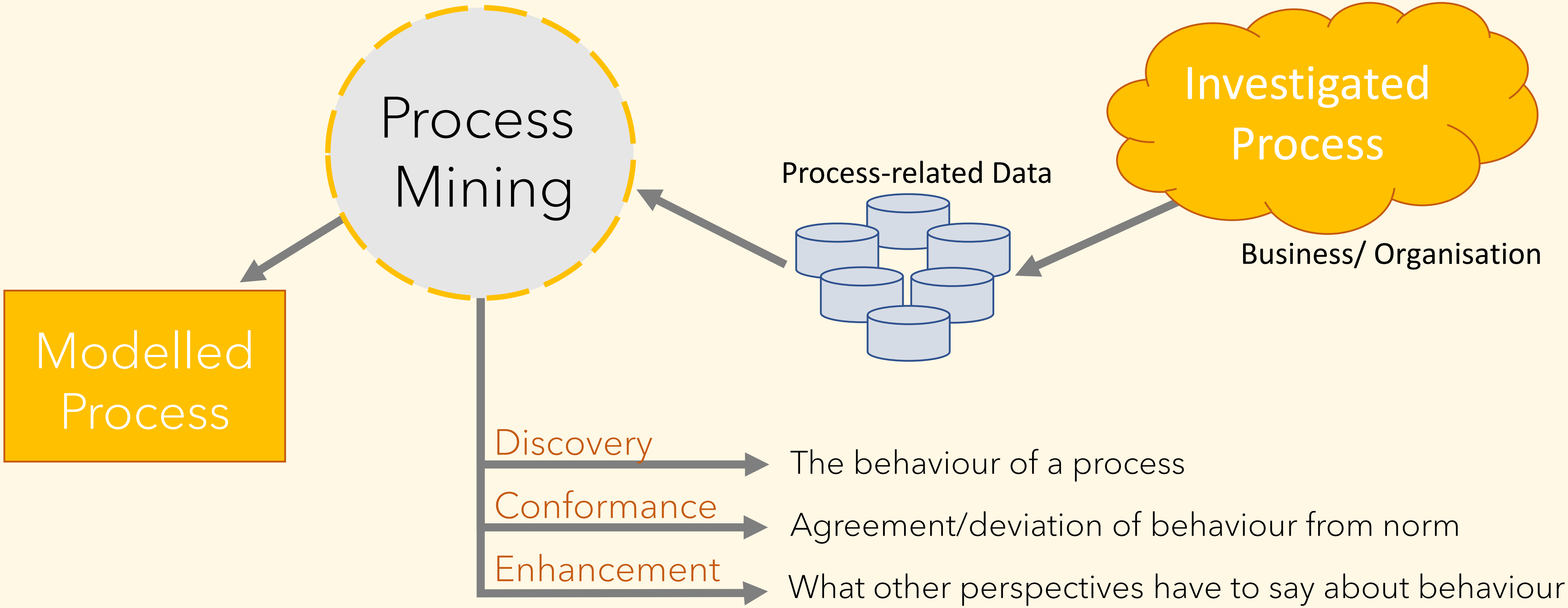
# Process Mining
# with
# Exogenous Data

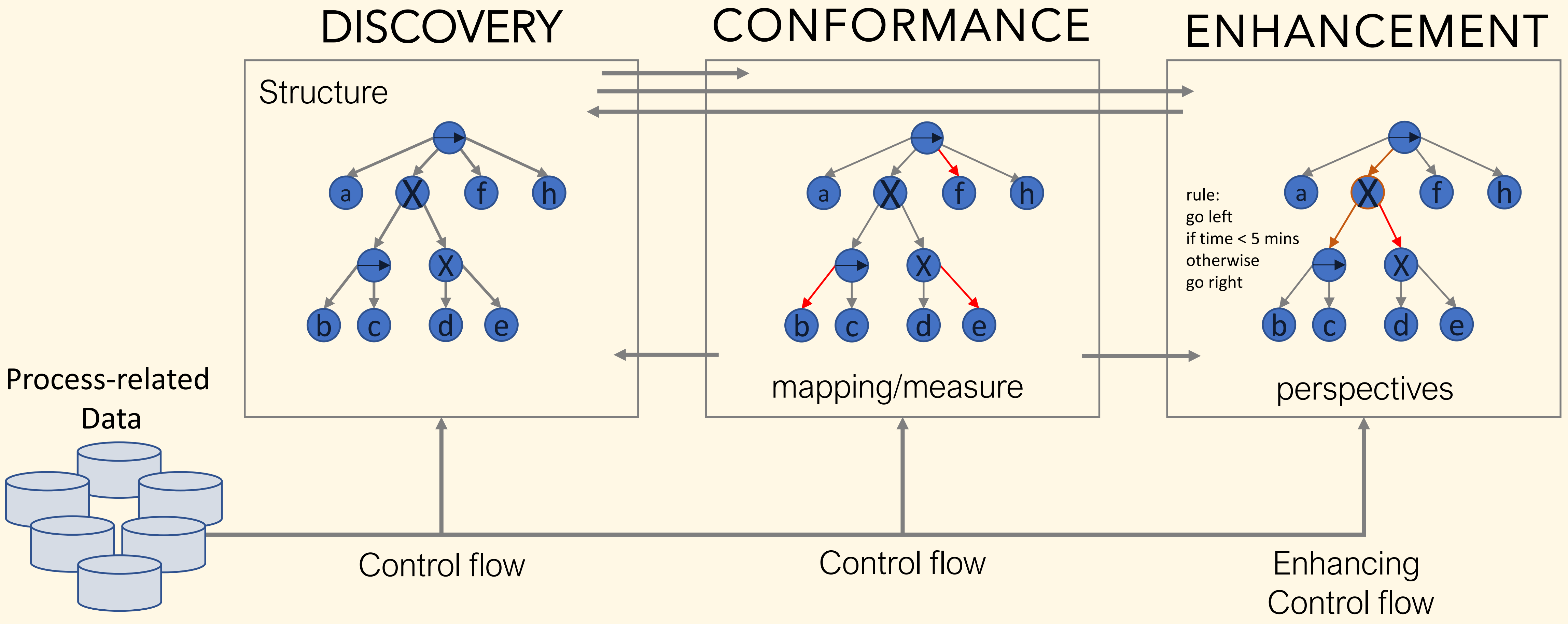
Traditional Thesis by Monograph

Adam Banham, PhD Candidate
Principal: Prof Moe Thandar Wynn
Associate: Dr Robert Andrews
External: Prof Sander Leemans

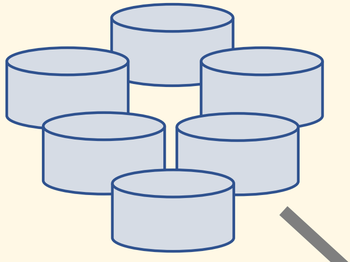
# Process Mining



# Process Mining



# Endogenous Data



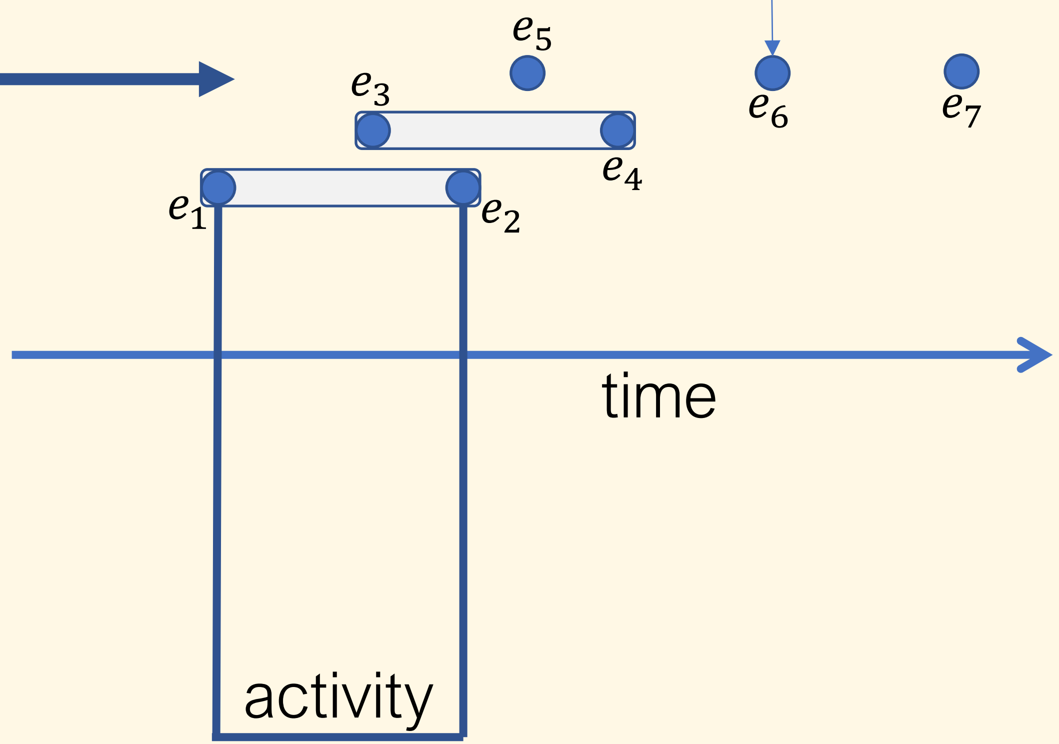
Process-related Data

Event log

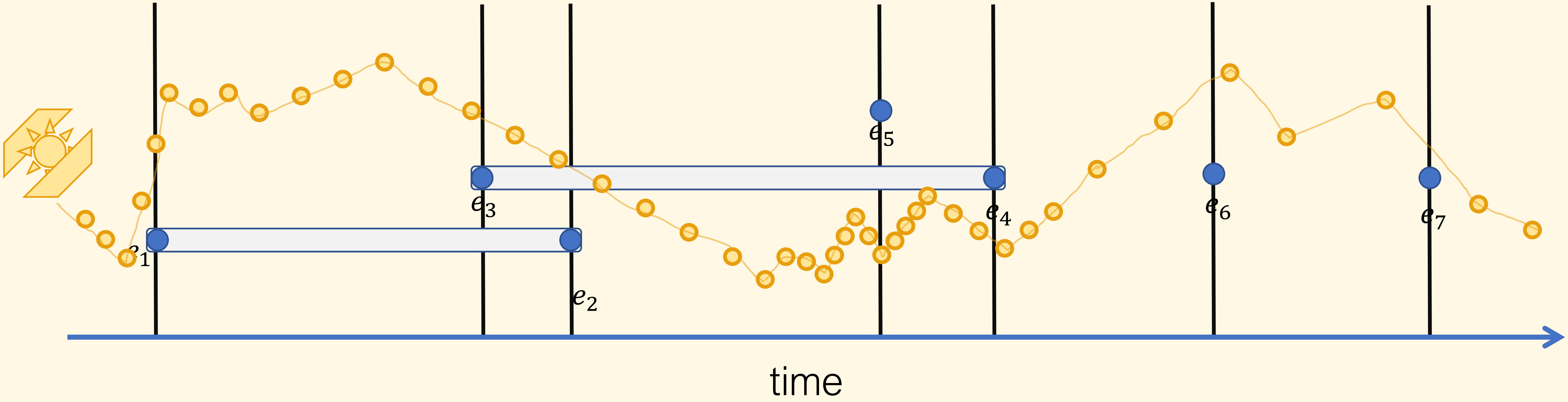
| case | act | res | cst | time | lyfc | $e_{\#}$ |
|---|---|---|---|---|---|---|
| Trace one | A | #13 | $40 | 04:00 | start | 1 |
| | A | #13 | | 04:30 | end | 2 |
| | B | #14 | $25 | 04:20 | start | 3 |
| | B | #13 | | 04:40 | end | 4 |
| | C | #15 | $12 | 04:50 | end | 5 |
| | D | #15 | $4 | 05:00 | end | 6 |
| | E | #16 | $2 | 05:15 | end | 7 |
| Trace two | A | #14 | $40 | 07:00 | start | 8 |
| | A | #14 | | 07:33 | end | 9 |
| | B | #14 | $25 | 07:28 | start | 10 |
| | B | #14 | $2 | 07:38 | end | 11 |
| | C | #15 | $12 | 07:40 | start | 12 |
| | C | #15 | | 07:59 | end | 13 |
| | E | #16 | $2 | 08:15 | end | 14 |

{ activity: D,
time: 05:00,
resource: #15,
cost: $4,
lifecycle: end }

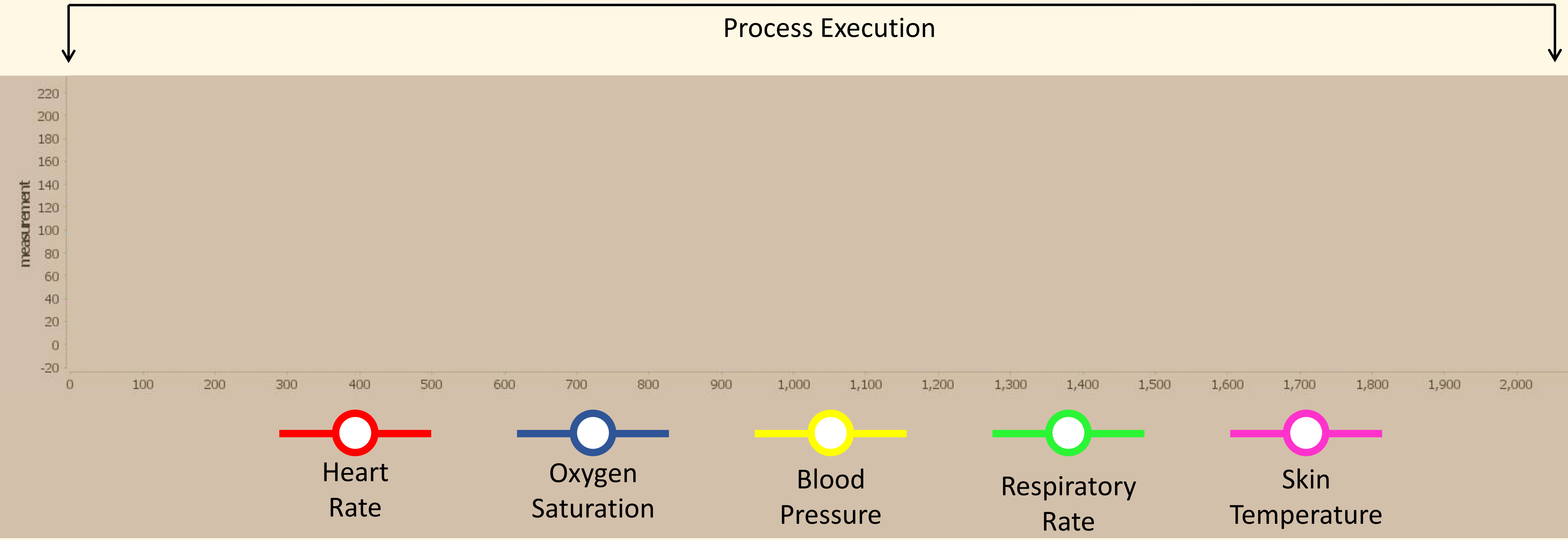
even



# Exogenous Data



# Exogenous Data



# Research Gaps

*Only event attributes are considered…*

- Data Source types around processes: Roseman et al (2008) [1] & van der Aalst and Dustdar (2012) [2]
- Intra/inter-trace attributes: Senderovich et al (2019) [3]
- Intra trace temporal attributes: Pourbafrani et al (2021) [4]
- Contextual events: Dees et al (2020) [5]
- Mutli-level processes based on event attributes: Leemans et al (2020) [6]
- Data expression restrictions on behaviour: Shraga et al (2019) [7]
- Challenges for process mining adoption: Munoz-Gama et al (2021) et al [8]

*Process-decision rules only consider logical operators consisting of {=,≠,<,>}*
*No temporal operators…*

- Decision Miner: Rozinat (thesis) (2010) [9]
- Alignment based Decision Miner: De Leoni and van der Aalst (2013) [10]
- Linear equalities with multiple variables for decision expressions: De Leoni et al (2013) [11]
- Non-exclusive decision mining: Mannhardt et al (2016) [12]
- Conformance measures to correct for decision expressions: Mannhardt (thesis) (2018) [13]
- Variable to variable expressions (read/written): Felli et al (2019) [14]

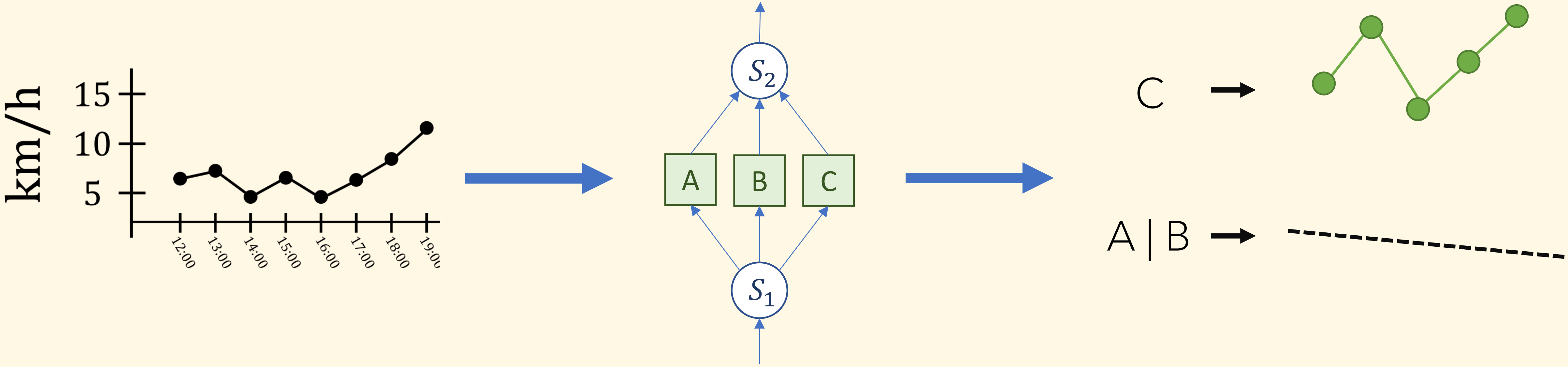
*No studies were found that focus on the temporal changes of decision-making…*

- Global decision making: Winter and Rinderle-Ma (2017) [15] and Winter et al (2020) [16]
- Declarative constraints: Pesic et al (2007) [17] and Leno et al (2020) [18]
- Temporal changes to intermediately available data: van Zelst (thesis) (2019) [19]
- Temporal changes to structure of process behaviour: Stertz et al (2020) [20] and Brockhoff et al (2020) [21]
- High velocity of events requiring abstractions: Mannhardt et al (2018) [22] and Smedt et al (2021) [23]

# Process Mining
## with
# Exogenous Data

*Research Aim*

*How can analysts study the influence of exogenous data on decisions in processes?*



# Process Mining
## with
# Exogenous Data

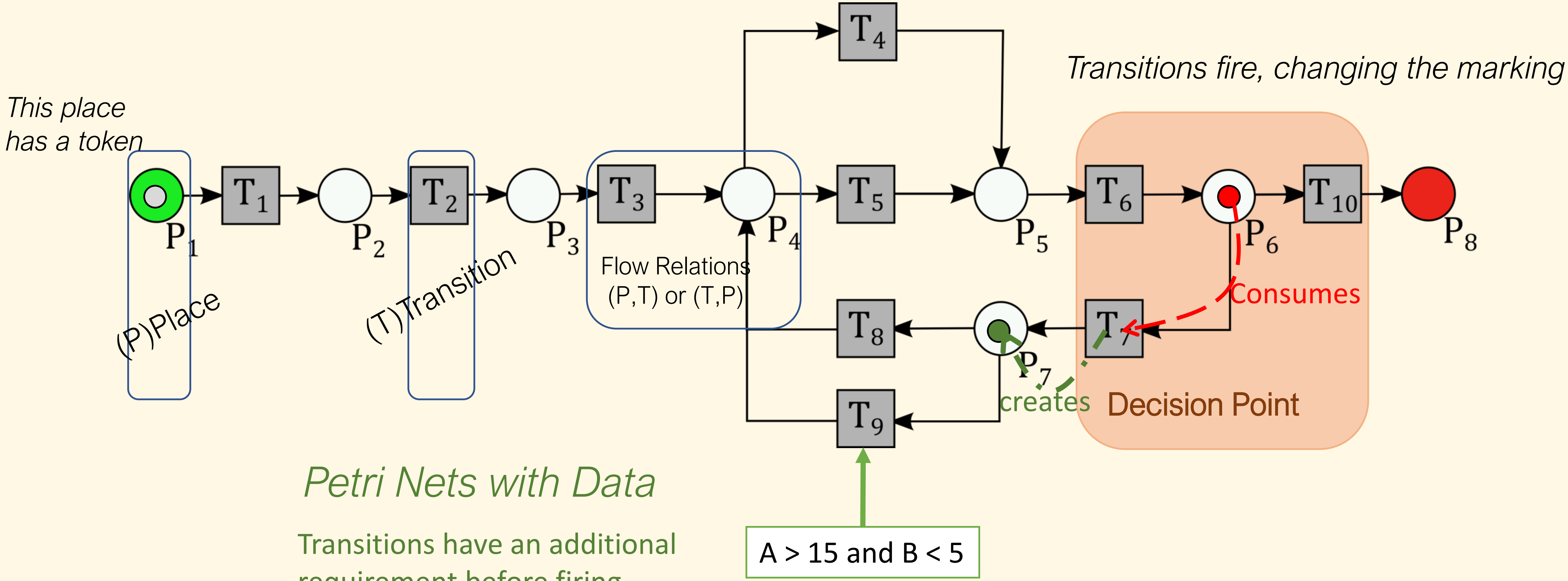
*Research Questions*

*How can we represent and use exogenous data in existing process mining techniques?*

*How can patterns/artefacts in exogenous data be identified and represented for decision-making?*

*How can analysts study the influence of exogenous data on decision-making variance in processes?*

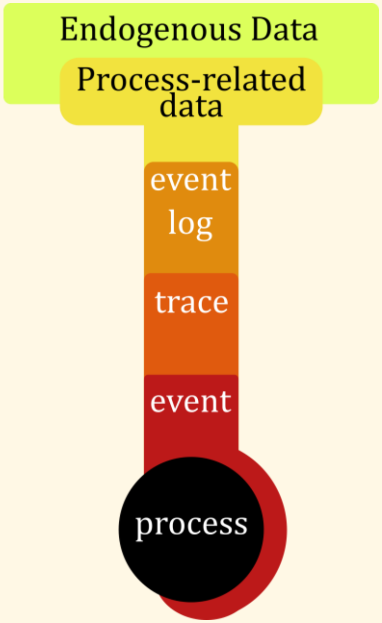
# Petri Nets

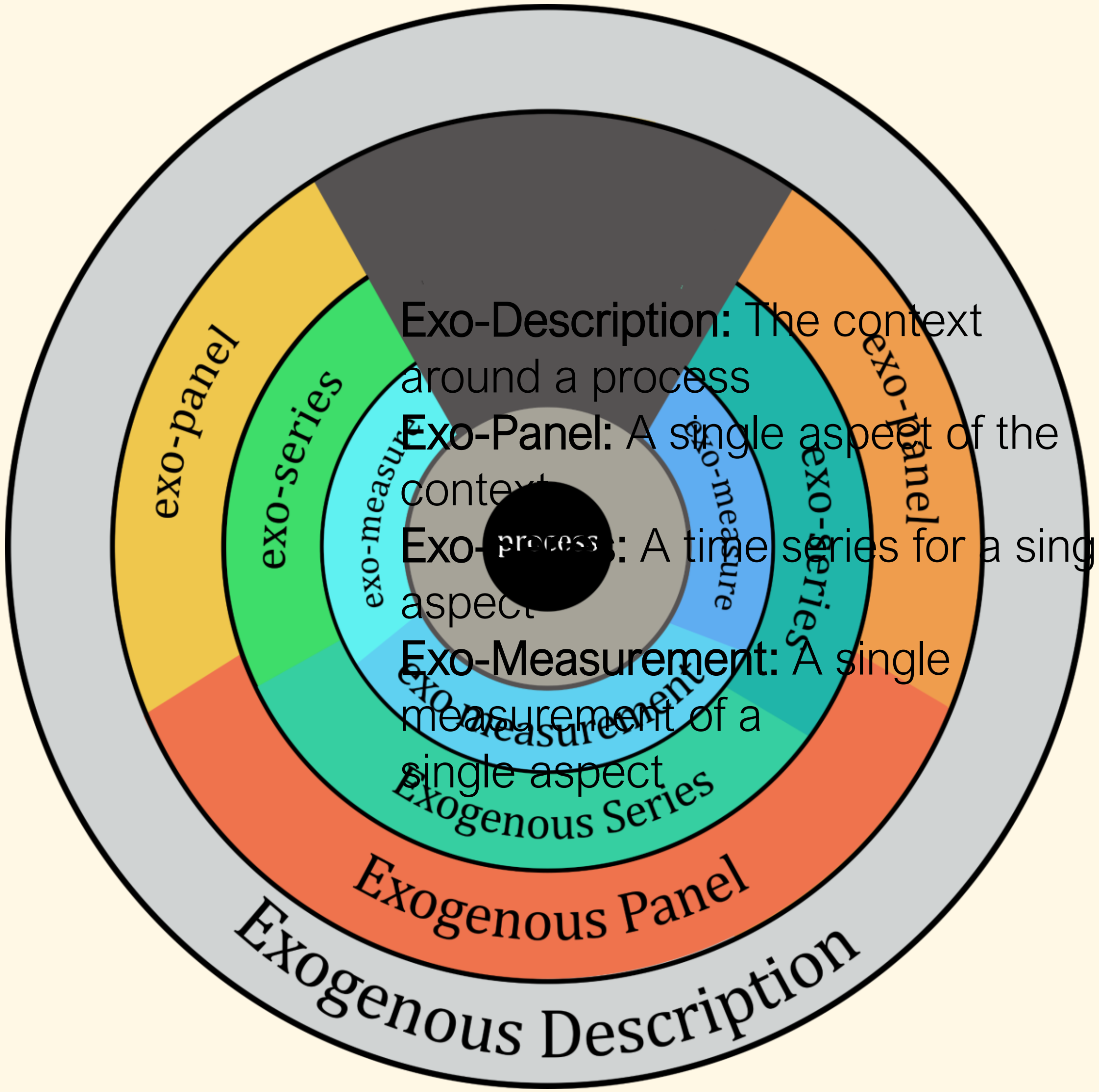



*Petri Nets with Data*

Transitions have an additional requirement before firing...

# Exogenous Data




**Exo-Description:** The context around a process

**Exo-Panel:** A single aspect of the context

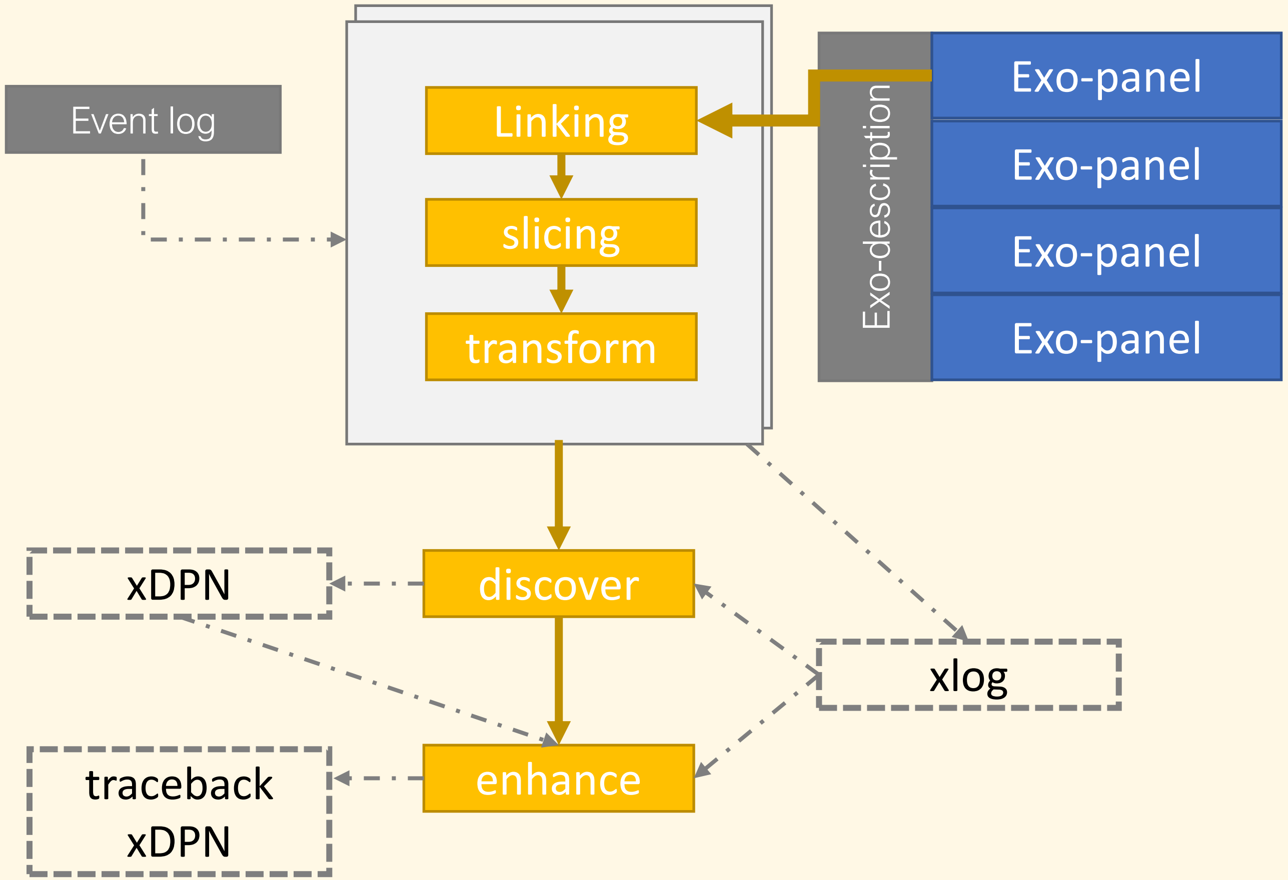
**Exo-Series:** A time series for a single aspect

**Exo-Measurement:** A single measurement of a single aspect

# Exogenous Data

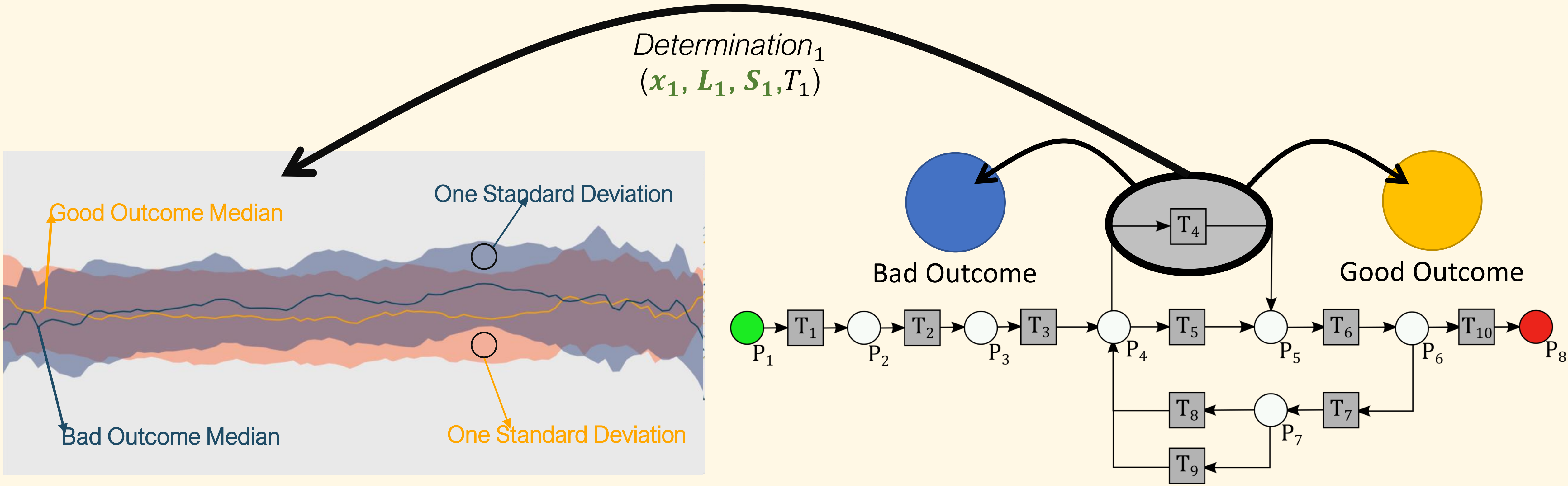
*RQ ONE*
*How can we represent and use exogenous data in existing process mining techniques?*

# xPM
# Framework for process mining with exogenous data




- By applying several *determinations*, we describe the possible influences affecting process behaviour
- *Determinations* are tuples of (**x, L, S, T**), consisting of:
  - an exogenous data source (exo–panel or **x**),
  - a linking function to connect a trace and an data source, to find a time series (exo-series or **L**),
  - a slicing function to create a subset of an exo–series (*slice*) for an event (**S**),
  - a transformation function to create transformed attributes for an event (**T**).

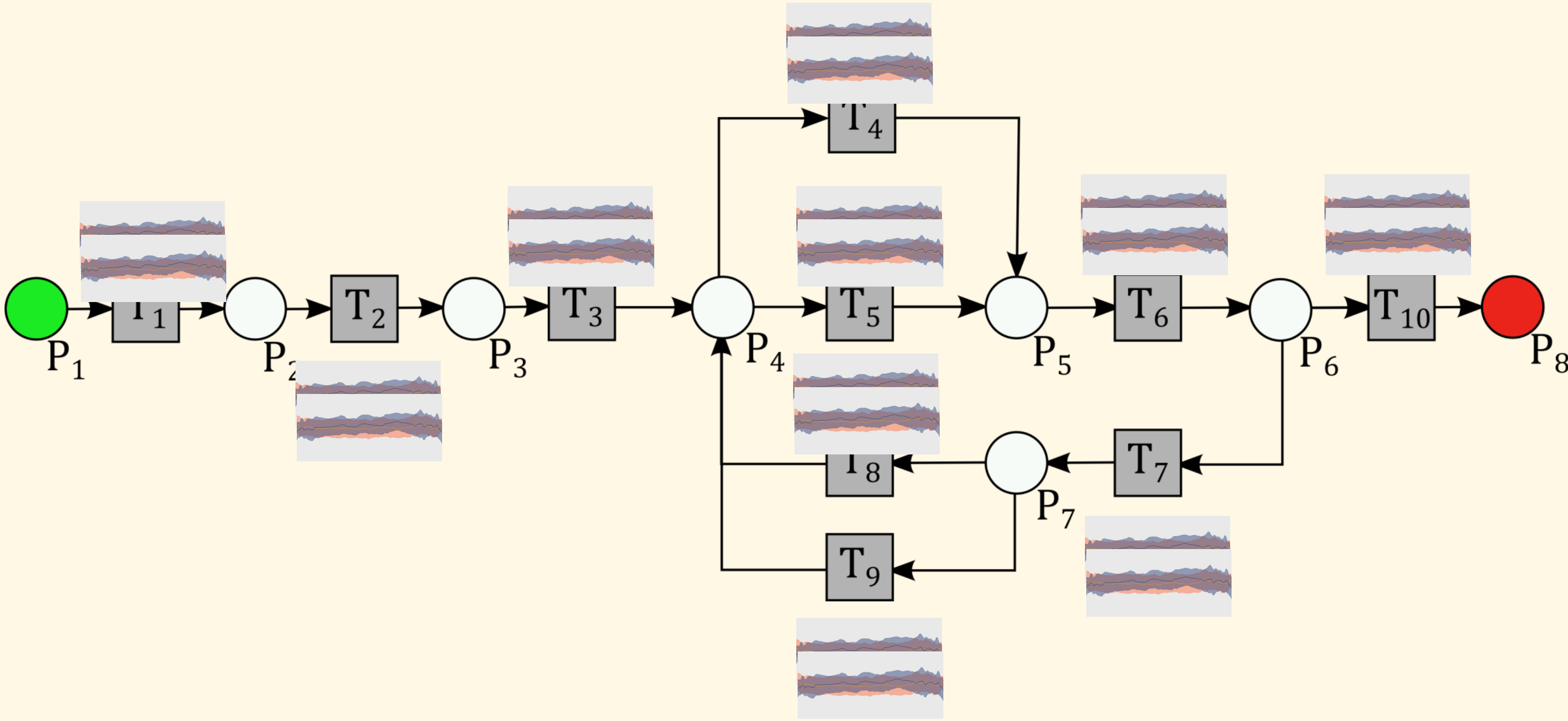
# Explorative Exogenous Series Analysis (EESA)



# Ranking a Collection of EESAs

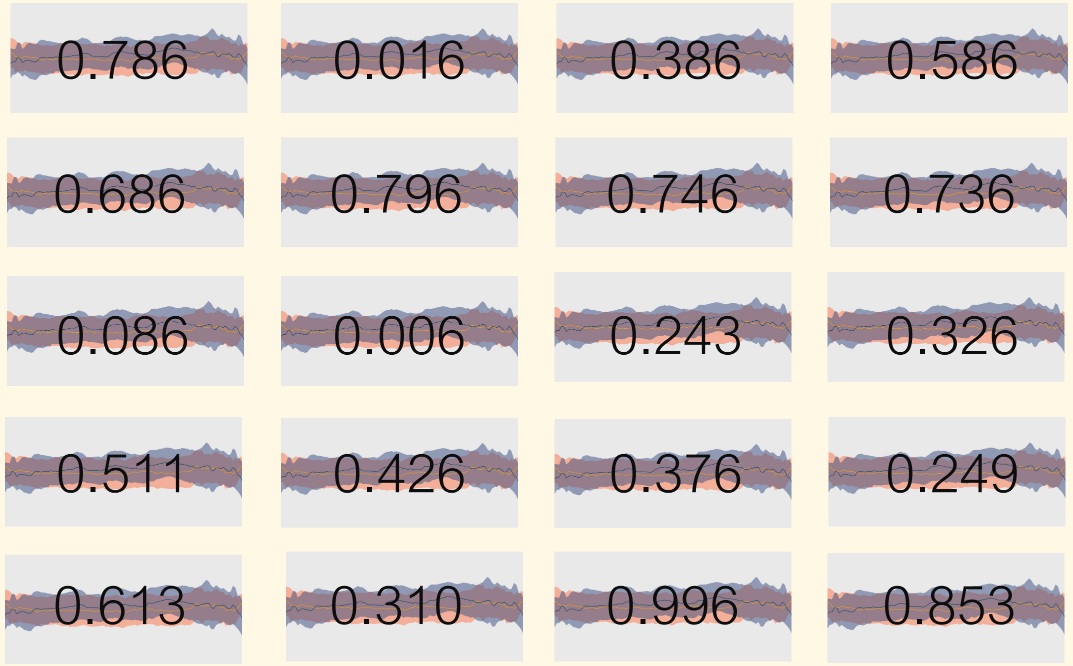
*Determination*$_1$ $(\boldsymbol{x_1}, \boldsymbol{L_1}, \boldsymbol{S_1}, T_1)$

*Determination*$_2$ $(\boldsymbol{x_2}, \boldsymbol{L_2}, \boldsymbol{S_1}, T_1)$



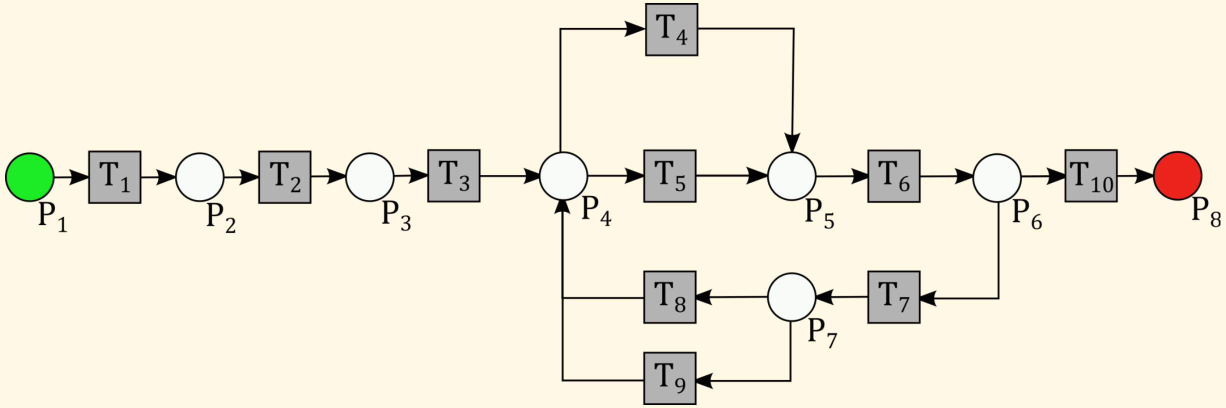
# Ranking a Collection of EESAs

Applying Dynamic Time Warping between
median trend lines for outcomes



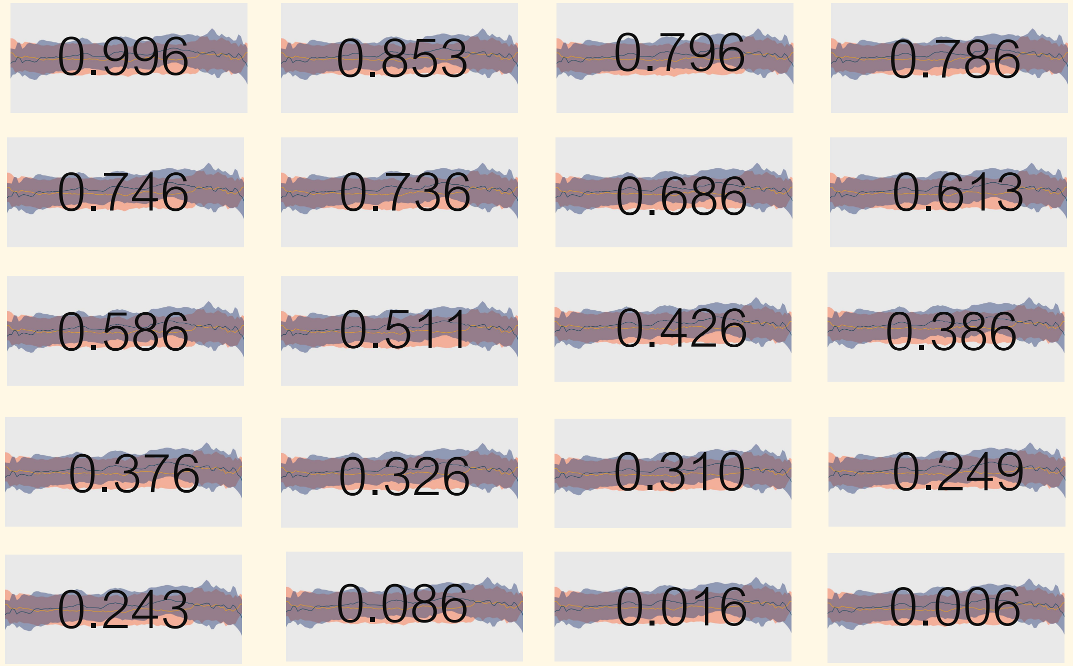

*Determination*$_1$
$(x_1, L_1, S_1, T_1)$

*Determination*$_2$
$(x_2, L_2, S_1, T_1)$



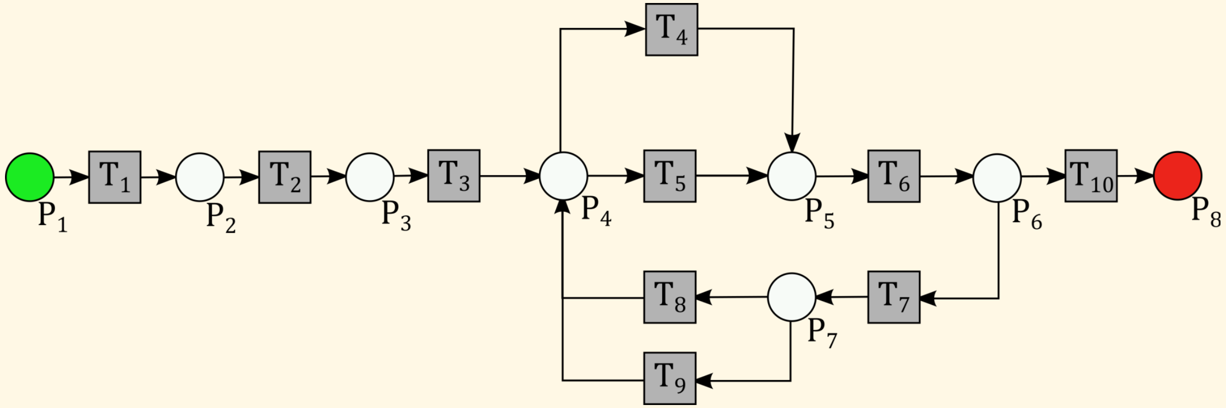
# Ranking a Collection of EESAs

Then we sort by descending value,
and an assign a rank



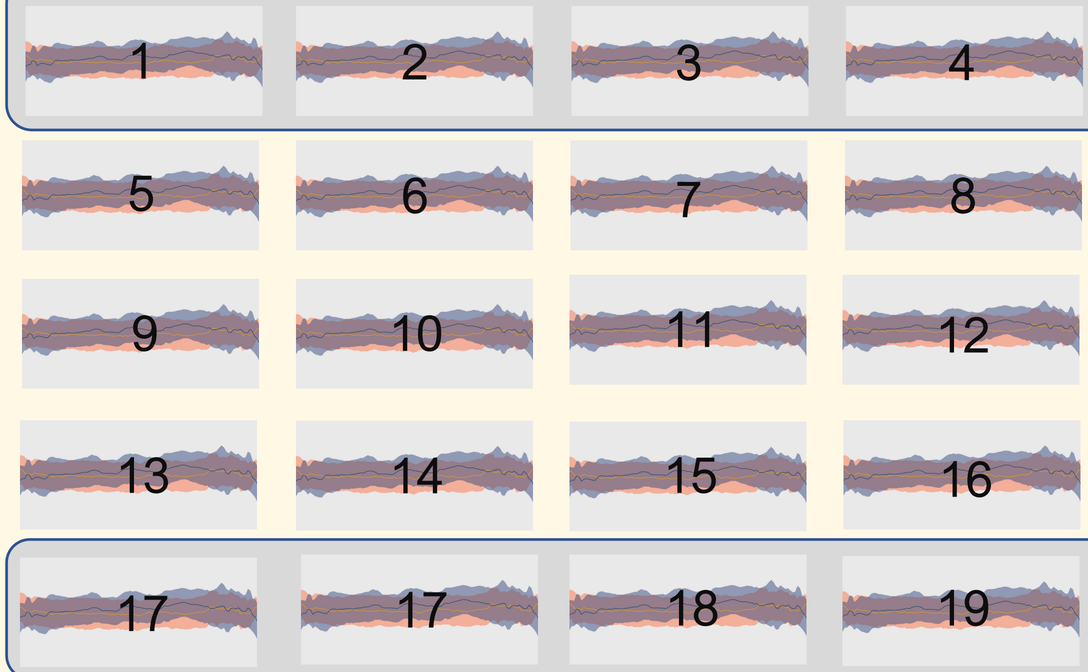

*Determination*$_1$ $(\boldsymbol{x_1}, \boldsymbol{L_1}, \boldsymbol{S_1}, T_1)$

*Determination*$_2$ $(\boldsymbol{x_2}, \boldsymbol{L_2}, \boldsymbol{S_1}, T_1)$



# Ranking a Collection of EESAs

Then we sort by descending value,
and an assign a rank

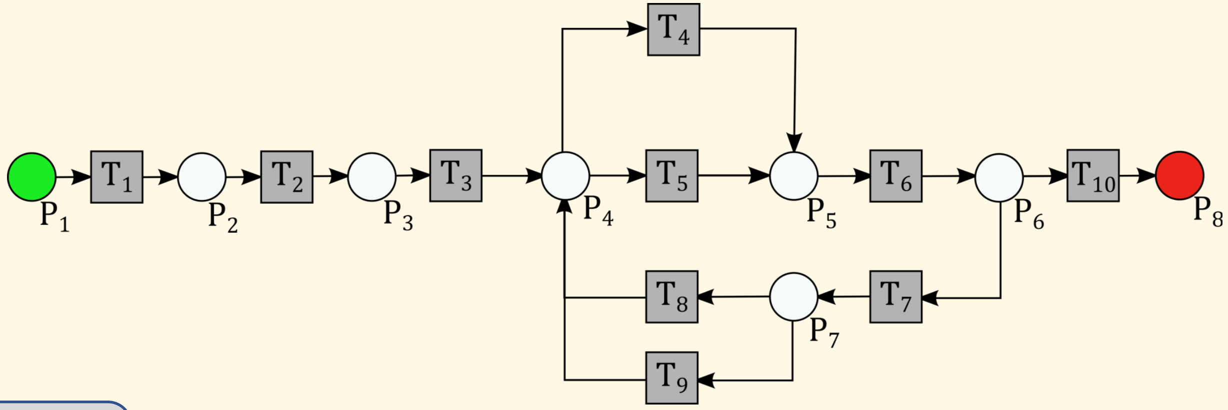



most different

most similar

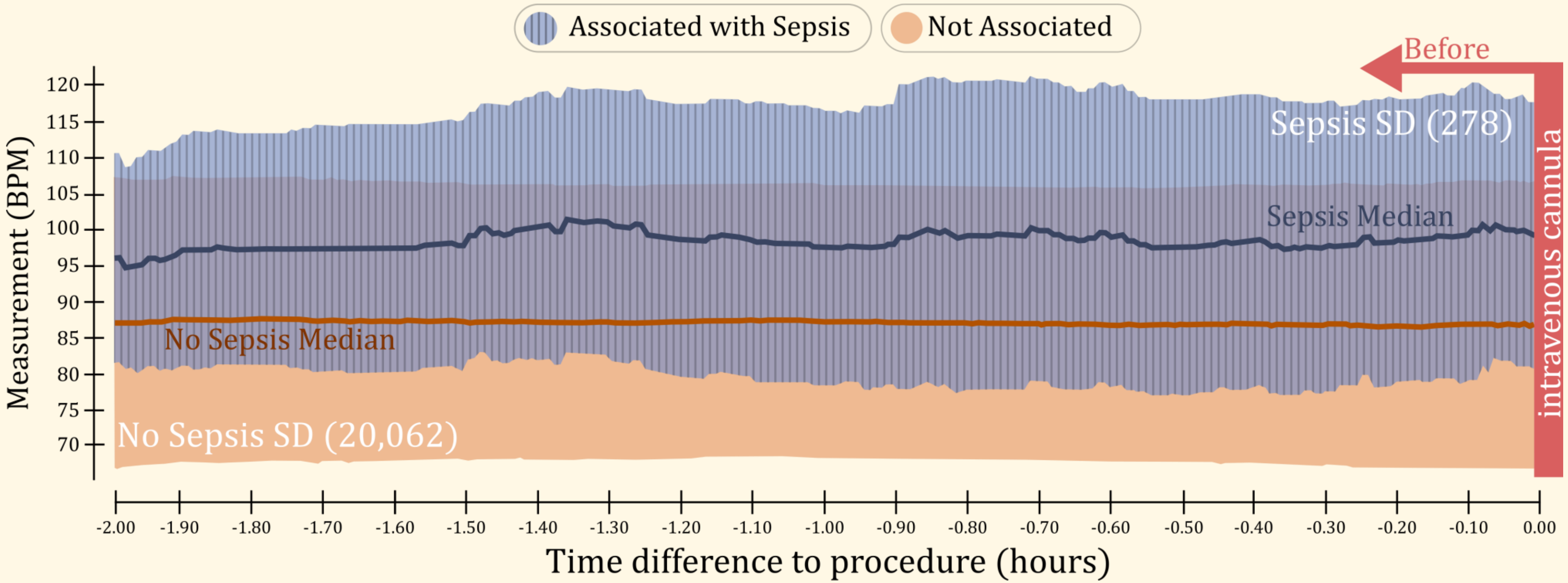
*Determination*$_1$ $(\boldsymbol{x_1}, \boldsymbol{L_1}, \boldsymbol{S_1}, T_1)$

*Determination*$_2$ $(\boldsymbol{x_2}, \boldsymbol{L_2}, \boldsymbol{S_1}, T_1)$

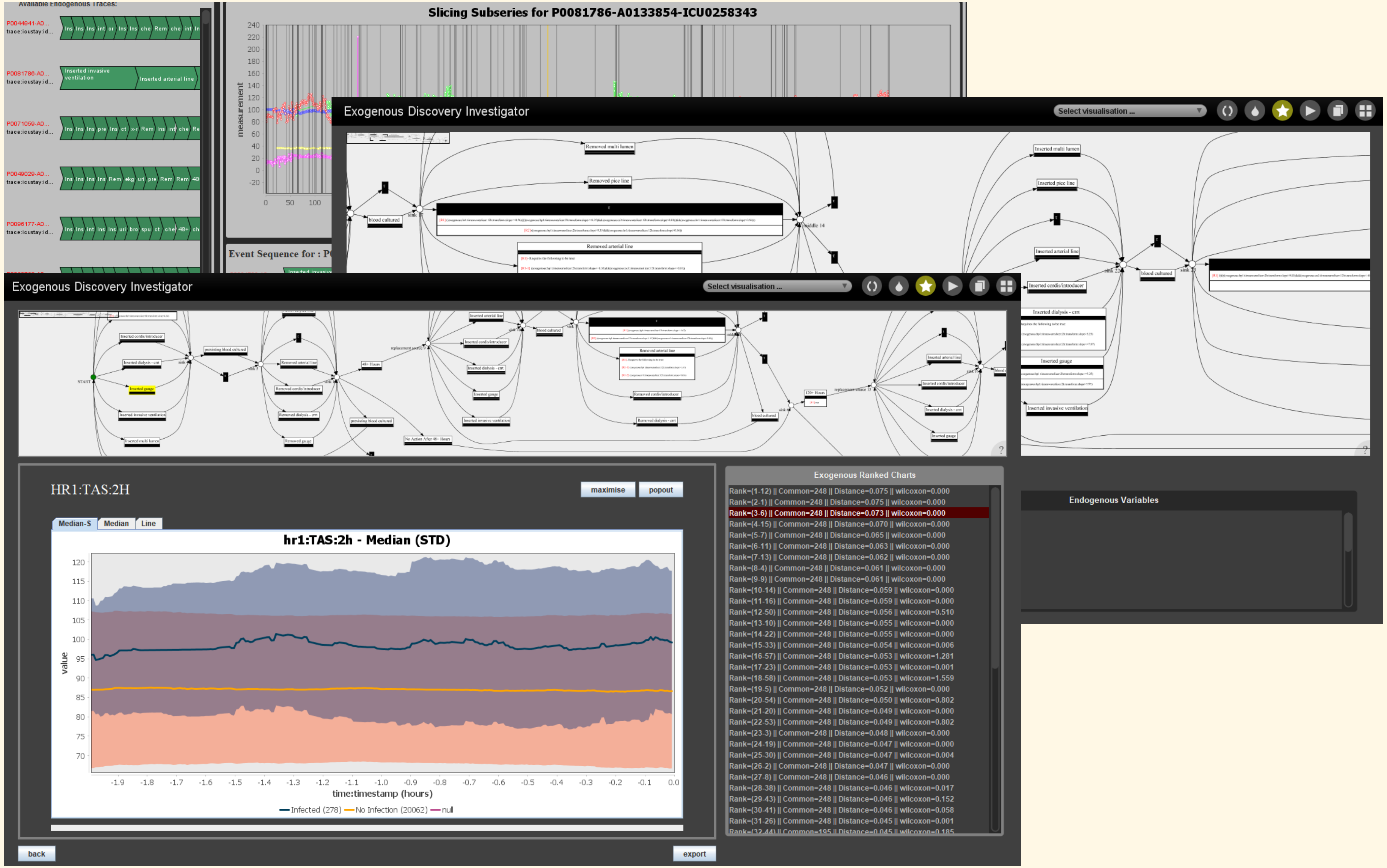


# xPM for Healthcare

- Used the publicly available dataset, MIMIC-III [27]
- Published workshop paper [26], at Event Data and Behaviour Analysis workshop, co-located at ICPM2021
- Worked with healthcare domain experts to investigate:
    - *Recognising patients with early or incubating infections and/or predicting those who will develop subsequent infections/sepsis*.
- Submission to special issue for the Journal of Artificial Intelligence for Medicine (Q1)



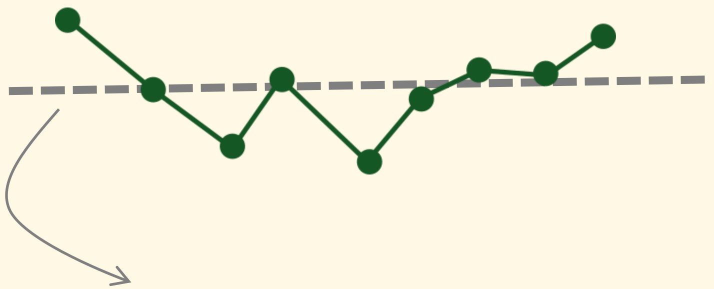
# Tool Support



# Patterns & Artefacts

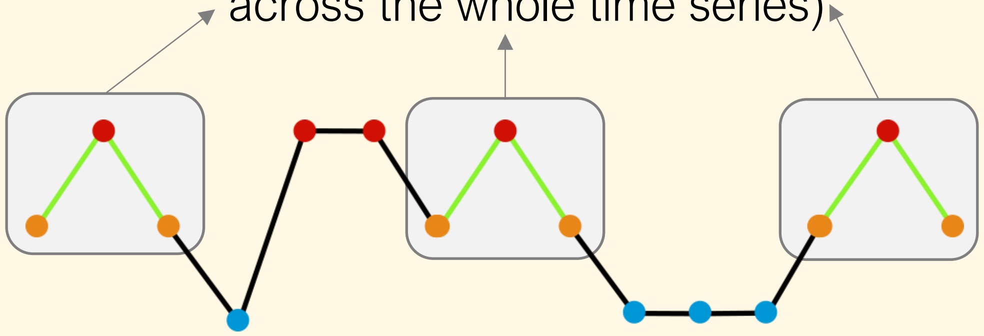
*RQ TWO*

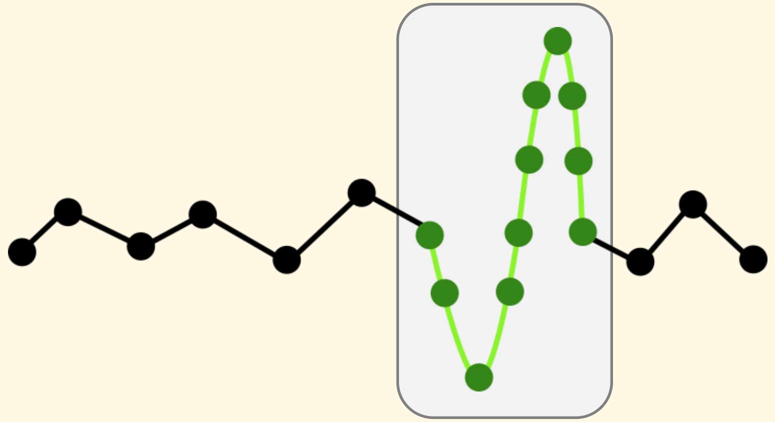
*How can patterns/artefacts in exogenous data be identified and represented for decision-making?*



Long running trends

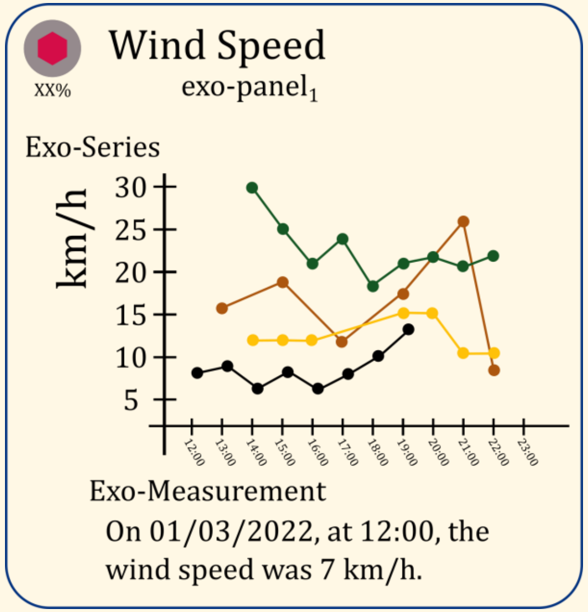
Patterns
(presented in many values across the whole time series)

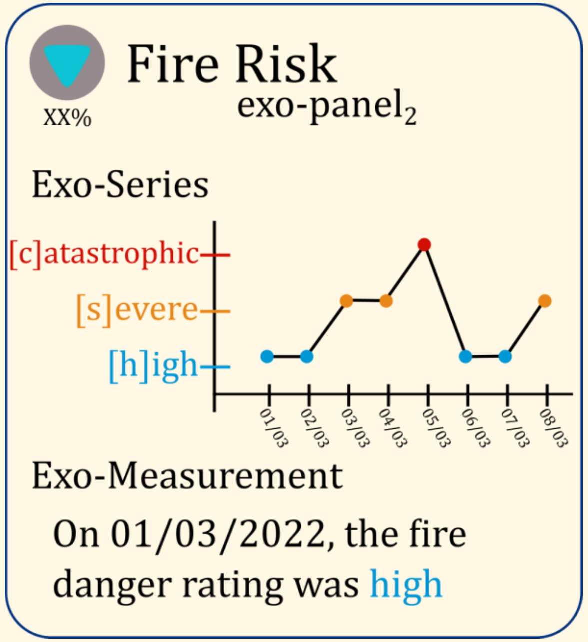


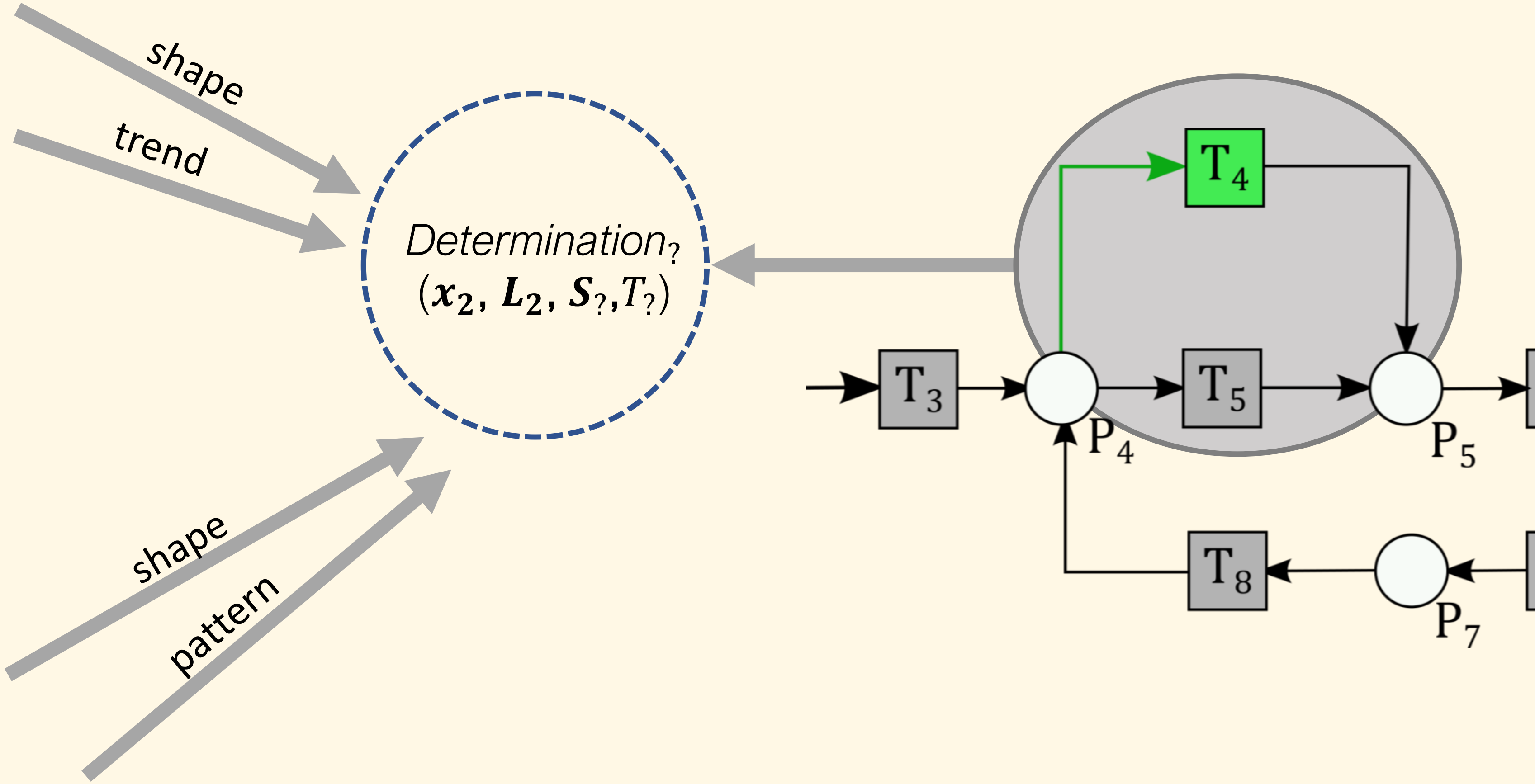


Shape
(presented in a unbroken sequence of values)

# Problem

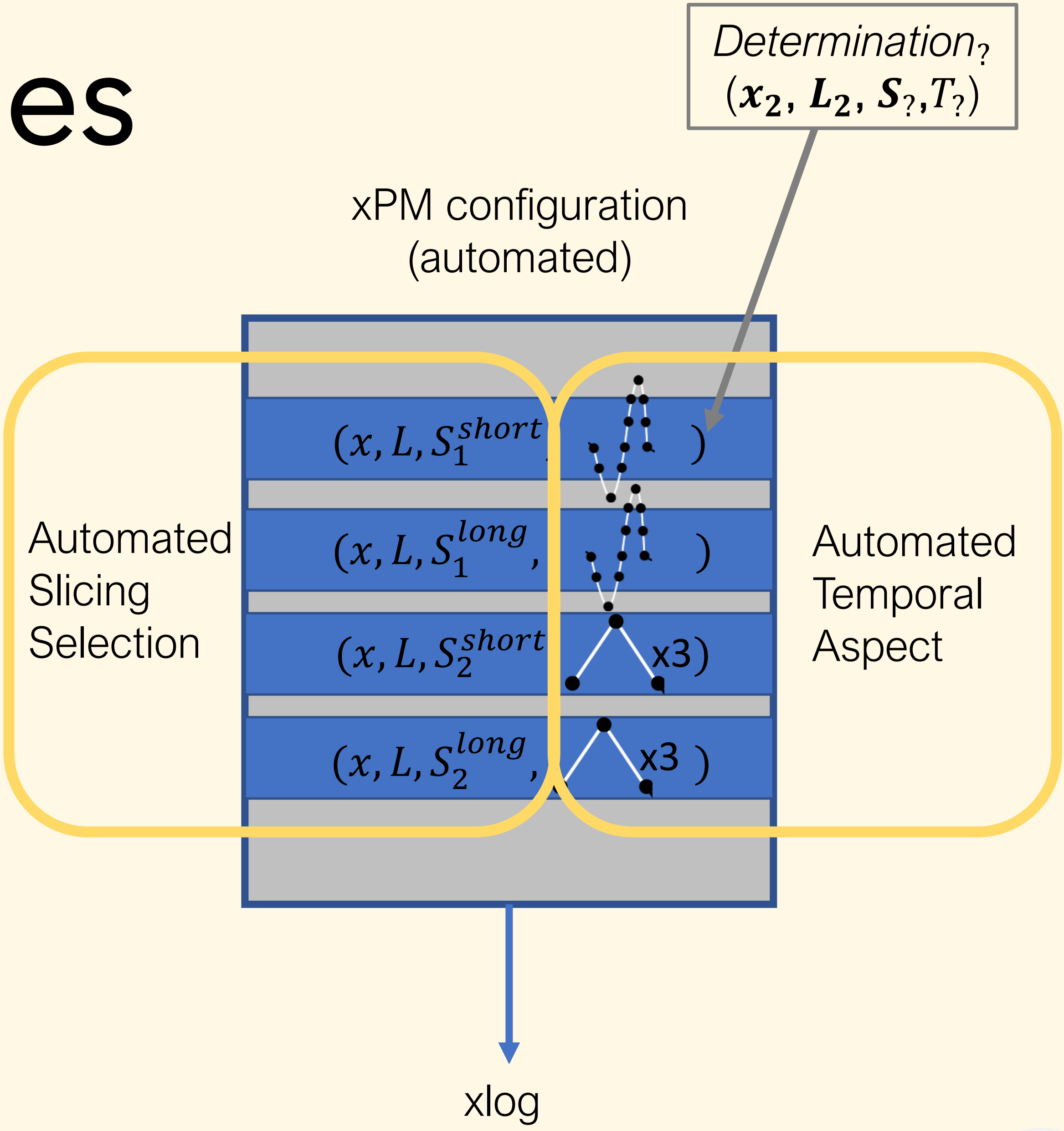

Wind Speed
XX%
exo-panel1
Exo-Series
km/h
30
25
20
15
10
5
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Exo-Measurement
On 01/03/2022, at 12:00, the wind speed was 7 km/h.


Fire Risk
XX%
exo-panel2
Exo-Series
[c]atastrophic
[s]evere
[h]igh
01/03
02/03
03/03
04/03
05/03
06/03
07/03
08/03
Exo-Measurement
On 01/03/2022, the fire danger rating was high


shape
trend
shape
pattern
Determination?
(x2, L2, S?, T?)
T3
T4
T5
T8
P4
P5
P7

# Planned Outcomes

- Automated analysis of finding process-relevant determinations
  - For transforms functions with guarantees
    - a shape operator
    - a sequential/pattern operator
    - a trend operator
  - For slicing functions with guarantees
    - uses the least amount of data [just in time detection]
    - uses the most amount of data [early detection]
- Automatic construction of an xPM configuration.



# Decision-Making Variance

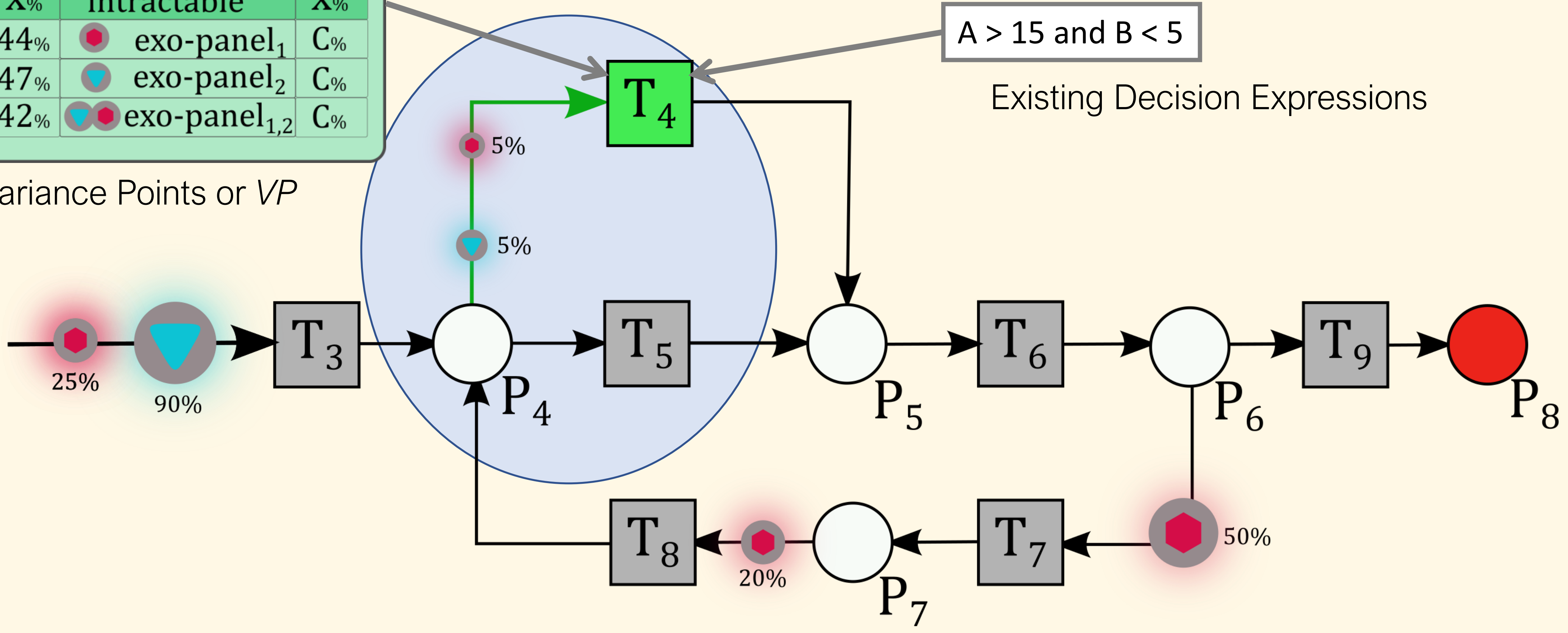
*RQ THREE*
*How can analysts study the influence of exogenous data on decision-making variance in processes?*

# Problem

**$T_4$ - First Rework**

| | P(VP) | Name | C(exp) |
|---|---|---|---|
| $VP_6$ | 50% | Endogenous | C% |
| $VP_7$ | X% | intractable | X% |
| $VP_8$ | 44% | exo-panel$_1$ | C% |
| $VP_9$ | 47% | exo-panel$_2$ | C% |
| $VP_{10}$ | 42% | exo-panel$_{1,2}$ | C% |

Variance Points or *VP*

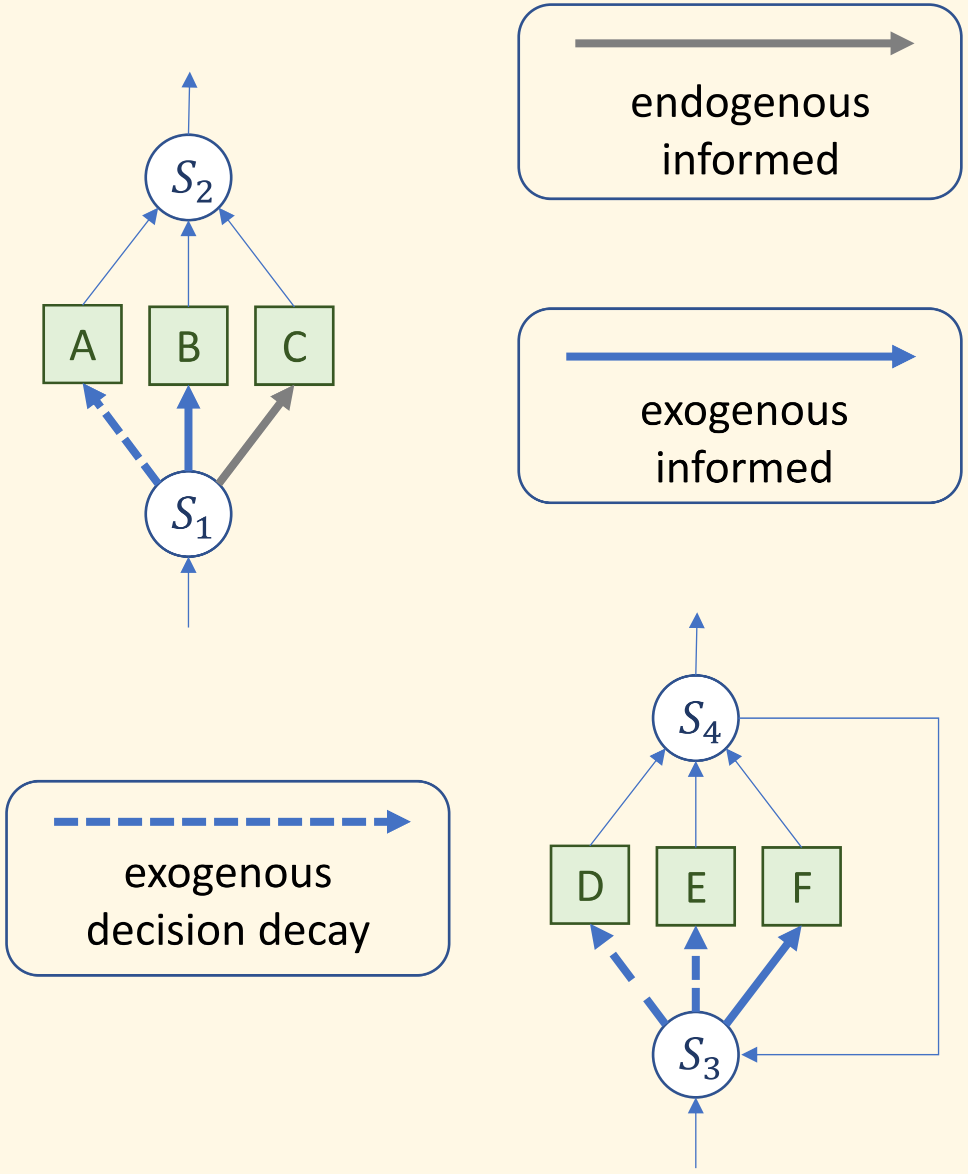



Wind Speed
XX%
exo-panel$_1$

Fire Risk
XX%
exo-panel$_2$

# Planned Outcomes

- A explicit understanding of discovered decision expressions
- Variance points are not static elements
    - Analysis for ‘decision decay’; does a decision expression change if a decision is repeated within an execution or has decision-making changed over time
    - Understanding of  how decision-making occurs in contextual situations based on exogenous data



# Process Mining
## with
# Exogenous Data

*Research Questions*

*How can we represent and use exogenous data in existing process mining techniques?*

*How can patterns/artefacts in exogenous data be identified and represented for decision-making?*

*How can analysts study the influence of exogenous data on decision-making variance in processes?*

# Project Design

*Our research approach generally follows design science research methodology [24]*

**Rigor**
- Review literature
- Identify existing work
- Benchmark
- Present research gap
- Consider objectives

**Design**
- Design prototype
- Validating prototype addresses objectives
- Comparing with existing techniques

**Relevance**
- Test prototype on unknown processes
- Usefulness and effectiveness for domain uses objectives kept?

Incorporating guidelines for presenting process mining algorithms with guarantees [25]

Define Guarantees

Validate Guarantees

Noise relationship

Study Effectiveness

# Progress to Date

22/02/2021

- Literature review
- Initial design of framework (xPM)
- Initial thesis review
- Stage Two

11/05/2021

- ICPM2021
- Developed xPM
- *Article Submission*
  - *EDBA 2021 Successful*
  - *xPM accepted*
- Started Developing Tool Support

17/09/2021

- Contacted Domain Experts
  - Worked through MIMIC-III data to find exemplar
- Extension of workshop paper
  - Design and development of visualisation technique (EESA)
  - Evaluation of xPM and EESA
- *Workshop Presentation at EDBA, ICPM2021*

20/12/2021

- *Refinement of submission for special issue*
  - *Submitted 11th April*
- Quantitively feedback on approach was received positively
  - Being able to reason about the relevant time series data for decisions

22/03/2022

# Next Steps

**11/05/2022**

- Benchmark and analysis of existing techniques for RQ Two
- Initial design for automatic identification of temporal factors

**01/07/2022**

- Implementation of our design
- Evaluation of our implementation against existing
- Draft article for automatic identification of temporal factors

**01/09/2022**

- Design and Implementation of automated procedure to translate temporal factors to slicing and transforming (xPM)
- *Article for automatic construction of xPM (slicing+transform) configuration*

**01/12/2022**

**01/02/2023**

- Benchmark of probabilistic techniques for decision mining
- Design and Implementation for variance point algorithm

**01/05/2023**

- Evaluation of the algorithm for decision-making variance
- *ICPM2023, BPM2023 submission*
- Development of support tools

**01/09/2023**

- Thesis writing

**22/02/2024**

- Thesis Submission ☺

# Process Mining
## with
# Exogenous Data

**Adam Banham, PhD Candidate**
Principal: Prof Moe Thandar Wynn
Associate: Dr Robert Andrews
External: Prof Sander Leemans

?

- [1] M. Rosemann, J. Recker, and C. Flender, “Contextualisation of business processes,” International Journal of Business Process Integration and Management, vol. 3, Art. no. 1, 2008, doi: 10.1504/ijbpim.2008.019347.
- [2] W. M. P. van der Aalst and S. Dustdar, “Process Mining Put into Context,” IEEE Internet Computing, vol. 16, Art. no. 1, Jan. 2012, doi: 10.1109/mic.2012.12.
- [3] A. Senderovich, C. D. Francescomarino, and F. M. Maggi, “From knowledge-driven to data-driven inter-case feature encoding in predictive process monitoring,” Information Systems, vol. 84, pp. 255–264, Sep. 2019, doi: 10.1016/j.is.2019.01.007.
- [4] M. Pourbafrani, S. Jiao, and W. M. P. van der Aalst, “SIMPT: Process Improvement Using Interactive Simulation of Time-Aware Process Trees,” in Research Challenges in Information Science, Springer International Publishing, 2021, pp. 588–594. doi: 10.1007/978-3-030-75018-3_40.
- [5] M. Dees, B. Hompes, and W. M. P. van der Aalst, “Events Put into Context (EPiC),” in 2020 2nd International Conference on Process Mining (ICPM), Oct. 2020, pp. 65–72. doi: 10.1109/icpm49681.2020.00020.
- [6] S. J. J. Leemans, K. Goel, and S. J. van Zelst, “Using Multi-Level Information in Hierarchical Process Mining: Balancing Behavioural Quality and Model Complexity,” in 2020 2nd International Conference on Process Mining (ICPM), Oct. 2020, pp. 137–144. doi: 10.1109/icpm49681.2020.00029.
- [7] R. Shraga, A. Gal, D. Schumacher, A. Senderovich, and M. Weidlich, “Inductive Context-aware Process Discovery,” in 2019 International Conference on Process Mining (ICPM), Jun. 2019, pp. 33–40. doi: 10.1109/icpm.2019.00016.
- [8] J. Munoz-Gama et al., “Process mining for healthcare: Characteristics and challenges,” J. Biomed. Informatics, vol. 127, p. 103994, 2022, doi: 10.1016/j.jbi.2022.103994.

- [9] A. Rozinat, “Process mining:conformance and extension,” Technische Universiteit Eindhoven, 2010. doi: 10.6100/IR690060.
- [10] M. De Leoni and W. M. P. van der Aalst, “Data-aware process mining,” 2013. doi: 10.1145/2480362.2480633.
- [11] M. De Leoni, M. Dumas, and L. García-Bañuelos, “Discovering Branching Conditions from Business Process Execution Logs,” Springer Berlin Heidelberg, 2013, pp. 114–129. doi: 10.1007/978-3-642-37057-1_9.
- [12] F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. P. van der Aalst, “Decision Mining Revisited - Discovering Overlapping Rules,” in Advanced Information Systems Engineering, Cham, 2016, pp. 377–392.
- [13] F. Mannhardt, “Multi-perspective process mining,” Technische Universiteit Eindhoven, 2018. doi:
- [14] P. Felli, M. de Leoni, and M. Montali, “Soundness Verification of Decision-Aware Process Models with Variable-to-Variable Conditions,” in 2019 19th International Conference on Application of Concurrency to System Design (ACSD), Jun. 2019, pp. 82–91. doi: 10.1109/acsd.2019.00013.
- [15] K. Winter and S. Rinderle-Ma, “Discovering Instance-Spanning Constraints from Process Execution Logs Based on Classification Techniques,” Oct. 2017. doi: 10.1109/edoc.2017.20.

- [16] K. Winter, F. Stertz, and S. Rinderle-Ma, “Discovering instance and process spanning constraints from process execution logs,” Information Systems, vol. 89, p. 101484, Mar. 2020, doi: 10.1016/j.is.2019.101484.
- [17] M. Pesic, H. Schonenberg, and W. M. P. van der Aalst, “DECLARE: Full Support for Loosely-Structured Processes,” in 11th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2007), Oct. 2007, p. 287. doi: 10.1109/edoc.2007.14.
- [18] V. Leno, M. Dumas, F. M. Maggi, M. L. Rosa, and A. Polyvyanyy, “Automated discovery of declarative process models with correlated data conditions,” Information Systems, vol. 89, p. 101482, Mar. 2020, doi: 10.1016/j.is.2019.101482.
- [19] S. J. van Zelst, “Process mining with streaming data,” Technische Universiteit Eindhoven, 2019.
- [20] F. Stertz, S. Rinderle-Ma, and J. Mangler, “Analyzing Process Concept Drifts Based on Sensor Event Streams During Runtime,” in Lecture Notes in Computer Science, Springer International Publishing, 2020, pp. 202–219. doi: 10.1007/978-3-030-58666-9_12
- [21] T. Brockhoff, M. S. Uysal, and W. M. P. van der Aalst, “Time-aware Concept Drift Detection Using the Earth Mover’s Distance,” Oct. 2020. doi: 10.1109/icpm49681.2020.00016.
- [22] F. Mannhardt, M. de Leoni, H. A. Reijers, W. M. P. van der Aalst, and P. J. Toussaint, ”Guided Process Discovery – A pattern-based approach,” Information Systems, vol. 76, pp. 1–18, Jul. 2018, doi: 10.1016/j.is.2018.01.009.
- [23] J. D. Smedt, A. Yeshchenko, A. Polyvyanyy, J. D. Weerdt, and J. Mendling, “Process Model Forecasting Using Time Series Analysis of Event Sequence Data,” Springer International Publishing, 2021, pp. 47–61. doi: 10.1007/978-3-030-89022-3_5.

- [24] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” Journal of Management Information Systems, vol. 24, Art. no. 3, Dec. 2007, doi: 10.2753/mis0742-1222240302.
- [25] J. M. E. M. van der Werf, A. Polyvyanyy, B. R. van Wensveen, M. Brinkhuis, and H. A. Reijers, “All that Glitters Is Not Gold,” in Advanced Information Systems Engineering, Springer International Publishing, 2021, pp. 141–157. doi: 10.1007/978-3-030-79382-1_9.
- [26] A. Banham, S. J. J. Leemans, M. T. Wynn, and R. Andrews, “xPM: A Framework for Process Mining with Exogenous Data,” in Lecture Notes in Business Information Processing, Springer International Publishing, 2022, pp. 85–97. doi: 10.1007/978-3-030-98581-3_7.
- [27] A. E. W. Johnson et al., “MIMIC-III, a freely accessible critical care database,” Scientific Data, vol. 3, Art. no. 1, May 2016, doi: 10.1038/sdata.2016.35.