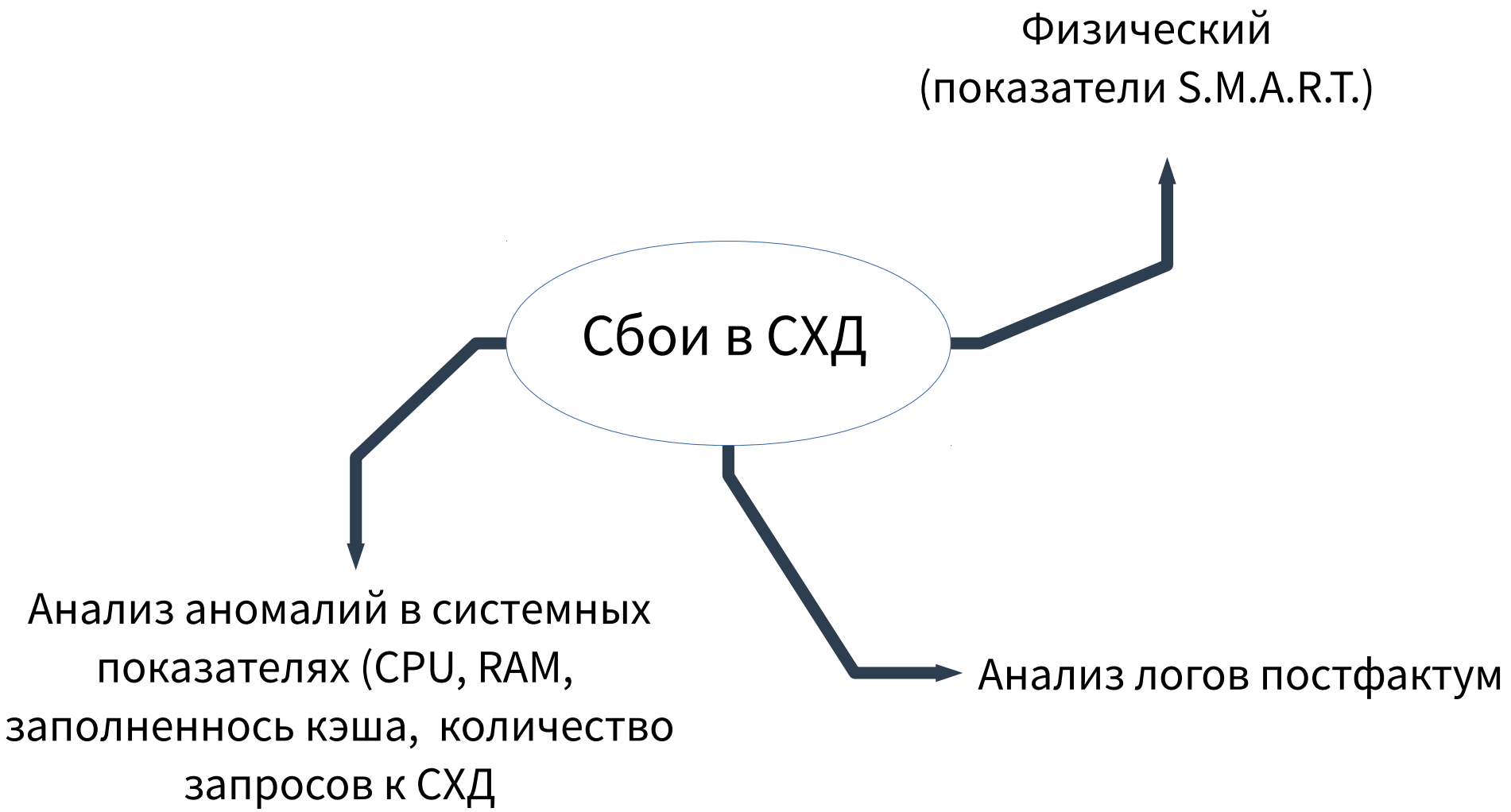
RAIDIX

Использование методов машинного обучения для предсказания сбоев в системах хранения данных

Автор: Васенина Анна Игоревна, 344 группа
Научный руководитель: д.ф.-м.н., проф. Терехов А.Н.
Консультант: руководитель исследовательской лаборатории RAIDIX к.т.н. Лазарева С.В.

Санкт-Петербургский государственный университет
Кафедра системного программирования
22 мая 2018

Введение


Физический
(показатели S.M.A.R.T.)
Сбои в СХД
Анализ аномалий в системных показателях (CPU, RAM, заполненность кэша, количество запросов к СХД
Анализ логов постфактум

Постановка задачи

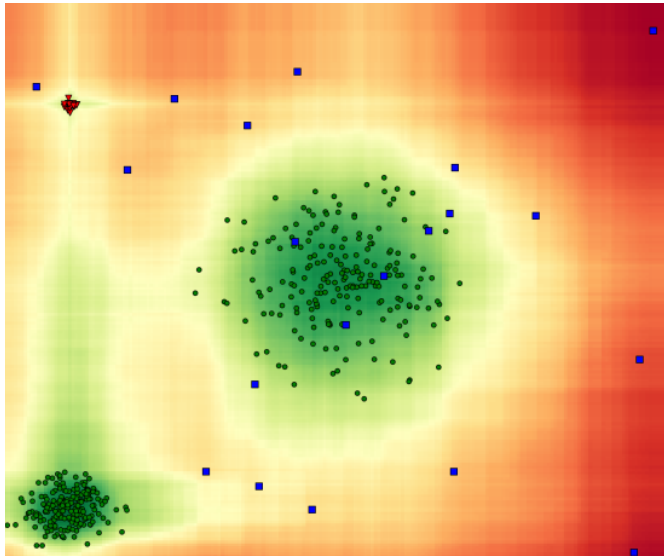
Целью работы является исследование применимости алгоритмов машинного обучения на базе ансамблей решающих деревьев к анализу логов, собираемых в системах хранения данных RAIDIX

Задачи:

- Изучить существующие методы машинного обучения на базе решающих деревьев
- Выработать методику анализа сообщений в логах
- Подготовить обучающую выборку
- Провести эксперименты
- Проанализировать результаты

Методы машинного обучения

Наша задача — бинарная классификация



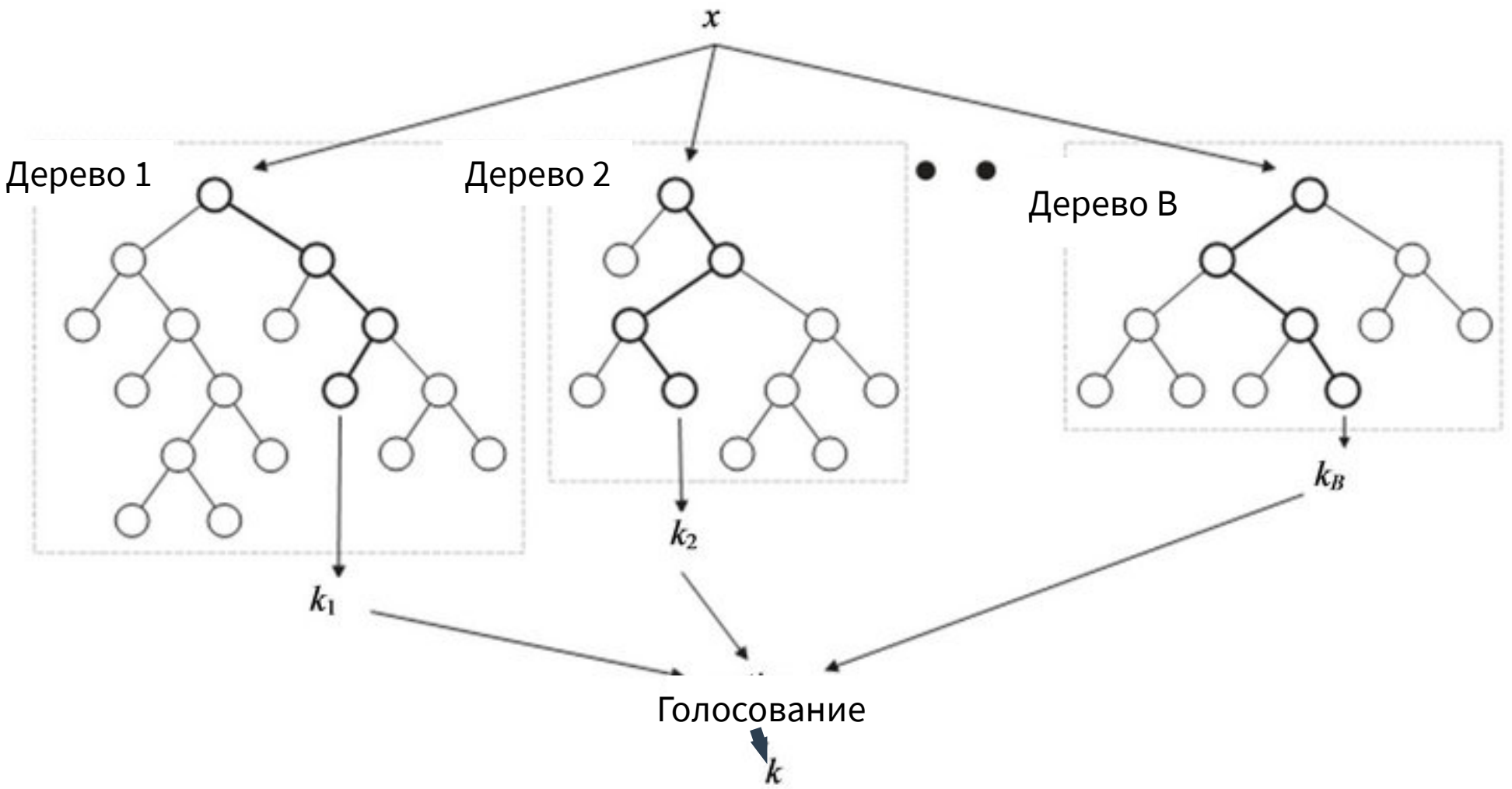
- Без учителя
  - Алгоритмы поиска аномалий (Isolation forest)

  Для обучения нужен только вектор признаков X

- С учителем
  - Алгоритмы классификации (Random Forest, XGBoost)

  Для обучения нужен вектор признаков X и ответ У

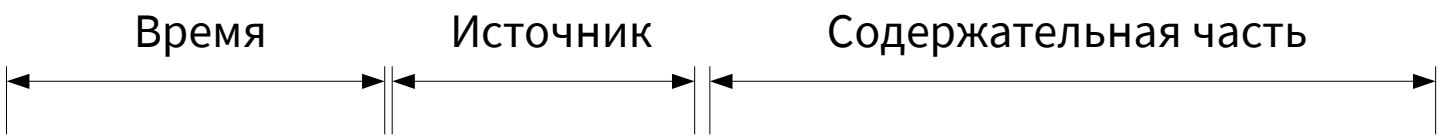
Леса бинарных решающих деревьев



Описание решения

Данные:

- Всего строк — 772 000
- Строк с ошибками — 25 000 (~3%)
- Общее время логов — 155 дней



Aug 31 19:09:27 GRX3 kernel: unable to read partition table

Строки находятся во временной зависимости друг от друга

Описание решения

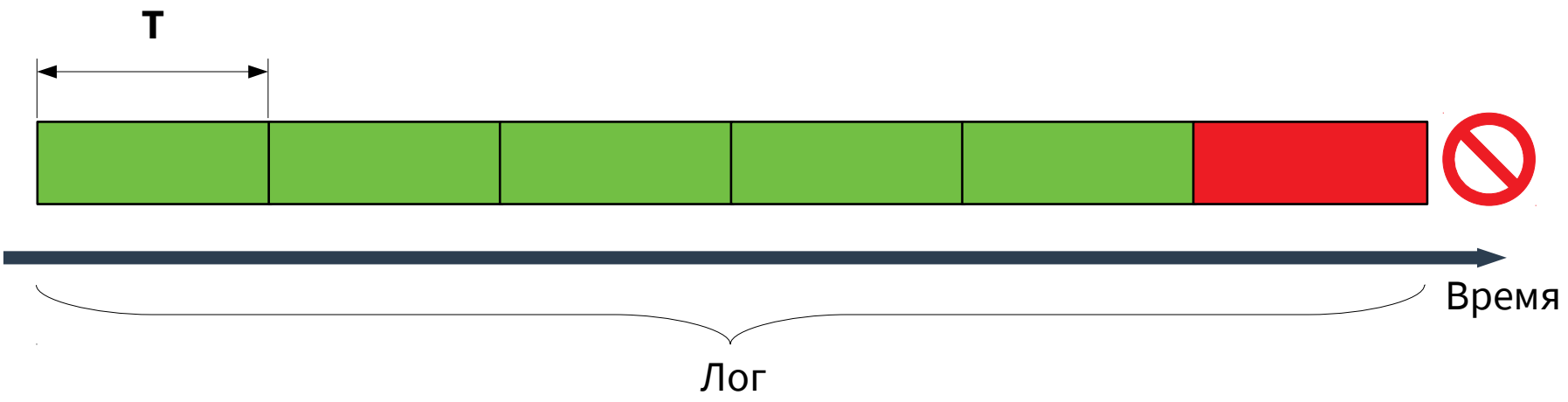
Первая мысль — анализ отдельных сообщений

Проблемы:

- Сложные связи
- Недостаточное количество данных

Описание решения

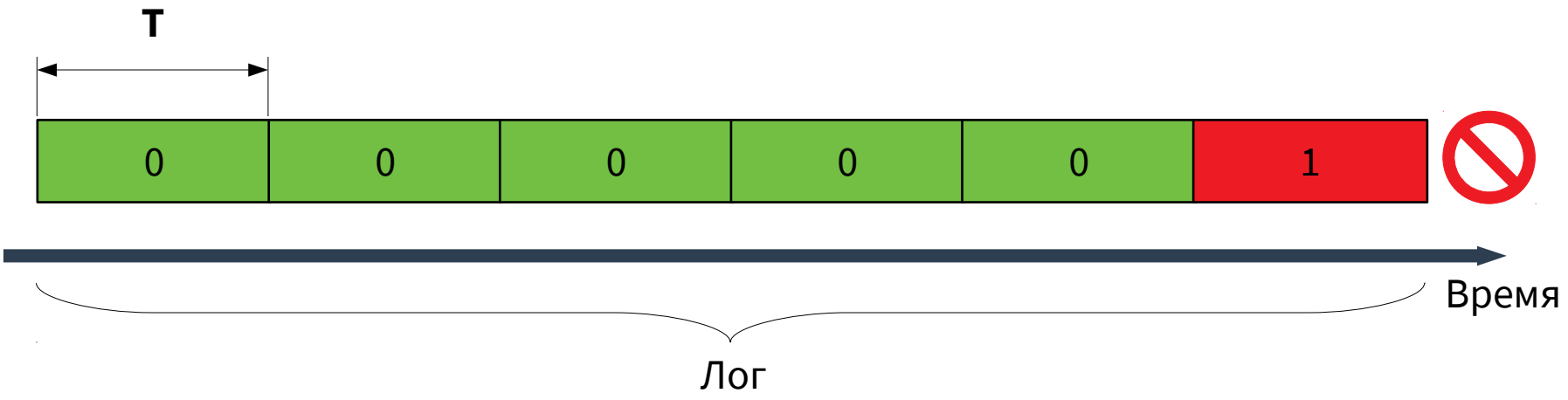
Для решения поставленной задачи разобьём логи на промежутки по времени T



Результат разбиения — набор сообщений (возможно, пустой)

Описание решения

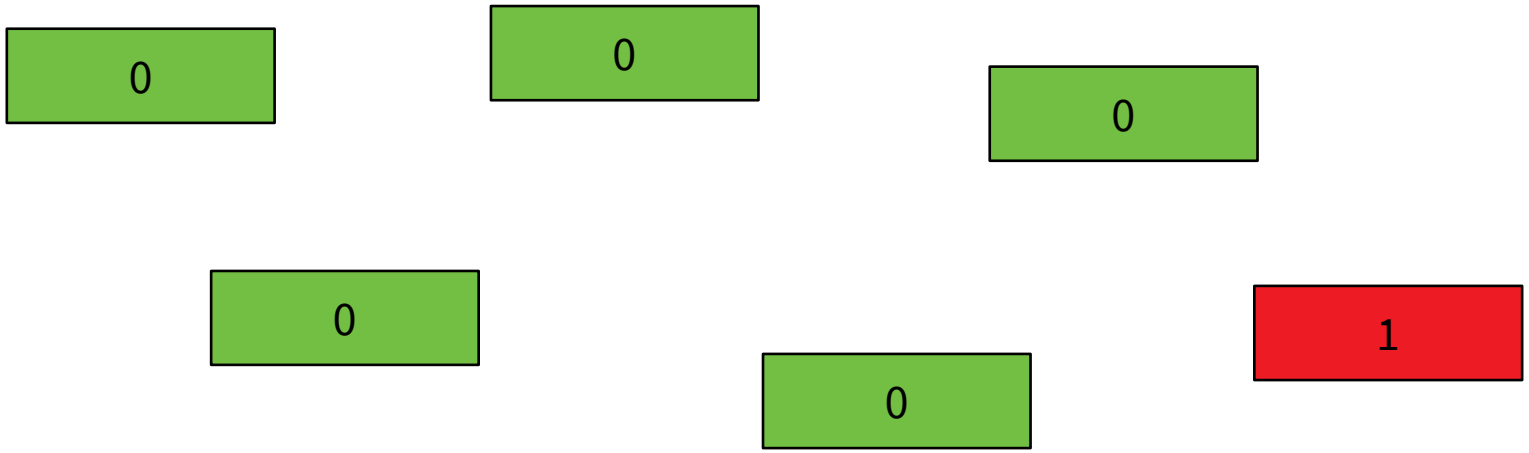
Для решения поставленной задачи разобьём логи на промежутки по времени T



Результат разбиения — набор сообщений (возможно, пустой)

Описание решения

Для решения поставленной задачи разобьём логи на промежутки по времени T



Результат разбиения — набор сообщений (возможно, пустой)

Выделение признаков

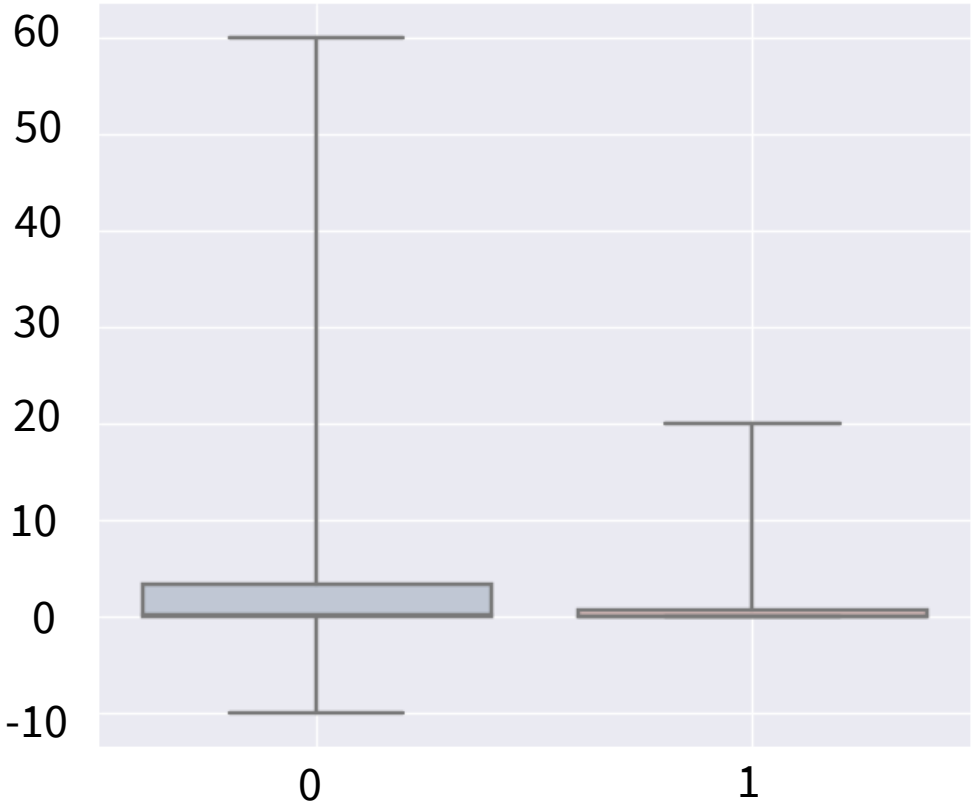
Вектором X для набора сообщений назовём следующую совокупность:

- Процентное содержание плохих сообщений
- Среднее расстояние между плохими/хорошими сообщениями (в минутах)
- Среднее количество хороших сообщений между плохими
- Однородность плохих/хороших сообщений

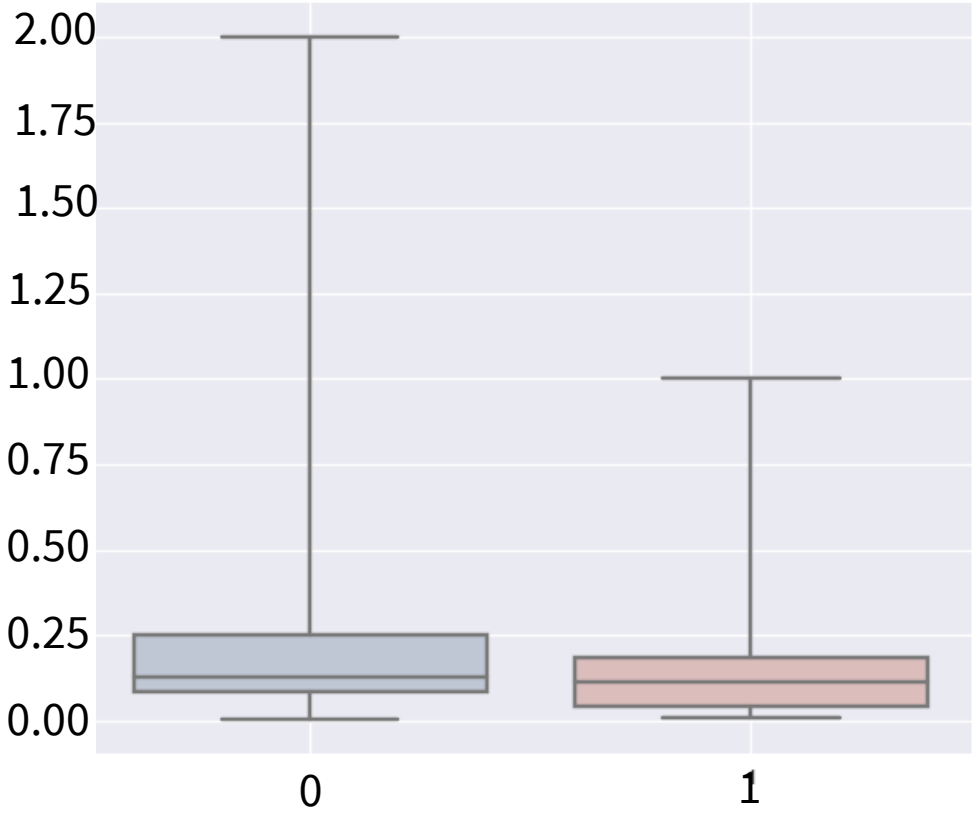
$$\text{Однородность} = \max \left(\frac{\text{Число повторений сообщения M}}{\text{Число всех сообщений}} \right)$$

Результатом Y: 0, если в следующем участке не было сбоя
1, если в следующем участке был сбой

Различие между классами по признакам (пример, T = 2 часа)



Среднее время между хорошими сообщениями



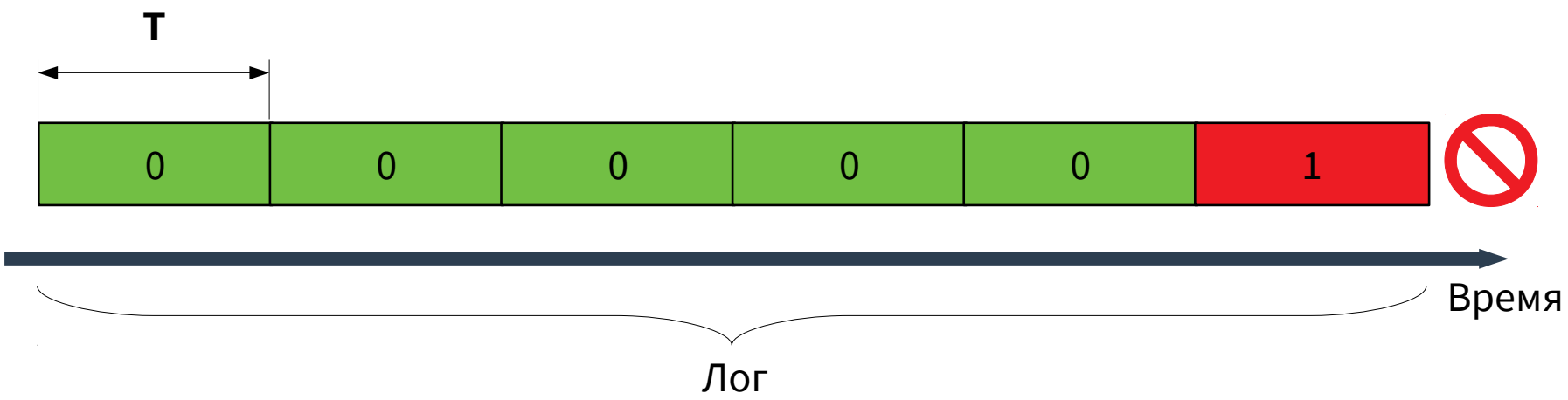
Однородность хороших сообщений

Критерии оценки

$$TNR = \frac{TN}{TN + FP} \qquad (1)$$

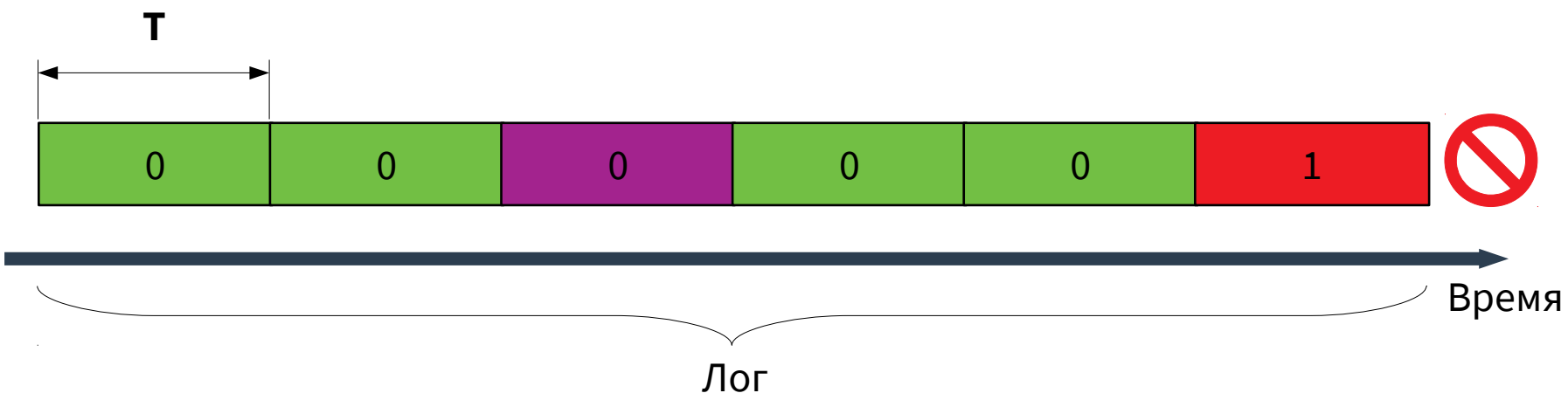
$$TPR = \frac{TP}{TP + FN} \qquad (2)$$

Формирование базы векторов



- **Мало плохих векторов**
  - 80% в обучающую, 20% в тестирующую
- **Много хороших**
  - 50% в обучающую, 50% в тестирующую

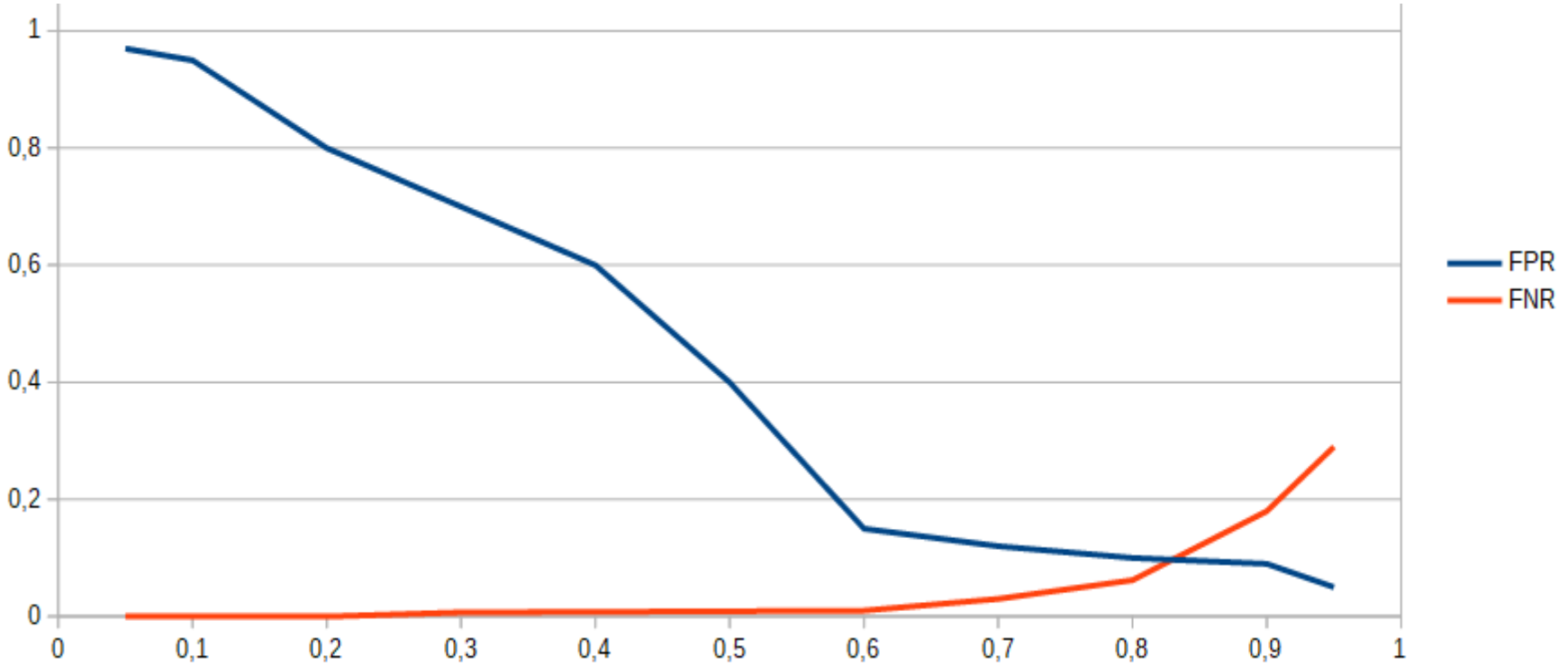
Формирование базы векторов



0 **- один из «ключевых» векторов**

- **Мало плохих векторов**
  - 80% в обучающую, 20% в тестирующую
- **Много хороших**
  - 50% в обучающую, 50% в тестирующую

Выбор порога подозрительности



Зависимость вероятности выдачи ложного положительного и ложного отрицательного прогнозов в зависимости от барьера вероятности, после которого прогноз считается положительным

Оценка

Язык — Python
Библиотеки — Sklearn, XGBoost

Алгоритм	Random Forest	XGBoost	Isolation Forest
TNR	0,95	0,95	0,62
TPR	0,7	0,72	0,87

Результаты

- Изучены существующие методы машинного обучения на базе решающих деревьев
- Выработана методика анализа сообщений в логах
- Подготовлена обучающая выборка
- Проведены эксперименты
    - Получен результат, удовлетворяющий заказчика