On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery

Joos Buijs

Boudewijn van Dongen

Wil van der Aalst

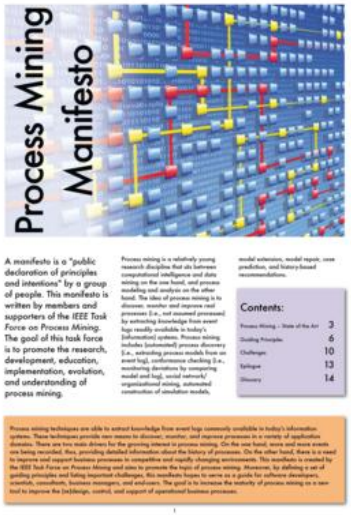
http://www.win.tue.nl/coselog/

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

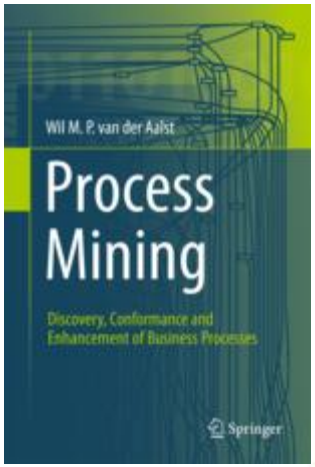
Advances in Process Mining

- **Many process discovery and conformance checking algorithms and tools are available (cf. the various ProM packages).**
- **Also commercial software based on these ideas:**
  Disco (Fluxicon), Reflect (Futura/Perceptive), BPMOne (Pallas Athena/Perceptive), ARIS Process Performance Manager (Software AG), Interstage Automated Process Discovery (Fujitsu), QPR ProcessAnalyzer/Analysis (QPR Software), flow (fourspark), Discovery Analyst (StereoLOGIC), etc.
- **We applied process mining in over 100 organizations.**



More than 75 people involving more than 50 organizations created the Process Mining Manifesto in the context of the IEEE Task Force on Process Mining.

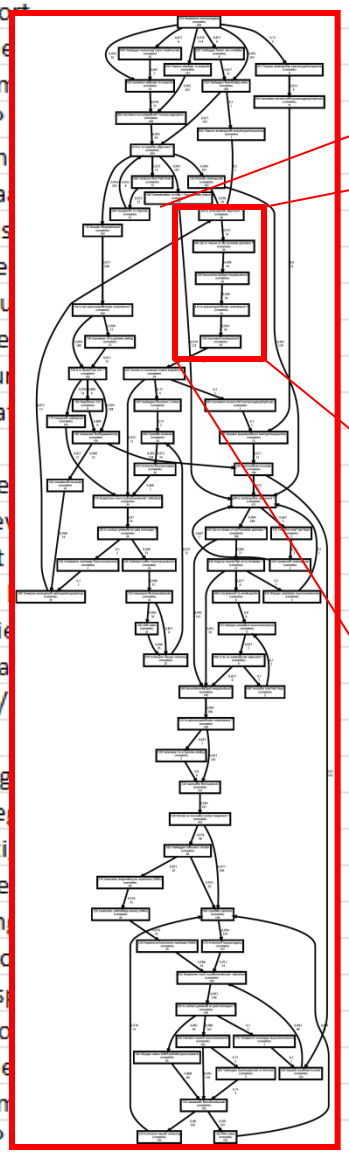
Available in 13 languages

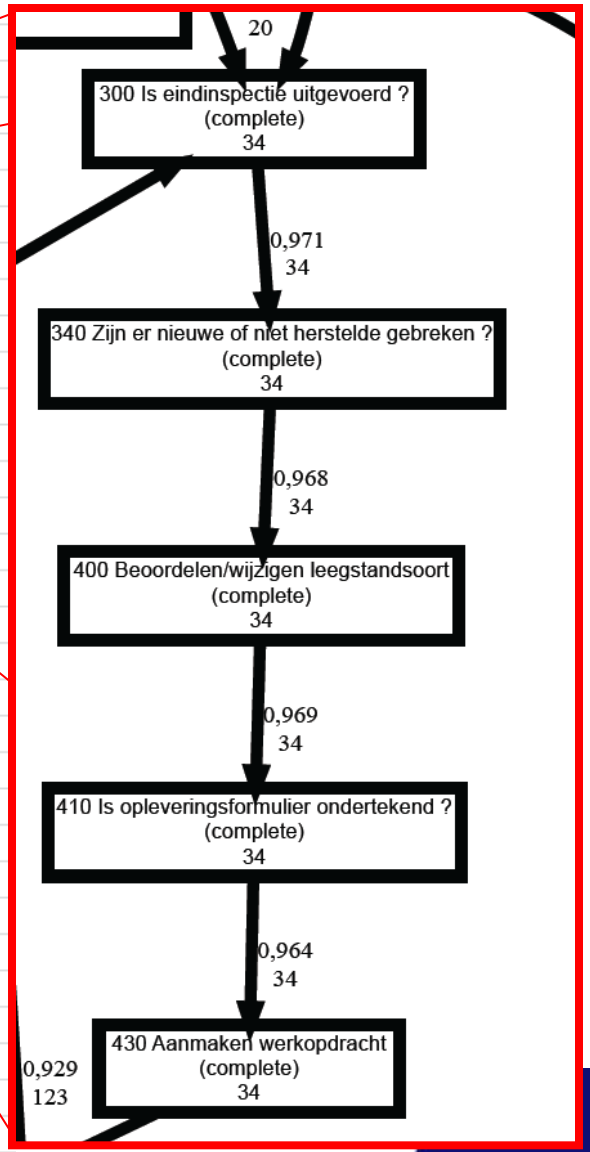


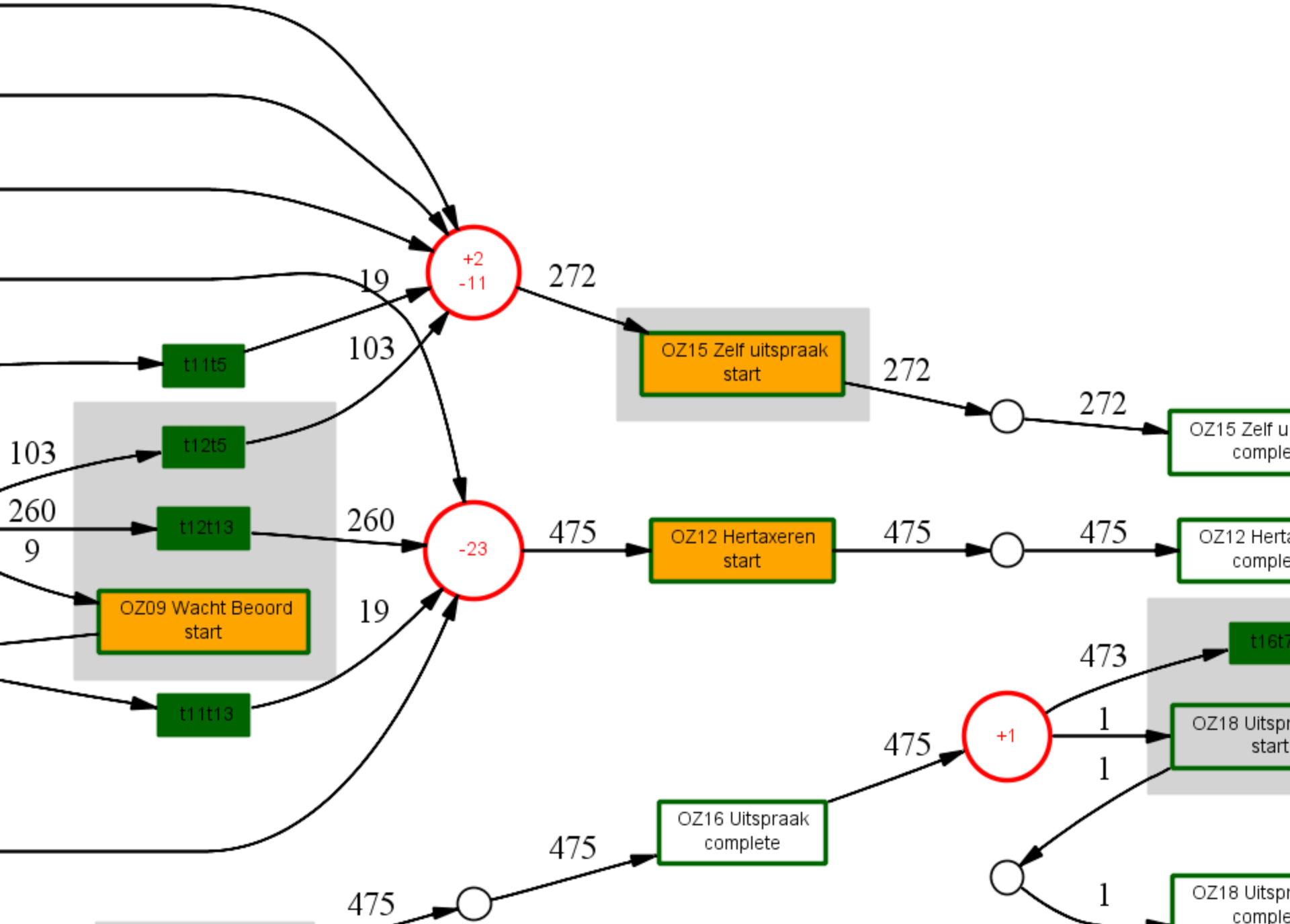
Example Process Discovery

(Vestia, Dutch housing agency, 208 cases, 5987 events)

117315	110 Bepalen leegstandsoo	16.05.2007 14:06:23
117315	120 Plannen eindinspectie	16.05.2007 14:36:01
117315	130 Is het opleveringsform	23.05.2007 09:41:40
117315	150 Is er sprake van ZAV ?	23.05.2007 09:41:51
117315	170 Aanpassen plattegron	23.05.2007 11:57:18
117315	180 Aanpassen woningwa	23.05.2007 09:42:37
117315	190 Actualiseren huurprijs	23.05.2007 09:48:23
117315	200 Toewijzen woning/be	23.05.2007 09:48:29
117315	210 Registreren voorl. huu	10.09.2007 16:24:36
117315	220 Is contract getekend e	11.09.2007 14:56:18
117315	240 Definitief maken Huur	31.03.2008 16:17:12
117315	250 Aanpassen factureera	09.09.2008 15:39:59
117315	260 After sales	09.09.2008 16:51:24
117315	270 Archiveren nieuwe ve	10.09.2008 07:52:08
117315	300 Is eindinspectie uitgev	07.06.2007 14:47:04
117315	340 Zijn er nieuwe of niet	07.06.2007 14:47:06
117315	400 Beoordelen/wijzigen l	07.06.2007 14:51:16
117315	410 Is opleveringsformulie	07.06.2007 14:51:26
117315	430 Aanmaken werkopdra	11.06.2007 09:21:39
117315	440 Worden er bonussen/	11.06.2007 09:21:49
117315	460 Opstellen eindnota	08.08.2007 16:18:26
117315	470 Archiveren huuropzeg	09.08.2007 14:42:23
119763	010 Registreren huuropze	09.05.2007 11:19:14
119763	030 Vastleggen toekomsti	09.05.2007 12:25:01
119763	050 Inplannen afspraak 1e	09.05.2007 11:59:52
119763	060 Aanmaken bevestigin	09.05.2007 12:31:57
119763	070 Is 1e inspectie uitgevo	16.05.2007 13:04:26
119763	100 Gereedmelden 1e insp	16.05.2007 13:43:39
119763	110 Bepalen leegstandsoo	16.05.2007 13:43:28
119763	120 Plannen eindinspectie	16.05.2007 13:42:58
119763	130 Is het opleveringsform	16.05.2007 13:34:49
119763	150 Is er sprake van ZAV ?	16.05.2007 13:34:56






+2
-11
272
19
103
t11t5
t12t5
103
260
t12t13
260
9
OZ09 Wacht Beoord start
19
t11t13
-23
475
OZ15 Zelf uitspraak start
272
272
OZ15 Zelf u
comple
OZ12 Hertaxeren start
475
475
475
OZ12 Herta
comple
473
1
1
+1
475
OZ16 Uitspraak complete
475
475
1
OZ18 Uitspr
start
OZ18 Uitspr
comple

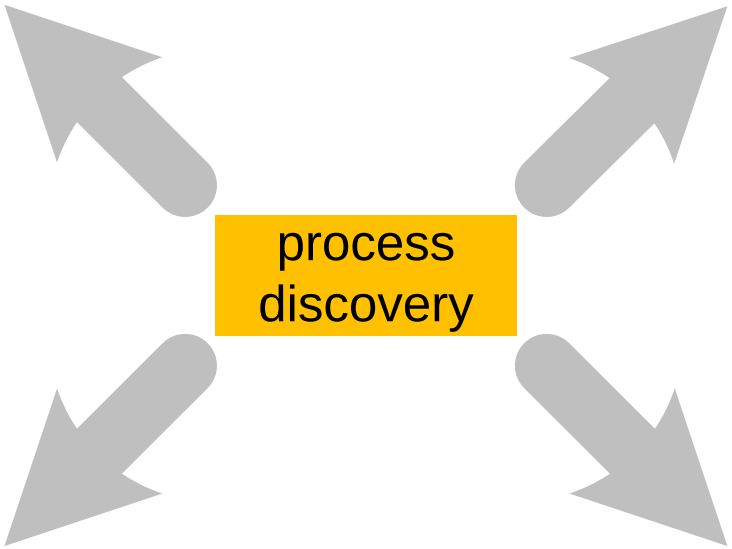
Challenge:
Four Competing Quality Criteria

“able to replay event log”

replay fitness

“Occam’s razor”

simplicity



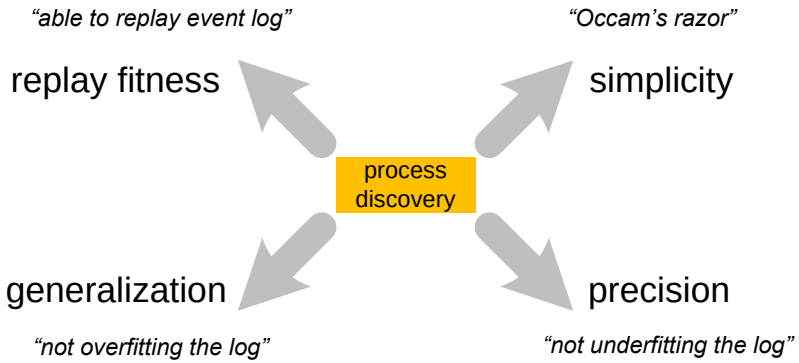
generalization

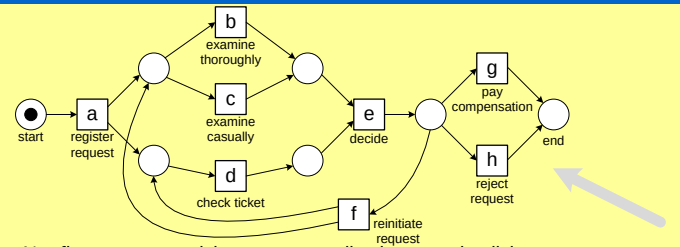
“not overfitting the log”

precision

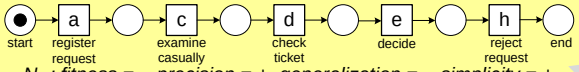
“not underfitting the log”

Example: one log four models

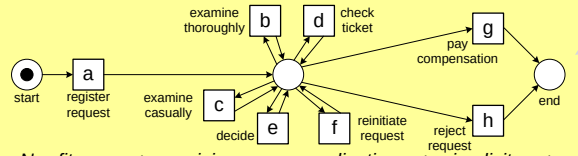




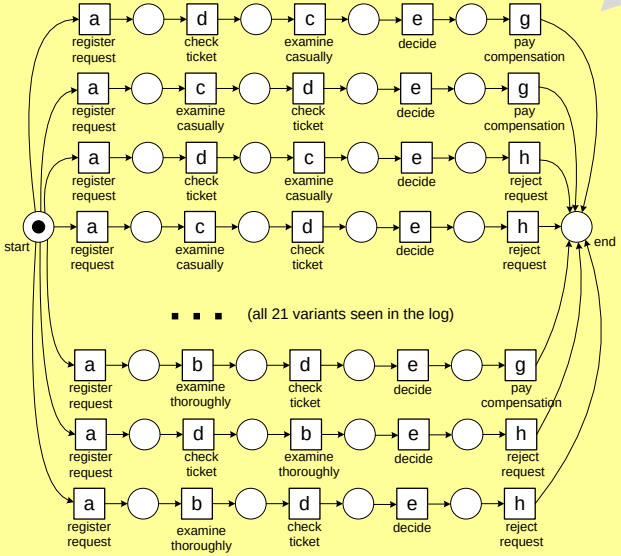
N_1 : *fitness* = +, *precision* = +, *generalization* = +, *simplicity* = +



N_2 : *fitness* = -, *precision* = +, *generalization* = -, *simplicity* = +



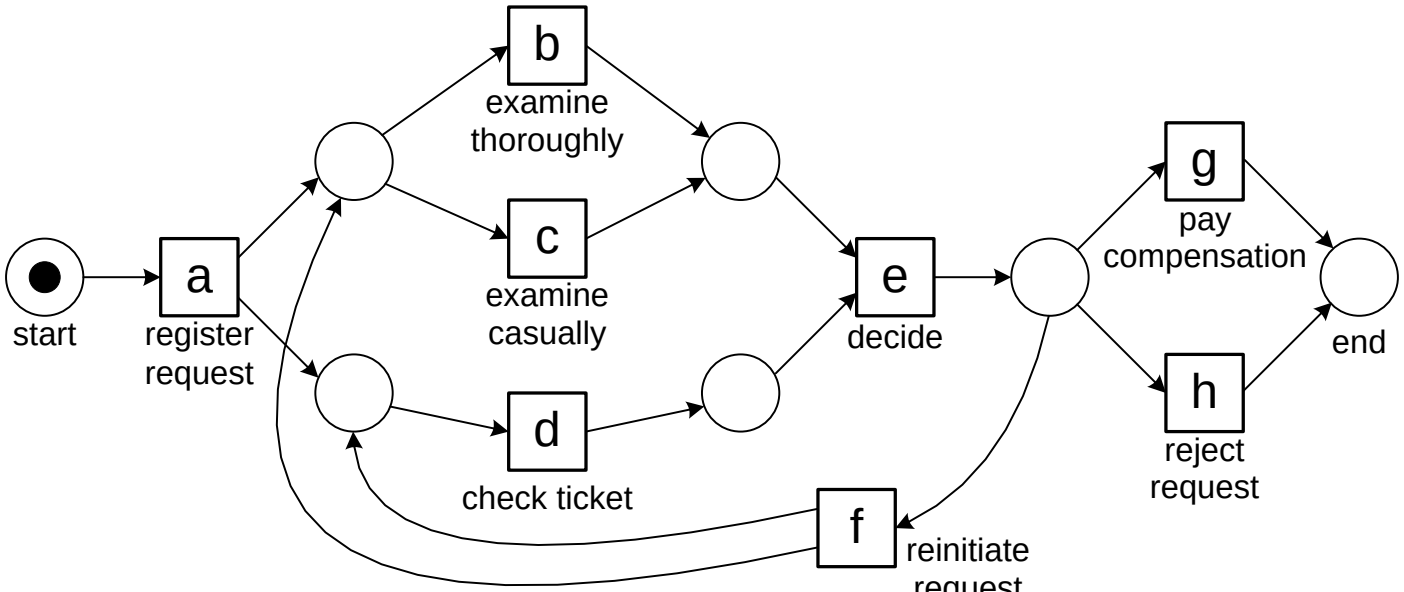
N_3 : *fitness* = +, *precision* = -, *generalization* = +, *simplicity* = +



N_4 : *fitness* = +, *precision* = +, *generalization* = -, *simplicity* = -

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

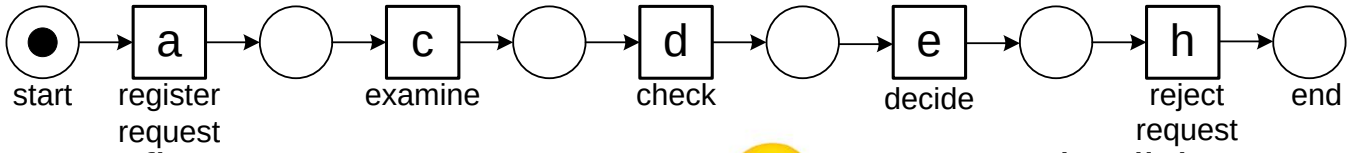
Model N_1



N_1 : fitness = +, precision = +, generalization = +, simplicity = +

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

Model N_2



N_2 : *fitness* =

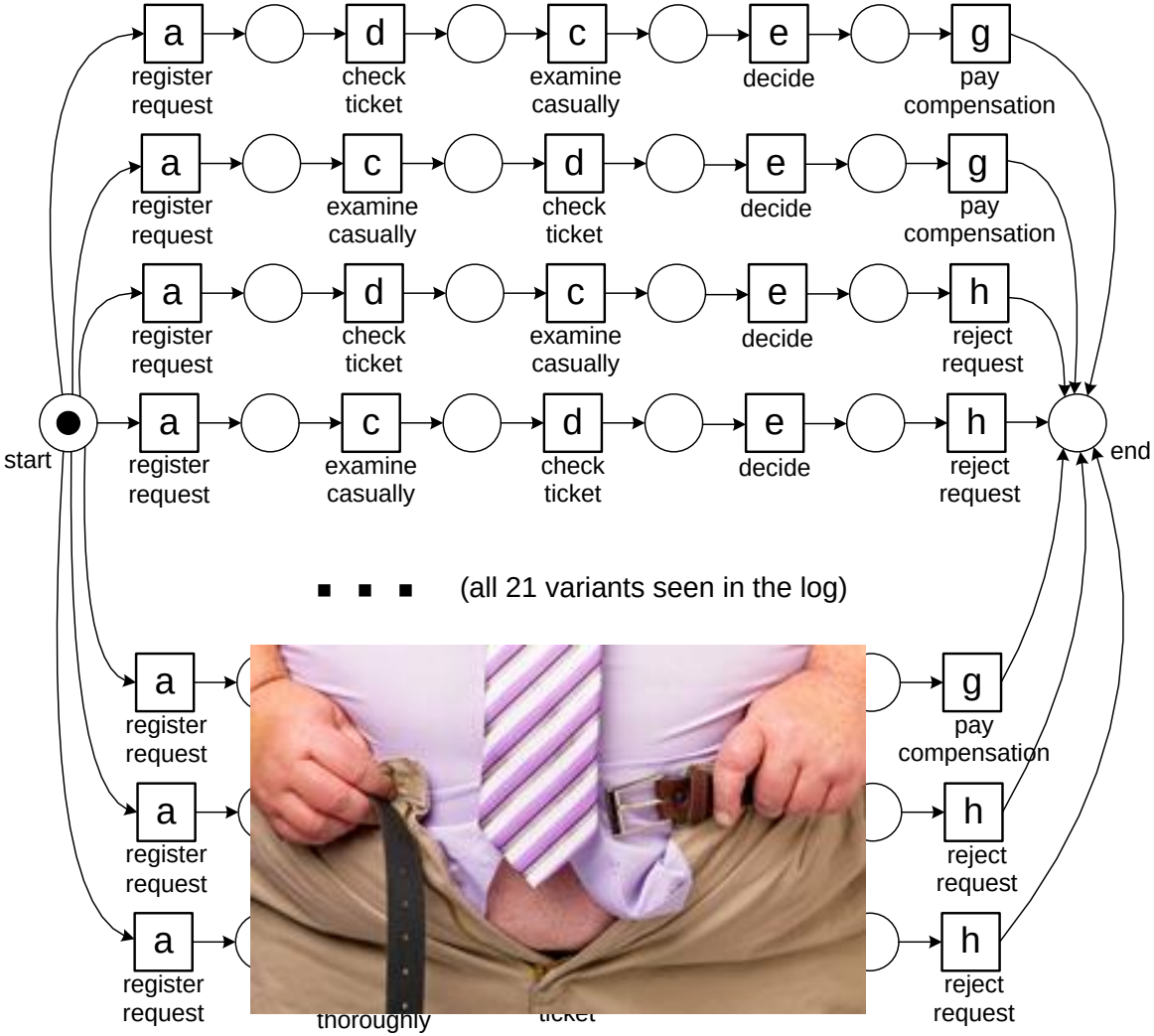
, *simplicity* = +

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

Model N_3

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

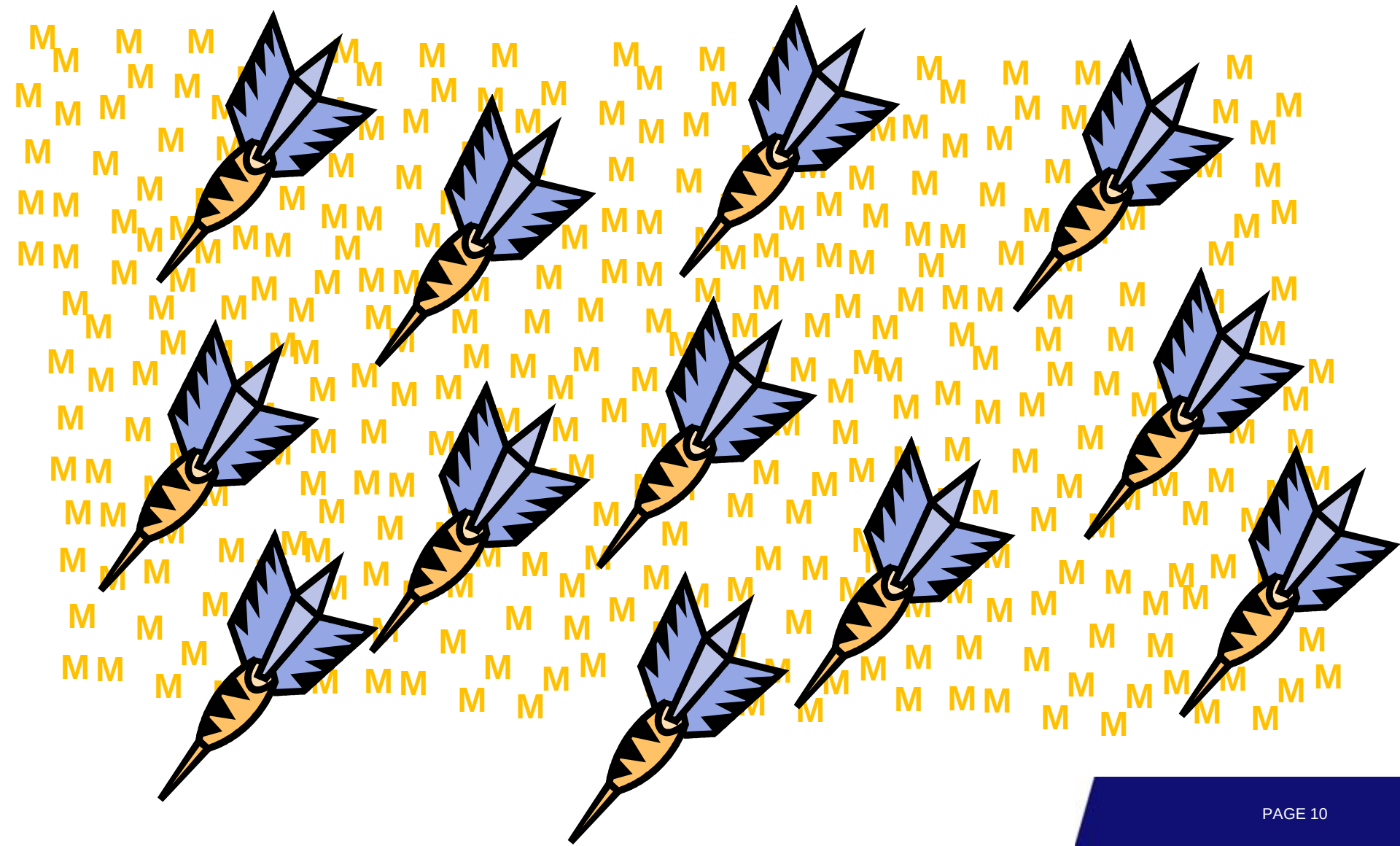
Model N_4



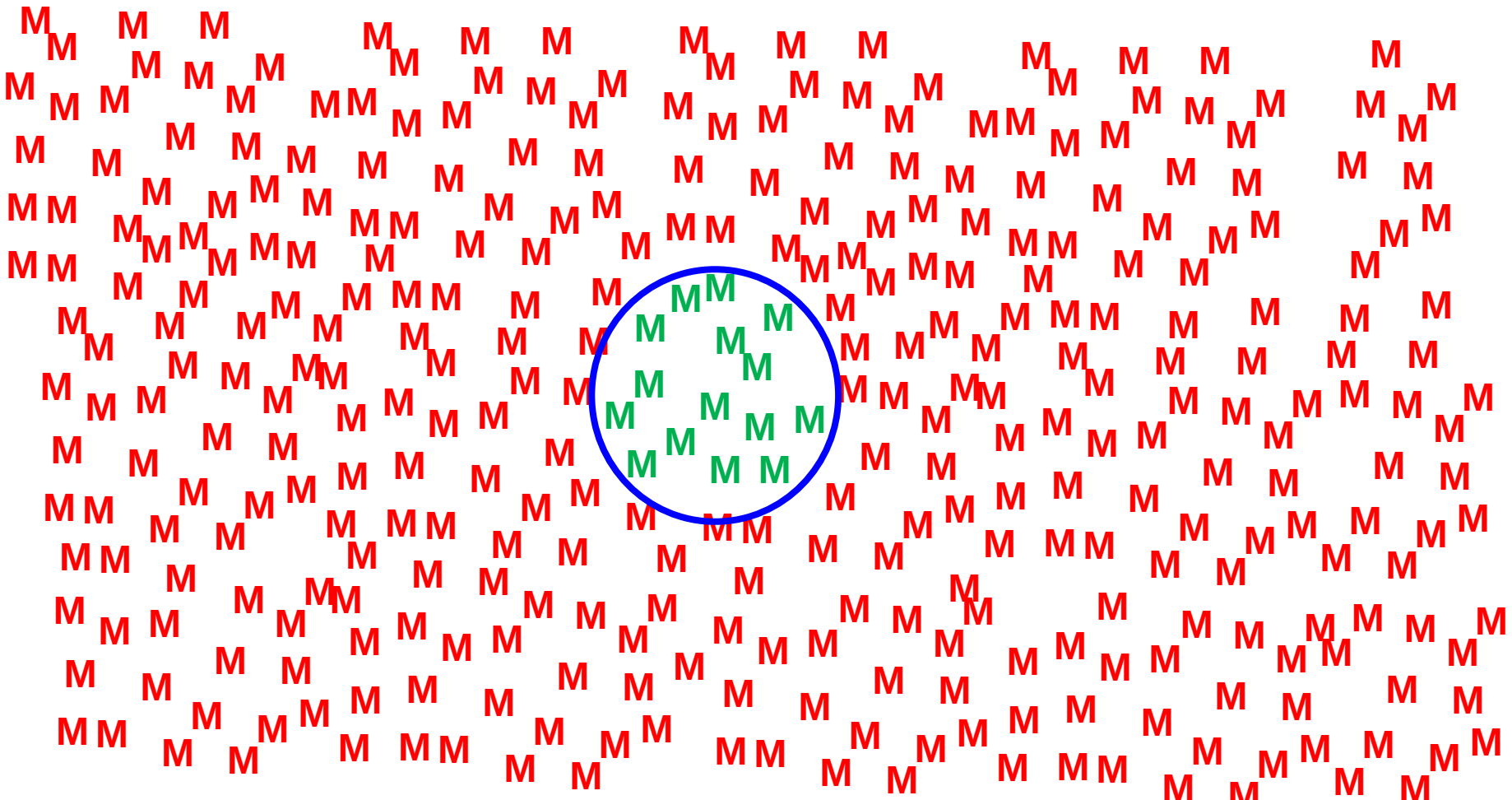
N_4 : fitness = +, precision = +, generalization = -, simplicity = -

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

Another challenge: Huge search space

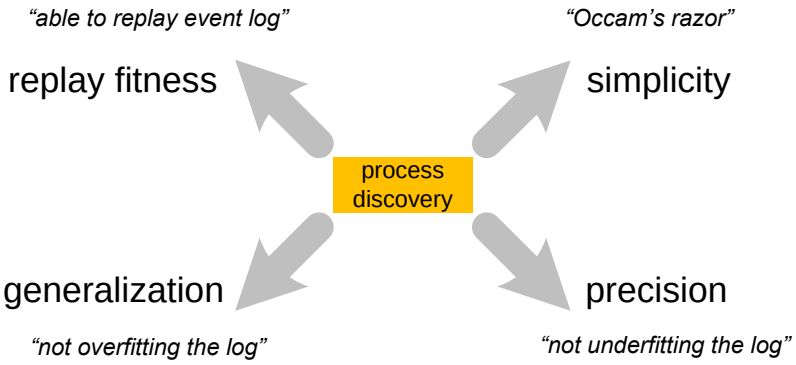


… with just a few interesting candidates

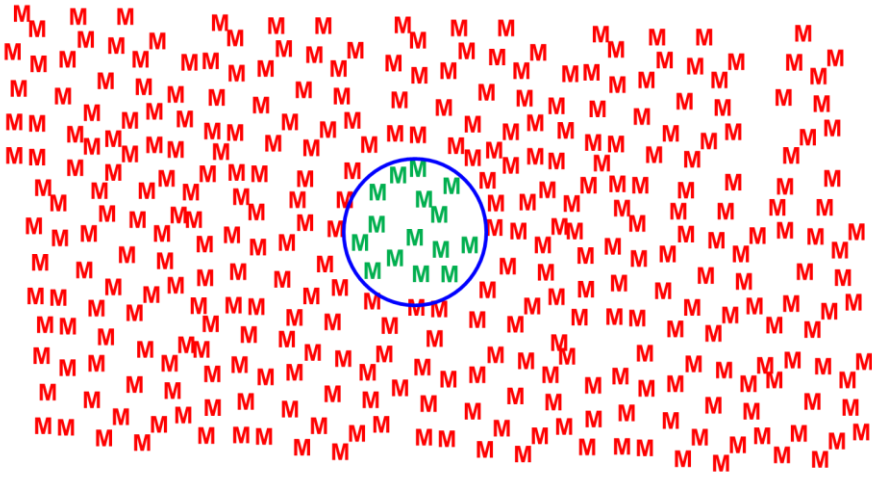


Two requirements

1. **It should be possible to seamlessly balance the different quality criteria based on user-defined preferences.**



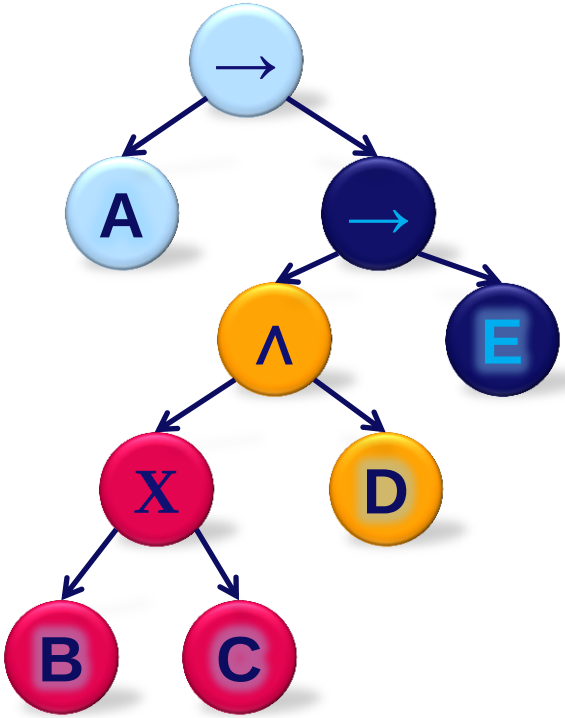
2. **The algorithm should always return a "correct" process model and not waste time on model having deadlocks and other anomalies.**



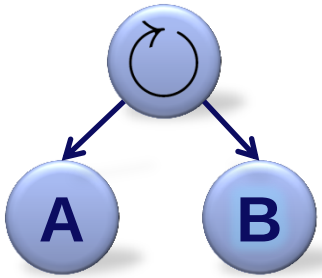
Proposal: Evolutionary Tree Miner (ETM)

- **Process trees as representation (= limit search space to "good" models).**
- **Genetic approach (= very flexible)**
- **Fitness function uses all four criteria (= seamlessly balance the different "forces")**

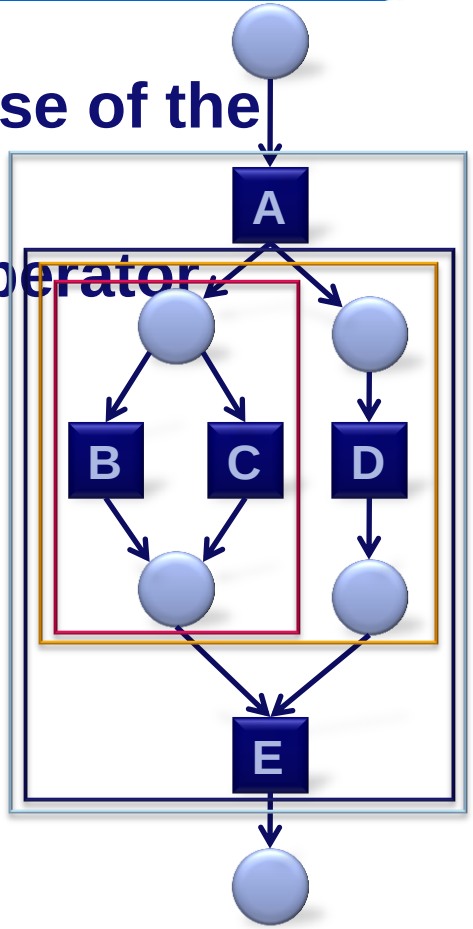
Representational Bias: Process Trees



- Always sound because of the block structure
- Also Loop and OR operator

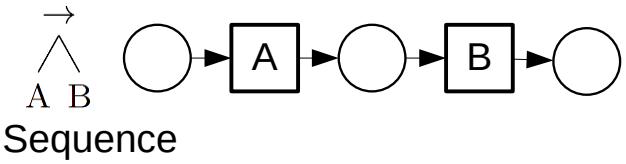


A (BA)*

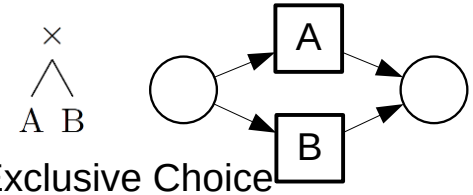


Petri Net Semantics

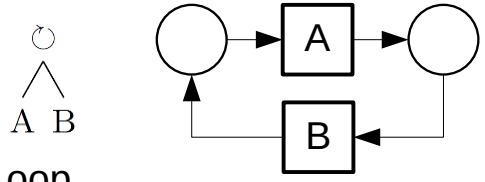
(used for comparison and conformance checking only)



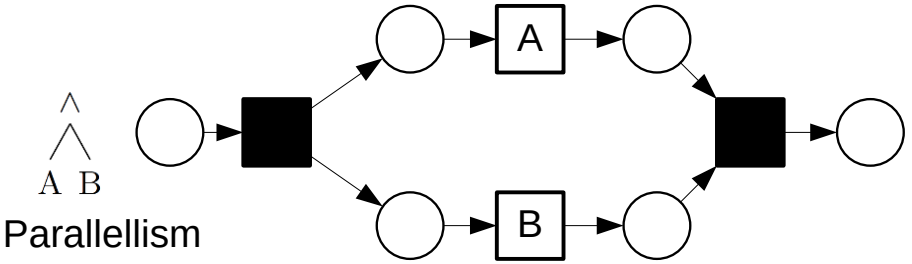
Sequence



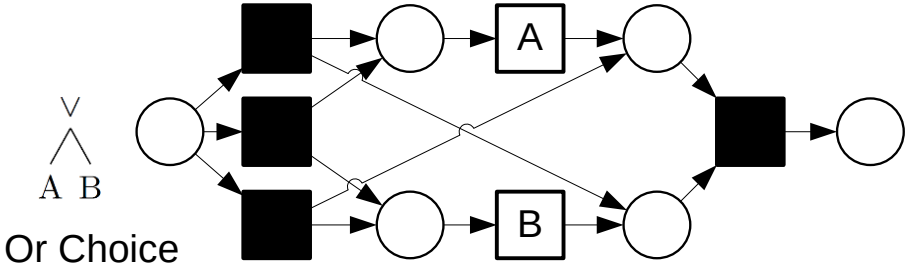
Exclusive Choice



Loop

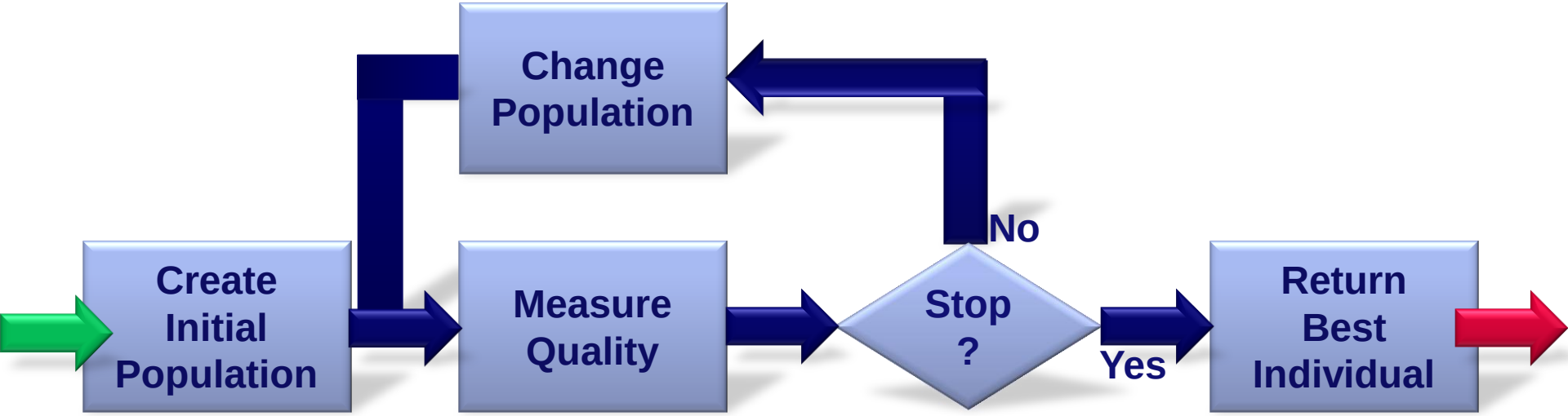


Parallellism

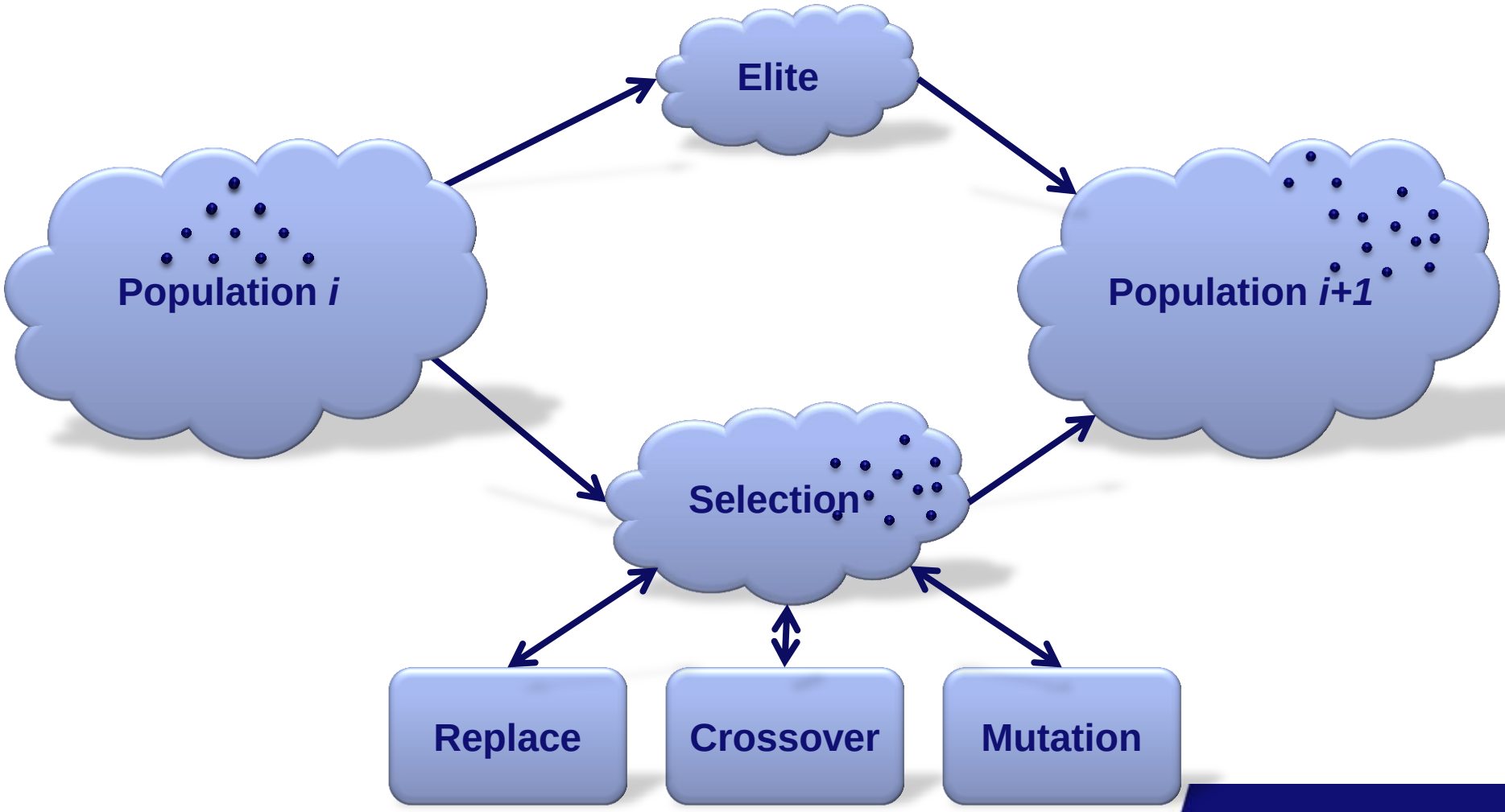


Or Choice

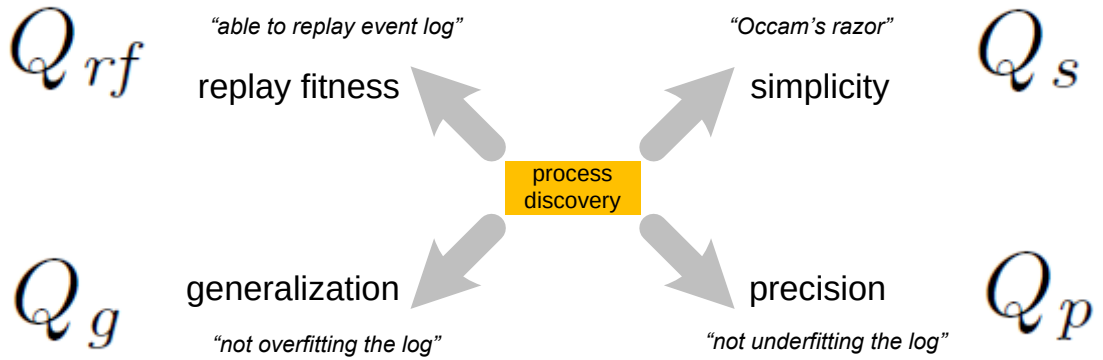
Steps of the Genetic ETM Algorithm



Population Change



Four Metrics (see paper)



$$Q_{rf} = 1 - \frac{\text{cost for aligning model and event log}}{\text{Minimal cost to align arbitrary event log on model and vice versa}}$$

$$Q_s = 1 - \frac{\#\text{duplicate activities} + \#\text{missing activities}}{\#\text{nodes in process tree} + \#\text{event classes in event log}}$$

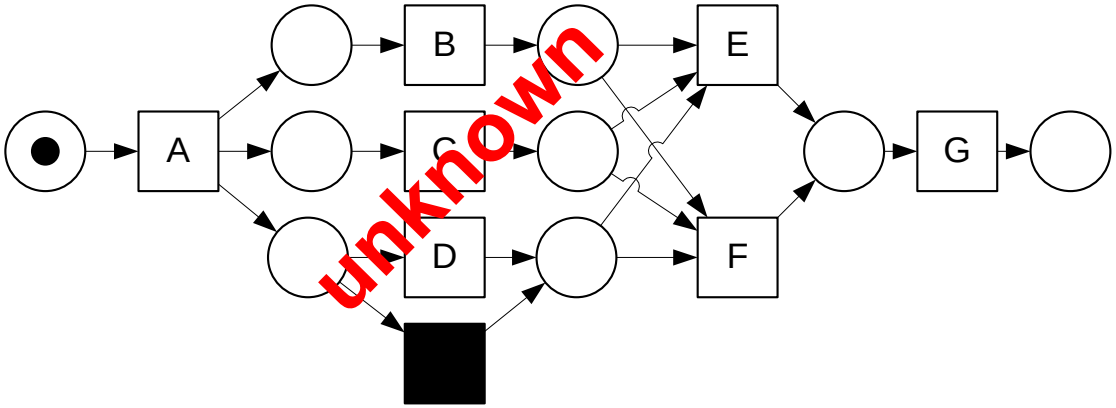
$$Q_p = 1 - \frac{\sum_{\text{visited markings}} \#\text{visits} * \frac{\#\text{outgoing edges} - \#\text{used edges}}{\#\text{outgoing edges}}}{\#\text{total marking visits over all markings}}$$

$$Q_g = 1 - \frac{\sum_{\text{nodes}} (\sqrt{\#\text{executions}})^{-1}}{\#\text{nodes in tree}}$$

1 = optimal
0 = very bad

Example

Trace	#
A B C D E G	6
A B C D F G	38
A B D C E G	12
A B D C F G	26
A B C F G	8
A C B E G	1
A D B C F G	1
A D B C E G	1
A D C B F G	4
A C D B F G	2
A C B F G	1

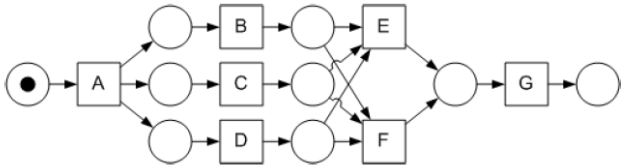


A = send e-mail, B = check credit,
C = calculate capacity, D = check system,
E = accept, F = reject, G = send e-mail

Conventional Algorithms (1/3)

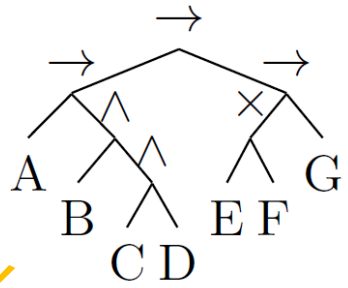
("best effort" mapping to process trees to allow for comparison)

alpha miner



sound

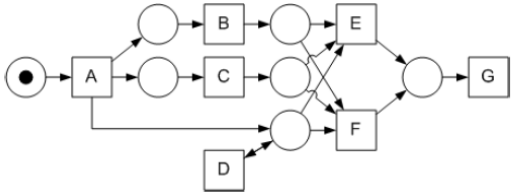
… lucky



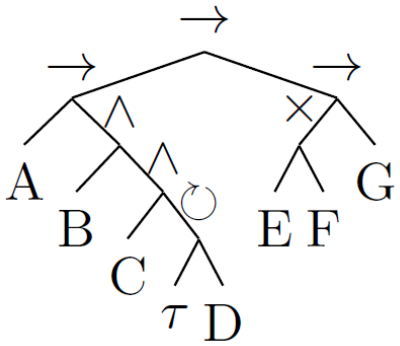
f: 0,992	p: 0,995
s: 1,000	g: 0,889

low fitness

ILP miner



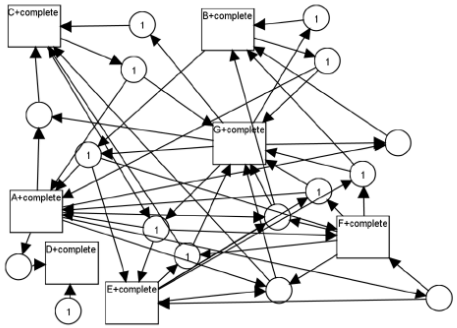
sound



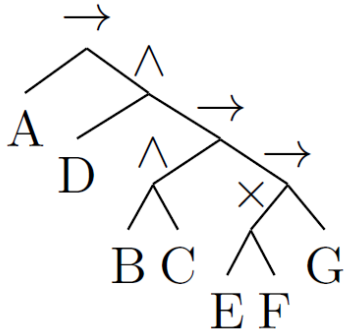
f: 1,000	p: 0,784
s: 0,933	g: 0,830

low precision

language-based region miner



sound

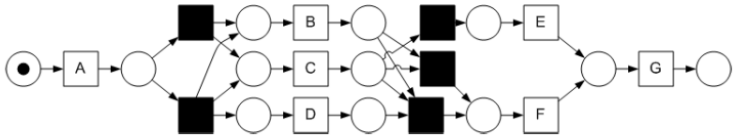


f: 0,992	p: 0,957
s: 1,000	g: 0,889

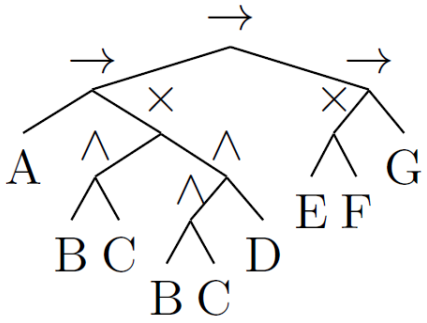
low fitness

Conventional Algorithms (2/3)

heuristic miner

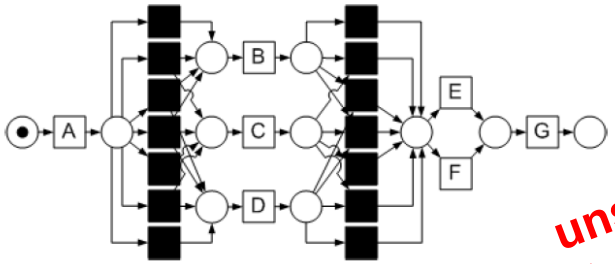


unsound

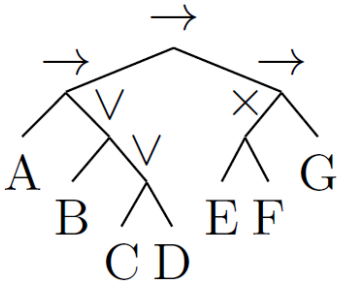


f: 1,000	p: 0,986
s: 0,875	g: 0,852

multi-phase miner



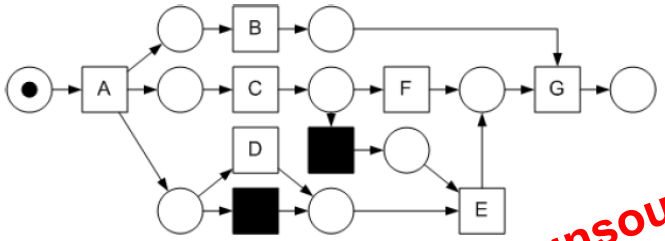
unsound (relaxed sound)



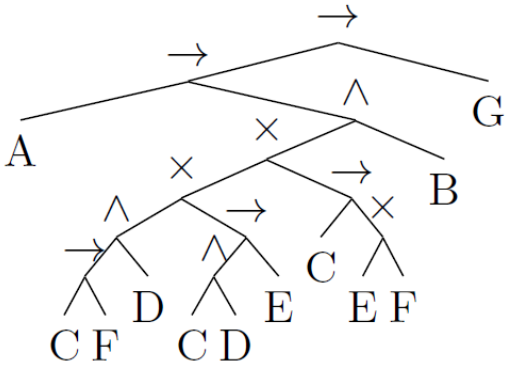
f: 1,000	p: 0,830
s: 1,000	g: 0,889

Conventional Algorithms (3/3)

genetic miner

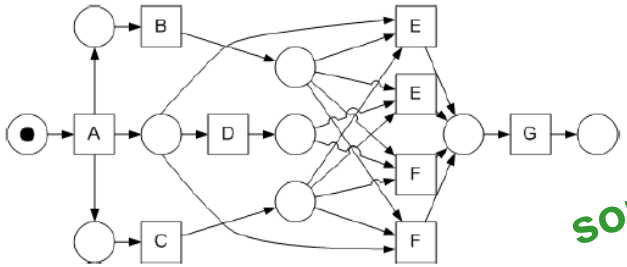


unsound

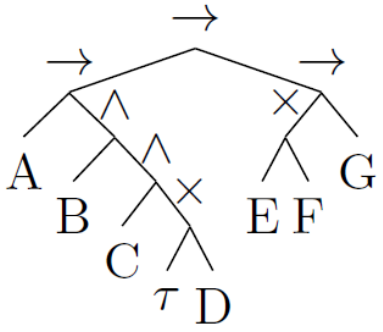


f: 1,000	p: 0,922
s: 0,737	g: 0,790

state-based region miner

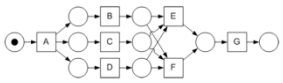


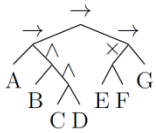
sound



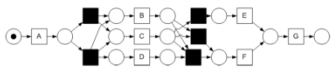
f: 1,000	p: 0,893
s: 0,933	g: 0,830

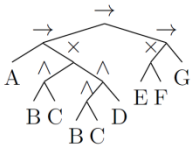
Often unsound result and no mechanism to seamlessly balance the four criteria



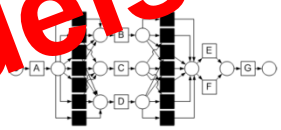


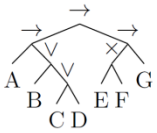
f: 0,992	p: 0,995
s: 1,000	g: 0,889



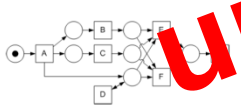


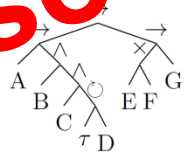
f: 1,000	p: 0,986
s: 0,875	g: 0,875



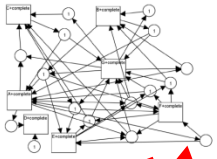


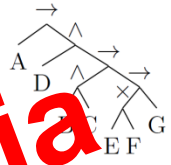
f: 1,000	p: 0,830
s: 1,000	g: 0,889



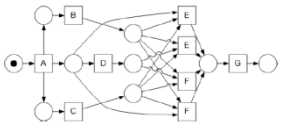


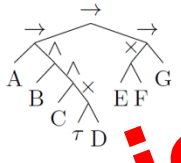
f: 1,000	p: 0,784
s: 0,933	g: 0,830



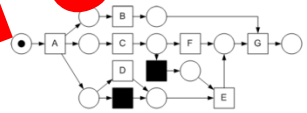


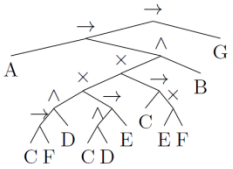
f: 0,992	p: 0,957
s: 1,000	g: 0,889





f: 1,000	p: 0,893
s: 0,933	g: 0,[illegible]



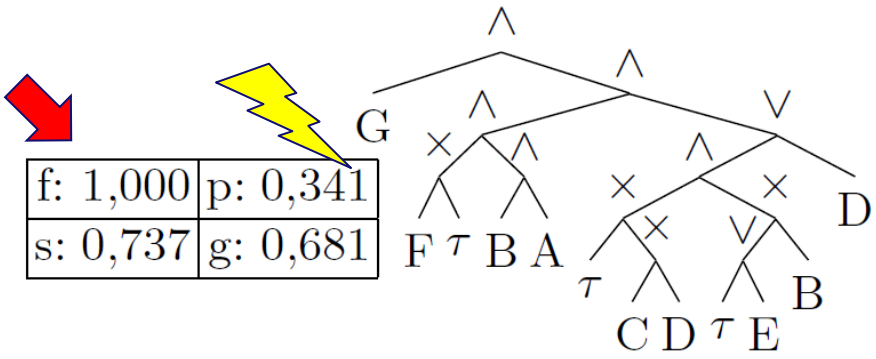


f: 1,000	p: 0,922
s: 0,737	g: 0,790

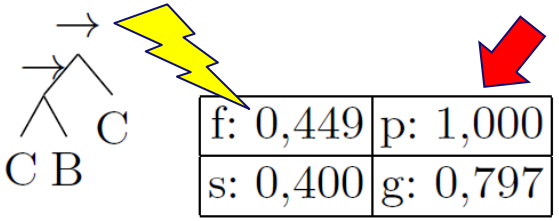
unsound models

ignored criteria

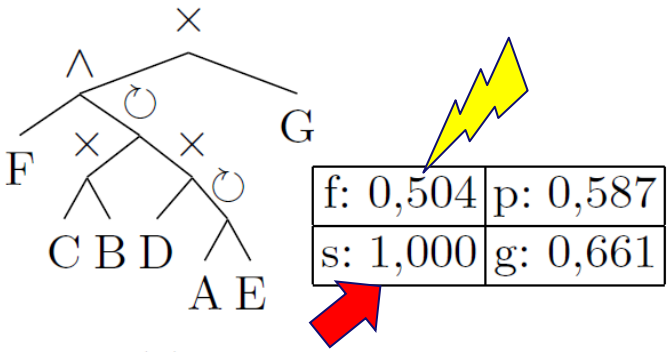
Genetic Mining (ETM) While Considering Only One Criterion



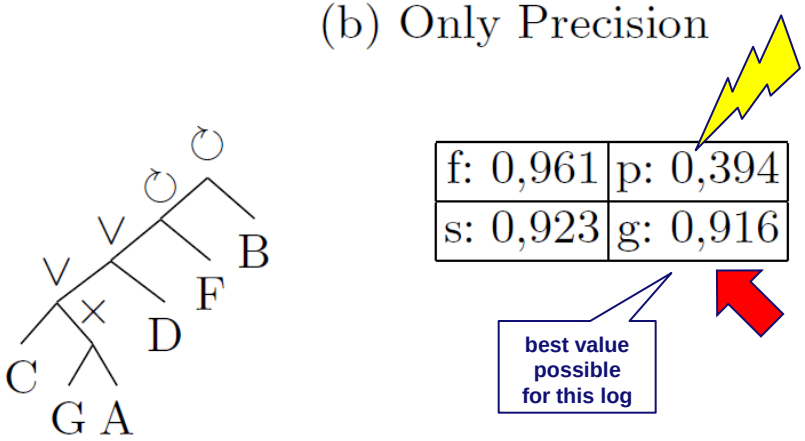
(a) Only replay fitness



(b) Only Precision



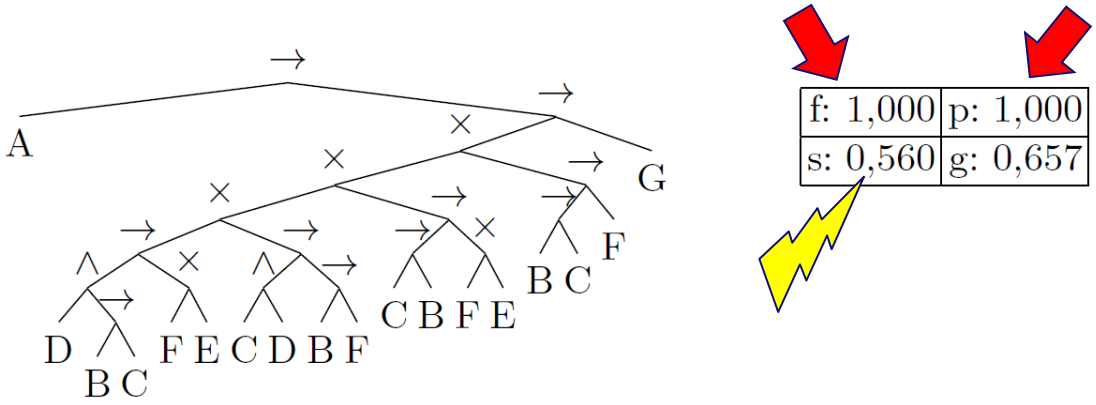
(c) Only Simplicity



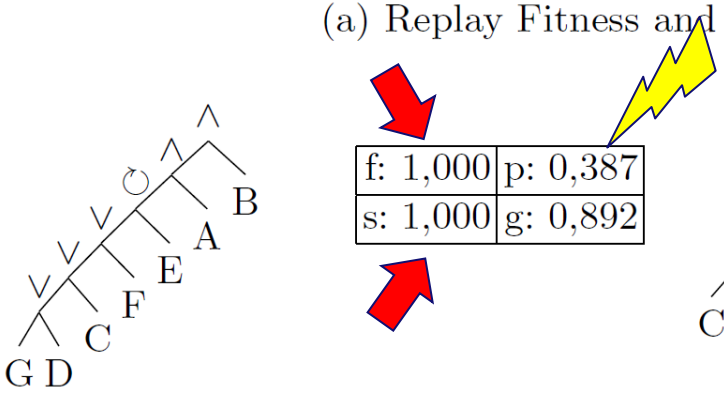
(d) Only Generalization

ETM with weight zero to three out of four perspectives.

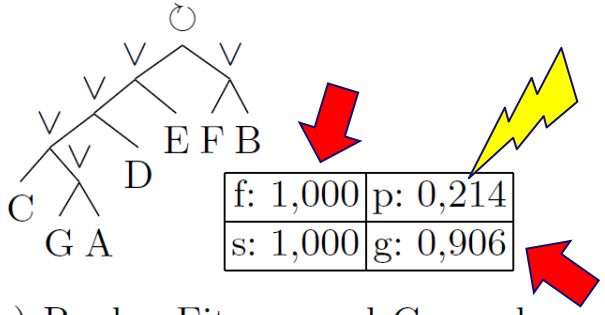
Considering Replay Fitness and One Other Criterion



(a) Replay Fitness and Precision



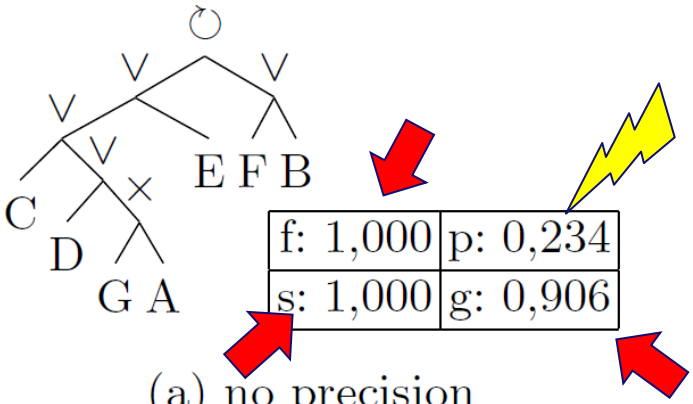
(b) Replay Fitness and Simplicity



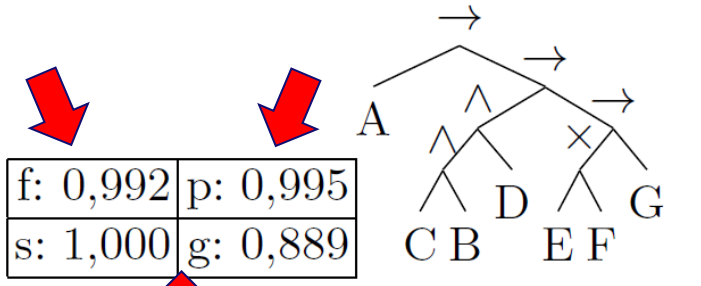
(c) Replay Fitness and Generalization

ETM with weight zero to two out of four perspectives.

Considering 3 of 4 Criteria



(a) no precision

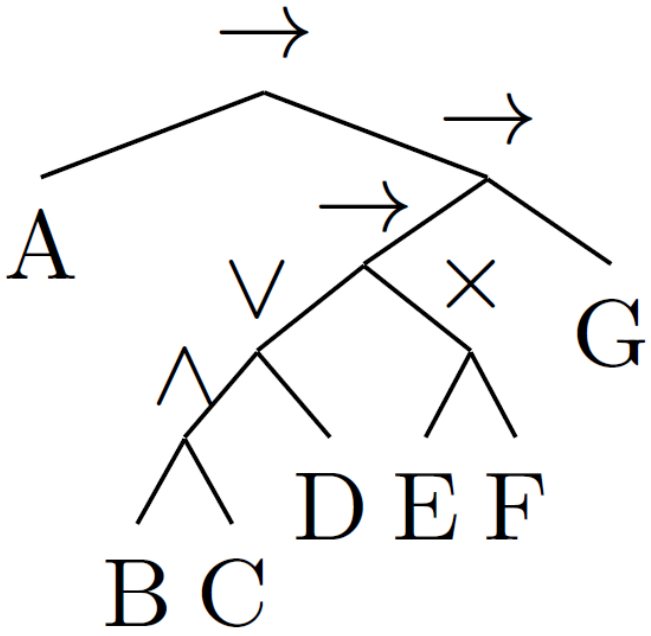


(b) no generalization or no simplicity

replay fitness needs to have a larger weight

Considering All Four Criteria with Emphasis on Fitness

f: 1,000	p: 0,923
s: 1,000	g: 0,889



fitness has weight 10

Initial Model Versus Discovered Model

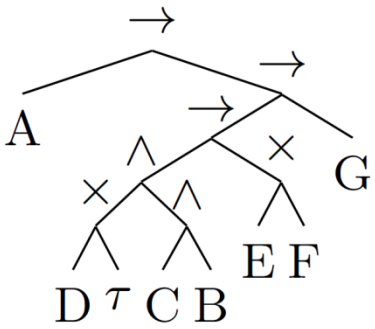
Better than existing algorithms (but patience is needed)!

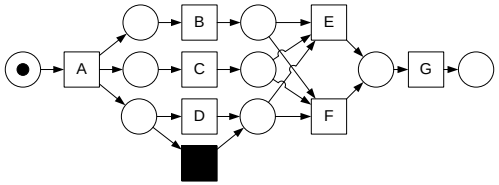
Trace	#
A B C D E G	6
A B C D F G	38
A B D C E G	12
A B D C F G	26
A B C F G	8
A C B E G	1
A D B C F G	1
A D B C E G	1
A D C B F G	4
A C D B F G	2
A C B F G	1

Discovered model outperforms initial model with respect too all criteria!

simulated

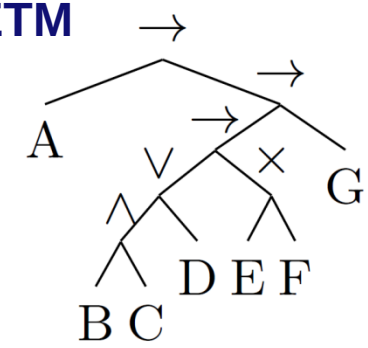
f: 1,000	p: 0,893
s: 0,933	g: 0,830





discovered by ETM

f: 1,000	p: 0,923
s: 1,000	g: 1,000



Real-Life Event Logs

- Event log L0 is the event log used before. L0 contains 100 traces, 590 events and 7 activities.
- Event Log L1 contains 105 traces, 743 events in total, with 6 different activities.
- Event Log L2 contains 444 traces, 3.269 events in total, with 6 different activities.
- Event Log L3 contains 274 traces, 1:582 events in total, with 6 different activities.

Event logs L1, L2 and L3 are extracted from the information systems of municipalities participating in the CoSeLoG project (http://www.win.tue.nl/coselog/).

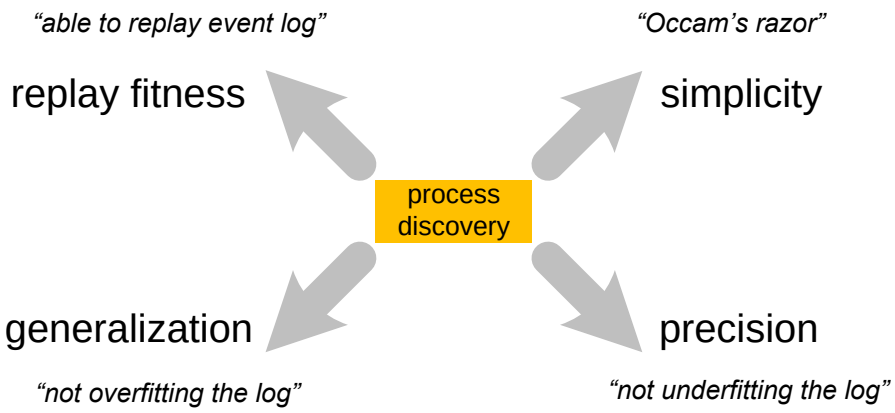
Results

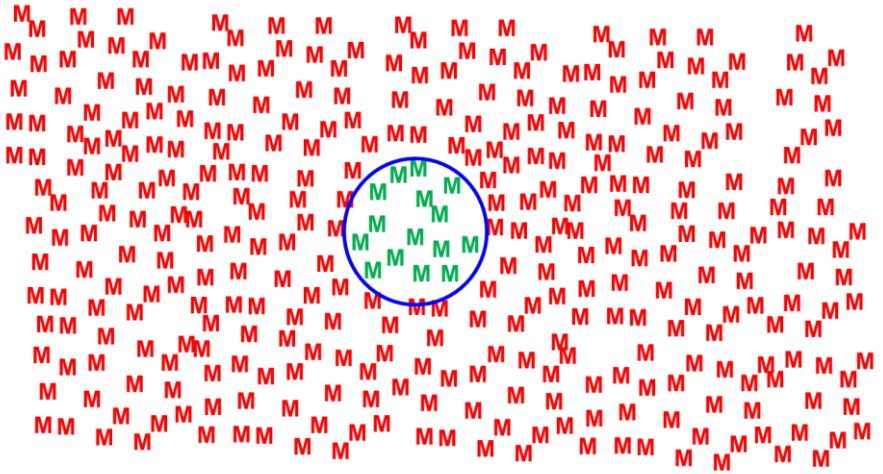
	L0		L1		L2		L3	
α-algorithm	f: 0,992	p: 0,995	f: 1,000	p: 0.510	f: 1.000	p: 0.468	f: 0.976	p: 0.532
	s: 1,000	g: 0,889	s: 0.923	g: 0.842	s: 0.923	g: 0.885	s: 0.923	g: 0.866
	overall: 0,969		overall: 0,819		overall: 0,819		overall: 0,824	
ILP Miner	f: 1,000	p: 0,748	f: 1.000	p: 0.551	f: 1.000	p: 0.752	f: 1.000	p: 0.479
	s: 0,933	g: 0,830	s: 0.857	g: 0.775	s: 0.923	g: 0.885	s: 0.857	g: 0.813
	overall: 0,887		overall: 0,796		overall: 0,890		overall: 0,787	
Heuristics	*f: 1,000*	*p: 0,986*	f: 0.966	p: 0.859	f: 0.917	p: 0.974	*f: 0.995*	*p: 1.000*
	s: 0,875	*g: 0,852*	s: 0.750	g: 0.746	s: 0.706	g: 0.716	*s: 1.000*	*g: 0.939*
	overall: 0,928		overall 0,830		overall: 0,828		***overall: 0,983***	
Genetic	*f: 1,000*	*p: 0,922*	*f: 0,997*	*p: 0,808*	*f: 0.905*	*p: 0.808*	*f: 0.987*	*p: 0.875*
	s: 0,737	*g: 0,790*	*s: 0,750*	*g: 0.707*	*s: 0,706*	*g: 0.717*	*s: 0.750*	*g: 0.591*
	overall: 0,862		*overall: 0,815*		*overall: 0,784*		*overall: 0,801*	
ETM	f: 0,992	p: 0,995	f: 0,901	p: 0,989	f: 0,863	p: 0,982	f: 0,995	p: 1,000
	s: 1,000	g: 0,889	s: 0,923	g: 0,894	s: 0,923	g: 0,947	s: 1,000	g: 0,939
	overall: 0,969		**overall: 0,927**		**overall: 0,929**		**overall: 0,983**	

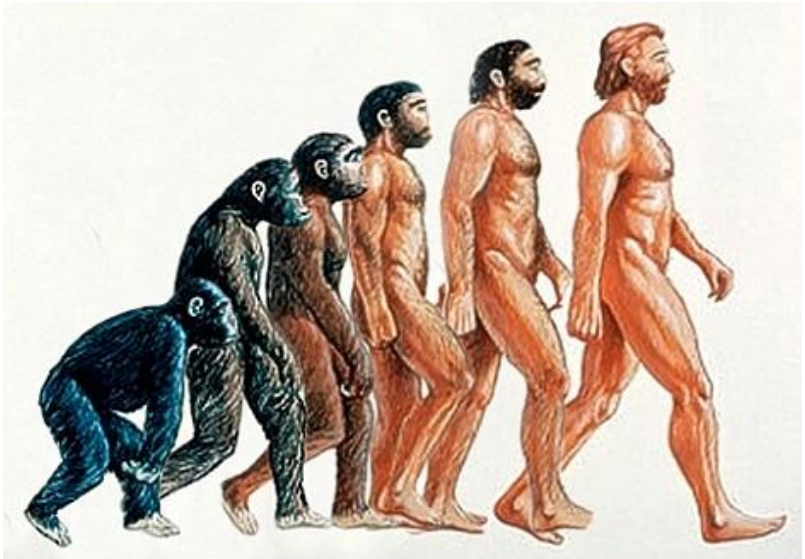
If unsound , the sound behavior is approximated when creating the process tree.

Equal weights for all criteria.

Conclusion







- First algorithm that allows for balancing all four perspectives.
- Genetic algorithm is very flexible, but also very slow.
- Process trees only used internally (choose your favorite representation)
- Future work:
  - Improve speed
  - Distribute PM tasks
  - Discover configurable process trees

and data mining necessary to understand the remainder of the book. Part II focuses on process discovery as the most important process mining task. Part III moves beyond discovering the control flow of processes and highlights conformance checking, and organizational and time perspectives. Part IV guides the reader in successfully applying process mining in practice, including an introduction to the widely used open-source tool ProM. Finally, Part V takes a step back, reflecting on the material presented and the key open challenges.

Overall, this book provides a comprehensive overview of the state of the art in process mining. It is intended for business process analysts, business consultants, process managers, graduate students, and BPM researchers.

Features and Benefits:

- First book on process mining, bridging the gap between business process modeling and business intelligence.
- Written by one of the most influential and most-cited computer scientists and the best-known BPM researcher.
- Self-contained and comprehensive overview for a broad audience in academia and industry.
- The reader can put process mining into practice immediately due to the applicability of the techniques and the availability of the open-source process mining software ProM.

Process Mining

Computer Science

springer.com

Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

www.processmining.org

www.win.tue.nl/ieeetfpm/