

# Машинное обучение

Лекция 9. Методы кластеризации



# Содержание лекции

- I. Задача кластеризации
- II. Чем могут отличаться задачи кластеризации
- III. Kmeans
- IV. EM-алгоритм
- V. Иерархическая агglomerативная кластеризация
- VI. Простые графовые методы кластеризации
- VII. Density-based методы

## 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

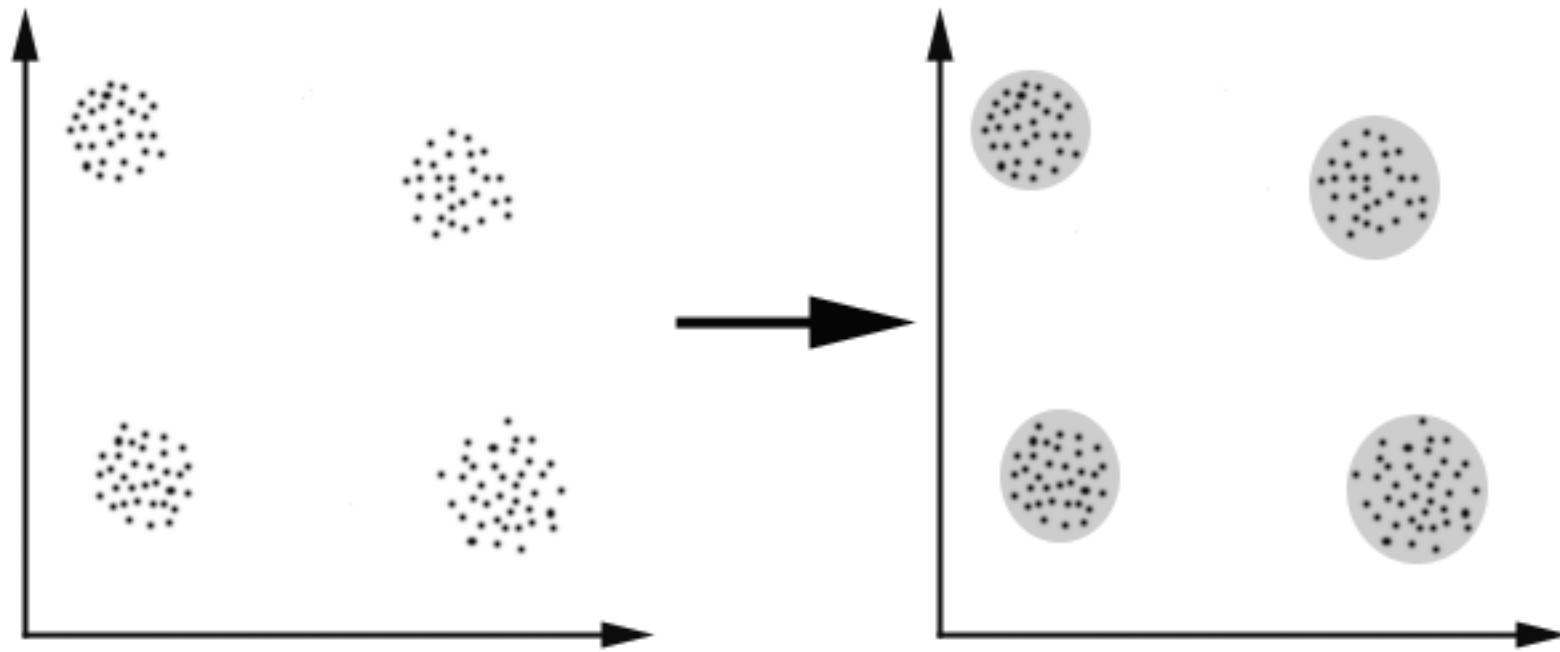
«Обучающая» выборка:

$x_1, \dots, x_l$  - 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 внутрикластерное расстояние

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

## 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

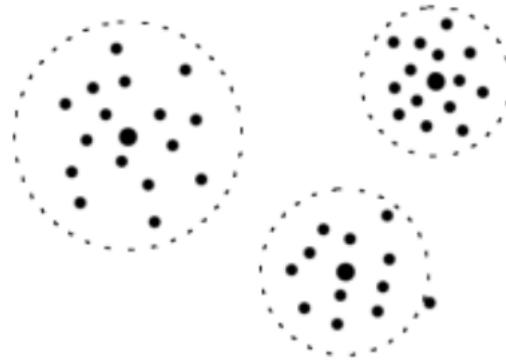
$$F_0 = \frac{\sum_{i < j} [y_i = y_j] \rho(x_i, x_j)}{\sum_{i < j} [y_i = y_j]}$$

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

$$F_0/F_1 \rightarrow \min$$

## II. Разнообразие задач кластеризации

# Форма кластеров



# Форма кластеров



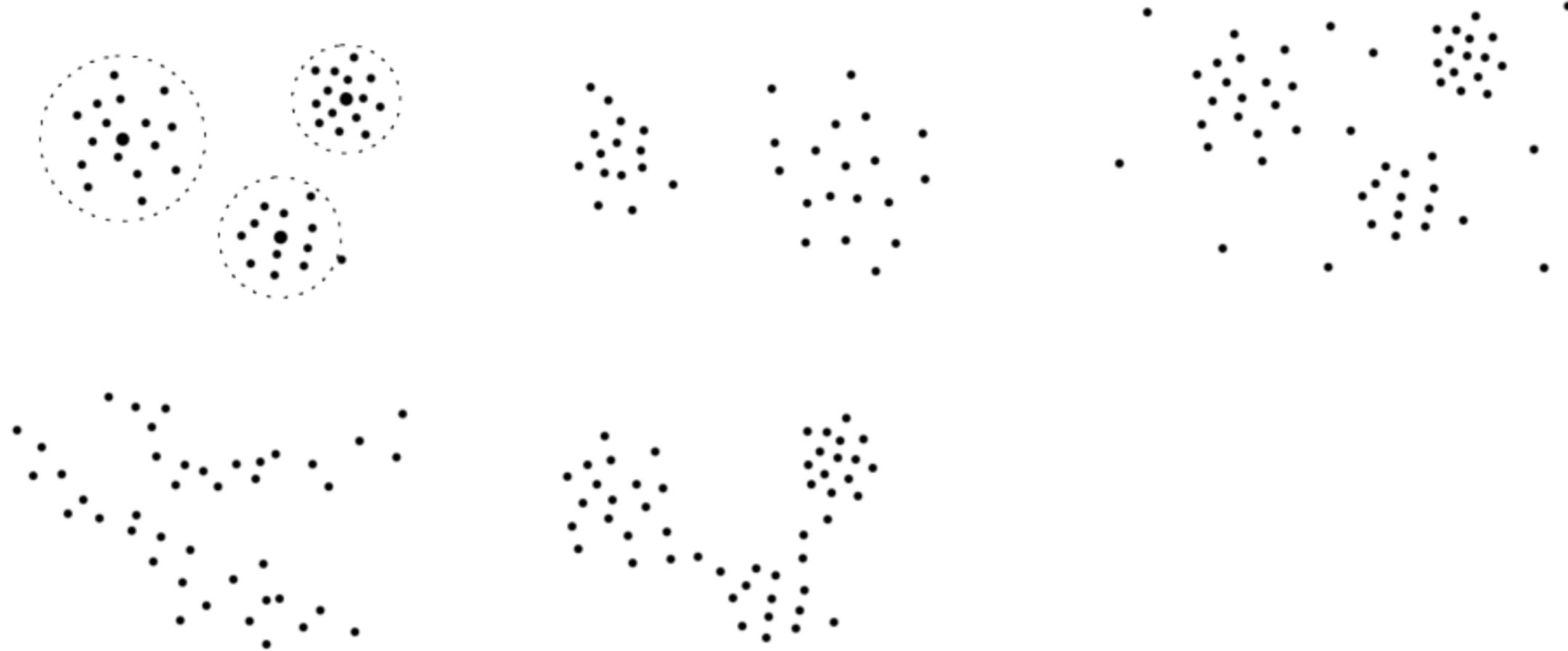
# Форма кластеров



# Форма кластеров



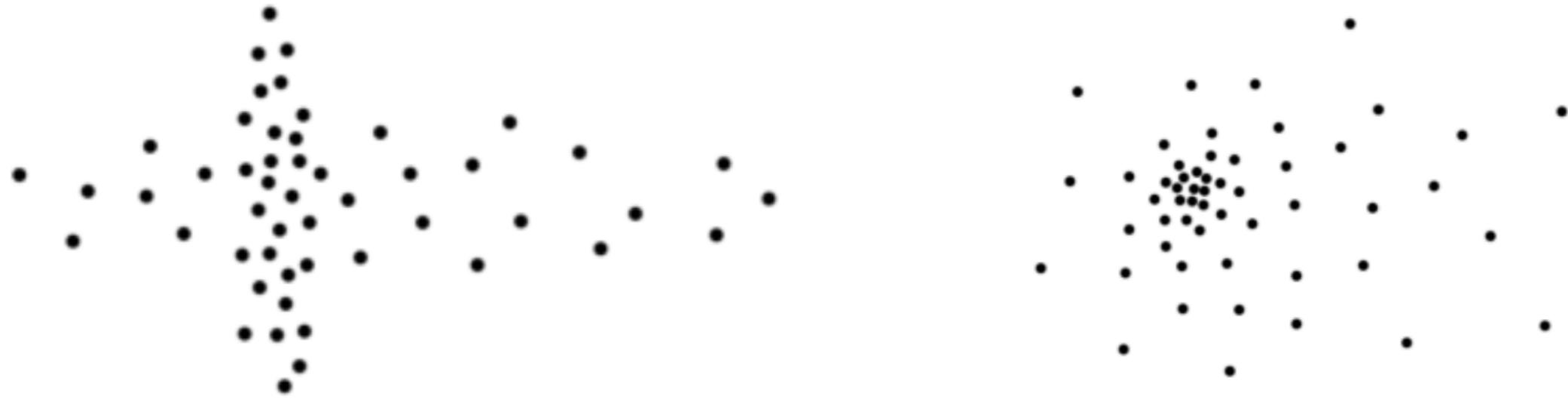
# Форма кластеров



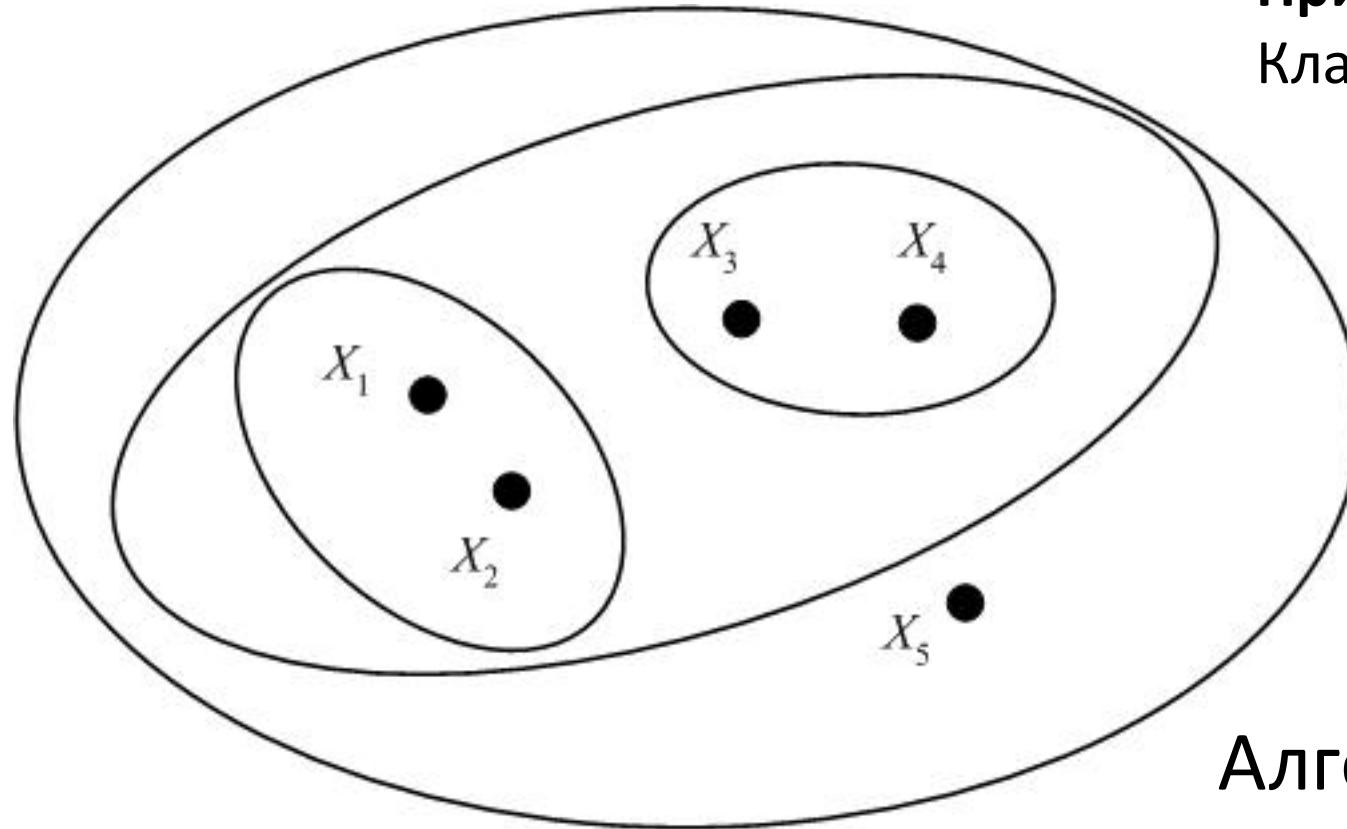
# Форма кластеров



# Форма кластеров

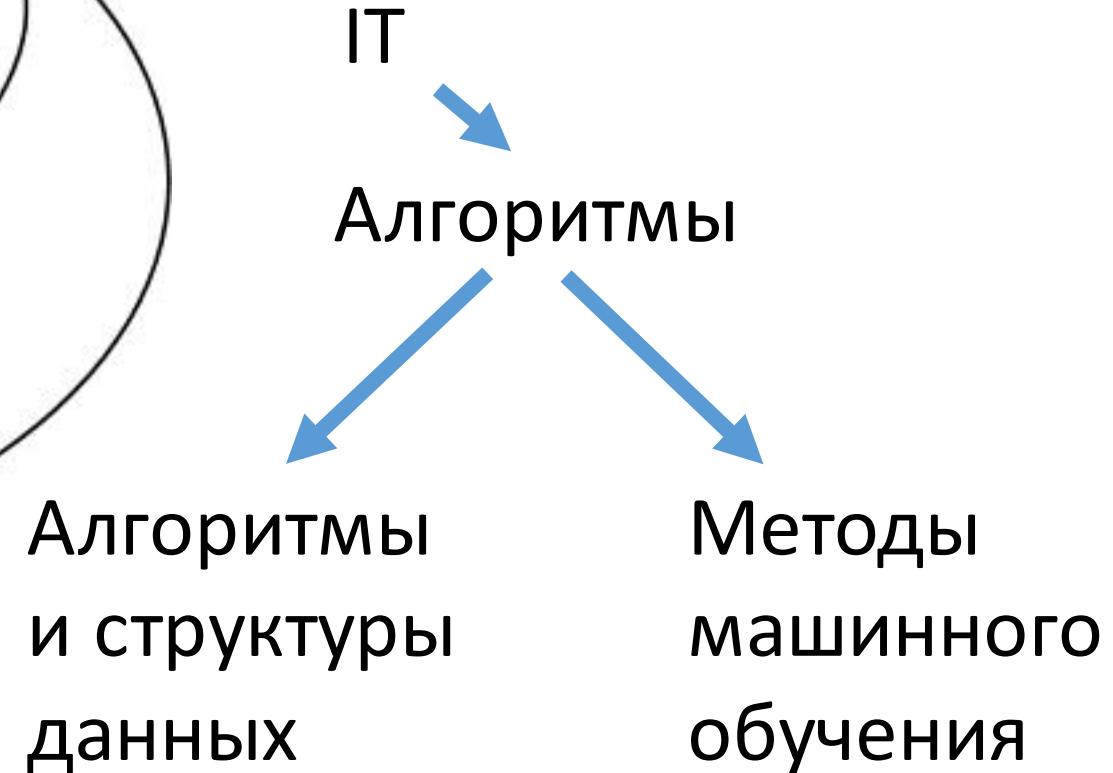


# Вложенность кластеров



**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

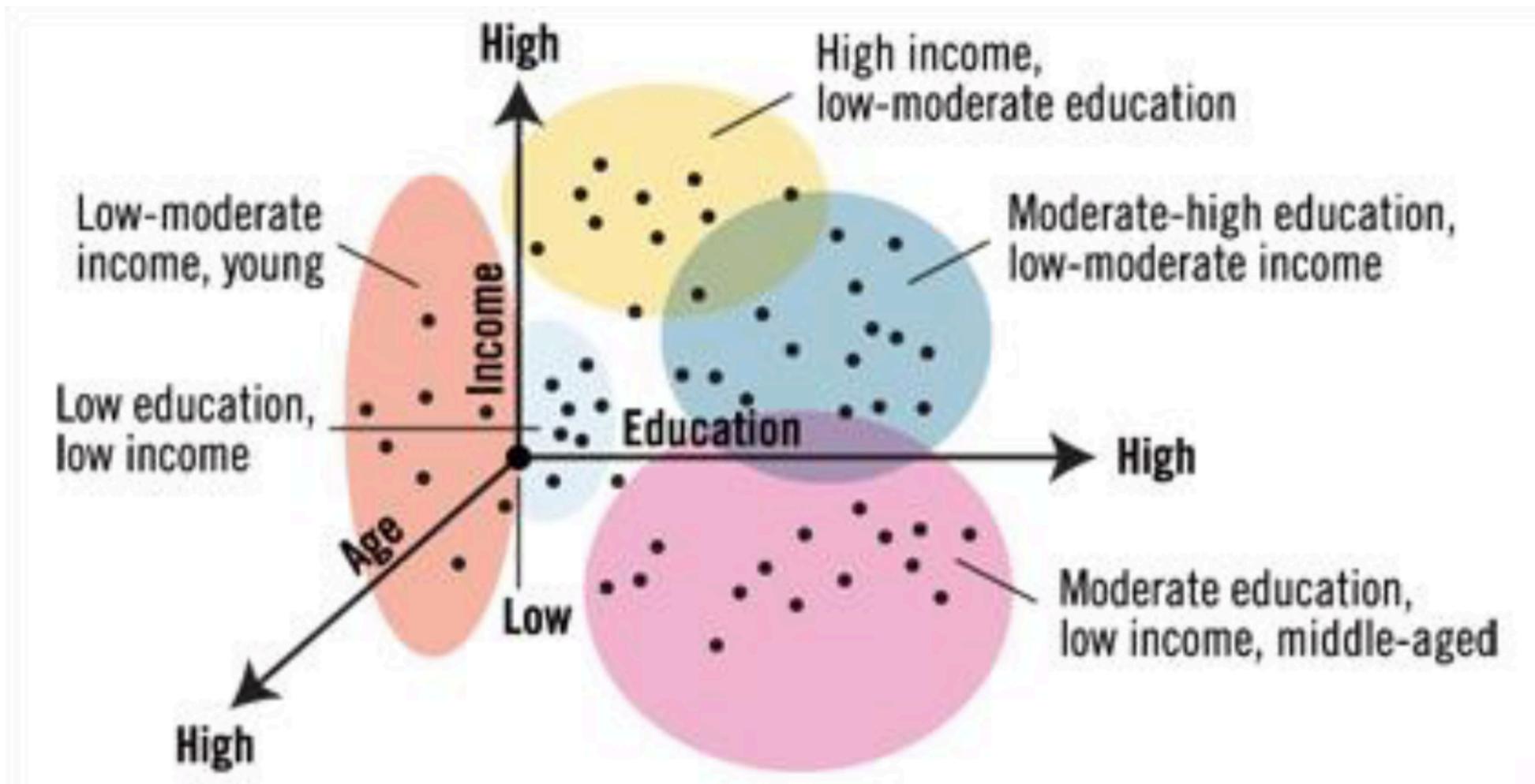
Открытие Олимпиады в Сочи  
посмотрели несколько миллиардов  
человек

**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

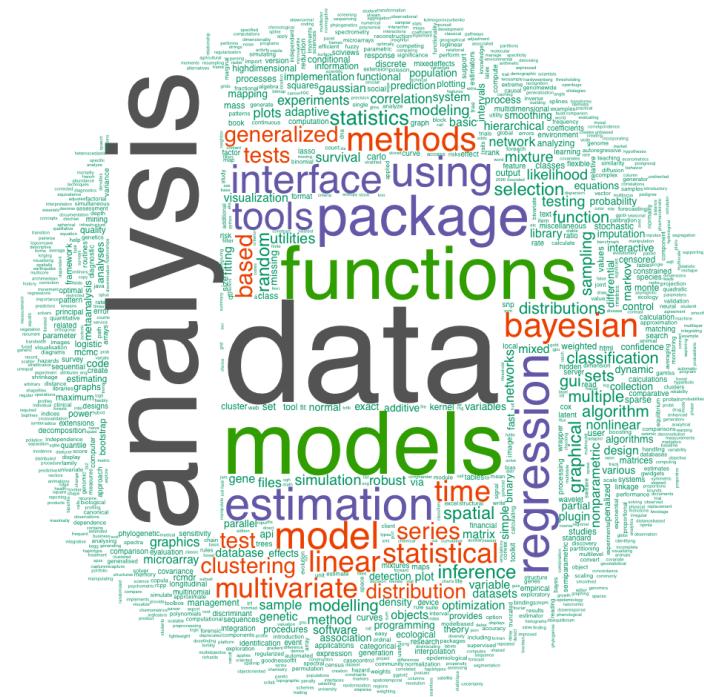
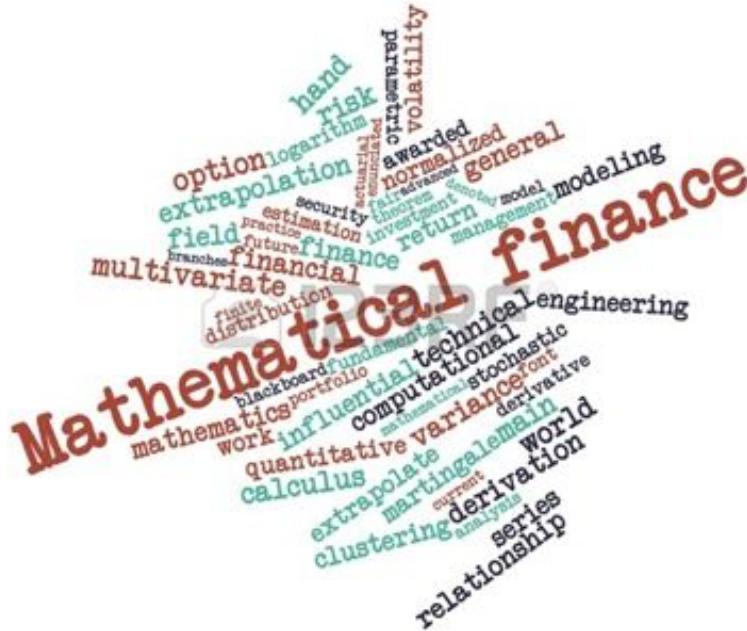
5

5

5

# «Жесткая» и «мягкая» кластеризации

# Кластеризация для выделения «тем»

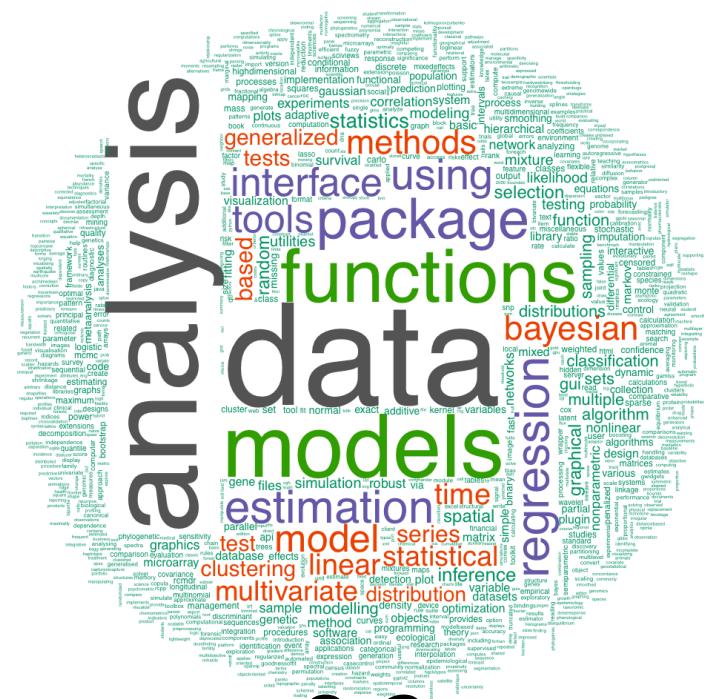


# «Жесткая» и «мягкая» кластеризации

## Кластеризация для выделения «тем»



0.2



0.3

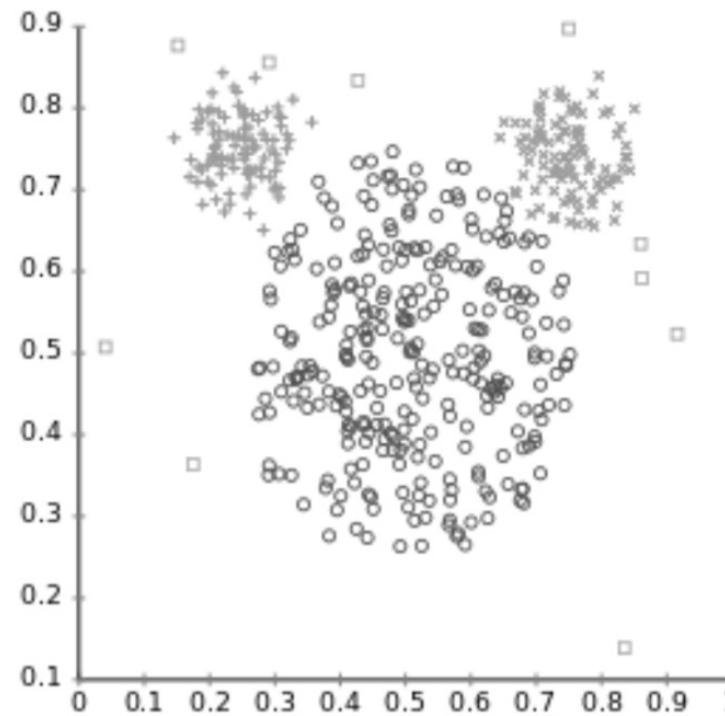


0.5

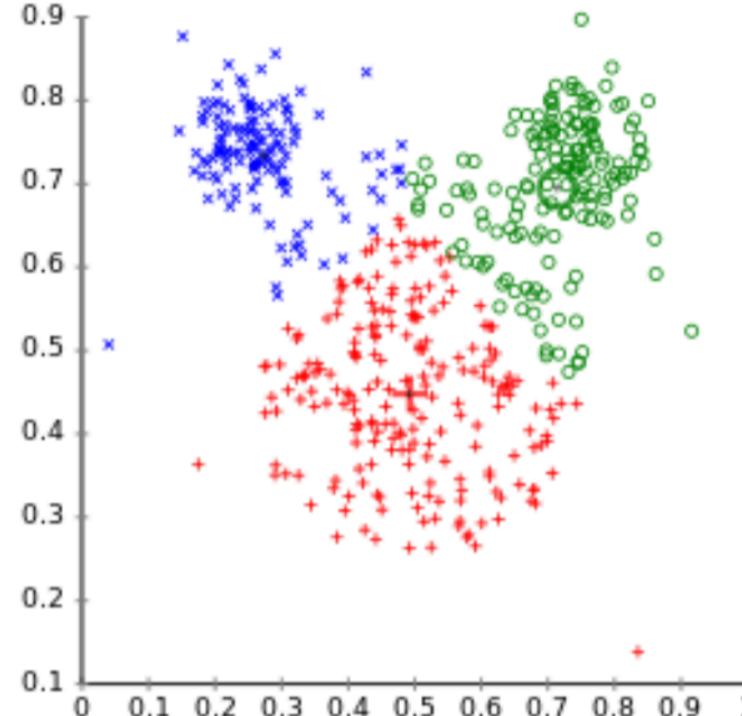
# Резюме: чем могут отличаться задачи кластеризации

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

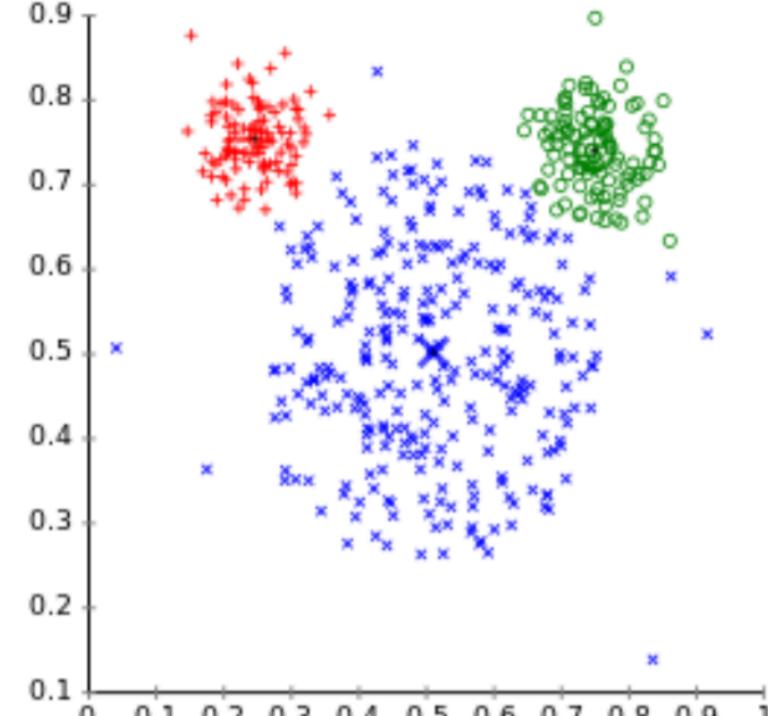
# Различия в результатах работы методов



Исходная выборка  
("Mouse" dataset)



Метод k средних  
(K-Means)



ЕМ-алгоритм

### **III. Метод К средних (K-Means)**

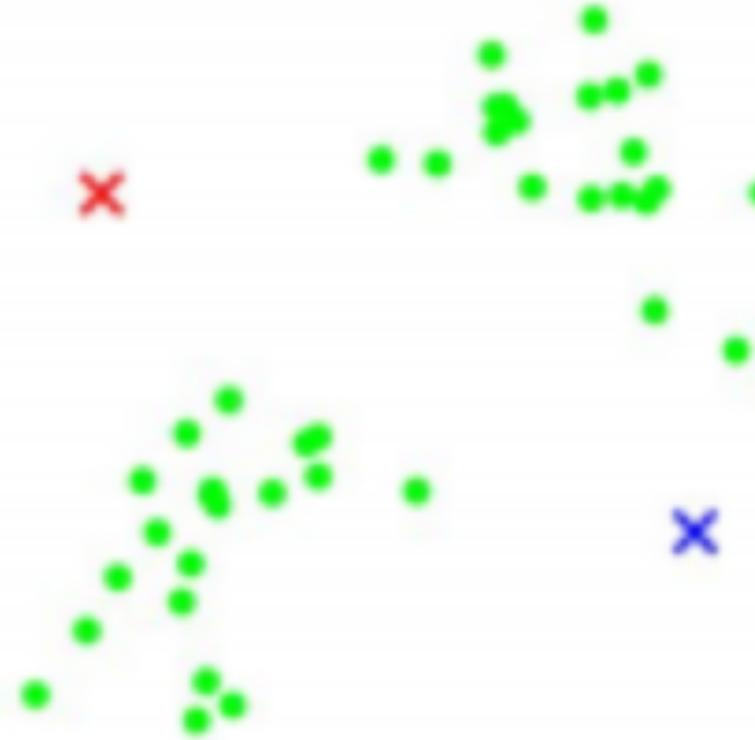
# План

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

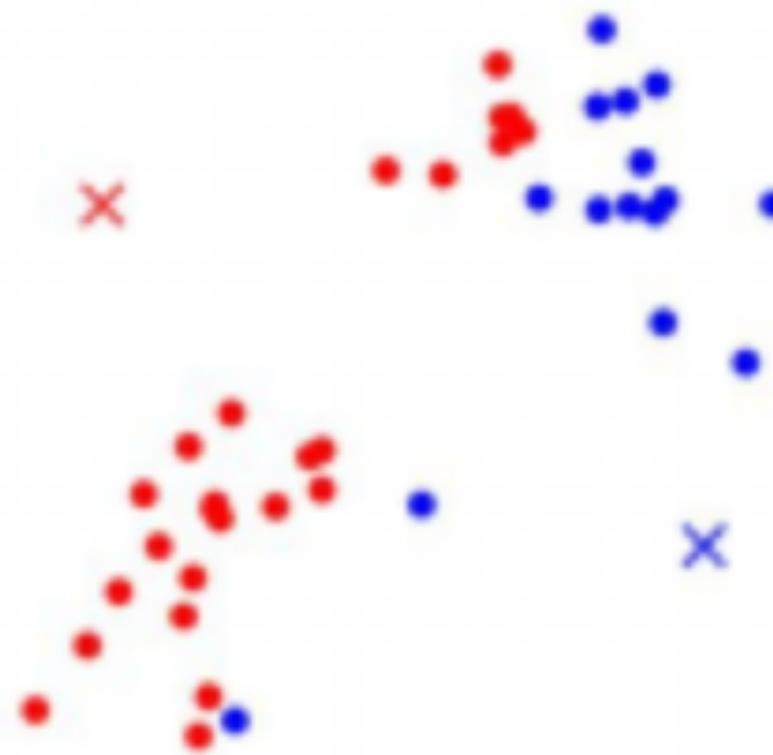
# Как работает K-Means



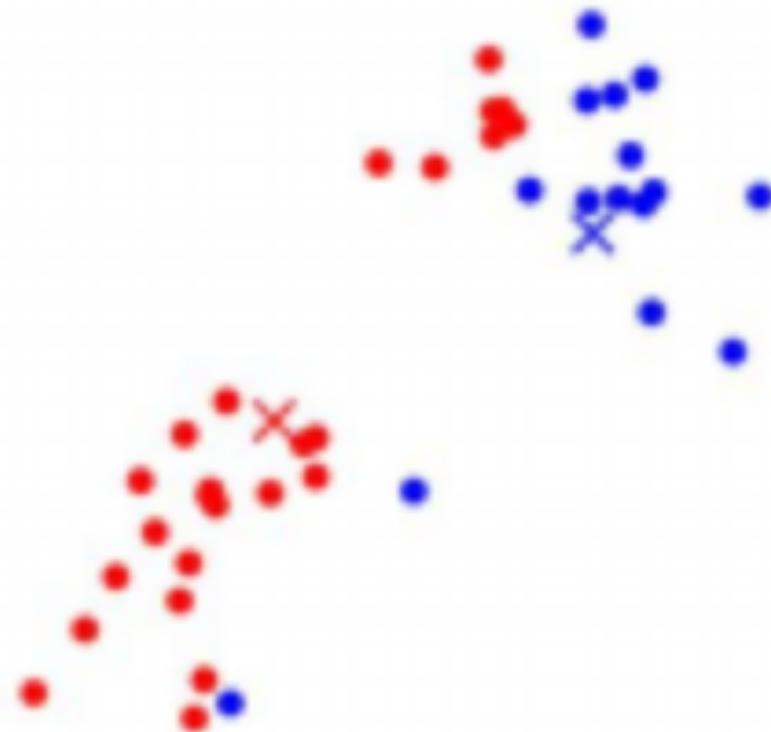
# Как работает K-Means



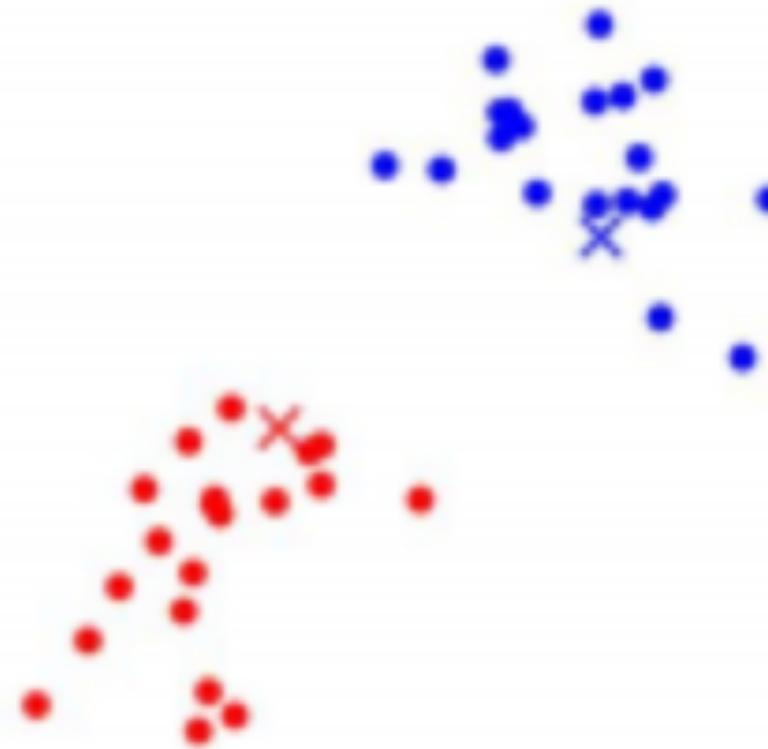
# Как работает K-Means



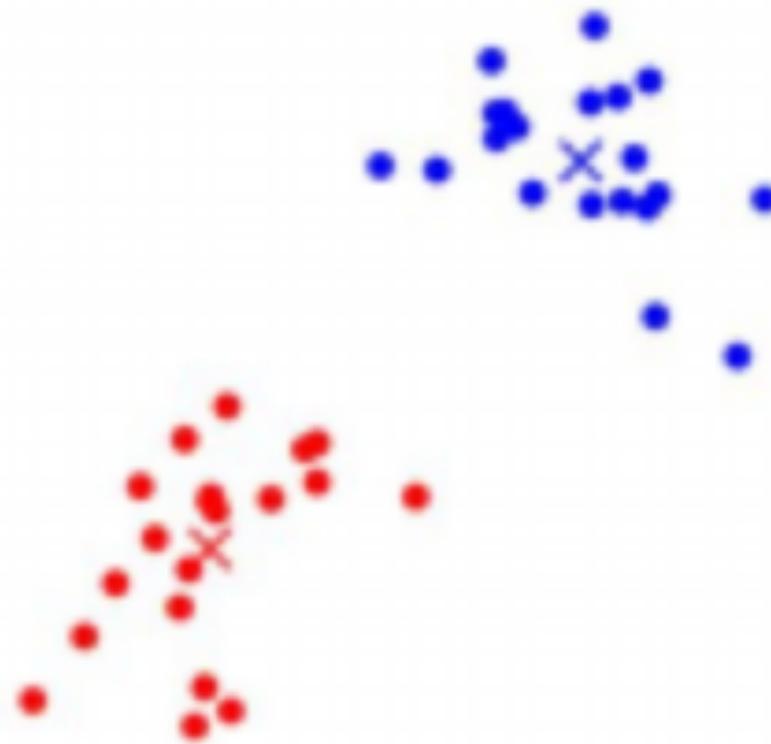
# Как работает K-Means



# Как работает K-Means



# Как работает K-Means



## Вариации K-Means

- В версии Болла Холла: уже рассказанный метод
- В версии Мак Кина: каждый раз, когда объект переходит из одного кластера в другой – центры кластеров пересчитываются

# 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)

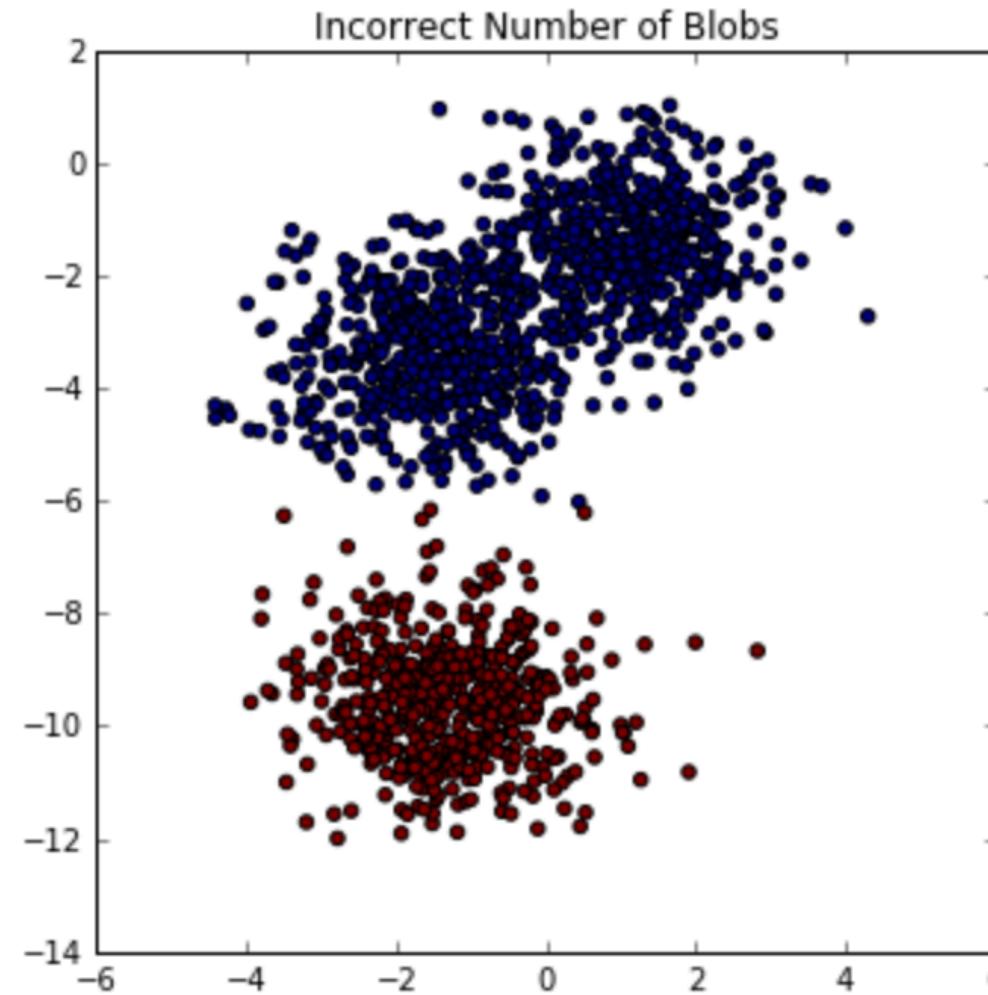


# Пример: квантизация изображений

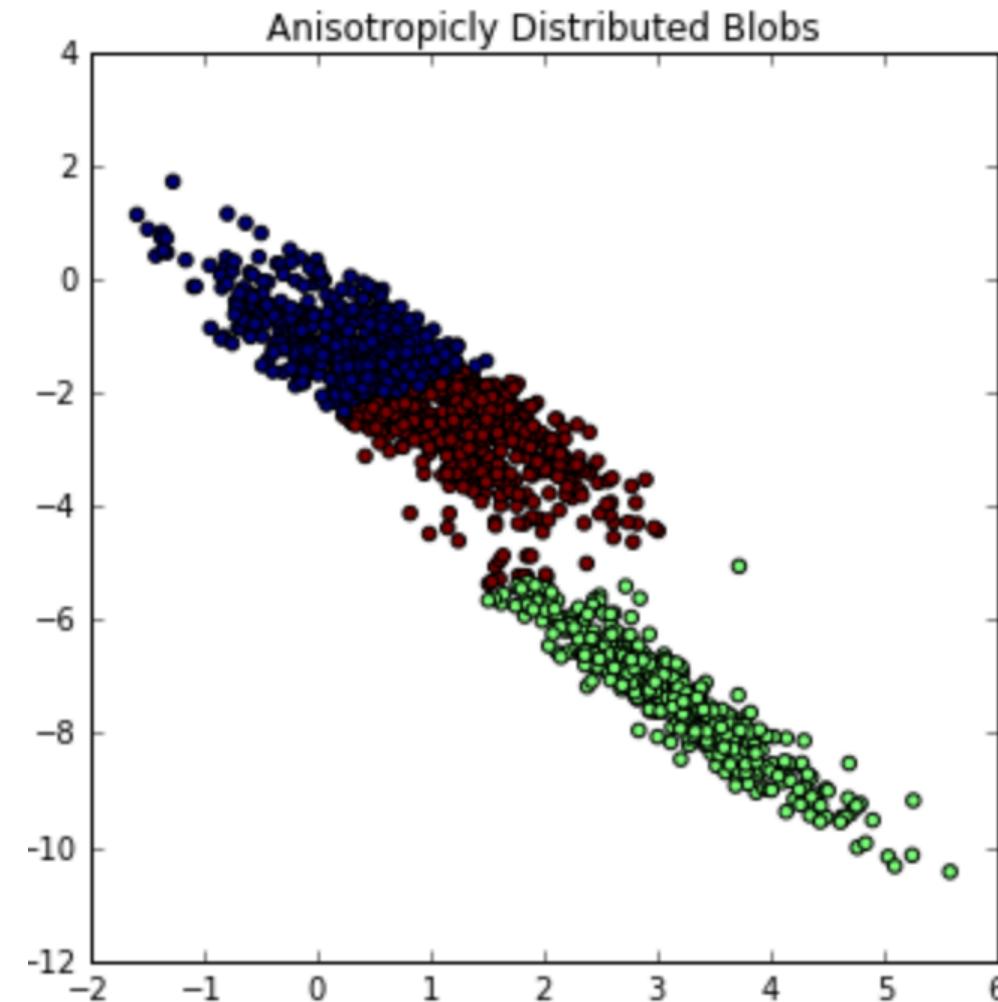
Quantized image (64 colors, K-Means)



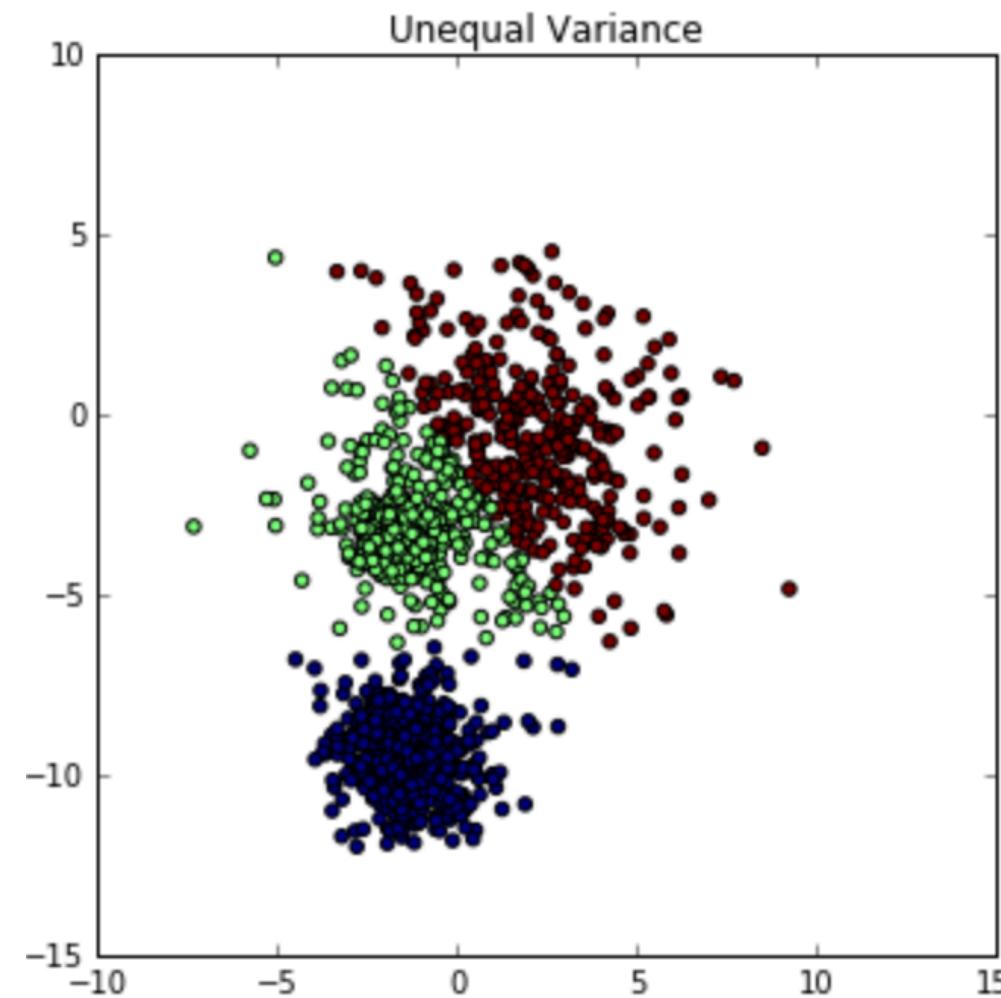
# KMeans и разные формы кластеров



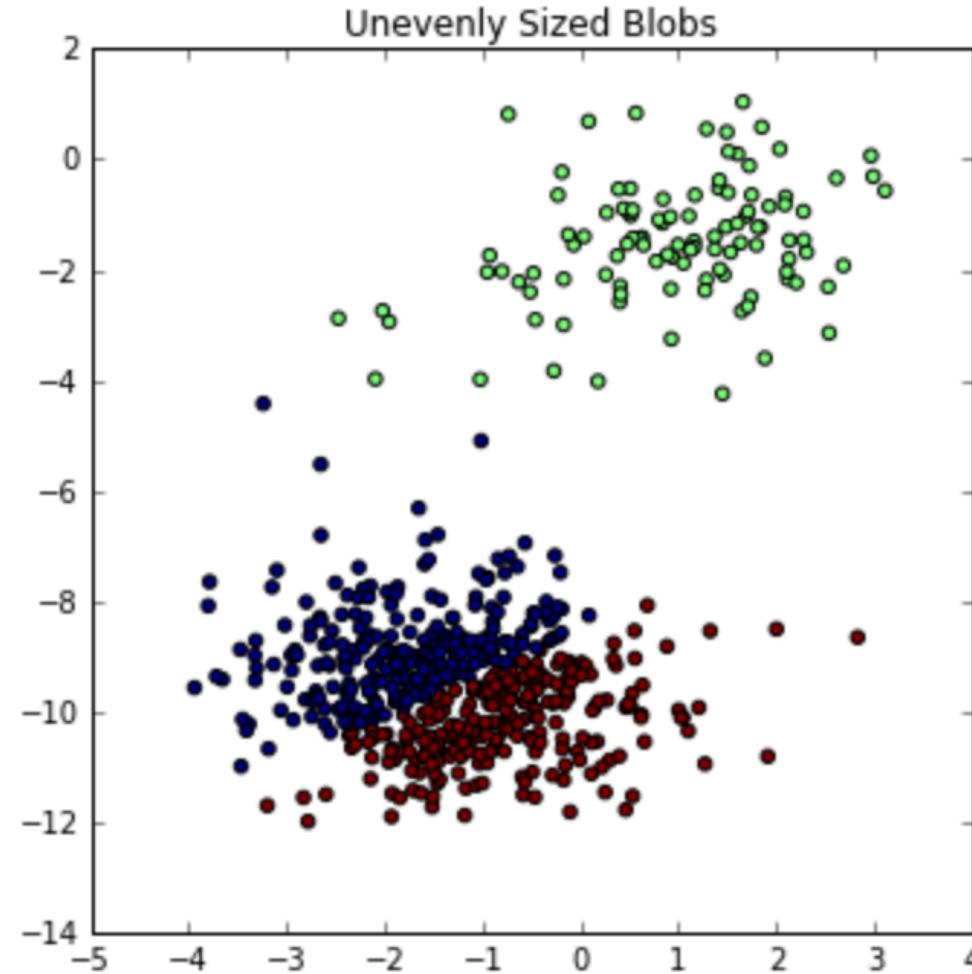
# KMeans и разные формы кластеров



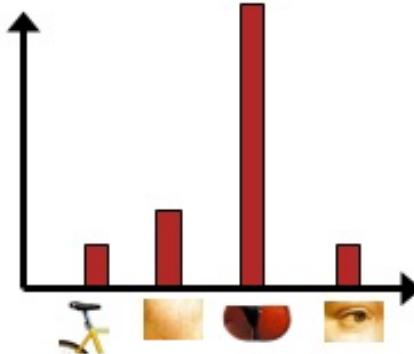
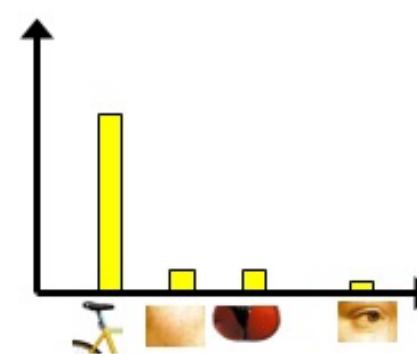
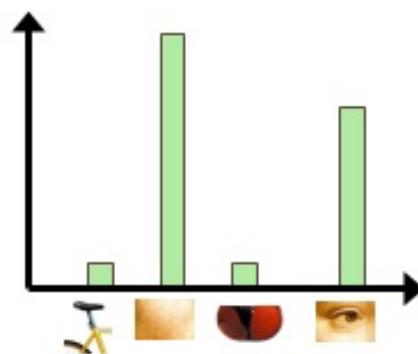
# KMeans и разные формы кластеров



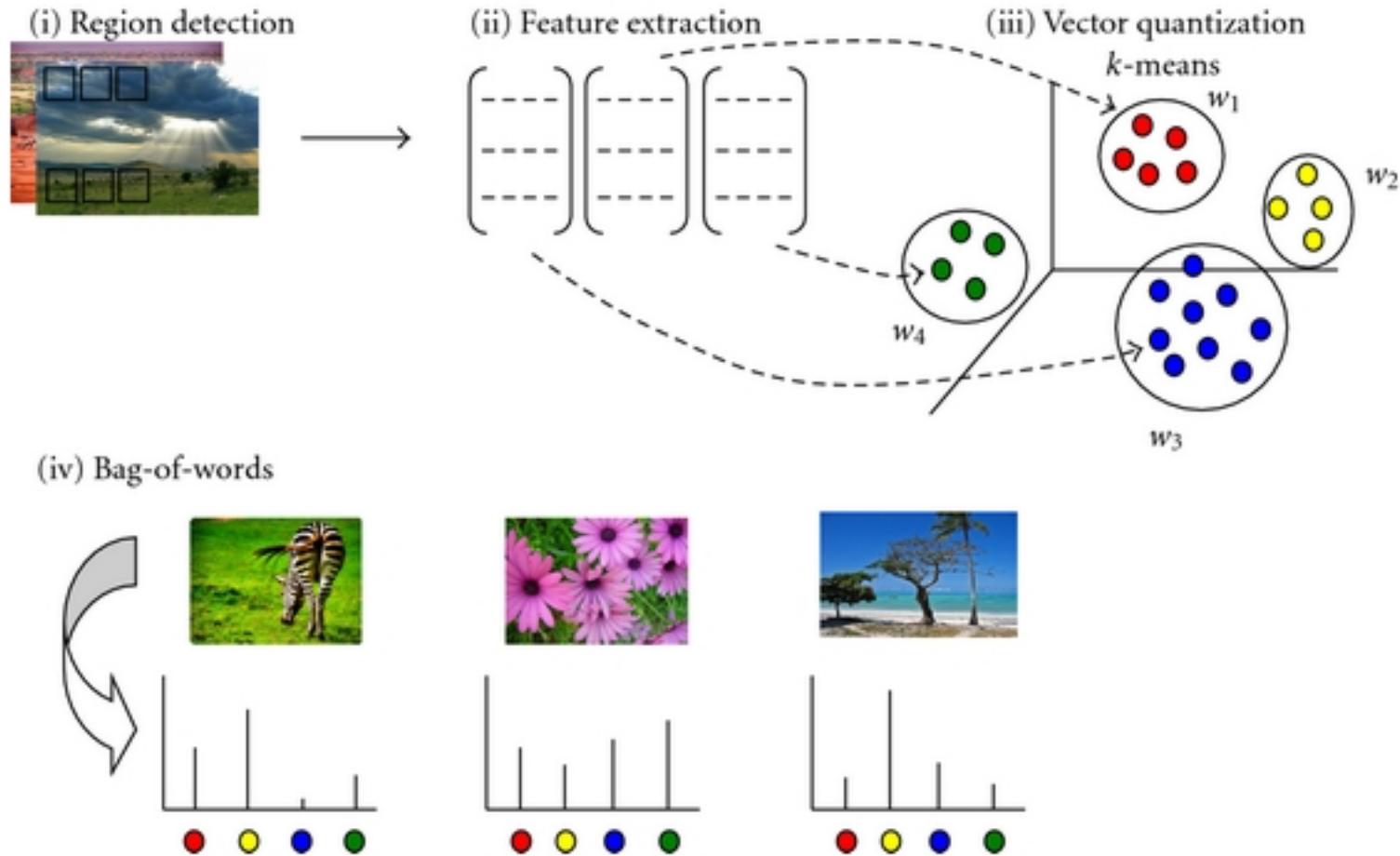
# KMeans и разные формы кластеров



# 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

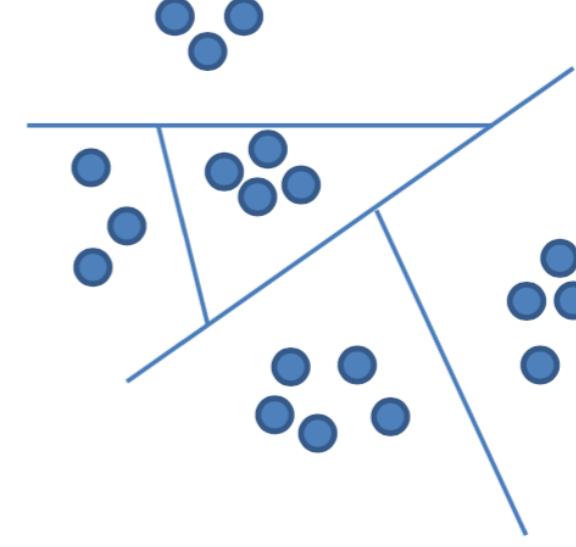
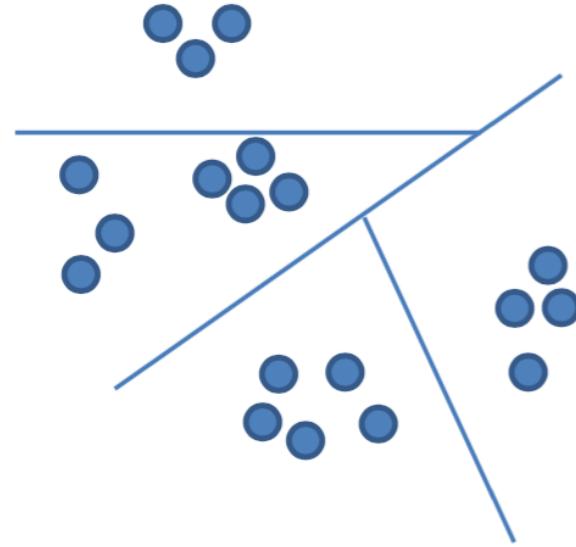
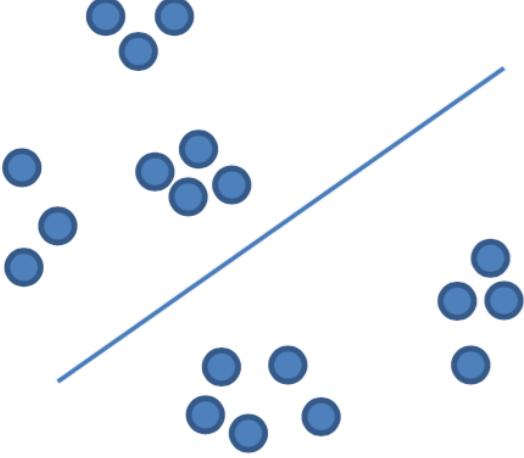
K Means итеративно минимизирует среднее внутрикластерное расстояние:

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

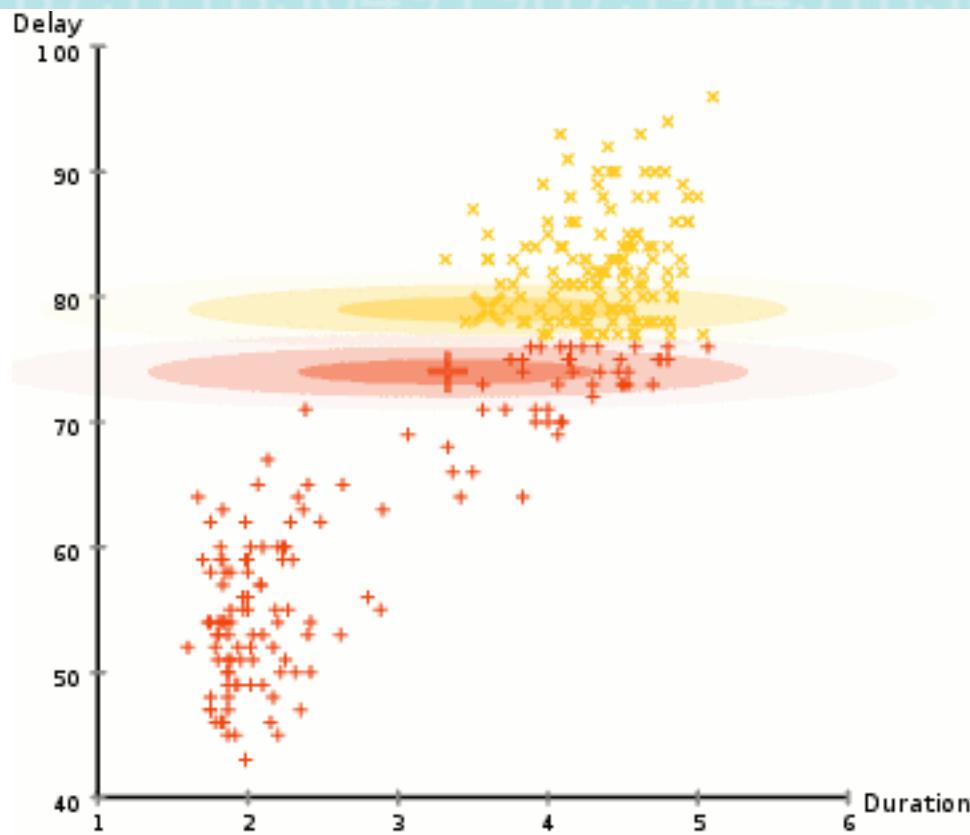
1. Как работает K Means
2. Вариации: K Means Болла-Холла и Мак Кина
3. Что делать, когда данных много: Mini Batch K-Means
4. Что делать, когда много признаков: понижение размерности
5. Выбор начальных приближений: Kmeans++
6. Пример: квантизация изображений
7. Работа K means с разными формами кластеров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

### III. Expectation-Maximization (ЕМ-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



## 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

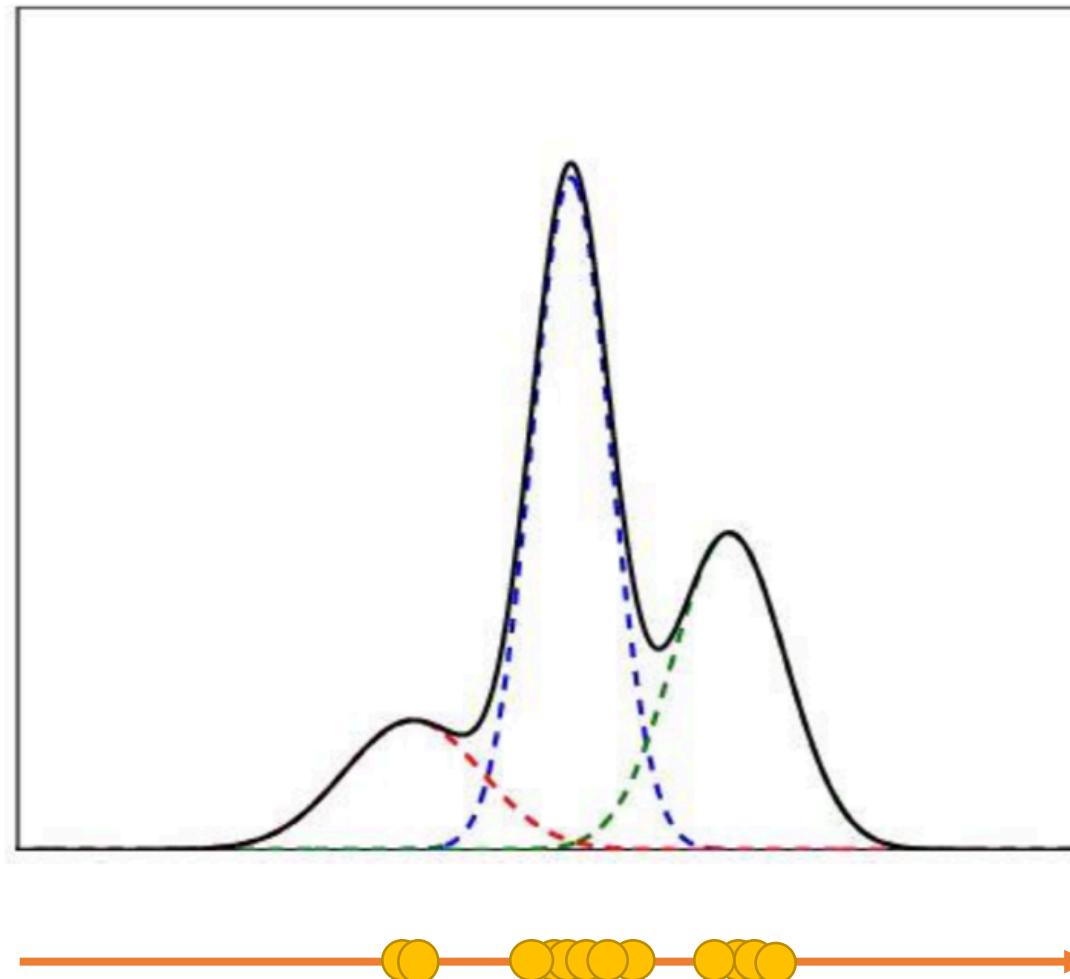
## Постановка задачи: разделение смеси

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

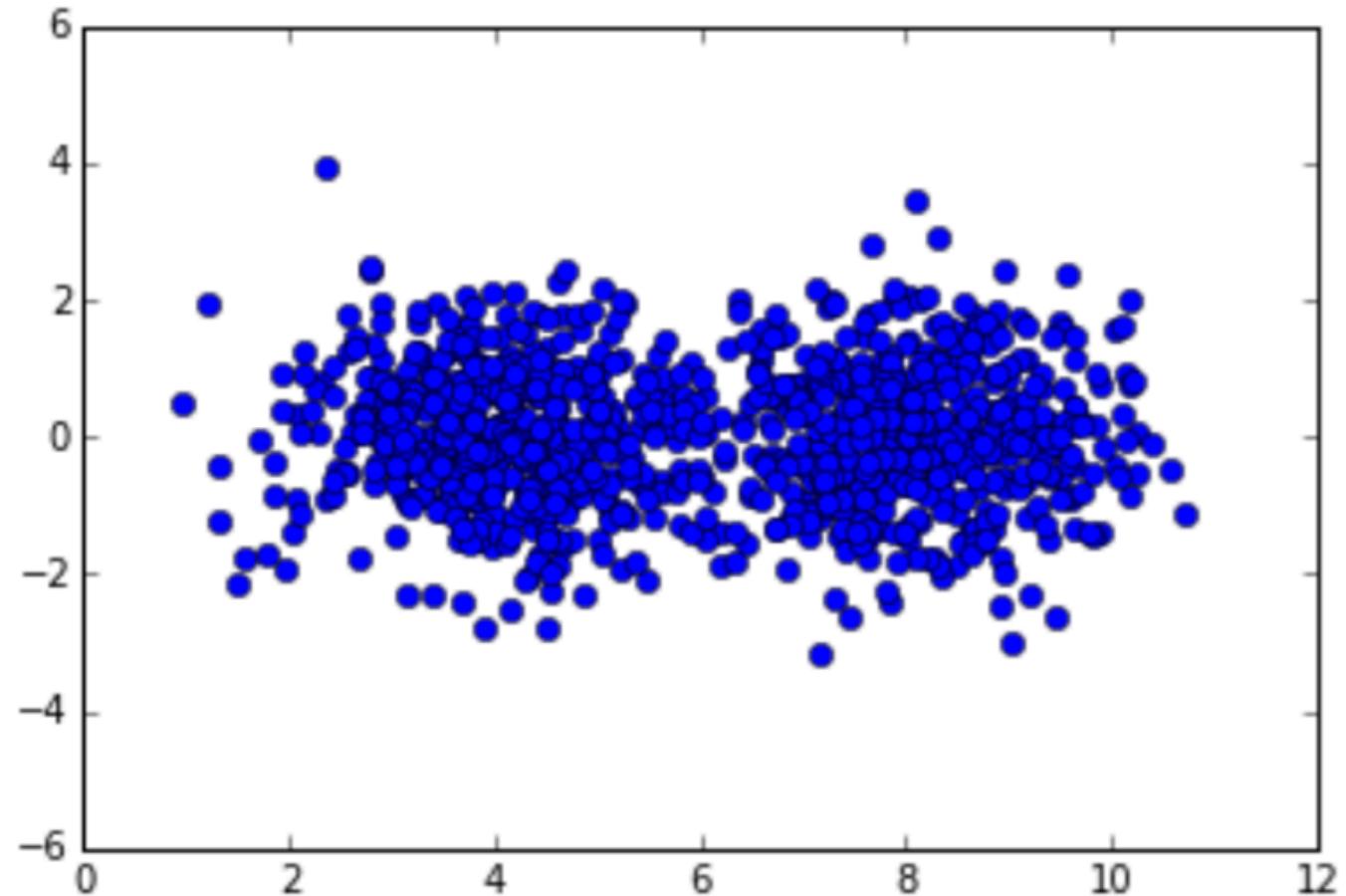
$$p_j(x) = \varphi(\theta_j; x)$$

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

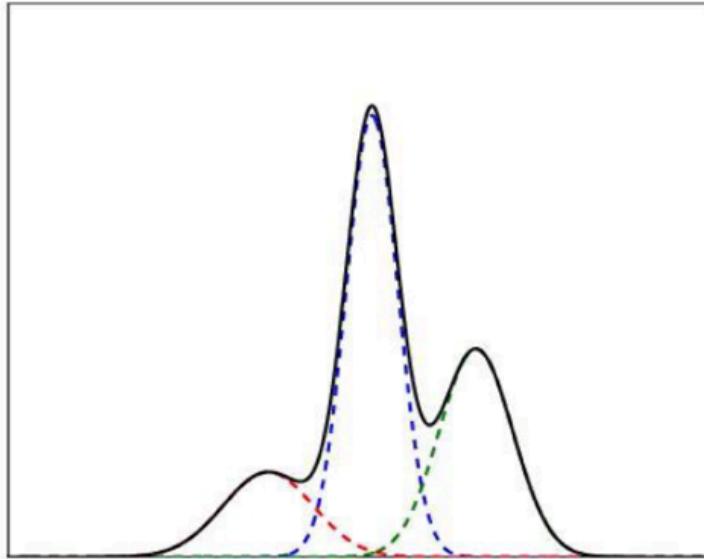
Как выглядит смесь распределений



# Как выглядит смесь распределений



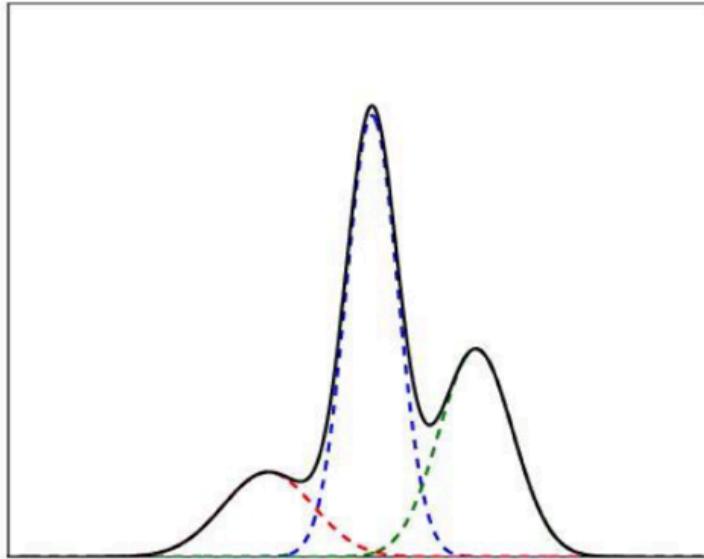
# Почему не решить задачу «в лоб»



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

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

# 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

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

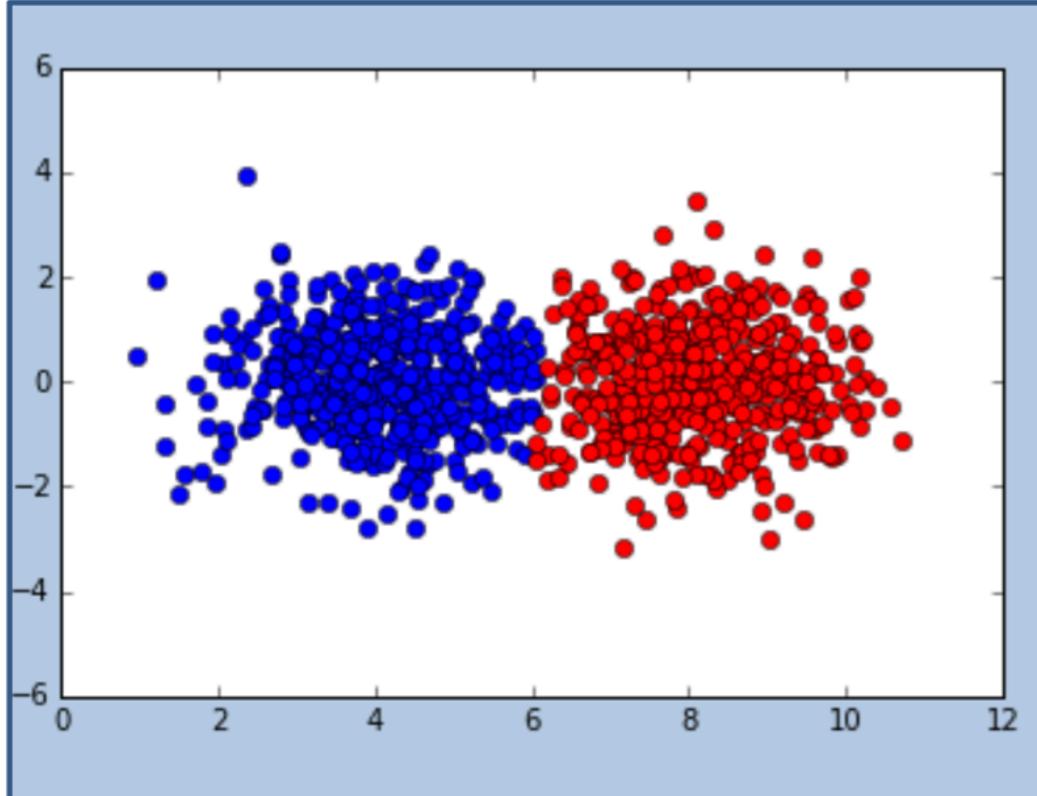
E-шаг:

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

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:

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

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

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

# 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:

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

# Классическое объяснение EM-алгоритма

$$L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}),$$

$$L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X} | \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})$$

**E-шаг:** 
$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}^{(t)}} [\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]$$

**M-шаг:** 
$$\boldsymbol{\theta}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)})$$

# 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. Классическое объяснение метода
8. Для чего еще используют алгоритм

## V. Агломеративная иерархическая кластеризация

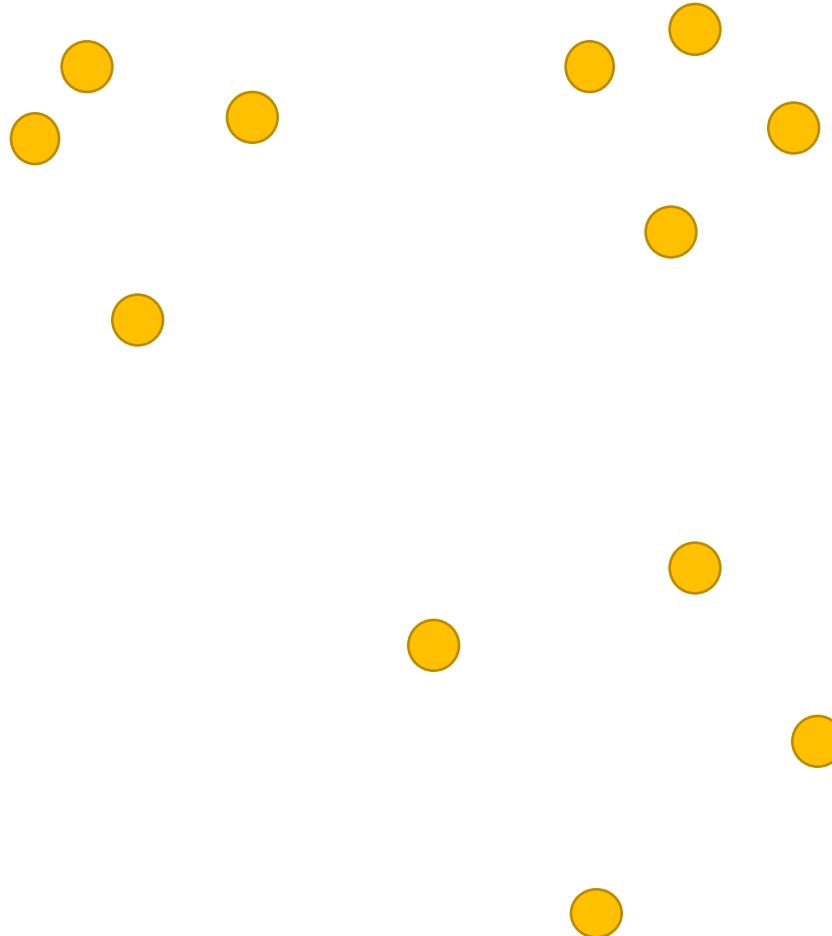
# План

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

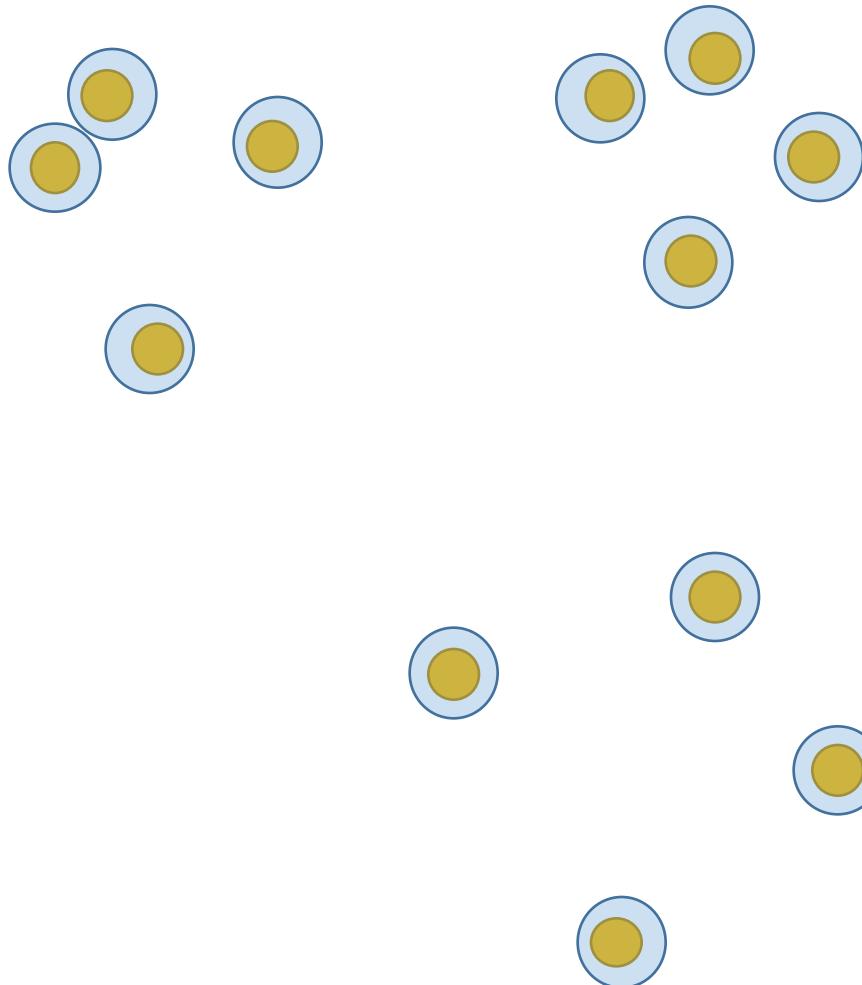
# Иерархическая кластеризация

- Агломеративная
- Дивизионная или дивизимная

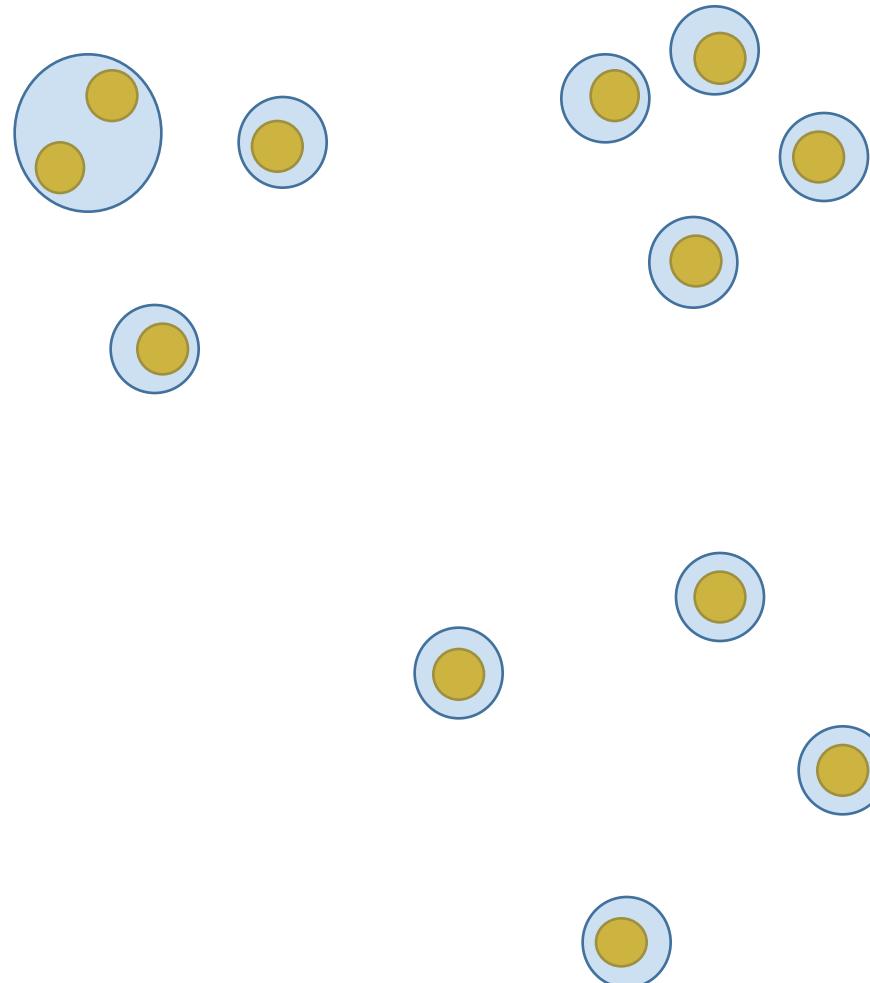
# Агломеративная кластеризация



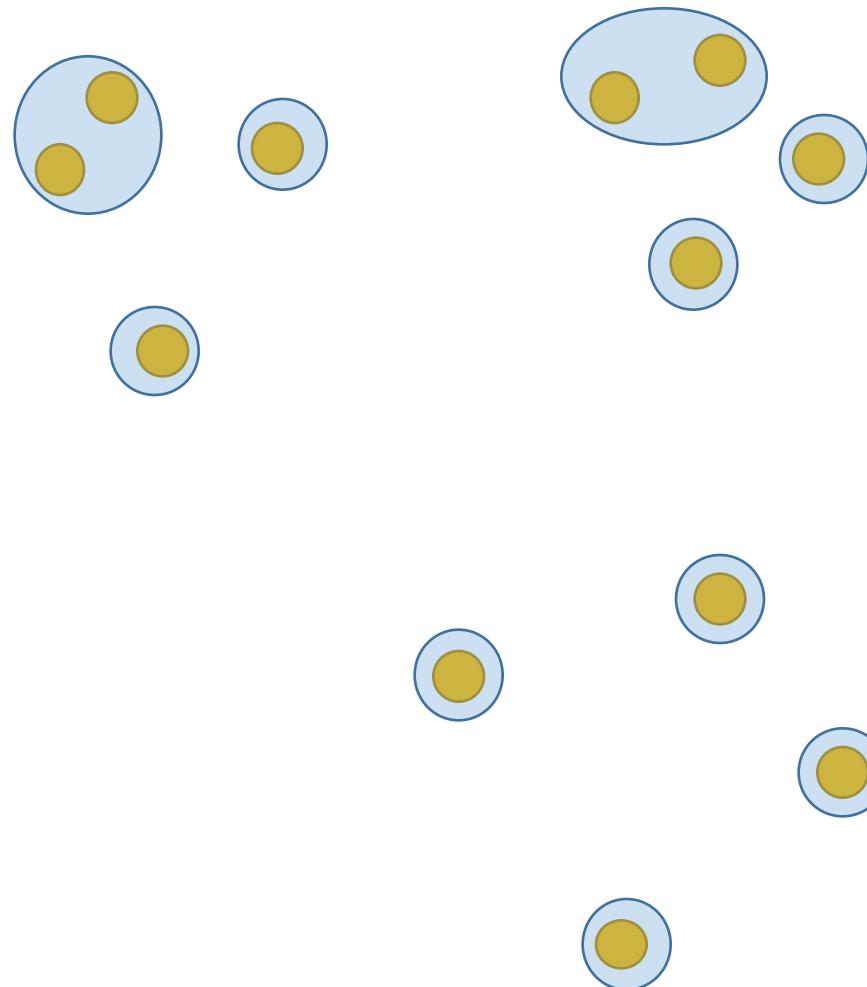
# Агломеративная кластеризация



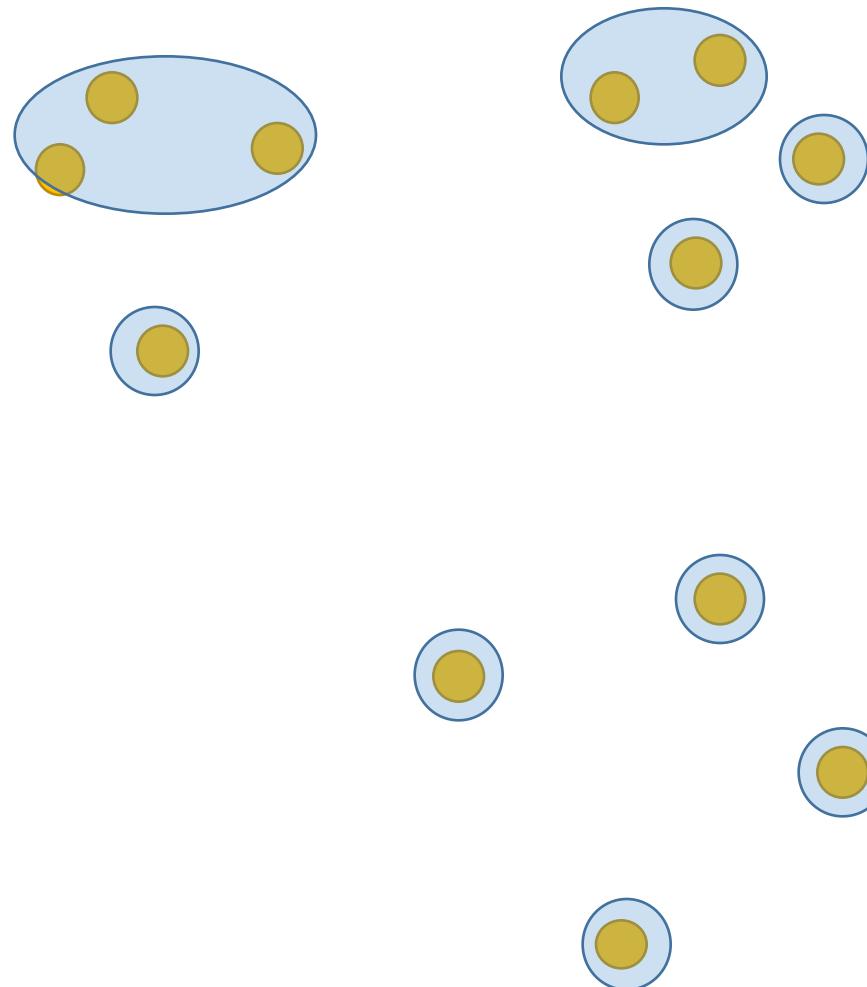
# Агломеративная кластеризация



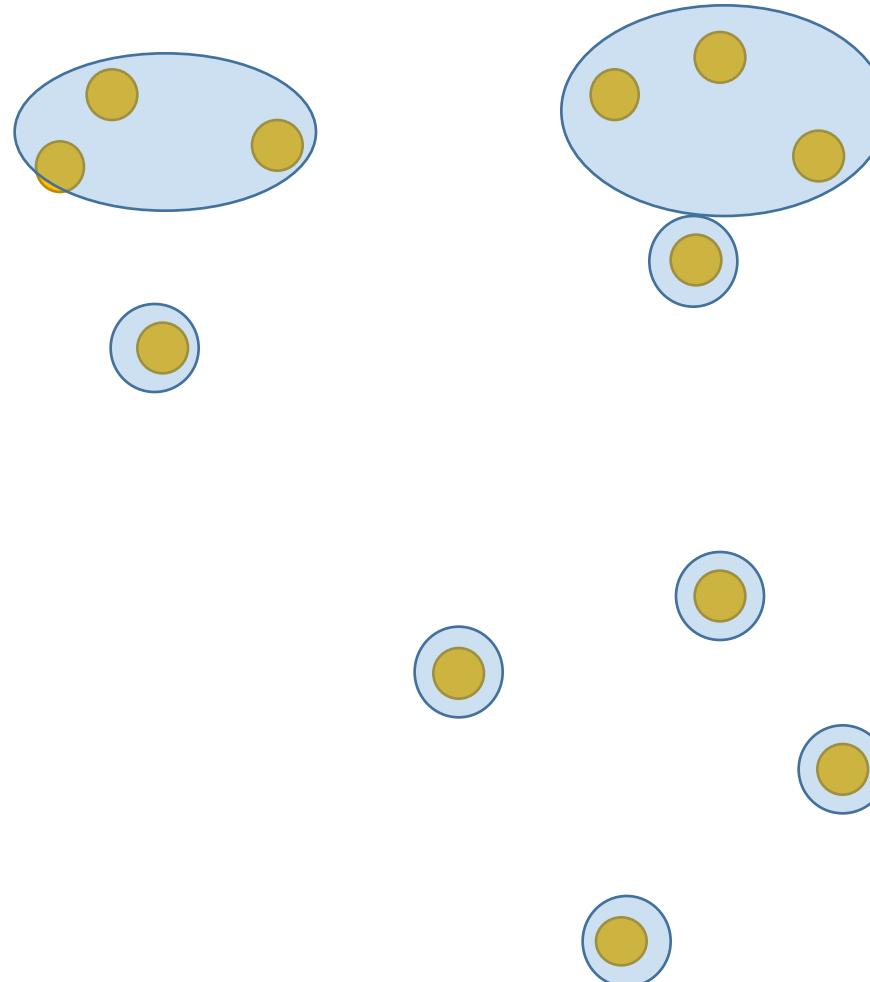
# Агломеративная кластеризация



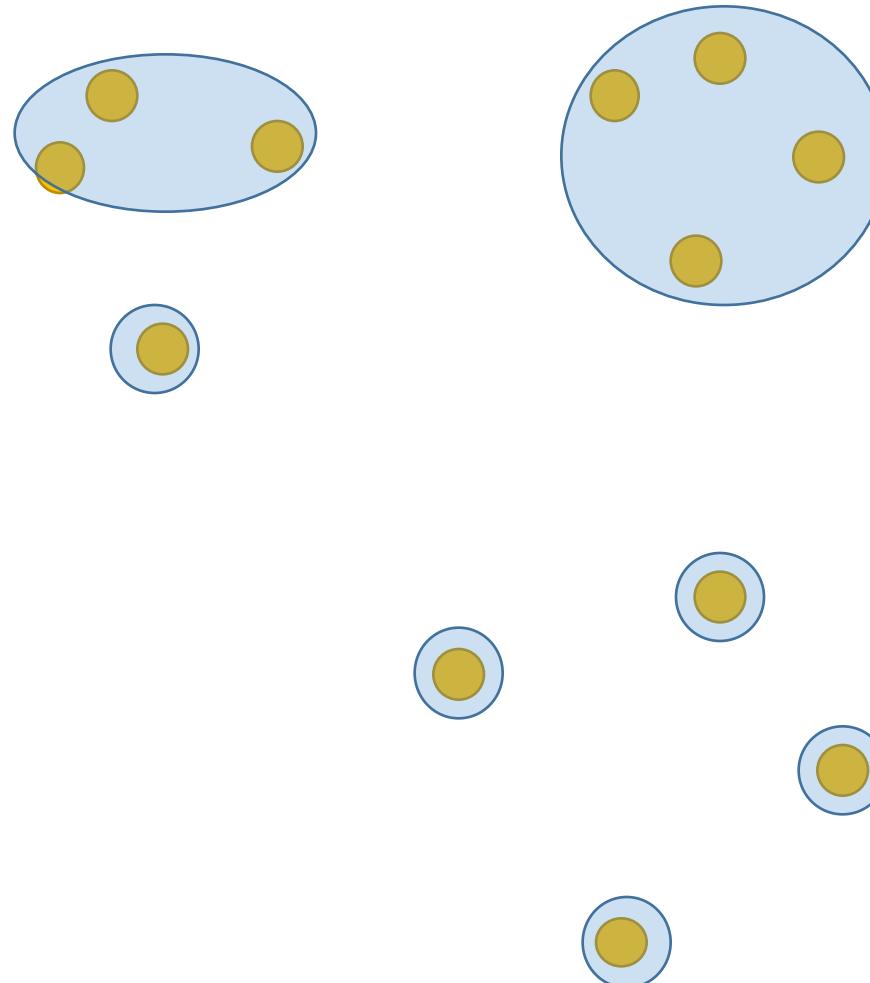
# Агломеративная кластеризация



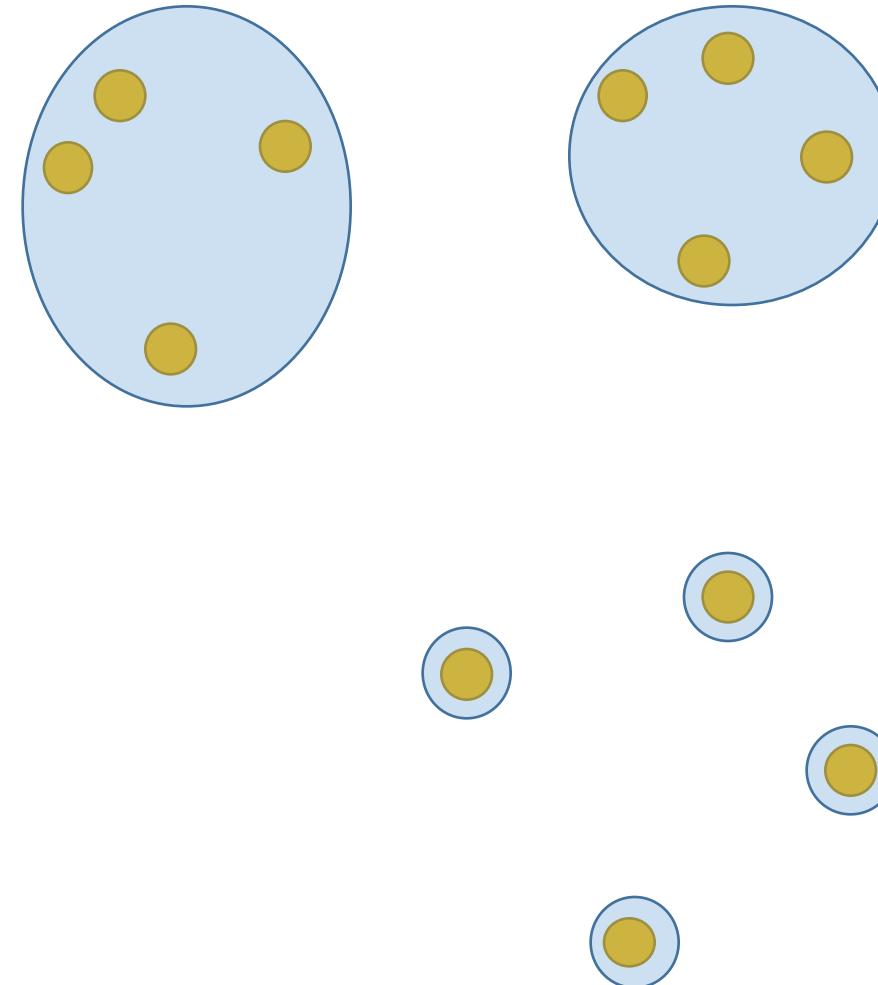
# Агломеративная кластеризация



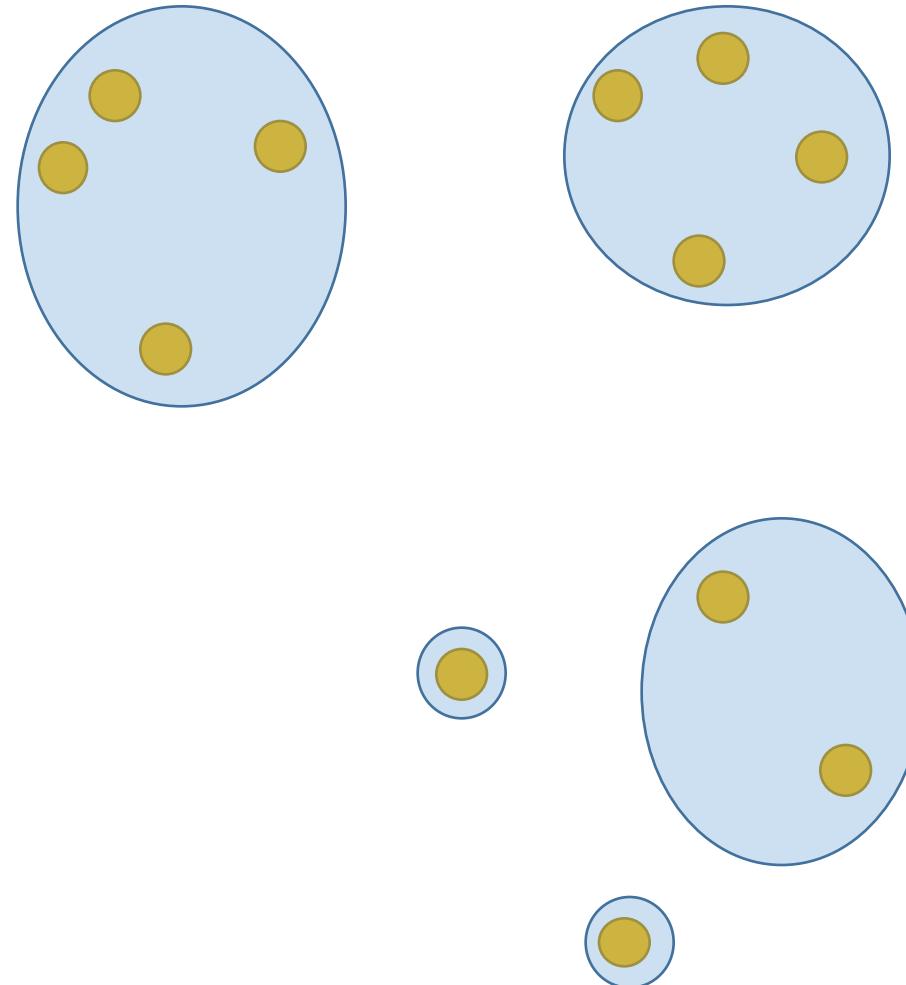
# Агломеративная кластеризация



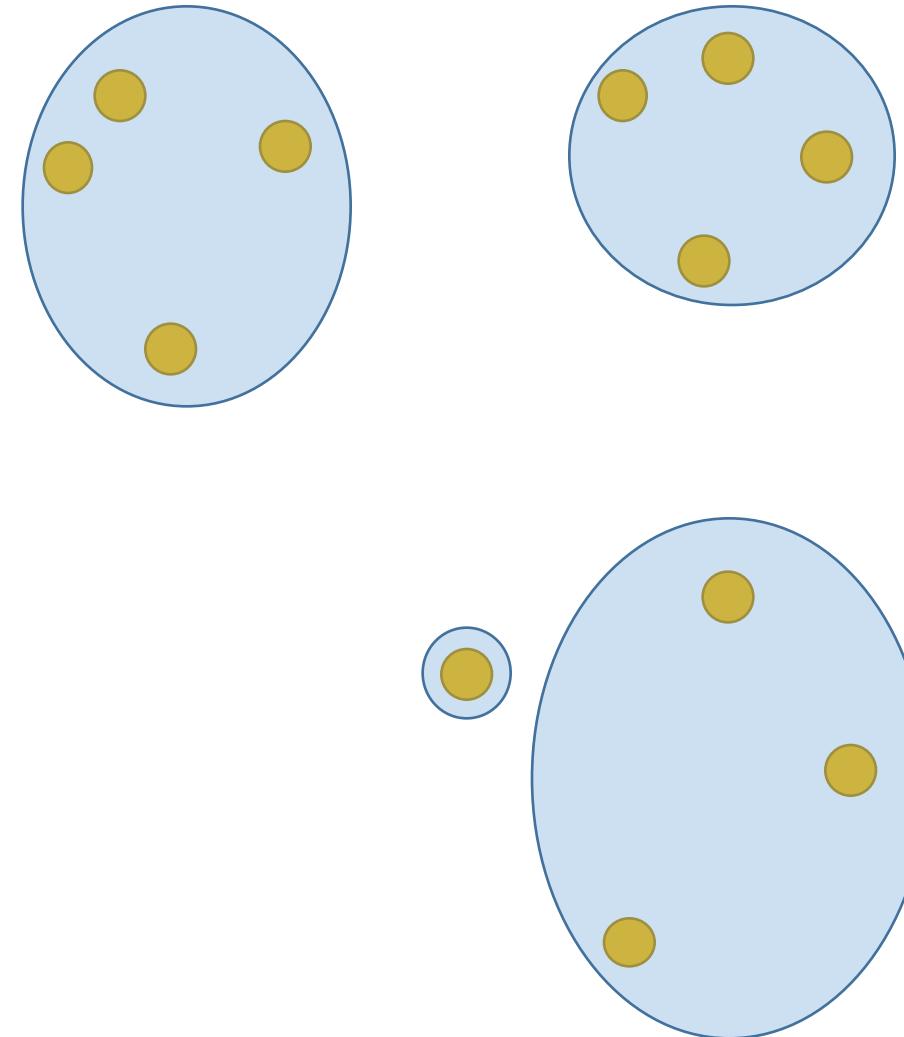
# Агломеративная кластеризация



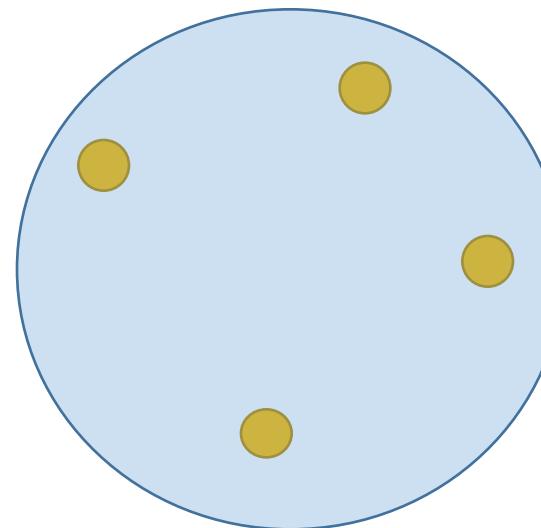
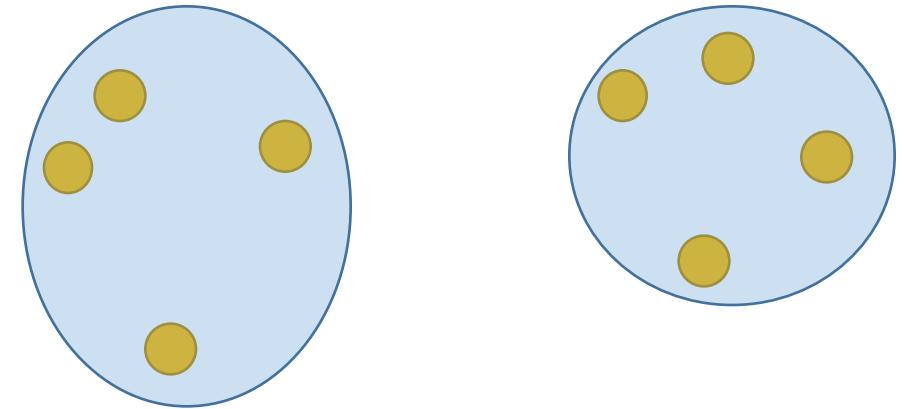
# Агломеративная кластеризация



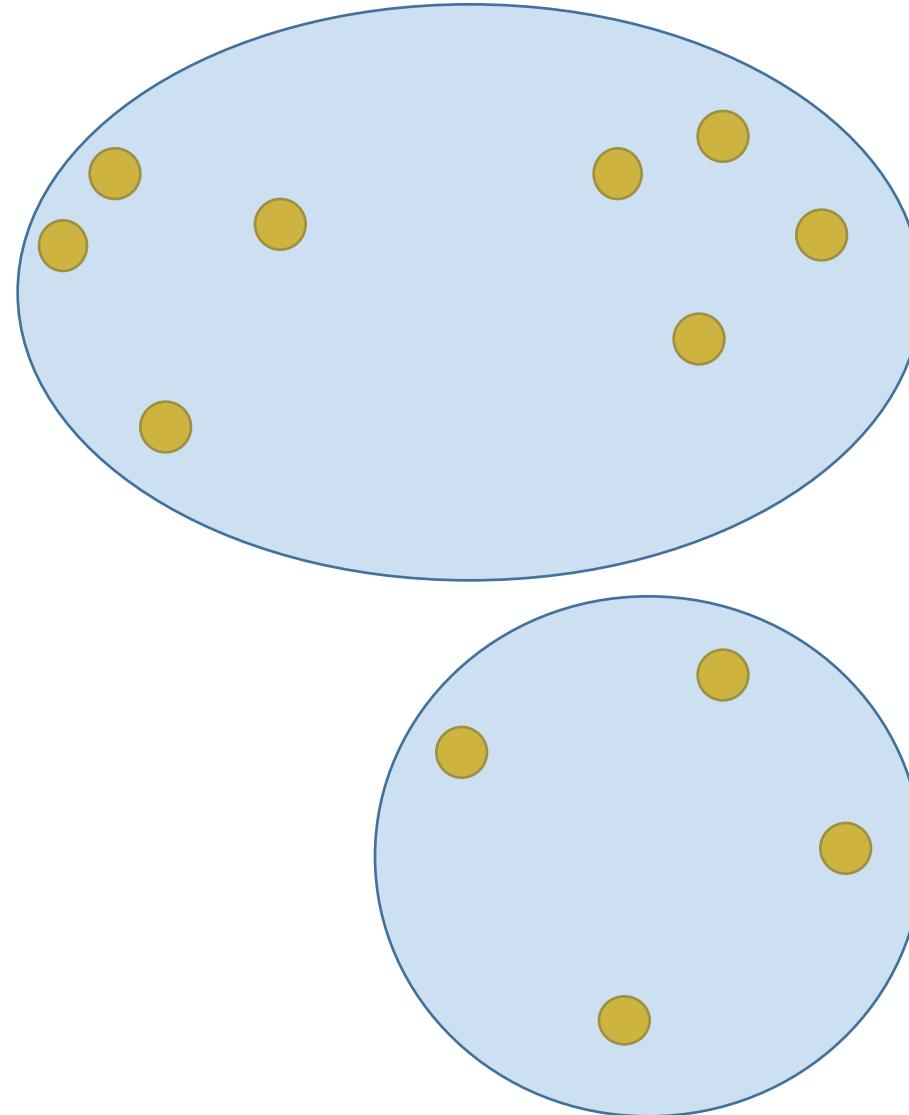
# Агломеративная кластеризация



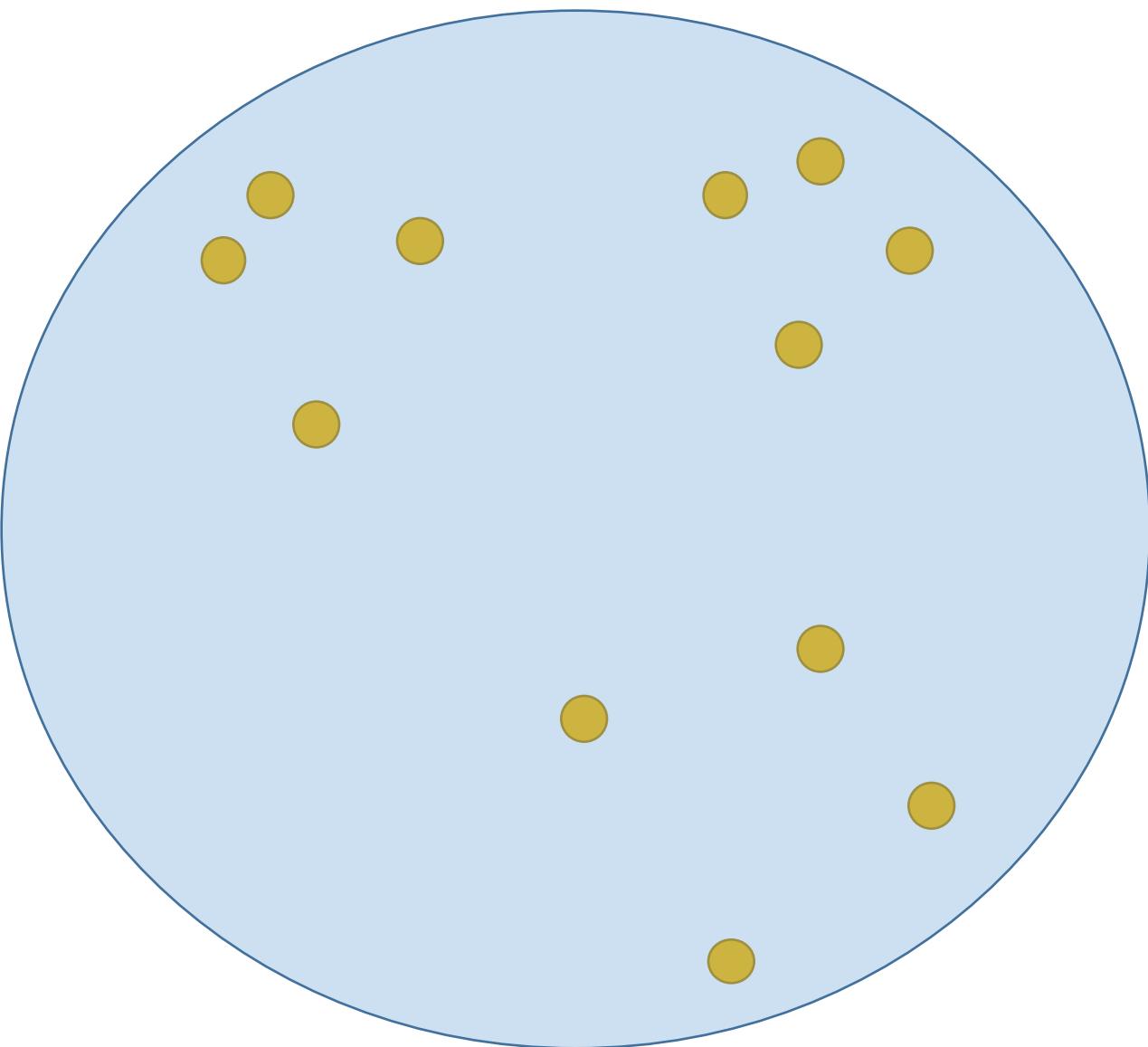
# Агломеративная кластеризация



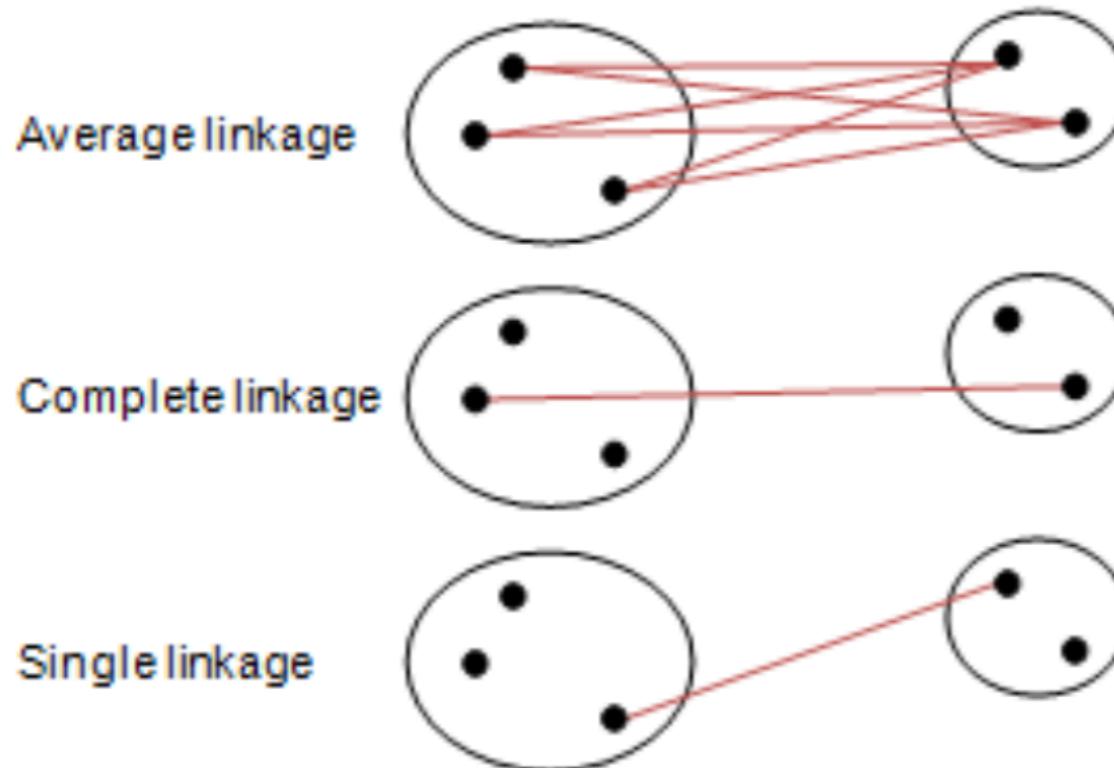
# Агломеративная кластеризация



# 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

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

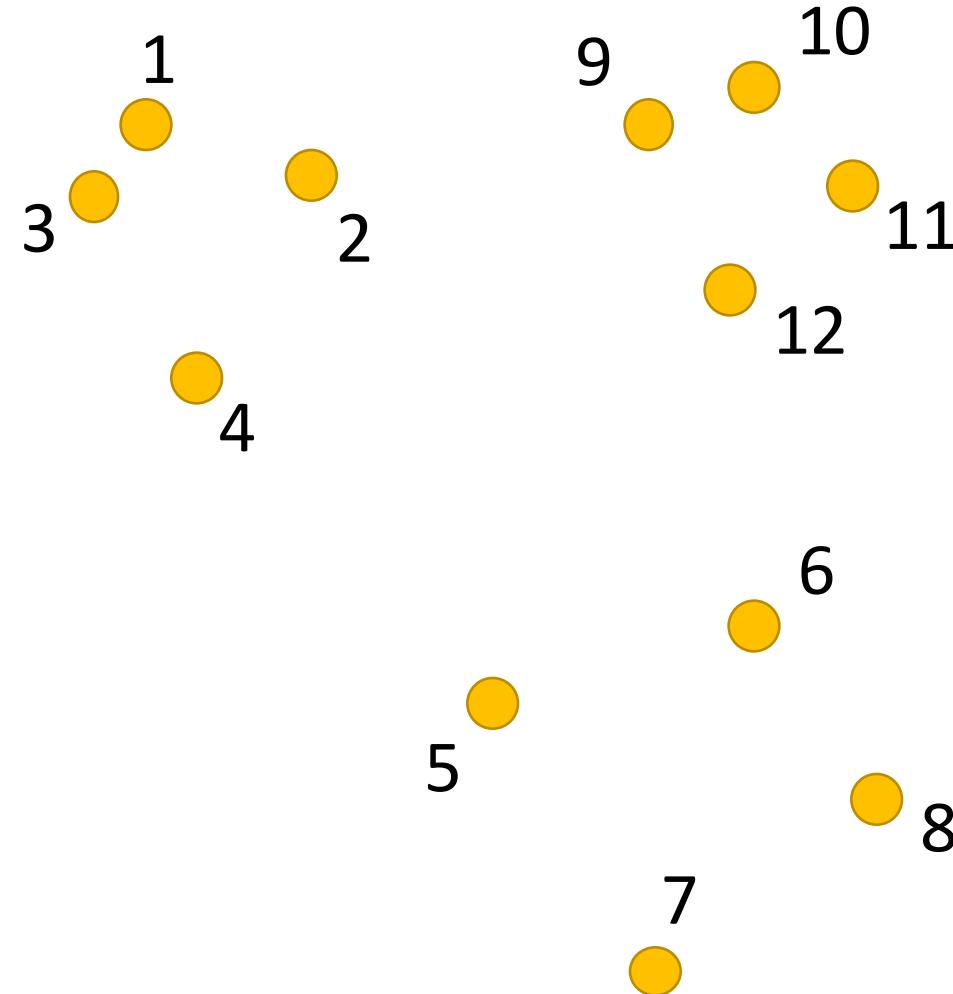
*Расстояние между центрами:*

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

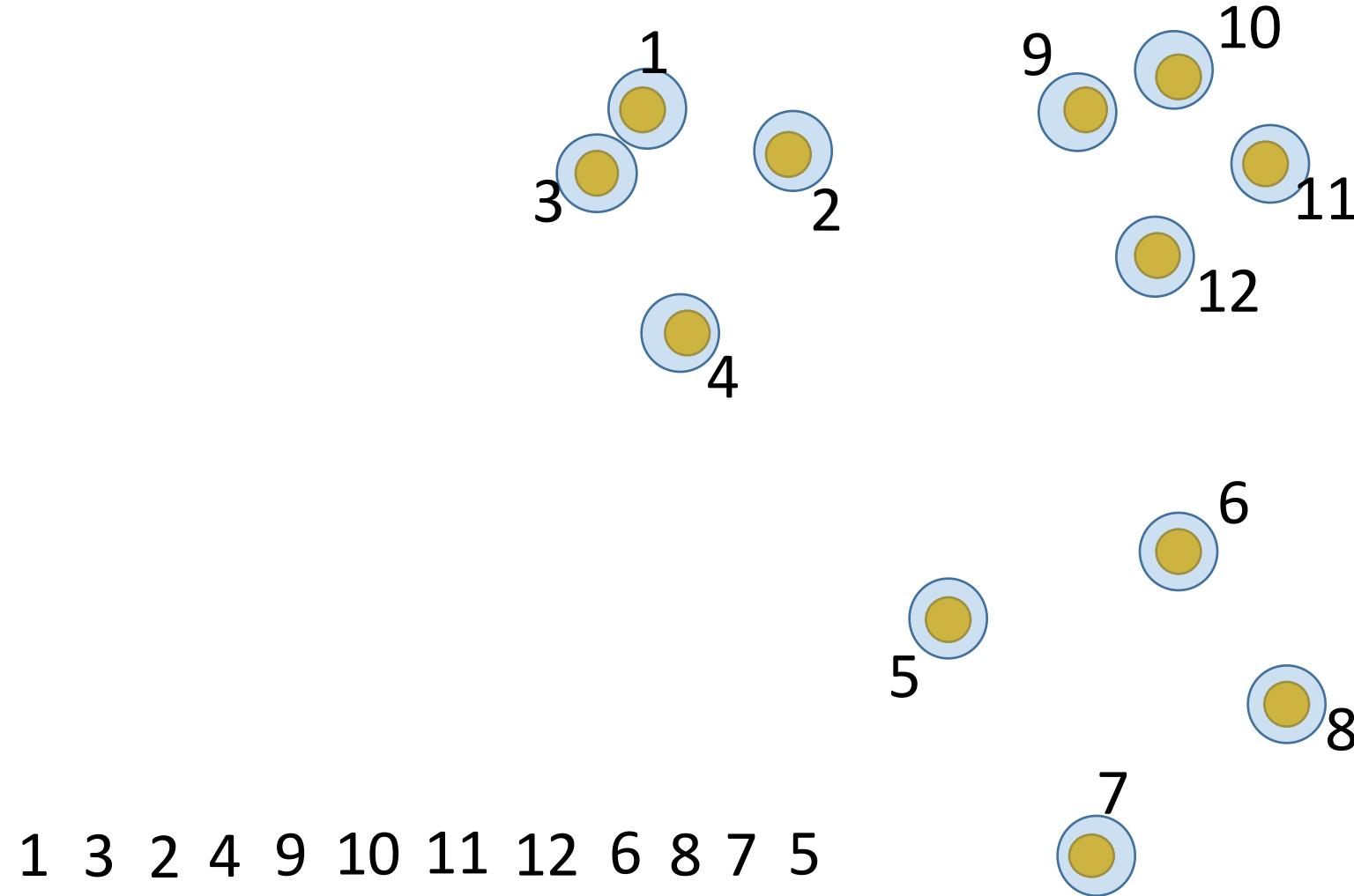
*Расстояние Уорда:*

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

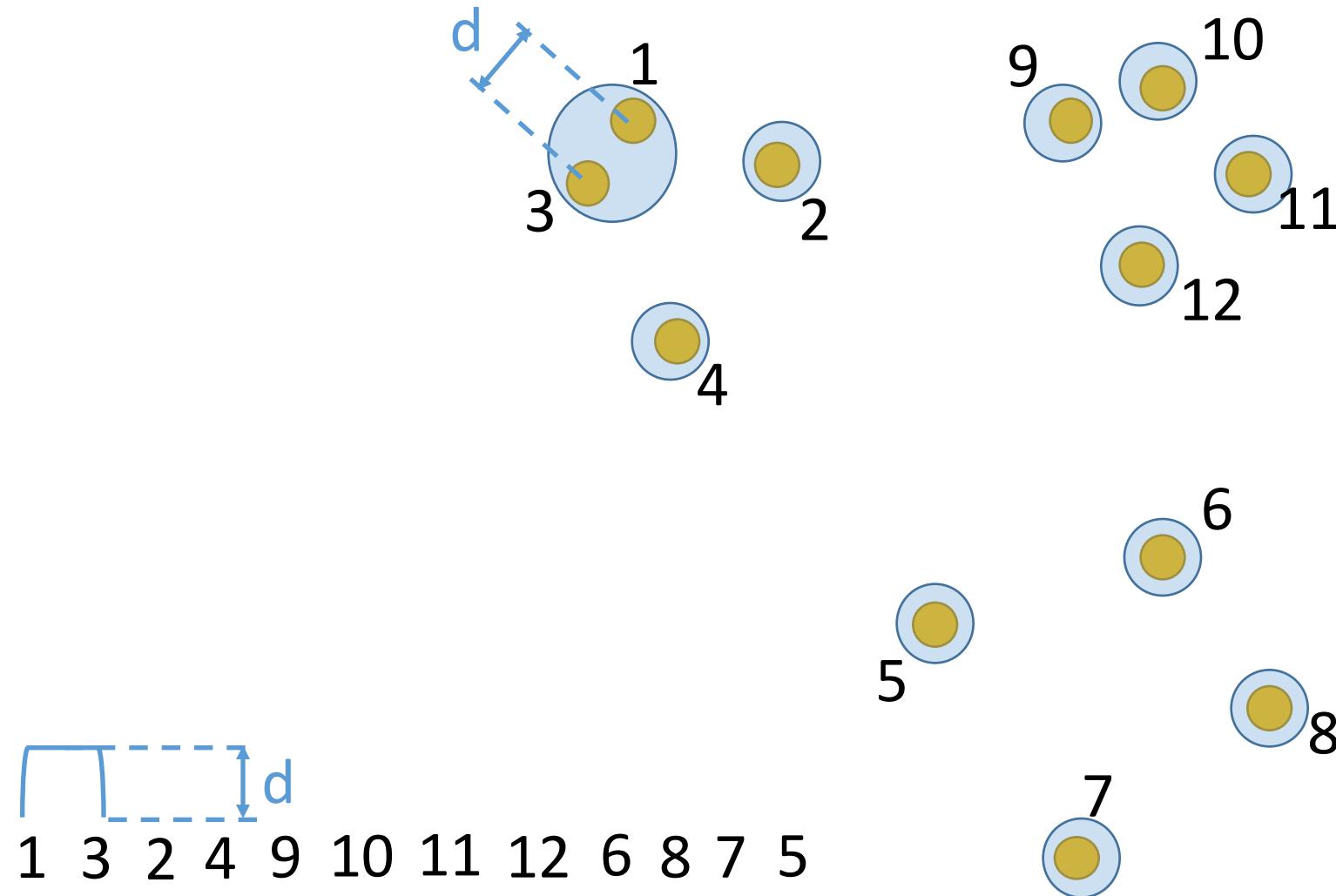
# Дендрограмма



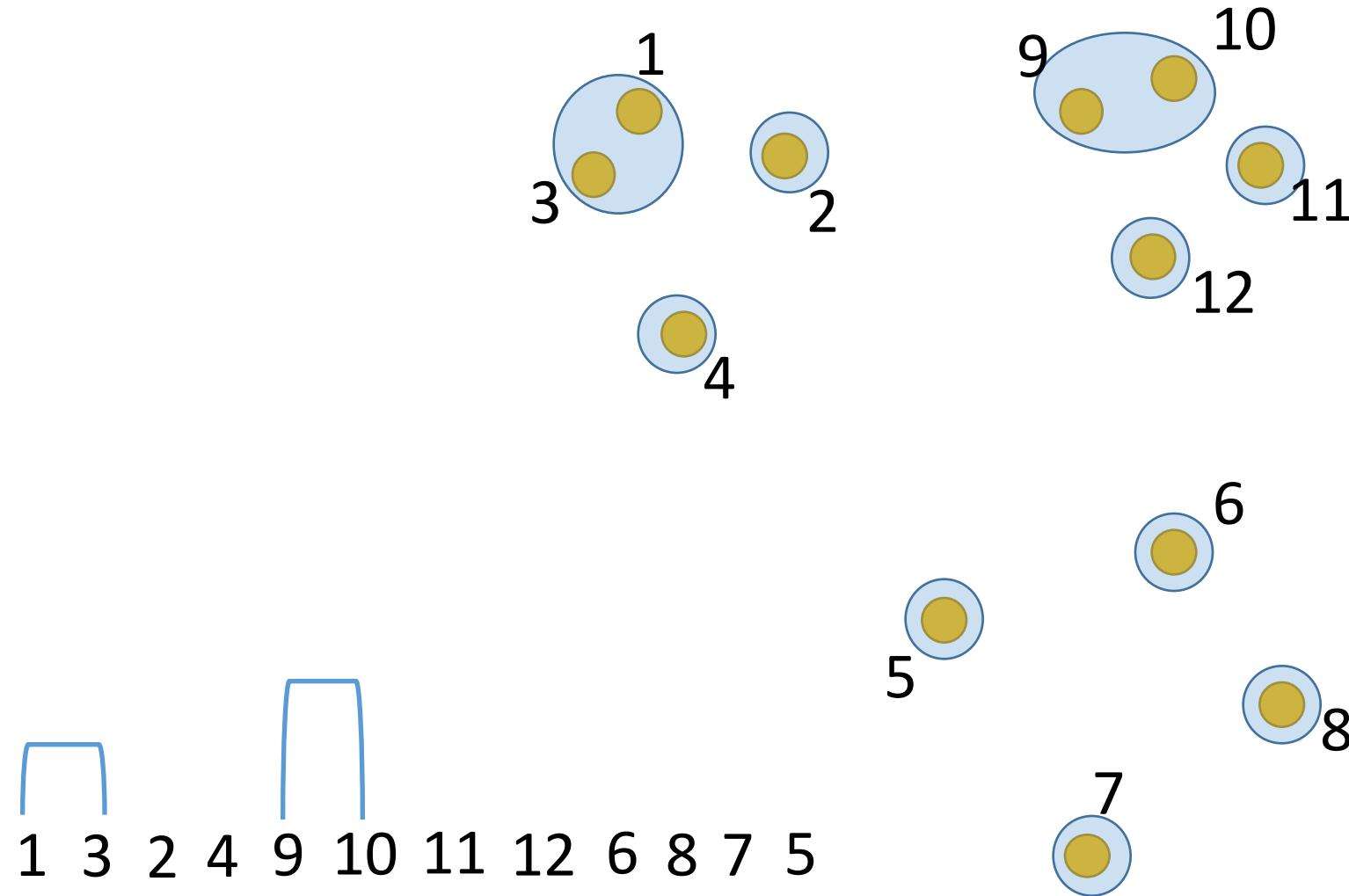
# Дендрограмма



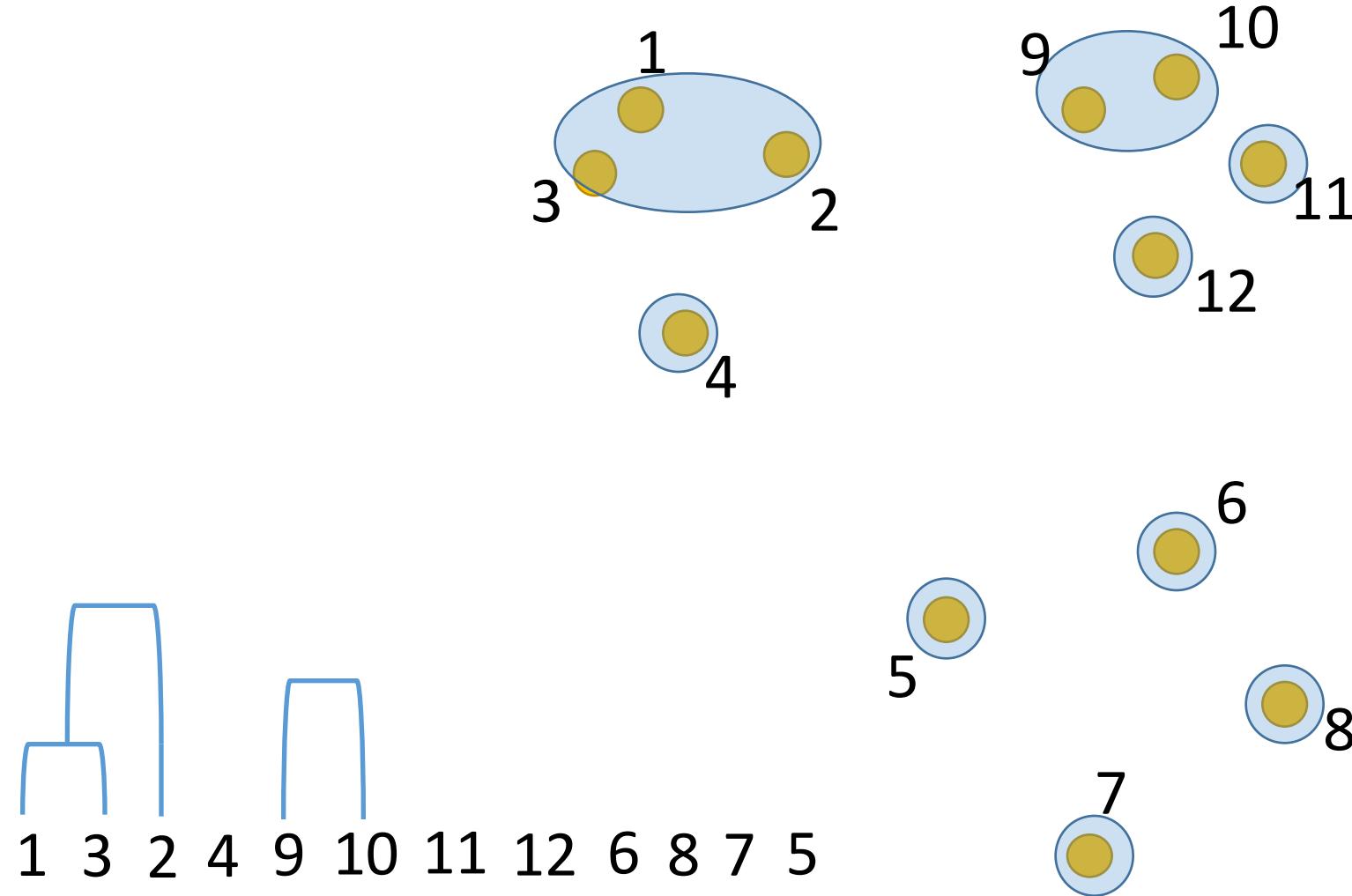
# Дендрограмма



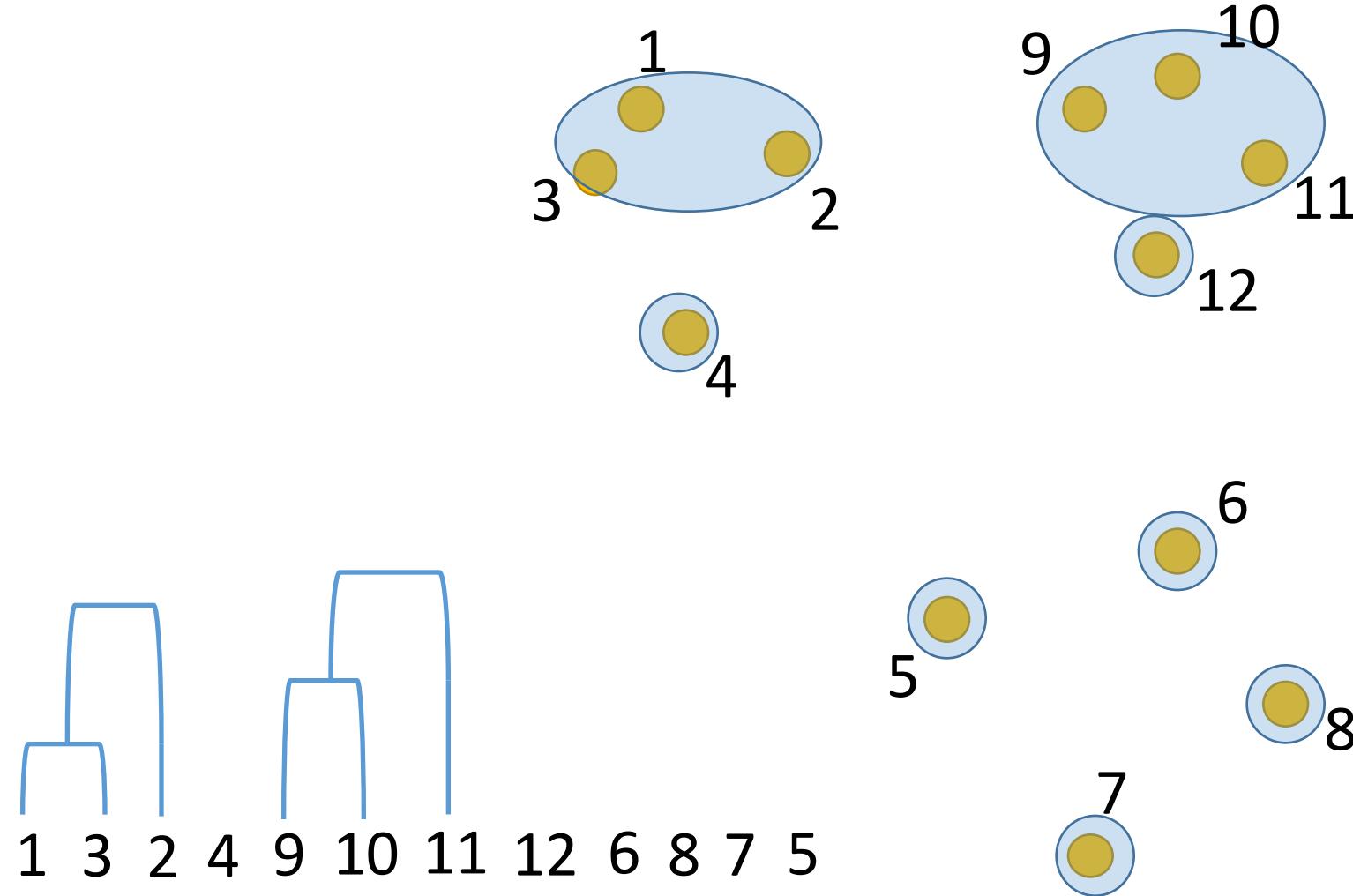
# Дендрограмма



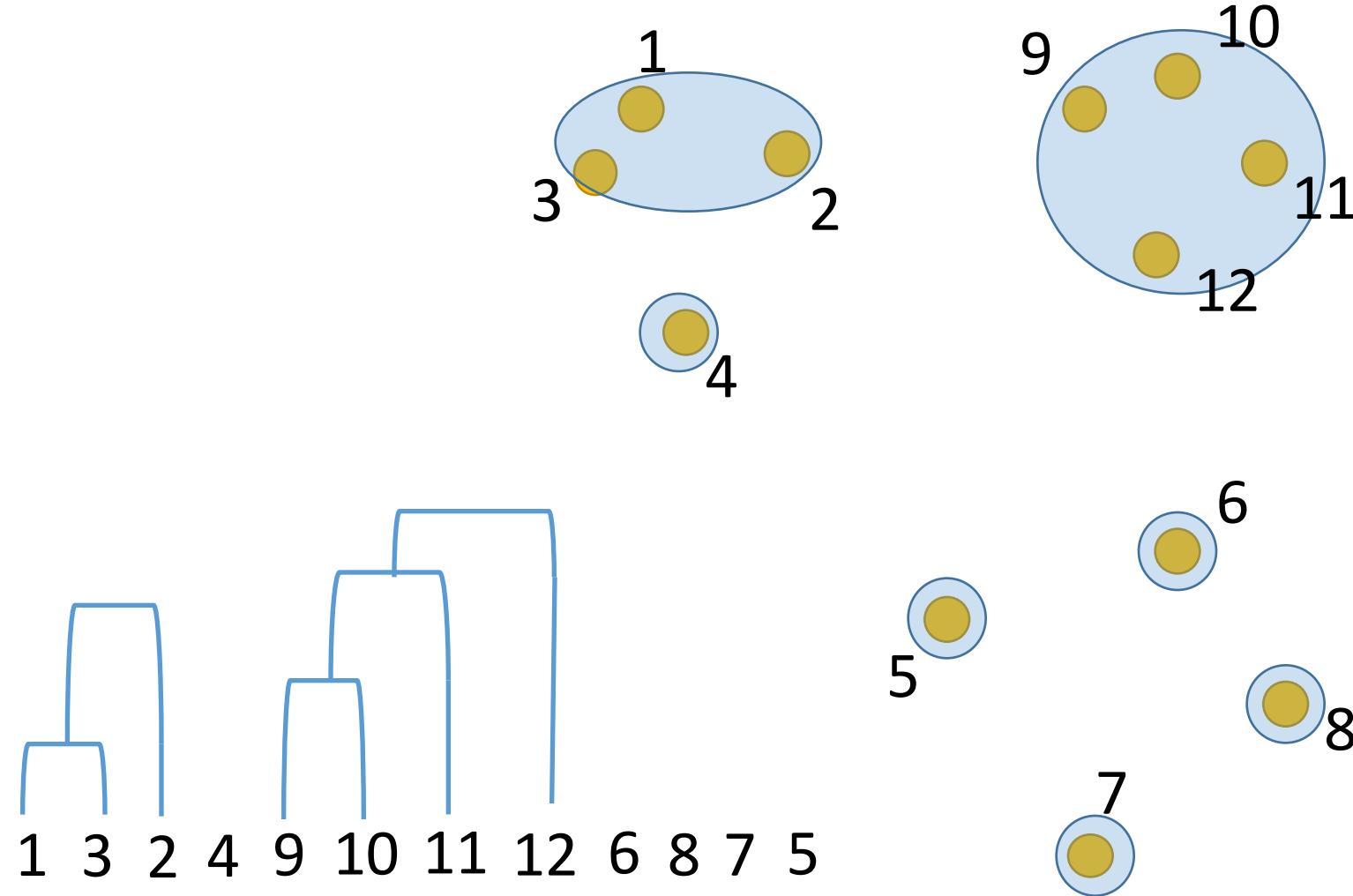
# Дендрограмма



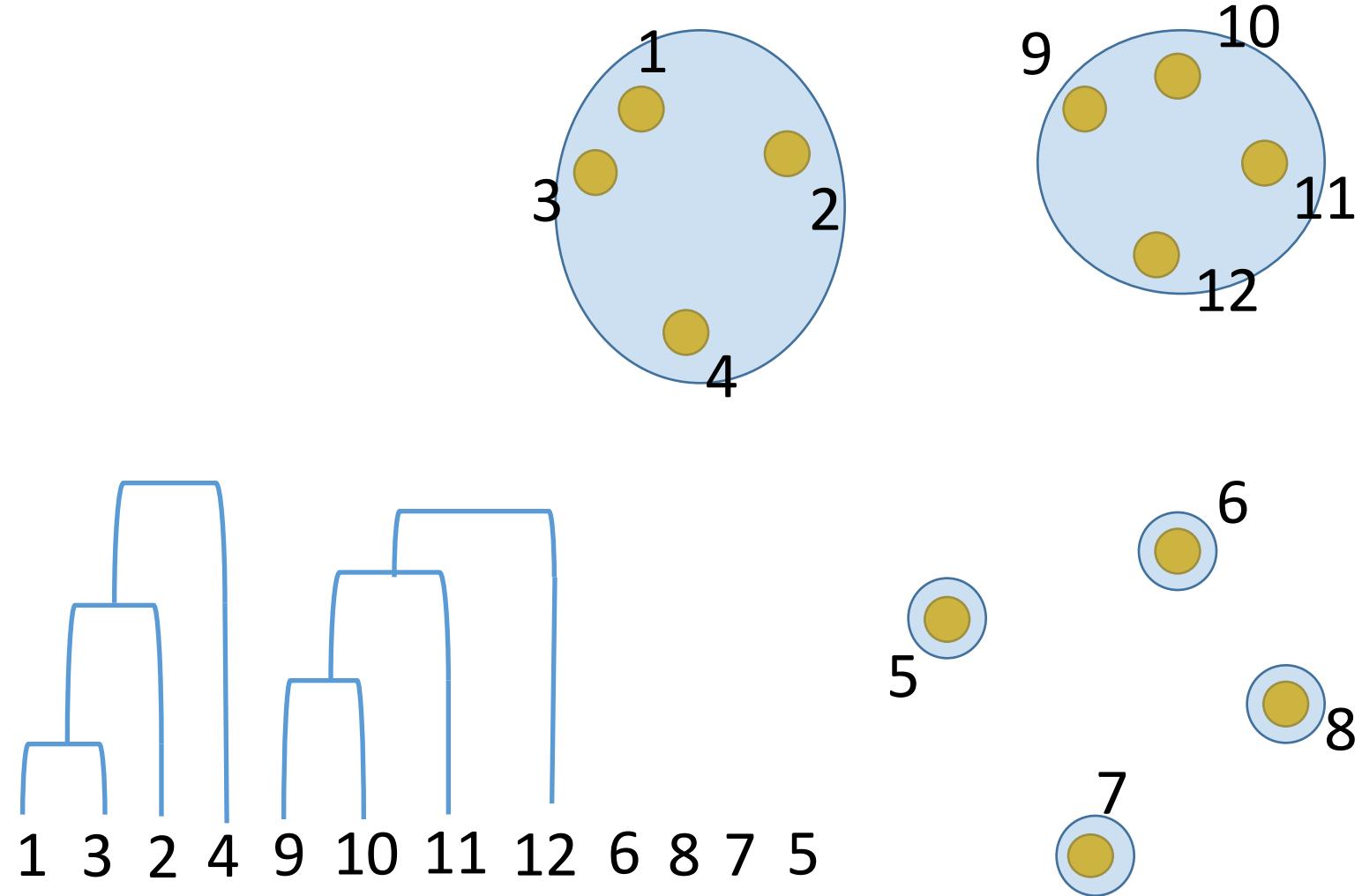
# Дендрограмма



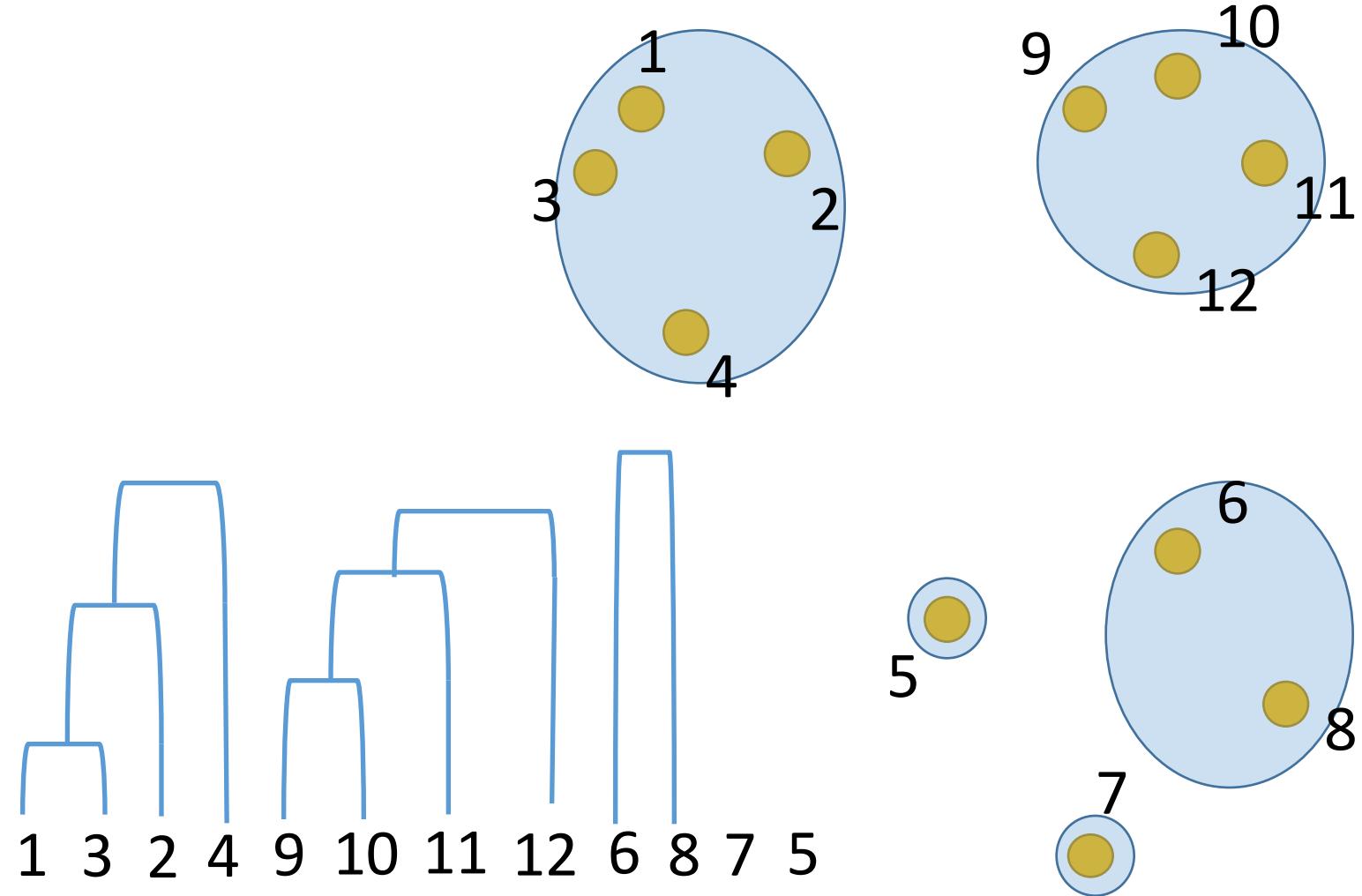
# Дендрограмма



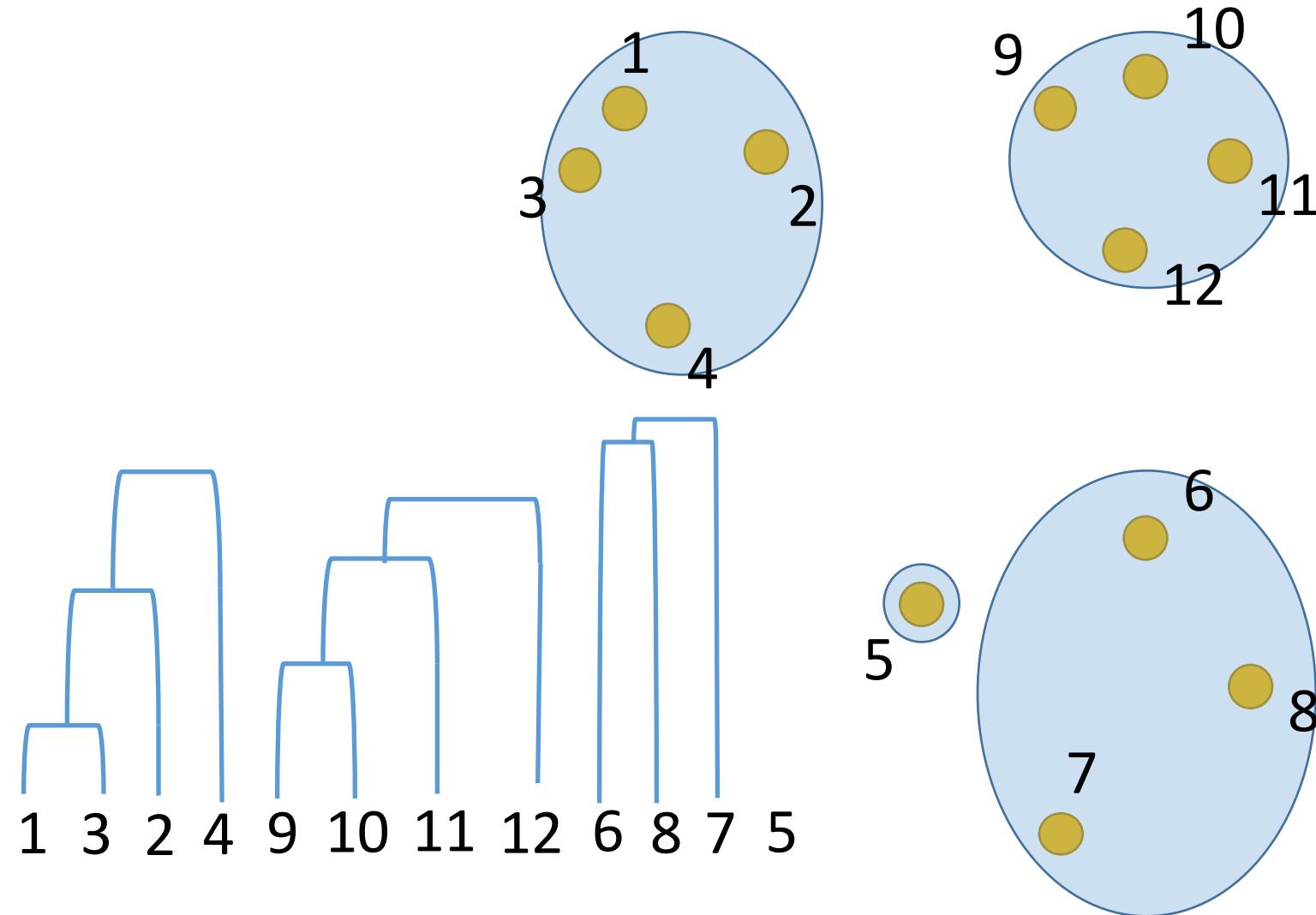
# Дендрограмма



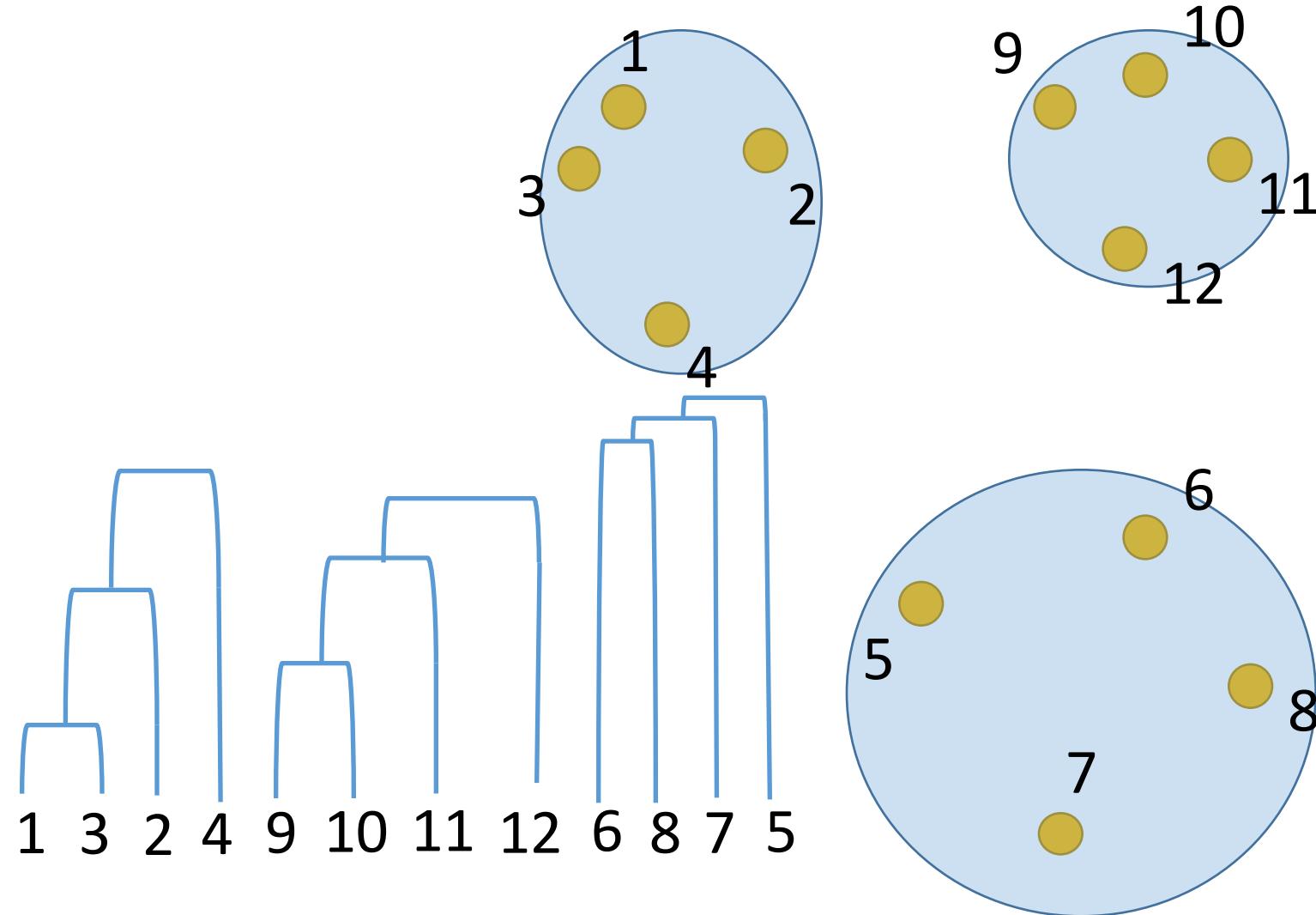
# Дендрограмма



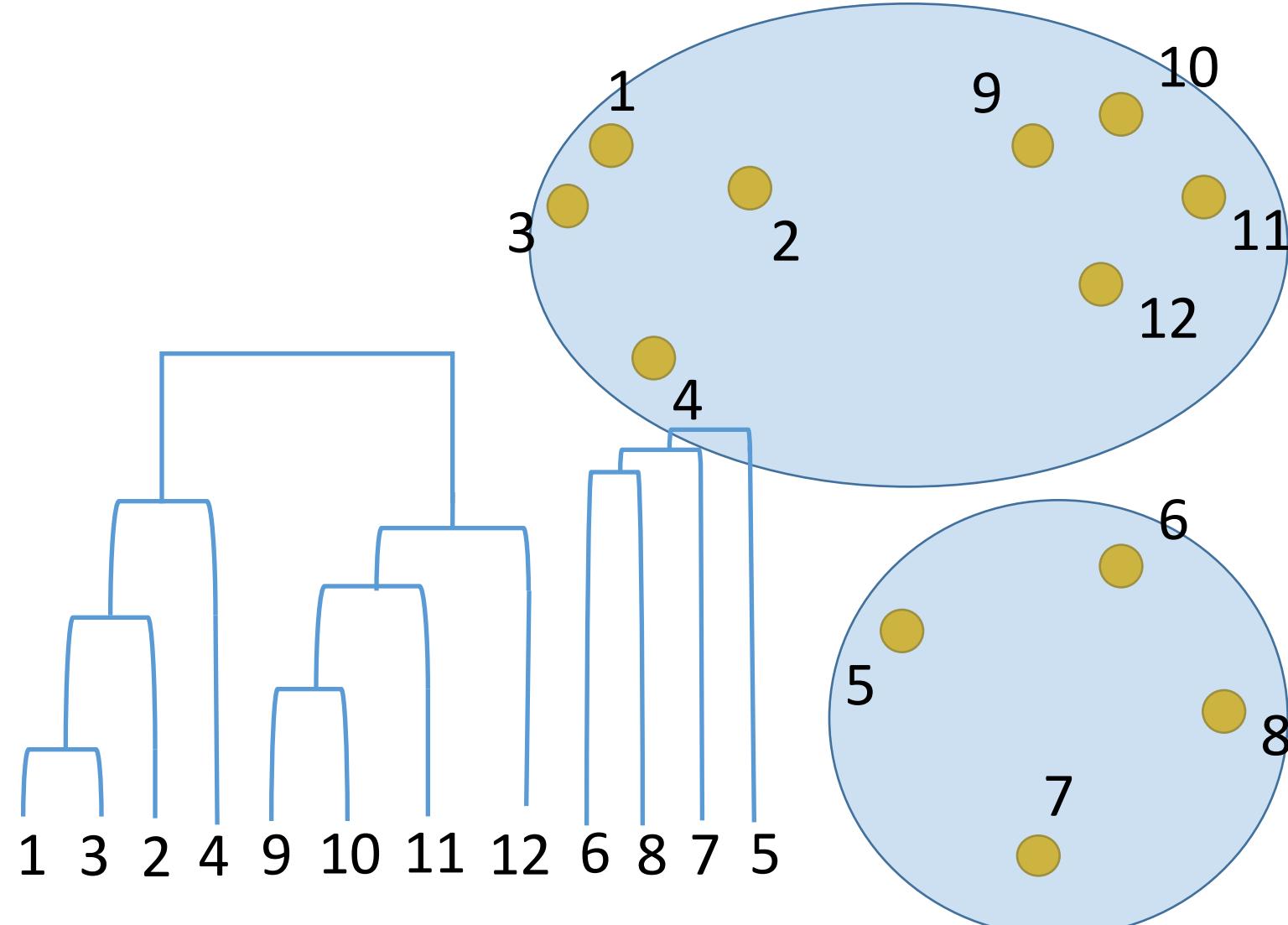
# Дендрограмма



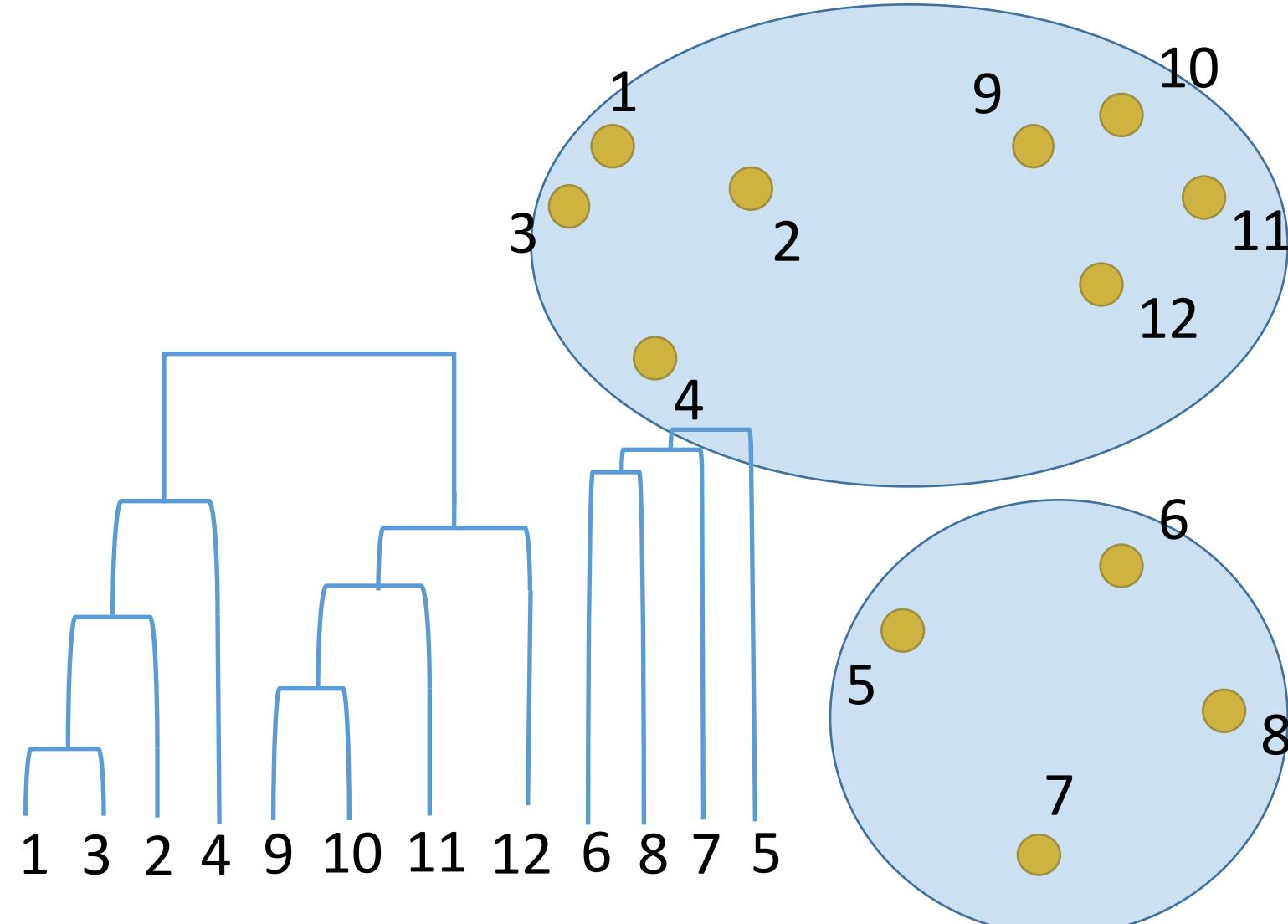
# Дендрограмма



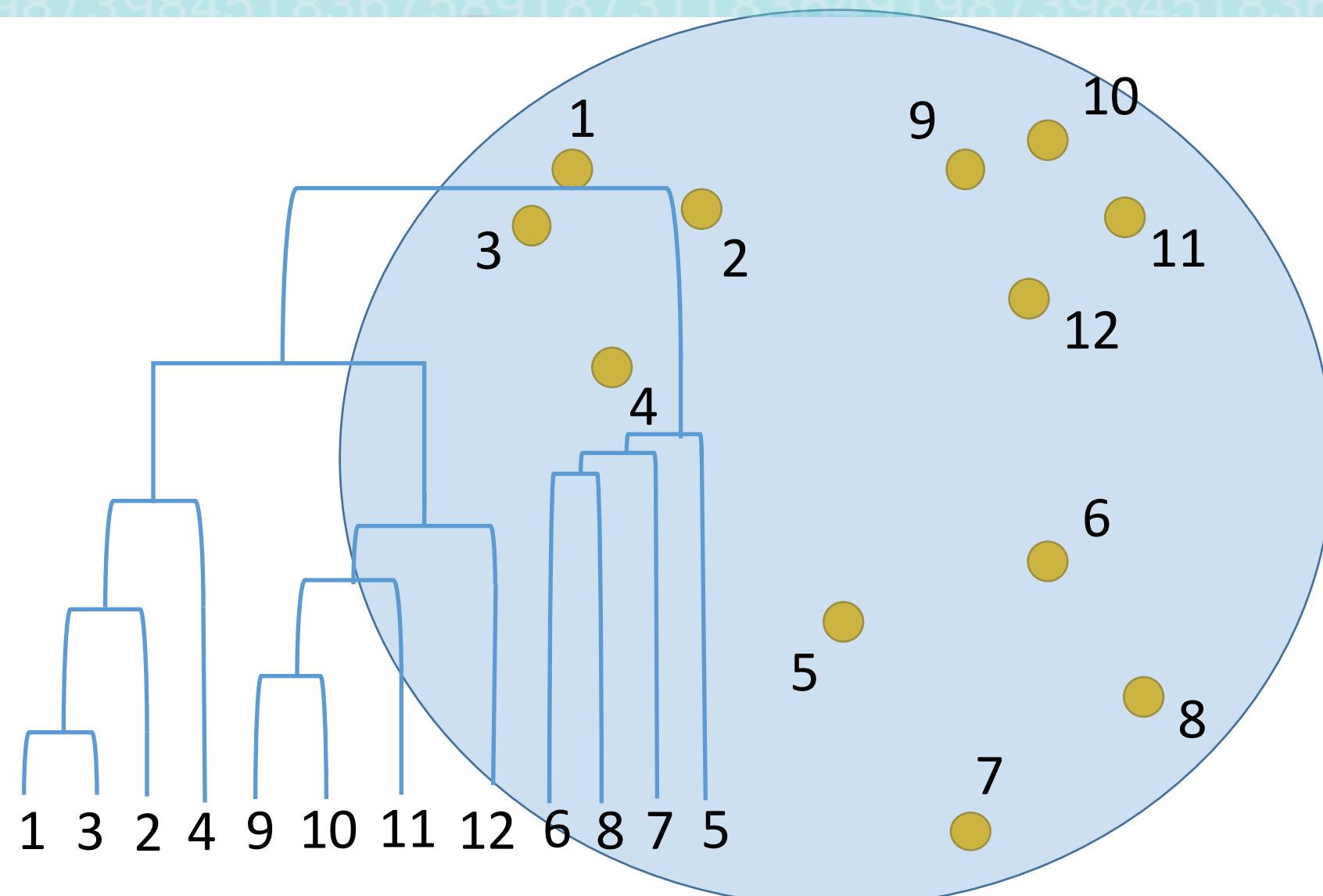
# Дендрограмма



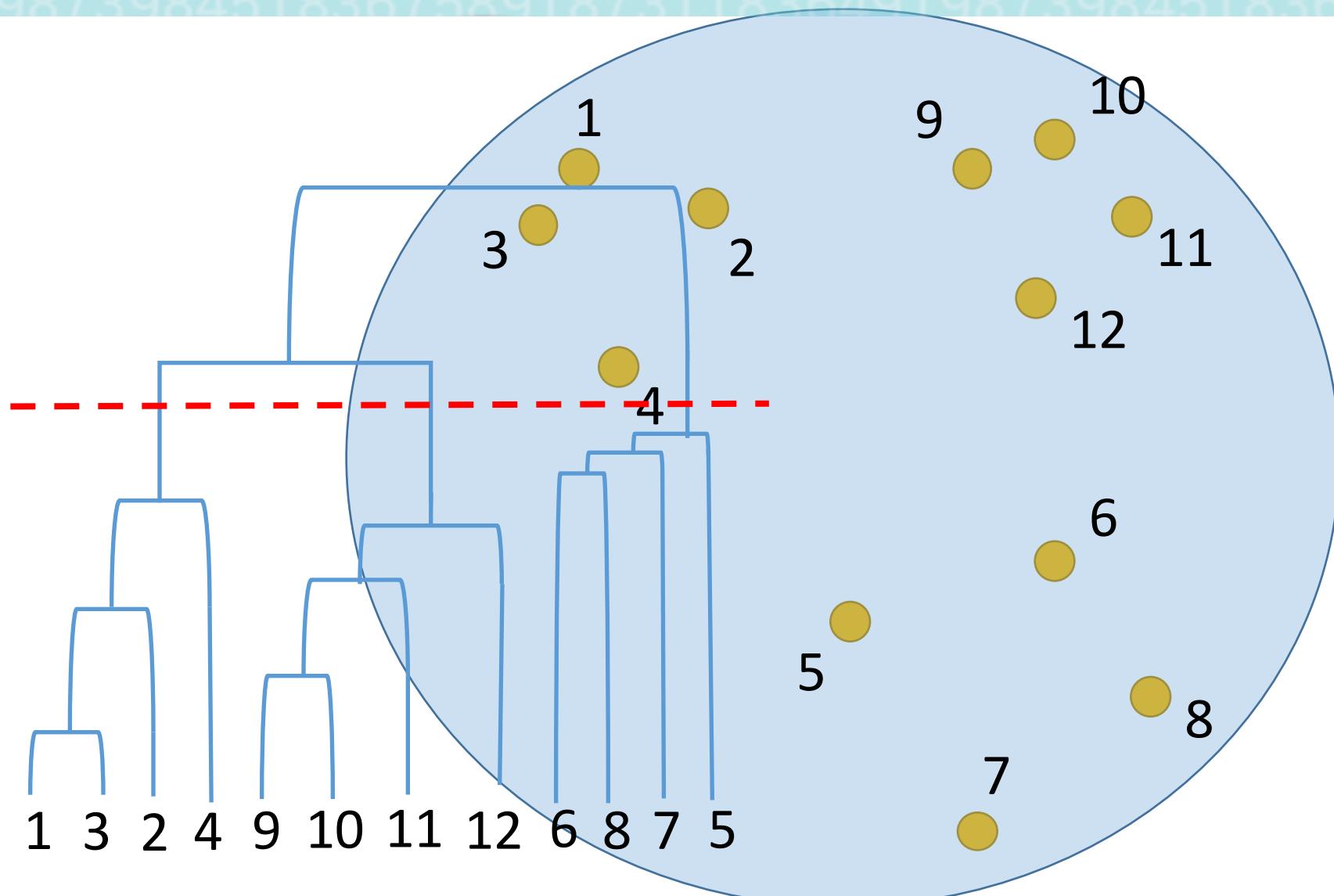
# Дендрограмма



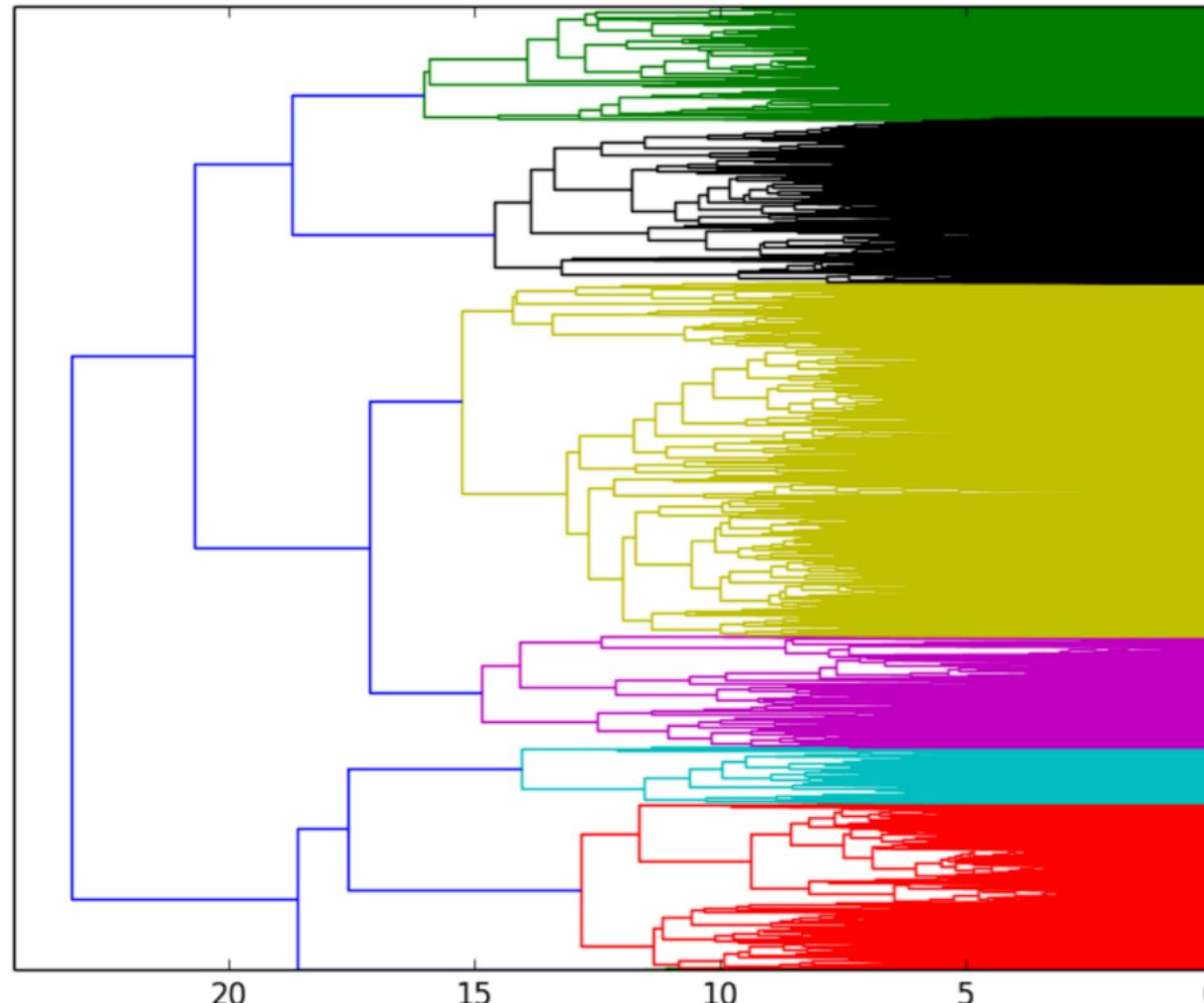
# Дендрограмма



# Дендрограмма

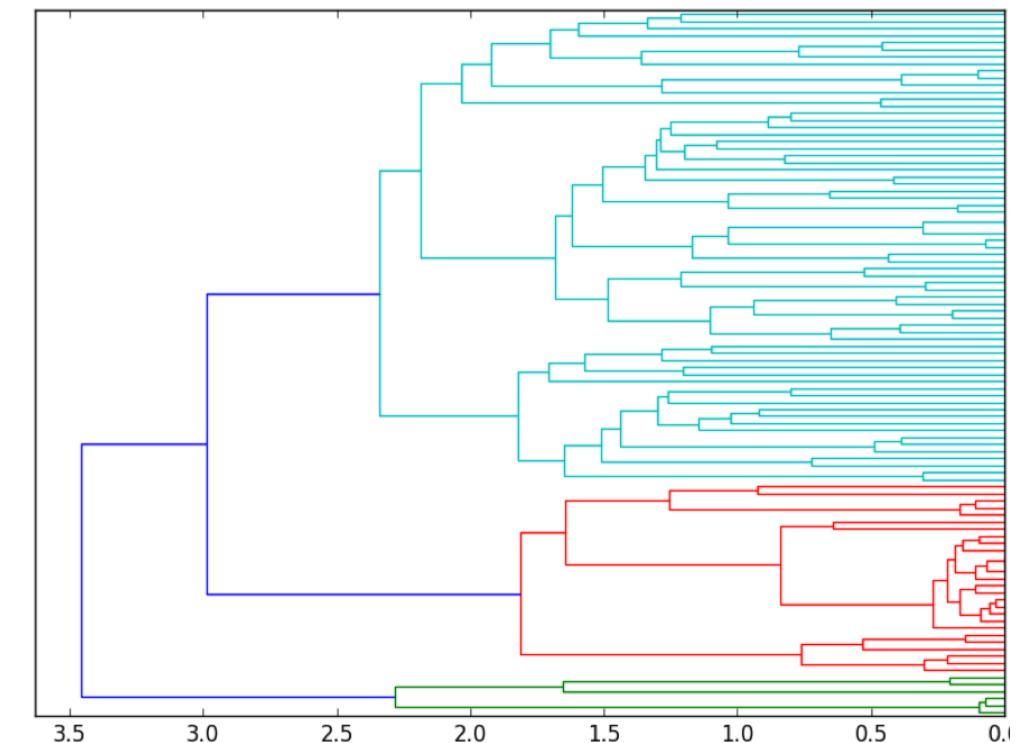
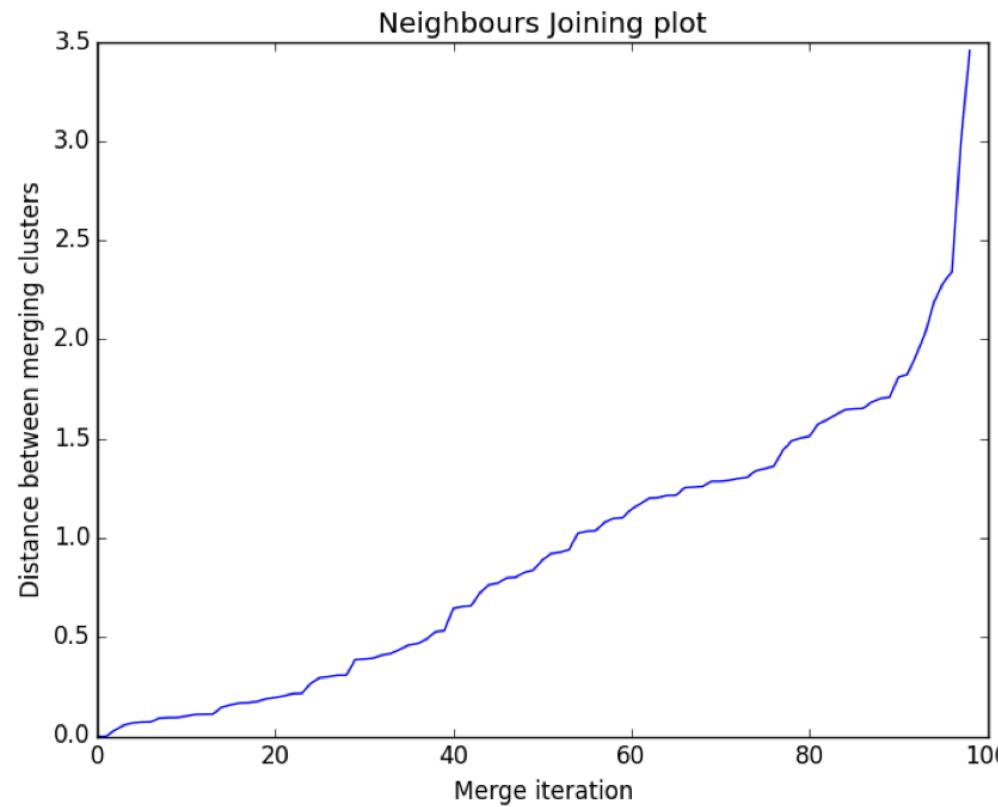


# Пример: кластеризация писем



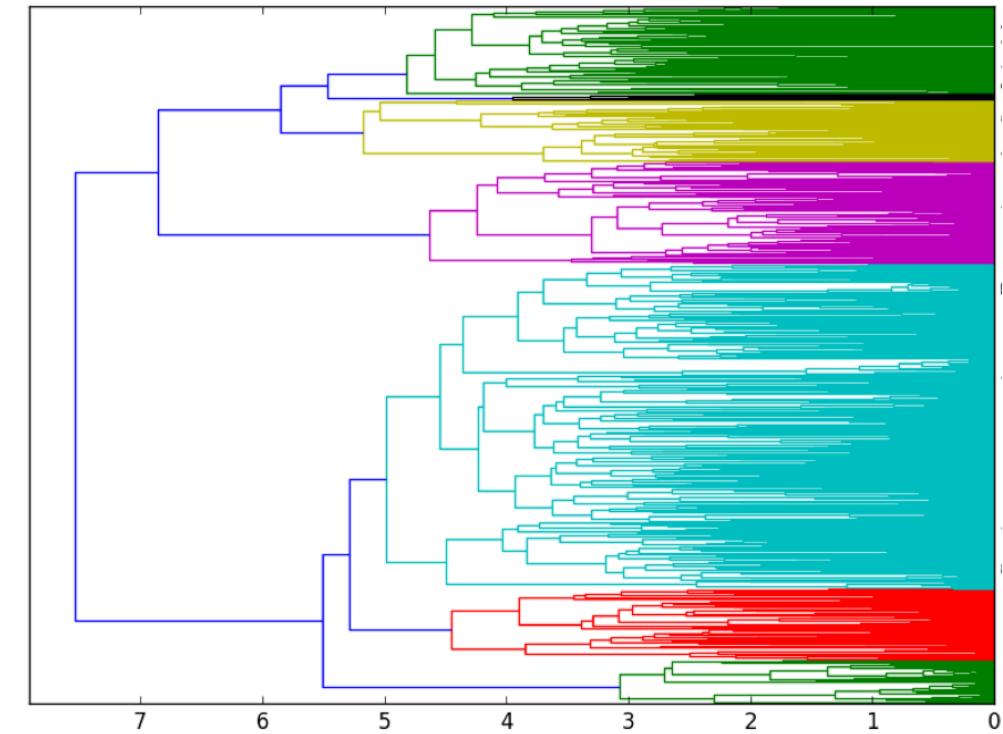
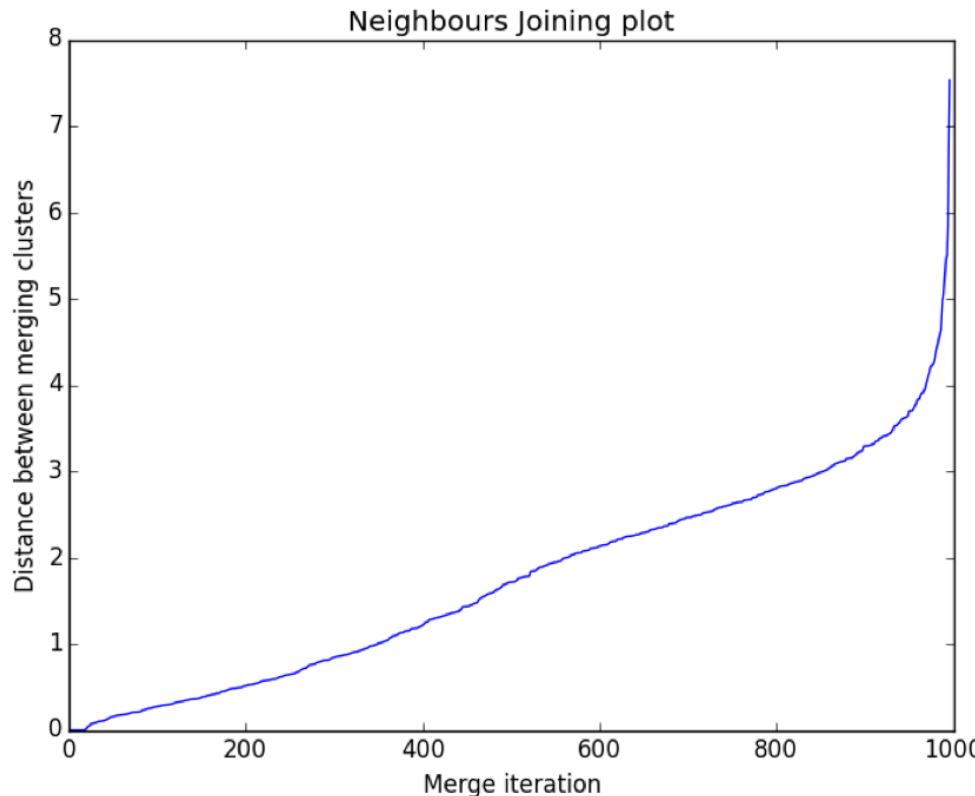
# Пример: расстояние между кластерами

- На подвыборке из 100 писем



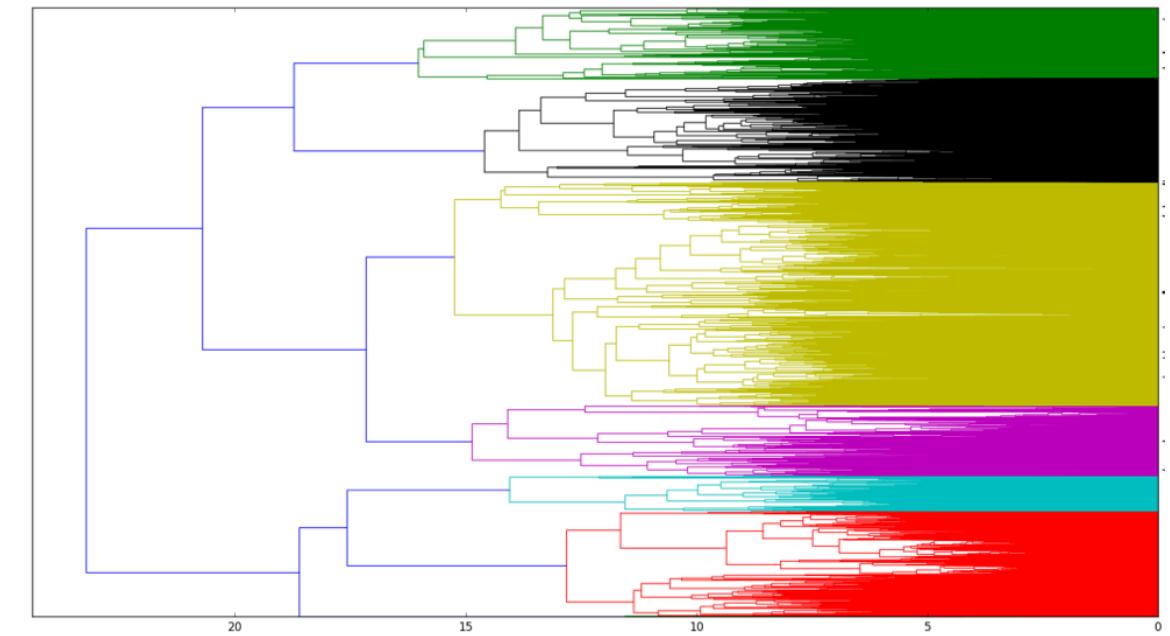
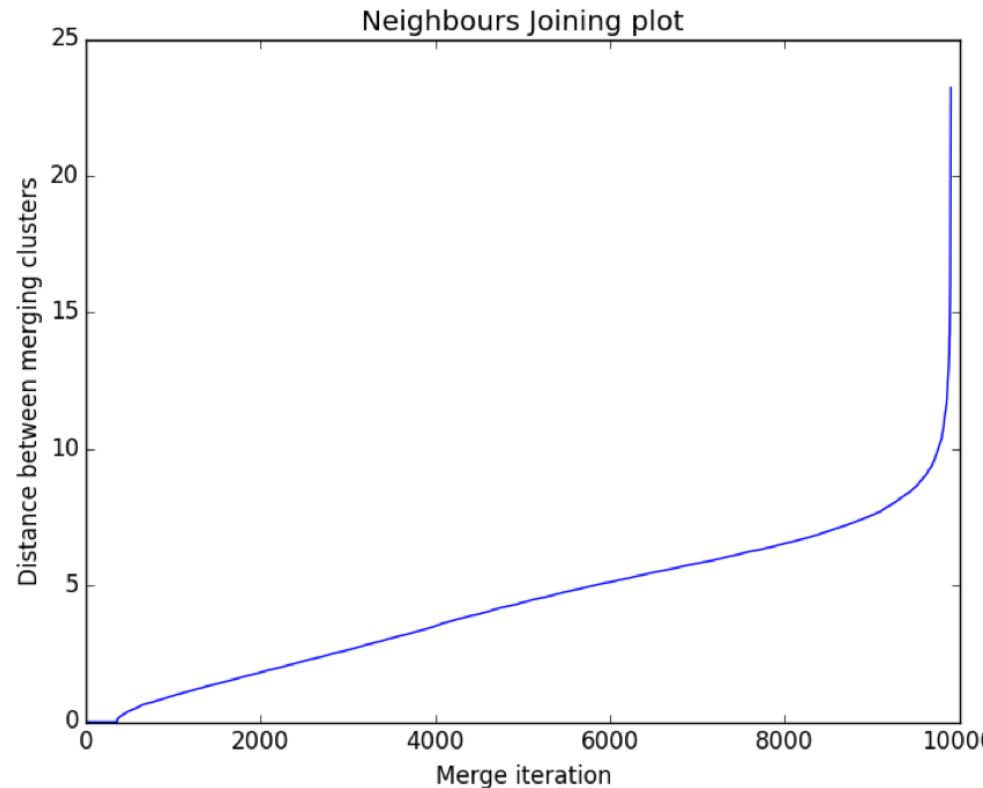
# Пример: расстояние между кластерами

- На подвыборке из 1000 писем



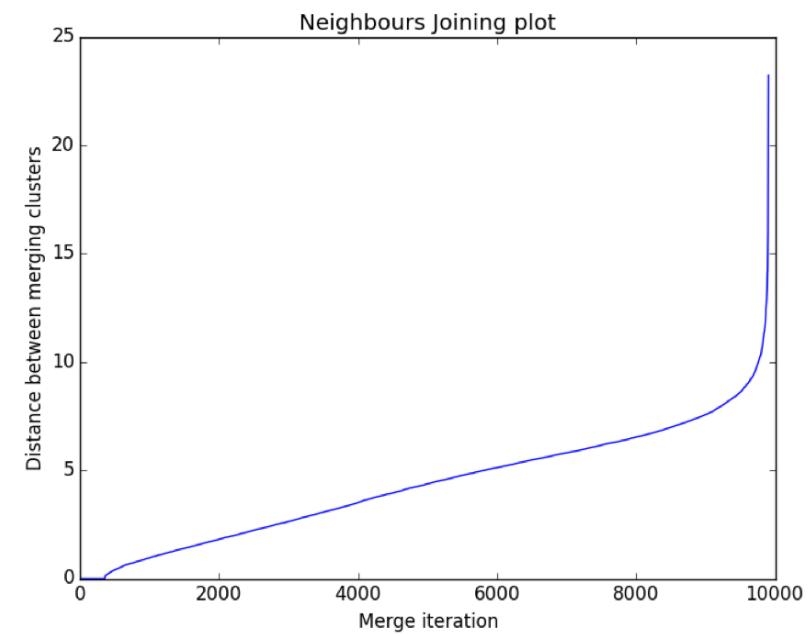
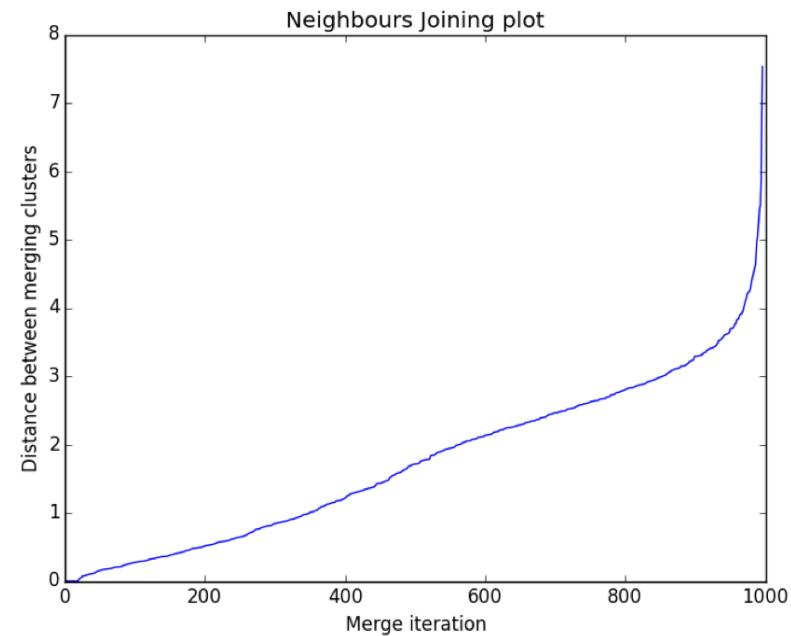
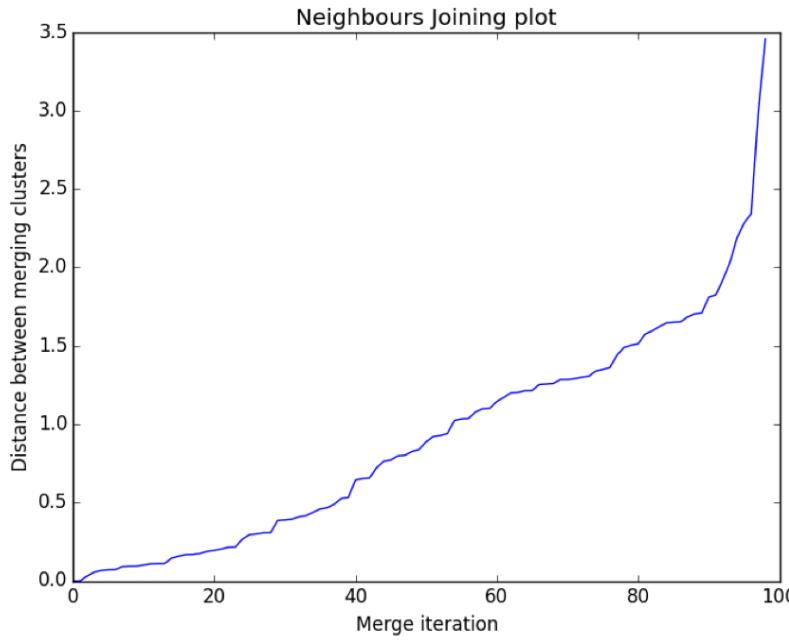
# Пример: расстояние между кластерами

- На подвыборке из 10000 писем



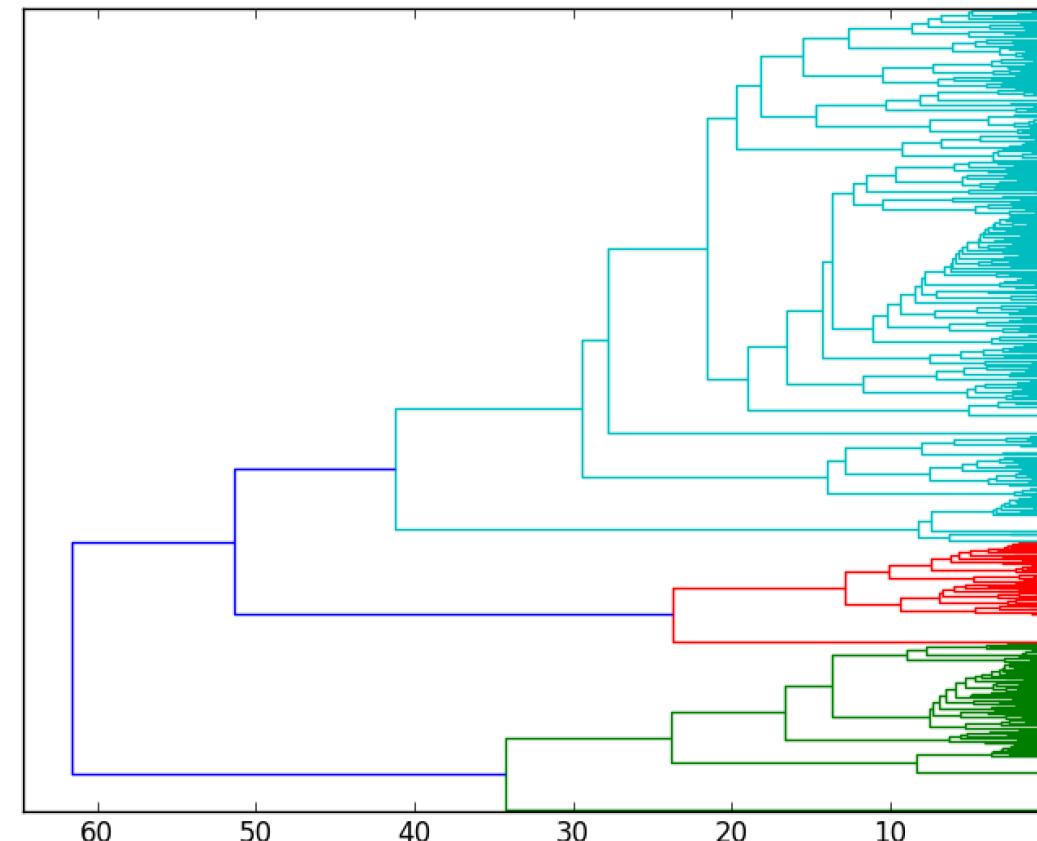
# Пример: расстояние между кластерами

- Сравним графики: 100, 1000, 10000 писем

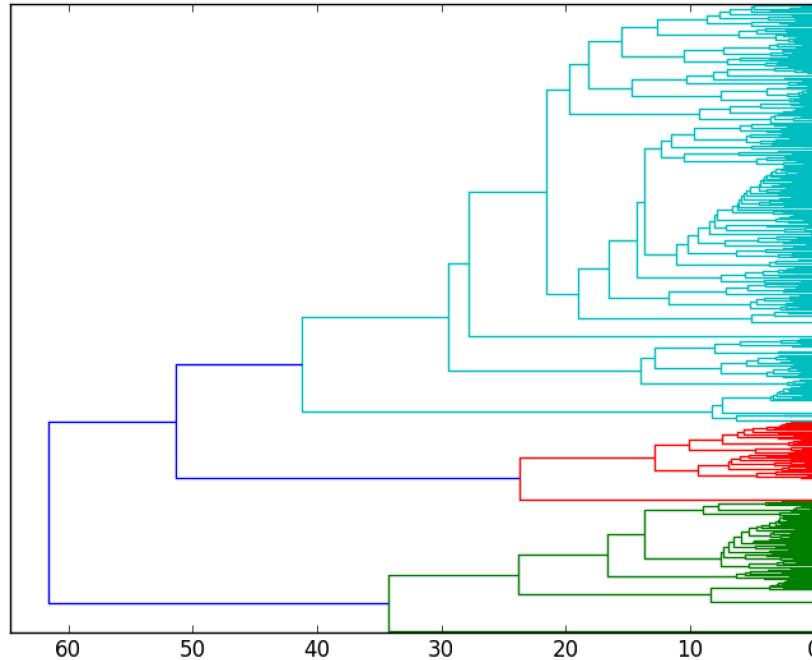


## Пример: перекос в размерах кластеров

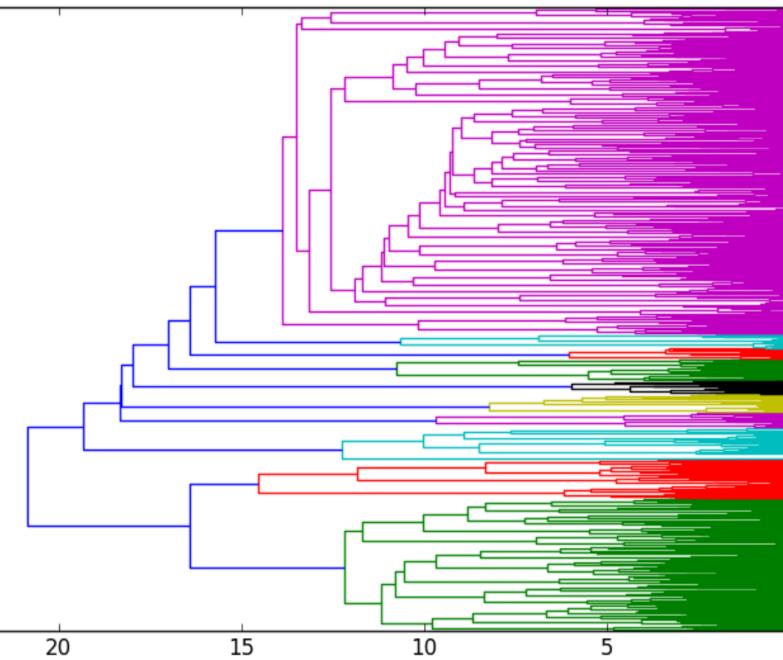
- Дендрограмма, построенная для другой выборки текстов:



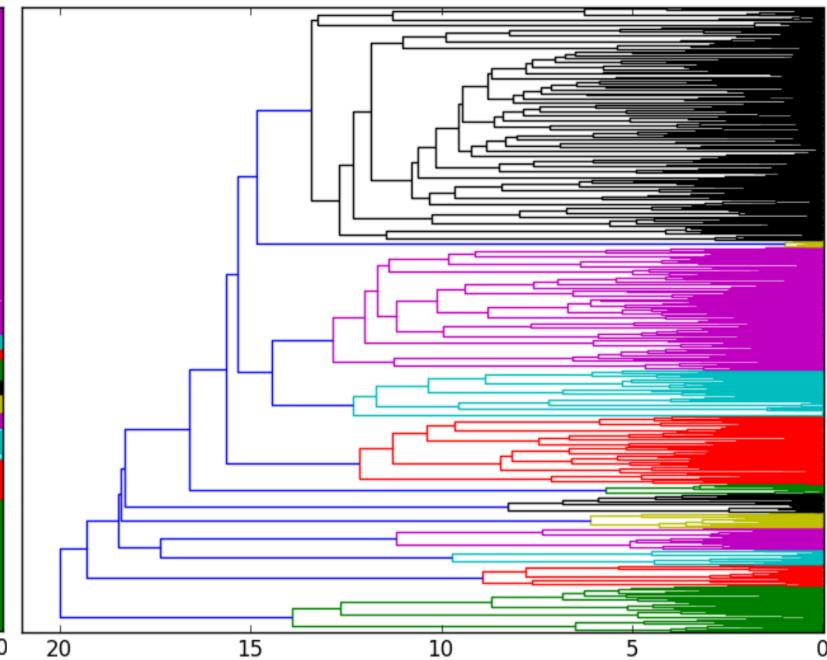
# Пример: добавляем SVD



Исходные признаки

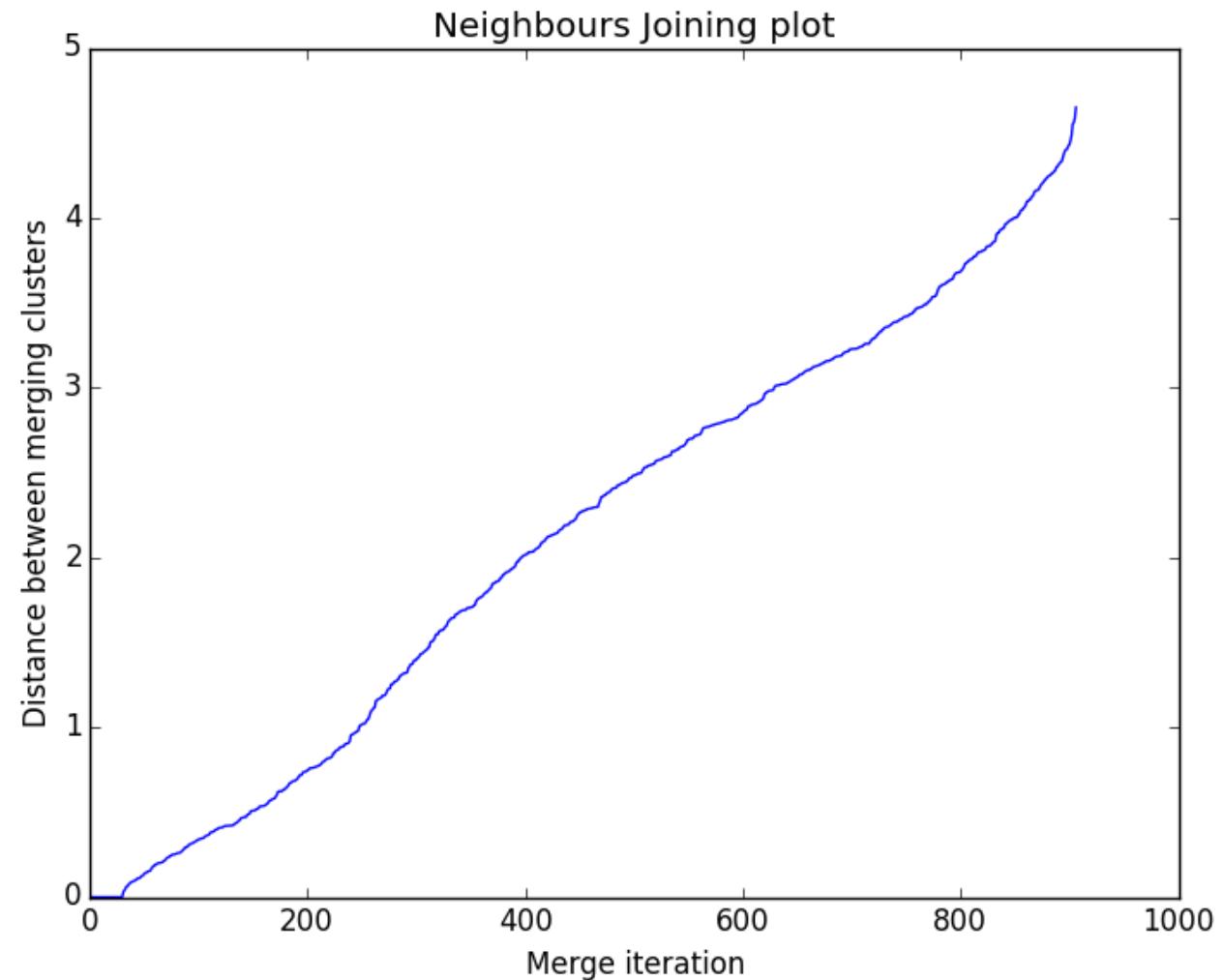


SVD



SVD (еще меньше компонент)

# Пример: SVD и расстояние при слиянии



# Резюме

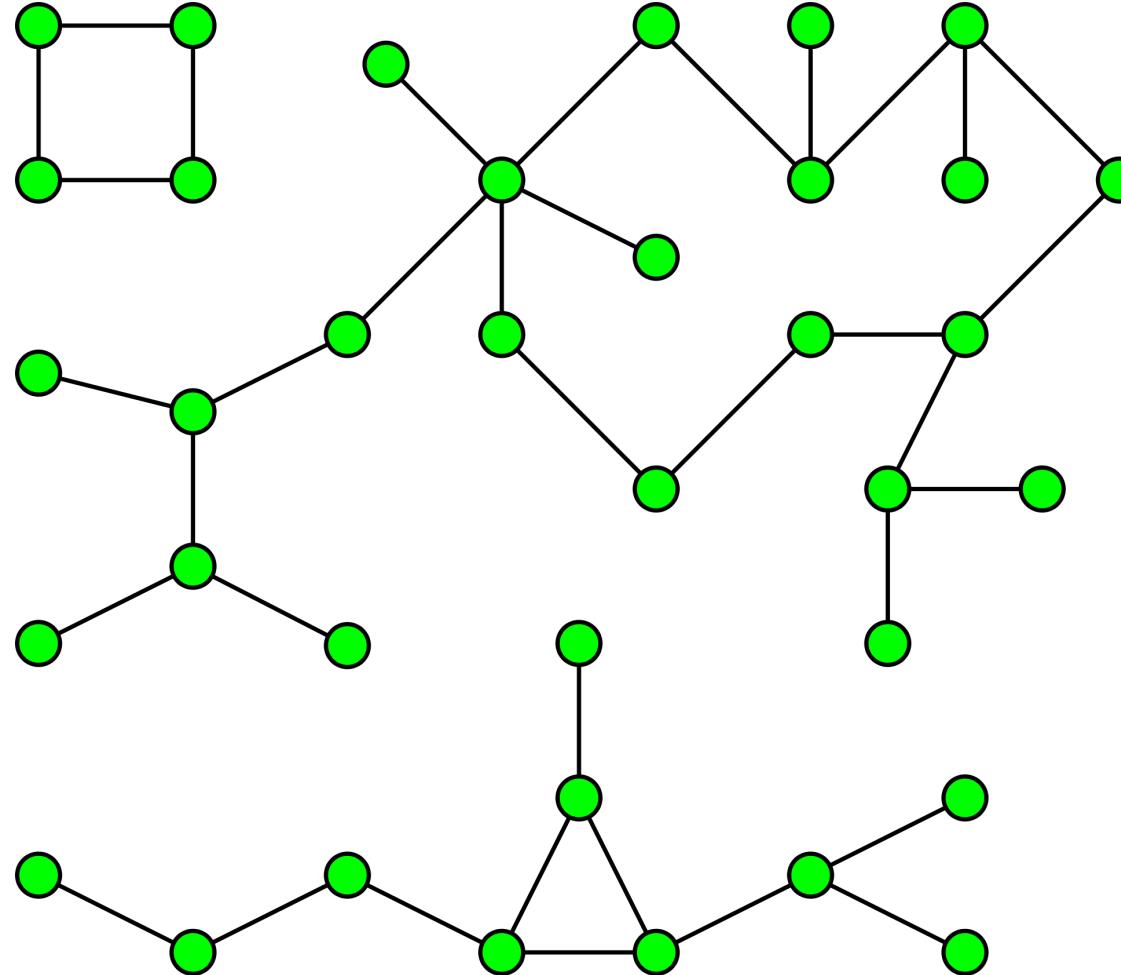
1. Иерархическая кластеризация
2. Как устроена агломеративная кластеризация
3. Расстояние между кластерами
4. Формула Ланса-Уильямса
5. Дендрограммы
6. Примеры работы

## 6. Простые графовые методы

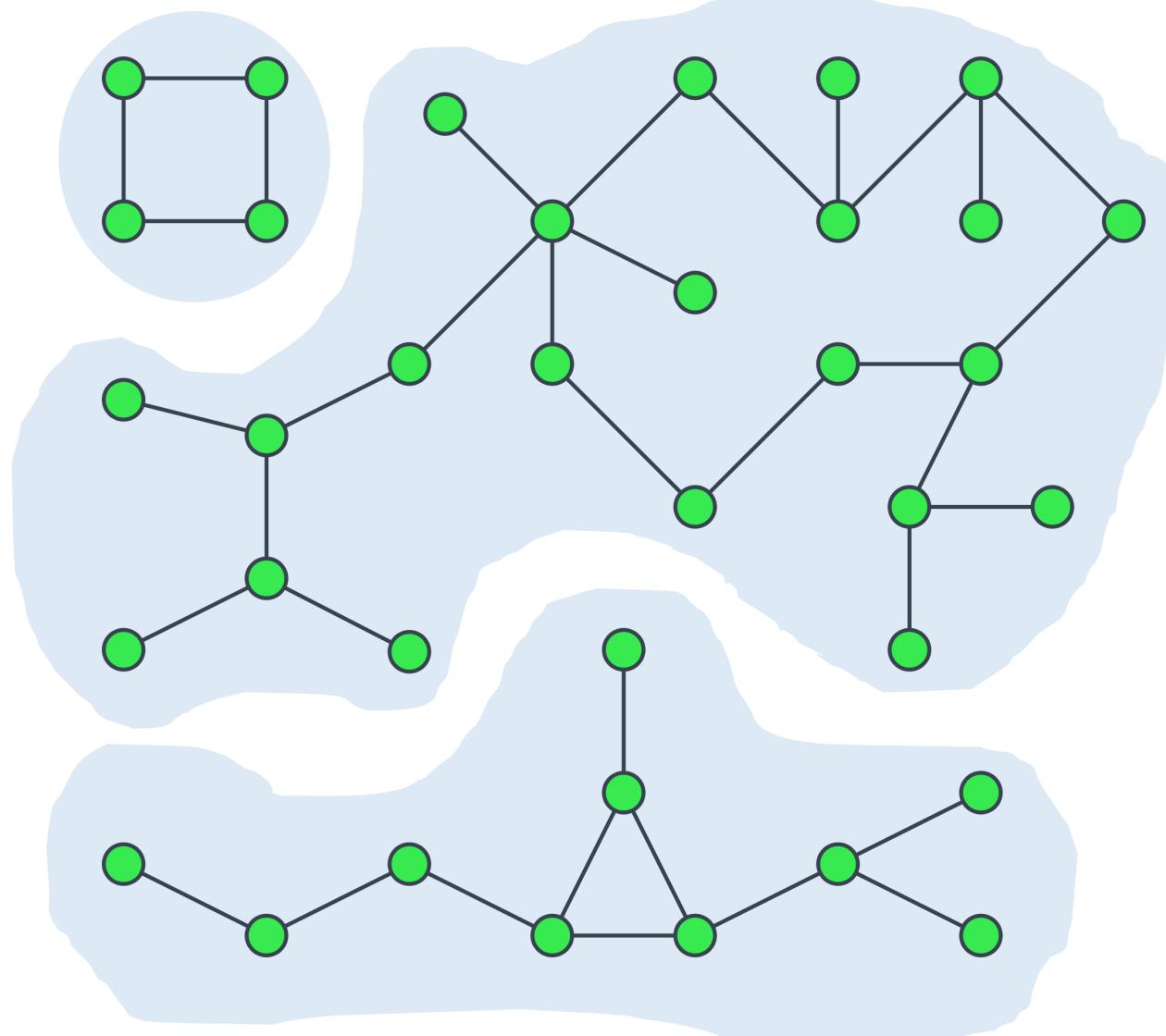
# План

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



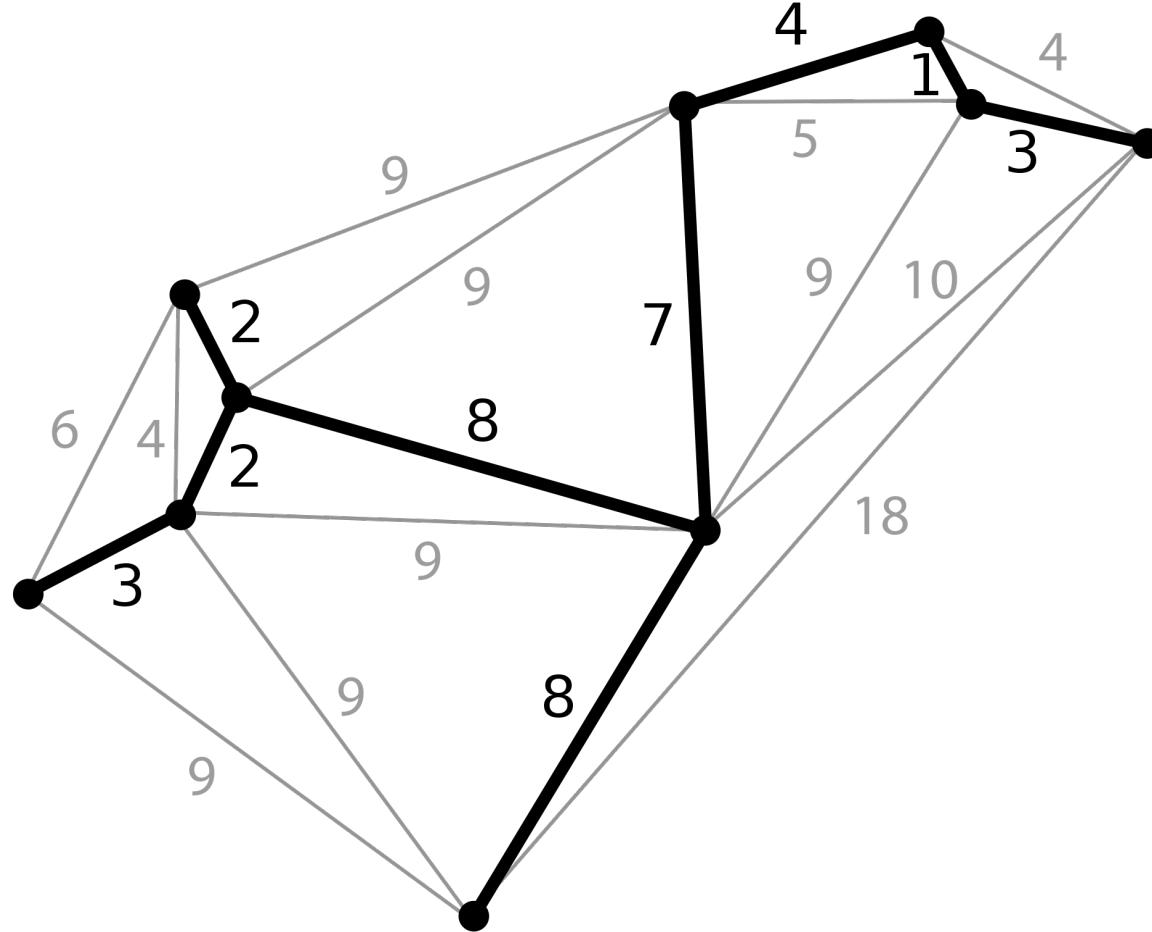
# Выделение связных компонент



## 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 оствового дерева

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

# Резюме

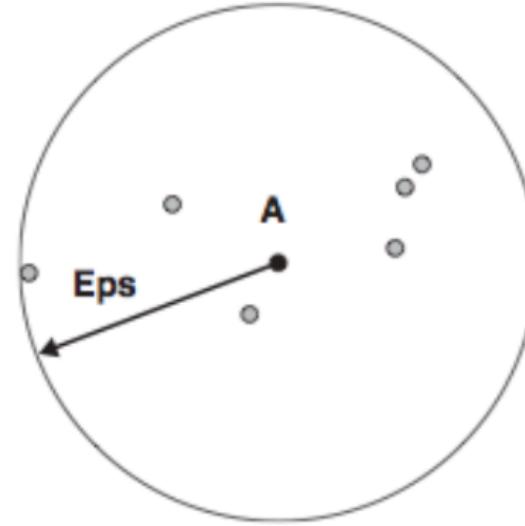
1. Связные компоненты
2. Кластеризация с помощью выделения связных компонент
3. Минимальное остовное дерево
4. Алгоритм Крускала
5. Кластеризация с помощью минимального остовного дерева

## VII. Кластеризация на основе плотности точек (density based clustering)

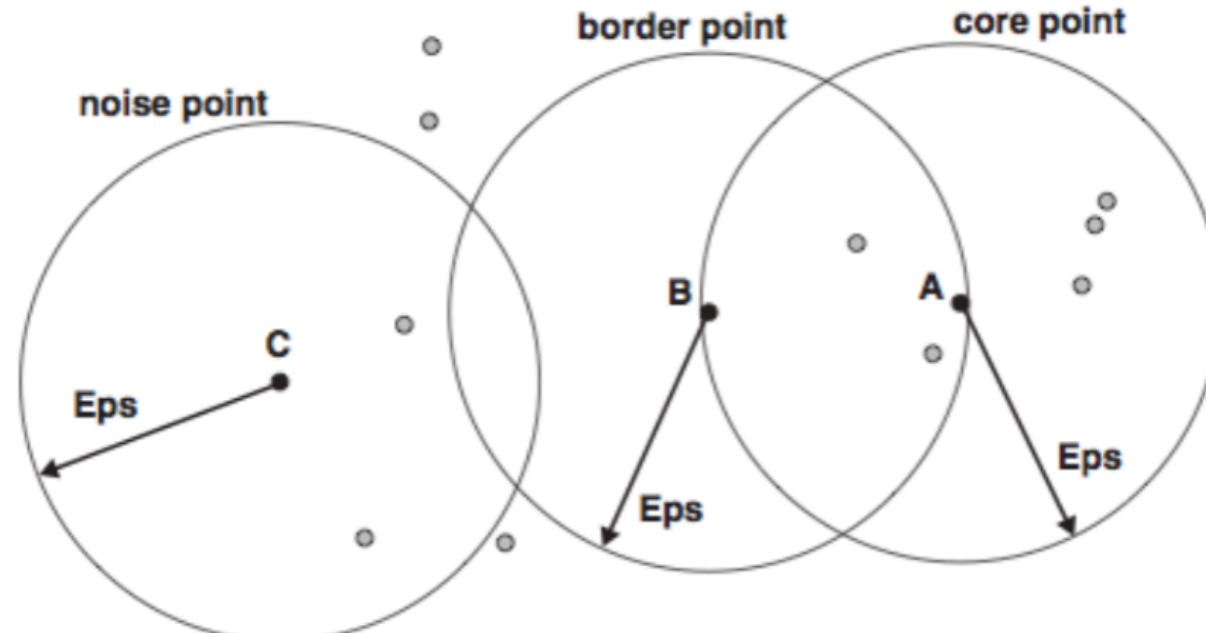
# План

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

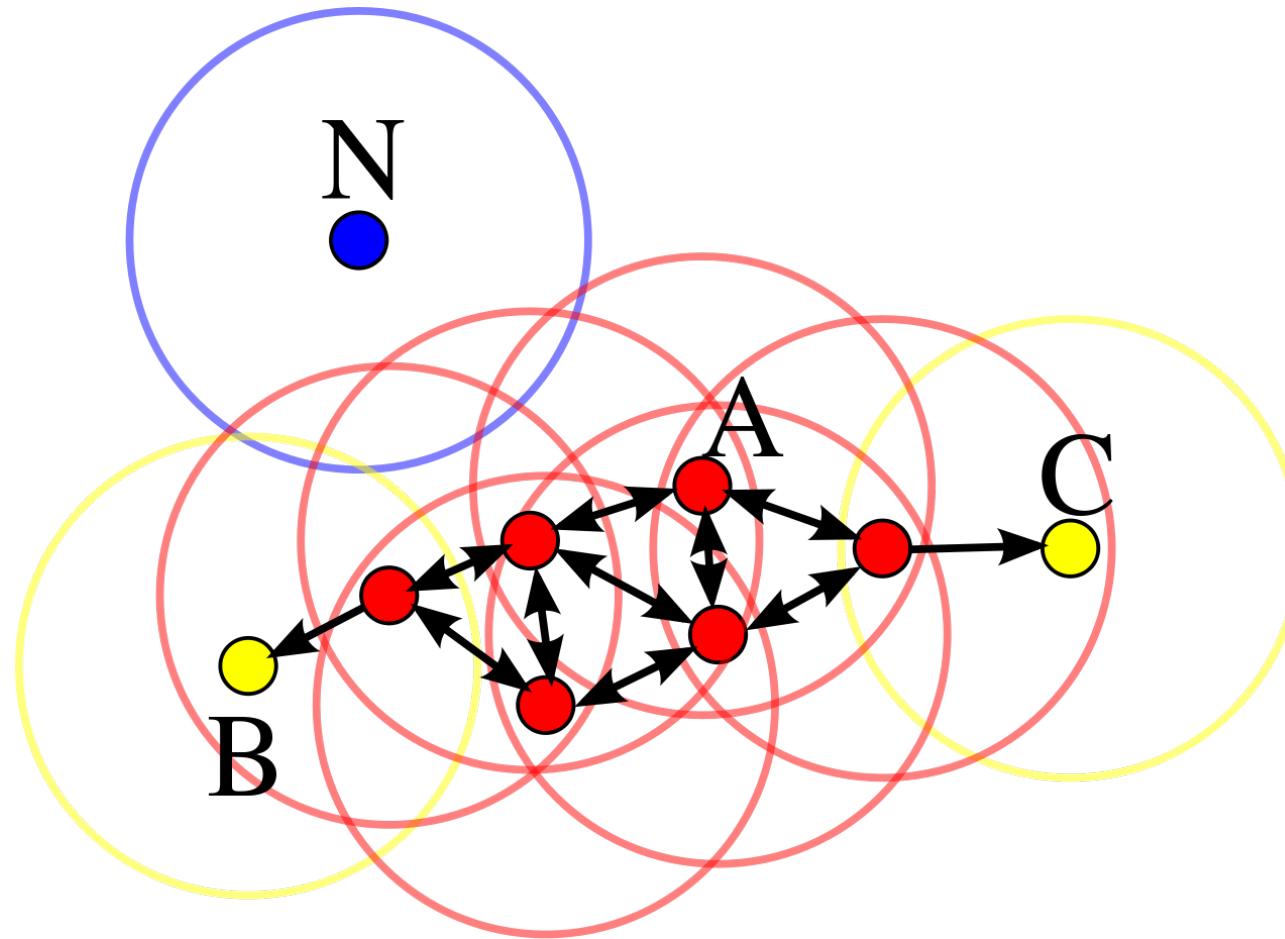


**Figure 8.20.** Center-based density.



**Figure 8.21.** Core, border, and noise points.

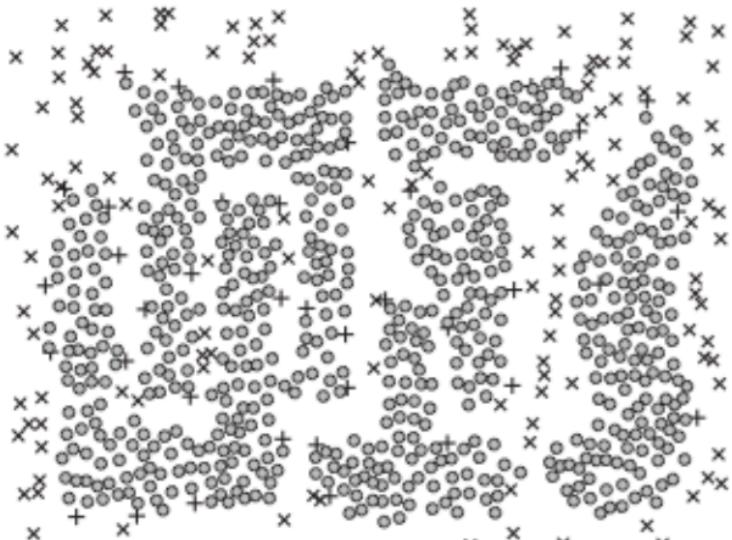
# Основные, шумовые и граничные точки



# DBSCAN



(a) Clusters found by DBSCAN.



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

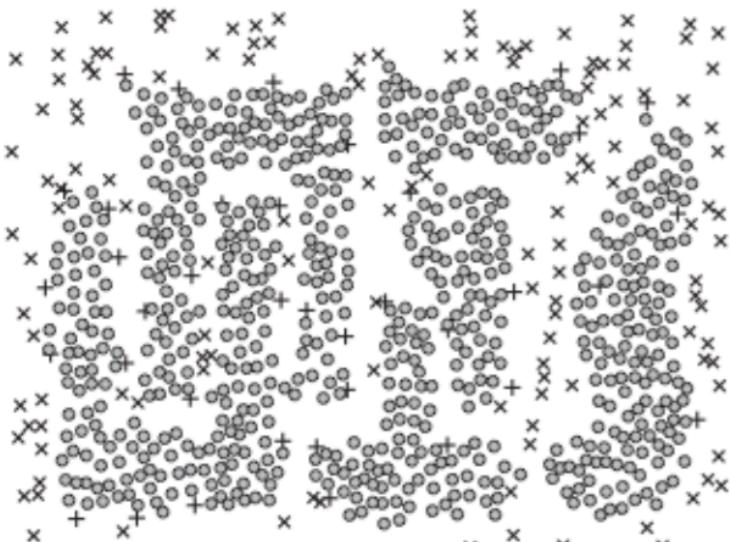
(b) Core, border, and noise points.

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.

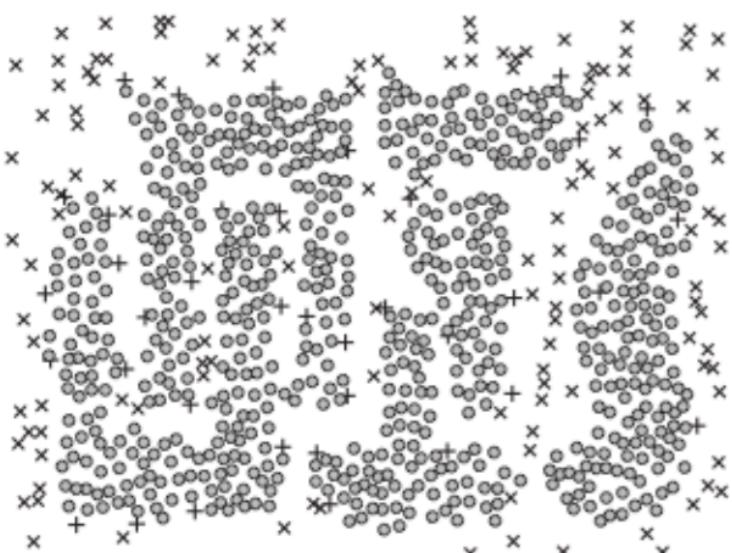
1: Пометить все точки, как основные, пограничные или шумовые.

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.

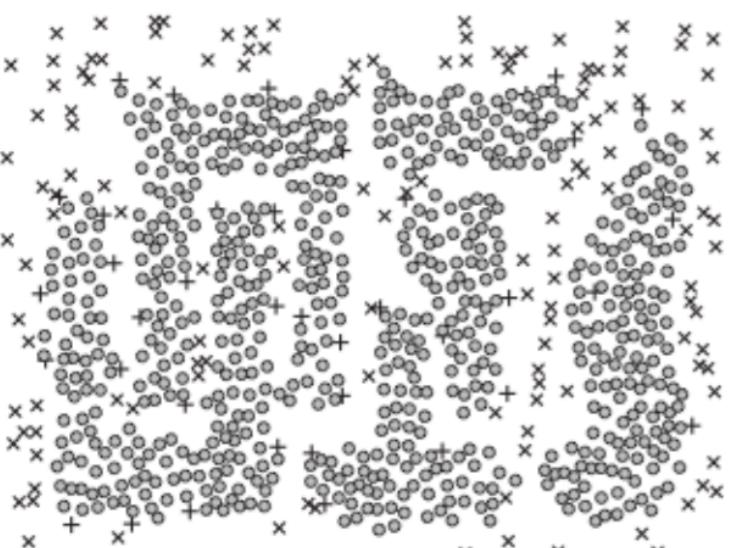
2: Отбросить точки шума.

3: Соединить все основные точки, находящиеся на расстоянии  $\text{Eps}$  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.

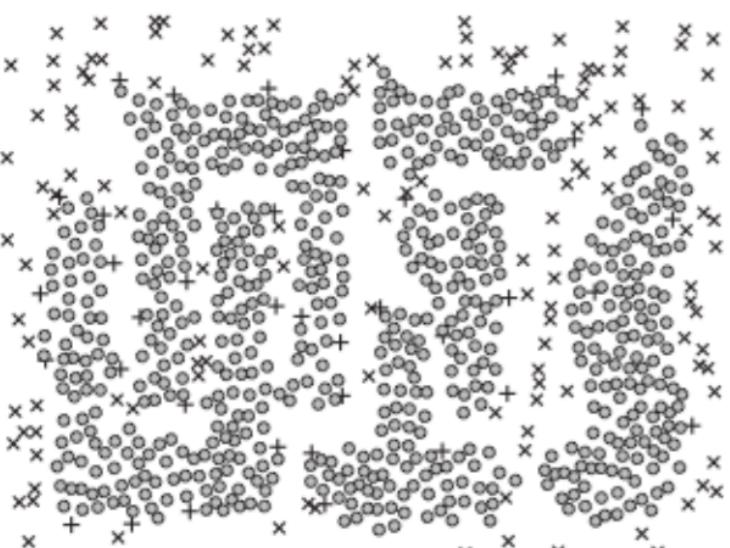
3: Соединить все основные точки, находящиеся на расстоянии  $Eps$  радиуса одна от другой.

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.

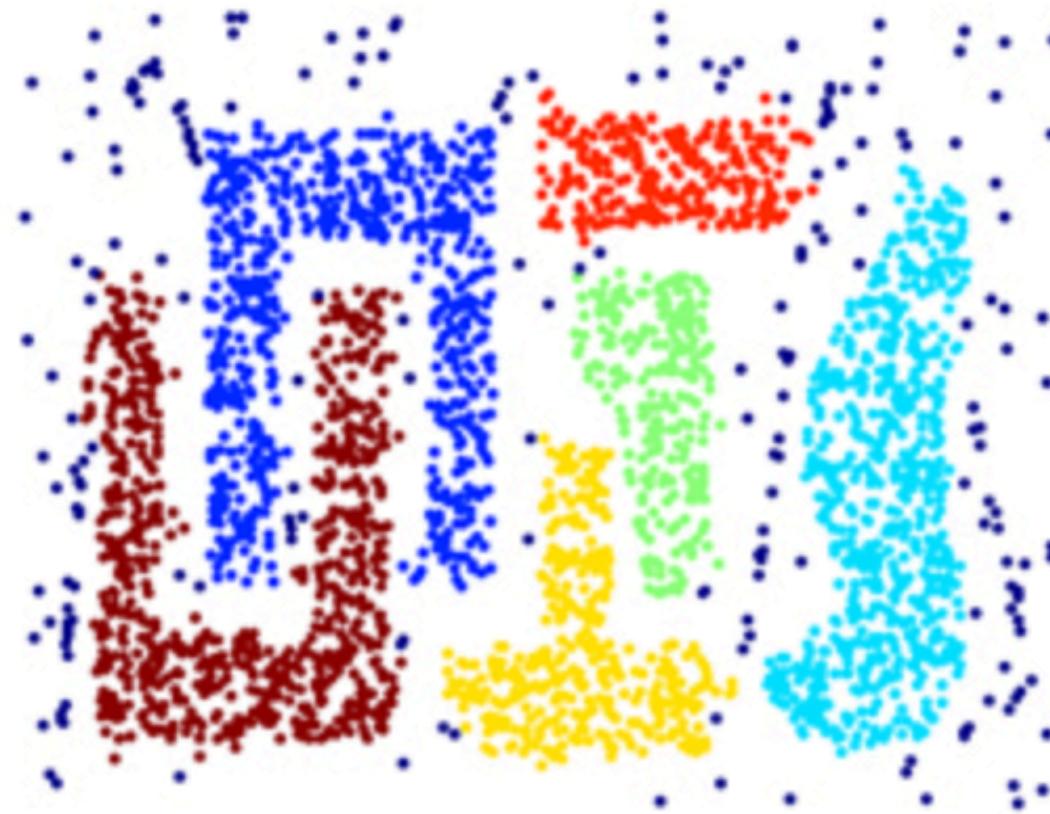
2: Отбросить точки шума.

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

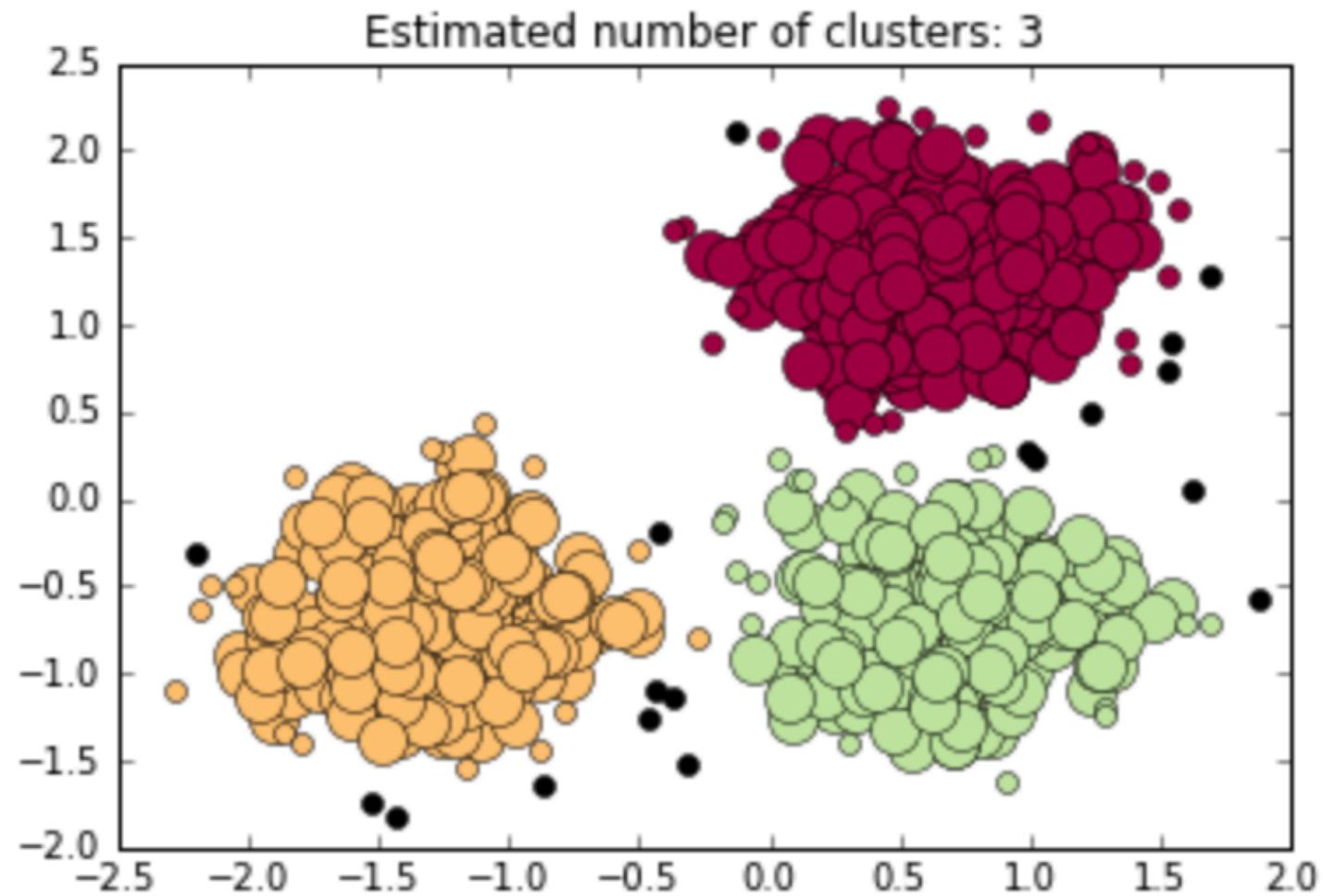
4: Объединить каждую группу соединенных основных точек в отдельный кластер.

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

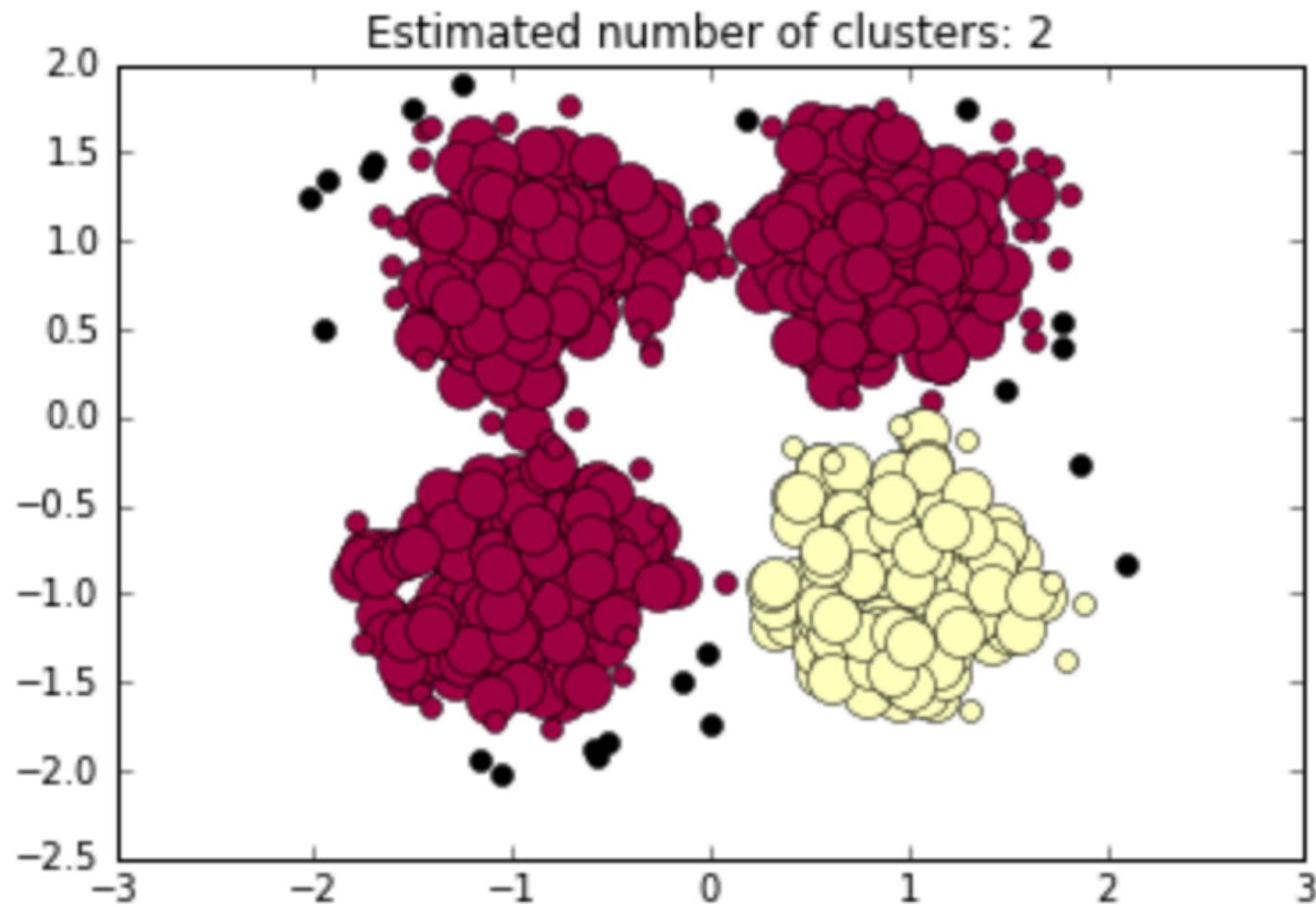
# DBSCAN: результаты работы



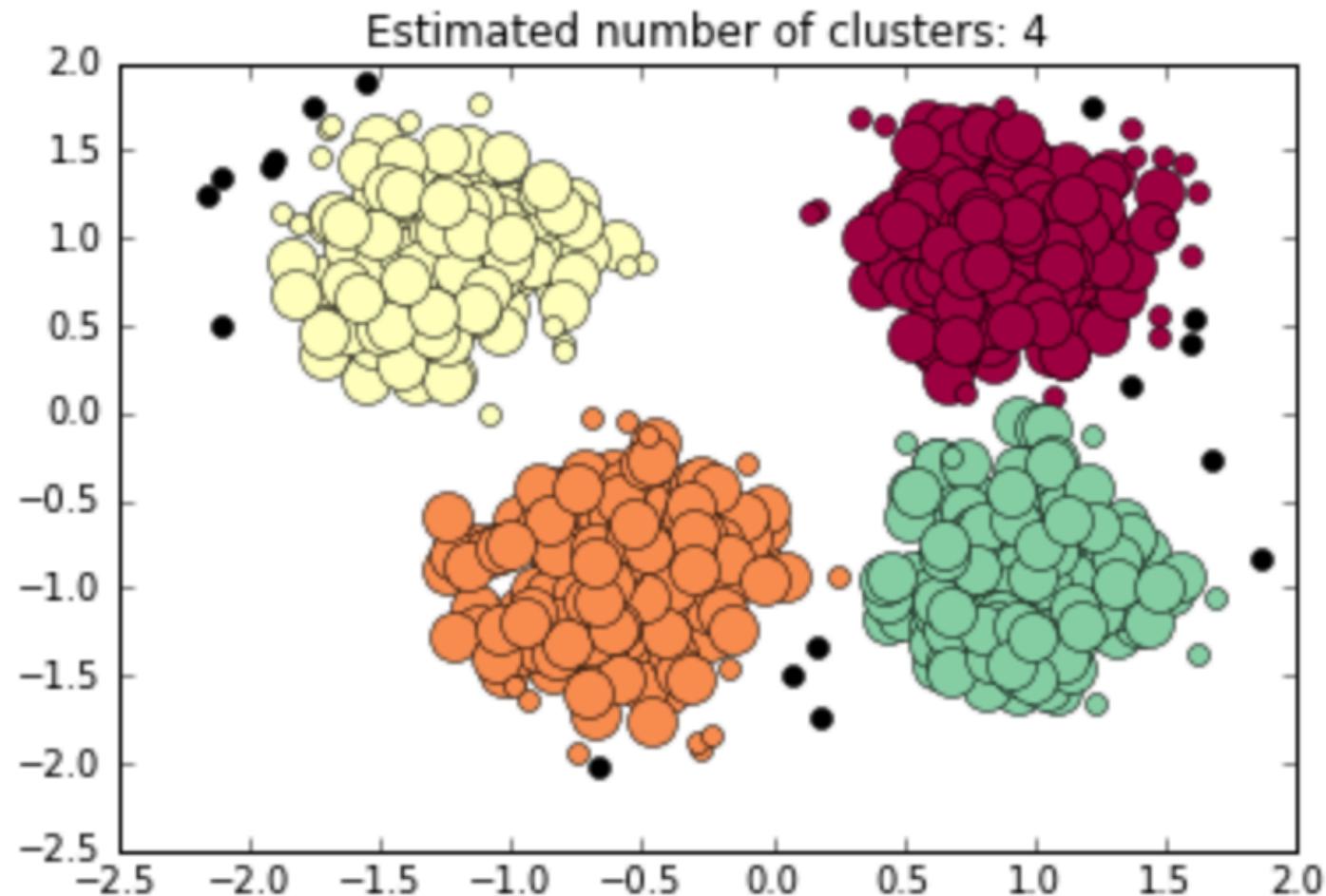
# Определение числа кластеров



# Определение числа кластеров



# Определение числа кластеров



# 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

# Рассмотрели на этой лекции:

- I. Задача кластеризации
- II. Чем могут отличаться задачи кластеризации
- III. Kmeans
- IV. EM-алгоритм
- V. Иерархическая агglomerативная кластеризация
- VI. Простые графовые методы кластеризации
- VII. Density-based методы

## В следующий раз:

- Оценка качества и подбор количества кластеров в задаче кластеризации
- Работа с признаками: извлечение, отбор, преобразование (в том числе PCA, SVD и t-SNE)

# Спасибо за внимание



[info@applieddatascience.ru](mailto:info@applieddatascience.ru)



[https://t.me/joinchat/B10lThC96v0BQCvs\\_joNew](https://t.me/joinchat/B10lThC96v0BQCvs_joNew)



[https://github.com/vkantor/ml2018jan\\_feb](https://github.com/vkantor/ml2018jan_feb)