

Data Mining in Action

Лекция 4. Обучение без учителя (unsupervised learning)



Упражнение с прошлой лекции: поиск ошибок

```
from random import randint

def loss(x, answer):
    return max([0, 1 - answer * f(x)])

def der_loss(x, answer):
    return -1.0 if 1 - answer * f(x) > 0 else 0.0

def fit(X_train, y_train):
    for k in range(10000):
        rand_index = randint(0, len(X_train))
        x = X_train[rand_index]
        y = y_train[rand_index]

        step = 0.01

        w -= x * step * y * der_loss(x, y)
```

Упражнение с прошлой лекции: поиск ошибок

```
from random import randint

def loss(x, answer):
    return max([0, 1 - answer * f(x)])

def der_loss(x, answer):
    return -1.0 if 1 - answer * f(x) > 0 else 0.0

def fit(X_train, y_train):
    for k in range(10000):
        rand_index = randint(0, len(X_train))
        x = X_train[rand_index]
        y = y_train[rand_index]

        step = 0.01

        w -= x * step * y * der_loss(x, y)
```

Ответы 0 и 1, а формулы – для +1 и -1

Упражнение с прошлой лекции: поиск ошибок

```
from random import randint

def loss(x, answer):
    return max([0, 1 - answer * f(x)])

def der_loss(x, answer):
    return -1.0 if 1 - answer * f(x) > 0 else 0.0

def fit(X_train, y_train):
    for k in range(10000):
        rand_index = randint(0, len(X_train))
        x = X_train[rand_index]
        y = y_train[rand_index]

        step = 0.01
        w -= x * step * y * der_loss(x, y)

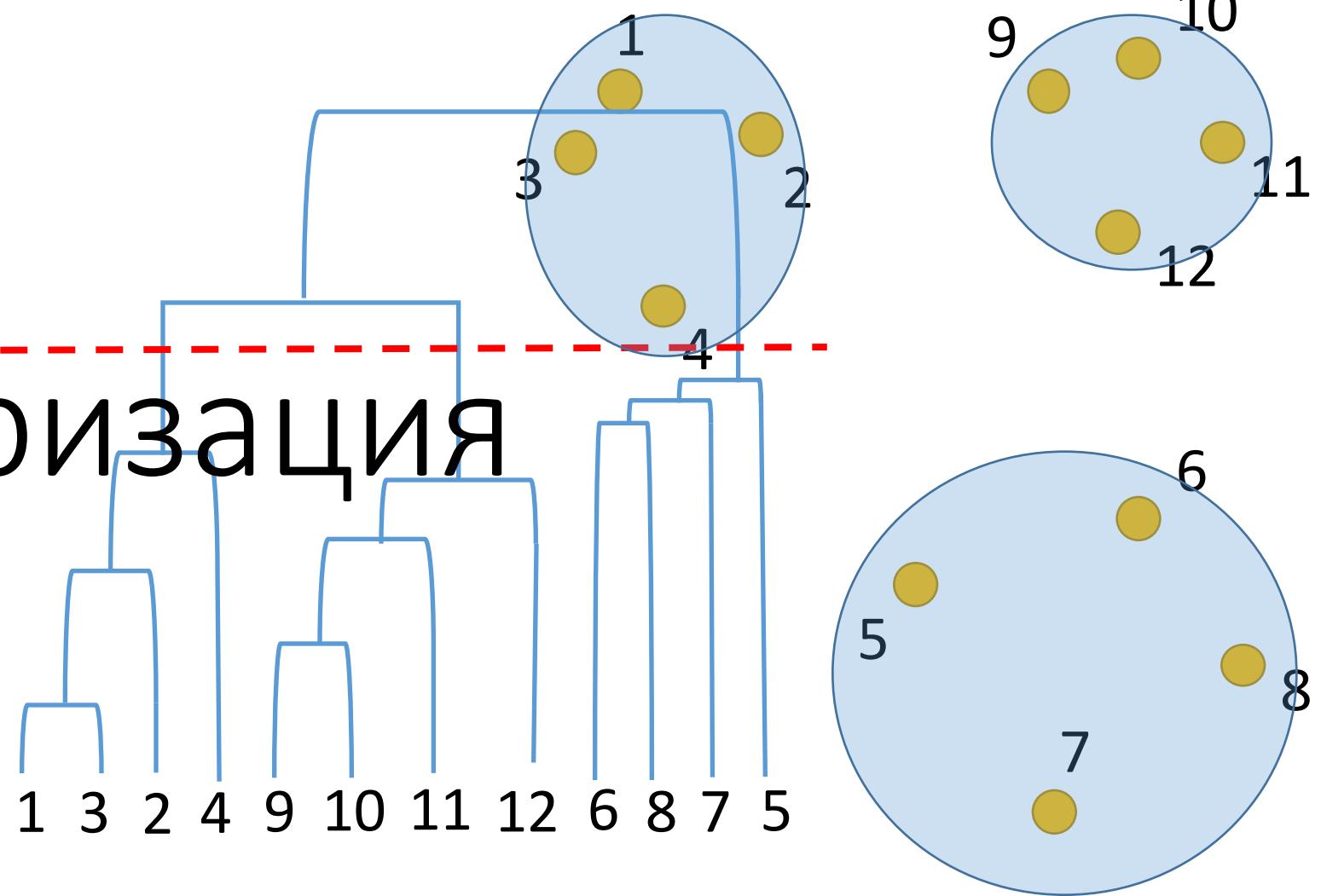
        Ответы 0 и 1, а формулы – для +1 и -1

        Обновляем только w – надо либо обновлять
        и w0, либо добавлять фиктивный признак
```

Сегодня на лекции

- Кластеризация
- Преобразование признаков
- Другие примеры unsupervised learning

Кластеризация



План

- I. Задача кластеризации
- II. Kmeans и EM-алгоритм
- III. Иерархические и графовые методы
- IV. Сравнение алгоритмов и оценка качества

I. 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

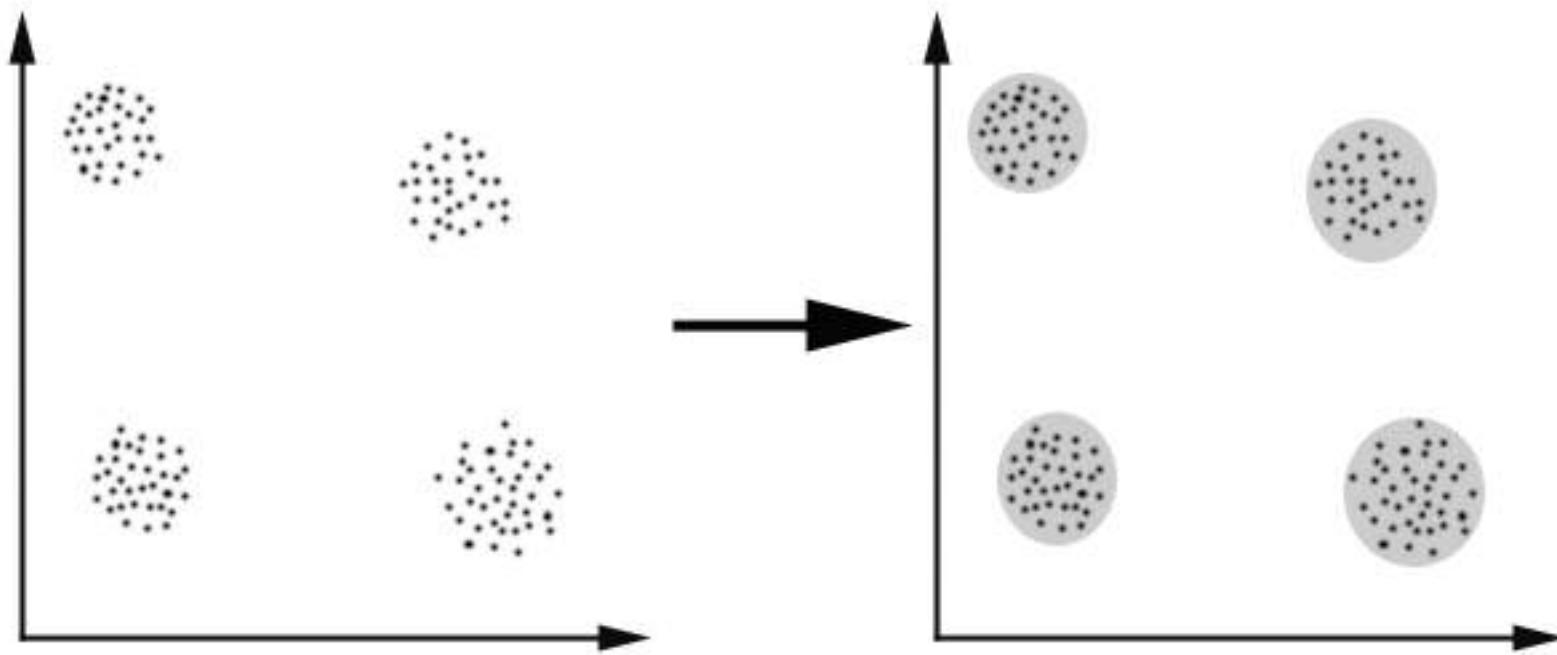
«Обучающая» выборка:

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

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

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

$$F_0/F_1 \rightarrow \min$$

Форма кластеров



Рисунок взят из курса лекций по машинному обучению К.В. Воронцова

Форма кластеров

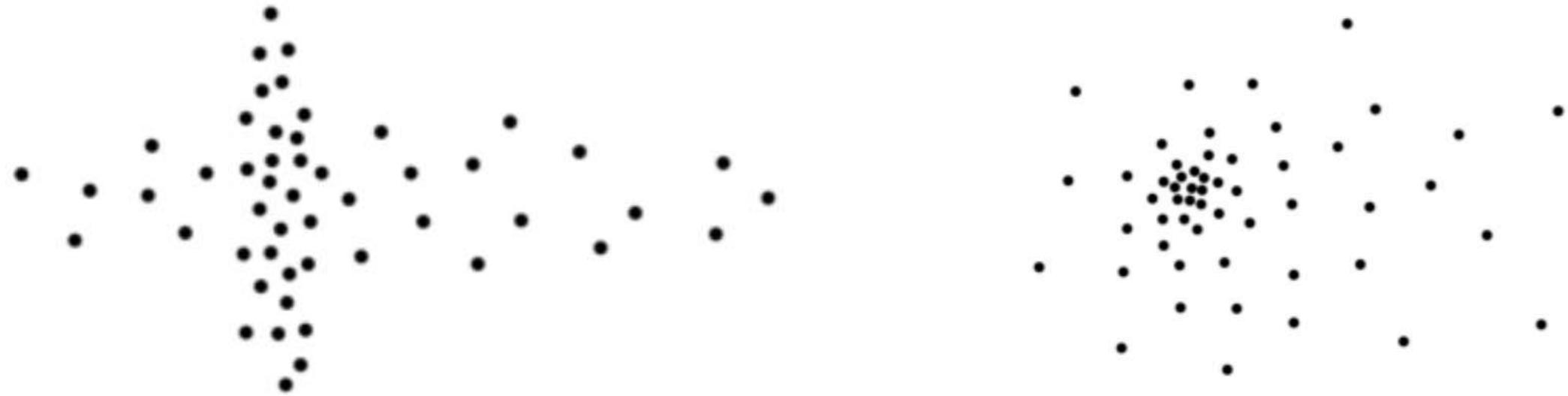
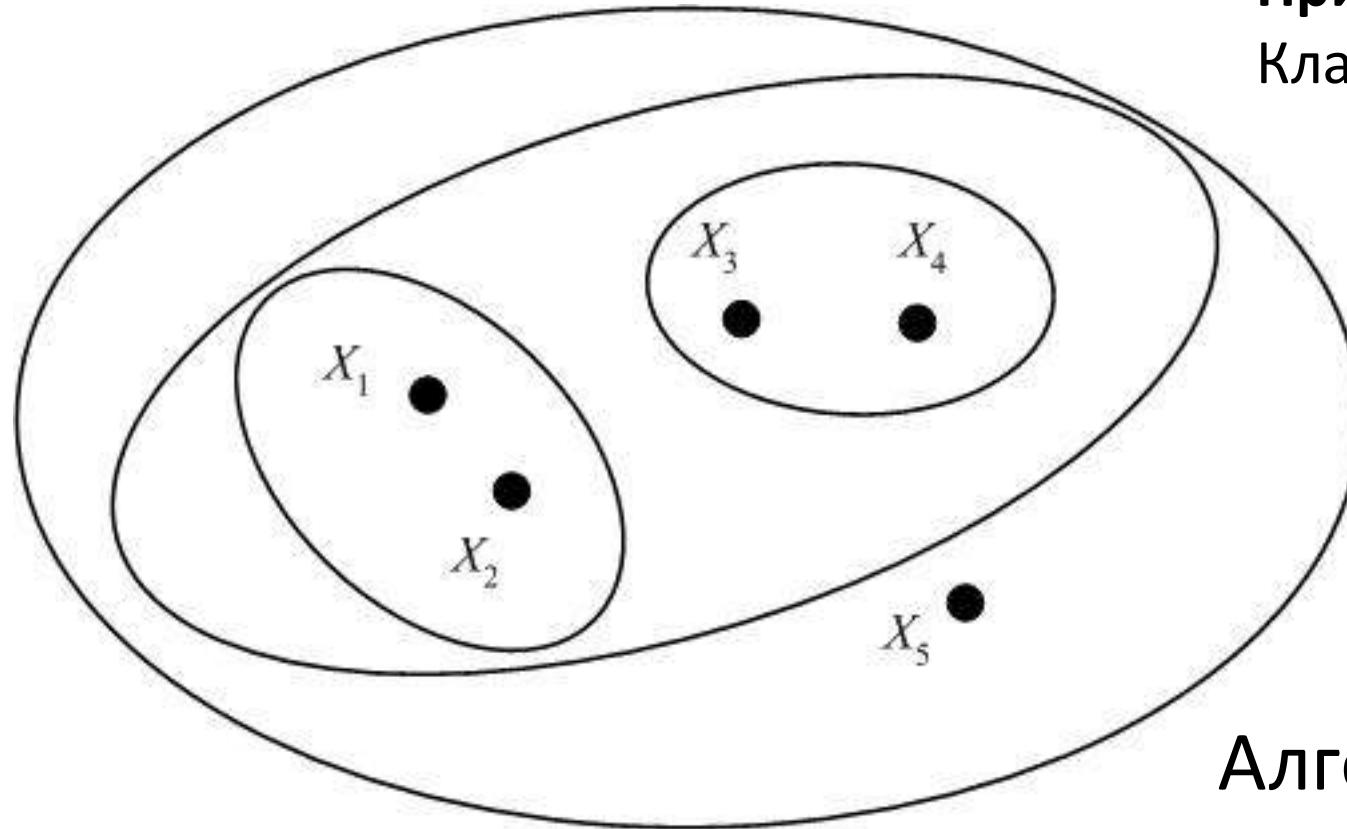


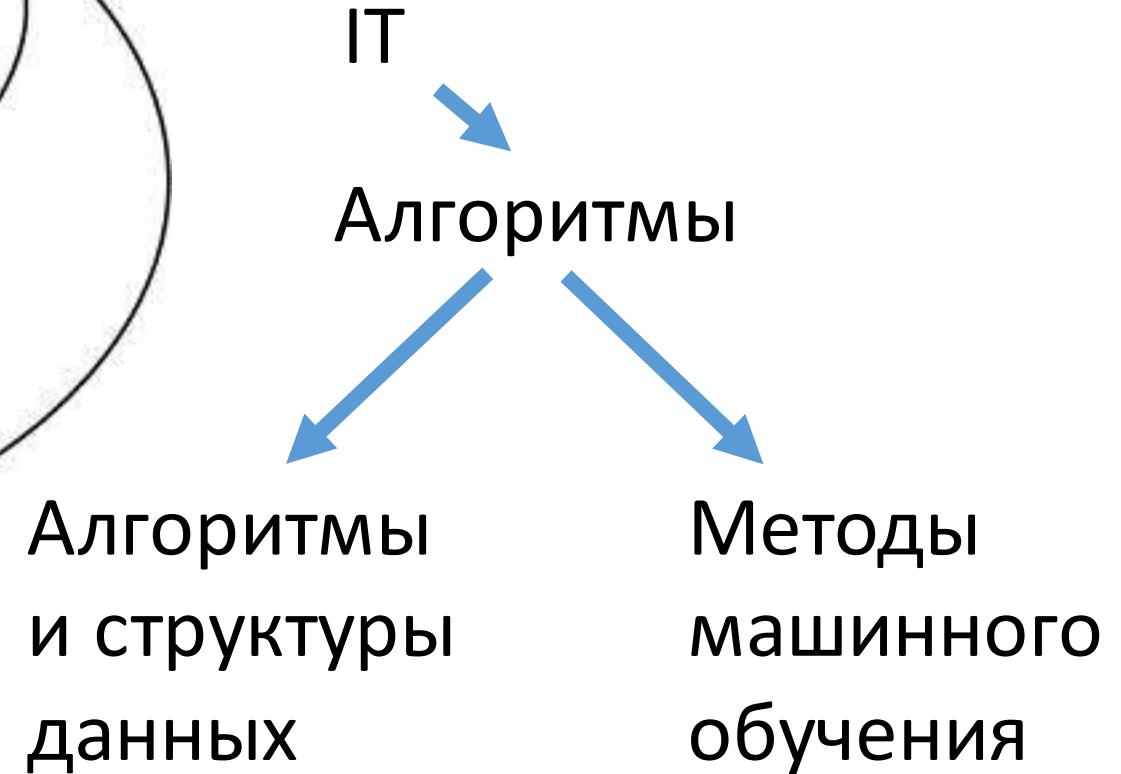
Рисунок взят из курса лекций по машинному обучению К.В. Воронцова

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

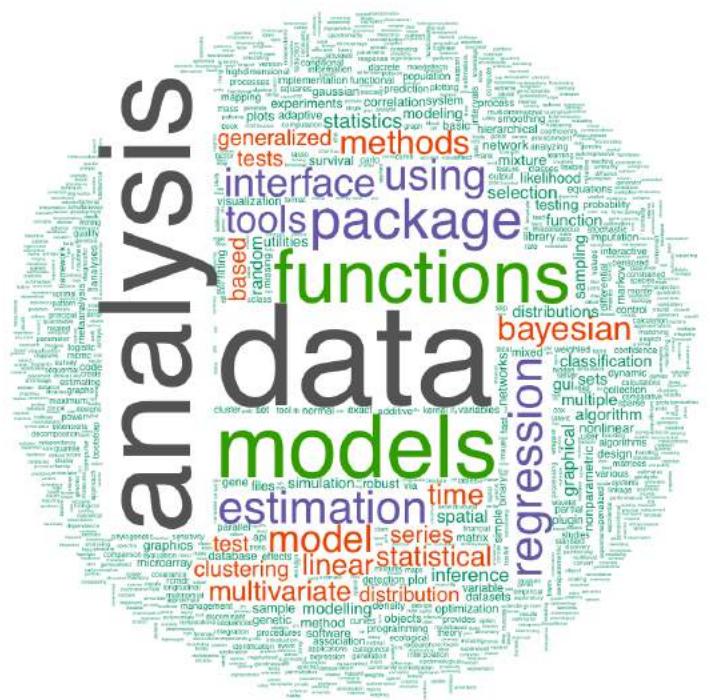
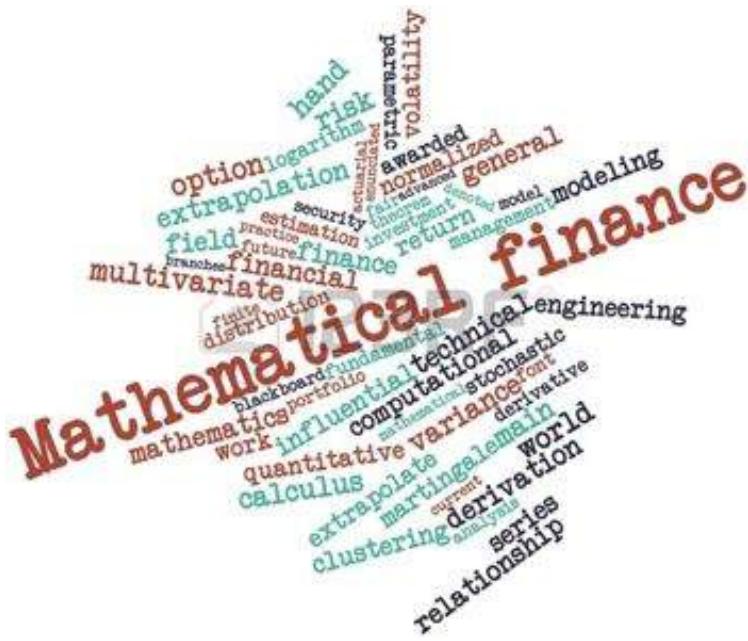
Открытие Олимпиады в Сочи
посмотрели несколько миллиардов
человек

Олимпиада в Сочи открыта

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

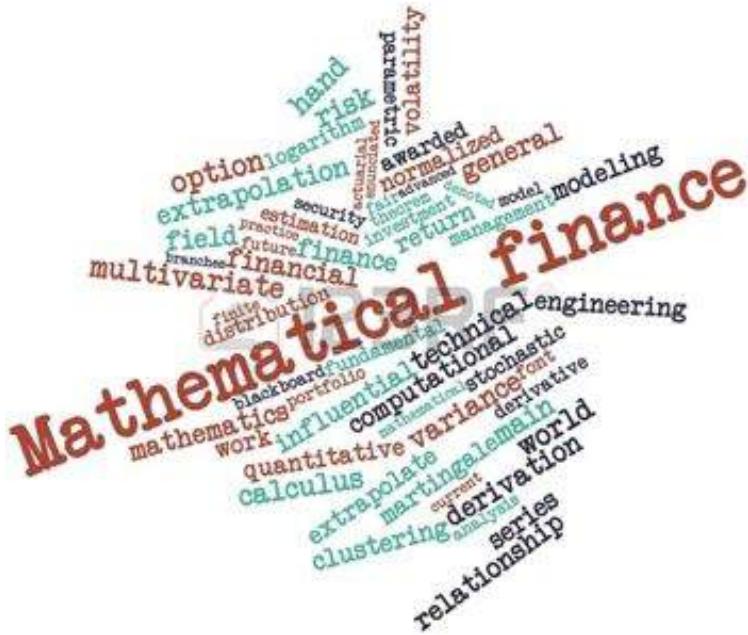
«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»

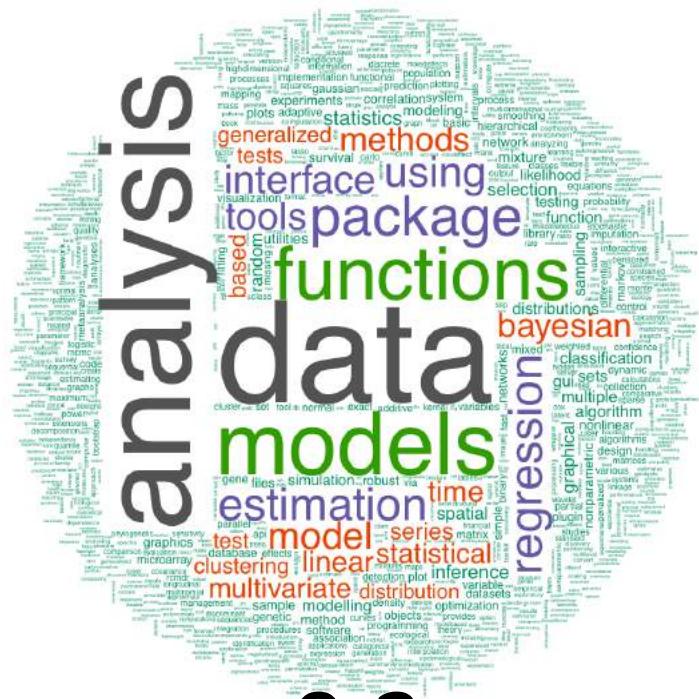


«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»



0.2



0.3

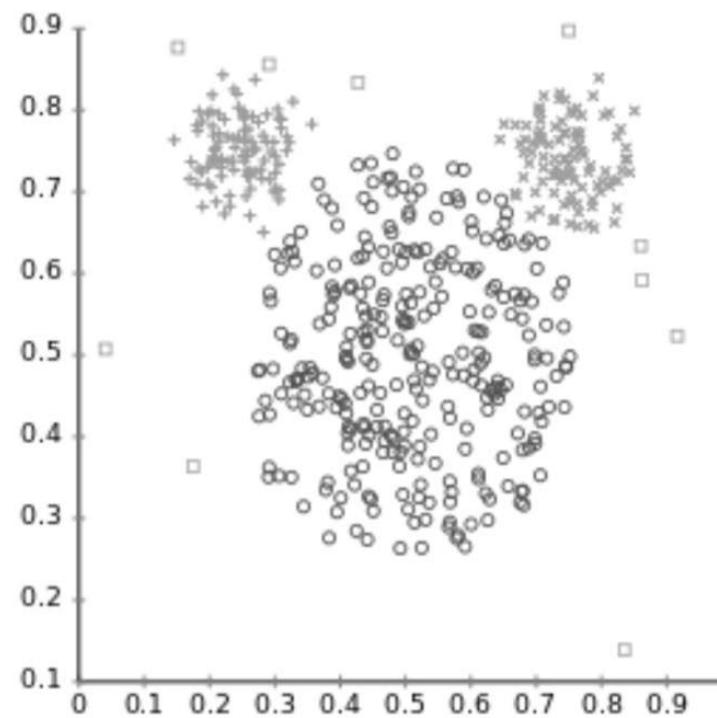


0.5

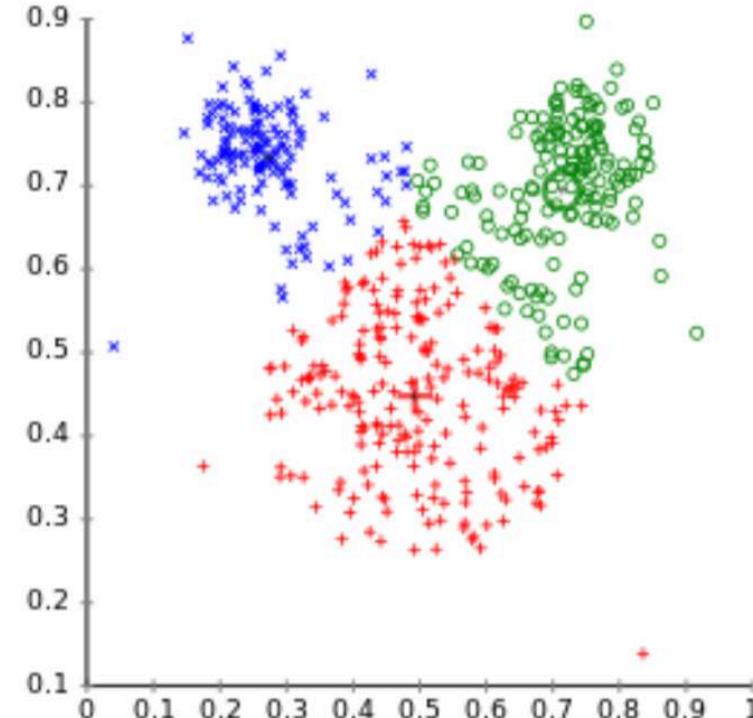
Резюме: чем могут отличаться задачи кластеризации

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

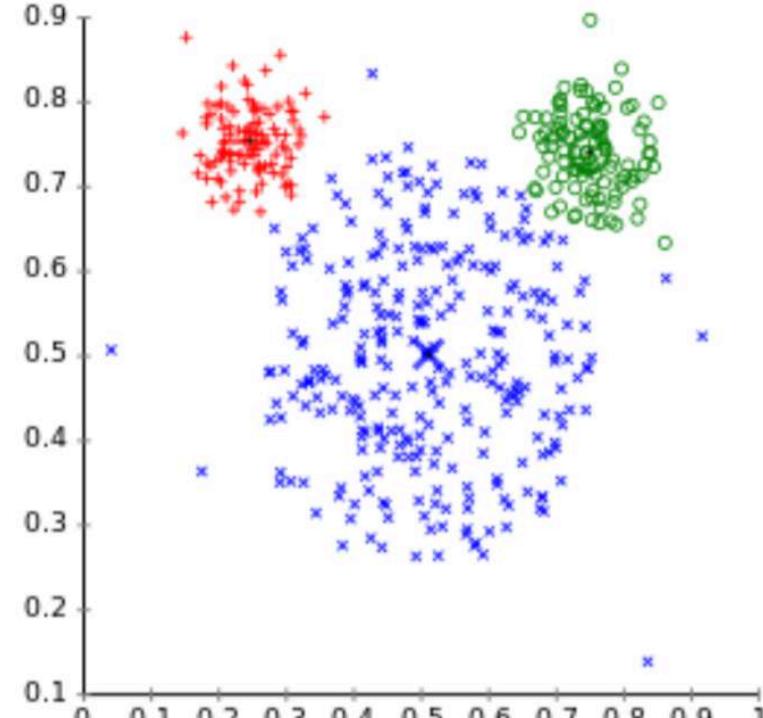
Различия в результатах работы методов



Исходная выборка
("Mouse" dataset)



Метод k средних
(K-Means)



ЕМ-алгоритм

II. K-Means и EM

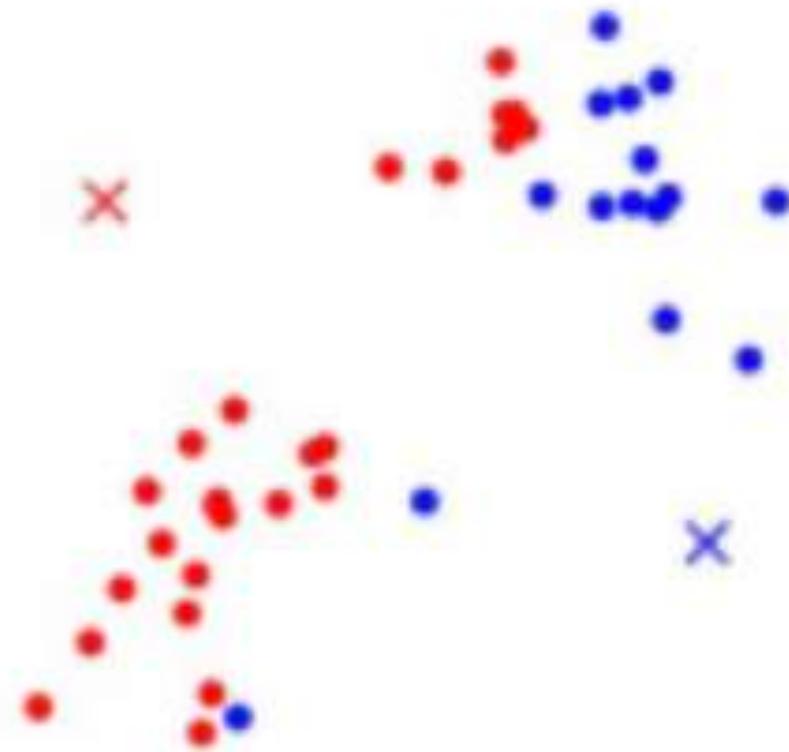
Как работает K-Means



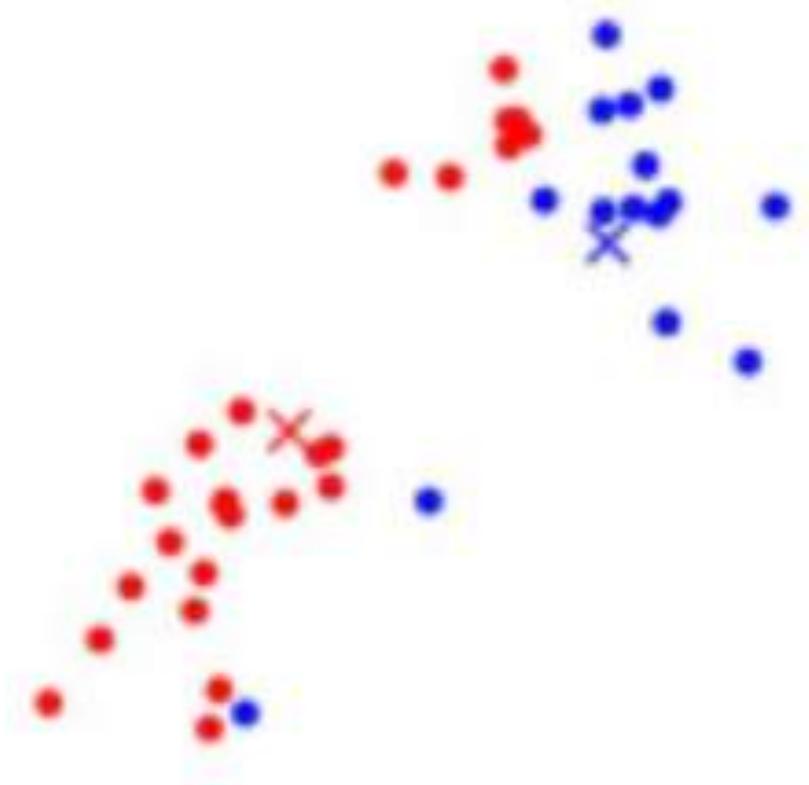
Как работает K Means



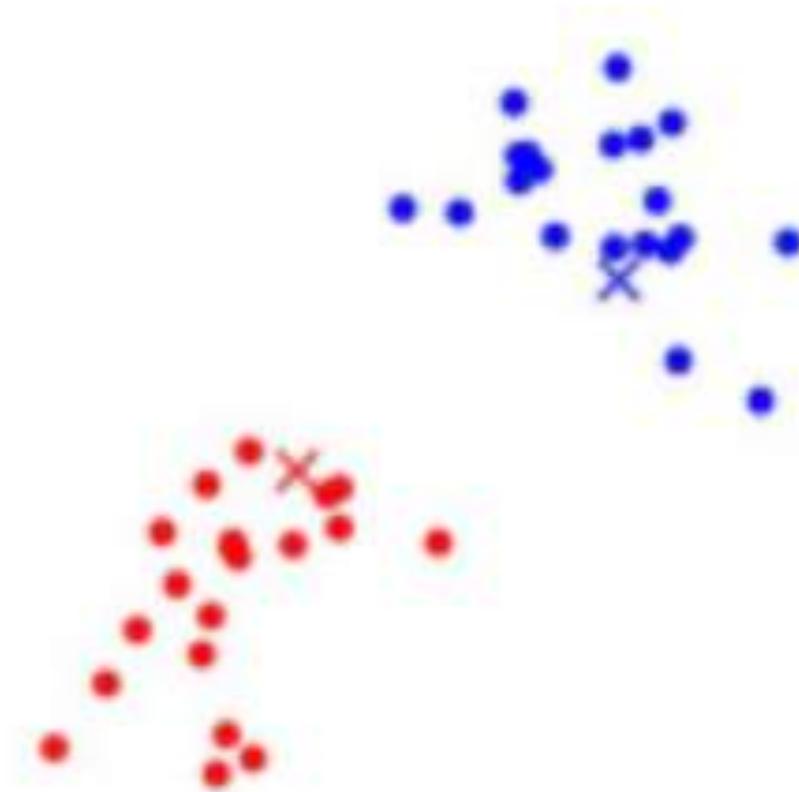
Как работает K-Means



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)

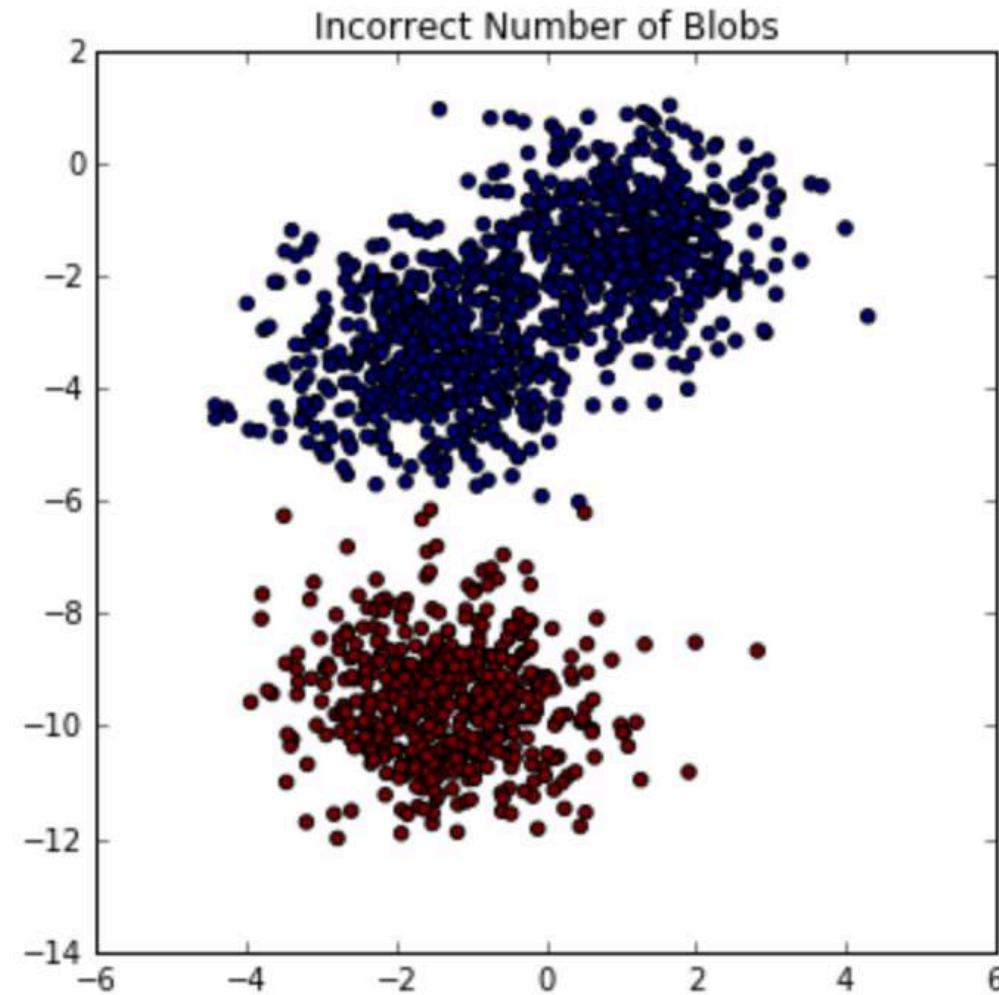


Пример: квантизация изображений

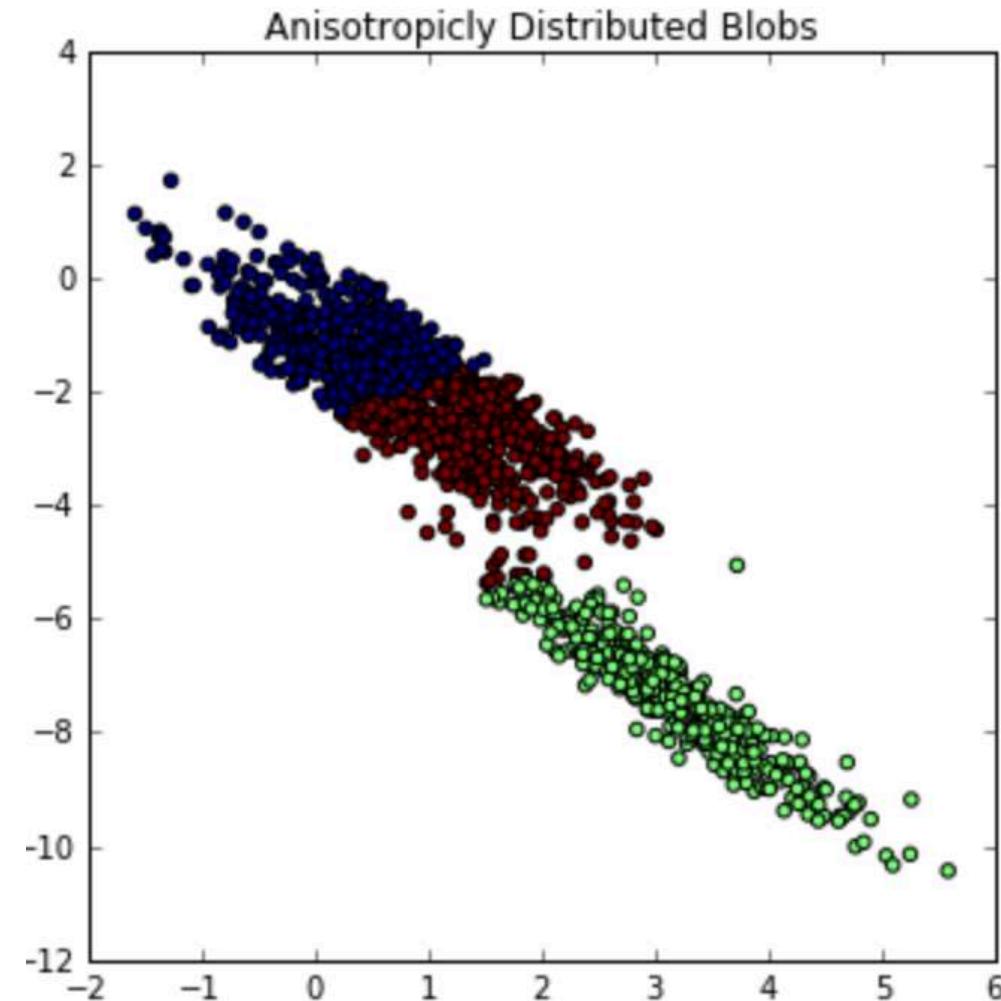
Quantized image (64 colors, K-Means)



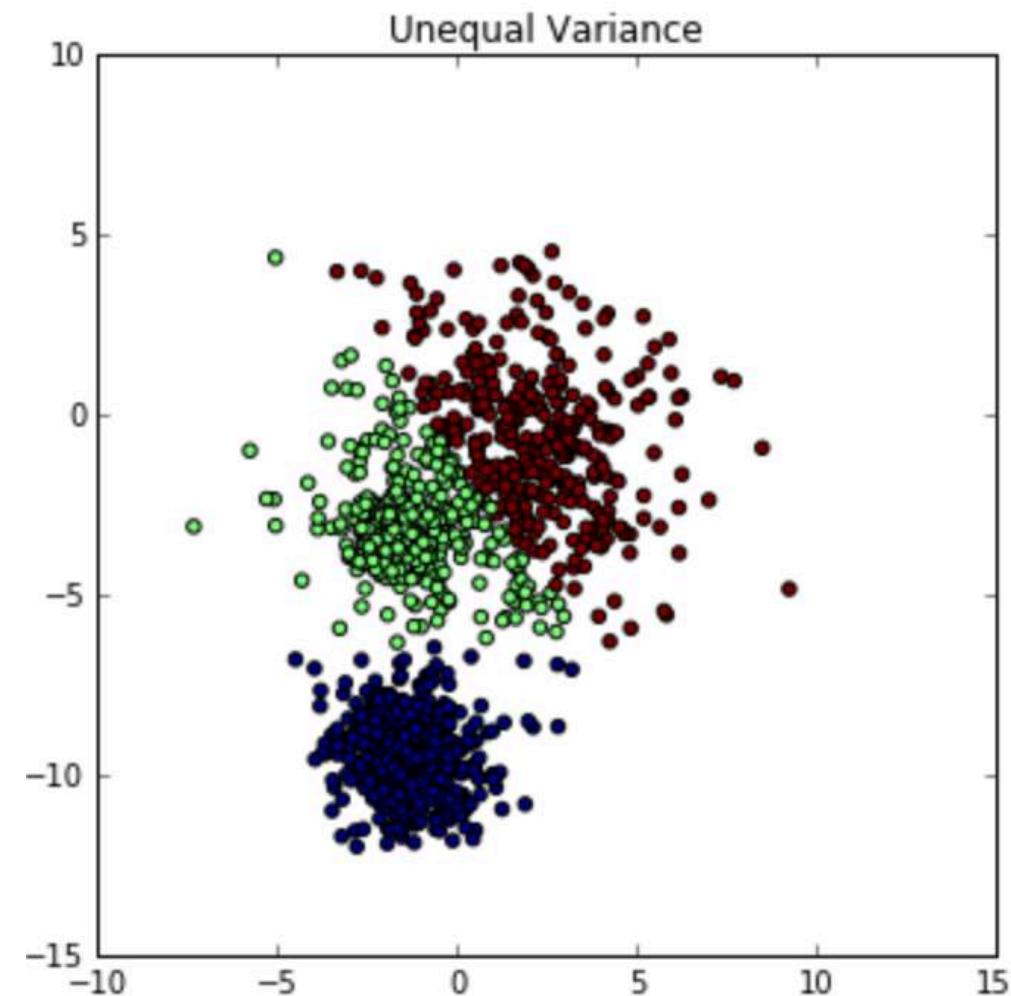
K Means и разные формы кластеров



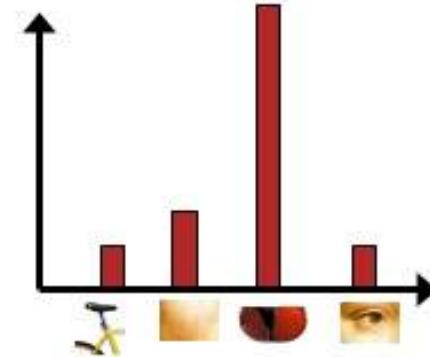
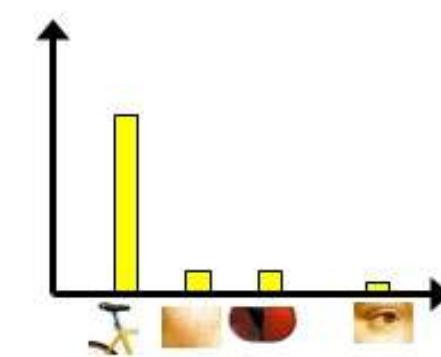
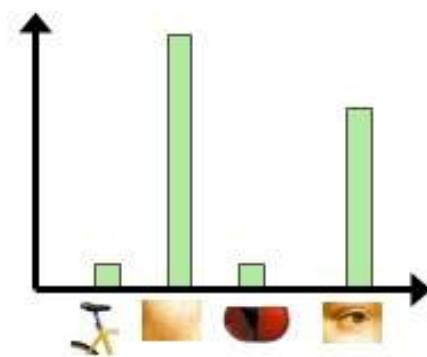
K Means и разные формы кластеров



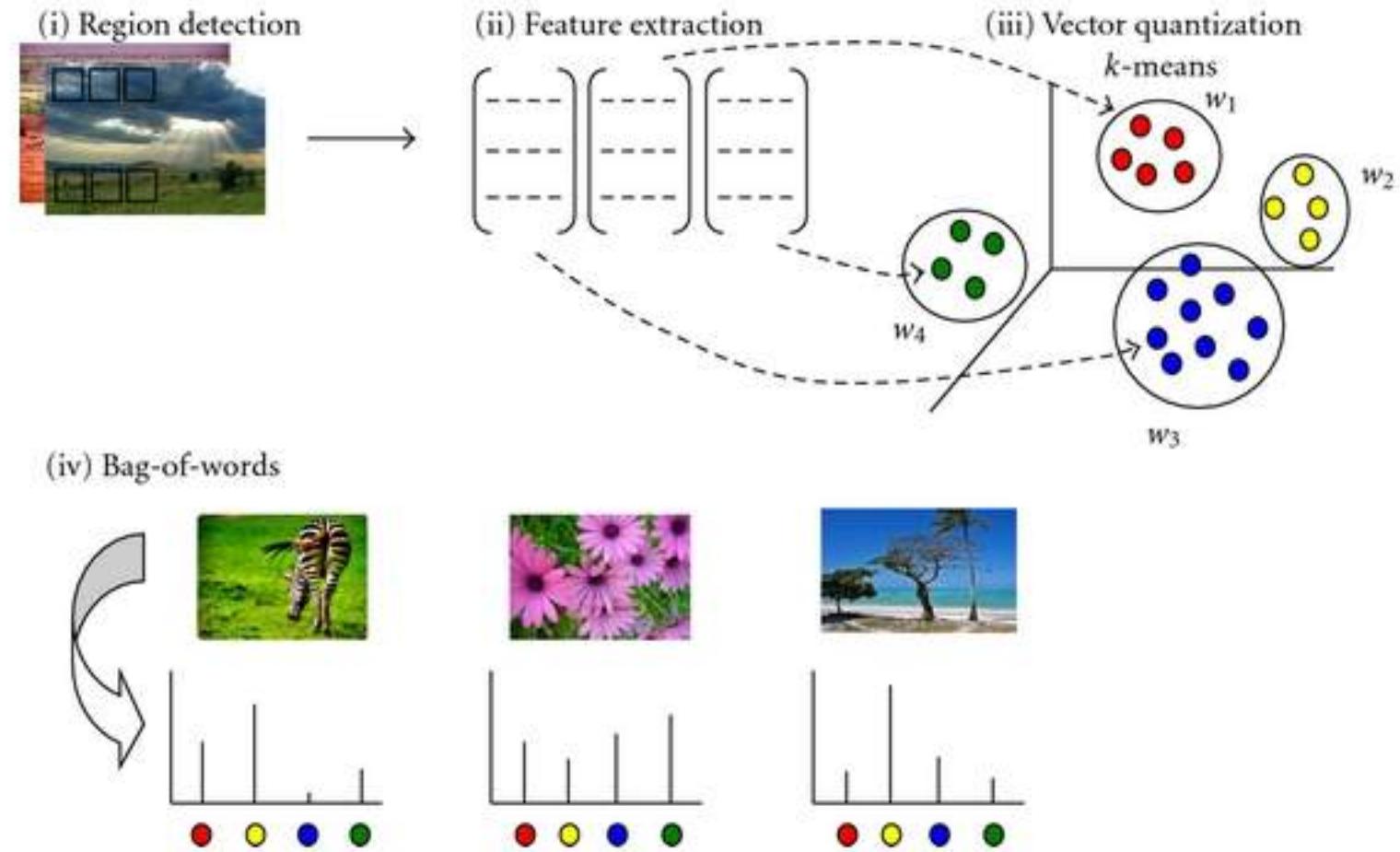
K Means и разные формы кластеров



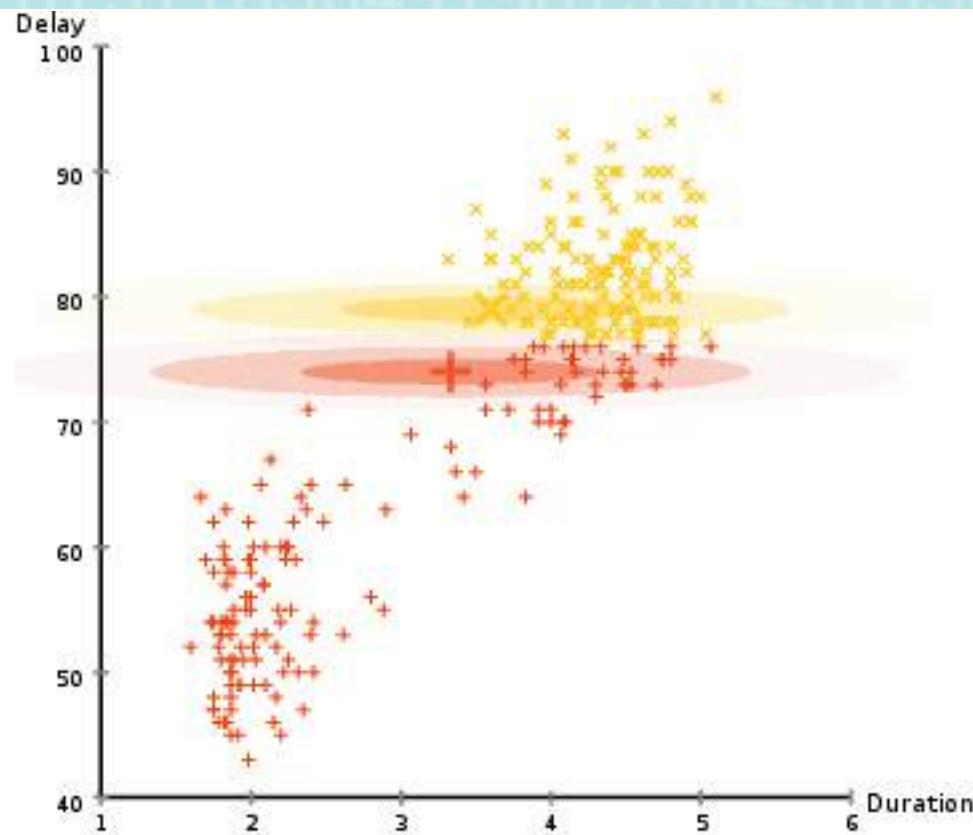
Пример: мешок визуальных слов



Пример: мешок визуальных слов



EM (Expectation-Maximization)



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

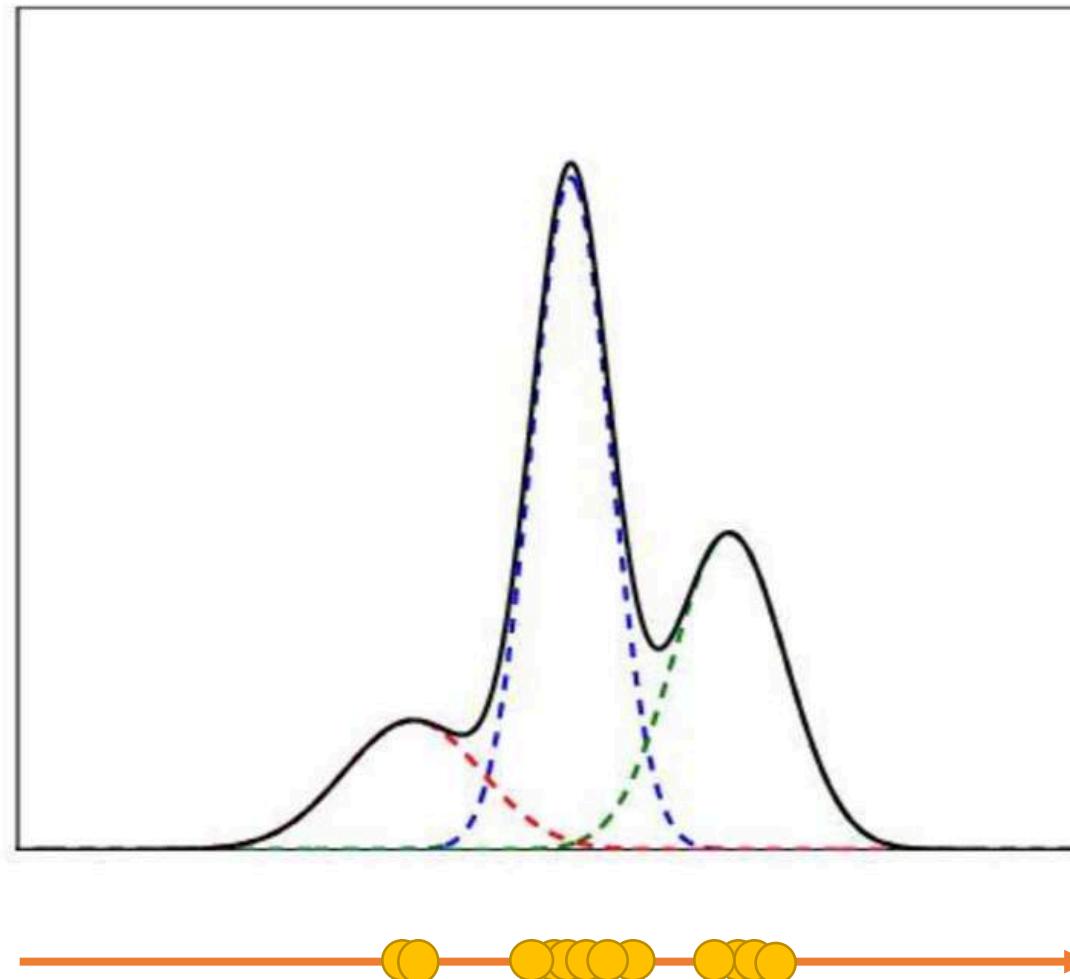
Постановка задачи: разделение смеси

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

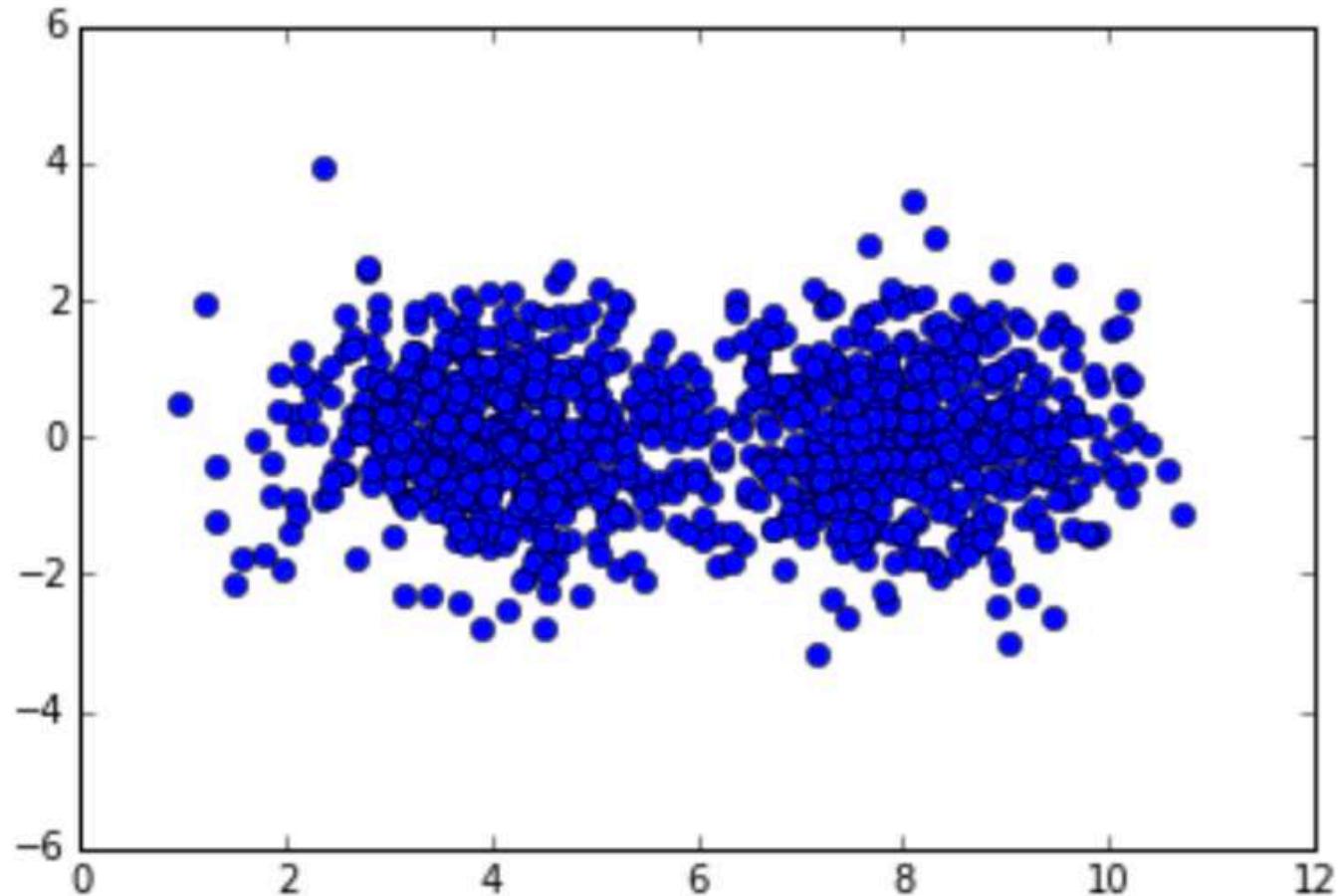
$$p_j(x) = \varphi(\theta_j; x)$$

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

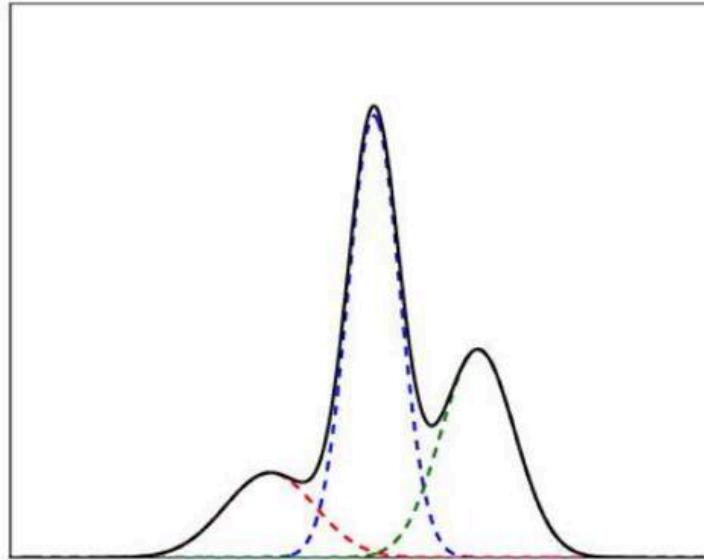
Как выглядит смесь распределений



Как выглядит смесь распределений



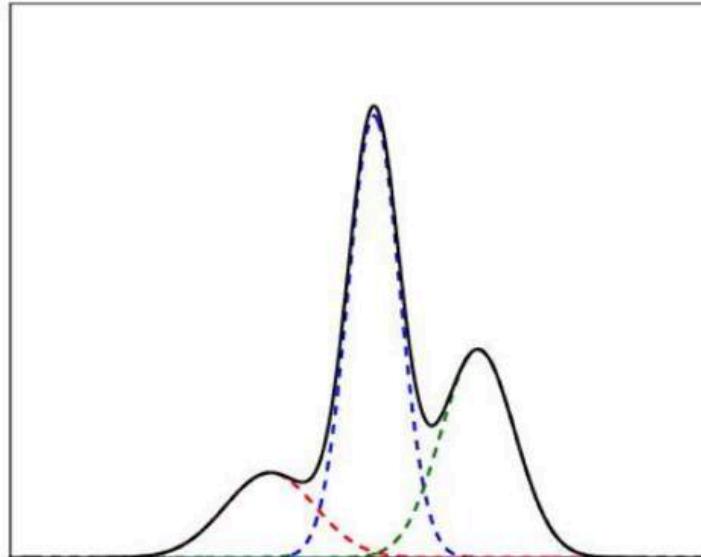
Почему не решить задачу «в лоб»



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

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

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

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

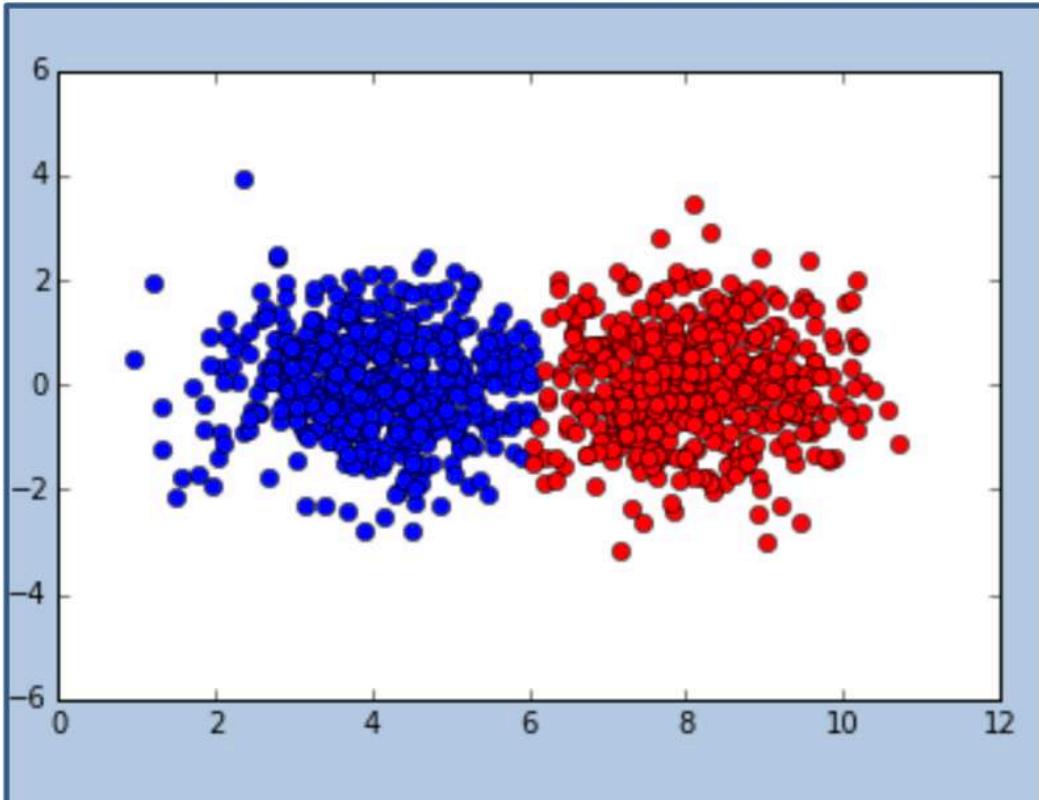
E-шаг:

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

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:

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

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

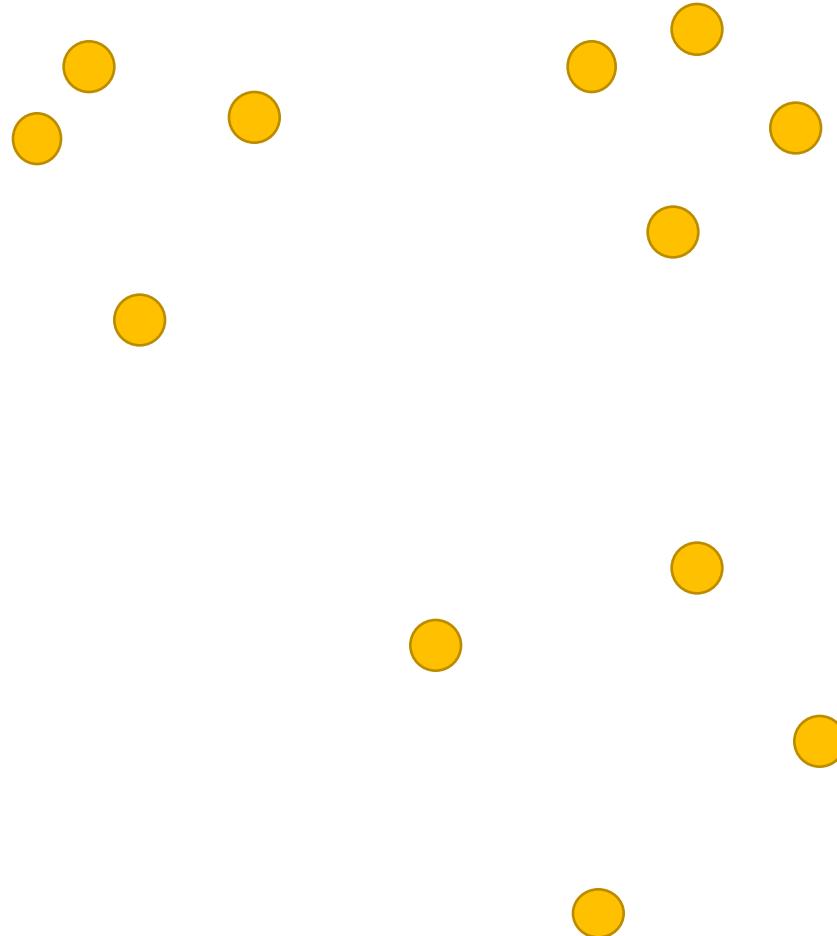
$$\Sigma_j = \frac{1}{N w_j - 1} \sum_{i=1}^N g_{ij} (x_i - \mu_j)(x_i - \mu_j)^T$$

III. Иерархические и графовые методы

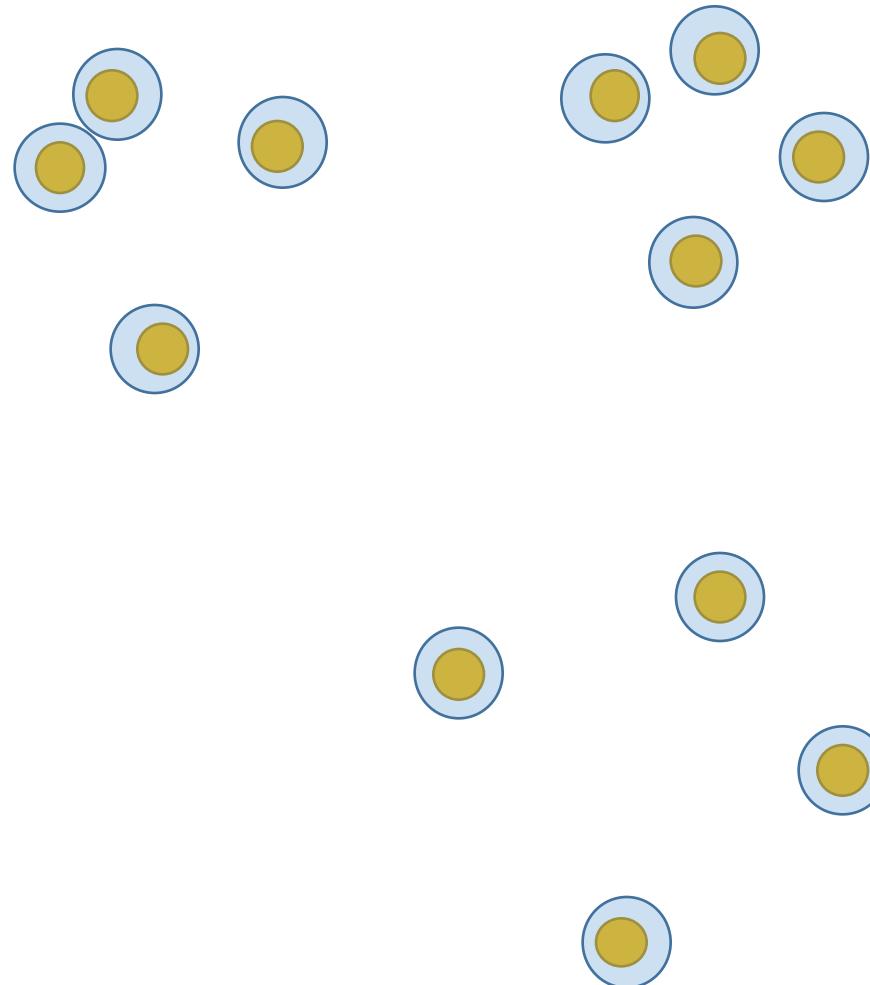
Иерархическая кластеризация

- Agglomerative clustering
- Divisive clustering

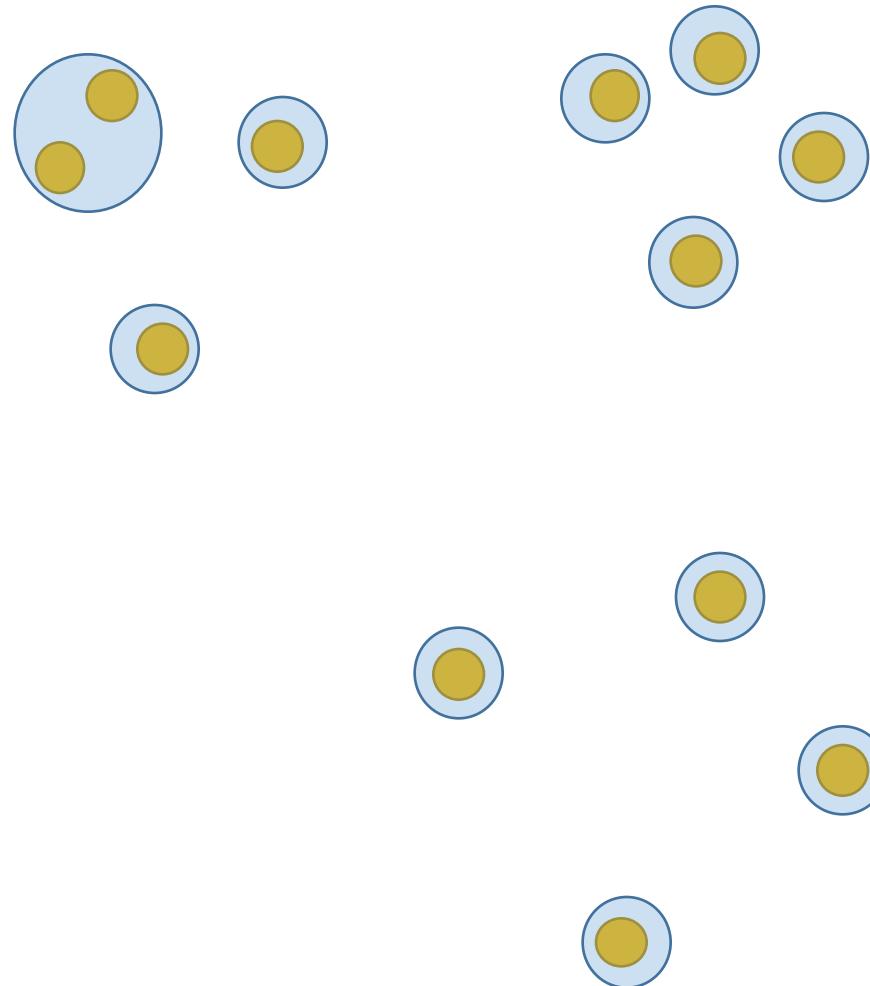
Агломеративная кластеризация



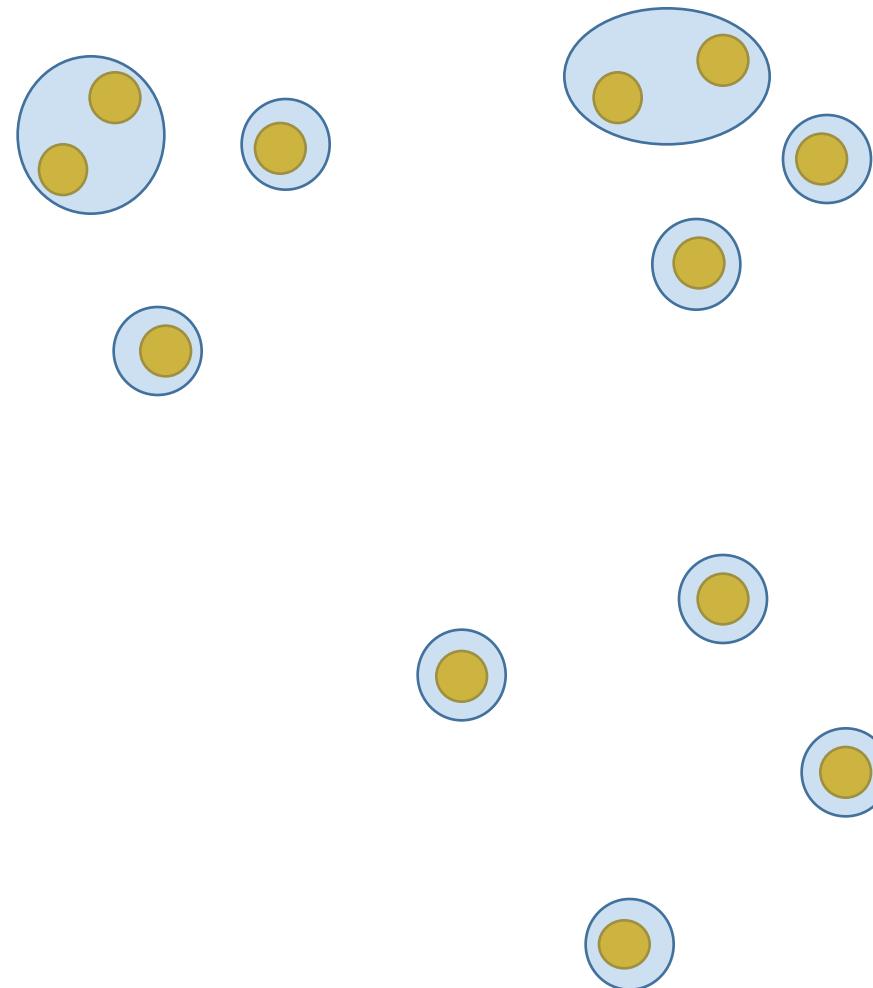
Агломеративная кластеризация



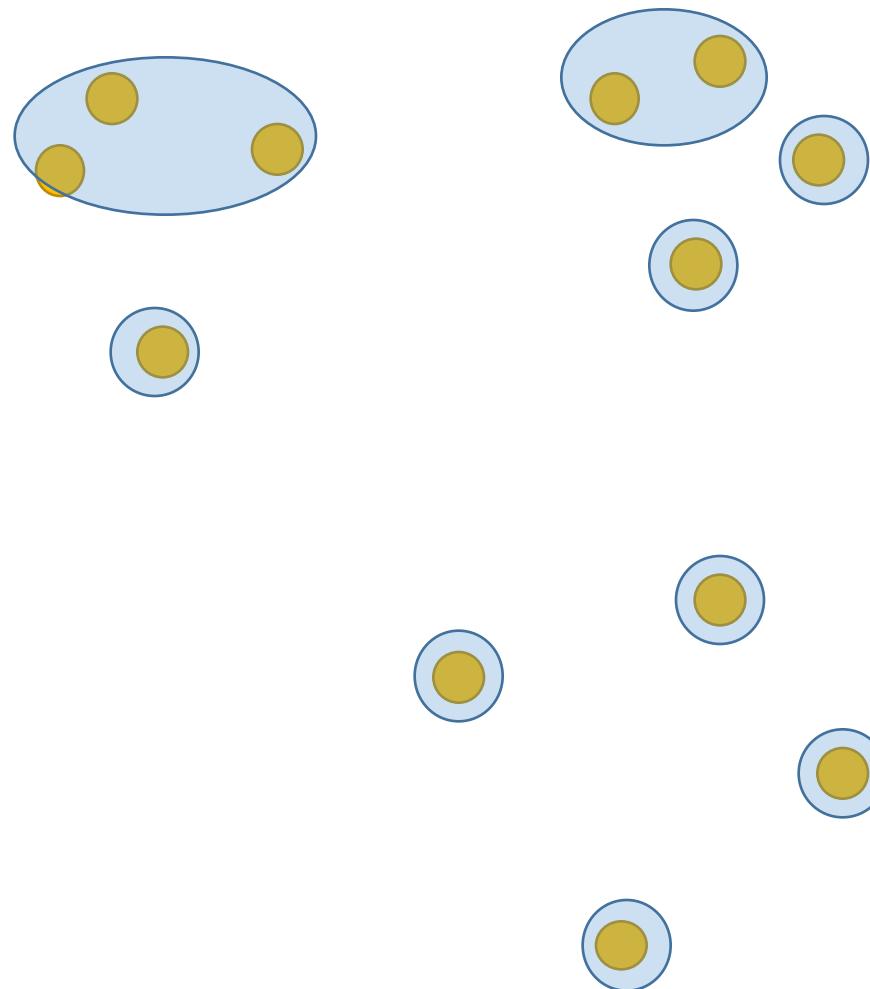
Агломеративная кластеризация



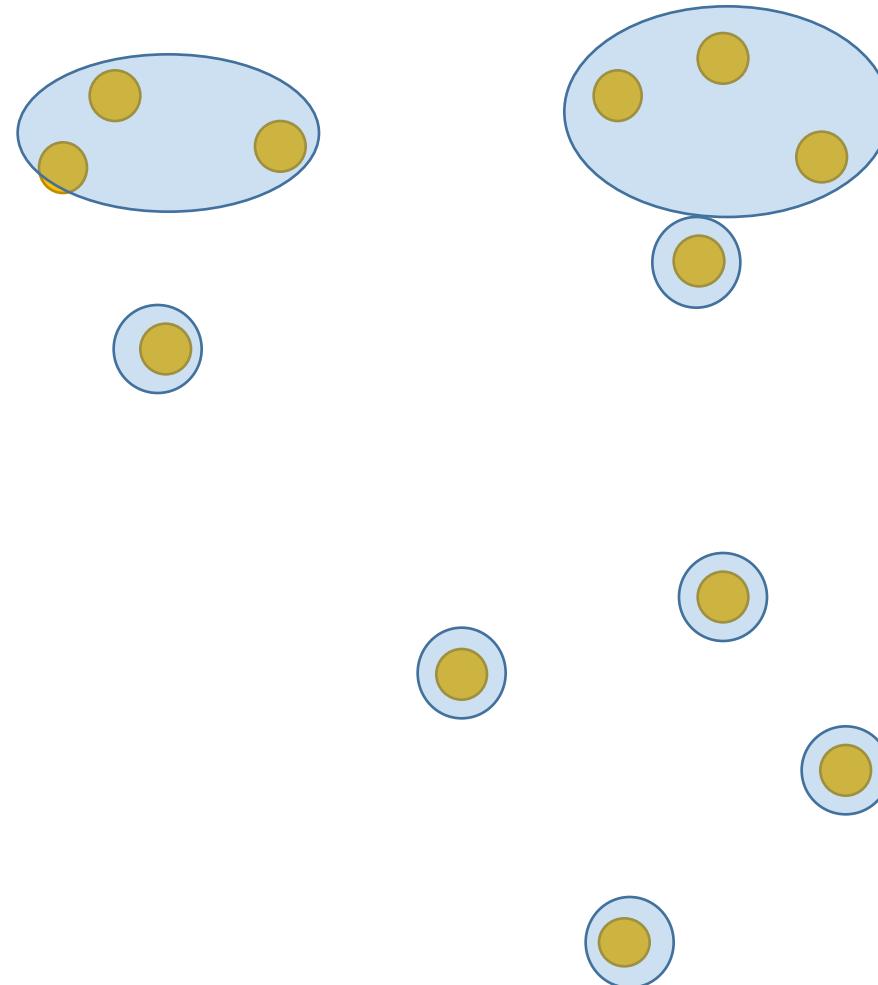
Агломеративная кластеризация



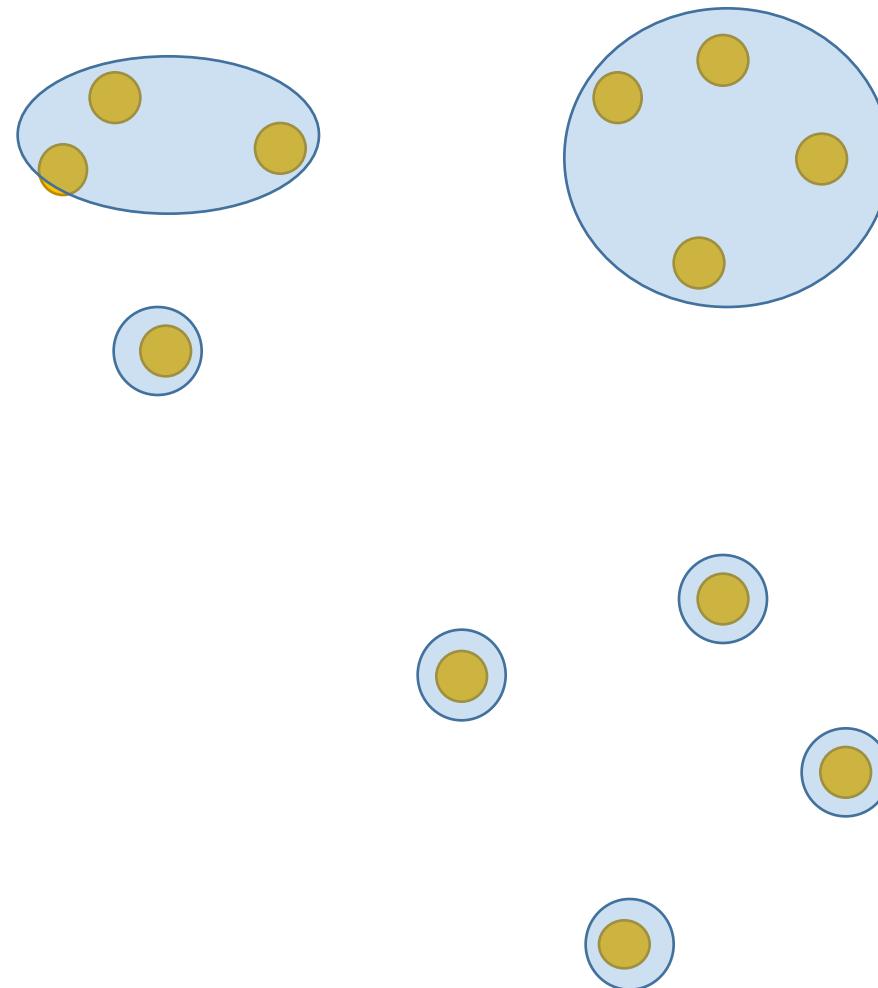
Агломеративная кластеризация



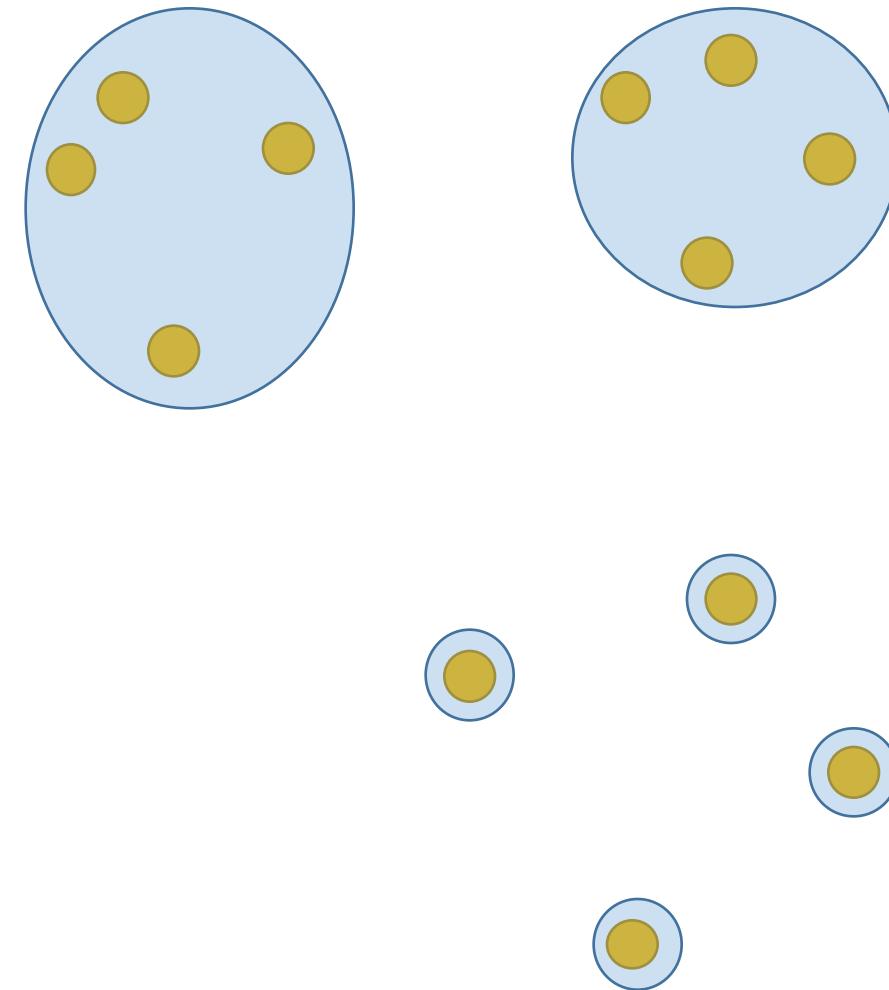
Агломеративная кластеризация



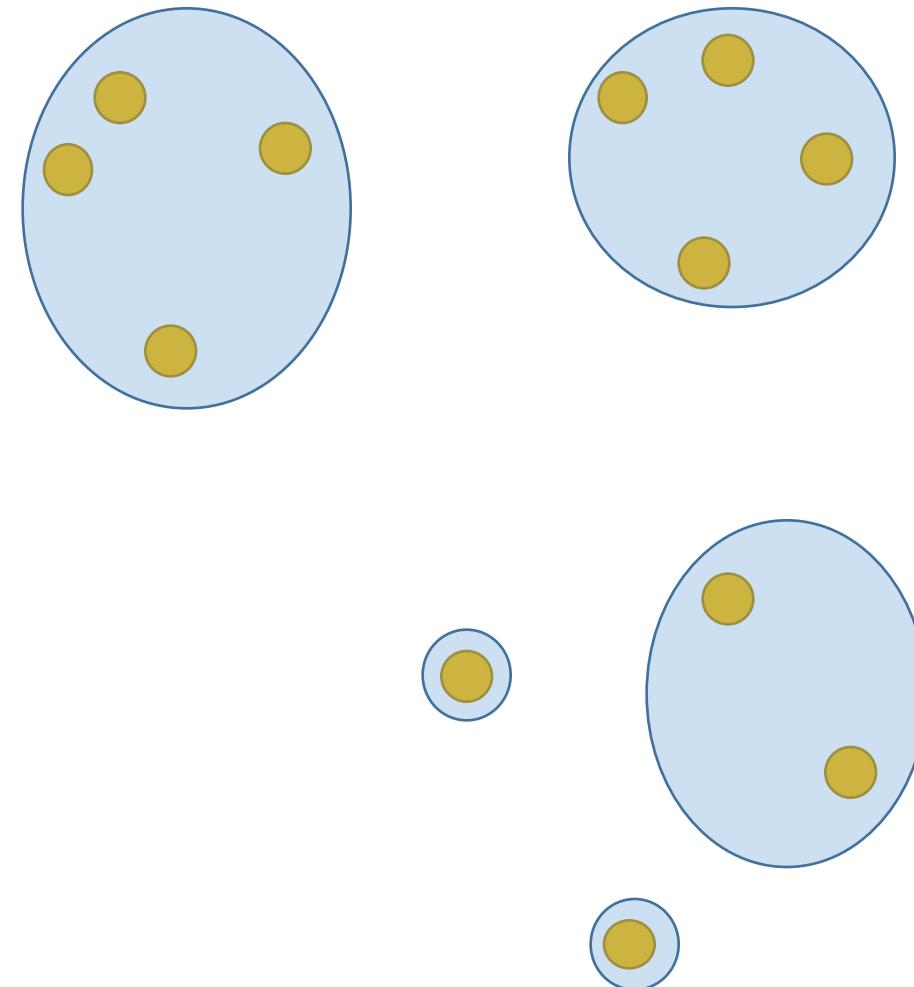
Агломеративная кластеризация



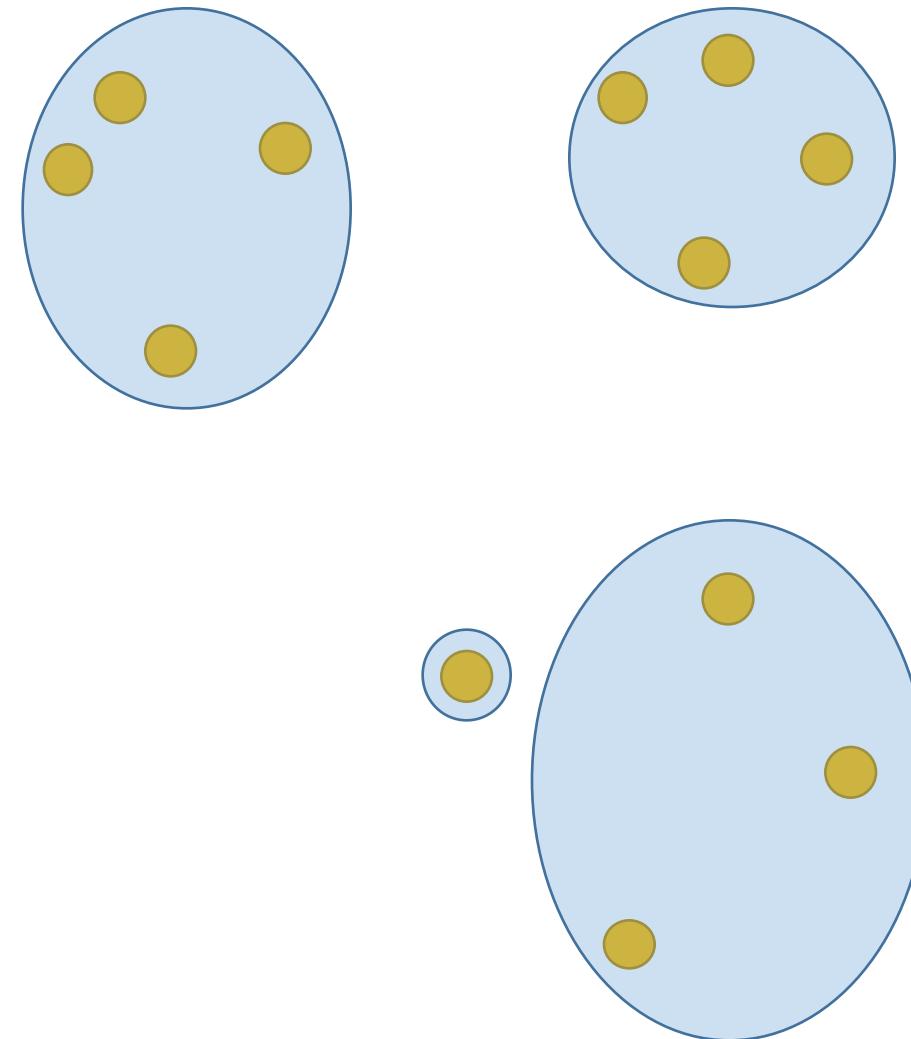
Агломеративная кластеризация



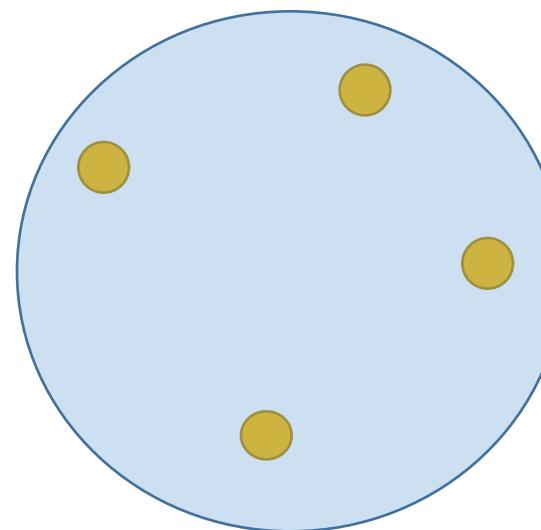
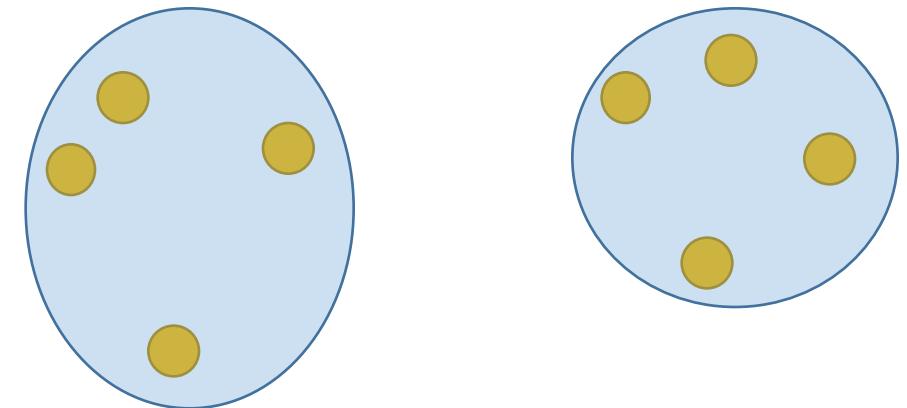
Агломеративная кластеризация



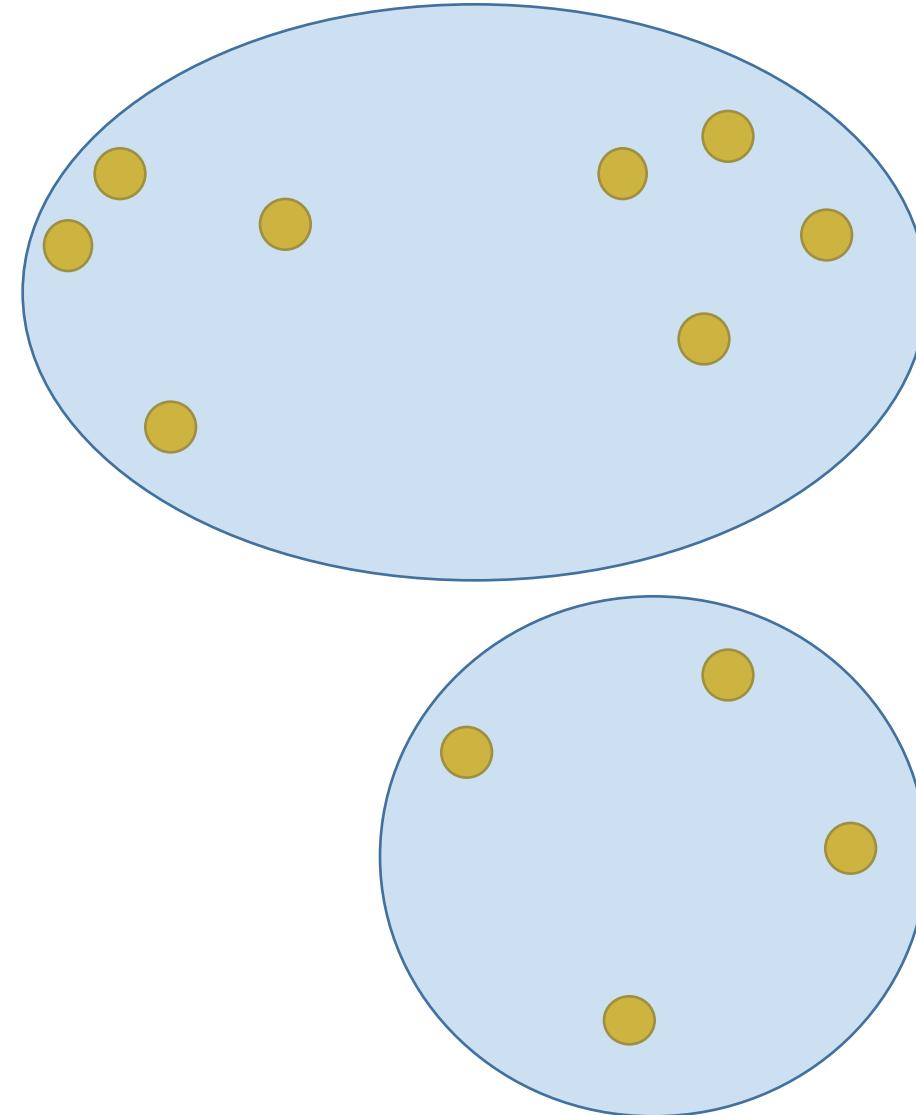
Агломеративная кластеризация



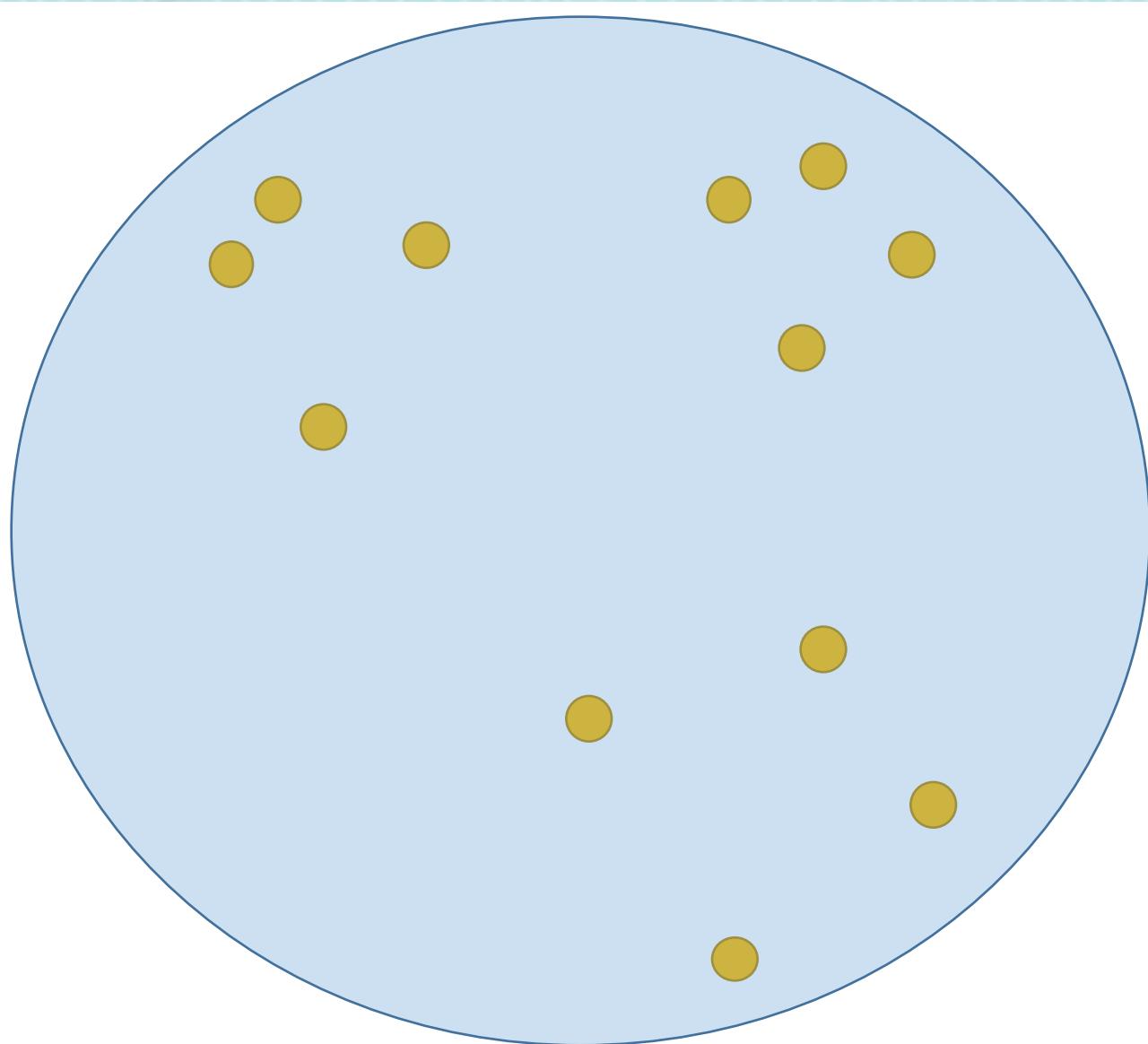
Агломеративная кластеризация



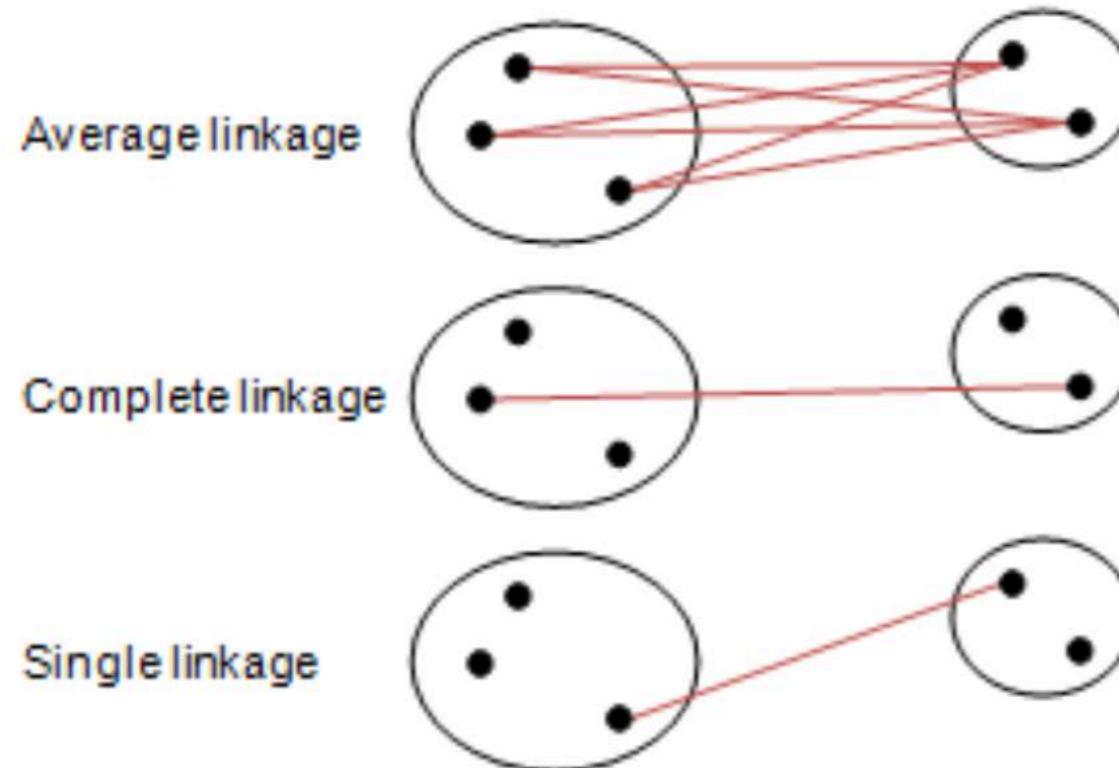
Агломеративная кластеризация



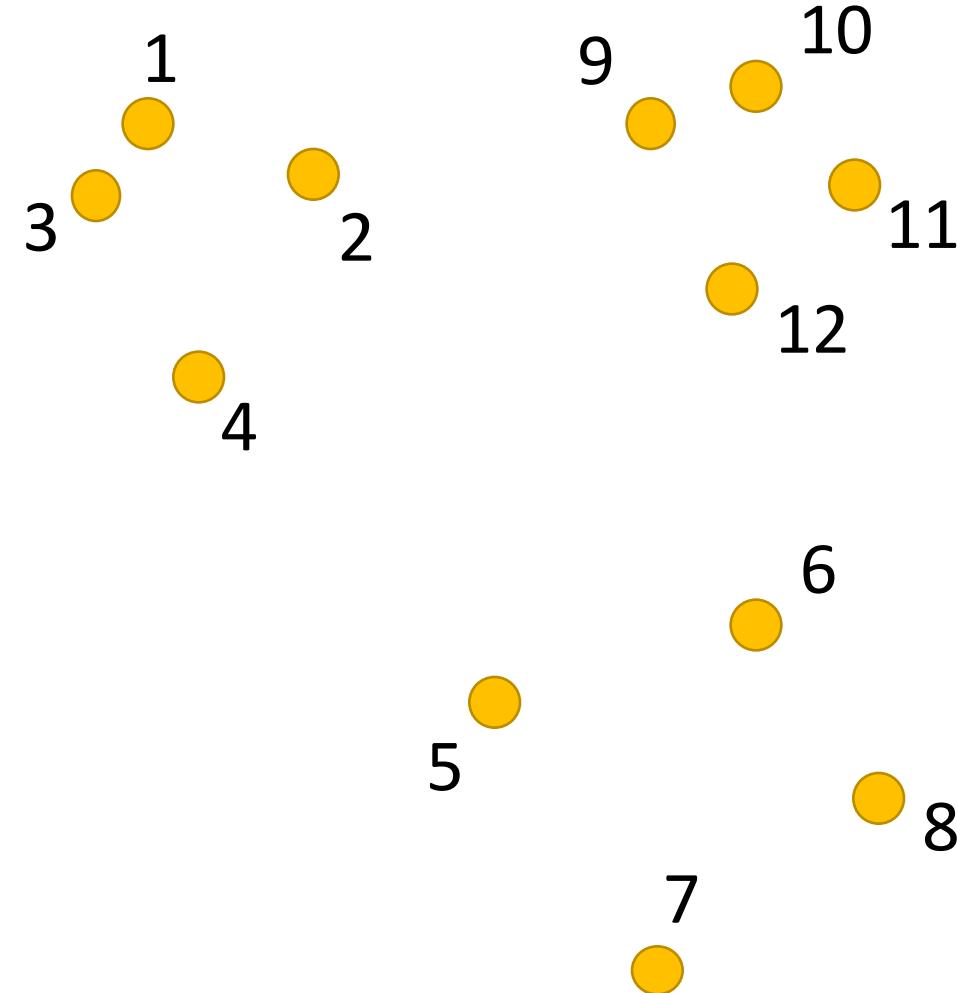
Агломеративная кластеризация



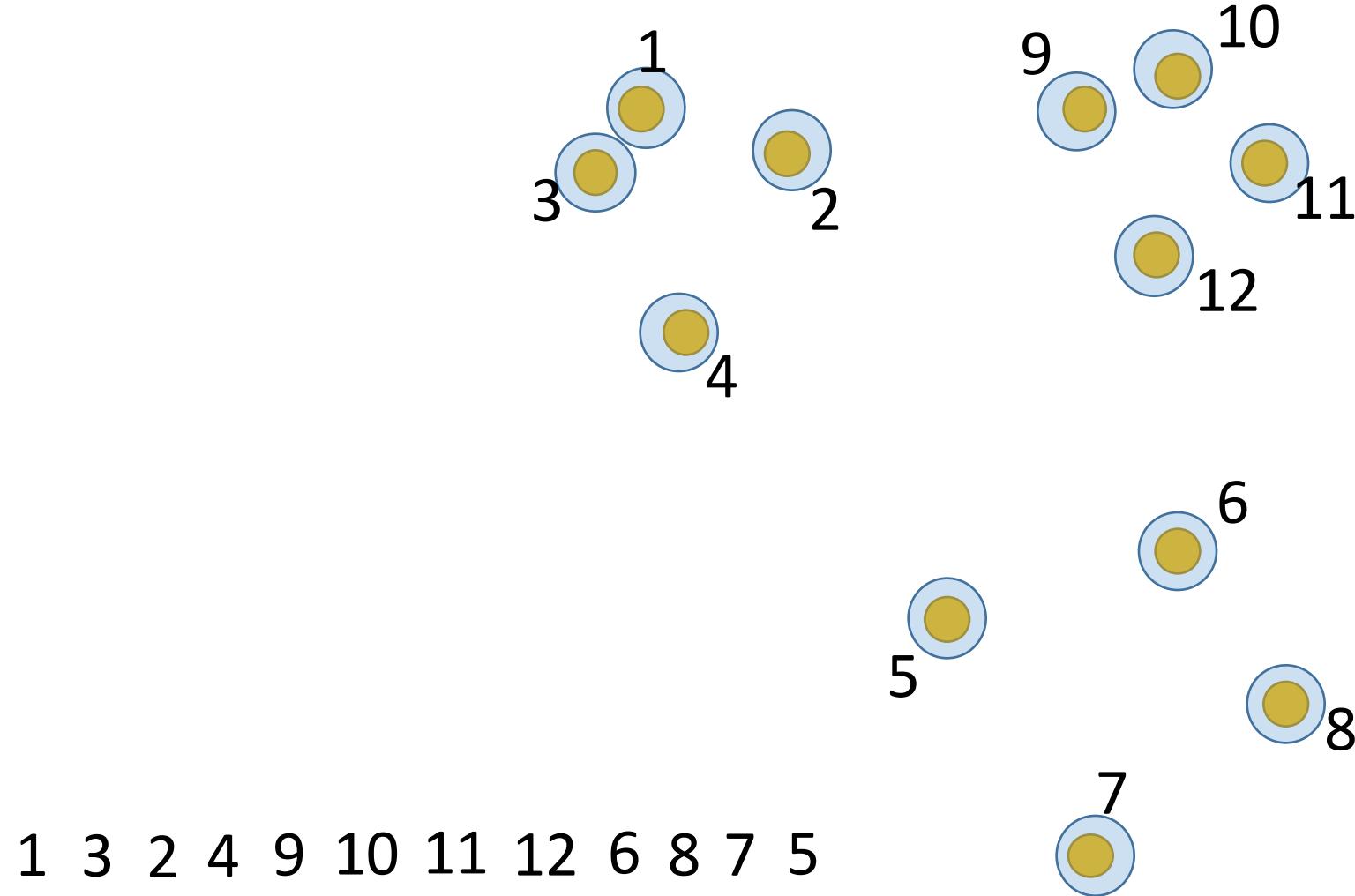
Расстояния между кластерами



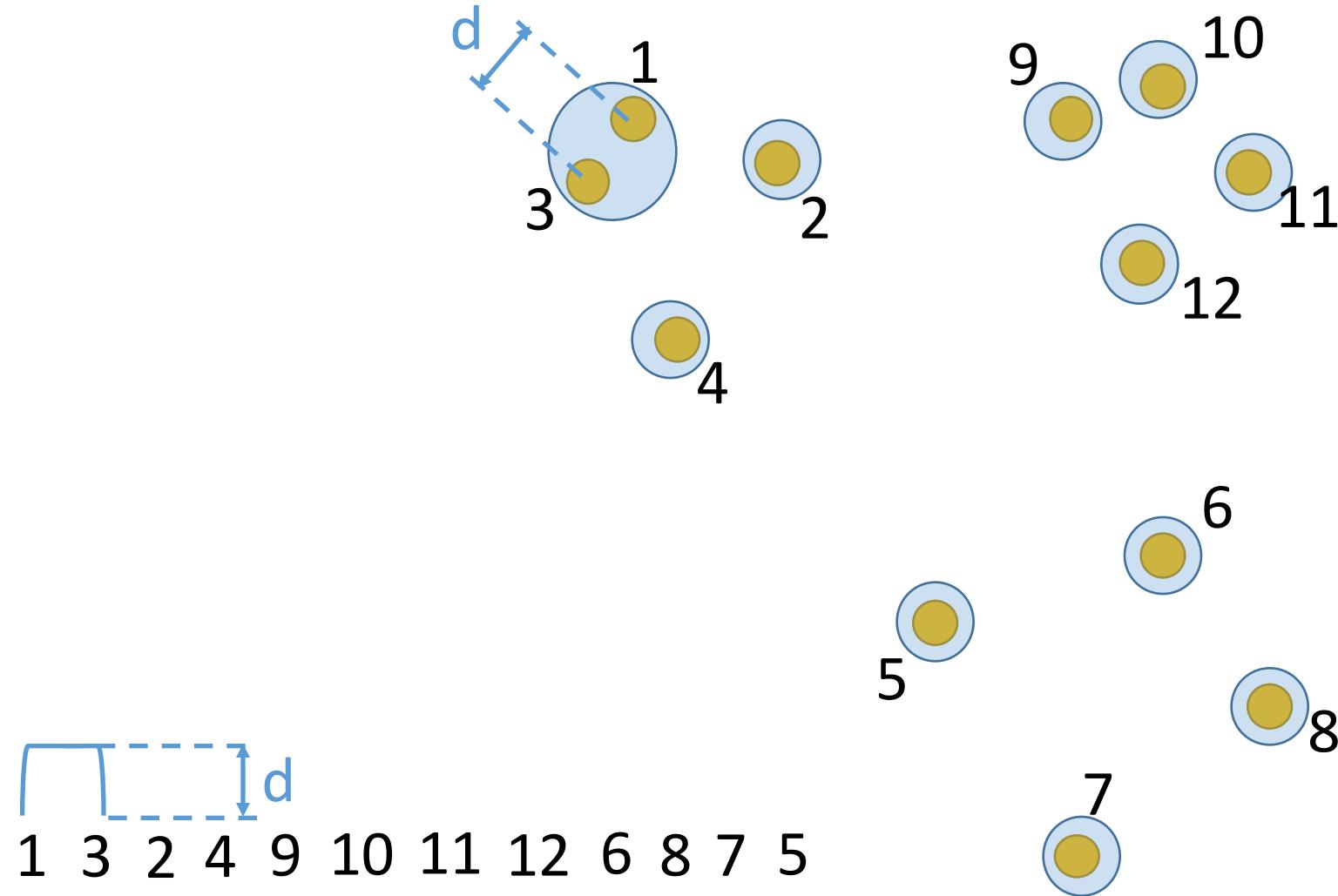
Дендрограмма



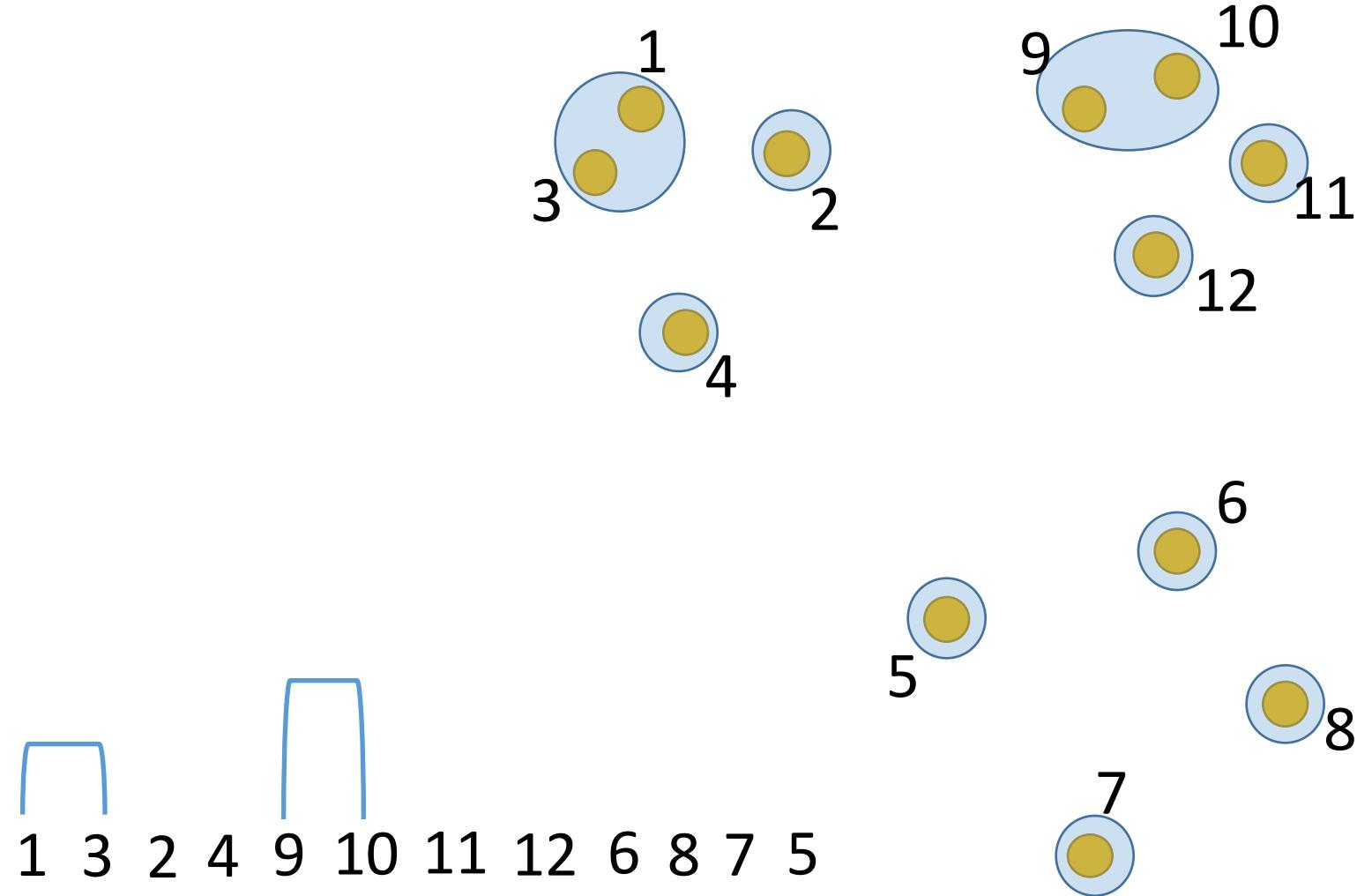
Дендрограмма



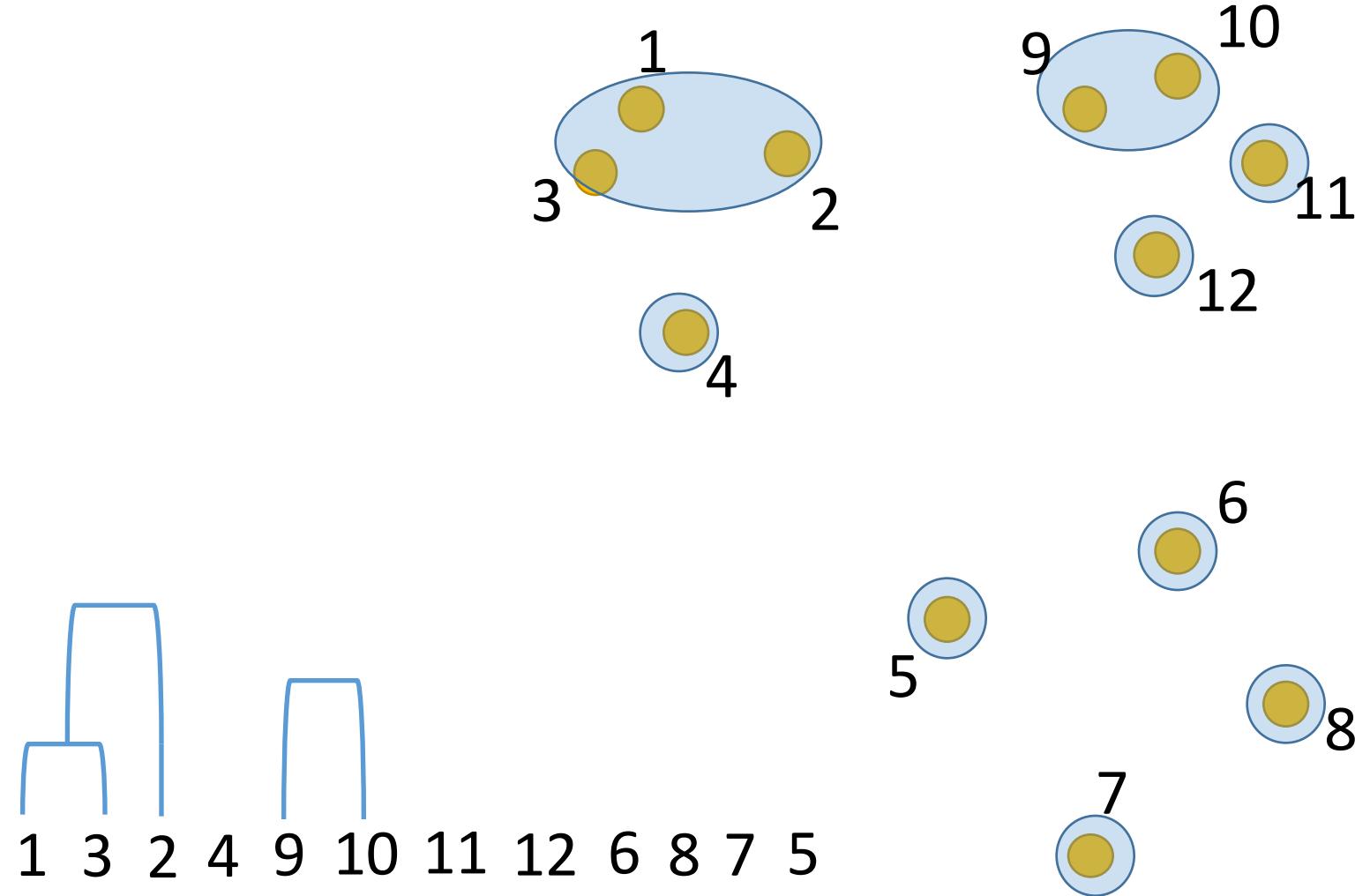
Дендрограмма



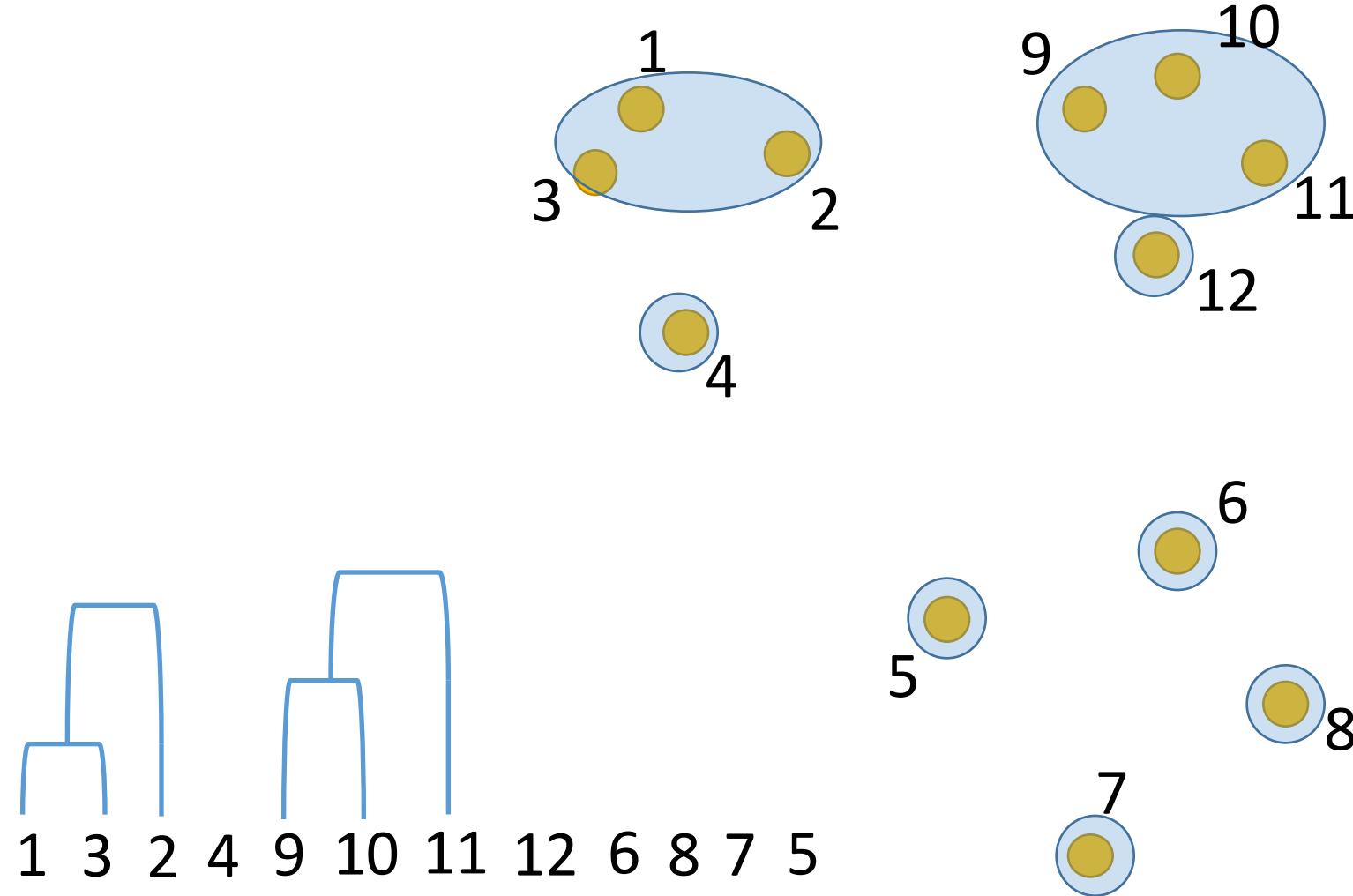
Дендрограмма



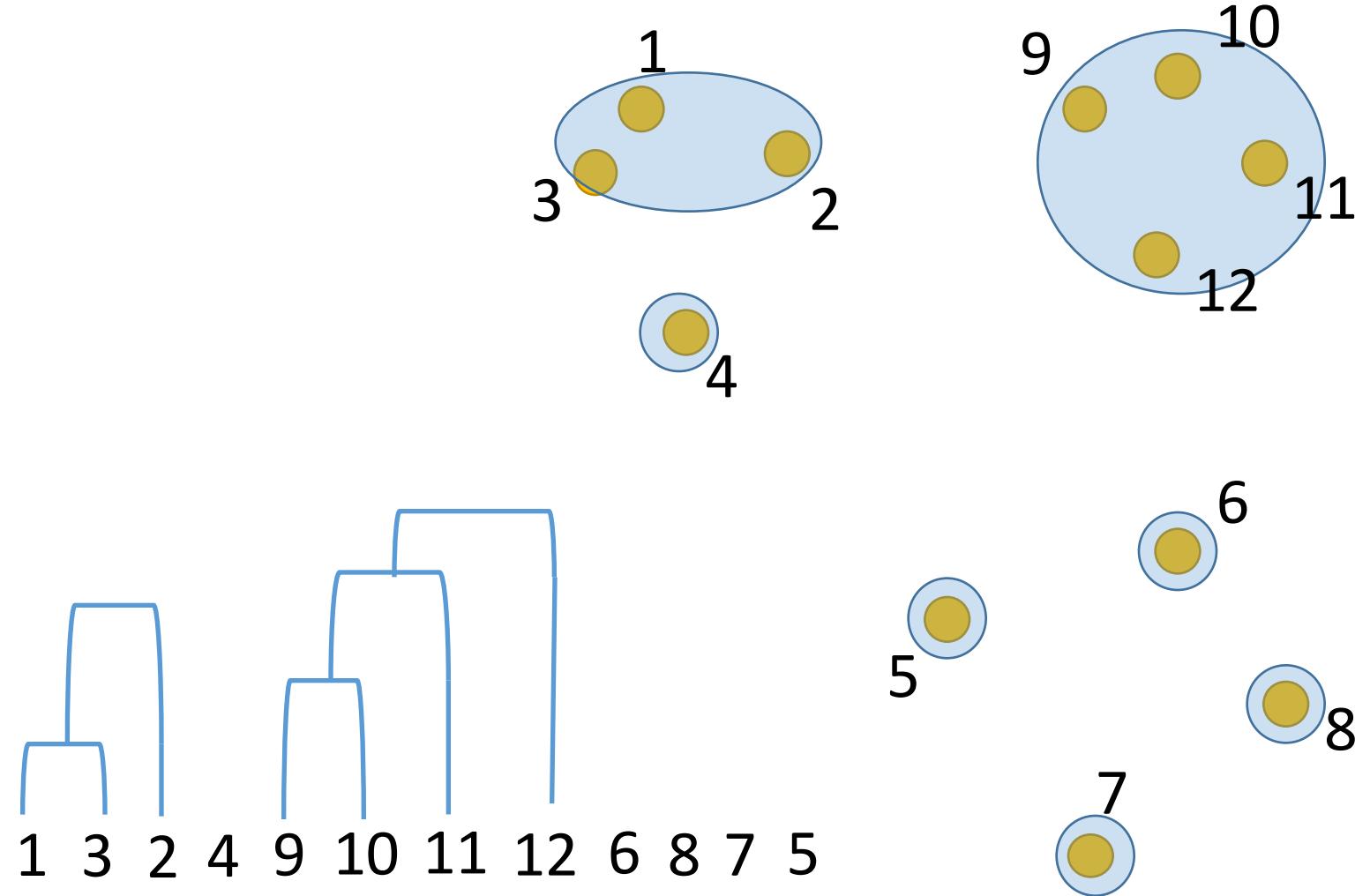
Дендрограмма



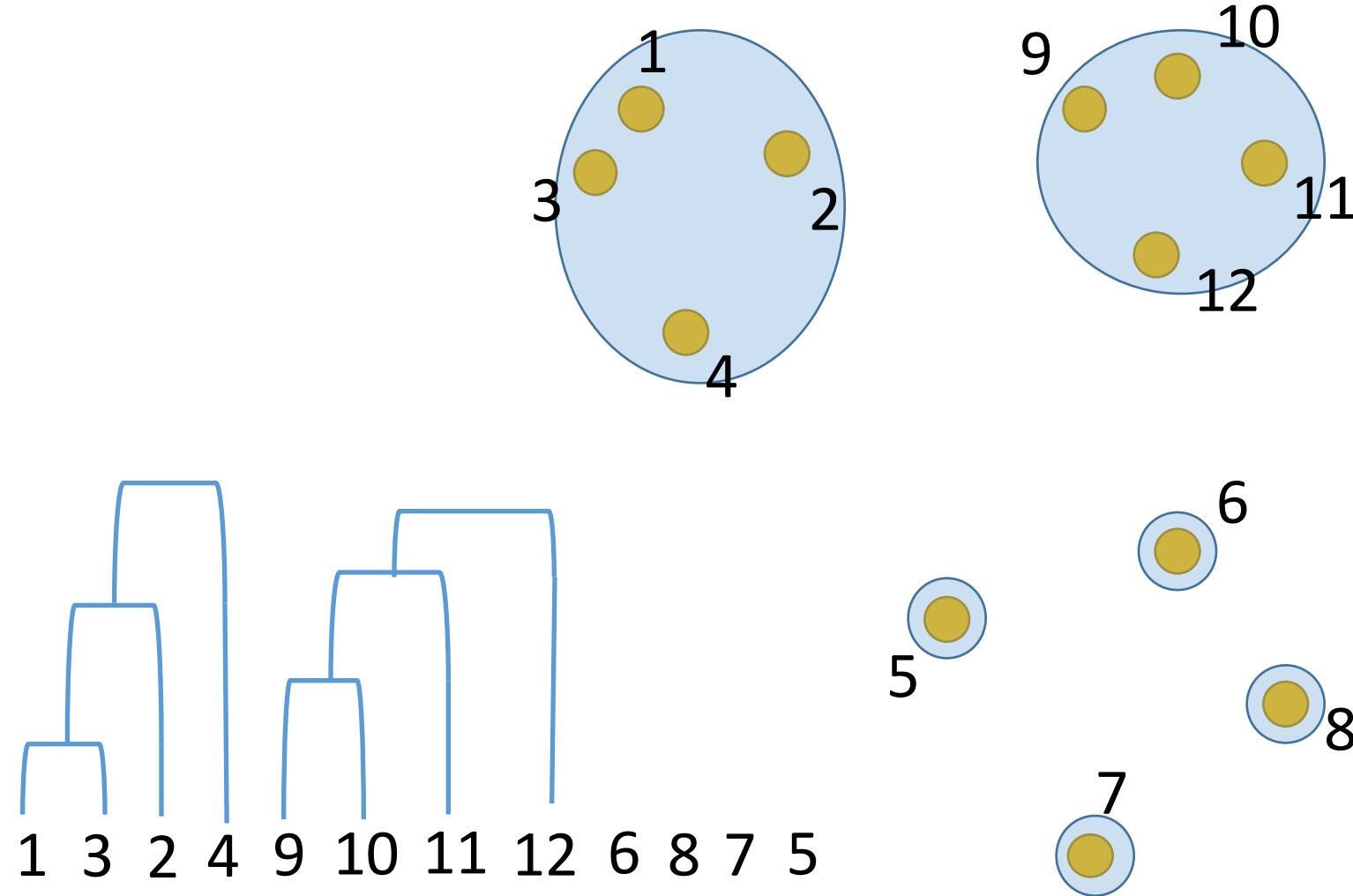
Дендрограмма



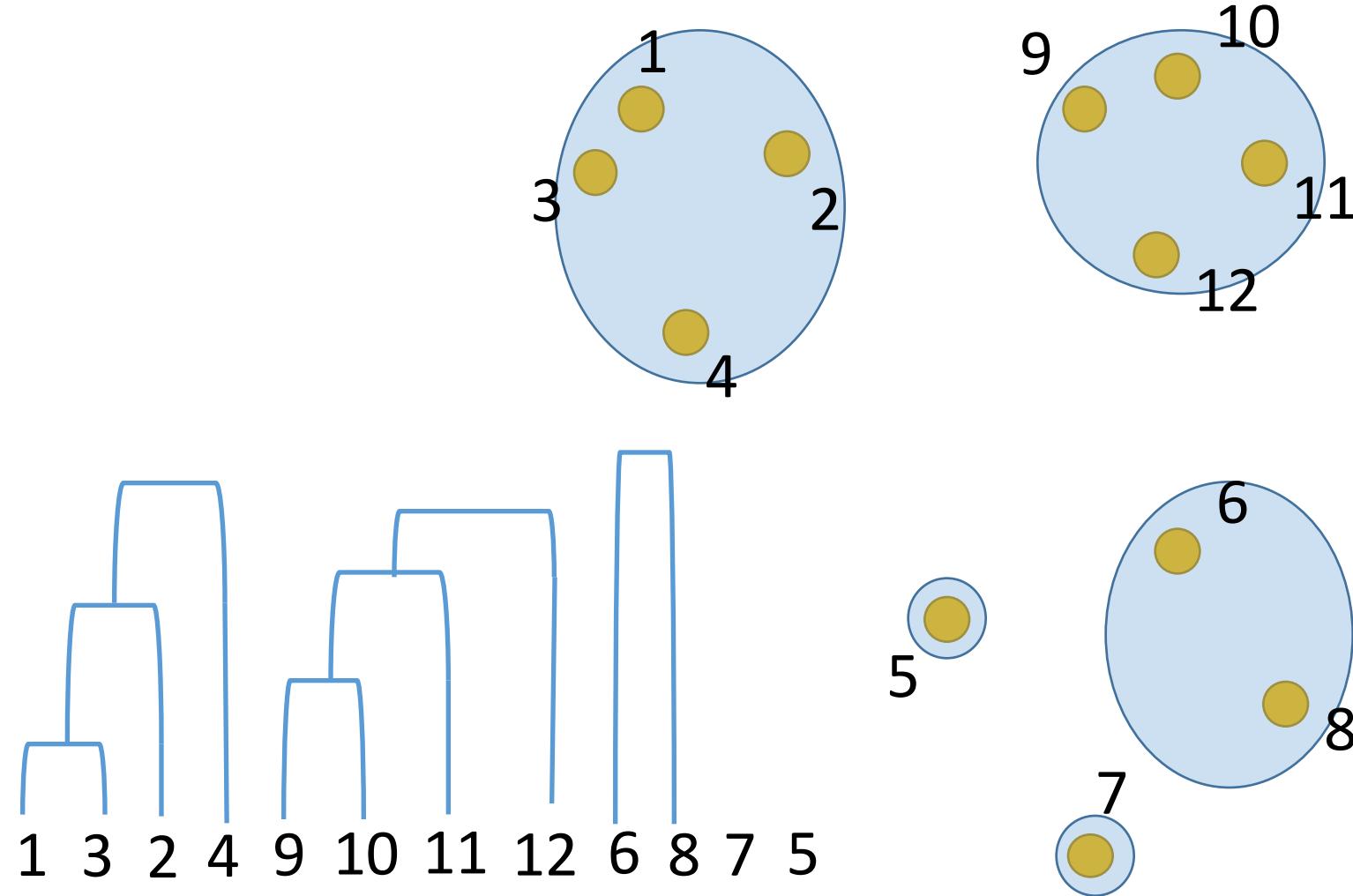
Дендрограмма



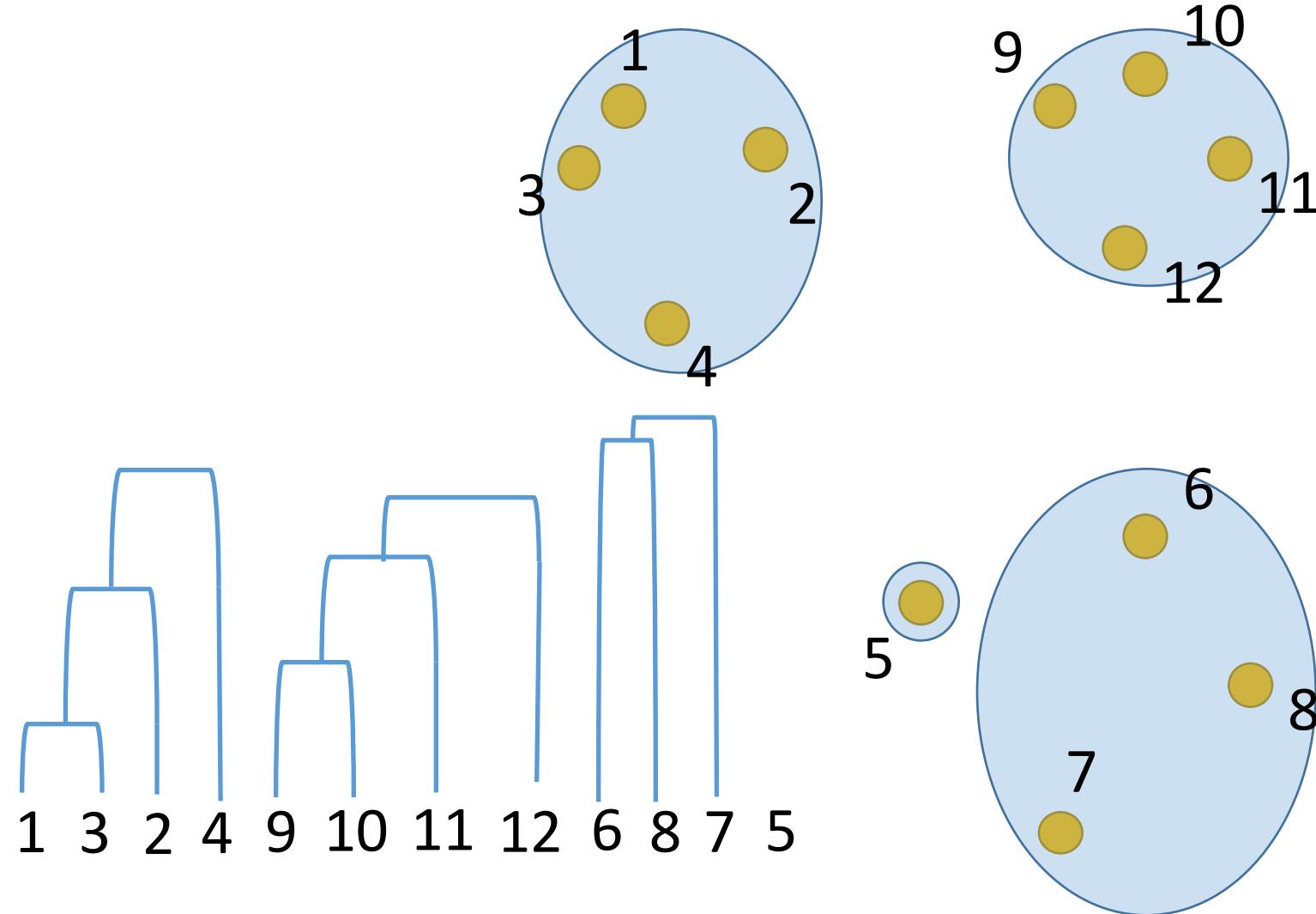
Дендрограмма



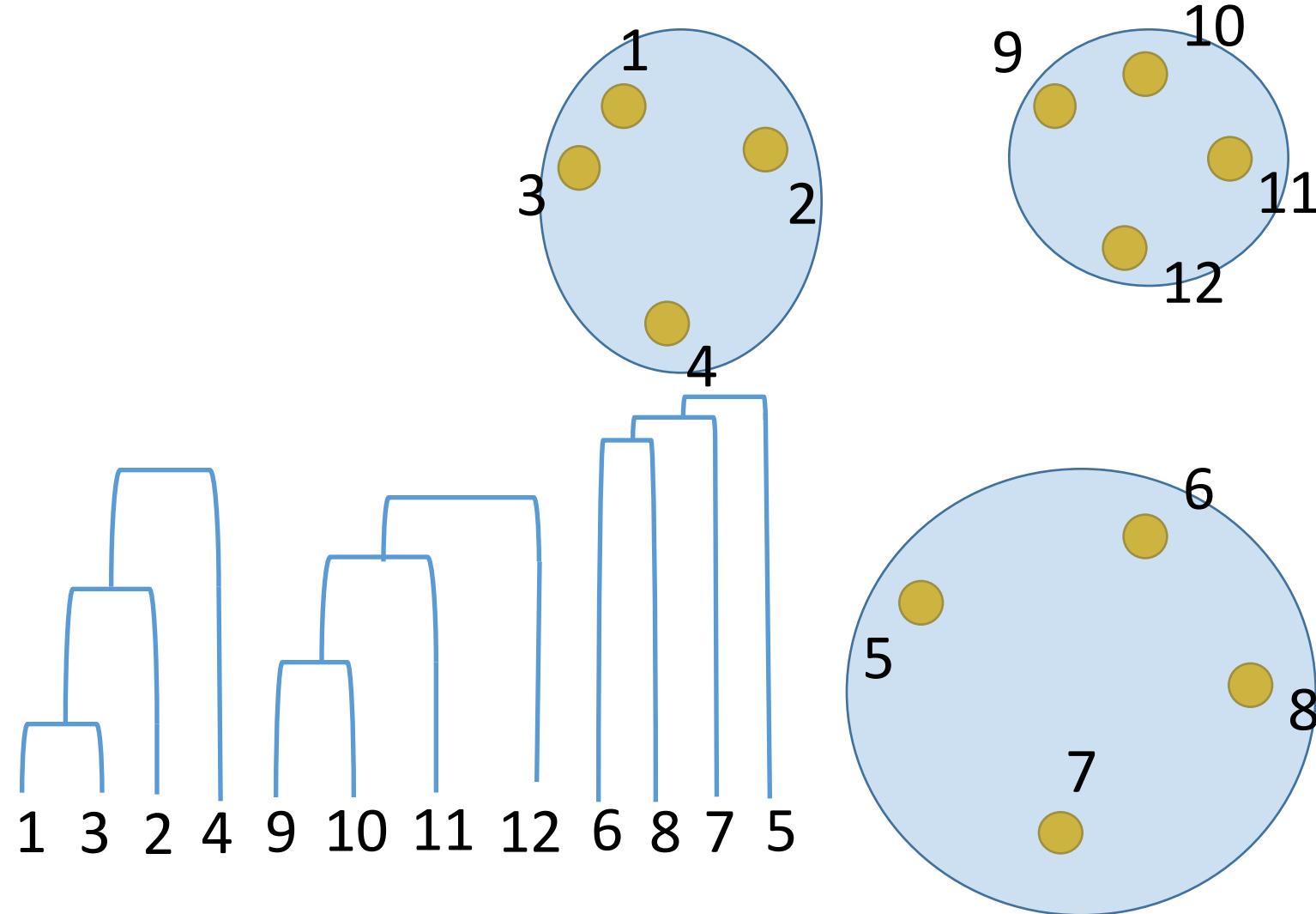
Дендрограмма



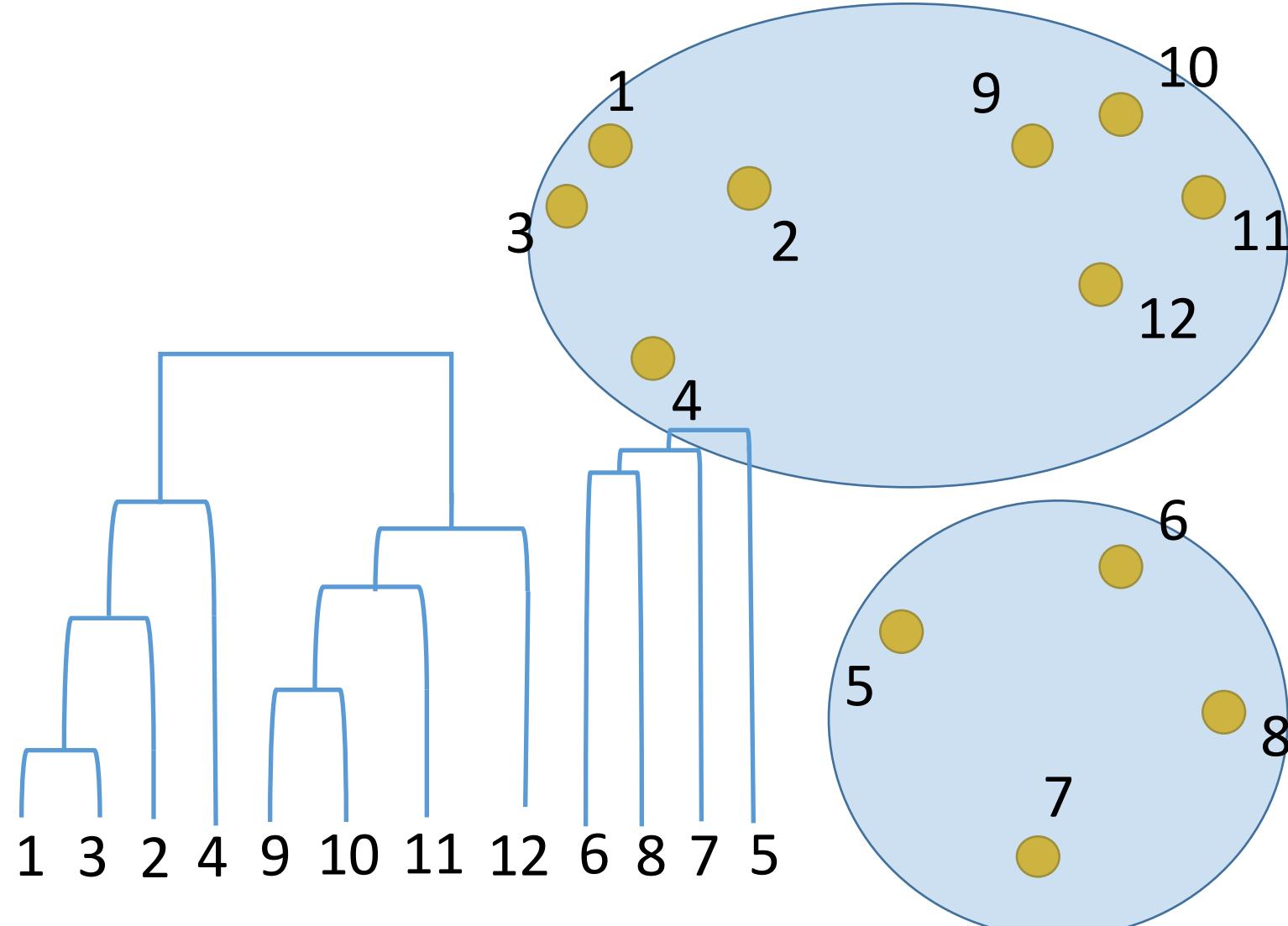
Дендрограмма



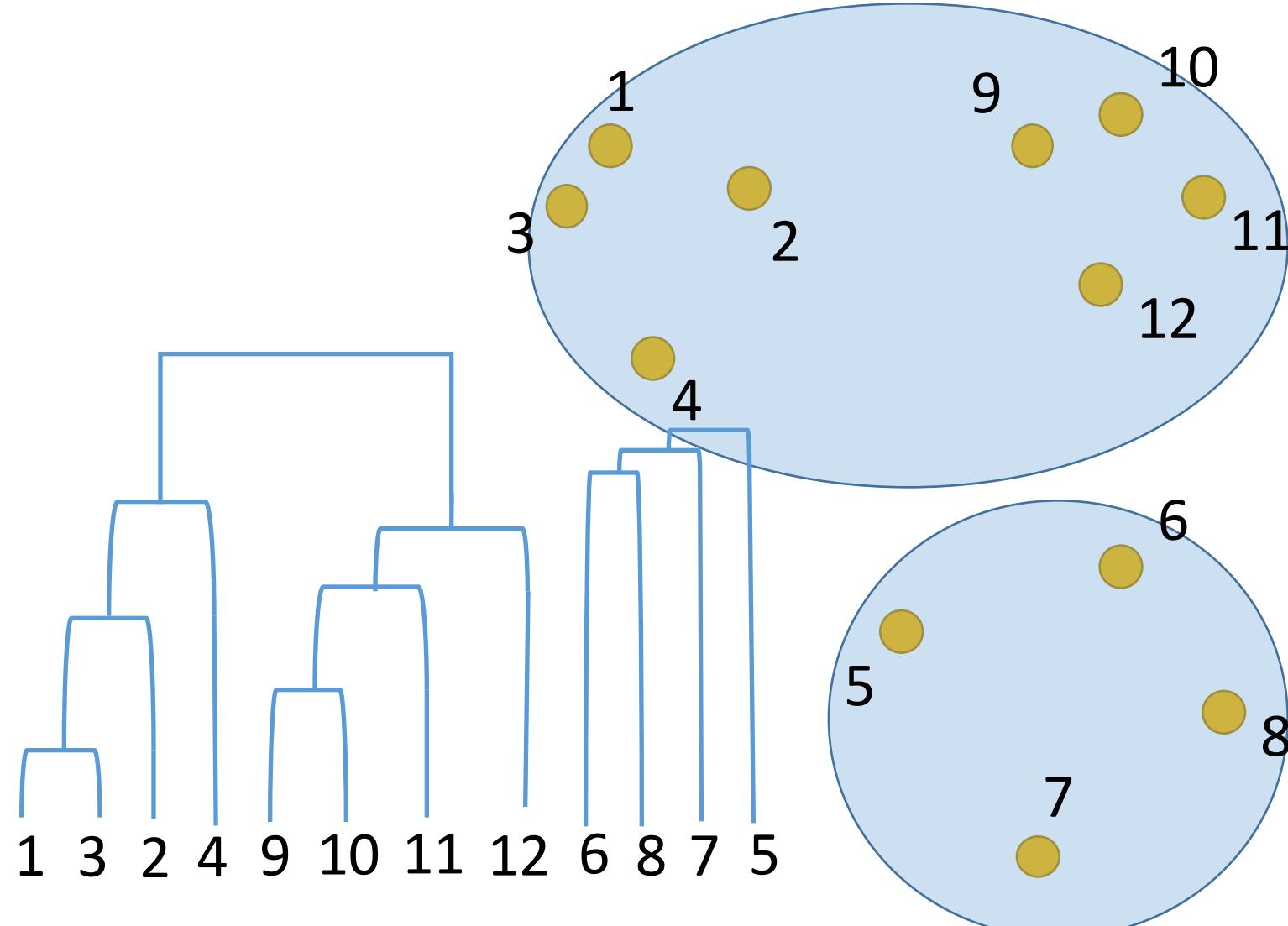
Дендрограмма



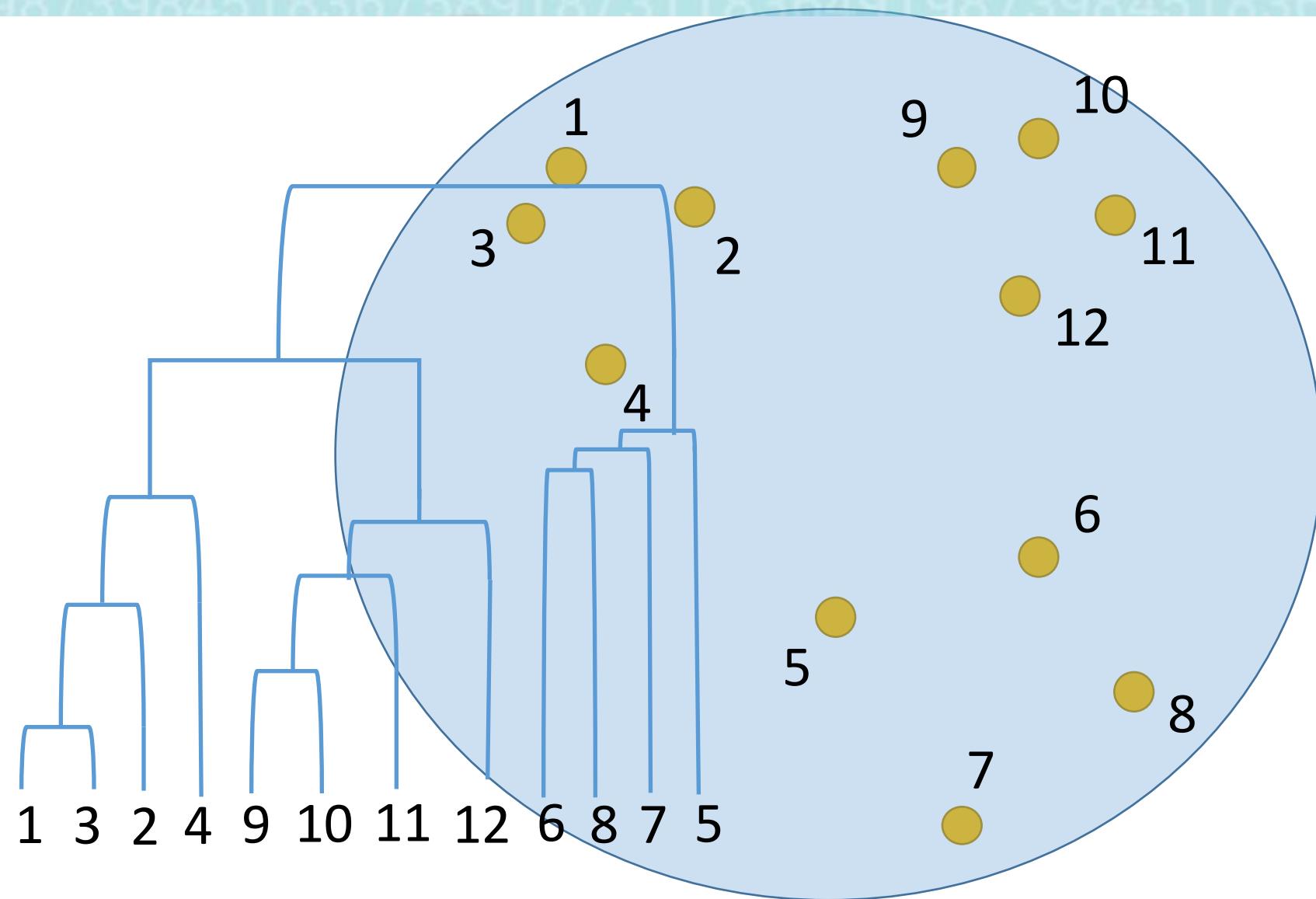
Дендрограмма



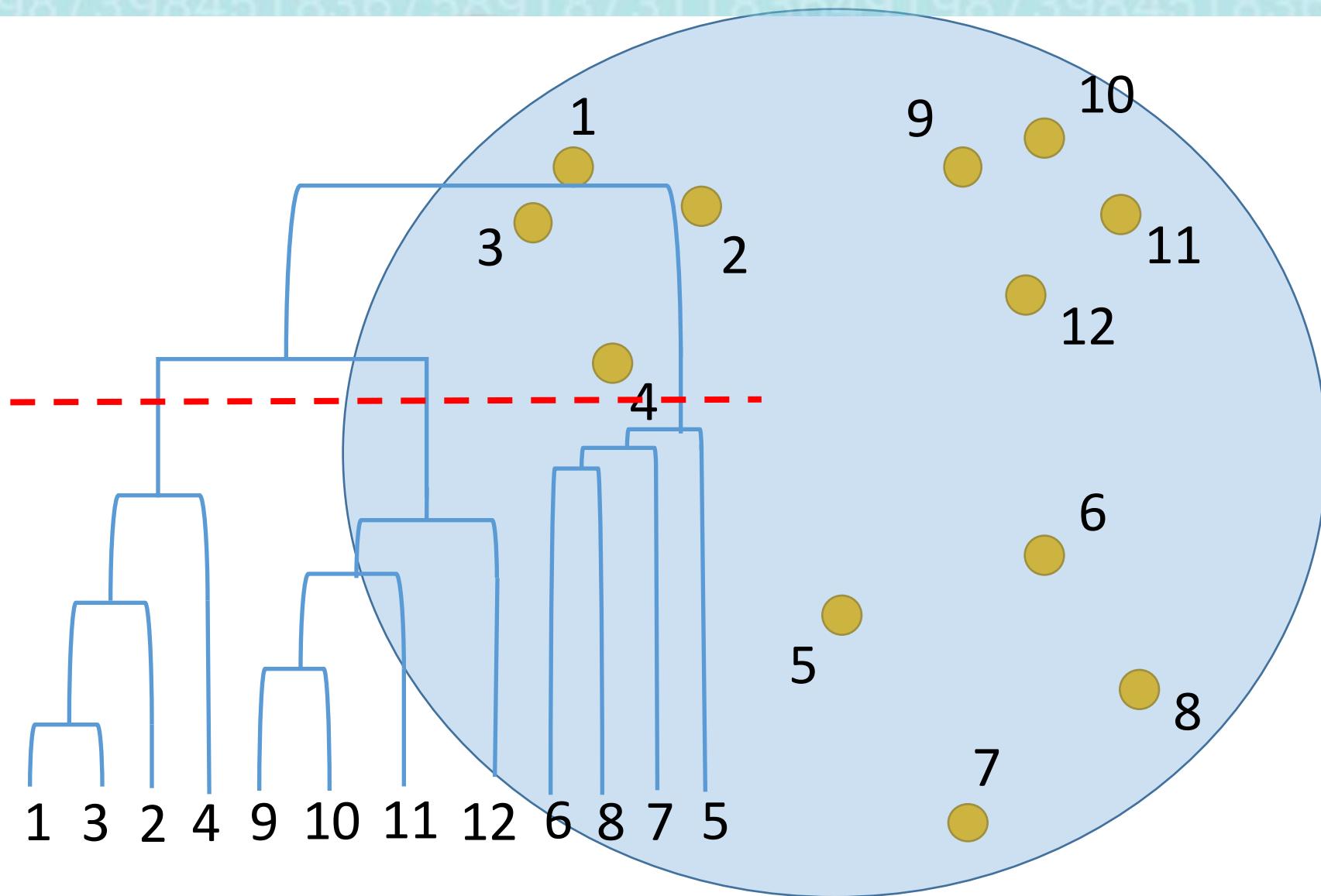
Дендрограмма



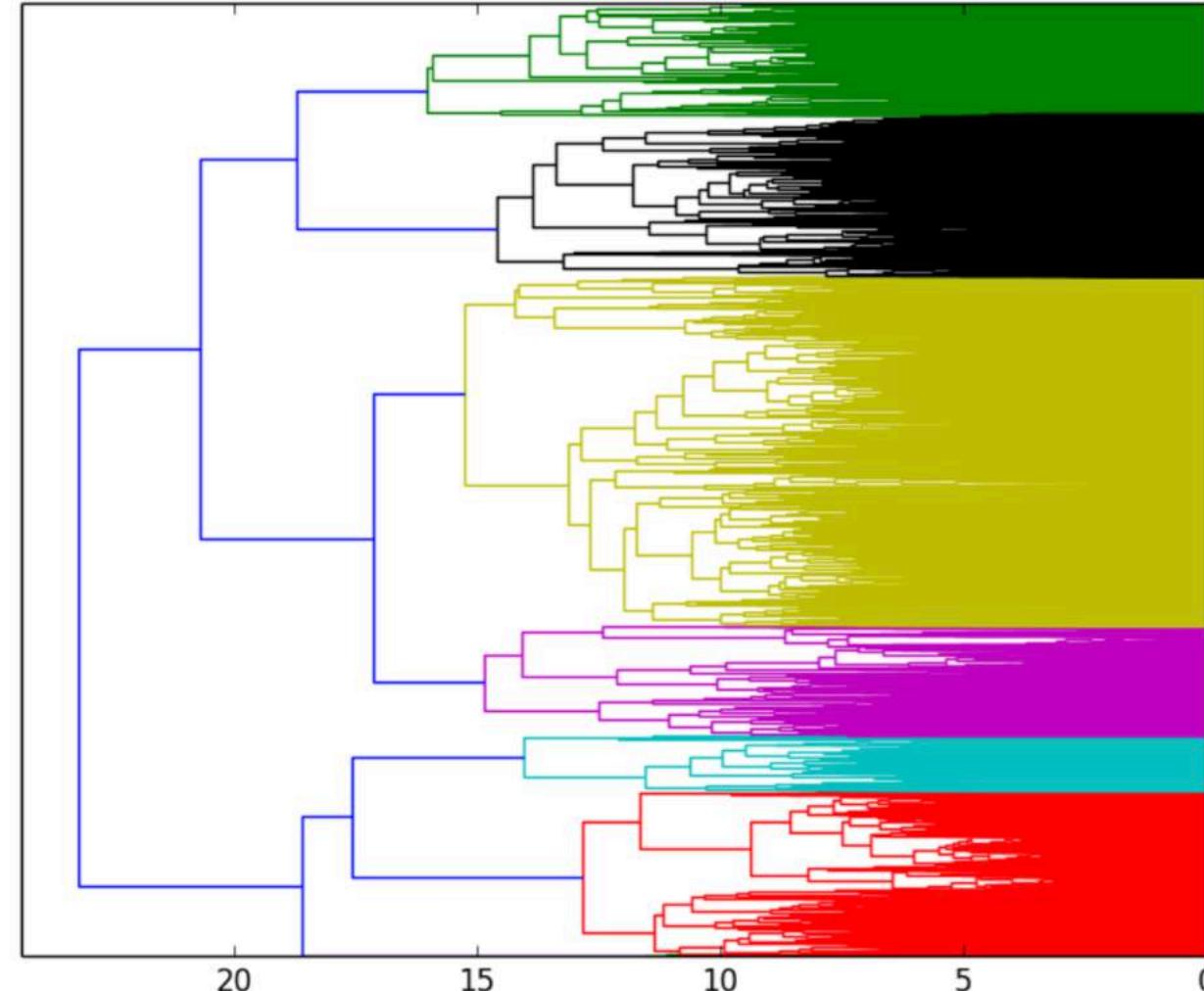
Дендрограмма



Дендрограмма

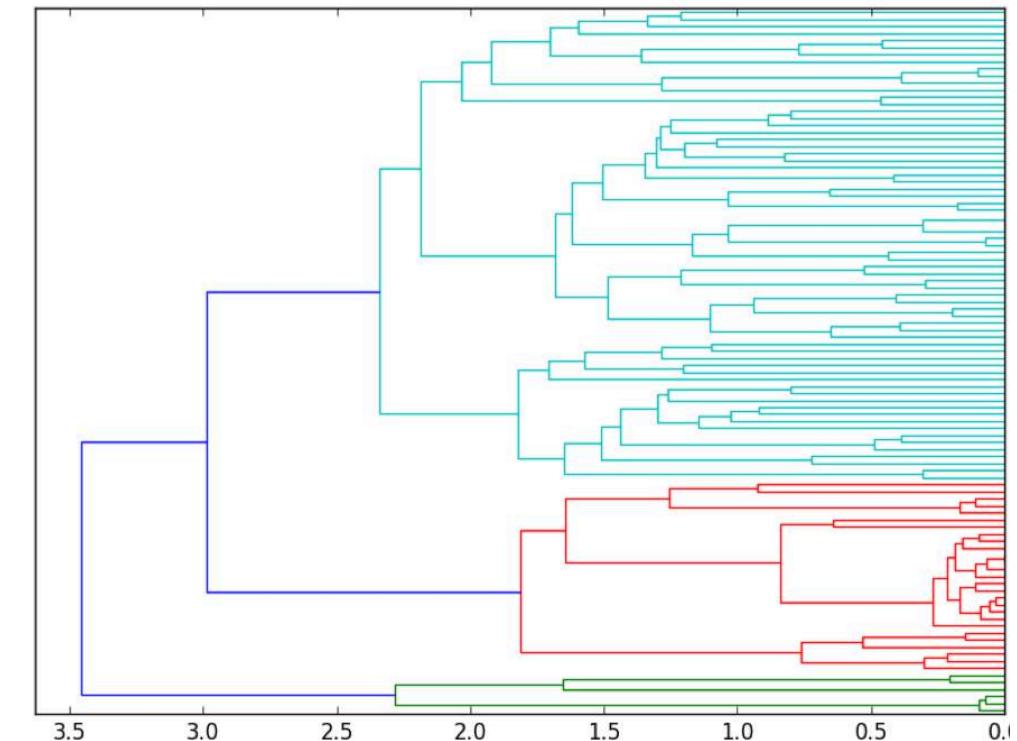
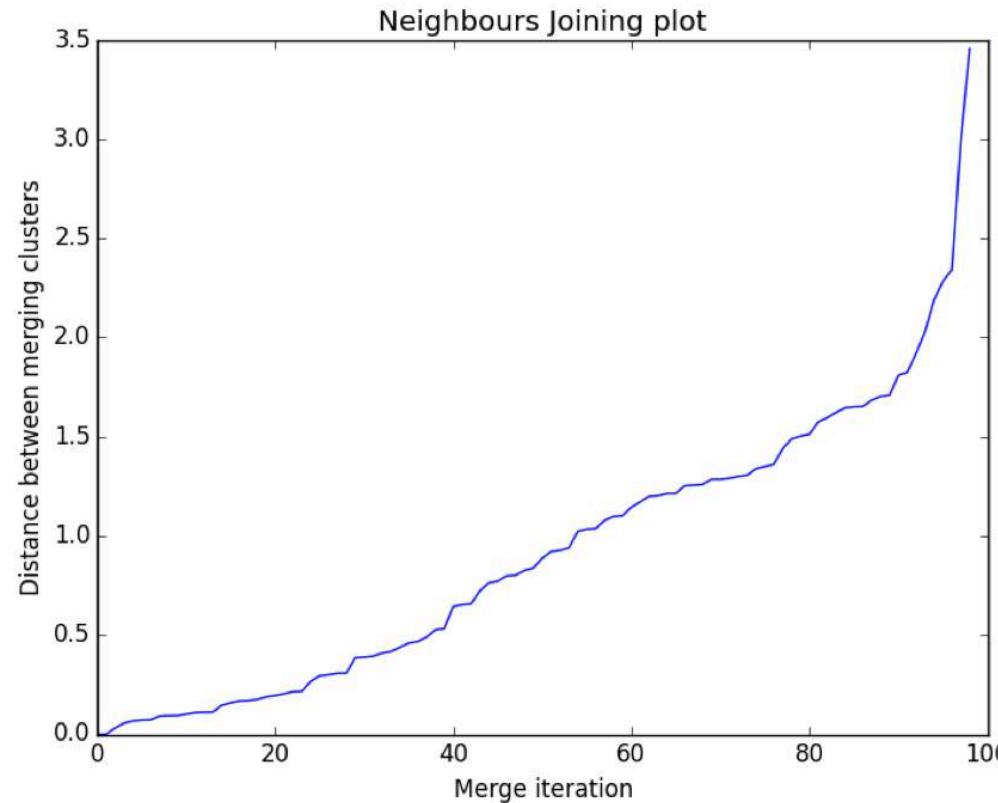


Пример: кластеризация писем



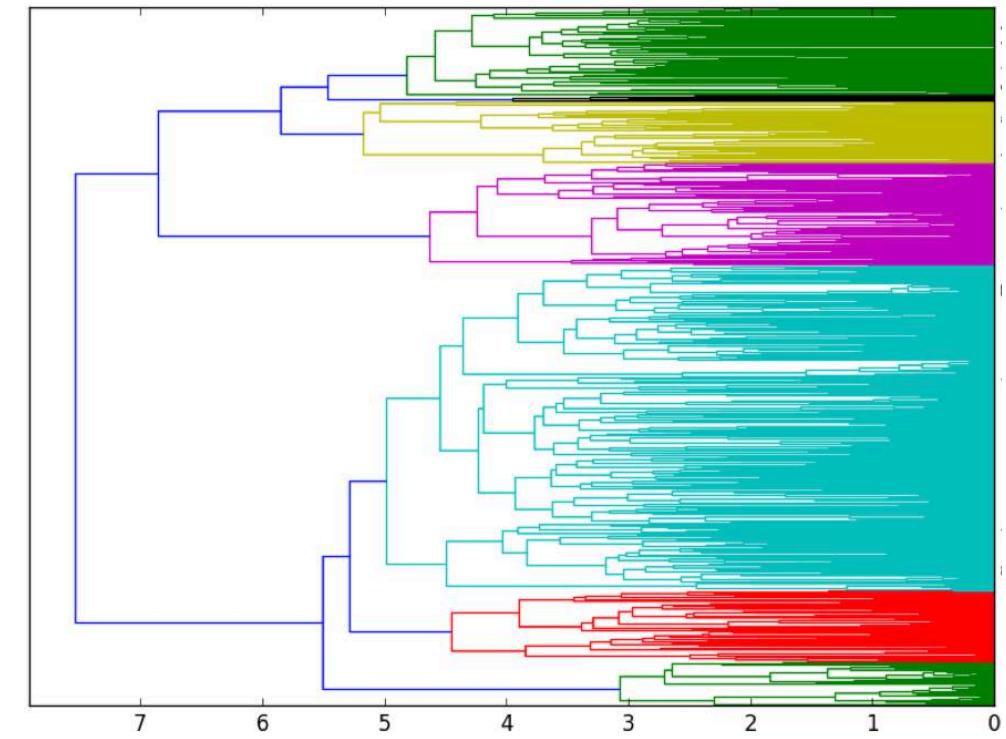
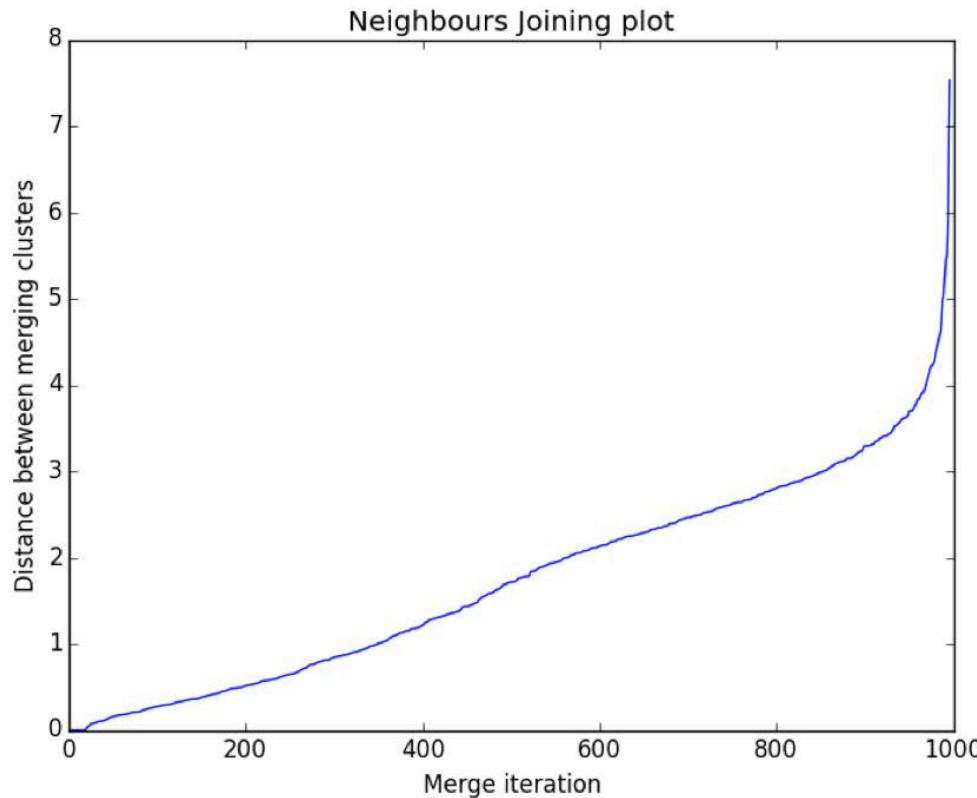
Пример: расстояние между кластерами

- На подвыборке из 100 писем



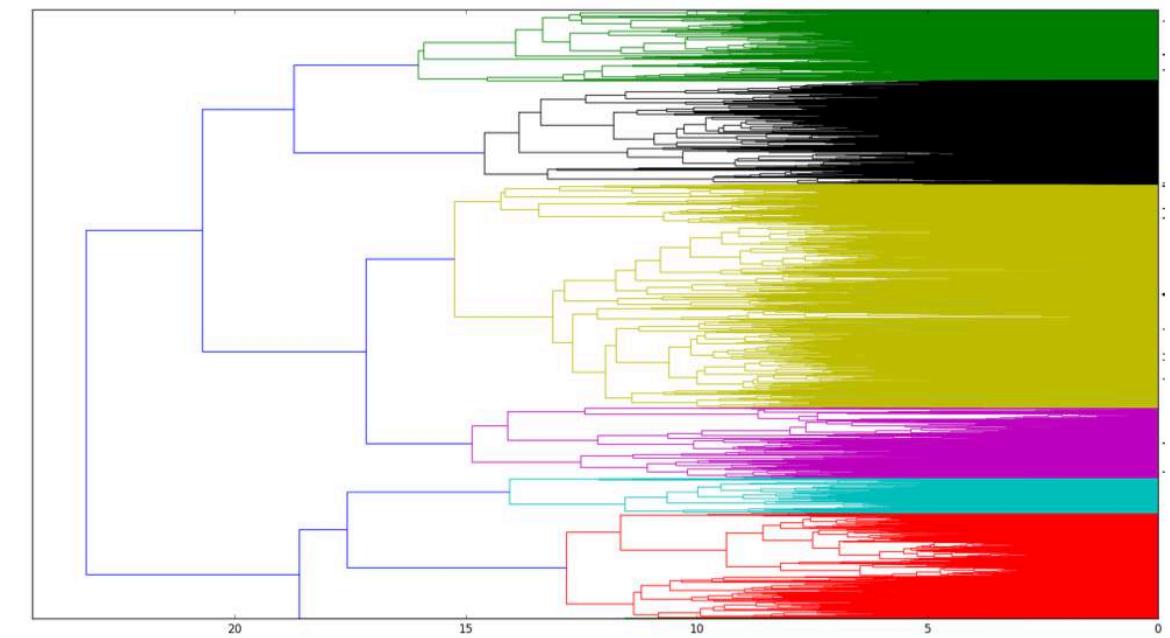
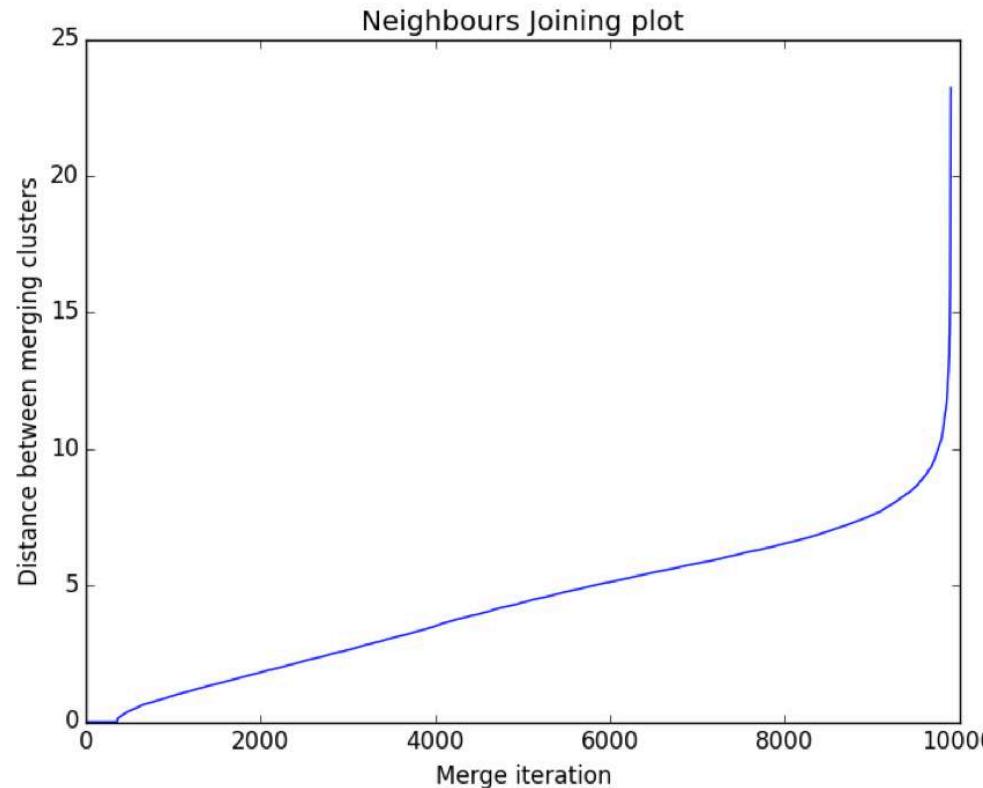
Пример: расстояние между кластерами

- На подвыборке из 1000 писем



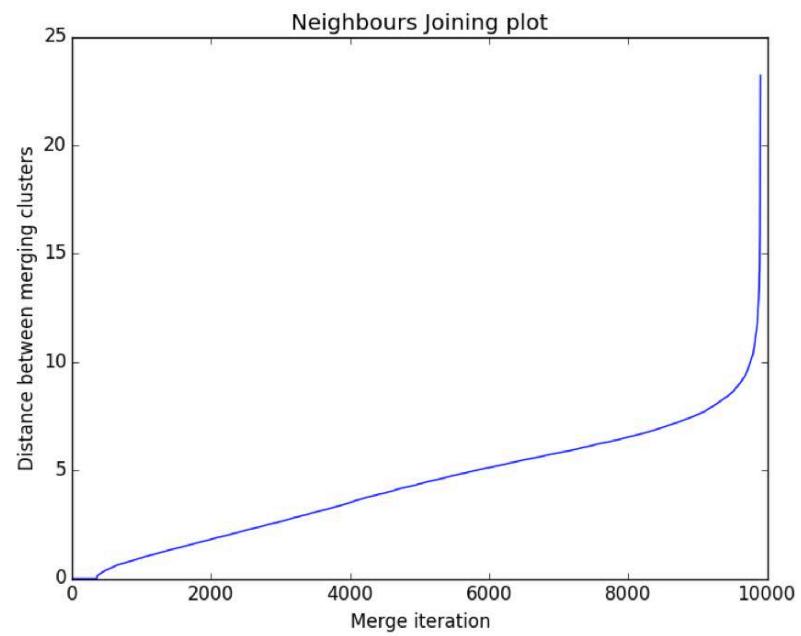
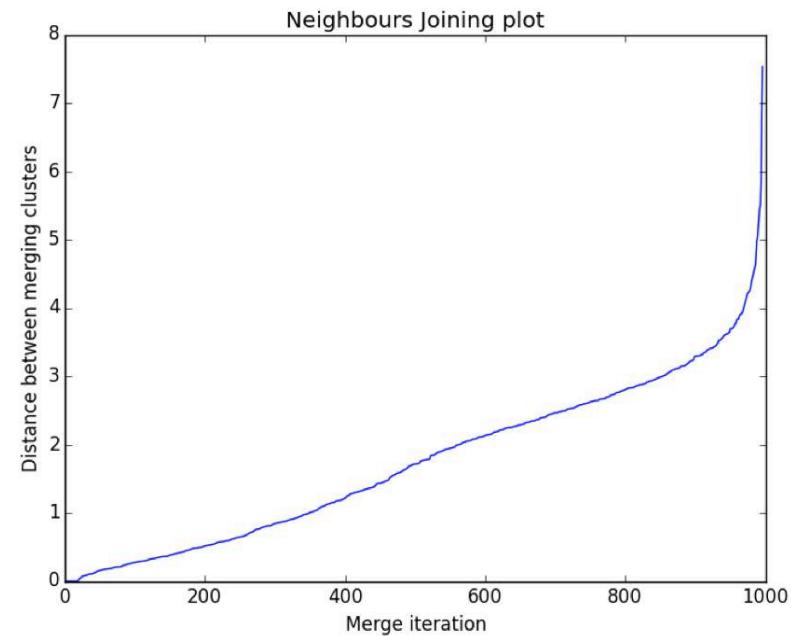
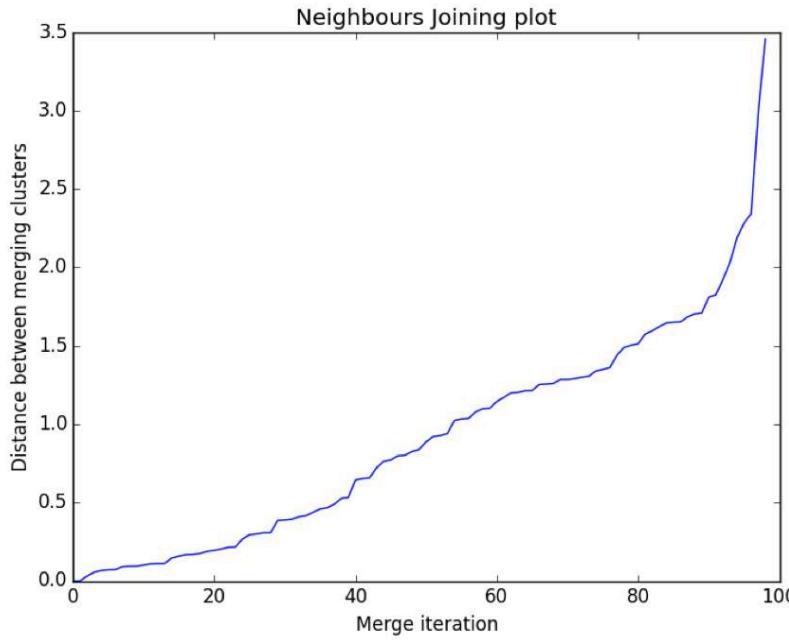
Пример: расстояние между кластерами

- На подвыборке из 10000 писем



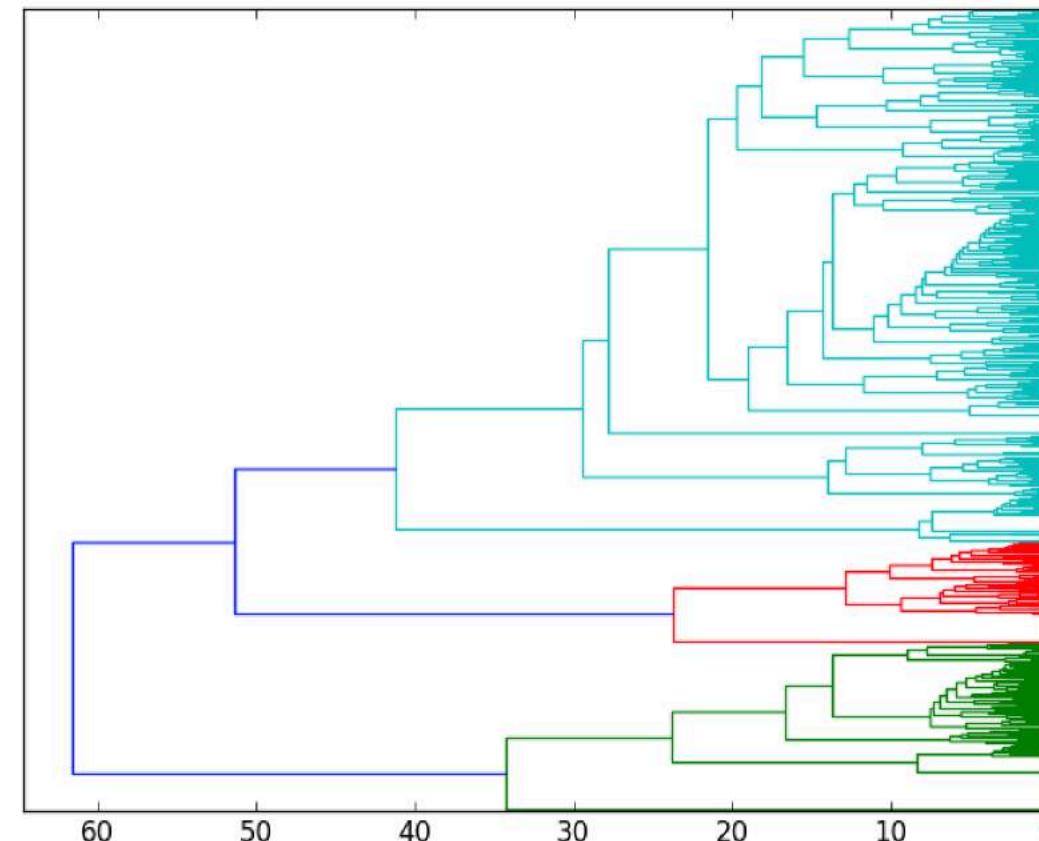
Пример: расстояние между кластерами

- Сравним графики: 100, 1000, 10000 писем

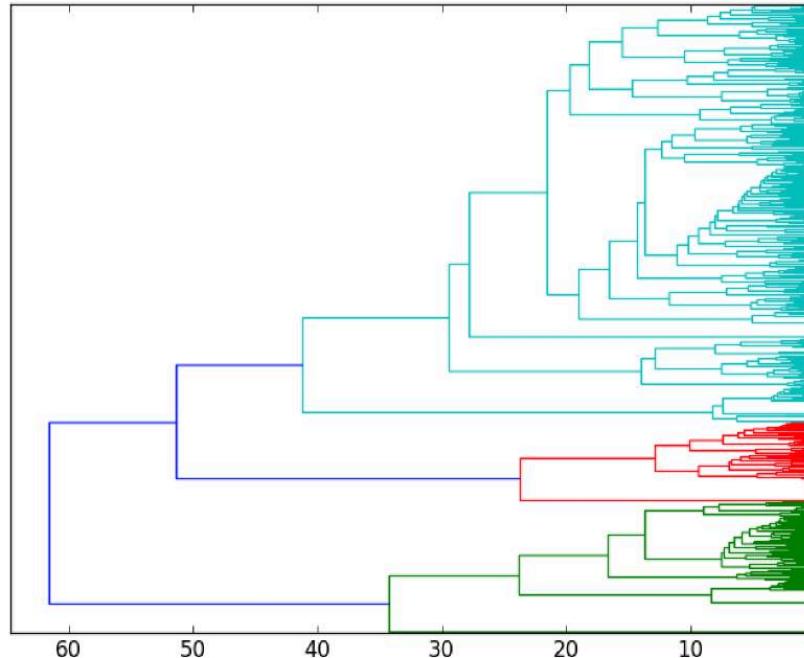


Пример: перекос в размерах кластеров

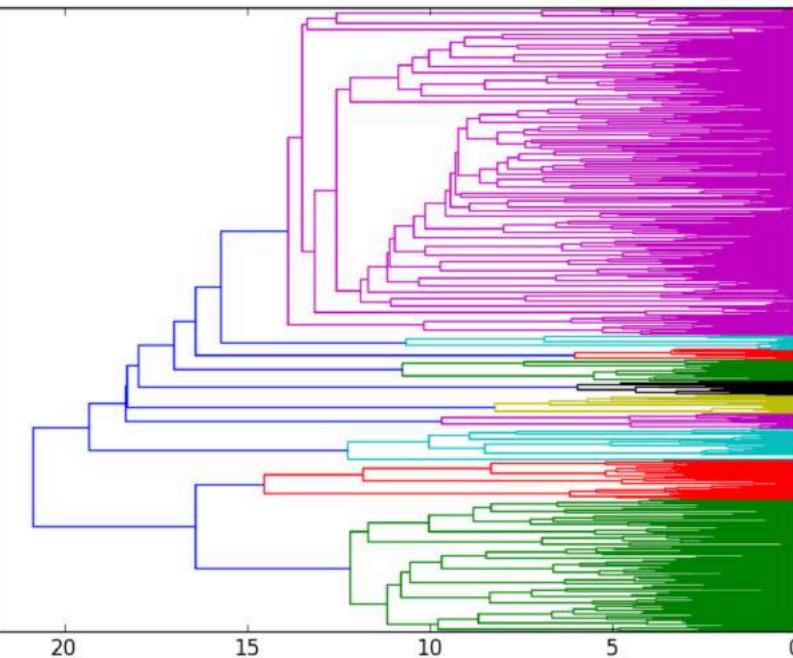
- Дендрограмма, построенная для другой выборки текстов:



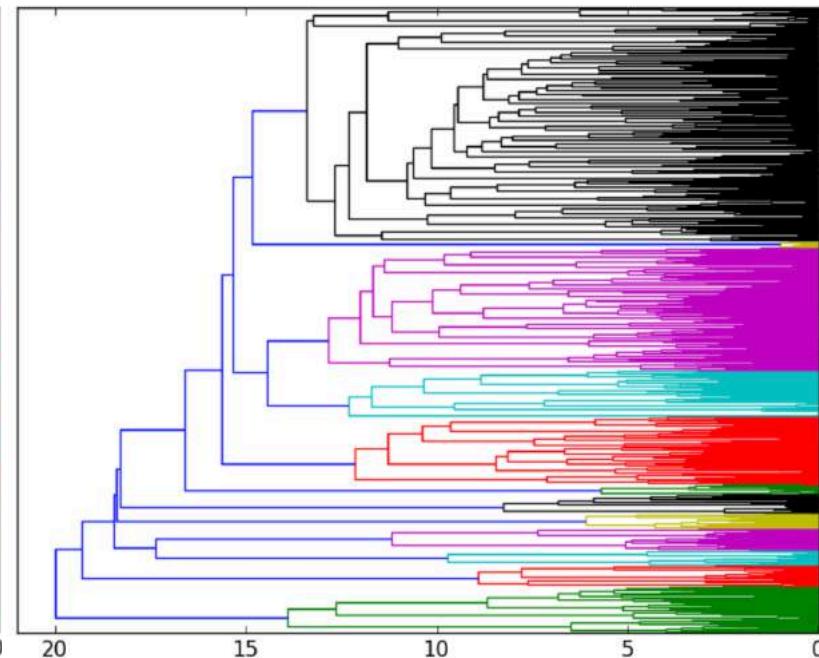
Пример: добавляем SVD



Исходные признаки

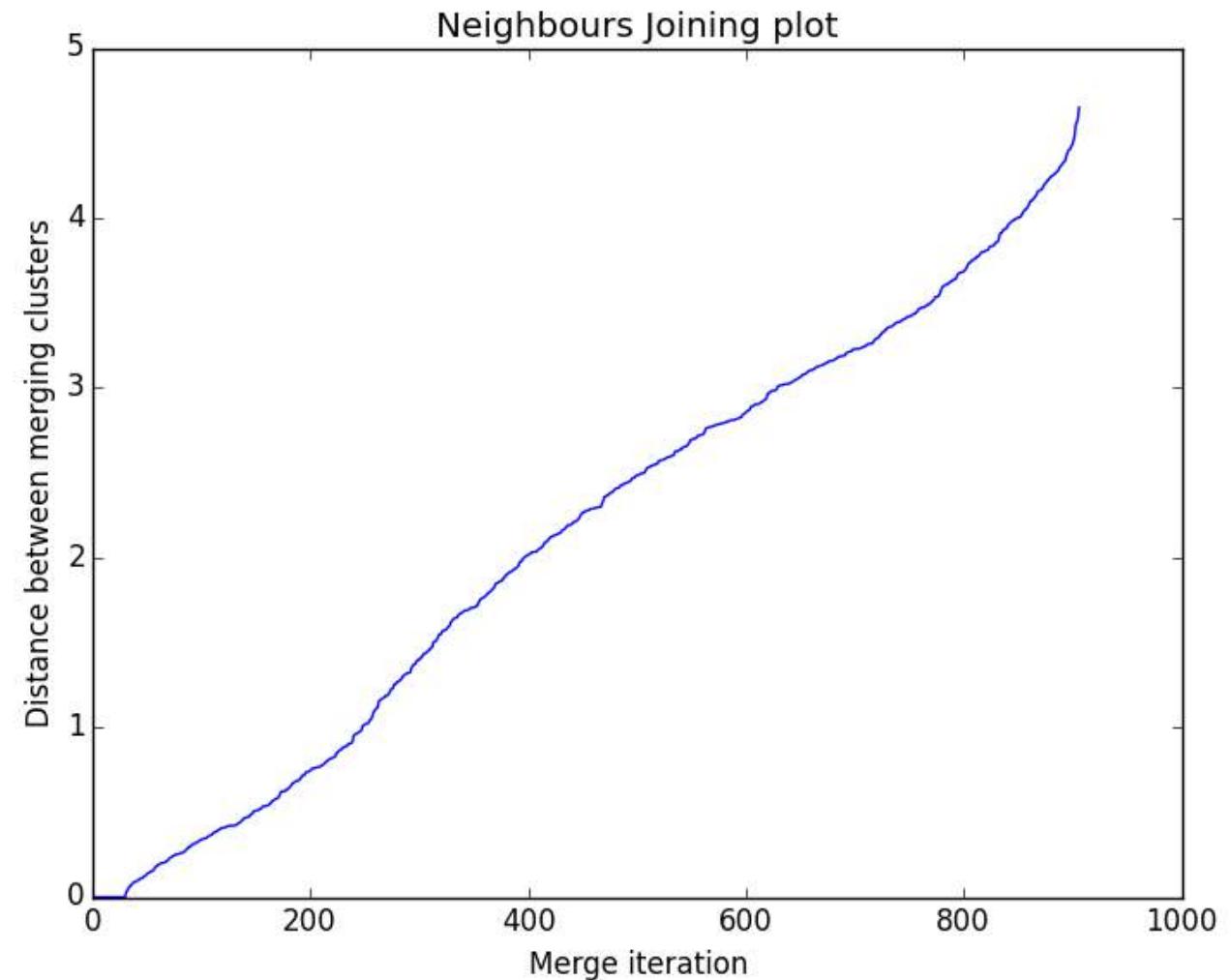


SVD

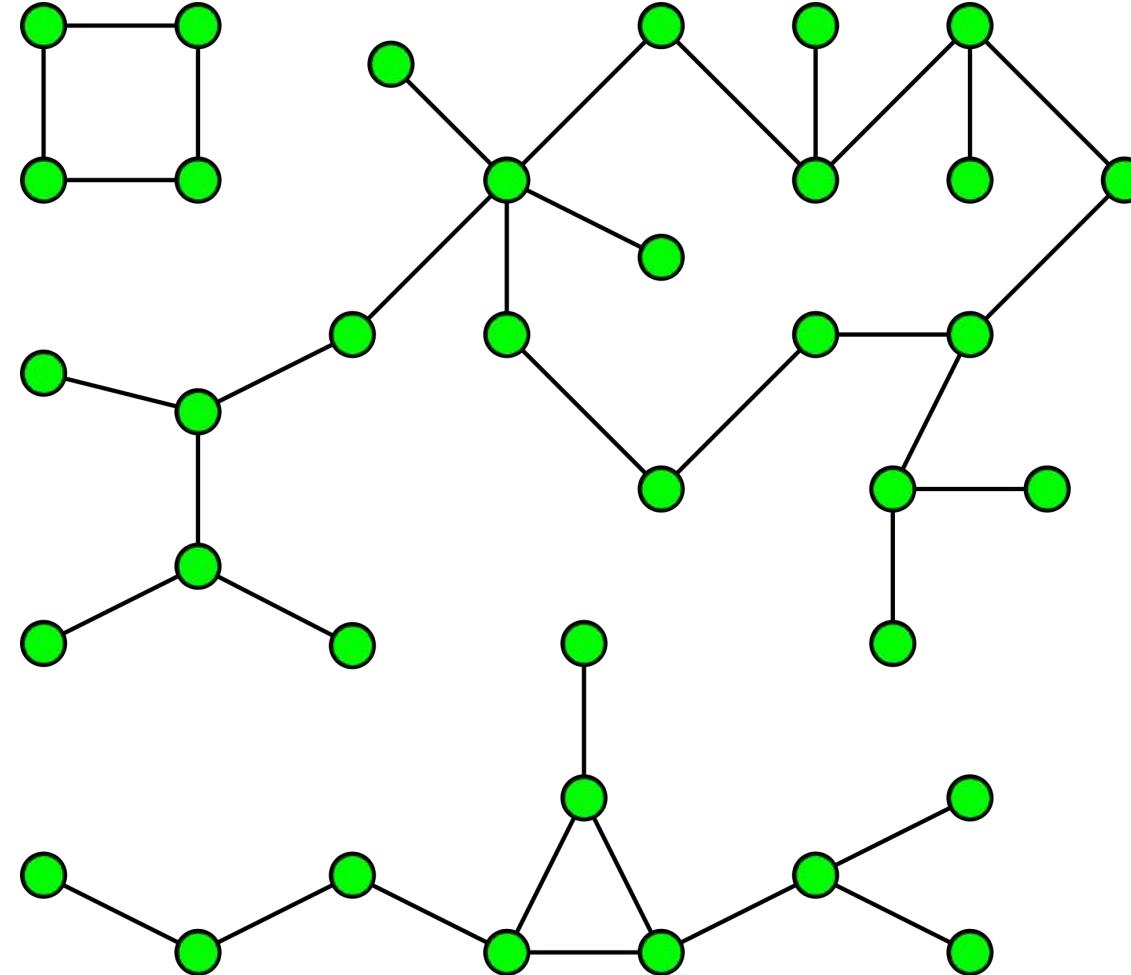


SVD (еще меньше компонент)

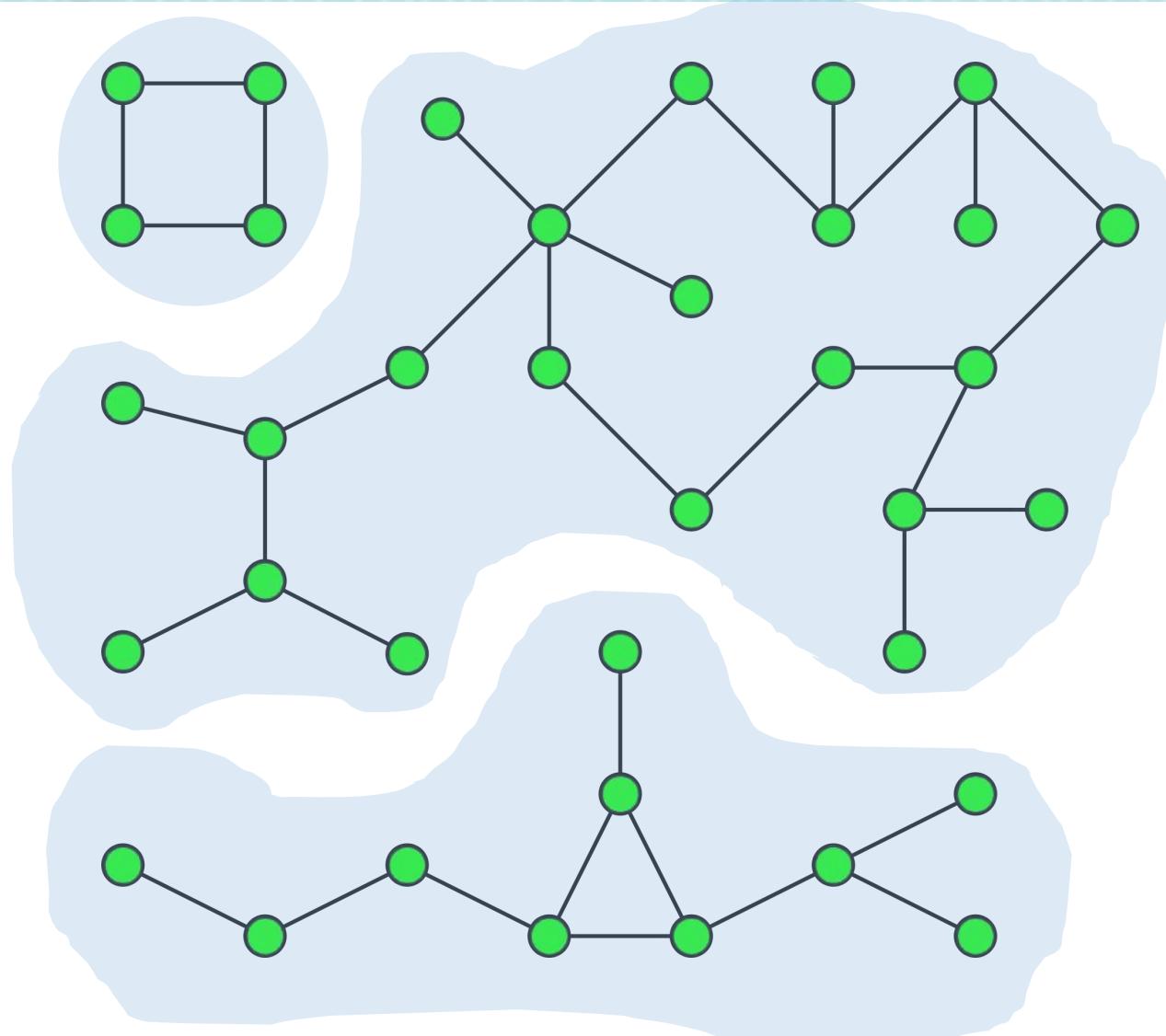
Пример: SVD и расстояние при слиянии



Простые графовые подходы



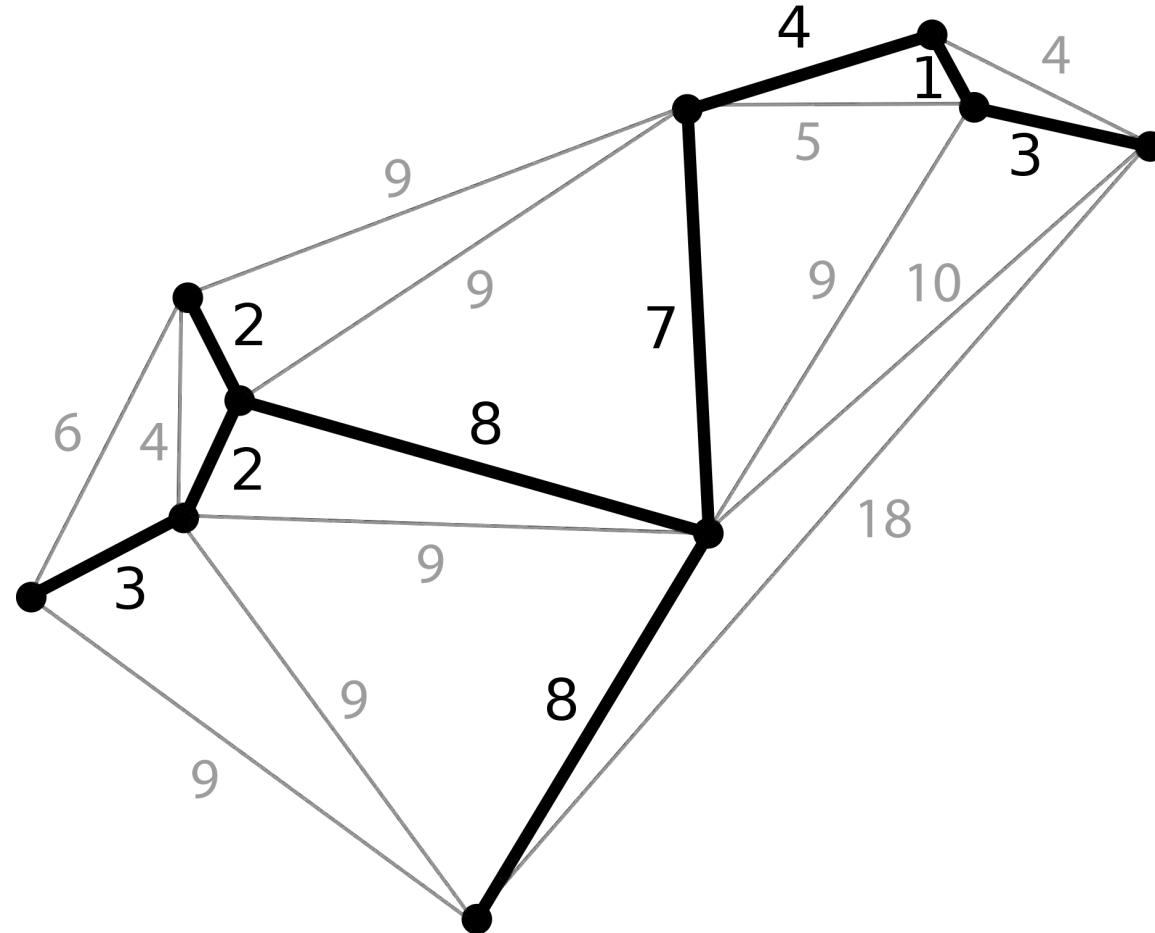
Выделение связных компонент



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 оствовного дерева

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

Идея density-based методов

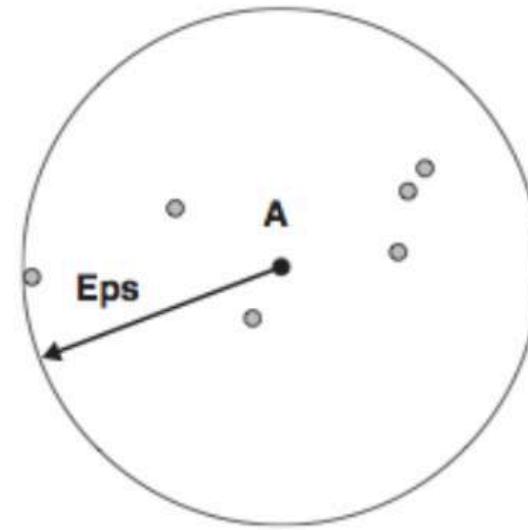


Figure 8.20. Center-based density.

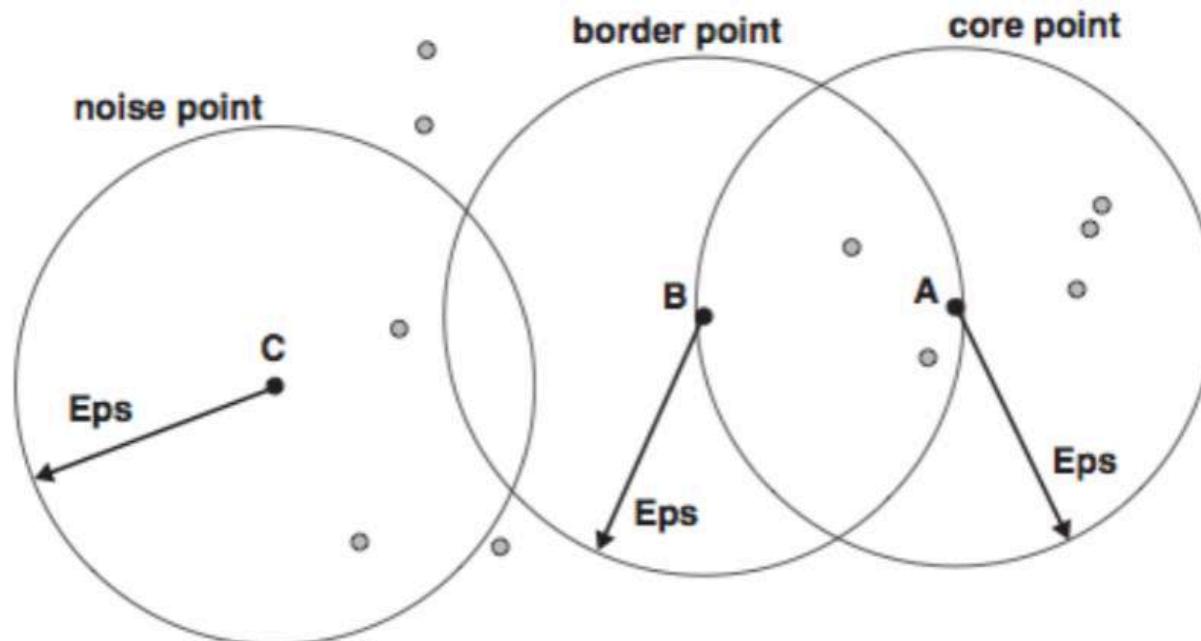
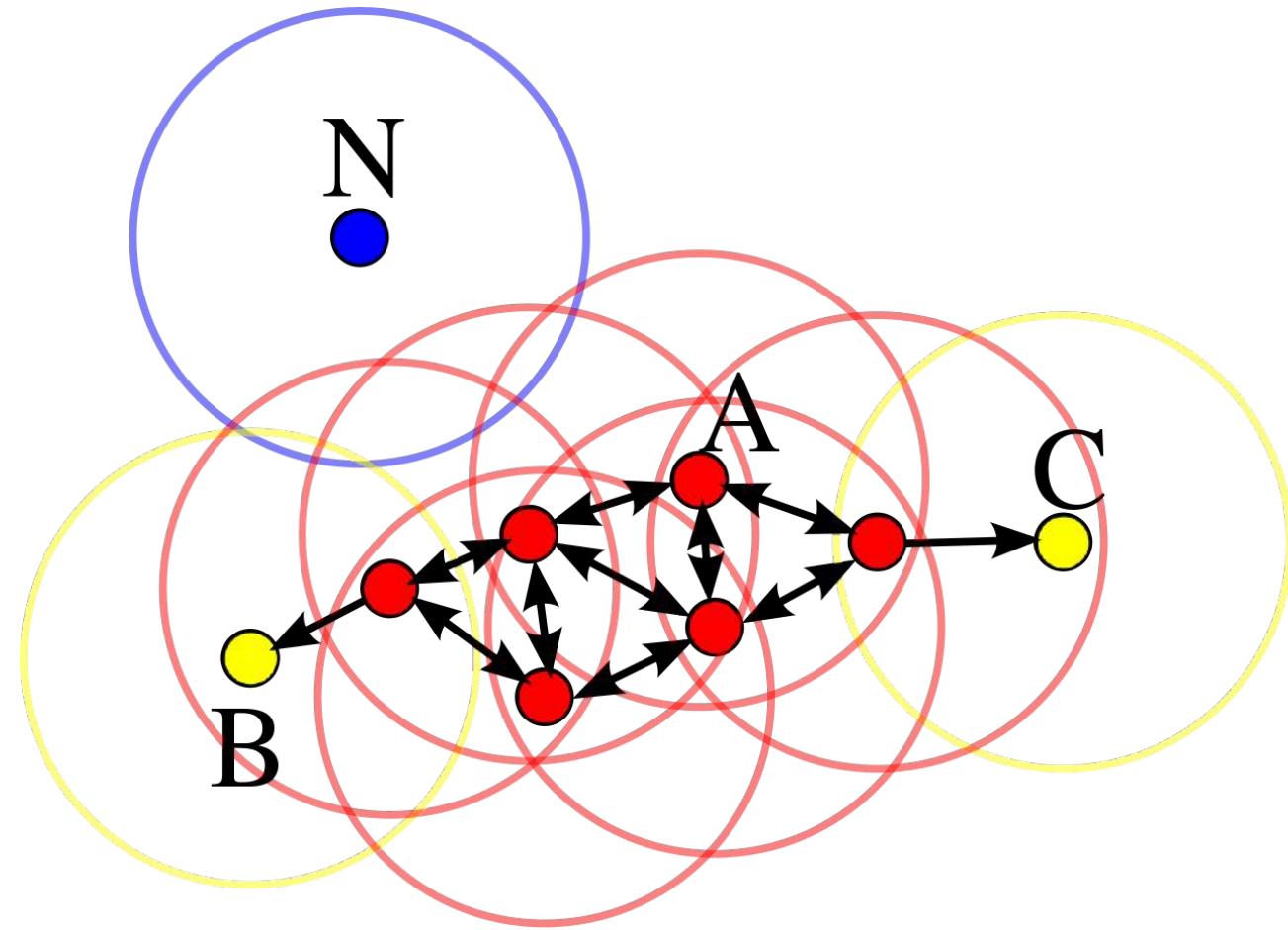


Figure 8.21. Core, border, and noise points.

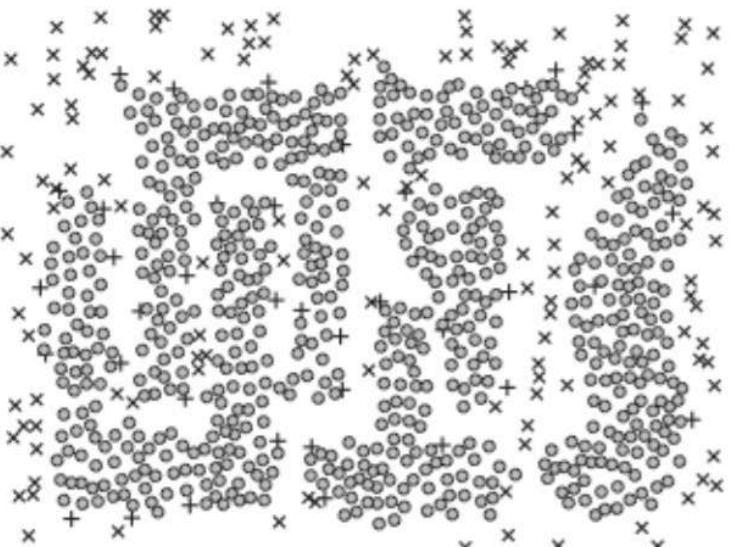
Основные, шумовые и граничные точки



DBSCAN



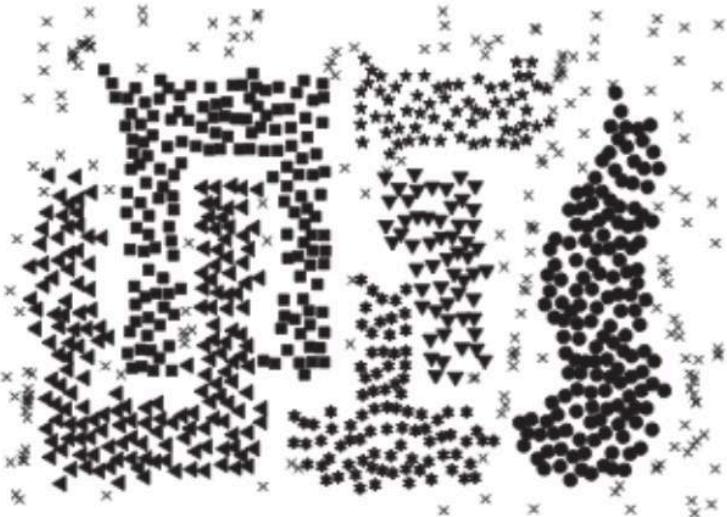
(a) Clusters found by DBSCAN.



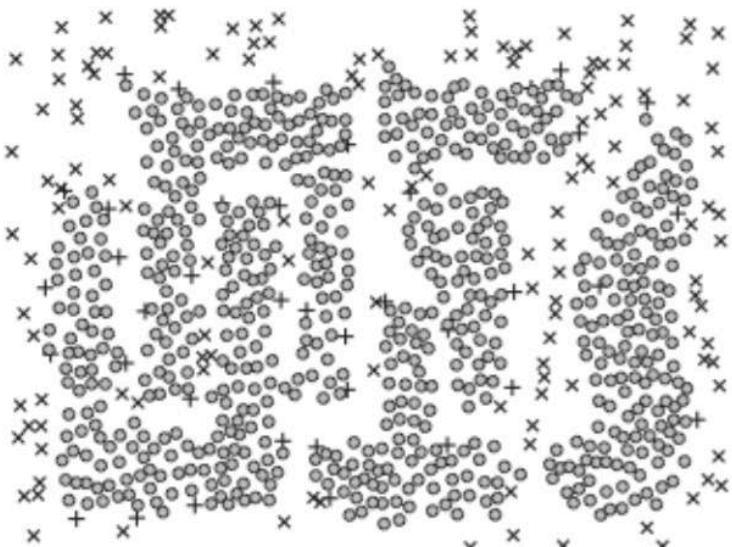
(b) Core, border, and noise points.

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

DBSCAN



(a) Clusters found by DBSCAN.



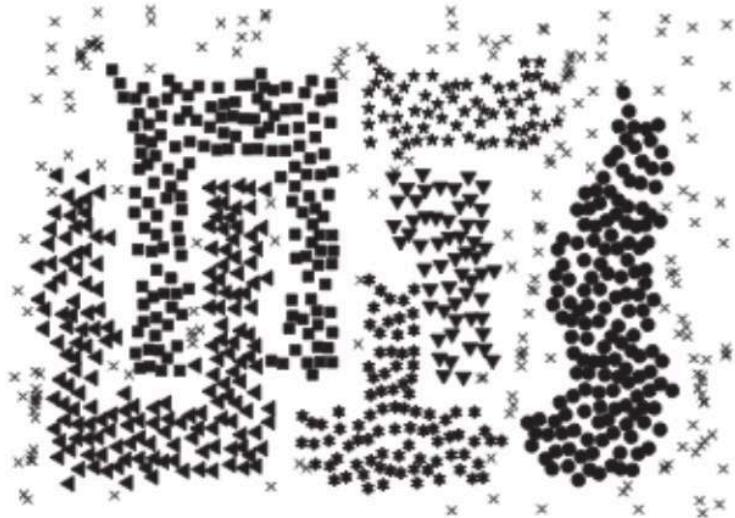
x – Noise Point + – Border Point o – Core Point

(b) Core, border, and noise points.

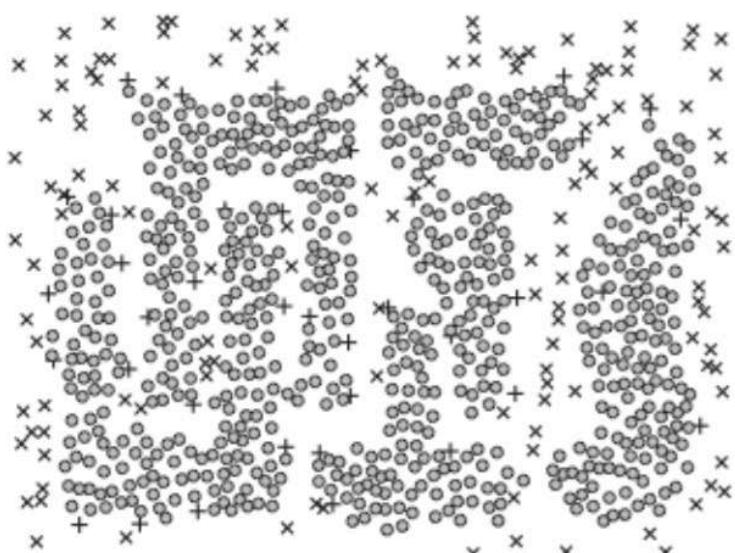
1: Пометить все точки, как основные, пограничные или шумовые.

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.

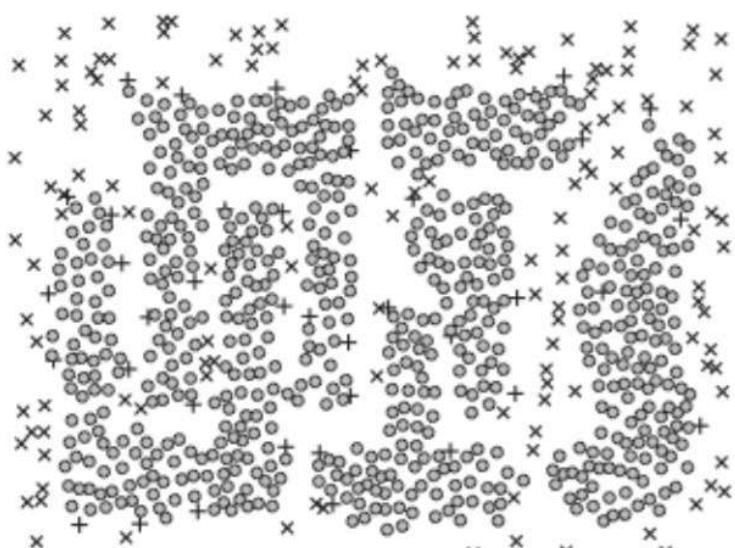
2: Отбросить точки шума.

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.

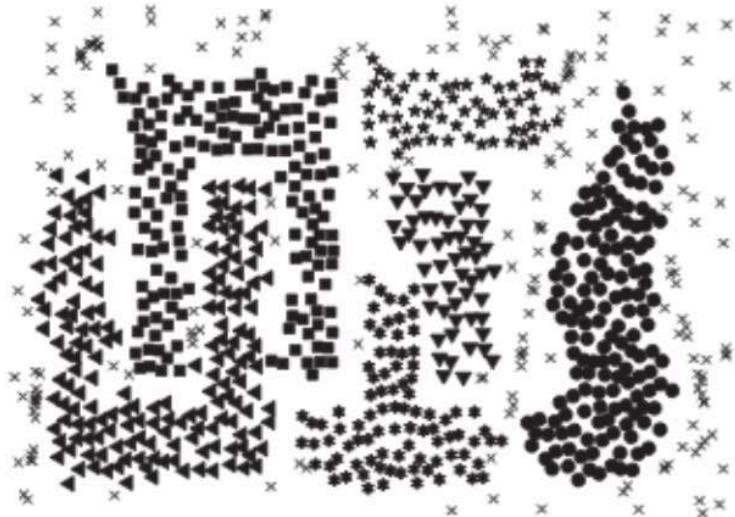
1: Пометить все точки, как основные, пограничные или шумовые.

2: Отбросить точки шума.

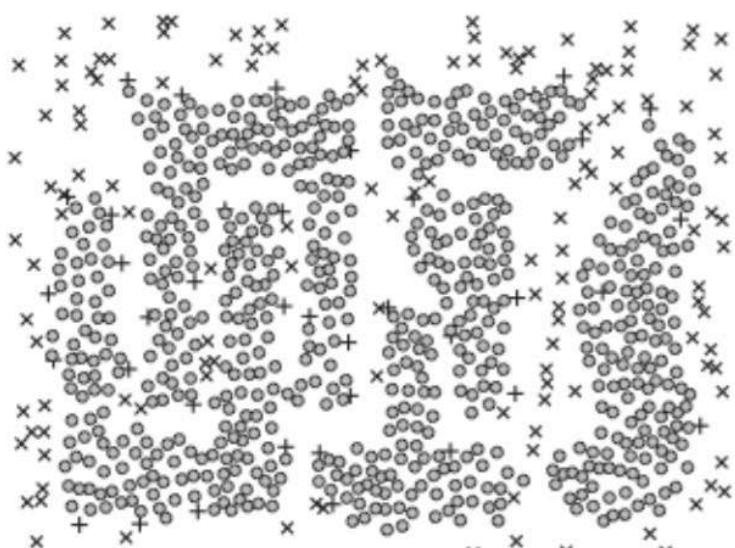
3: Соединить все основные точки, находящиеся на расстоянии Eps радиуса одна от другой.

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.

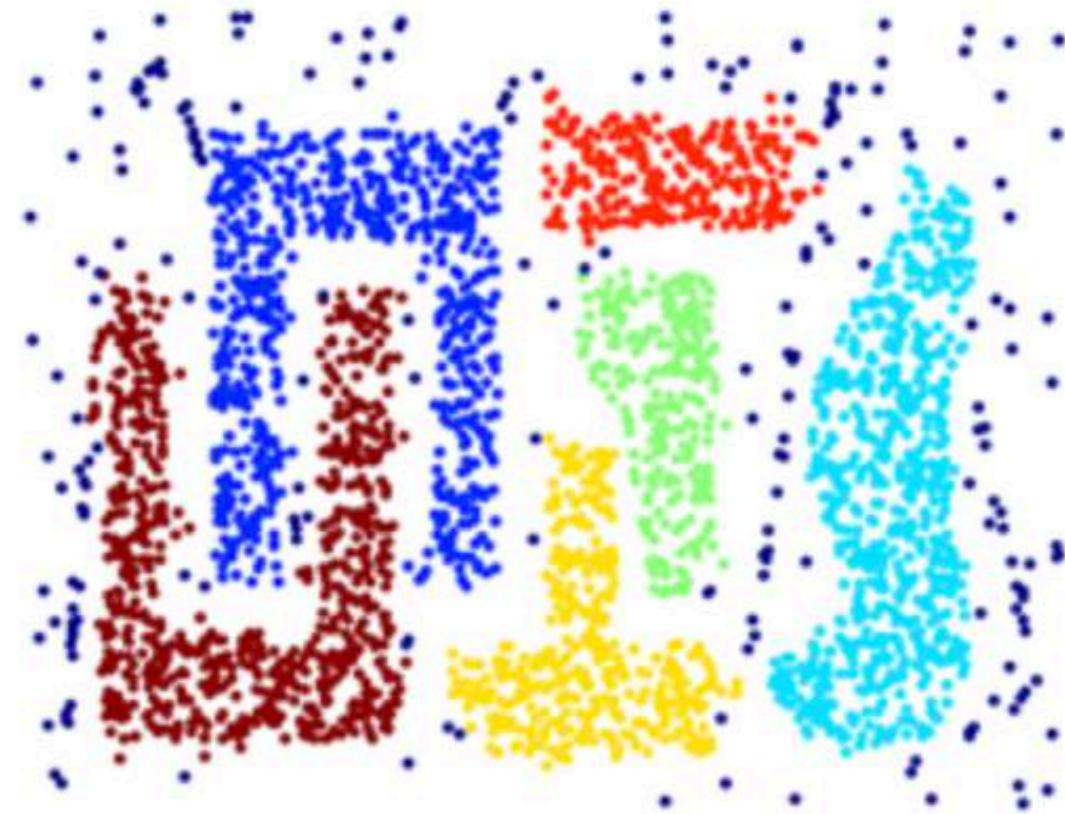
2: Отбросить точки шума.

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

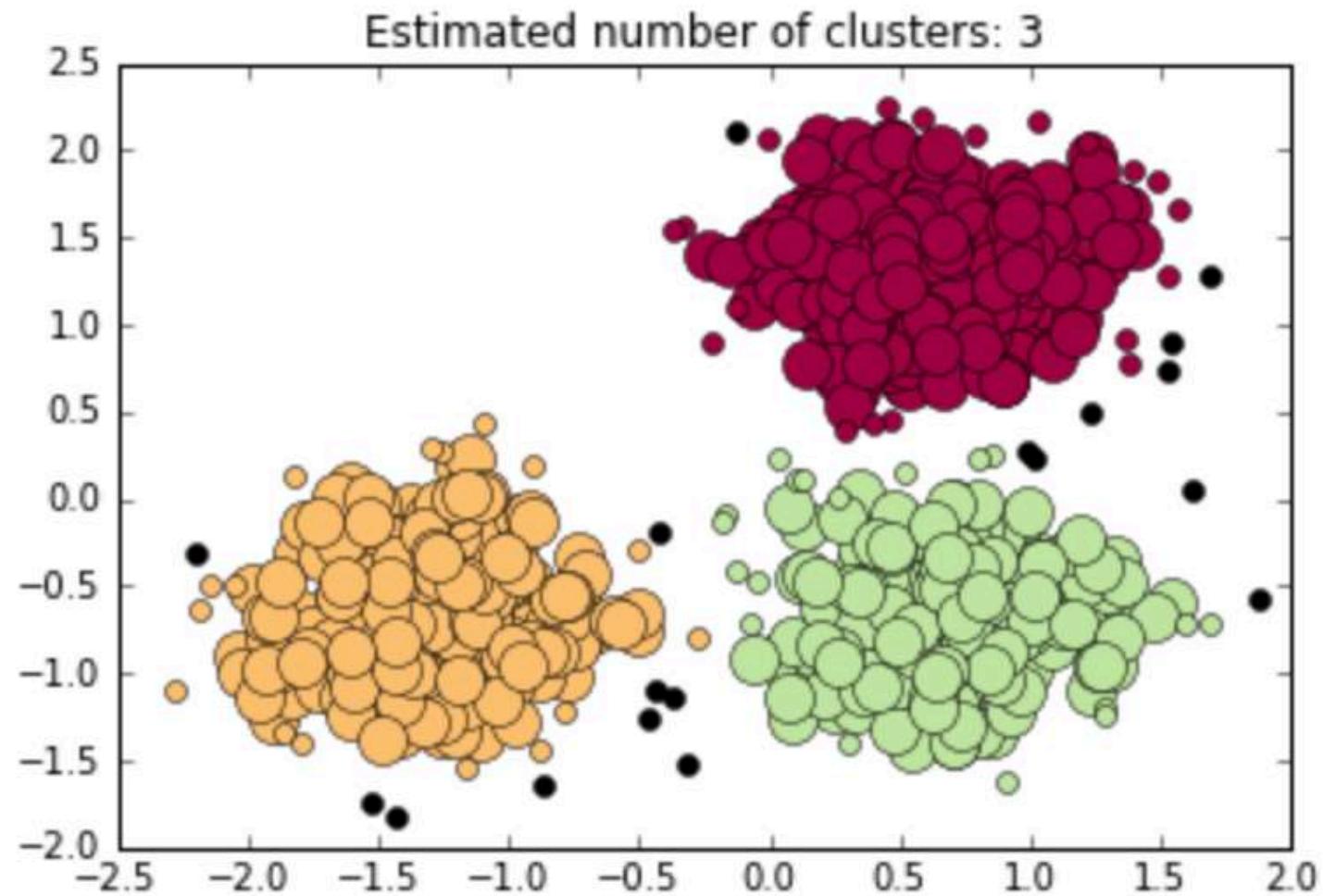
4: Объединить каждую группу соединенных основных точек в отдельный кластер.

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

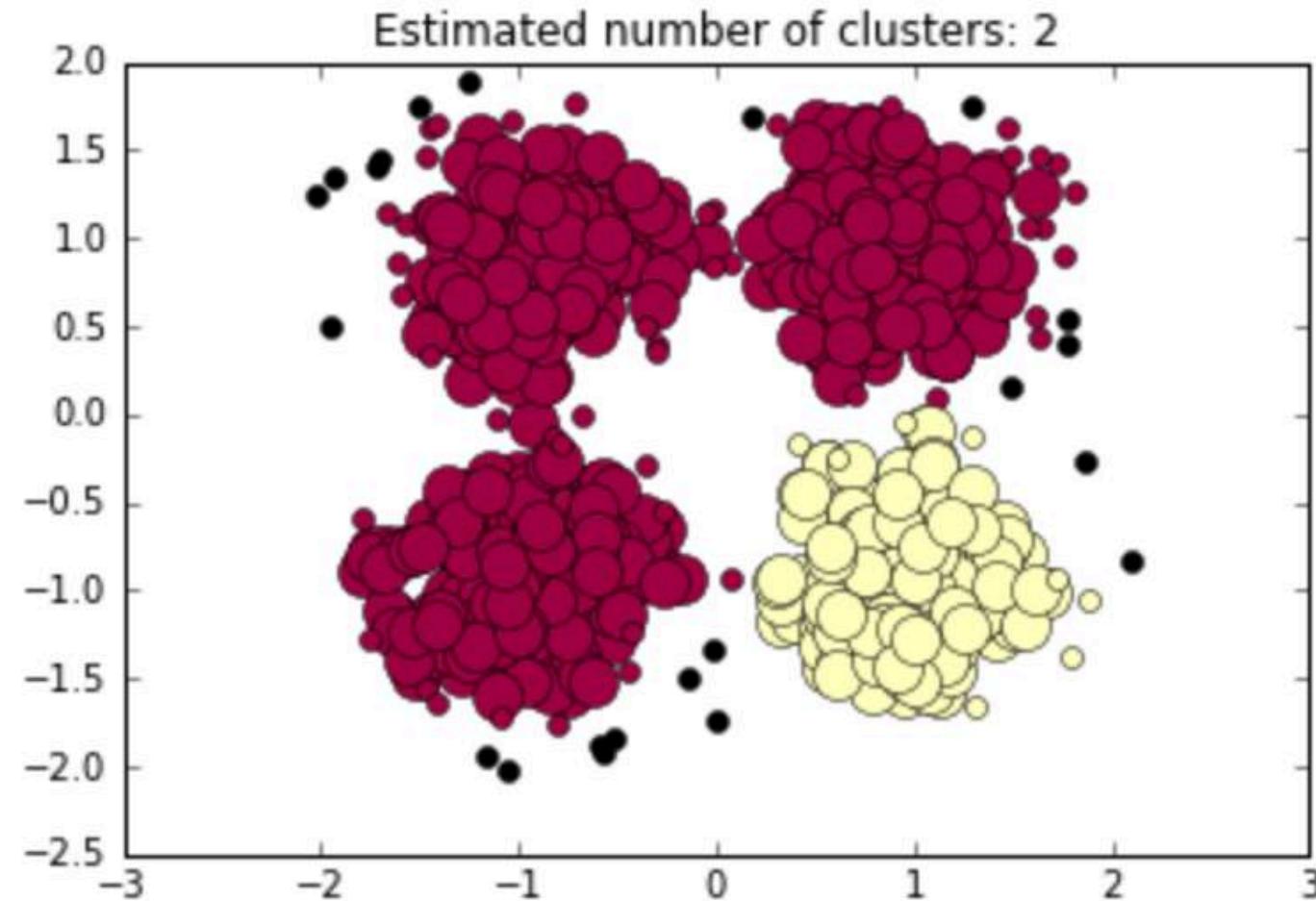
DBSCAN: результаты работы



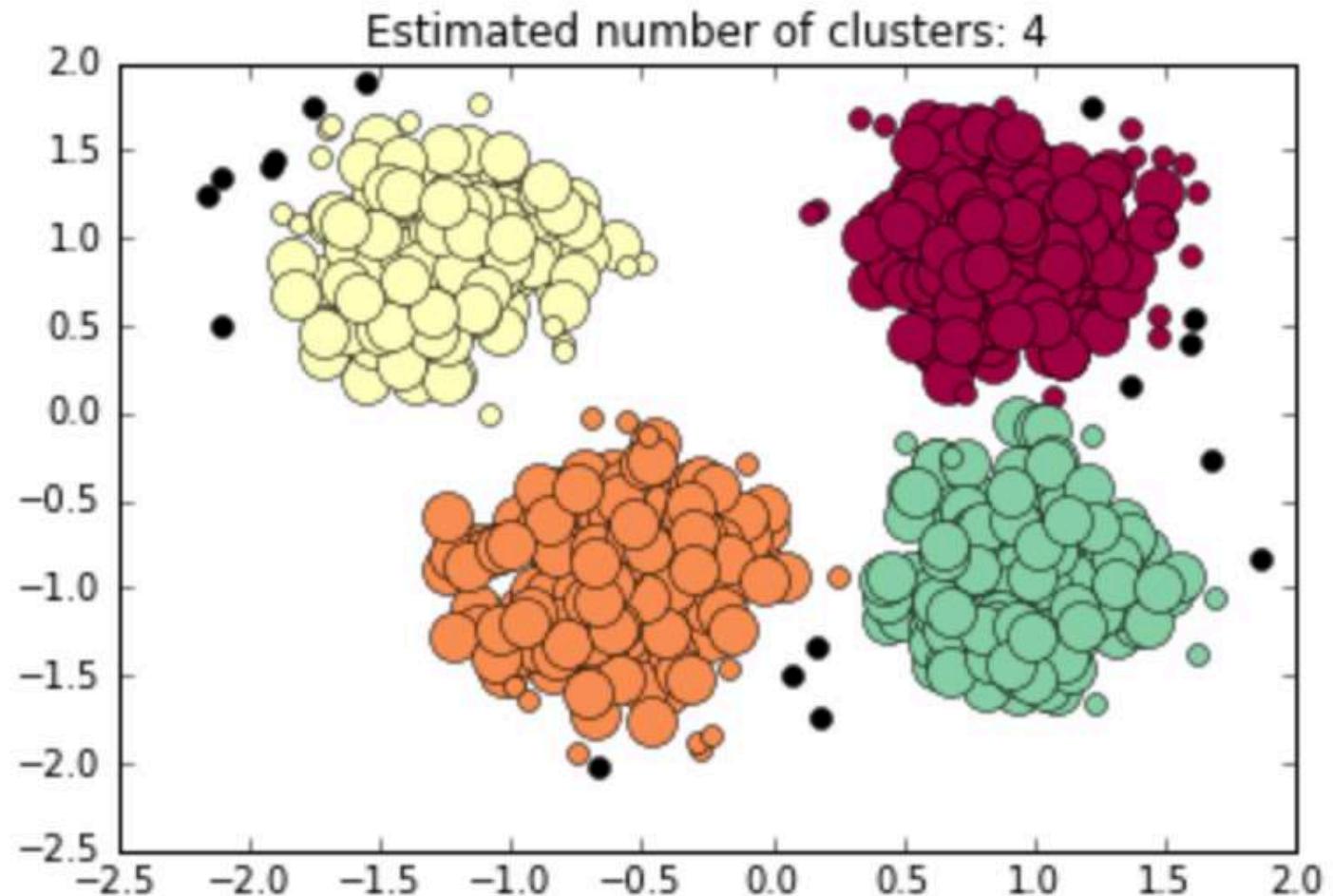
Определение числа кластеров



Определение числа кластеров



Определение числа кластеров



IV. Сравнение алгоритмов и оценка качества

Алгоритмы

Рассмотренные нами:

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

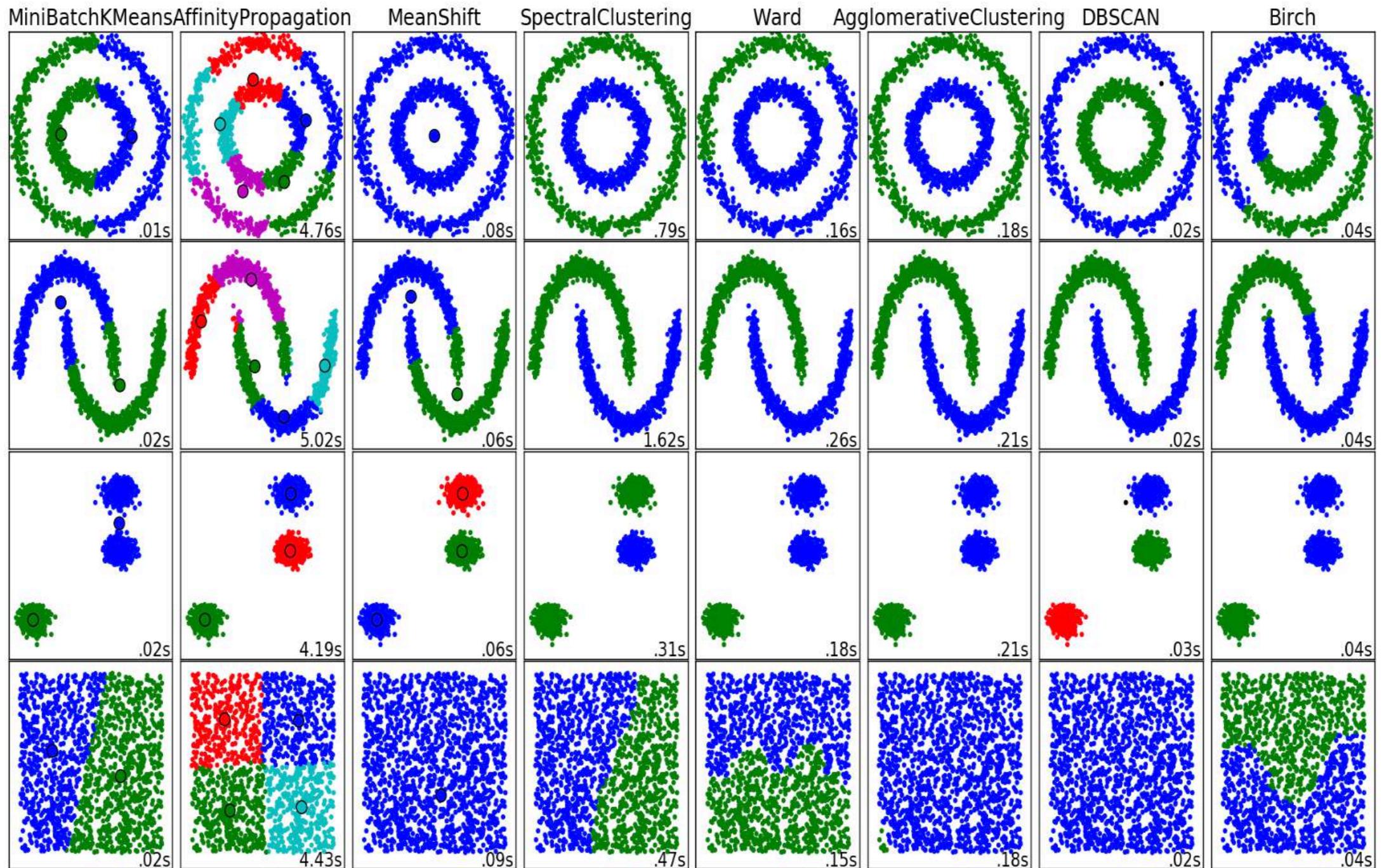
Алгоритмы

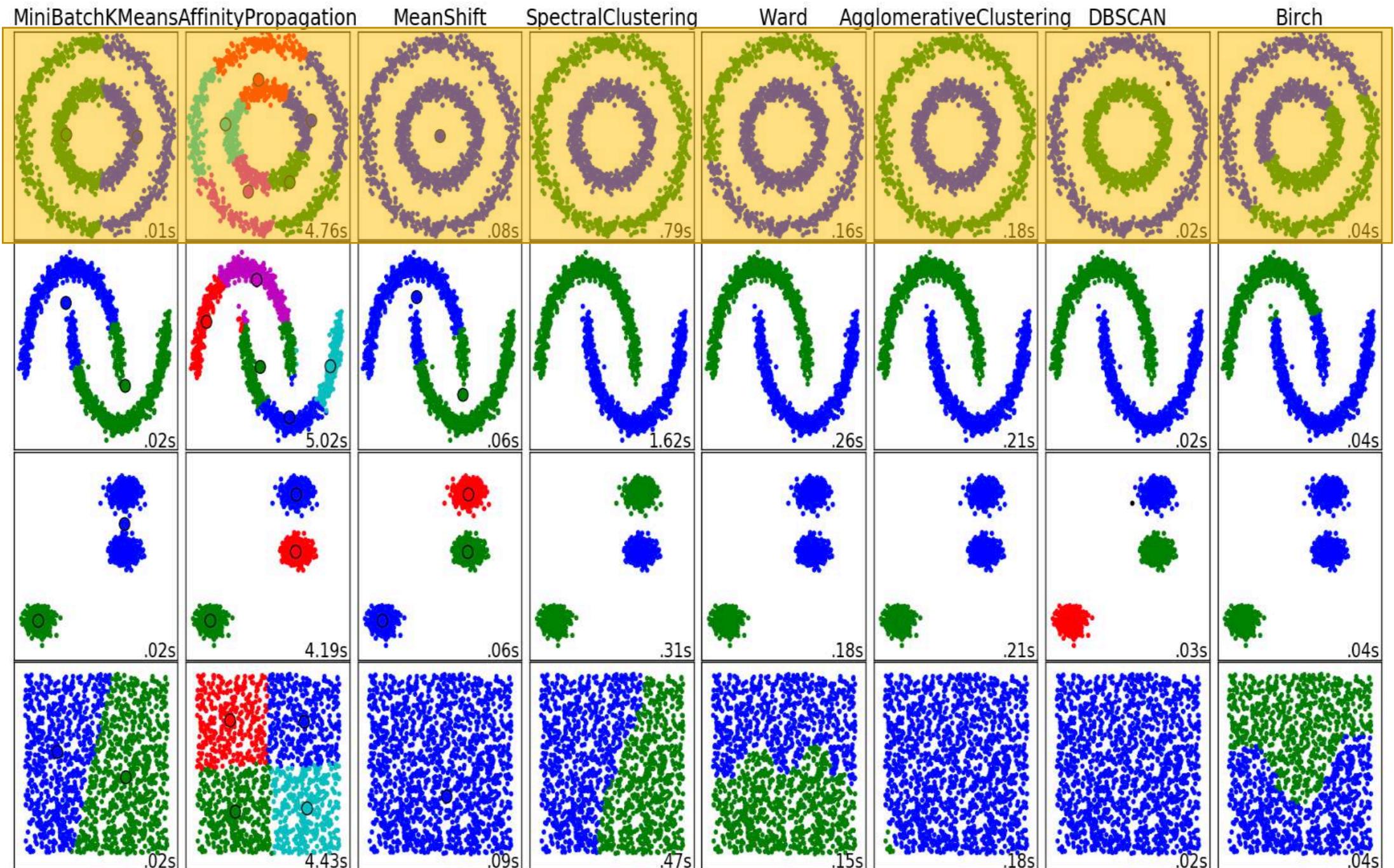
Рассмотренные нами:

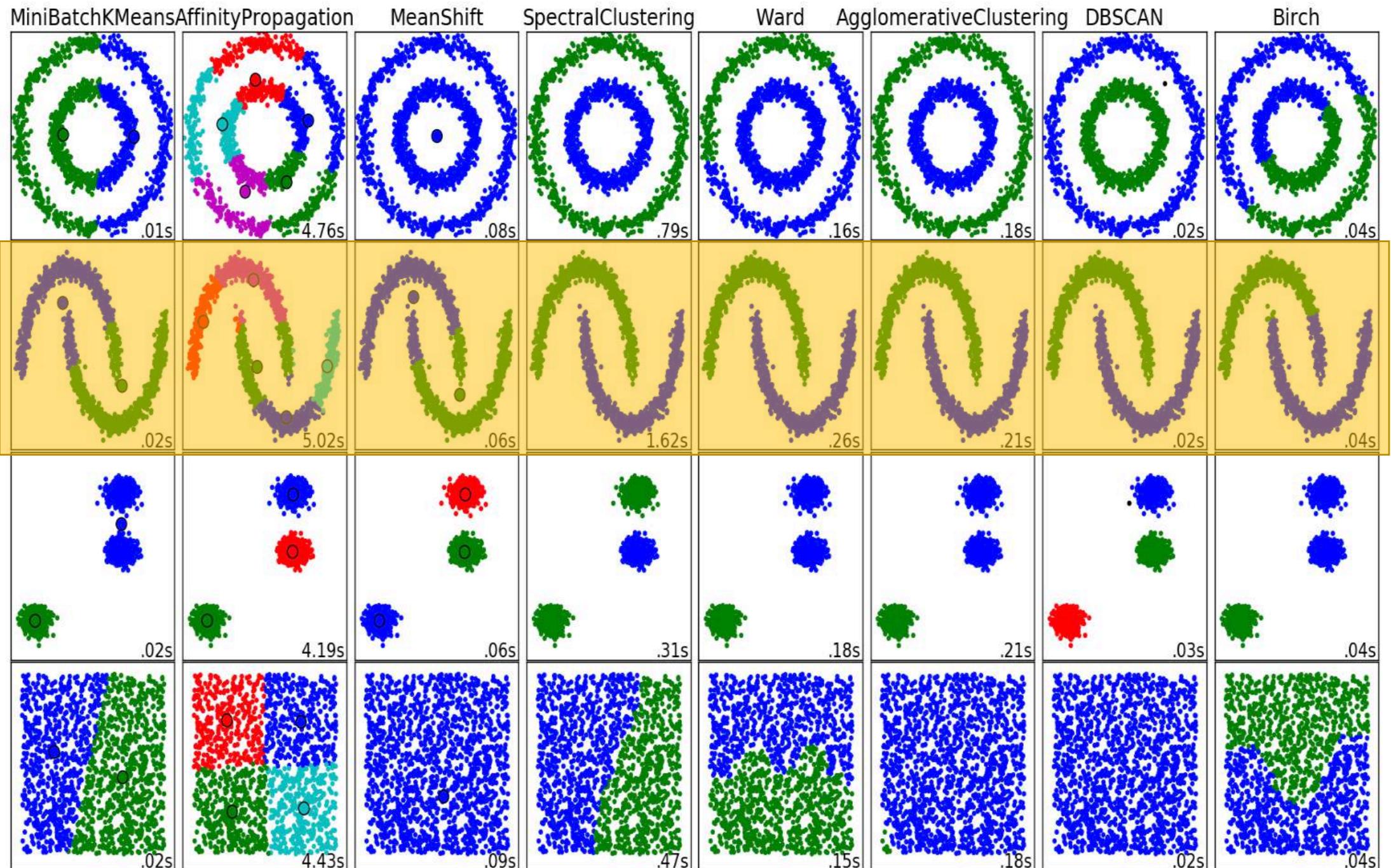
- К-средних
- ЕМ-алгоритм
- Агglomerативная иерархическая кластеризация
- DBSCAN

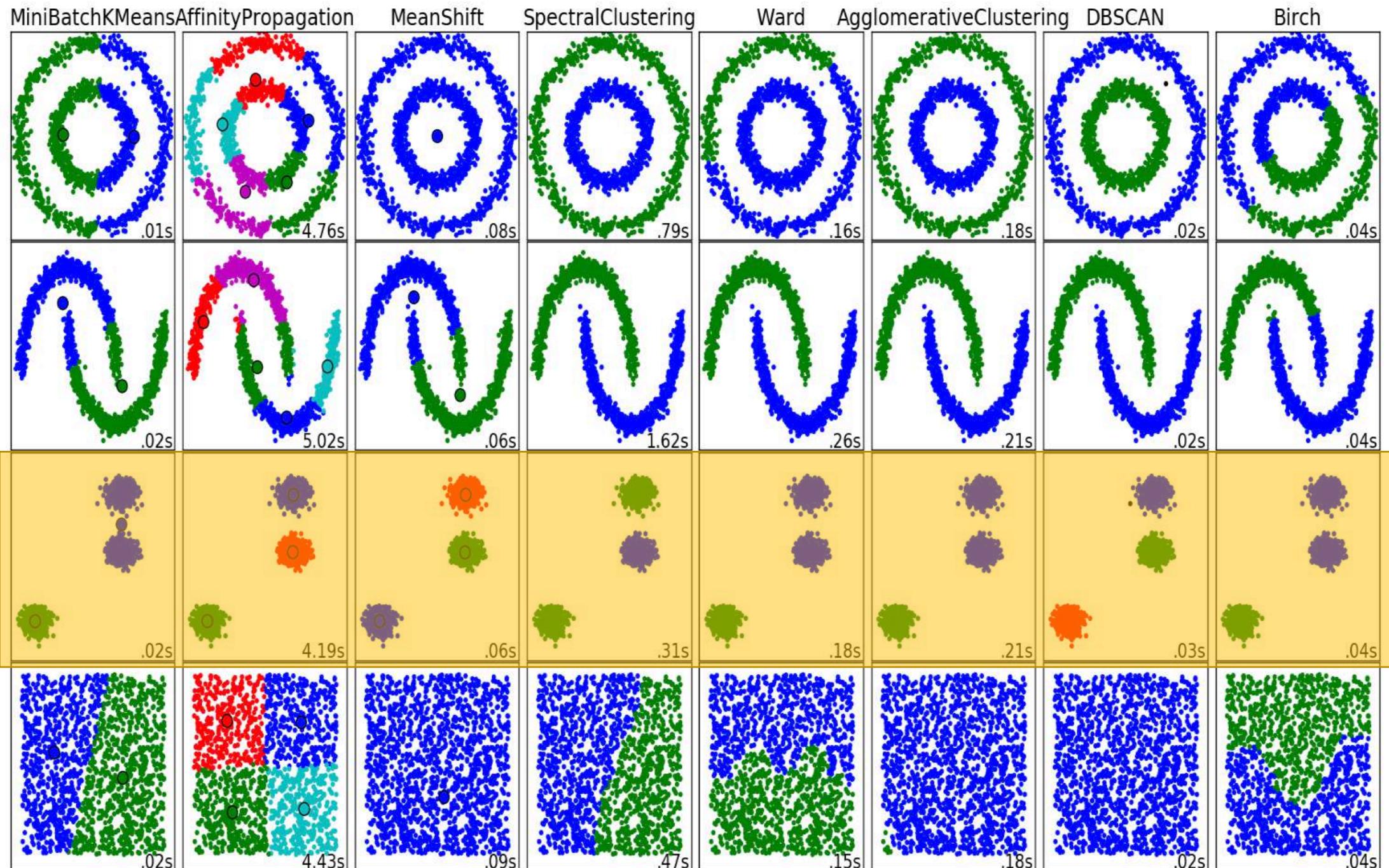
В scikit-learn:

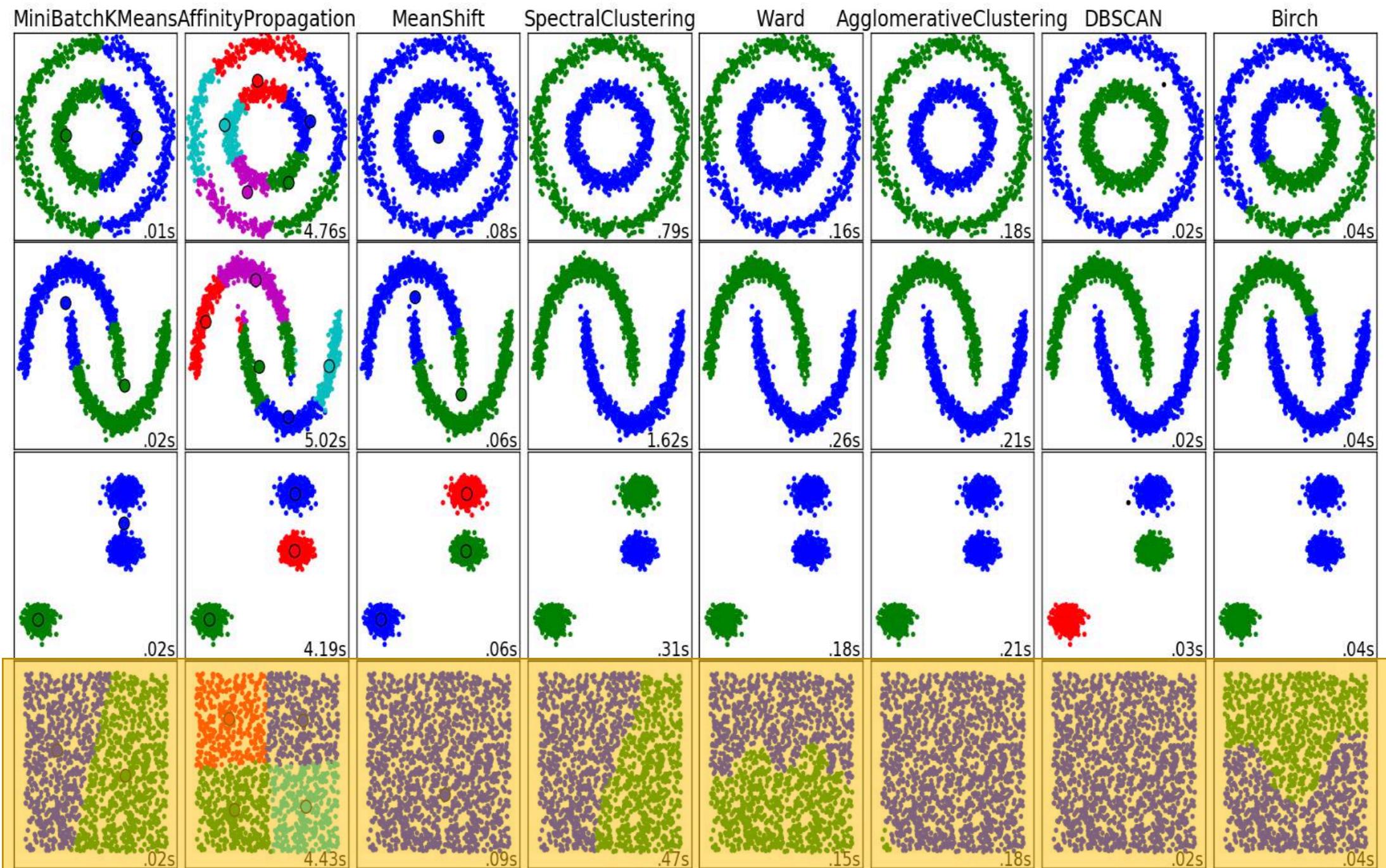
KMeans, MiniBatchKMeans, GaussianMixture,
AgglomerativeClustering, Ward, DBSCAN, MeanShift,
AffinityPropagation, SpectralClustering, Birch

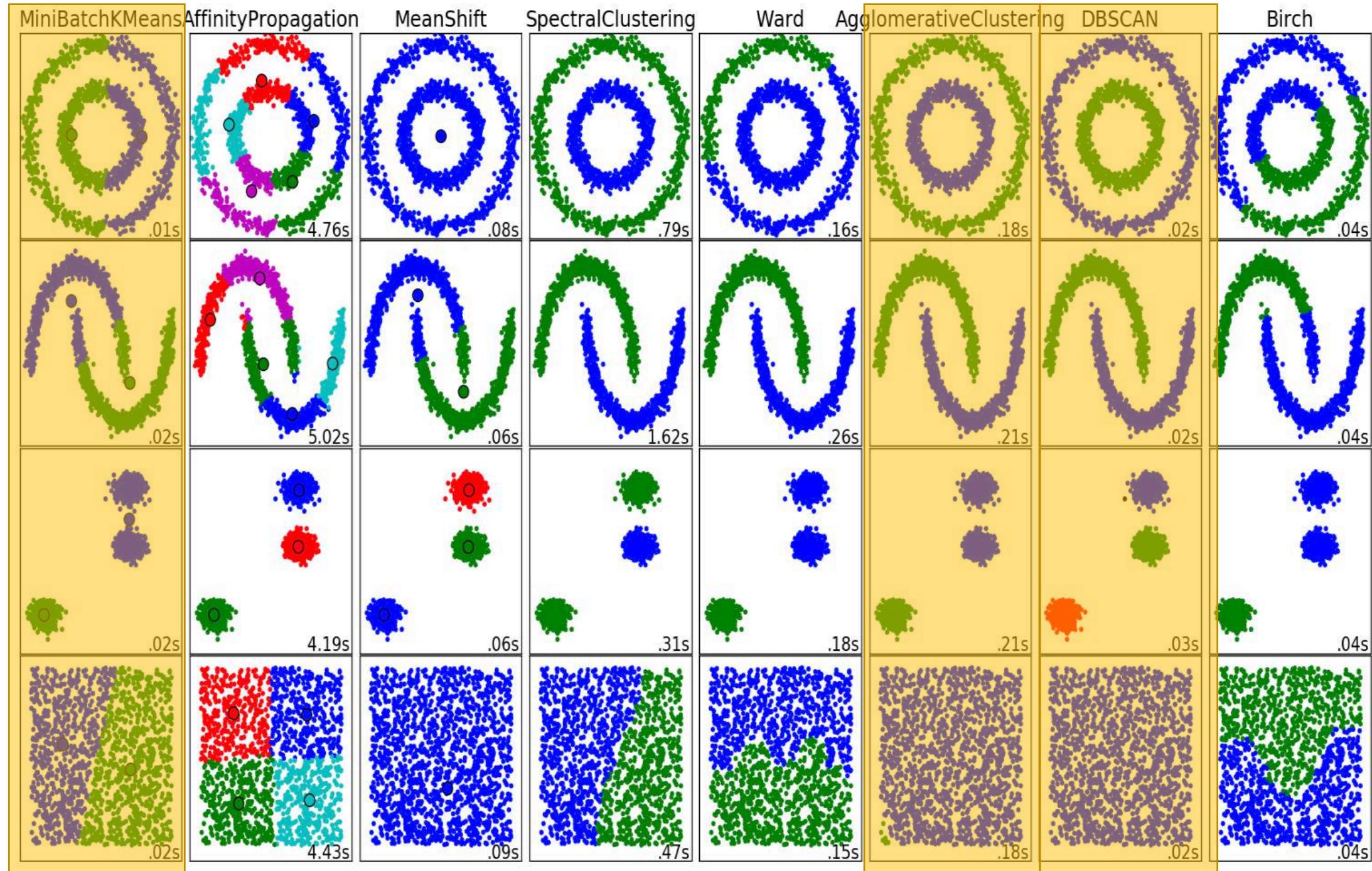












Оценка качества

С фиксированным количеством кластеров и признаками:

- Среднее межкластерное/внутрикластерное расстояние

Оценка качества

С фиксированным количеством кластеров и признаками:

- Среднее межклusterное/внутрикластерное расстояние

Если нужно подобрать количество кластеров:

- Silhouette coefficient

Оценка качества

С фиксированным количеством кластеров и признаками:

- Среднее межкластерное/внутрикластерное расстояние

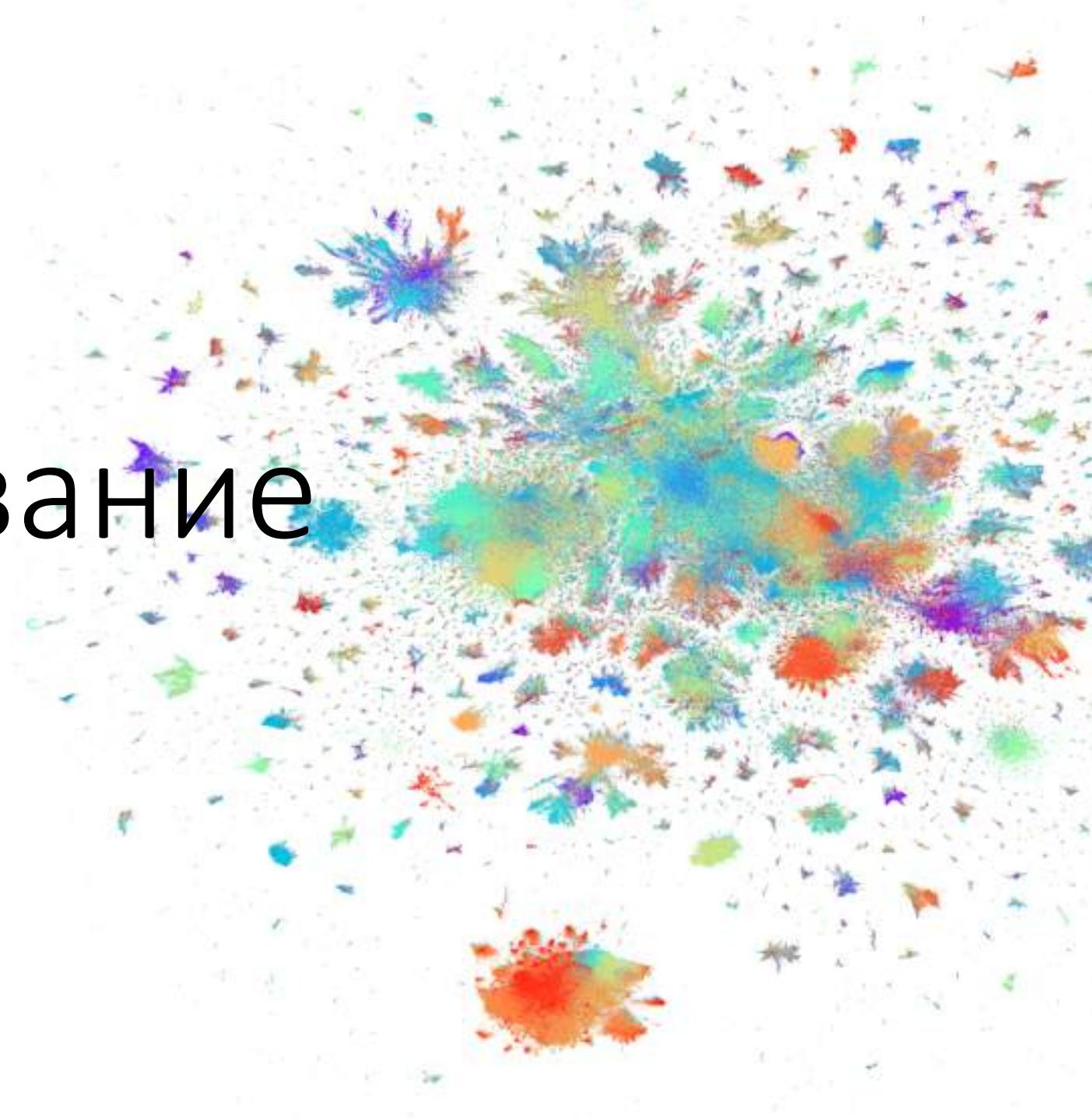
Если нужно подобрать количество кластеров:

- Silhouette coefficient

Если нужно подобрать признаки:

- Однородность и полнота (homogeneity & completeness)
- V-мера
- Ответы ассессоров на вопросы

Преобразование признаков



План

1. Генерация признаков
2. Отбор признаков
3. Преобразование признаков:
 - Метод главных компонент и SVD,
 - Manifold learning

I. Генерация признаков

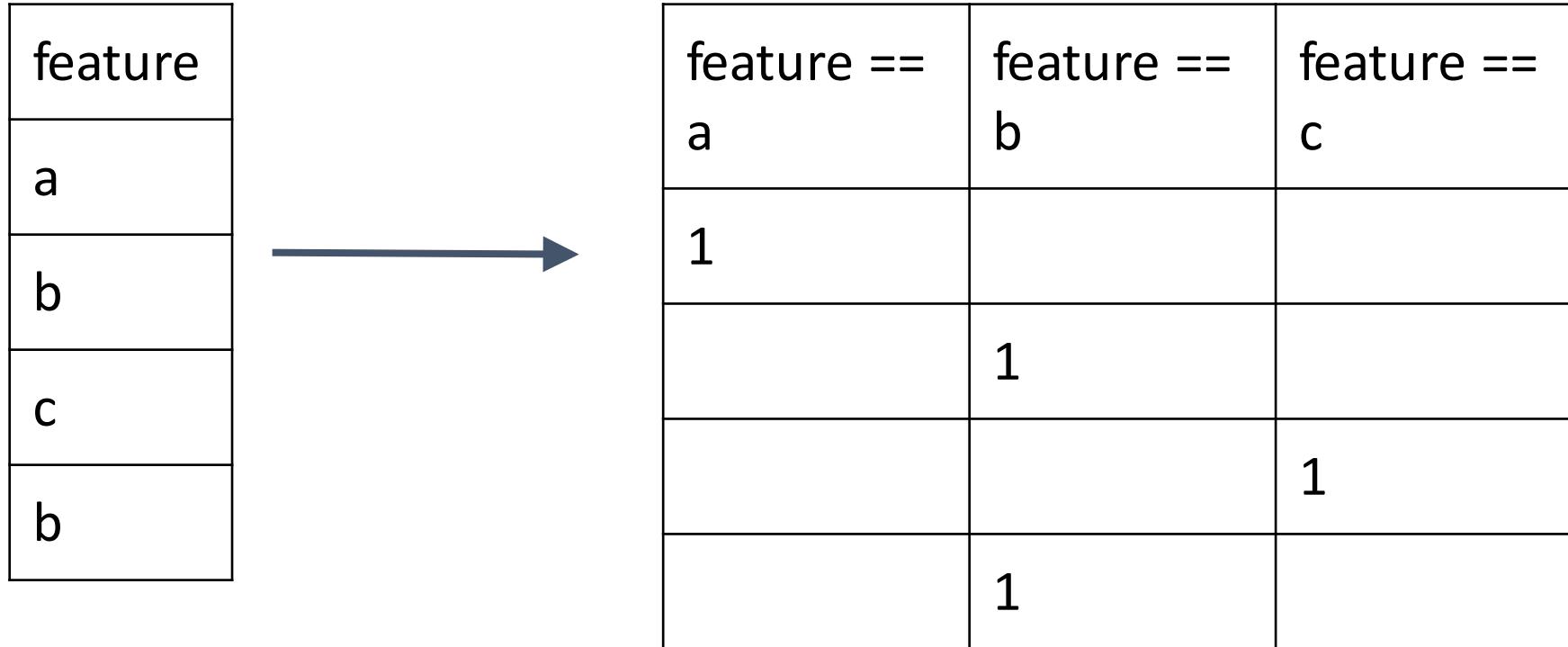
Виды признаков

Какие бывают признаки:

1. Числовые
2. Порядковые
3. Категориальные
4. Даты и время
5. Координаты

Категориальные признаки

Бинаризация



Категориальные признаки

№ склада	Город	...	Продано вина (ящиков)
2343	Москва	...	56000
185	Самара	...	10500
121	Ростов	...	11300
...

Категориальные признаки

№ склада	В среднем продают в городе	...	Продано вина (ящиков)
2343	59000	...	56000
185	11200	...	10500
121	12100	...	11300
...

Feature engineering

- Генерация признаков (feature generation) – генерация признаков по известным данным
- Отбор признаков (feature selection) – ранжирование признаков по «полезности» и выкидывание наименее полезных (или даже наоборот «вредных»)
- Преобразование признаков (feature transform) – создание новых признаков на основе имеющихся

Пример 1: текстовые признаки

- Dataset: 20news_groups
- Электронные письма, разбитые по 20 темам (классам)
- Попробуем придумать классификатор, который различает две темы: **auto** и **politics.mideast**

Извлечение текстовых признаков

- Пример письма 1:

From: carl_f_hoffman@cup.portal.com

Newsgroups: rec.autos

Subject: 1993 Infiniti G20

Message-ID: <78834@cup.portal.com>

Date: Mon, 5 Apr 93 07:36:47 PDT

Organization: The Portal System (TM)

Lines: 26

I am thinking about getting an Infiniti G20.

In consumer reports it is ranked high in many
catagories including highest in reliability index for compact cars.
(Mitsubishi Galant was second followed by Honda Accord)

Извлечение текстовых признаков

- Пример письма 2:

From: Bob.Waldrop@f418.n104.z1.fidonet.org (Bob Waldrop)

Subject: Celebrate Liberty! 1993

Message-ID: <1993Apr5.201336.16132@dsd.es.com>

Followup-To:talk.politics.misc

Announcing... Announcing... Announcing... Announcing...

CELEBRATE LIBERTY!
1993 LIBERTARIAN PARTY NATIONAL CONVENTION
AND POLITICAL EXPO

THE MARRIOTT HOTEL AND THE SALT PALACE
SALT LAKE CITY, UTAH

Текстовые признаки: bag-of-words



The world of **TOTAL**

- » All About The Company
 - Global Activities
 - Corporate Structure
 - TOTAL's Story
 - Upstream Strategy
 - Downstream Strategy
 - Chemicals Strategy
 - TOTAL Foundation
 - Homepage

all about the company

Our strong exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

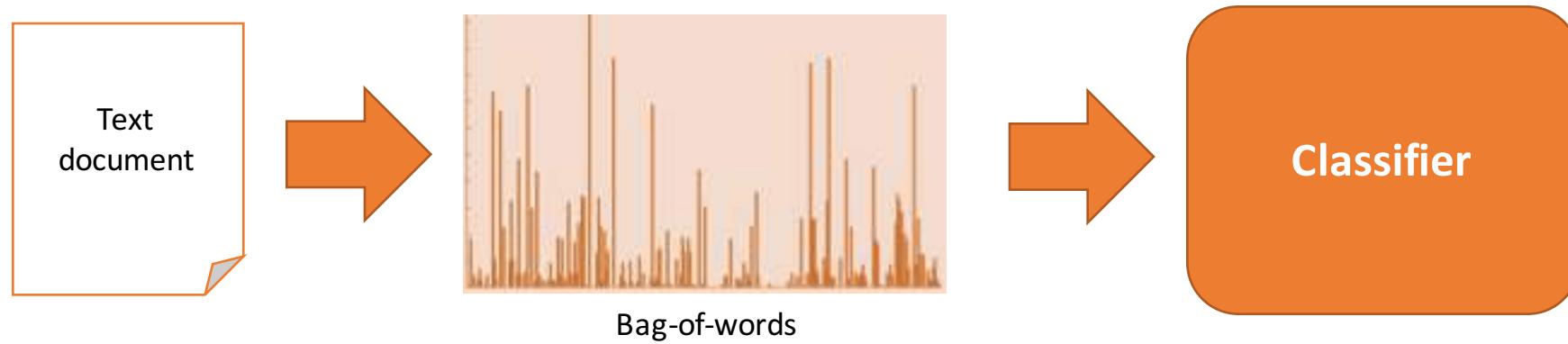
At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Sea complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profits to the core energy business.

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	0

Простой классификатор текстов



Взвешивание частот слов в текстах

Term Frequency

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

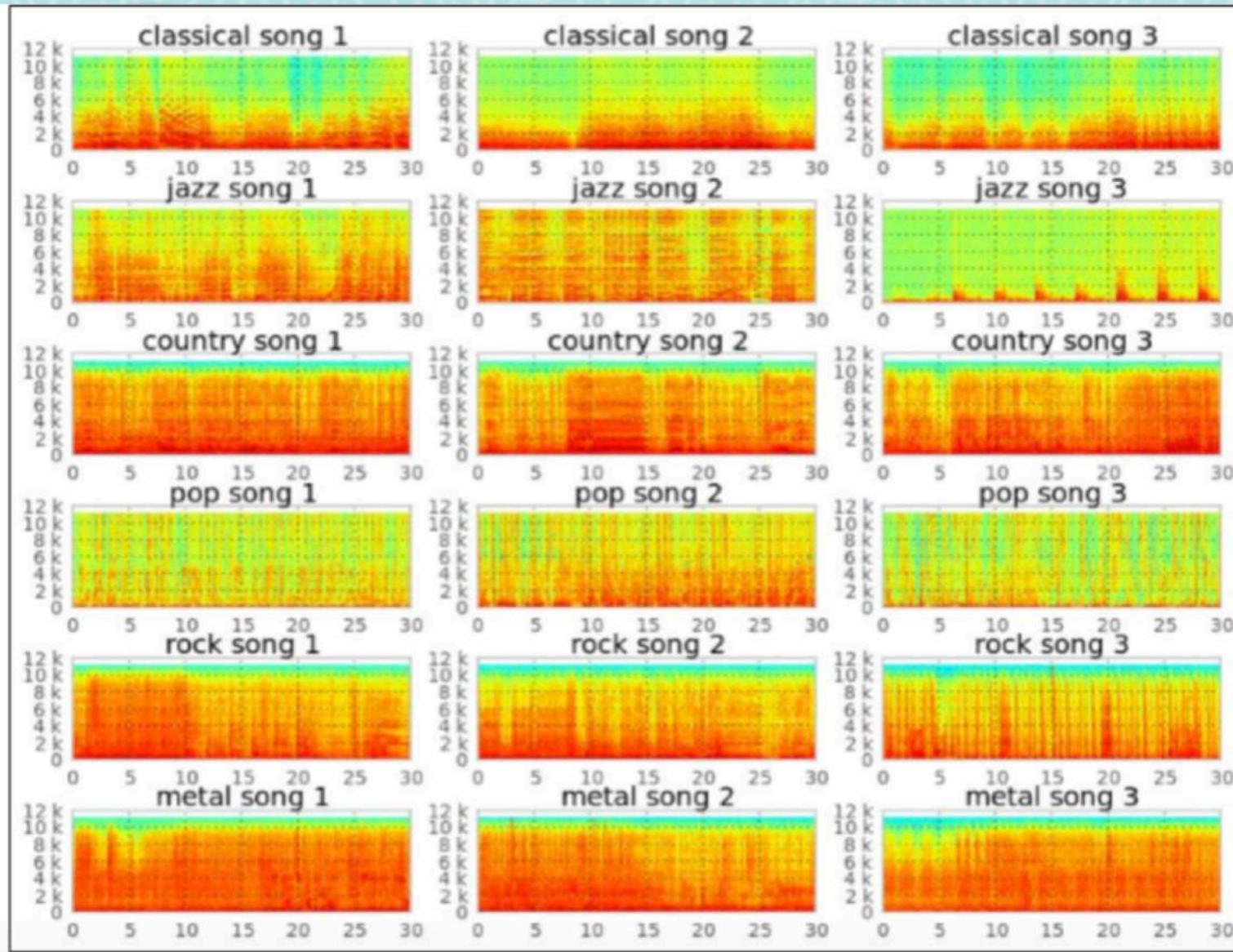
Inverse Document Frequency

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

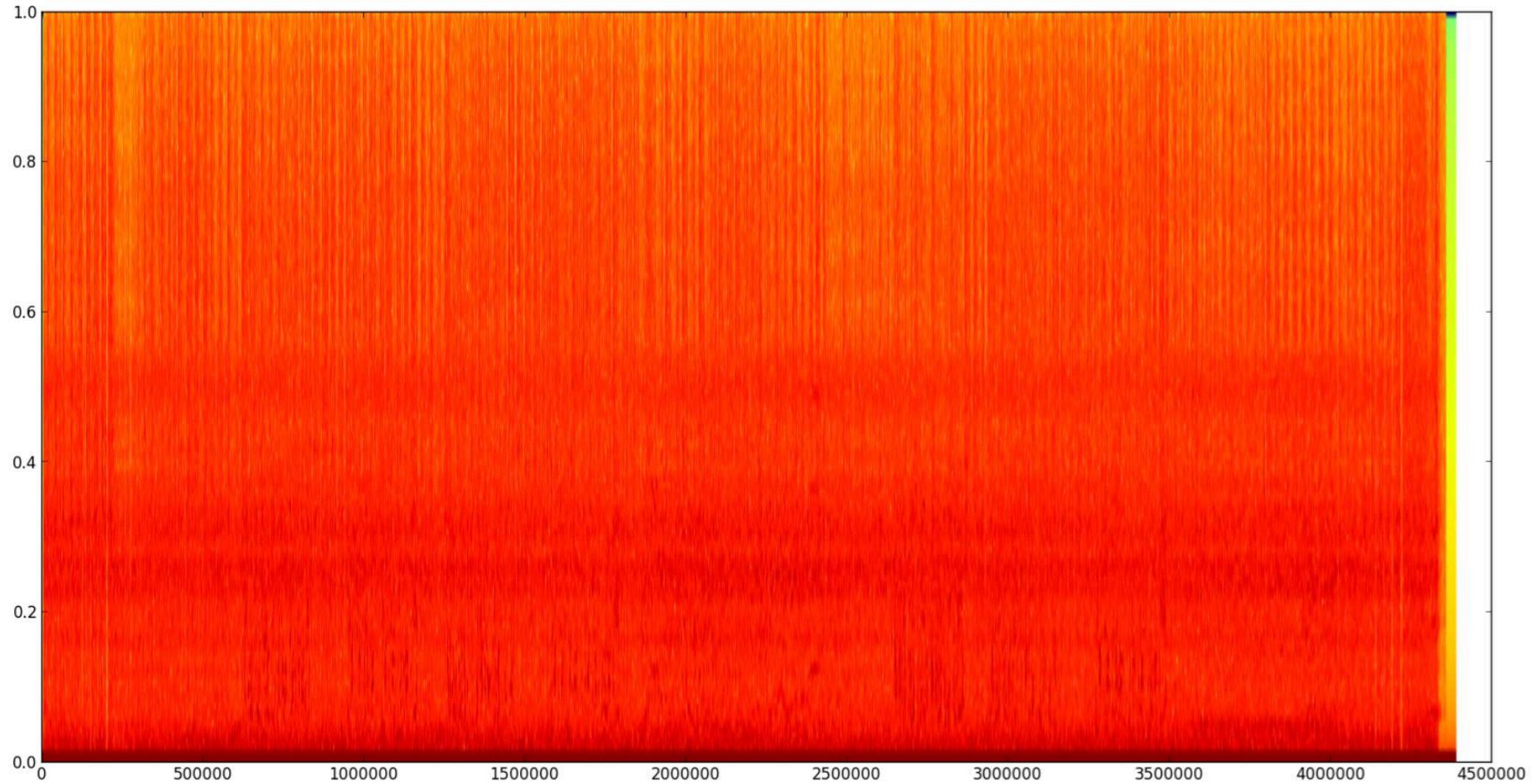
Частоты N-грамм

- $N = 1$: This is a sentence *unigrams:* this,
is,
a,
sentence
- $N = 2$: This is a sentence *bigrams:* this is,
is a,
a sentence
- $N = 3$: This is a sentence *trigrams:* this is a,
is a sentence

Пример 2: признаки аудиофайла

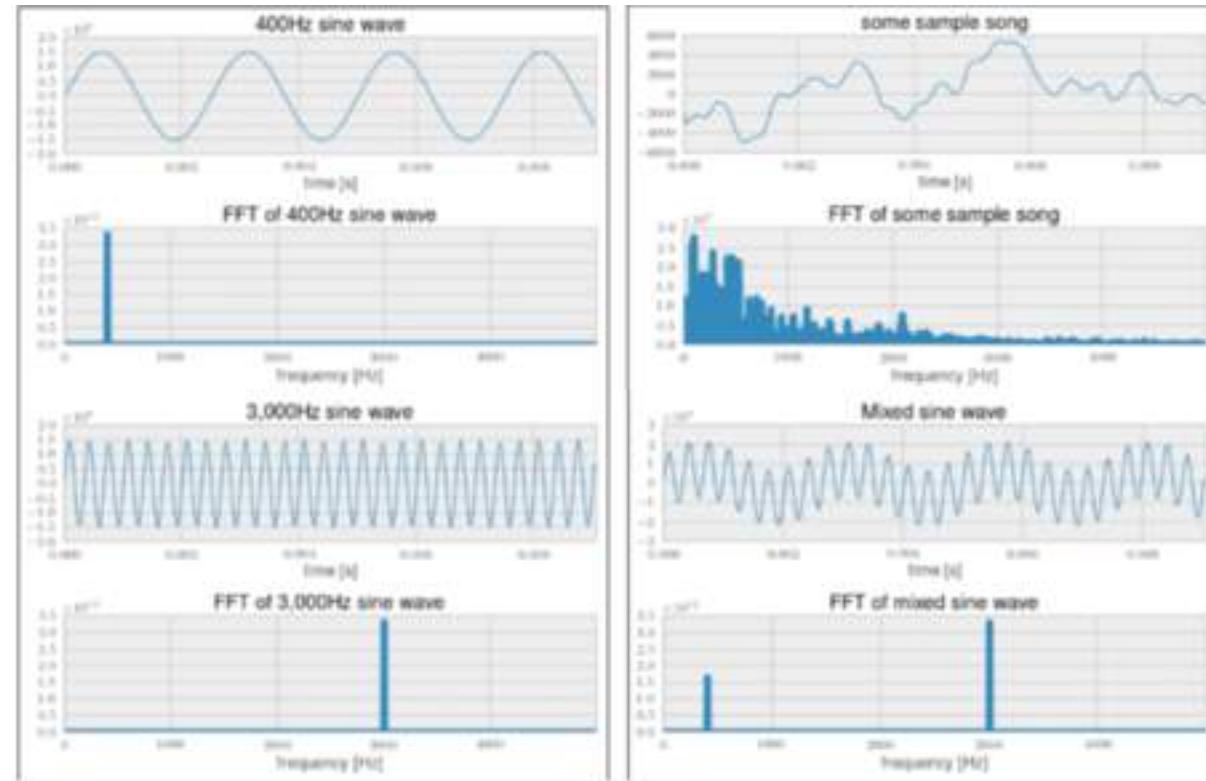


Пример 2: признаки аудиофайла



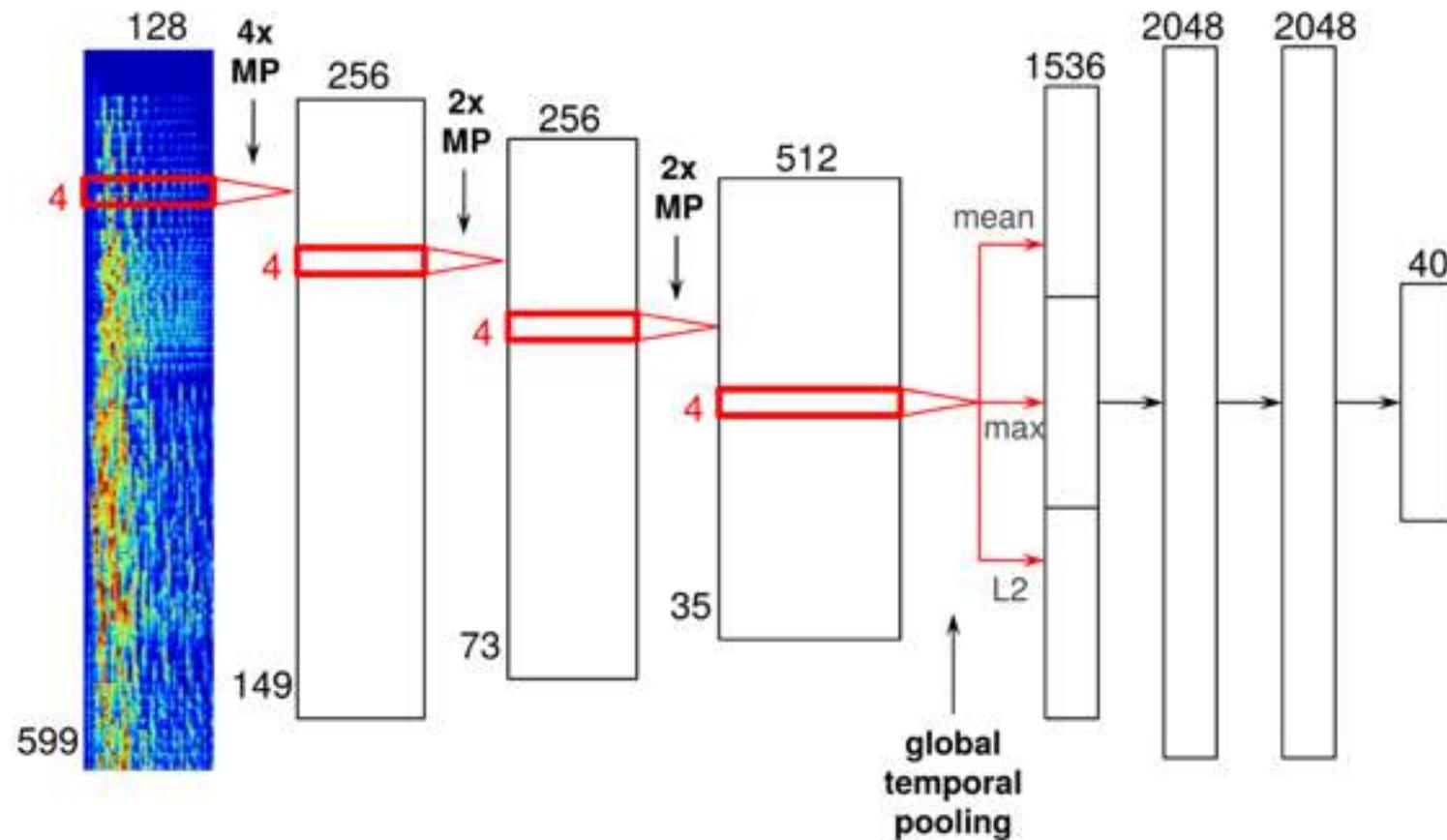
Пример 2: признаки аудиофайла

MFCC - преобразование Фурье логарифма спектра



Пример 2: признаки аудиофайла

Embeddings с помощью нейронных сетей:



Пример 3: признаки изображения

Input image

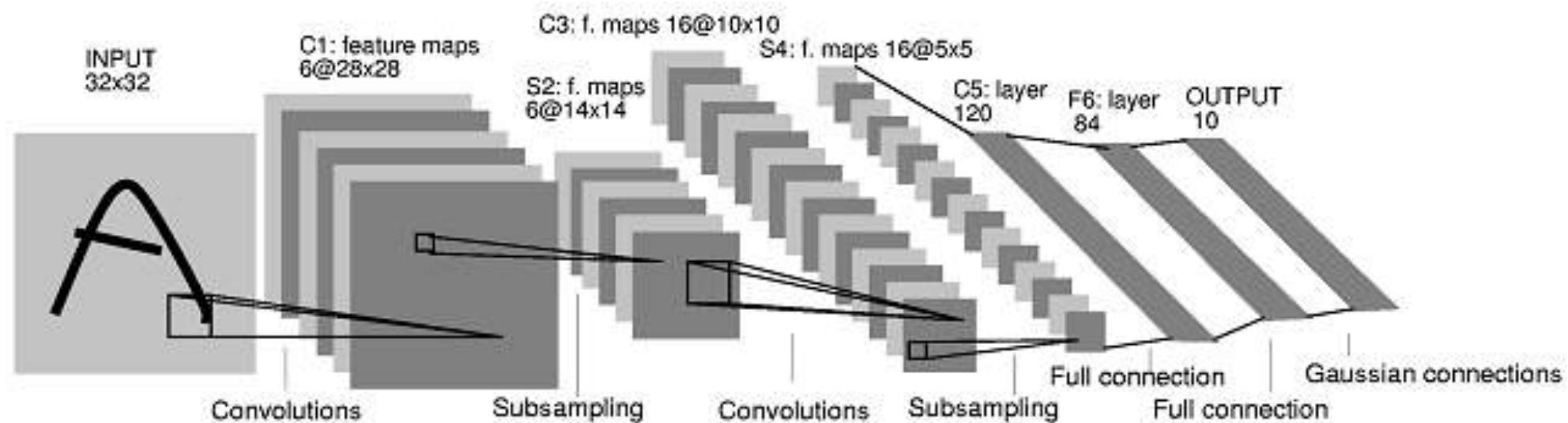


Histogram of Oriented Gradients



Пример 3: признаки изображения

Выходы слоев из нейросети



II. Отбор признаков

Отбор признаков

1. Статистические методы
2. С помощью регуляризации L1
3. Жадный отбор
4. С помощью моделей

Отбор признаков по статистическим критериям

Пример: критерий хи-квадрат позволяет отобрать лучшие бинарные признаки для каждого класса

	Значение признака 1	Значение признака 0
Объект принадлежит классу	A	B
Объект не принадлежит классу	C	D

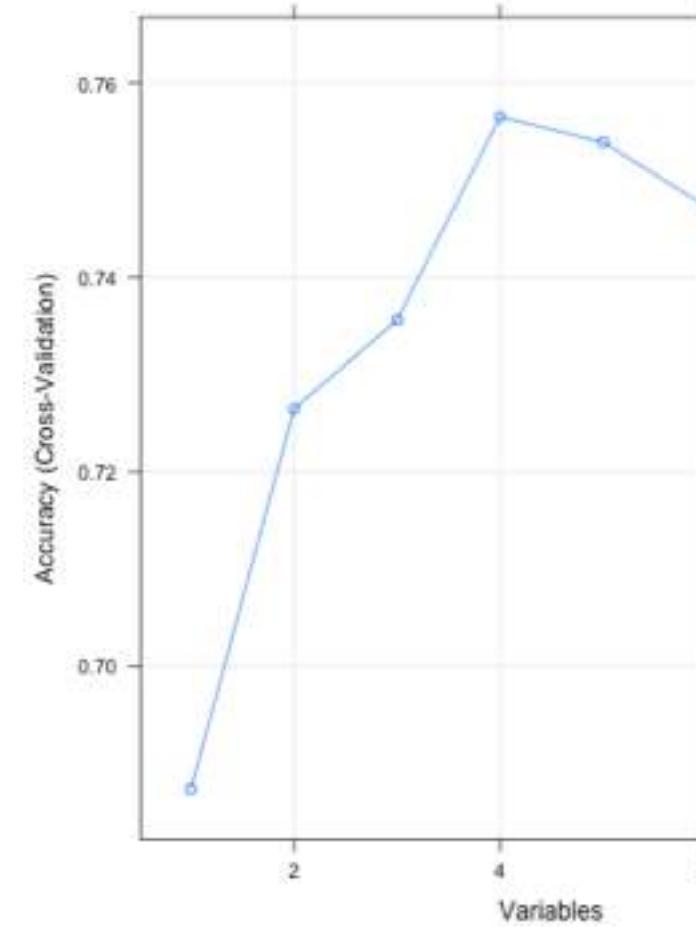
$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(A + C)(B + D)(C + D)}$$

Отбор признаков с помощью |1-регуляризации

$$\sum_{i=1}^l L(M_i) + \gamma \sum_{k=1}^m |w_k| \rightarrow \min$$

Жадный отбор признаков

Чередование добавления и удаления
признаков

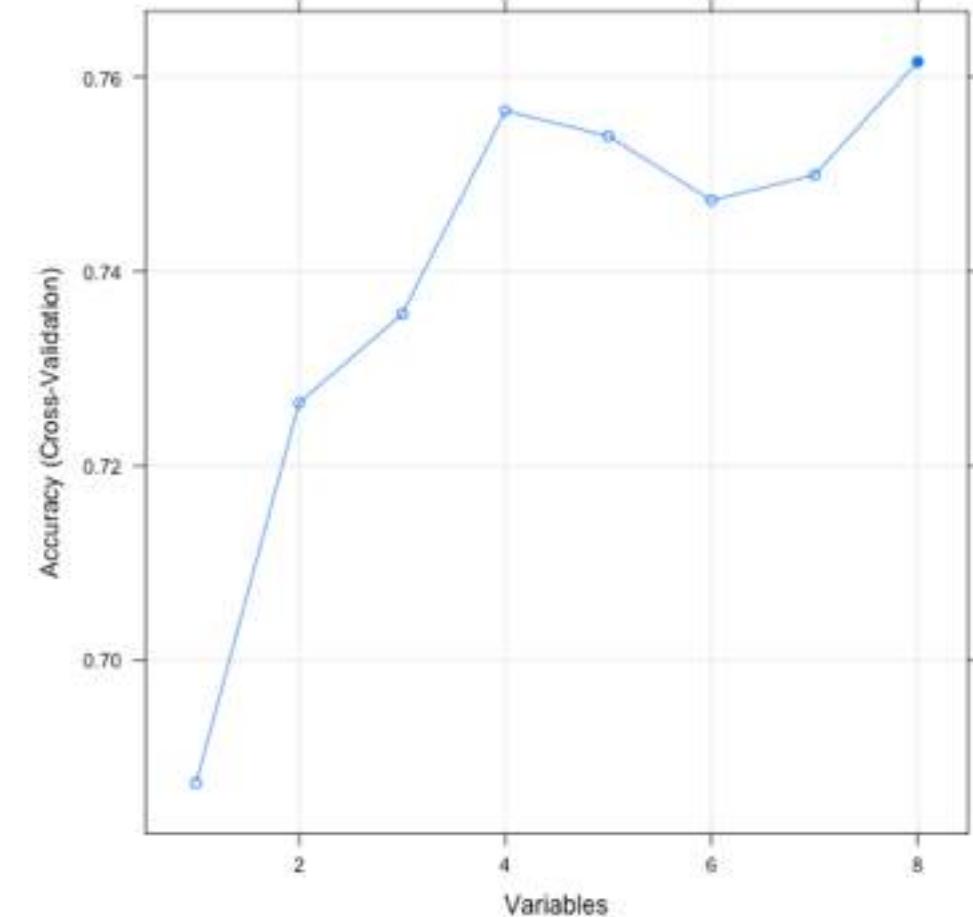


Жадный отбор признаков

Чередование добавления и удаления
признаков

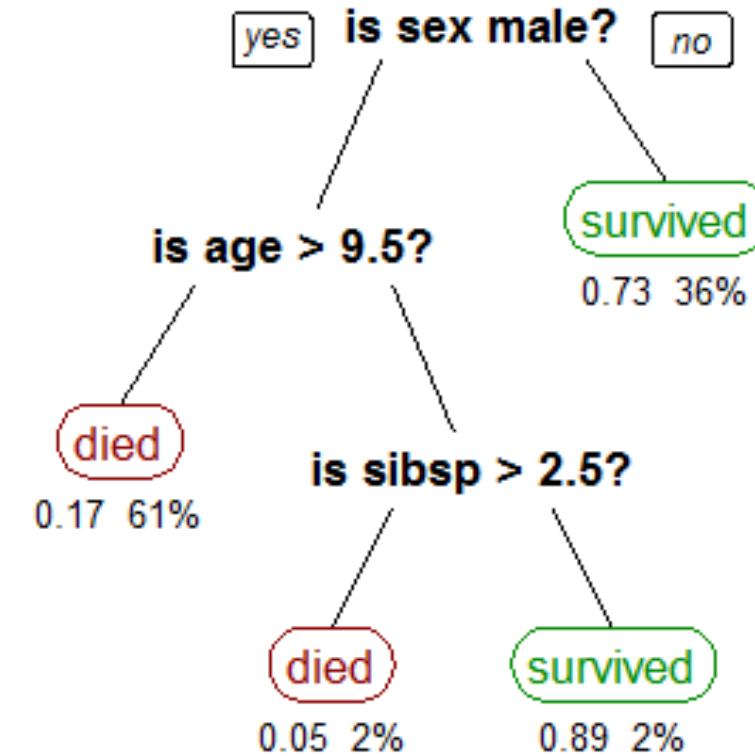
Этап добавления: добавляем лучшие
признаки

Этап удаления: удаляем худшие
признаки



Отбор признаков с помощью моделей

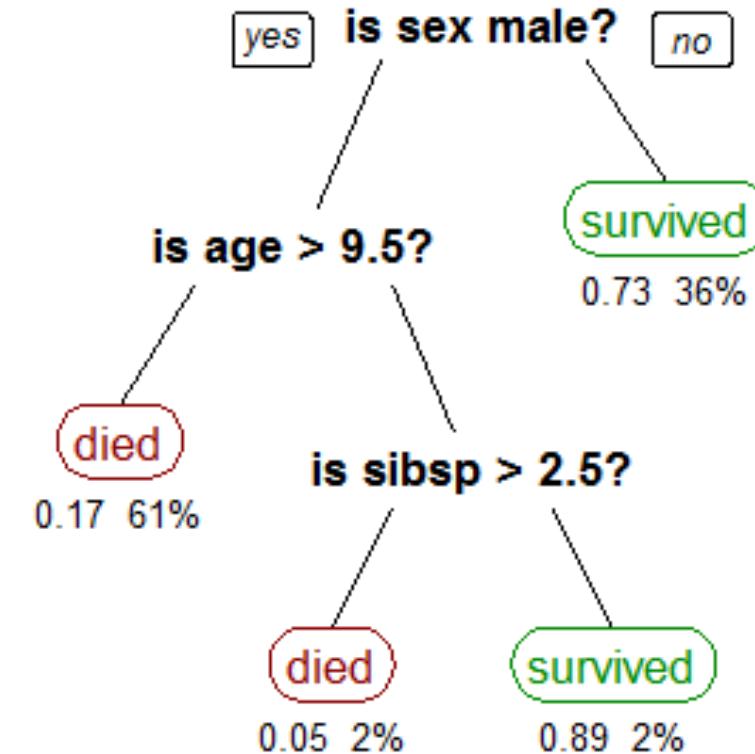
Вопрос: как можно оценивать
важность признака в решающих
деревьях?



Отбор признаков с помощью моделей

Вопрос: как можно оценивать
важность признака в решающих
деревьях?

А в линейных моделях?



III. Преобразование признаков

Преобразование признаков

1. Задача понижения размерности
2. Метод главных компонент и SVD
3. Manifold learning

Как выглядит обучающая выборка

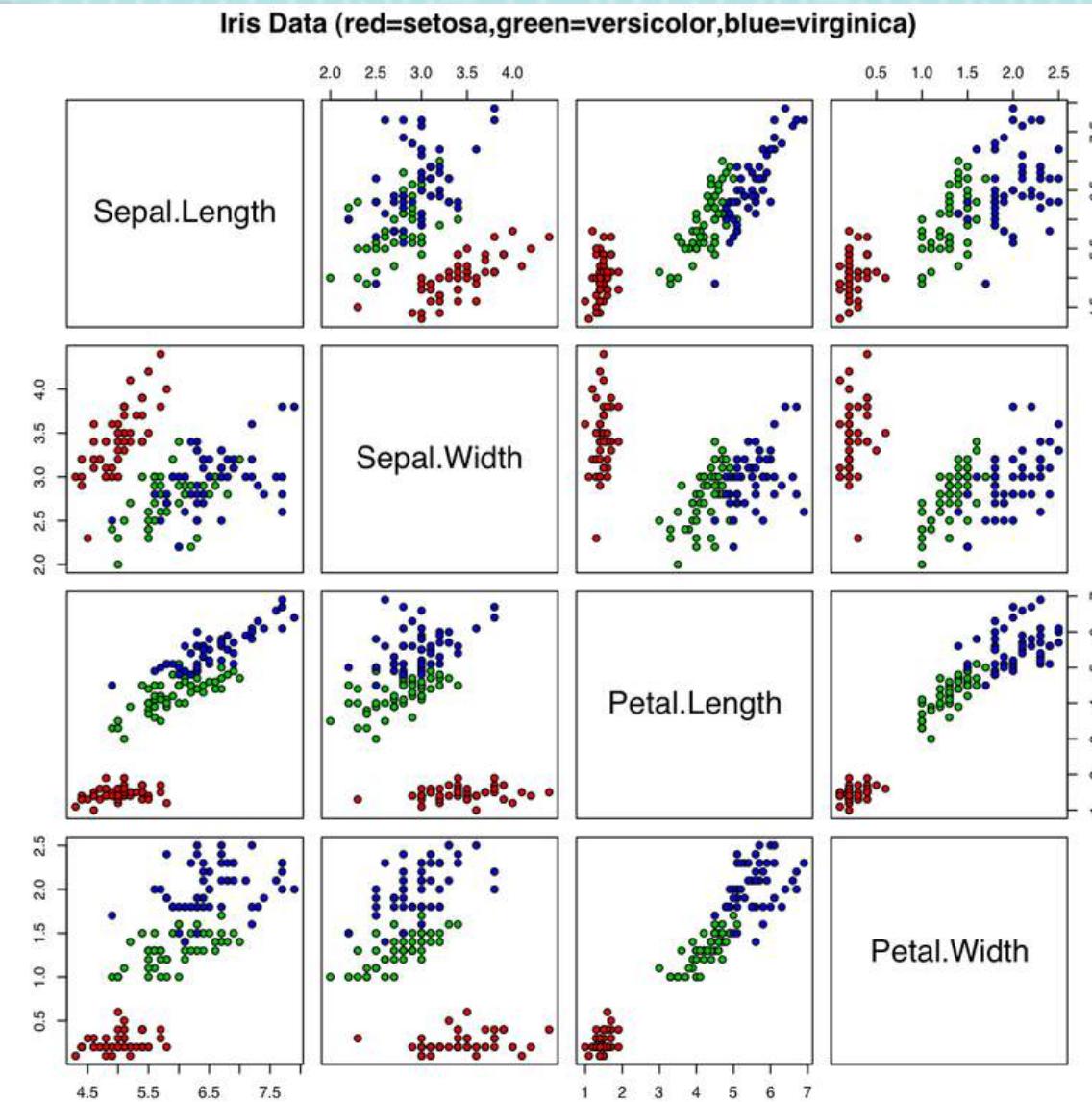
Fisher's Iris Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	<i>I. versicolor</i>
6.0	2.2	5.0	1.5	<i>I. virginica</i>
6.2	2.2	4.5	1.5	<i>I. versicolor</i>
6.0	2.2	4.0	1.0	<i>I. versicolor</i>
6.3	2.3	4.4	1.3	<i>I. versicolor</i>
5.5	2.3	4.0	1.3	<i>I. versicolor</i>
5.0	2.3	3.3	1.0	<i>I. versicolor</i>
4.5	2.3	1.3	0.3	<i>I. setosa</i>
5.5	2.4	3.8	1.1	<i>I. versicolor</i>
5.5	2.4	3.7	1.0	<i>I. versicolor</i>
4.9	2.4	3.3	1.0	<i>I. versicolor</i>
6.7	2.5	5.8	1.8	<i>I. virginica</i>
5.7	2.5	5.0	2.0	<i>I. virginica</i>
6.3	2.5	5.0	1.9	<i>I. virginica</i>
6.3	2.5	4.9	1.5	<i>I. versicolor</i>
4.9	2.5	4.5	1.7	<i>I. virginica</i>

Что хотелось бы уметь

- Визуализировать обучающую выборку, когда признаков больше трёх
- Уменьшать количество признаков, переходя к новым, более информативным

Визуализируем выборку

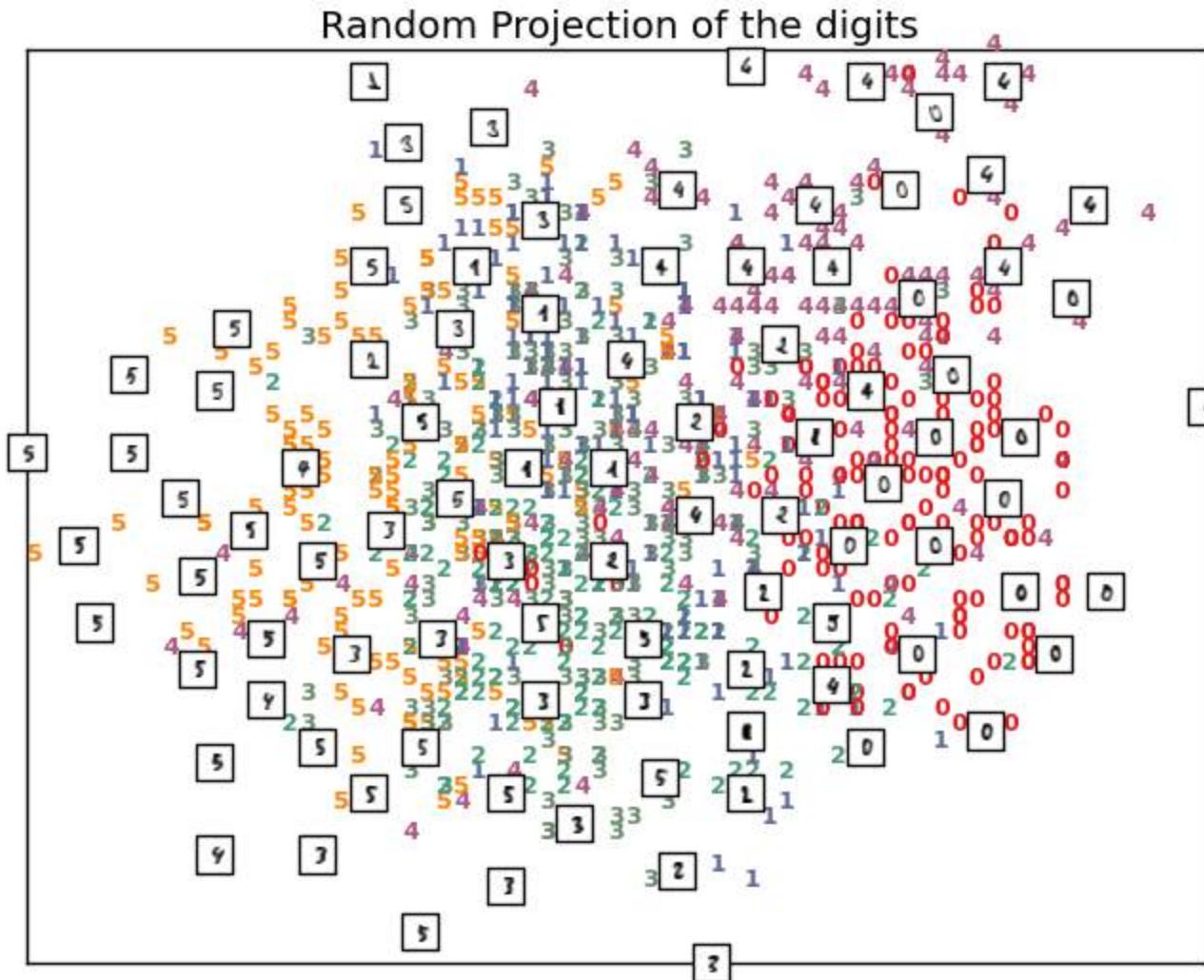


Более сложный случай

Что делать, если признаков еще больше?

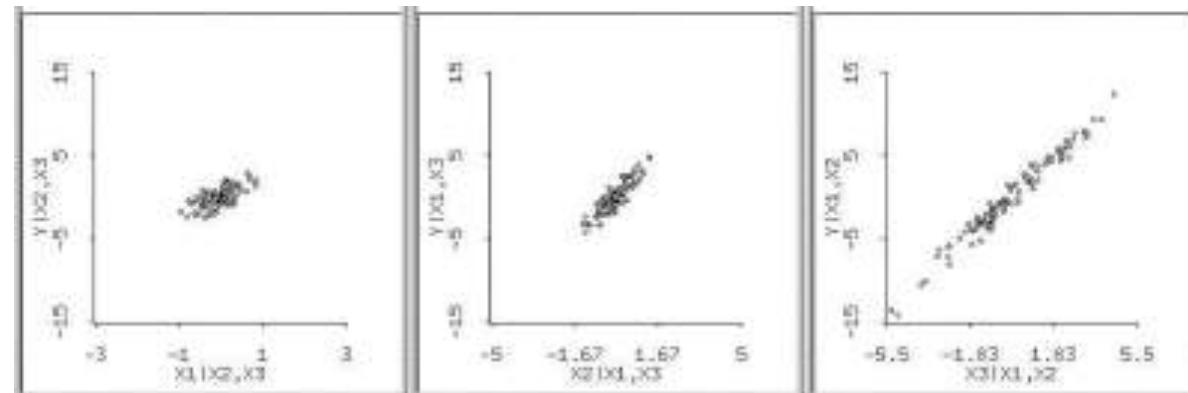
1	6	4	12	5	5	3	4	1	67	3	2	1	2	1	0	0	1	0	0	0	1	0	0	1	1	
2	48	2	60	1	3	2	2	1	22	3	1	1	1	1	0	0	1	0	0	0	1	0	0	1	2	
4	12	4	21	1	4	3	3	1	49	3	1	2	1	1	0	0	1	0	0	0	1	0	0	1	1	
1	42	2	79	1	4	3	4	2	45	3	1	2	1	1	0	0	0	0	0	0	0	0	0	1	1	
1	24	3	49	1	3	3	4	4	53	3	2	2	1	1	1	0	1	0	0	0	0	0	0	1	2	
4	36	2	91	5	3	3	4	4	35	3	1	2	2	1	0	0	1	0	0	0	0	0	1	0	1	
4	24	2	28	3	5	3	4	2	53	3	1	1	1	1	0	0	1	0	0	0	1	0	0	1	1	
2	36	2	69	1	3	3	2	3	35	3	1	1	2	1	0	1	1	0	1	0	0	0	0	0	1	
4	12	2	31	4	4	1	4	1	61	3	1	1	1	1	0	0	1	0	0	0	1	0	0	1	0	
2	30	4	52	1	1	4	2	3	28	3	2	1	1	1	1	0	1	0	0	0	1	0	0	0	2	
2	12	2	13	1	2	2	1	3	25	3	1	1	1	1	1	0	1	0	0	1	0	0	0	1	2	
1	48	2	43	1	2	2	4	2	24	3	1	1	1	1	0	0	1	0	0	1	0	0	0	1	2	
2	12	2	16	1	3	2	1	3	22	3	1	1	2	1	0	0	1	0	0	0	1	0	0	1	1	
1	24	4	12	1	5	3	4	3	60	3	2	1	1	1	1	0	1	0	0	0	1	0	0	2		
1	15	2	14	1	3	2	4	3	28	3	1	1	1	1	1	0	1	0	0	1	0	0	0	1	1	
1	24	2	13	2	3	2	2	3	32	3	1	1	1	1	1	0	0	1	0	0	0	1	0	0	2	
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	0	0	1	0	0	0	1	0	0	1	1	
1	30	0	81	5	2	3	3	3	25	1	3	1	1	1	1	0	0	1	0	0	0	1	0	0	1	1
2	24	2	126	1	5	2	2	4	44	3	1	1	2	1	0	1	1	0	0	0	0	0	0	0	0	2
4	24	2	34	3	5	3	2	3	31	3	1	2	2	1	0	0	1	0	0	0	1	0	0	0	1	1
4	9	4	21	1	3	3	4	3	48	3	3	1	2	1	1	0	1	0	0	0	1	0	0	1	1	
1	6	2	26	3	3	3	3	1	44	3	1	2	1	1	0	0	1	0	0	1	0	0	0	1	1	
1	10	4	22	1	2	3	3	1	48	3	2	2	1	2	1	0	1	0	0	1	0	0	0	1	0	
2	12	4	18	2	2	3	4	2	44	3	1	1	1	1	0	1	1	0	0	0	1	0	0	0	1	1
4	10	4	21	5	3	4	1	3	26	3	2	1	1	2	0	0	1	0	0	0	1	0	0	0	1	1
1	6	2	14	1	3	3	2	1	36	1	1	1	2	1	0	0	1	0	0	0	1	0	0	1	0	
4	6	0	4	1	5	4	4	3	39	3	1	1	1	1	0	0	1	0	0	0	1	0	0	1	0	
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	1	0	0	1	0	0	1	0	0	0	1	1
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	1	0	0	0	0	0	1	0	0	0	1	1
1	60	3	68	1	5	3	4	4	63	3	2	1	2	1	0	0	1	0	0	0	1	0	0	0	1	2
2	18	2	19	4	2	4	3	1	36	1	1	1	2	1	0	0	1	0	0	0	1	0	0	0	1	1
1	24	2	40	1	3	3	2	3	27	2	1	1	1	1	1	0	0	1	0	0	0	1	0	0	1	1

Пример случайной проекции для рукописных цифр



Проблемы «лишних признаков»

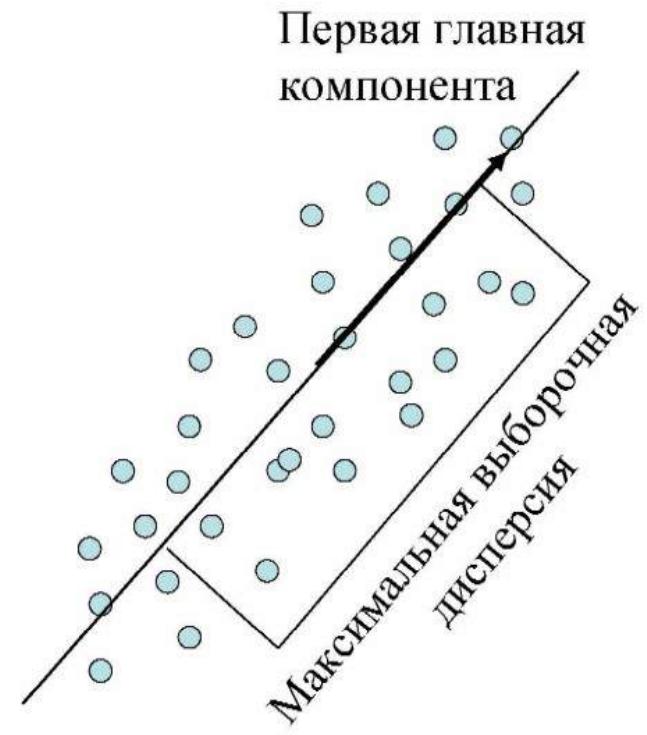
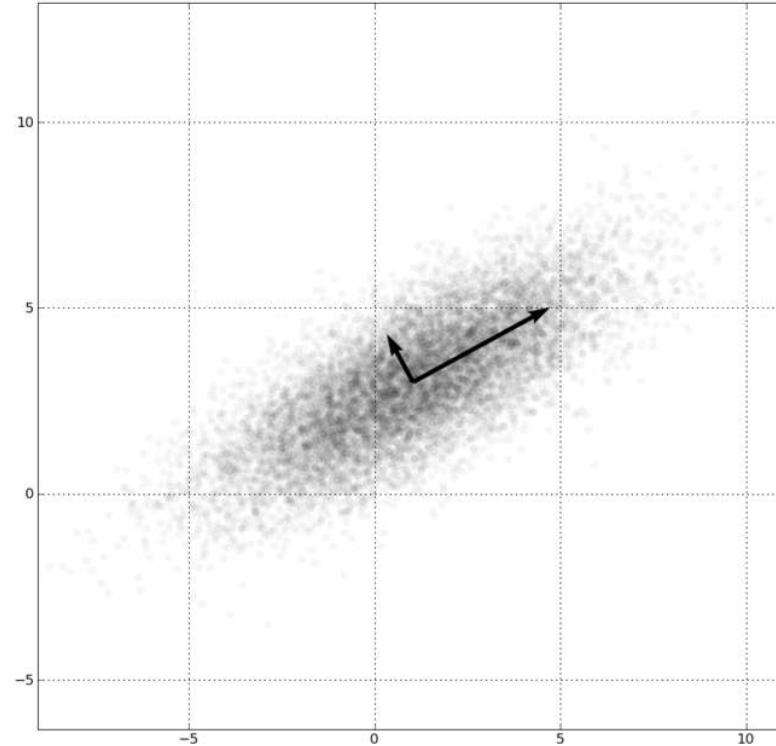
Если признаки сильно коррелированы, то у многих методов машинного обучения будут проблемы (например, из-за неустойчивости обращения матрицы ковариаций, где это нужно)



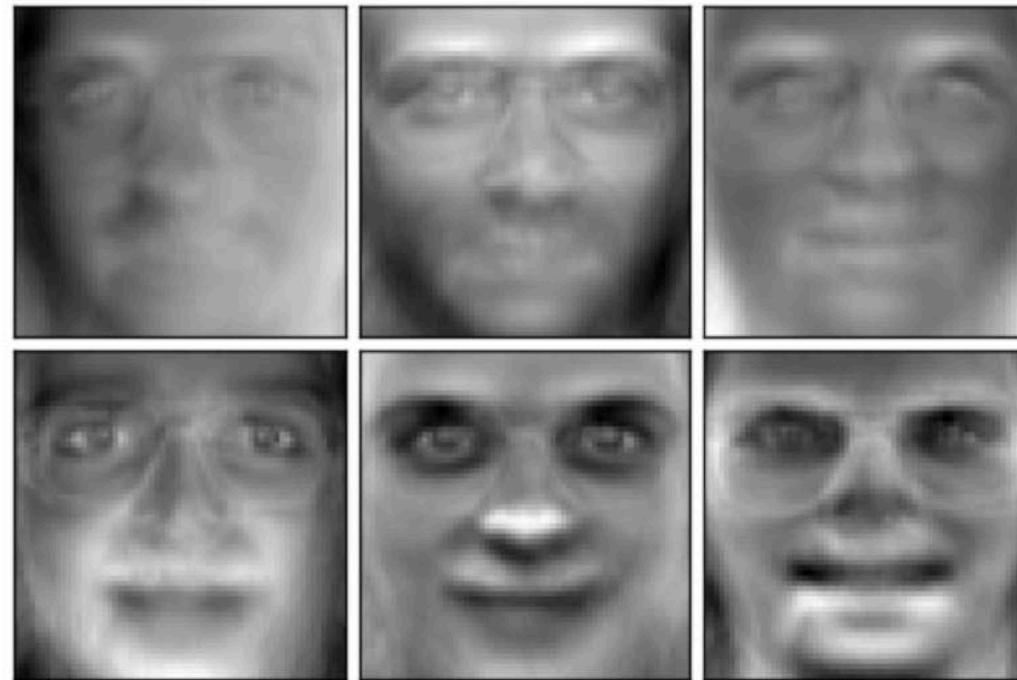
Principal Component Analysis 1

- Идея 1: давайте выделять в пространстве признаков направления, вдоль которых разброс точек наибольший (они кажутся наиболее информативными)

PCA (интерпретация 1)



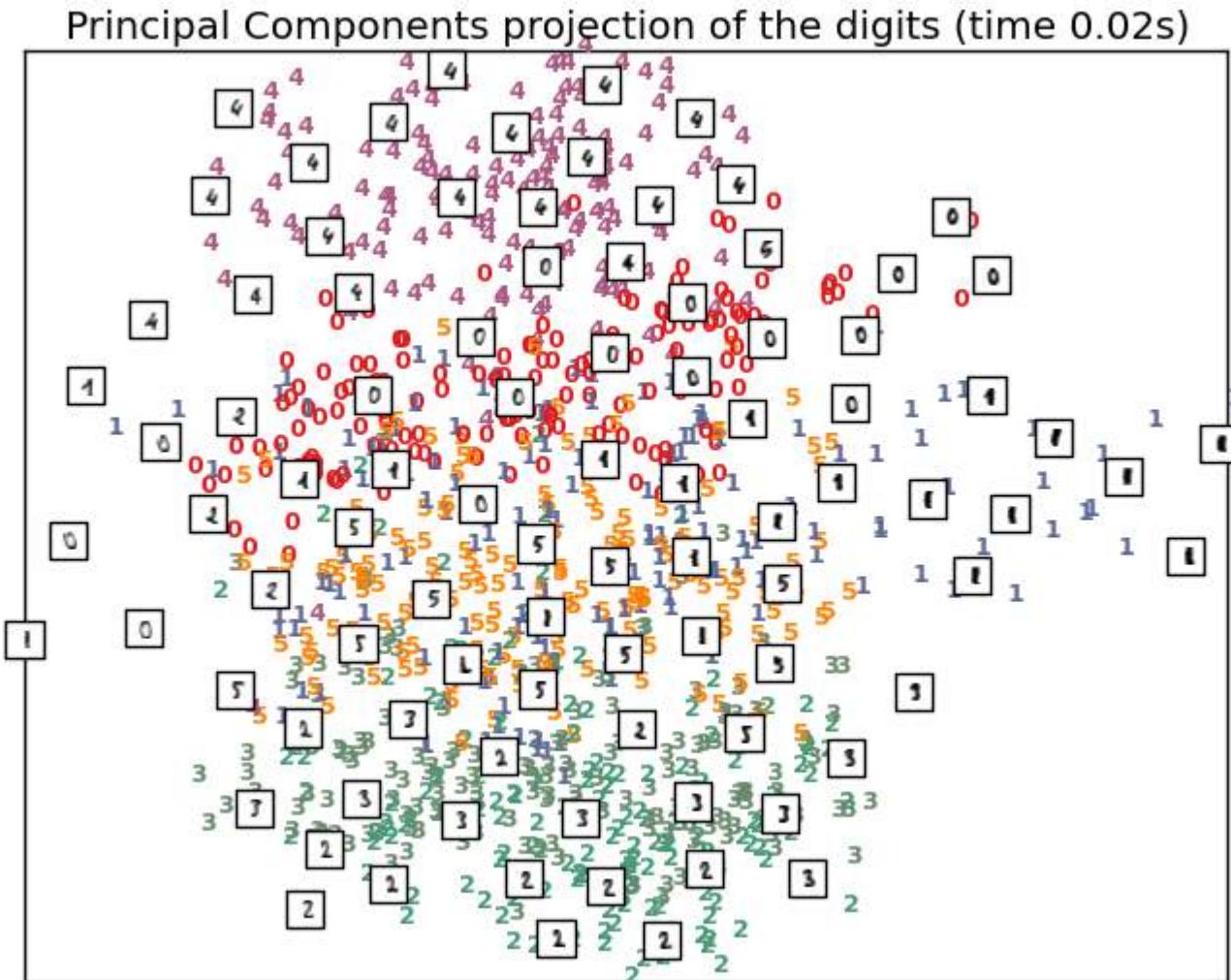
Пример: eigenfaces



$$= \text{mean} + 0.9 * \text{face 1} - 0.2 * \text{face 2} + 0.4 * \text{face 3} + \dots$$



Рукописные цифры: проекция на главные компоненты



PCA (интерпретация 2)

Приблизим исходную матрицу признаков произведением двух матриц:

$$X \approx U \cdot V^T$$

$l \times n \quad l \times k \quad k \times n$

$$\| X - U \cdot V^T \| \rightarrow \min$$

PCA: как сделать?

- Центрируем выборку (из каждого признака вычитаем среднее значение), получаем матрицу X с новыми значениями признаков
- Делаем SVD-разложение матрицы X :

$$X \approx A \cdot \Lambda \cdot B^T$$

Выбираем $U = A \cdot \Lambda$, $V = B$

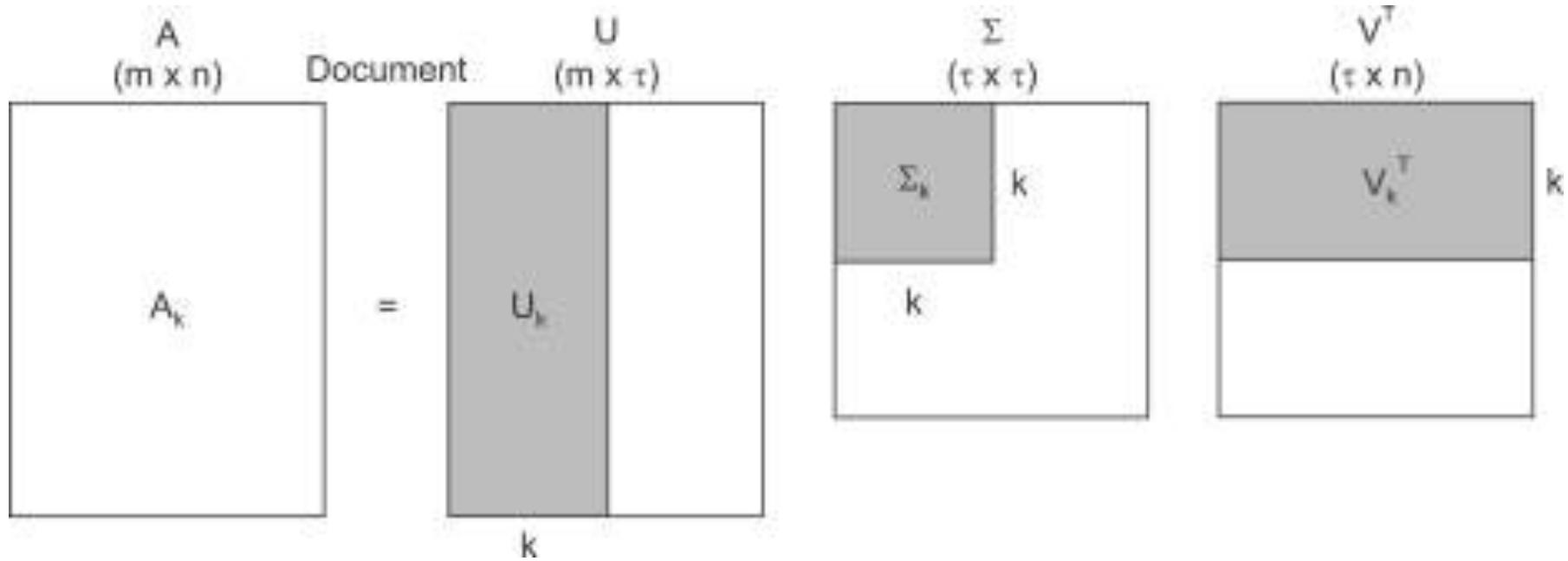
SVD

SVD = Singular Vector Decomposition (сингулярное разложение матриц)

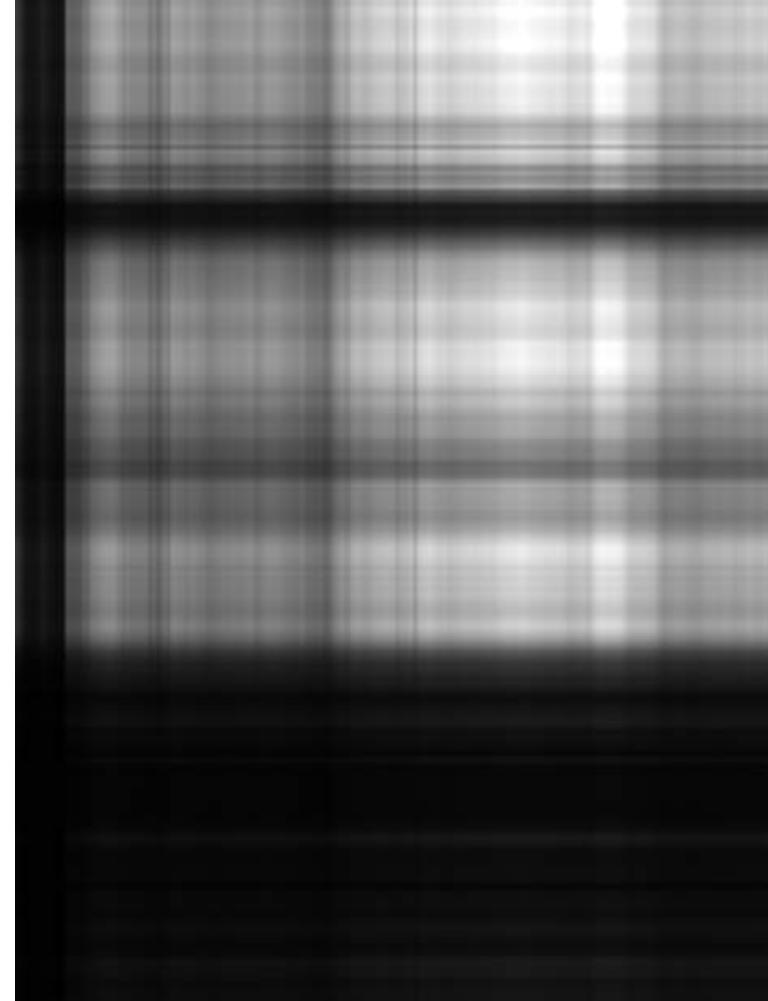
Позволяет получить наилучшее приближение исходной матрицы X матрицей X' ранга k.

Применяется для снижения размерности пространства признаков.

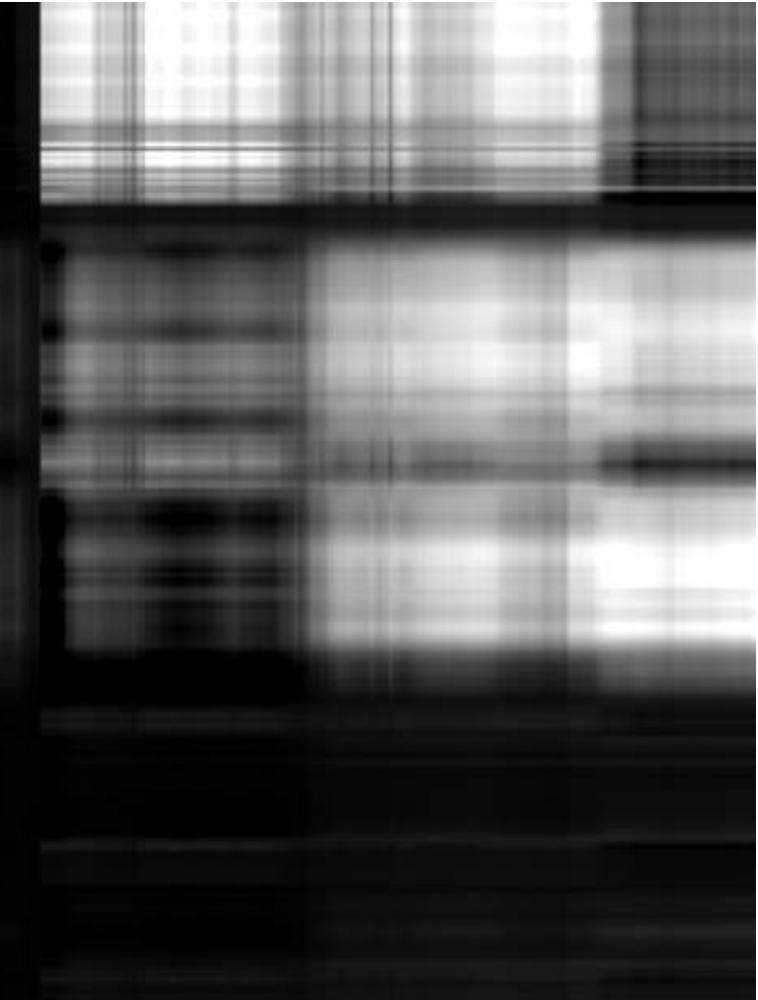
SVD



SVD: пример



SVD: пример

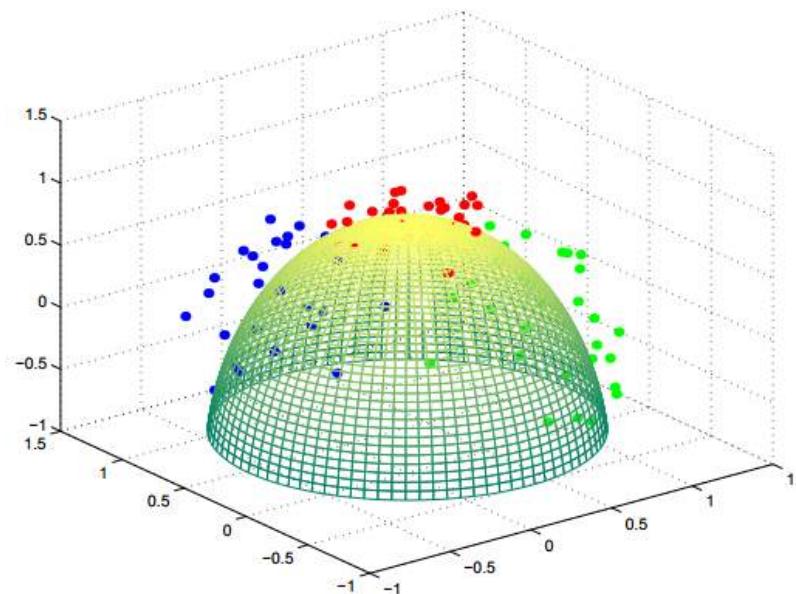


SVD: пример



А что, если линейных преобразований признаков мало?

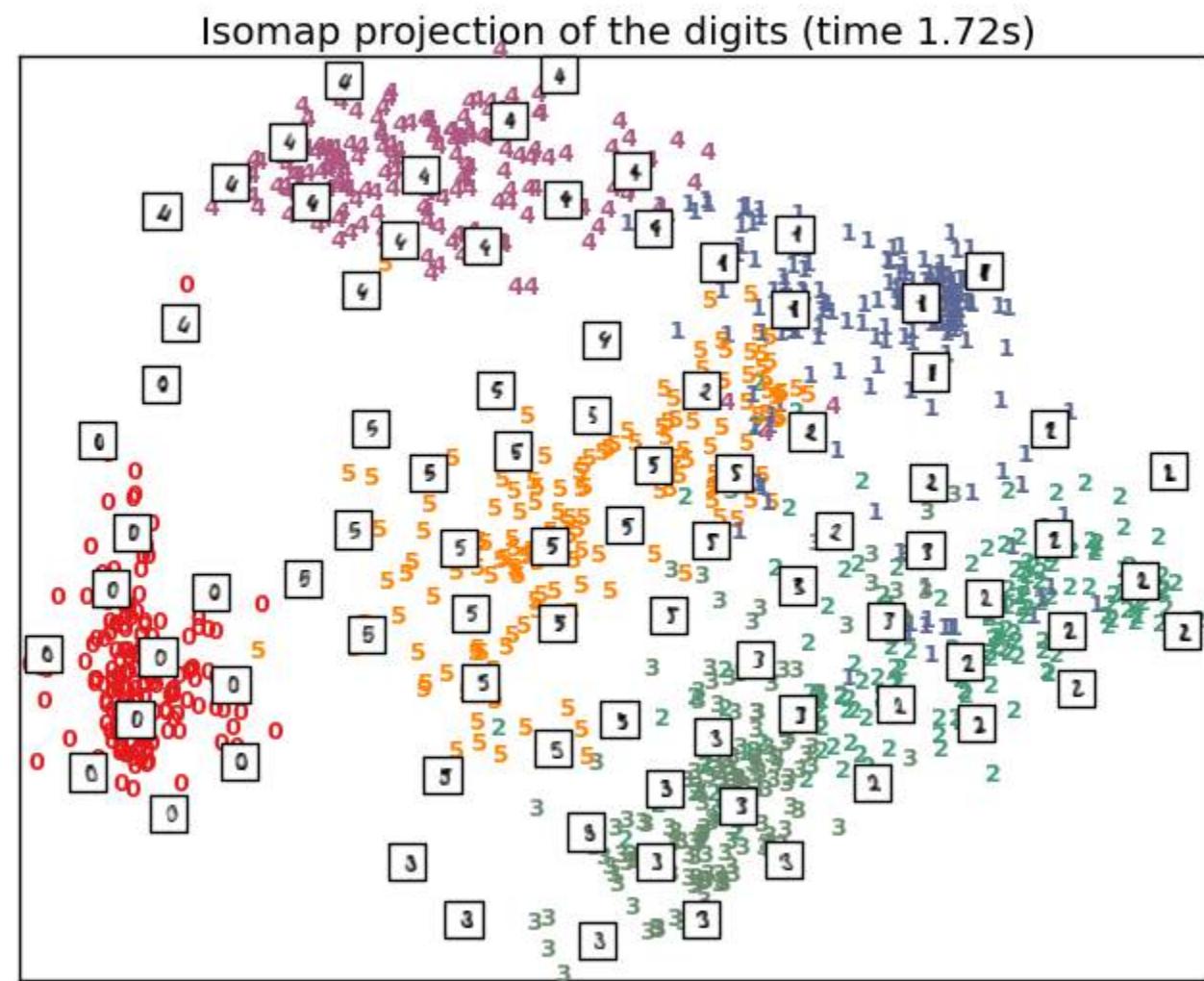
- Идея 1: объекты могут лежать в пространстве признаков на поверхности малой размерности.
- Идея 2: эта поверхность может быть нелинейной.



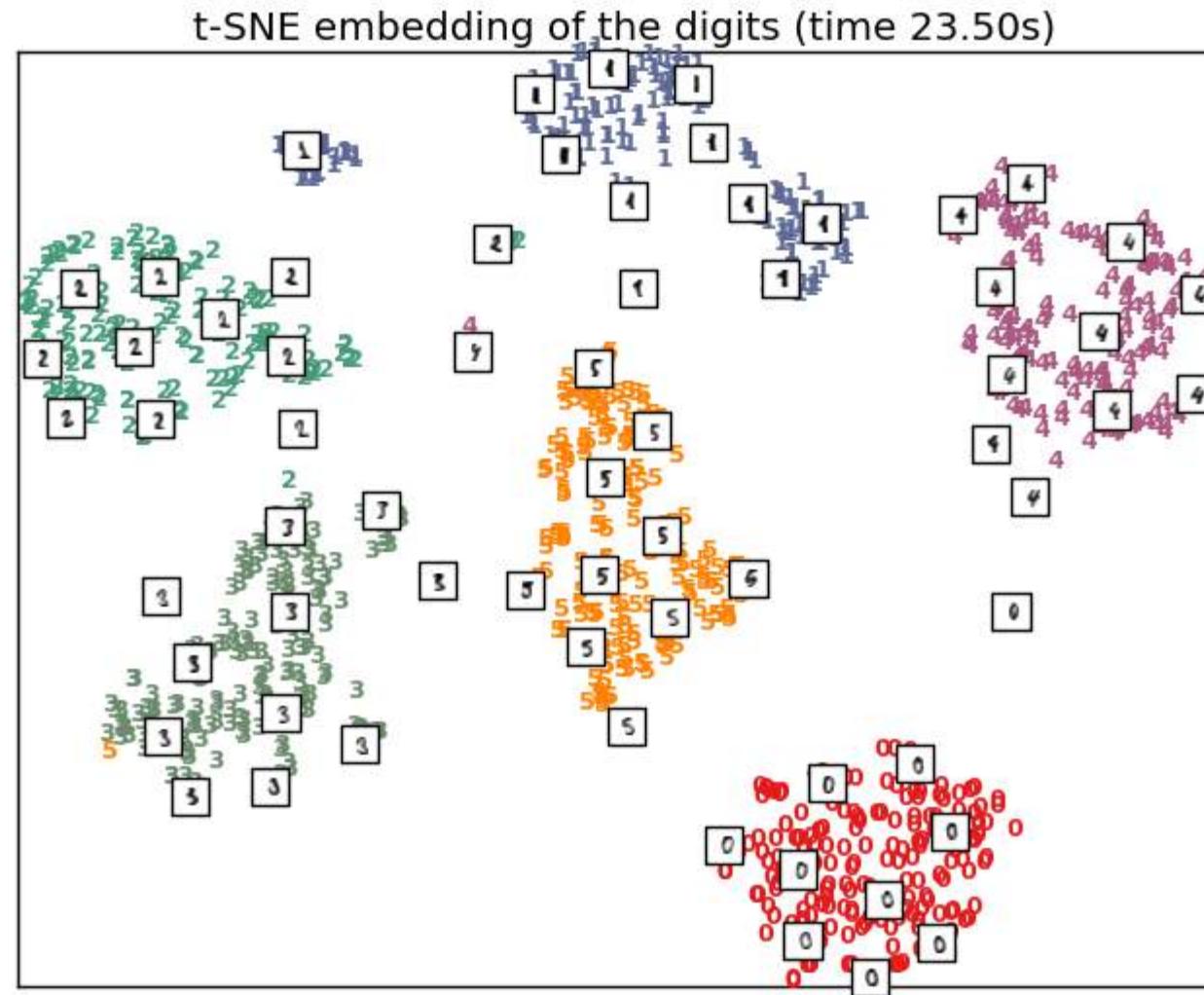
Нелинейное преобразование признаков

- SOM (Self-Organizing Maps) – самоорганизующиеся карты Кохонена. Не самый новый алгоритм, но идейно очень прост.
- Есть целое направление Manifold Learning

Manifold learning: Isomap



Manifold learning: t-SNE



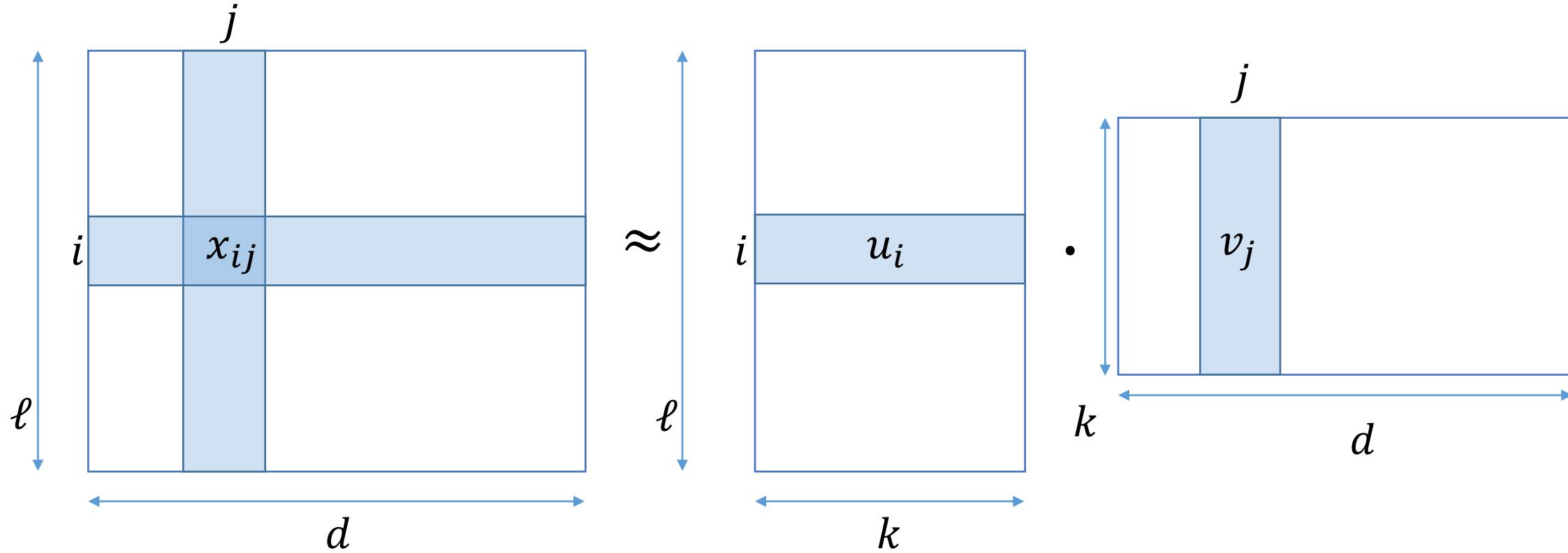
Резюме

1. Генерация признаков
2. Отбор признаков
3. Преобразование признаков:
 - Метод главных компонент и SVD,
 - Manifold learning

Другие задачи



Матричные разложения



$$x_{ij} \approx \langle u_i, v_j \rangle$$

Извлечение ассоциативных правил

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

market
basket
transactions

{Diapers, Beer}

Example of a frequent itemset

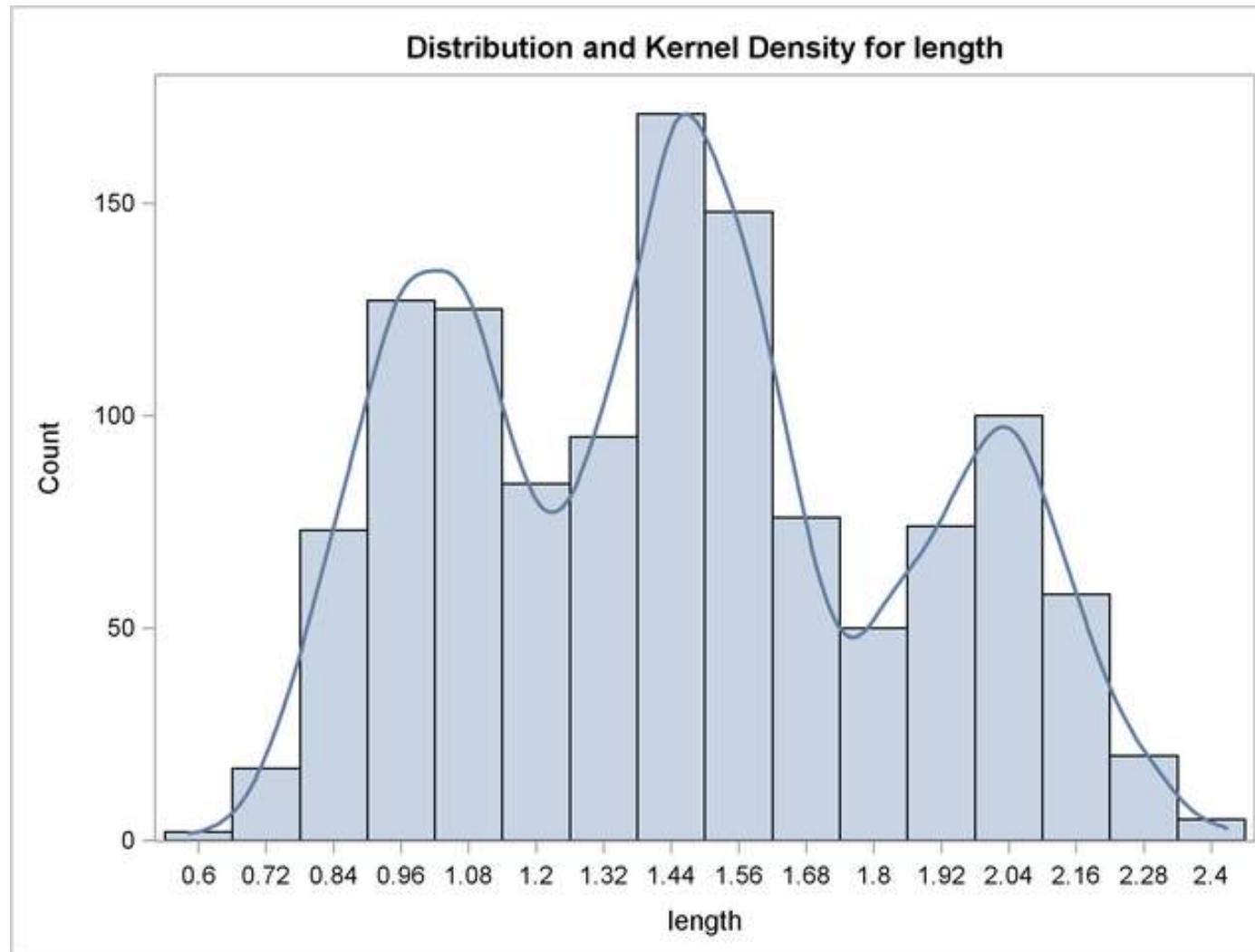
{Diapers} → {Beer}

Example of an association rule

Детектирование аномалий



Восстановление плотности



Спасибо за внимание



dmia@applieddatascience.ru



<https://t.me/joinchat/B10ITk74nRV56Dp1TDJGNA>



https://github.com/applied-data-science/Data_Mining_in_Action_2018_Spring

Анонс планов

1. Переформатирование DMIA
2. Запуск офлайн и онлайн курсов:
 - ML Intro (Supervised + Оценка качества + Основы постановки задач),
 - Advanced ML (Unsupervised + Доп. темы + кейсы),
 - ML для Senior & Lead DS (Постановка задач)

Благодарности

Курс читается при спонсорской поддержке компании:



**Райффайзен
БАНК**