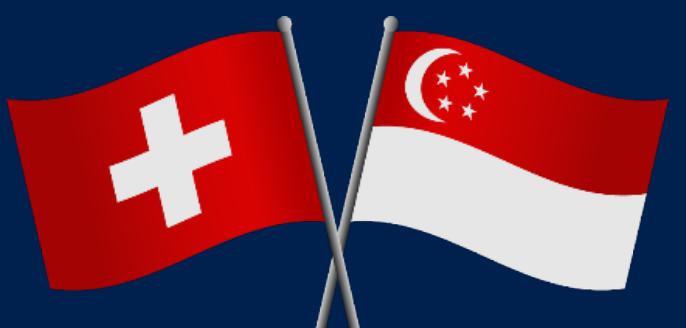


## Поиск аномалий во временных рядах системы Acronis Storage

Научный руководитель: к.ф.-м.н., Кулага Андрей Александрович

Студент: Угнивенко Виталий, 673

Москва 2020



Dual headquarters  
in Switzerland and Singapore

# Актуальность работы

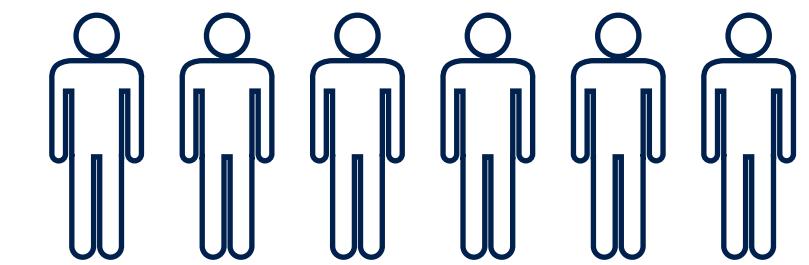
10-ки

Хранилищ данных  
по всему миру



500 000 000

Частных  
клиентов



500 000

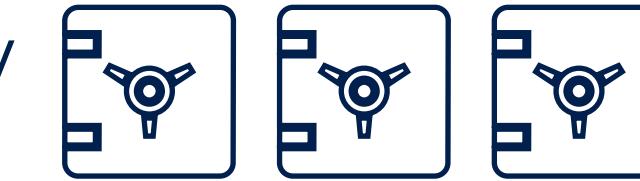
Корпоративных  
клиентов



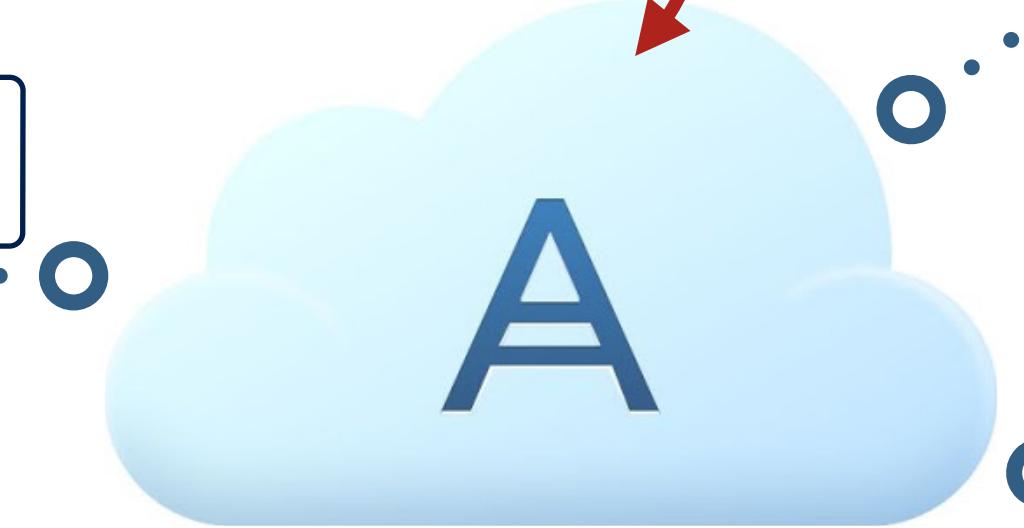
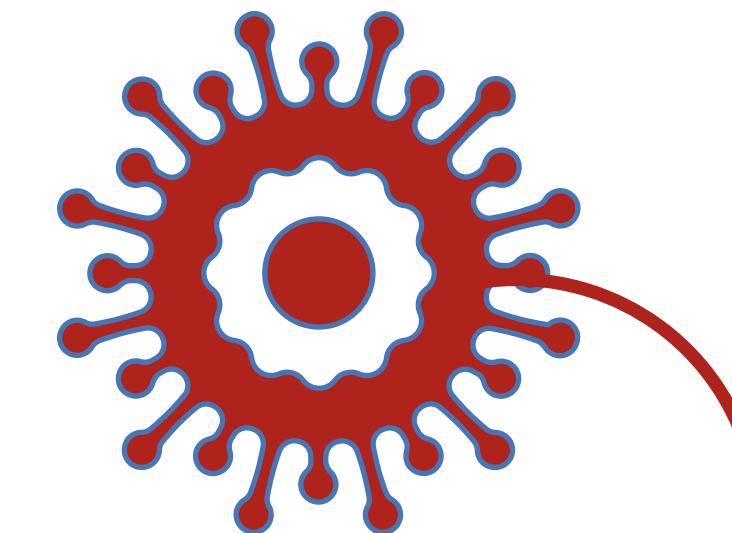
# Актуальность работы

10-ки

Хранилищ данных  
по всему миру

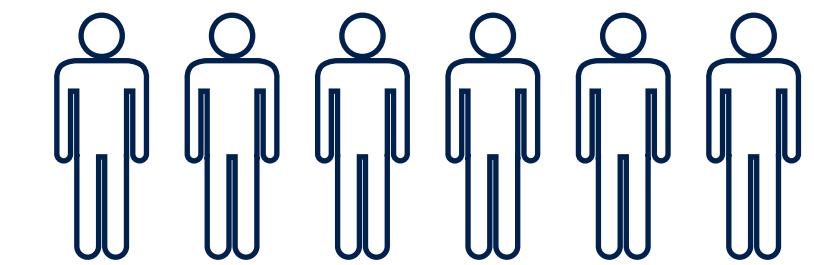


Аномалии



500 000 000

Частных  
клиентов



500 000

Корпоративных  
клиентов



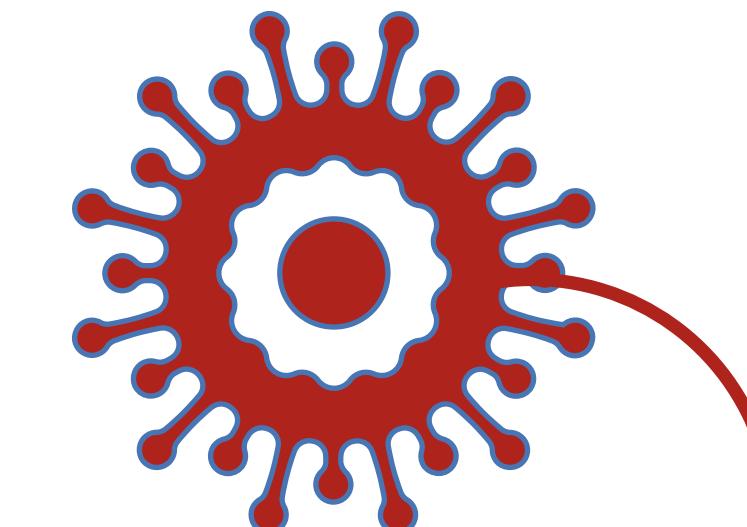
# Актуальность работы

10-ки

Хранилищ данных  
по всему миру

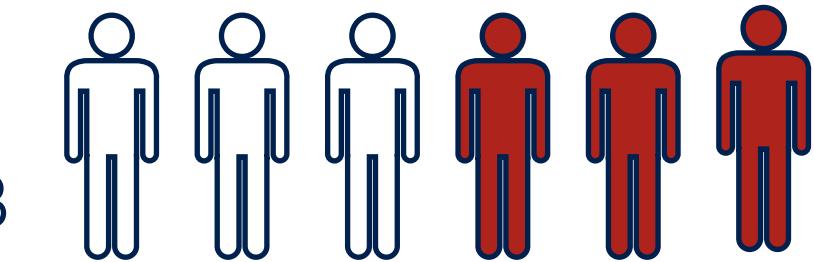


Аномалии



500 000 000

Частных  
клиентов



500 000

Корпоративных  
клиентов



# Актуальность работы

10-ки

Хранилищ данных  
по всему миру



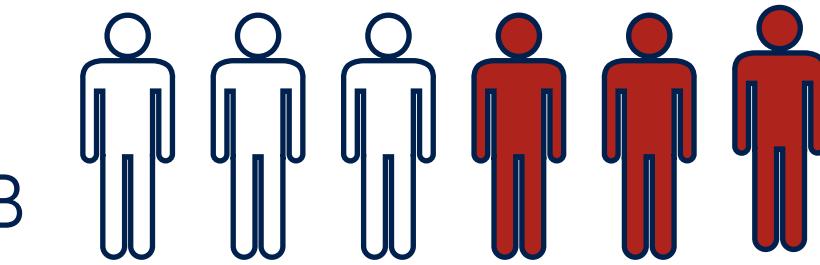
Аномалии



Идентификация  
аномалий

500 000 000

Частных  
клиентов



500 000

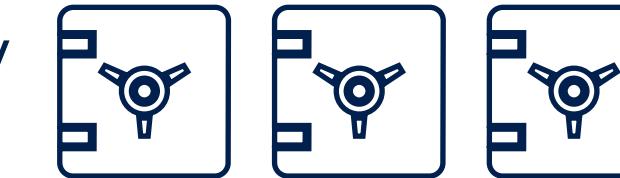
Корпоративных  
клиентов



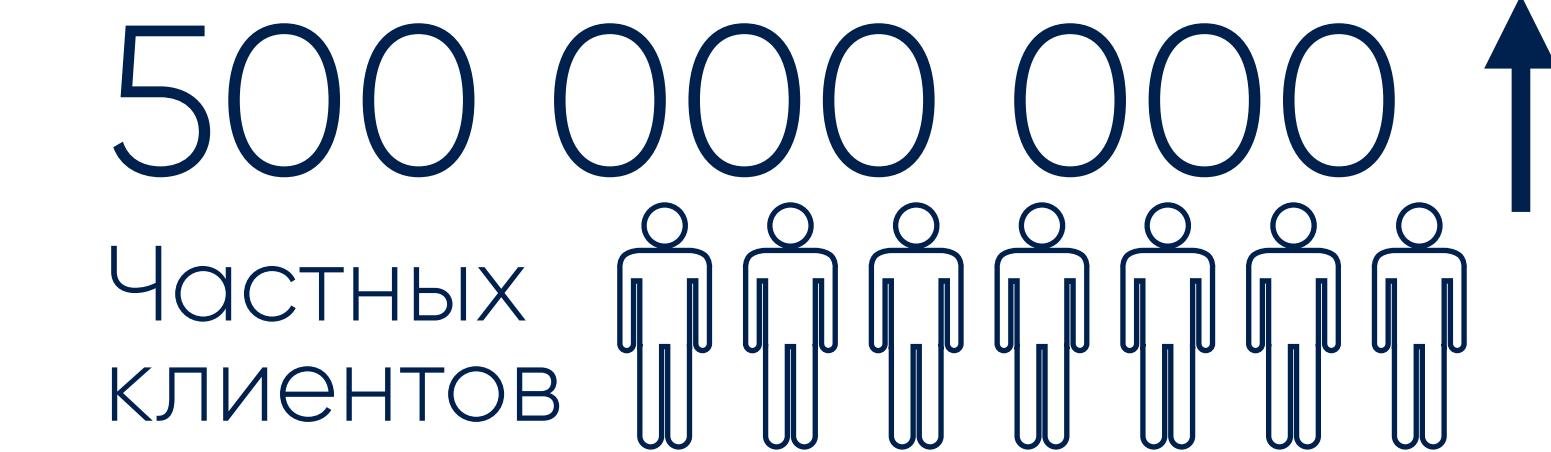
# Актуальность работы

10-ки

Хранилищ данных  
по всему миру

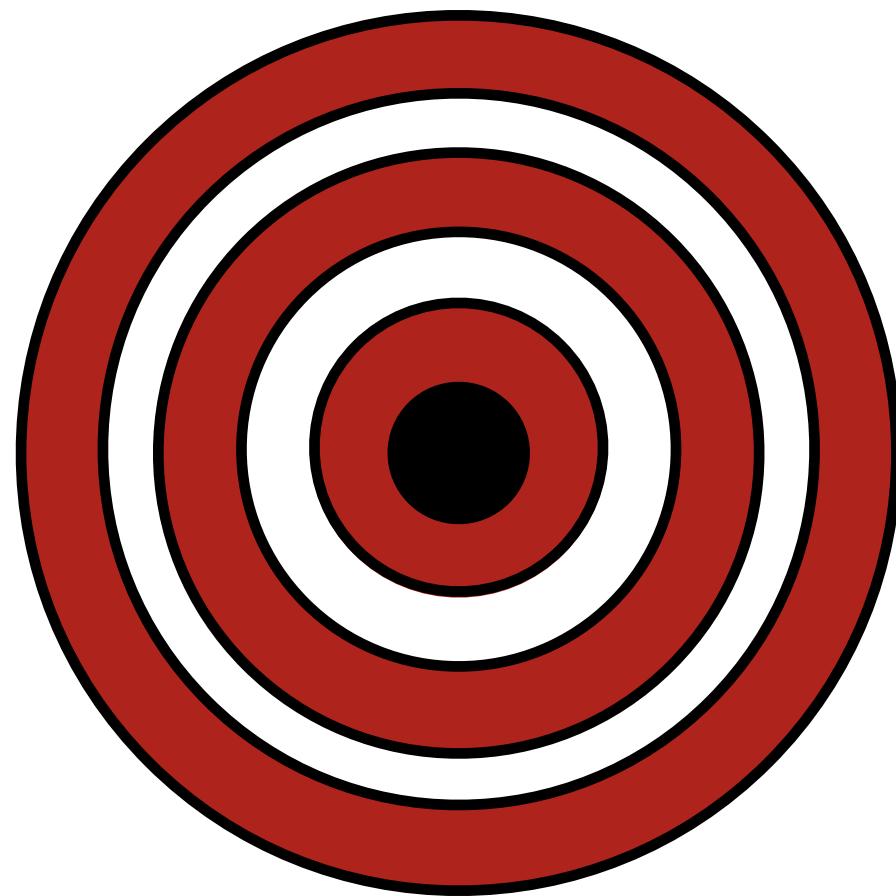


Идентификация  
аномалий

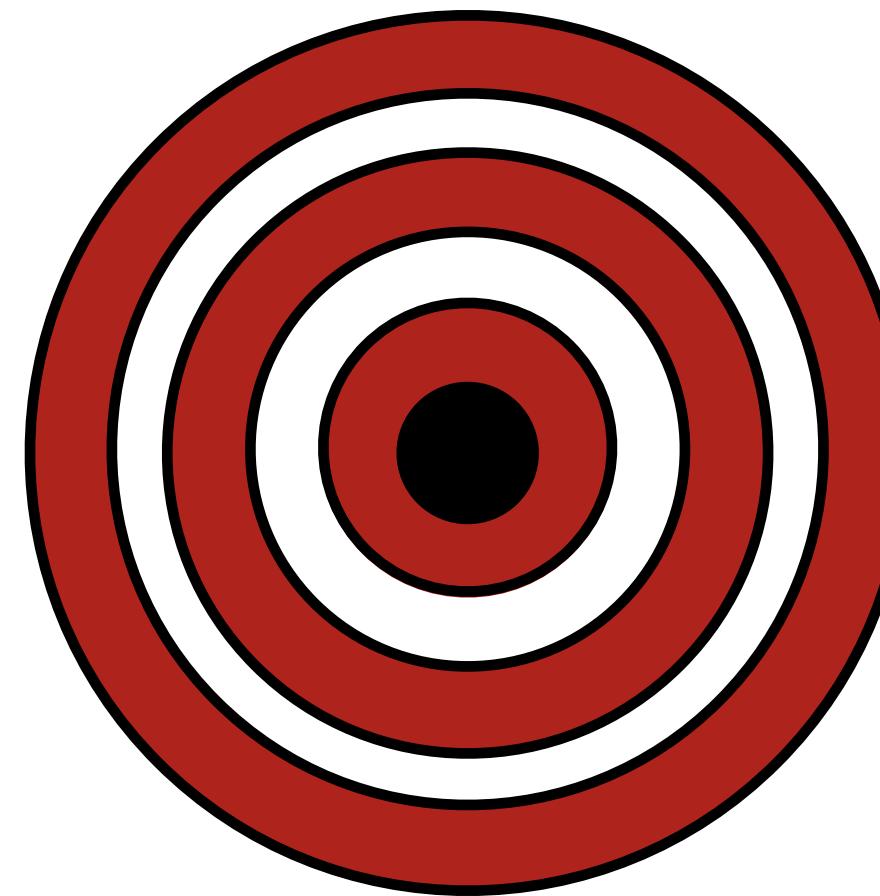


# Задача

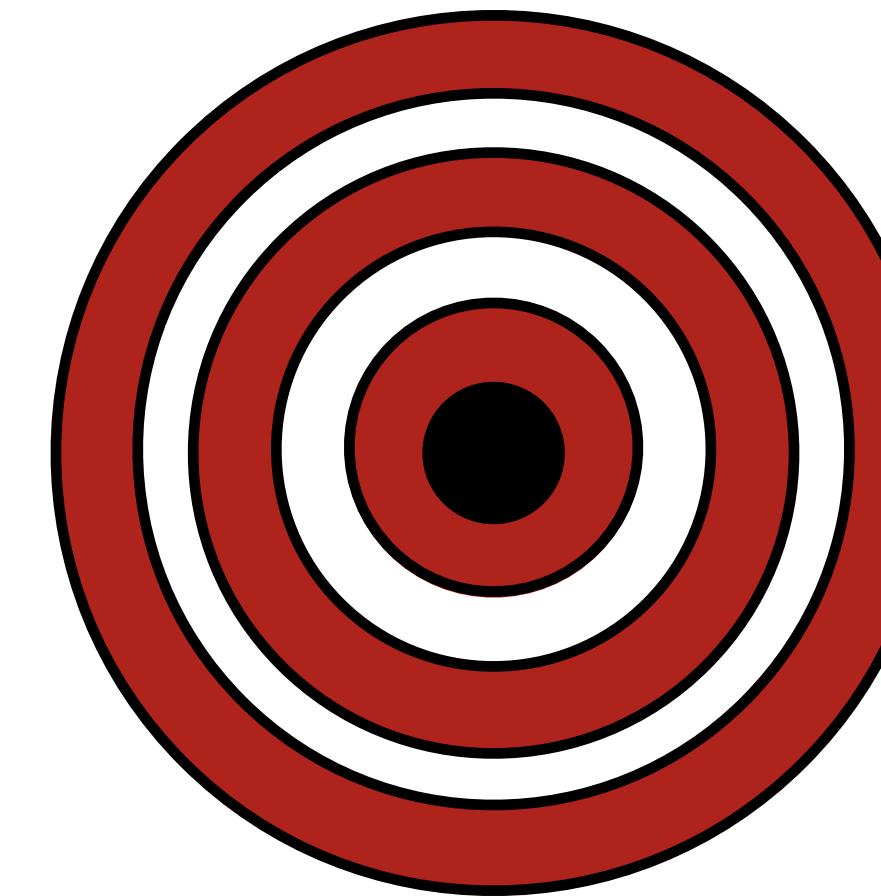
Создать систему поиска аномалий со следующими  
характеристиками



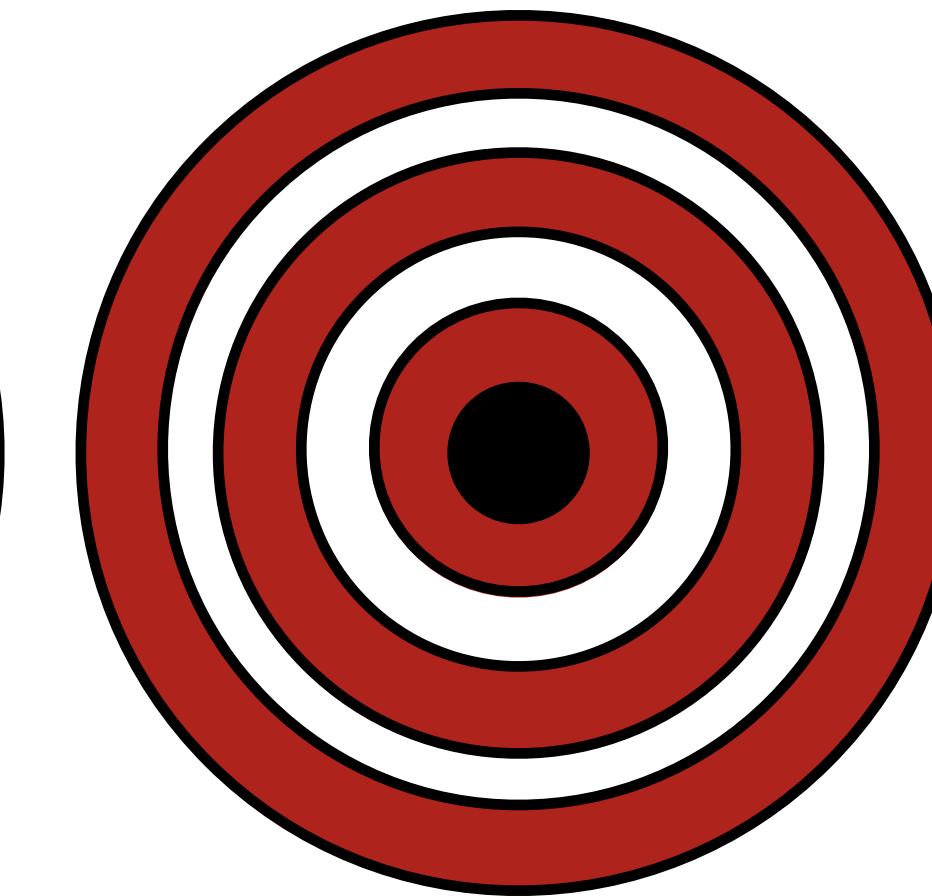
Полное отсутствие  
ручной работы



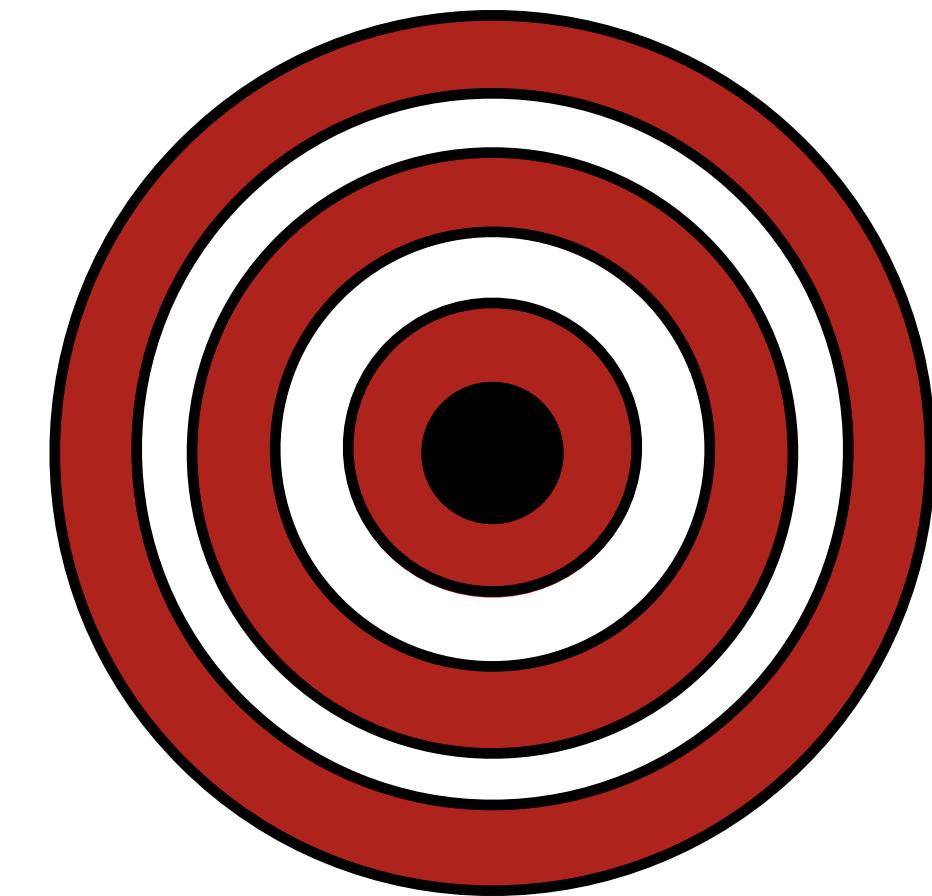
Универсальность  
для всего  
многообразия  
хранилищ данных  
компании Acronis



Легко настраивается  
под требования  
заинтересованных  
сторон

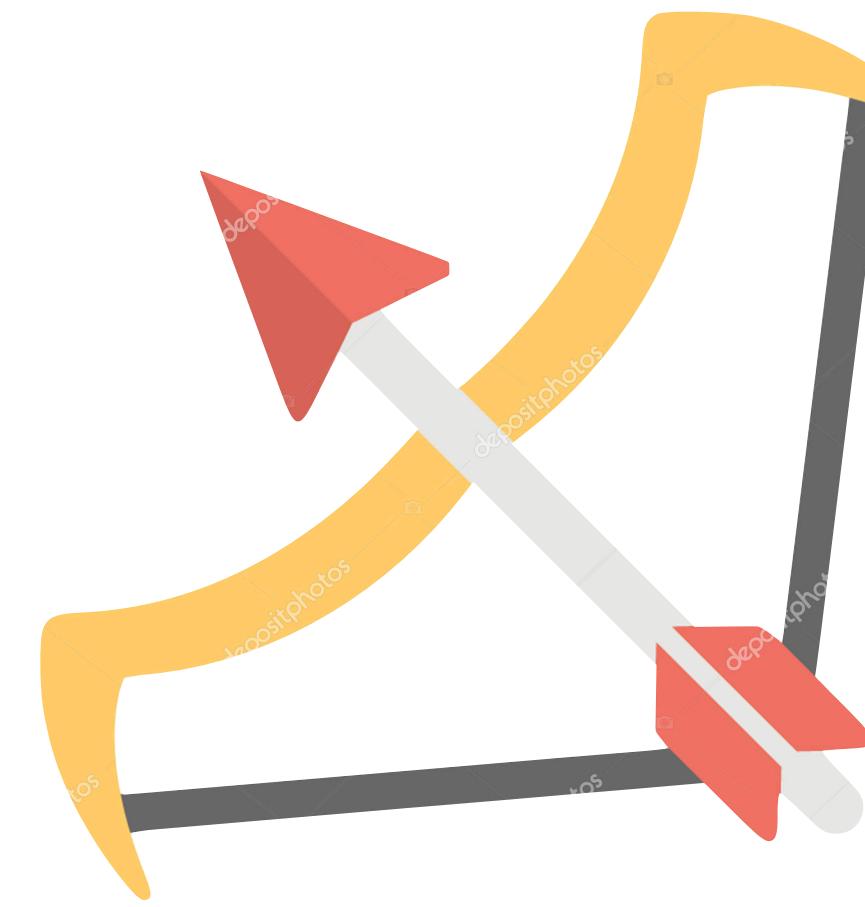


Однаково  
корректная работа  
при любой нагрузке  
системы

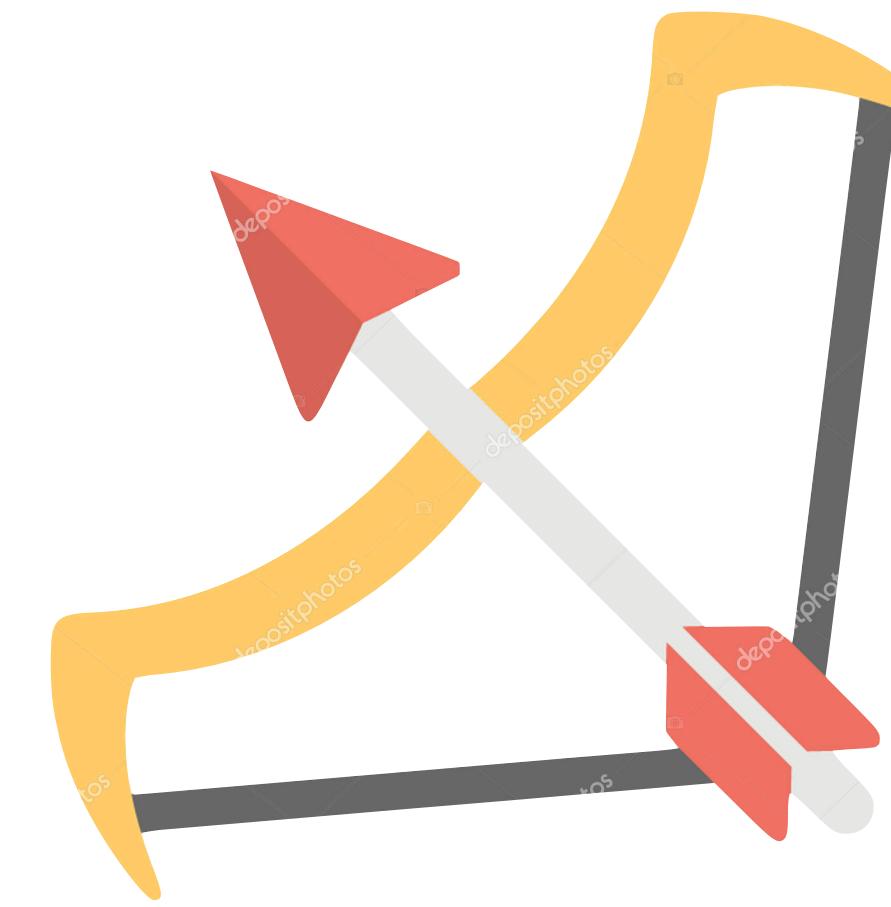


Высокая точность  
и низкий уровень  
ложных  
срабатываний

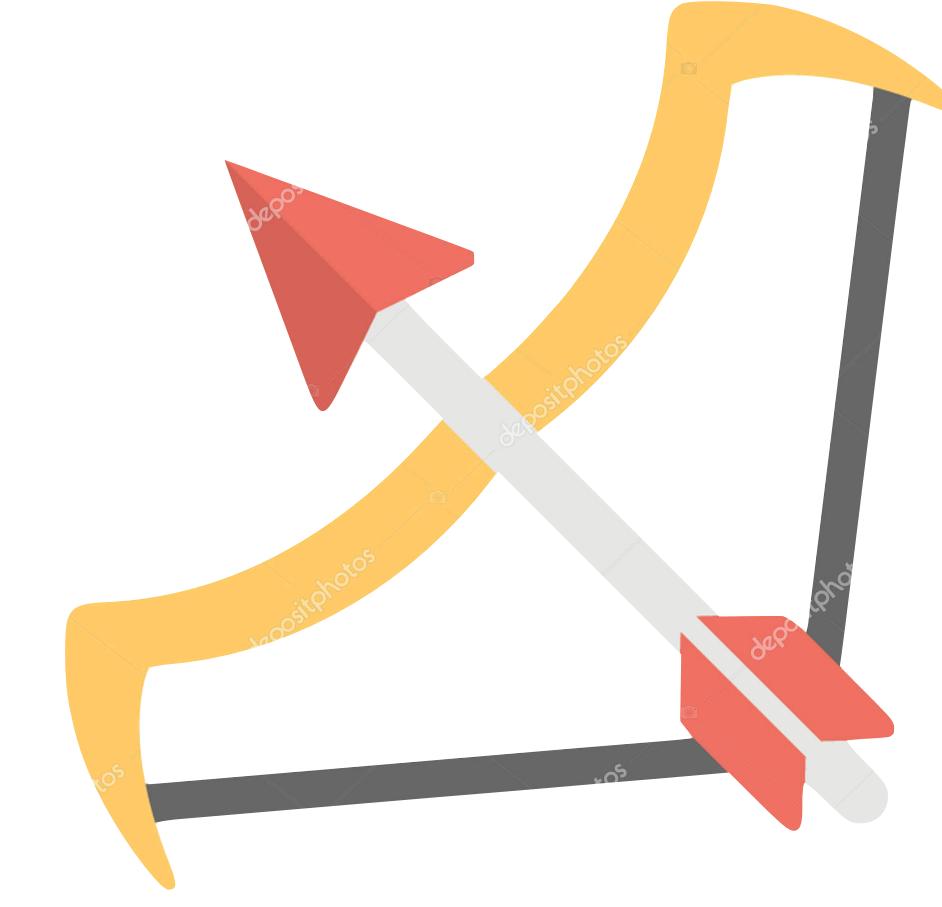
# Инструментарий



Python 3.7



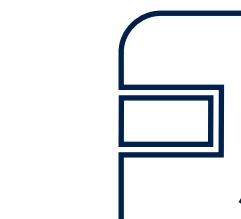
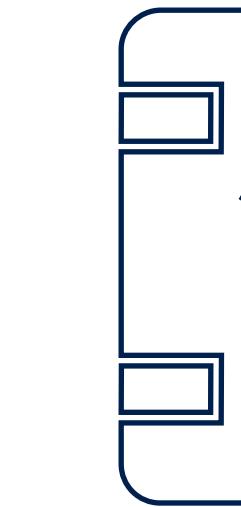
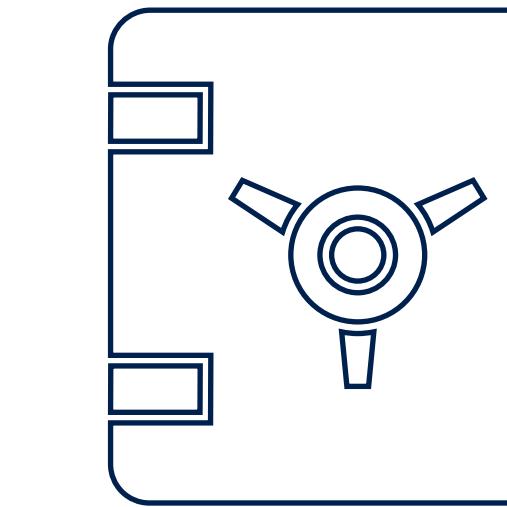
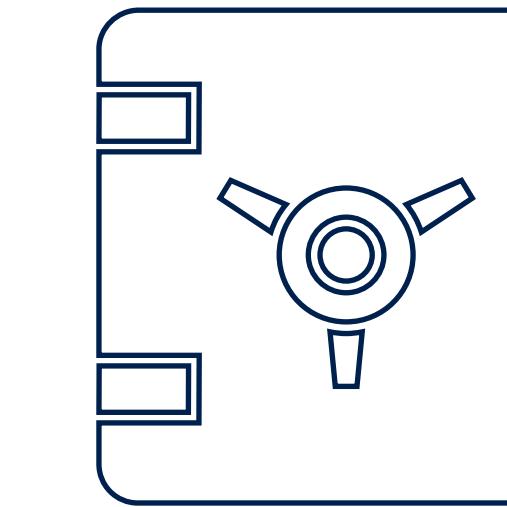
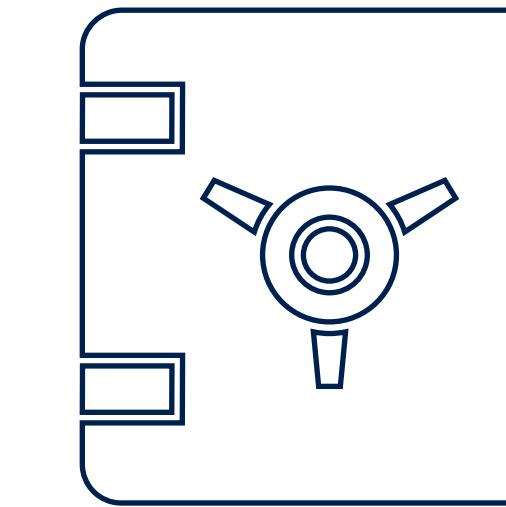
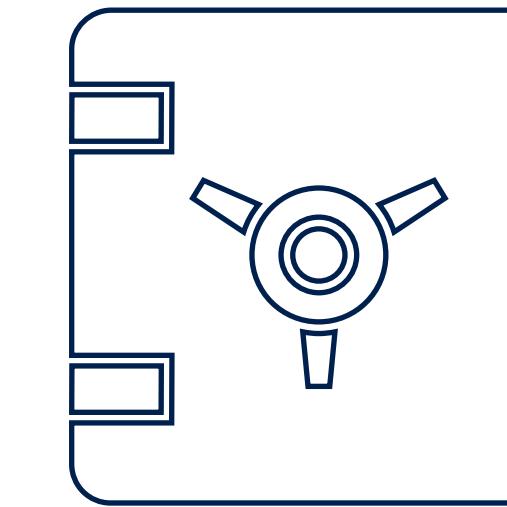
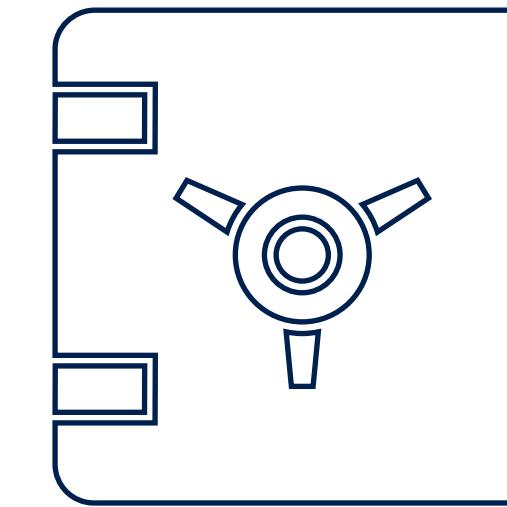
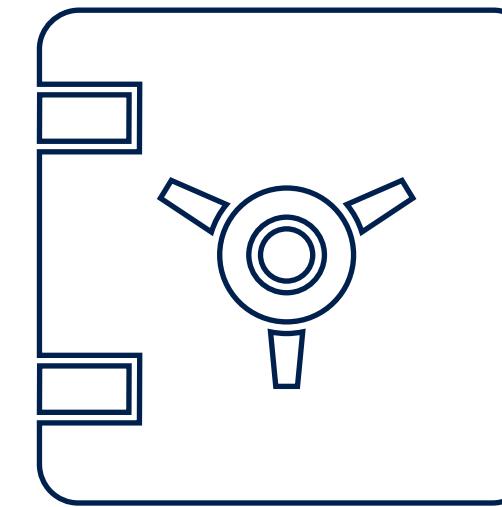
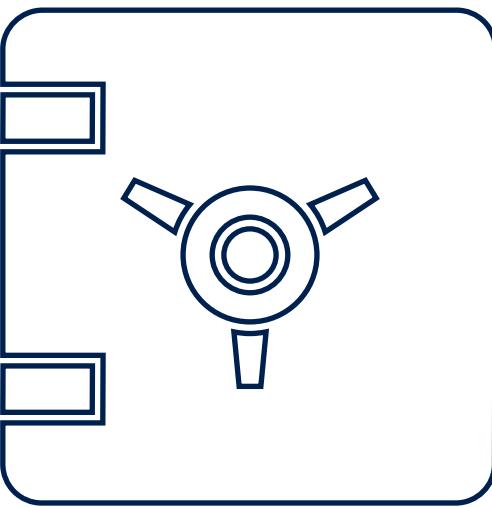
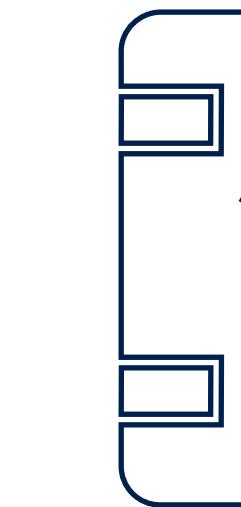
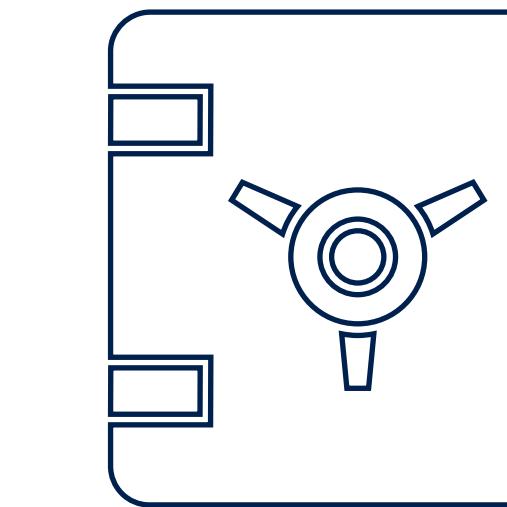
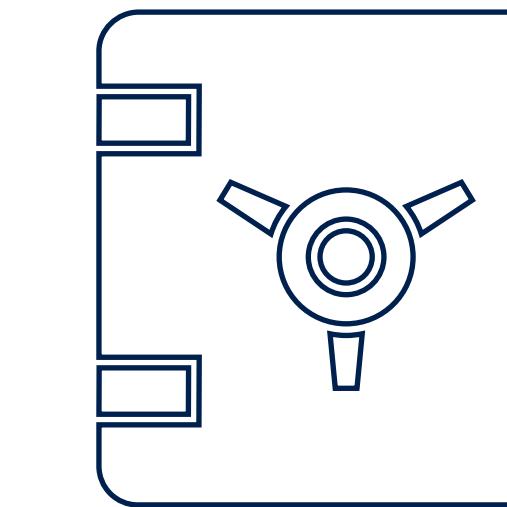
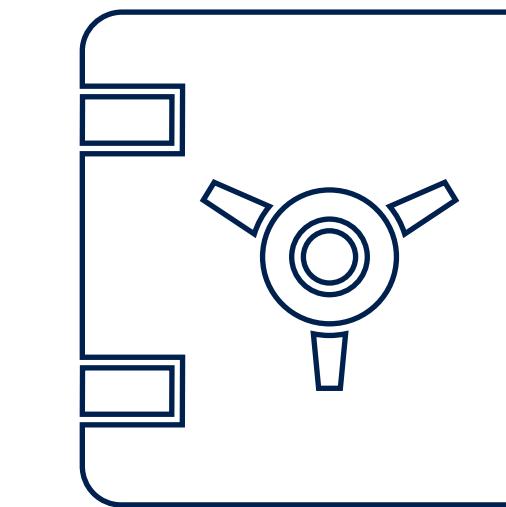
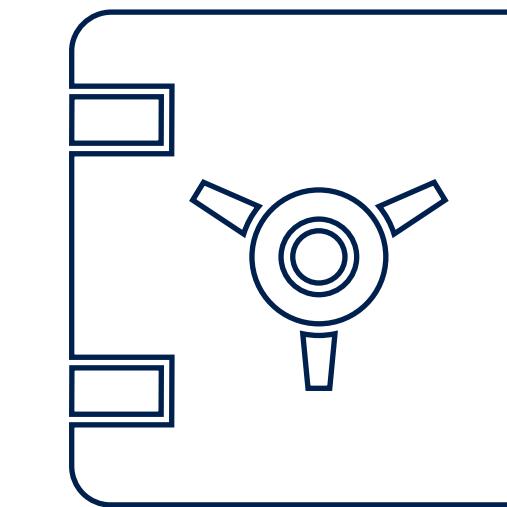
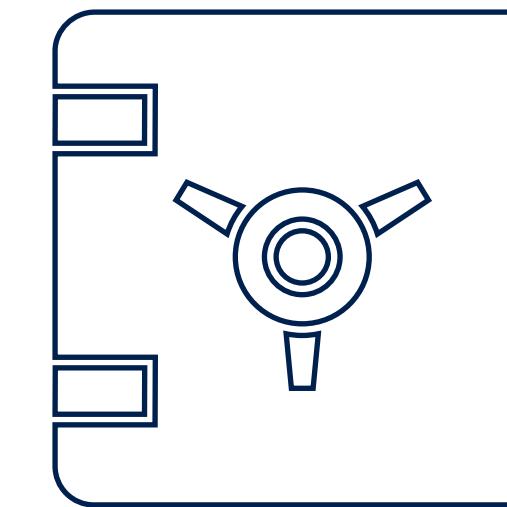
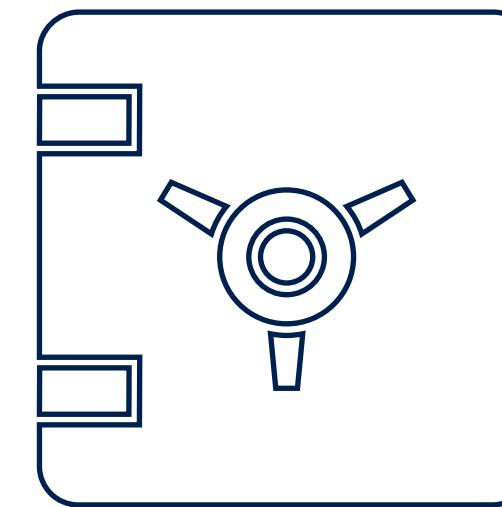
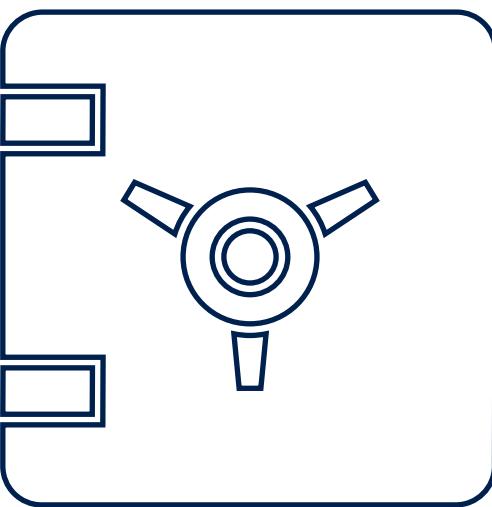
Prometheus



Grafana

# Хранилища данных

# Acronis Storage

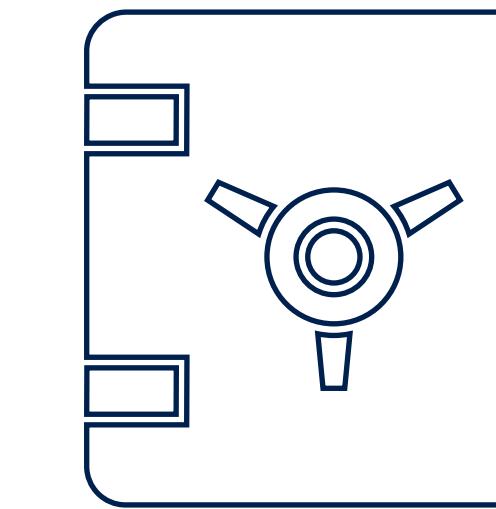
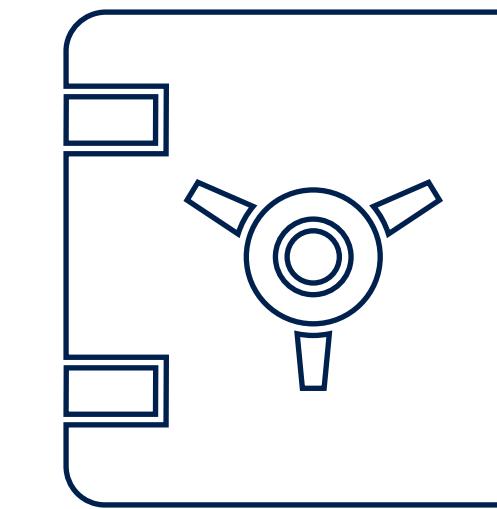
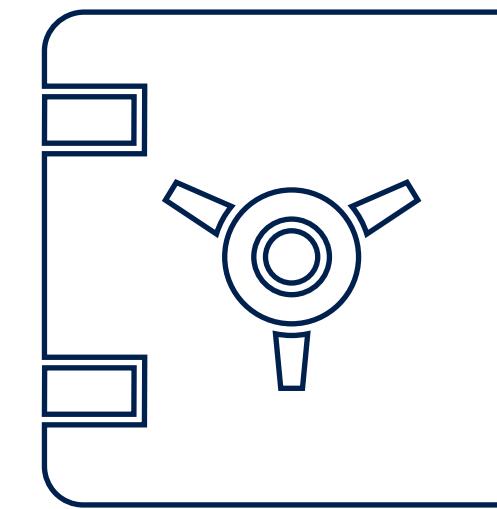


# Хранилища данных

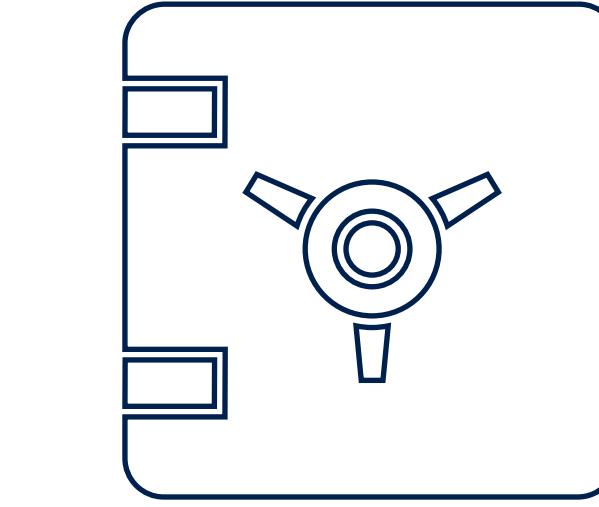
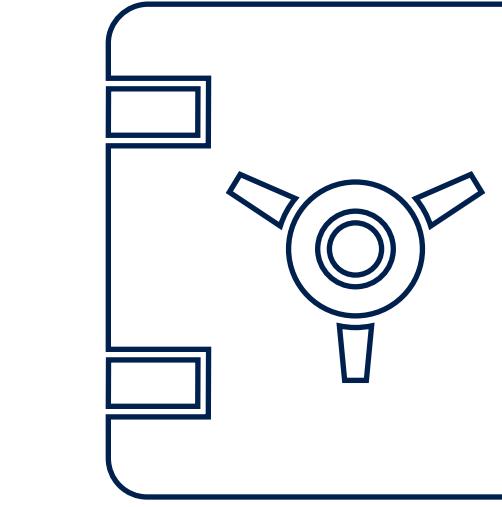
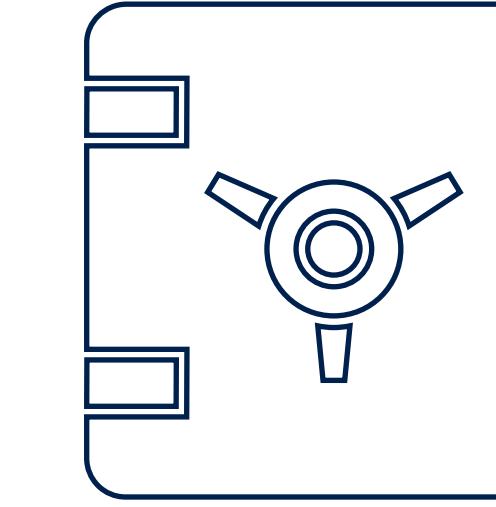
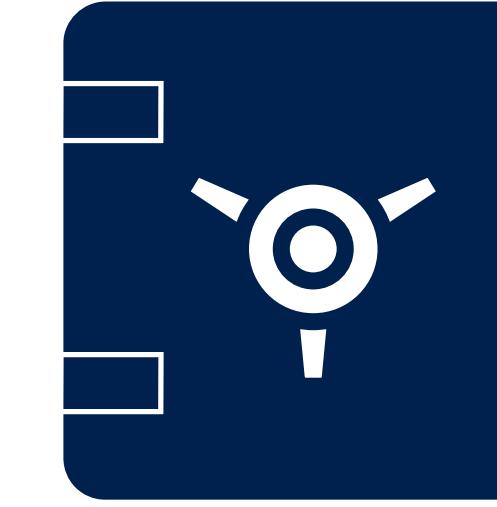
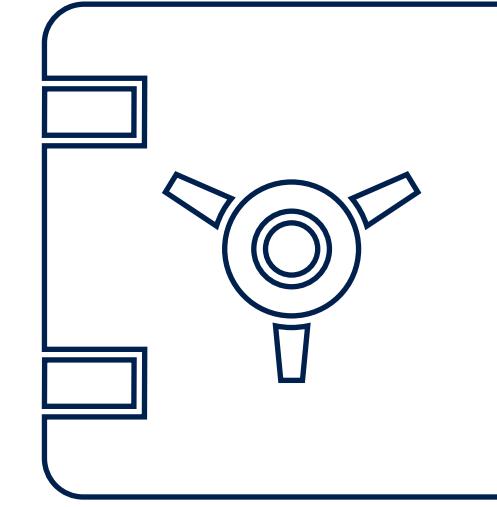
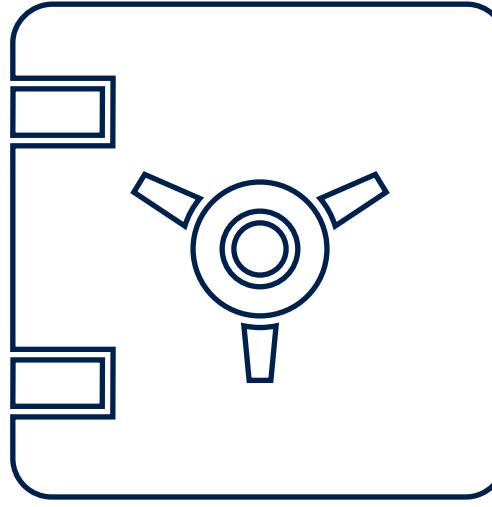
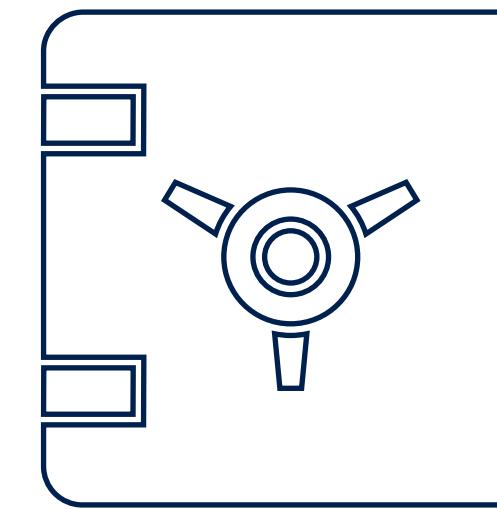
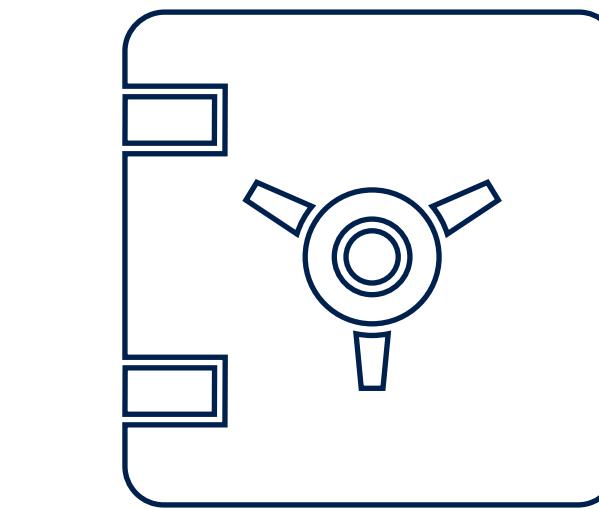
## Acronis Storage



us3



au2-acs1

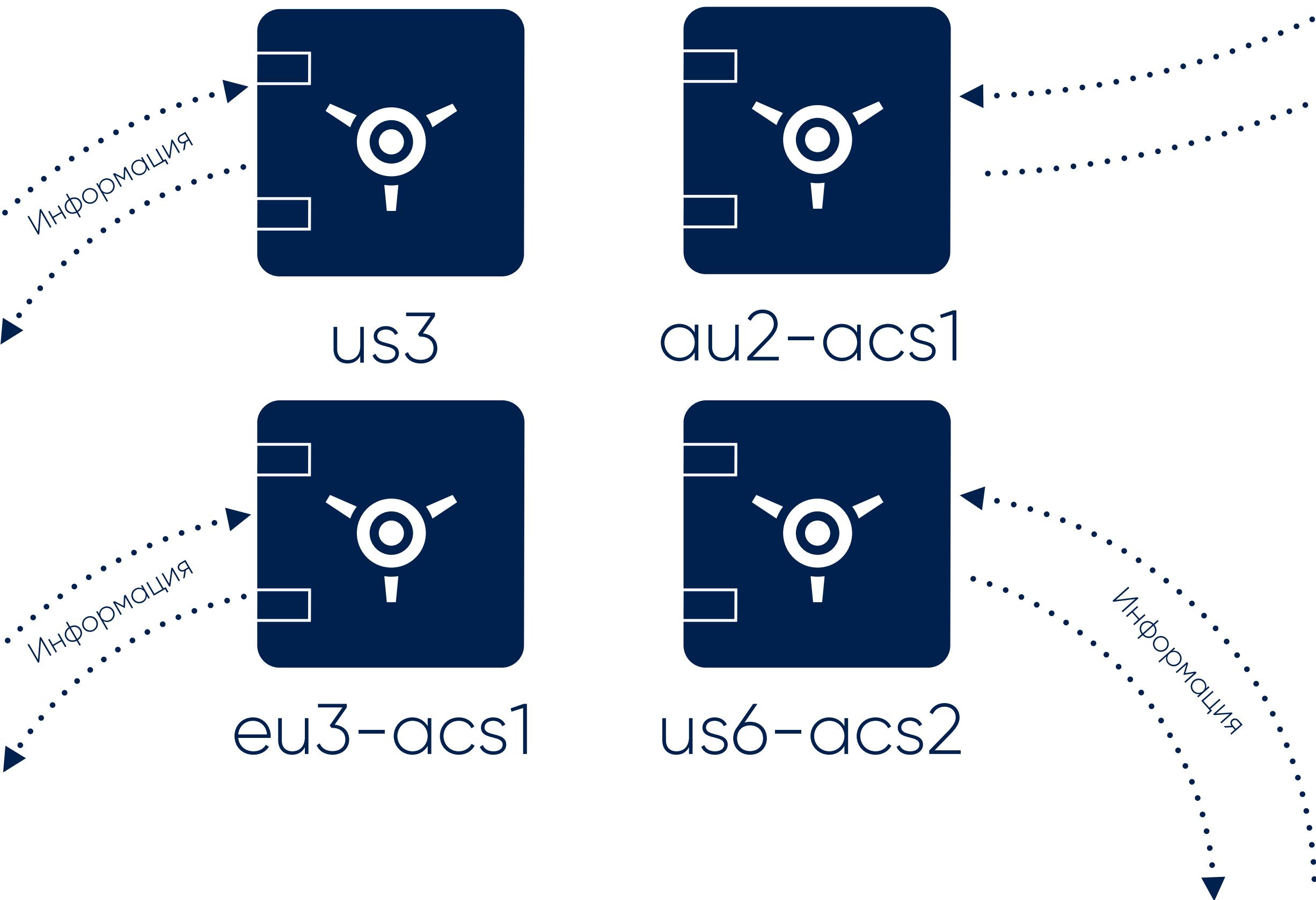


eu3-acs1

us6-acs2

# Хранилища данных

## Acronis Storage

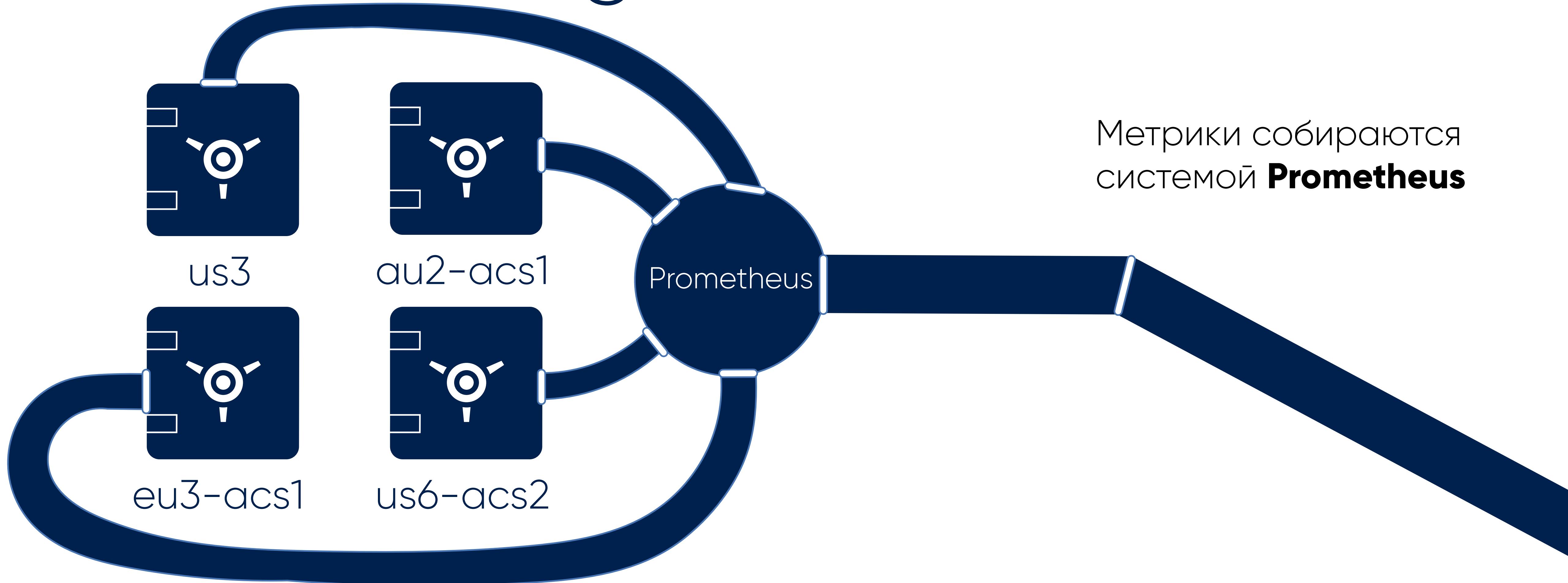


**Информация о функционировании**  
хранилищ данных компаний Acronis  
хранится в виде временных рядов –  
метрик. Их мы и будем рассматривать  
для поиска аномалий.

**abwg\_req\_latency\_ms\_sum** – целевая  
метрика. Её смысл – задержка  
запросов нарастающим итогом. По  
ней будем детектировать аномалии.

# Хранилища данных

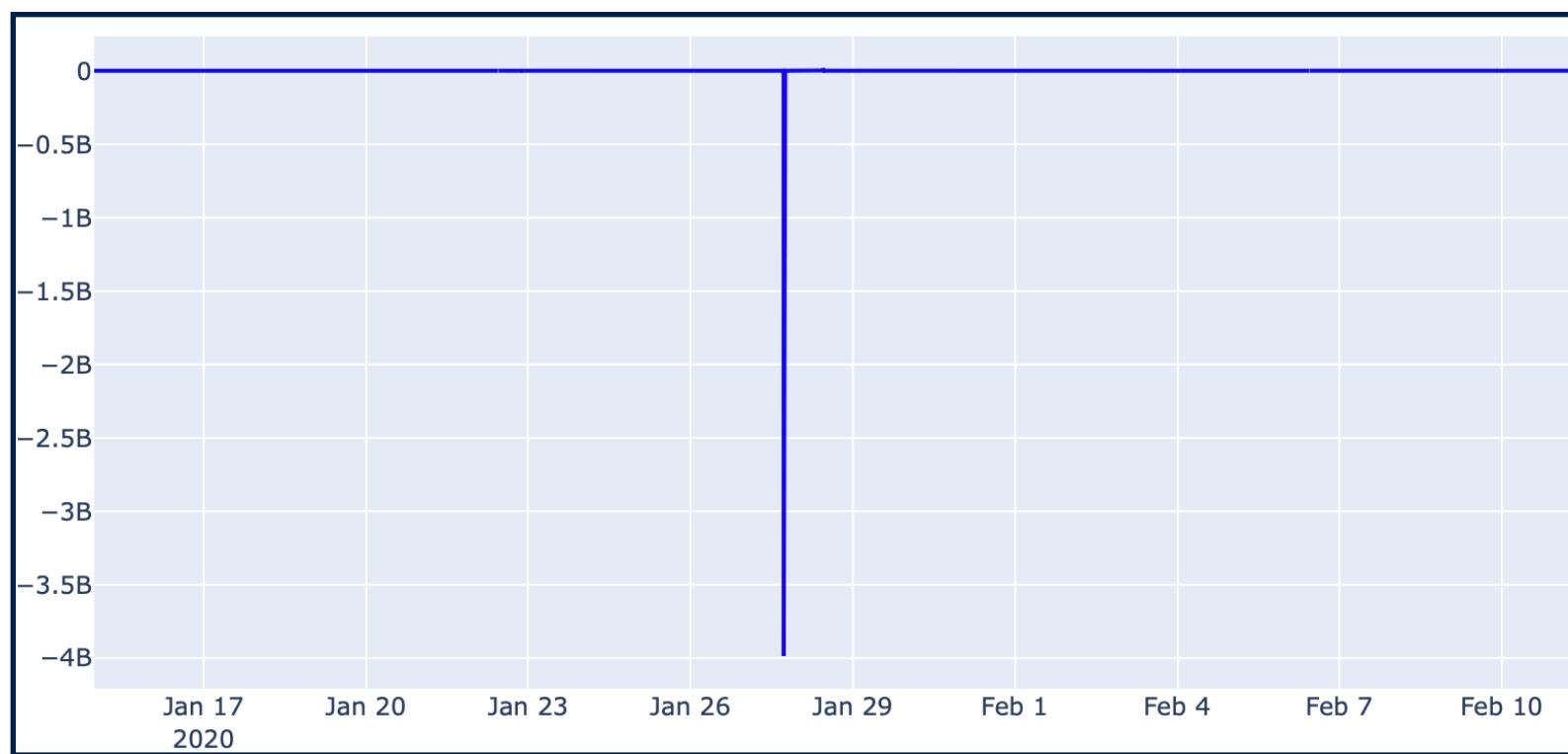
## Acronis Storage



# Схема

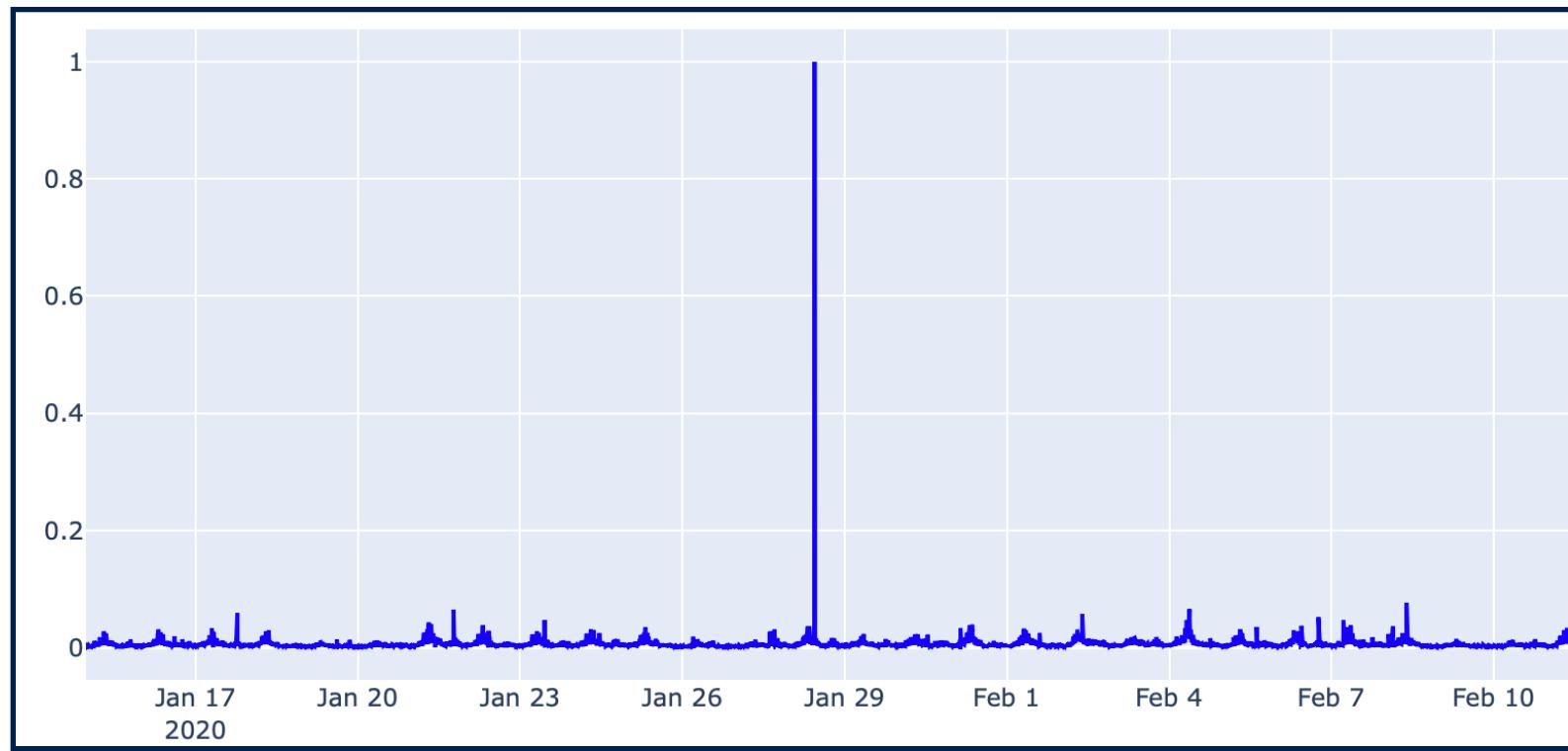


# Предобработка данных

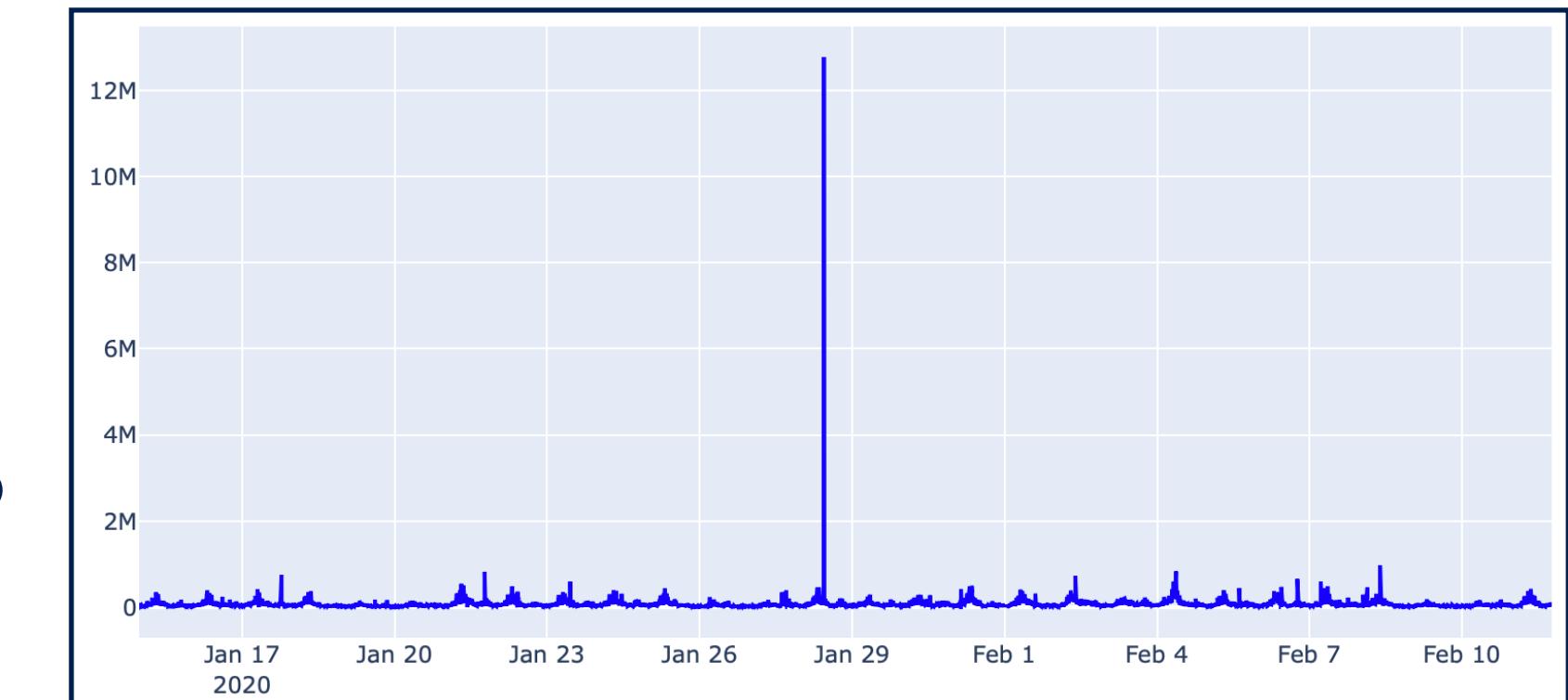
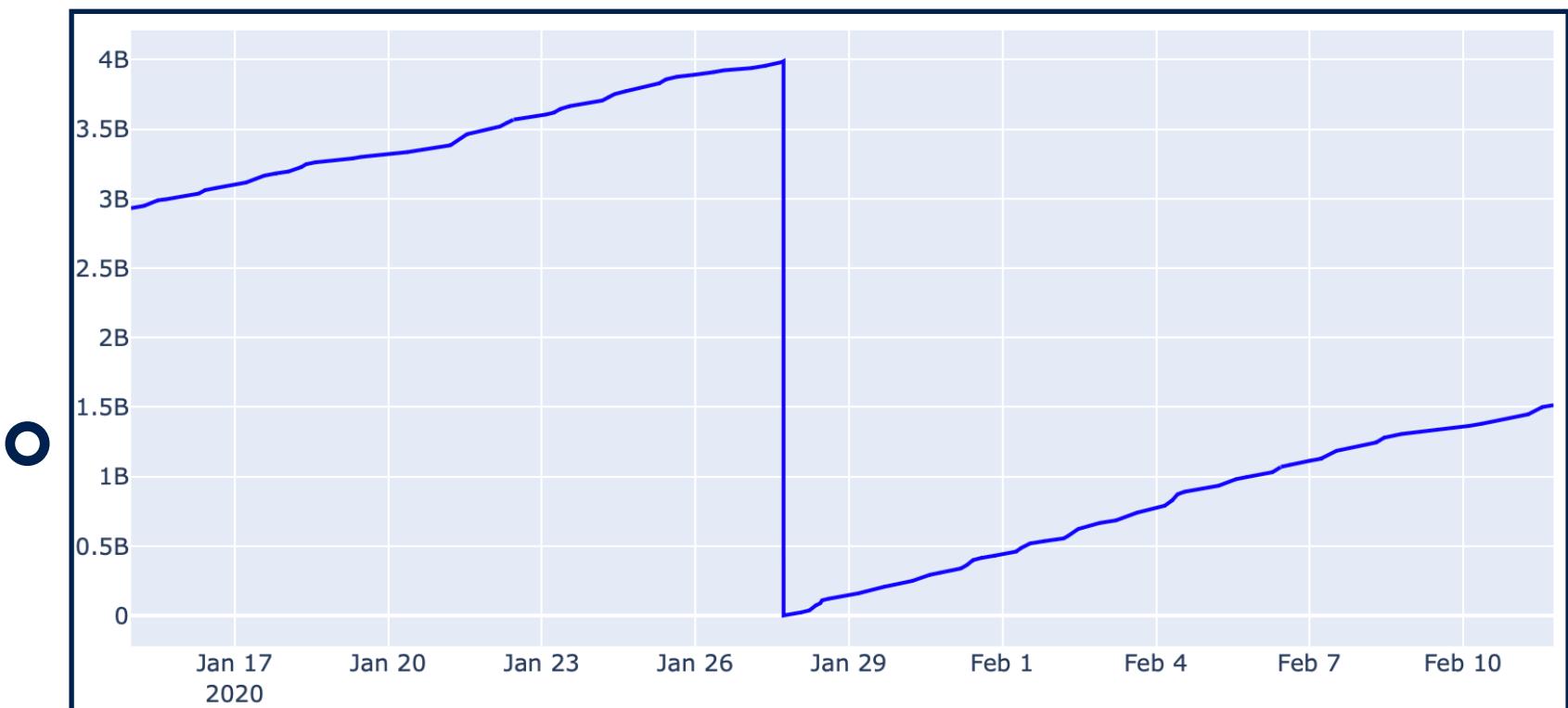


Убираем  
тренд

Удаление  
участков с  
перезагрузками



Масштабирование



# Предиктивная модель



Усовершенствование ARIMA( $p, d = 1, q$ )

$$\hat{y}_t = \sum_{i=1}^p \alpha_i \cdot y_{t-i} + \sum_{i=1}^q \beta_i \cdot \hat{\varepsilon}_{t-i} + \sum_{k=1}^r \sum_{i=1}^p \gamma_i x_{t-i}^k$$

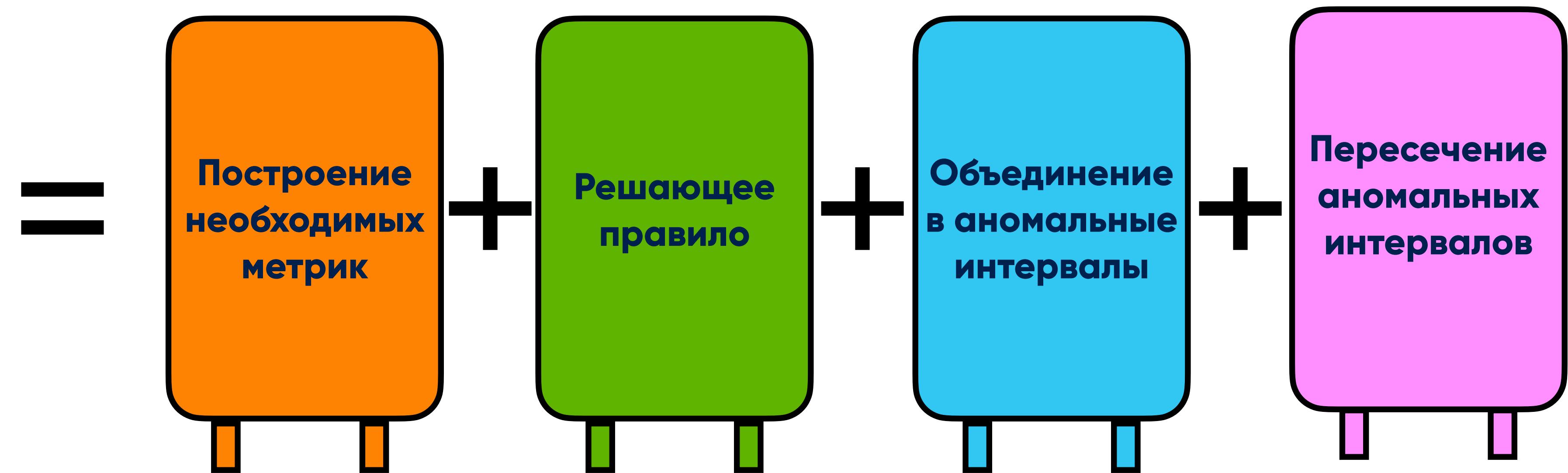
значение целевой переменной в момент времени  $(t - i)$

значение ошибки модели в момент времени  $(t - i)$

значение  $k$ -го признака в момент времени  $(t - i)$

Подбирали параметры  $\{\alpha_i\}$ ,  $\{\beta_i\}$  и  $\{\gamma_i\}$  на обучении, а параметры  $p = 2$ ,  $q = 2$  и  $r = 5$  на валидации.

# Составляющие алгоритма поиска аномалий

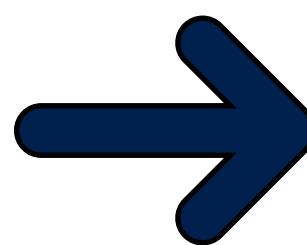
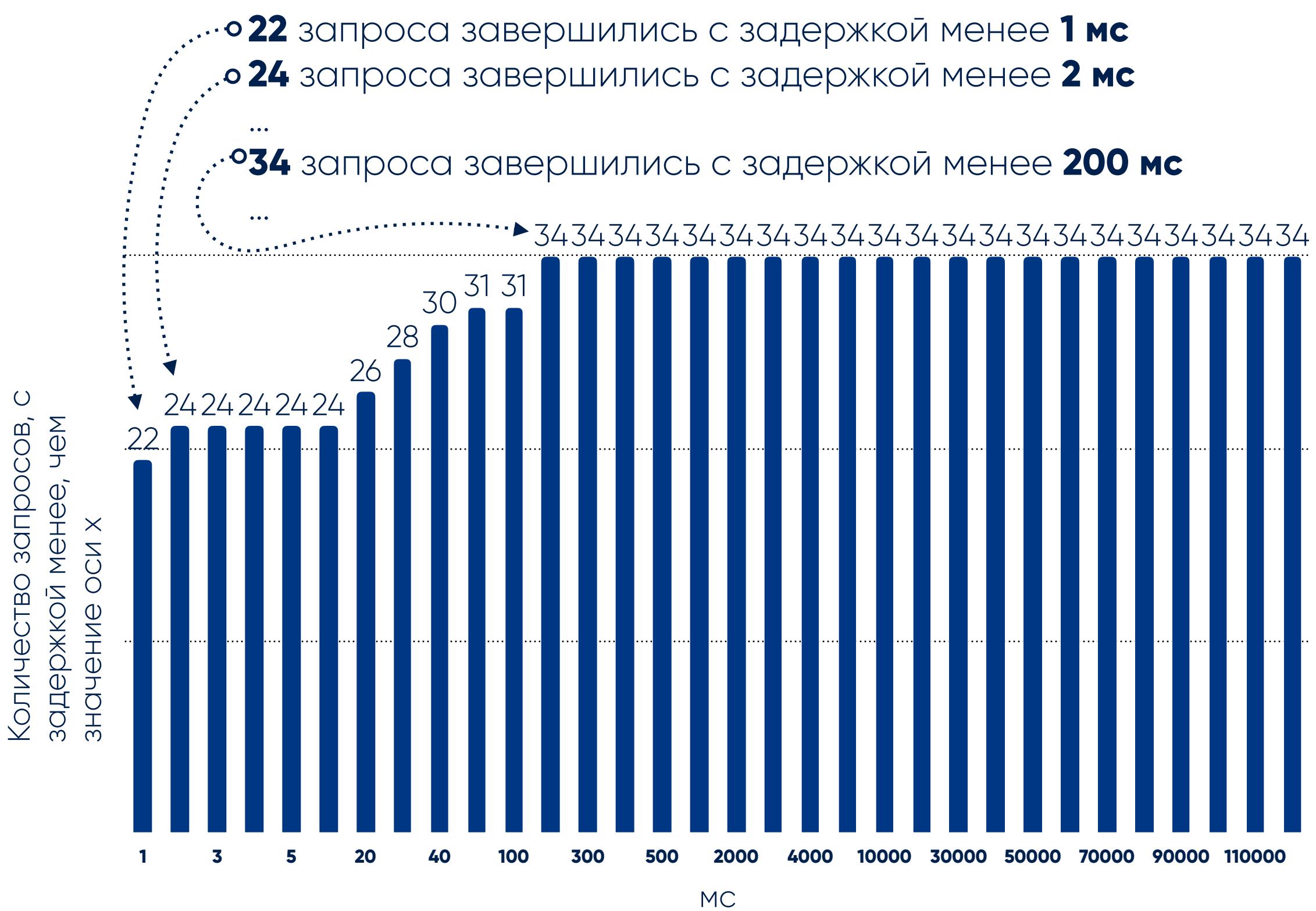


`abgw_iop_latency_ms_bucket{err="OK", iop="pread", proxied="0"}`

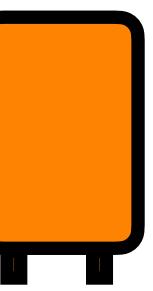
DC = au2-ac51

INSTANCE = 1

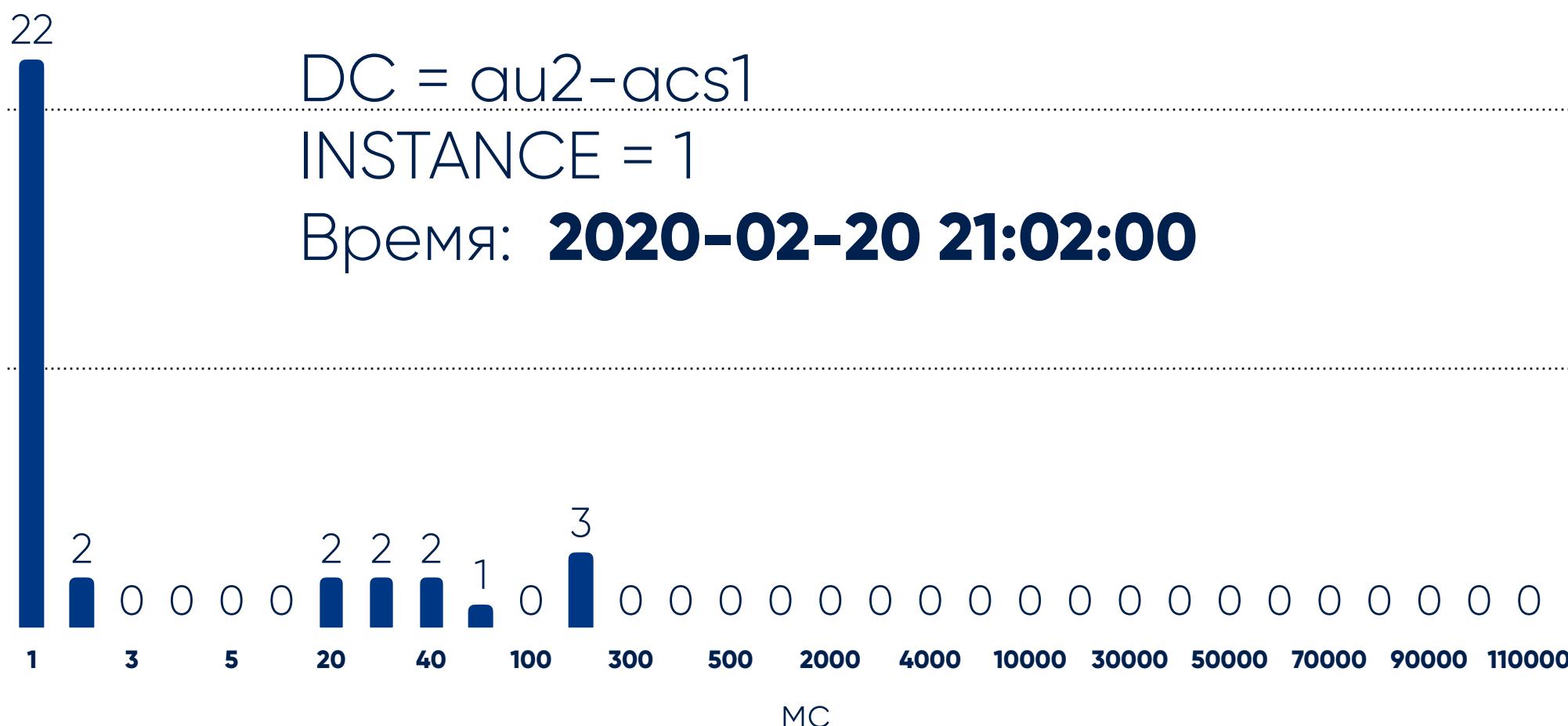
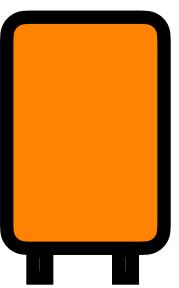
Время: **2020-02-20 21:02:00**



Количество запросов, с задержкой между соседними значениями оси X



# Построение метрик



mean = 20.1  
variance = 3166.7

$$mean_k = \frac{1}{\mathbb{I}([0, \infty])} \cdot \sum_{i,j>i} \mathbb{I}([i, j]) \cdot \frac{j - i}{2}$$

$$variance_k = \frac{1}{\mathbb{I}([0, \infty])} \cdot \sum_{i,j>i} \mathbb{I}([i, j]) \cdot \left( \frac{j - i}{2} - mean_k \right)^2$$

# Построение метрик

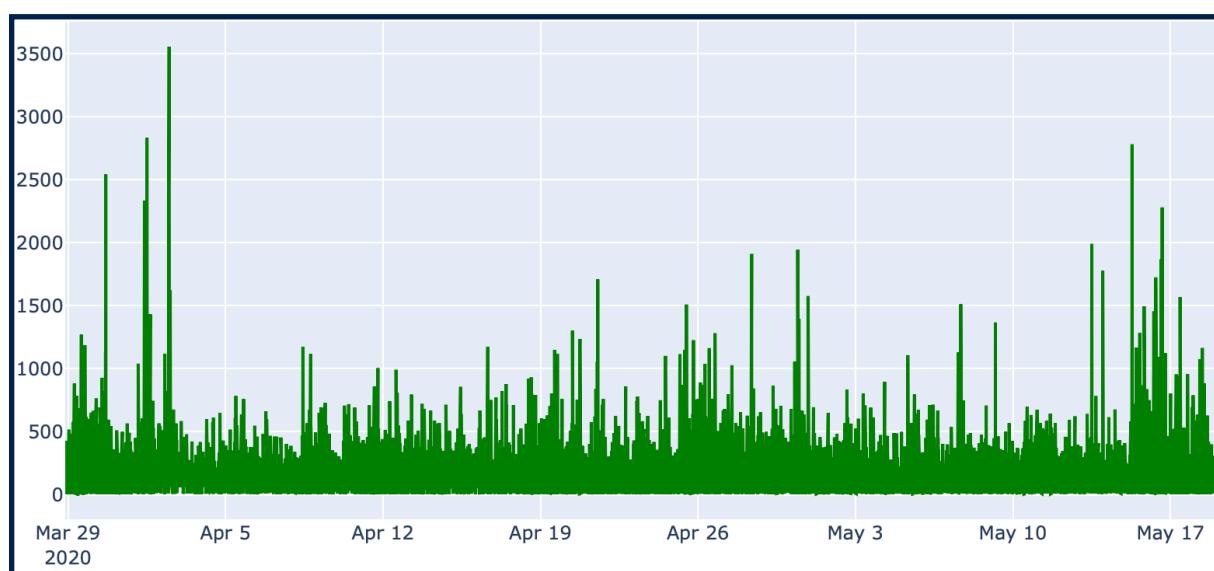


DC = au2-acs1

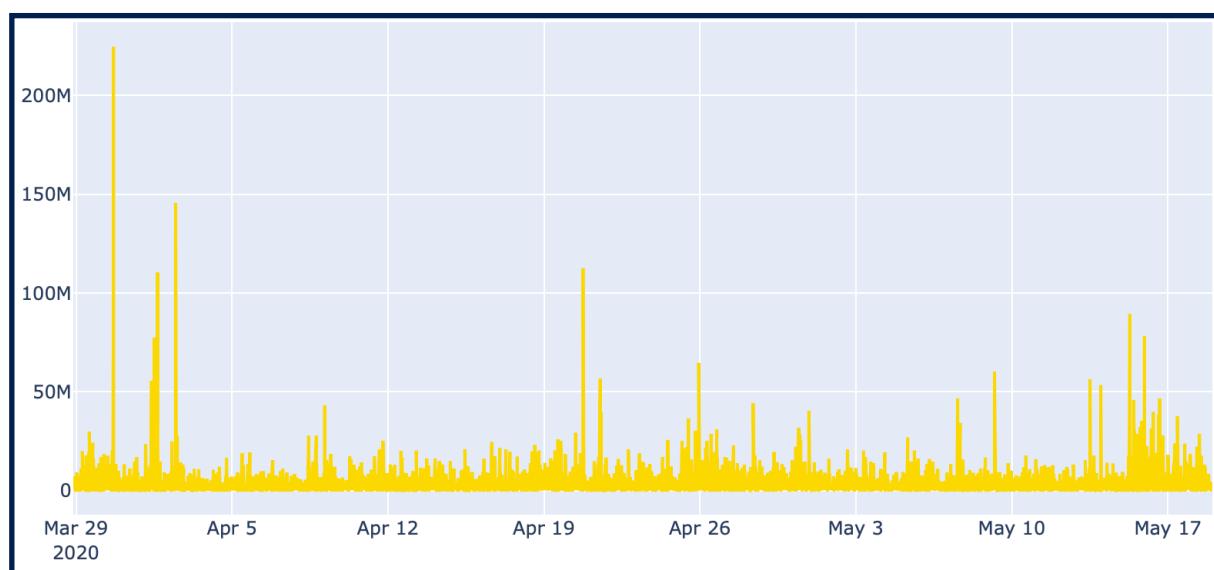
INSTANCE = 1

Время: **2020-02-20 21:02:00 - 2020-05-18 23:44:00**

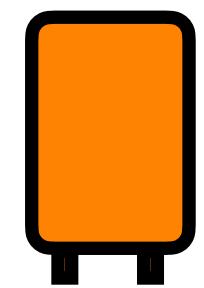
Среднее  
значение  
задержки



Дисперсия  
задержки

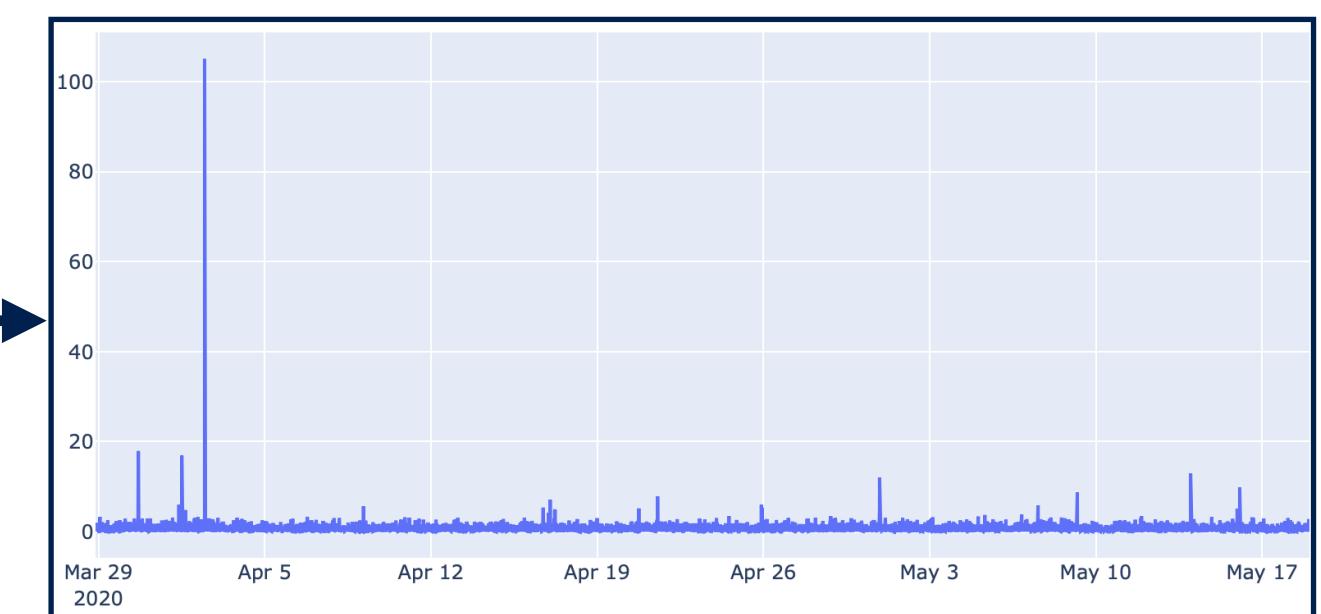


# Построение метрик

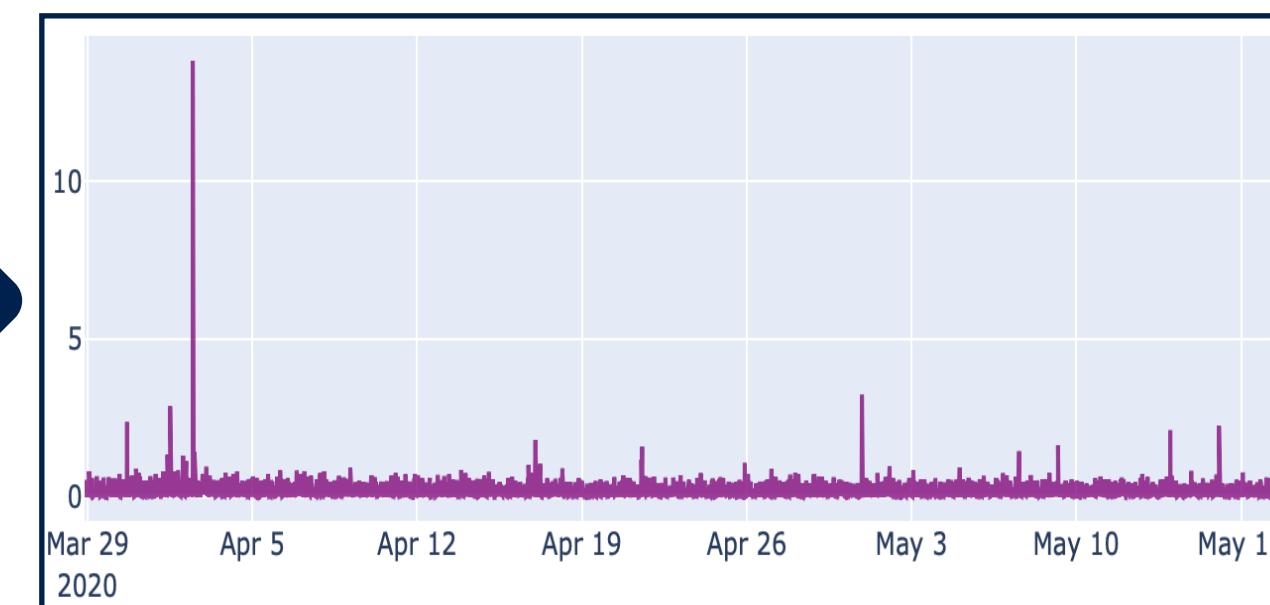


Ошибка модели

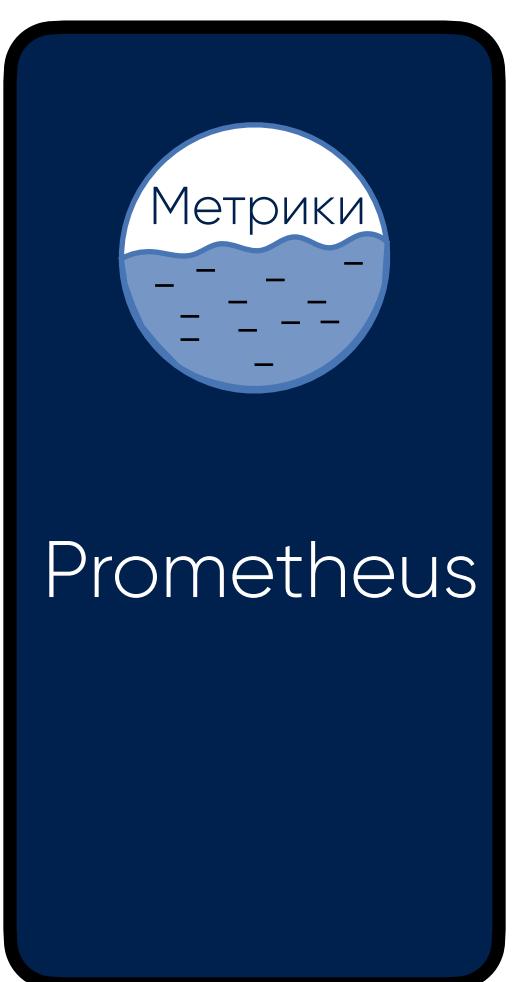
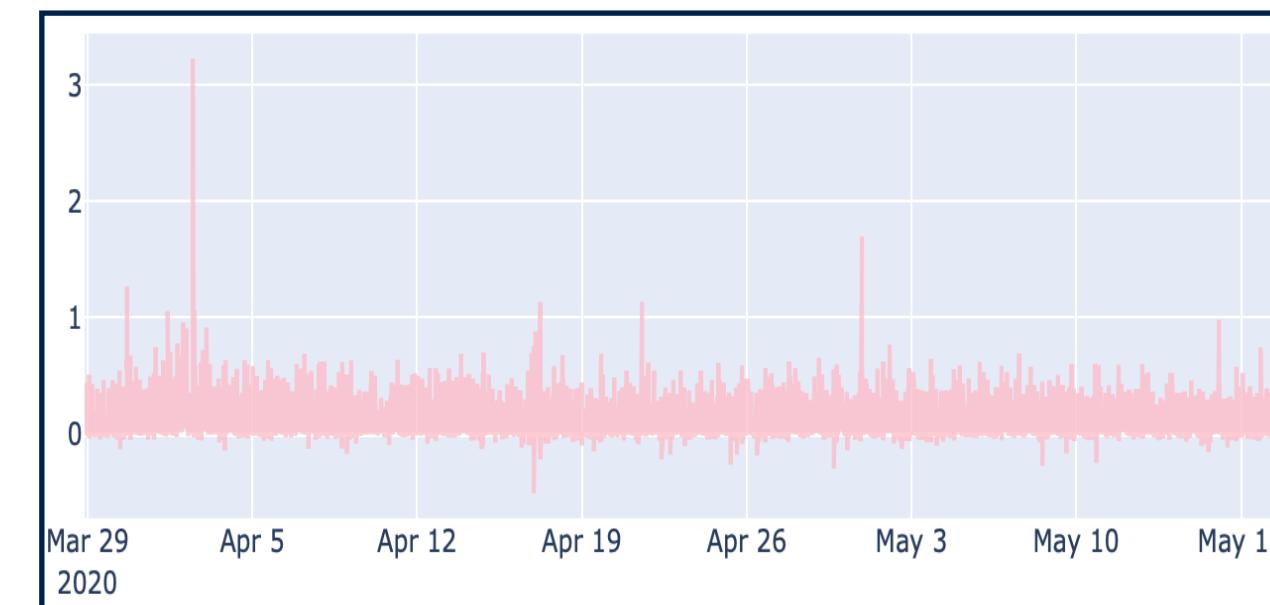
Вычитаем и нормируем



Целевая переменная

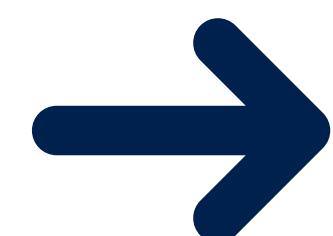


Предсказание целевой переменной

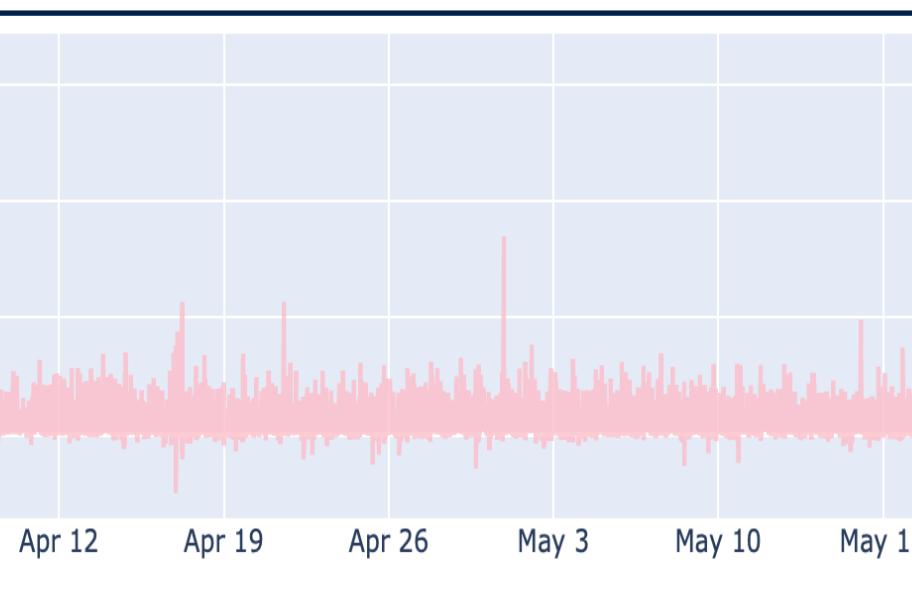


Prometheus

Предиктивная модель



Предсказание целевой переменной



П

о

с

т

р

о

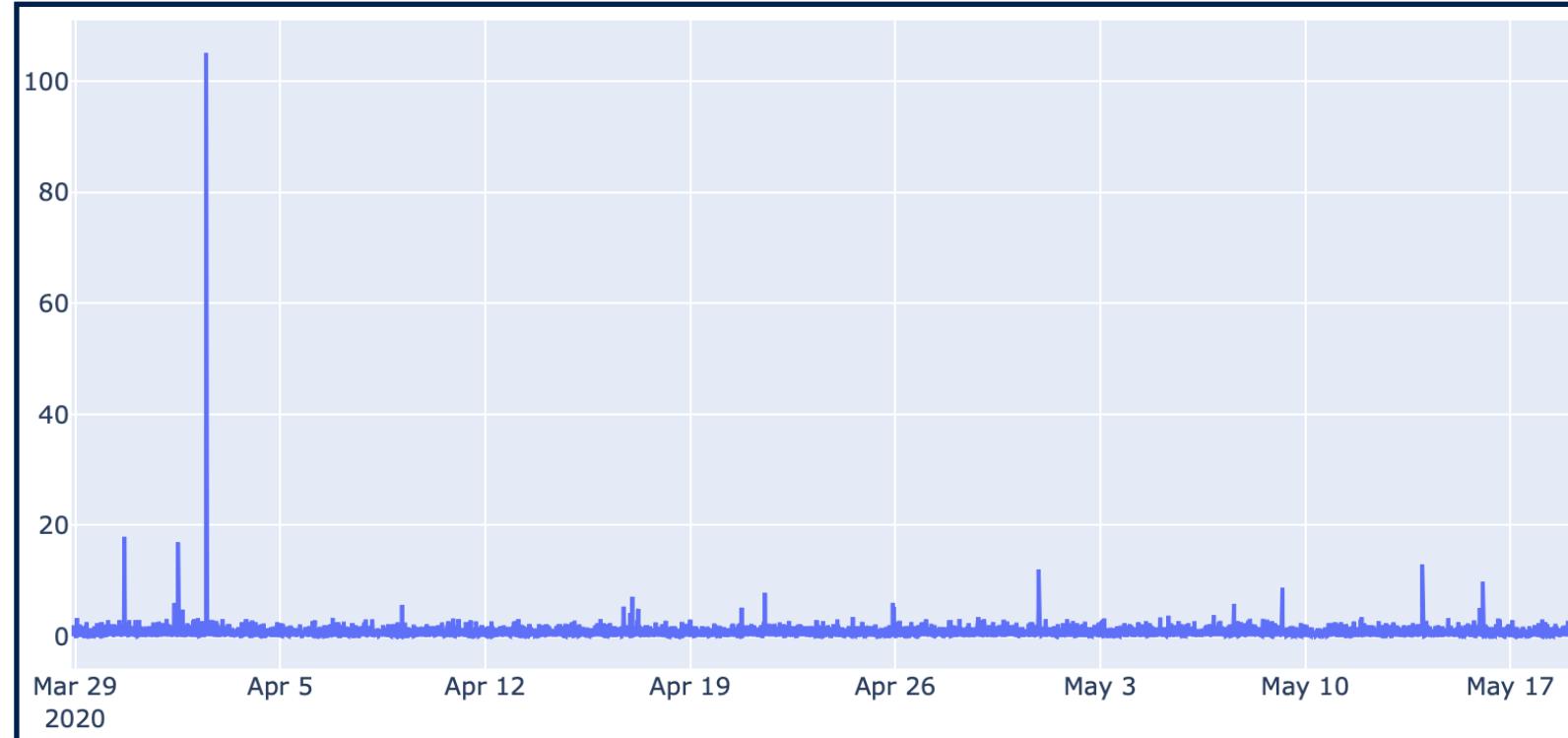
е

н

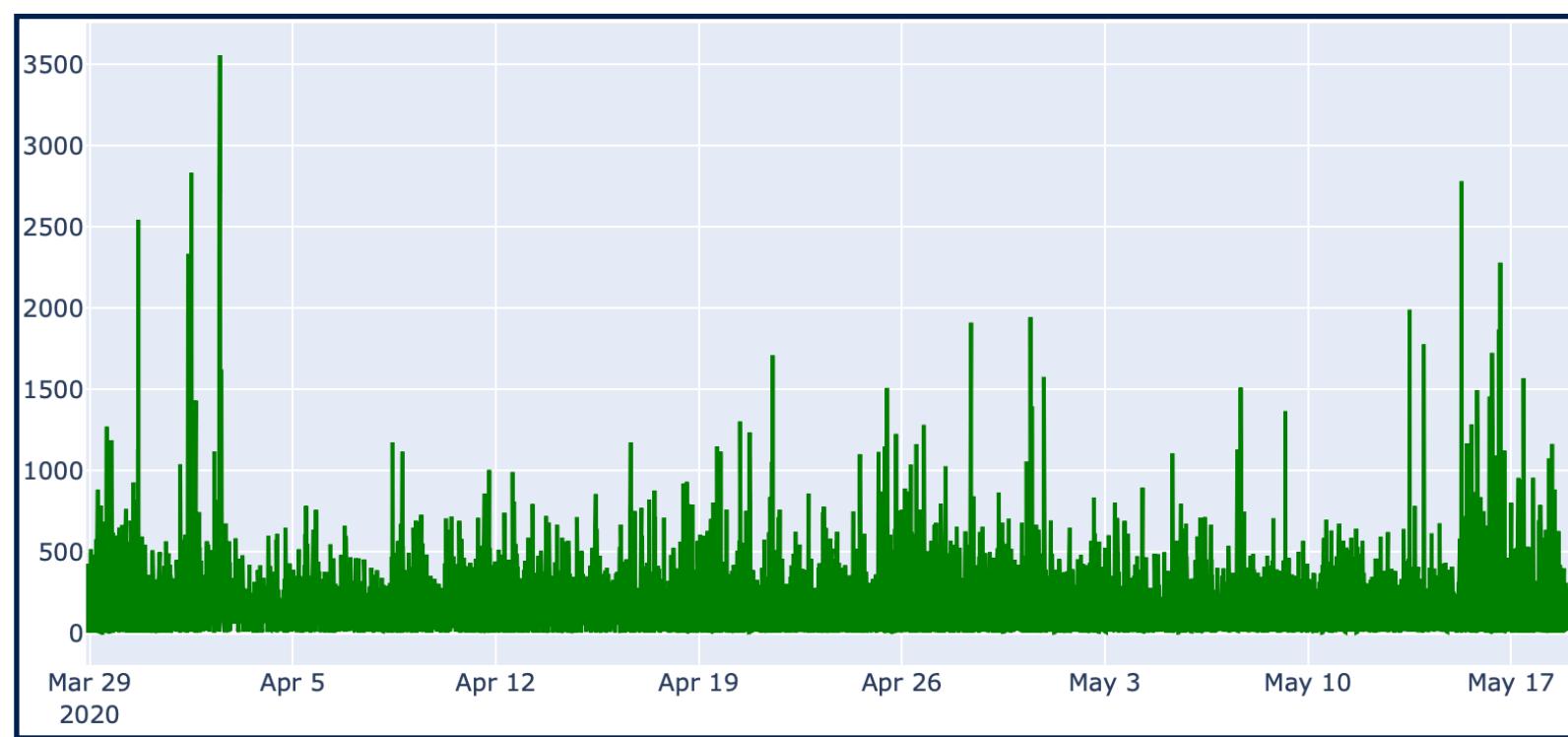
# Построение метрик



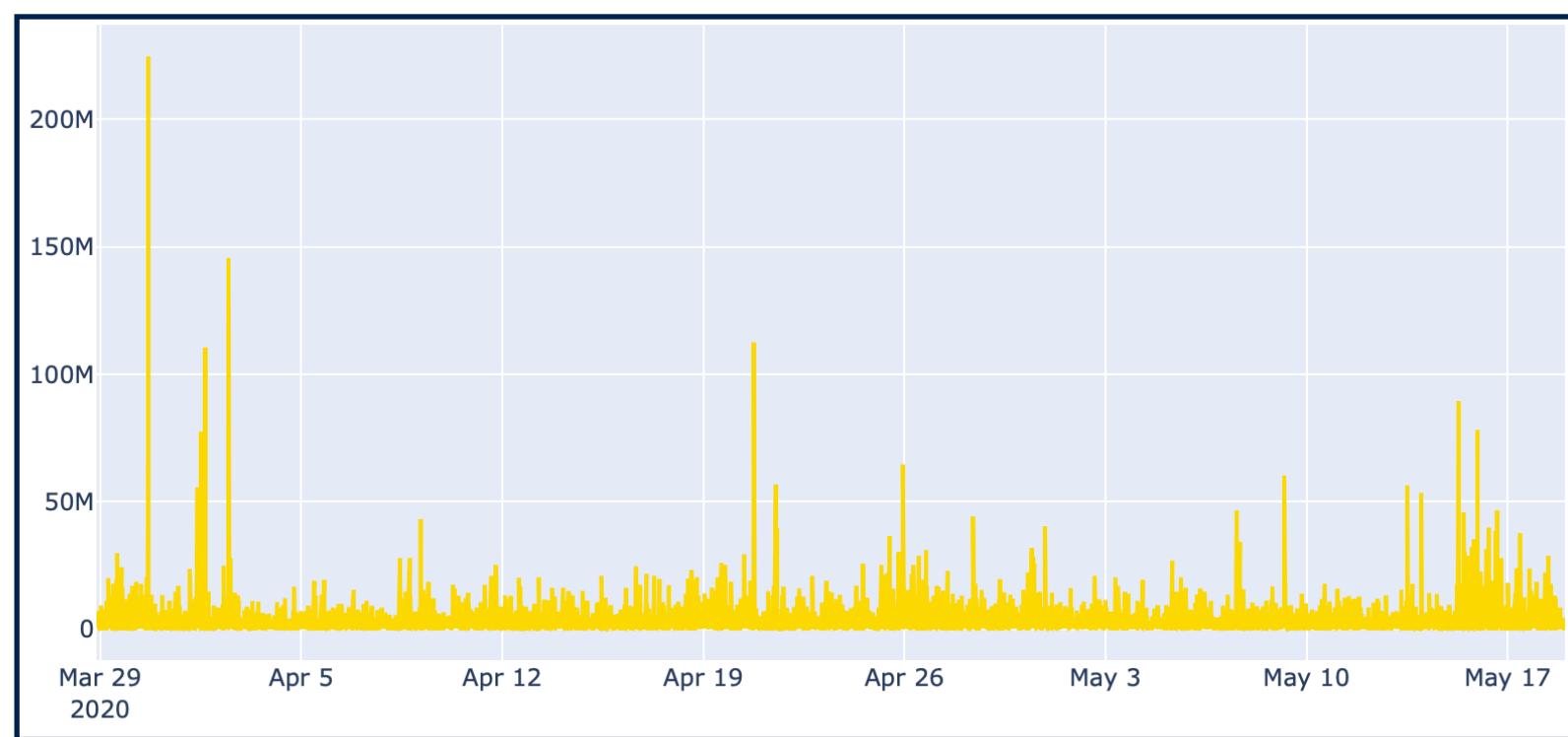
Ошибка  
модели



Среднее  
значение  
задержки



Дисперсия  
задержки



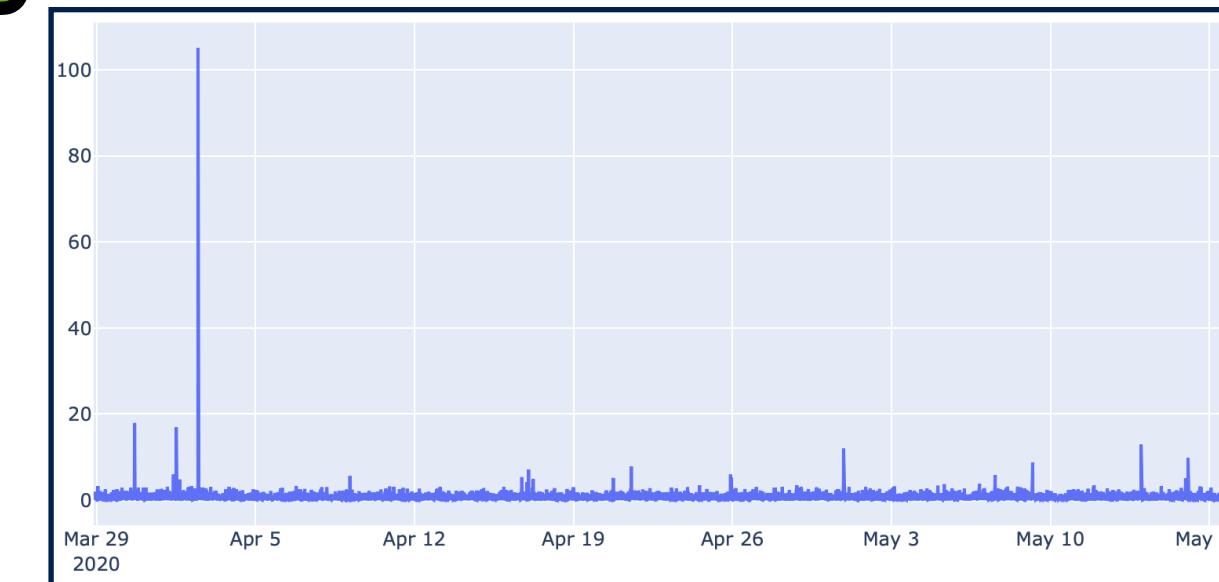
**Нормально  
распределённые  
метрики**, с  
которыми будем  
работать для  
поиска аномалий.

# Решающее правило

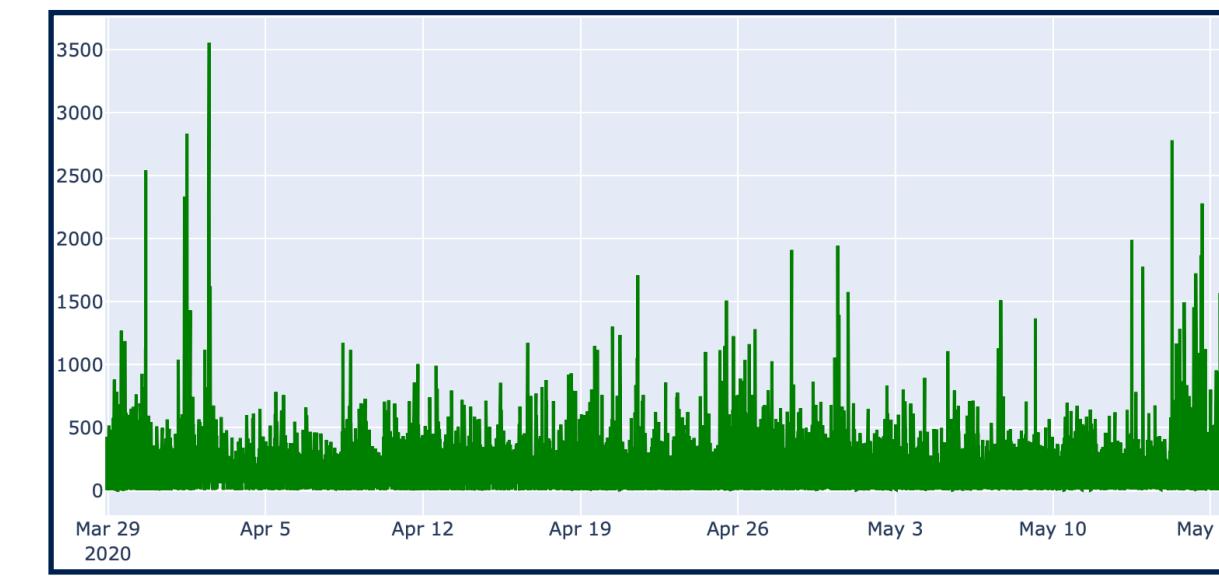


Значение  $\notin (\mu - k \cdot \sigma, \mu + k \cdot \sigma)$  - **аномальный момент**  
Значение  $\in (\mu - k \cdot \sigma, \mu + k \cdot \sigma)$  - **не аномальный момент**

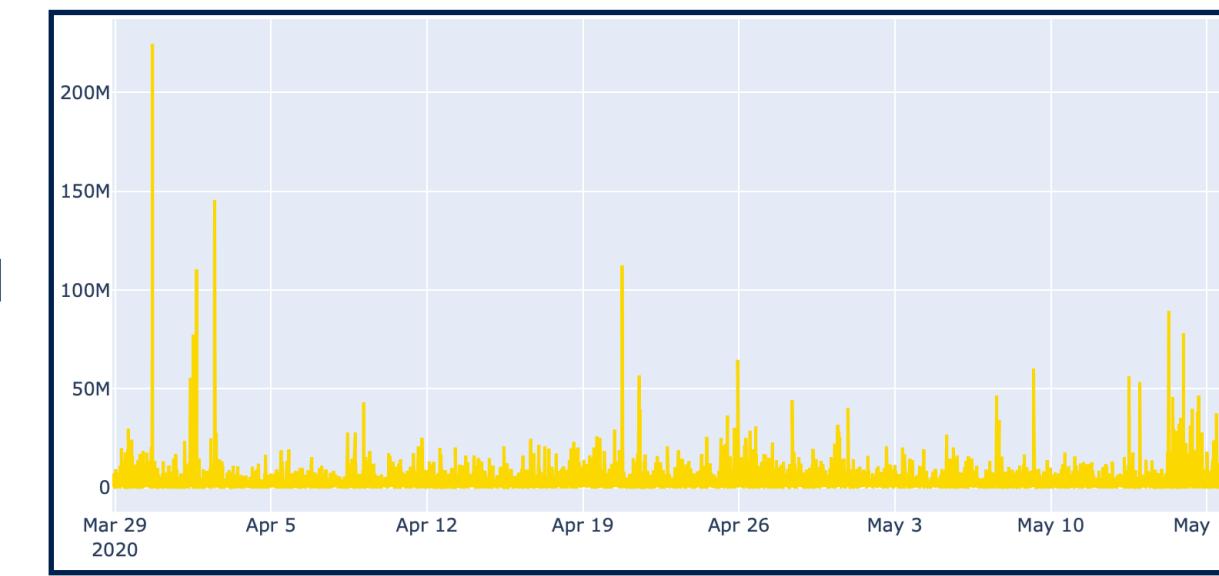
Ошибка модели



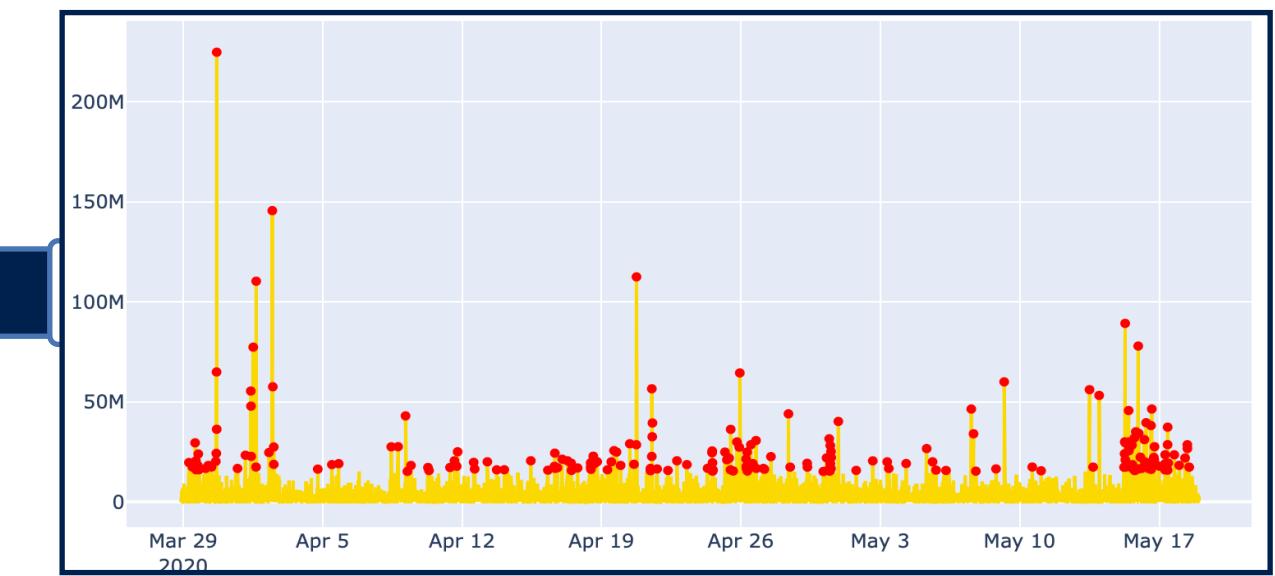
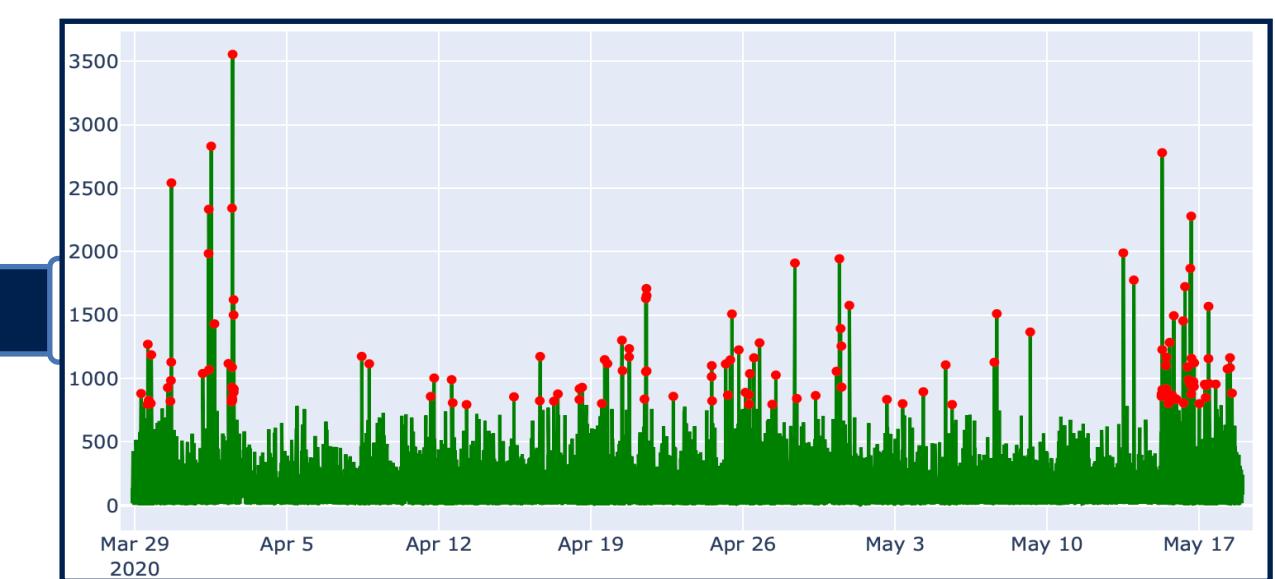
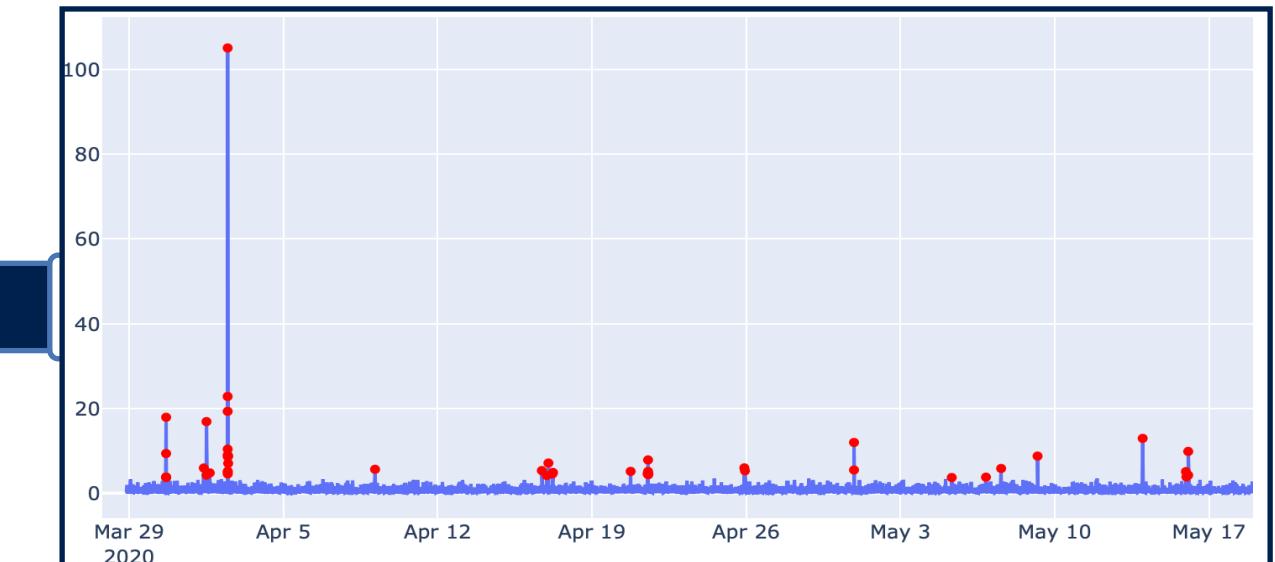
Среднее значение задержки



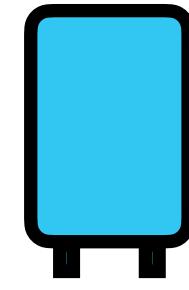
Дисперсия задержки



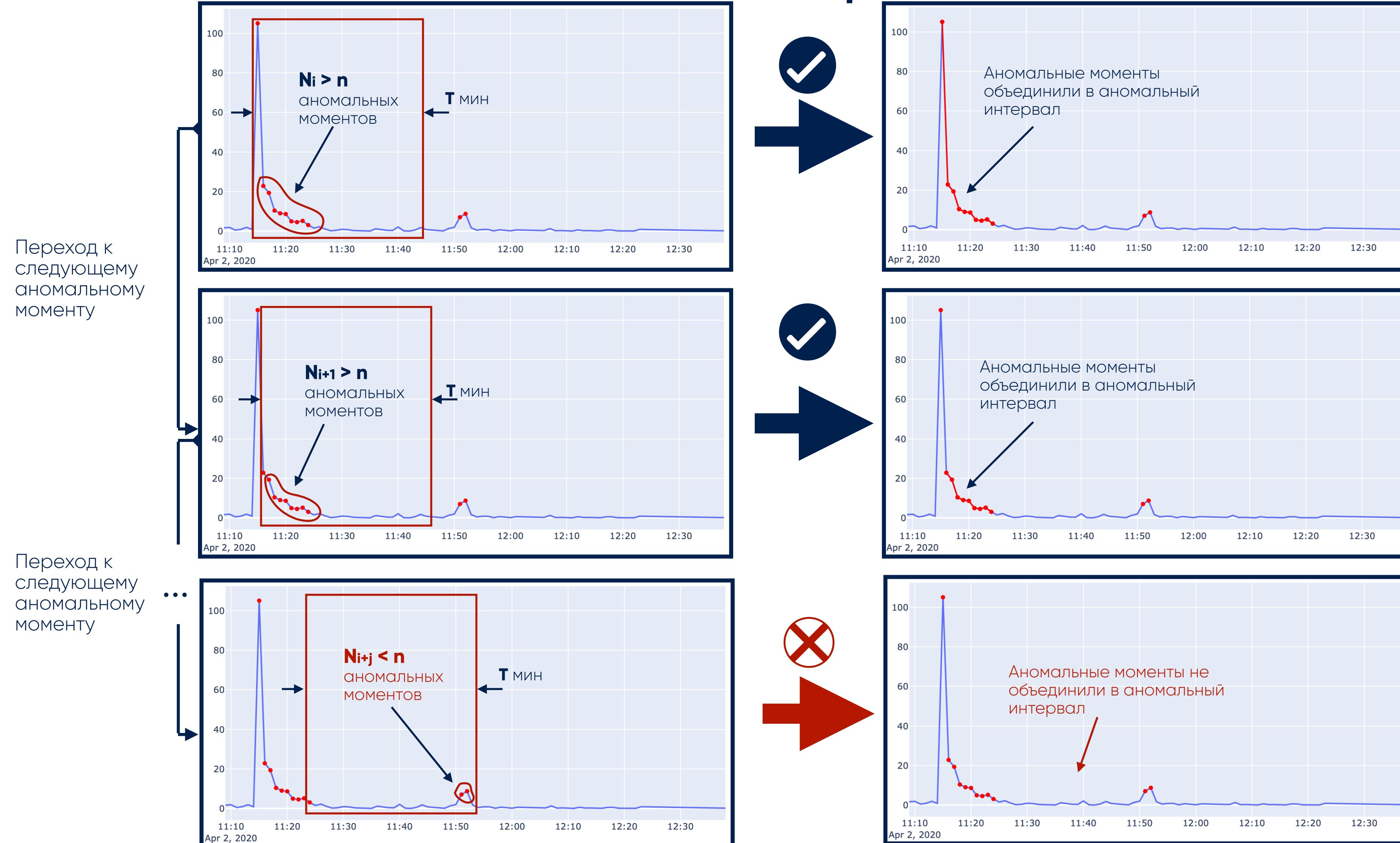
Решающее правило



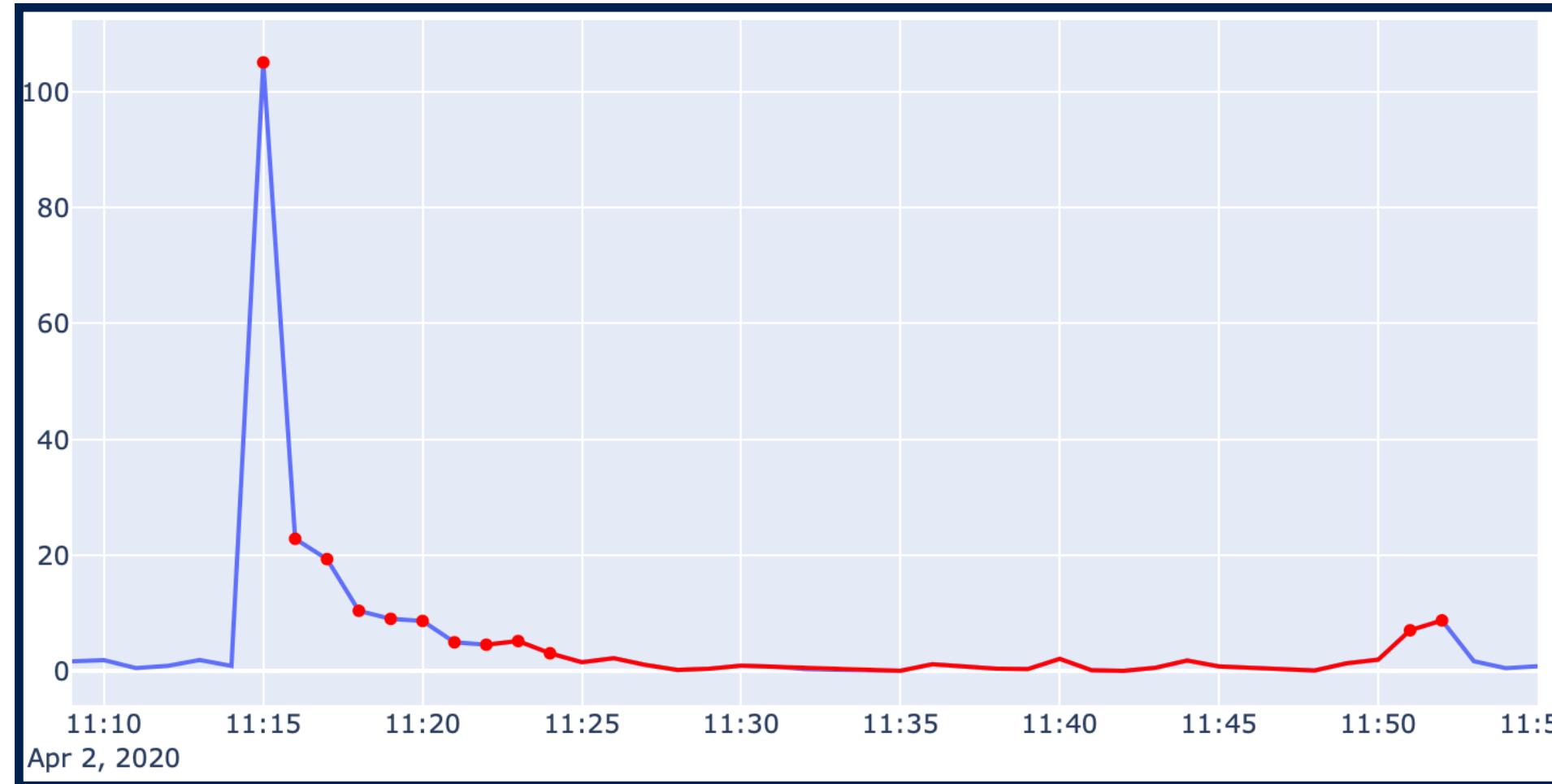
# Объединение в интервалы



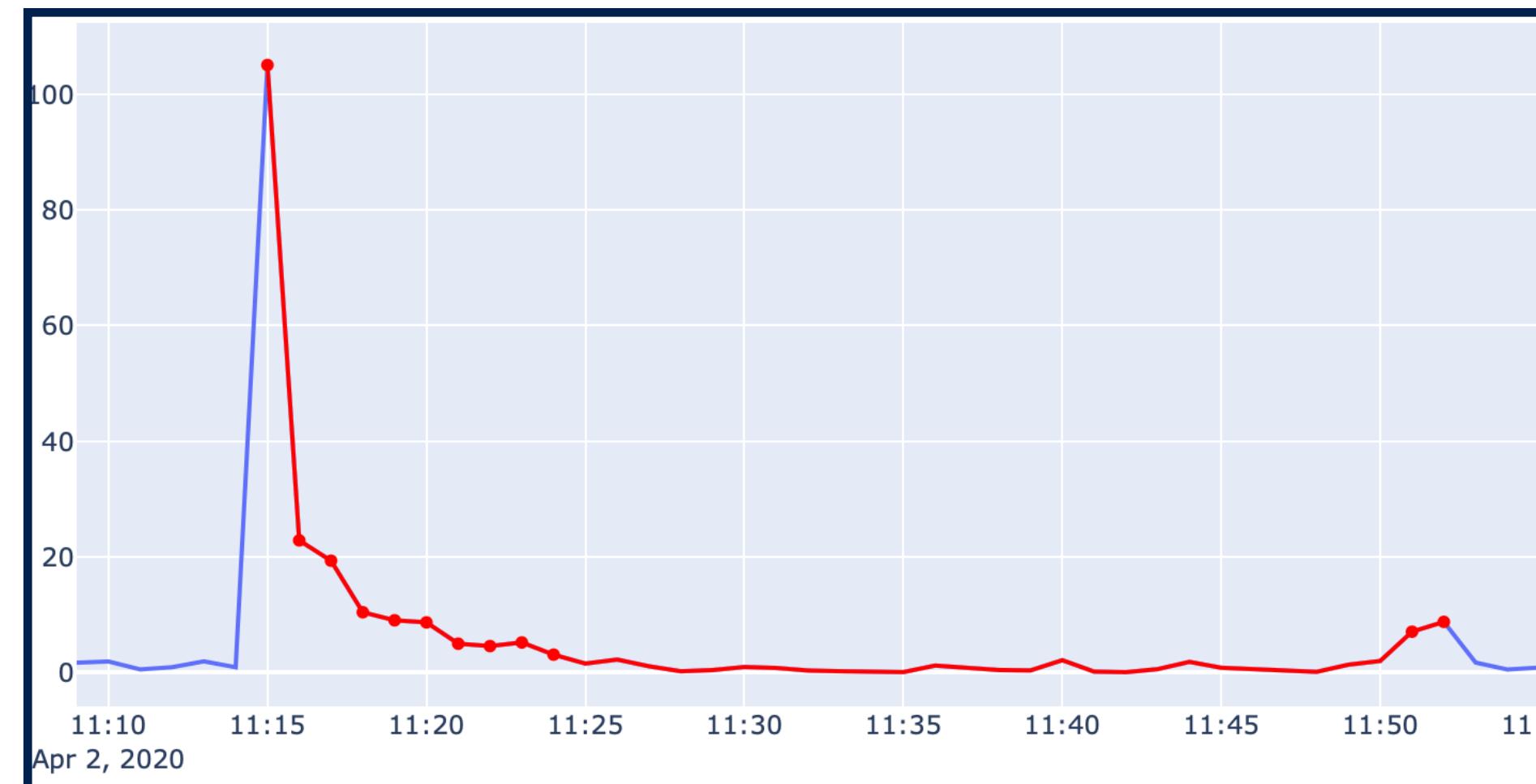
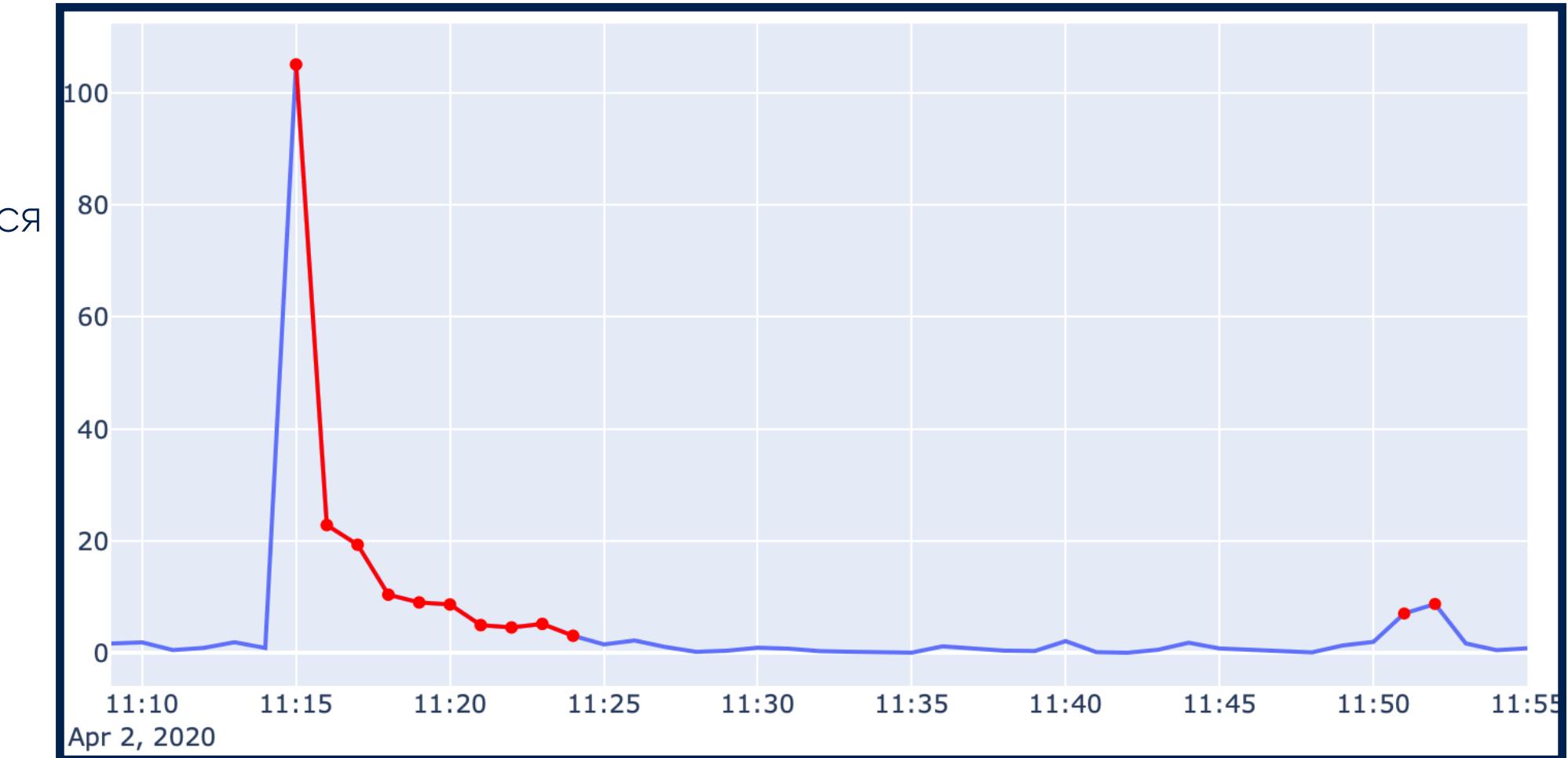
# Объединение в интервалы



# Объединение в интервалы



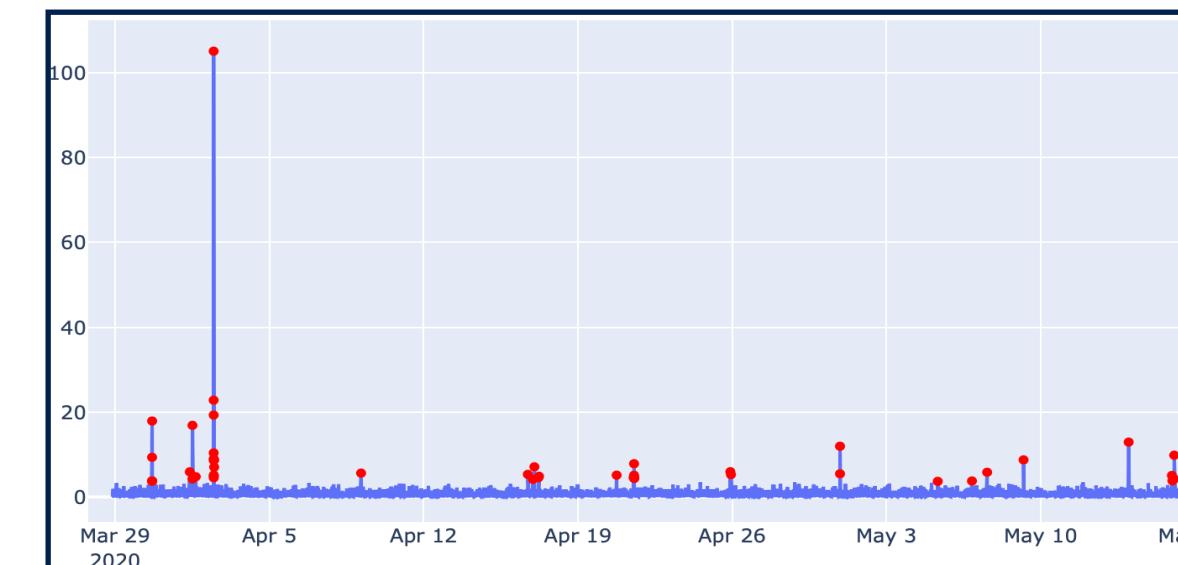
Объединяем  
пересекающиеся  
аномальные  
интервалы



# Объединение в интервалы

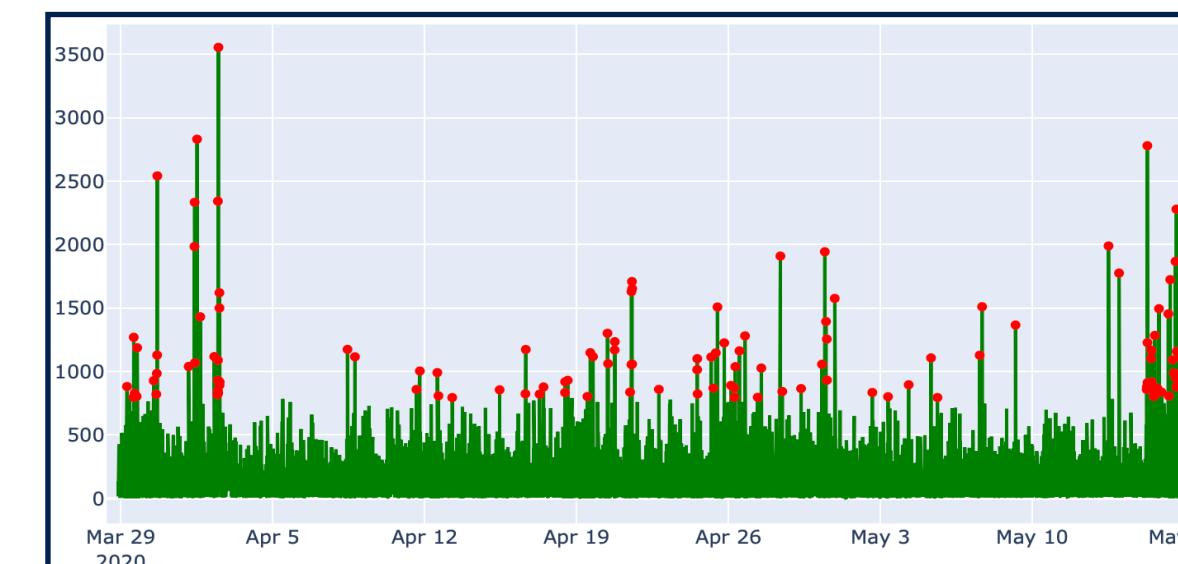
## overview

Ошибка  
модели

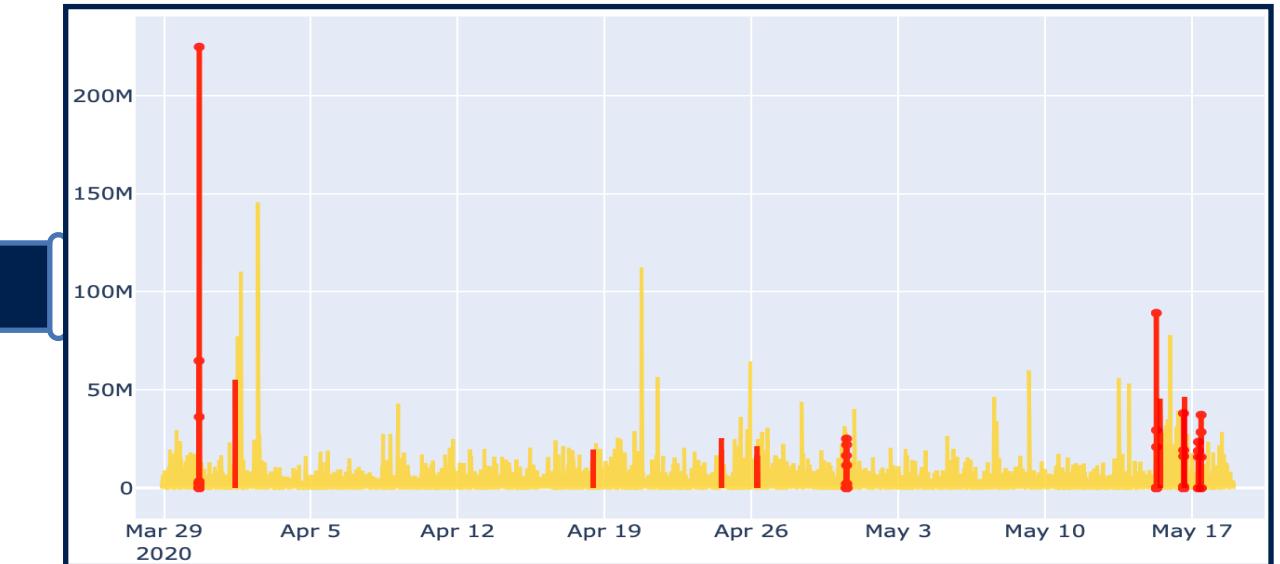
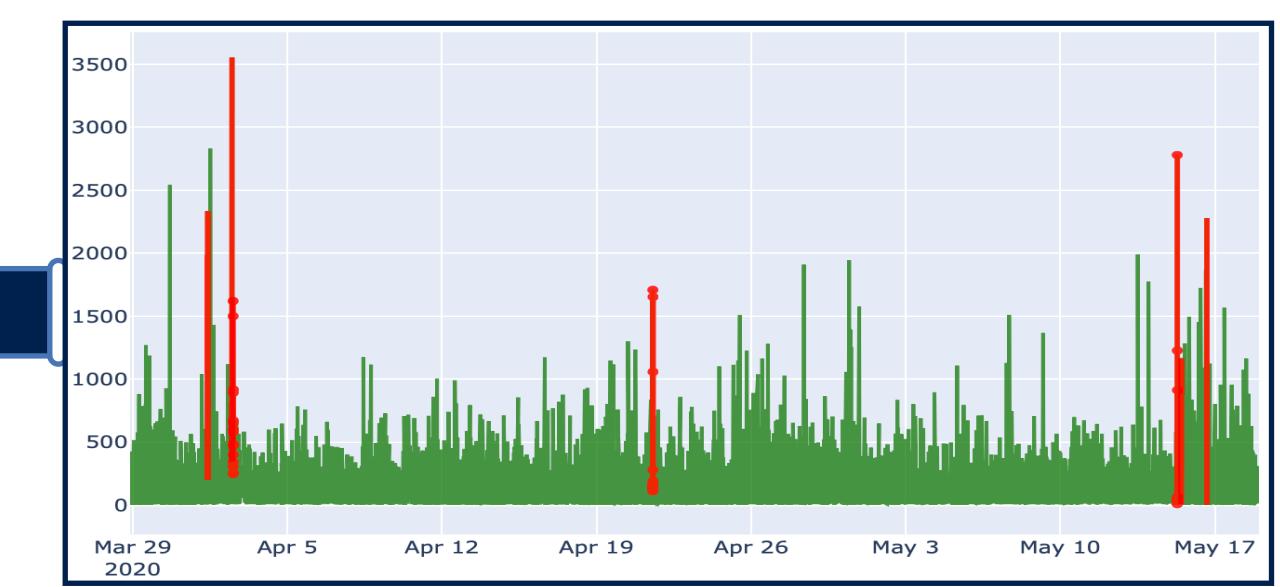
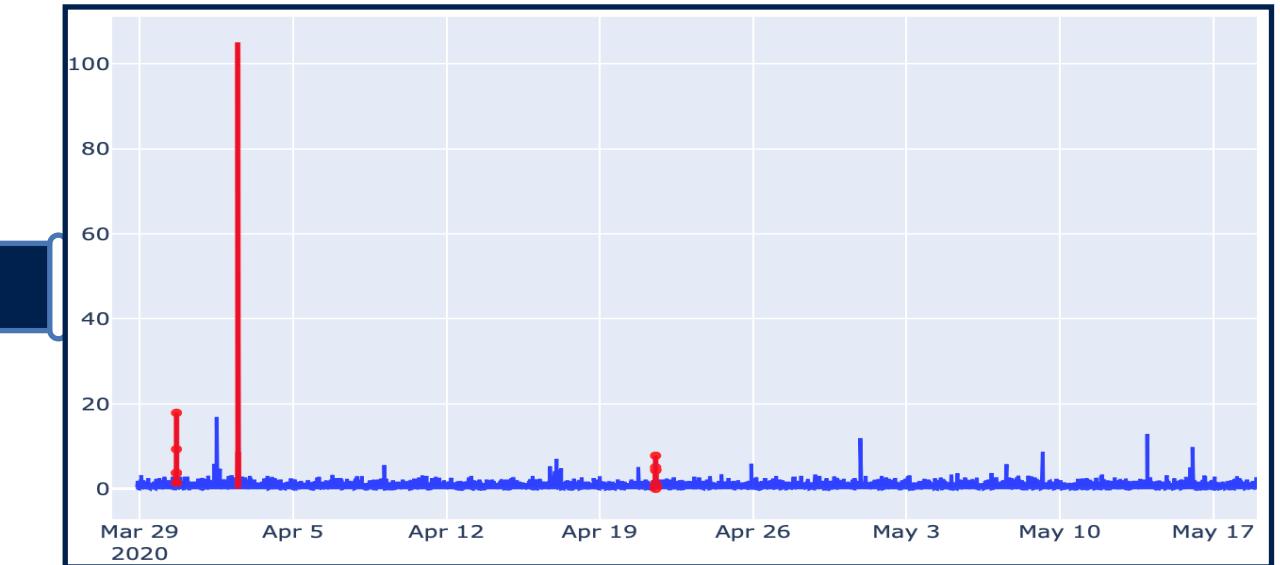
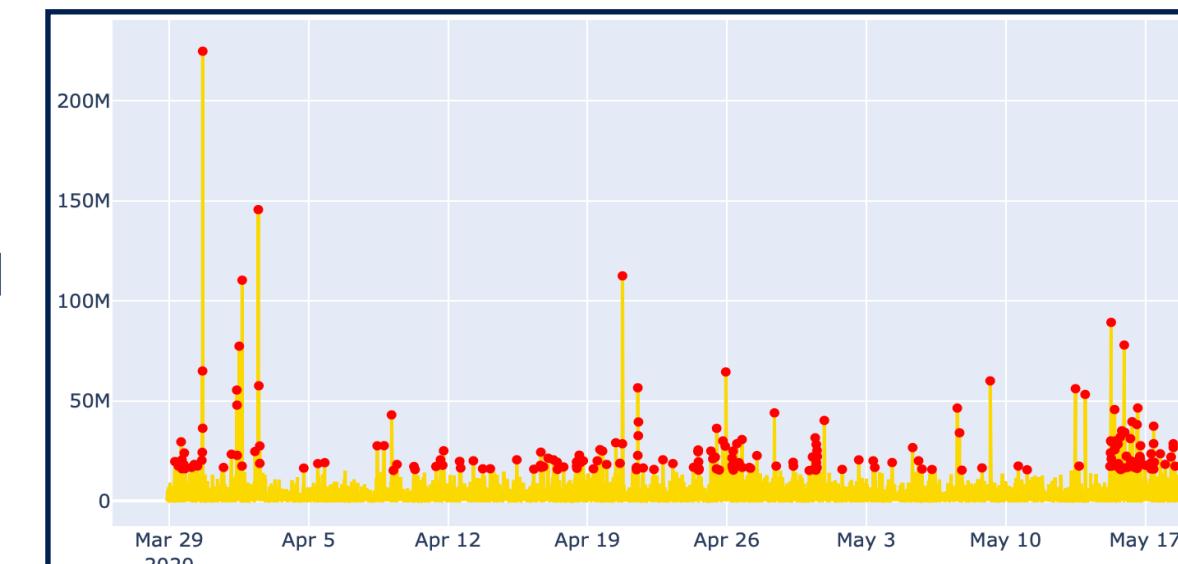


Алгоритм  
кластеризации

Среднее  
значение  
задержки



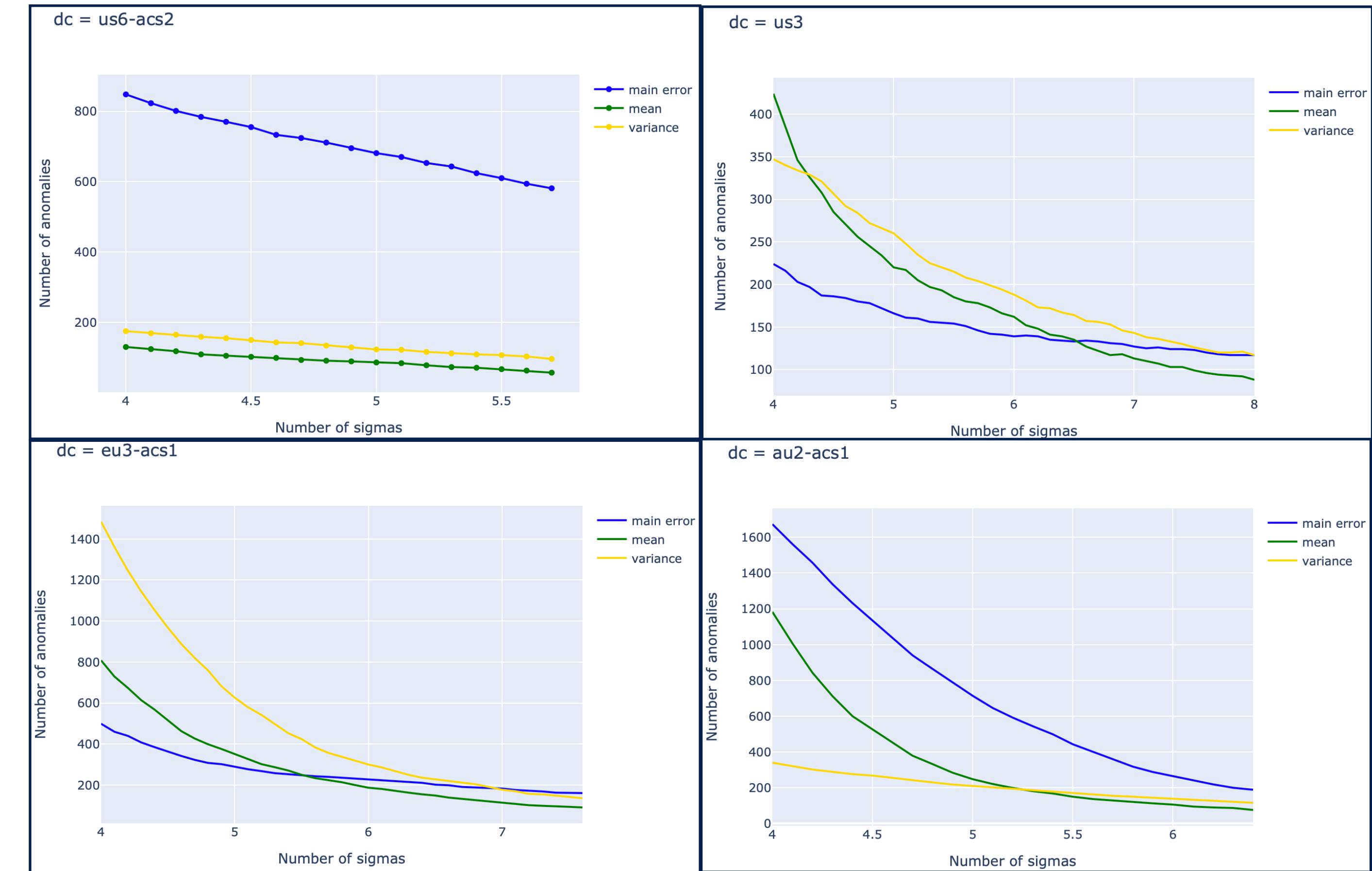
Дисперсия  
задержки



# Какую метрику использовать для поиска аномалий?

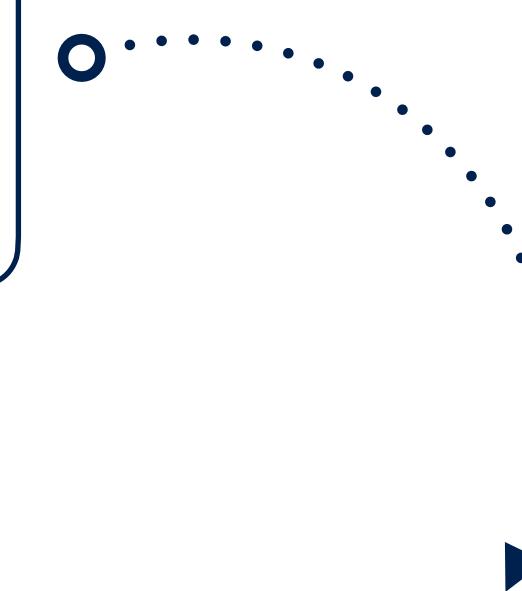
Рассмотрим зависимость количества аномальных интервалов, от параметра  $k$  решающего правила.

**Вывод:** отдельные метрики не являются универсальными



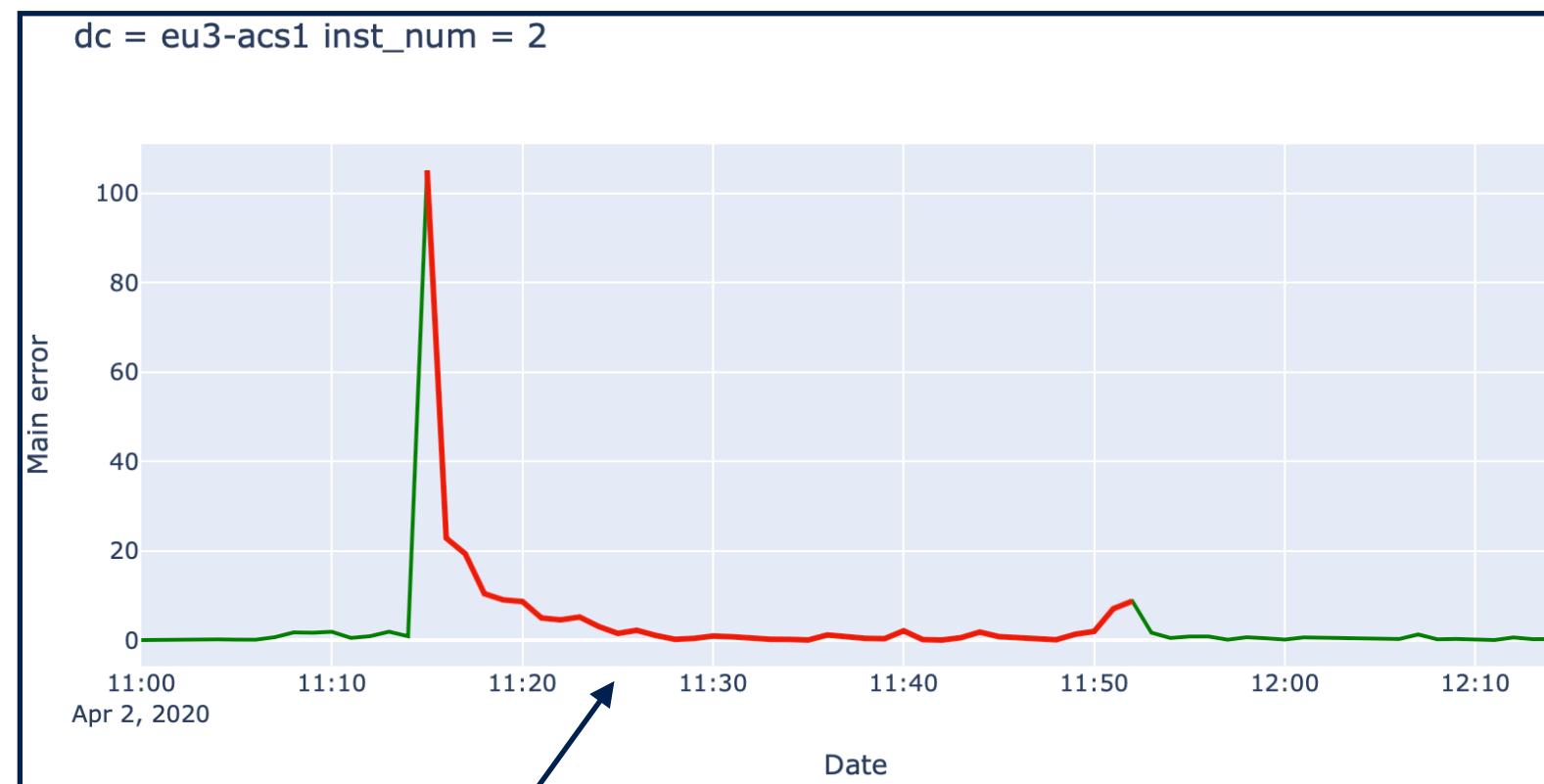
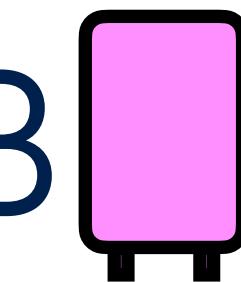
# Какую метрику использовать для поиска аномалий?

**Предположение:** во время аномального поведения сразу несколько из рассматриваемых метрик должны отклоняться от нормального поведения. То есть аномальные интервалы, найденные по разным метрикам, должны пересекаться.



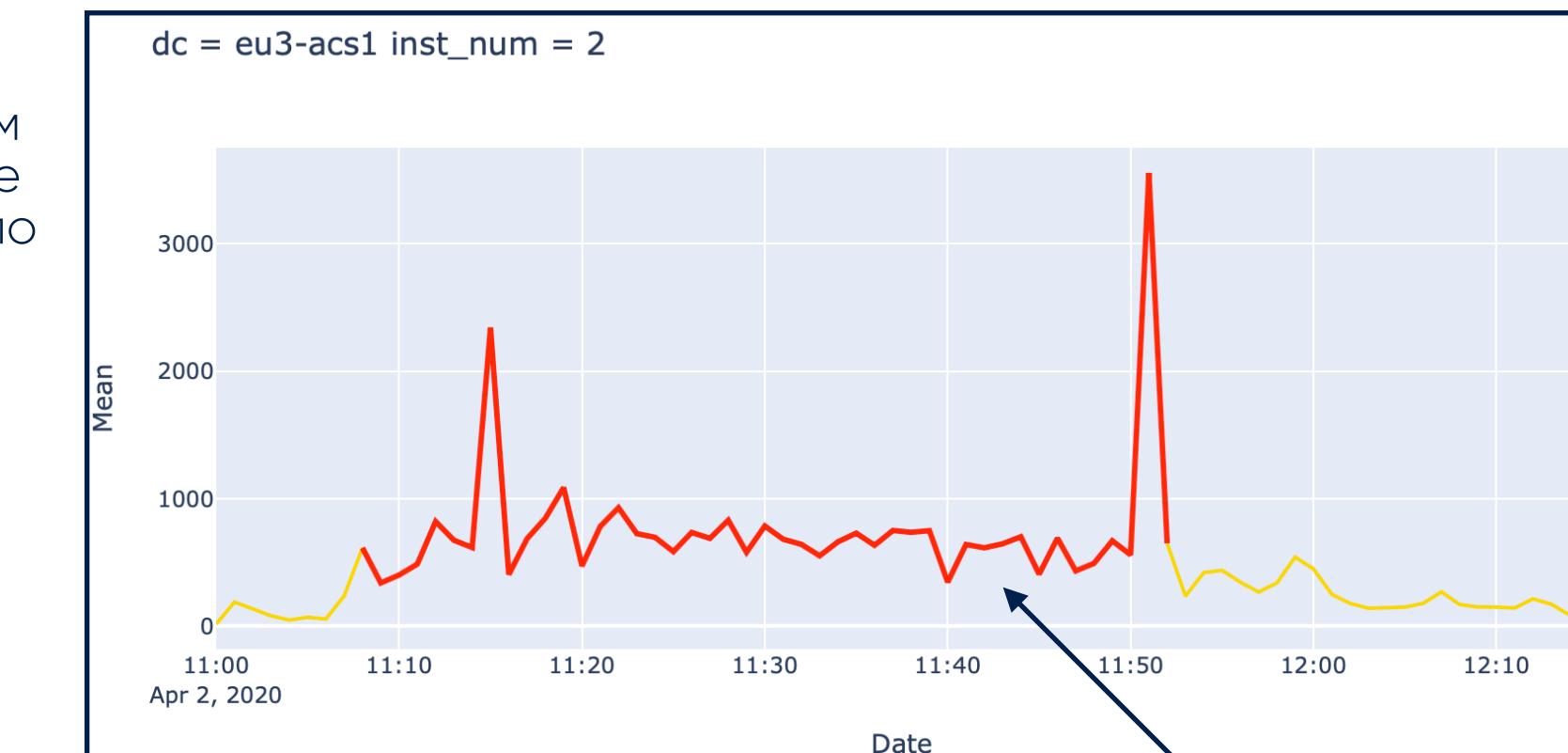
**Аномальными интервалами** будем считать все не пустые пересечения аномальных интервалов, найденных с помощью среднего значения задержек процессов и ошибки модели.

# Пересечение аномальных интервалов



Аномальный интервал,  
найденный с помощью  
ошибки модели

Пересекаем  
аномальные  
интервалы по  
разным  
метрикам



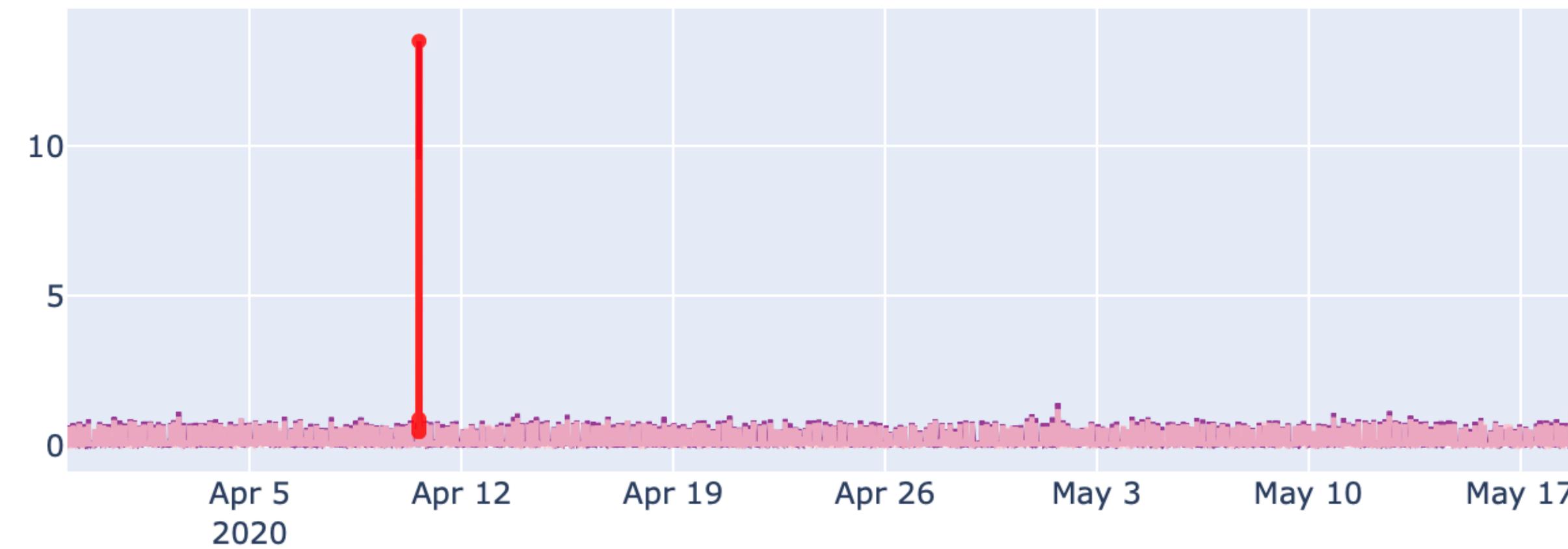
Аномальный интервал,  
найденный с помощью  
среднего значения  
запросов



Итоговый аномальный  
интервал на целевой  
переменной

# Примеры найденных аномалий

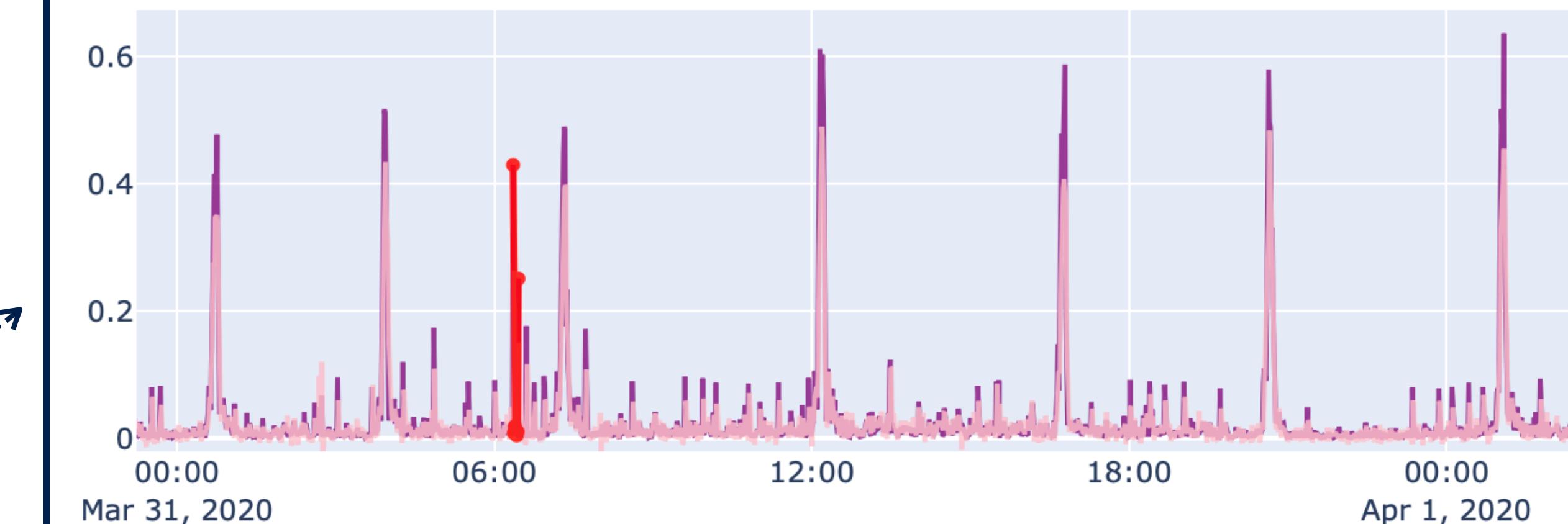
dc = au2-ac51 inst\_num = 14



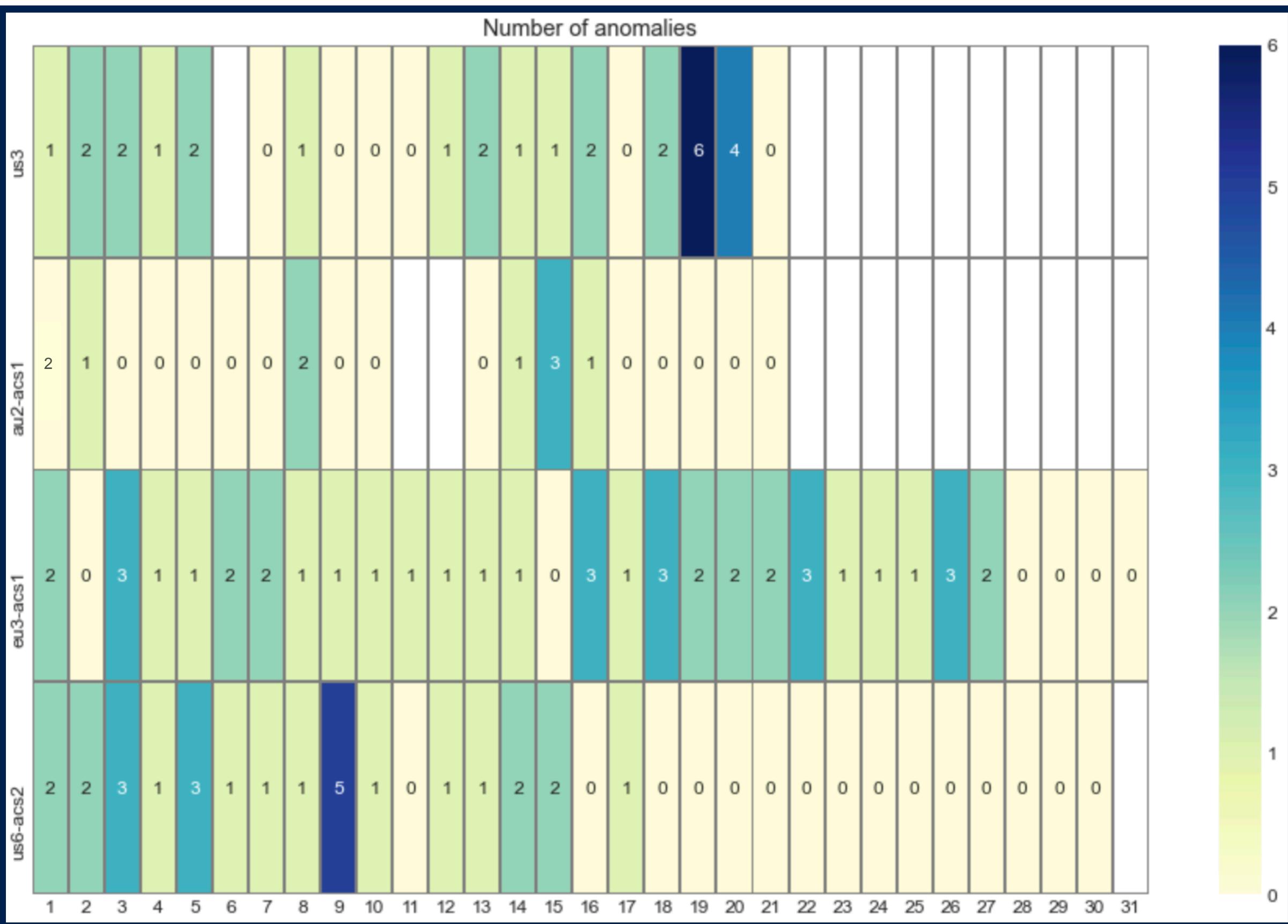
Очевидно

Рост задержки не  
соответствует  
периоду

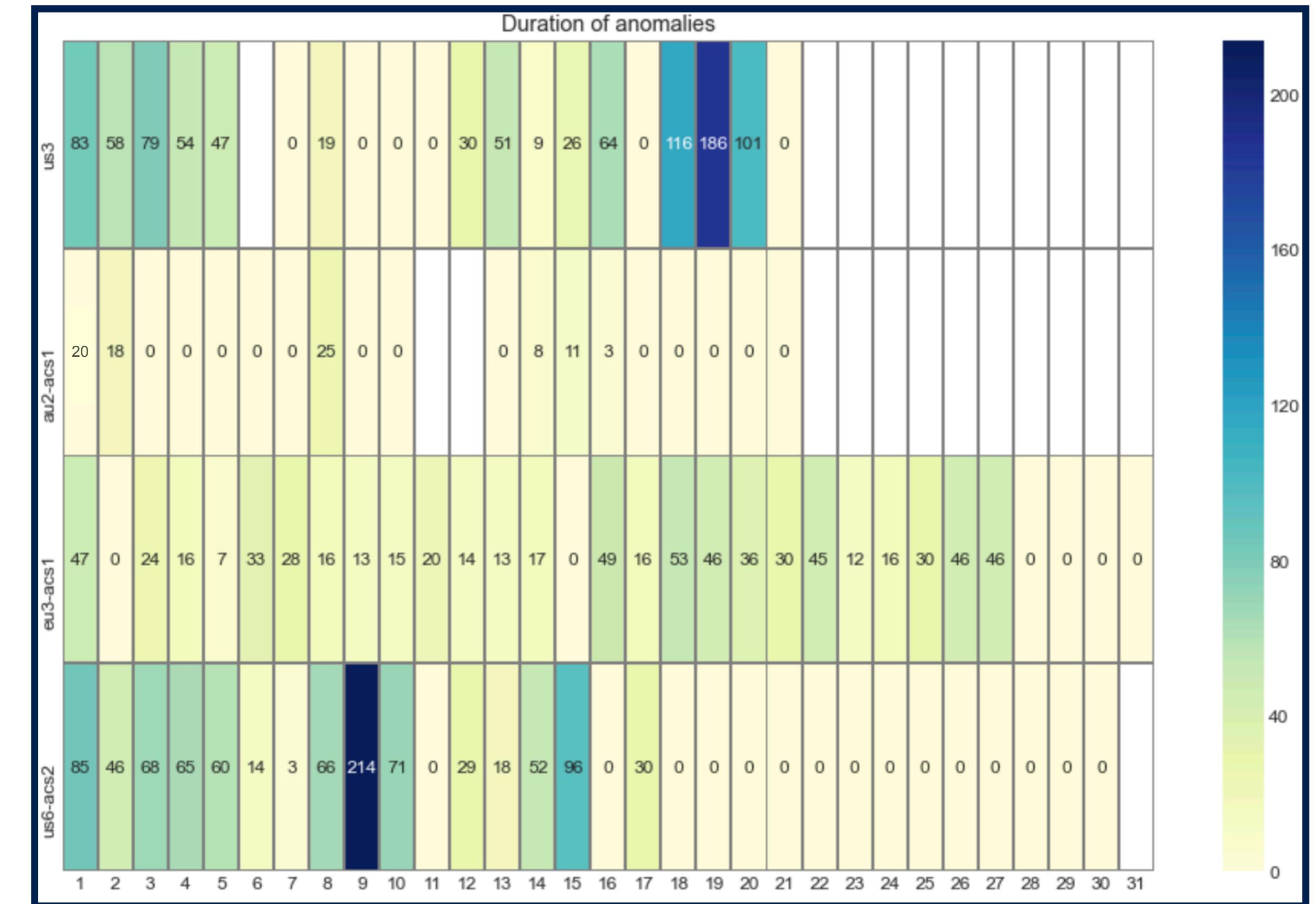
dc = au2-ac51 inst\_num = 8



# Тепловая карта количества аномалий за 2 месяца

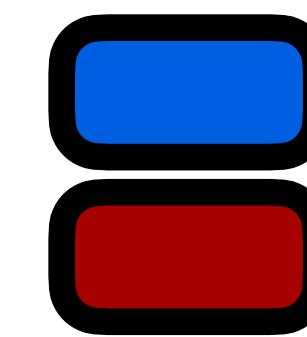


# Тепловая карта суммарной длительности аномалий за 2 месяца

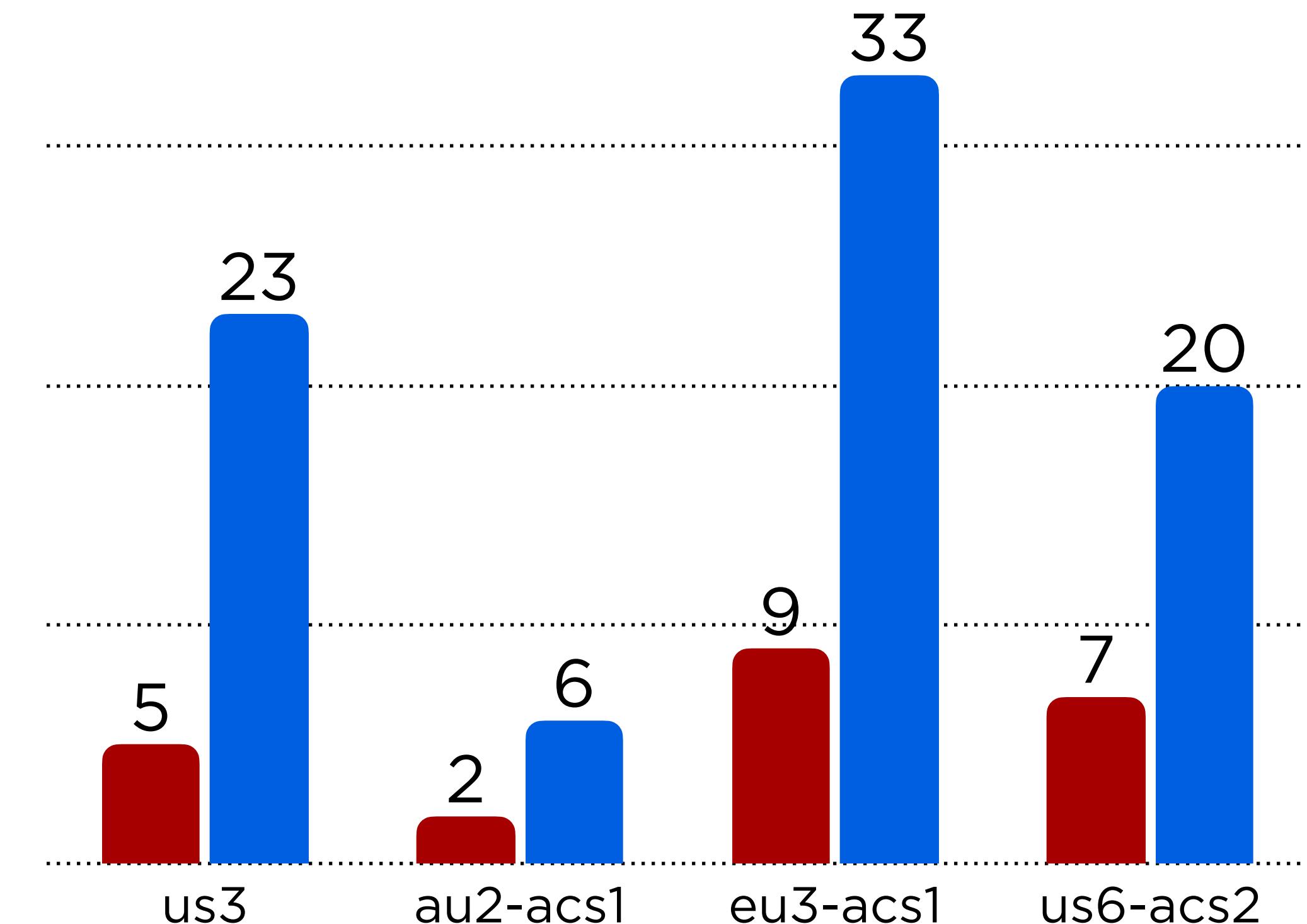


# Корректность алгоритма

78%  
точность  
алгоритма



- True Positive
- False Positive



# Заключение



Универсальность  
для всего  
многообразия  
хранилищ данных  
компании Acronis



Полное отсутствие  
ручной работы



Легко  
настраивается  
под требования  
заинтересованных  
сторон



Однако  
корректная работа  
при любой нагрузке  
системы.



Высокая точность  
и низкий уровень  
ложных  
срабатываний

# Спасибо за внимание!