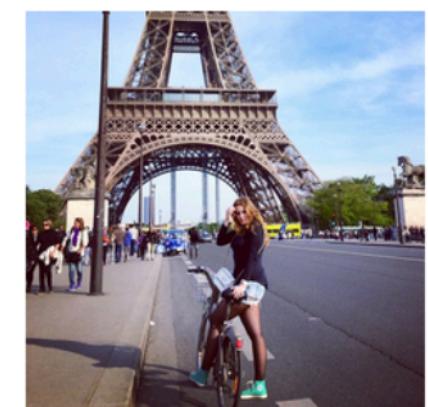
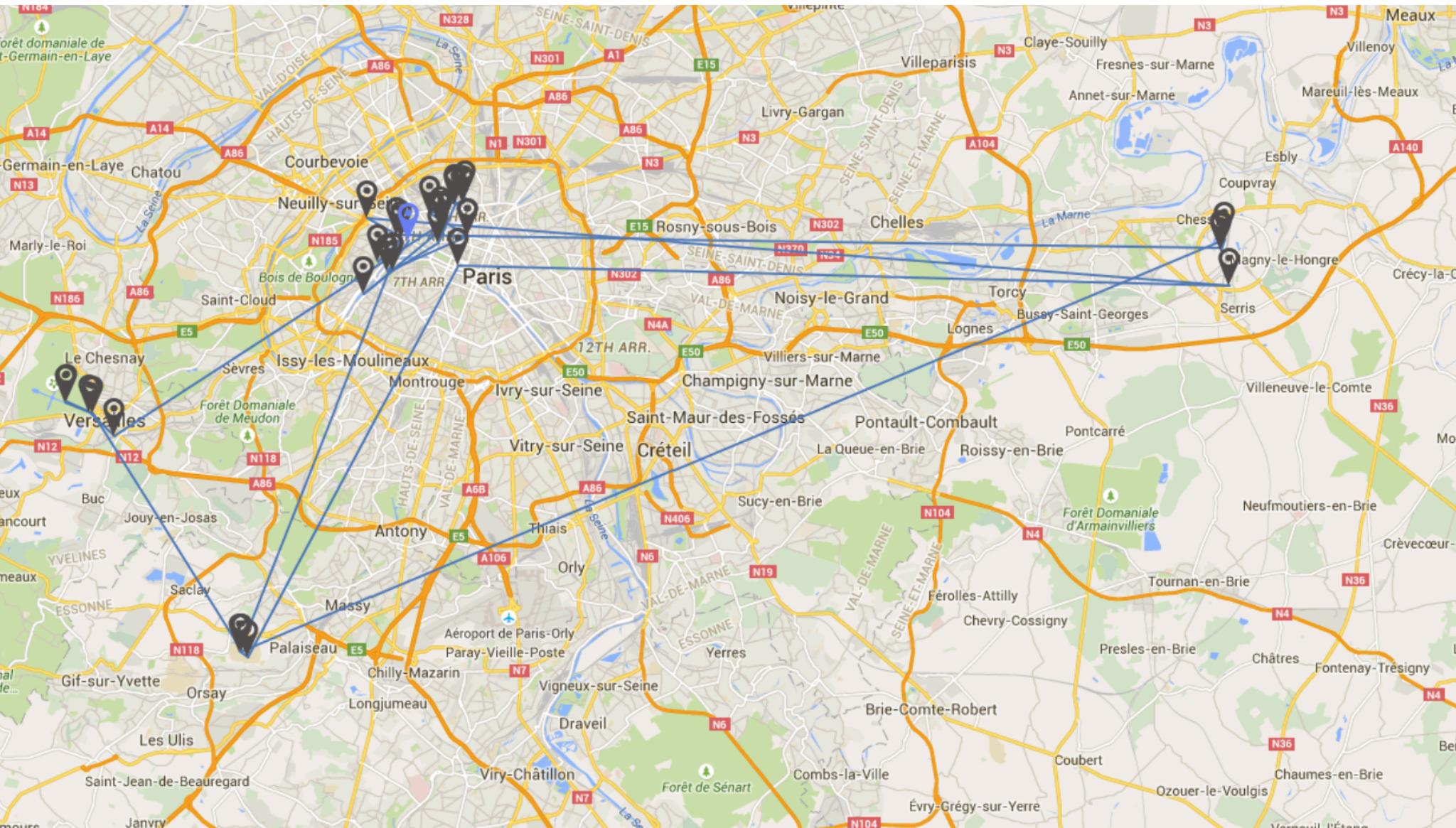


Data Mining in Action

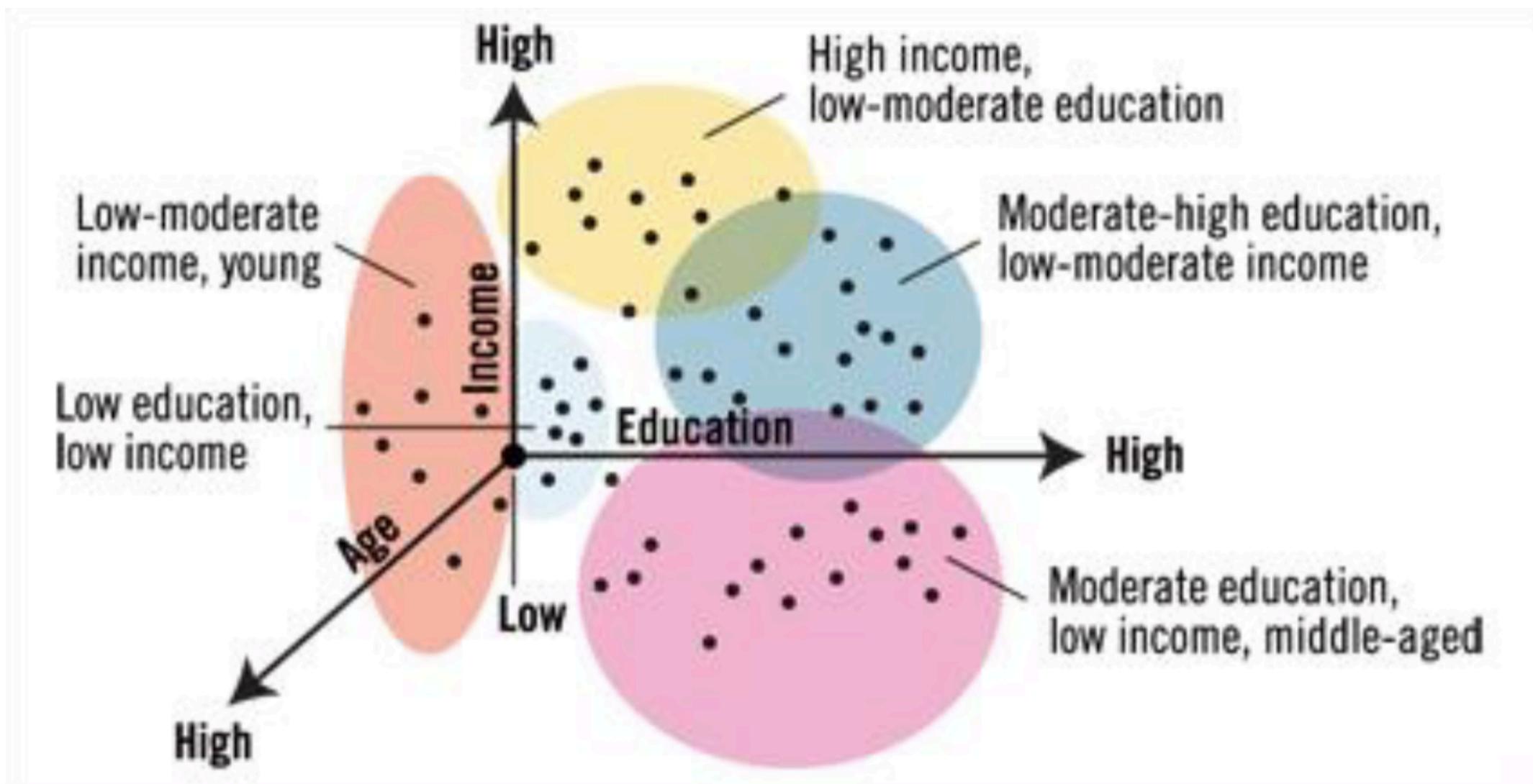
Лекция 3 Кластеризация

Виктор Кантор

Приимер: анализ геоданных



Пример: сегментация рынка



Пример: кластеризация текстов по теме



Керлингистки сборной РФ сделали
правильные выводы после ОИ -
Сидорова

10:38 26.03.2014



Путин призвал МВД использовать в
Крыму опыт работы на Олимпиаде

14:13 21.03.2014



Два "олимпийских" спецавтопарка
останутся в Сочи как наследие Игр

11:50 26.03.2014

Скриншот с сайта РИА Новости (ria.ru)

Кластеризация

1. Задача кластеризации
2. Разнообразие задач
3. K-Means
4. EM-алгоритм
5. Иерархическая кластеризация
6. Простые графовые методы
7. Density-based кластеризация
8. Выбор метода
9. Оценка качества

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 $a(x)$:

$$a(x) \approx y$$

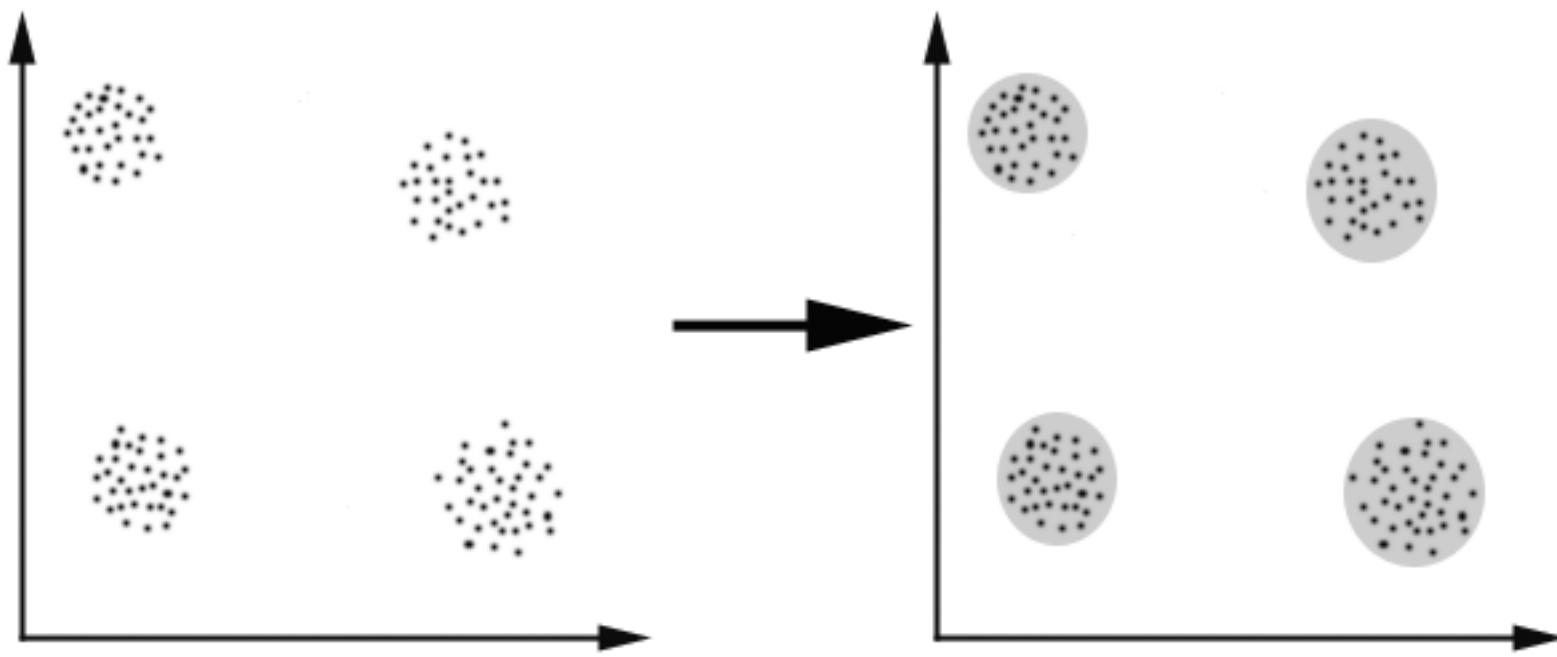
Кластеризация

«Обучающая» выборка:
 x_1, \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 внутриклusterное расстояние

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

Среднее межклusterное расстояние

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

Придумываем метрику качества

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

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

$$F_0/F_1 \rightarrow \min$$

2. 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

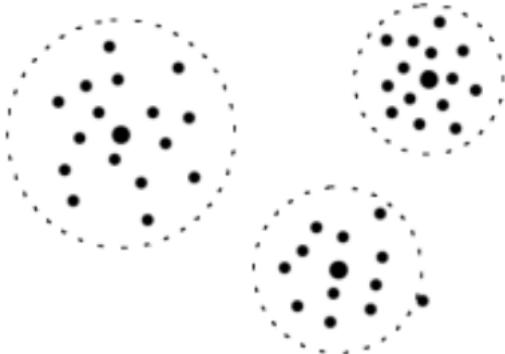
Форма кластеров



Форма кластеров



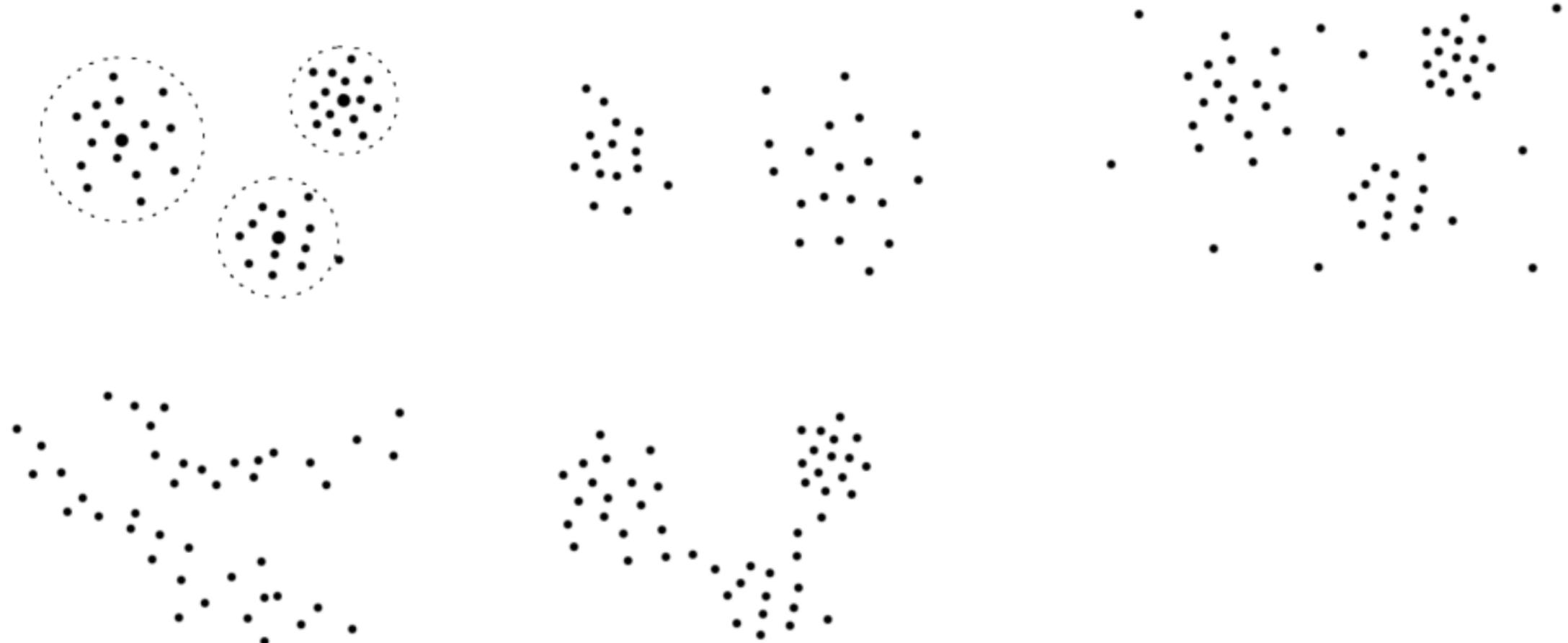
Форма кластеров



Форма кластеров



Форма кластеров



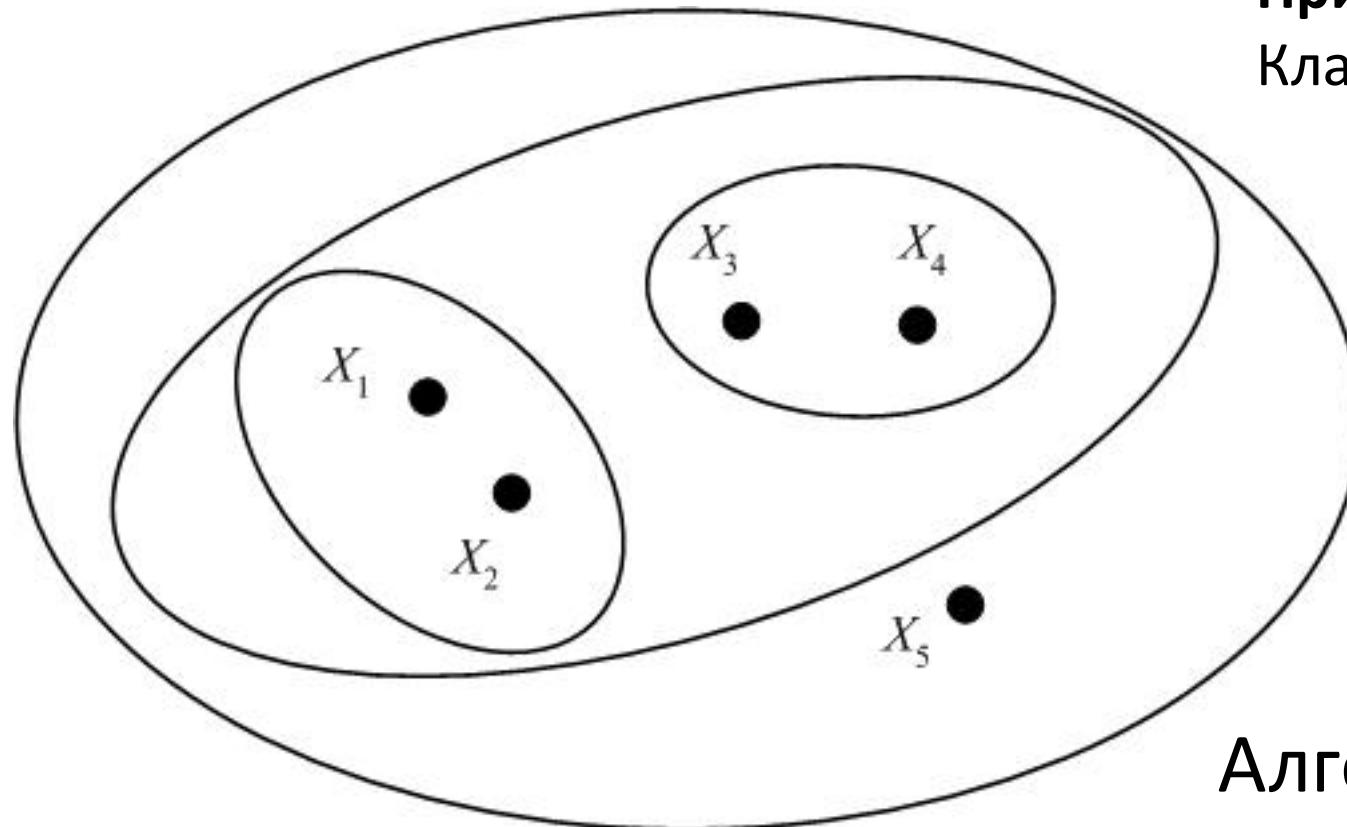
Форма кластеров



Форма кластеров

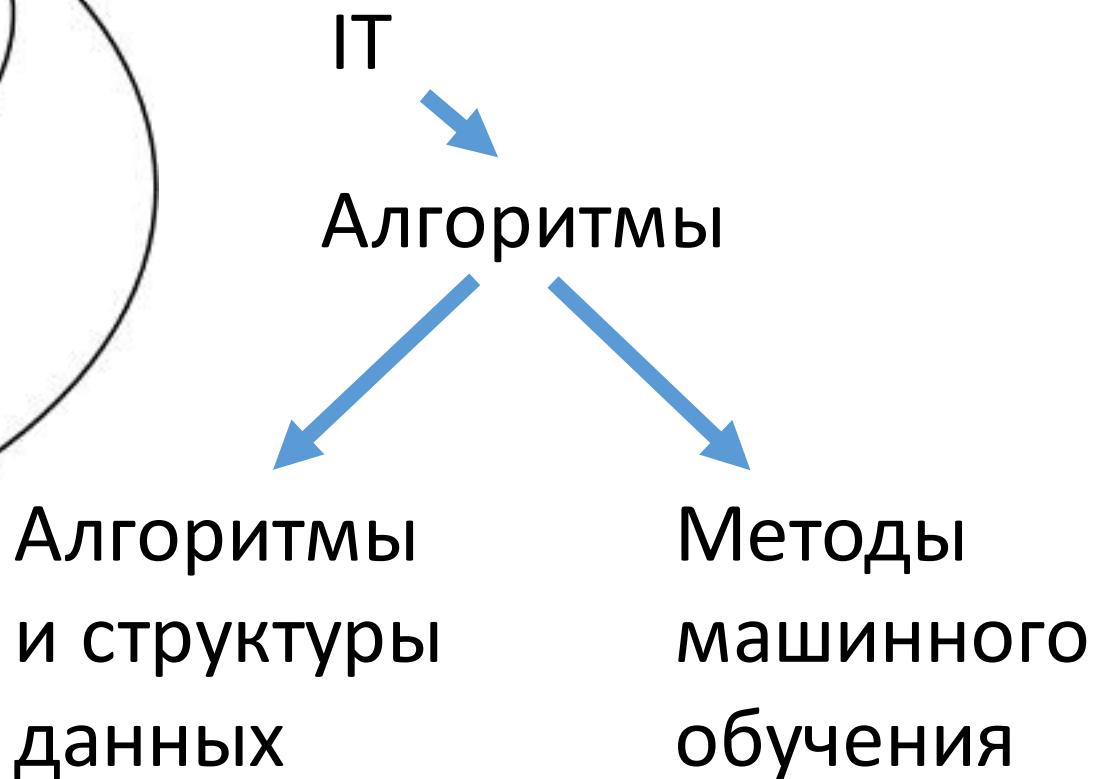


Вложенность кластеров



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



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



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



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

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

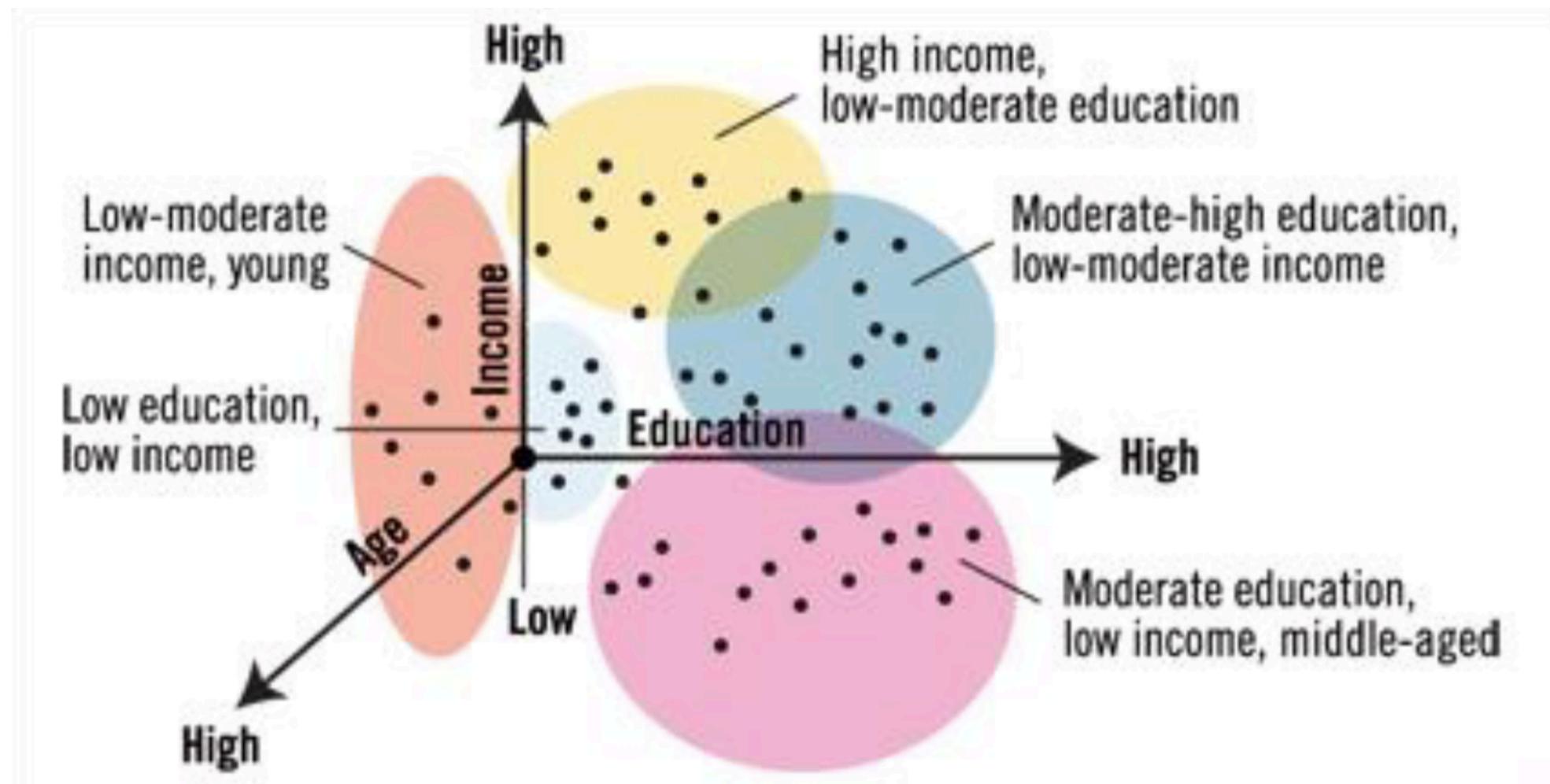
Открытие Олимпиады в Сочи
посмотрели несколько миллиардов
человек

Олимпиада в Сочи открыта

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

Основная задача или вспомогательная

Сегментация целевой аудитории



Основная задача или вспомогательная

Кластеризация символов по написанию для улучшения
распознавания

5

5

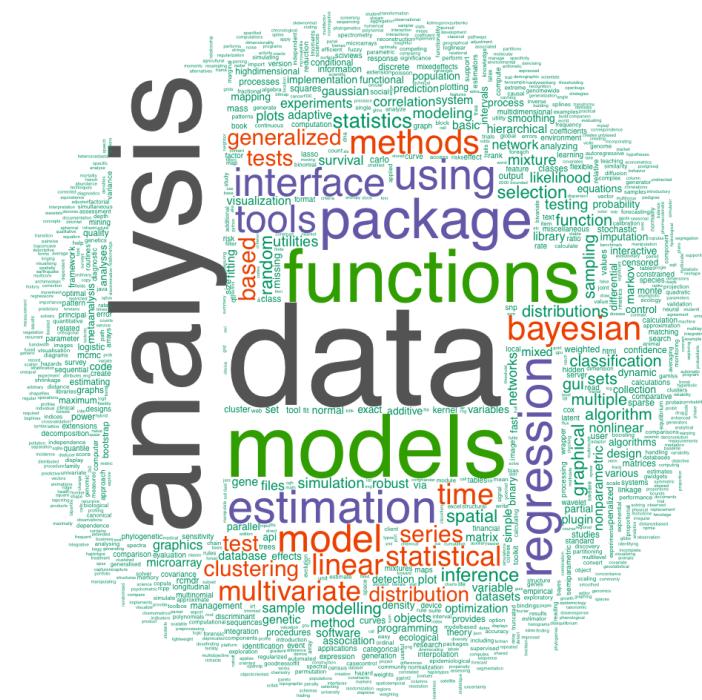
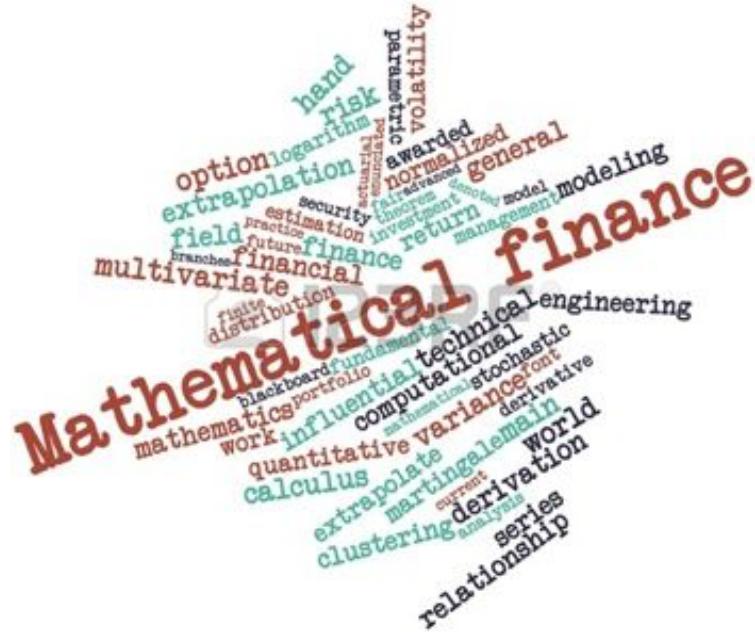
5

5

5

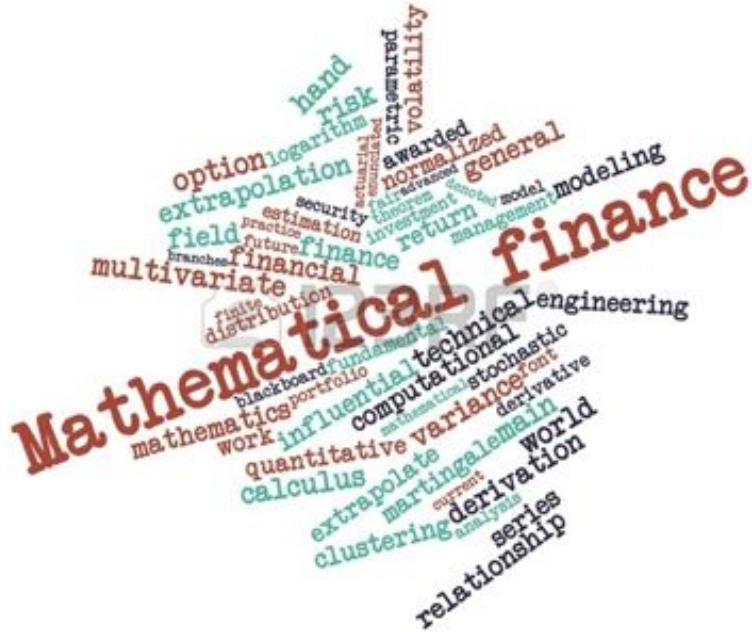
«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»

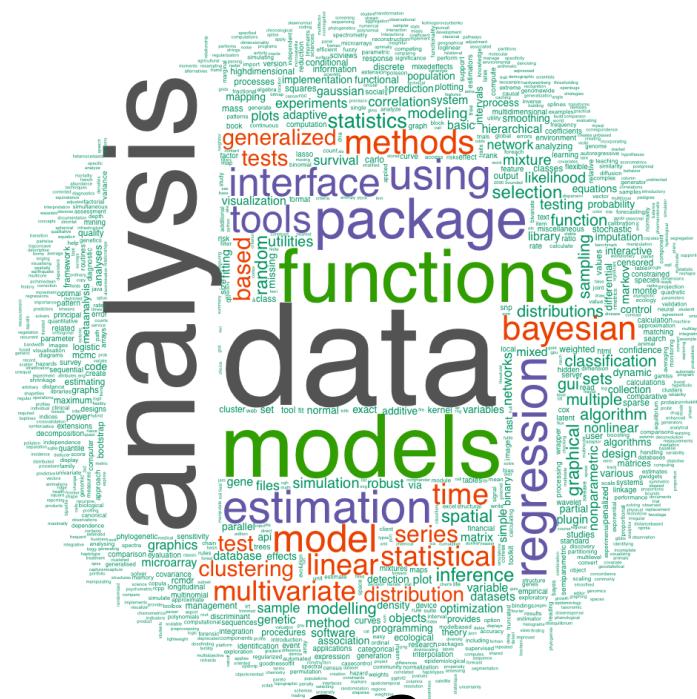


«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»



0.2



0.3

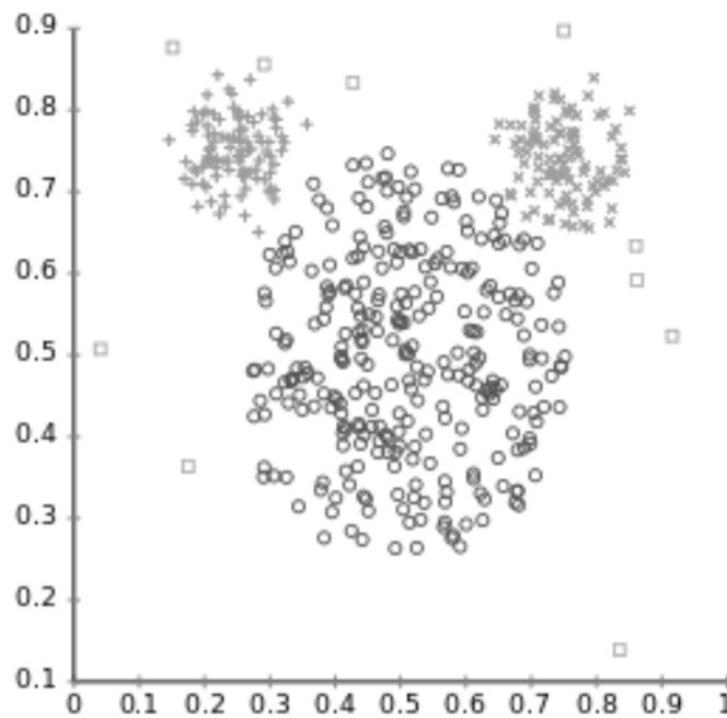


0.5

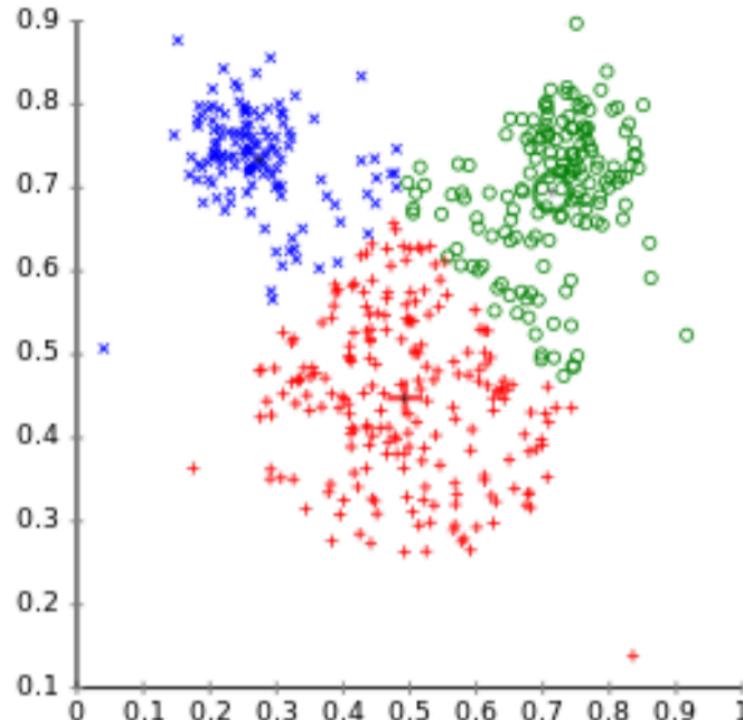
Резюме: чем могут отличаться задачи кластеризации

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

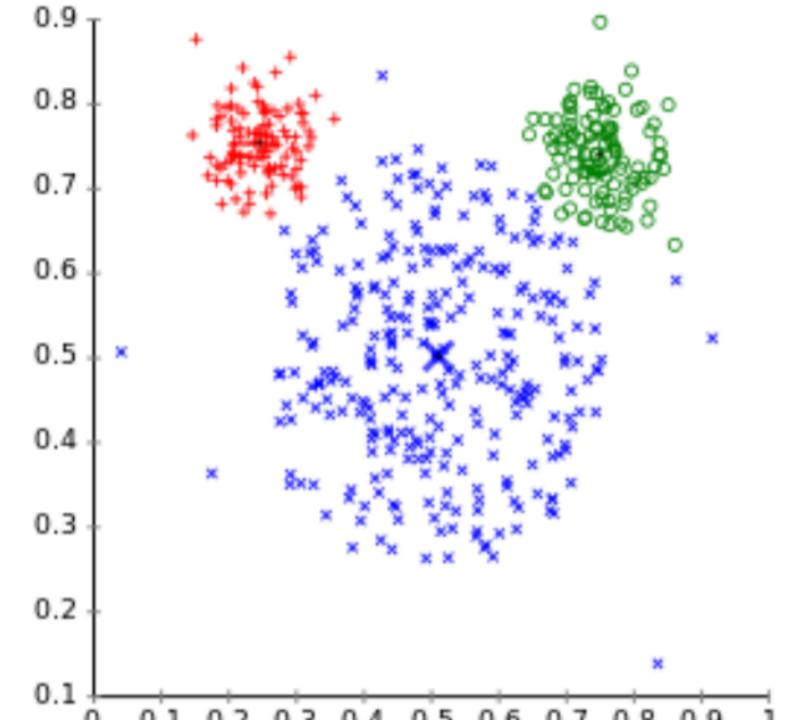
Различия в результатах работы методов



Исходная выборка
("Mouse" dataset)



Метод k средних
(K-Means)



ЕМ-алгоритм

3. Метод К средних (K-Means)

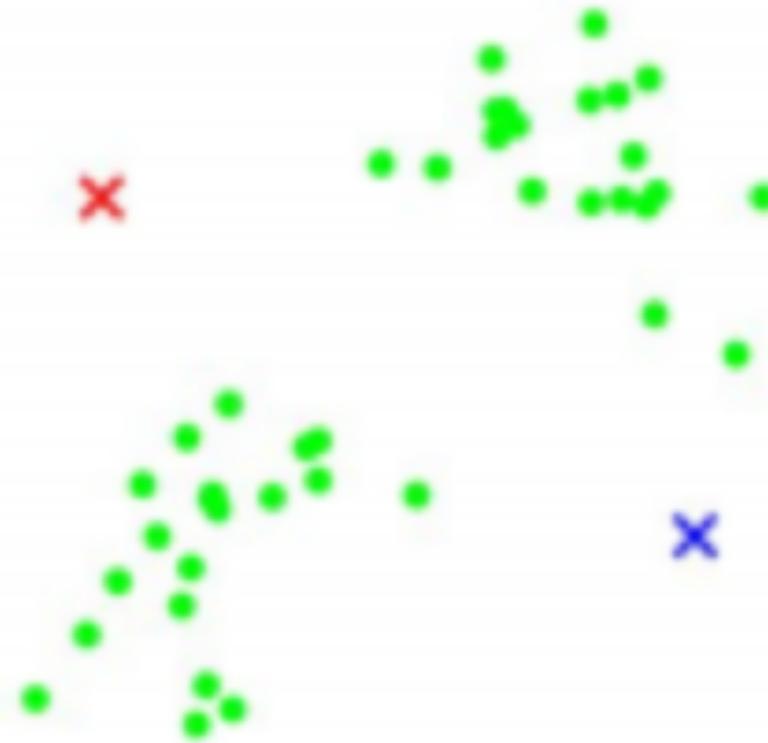
План

1. Как работает K-Means
2. Что делать, когда данных много: Mini Batch K-Means
3. Что делать, когда много признаков
4. Выбор начальных приближений: Kmeans++
5. Пример: уменьшение количества цветов в изображении
6. Работа K means с разными формами кластеров
7. Пример: мешок визуальных слов (bag of visual words)
8. Что оптимизирует K means

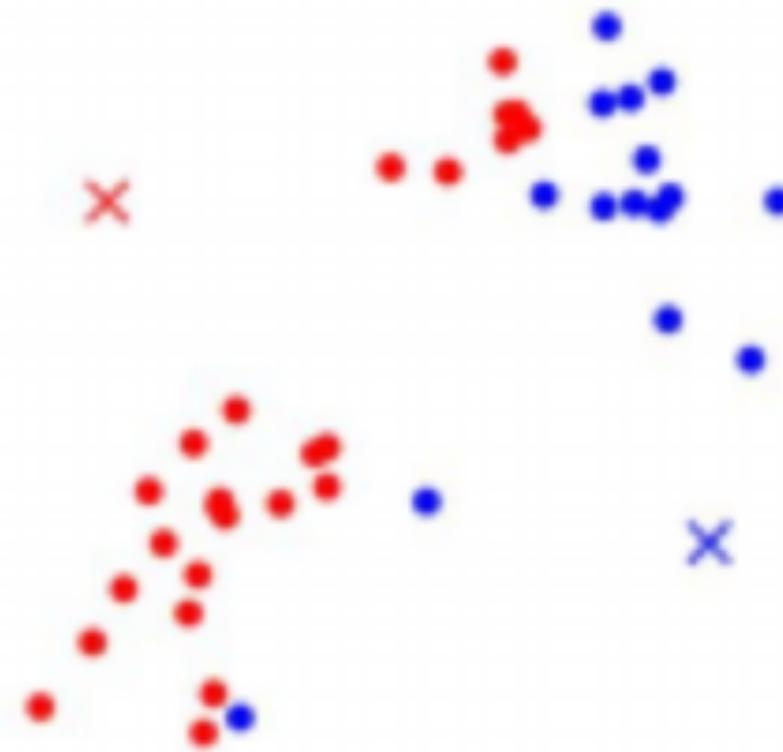
Как работает K Means



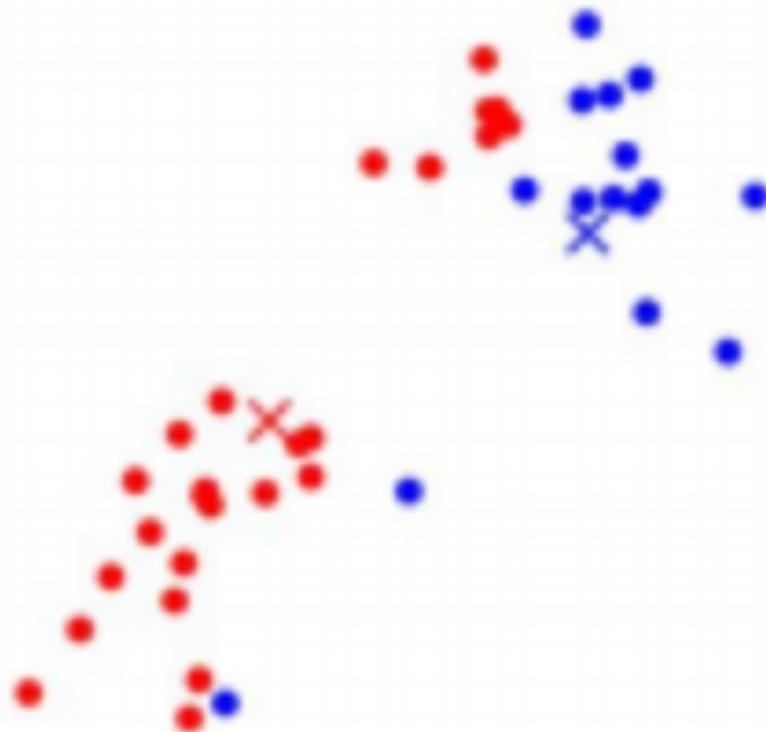
Как работает K Means



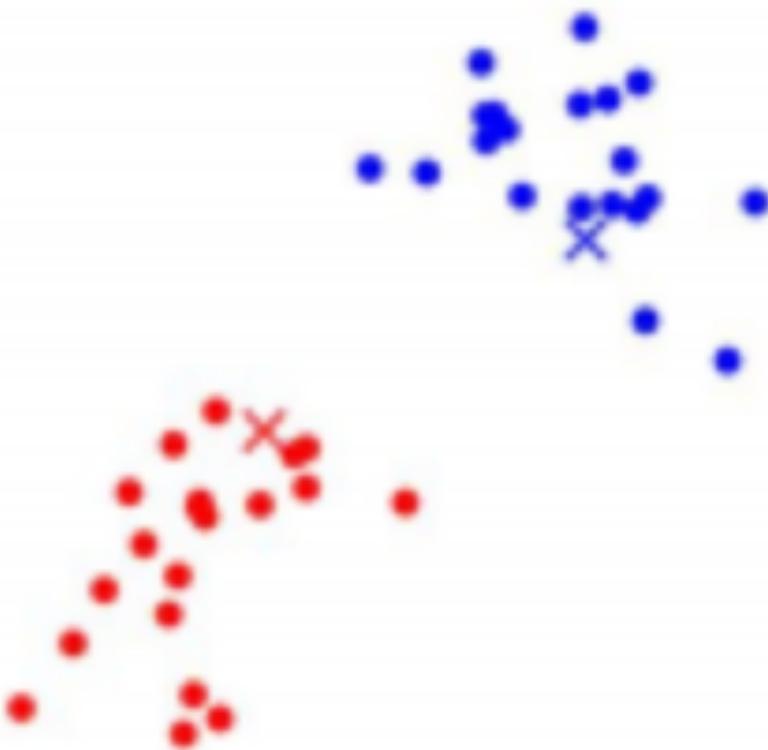
Как работает K Means



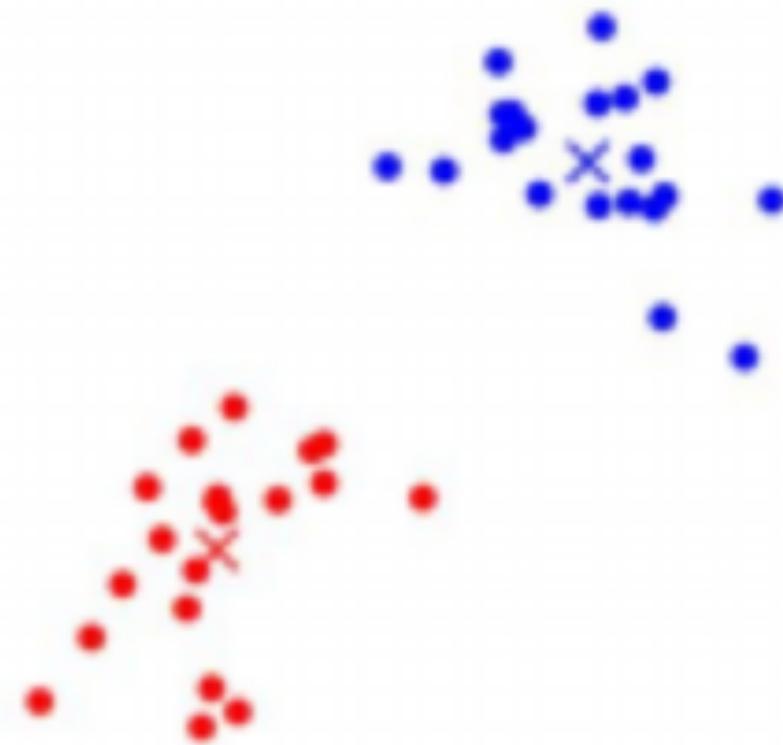
Как работает K Means



Как работает 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
- 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 случайно 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: квантизация изображений

Original image (96,615 colors)



Пример: квантизация изображений

Quantized image (64 colors, Random)

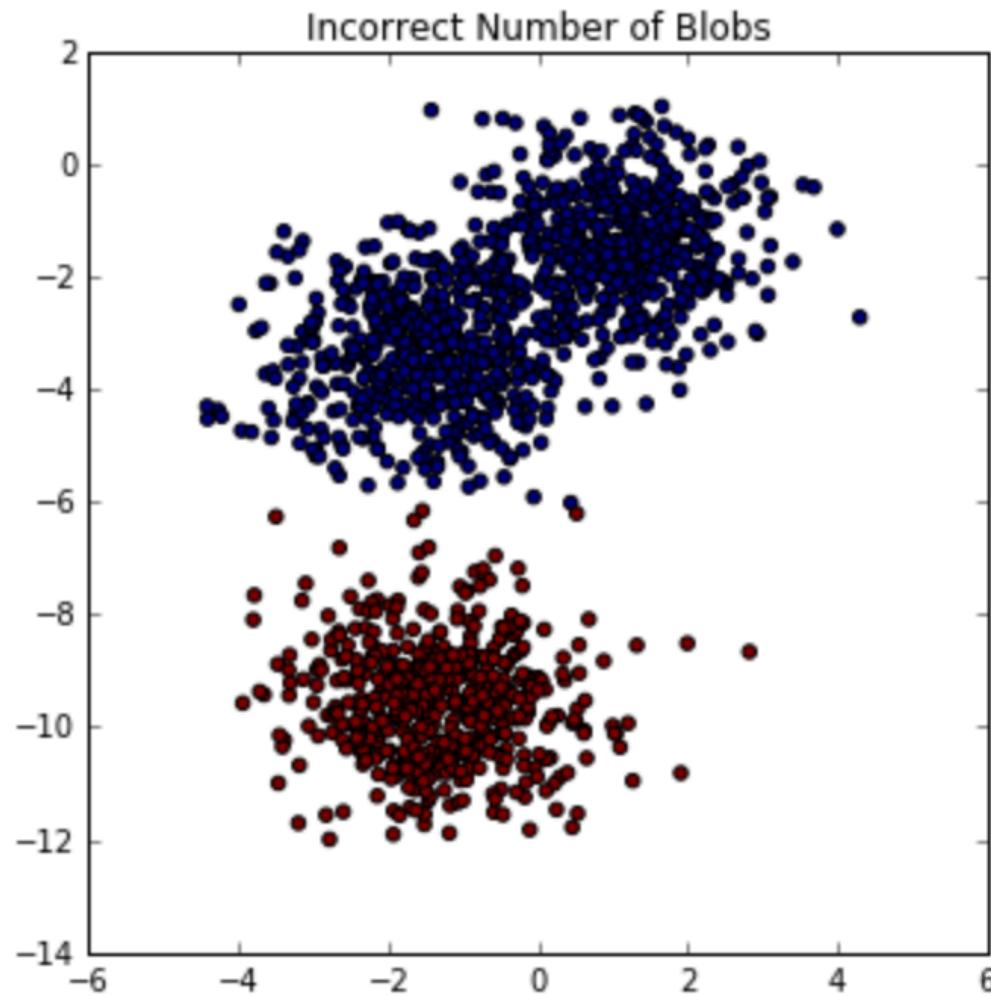


Пример: квантизация изображений

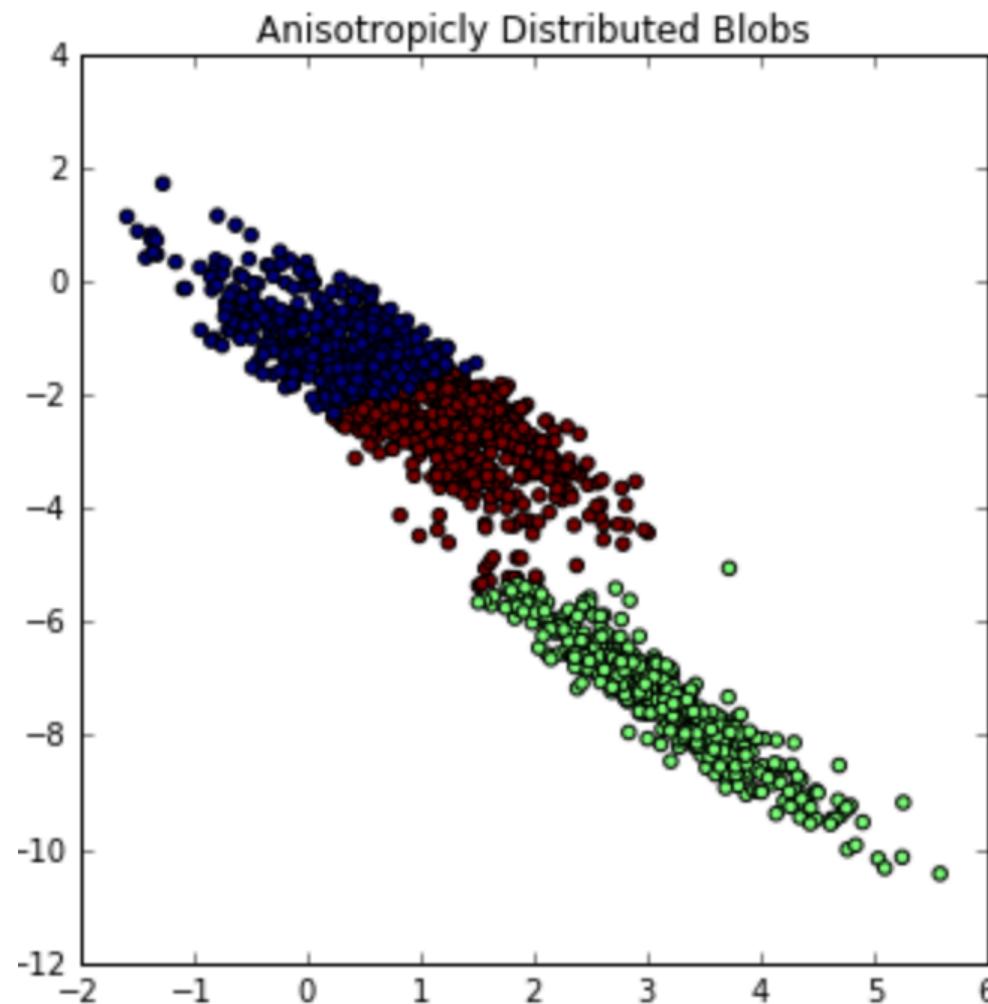
Quantized image (64 colors, K-Means)



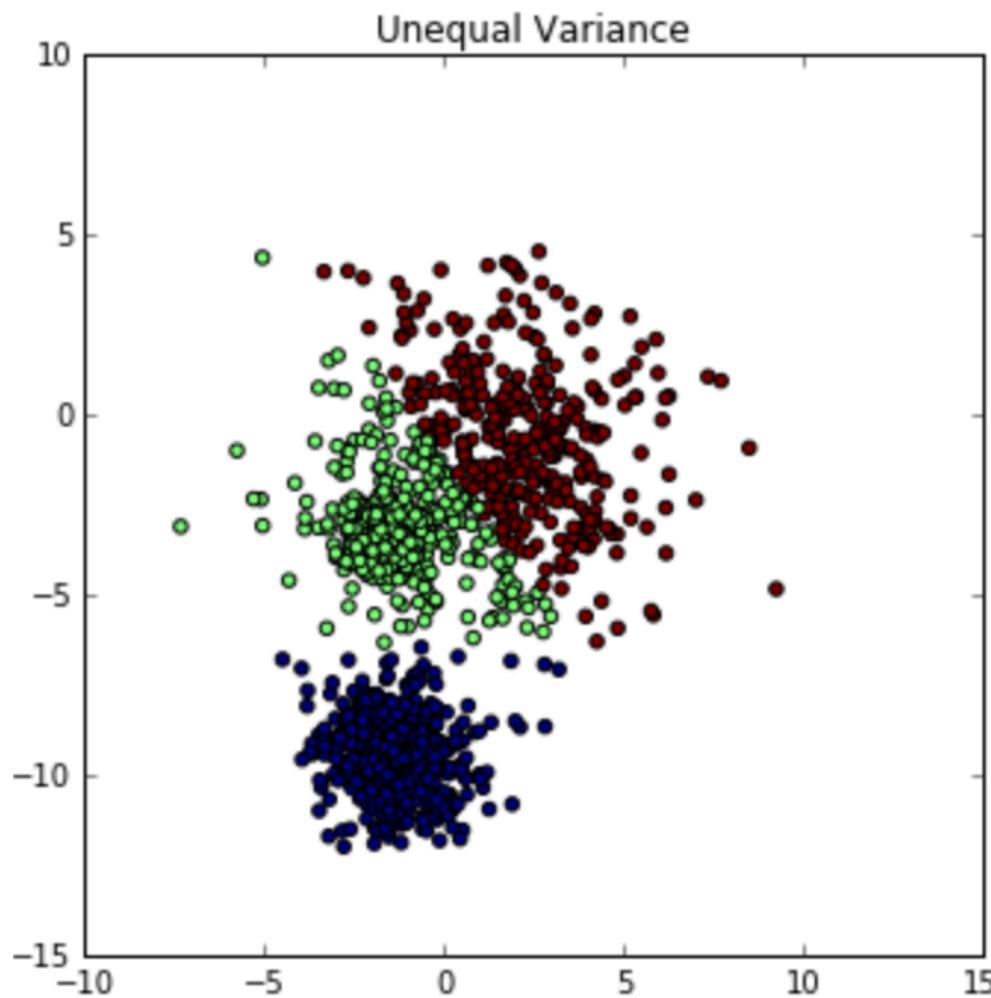
K Means и разные формы кластеров



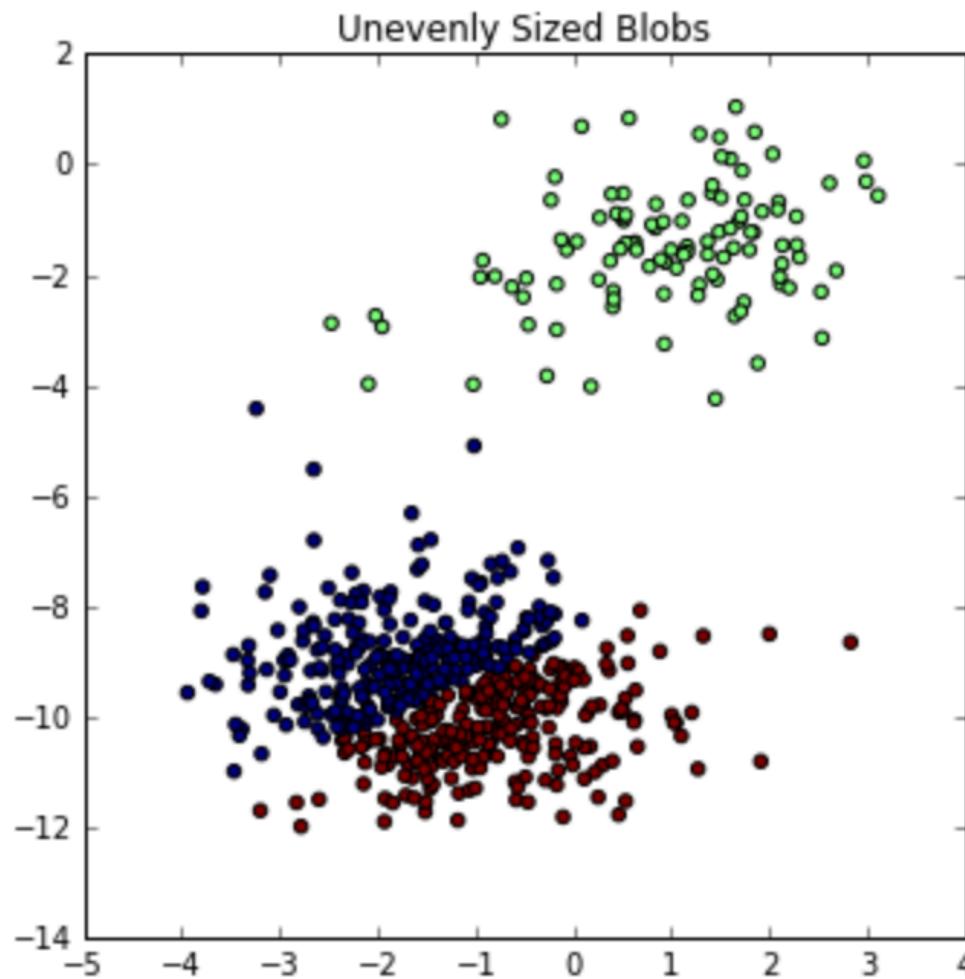
K Means и разные формы кластеров



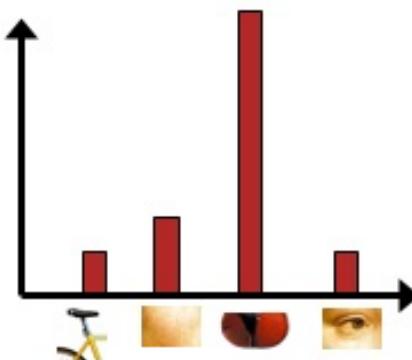
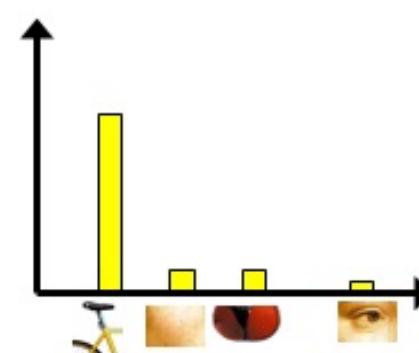
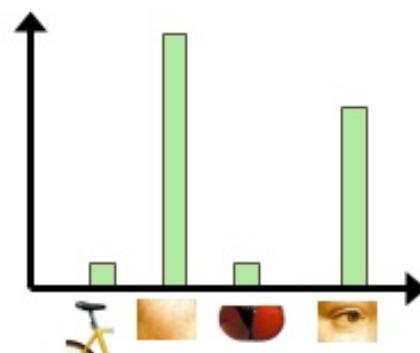
K Means и разные формы кластеров



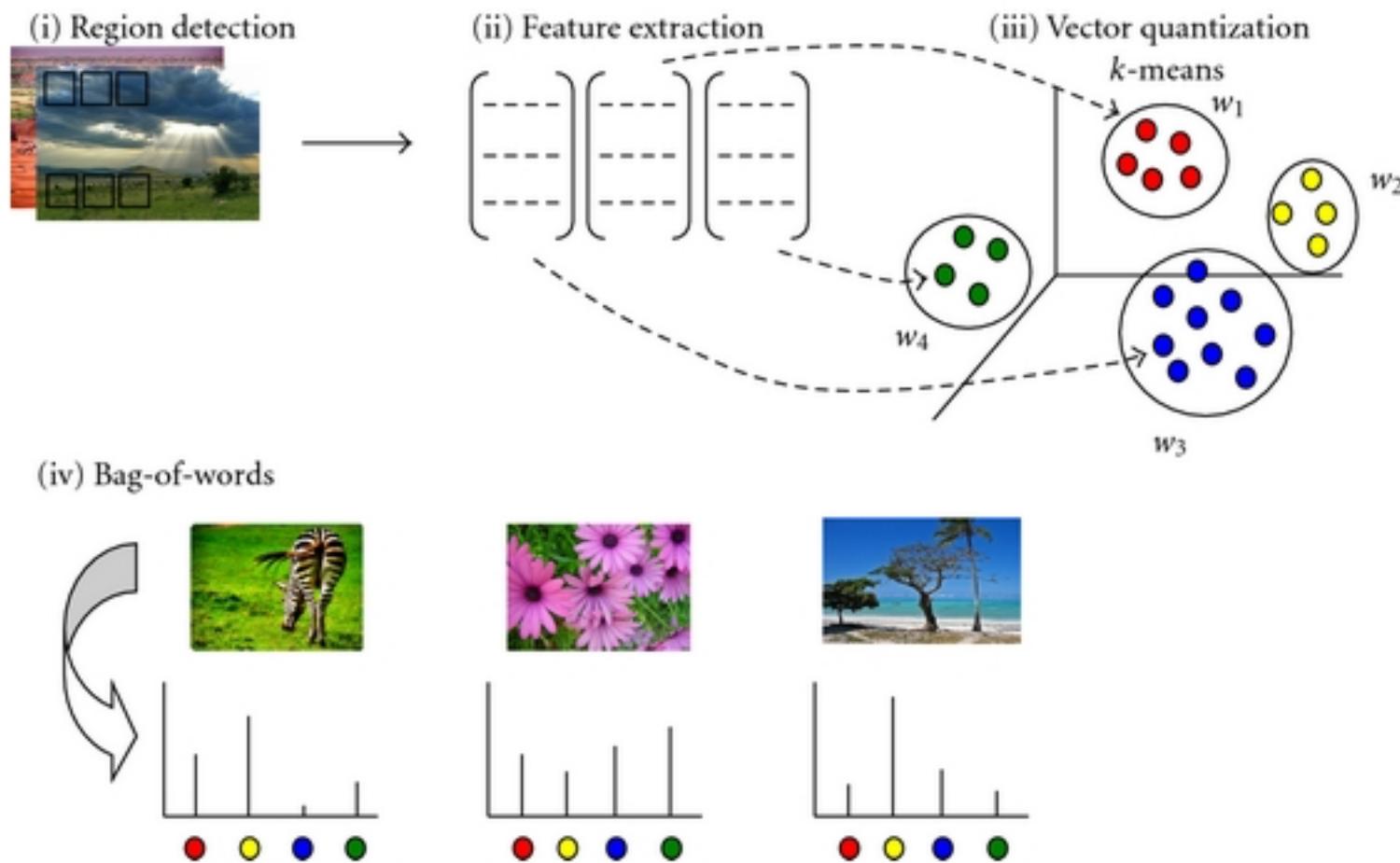
K Means и разные формы кластеров



Пример: мешок визуальных слов



Пример: мешок визуальных слов



Что оптимизирует K Means

Среднее внутрикластерное расстояние:

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

Что оптимизирует K Means

Среднее внутрикластерное расстояние:

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

Альтернативный вариант, если есть центры кластеров:

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

Что оптимизирует K Means

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

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

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

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

Что оптимизирует K Means

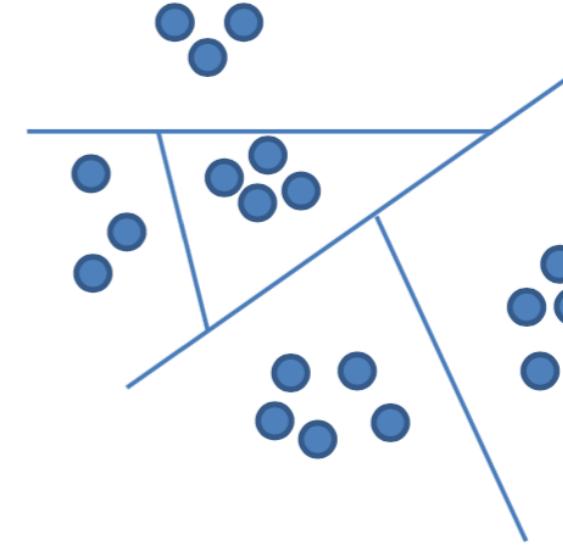
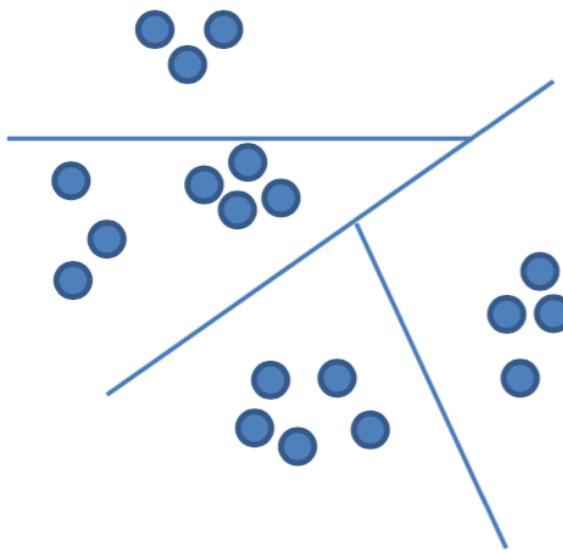
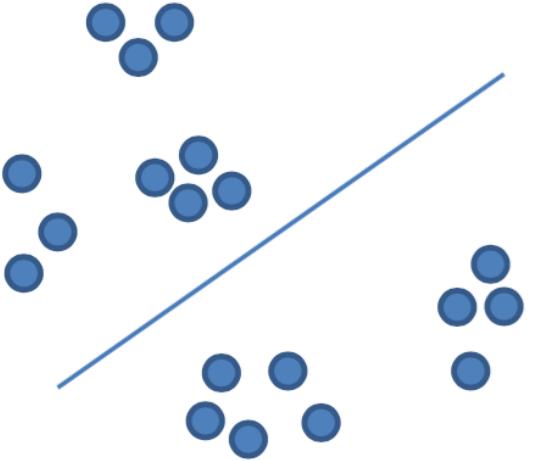
K Means итеративно минимизирует среднее внутрикластерное расстояние:

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

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

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

Подбор числа кластеров: BisectKMeans



ИТОГИ

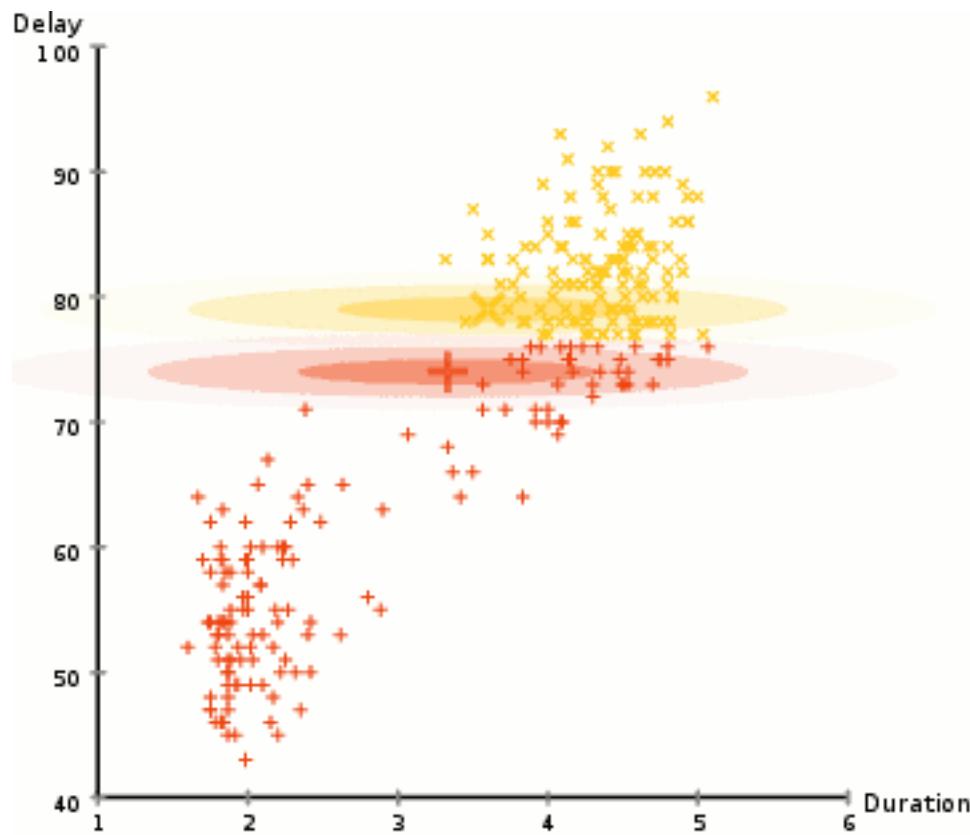
1. Как работает K Means
2. Что делать, когда данных много: Mini Batch K-Means
3. Что делать, когда много признаков: понижение размерности
4. Выбор начальных приближений: Kmeans++
5. Пример: квантизация изображений
6. Работа K means с разными формами кластеров
7. Пример: мешок визуальных слов (bag of visual words)
8. Что оптимизирует K means

4. Expectation-Maximization (EM-алгоритм)

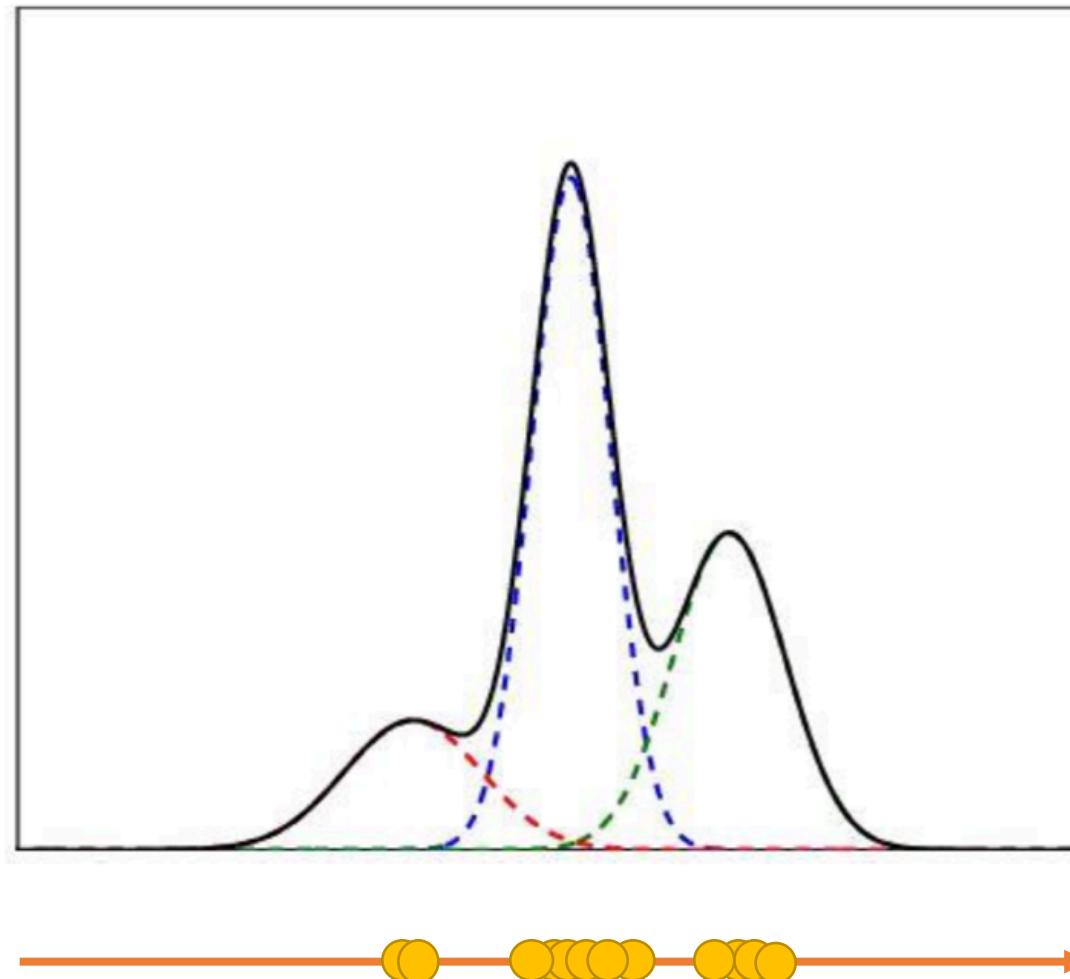
План

1. Как выглядит кластеризация с помощью ЕМ-алгоритма
2. Постановка задачи
3. Почему не решить «в лоб»
4. Описание ЕМ алгоритма
5. ЕМ-алгоритм в случае гауссовских распределений
6. Простое объяснение метода
7. Классическое объяснение метода
8. Для чего еще используют алгоритм

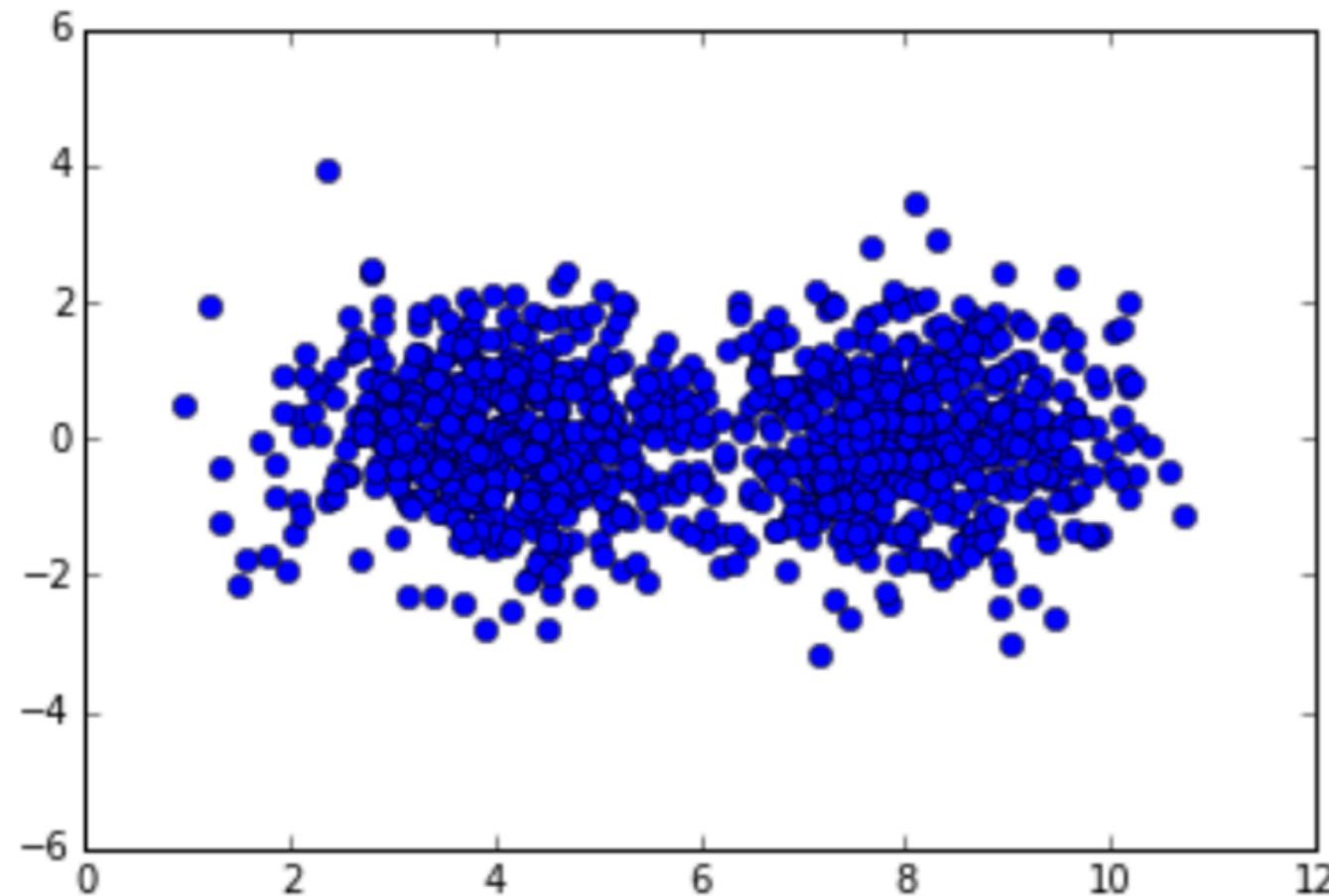
Как это выглядит



Как выглядит смесь распределений



Как выглядит смесь распределений



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 :

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

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 $p_1(x), \dots, p_K(x)$

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

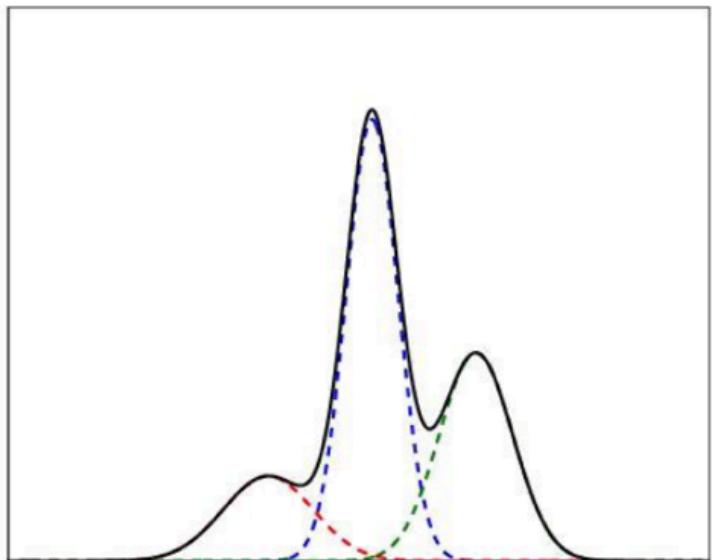
Постановка задачи: разделение смеси

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

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

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

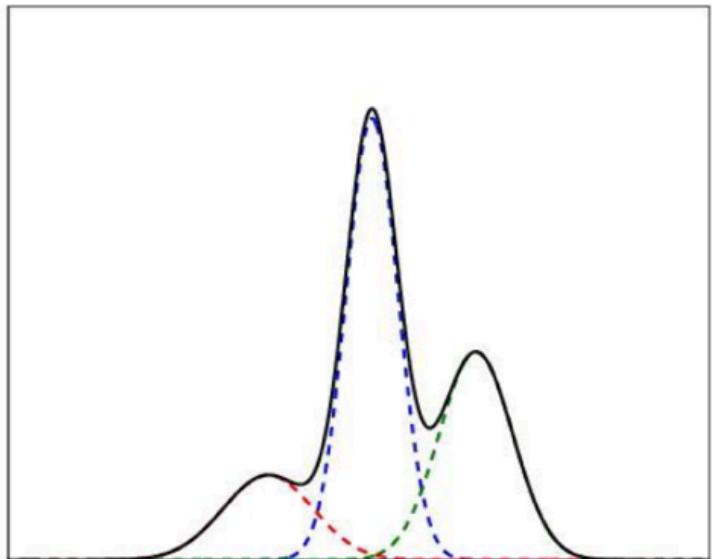
Почему не решить задачу «в лоб»



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

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

Почему не решить задачу «в лоб»



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

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

EM-алгоритм

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

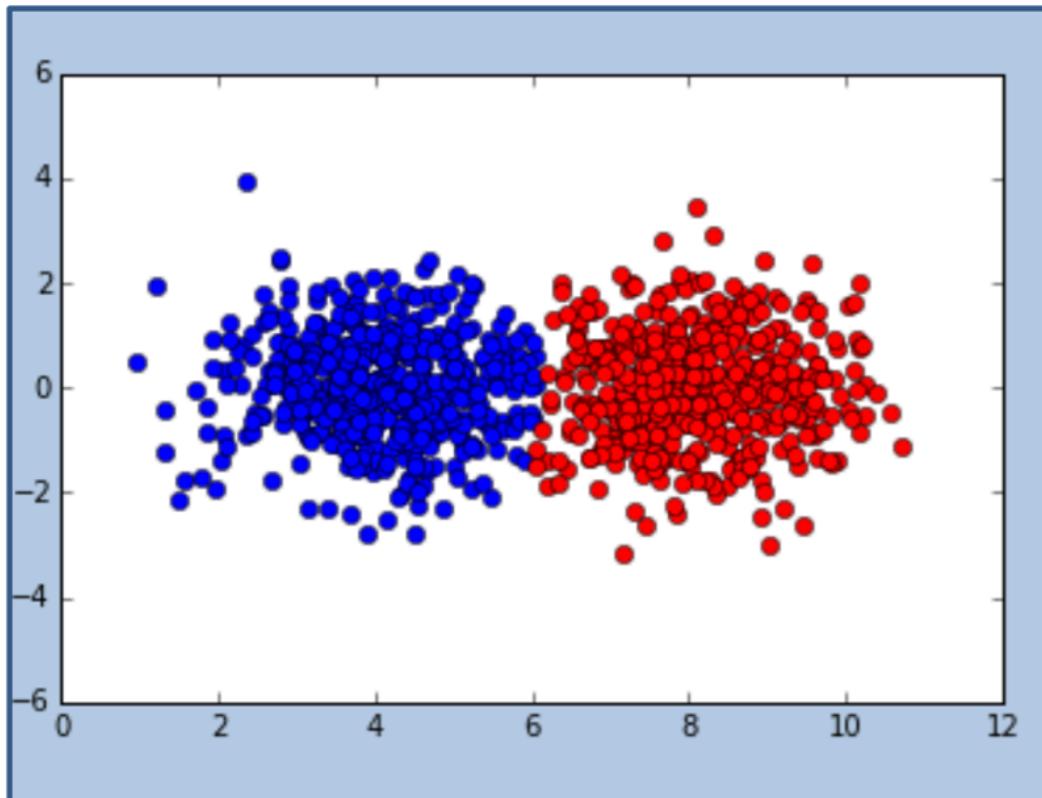
E-шаг:

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

M-шаг:

$$w_j = \frac{1}{N} \sum_{i=1}^N g_{ji} \quad \theta_j = \operatorname{argmax}_{\theta} \sum_{i=1}^N g_{ji} \ln \varphi(\theta; x)$$

Пример: 2 кластера с гауссовой плотностью



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

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

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

M-шаг:

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

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

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

Простое объяснение ЕМ-алгоритма

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

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:

$$w_j = \frac{1}{N} \sum_{i=1}^N g_{ji} \quad \theta_j = \operatorname{argmax}_{\theta} \sum_{i=1}^N g_{ji} \ln \varphi(\theta; x)$$

Какие еще задачи решаются с помощью ЕМ-алгоритма

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

Резюме

1. Как выглядит кластеризация с помощью ЕМ-алгоритма
2. Постановка задачи
3. Почему не решить «в лоб»
4. Описание ЕМ алгоритма
5. ЕМ-алгоритм в случае гауссовских распределений
6. Простое объяснение метода
7. Для чего еще используют алгоритм

5. Агломеративная иерархическая кластеризация

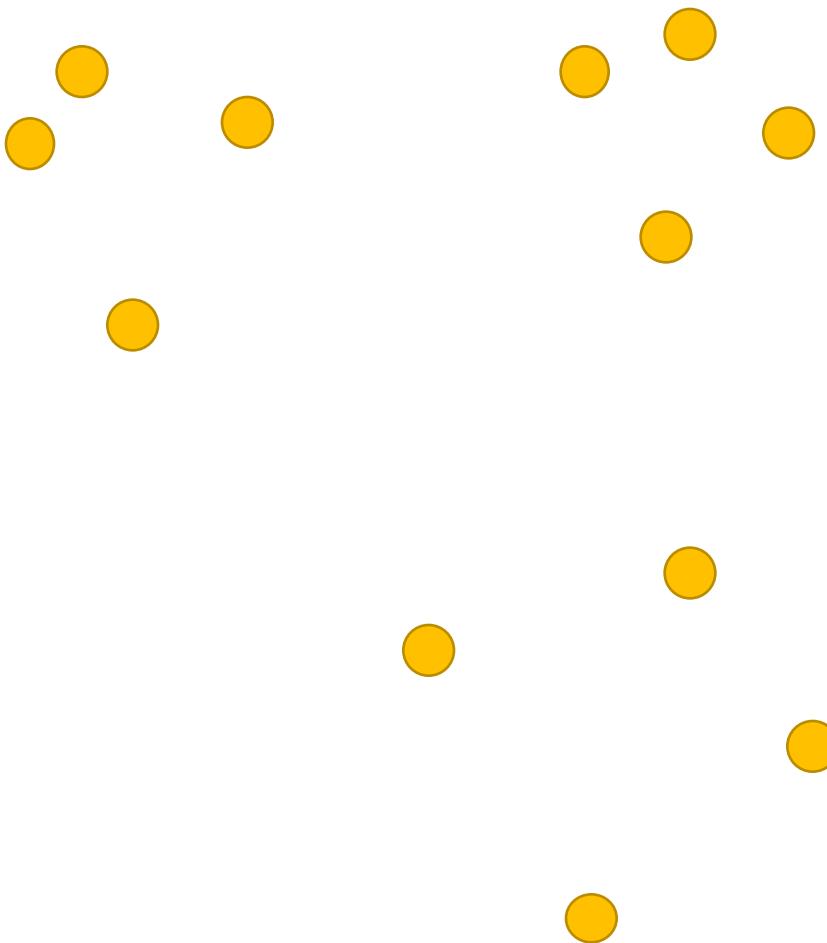
План

1. Иерархическая кластеризация
2. Как устроена агломеративная кластеризация
3. Расстояние между кластерами
4. Формула Ланса-Уильямса
5. Дендрограммы
6. Примеры работы

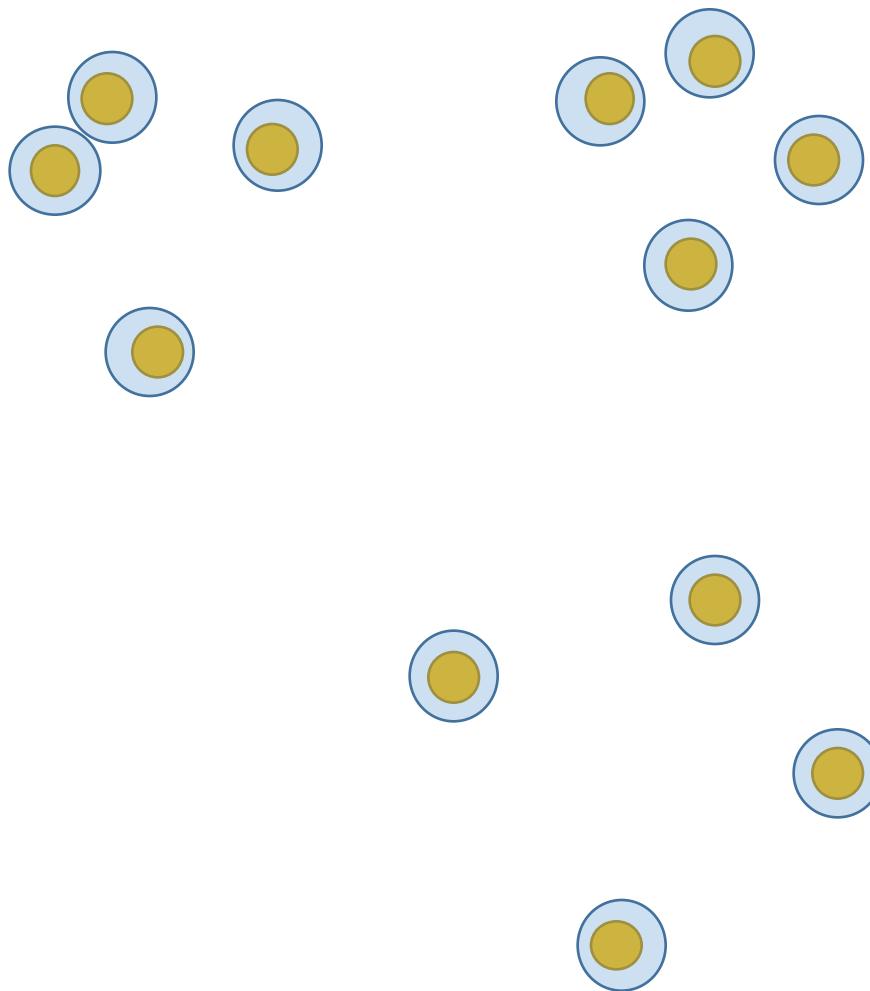
Иерархическая кластеризация

- Агломеративная
- Дивизионная или дивизимная

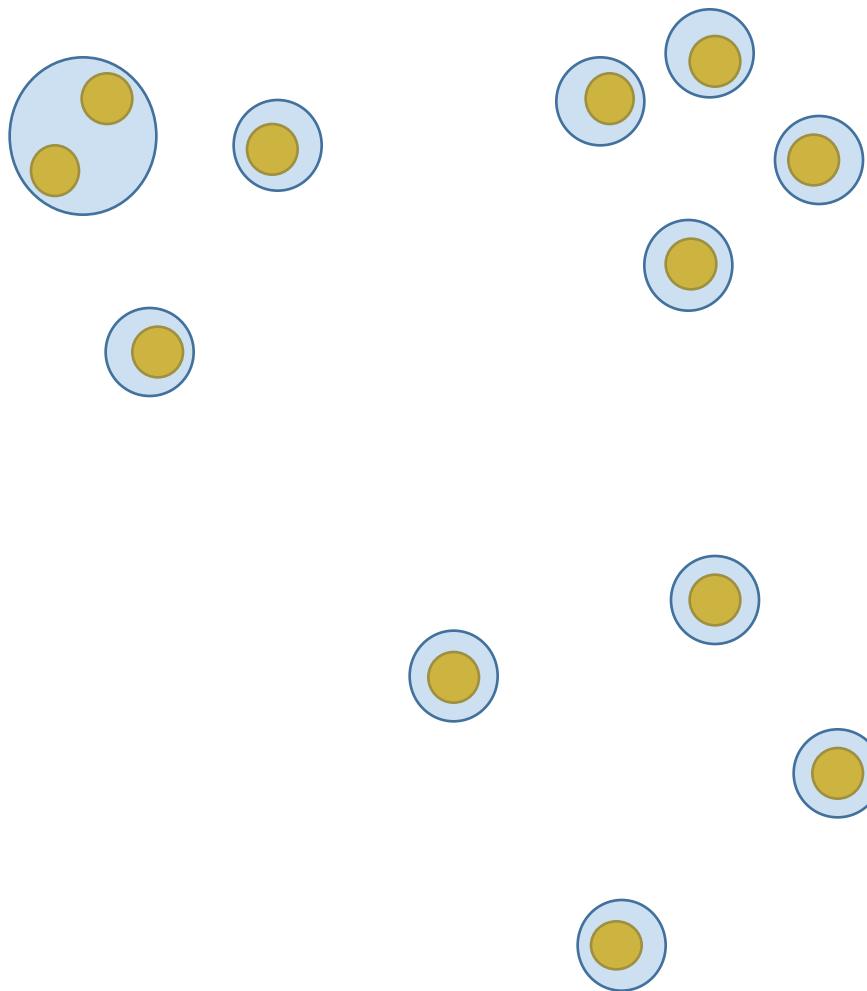
Агломеративная кластеризация



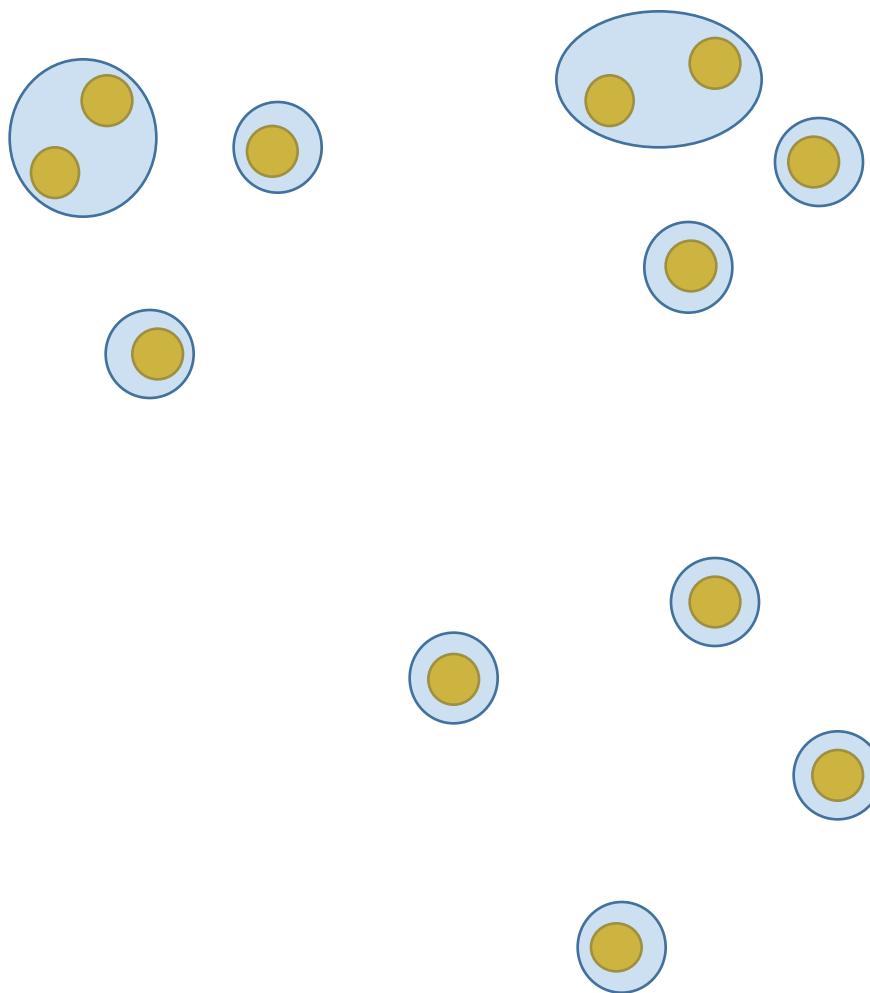
Агломеративная кластеризация



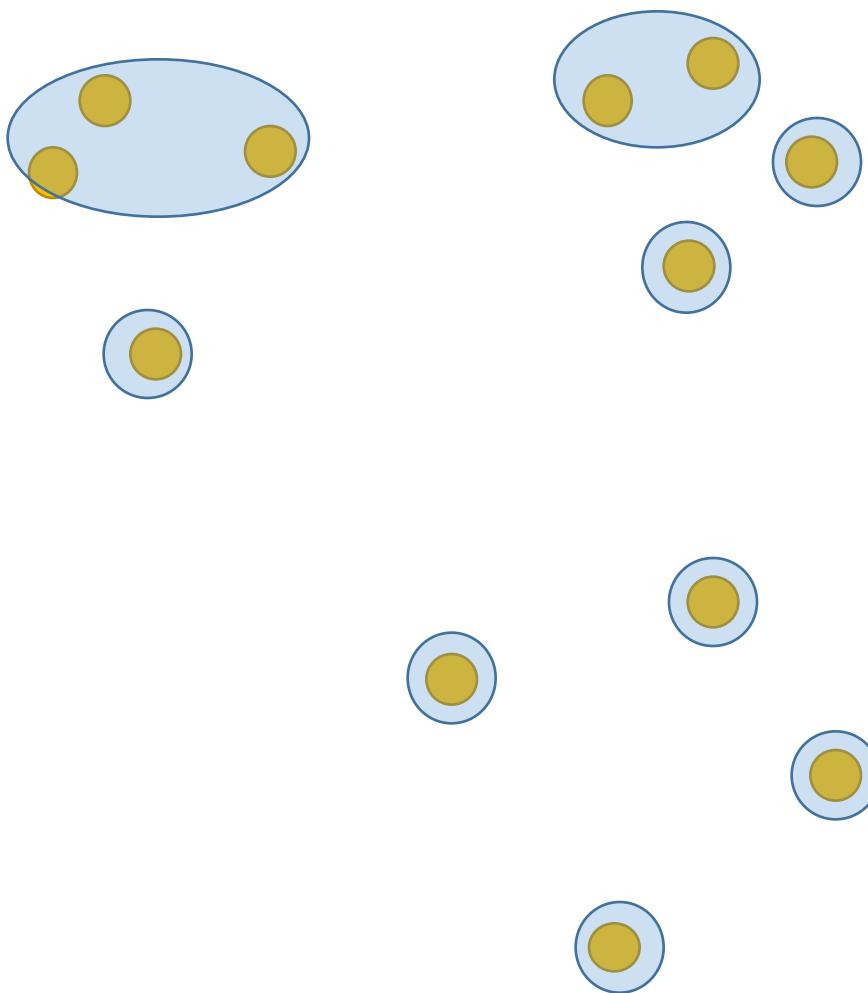
Агломеративная кластеризация



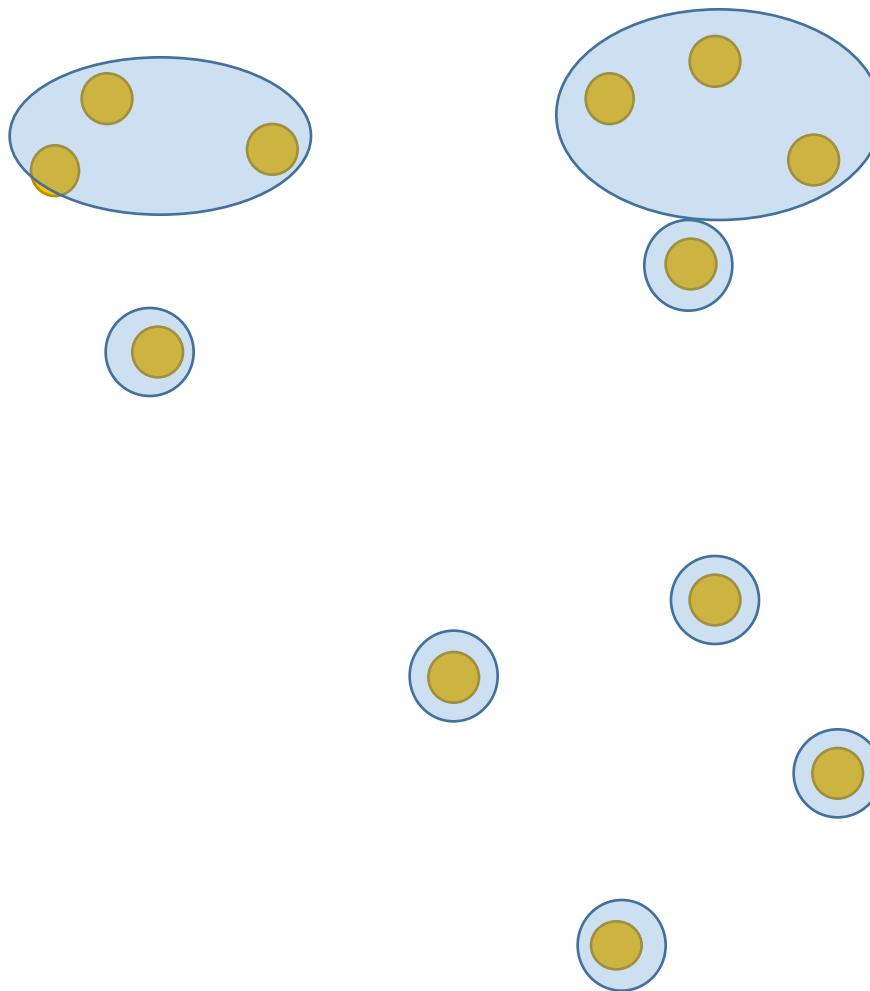
Агломеративная кластеризация



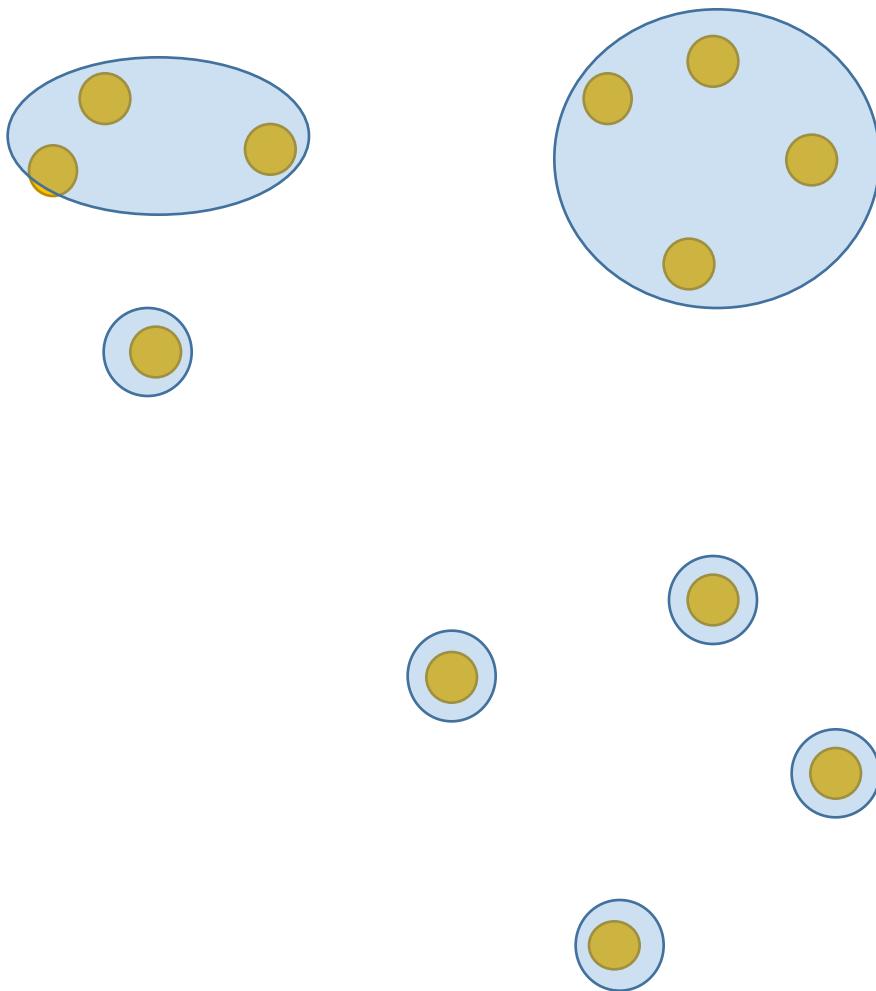
Агломеративная кластеризация



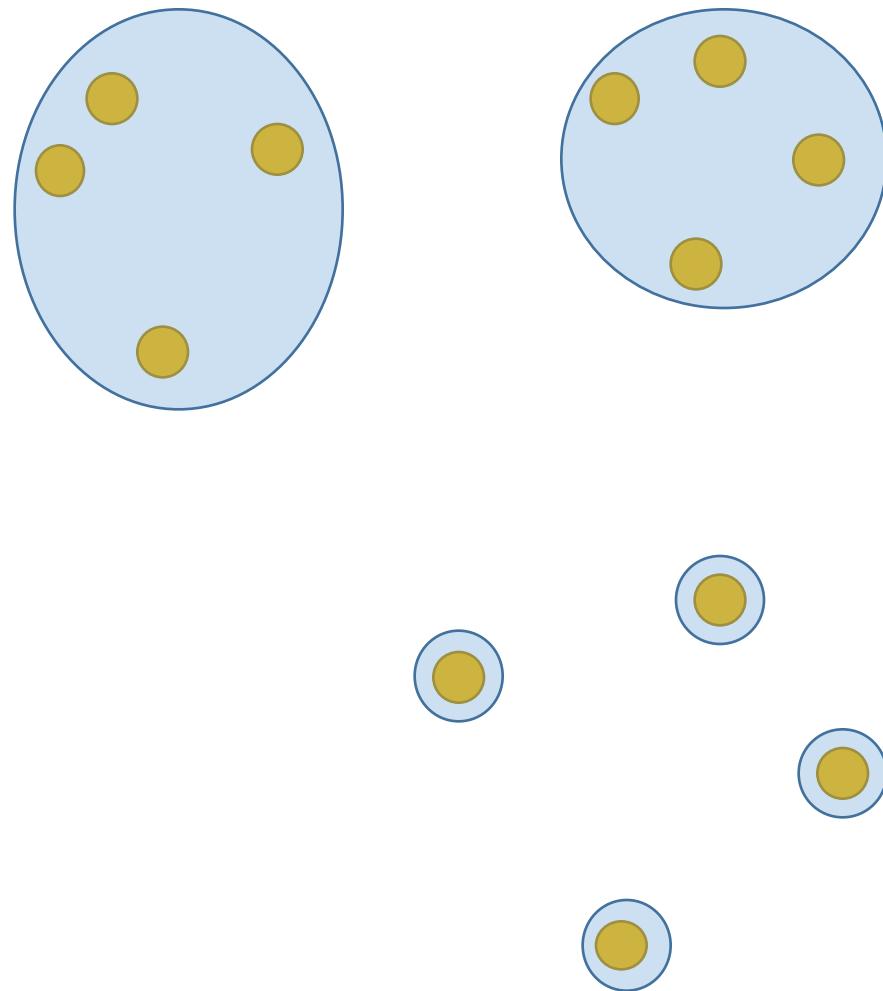
Агломеративная кластеризация



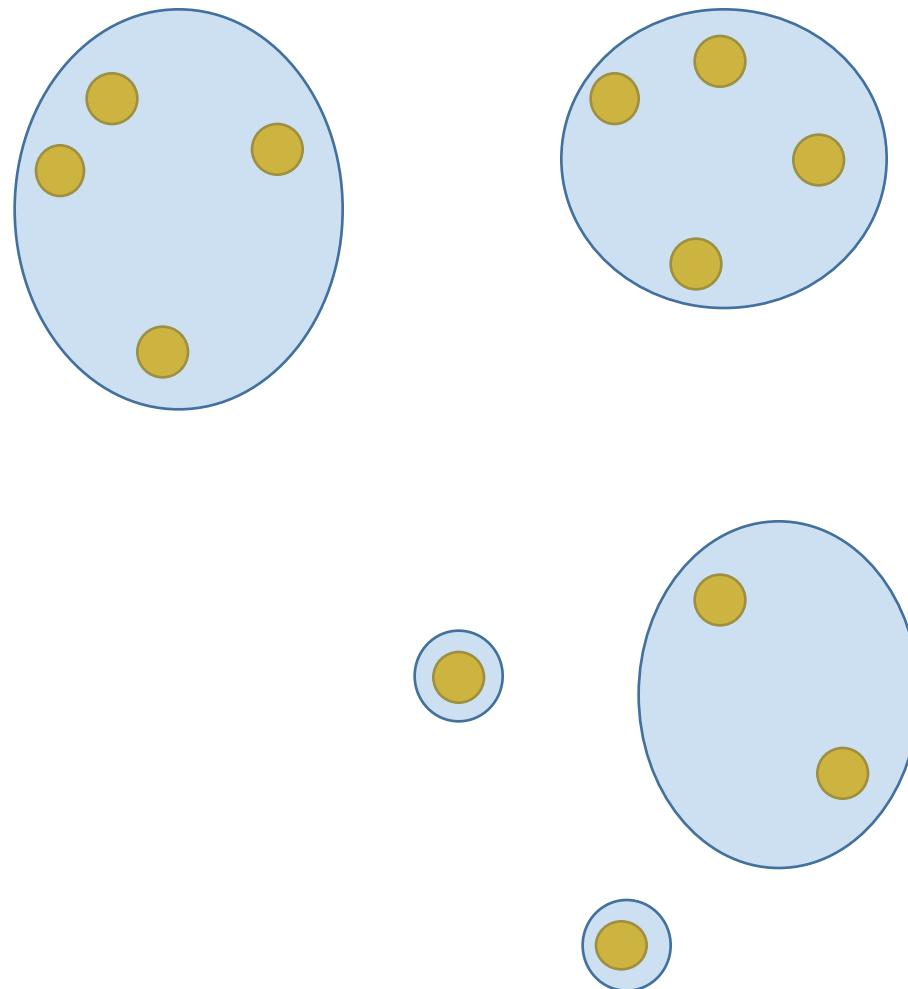
Агломеративная кластеризация



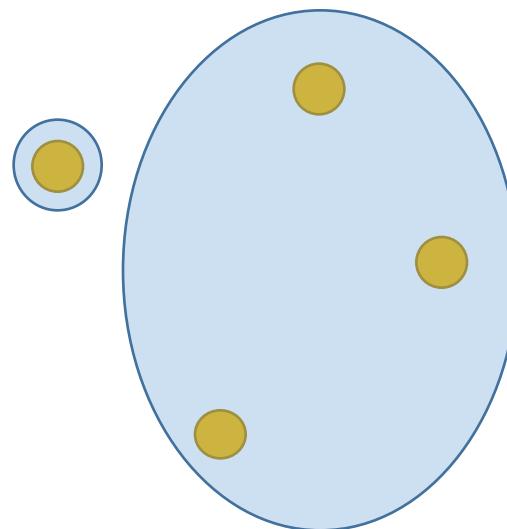
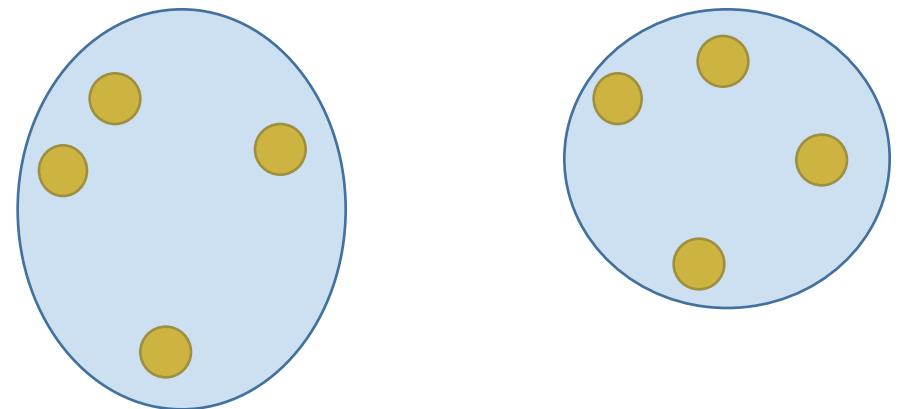
Агломеративная кластеризация



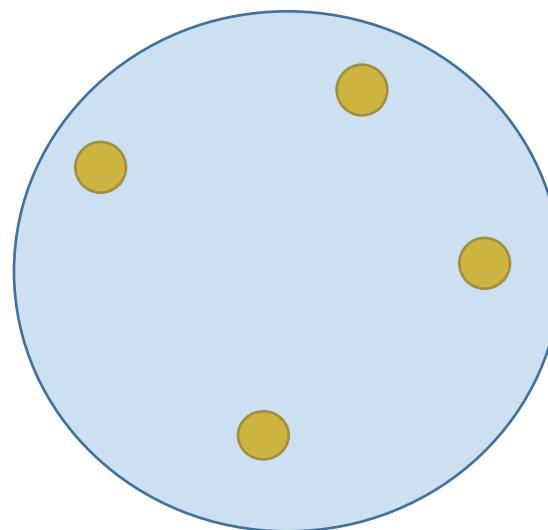
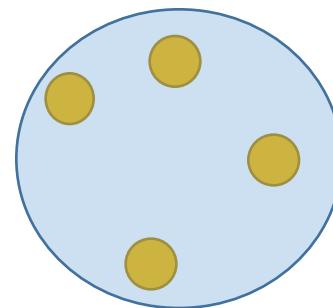
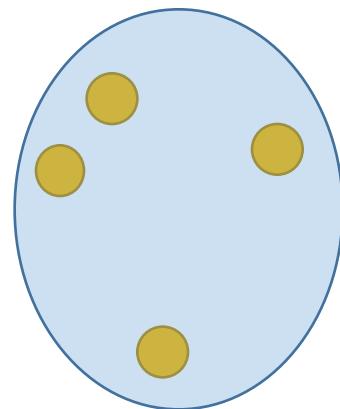
Агломеративная кластеризация



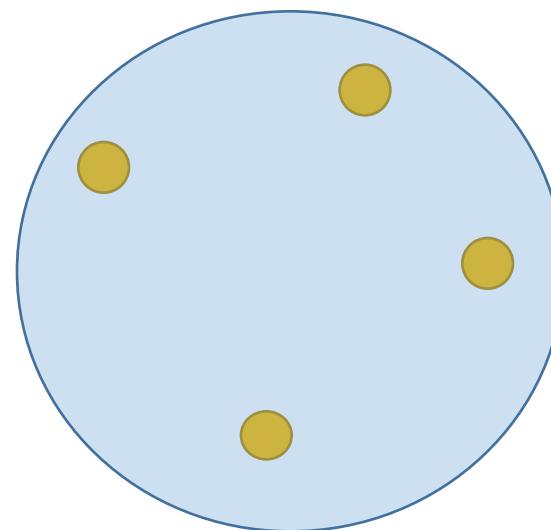
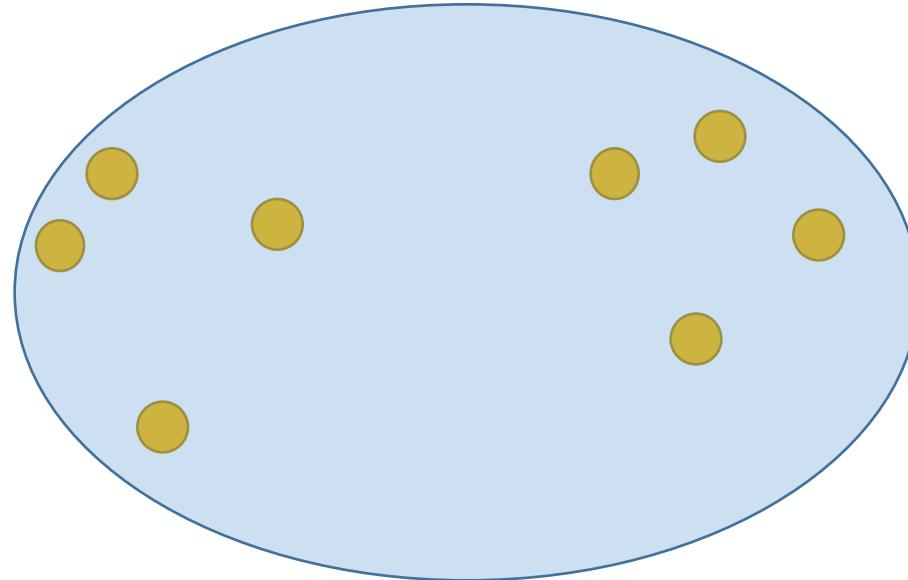
Агломеративная кластеризация



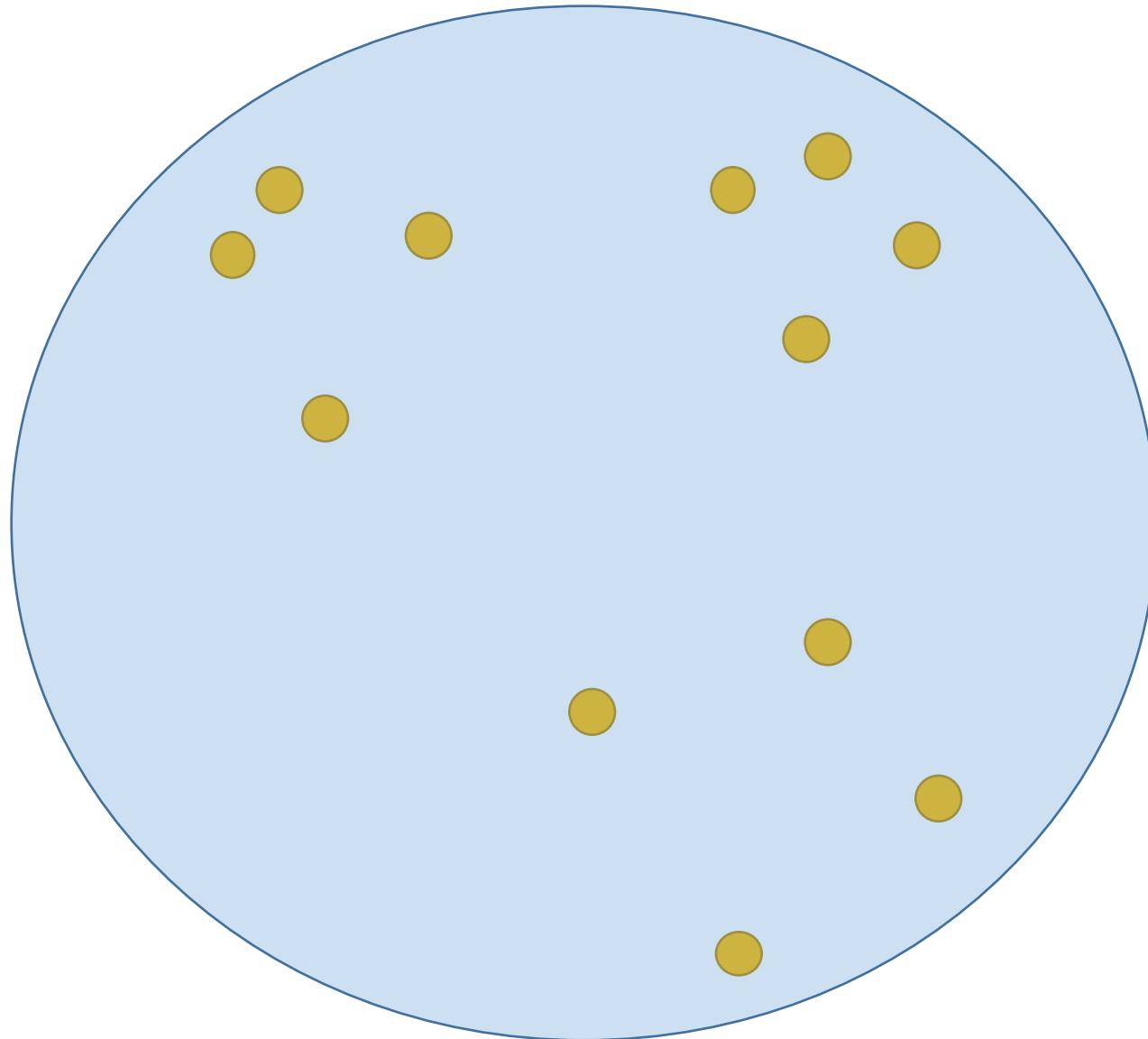
Агломеративная кластеризация



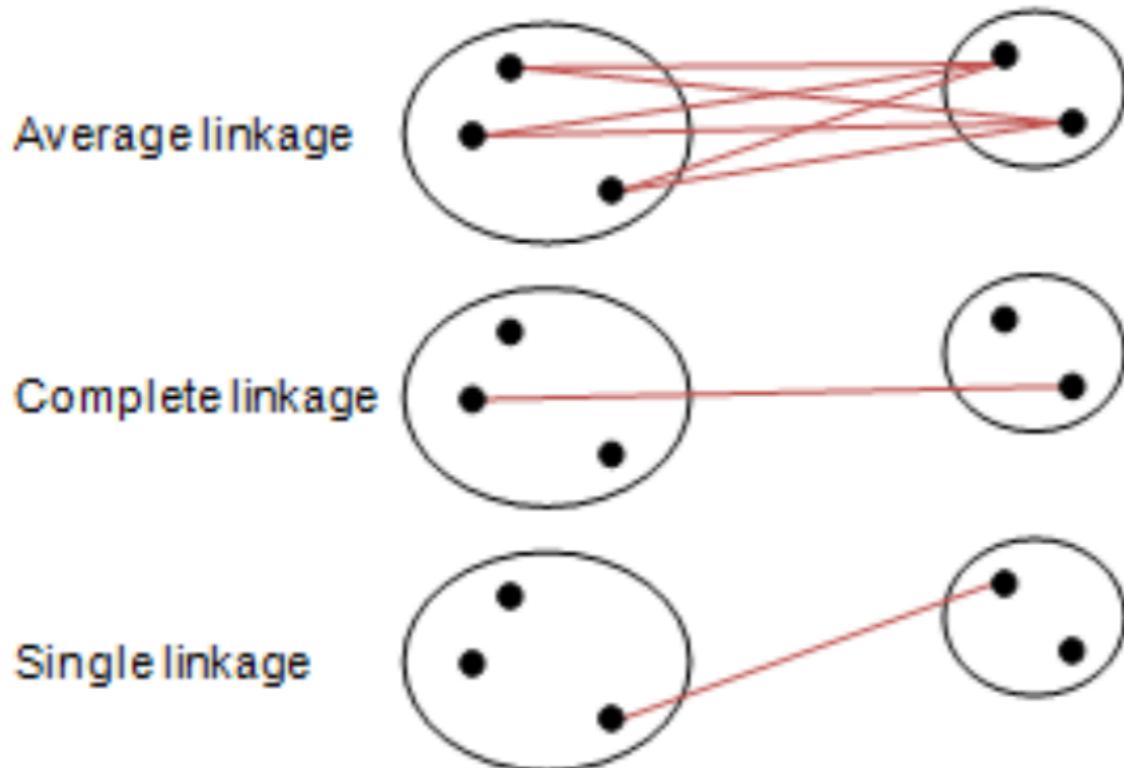
Агломеративная кластеризация



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:

$$R^6(W, S) = \min_{w \in W, s \in S} \rho(w, s); \quad \alpha_U = \alpha_V = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}.$$

Расстояние дальнего соседа:

$$R^\Delta(W, S) = \max_{w \in W, s \in S} \rho(w, s); \quad \alpha_U = \alpha_V = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}.$$

Среднее расстояние:

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

Формула Ланса-Уильямса

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

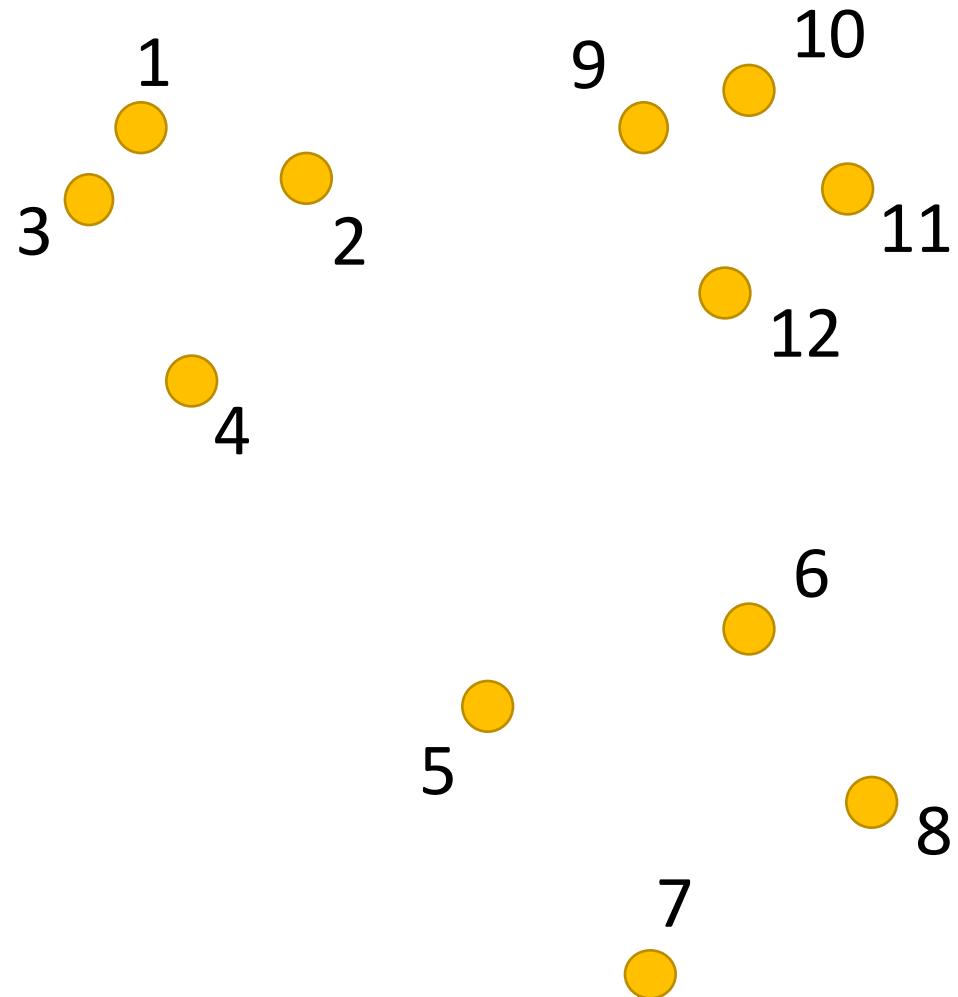
Расстояние между центрами:

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

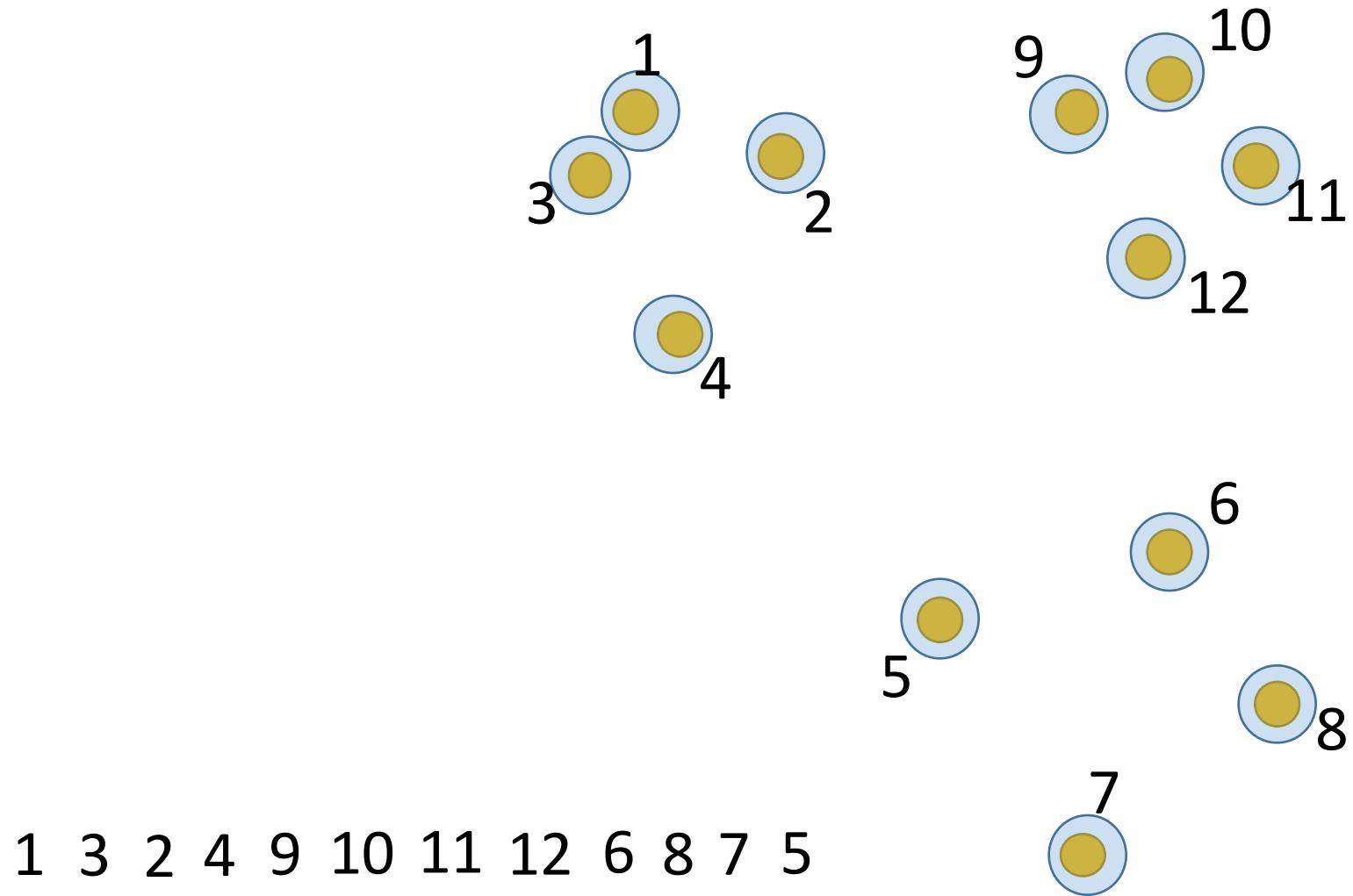
Расстояние Уорда:

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

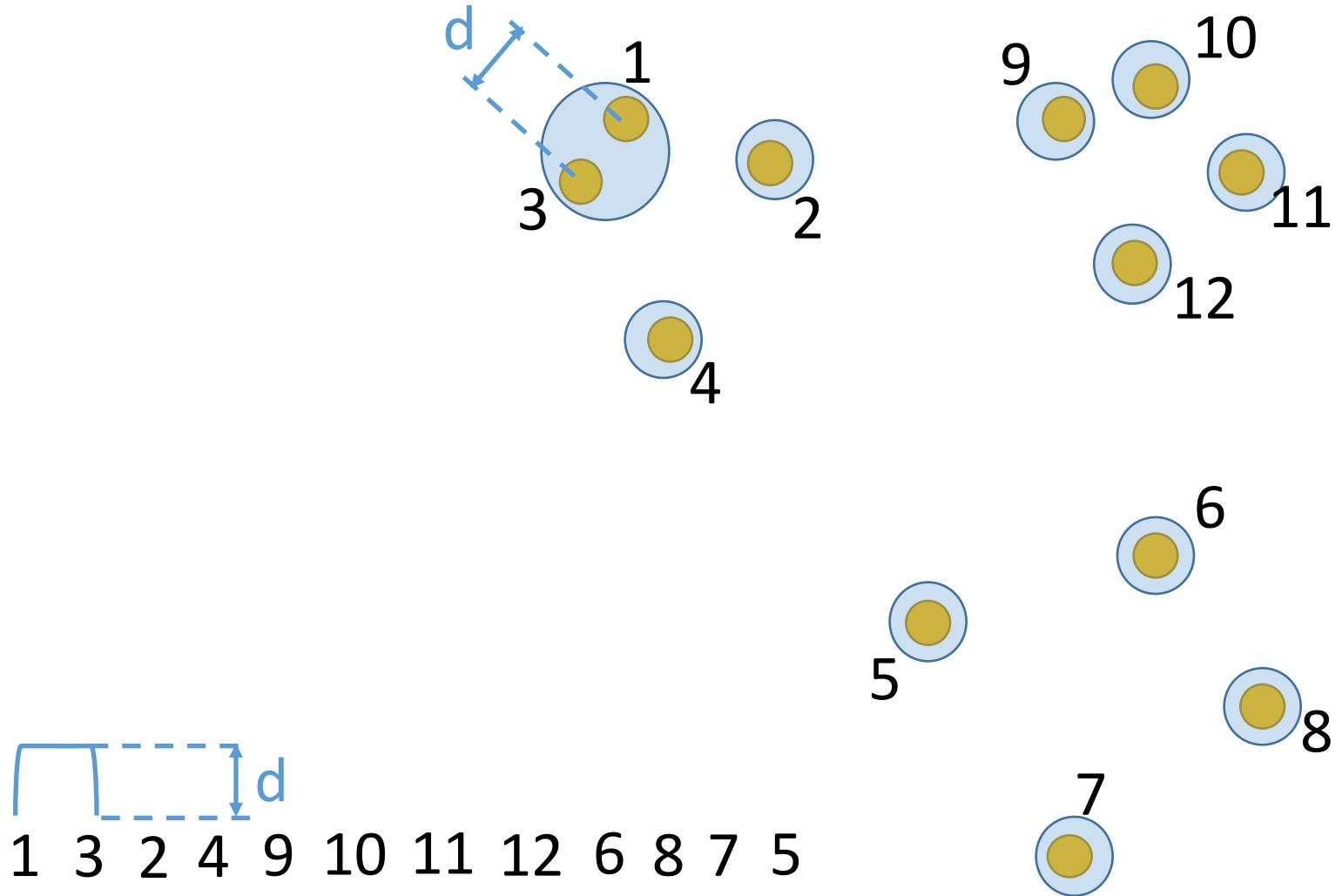
Дендрограмма



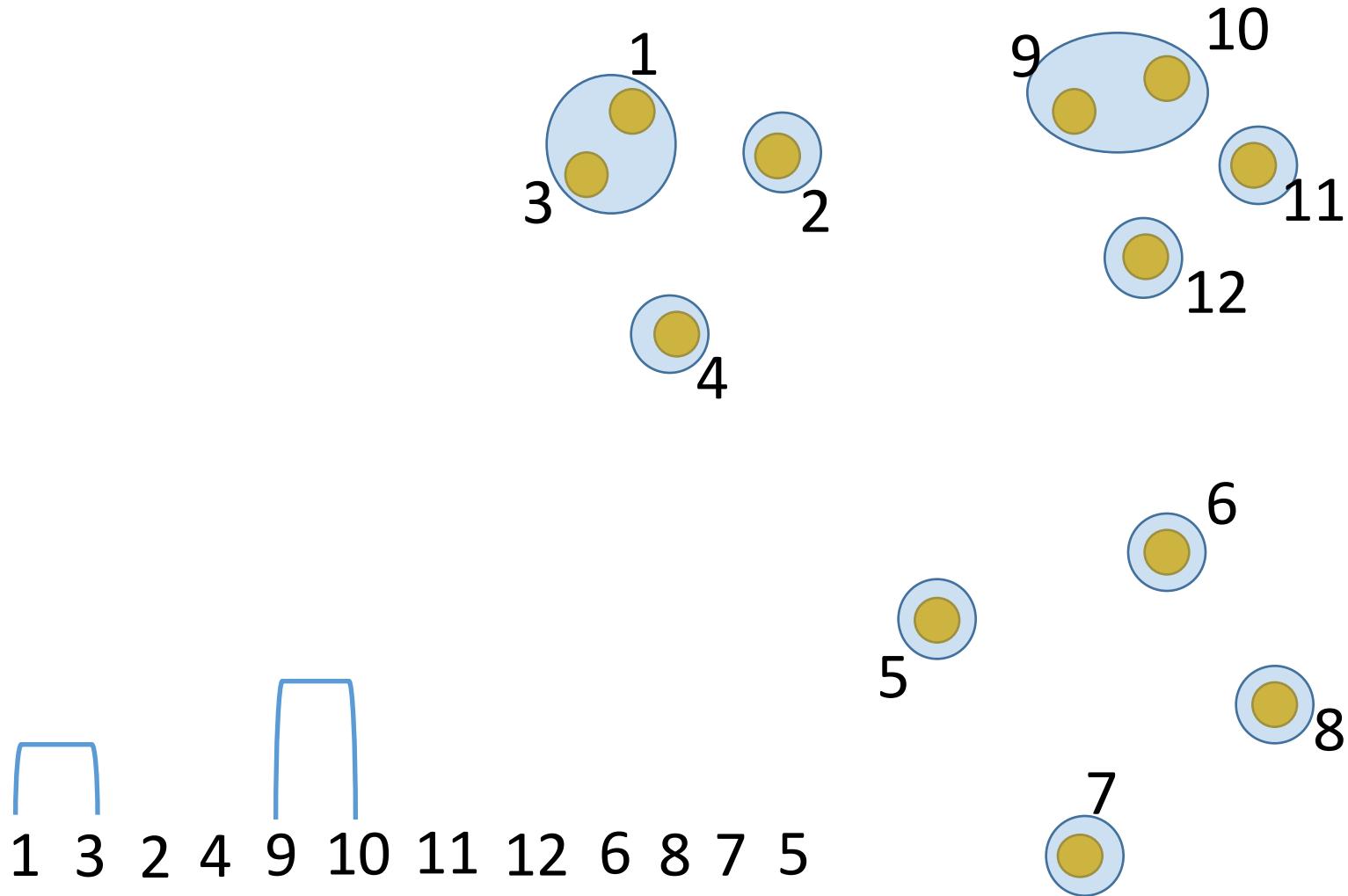
Дендрограмма



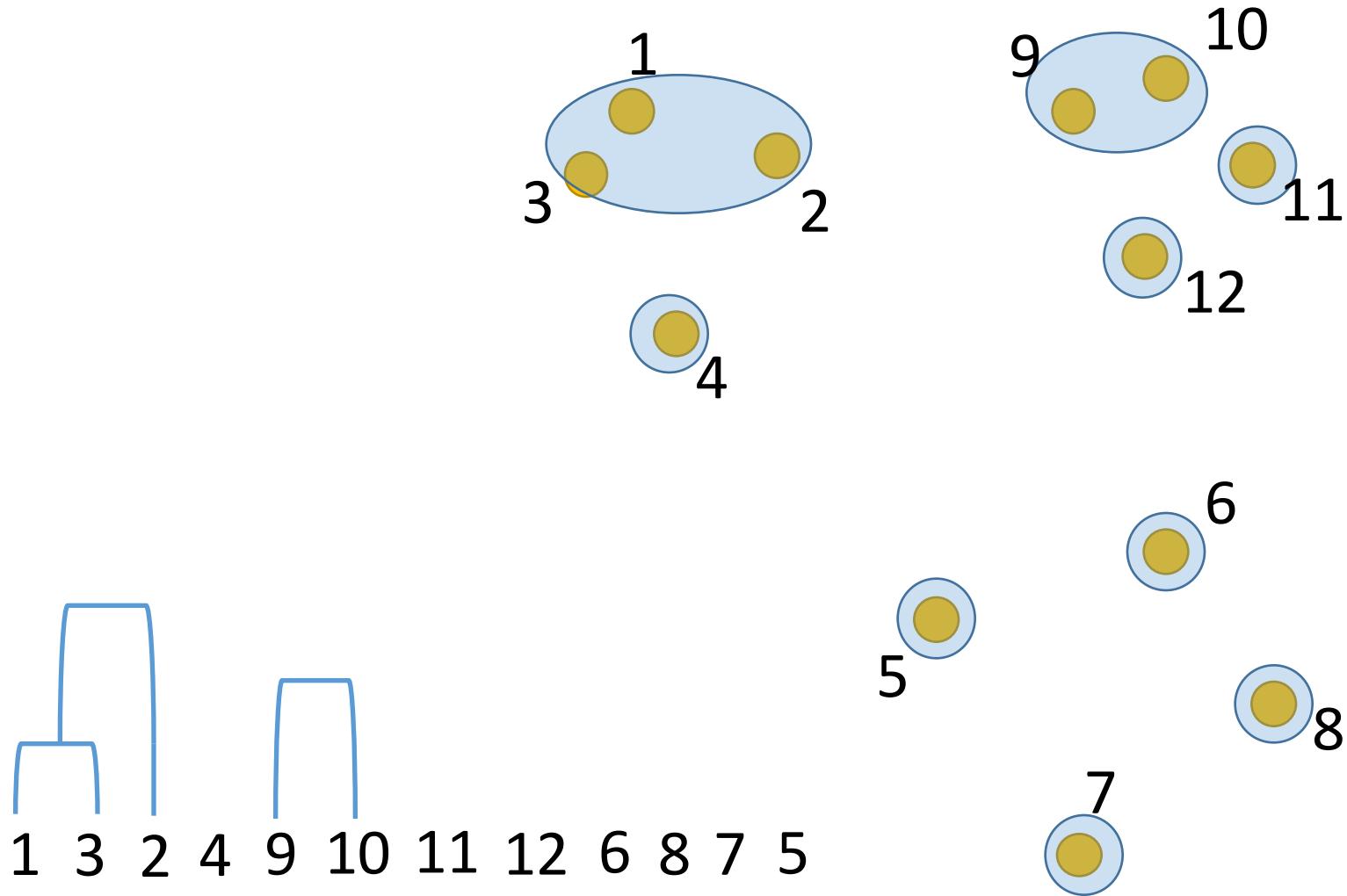
Дендрограмма



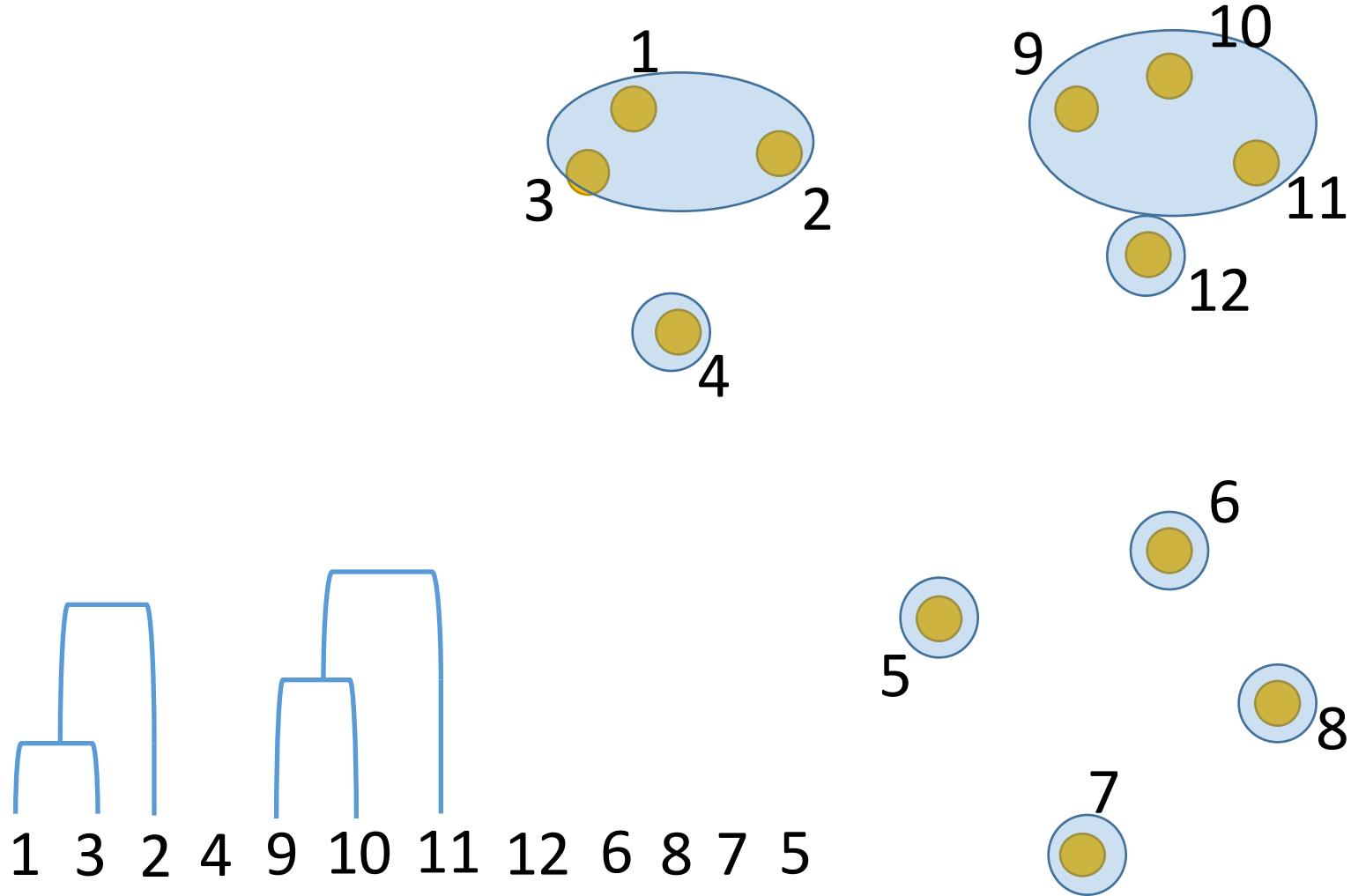
Дендрограмма



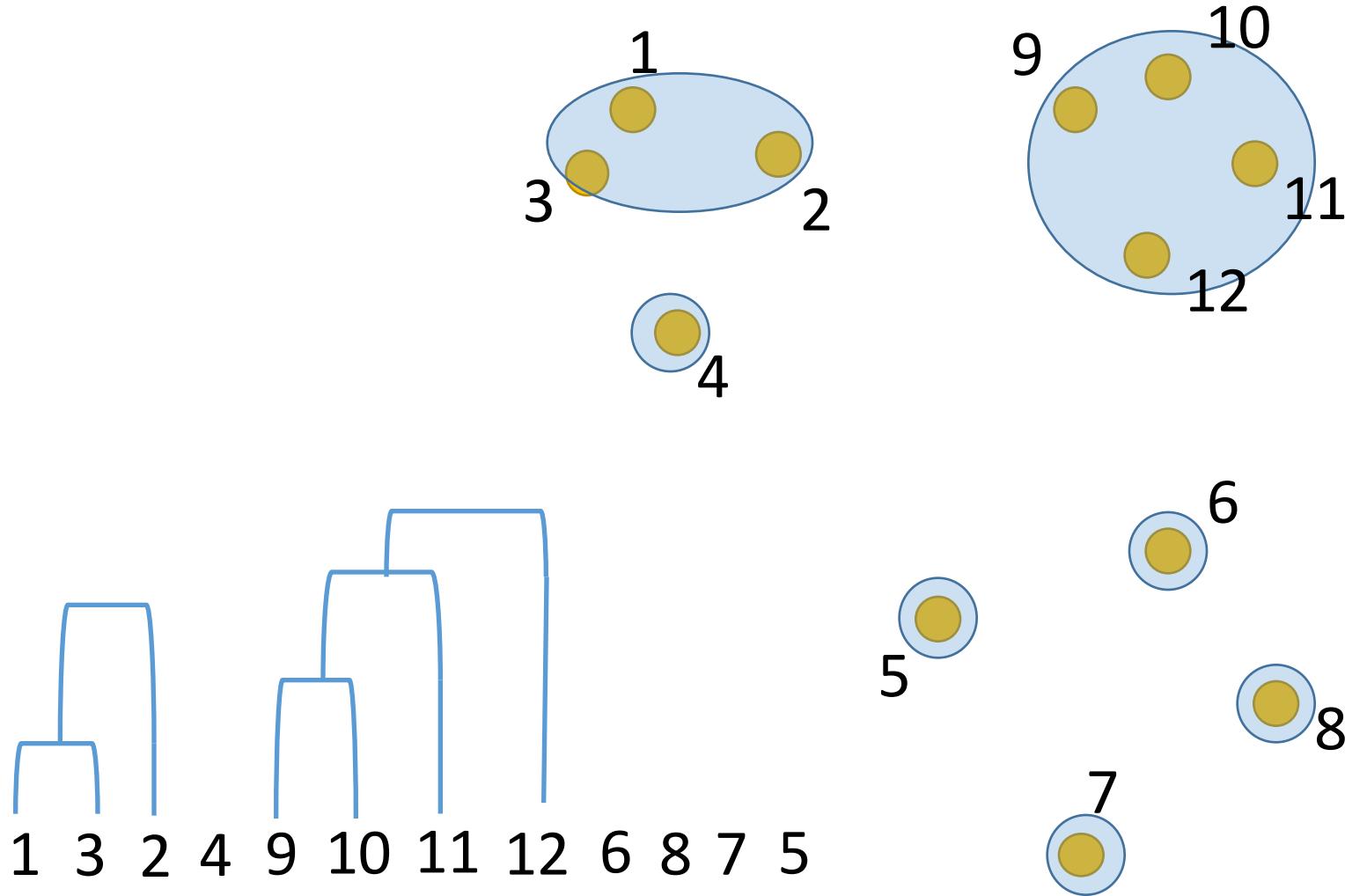
Дендрограмма



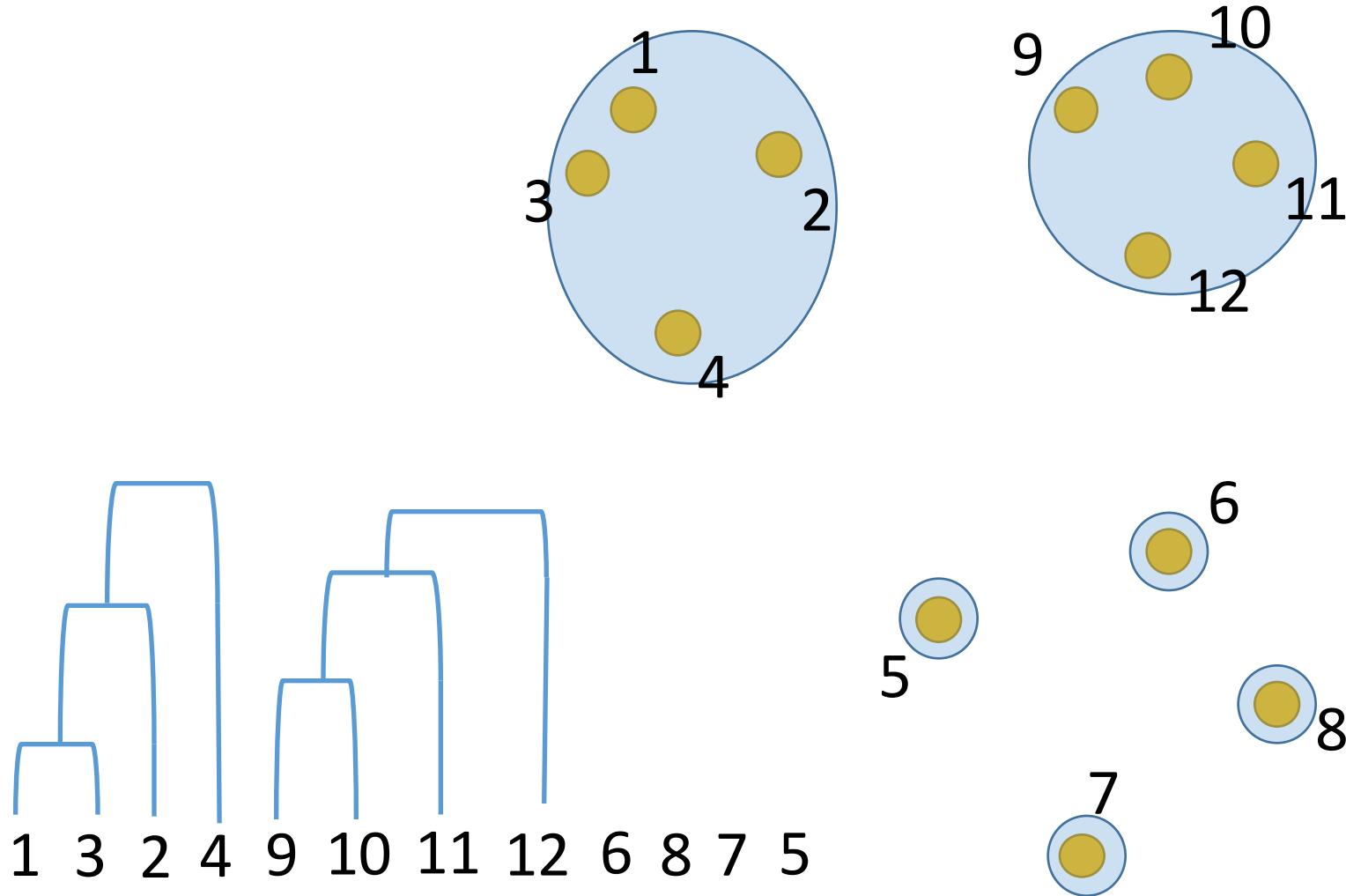
Дендрограмма



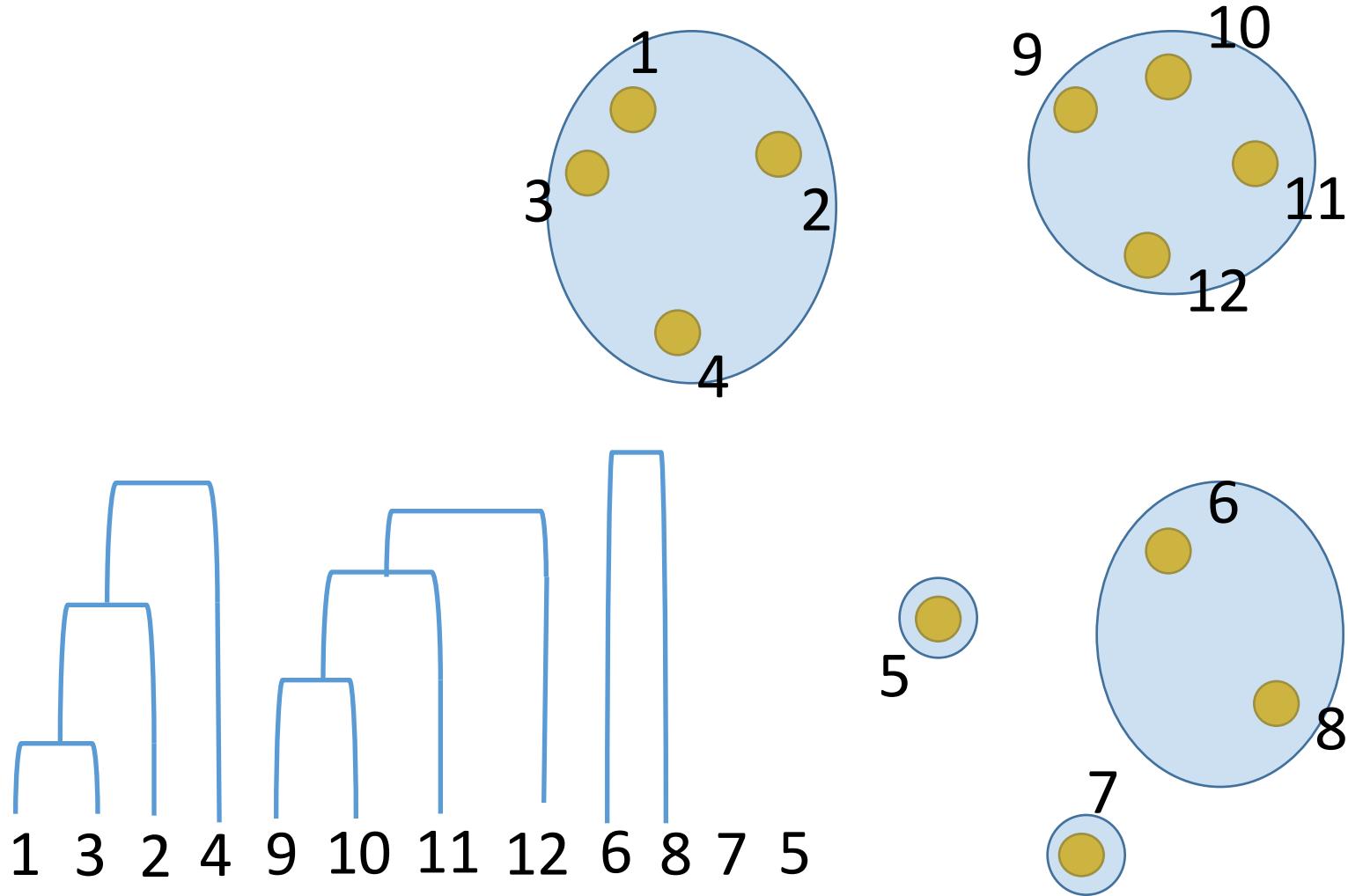
Дендрограмма



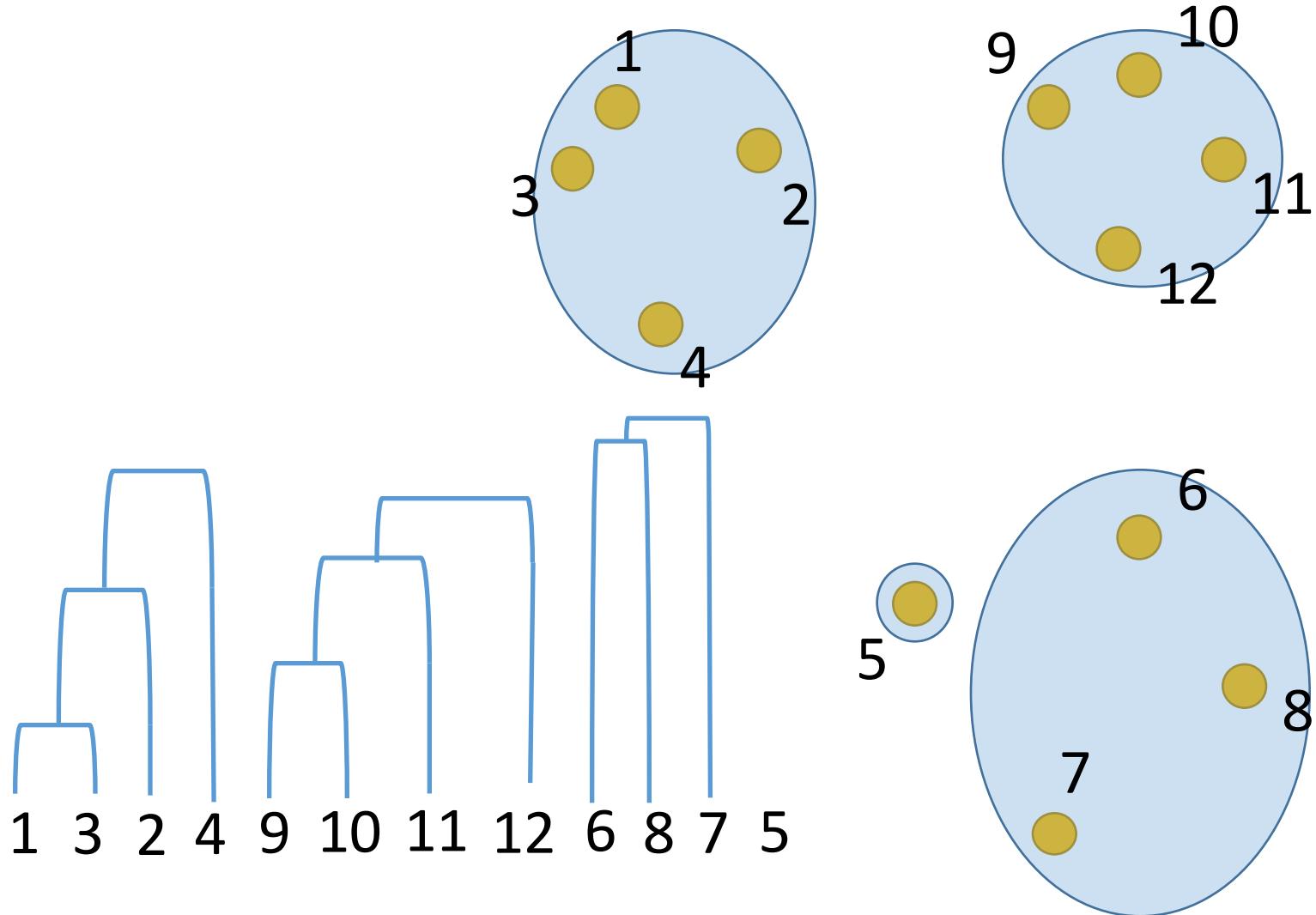
Дендрограмма



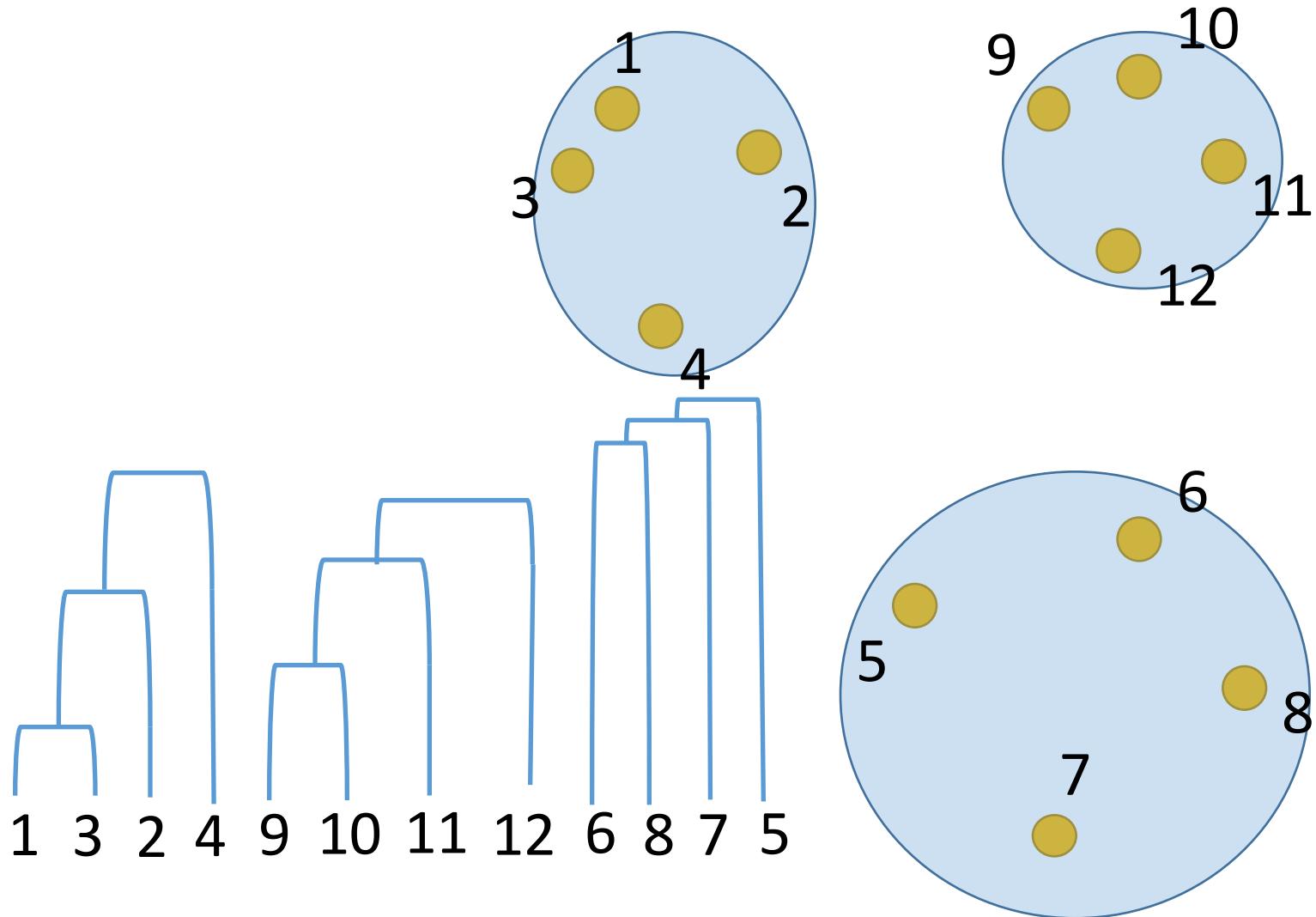
Дендрограмма



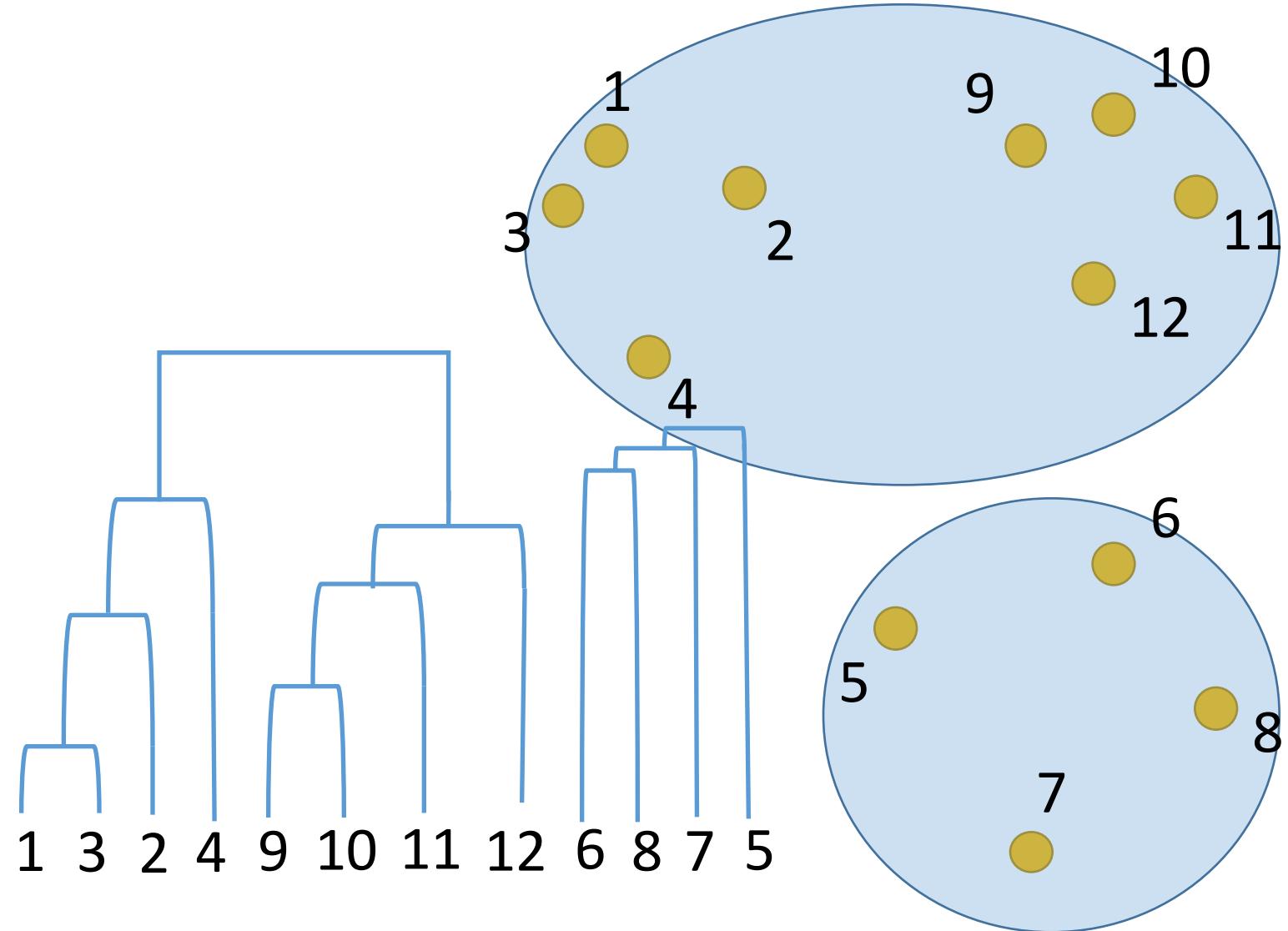
Дендрограмма



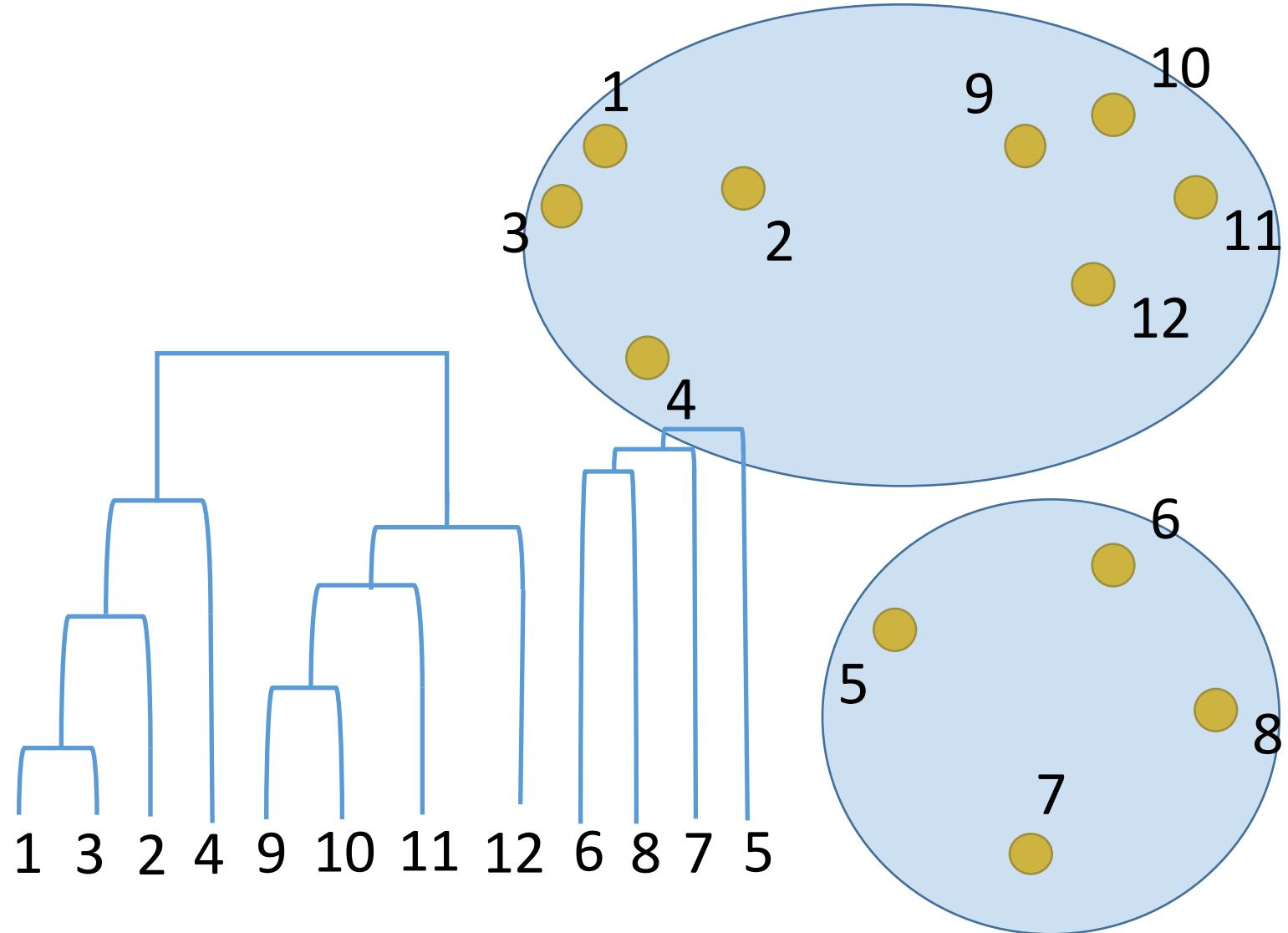
Дендрограмма



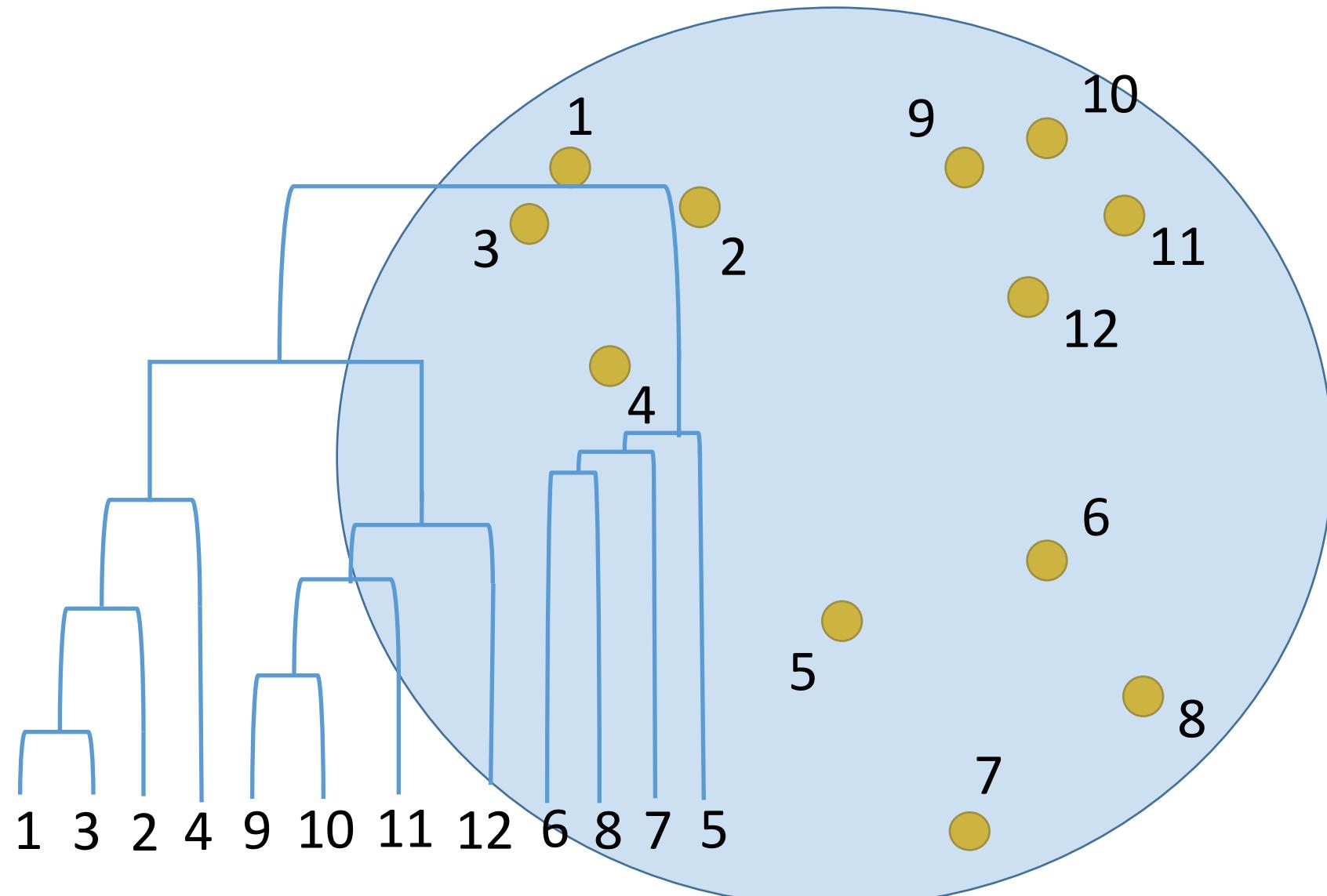
Дендрограмма



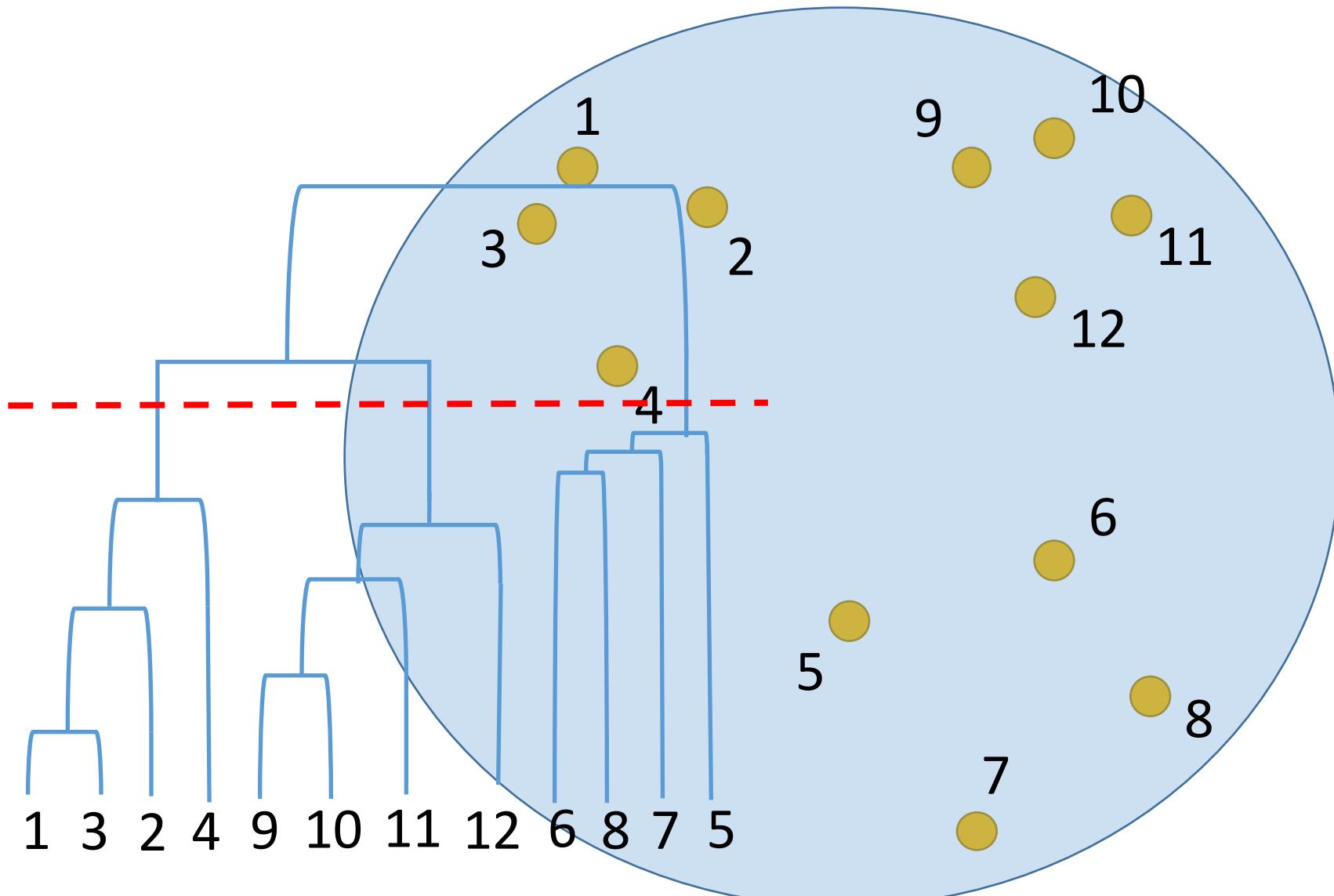
Дендрограмма



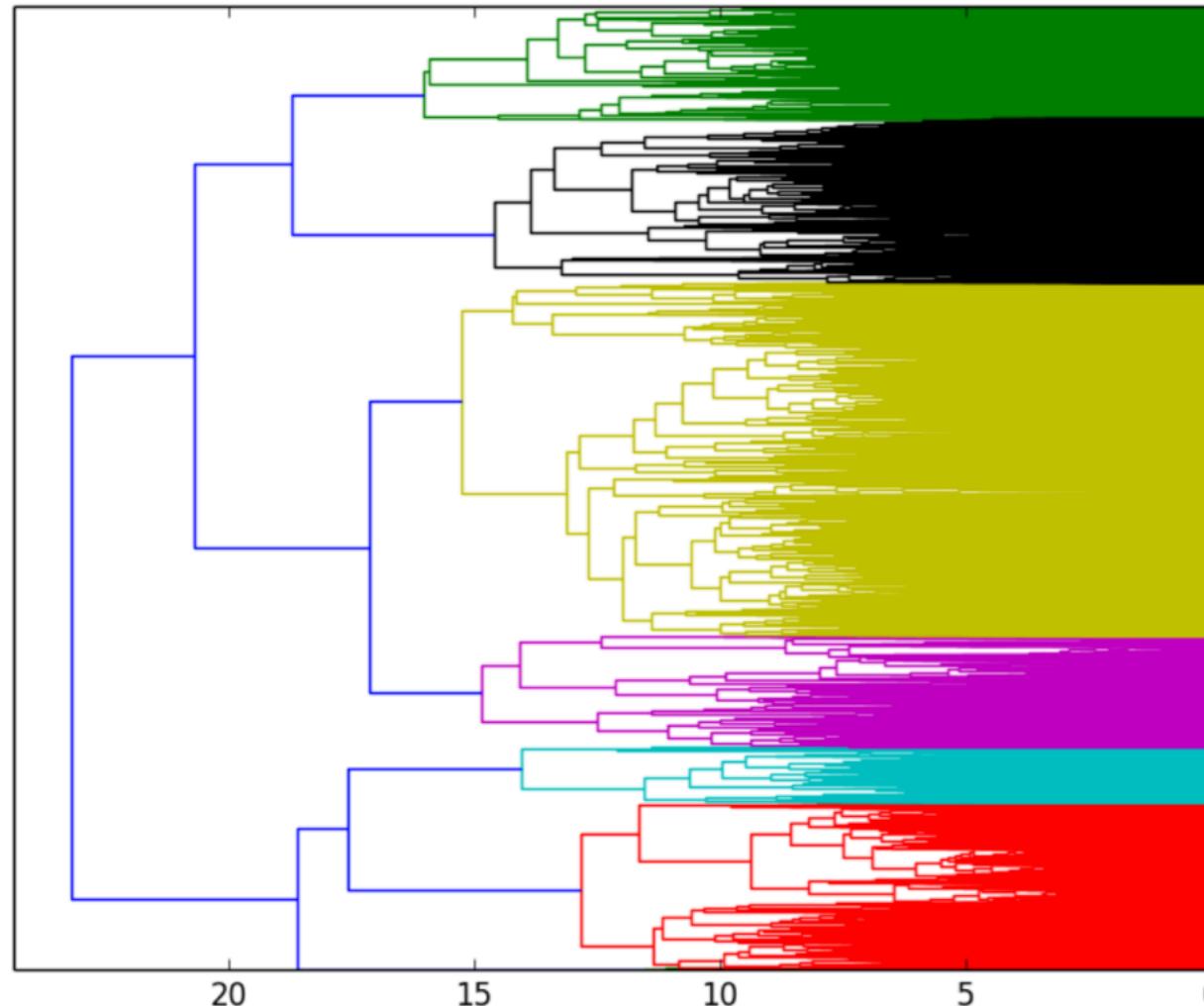
Дендрограмма



Дендрограмма

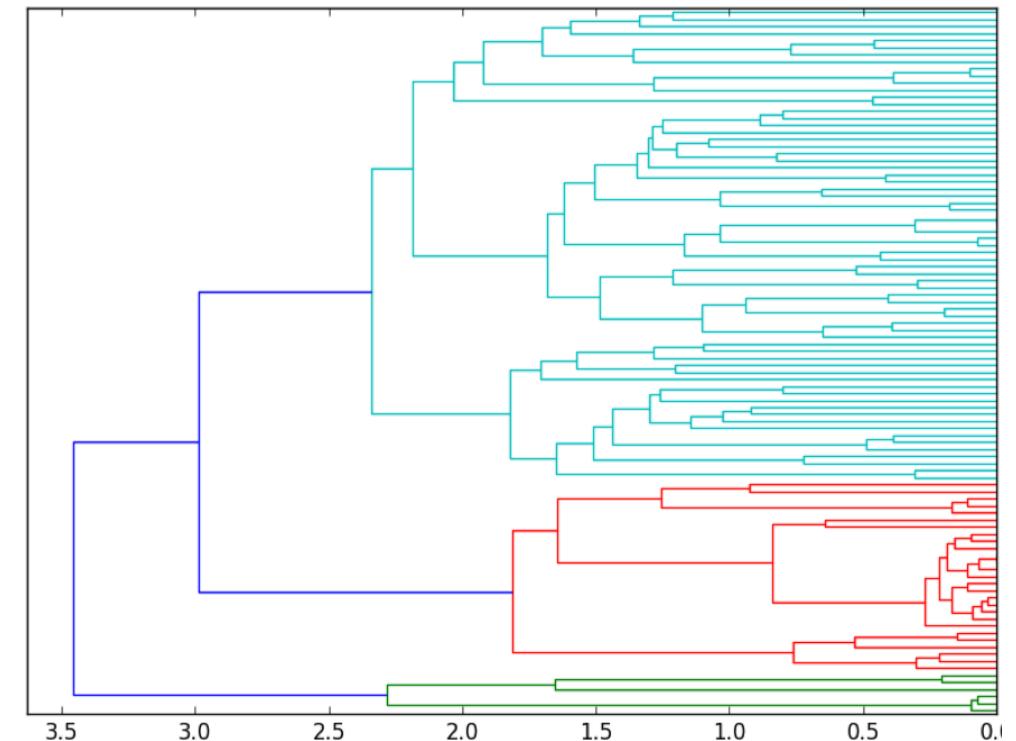
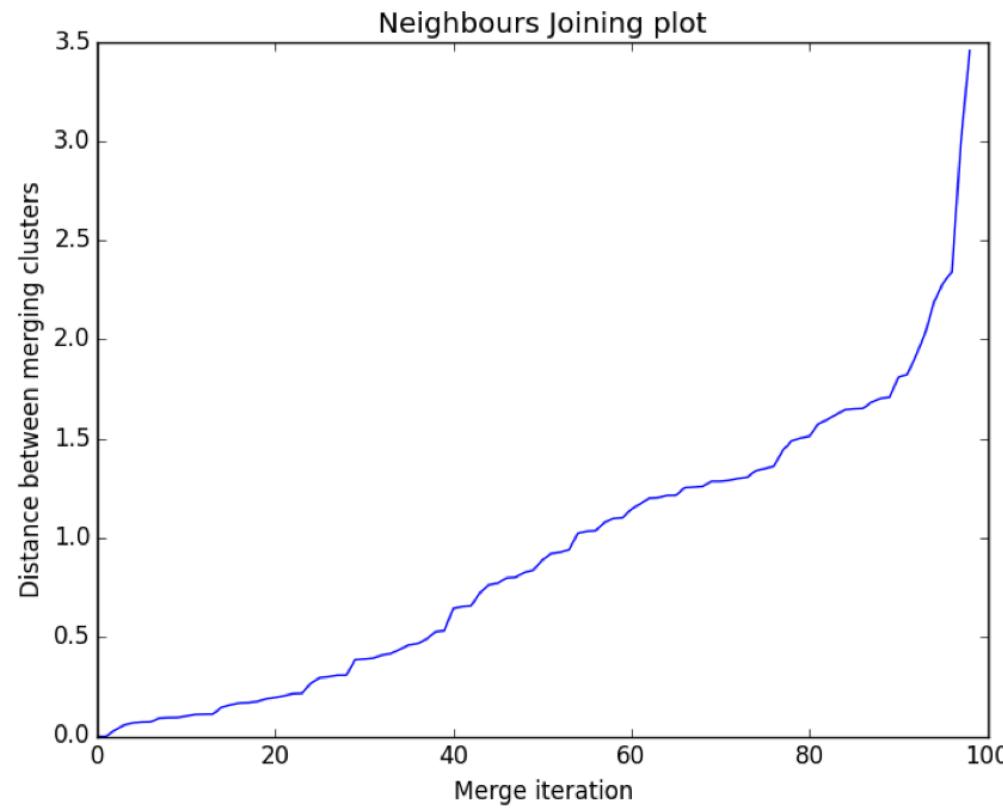


Пример: кластеризация писем



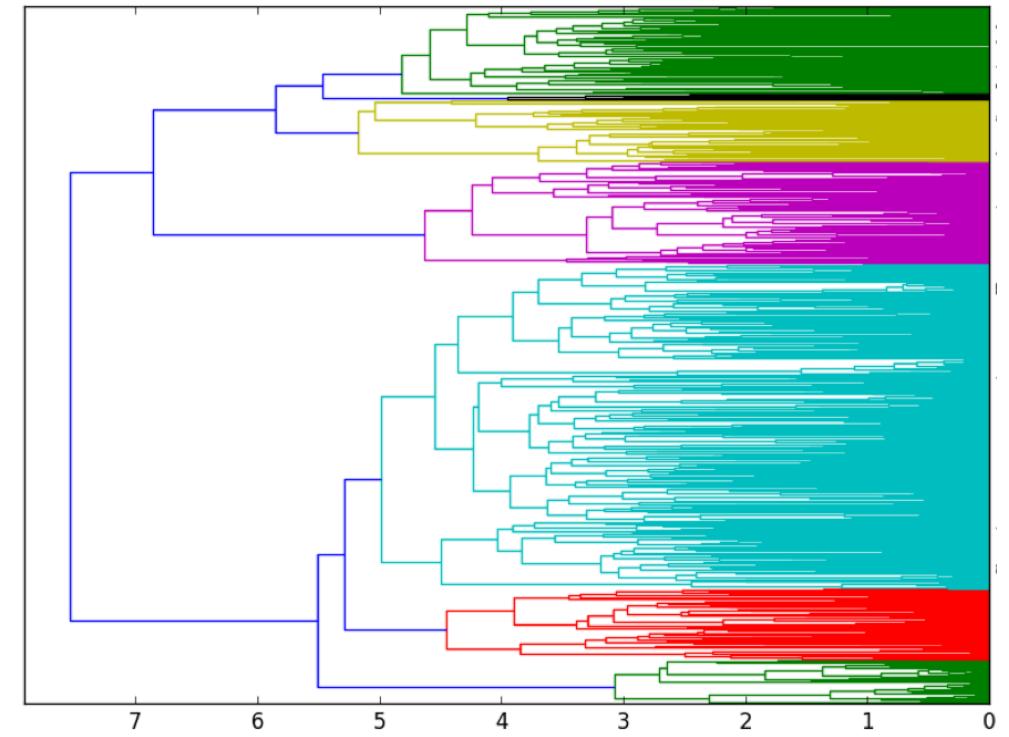
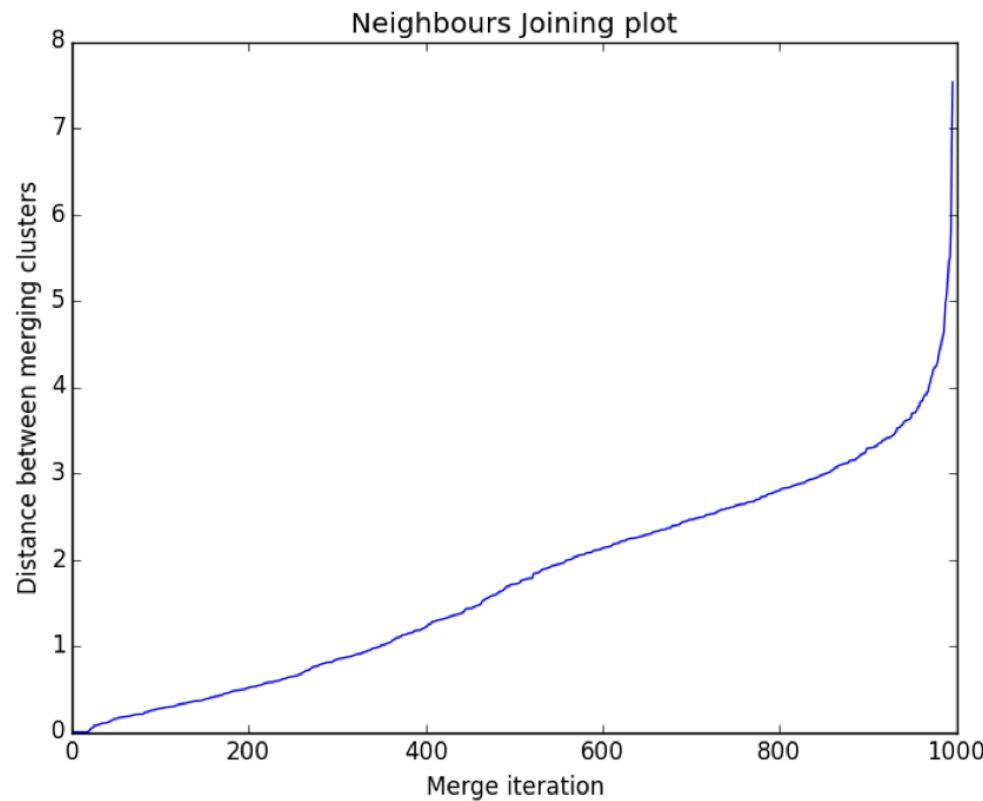
Пример: расстояние между кластерами

- На подвыборке из 100 писем



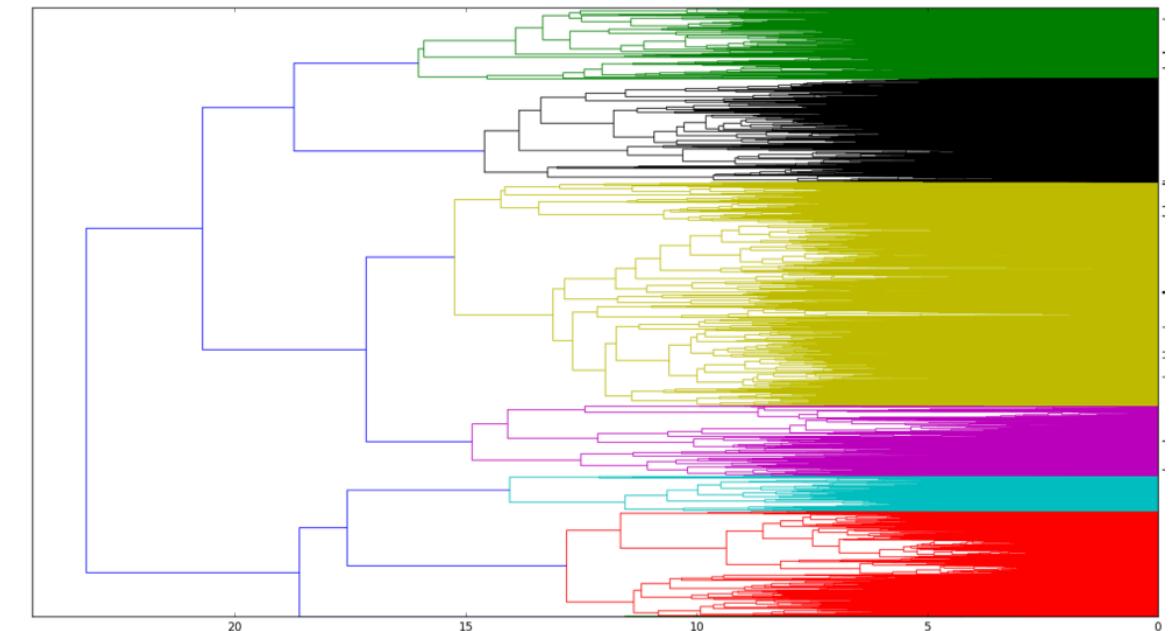
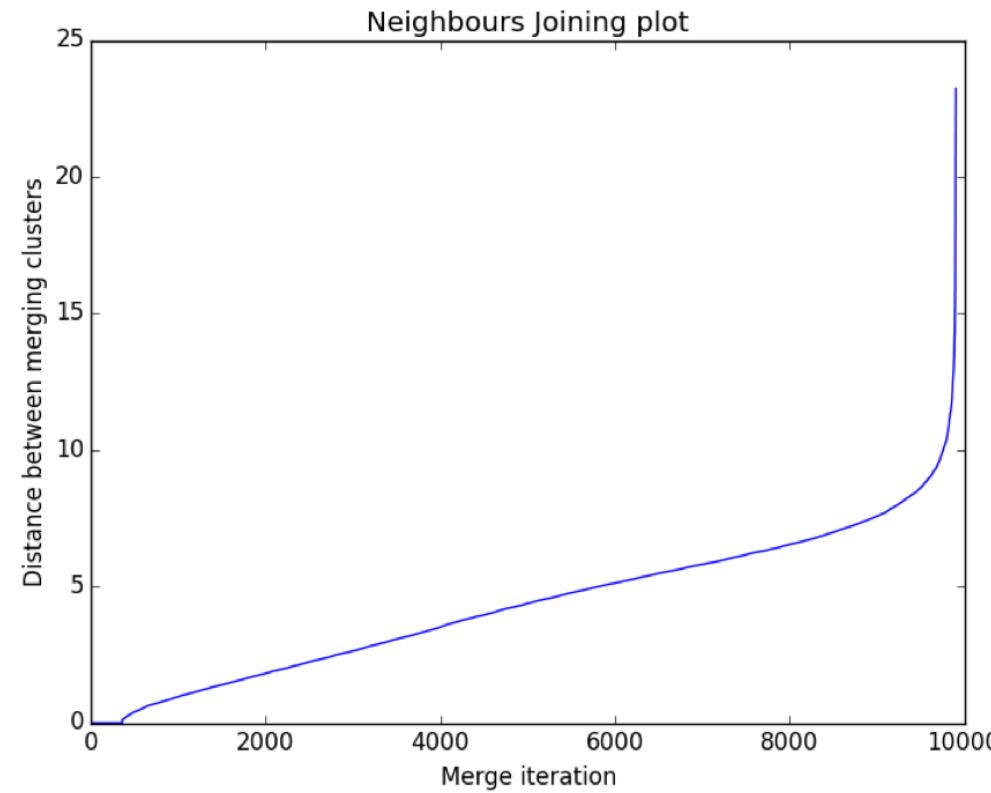
Пример: расстояние между кластерами

- На подвыборке из 1000 писем



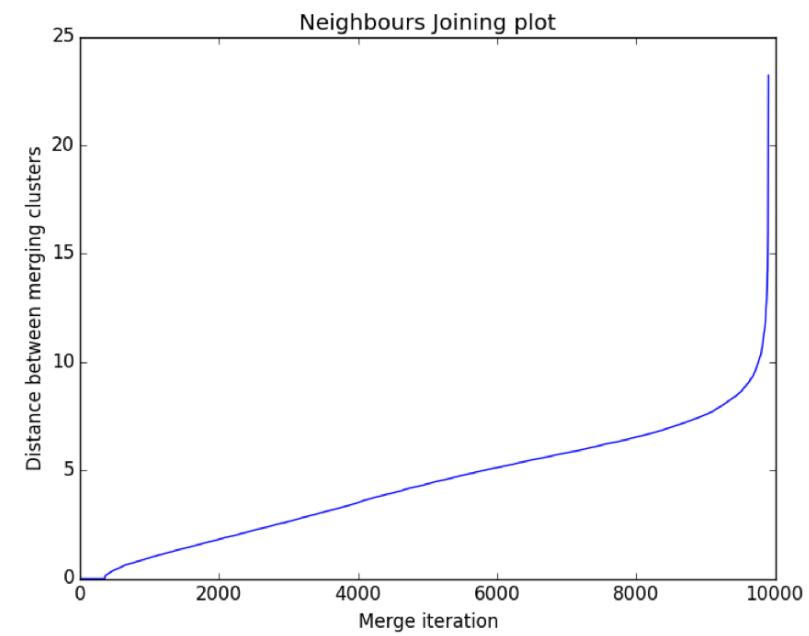
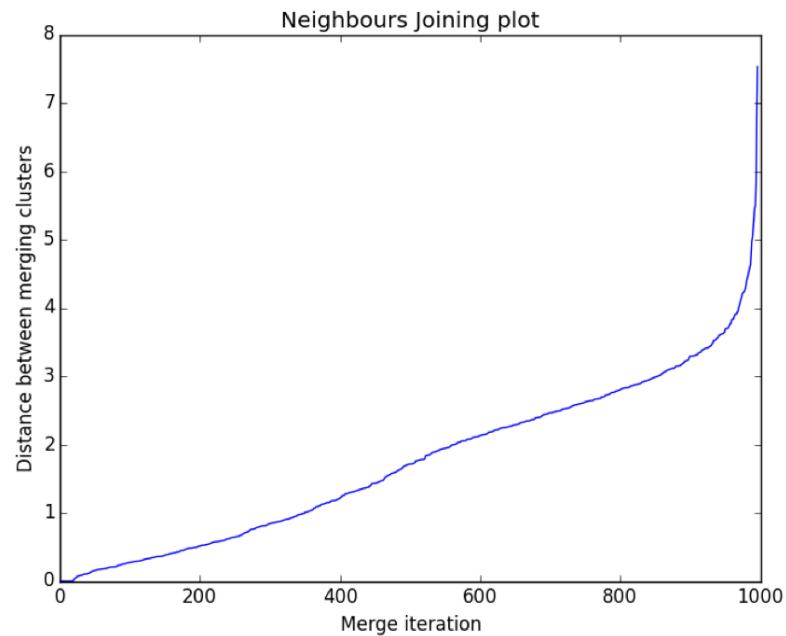
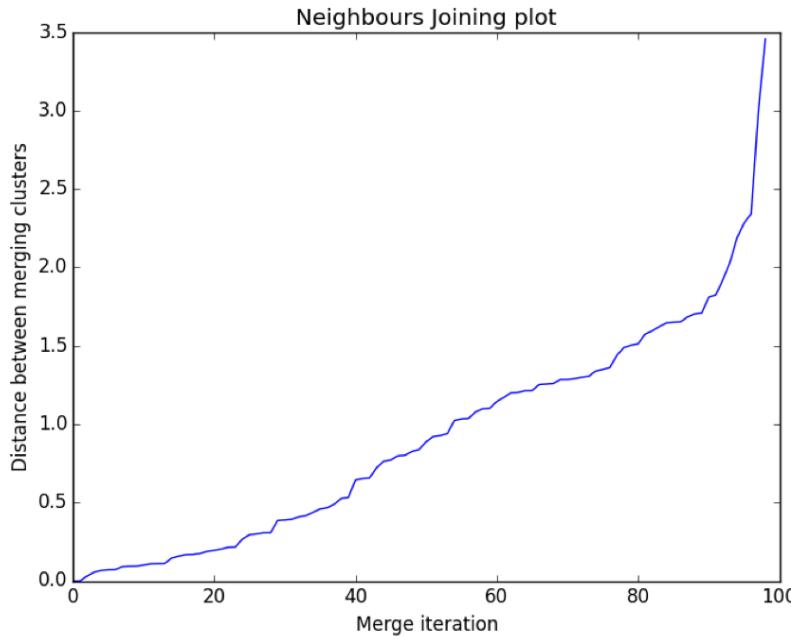
Пример: расстояние между кластерами

- На подвыборке из 10000 писем



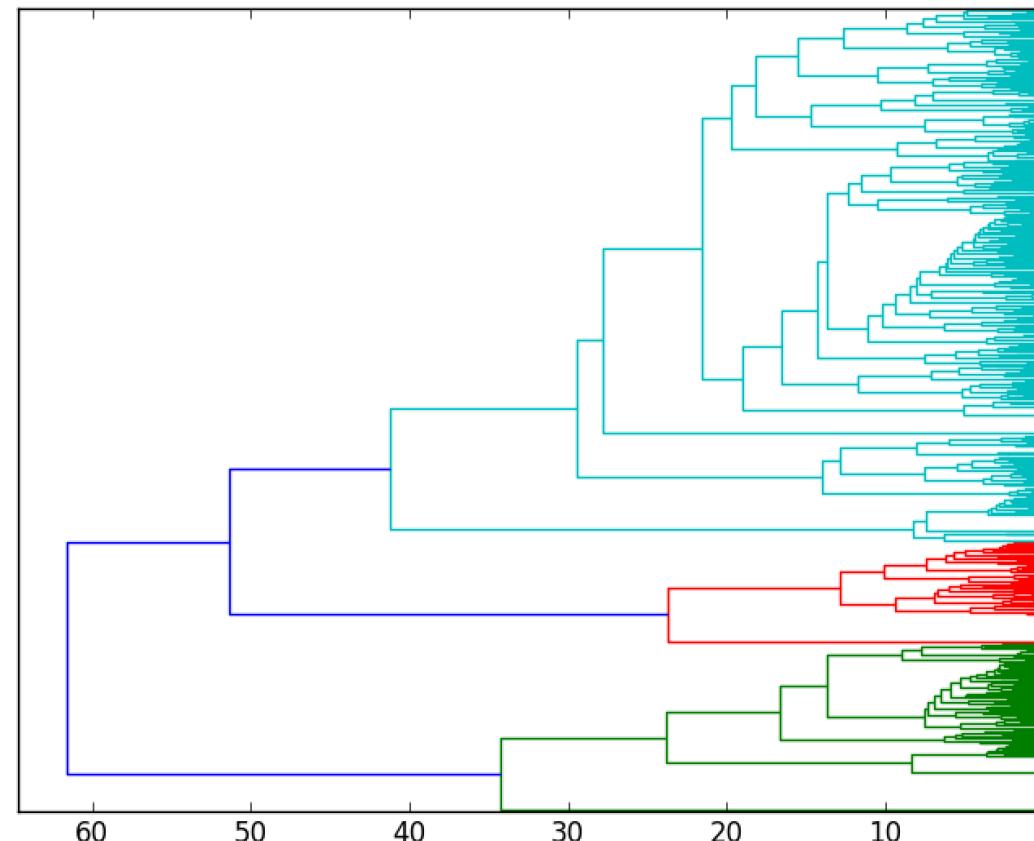
Пример: расстояние между кластерами

- Сравним графики: 100, 1000, 10000 писем

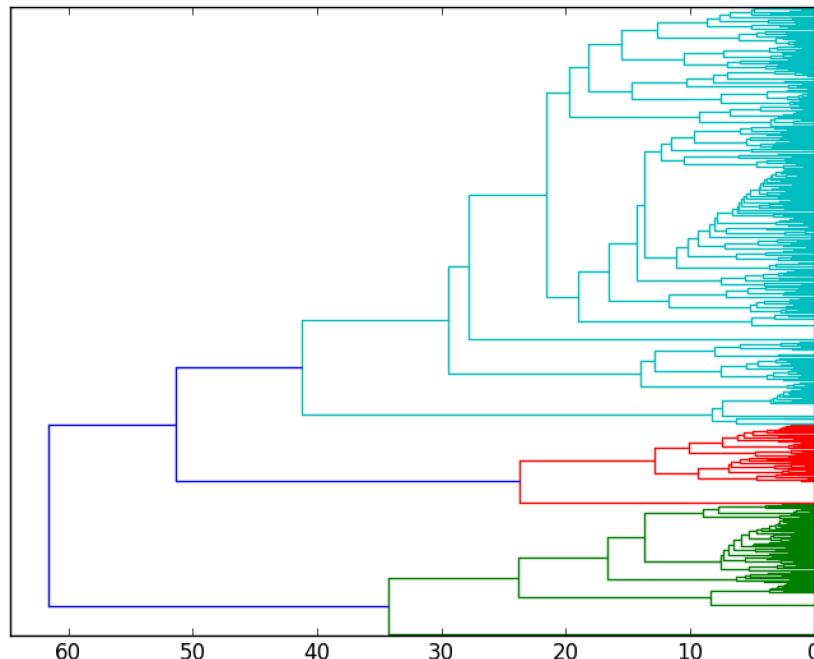


Пример: перекос в размерах кластеров

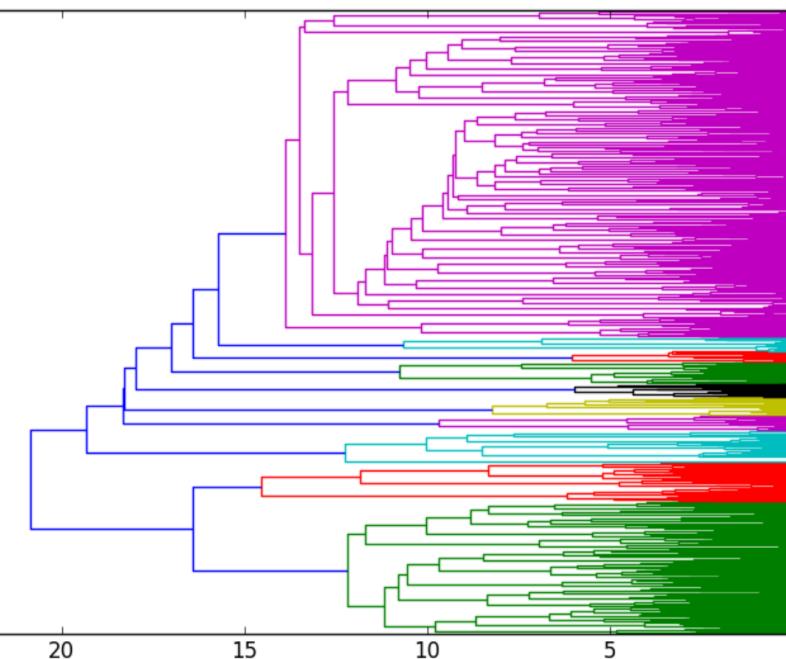
- Дендрограмма, построенная для другой выборки текстов:



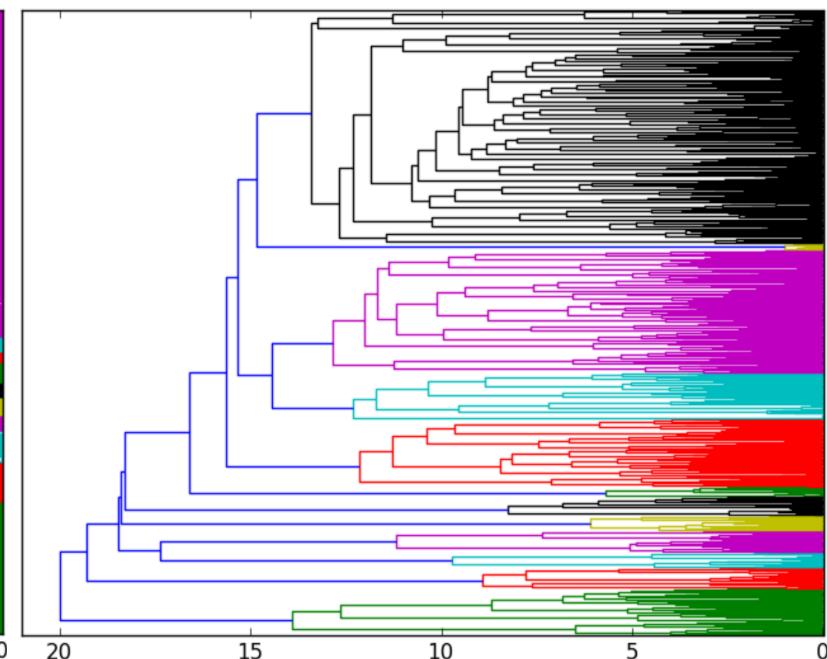
Пример: добавляем SVD



Исходные признаки

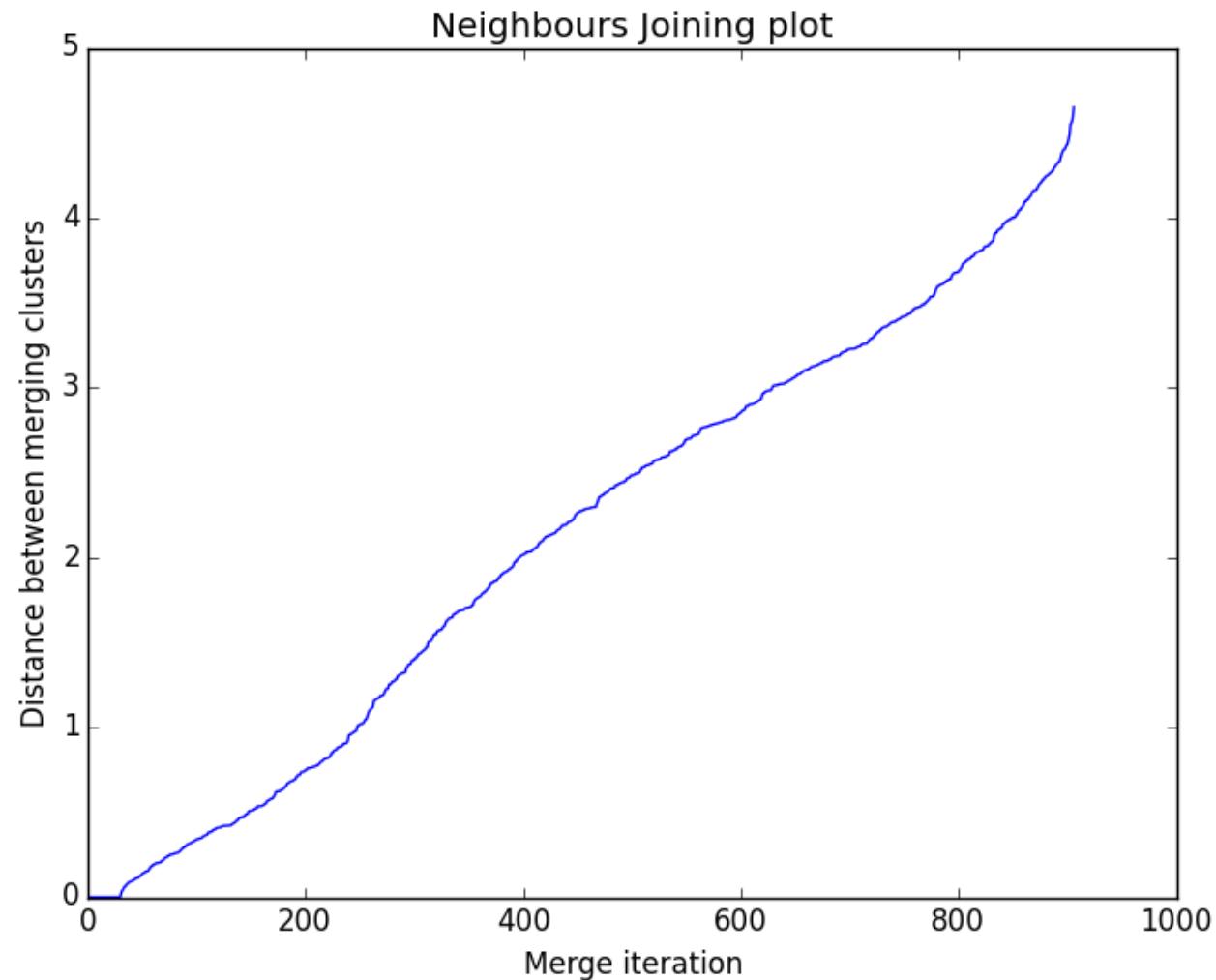


SVD



SVD (еще меньше компонент)

Пример: SVD и расстояние при слиянии



Резюме

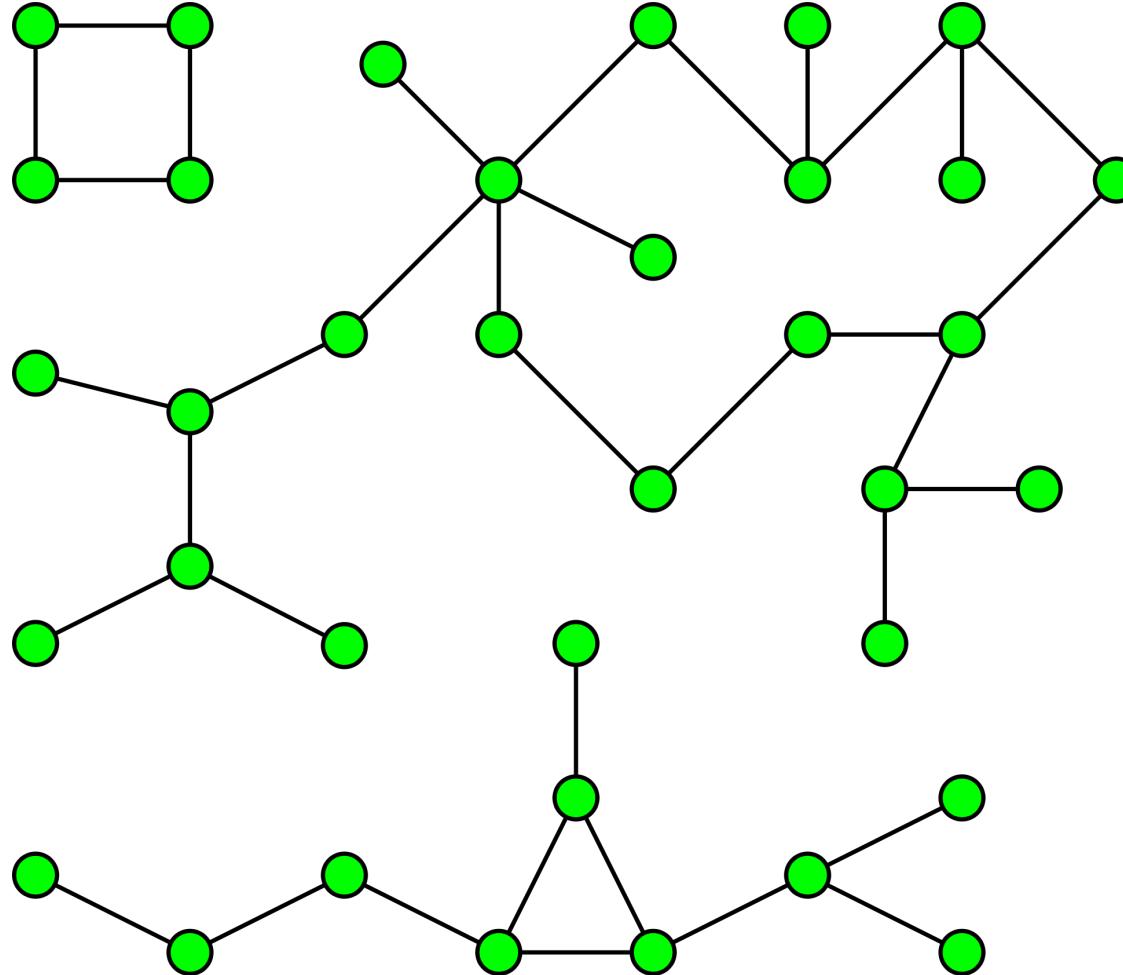
1. Иерархическая кластеризация
2. Как устроена агломеративная кластеризация
3. Расстояние между кластерами
4. Формула Ланса-Уильямса
5. Дендрограммы
6. Примеры работы

6. Простые графовые методы

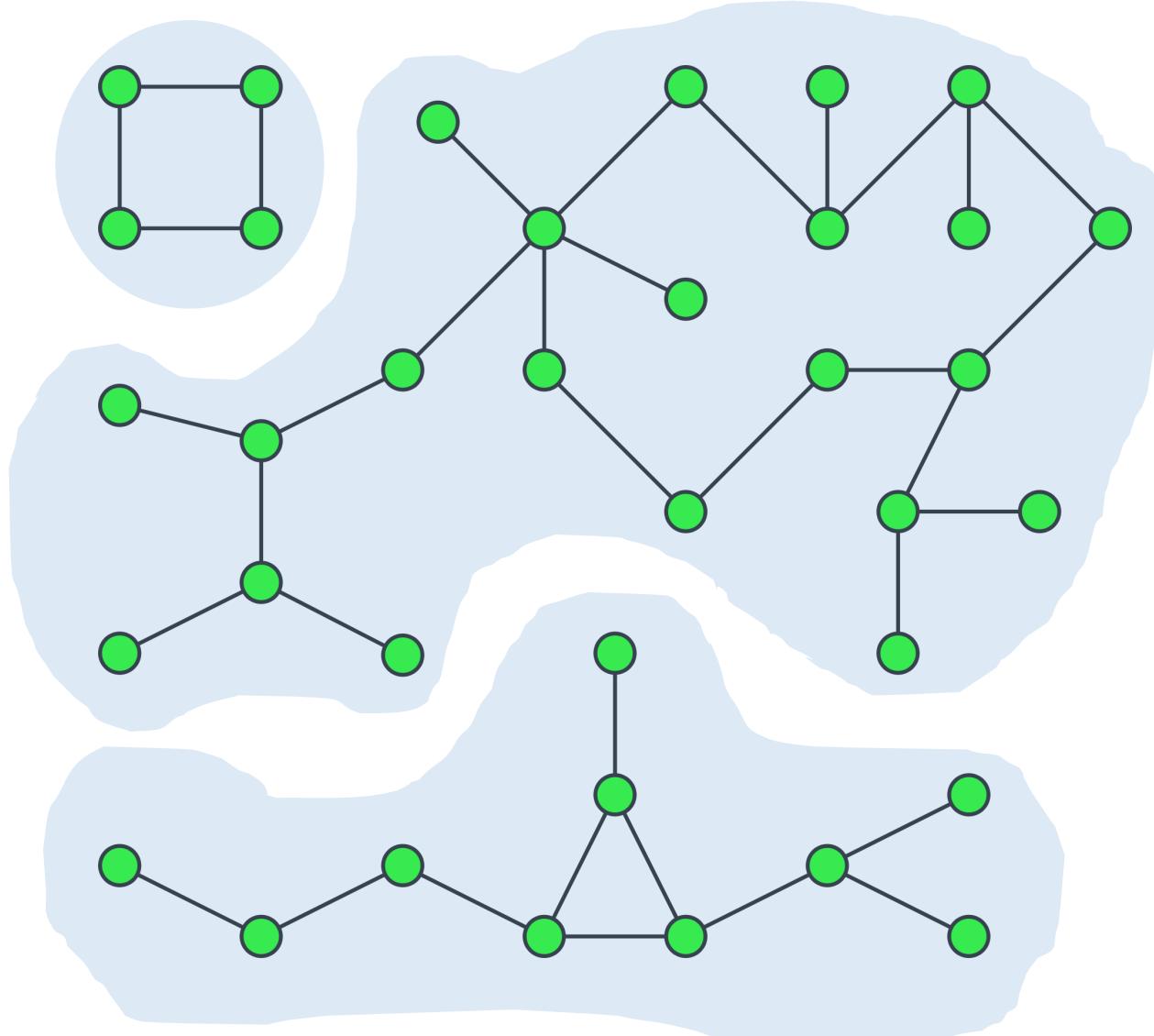
План

1. Связные компоненты
2. Кластеризация с помощью выделения связных компонент
3. Минимальное остовное дерево
4. Алгоритм Крускала
5. Кластеризация с помощью минимального остовного дерева

Выделение связных компонент



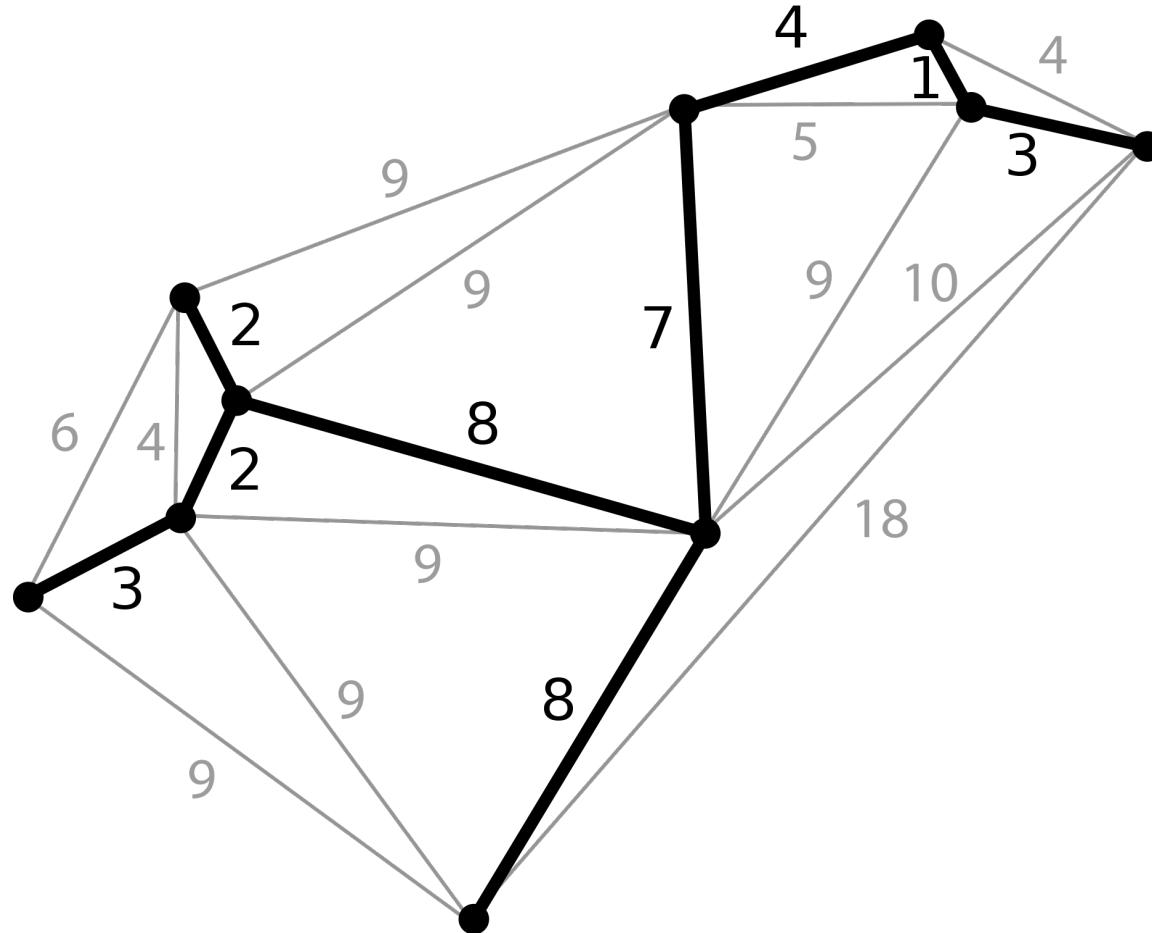
Выделение связных компонент



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



Минимальное оставное дерево

Алгоритм Крускала (Kruskal):

1. Изначально множество уже найденных ребер пустое
2. На первом шаге добавляем ребро с минимальным весом
3. На каждом шаге добавляем ребро, одна из вершина которого лежит в множестве выбранных вершин, а другая – нет, при этом среди всех таких ребер выбираем ребро с наименьшим весом
4. В тот момент, когда задействованы все вершины графа – выбранные ребра образуют минимальное оставное 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

Резюме

1. Связные компоненты
2. Кластеризация с помощью выделения связных компонент
3. Минимальное остовное дерево
4. Алгоритм Крускала
5. Кластеризация с помощью минимального остовного дерева

7. Кластеризация на основе плотности точек (density based clustering)

План

1. Идея методов на основе плотности точек
2. Пример основных, граничных и шумовых точек
3. DBSCAN
4. Пример работы DBSCAN
5. Определение числа кластеров
6. Настройка параметров DBSCAN

Идея density-based методов

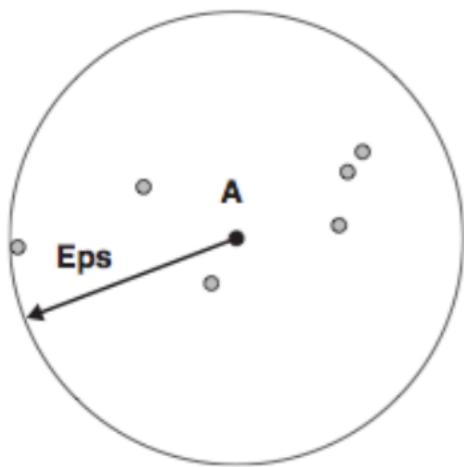


Figure 8.20. Center-based density.

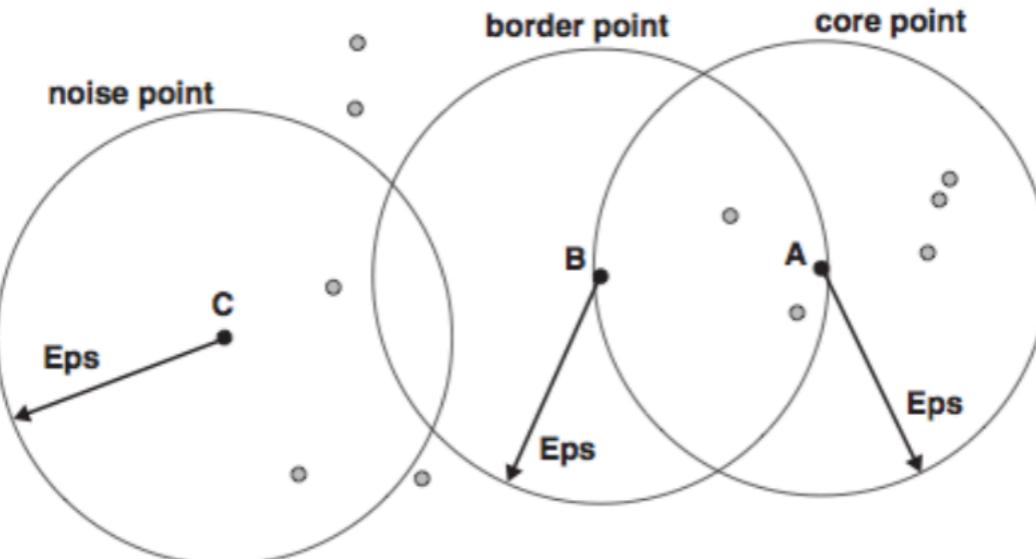
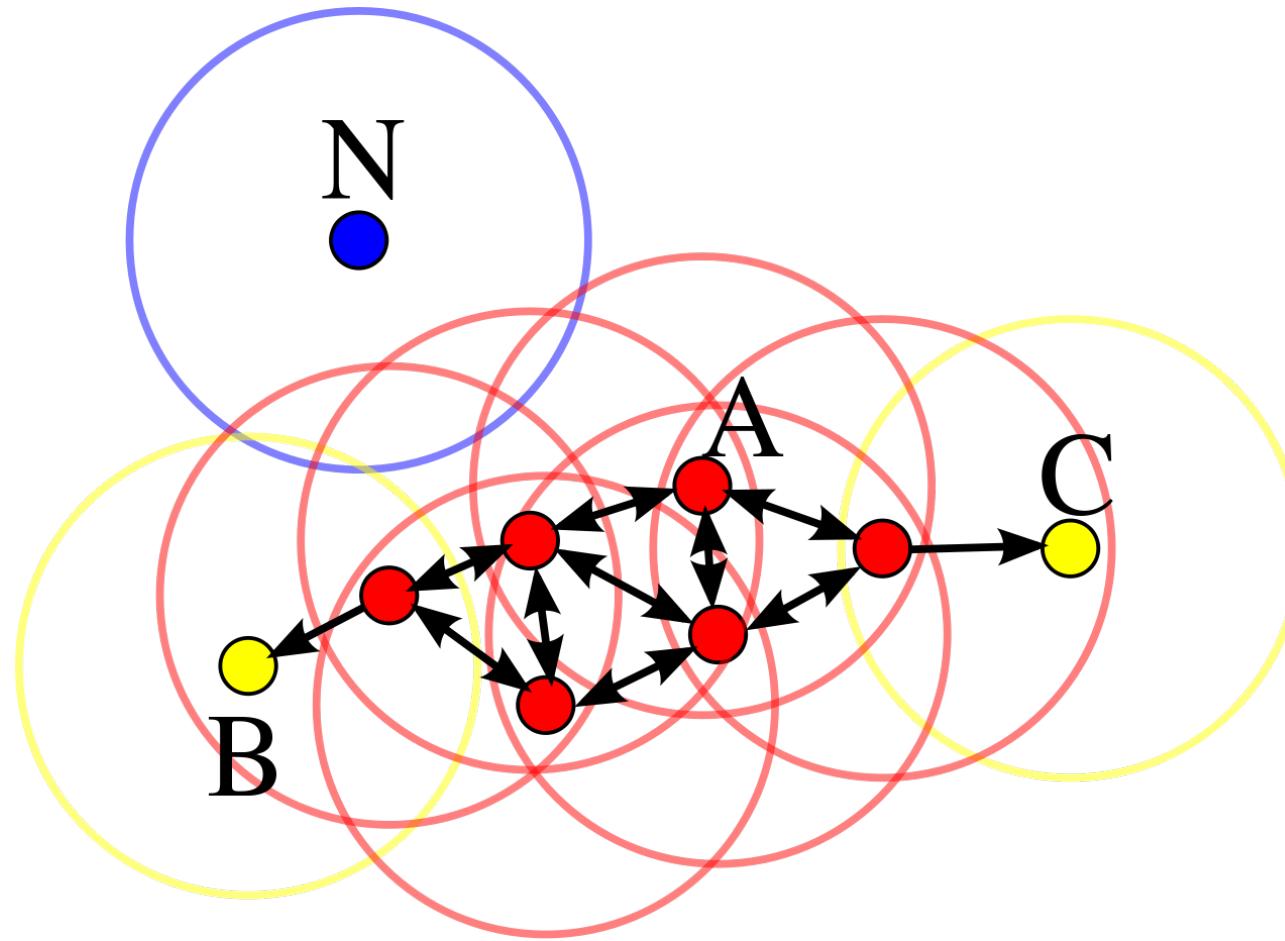


Figure 8.21. Core, border, and noise points.

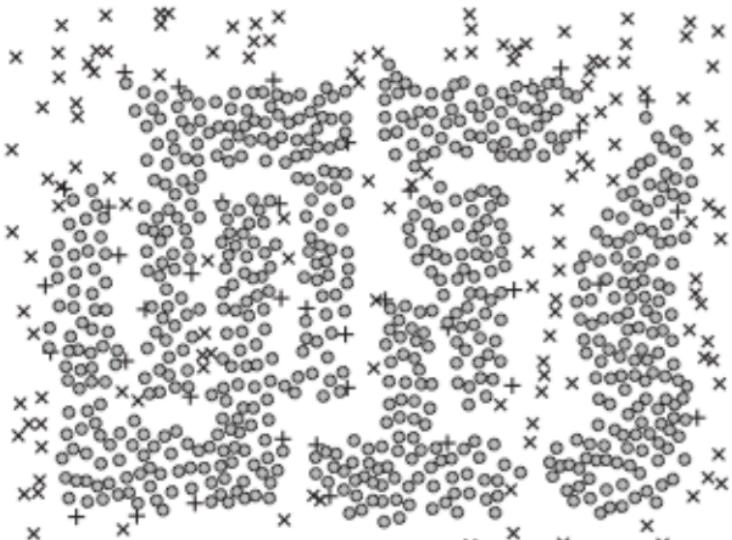
Основные, шумовые и граничные точки



DBSCAN



(a) Clusters found by DBSCAN.



x – Noise Point + – Border Point o – Core Point

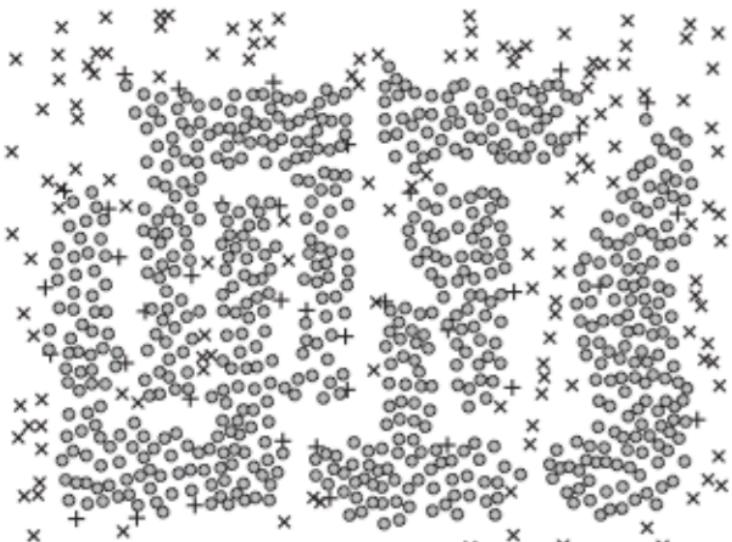
(b) Core, border, and noise points.

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

DBSCAN



(a) Clusters found by DBSCAN.



x – Noise Point + – Border Point o – Core Point

(b) Core, border, and noise points.

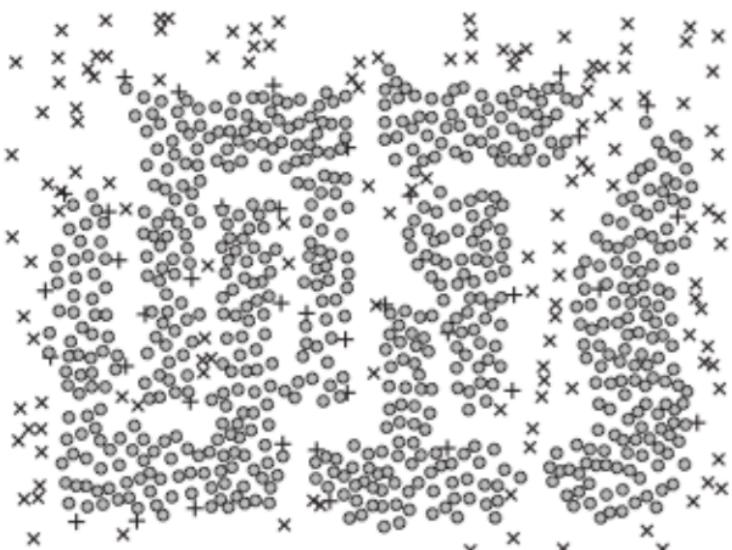
1: Пометить все точки, как основные, пограничные или шумовые.

2: Отбросить точки шума.

DBSCAN



(a) Clusters found by DBSCAN.



x – Noise Point + – Border Point o – Core Point

(b) Core, border, and noise points.

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

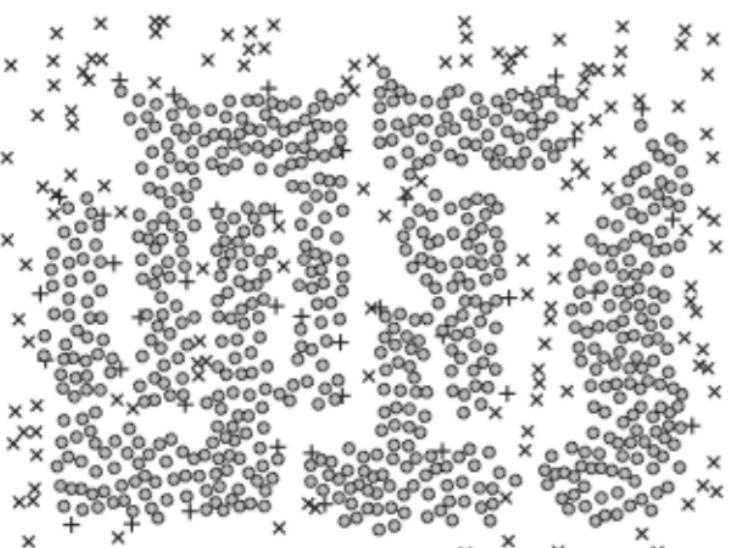
2: Отбросить точки шума.

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

DBSCAN



(a) Clusters found by DBSCAN.



x – Noise Point + – Border Point o – Core Point

(b) Core, border, and noise points.

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

2: Отбросить точки шума.

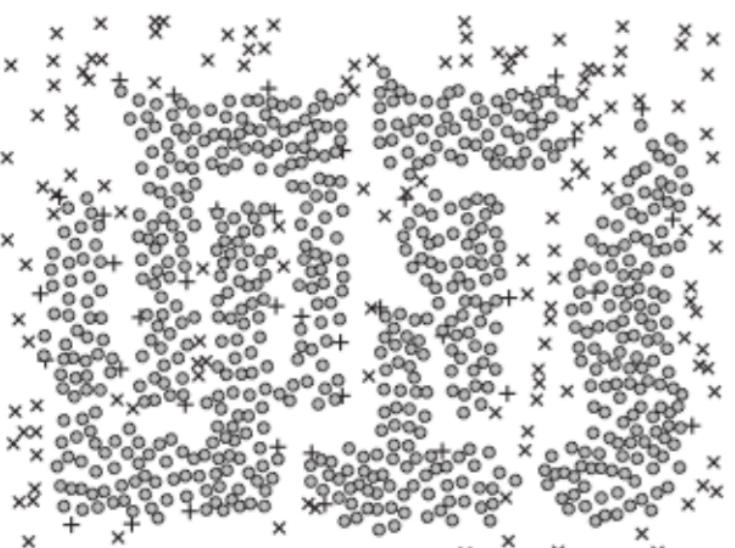
3: Соединить все основные точки, находящиеся на расстоянии Eps радиуса одна от другой.

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

DBSCAN



(a) Clusters found by DBSCAN.



x – Noise Point + – Border Point o – Core Point

(b) Core, border, and noise points.

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

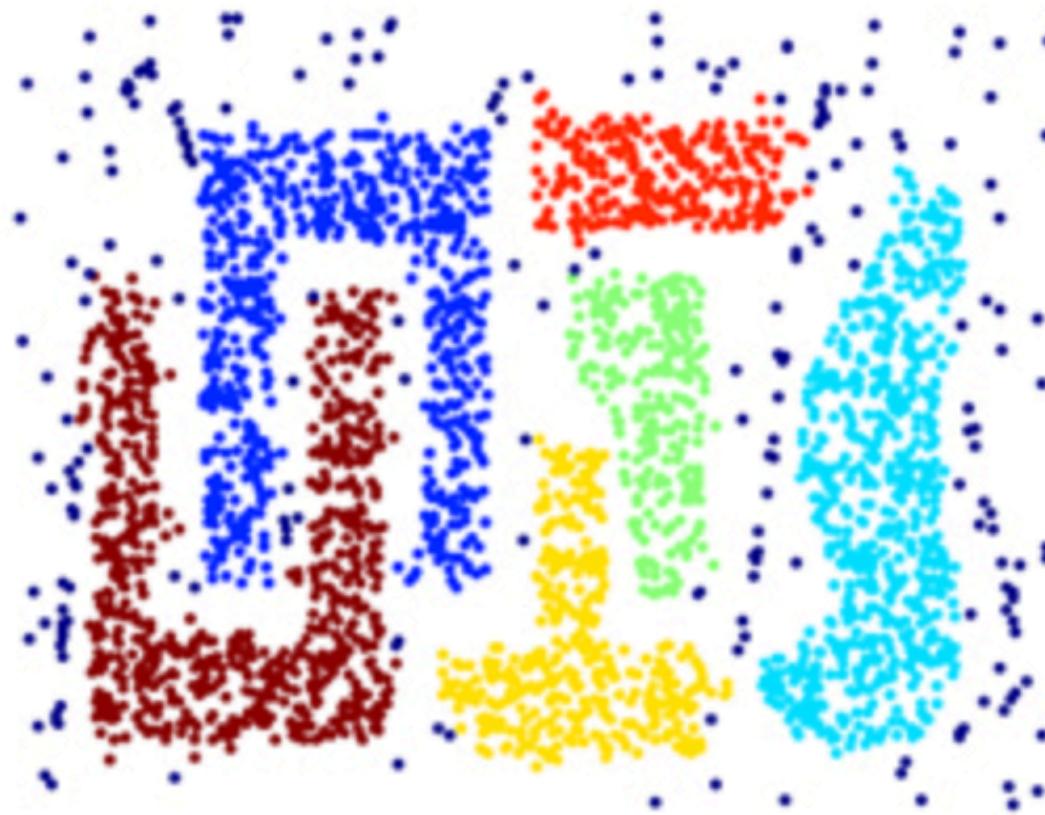
2: Отбросить точки шума.

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

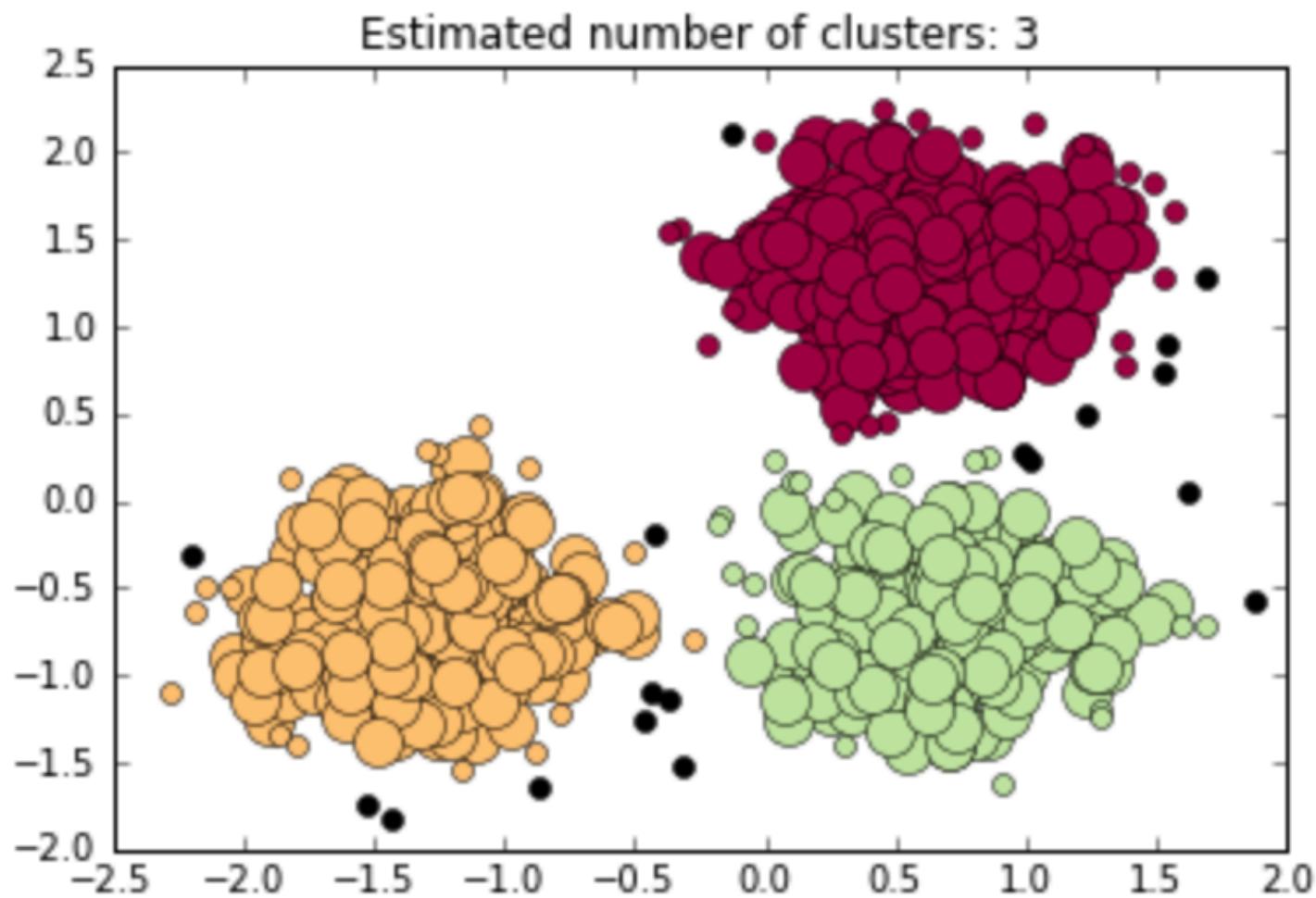
4: Объединить каждую группу соединенных основных точек в отдельный кластер.

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

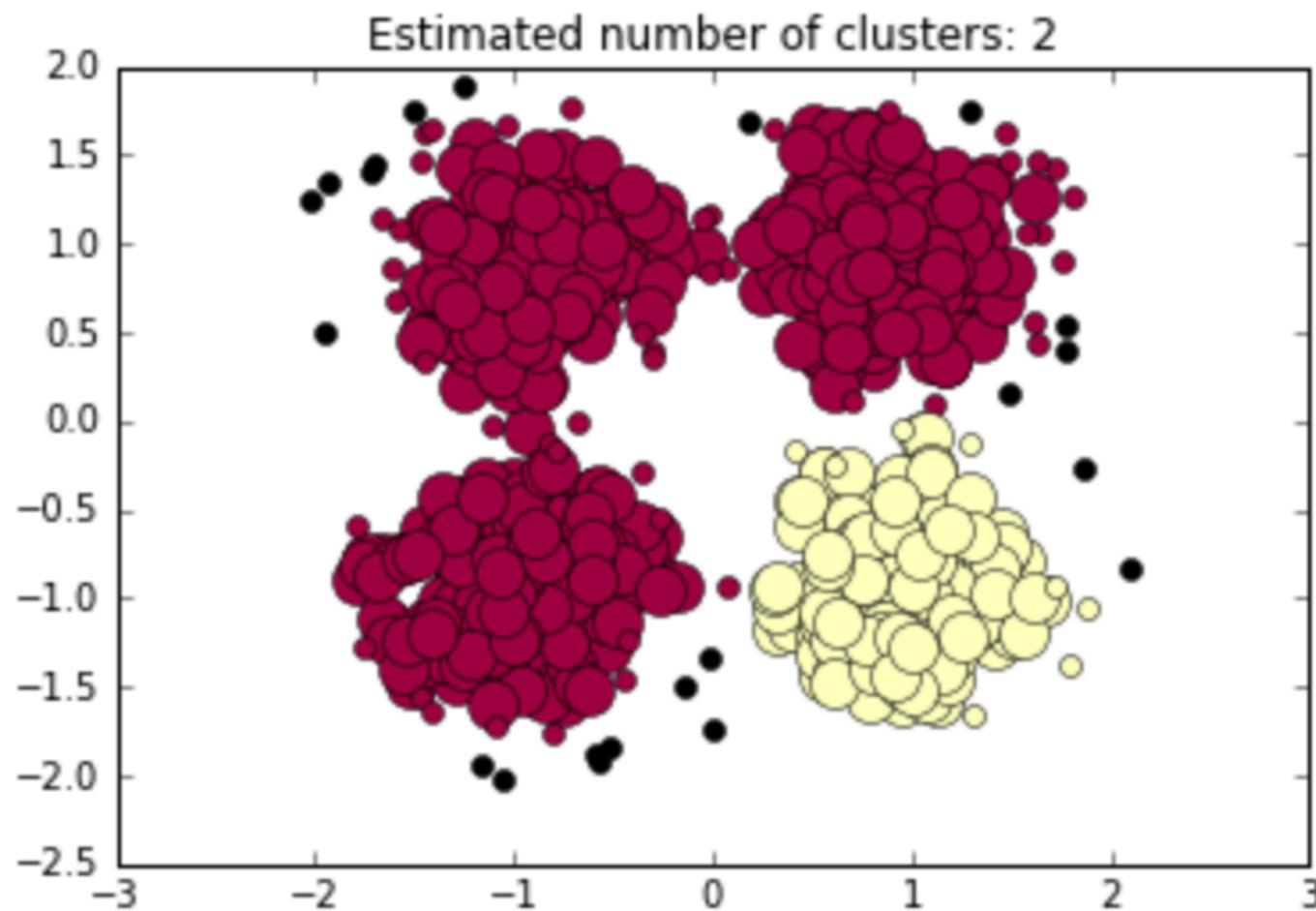
DBSCAN: результаты работы



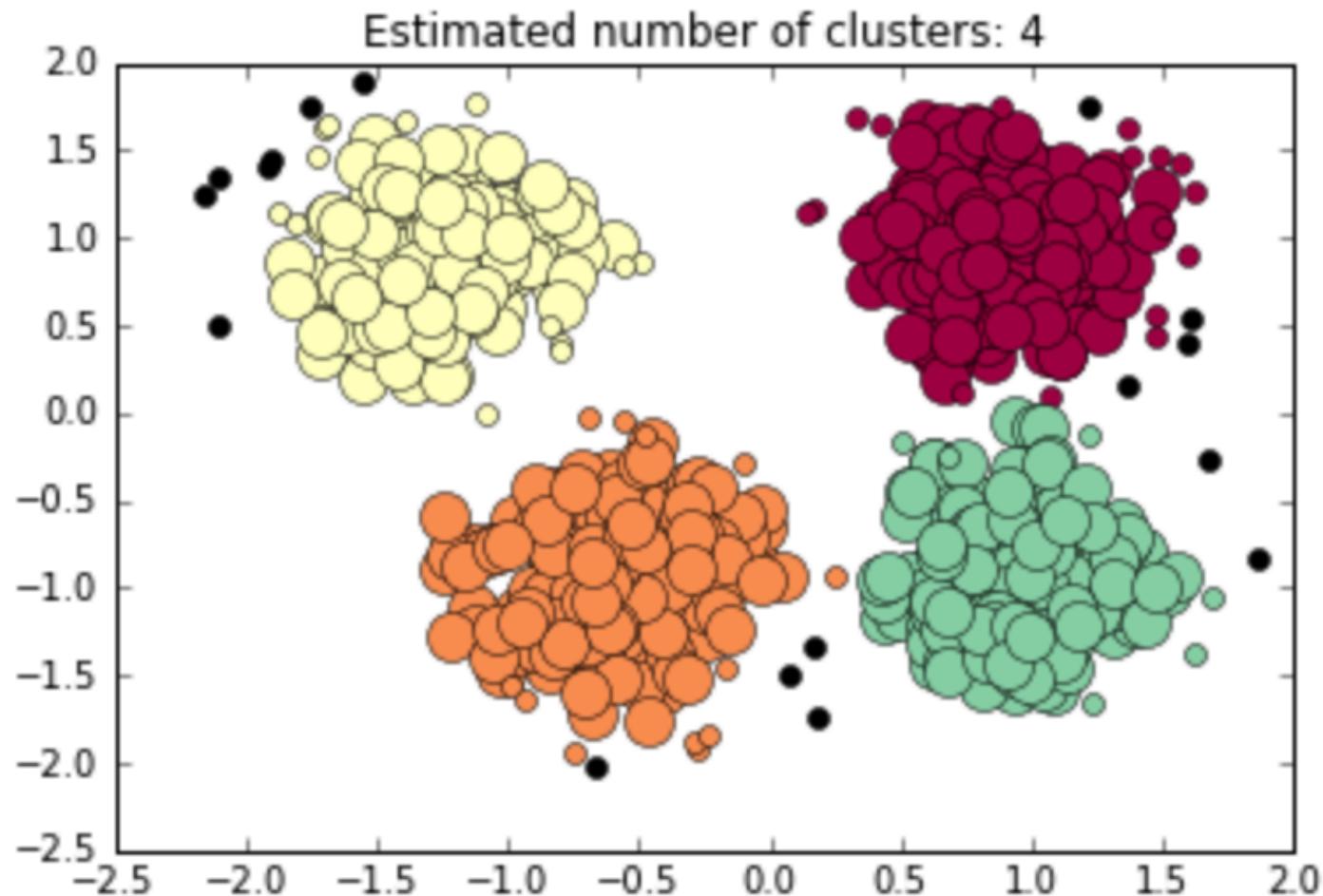
Определение числа кластеров



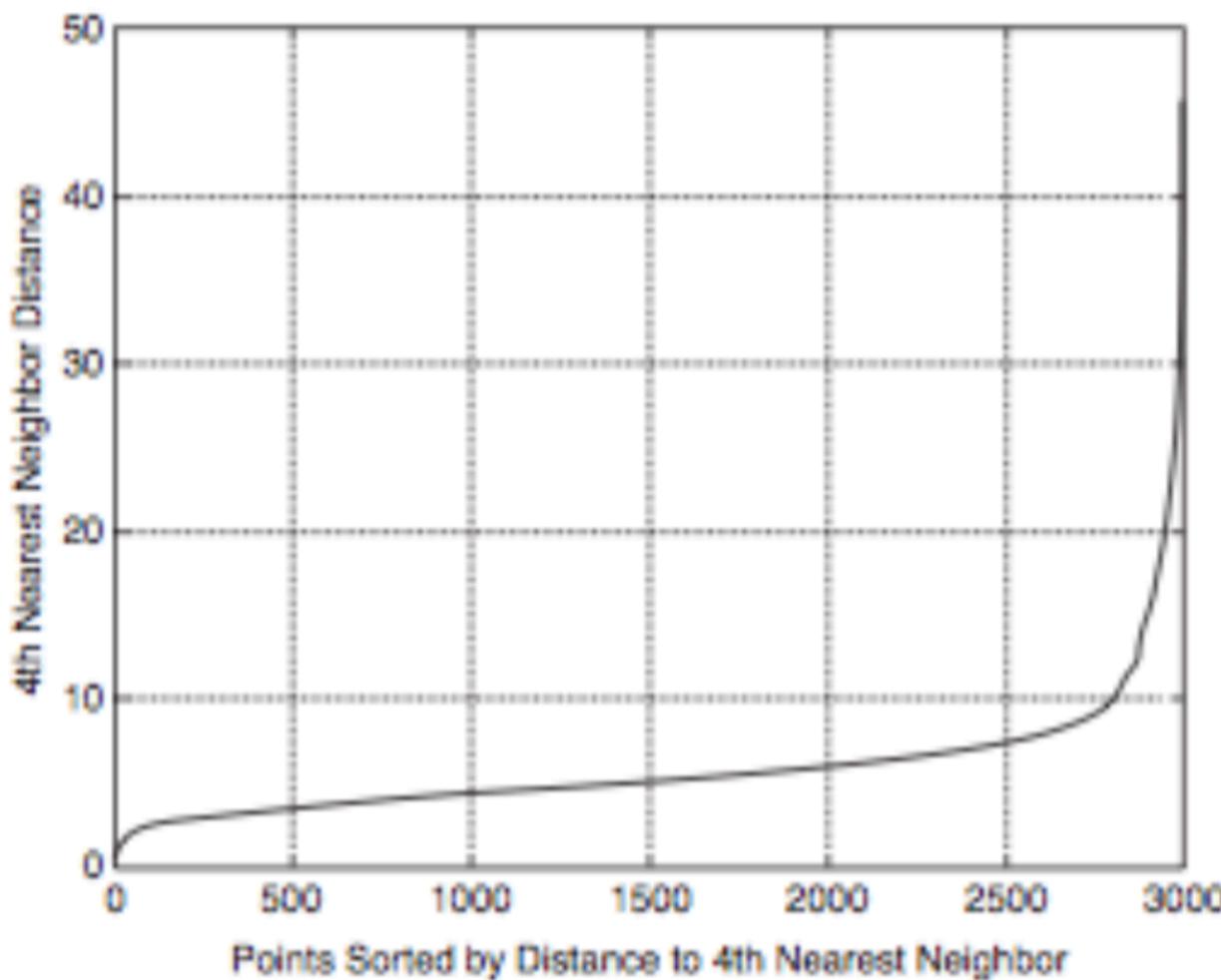
Определение числа кластеров



Определение числа кластеров



DBSCAN: подбор параметров



Резюме

1. Идея методов на основе плотности точек
2. Пример основных, граничных и шумовых точек
3. DBSCAN
4. Пример работы DBSCAN
5. Определение числа кластеров
6. Настройка параметров DBSCAN

8. Выбор метода кластеризации

Алгоритмы

Рассмотренные нами:

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

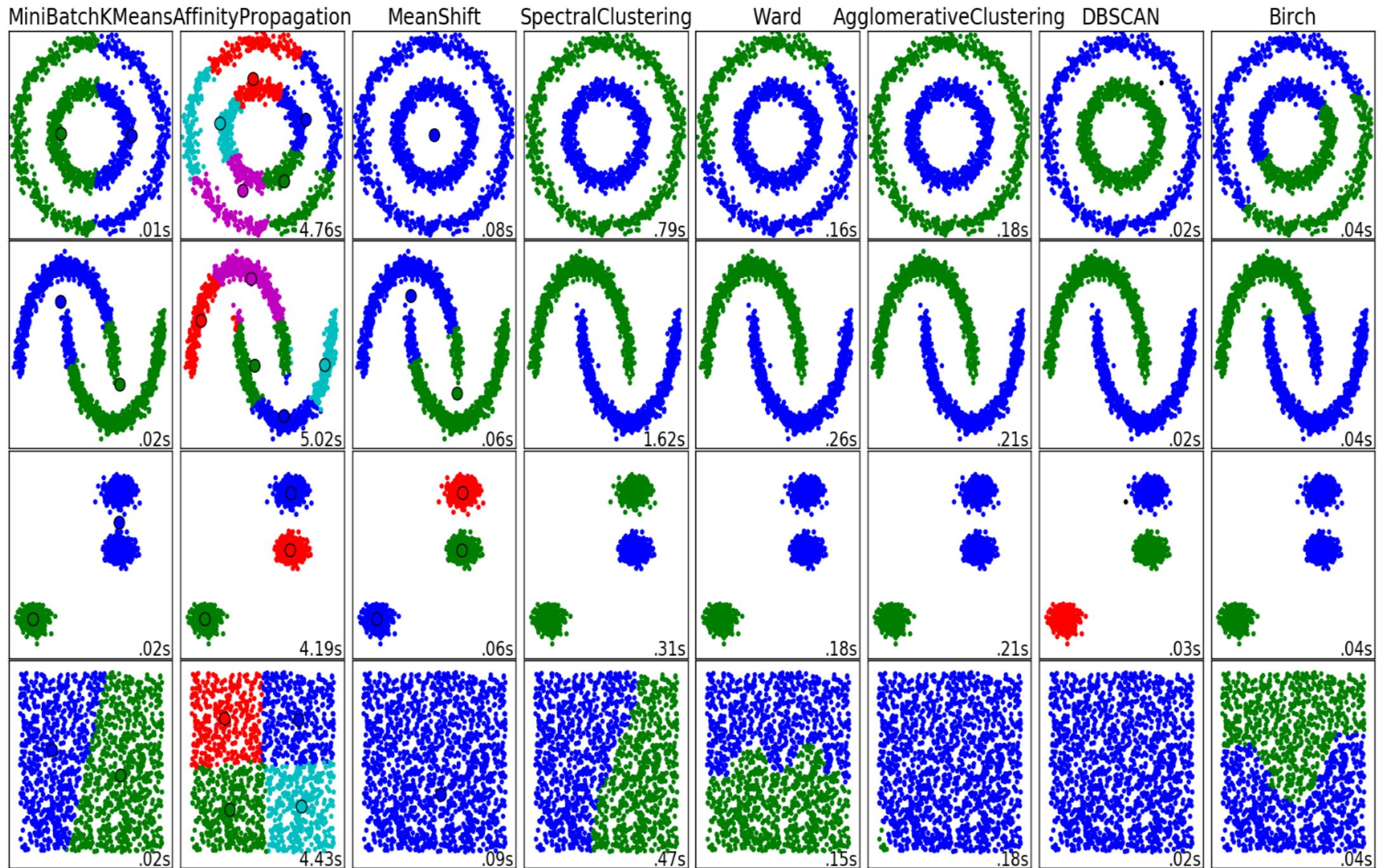
Алгоритмы

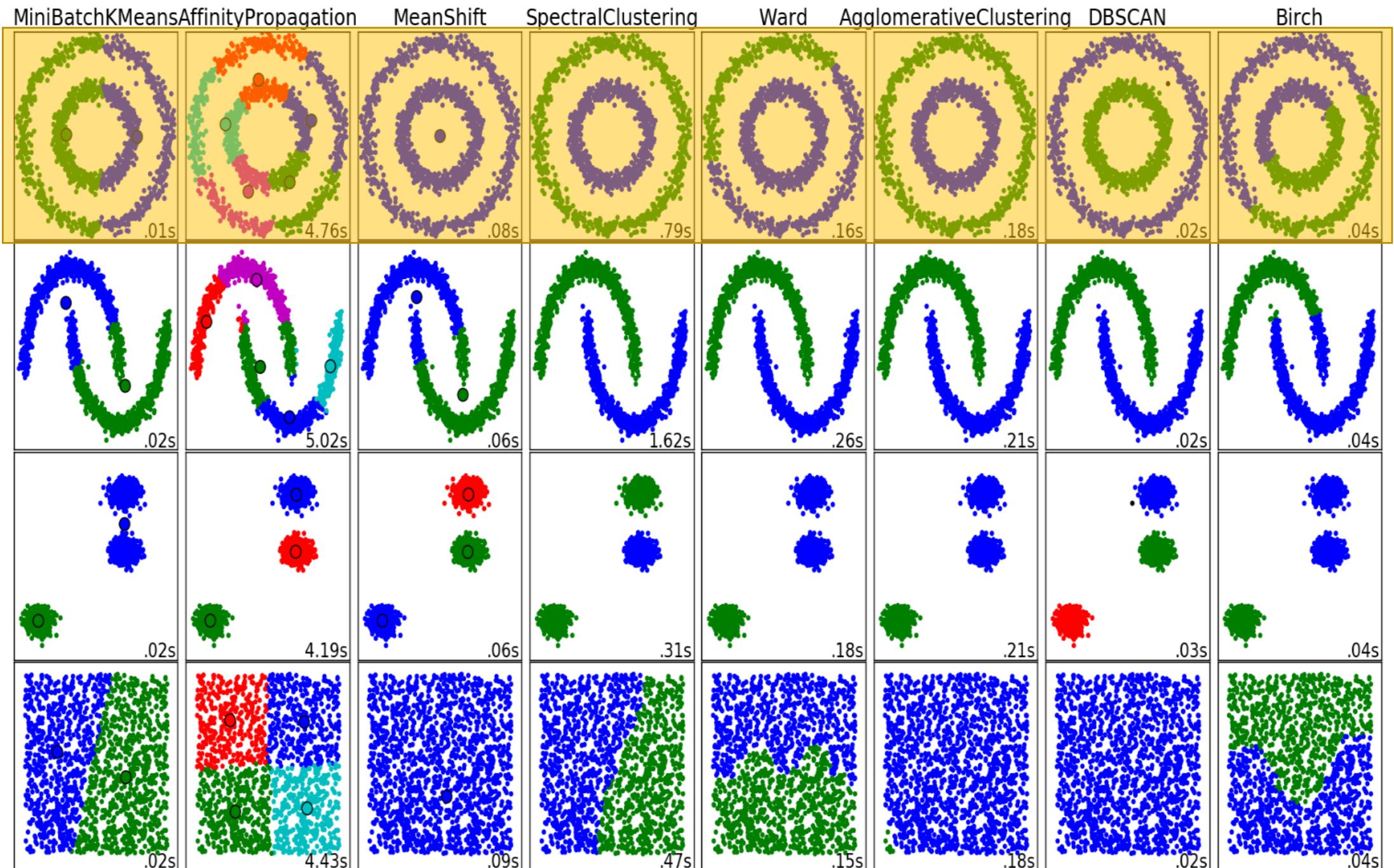
Рассмотренные нами:

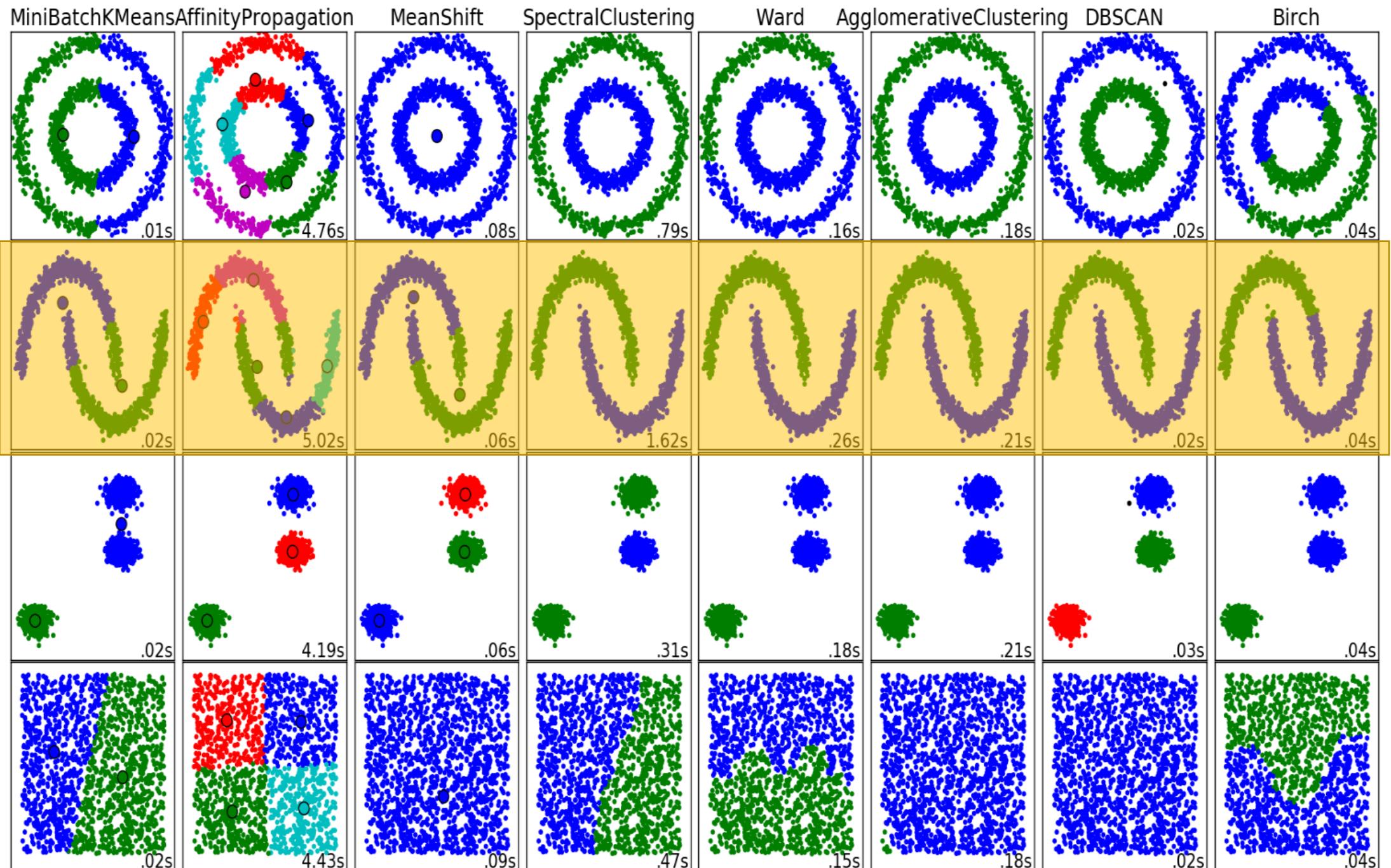
- К-средних
- ЕМ-алгоритм
- Агglomerативная иерархическая кластеризация
- DBSCAN

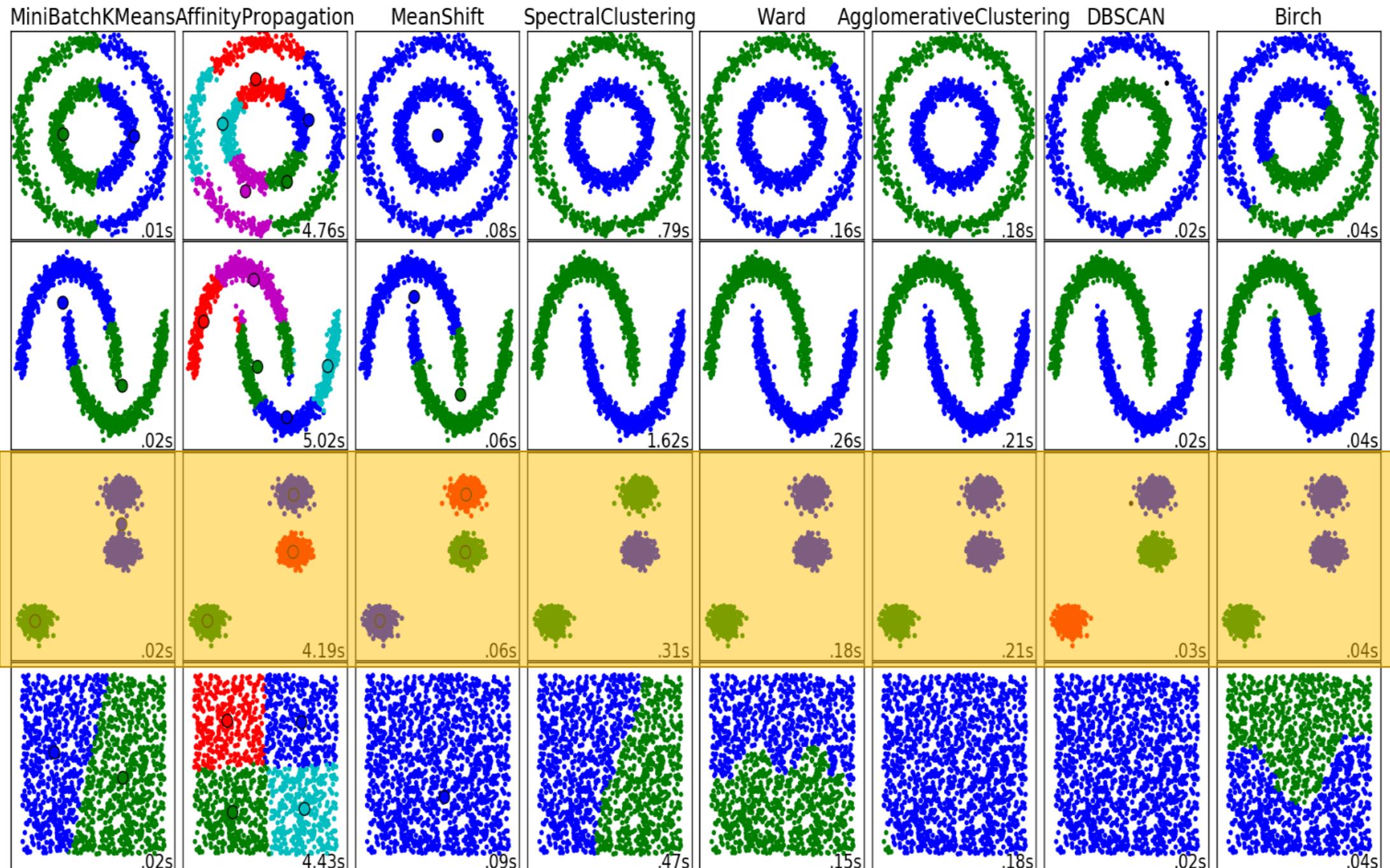
В scikit-learn:

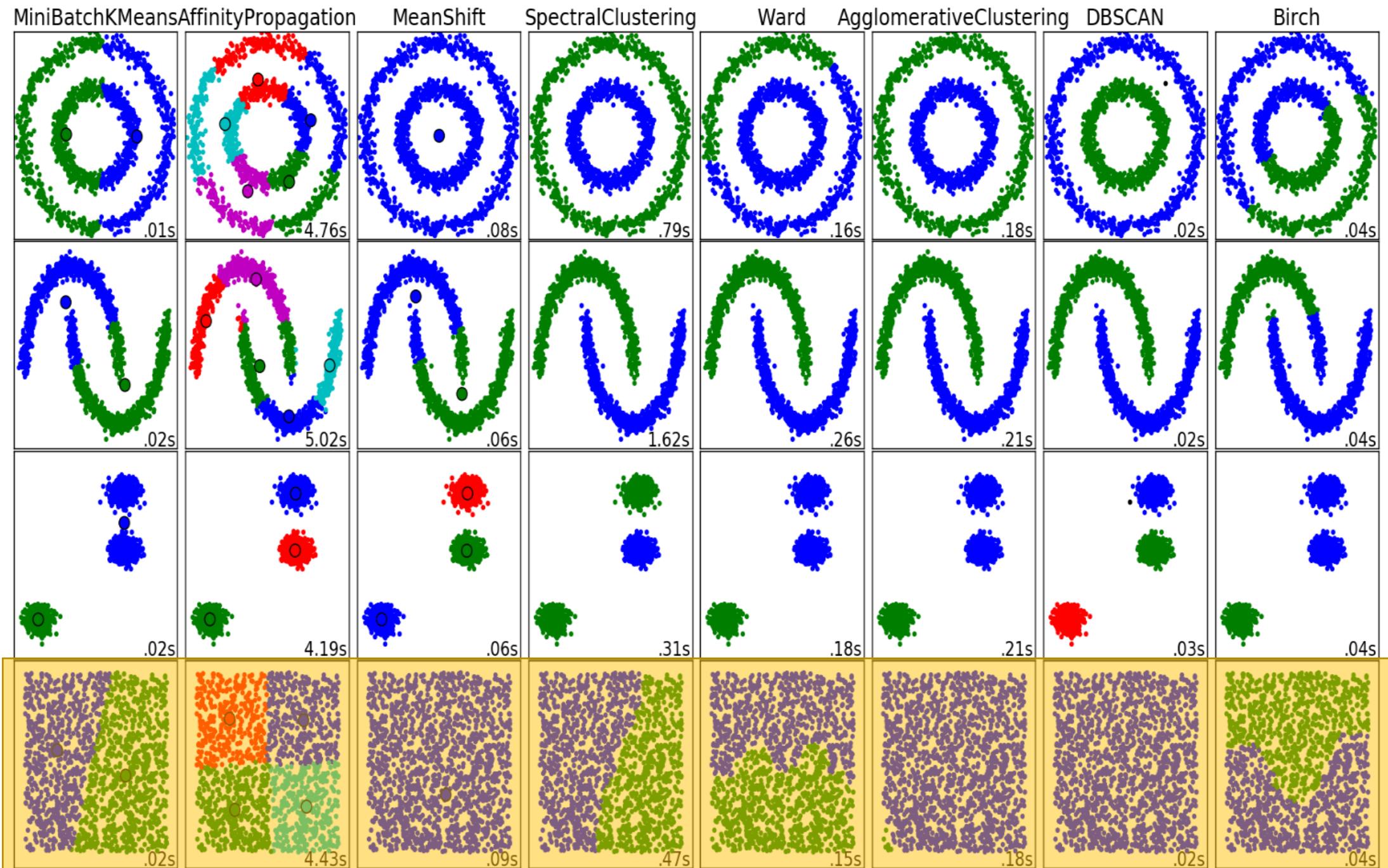
KMeans, MiniBatchKMeans, GaussianMixture,
AgglomerativeClustering, Ward, DBSCAN, MeanShift,
AffinityPropagation, SpectralClustering, Birch

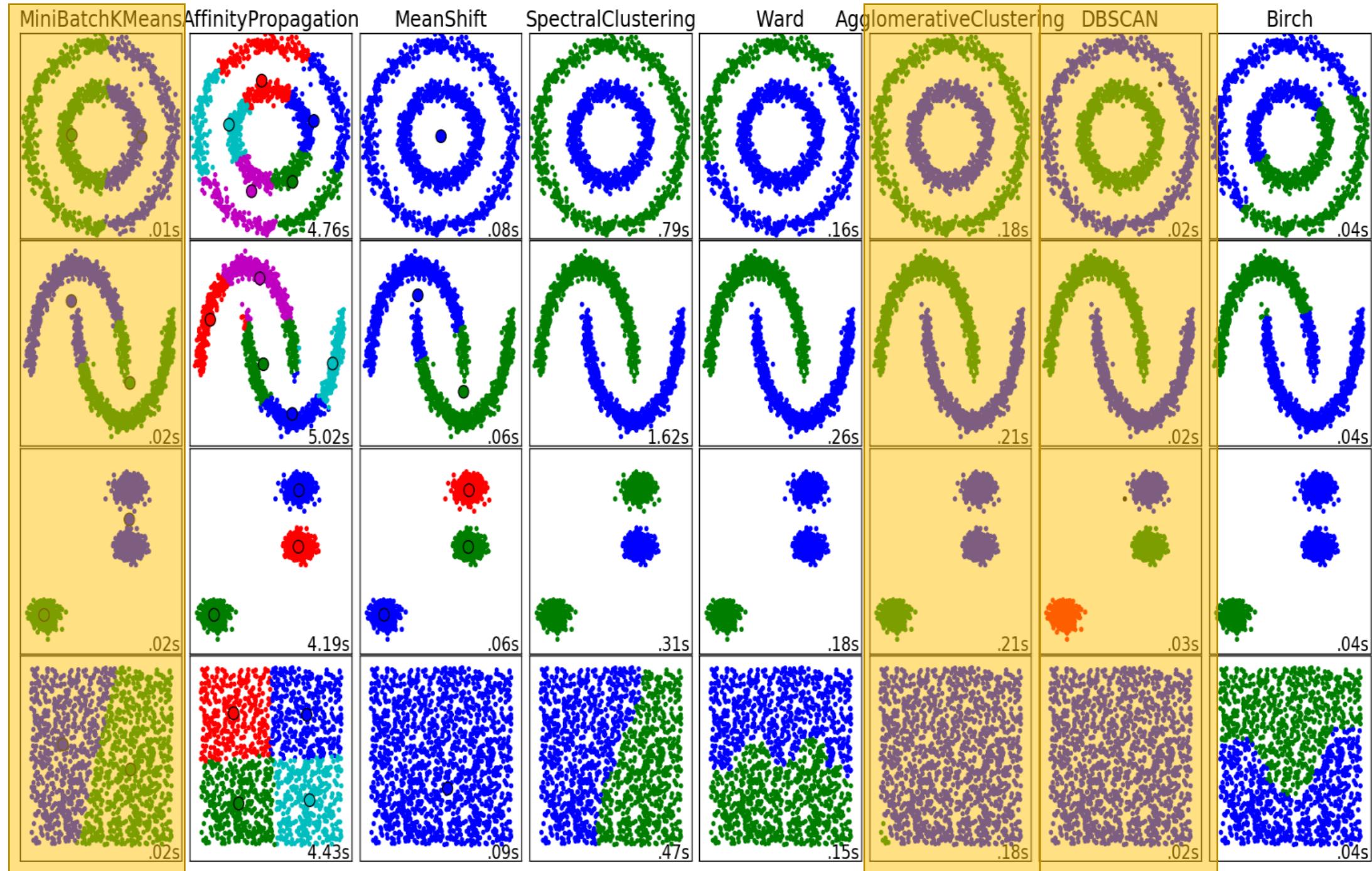












| Метод | Параметры | Масштабируемость | Use-case | Геометрия |
|--------|-----------------|---|--|----------------------|
| KMeans | Число кластеров | Очень много примеров (MiniBatch), среднее число кластеров | Выпуклые, примерно одинаковые кластеры | Евклидово расстояние |
| | | | | |
| | | | | |

| 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 |

Попробуем систематизировать

- По структуре кластеров:
 - Иерархические
 - Агломеративные
 - Дивизионные
 - Плоские
- По форме
 - Кластеры выпуклой формы
 - Кластеры-ленты
 - Сгустки на «фоне»
 - ...
- По присвоению объектов к кластерам:
 - Жесткая кластеризация
 - Мягкая кластеризация

Резюме

- К-средних
- EM с нормальным распределением
- Иерархическая агglomerативная кластеризация
- DBSCAN

9. Оценка качества и рекомендации по решению задачи кластеризации

План

1. Среднее внутрикластерное и межкластерное расстояние
2. Силуэт (silhouette coefficient)
3. Подбор количества кластеров по силуэту
4. Проверка наличия кластерной структуры
5. Проблема выбора хороших признаков
6. Полнота и однородность (completeness & homogeneity)
7. Оценка качества с привлечением ассессоров

Среднее внутриклusterное расстояние

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

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

Среднее межклusterное расстояние

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

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

Комбинируем функционалы

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

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

$$F_0/F_1 \rightarrow \min$$

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

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

$$\Phi_0/\Phi_1 \rightarrow \min$$

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*

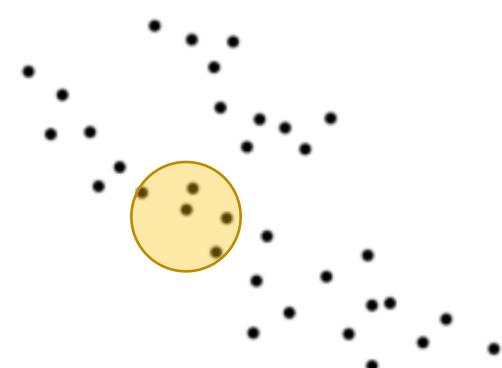
$$s = \frac{b - a}{\max(a, b)}$$



Коэффициент силуэта

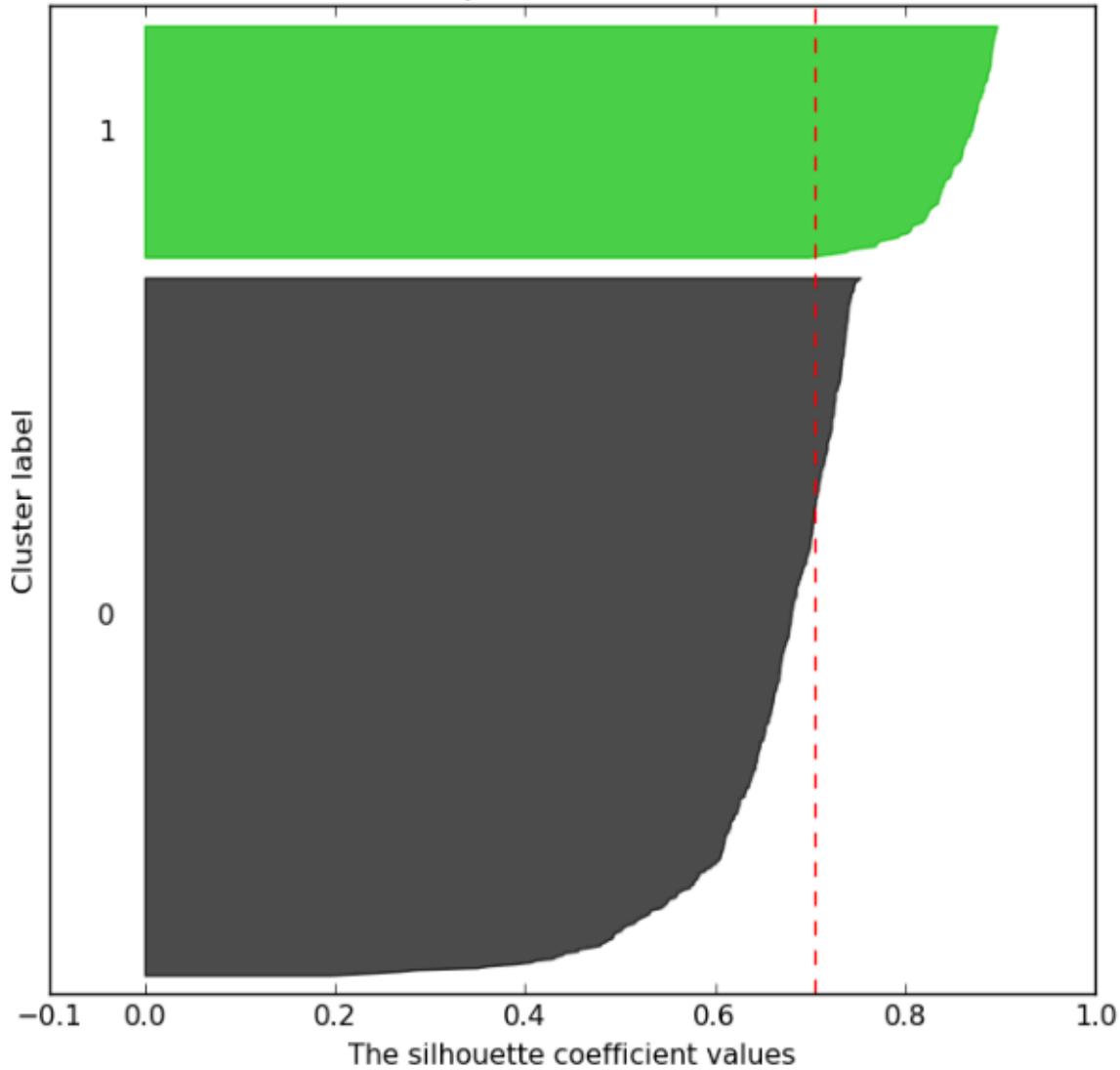
- **a**: Среднее расстояние от данного объекта до всех других объектов из того же кластера
- **b**: Среднее расстояние от данного объекта до всех объектов из *ближайшего другого кластера*

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

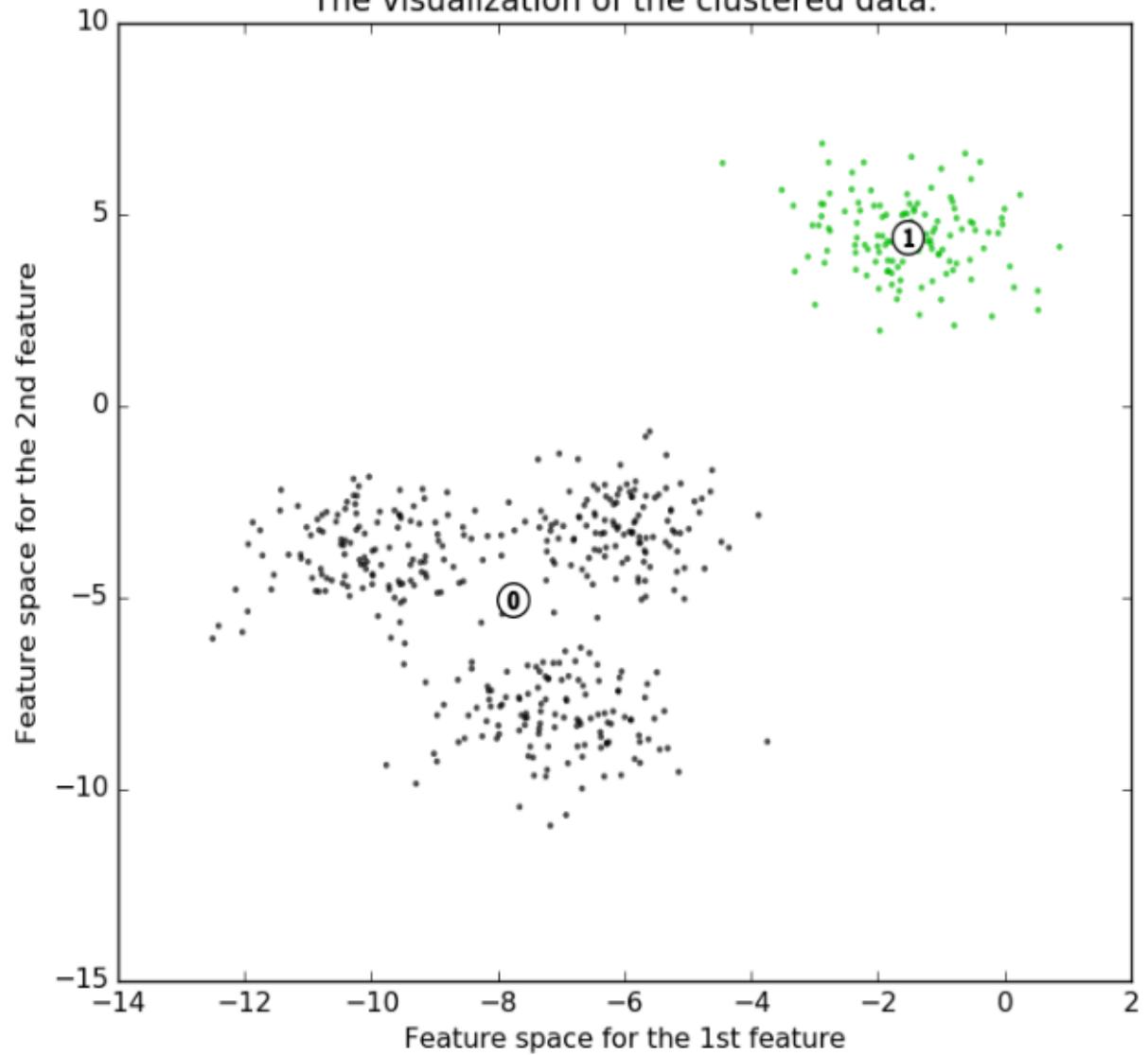


Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

The silhouette plot for the various clusters.

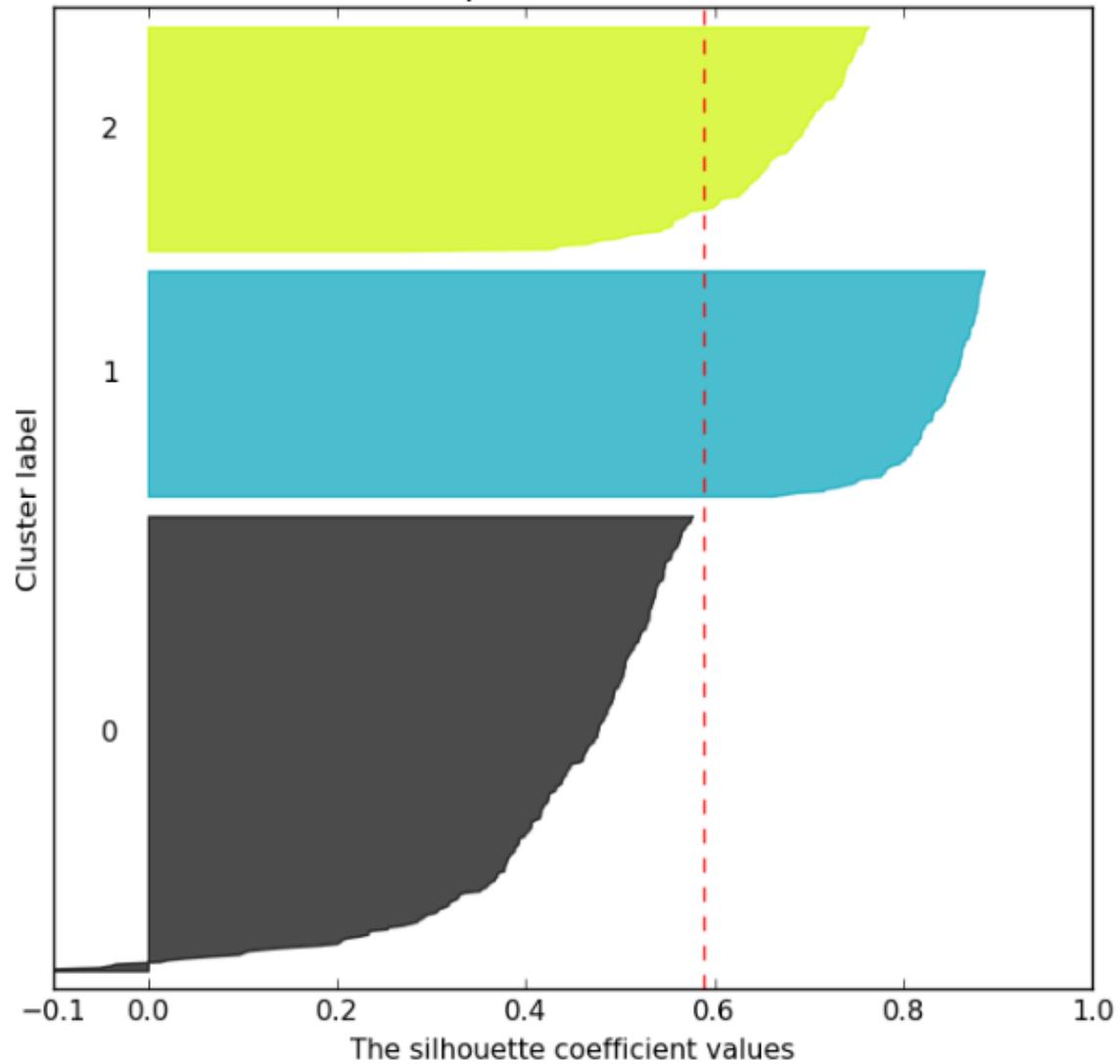


The visualization of the clustered data.

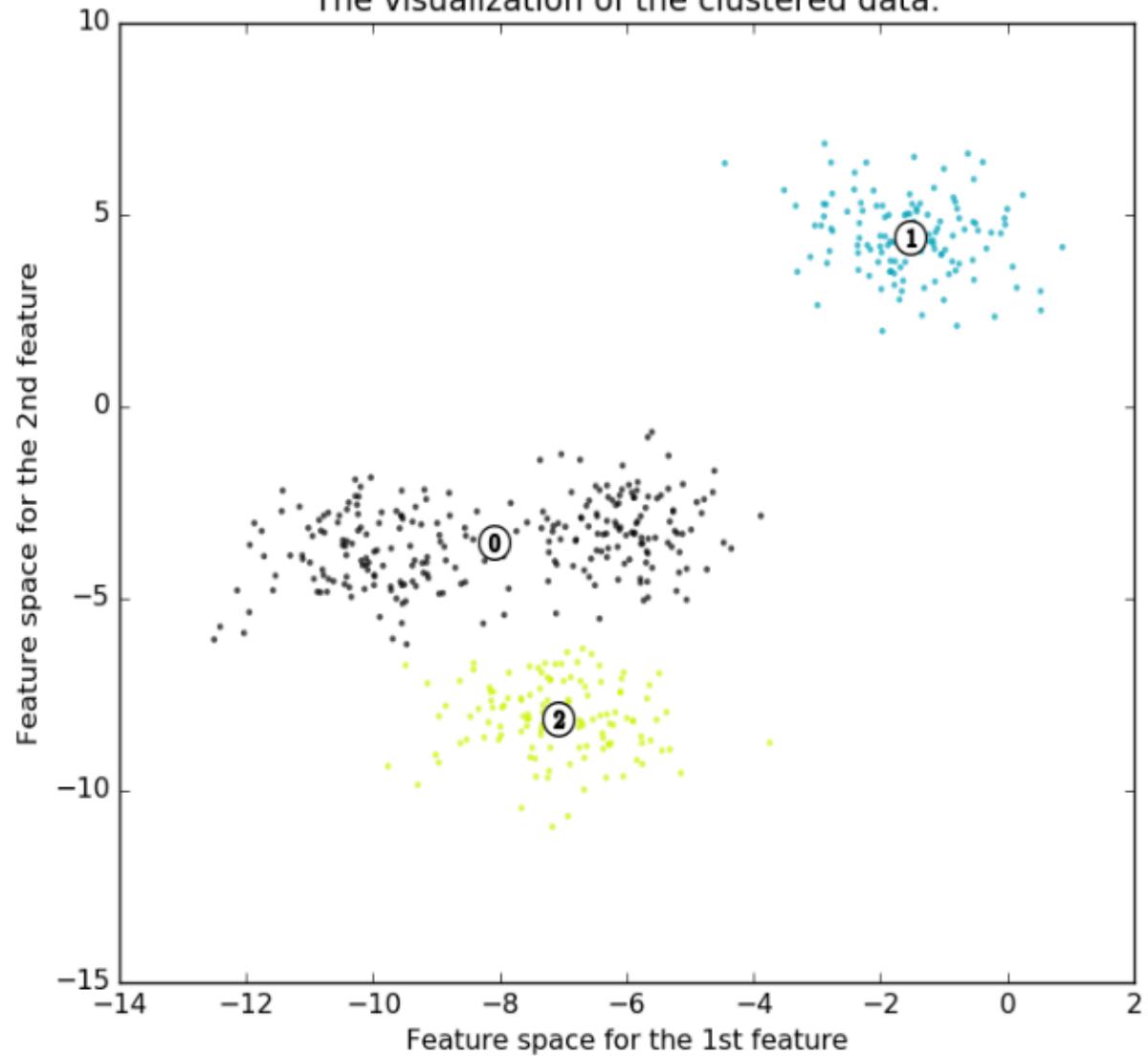


Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

The silhouette plot for the various clusters.

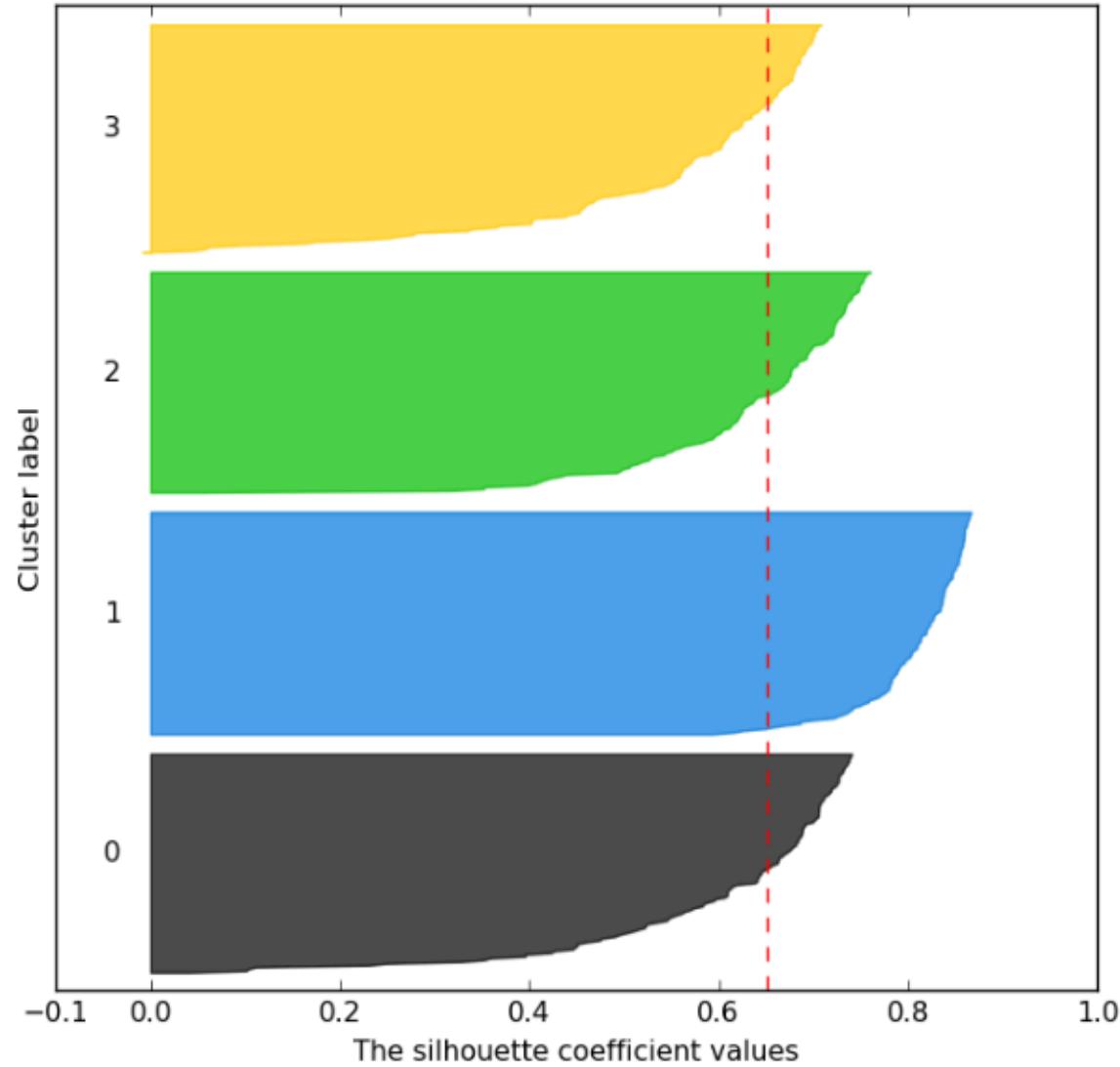


The visualization of the clustered data.

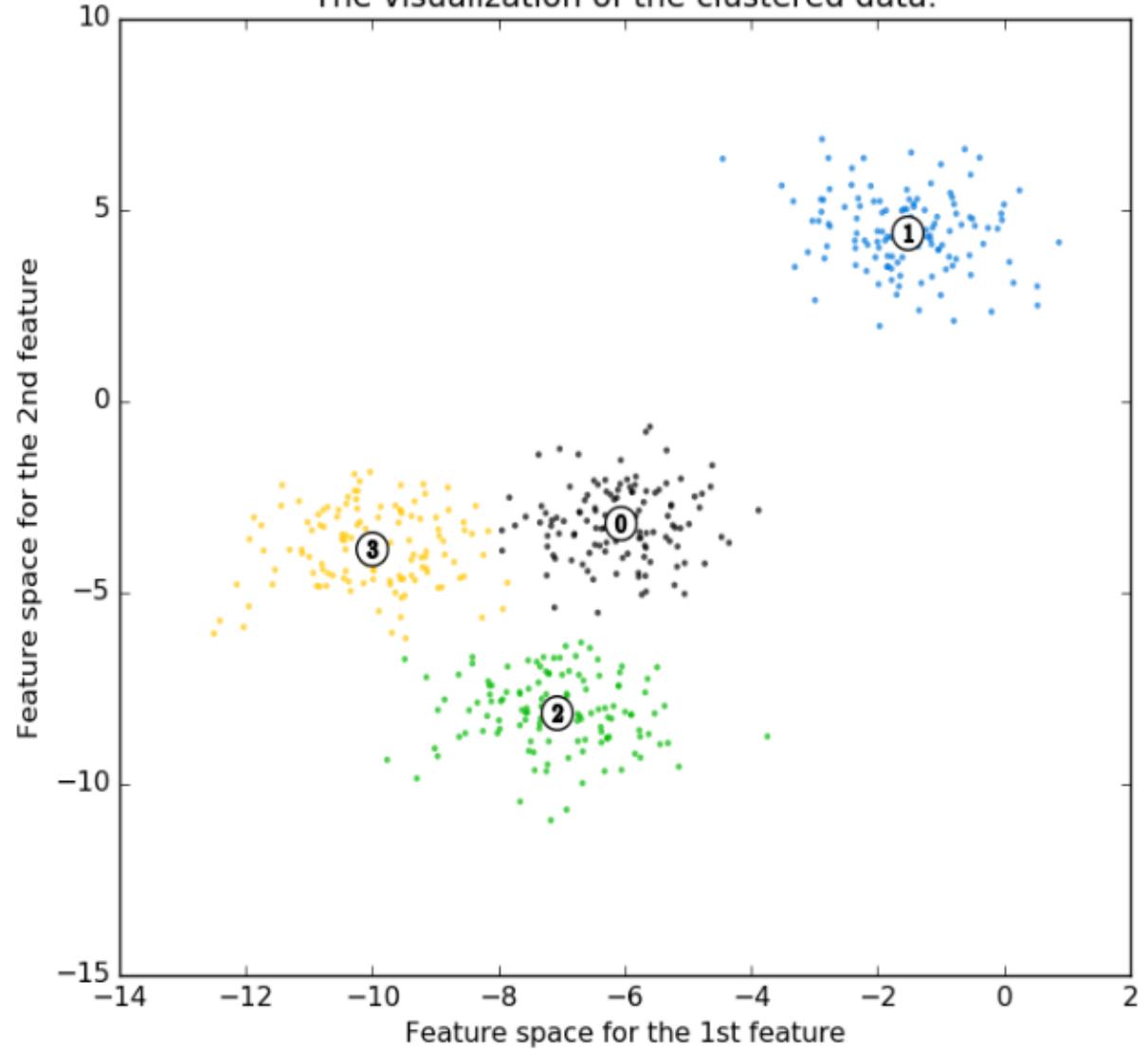


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4

The silhouette plot for the various clusters.

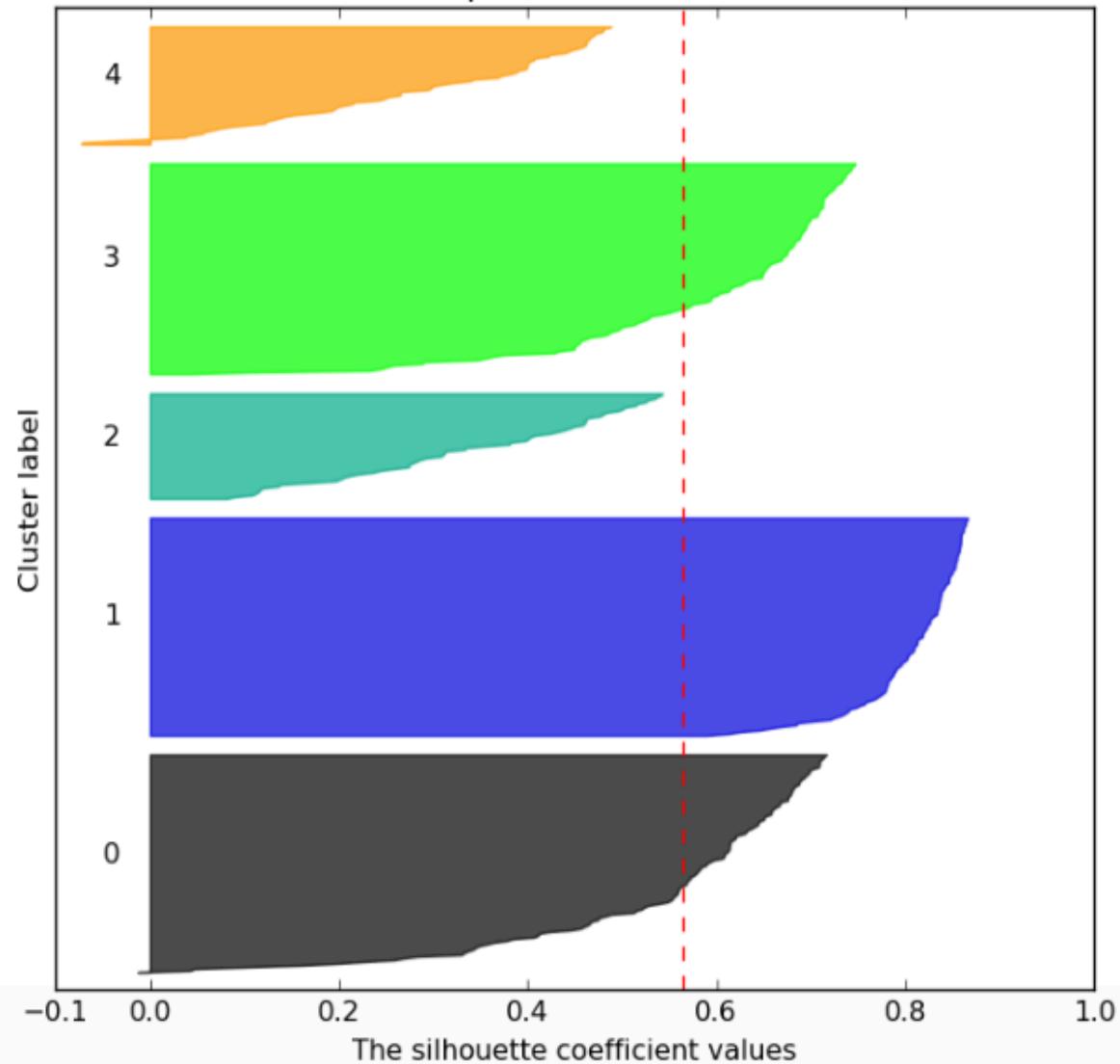


The visualization of the clustered data.

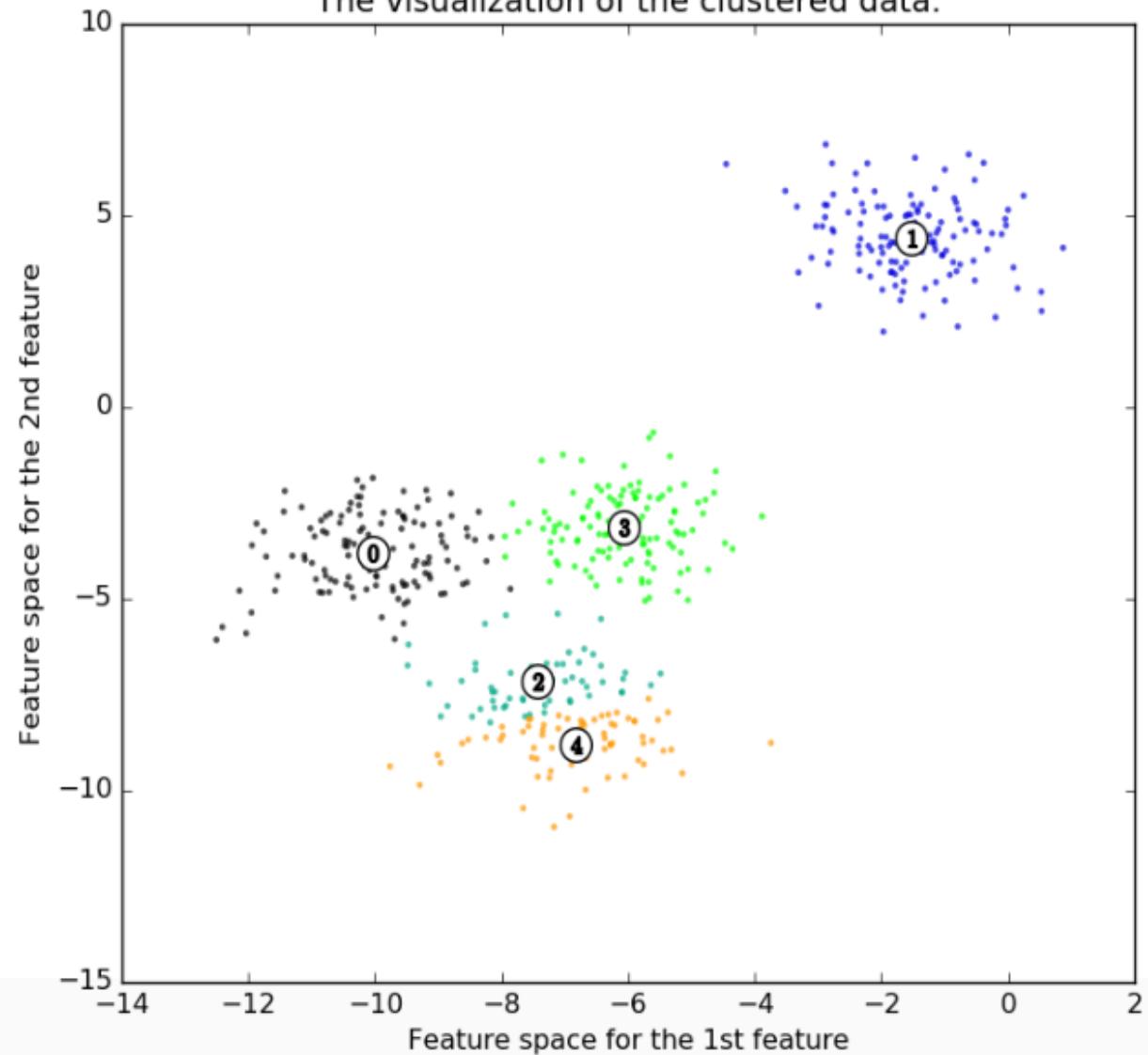


Silhouette analysis for KMeans clustering on sample data with n_clusters = 5

The silhouette plot for the various clusters.

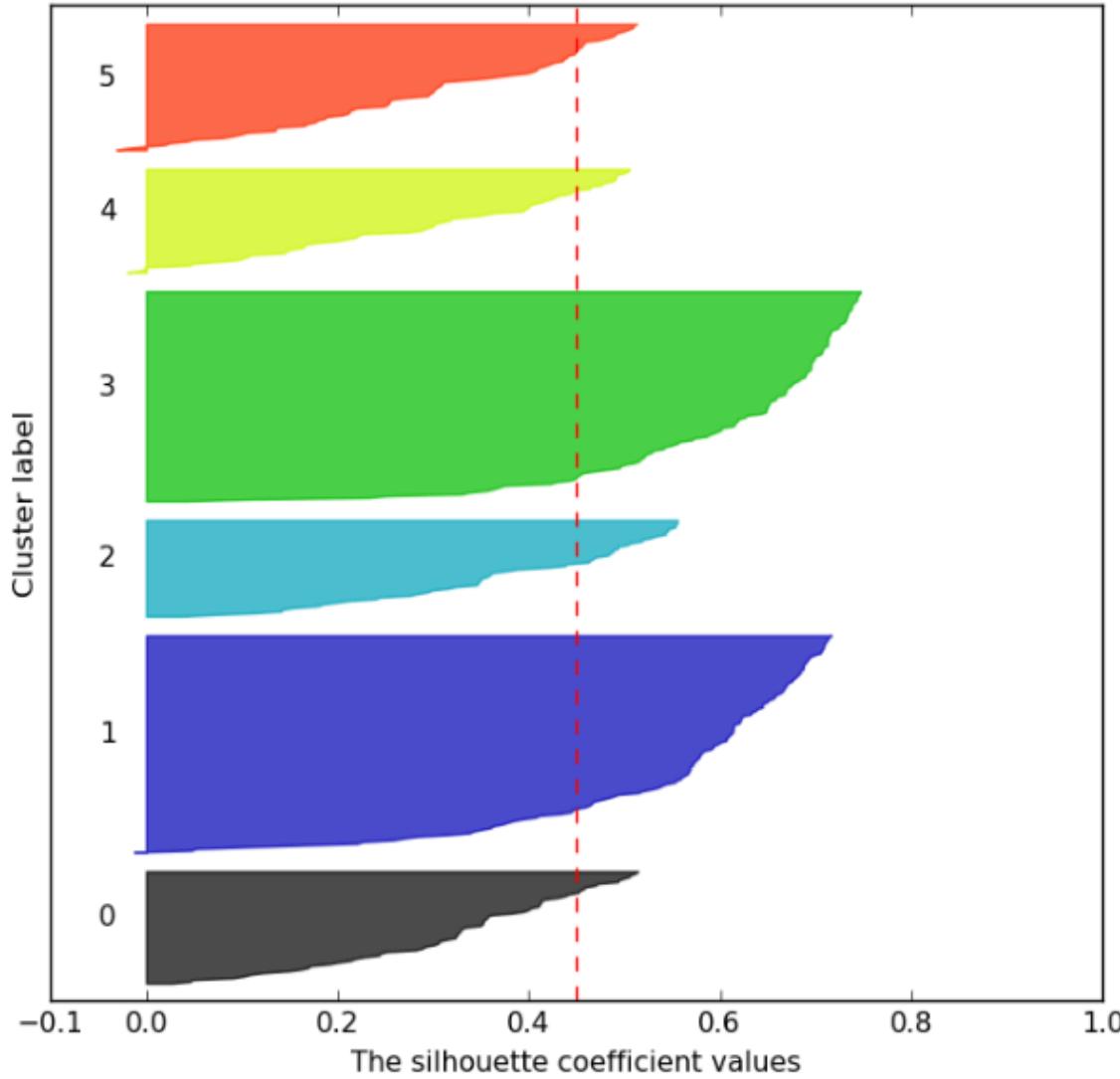


The visualization of the clustered data.

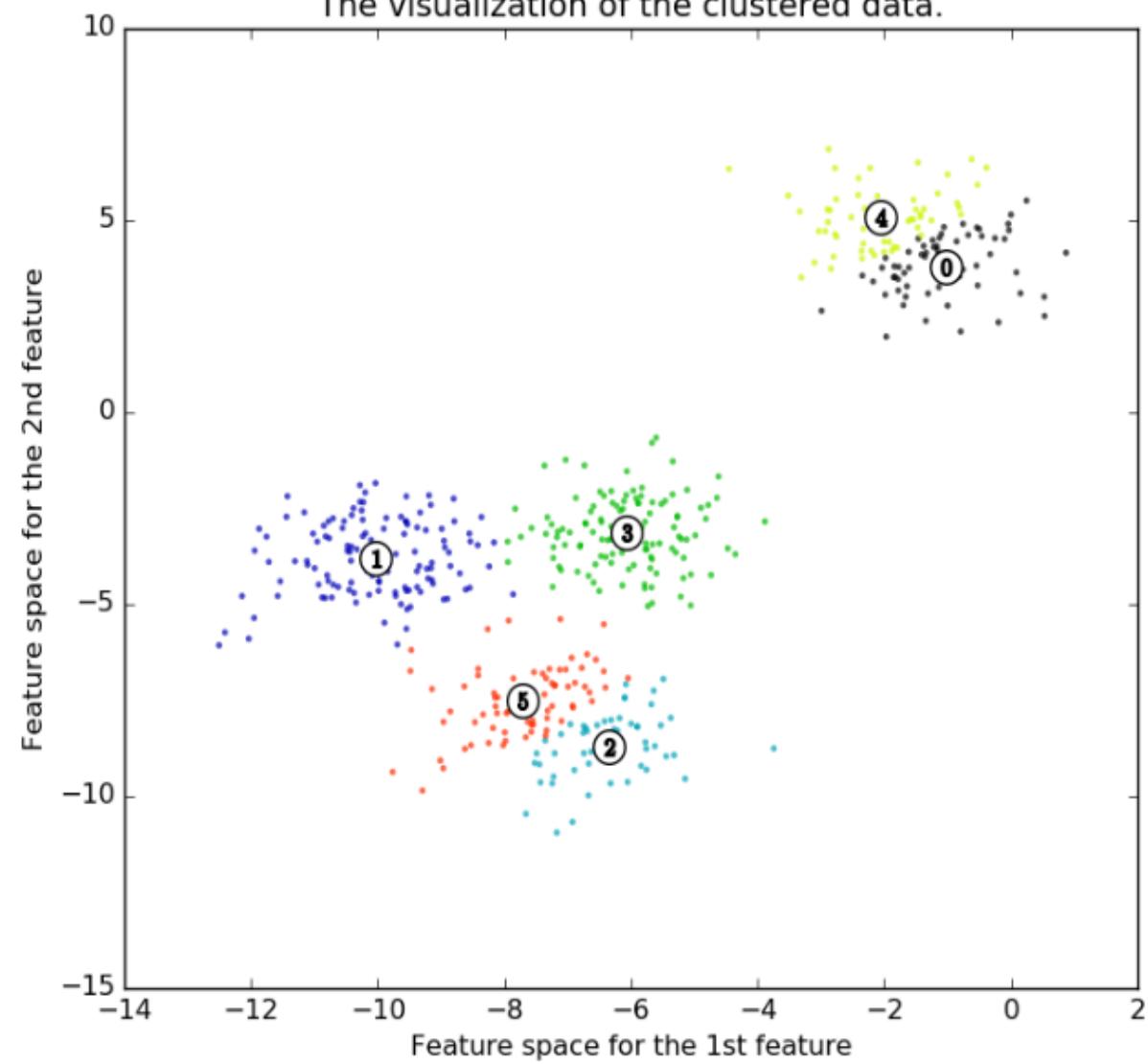


Silhouette analysis for KMeans clustering on sample data with n_clusters = 6

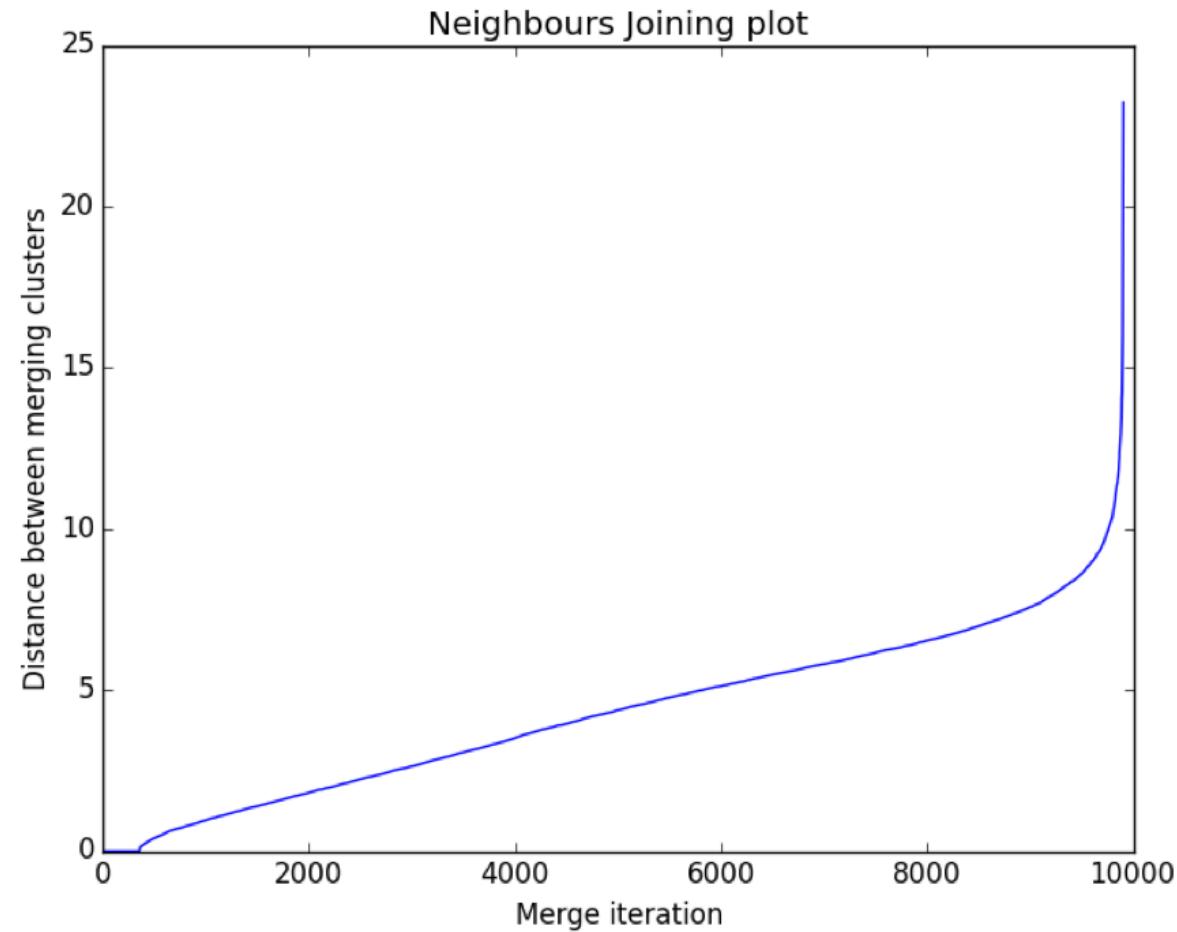
The silhouette plot for the various clusters.



The visualization of the clustered data.



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, F-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

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

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

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

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

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

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

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

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

Однородность, полнота, V-мера

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

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

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

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

Однородность, полнота, V-мера

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

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

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

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

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

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»

Резюме

1. Среднее внутрикластерное и межкластерное расстояние
2. Силуэт (silhouette coefficient)
3. Подбор количества кластеров по силуэту
4. Проверка наличия кластерной структуры
5. Проблема выбора хороших признаков
6. Полнота и однородность (completeness & homogeneity)
7. Оценка качества с привлечением ассессоров

На следующей лекции:
метрики качества и работа с признаками