Dataset:

Japan wiki, has a spike; grouped by time window; 5 – 10 – 15 min ; with 10min used for tuning

(t, t – 1, … t- n) to predict (t + 2)

The prediction is done as we close moment t, therefore if we predict moment t+1 and most requests are in the beginning of the interval the system would not have time to react.

Baseline

The naïve approach: predict that the workload won’t change, so nr requests at moment t+2 will be nr requests at moment t.