# Proposal

## Proposed topic: Analysis of event log properties in process prediction

## Background

Process mining is a family of methods to analyze business processes based on event logs, which is a collection of traces, each representing the execution of the process also called a case. A trace or case contains a sequence of timestamped activities and may carry other attributes. Since process mining is a relatively young discipline, the open challenges in this field are still many (van der Aalst, 2016). In particular, one capability called predictive process monitoring has been emphasized by researchers recently that providing operational support based on real time detection, prediction, and recommendation, since being able to predict future behavior of a business process is an important business capability (Everman, 2017). It can be applied to predict various attributes at real time, such as numerical value (time, cost), categorical value (risk, next activity, event outcome) and a sequence of activities for a running trace. To complete such tasks, a wide range of methods have been proposed, including machine learning algorithms (like decision trees, random forest or HMM), and deep neural networks architectures (like RNN especially LSTM, and CNN etc.).

## Motivation/Problem Statement

Though a board range of techniques have been proposed to complete process prediction tasks, litter of the study has focused on the comparison of the existing approaches to answer questions like, what can be predicted, and which methods suits the data best, given certain data with specific properties, such as size, dependency among cases, and sparsity.

For example, Everman et al. (2017) apply LSTM networks with two hidden layers to predict the type of the next activity of a case using embedded dimension of LSTMs to reduce the inputs ‘size and to include additional attuites such as associated recourses. However, it cannot handle numerical variables and predict timestep. Also, the model is only evaluated on data with limited number of activities.

Similarly, Tax et al. (2017) also use LSTM networks but with a shared layer that feeds two independent LSTM layers specialized in predicting the next event and its timestamp. By repeatedly performing the prediction task, it can be extended for predicting the remaining sequence of events until termination. However, this approach is based on one-hot encoding to capture the activity type, thus it’s not suitable for event log with high sparsity when performing real time prediction.

While Sendrovich et al. (2017) consider the dependency of cases, they encode features within a bi-dimensional space characterized by both intra-case and inter-case information to predict the remaining times of running case using machine learning methods like random forest. However, it is not extended for predicting categorical values and unsure whether it can outperform than deep neural network approaches on specific event logs.

To answer questions like which methods suits the data best, Francescomarino et al. (2018) conduct a systematically literature review to provide a framework for classifying the existing predictive process monitoring approaches. However, they focused on four dimensions, including type of prediction, type of input, family of algorithms and tool availability. The type of input dimension is only roughly investigated to show which kind of data is needed for certain algorithms but not taking steps deeper to look at the specific properties of the event log.

Another comparison is conducted by Heinrich et al. (2020), they investigate how the hybrid deep learning architecture like Gated Recurrent Unit Network (GRU), Gated Convolutional Neural Network (GCNN) and Key-Value-Predict Attention Network (KVP) can be migrated to complete business process prediction. The performance is evaluated on various datasets with different size and sparsity considering the effect of preprocessing. They find the number of activities and number of process instances negatively correlate to the F1 score. However, this study does not include traditional machine learning approaches but rather focusing on the family of deep learning architectures.

As such, this paper has the objective to address the research gap illustrated above and to investigate how the properties of event logs affect the process prediction performance considering both traditional machine learning approaches and deep learning architectures. In other words, it aims to investigate if one method always outperforms than others in certain prediction task based on event log datasets that share common properties. Thus, it supports, one the one hand, researchers who would like to gain insights on the effect of event log properties, on the other hand, companies that need to be guided to find the solution that best fits their business processes and event log data.

## Research questions

The research question of this paper is *how the properties of event logs affect the process prediction performance of machine learning and deep learning approaches*. Considering the above discussed three properties, including size, sparsity and dependency, this main RQ can be extended to preliminary sub-questions in below:

* How the size of event logs (like number of events, number of traces etc.) affect the *process prediction performance of ML and DL approaches?*
* How the sparsity of event logs (number of activity/ number of traces) affect the *process prediction performance of ML and DL approaches?*
* How the (inter-case) dependency of event logs (determined by the real business scenario) affect the *process prediction performance of ML and DL approaches?*

## Research methods

To answer the defined research questions, an extensive literature review needs to be conducted to gain a comprehensive overview on the existing approaches for process prediction, and thus guide further experiments. A list of preliminary referencing articles is listed at the end of the proposal.

A series of experiments is also needed to test the hypotheses (if necessary) and evaluate the perdition performance on various datasets. The key decisions need to be concerned further are summarized below:

* Which properties should be considered other than the three mentioned above? (like variability, but I am not sure what it exactly means)
* Which datasets can fulfill the requirements and can be selected for evaluation? (most of study use data from BPI)
* How to treat different pre- and postprocessing steps and encoding approaches of each method?
* Shall I include all types of prediction (classification, regression etc.)?

## Proposed timeline

7.2020 - 8.2020 Refine proposal and register thesis

8.2020 - 10.2020 Literature review and refine research questions and hypotheses

10.2020 - 11.2020 Select datasets, design and conduct experiment

11.2020 -12.2020 complete other sections like discussion

1.2020 finetune

## Preliminary list of references

1. Baars, H., & Kemper, H. G. (2008). Management support with structured and unstructured data—an integrated business intelligence framework. *Information Systems Management*, *25*(2), 132-148.
2. Breuker, D., Matzner, M., Delfmann, P., & Becker, J. (2016). Comprehensible Predictive Models for Business Processes. *MIS Q.*, *40*(4), 1009-1034.
3. Camargo, M., Dumas, M., & González-Rojas, O. (2019, September). Learning accurate LSTM models of business processes. In *International Conference on Business Process Management* (pp. 286-302). Springer, Cham.
4. Conforti, R., de Leoni, M., La Rosa, M., van der Aalst, W. M., & ter Hofstede, A. H. (2015). A recommendation system for predicting risks across multiple business process instances. *Decision Support Systems*, *69*, 1-19.
5. Conforti, R., Fink, S., Manderscheid, J., & Röglinger, M. (2016, September). PRISM–a predictive risk monitoring approach for business processes. In *International Conference on Business Process Management* (pp. 383-400). Springer, Cham.
6. Di Francescomarino, C., Ghidini, C., Maggi, F. M., & Milani, F. (2018, September). Predictive process monitoring methods: Which one suits me best?. In *International Conference on Business Process Management* (pp. 462-479). Springer, Cham.
7. Evermann, J., Rehse, J. R., & Fettke, P. (2017). Predicting process behaviour using deep learning. *Decision Support Systems*, *100*, 129-140.
8. Heinrich, K., Zschech, P., Janiesch, C., & Bonin, M. Ein Vergleich aktueller Deep-Learning-Architekturen zur Prognose von Prozessverhalten.
9. Maggi, F. M., Di Francescomarino, C., Dumas, M., & Ghidini, C. (2014, June). Predictive monitoring of business processes. In *International conference on advanced information systems engineering* (pp. 457-472). Springer, Cham.
10. Márquez-Chamorro, A. E., Resinas, M., Ruiz-Cortés, A., & Toro, M. (2017). Run-time prediction of business process indicators using evolutionary decision rules. *Expert Systems with Applications*, *87*, 1-14.
11. Márquez-Chamorro, A. E., Resinas, M., & Ruiz-Cortes, A. (2017). Predictive monitoring of business processes: a survey. *IEEE Transactions on Services Computing*, *11*(6), 962-977.
12. Mehdiyev, N., Evermann, J., & Fettke, P. (2018). A novel business process prediction model using a deep learning method. *Business & information systems engineering*, 1-15.
13. Pasquadibisceglie, V., Appice, A., Castellano, G., & Malerba, D. (2019, June). Using convolutional neural networks for predictive process analytics. In *2019 International Conference on Process Mining (ICPM)* (pp. 129-136). IEEE.
14. Polato, M., Sperduti, A., Burattin, A., & de Leoni, M. (2018). Time and activity sequence prediction of business process instances. *Computing*, *100*(9), 1005-1031.
15. Senderovich, A., Di Francescomarino, C., Ghidini, C., Jorbina, K., & Maggi, F. M. (2017, September). Intra and inter-case features in predictive process monitoring: A tale of two dimensions. In *International Conference on Business Process Management* (pp. 306-323). Springer, Cham.
16. Tax, N., Verenich, I., La Rosa, M., & Dumas, M. (2017, June). Predictive business process monitoring with LSTM neural networks. In *International Conference on Advanced Information Systems Engineering* (pp. 477-492). Springer, Cham.
17. Tax, N., Teinemaa, I., & van Zelst, S. J. (2018). An interdisciplinary comparison of sequence modeling methods for next-element prediction. *CoRR abs/1811.00062*, *302*.
18. Verenich, I., Dumas, M., Rosa, M. L., Maggi, F. M., & Teinemaa, I. (2019). Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(4), 1-34.