A genetic algorithm for process discovery guided by completeness, precision and simplicity

Borja Vázquez-Barreiros, Manuel Mucientes, Manuel Lama

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA - SPAIN

Focus on complete, precise and simple models… Why?

Some domains, like **e-learning**, require:

- **High** levels of **completeness**
  - The teacher needs to know **all the behavior** undertaken by the students to make a correct assesment
- **Precise** models
  - The teacher **only needs** to know what the students did
- **Simple** models
  - The model must be **readable**

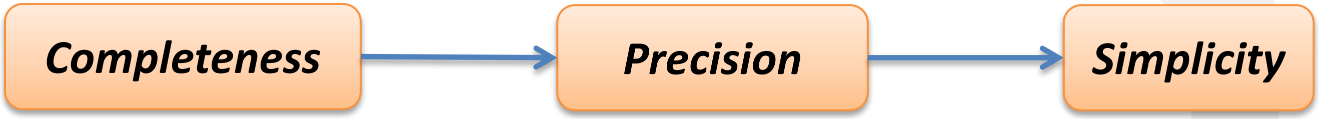
Process Mining

Motivation

FOCUS ON COMPLETE, PRECISE AND SIMPLE MODELS… WHY?

Some domains, like **e-learning**, require:

- **High** levels of **completeness**
  - The teacher needs to know **all the behavior** undertaken by the students to make a correct assesment
- **Precise** models
  - The teacher **only needs** to know what the students did
- **Simple** models
  - The model must be **readable**



Completeness → ***Precision*** → ***Simplicity***

CiTIUS

ProDiGen: **Process Discovery** through a **Genetic** algorithm

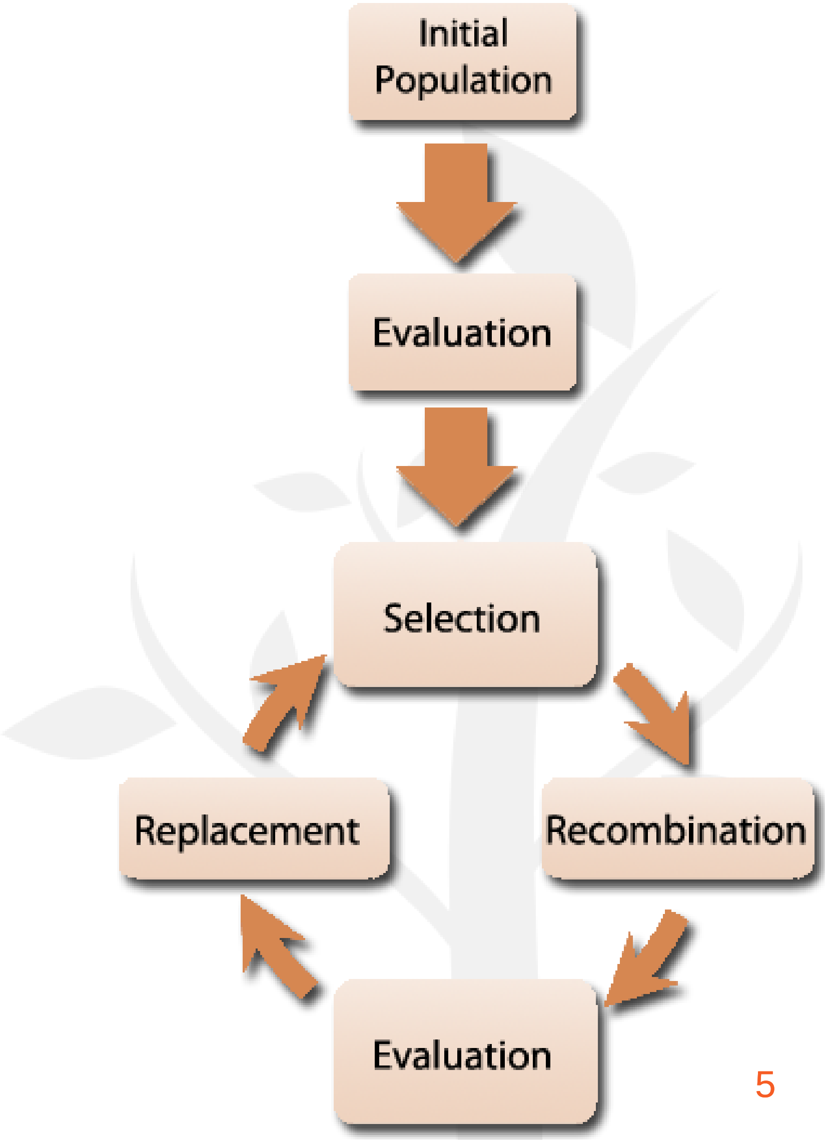
- Genetic algorithm that **searches** complete, precise and simple models
  - **Hierarchical fitness** based on completeness, precision and simplicity
  - New criteria for **precision** and **simplicity**
  - **Heuristics** in the **crossover** and **mutation**
  - Heuristics Miner's solution is **incorporated to the initial population**
  - **Binary tournament** selection
  - **Steady-state replacement** with **reinitializations**

ProDiGen

What is a genetic algorithm?

- Optimization algorithm
- Components
  - **Individuals**: potential solutions
  - **Population**: set of individuals
- The **population evolves** until obtain an optimal solution

Evolutionary cycle

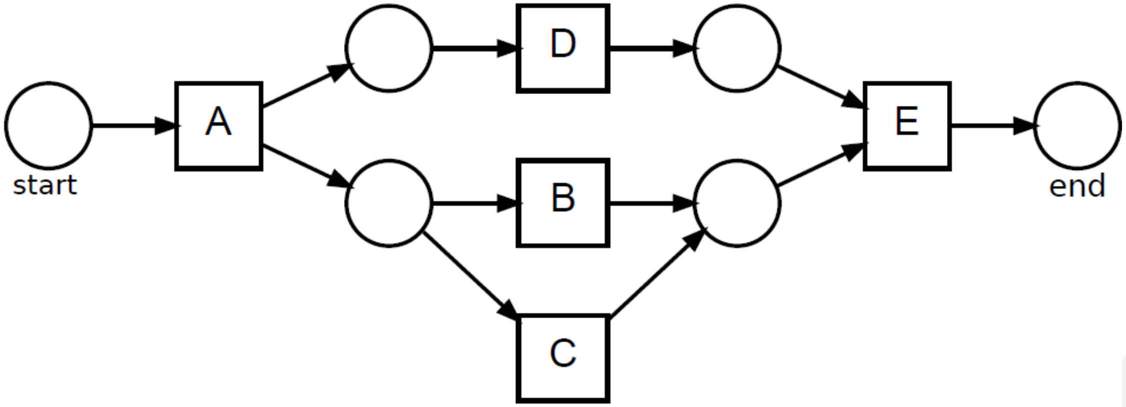


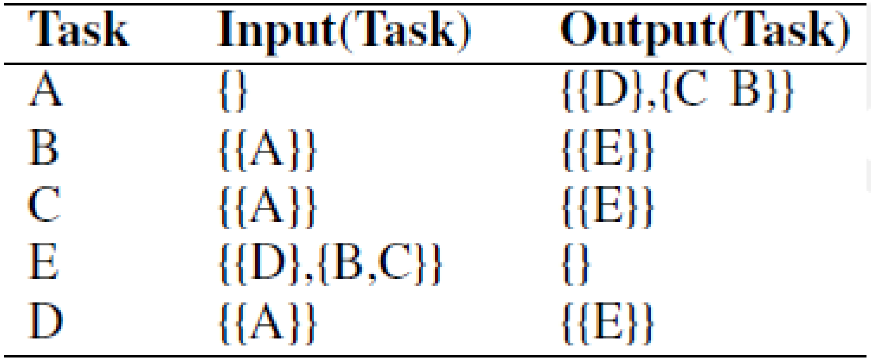
CiTIUS

ProDiGen

Genetic algorithm – **Internal representation**

- **Causal Matrix** [1]: maps any Petri net in terms of causal dependencies





Task	Input(Task)	Output(Task)
A	{}	{{D},{C B}}
B	{{A}}	{{E}}
C	{{A}}	{{E}}
E	{{D},{B,C}}	{}
D	{{A}}	{{E}}

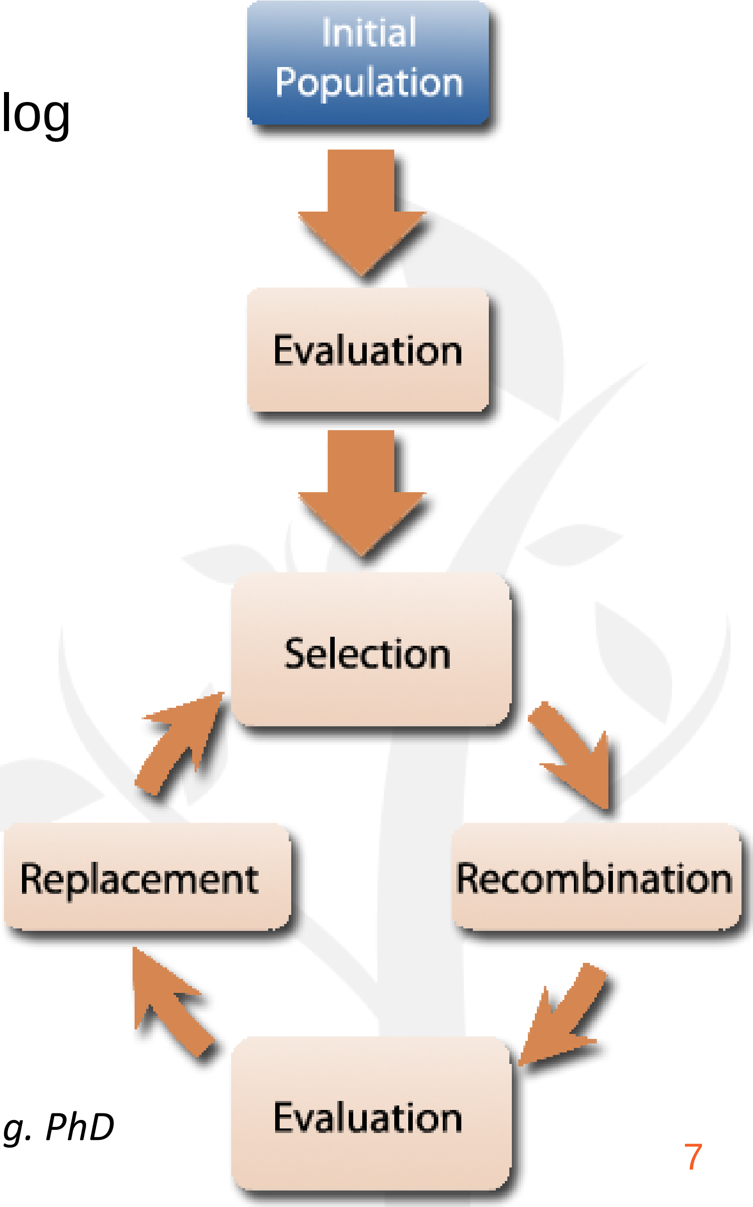
[1] *de Medeiros, A.K.A : Genetic Process Mining. PhD thesis, Technische Universiteit Eindhoven*

ProDiGen

Initial Population – **Building the initial population**

- Based on the **local information** [1] of the log

$$\begin{cases} \frac{\#aba+\#bab}{\#aba+\#bab+1} & \text{if } a \neq b \text{ and } \#aba > 0; \\ \frac{\#ab-\#ba}{\#ab+\#ba+1} & \text{if } a \neq b \text{ and } \#aba = 0; \\ \frac{\#ab}{\#ab+1} & \text{if } a = b. \end{cases}$$



[1] *de Medeiros, A.K.A : Genetic Process Mining. PhD thesis, Technische Universiteit Eindhoven*

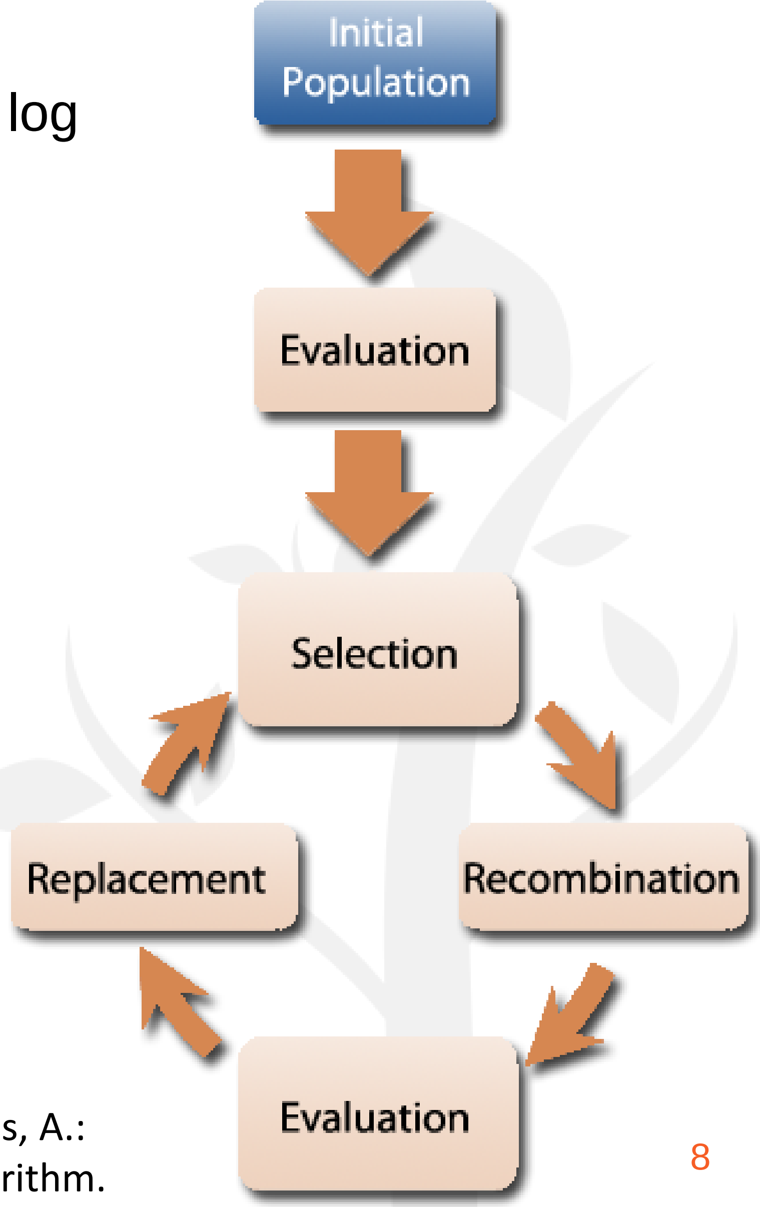
ProDiGen

Initial Population – **Building the initial population**

- Based on the **local information** [1] of the log

$$\begin{cases} \frac{\#aba+\#bab}{\#aba+\#bab+1} & \text{if } a \neq b \text{ and } \#aba > 0; \\ \frac{\#ab-\#ba}{\#ab+\#ba+1} & \text{if } a \neq b \text{ and } \#aba = 0; \\ \frac{\#ab}{\#ab+1} & \text{if } a = b. \end{cases}$$

- The solution of the Heuristics Miner [2] **is also added to the initial population**
  - It **does not affect the final solution**, only in **the convergence speed**



Initial Population

Evaluation

Selection

Recombination

Evaluation

Replacement

[2] Weijters, A., van der Aalst, W., de Medeiros, A.: Process mining with the heuristics miner-algorithm.

ProDiGen

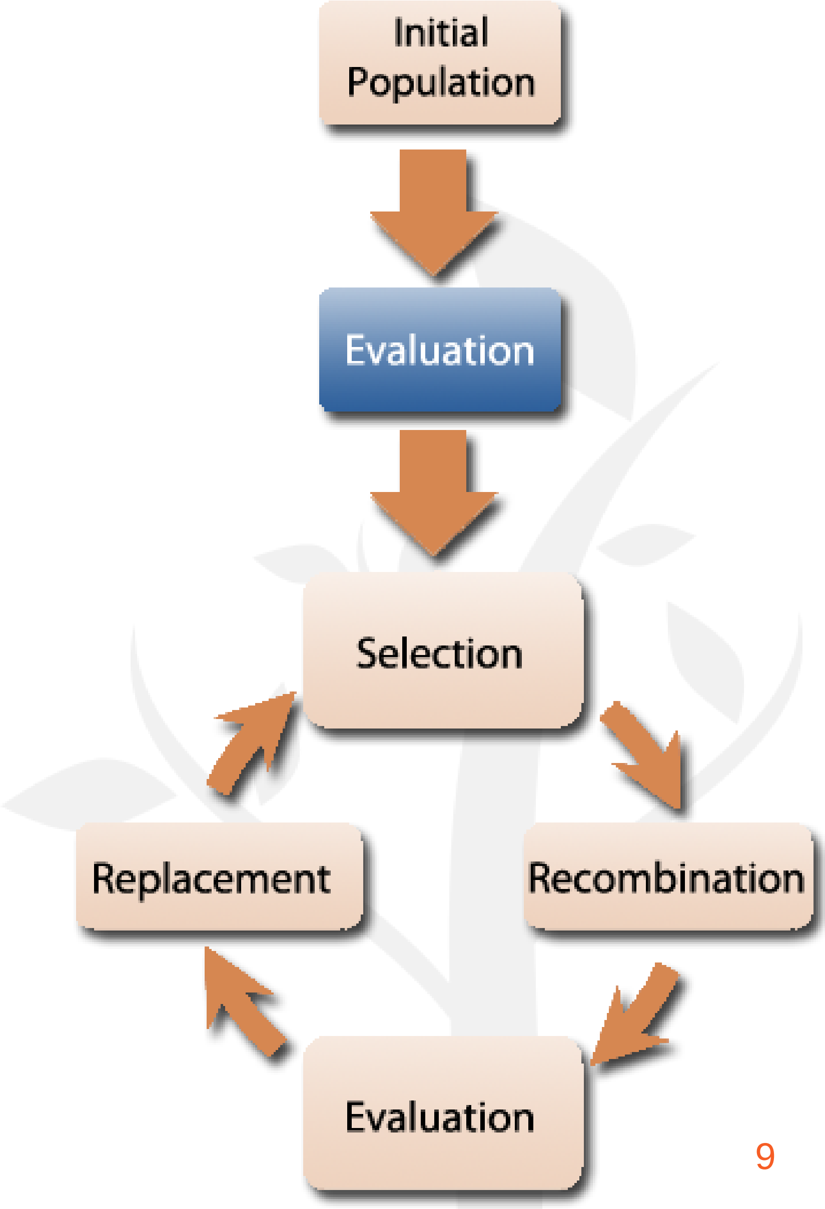
Evolutionary cycle – **Evaluation**

- Each individual is evaluated with **three** objectives:

Completeness

Precision

Simplicity



CiTIUS

ProDiGen

Evolutionary cycle – *Evaluation:* **Completeness**

- **Completeness** [1]**:** the retrieved model can reproduce **all the behavior** of the log

- Punish individuals with:
  - Tasks with **incorrect input arcs** (missing tokens)
  - Tasks with **incorrect output arcs** (extra tokens)

$$C_f\,(L,\ CM) = \frac{allParsedActivities(L,CM) - punishment}{numActivitiesLog(L)}$$

$$punishment = \frac{allMissingTokens(L,CM)}{numTracesLog(L) - numTracesMissingTokens(L,CM) + 1} + \frac{allExtraTokensLeftBehind(L,CM)}{numTracesLog(L) - numTracesExtraTokensLeftBehind(L,CM) + 1}$$

CiTIUS

[1] *de Medeiros, A.K.A : Genetic Process Mining. PhD thesis, Technische Universiteit Eindhoven*

ProDiGen

Evolutionary cycle – *Evaluation:* **Precision**

- **Precision:** the retrieved model **avoids additional behavior** , i.e, behavior not represented in the log

- Punish those individuals that **enable too many activities** during the parsing of the log
  - For each enabled activity → one possible path of execution

$$P_f(L,\ CM) = \frac{1}{allEnabledActivities(L,\ CM)}$$

- Each individual's precision **evolves regardless the rest of the population**

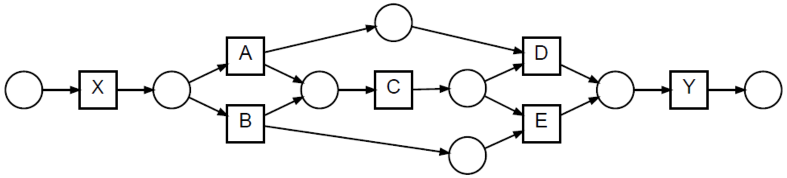
ProDiGen

Evolutionary cycle – *Evaluation:* **Simplicity**

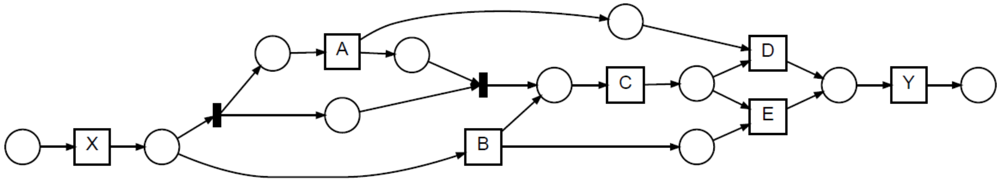
- **Simplicity:** discover learning paths with the **minimal structure**

Trace user 1 = (X,A,C,D,Y) **Trace user 2 = (X,B,C,E,Y)**

Model a



Model b



$C_f(a) = 1 ; P_f(a) = 1/12$

$C_f(b) = 1 ; P_f(b) = 1/12$

- Both models have the **same completeness and precision,** but **different simplicity**

CiTIUS

ProDiGen

Evolutionary cycle – *Evaluation:* **Simplicity**

- Counts the **number of relations** of the causal matrix

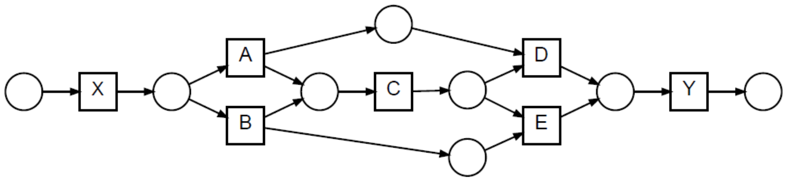
$$S_f\left(CM\right) = \frac{1}{\sum_{t \in CM}\left(\sum_{\Phi \in I(t)} |\Phi| + \sum_{\Psi \in O(t)} |\Psi|\right)}$$

ProDiGen

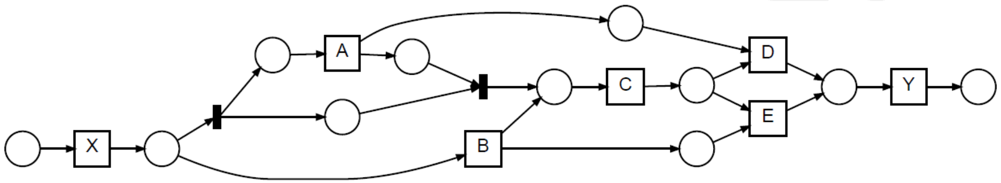
Evolutionary cycle – *Evaluation:* **Simplicity**

- Counts the **number of relations** of the causal matrix

$$S_f\left(CM\right) = \frac{1}{\sum_{t\in CM}\left(\sum_{\Phi\in I(t)}|\Phi| + \sum_{\Psi\in O(t)}|\Psi|\right)}$$



Task	I(Task)	O(Task)
X	{}	{{A,B}}
A	{{X}}	{{C},{D}}
B	{{X}}	{{C},{E}}
C	{{A,B}}	{{D,E}}
D	{{A},{C}}	{{Y}}
E	{{B},{C}}	{{Y}}
Y	{{D,E}}	{}



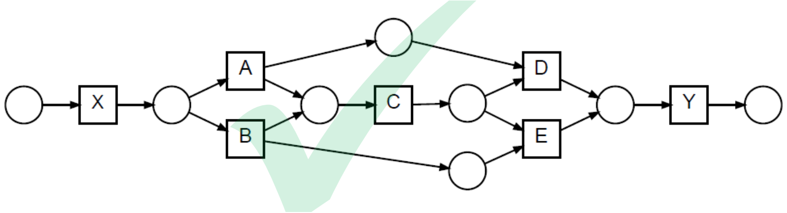
Task	I(Task)	O(Task)
X	{}	{{A,B},{B,C}}
A	{{X}}	{{C},{D}}
B	{{X}}	{{C},{E}}
C	{{A,B},{B,X}}	{{D,E}}
D	{{A},{C}}	{{Y}}
E	{{B},{C}}	{{Y}}
Y	{{D,E}}	{}

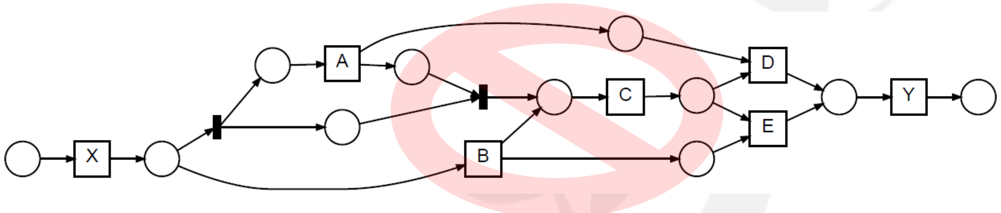
ProDiGen

Evolutionary cycle – *Evaluation:* **Simplicity**

- Counts the **number of relations** of the causal matrix

$$S_f\left(CM\right) = \frac{1}{\sum_{t \in CM}\left(\sum_{\Phi \in I(t)} |\Phi| + \sum_{\Psi \in O(t)} |\Psi|\right)}$$





Task	I(Task)	O(Task)
X	{}	{{A,B}}
A	{{X}}	{{C},{D}}
B	{{X}}	{{C},{E}}
C	{{A,B}}	{{D,E}}
D	{{A},{C}}	{{Y}}
E	{{B},{C}}	{{Y}}
Y	{{D,E}}	{}

Task	I(Task)	O(Task)
X	{}	{{A,B},{B,C}}
A	{{X}}	{{C},{D}}
B	{{X}}	{{C},{E}}
C	{{A,B},{B,X}}	{{D,E}}
D	{{A},{C}}	{{Y}}
E	{{B},{C}}	{{Y}}
Y	{{D,E}}	{}

$S_f(a) = 1/20$

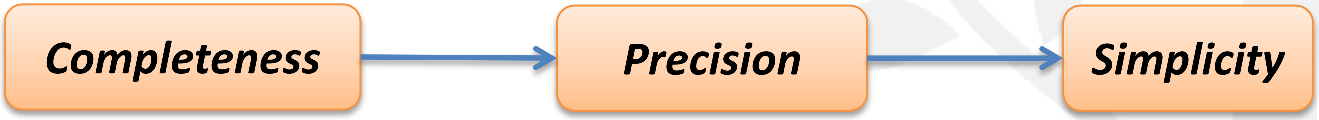
$S_f(b) = 1/24$

ProDiGen

Evolutionary cycle – **Evaluation**

Hierarchical fitness function

$$
\begin{aligned}
F(a) > F(b) \iff & \{C_f(a) > C_f(b)\} \lor \{C_f(a) = C_f(b) \land P_f(a) > P_f(b)\} \\
& \lor \{C_f(a) = C_f(b) \land P_f(a) = P_f(b) \land S_f(a) > S_f(b)\}
\end{aligned}
$$



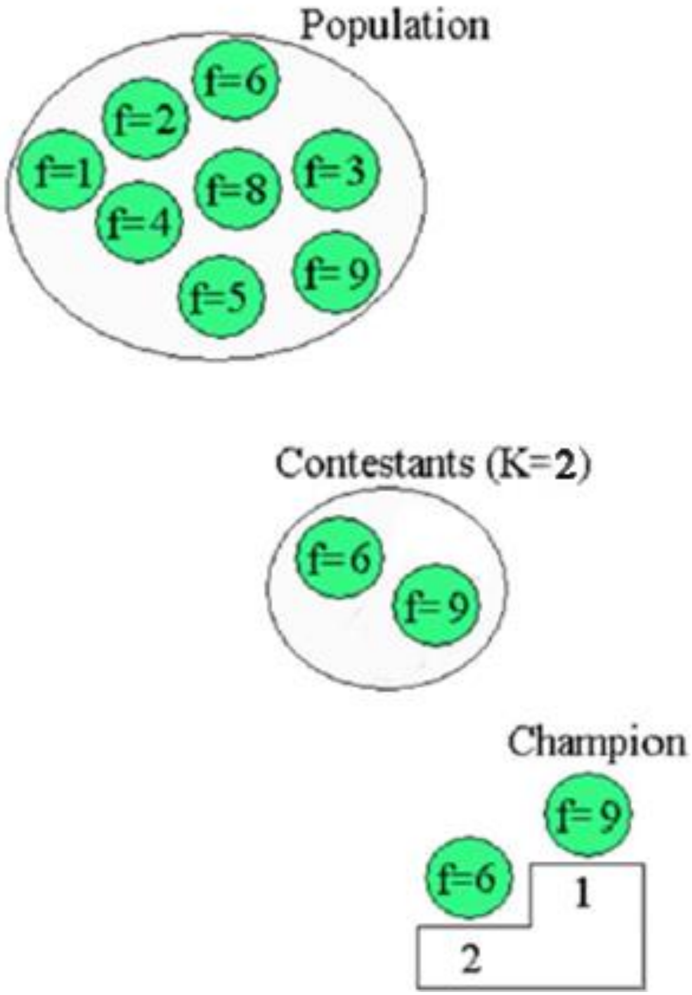
Completeness → ***Precision*** → ***Simplicity***

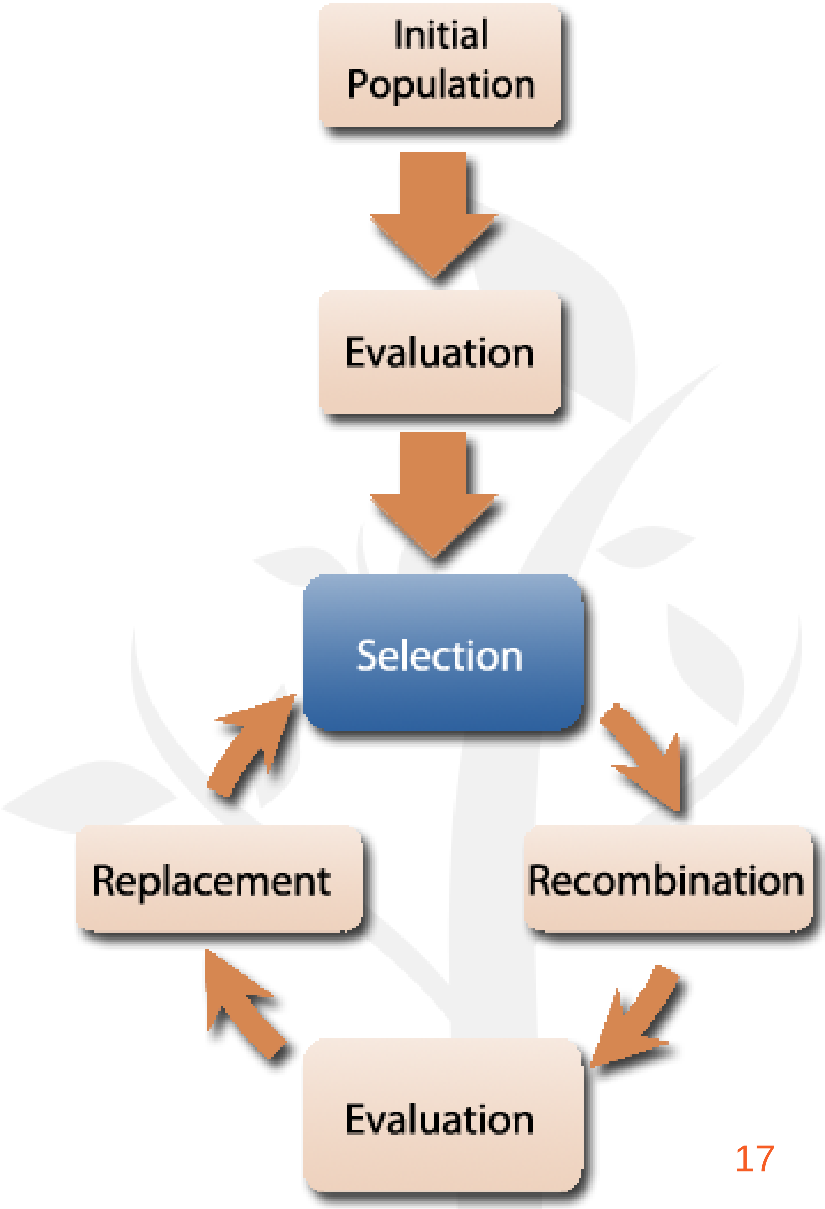
CiTIUS

ProDiGen

Evolutionary cycle – **Selection**

- Binary Tournament





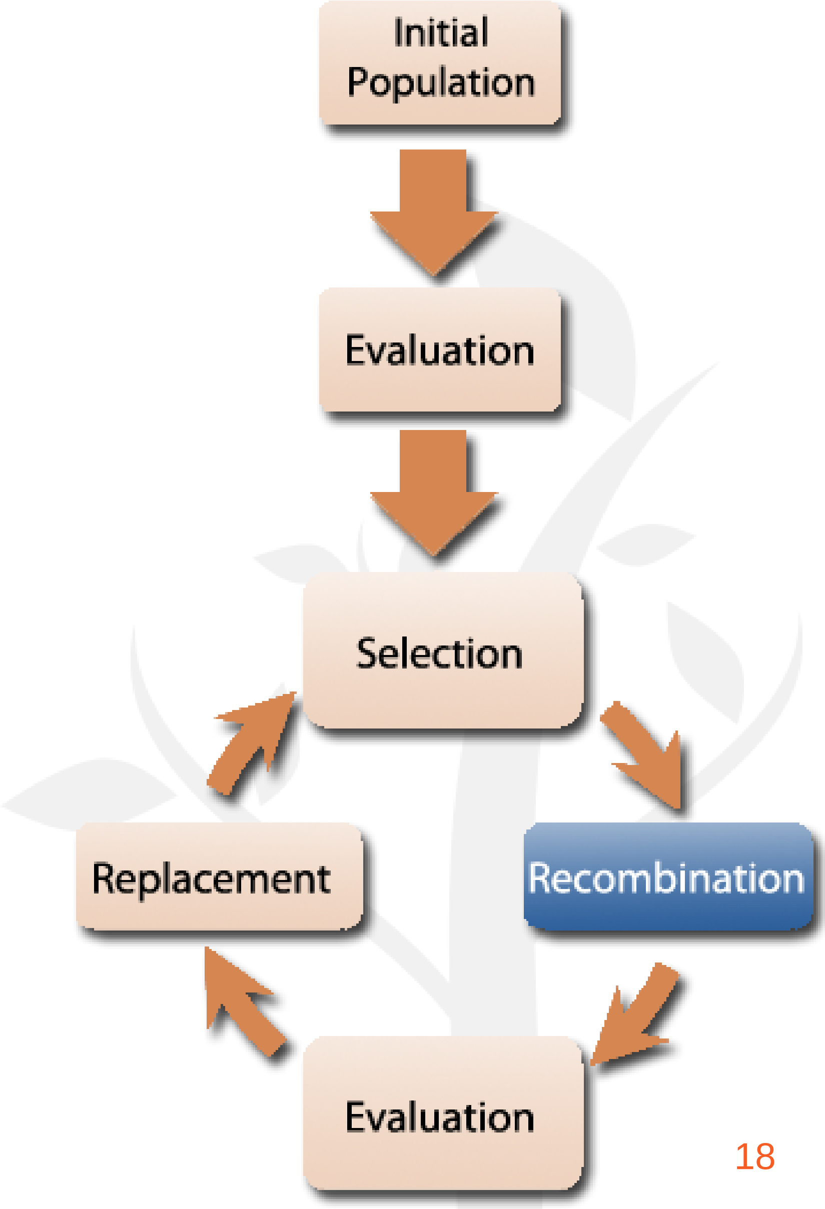
ProDiGen

Evolutionary cycle – **Recombination**

- Generate new individuals

- **Crossover**
  - Combines the characteristics of two parents into two offspring

- **Mutation**
  - Adds or removes characteristics from an individual

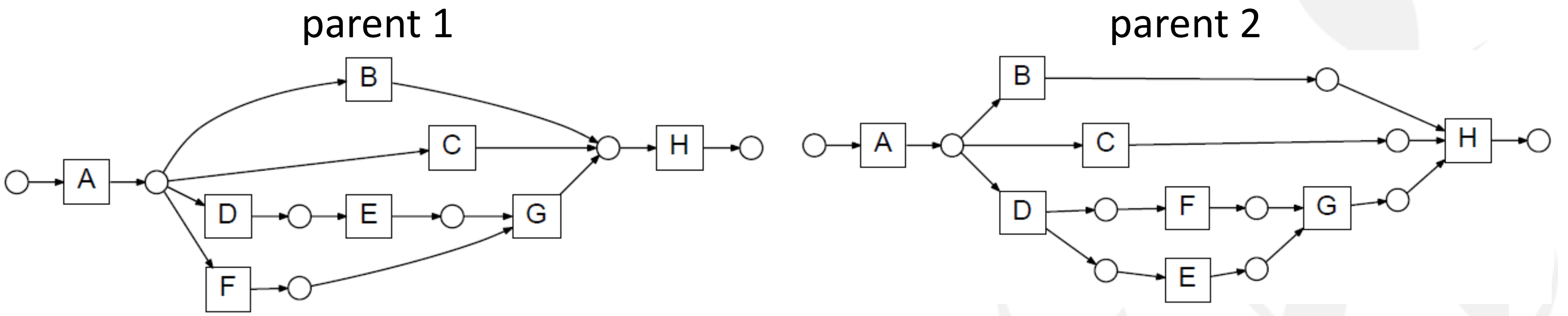


CiTIUS

ProDiGen

Evolutionary cycle – *Recombination*: **Crossover**

- **The crossover operator picks one task** of the parent 1 and exchange the input and output dependencies with the same task of the parent 2



- The size of the causal matrix **increases** with the number of activities in the log.
  - ▷ Picking the crossover point **at random** produces a poor performance of the crossover

ProDiGen

Evolutionary cycle – *Recombination*: **Crossover**

Guided by a *Probability Density Function* generated from the errors

- While parsing the log, each individual **records the incorrectly parsed tasks**
  - The **correctly parsed tasks** have a null chance for being crossed
  - The **incorrectly parsed tasks** have an uniform probability for being crossed
- The crossover point is selected from the incorrectly parsed tasks of **the individual with the higher completeness**
- The crossover is performed as defined in [1]

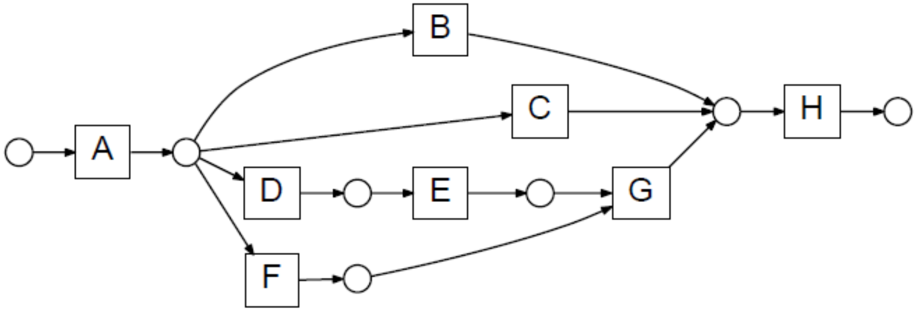
[1] *de Medeiros, A.K.A : Genetic Process Mining. PhD thesis, Technische Universiteit Eindhoven*

ProDiGen

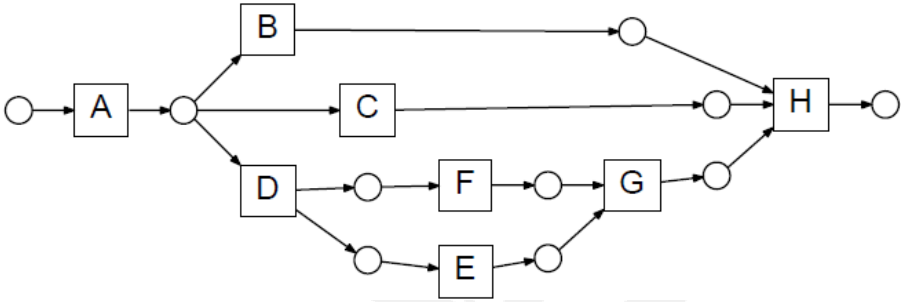
Evolutionary cycle – *Recombination*: **Crossover**

Log = { ***ABH*, *ACH*, *ADEFGH*, *ADFEGH*** }

parent 1



parent 2



C_f=0.85

incorrectlyParsedTasks = {F}

>

C_f=0.33

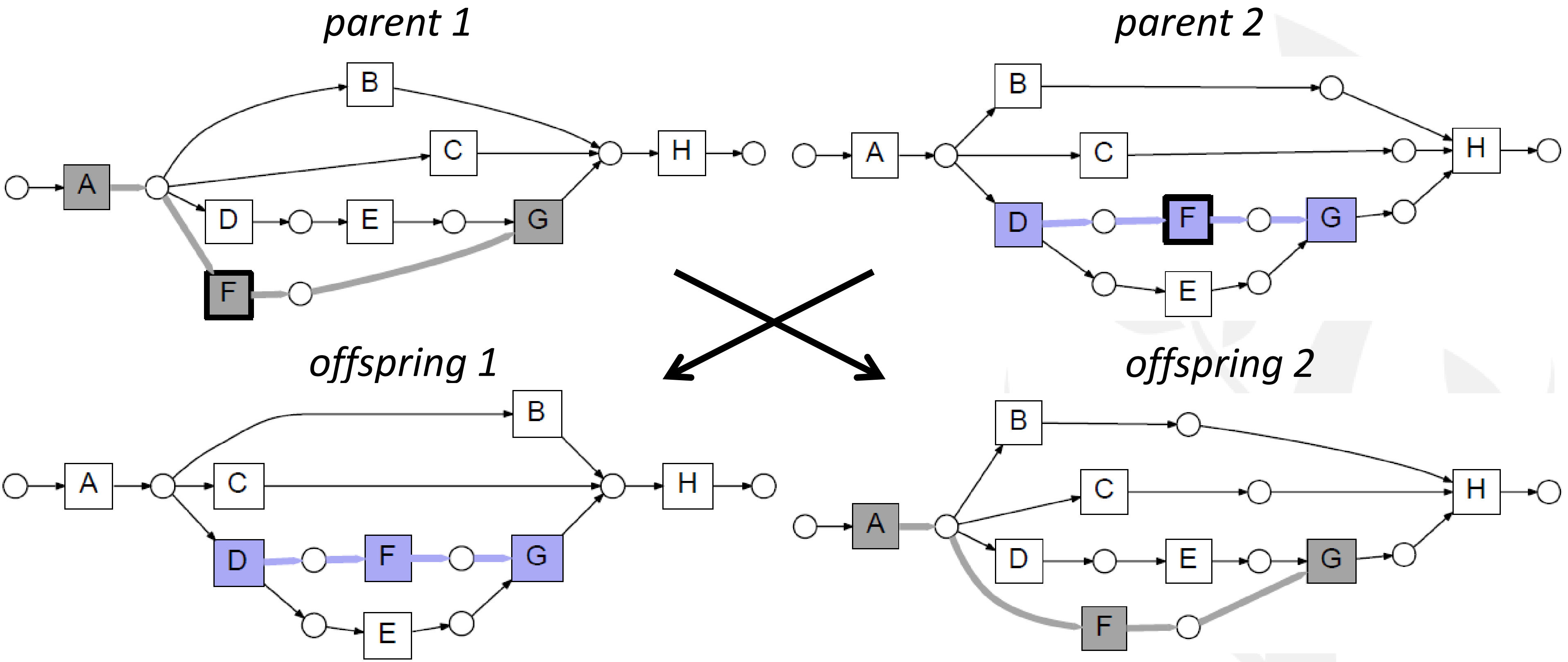
incorrectlyParsedTasks = {H}

- The crossover point is selected from *incorrectlyParsedTasks* of parent1**:**
  - ▷ **Task F**

ProDiGen

Evolutionary cycle – *Recombination*: **Crossover**

Log = { ***ABH*****,** ***ACH*****,** ***ADEFGH*****,** ***ADFEGH*** }



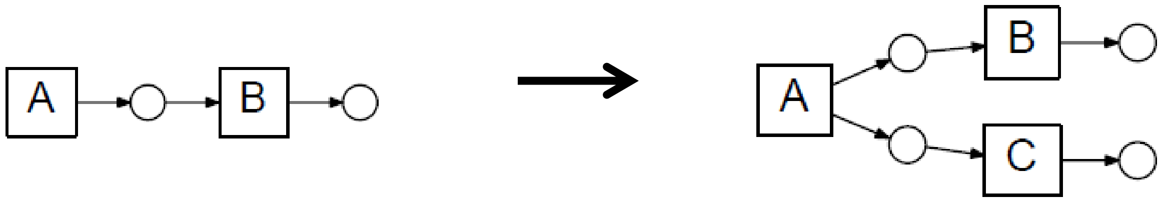
C_f= 1.0

C_f = 0.11

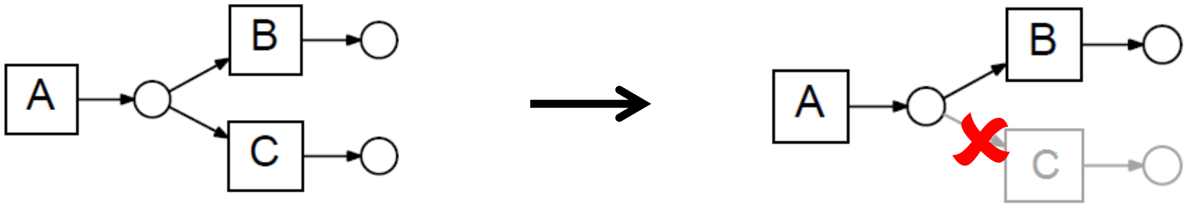
ProDiGen

Evolutionary cycle – *Recombination*: **Mutation**

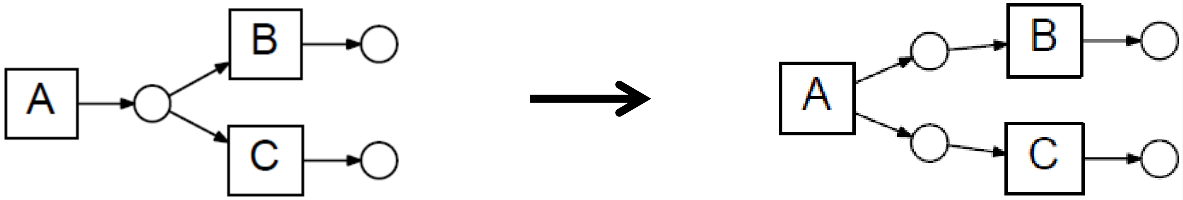
- The mutation operator can:
  - **Add a new task** to the input and/or output function of task



  - **Remove a task** from the input and/or output function of task



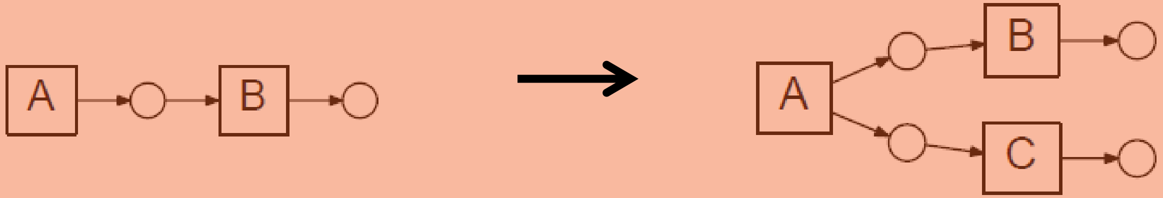
  - **Redistribute the elements** of the input and/or output function of task



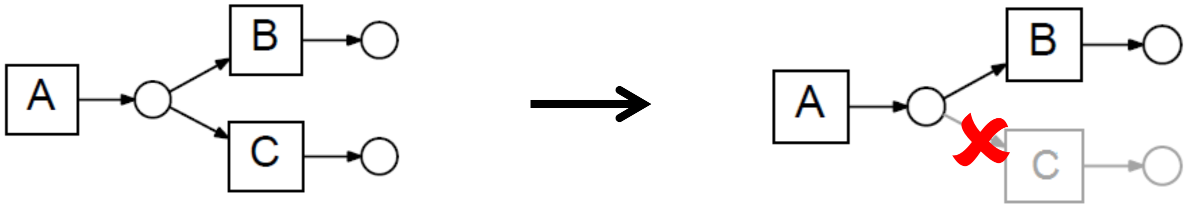
ProDiGen

Evolutionary cycle – *Recombination*: **Mutation**

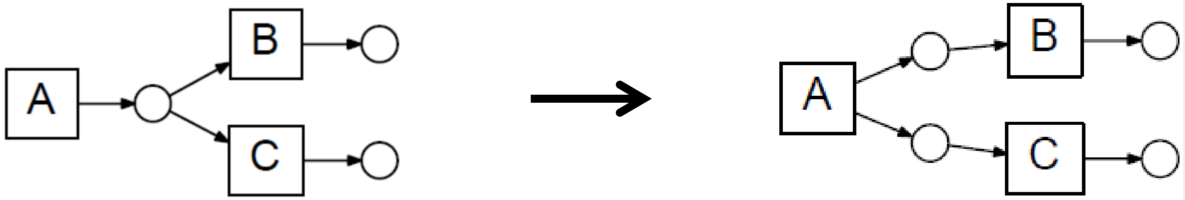
- The mutation operator can:
  - **Add a new task** to the input and/or output function of task



  - **Remove a task** from the input and/or output function of task



  - **Redistribute the elements** of the input and/or output function of task



Evolutionary cycle – *Recombination*: **Mutation**

Guided by the causal dependencies of the log

Two sets for each task:

- inputDependencies(t):
  - The set of activities that appear ***before t*** in any trace of the log
- outputDependencies(t):
  - The set of activities that appear ***after t*** in any trace of the log
- **Reduce the search space** to those models that are supported by the information in the log

ProDiGen

Evolutionary cycle – *Recombination*: **Mutation**

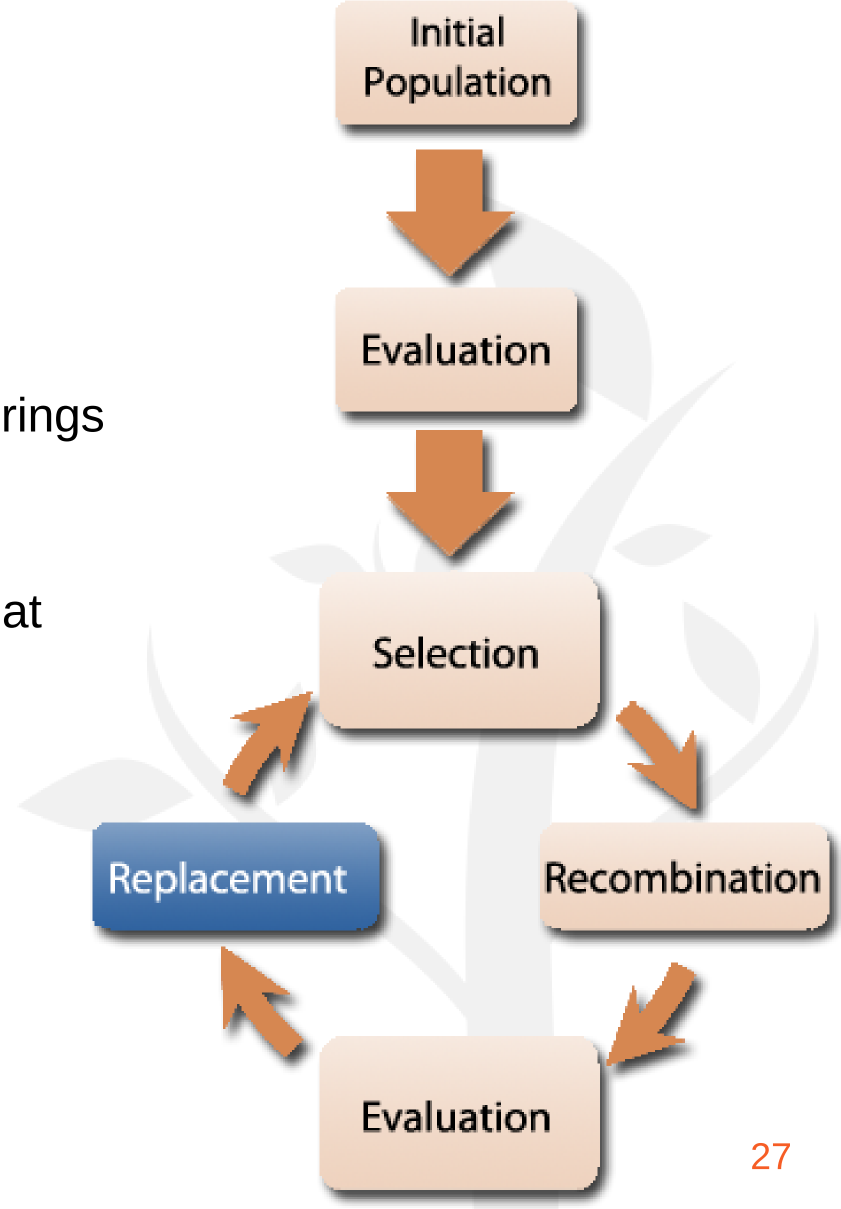
Additionally, to **minimize duplicate** individuals:

- The individual is **iteratively modified** until it is different from its parent
- **Only one task is affected** by the mutation process
- Individuals are **always forced to mutate**
  - Mutation rate = 1

ProDiGen

Evolutionary cycle – **Replacement**

- Update of the population:
  - Combines and sorts parents and offsprings (2N size population)
  - The **repeated individuals** are placed at the bottom of the ranking
  - The **N best individuals survive** to the next cycle

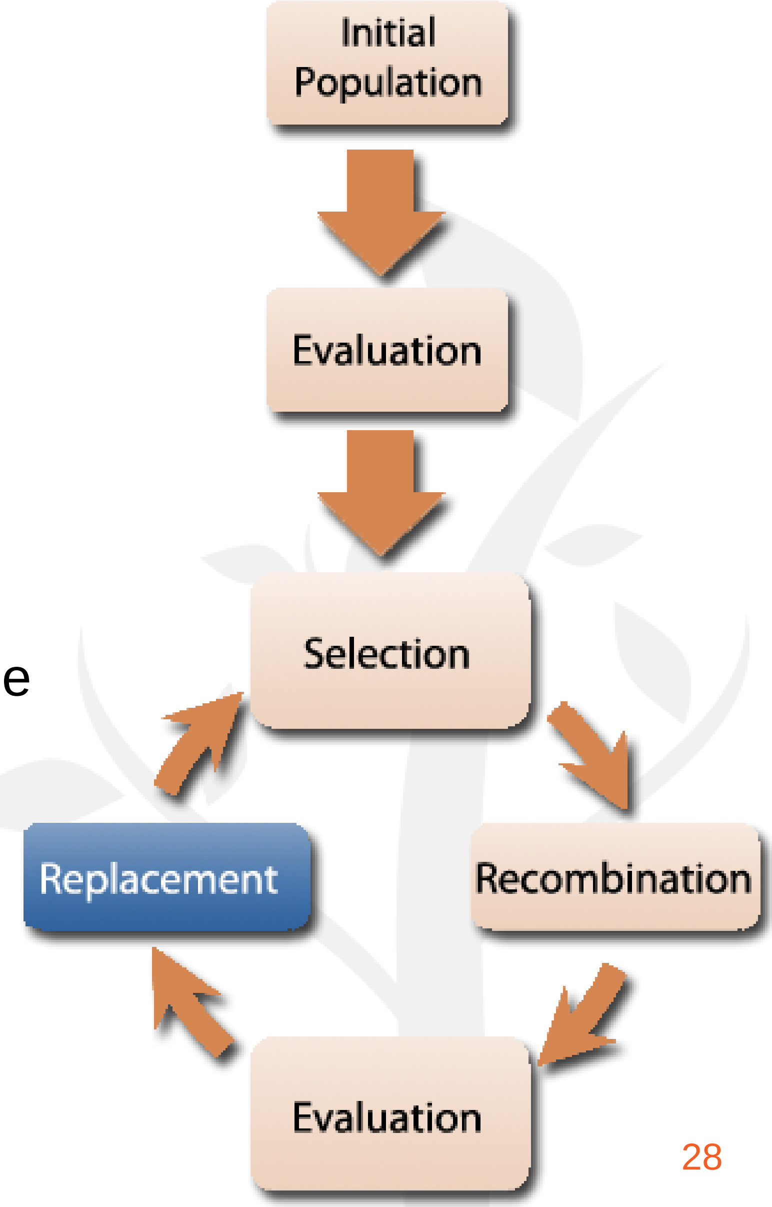


CiTIUS

ProDiGen

Evolutionary cycle – **Reinitialization**

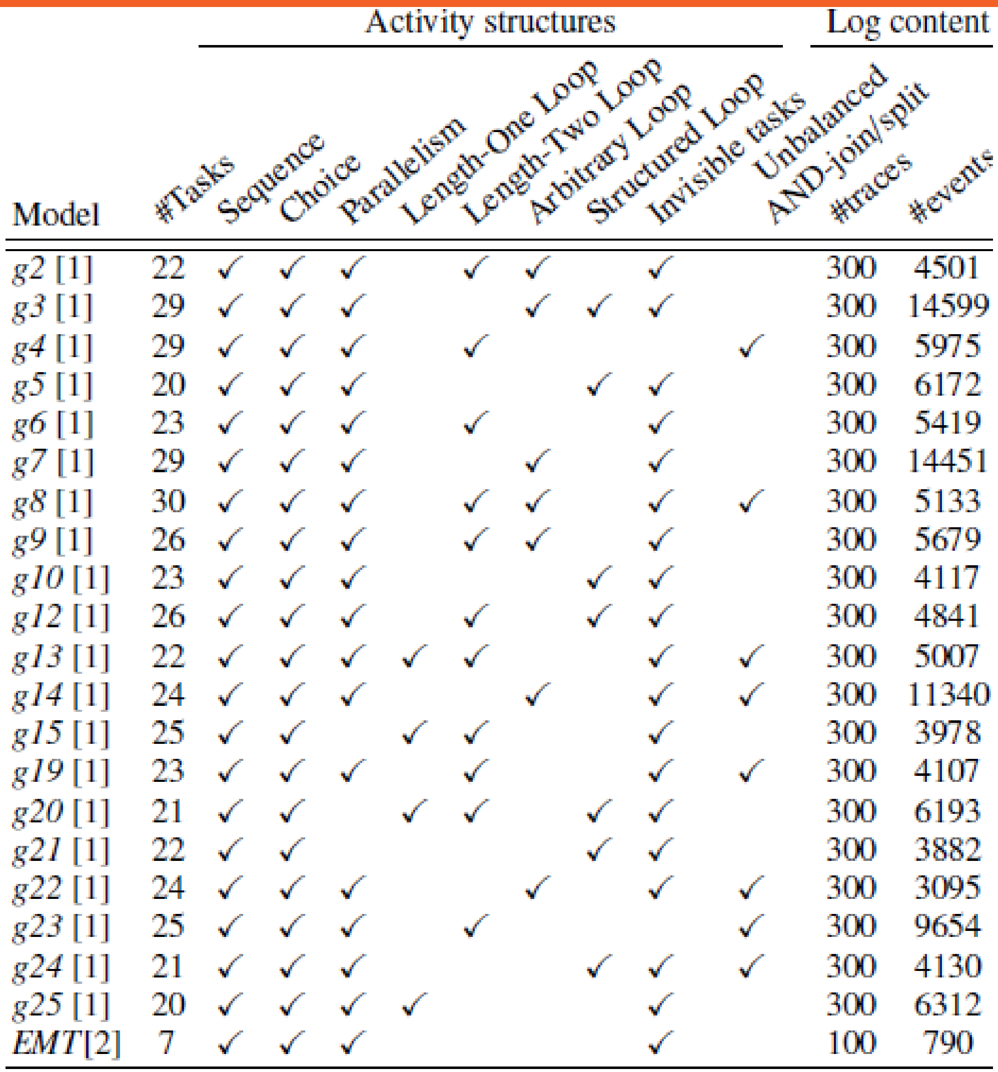
- Indicators:
  - If the best solution does not change
  - If there are not new individuals in the population
- Population generated as in the initial stage
- Adds a **mutation of the best individual**



Experimentation

Logs

- **21 models** with their respective logs
- Both logs and models **created by other authors [1,2]**
- Used to validate:
  - **Genetic Miner [1]**
  - **ETM [2]**



Model	#Tasks	Sequence	Choice	Parallelism	Length-One Loop	Length-Two Loop	Arbitrary Loop	Structured Loop	Invisible tasks	Unbalanced AND-join/split	#traces	#events
	Activity structures										Log content	
g2 [1]	22	✓	✓	✓		✓	✓		✓		300	4501
g3 [1]	29	✓	✓	✓			✓	✓	✓		300	14599
g4 [1]	29	✓	✓	✓		✓				✓	300	5975
g5 [1]	20	✓	✓	✓				✓	✓		300	6172
g6 [1]	23	✓	✓	✓		✓			✓		300	5419
g7 [1]	29	✓	✓	✓			✓		✓		300	14451
g8 [1]	30	✓	✓	✓		✓	✓		✓	✓	300	5133
g9 [1]	26	✓	✓	✓		✓	✓		✓		300	5679
g10 [1]	23	✓	✓	✓				✓	✓		300	4117
g12 [1]	26	✓	✓	✓		✓		✓	✓		300	4841
g13 [1]	22	✓	✓	✓	✓	✓			✓	✓	300	5007
g14 [1]	24	✓	✓	✓			✓		✓	✓	300	11340
g15 [1]	25	✓	✓		✓	✓			✓		300	3978
g19 [1]	23	✓	✓	✓		✓			✓	✓	300	4107
g20 [1]	21	✓	✓		✓	✓		✓	✓		300	6193
g21 [1]	22	✓	✓					✓	✓		300	3882
g22 [1]	24	✓	✓	✓			✓		✓	✓	300	3095
g23 [1]	25	✓	✓	✓		✓				✓	300	9654
g24 [1]	21	✓	✓	✓				✓	✓	✓	300	4130
g25 [1]	20	✓	✓	✓	✓				✓		300	6312
EMT[2]	7	✓	✓	✓					✓		100	790

[1] *de Medeiros, A.K.A : Genetic Process Mining. PhD thesis*

[2] *Buijs, J., van Dongen, B., van der Aalst, W.: On the role of fitness, precision, generalization and simplicity in process discovery.*

Experimentation

Metrics – **Based on the original model** [1]

- To quantify the **behavior similarity** :
  - Behavioral precision (B_p)
  - Behavioral recall (B_r)
- To quantify the **structural similiarty :**
  - Structural precision (S_p)
  - Structural recall (S_r)

[1] *de Medeiros, A.K.A : Genetic Process Mining. PhD thesis, Technische Universiteit Eindhoven*

Experimentation

Metrics – **Based on the log [4]**

- To measure the completeness:
  - **Proper Completion (C):** Percentage of correctly parsed traces. If all the traces are correctly parsed, then **C = 1**

- To measure the precision:
  - **Alignment precision (P):** If all the behavior allowed by the model is actually observed, then **P = 1**

- To measure the simplicity:
  - The **weighted P/T average arc degree (S'):** The lower is S, the higher the complexity

$$S = \frac{1}{1 + S'}$$

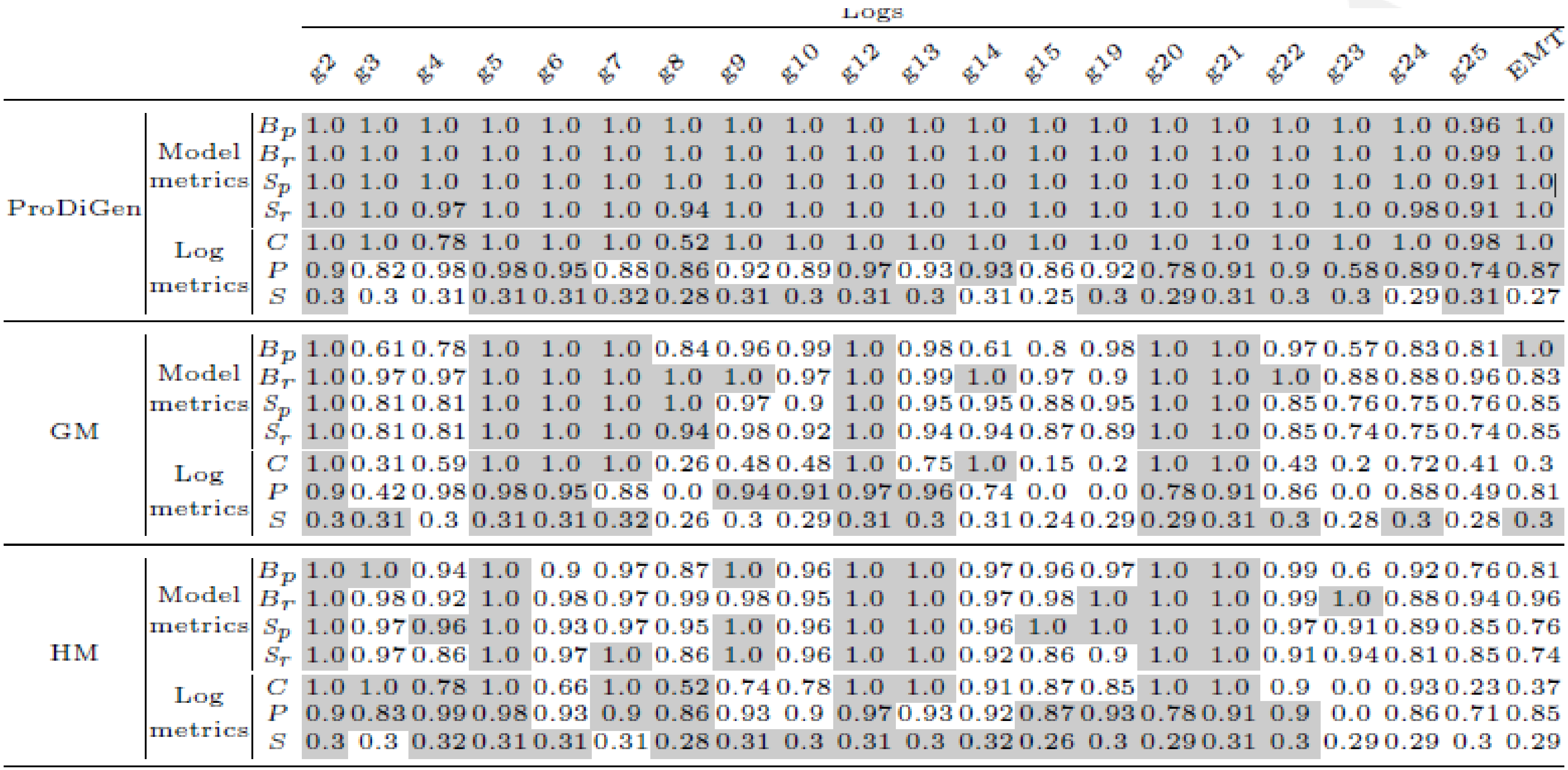
CiTIUS

[4] van der Aalst, W., Adriansyah, A., van Dongen, B.: Replaying history on process models for conformance checking and performance analysis.

Experimentation

Results

- ProDiGen **finds the original model in the 81%** (17 out of 21) of the cases



			Logs																				
			g2	g3	g4	g5	g6	g7	g8	g9	g10	g12	g13	g14	g15	g19	g20	g21	g22	g23	g24	g25	EMT
ProDiGen	Model metrics	B_p	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.96	1.0
		B_r	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99	1.0
		S_p	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.91	1.0
		S_r	1.0	1.0	0.97	1.0	1.0	1.0	0.94	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.98	0.91	1.0
	Log metrics	C	1.0	1.0	0.78	1.0	1.0	1.0	0.52	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.98	1.0
		P	0.9	0.82	0.98	0.98	0.95	0.88	0.86	0.92	0.89	0.97	0.93	0.93	0.86	0.92	0.78	0.91	0.9	0.58	0.89	0.74	0.87
		S	0.3	0.3	0.31	0.31	0.31	0.32	0.28	0.31	0.3	0.31	0.3	0.31	0.25	0.3	0.29	0.31	0.3	0.3	0.29	0.31	0.27
GM	Model metrics	B_p	1.0	0.61	0.78	1.0	1.0	1.0	0.84	0.96	0.99	1.0	0.98	0.61	0.8	0.98	1.0	1.0	0.97	0.57	0.83	0.81	1.0
		B_r	1.0	0.97	0.97	1.0	1.0	1.0	1.0	1.0	0.97	1.0	0.99	1.0	0.97	0.9	1.0	1.0	1.0	0.88	0.88	0.96	0.83
		S_p	1.0	0.81	0.81	1.0	1.0	1.0	1.0	0.97	0.9	1.0	0.95	0.95	0.88	0.95	1.0	1.0	0.85	0.76	0.75	0.76	0.85
		S_r	1.0	0.81	0.81	1.0	1.0	1.0	0.94	0.98	0.92	1.0	0.94	0.94	0.87	0.89	1.0	1.0	0.85	0.74	0.75	0.74	0.85
	Log metrics	C	1.0	0.31	0.59	1.0	1.0	1.0	0.26	0.48	0.48	1.0	0.75	1.0	0.15	0.2	1.0	1.0	0.43	0.2	0.72	0.41	0.3
		P	0.9	0.42	0.98	0.98	0.95	0.88	0.0	0.94	0.91	0.97	0.96	0.74	0.0	0.0	0.78	0.91	0.86	0.0	0.88	0.49	0.81
		S	0.3	0.31	0.3	0.31	0.31	0.32	0.26	0.3	0.29	0.31	0.3	0.31	0.24	0.29	0.29	0.31	0.3	0.28	0.3	0.28	0.3
HM	Model metrics	B_p	1.0	1.0	0.94	1.0	0.9	0.97	0.87	1.0	0.96	1.0	1.0	0.97	0.96	0.97	1.0	1.0	0.99	0.6	0.92	0.76	0.81
		B_r	1.0	0.98	0.92	1.0	0.98	0.97	0.99	0.98	0.95	1.0	1.0	0.97	0.98	1.0	1.0	1.0	0.99	1.0	0.88	0.94	0.96
		S_p	1.0	0.97	0.96	1.0	0.93	0.97	0.95	1.0	0.96	1.0	1.0	0.96	1.0	1.0	1.0	1.0	0.97	0.91	0.89	0.85	0.76
		S_r	1.0	0.97	0.86	1.0	0.97	1.0	0.86	1.0	0.96	1.0	1.0	0.92	0.86	0.9	1.0	1.0	0.91	0.94	0.81	0.85	0.74
	Log metrics	C	1.0	1.0	0.78	1.0	0.66	1.0	0.52	0.74	0.78	1.0	1.0	0.91	0.87	0.85	1.0	1.0	0.9	0.0	0.93	0.23	0.37
		P	0.9	0.83	0.99	0.98	0.93	0.9	0.86	0.93	0.9	0.97	0.93	0.92	0.87	0.93	0.78	0.91	0.9	0.0	0.86	0.71	0.85
		S	0.3	0.3	0.32	0.31	0.31	0.31	0.28	0.31	0.3	0.31	0.3	0.32	0.26	0.3	0.29	0.31	0.3	0.29	0.29	0.3	0.29

Current applications of ProDiGen

SoftLearn

- Is being used to discover the workflow that **represents the learning path followed by the learners** during the course

SoftLearn

SoftLearn demo Web Page

tec.citius.usc.es/SoftLearn

Conclusions

- Genetic algorithm for process discovery guided by **completeness, precision and simplicity**
- New criteria for **precision** and **simplicity**
- Recombination guided by **heuristics**
- Heuristics Miner's solution is **incorporated to the initial population**
- **Great** performance

THANKS FOR YOUR ATTENTION

borja.vazquez@usc.es

ProDiGen

tec.citius.usc.es/SoftLearn/ProDiGen.html

SoftLearn demo Web Page

tec.citius.usc.es/SoftLearn

CiTIUS