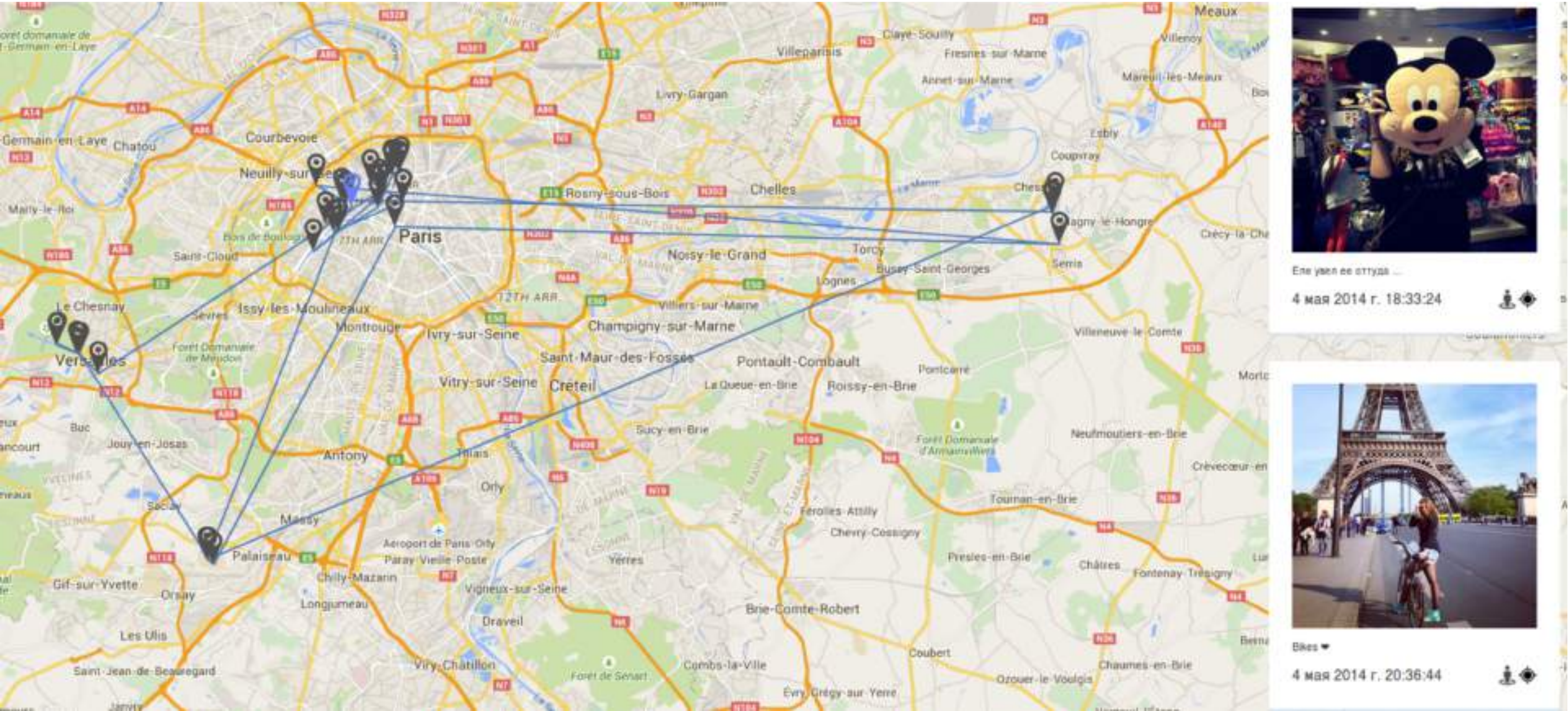
Машинное обучение

Лекция 6. Кластеризация

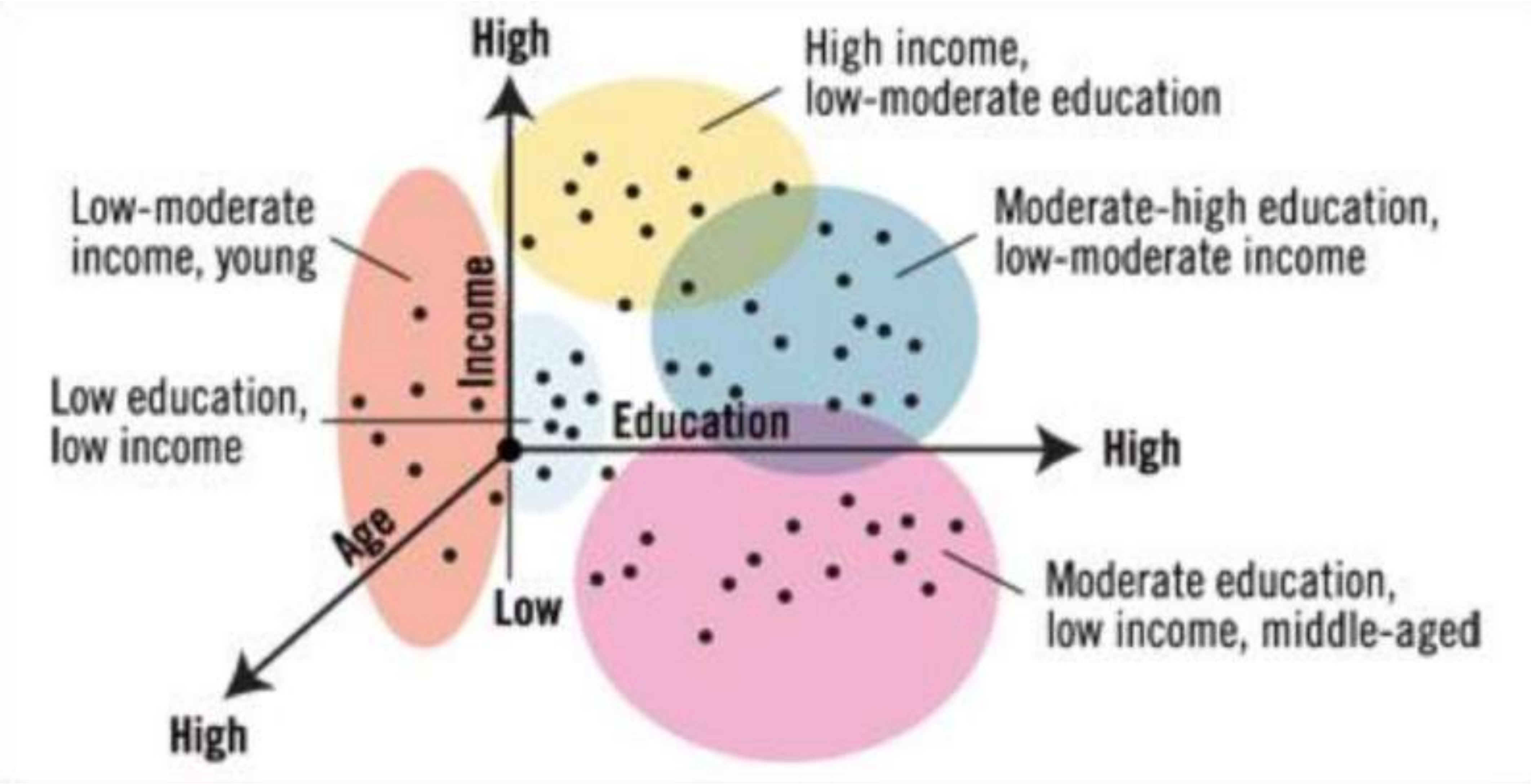
Кантор Виктор

Программный директор академии MADE

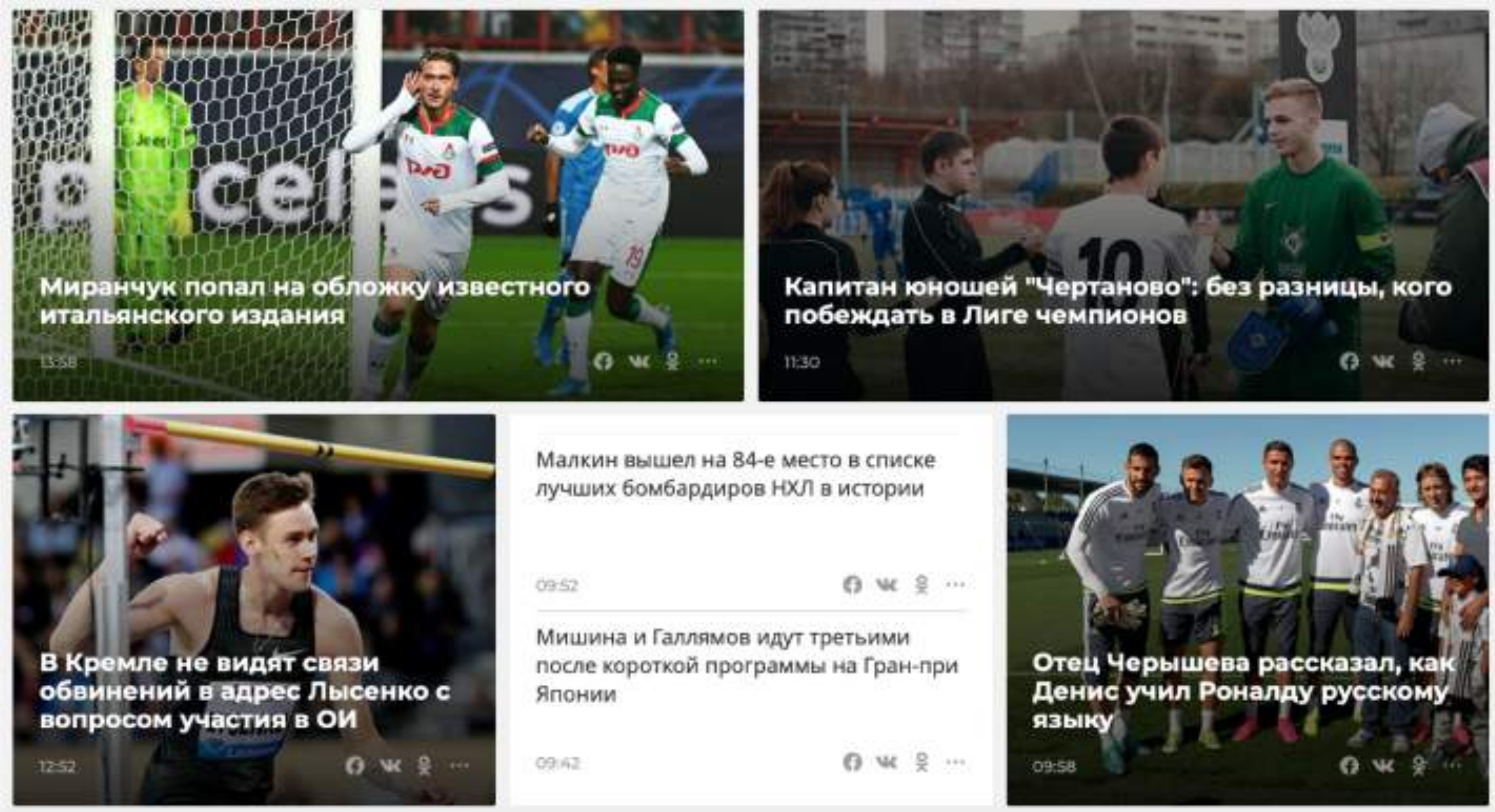
Пример: анализ геоданных



Пример: сегментация рынка



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:

$x_1, \dots, x_l$ - объекты

$y_1, \dots, y_l$ - 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 $a(x)$:

$$a(x) \approx y$$

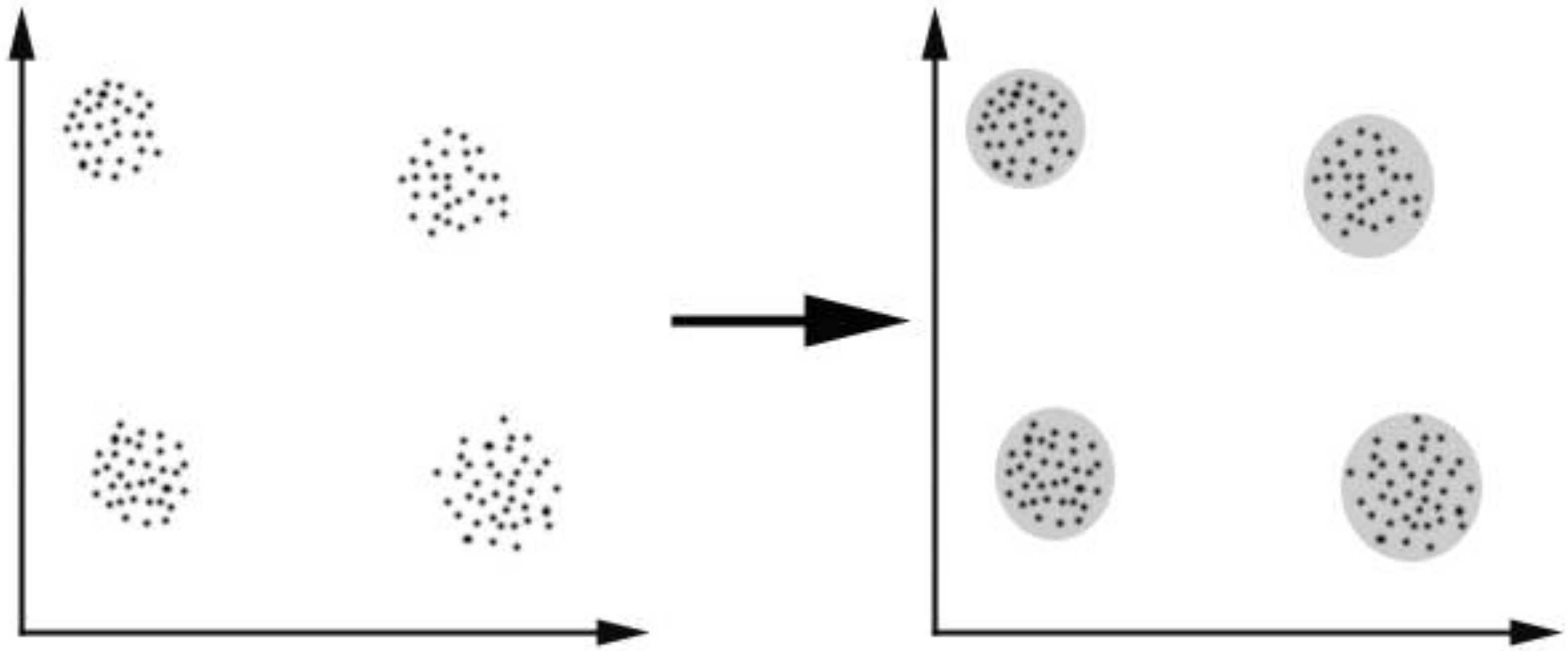
Кластеризация

«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

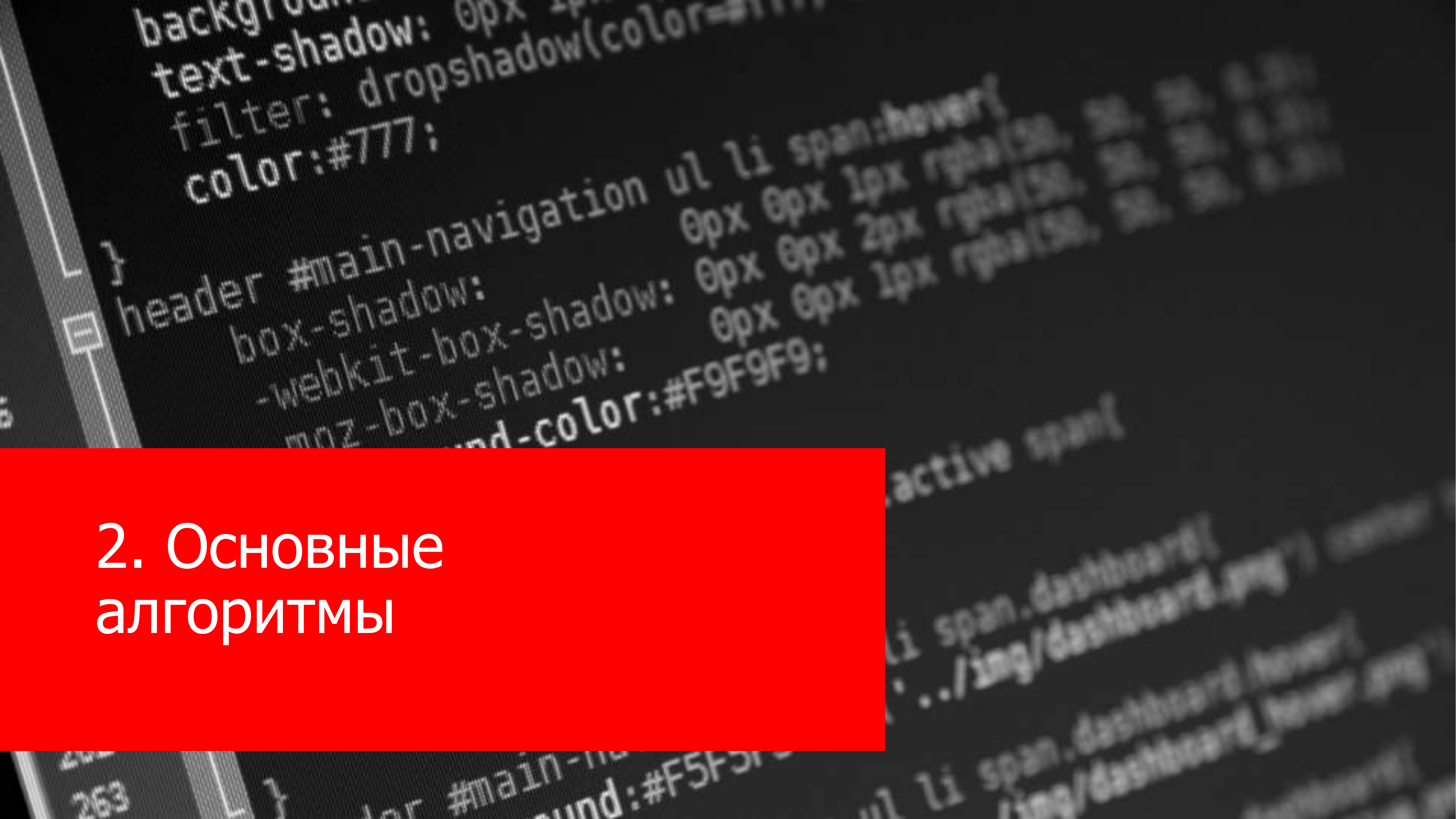
$$F_1 = \frac{\sum\limits_{i<j} [y_i \neq y_j]\, \rho(x_i, x_j)}{\sum\limits_{i<j} [y_i \neq y_j]} \to \max$$

Придумываем метрику качества

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

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

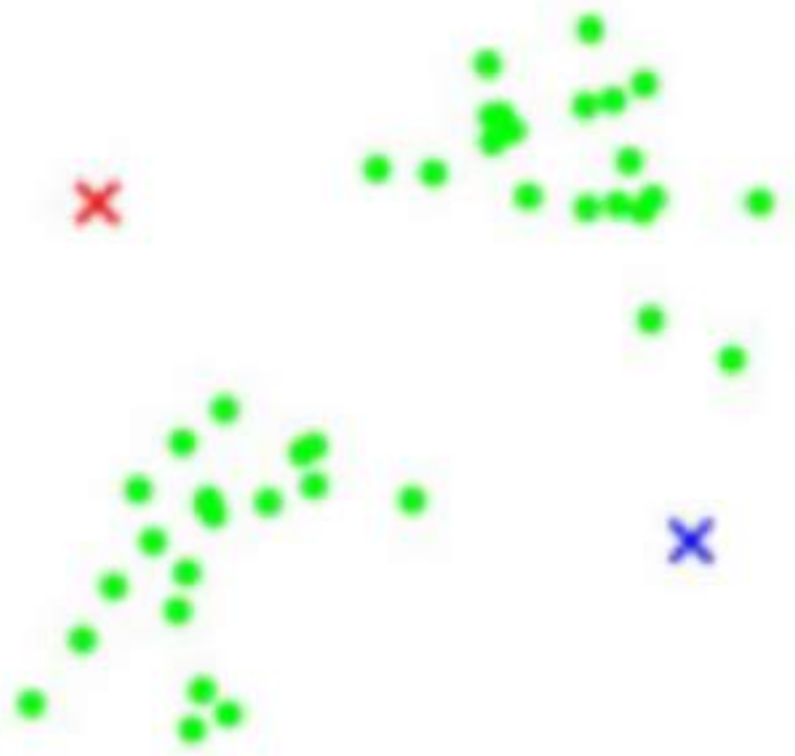
$$F_0 / F_1 \to \min$$



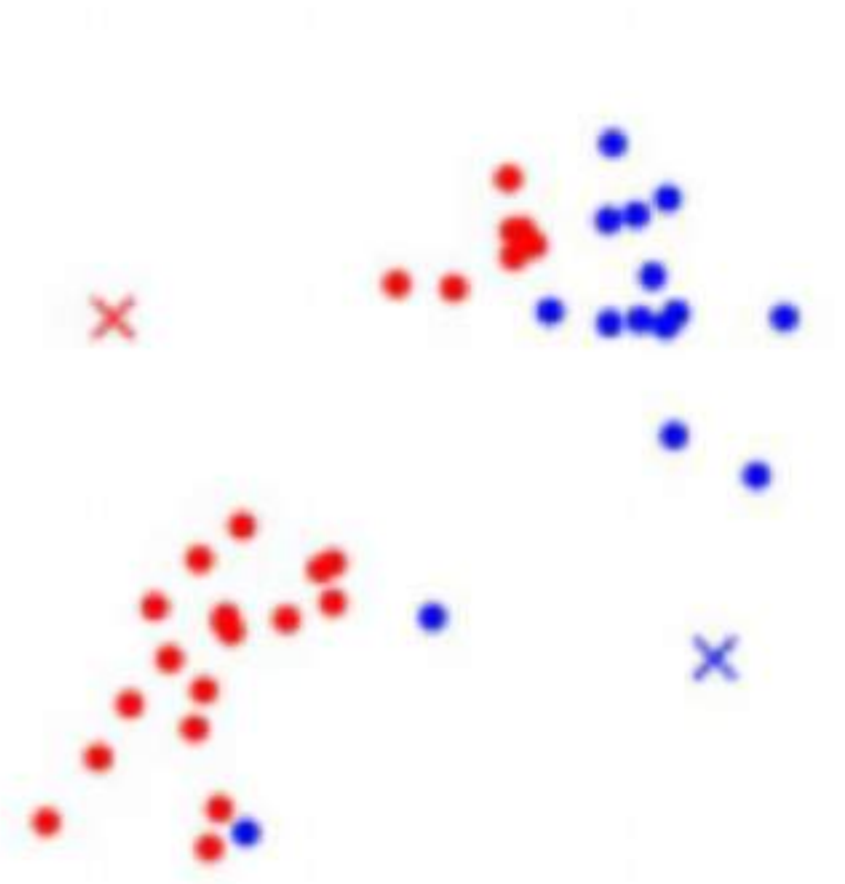
2. Основные алгоритмы

Напоминание: K Means

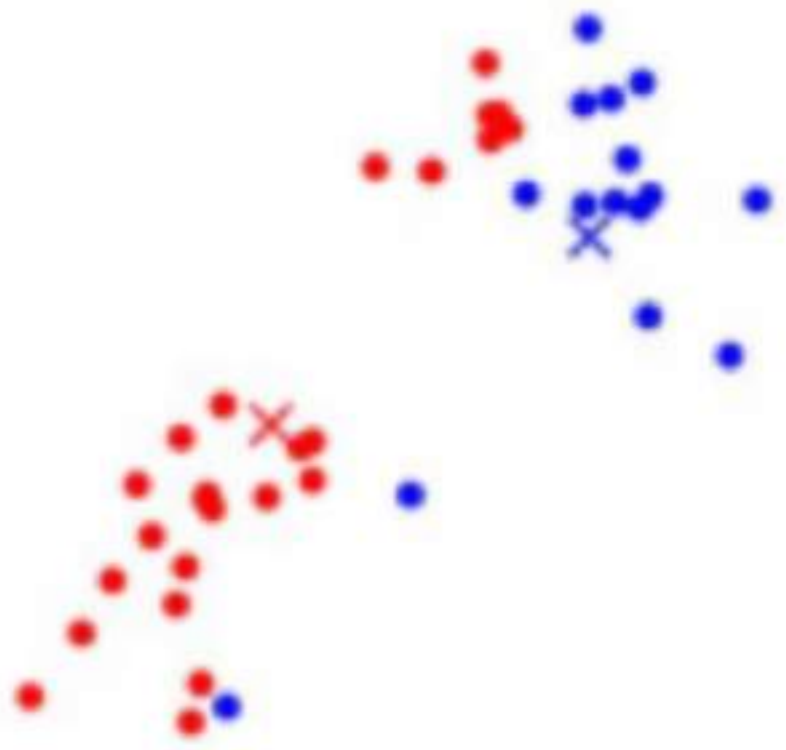
Как работает K Means



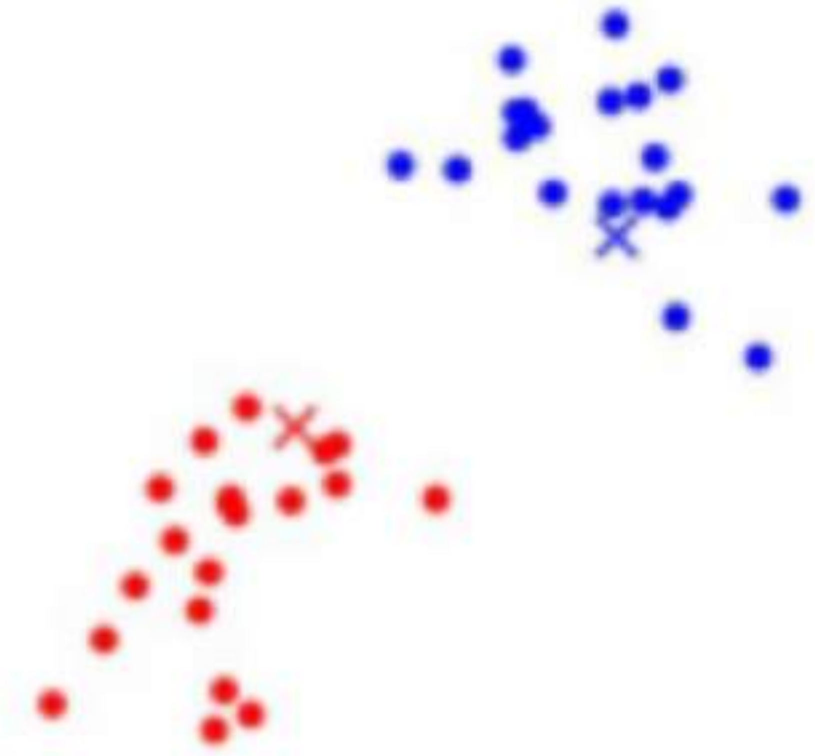
Как работает K Means



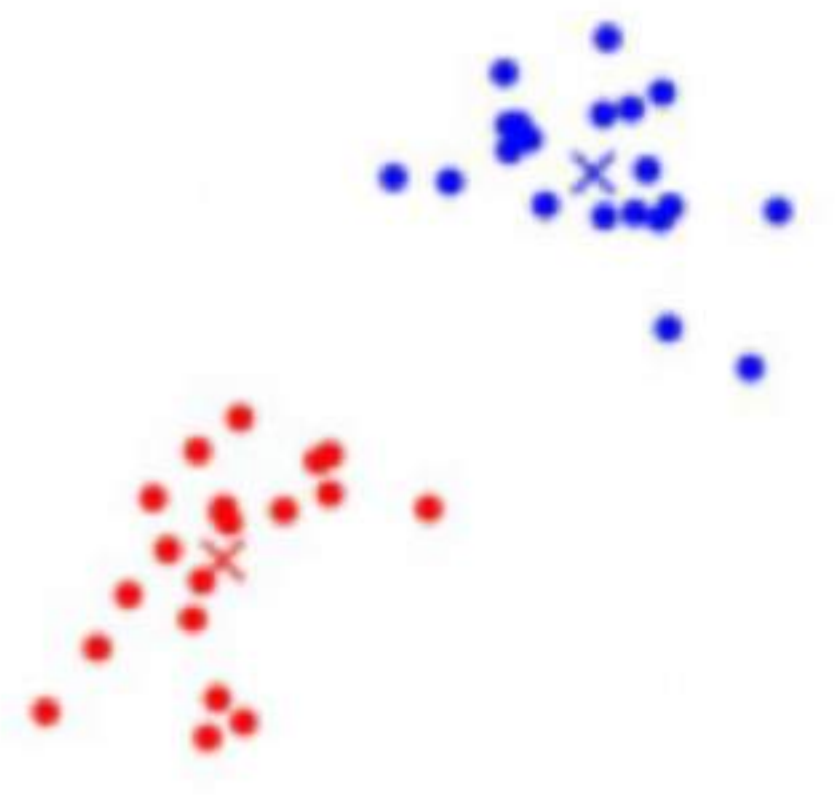
Как работает K Means



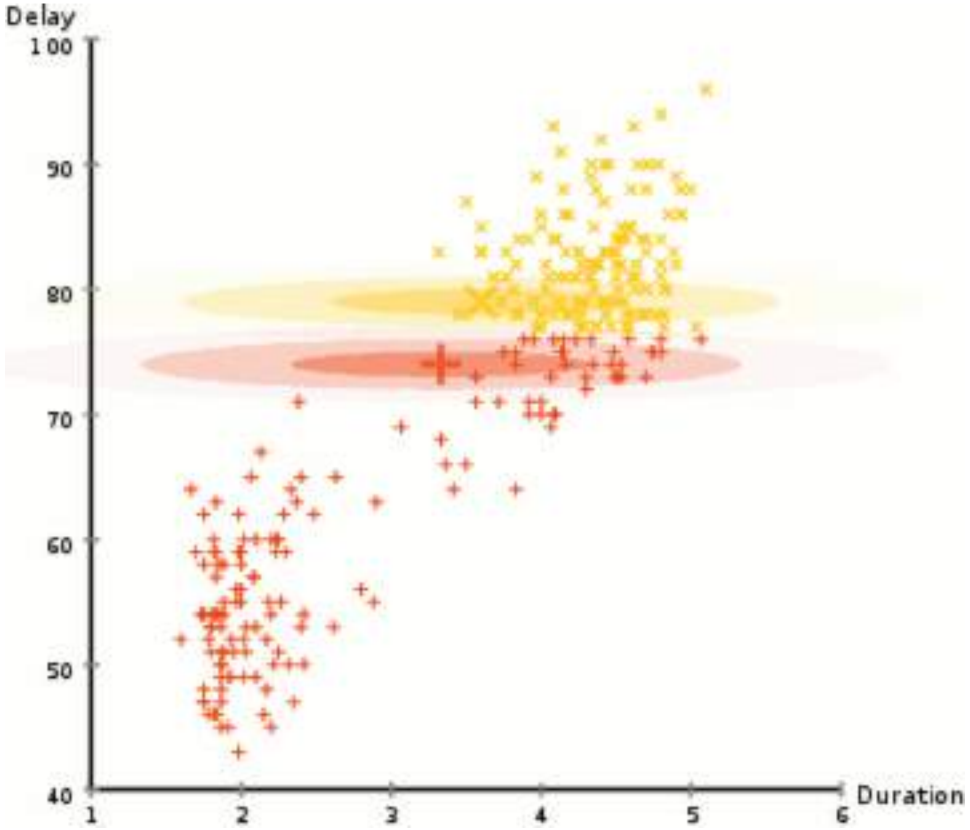
Как работает K Means



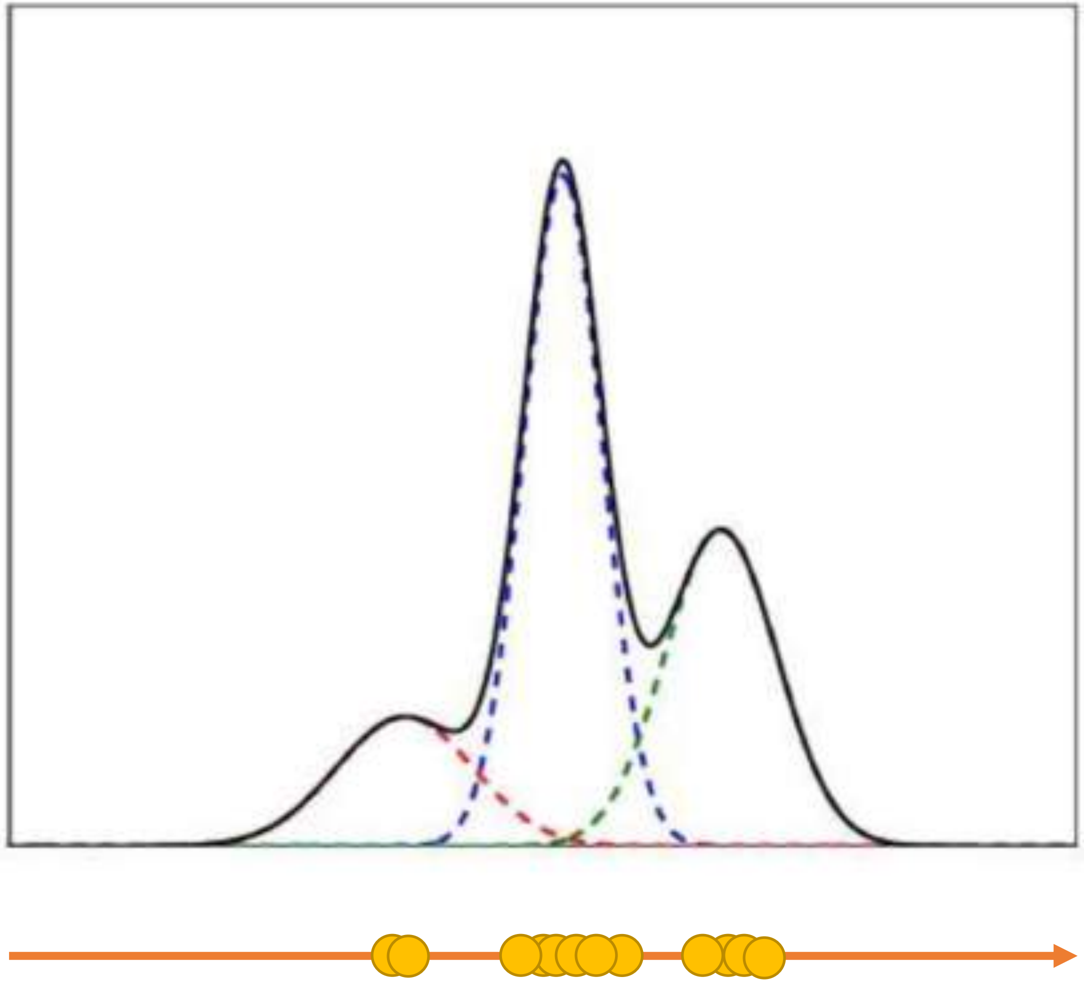
Как работает K Means



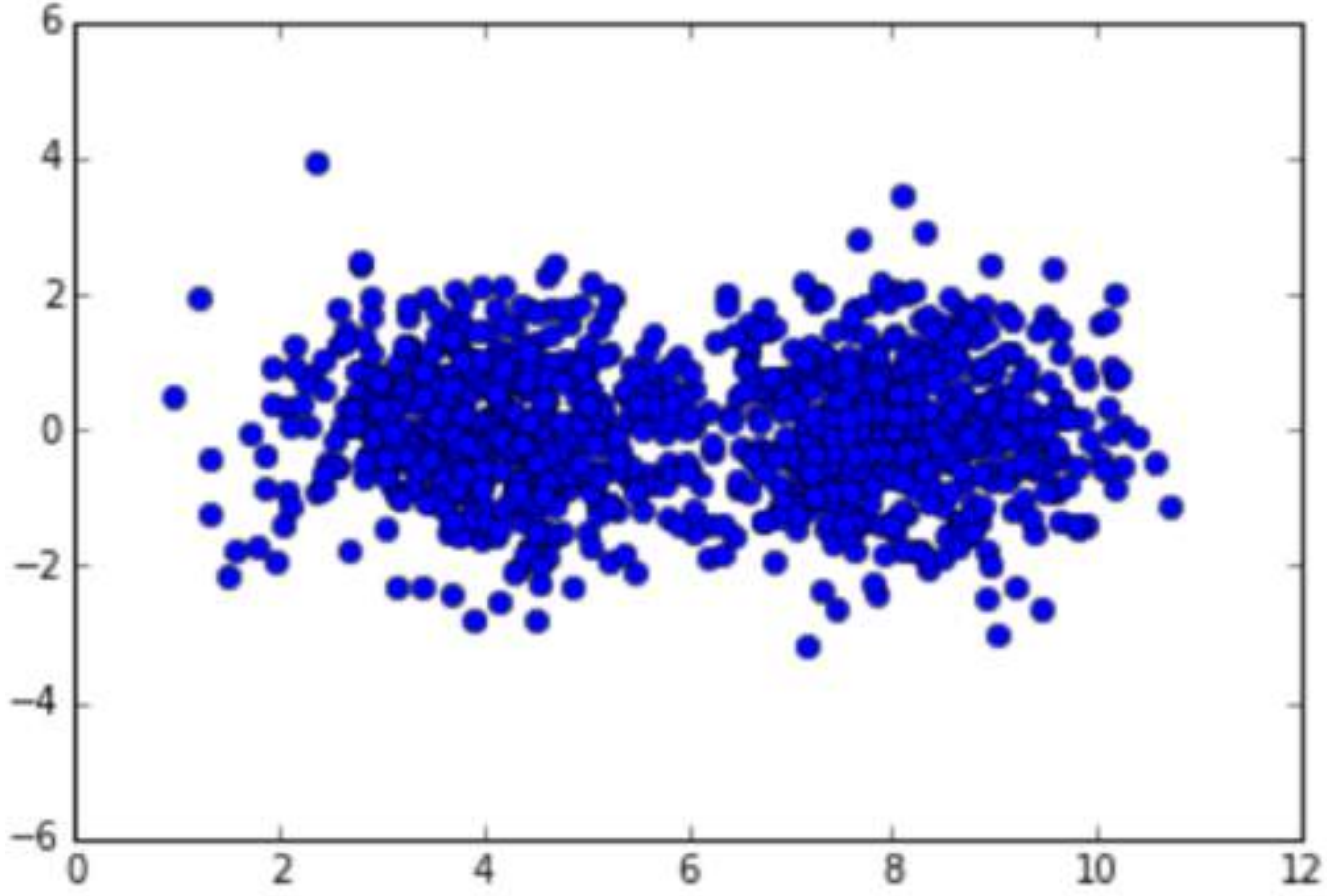
Кластеризация с помощью ЕМ-алгоритма



Как выглядит смесь распределений



Как выглядит смесь распределений



EM-алгоритм

$$p(x) = \sum_{j=1}^{K} w_j p_j(x), \qquad p_j(x) = \varphi(\theta_j; x)$$

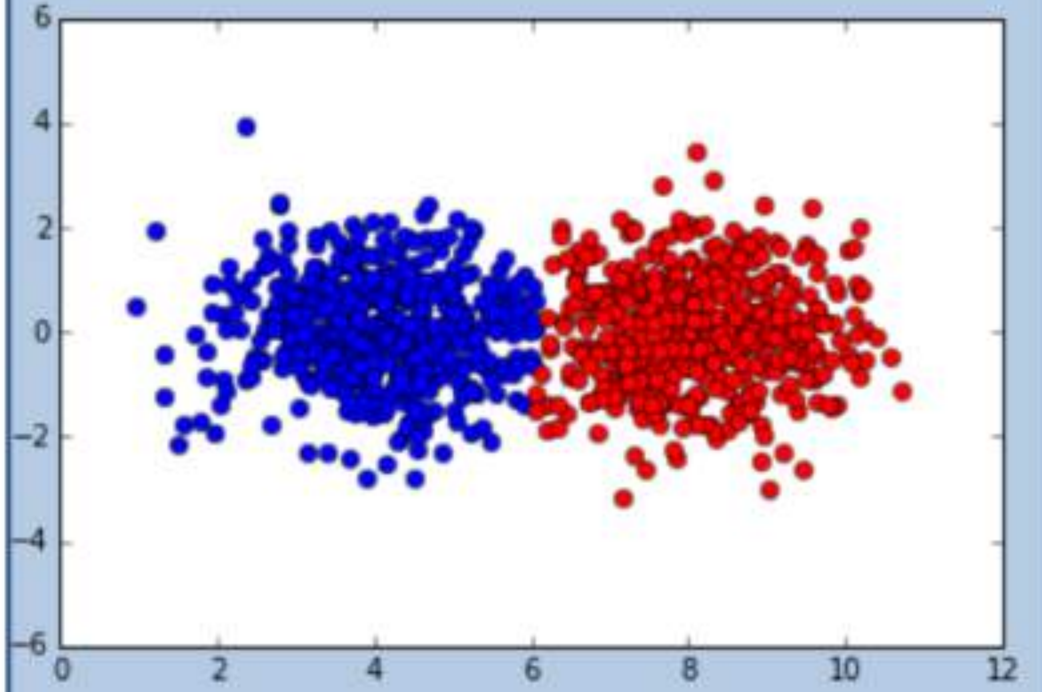
<u>Е-шаг:</u>

$$g_{ji} = p(j|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$$

<u>М-шаг:</u>

$$w_j = \frac{1}{N} \sum_{i=1}^{N} g_{ji} \qquad \theta_j = argmax_\theta \sum_{i=1}^{N} g_{ji} \ln \varphi(\theta; x)$$

Пример: 2 кластера с гауссовской плотностью



Относим x_i к кластеру j, для которого больше $p(j|x_i) = g_{ij}$

$$p(x) = w_1 p_1(x) + w_2 p_2(x)$$

<u>Е-шаг:</u> $g_{ji} = p(j|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$

<u>М-шаг:</u>

$$w_j = \frac{1}{N} \sum_{i=1}^{N} g_{ji}$$

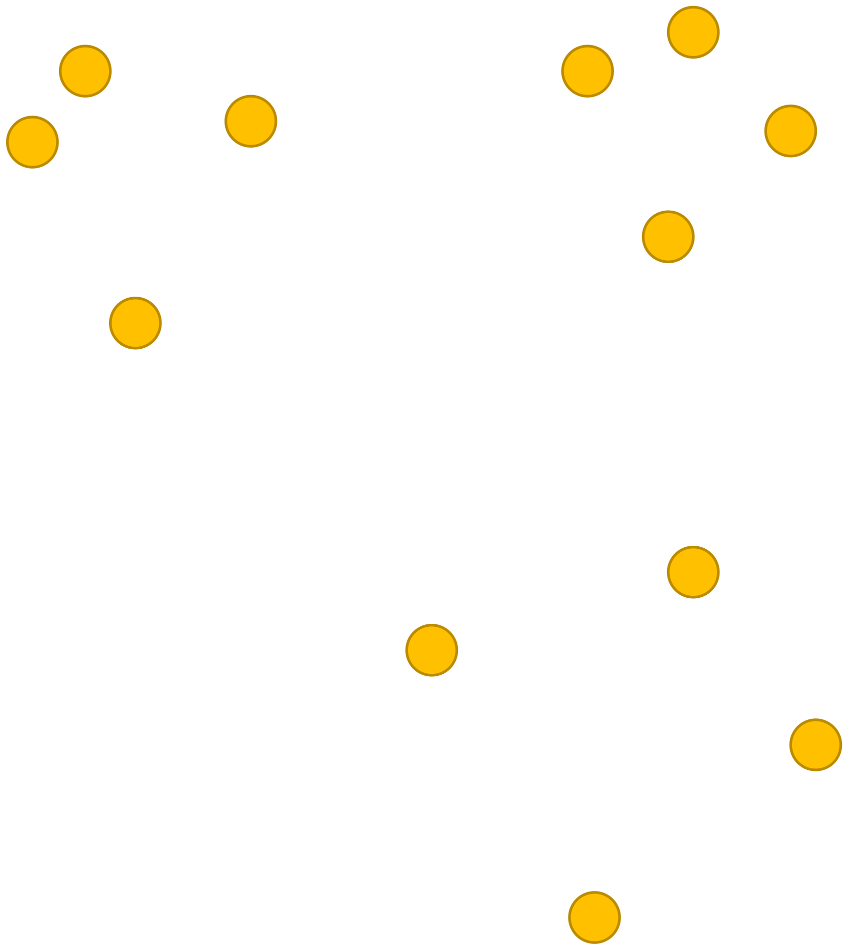
$$\mu_j = \frac{1}{N w_j} \sum_{i=1}^{N} g_{ij} x_i$$

$$\Sigma_j = \frac{1}{N w_j - 1} \sum_{i=1}^{N} g_{ij} (x_i - \mu_j)(x_i - \mu_j)^T$$

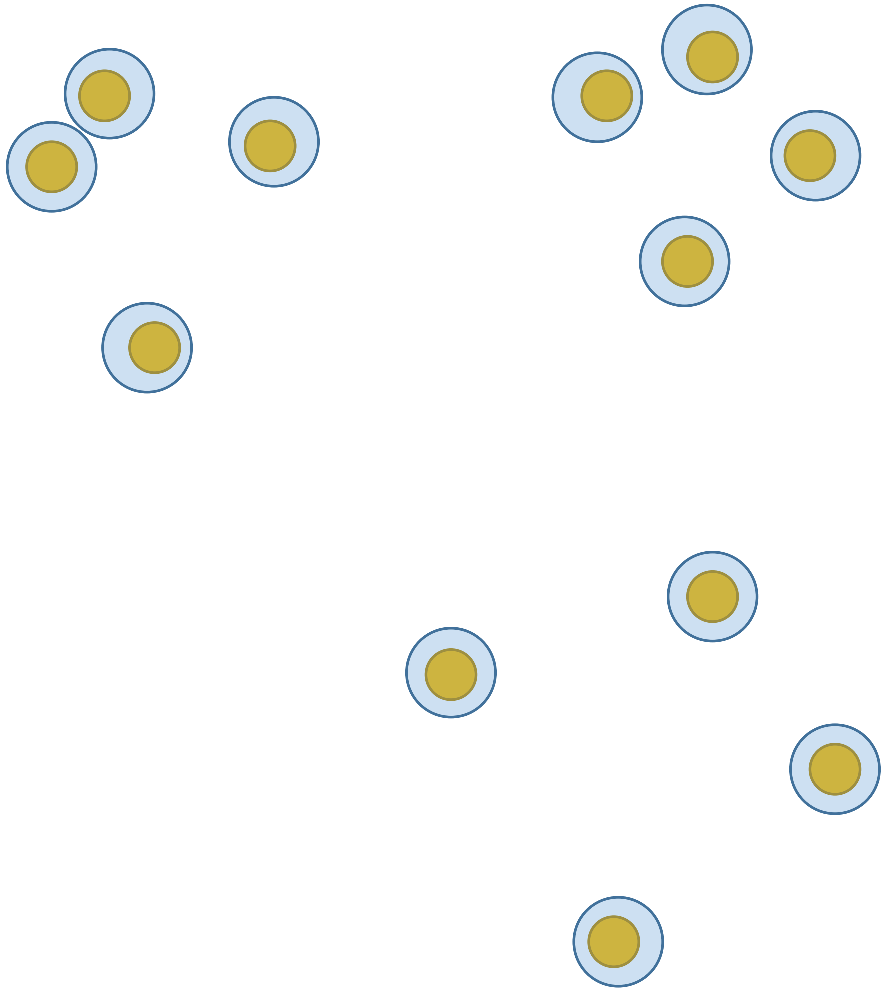
Простое объяснение EM-алгоритма

- Выбираем «скрытые переменные» таким образом, чтобы с ними было проще максимизировать правдоподобие
- E-шаг:
  - Оцениваем скрытые переменные
- M-шаг:
  - Оцениваем $w_1, \dots, w_K$ и $p_1(x), \dots, p_K(x)$, считая скрытые переменные зафиксированными

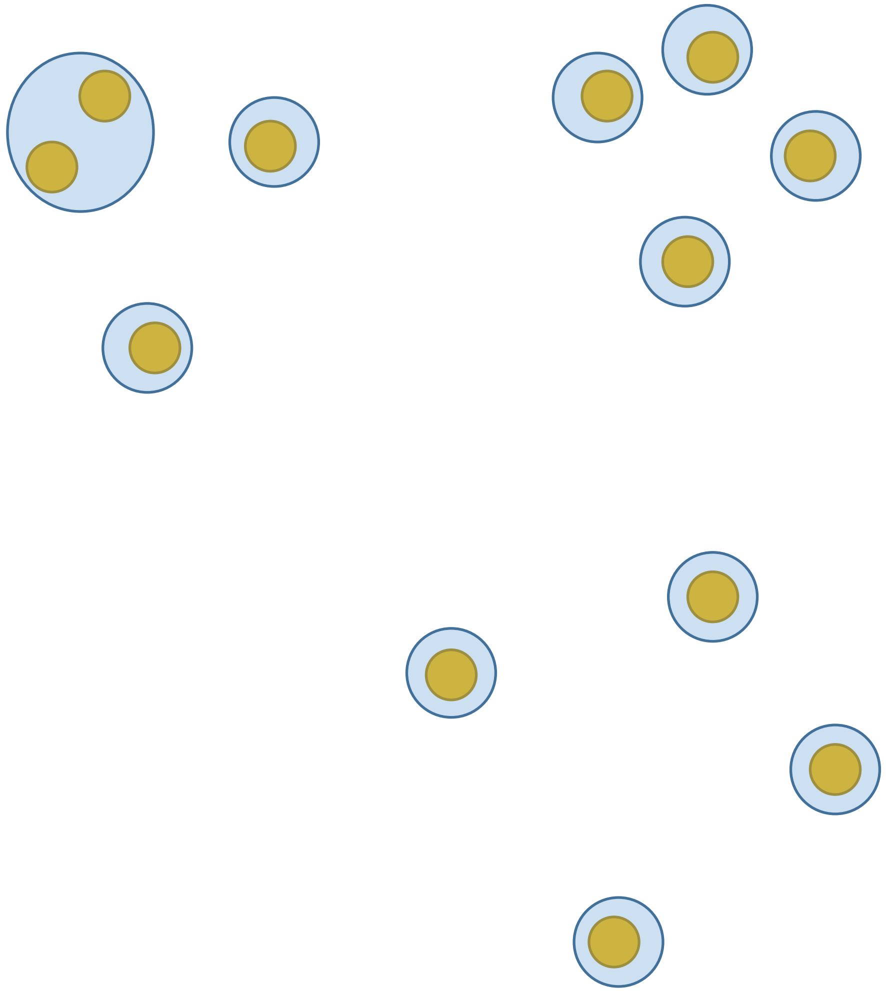
Агломеративная кластеризация



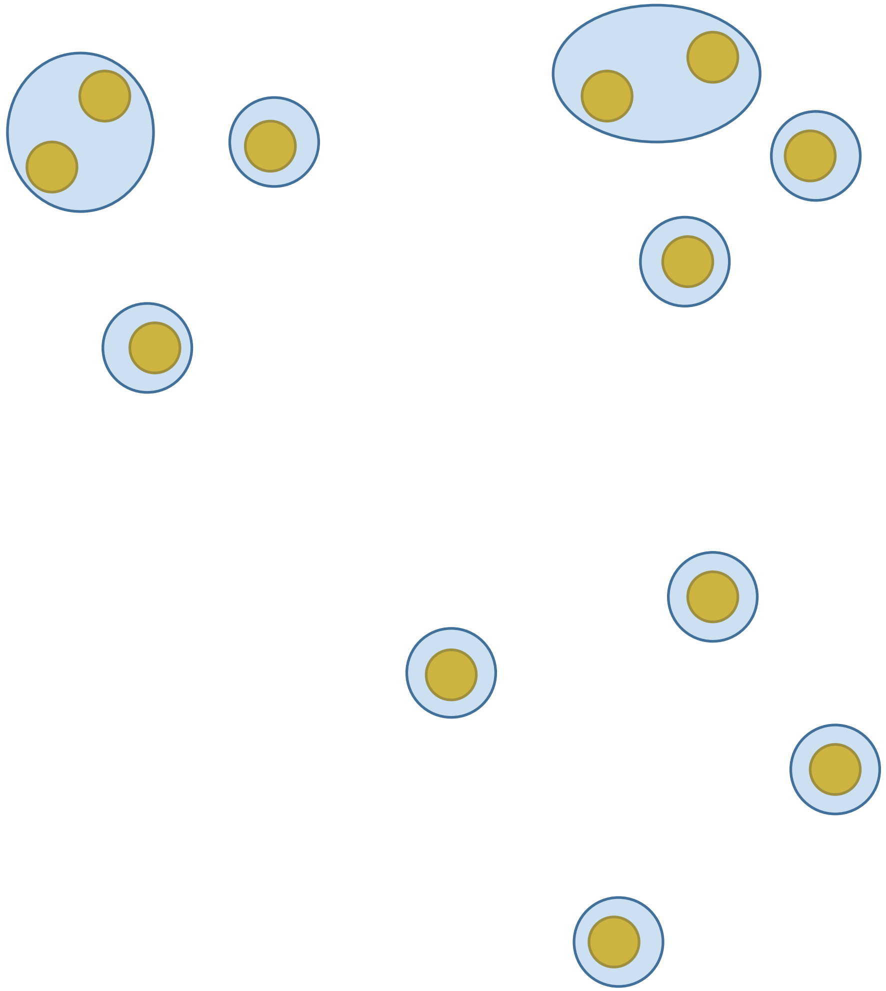
Агломеративная кластеризация



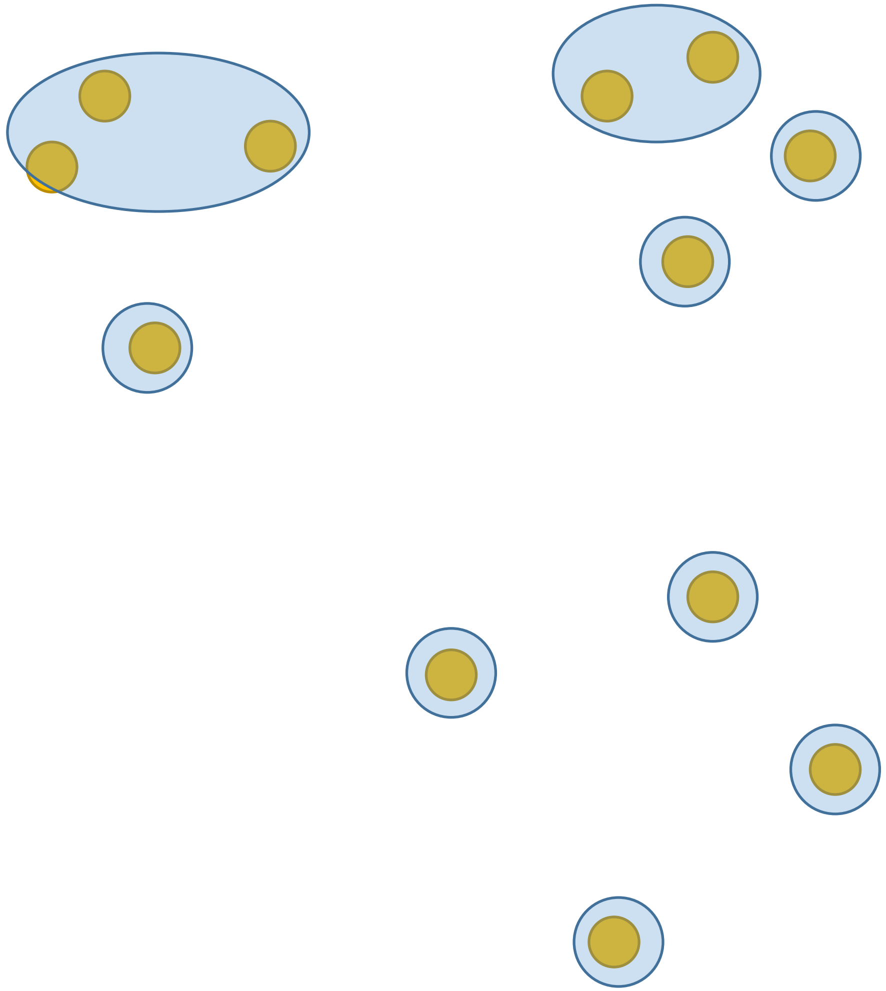
Агломеративная кластеризация



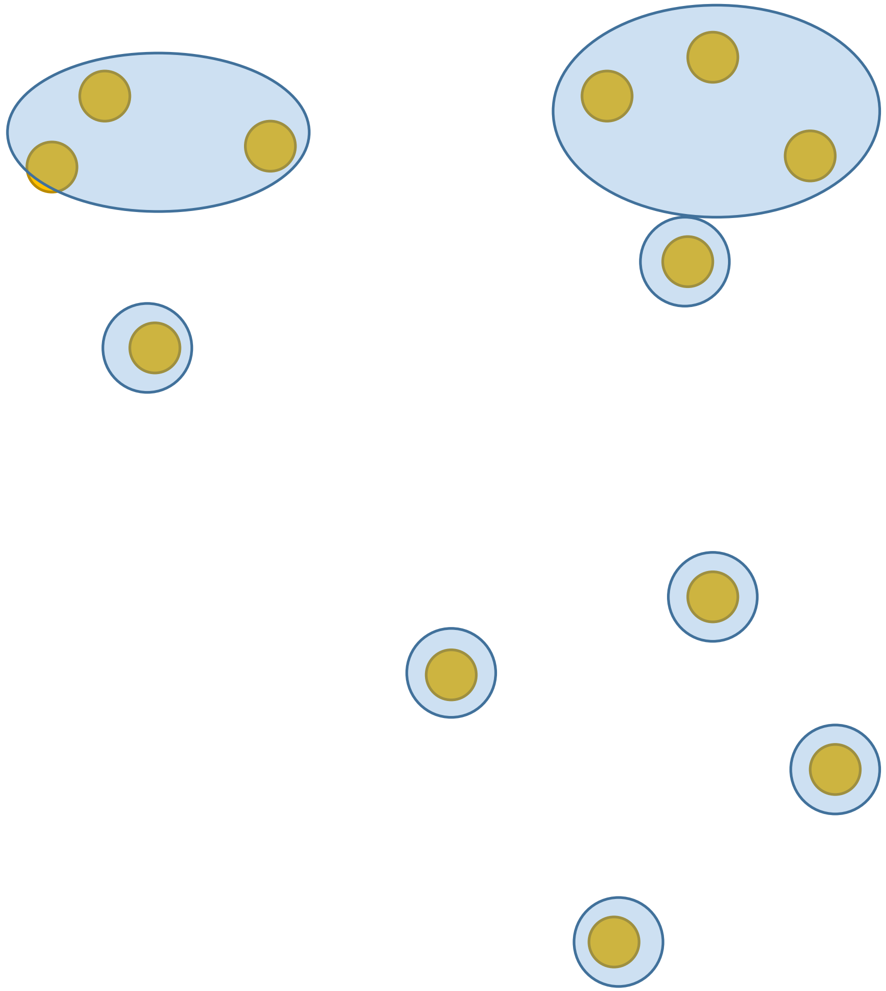
Агломеративная кластеризация



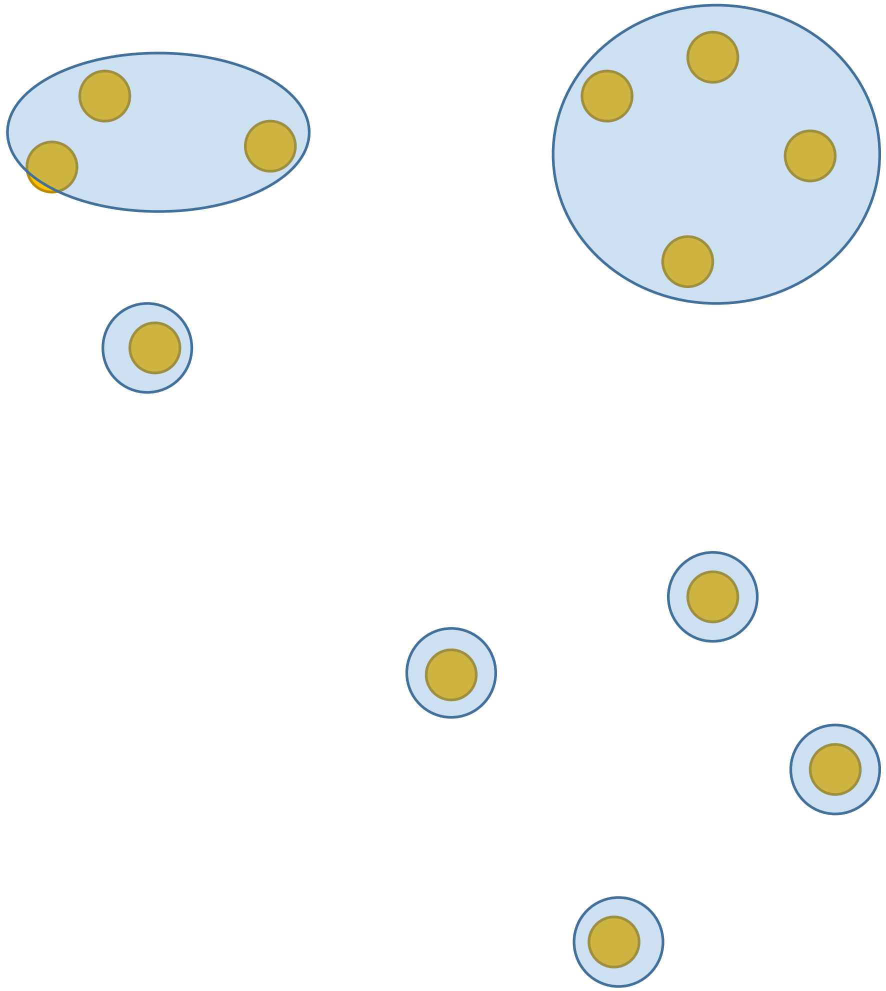
Агломеративная кластеризация



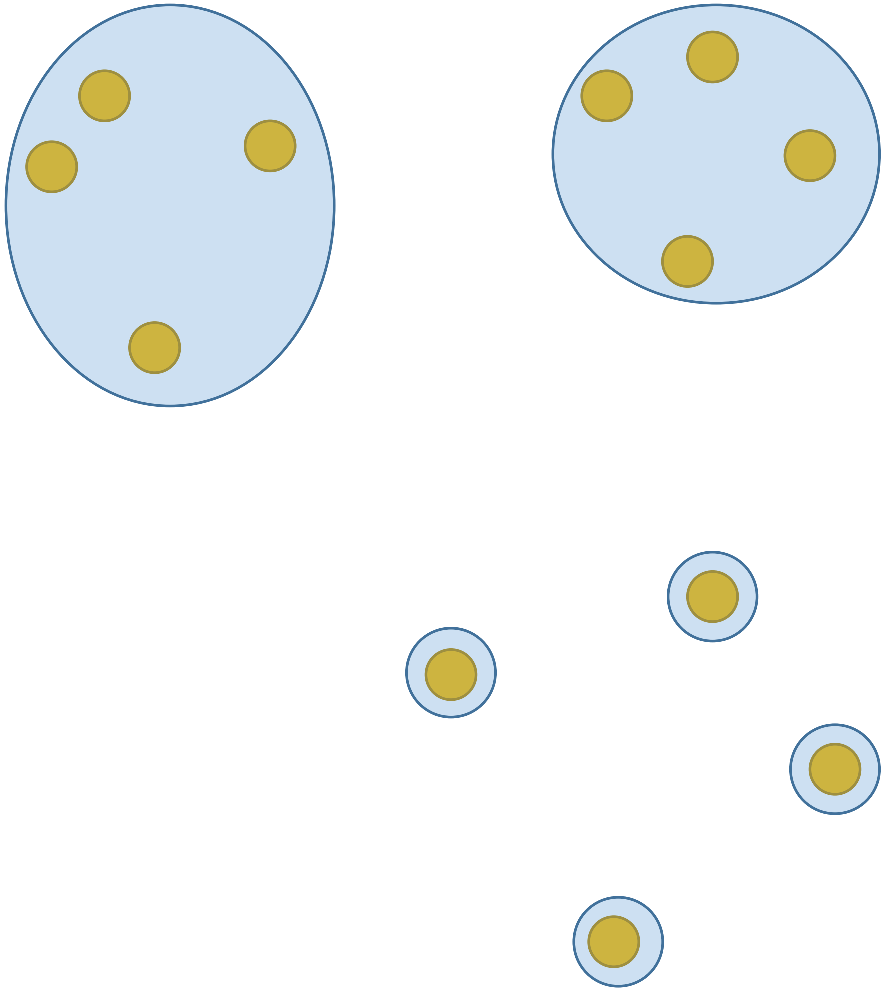
Агломеративная кластеризация



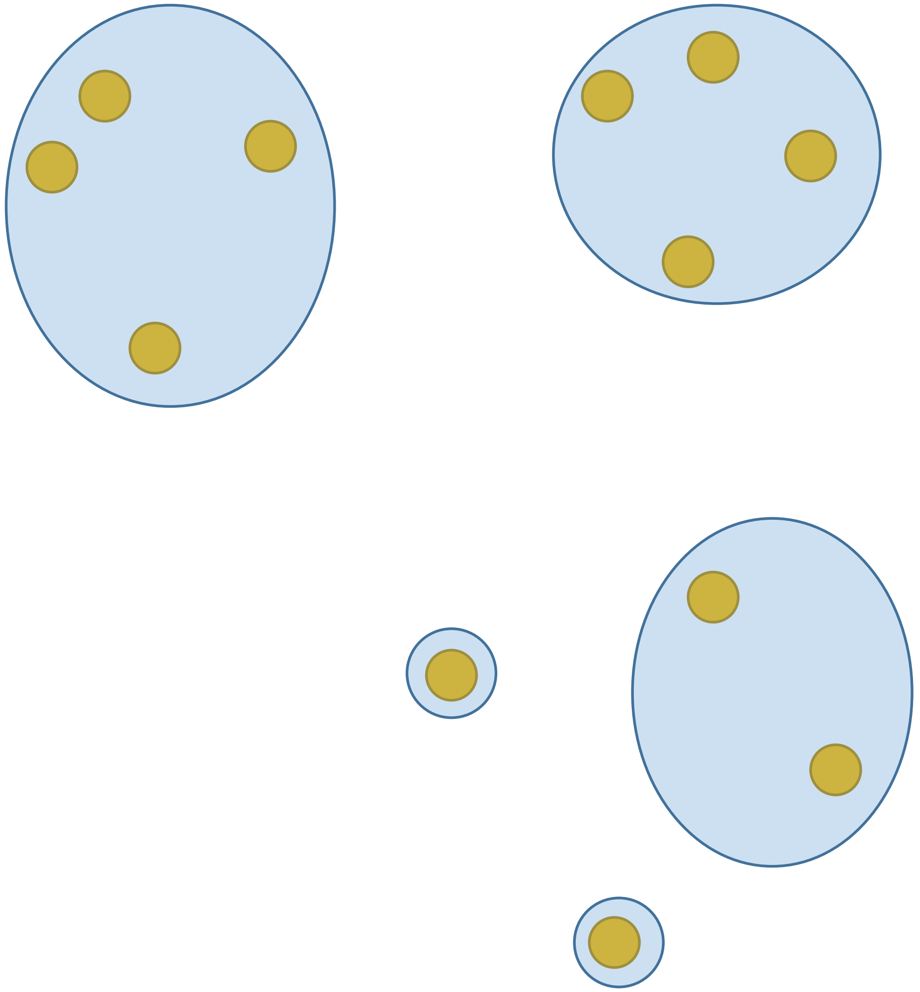
Агломеративная кластеризация



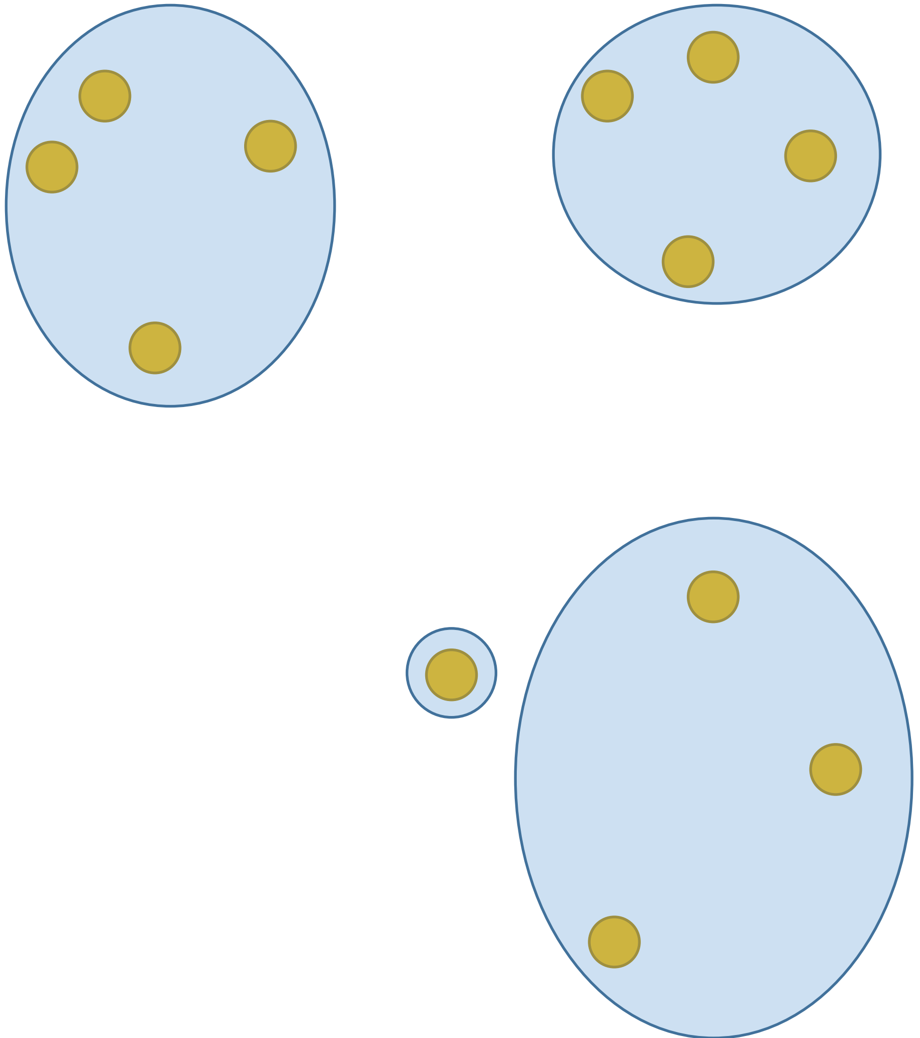
Агломеративная кластеризация



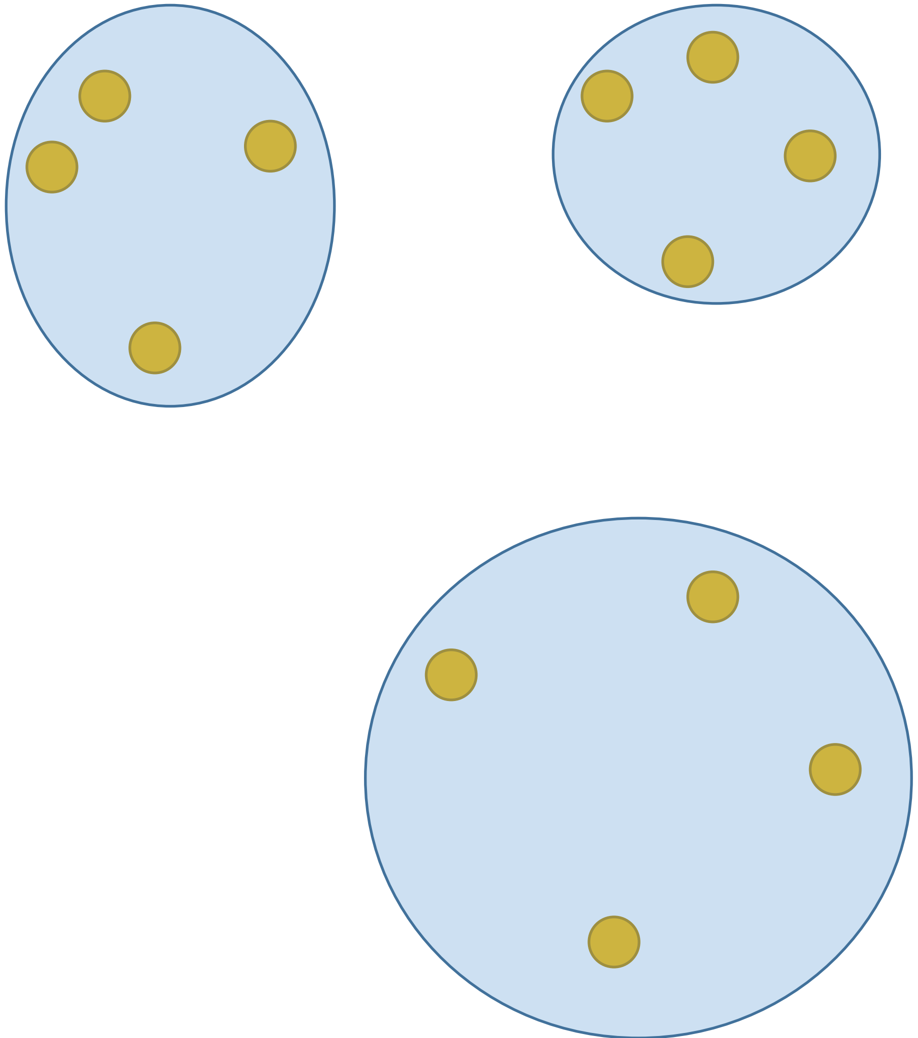
Агломеративная кластеризация



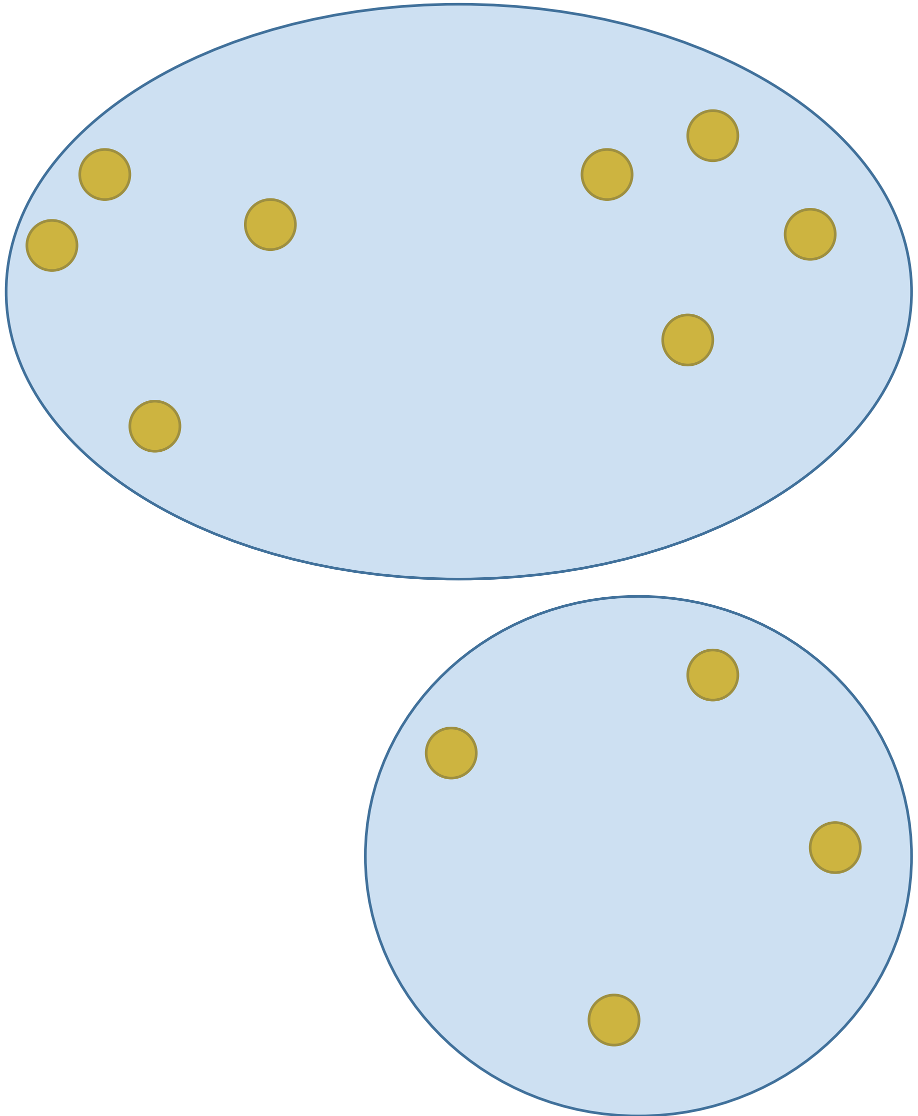
Агломеративная кластеризация



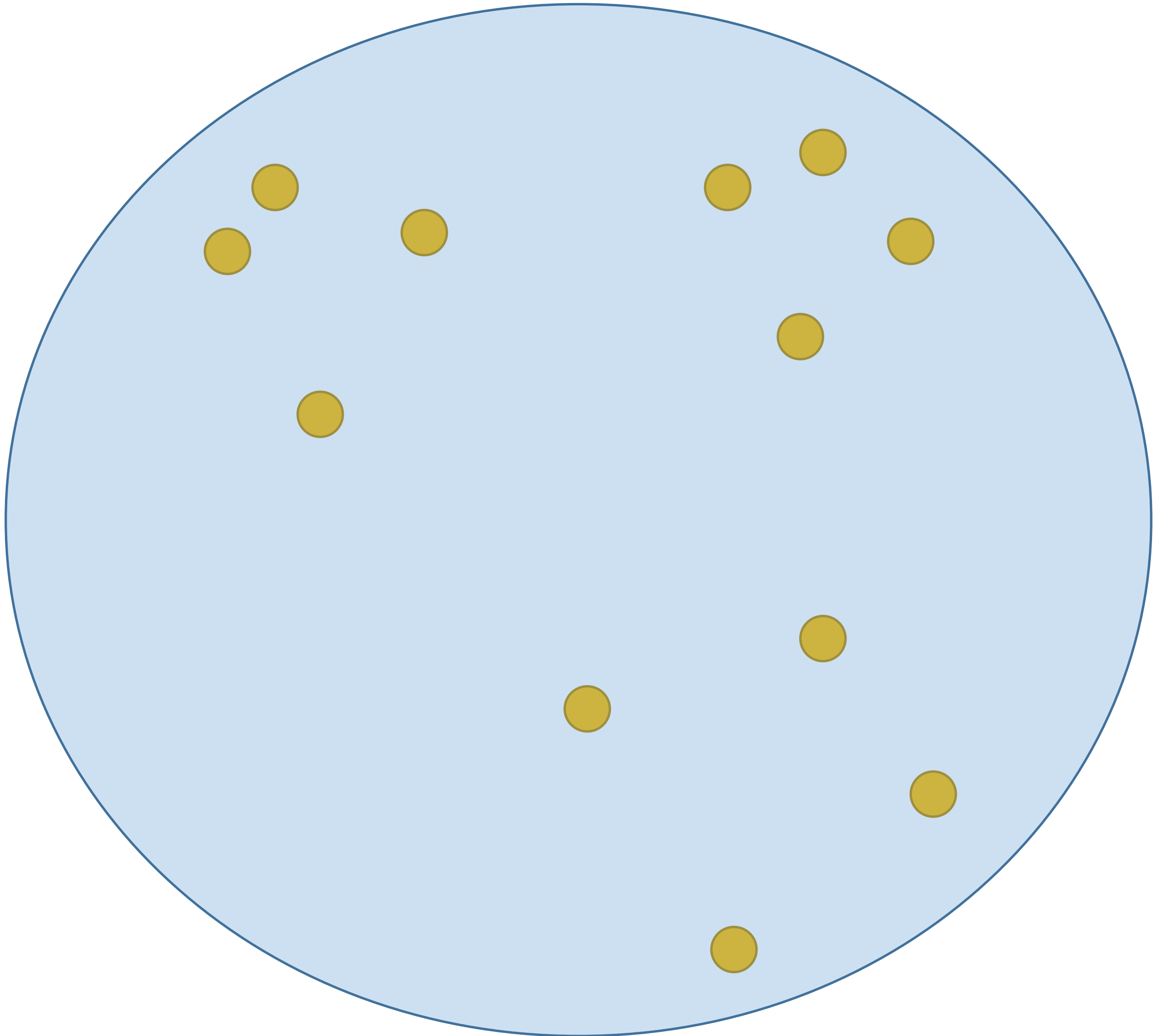
Агломеративная кластеризация



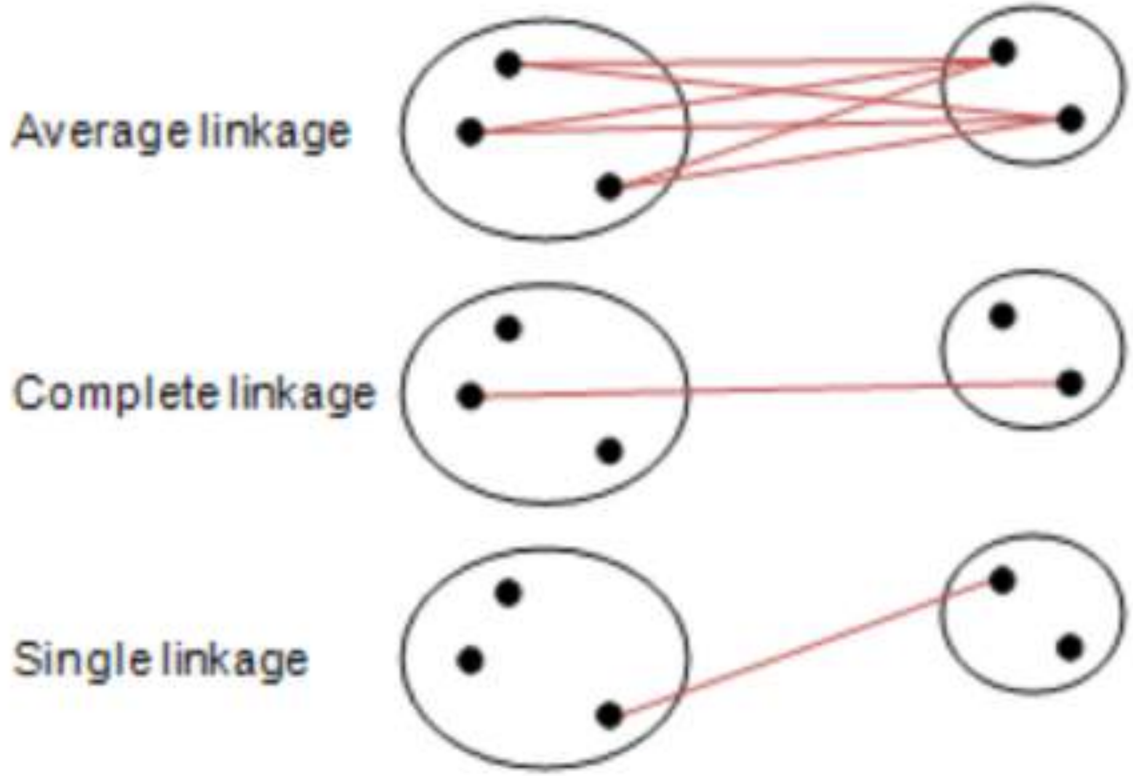
Агломеративная кластеризация



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:

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

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:

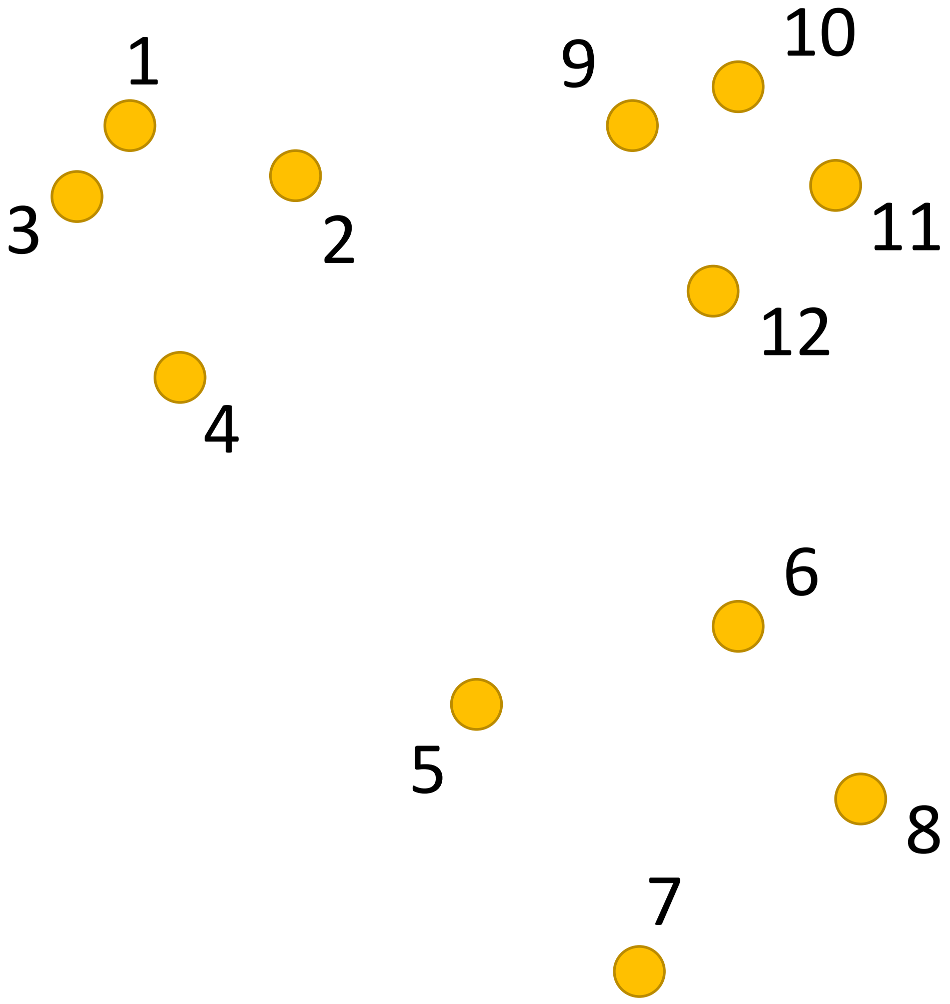
$$R^{\text{ц}}(W, S) = \rho^2\Big(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|}\Big); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = -\alpha_U \alpha_V,\ \gamma = 0.$$

Расстояние Уорда:

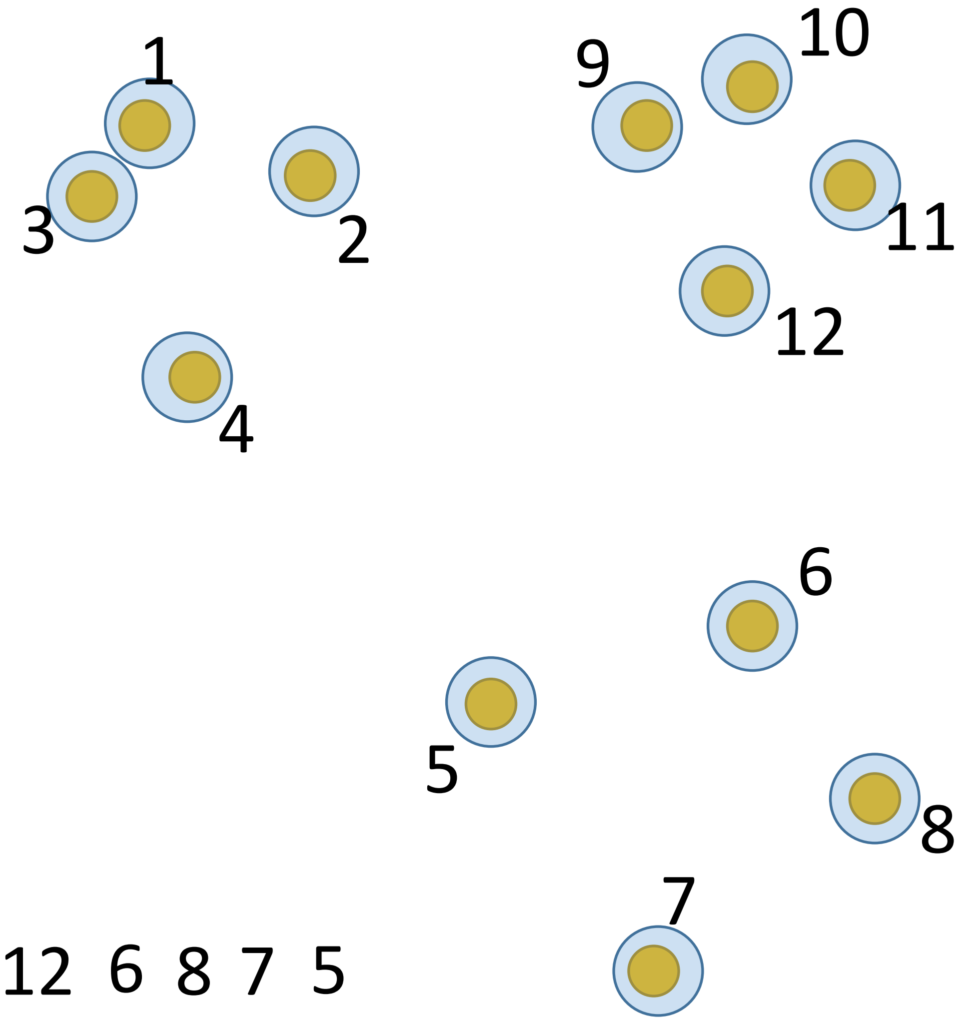
$$R^{\text{у}}(W, S) = \frac{|S||W|}{|S|+|W|} \rho^2\Big(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|}\Big); \quad \alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \beta = \frac{-|S|}{|S|+|W|},\ \gamma = 0.$$

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

Дендрограмма

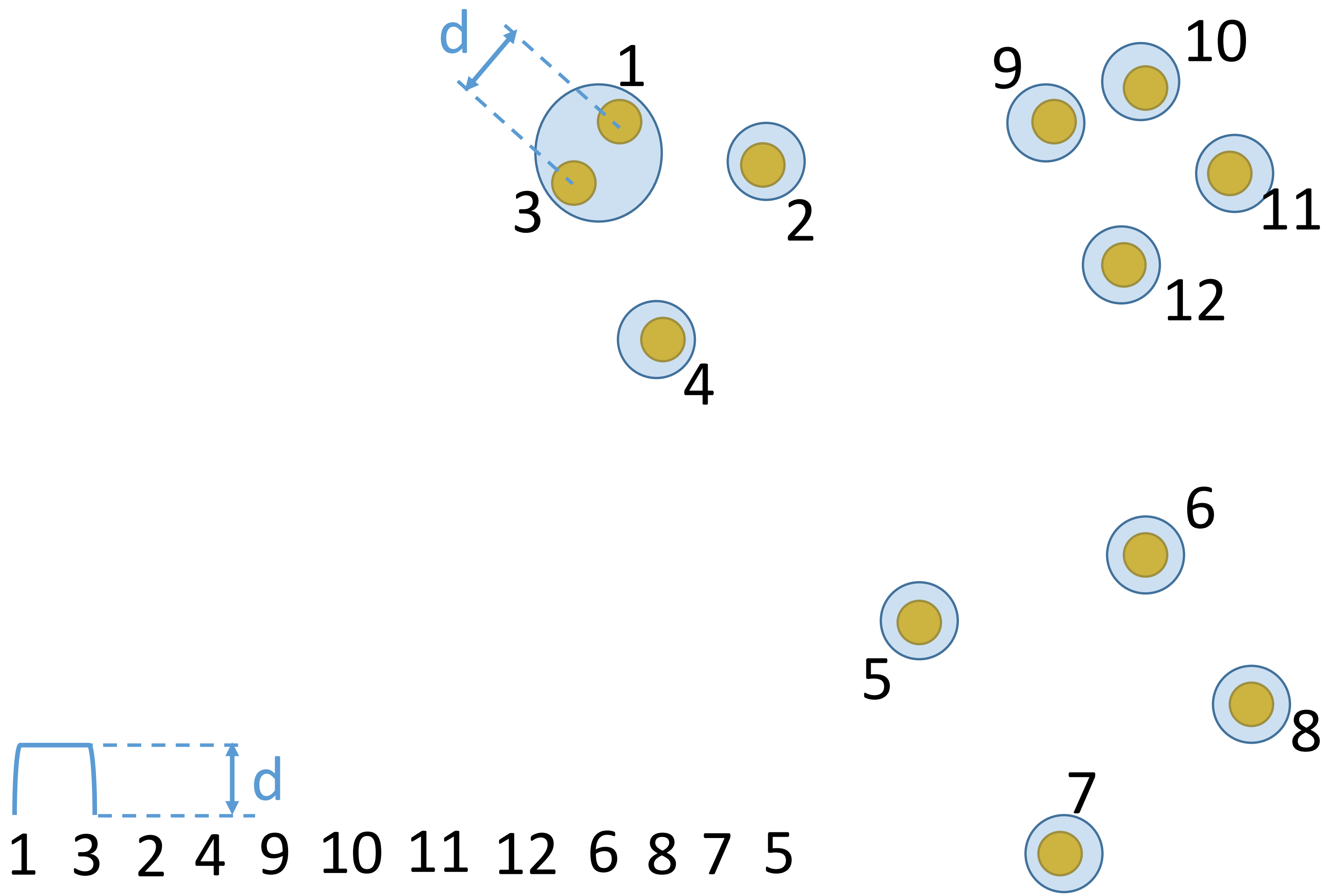


Дендрограмма

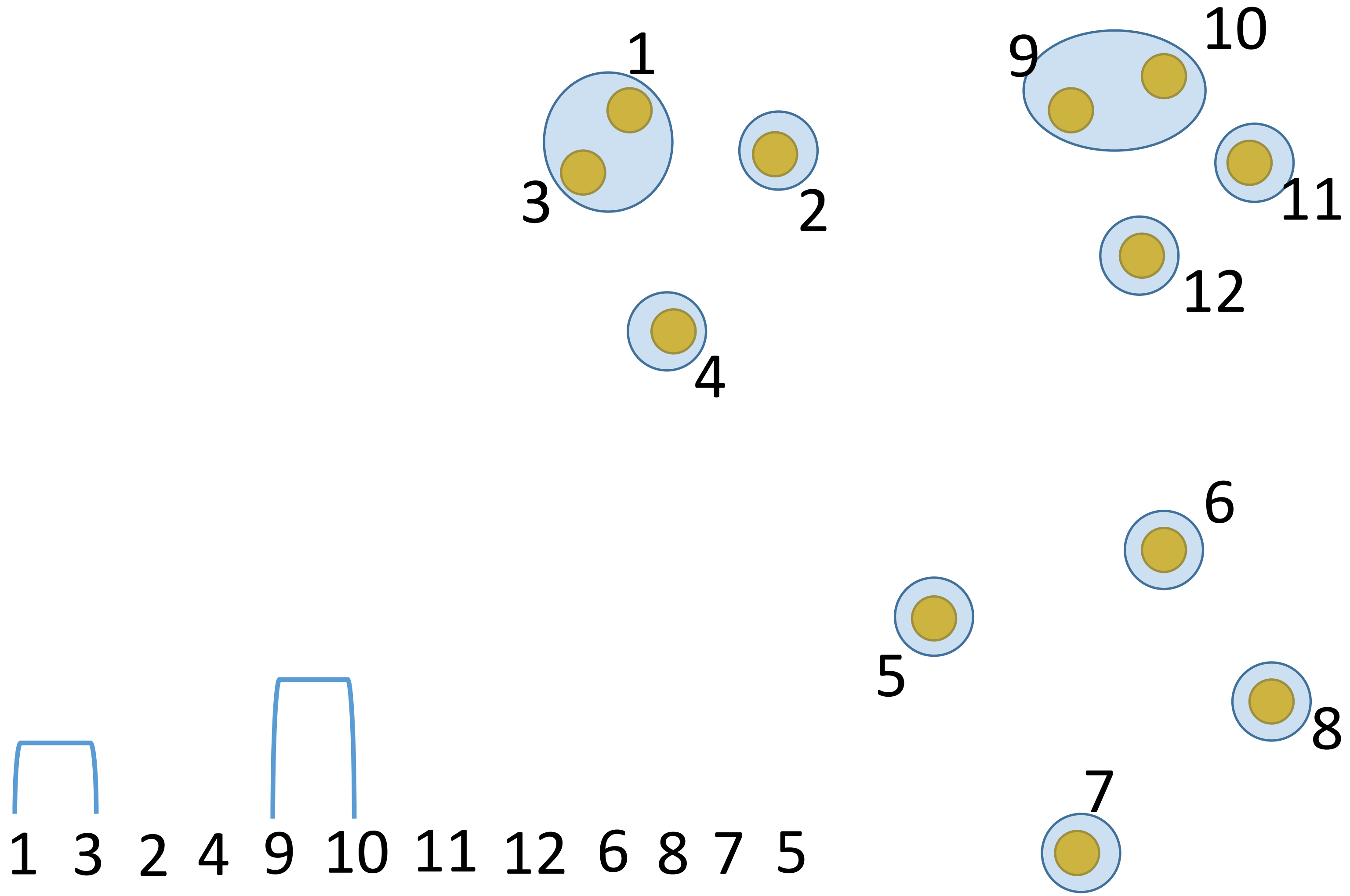


1 3 2 4 9 10 11 12 6 8 7 5

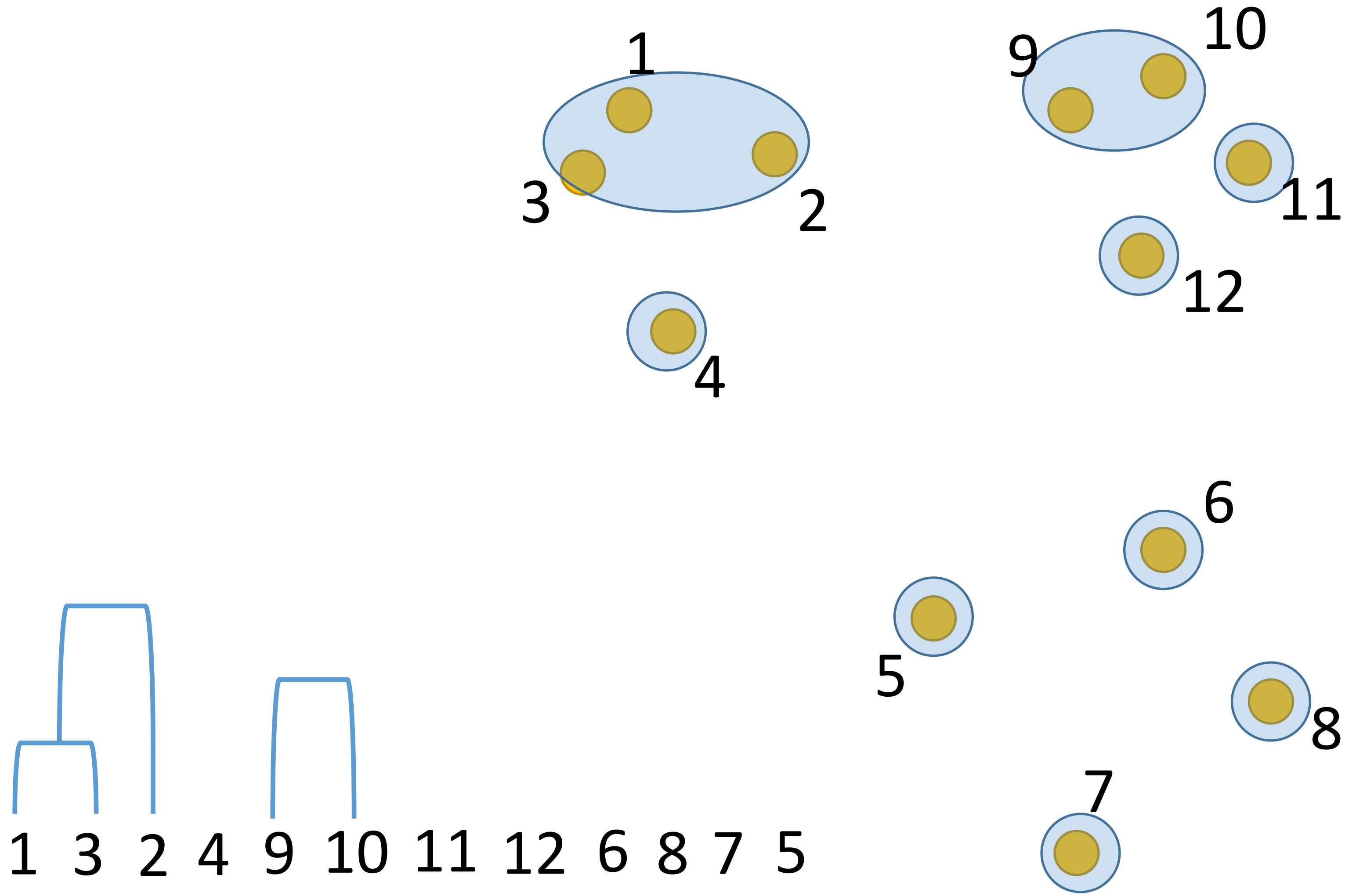
Дендрограмма



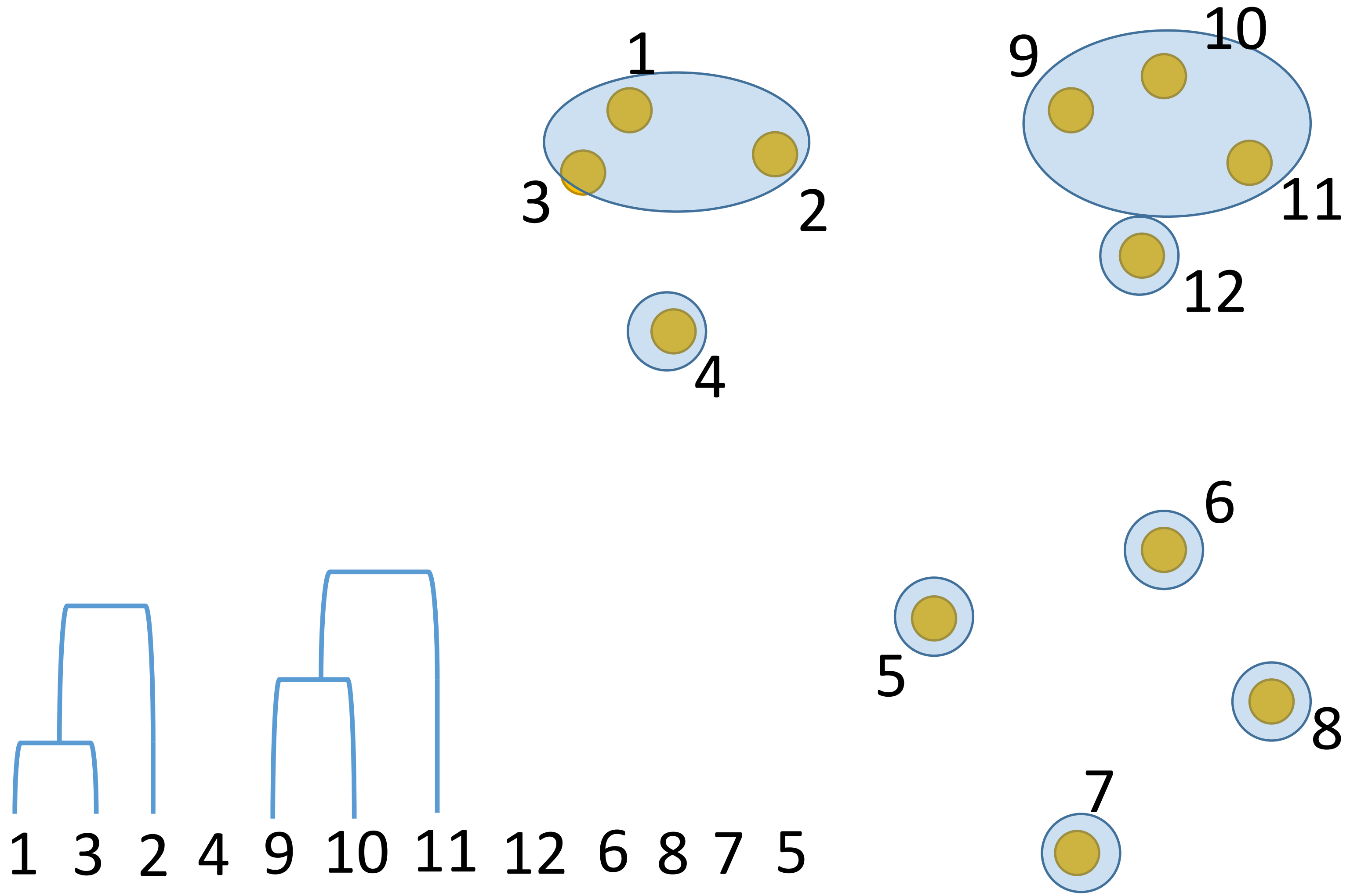
Дендрограмма



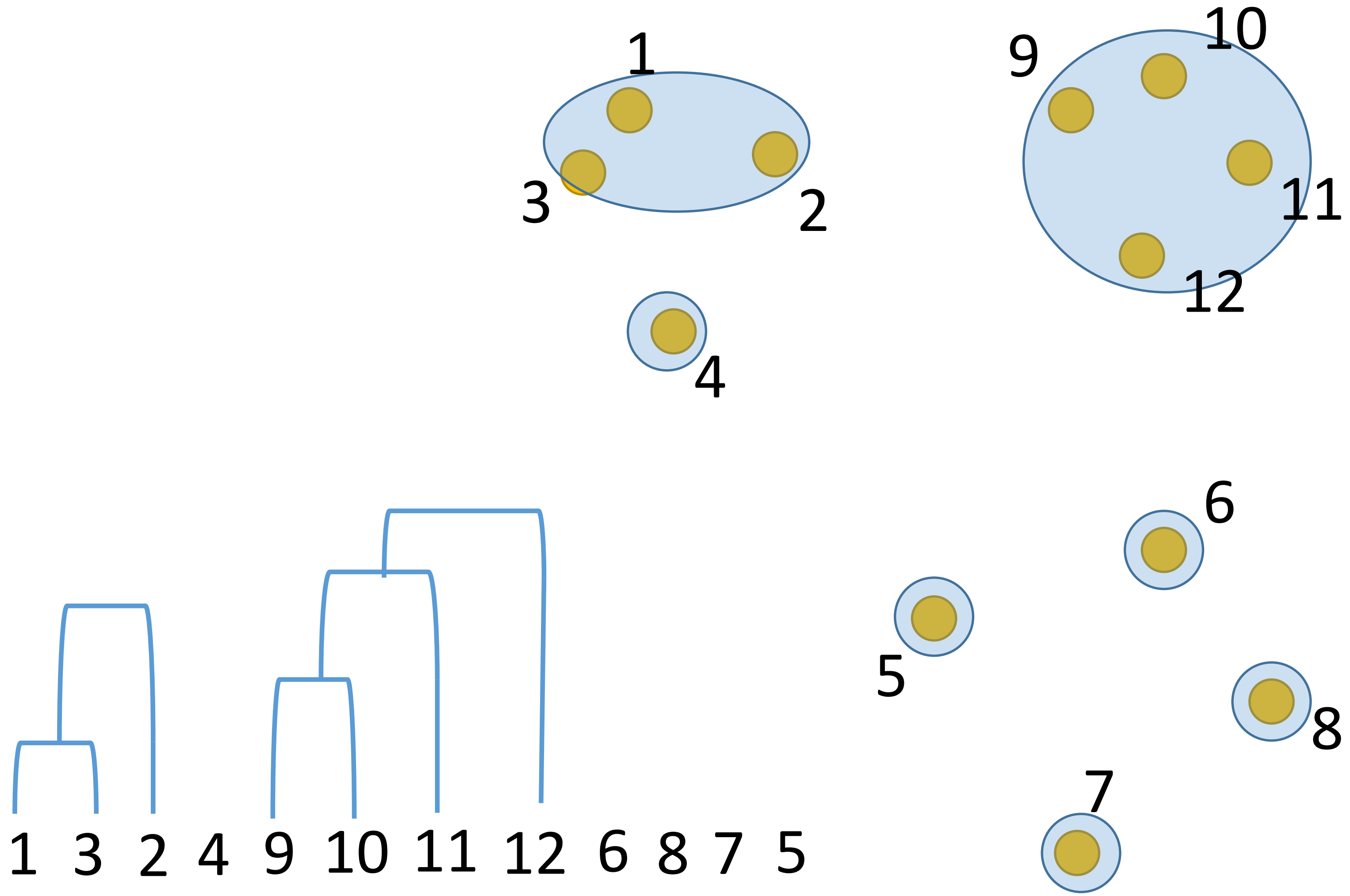
Дендрограмма



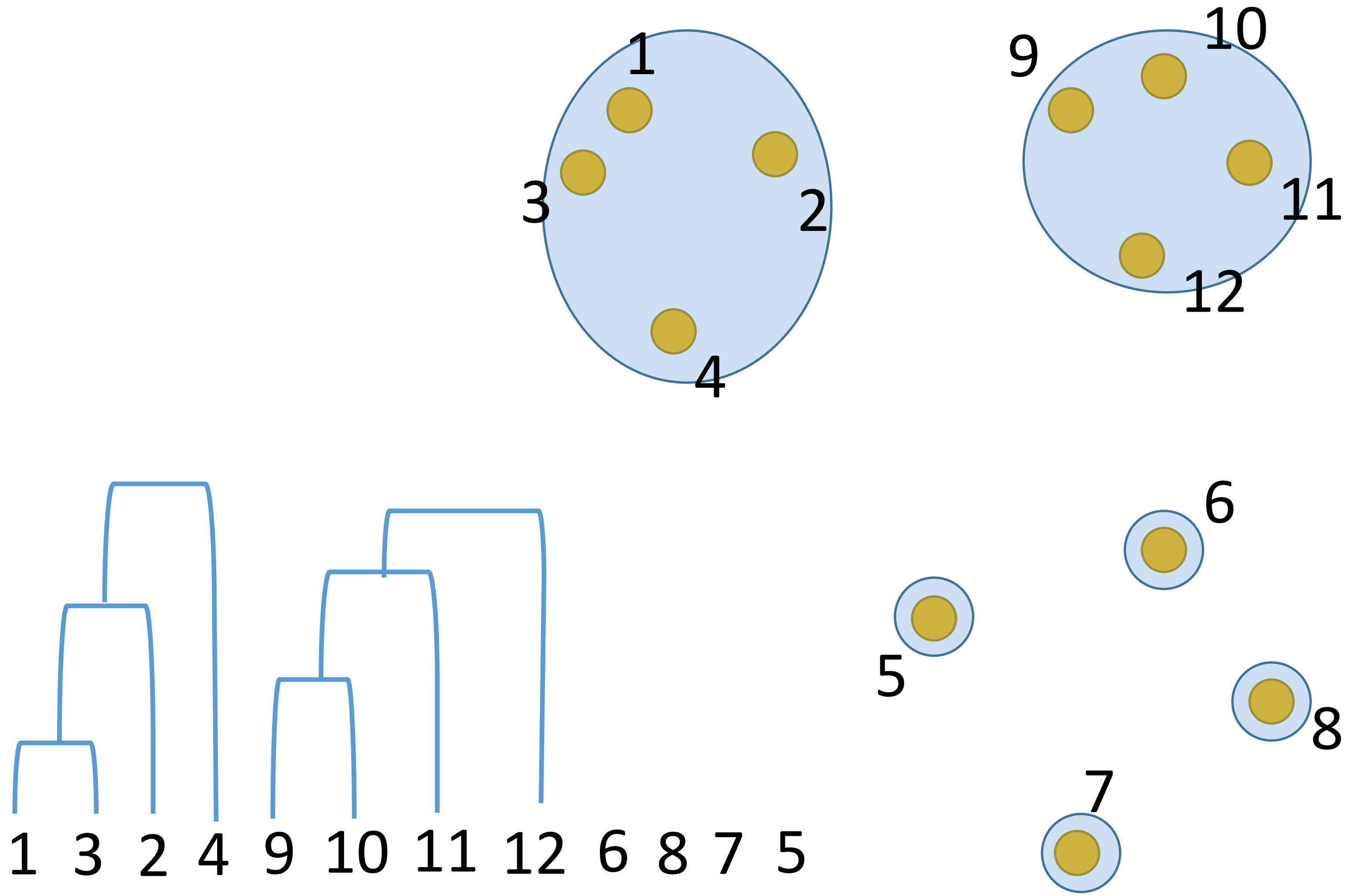
Дендрограмма



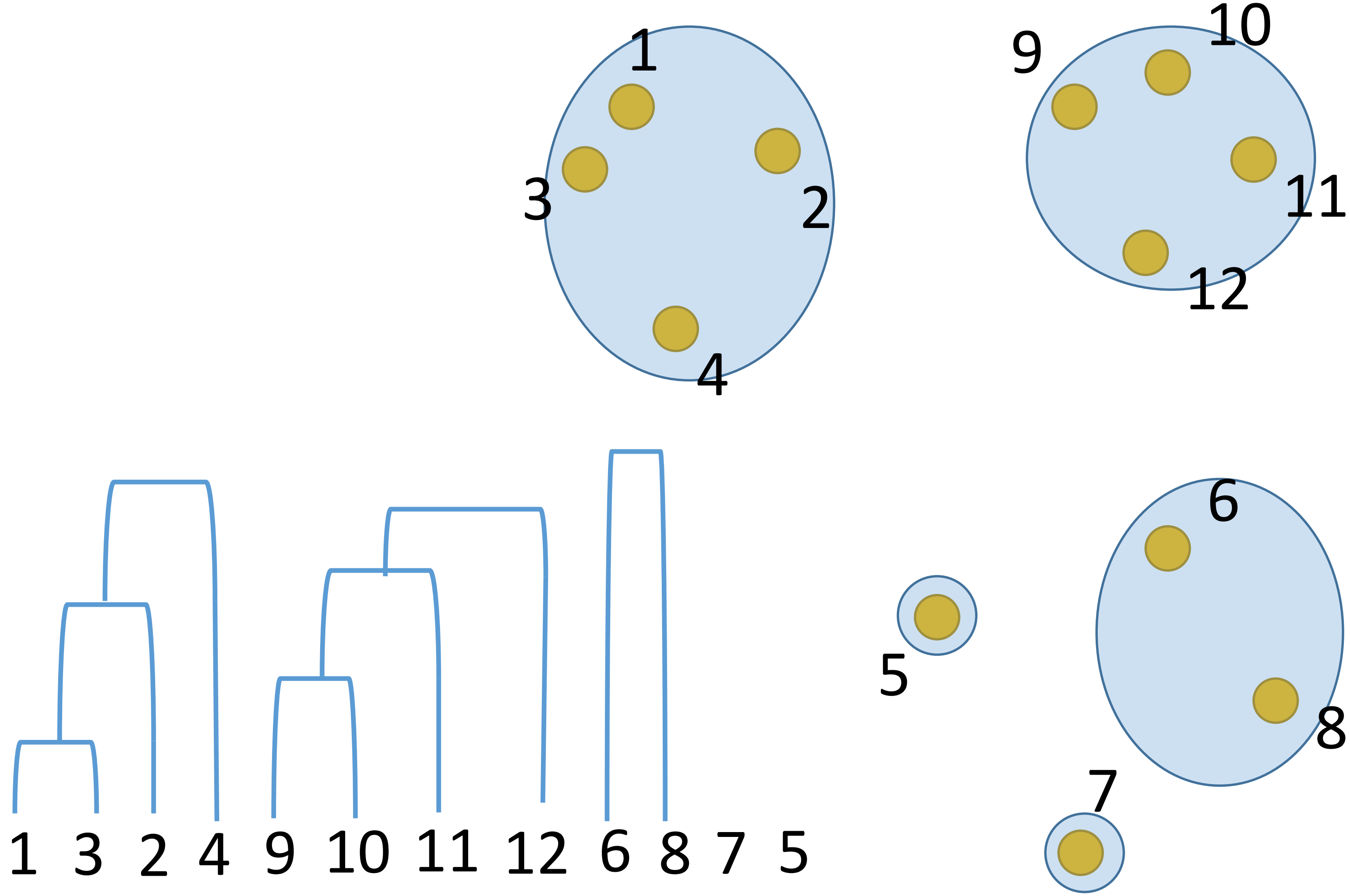
Дендрограмма



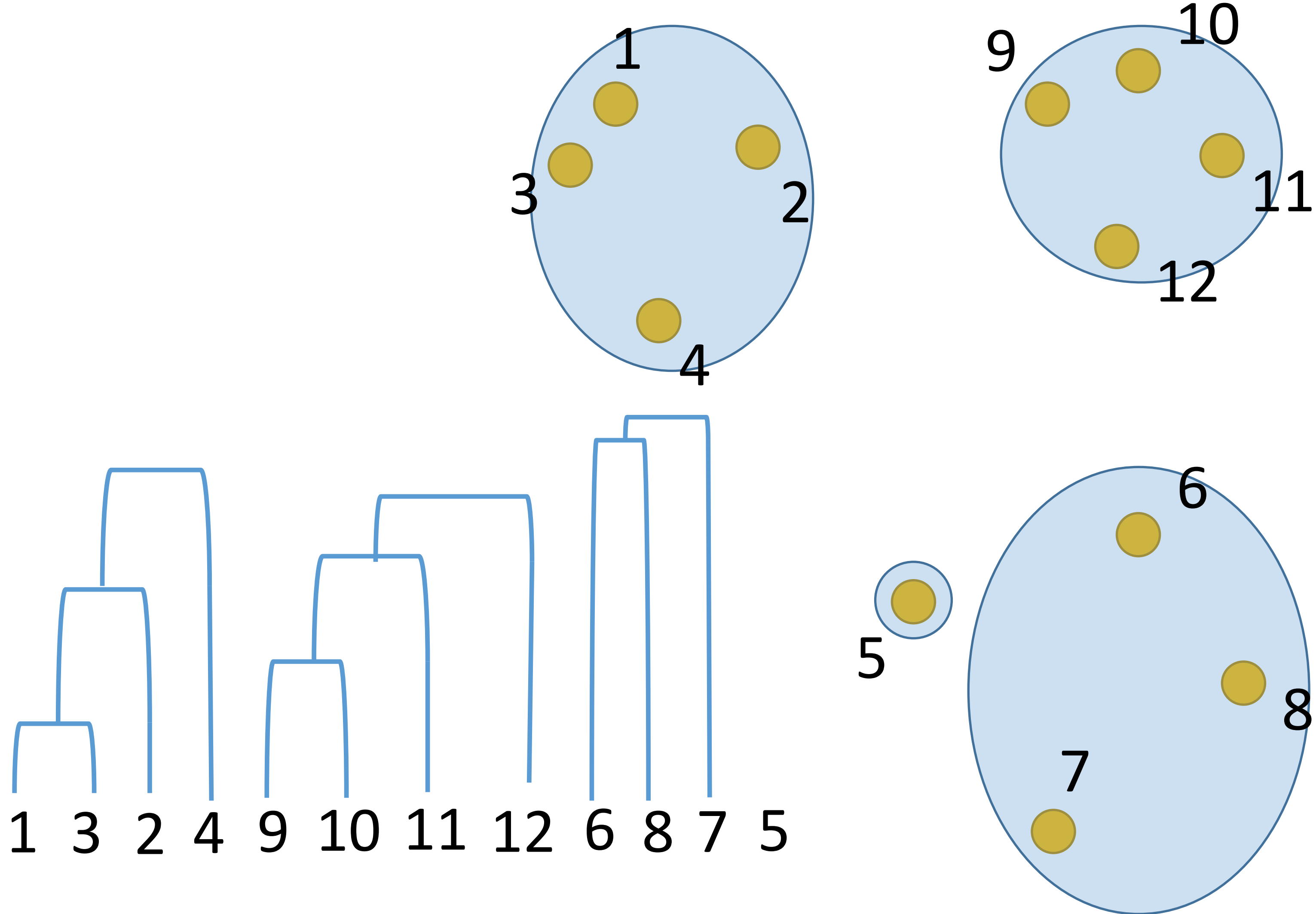
Дендрограмма



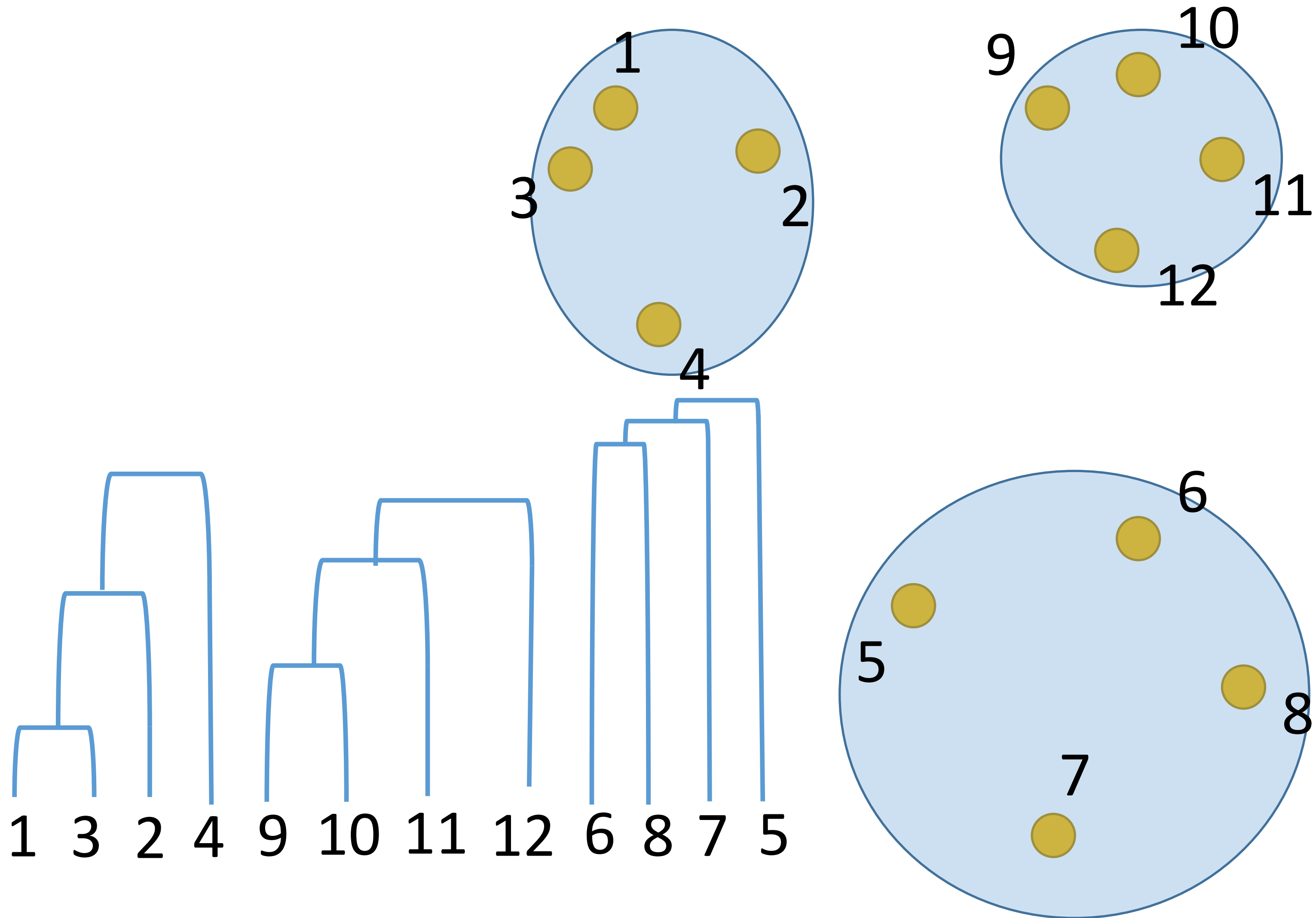
Дендрограмма



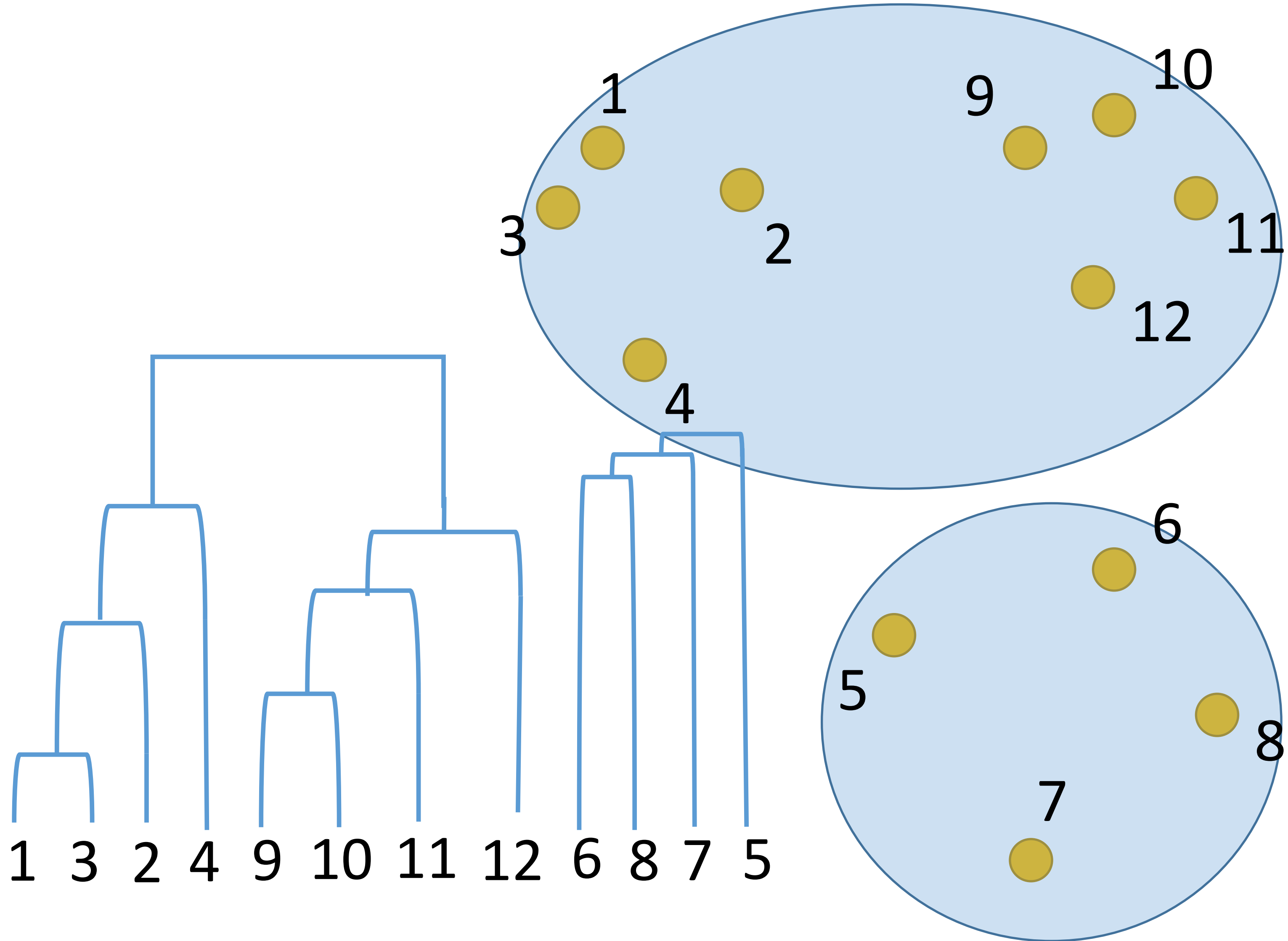
Дендрограмма



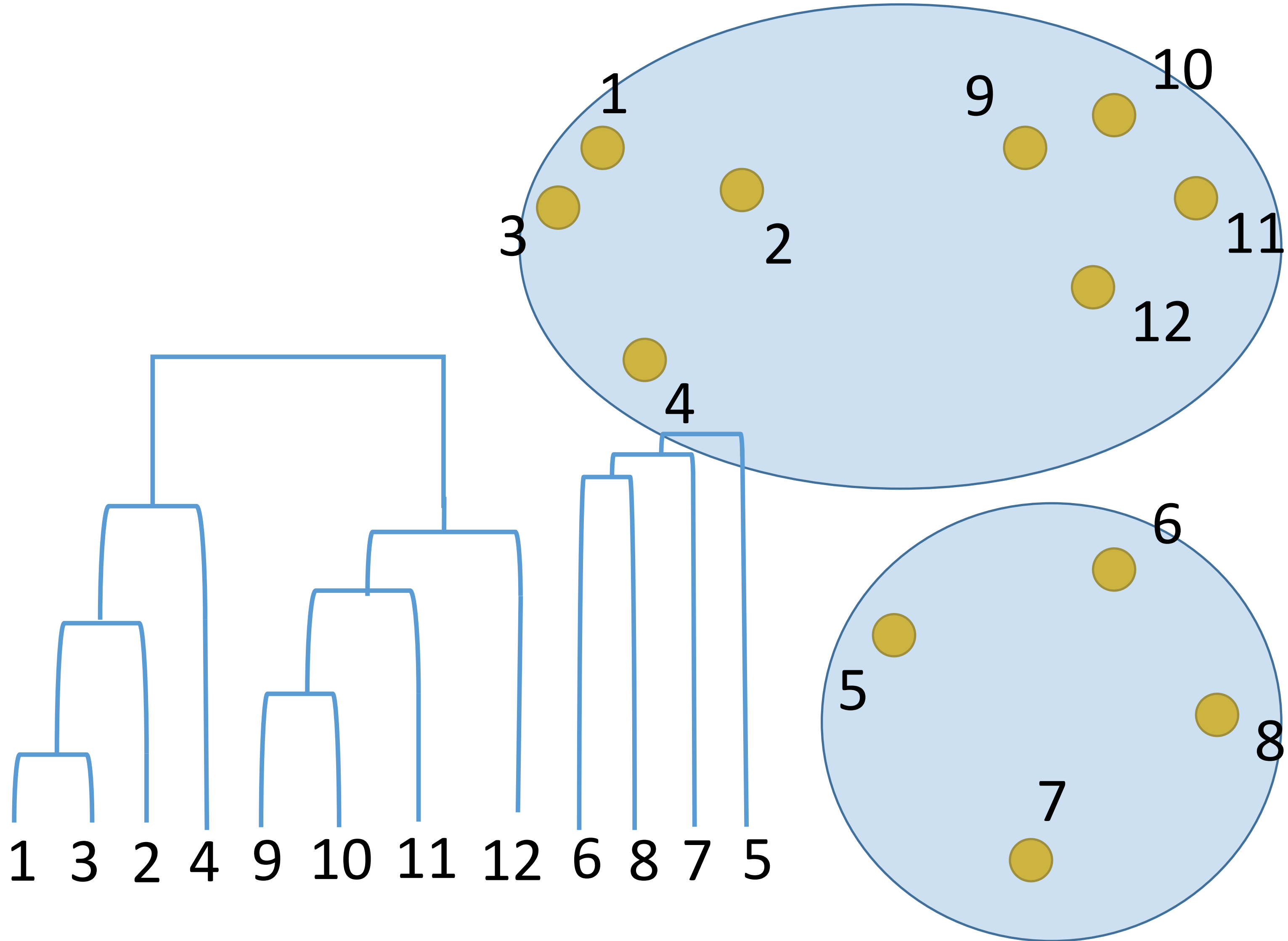
Дендрограмма



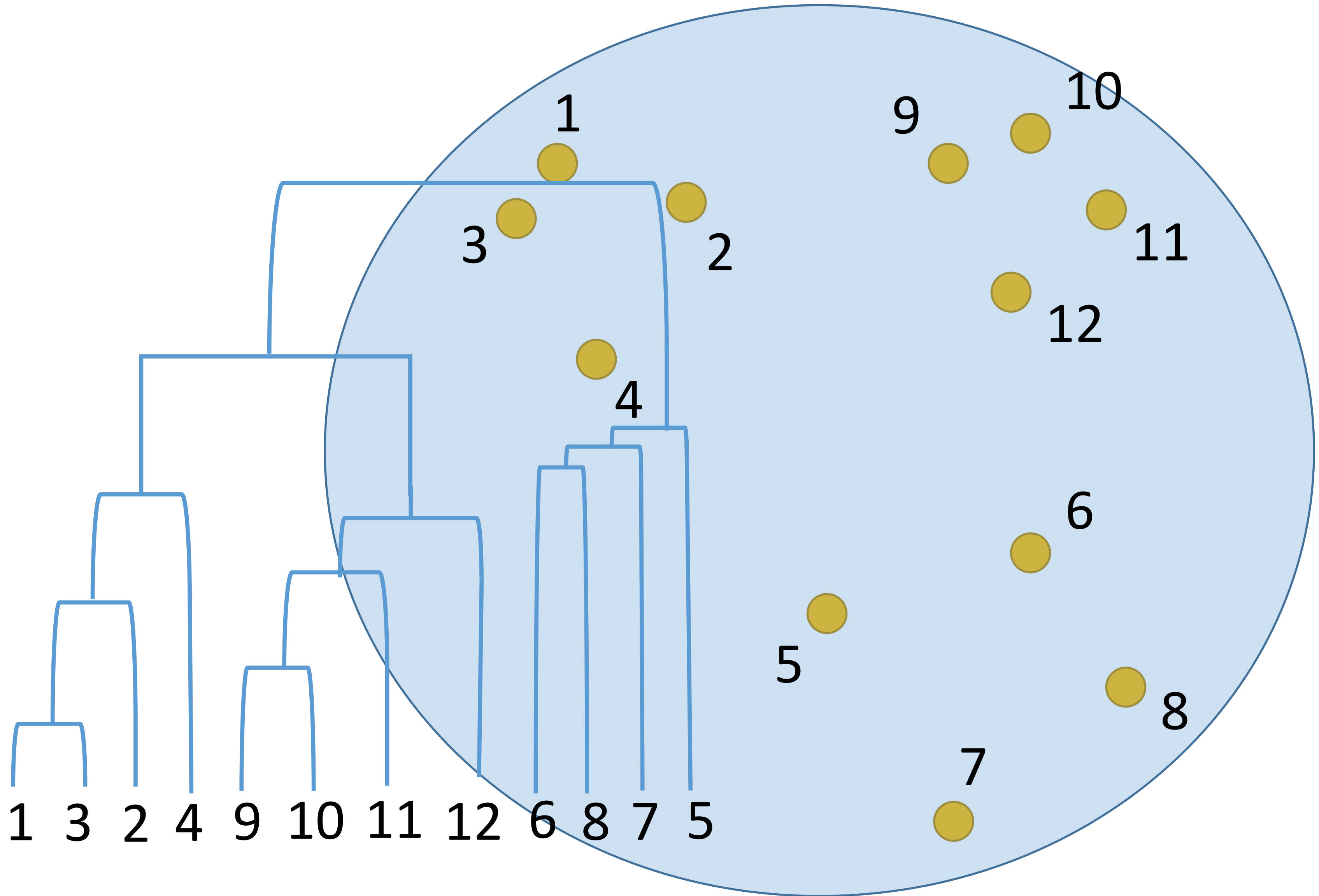
Дендрограмма



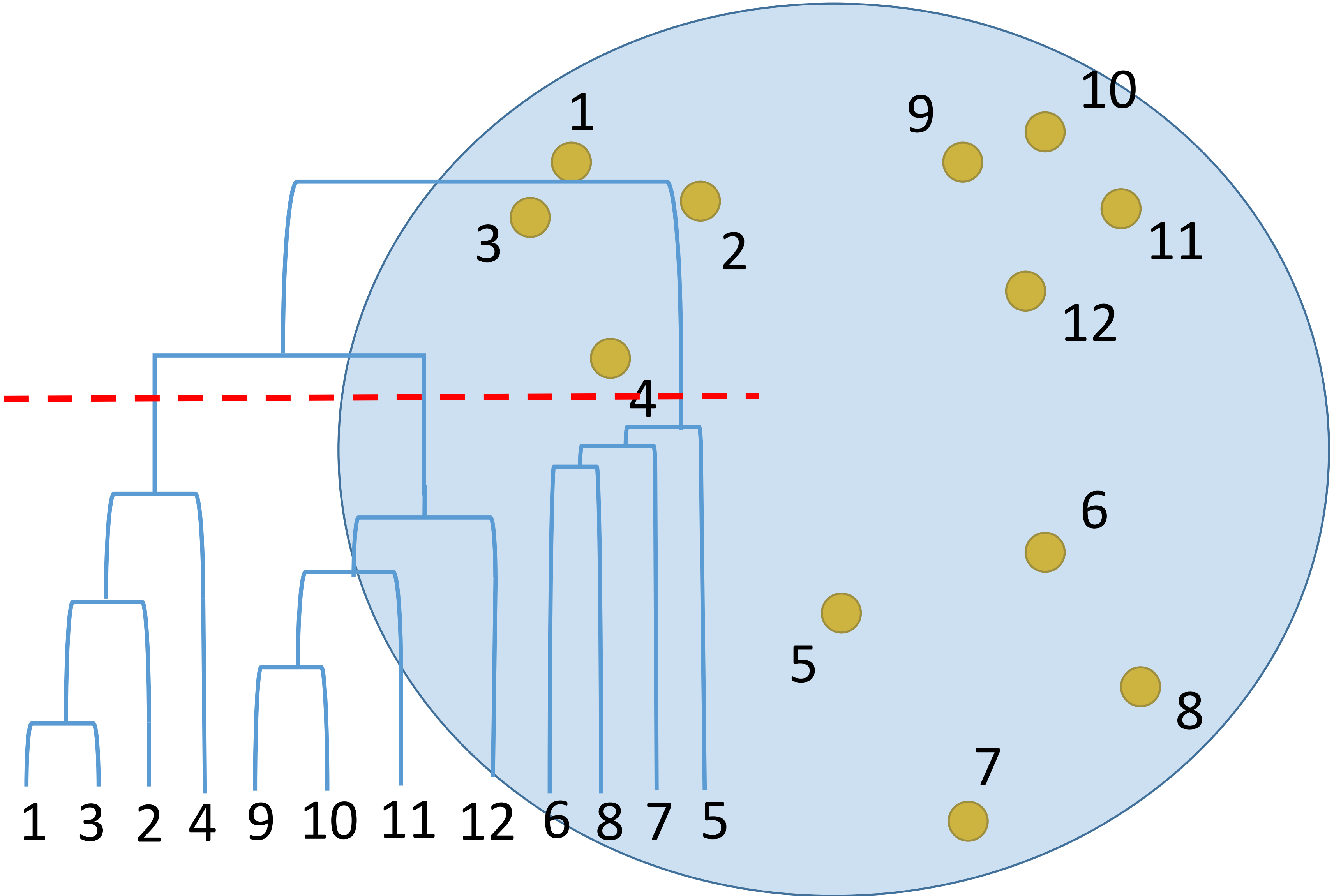
Дендрограмма



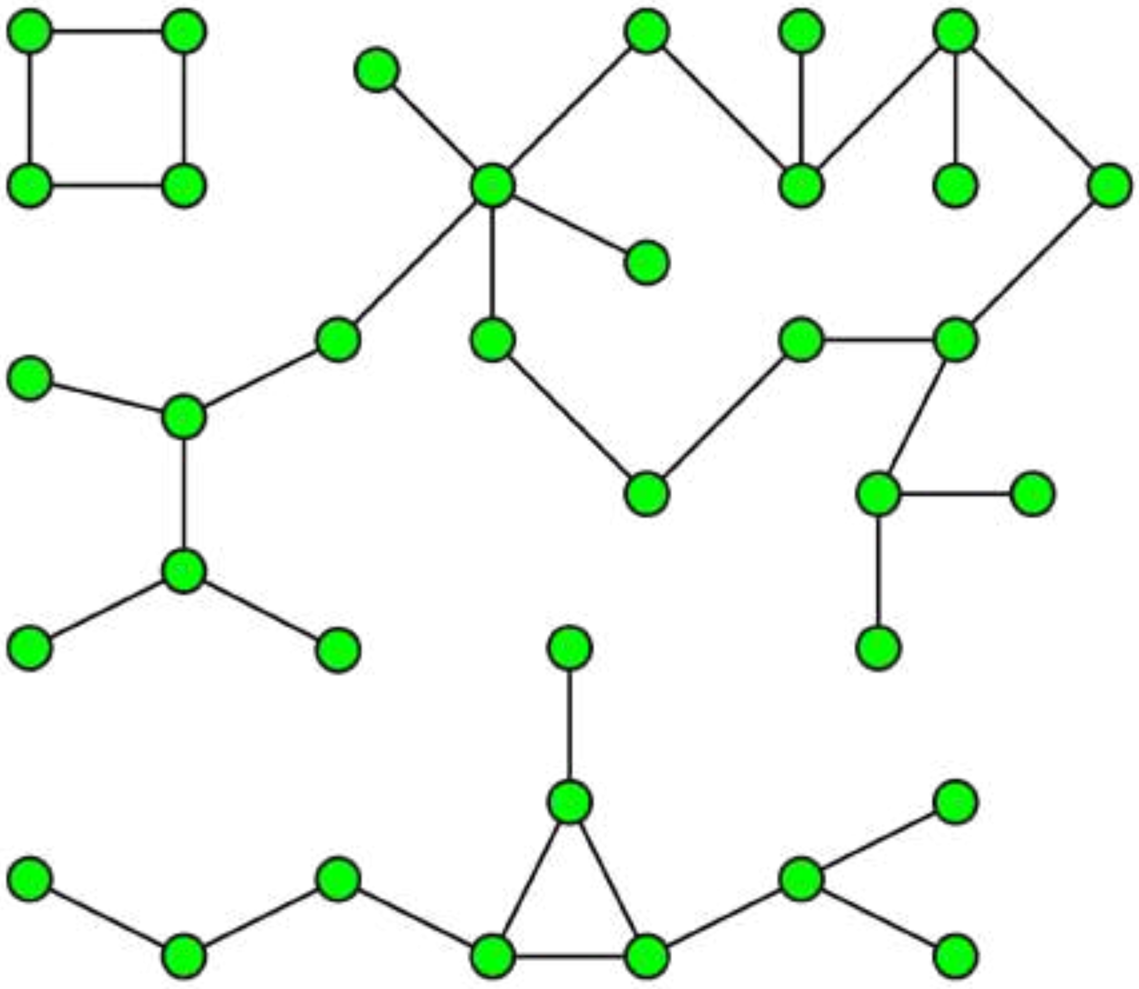
Дендрограмма



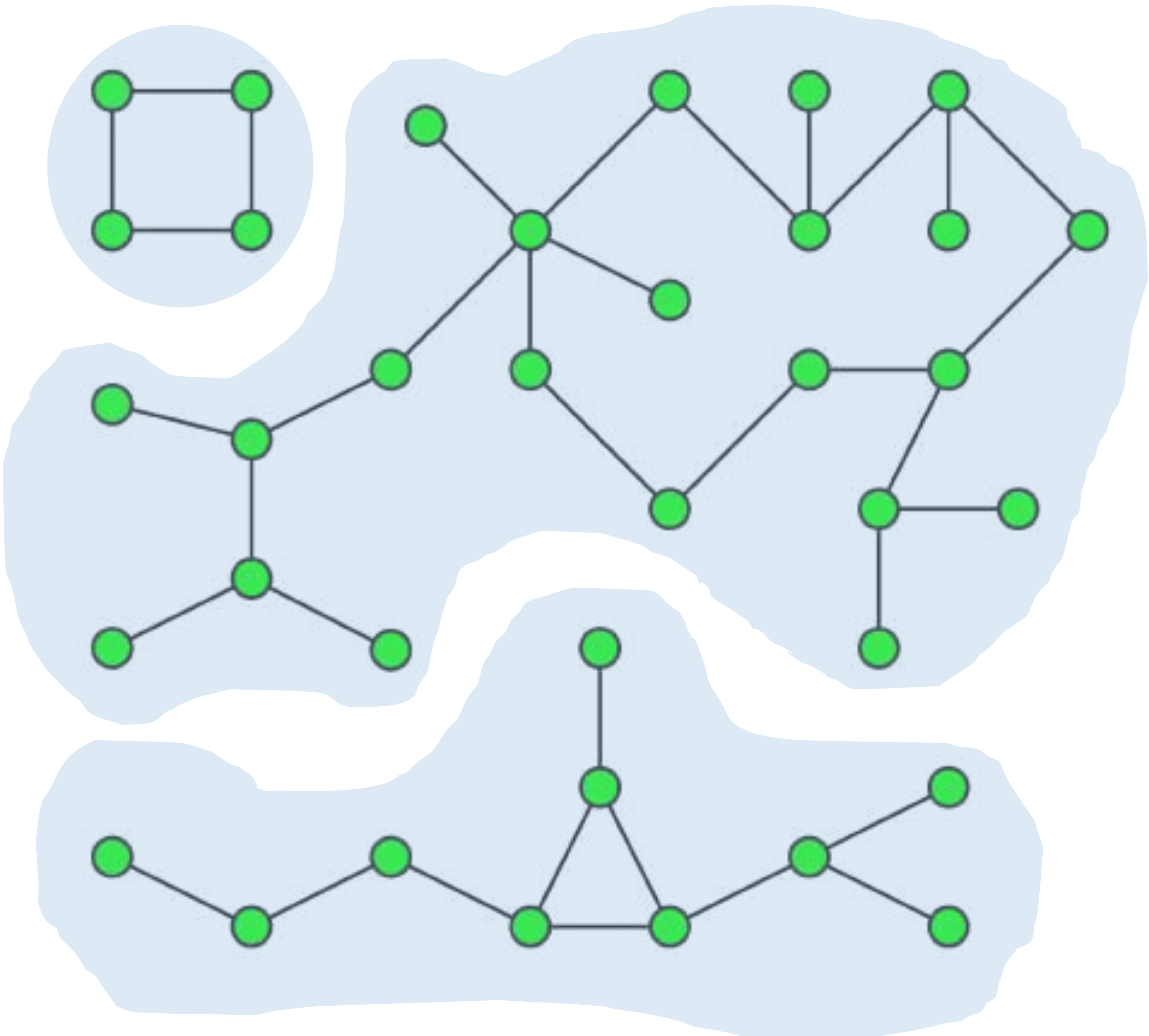
Дендрограмма



Выделение связных компонент



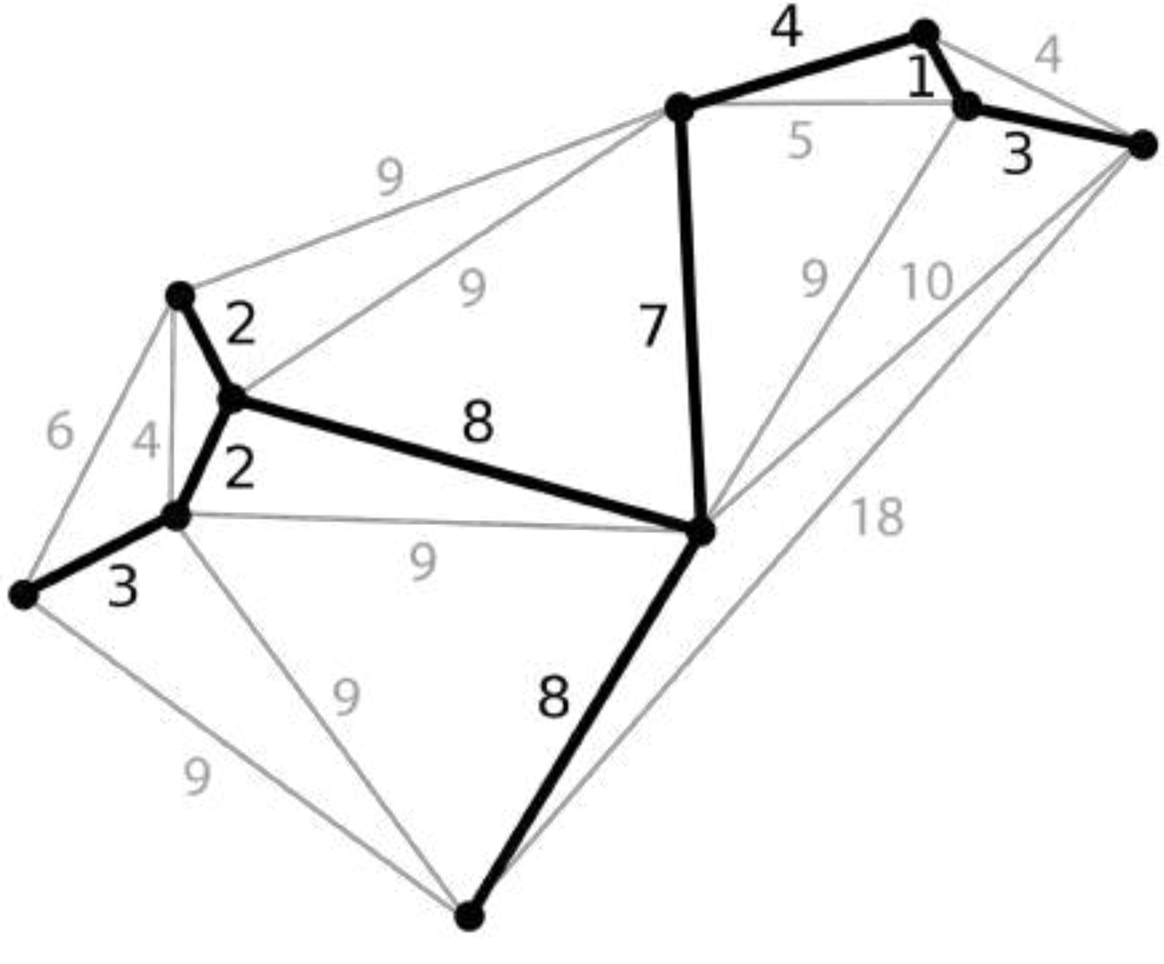
Выделение связных компонент



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 K-1 ребро с максимальным весом
- Получаем K компонент связности, которые интерпретируем как кластеры

Идея density-based методов

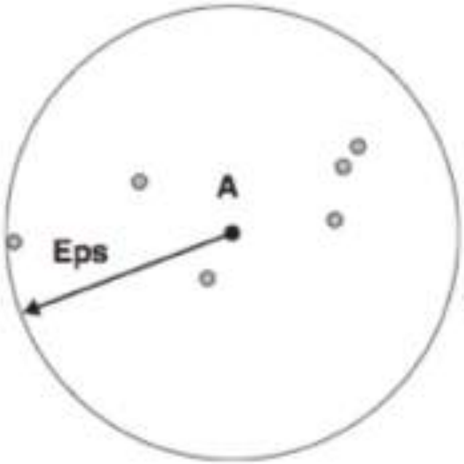


Figure 8.20. Center-based density.

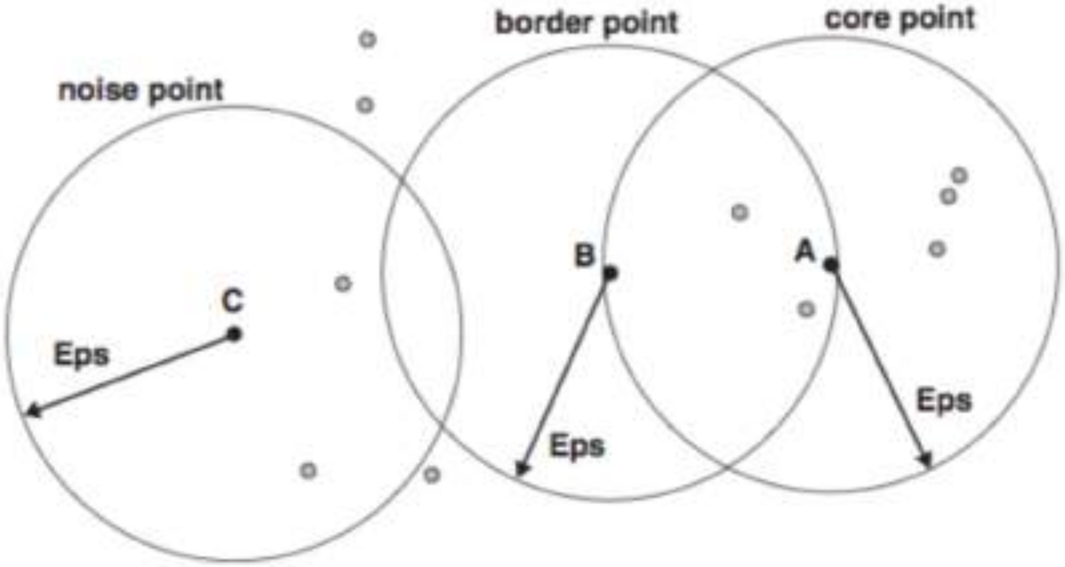
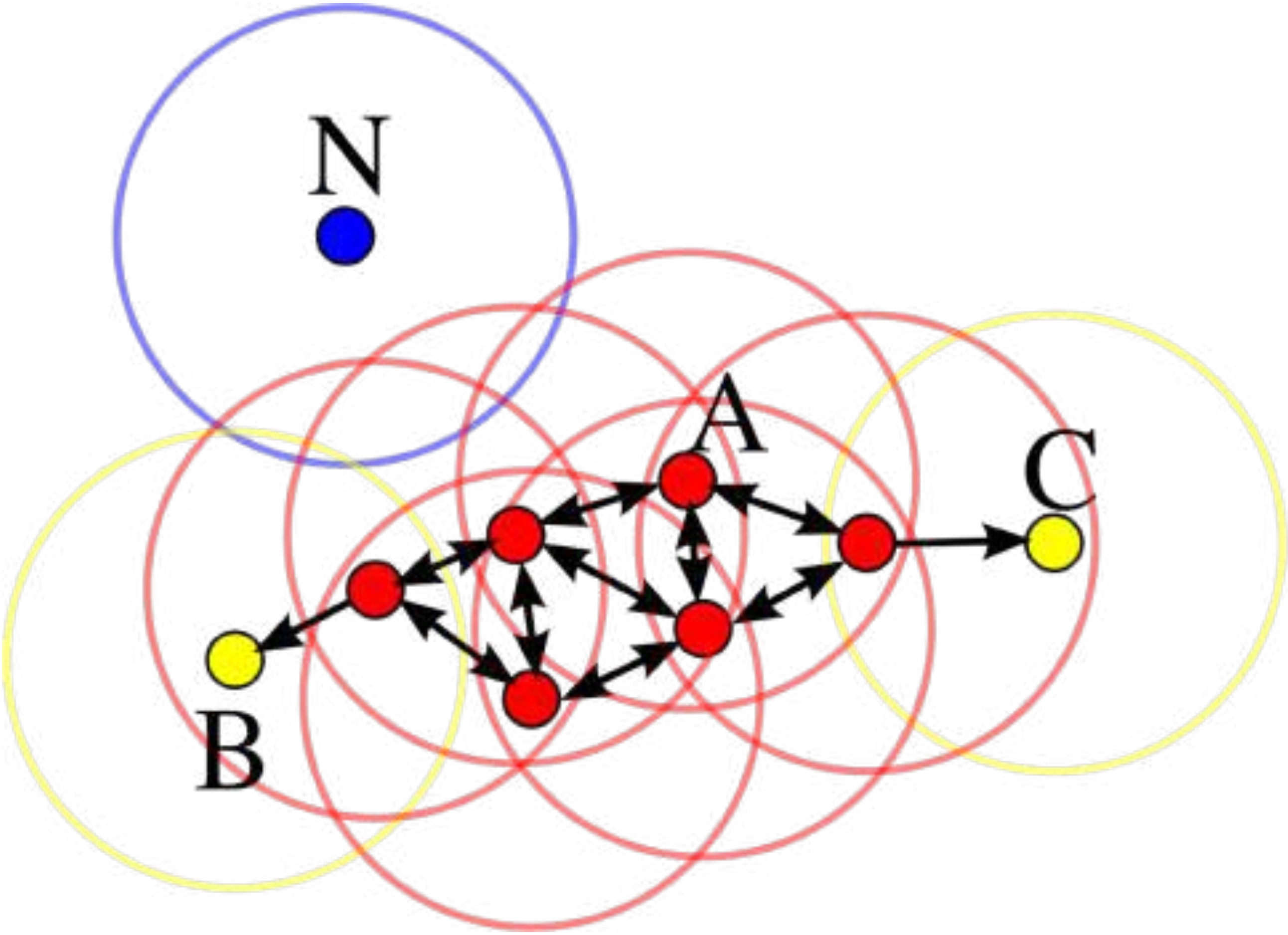


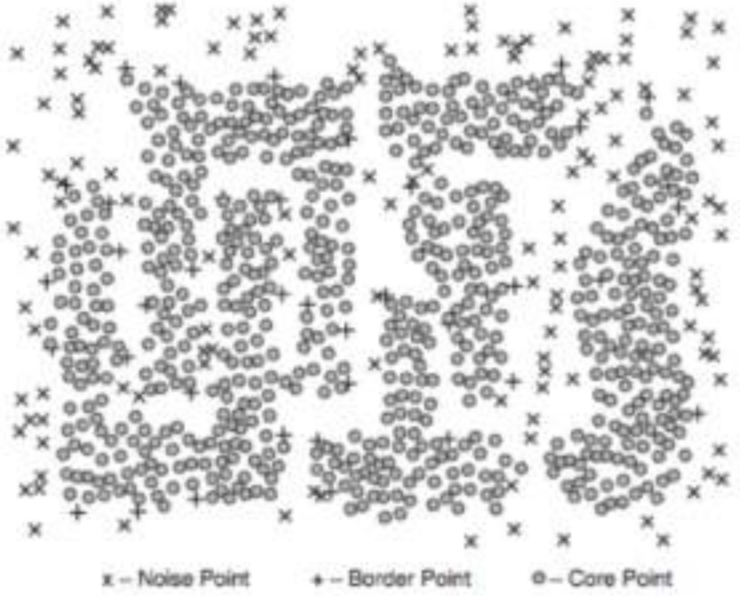
Figure 8.21. Core, border, and noise points.

Основные, шумовые и граничные точки



DBSCAN

(a) Clusters found by DBSCAN.

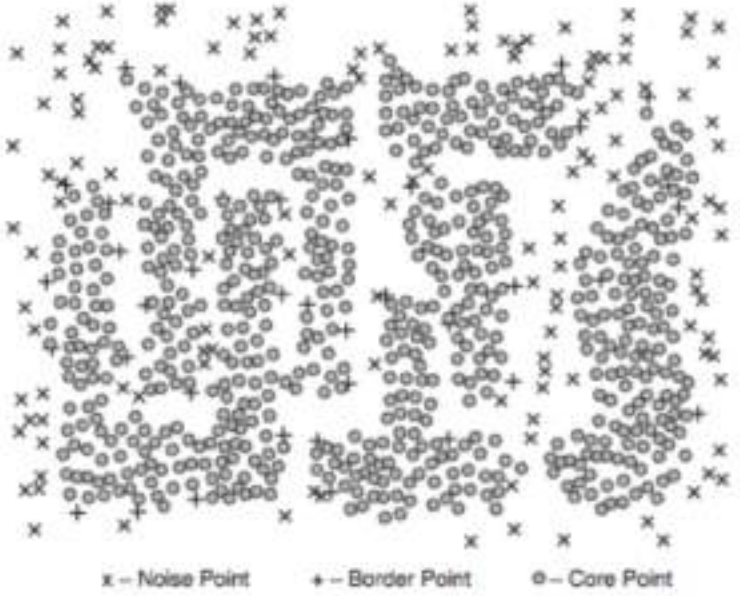


(b) Core, border, and noise points.

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

DBSCAN

(a) Clusters found by DBSCAN.



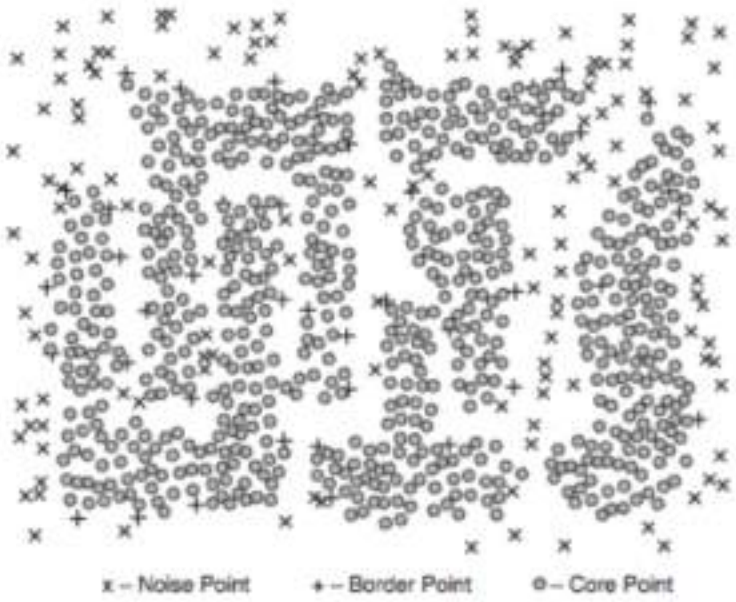
(b) Core, border, and noise points.

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.

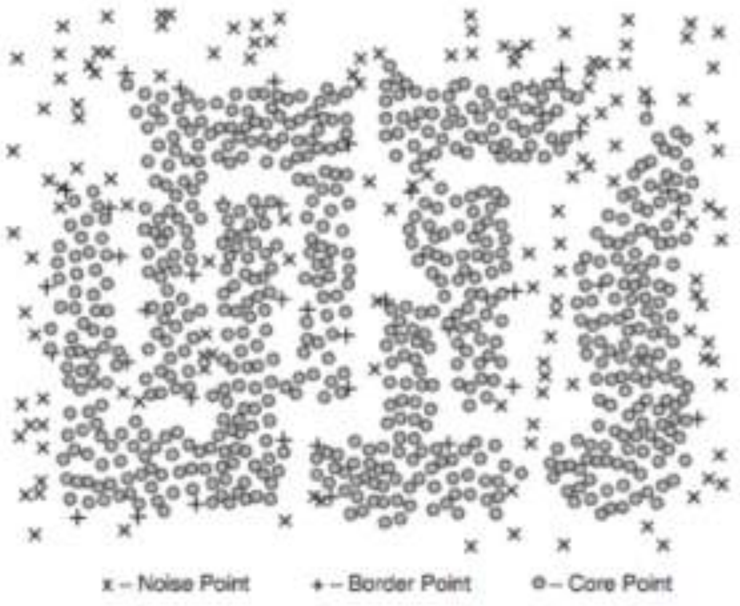
1: Пометить все точки, как основные, пограничные или шумовые.

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.

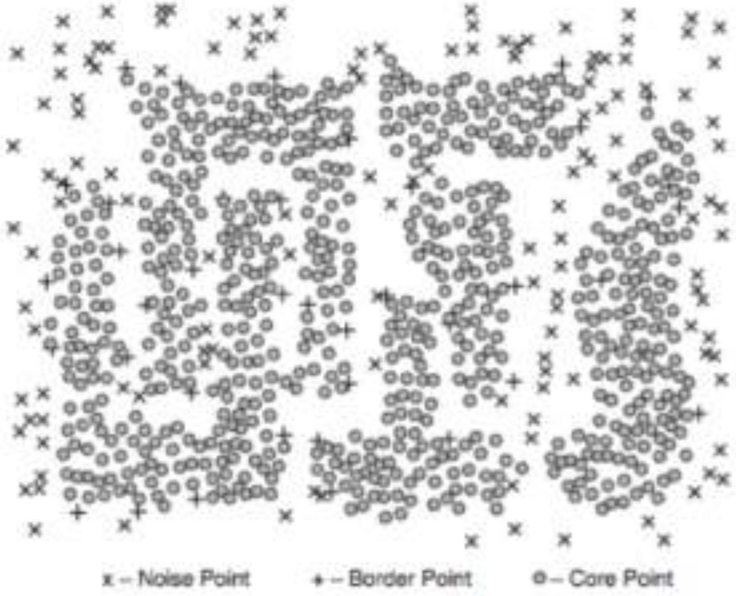
2: Отбросить точки шума.

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.

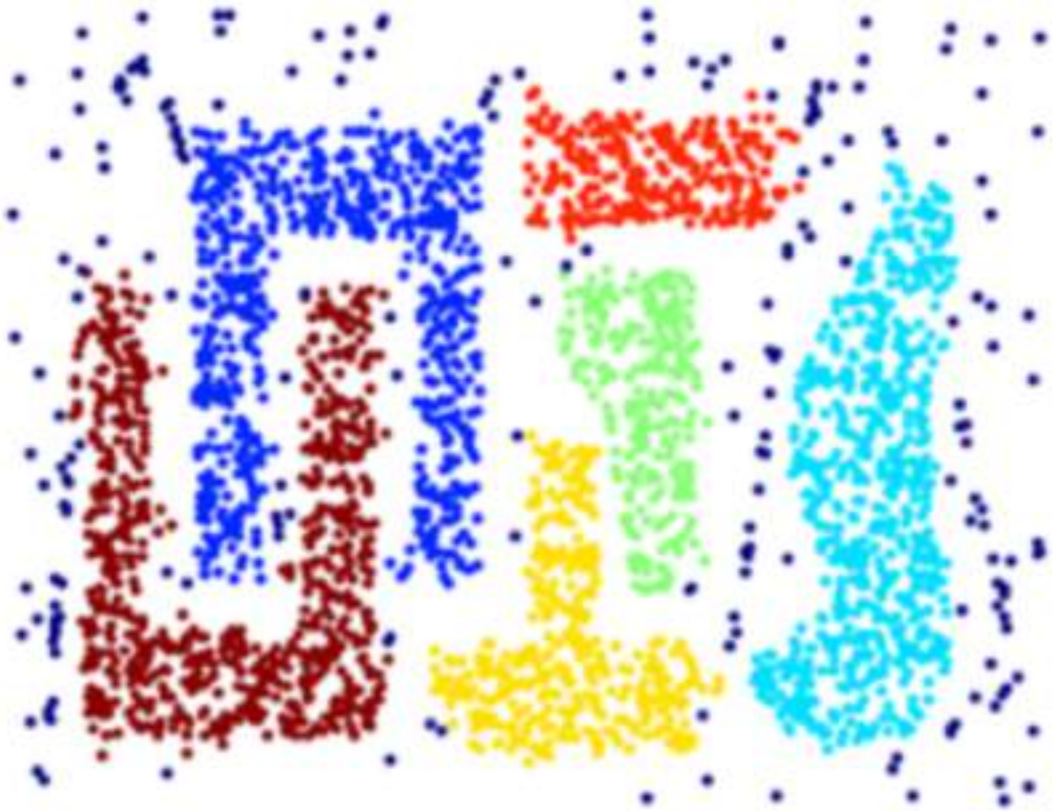
2: Отбросить точки шума.

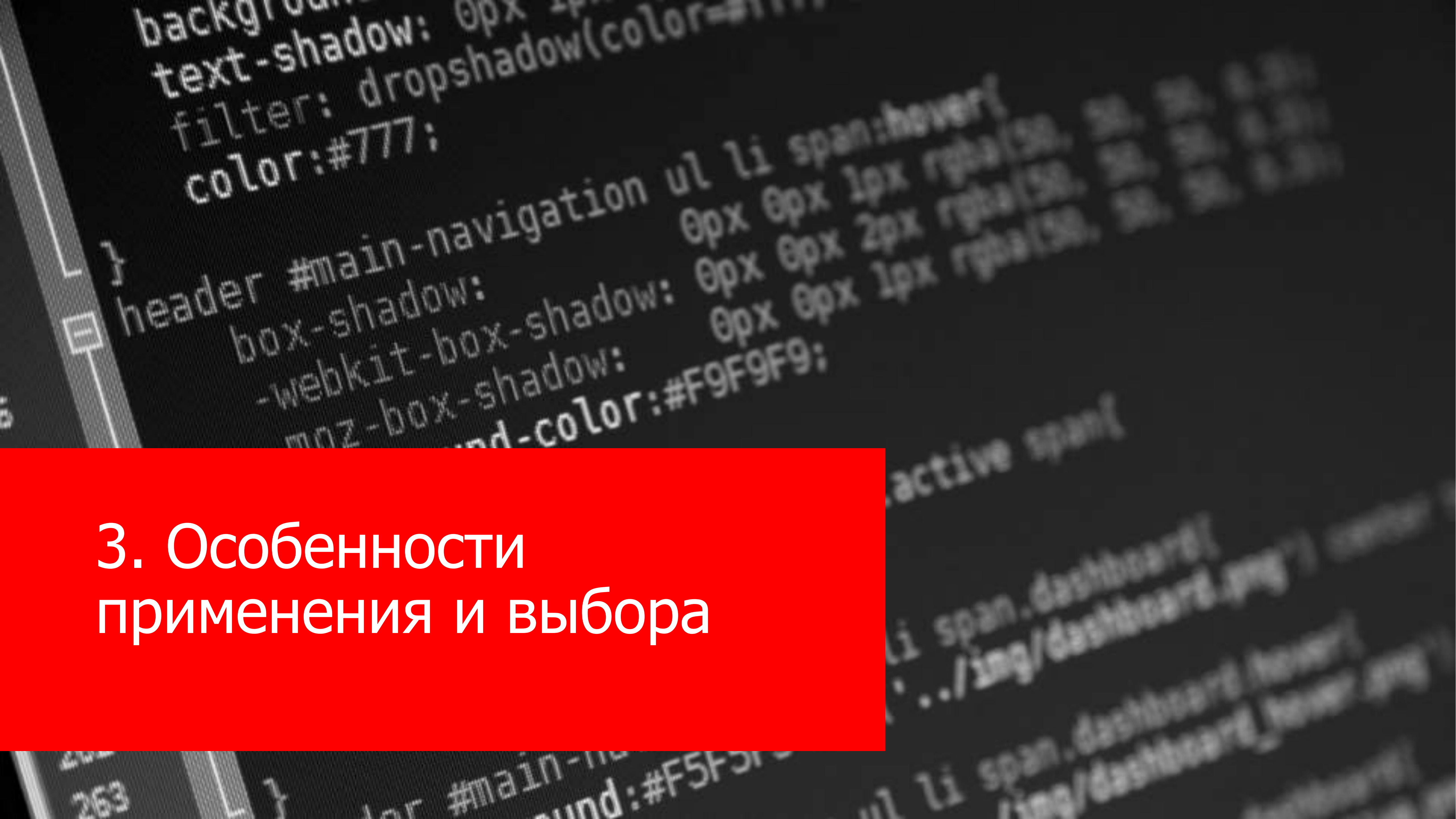
3: Соединить все основные точки, находящиеся на расстоянии Eps радиуса одна от другой.

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

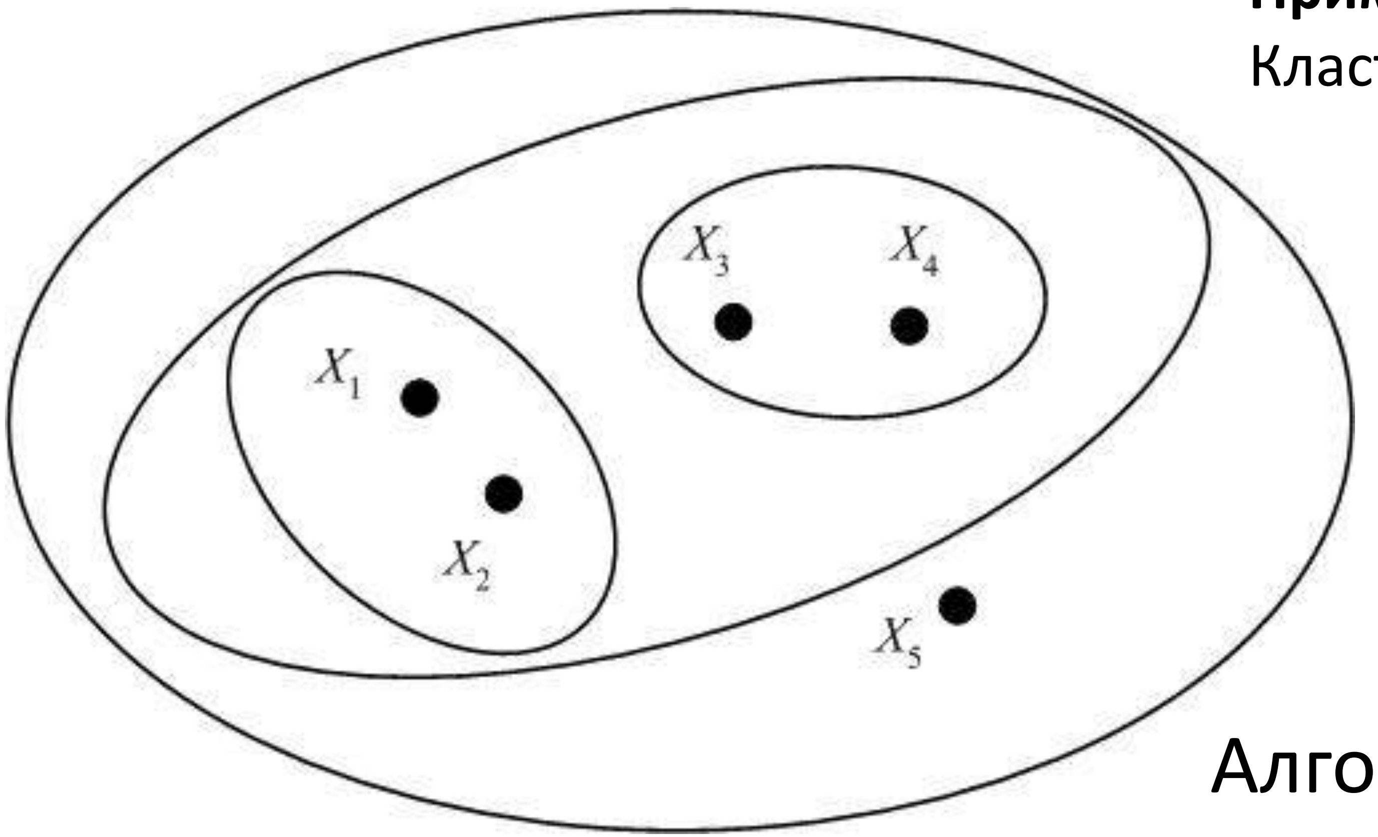
Зачем нужны разные алгоритмы кластеризации

- 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

IT

Алгоритмы

Алгоритмы и структуры данных

Методы машинного обучения

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

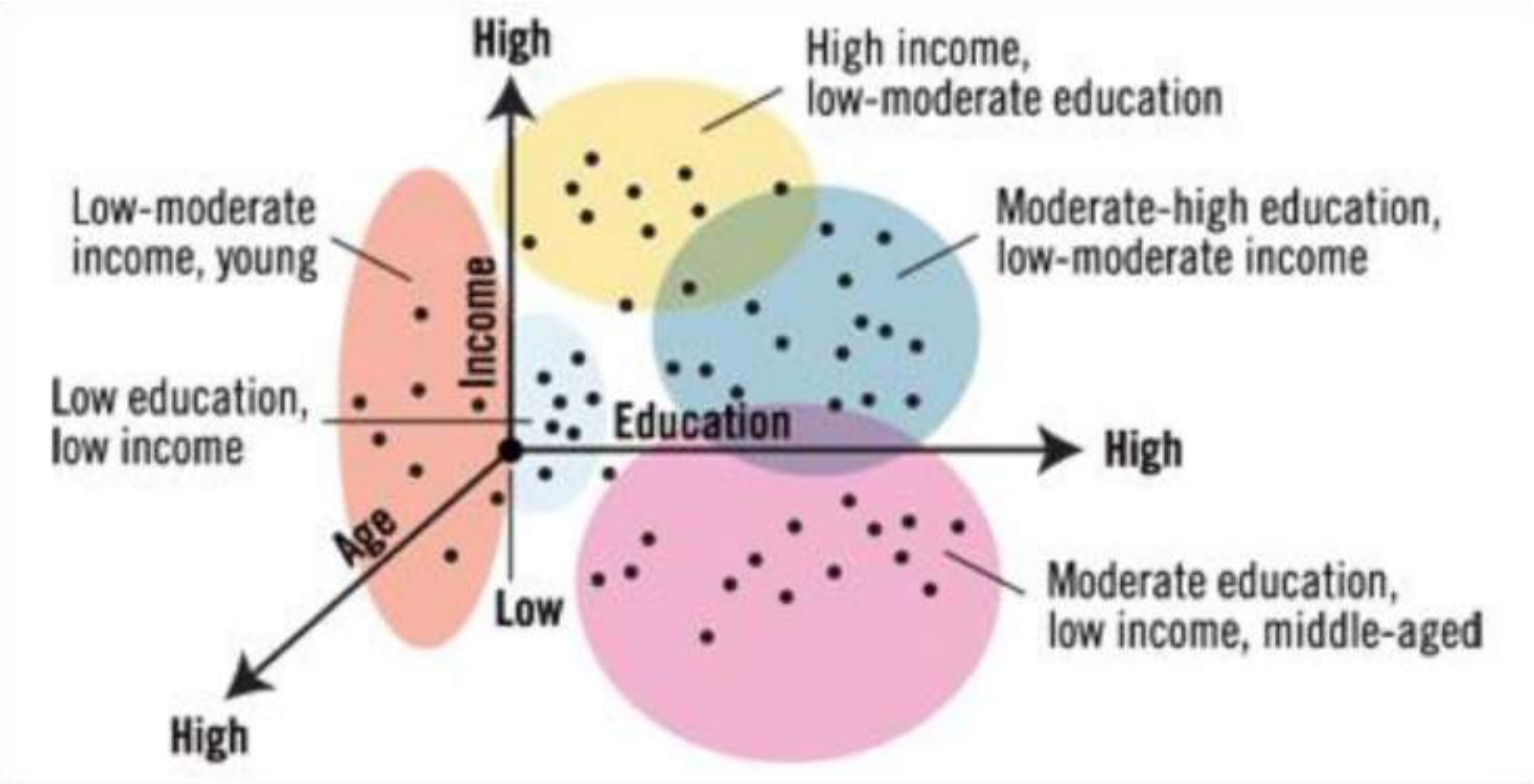
Открытие Олимпиады в Сочи посмотрели несколько миллиардов человек

Олимпиада в Сочи открыта

Церемония открытия Олимпиады в Сочи. Онлайн-репортаж

Основная задача или вспомогательная

Сегментация целевой аудитории



Основная задача или вспомогательная

Кластеризация символов по написанию для улучшения распознавания

Пример: квантизация изображений

Original image (96,615 colors)

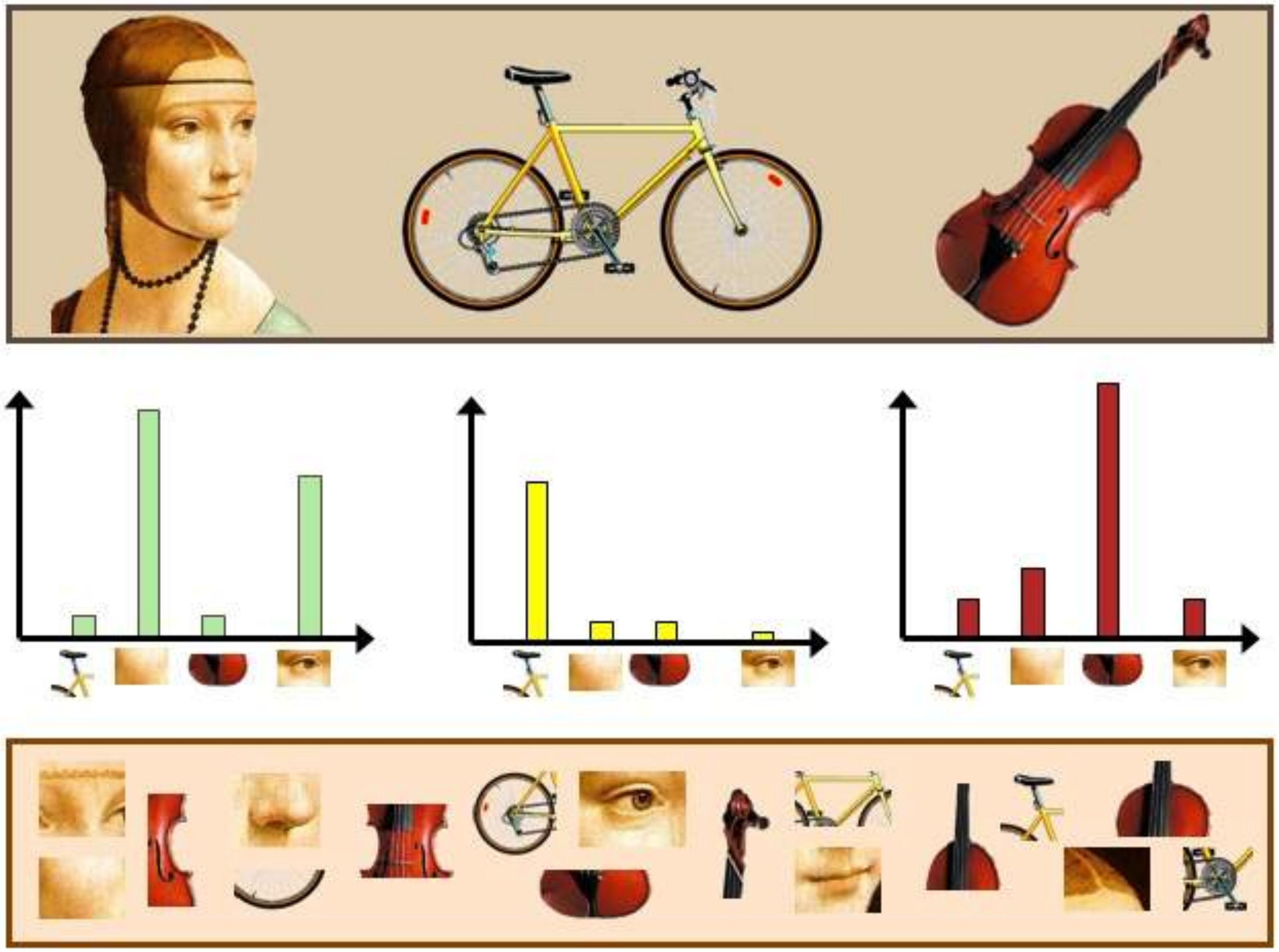
Пример: квантизация изображений

Quantized image (64 colors, Random)

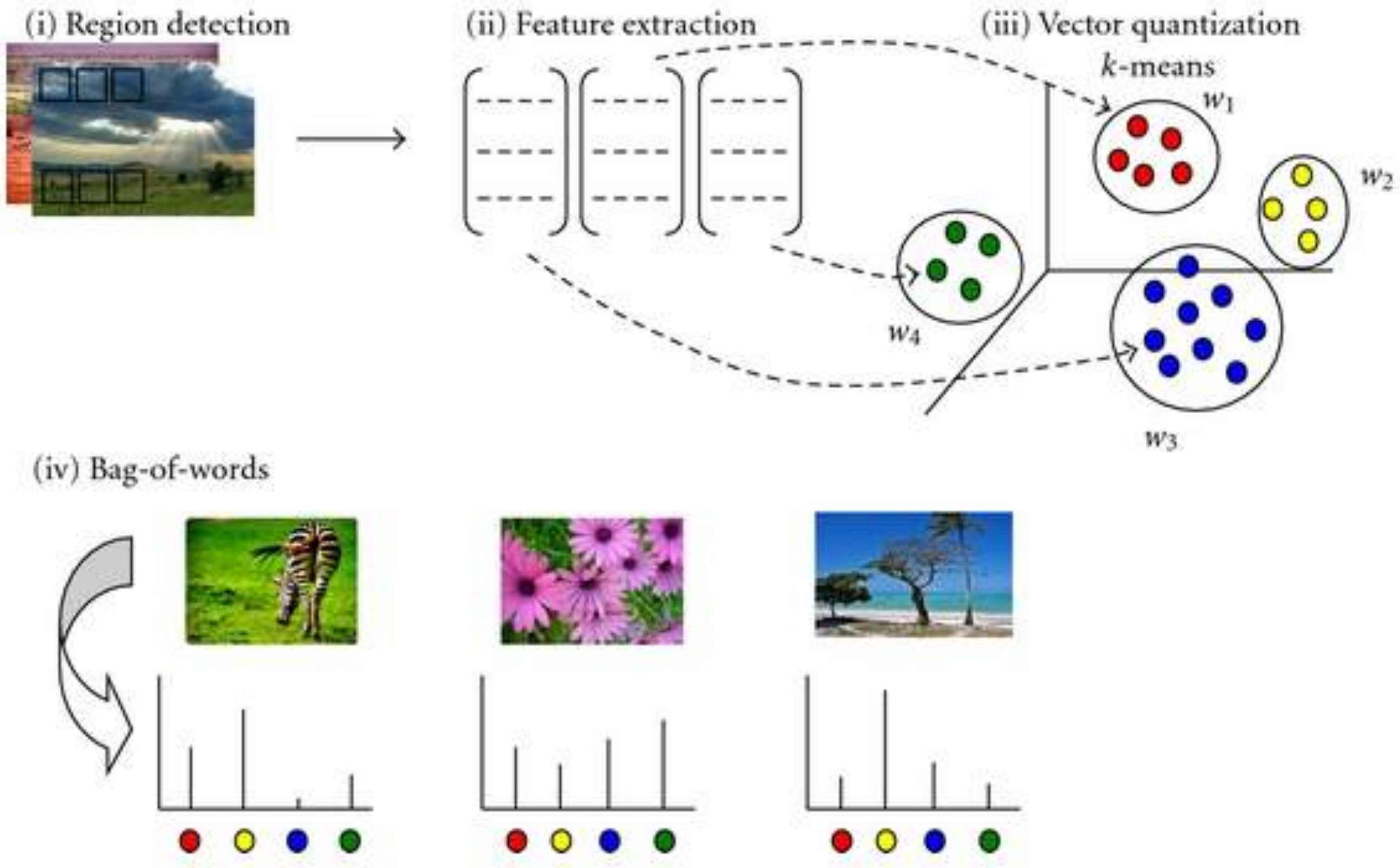
Пример: квантизация изображений

Quantized image (64 colors, K-Means)

Пример: мешок визуальных слов

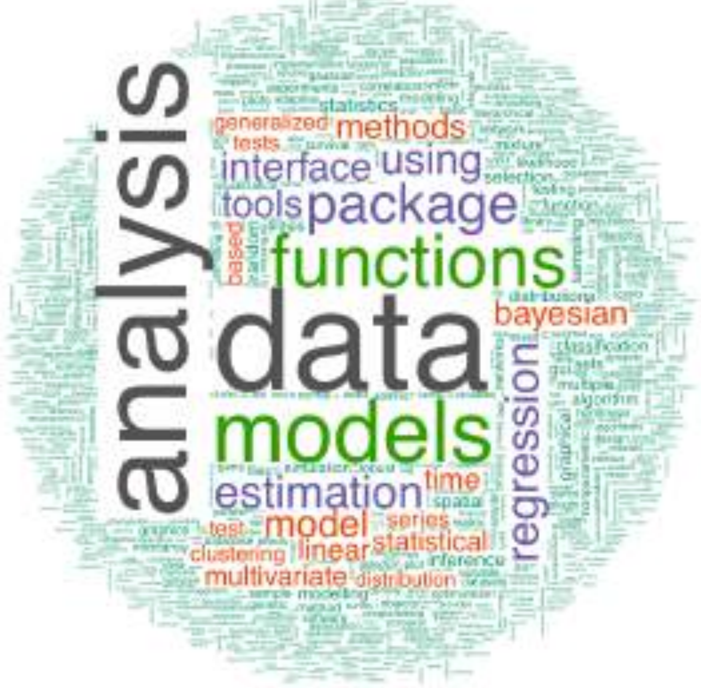


Пример: мешок визуальных слов



«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»



«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»

0.2

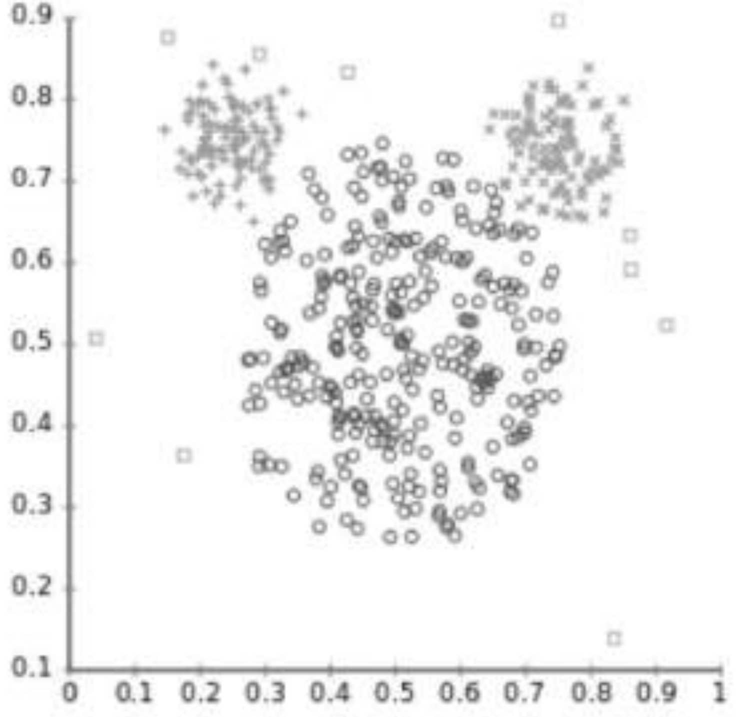
0.3

0.5

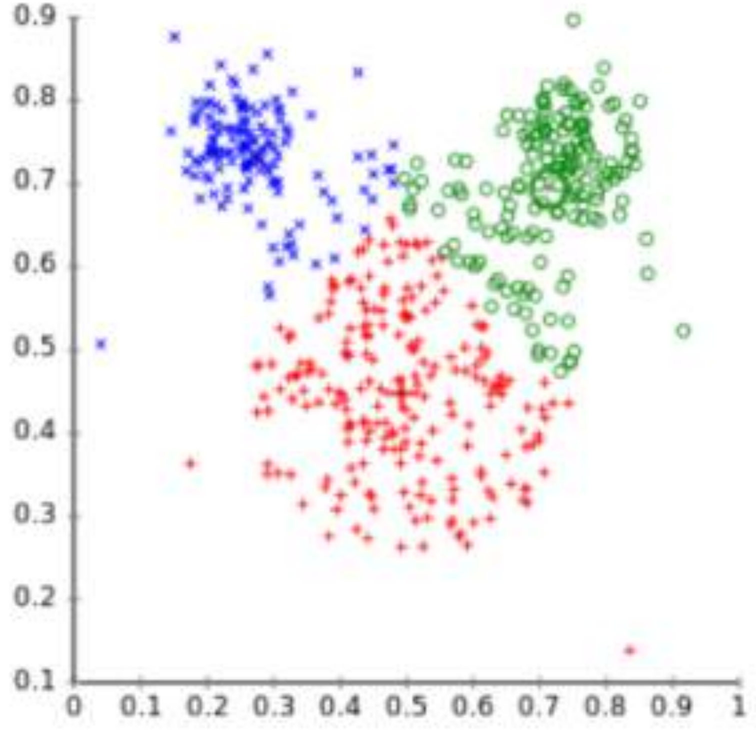
Резюме: чем могут отличаться задачи кластеризации

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

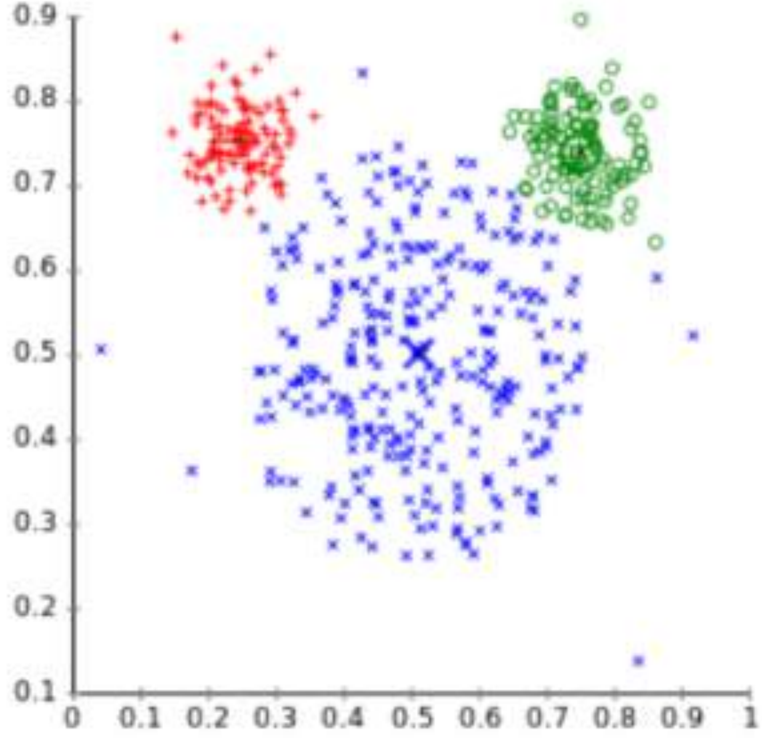
Различия в результатах работы методов



Исходная выборка
("Mouse" dataset)



Метод k средних
(K-Means)



EM-алгоритм

Алгоритмы

Рассмотренные нами:

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

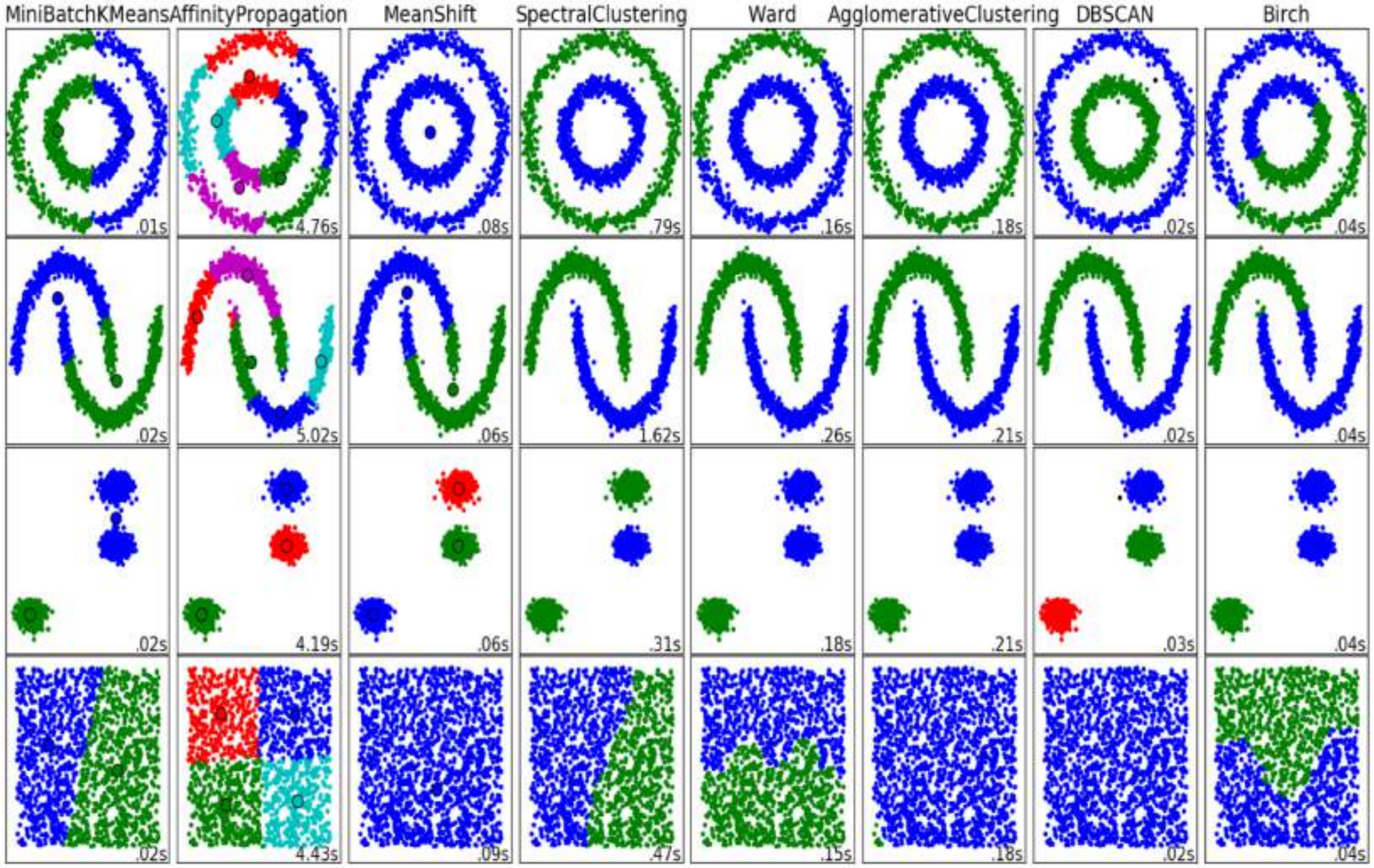
Алгоритмы

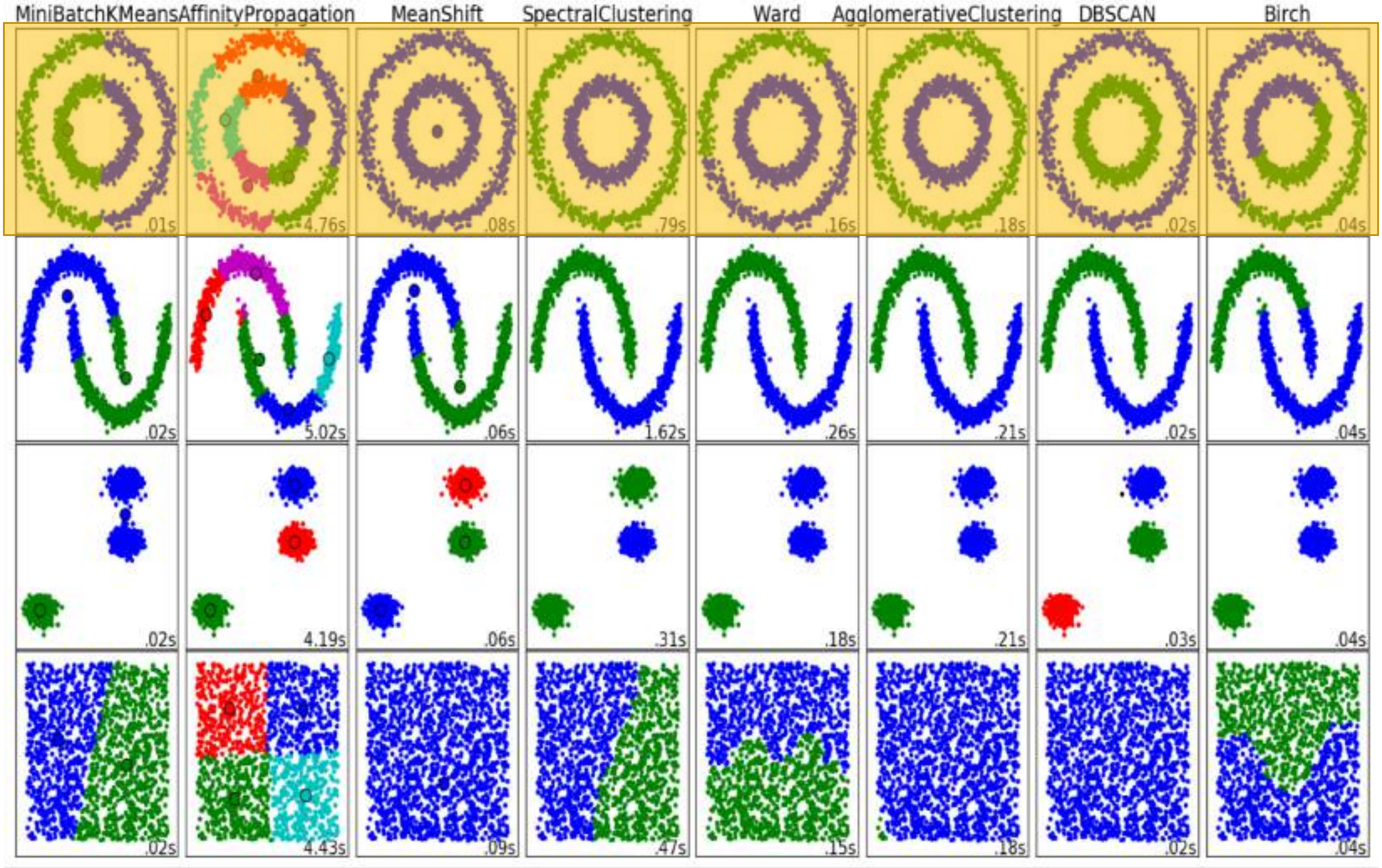
Рассмотренные нами:

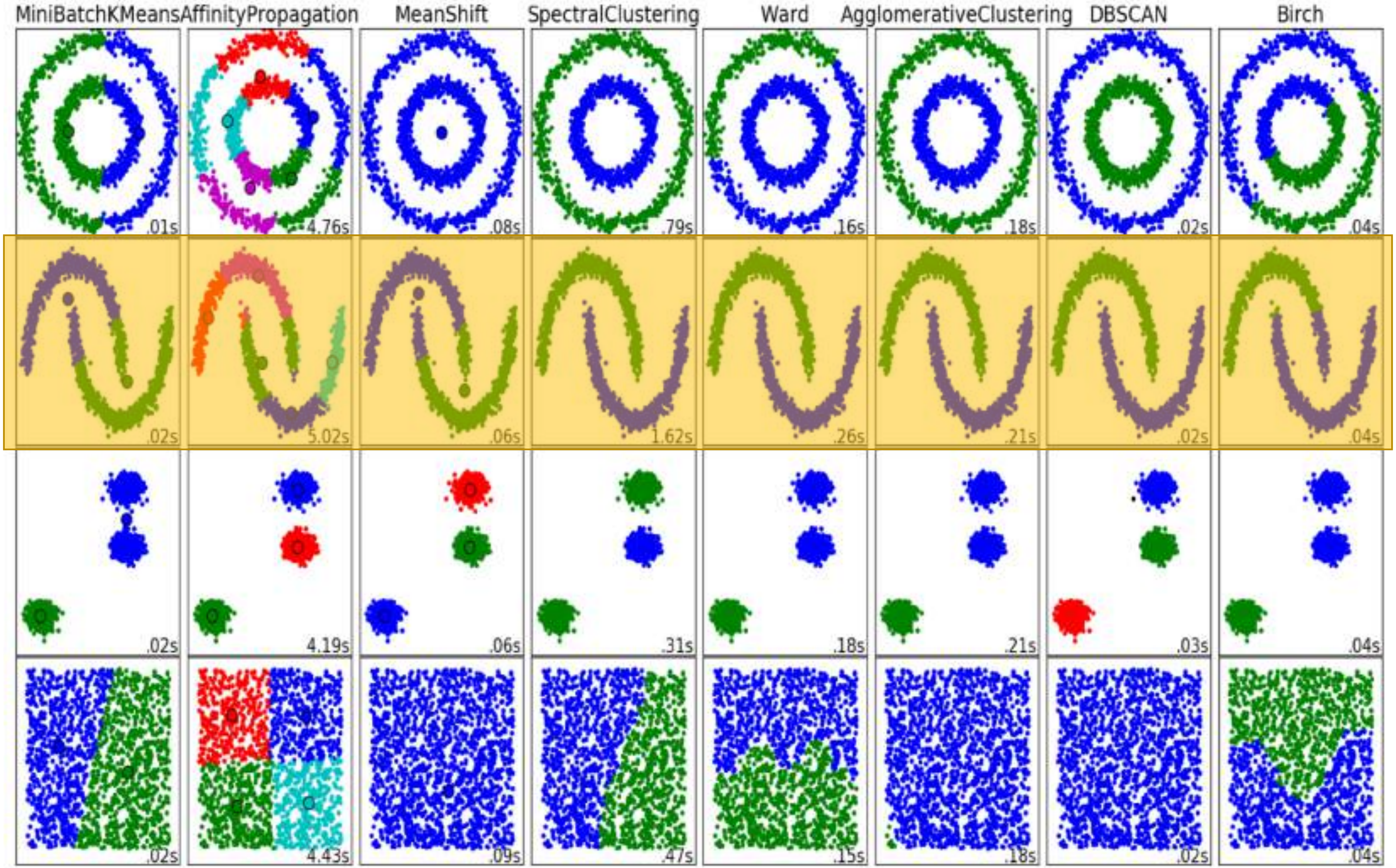
- К-средних
- ЕМ-алгоритм
- Агломеративная иерархическая кластеризация
- DBSCAN

В scikit-learn:

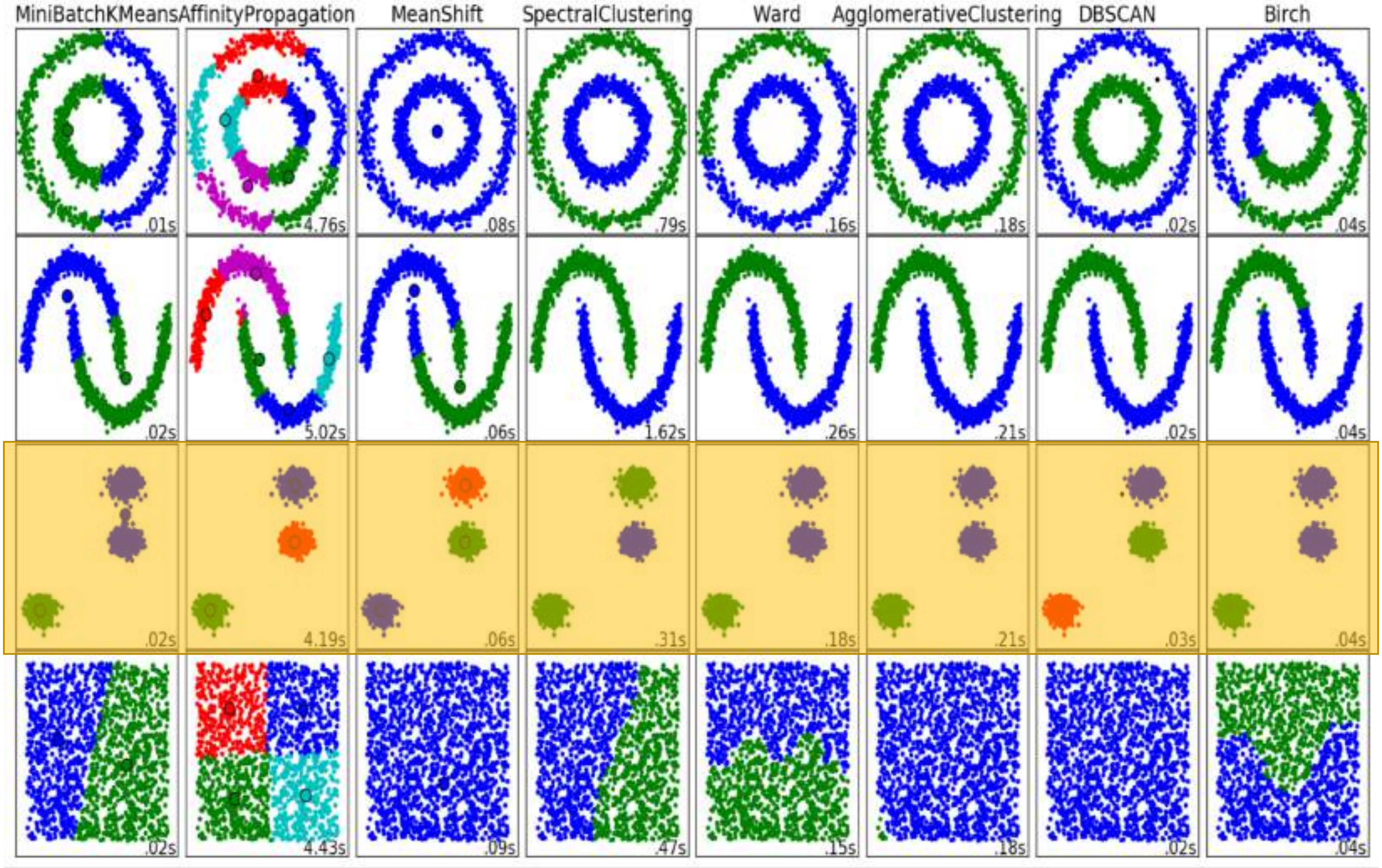
KMeans, MiniBatchKMeans, GaussianMixture, AgglomerativeClustering, Ward, DBSCAN, MeanShift, AffinityPropagation, SpectralClustering, Birch

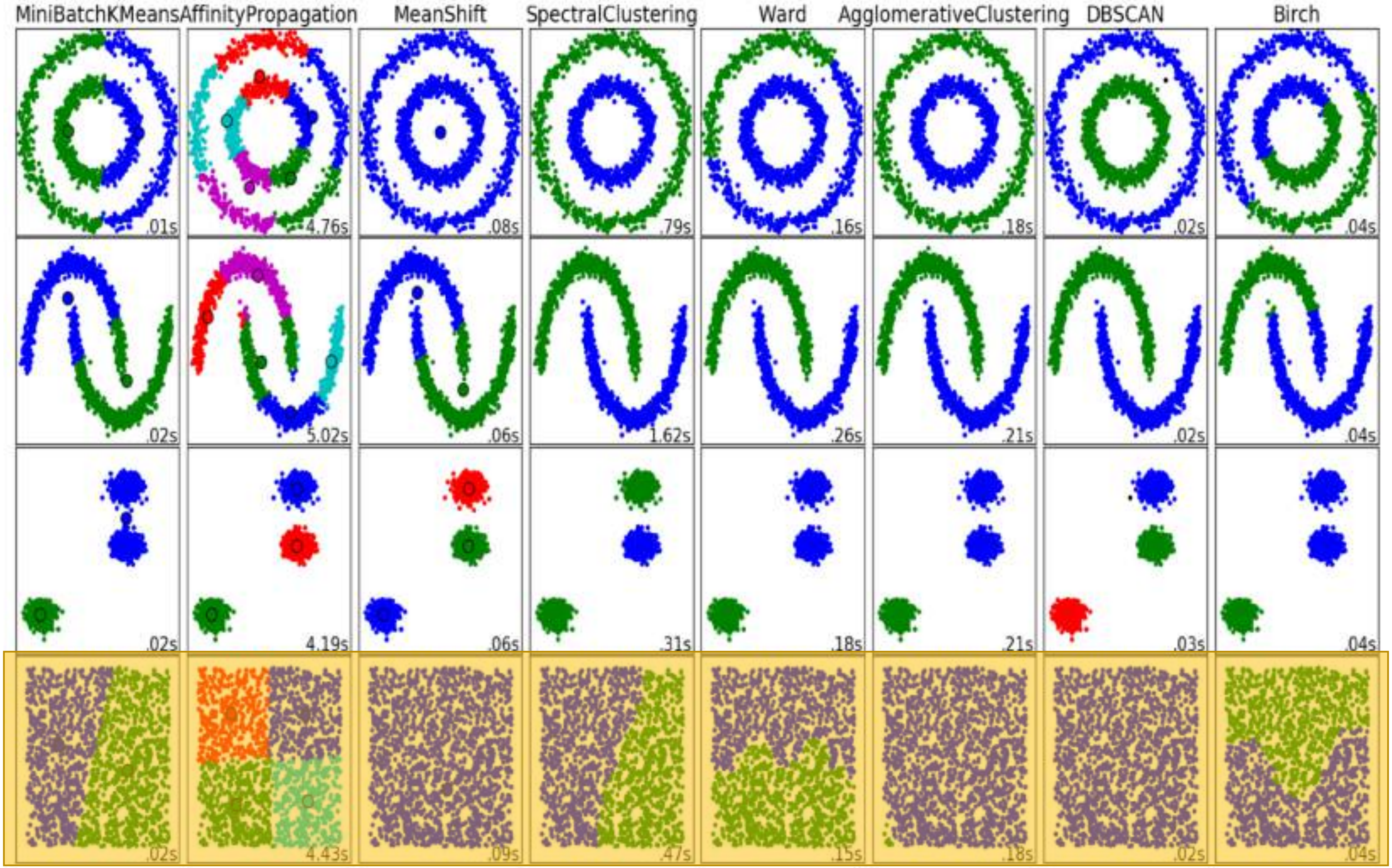

MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
4.76s
.08s
.79s
.16s
.18s
.02s
.04s
.02s
5.02s
.06s
1.62s
.26s
.21s
.02s
.04s
.02s
4.19s
.06s
.31s
.18s
.21s
.03s
.04s
.02s
4.43s
.09s
.47s
.15s
.18s
.02s
.04s

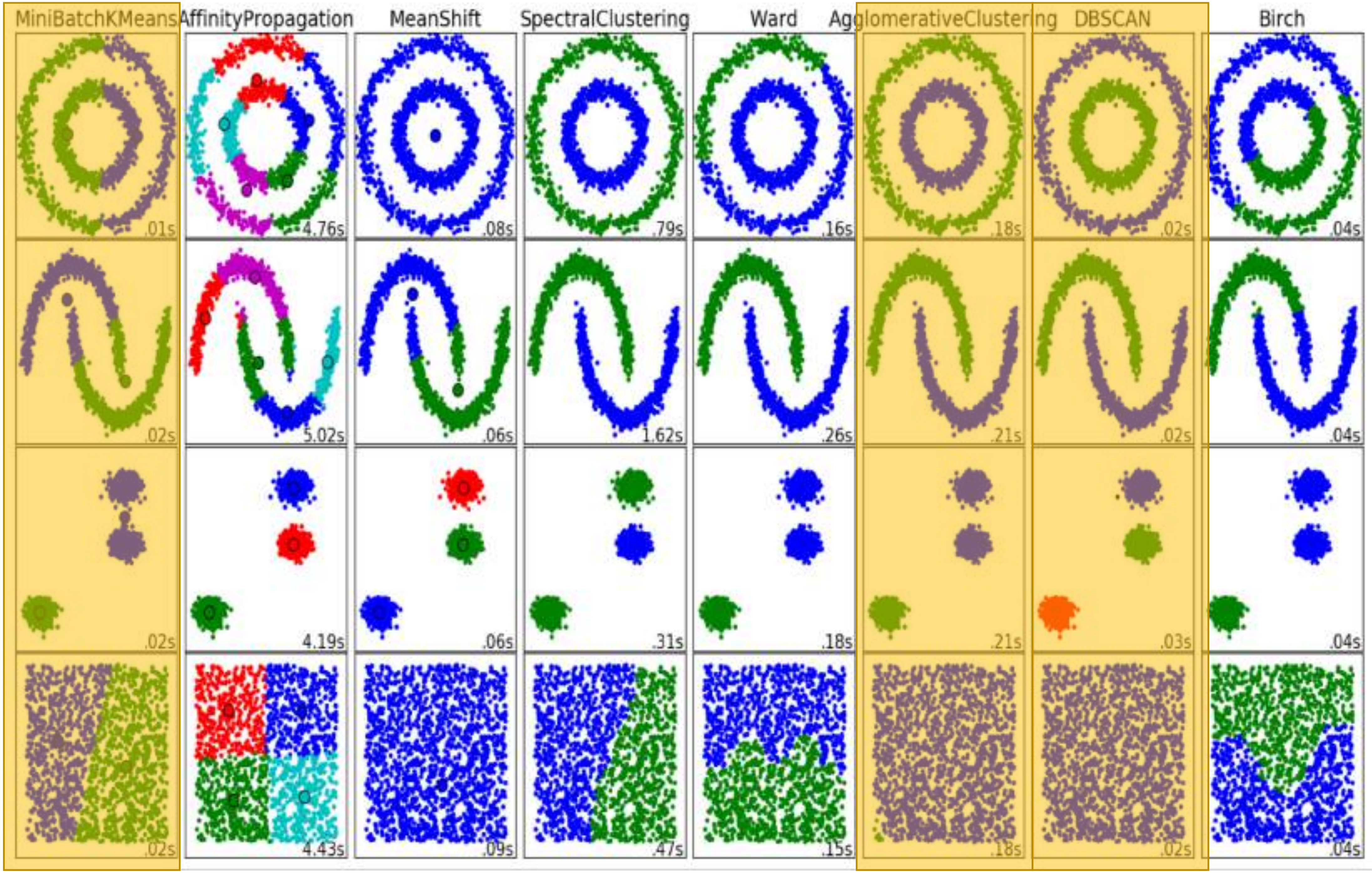

MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
4.76s
.08s
.79s
.16s
.18s
.02s
.04s
.02s
5.02s
.06s
1.62s
.26s
.21s
.02s
.04s
.02s
4.19s
.06s
.31s
.18s
.21s
.03s
.04s
.02s
4.43s
.09s
.47s
.15s
.18s
.02s
.04s



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


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


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

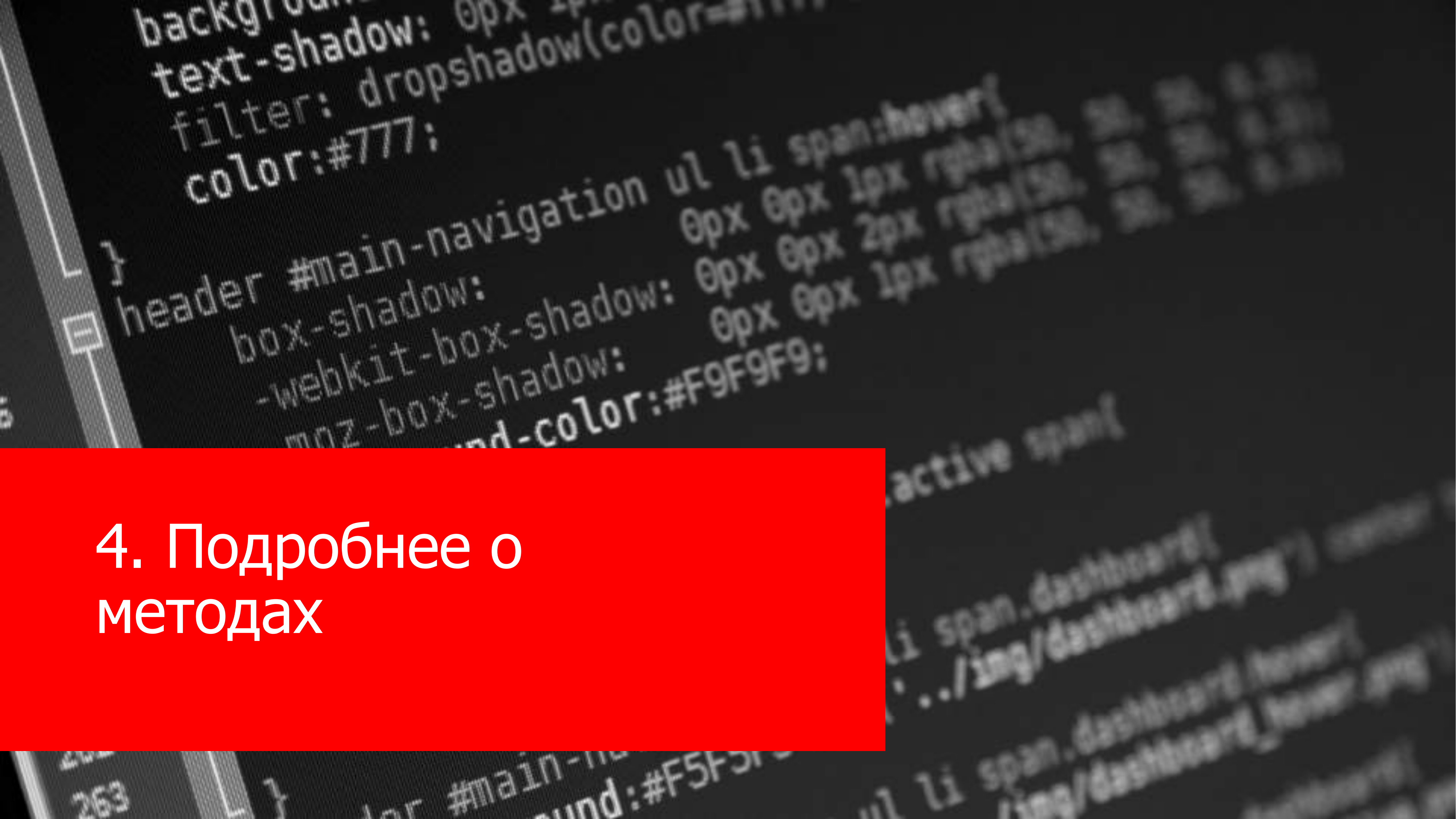

MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
4.76s
.08s
.79s
.16s
.18s
.02s
.04s
.02s
5.02s
.06s
1.62s
.26s
.21s
.02s
.04s
.02s
4.19s
.06s
.31s
.18s
.21s
.03s
.04s
.02s
4.43s
.09s
.47s
.15s
.18s
.02s
.04s

Метод	Параметры	Масштабируемость	Use-case	Геометрия
KMeans	Число кластеров	Очень много примеров (MiniBatch), среднее число кластеров	Выпуклые, примерно одинаковые кластеры	Евклидово расстояние

Метод	Параметры	Масштабируемость	Use-case	Геометрия
KMeans	Число кластеров	Очень много примеров (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)

Метод	Параметры	Масштабируемость	Use-case	Геометрия
KMeans	Число кластеров	Очень много примеров (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 примеров и много кластеров	Много кластеров, нужно задавать метрику	Любая метрика/функция близости

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



4. Подробнее о методах

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

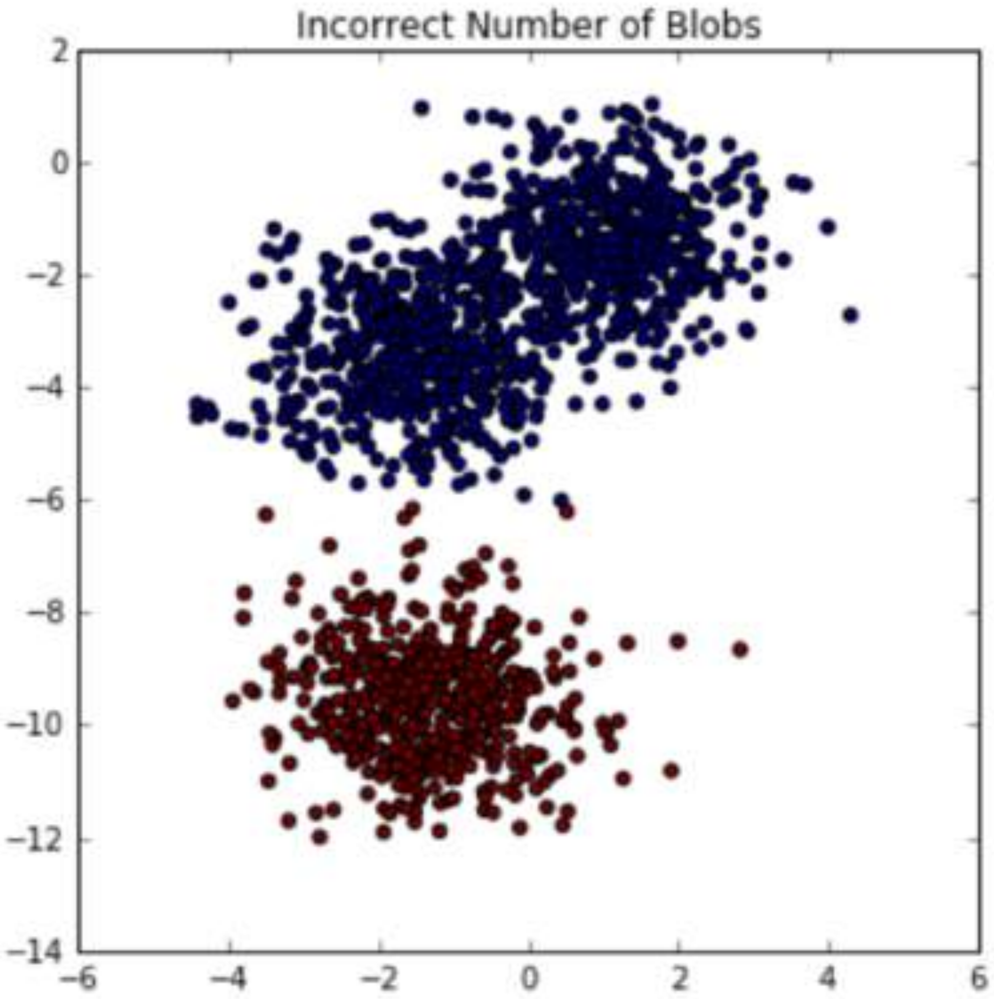
Понижение размерности пространства

- 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
- Варианты: отбор признаков, метод главных компонент (PCA), сингулярное разложение (SVD) – об этом – далее в курсе

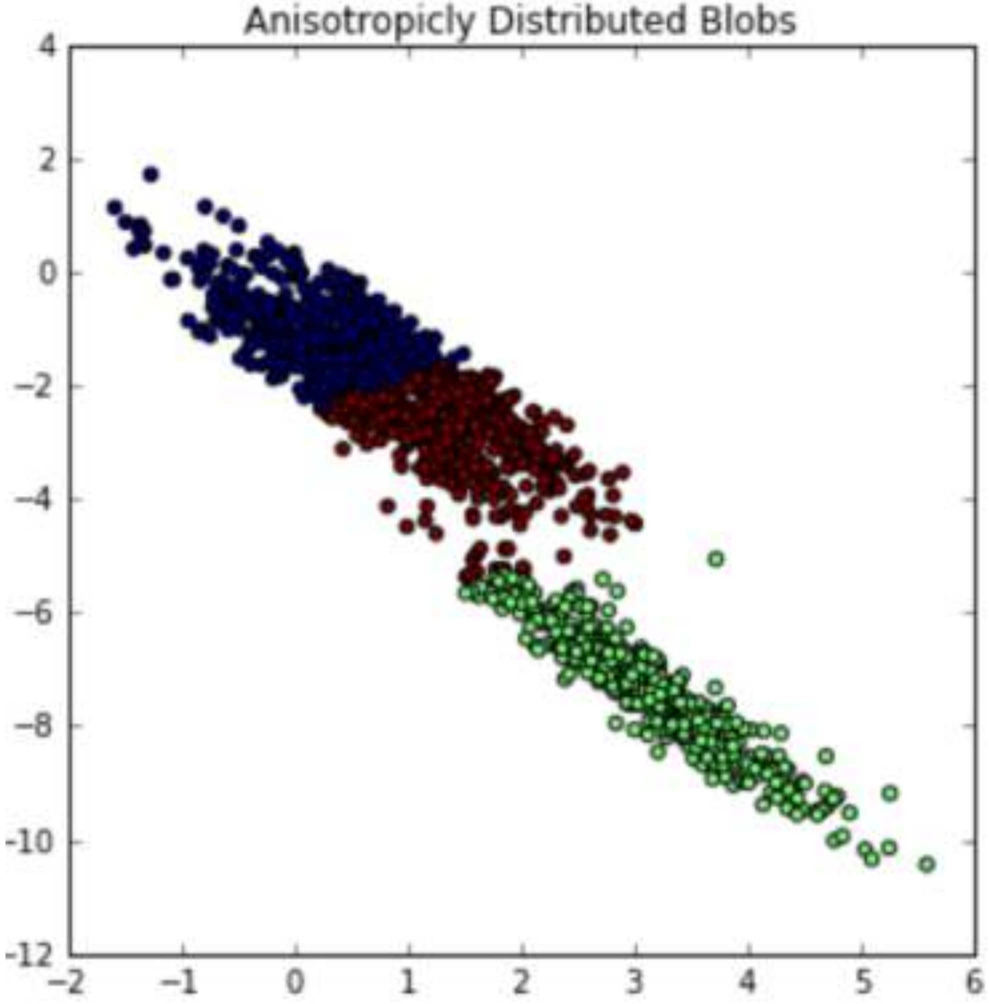
K Means++

- 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 K?
- 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

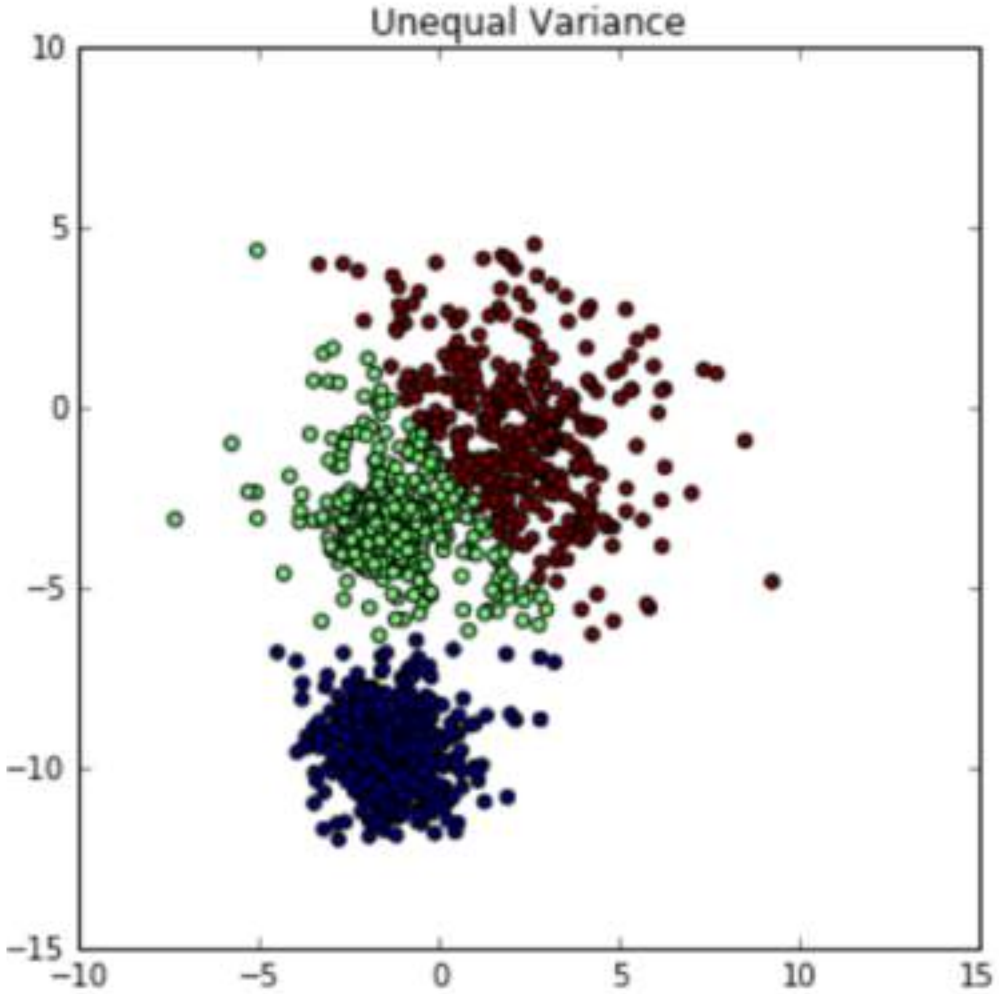
К Means и разные формы кластеров



K Means и разные формы кластеров



K Means и разные формы кластеров



Что оптимизирует K Means

Среднее внутрикластерное расстояние:

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

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

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

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

Что оптимизирует K Means

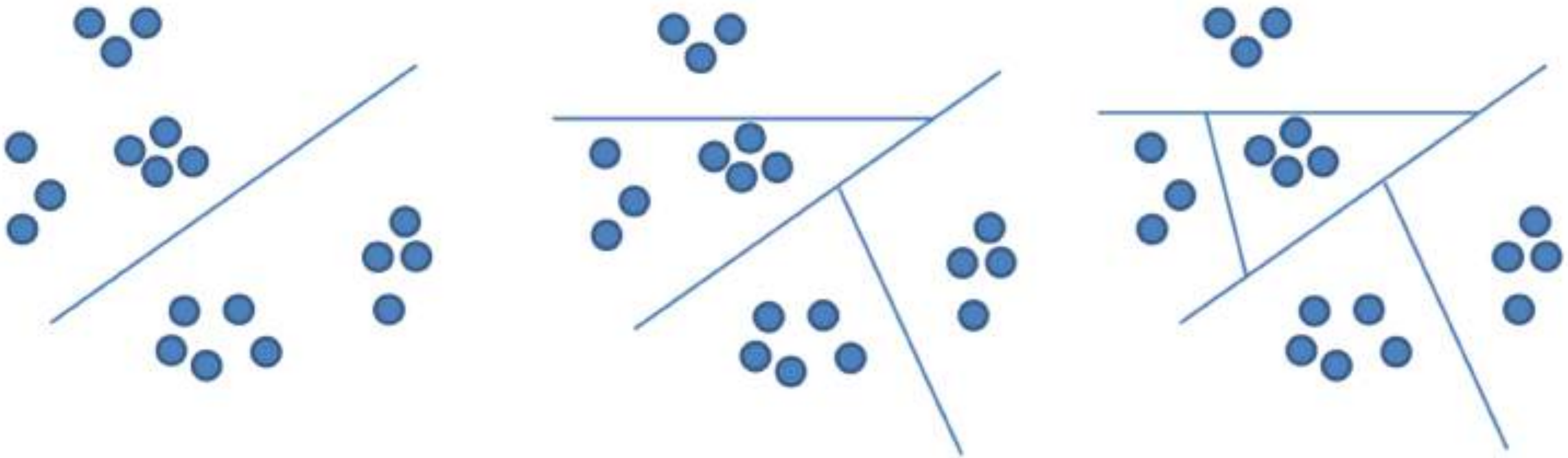
K Means итеративно минимизирует среднее внутрикластерное расстояние:

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

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

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

Подбор числа кластеров: BisectKMeans



Кластеризация в EM: постановка задачи

Модель порождения данных:

- Априорные вероятности кластеров - $w_1, \ldots, w_K$
- Плотности распределения кластеров - $p_1(x), \ldots, p_K(x)$
- Плотность распределения вектора признаков x:

$$p(x) = \sum_{j=1}^{K} w_j p_j(x)$$

Кластеризация в ЕМ: постановка задачи

Модель порождения данных:

- Априорные вероятности кластеров - $w_1, \dots, w_K$
- Плотности распределения кластеров - $p_1(x), \dots, p_K(x)$
- Плотность распределения вектора признаков x:

$$p(x) = \sum_{j=1}^{K} w_j p_j(x)$$

Что будем делать:

По выборке оценим параметры модели: $w_1, \dots, w_K$ и $p_1(x), \dots, p_K(x)$

Кластеризация в EM: постановка задачи

Модель порождения данных:

- Априорные вероятности кластеров - $w_1, \dots, w_K$
- Плотности распределения кластеров - $p_1(x), \dots, p_K(x)$
- Плотность распределения вектора признаков x:

$$p(x) = \sum_{j=1}^{K} w_j p_j(x)$$

Что будем делать:

По выборке оценим параметры модели: $w_1, \dots, w_K$ и $p_1(x), \dots, p_K(x)$

Зачем:

Сможем оценивать вероятность принадлежности к кластеру

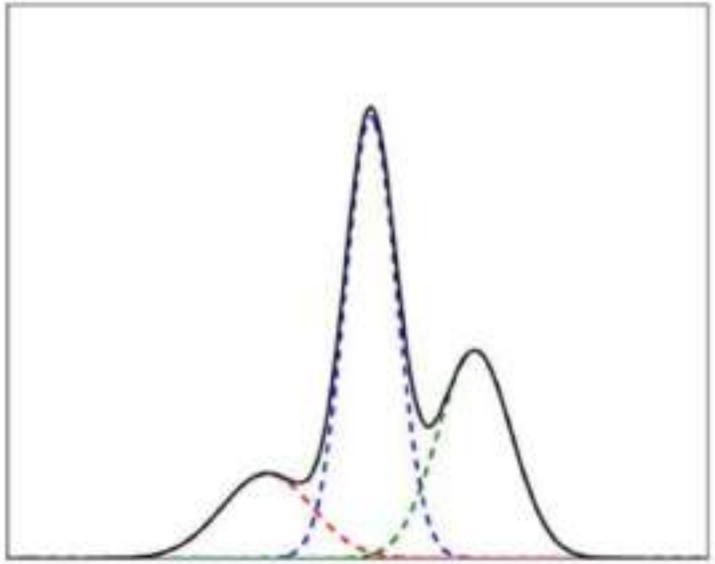
Постановка задачи: разделение смеси

$$p(x) = \sum_{j=1}^{K} w_j p_j(x)$$ $\Rightarrow$ Оценить: $w_1, \dots, w_K$ и $p_1(x), \dots, p_K(x)$

$$p_j(x) = \varphi(\theta_j; x)$$

Например, $p_j(x)$ - плотность нормального распределения (со своими параметрами для каждой компоненты)

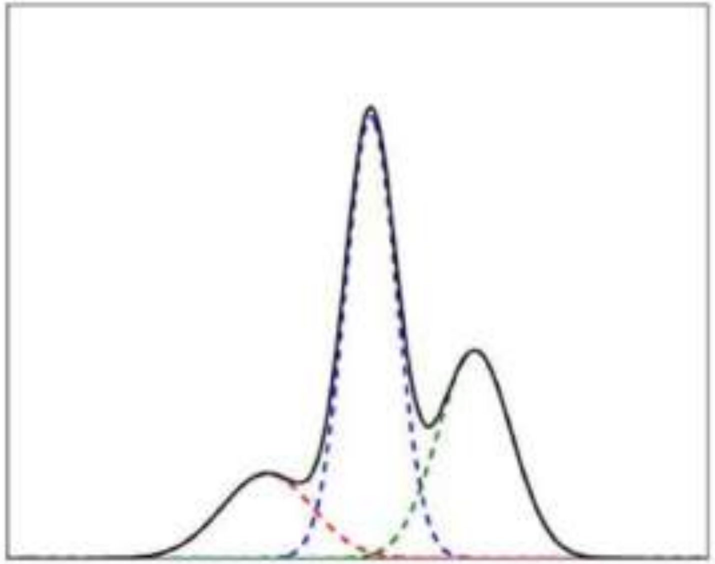
Почему не решить задачу «в лоб»



$$p(x) = \sum_{j=1}^{K} w_j p_j(x), \qquad p_j(x) = \varphi(\theta_j; x)$$

$$w, \theta = argmax_{\theta, w} \sum_{j=1}^{K} \ln p(x_i)$$

Почему не решить задачу «в лоб»



$$p(x) = \boxed{\sum_{j=1}^{K} w_j p_j(x)}, \qquad p_j(x) = \varphi(\theta_j; x)$$

$$w\,,\theta = argmax_{\theta,w} \sum_{j=1}^{K} \ln \boxed{p(x_i)}$$

Простое объяснение ЕМ-алгоритма

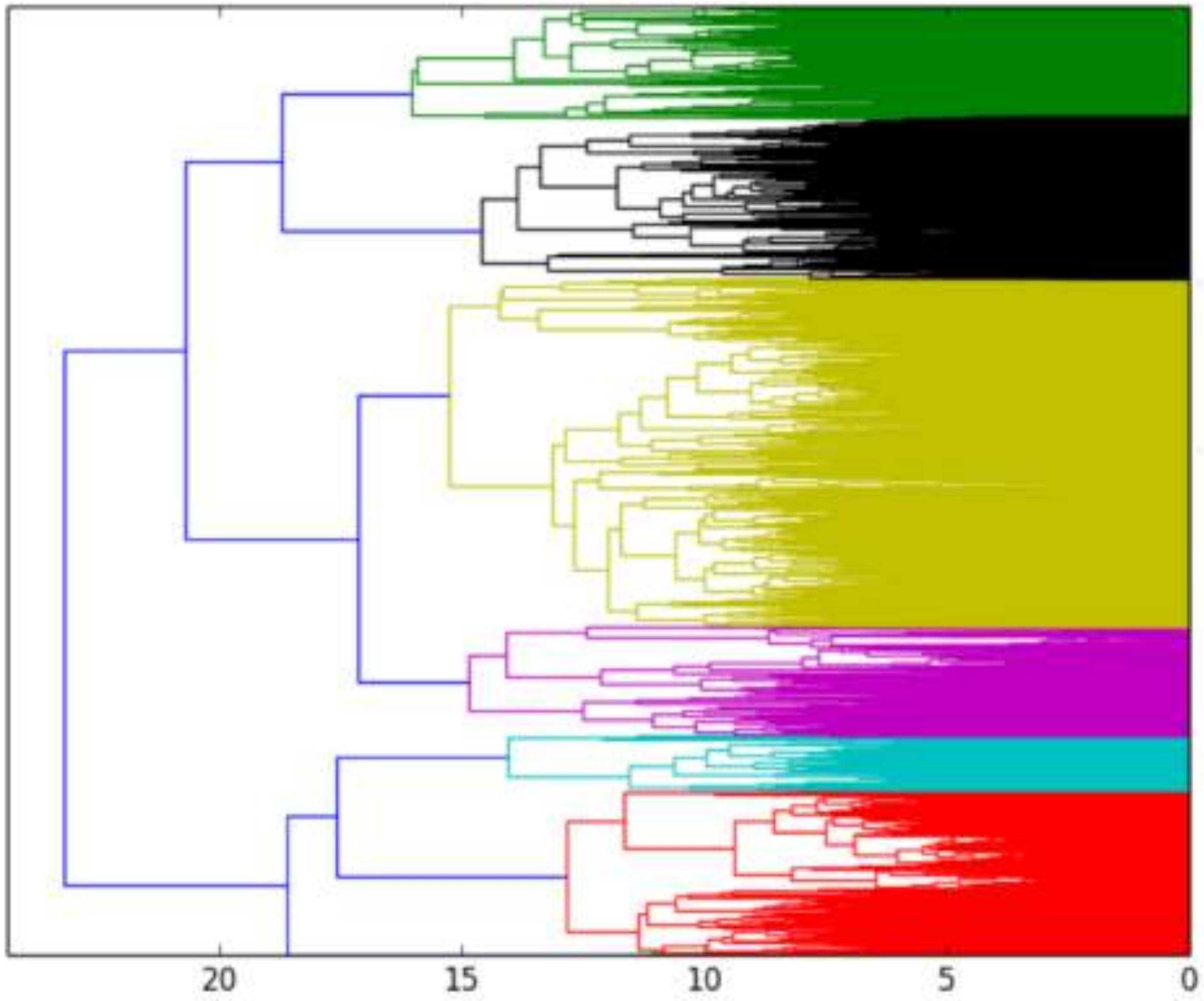
- Е-шаг:
  - Для задачи разделения смеси подходят $P(j|x_i)$
  - Расписав по формуле Байеса, получаем: $P(j|x_i) = \frac{w_j p_j(x_i)}{\sum_{k=1}^{K} w_k p_k(x_i)}$
- М-шаг:
  - Максимизируем правдоподобие по $w_1, \dots, w_K$ и $p_1(x), \dots, p_K(x)$, считая $P(j|x_i)$ константами
  - Если выписать производные по параметрам и приравнять к нулю, получаем:

$$w_j = \frac{1}{N}\sum_{i=1}^{N} g_{ji} \qquad \theta_j = argmax_\theta \sum_{i=1}^{N} g_{ji} \ln \varphi(\theta; x)$$

Какие еще задачи решаются с помощью ЕМ-алгоритма

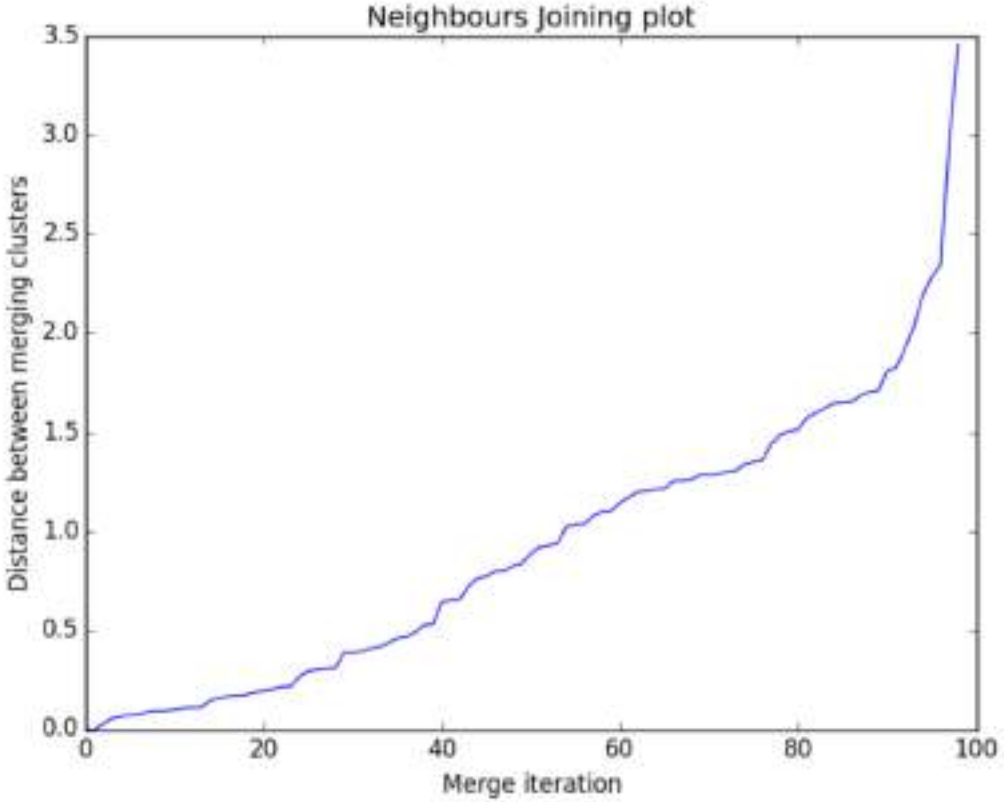
- Оценка параметров в других вероятностных моделях (не только в смеси распределений)
- Восстановление плотности распределения
- Классификация

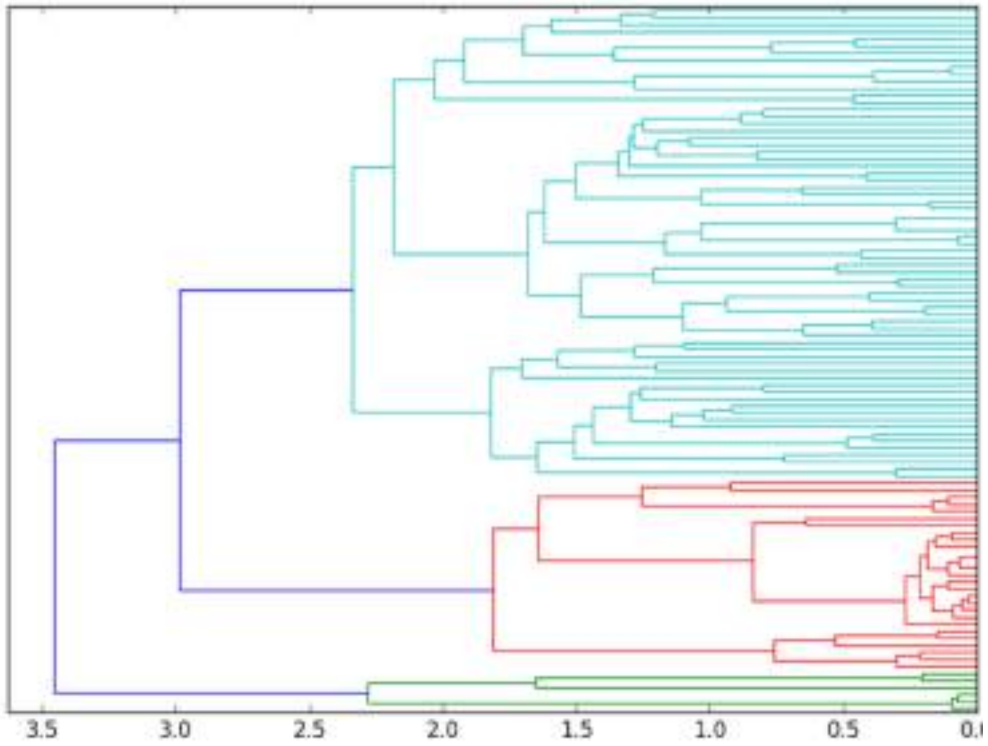
Пример: проблемы иерархической кластеризации



Пример: расстояние между кластерами

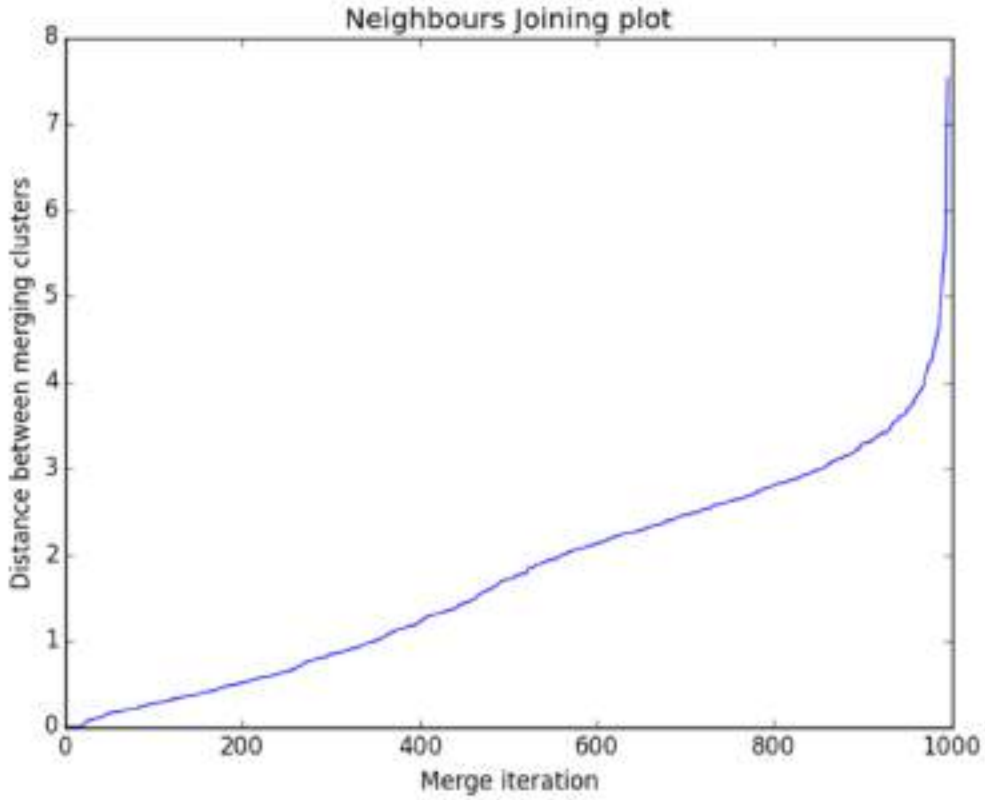
- На подвыборке из 100 писем

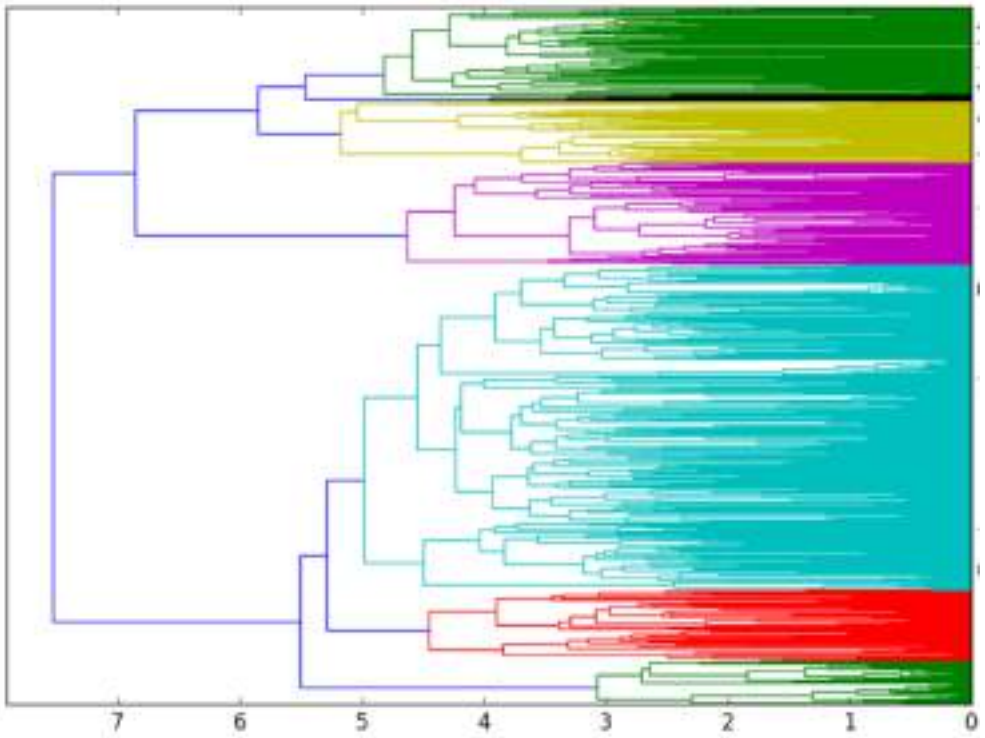




Пример: расстояние между кластерами

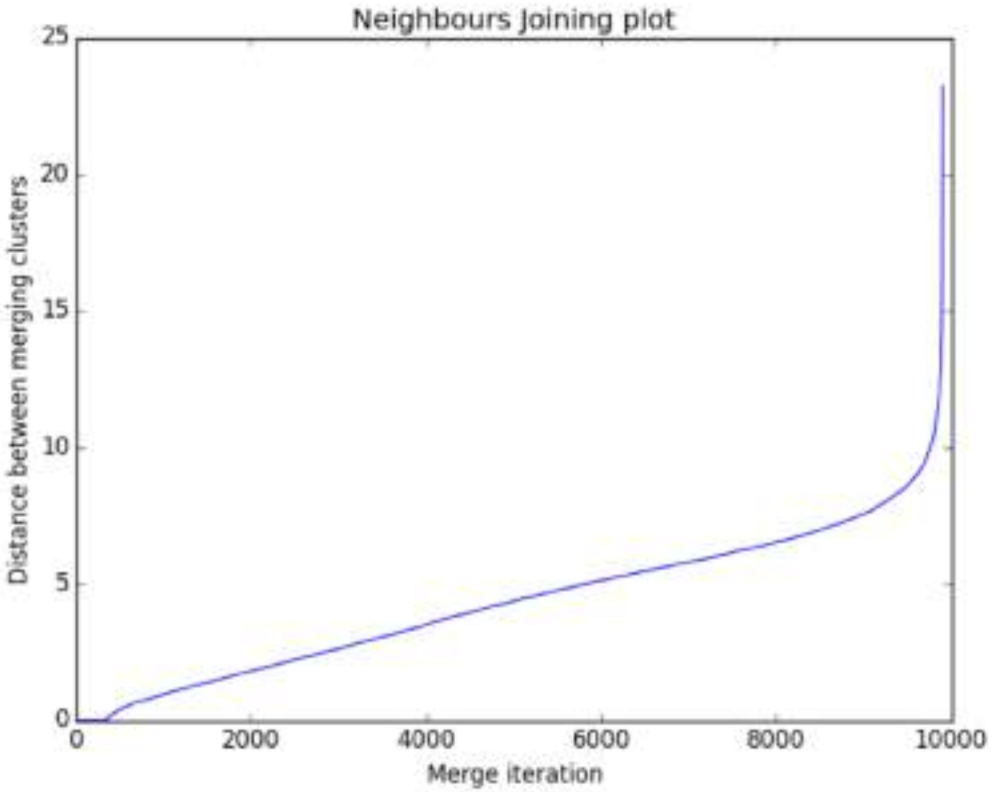
- На подвыборке из 1000 писем

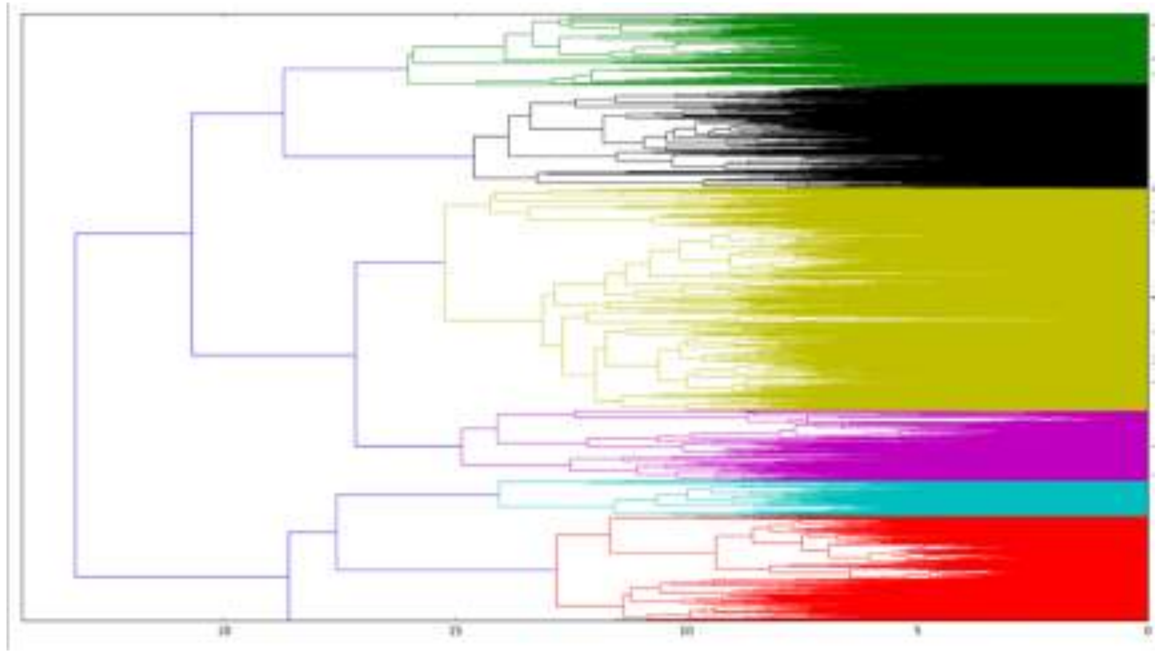




Пример: расстояние между кластерами

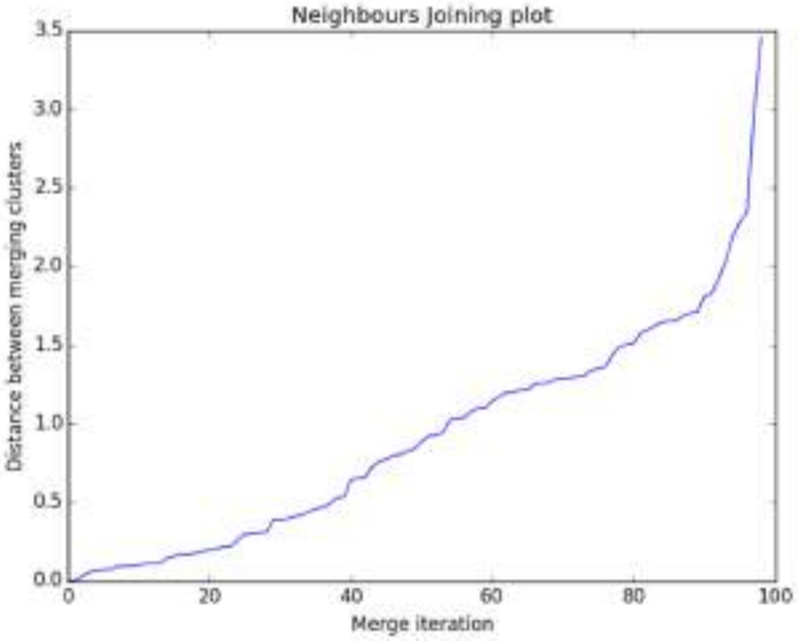
- На подвыборке из 10000 писем

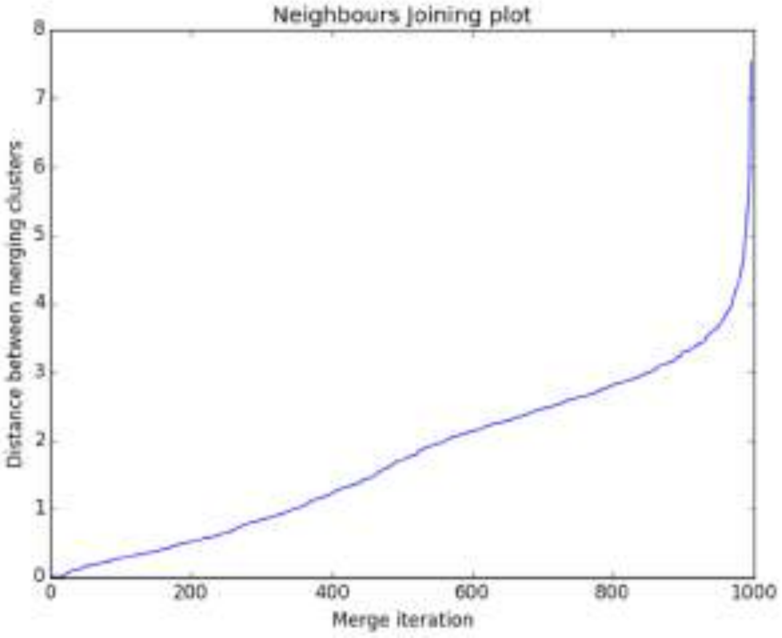


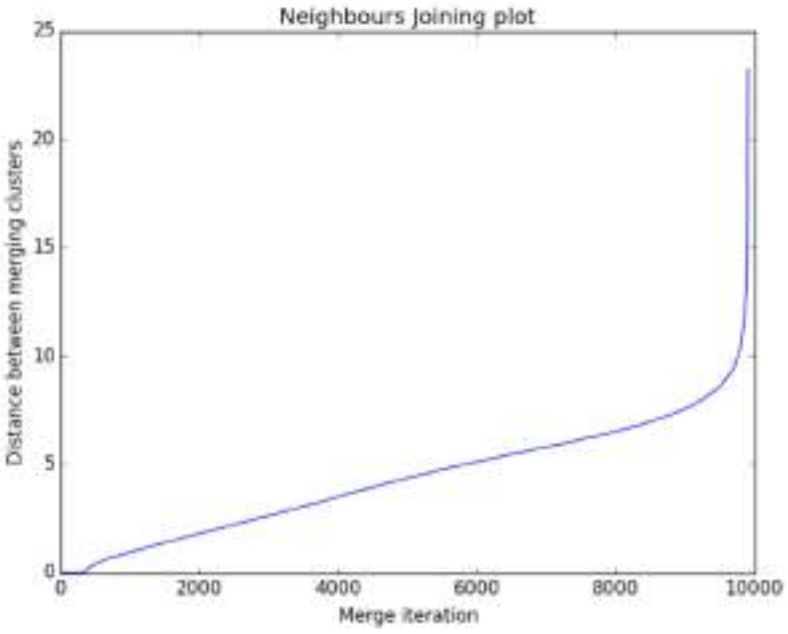


Пример: расстояние между кластерами

- Сравним графики: 100, 1000, 10000 писем

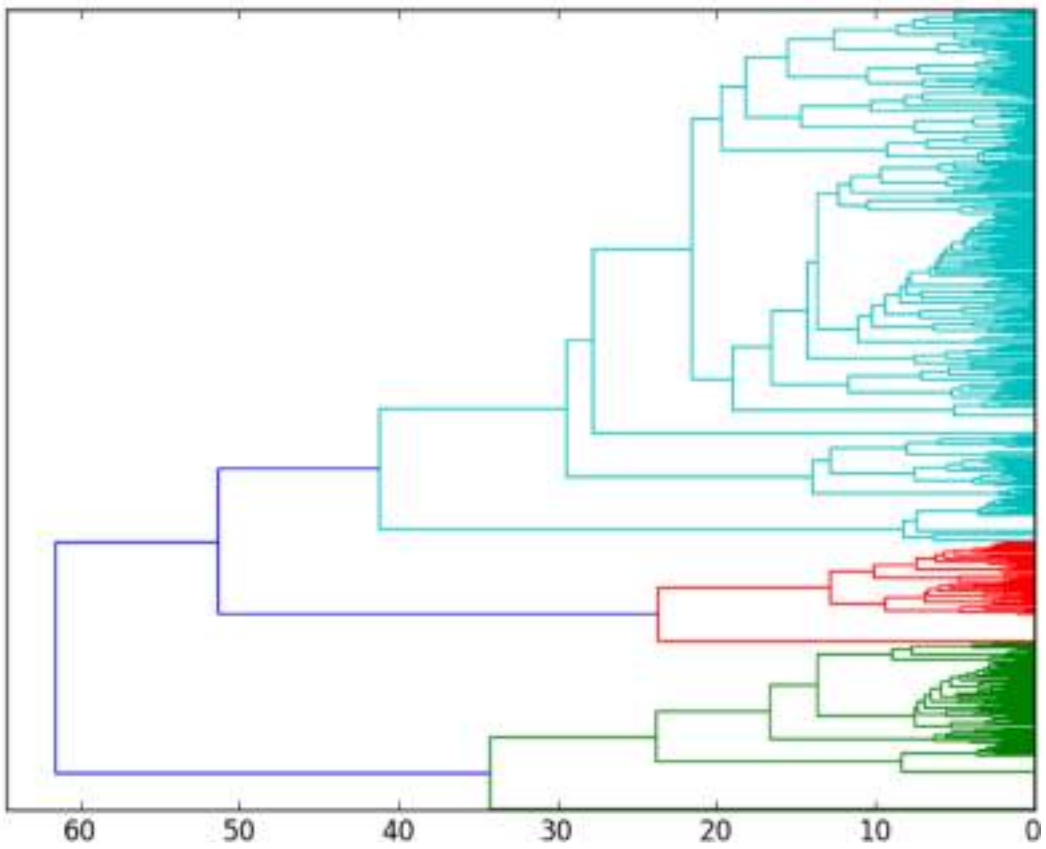




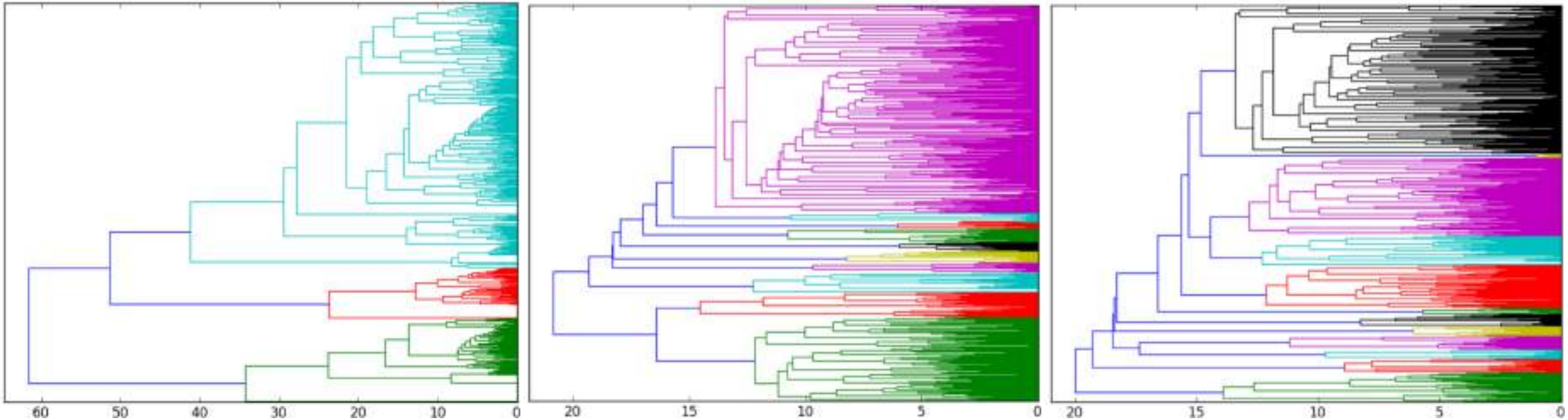


Пример: перекос в размерах кластеров

- Дендрограмма, построенная для другой выборки текстов:

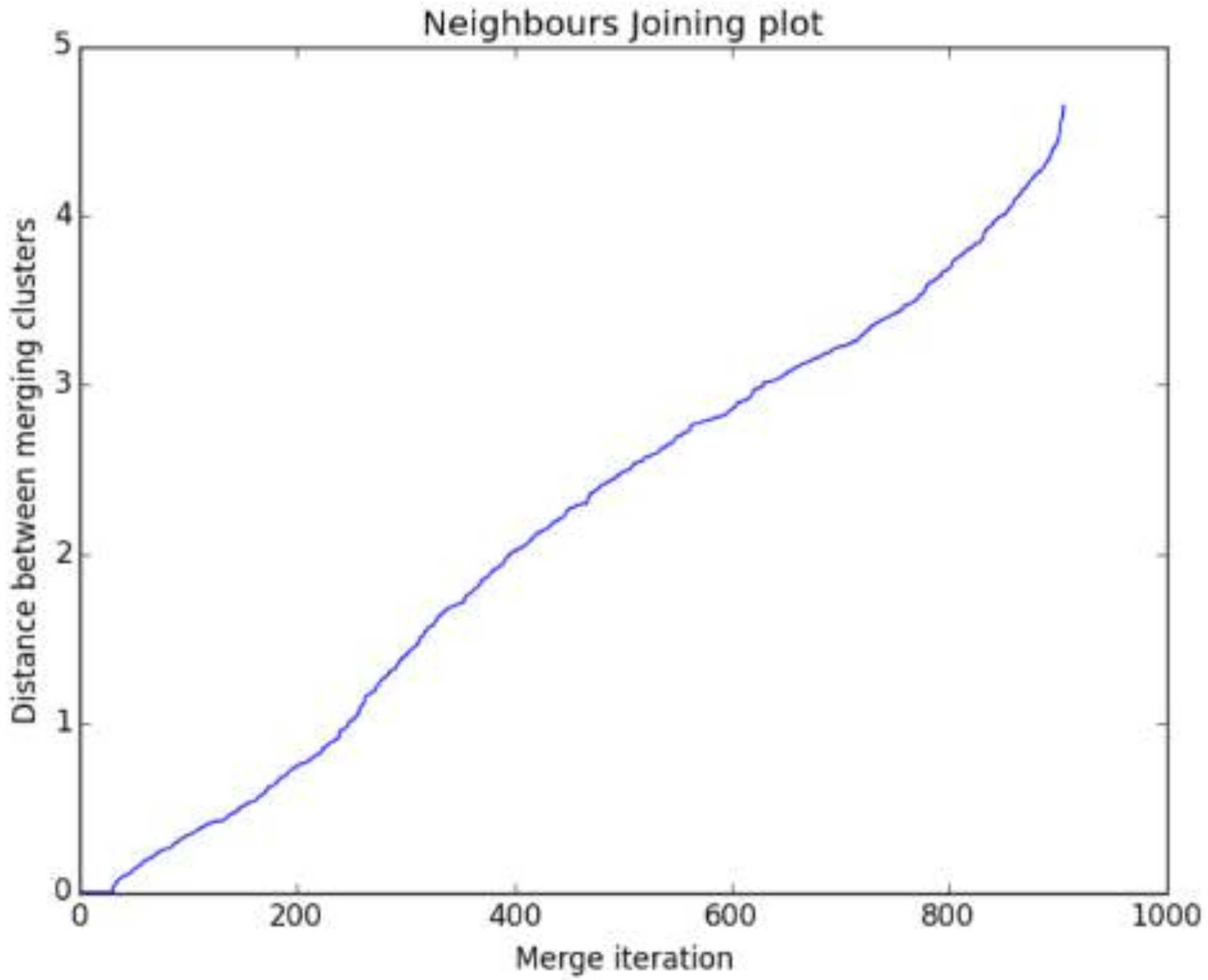


Пример: добавляем SVD

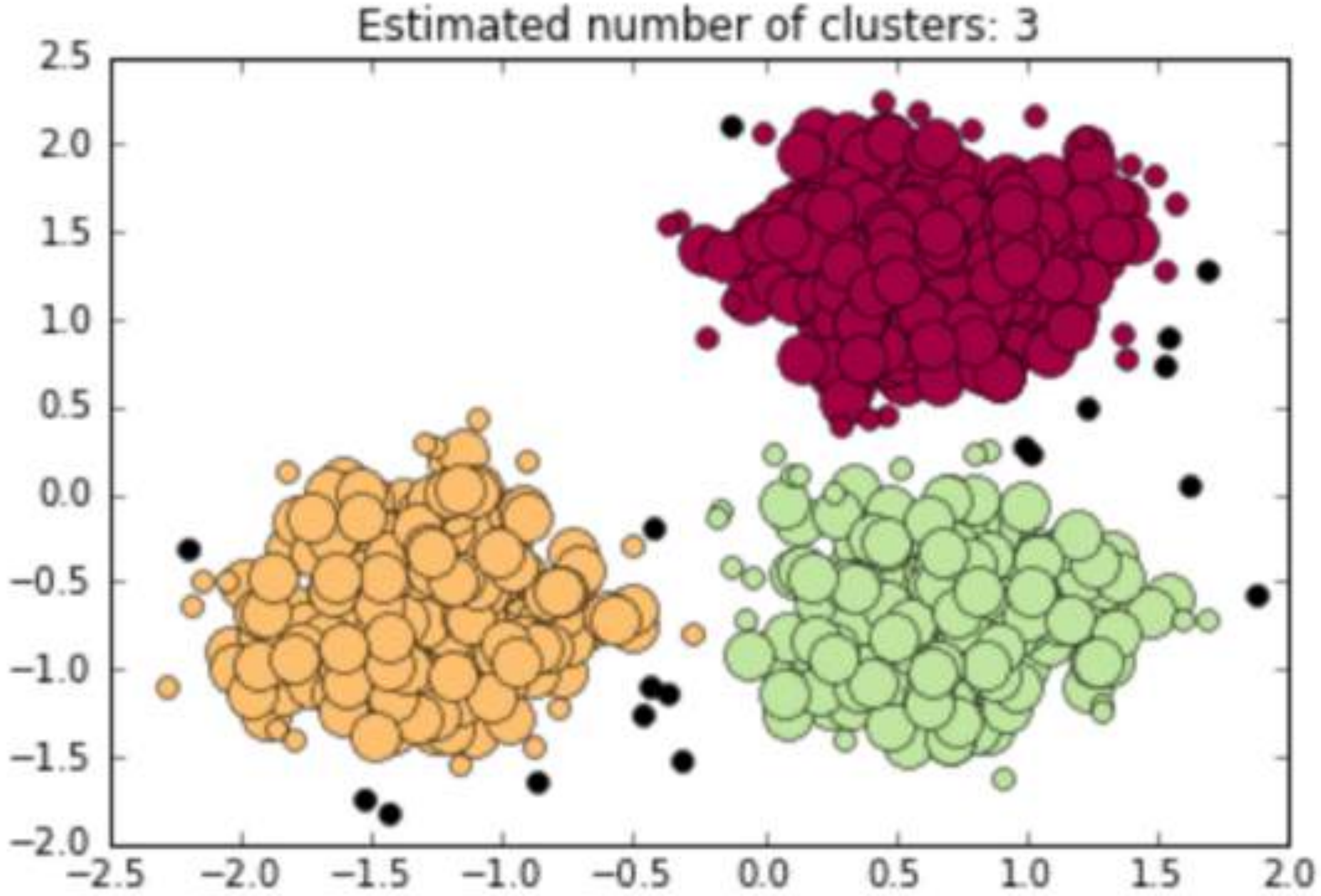


Исходные признаки SVD SVD (еще меньше компонент)

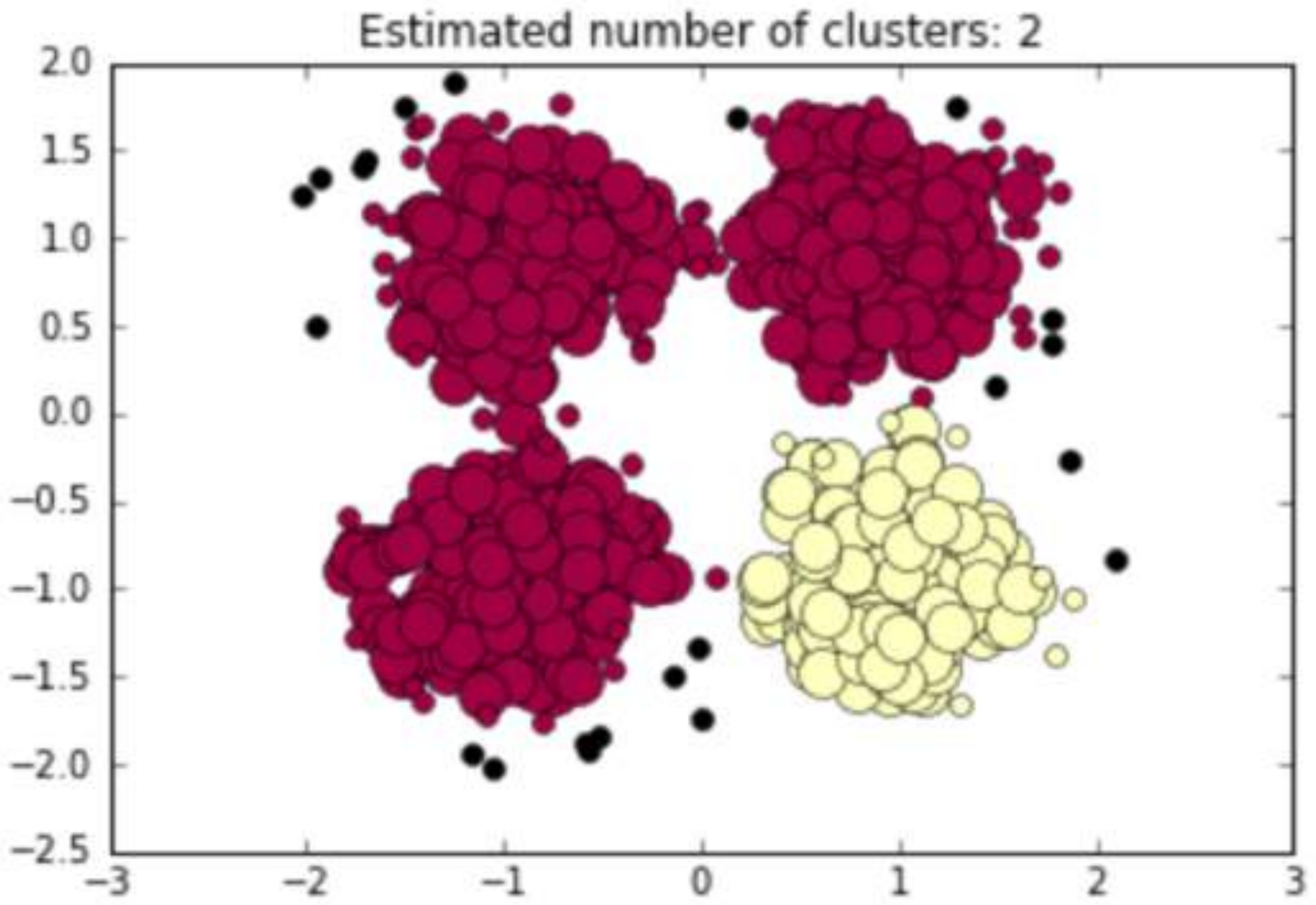
Пример: SVD и расстояние при слиянии



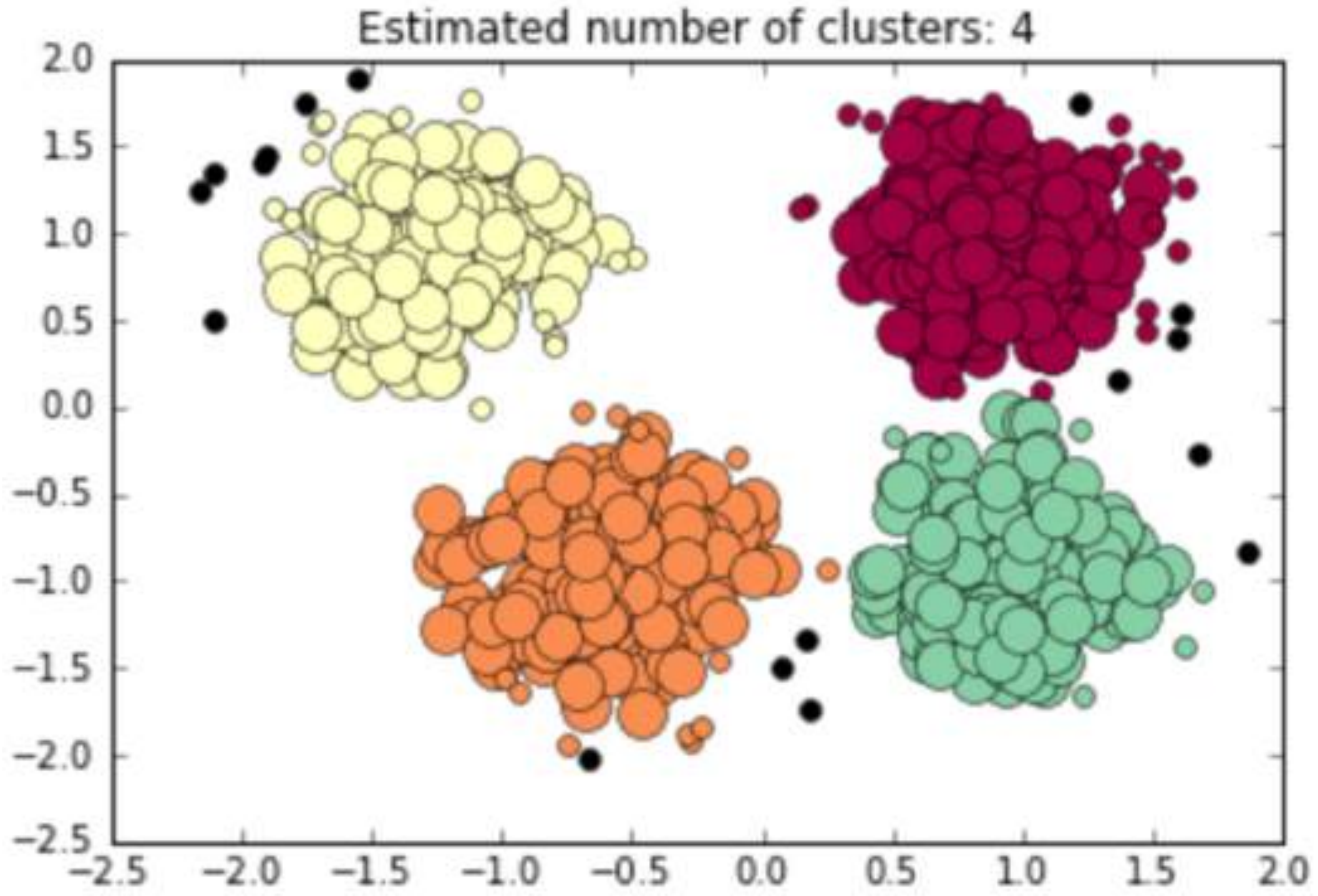
Пример: число кластеров в DBSCAN



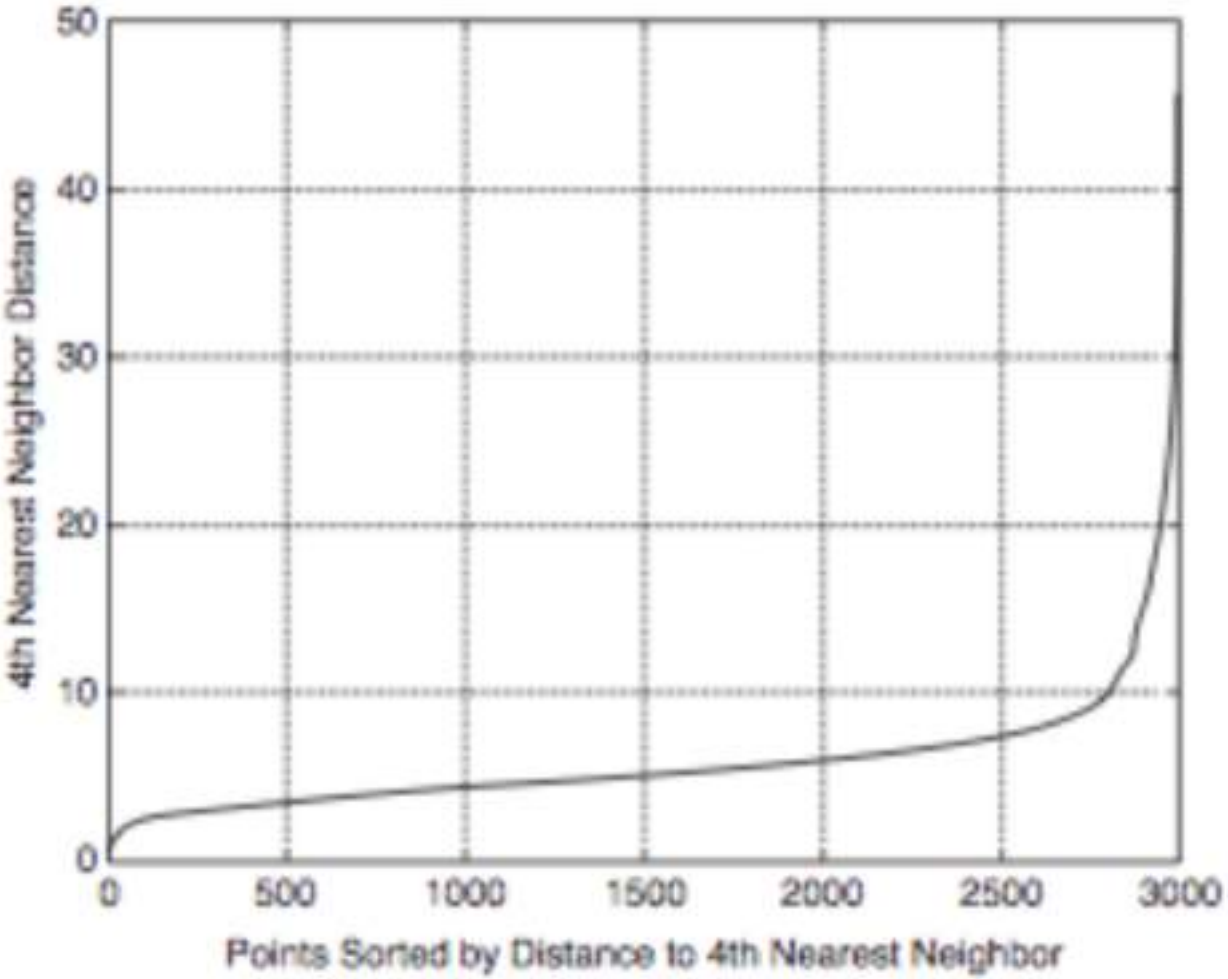
Пример: число кластеров в DBSCAN

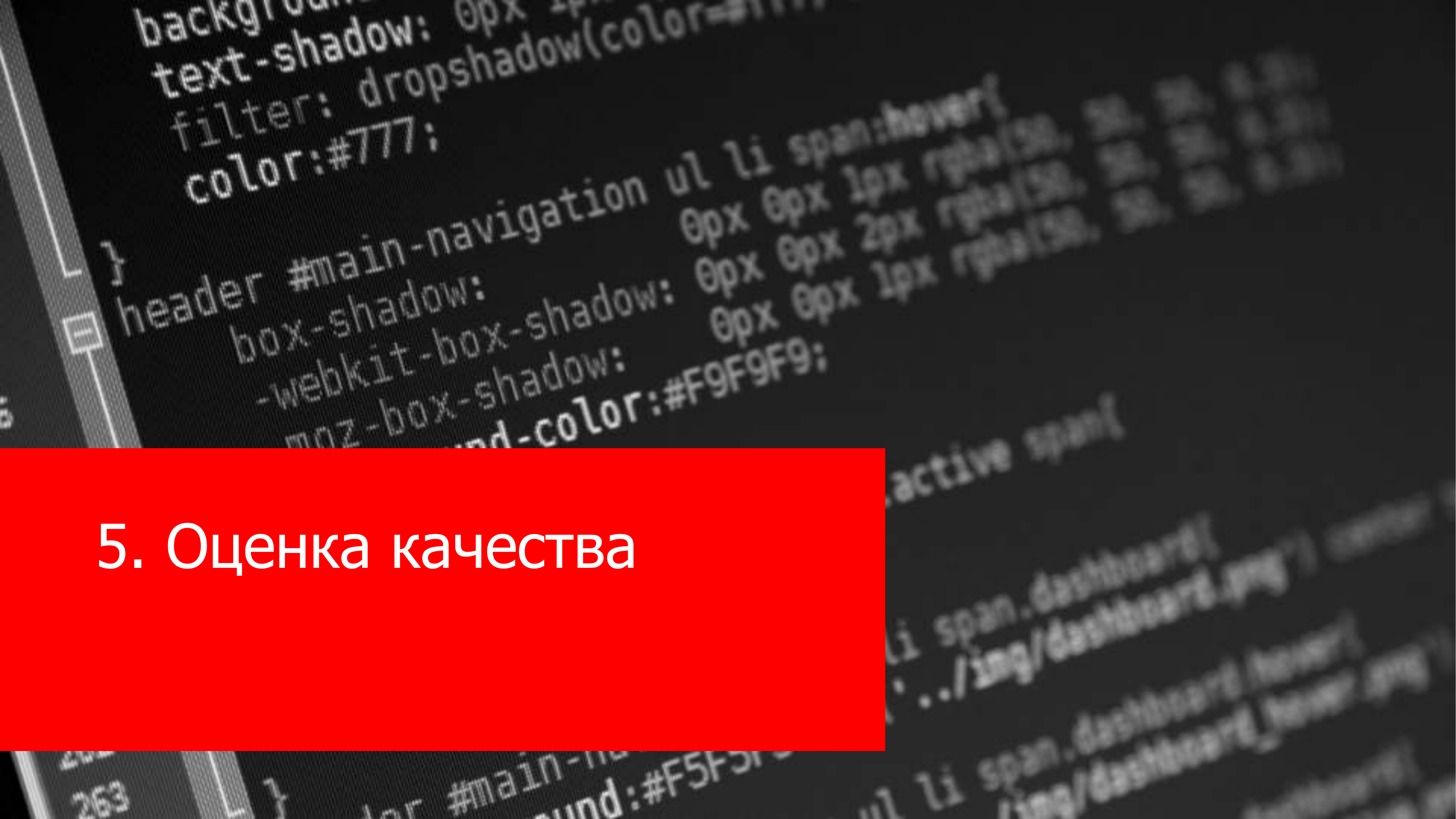


Пример: число кластеров в DBSCAN



DBSCAN: подбор параметров





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

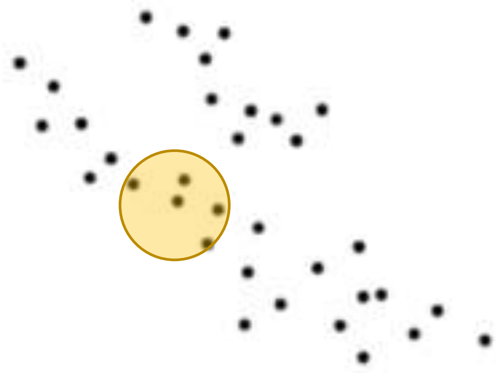
- **a**: Среднее расстояние от данного объекта до всех других объектов из того же кластера
- **b**: Среднее расстояние от данного объекта до всех объектов из *ближайшего другого кластера*

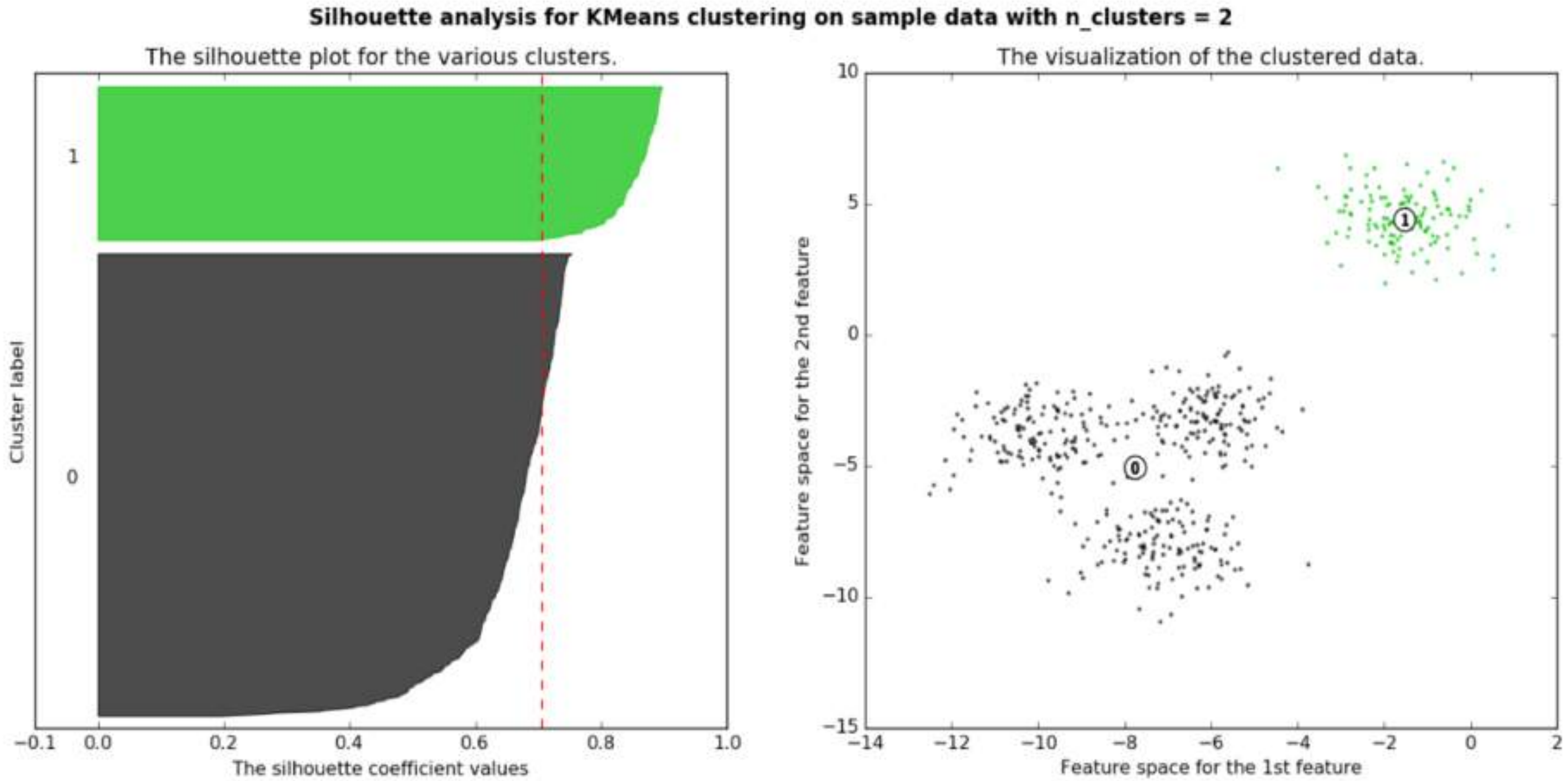
$$s = \frac{b - a}{max(a, b)}$$

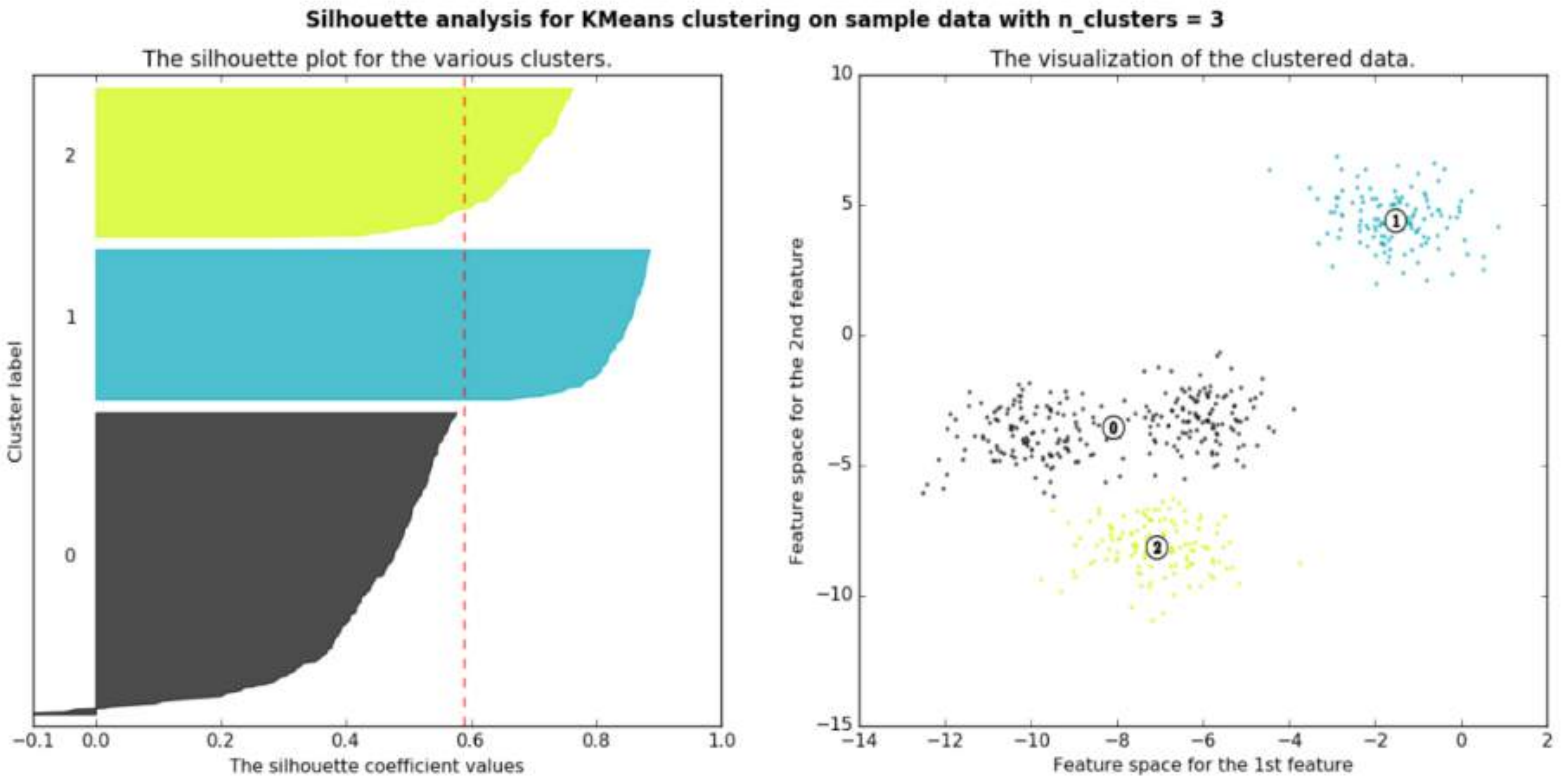
Коэффициент силуэта

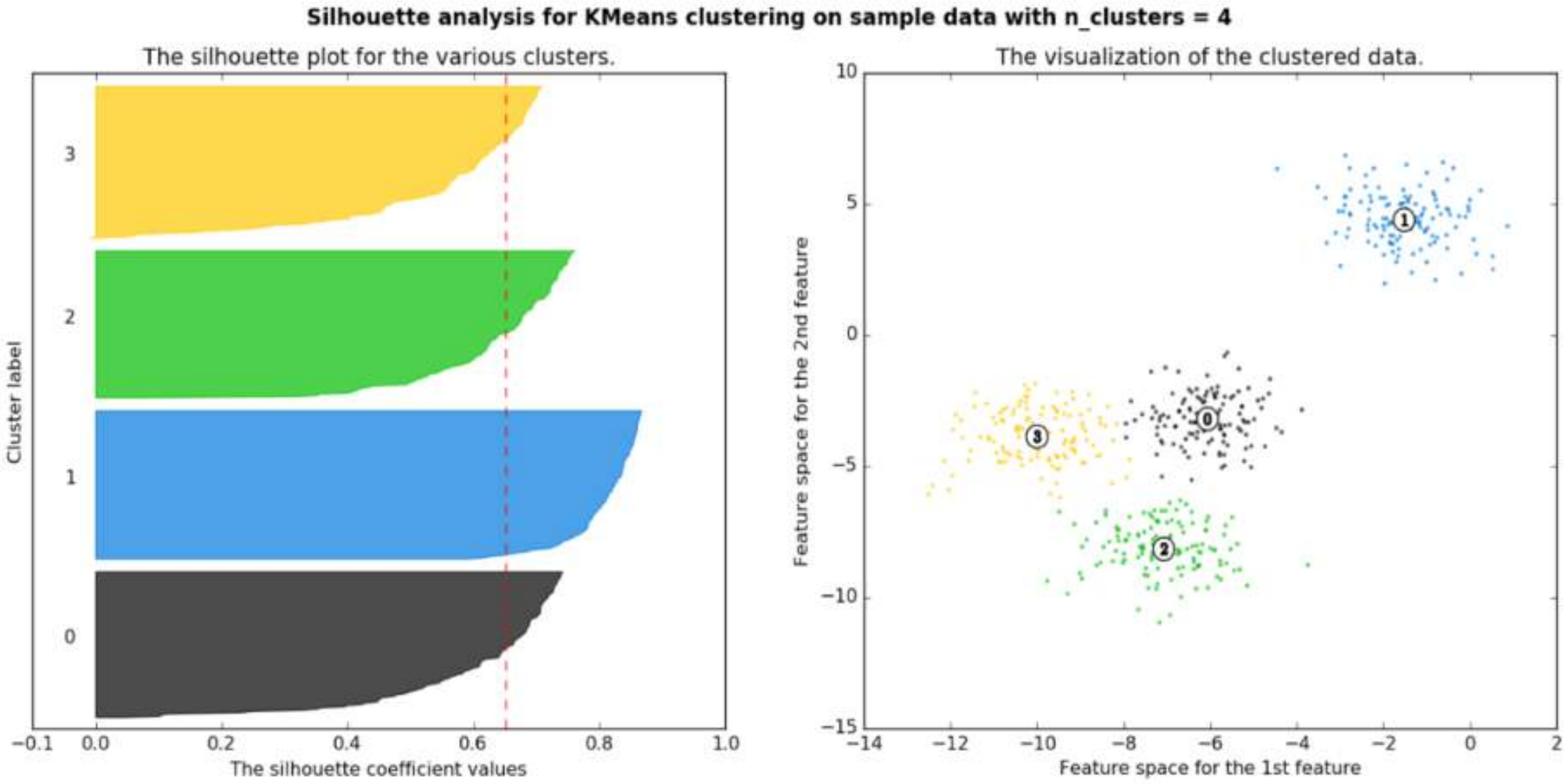
- **a**: Среднее расстояние от данного объекта до всех других объектов из того же кластера
- **b**: Среднее расстояние от данного объекта до всех объектов из *ближайшего другого кластера*

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

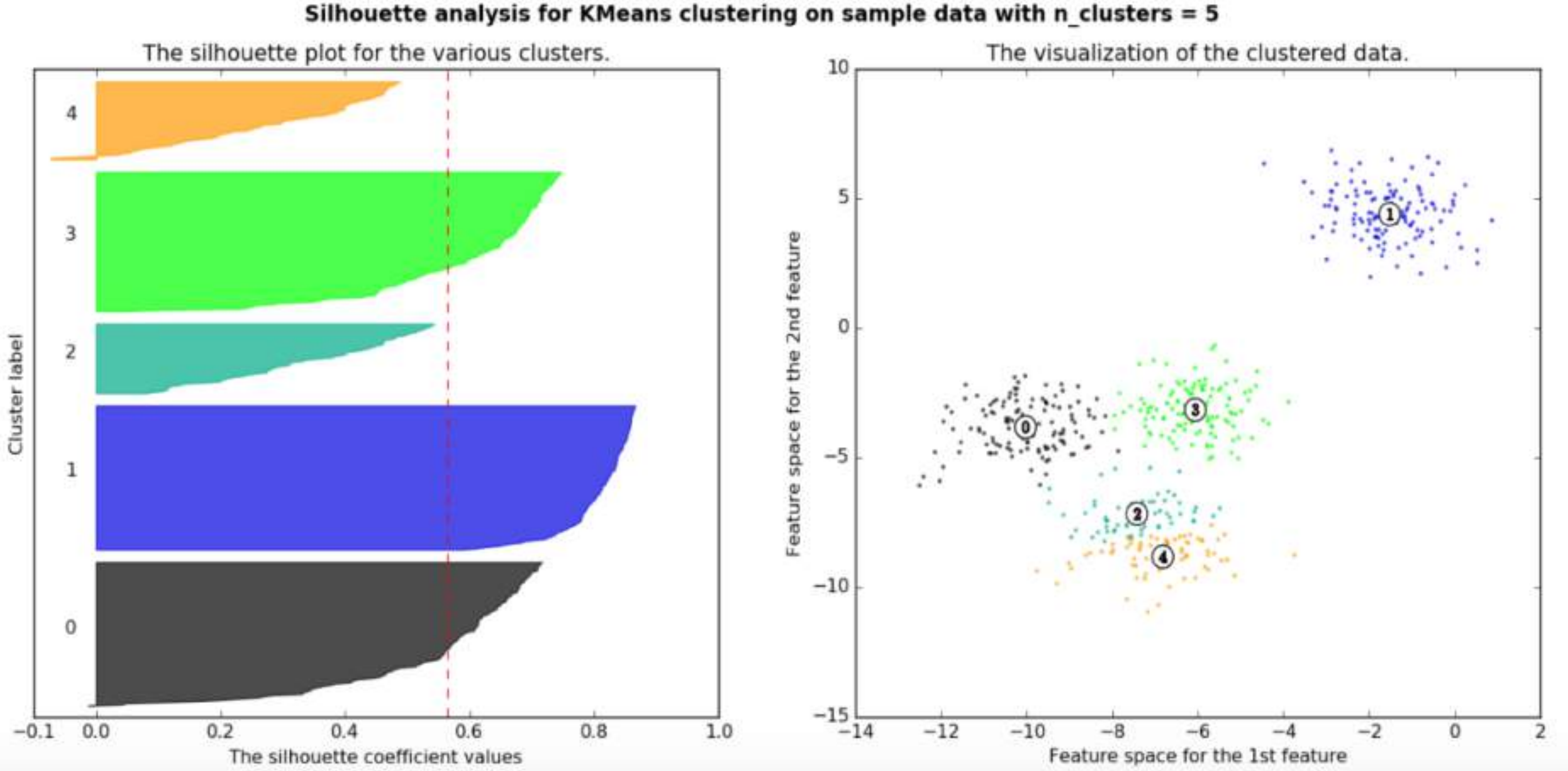



Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
The silhouette plot for the various clusters.
Cluster label
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature

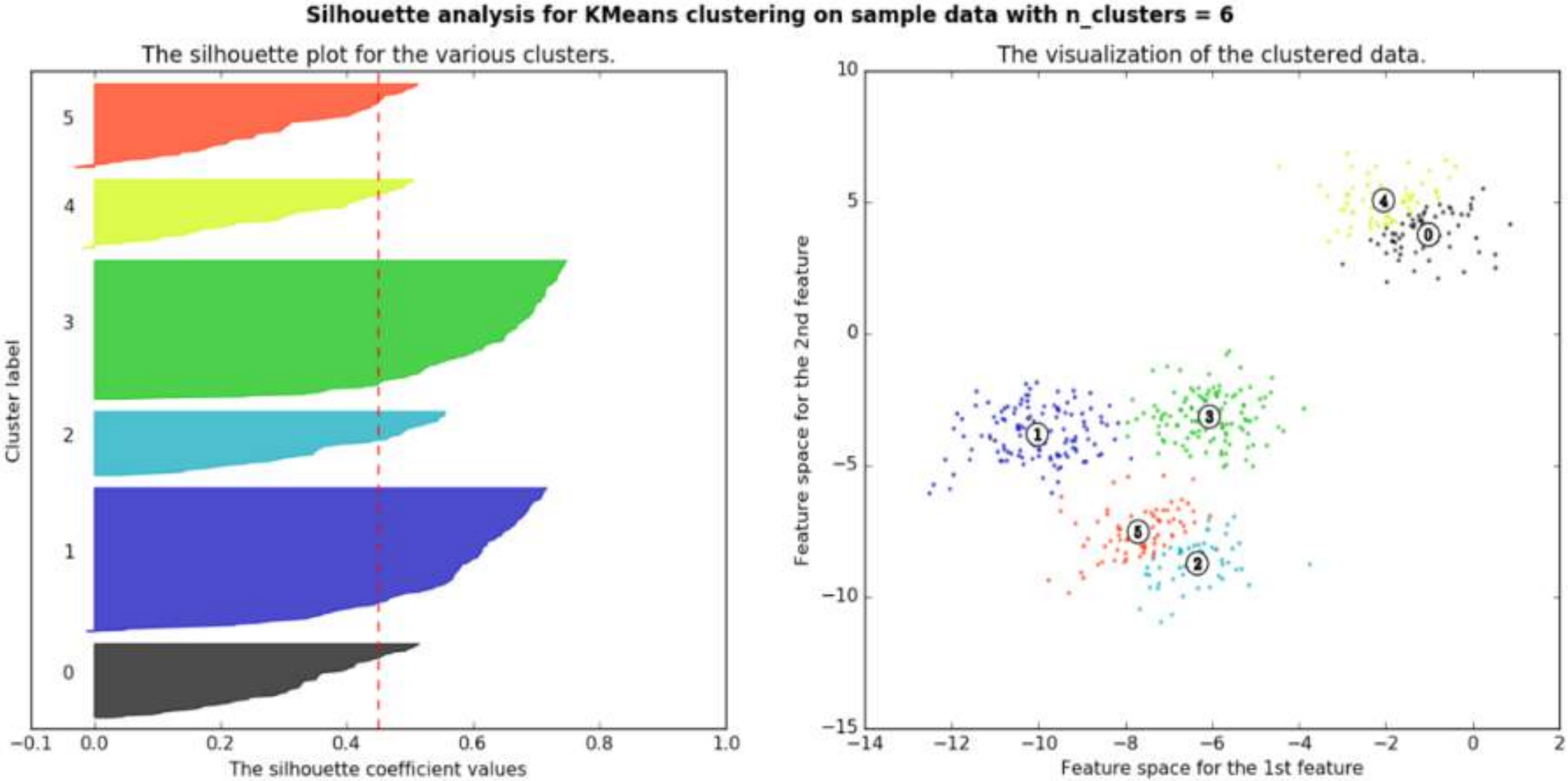

Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
Cluster label
2
1
0
-0.1 0.0 0.2 0.4 0.6 0.8 1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14 −12 −10 −8 −6 −4 −2 0 2
Feature space for the 1st feature



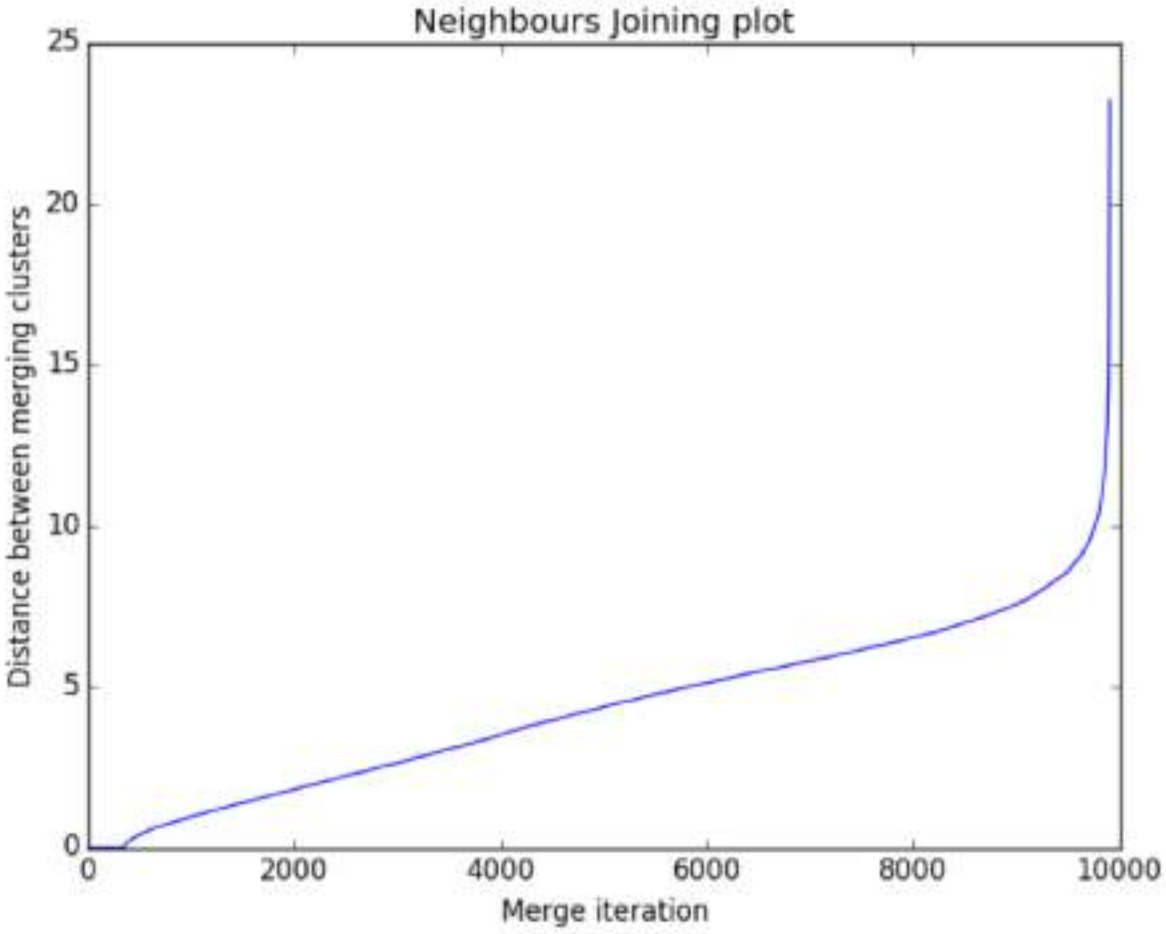
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature


Silhouette analysis for KMeans clustering on sample data with n_clusters = 6
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2

Проверка наличия кластерной структуры



Проверка наличия кластерной структуры

1. Генерируем p случайных точек из равномерного распределения и p случайных из обучающей выборки
2. Вычисляем величину (статистика Хопкинса):

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

Выбор признаков

Что хотим уметь делать:

Для разных признаков понимать, насколько хорошо решена задача кластеризации

Зачем:

Тогда сможем выбирать наиболее адекватные признаки

В чем проблема:

Текущие метрики зависят от признакового пространства

Однородность, полнота, V-мера

В каких случаях значения метрик максимальны:

- **Однородность**: кластер состоит только из объектов одного класса
- **Полнота**: все объекты из класса принадлежат к одному кластеру

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$H = -\sum_i p_i \ln p_i$$

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$H = -\sum_i p_i \ln p_i$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right)$$

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$H = -\sum_i p_i \ln p_i$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right)$$

$$P(c) = \frac{n_c}{n}$$

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$H = -\sum_i p_i \ln p_i$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right)$$

$$P(c) = \frac{n_c}{n}$$

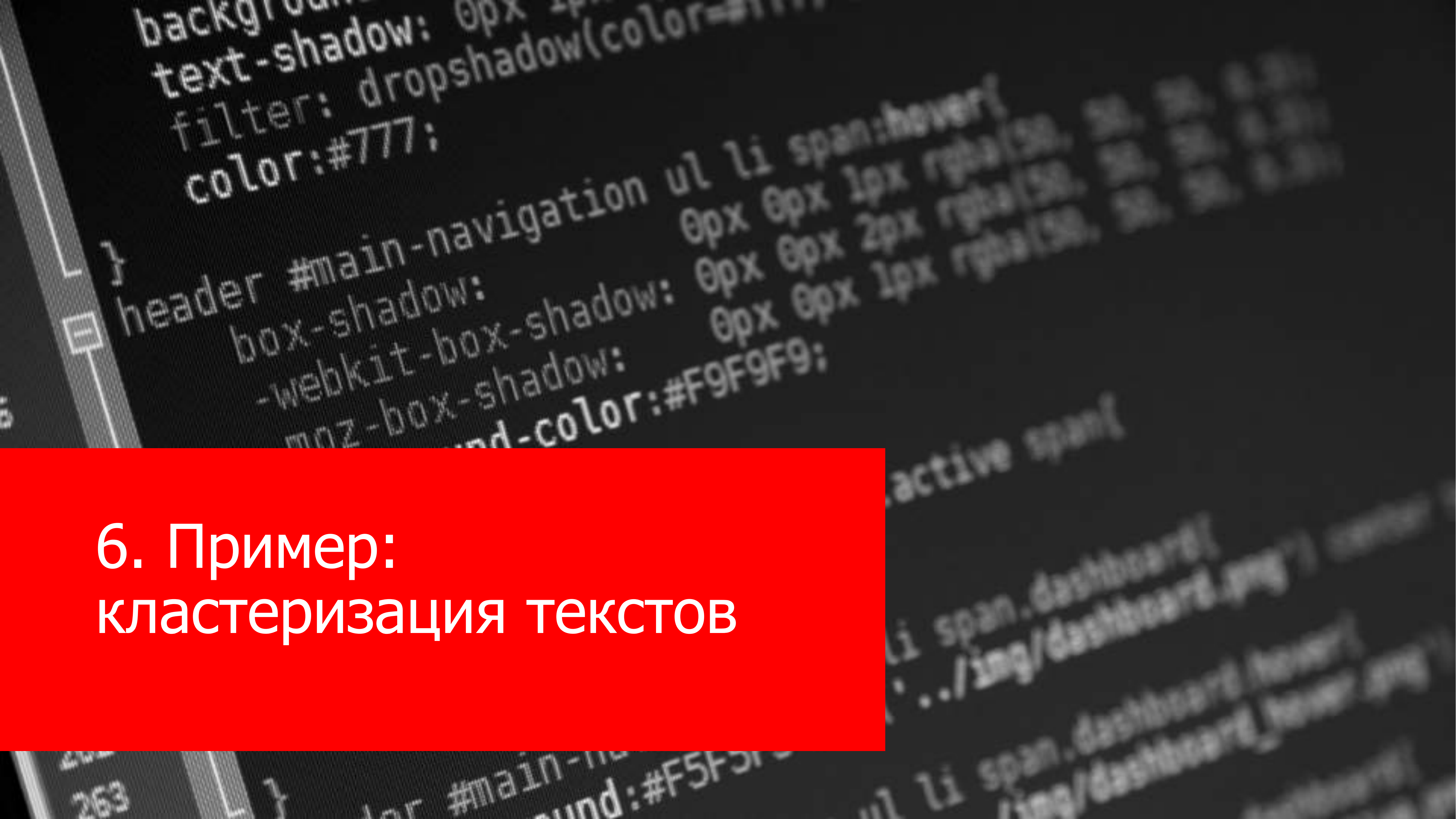
$$H(C|K) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{n_{c,k}}{n_k} \cdot \log\left(\frac{n_{c,k}}{n_k}\right)$$

$$P(c|k) = \frac{n_{c,k}}{n_k}$$

Привлечение асессоров для оценки качества

Если разметки нет, можно:

1. Использовать метрики без разметки
2. Создать разметку с помощью асессоров и использовать ее
3. Предложить асессорам отвечать на вопросы вида «допустимо ли эти объекты относить в один/в разные кластеры» или решать задания вида «найти лишний объект»



6. Пример: кластеризация текстов

Идеи кластеризации

- K-means на tf*idf
- Иерархическая кластеризация на tf*idf
- EM-алгоритм

Идеи кластеризации

- K-means на tf*idf – просто и более-менее работает
- Иерархическая кластеризация на tf*idf – часто возникает «гигантский кластер»
- EM-алгоритм – нет причин для использования именно гауссиан (хотя можно) + медленно

Последнее – повод применить EM-алгоритм как-то иначе

Вероятность встретить слово в документе

$$p(w|d) = \sum_{t=1}^{T} p(w|t,d)p(t|d)$$

Предположение условной независимости:

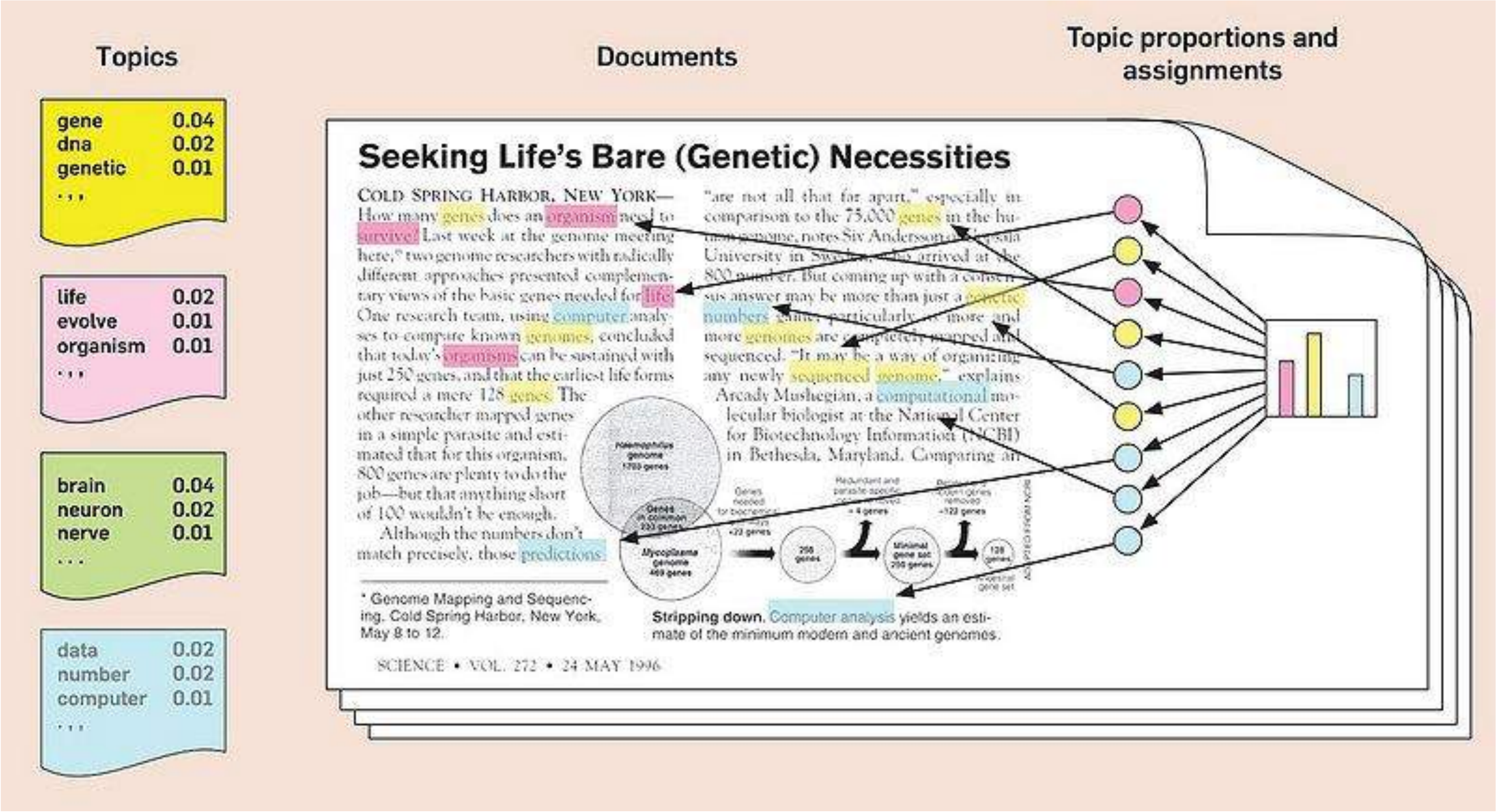
$$p(w|t,d) = p(w|t\ )$$

Вероятностная модель

$$p(w|d) = \sum_{t=1}^{T} p(w|t)p(t|d)$$

- $p(t|d)$ – результаты мягкой кластеризации документов по T кластерам
- $p(w|t)$ – распределение слов в кластере
- $p(w|d) \approx \frac{n_{dw}}{\sum_w n_{dw}}$

Мягкая кластеризация по кластерам-темам



Максимизация правдоподобия

$$L = \sum_{d,w} n_{dw} \ln \sum_{t=1}^{T} p(w|t)p(t|d) \to \max_{p(w|t),p(t|d)}$$

Максимизация правдоподобия

$$L = \sum_{d,w} n_{dw} \ln \sum_{t=1}^{T} \underbrace{p(w|t)p(t|d)}_{p_{tdw}} \to \max_{p(w|t),p(t|d)}$$

Максимизация правдоподобия

$$L = \sum_{d,w} n_{dw} \ln \sum_{t=1}^{T} \underbrace{p(w|t)p(t|d)}_{*} \to \max_{p(w|t),p(t|d)}$$

Упражнение: как выражается $p_{tdw} = p(t|d,w)$ через $*$?

Ответ:

$$p_{tdw} = p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t,d)p(t|d)}{p(w|d)} = \frac{*}{\sum_t *} = \operatorname*{norm}_t *$$

Максимизация правдоподобия

$$L = \sum_{d,w} n_{dw} \ln \sum_{t=1}^{T} \underbrace{p(w|t)p(t|d)}_{*} \to \max_{p(w|t),p(t|d)}$$

Упражнение: как выражается $p_{tdw} = p(t|d,w)$ через $*$?

Ответ:

$$p_{tdw} = p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t,d)p(t|d)}{p(w|d)} = \frac{*}{\sum_t *} = \operatorname*{norm}_t *$$

EM-алгоритм и PLSA (Probabilistic Latent Semantic Analysis)

Е-шаг:

$$p_{tdw} = \operatorname*{norm}_{t} p(w|t)p(t|d)$$

М-шаг:

$$p(w|t) = \operatorname*{norm}_{w} \sum_{d} n_{dw} \, p_{tdw}$$

$$p(t|d) = \operatorname*{norm}_{t} \sum_{w} n_{dw} \, p_{tdw}$$

Проблема PLSA

- PLSA делает матричное разложение матрицы $\hat{p}(w|d) = \frac{n_{dw}}{\sum_w n_{dw}}$
- Но разложение не единственно

Проблема PLSA

- PLSA делает матричное разложение матрицы $\hat{p}(w|d) = \frac{n_{dw}}{\sum_w n_{dw}}$
- Но разложение не единственно

В таких случаях в ML делают одно из двух:

1. Добавляют регуляризатор
2. Добавляют априорное распределение на параметры модели

Проблема PLSA

- PLSA делает матричное разложение матрицы $\hat{p}(w|d) = \frac{n_{dw}}{\sum_w n_{dw}}$
- Но разложение не единственно

В таких случаях в ML делают одно из двух:

1. Добавляют регуляризатор
2. Добавляют априорное распределение на параметры модели

Что во многом примерно одно и то же

Latent Dirichlet Allocation

- Добавляем априорные распределения параметров:

$$\big(p(w|t)\big)_{w=1,\dots,W} \sim Dir(\beta),$$
$$\big(p(t|d)\big)_{t=1,\dots,T} \sim Dir(\alpha)$$

- Почему распределение Дирихле? Потому что сопряженное к мультиномиальному

- Применяем Байесовский вывод, получаем готовые итерации для алгоритма

LDA как регуляризованный PLSA

$$L = \sum_{d,w} n_{dw} \ln \sum_{t=1}^{T} p(w|t)p(t|d) + \sum_{t,w} \ln p(w|t)^{\beta_w - 1} + \sum_{d,t} \ln p(t|d)^{\alpha_w - 1}$$

EM-алгоритм в LDA

Е-шаг:

$$p_{tdw} = \operatorname*{norm}_{t} p(w|t)p(t|d)$$

М-шаг:

$$p(w|t) = \operatorname*{norm}_{w} \left(\sum_{d} n_{dw} \, p_{tdw} + \beta_w - 1 \right)$$

$$p(t|d) = \operatorname*{norm}_{t} \left(\sum_{w} n_{dw} \, p_{tdw} + \alpha_t - 1 \right)$$

Пример результатов LDA

0.016*"\'s" + 0.009*"medical" + 0.008*"number" + 0.007*"gm" + 0.007*"disease" + 0.007*"year" + 0.007*"health" + 0.006*"study" + 0.006*"patients" + 0.006*"aids"

0.020*"\'s" + 0.010*"game" + 0.009*"one" + 0.008*"n\'t" + 0.007*"year" + 0.007*"team" + 0.007*"last" + 0.006*"first" + 0.006*"games" + 0.006*"car"

0.009*"\'s" + 0.009*"one" + 0.009*"people" + 0.007*"would" + 0.006*"said" + 0.005*"children" + 0.005*"could" + 0.005*"gun" + 0.004*"us" + 0.004*"n\'t"'),

0.010*"public" + 0.010*"information" + 0.009*"government" + 0.009*"new" + 0.007*"use" + 0.007*"1993" + 0.007*"national" + 0.006*"encryption" + 0.006*"security" + 0.005*"law"'),

0.114*"..." + 0.006*".." + 0.006*"new" + 0.006*"_/" + 0.005*"canada" + 0.005*"newsletter" + 0.005*"insurance" + 0.004*"10" + 0.004*"apr" + 0.004*"1993"

Пример результатов LDA

0.016*"\'s" + 0.012*"q" + 0.011*"mr." + 0.010*"people" + 0.010*"israel" + 0.009*"president" + 0.008*"would" + 0.007*"jews" + 0.006*"think" + 0.006*"israeli"

0.054*"god" + 0.016*"jesus" + 0.013*"bible" + 0.013*"church" + 0.012*"christian" + 0.011*"christ" + 0.009*"christians" + 0.009*"lord" + 0.008*"\'s" + 0.007*"faith"

0.015*"image" + 0.011*"file" + 0.010*"available" + 0.010*"software" + 0.009*"files" + 0.008*"version" + 0.007*"ftp" + 0.007*"data" + 0.007*"\'s" + 0.007*"also"

0.016*"armenian" + 0.014*"armenians" + 0.013*"turkish" + 0.010*"turkey" + 0.008*"greek" + 0.007*"history" + 0.007*"turks" + 0.007*"people" + 0.006*"russian" + 0.006*"greece"

0.013*"space" + 0.007*"power" + 0.006*"earth" + 0.006*"also" + 0.006*"\'s" + 0.005*"new" + 0.005*"used" + 0.005*"high" + 0.005*"sale" + 0.005*"system"

Проблема интерпретируемости LDA

- Как правило, LDA не дает интерпретируемость тем человеком
- Проблема в том, что решение задачи матричного разложения не обязывает темы быть «осмысленными»

Вариант 1: делать более «умную» регуляризацию

Вариант 2: искать модели и подходы, у которых «из коробки» более интерпретируемый результат

Вариант 3: вернуться к простым методам

Более «умная» регуляризация

Максимизация логарифма правдоподобия с регуляризатором:

$$\sum_{d,w} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \quad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{ll} \text{Е-шаг:} & \left\{ \begin{array}{l} p_{tdw} \equiv p(t|d,w) = \operatorname*{norm}_{t\in T}\left(\phi_{wt}\theta_{td}\right) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \quad n_{wt} = \sum_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname*{norm}_{t\in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum_{w\in d} n_{dw} p_{tdw} \end{array} \right. \\ \text{М-шаг:} & \end{array}$$

где $\operatorname*{norm}_{t\in T}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$ — операция нормировки вектора.

Additive Regularization of Topic Models

К.В. Воронцов

См. подробнее по ссылке

Более «умная» регуляризация

1. декоррелирование тем:

$$R(\Phi) = -\tau \sum_{s,t \in T} \sum_{w \in W} \phi_{wt} \phi_{ws}$$

2. разреживание распределений $p(t|d)$:

$$R(\Theta) = -\alpha \sum_{d,t} \ln \theta_{td}$$

3. сглаживание распределений $p(w|t)$:

$$R(\Phi) = \beta \sum_{t,w} \ln \phi_{wt}$$

Additive Regularization of Topic Models

К.В. Воронцов

[См. подробнее по ссылке]

Модели с более интерпретируемыми результатами

- Word2vec и его модификации лучше моделируют семантическую близость слов и документов
- Кластеризация в пространстве эмбеддингов как правило более интерпретируема

Напоминание: оптимизационная задача в word2vec с negative sampling

$$P(D=1|w,c) = \sigma(\vec{w}\cdot\vec{c}) = \frac{1}{1+e^{-\vec{w}\cdot\vec{c}}}$$

$$\log\sigma(\vec{w}\cdot\vec{c}) + k\cdot\mathbb{E}_{c_N\sim P_D}\left[\log\sigma(-\vec{w}\cdot\vec{c}_N)\right]$$

$$P_D(c) = \frac{\#(c)}{|D|}$$

$$\ell = \sum_{w\in V_W}\sum_{c\in V_C}\#(w,c)\left(\log\sigma(\vec{w}\cdot\vec{c}) + k\cdot\mathbb{E}_{c_N\sim P_D}\left[\log\sigma(-\vec{w}\cdot\vec{c}_N)\right]\right)$$

Word2Vec и матричные разложения

$PMI(w_i, c_j)$ - совместная встречаемость w_i и c_j

Word2Vec и матричные разложения

$PMI(w_i, c_j)$ - совместная встречаемость w_i и c_j

Измеряется так:

$$PMI(w_i, c_j) = \ln \frac{p(w_i)p(c_j)}{p(w_i, c_j)}$$

Word2Vec и матричные разложения

$PMI(w_i, c_j)$ - совместная встречаемость w_i и c_j

Измеряется так:

$$PMI(w_i, c_j) = \ln \frac{p(w_i)p(c_j)}{p(w_i, c_j)}$$

Оказывается (Levi, NIPS 2014), Word2Vec выполняет матричное разложение матрицы, заполненной числами $PMI(w_i, c_j) - \ln k$ (k – количество примеров в Negative Sampling)

Общая идея эмбеддингов

1. Есть объекты, для которых вам нужно обучить векторные представления v_i
2. Из каких соображений обучать представления – формулируется *оптимизационной задачей,* составленной из неких разумных соображений
3. Оптимизационная задача решается некоторым методом численной оптимизации (например, SGD)

Возможные причины интерпретируемости

Очень вероятно, что интерпретируемость обоснована именно удачной постановкой оптимизационной задачи в word2vec

Если рассмотреть как документ множество всех контекстов слова и в оптимизационной задаче вместо оценки

$P(D = 1|w, c) = \sigma(< w, c >)$ оценивать (как в topic models)

$$P(D = 1|w, c) = \sum_t p(w|t)p(t|c)$$

и применить все то же тематическое моделирование – регулярности word2vec остаются

Более простые подходы

Обсудим на семинаре

Обсудили на лекции

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