

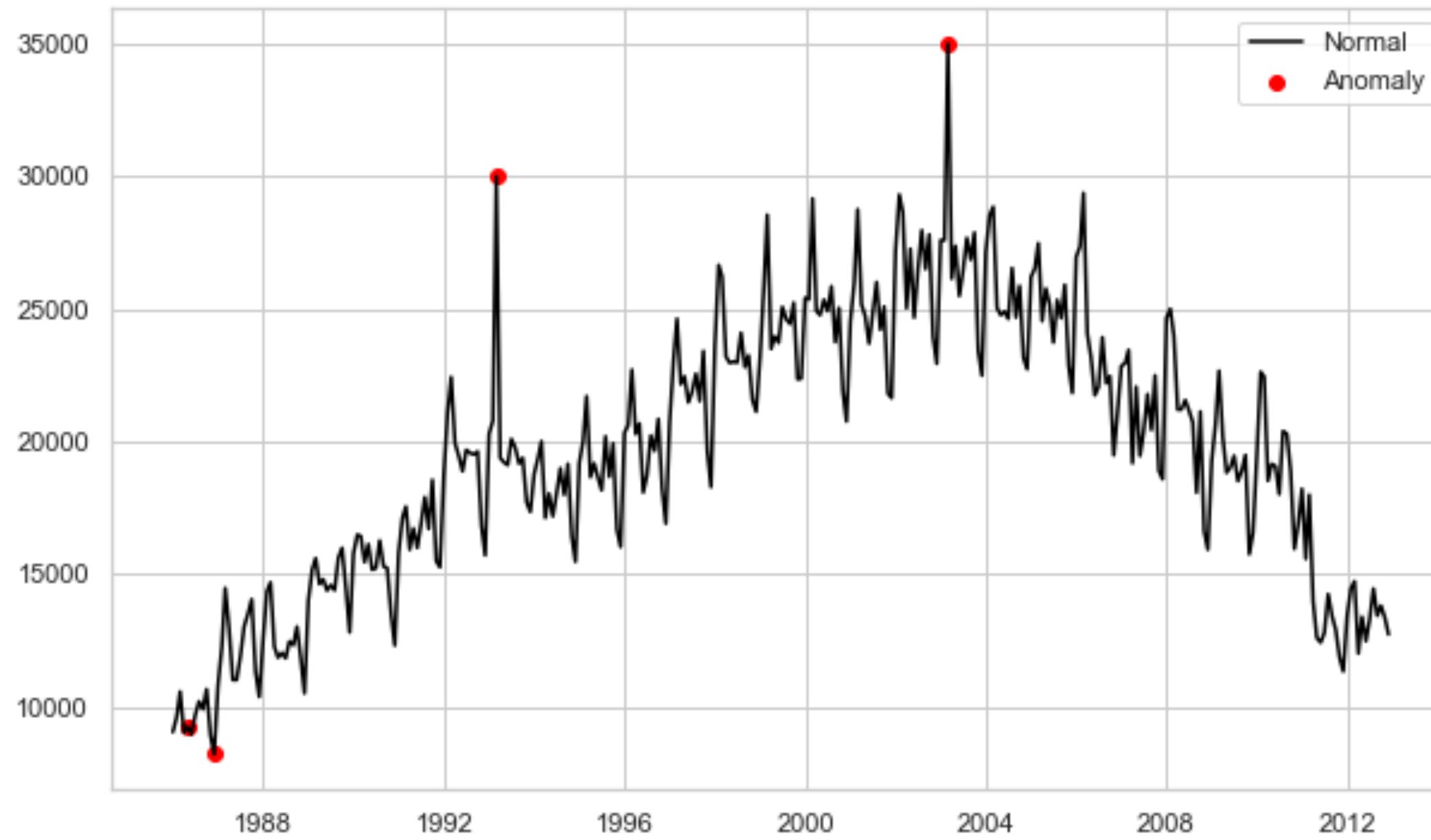
Поиск аномалий и точек смены поведения

День 3

Введение

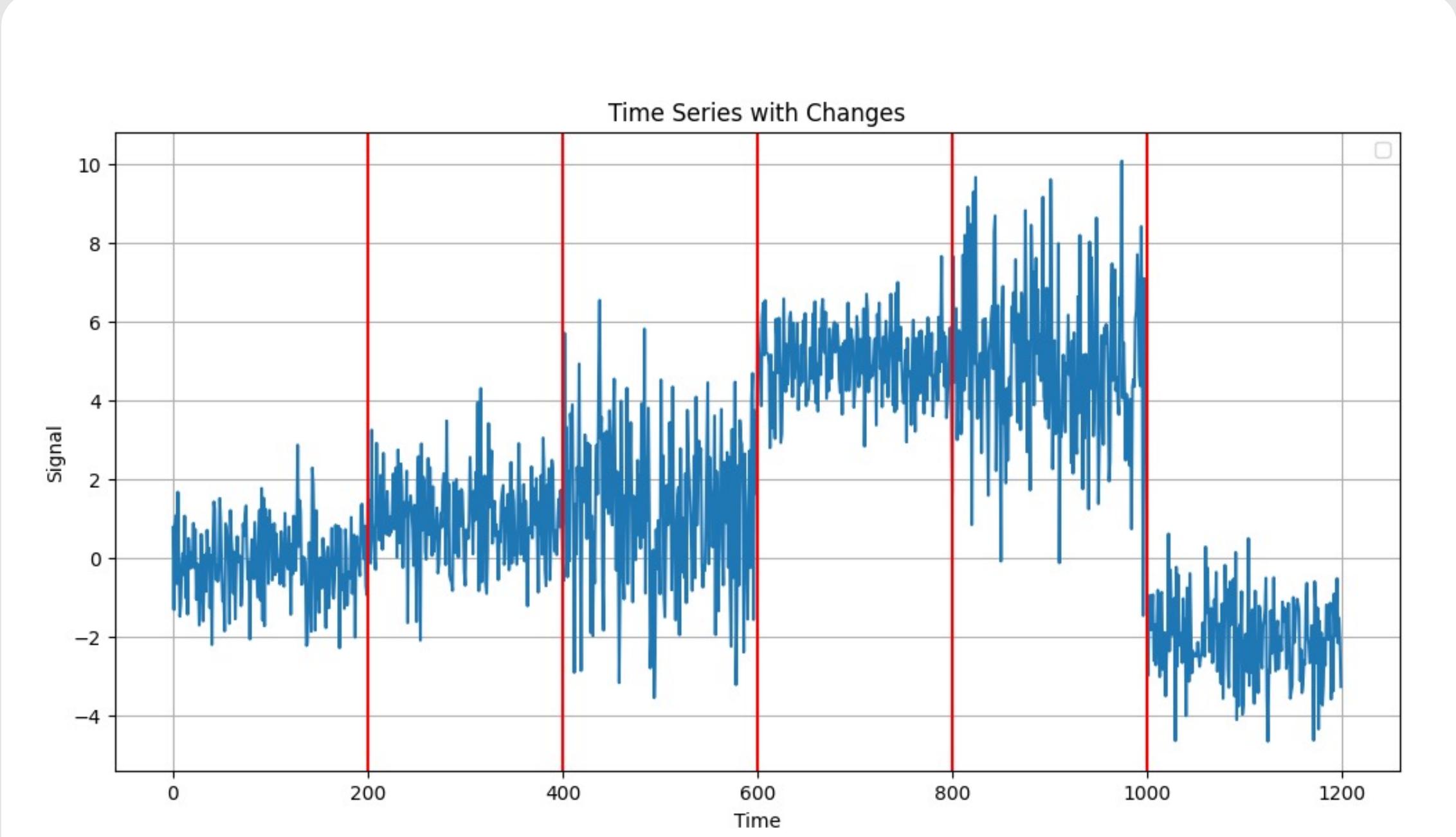
Откуда взялась такая задача?

Аномальное поведение



Аномальные точки

Произошло что-то непонятное
единоразово.

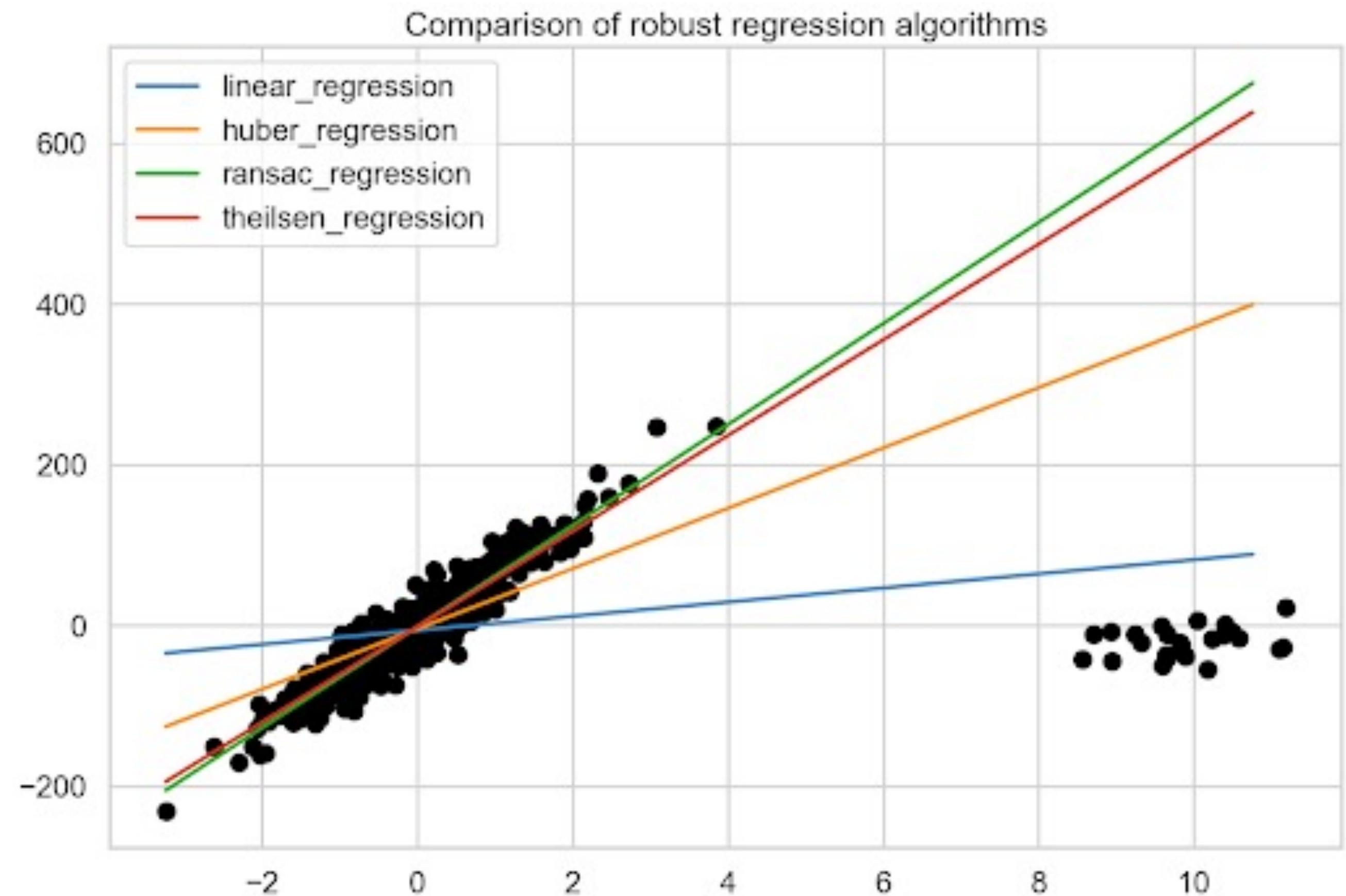


Точки смены поведения

Произошло что-то непонятное, и ряд
начал вести себя по-другому.

Задача прогнозирования

Идея: удаление аномалий из данных улучшает качество решения целевой задачи (прогнозирования).

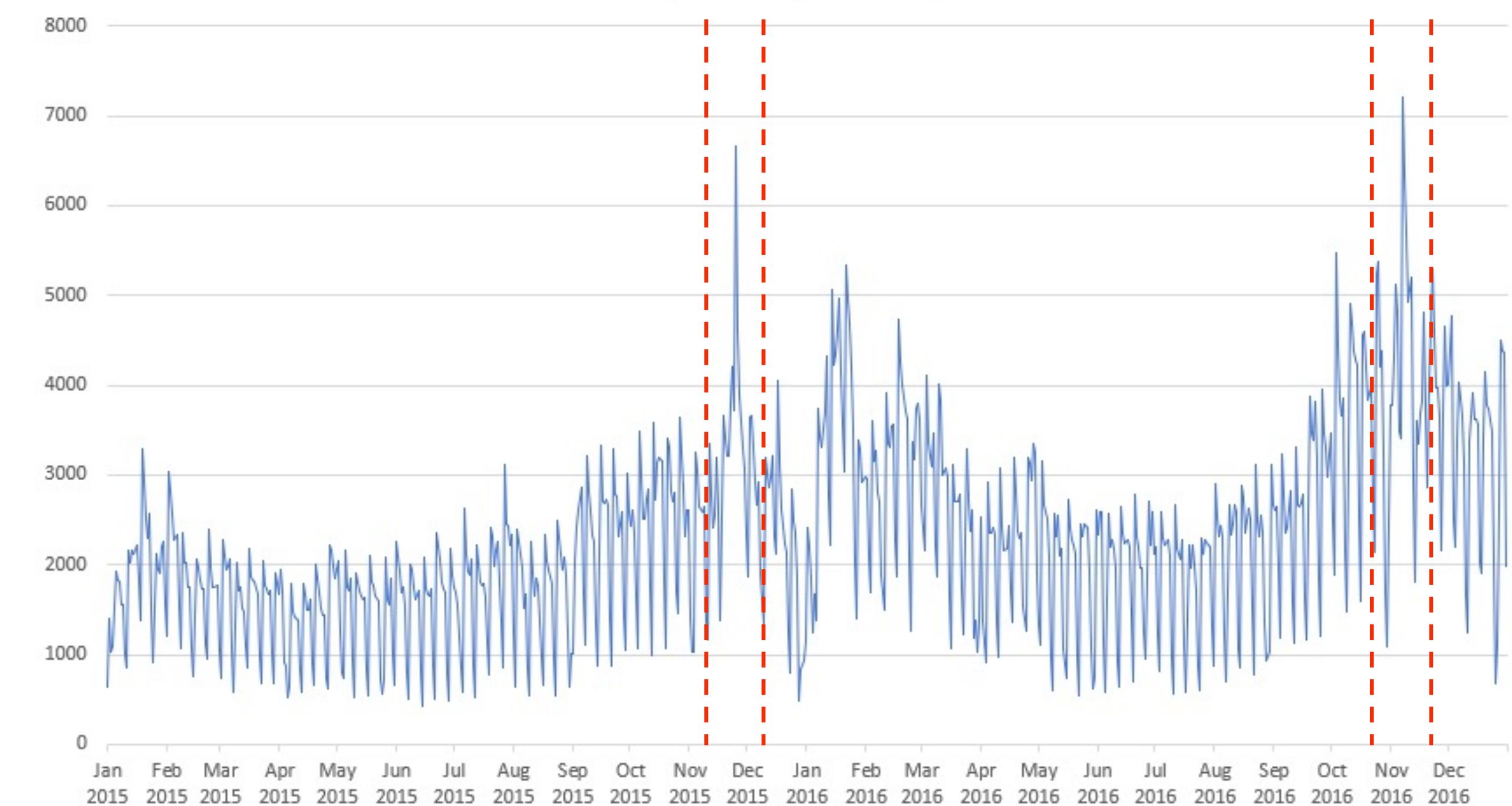


Мониторинг бизнес-процессов

Идея: хотим искать аномалии сами по себе — отражают сбои/изменения в бизнес-процессах.

ПРИМЕРЫ

1. Резко выросло количество обращений в кол-центр (срочно ищем проблему).
2. У метрики лояльности поменялся тренд (почему?).



Мониторинг бизнес-процессов



От данных к инсайту:

стратегия эффективного
оффлайн мониторинга



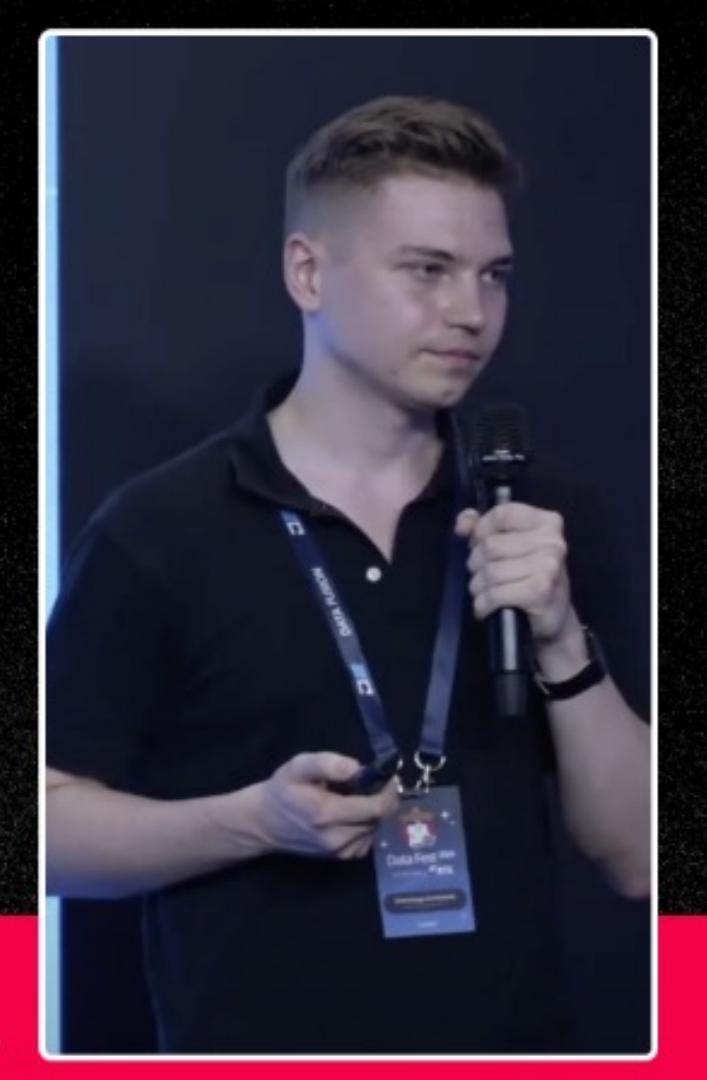
Максим Жерело | От данных к инсайту:
стратегия эффективного
оффлайн-мониторинга

Скоринг аномалий для
превентивного выявления
сбоев информационных
систем

Алгазинов Александр



Data Fest²⁰²⁴ • RANDOM DS/ML



Александр Алгазинов | Скоринг аномалий
для превентивного выявления сбоев
информационных систем

Мониторинг промышленных процессов



Анализ поведения показателей с различных датчиков в рамках промышленного процесса

ПРИМЕРЫ

1. Резкое повышение температуры
2. Скачки напряжения

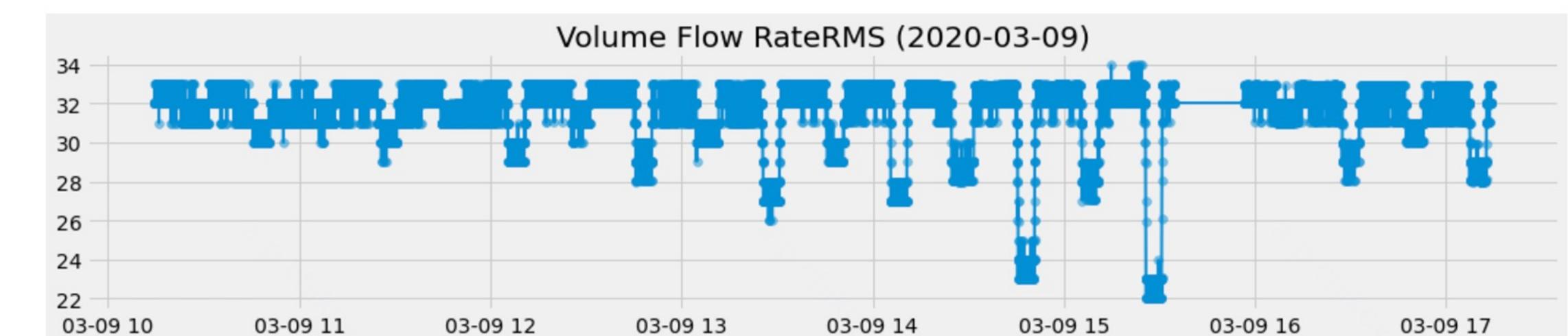
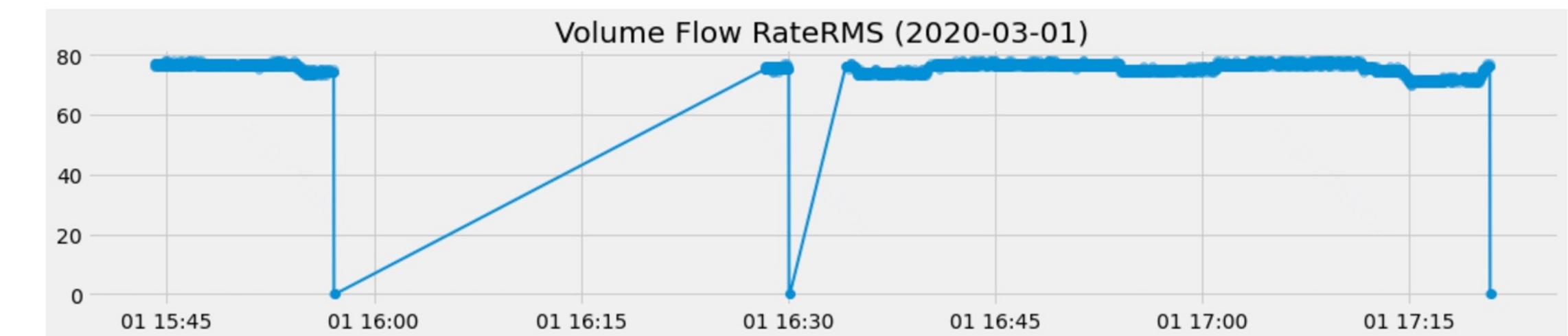
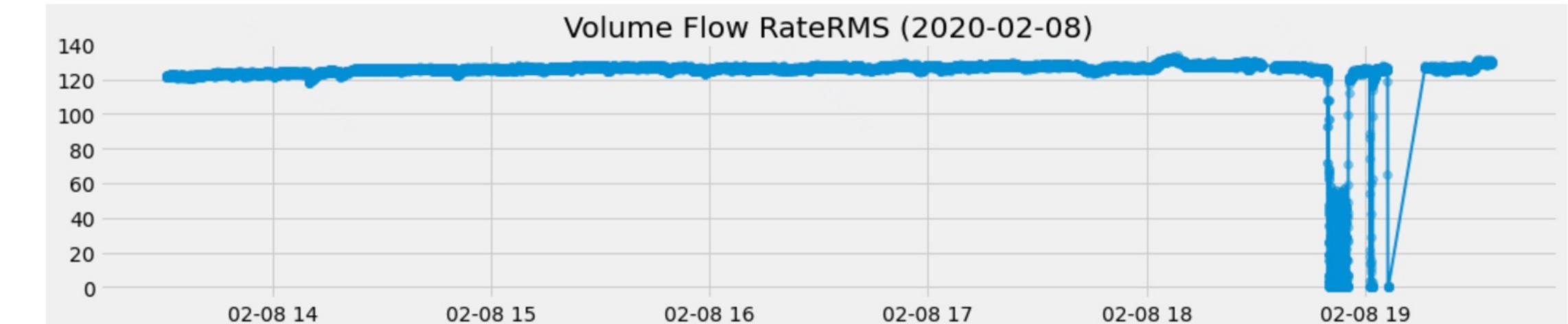


Юрий Кацер | SKAB и другие open-source-бенчмарки для задач обнаружения аномалий в промышленности

Мониторинг промышленных процессов

SKAB – Skoltech Anomaly Benchmark

- 34 штуки
- Аномалии в метриках промышленной установки
- Температура, вибрация, напряжение...



Испытательный стенд

Стенд состоит из следующий систем:

- 01 Система циркуляции воды
- 02 Система мониторинга состояния системы циркуляции воды
- 03 Система контроля и управления системой циркуляции воды
- 04 Система демонстрации технологии Time-Sensitive Networking (TSN)
- 05 Система сбора, обработки и визуализации данных

Фокус



Передняя панель и состав систем циркуляции воды, мониторинга, контроля и управления:
1 - водяной бак; 2 - водяной насос; 3 - электродвигатель; 4 - клапаны; 5 - механический рычаг для обеспечения нессоюности валов; 6 - кнопка аварийной остановки.

Датчики:
7 - датчик расхода (NI 9401 8-channel); 8 - датчик давления (NI 9203 8-channel); 9, 10 - вибродатчики (NI 9232 3-channel); 11, 12 - термопары (NI 9213 Spring Terminal 16-channel thermocouple).

Мониторинг ИТ-систем

Детектируем аномалии
в показателях технических метрик
(часто real-time).

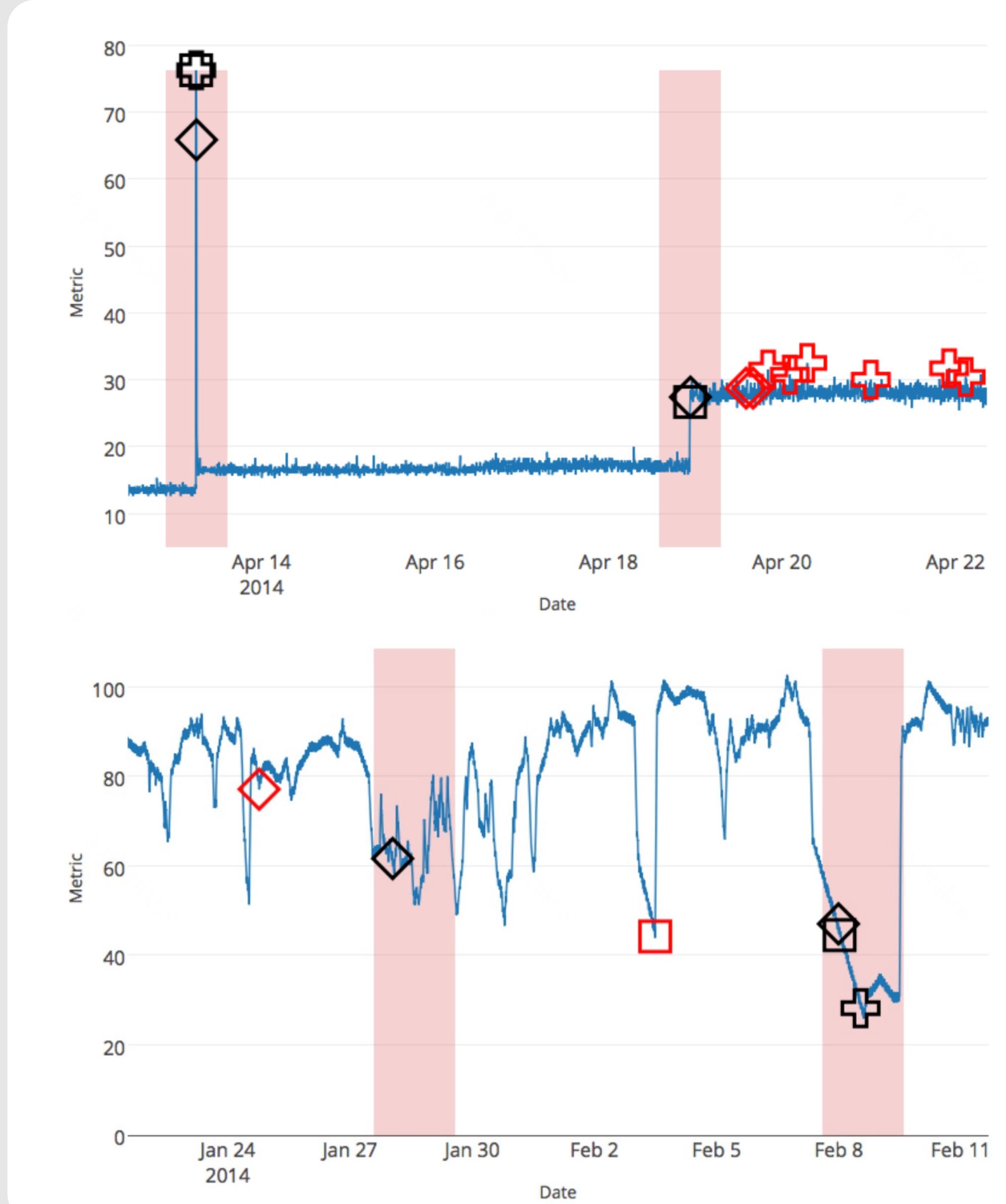
- Потребление CPU/RAM
- Количество запросов к системе (RPS)
- Фон ошибок/логов



Мониторинг IT-систем

NAB — Numenta Anomaly Benchmark

- 58 штук — синтетика + реальные
- AWS server metrics, Twitter volume, advertisement clicking metrics, traffic data
- Streaming anomaly detection



И другие...

Детекция фрода

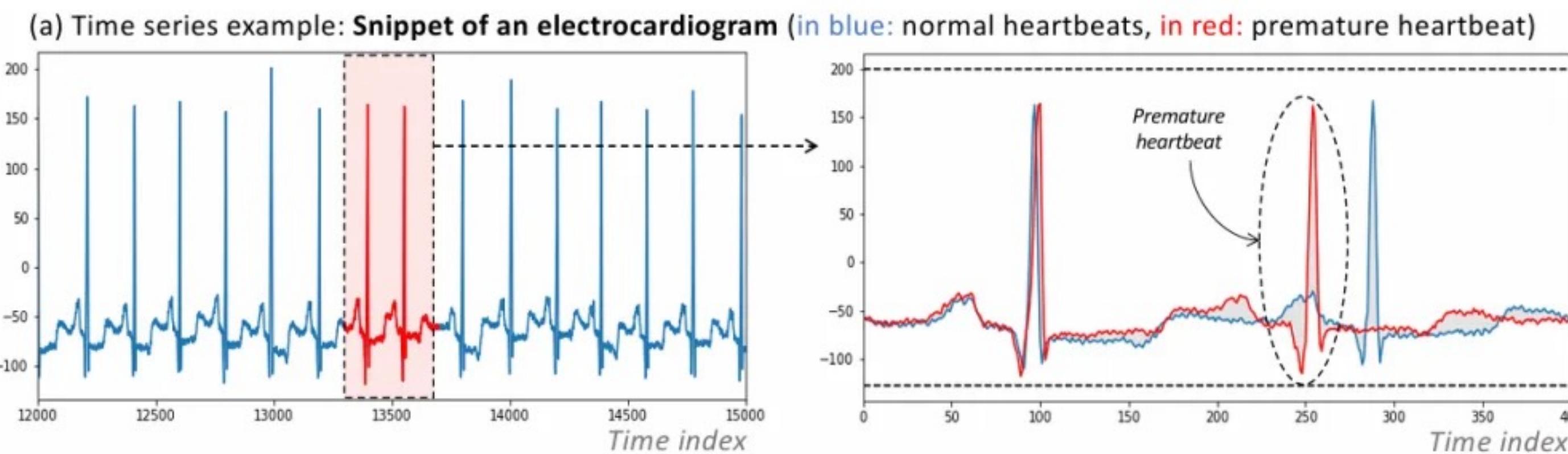
ПРИМЕРЫ

1. Большая сумма транзакции за границей
2. Необычные паттерны транзакций — частые мелкие транзакции

Мониторинг в здравоохранении

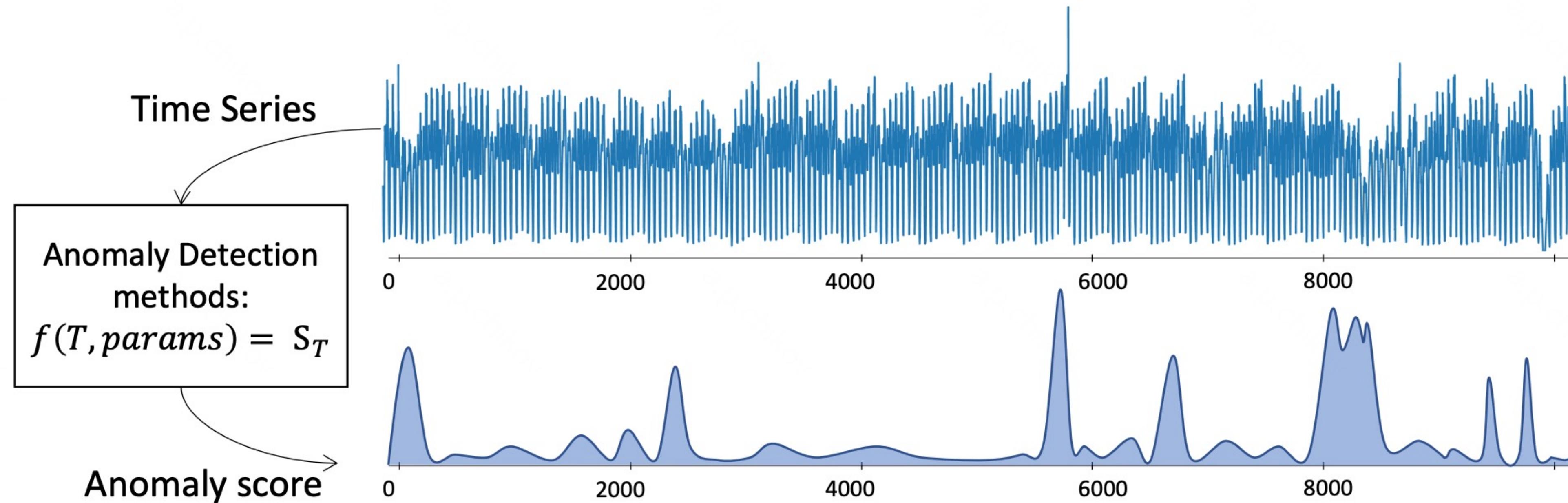
ПРИМЕРЫ

1. Нехарактерный ритм биения сердца на ЭКГ
2. Смена уровня показателей пациента (пульс/дыхание/давление...)



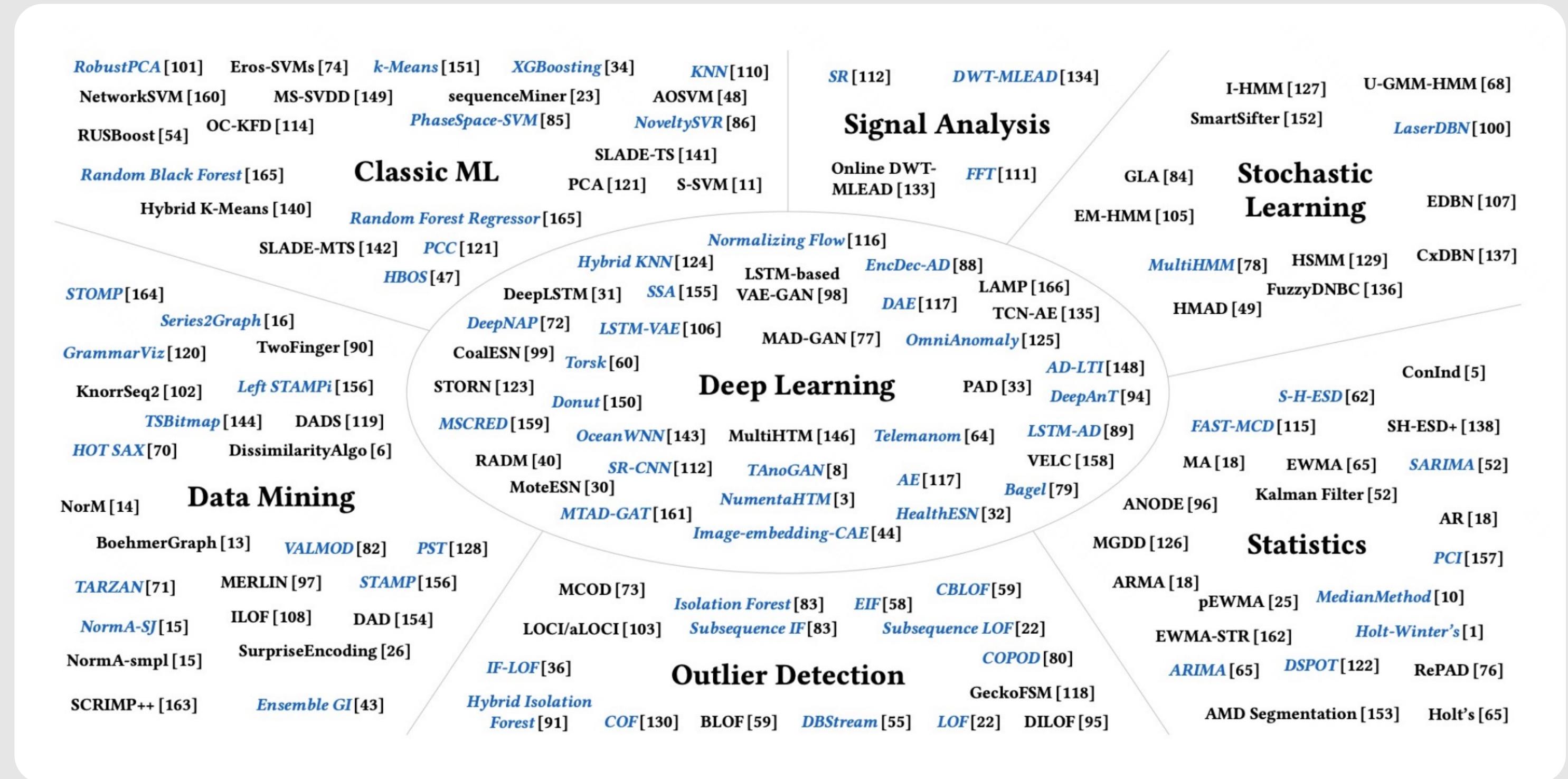
Методы поиска точечных аномалий

Формализация



Хотим создать некоторый оракул (f), который выставляет высокий скор для аномалий и низкий для нормальных наблюдений.

Таксономия (по дому)



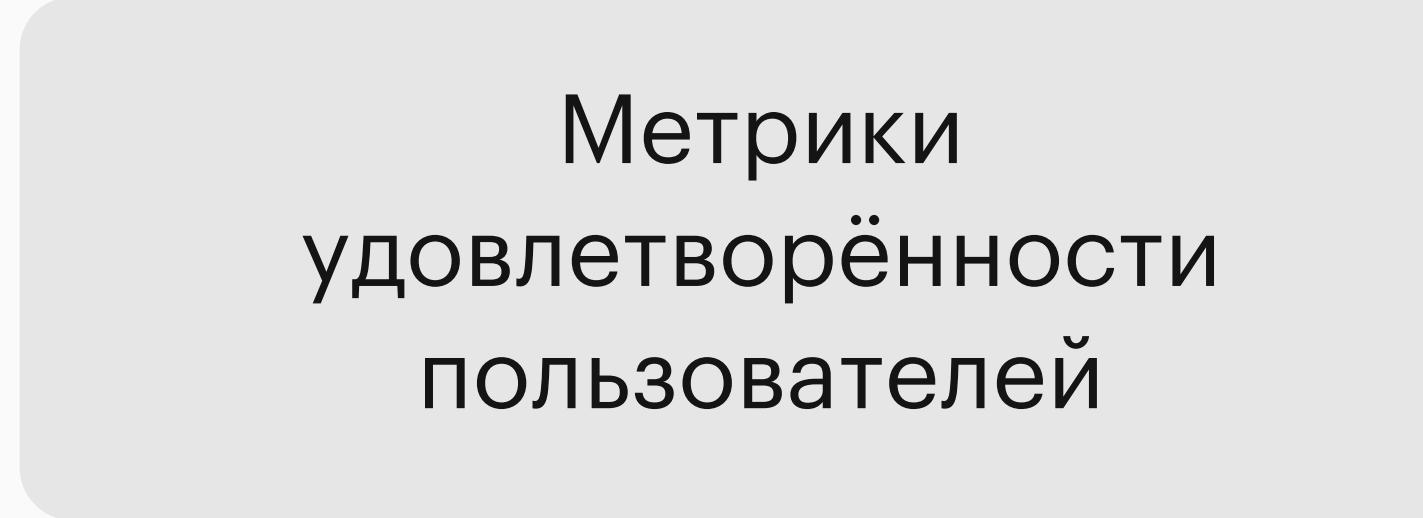
Методов поиска аномалий
бесконечное множество:

- часть специфична для Time Series;
- часть строится на сведениях к табличным данным.

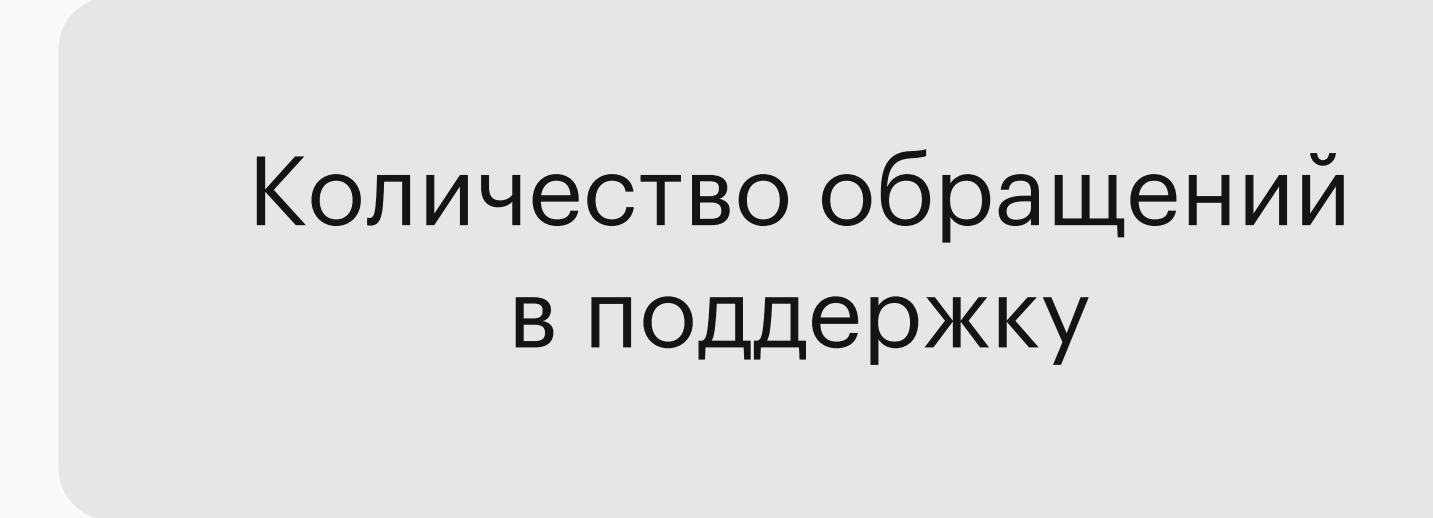
Таксономия (по стратегии обнаружения)

offline

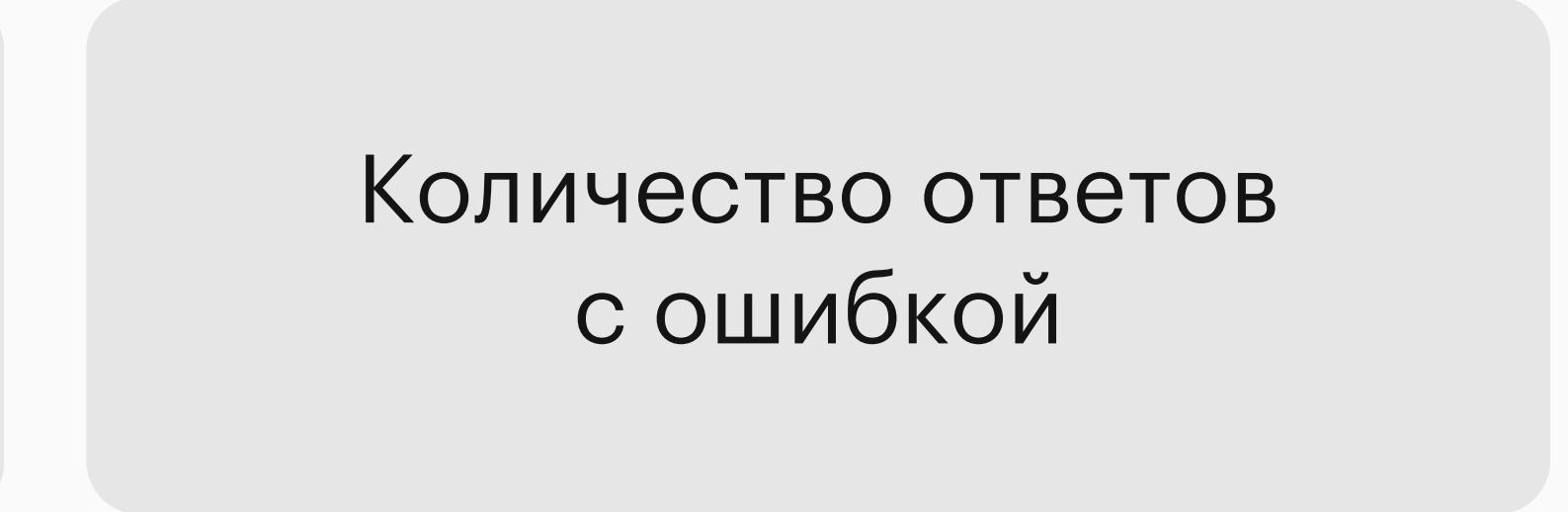
online



Хотим понимать, всё ли в порядке с продуктом.



Хотим своевременно распределять нагрузку.



Хотим моментально реагировать на сбои.

Таксономия (по стратегии обнаружения)

Offline

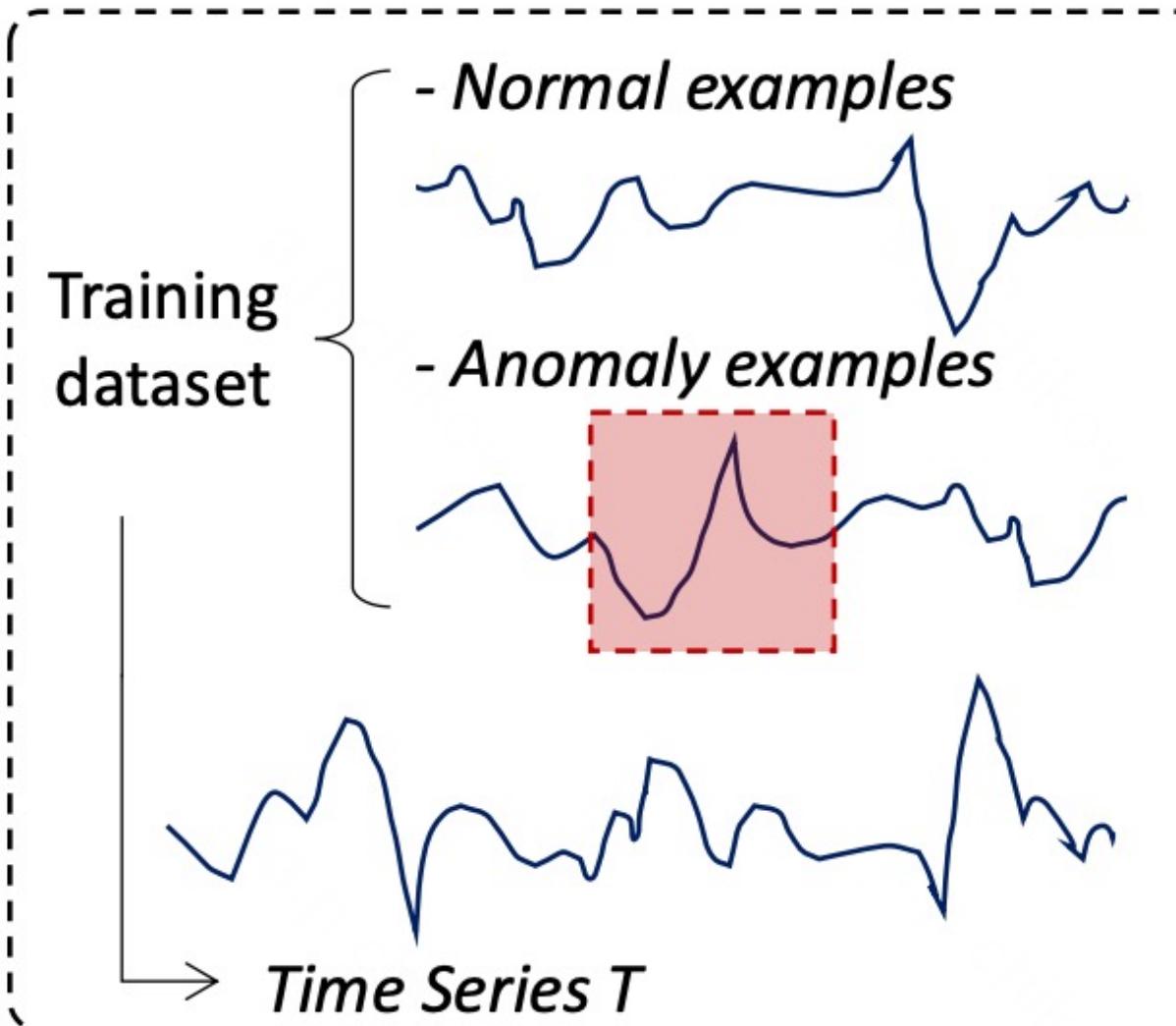
- Меньше требований к скорости работы.
- Доступны данные за весь анализируемый промежуток.
- Частотность:
 - дневная,
 - недельная,
 - месячная.

Online

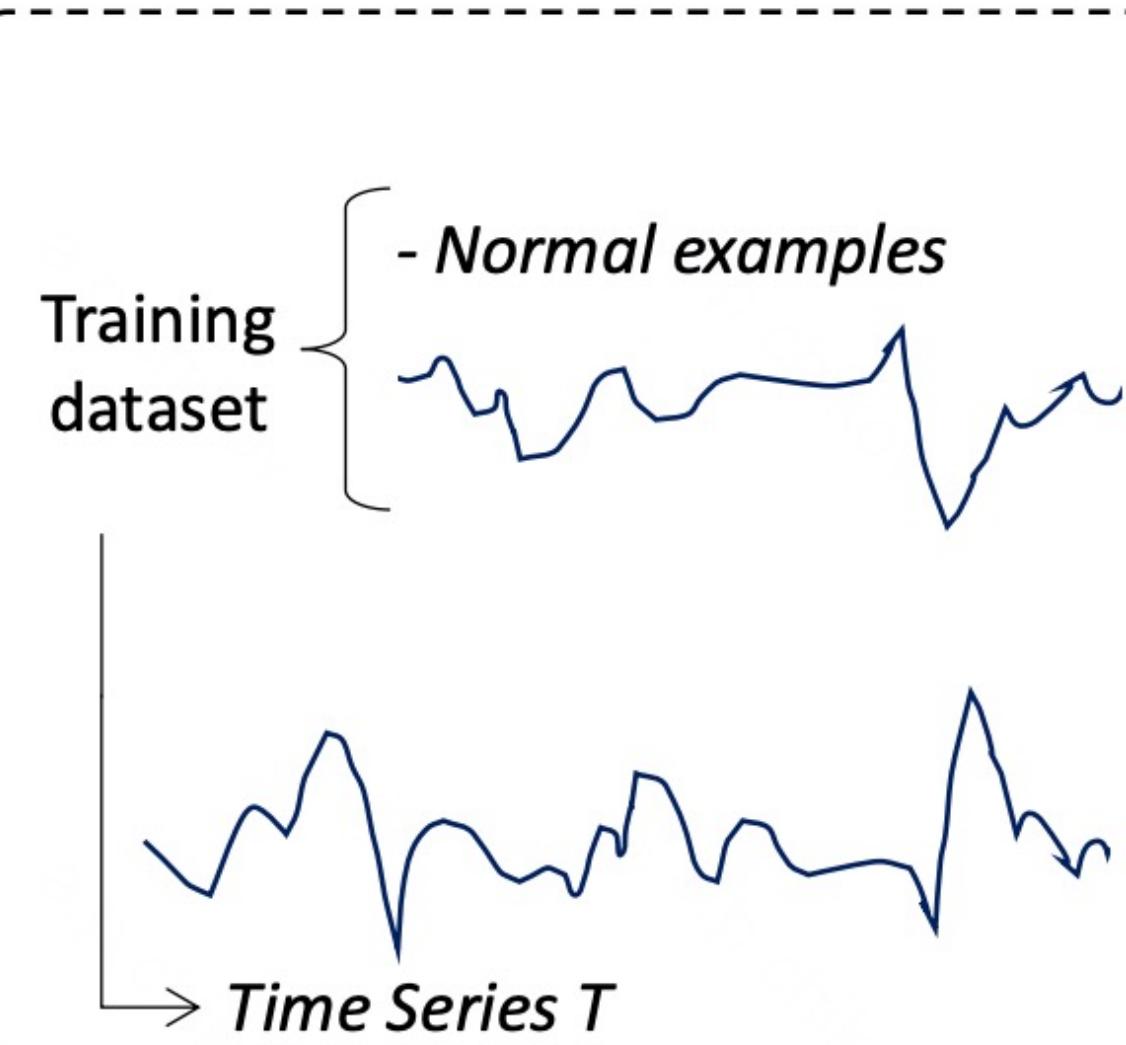
- Жёсткие ограничения на время работы: должны обнаружить аномалию как можно быстрее.
- Доступны только исторические данные.
- Частотность:
 - минутные,
 - часовые.

Таксономия (по наличию разметки)

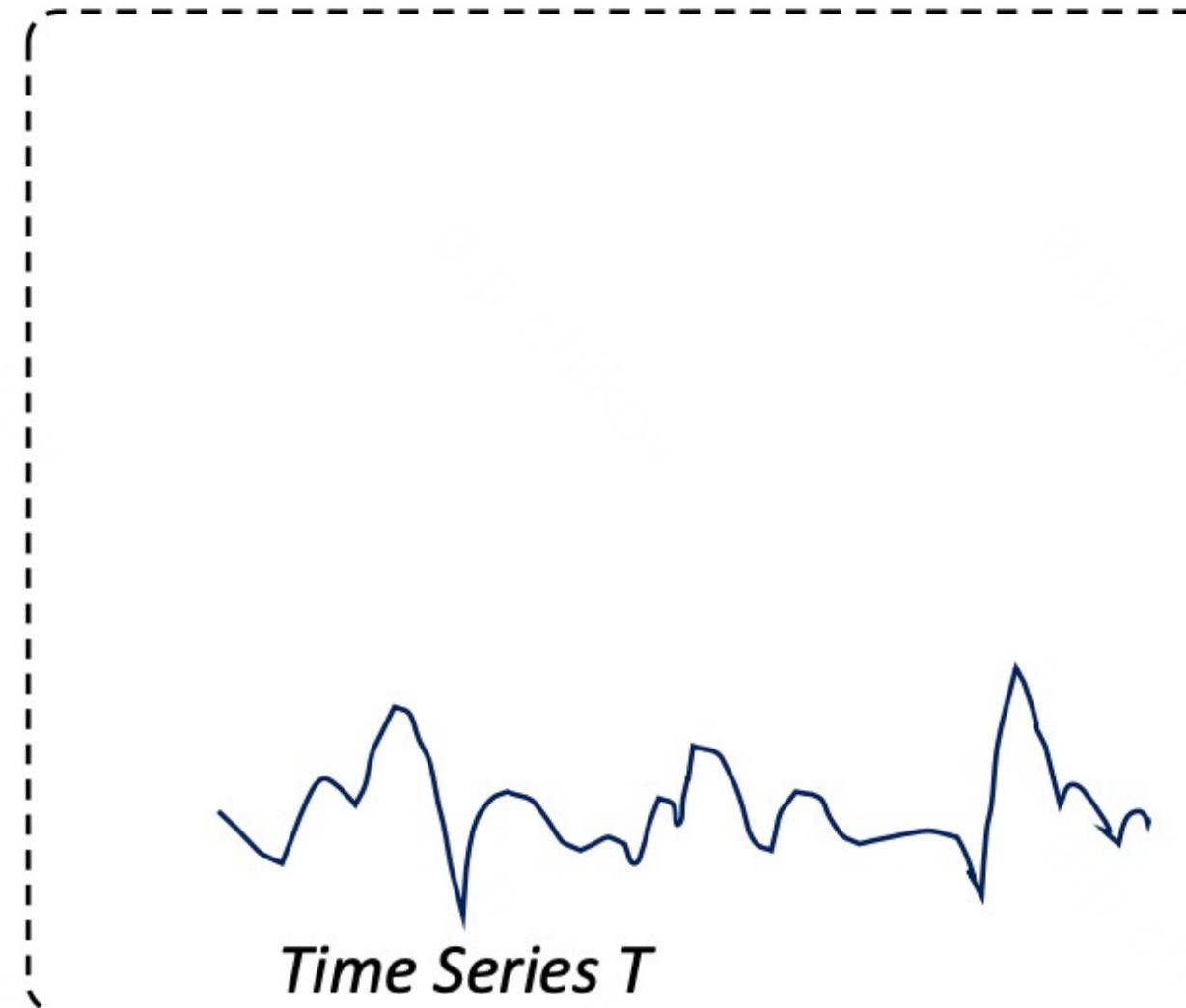
Supervised



Semi-supervised



Unsupervised



Чаще используются эти подходы из-за проблем с разметкой.

Оценка качества

1. Метрики классификации

→ Precision

$$Precision = \frac{TP}{TP + FP}$$

→ Recall

$$Recall = \frac{TP}{TP + FN}$$

→ F-beta

$$F_{\beta} = \frac{(1 + \beta^2) Precision \ Recall}{\beta^2 Precision + Recall}$$

2. Метрики downstream-задач

→ Прогнозирование → MAE

А разметку где брать?

Разметка не нужна.

Supervised-методы и разметка данных

Classification Model

timestamp	target	features	class
01-01-24	2	4	0
02-01-24	10	3	0
03-01-24	100	2	1
04-01-24	3	4	0

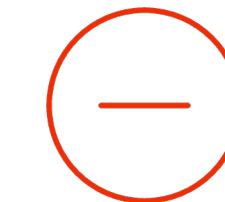


CatBoost



Плюсы:

подстраивается под конкретную задачу.

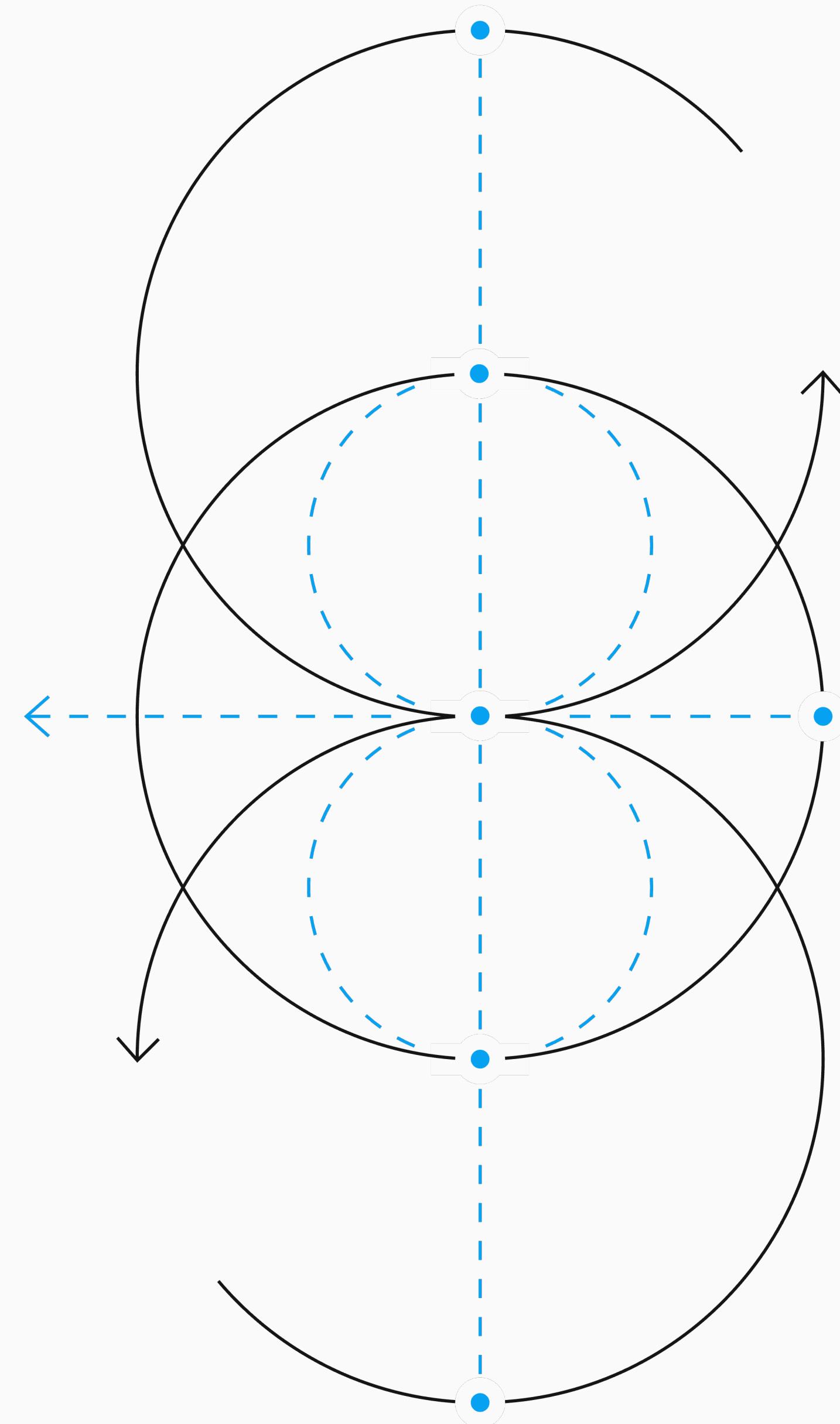


Минусы:

- нужна разметка;
- мало позитивов.

В чём сложность?

- Готовой разметки чаще всего нет.
 - Не можем учить supervised.
 - Не можем оценить качество.
- Нет чёткой формализации, что такое аномалия.
 - Ищем что-то неопределённое.
- Долго и дорого собирать разметку.



Search Method



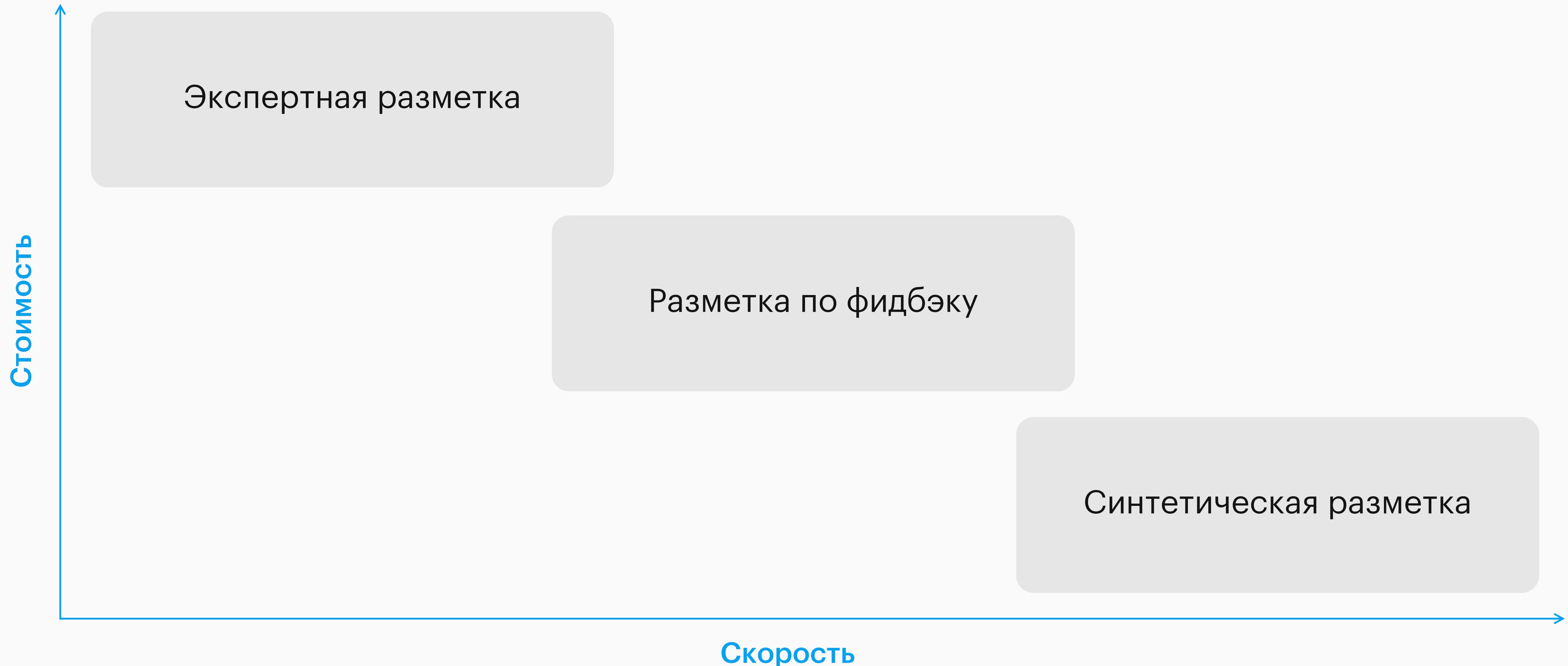
Реально собираем разметку

- Экспертная разметка:** просим эксперта разметить его данные.
- Разметка по фидбэку:** просим эксперта провалидировать данные с предразметкой.

Готовая разметка

- Открытые датасеты:** сложно найти данные, похожие на ваш домен.
- Синтетические данные:** нужно потратить время на настройку процесса генерации.

Основные подходы

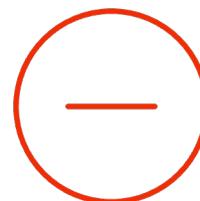


Синтетические данные

Идея: сгенерируем датасет на основе реальных данных, добавим аномалии, характерные для целевых данных.



Известна точная разметка.



Синтетика не всегда полностью описывает реальные данные.

RecSys & TS

Генерация синтетических временных рядов для решения задачи поиска аномалий

Владислав Власов
исследователь-разработчик,
Anomaly Analyzer, Т-Банк

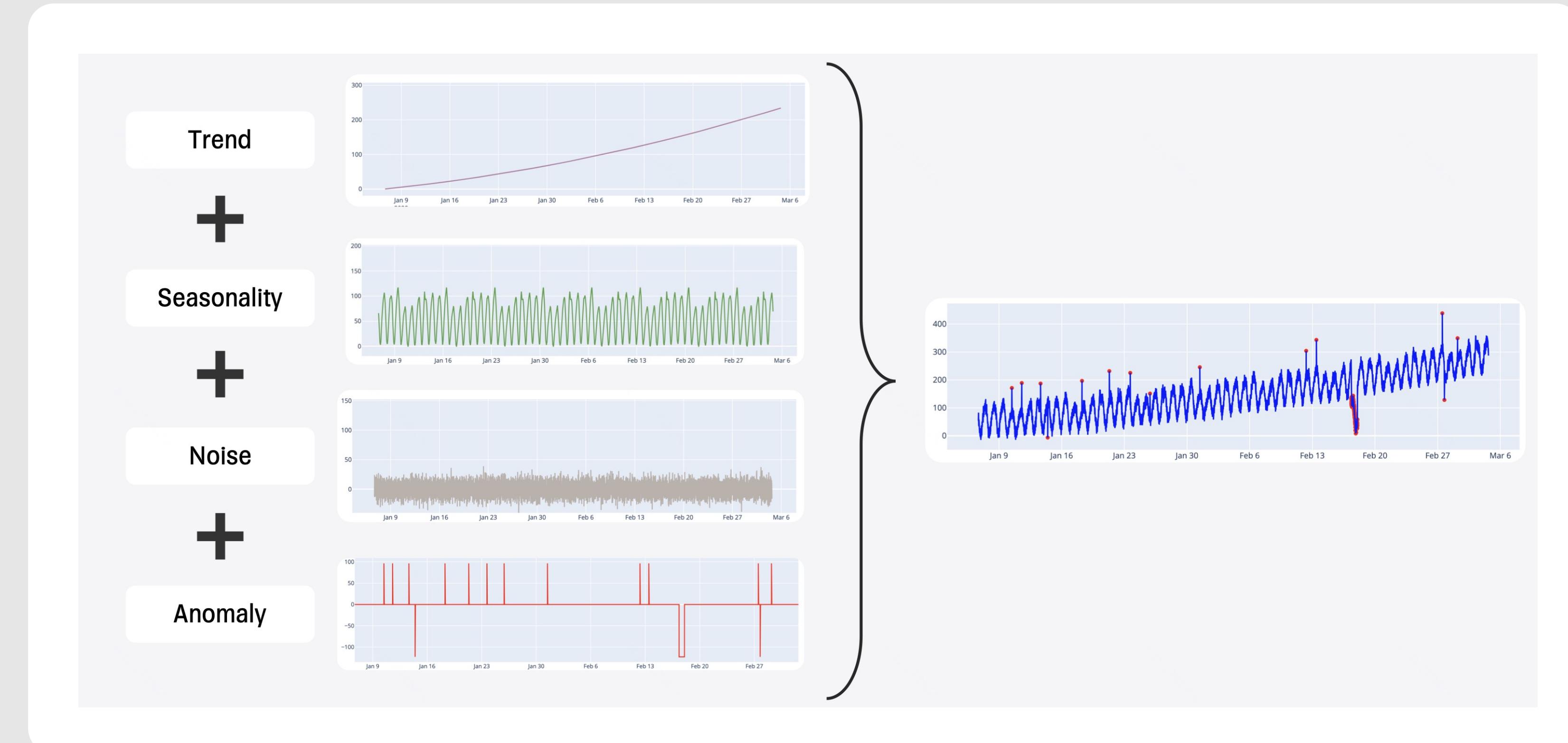


Генерация синтетических временных рядов для решения задачи поиска аномалий

Синтетические данные

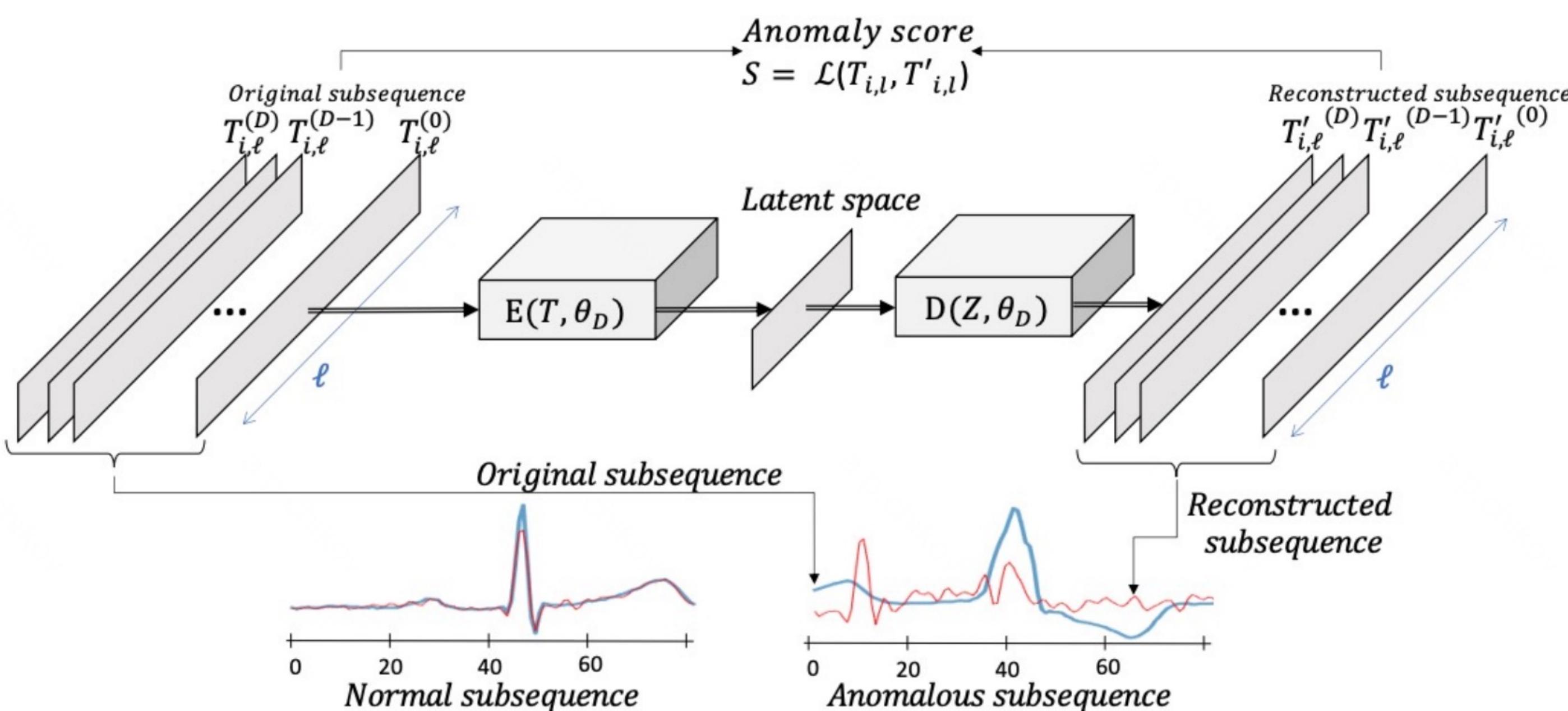
Можно генерировать синтетику по слоям, аномалии — это отдельный слой.

(Будем так делать на семинаре.)



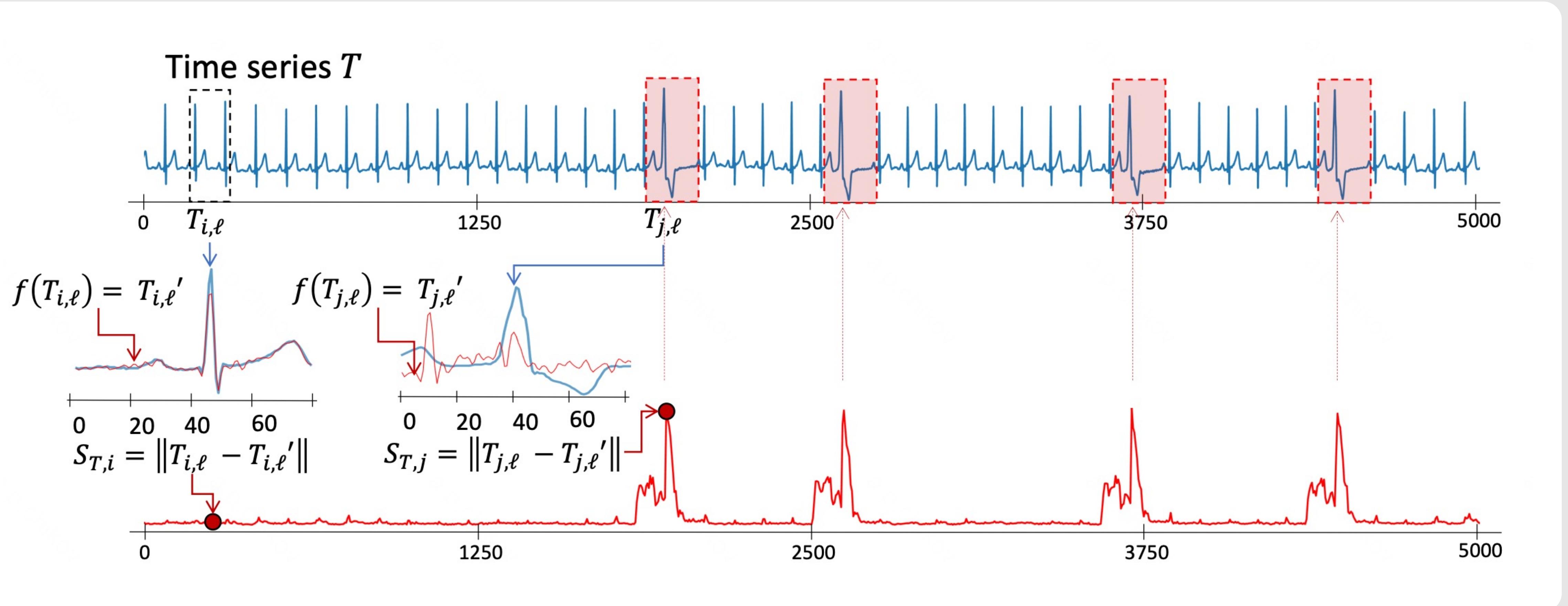
Semi-supervised и Unsupervised

AutoEncoders

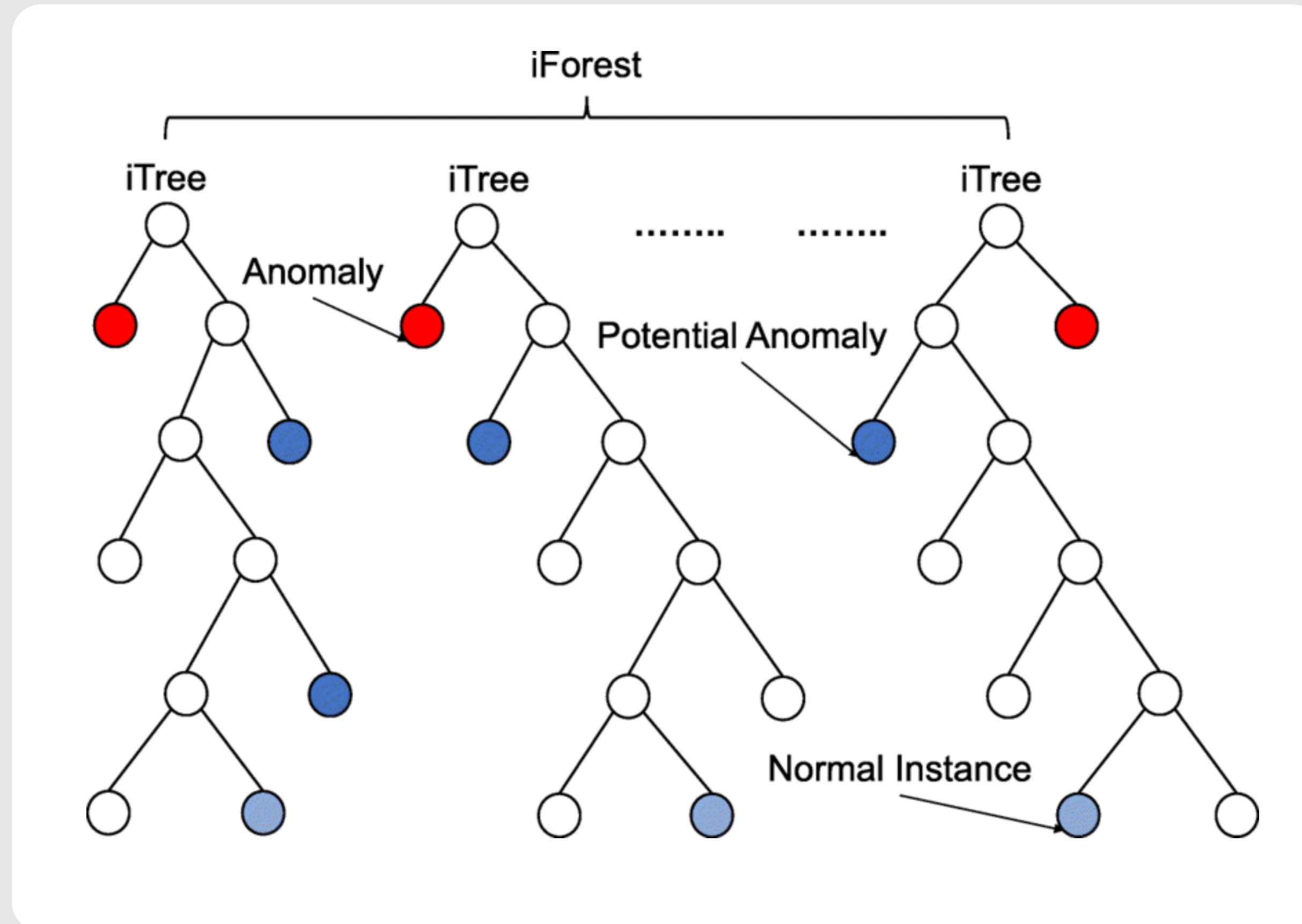


- Обучим автоэнкодер минимизировать reconstruction loss.
- Считаем аномалией всё выше порога реконструкции.

AutoEncoders



Isolation Forest

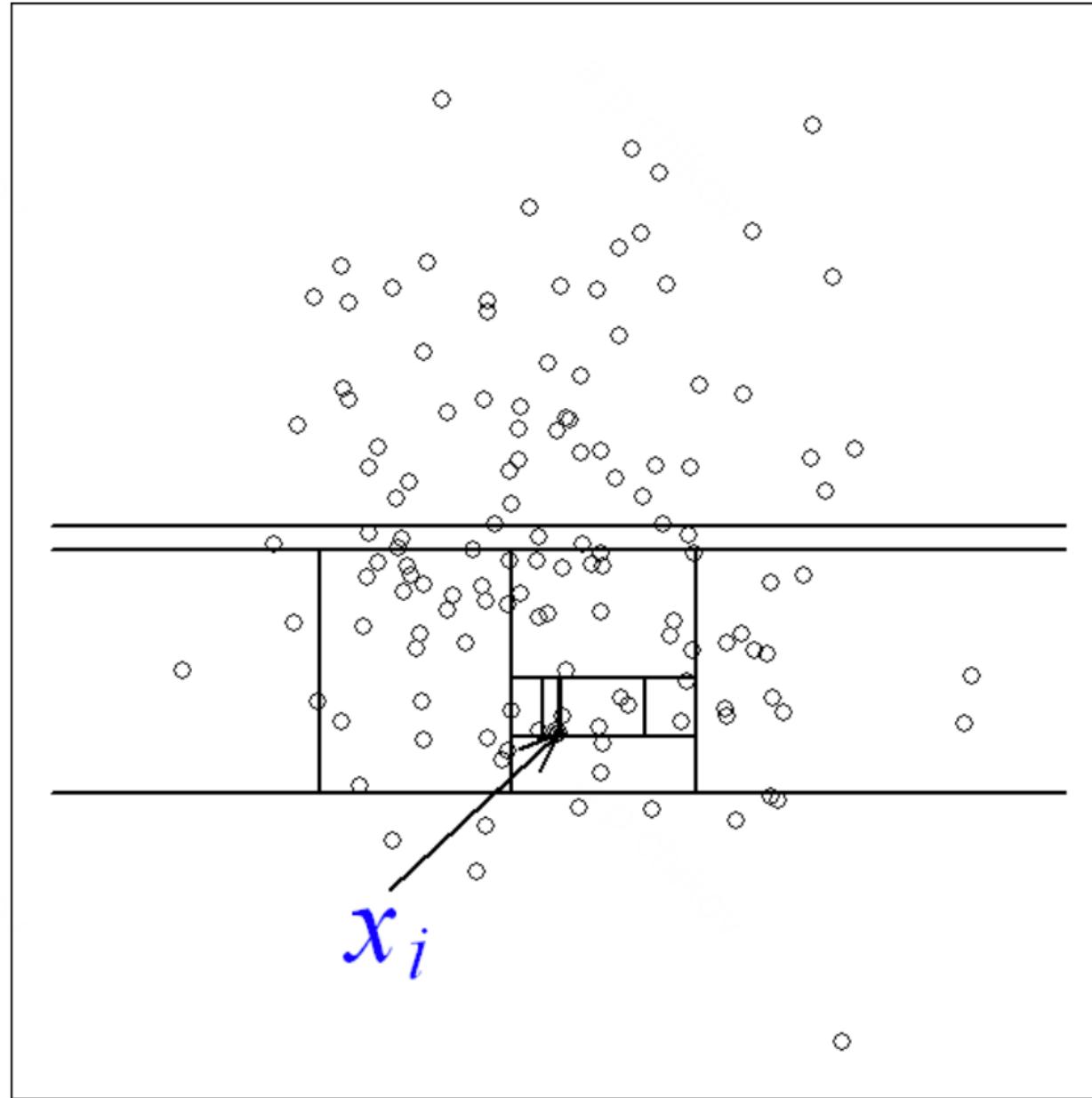


Повторяем рекурсивно +
много деревьев.

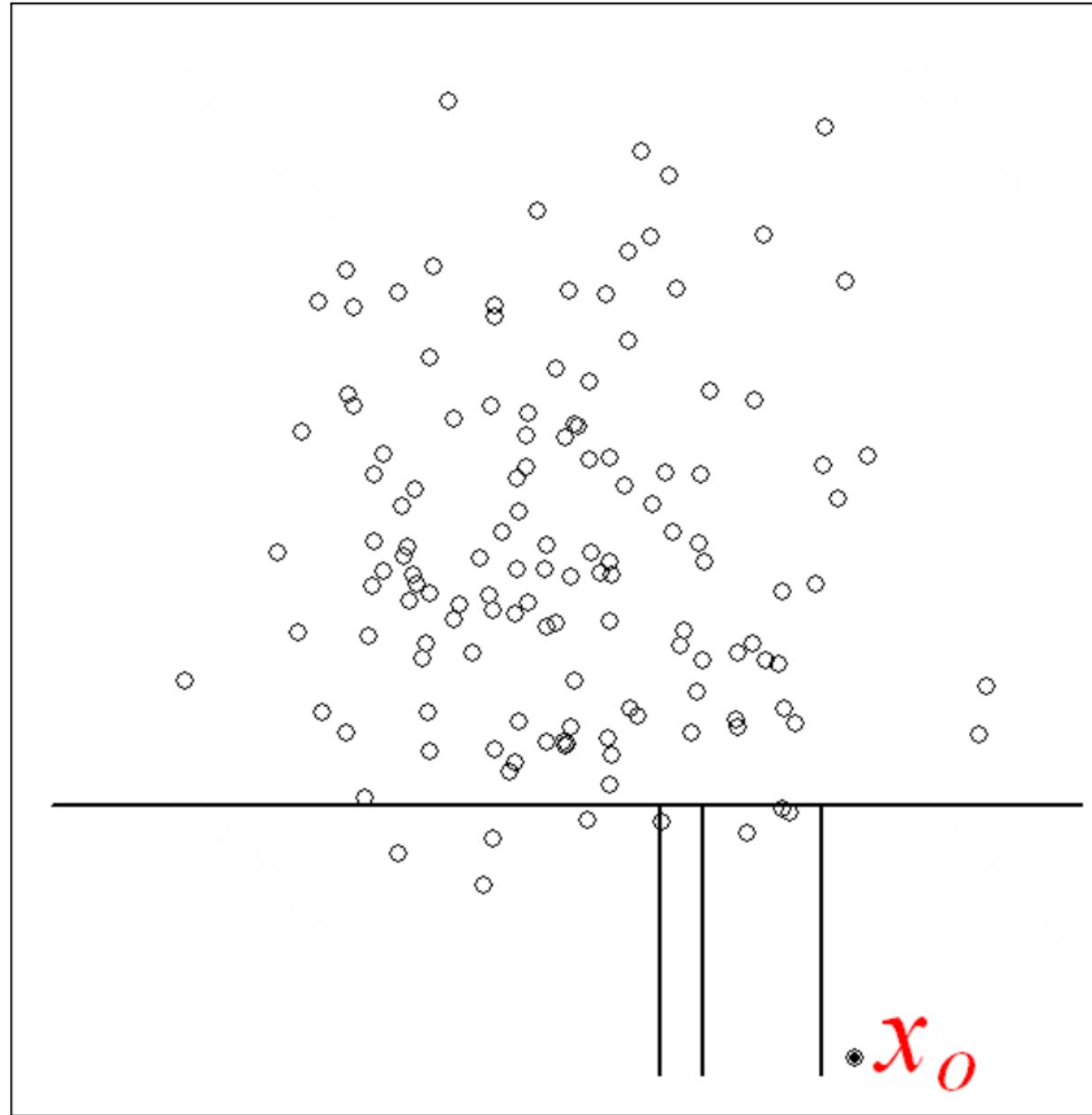
1. Выбираем случайный признак.
2. Выбираем случайный порог из значений признака.
3. Разделяем семплы по порогу.

Идея: аномальные точки будут
отделяться в отдельные вершины
быстрее нормальных.

Isolation Forest

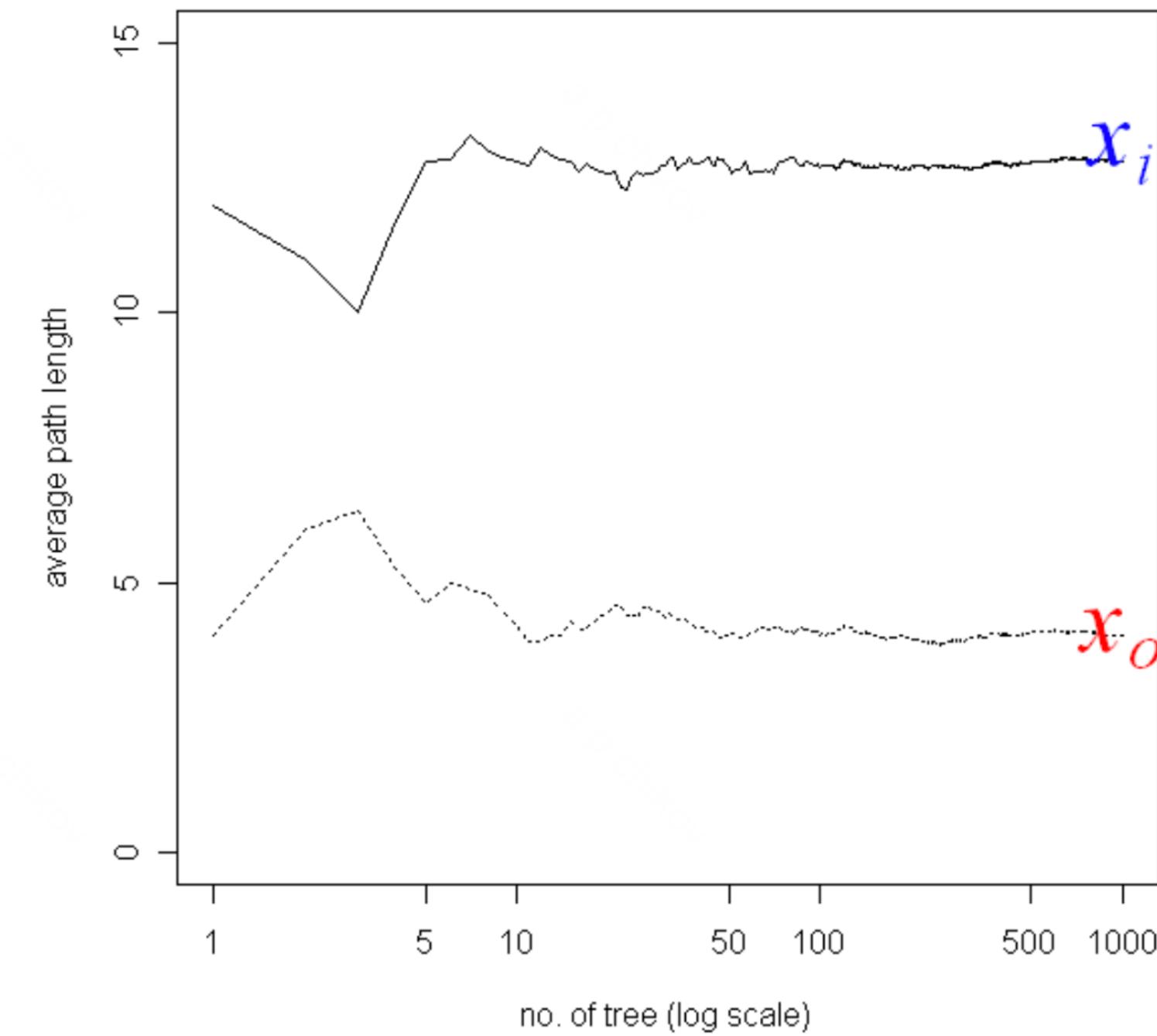


(a) Isolating x_i



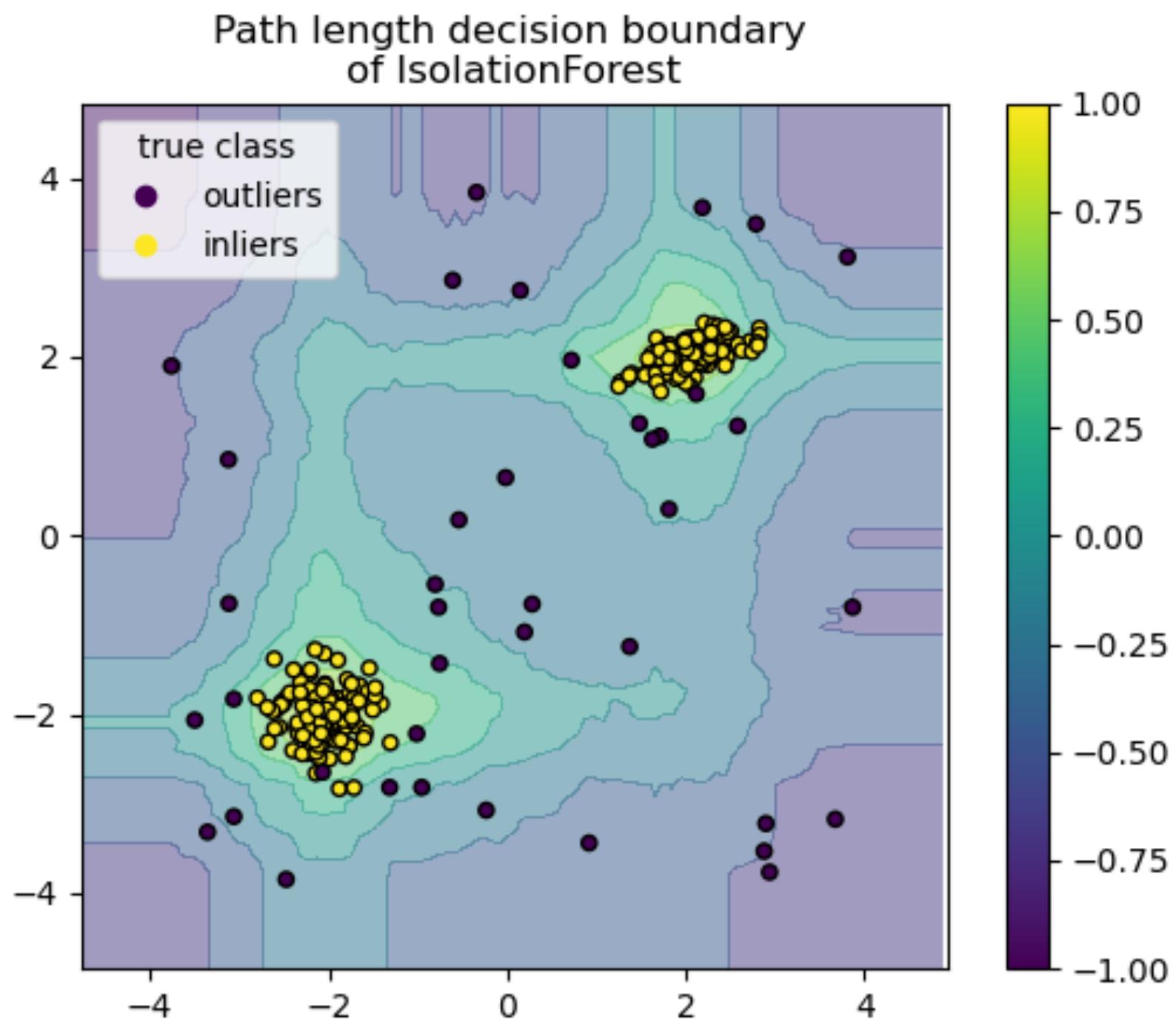
(b) Isolating x_o

Сплиты в дереве для нормальной
и аномальной точек

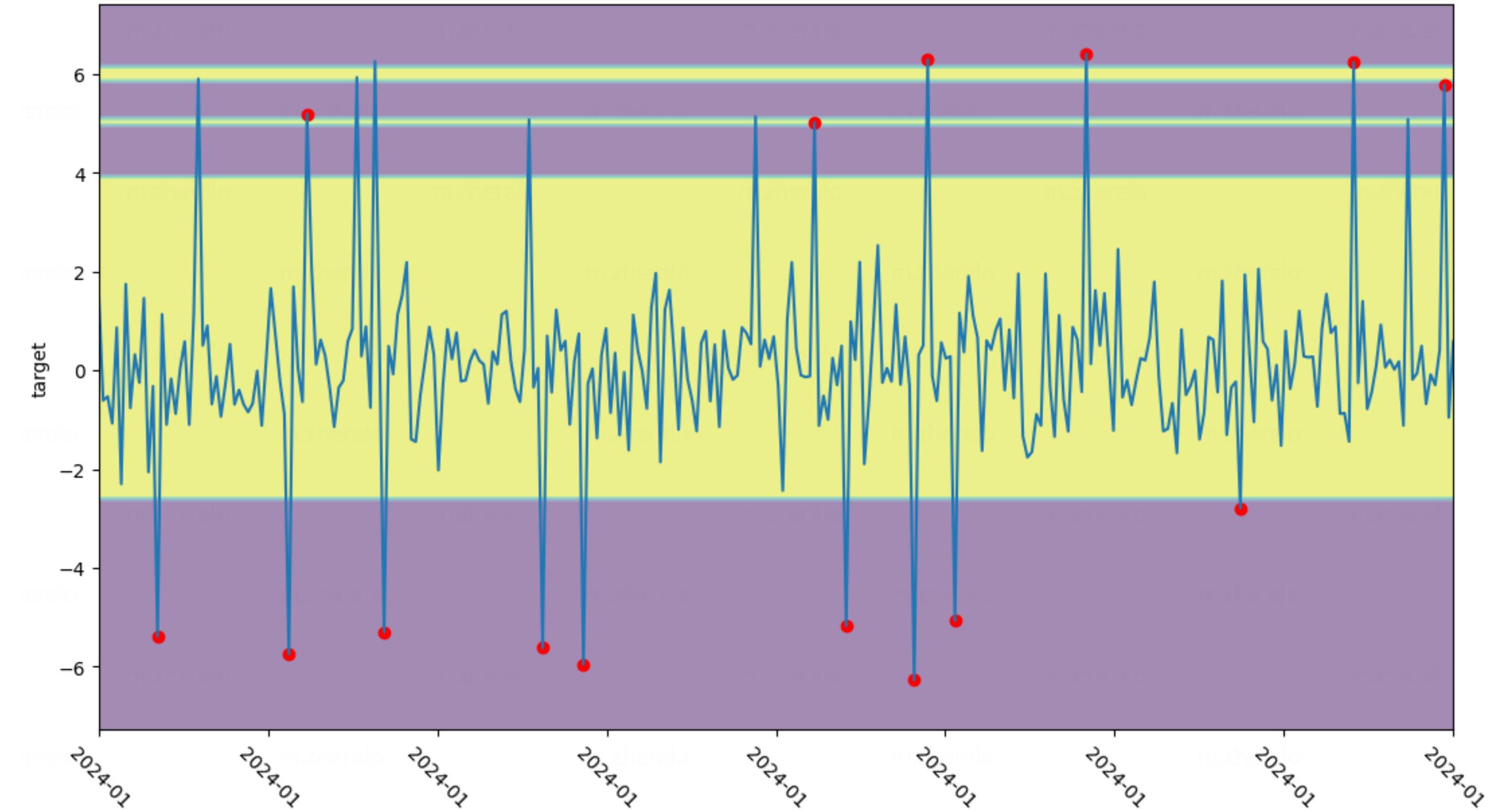


Средняя длина пути для нормальной
и аномальной точек с ростом
числа деревьев

Isolation Forest

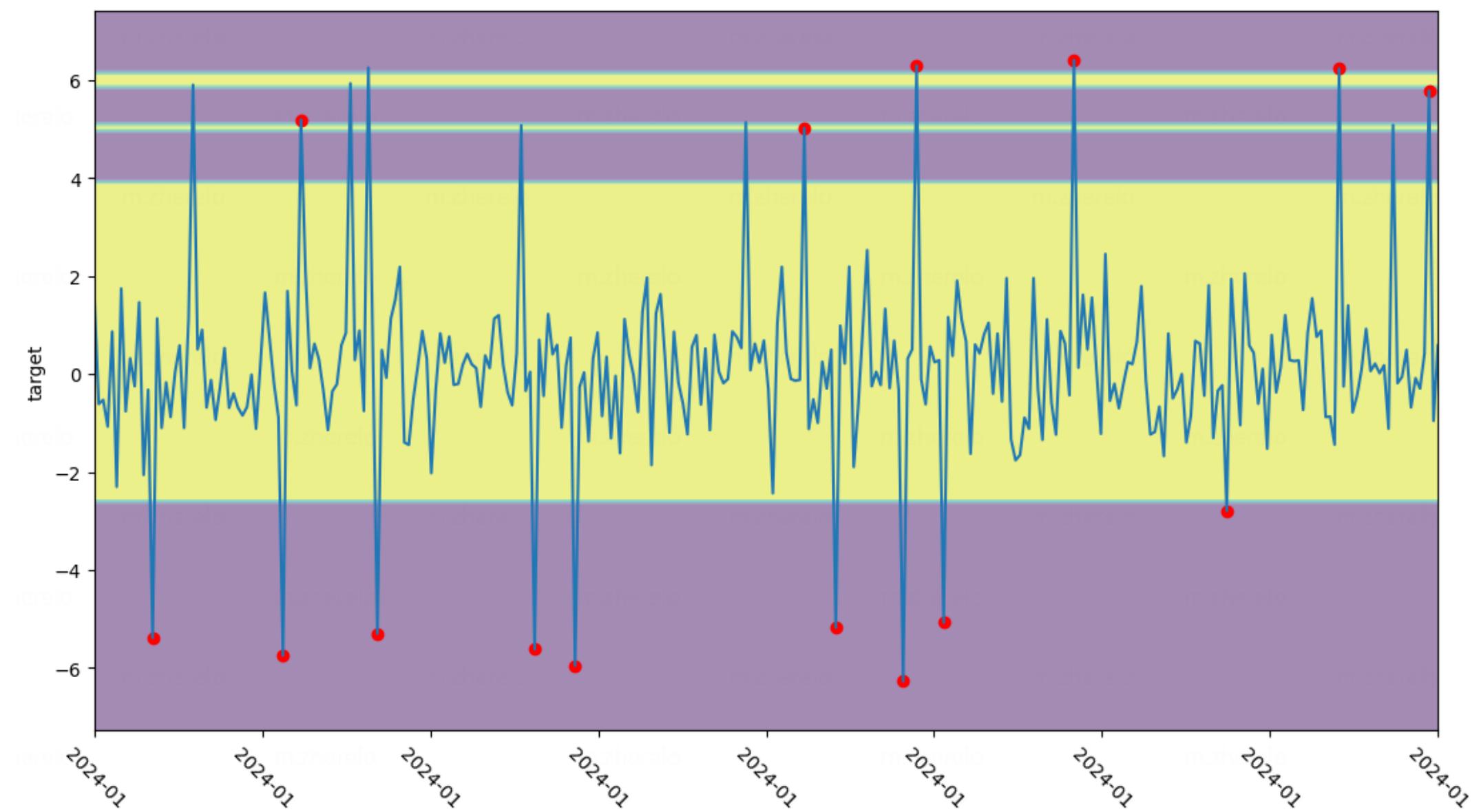


Средняя длина путей в 2d

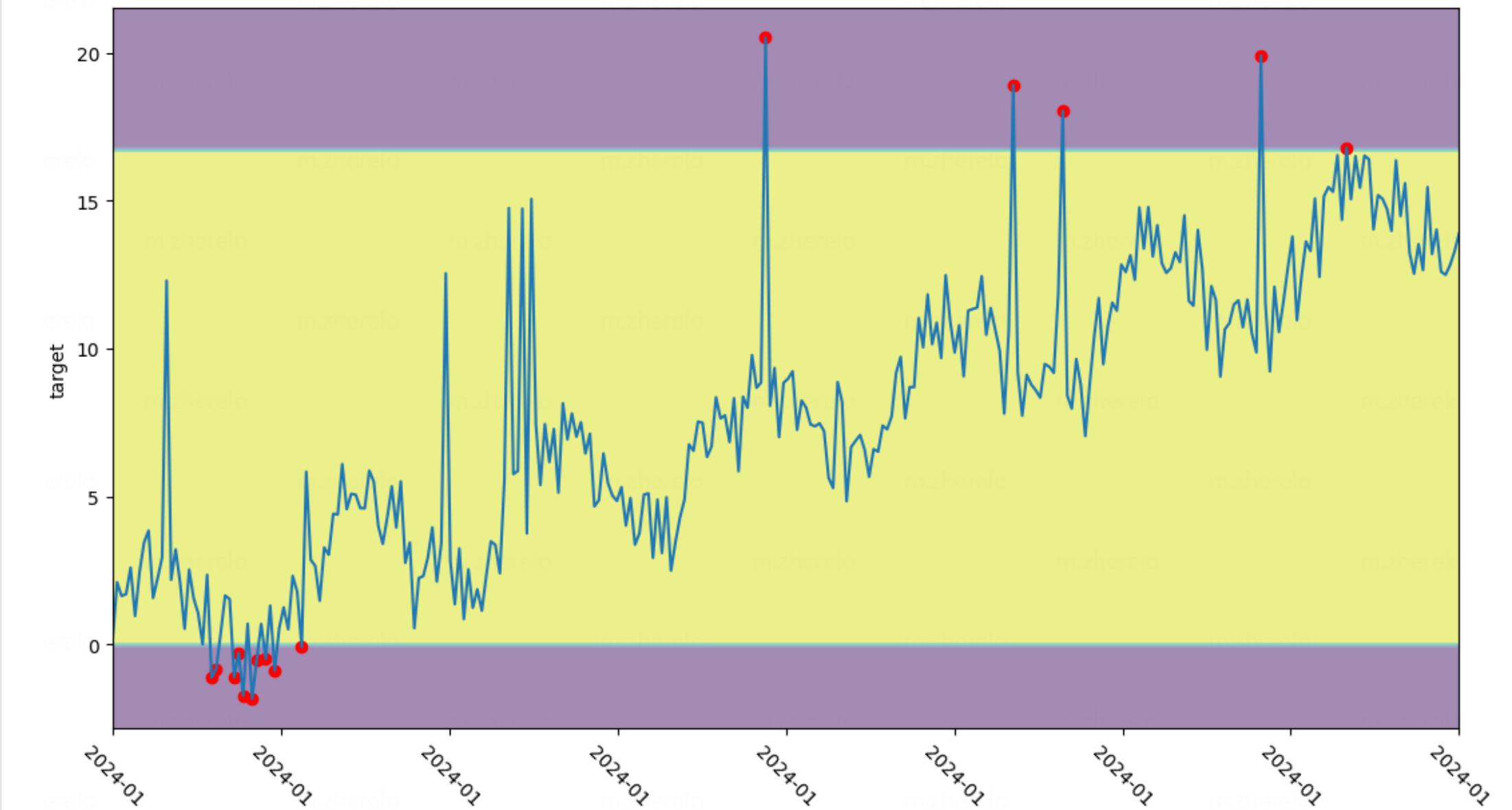


Решающее правило в 1d
(Сплитимся только по target.)

Isolation Forest

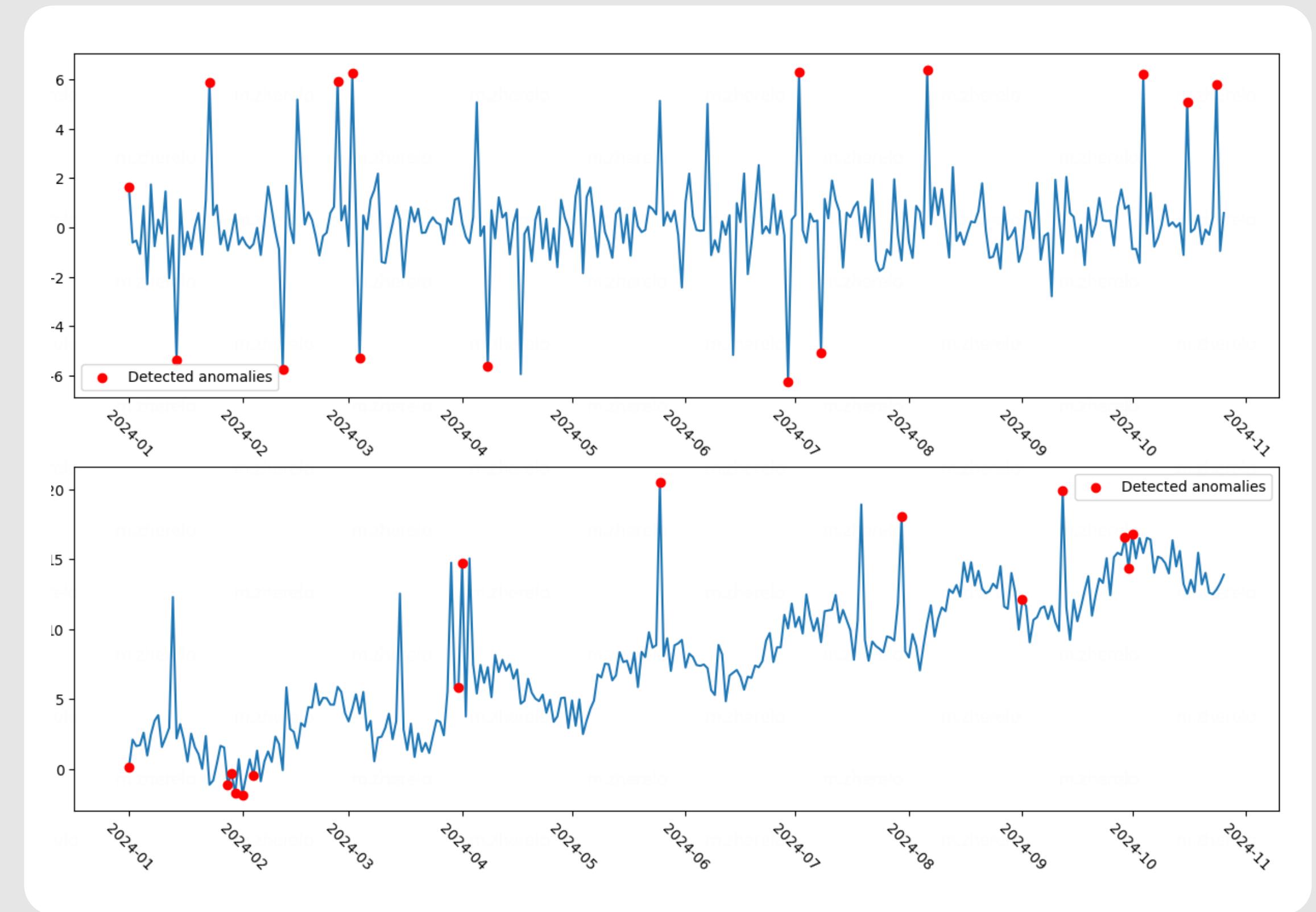
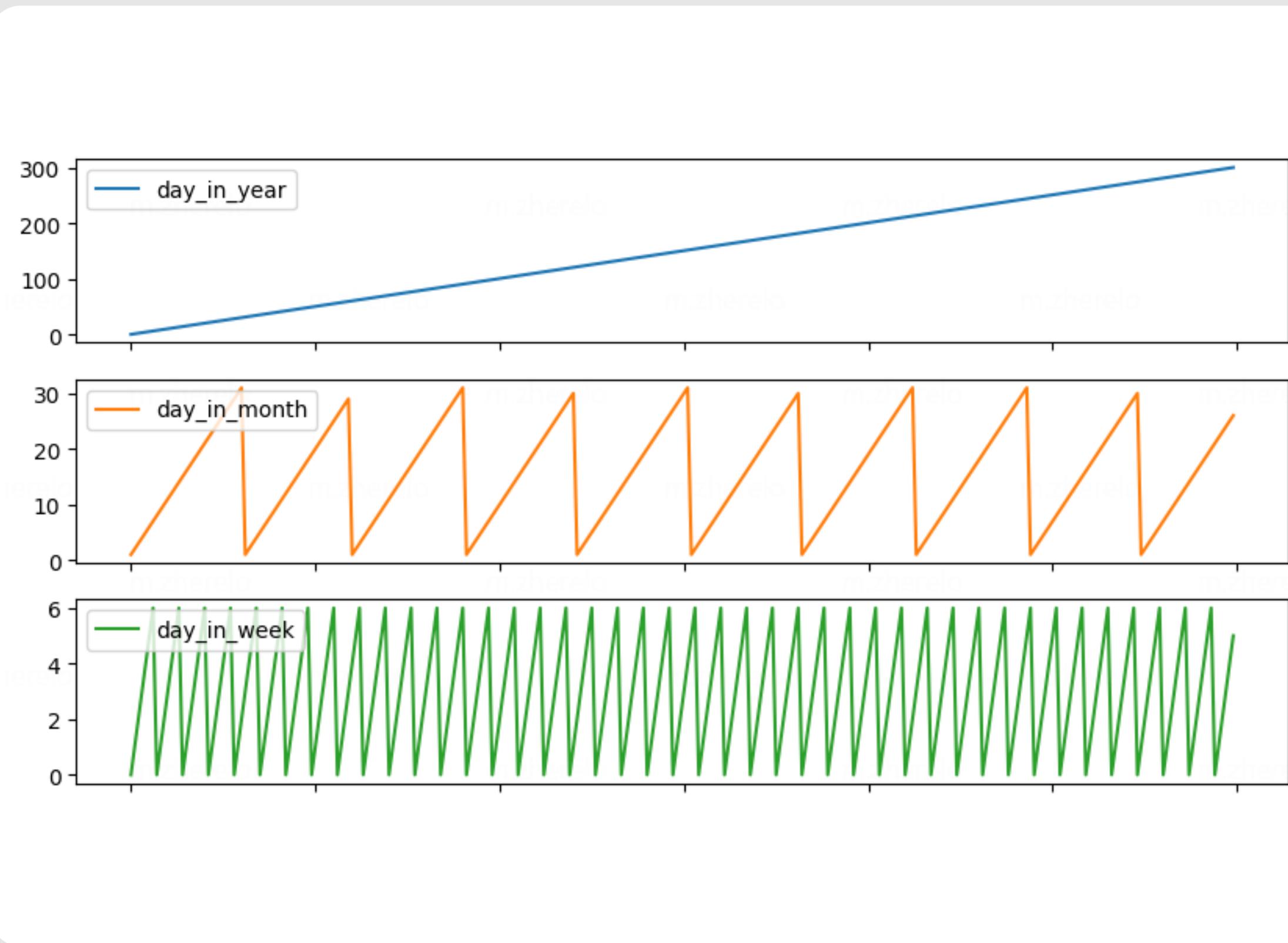


OK

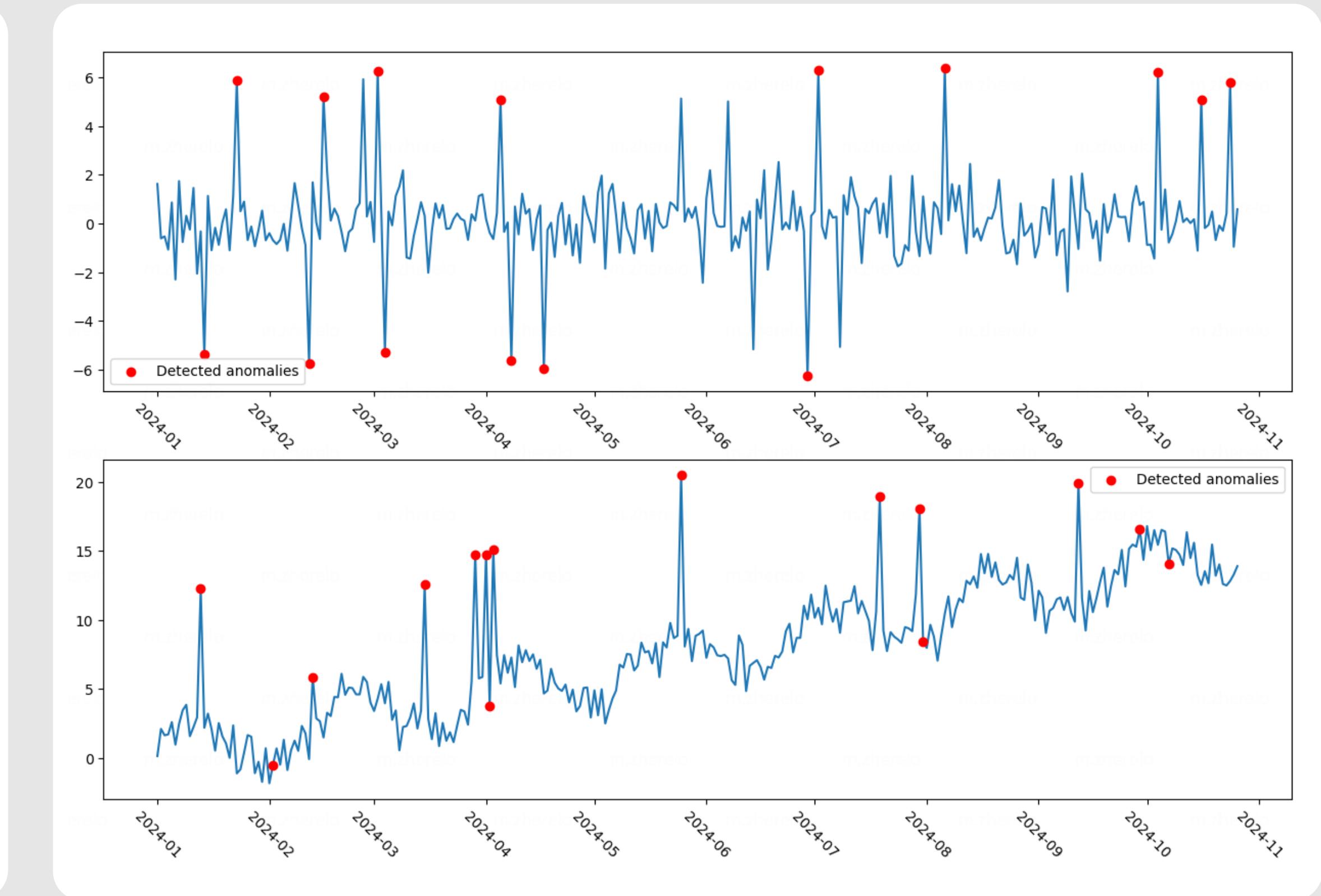
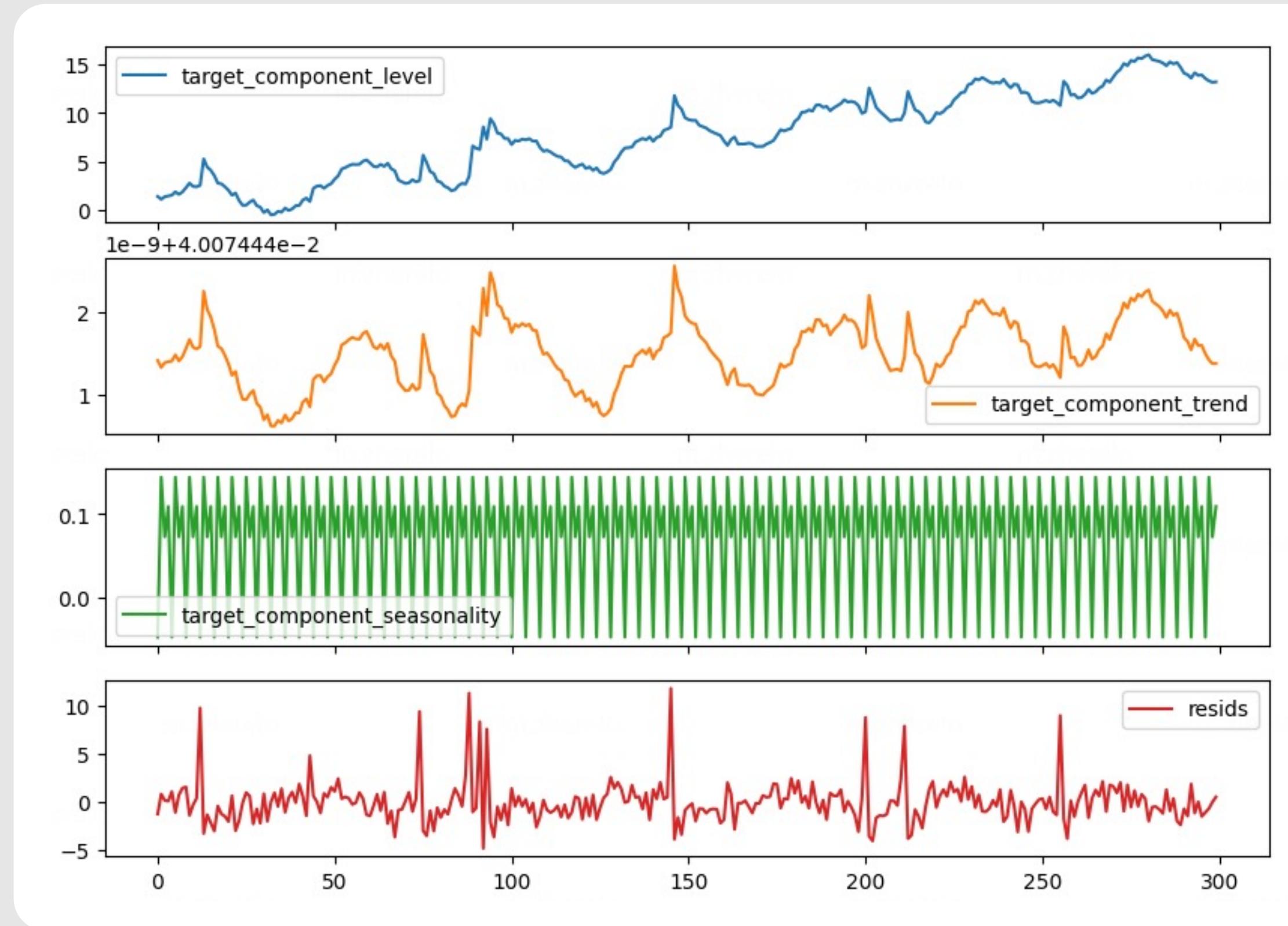


Нужна более сложная решающая поверхность.

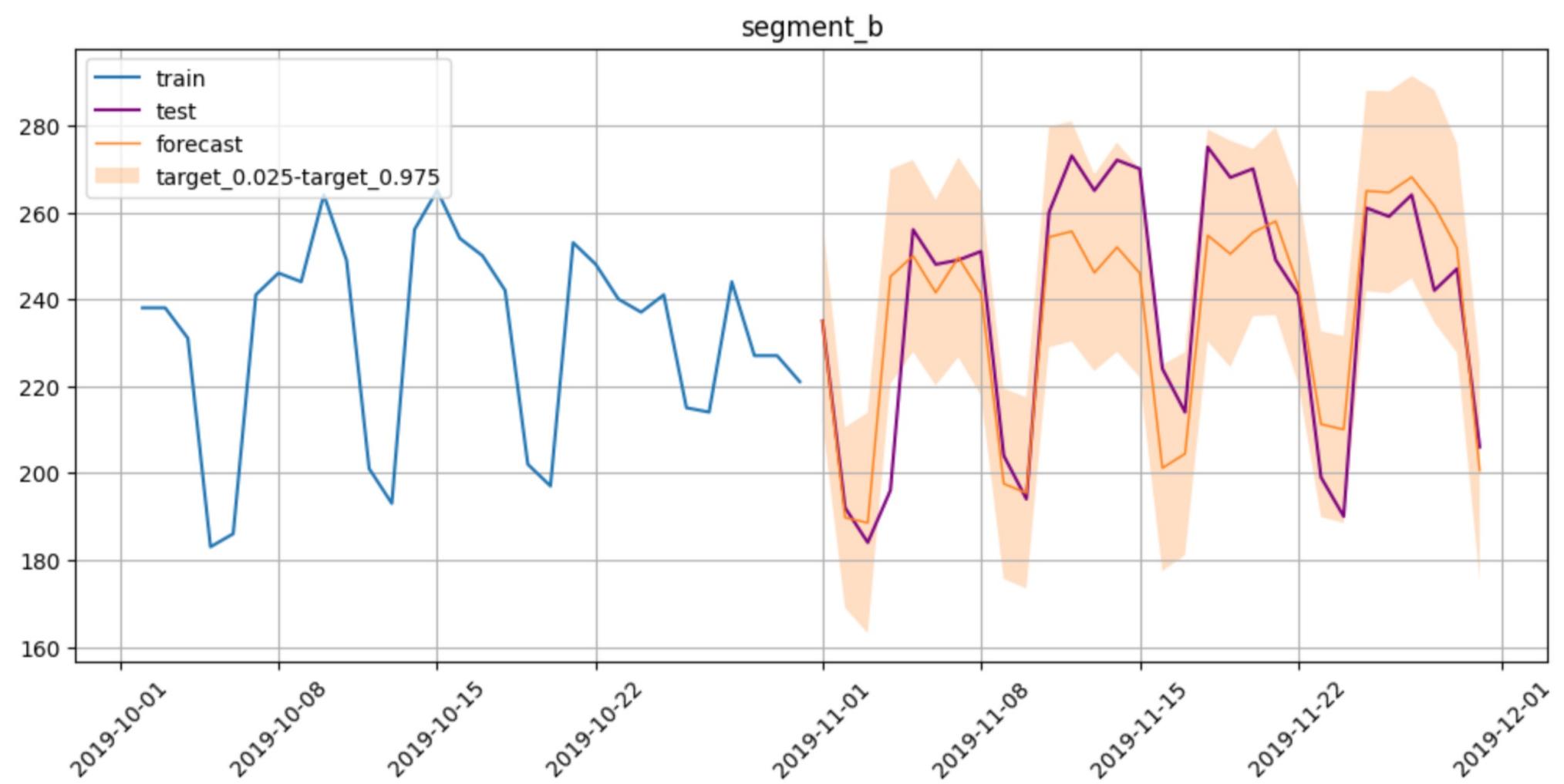
Isolation Forest



Isolation Forest



Forecasting Model



1. Обучаем модель.
2. Строим прогнозы.
3. (online) Ждём фактические данные.



Плюсы:

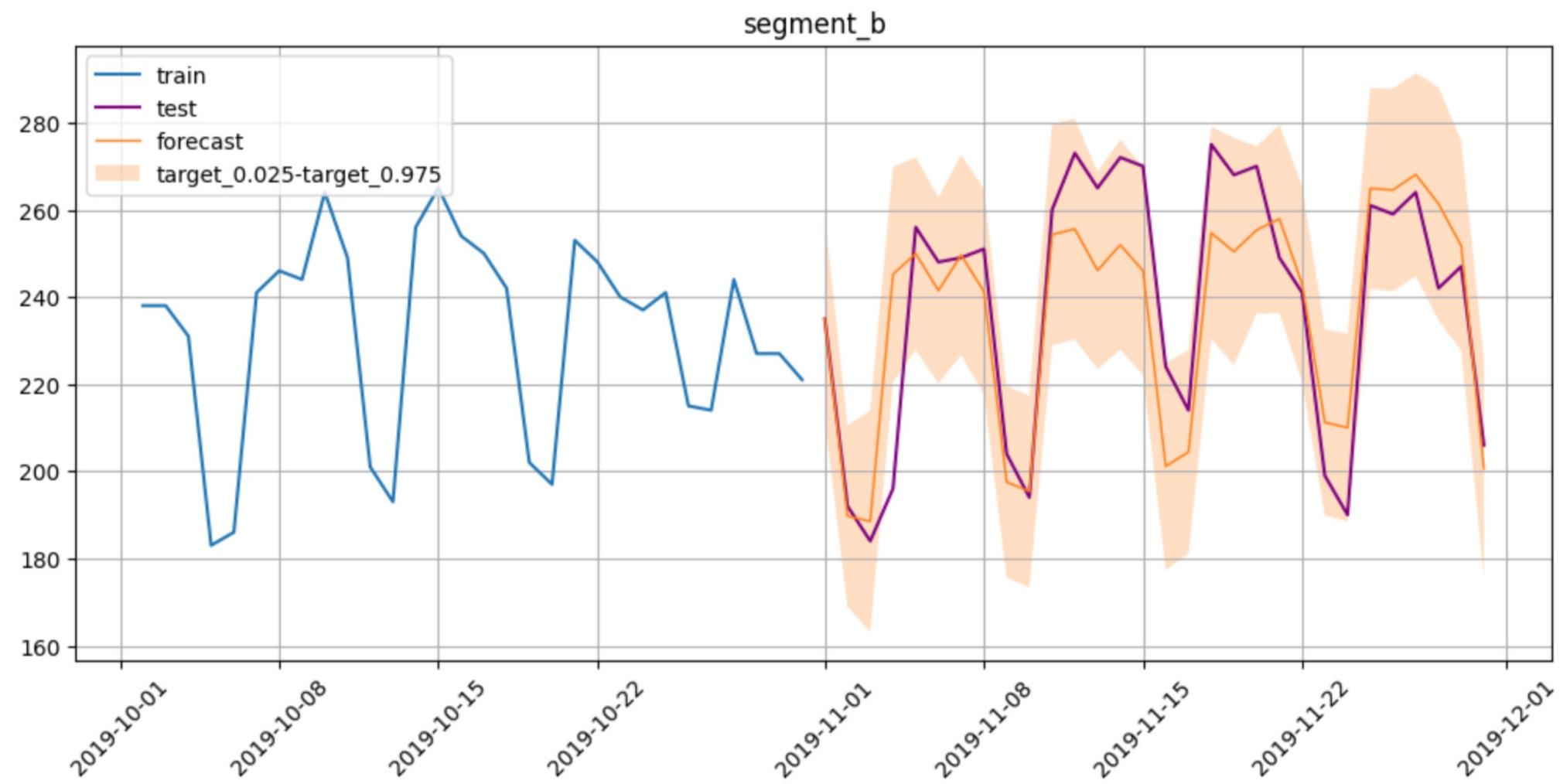
- не нужна разметка;
- учитывает особенности временных рядов (тренд/сезонность/...).



Минусы:

работает при условии, что модель хорошая.

Forecasting Model



Вариации

1. Доверительные интервалы.
2. Прогноз vs факт: нужно подбирать порог отклонения.
3. Двухгоризонтная модель: раннее оповещение о возможных разладках в будущем.

Rule Based (Global Statistics)

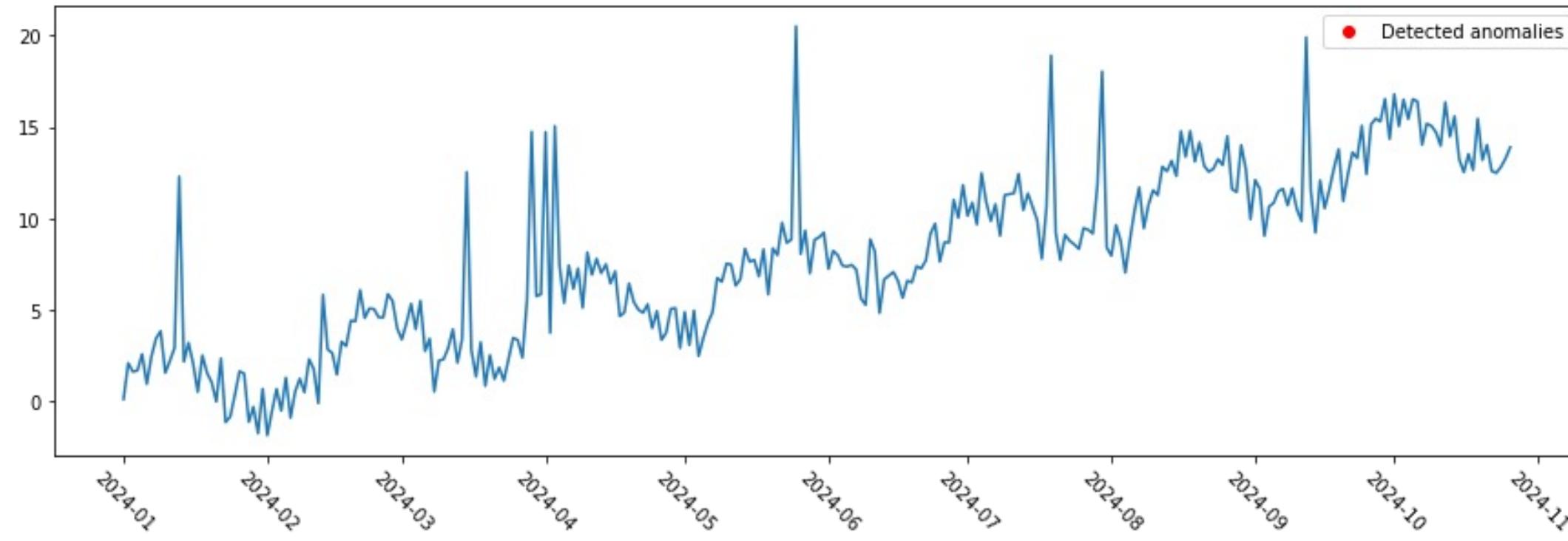
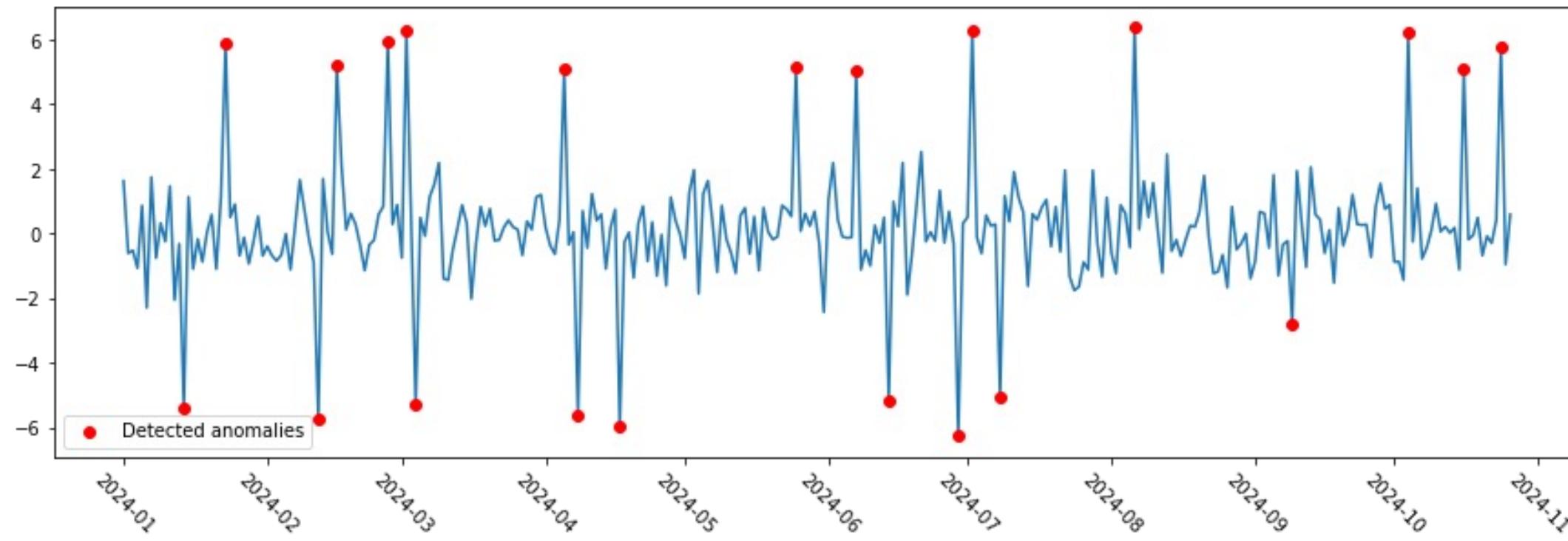


Будем считать аномалией всё, что попадает в интервал.

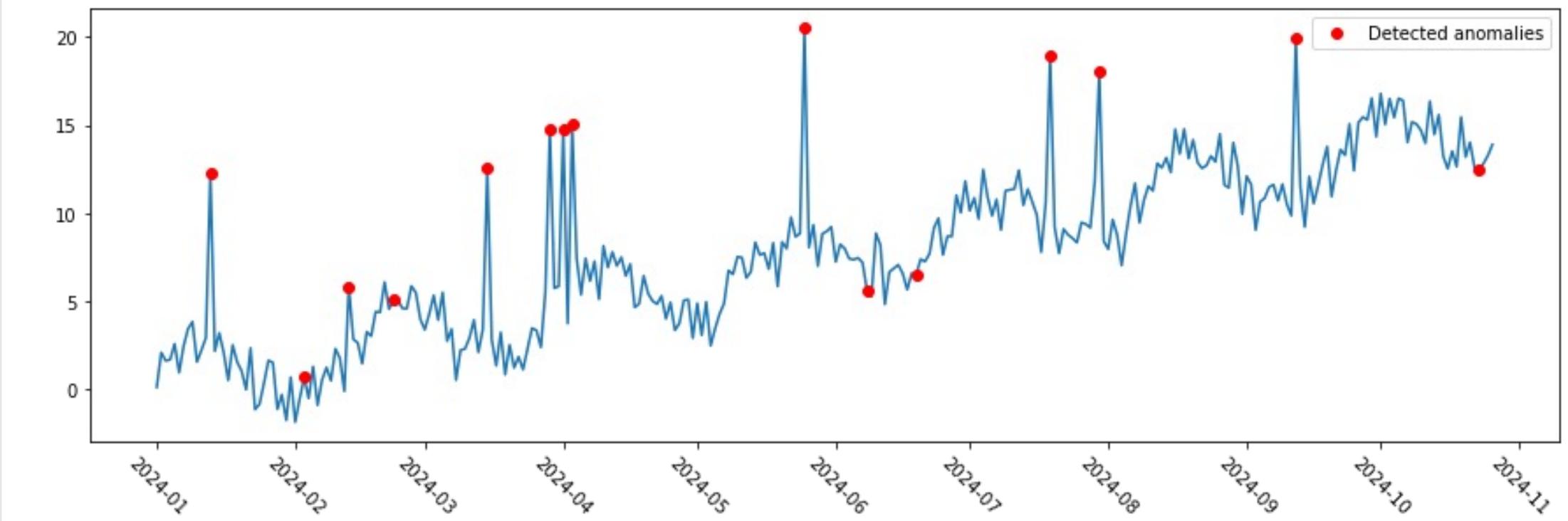
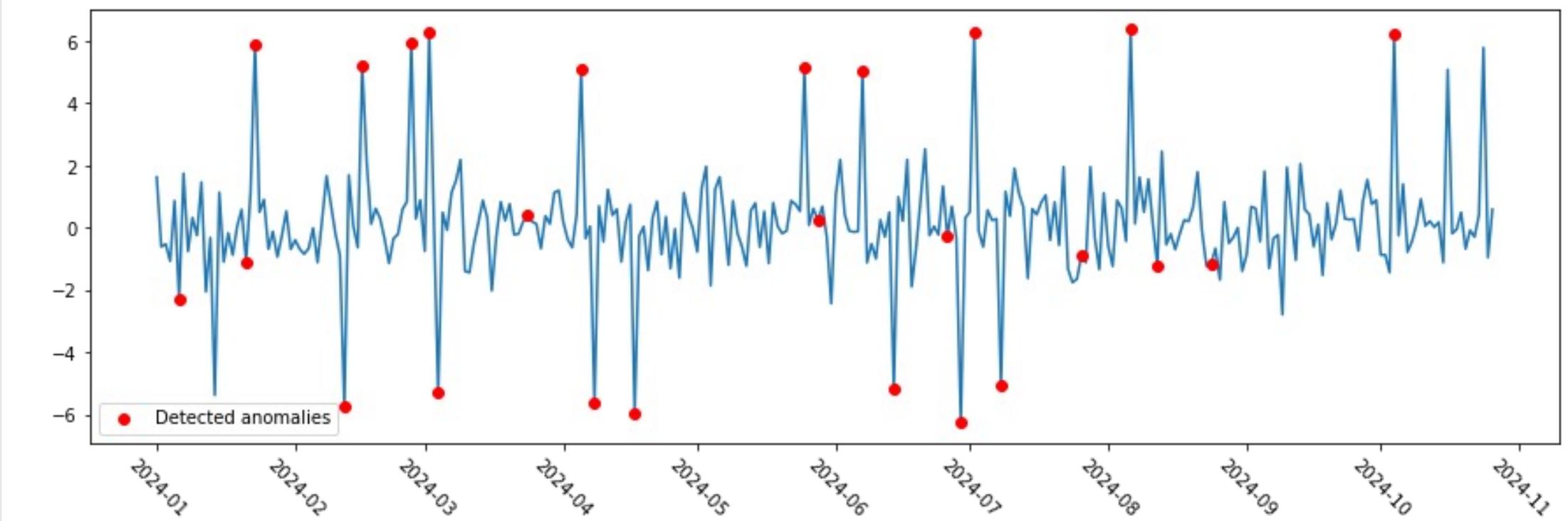
$$Anomaly(t) = I[y_t \in [LowerBound, UpperBound]]$$

- **Threshold** — подбираем пороги руками (бизнес-логика).
- **Quantile** — используем квантили в качестве порогов (статистика).
- **IQR** — используем интерквантильный размах как коридор.

Rule Based (STL + IQR)

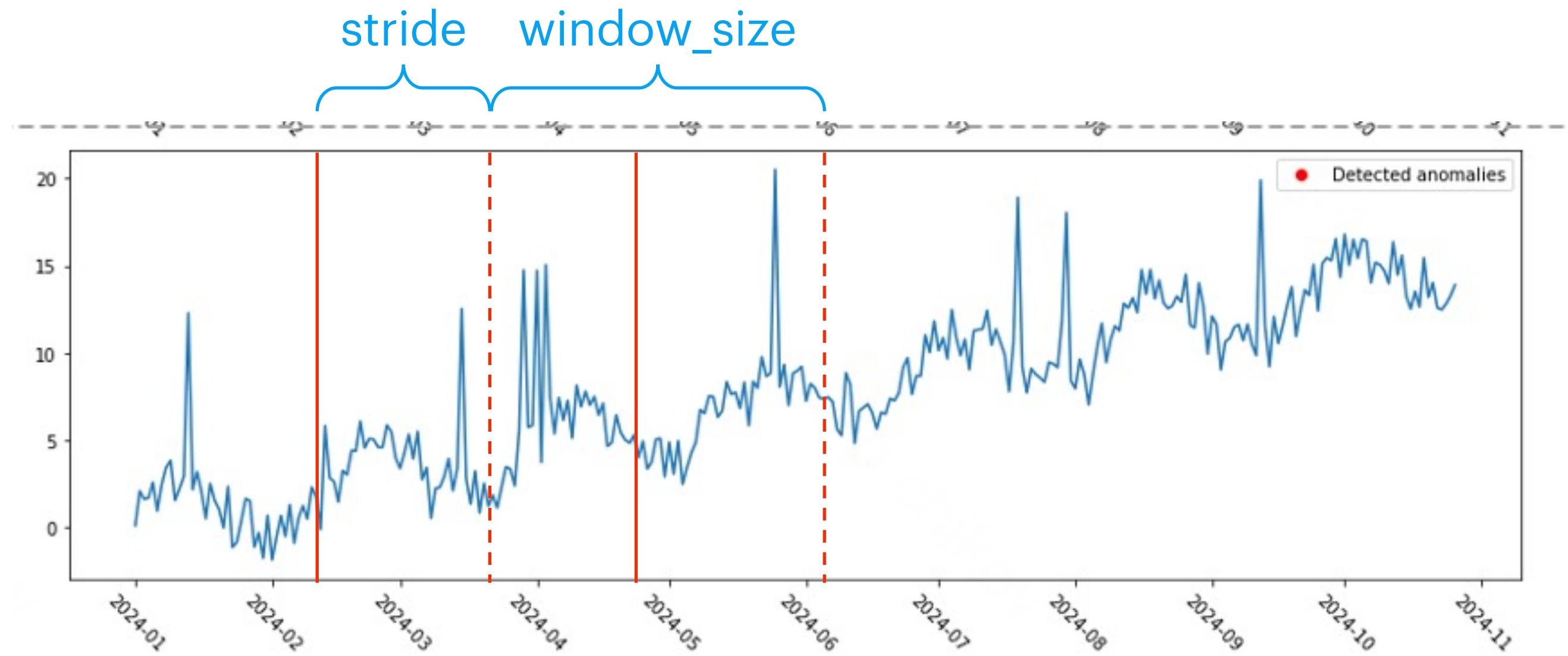


IQR (y)



IQR ($y - \text{seasonal} - \text{trend}$)

Rule Based (Sliding Window Statistics)

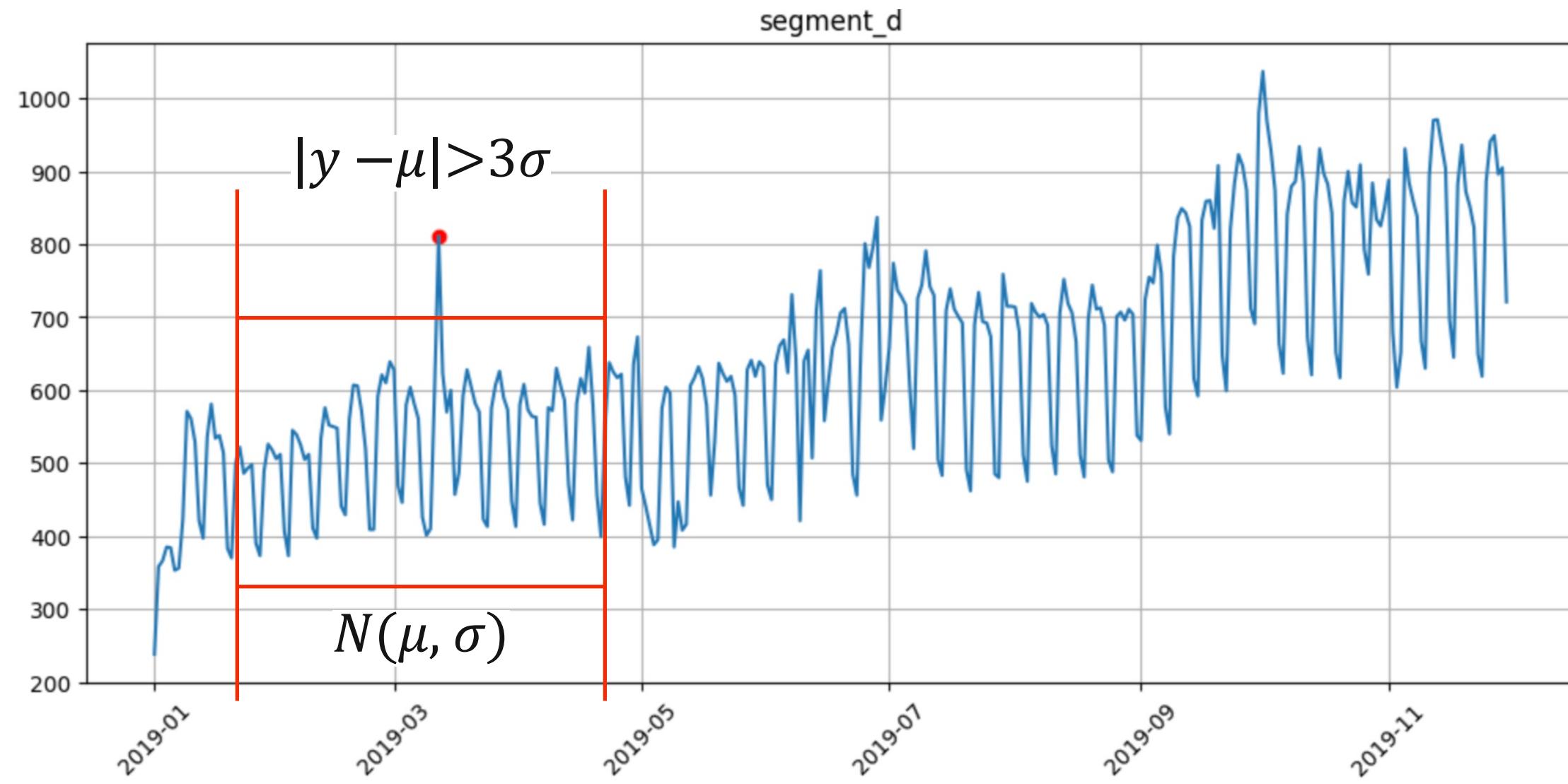


- Сгенерируем набор окон.
- Будем считать статистики внутри окон.
- Каждая точка может быть в нескольких окошках → ансамблируем результаты.

Можем оценивать пороги динамически.

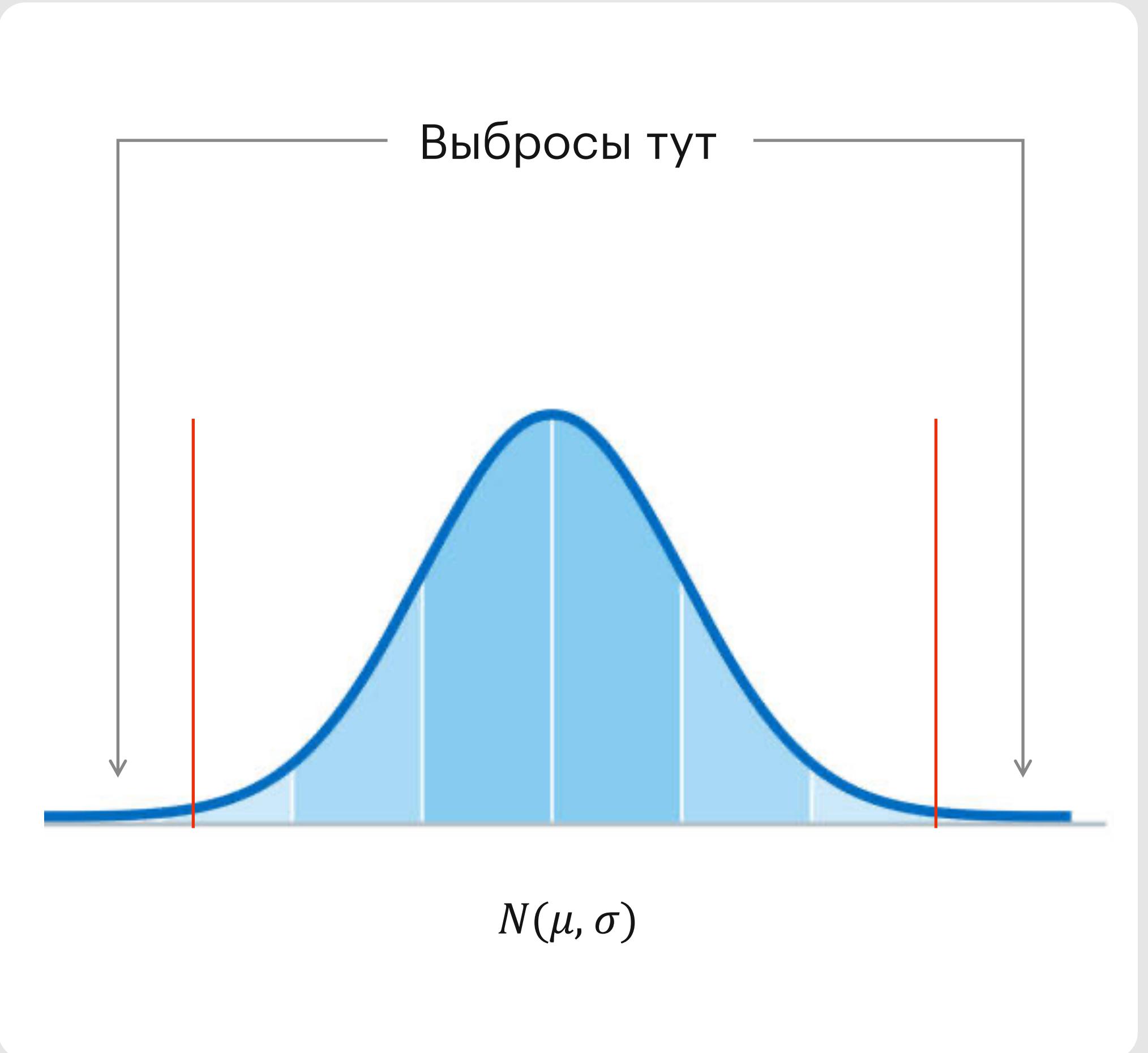
$$\text{Anomaly}(t) = I[y_t \in [\text{LowerBound}_t, \text{UpperBound}_t]]$$

Rule Based (Sliding Window Statistics)

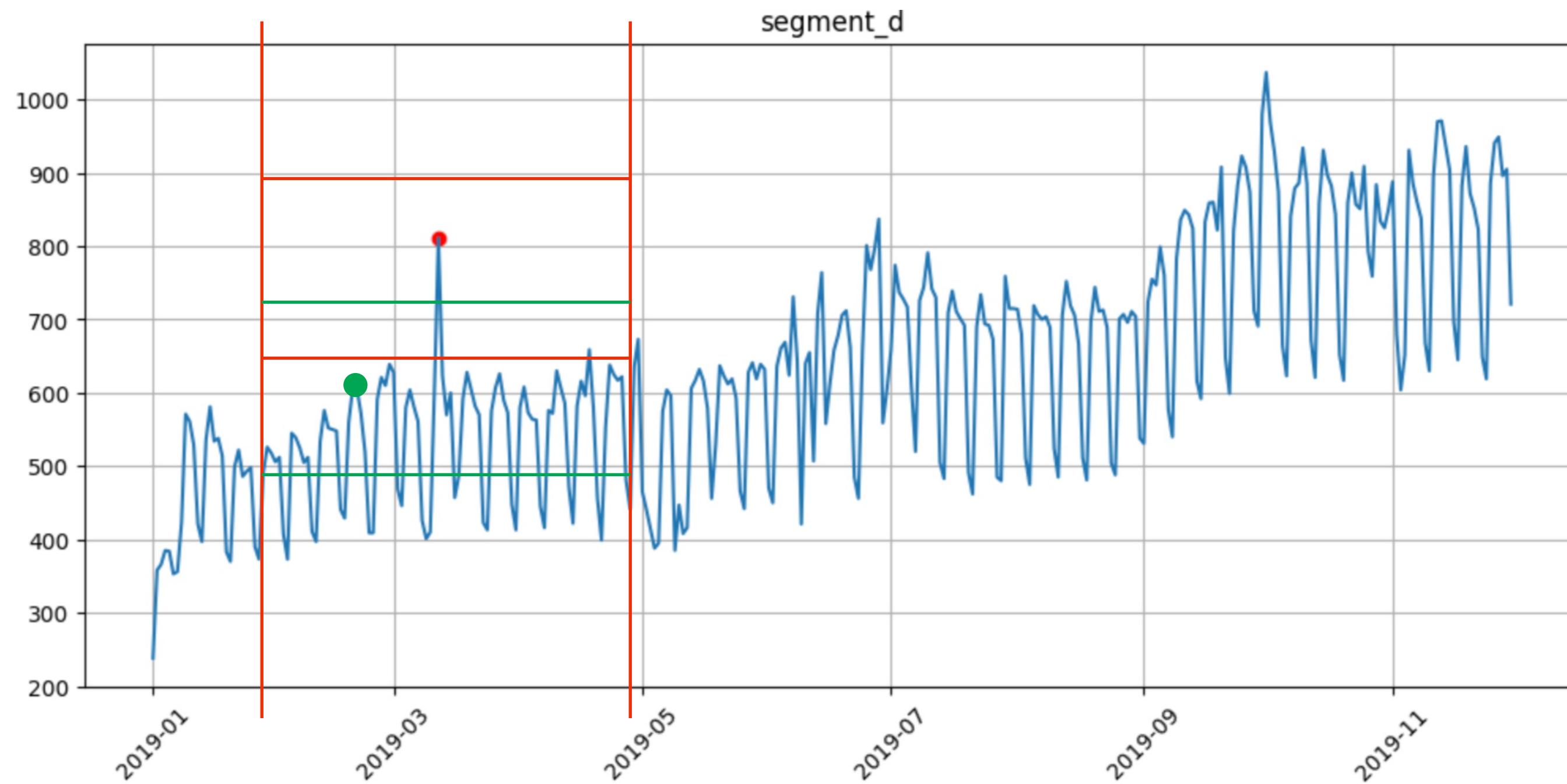


Аномалии из локального нормального распределения

$$Anomaly(t) = I[y_t \in [\mu_t - \alpha * \sigma_t, \mu_t + \alpha * \sigma_t]]$$



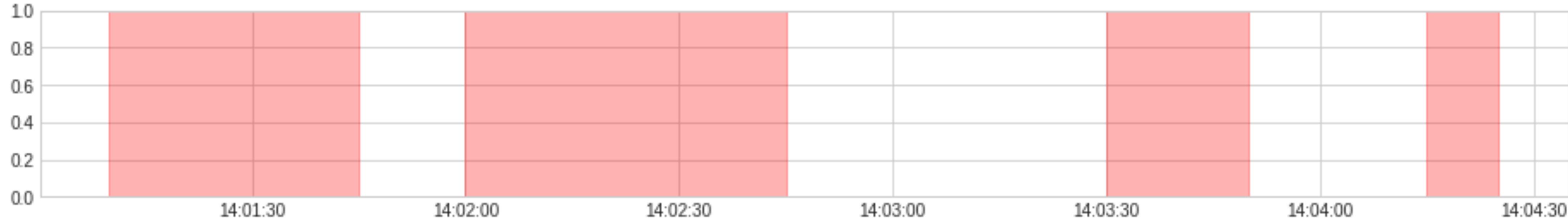
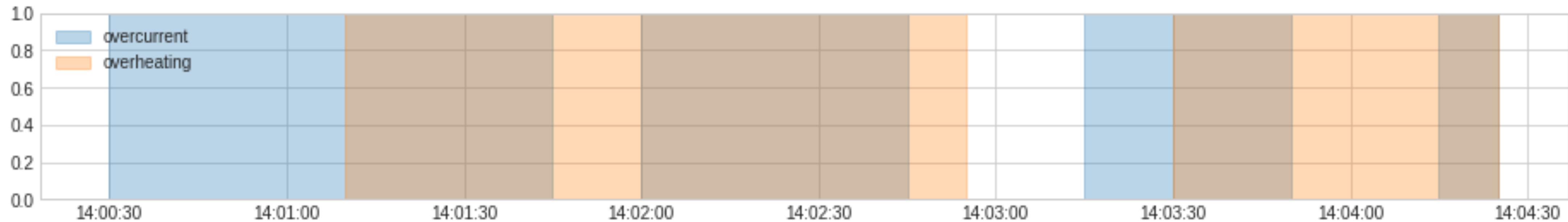
Rule Based (Density)



- Для каждой точки строим окрестность.
- Считаем количество соседей в этой окрестности.
- Мало соседей → аномалия.
- Для каждой точки — несколько окон.

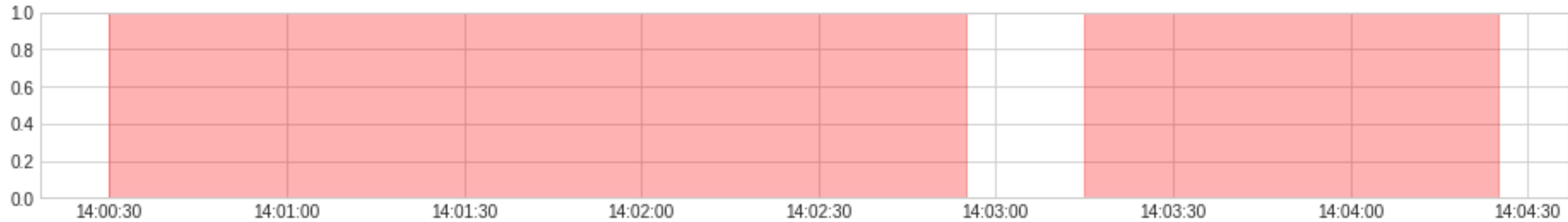
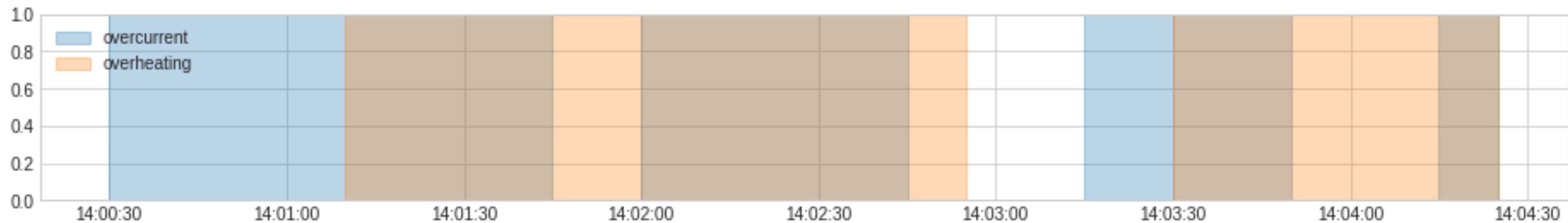
Ансамблирование (AND)

$$\text{Anomaly}(t) = \text{AND}(\text{Anomaly}_1(t), \dots, \text{Anomaly}_n(t))$$



Ансамблирование (OR)

$$\text{Anomaly}(t) = \text{OR}(\text{Anomaly}_1(t), \dots, \text{Anomaly}_n(t))$$



Ансамблирование (Model)

$$\text{Anomaly}(t) = \text{Model}(\text{Anomaly}_1(t), \dots, \text{Anomaly}_n(t))$$

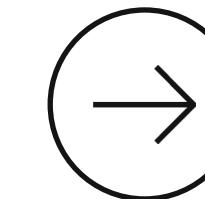
01

Обучаем базовые детекторы на train.

02

Генерируем разметку на test.

	timestamp	model_1	model_2	model_3	model_4	target
0	2000-01-01	0	0	1	0	0
1	2000-01-02	0	0	0	1	1
2	2000-01-03	1	1	0	1	1
3	2000-01-04	1	0	1	1	1
4	2000-01-05	0	0	0	0	0
5	2000-01-06	1	1	1	1	1





Ансамблирование

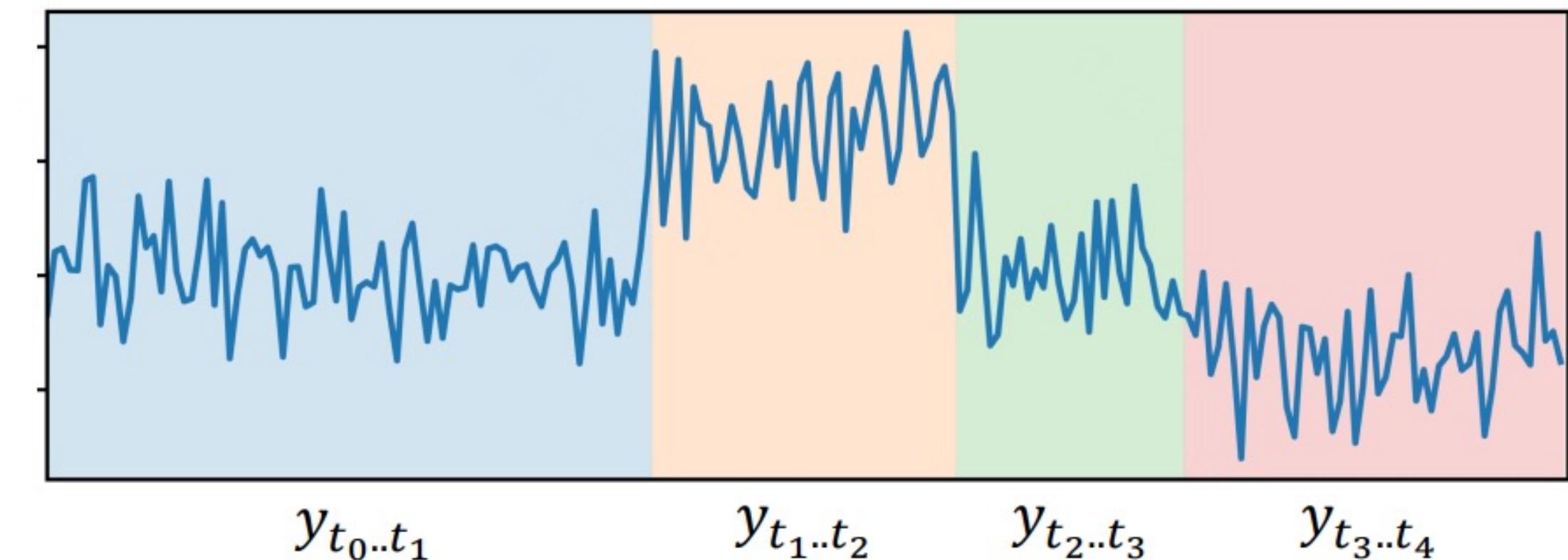
- AND — максимизируем precision (только уверенные аномалии).
- OR — максимизируем recall (все подозрительные точки).
- Model — «оптимально взвешиваем» мнения алгоритмов.
-

Методы поиска точек смены поведения

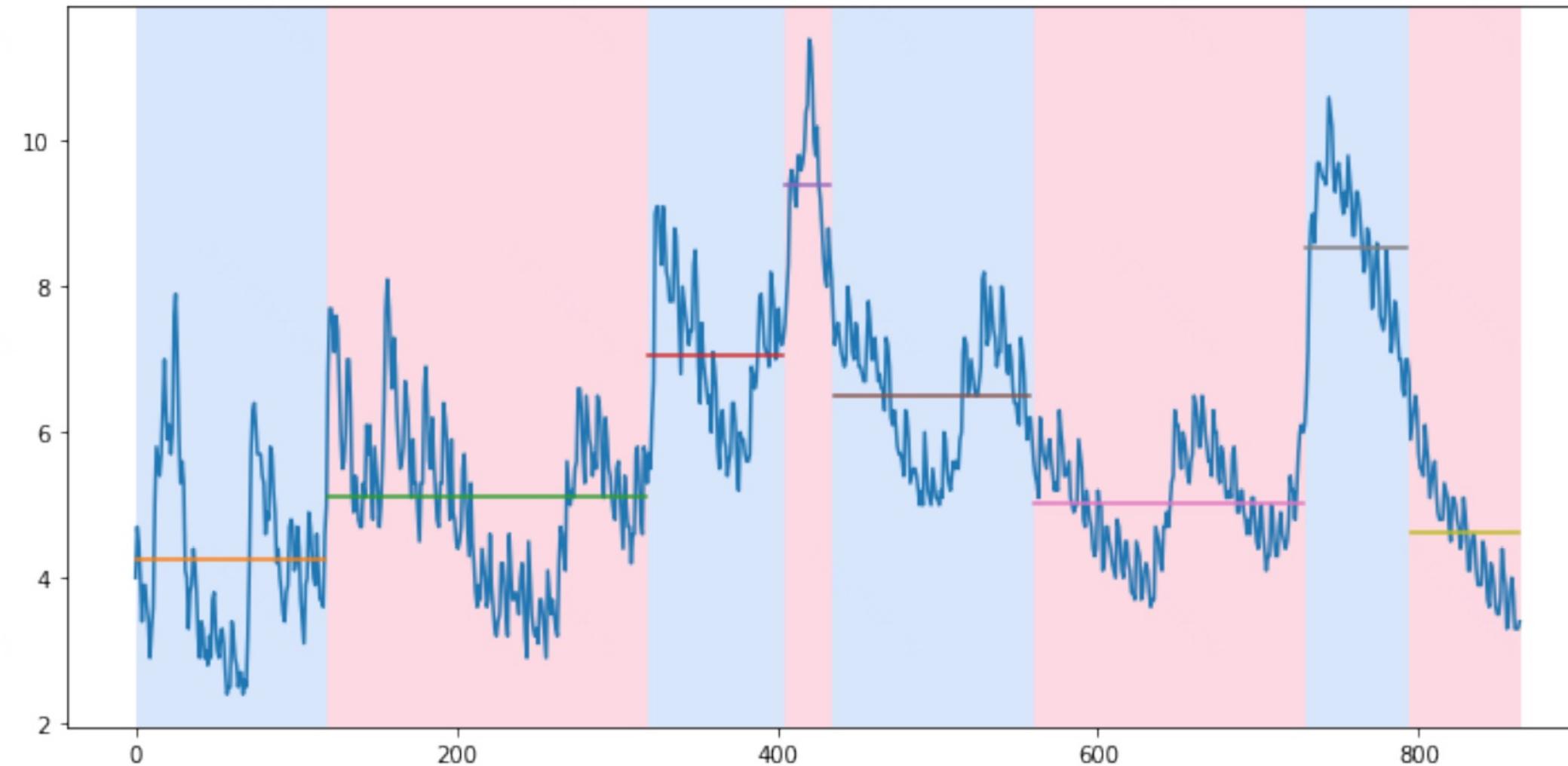
Формализация

- Необходимо найти набор точек $t_1 \dots t_k$, которые максимизируют гомогенность сегментов.
- Функция стоимости (**cost function**) — некоторая оценка гомогенности сегмента.

$$(\hat{t}_1, \dots, \hat{t}_K) = \underset{(t_1, \dots, t_K)}{\operatorname{argmin}} \sum_{k=0}^K c(x[t_k : t_{k+1}])$$

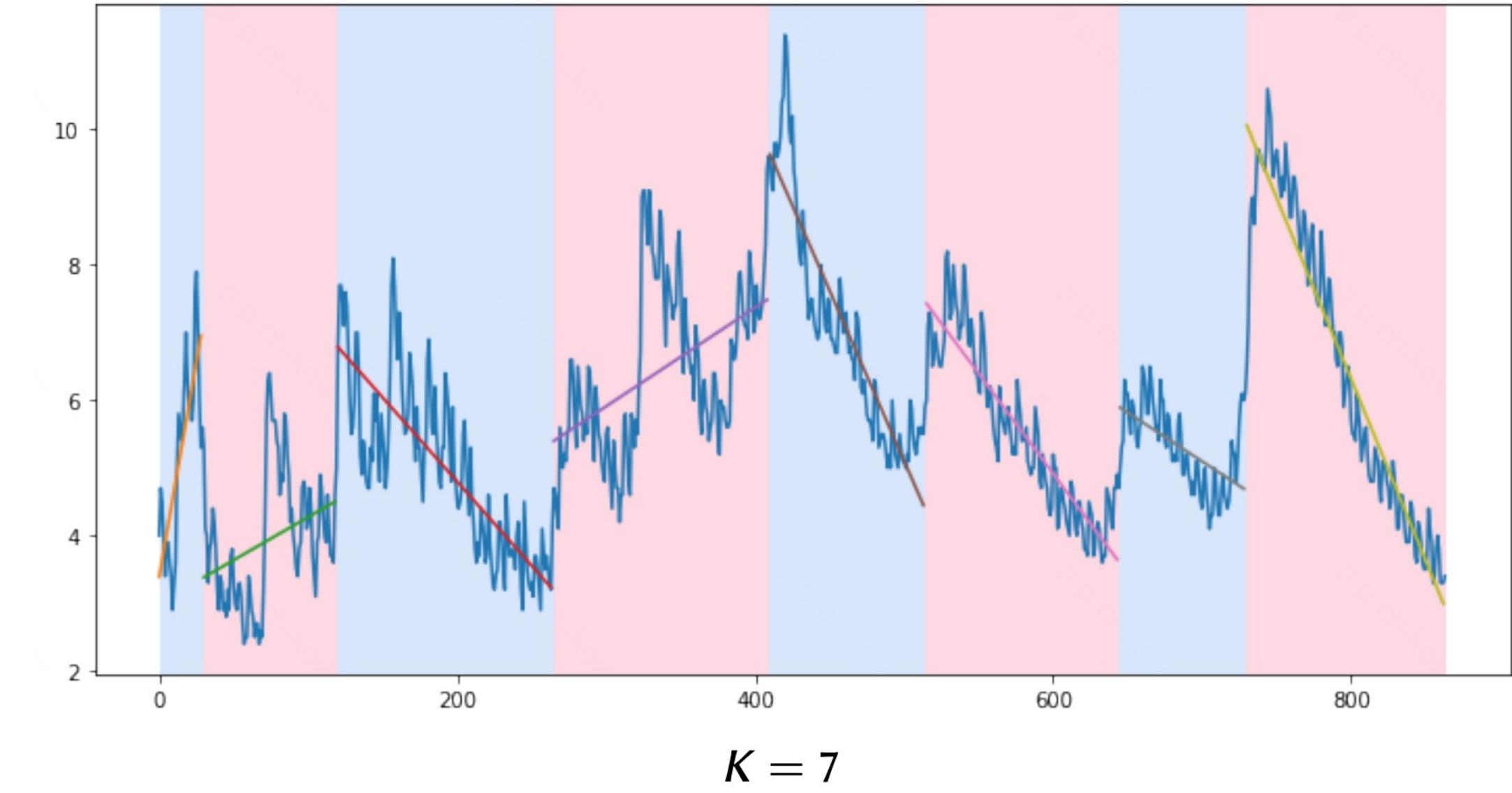


Cost Function



L2 — смена уровня

$$c(y_I) = \sum_{t \in I} \|y_t - \bar{y}\|_2^2$$



Linear — смена тренда

$$c(y_I) = \min_{\delta \in \mathbb{R}^p} \sum_{t \in I} \|y_t - \delta' x_t\|_2^2.$$

Search Method

Точное решение:

- основано на методе динамического программирования,
- находит истинное решение оптимизационной задачи,
- долго работает.

Приближённые решения:

- основаны на оценке «аномальности» каждой точки в отдельности,
- находят субоптимальные решения,
- работают быстро.

[\(Дальше — про них.\)](#)

Логика приближённых решений



Основные компоненты

01

Стратегия перебора точек.

02

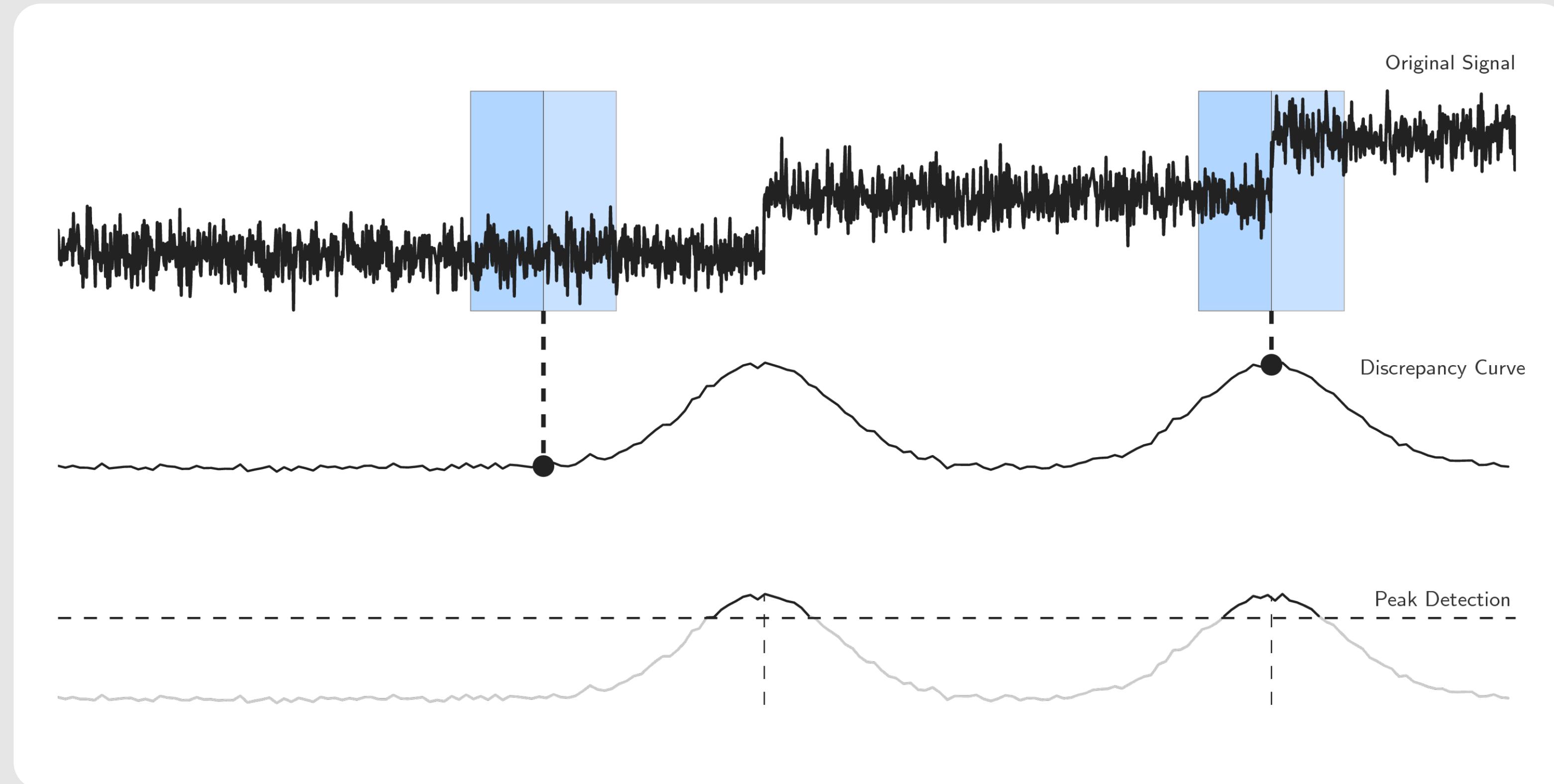
Discrepancy function — функция оценки стоимости разделения/слияния отрезков в смысле гомогенности.

03

Точки с наибольшим/наименьшим значением discrepancy function и будут искомыми.

$$d(y_{u..v}, y_{v..w}) = c(y_{u..w}) - c(y_{u..v}) - c(y_{v..w})$$

Rolling Window



- Двигаемся окошком размера $2w$ вдоль ряда.
- Делаем оценку для середины окошка.

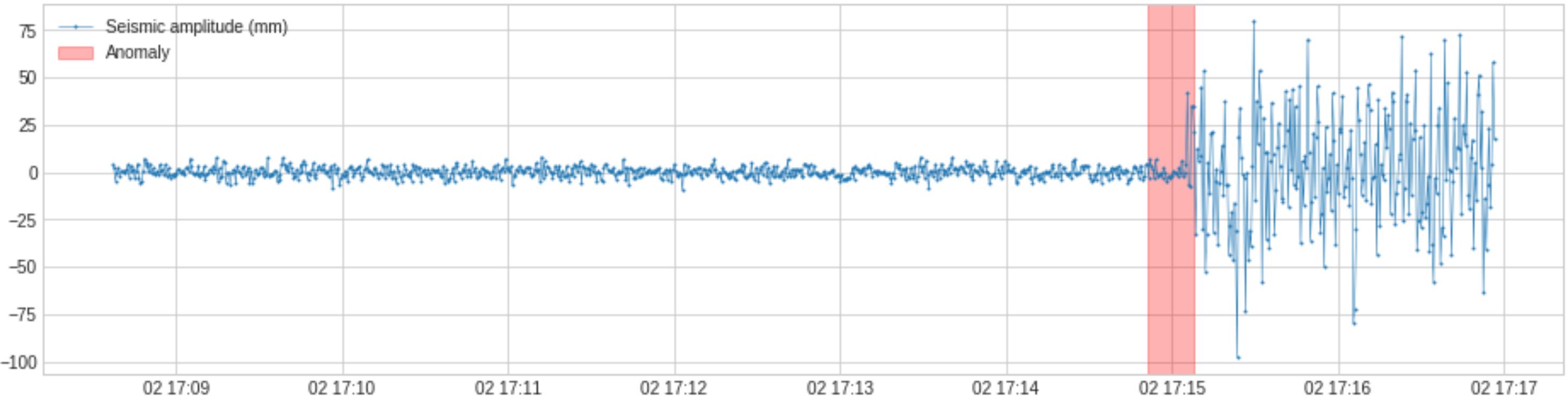
Rolling Window



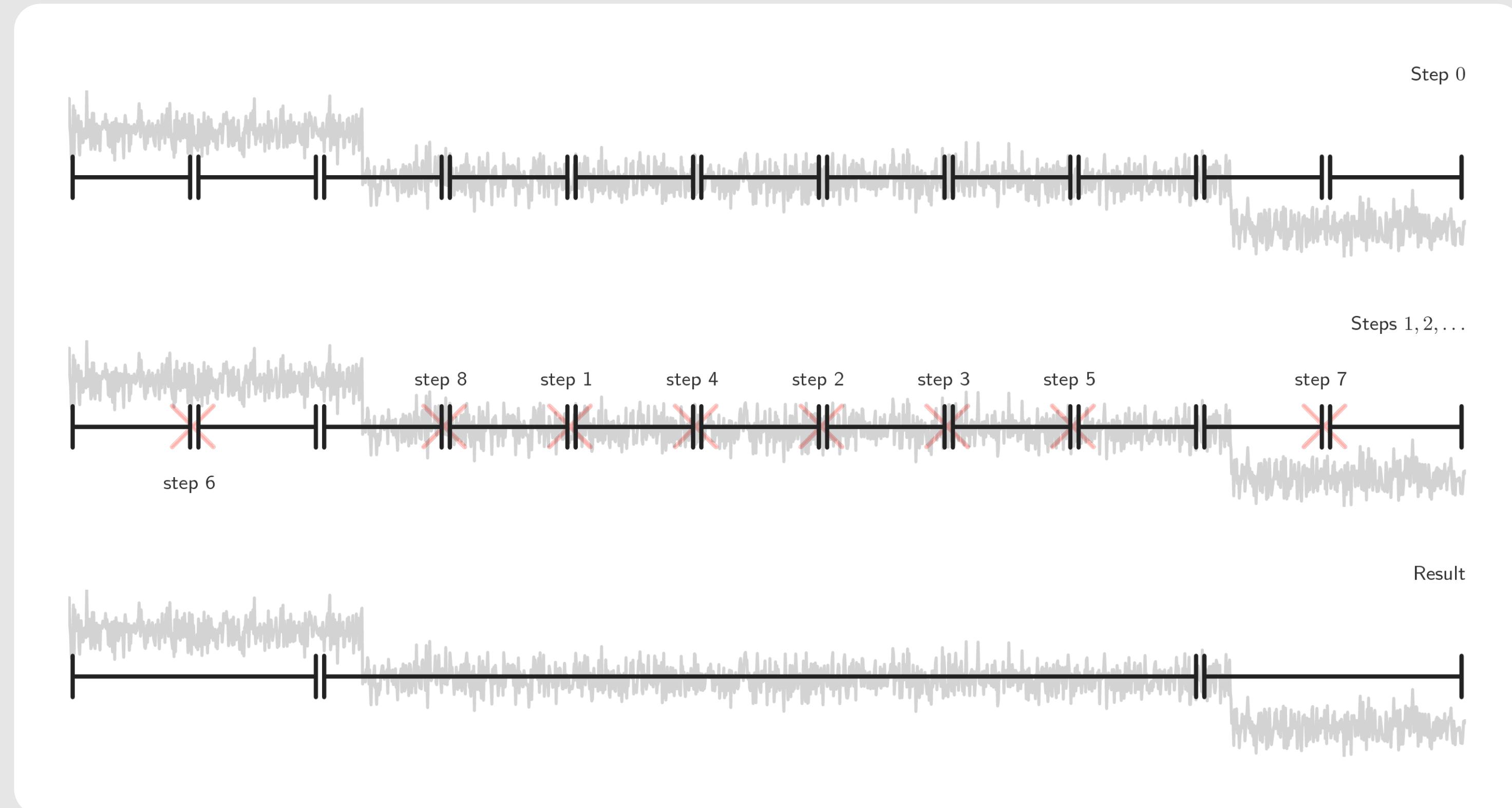
Level Shift
(mean)



Volatility Shift
(std)

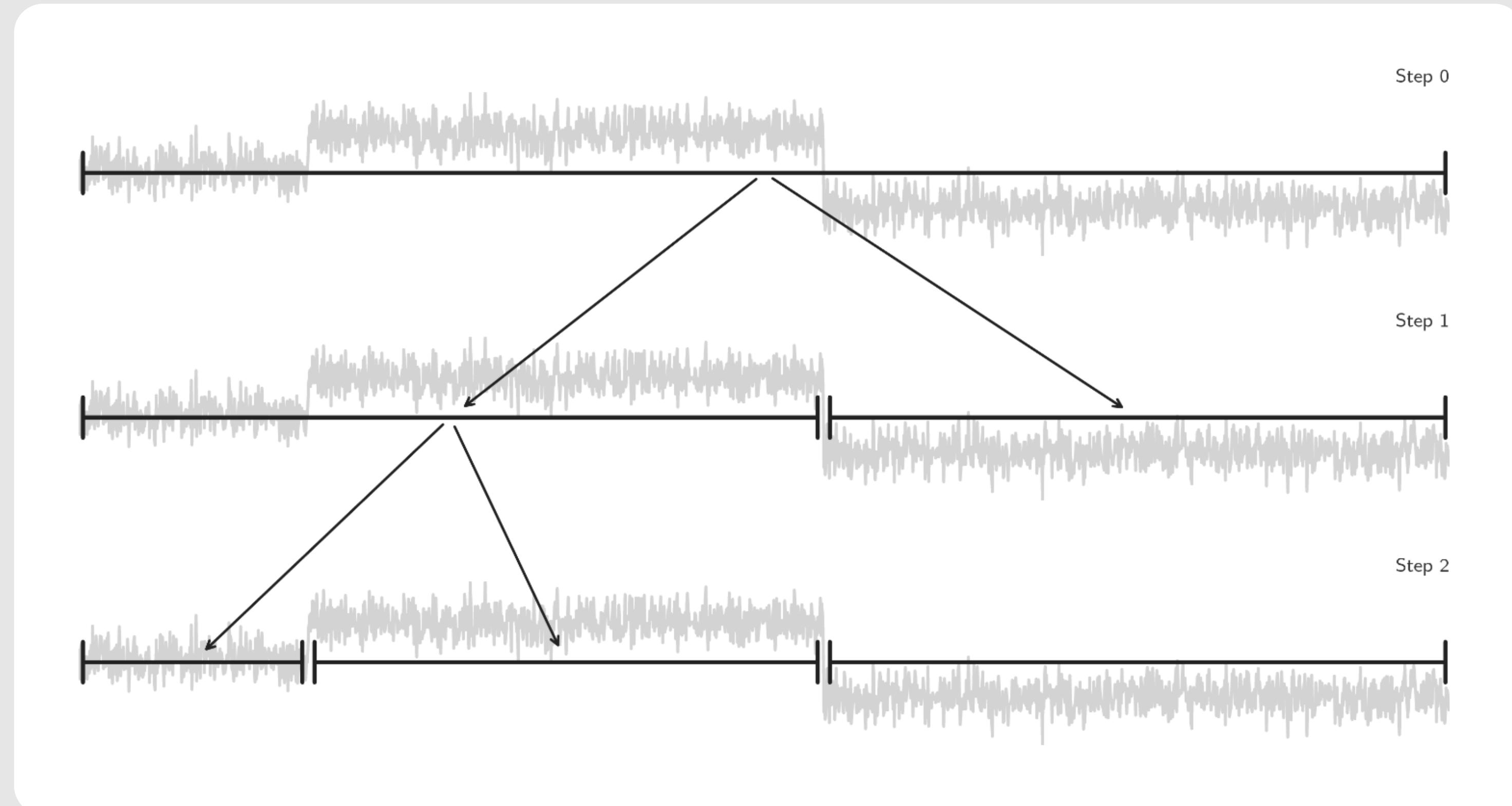


Bottom UP Segmentation



1. Равномерно раскидываем потенциальные точки N .
2. Для всех соседних отрезков считаем стоимость слияния.
3. Сливаем самую дешёвую пару.
4. Останавливаемся по критерию.

Binary Segmentation



Повторяем рекурсивно.

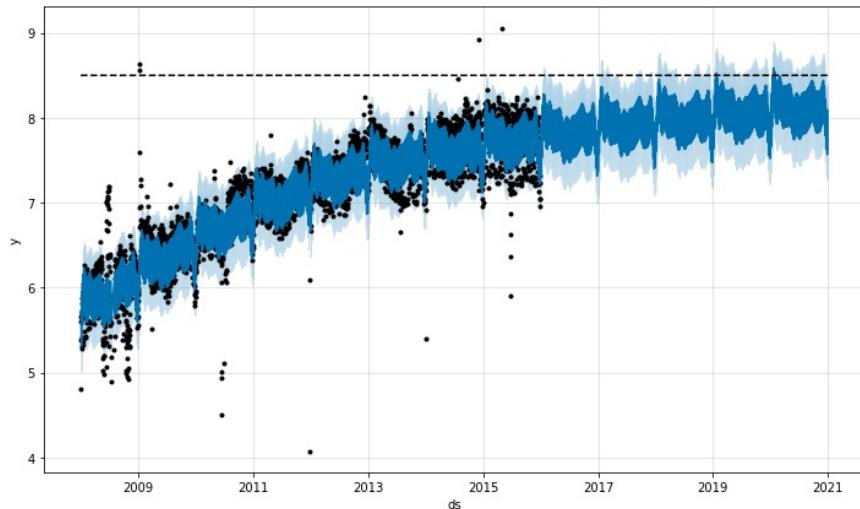
1. Делим порезок пополам всеми возможными способами.
2. Считаем стоимость разбиения для каждого способа.
3. Выбираем самый дорогой.

Prophet

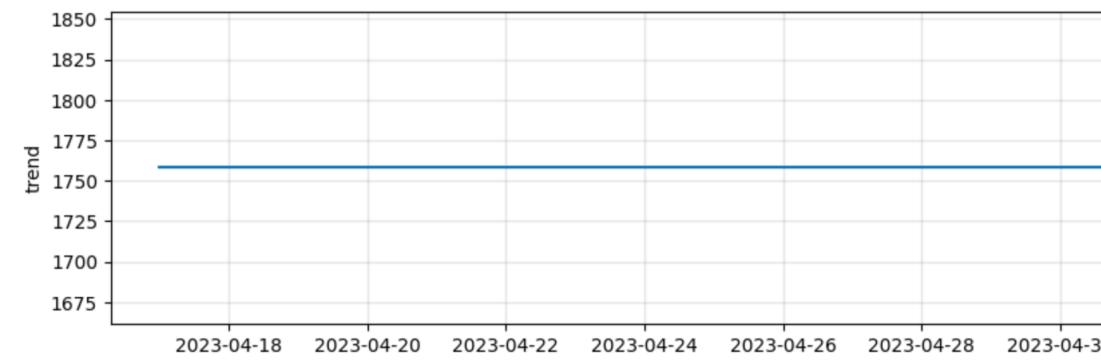
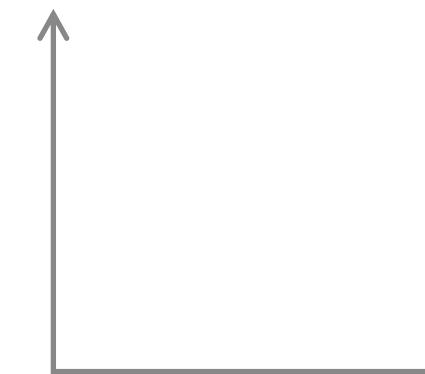
```

1 ProphetModel(
2     growth: str = "linear",
3     changepoints: Optional[List[datetime]] = None,
4     n_changepoints: int = 25,
5     changepoint_range: float = 0.8,
6     changepoint_prior_scale: float = 0.05,
7 )

```



logistic

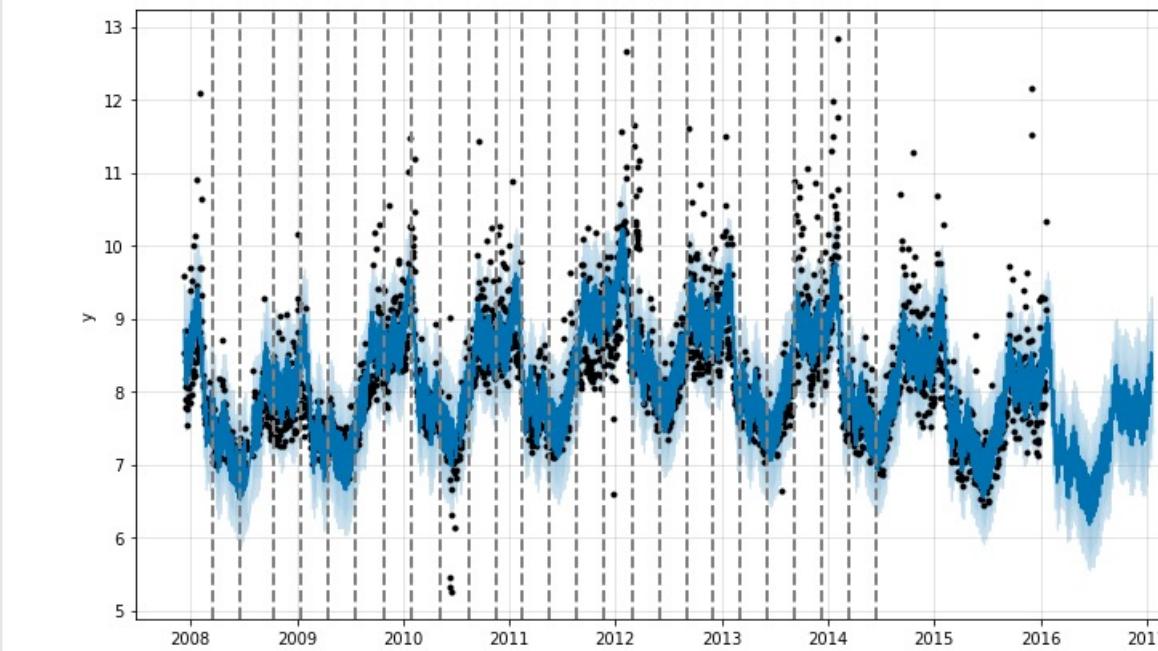


flat

growth

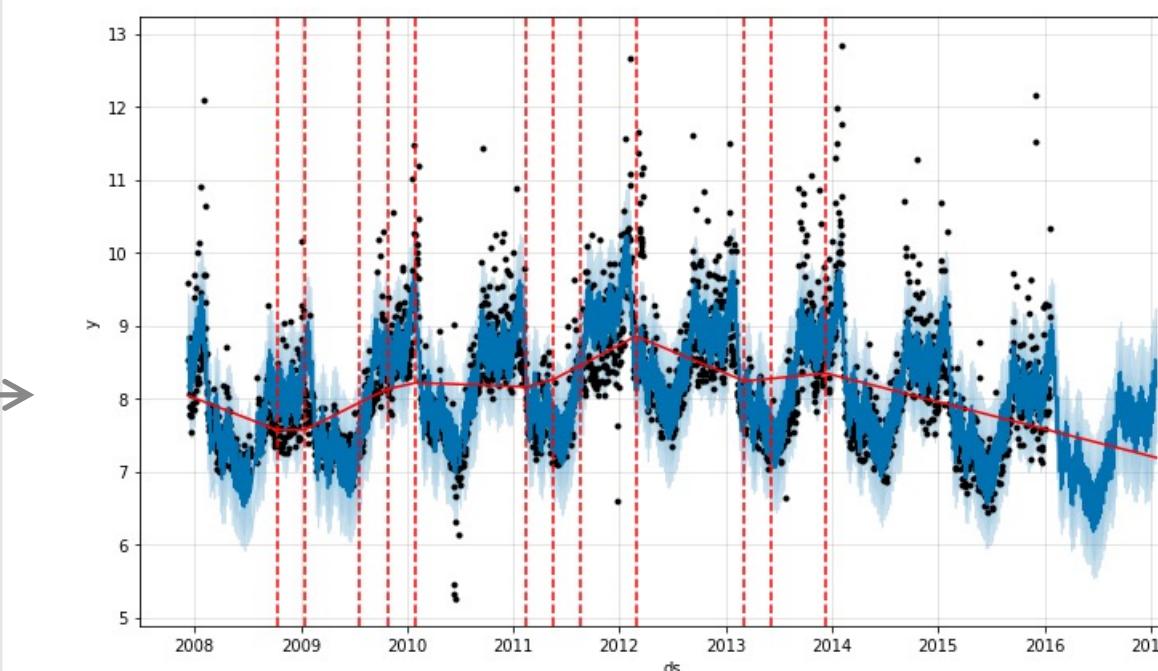
linear

change_point_range



01

Равномерно
разбрасываем
n_changepoints
на **change_point_range**
процентов истории.



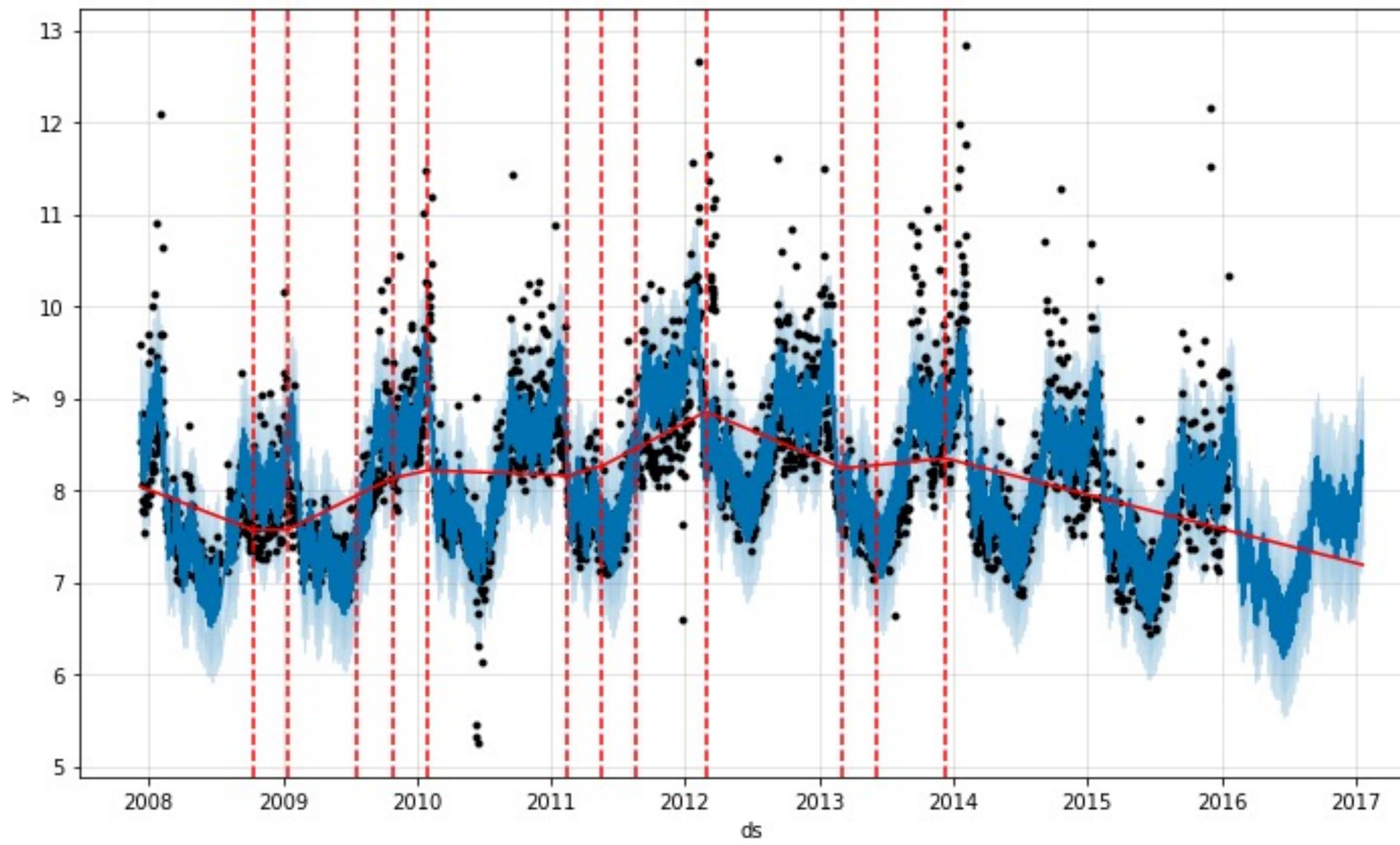
02

Количество выбранных
точек зависит
от **changepoint_prior_scale**.

Prophet

$$g(t) = (k + \mathbf{a}(t)^\top \boldsymbol{\delta})t + (m + \mathbf{a}(t)^\top \boldsymbol{\gamma})$$

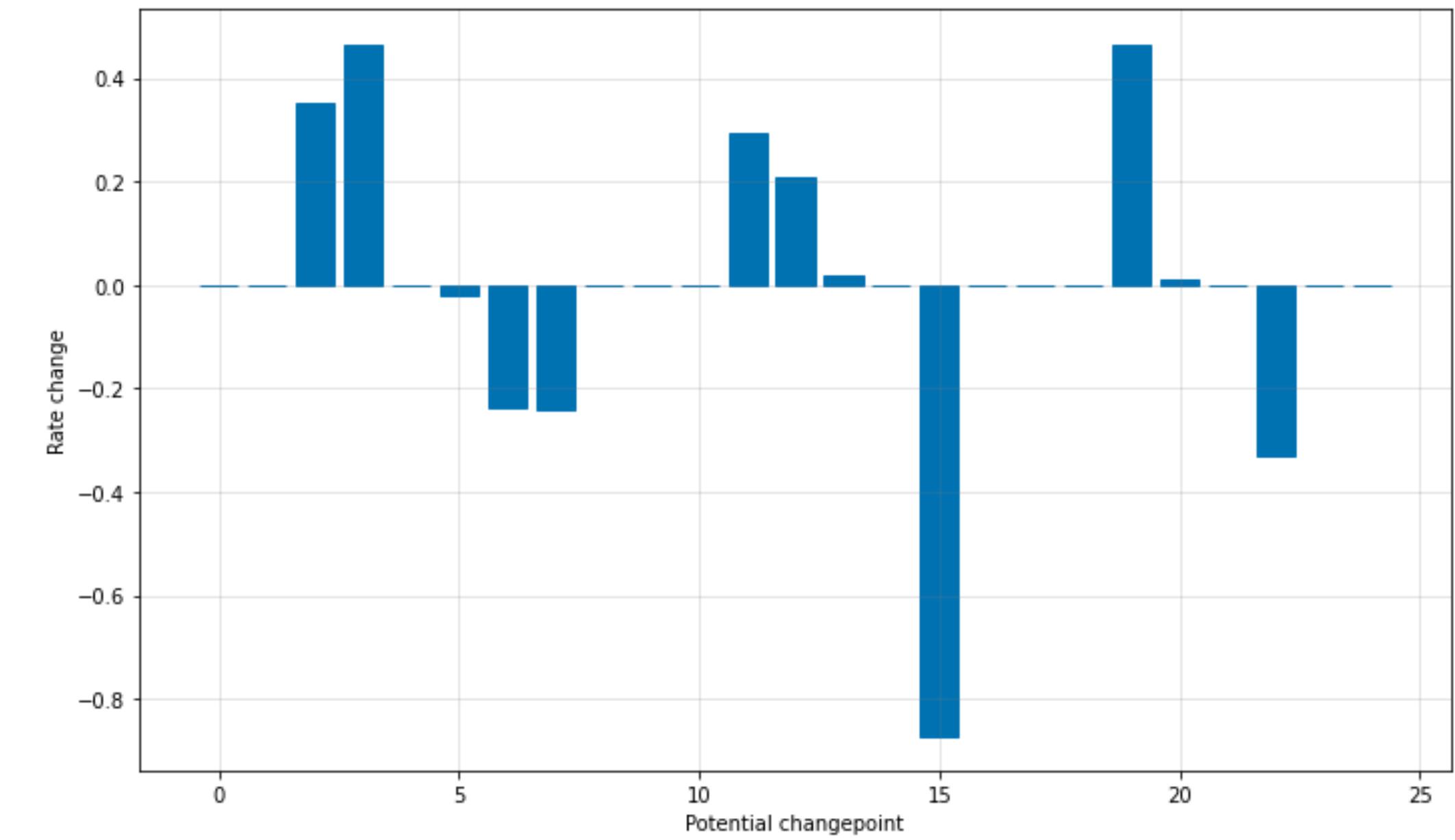
$$\delta_j \sim \text{Laplace}(0, \tau)$$



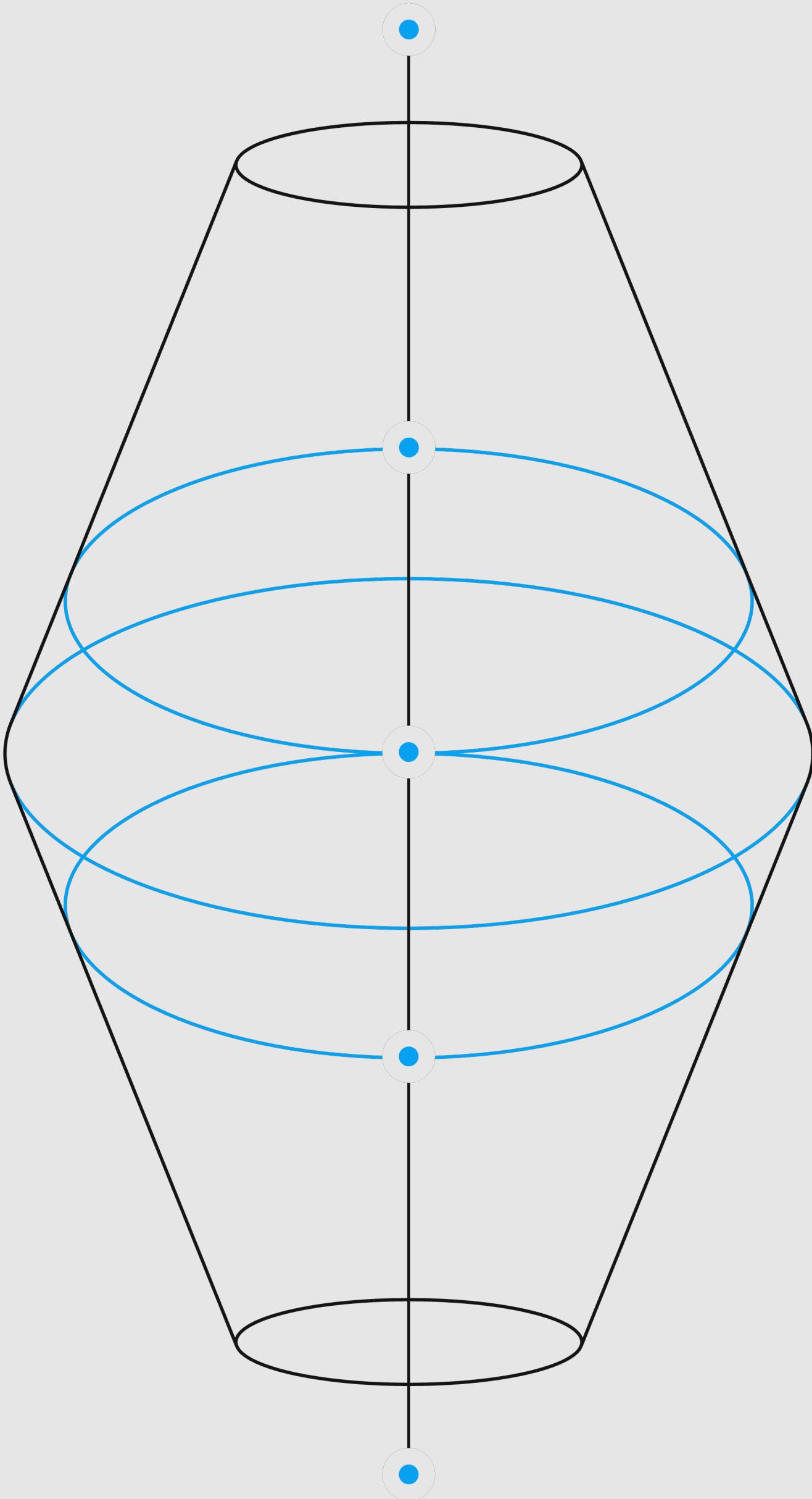
MAP оценка

$$\hat{\theta}_{\text{MAP}}(x) = \arg \max_{\theta} f(\theta \mid x)$$

(Аналогично Lasso-регрессии.)



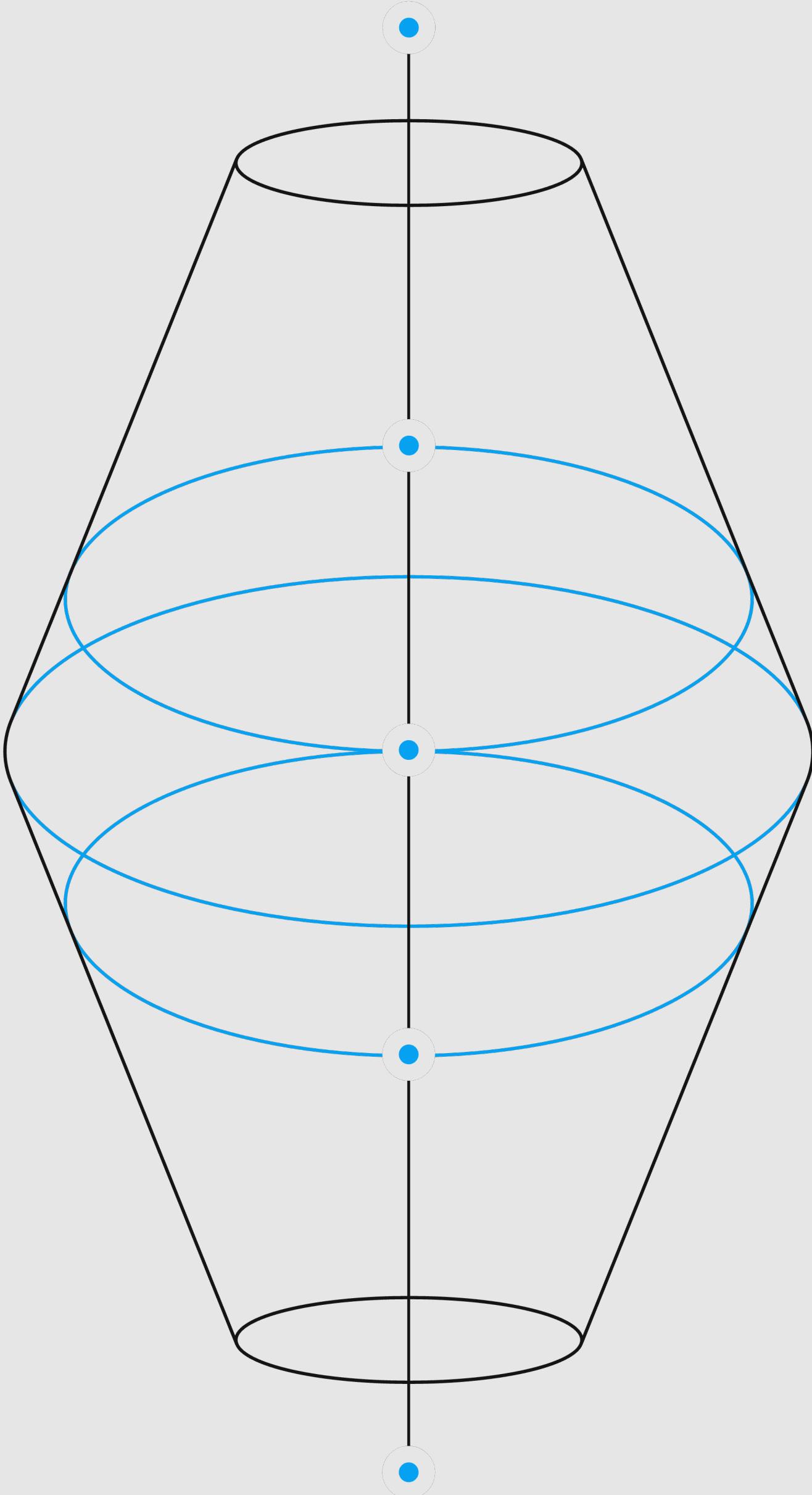
Заключение



Особенности задачи

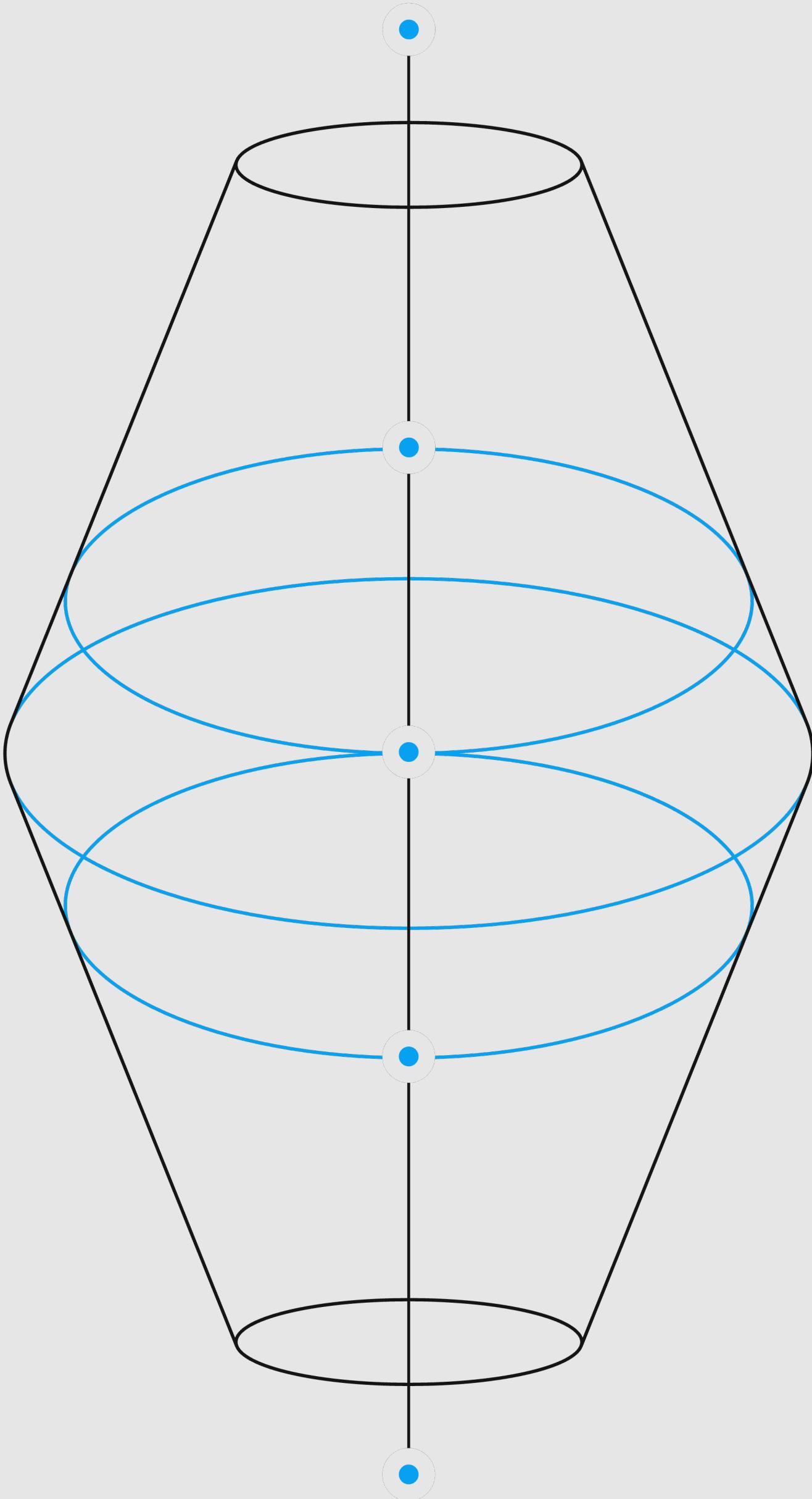
- Сложно оценивать качество.
- Нет общепринятых бенчмарков.
- Тонкая настройка под заказчика.
- SLA на время отклика.
- Необходима адаптация методов под особенности Time Series.

Про что поговорили



- Где возникает задача обнаружения аномалий
- Как собирать разметку
- Методы обнаружения точечных аномалий
- Методы обнаружения точек смены поведения

Библиотеки



- [ruptures](#) – больше про сегментацию.
- [ADTK](#) – rule-based-подходы.
- [PyOD](#) – методы для табличек.
- Фреймворки....