Лекция 11
«Кластеризация»

Автор: д.т.н Джомартова Ш.А.

Кластеризация

Кластер (англ. *cluster*) переводится как «группа», «скопление». Этот термин означает множество объектов функционально схожих между собой и собранных в одну связку.

Применительно к интеллектуальному анализу данных под **кластерным анализом** (data clustering) понимается задача разбиения выборки объектов на непересекающиеся подмножества, называемые **кластерами**, так чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Формальная постановка задачи

- Дано множество данных, состоящее из *N* объектов (векторов):

$$S_1, S_2, \ldots, S_N$$

- Каждый объект описывается набором признаков:

$$x_1, x_2, \ldots, x_m,$$

где *m* – размерность пространства признаков

Формальная постановка задачи

- Таким образом, *i*-й объект можно записать в виде:

$$S_i = (x_{i1}, x_{i2}, \ldots, x_{im})$$

- Класс для каждого объекта неизвестен

Формальная постановка задачи

Требуется:

- найти способ сравнения $d(S_p, S_q)$ объектов между собой (меру сходства, функцию расстояния)
- определить множество кластеров

$$C_1, C_2, \ldots, C_r$$

причем количество кластеров r – неизвестно

- разбить данные по кластерам

Метрики расстояния между объектами

- евклидово расстояние

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}$$

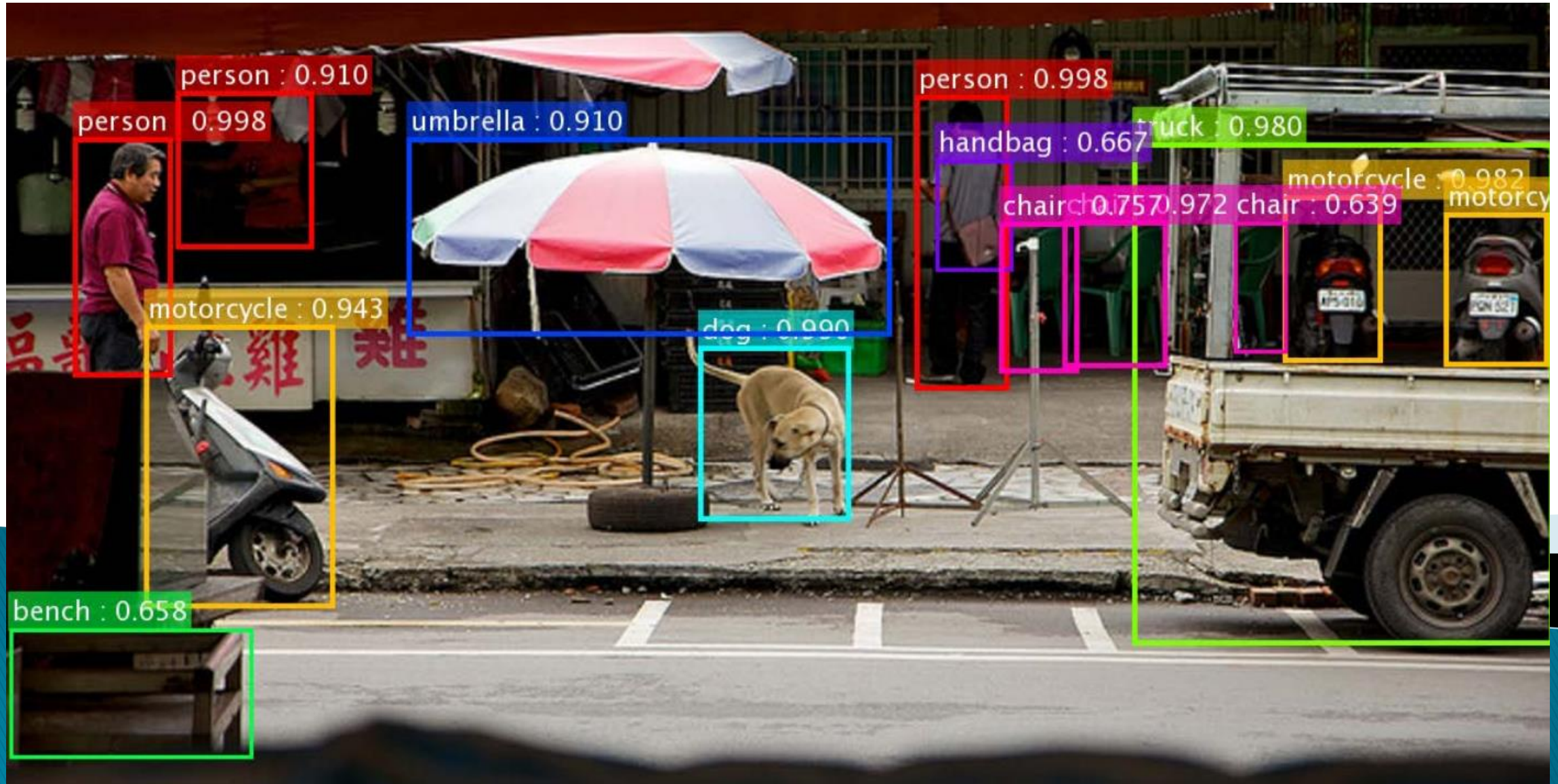
- Манхэттенское расстояние

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

- расстояние Чебышева

$$d(x,y) = \max_{1 \le i \le n} |x_i - y_i|$$

Пример: Распознавание образов



Пример: Классификация результатов поиска























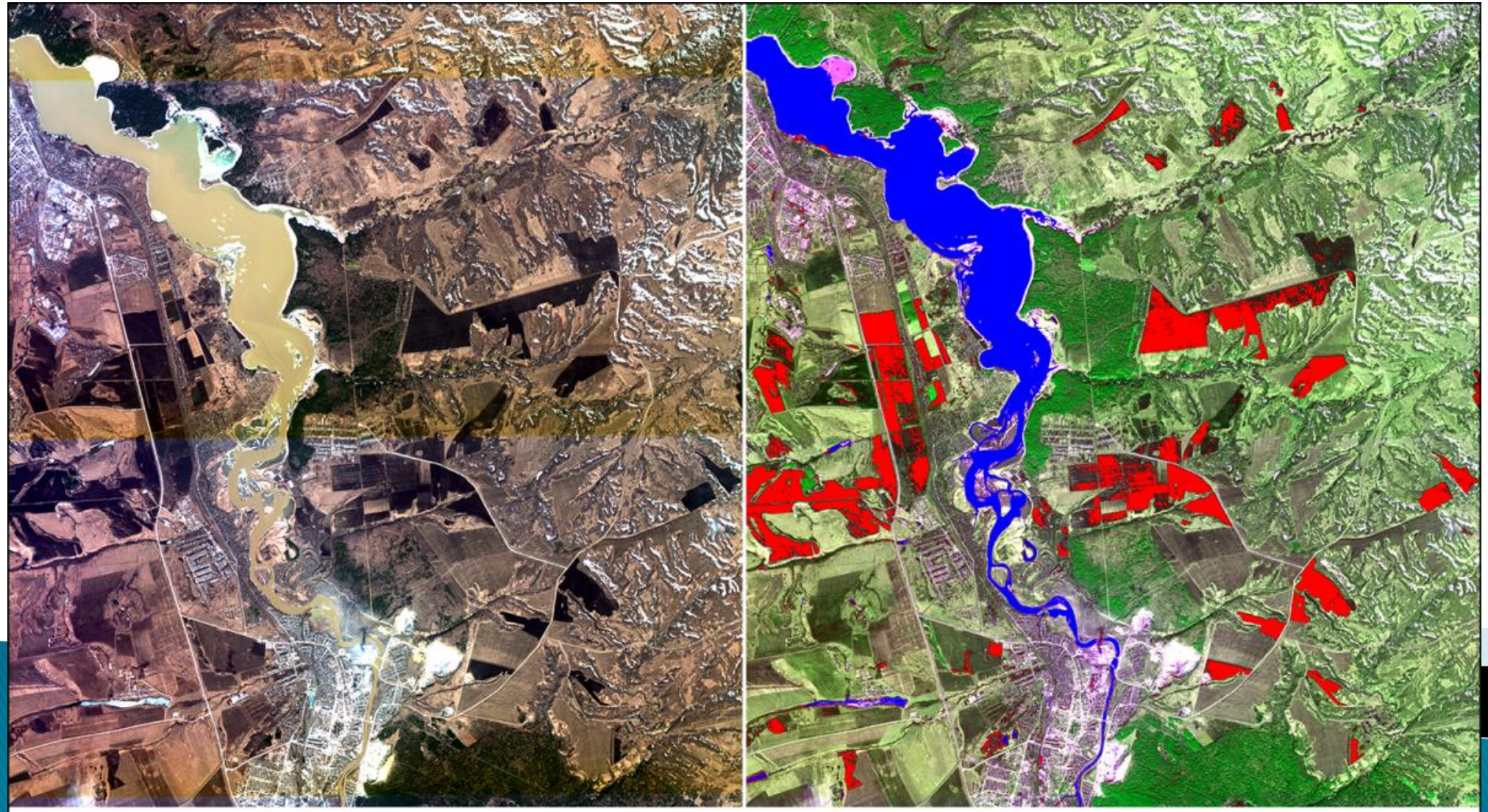






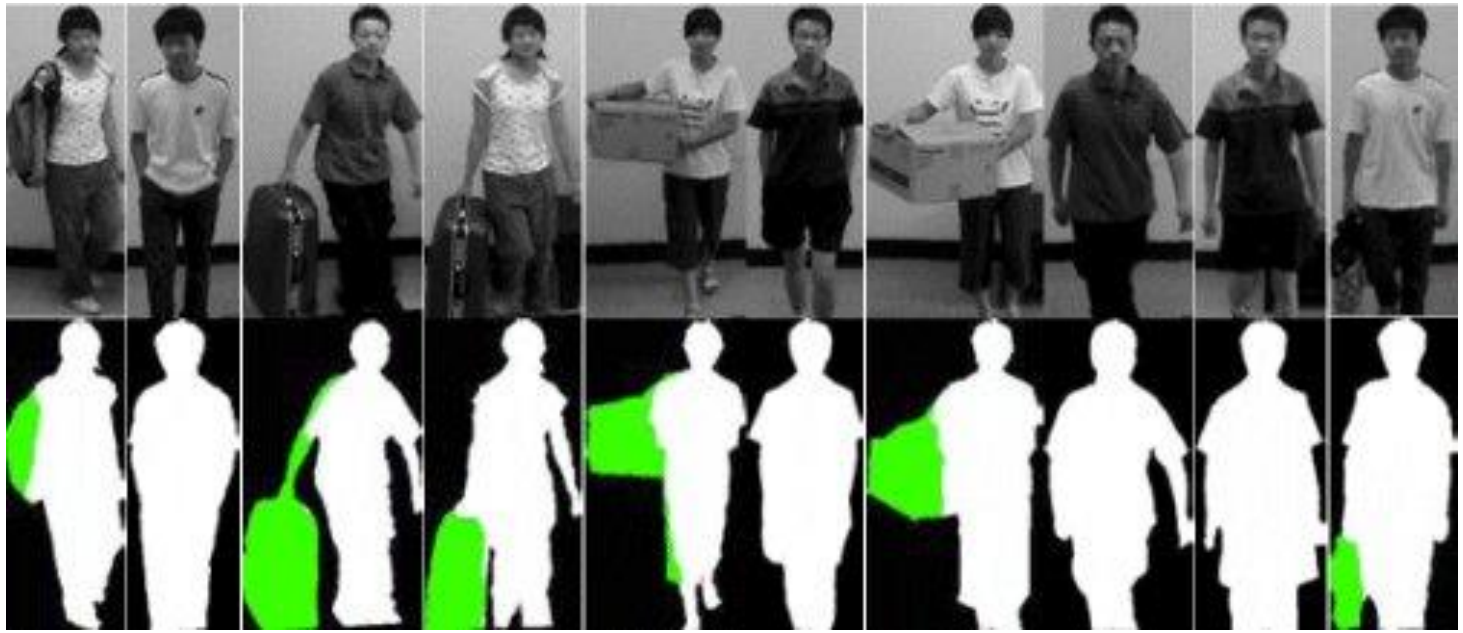


Пример: Сегментация изображений



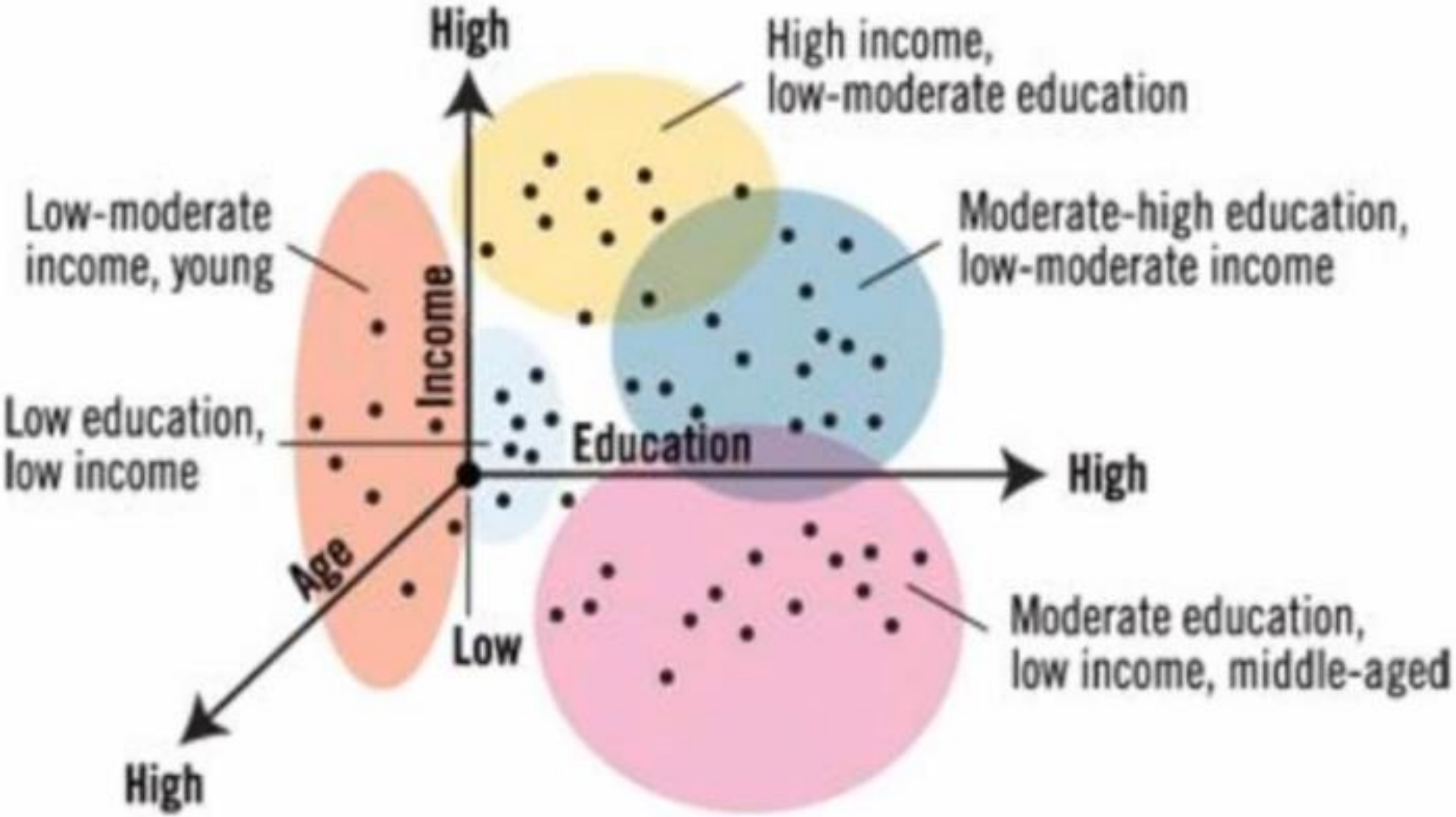
RGB-композит изображения, полученного со спутника «Канопус-В»,
и результат выделения водных объектов и переувлажненных почв

Пример: Сегментация изображений





Пример: Сегментация рынка



Пример: класстеризация текста по темам



Скриншот с сайта РИА Новости (ria.ru)

Ранее: обучение на размеченных данных (supervised learning)

Обучающая выборка:

$x_1, \dots, x_l$ - объекты

$y_1, \dots, y_l$ - ответы

Тестовая выборка:

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

В регрессии: y_i - прогнозируемая величина

В классификации: y_i - метка класса

Восстановление отображения

Считаем, что есть отображение:

$$x \mapsto y$$

Обучающая выборка – это примеры значений, по которым мы пытаемся построить $a(x)$:

$$a(x) \approx y$$

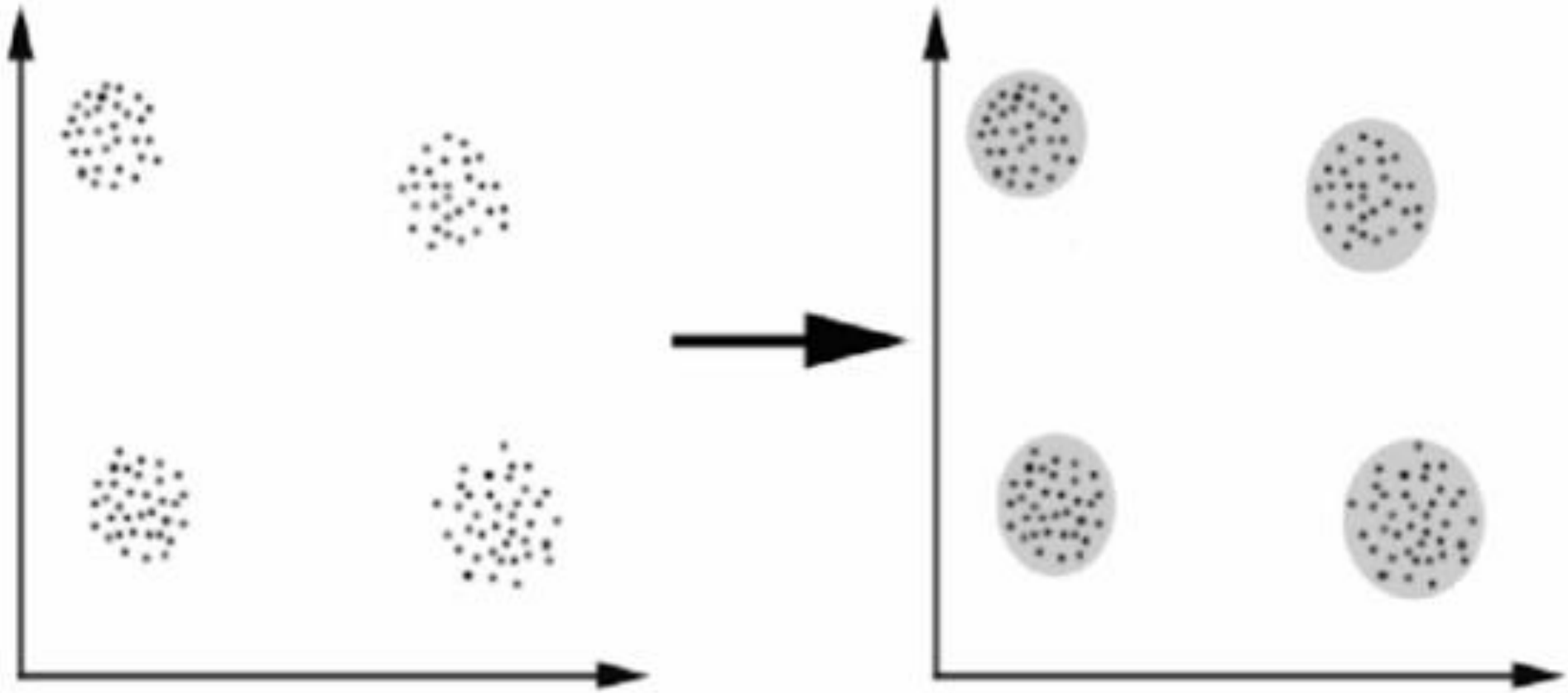
Кластеризация

«Обучающая» выборка:
$x_1, \dots, x_l$ - объекты

Она же и тестовая

Нужно поставить метки $y_1, \dots, y_l$, так, чтобы объекты с одной и той же меткой были похожи, а с разными метками – не очень похожи

Как это выглядит



Восстановление отображения в кластеризации

Считаем, что есть отображение:

$$x \mapsto y$$

Пытаемся построить $a(x)$, но примеров y теперь нет.

Нужно не приближать известные значения, а строить отображение с некоторыми хорошими свойствами.

Среднее внутрикластерное расстояние

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

Среднее межкластерное расстояние

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

Придумываем метрику качества

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

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

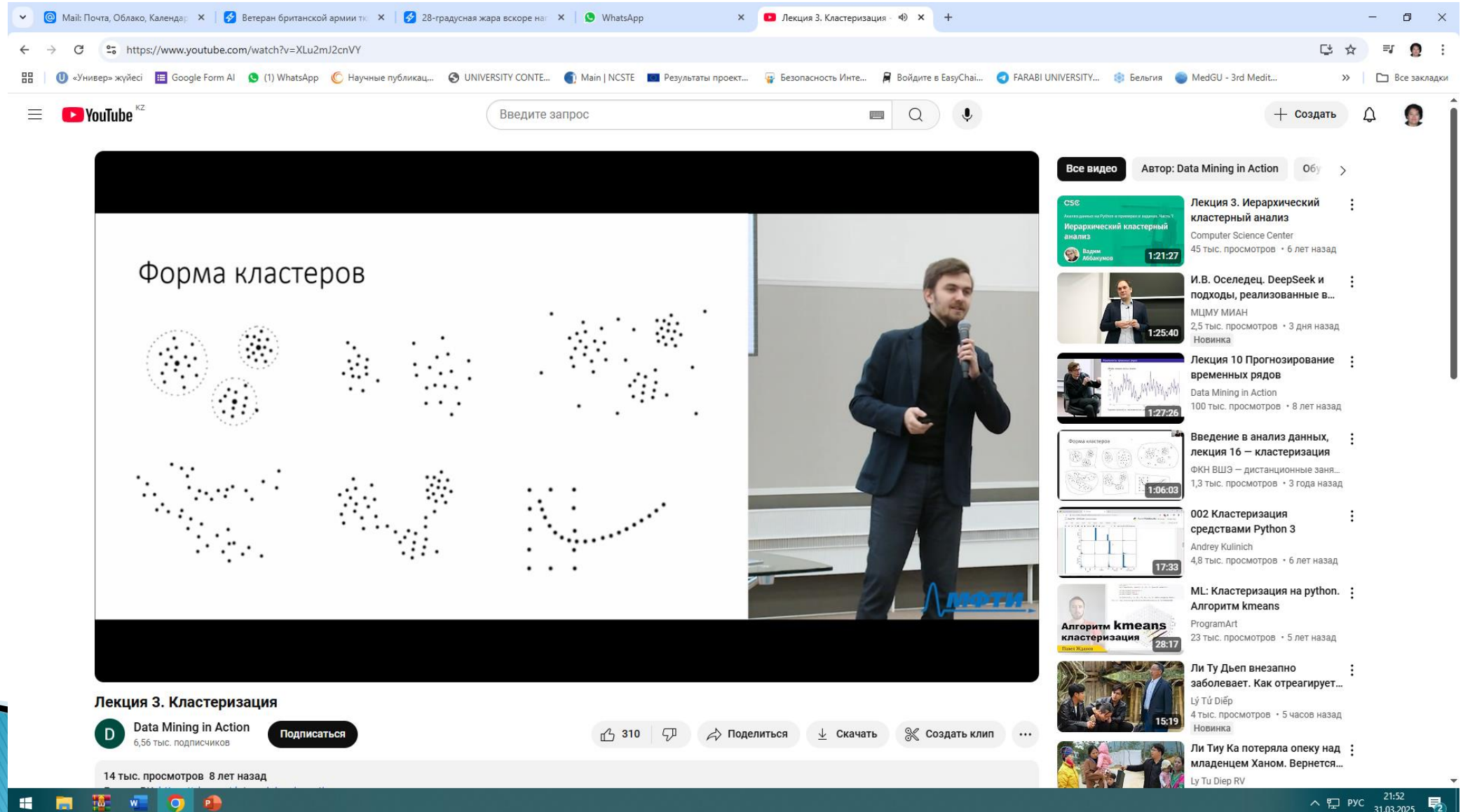
$$F_0 / F_1 \to \min$$

Разнообразие задач кластеризации

Зачем нужны разные алгоритмы кластеризации

- Каждые данные в чем-то «особенные»
- Каждая задача кластеризации тоже
- В разных задачах кластеризации могут быть отличия:
  - Форма кластеров
  - Необходимость делать кластеры вложенными друг в друга
  - Размер кластеров
  - Кластеризация - основная задача или побочная
  - «Жесткая» или «мягкая» кластеризация
- В задачах с разными особенностями могут быть уместны разные методы

Форма кластеров

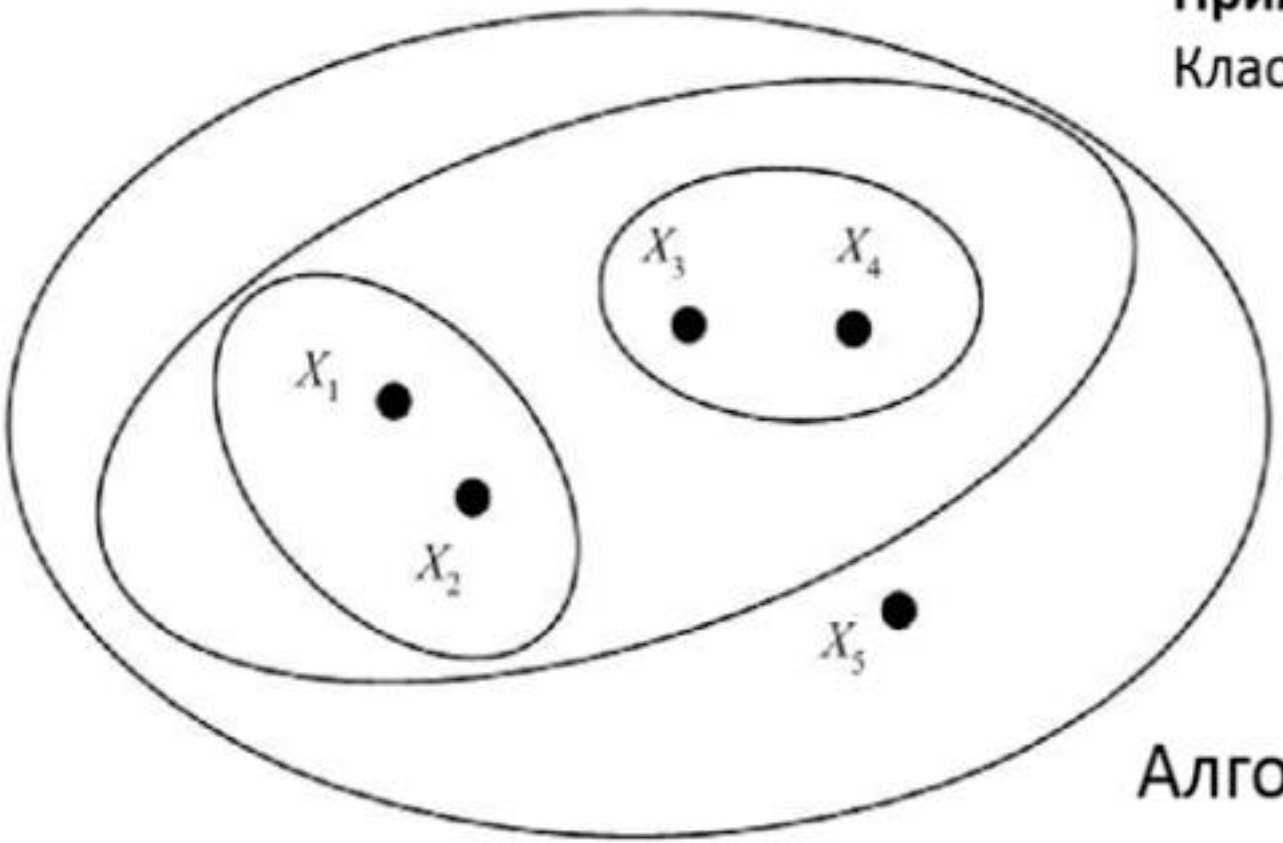


Форма кластеров

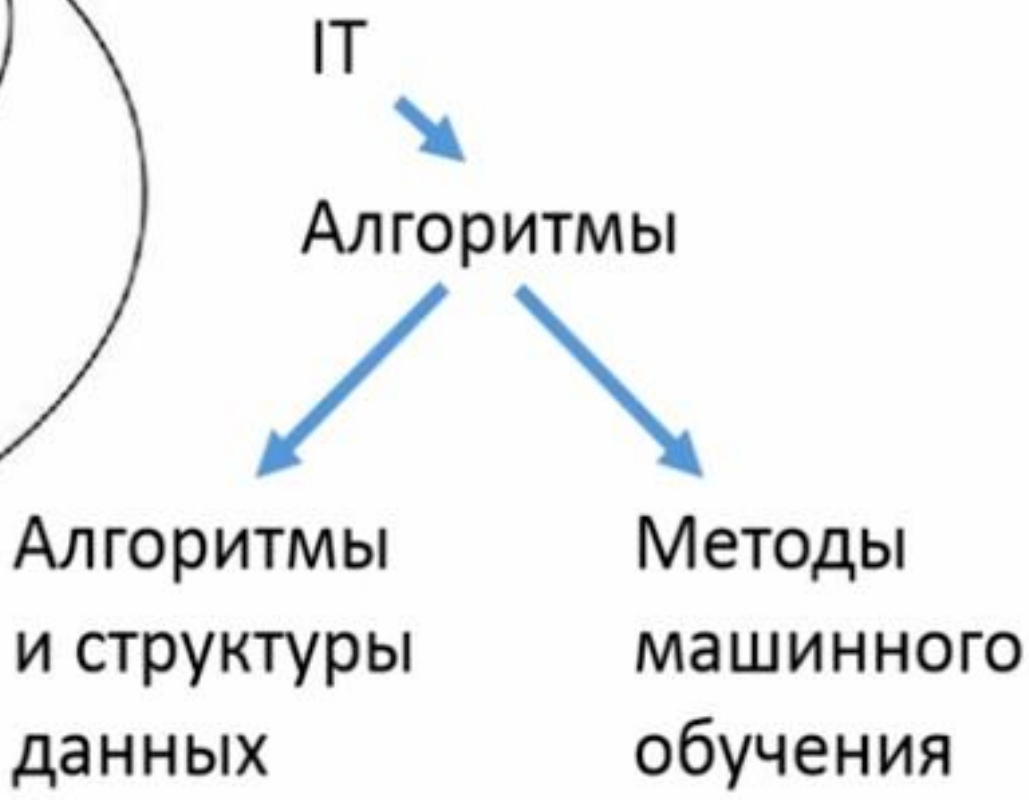




Вложенность кластеров



Пример:
Кластеризация статей с Хабрахабра



Размер кластеров

- 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).

Размер кластеров

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



Скриншот с сайта РИА Новости (ria.ru)

Размер кластеров

- 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

Основная задача или вспомогательная

Кластеризация новостей

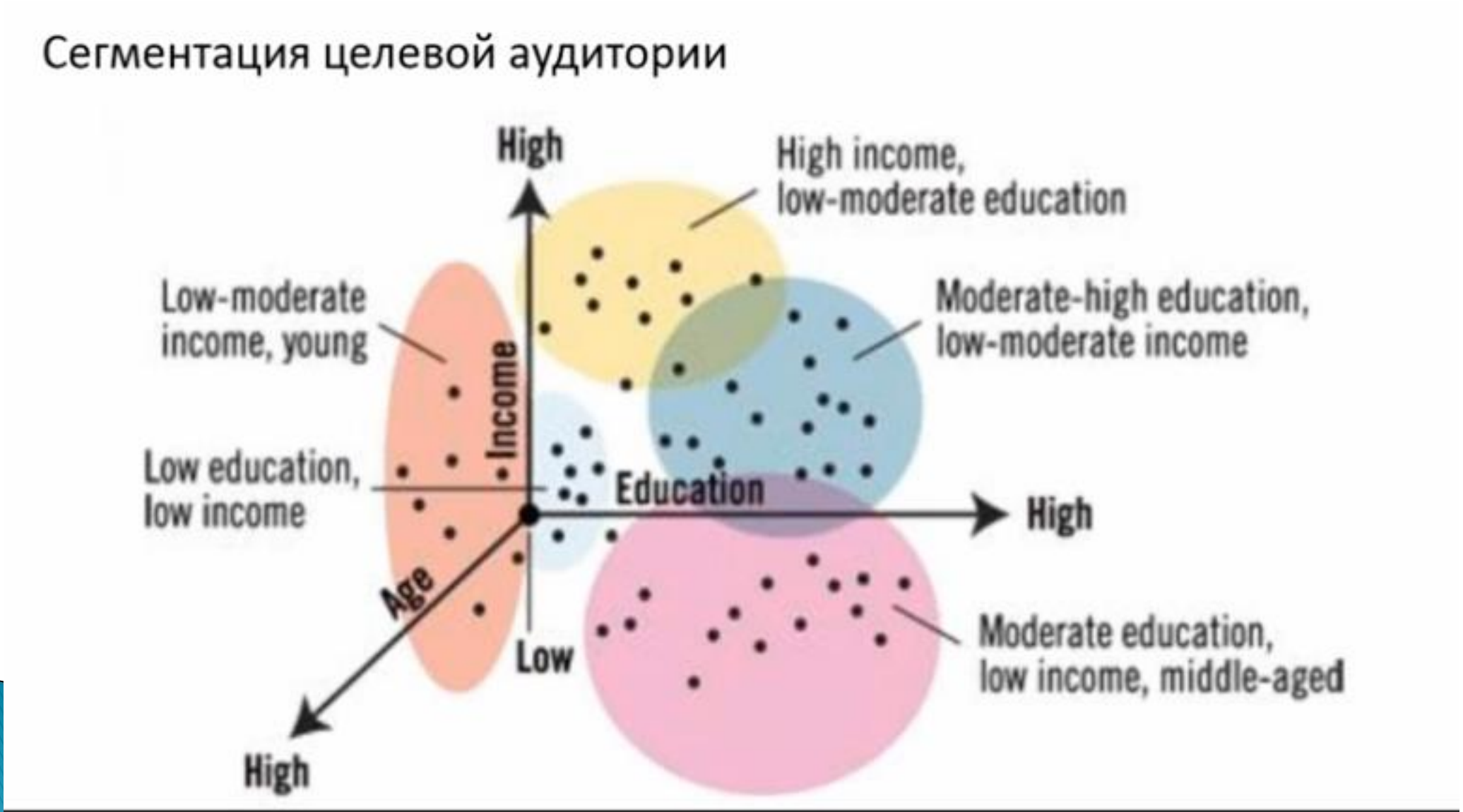
11:41, 08 ФЕВРАЛЯ 2014

Открытие Олимпиады в Сочи посмотрели несколько миллиардов человек

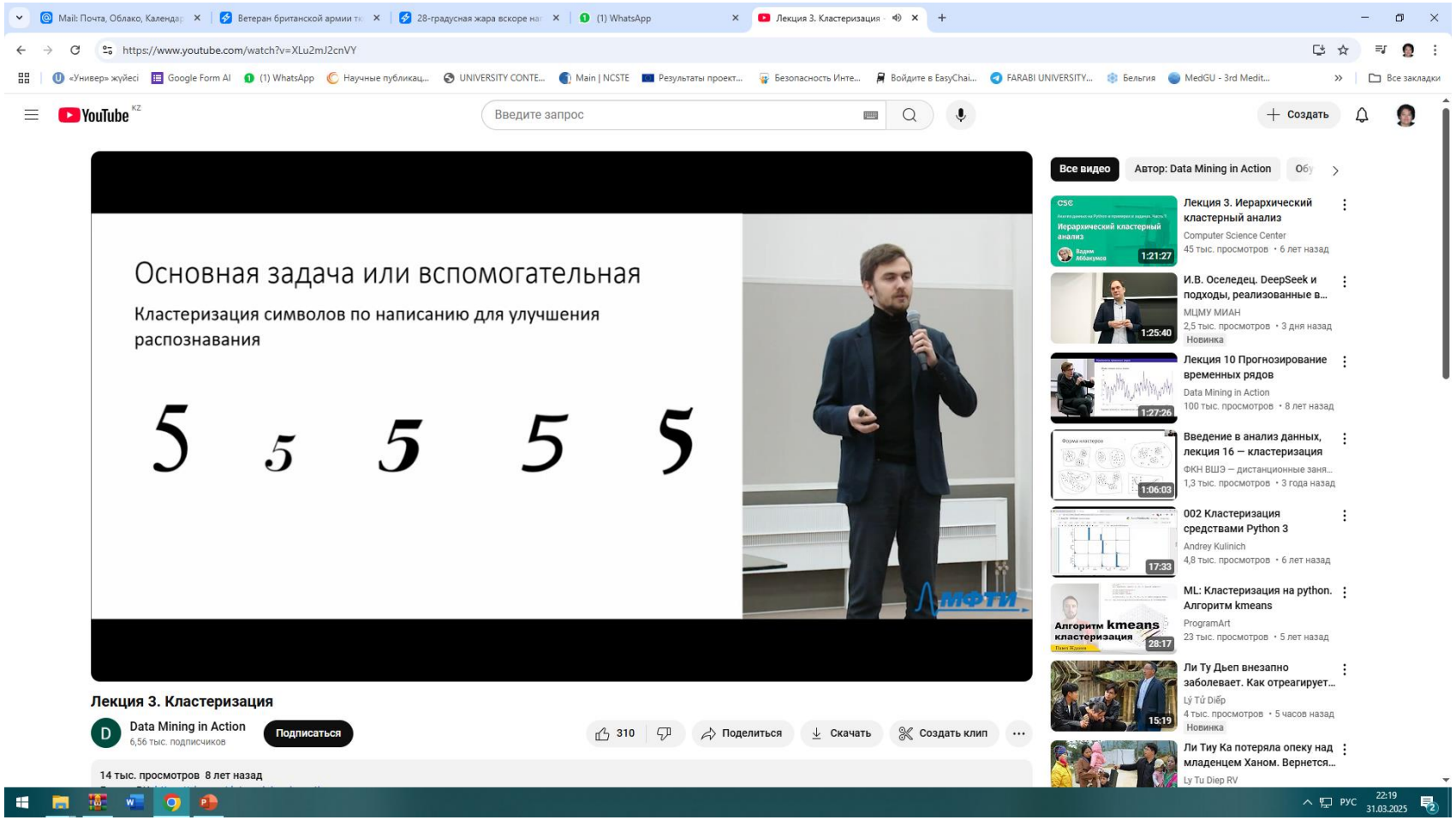
Олимпиада в Сочи открыта

Церемония открытия Олимпиады в Сочи. Онлайн-репортаж

Основная задача или вспомогательная



Основная задача или вспомогательная



«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»



0.2



0.3



0.5

Метод К средних (K-Means)

Как работает K-Means



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

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 К?
- Вариант выбора начальных приближений:
  - первый центр выбираем случано 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

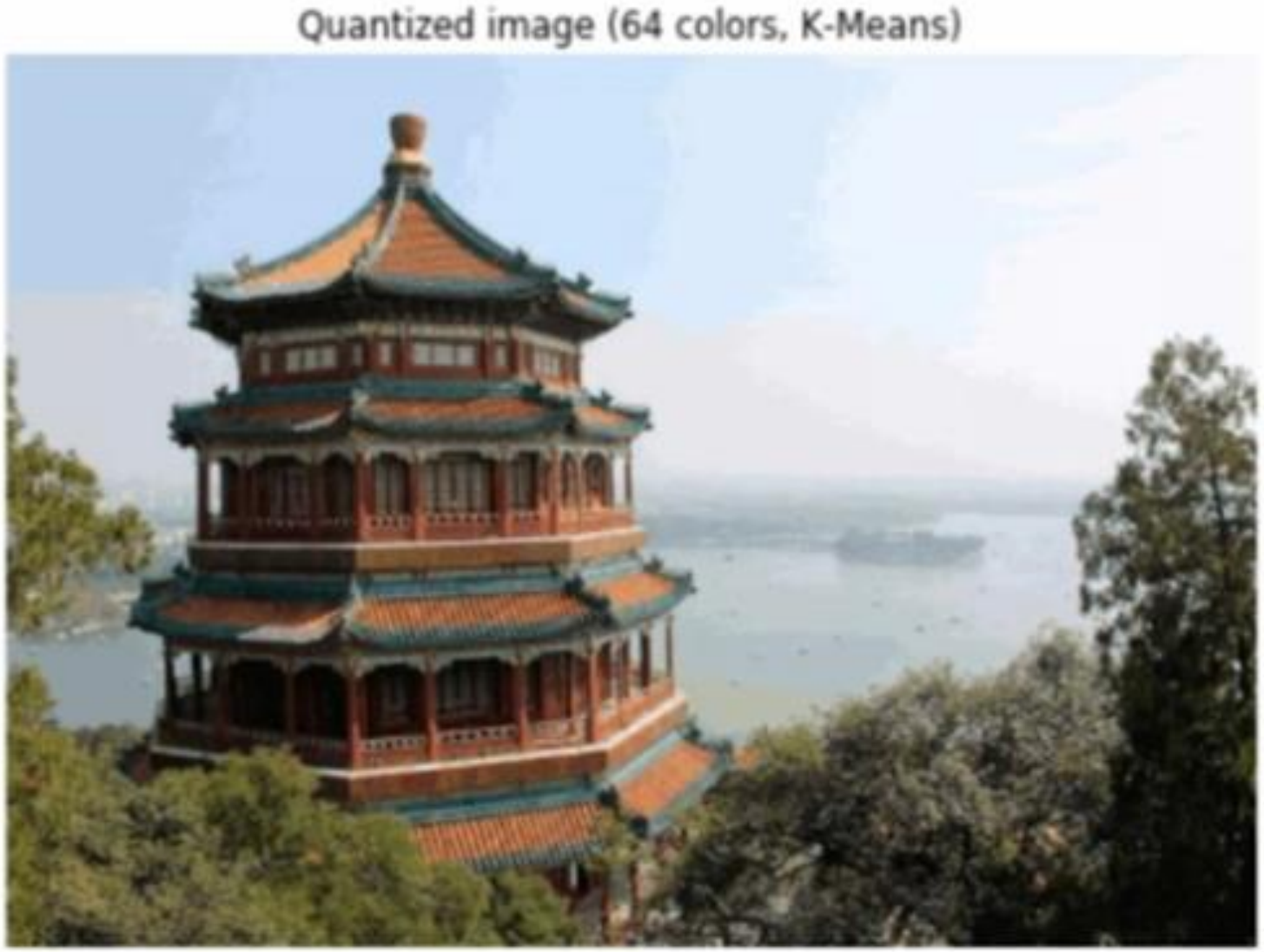
Пример



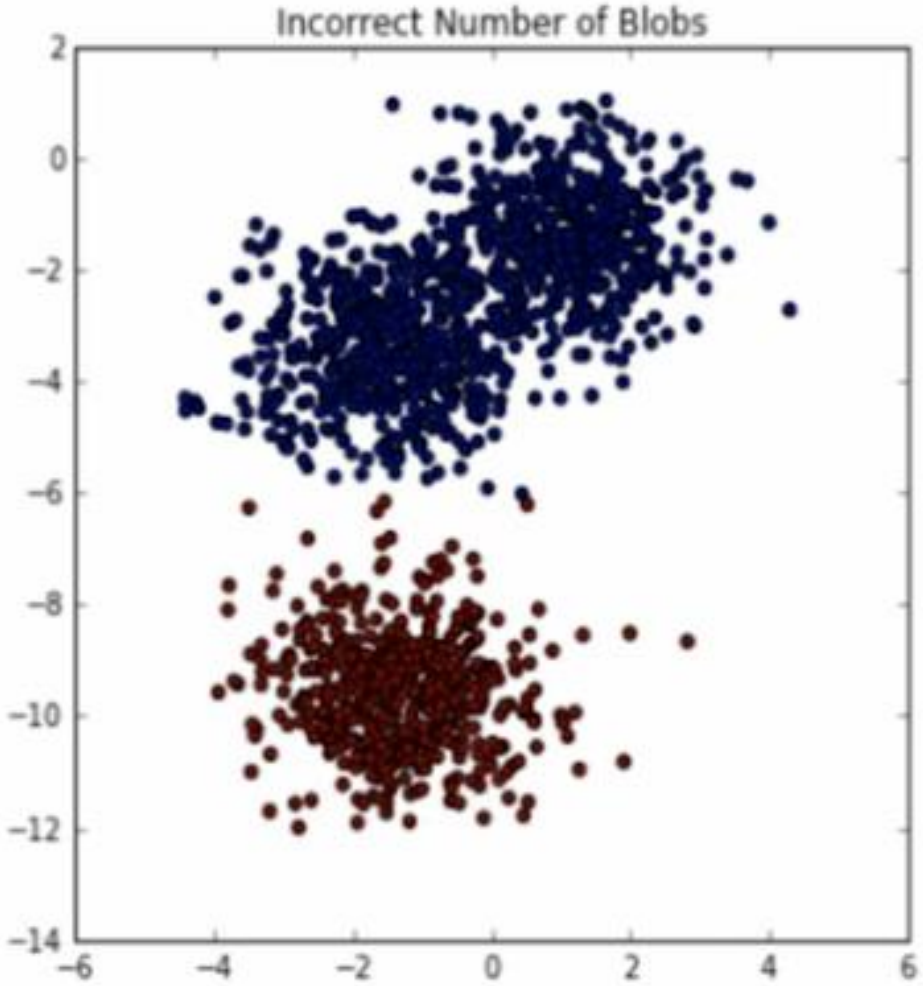
Пример



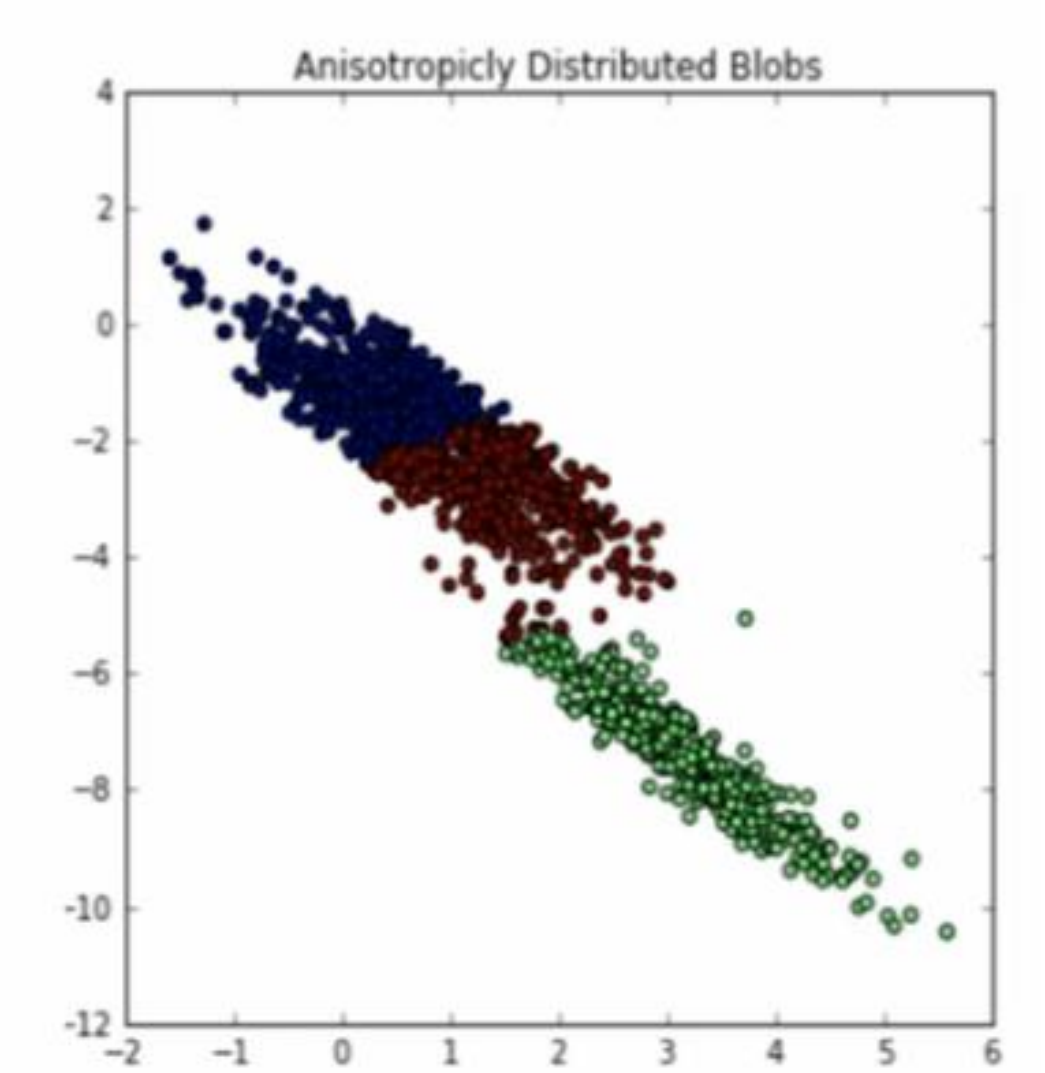
Пример



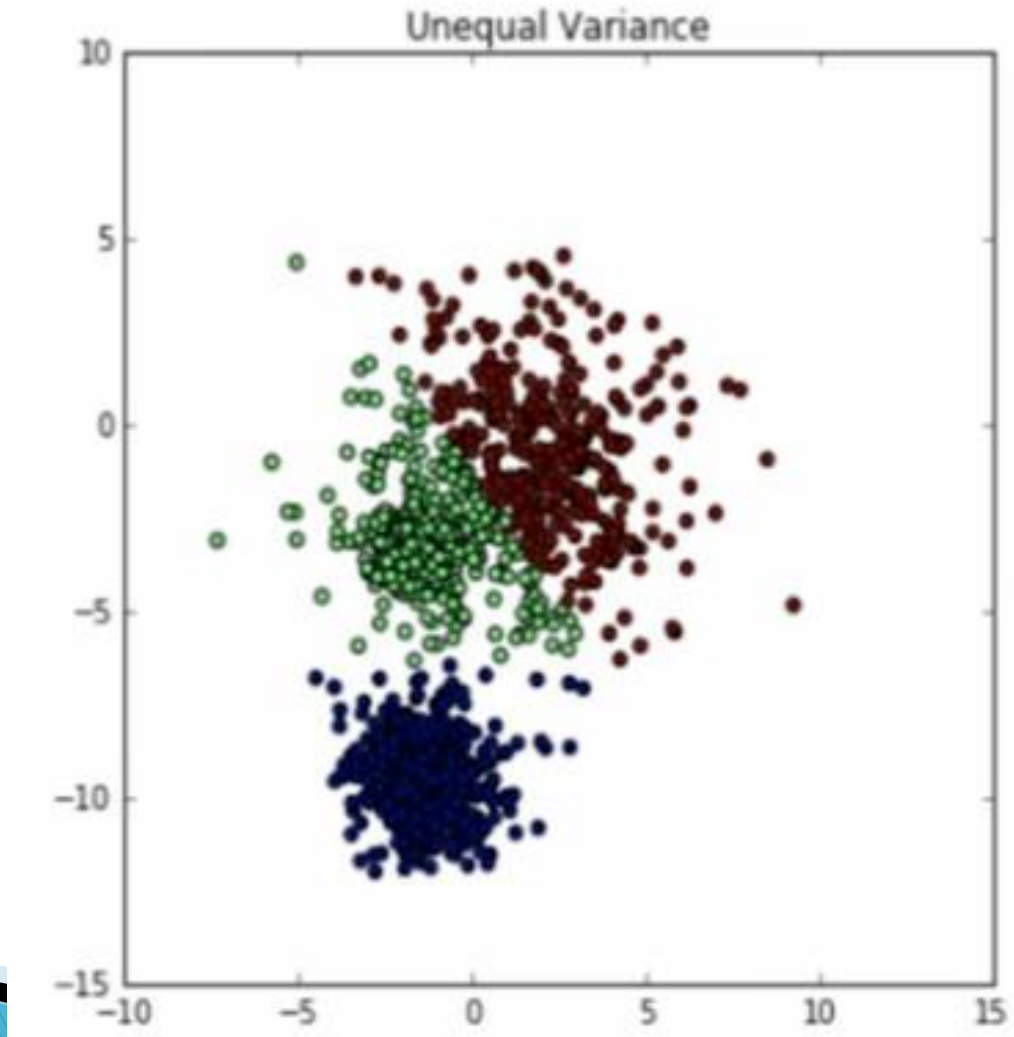
K-Means и разные формы кластеров



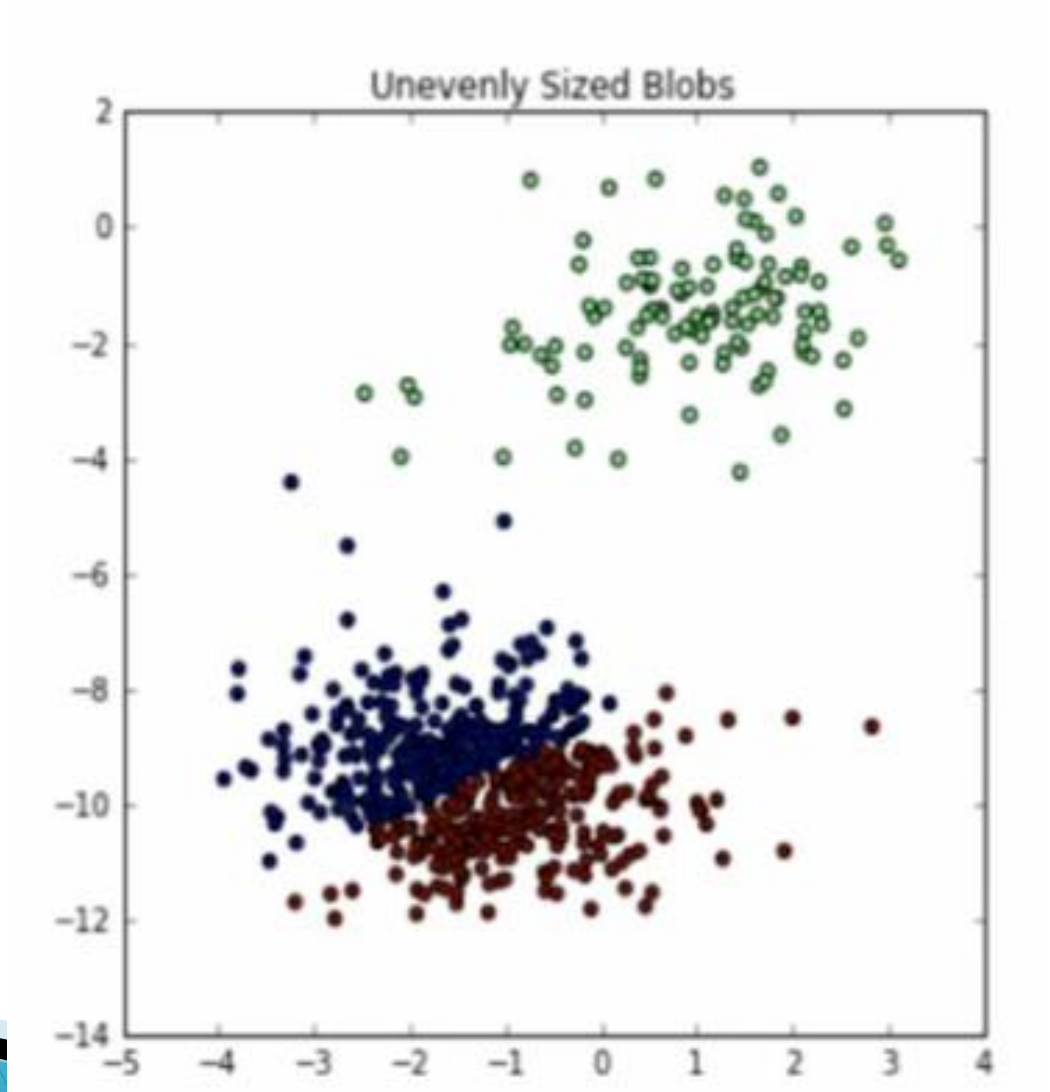
K-Means и разные формы кластеров



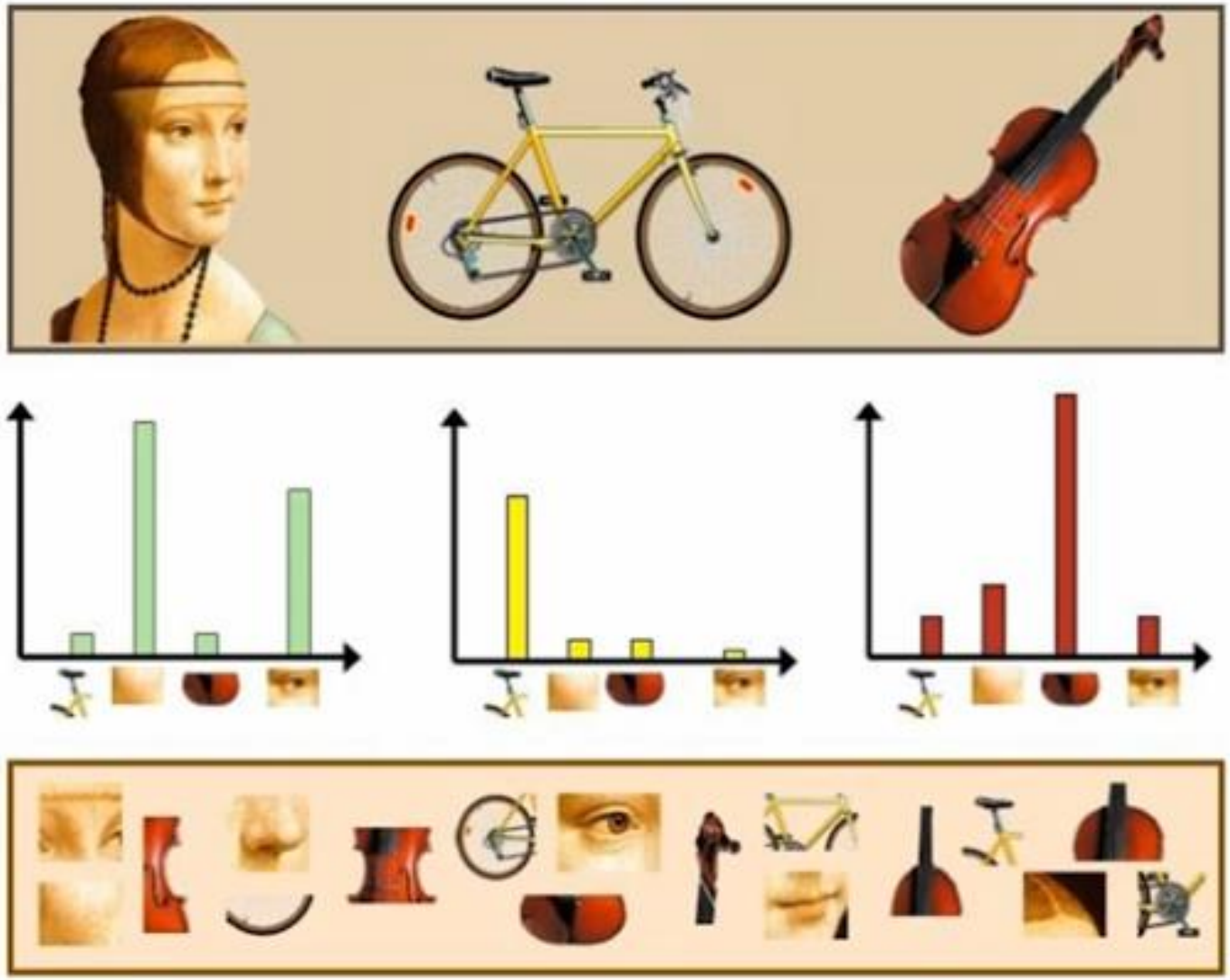
K-Means и разные формы кластеров



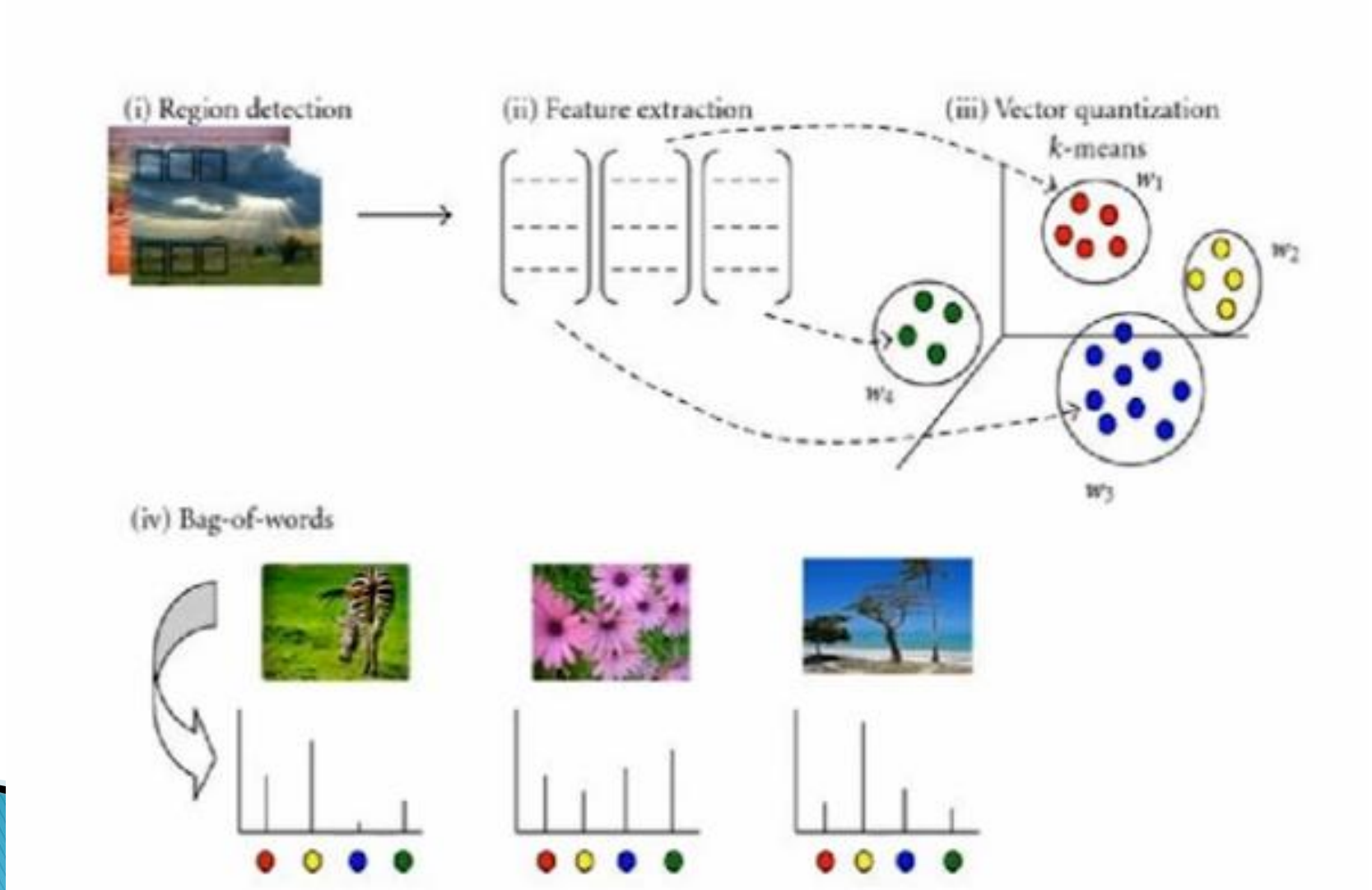
K-Means и разные формы кластеров



Пример: мешок визуальных слов



Пример: мешок визуальных слов



Что оптимизирует K- Means

Среднее внутрикластерное расстояние:

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

Альтернативный вариант, если есть центры кластеров:

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

Что оптимизирует K- Means

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

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

Что оптимизирует K- Means

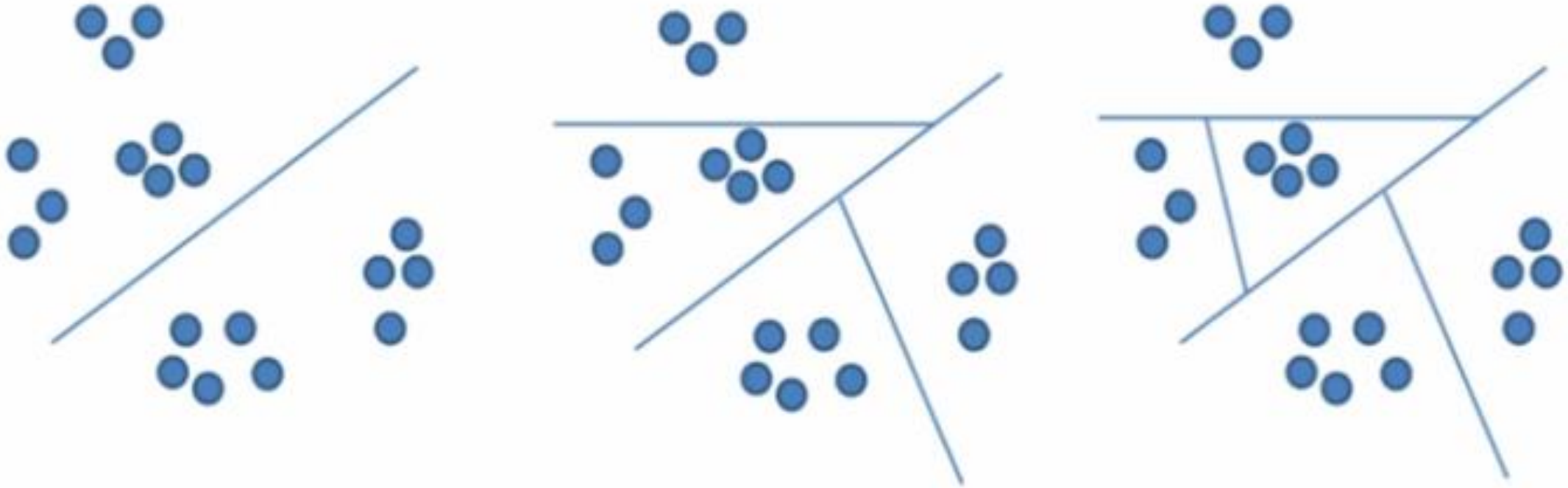
K Means итеративно минимизирует среднее внутрикластерное расстояние:

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

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

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

Подбор числа кластеров: BisectKMeans



Особенности K-Means

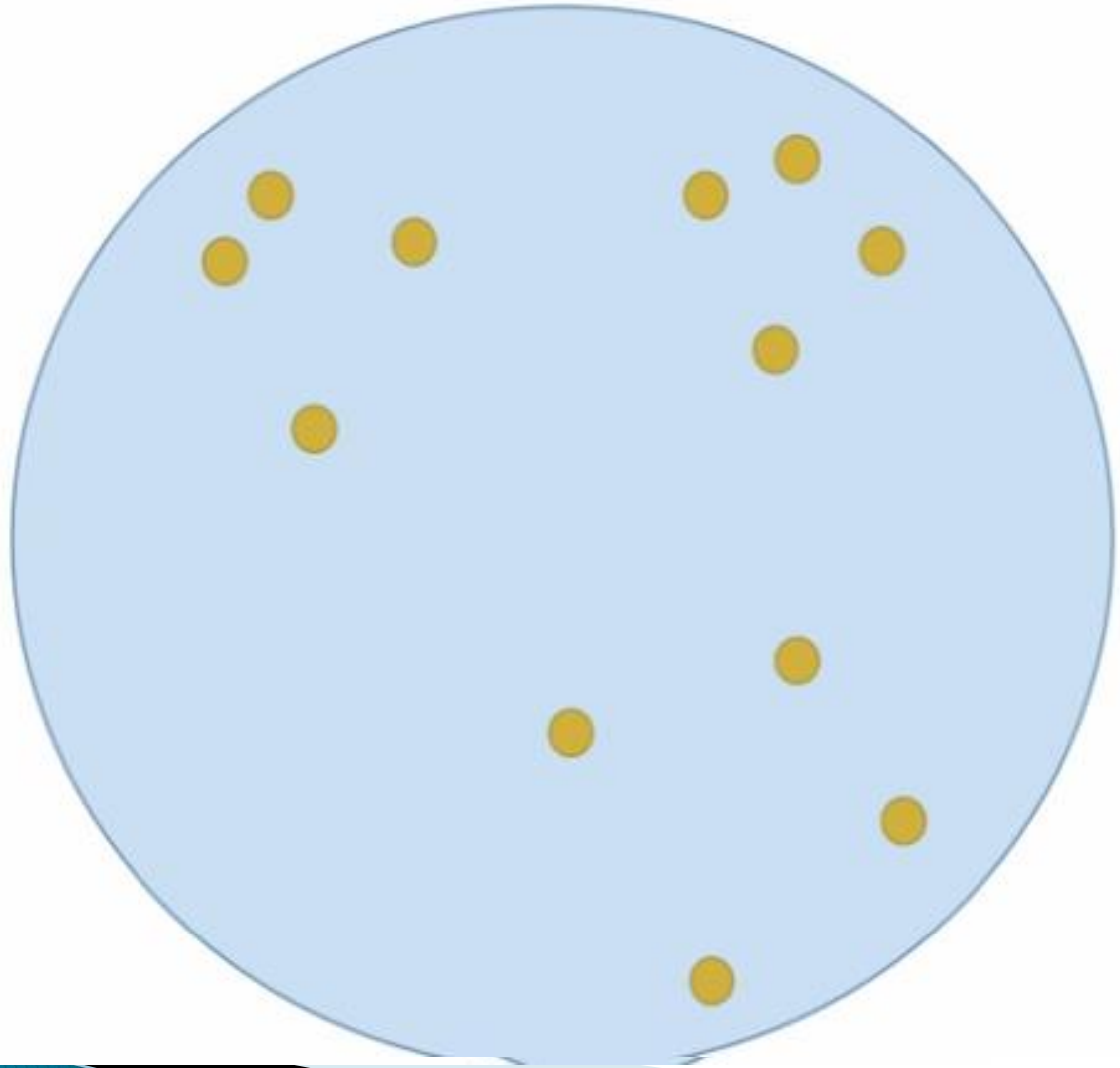
- Может работать с большими объёмами данных
- Подходит для кластеров с простой геометрией
- Требует выбора числа кластеров

Агломеративная иерархическая кластеризация

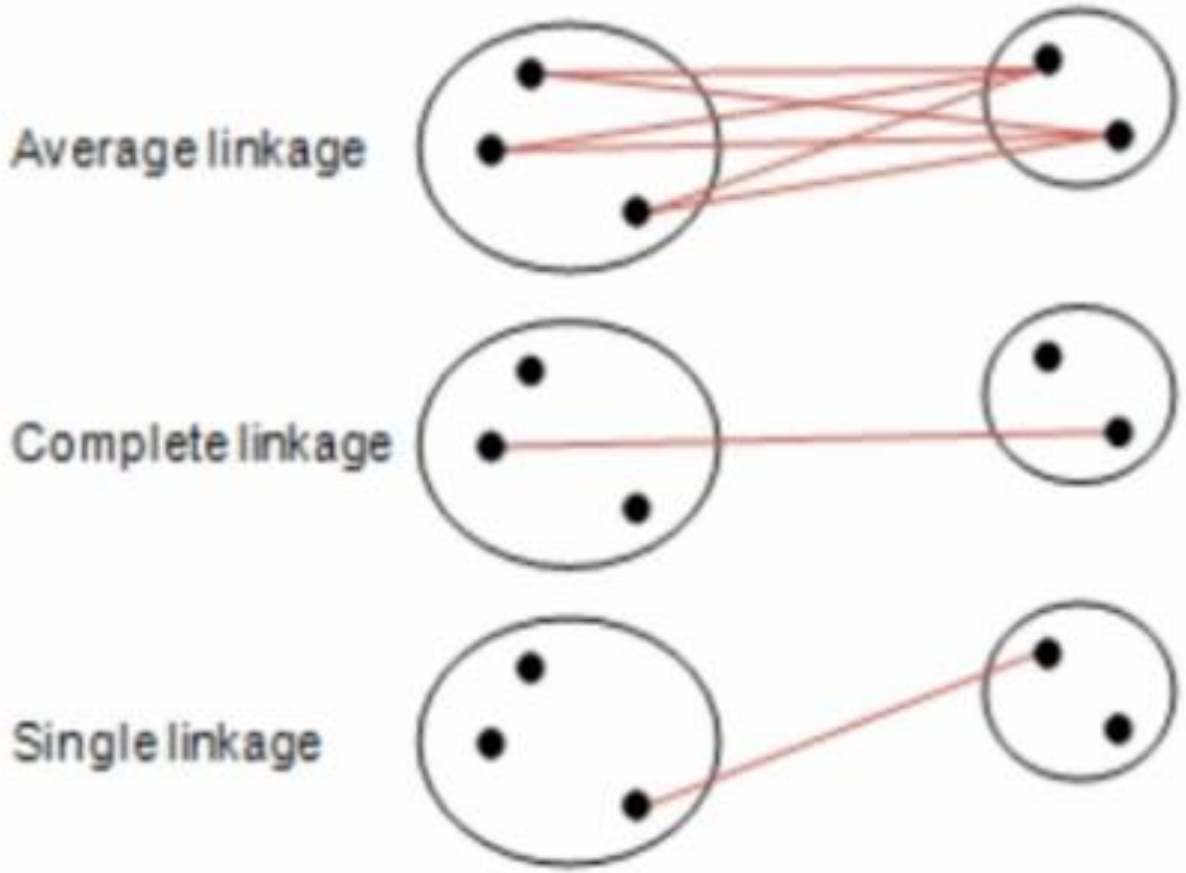
Иерархическая кластеризация

- Агломеративная
- Дивизионная или дивизимная

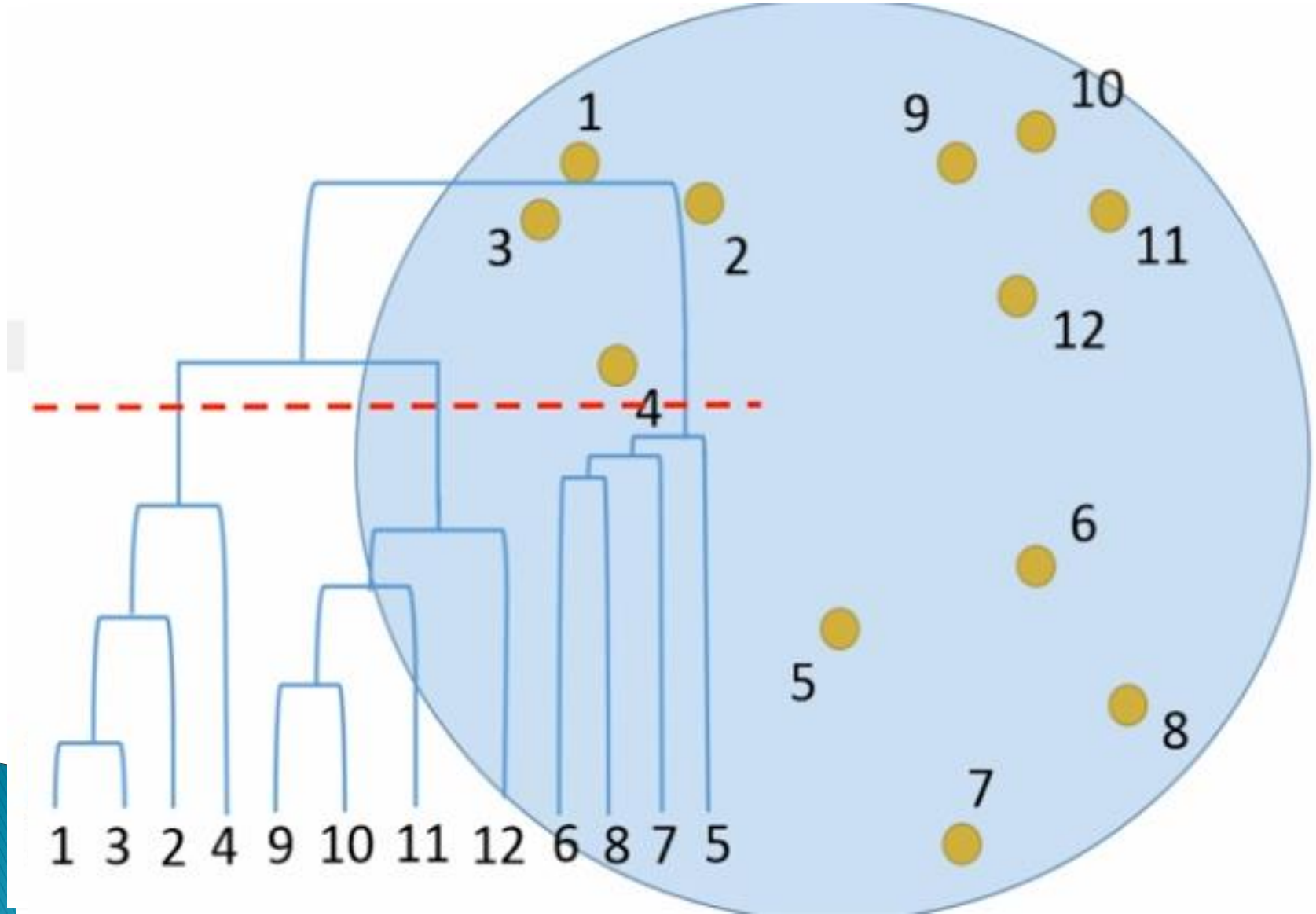
Агломеративная кластеризация

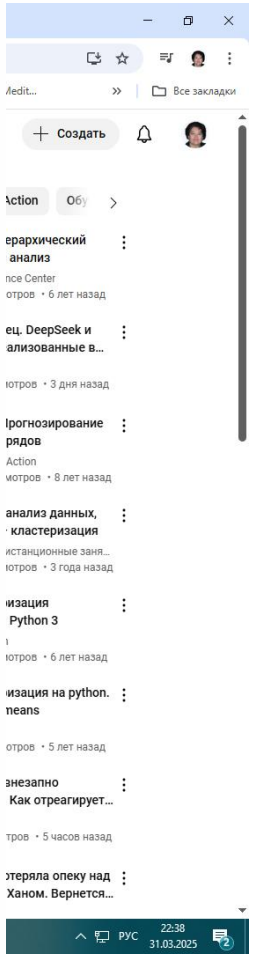


Расстояние между кластерами

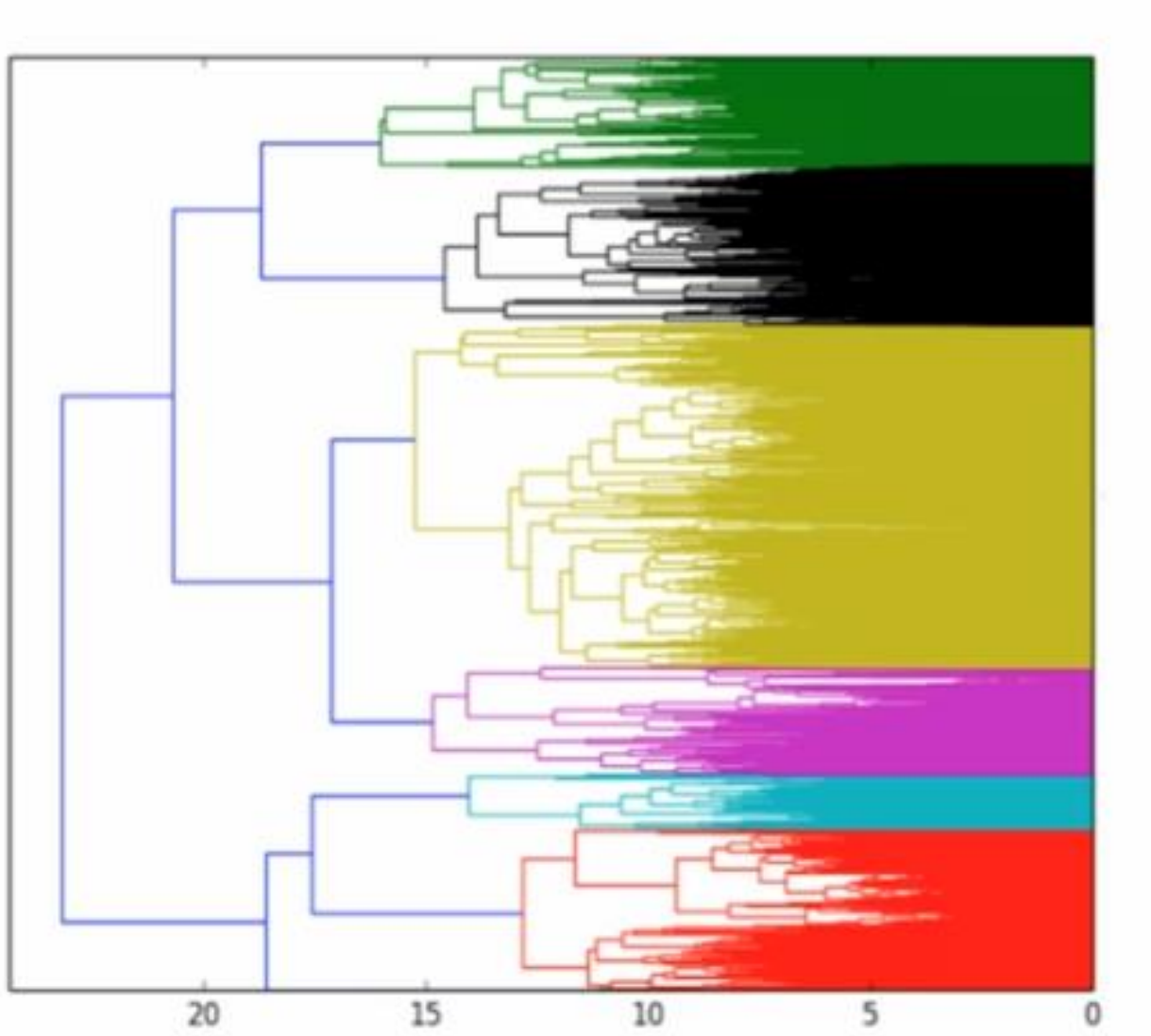


Дендрограмма



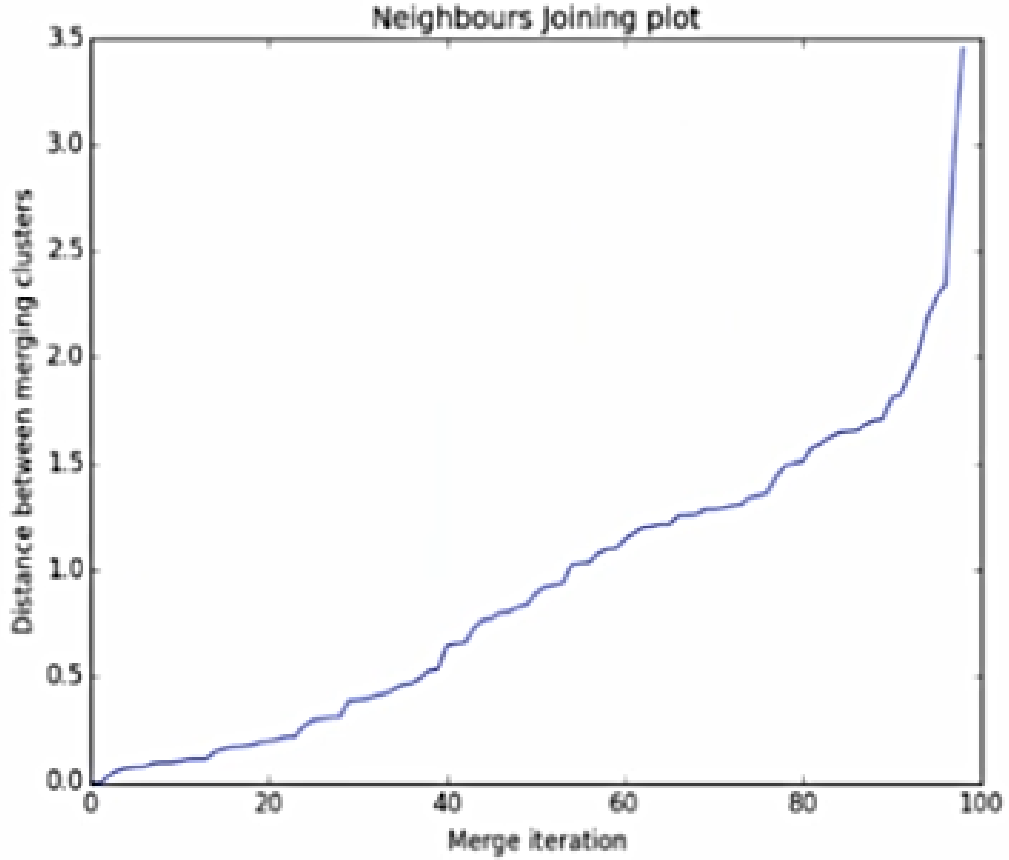


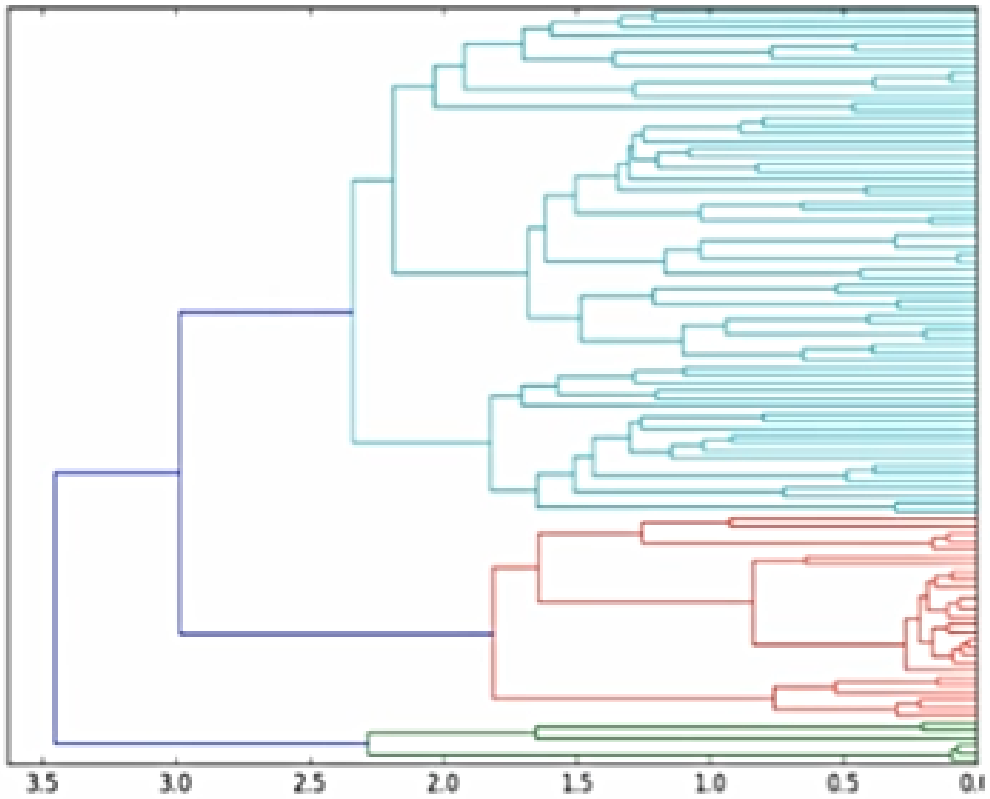
Пример: кластеризация писем



Пример: расстояние между кластерами

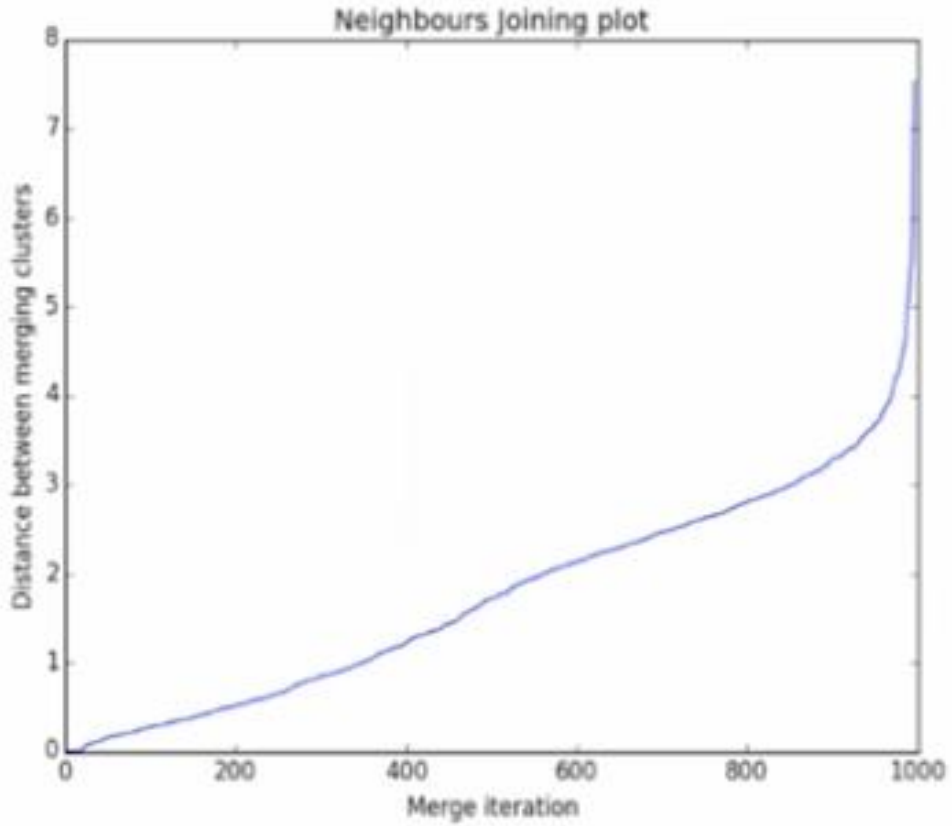
На подвыборке из 100 писем

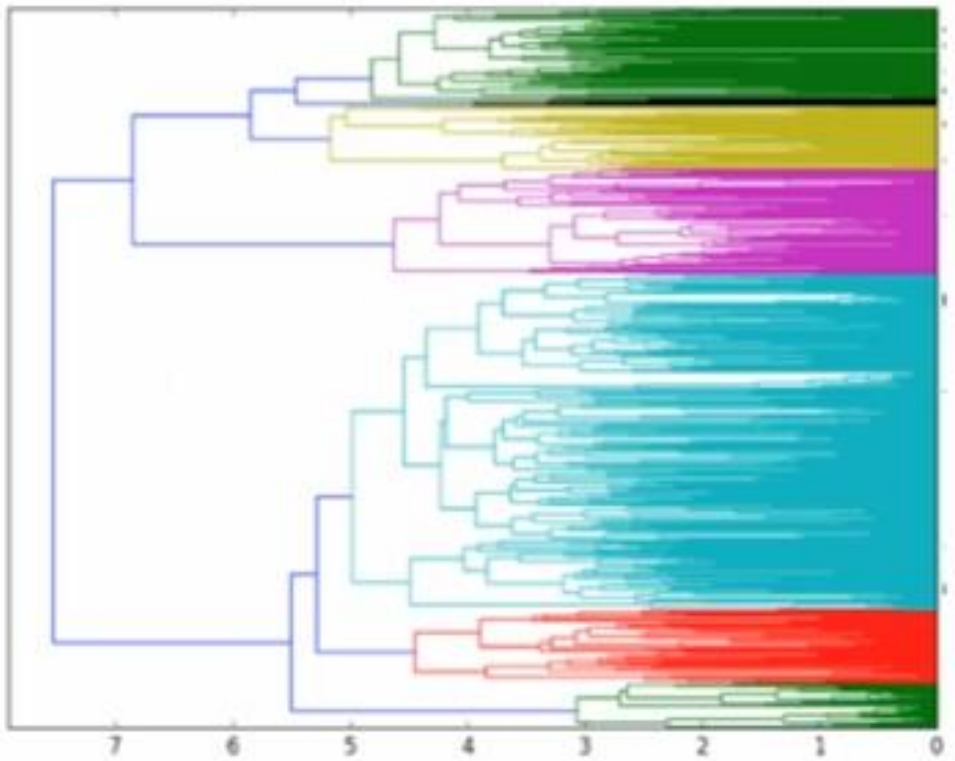




Пример: расстояние между кластерами

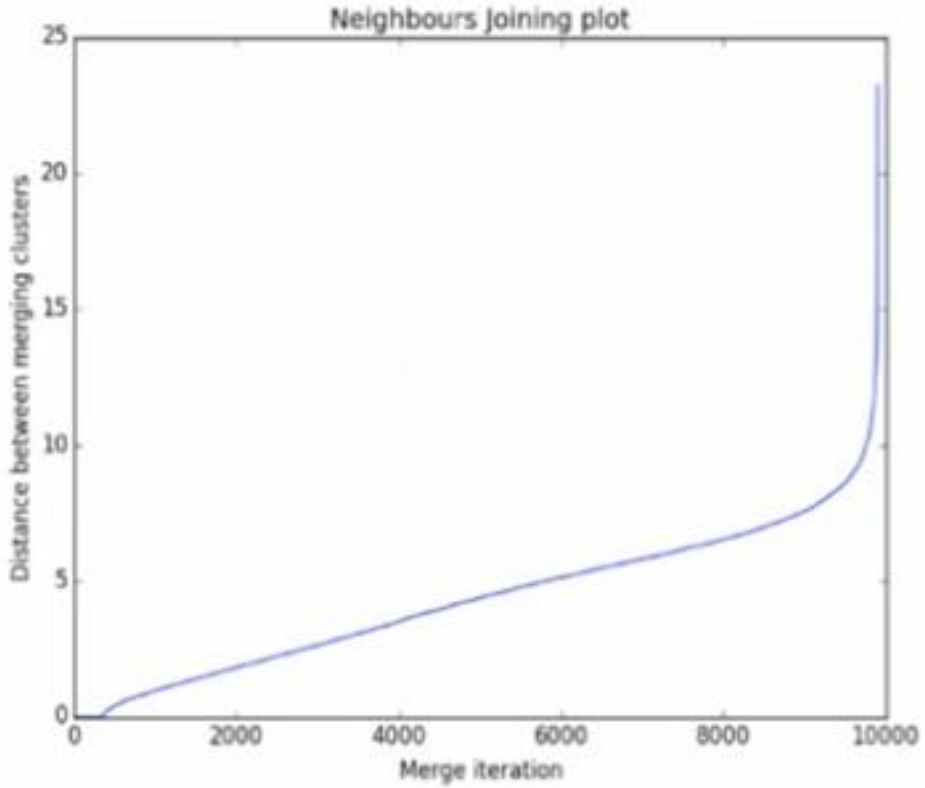
На подвыборке из 1000 писем

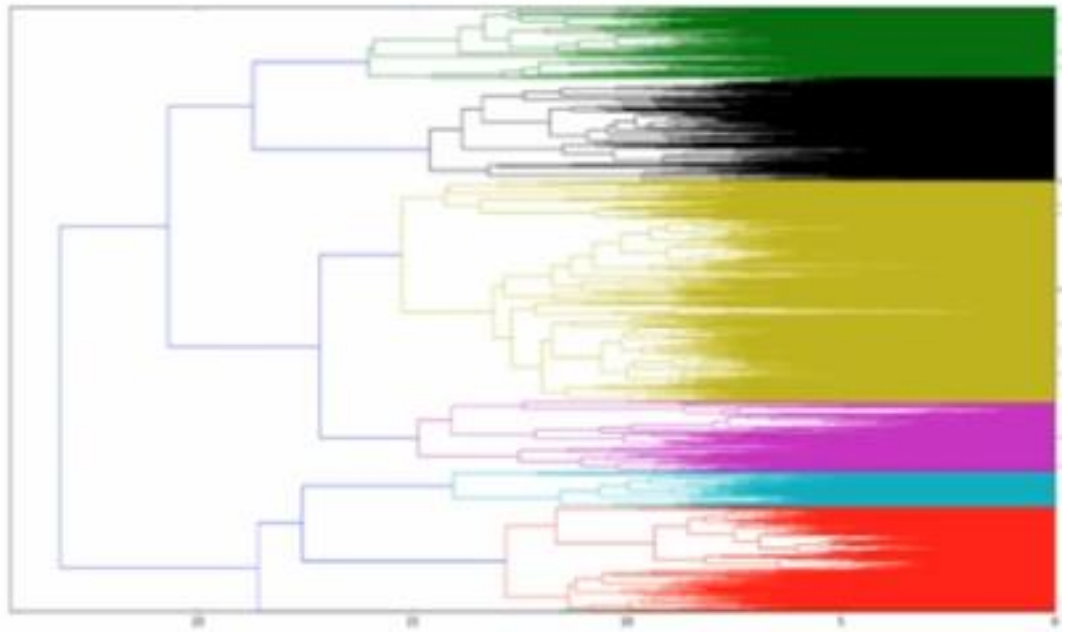




Пример: расстояние между кластерами

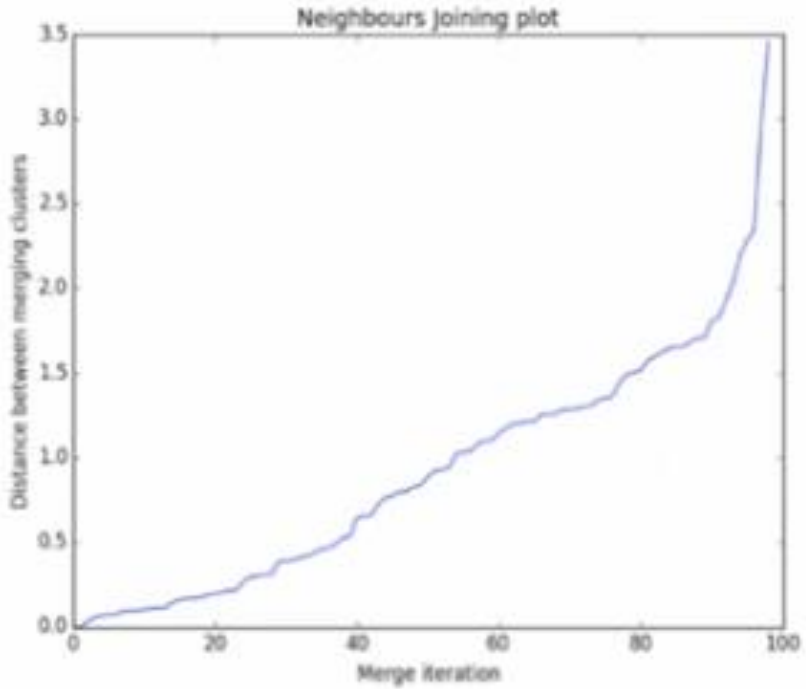
- На подвыборке из 10000 писем

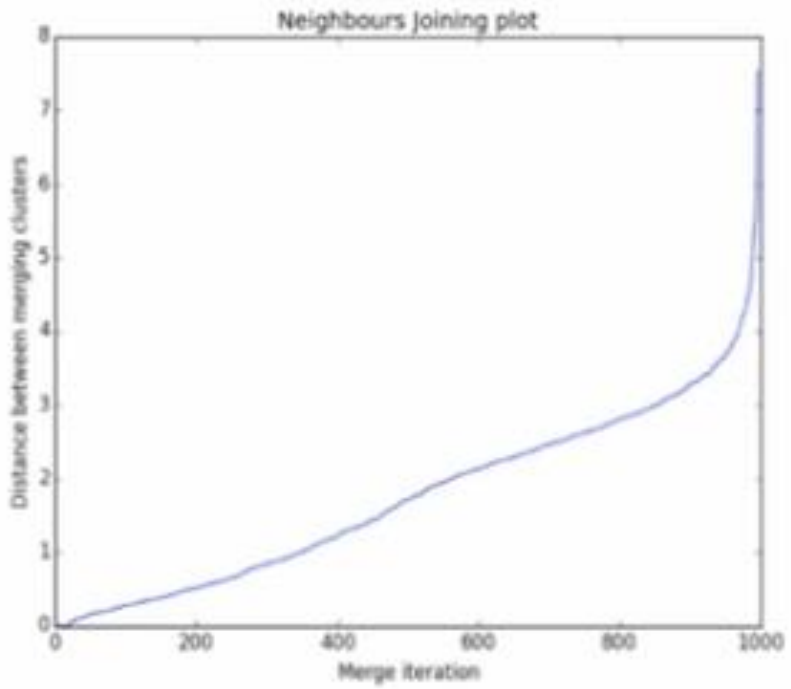


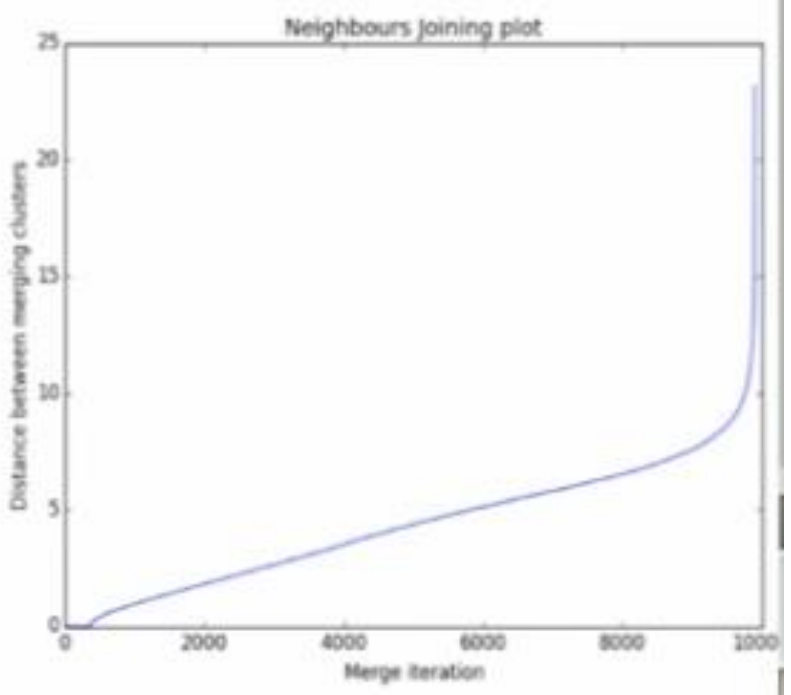


Пример: расстояние между кластерами

- Сравним графики: 100, 1000, 10000 писем







Кластеризация на основе плотности точек (Density Based Clustering) DBSCAN

Идея Density-Based методов



Figure 8.20. Center-based density.

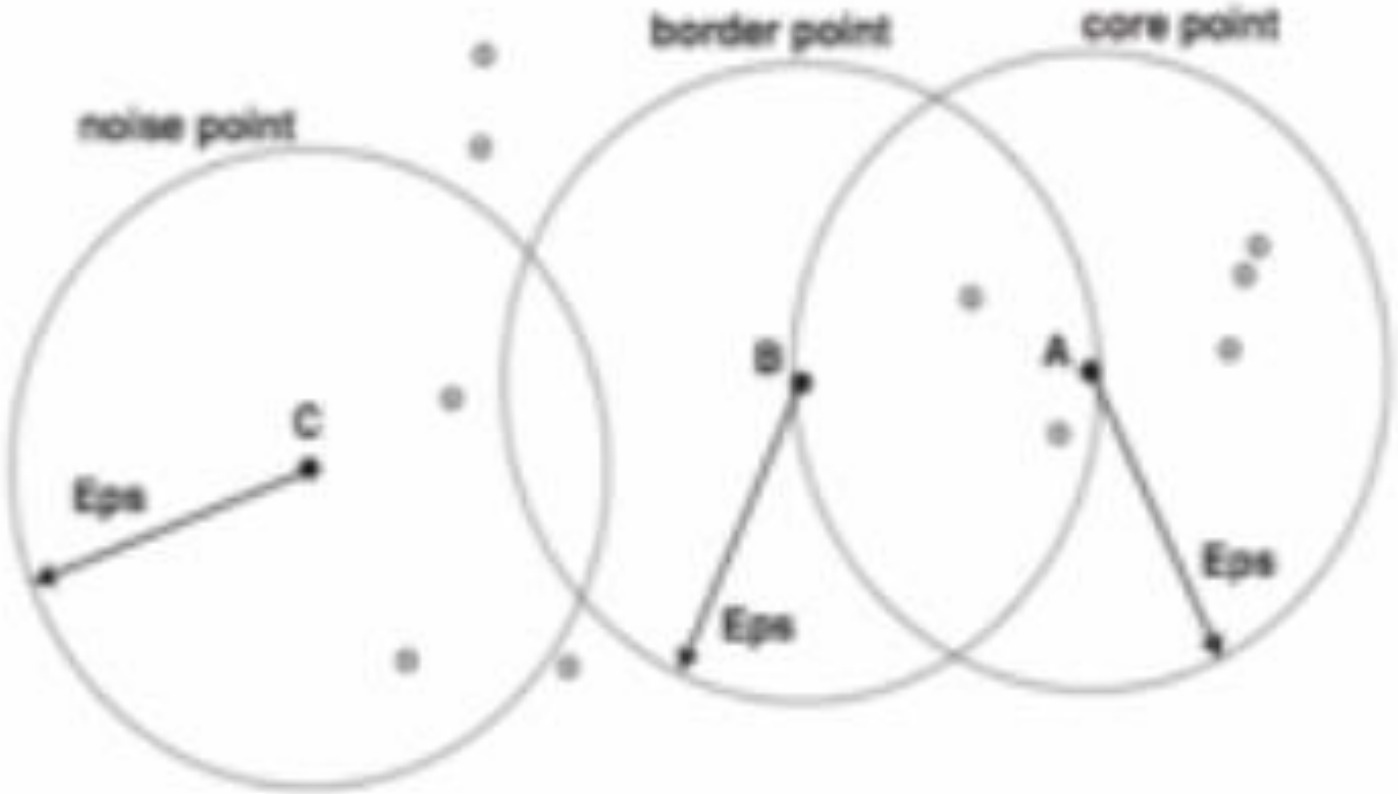
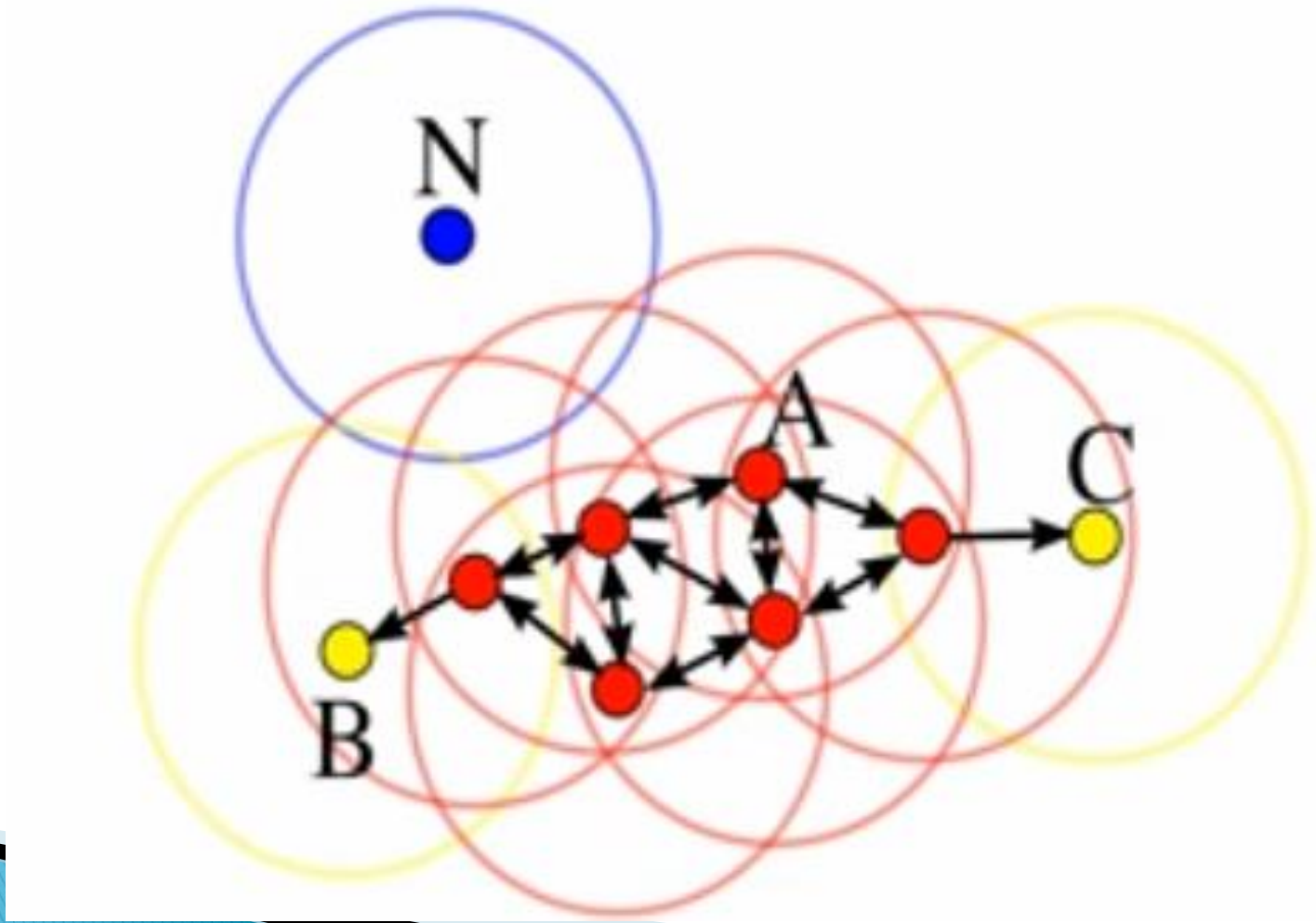


Figure 8.21. Core, border, and noise points.

Основные, шумовые и граничные точки



DBSCAN



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

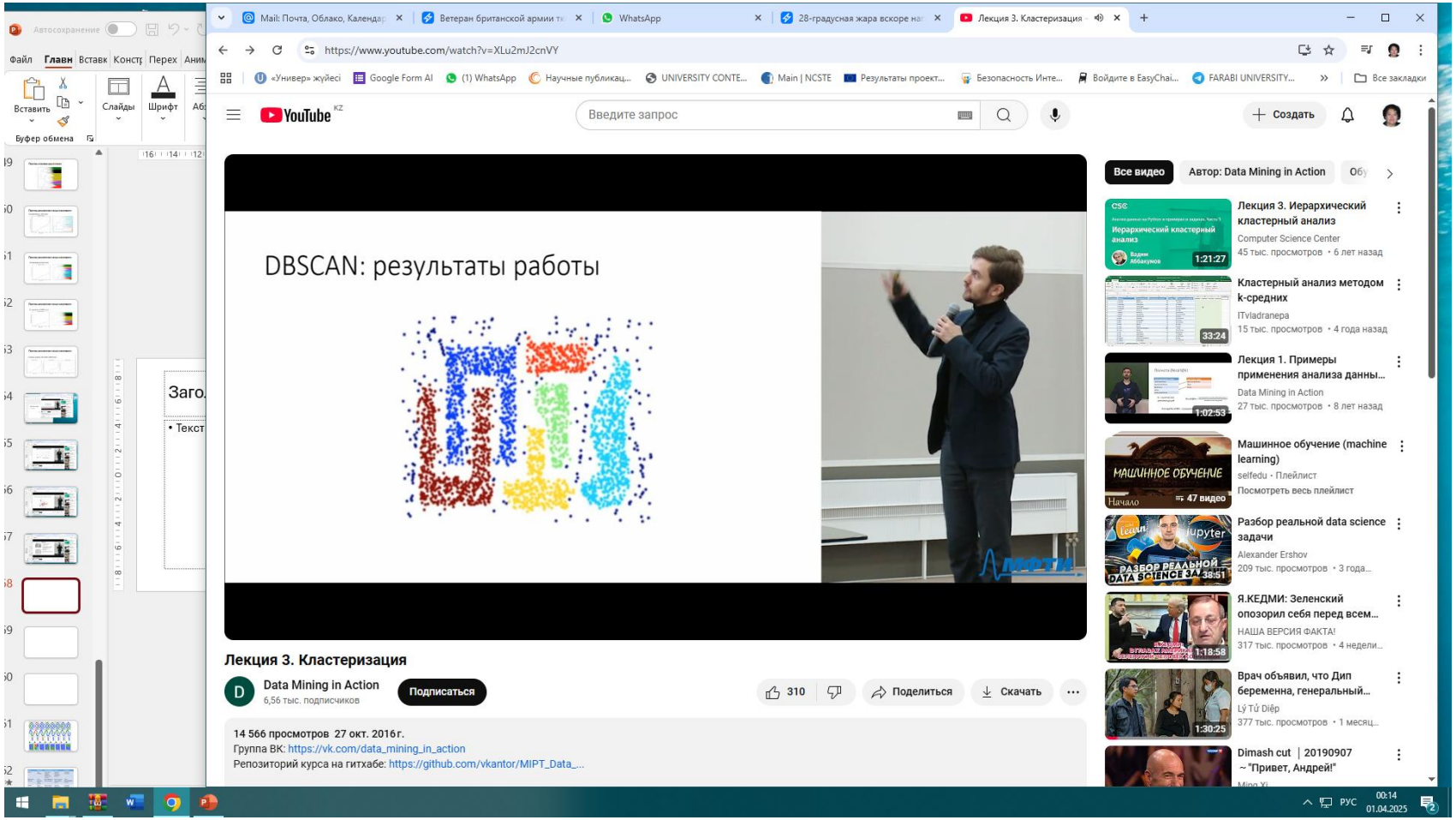
2: Отбросить точки шума.

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

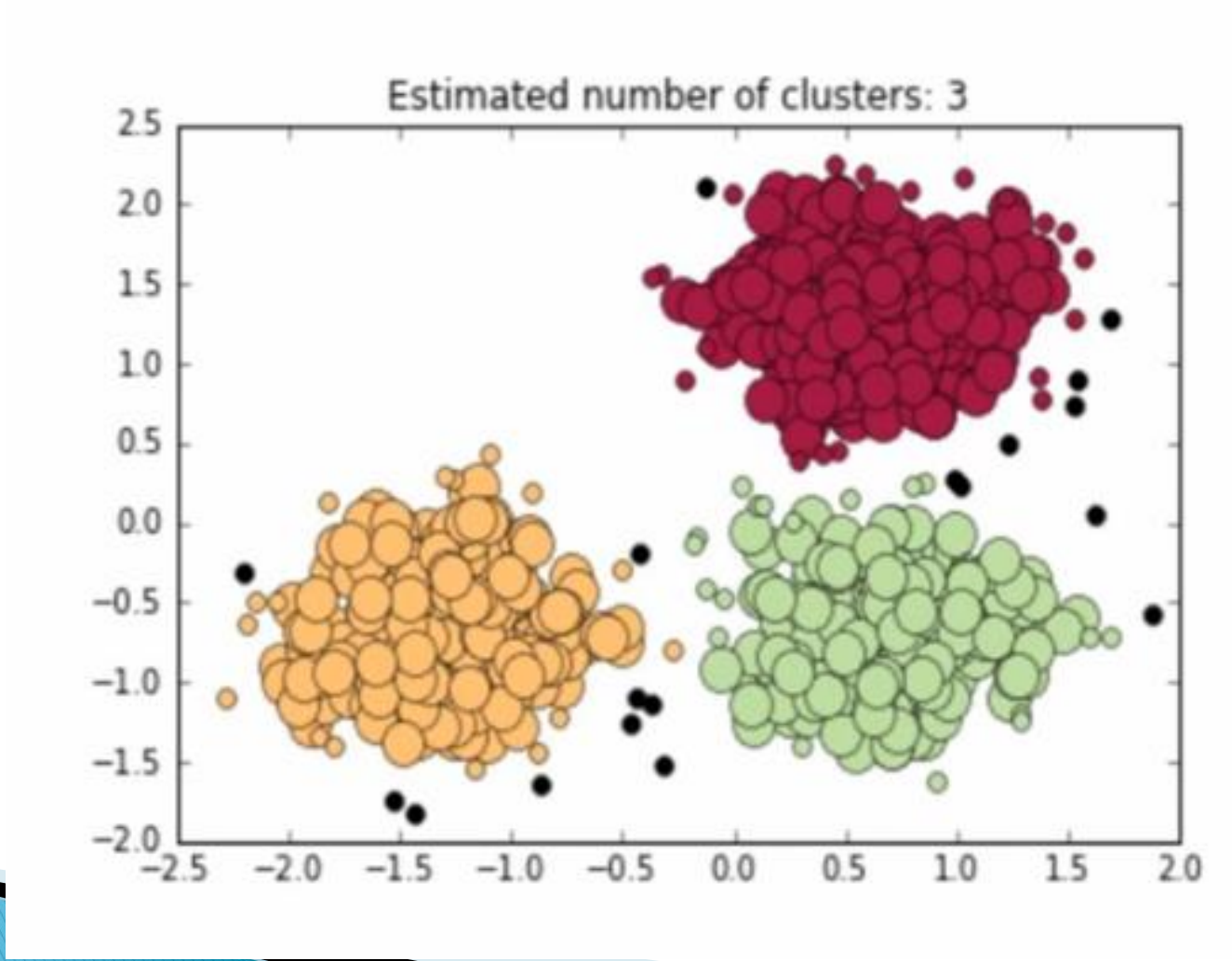
4: Объединить каждую группу соединенных основных точек в отдельный кластер.

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

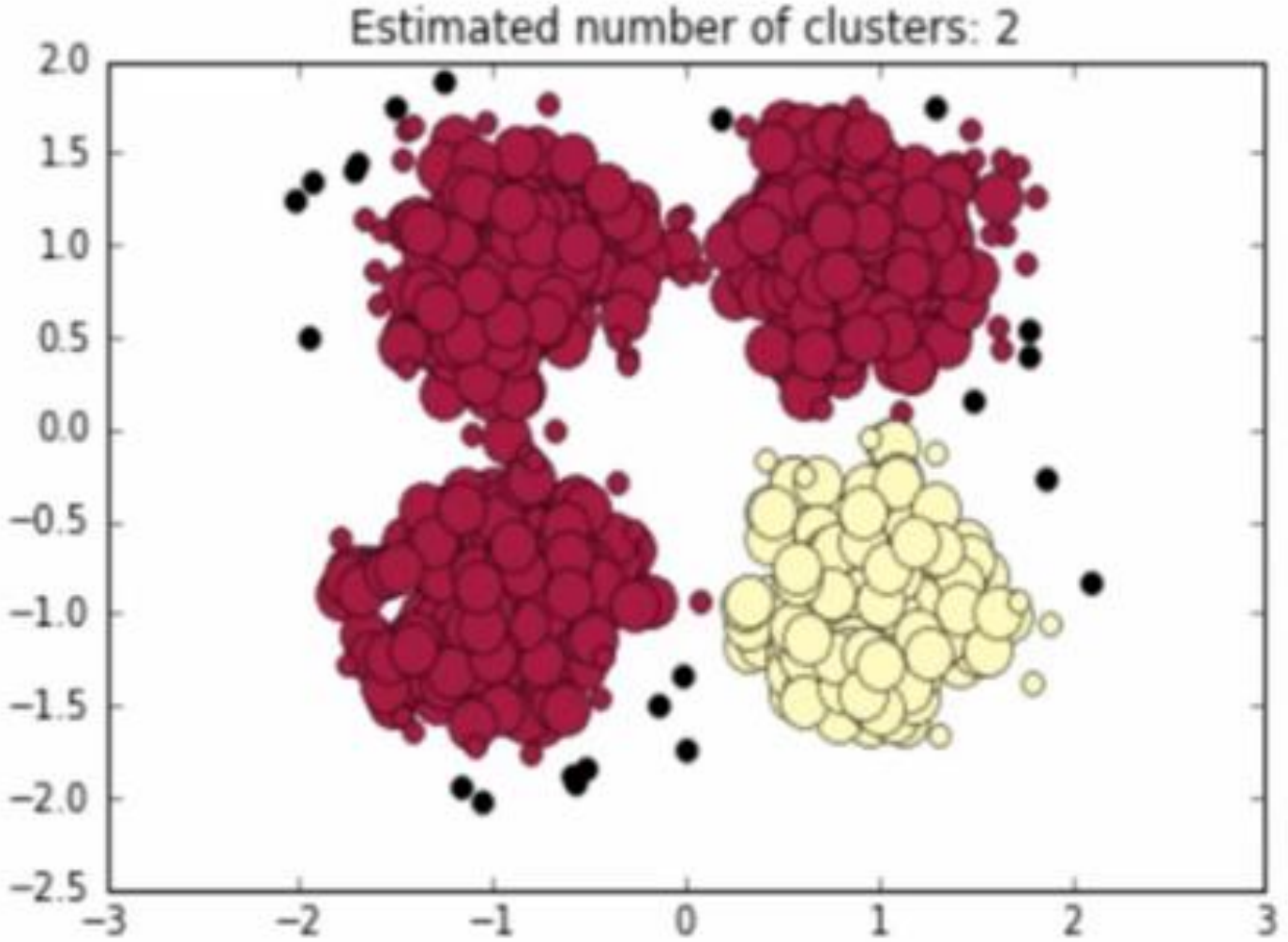
DBSCAN: результаты работы



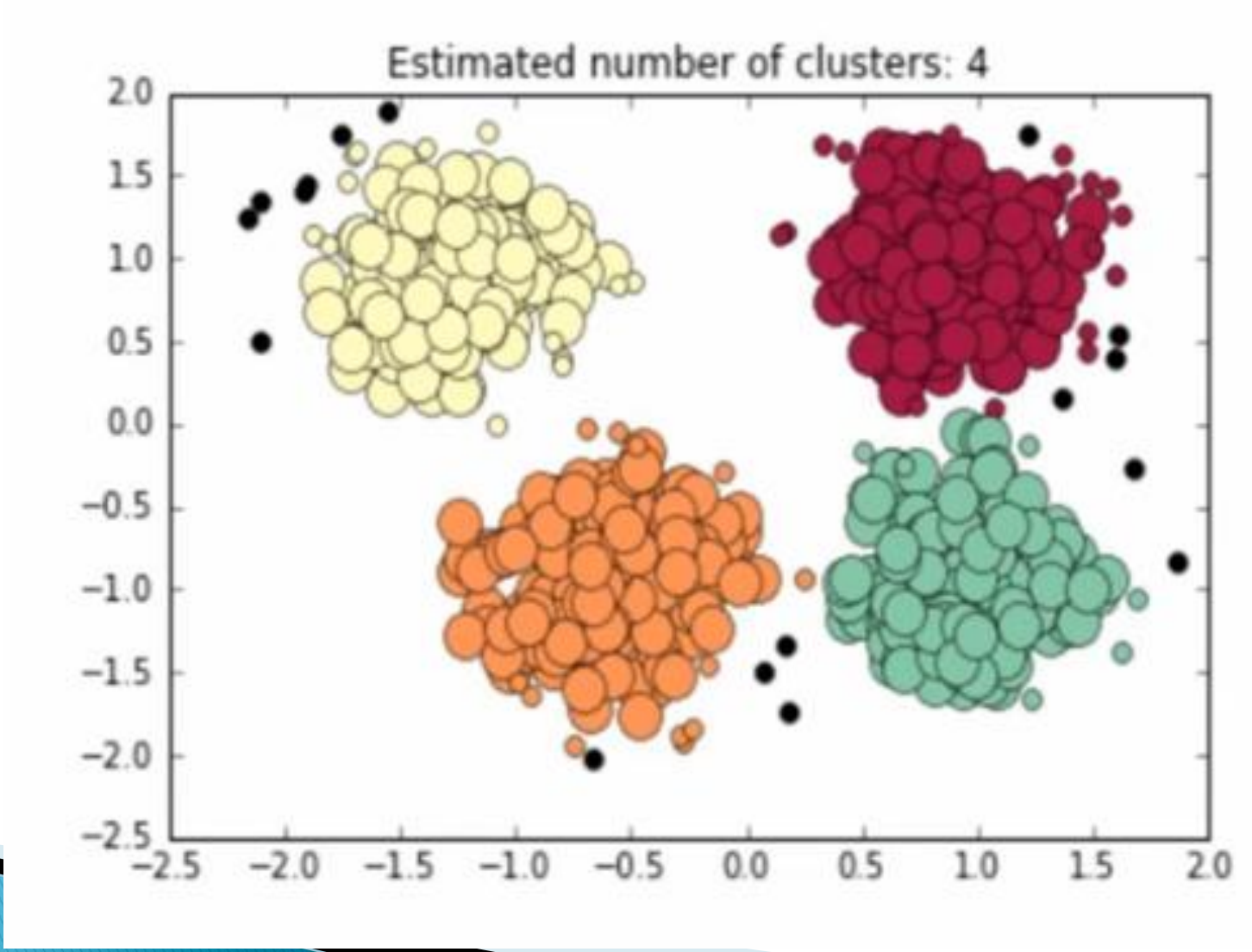
Определение числа кластеров



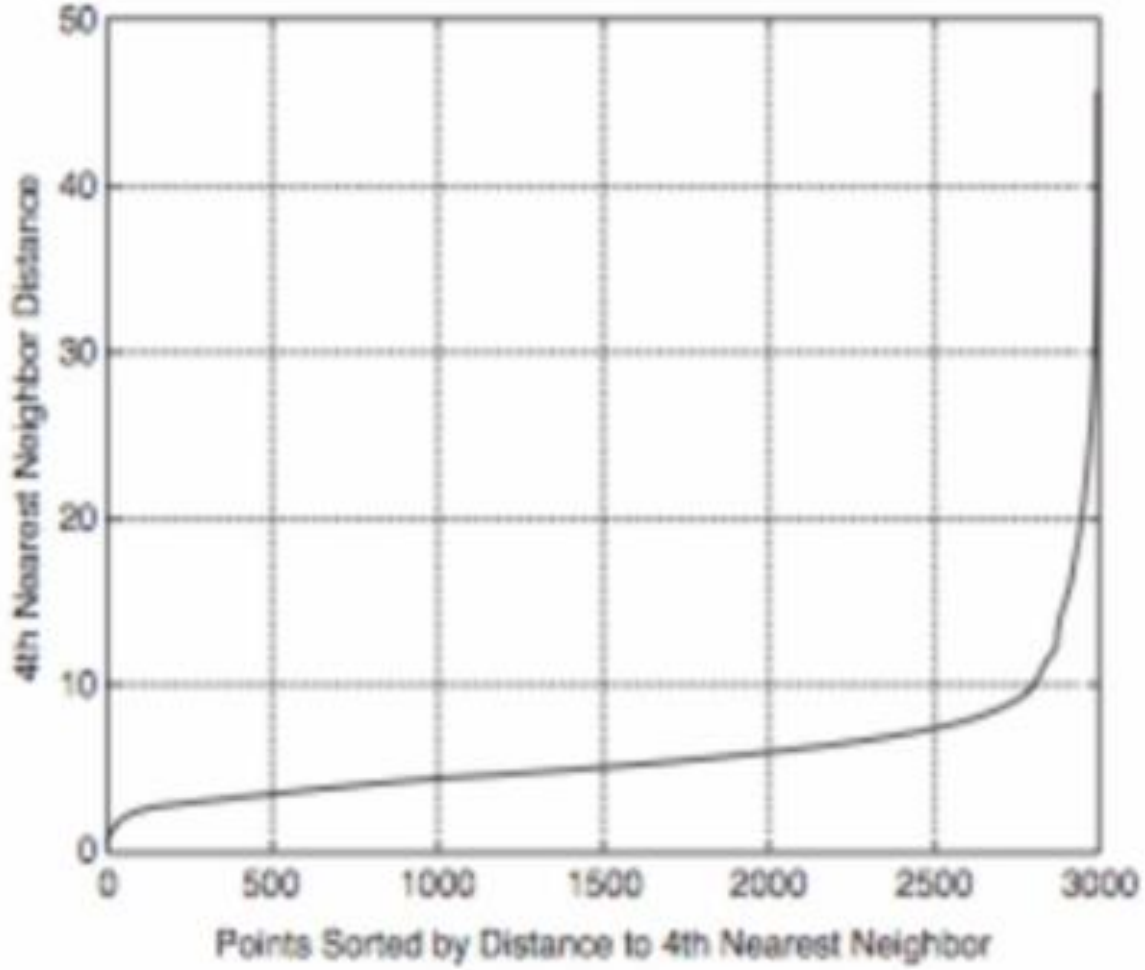
Определение числа кластеров



Определение числа кластеров



DBSCAN: подбор параметров



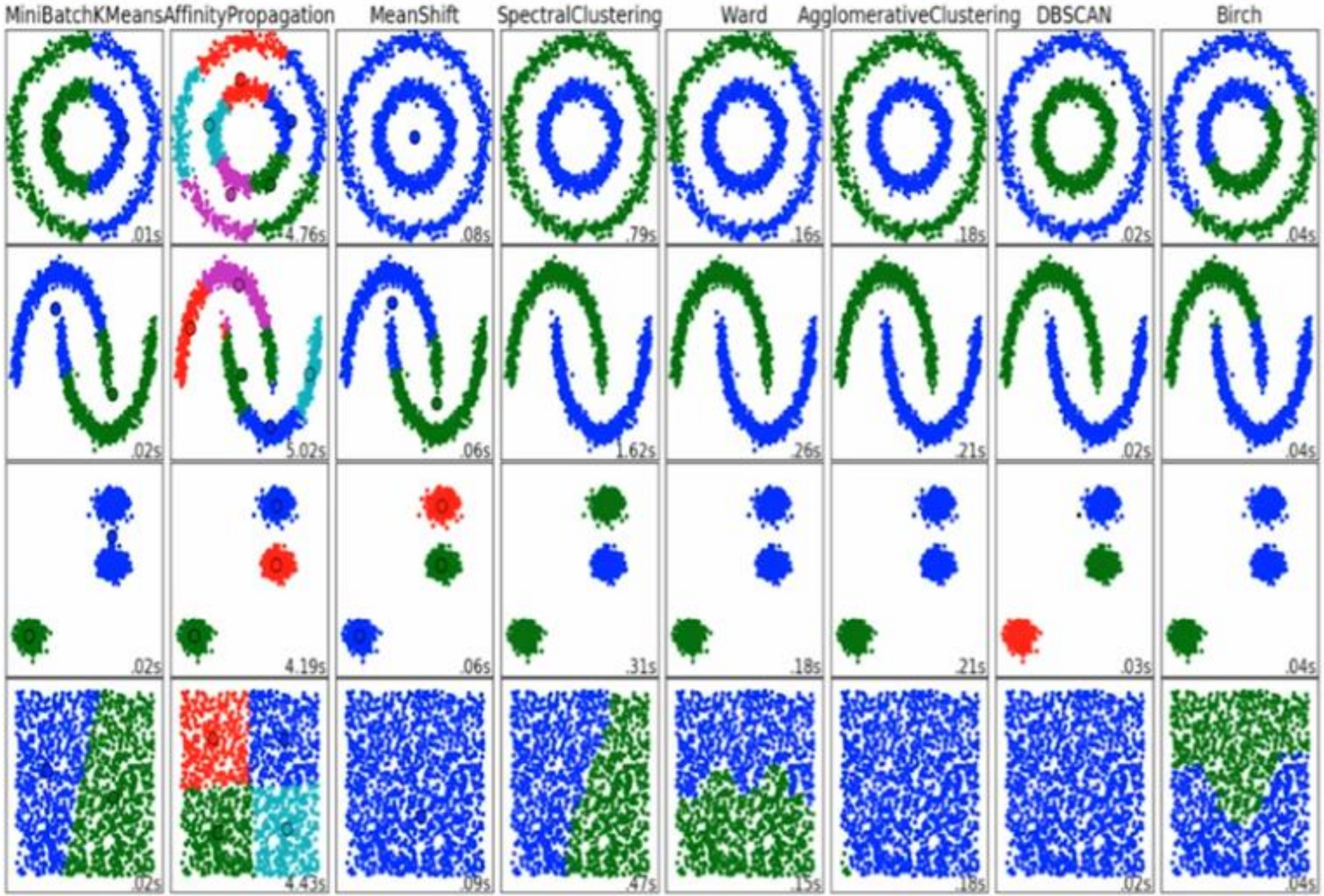
Особенности DBSCAN

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

Выбор метода кластеризации

Алгоритмы

- К – средних
- Агломеративная иерархическая кластеризация
- DBSCAN

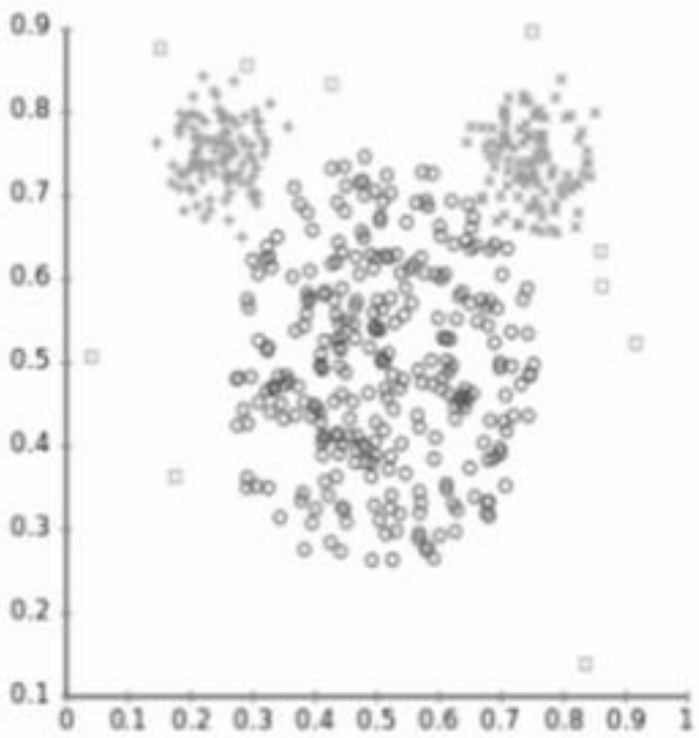

MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
4.76s
.08s
.79s
.16s
.18s
.02s
.04s
.02s
5.02s
.06s
1.62s
.26s
.21s
.02s
.04s
.02s
4.19s
.06s
.31s
.18s
.21s
.03s
.04s
.02s
4.43s
.09s
.47s
.15s
.18s
.02s
.04s

Метод	Параметры	Масштабируемость	Use-case	Геометрия
KMeans	Число кластеров	Очень много примеров (MiniBatch), среднее число кластеров	Выпуклые, примерно одинаковые кластеры	Евклидово расстояние
Agglomerative Clustering	Число кластеров, linkage, метрика	Много примеров и много кластеров	Много кластеров, нужно задавать метрику	Любая метрика/функция близости
DBSCAN	Радиус окрестности, число соседей	Много объектов, среднее число кластеров	Неравные невыпуклые кластеры, выбросы,	Евклидово расстояние

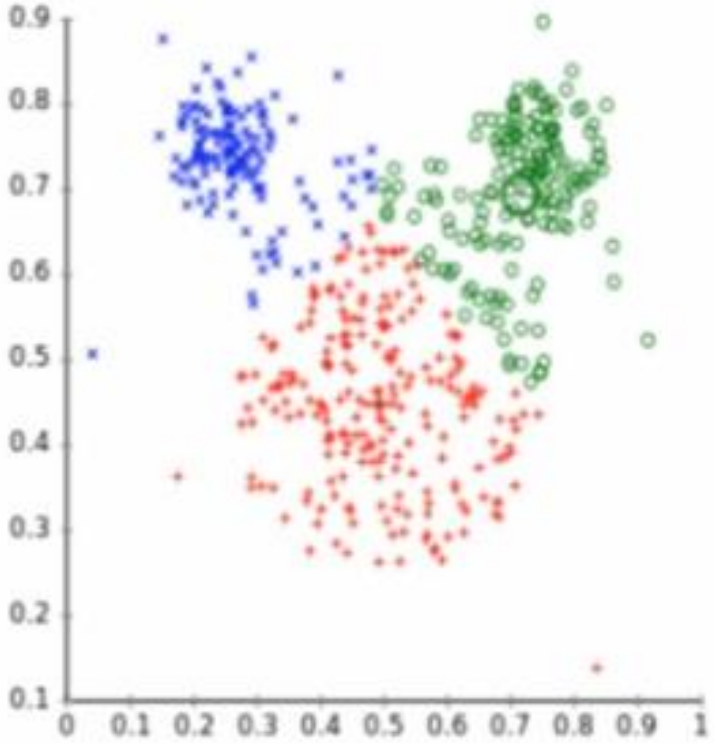
Резюме: чем могут отличаться задачи кластеризации

- Форма кластеров, которые нужно выделять
- Необходимость «вложенности» кластеров
- Размер кластеров
- Конечная задача или вспомогательная
- Жесткая или мягкая кластеризация

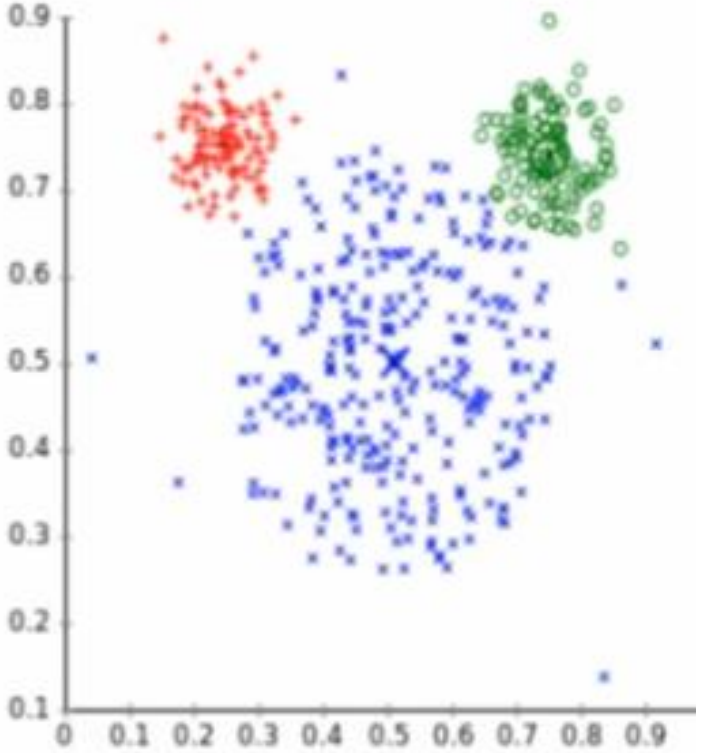
Различия в результатах работы методов



Исходная выборка
("Mouse" dataset)



Метод k средних
(K-Means)



EM-алгоритм



Спасибо за внимание!