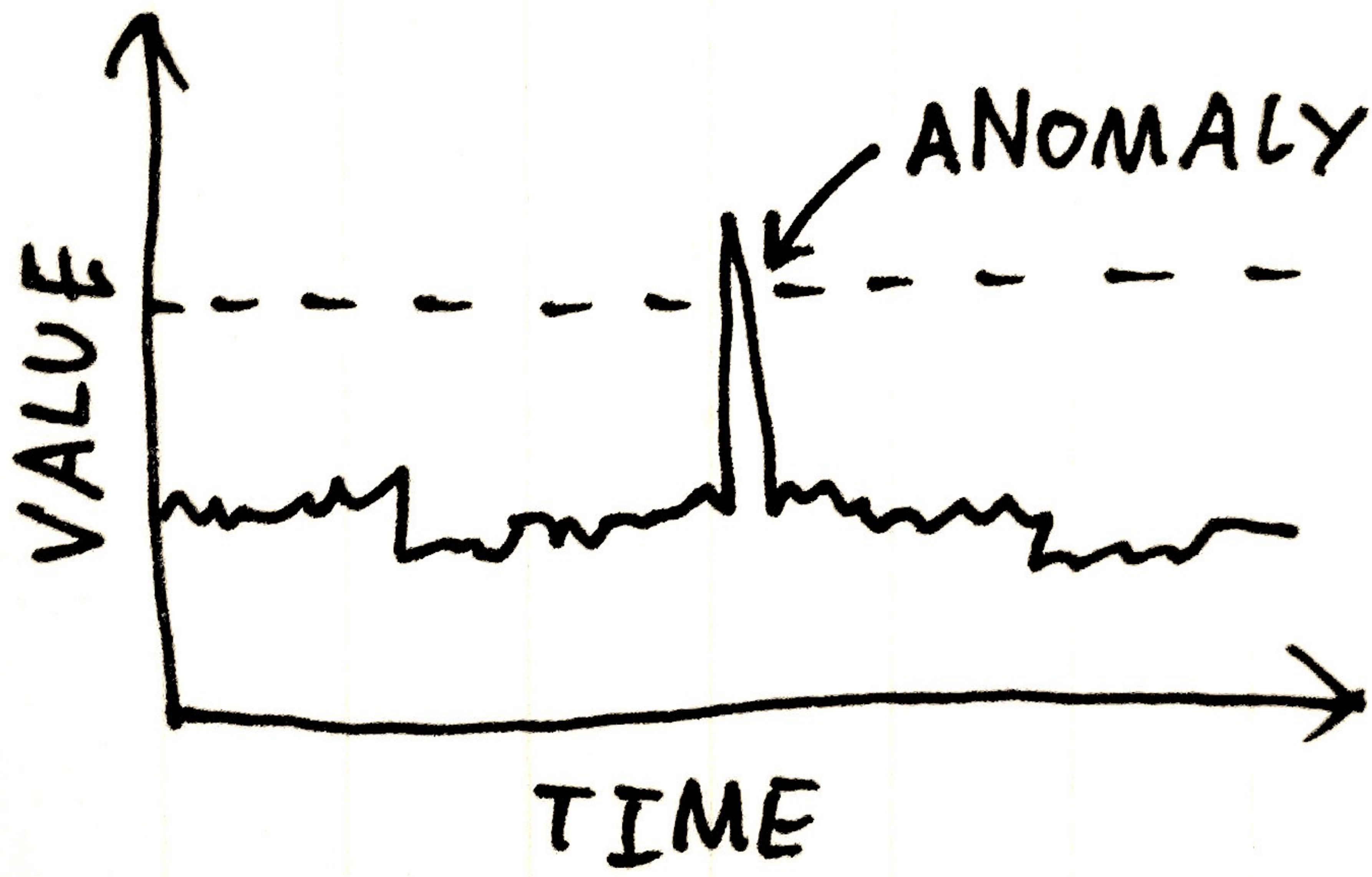
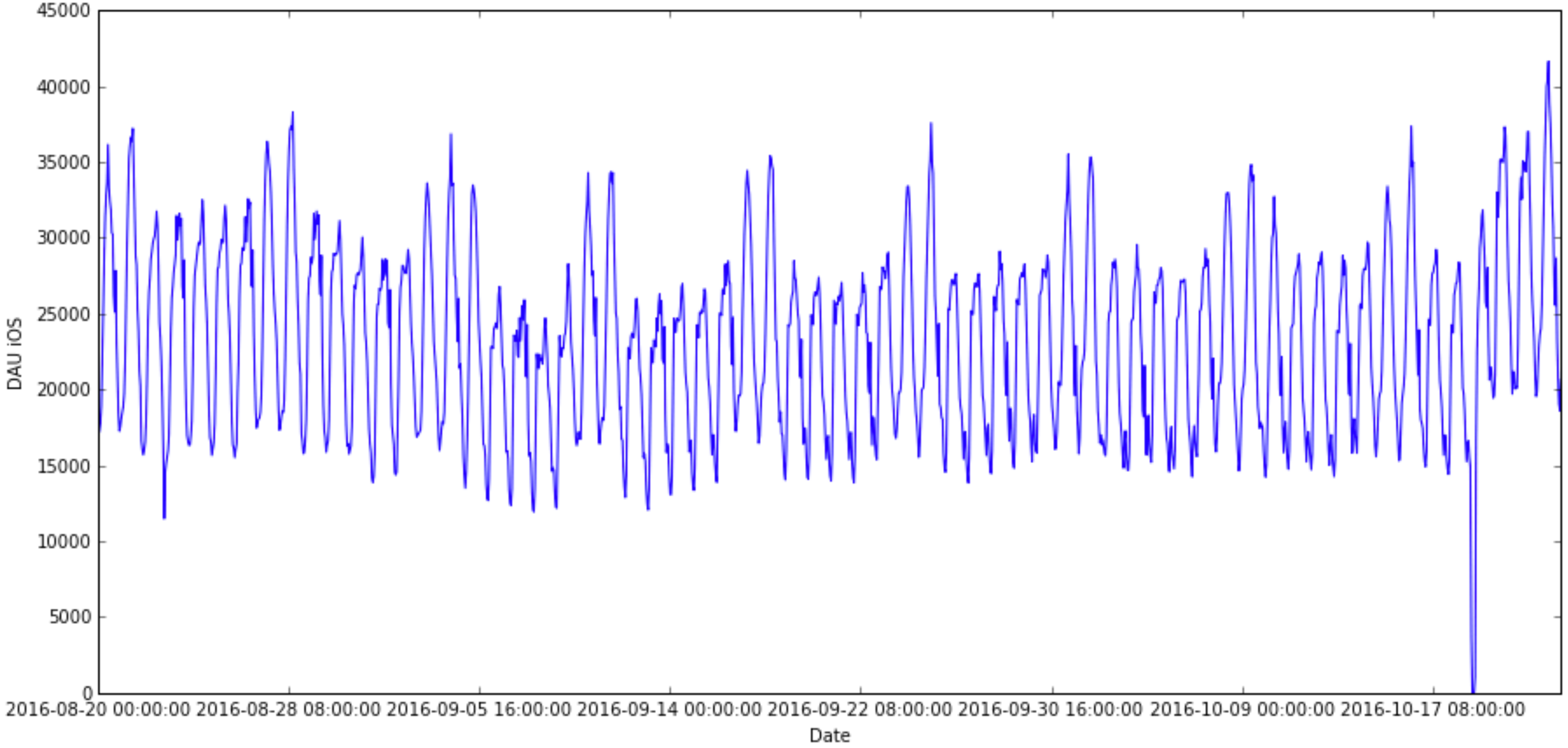
Anomaly Detection in Time Series

Dmitry Sergeyev

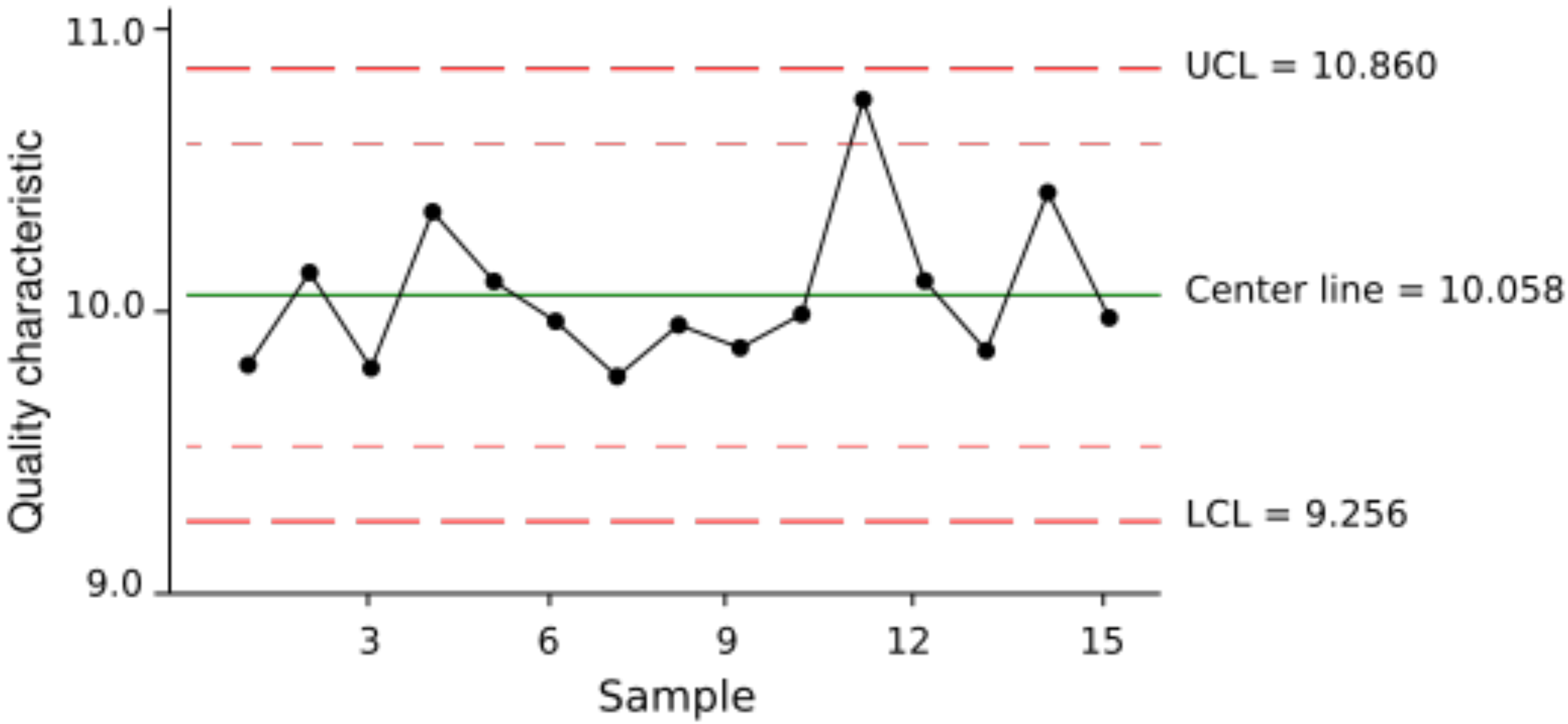
Why?



Why?


45000
40000
35000
30000
25000
20000
15000
10000
5000
0
DAU iOS
2016-08-20 00:00:00
2016-08-28 08:00:00
2016-09-05 16:00:00
2016-09-14 00:00:00
2016-09-22 08:00:00
2016-09-30 16:00:00
2016-10-09 00:00:00
2016-10-17 08:00:00
Date

Statistical Process Control



https://www.moresteam.com/toolbox/statistical-process-control-spc.cfm

Statistical Process Control

The application of SPC involves three main phases of activity:

- Understanding the process and the specification limits
- Eliminating assignable (special) sources of variation, so that the process is stable
- Monitoring the ongoing production process, assisted by the use of control charts, to detect significant changes of mean or variation.

Window estimations

Window estimations

- Simple moving average
- Weighted moving average
- Exponentially weighted moving average
- etc.

Window estimations

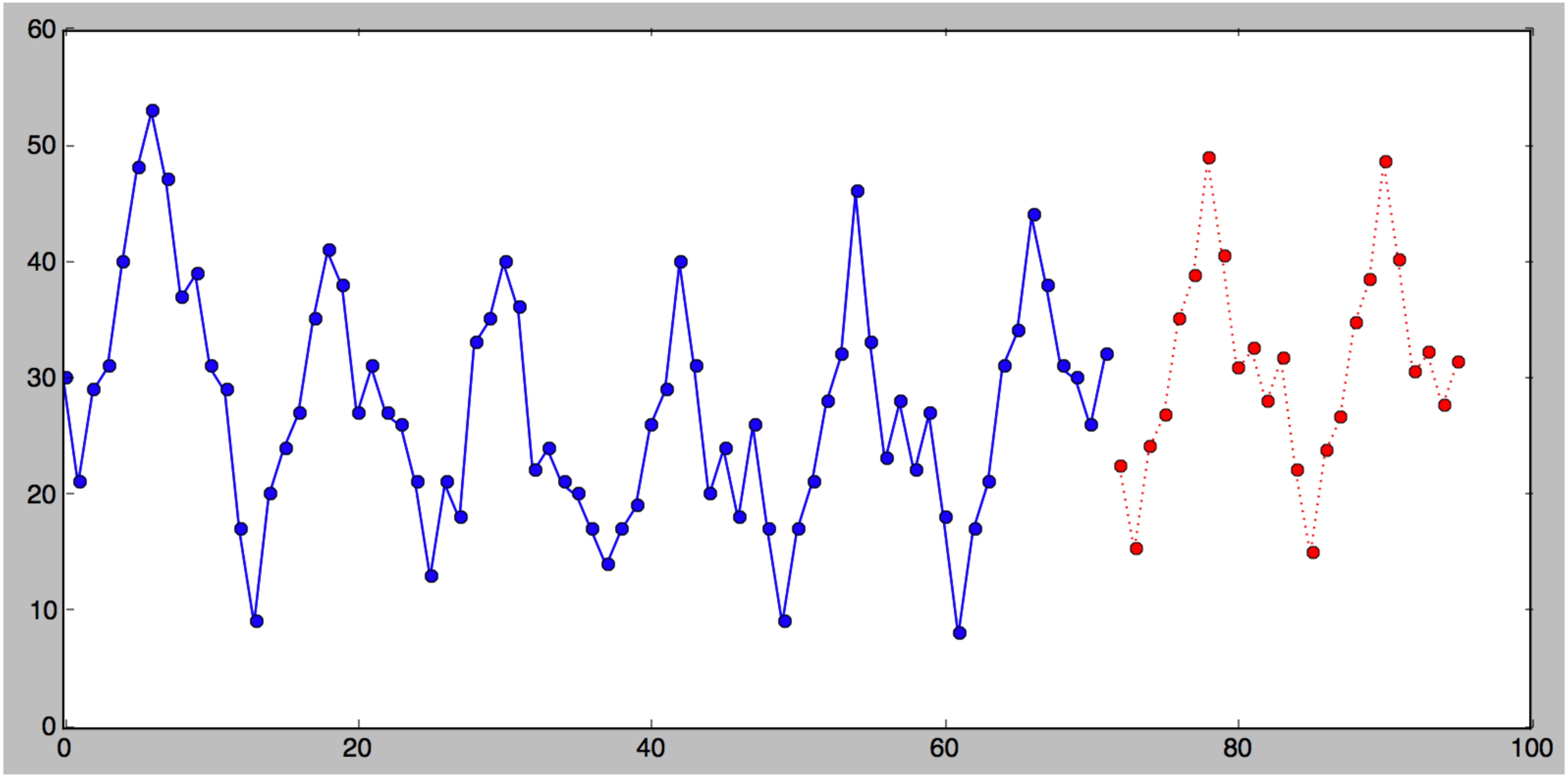
Good:

- Easy to use
- Fast calculations

Bad:

- One-step ahead predictions
- False positives

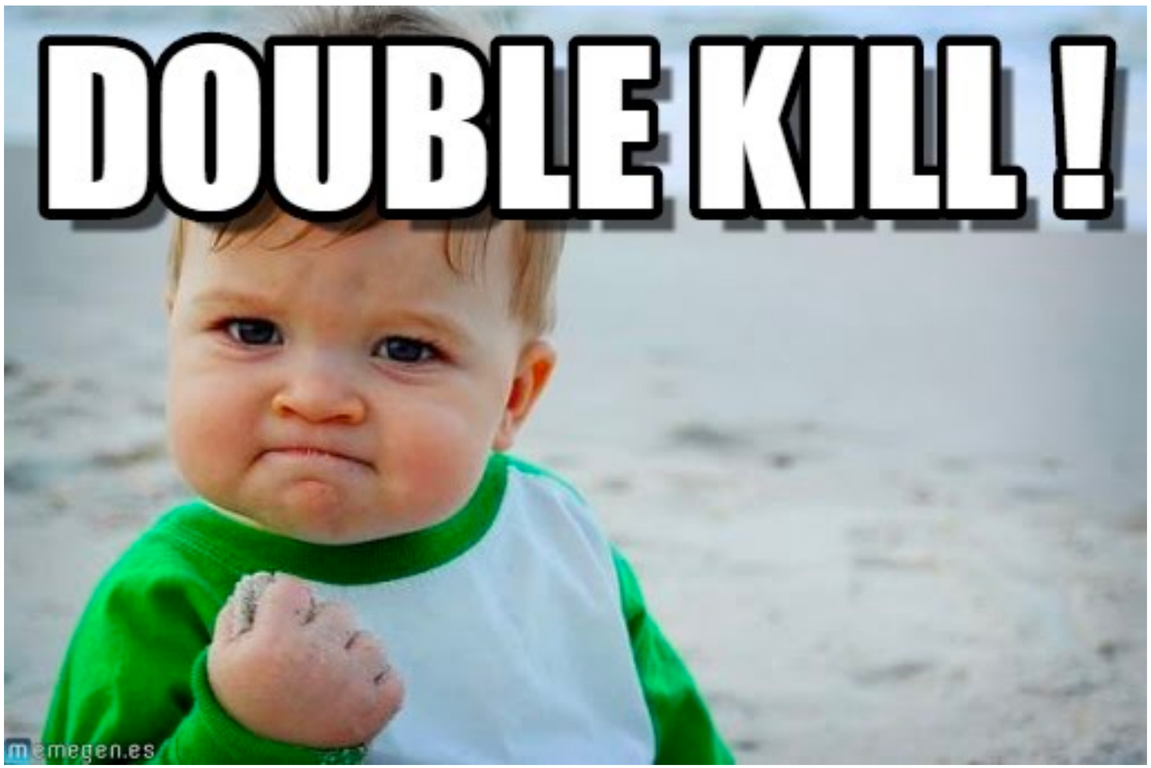
Exponential Smoothing



https://grisha.org/blog/2016/01/29/triple-exponential-smoothing-forecasting/

Exponential Smoothing

Double Exponential Smoothing



Exponential Smoothing

Double Exponential Smoothing

Double exponential smoothing is nothing more than exponential smoothing applied to both level and trend.

- Level – expected value/slope/intercept
- Trend or Slope – $y_t - y_{t-1}$

Exponential Smoothing

Double Exponential Smoothing

$$
\begin{aligned}
\ell_x &= \alpha y_x + (1-\alpha)(\ell_{x-1} + b_{x-1}) && \text{level} \\
b_x &= \beta(\ell_x - \ell_{x-1}) + (1-\beta) b_{x-1} && \text{trend} \\
\hat{y}_{x+1} &= \ell_x + b_x && \text{forecast}
\end{aligned}
$$

- In the first equation we're using ℓ instead of ŷ and on the right side the expected value becomes the sum of level end trend.

- The second equation introduces β, the trend factor (or coefficient). As with α, some values of β work better than others depending on the series.

Holt-Winters

Triple Exponential Smoothing

Holt-Winters

Triple Exponential Smoothing

$$\ell_x = \alpha(y_x - s_{x-L}) + (1-\alpha)(\ell_{x-1} + b_{x-1}) \qquad \text{level}$$
$$b_x = \beta(\ell_x - \ell_{x-1}) + (1-\beta)b_{x-1} \qquad \text{trend}$$
$$s_x = \gamma(y_x - \ell_x) + (1-\gamma)s_{x-L} \qquad \text{seasonal}$$
$$\hat{y}_{x+m} = \ell_x + mb_x + s_{x-L+1+(m-1)modL} \qquad \text{forecast}$$

- γ - the smoothing factor for the seasonal component
- We can forecast any number of points into the future
- The forecast equation now consists of level, trend and the seasonal component

Brutlag

$$\hat{y}_{max_t} = L_{t-1} + P_{t-1} + S_{t-T} + m \cdot d_{t-T}$$

$$\hat{y}_{min_t} = L_{t-1} + P_{t-1} + S_{t-T} - m \cdot d_{t-T}$$

d is predicted deviation given by:

$$d_t = \gamma \left| y_t - \hat{y}_t \right| + (1 - \gamma) d_{t-T}$$

γ is the seasonal change smoothing factor
m is the scaling factor for Brutlags confidence bands

https://fedcsis.org/proceedings/2012/pliks/118.pdf

Kolmogorov-Smirnov

Good:

- works on periodic data when its period is smaller than the window size
- outlier resistant (less sensitive to small-scale changes with larger windows)
- good for data exploration (allows top-down exploration of large-scale changes first)

Bad:

- produces false positives on trends and seasonal changes
- needs many (hundreds) unique values (it won't work on metrics gathered with 5 minutes interval)
- bad for alerting because of false positives

http://mabrek.github.io/blog/statistics-for-monitoring-anomaly-p2/

SARIMA

$SARIMA(p,d,q)(P,D,Q)_s$

- p – AR(p)
- d – порядок интегрирования
- q – MA(q)
- P – порядок сезонной составляющей SAR(P)
- D – порядок интегрирования сезонной составляющей
- Q – порядок сезонной составляющей SMA(Q)
- s – размерность сезонности (месяц, квартал и т.д.)

https://habrahabr.ru/post/210530/

Prophet

https://facebookincubator.github.io/prophet/static/prophet_paper_20170113.pdf

https://facebookincubator.github.io/prophet/

Prophet

Good:

- Stuff is done for you
- Custom parameters
- Easy to use

Bad:

- Less accurate
- Doesn’t work with non-daily/monthly data