

# Data Preprocessing for Goal-oriented Process Discovery

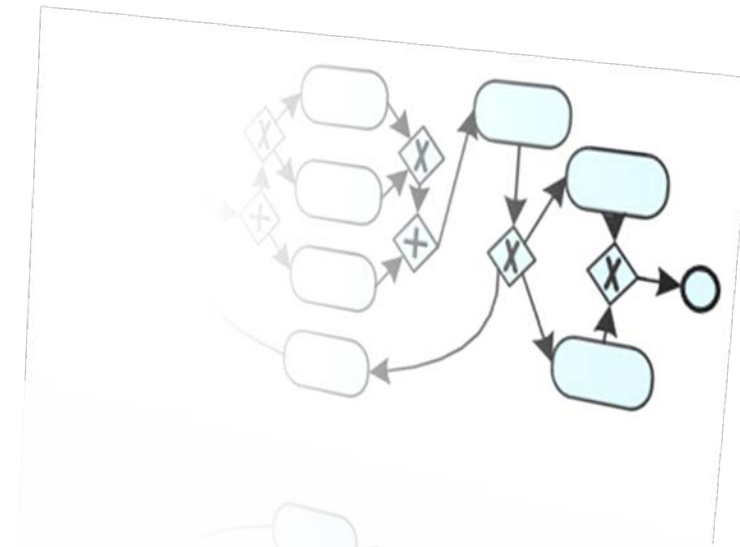
Mahdi Ghasemi, Daniel Amyot

University of Ottawa, Canada

{mghasemi, damyot}@uottawa.ca

CrowdRE'19, Jeju, South Korea

September 24, 2019



# What is *process mining*?

**Event log**

Case	Activity	Timestamp	Resource
432	register travel request (a)	18-3-2014:9.15	John
432	get support from local manager (b)	18-3-2014:9.25	Mary
432	check budget by finance (d)	19-3-2014:8.55	John
433	register travel request (a)	19-3-2014:9.02	Adrian
432	decide (e)	19-3-2014:9.36	Sue
433	get support from local manager (b)	19-3-2014:9.42	Mary
432	accept request (g)	19-3-2014:9.48	Mary

Case ID

Activity

Timestamp

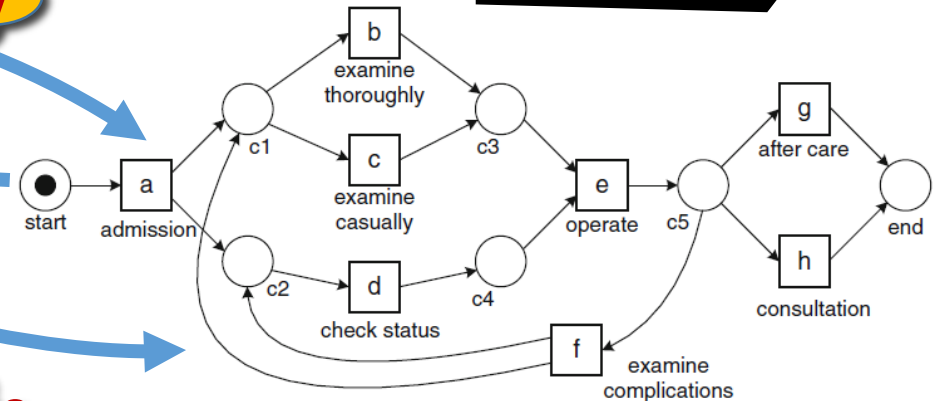
Three required fields to discover a model from crowd data logs

**Discovery**

**Enhancement**

**Conformance**

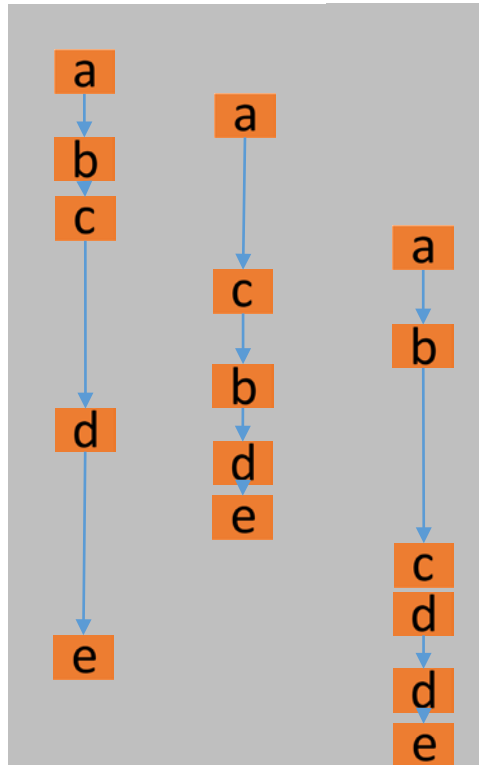
**Process model**



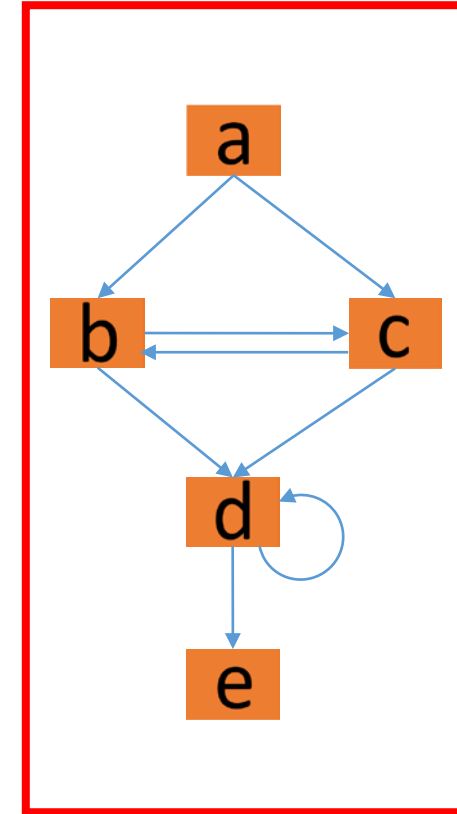
## Event Logs

Time	Case	Activity
1:00	1	a
1:20	2	a
1:22	1	b
1:25	1	c
1:29	3	a
1:32	2	c
1:35	3	b
1:40	2	b
1:57	1	d
1:53	2	d
1:59	2	e
2:25	3	c
2:35	3	d
2:36	1	e
2:40	3	d
2:45	3	e

Case1 Case2 Case3

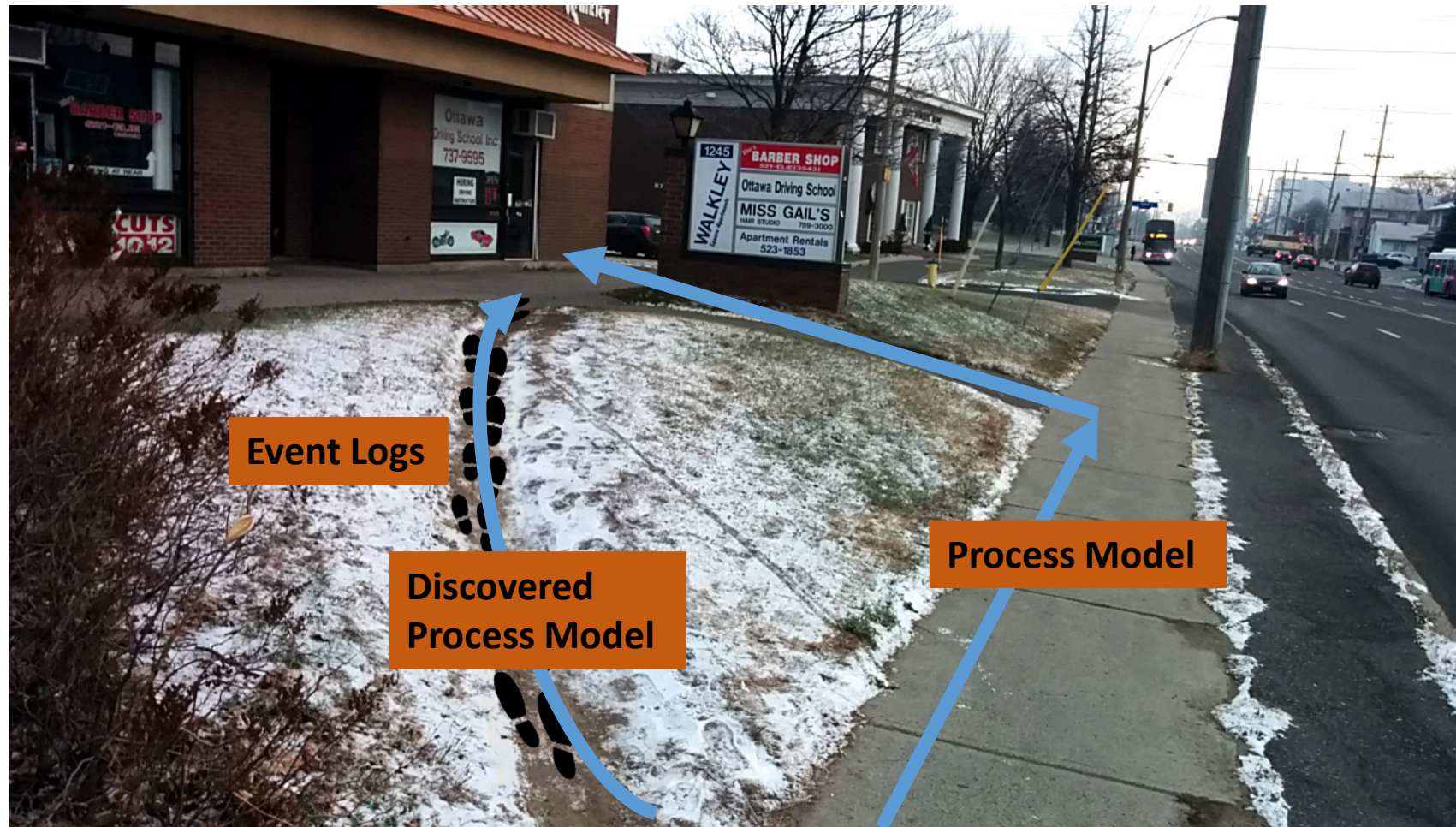


## Process Model

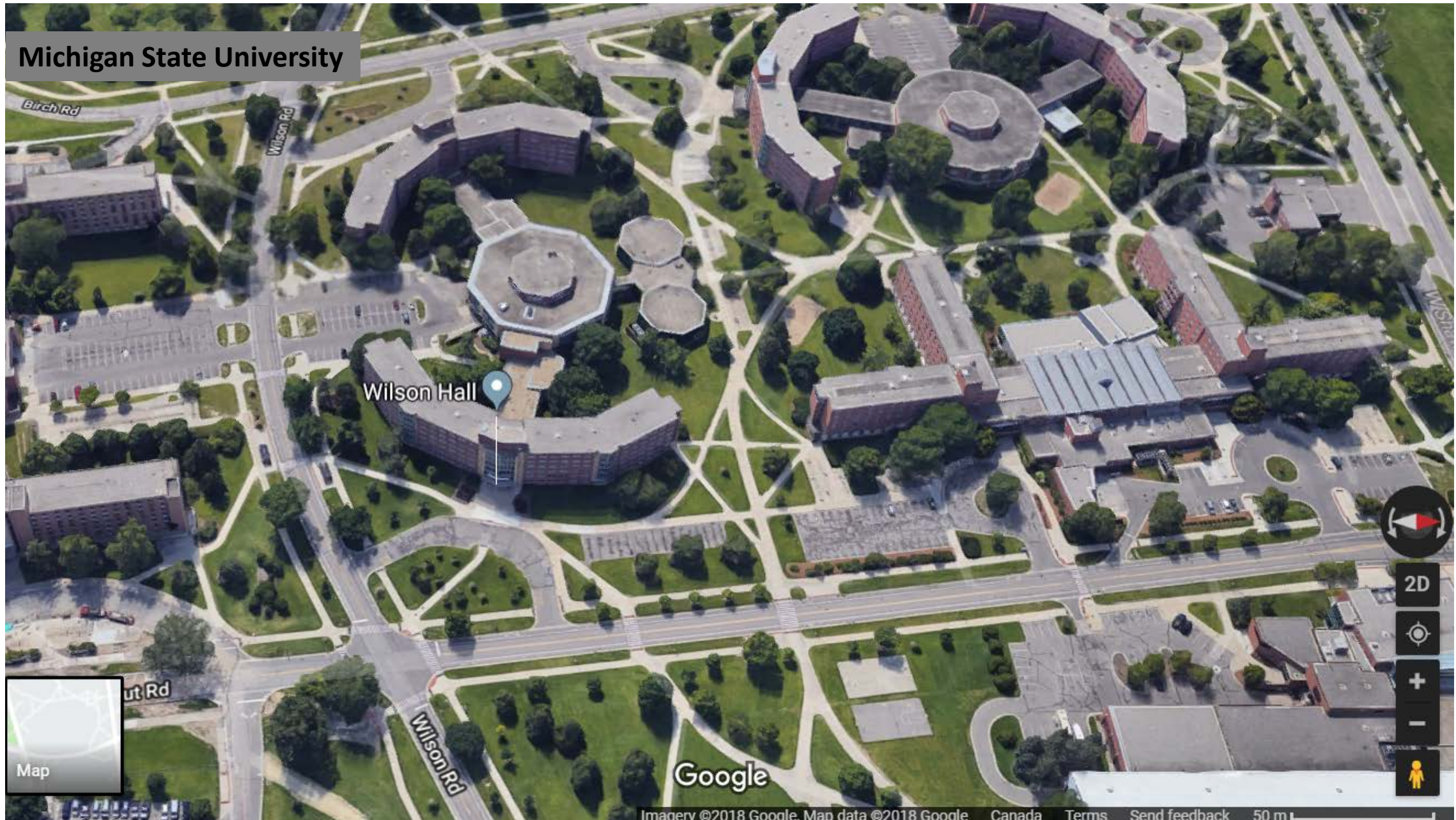


Process Discovery





## Michigan State University



**Process Mining:** activity-oriented approach, focuses on “how”, “what”, “where”, “who”, and especially “when” questions

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

not considered by process mining

● **Potential for synergy**



## Problems of two complementary domains:

### Process mining:

1. Unreliable rationality behind the discovered models
2. Unstructured “spaghetti-like” processes

### Requirements Engineering (RE):

1. Using huge crowd-based data logs generated from organization processes throughout the RE lifecycle



# Goal-oriented Process Enhancement and Discovery (GoPED)

## Current process mining:

The whole cases' log



→ Model

## Our approach:

Satisfied cases' log:



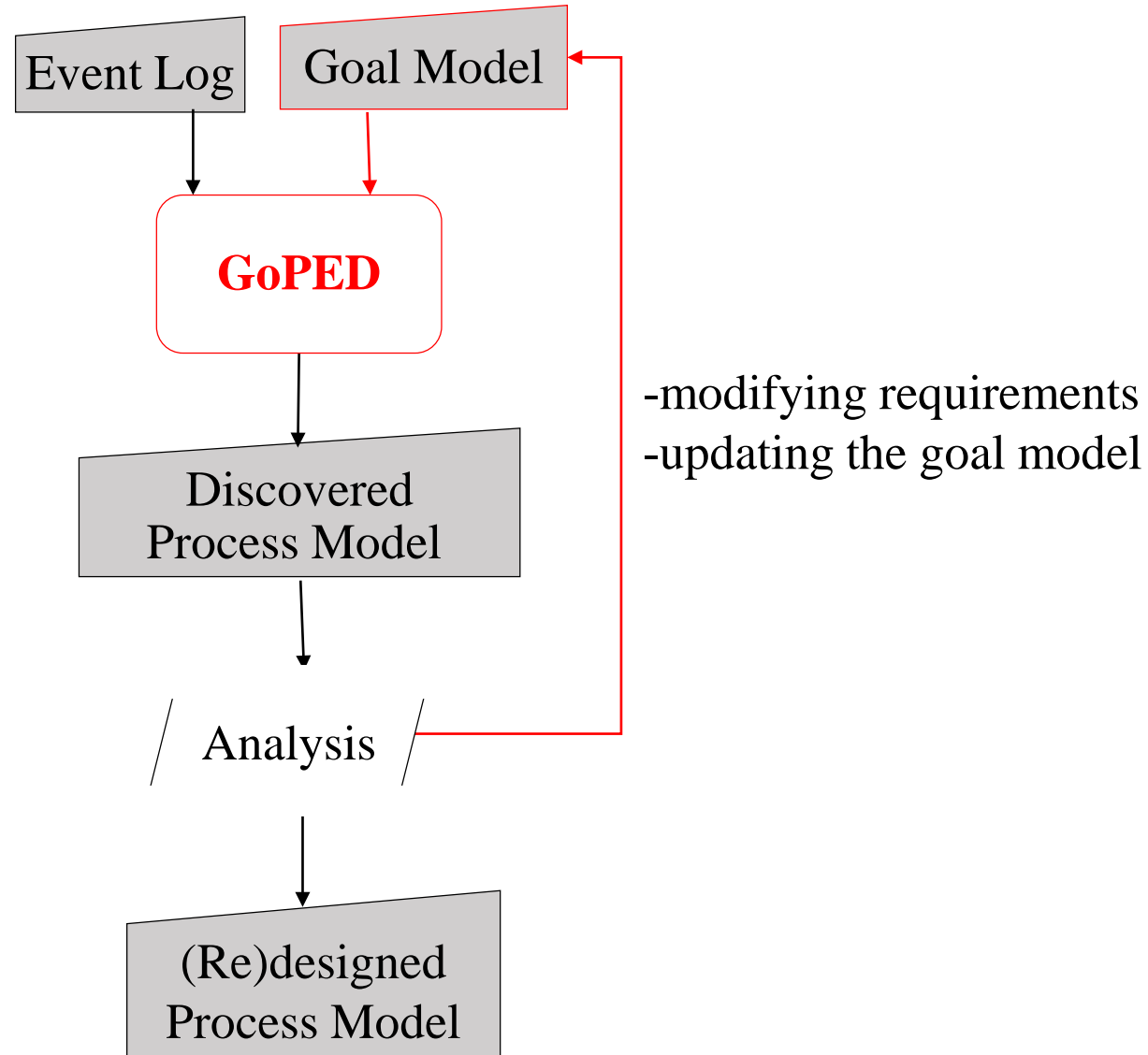
→ Good model  
(Goal-aligned model)

Dis-satisfied cases' log:



→ Bad model





# Event log enhanced with goal-related attributes (*EnhancedLog*)

Goal-related attributes

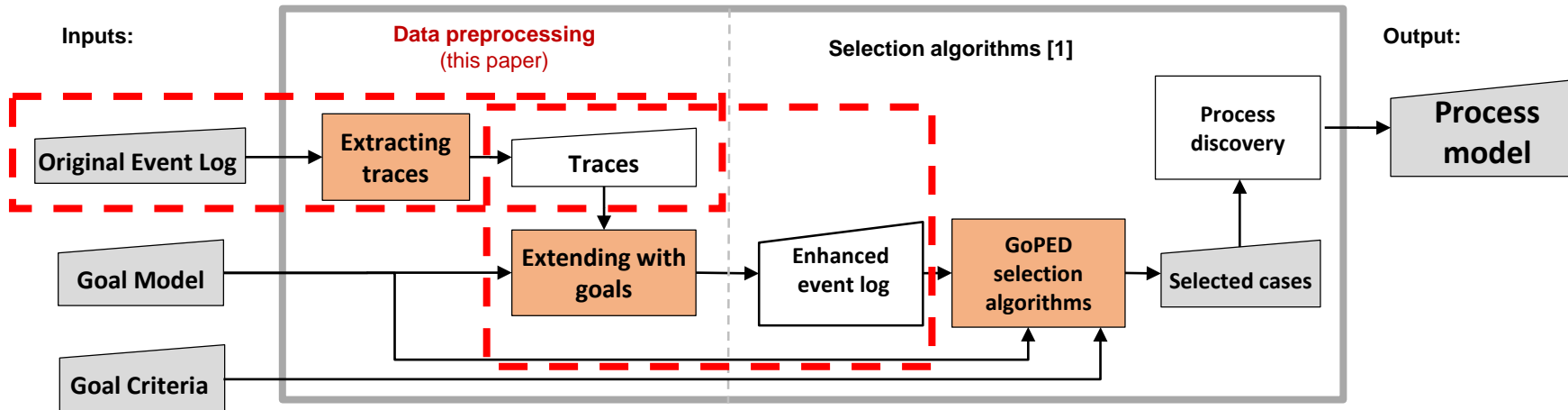
<i>Case</i>	<i>Trace</i>	<i>Goal 1</i>	<i>Goal 2</i>	...	<i>Goal n</i>	<i>Overall</i>
$c_1$	$t_1$	$s_{1,1}$	$s_{1,2}$	...	$s_{1,n}$	$s_{1.Ove}$
$c_2$	...	$s_{2,1}$	$s_{2,2}$	...	$s_{2,n}$	$s_{2.Ove}$
...	...	...	...	...	...	...
$c_m$	$t_m$	$s_{m,1}$	$s_{m,2}$	...	$s_{m,n}$	$s_{m.Ove}$
<i>Aggregated satisfaction: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

Aggregated satisfaction level  
(e.g., mean, median, ...)

Comprehensive  
satisfaction level

Overall satisfaction level  
(using the goal model)

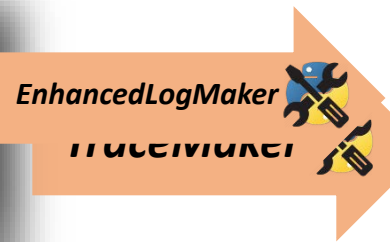
# Overview of GoPED steps and the position of data preprocessing



Case	Trace
$c_1$	$\langle a, b, c, g \rangle$
$c_2$	$\langle a, b, c, g \rangle$
$c_3$	$\langle a, b, c, d, e, c, g \rangle$

3	a: admission	12/03 09:45	Rose	\$150
3	b: regular test	12/03 11:32	Tina	\$100
3	c: check the result	15/03 11:39	Hannah	\$50
3	d: ask advanced test	15/03 11:41	Hannah	\$0
3	e: advanced test	20/03 09:35	Linda	\$400
3	c: check the result	22/03 14:12	Hannah	\$50
3	g: send the result	23/03 08:17	Jane	\$100



Case	Trace	G1	G2	G3	Overall
$c_1$	$\langle a, b, c, g \rangle$	100	100	100	100
$c_2$	$\langle a, b, c, g \rangle$	94	100	100	97
$c_3$	$\langle a, b, c, d, e, c, g \rangle$	61	59	100	59

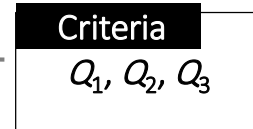
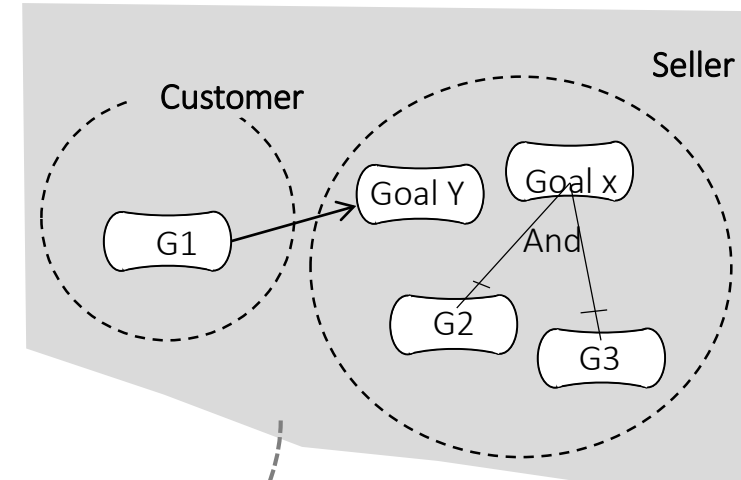
  

$c_3$	$\langle a, b, c, d, e, c, g \rangle$
-------	---------------------------------------

## Event Log:

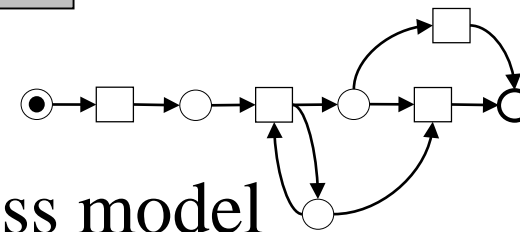
Case ID	Trace	Goal1: Customer's satisfaction	KPI of Process time	Goal2: Process time	KPI of Resource usage	Goal3: Resource usage	....
1	<a,b,b,c,d,e>	90	14 days	80	...	20	
2	<a,b,c,b,d>	20	44 days	10	...	80	.....
3	<a,b,c,e>	60	15 days	75	...	80	
4	<a,b,b,b,b,e>	10	38 days	20	...	25	
5	<a,b,b,c,c,e>	30	38 days	20	...	40	
...	...	...	...	...	...	...	
m	...	...	...	...	...	...	
Aggregated:		45		30		65	.....

## Goal model:



Process Discovery

Discovered process model



1. M. Ghasemi, "Towards Goal-oriented Process Mining", [Doctoral Symposium paper], in *Proceedings of the International Requirements Engineering Conference*, IEEE CS, 2018, pp. 484-489.

	<i>Case</i>	<i>Trace</i>	<i>Goal 1</i>	<i>Goal 2</i>	<i>...</i>	<i>Goal n</i>	<i>Overall</i>
<b>1</b>	$c_1$	$t_1$	$s_{1,1}$	$s_{1,2}$	$\dots$	$s_{1,n}$	$s_{1.Ove}$
	$c_2$	$\dots$	$s_{2,1}$	$s_{2,2}$	$\dots$	$s_{2,n}$	$s_{2.Ove}$
	$c_m$	$t_m$	$s_{m,1}$	$s_{m,2}$	$\dots$	$s_{m,n}$	$s_{m.Ove}$
	<i>Aggregated satisfaction:</i>		$s_{Agg.1}$	$s_{Agg.2}$	$\dots$	$s_{Agg.n}$	$s_{Comp}$

Diagram annotations: Red boxes highlight  $s_{1,2}$ ,  $s_{2,2}$ ,  $s_{m,2}$ ,  $s_{Agg.2}$ ,  $s_{Agg.n}$ , and  $s_{Comp}$ . A red box with '1' and arrows points to the first three rows. A red box with '2' and a curved arrow points from  $s_{Agg.2}$  to  $s_{Agg.n}$ . A red box with '3' and an arrow points to  $s_{Comp}$ .

## Three criteria for model discovery in GoPED:

1. To guarantee the satisfaction level for one or multiple goals for all cases
2. To guarantee the aggregated satisfaction level of one or multiple goals
3. To guarantee the comprehensive satisfaction level

**An example in healthcare:  
Screening and Diagnosis of Gestational Diabetes (DGD)**



## KPIs

## Event log of 10 patients

Case	Trace	Process Time (day)	Cost (\$)	Patient rating (1-10)	Accuracy of results
Patient_1	⟨a, b, c, g⟩	4	400	9	1
Patient_2	⟨a, b, c, g⟩	5	400	9	1
Patient_3	⟨a, b, c, g⟩	5	400	9	0
Patient_4	⟨a, b, c, d, e, c, g⟩	11	850	8	1
Patient_5	⟨a, b, c, d, e, c, g⟩	9	850	7	1
Patient_6	⟨a, b, c, d, e, c, g⟩	10	850	8	1
Patient_7	⟨a, b, c, f, b, c, g⟩	8	600	7	1
Patient_8	⟨a, b, c, f, b, c, d, e, c, g⟩	17	1100	6	1
Patient_9	⟨a, b, c, f, b, c, d, e, c, g⟩	16	1100	5	1
Patient_10	⟨a, b, c, d, b, c, d, e, c, d, e, c, g⟩	31	1150	4	1



## Goals

EnhancedLog					
Case	Trace	G <sub>1</sub> : To decrease process time	G <sub>2</sub> : To decrease cost	G <sub>3</sub> : To do a smooth process	G <sub>4</sub> : To screen accurately
Patient_1	⟨a, b, c, g⟩	100	100	88	100
Patient_2	⟨a, b, c, g⟩	94	100	88	100
Patient_3	⟨a, b, c, g⟩	94	100	88	0
Patient_4	⟨a, b, c, d, e, c, g⟩	61	59	75	100
Patient_5	⟨a, b, c, d, e, c, g⟩	72	59	63	100
Patient_6	⟨a, b, c, d, e, c, g⟩	67	59	75	100
Patient_7	⟨a, b, c, f, b, c, g⟩	78	82	63	100
Patient_8	⟨a, b, c, f, b, c, d, e, c, g⟩	41	20	50	100
Patient_9	⟨a, b, c, f, b, c, d, e, c, g⟩	43	20	40	100
Patient_10	⟨a, b, c, d, b, c, d, e, c, d, e, c, g⟩	9	10	30	100





**L:**

$\langle a, b, c, g \rangle^3,$

$\langle a, b, c, d, e, c, g \rangle^3,$

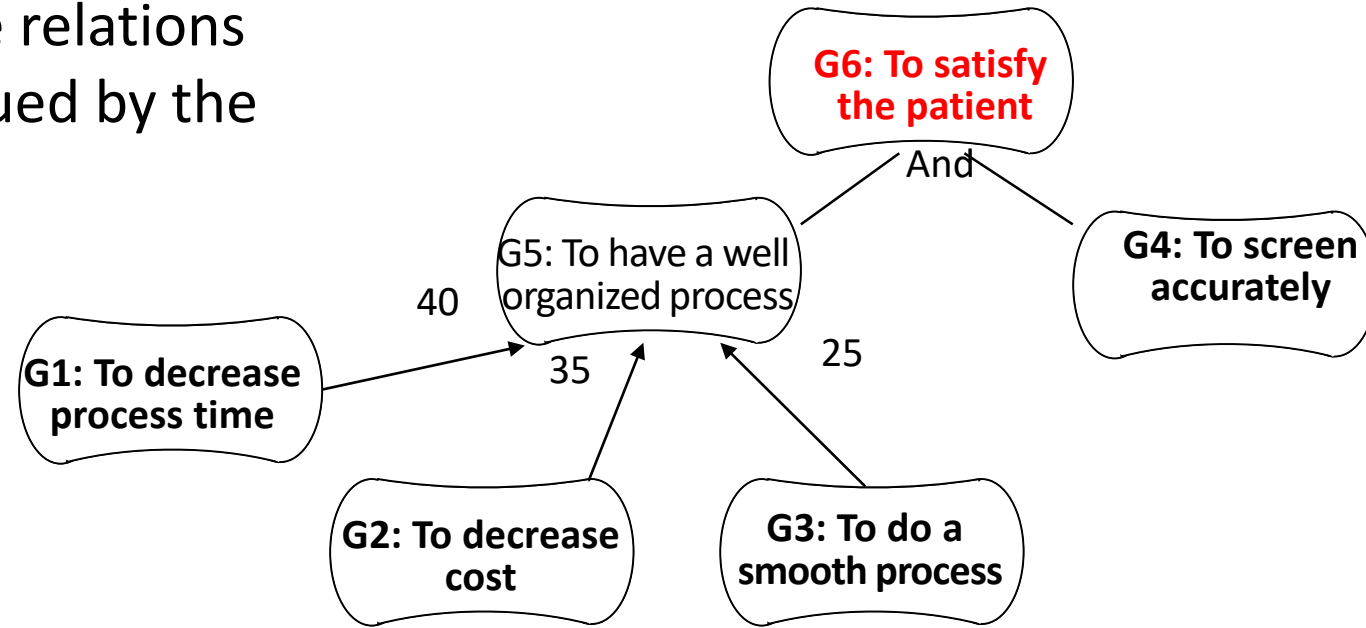
$\langle a, b, c, f, b, c, g \rangle^1,$

$\langle a, b, c, f, b, c, d, e, c, g \rangle^2,$

$\langle a, b, c, d, b, c, d, e, c, d, e, c, g \rangle^1$



Goal model showing the relations between the goals pursued by the DGD process:



$$\text{Overall} = \text{Sat}(G_6) = \text{Minimum} \left( \text{SL}(G_4), 0.4 \text{SL}(G_1) + 0.35 \times \text{SL}(G_2) + 0.25 \times \text{SL}(G_3) \right)$$

EnhancedLog						
Case	Trace	G <sub>1</sub> : To decrease process time	G <sub>2</sub> : To decrease cost	G <sub>3</sub> : To do a smooth process	G <sub>4</sub> : To screen accurately	Overall:
Patient_1	⟨a, b, c, g⟩	100	100	88	100	97
Patient_2	⟨a, b, c, g⟩	94	100	88	100	95
Patient_3	⟨a, b, c, g⟩	94	100	88	0	0
Patient_4	⟨a, b, c, d, e, c, g⟩	61	59	75	100	62
Patient_5	⟨a, b, c, d, e, c, g⟩	72	59	63	100	65
Patient_6	⟨a, b, c, d, e, c, g⟩	67	59	75	100	66
Patient_7	⟨a, b, c, f, b, c, g⟩	78	82	63	100	75
Patient_8	⟨a, b, c, f, b, c, d, e, c, g⟩	41	20	50	100	36
Patient_9	⟨a, b, c, f, b, c, d, e, c, g⟩	43	20	40	100	34
Patient_10	⟨a, b, c, d, b, c, d, e, c, d, e, c, g⟩	9	10	30	100	15
<b>Aggregated:</b>		66	61	66	90	64



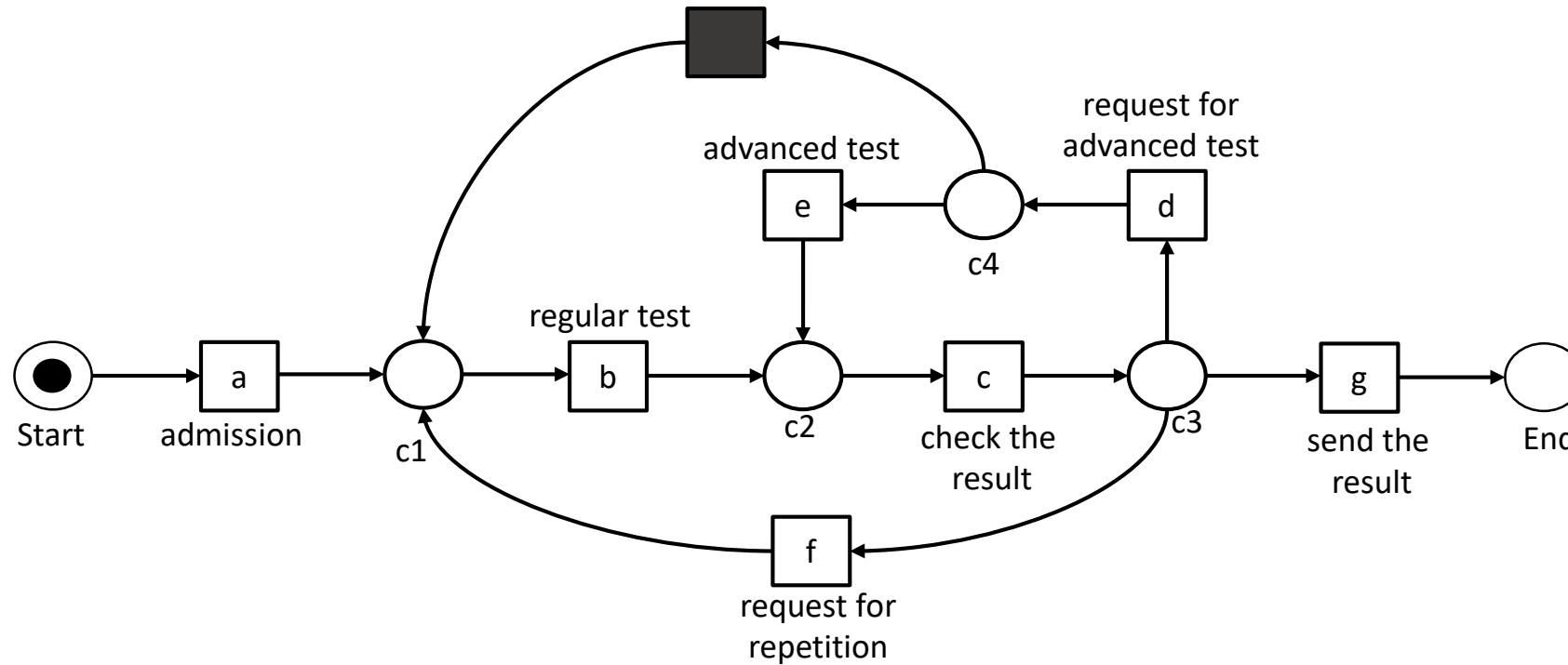
EnhancedLog							Overall:
Case	Trace	1	G <sub>1</sub> : To decrease process time	G <sub>2</sub> : To decrease cost	G <sub>3</sub> : To do a smooth process	G <sub>4</sub> : To screen accurately	
→ Patient_1	⟨a, b, c, g⟩		100	100	88	100	97
→ Patient_2	⟨a, b, c, g⟩		94	100	88	100	95
→ Patient_3	⟨a, b, c, g⟩		94	100	88	0	0
→ Patient_4	⟨a, b, c, d, e, c, g⟩		61	59	75	100	62
→ Patient_5	⟨a, b, c, d, e, c, g⟩		72	59	63	100	65
→ Patient_6	⟨a, b, c, d, e, c, g⟩		67	59	75	100	66
→ Patient_7	⟨a, b, c, f, b, c, g⟩		78	82	63	100	75
→ Patient_8	⟨a, b, c, f, b, c, d, e, c, g⟩		41	20	50	100	36
→ Patient_9	⟨a, b, c, f, b, c, d, e, c, g⟩		43	20	40	100	34
→ Patient_10	⟨a, b, c, d, b, c, d, e, c, d, e, c, g⟩		9	10	30	100	15
aggregated			66	61	66	90	64

Diagram annotations: A red box labeled '1' highlights the G<sub>1</sub> column. A red box labeled '2' is positioned below the aggregated row, with yellow arrows pointing to the G<sub>2</sub> (61) and G<sub>3</sub> (66) cells. A red box labeled '3' is positioned to the right of the aggregated row, with a grey arrow pointing to the Overall: (64) cell.

## Three criteria for model discovery:

1. To guarantee the satisfaction level for one or multiple goals for all cases
2. To guarantee the aggregated satisfaction level of one or multiple goals
3. To guarantee the comprehensive satisfaction level

## Process model discovered using *all* cases



1

**Case perspective:** generate a model that guarantees (with a confidence of 90%) that the satisfaction level for all patients in terms of goal “To decrease process time” will be at least 75 and that in terms of goal “To do a smooth process” will be at least 80.

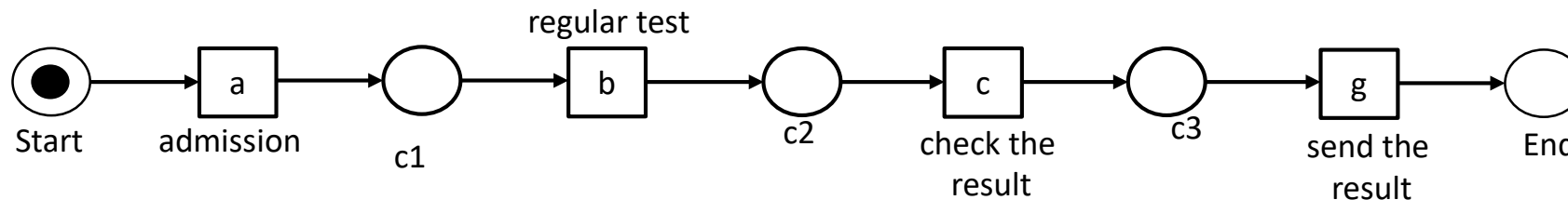
$Subset = \{Patient\_1, Patient\_2, Patient\_3\}$

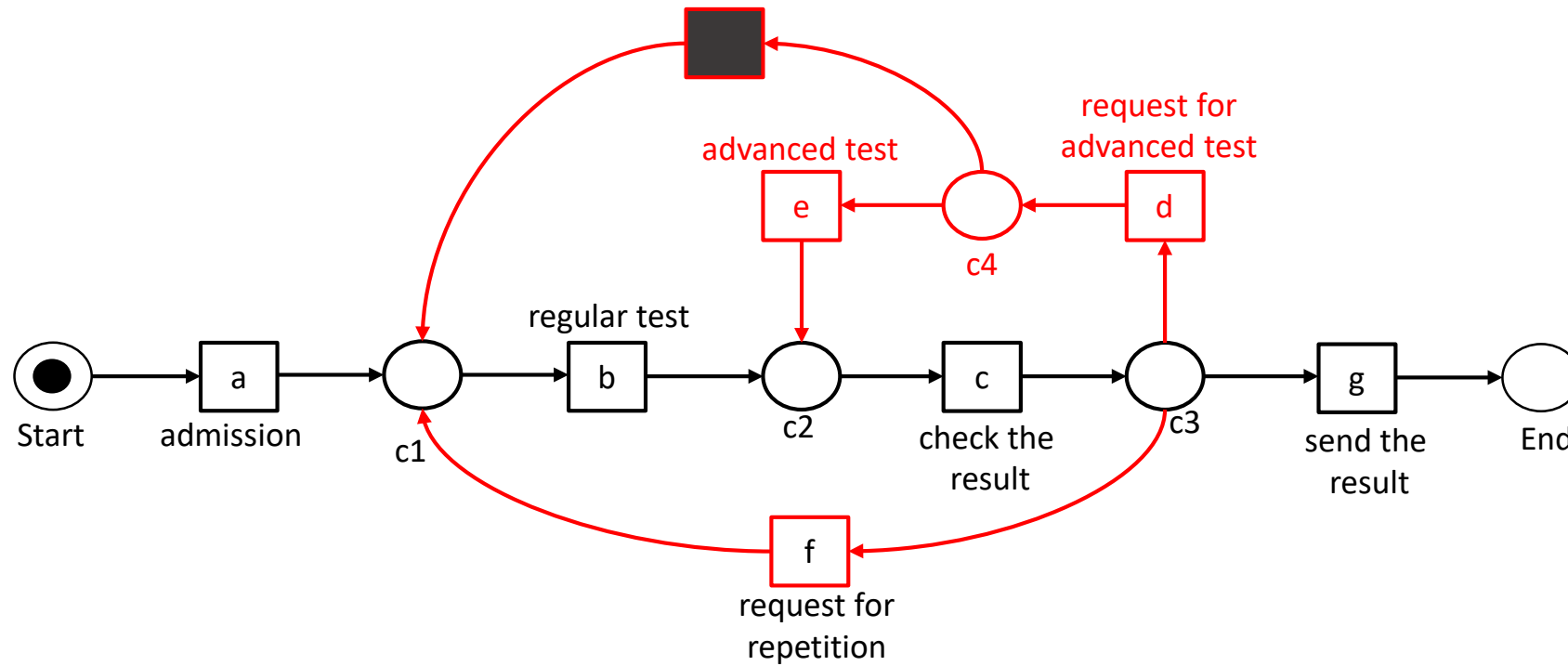
$Log = \{\langle a, b, c, g \rangle^3\}$

Event log and satisfaction level of goals [0-100]					
Case	Trace	G <sub>1</sub> : To decrease process time	G <sub>2</sub> : To decrease cost	G <sub>3</sub> : To do a smooth process	G <sub>4</sub> : To screen accurately
Patient_1	$\langle a, b, c, g \rangle$	100	100	88	100
Patient_2	$\langle a, b, c, g \rangle$	94	100	88	100
Patient_3	$\langle a, b, c, g \rangle$	94	100	88	0
Patient_4	$\langle a, b, c, d, e, c, g \rangle$	61	59	75	100
Patient_5	$\langle a, b, c, d, e, c, g \rangle$	72	59	63	100
Patient_6	$\langle a, b, c, d, e, c, g \rangle$	67	59	75	100
Patient_7	$\langle a, b, c, f, b, c, g \rangle$	78	82	63	100
Patient_8	$\langle a, b, c, f, b, c, d, e, c, g \rangle$	41	20	50	100
Patient_9	$\langle a, b, c, f, b, c, d, e, c, g \rangle$	43	20	40	100
Patient_10	$\langle a, b, c, d, b, c, d, e, c, d, e, c, g \rangle$	9	10	30	100



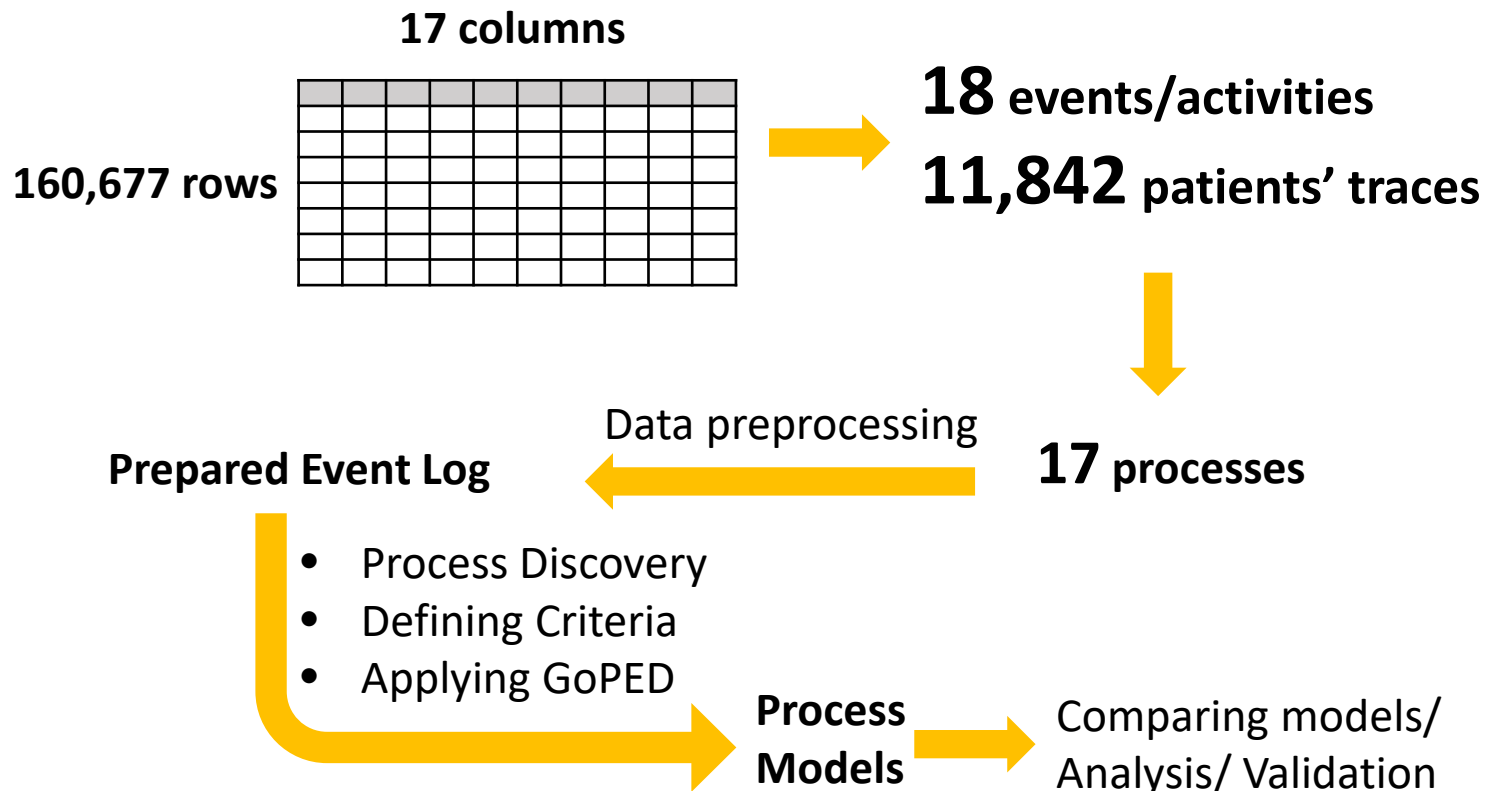
**Case perspective:** generate a model that guarantees (with a confidence of **90%**) that the satisfaction level for all patients in terms of goal “**To decrease process time**” will be at least **75** and that in terms of goal “**To do a smooth process**” will be at least **80**.







# Data log from the Children Hospital of Eastern Ontario (CHEO) :

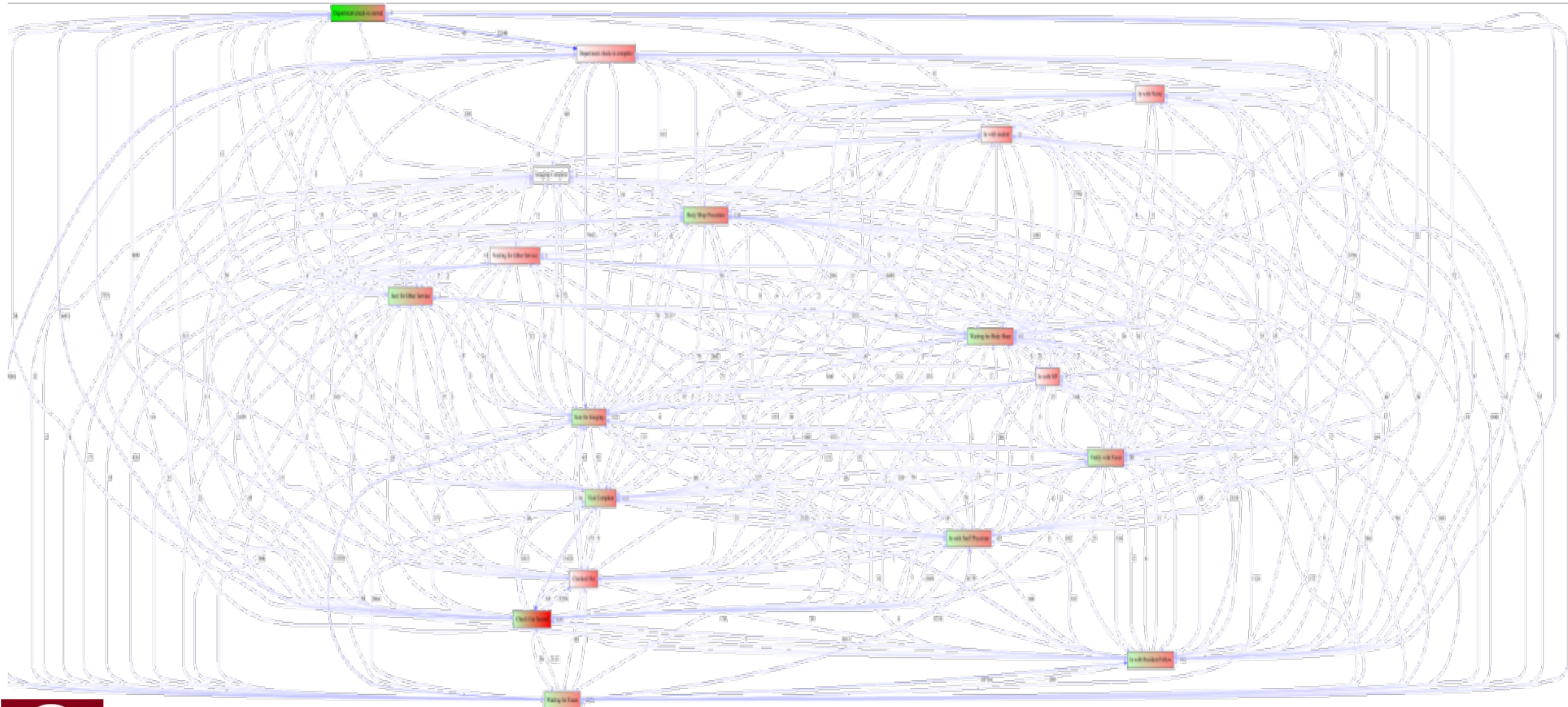


Processes' name and the number of their events


	Process Name	Number of events (absolute)	Number of events (relative)
1	NEW PLASTER CLINIC	73991	46.05%
2	RETURN PLASTER CLINIC	22688	14.12%
3	NEW ORTHO	19459	12.11%
4	RETURN ORTHO	18250	11.36%
5	RETURN SCOLIOSIS	7642	4.76%
6	NEW SCOLIOSIS	5247	3.27%
7	NEW SPORTS MEDICINE	4934	3.07%
8	NEW DDH	3074	1.91%
9	NEW BODY SHOP	1545	0.96%
10	RETURN BODY SHOP	1465	0.91%
11	RETURN SPORTS MEDICINE	1132	0.70%
12	NEW CLUB FOOT	817	0.51%
13	RETURN DDH	188	0.12%
14	RETURN CLUB FOOT	175	0.11%
15	RETURN TELEHEALTH	59	0.04%
16	NEW TELEHEALTH	6	0.00%
17	TELEHEALTH	3	0.00%
	<b>Grand Total</b>	<b>160675</b>	<b>100.00%</b>



# Spaghetti-like CHEO process model

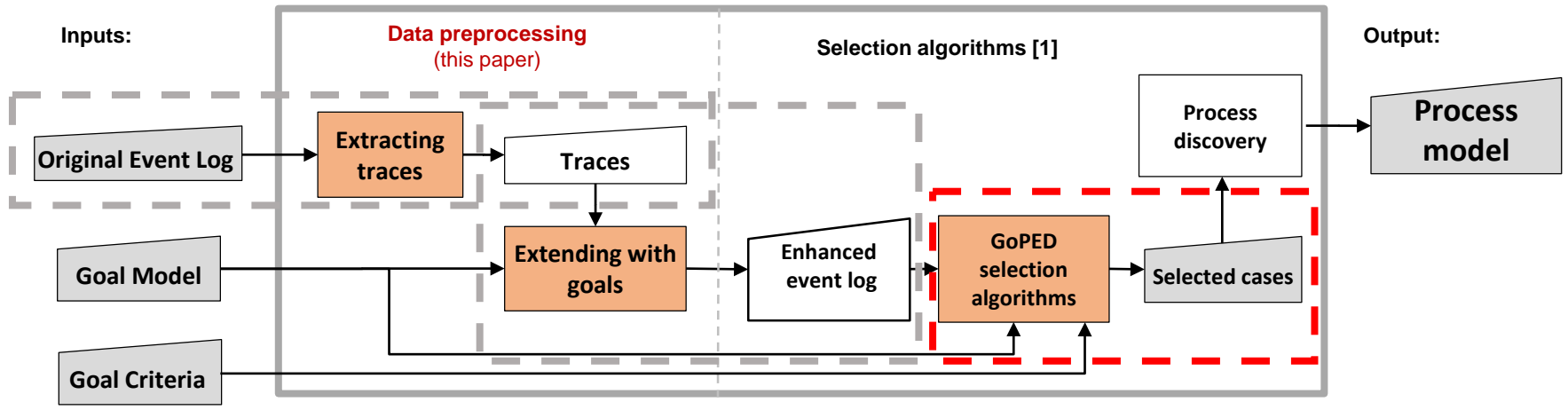


# Performance of tools:

<b>TraceMaker</b>		Small sample 15 events 6 cases		CHEO Hospital(CSV=7.4 M) 160,677 events 11,842 cases		Synthetic large sample(CSV=40.6M) 1,025,000 events 50,000 cases	
	List()	Panadas	List()	Panadas	List()	Panadas	
	<b>Time</b>	0.003 s	0.030 s	0.585 s	38.889 s	6.337 s	308.543 s

<b>EnhancedLogMaker</b>		CHEO Hospital(CSV=7.4 M) 11,842 cases 3 goals		Synthetic large sample(CSV=40.6M) 50,000 cases 3 goals	
	List()	List()	List()		
	<b>Time (s)</b>	0.48 s	1.687 s		

# Future work related to GoPED



**Algorithm 1** Selecting a subset of an event log to infer a process model that guarantees a minimum satisfaction level for one or multiple goals in each selected case

**Input:** *EnhancedLog*: An enhanced structured event log ▷ explained in Definition 2  
*Q<sub>goal</sub>*: A set of criteria, ▷ explained in Definition 3  
*conf*: a confidence level ▷ explained in Definition 3  
**Output:** *SelectedCases* ▷ a subset of cases selected according to the criteria and the all-or-none rule

```

1  sort_by_trace(EnhancedLog) ▷ sort the cases based on their traces
2  trace(cases0) ← {} ▷ is an empty trace, which cannot happen in reality
3  trace(cases[NumberOfCases+1]) ← {} ▷ also flag the end of the log
4  SelectedCases ← {}
5  index ← 1
6  while index ≤ NumberOfCases ▷ NumberOfCases is m in Table 1
7  SameTraceCases ← {} ▷ a set of cases whose traces are the same
8  NumberOfSatisfiedCasesOfTrace ← 0 ▷ counts the satisfied cases of a trace
9  do
10 SameTraceCases ← SameTraceCases ∪ {casesindex}
11 if caseindex meets all criteria of Qgoal then
12   NumberOfSatisfiedCasesOfTrace ++
13 end if
14 index ++
15 while trace(casesindex) = trace(casesindex-1)
16 if NumberOfSatisfiedCasesOfTrace / size(SameTraceCases) ≥ conf then
17   SelectedCases ← SelectedCases ∪ SameTraceCases
18 end if
19 end while
20 return SelectedCases ▷ the resulting subset of cases
    
```

**Algorithm 2** Selecting a subset of an event log to infer a process model that guarantees the overall satisfaction level(s) of one or multiple goals

**Input:** *EnhancedLog*: An enhanced structured event log ▷ explained in Definition 2  
*Q<sub>goal</sub>*: A set of criteria (some goals and thresholds for their satisfaction level) ▷ Definition 3  
*g*: A function computing the satisfaction of the whole goal model ▷ Definition 3  
**Output:** *SelectedCases* ▷ a subset of all cases selected regarding the criteria and the all-or-none rule, *SelectedCases* ⊆ *C*, Definition 4

```

1  SelectedCases ← {}
2  Solve the binary optimization below: (xi is a flag for either selecting case, or not)
   Max Z = ∑i=1m xi ▷ this is to find the largest subset
   s.t.
   ∀ r, t 1 ≤ r < t ≤ m : if trace(cr) = trace(ct) xr = xt ▷ all-or-none rule
   ∀ j where Gj ∈ G : ∑i=1m xi sij ≥ Sij ▷ |Qgoal| constraints
   xi ∈ {0, 1} ▷ if xi = 1, case i should be selected
3  end of binary optimization
4  for i = 1 to NumberOfCases do ▷ NumberOfCases is m in Table 1
5   if xi = 1 then
6    SelectedCases ← SelectedCases ∪ {ci}
7   end
8  end
9  return SelectedCases ▷ the resulting subset of cases that meets the criteria
    
```

**Algorithm 3** Selecting the largest subset of an event log to infer a process model that guarantees a comprehensive satisfaction level

**Input:** *EnhancedLog*: An enhanced structured event log ▷ as explained by Definition 2  
*S<sub>Comp</sub>*:  $\bar{s}l_{Comp}$  ▷ a minimum threshold for comprehensive satisfaction level  
*F*: a tuple of functions  $(f_1, f_2, \dots, f_n) | s_{ij} = f_j(s_{1j}, s_{2j}, \dots, s_{mj}), j = 1, \dots, n$   
*g*: a function derived from the goal model,  $s_{i,goal} = g(s_{i1}, s_{i2}, \dots, s_{in})$   
**Output:** *SelectedCases* ▷ a subset of all cases selected regarding the criteria and the all-or-none rule, *SelectedCases* ⊆ *C*

```

1  SelectedCases ← {}
2  if SComp = f1}(s1,1, s2,1, ..., sm,1}) then
3   use Algorithm 2 and exit.
4  else
5   Solve the binary optimization below:
   Max Z = ∑i=1m xi ▷ this is to find the largest subset.
   s.t.
   ∀ r, t 1 ≤ r < t ≤ m : if trace(cr) = trace(ct) xr = xt ▷ all-or-none rule
   g(∑i=1m xi sij / ∑i=1m xi, ∑i=1m xi sij / ∑i=1m xi, ..., ∑i=1m xi sin / ∑i=1m xi) ≥ slComp
   xi ∈ {0, 1}
6  end
7  for i = 1 to NumberOfCases do ▷ NumberOfCases is m in Table 1
8   if xi = 1 then
9    SelectedCases ← SelectedCases ∪ {ci}
10  end
11 end
12 return SelectedCases ▷ the resulting subset of cases that meets the criterion.
    
```



Thank you!

