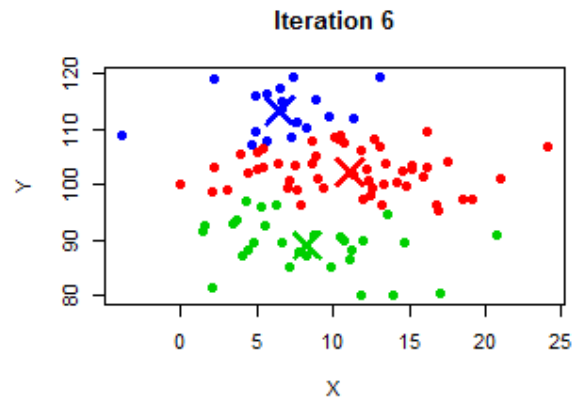
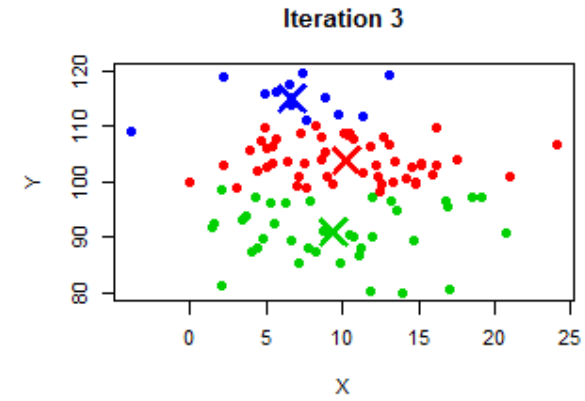
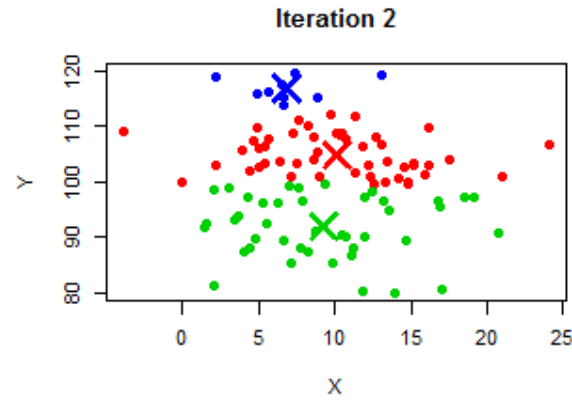
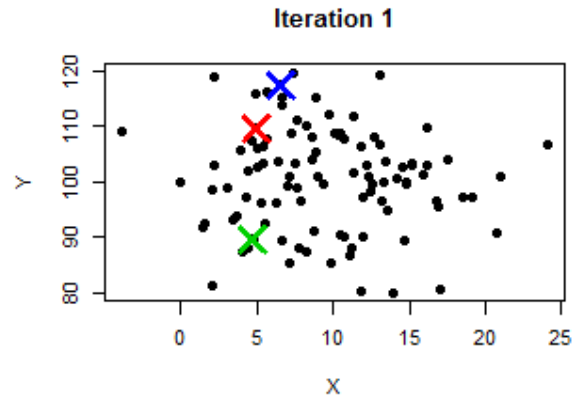
- Дано: выборка $x_1, \dots, x_\ell$
- Параметр: число кластеров K
- Начало: случайно выбрать K центров кластеров $c_1, \dots, c_K$
- Повторять по очереди до сходимости:
 - Шаг А: отнести каждый объект к ближайшему центру
$$y_i = \arg \min_{j=1, \dots, K} \rho(x_i, c_j)$$
 - Шаг Б: переместить центр каждого кластера в центр тяжести

$$c_j = \frac{\sum_{i=1}^{\ell} x_i [y_i = j]}{\sum_{i=1}^{\ell} [y_i = j]}$$

K-Means



K-Means



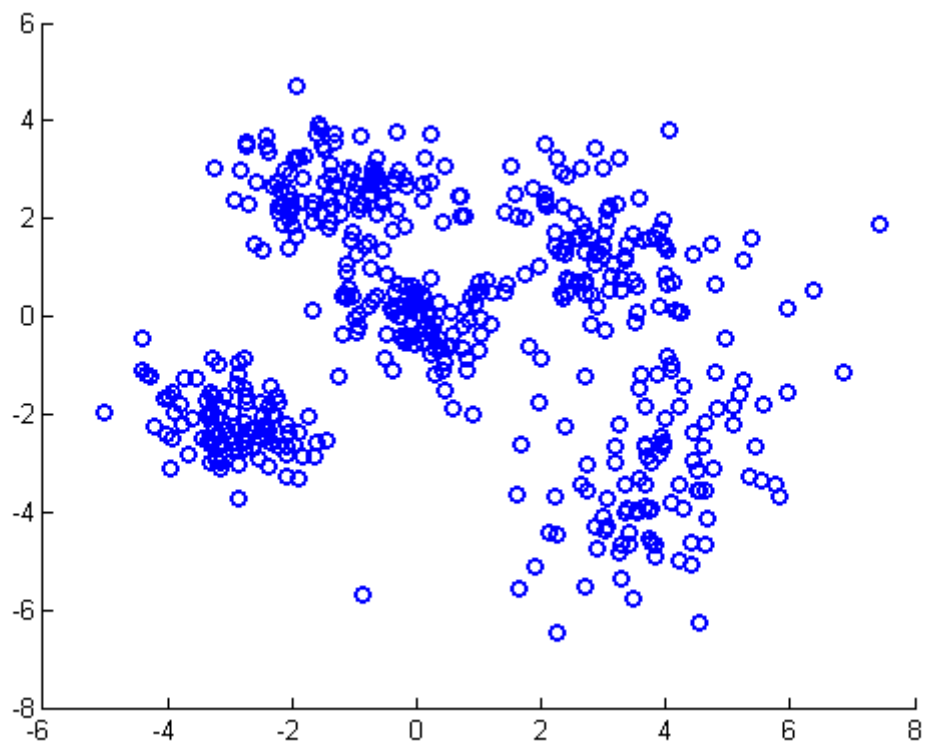
Выбор числа кластеров

- Качество кластеризации: внутрикластерное расстояние

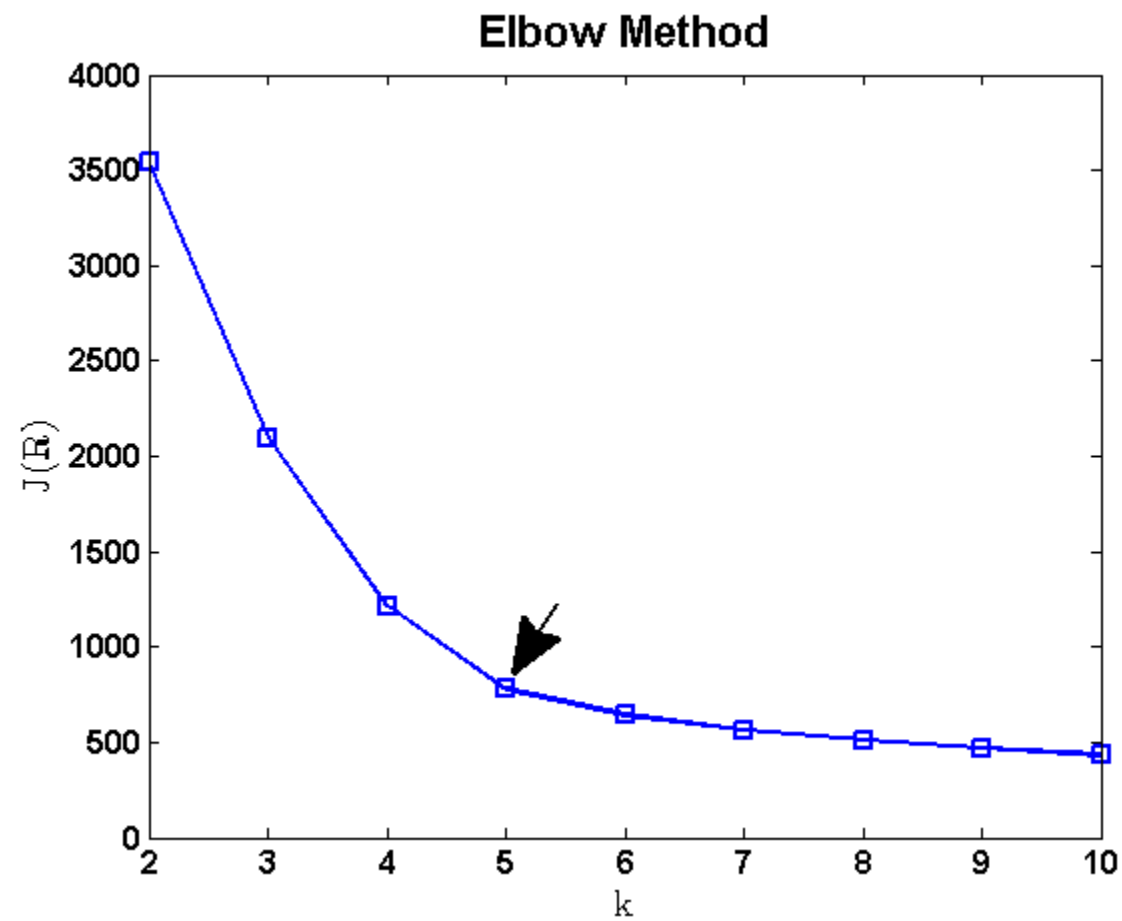
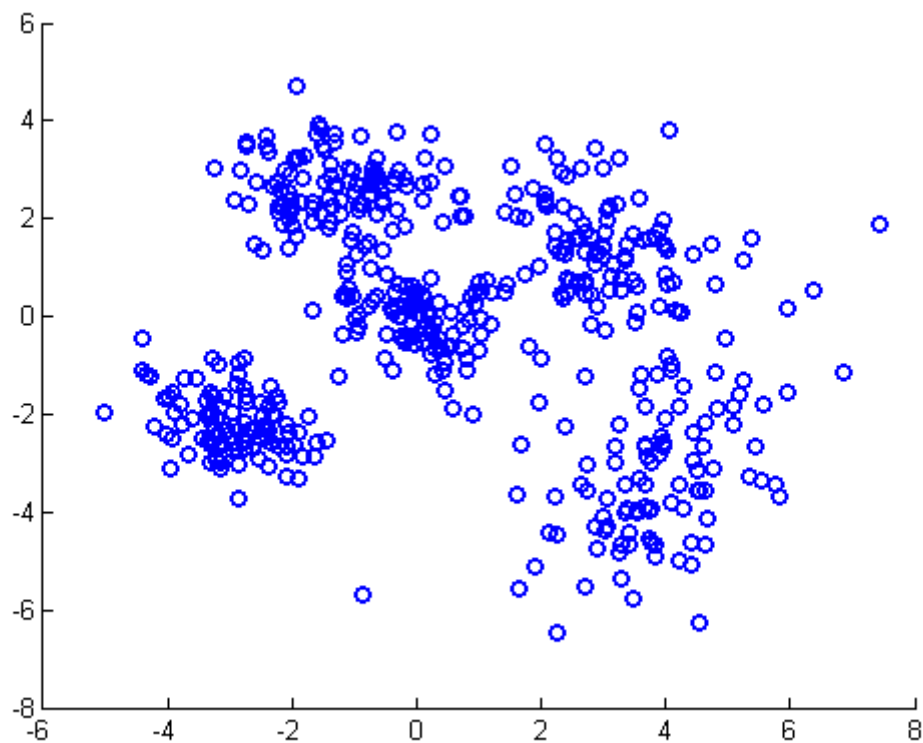
$$J(C) = \sum_{i=1}^{\ell} \rho(x_i, c_{y_i})$$

- Зависит от K
- Нужно подобрать такое K , после которого качество меняется не слишком сильно

Выбор числа кластеров



Выбор числа кластеров

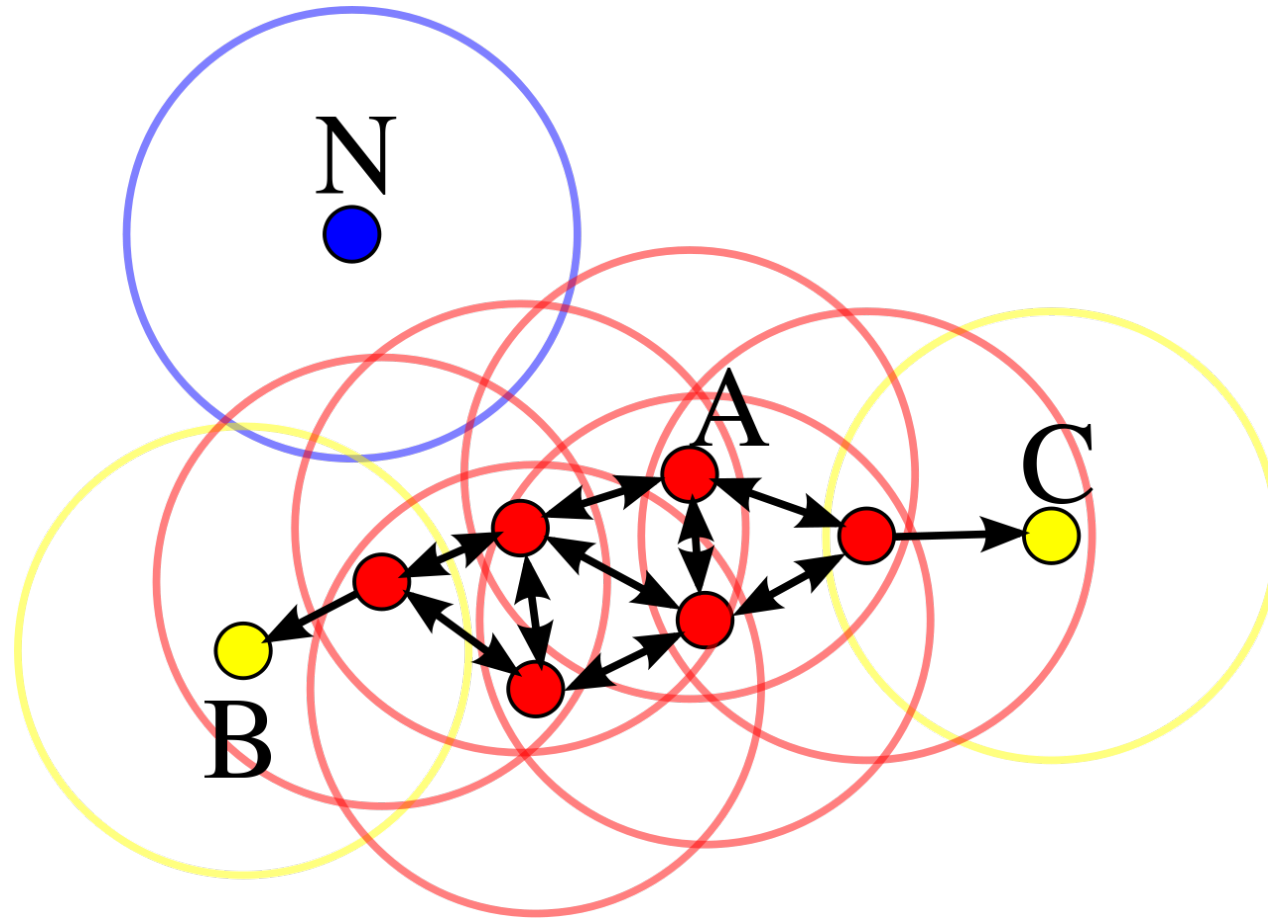


Особенности K-Means

- Может работать с большими объёмами данных
- Подходит для кластеров с простой геометрией
- Требуется выбора числа кластеров

Density-based clustering

Основные, граничные и шумовые точки



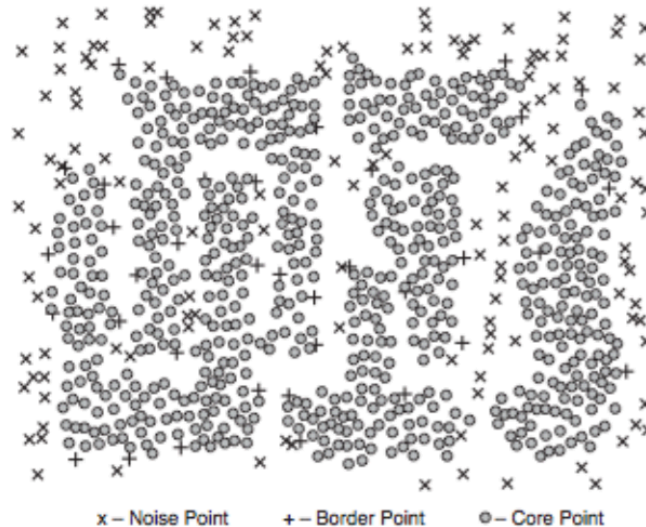
Параметры DBSCAN

- Размер окрестности (eps)
- Минимальное число объектов в окрестности

DBSCAN



(a) Clusters found by DBSCAN.



(b) Core, border, and noise points.

1: Пометить все точки, как основные, пограничные или шумовые.

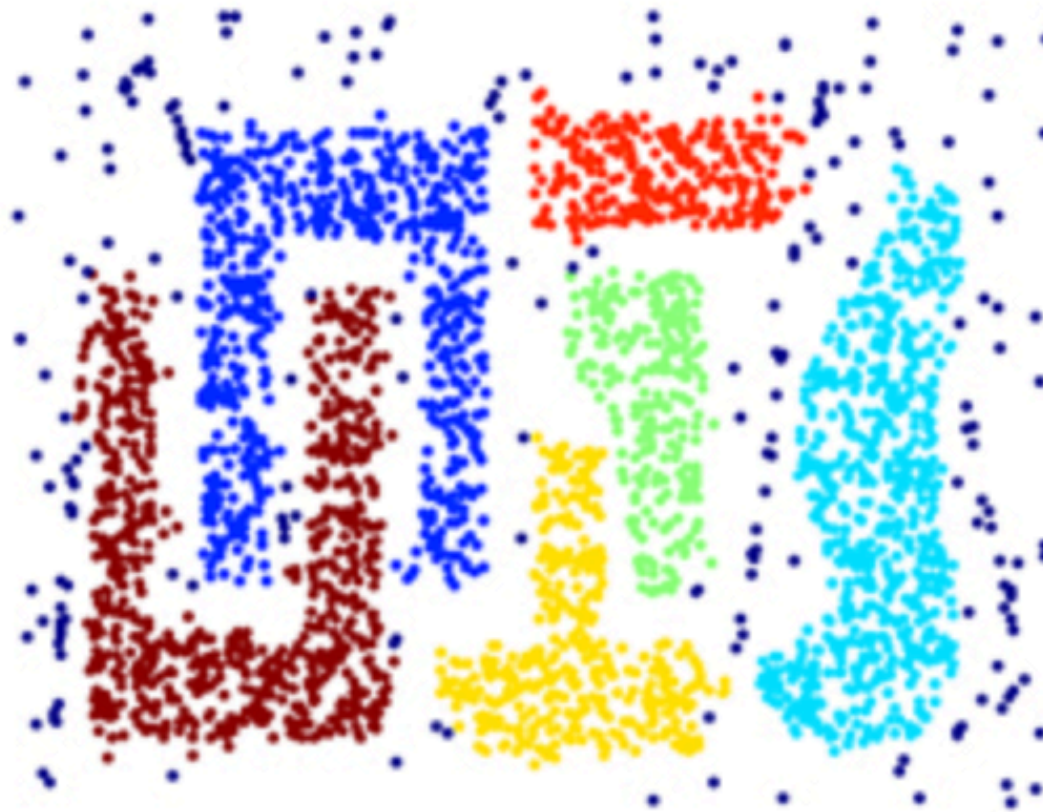
2: Отбросить точки шума.

3: Соединить все основные точки, находящиеся на расстоянии Eps радиуса одна от другой.

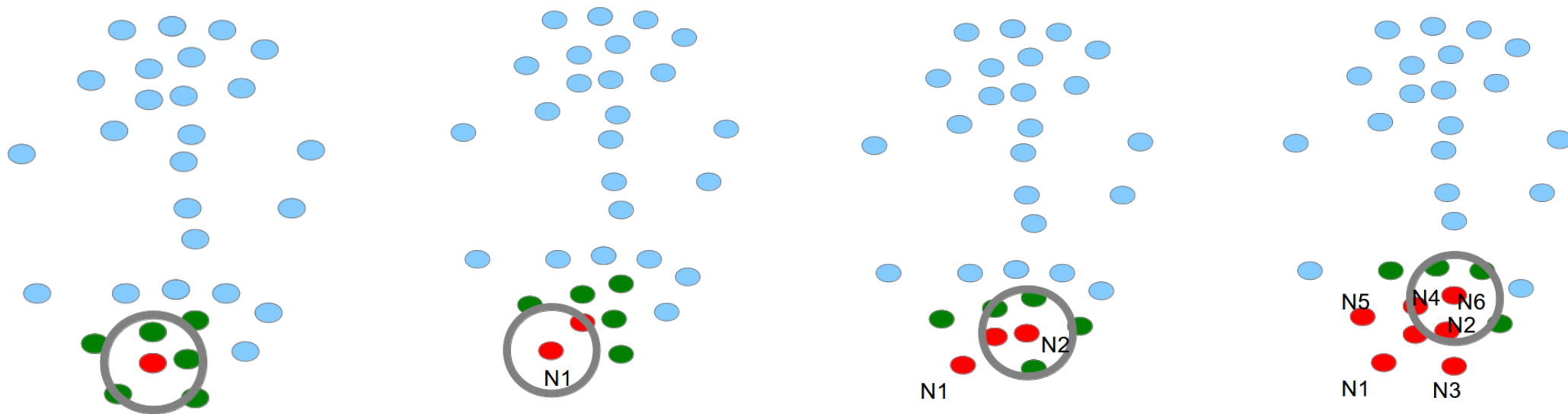
4: Объединить каждую группу соединенных основных точек в отдельный кластер.

5: Назначить каждую пограничную точку одному из кластеров, ассоциированных с ней основных точек.

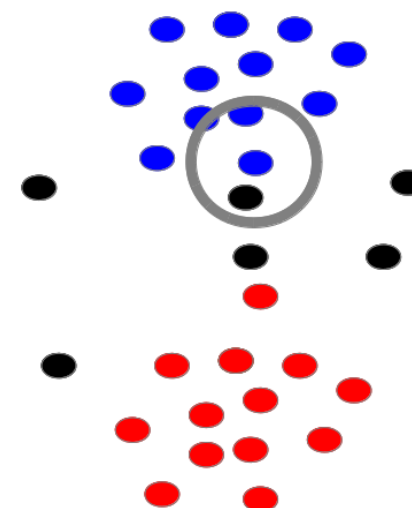
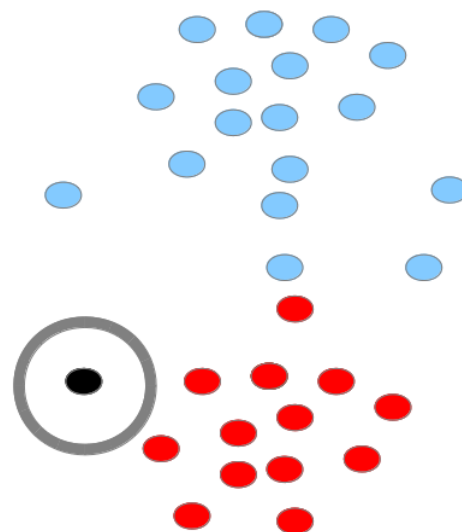
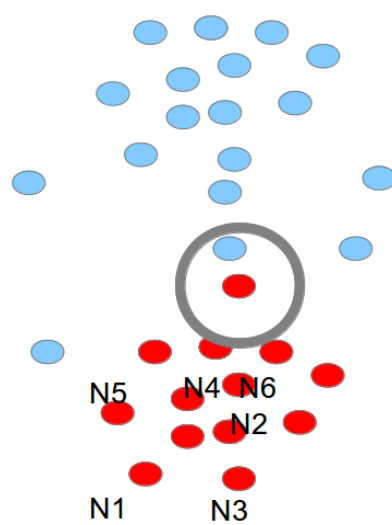
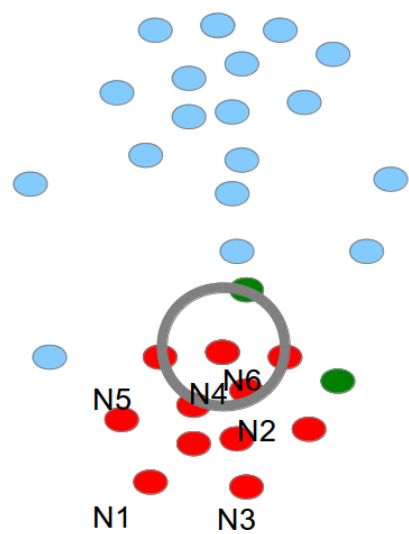
DBSCAN: результаты работы



Пример



Пример



Особенности DBSCAN

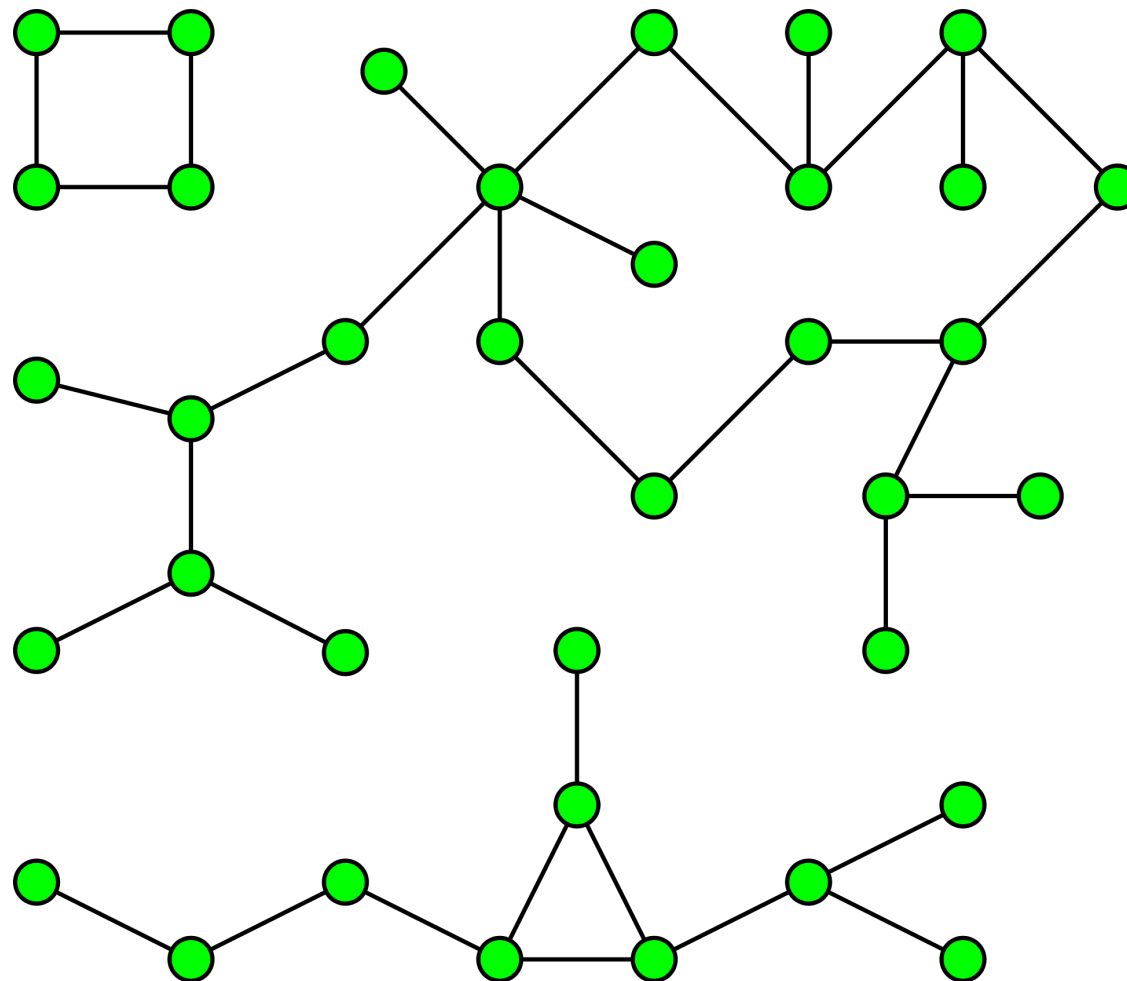
- Находит кластеры произвольной формы
- Может работать с большими объёмами данных
- Нужно подбирать размер окрестности (eps) и минимальное число объектов в окрестности

Графовые методы

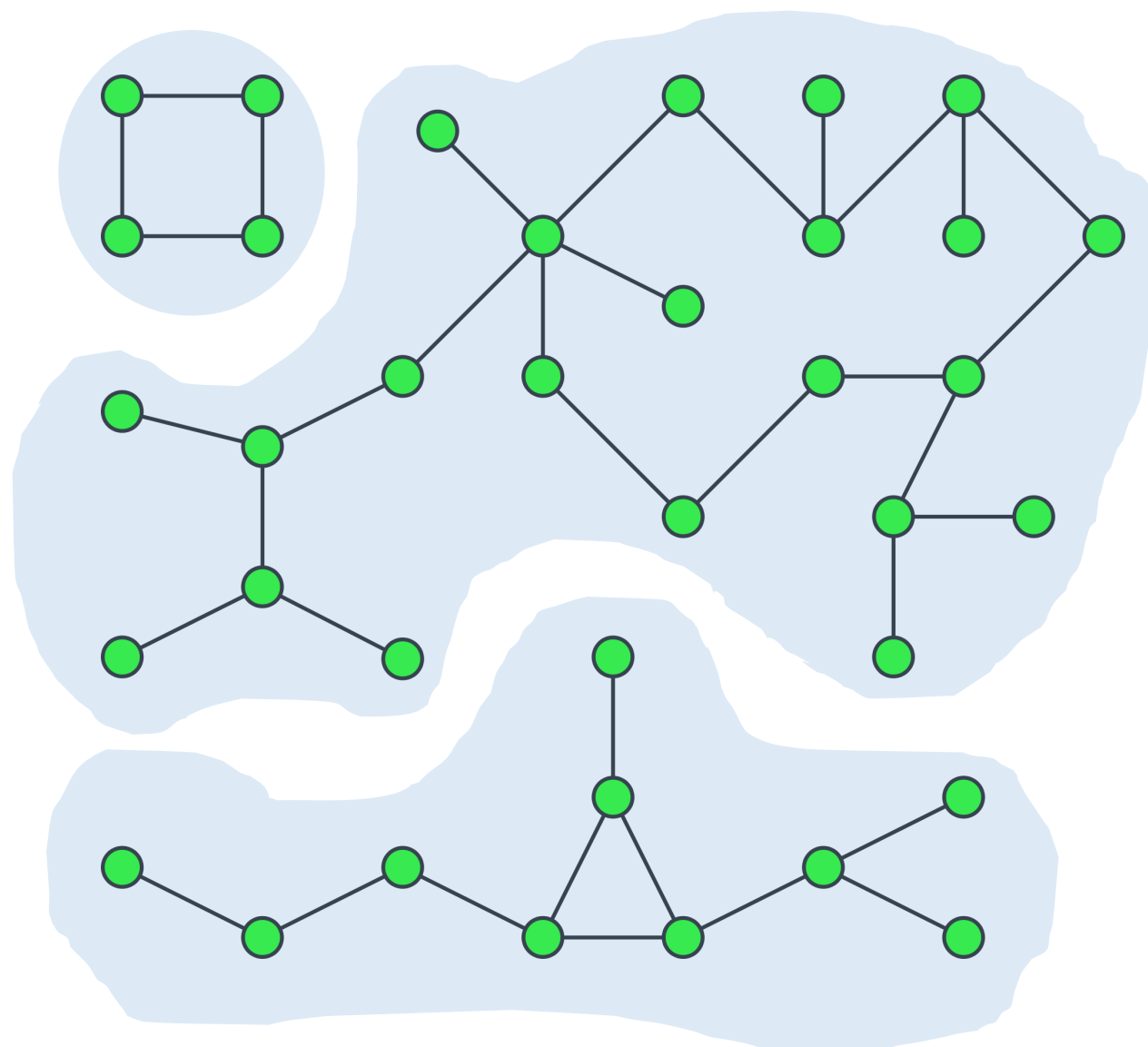
Кластеризация по компонентам связности

- Граф: вершины соответствуют объектам
- Соединяем ребром объекты, расстояние между которыми меньше R
- Выделяем компоненты связности

Выделение связанных компонент



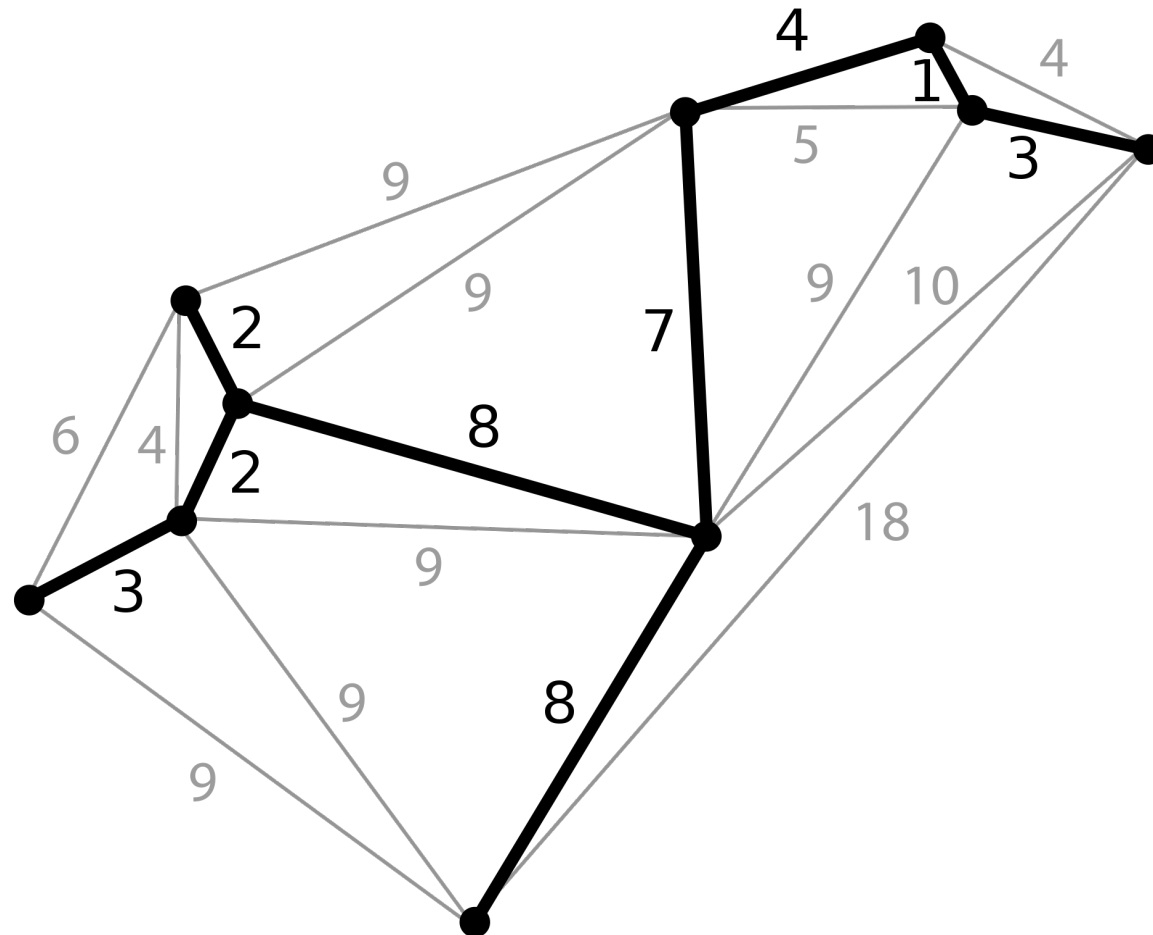
Выделение связанных компонент



Кластеризация по компонентам связности

- Быстрая и простая
- Параметр — минимальное расстояние R
- Проблема: непонятно, как выбрать R , если нужно получить K кластеров

Минимальное остовное дерево



Минимальное остовное дерево

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

Резюме

- Кластеризация — задача без строгой постановки и без строгих критериев качества
- Много разновидностей в подходах
- Методы: K-Means, DBSCAN, графовые и т.д.