Яндекс

Практическое прогнозирование

Максим Гришин, 5 июня 2023

Давайте представлюсь

Максим Гришин
Старший аналитик-разработчик

- 4.5 года в Яндексе
- Прогнозирую ключевые метрики поискового портала
- Разрабатываю внутренние инструменты для прогнозирования

Содержание

Туториалы про прогнозы замалчивают главное

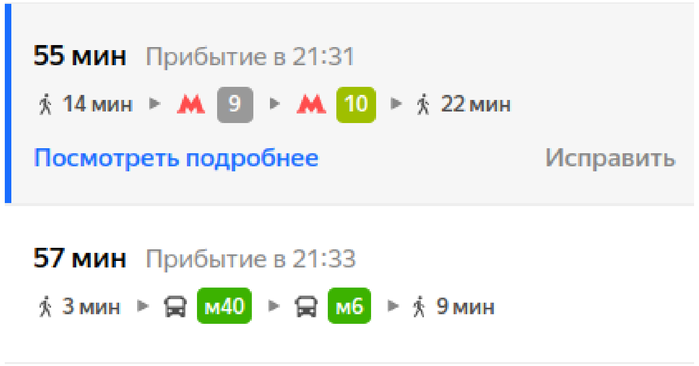
01

О чём речь?

Виды аналитики

- Дескриптивная / описательная: что было?
- Предиктивная / предсказательная: что будет?
- Прескриптивная / предписывающая: что делать?

Прогнозы в быту



Images by vectorjuice on Freepik: 1, 2

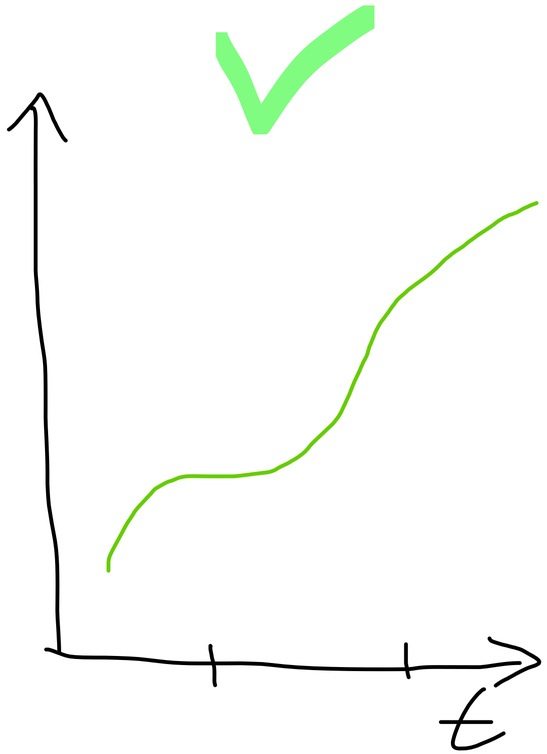
Ключевые показатели

- Аудитория сервиса, показатели её оттока (churn) и удержания (retention)
- Установки, активации
- Выручка, расходы, доходы
- Показы, клики, конверсии в покупку

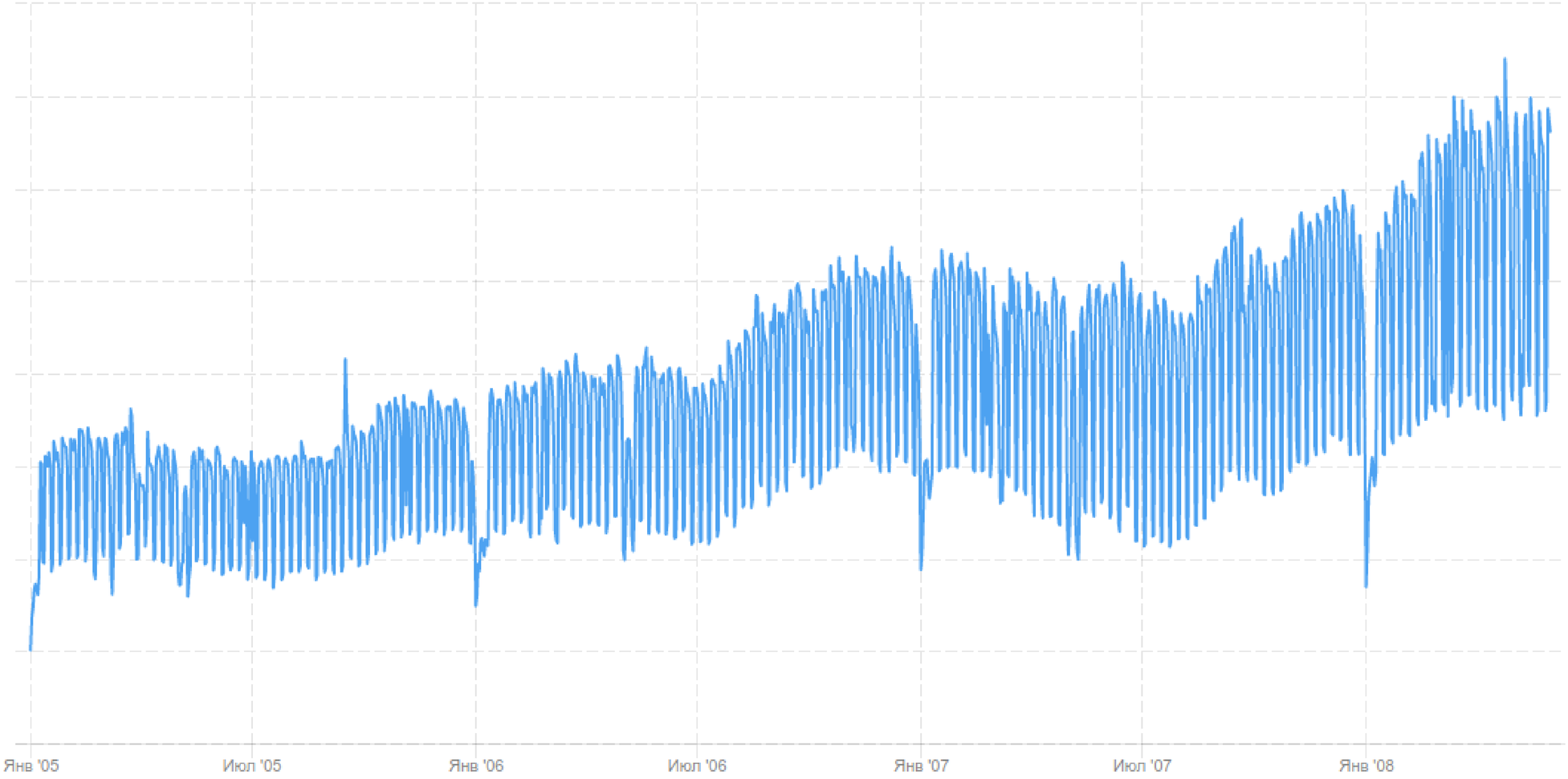
Временные ряды

Значение зависит от времени $\mathbb{R} = f(t)$

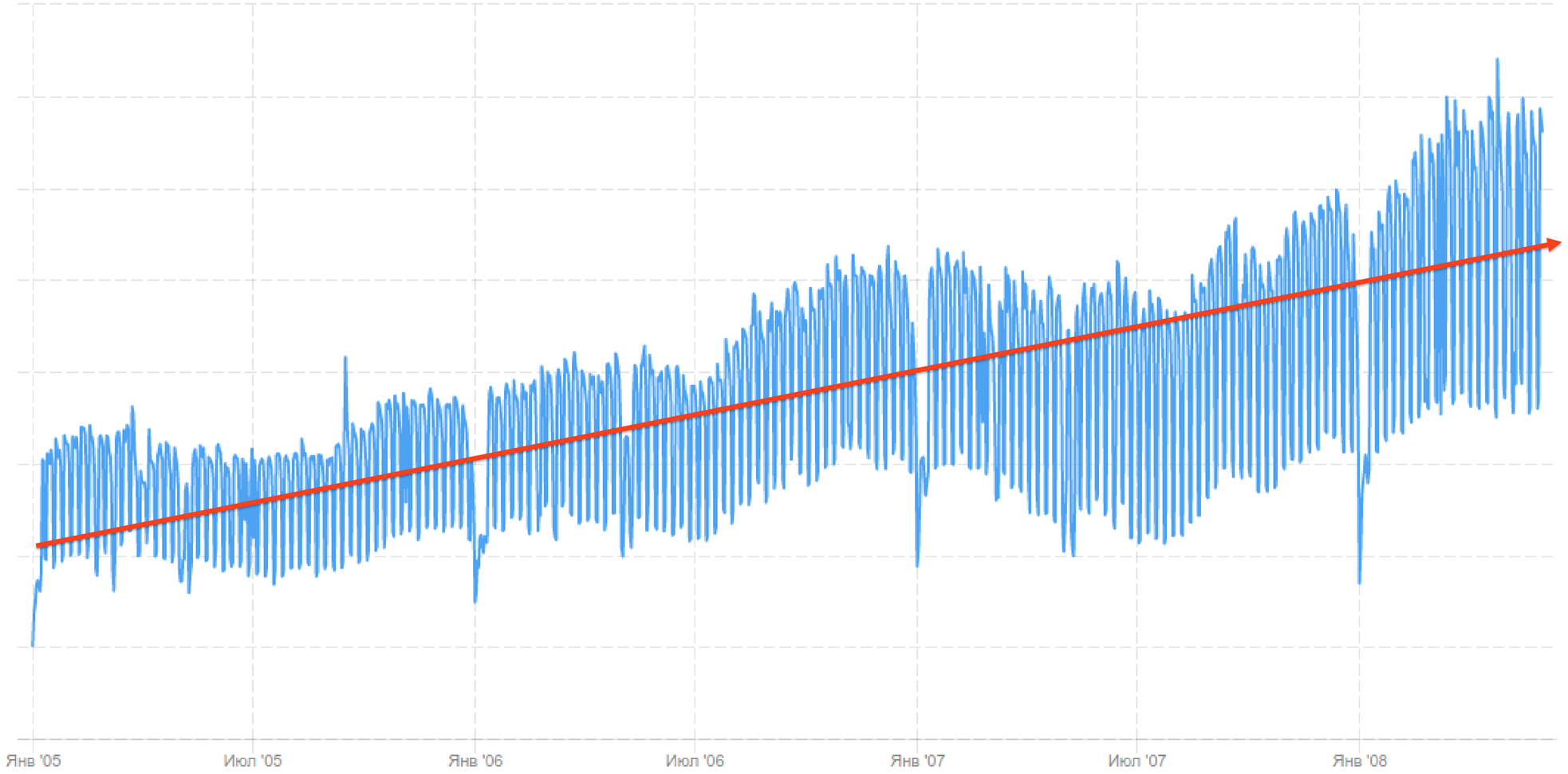
Каждому моменту времени соответствует одно значение



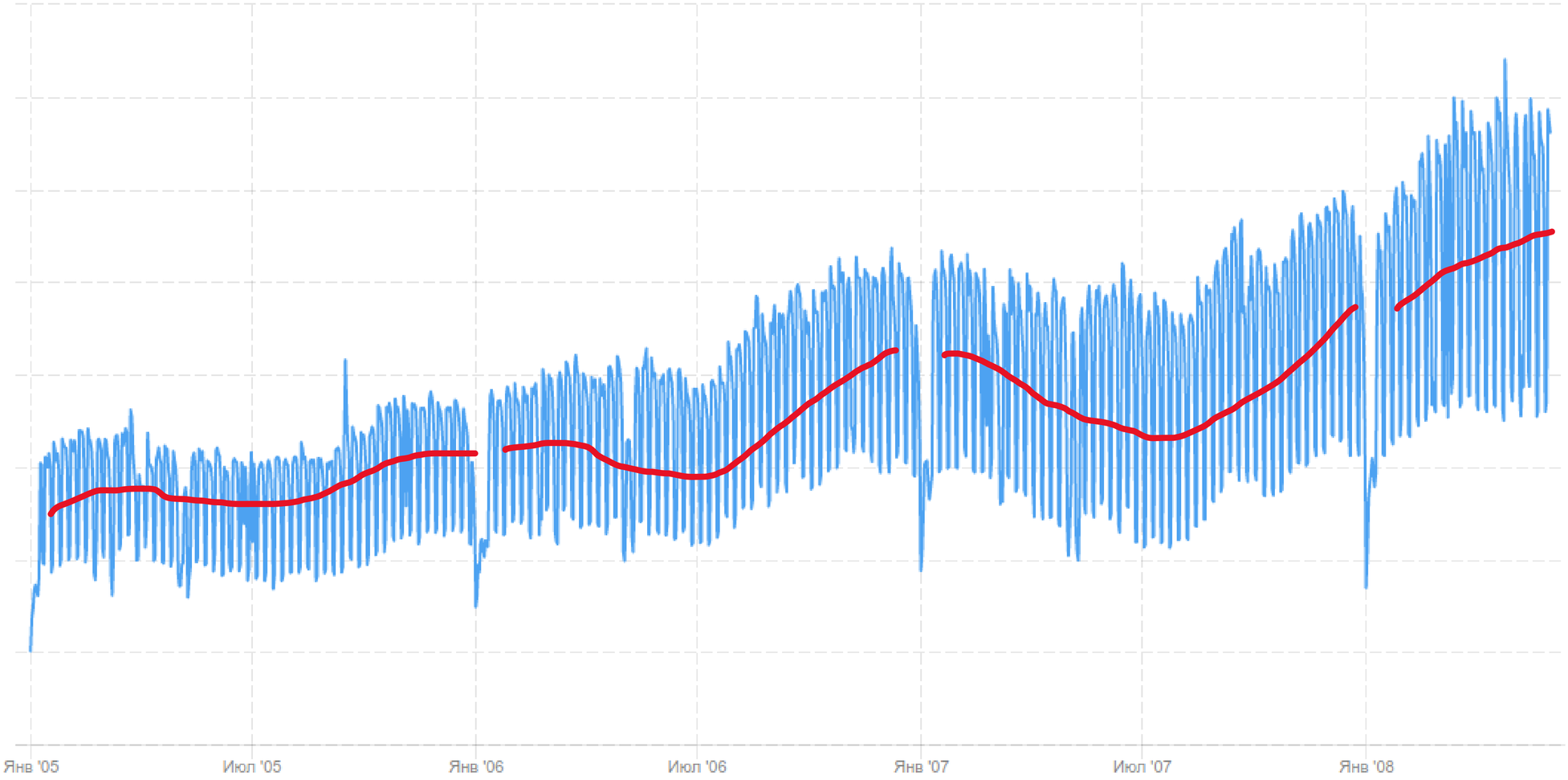
Временной ряд



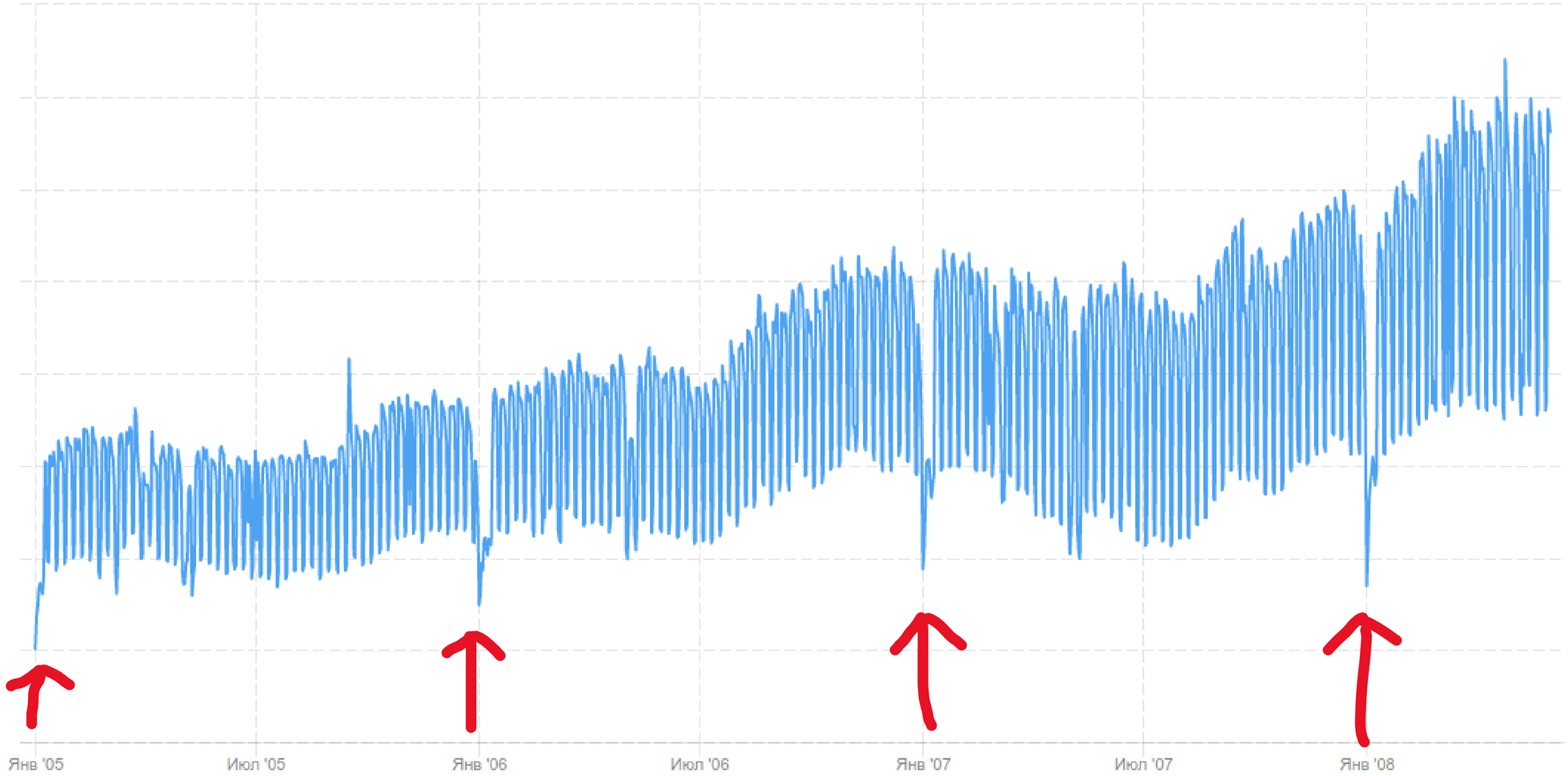
Тренд



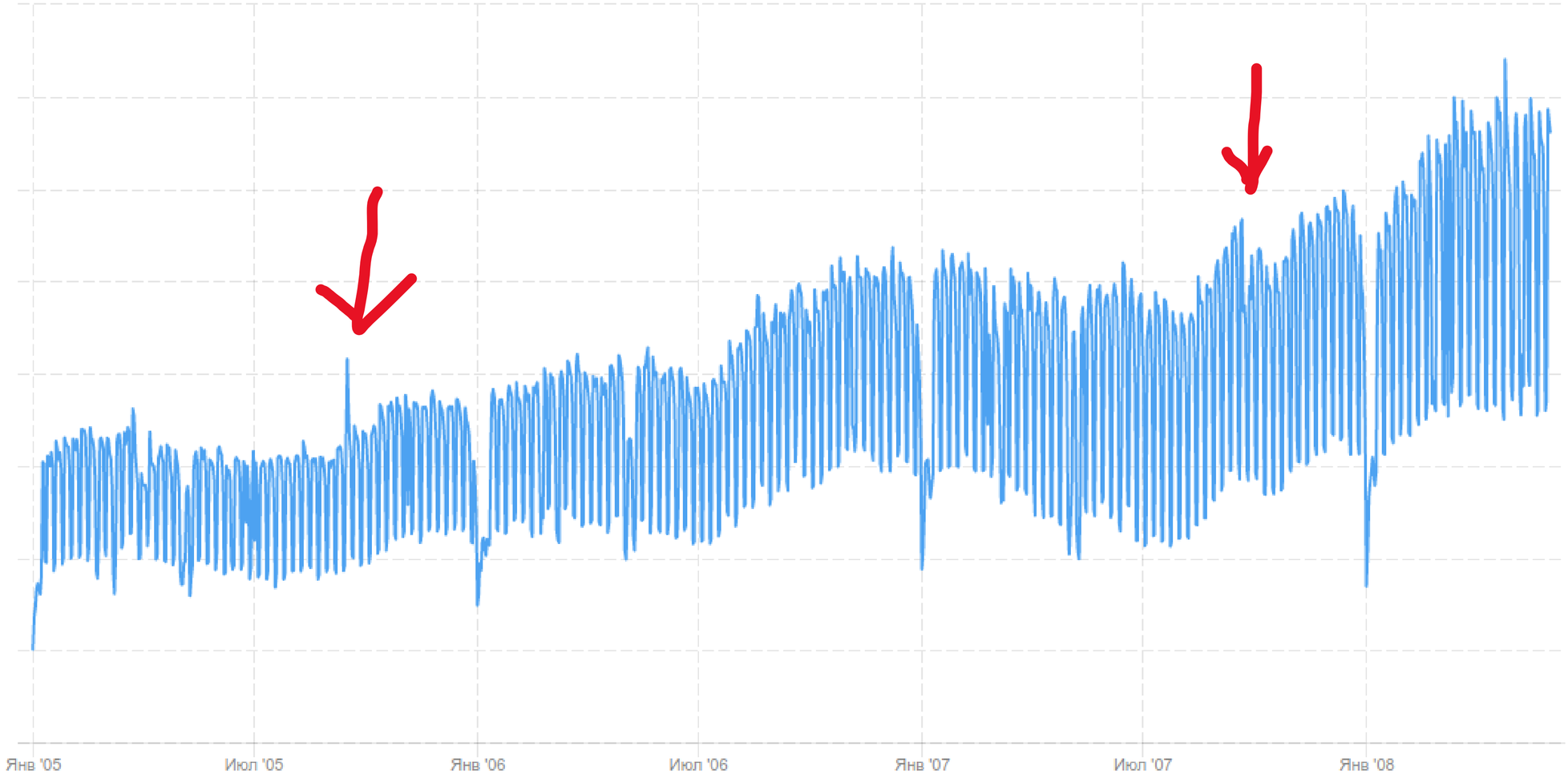
Внутригодовая сезонность



Праздники



Аномалии

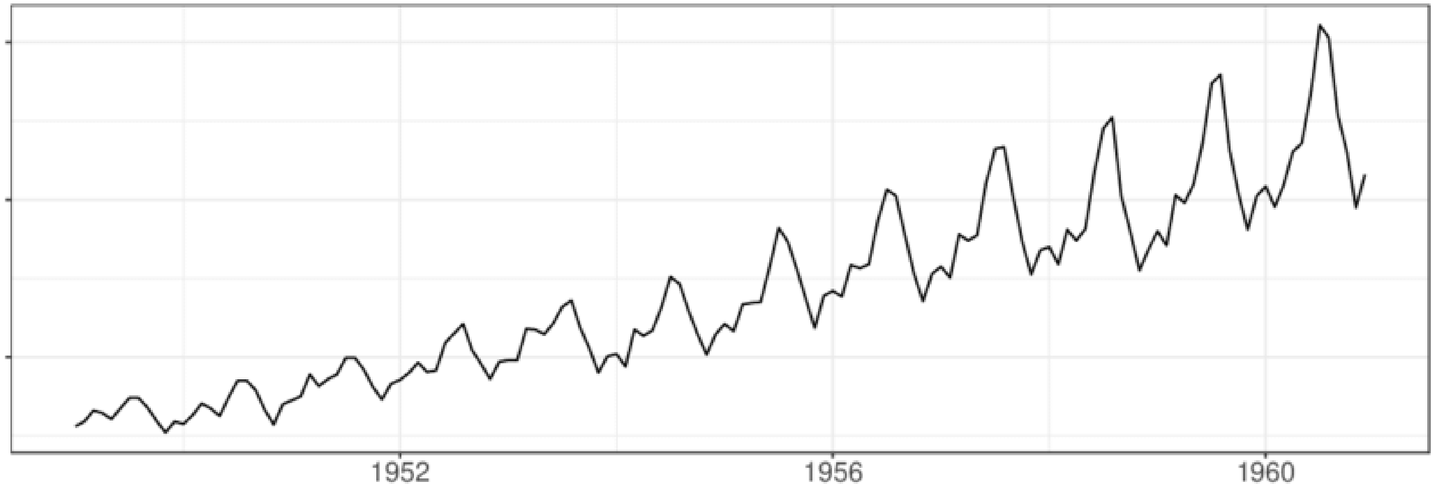


Внутринедельная сезонность

Виды сезонности

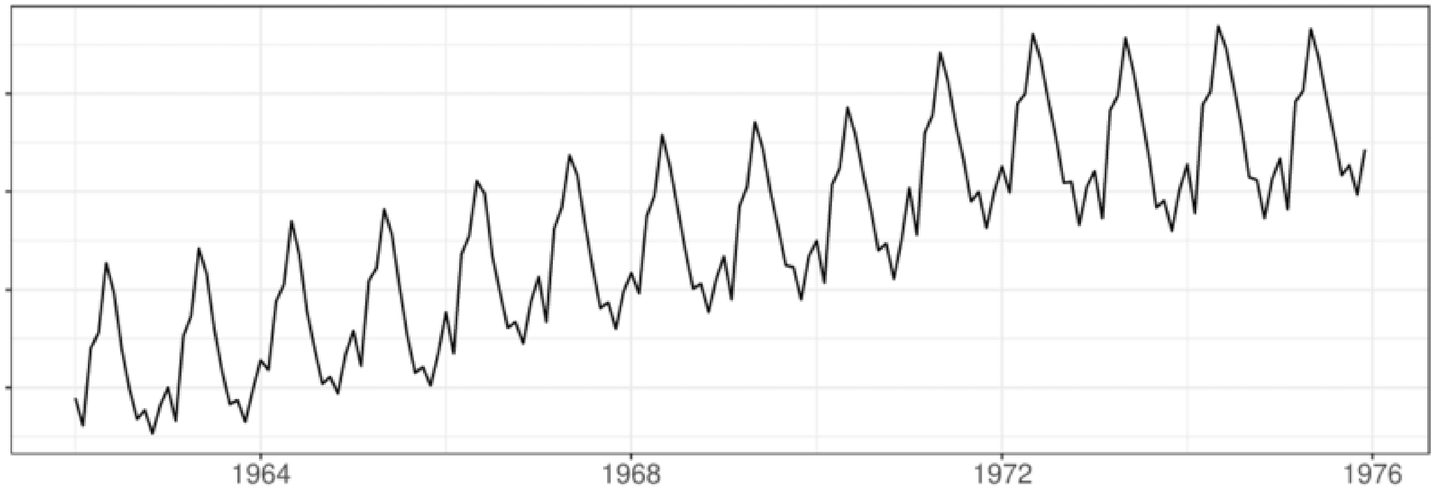
- Мультипликативная

  тренд * сезонность



- Аддитивная

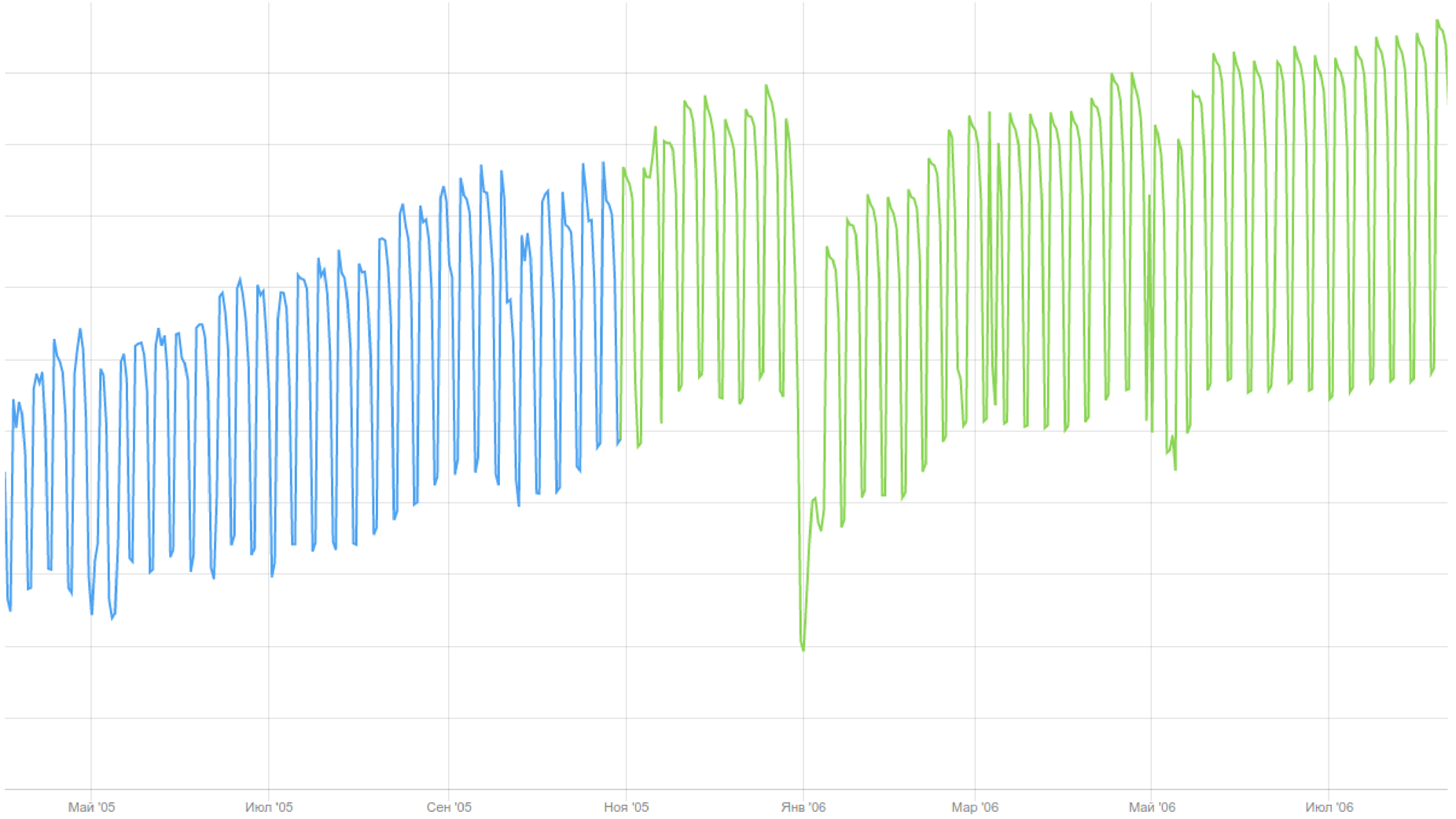
  тренд + сезонность



Credit: https://www.researchgate.net/figure/Examples-for-multiplicative-and-additive-relationship-between-time-series-components_fig2_348592737

Прогноз - экстраполяция

факт



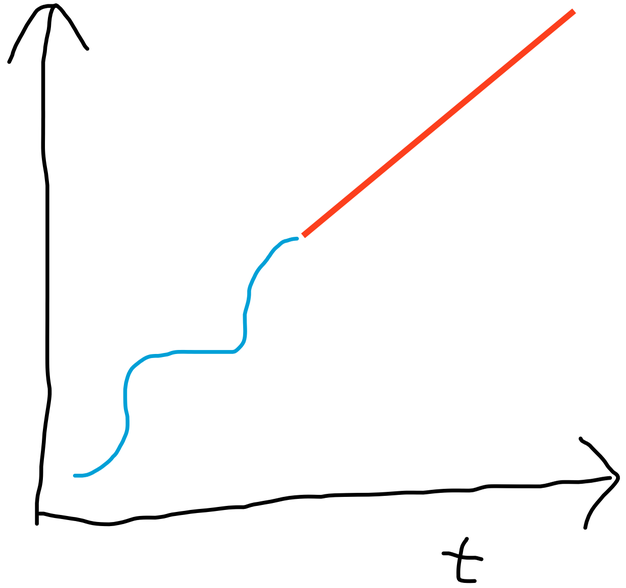
прогноз

02

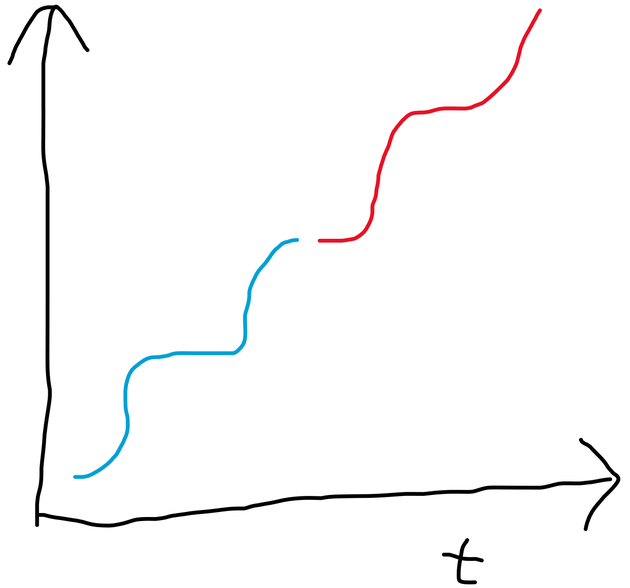
Цели, методы, качество

Как понять, хороший ли прогноз?

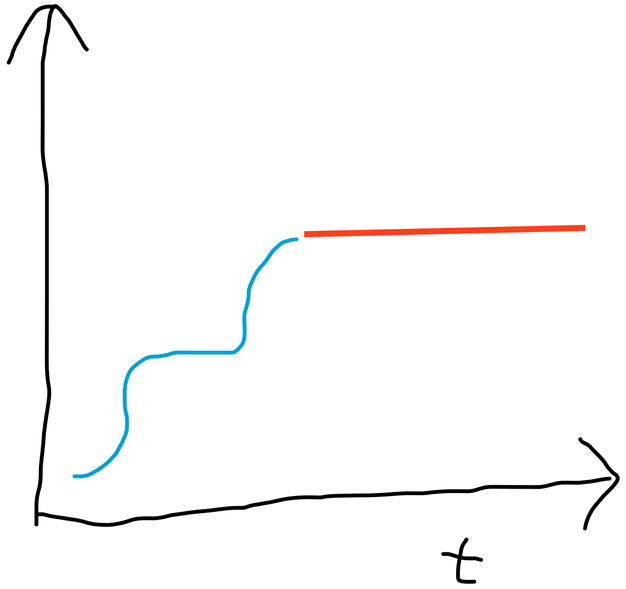
1



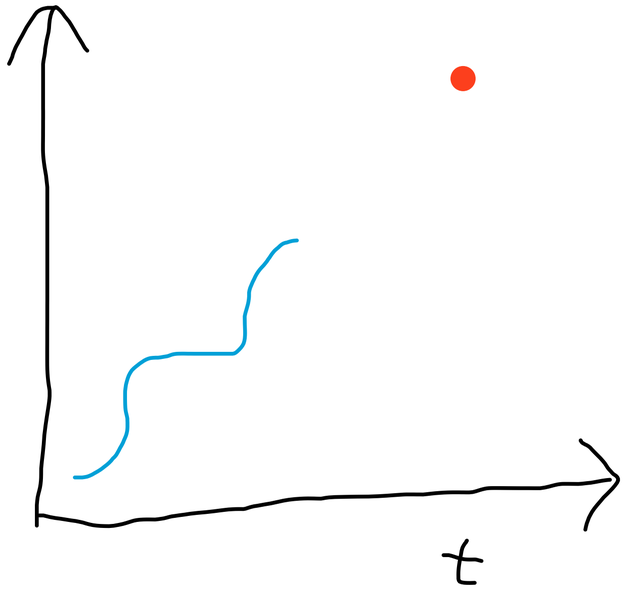
3



2



4



Цели

- Развитие продукта, новые фичи, конкурентная аналитика
- Запуск в новом городе, стране, сегменте аудитории
- Построение планов менеджерам продаж
- Постановка KPI
- Поддержка продукта: сколько заказывать серверов, сотрудников call-центра и пр.
- Исполнителю неизвестно, какая цель, просто нужно срочно сделать прогноз

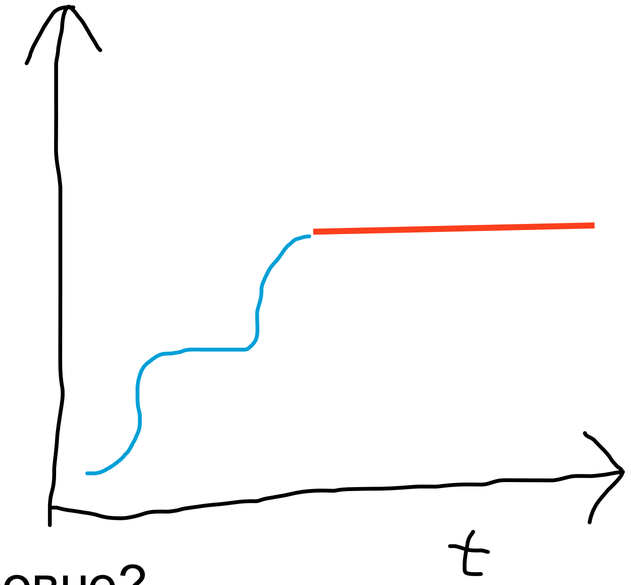
Заказчик: нужен прогноз

- Для чего? Какую задачу решает заказчик? Можно ли её решить без составления прогноза?
- Насколько срочная эта задача? Можно ли отложить или сначала показать черновой вариант,
  а позже уточнить?
- Есть ли фактические данные, или их нужно подготовить?
- Обсудить риски: сроки, степень проработки, точность

Методы прогнозирования

- Прогноз константой или линейкой
- Экспертный
- Статистический, эконометрический
  - Excel
  - Код на Python, R и пр.
  - Другой знакомый вам инструмент
- Статистический с корректировкой экспертов

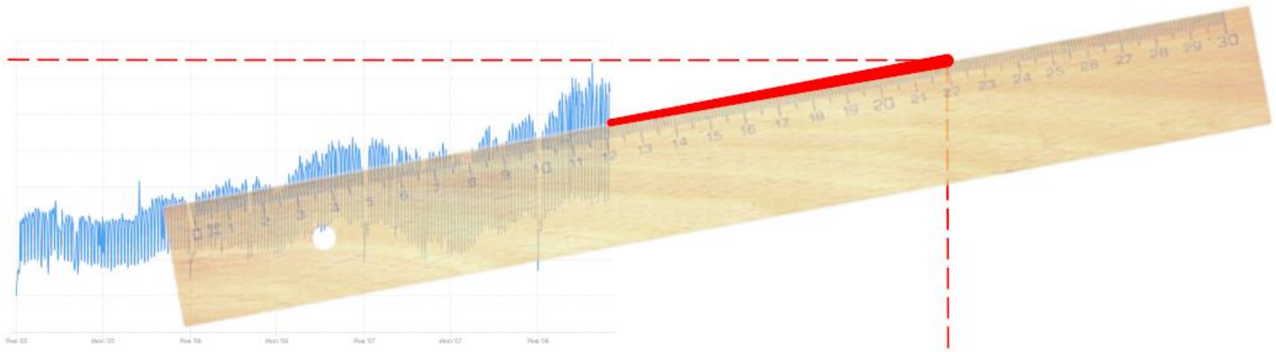
Константой или линейкой



- **Когда**: нужно прямо сейчас или вопросы простые

- **Подходит для ответов на вопросы**:
  - Линейка: куда мы трендово придём к концу года?
  - Константа: что будет, если мы останемся на текущем уровне?

- **Плюсы**: очень быстрый метод
- **Минусы**: отвечает только на ряд вопросов либо неточный



Экспертный прогноз

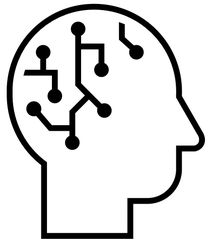
- **Когда**: нет данных, они плохие; нужно мнение или тайное знание

- **Подходит для ответов на вопросы**:
  - Что будет, если...?
  - Какая примерная оценка ...?

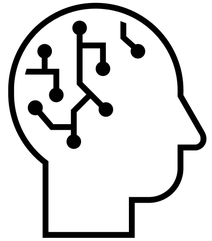
- **Плюсы**:
  - Есть оценка, когда другие методы неприменимы
  - Учтено недоступное вам знание

- **Минусы**:
  - Сильно зависит от квалификации эксперта и от степени доверия к нему
  - У разных экспертов - разная оценка
  - Может быть медленным

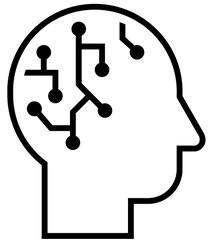
Экспертный прогноз



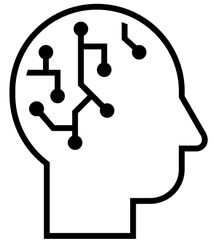
42



30



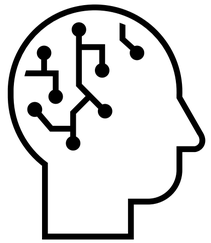
55



45

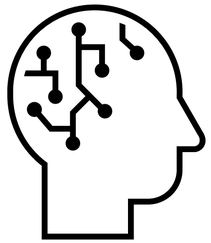
$$\frac{30 + 55 + 45}{3} = 43.33$$

Экспертный прогноз с уверенностью



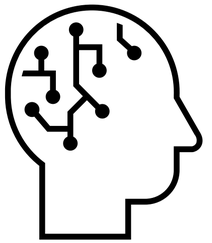
30

уверен на 70%



55

уверен на 90%



45

уверен на 80%

$$\frac{30 \cdot 0.7 + 55 \cdot 0.9 + 45 \cdot 0.8}{0.7 + 0.8 + 0.9} = 44.375$$

Статистический прогноз

- **Когда**:
  - нужна высокая точность
  - есть время, или вы отлично владеете инструментом

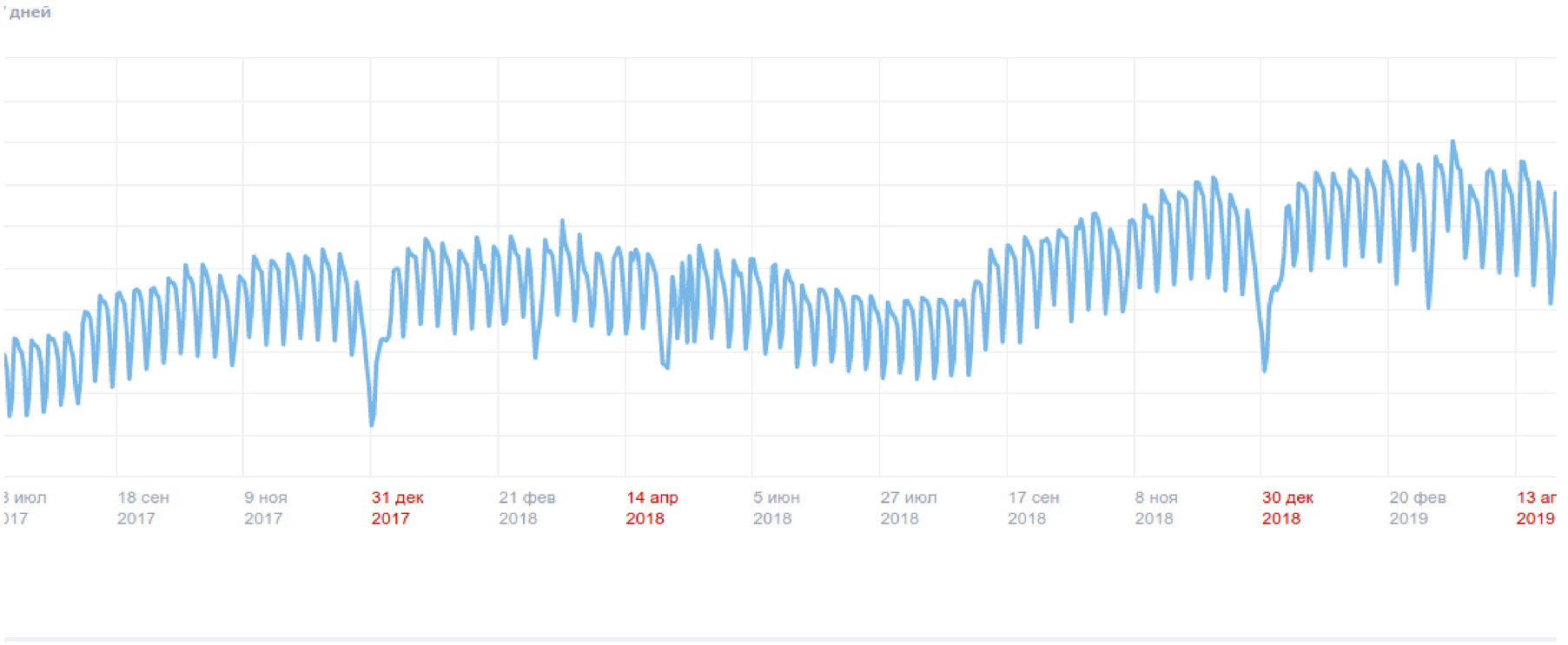
- **Подходит для ответов на вопросы (сложные)**:
  - Каким станет показатель, если мы будем работать так же хорошо, как и раньше?
  - Каким станет показатель на дату X?
  - Какое вероятностное распределение показателя на дату X?

- **Плюсы**: точный, если сделать правильно
- **Минусы**: высокий порог вхождения, если не знакомы с Excel или программированием

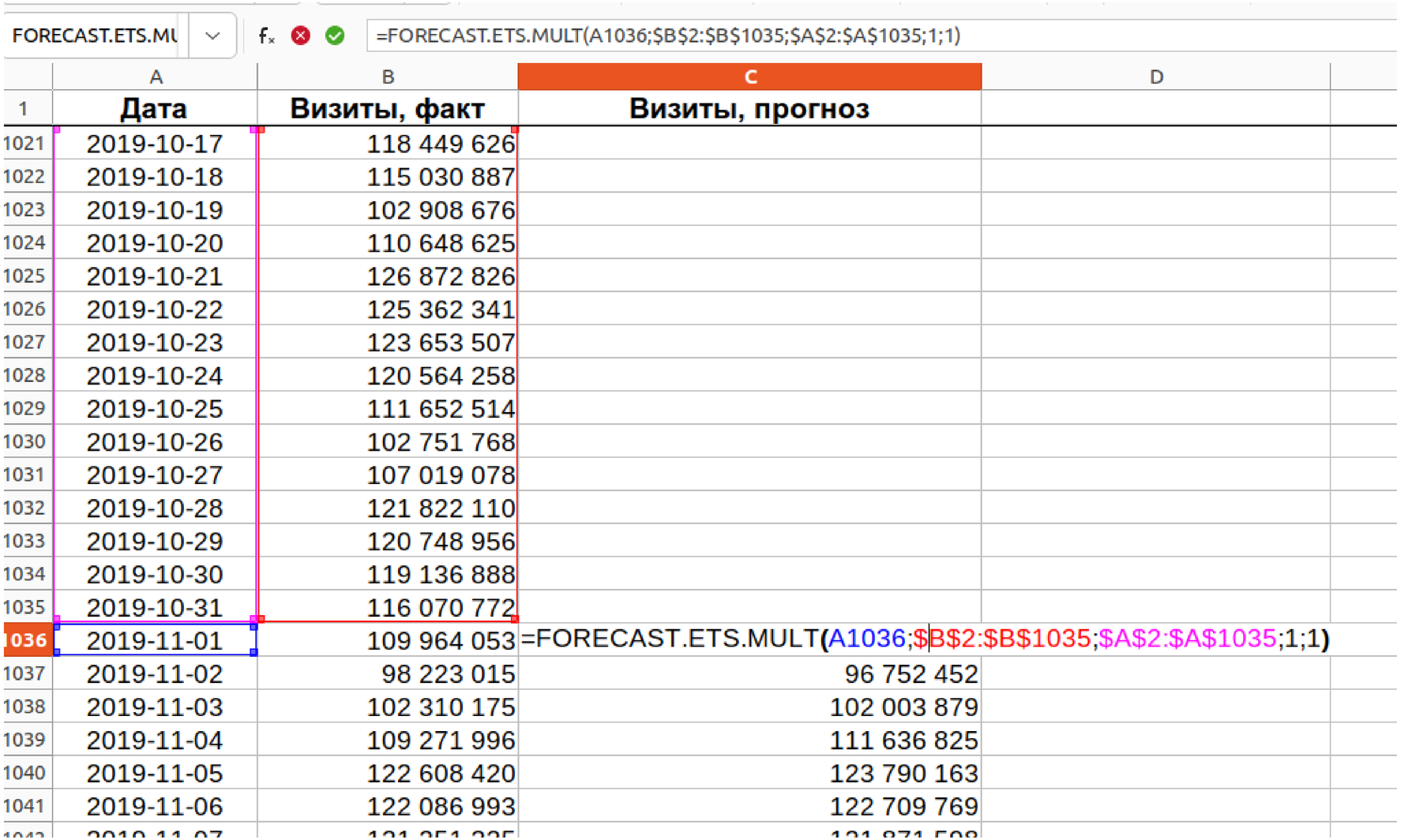
Яндекс Радар

Поисковые системы в России

1 янв 2017 — 31 окт 2019 | Все типы устройств | Все платформы



Прогноз в Excel

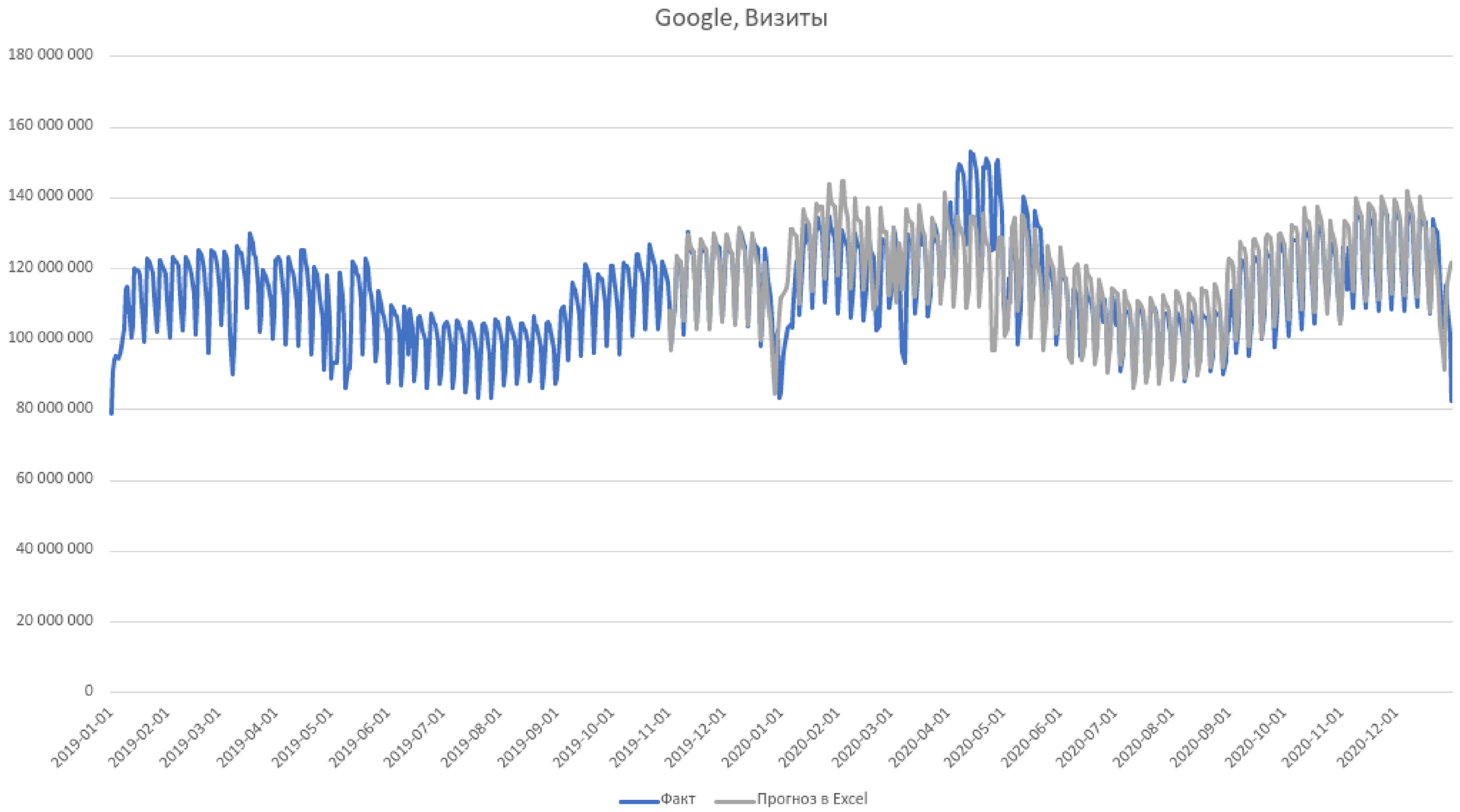


	A	B	C	D
1	Дата	Визиты, факт	Визиты, прогноз	
1021	2019-10-17	118 449 626		
1022	2019-10-18	115 030 887		
1023	2019-10-19	102 908 676		
1024	2019-10-20	110 648 625		
1025	2019-10-21	126 872 826		
1026	2019-10-22	125 362 341		
1027	2019-10-23	123 653 507		
1028	2019-10-24	120 564 258		
1029	2019-10-25	111 652 514		
1030	2019-10-26	102 751 768		
1031	2019-10-27	107 019 078		
1032	2019-10-28	121 822 110		
1033	2019-10-29	120 748 956		
1034	2019-10-30	119 136 888		
1035	2019-10-31	116 070 772		
1036	2019-11-01	109 964 053	=FORECAST.ETS.MULT(A1036;$B$2:$B$1035;$A$2:$A$1035;1;1)	
1037	2019-11-02	98 223 015	96 752 452	
1038	2019-11-03	102 310 175	102 003 879	
1039	2019-11-04	109 271 996	111 636 825	
1040	2019-11-05	122 608 420	123 790 163	
1041	2019-11-06	122 086 993	122 709 769	

https://support.microsoft.com/en-us/office/forecast-ets-function-15389b8b-677e-4fbd-bd95-21d464333f41

FORECAST.ETS.MULT(новая дата;значения;даты;1;1)

Прогноз в Excel: результат



Прогноз в Python + Facebook Prophet

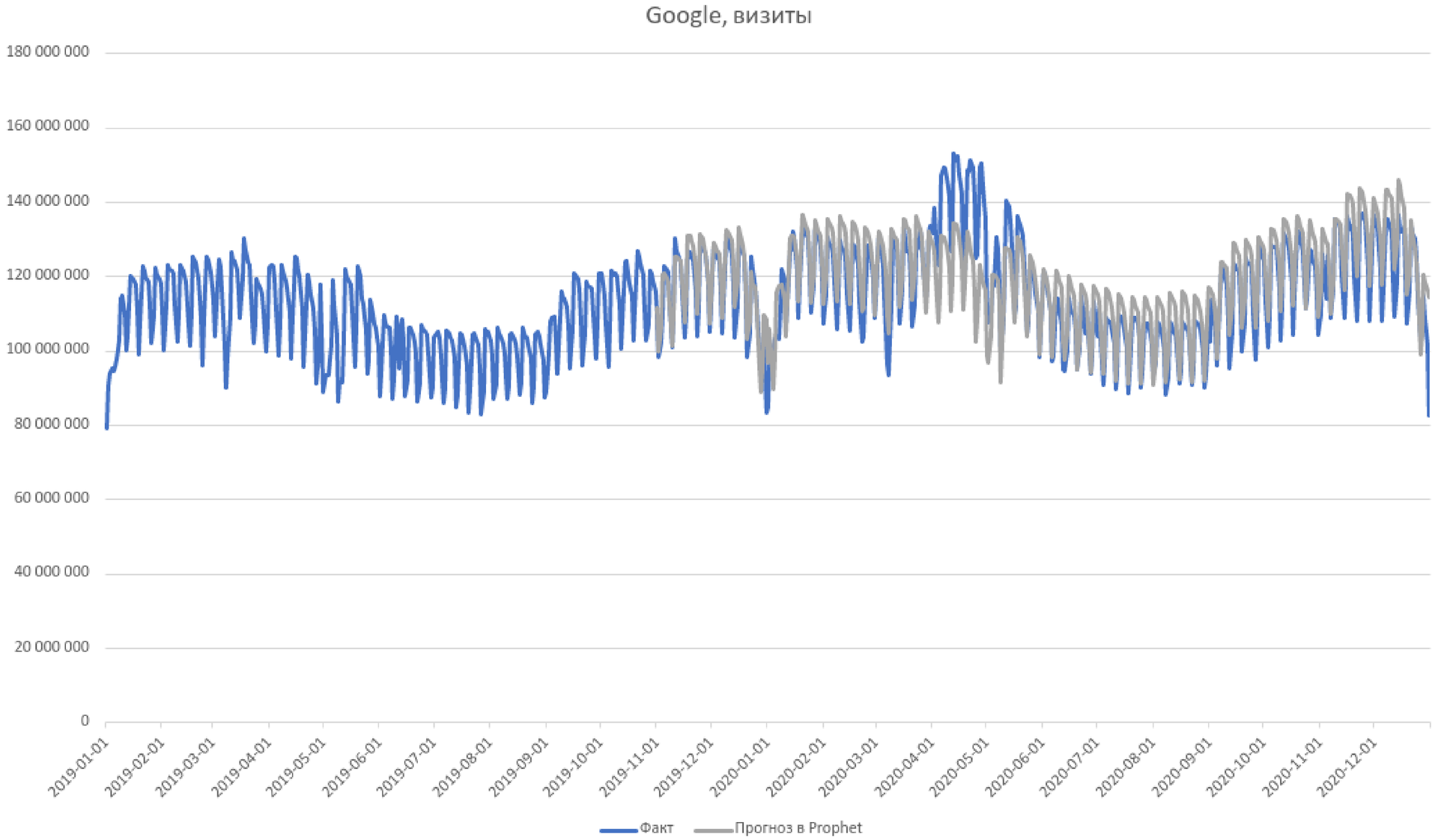
```
import pandas as pd
from prophet import Prophet

df = pd.read_csv('google_visits.csv')

model = Prophet(seasonality_mode='multiplicative',
                holidays_prior_scale=0.02, yearly_seasonality=15)
model.add_country_holidays(country_name='RU')
model.fit(df)

future_df = model.make_future_dataframe(periods=427, include_history=False)
forecast = model.predict(future_df)
```

Прогноз в Python + Facebook Prophet



Сравнение

Данные	Сумма за декабрь 2020	Разница	Ошибка в %
Факт	3 763 млн		
Прогноз в Excel	3 813 млн	50 млн	1.34%
Прогноз в Prophet	3 991 млн	229 млн	6.08%

Качество

- ML-разработчик: RMSE, log loss, Accuracy, F1 score, ROC-AUC
- Аналитик: отклонение между фактом и прогнозом в условных единицах или процентах
- Менеджер: помог ли прогноз принимать решения?
- Бизнес: оказались ли правильными решения, основанные на прогнозе?

Качество

Ситуация

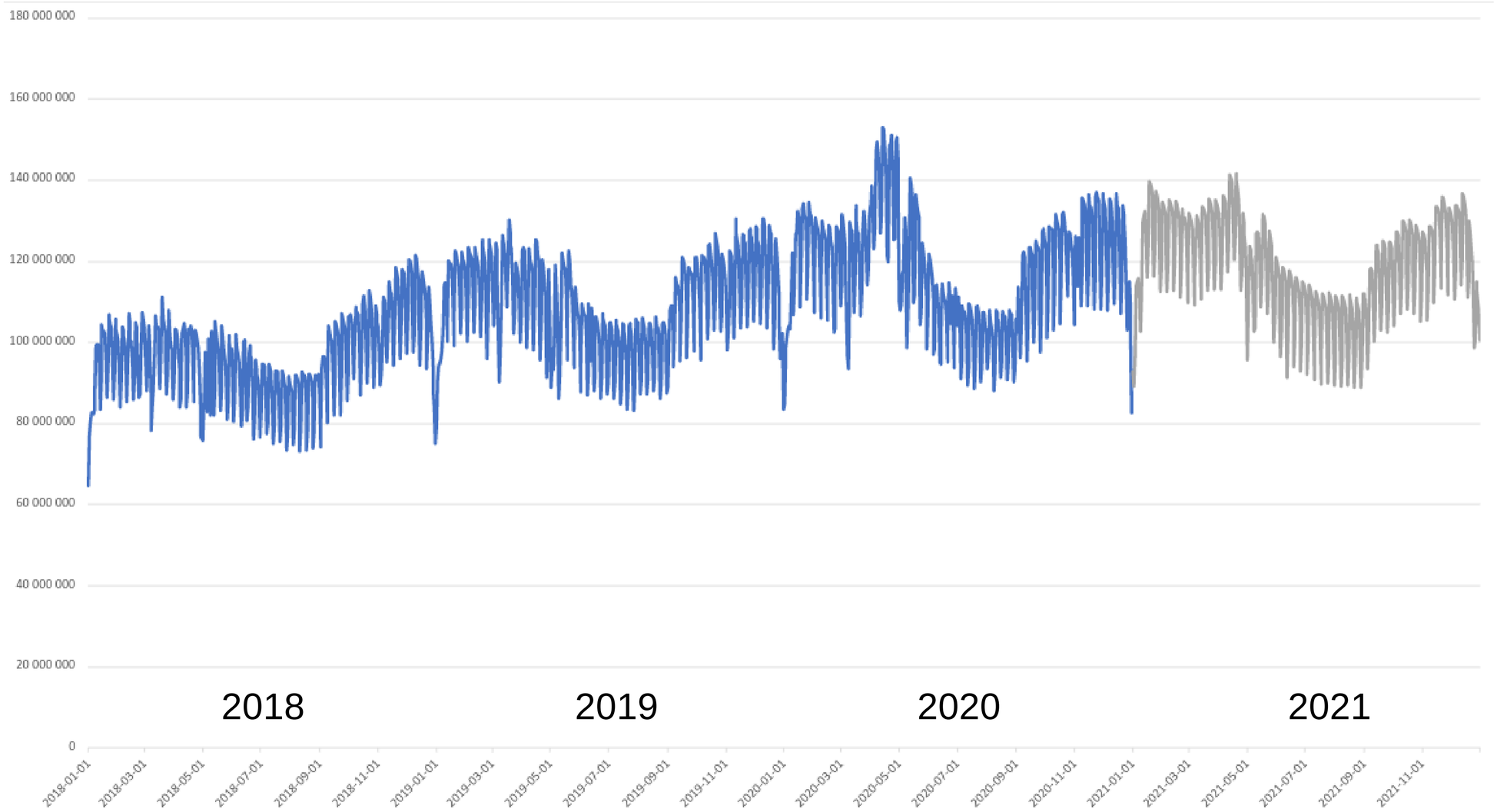
- Младший аналитик сделал прогноз выручки, который затем попал в факт. Получил плюс в конце квартала
- Старший аналитик сделал прогноз выручки, который затем попал в факт. Получил минус в конце квартала

03

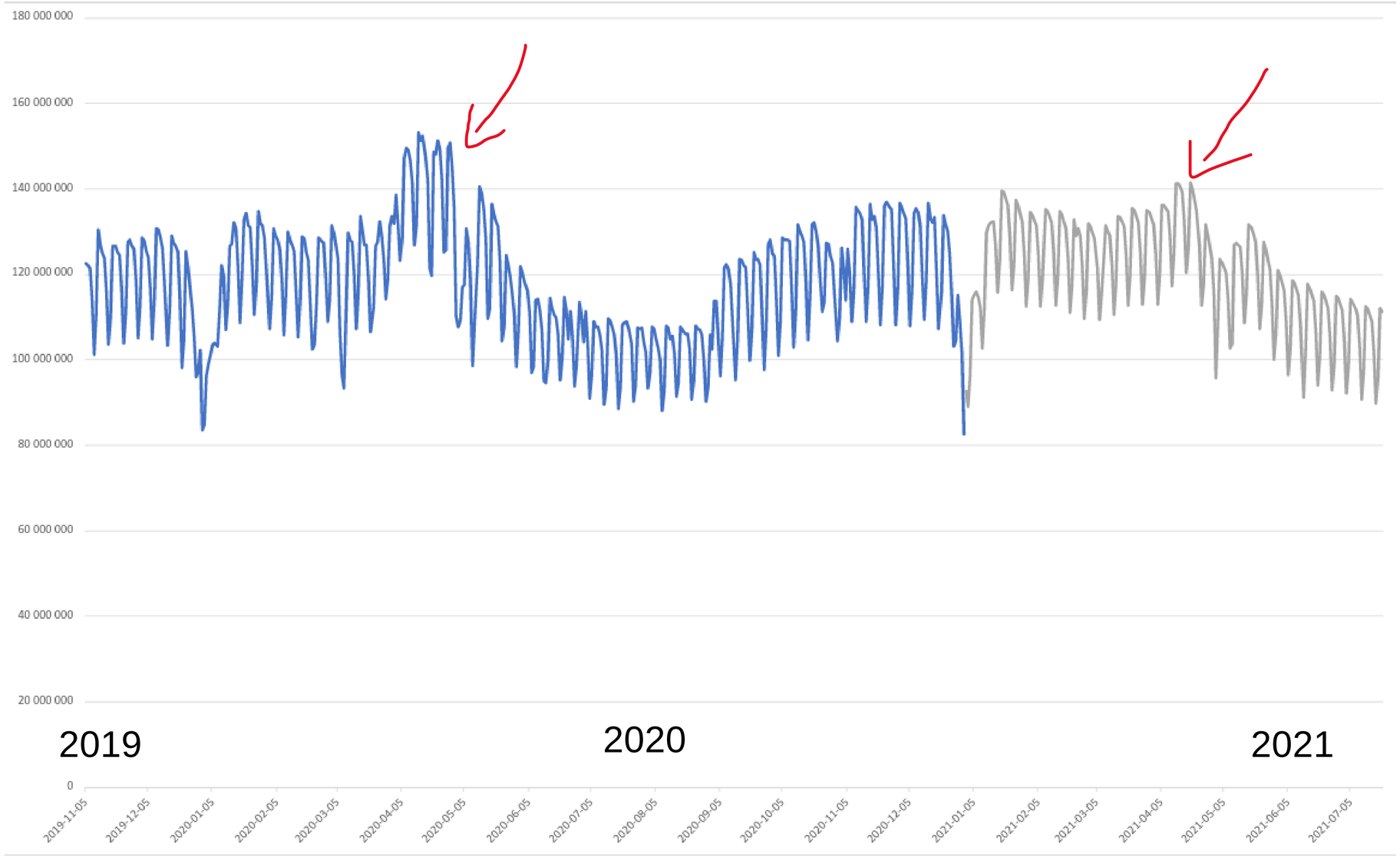
Очистка данных

Garbage in, garbage out

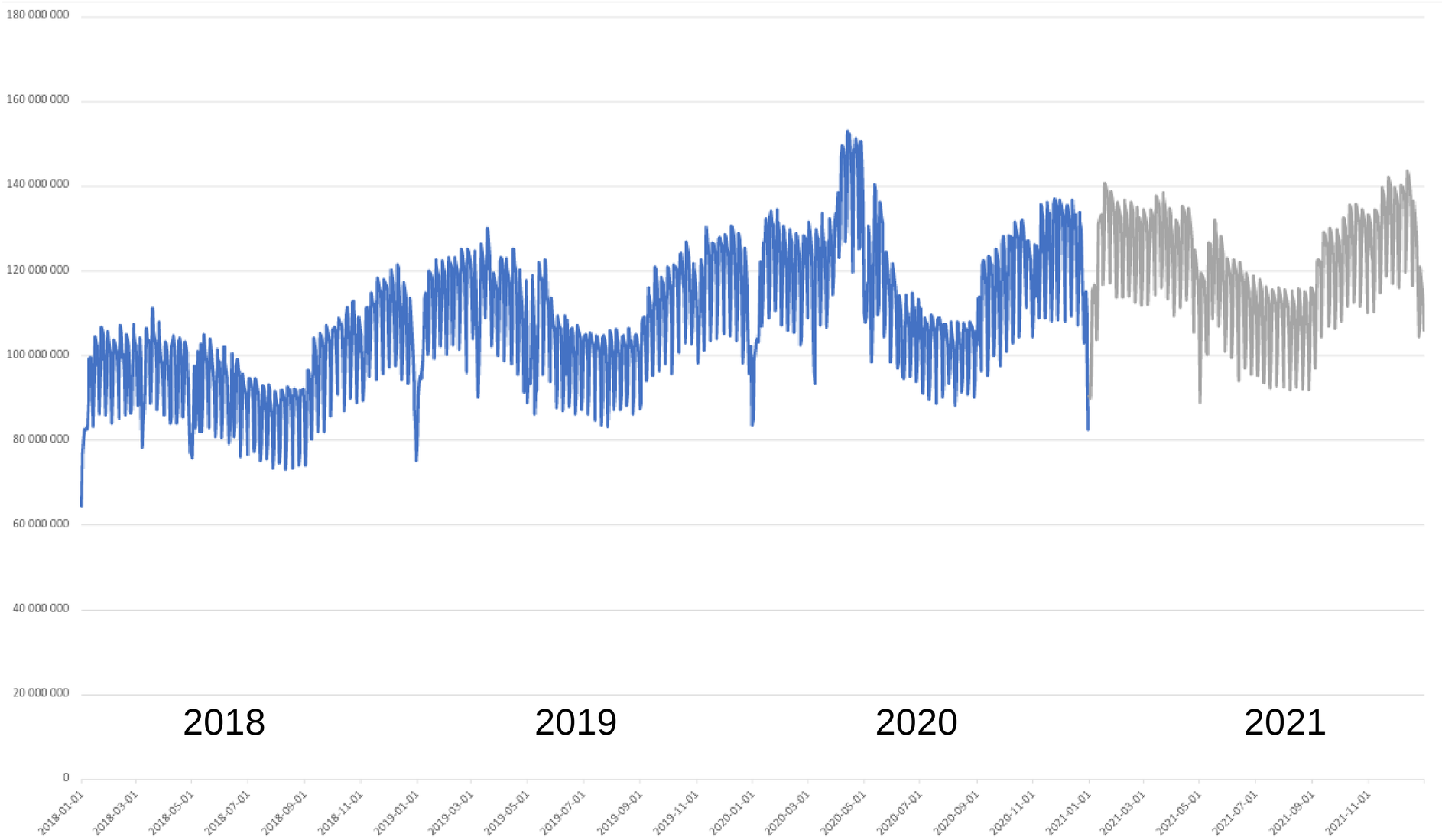
Прогноз визитов на 2021 год



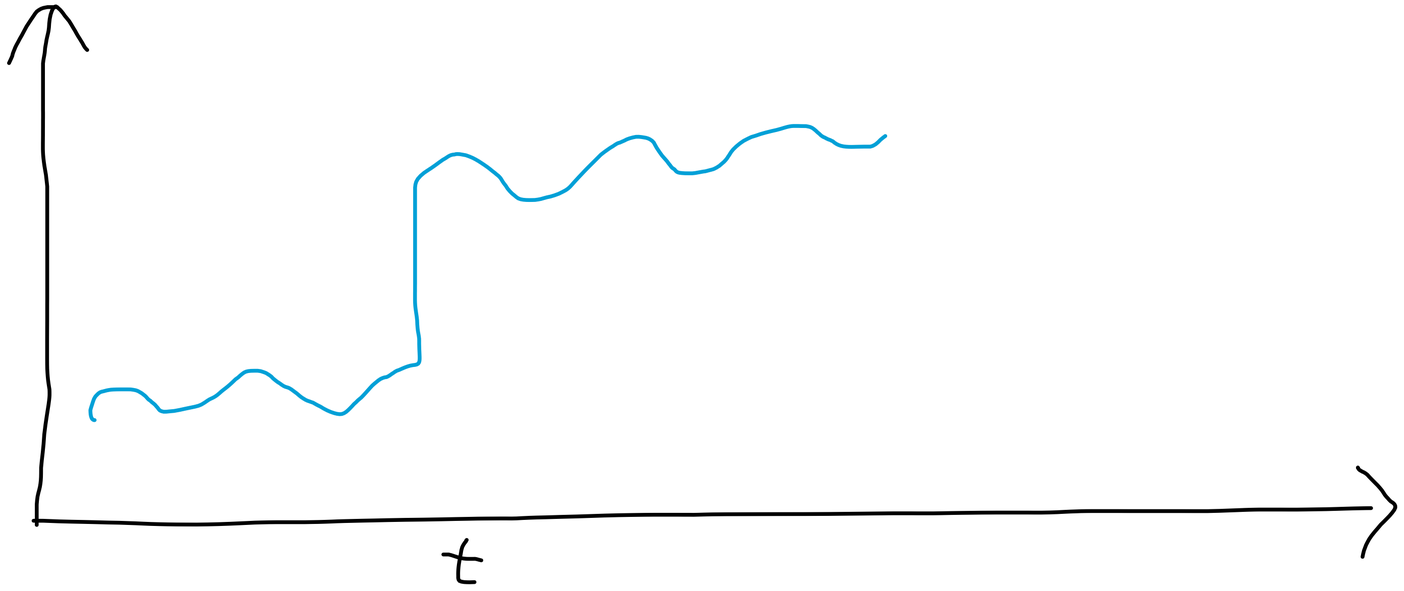
Прогноз визитов на 2021 год



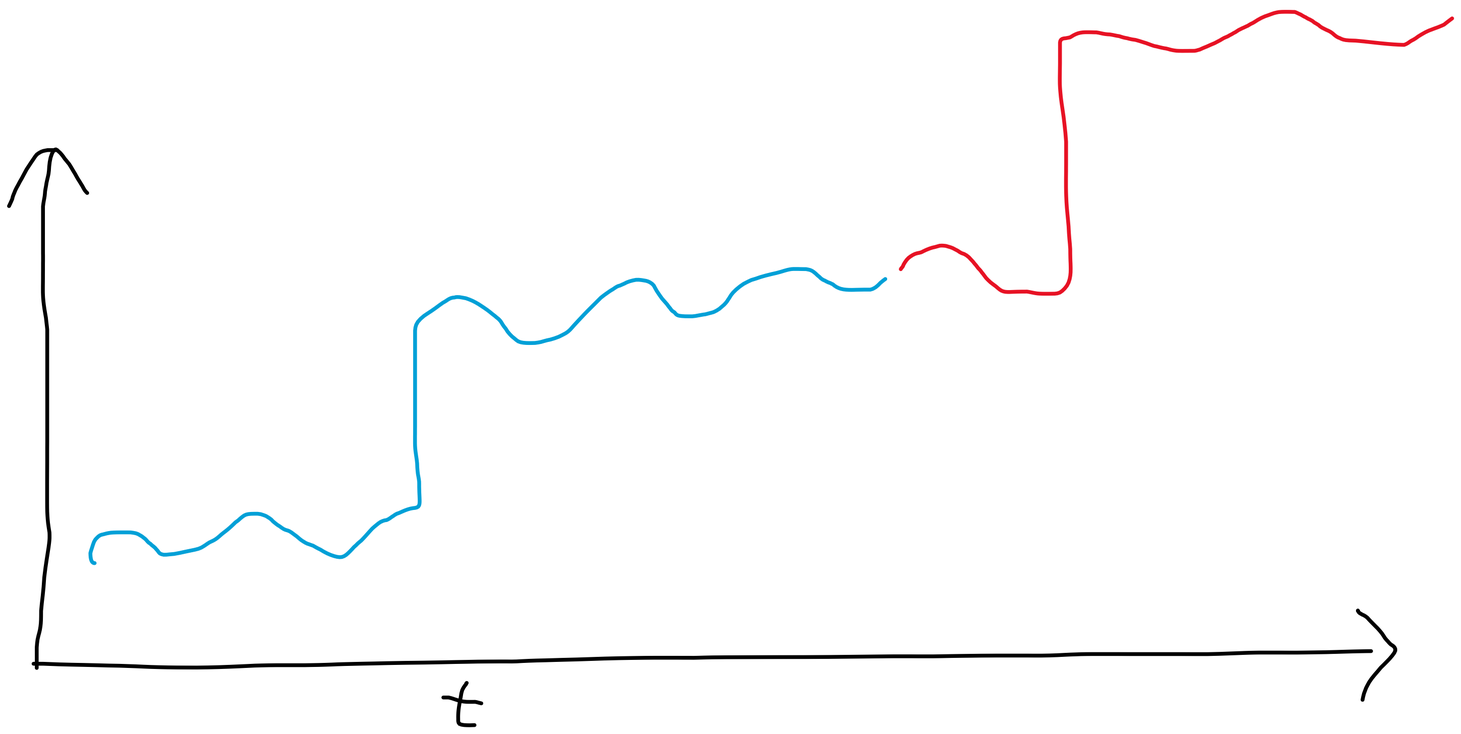
Прогноз визитов на 2021 год



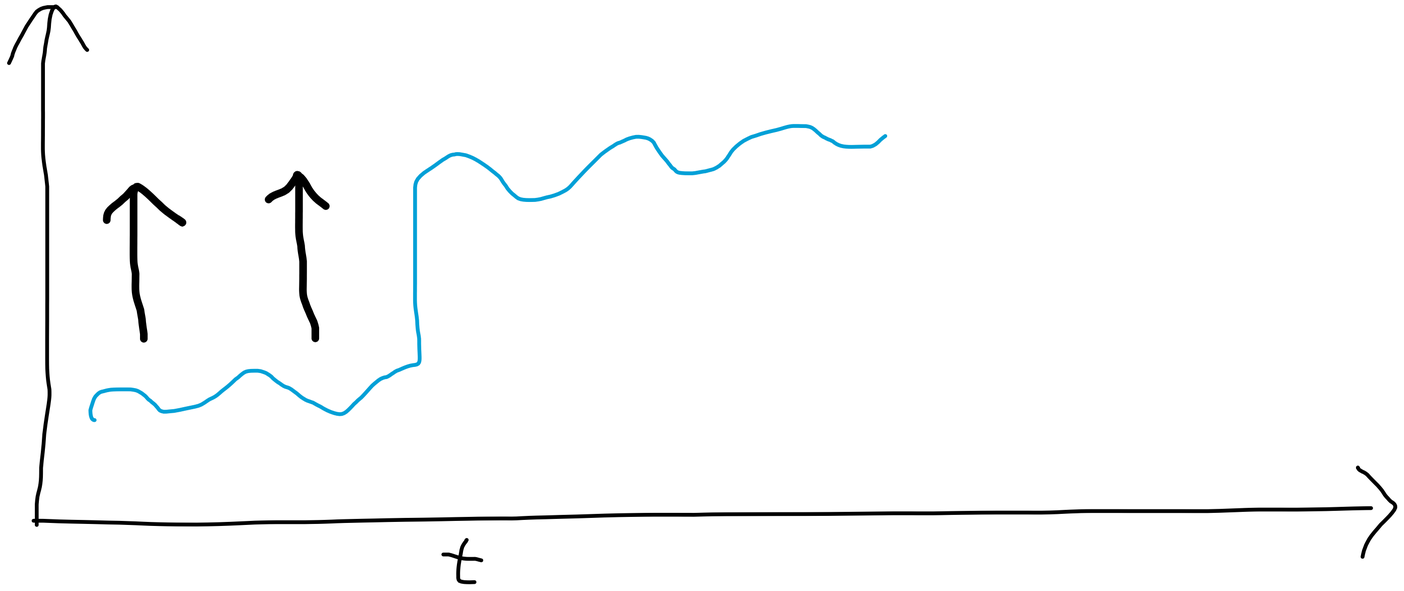
"Полочки", level change, trend shift



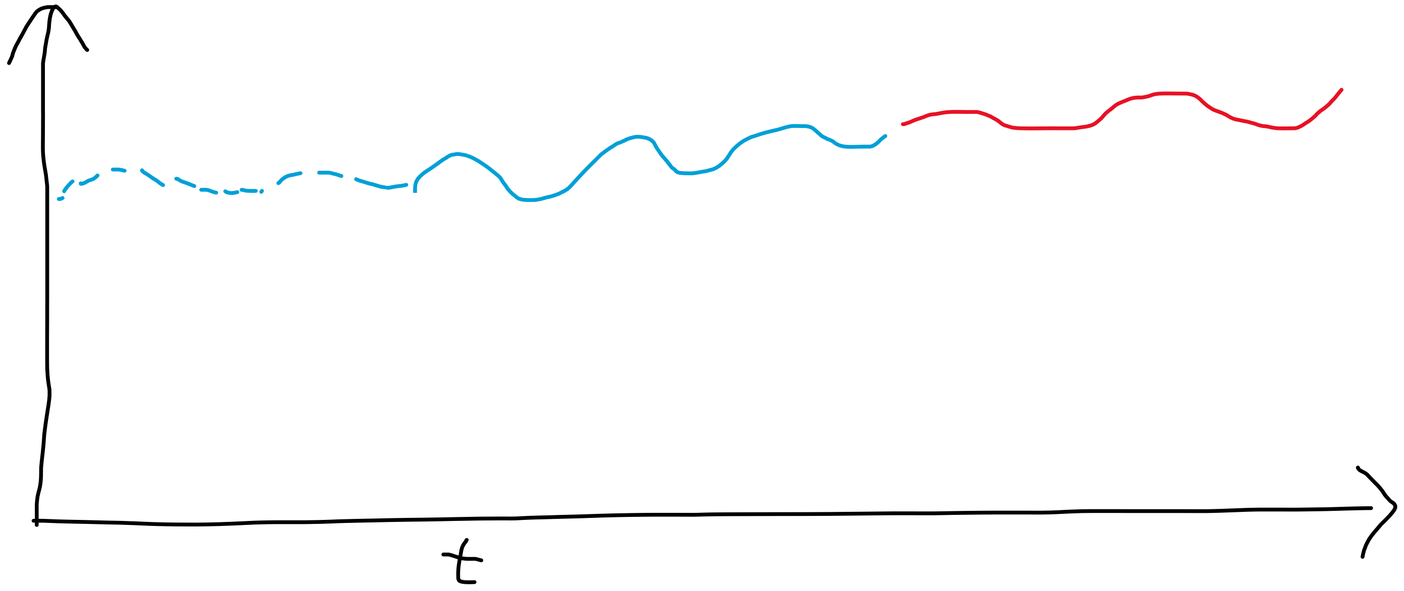
"Полочки", level change, trend shift



"Полочки", level change, trend shift



"Полочки", level change, trend shift



Аномалии в данных

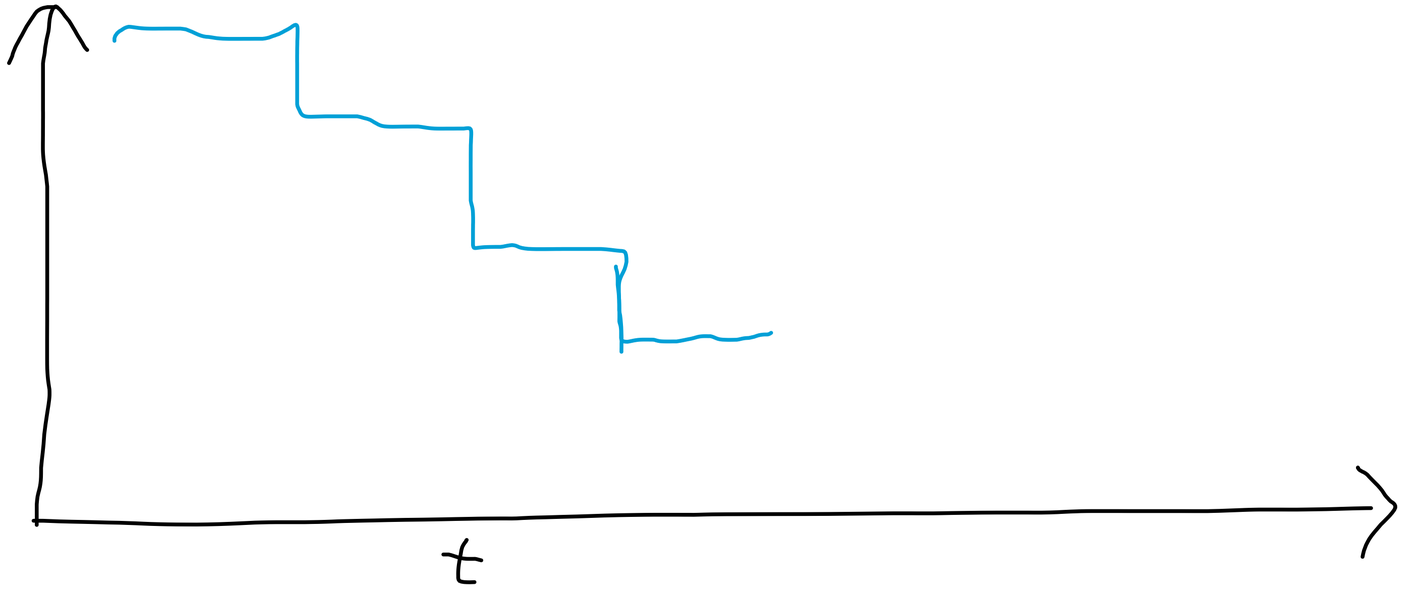
- Вырезать как можно меньше данных
- Не путать с праздниками и сезонностью
- Аномальные периоды можно вырезать или интерполировать
- Последние точки могут быть "новой реальностью", если причина продуктовая
- Последние точки могут быть некорректными, если причина техническая (проблемы с логированием), тогда от них не стоит строить прогноз

Задача - сконструировать такой синтетический временной ряд, чтобы модель смогла корректно учесть тренды и сезонность

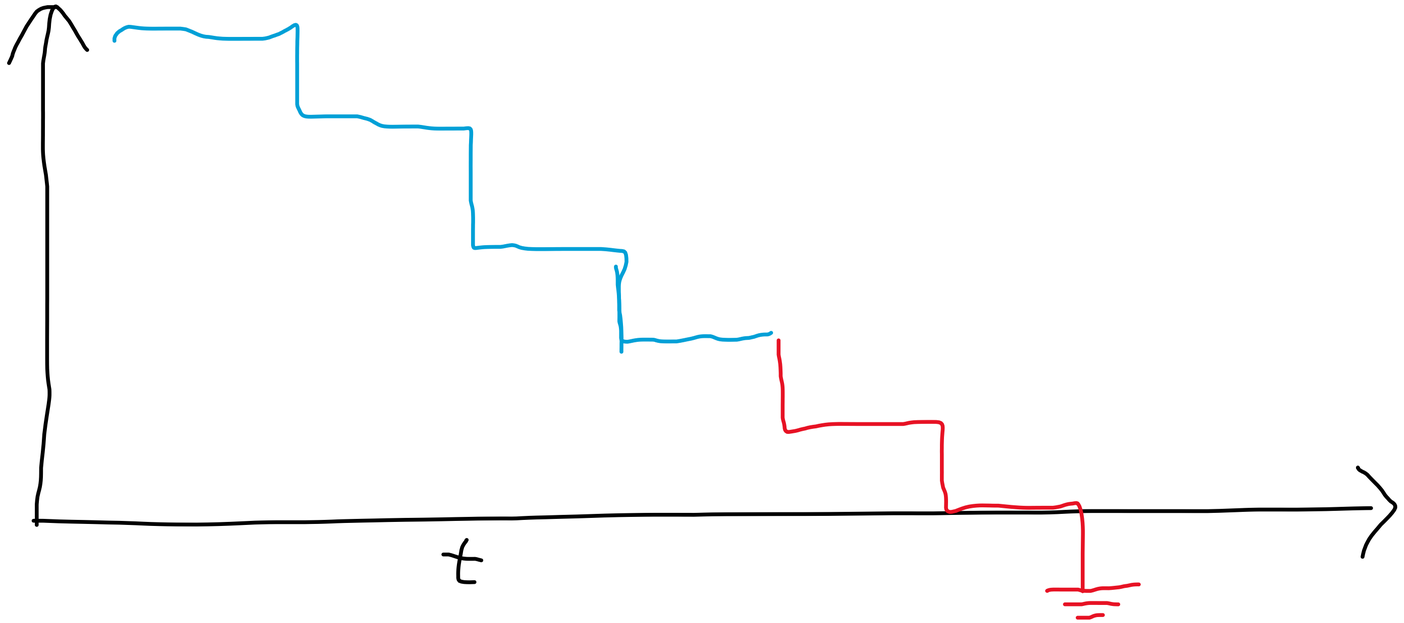
04

Декомпозиция

Постройте прогноз

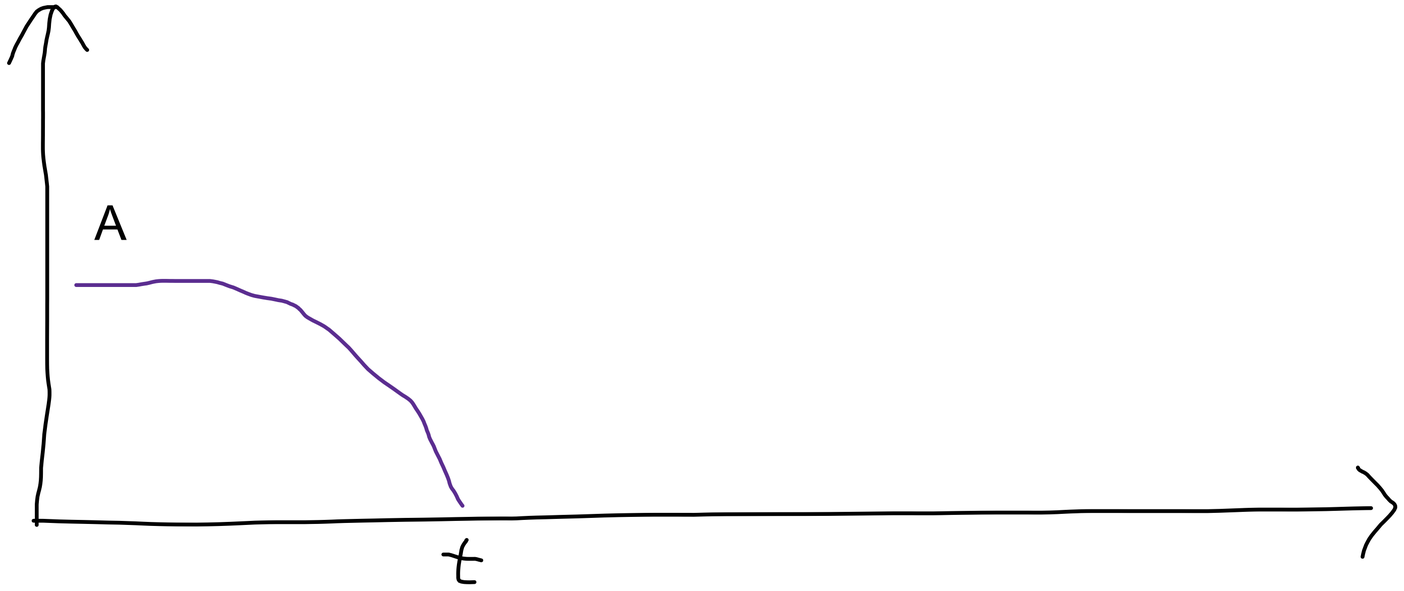


Постройте прогноз



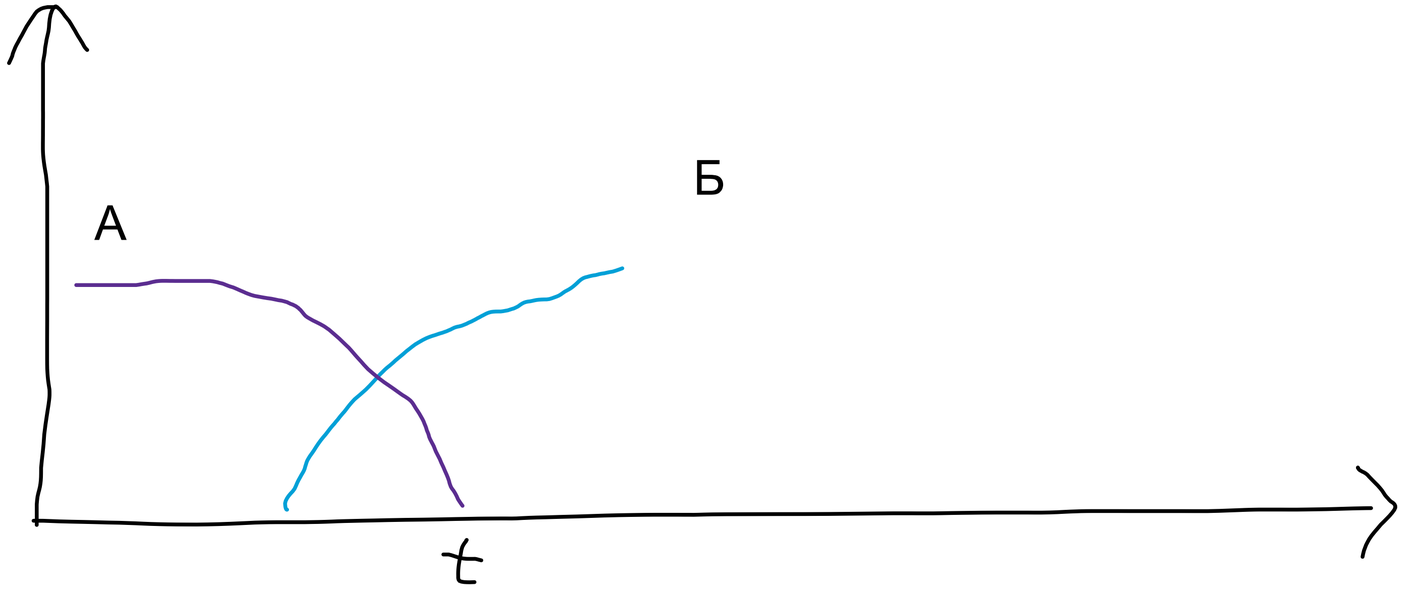
Декомпозиция для повышения точности

- A - старое десктопное приложение

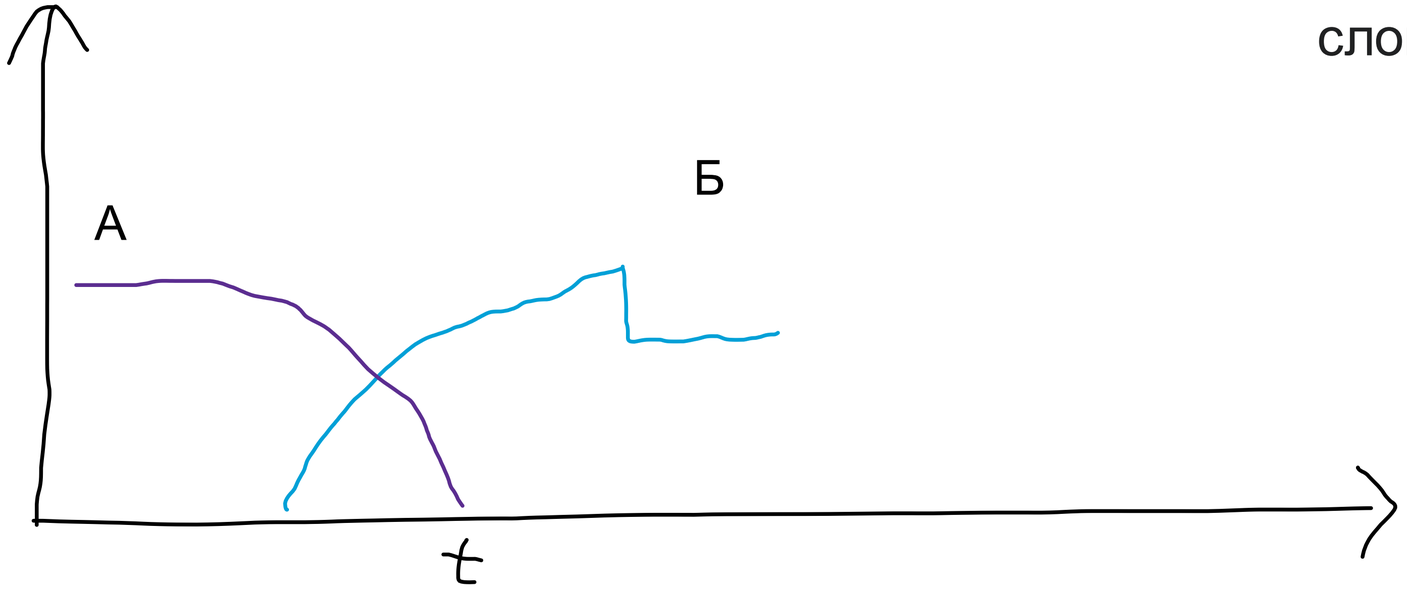


Декомпозиция для повышения точности

- А - старое десктопное приложение
- Б - новое десктопное приложение



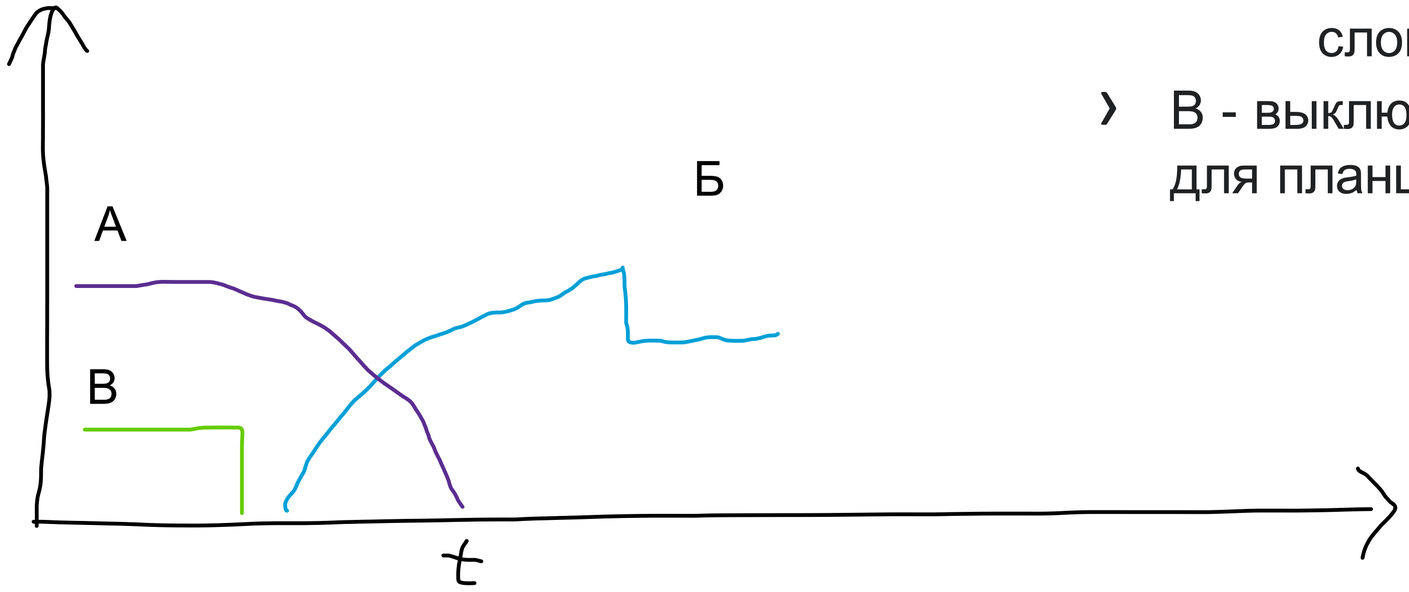
Декомпозиция для повышения точности



- А - старое десктопное приложение
- Б - новое десктопное приложение

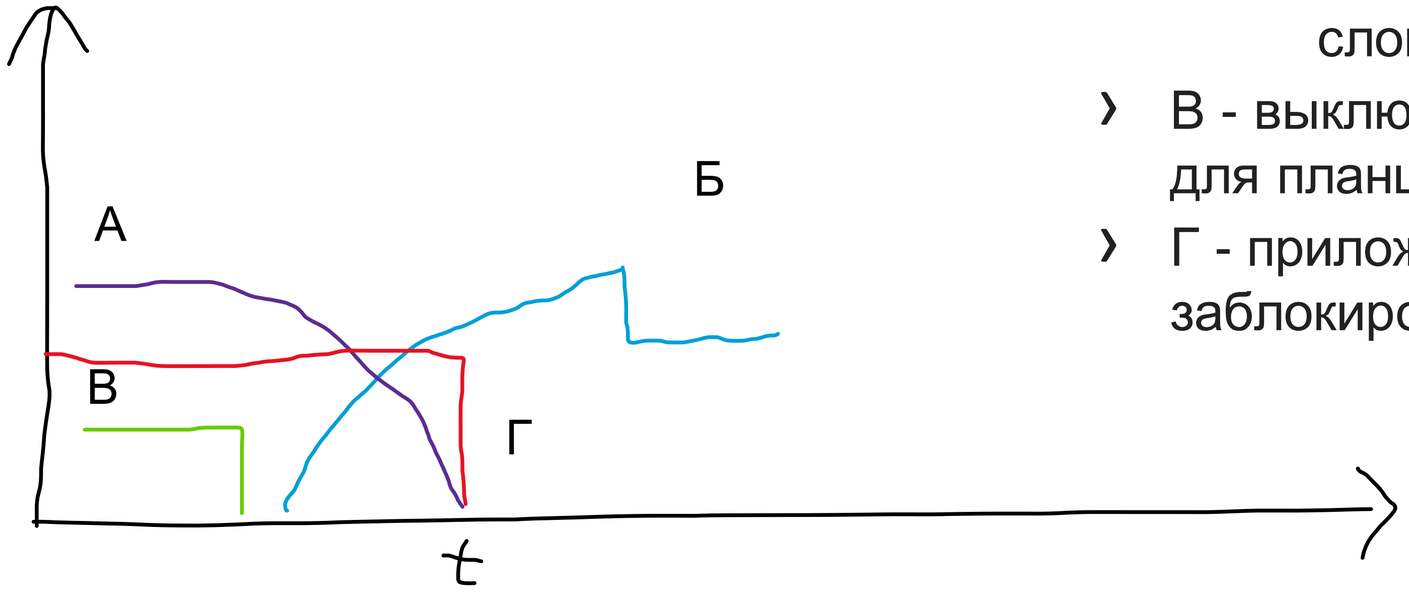
  сломали логирование!

Декомпозиция для повышения точности



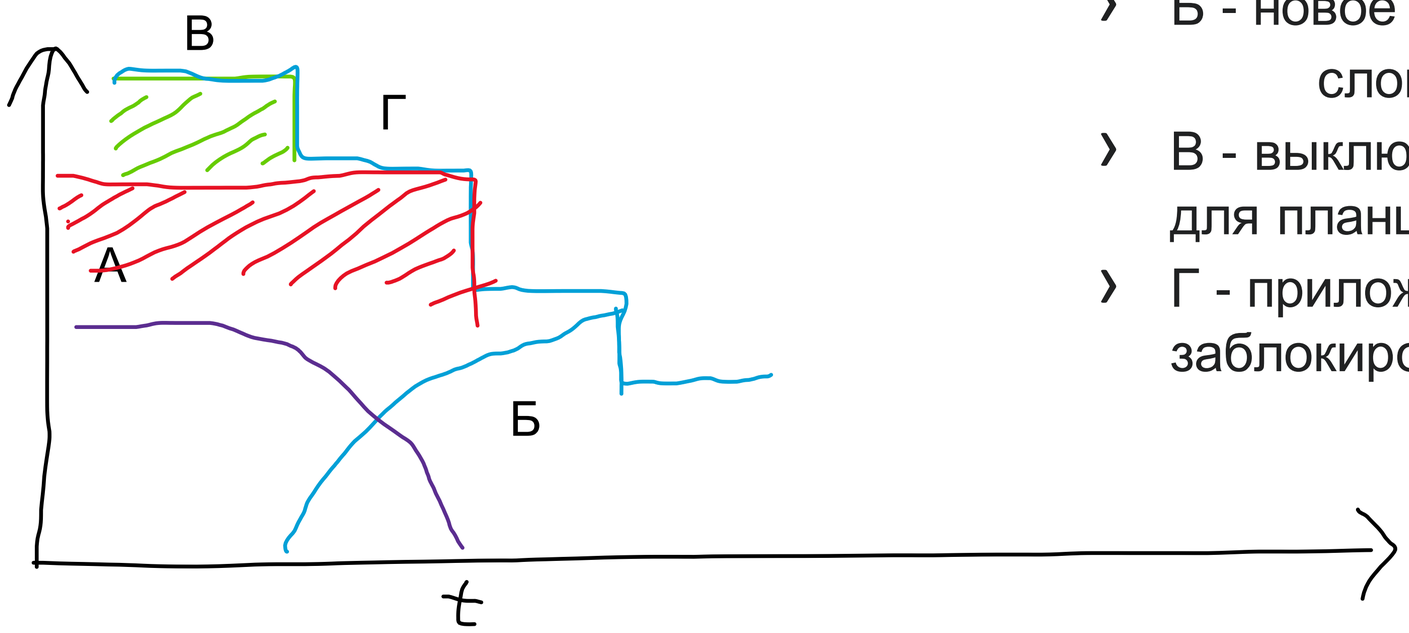
- А - старое десктопное приложение
- Б - новое десктопное приложение
  сломали логирование!
- В - выключили старое приложение для планшетов

Декомпозиция для повышения точности



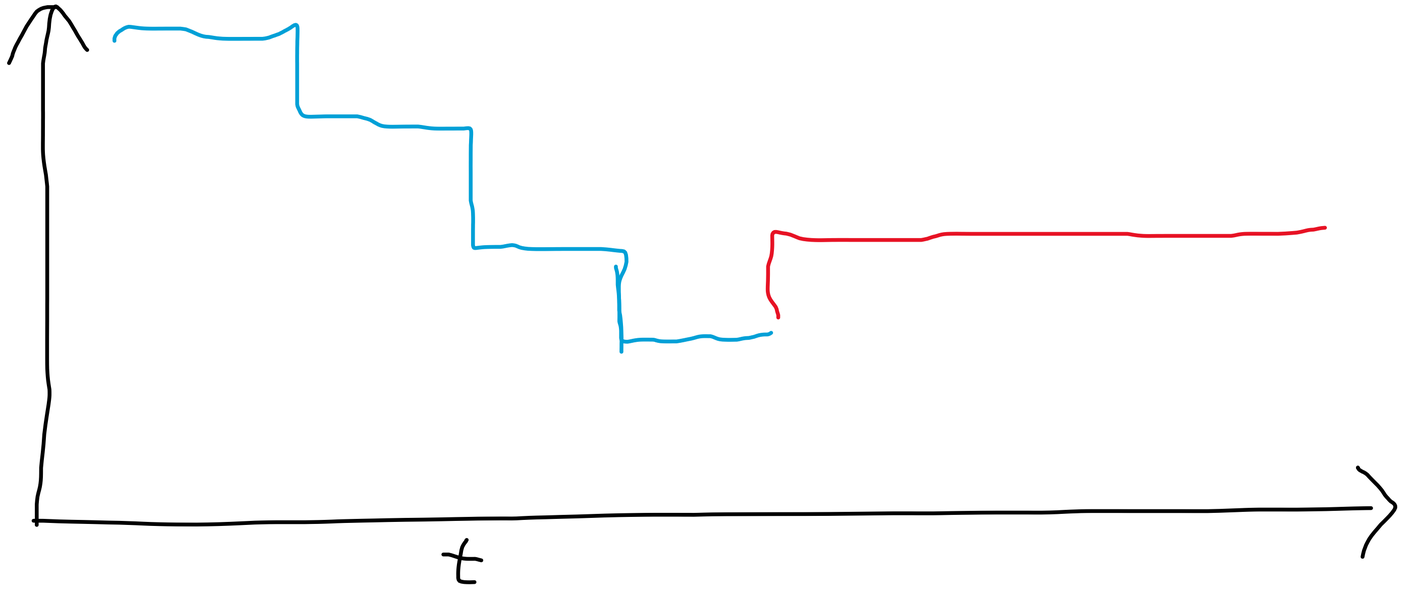
- А - старое десктопное приложение
- Б - новое десктопное приложение
  сломали логирование!
- В - выключили старое приложение для планшетов
- Г - приложение под iOS было заблокировано

Декомпозиция для повышения точности



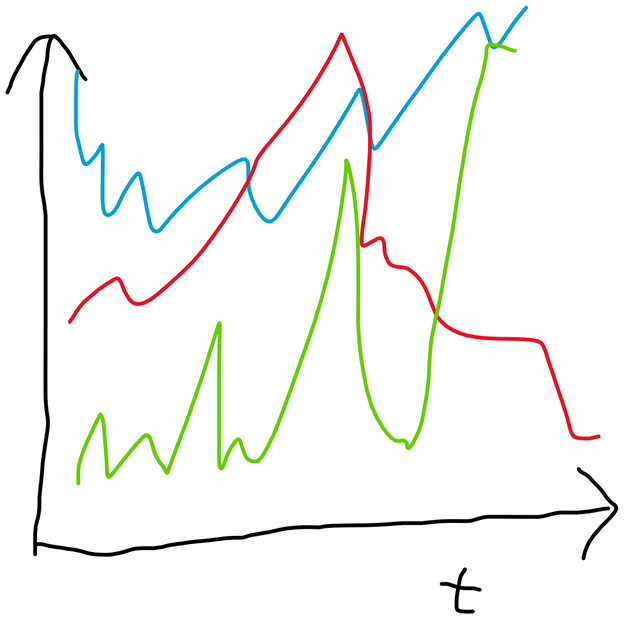
- А - старое десктопное приложение
- Б - новое десктопное приложение
  сломали логирование!
- В - выключили старое приложение для планшетов
- Г - приложение под iOS было заблокировано

Итоговый прогноз

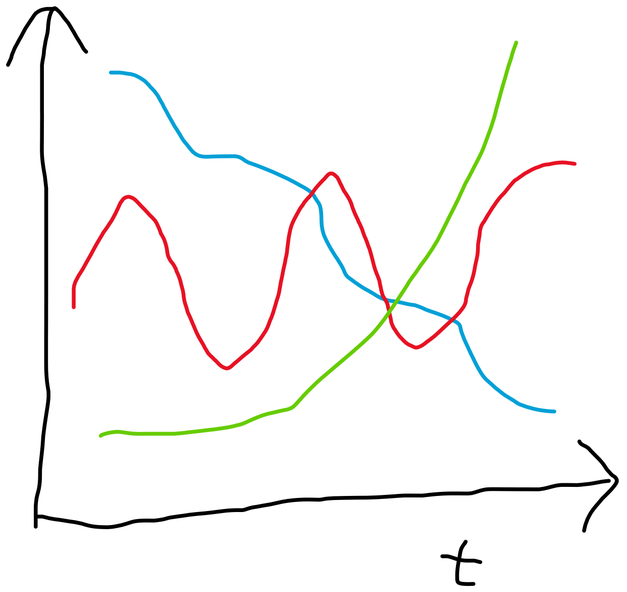


Декомпозиция для удобства

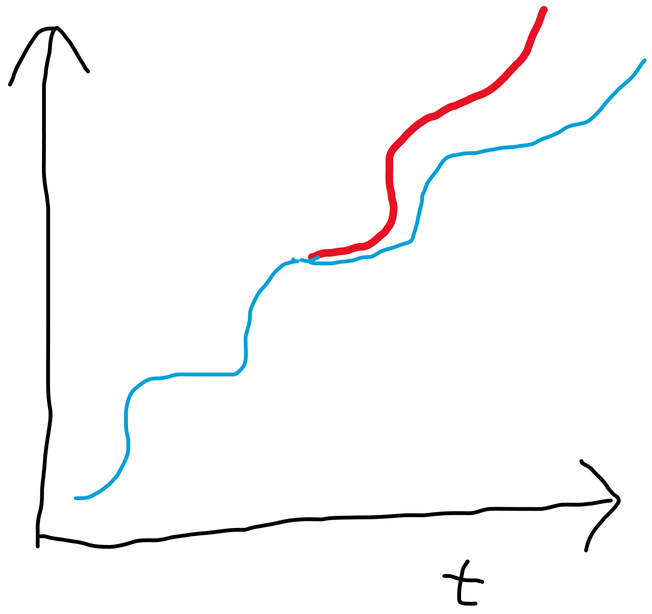
Как хочет заказчик



Как мне удобно

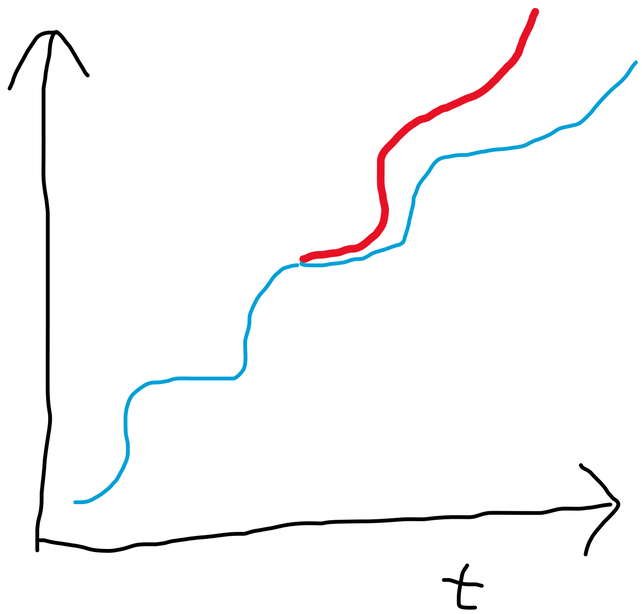


Почему прогноз не попал в факт?

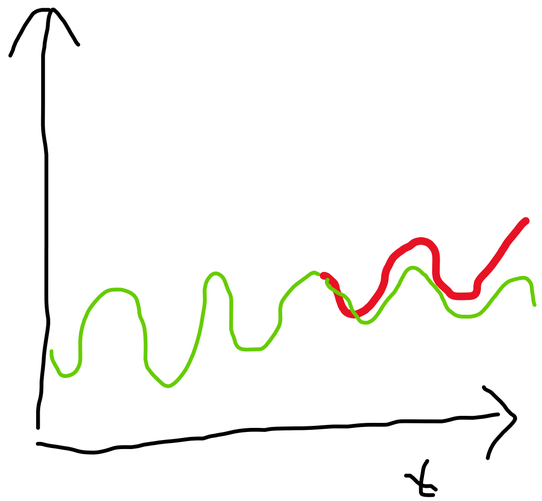


Выручка

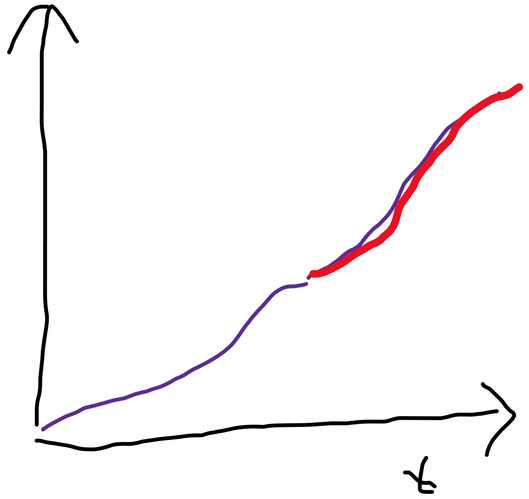
Декомпозиция для анализа ошибки



=



X



Выручка

DAU

ARPU

05

Вызовы

Трендовые модели перестают работать

- Кризисы
- Проблемы с логированием, новые аномалии
- Запускаются новые фичи сервиса
- Отключаются части сервиса, уходят клиенты
- Новые требования бизнеса, например, новые разбивки на сегменты

06

Заключение

Что запомнить

- Прогнозирование – это интересно!

- Уточнять требования заказчика, узнавать, для чего нужен прогноз
- Обсуждать риски, идти на компромиссы
- Не использовать сложные методы, когда это не нужно
- Делать предварительный анализ данных, декомпозировать
- Очищать данные от аномалий и полочек
- Отслеживать качество прогноза, корректировать модель

Что дальше

- Почитать про автоматические методы очистки данных
- Попробовать сделать прогноз с помощью линейной регрессии
- Попробовать библиотеки для прогнозирования: Darts, Kats, Merlion
- Познакомиться с Multivariate forecasting
- Разобраться с прогнозированием в Excel
- Узнать, что такое автокорреляция, анализ остатков и кросс-валидация модели

Стажировка

https://yandex.ru/yaintern/

- Стажировка в течение года / летняя и Deep dive от Яндекс Маркета
- 3, 4 или 6 месяцев

Вопросы