

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:

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

РАНЕЕ: ОБУЧЕНИЕ НА РАЗМЕЧЕННЫХ ДАННЫХ (SUPERVISED LEARNING)

- » Обучающая выборка:
 x_1, \dots, x_ℓ — объекты
 y_1, \dots, y_ℓ — ответы
- » Тестовая выборка:
 $x_{\ell+1}, \dots, x_{\ell+u}$
- » В регрессии: y_i — прогнозируемая величина
В классификации: y_i — метка класса

ВОССТАНОВЛЕНИЕ ОТОБРАЖЕНИЙ

- Считаем, что есть отображение:

$$\textcolor{red}{x} \mapsto y$$

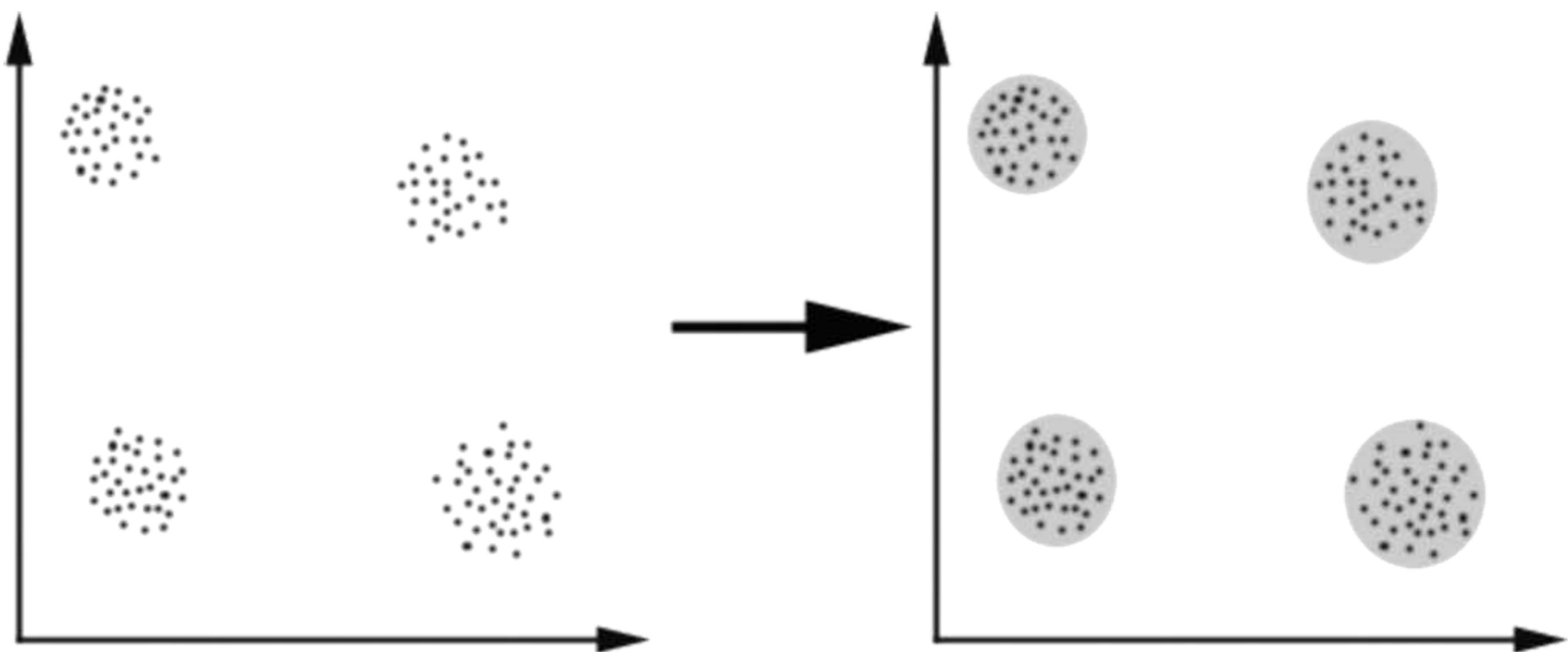
- Обучающая выборка — это примеры значений, по которым мы пытаемся построить $\textcolor{red}{a(x)}$:

$$a(x) \approx y$$

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:

$$\textcolor{red}{x} \mapsto y$$

- Пытаемся построить $a(\textcolor{red}{x})$, но примеров $\textcolor{red}{y}$ 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 МЕЖКЛАСТЕРНОЕ 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

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

$$\frac{F_0}{F_1} \rightarrow \min$$

РЕЗЮМЕ

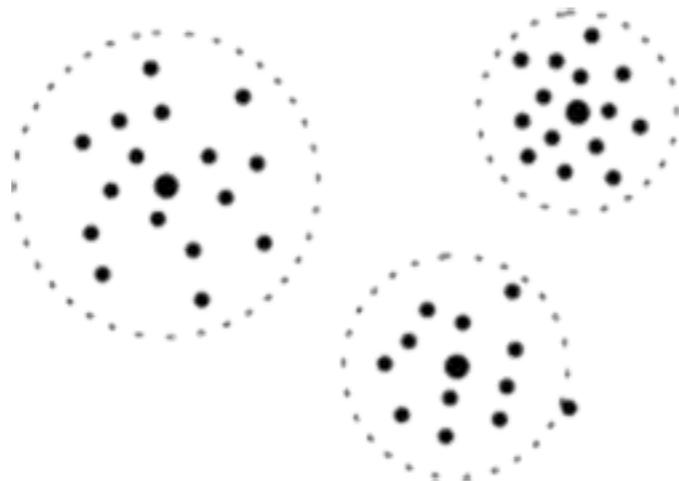
- Отличия от обучения на размеченных данных
- Постановка задачи кластеризации
- Простые способы оценить качество кластеризации
- В следующем видео: о том, какими бывают задачи кластеризации

ПРИМЕРЫ ЗАДАЧ КЛАСТЕРИЗАЦИИ

ЗАЧЕМ НУЖНЫ РАЗНЫЕ АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ

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

ФОРМА КЛАСТЕРОВ



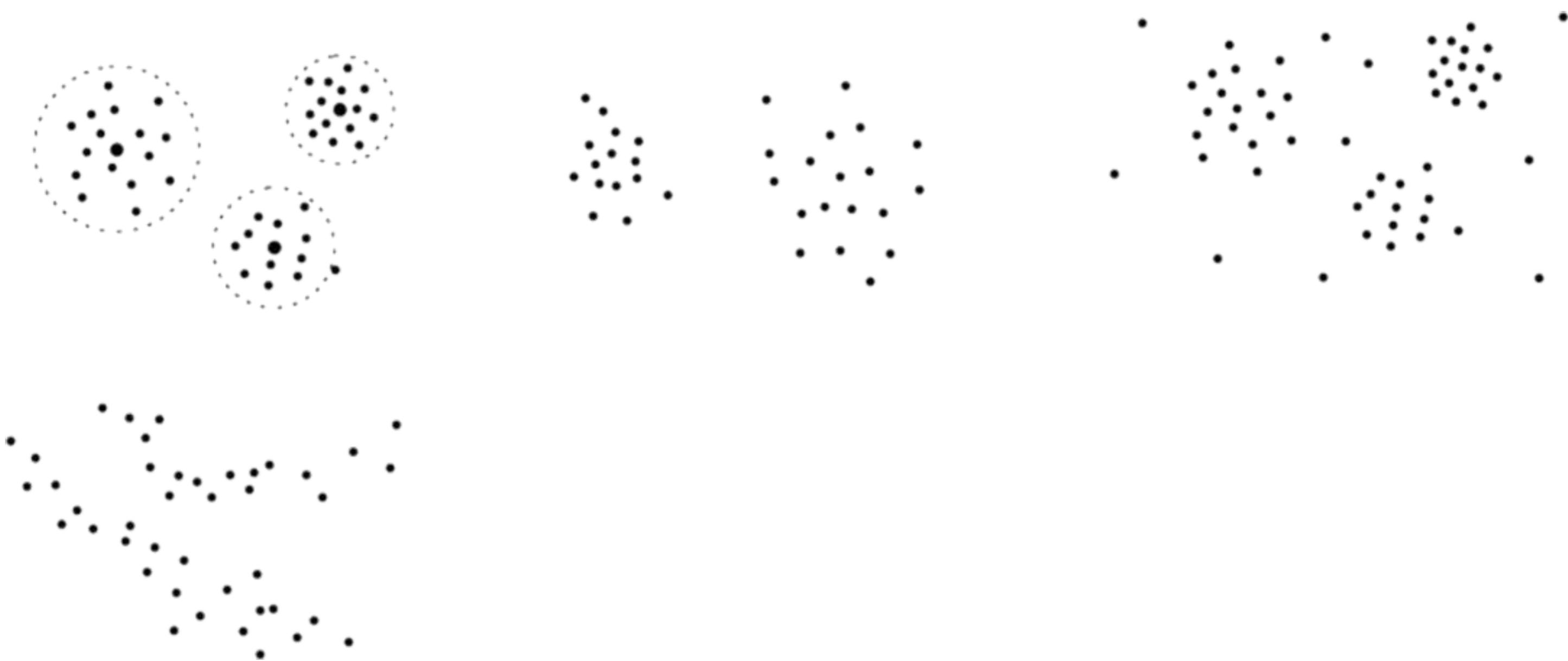
ФОРМА КЛАСТЕРОВ



ФОРМА КЛАСТЕРОВ



ФОРМА КЛАСТЕРОВ



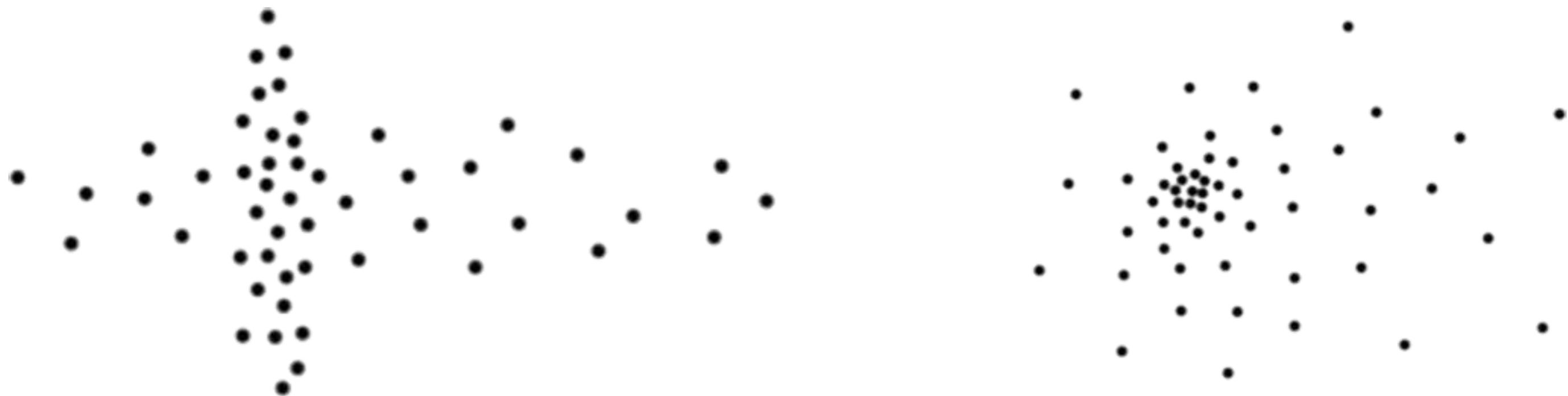
ФОРМА КЛАСТЕРОВ



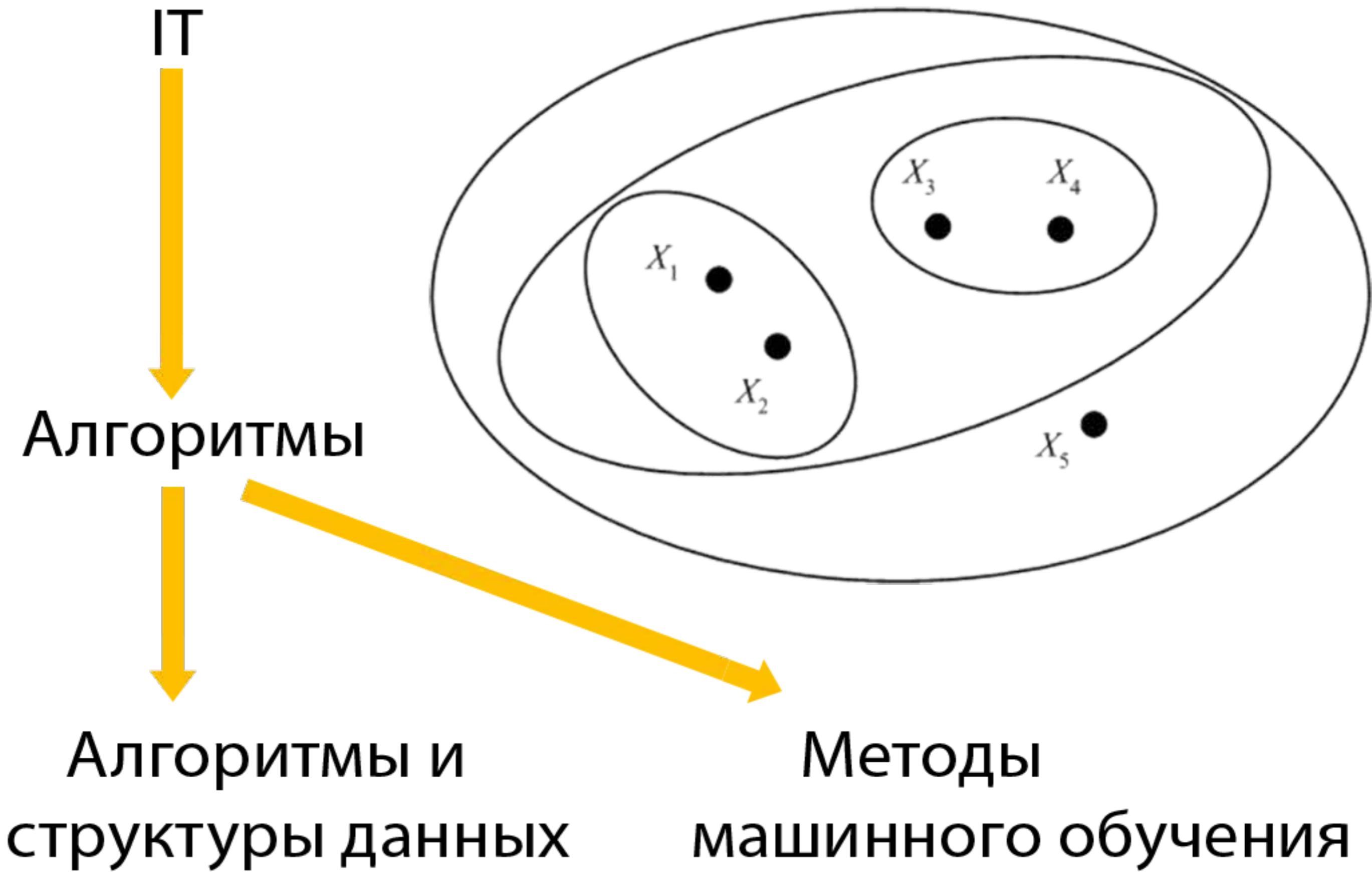
ФОРМА КЛАСТЕРОВ



ФОРМА КЛАСТЕРОВ



ФОРМА КЛАСТЕРОВ



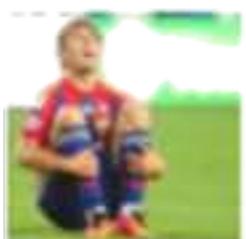
РАЗМЕР КЛАСТЕРОВ

- 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

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

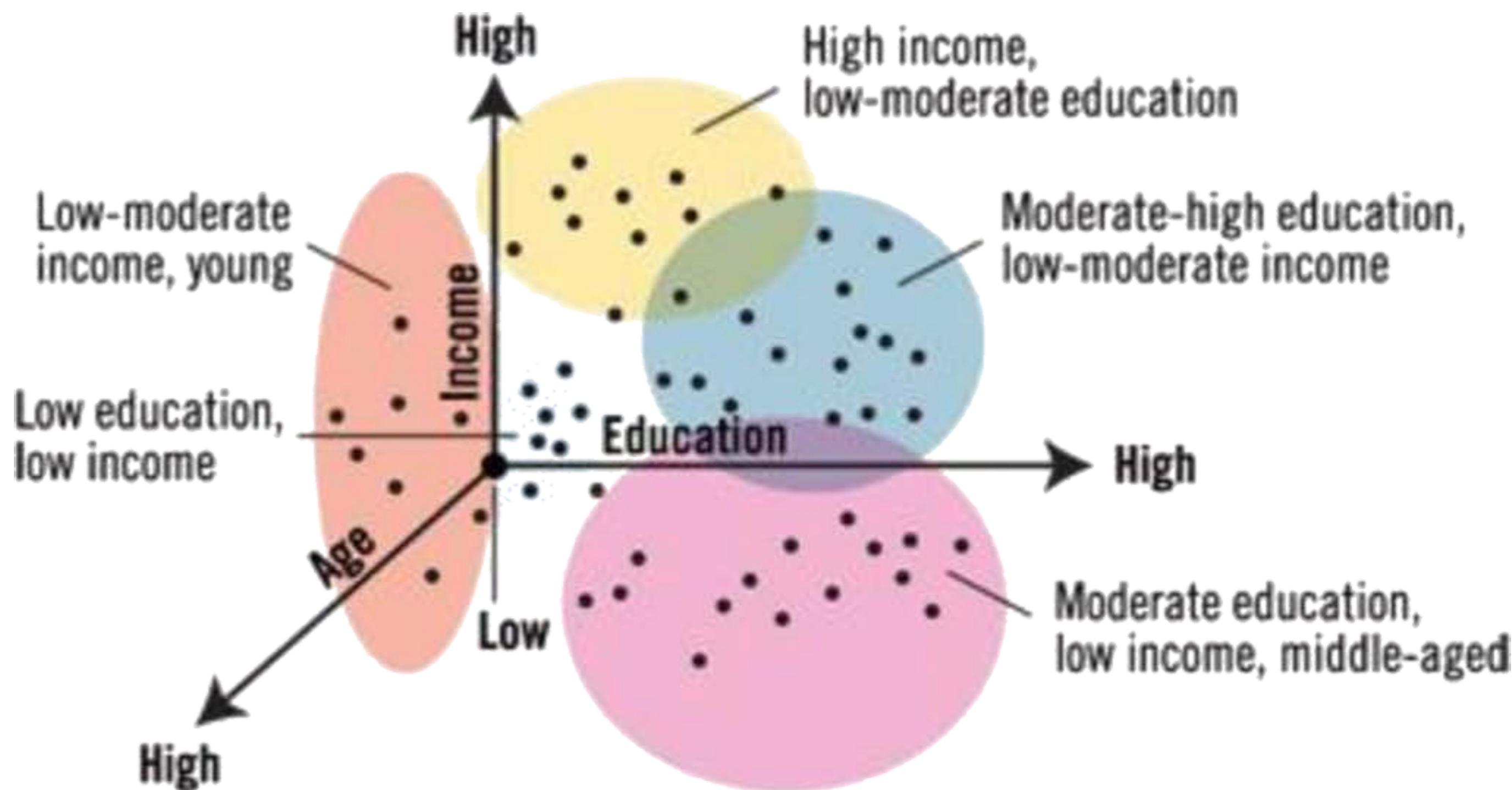
Открытие Олимпиады в Сочи
посмотрели несколько миллиардов
человек

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

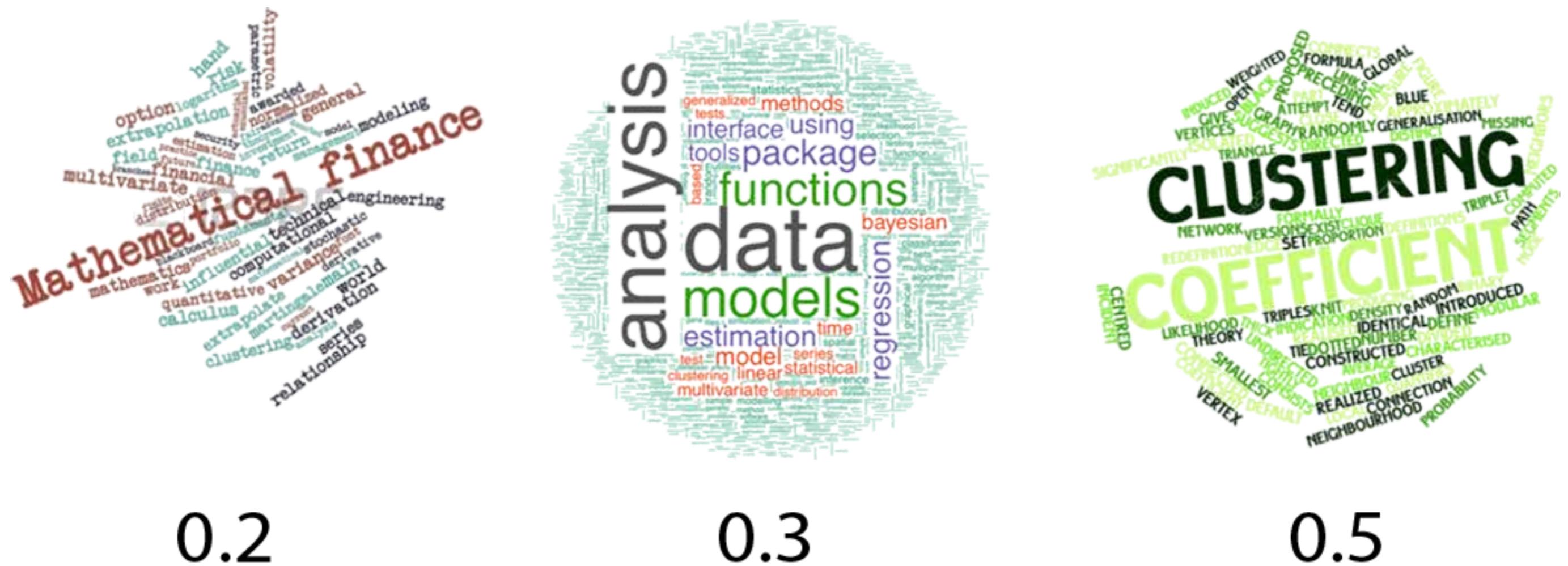
«ЖЁСТКАЯ» И «МЯГКАЯ» КЛАСТЕРИЗАЦИИ

- ## › Кластеризация для выделения «тем»



«ЖЁСТКАЯ» И «МЯГКАЯ» КЛАСТЕРИЗАЦИИ

- ## › Кластеризация для выделения «тем»



РЕЗЮМЕ: ЧЕМ МОГУТ ОТЛИЧАТЬСЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ

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

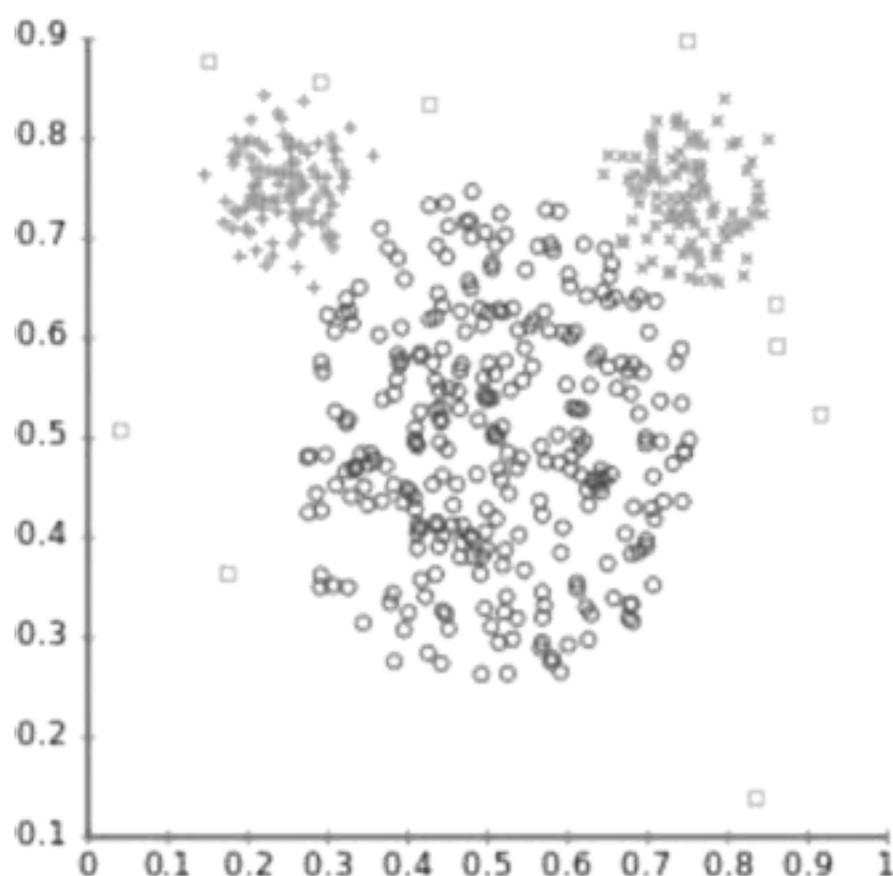
ЗНАКОМСТВО С МЕТОДАМИ КЛАСТЕРИЗАЦИИ

КАК МОГУТ ВЫГЛЯДЕТЬ КЛАСТЕРЫ

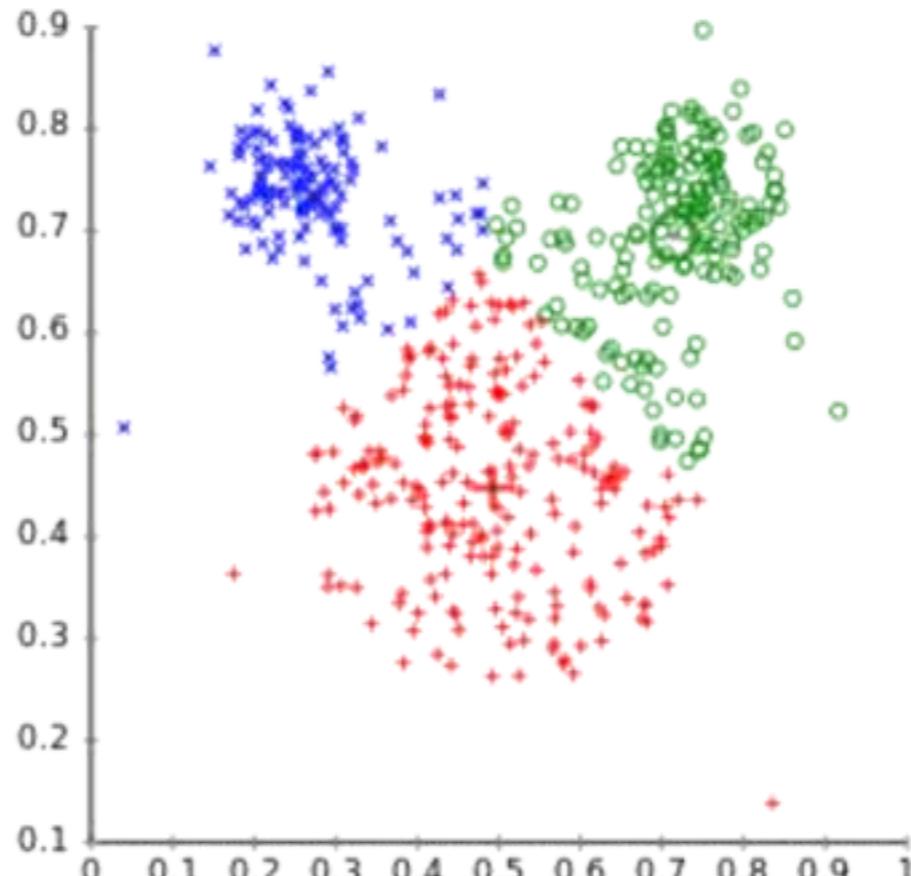


Универсального метода кластеризации **нет**

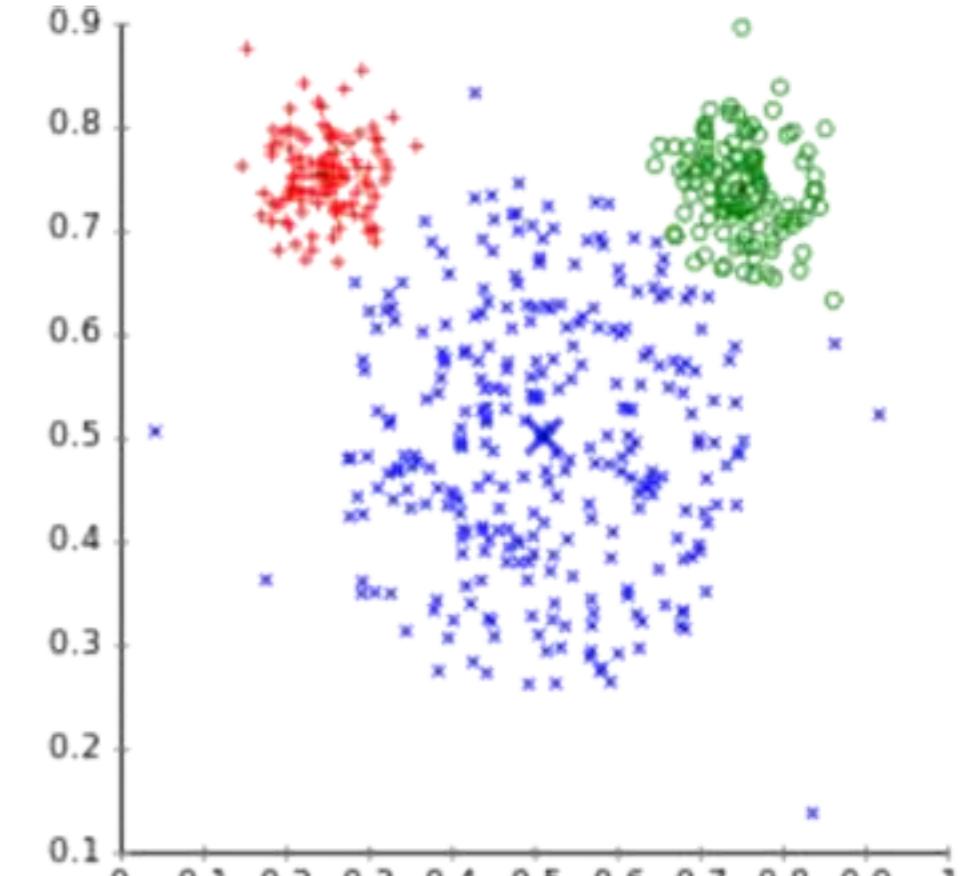
РАЗЛИЧИЯ В РЕЗУЛЬТАТАХ РАБОТЫ



Исходная выборка
("Mouse" dataset)



Метод k средних
(K-Means)

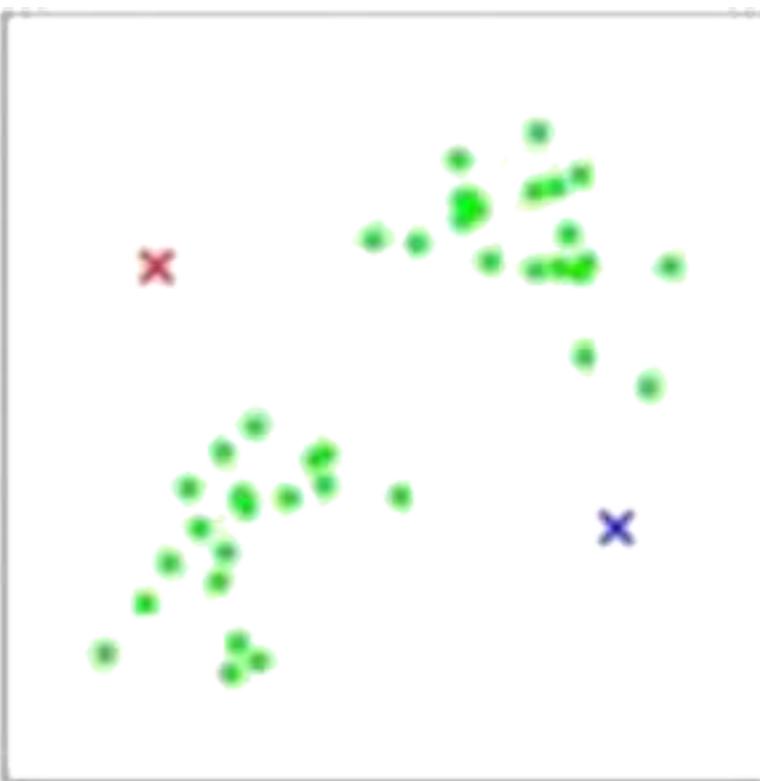


EM-алгоритм

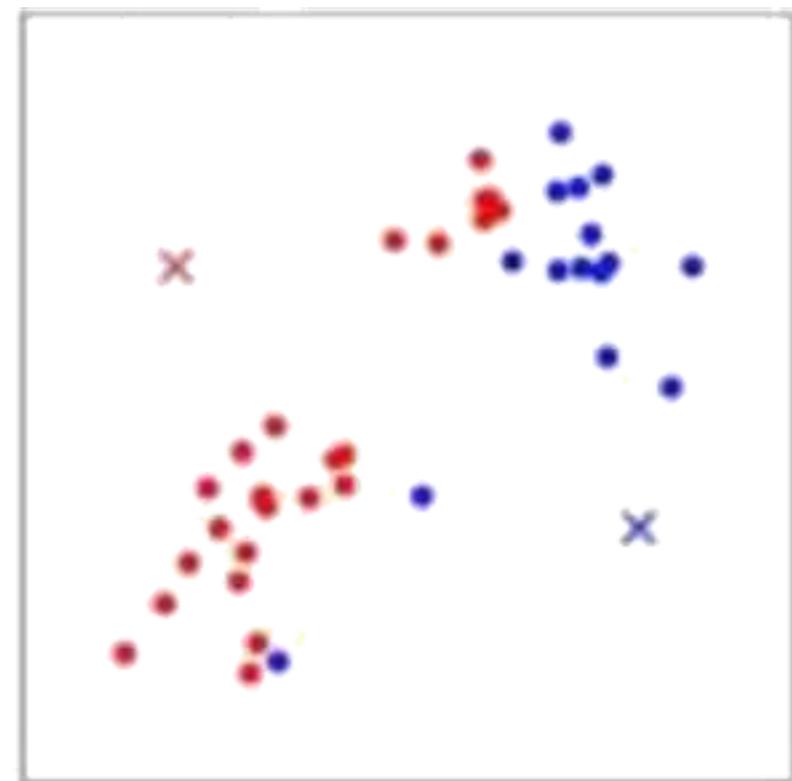
ПРОСТОЙ МЕТОД: k-Means



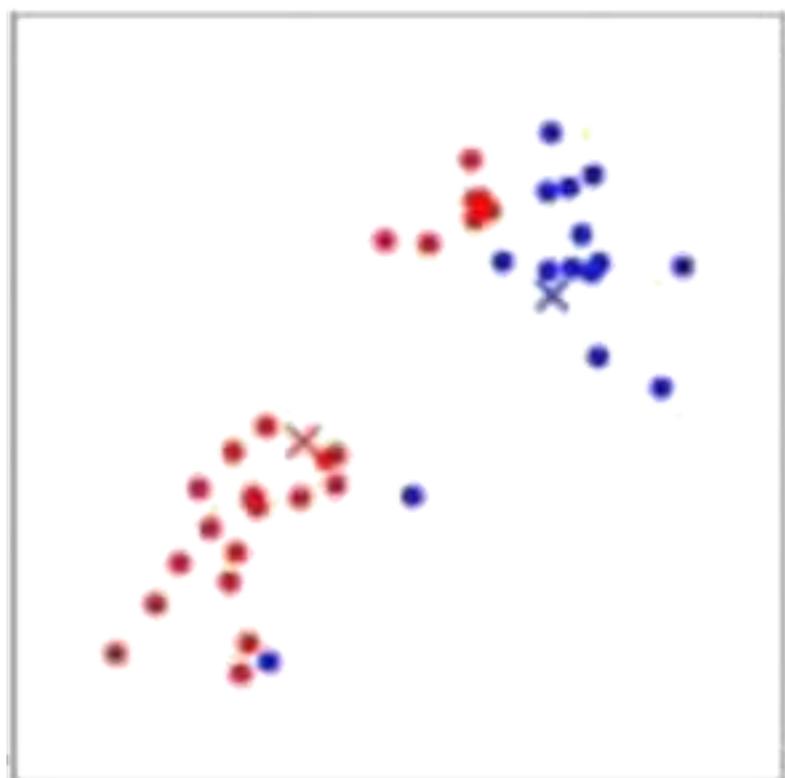
(a)



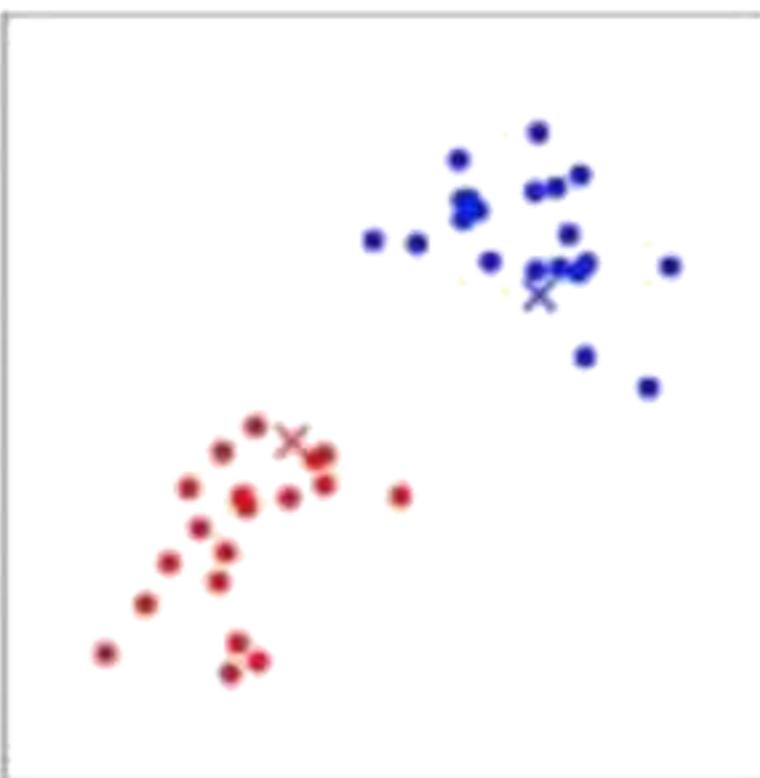
(b)



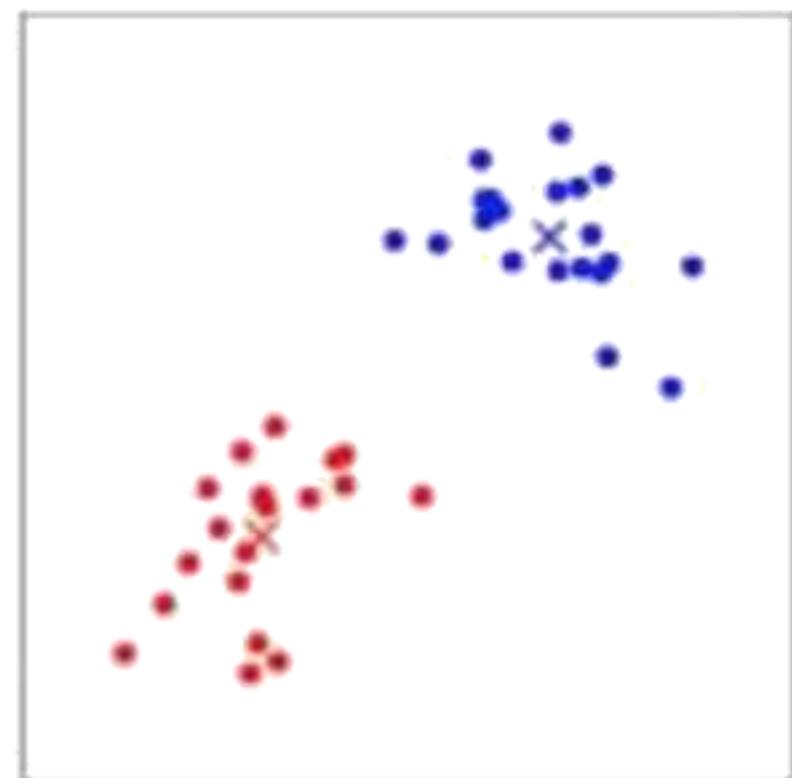
(c)



(d)

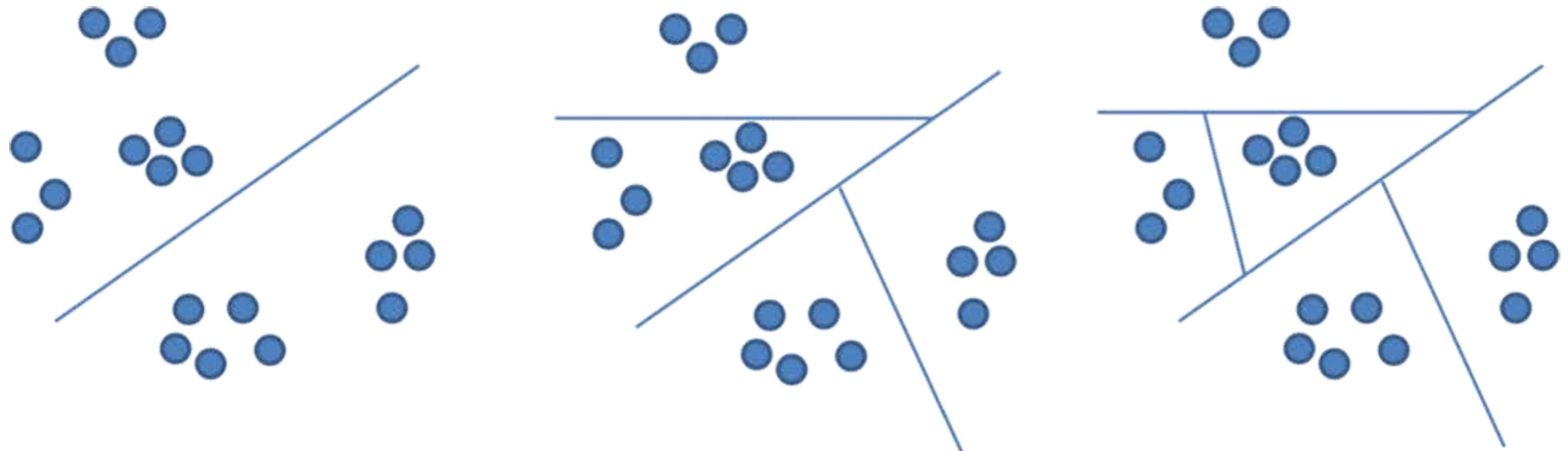


(e)



(f)

ПОДБОР ЧИСЛА КЛАСТЕРОВ: BisectMeans



БОЛЕЕ СЛОЖНЫЙ МЕТОД: ЕМ

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

› Е-шаг:

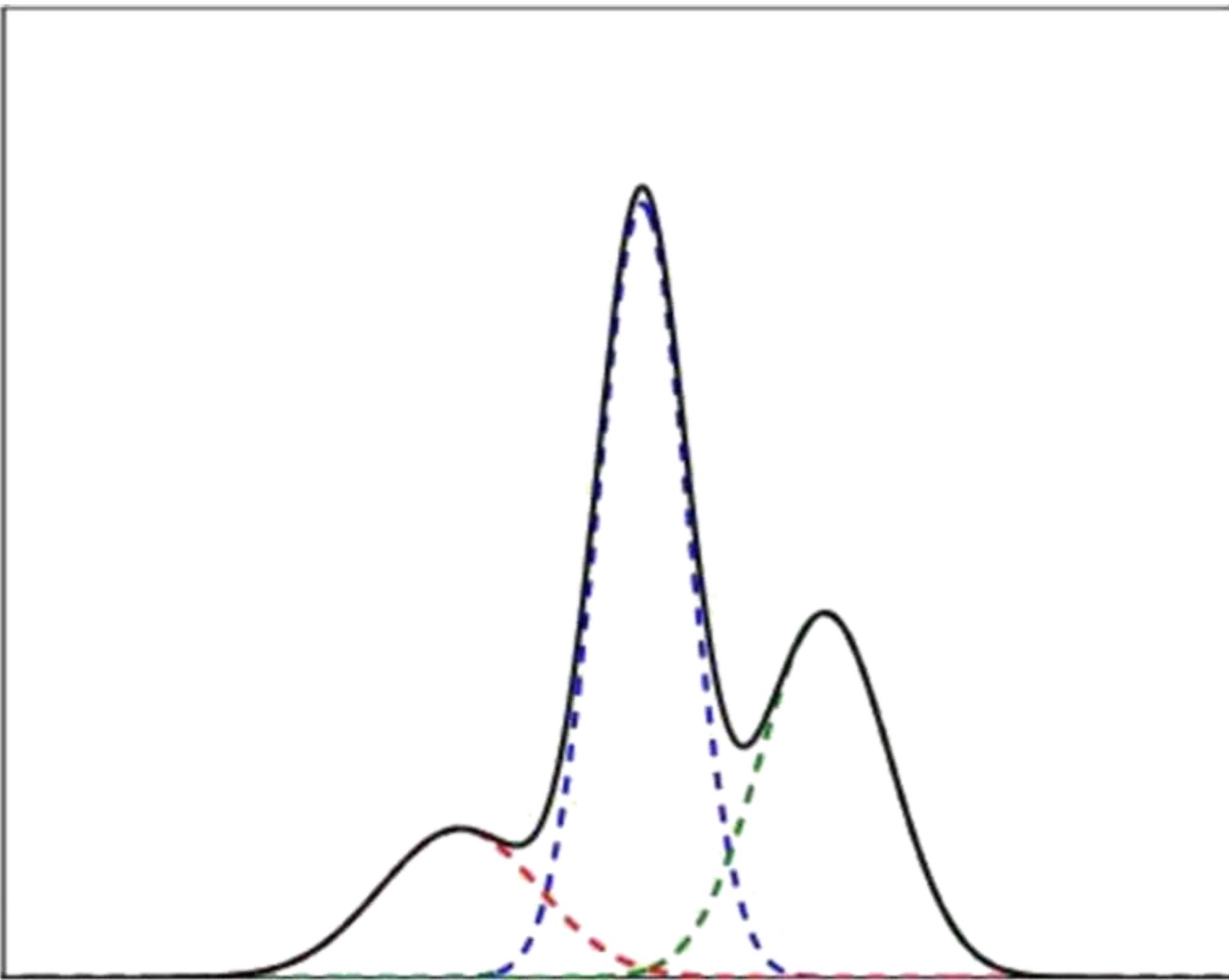
$$g_{ji} = p(j|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$$

› М-шаг:

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

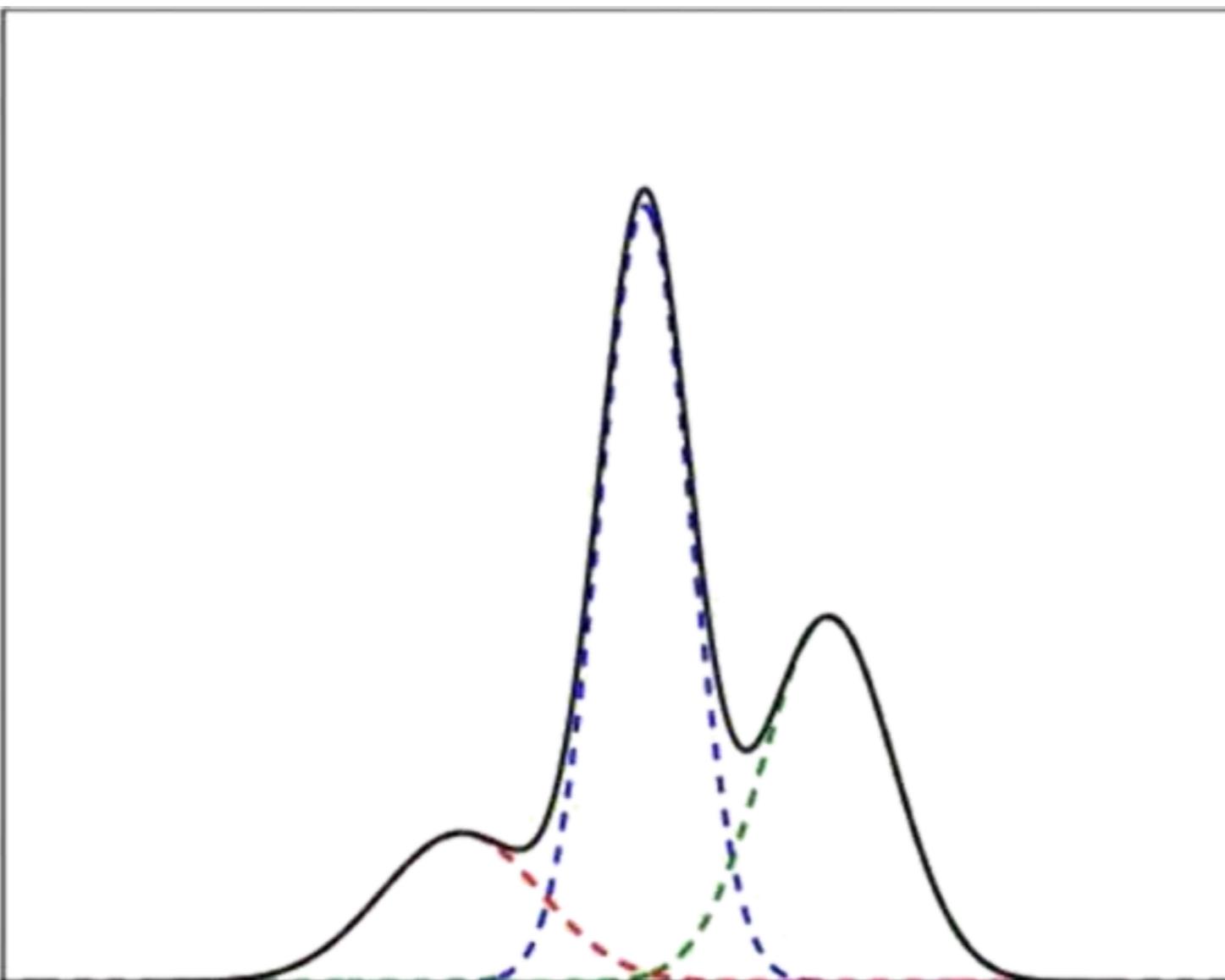
$$\theta_j = \operatorname{argmax}_{\theta} \sum_{i=1}^N g_{ji} \ln \varphi(\theta; x)$$

ЗАДАЧА РАЗДЕЛЕНИЯ СМЕСИ РАСПРЕДЕЛЕНИЙ



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

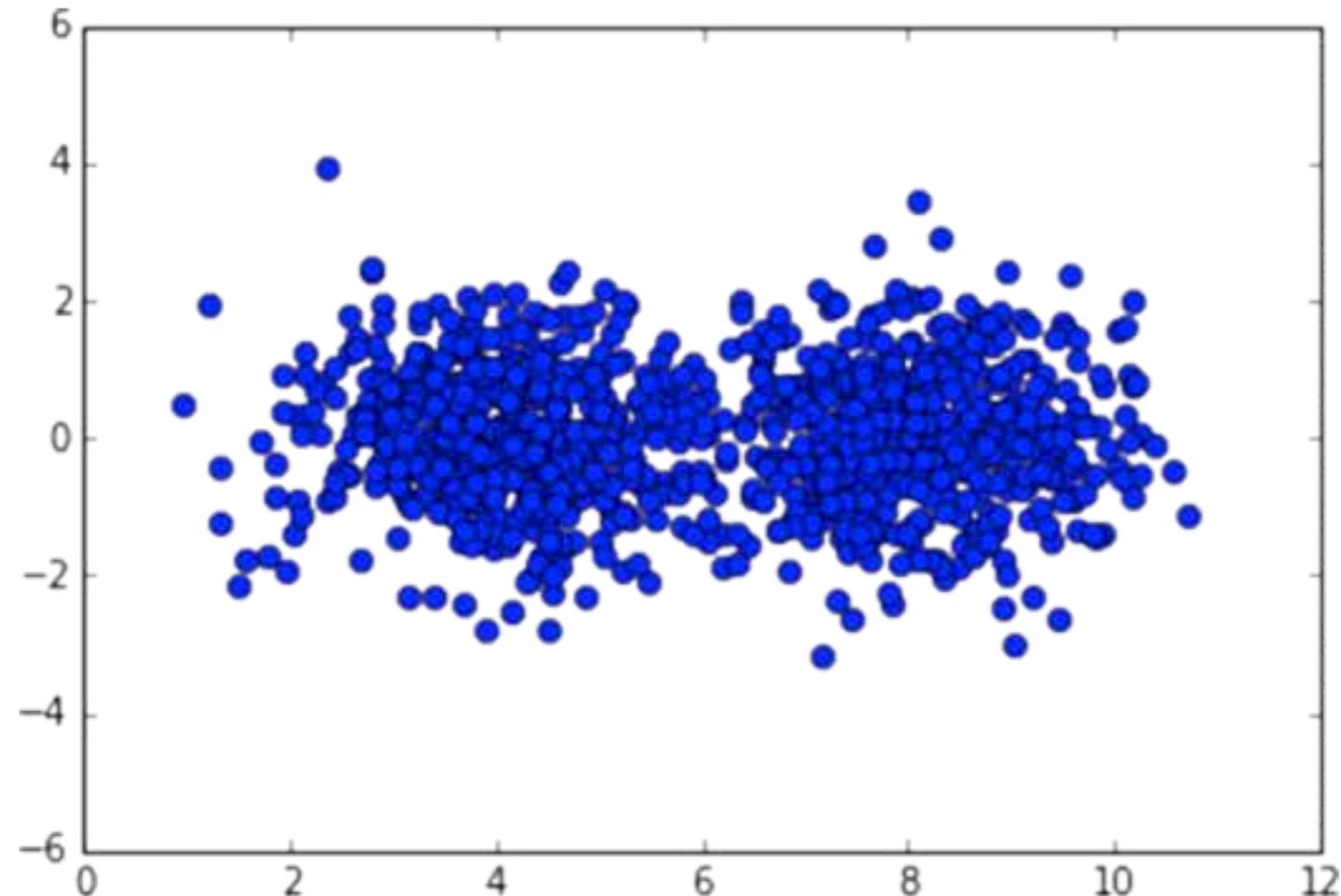
ЗАДАЧА РАЗДЕЛЕНИЯ СМЕСИ РАСПРЕДЕЛЕНИЙ



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

$$\mathbf{w}, \boldsymbol{\theta} = \operatorname{argmax}_{\boldsymbol{\theta}, \mathbf{w}} \sum_{j=1}^K \ln p(x_i)$$

ВЫБОРКА ИЗ СМЕСИ ГАУССОВСКИХ РАСПРЕДЕЛЕНИЙ

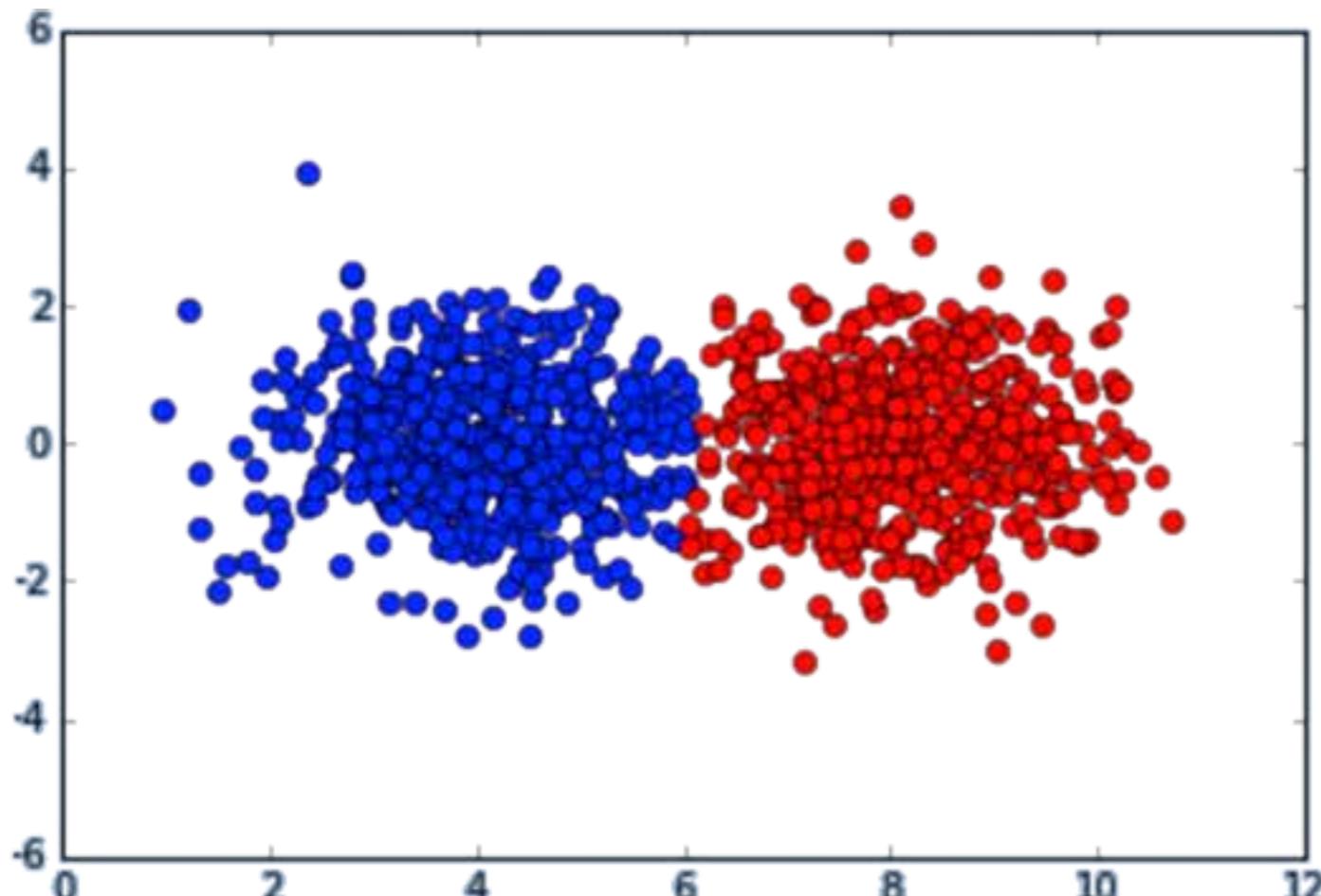


Точки сгенерированы из смеси:

$$p(x) = \frac{1}{2}\mathcal{N}\left(\begin{pmatrix} 4 \\ 0 \end{pmatrix}, 1\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix} 8 \\ 0 \end{pmatrix}, 1\right)$$

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

КЛАСТЕРИЗАЦИЯ ЕМ-АЛГОРИТМА



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

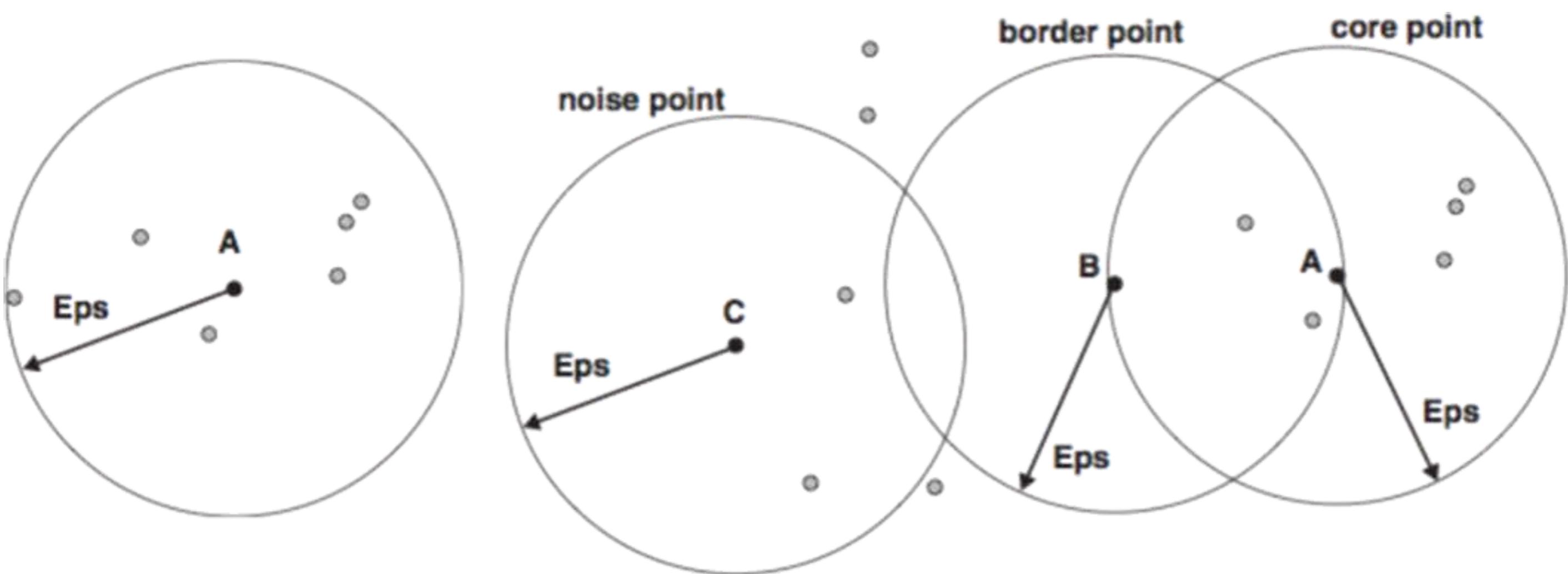
$$p(x) = \mathbf{w}_1 p_1(x) + \mathbf{w}_2 p_2(x)$$

» Е-шаг: $g_{ji} = p(j|x_i) = \frac{\mathbf{w}_j p_j(x_i)}{p(x_i)}$

» М-шаг: $\mathbf{w}_j = \frac{1}{N} \sum_{i=1}^N g_{ji}, \quad \mu_j = \frac{1}{N\mathbf{w}_j} \sum_{i=1}^N g_{ji} x_i$

$$\Sigma_j = \frac{1}{N\mathbf{w}_j - 1} \sum_{i=1}^N g_{ji} (x_i - \mu_i)(x_i - \mu_j)^T$$

ИДЕЯ DENSITY-BASED МЕТОДОВ



DBSCAN

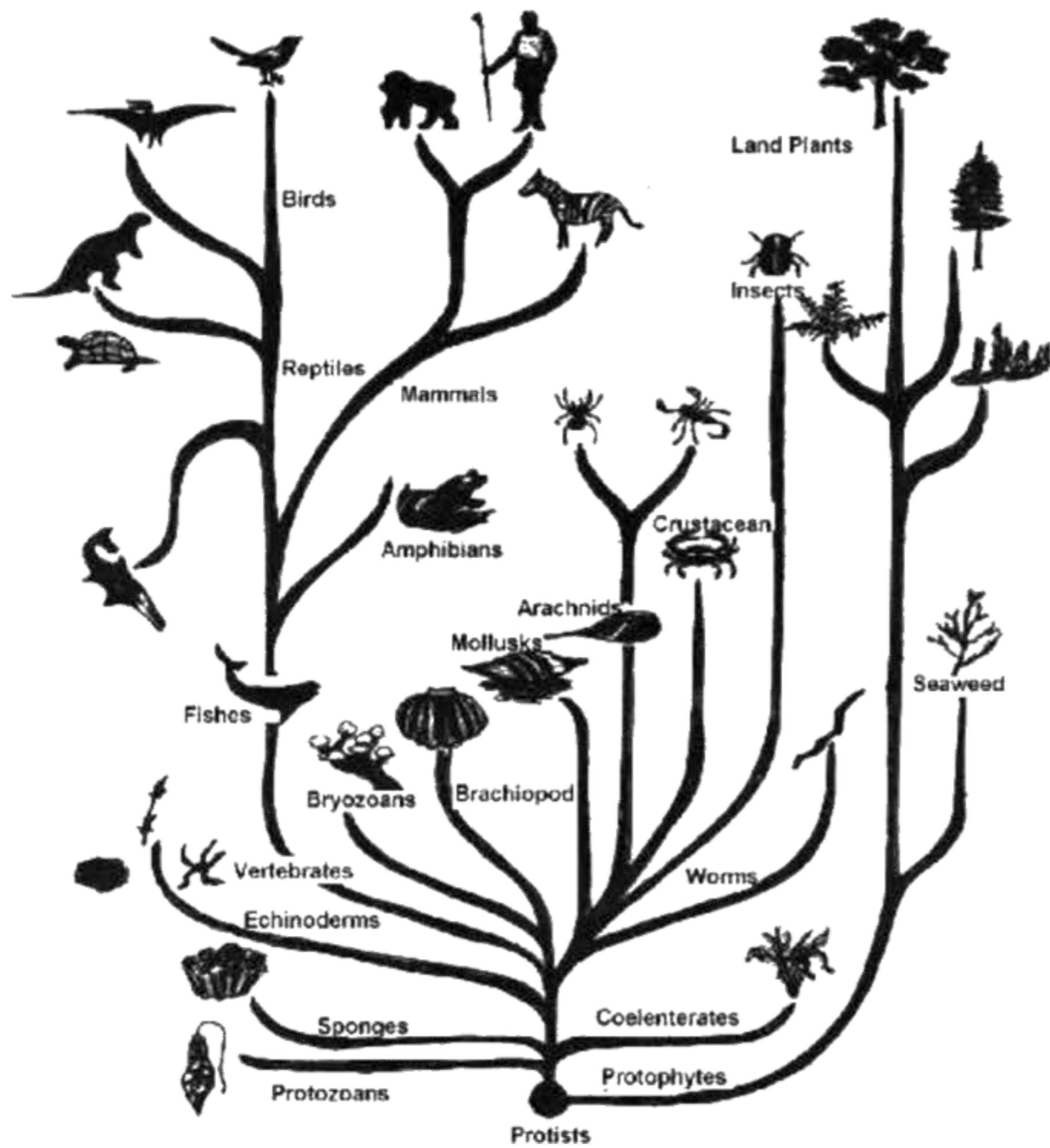
1. Пометить все точки как основные, пограничные или шумовые.
2. Отбросить точки шума.
3. Соединить все основные точки, находящиеся на расстоянии ϵ одна от другой.
4. Объединить каждую группу соединённых основных точек в отдельный кластер
5. Назначить каждую пограничную точку одному из кластеров, ассоциированных с ней основных точек.



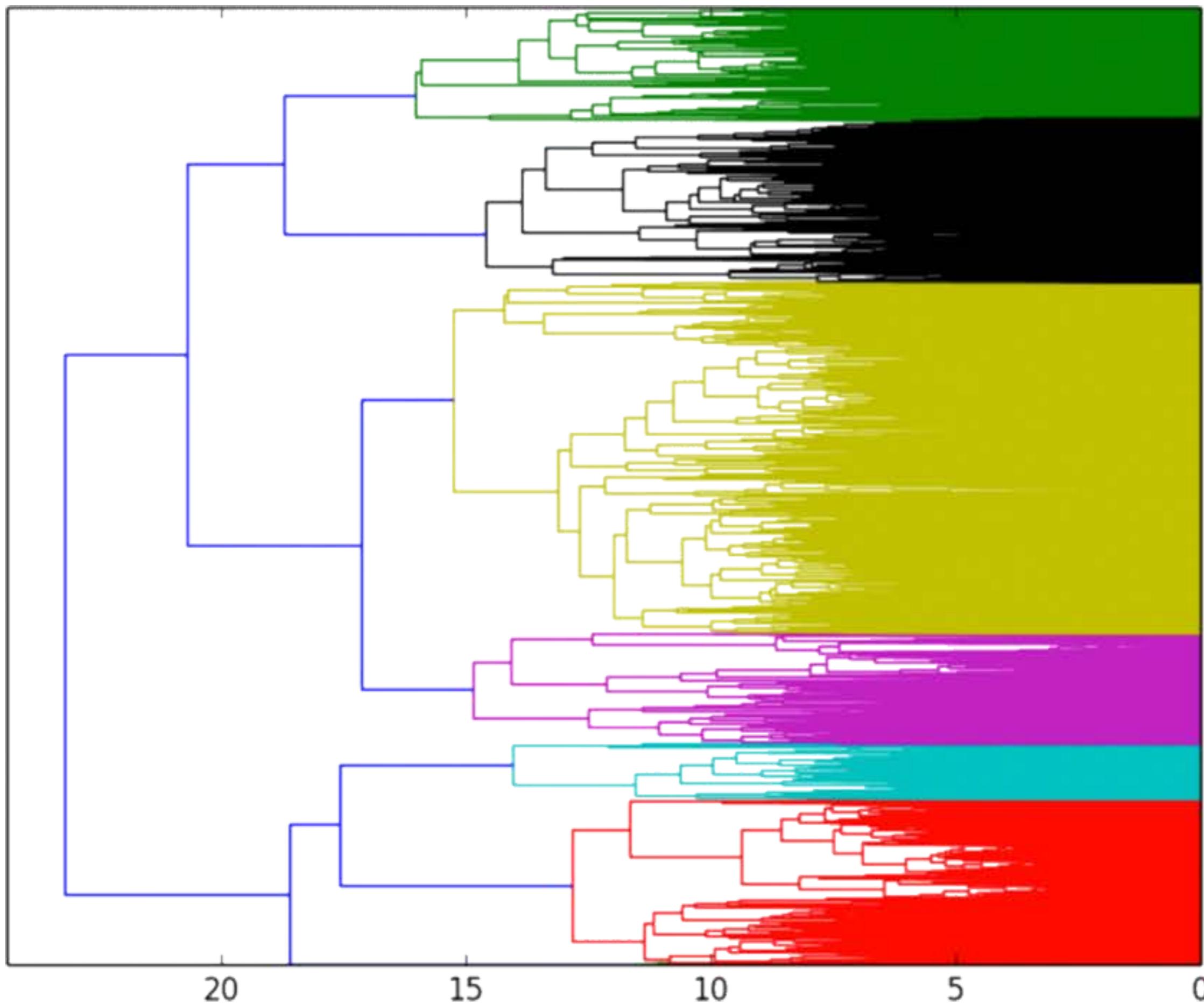
ИДЕЯ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

- › Вводим расстояние на объектах
- › Пытаемся выстроить «иерархию» вложенных друг в друга кластеров
- › Получаем дерево, вершины в котором кластеры
- › Дерево можно «обрезать» на какой-то фиксированной глубине и получить нужное число кластеров. Или оставить только достаточно большие кластеры

АНАЛОГИЯ ИЗ БИОЛОГИИ

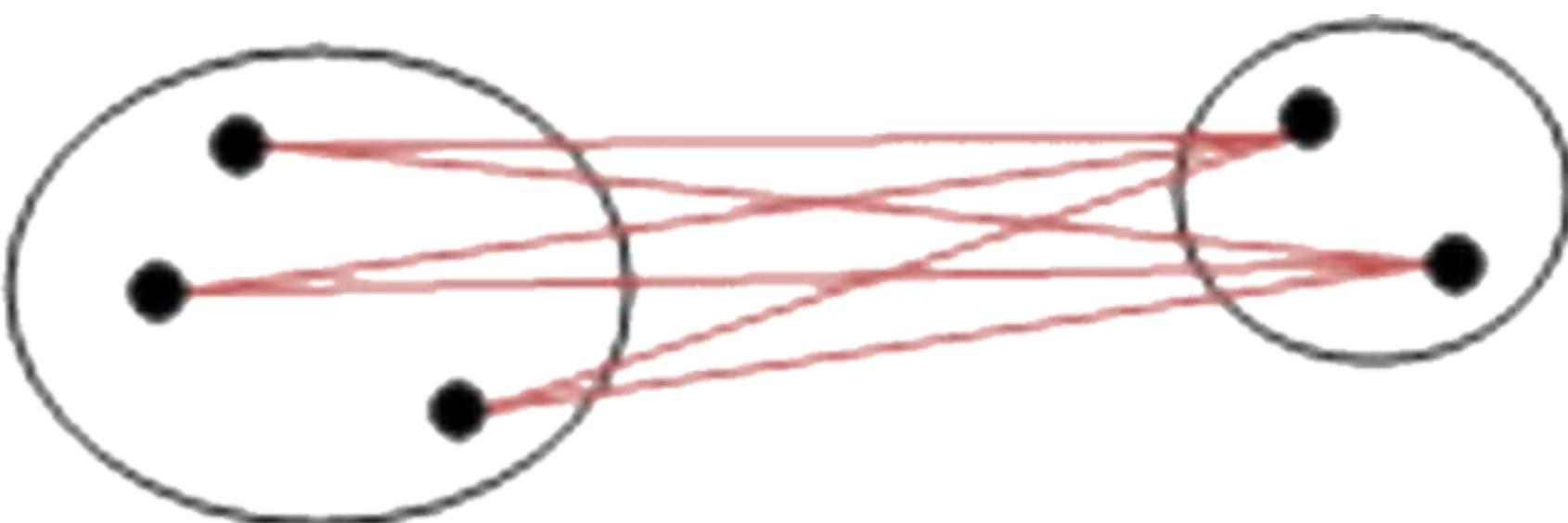


ДЕНДРОГРАММЫ

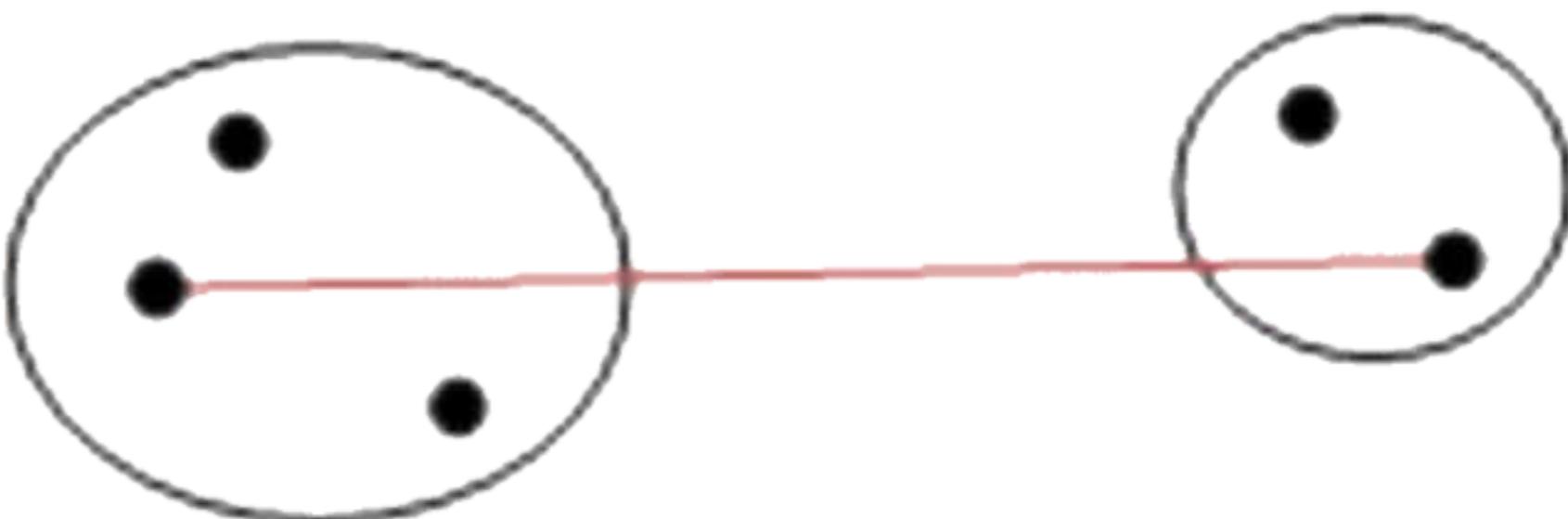


РАССТОЯНИЯ МЕЖДУ КЛАСТЕРАМИ

Average linkage



Complete linkage



Single linkage



ПОПРОБУЕМ СИСТЕМАТИЗИРОВАТЬ

По структуре кластеров:

- › Иерархические
 - ▶ Агломеративные
 - ▶ Дивизионные
- › Плоские

ПОПРОБУЕМ СИСТЕМАТИЗИРОВАТЬ

По форме:

- › Кластеры выпуклой формы
- › Кластеры-ленты
- › Сгустки на «фоне»
- › ...

ПОПРОБУЕМ СИСТЕМАТИЗИРОВАТЬ

По присвоению объектов к кластерам:

- › Жёсткая кластеризация
- › Мягкая кластеризация

РЕЗЮМЕ

- Метод K средних
- EM-алгоритм
- Методы, основанные на плотности точек
(density based)
- Иерархическая кластеризация