Event Log Sampling for Predictive Process Monitoring

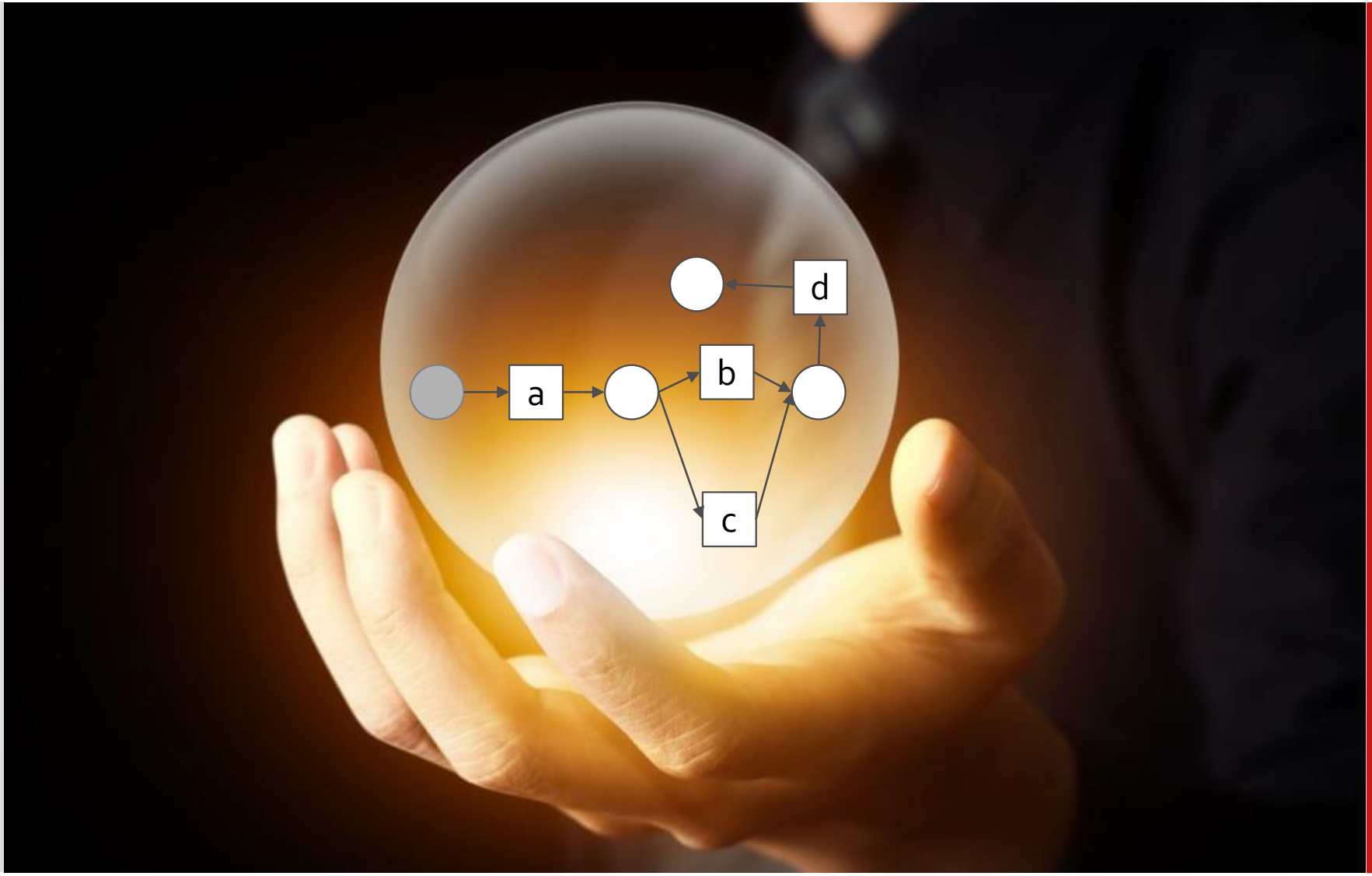
Authors:

Mohammadreza Fani Sani,
Mozhgan Vazifehdoostirani,
Gyunam Park,
Marco Pegoraro,
Sebastiaan J. van Zelst,
and Wil M.P. van der Aalst

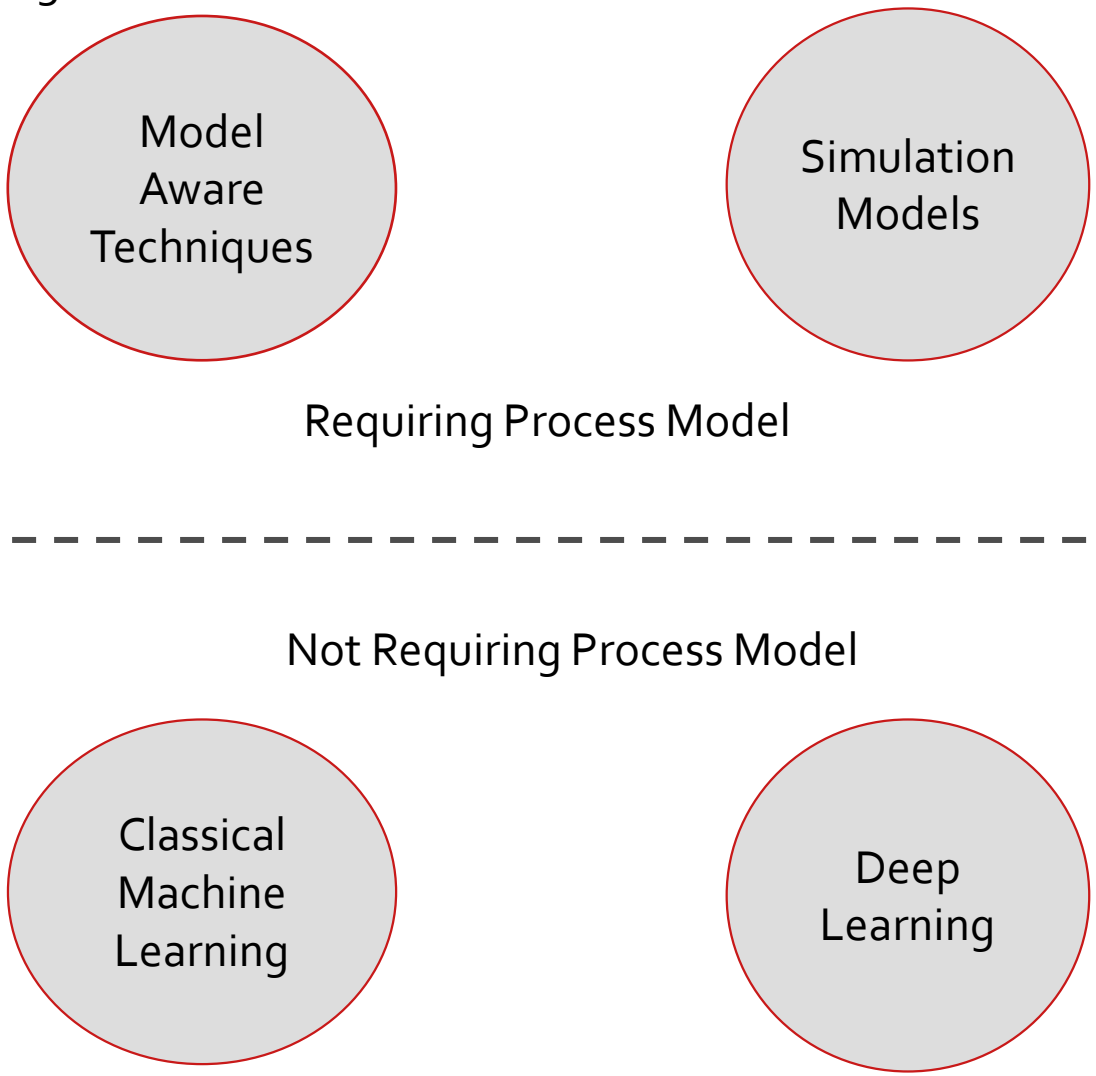
Presenter:
Mozhgan Vazifehdoostirani

m.vazifehdoostirani@tue.nl

What would happen in the future?



It is an extremely active field of research!



Motivation!

Expensive computational costs!

- Hyperparameters tunning
- Repeating training with different parameters
- Changes in dynamic processes
- Hardware limitation

Research Question

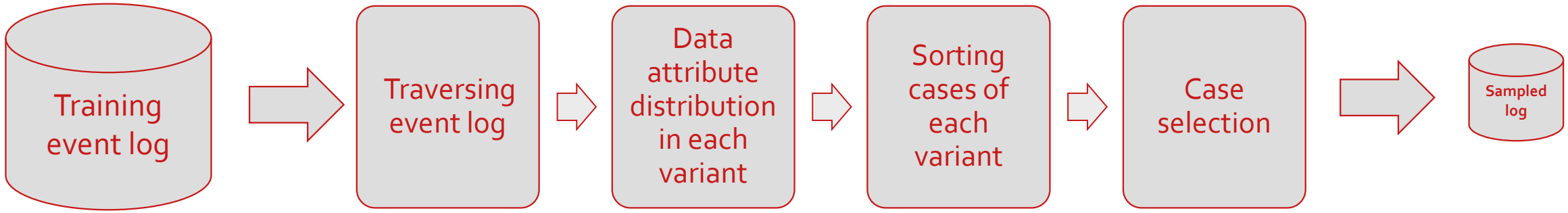
Can we improve the computational performance by using sampled event logs, while maintaining the accuracy?

Our Hypothesis

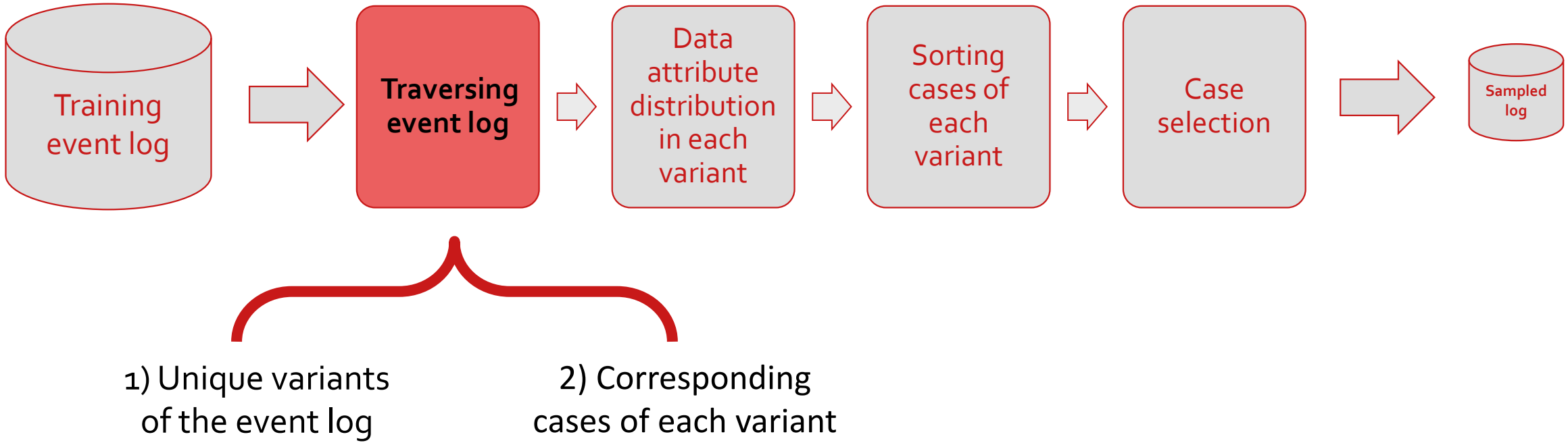
Select the best Candidates
Reduce the training time

Maintain accuracy
We do not loose too much information

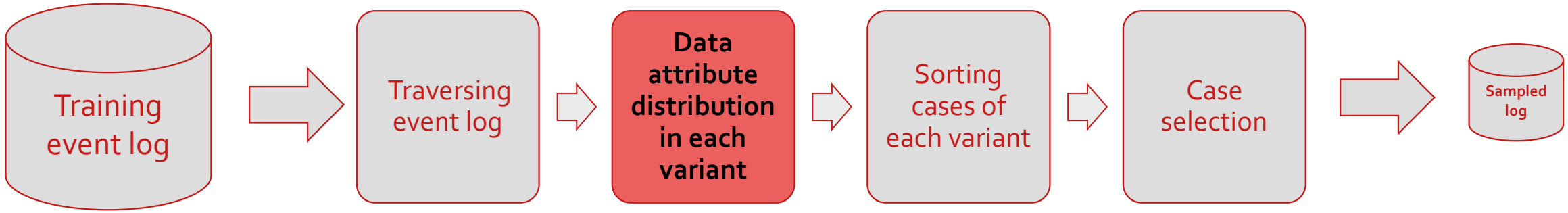
Proposed Sampling Procedure

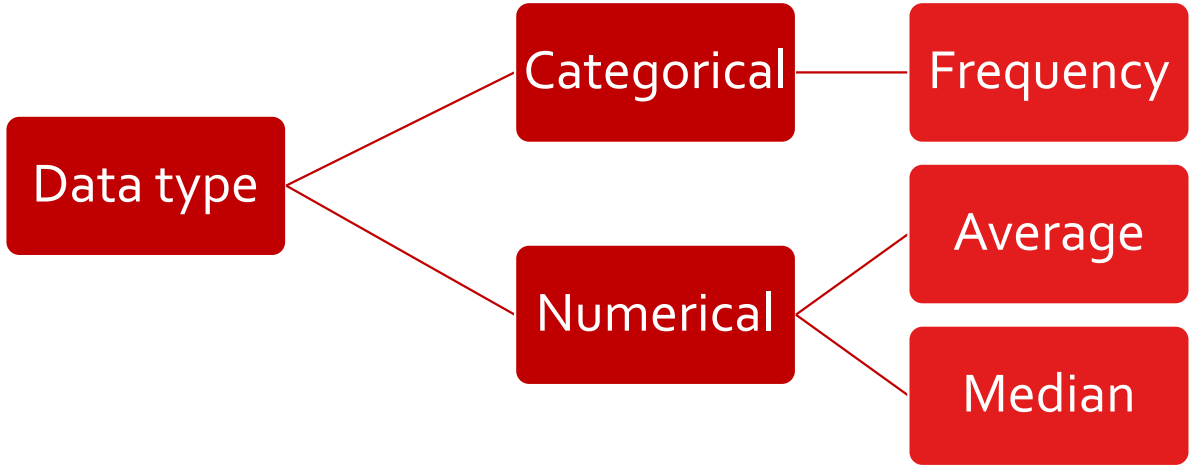


Proposed Sampling Procedure

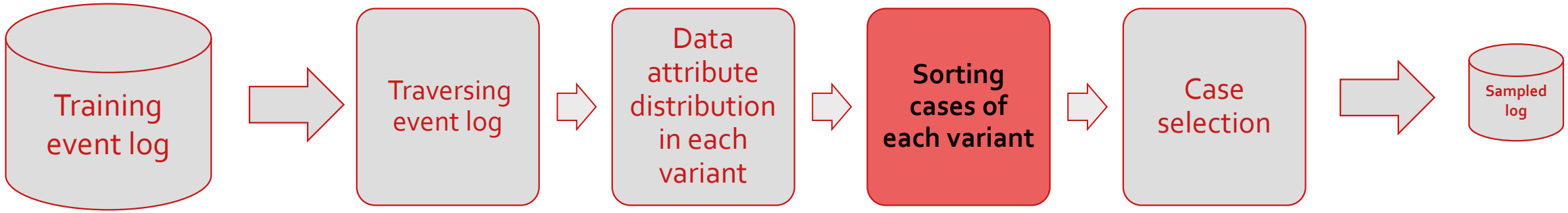


Proposed Sampling Procedure

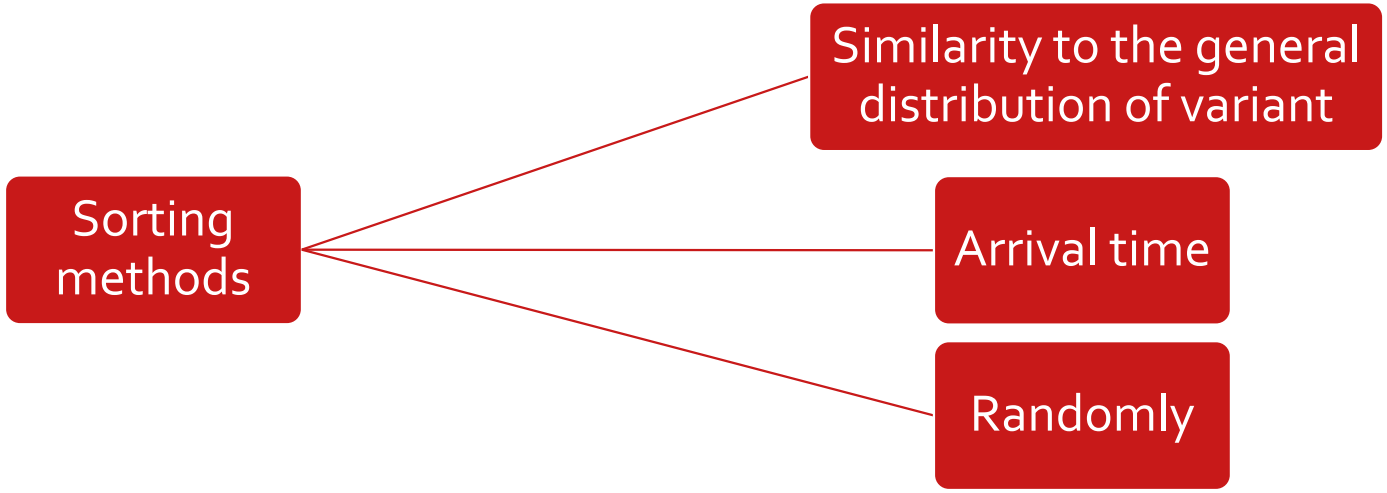




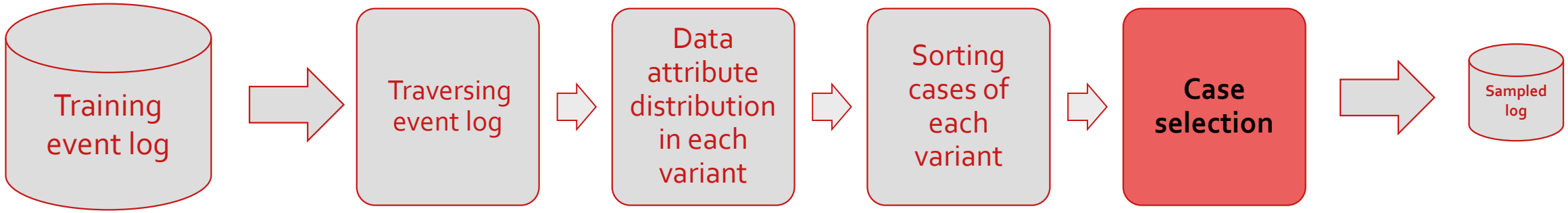
Proposed Sampling Procedure



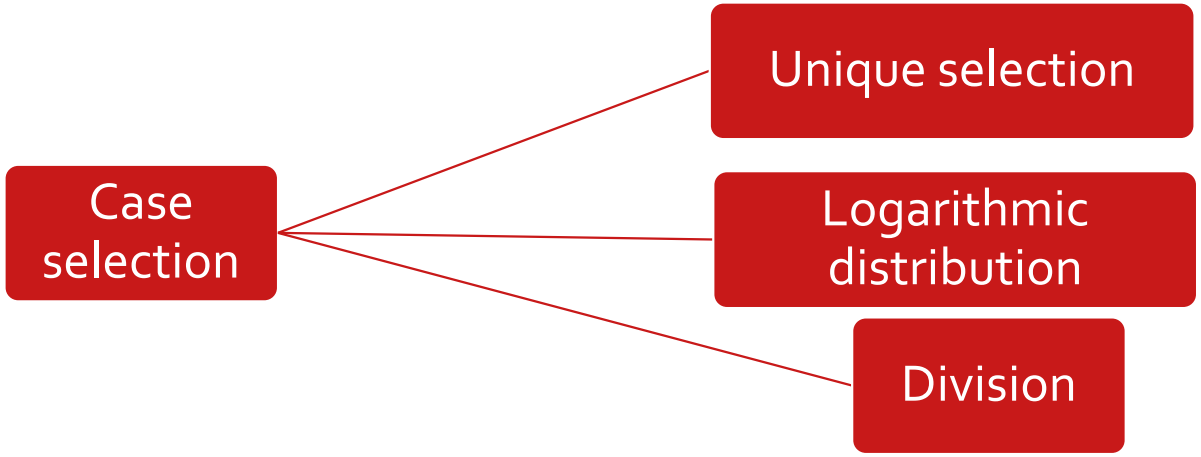
Which cases can represent their variant better?



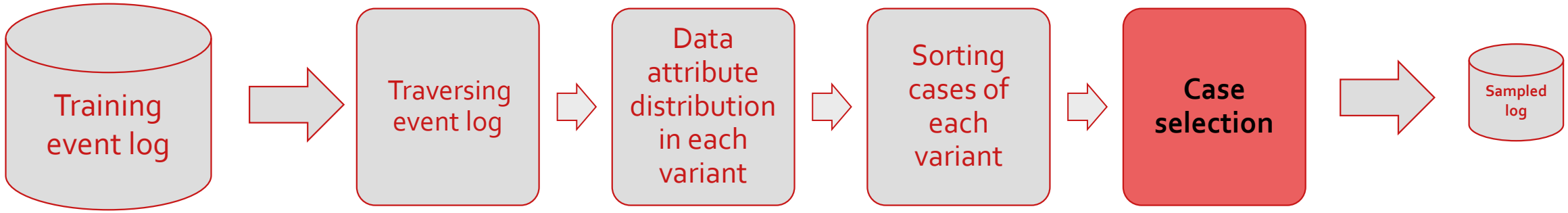
Proposed Sampling Procedure



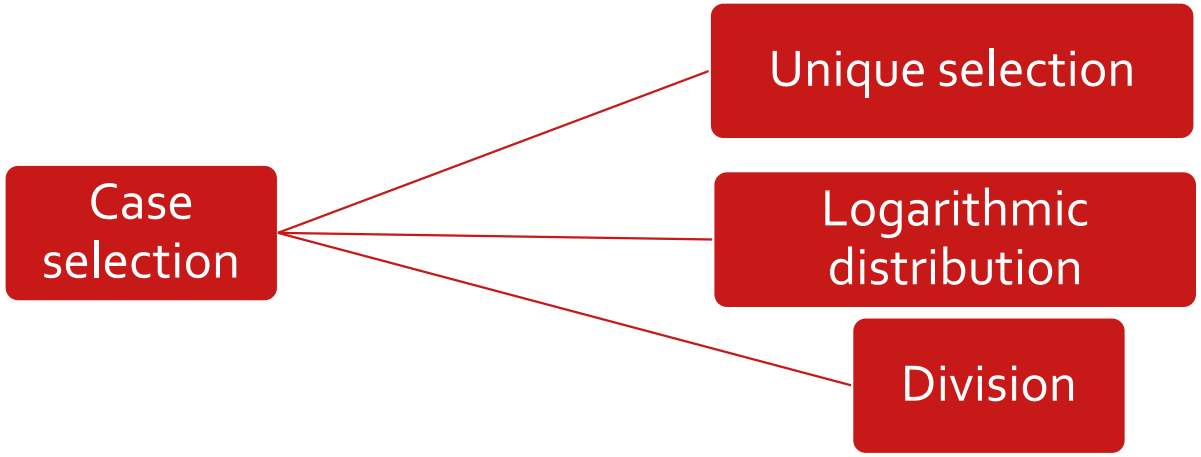
How many cases should be chosen?



Proposed Sampling Procedure



How many cases should be chosen?



If we have 100 cases from a variant:

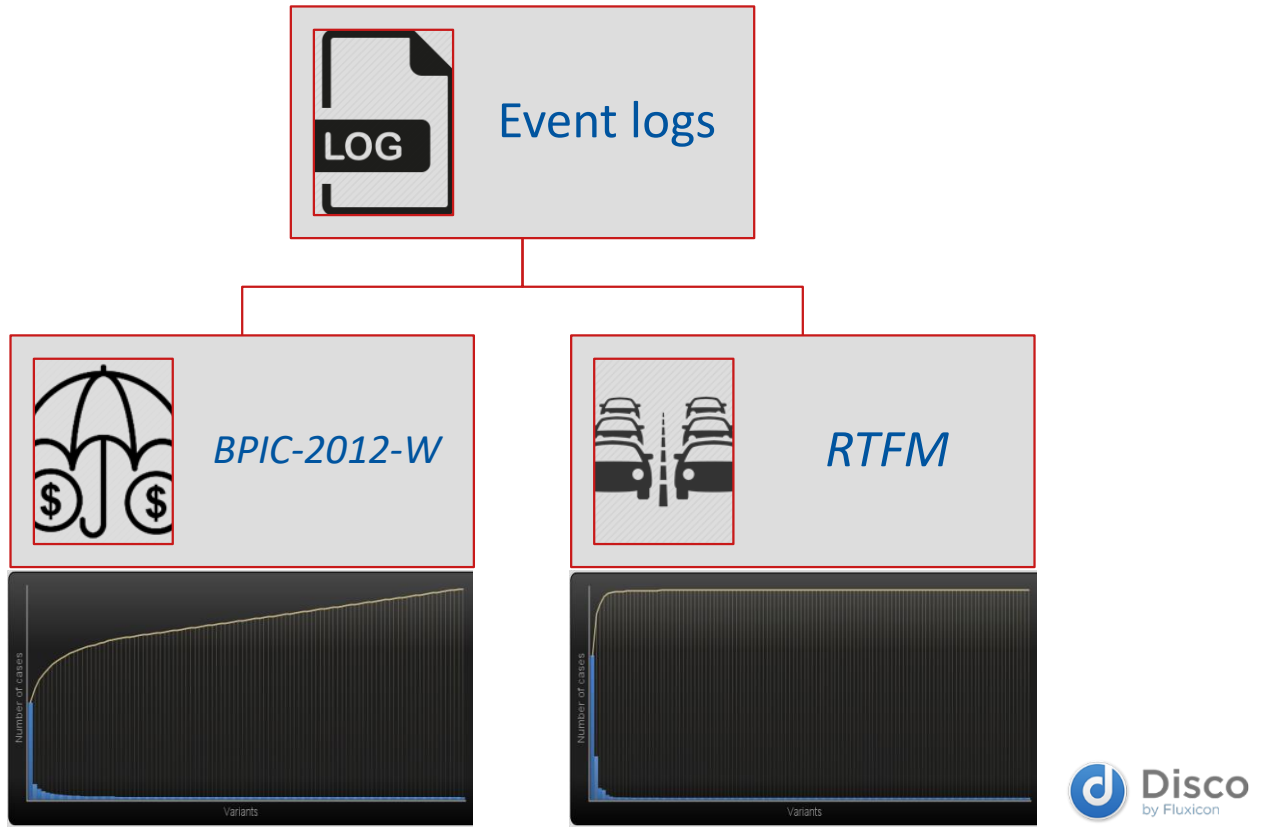
➡ Unique selection ➔ 1 case

➡ Logarithmic of base k ➔ $[log_k^{100}]$ cases

➡ Division by k ➔ $\left\lceil \frac{100}{k} \right\rceil$ cases

Evaluation

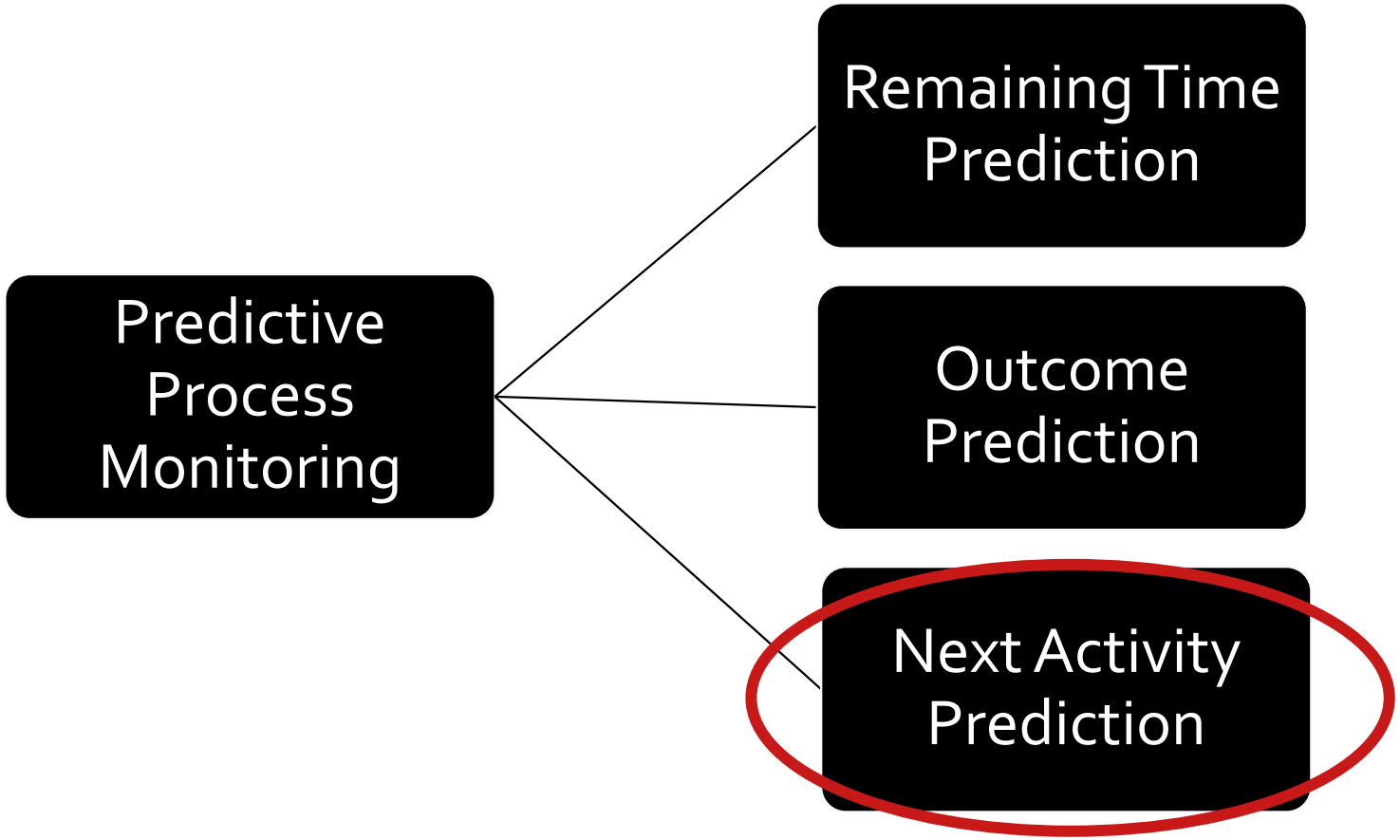
1) Datasets



Event Log	Cases	Activities	Variants	Attributes	FE Time
RTFM	150370	11	231	1	73649 s
BPIC-2012-W	9658	6	2643	2	1212 s

Evaluation

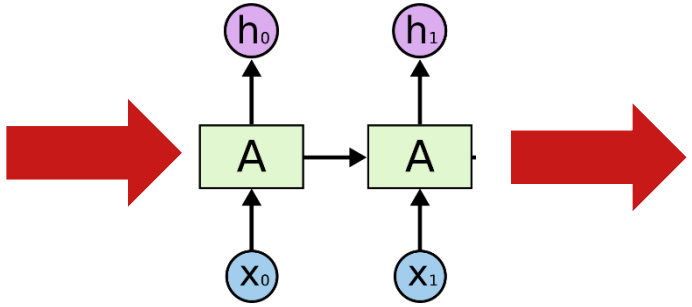
2) Prediction objective



Evaluation

3) Prediction Models

LSTM



Event Log	LSTM Train Time	LSTM Acc
RTFM	3021	0.791
BPIC-2012-W	3344	0.68

XG Boost

Event Log	XGB Train Time	XGB Acc
RTFM	11372	0.814
BPIC-2012-W	2011	0.685

Evaluation

4) Metrics

$$R_s = \frac{\text{Size of the } \textit{whole event log}}{\text{Size of the sampled event log}}$$

$$R_t = \frac{\text{Training time using whole data}}{\text{Training time using the sampled data}}$$

$$R_{Acc} = \frac{\text{Accuracy using the sampled training log}}{\text{Accuracy using the whole training log}}$$

$$R_{FE} = \frac{\text{Feature extraction time using whole data}}{\text{Feature extraction time using the sampled data}}$$

Results

The reduction in the **size of training logs** and the improvement in the **performance of feature extraction**

Sample Methods	Division						Logarithmic distribution						unique	
	K = 2		K = 3		K = 10		log2		log3		log10			
Event Log	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}
RTFM	1.99	4.8	3	11.1	9.8	106.9	153.5	12527.6	236.3	23699.2	572.3	74912.8	285.1	24841.8
BPIC-2012-W	1.22	1.37	1.41	1.8	1.66	2.51	6.06	22.41	9.05	37.67	28.5	208.32	1.73	2.36

Results

The reduction in the **size of training logs** and the improvement in the **performance of feature extraction**

Sample Methods	Division						Logarithmic distribution						unique	
	K = 2		K = 3		K = 10		log2		log3		log10			
Event Log	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}	R_S	R_{FE}
RTFM	1.99	4.8	3	11.1	9.8	106.9	153.5	12527.6	236.3	23699.2	572.3	74912.8	285.1	24841.8
BPIC-2012-W	1.22	1.37	1.41	1.8	1.66	2.51	6.06	22.41	9.05	37.67	28.5	208.32	1.73	2.36

Lowest

Highest

Results

Lowest

Highest

- **LSTM**

Sample Methods	Division						Logarithmic distribution						unique	
	K = 2		K = 3		K = 10		log2		log3		log10			
Event Log	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t
RTFM	**1.001**	**2**	1.004	2.9	0.99	9	0.716	26.7	0.724	33	**0.767**	**41.8**	0.631	29.1
BPIC-2012-W	**1**	**1.4**	0.985	1.3	0.938	1.3	0.977	4.7	0.97	5.8	**0.876**	**11.9**	0.996	1.6

- **XG Boost**

Sample Methods	Division						Logarithmic distribution						unique	
	K = 2		K = 3		K = 10		log2		log3		log10			
Event Log	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t
RTFM	**1**	**2.4**	1	1.4	1	84.1	0.686	126.4	0.706	191.8	**0.772**	**355**	0.582	297.7
BPIC-2012-W	**0.999**	**2.3**	0.998	2.4	0.997	3.4	0.923	10.7	0.97	16.7	**0.883**	**64.8**	0.997	2.8

Results

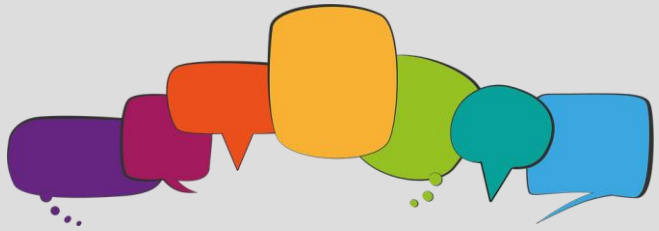
Accuracy improvement

- **LSTM**

Sample Methods	Division						Logarithmic distribution						unique	
	K = 2		K = 3		K = 10		log2		log3		log10			
Event Log	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t
RTFM	**1.001**	**2**	1.004	2.9	0.99	9	0.716	26.7	0.724	33	**0.767**	**41.8**	0.631	29.1
BPIC-2012-W	**1**	**1.4**	0.985	1.3	0.938	1.3	0.977	4.7	0.97	5.8	**0.876**	**11.9**	0.996	1.6

- **XG Boost**

Sample Methods	Division						Logarithmic distribution						unique	
	K = 2		K = 3		K = 10		log2		log3		log10			
Event Log	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t	R_{Acc}	R_t
RTFM	**1**	**2.4**	1	1.4	1	84.1	0.686	126.4	0.706	191.8	**0.772**	**355**	0.582	297.7
BPIC-2012-W	**0.999**	**2.3**	0.998	2.4	0.997	3.4	0.923	10.7	0.97	16.7	**0.883**	**64.8**	0.997	2.8



Discussion

- Sampling event logs could increase the performance and keep the accuracy in some cases

Discussion

- Sampling event logs could increase the performance and keep the accuracy in some cases
- We observed that different event logs needs different sampling methods

Discussion

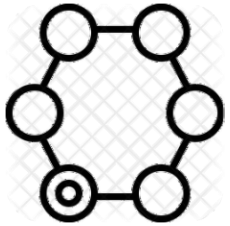
- Sampling event logs could increase the performance and keep the accuracy in some cases
- We observed that different event logs needs different sampling methods
- Characteristics of the given event log and suitable sampling parameters has more effect than number of sampled cases or prediction models

Discussion

- Sampling event logs could increase the performance and keep the accuracy in some cases
- We observed that different event logs needs different sampling methods
- Characteristics of the given event log and suitable sampling parameters has more effect than number of sampled cases or prediction models
- Using the proposed sampling method, we could speed up hyperparameters tunning and adapting with changes due to concept drift

Future work

Relationship between the event log characteristics and the sampling parameters



Predicting critical infrequent activities

Sampling methods for outcome and remaining time prediction

Question?

Thank you

RWTH AACHEN UNIVERSITY TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY