Measuring the Stability of Process Outcome Predictions in Online Settings

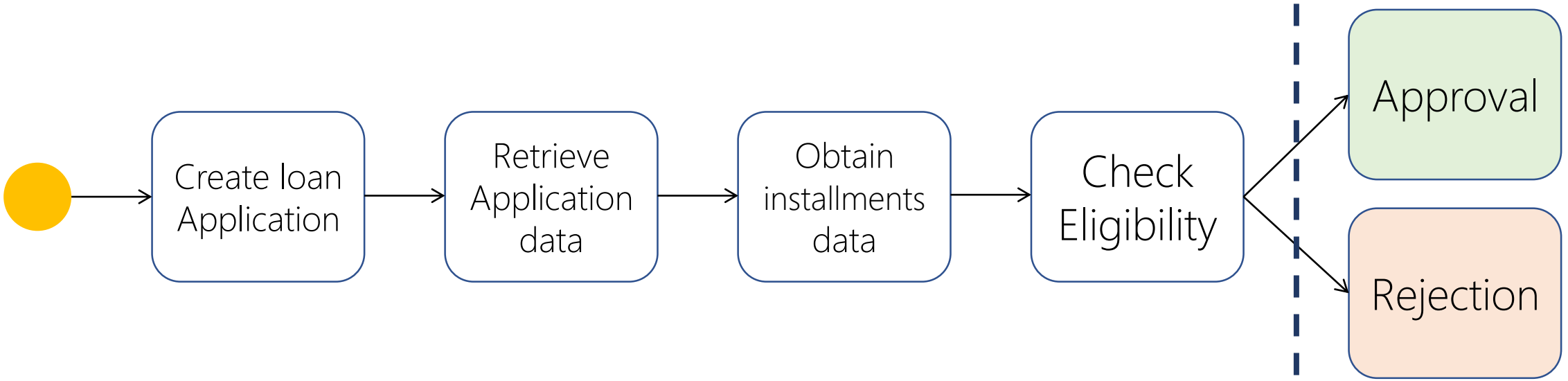
Suhwan Lee[1] Marco Comuzzi[2] Xixi Lu[1] Hajo A. Reijers[1]

[1]Utrecht University, Utrecht, The Netherlands
[2]Ulsan National Institute of Science and Technology

Outline

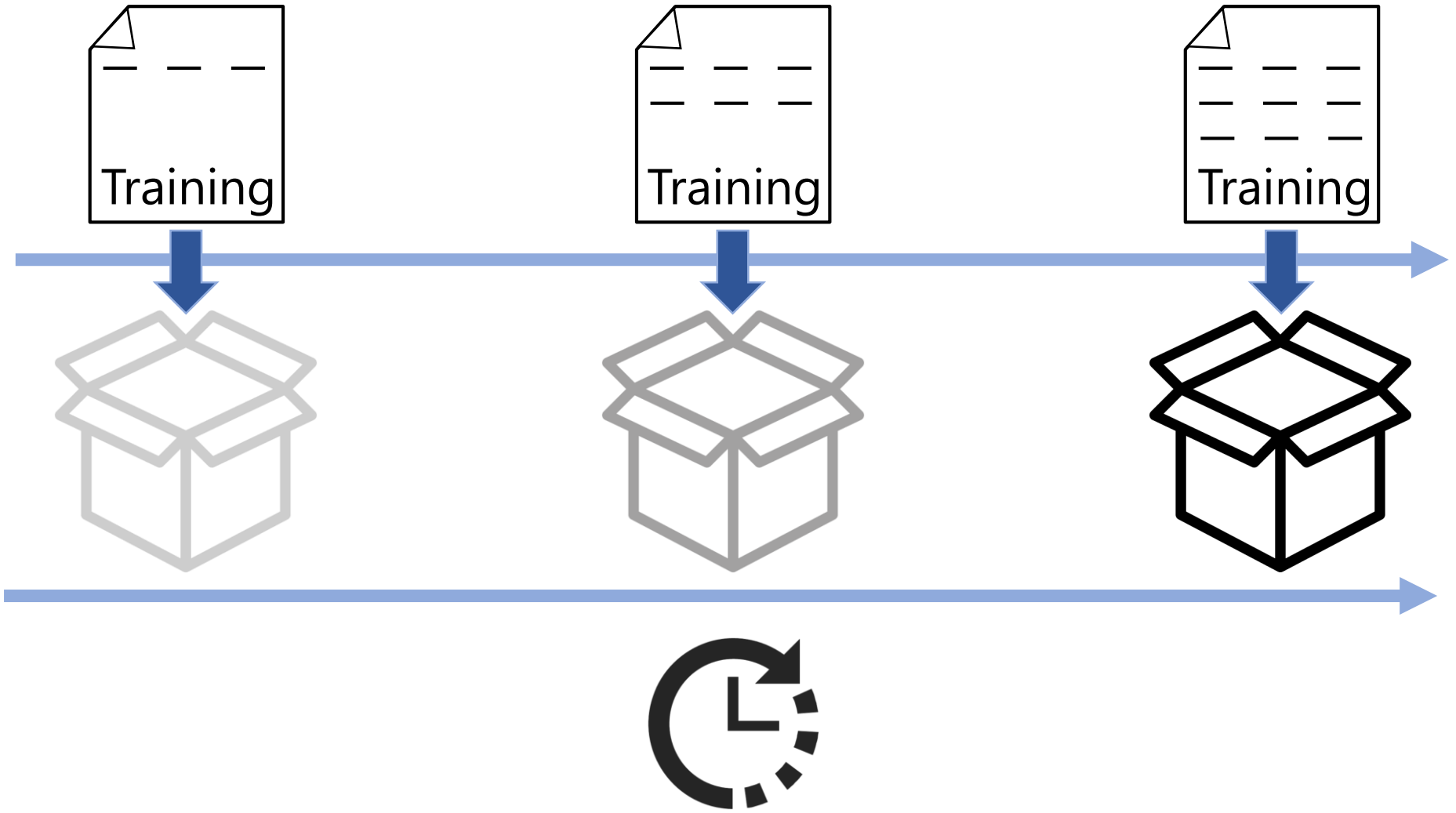
1. Introduction
2. Related works & Research question
3. Meta-measures
4. Experiment results
5. Conclusion & Future works

Predictive process monitoring

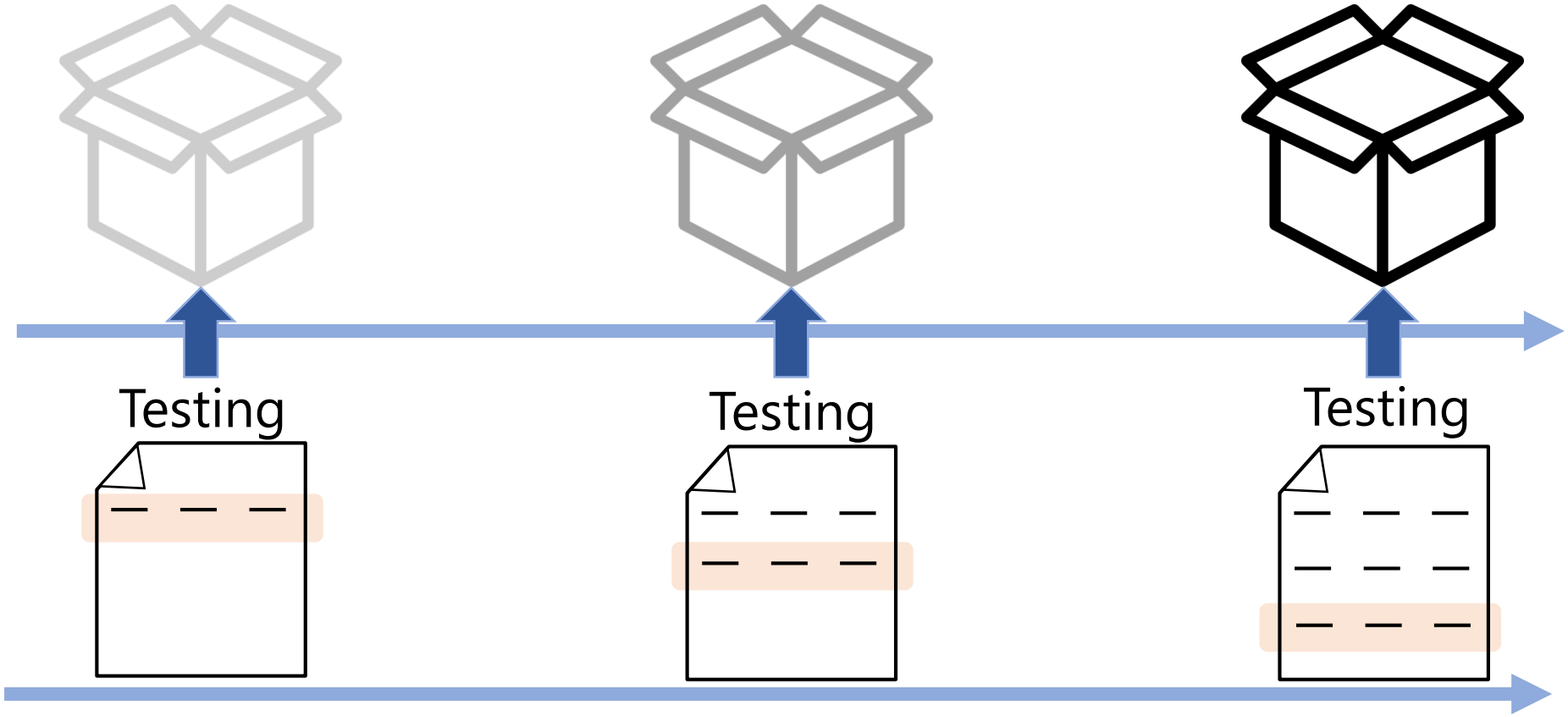


Prediction model

Online predictive process monitoring

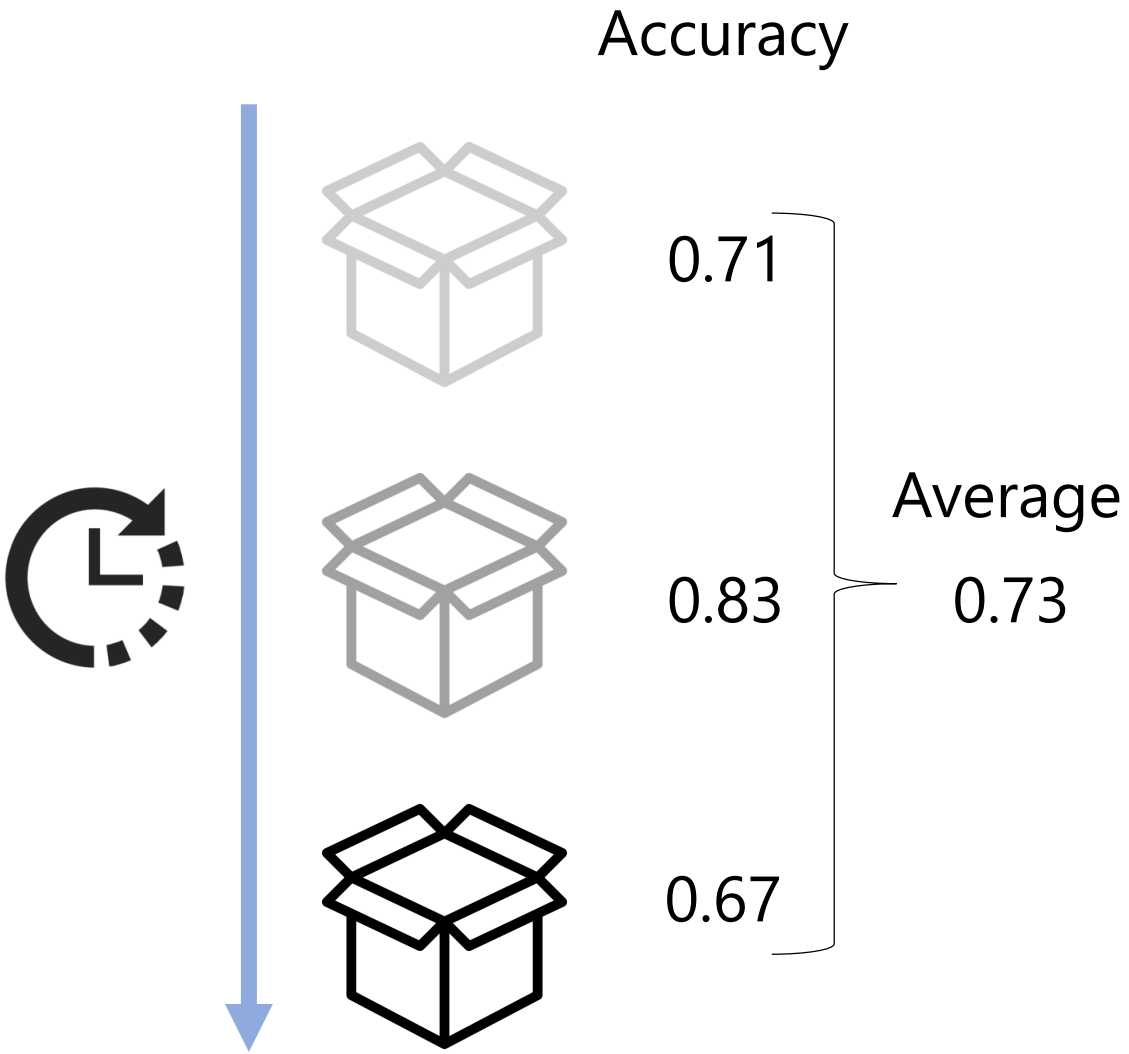


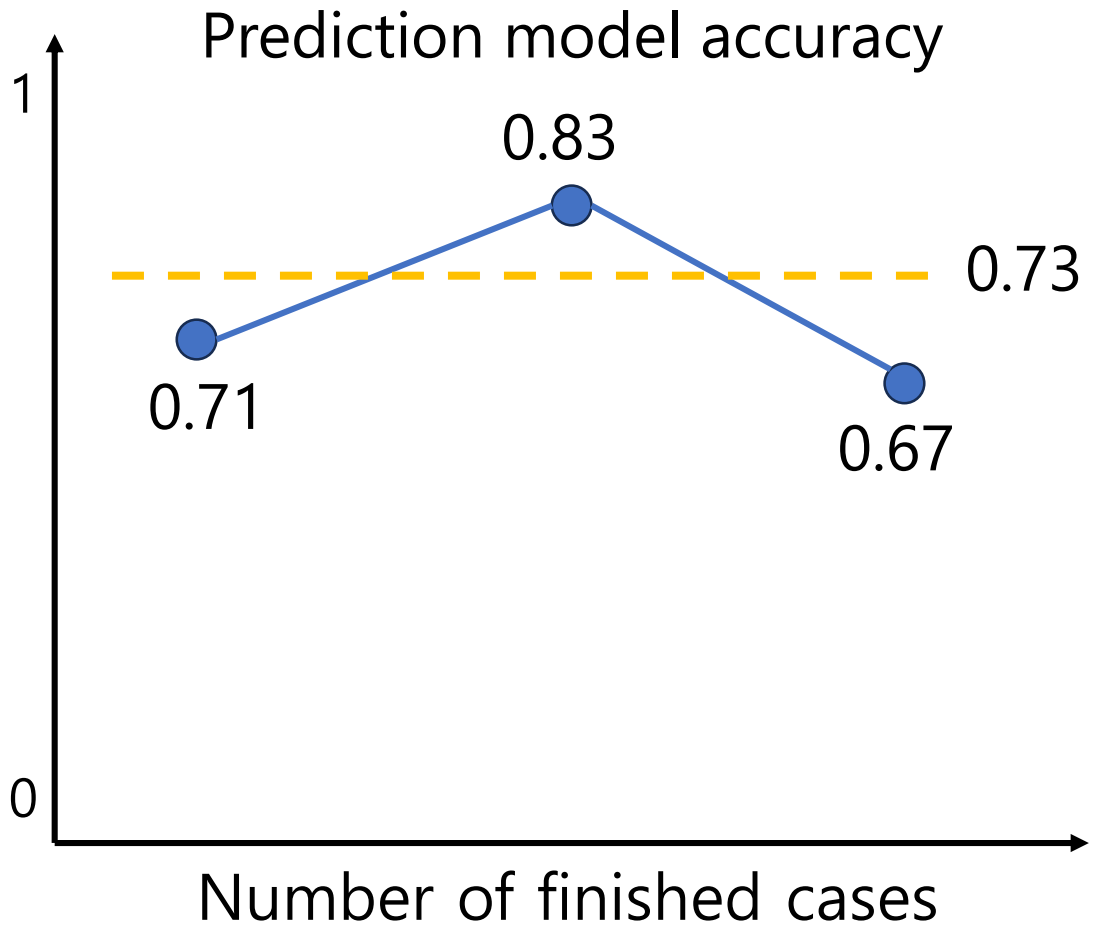
Model is updating?



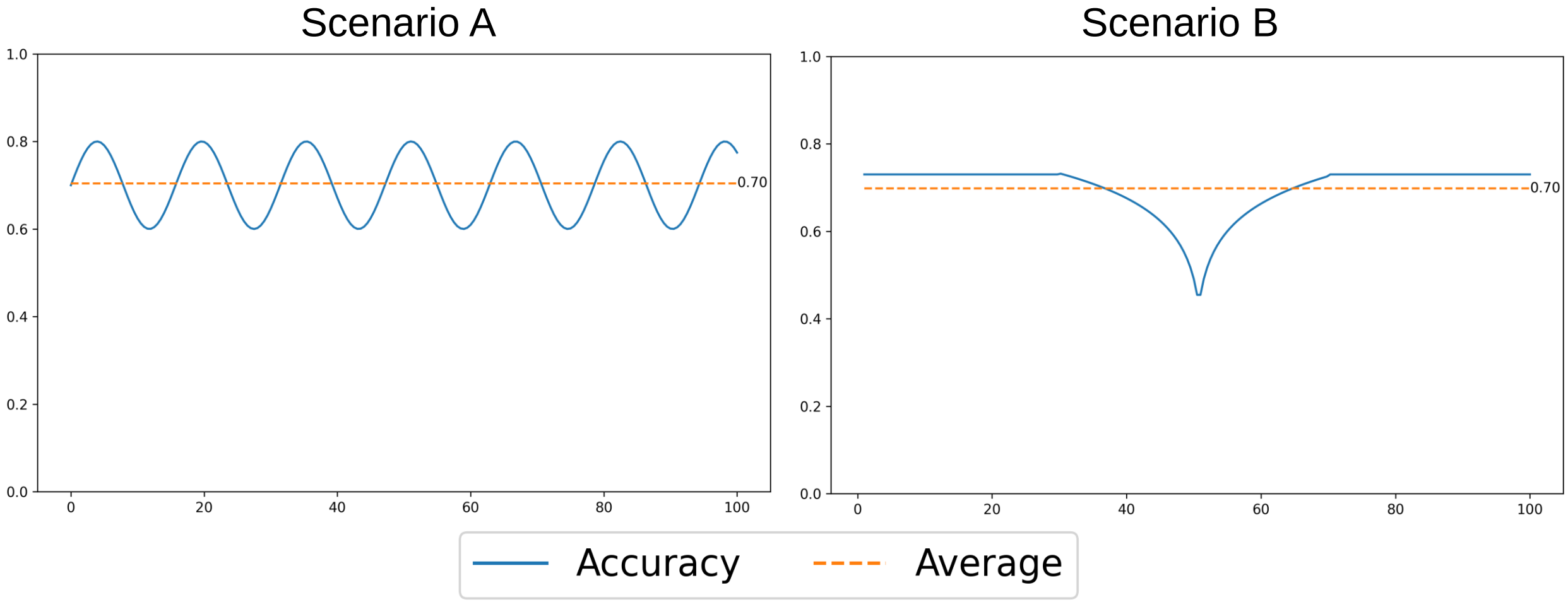
Model performance is also updated

Performance of online predictive model





What is the best model?

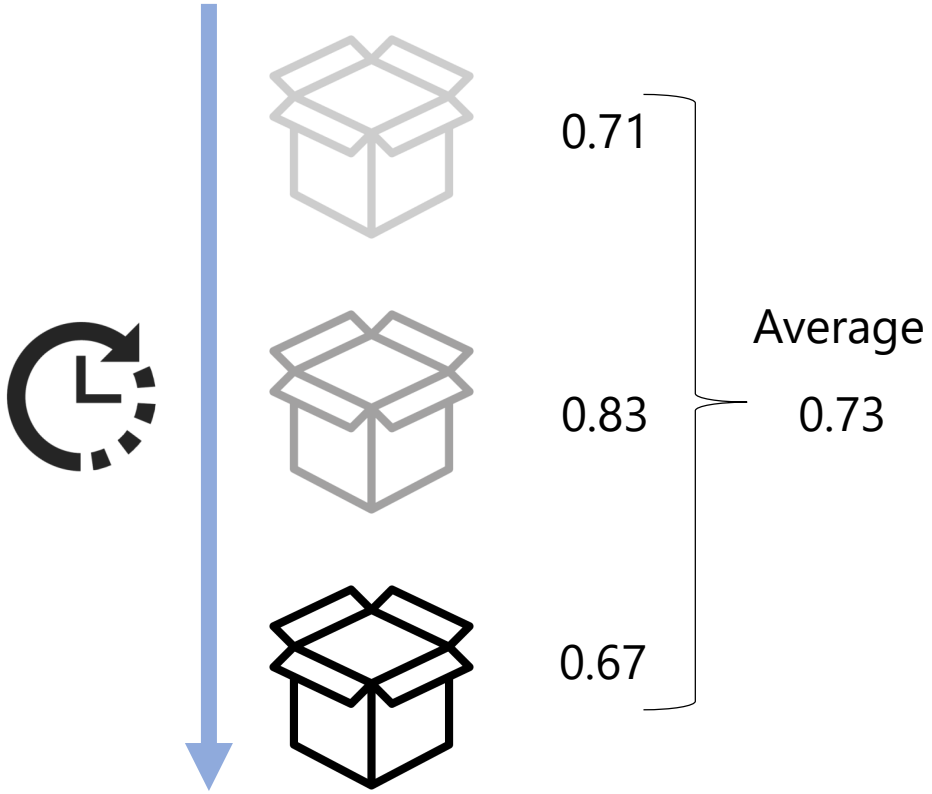


Research question

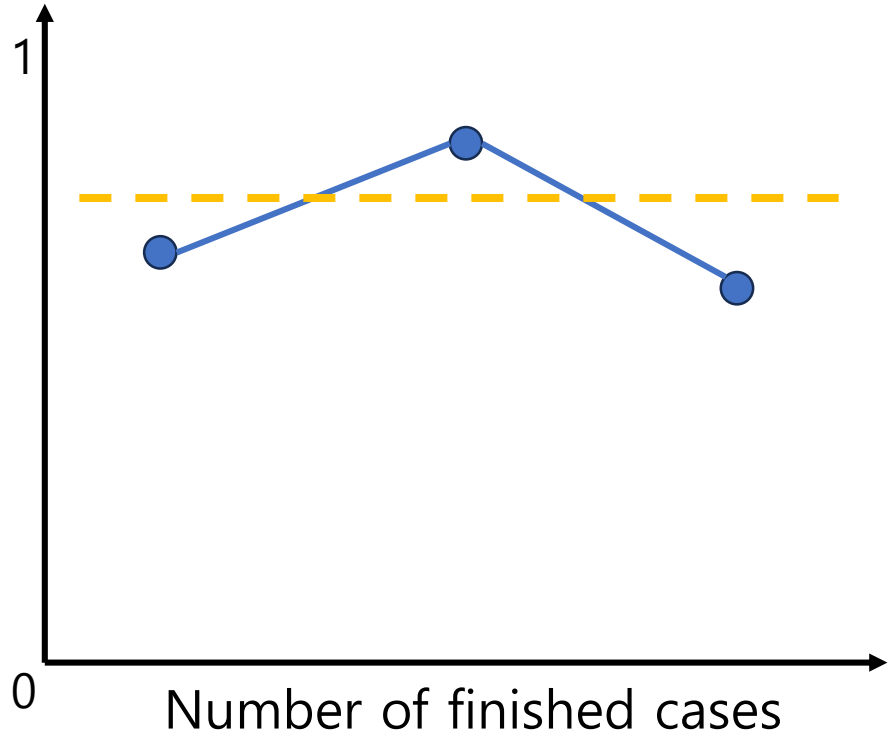
How to assess the **stability** of models for online predictive process monitoring?

Related works

Single aggregated value

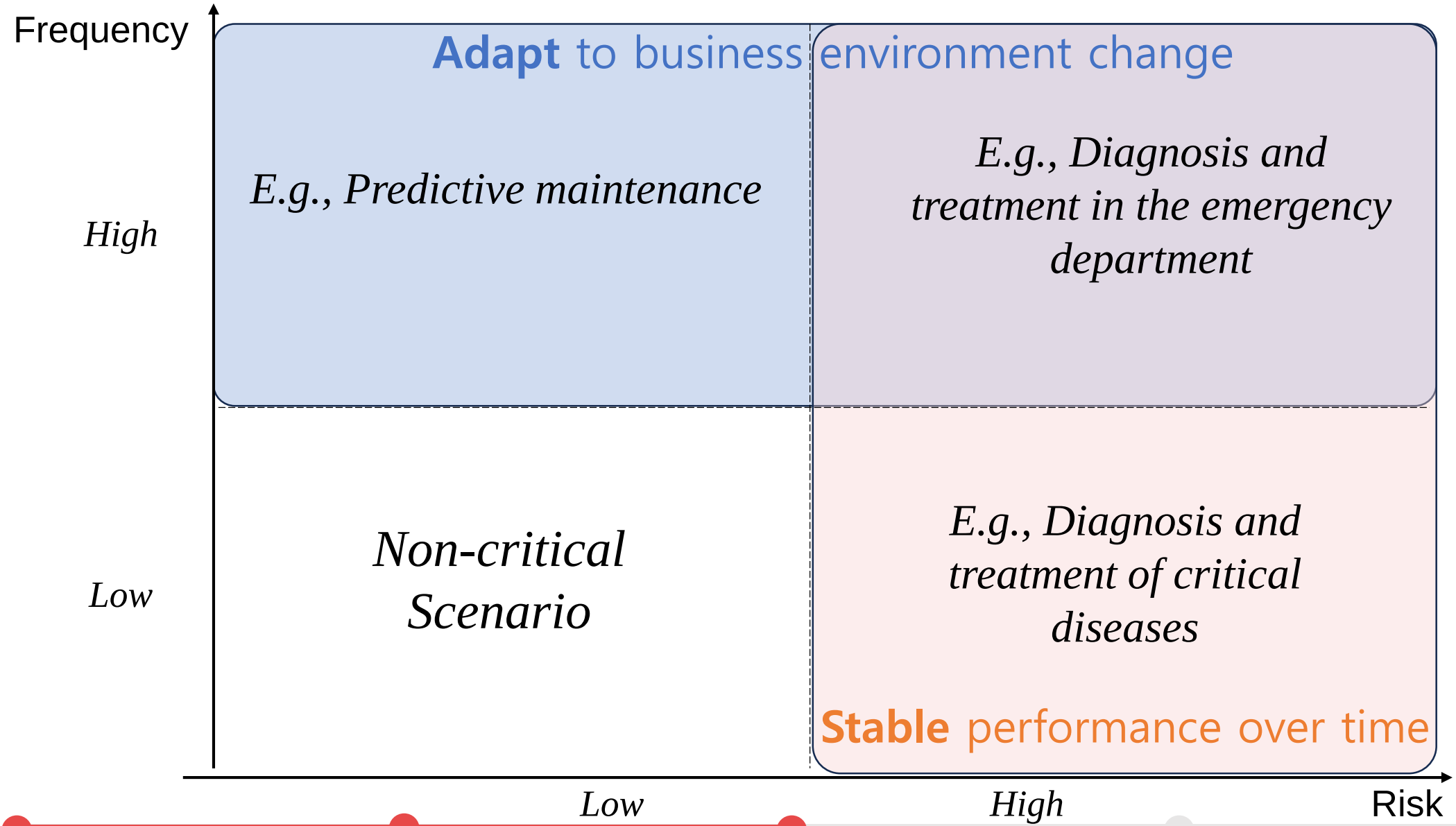


Time-series visualization

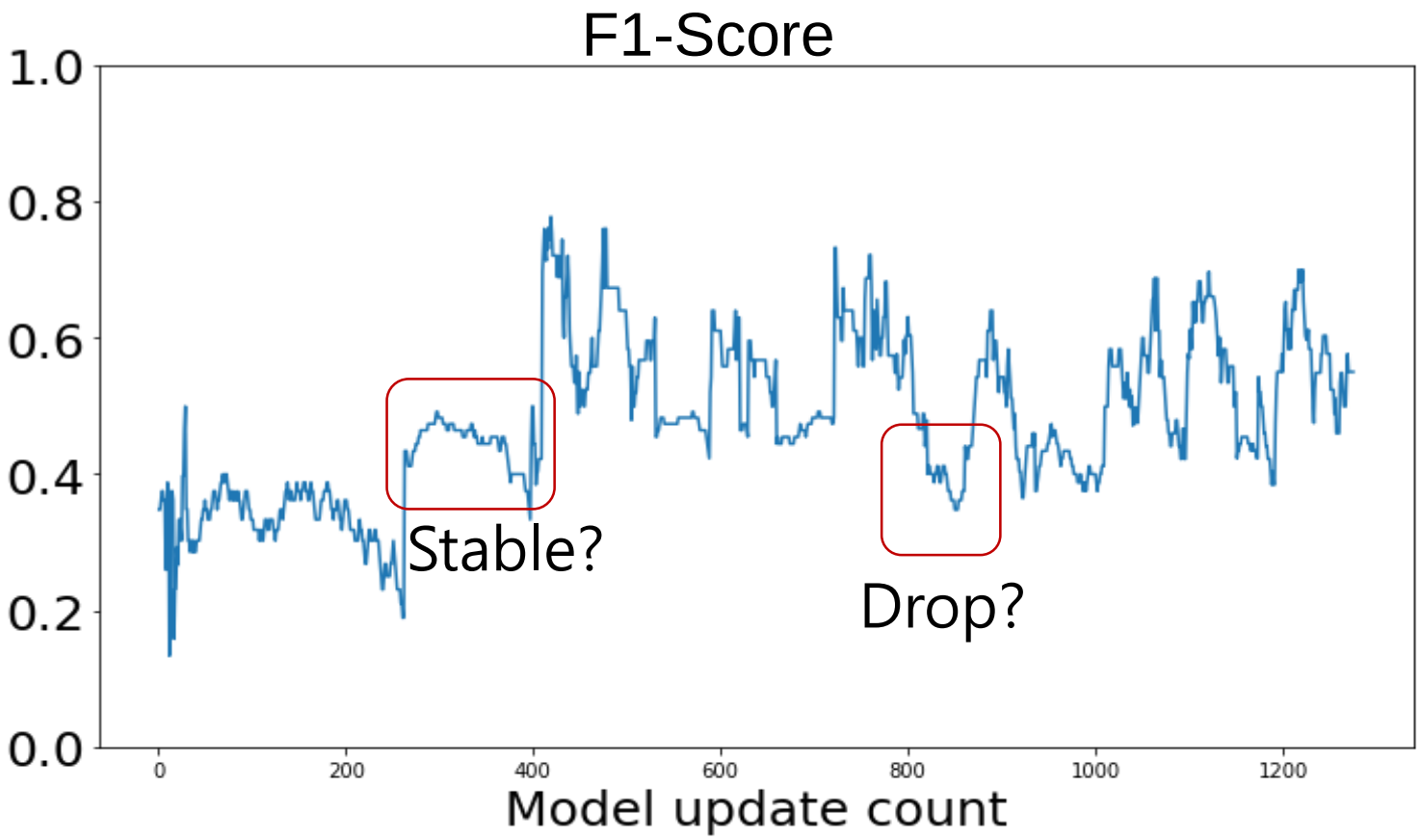


Do not assess the model's fluctuation in performance

Motivation - Business scenario



Continuous performance evaluation



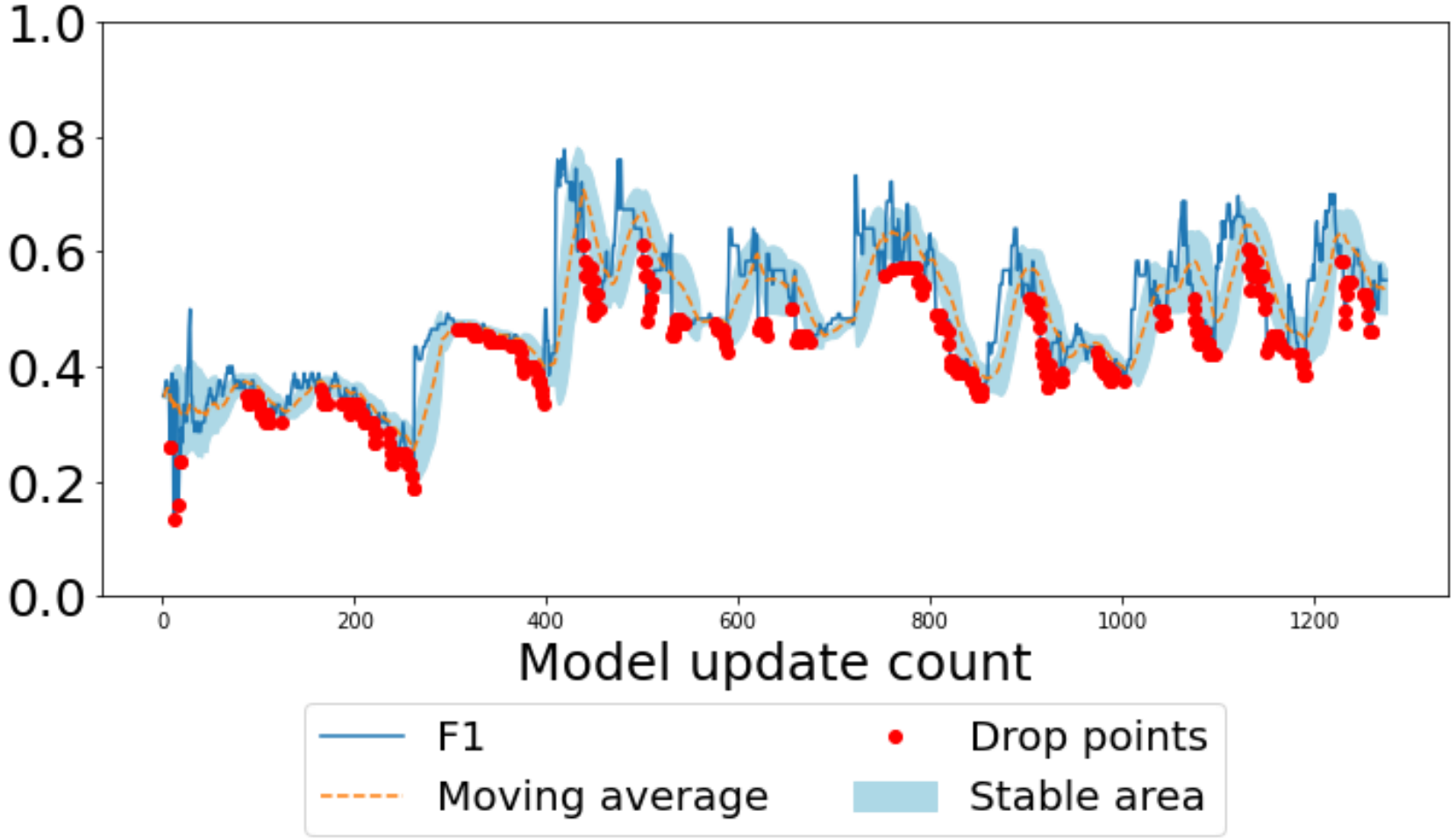
Continuous performance evaluation

F1-Score (p_t)	0.550	0.490	0.520	0.520	0.520	0.500
Moving Average (ma_t)		0.520	0.510	0.520	0.513	
Moving Standard deviation (φ_t)		0.030	0.017	0.000	0.012	

Stable area $= (ma_t - \varphi_t,\ ma_t + \varphi_t)$

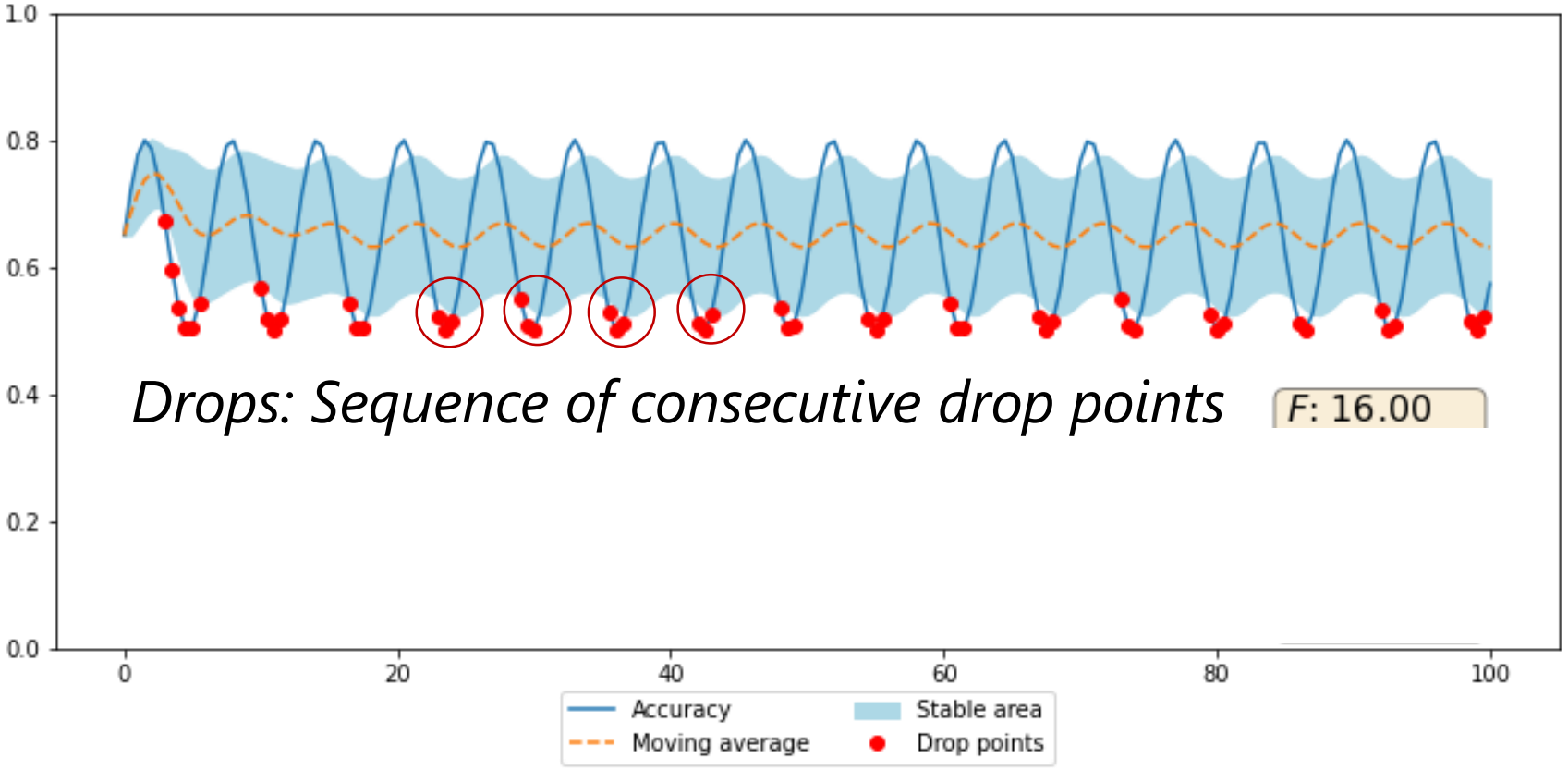
Drop point $(d_t) = p_t < (ma_t - \varphi_t)$

Continuous performance evaluation



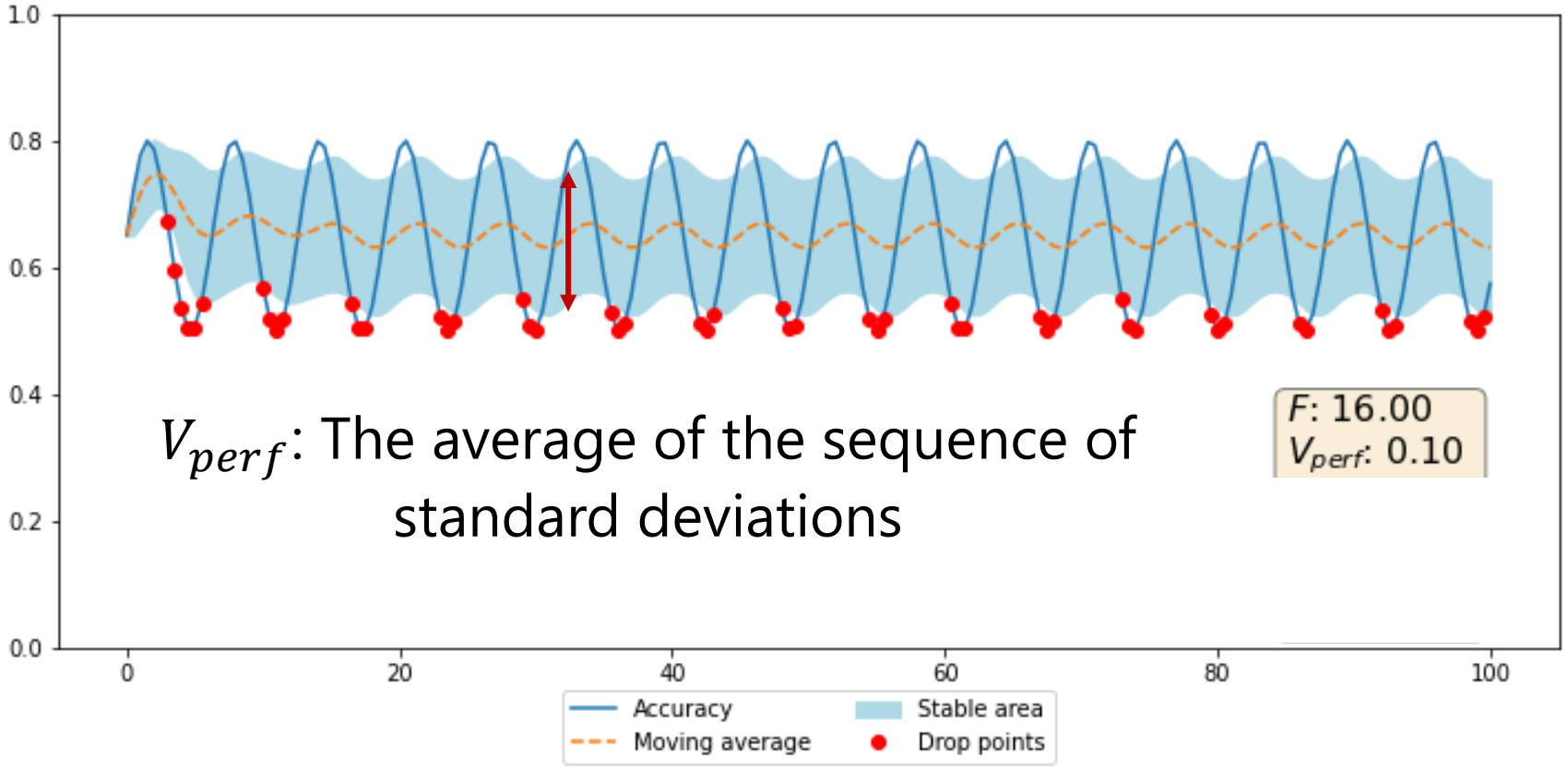
Meta-measures

1. Frequency of relevant performance drops (F)



Meta-measures

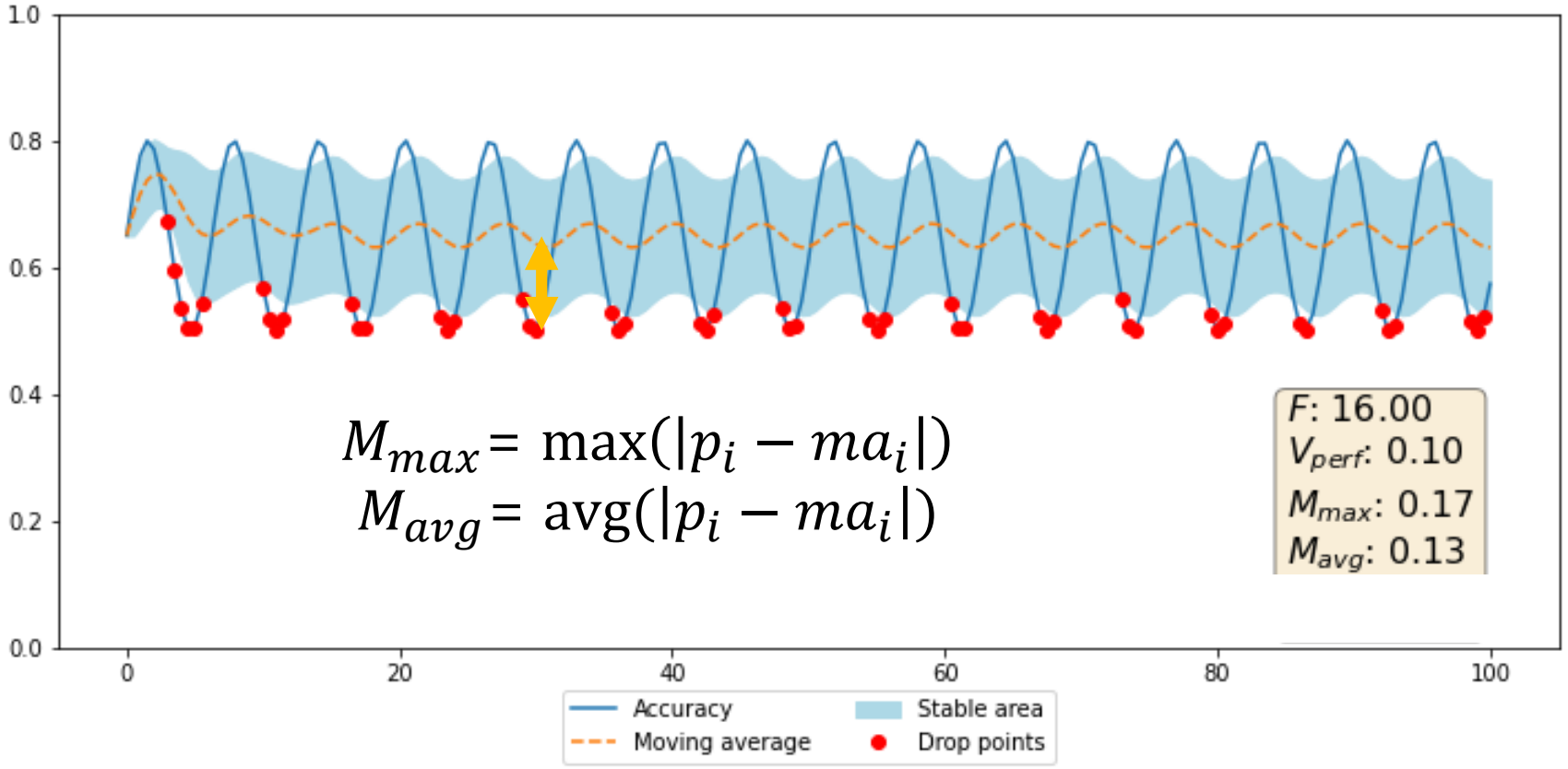
2. Volatility of the performance (V_{perf})



V_{perf}: The average of the sequence of standard deviations

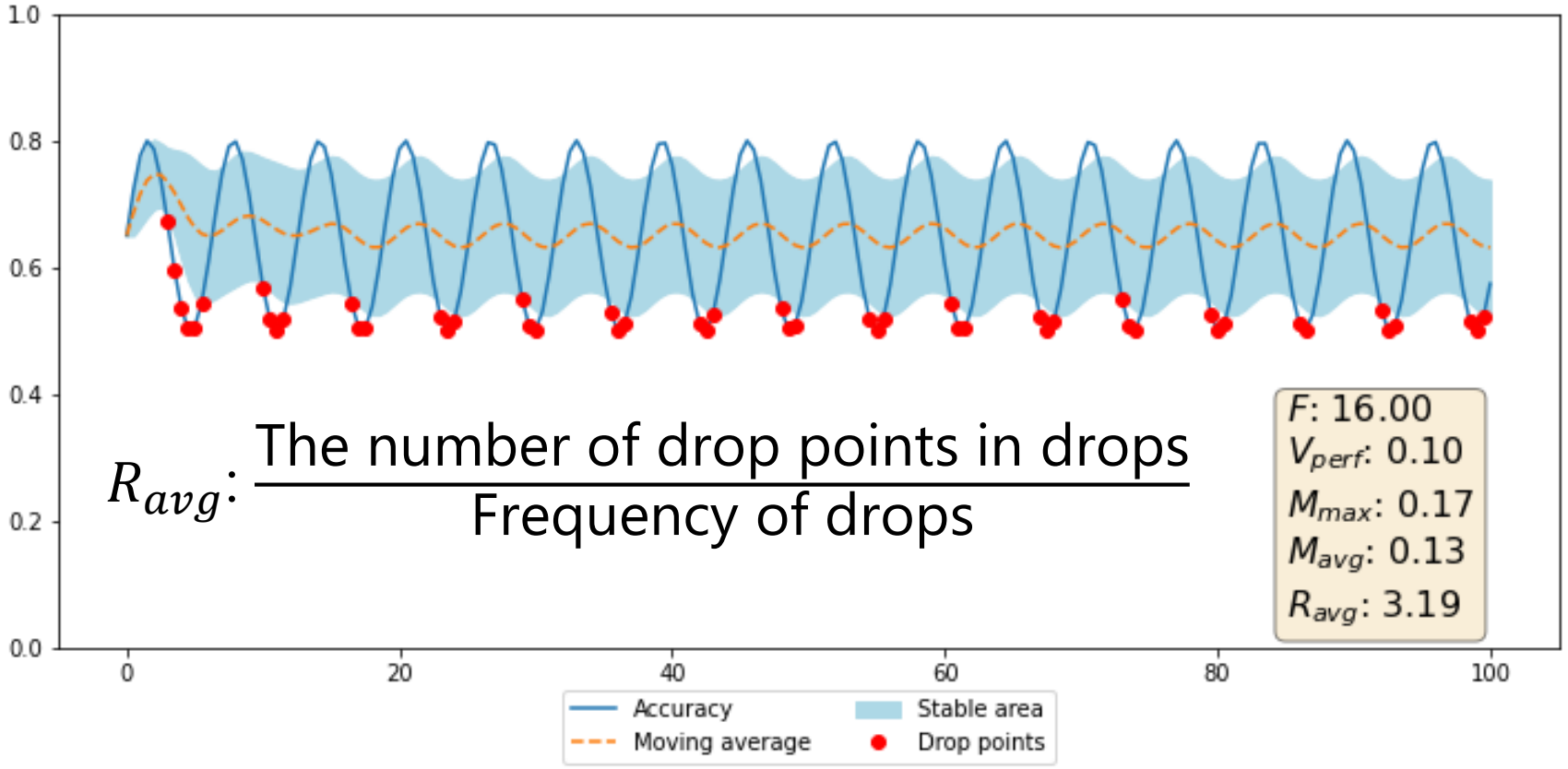
Meta-measures

3. Magnitude of performance drop ($M_{max,avg}$)



Meta-measures

4. Recovery rate (R_{avg})

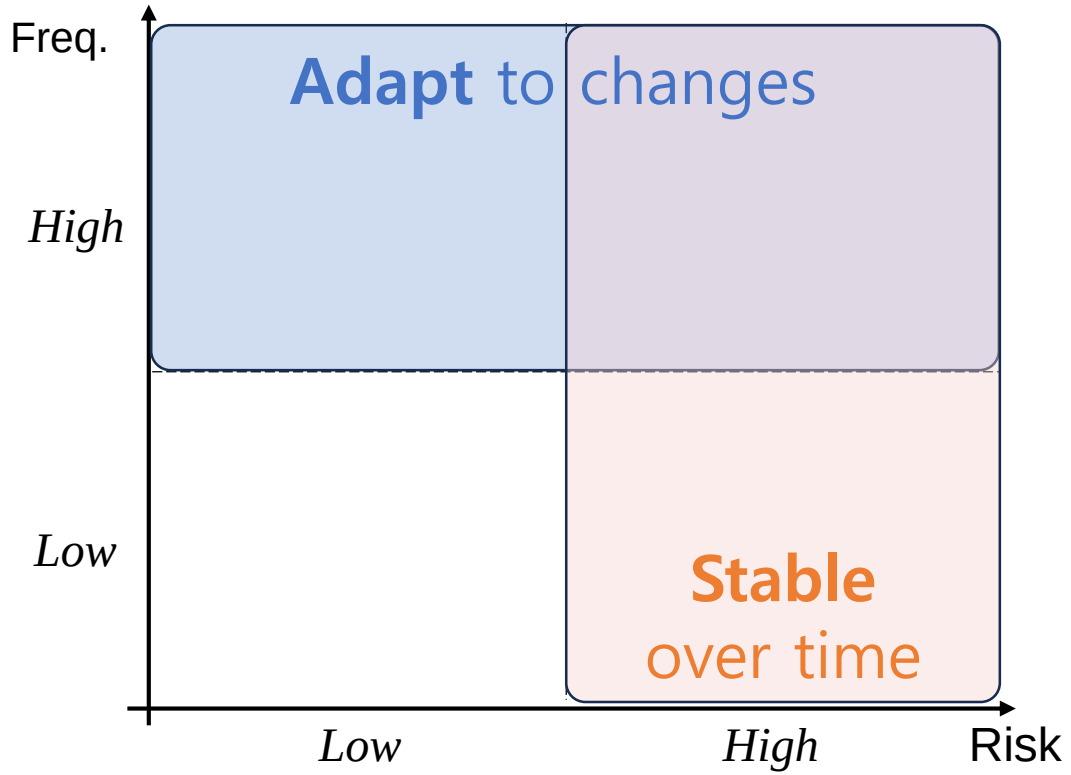


$$R_{avg}: \frac{\text{The number of drop points in drops}}{\text{Frequency of drops}}$$

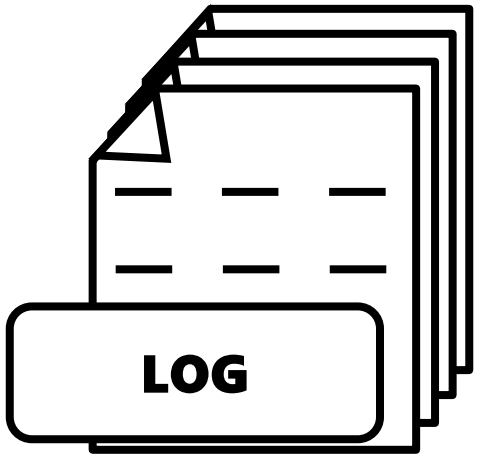
Experiment setting

How to use the meta-measures?

Let's look at the business scenarios again

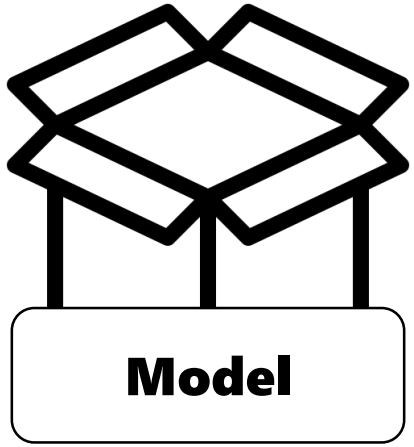


Experiment setting

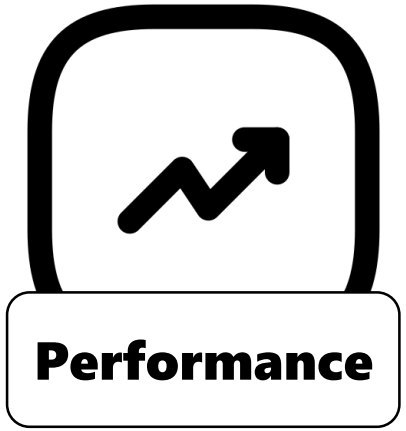


Two real-life logs
- BPIC 2015 & BPIC 2017

Two synthetic logs
- Different concept drift

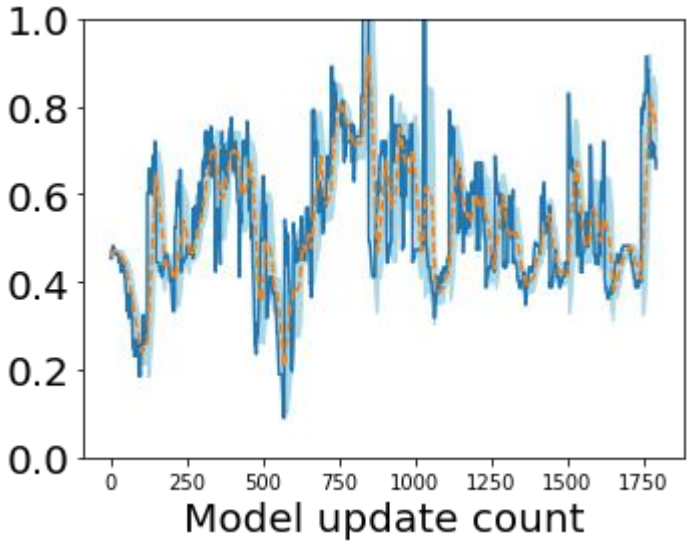
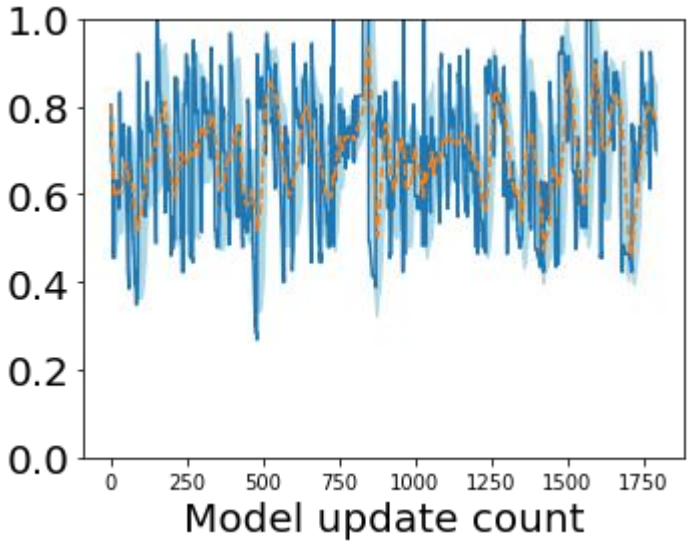
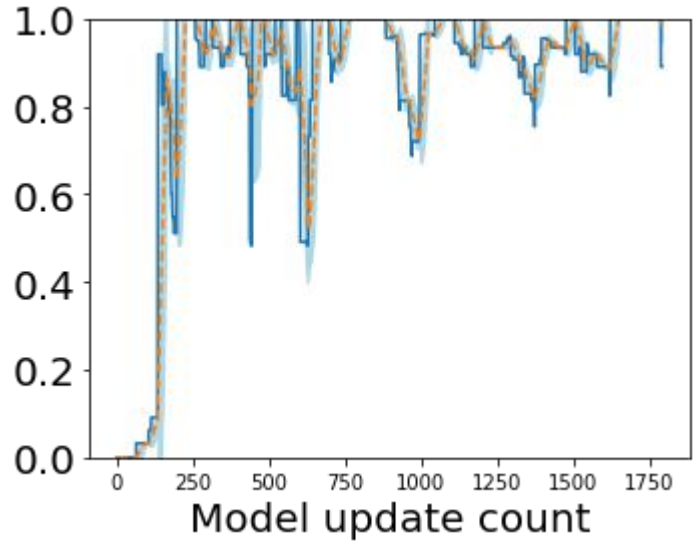


Three algorithms
(Binary outcome prediction)
- Incremental (HATC)
- Sliding window (XGB)
- Train-once (LSTM)

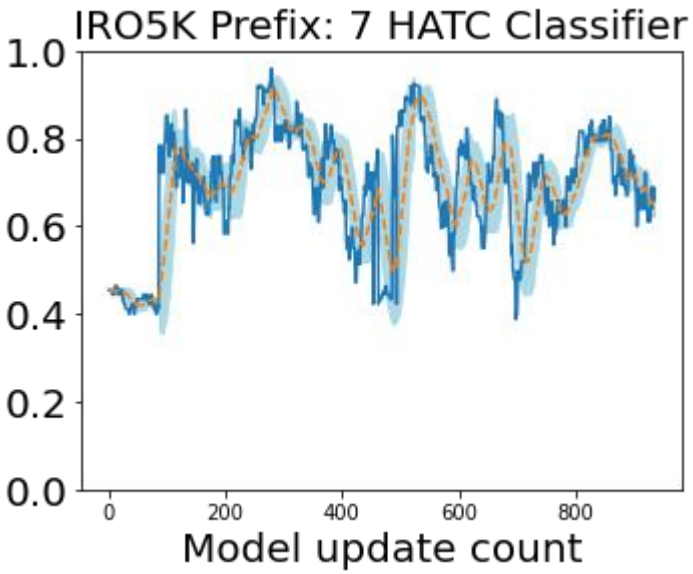


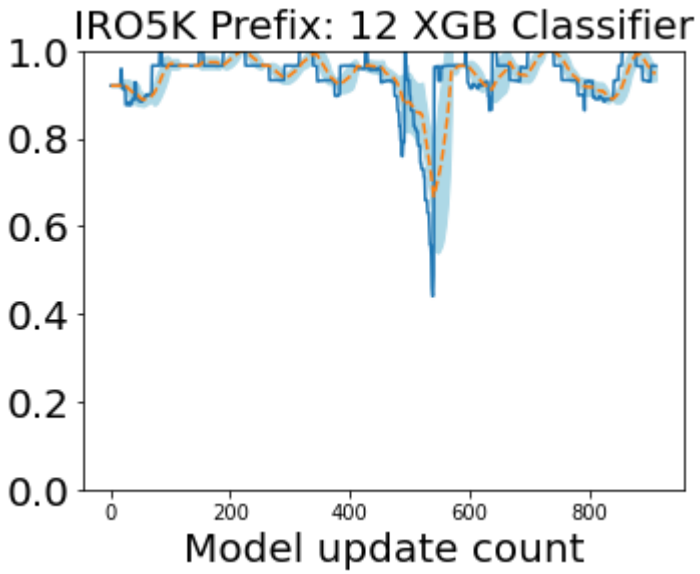
Four measures
- Accuracy
- Precision, Recall, & F1-Score

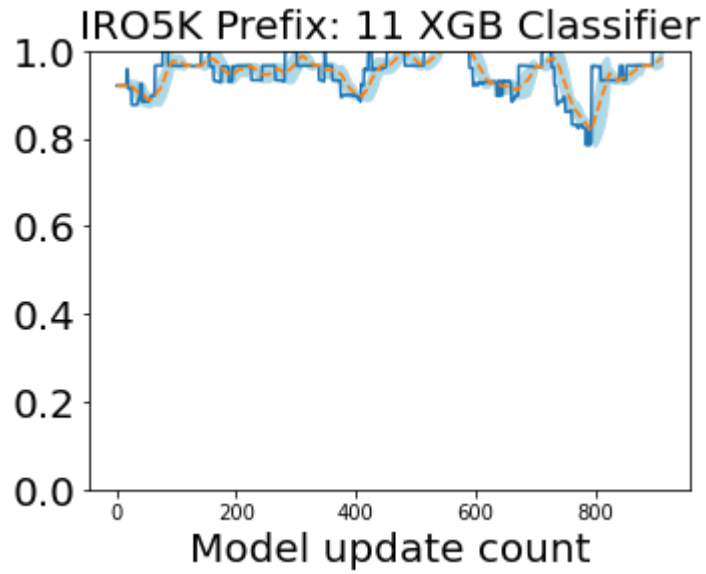
Result

	BPIC17 Prefix 2 XGB	BPIC17 Prefix 7 XGB	BPIC17 Prefix 14 XGB
Average F1-Score	0.54	0.69	0.86
			
Freq. of Drops	54	67	31
Volatility of perf.	0.074	0.114	0.037
Avg. Magnitude	0.115	0.166	0.068
Recovery rate	8.556	6.776	13.194

Result

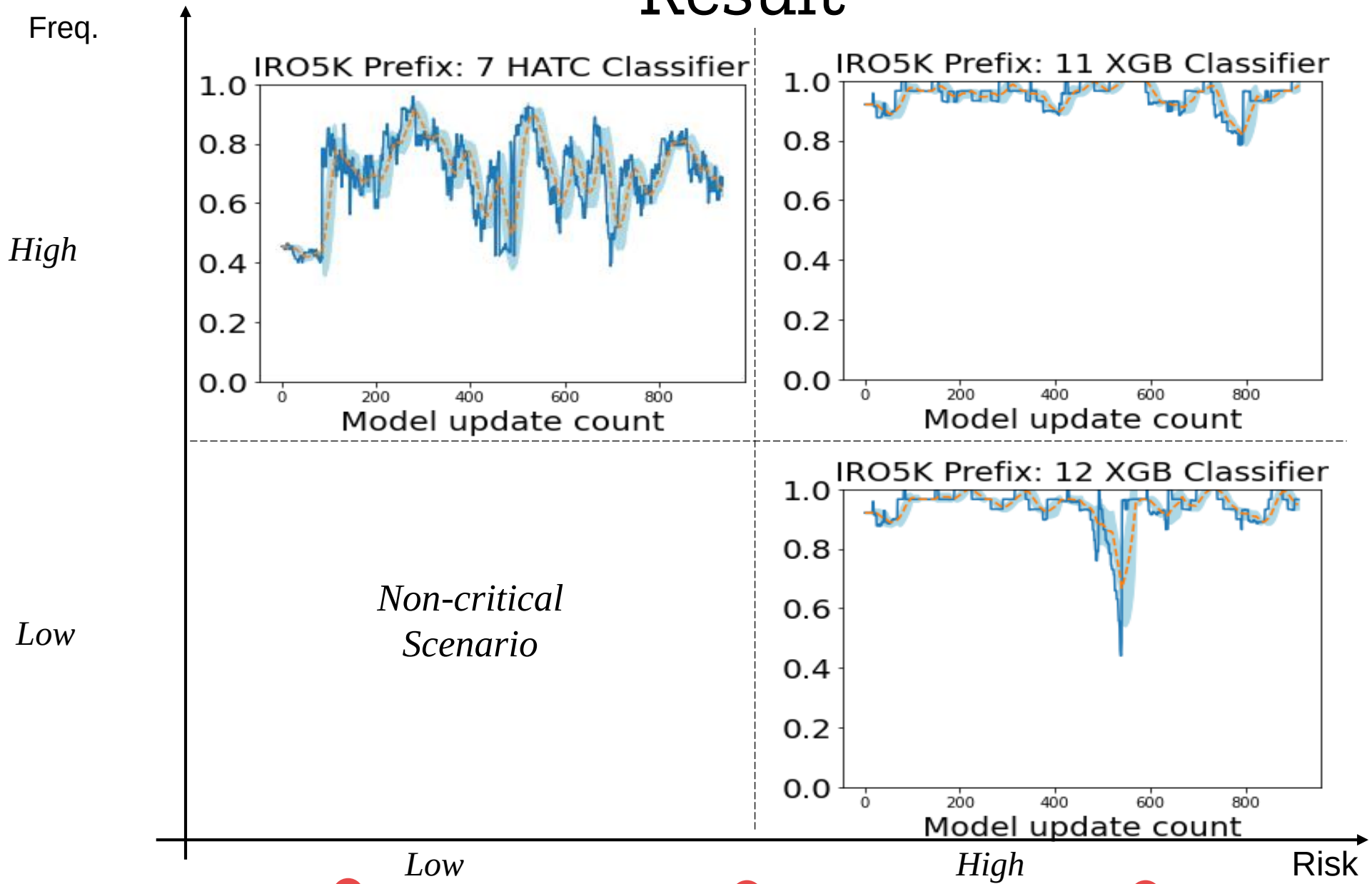






Freq. of Drops	44	24	26
Volatility of perf.	0.06	0.02	0.02
Max. Magnitude	0.29	0.25	0.10
Recovery rate	6.57	11.04	7.69

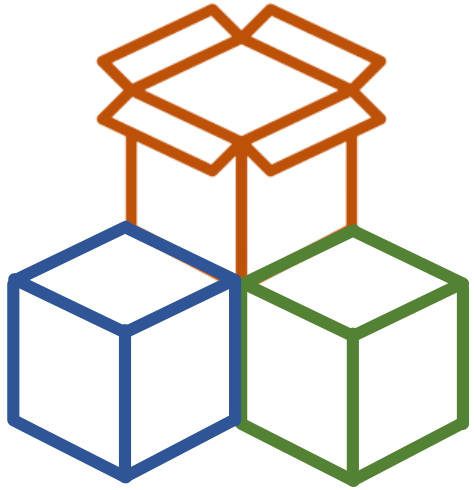
Result


Freq.
High
Low
IRO5K Prefix: 7 HATC Classifier
IRO5K Prefix: 11 XGB Classifier
IRO5K Prefix: 12 XGB Classifier
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
Model update count
Non-critical Scenario
Low
High
Risk

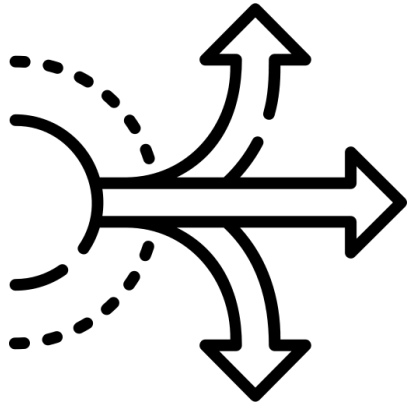
Conclusion & Future works

1) We develop **Meta-measures** for online process outcome predictive monitoring

2) We assess the **performance stability** in various business scenarios



In-depth analysis with benchmark test



Uncover the causes of the performance drop

Thank you