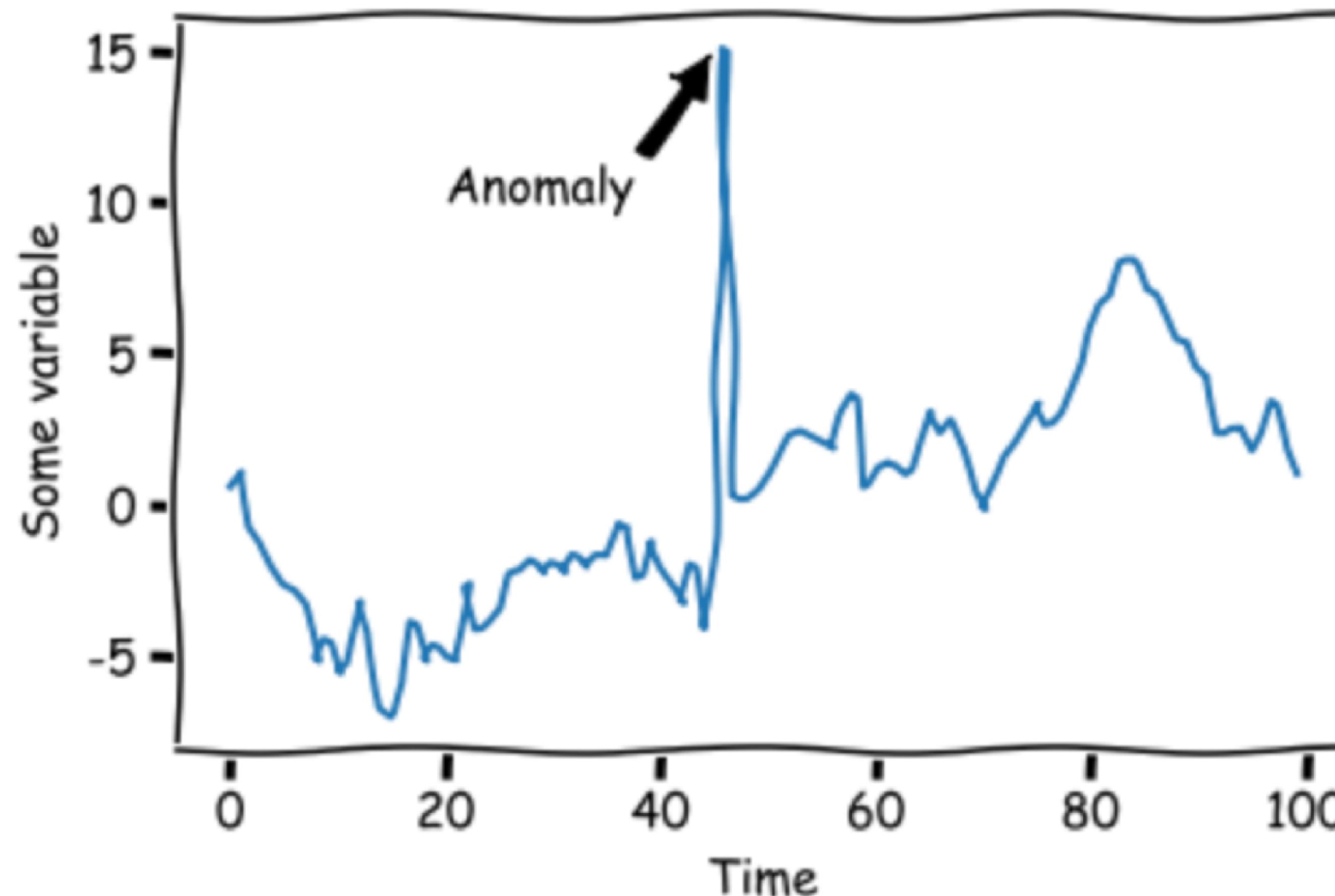


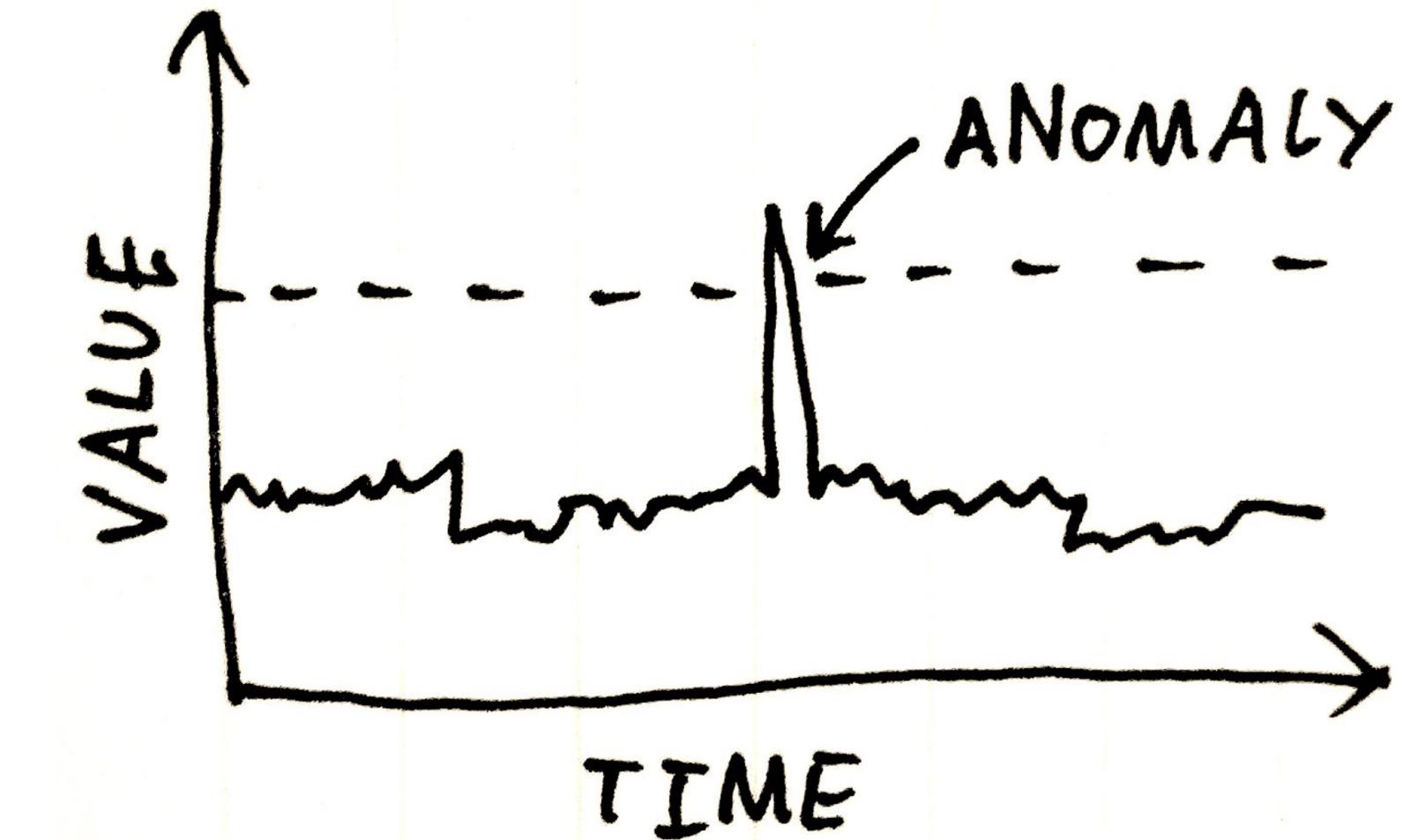
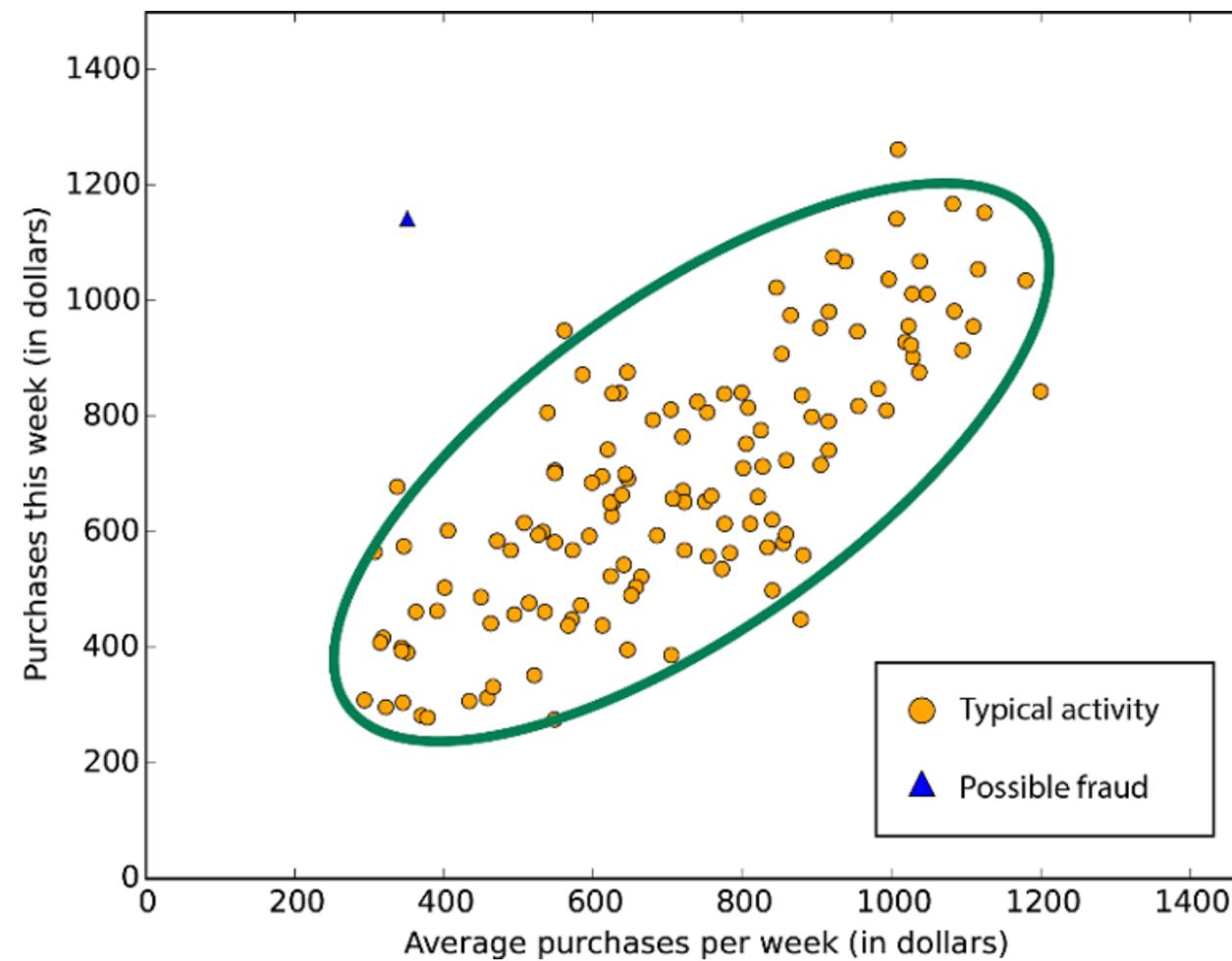
ОТКРИВАНЕ НА АНОМАЛИИ



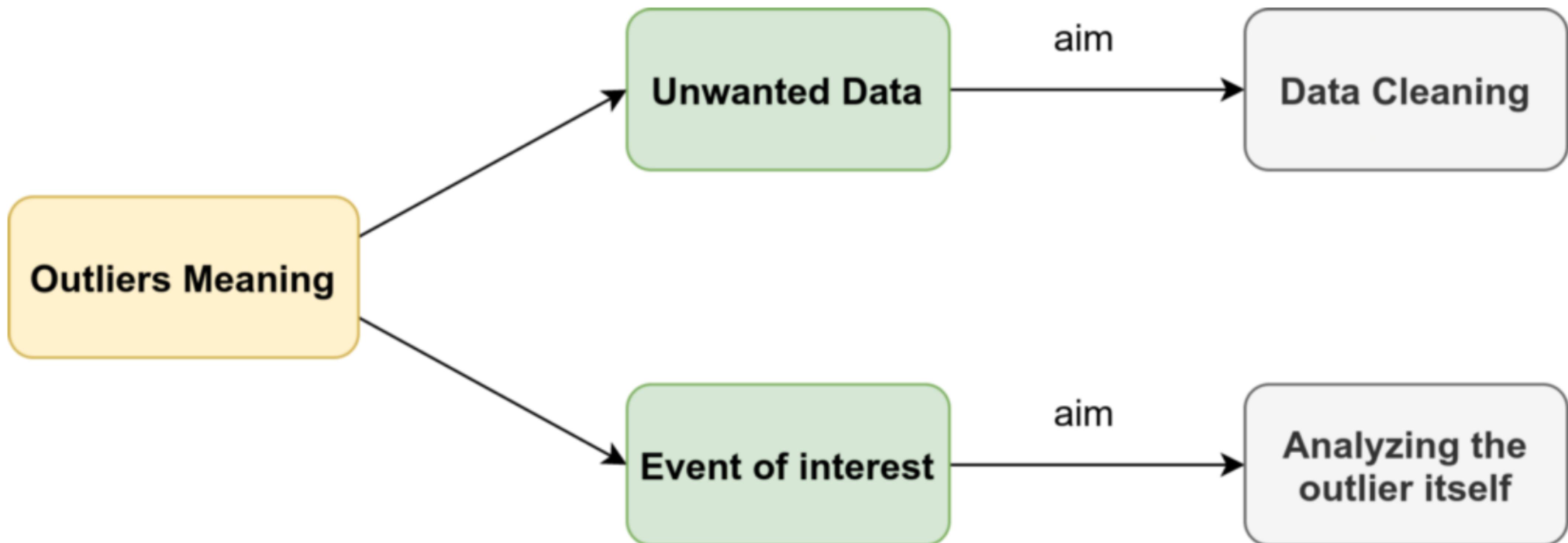
ИВАН ИВАНОВ 1МІ340085

Въведение

- Откриване на аномалии == Откриване на данни/събития/наблюдения, които се отклоняват от стандартното поведение на данните
- Аномалия == Outlier
- Дисбаланс на класовете - аномалиите са много по-малко от останалите данни
- Различни техники - Учене с учител и Учене без учител
- Различни видове данни, в които можем да търсим аномалии:



Защо търсим аномалии в данните?



Времеви редове

- Съвкупността от наблюденията над даден обект или явление през равни интервали във времето. Най-често времевите редове са описани като списък от двойки време-стойност: $[(2022-01-01, 5.5), (2022-01-02, 10.4), (2022-01-03, 7.13)]$
- Примери (за всеки пример взимаме по една стойност на час/ден/месец):
 - Брой посещения на уеб страница
 - Брой инсталации на приложение
 - Средна стойност (в лева) на поръчка
 - Bounce rate (брой посещения, които посещават една страница)
 - Churn rate ($B-E/B$; B - #клиенти в началото, E - #клиенти в края)
 - Cost per click/Revenue per click
 - Използвани CPU/RAM/IOPS ресурси

Характеристики на времевите редове - I

- Силна Стационарност (Stationarity) (свойство на стохастичния процес, който реализира времевия ред)

Средната стойност и дисперсието не се променят с времето. Ако данните не са стационарни, то можем да ги направим такива с някаква трансформация, например:
 $Y_t = Z_t - Z_{t-1}$ (и премахване на първия елемент), прилагане на логаритъм и др.

- Тенденция (Trend)

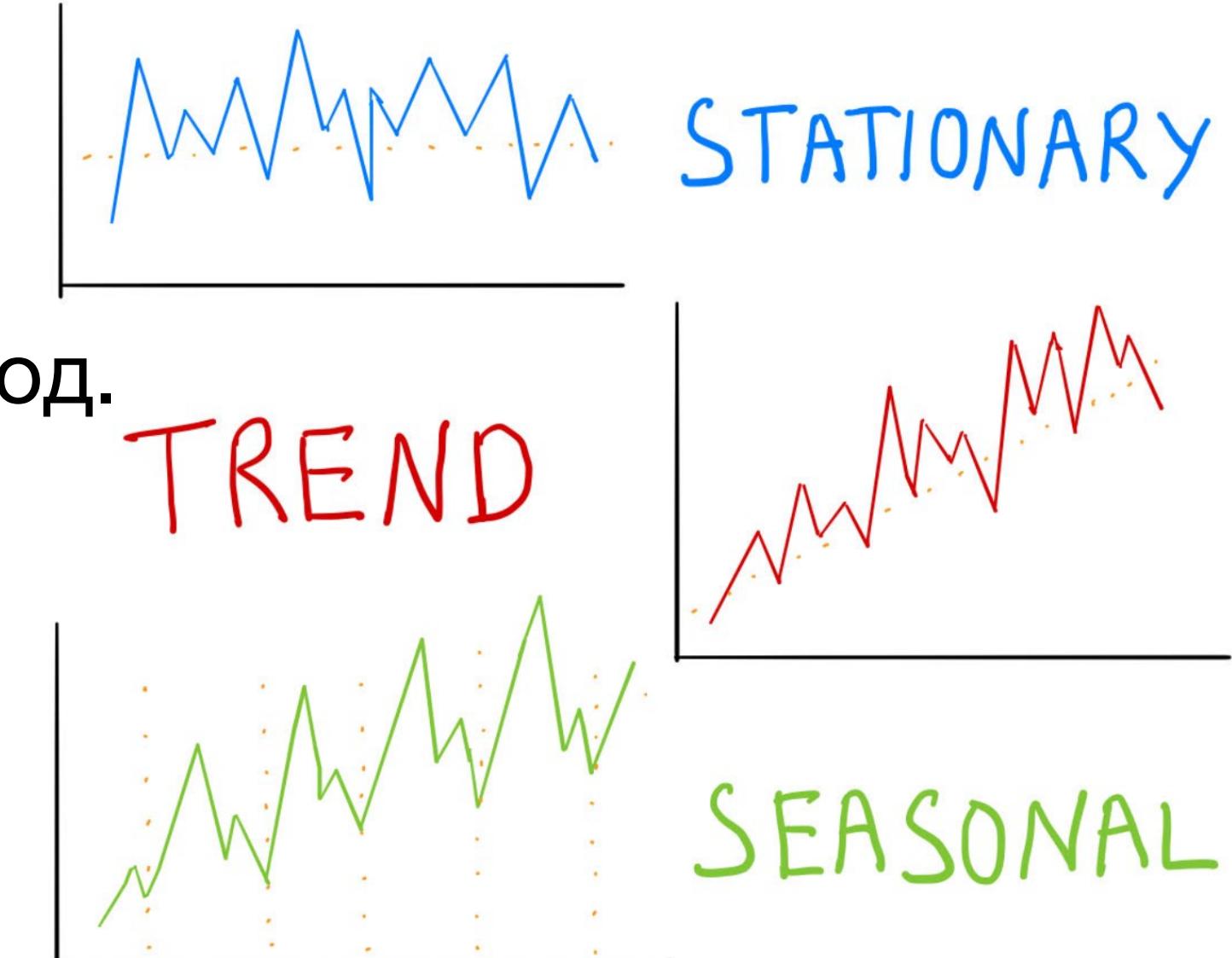
Плавна промяна под въздействието на закономерни трайно действащи причини, проявяващи се в целия изследван период.

- Сезонност (Seasonality)

Промяна в следствие на циклични фактори, проявяващи се с определена честота.

- Случайност (Irregularity)

Случайни отклонения, които се обясняват с действието на причини с несистематичен стохастичен характер. Тези отклонения не могат да се моделират и предсказват.



Характеристики на времевите редове - II

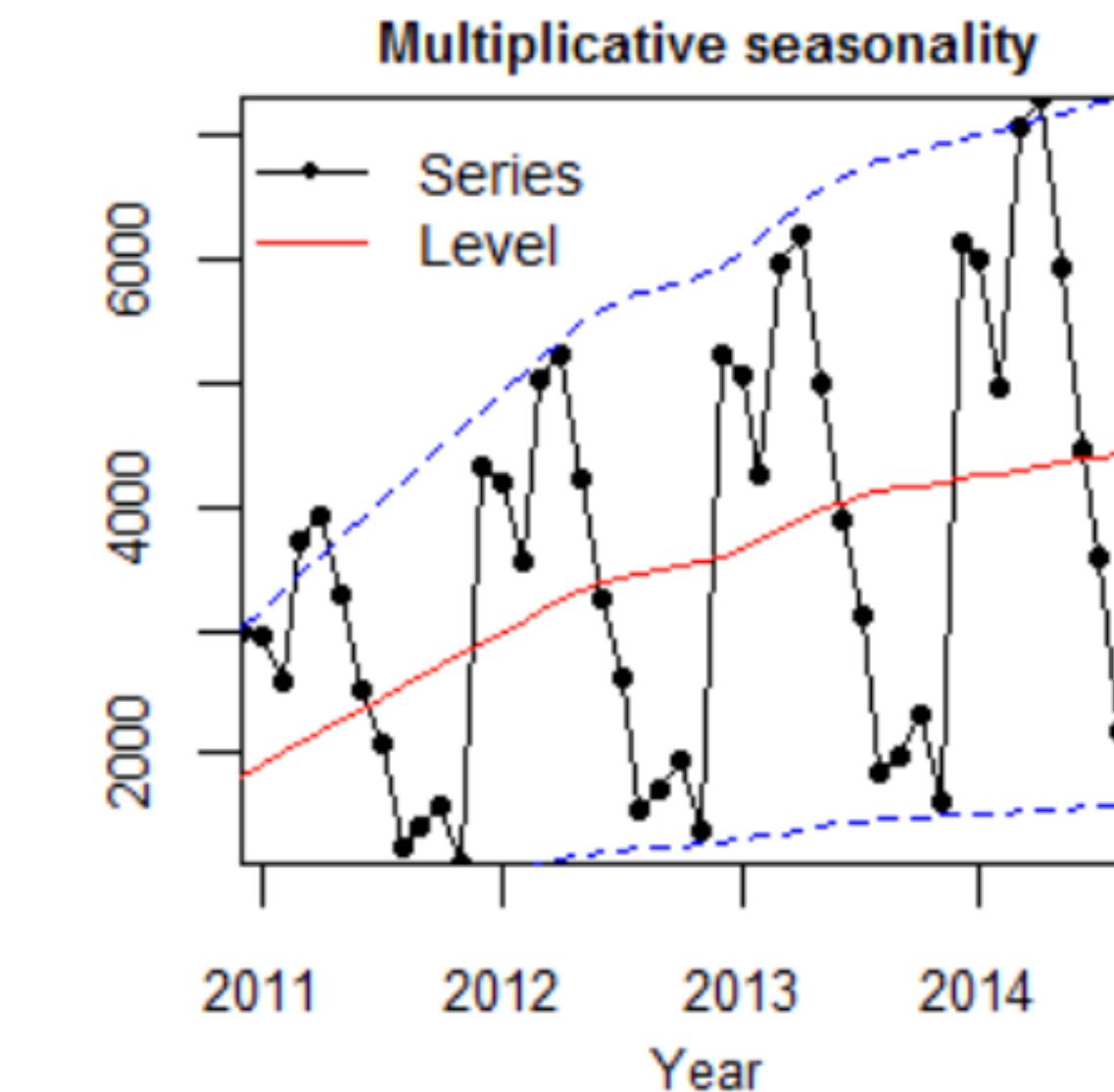
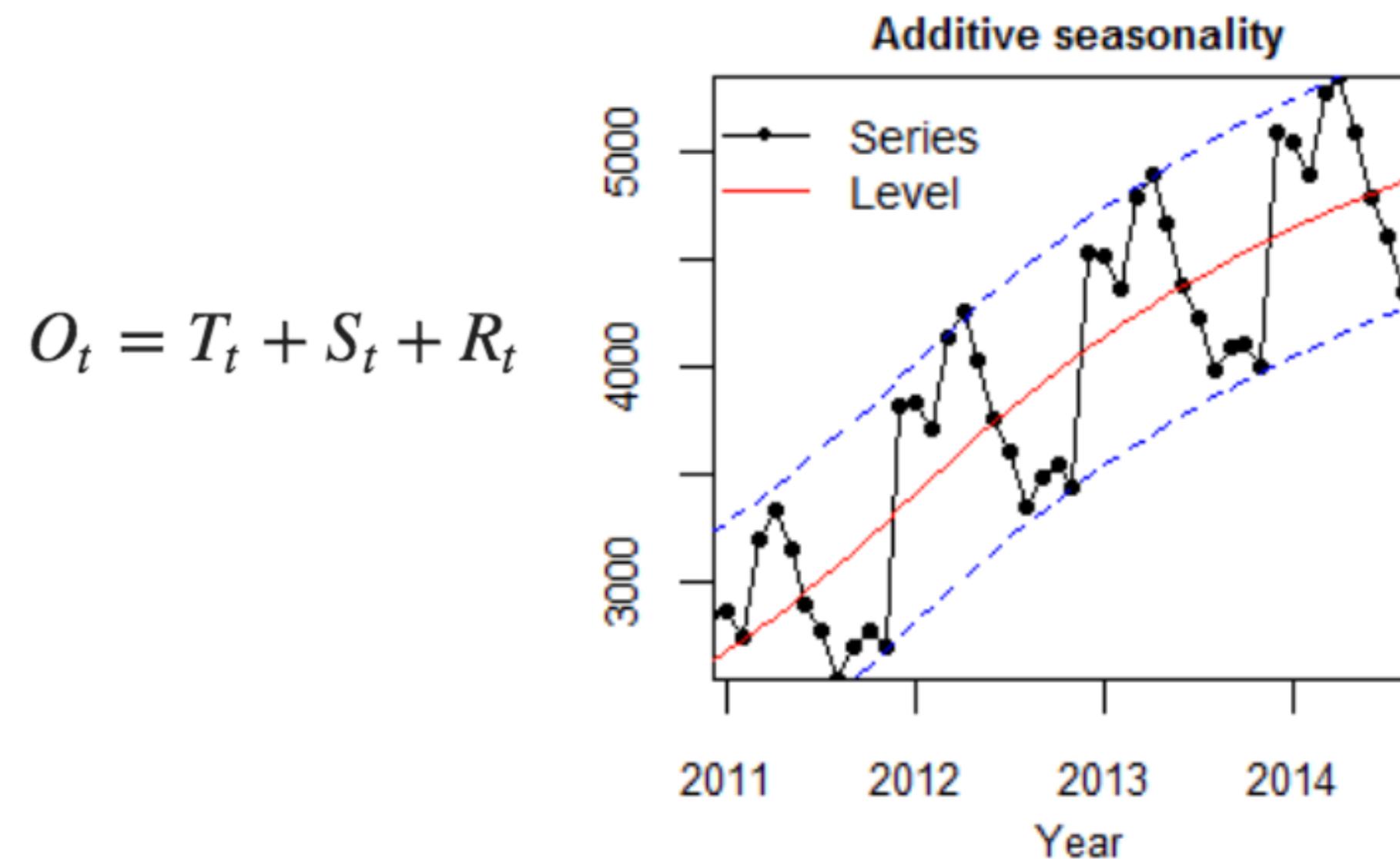
- Слаба Стационарност

Функцията на средната стойност μ_t не зависи от времето, а функцията на автокорелацията $\gamma(s, t)$ зависи само от s и t чрез тяхната абсолютна разлика $|s - t|$

- Адитивна и Мултипликативна Сезонност

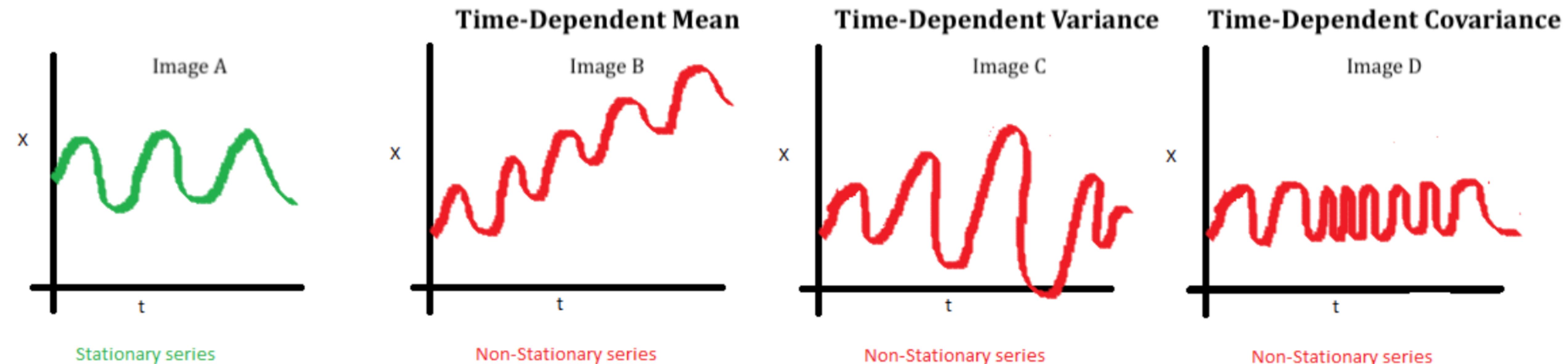
При адитивната сезонност, ефектът на сезонността се добавя към стойностите.

При мултипликативната, ефектът на сезонността е множител.

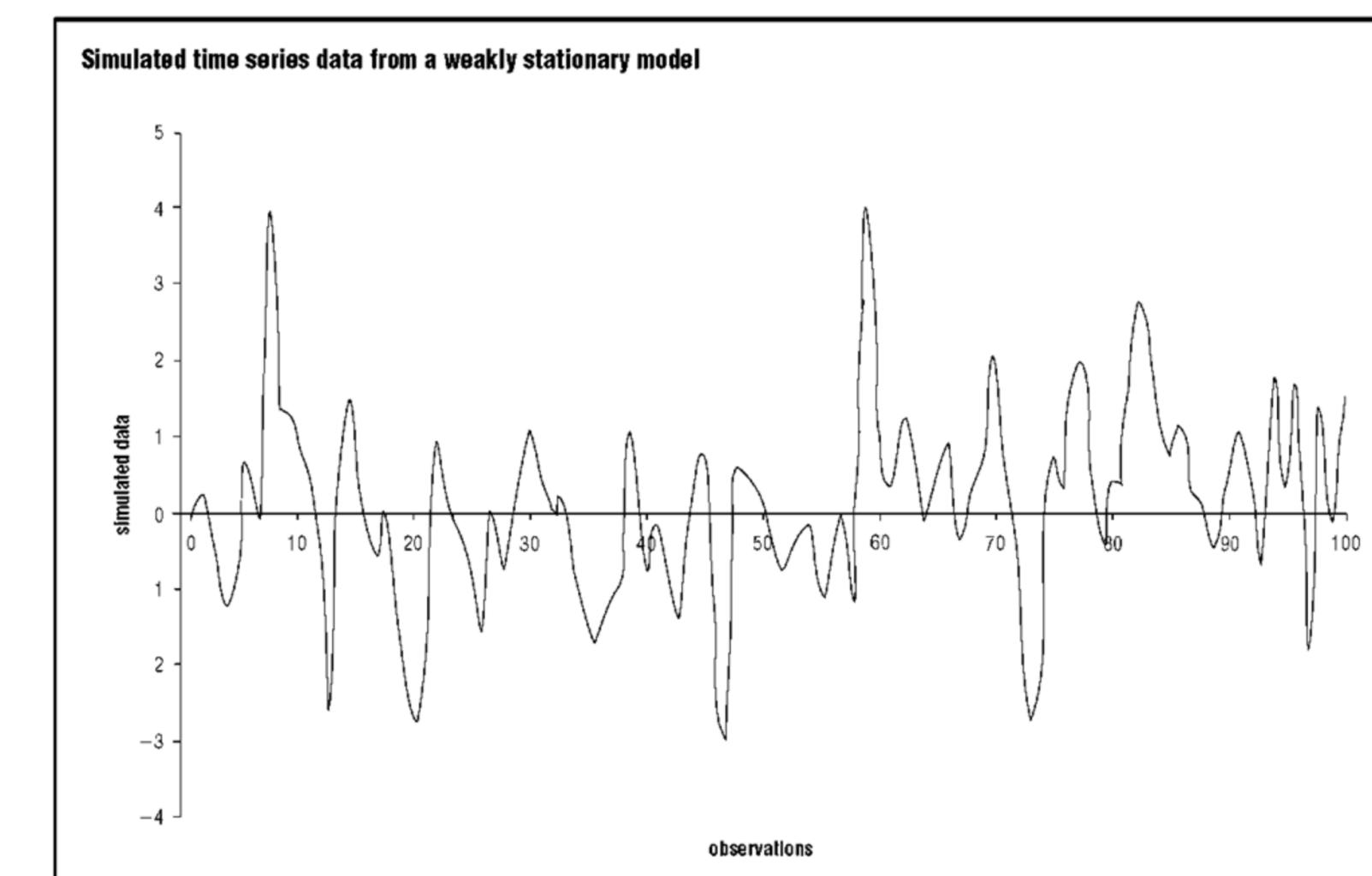


<https://stats.stackexchange.com/questions/282635/why-is-weak-stationarity-equivalent-to-strict-stationarity-only-when-distributio>

Характеристики на времевите редове - III



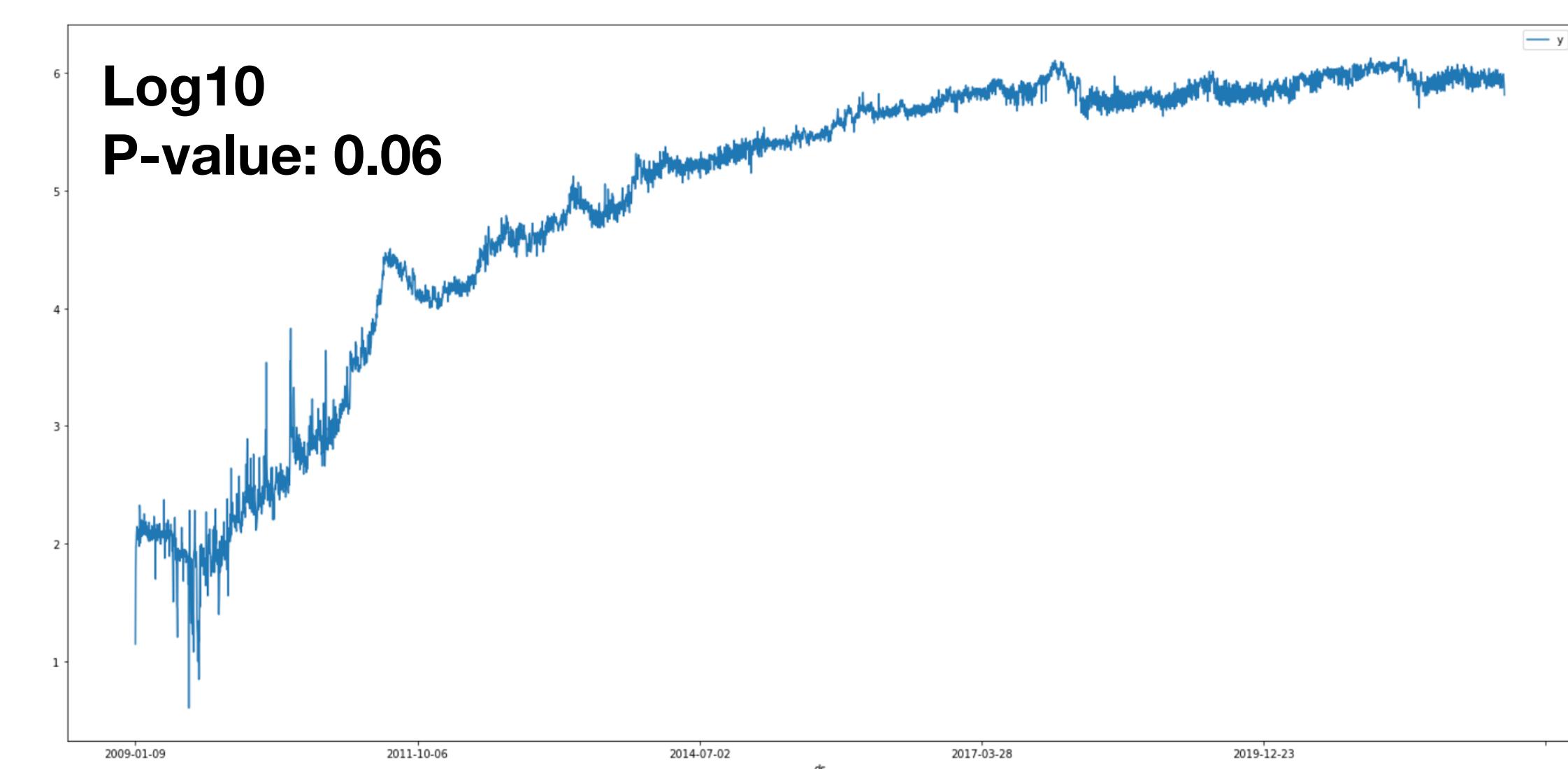
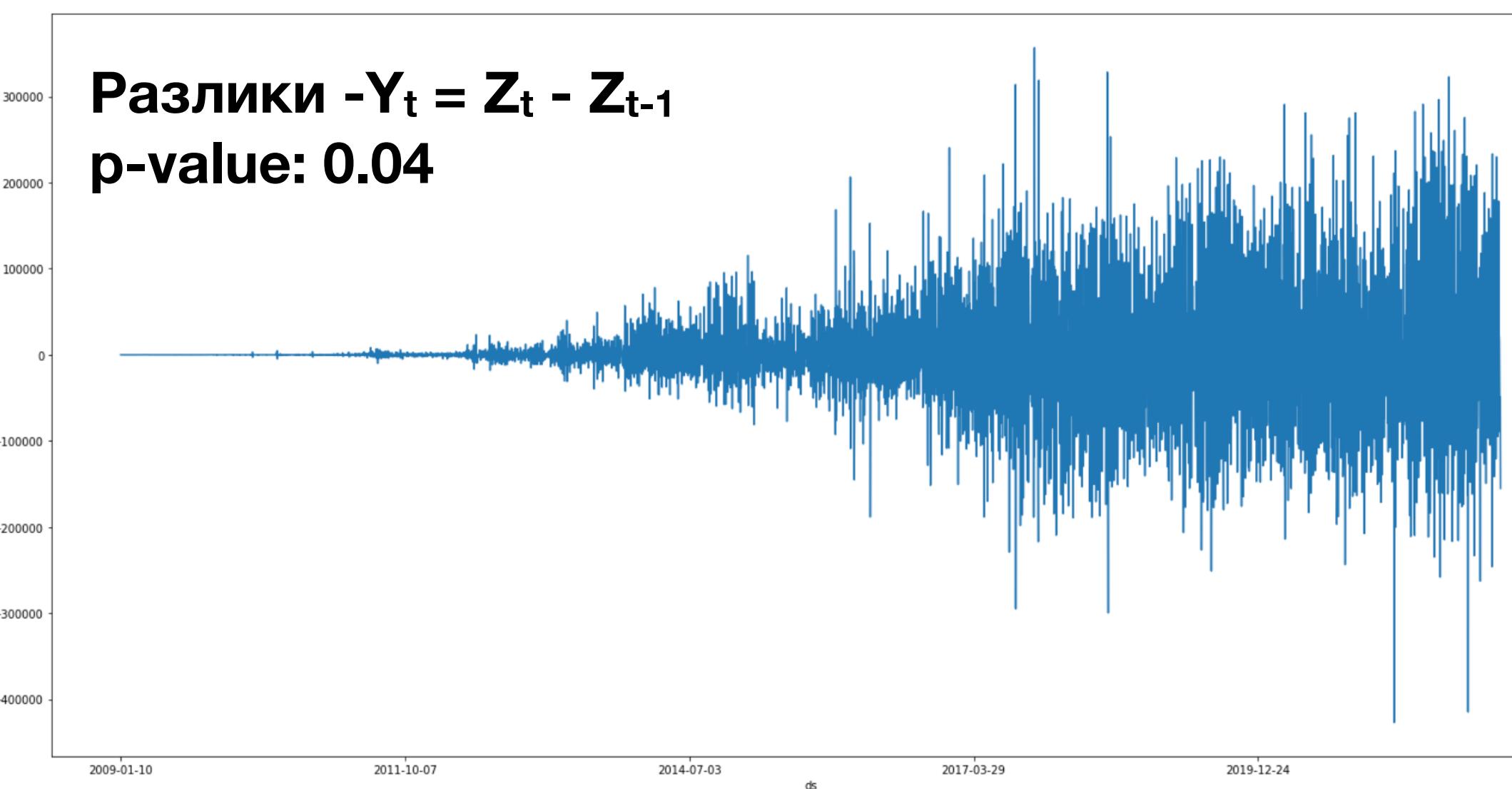
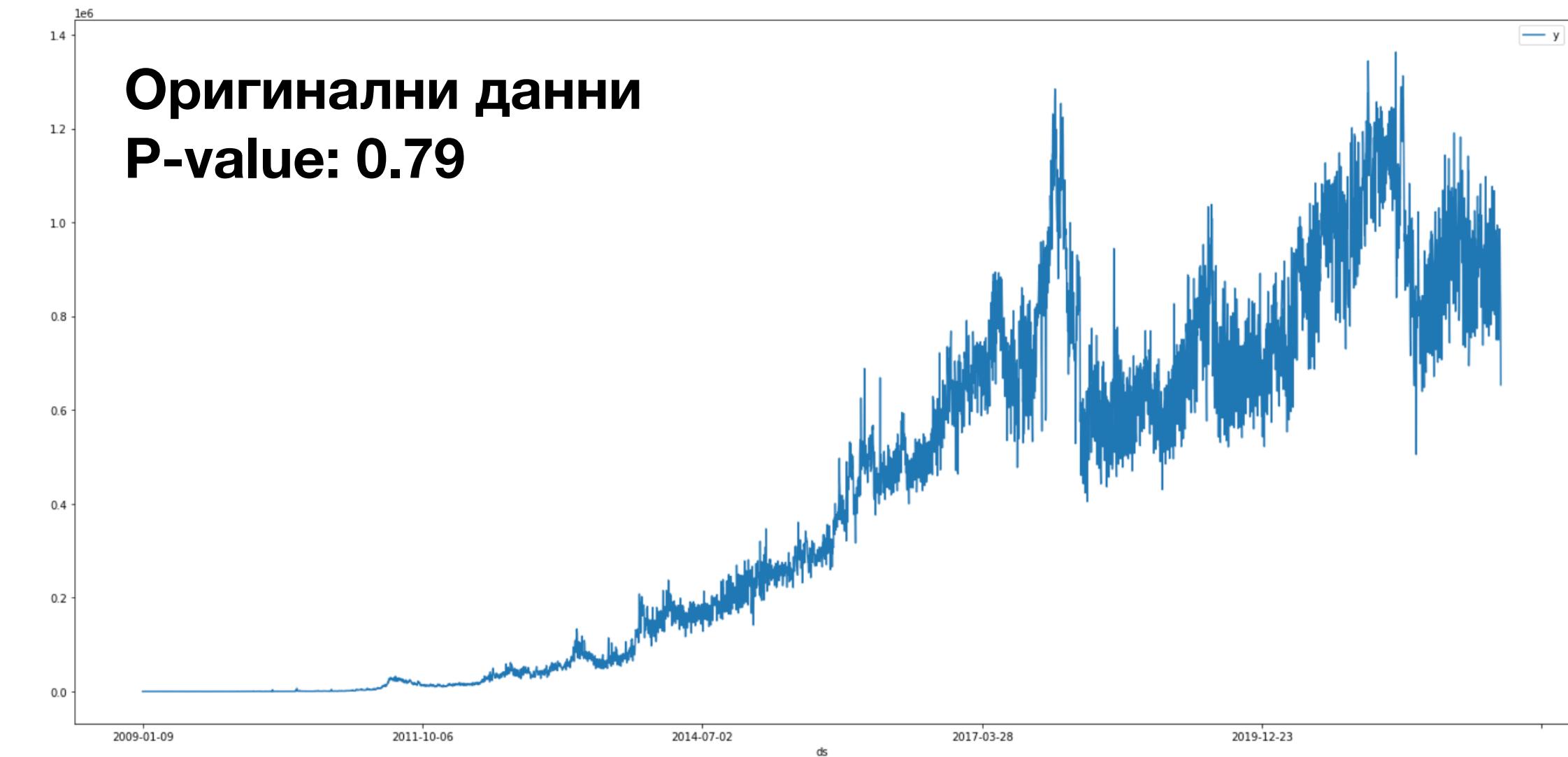
<https://towardsdatascience.com/achieving-stationarity-with-time-series-data-abd59fd8d5a0>



Характеристики на времевите редове - IV

Тест за стационарност:
Augmented Dickey-Fuller (adfuller)

- p-value ≤ 0.05 - стационарни
- P-value > 0.05 - нестационарни

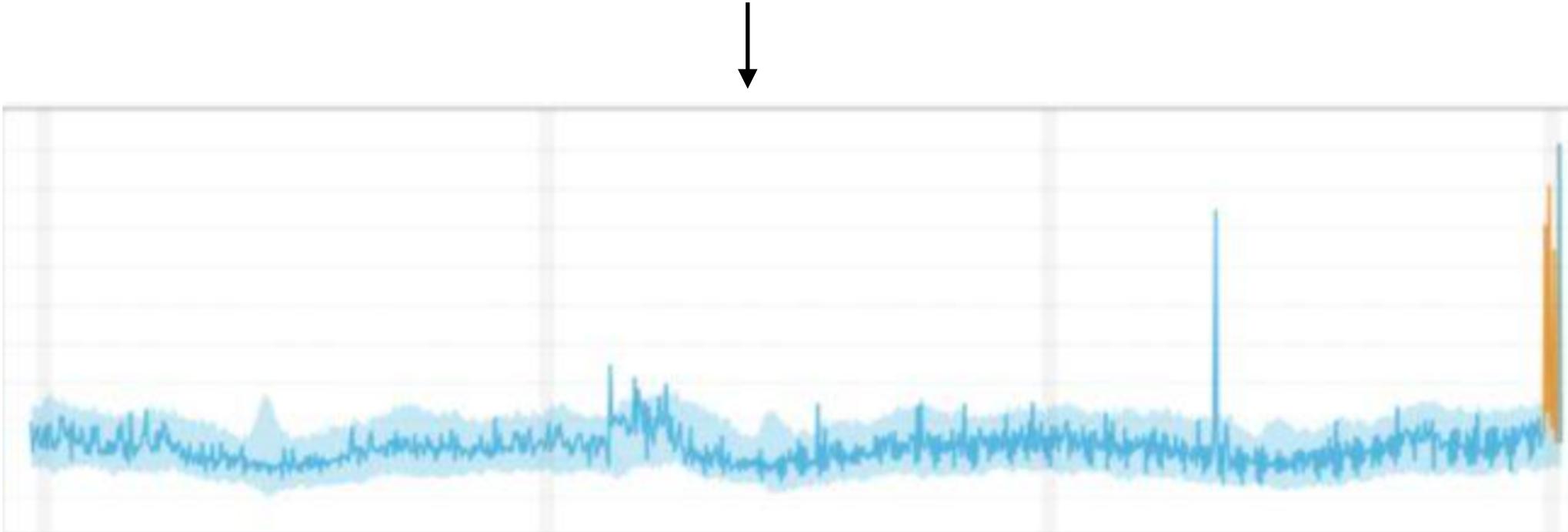


Видове аномалии във времеви редове

Различаваме три главни групи аномалии във времеви редове

Глобални

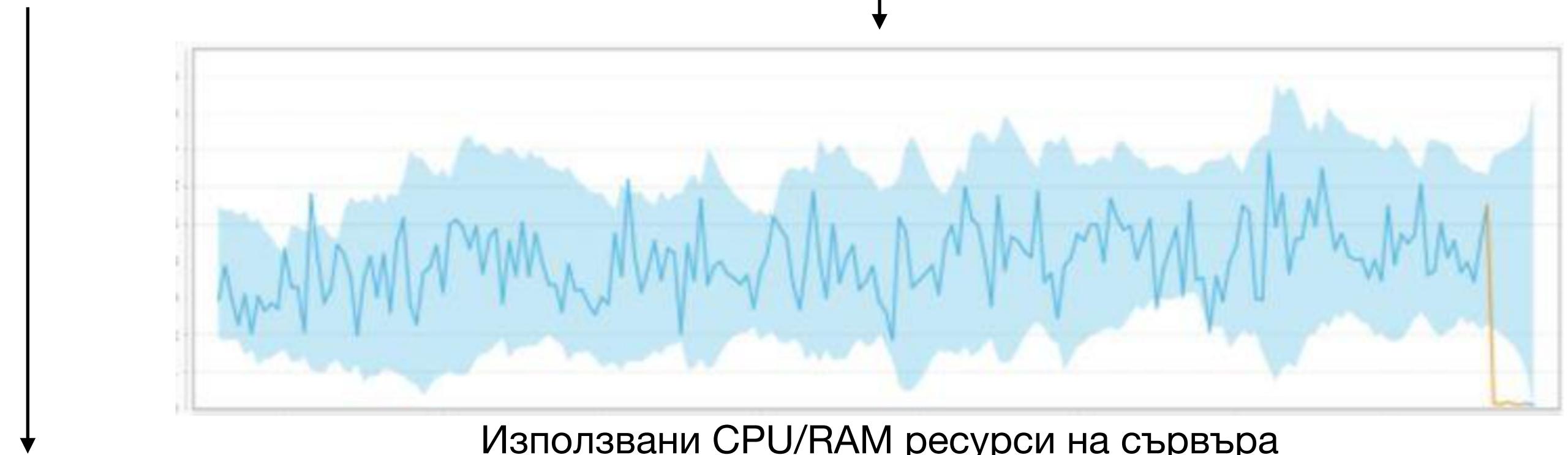
(global; point anomaly)



Време за отговор на заявка на HTTP сървър

Контекстуални

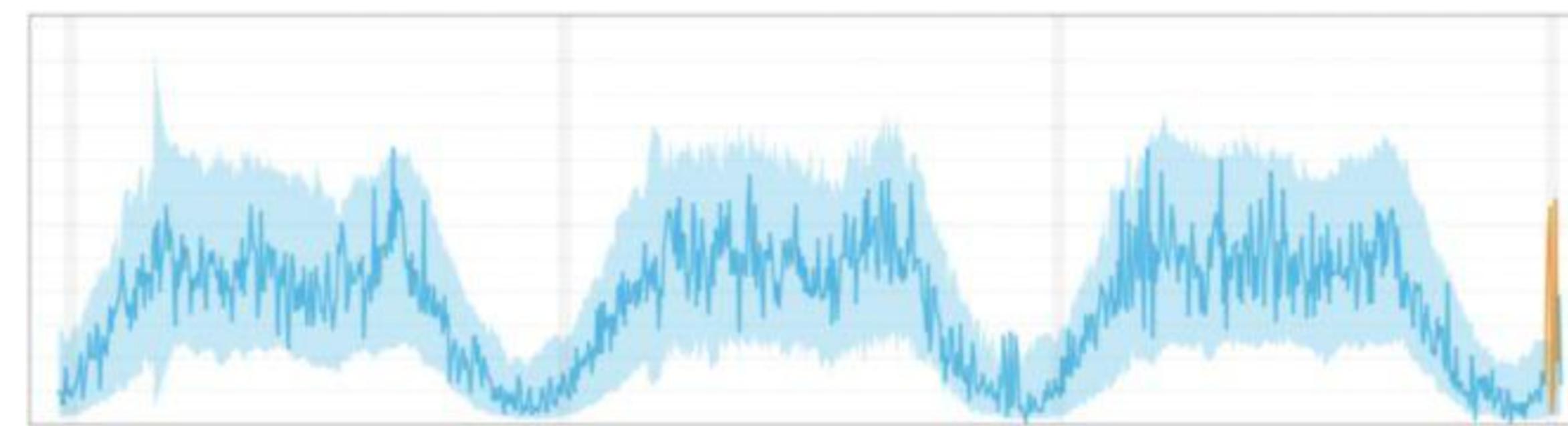
(contextual; conditional anomaly)



Използвани CPU/RAM ресурси на сървъра

Колективни

(collective)



Активни потребители - увеличение през деня и спад през нощта

<https://www.anodot.com/blog/quick-guide-different-types-outliers/>

Откриване на аномалии във времеви редове

- Типове данни
 - Анотирани (всички данни имат клас - аномалия или не) - могат да се използват познати техники като clustering analysis, isolation forests, neural network classifiers и др.
 - Неанотирани - това, което ни интересува в тази презентация.
- Типове техники за откриване на аномалии
 - Ръчно от човек
 - Базирано на правила (*if disk_usage > 95% and queries_count < 100; if payments_count = 0*)
 - “Умни” техники, базиращи се на статистически методи и/или машинно самообучение
 - Известни техники/алгоритми са: ARIMA, регресионни дървета, RNN/LSTM, движеща се средна стойност (Moving Average - SMA, EMA, WMA).
 - В общия случай, методите за откриване на аномалии без учител работят като строят опростен модел на данните и предсказват интервали, в които те могат да варират. Всичко извън тези предсказани граници се счита за аномалия.

Facebook Prophet - I

- Библиотека с отворен код - <https://github.com/facebook/prophet>
- Библиотека за прогнозиране (forecasting). Прогнозирането е основна задача в Data Science, чието решаване в частност спомага за откриване на аномалии.
- Работи без нужда от прецизна настройка. Това е голям плюс в сравнение с другите популярни методи.
- Напълно автоматичното прогнозиране е трудно за настройване и не е гъвкаво, а хората с добри познания в изследваната област имат малки познания и опит в прогнозирането на времеви редове.



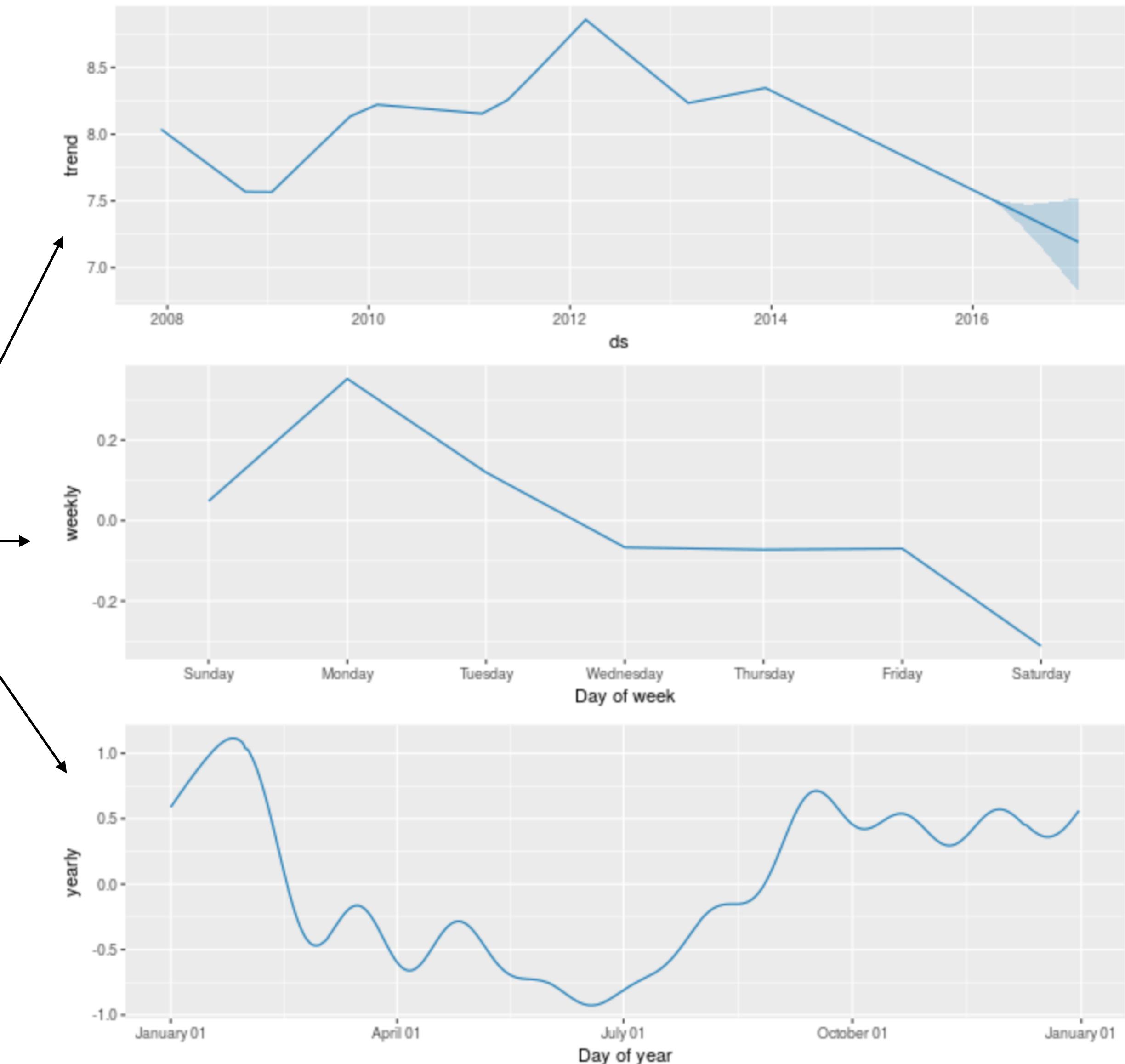
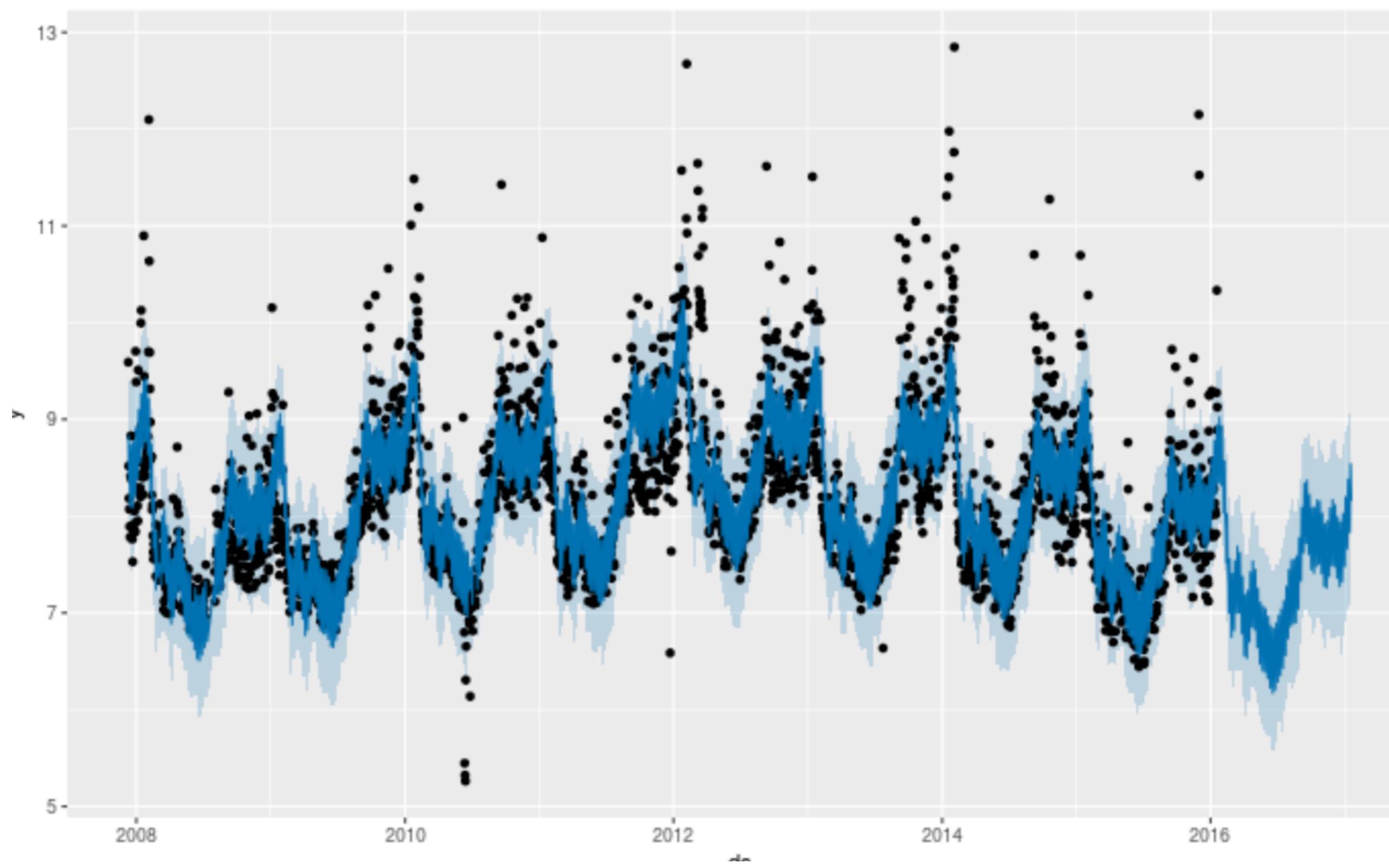
Facebook Prophet - II

Prophet е адитивен регресивен модел с 4 главни компоненти:

1. Компонента за откриване на тенденция (чрез *piecewise linear* или *logistic growth curve*). Prophet автоматично намира точките, в които се променя тенденцията - точки на промяна. Те могат да бъдат подадени и от потребителя.
2. Компонента за откриване на седмична сезонност, използваща фиктивни (dummy) променливи
3. Компонента за откриване на годишна сезонност, използваща трансформации на Фурье
4. Използване на подаден от потребителя списък с важни дати (holidays)

Facebook Prophet - III

Брой прегледи на уикипедия страницата на спортист по американски футбол (Пейтън Манинг)



Facebook Prophet - IV

<https://www.kaggle.com/code/vinayjaju/anomaly-detection-using-facebook-s-prophet/notebook>

- Подготвяне на данните в две колони: **ds** и **y** (*няма изискване за стационарност*)
- Конфигуриране на модела (seasonality_mode, daily/monthly/yearly seasonality, etc.)
- Трениране
- Предсказване

```
def fit_predict_model(dataframe, interval_width = 0.99, changepoint_range = 0.8):
    m = Prophet(daily_seasonality = False,
                yearly_seasonality = False,
                weekly_seasonality = False,
                seasonality_mode = 'multiplicative',
                interval_width = interval_width,
                changepoint_range = changepoint_range)
    m = m.fit(dataframe)

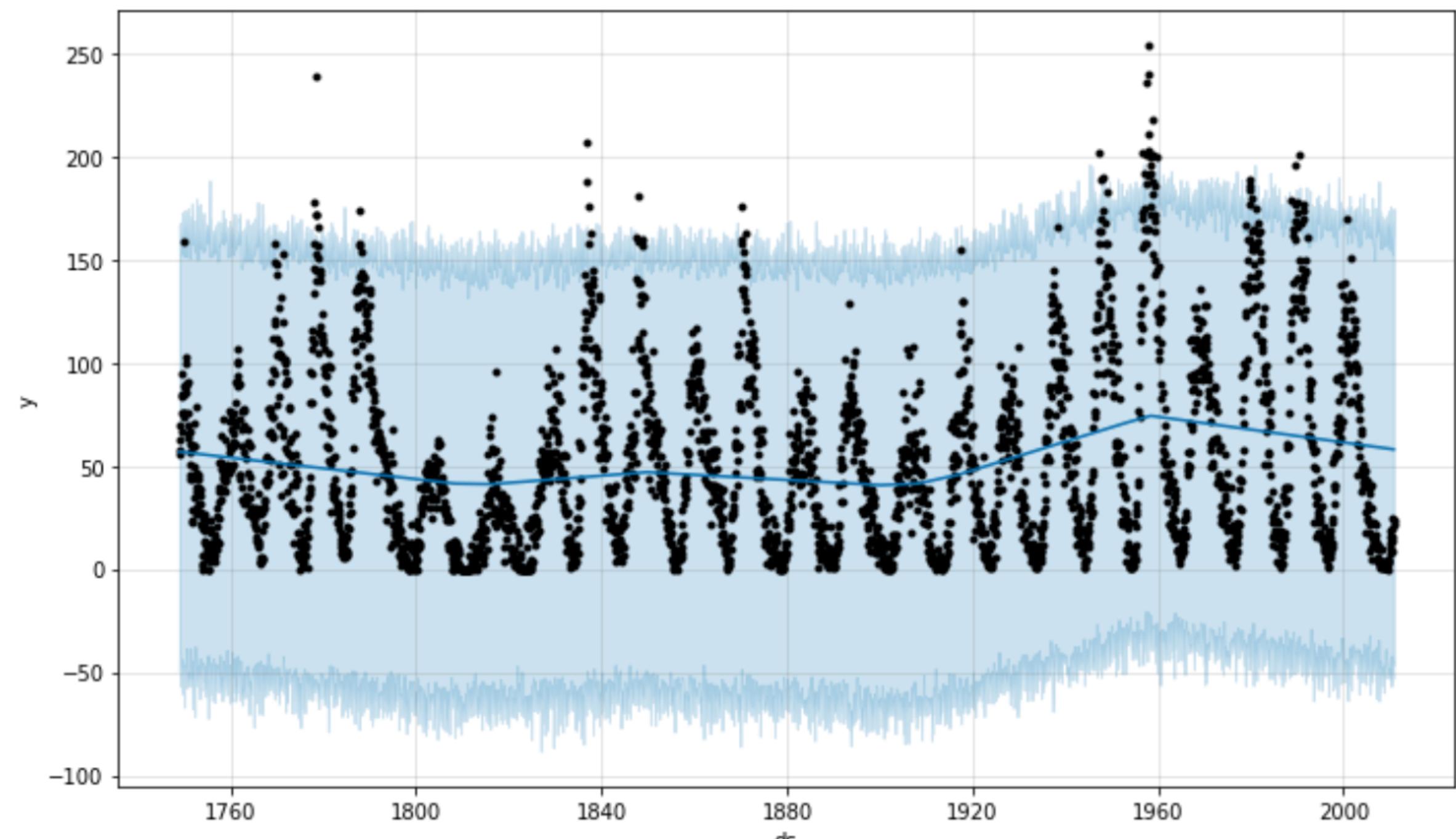
    forecast = m.predict(dataframe)
    forecast['fact'] = dataframe['y'].reset_index(drop = True)

    return forecast
```

Facebook Prophet - V

Резултатът от изпълнението добавя няколко колони. Те са налични и за историческите данни, а не само за предсказаните нови.

- **y** - истинската стойност
- **yhat_lower** - долната стойност на интервала на достоверност
- **yhat_upper** - горната стойност на интервала на достоверност
- **yhat** - предсказана стойност



Facebook Prophet - VI

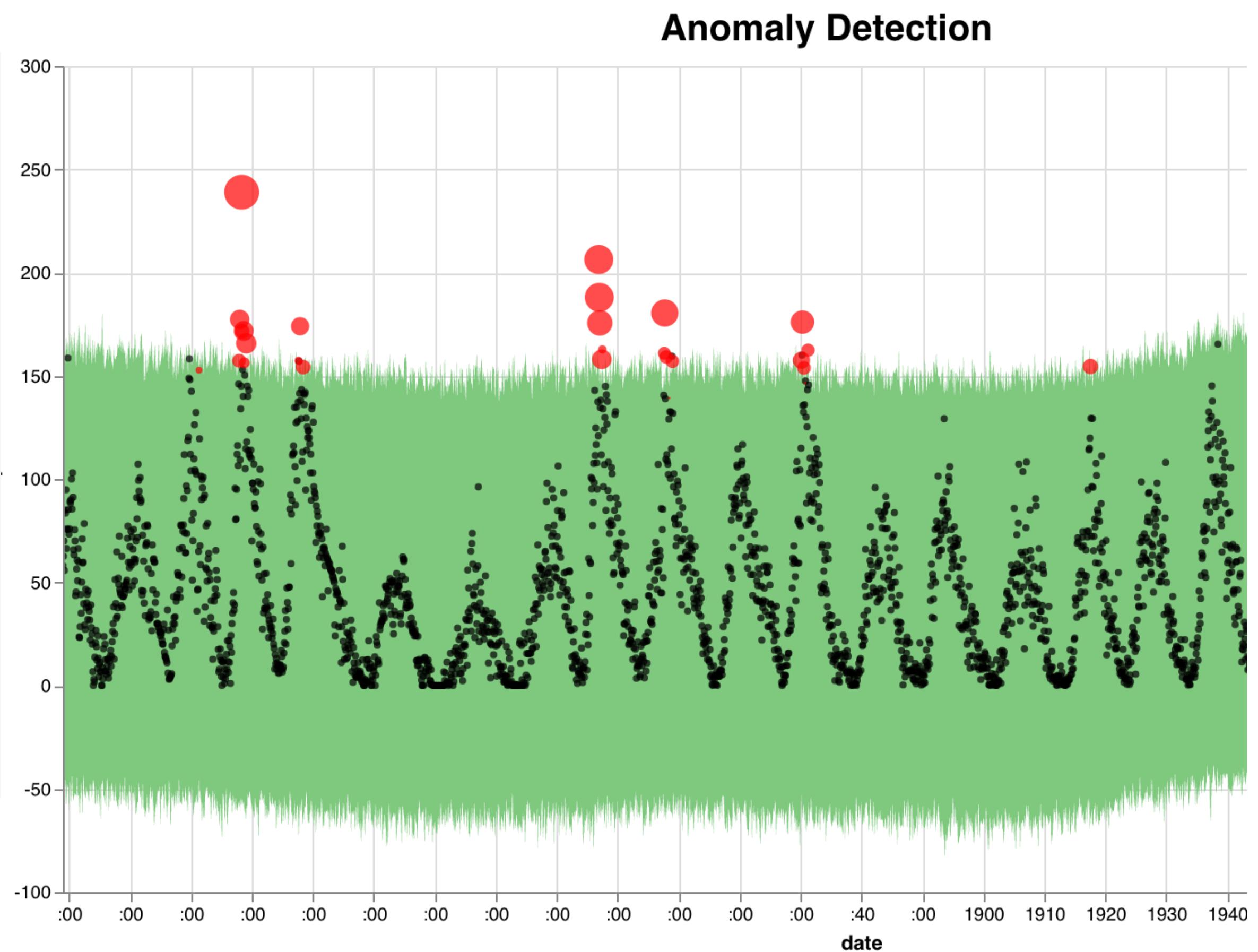
От получените данни извличаме аномалиите. Това са всички данни, за които $y > yhat_{upper}$ или $y < yhat_{lower}$

```
def detect_anomalies(forecast):
    forecasted = forecast[['ds', 'trend', 'yhat', 'yhat_lower', 'yhat_upper', 'fact']]
    forecasted['fact'] = df['y']

    forecasted['anomaly'] = 0
    forecasted.loc[forecasted['fact'] > forecasted['yhat_upper'], 'anomaly'] = 1
    forecasted.loc[forecasted['fact'] < forecasted['yhat_lower'], 'anomaly'] = -1

    #anomaly importances
    forecasted['importance'] = 0
    forecasted.loc[forecasted['anomaly'] == 1, 'importance'] = \
        (forecasted['fact'] - forecasted['yhat_upper'])/forecasted['fact']
    forecasted.loc[forecasted['anomaly'] == -1, 'importance'] = \
        (forecasted['yhat_lower'] - forecasted['fact'])/forecasted['fact']

    return forecasted
```



ФИНАЛ