



# **О задаче кластеризации**

Евгений Борисов

# метрические методы : кластеризация

метрический подход - использование расстояний между объектами

метрика - функция расстояния

$$\rho: X \times X \rightarrow [0, \infty)$$

аксиома тождества :  $\rho(x, y) = 0 \Leftrightarrow x = y$

симметрия:  $\rho(x, y) = \rho(y, x)$

неравенство треугольника:  $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$

# метрические методы : кластеризация

метрика - функция расстояния

Евклидова метрика:  $\rho(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$

метрика Минковского:  $\rho(x, y) = \sqrt[n]{\sum_i w_i |x_i - y_i|^n}$

метрика Чебышева:  $\rho(x, y) = \max_i |x_i - y_i|$

косинусная метрика:  $\rho(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$

# метрические методы : кластеризация

**о задаче:** обучение «без учителя» (unsupervised learning)

дано:

$X$  - объекты

$\rho: X \times X \rightarrow [0, \infty)$  - функция расстояния (метрика)

найти:

$Y$  - кластеры (метки)

$a: X \rightarrow Y$  - кластеризатор

- кластер состоит из близких объектов

- объекты разных кластеров существенно разные

# метрические методы : кластеризация

## **о некорректности (размытости) задачи кластеризации**

- недостаточно точная постановка задачи
- много разных критериев качества
- число кластеров обычно заранее не известно
- результат сильно зависит от метрики
- нормировка данных может существенно изменять результат

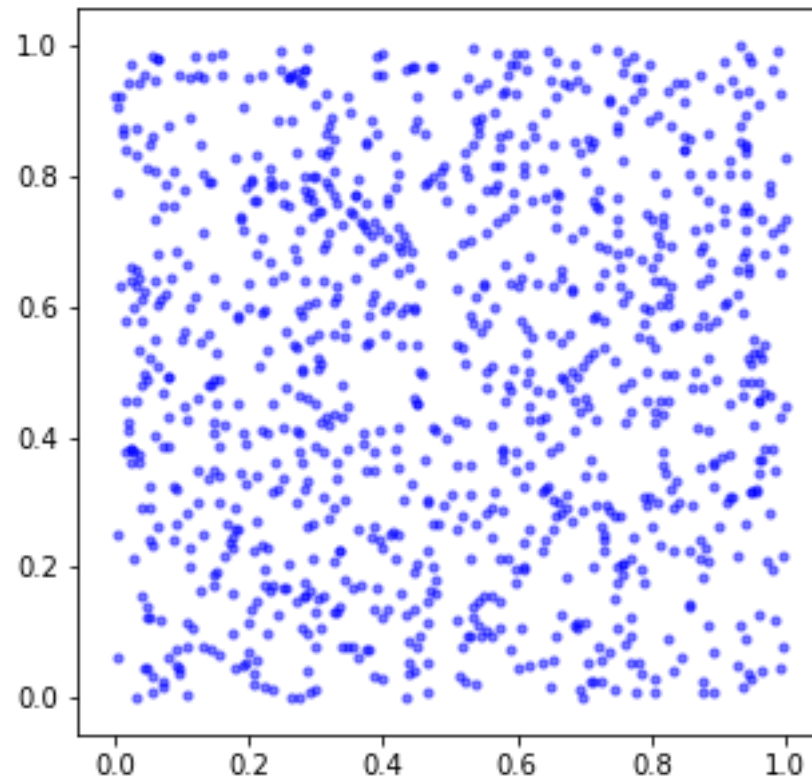
# метрические методы : кластеризация

## цели кластеризации

- предварительная обработка данных для упрощения основной задачи
- сжатие данных (оставляем один или несколько объектов от кластера)
- выделить нетипичные объекты
- построение иерархии объектов

# метрические методы : кластеризация

**тип кластера:** кластеры могут отсутствовать совсем

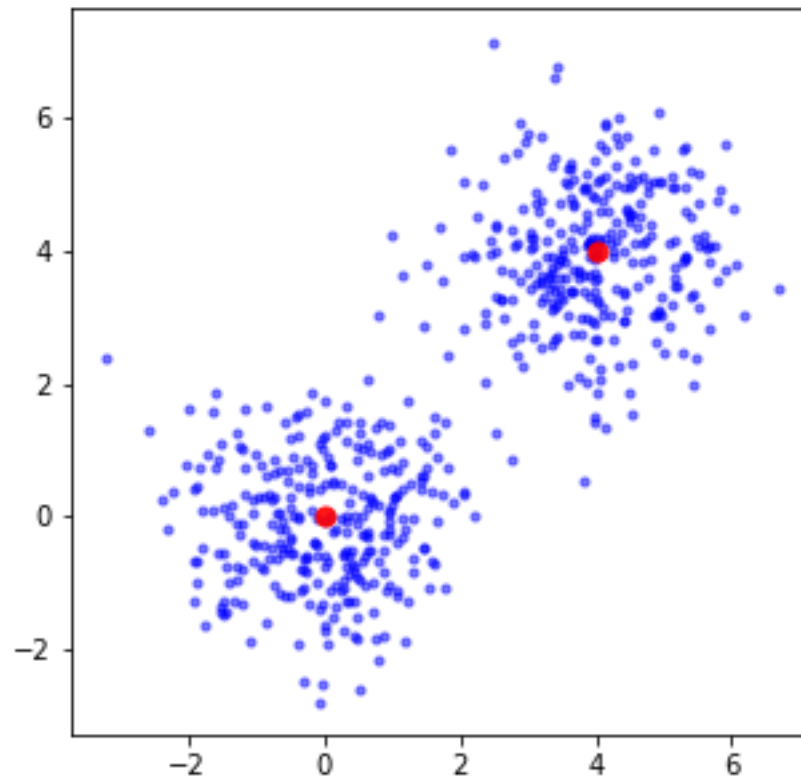


# метрические методы : кластеризация

типы кластера: кластеры с центром

другие типы кластеров:

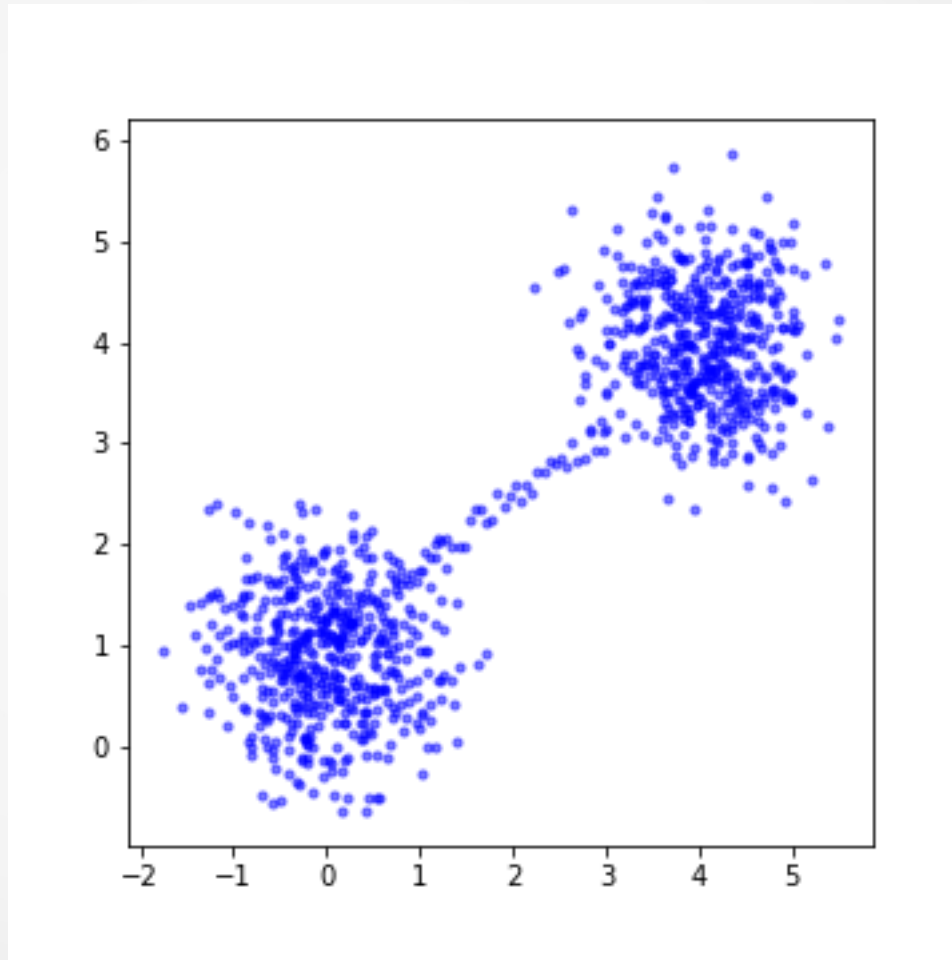
- могут отсутствовать





# метрические методы : кластеризация

тип кластера: кластеры с перемычками



**другие типы кластеров:**

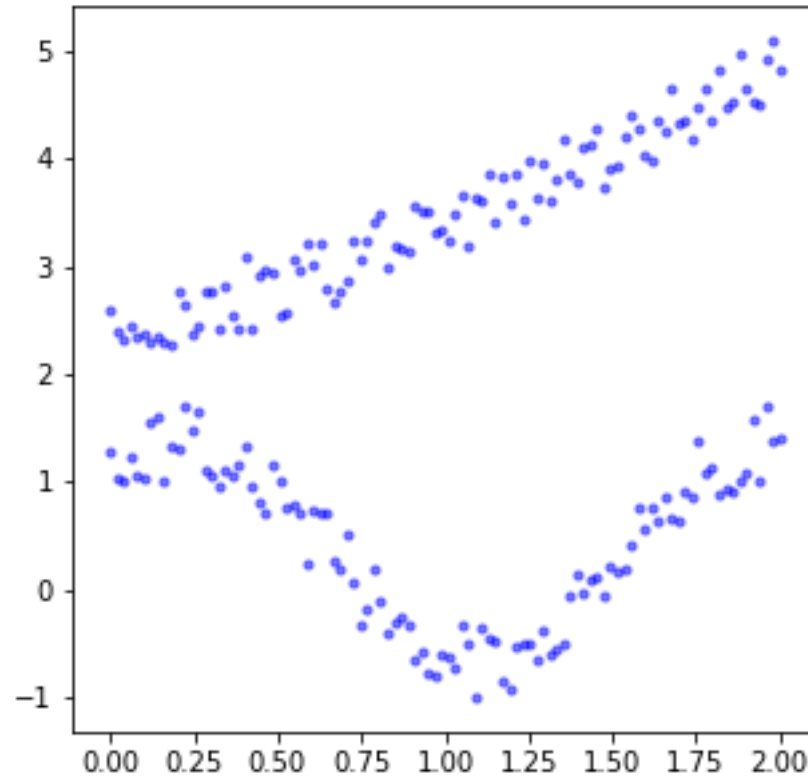
- могут отсутствовать
- кластеры с центром

# метрические методы : кластеризация

тип кластера: кластеры ленточные

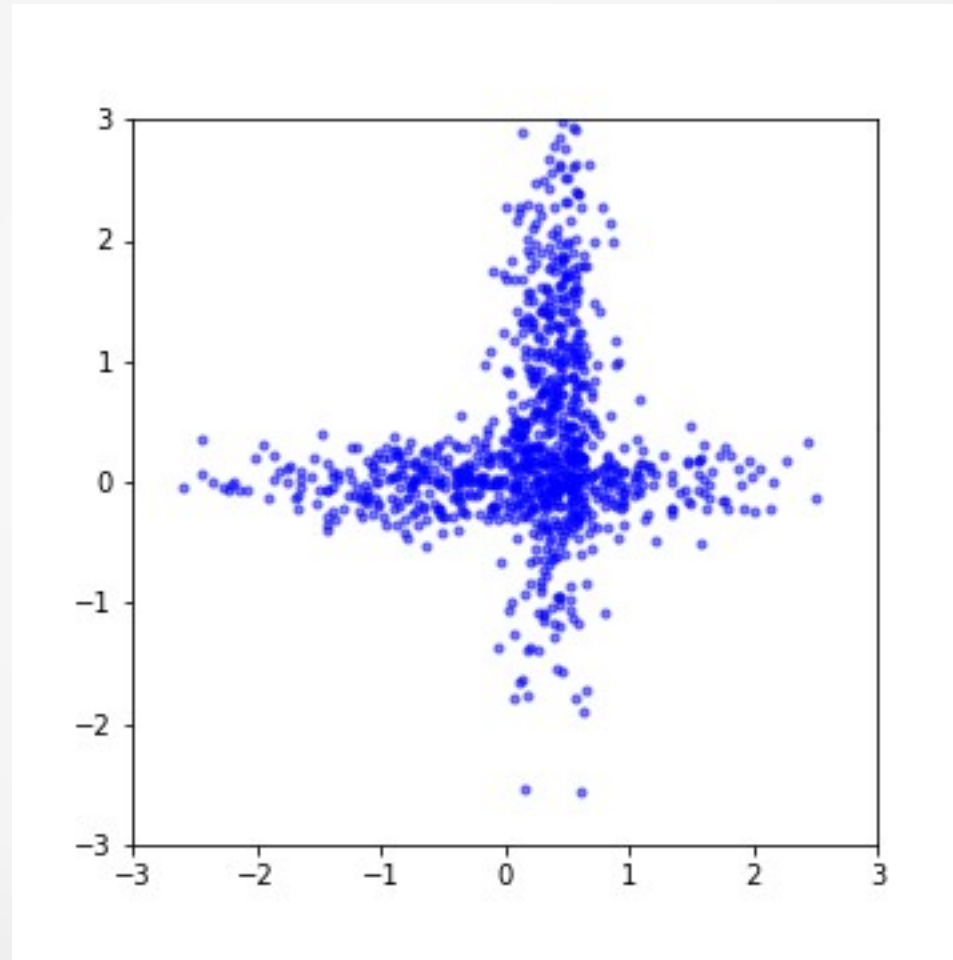
**другие типы кластеров:**

- могут отсутствовать
- кластеры с центром
- кластеры с перемычками



# метрические методы : кластеризация

тип кластера: кластеры с наложением



**другие типы кластеров:**

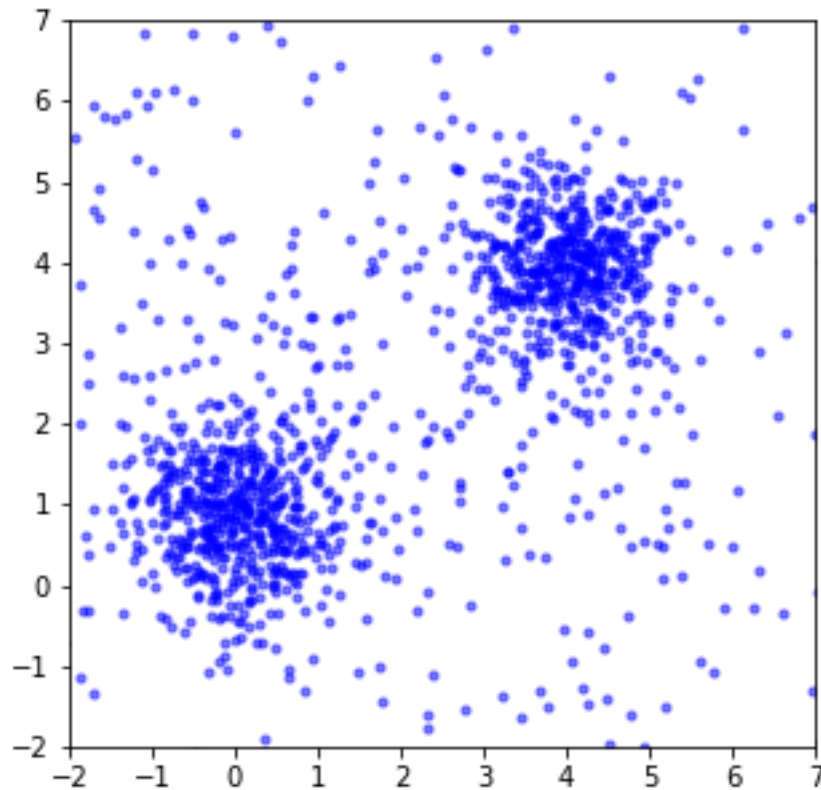
- могут отсутствовать
- кластеры с центром
- кластеры с перемычками
- кластеры ленточные

# метрические методы : кластеризация

тип кластера: кластеры с шумом

**другие типы кластеров:**

- могут отсутствовать
- кластеры с центром
- кластеры с перемычками
- кластеры ленточные
- кластеры с наложением

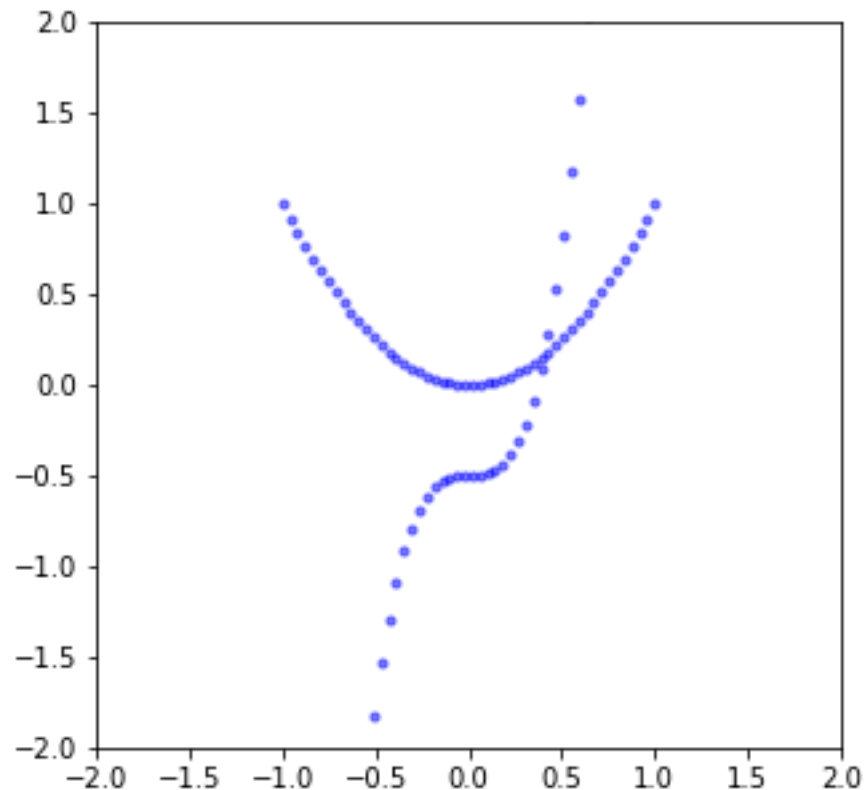


# метрические методы : кластеризация

тип кластера: кластеры по типу регулярности

**другие типы кластеров:**

- могут отсутствовать
- кластеры с центром
- кластеры с перемычками
- кластеры ленточные
- кластеры с наложением
- кластеры с шумом



# метрические методы : кластеризация

**оценки кластеризации**  $a: X \rightarrow Y$

$$ri = \frac{\sum_{i < j} [a_i = a_j] \rho(x_i, x_j)}{\sum_{i < j} [a_i = a_j]} \rightarrow \min$$

среднее внутрикластерное расстояние

$$ro = \frac{\sum_{i < j} [a_i \neq a_j] \rho(x_i, x_j)}{\sum_{i < j} [a_i = a_j]} \rightarrow \max$$

среднее межкластерное расстояние

отношение внутрикластерного и межкластерного расстояний

$$\frac{ri}{ro} \rightarrow \min$$

# метрические методы : кластеризация

## **метод к-средних (k-means)**

количество кластеров как параметр,

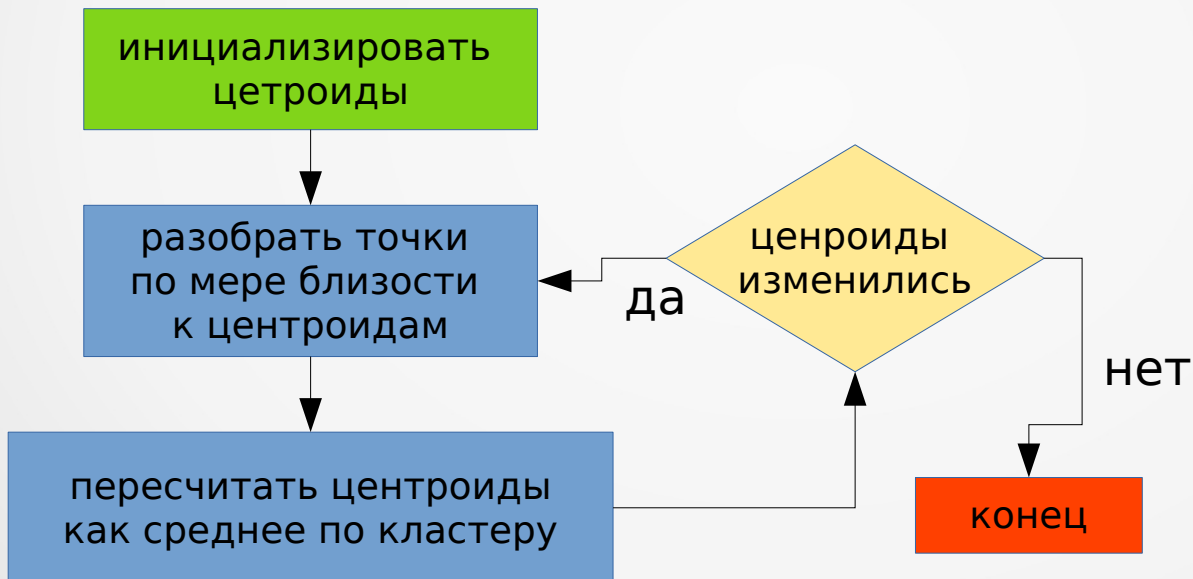
цель - найти точки-центроиды

# метрические методы : кластеризация

## метод к-средних (k-means)

количество кластеров как параметр,

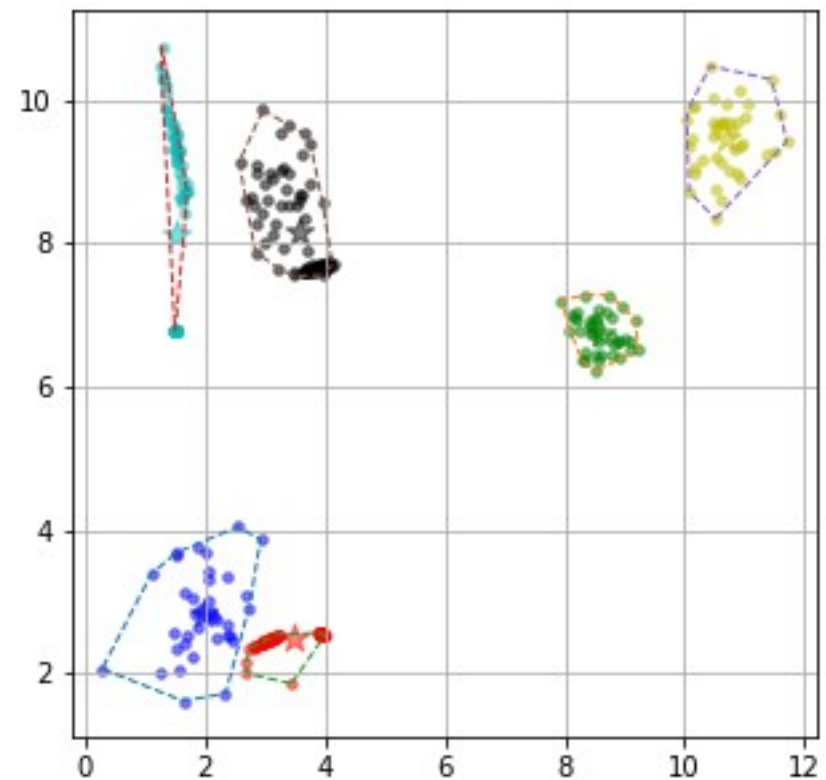
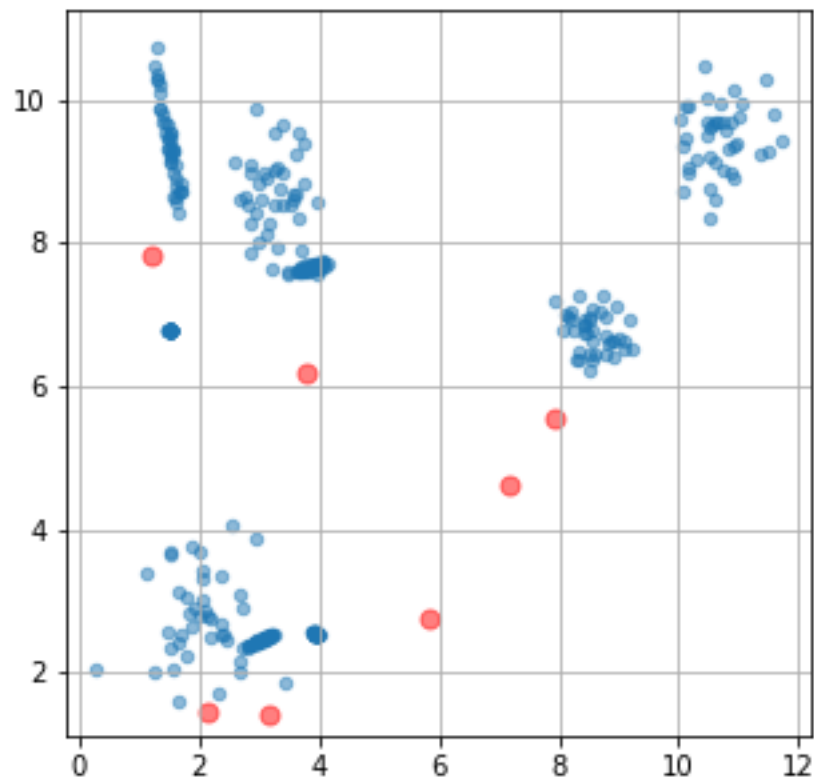
цель - найти точки-центроиды





# метрические методы : кластеризация

**k-means:** начальное состояние и результат



# метрические методы : кластеризация

## метод MeanShift (ФорЭл - ФОРмальные Элементы)

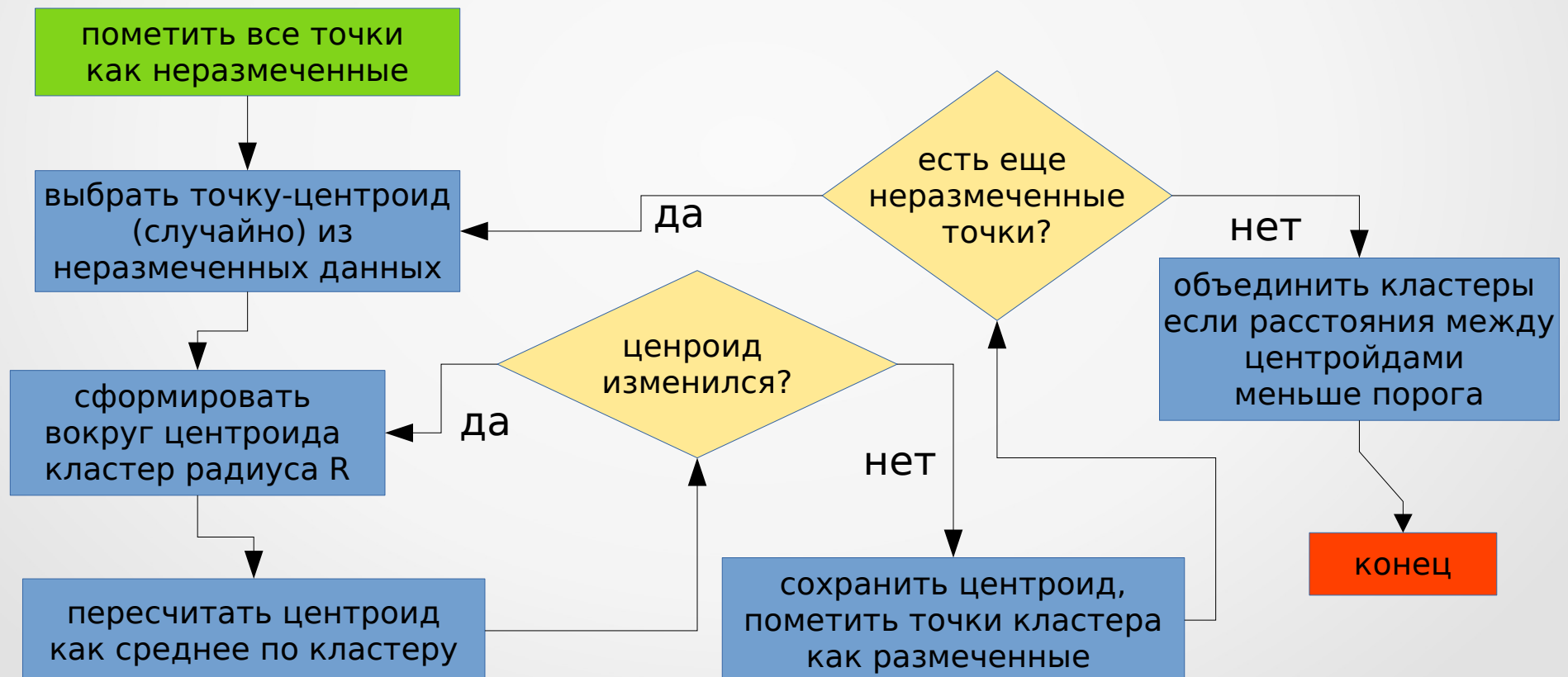
фиксируем радиус  $R$  кластеров,

цель - найти точки-центроиды

# метрические методы : кластеризация

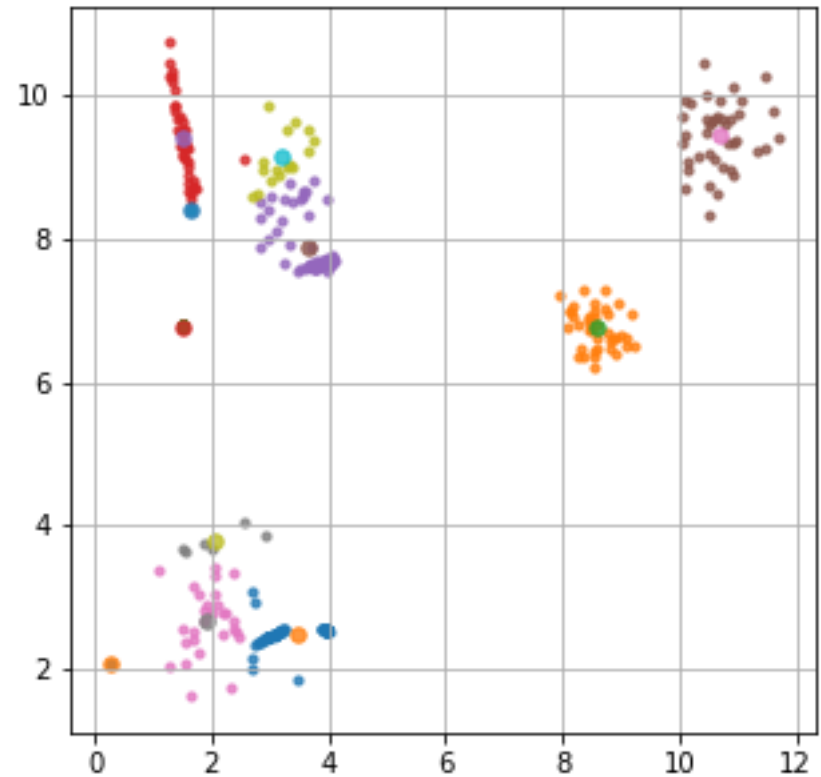
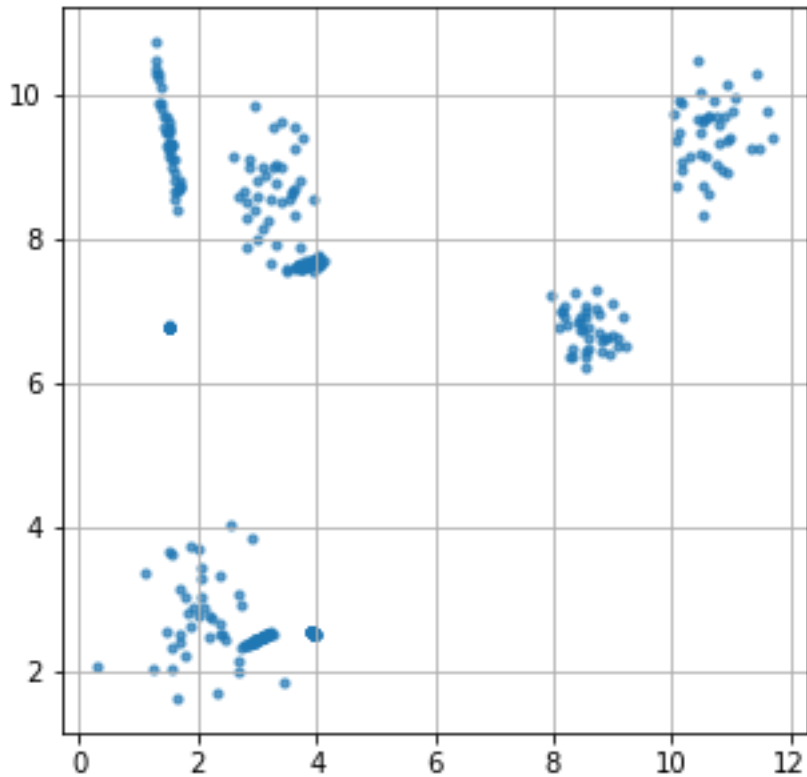
## метод MeanShift (ФорЭл - ФОРмальные Элементы)

фиксируем радиус  $R$  кластеров, цель - найти точки-центроиды



# метрические методы : кластеризация

**MeanShift (ФорЭл)** начальное состояние и результат



# метрические методы : кластеризация

## **метод КНП (Кратчайший Незамкнутый Путь)**

параметр - количество кластеров  $k$

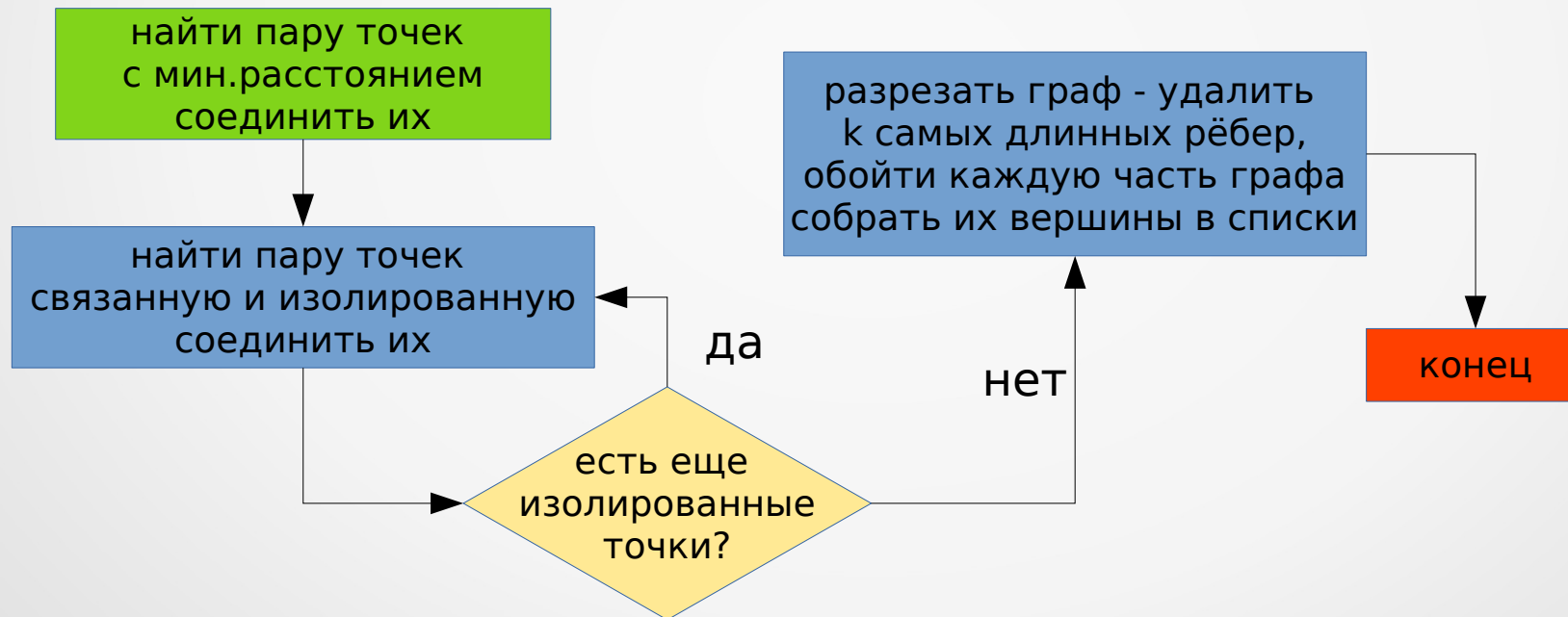
цель - построить ациклический граф на точках

# метрические методы : кластеризация

## метод КНП (Кратчайший Незамкнутый Путь)

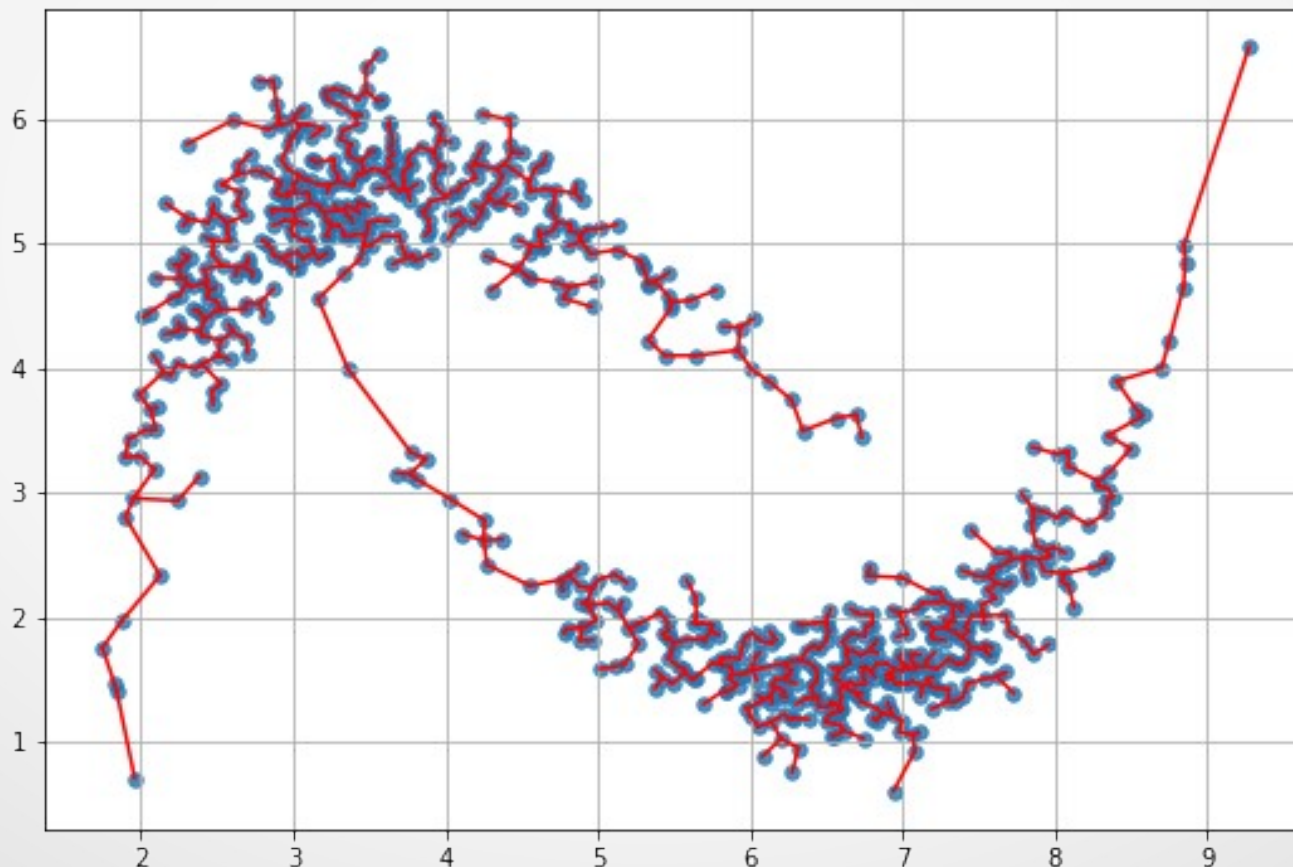
параметр - количество кластеров  $k$

цель - построить ациклический граф на точках



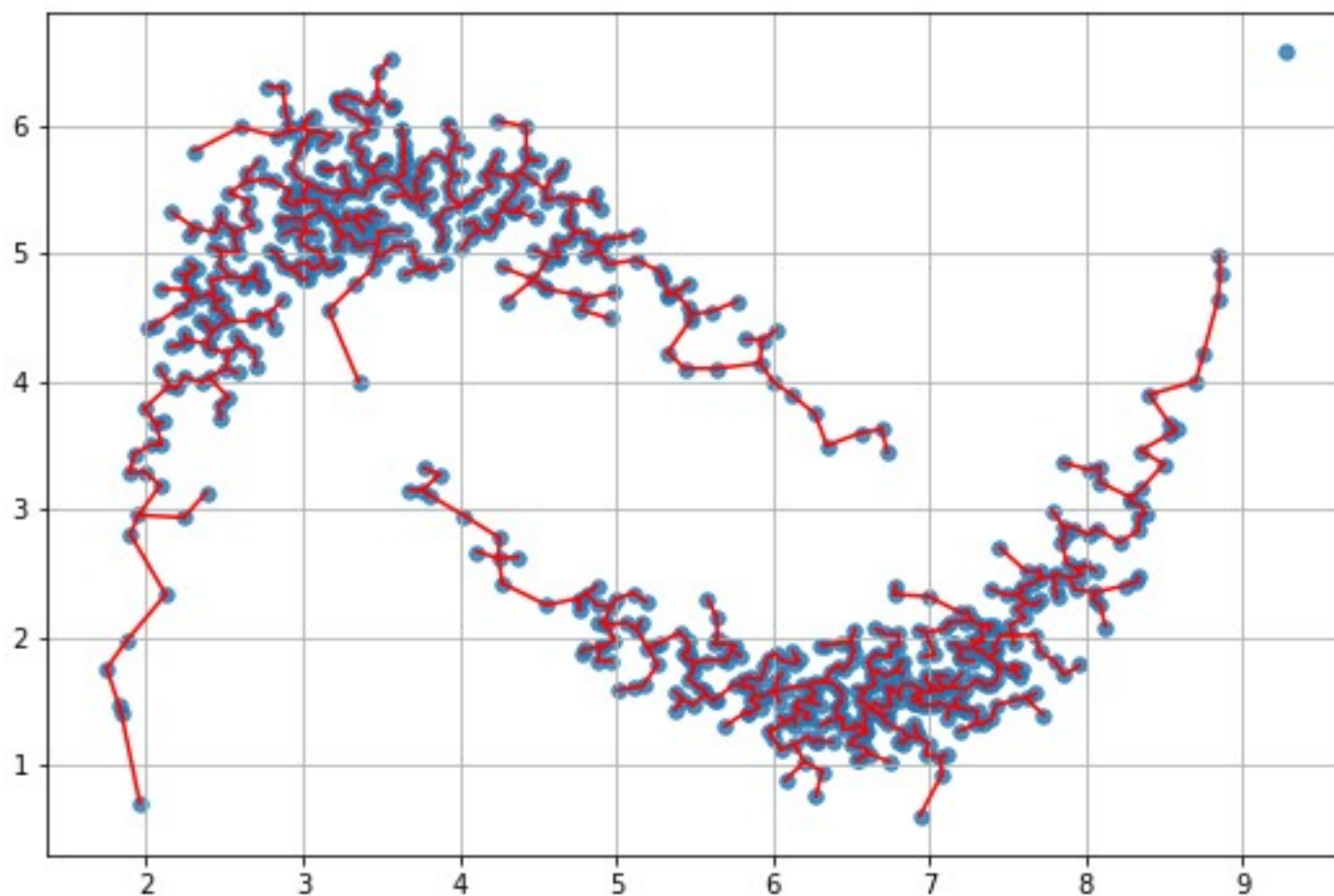
# метрические методы : кластеризация

**КНП:** полный граф



# метрические методы : кластеризация

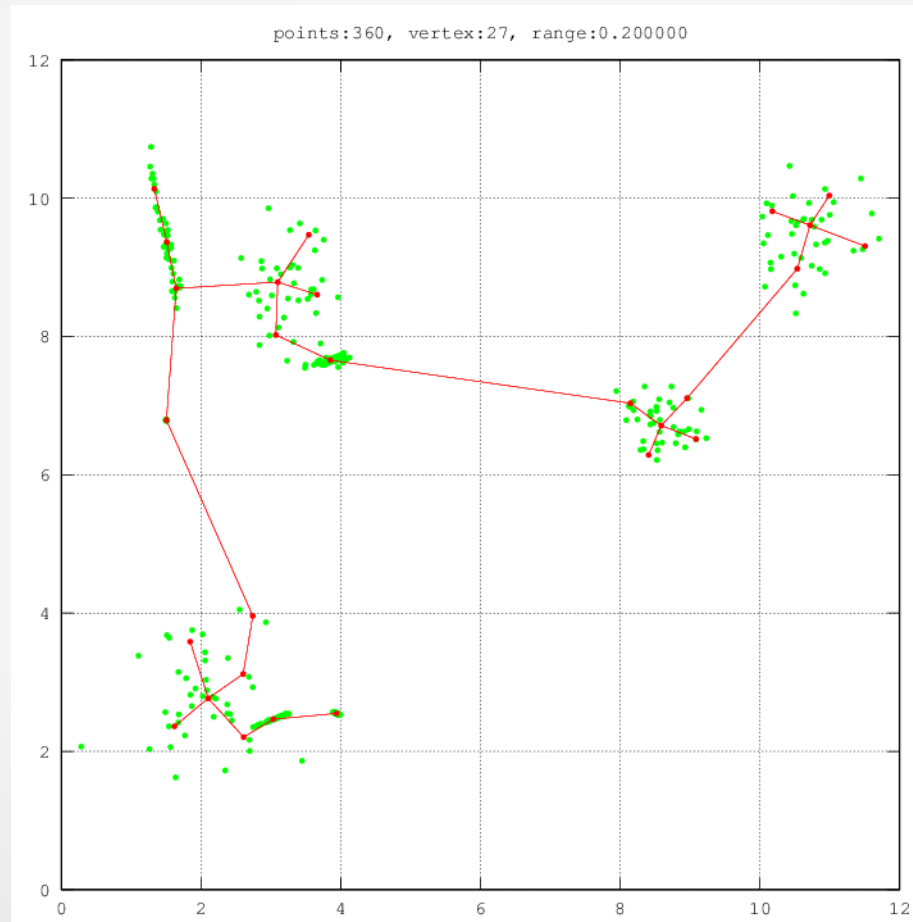
**КНП:** разрезанный граф





# метрические методы : кластеризация

ФорЭл + КНП



# метрические методы : кластеризация

## метод DBSCAN

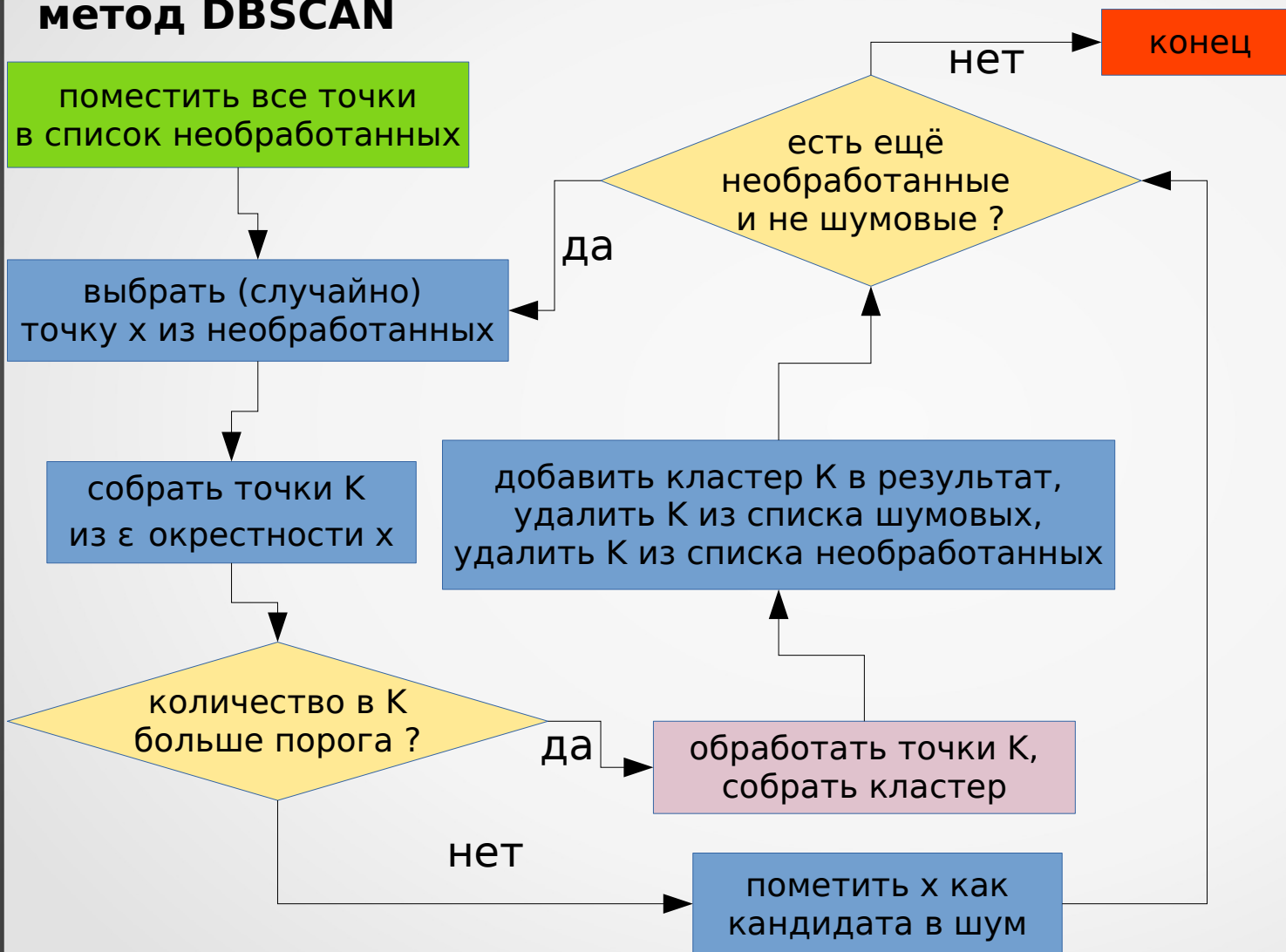
фиксируем размер окрестности точки

минимальное количество объектов в кластере

точки делим на корневые, граничные и шум

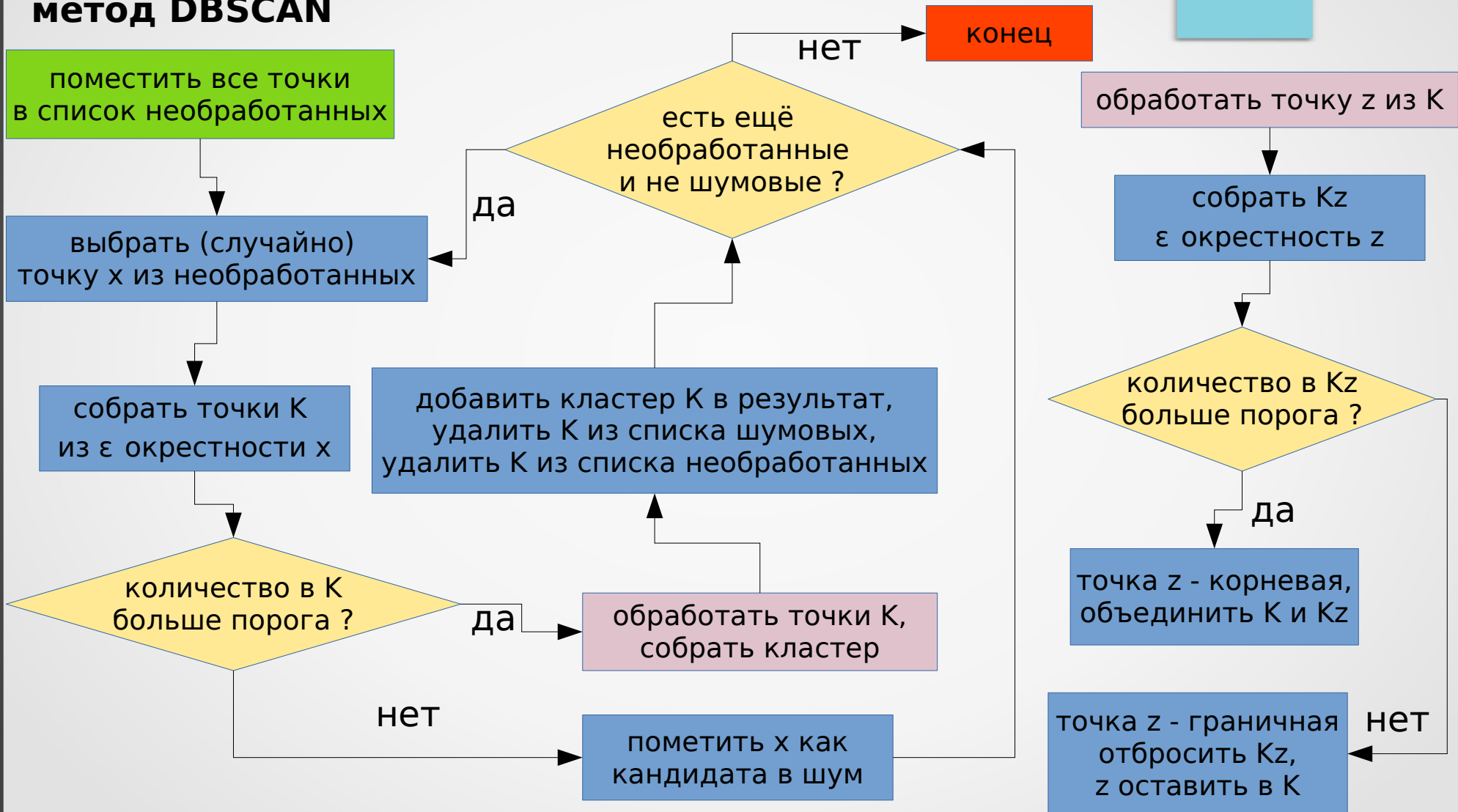
# кластеризация

## метод DBSCAN



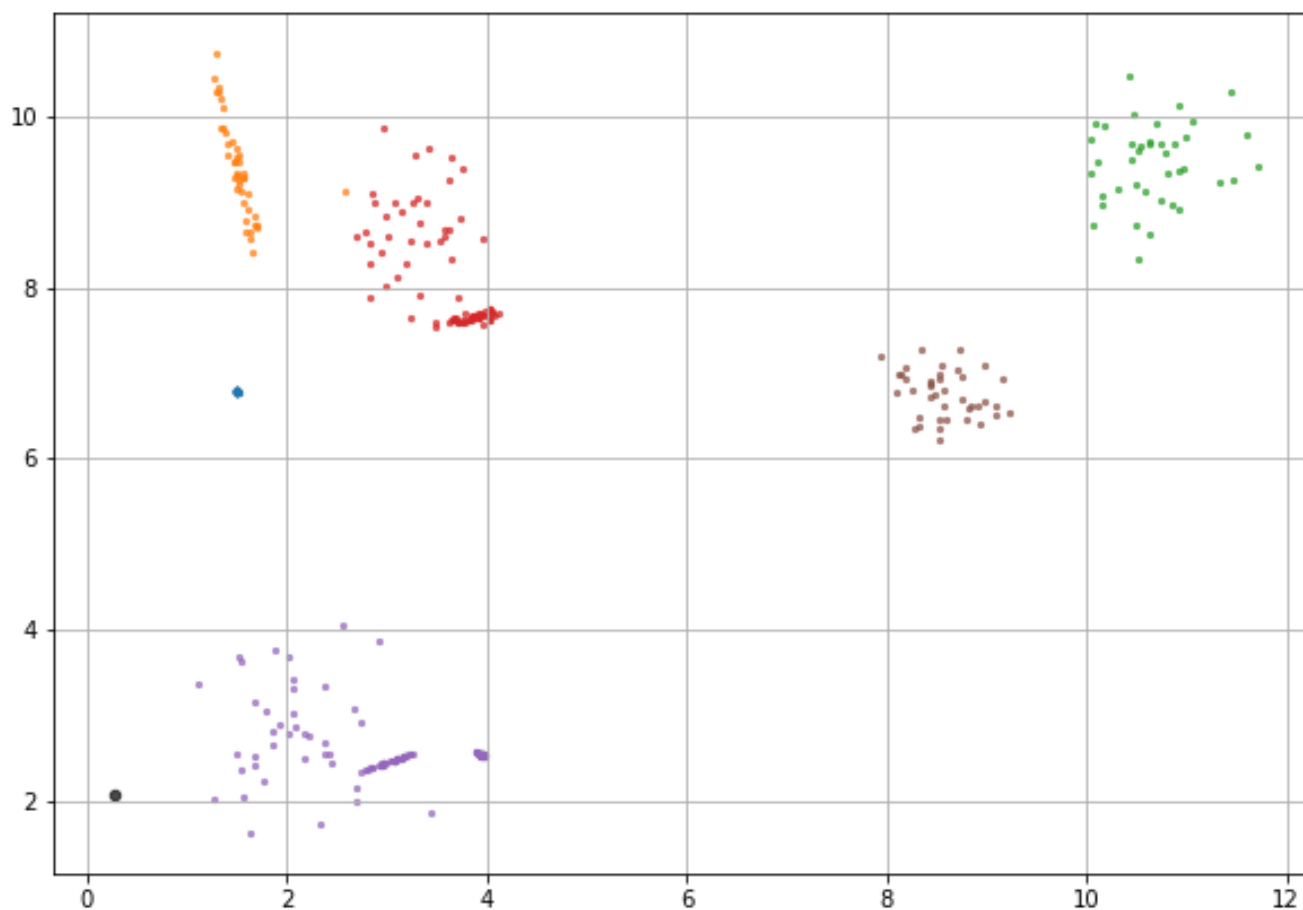
# метрические методы : кластеризация

## метод DBSCAN



# метрические методы : кластеризация

метод **DBSCAN** - результат



# метрические методы : кластеризация

## **иерархическая кластеризация**

идея - последовательное объединение близких групп объектов

вход: данные

выход: история объединения групп в виде дерева (дендрограмма)

нужен метод оценки расстояний между множествами точек

# метрические методы : кластеризация

## **иерархическая кластеризация**

идея - последовательное объединение близких групп объектов

вход: данные

выход: история объединения групп в виде дерева (дендрограмма)

нужен метод оценки расстояний между множествами точек

- расстояние между центрами множеств
- наибольшее расстояние среди всех точек множеств
- наименьшее расстояние между всеми точками множеств
- среднее расстояние между всеми точками множеств

# метрические методы : кластеризация

## иерархическая кластеризация

последовательное объединение близких групп объектов

метод оценки расстояний между множествами точек

расстояние Уорда

$$\rho_{ward}(W, S) = \frac{|W| \cdot |S|}{|W| + |S|} \cdot \rho \left( \frac{\sum_{w \in W} w}{|W|}, \frac{\sum_{s \in S} s}{|S|} \right)$$



# метрические методы : кластеризация

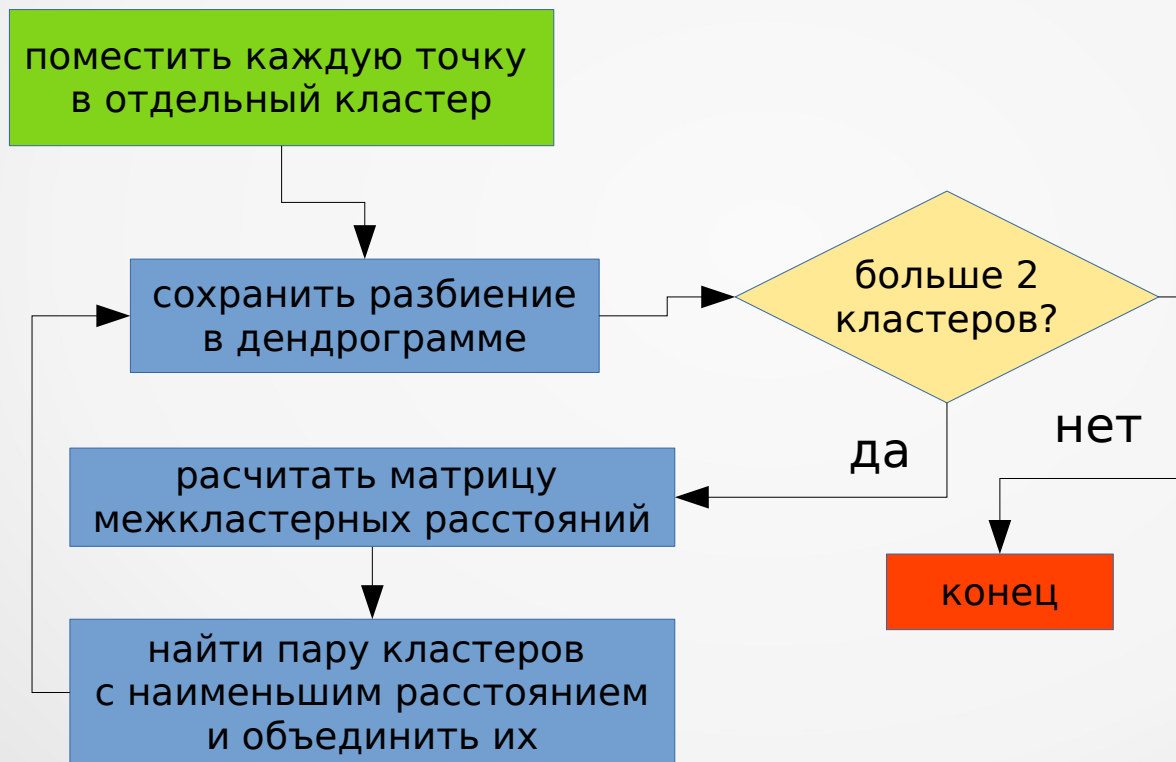
**иерархическая кластеризация:** базовый алгоритм

последовательное объединение близких групп объектов

# метрические методы : кластеризация

**иерархическая кластеризация:** базовый алгоритм

последовательное объединение близких групп объектов



# метрические методы : кластеризация

**иерархическая кластеризация:** регулировка глубины

последовательное объединение близких групп объектов

введём параметр - порог межкластерного расстояния  $\delta$

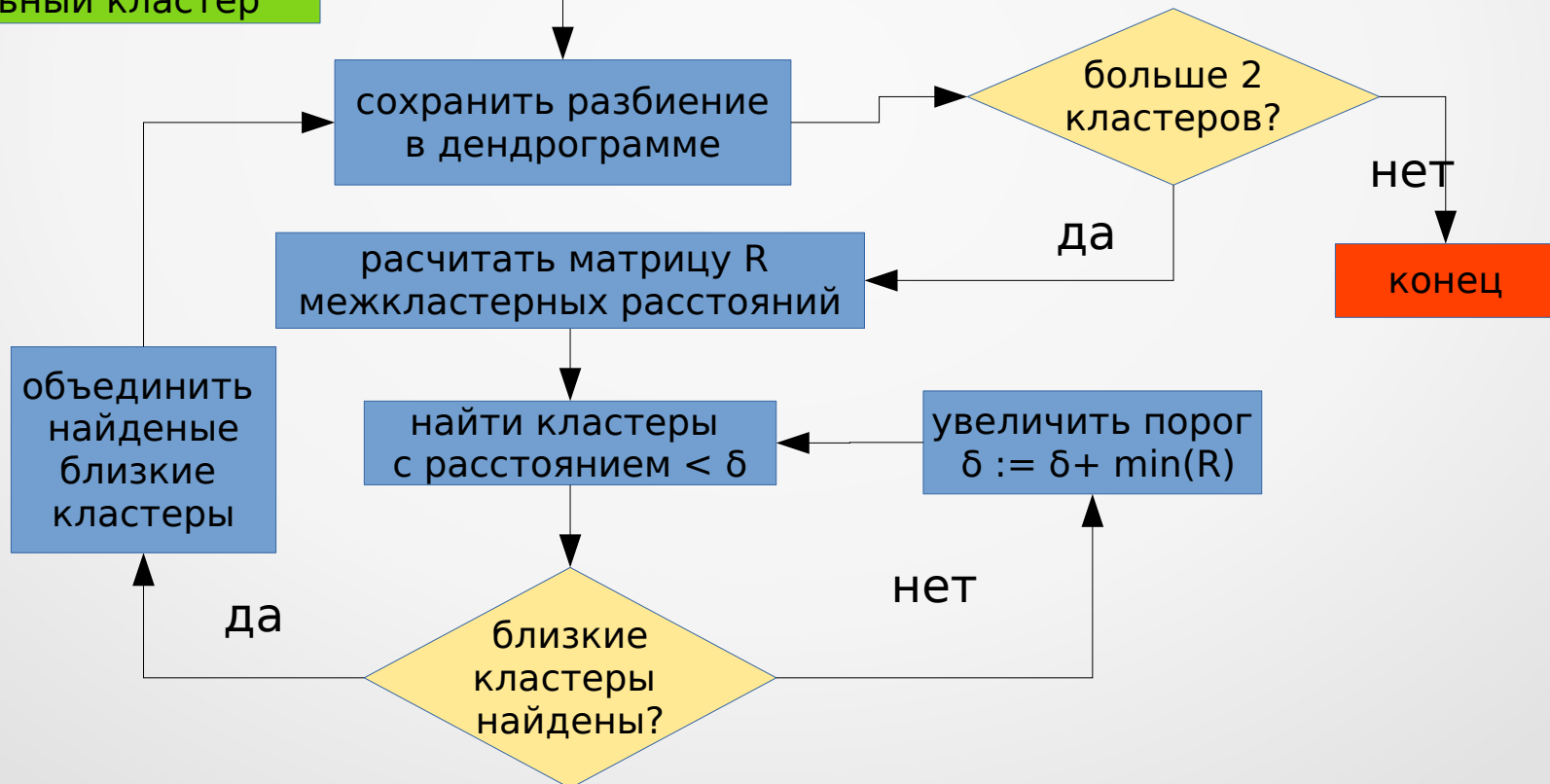
# метрические методы : кластеризация

**иерархическая кластеризация:** регулировка глубины

последовательное объединение близких групп объектов

введём параметр - порог межкластерного расстояния  $\delta$

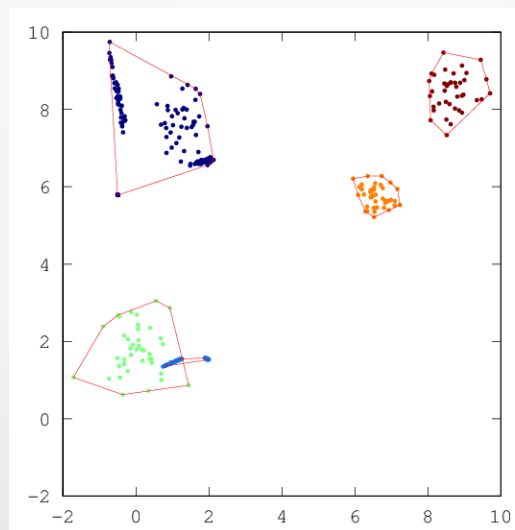
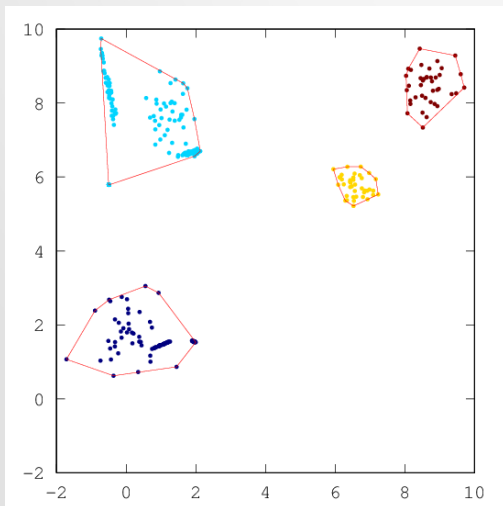
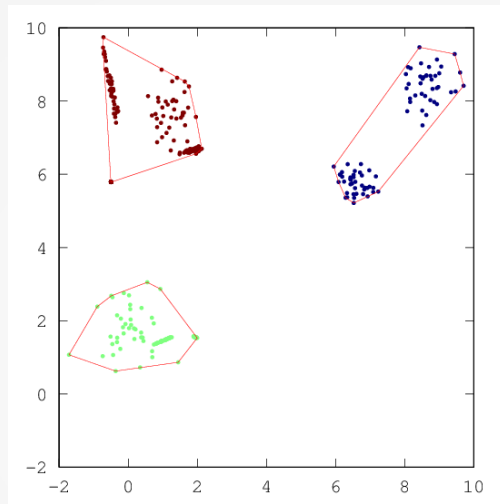
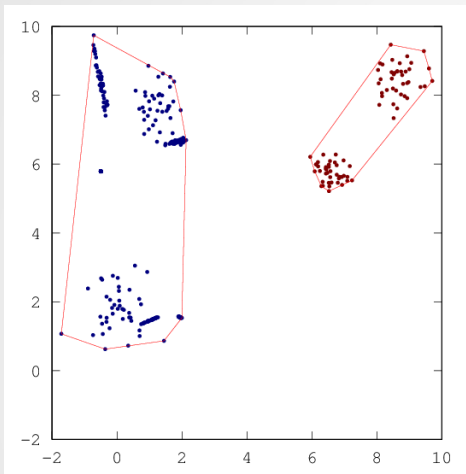
поместить каждую точку  
в отдельный кластер



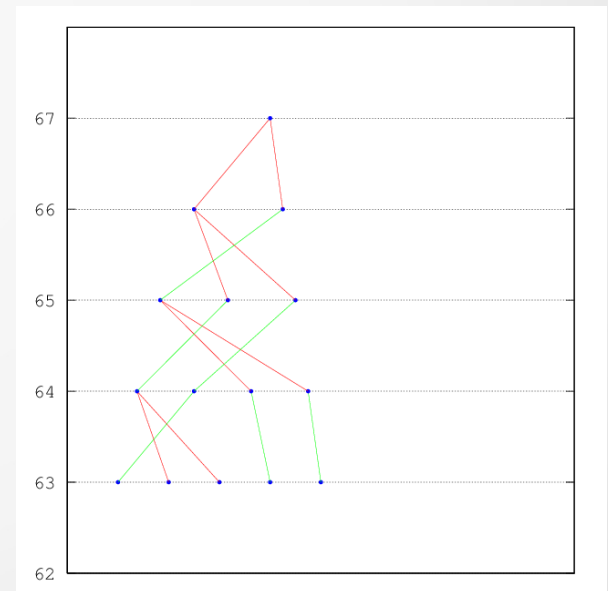
# метрические методы : кластеризация

## иерархическая кластеризация

последовательное объединение близких групп объектов



вершина дендрограммы  
в 67 слоёв



# метрические методы : кластеризация

## приложение - автоматический агрегатор новостей

### КЛАСТЕР 1:

Около 18 тысяч человек покинули подконтрольные боевикам районы Алеппо За минувшие сутки из подконтрольных боевикам районов сирийского города Алеппо было выведено около 17,971 тысячи жителей, в их числе 7,542 тысячи детей. Об этом в субботу, 10 декабря, сообщает ТАСС со ссылкой на российский Центр примирения враждующих сторон в Арабской Республике.

Битва за Алеппо: повстанцы просят дать им вывезти раненых Сирийские повстанцы просят о пятидневном перемирии, чтобы эвакуировать раненых из районов в восточной части Алеппо, после того как они вывели все свои отряды из исторического центра — Старого города.

### КЛАСТЕР 2:

Финальная распродажа! Chery Tiggo от 19990 руб (199,9 млн)  
«Китайские автомобили» объявляют финальную распродажу популярных кроссоверов Chery Tiggo FL! На автомобили в максимальной комплектации установлена специальная цена 19 990 рублей (199,9 млн). Количество автомобилей ограничено!

Не успели купить новый автомобиль в «черную пятницу»? Не нашли ничего подходящего в дилерских автосалонах? Не беда: автосалон «Китайские автомобили» объявляет «черные субботы»! «Черная суббота» — это не шестой трудовой день в советской стране, а желанный праздник для покупателей новеньких авто!

# метрические методы : кластеризация

git clone [https://github.com/mechanoid5/ml\\_lectorium.git](https://github.com/mechanoid5/ml_lectorium.git)

- К.В. Воронцов Методы кластеризации. - курс "Машинное обучение" ШАД Яндекс 2014
- Е.С.Борисов Кластеризатор на основе алгоритма k-means.  
<http://mechanoid.su/ml-k-means.html>
- Е.С.Борисов Метод кластеризации КНП.  
<http://mechanoid.su/ml-knp.html>
- Е.С.Борисов Метод кластеризации ФОРЭЛ.  
<http://mechanoid.su/ml-forel.html>
- Е.С.Борисов Метод иерархической кластеризации.  
<http://mechanoid.su/ml-lnwl.html>
-

# метрические методы : кластеризация



**Вопросы ?**



# метрические методы : кластеризация

## источники данных для экспериментов



sklearn.datasets  
UCI Repository  
kaggle



## задание

- реализовать итоговый обход графа для КНП
- реализовать комбинированный метод ФОРЭЛ+КНП
- реализовать иерархический кластеризатор
- применить кластеризаторы для разных наборов данных
- посчитать оценки результатов кластеризации