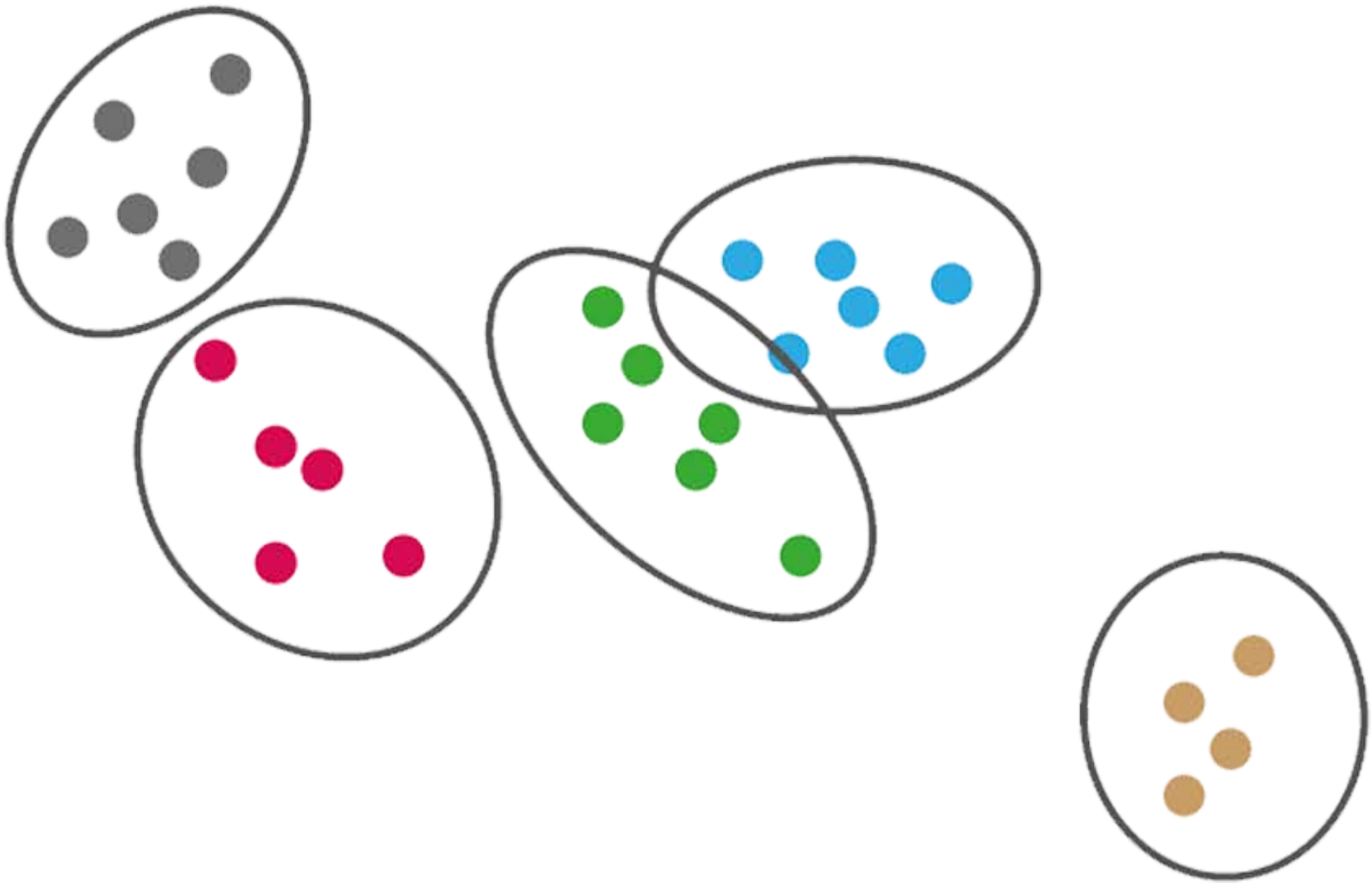
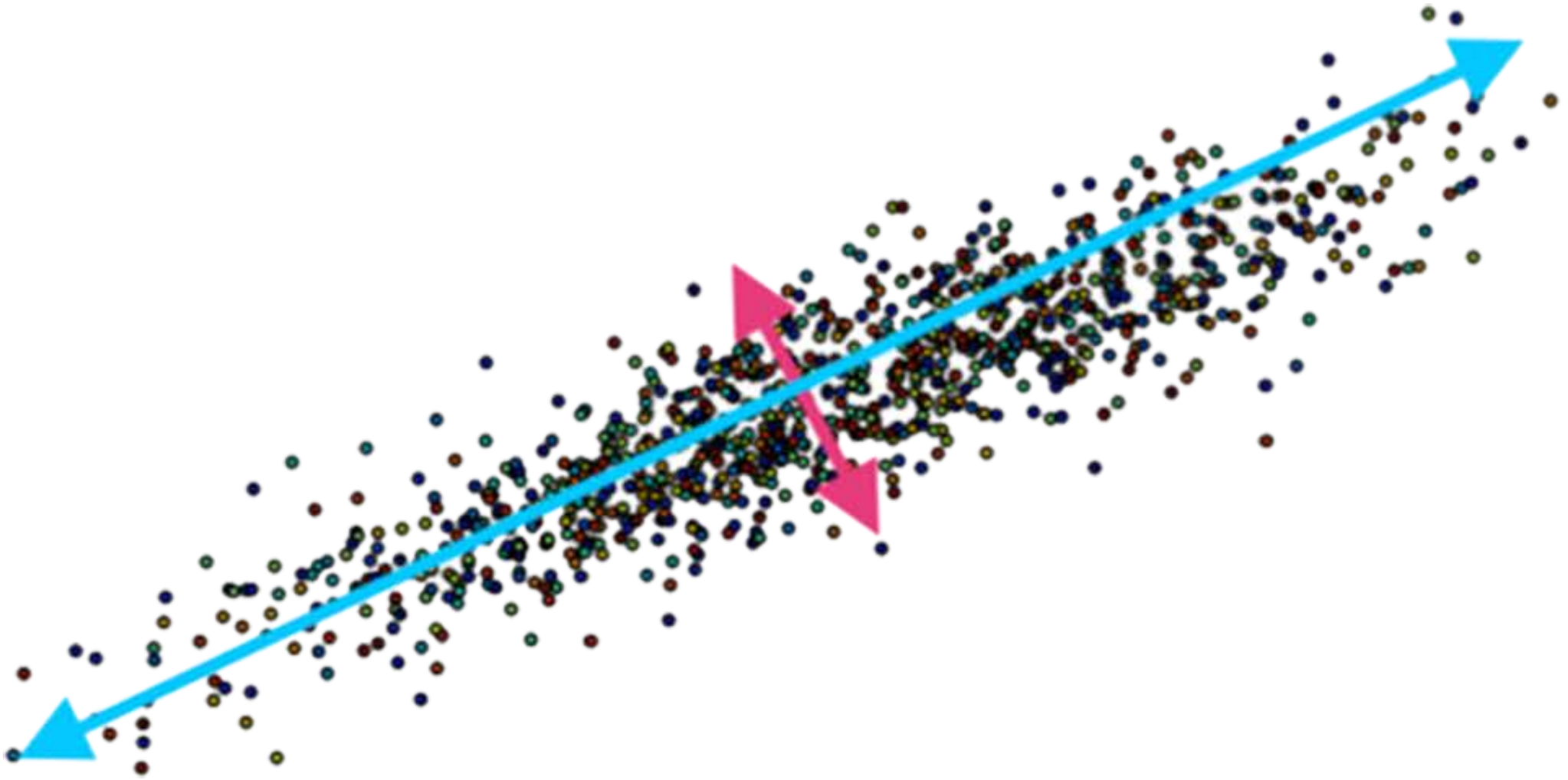
ЗАДАЧА ОБНАРУЖЕНИЯ АНОМАЛИЙ

КЛАСТЕРИЗАЦИЯ



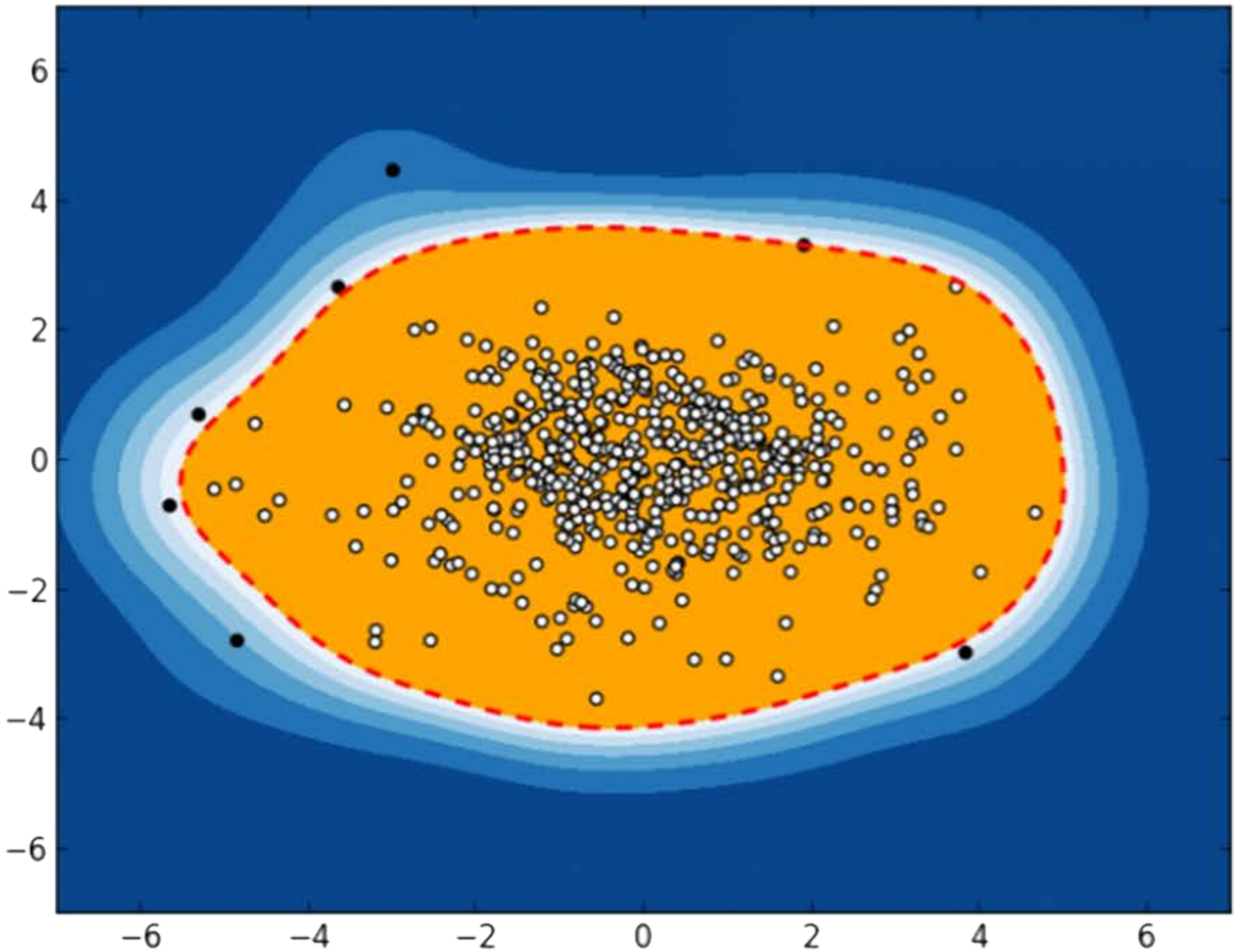
МФТИ

ПОНИЖЕНИЕ РАЗМЕРНОСТИ



МФТИ

ПОИСК АНОМАЛИЙ



МФТИ

ПОИСК АНОМАЛИЙ

- Похож ли новый объект на остальных?

ПОИСК АНОМАЛИЙ

- Объект — клиент банка в текущий момент времени
- Не выделяется ли его поведение?
- Не мошенник ли это?

МФТИ

ПОИСК АНОМАЛИЙ

- Объект — показатели сложной компьютерной системы
- Загрузка процессоров, памяти, сети и т.д.
- Отличается ли текущее состояние системы от тех, которые мы наблюдали ранее?

МФТИ

ПОИСК АНОМАЛИЙ

- Задача: определение тональности отзыва на банк
- Можно ли к новому отзыву применять модель, обученную на прошлых данных?
- Не изменилось ли распределение признаков?

МФТИ

МЕТОДЫ ОБНАРУЖЕНИЯ АНОМАЛИЙ

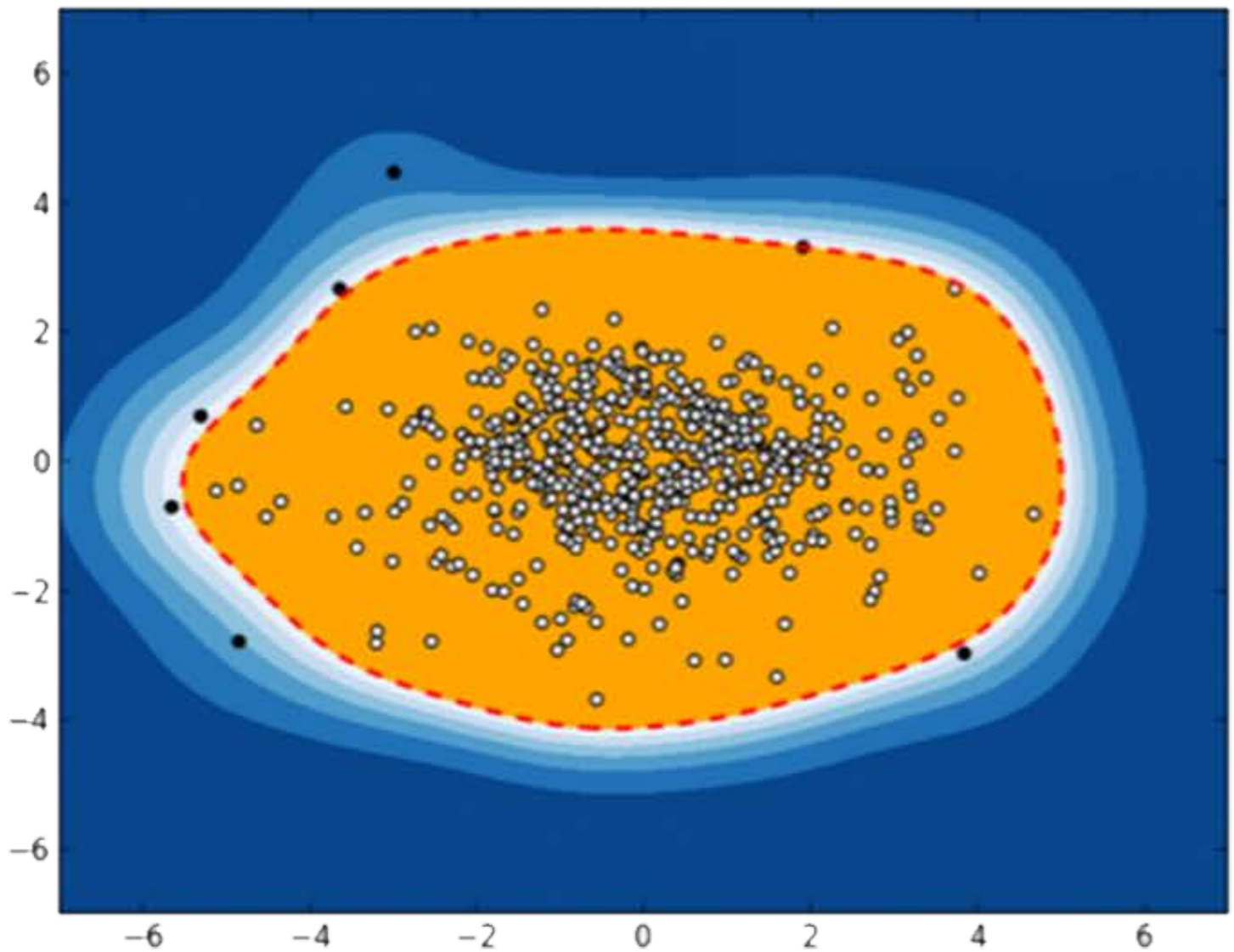
- Методы, основанные на восстановлении плотности
- Методы, основанные на классификации

РЕЗЮМЕ

- Поиск аномалий — обнаружение объектов, которые существенно отличаются от других

ПАРАМЕТРИЧЕСКОЕ ВОССТАНОВЛЕНИЕ ПЛОТНОСТИ

ПОИСК АНОМАЛИЙ



МФТИ

АНОМАЛИЯ

- Объект, который получен из другого распределения на пространстве объектов
- Как восстановить распределение?

МФТИ

ВОССТАНОВЛЕНИЕ РАСПРЕДЕЛЕНИЙ

- Параметрические методы
- Непараметрические методы
- Восстановление смесей

МФТИ

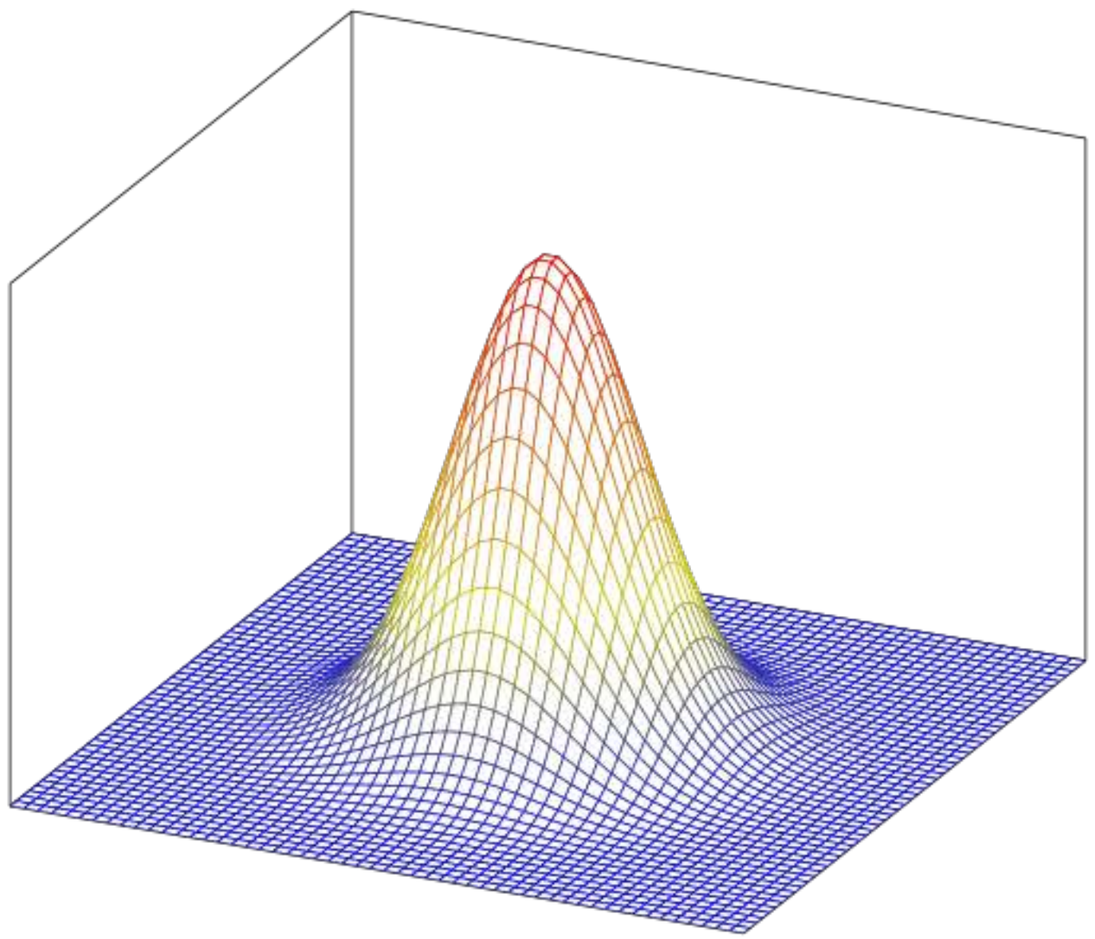
ПАРАМЕТРИЧЕСКИЙ ПОДХОД

- $p(x) = \phi(x|\theta)$
- θ — параметры распределения

ПРИМЕР

$$\phi(x|\theta) = \mathcal{N}(\mu, \Sigma)$$

- Нормальное распределение
- Параметры: $\theta = (\mu, \Sigma)$



МФТИ

ОБУЧЕНИЕ

- Нужно подобрать параметры распределения θ
- Вероятность у объектов из выборки должна быть как можно выше

МФТИ

МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

$$\sum_{i=1}^{\ell} \log\phi(x_i|\theta) \to \max_{\theta}$$

- Для некоторых распределений решается аналитически

МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

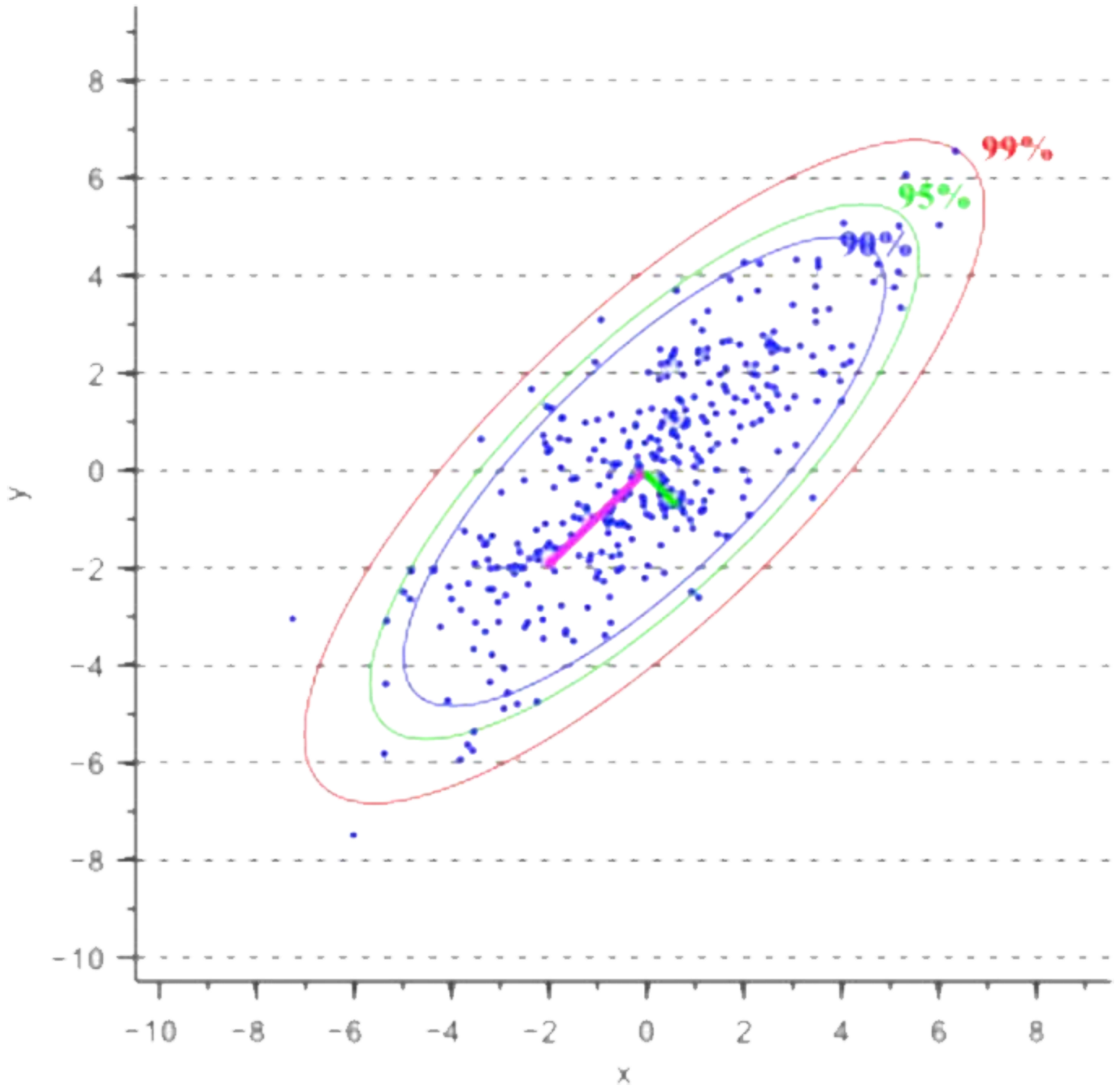
- Для нормального распределения:

$$\mu = \frac{1}{\ell}\sum_{i=1}^{\ell} x_i$$

$$\Sigma = \frac{1}{\ell}\sum_{i=1}^{\ell}(x_i - \mu)(x_i - \mu)^T$$

МФТИ

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

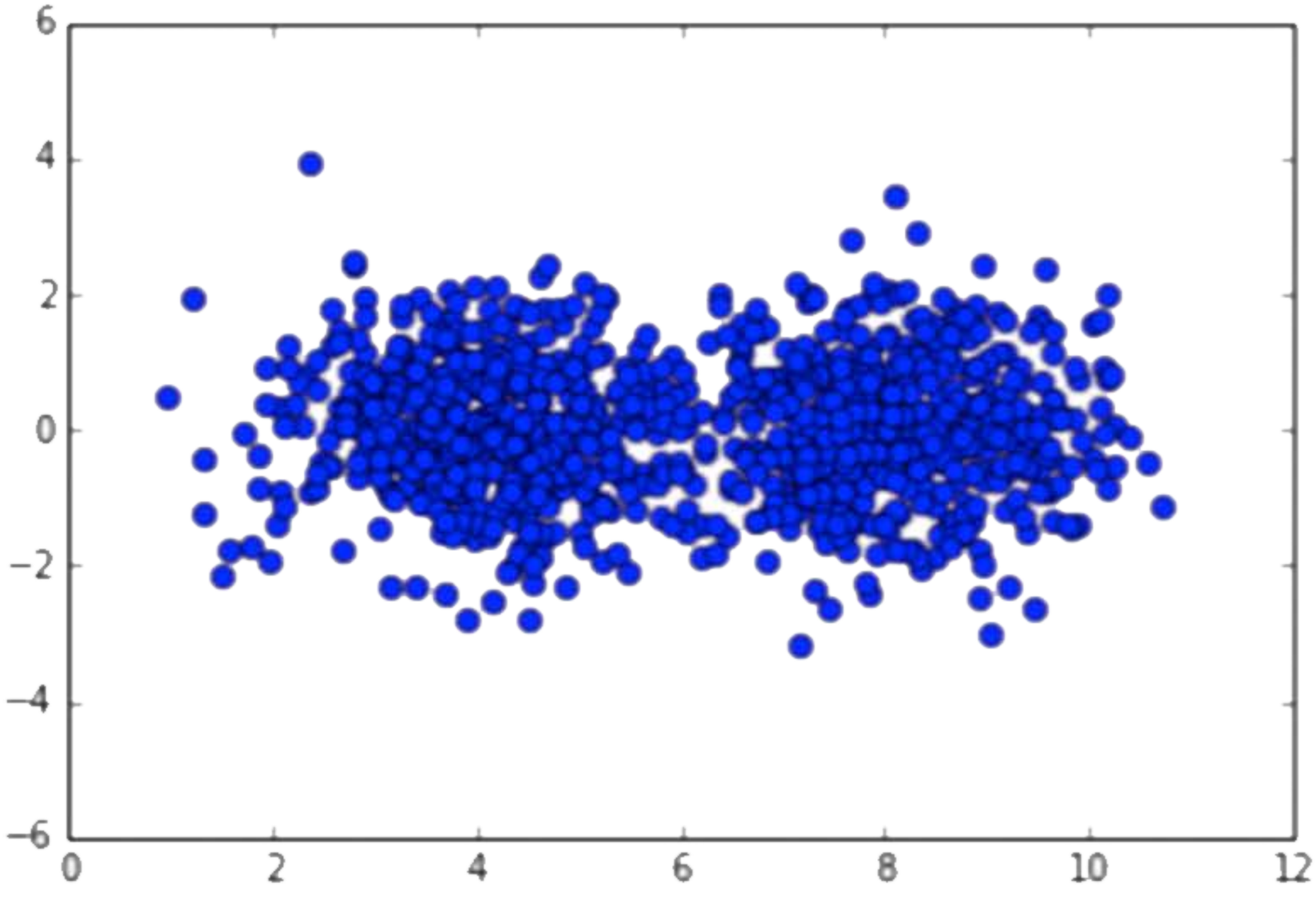


ОБНАРУЖЕНИЕ АНОМАЛИЙ

- Новый объект x
- Если $p(x) < t$, то это аномалия
- Порог t:
  - Из априорных соображений
  - По известным аномалиям

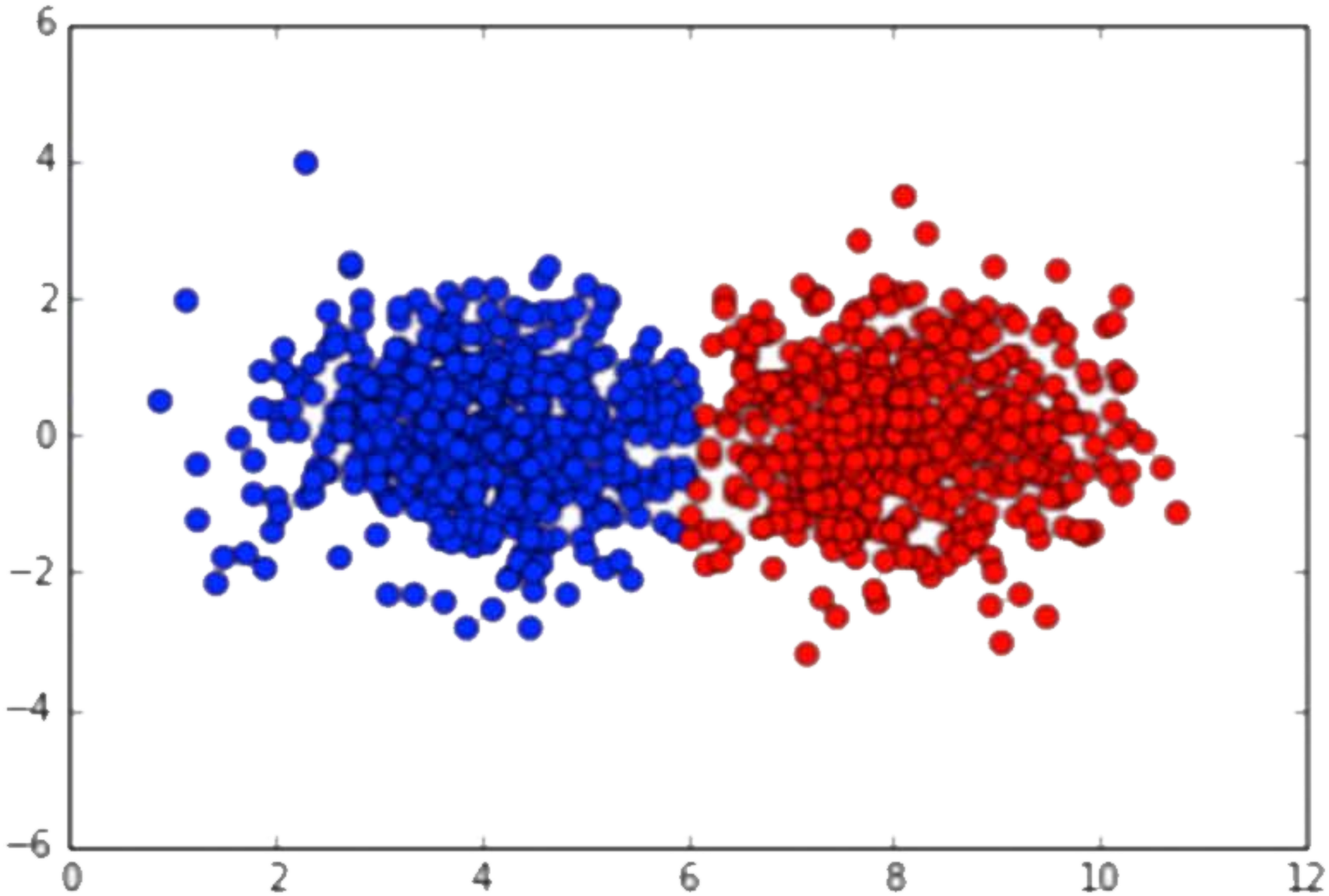
МФТИ

СМЕСИ РАСПРЕДЕЛЕНИЙ



МФТИ

СМЕСИ РАСПРЕДЕЛЕНИЙ



МФТИ

EM-АЛГОРИТМ

$$p(x) = \sum_{j=1}^{K} \mathbf{w}_j \, p_j(x) \qquad p_j(x) = \phi(x|\theta_j)$$

- E-шаг:

$$g_{ji} = p(j|x_i) = \frac{\mathbf{w}_j \, p_j(x_i)}{p(x_i)}$$

- M-шаг:

$$\mathbf{w}_j = \frac{1}{N} \sum_{i=1}^{N} g_{ji}$$

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

МФТИ

РЕЗЮМЕ

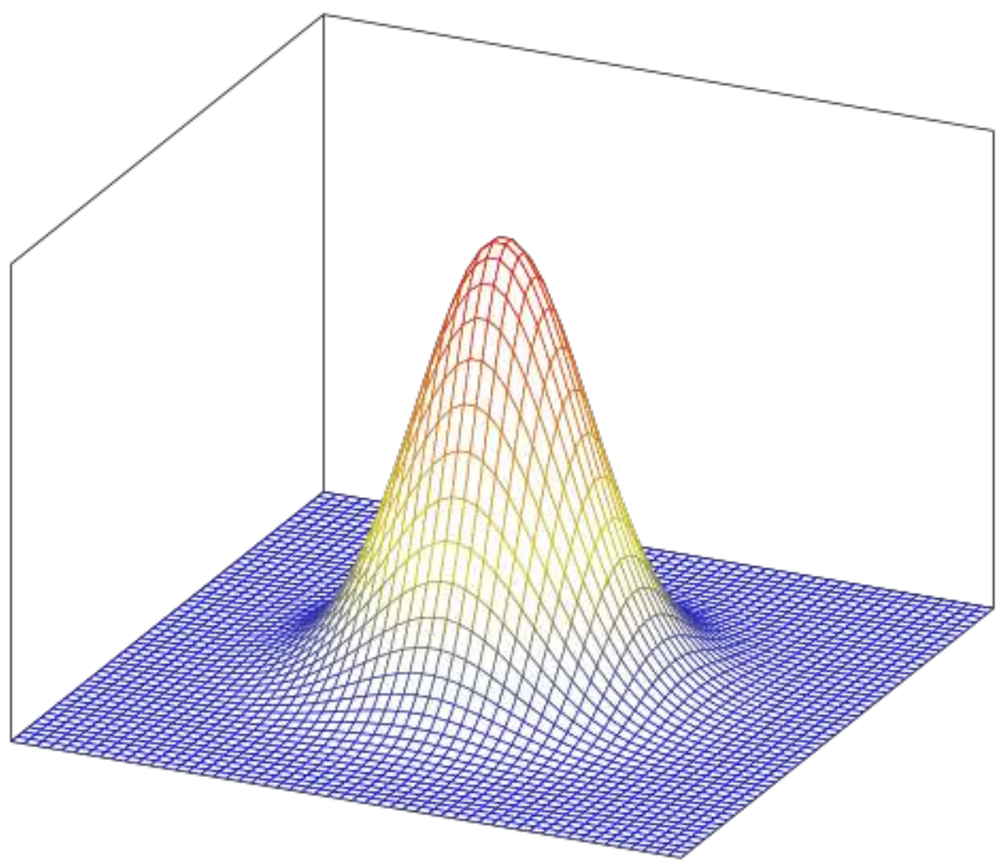
- Аномалия — объект из другого распределения
- Параметрические методы восстановления

НЕПАРАМЕТРИЧЕСКОЕ ВОССТАНОВЛЕНИЕ ПЛОТНОСТИ

ПАРАМЕТРИЧЕСКОЕ ВОССТАНОВЛЕНИЕ

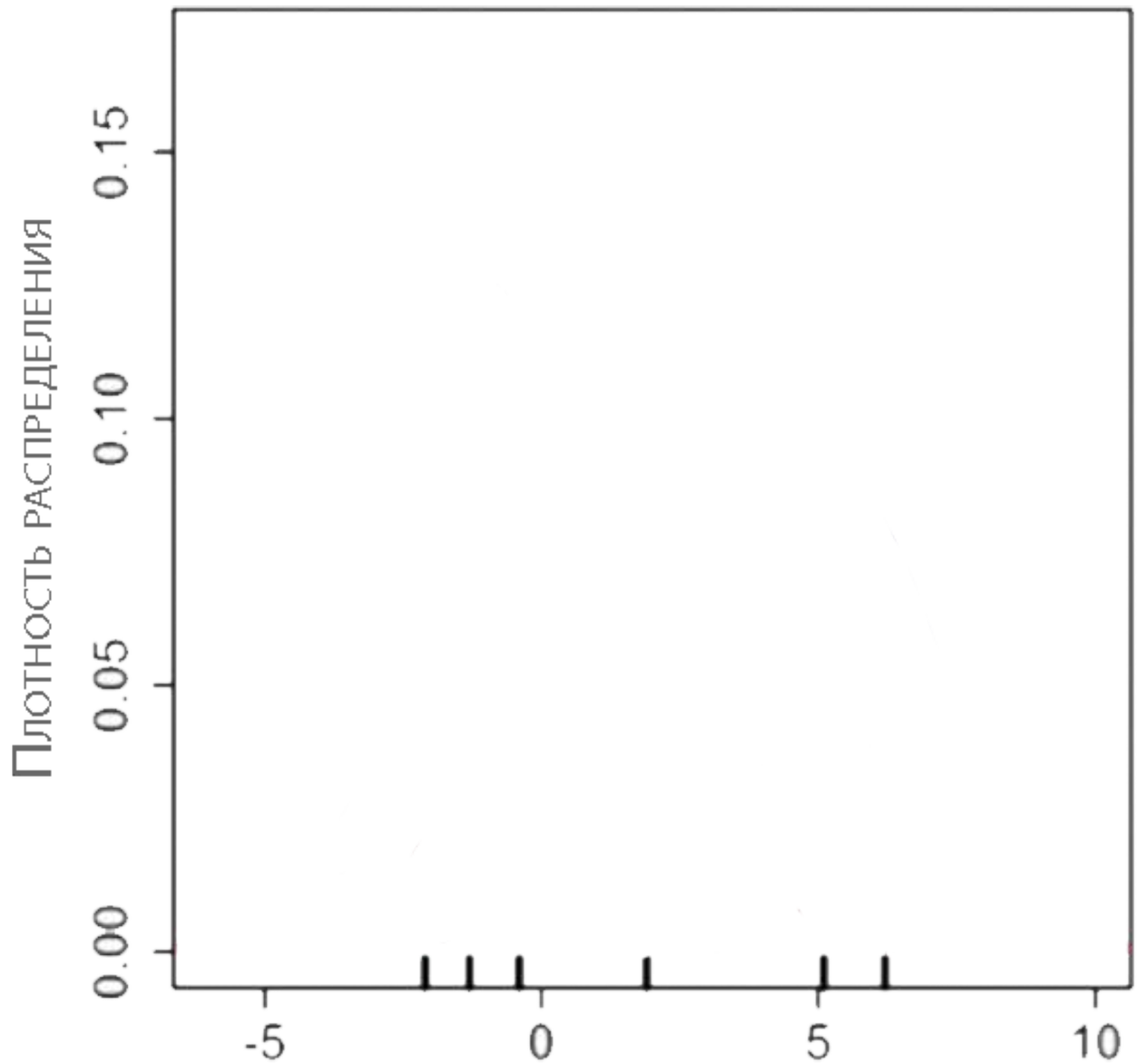
$$\phi(x|\theta) = \mathcal{N}(\mu, \Sigma)$$

- Нормальное распределение
- Параметры: $\theta = (\mu, \Sigma)$



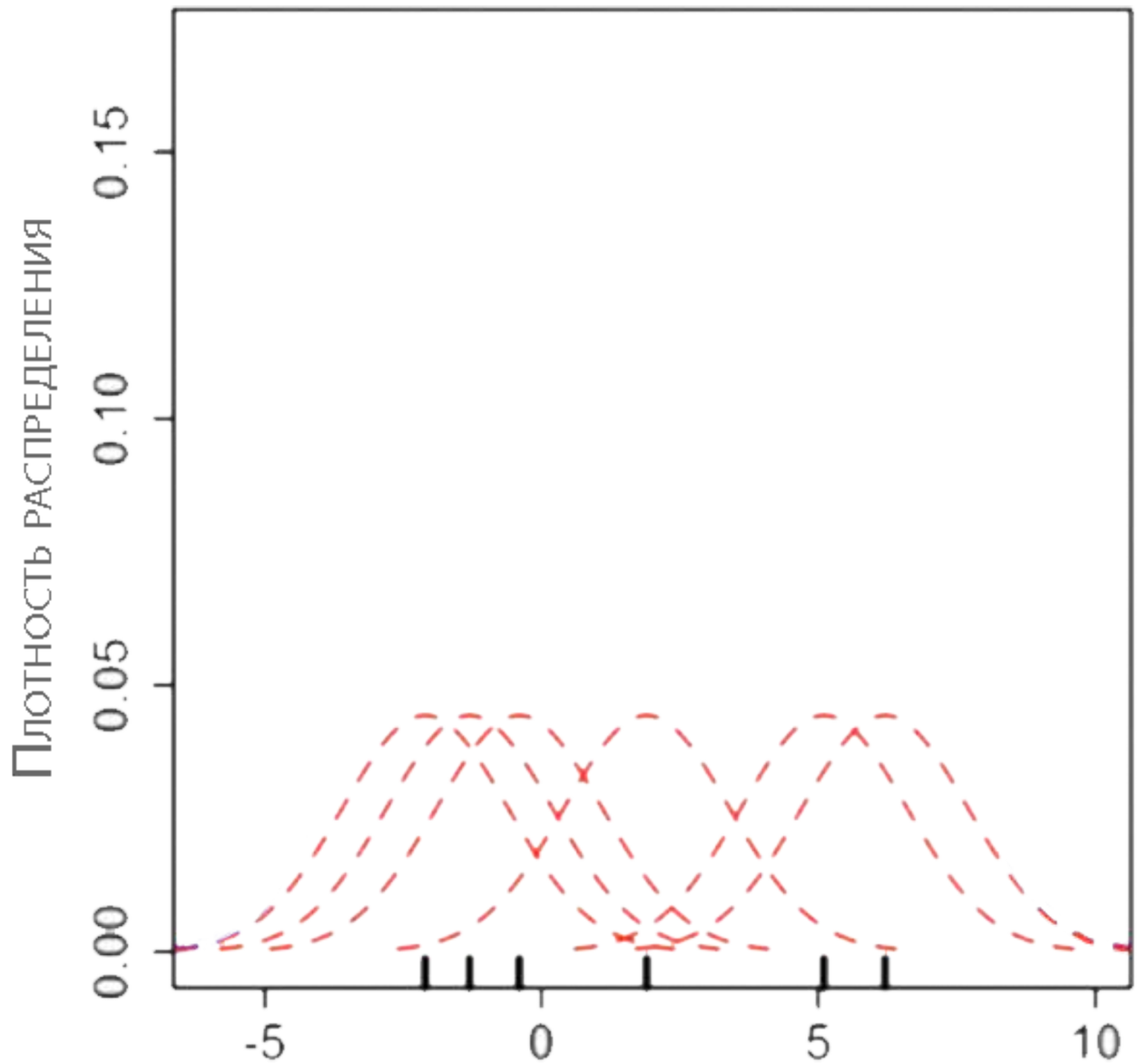
МФТИ

НЕПАРАМЕТРИЧЕСКОЕ ВОССТАНОВЛЕНИЕ



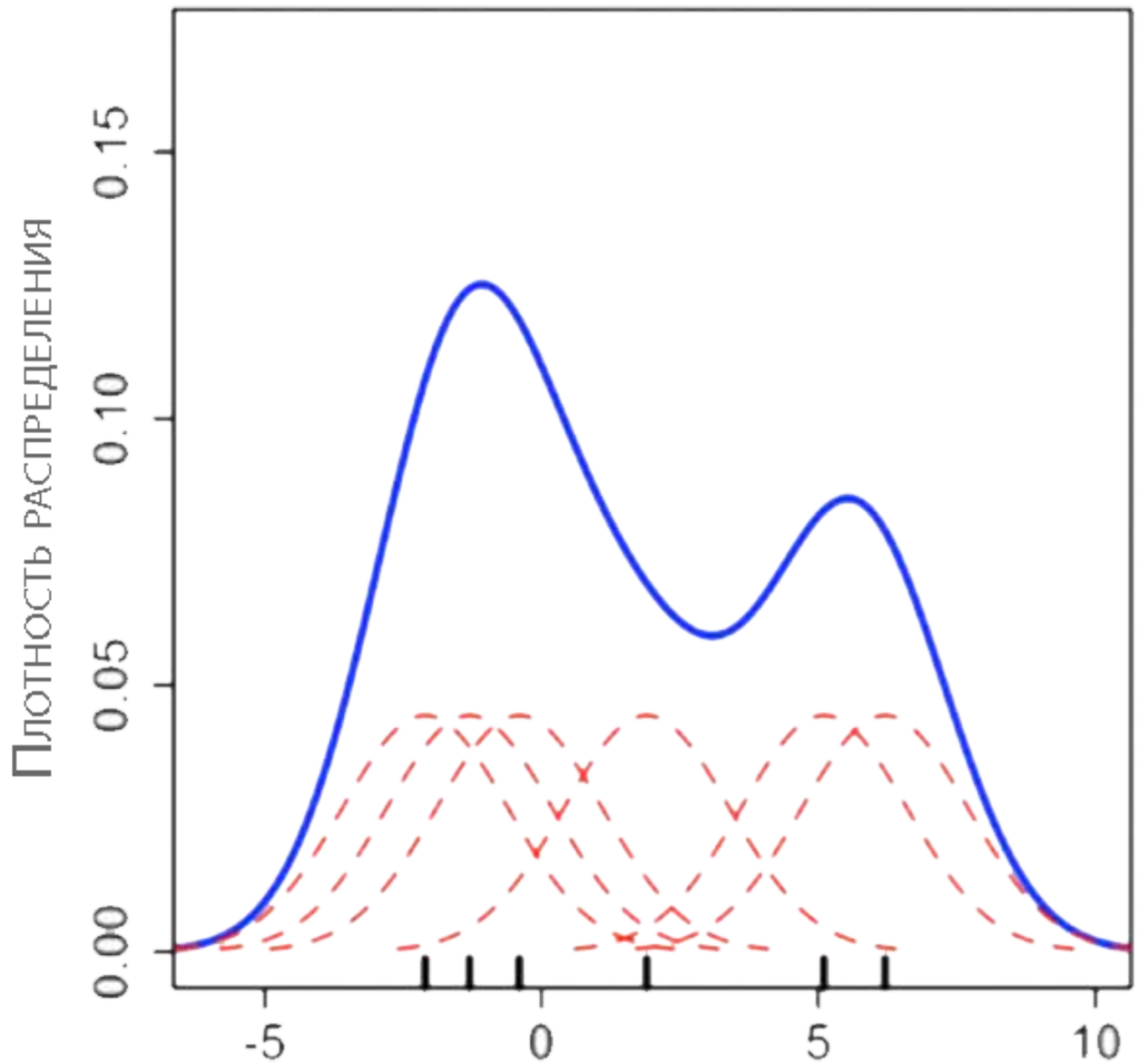
МФТИ

НЕПАРАМЕТРИЧЕСКОЕ ВОССТАНОВЛЕНИЕ



МФТИ

НЕПАРАМЕТРИЧЕСКОЕ ВОССТАНОВЛЕНИЕ



МФТИ

ОЦЕНКА ПАРЗЕНА-РОЗЕНБЛАТТА

$$p_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right)$$

ОЦЕНКА ПАРЗЕНА-РОЗЕНБЛАТТА

$$p_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right)$$

Ядро

ОЦЕНКА ПАРЗЕНА-РОЗЕНБЛАТТА

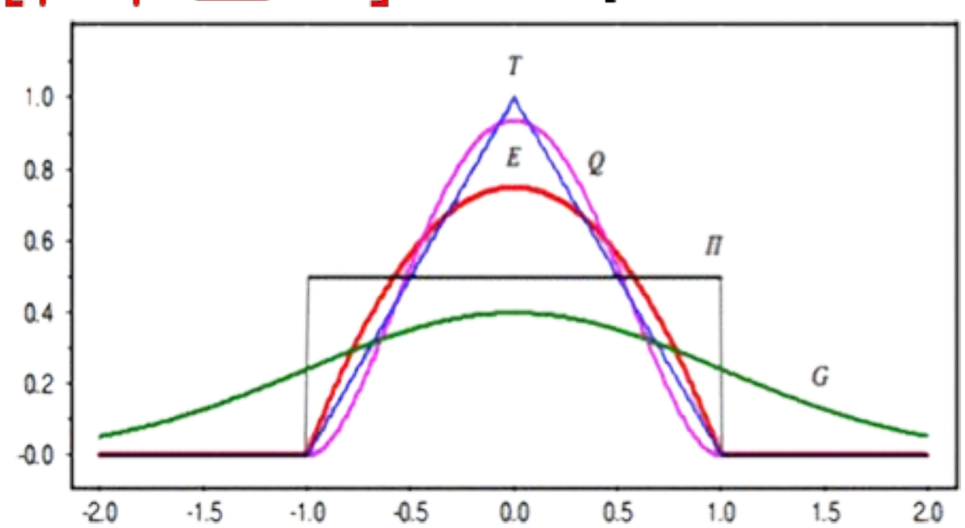
$$p_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right)$$

- $K(r)$ — ядро
- Чётная функция
- $\int K(r)dr = 1$

МФТИ

ЯДРО

- $E(r) = \frac{3}{4}(1 - r^2)[|r| \le 1]$ — оптимальное
- $Q(r) = \frac{15}{16}(1 - r^2)^2[|r| \le 1]$ — квартическое
- $T(r) = (1 - |r|)[|r| \le 1]$ —треугольное
- $G(r) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}r^2\right)$ —гауссовское
- $\Pi(r) = \frac{1}{2}[|r| \le 1]$ —прямое



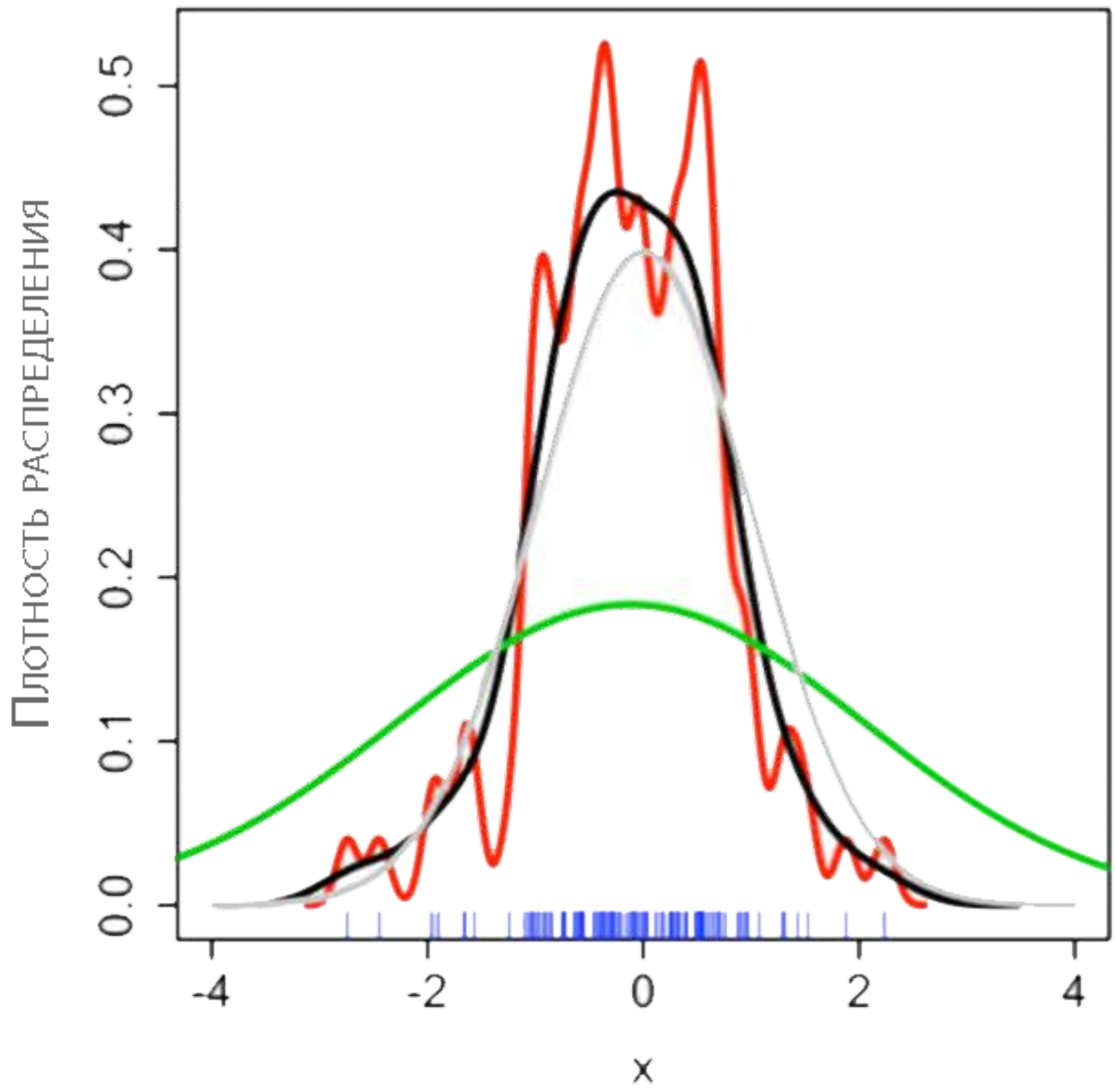
МФТИ

ОЦЕНКА ПАРЗЕНА-РОЗЕНБЛАТТА

$$p_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right)$$

- h — ширина окна

ШИРИНА ОКНА



МФТИ

МНОГОМЕРНЫЙ СЛУЧАЙ

$$p_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)$$

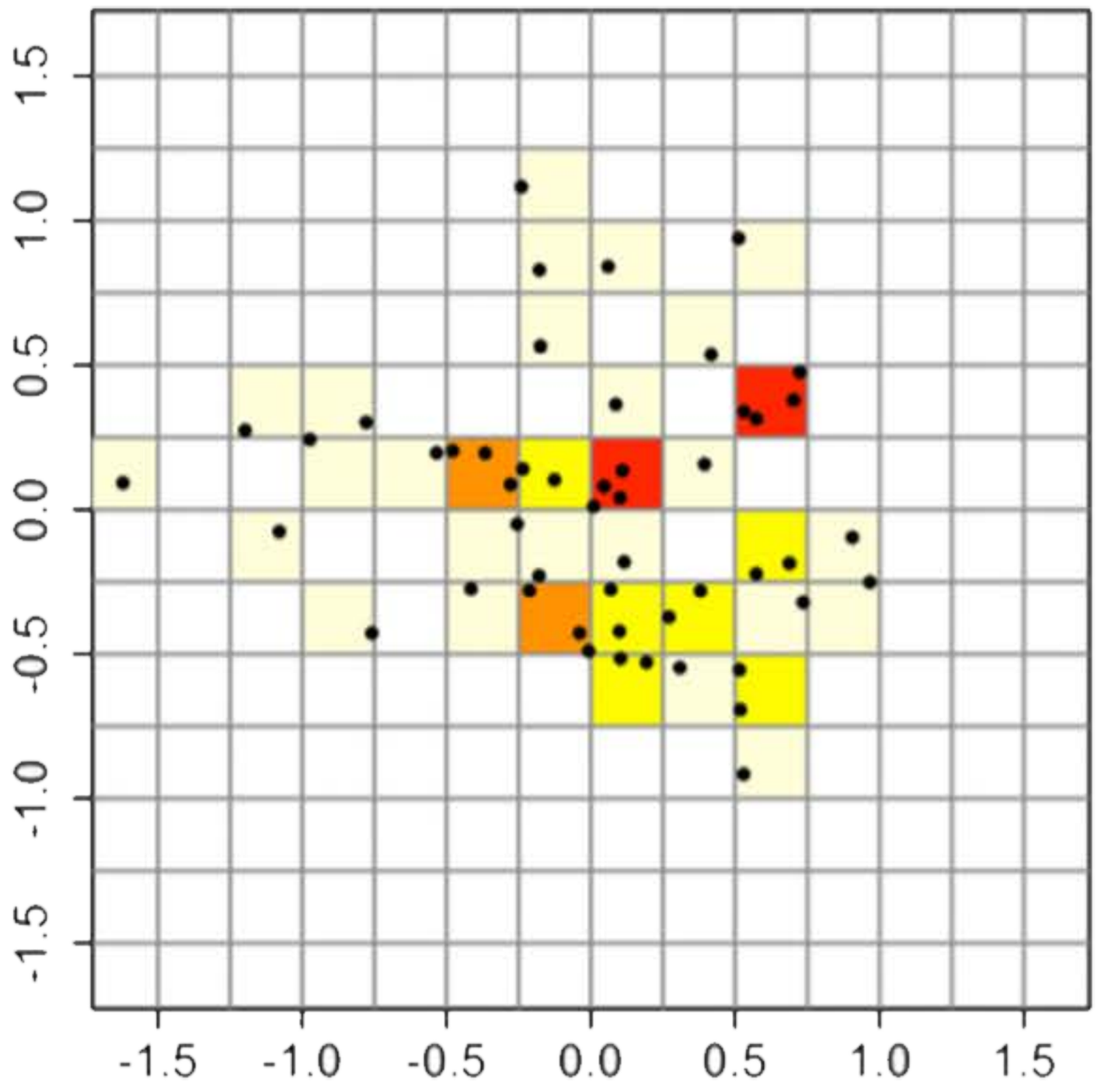
› $V(h) = \int K\left(\frac{\rho(x, x_i)}{h}\right) dx$ —

нормировочная константа

МФТИ

МНОГОМЕРНЫЙ СЛУЧАЙ

- Чем выше размерность, тем больше объектов нужно



МФТИ

МНОГОМЕРНЫЙ СЛУЧАЙ

- Число объектов, необходимых для качественного оценивания, растёт экспоненциально с ростом размерности

ОБНАРУЖЕНИЕ АНОМАЛИЙ

- Новый объект x
- Если $p(x) < t$, то это аномалия
- Порог t:
  - Из априорных соображений
  - По известным аномалиям

МФТИ

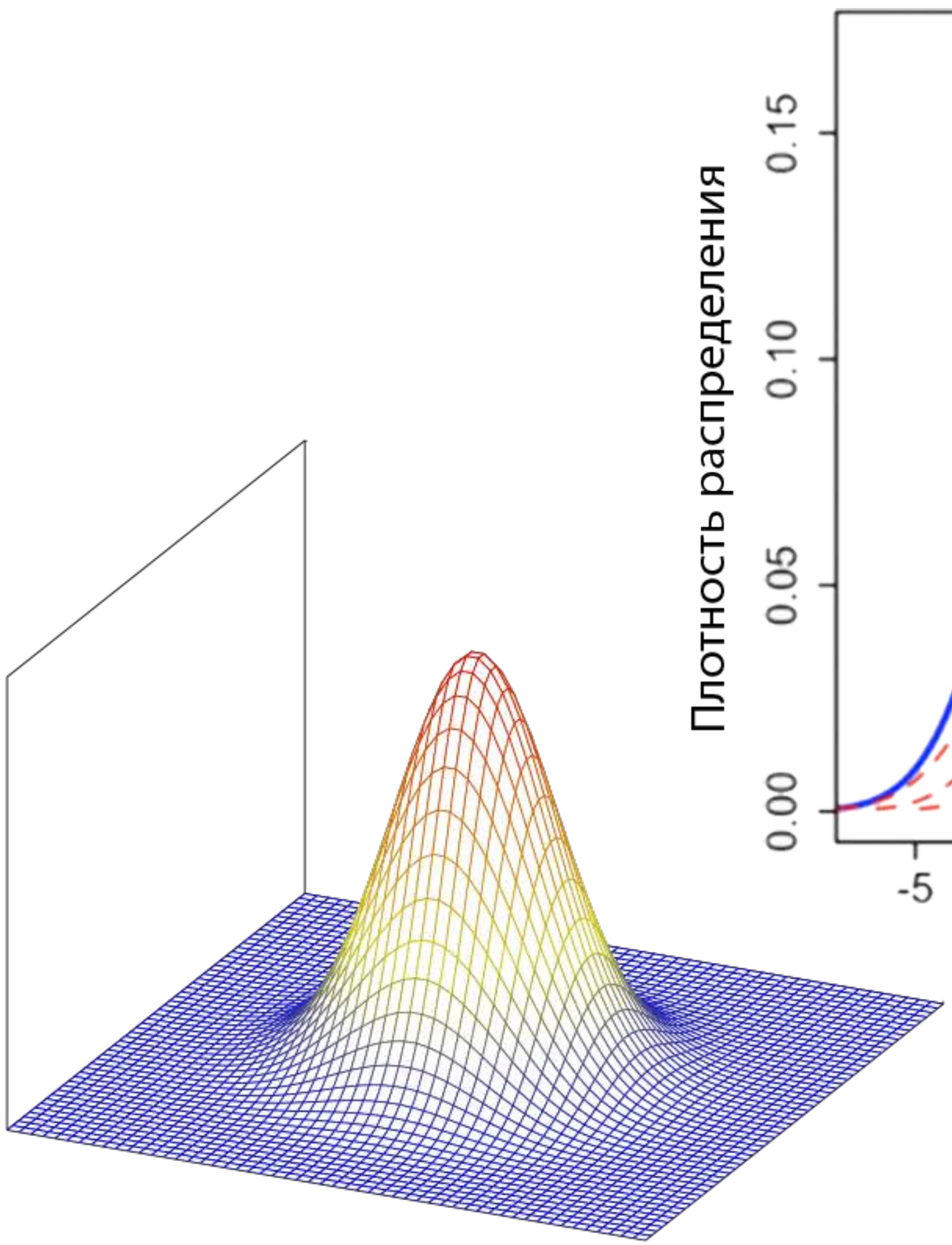
РЕЗЮМЕ

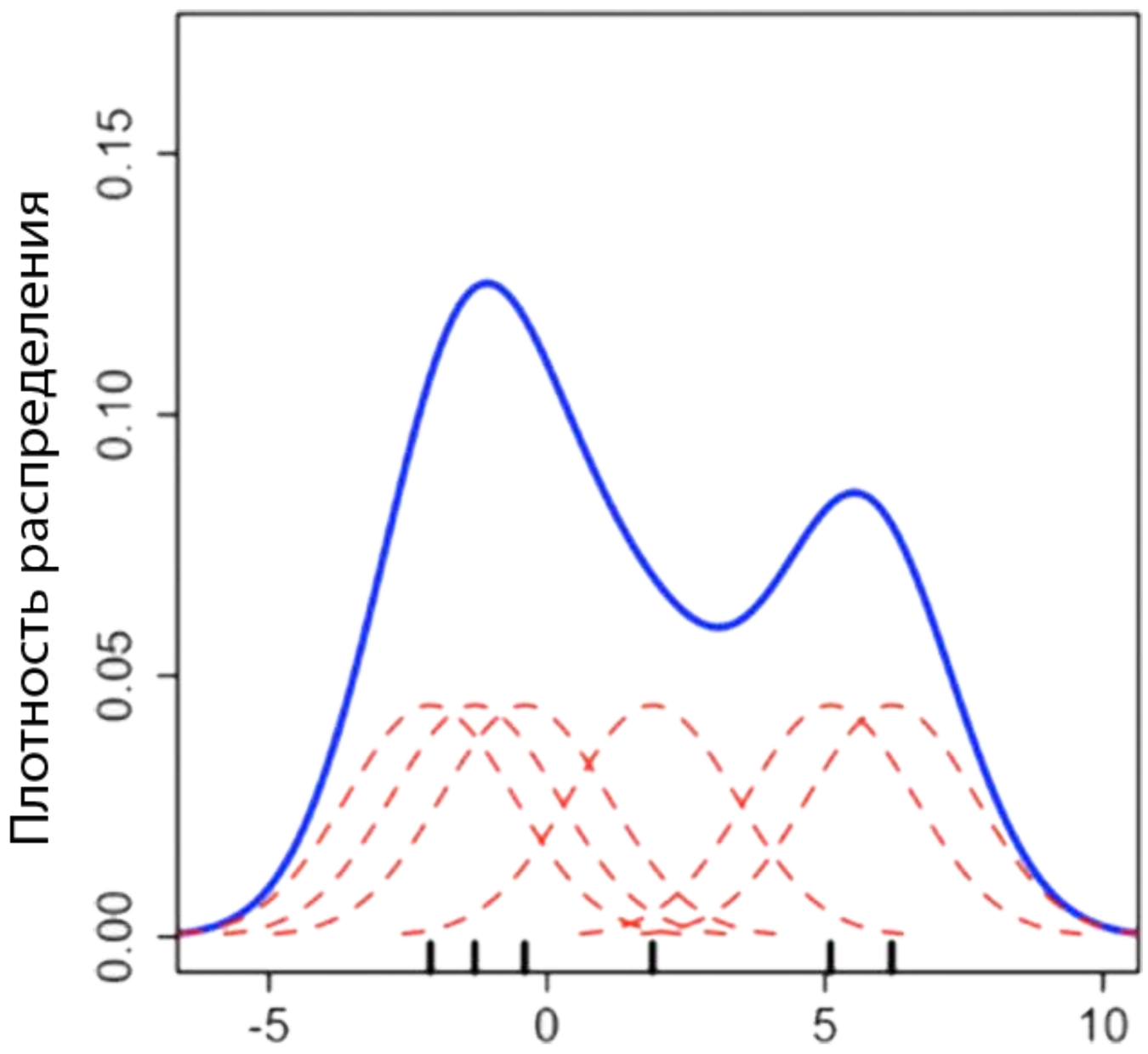
- Непараметрический подход — для восстановления сложных плотностей
- Параметры: ядро и ширина окна
- В многомерно случае требуется большая выборка

МФТИ

ОДНОКЛАССОВЫЙ SVM

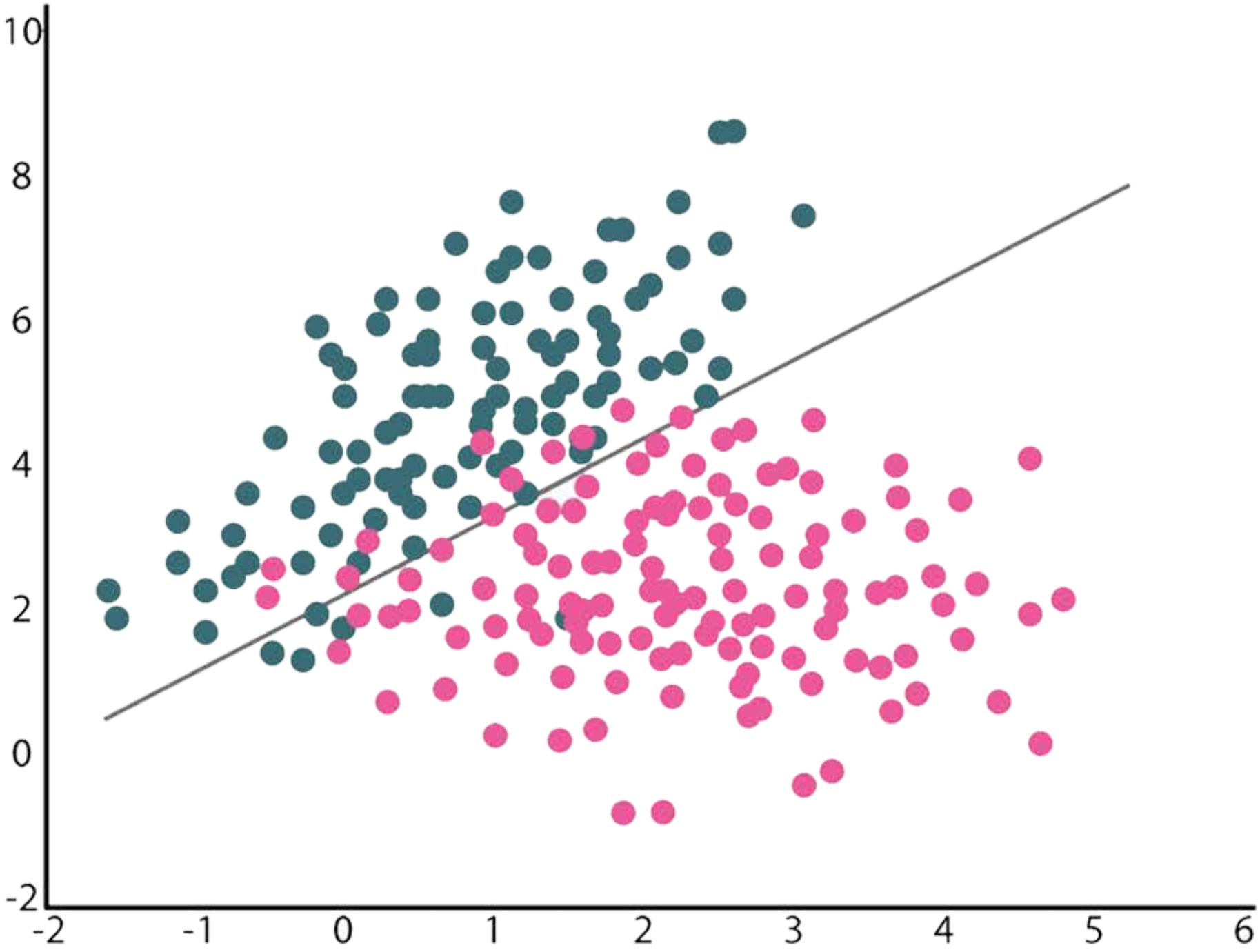
ВОССТАНОВЛЕНИЕ ПЛОТНОСТИ





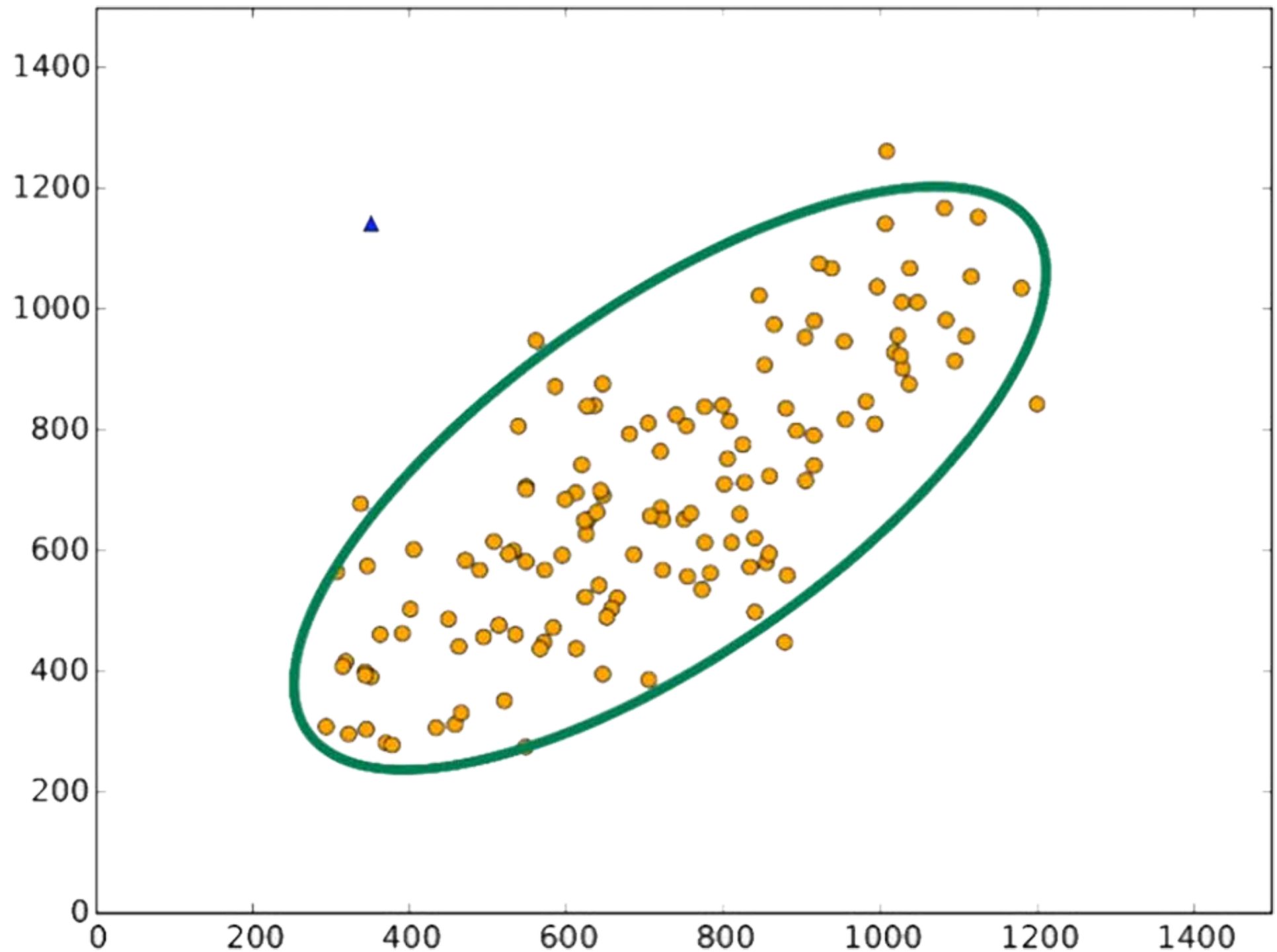
МФТИ

КЛАССИФИКАЦИЯ



МФТИ

ОБНАРУЖЕНИЕ АНОМАЛИЙ



МФТИ

ОБНАРУЖЕНИЕ АНОМАЛИЙ

- Обучающая выборка — нормальные объекты
- Начало координат — аномалия
- Задача: отделить выборку гиперплоскостью от нуля
- Максимизация отступа

МФТИ

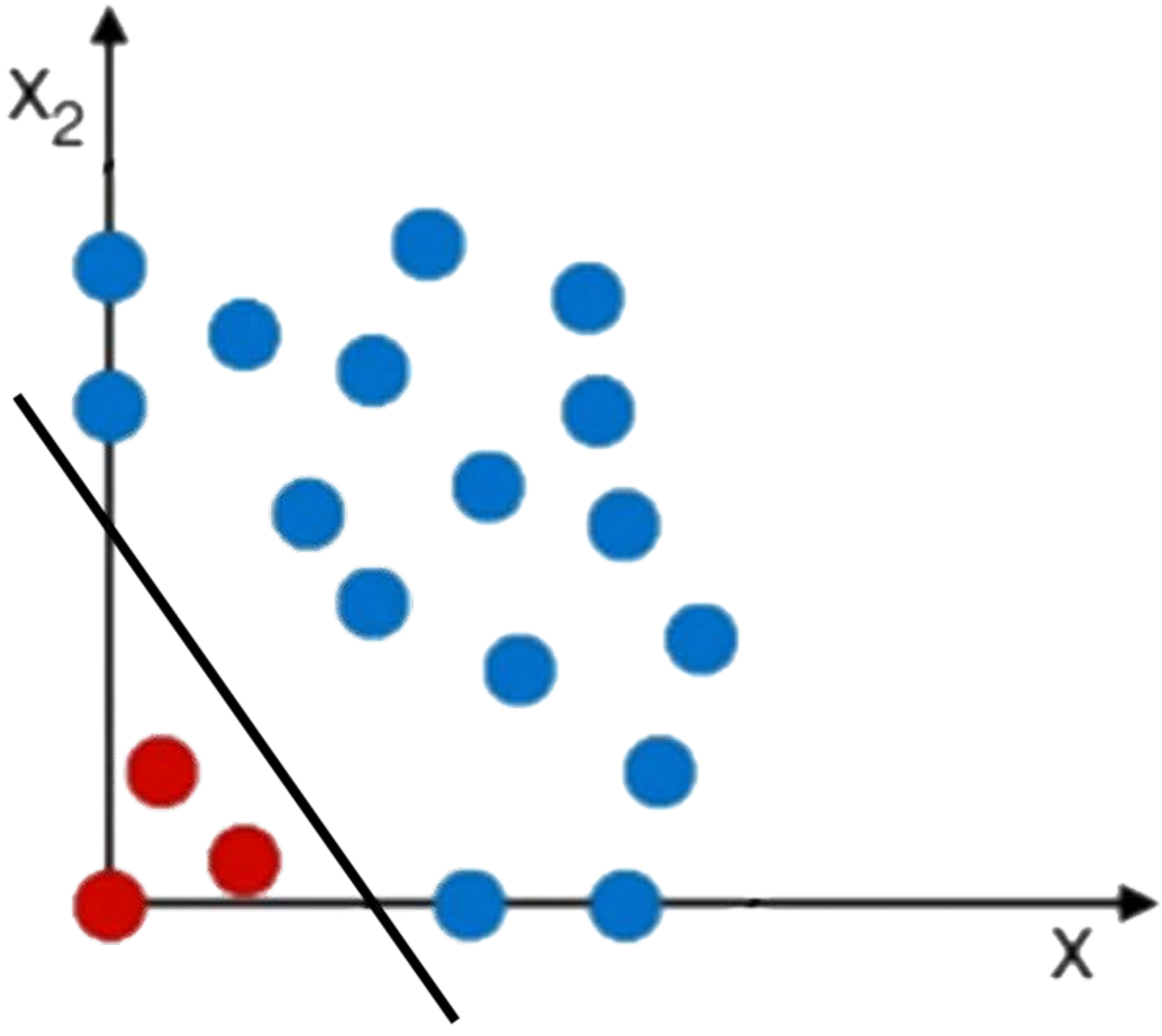
ОДНОКЛАССОВЫЙ SVM

$$\begin{cases} \frac{1}{2}\| \mathbf{w} \|^2 + \frac{1}{v\ell}\sum_{i=1}^{\ell}\xi_i - \rho \to \min\limits_{\mathbf{w},\xi,\rho} \\ \langle \mathbf{w}, x_i \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0 \end{cases}$$

- v — верхняя оценка на долю аномалий на выборке

МФТИ

ОДНОКЛАССОВЫЙ SVM



МФТИ

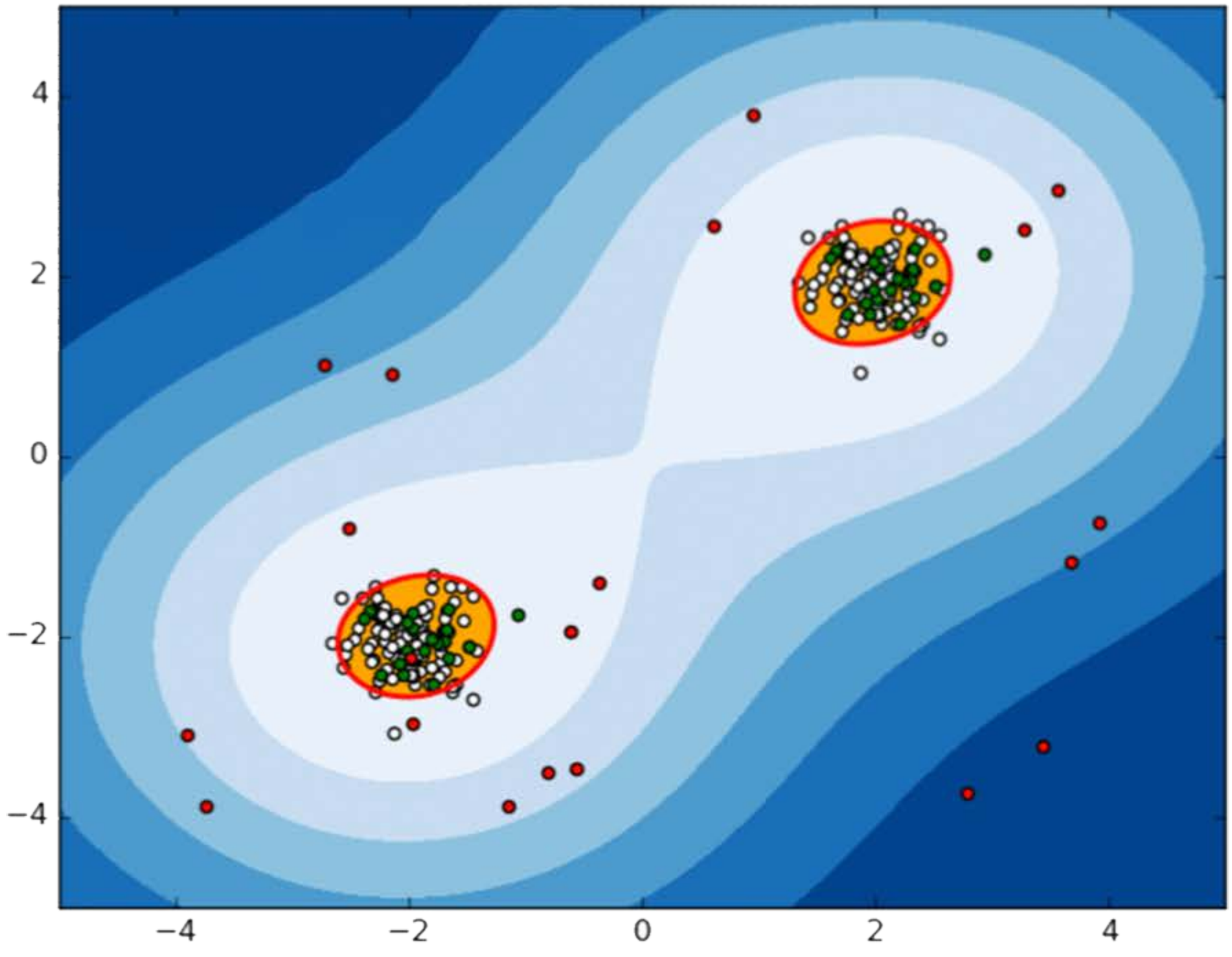
ЯДРОВОЙ ПЕРЕХОД

- SVM позволяет строить нелинейные разделяющие поверхности
- Ядровой переход (kernel trick)
- Популярный выбор: RBF-ядро

$$K(x, z) = \exp\left(\frac{\|x - z\|^2}{\sigma^2}\right)$$

МФТИ

ЯДРОВОЙ ПЕРЕХОД



МФТИ

РЕЗЮМЕ

- Одноклассовый SVM отделяет выборку от начала координат
- Имеет смысл при использовании ядер