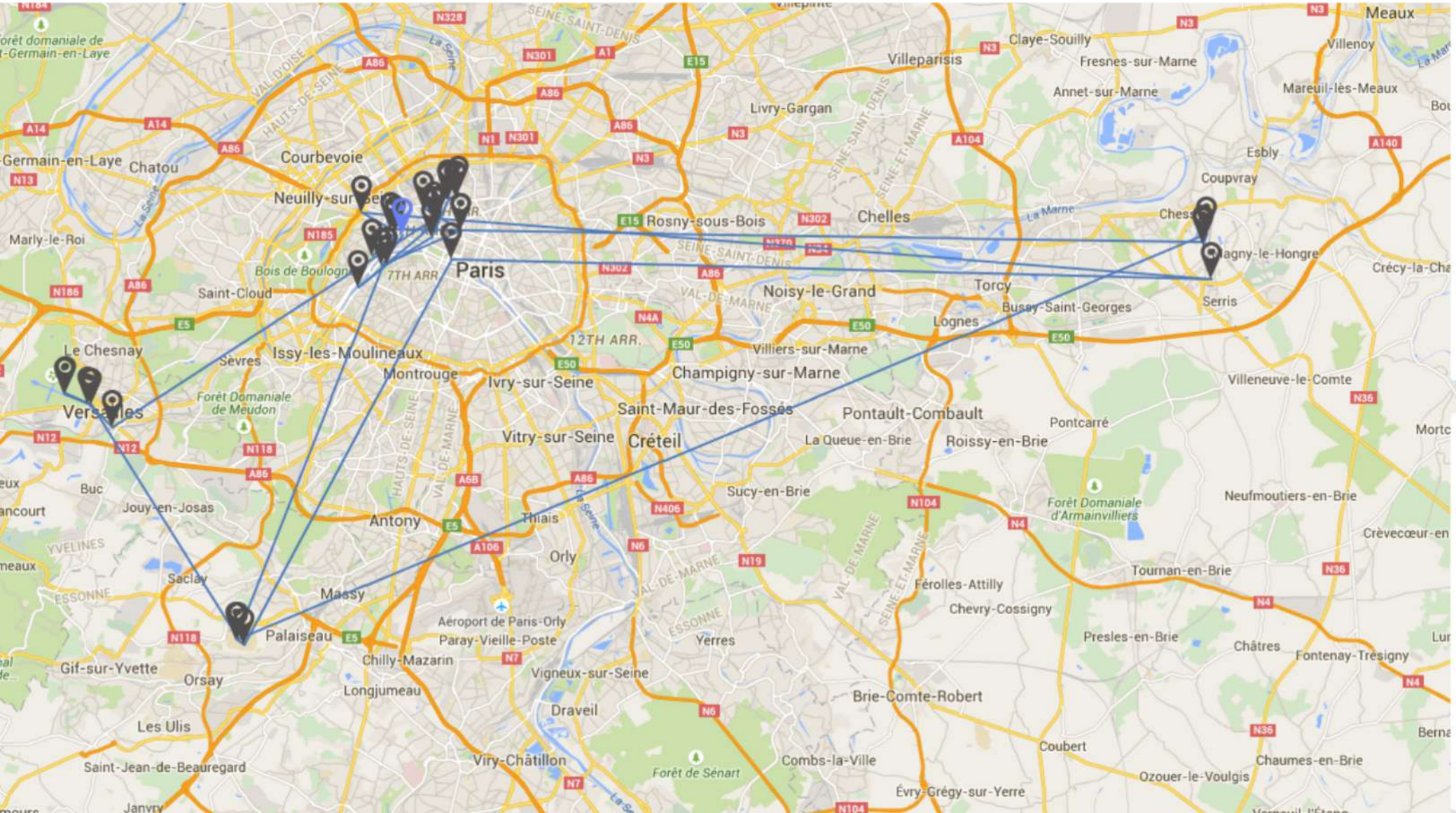
Машинное обучение

Лекция 11
Кластеризация

Андрей Нарцев
andrei.nartsev@gmail.com
anartsev@hse.ru

НИУ ВШЭ, 2025

Пример: анализ геоданных



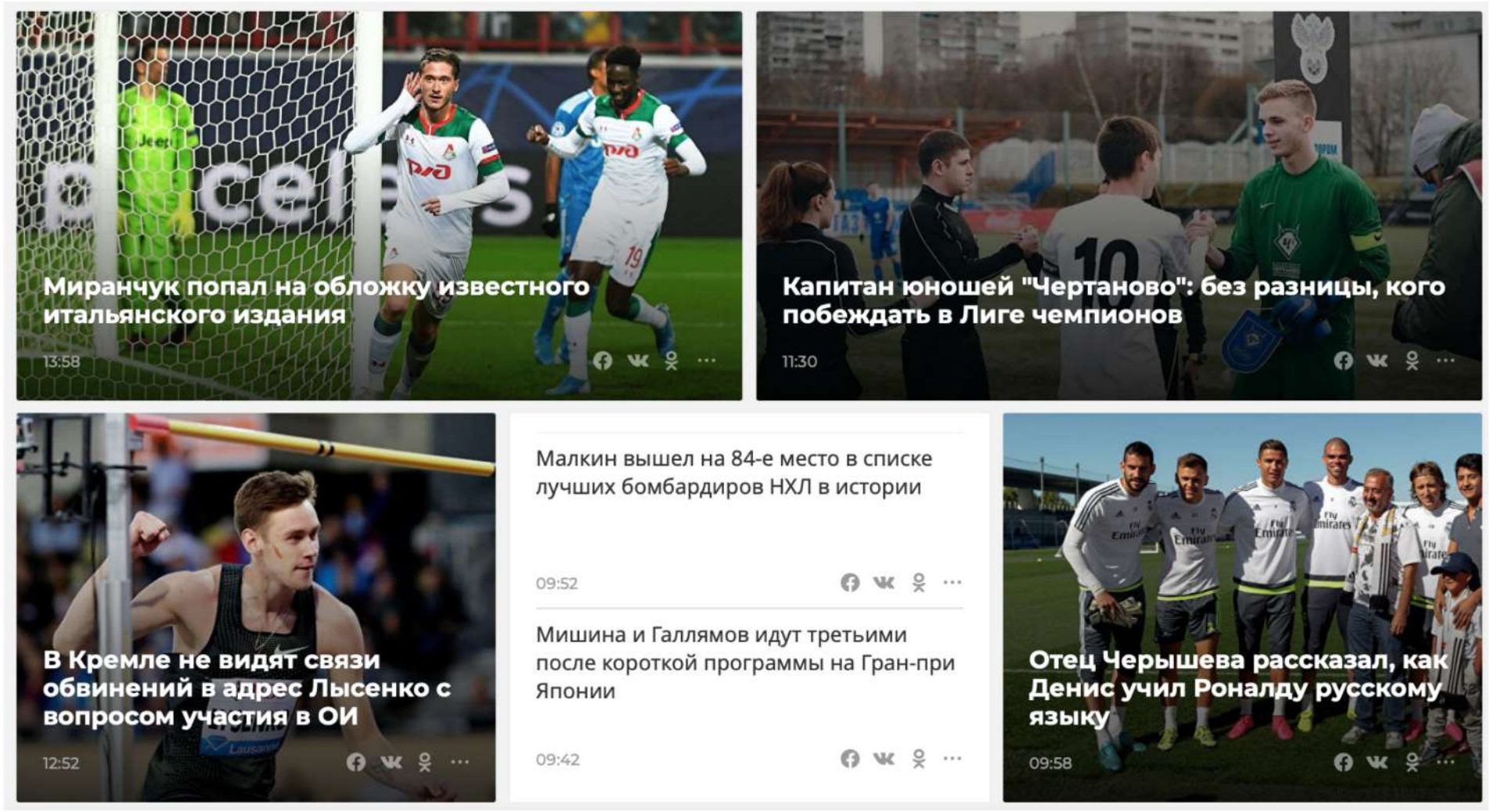
Еле увел ее оттуда ...

4 мая 2014 г. 18:33:24

Bikes ❤

4 мая 2014 г. 20:36:44

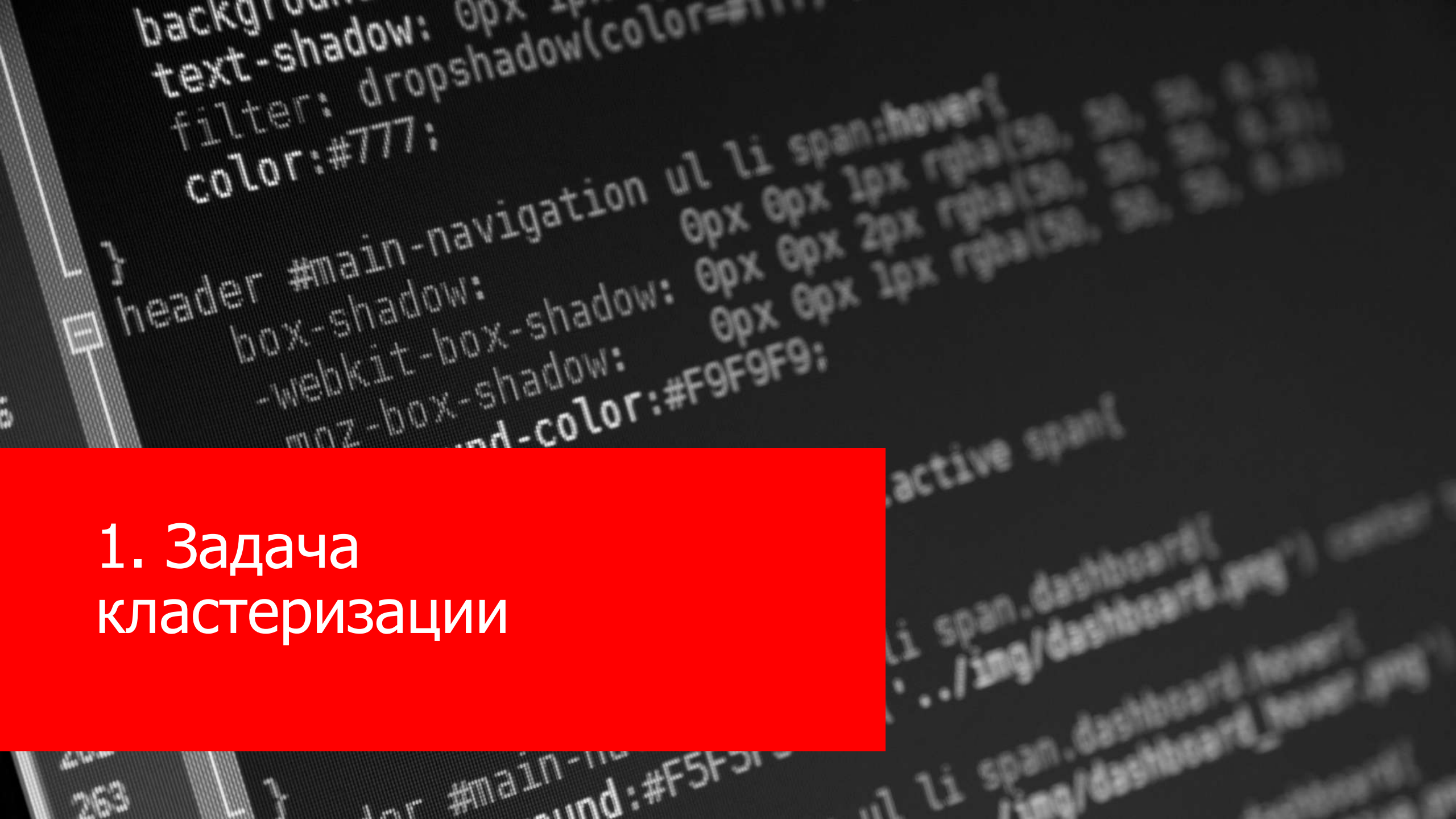
Пример: кластеризация текстов по теме



Скриншот с сайта РИА Новости (ria.ru)

Сегодня на лекции

1. Задача кластеризации
2. Основные методы
3. Особенности применения и выбора
4. Подробнее об алгоритмах
5. Оценка качества
6. Пример: кластеризация текстов



1. Задача кластеризации

Ранее: обучение на размеченных данных (supervised learning)

Обучающая выборка:

$x_1, \ldots, x_l$ - объекты

$y_1, \ldots, y_l$ - ответы

Ранее: обучение на размеченных данных (supervised learning)

Обучающая выборка:

$x_1, \dots, x_l$ - объекты

$y_1, \dots, y_l$ - ответы

Тестовая выборка:

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

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 $a(x)$:

$$a(x) \approx y$$

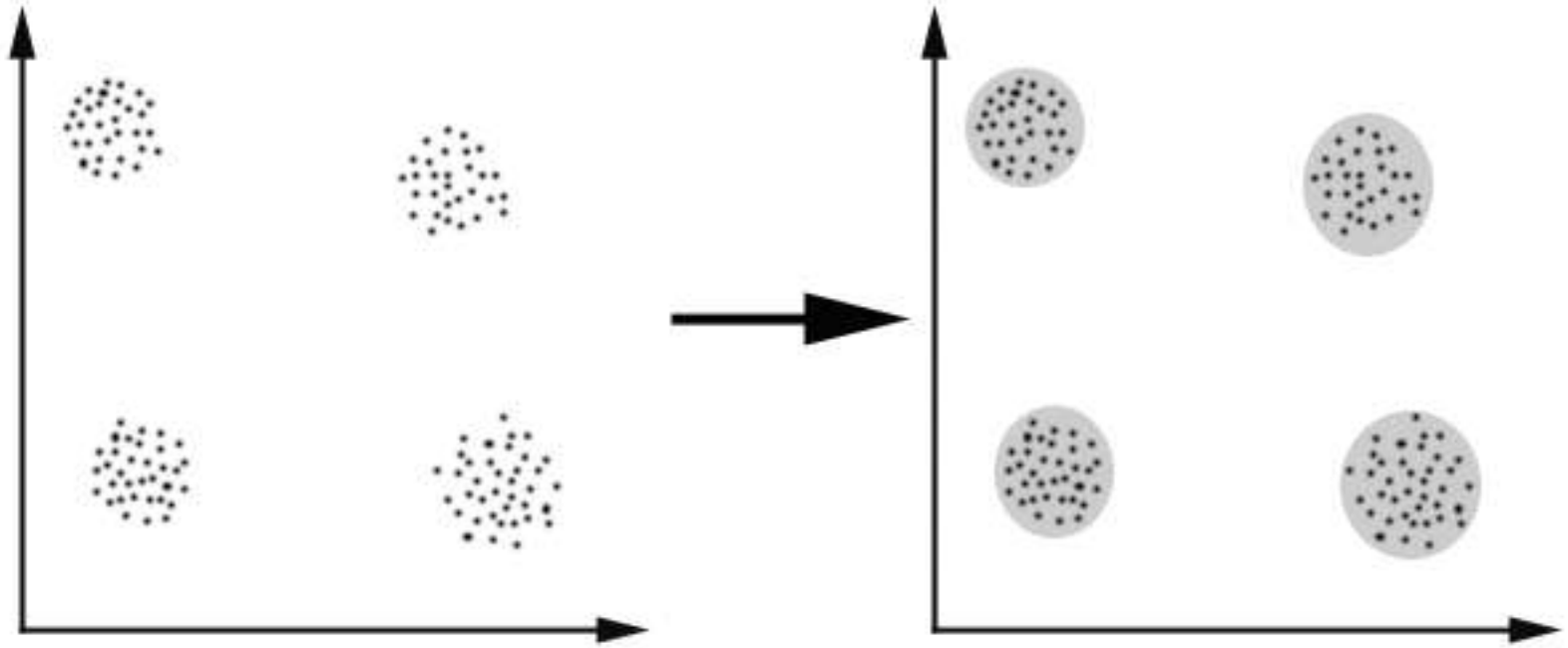
Кластеризация

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

Она же и тестовая

Нужно поставить метки $y_1, \ldots, y_l$, 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

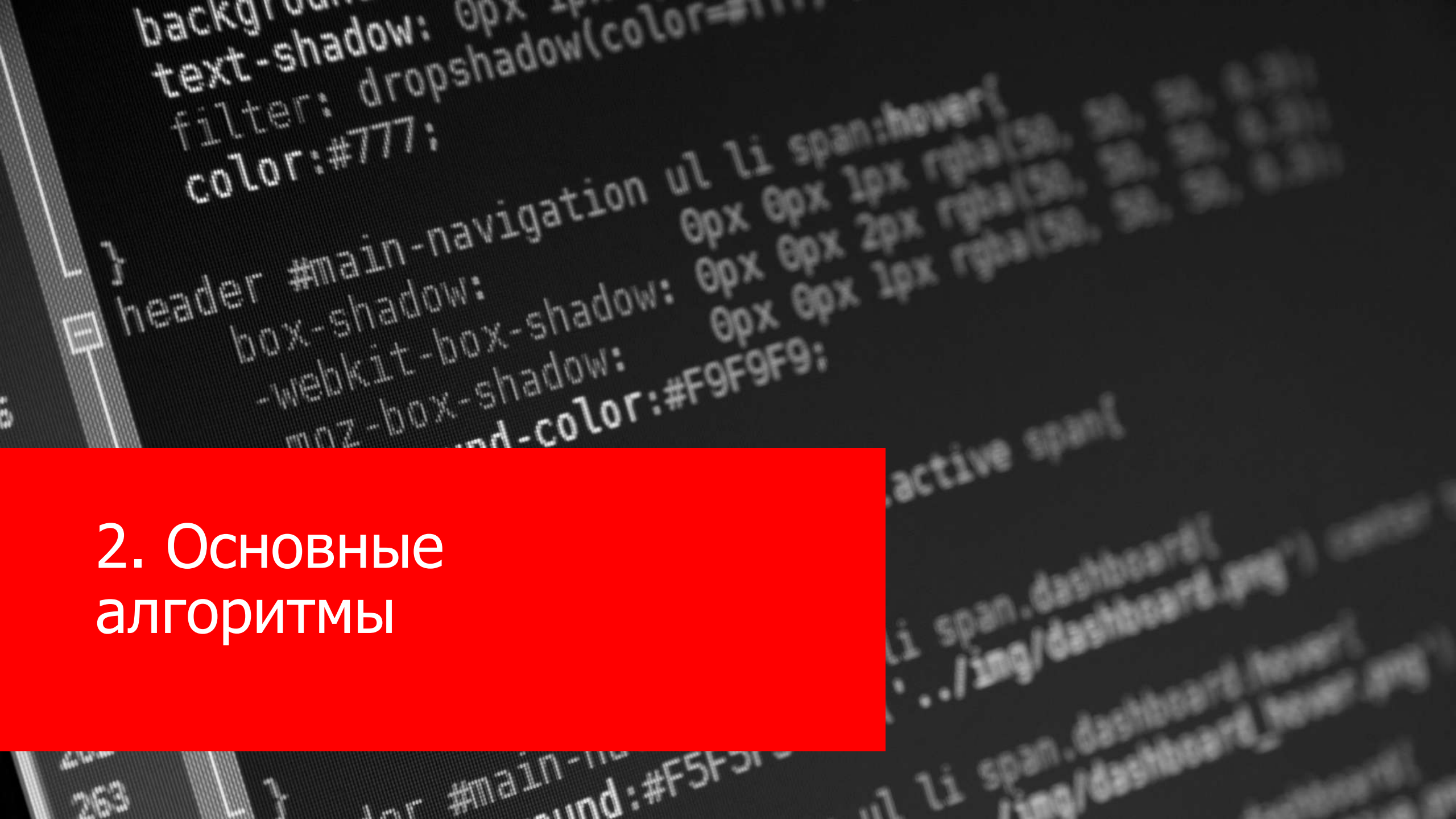
$$F_1 = \frac{\sum\limits_{i<j}[y_i \neq y_j]\,\rho(x_i, x_j)}{\sum\limits_{i<j}[y_i \neq y_j]} \to \max$$

Придумываем метрику качества

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

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

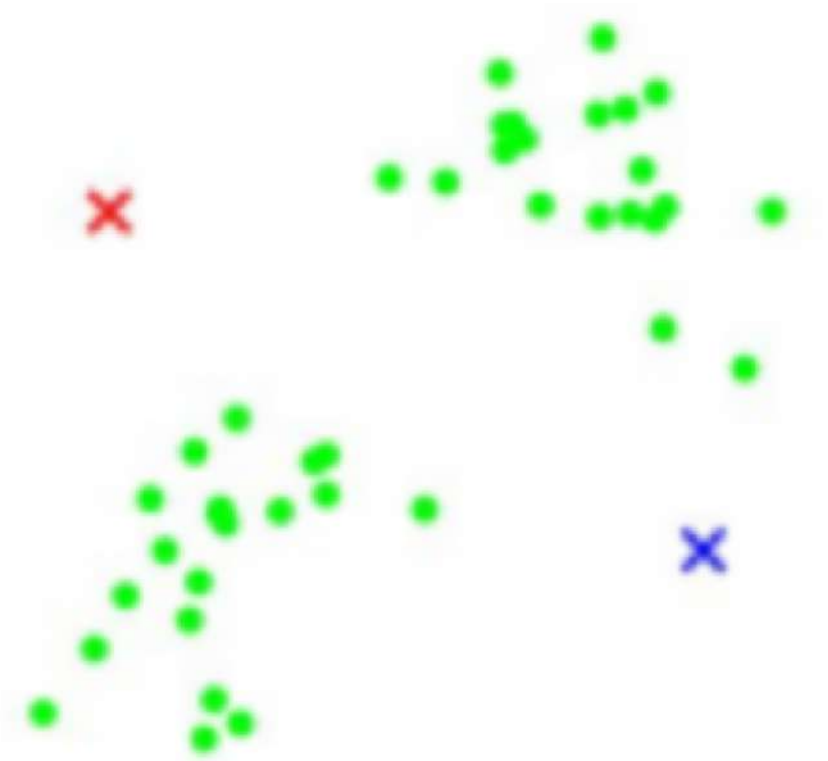
$$F_0 / F_1 \to \min$$



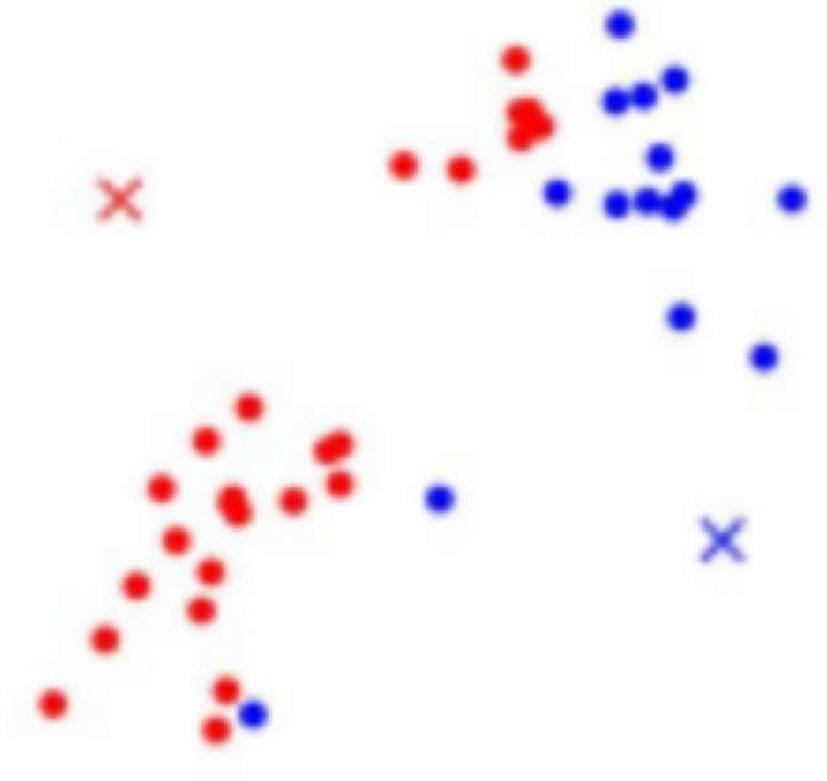
2. Основные алгоритмы

Напоминание: K Means

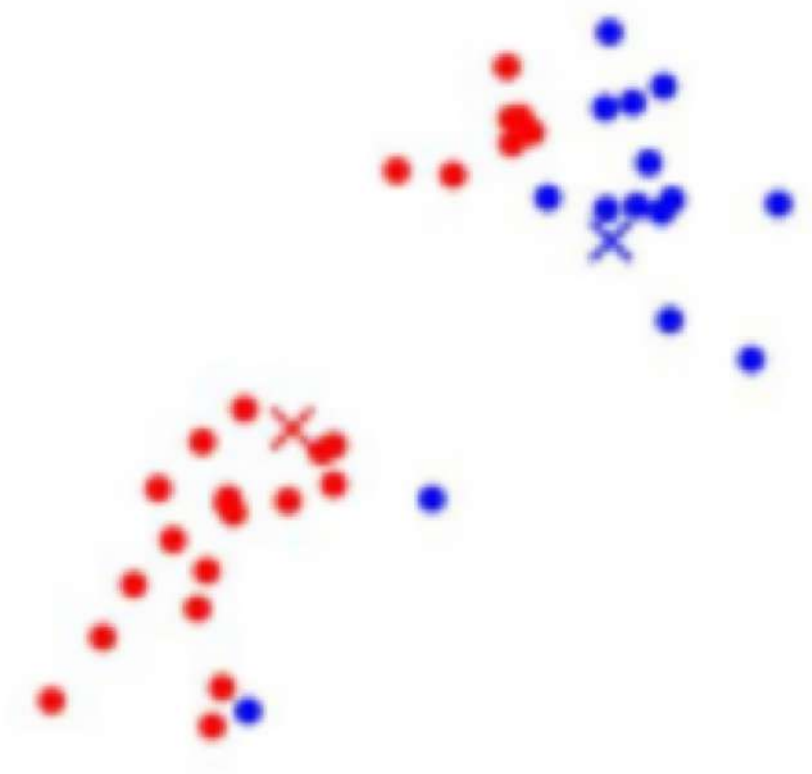
Как работает K Means



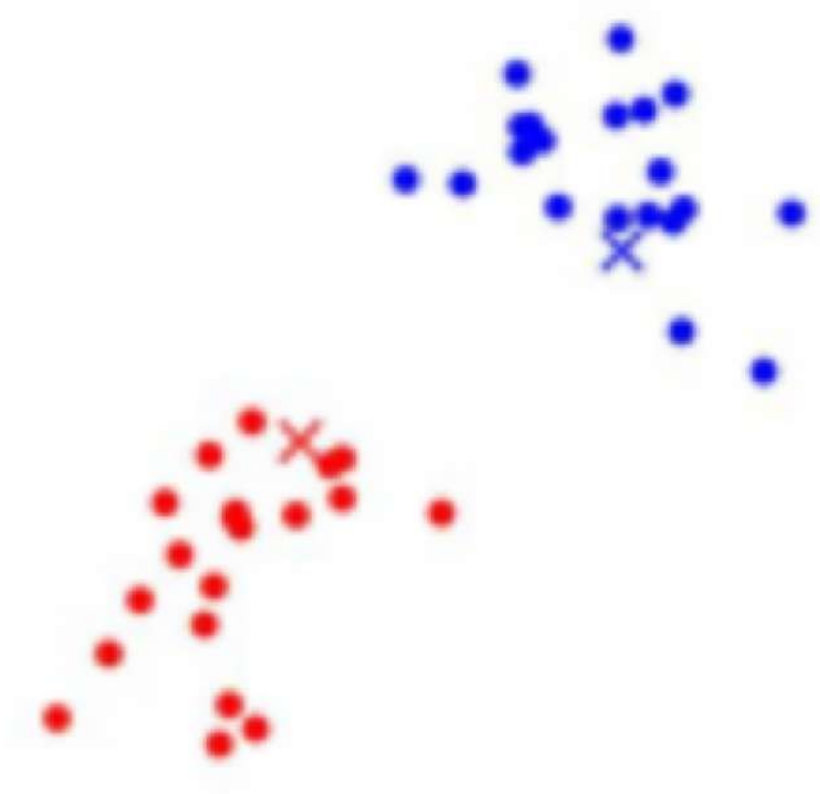
Как работает K Means



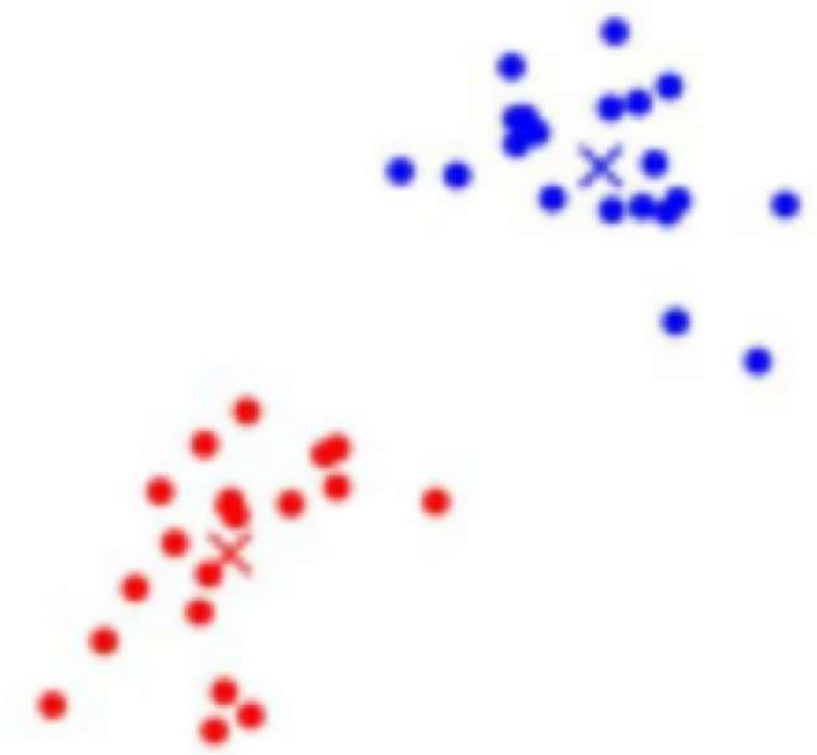
Как работает K Means



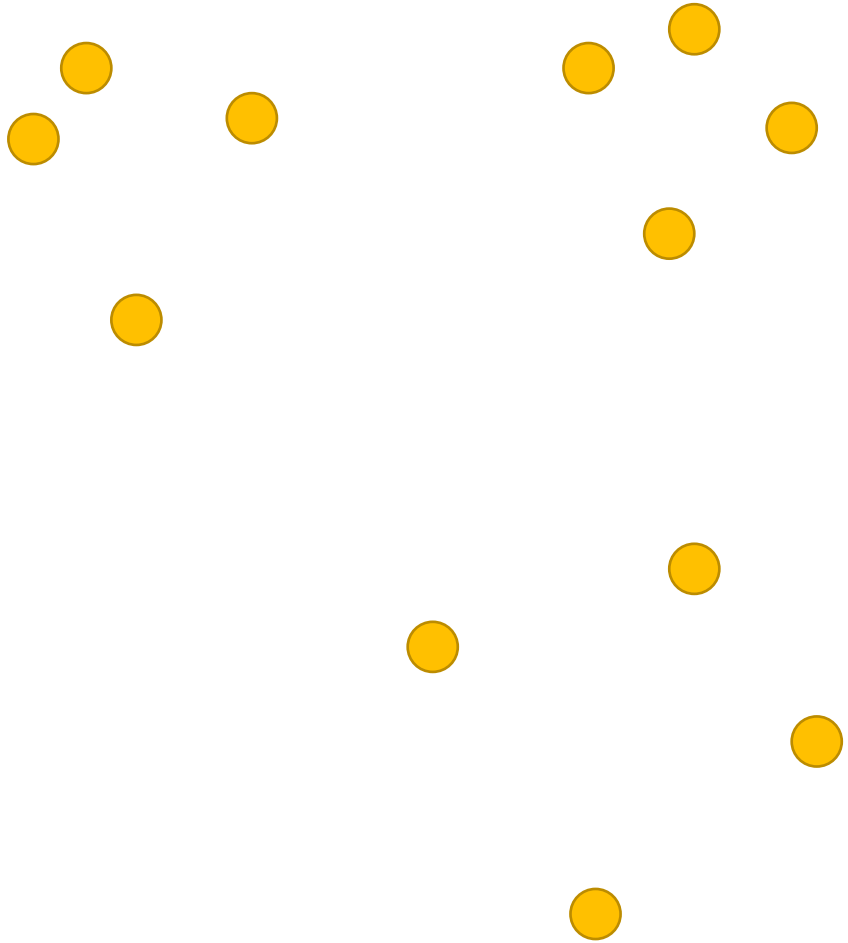
Как работает K Means



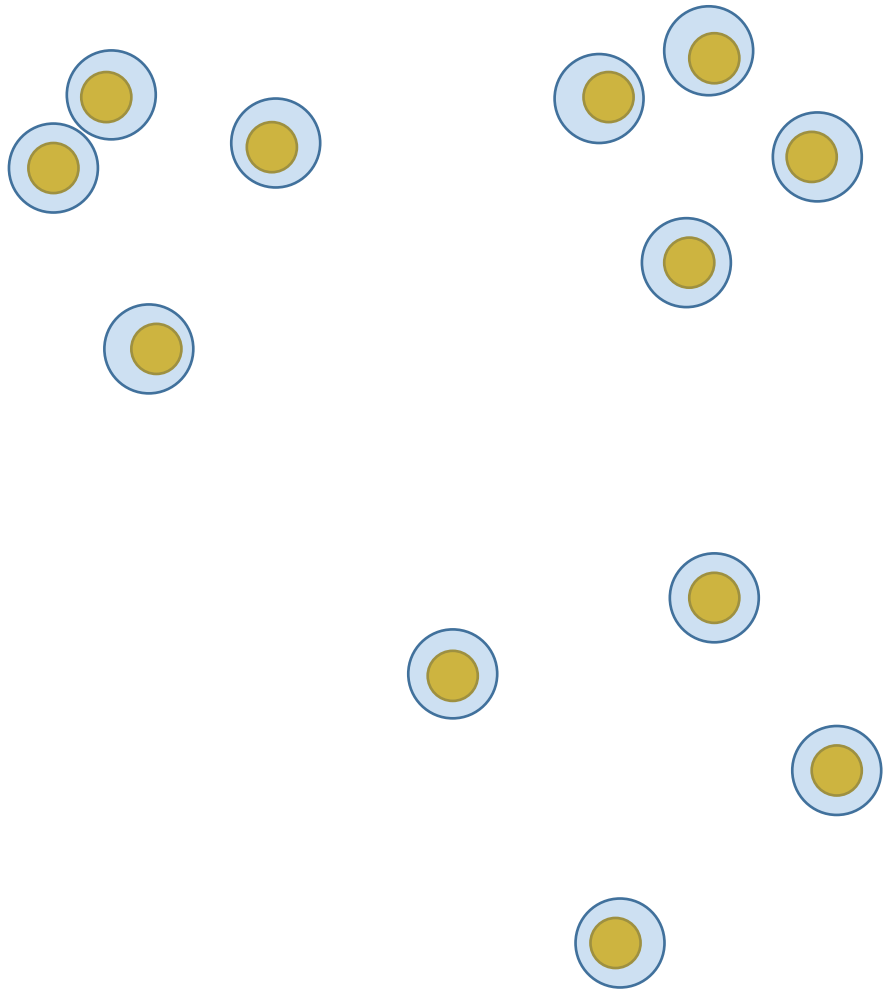
Как работает K Means



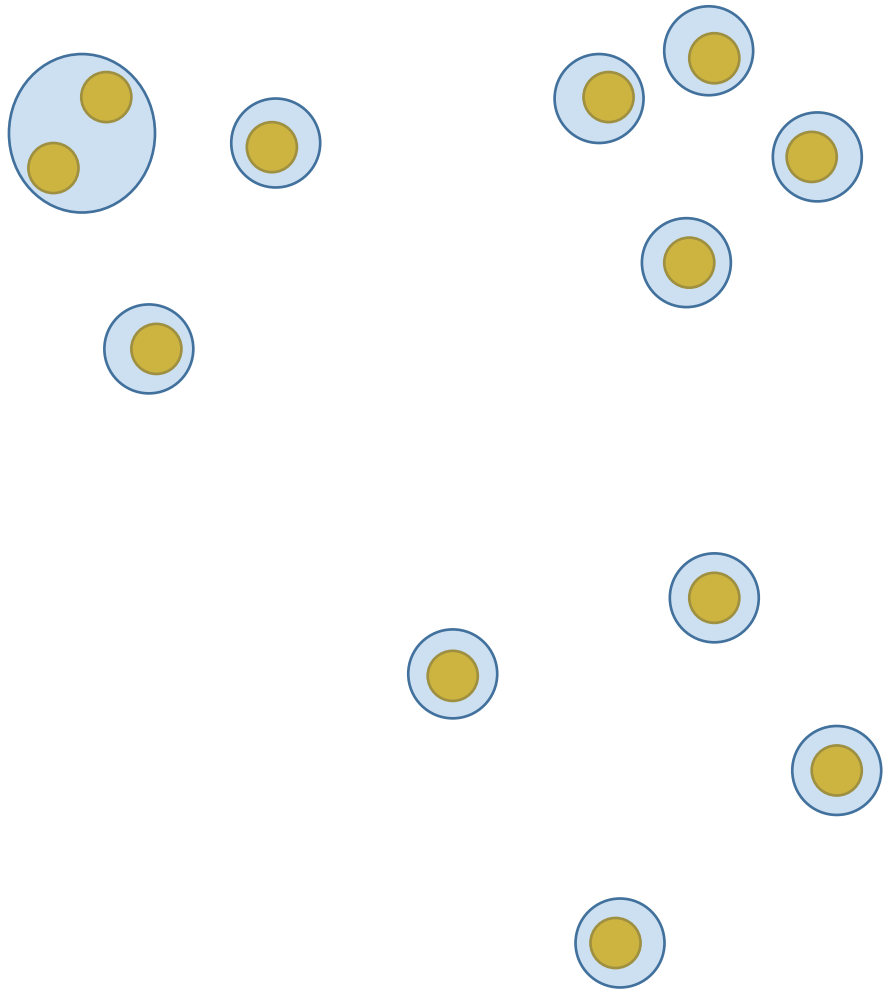
Агломеративная кластеризация



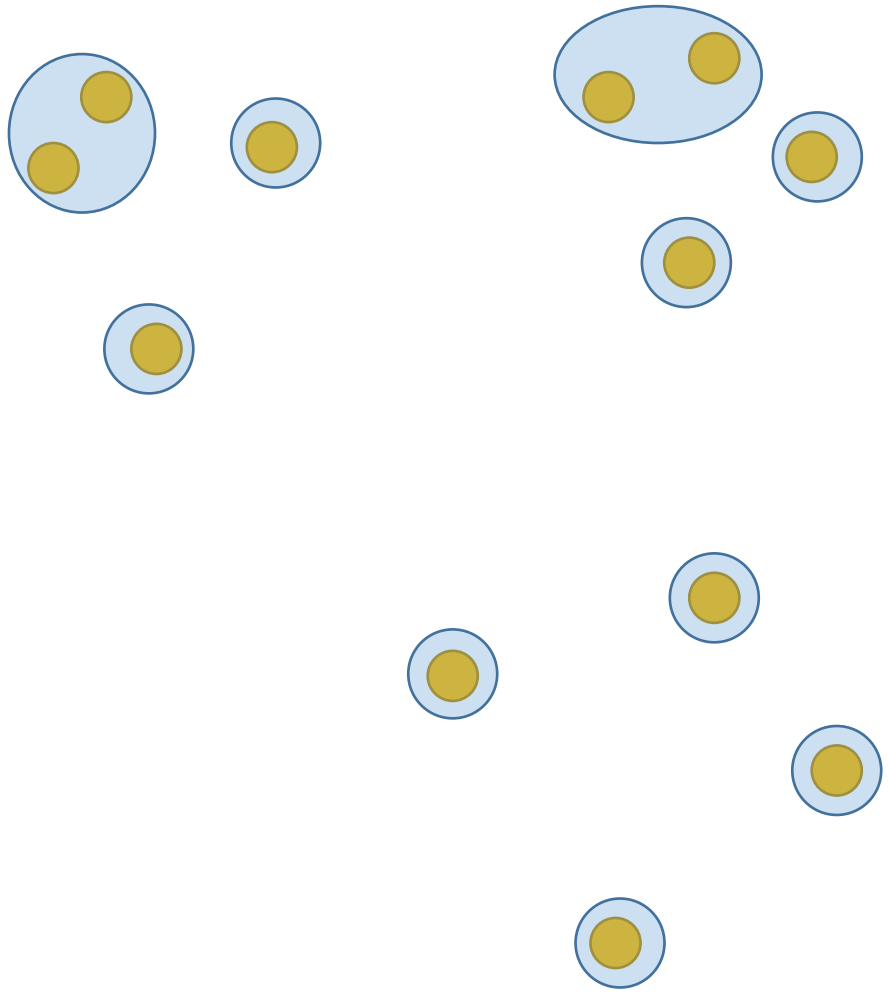
Агломеративная кластеризация



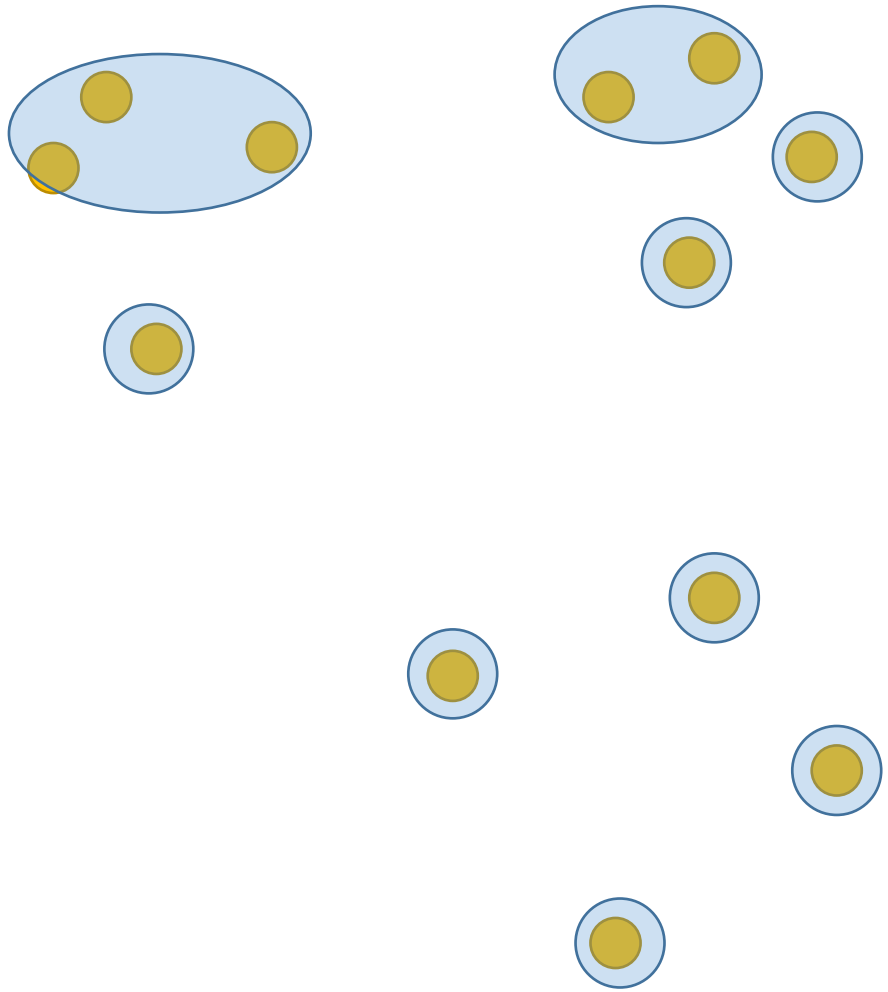
Агломеративная кластеризация



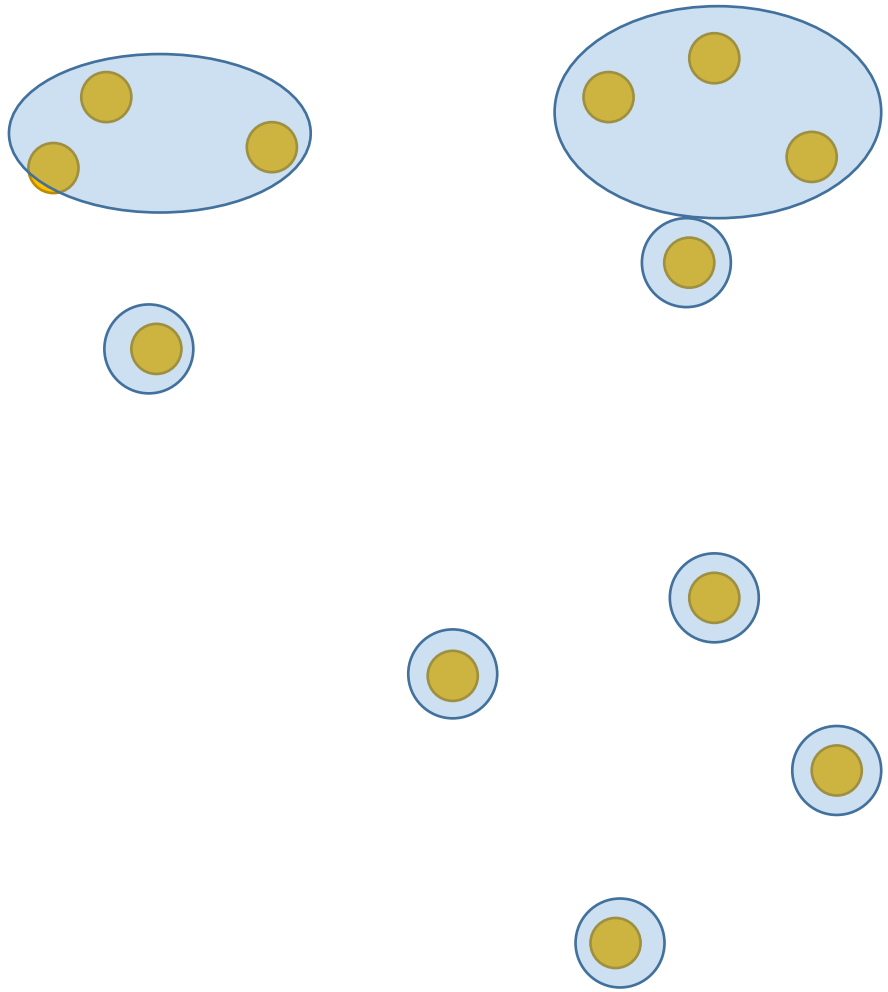
Агломеративная кластеризация



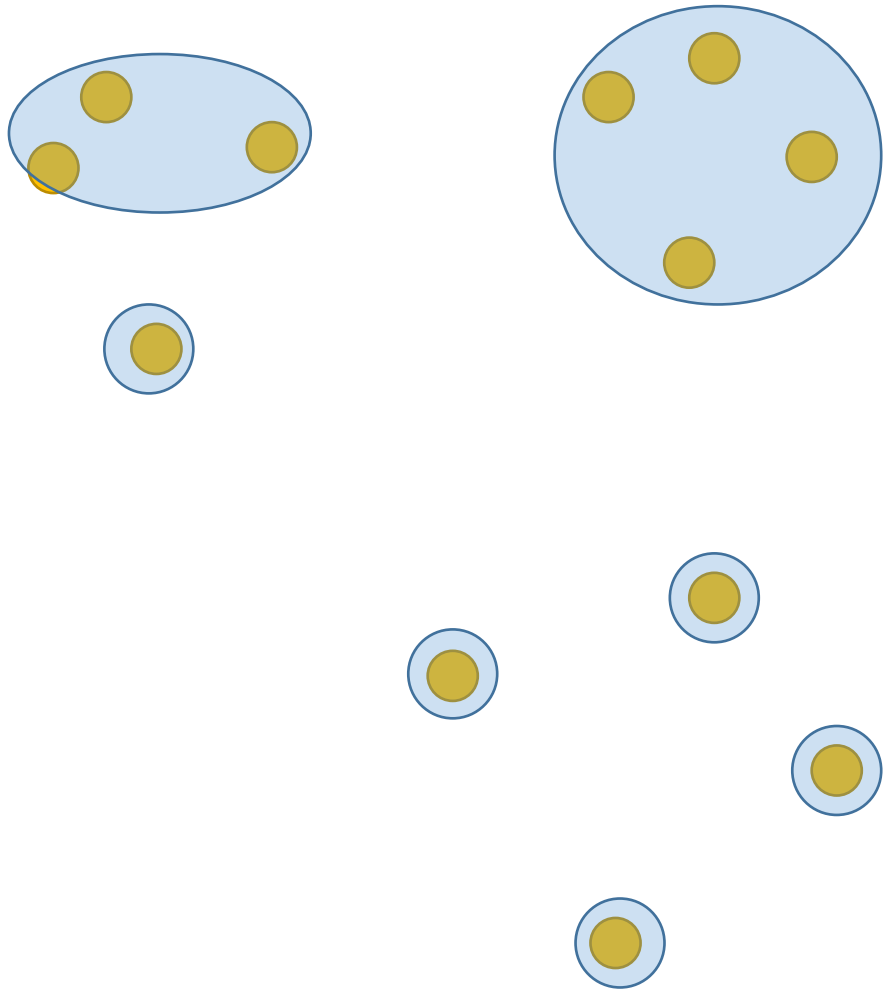
Агломеративная кластеризация



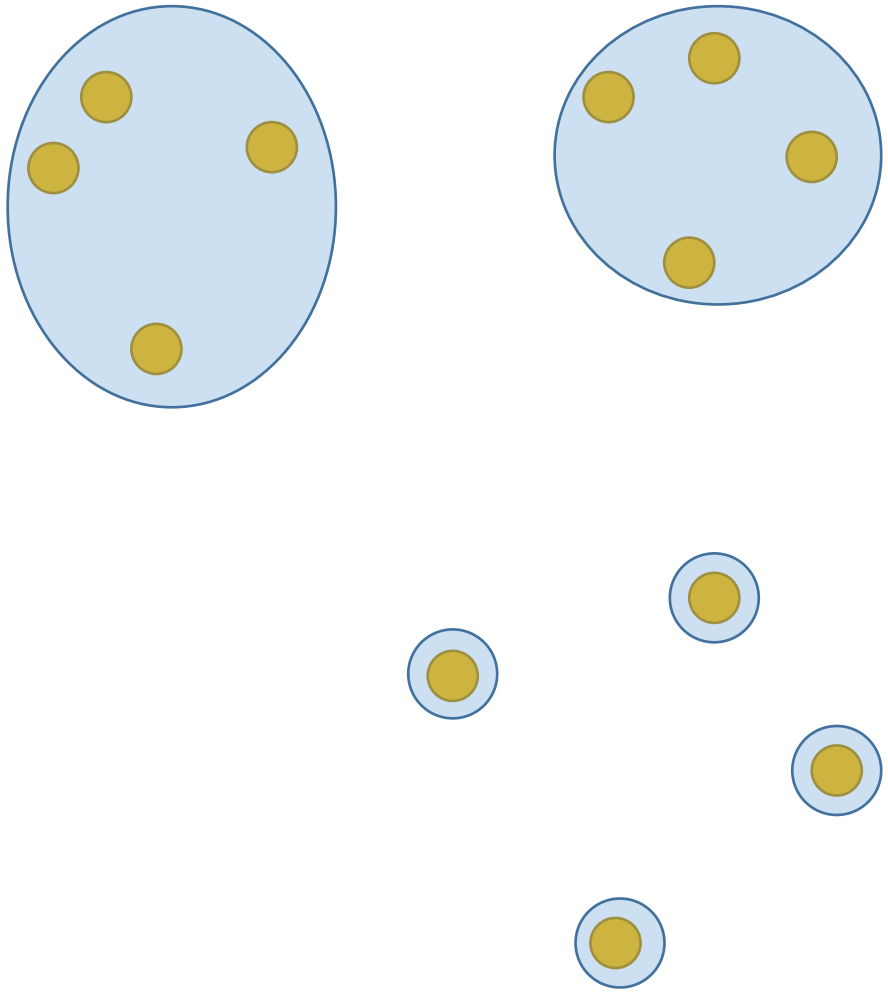
Агломеративная кластеризация



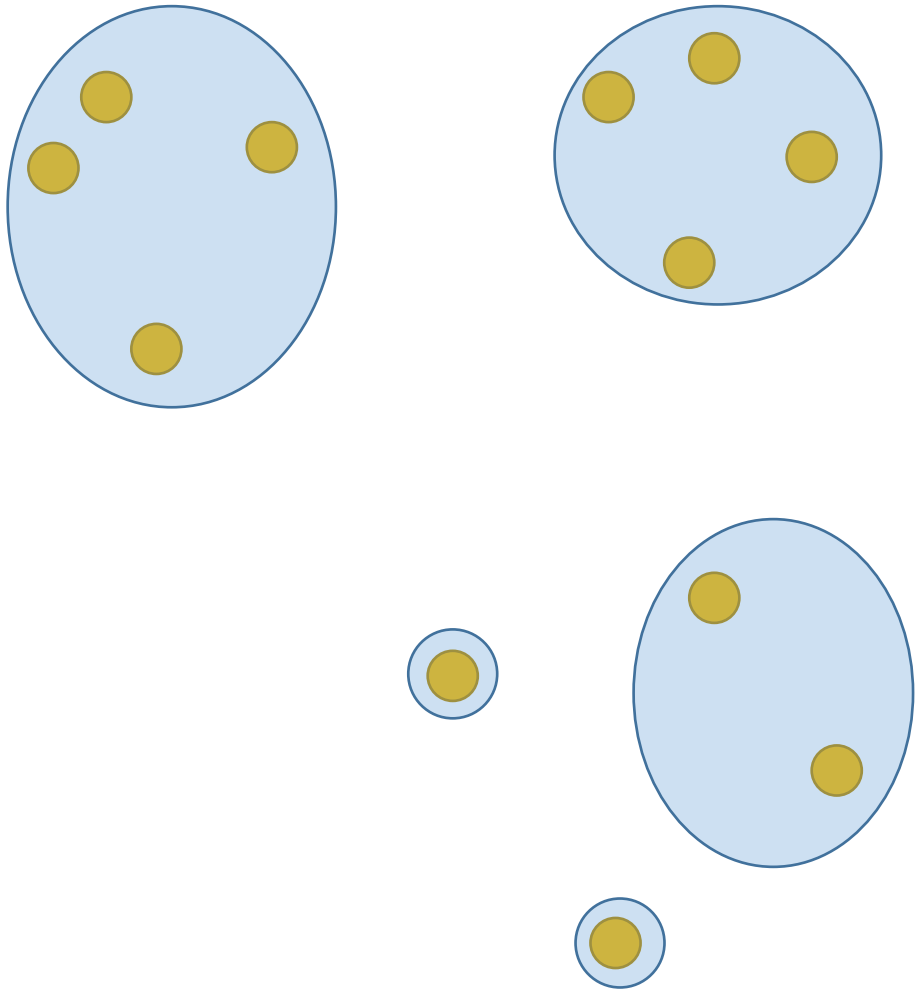
Агломеративная кластеризация



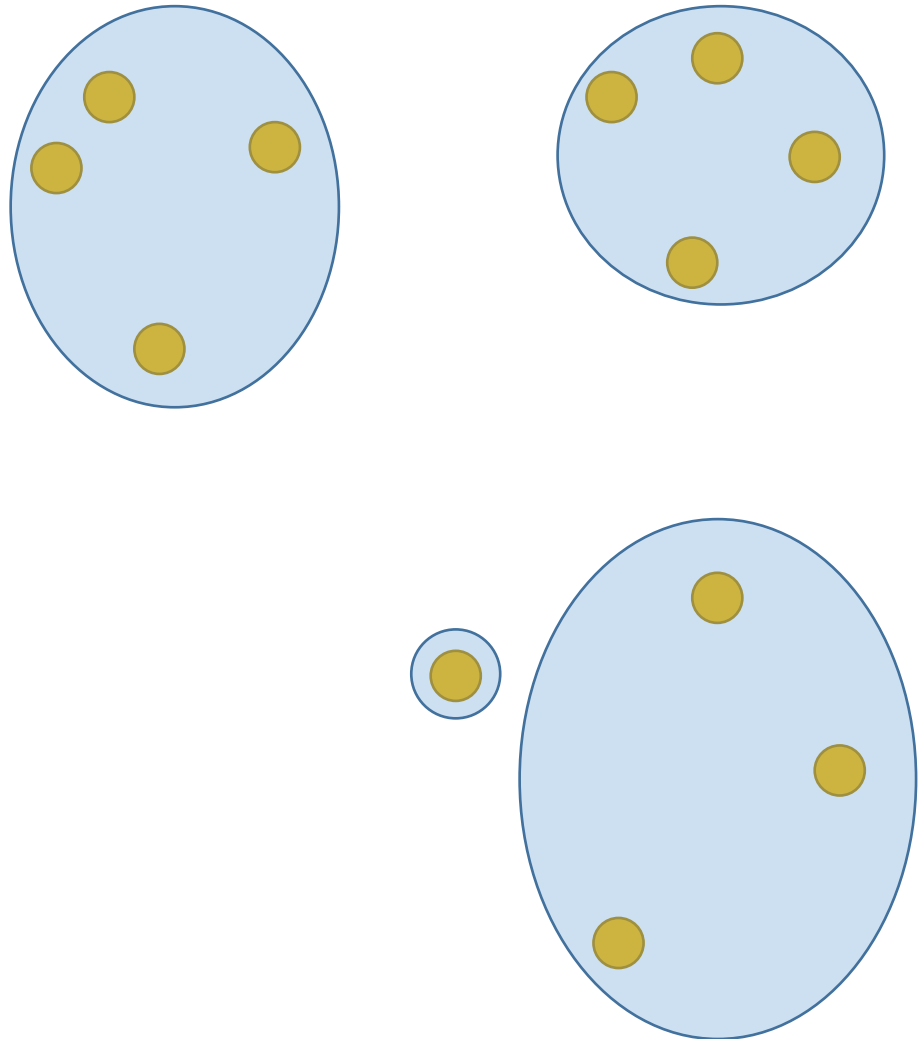
Агломеративная кластеризация



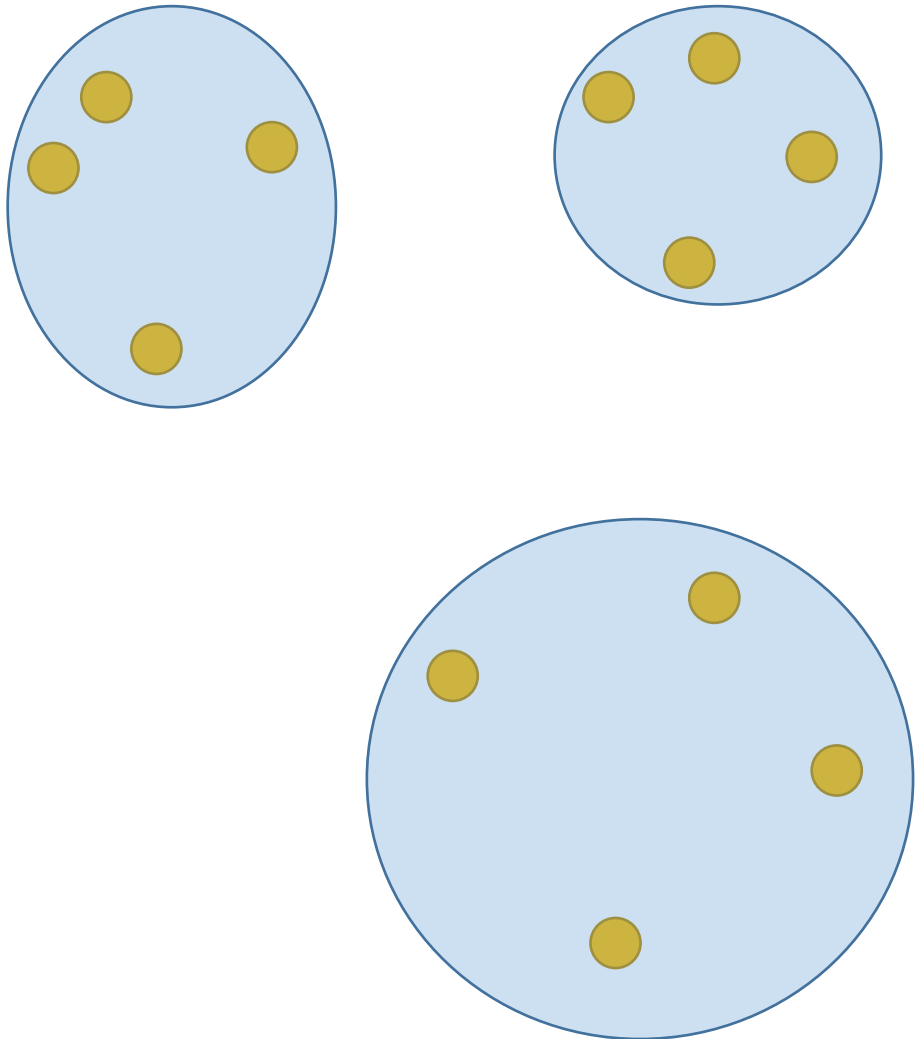
Агломеративная кластеризация



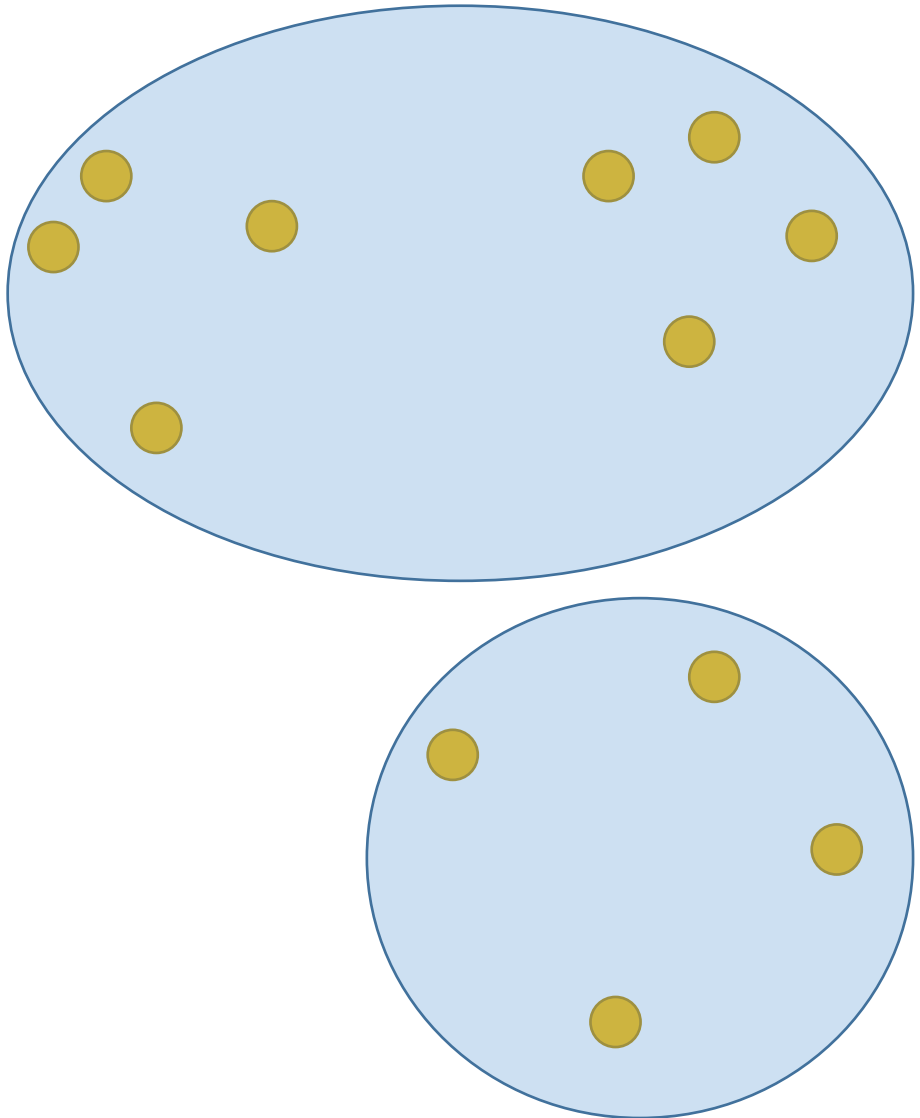
Агломеративная кластеризация



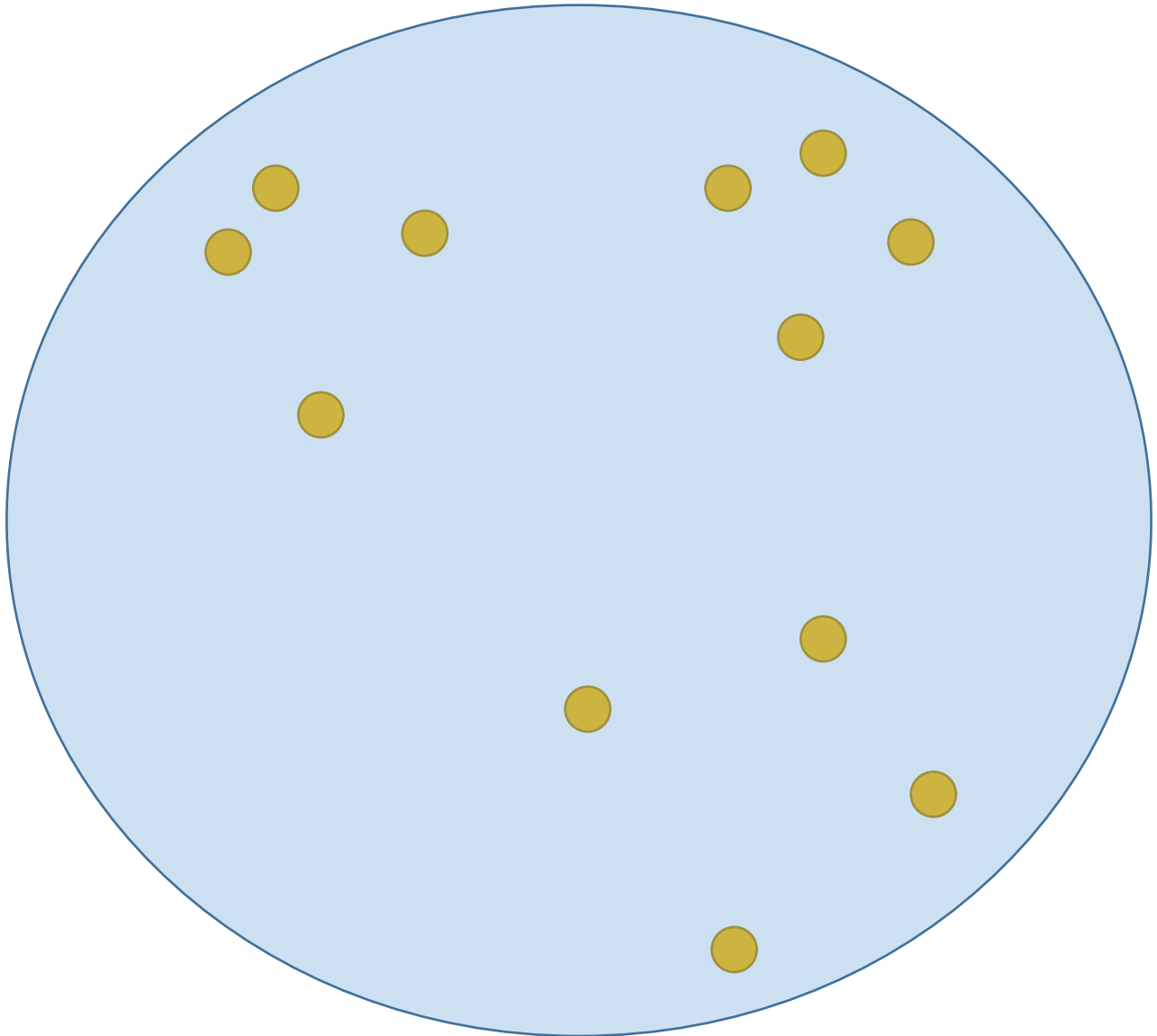
Агломеративная кластеризация



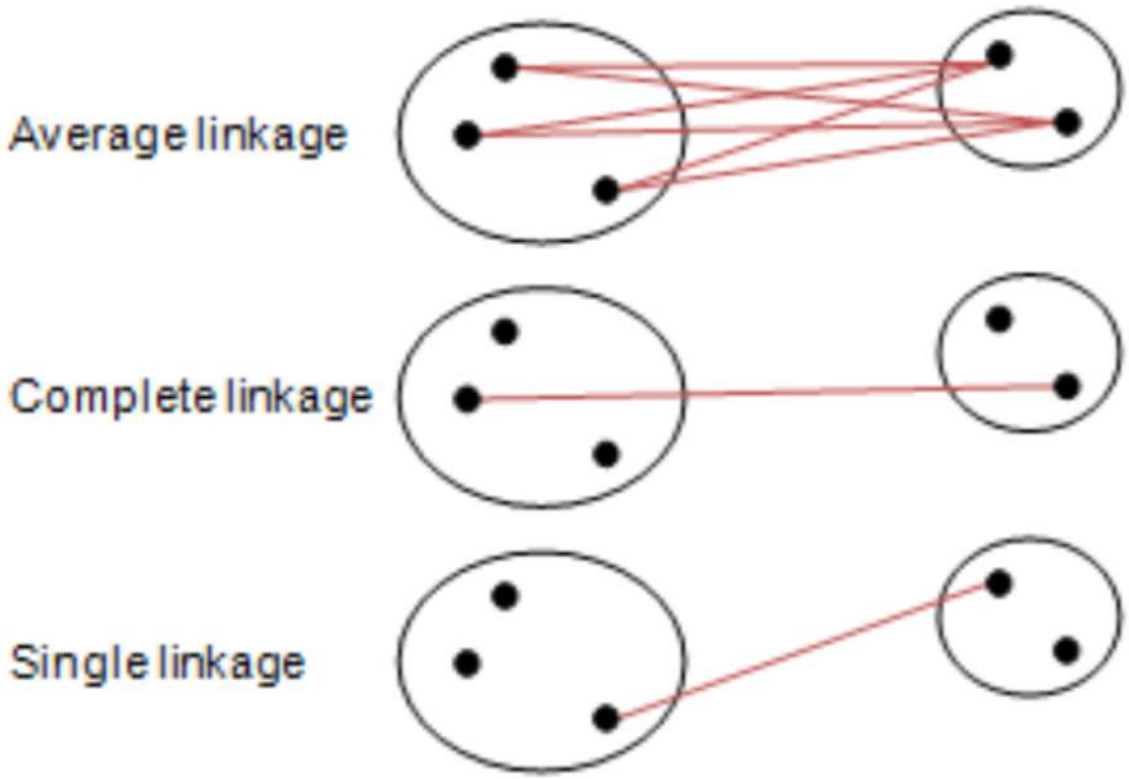
Агломеративная кластеризация



Агломеративная кластеризация



Расстояния между кластерами



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:

$$R^{\text{б}}(W, S) = \min_{w \in W, s \in S} \rho(w, s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = -\tfrac{1}{2}.$$

Расстояние дальнего соседа:

$$R^{\text{д}}(W, S) = \max_{w \in W, s \in S} \rho(w, s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = \tfrac{1}{2}.$$

Среднее расстояние:

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

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 между центрами:

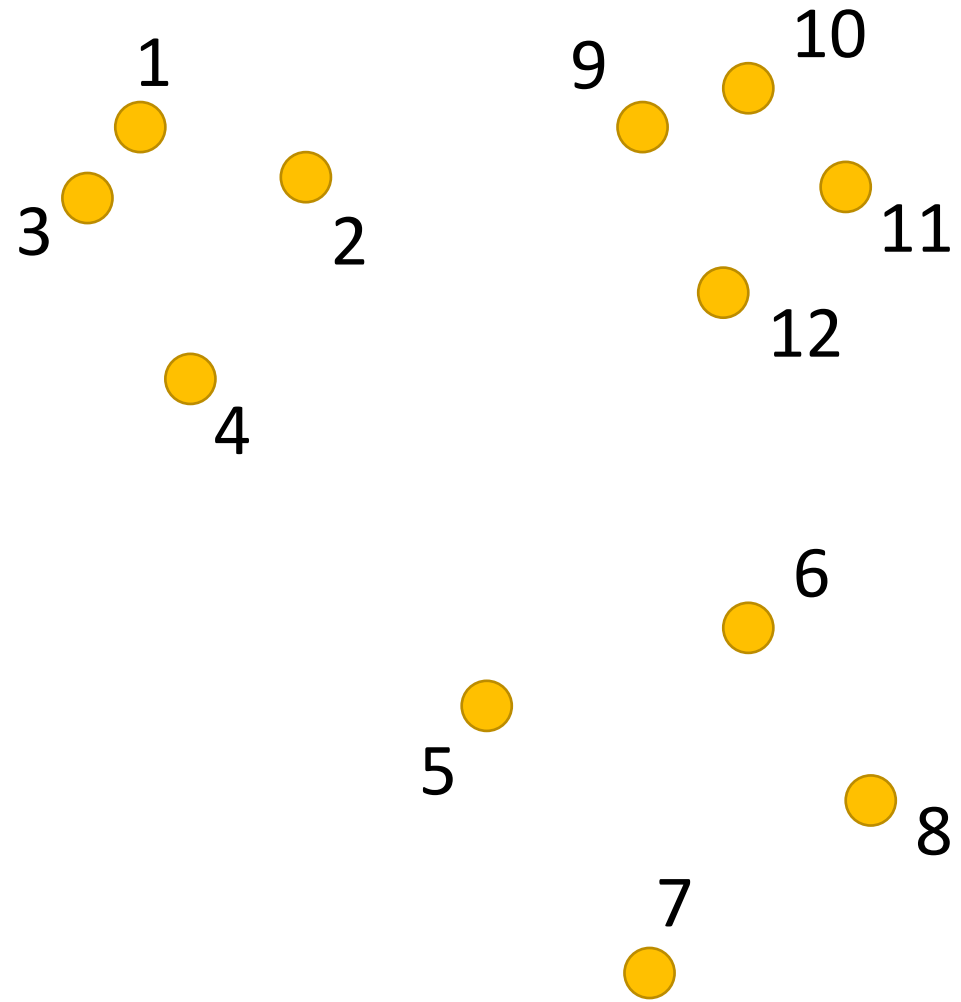
$$R^{\text{ц}}(W, S) = \rho^2\Big(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|}\Big); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = -\alpha_U \alpha_V,\ \gamma = 0.$$

Расстояние Уорда:

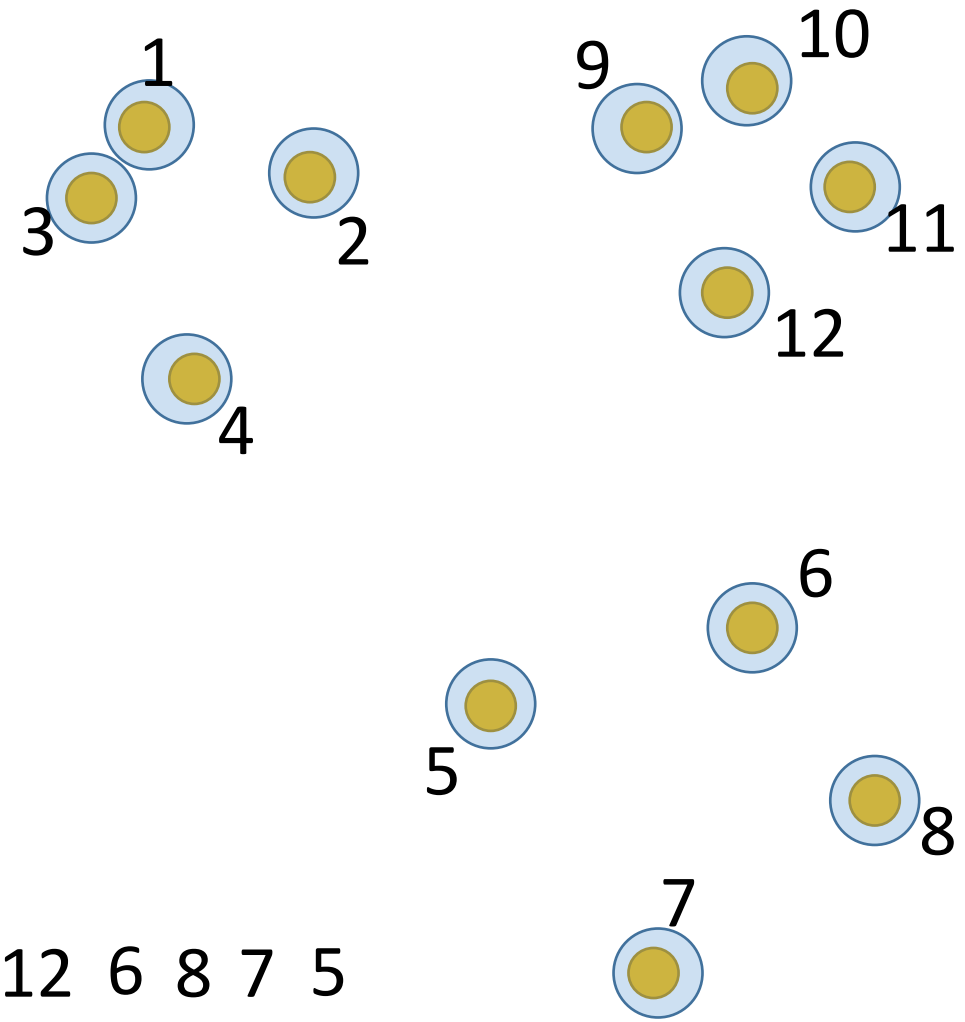
$$R^{\text{у}}(W, S) = \frac{|S||W|}{|S|+|W|}\rho^2\Big(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|}\Big); \quad \alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \beta = \frac{-|S|}{|S|+|W|},\ \gamma = 0.$$

Source:
http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_%28курс_лекций%2C_К.В.Воронцов%29

Дендрограмма

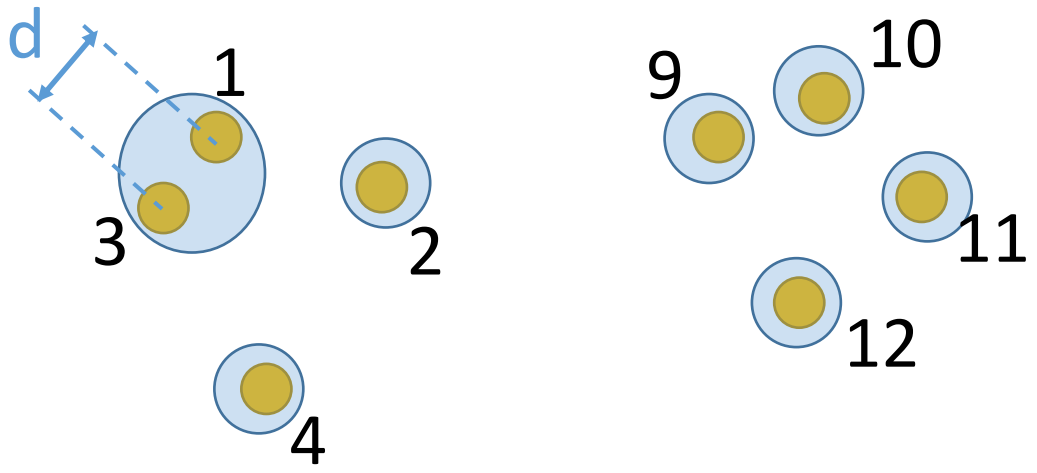


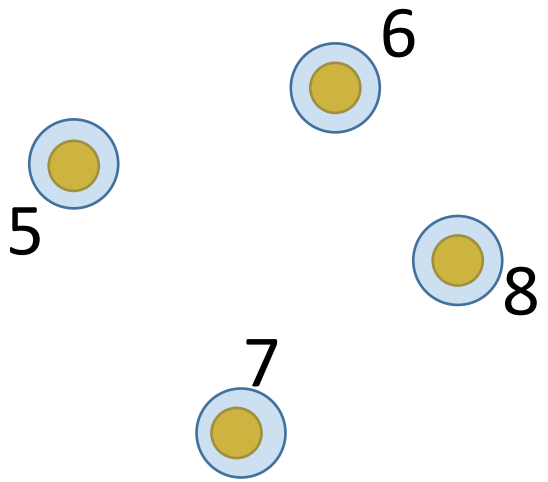
Дендрограмма



1 3 2 4 9 10 11 12 6 8 7 5

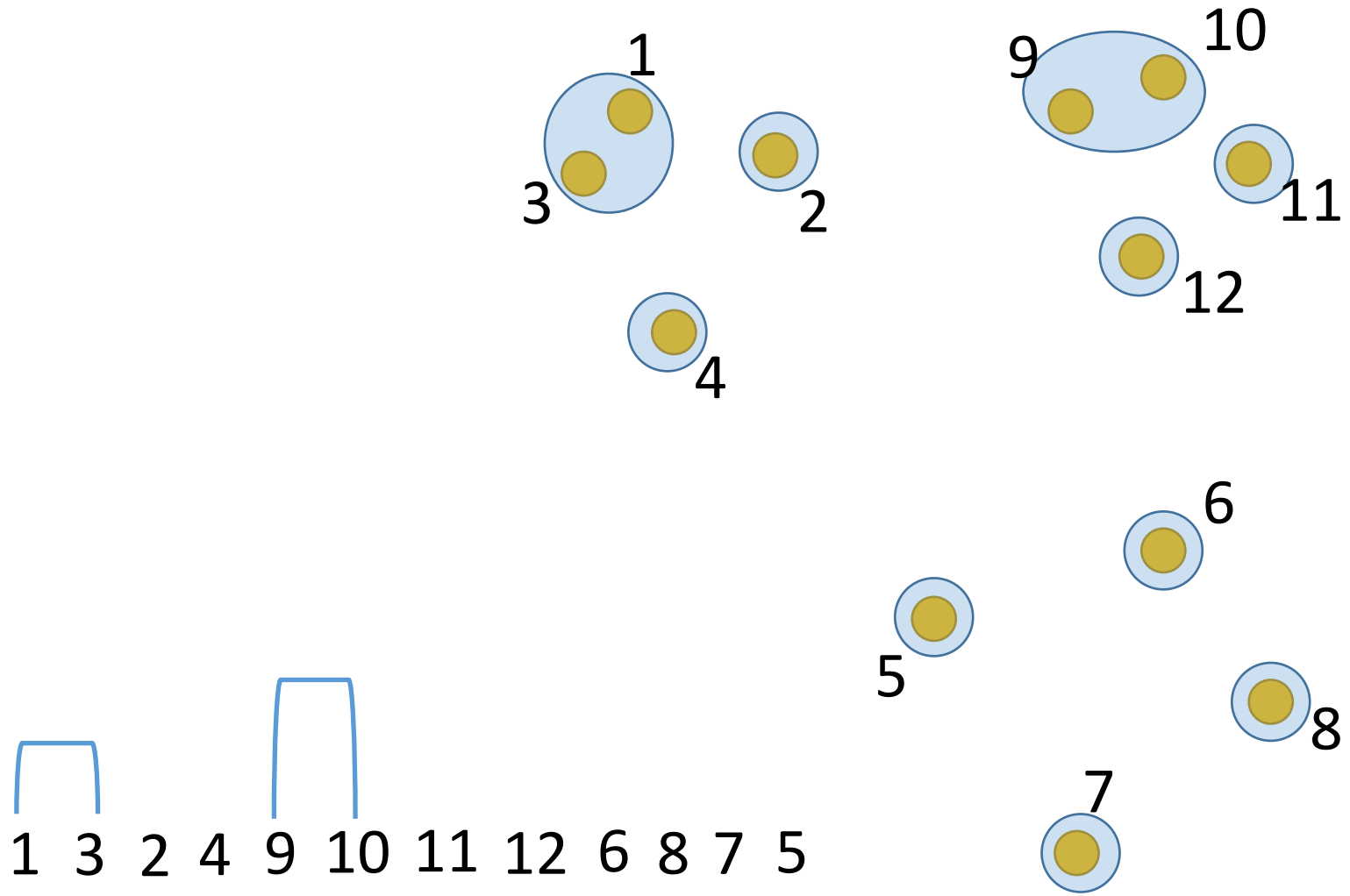
Дендрограмма



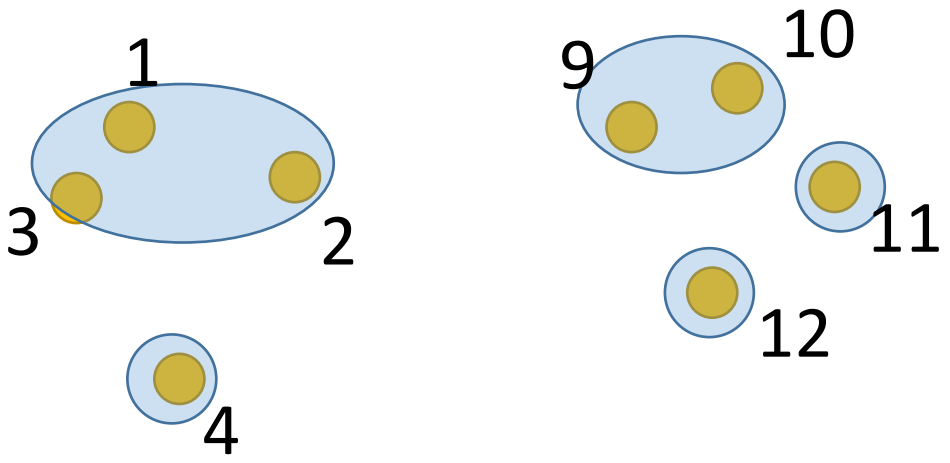


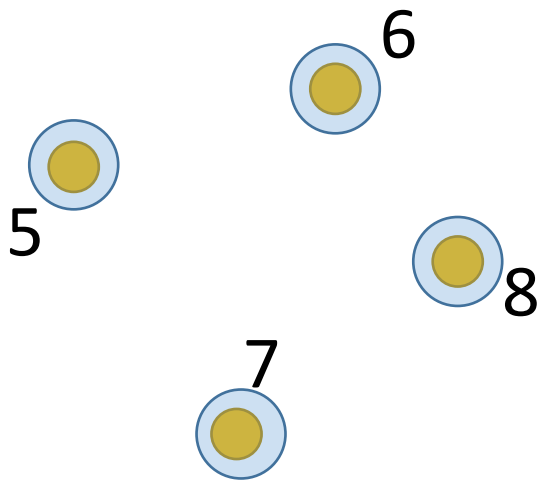
1 3 2 4 9 10 11 12 6 8 7 5

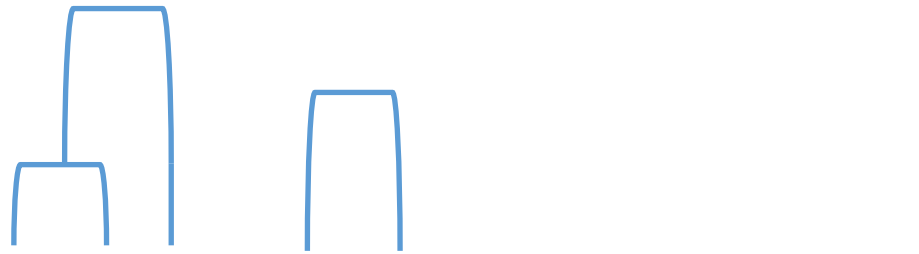
Дендрограмма



Дендрограмма

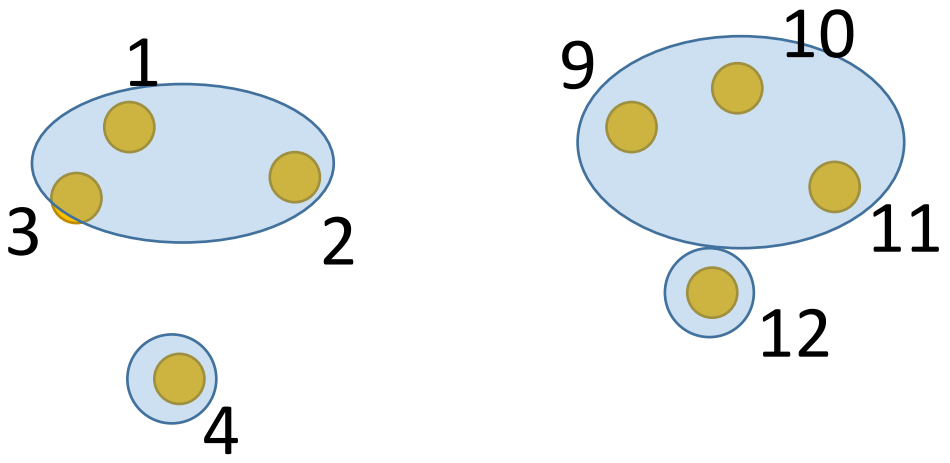


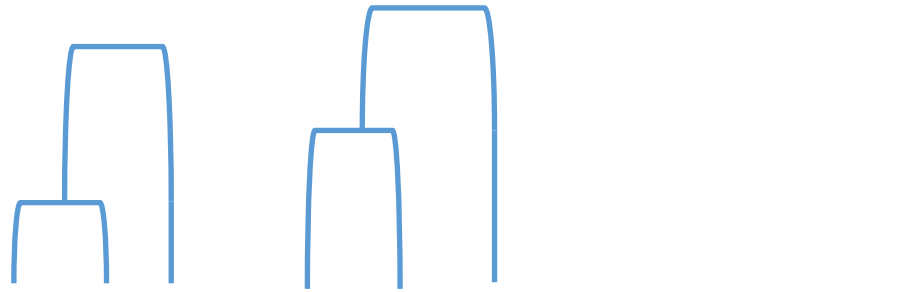


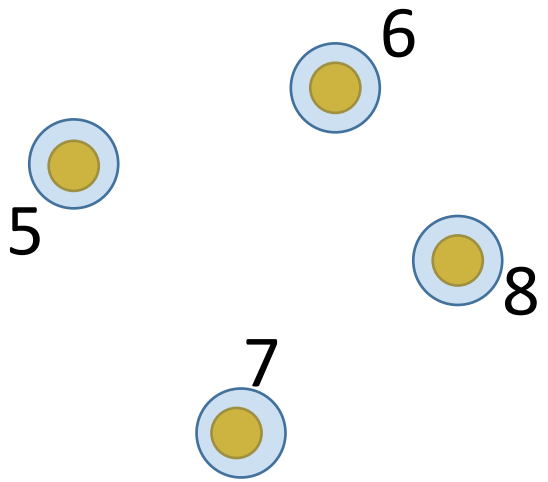


1 3 2 4 9 10 11 12 6 8 7 5

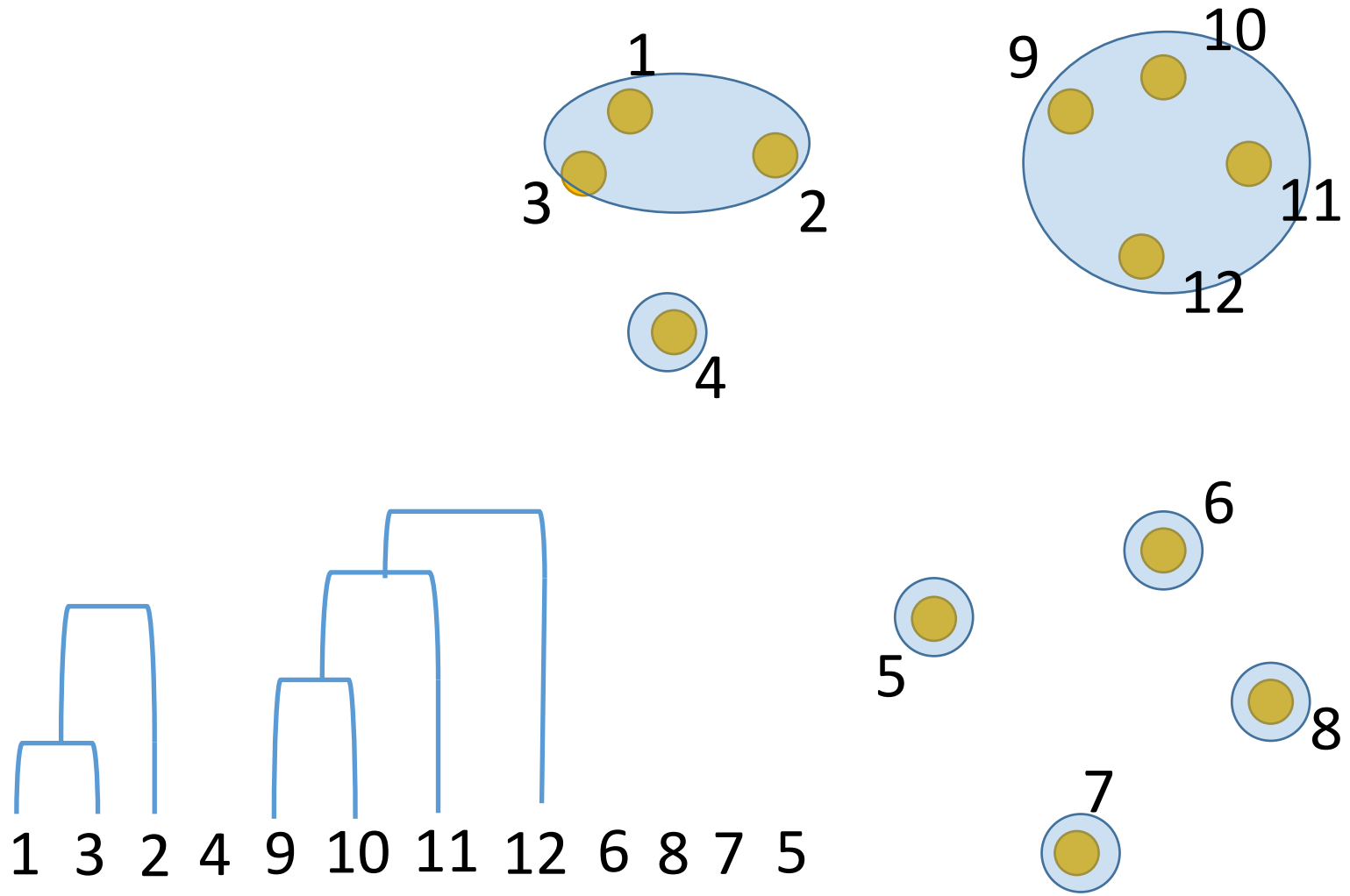
Дендрограмма



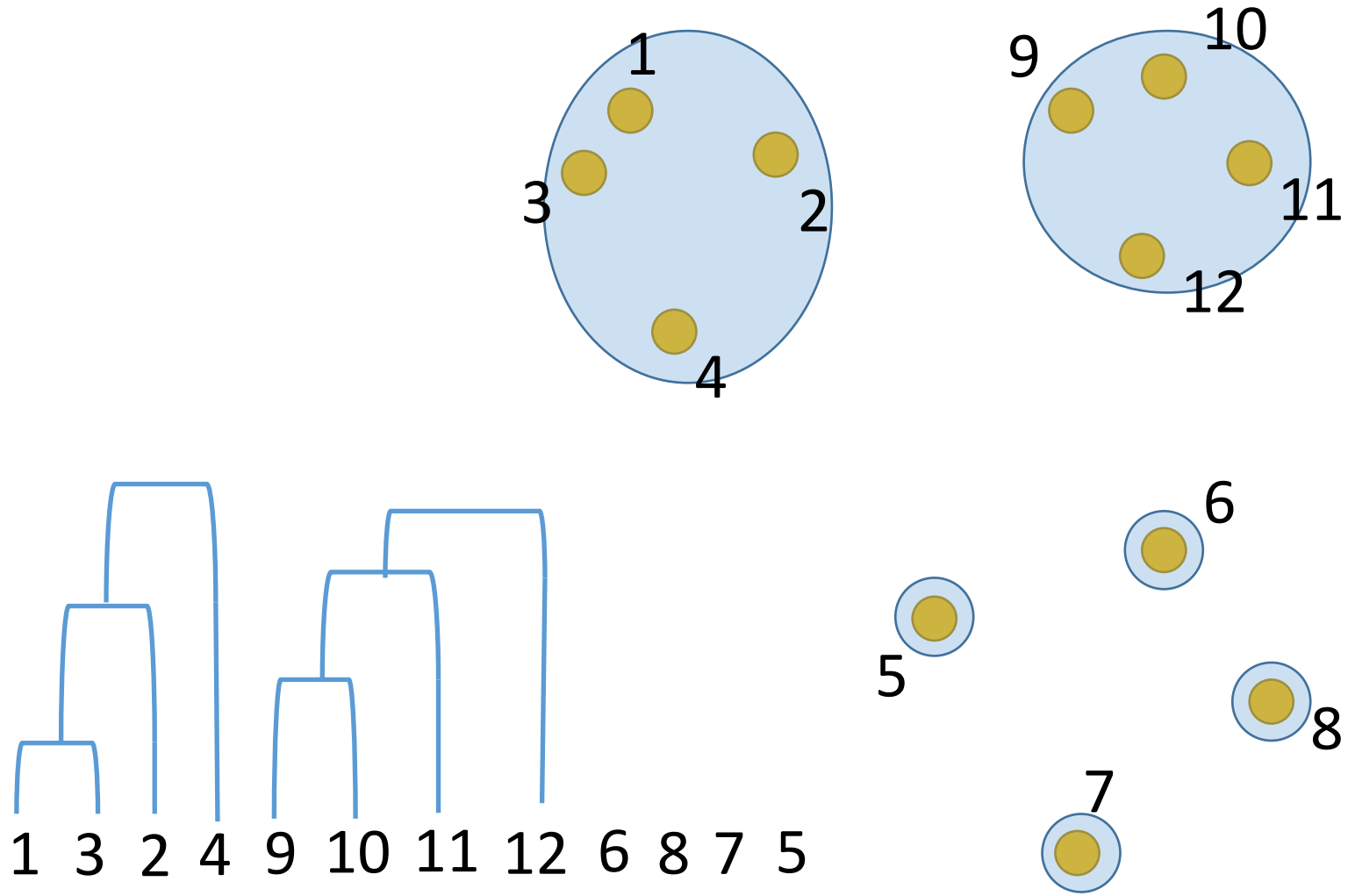




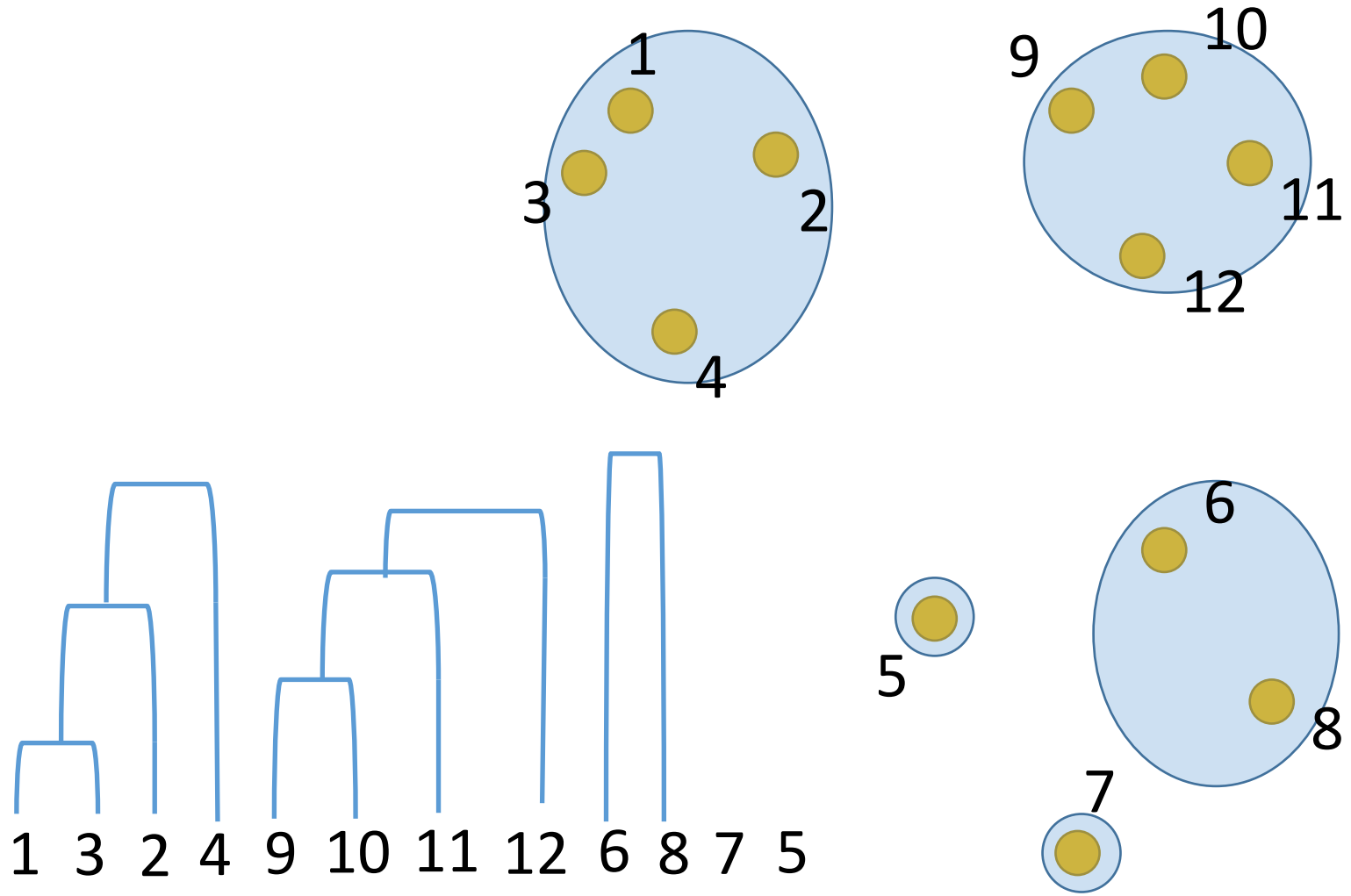
Дендрограмма



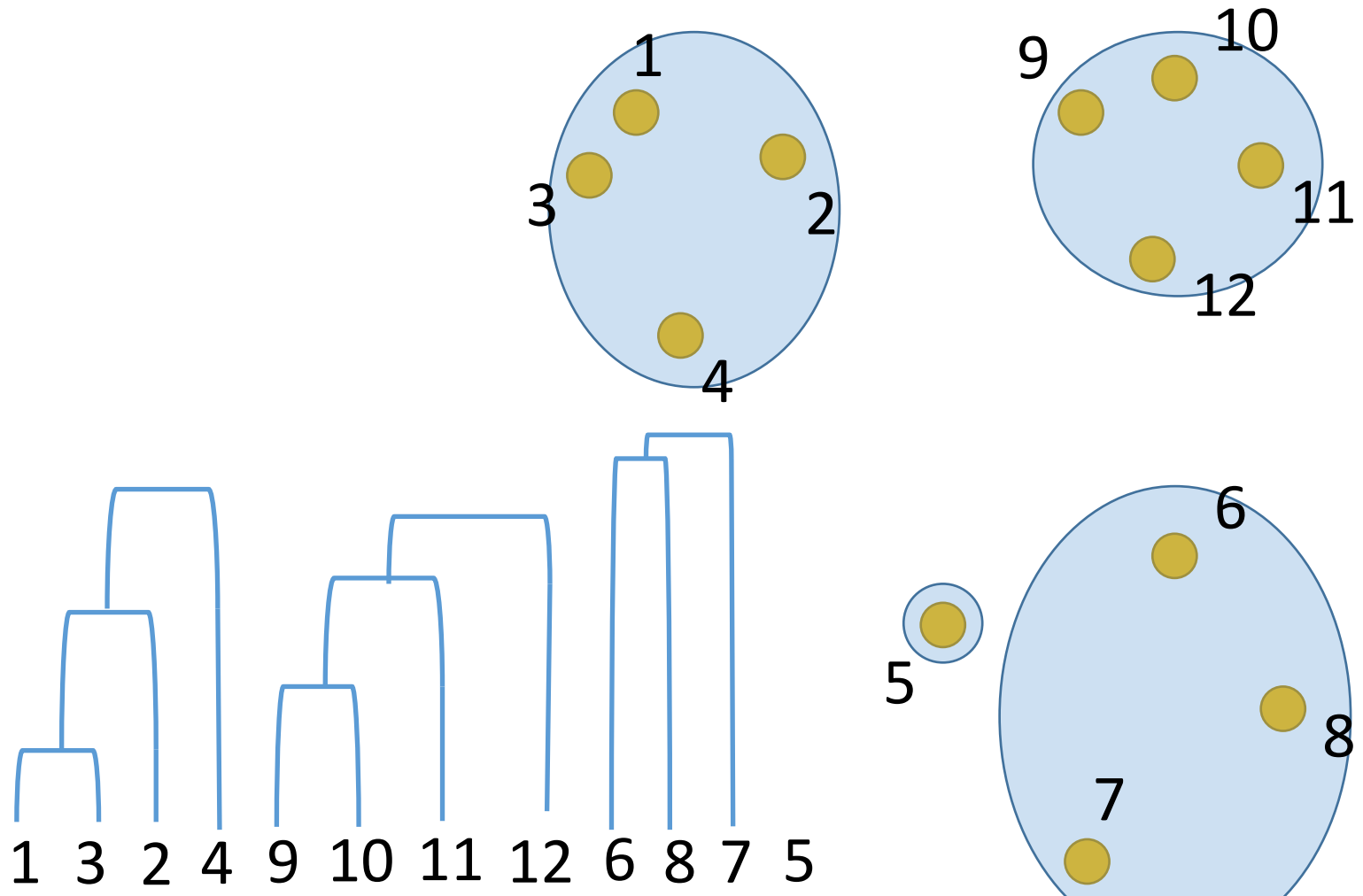
Дендрограмма



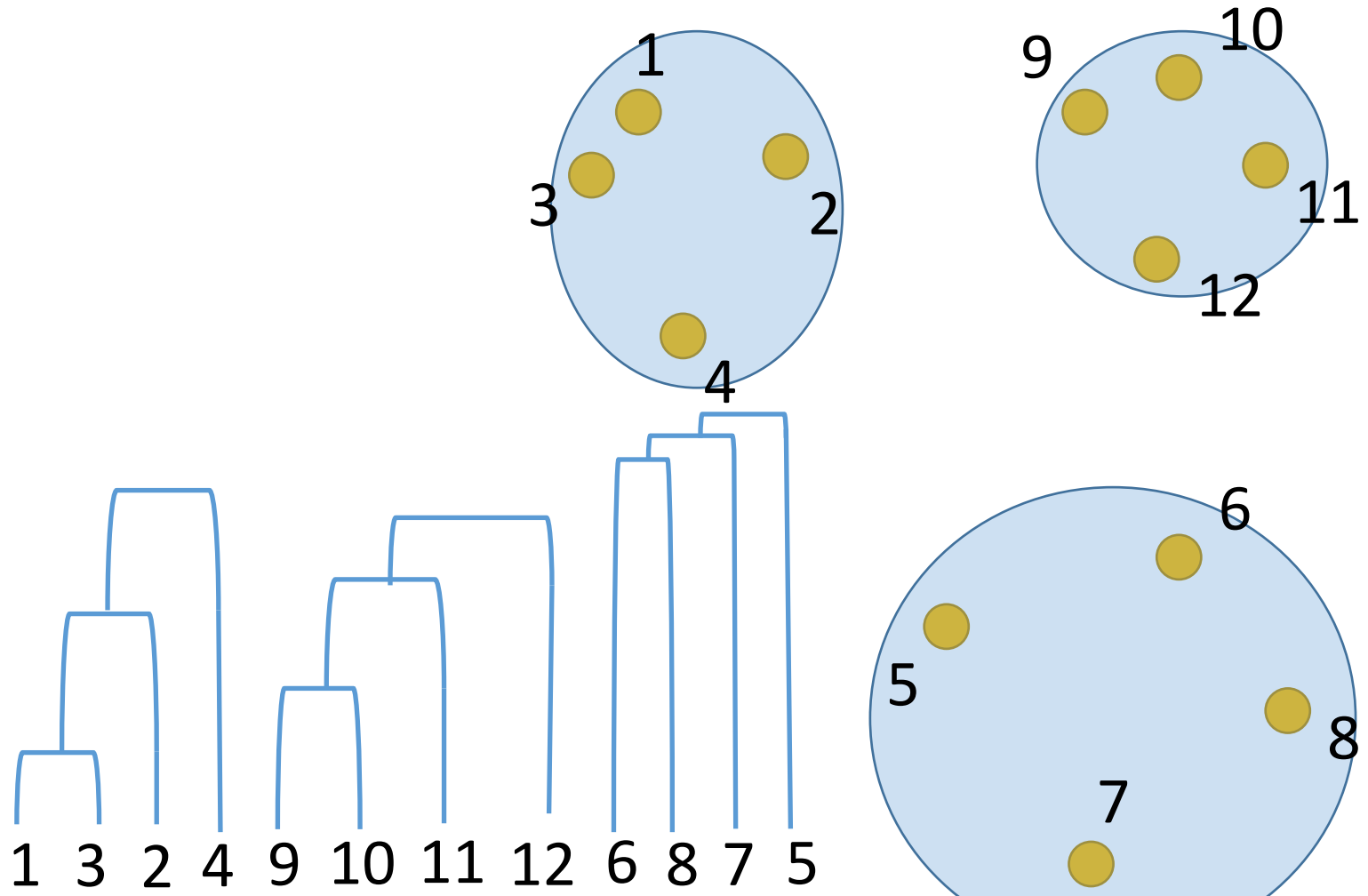
Дендрограмма



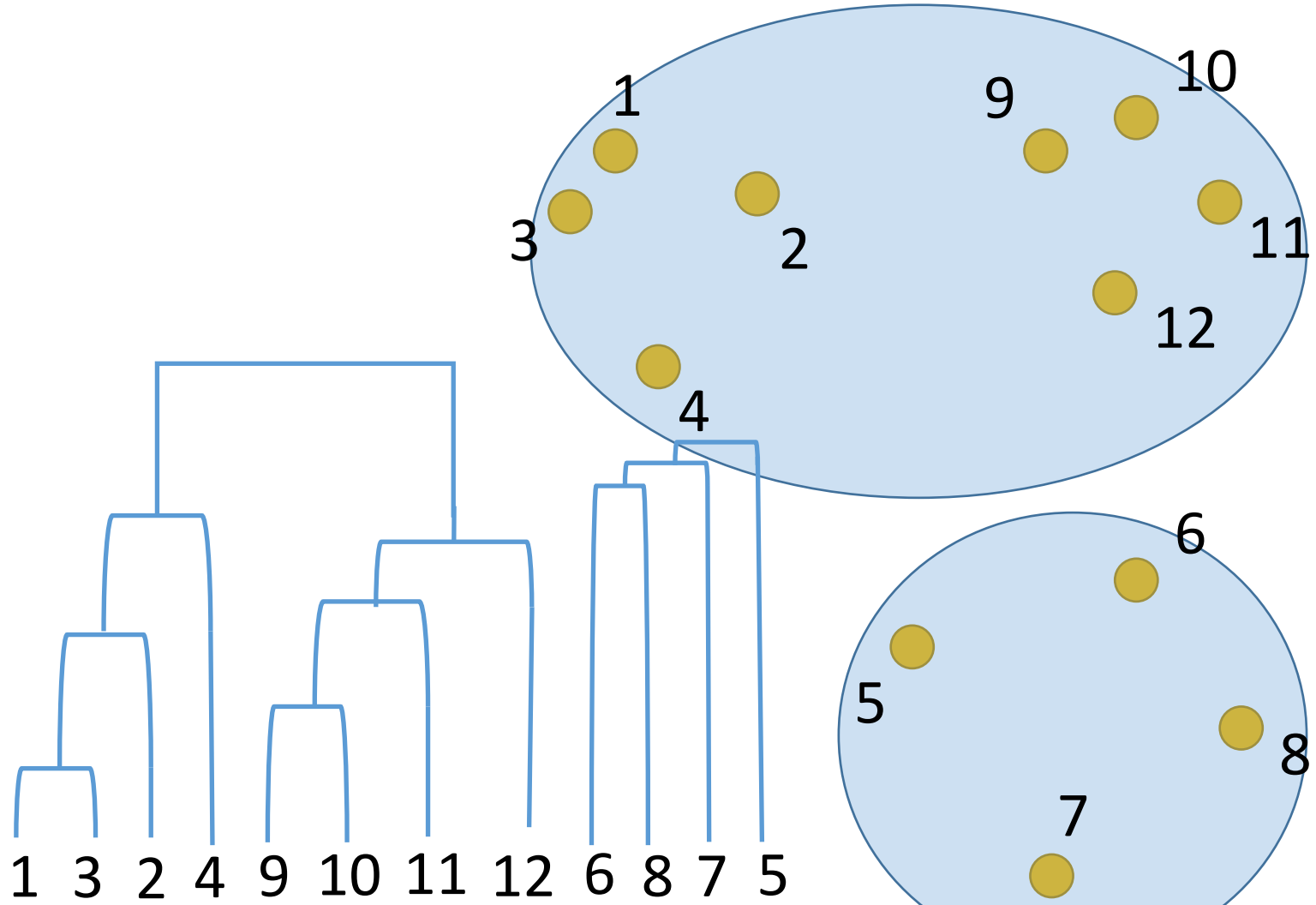
Дендрограмма



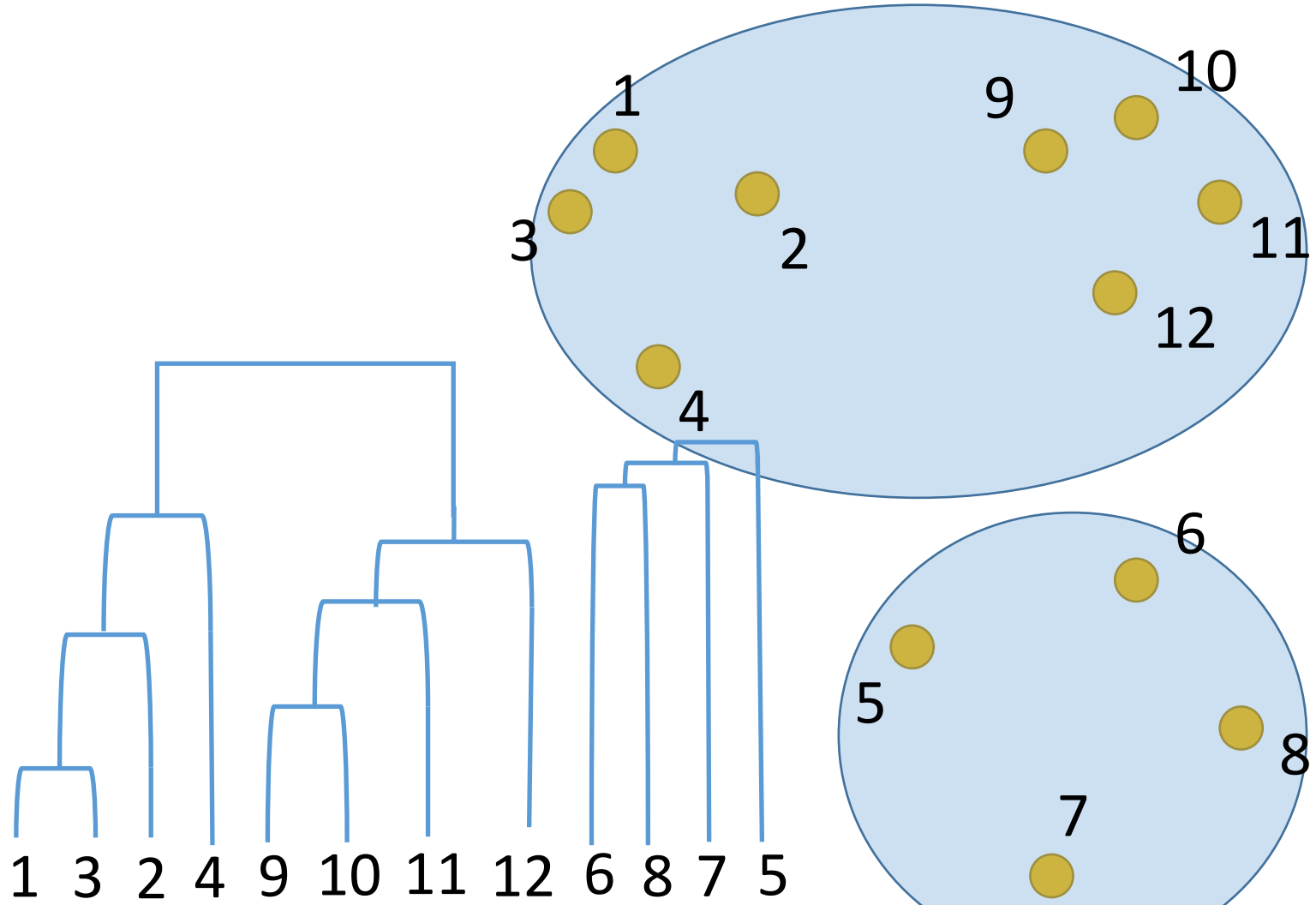
Дендрограмма



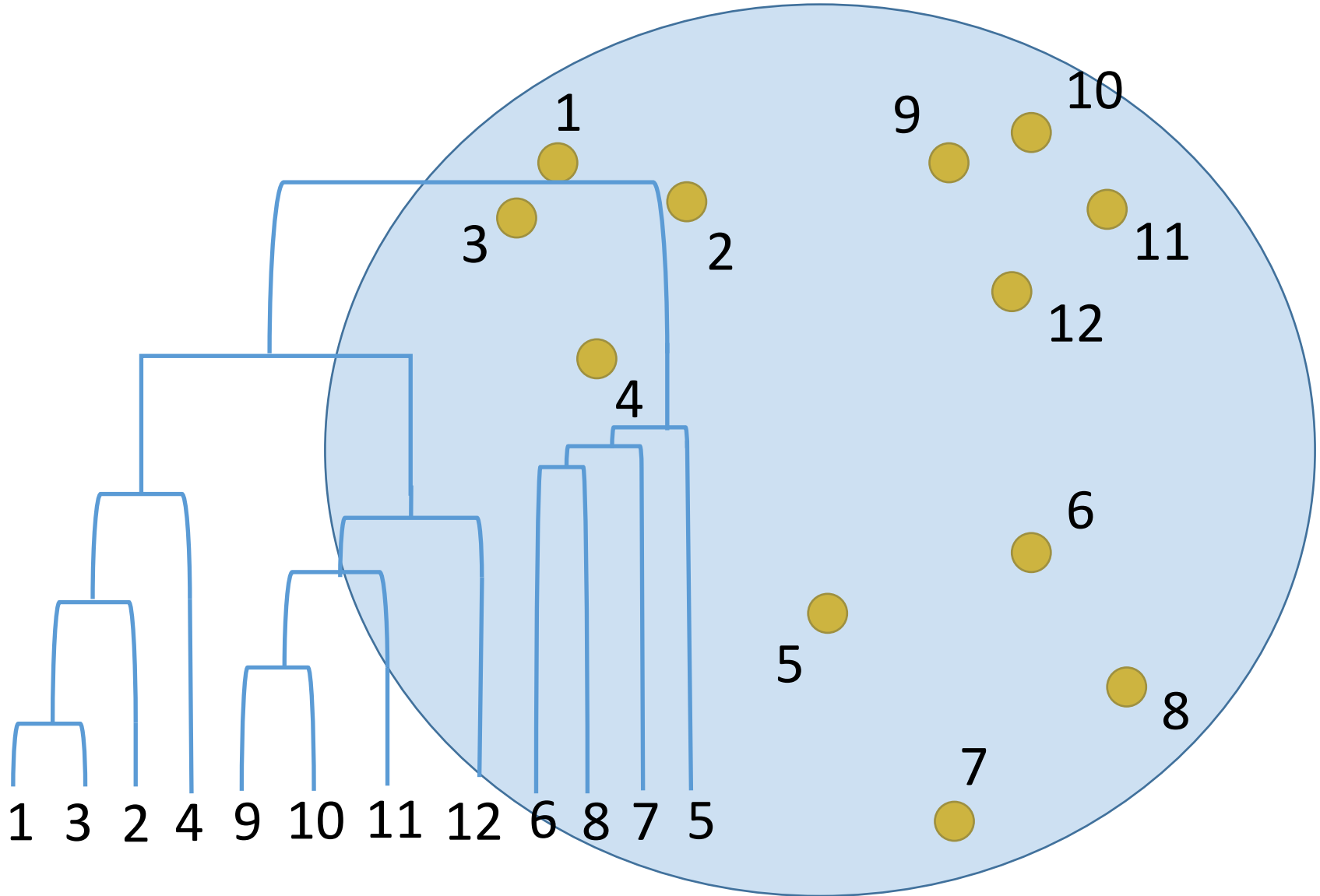
Дендрограмма



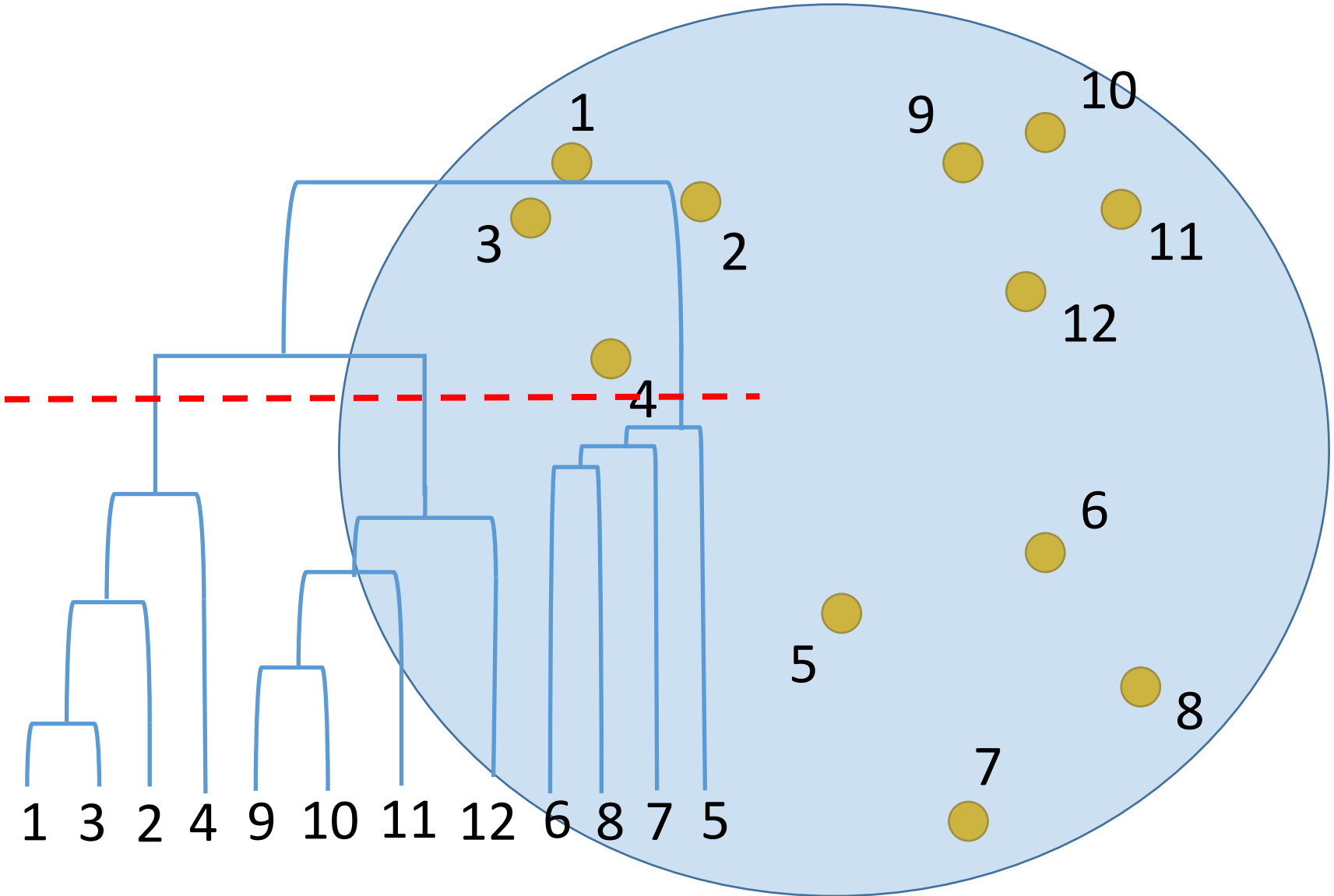
Дендрограмма



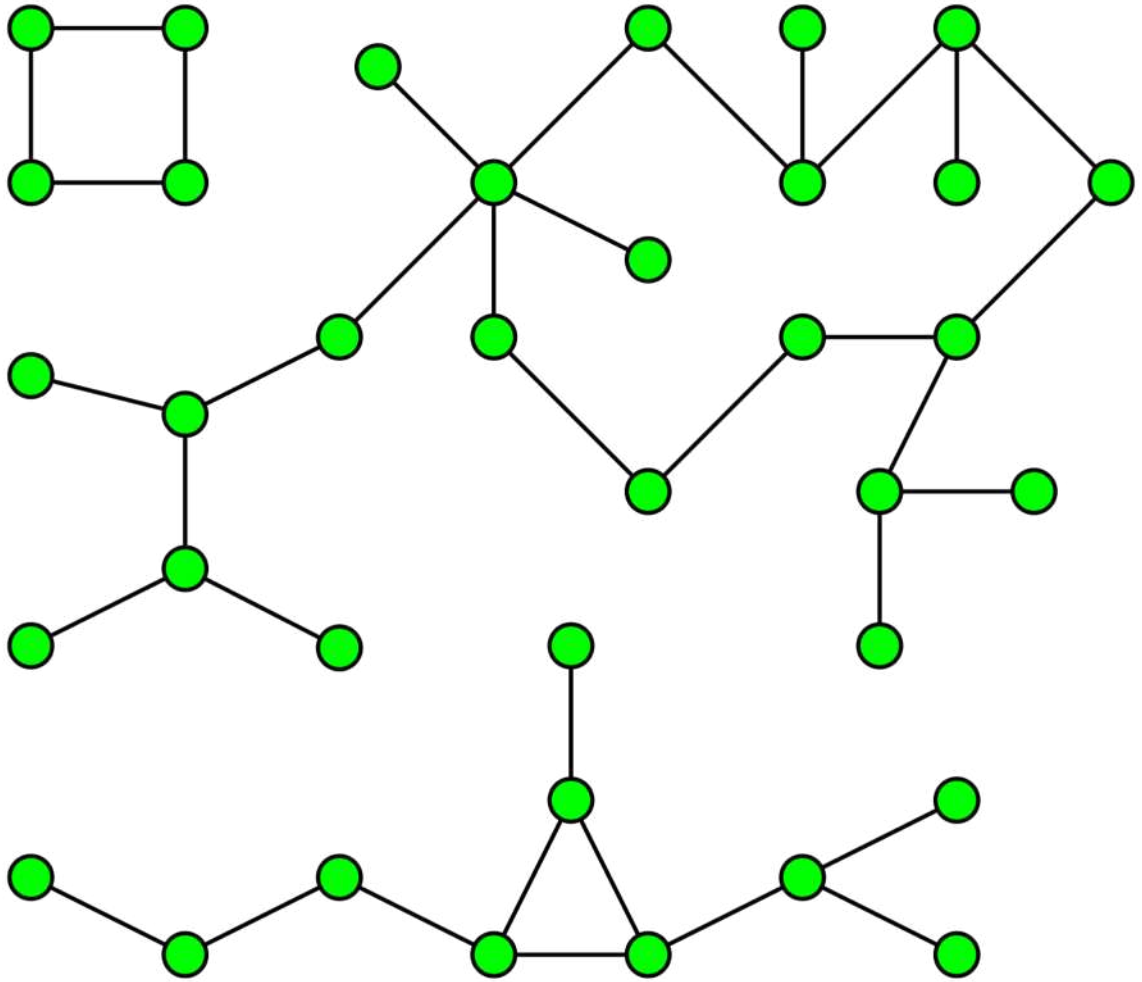
Дендрограмма



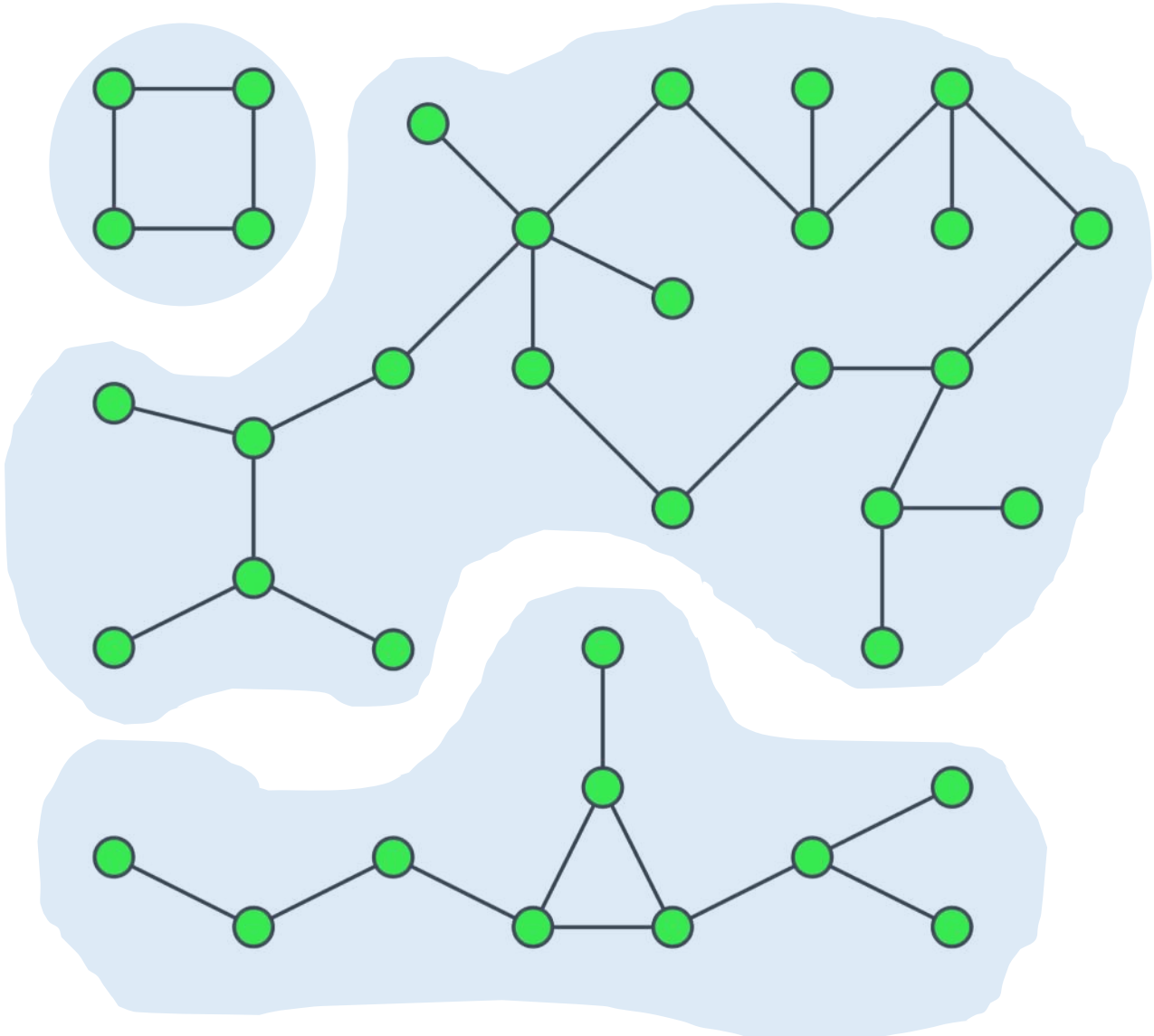
Дендрограмма



Выделение связных компонент



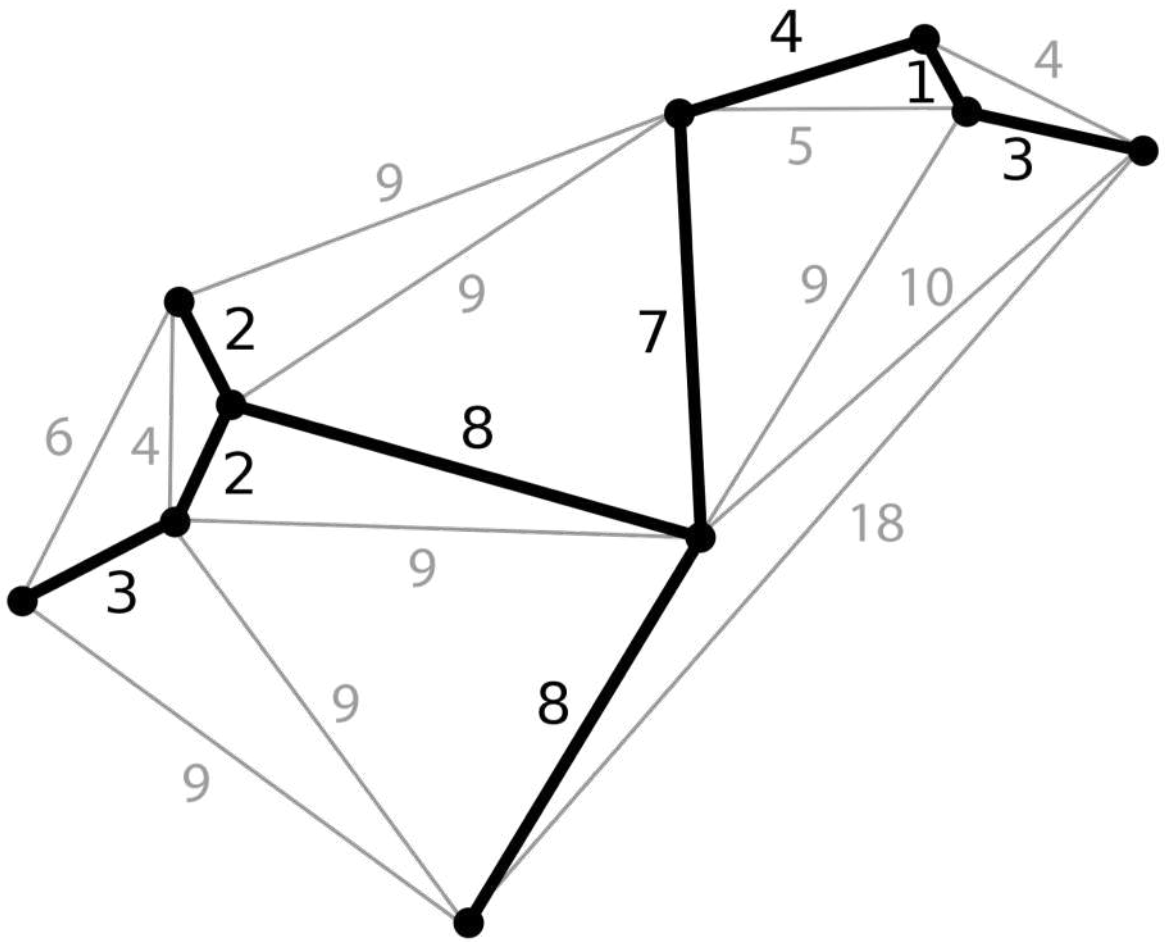
Выделение связных компонент



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



Кластеризация с помощью минимального остовного дерева

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

Идея density-based методов

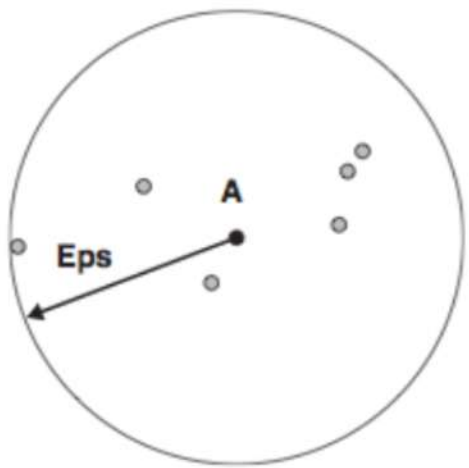


Figure 8.20. Center-based density.

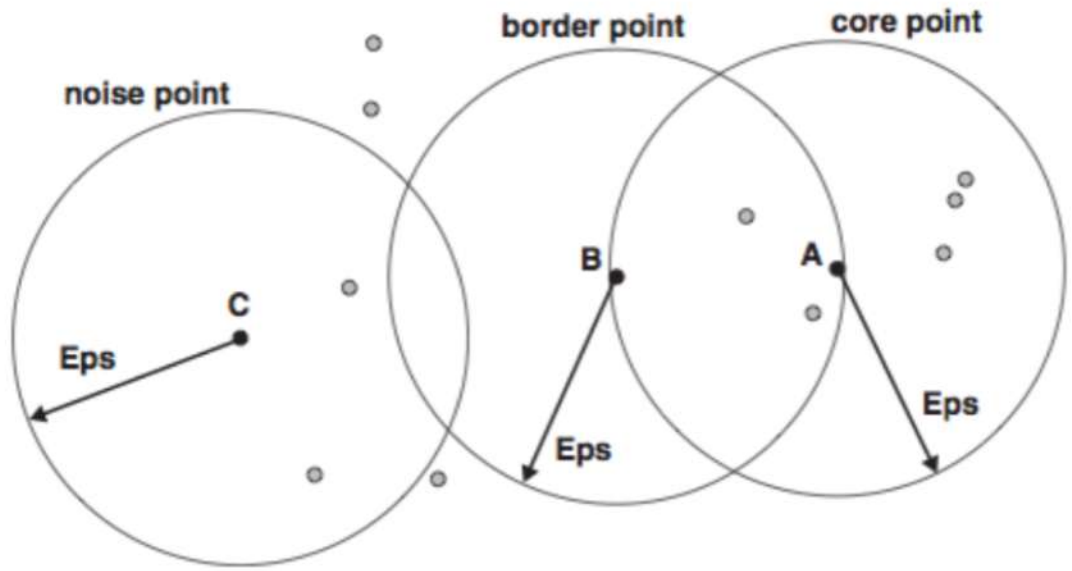
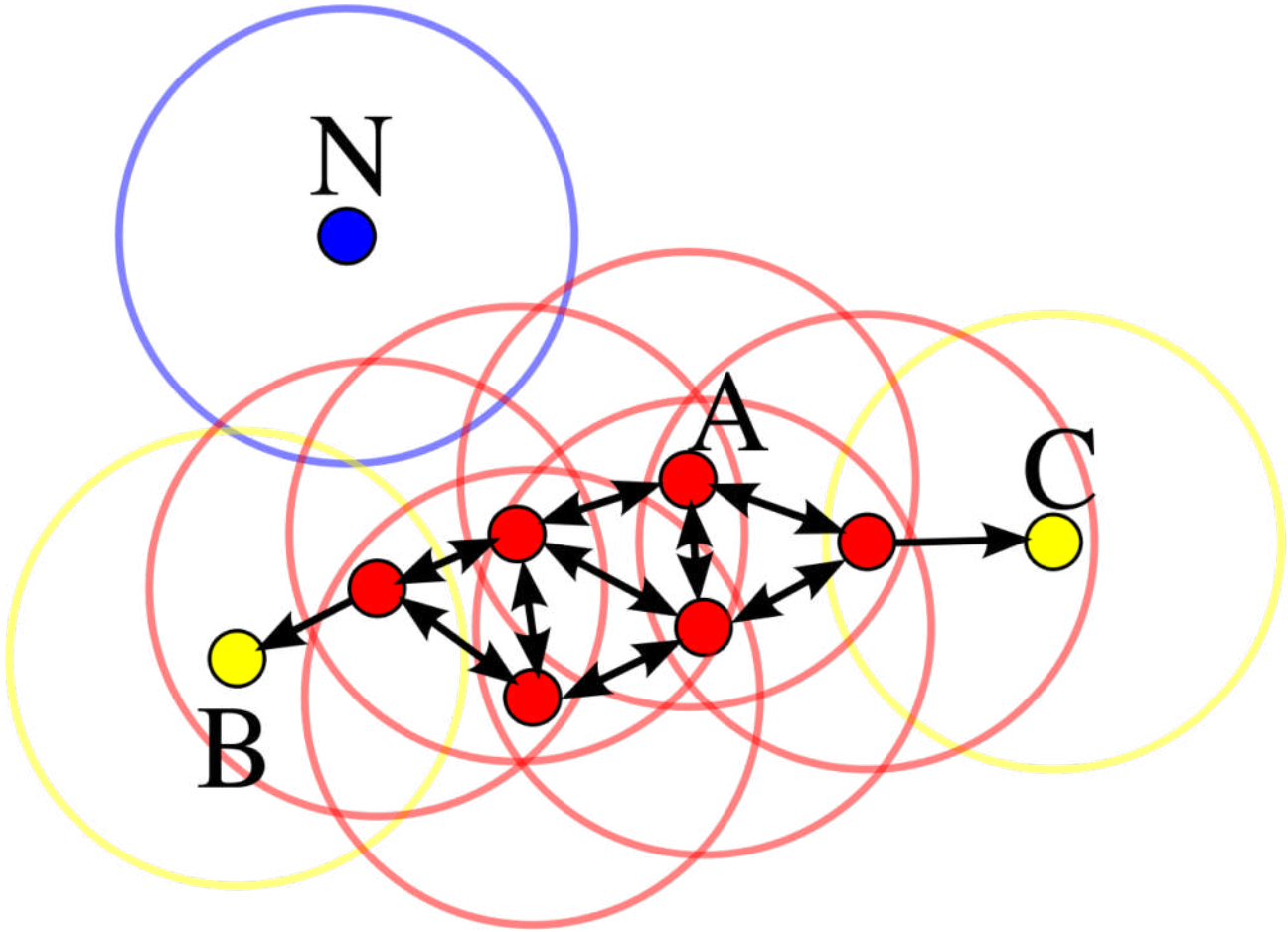


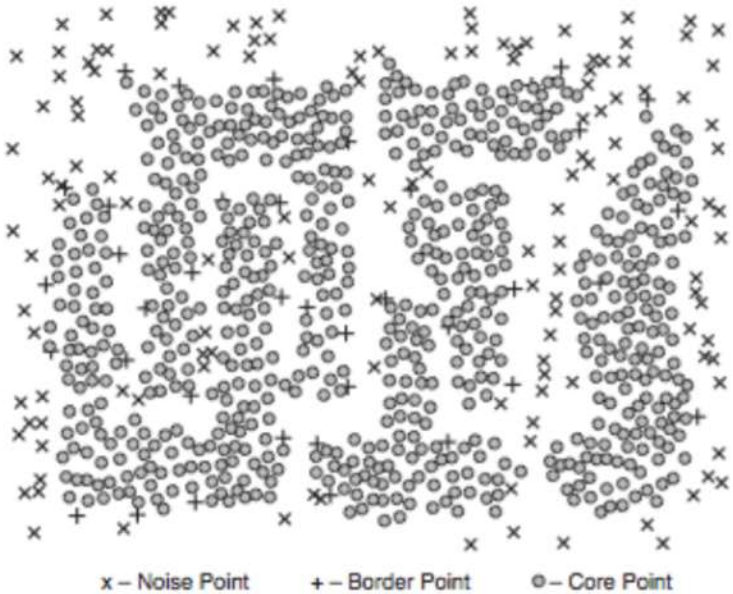
Figure 8.21. Core, border, and noise points.

Основные, шумовые и граничные точки



DBSCAN

(a) Clusters found by DBSCAN.

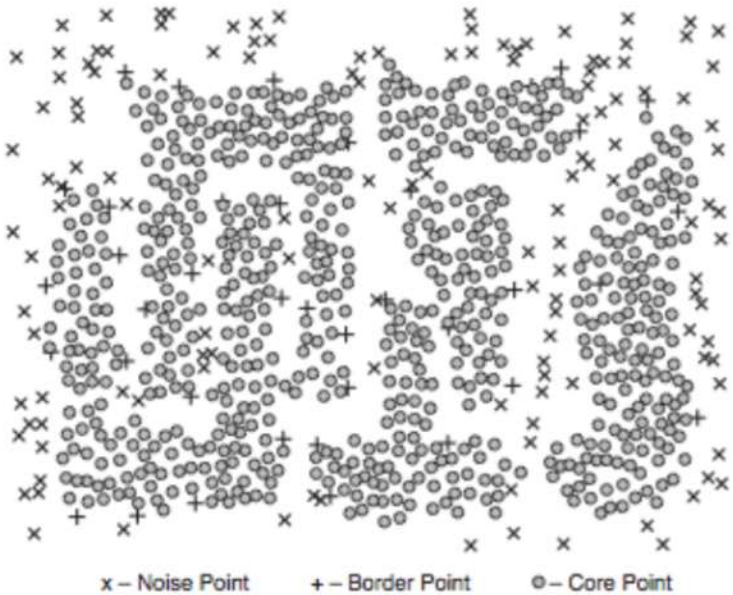


(b) Core, border, and noise points.

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

DBSCAN

(a) Clusters found by DBSCAN.



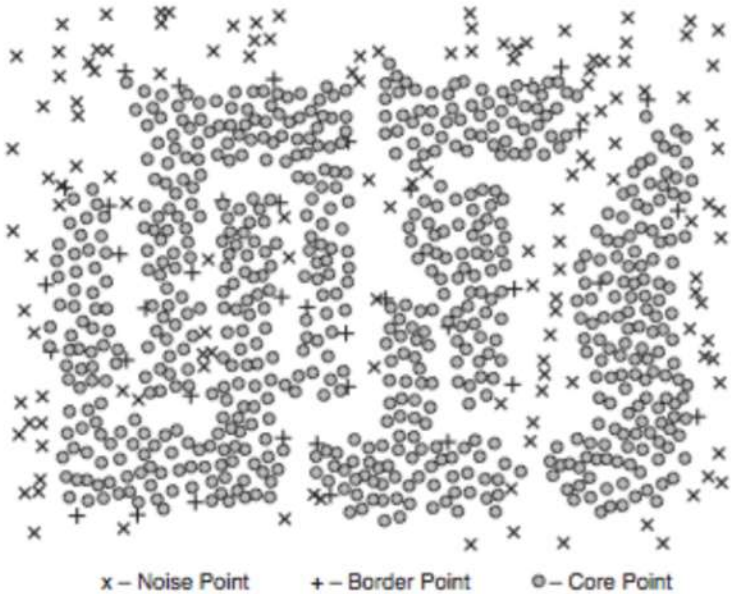
(b) Core, border, and noise points.

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

2: Отбросить точки шума.

DBSCAN

(a) Clusters found by DBSCAN.



(b) Core, border, and noise points.

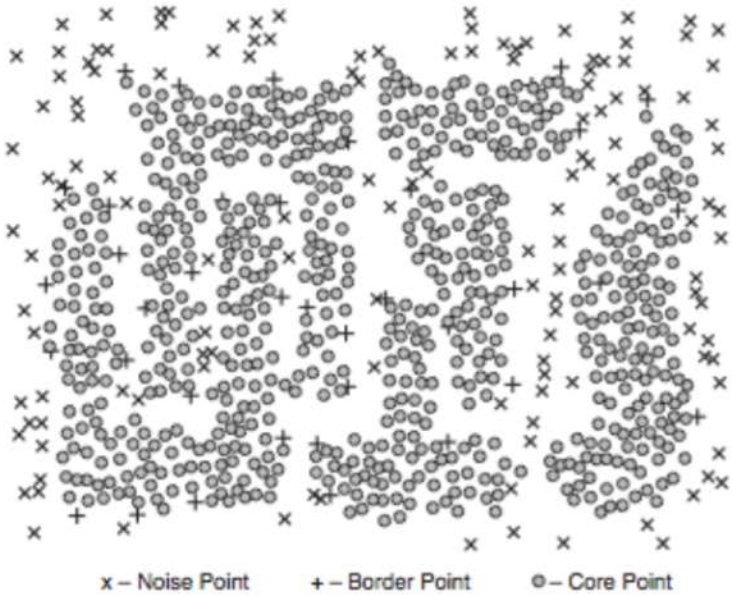
1: Пометить все точки, как основные, пограничные или шумовые.

2: Отбросить точки шума.

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

DBSCAN

(a) Clusters found by DBSCAN.



(b) Core, border, and noise points.

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

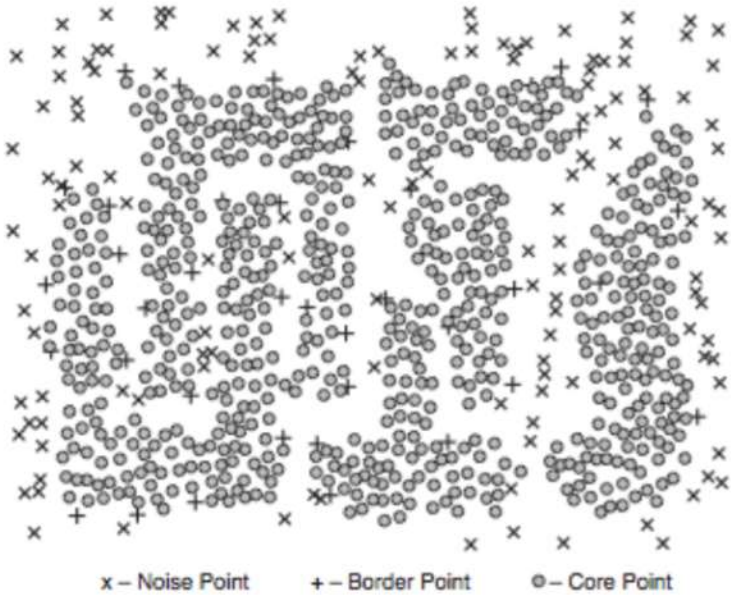
2: Отбросить точки шума.

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

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

DBSCAN

(a) Clusters found by DBSCAN.



(b) Core, border, and noise points.

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

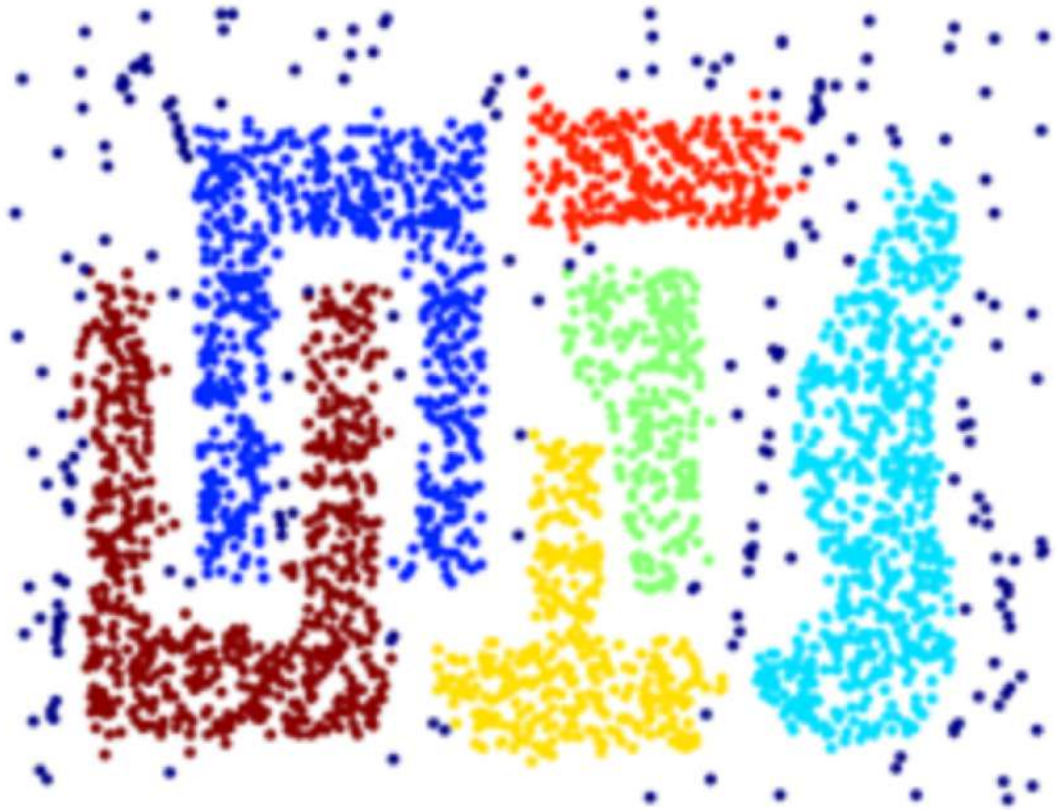
2: Отбросить точки шума.

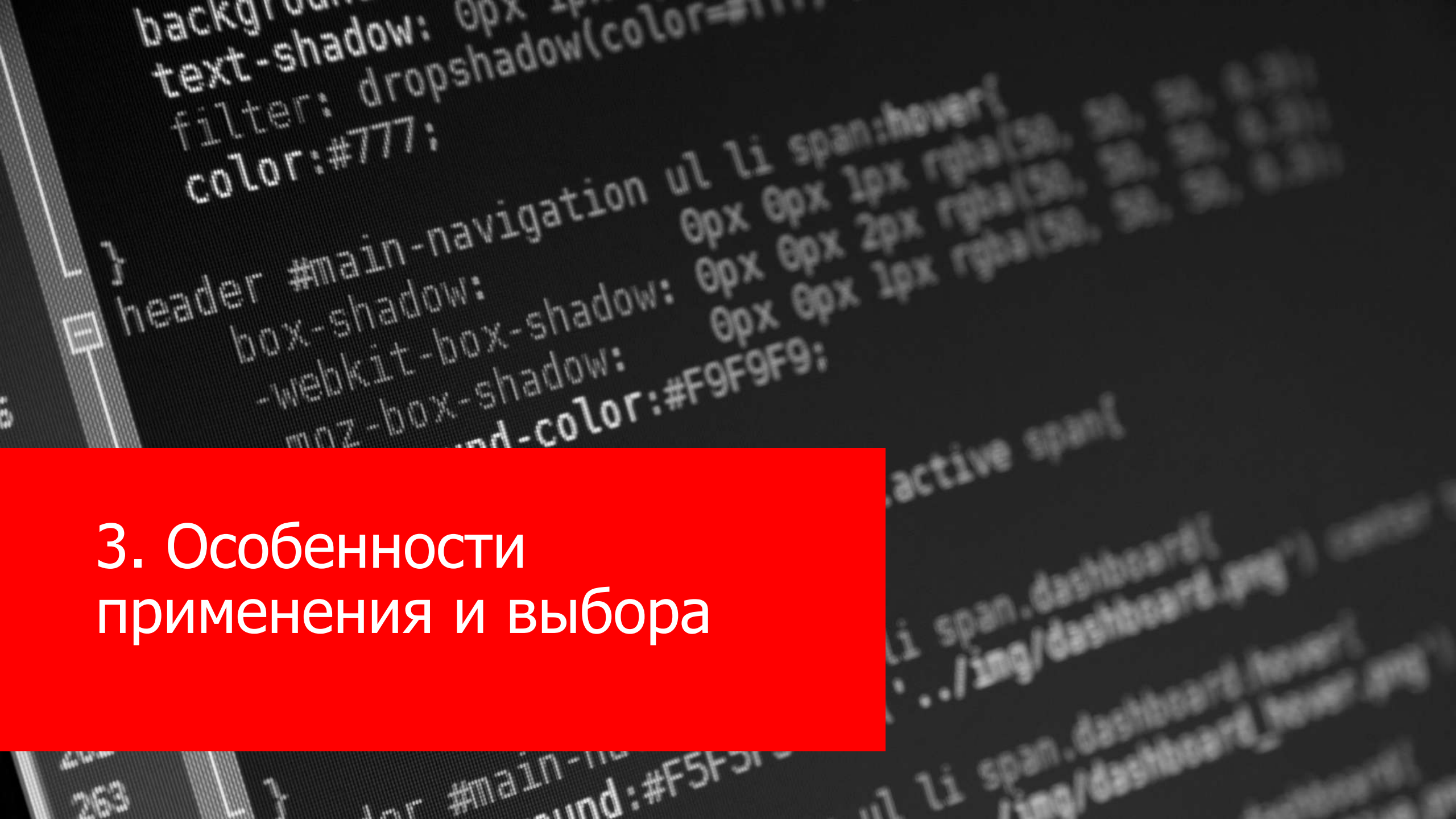
3: Соединить все основные точки, находящиеся на расстоянии Eps радиуса одна от другой.

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





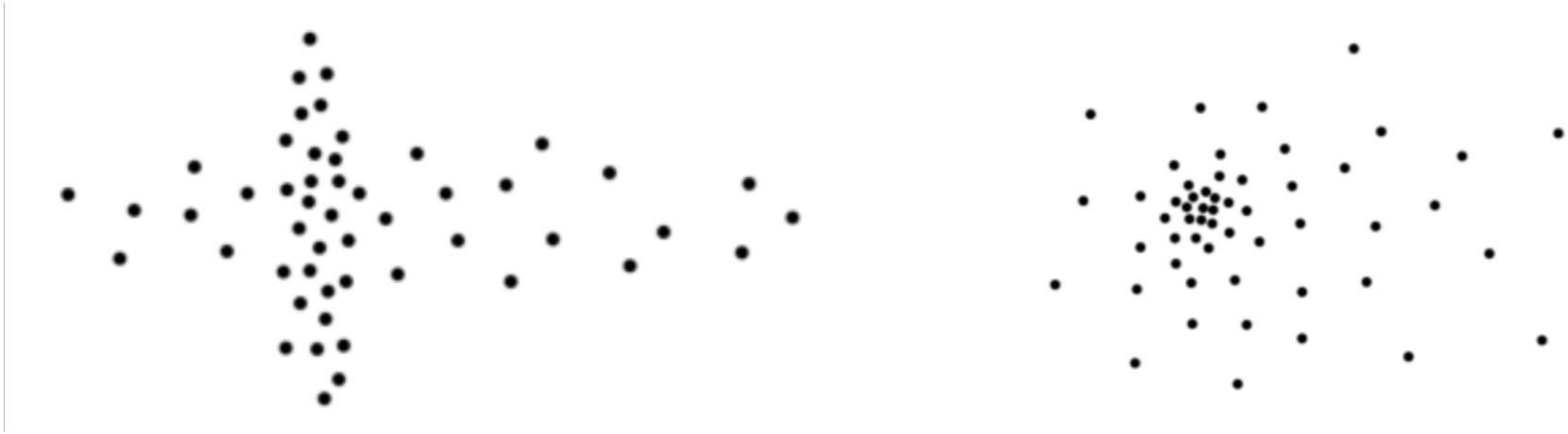
3. Особенности применения и выбора

Зачем нужны разные алгоритмы кластеризации

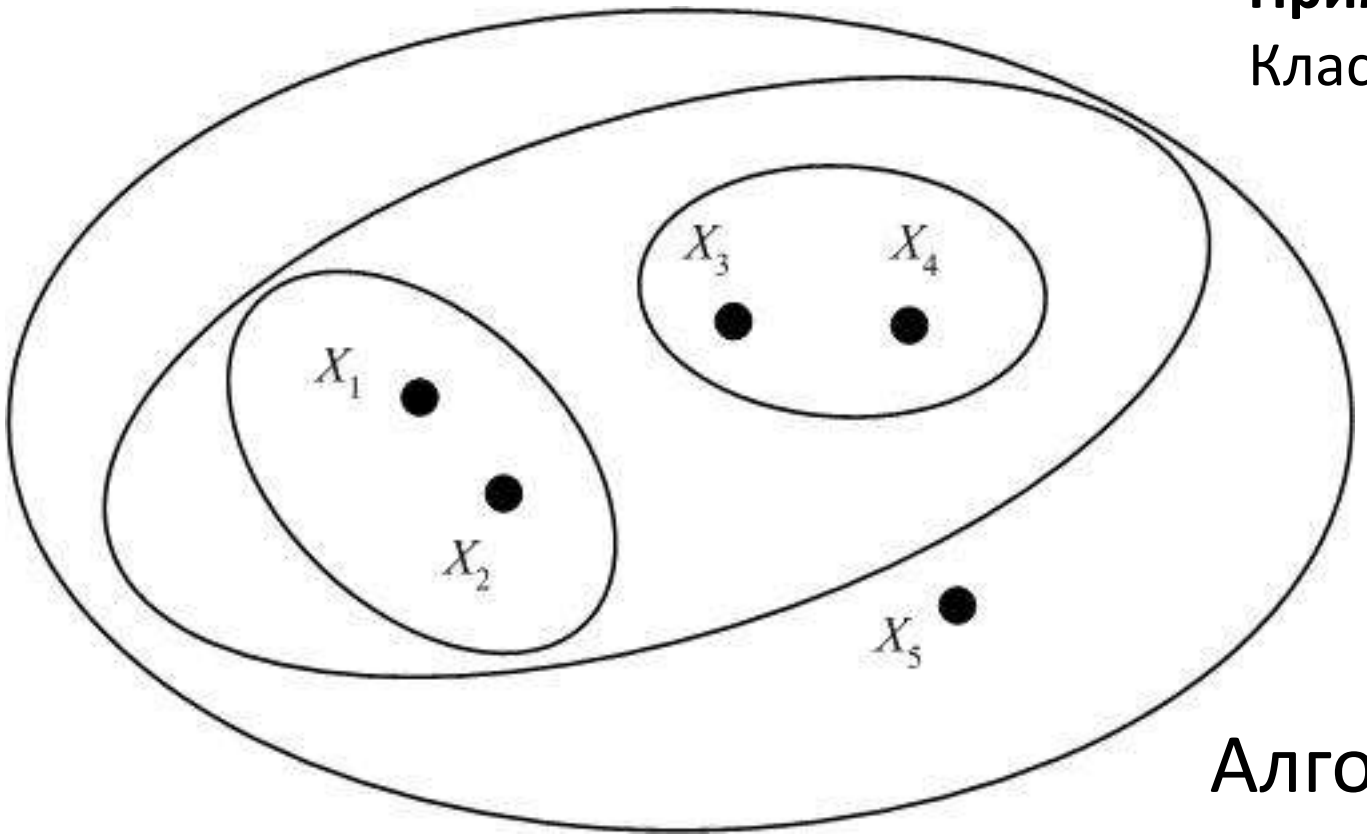
- Каждые данные в чем-то «особенные»
- Каждая задача кластеризации тоже
- В разных задачах кластеризации могут быть отличия:
  - Форма кластеров
  - Необходимость делать кластеры вложенными друг в друга
  - Размер кластеров
  - Кластеризация - основная задача или побочная
  - «Жесткая» или «мягкая» кластеризация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

Форма кластеров

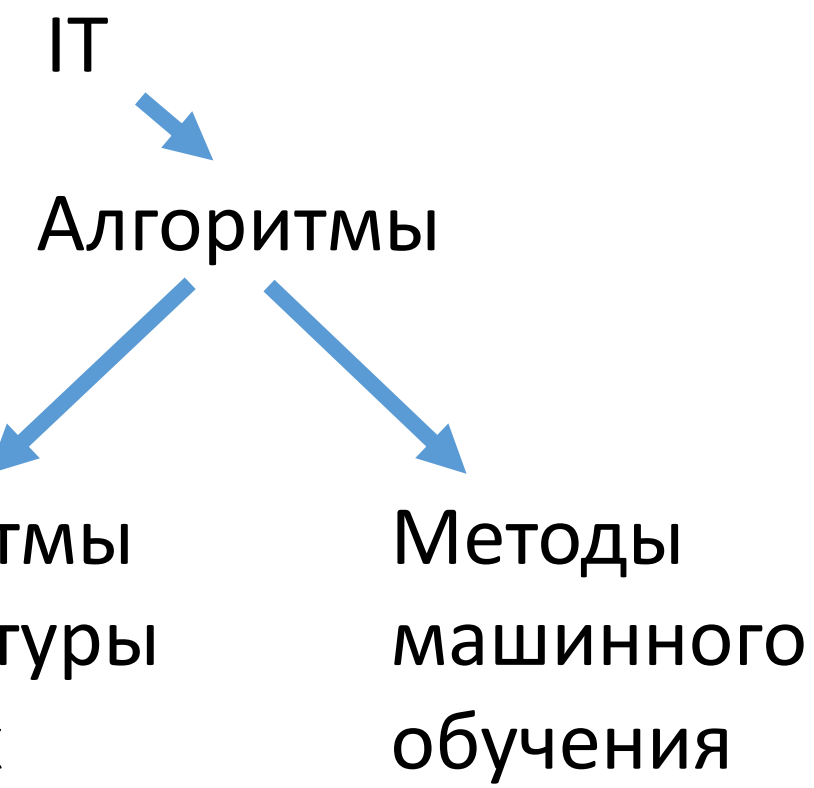
Форма кластеров



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).

Скриншот с сайта Яндекс.Новости (news.yandex.ru)

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

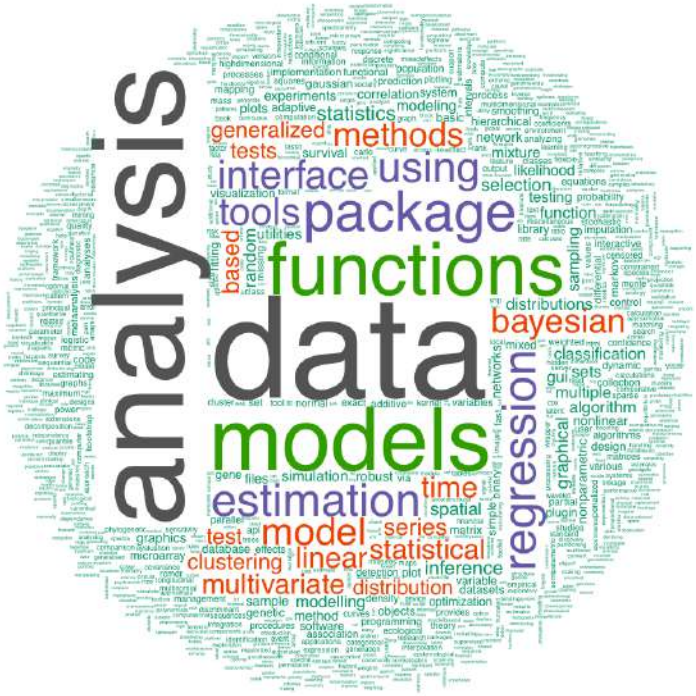
Церемония открытия Олимпиады в Сочи. Онлайн-репортаж

Основная задача или вспомогательная

Кластеризация символов по написанию для улучшения распознавания

«Жесткая» и «мягкая» кластеризации

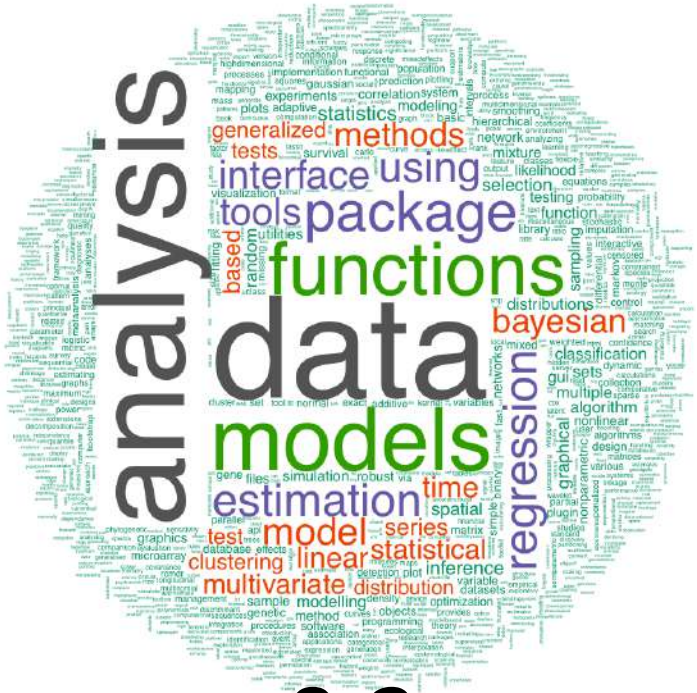
Кластеризация для выделения «тем»



«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»

0.2



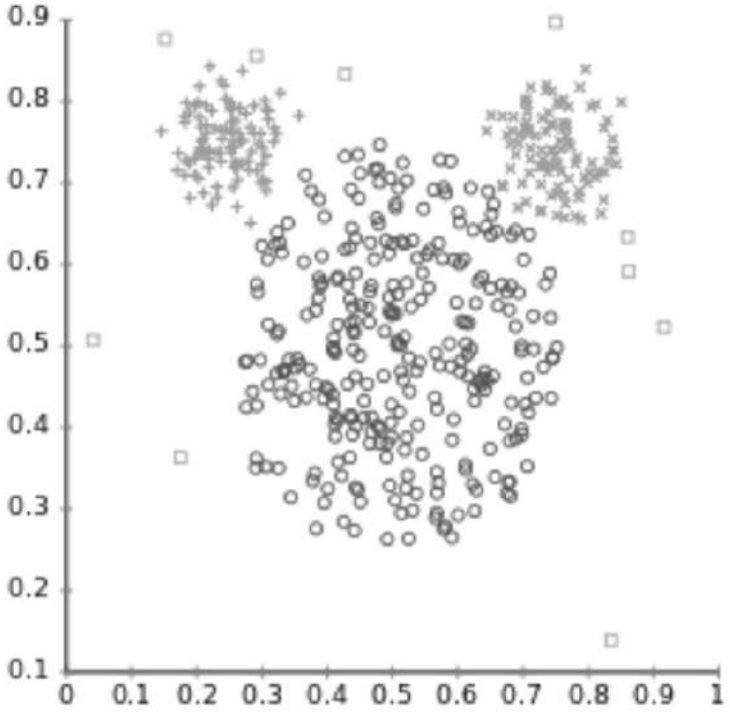
0.3

0.5

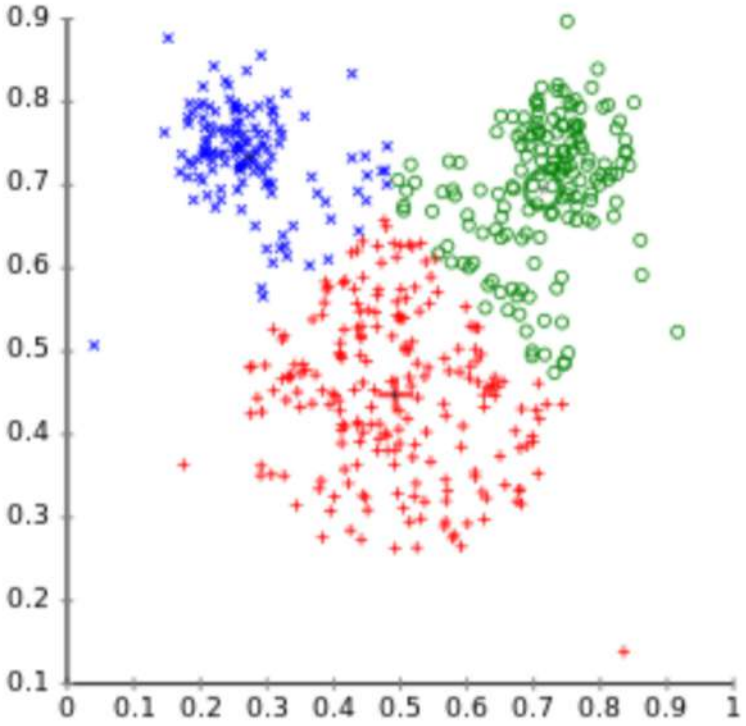
Резюме: чем могут отличаться задачи кластеризации

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

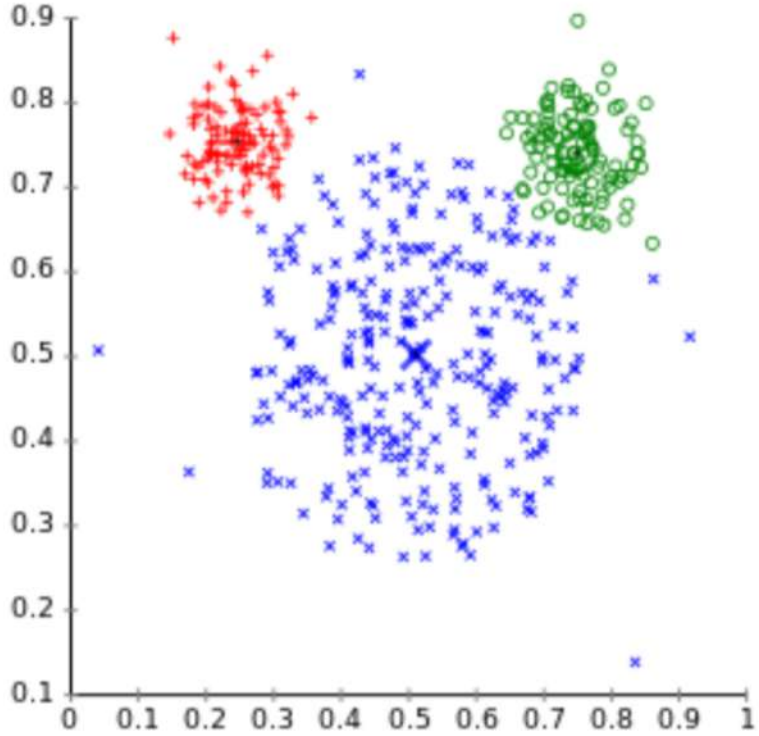
Различия в результатах работы методов



Исходная выборка
("Mouse" dataset)



Метод k средних
(K-Means)



EM-алгоритм

Алгоритмы

Рассмотренные нами:

- K-средних
- EM-алгоритм
- Агломеративная иерархическая кластеризация
- DBSCAN

Алгоритмы

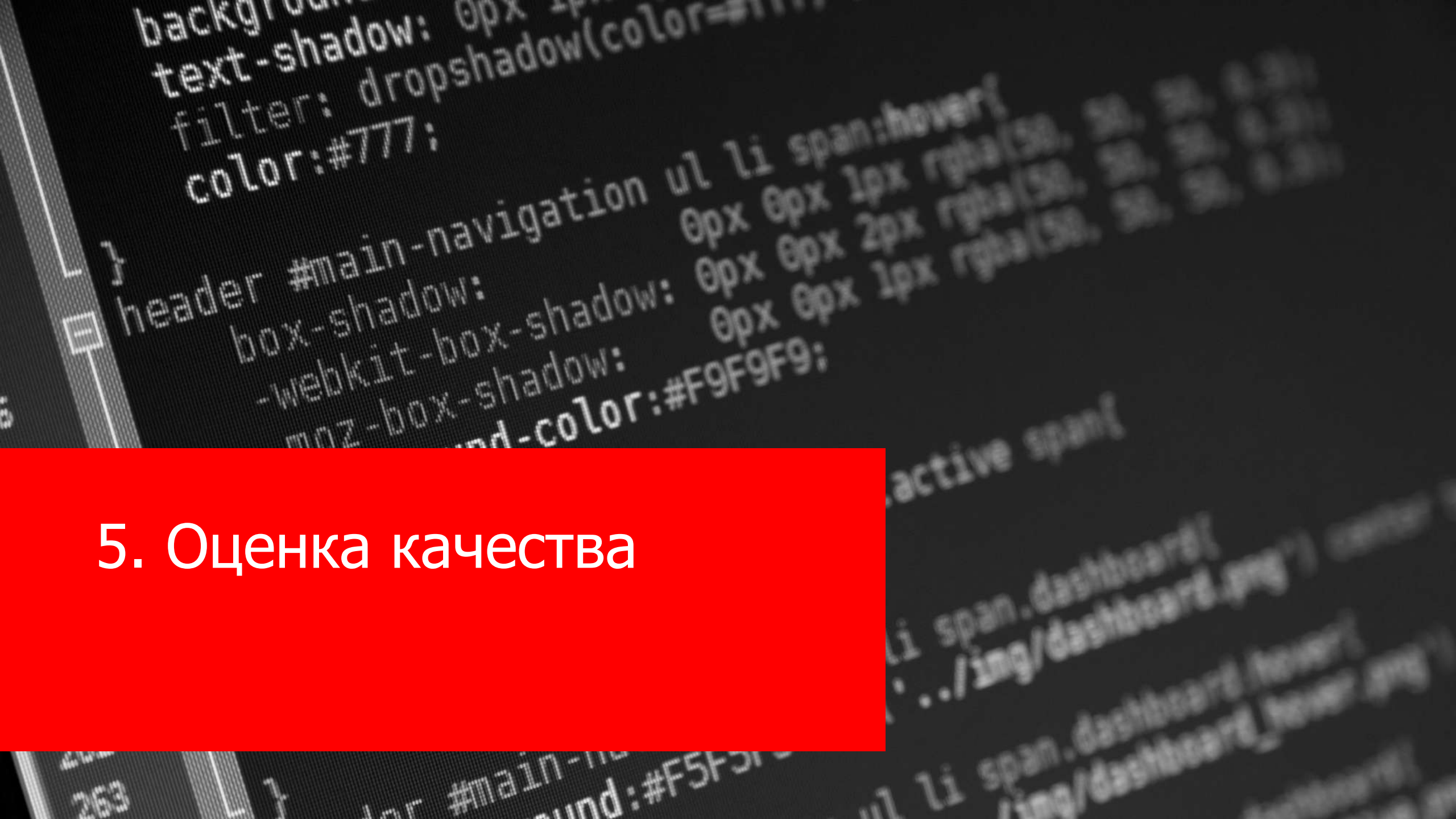
Рассмотренные нами:

- K-средних
- EM-алгоритм
- Агломеративная иерархическая кластеризация
- DBSCAN

В scikit-learn:

KMeans, MiniBatchKMeans, GaussianMixture, AgglomerativeClustering, Ward, DBSCAN, MeanShift, AffinityPropagation, SpectralClustering, Birch

Метод	Параметры	Масштабируемость	Use-case	Геометрия
KMeans	Число кластеров	Очень много объектов (MiniBatch), среднее число кластеров	Выпуклые, примерно одинаковые кластеры	Евклидово расстояние
GaussianMixture	Веса, векторы средних, матрицы ковариаций	-	Восстановление плотности, выпуклые кластеры	Обобщение евклидовой метрики (с весами)
Agglomerative Clustering	Число кластеров, linkage, метрика	Много объектов и много кластеров	Много кластеров, нужно задавать метрику/близость (например, косинусную)	Любая метрика/функция близости, для евклидовой - Ward
DBSCAN	Радиус окрестности, число соседей	Много объектов, среднее число кластеров	Неравные невыпуклые кластеры, выбросы,	Евклидово расстояние


5. Оценка качества

Среднее внутрикластерное расстояние

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

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

Среднее межкластерное расстояние

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

$$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to \max$$

Комбинируем функционалы

$$F_0 = \frac{\sum\limits_{i<j}[y_i = y_j]\,\rho(x_i, x_j)}{\sum\limits_{i<j}[y_i = y_j]} \qquad F_1 = \frac{\sum\limits_{i<j}[y_i \neq y_j]\,\rho(x_i, x_j)}{\sum\limits_{i<j}[y_i \neq y_j]}$$

$$F_0 / F_1 \to \min$$

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:\ y_i = y} \rho^2(x_i, \mu_y) \qquad \Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu)$$

$$\Phi_0 / \Phi_1 \to \min$$

Коэффициент силуэта

- **a**: Среднее расстояние от данного объекта до всех других объектов из того же кластера
- **b**: Среднее расстояние от данного объекта до всех объектов из *ближайшего другого кластера*

$$s = \frac{b - a}{max(a, b)}$$

Коэффициент силуэта

- **a**: Среднее расстояние от данного объекта до всех других объектов из того же кластера
- **b**: Среднее расстояние от данного объекта до всех объектов из *ближайшего другого кластера*

$$s = \frac{b - a}{max(a, b)}$$

Коэффициент силуэта

- **a**: Среднее расстояние от данного объекта до всех других объектов из того же кластера
- **b**: Среднее расстояние от данного объекта до всех объектов из *ближайшего другого кластера*

$$s = \frac{b - a}{max(a, b)}$$