ЗАДАЧА КЛАСТЕРИЗАЦИИ

РАНЕЕ: ОБУЧЕНИЕ НА РАЗМЕЧЕННЫХ ДАННЫХ (SUPERVISED LEARNING)

- Обучающая выборка:

  $x_1, \ldots, x_\ell$ — объекты

  $y_1, \ldots, y_\ell$ — ответы

МФТИ

РАНЕЕ: ОБУЧЕНИЕ НА РАЗМЕЧЕННЫХ ДАННЫХ (SUPERVISED LEARNING)

- Обучающая выборка:

  $x_1, \ldots, x_\ell$ — объекты

  $y_1, \ldots, y_\ell$ — ответы

- Тестовая выборка:

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

МФТИ

РАНЕЕ: ОБУЧЕНИЕ НА РАЗМЕЧЕННЫХ ДАННЫХ (SUPERVISED LEARNING)

- Обучающая выборка:

  $x_1, \ldots, x_\ell$ — объекты

  $y_1, \ldots, y_\ell$ — ответы

- Тестовая выборка:

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

- В регрессии: y_i — прогнозируемая величина
  В классификации: y_i — метка класса

МФТИ

КЛАСТЕРИЗАЦИЯ

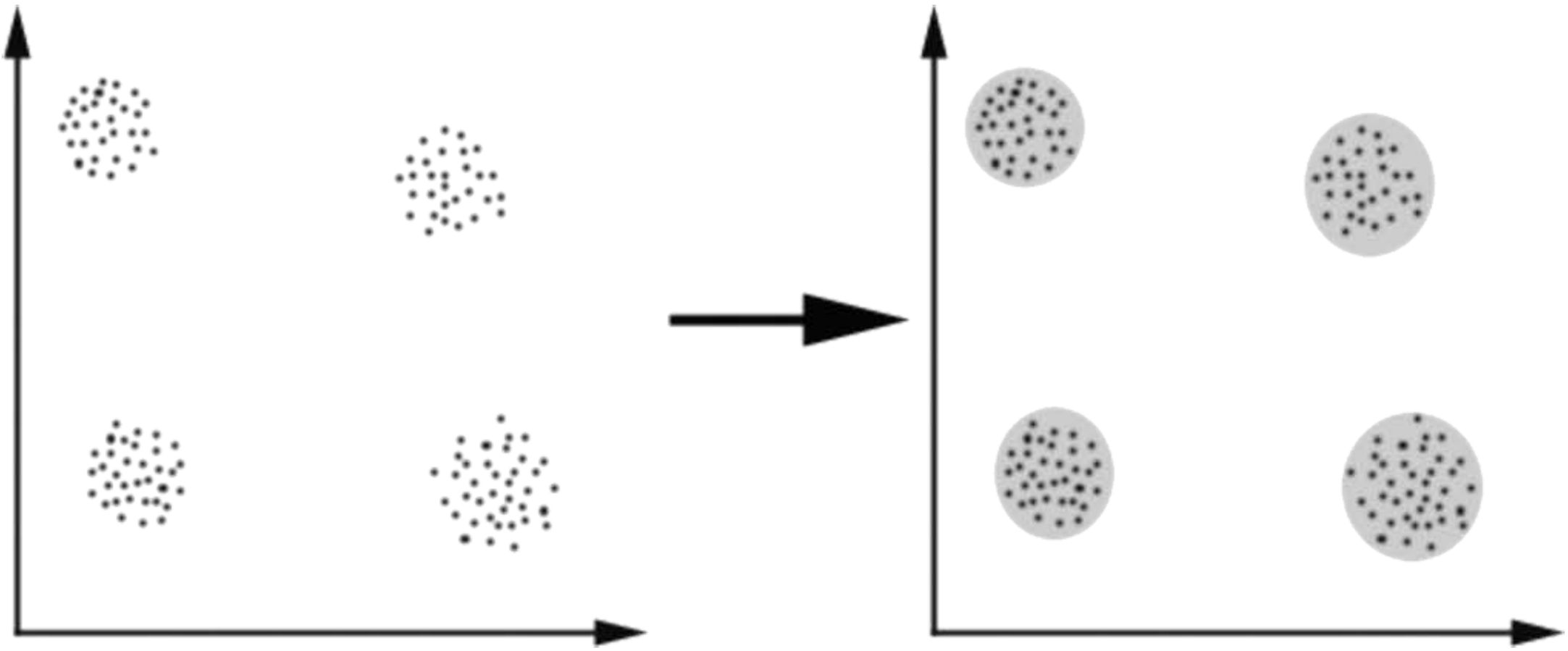
- «Обучающая» выборка:
  $x_1, \ldots, x_\ell$ — объекты

- Она же и тестовая

- Нужно поставить метки $y_1, \ldots, y_\ell$ так, чтобы объекты с одной и той же меткой были похожи, а с разными метками — не очень похожи

МФТИ

КАК ЭТО ВЫГЛЯДИТ



МФТИ

СРЕДНЕЕ ВНУТРИКЛАСТЕРНОЕ РАССТОЯНИЕ

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

СРЕДНЕЕ МЕЖКЛАСТЕРНОЕ РАССТОЯНИЕ

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

ПРИДУМЫВАЕМ МЕТРИКУ КАЧЕСТВА

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

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

МФТИ

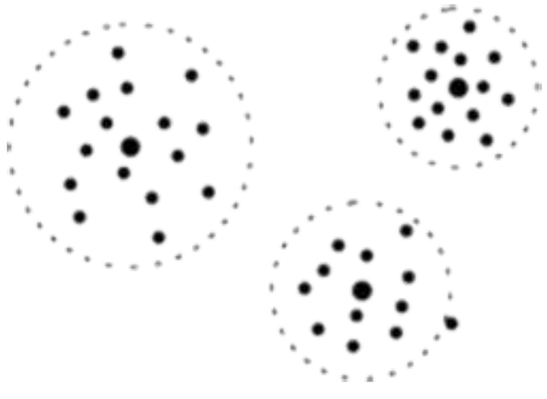
ПРИМЕРЫ ЗАДАЧ КЛАСТЕРИЗАЦИИ

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

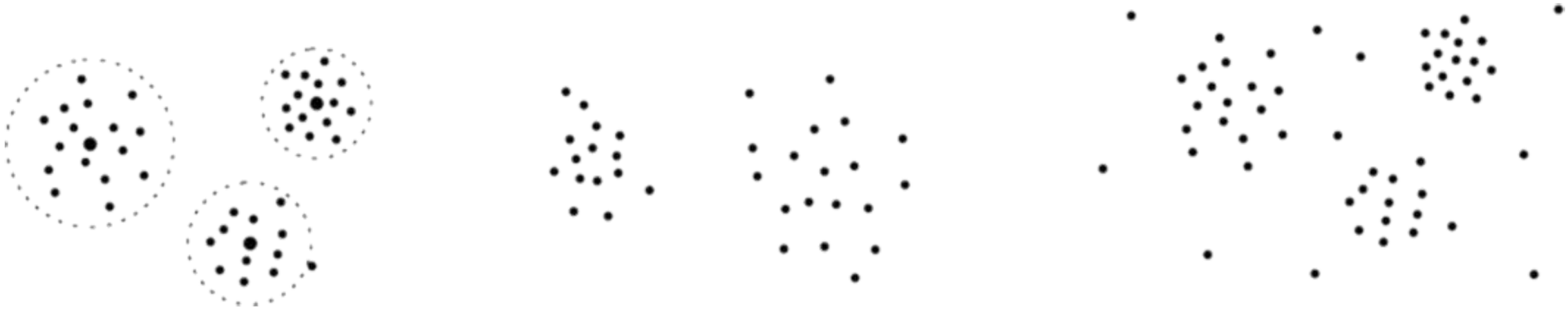
МФТИ

ФОРМА КЛАСТЕРОВ

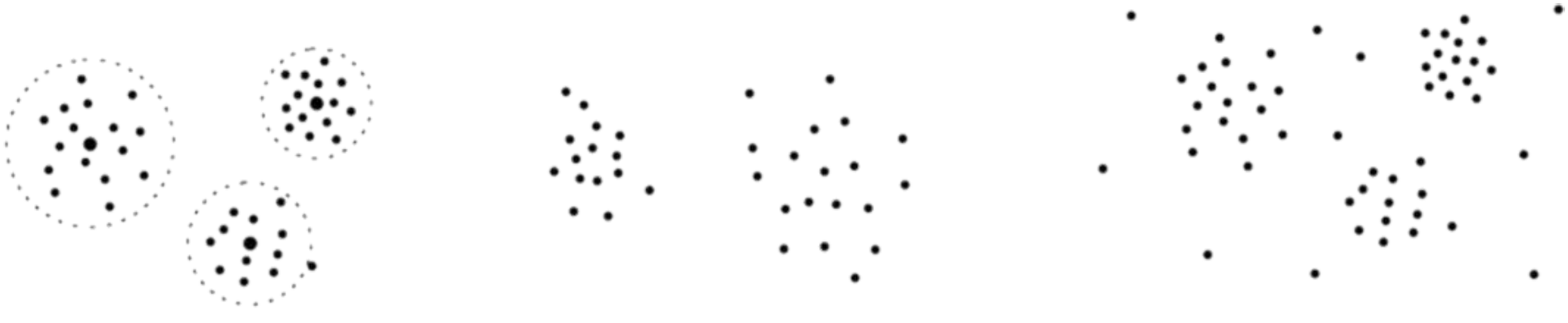


ФОРМА КЛАСТЕРОВ

ФОРМА КЛАСТЕРОВ

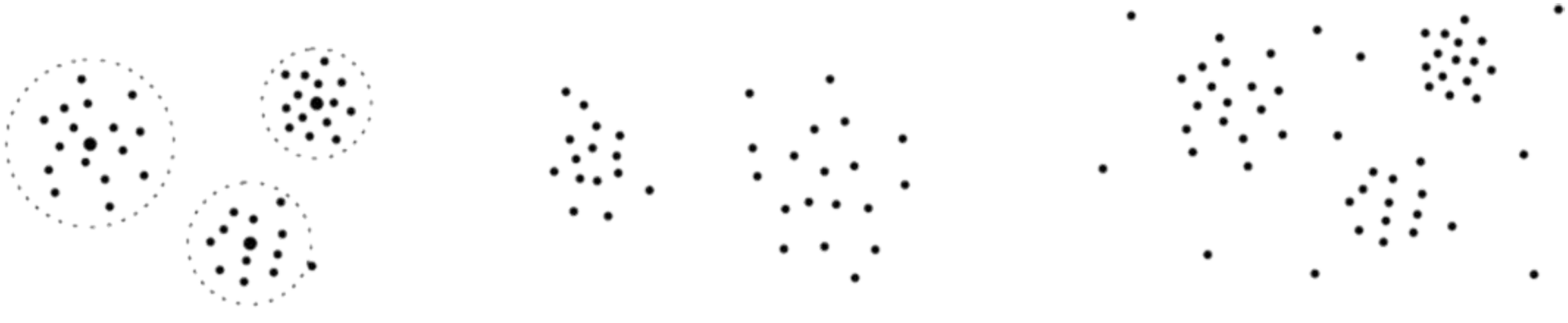


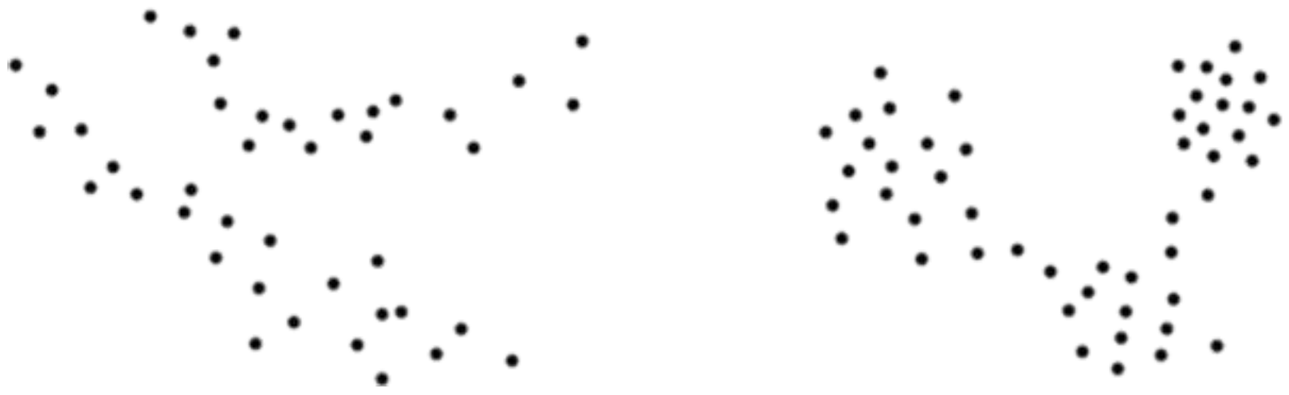
ФОРМА КЛАСТЕРОВ



МФТИ

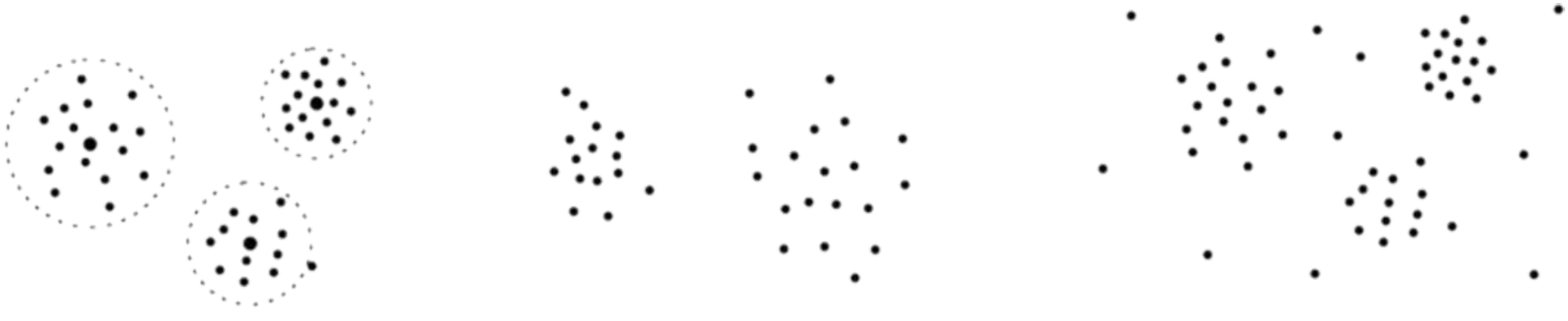
ФОРМА КЛАСТЕРОВ



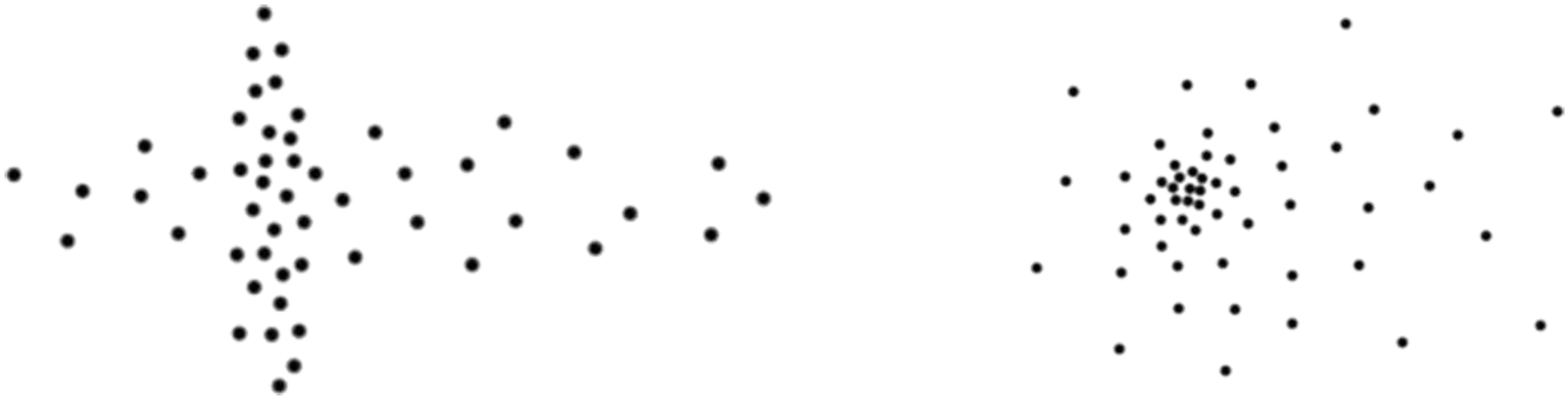


МФТИ

ФОРМА КЛАСТЕРОВ

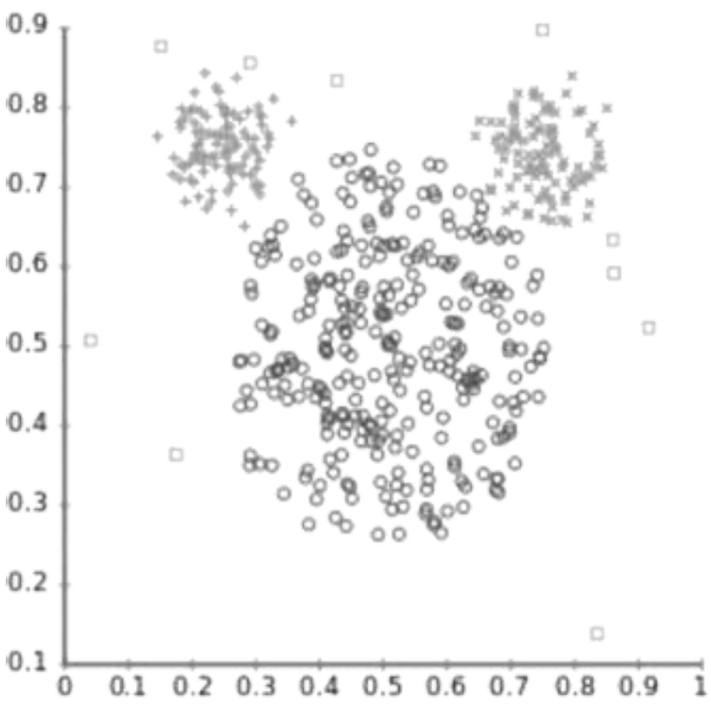


ФОРМА КЛАСТЕРОВ

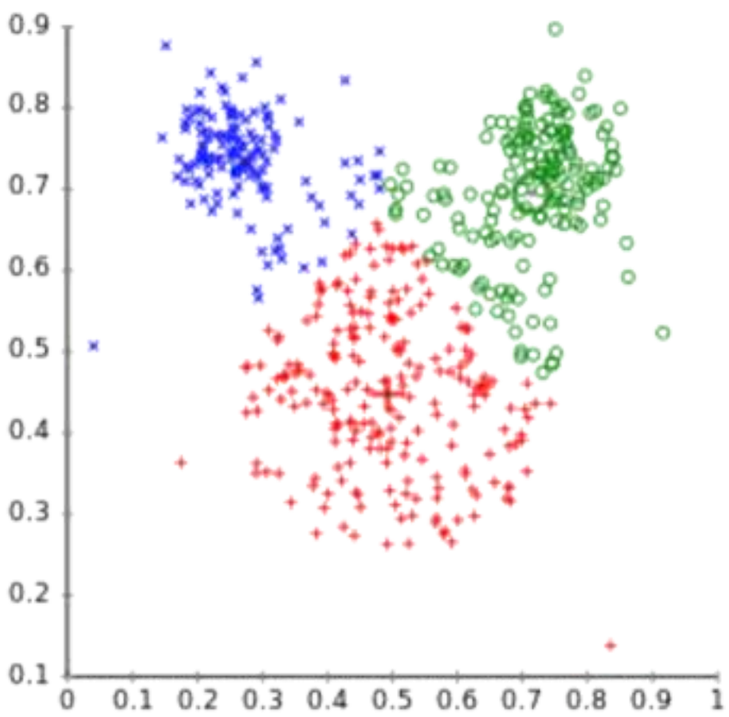


МФТИ

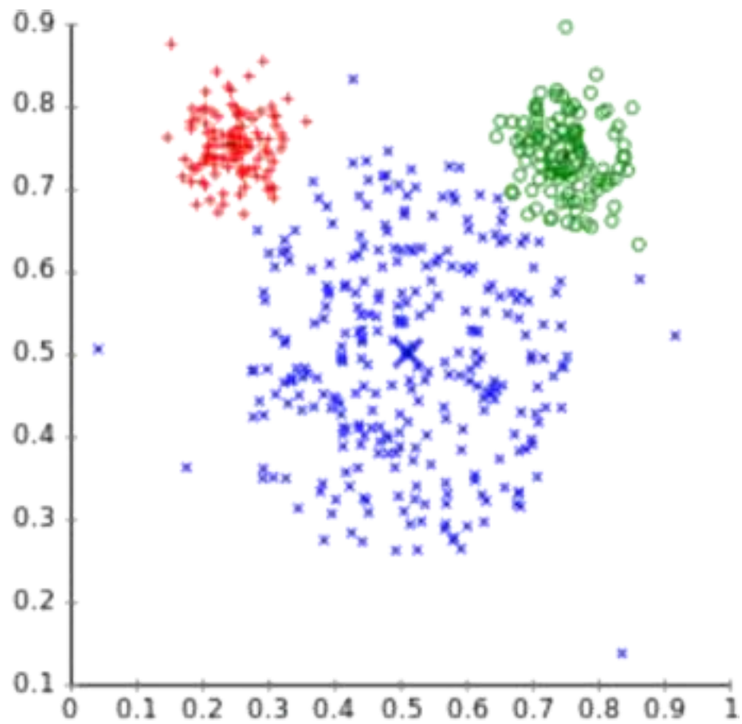
РАЗЛИЧИЯ В РЕЗУЛЬТАТАХ РАБОТЫ



Исходная выборка
("Mouse" dataset)



Метод k средних
(K-Means)

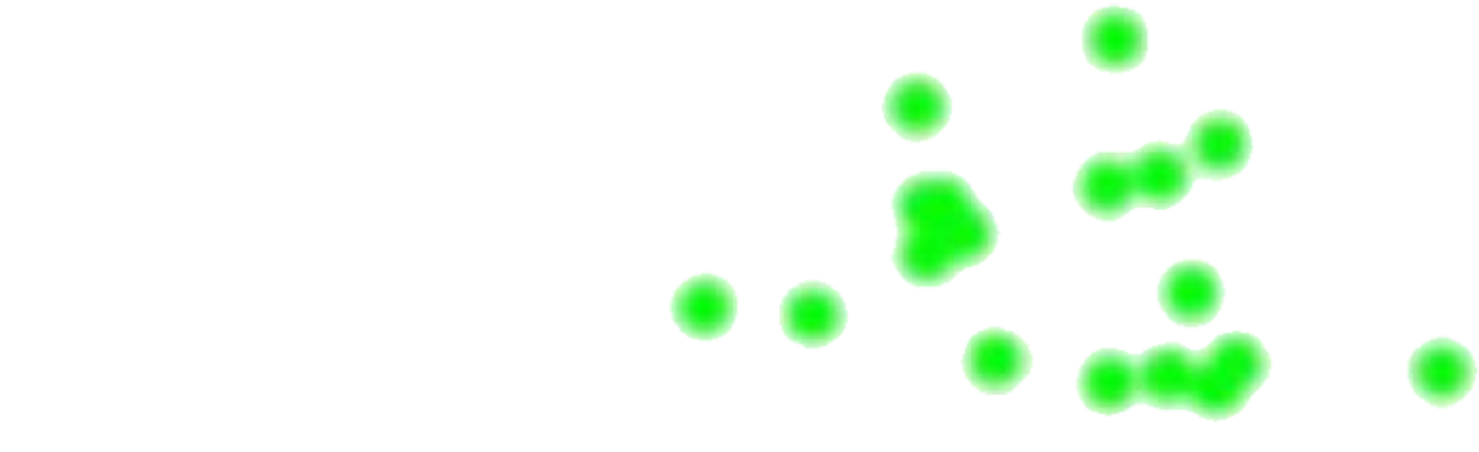


EM-алгоритм

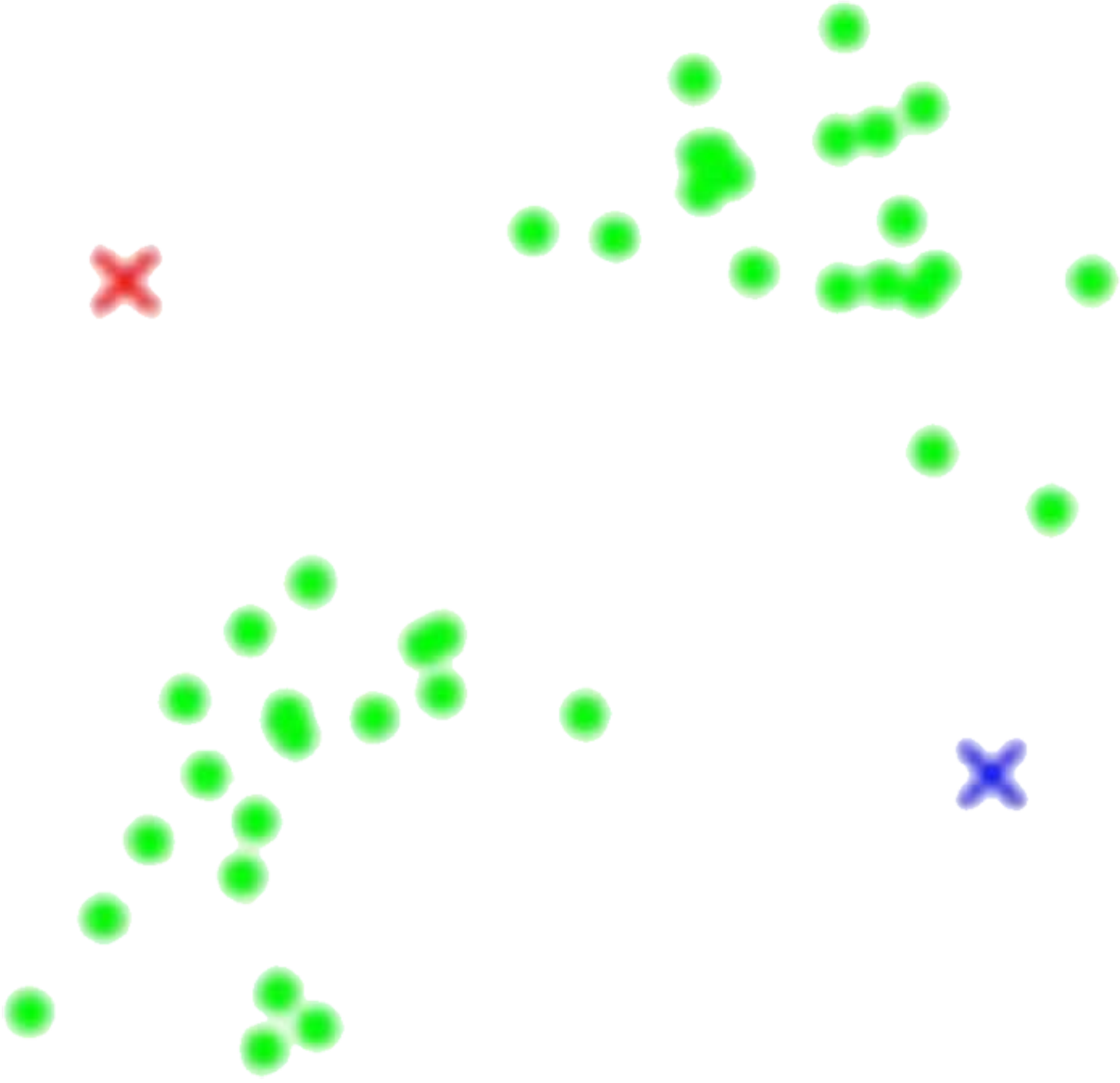
МФТИ

МЕТОД К СРЕДНИХ (K MEANS)

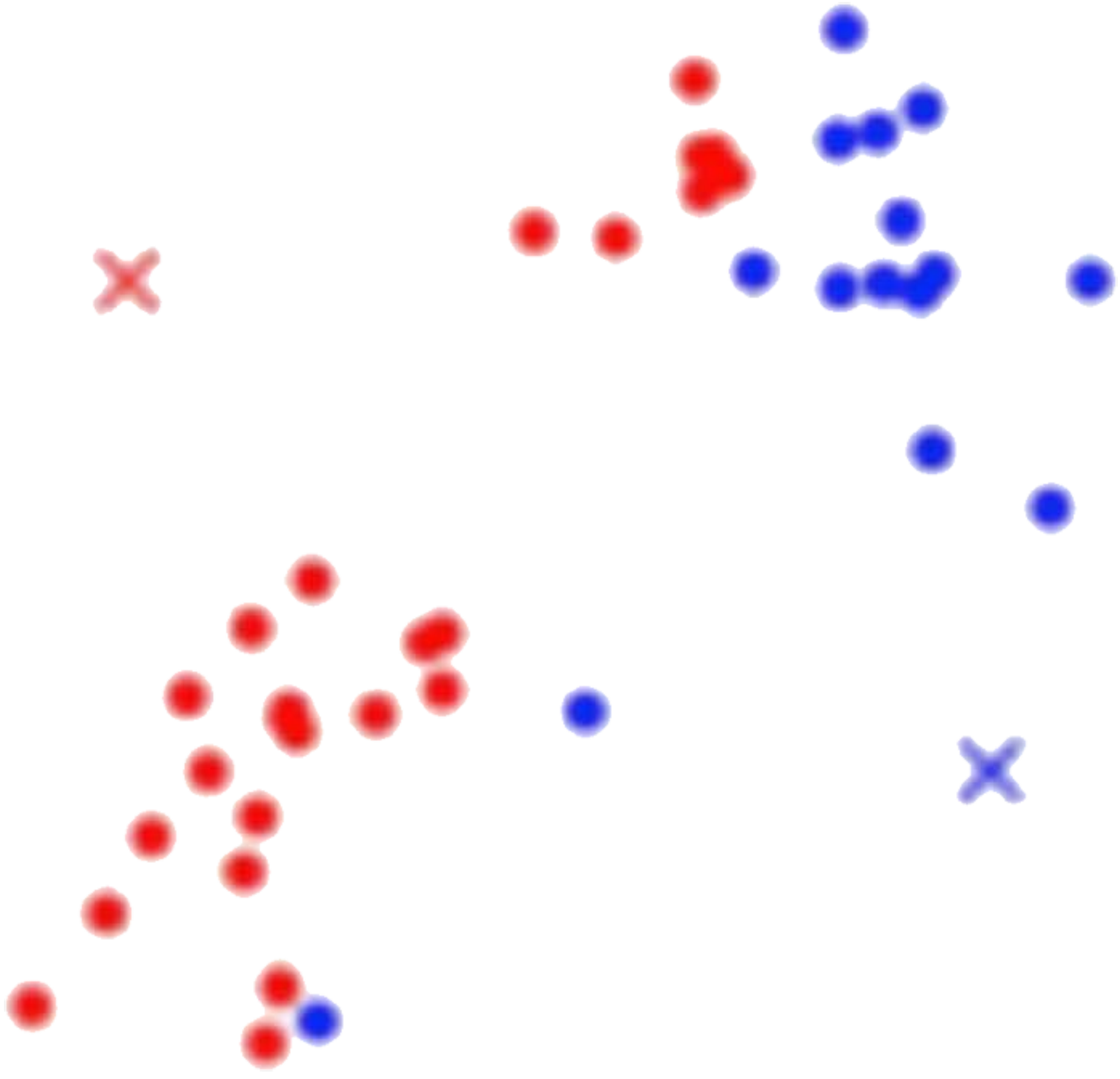
КАК РАБОТАЕТ K MEANS



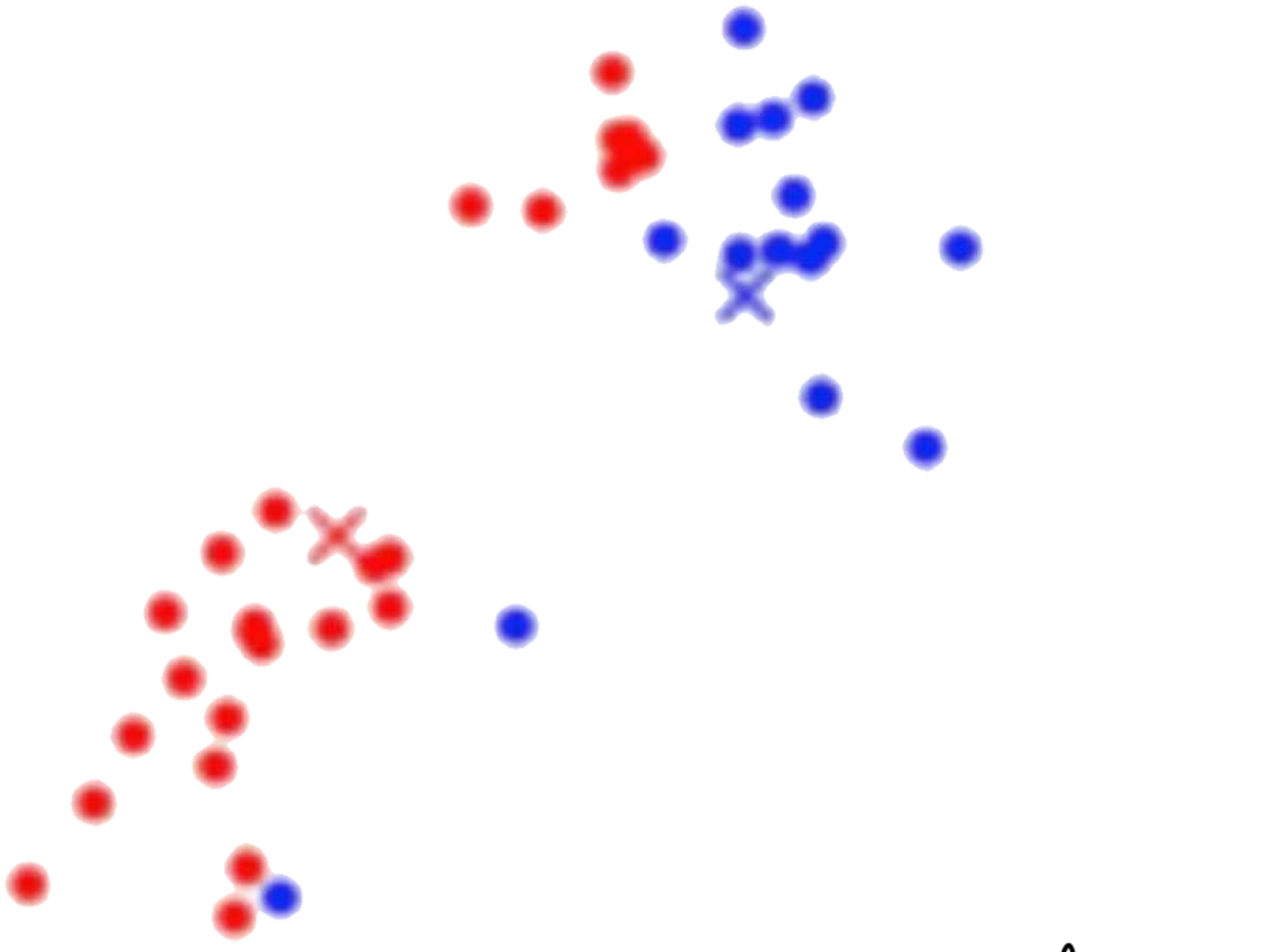
КАК РАБОТАЕТ K MEANS



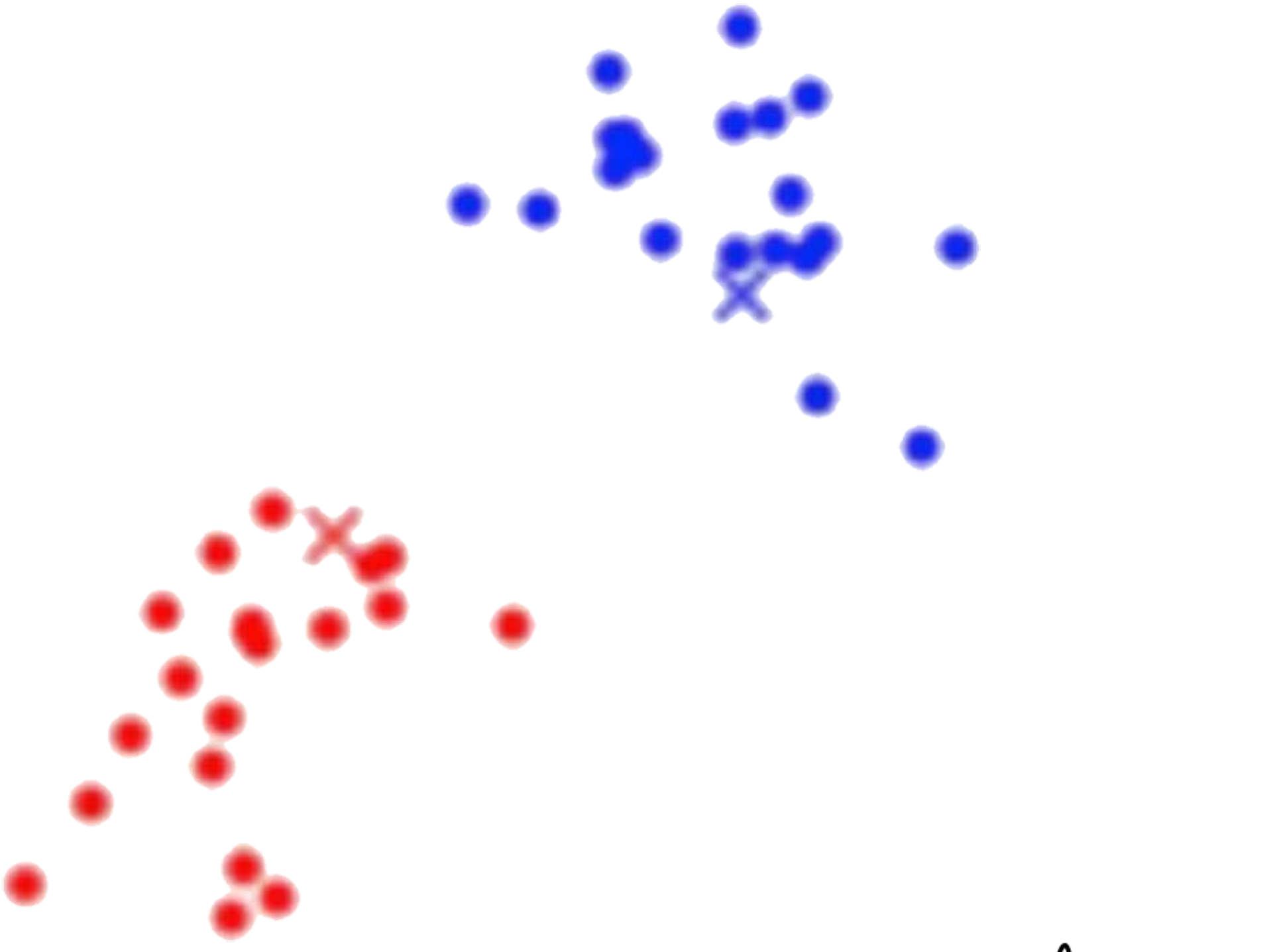
КАК РАБОТАЕТ K MEANS



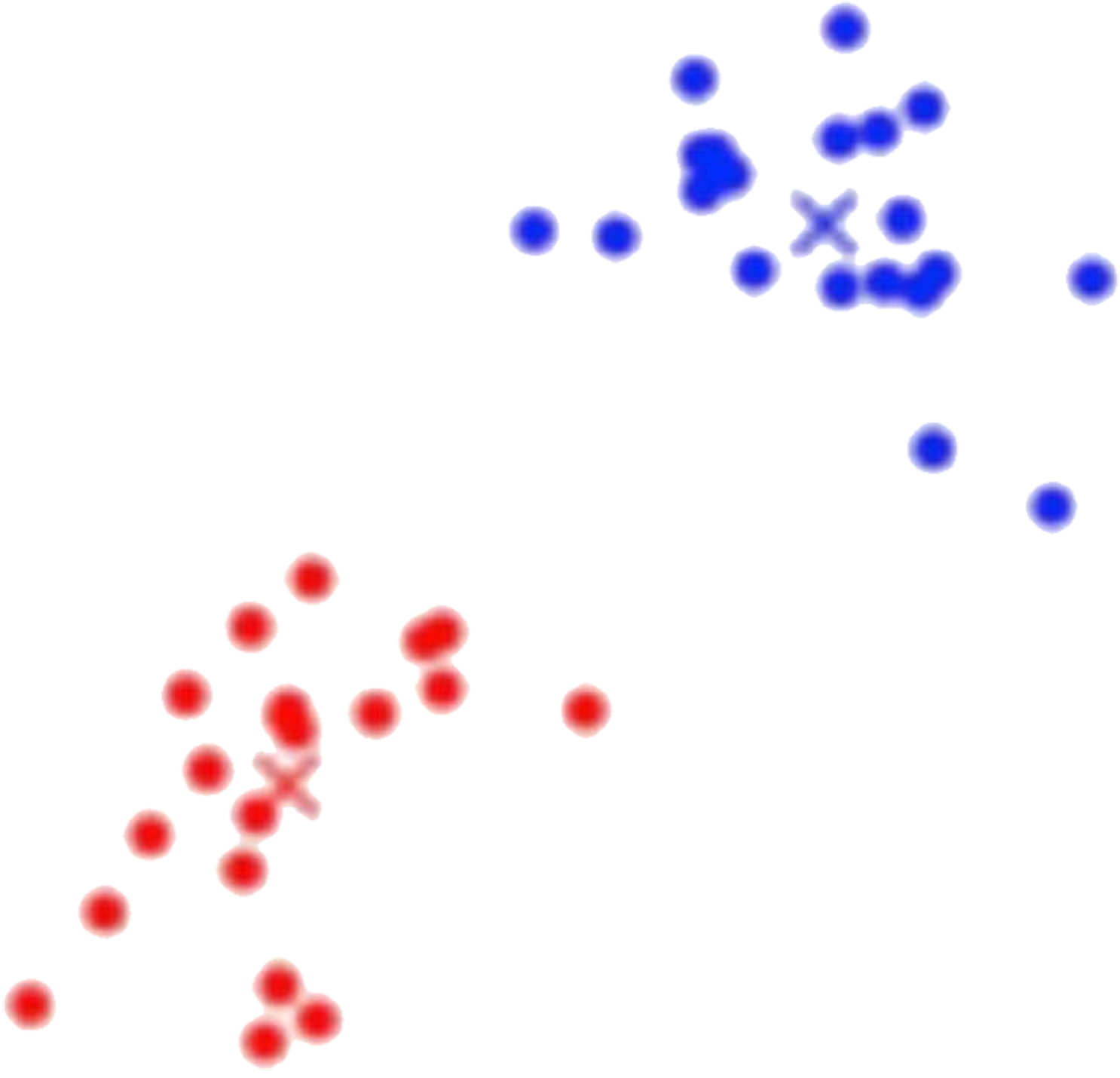
КАК РАБОТАЕТ K MEANS



КАК РАБОТАЕТ K MEANS



КАК РАБОТАЕТ K MEANS



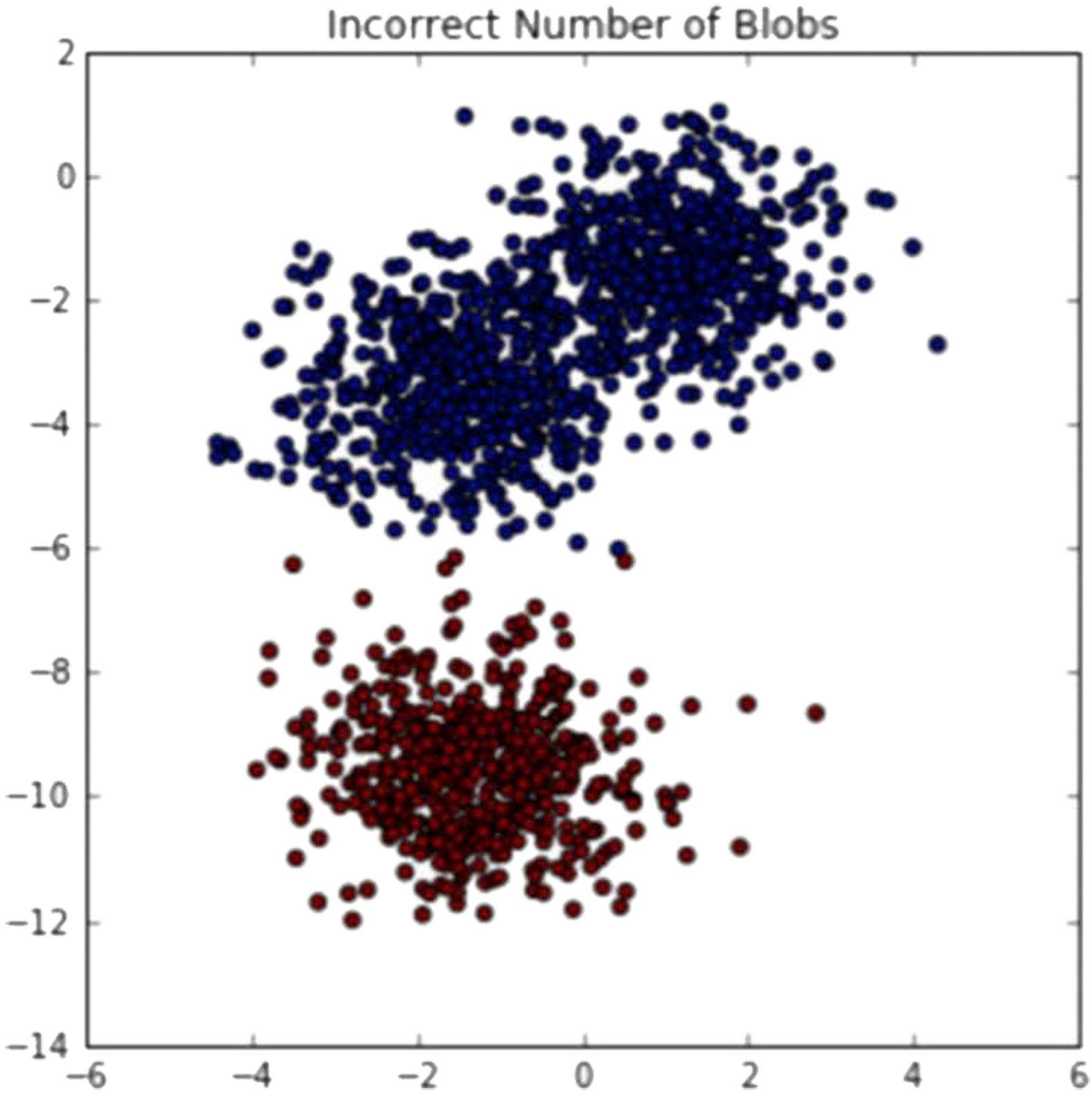
ПРИМЕР: КВАНТИЗАЦИЯ ИЗОБРАЖЕНИЙ

ПРИМЕР: КВАНТИЗАЦИЯ ИЗОБРАЖЕНИЙ

Quantized image (64 colors, Random)

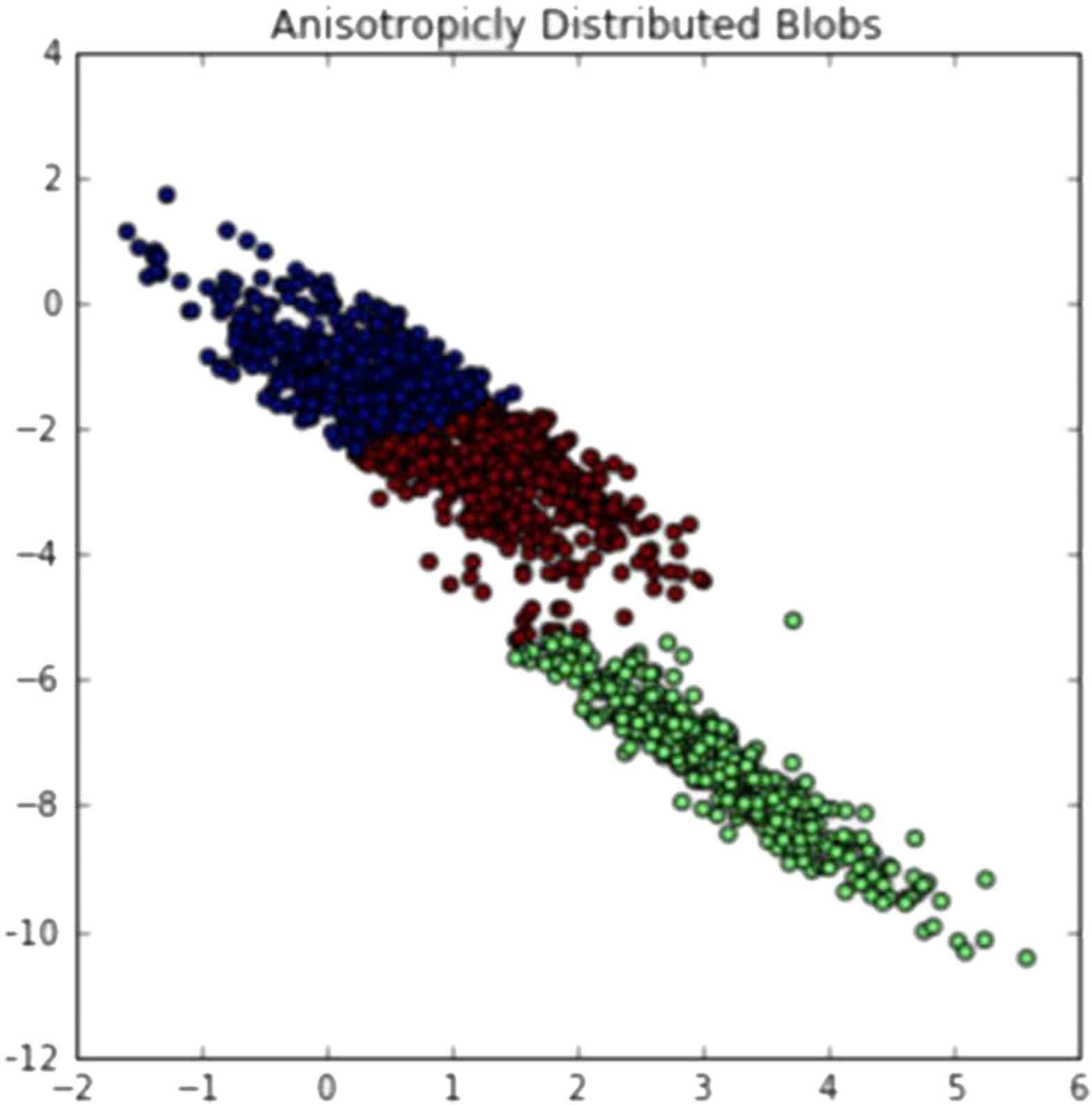
ПРИМЕР: КВАНТИЗАЦИЯ ИЗОБРАЖЕНИЙ

K MEANS И РАЗНЫЕ ФОРМЫ КЛАСТЕРОВ



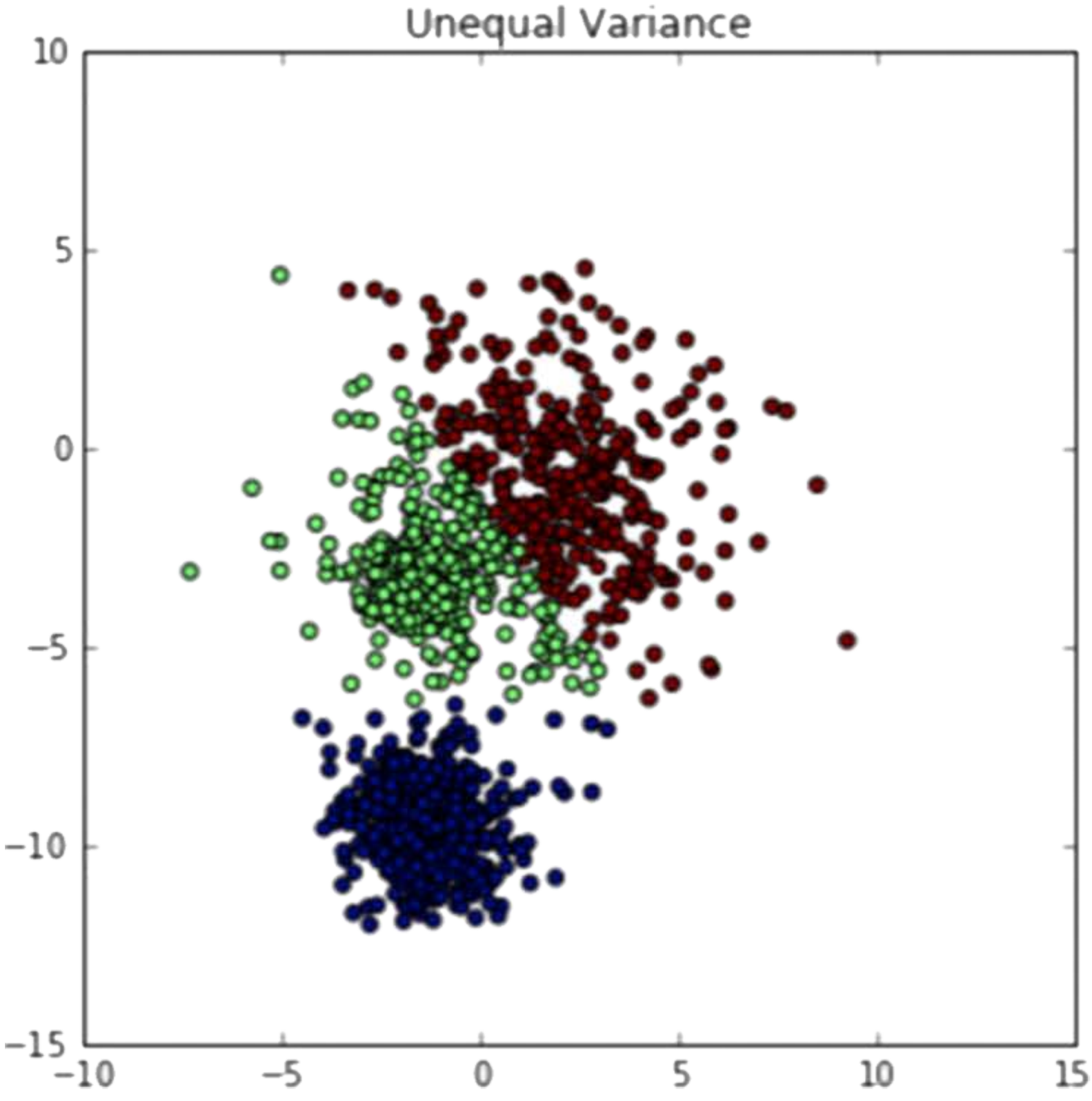
МФТИ

K MEANS И РАЗНЫЕ ФОРМЫ КЛАСТЕРОВ



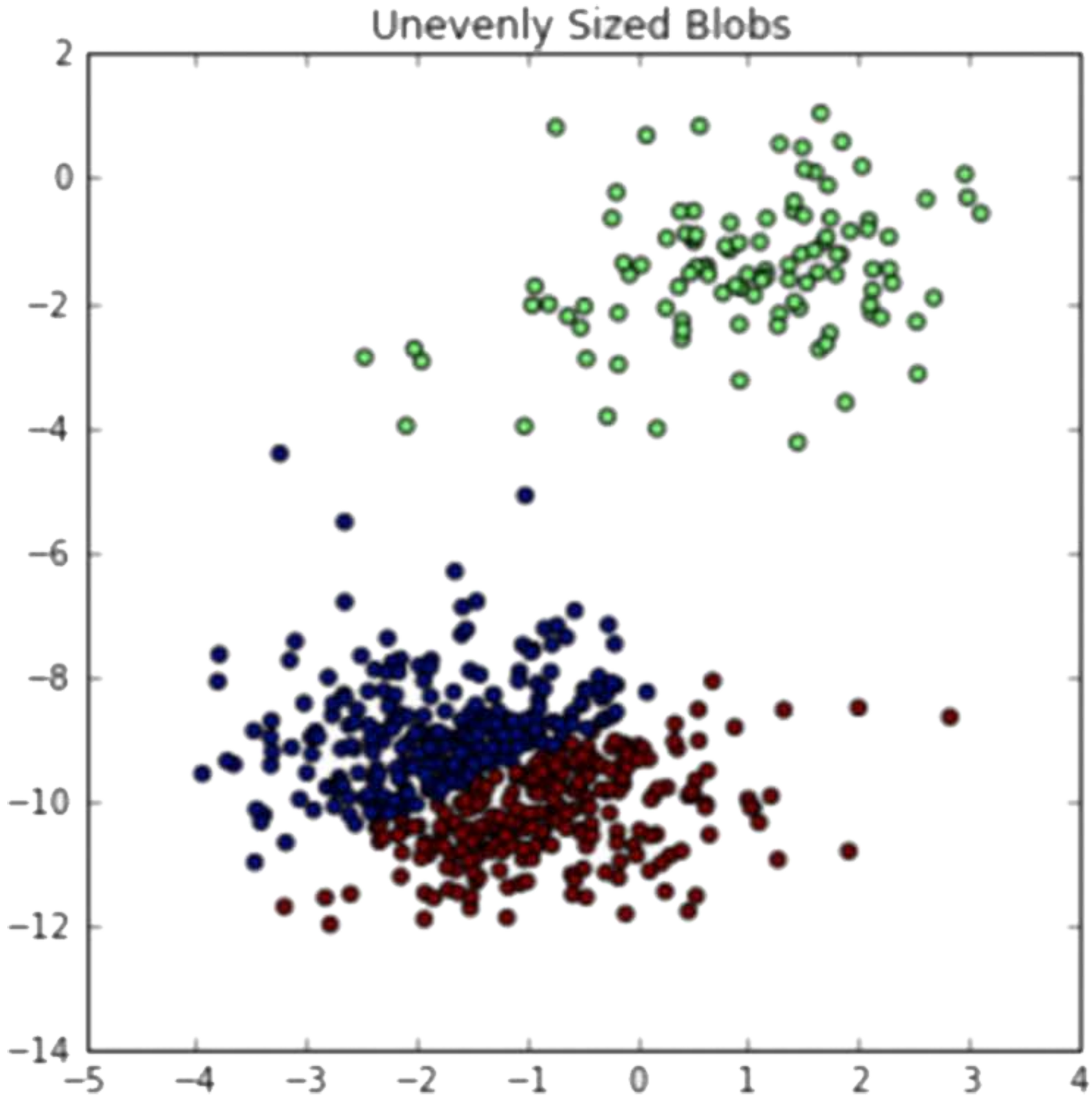
МФТИ

K MEANS И РАЗНЫЕ ФОРМЫ КЛАСТЕРОВ



МФТИ

K MEANS И РАЗНЫЕ ФОРМЫ КЛАСТЕРОВ



МФТИ

ЧТО ОПТИМИЗИРУЕТ K MEANS

- Среднее внутрикластерное расстояние:

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

МФТИ

ЧТО ОПТИМИЗИРУЕТ K MEANS

- Среднее внутрикластерное расстояние:

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

- Альтернативный вариант, если есть центры кластеров:

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

МФТИ