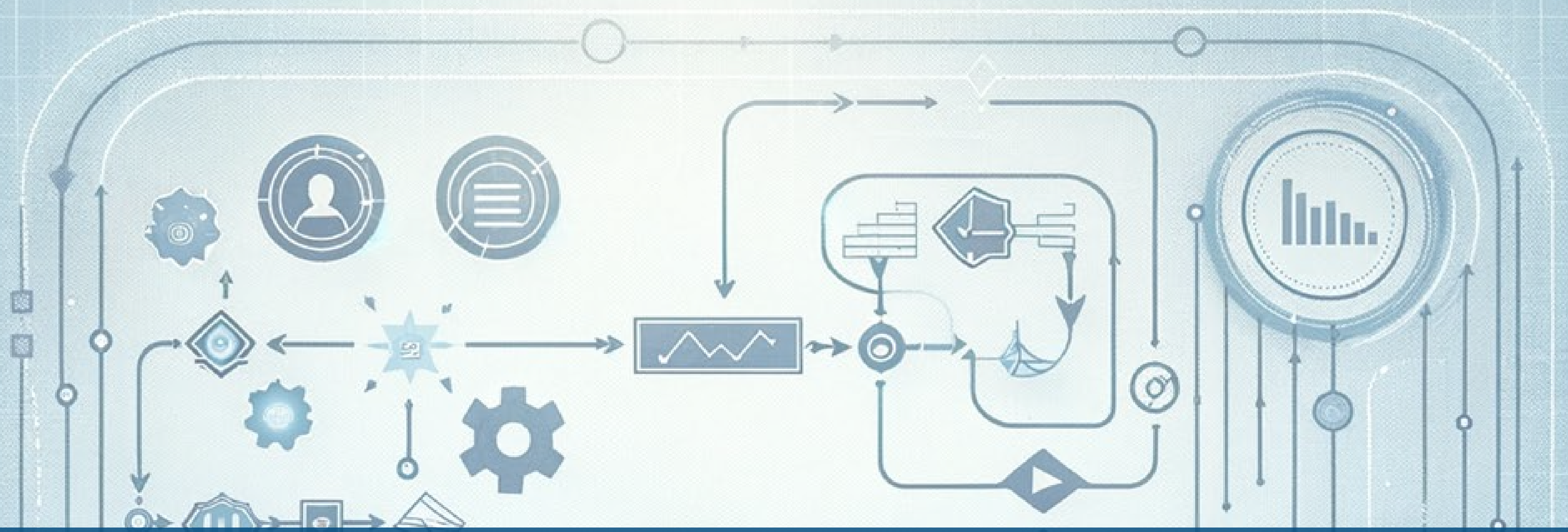


Goal-oriented Process Mining: A Scalability Experiment

Mahdi Ghasemi, **Daniel Amyot**, William Van Woensel (damyot@uottawa.ca)

15th International Model-Driven Requirements Engineering (MoDRE) Workshop
Valenciá, Spain, September 2, 2025

uOttawa

Process Mining Overview

Process Mining

Event data

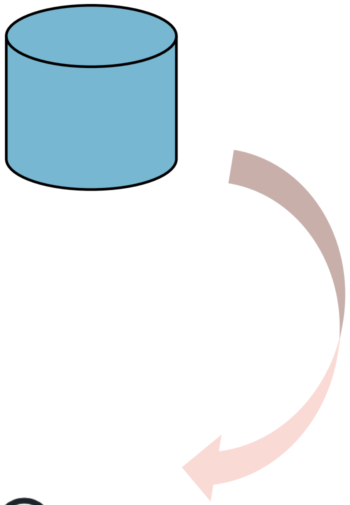
Case ID	Activity	Timestamp	Resource...
C0546	Confirm order	01/02/2025T08:29	Jane
C0546	Goods shipped	01/02/2025T09:02	Alex
C0546	Emit invoice	02/02/2025T07:35	Bob
C0479	Reject order	02/02/2025T08:25	Jane

event!

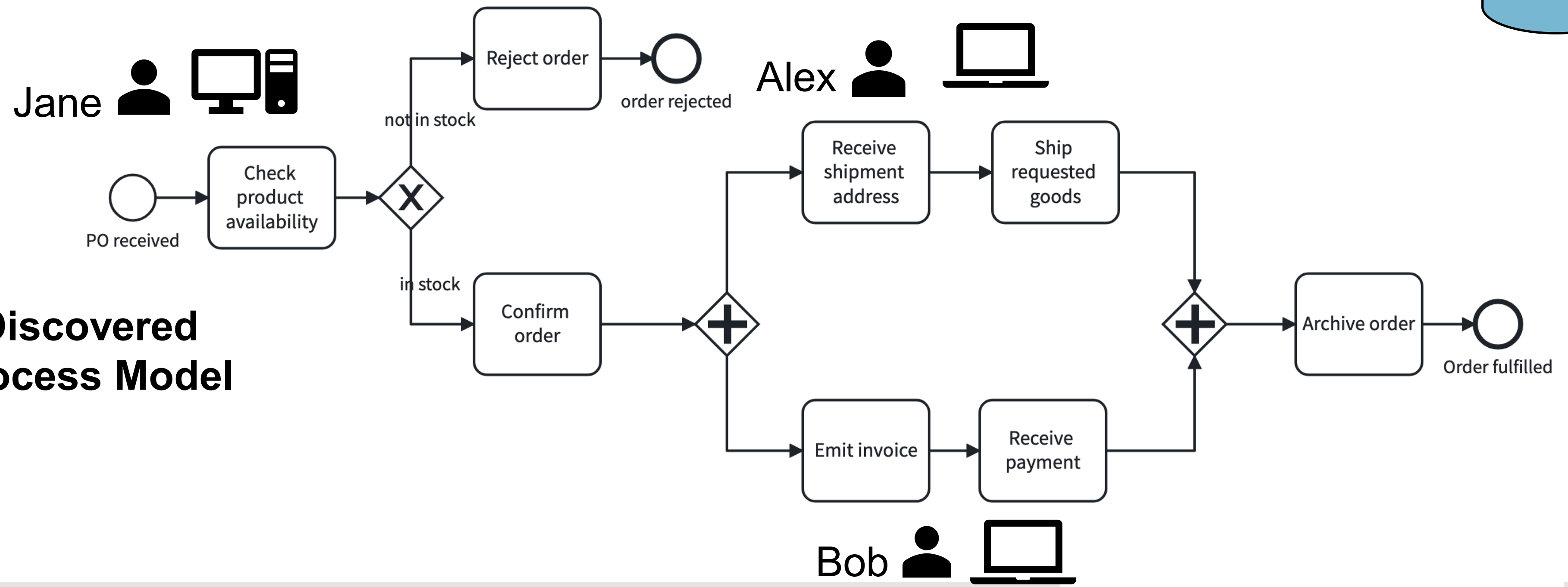
✓distinguish cases

✓ recorded activities

✓ timestamped
(when it took place!)



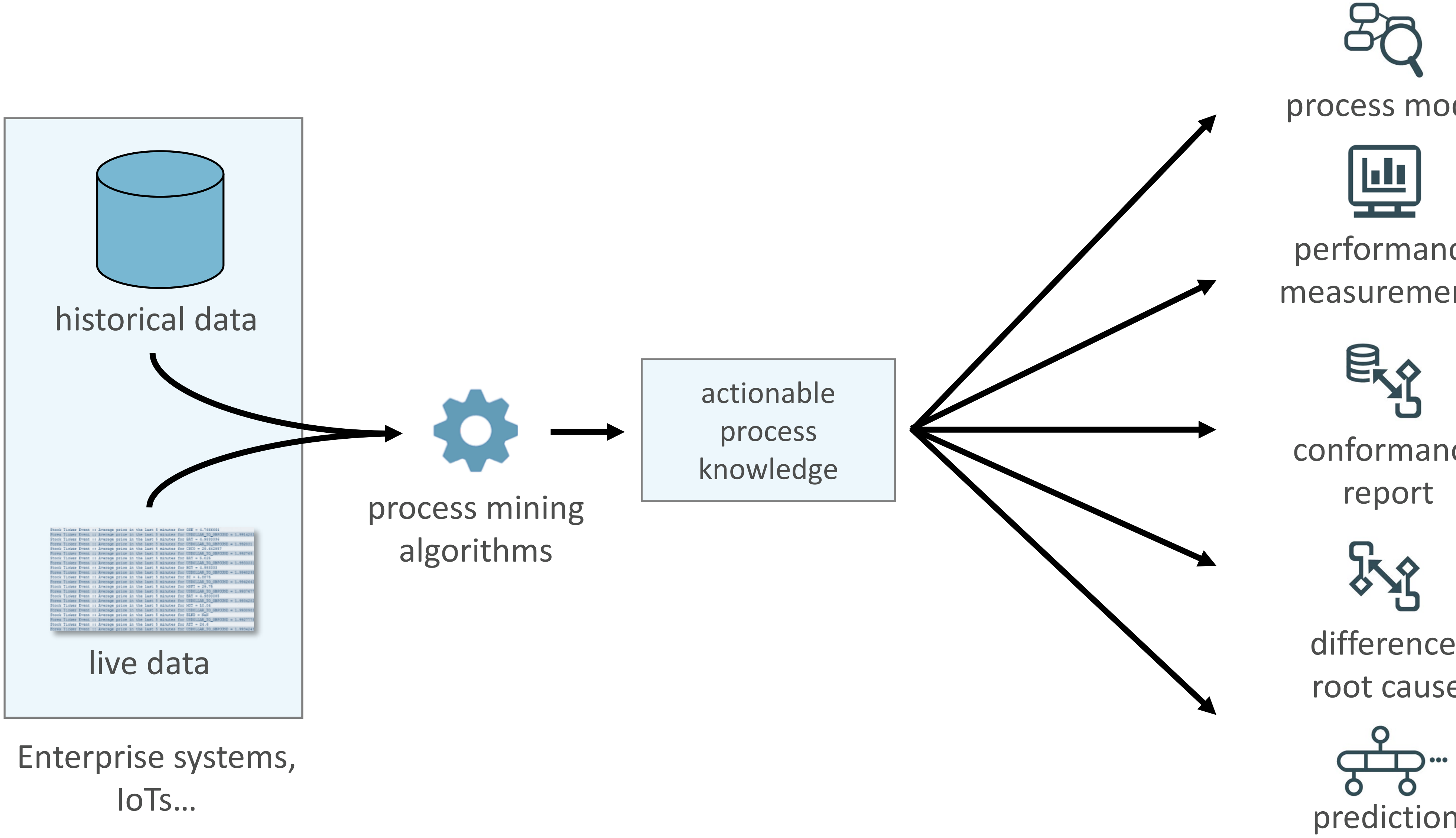
Discovered Process Model



Why? Data-Driven Requirements Elicitation!

Credit: Dr. Mahdi Ghasemi

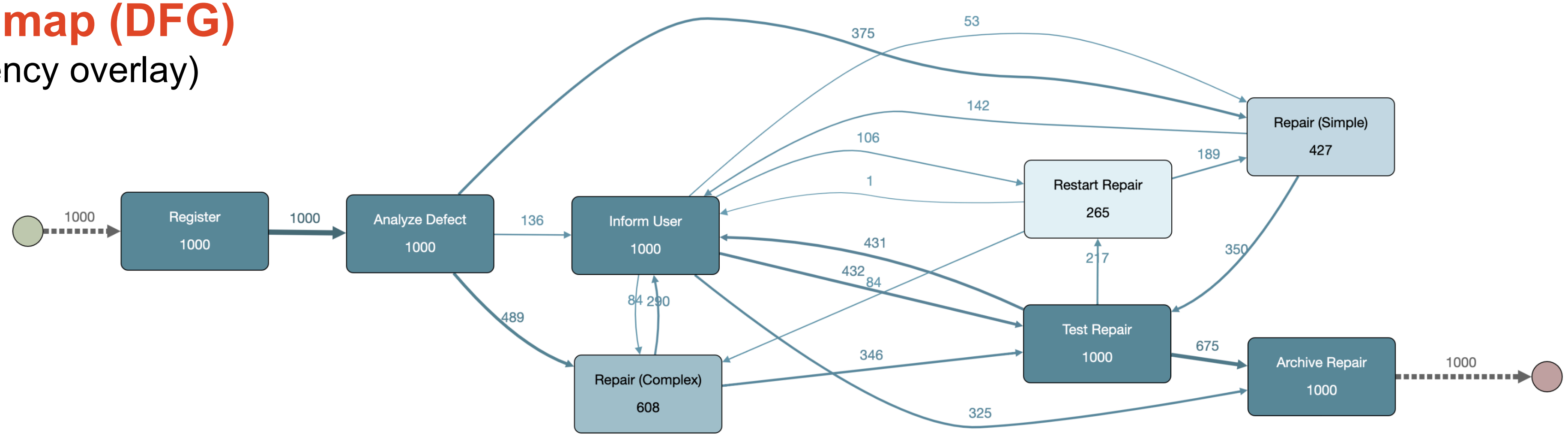
Process Mining



Mined Models with *Frequencies* as Overlay

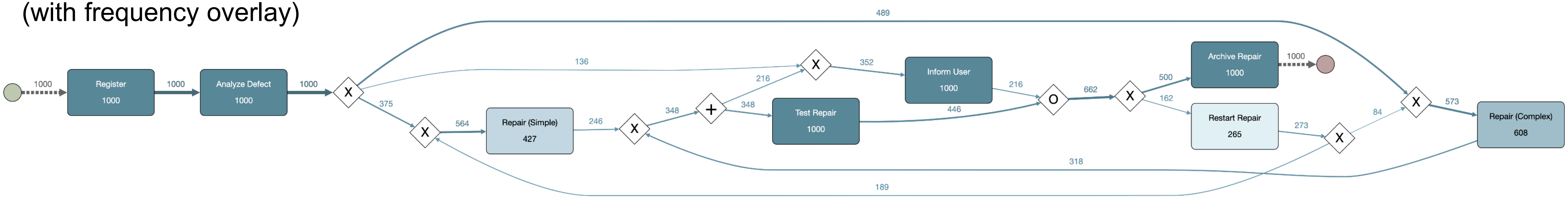
Process map (DFG)

(with frequency overlay)



BPMN model

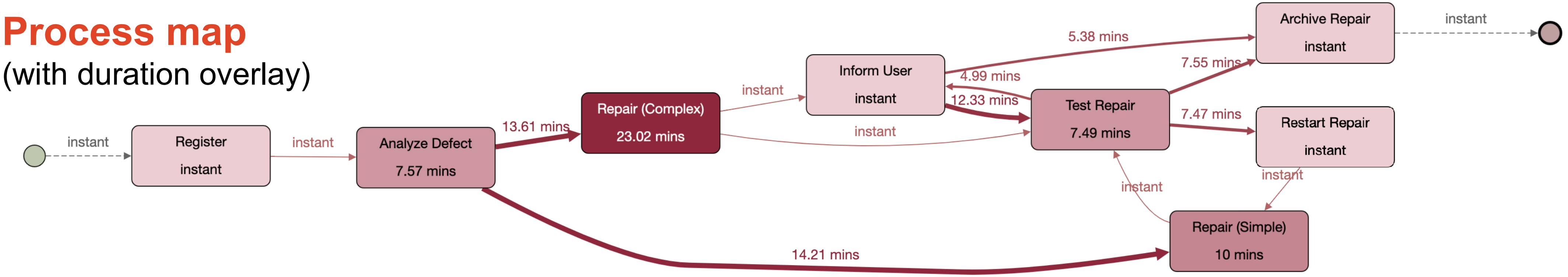
(with frequency overlay)



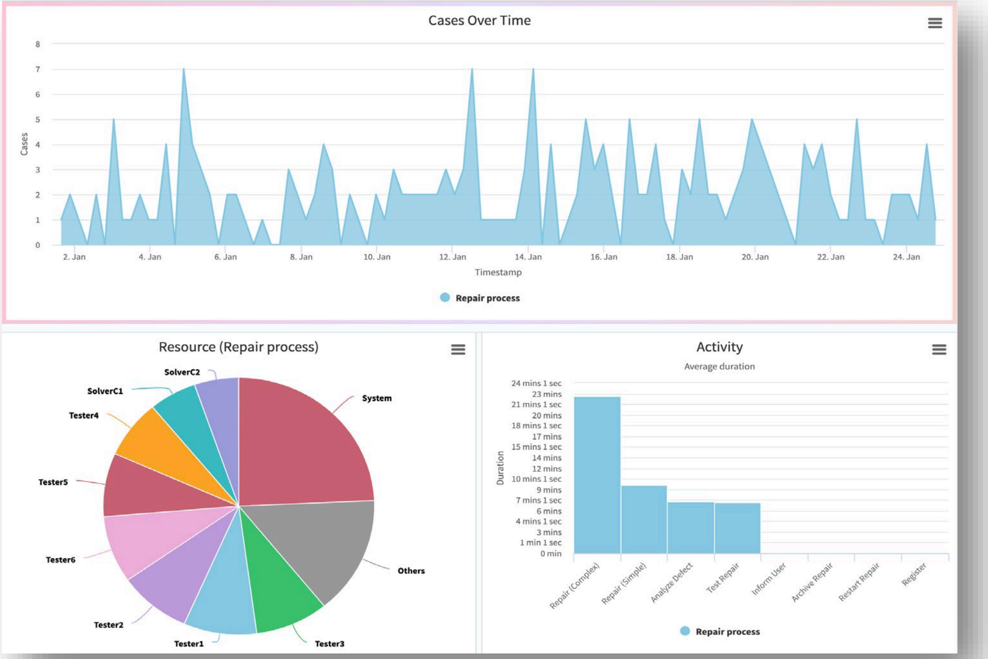
Mined Models with *Durations* as Overlay

Process map

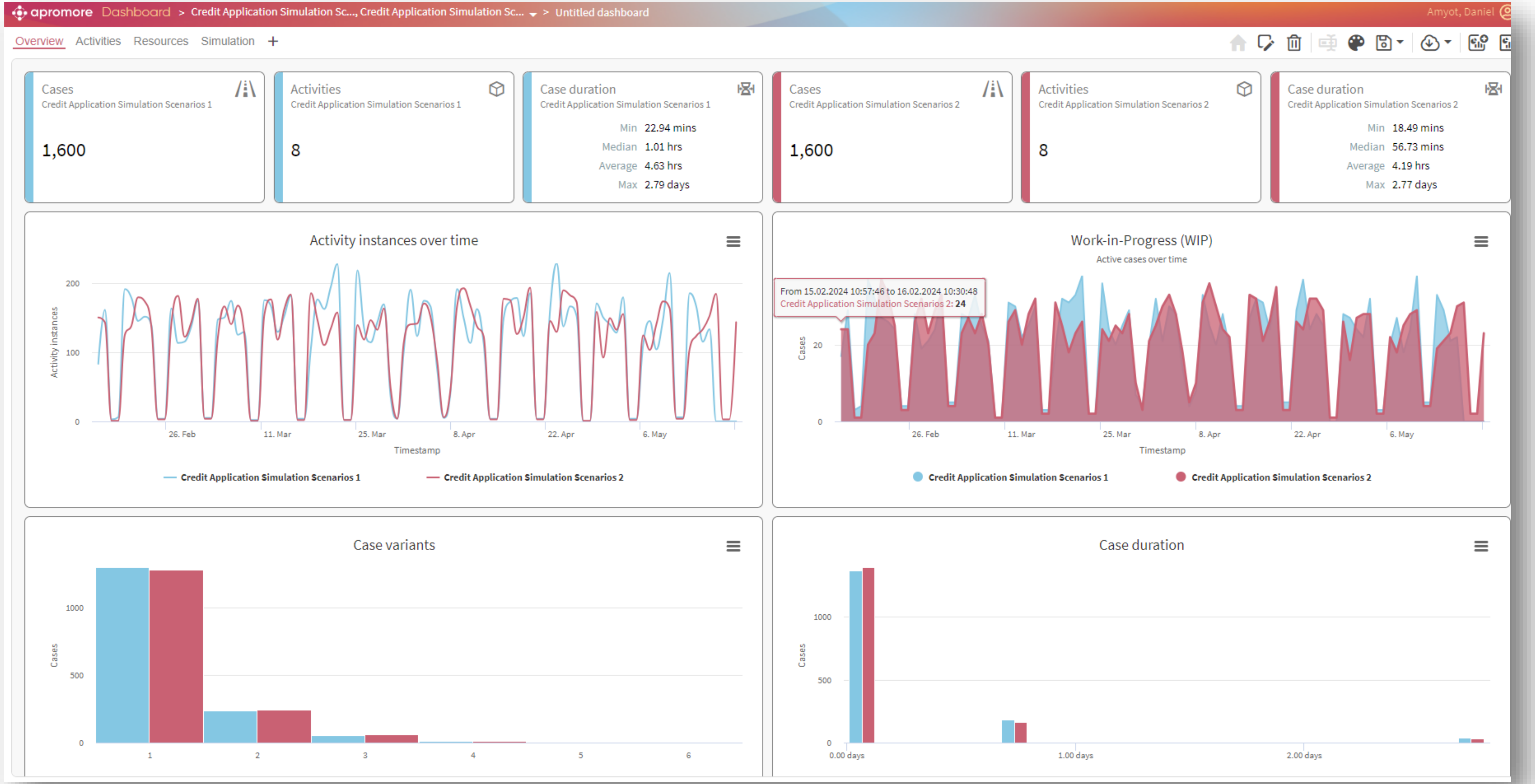
(with duration overlay)



Process performance dashboards

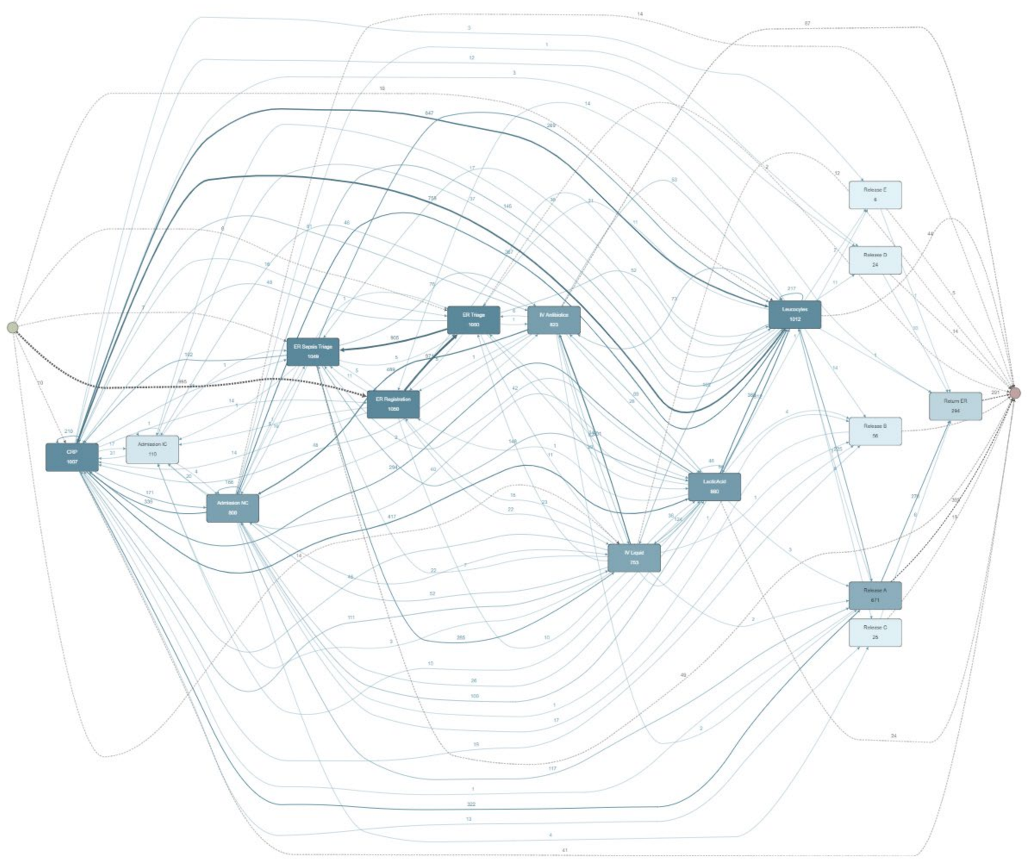


Simulations for What-If Analysis

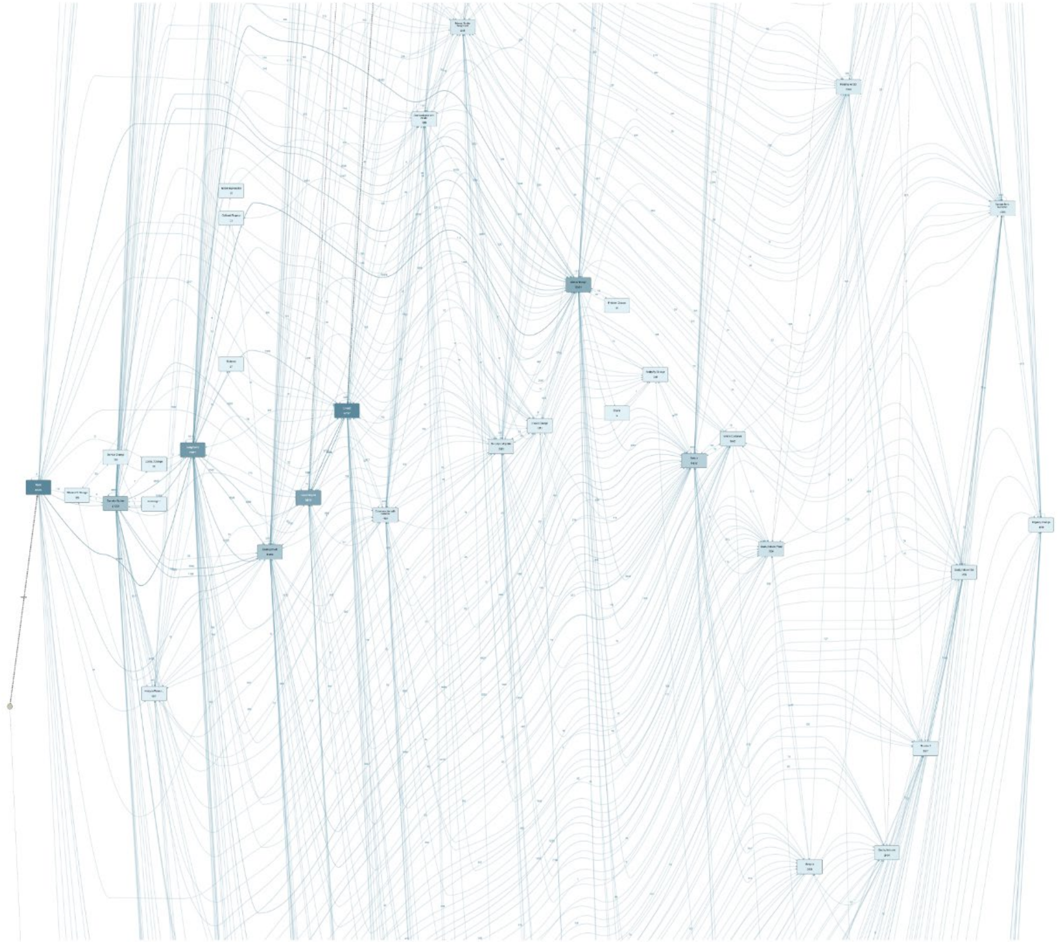


Goal-oriented Process Mining

Important Challenge: Complexity!



Patient Treatment Process @ Hospital (Sepsis infections)



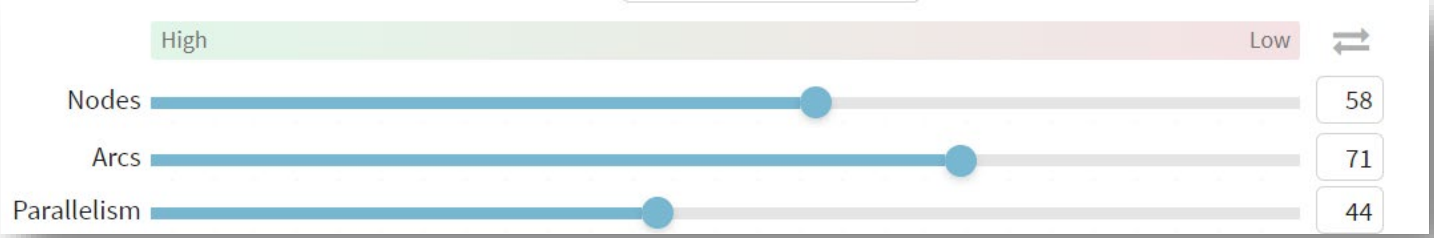
IT Incident Management @ Bank

Process Map: Abstraction and Filtering

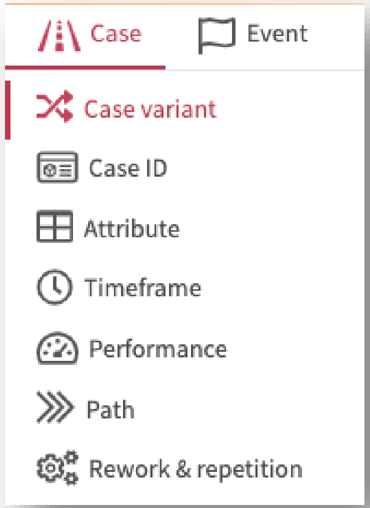
Abstraction: Group certain activities together

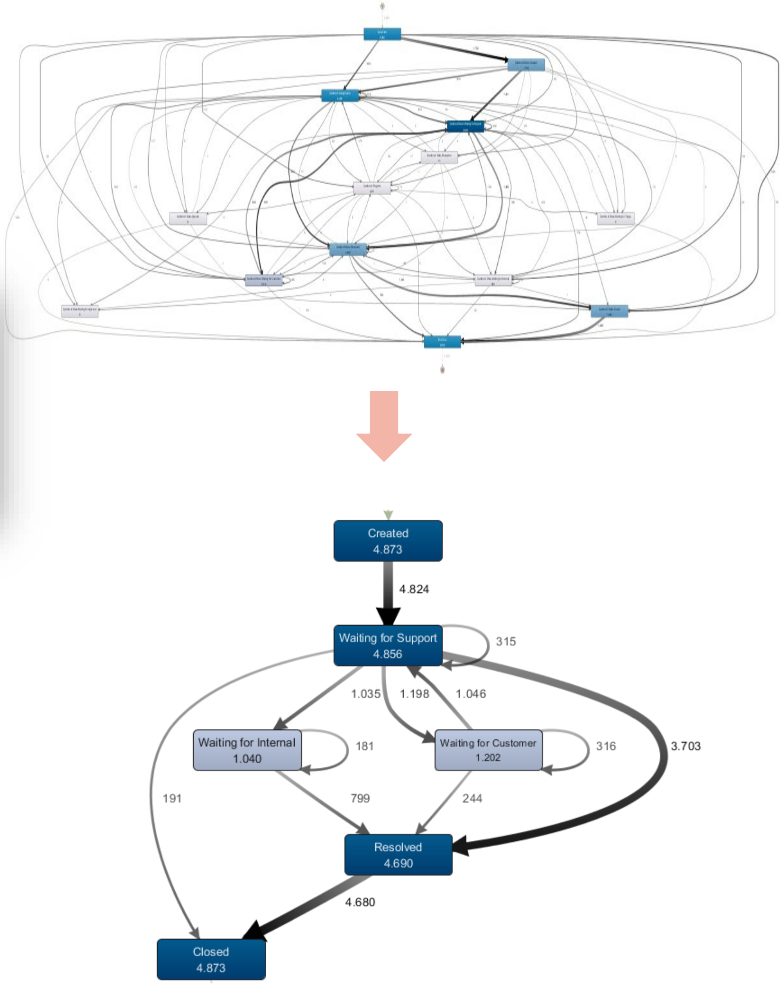
Filtering:

- Frequent activities or arcs



- Specific categories of cases or events





What about Filtering based on Goals?

Process Mining: focuses on “how”, “what”, “where”, “who”, and especially “when” questions

Goal-oriented modelling: focuses mainly on answering “why” questions

Potential for synergy!

Current process mining:

Log with all cases

→ Model

Goal-oriented Process Mining:

Log of cases that satisfy goals

→ Good model
(Goal-aligned model)

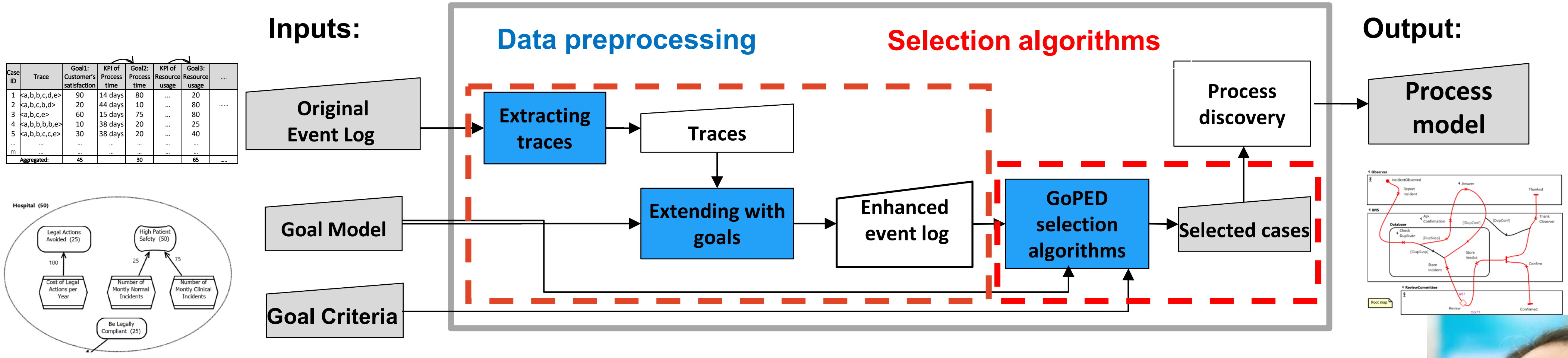
Log of cases that do not satisfy goals

→ “Bad” model

Ghasemi, M., Amyot D. (2020) [From event logs to goals: a systematic literature review of goal-oriented process mining](). *REJ*, 25(1), 67-93.

Goal-oriented Process Mining (GoPM)

GoPM enables the quantitative, goal-driven selection of relevant cases and variants in an event log.

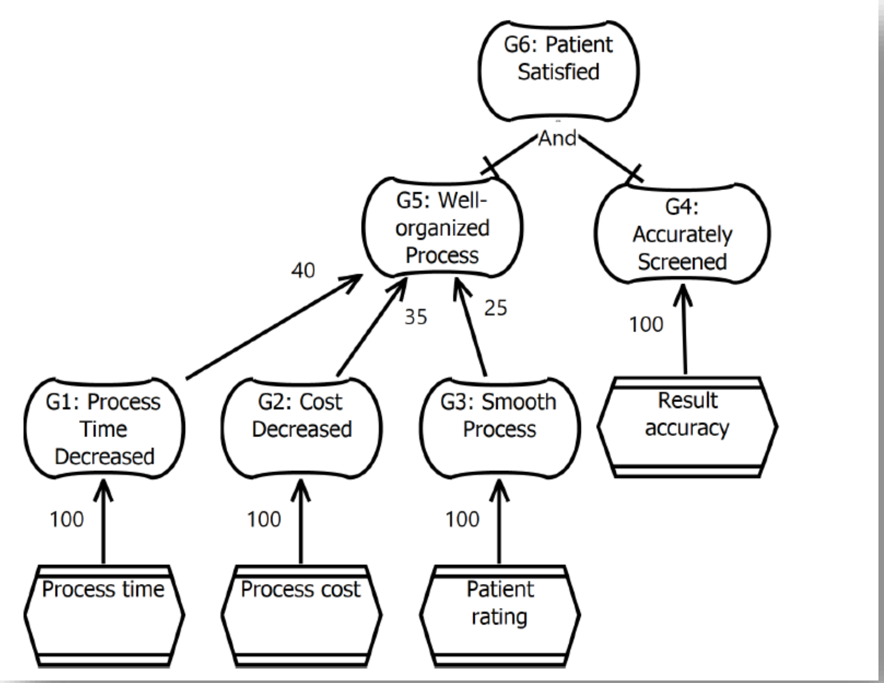


Ghasemi (2022) Goal-oriented Process Mining.
PhD thesis, DTI, University of Ottawa, 2022
http://dx.doi.org/10.20381/ruor-27301

GoPM Inputs

DGD EVENT LOG FOR 10 PATIENTS, WITH CURRENT DATA ATTRIBUTES.

Case	Trace	Days	Cost	Rating	Accuracy
C_1	$\langle a, b, c, g \rangle$	4	400	9	1
C_2	$\langle a, b, c, g \rangle$	5	400	9	1
C_3	$\langle a, b, c, g \rangle$	5	400	9	0
C_4	$\langle a, b, c, d, e, c, g \rangle$	11	850	8	1
C_5	$\langle a, b, c, d, e, c, g \rangle$	9	850	7	1
C_6	$\langle a, b, c, d, e, c, g \rangle$	10	850	8	1
C_7	$\langle a, b, c, f, b, c, g \rangle$	8	600	7	1
C_8	$\langle a, b, c, f, b, c, d, e, c, g \rangle$	17	1100	6	1
C_9	$\langle a, b, c, f, b, c, d, e, c, g \rangle$	16	1100	5	1
C_10	$\langle a, b, c, d, b, c, d, e, c, g \rangle$	31	1150	4	1

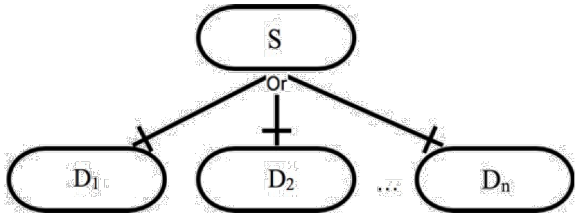


KPI DEFINITIONS FOR THE DGD PROCESS.

Indicator	Linked Goal	Worst v.	Threshold v.	Target v.
Process time (days)	Process time decreased	35	13	4
Process cost ($)	Cost decreased	1200	950	400
Patient rating	Smooth process	1	6	10
Result Accuracy	Accurately screened	0	–	1

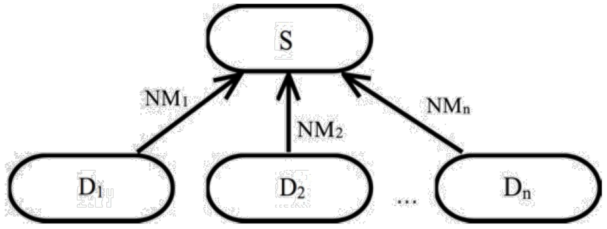
GRL Arithmetic Semantics

OR-Decomposition Links



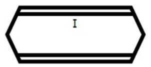
$$v(S) = Max(v(D_1), v(D_2), ..., v(D_n))$$

Contribution Links



$$v(S) = Max(0, Min(100, \frac{\sum_{x=1}^{n}(v(D_x) \times NM_x)}{100}))$$

Indicators



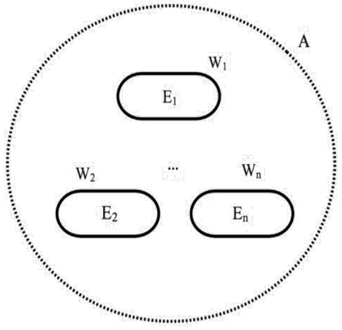
C: Current
T: Target
TH: Threshold
W: Worst

$$v(I) = \begin{cases} 100 & \text{if } C \geq T \\ 0 & \text{if } C \leq W \\ Abs(\frac{C-TH}{T-TH}) \times 50 + 50 & \text{if } TH \leq C < T \\ -Abs(\frac{C-TH}{W-TH}) \times 50 + 50 & \text{if } W < C < TH \end{cases}$$

When $T < W$

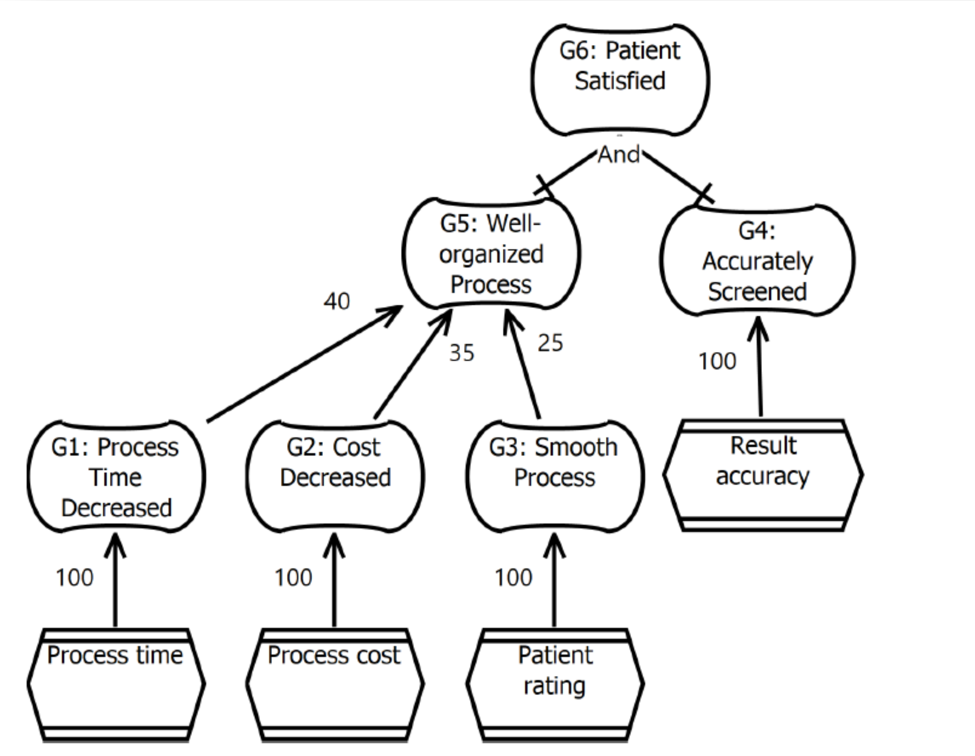
$$v(I) = \begin{cases} 100 & \text{if } C \leq T \\ 0 & \text{if } C \geq W \\ Abs(\frac{C-TH}{TH-T}) \times 50 + 50 & \text{if } T < C \leq TH \\ -Abs(\frac{C-TH}{TH-W}) \times 50 + 50 & \text{if } TH < C < W \end{cases}$$

Actors



$$v(A) = Max(0, Min(100, \frac{\sum_{x=1}^{n}(v(E_x) \times W_x)}{Max(100, \sum_{x=1}^{n}(W_x))}))$$

Log Enhanced with Goal Satisfactions



ENHANCED LOG OF THE DGD PROCESS: ADDITIONAL GOAL SATISFACTION LEVELS, WITH AGGREGATED VALUES.

Case	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**
C_1	100	100	88	100	97	97
C_2	94	100	88	100	95	95
C_3	94	100	88	0	95	0
C_4	61	59	75	100	64	64
C_5	72	59	63	100	65	65
C_6	67	59	75	100	66	66
C_7	78	82	63	100	76	76
C_8	41	20	50	100	36	36
C_9	43	20	40	100	34	34
C_10	9	10	30	100	15	15
Aggregate satisfaction:	**65.9**	**60.9**	**66**	**90**	**64.1**	**54.8**

GoPED Criteria

Case	Trace	Goal 1	Goal 2	…	Goal n	Overall
c_1	t_1	$s_{1,1}$	$s_{1,2}$	…	$s_{1,n}$	$s_{1.Ove}$
c_2	…	$s_{2,1}$	$s_{2,2}$	…	$s_{2,n}$	$s_{2.Ove}$
…	…					…
c_m	t_m	$s_{m,1}$	$s_{m,2}$	…	$s_{m,n}$	$s_{m.Ove}$
Aggregated satisfaction:		$s_{Agg.1}$	$s_{Agg.2}$	…	$s_{Agg.n}$	s_{Comp}

1 2 3

Three criteria for model discovery in GoPED:

1. **Case** perspective: satisfaction level for *one or multiple goals for all cases*
2. **Goal** perspective: *aggregated* satisfaction level of *one or multiple goals*
3. **Organization** perspective: *comprehensive* satisfaction level

Goal-oriented Process Enhancement and Discovery (GoPED) Algorithms

Algorithm 1: Case_Perspective

Input: *EnhancedLog* ; `// Log enhanced with goals`
Input: Q_{case}: Set(criteria) ; `// Each criterion is a triple <goal, operator, value>`
Input: *conf*: number ; `// Confidence level`
Output: *CasesKept*: Set(cases)

SortByTrace(*EnhancedLog*);
NumCases ← NumberOfCases(*EnhancedLog*);
$trace(case_0) \leftarrow \langle\rangle$; `// Add empty trace before log`
$trace(case_{NumCases+1}) \leftarrow \langle\rangle$; `// ... and after log`
CasesKept ← ∅;
index ← 1;
while $index \leq NumCases$ **do**
  SameTraceC ← ∅ ; `// Cases with same traces`
  NumSatCasesOfTrace ← 0;
  repeat
    $SameTraceC \leftarrow SameTraceC \cup \{case_{index}\}$;
    if $case_{index}$ *meets all criteria of* Q_{case} **then**
      NumSatCasesOfTrace++;
    index++;
  until $trace(case_{index}) \neq trace(case_{index-1})$;
  if $NumSatCasesOfTrace/|SameTraceC| \geq conf$ **then** ; `// Keep case if confidence level is met`
    $CasesKept \leftarrow CasesKept \cup SameTraceC$;
return *CasesKept*;

Algorithm 2: Goal_Perspective

Input: *EnhancedLog* ; `// Log enhanced with goals`
Input: Q_{goal}: Set(criteria) ; `// <goal, threshold>`
Input: *G* ; `// Aggregation funct., one per goal`
Output: *CasesKept*: Set(cases)

$m \leftarrow$ NumberOfCases(*EnhancedLog*) ; `// NumCases`
CasesKept ← ∅;
Solve this binary optimization ; `//` x_i`: when equal to 1, keep case` c_i `;` $s_{i,j}$`: satisfaction level of goal` j `for case` c_i

> **Maximize** $z = \sum_{i=1}^{m} x_i$ **s.t.**
> $\forall r, t,\ 1 \leq r < t \leq m :$ `// All-or-none rule`
> $\quad \text{trace}(c_r) = \text{trace}(c_t) \Rightarrow x_r = x_t$
> `// Ensure that` Q_{goal} `constraints are met`
> $\forall j$ where $G_j \in G : \frac{\sum_{i=1}^{m} x_i \cdot s_{i,j}}{\sum_{i=1}^{m} x_i} \geq \text{threshold}_j$
> $x_i \in \{0, 1\}$ `// Two potential values for` x_i

for $i = 1$ *to* m **do**
  if $x_i = 1$ **then**
    $CasesKept \leftarrow CasesKept \cup \{c_i\}$
return *CasesKept*;

Algorithm 3: Organization_Perspective

Input: *EnhancedLog* ; `// Log enhanced with goals`
Input: Q_{comp}: $<oper \in \{\leq, =, \geq\}, val \in [0..100]>$
Input: *G* ; `// Goal model function`
Output: *CasesKept*: Set(cases)

$m \leftarrow$ NumberOfCases(*EnhancedLog*) ; `// NumCases`
CasesKept ← ∅;
Solve this binary optimization ; `//` x_i`: when equal to 1, keep case` c_i `;` $s_{i,j}$`: satisfaction level of goal` j `for case` c_i

> **Maximize** $z = \sum_{i=1}^{m} x_i$ **s.t.**
> $\forall r, t,\ 1 \leq r < t \leq m :$ `// All-or-none rule`
> $\quad \text{trace}(c_r) = \text{trace}(c_t) \Rightarrow x_r = x_t$
> `// Ensure that` Q_{comp} `constraint is met`
> $G\left(\frac{\sum_{i=1}^{m} x_i \cdot s_{i,1}}{\sum_{i=1}^{m} x_i}, \ldots, \frac{\sum_{i=1}^{m} x_i \cdot s_{i,n}}{\sum_{i=1}^{m} x_i}\right) < oper >< val >$
> $x_i \in \{0, 1\}$ `// Two potential values for` x_i

for $i = 1$ *to* m **do**
  if $x_i = 1$ **then**
    $CasesKept \leftarrow CasesKept \cup \{c_i\}$
return *CasesKept*;

Variant Selection: All Cases or None

ENHANCED LOG OF THE DGD PROCESS: ADDITIONAL GOAL SATISFACTION LEVELS, WITH AGGREGATED VALUES.

Case	G1	G2	G3	G4	G5	G6
C_1	100	100	88	100	97	97
C_2	94	100	88	100	95	95
C_3	94	100	88	0	95	0
C_4	61	59	75	100	64	64
C_5	72	59	63	100	65	65
C_6	67	59	75	100	66	66
C_7	78	82	63	100	76	76
C_8	41	20	50	100	36	36
C_9	43	20	40	100	34	34
C_10	9	10	30	100	15	15
Aggregate satisfaction:	**65.9**	**60.9**	**66**	**90**	**64.1**	**54.8**

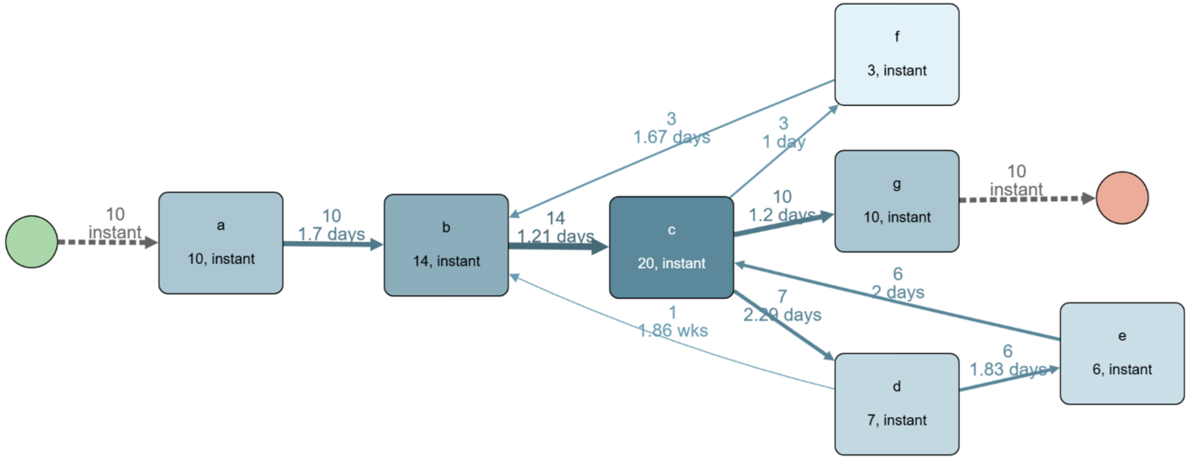
Filter for Case Perspective (Algorithm 1)

Select case variants where:

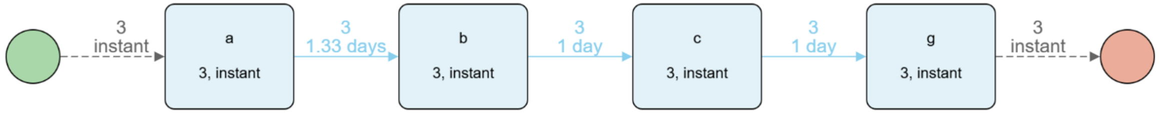
- **G1 ≥ 80 ∧ G4 ≥ 100**
- with a confidence **≥ 60%**

For the first variant (cases C_1 to C_3), 2 out of 3 cases (67%) satisfy these goal-oriented criteria; this variant meets the required confidence level. None of the other variants does.

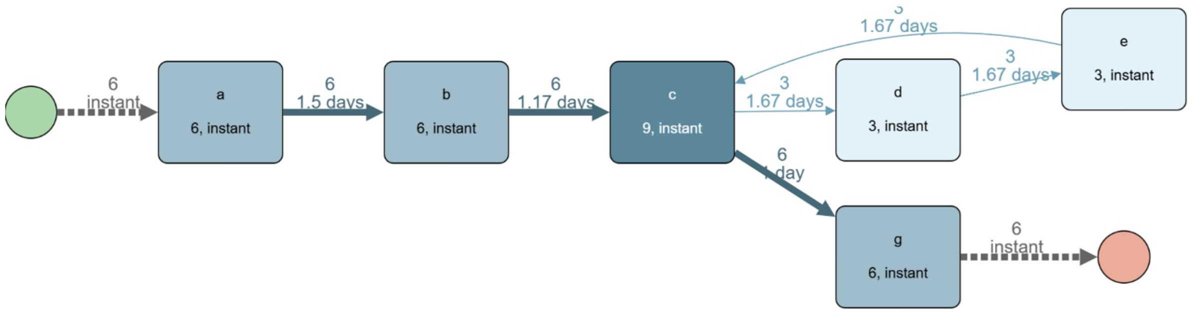
GoPED Application to Enhanced Event Log



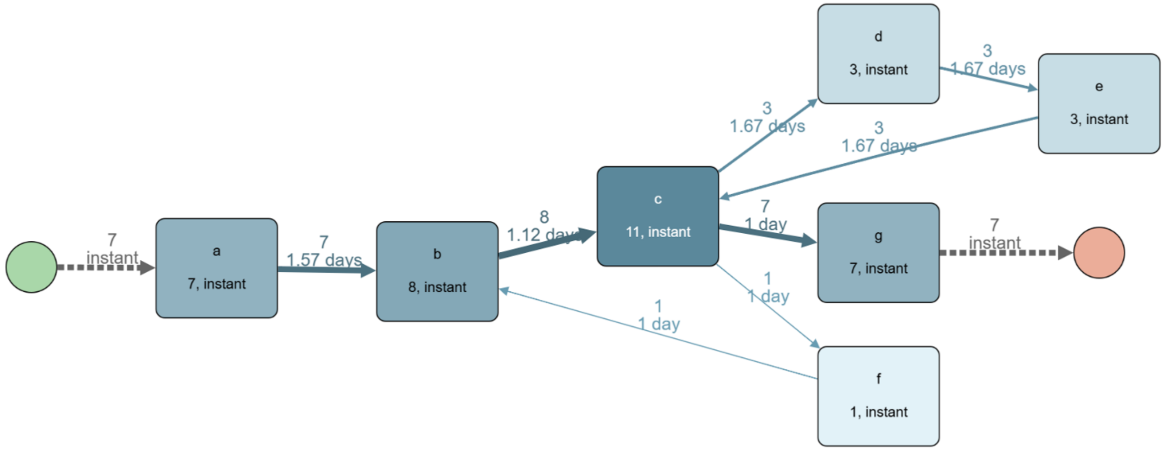
a) Process model discovered from the original event log by Apromore



b) Algo 1 (*case perspective*) with G1≥80, G4≥100, and confidence = 60%



c) Algo 2 (*goal perspective*) with aggregated G1 (=81.3) ≥80 and aggregated G3 (=79.5) ≥78



d) Algo 3 (*organization perspective*) with comprehensive satisfaction (=66.1) ≥ 65

Scalability Experiment

Objectives

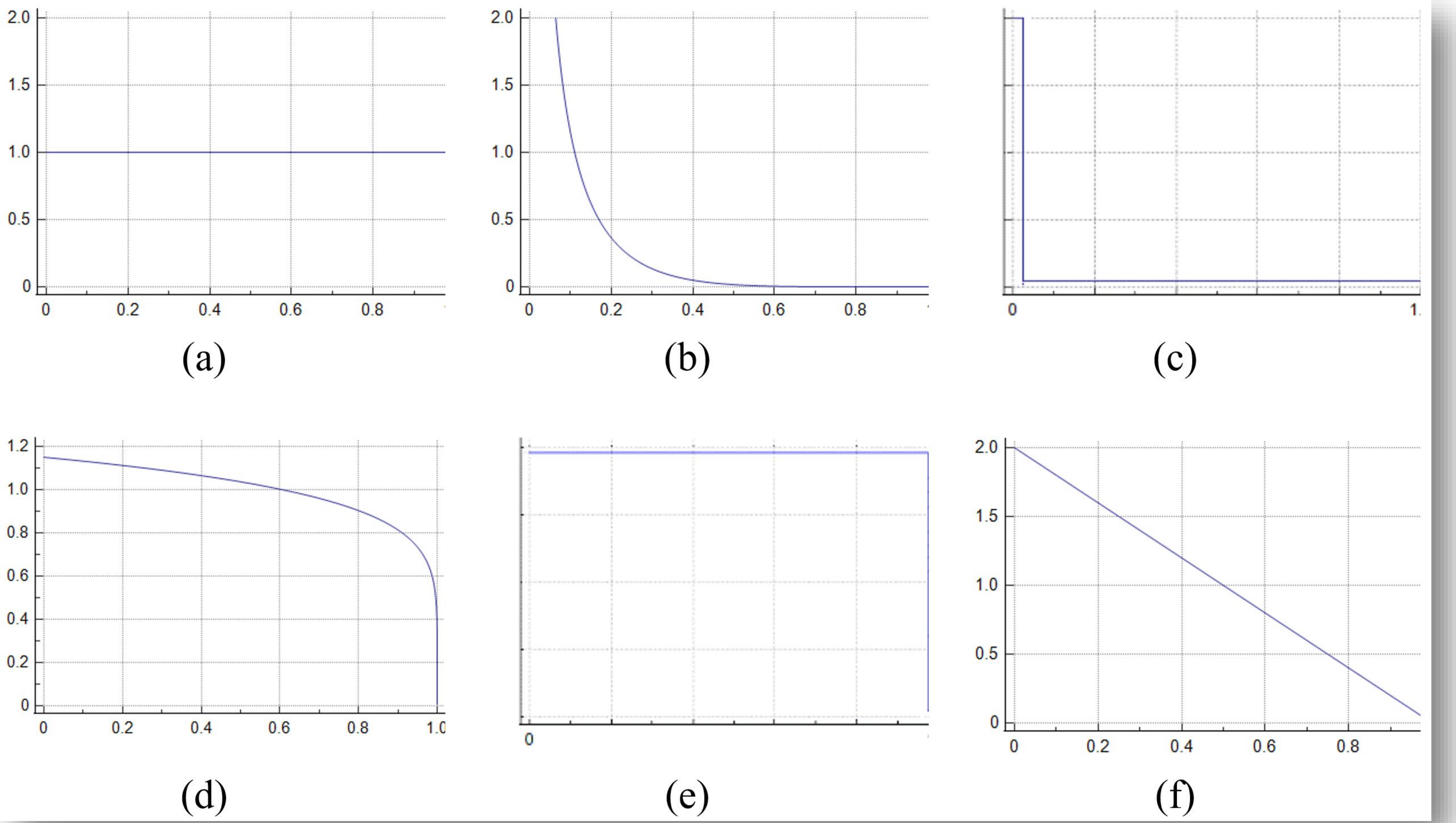
Evaluate the algorithms' sensitivity to 4 event log factors:

- (L1) distribution of cases among variants
- (L2) number of cases
- (L3) number of traces
- (L4) length of traces

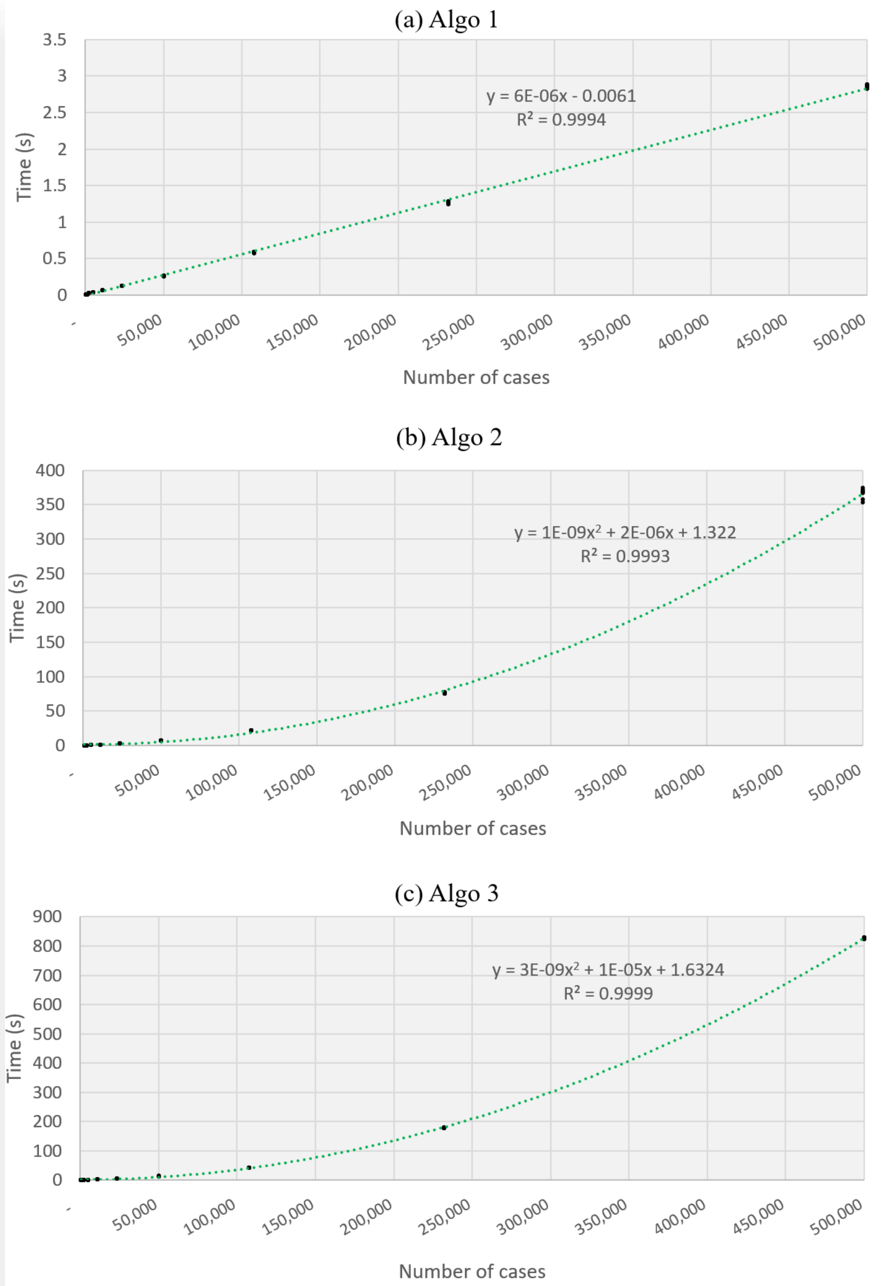
And 2 goal factors:

- (G1) number of goal criteria
- (G2) the goal criteria's boundaries on satisfaction levels

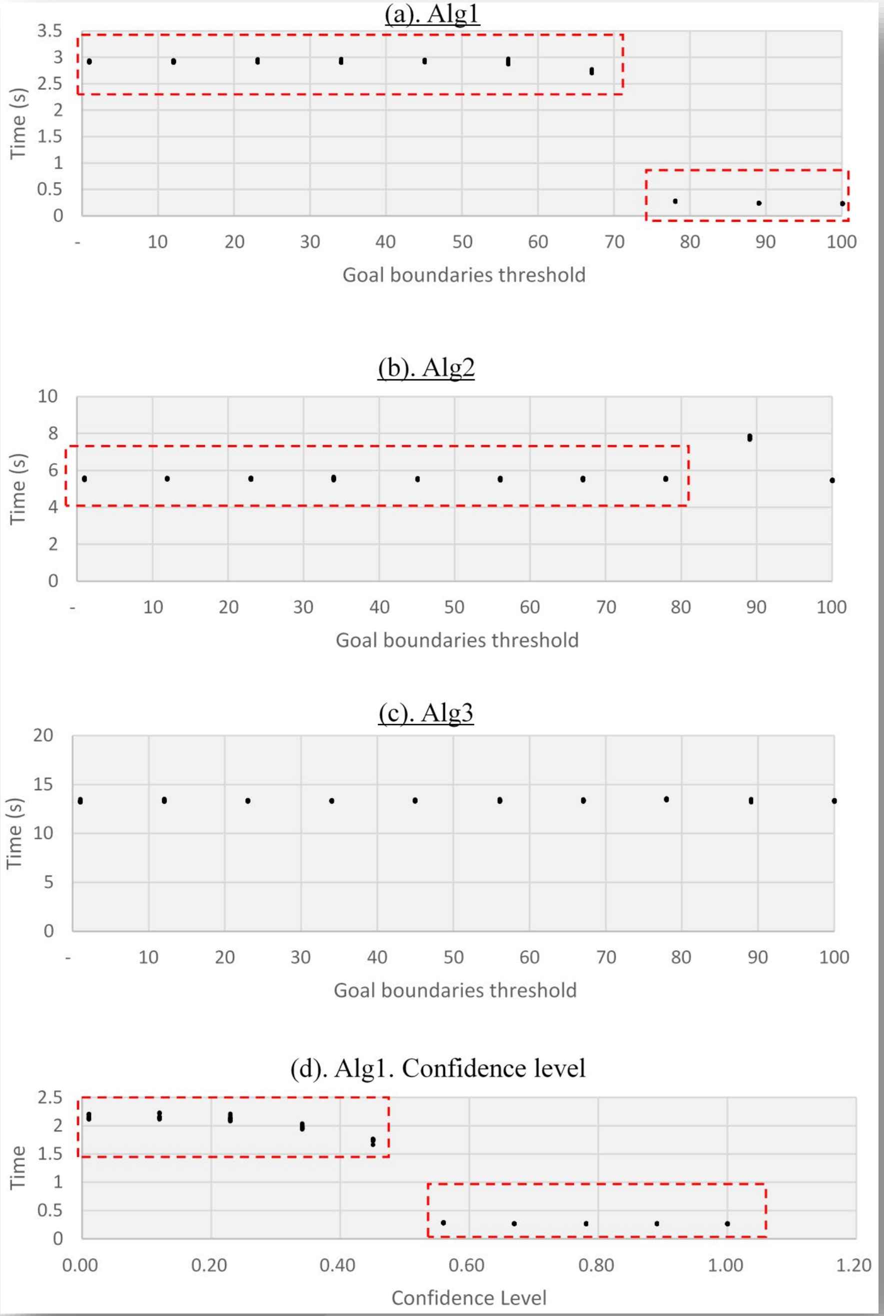
Distribution Formats



(L1) distribution of cases among variants



(G2) the goal criteria's boundaries on satisfaction levels



Correlation between Algorithm Runtimes and Factors

Factor	Algo. 1	Algo. 2	Algo. 3
(L1) Dist. cases / traces	No corr.	No corr.	No corr.
(L2) Number of cases	Pos., Linear	Pos., Quadr.	Pos., Quadr.
(L3) Number of traces	Pos., Quadr.	Neg., Linear	Neg., Linear
(L4) Length of traces	Pos., Linear	Pos., Linear	Pos., Linear
(G1) Num. cons. goals	Pos., Linear	Pos., Linear	Pos., Linear
(G2) Criteria's bounds	No corr.	No corr.	No corr.

… and fast enough to be used in practice

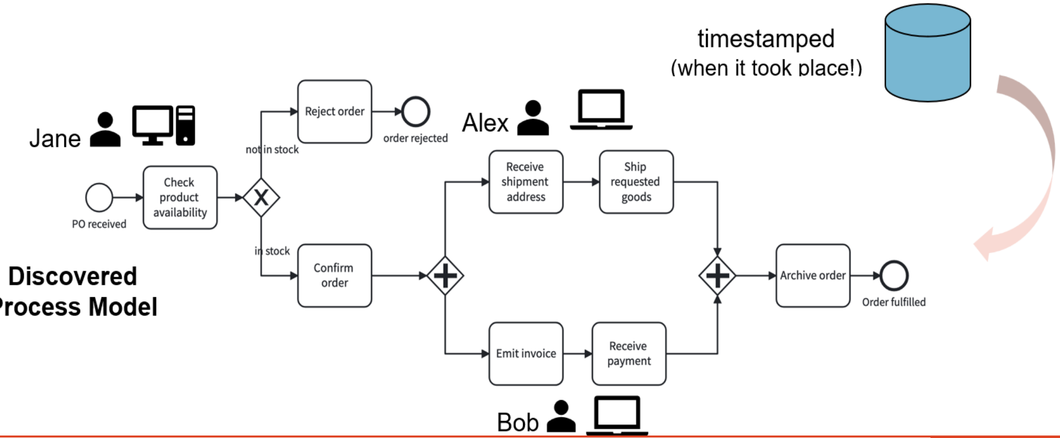
Some Challenges

- Common absence of goal-related information in event logs, and of goal models
- Tolerance to noise in event logs, in variant selection
- Concept drift in the requirements and goals.

Process Mining

Event data

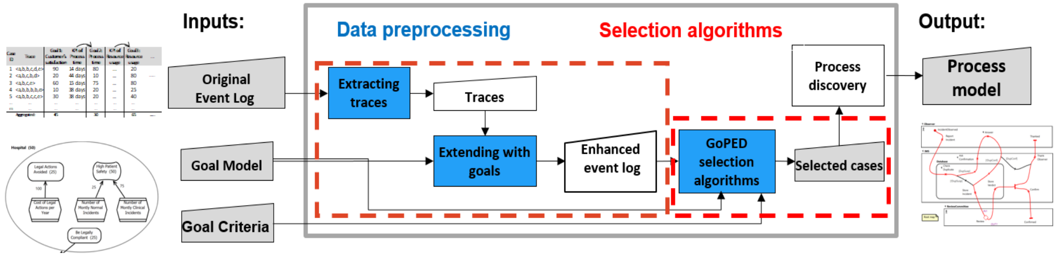
Case ID	Activity	Timestamp	Resource...
C0546	Confirm order	01/02/2025T08:29	Jane
C0546	Goods shipped	01/02/2025T09:02	Alex
C0546	Emit invoice	02/02/2025T07:35	Bob
C0479	Reject order	02/02/2025T08:25	Jane



Discovered Process Model

Goal-oriented Process Mining (GoPM)

GoPM enables the quantitative, goal-driven selection of relevant cases and variants in an event log.



Correlation between Algorithm Runtimes and Factors

Factor	Algo. 1	Algo. 2	Algo. 3
(L1) Dist. cases / traces	No corr.	No corr.	No corr.
(L2) Number of cases	Pos., Linear	Pos., Quadr.	Pos., Quadr.
(L3) Number of traces	Pos., Quadr.	Neg., Linear	Neg., Linear
(L4) Length of traces	Pos., Linear	Pos., Linear	Pos., Linear
(G1) Num. cons. goals	Pos., Linear	Pos., Linear	Pos., Linear
(G2) Criteria's bounds	No corr.	No corr.	No corr.

… and fast enough to be used in practice

GoPM: A Clever Way of Simplifying Models!