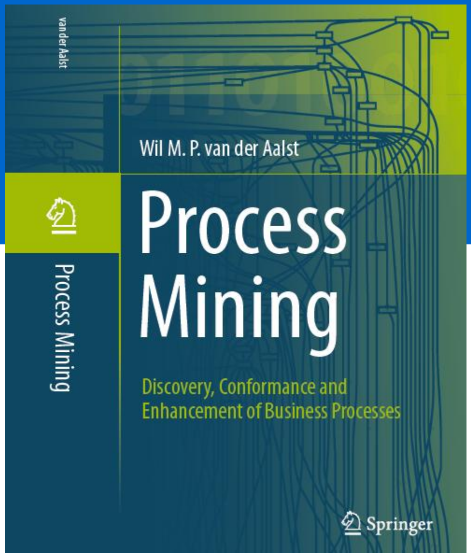
Third European Business Intelligence Summer School (eBISS 2013)

Process Mining: Making Sense of Processes Hidden in Big Event Data

prof.dr.ir. Wil van der Aalst
www.processmining.org

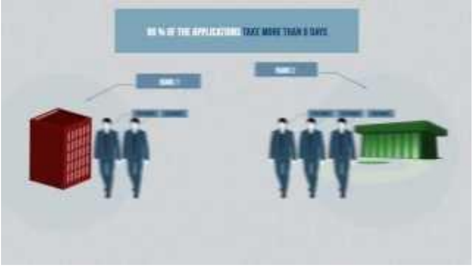


TU/e Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

Some Movies …

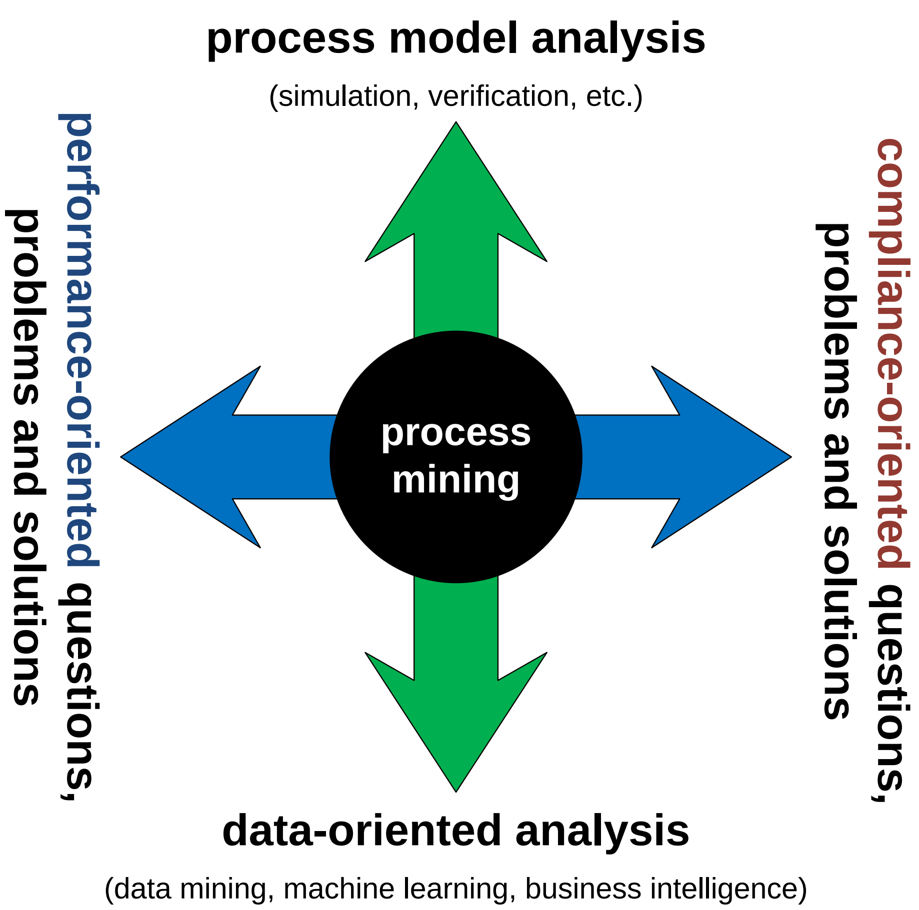
http://www.youtube.com/watch?v=mVvc6NUeoHo



http://www.youtube.com/watch?v=7oat7MatU_U

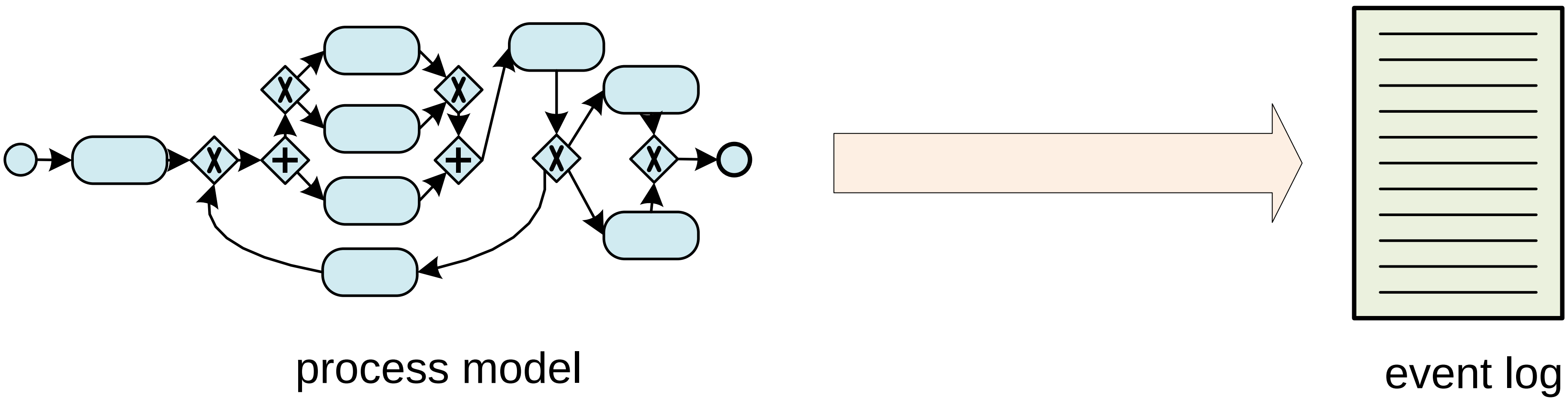
http://www.youtube.com/watch?v=nKy2Sx2WYRE

Positioning Process Mining

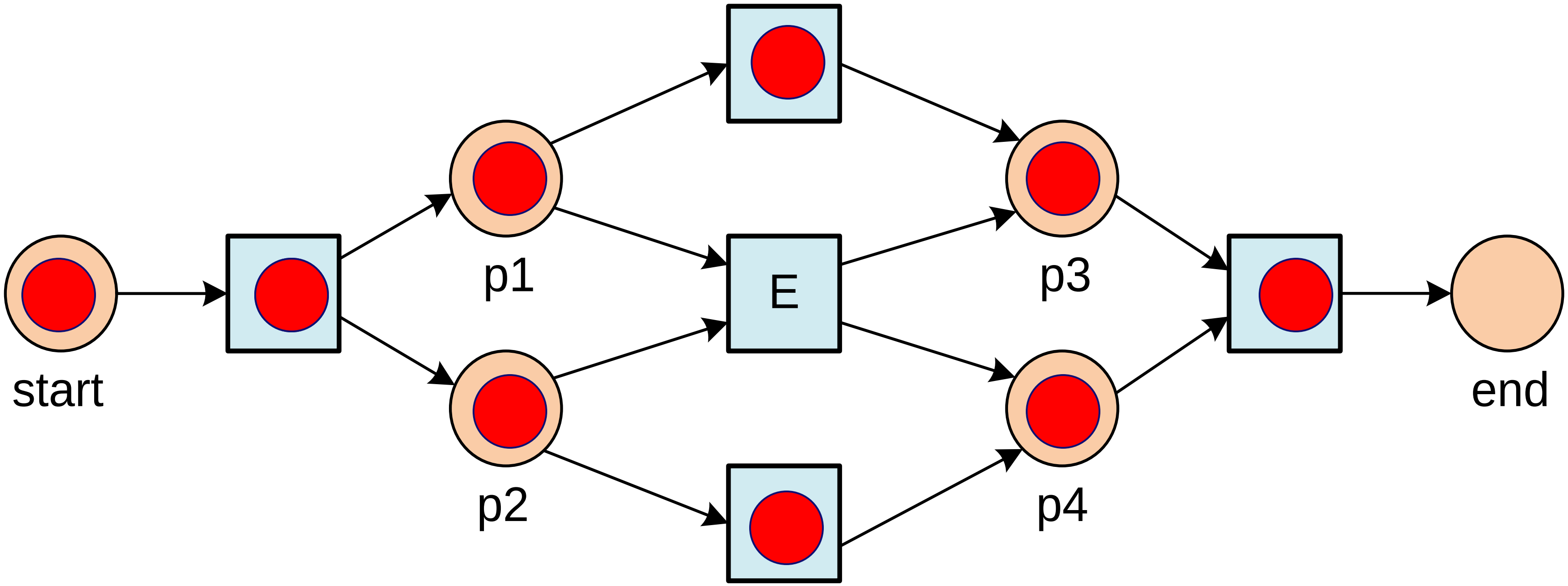


On the different roles of (process) models …

Play-Out



Play-Out (Classical use of models)



A B C D

A E D

A E D

A B C D

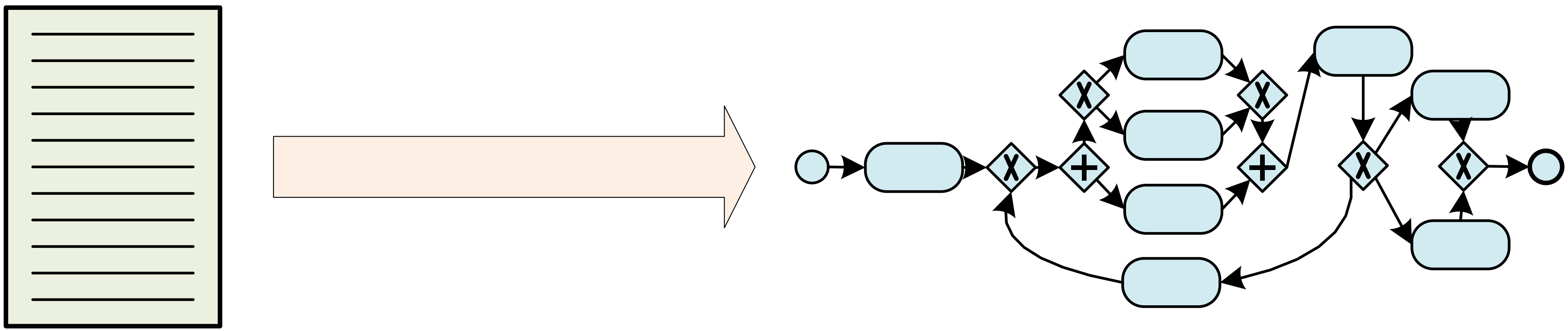
A C B D

A C B D

A E D

A C B D

Play-In



event log

process model

Play-In

A B C D

A E D

A E D

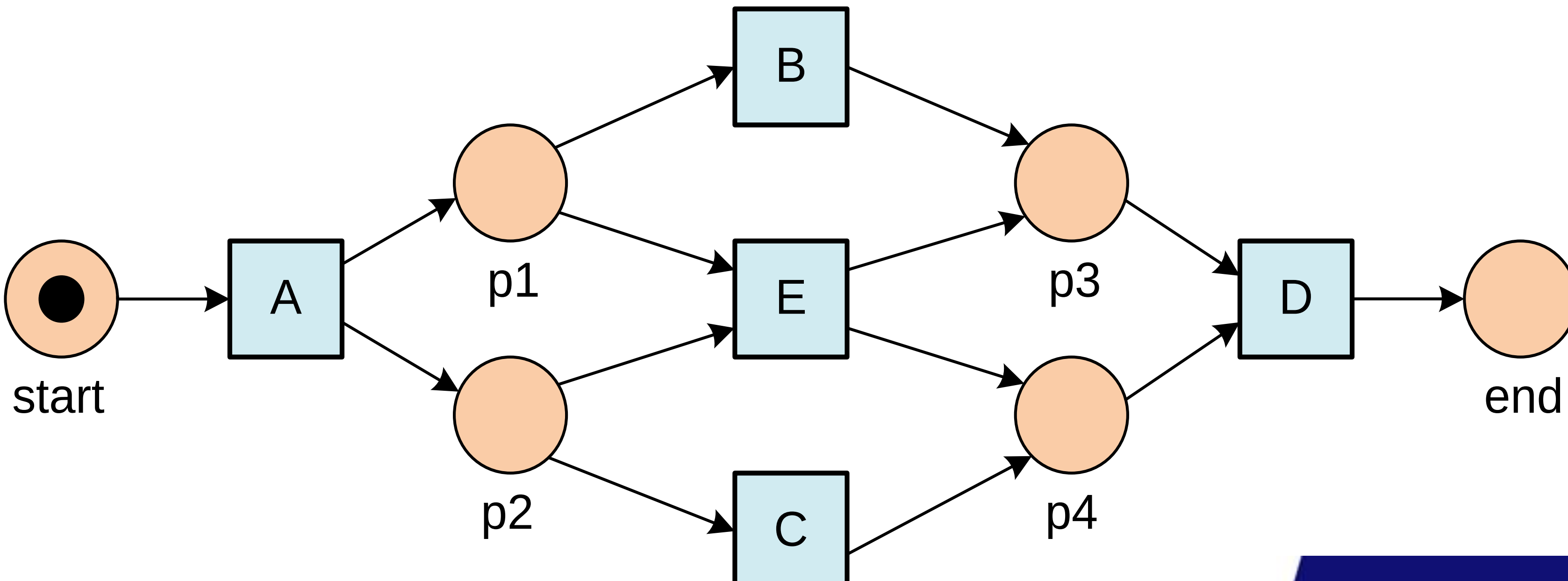
A B C D

A C B D

A C B D

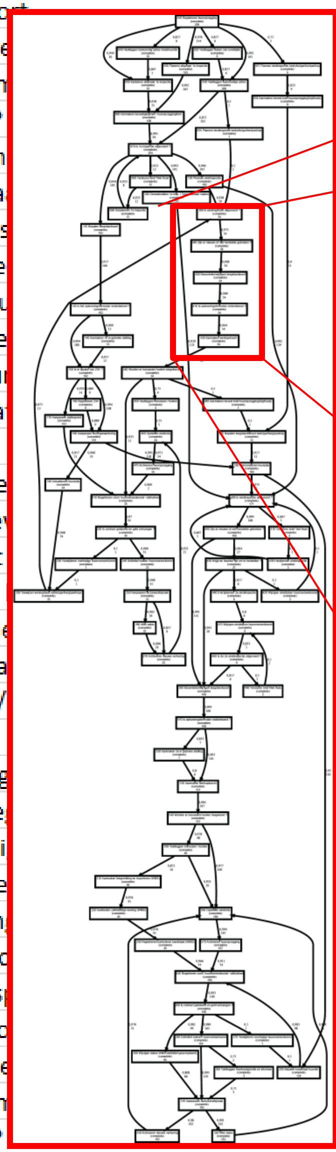
A E D

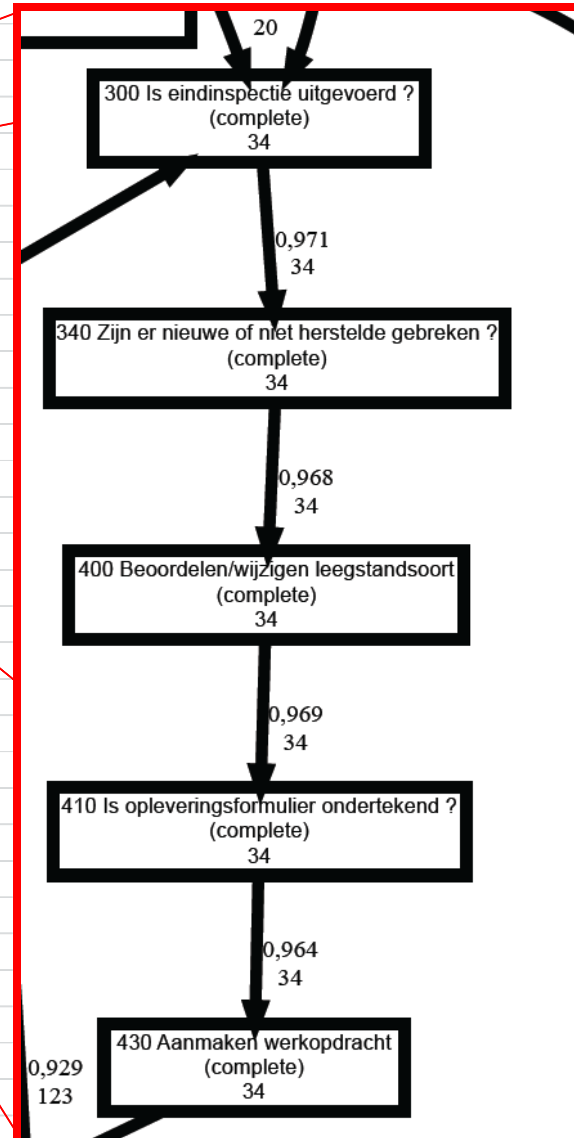
A C B D



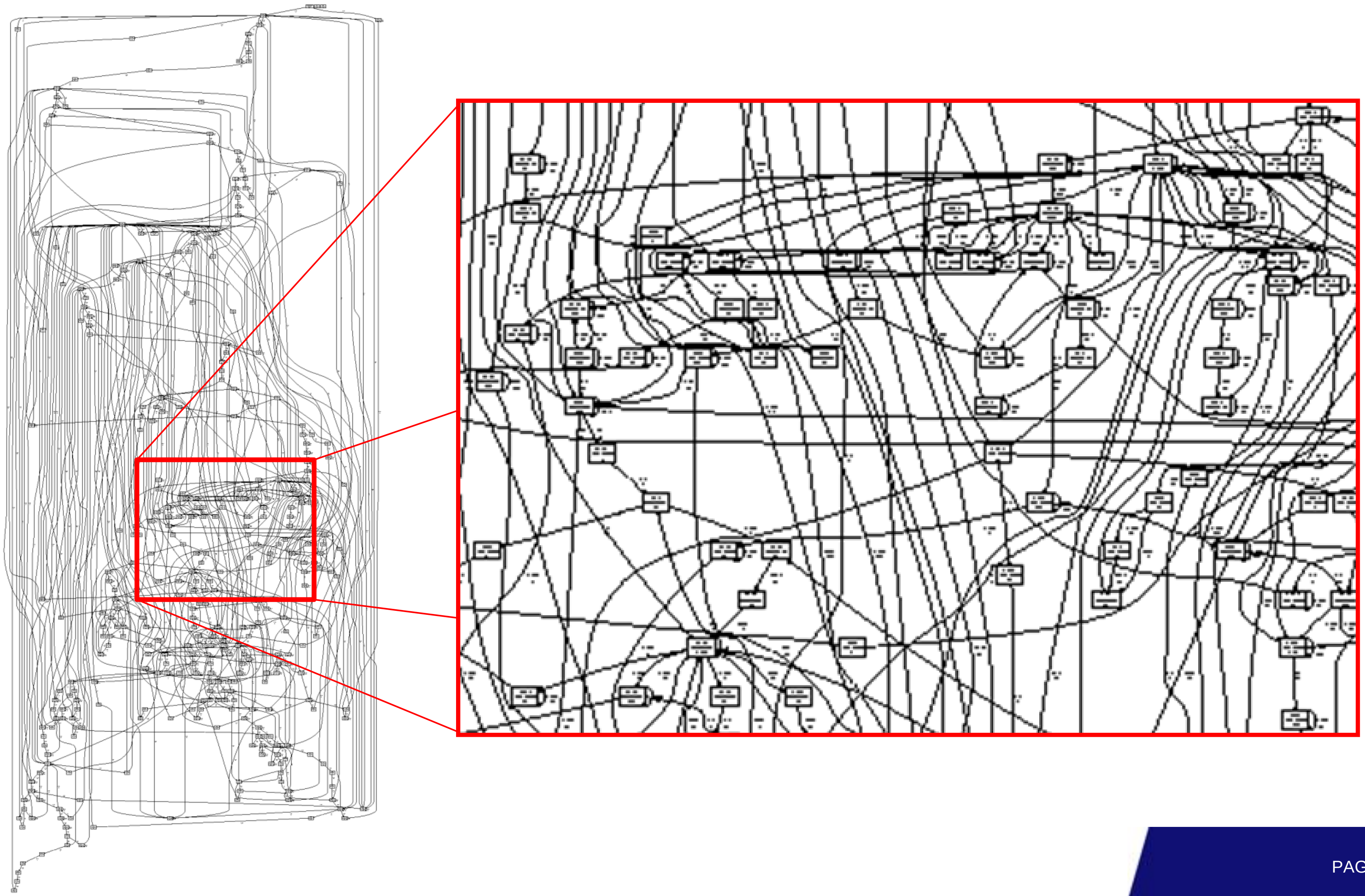
Example Process Discovery
(Vestia, Dutch housing agency, 208 cases, 5987 events)

117315	110 Bepalen leegstandsoort	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 uitge	07.06.2007 14:47:04
117315	340 Zijn er nieuwe of niet	07.06.2007 14:47:06
117315	400 Beoordelen/wijzigen	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 huuropzeg	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



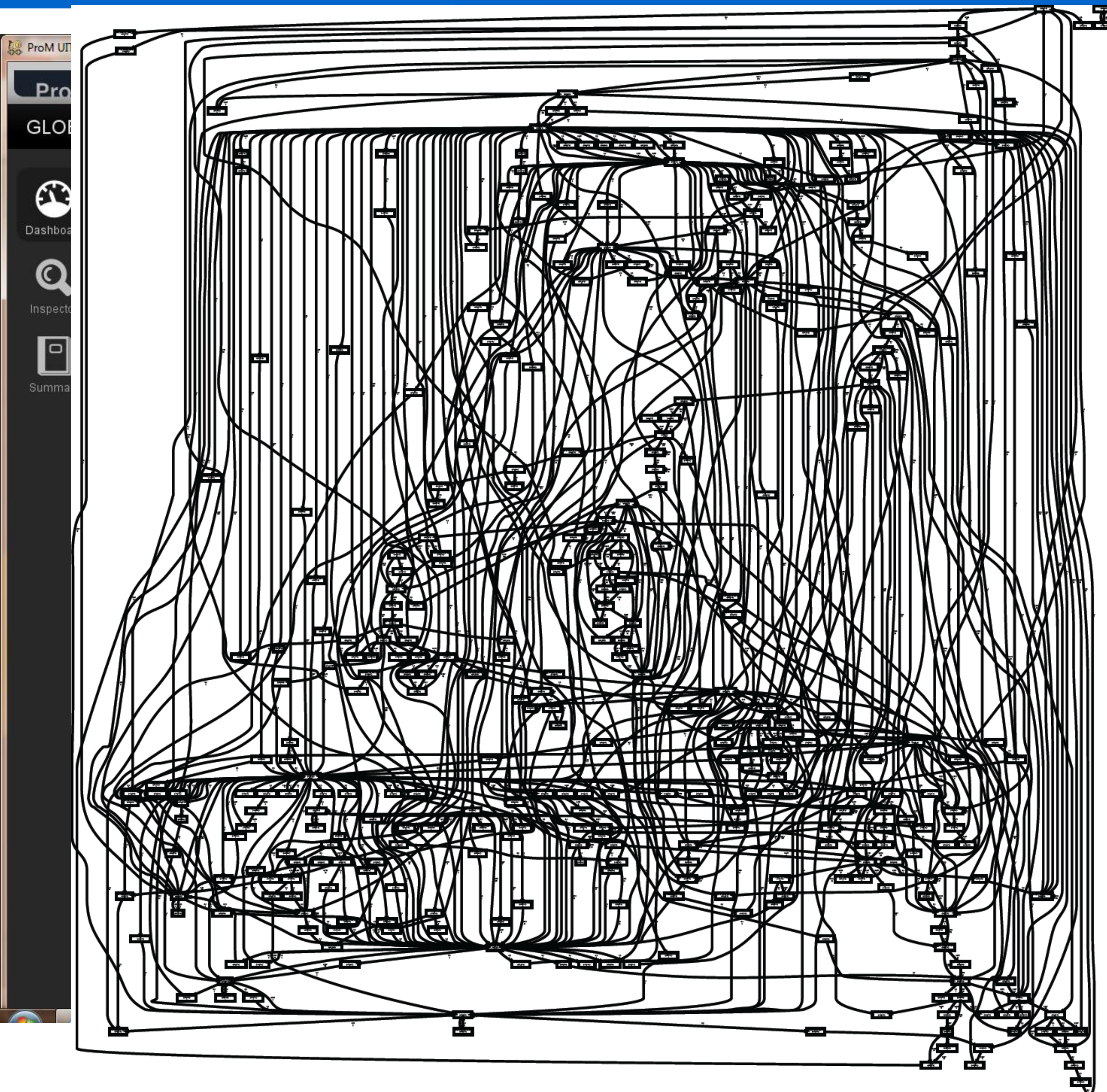


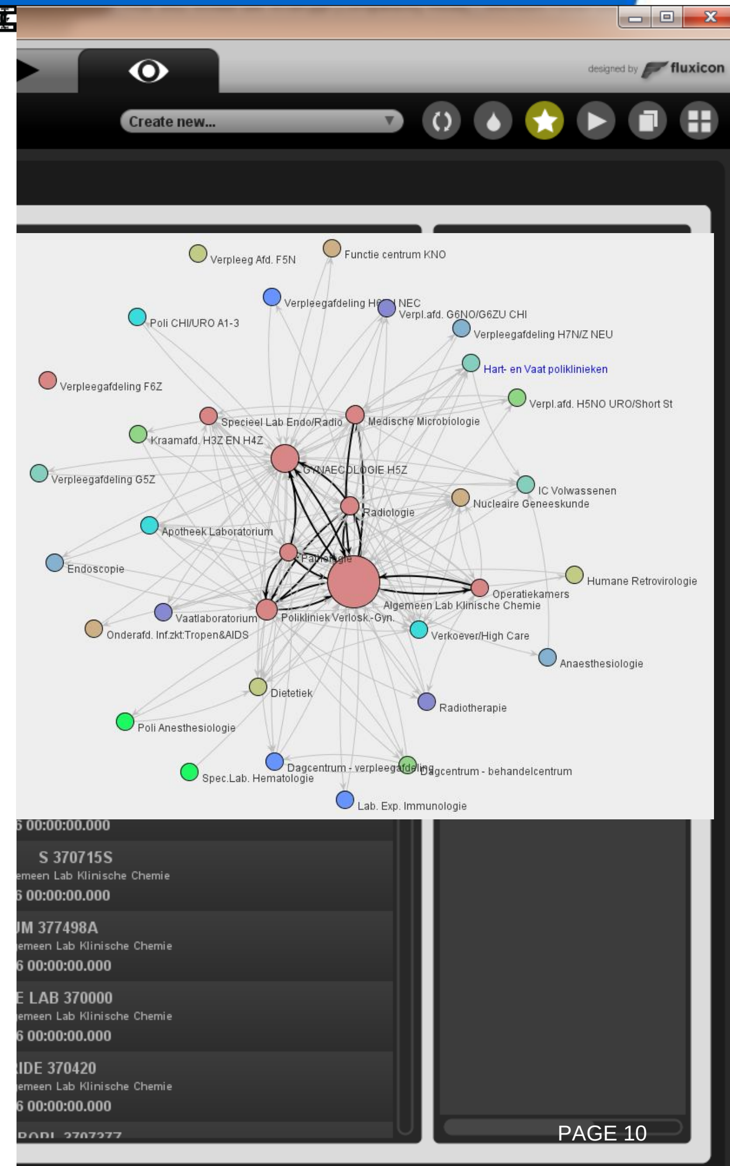
Example Process Discovery
(ASML, test process lithography systems, 154966 events)



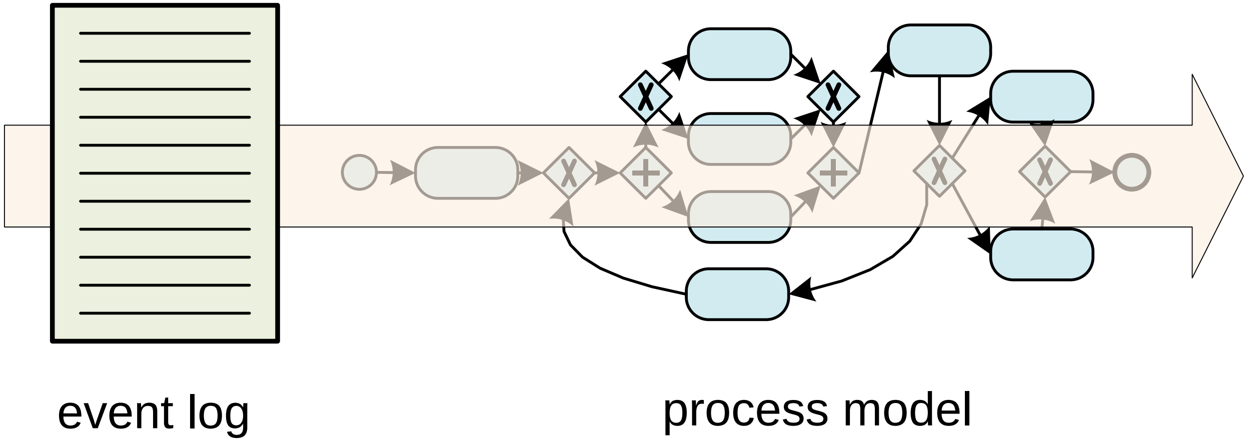
Example Process Discovery

(AMC, 627 gynecological oncology patients, 24331 events)





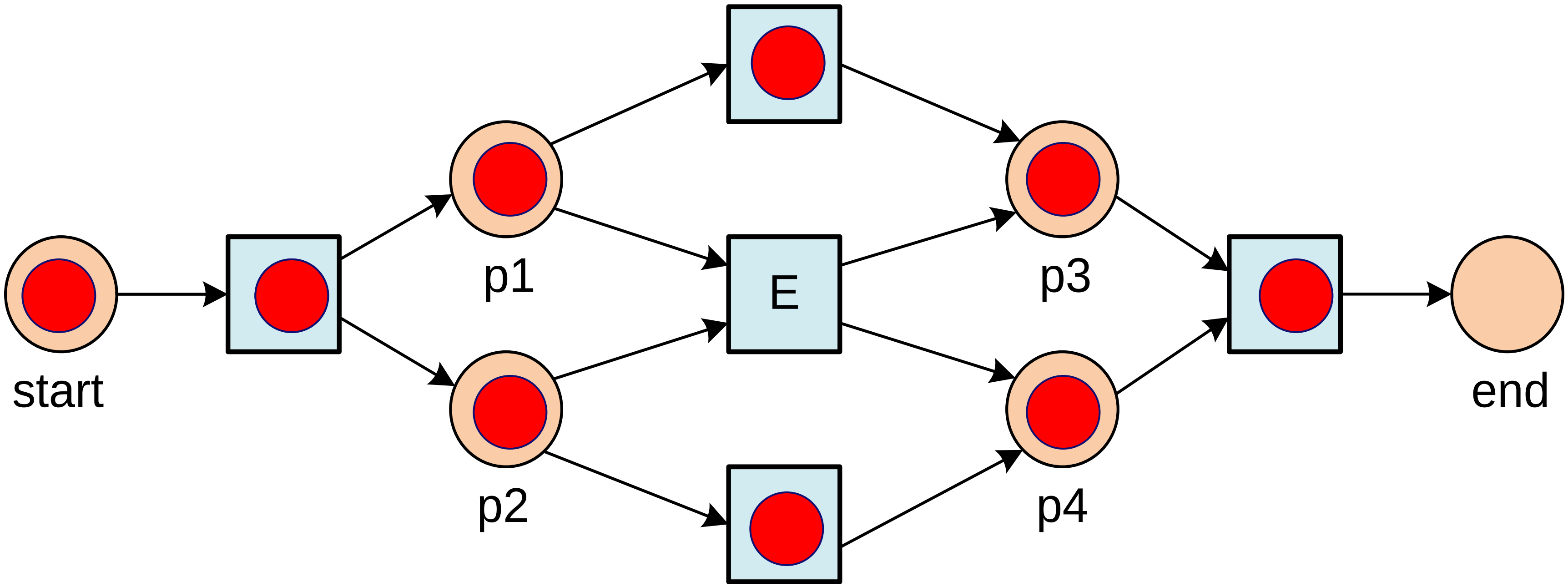
Replay



- extended model showing times, frequencies, etc.
- diagnostics
- predictions
- recommendations

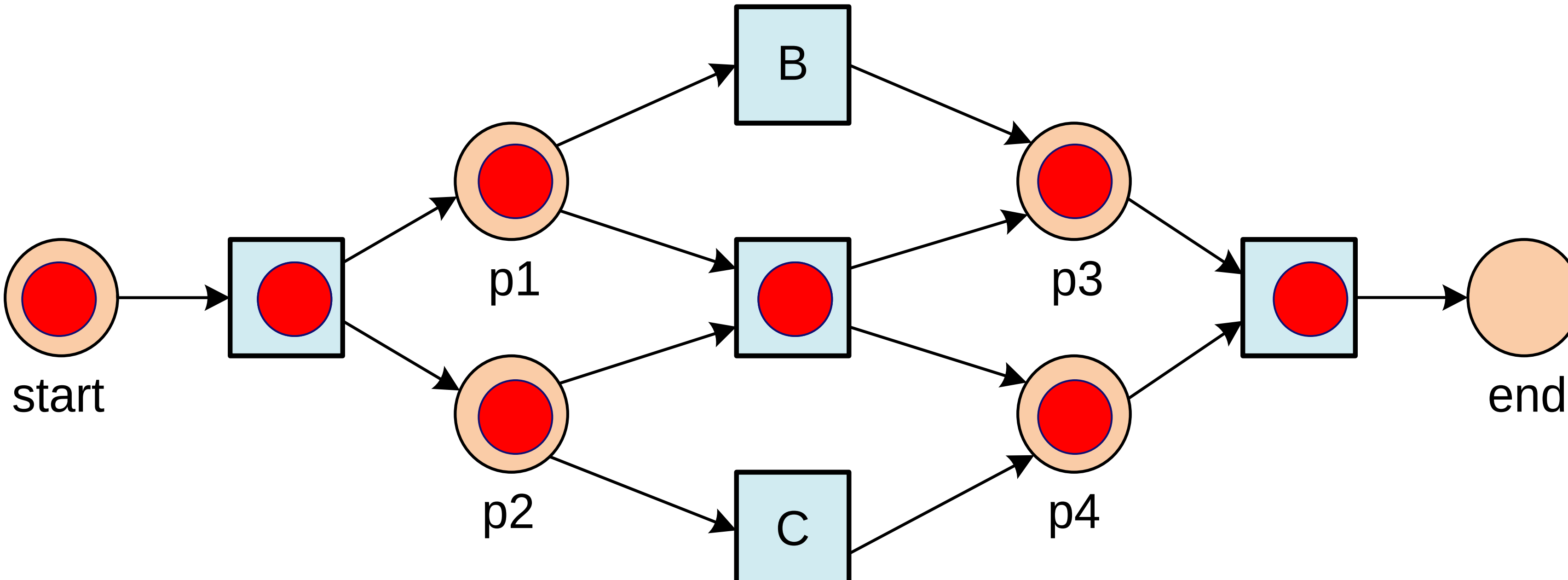
Replay

A B C D



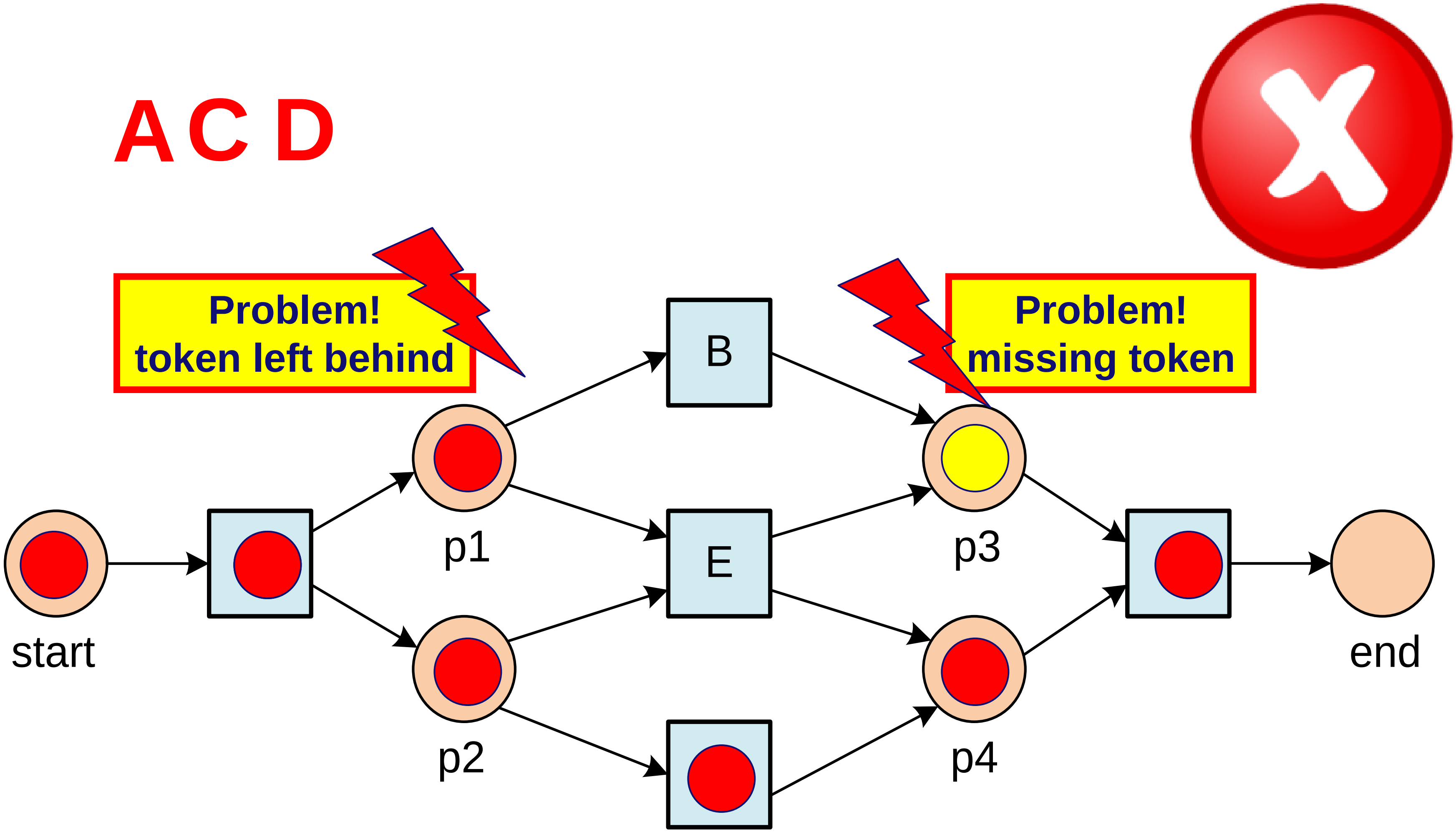
Replay

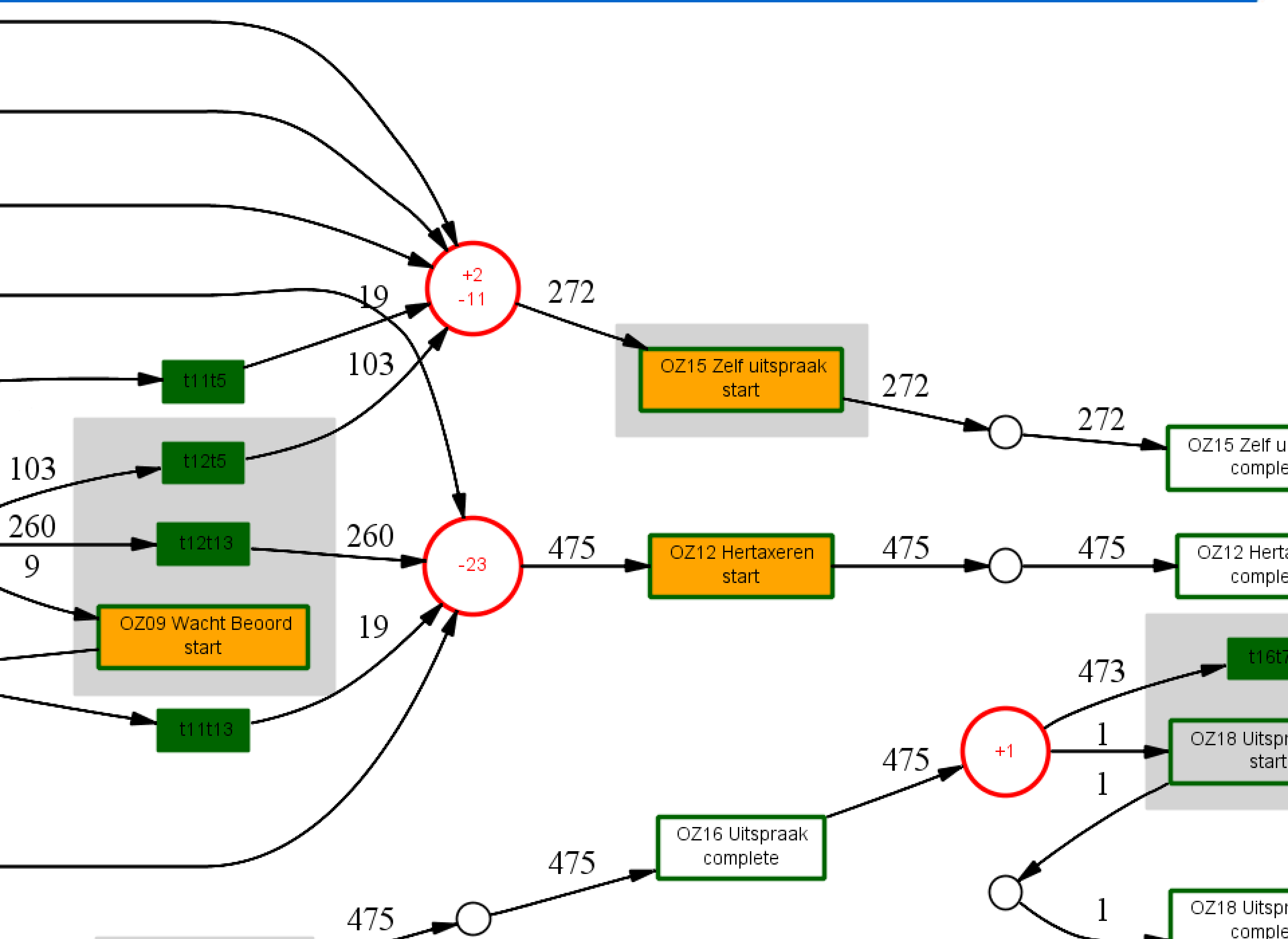
A E D



Replay can detect problems

A C D



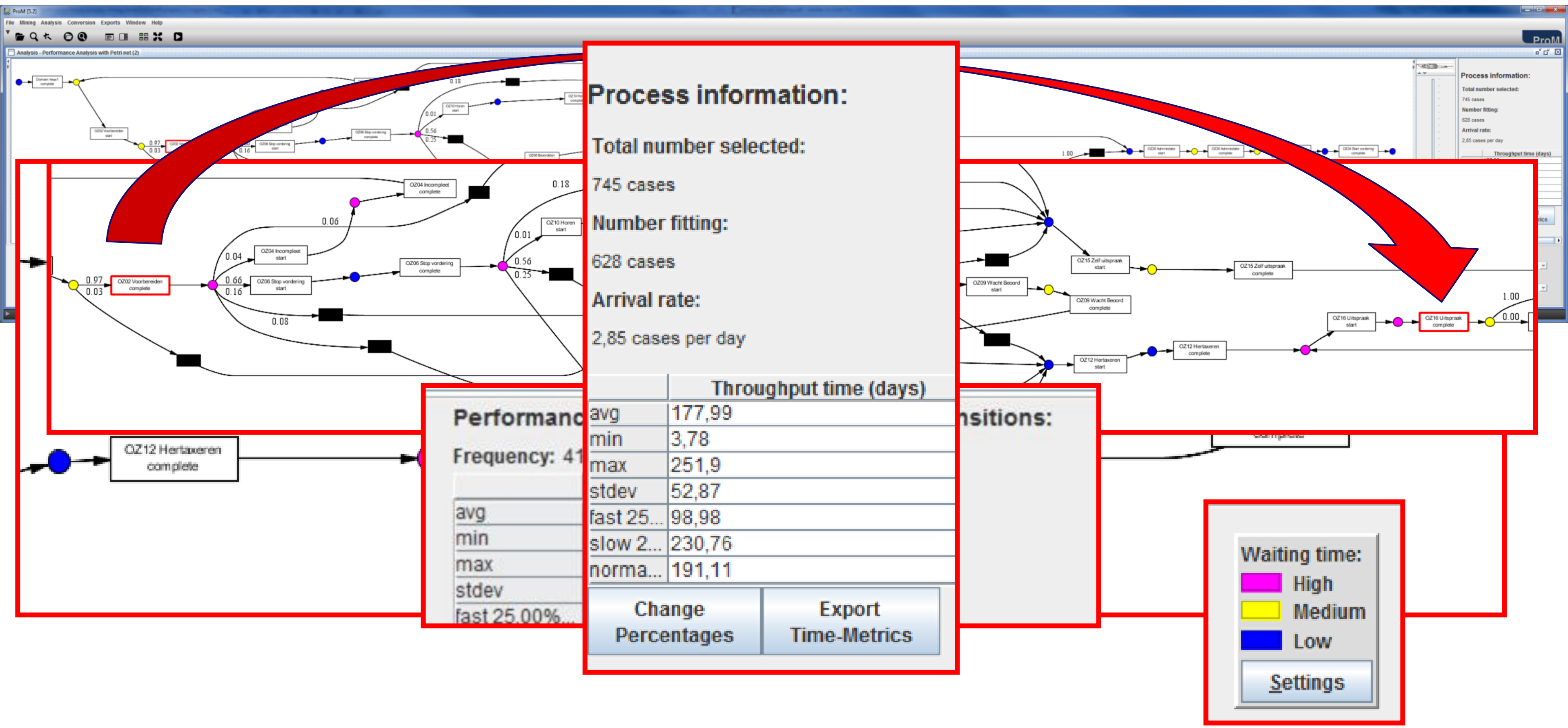


+2
-11
272
19
103
t1t5
OZ15 Zelf uitspraak start
272
272
OZ15 Zelf u
comple
103
t12t5
260
t12t13
260
9
-23
475
OZ12 Hertaxeren start
475
475
OZ12 Herta
comple
OZ09 Wacht Beoord start
19
t11t13
t16t7
473
+1
1
OZ18 Uitspr
start
475
1
OZ16 Uitspraak complete
475
475
1
OZ18 Uitspr
comple

Replay can extract timing information

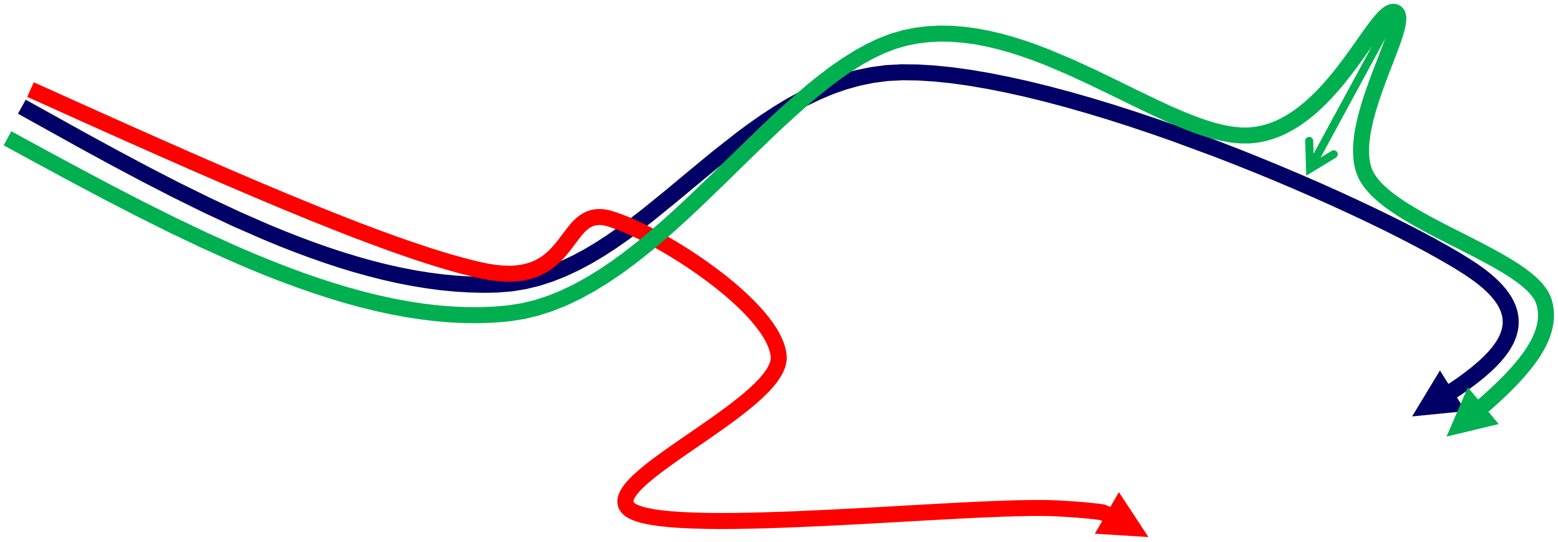
Performance Analysis Using Replay

(WOZ objections Dutch municipality, 745 objections, 9583 event, f= 0.988)



Models are like the glasses required to see and understand event data!

Alignments are essential!



- **conformance checking to diagnose deviations**
- **squeezing reality into the model to do model-based analysis**

a	c	$\gg$	d	$\gg$	f	$\gg$
a	c	b	d	τ	$\gg$	h
$t1$	$t4$	$t3$	$t5$	$t7$		$t10$

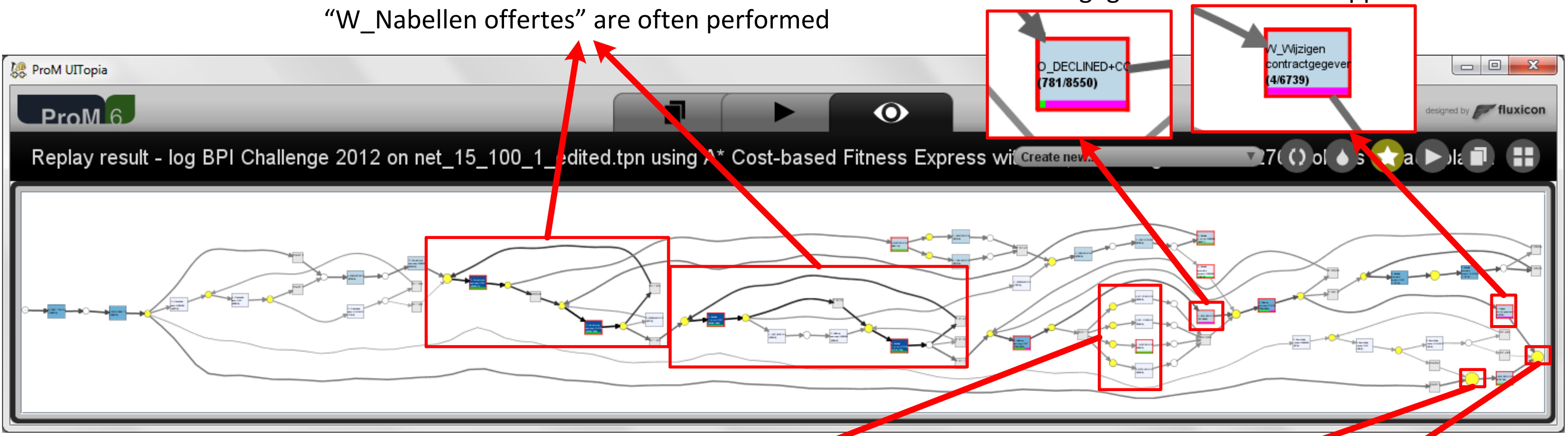
a	c	$\gg$	d	$\gg$	f	$\gg$
a $t1$	c $t4$	b $t3$	d $t5$	τ $t7$	$\gg$	h $t10$

Example: BPI Challenge 2012

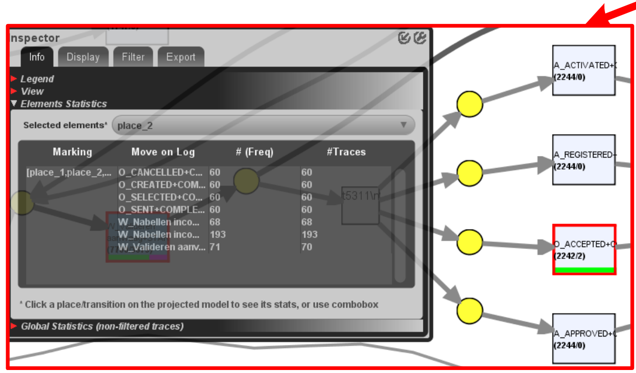
(Dutch financial institute, doi:10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f)

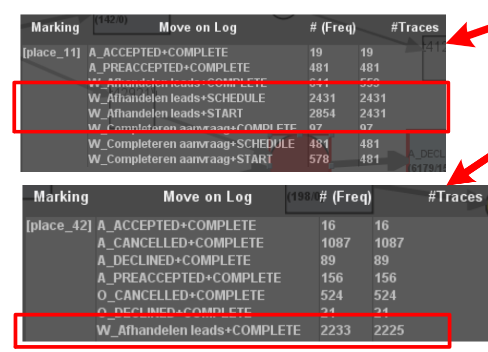
Loops of "W_Completeren aanvraag" and "W_Nabellen offertes" are often performed

"O_DECLINED" and "W_Wijzigen contractgegevens" are often skipped



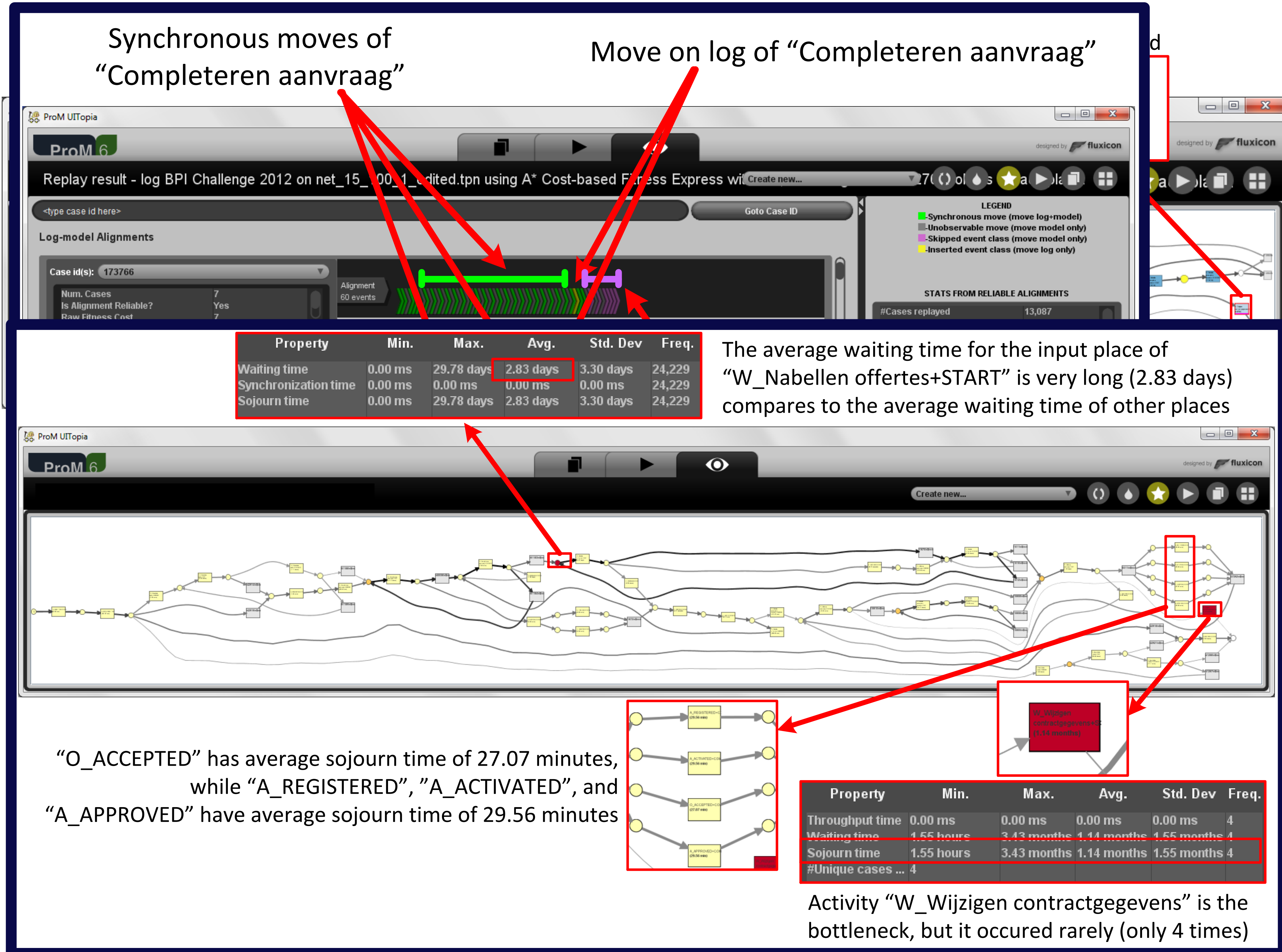
Many moves on log of "O_CANCELLED", "O_CREATED", "O_SELECTED", "O_SENT" occurred with the same frequency value (i.e. 60) before parallel branch





Many moves on log of "W_Afhandelen leads" (> 2200 times) occurred in the end of traces

Work of Arya Adriansyah (Replay project)



Property	Min.	Max.	Avg.	Std. Dev	Freq.
Waiting time	0.00 ms	29.78 days	2.83 days	3.30 days	24,229
Synchronization time	0.00 ms	0.00 ms	0.00 ms	0.00 ms	24,229
Sojourn time	0.00 ms	29.78 days	2.83 days	3.30 days	24,229

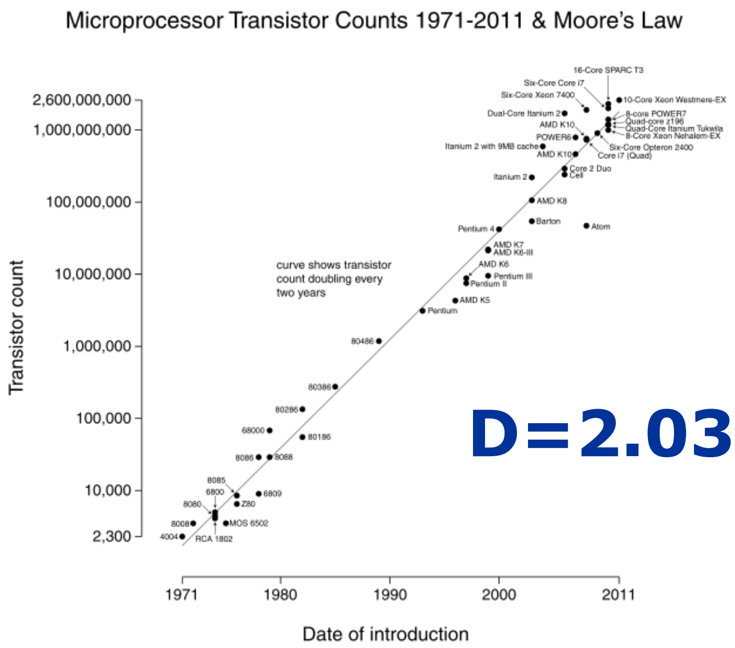
Property	Min.	Max.	Avg.	Std. Dev	Freq.
Throughput time	0.00 ms	0.00 ms	0.00 ms	0.00 ms	4
Waiting time	1.55 hours	3.43 months	1.14 months	1.55 months	4
Sojourn time	1.55 hours	3.43 months	1.14 months	1.55 months	4
#Unique cases ...	4				

Desire Lines in Big Data

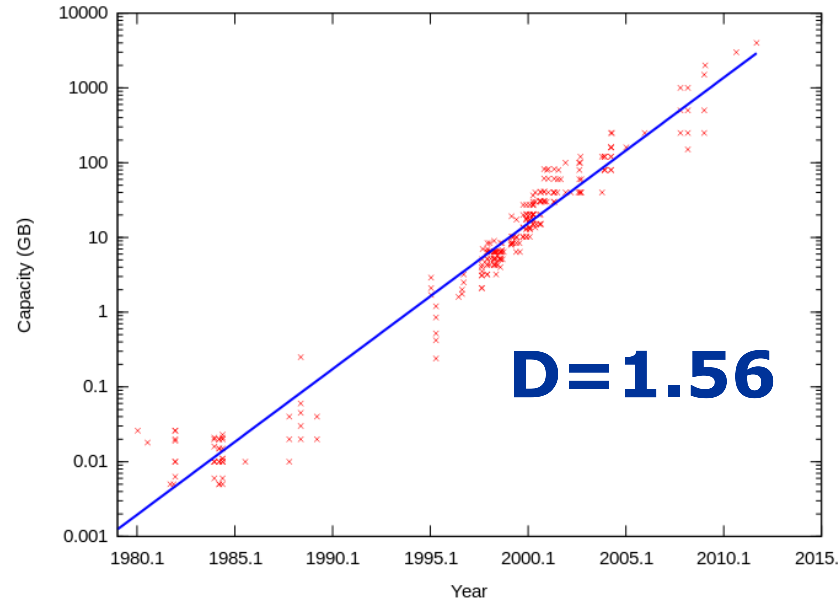

www.olifantenpaadjes.nl
0100110011010101010
010011010101010
60


0100110011010101010

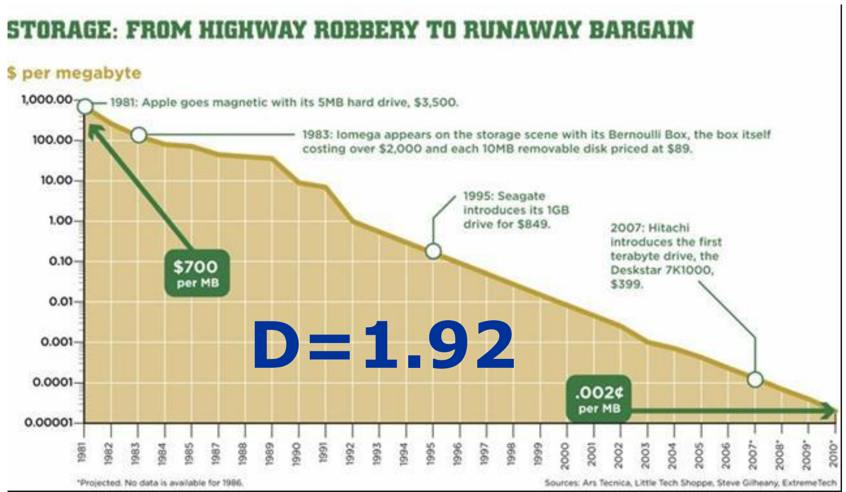
Moore's Law



Microprocessor Transistor Counts 1971-2011 & Moore's Law
Transistor count
2,600,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
2,300
curve shows transistor count doubling every two years
1971
1980
1990
2000
2011
Date of introduction
D=2.03



Capacity (GB)
10000
1000
100
10
1
0.1
0.01
0.001
1980.1
1985.1
1990.1
1995.1
2000.1
2005.1
2010.1
Year
D=1.56

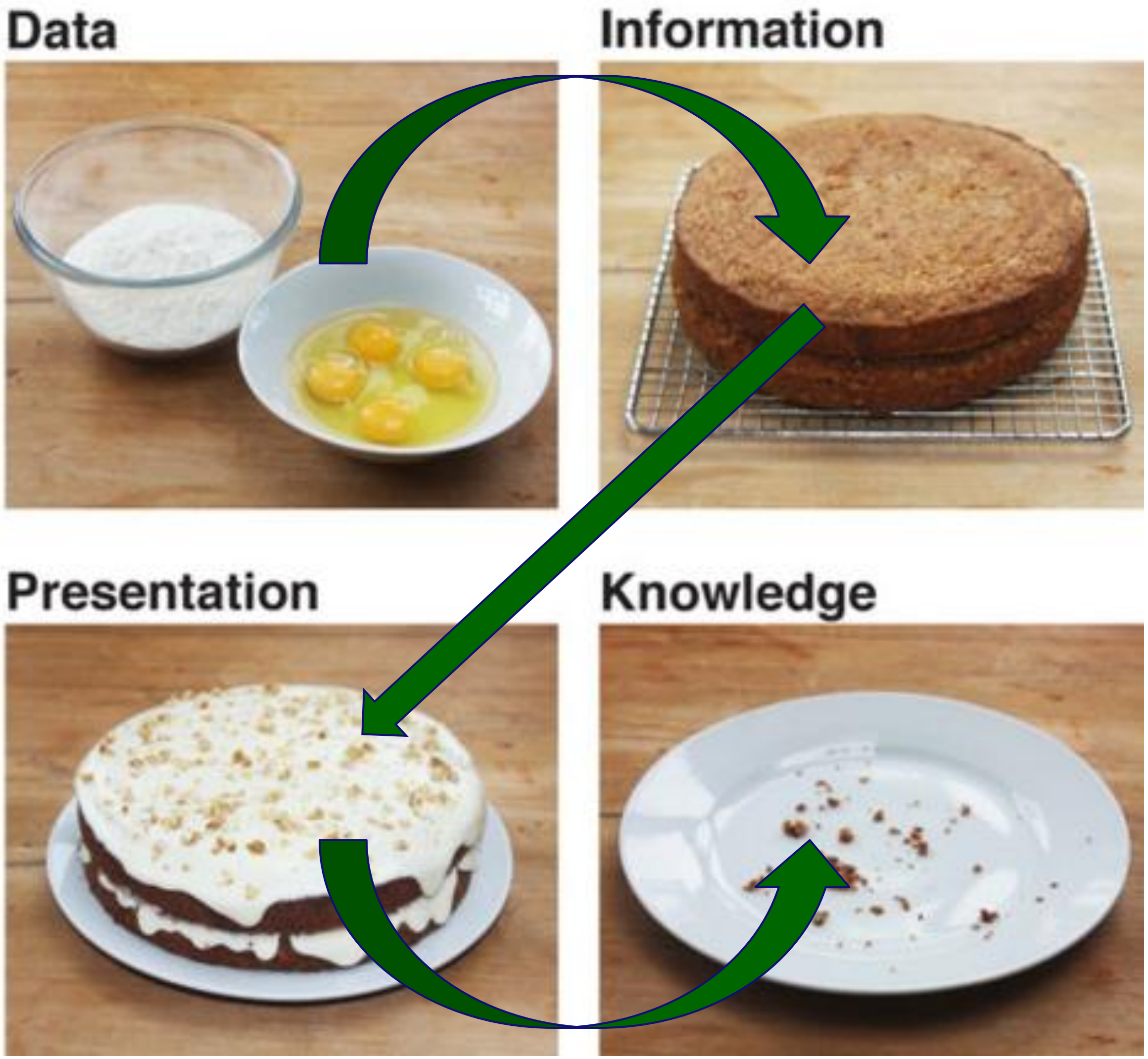


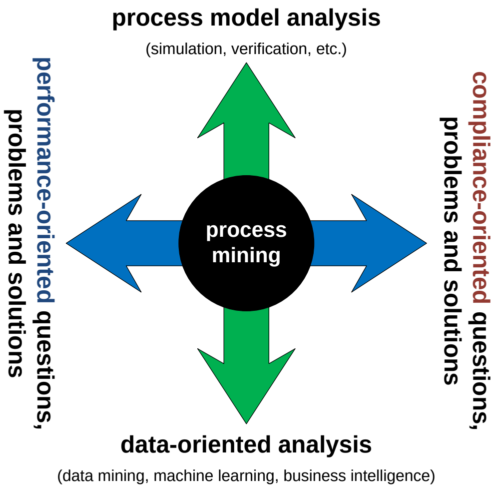
STORAGE: FROM HIGHWAY ROBBERY TO RUNAWAY BARGAIN
$ per megabyte
1981: Apple goes magnetic with its 5MB hard drive, $3,500.
1983: Iomega appears on the storage scene with its Bernoulli Box, the box itself costing over $2,000 and each 10MB removable disk priced at $89.
1995: Seagate introduces its 1GB drive for $849.
2007: Hitachi introduces the first terabyte drive, the Deskstar 7K1000, $399.
$700 per MB
.002¢ per MB
D=1.92

A simple calculation

- Starting point 2010:
  - Harddisk 1 Terabyte = 10^{12} bytes
  - Digital Universe 1.2 Zettabyte = $1.2*10^{21}$ bytes (estimate in IDC's annual report, "The Digital Universe Decade – Are You Ready?" May 2010)
- Disk needs to grow $2^{30.16} = 1.2* 10^9 = 1.2*10^{21}/ 10^{12}$ times its current size.
- Assuming D=1.56 this takes 30.16*1.56 = 47.05 years.
- Hence, in 2060 your laptop can contain all of today's digital universe (internet, computer files, transaction logs, movies, photos, music, books, databases, etc.)!

From Data to Actionable Knowledge

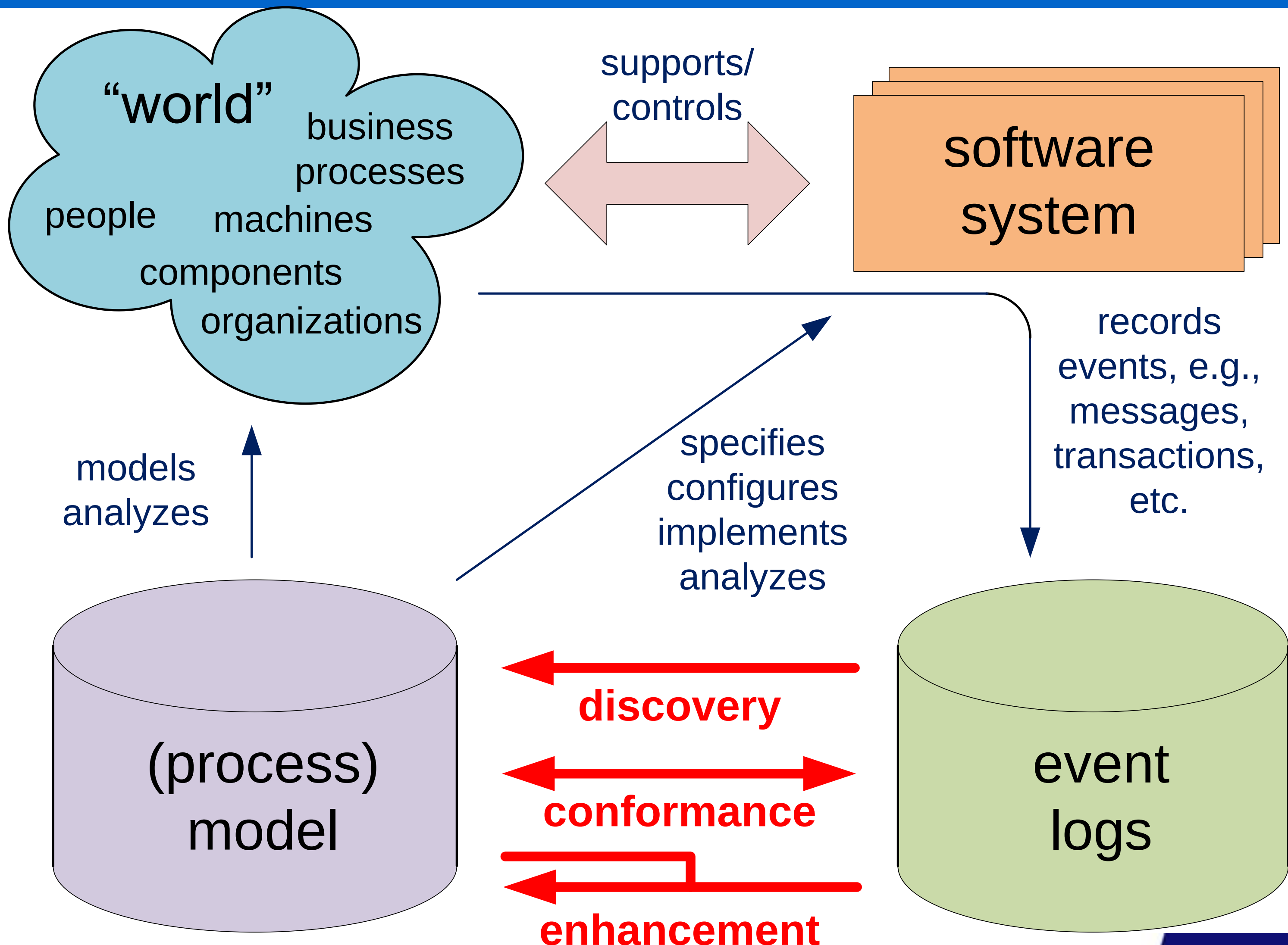




Process Mining

- Process discovery: "What is really happening?"
- Conformance checking: "Do we do what was agreed upon?"
- Performance analysis: "Where are the bottlenecks?"
- Process prediction: "Will this case be late?"
- Process improvement: "How to redesign this process?"
- Etc.

Process Mining



Starting point: event log

case id	event id	properties				
		timestamp	activity	resource	cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket			
	35654487	30-12-2010:14.16	examine casually			
	35654488	05-01-2011:11.22	decide			
	35654489	08-01-2011:12.05	pay compensation			
3	35654521	30-12-2010:14.32	register request			
	35654522	30-12-2010:15.06	examine casually			
	35654524	30-12-2010:16.34	check ticket			
	35654525	06-01-2011:09.18	decide			
	35654526	06-01-2011:12.18	reinitiate request			
	35654527	06-01-2011:13.06	examine thoroughly			
	35654530	08-01-2011:11.43	check ticket			
	35654531	09-01-2011:09.55	decide			
	35654533	15-01-2011:10.45	pay compensation			
4	35654641	06-01-2011:15.02	register request			
	35654643	07-01-2011:12.06	check ticket			
	35654644	08-01-2011:14.43	examine thoroughly			
	35654645	09-01-2011:12.02	decide			
	35654647	12-01-2011:15.44	reject request			
5	35654711	06-01-2011:09.02	register request			
	35654712	07-01-2011:10.16	examine casually			
	35654714	08-01-2011:11.22	check ticket			
	35654715	10-01-2011:13.28	decide			
	35654716	11-01-2011:16.18	reinitiate request			
	35654718	14-01-2011:14.33	check ticket			
	35654719	16-01-2011:15.50	examine casually			
	35654720	19-01-2011:11.18	decide	Sara	200	...
	35654721	20-01-2011:12.48	reinitiate request	Sara	200	...
	35654722	21-01-2011:09.06	examine casually	Sue	400	...
	35654724	21-01-2011:11.34	check ticket	Pete	100	...
	35654725	23-01-2011:13.12	decide	Sara	200	...
	35654726	24-01-2011:14.56	reject request	Mike	200	...
6	35654871	06-01-2011:15.02	register request	Mike	50	...
	35654873	06-01-2011:16.06	examine casually	Ellen	400	...
	35654874	07-01-2011:16.22	check ticket	Mike	100	...
	35654875	07-01-2011:16.52	decide	Sara	200	...
	35654877	16-01-2011:11.47	pay compensation	Mike	200	...
...	...	...	...	...	...	...

case id	event id	properties				
		timestamp	activity	resource	cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...

XES, MXML, SA-MXML, CSV, etc.

Simplified event log

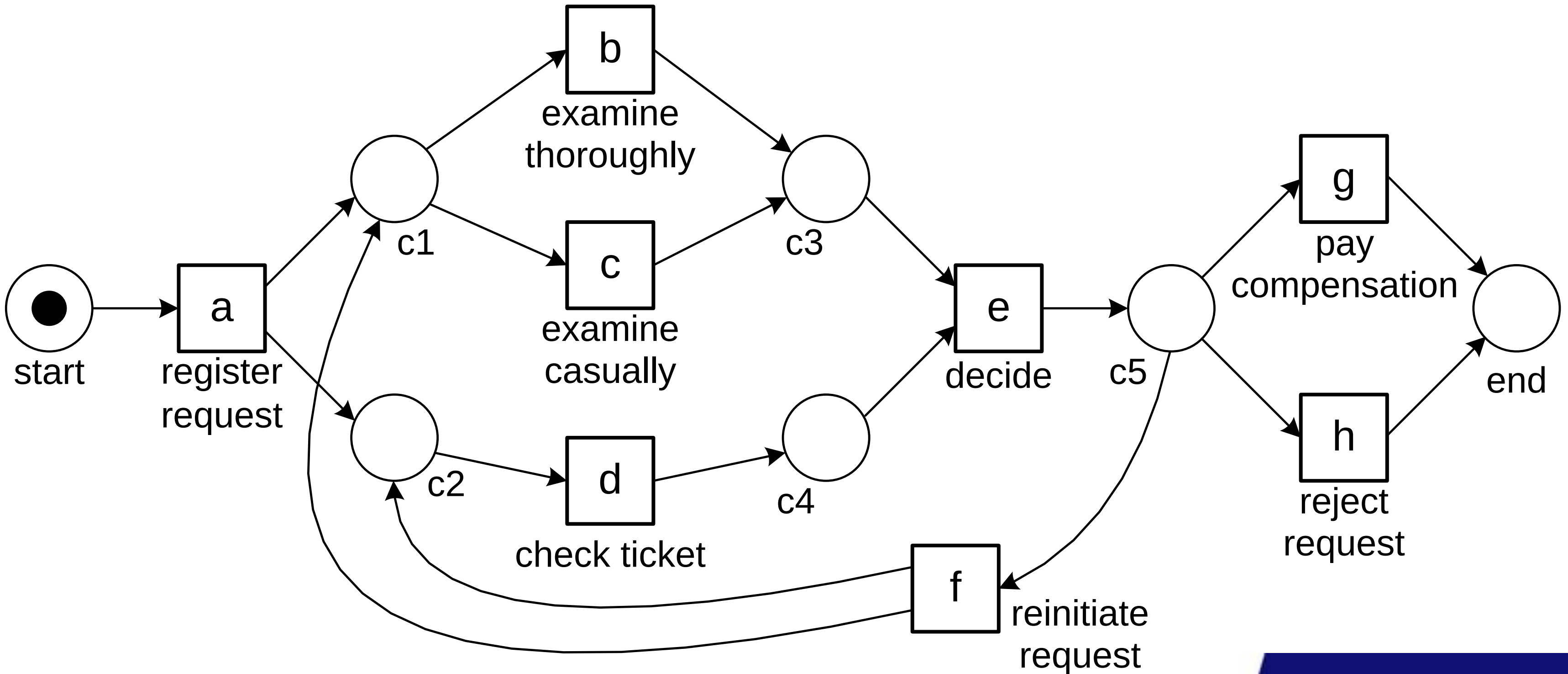
case id	event id	timestamp	activity	resource		
1	35654423	30-12-2010:11.02	register request	Pete		
	35654424	31-12-2010:10.06	examine thoroughly	Sue		
	35654425	05-01-2011:15.12	check ticket	Mike		
	35654426	06-01-2011:11.18	decide	Sara		
	35654427	07-01-2011:14.24	reject request	Pete		
2	35654483	30-12-2010:11.32	register request	Mike		
	35654485	30-12-2010:12.12	check ticket	Mike		
	35654487	30-12-2010:14.16	examine casually	Pete		
	35654488	05-01-2011:11.22	decide	Sara		
	35654489	08-01-2011:12.05	pay compensation	Ellen		
3	35654521	30-12-2010:14.32	register request	Pete		
	35654522	30-12-2010:15.06	examine casually	Mike		
	35654524	30-12-2010:16.34	check ticket	Ellen		
	35654525	06-01-2011:09.18	decide	Sara		
	35654526	06-01-2011:12.18	reinitiate request	Sara		
	35654527	06-01-2011:13.06	examine thoroughly	Sean		
	35654530	08-01-2011:11.43	check ticket	Pete		
	35654531	09-01-2011:09.55	decide	Sara		
	35654533	15-01-2011:10.45	pay compensation	Ellen		
4	35654641	06-01-2011:15.02	register request	Pete		...
	35654643	07-01-2011:12.06	check ticket	Mike	100	...
	35654644	08-01-2011:14.43	examine thoroughly	Sean	400	...
	35654645	09-01-2011:12.02	decide	Sara	200	...
	35654647	12-01-2011:15.44	reject request	Ellen	200	...
5	35654711	06-01-2011:09.02	register request	Ellen	50	...
	35654712	07-01-2011:10.16	examine casually	Mike	400	...
	35654714	08-01-2011:11.22	check ticket	Pete	100	...
	35654715	10-01-2011:13.28	decide	Sara	200	...
	35654716	11-01-2011:16.18	reinitiate request	Sara	200	...
	35654718	14-01-2011:14.33	check ticket	Ellen	100	...
	35654719	16-01-2011:15.50	examine casually	Mike	400	...
	35654720	19-01-2011:11.18	decide	Sara	200	...
	35654721	20-01-2011:12.48	reinitiate request	Sara	200	...
	35654722	21-01-2011:09.06	examine casually	Sue	400	...
	35654724	21-01-2011:11.34	check ticket	Pete	100	...
	35654725	23-01-2011:13.12	decide	Sara	200	...
	35654726	24-01-2011:14.56	reject request	Mike	200	...
6	35654871	06-01-2011:15.02	register request	Mike	50	...
	35654873	06-01-2011:16.06	examine casually	Ellen	400	...
	35654874	07-01-2011:16.22	check ticket	Mike	100	...
	35654875	07-01-2011:16.52	decide	Sara	200	...
	35654877	16-01-2011:11.47	pay compensation	Mike	200	...
...	...	...	...	...	...	...

case id	trace
1	$\langle a,b,d,e,h \rangle$
2	$\langle a,d,c,e,g \rangle$
3	$\langle a,c,d,e,f,b,d,e,g \rangle$
4	$\langle a,d,b,e,h \rangle$
5	$\langle a,c,d,e,f,d,c,e,f,c,d,e,h \rangle$
6	$\langle a,c,d,e,g \rangle$
...	...

a = register request,
b = examine thoroughly,
c = examine casually,
d = check ticket,
e = decide,
f = reinitiate request,
g = pay compensation,
and h = reject request

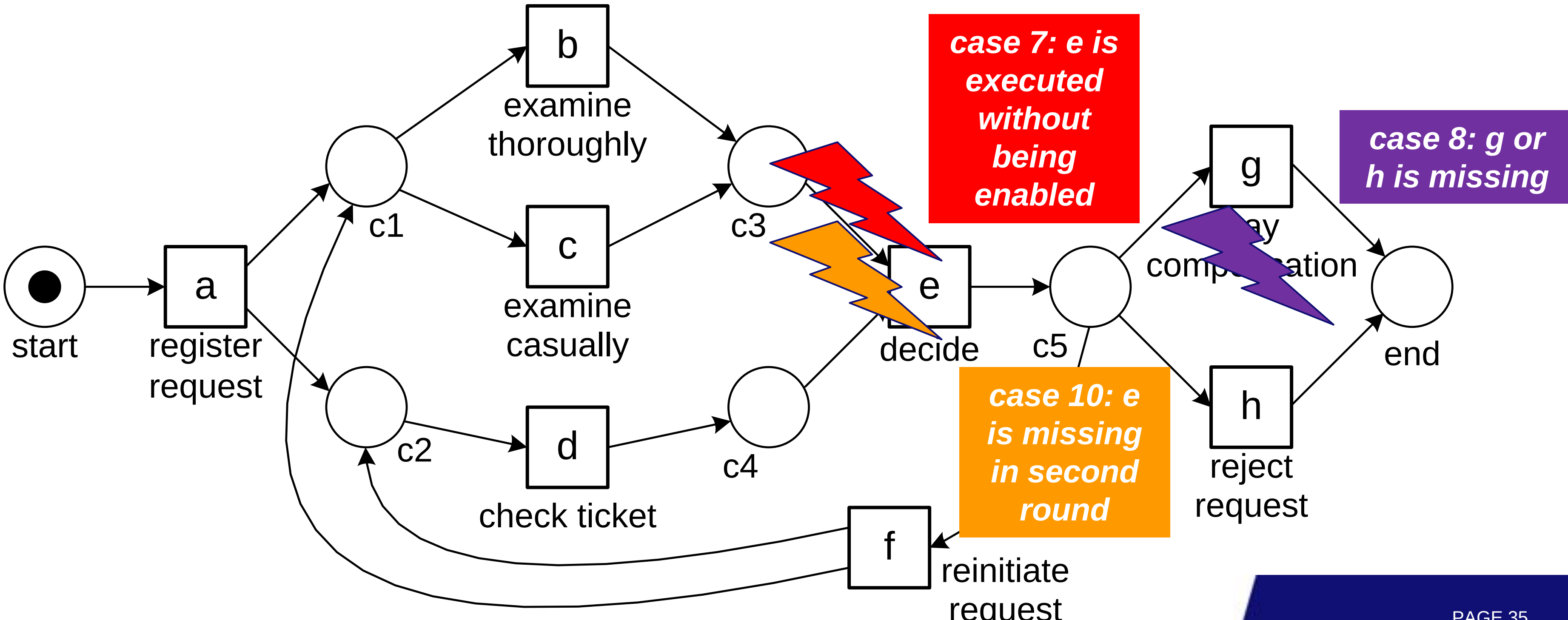
Process discovery

case id	trace
1	$\langle a,b,d,e,h \rangle$
2	$\langle a,d,c,e,g \rangle$
3	$\langle a,c,d,e,f,b,d,e,g \rangle$
4	$\langle a,d,b,e,h \rangle$
5	$\langle a,c,d,e,f,d,c,e,f,c,d,e,h \rangle$
6	$\langle a,c,d,e,g \rangle$
...	...



Conformance checking

case id	trace
1	$\langle a,b,d,e,h \rangle$
2	$\langle a,d,c,e,g \rangle$
3	$\langle a,c,d,e,f,b,d,e,g \rangle$
4	$\langle a,d,b,e,h \rangle$
5	$\langle a,c,d,e,f,d,c,e,f,c,d,e,h \rangle$
6	$\langle a,c,d,e,g \rangle$
7	$\langle \mathbf{a,b,e,g} \rangle$
8	$\langle \mathbf{a,b,d,e} \rangle$
9	$\langle a,d,c,e,f,d,c,e,f,b,d,e,h \rangle$
10	$\langle \mathbf{a,c,d,e,f,b,d,g} \rangle$

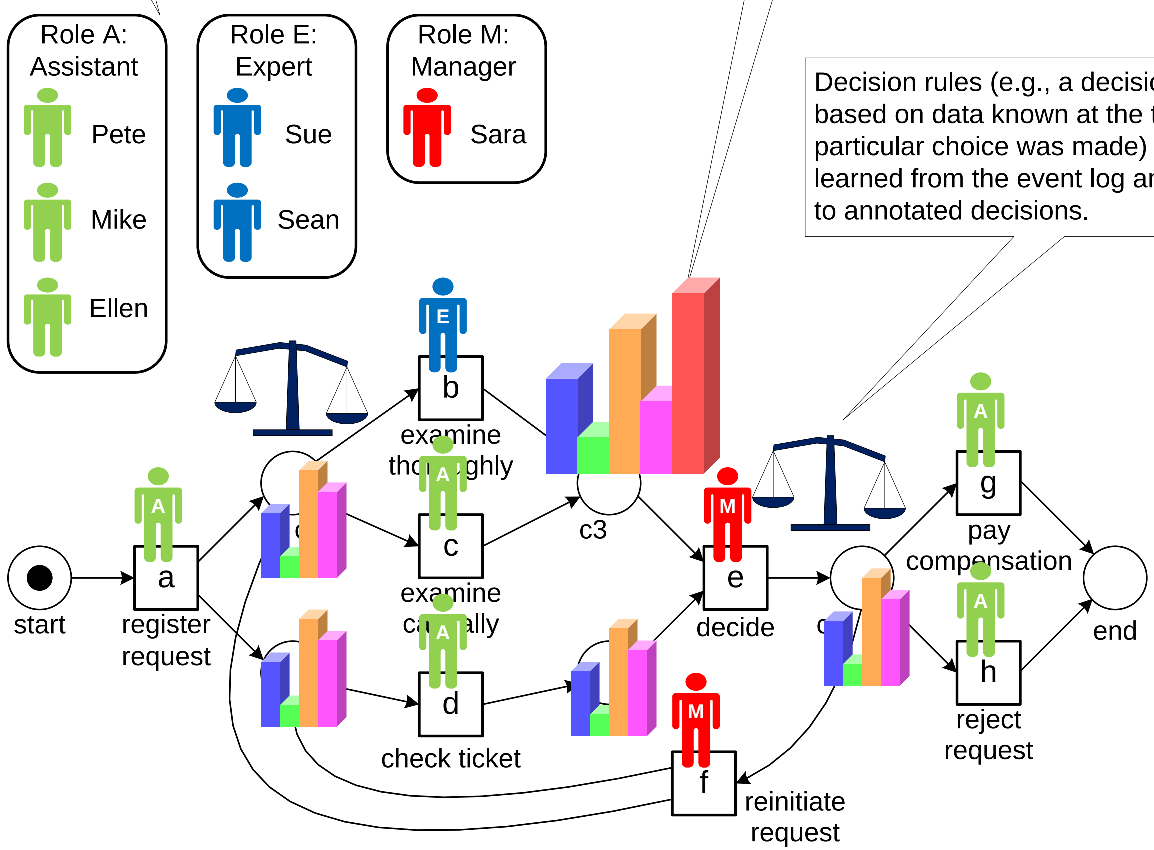


Extension: Adding perspectives to model based on event log

The event log can be used to discover roles in the organization (e.g., groups of people with similar work patterns). These roles can be used to relate individuals and activities.

Performance information (e.g., the average time between two subsequent activities) can be extracted from the event log and visualized on top of the model.

Decision rules (e.g., a decision tree based on data known at the time a particular choice was made) can be learned from the event log and used to annotated decisions.



We applied ProM in >100 organizations

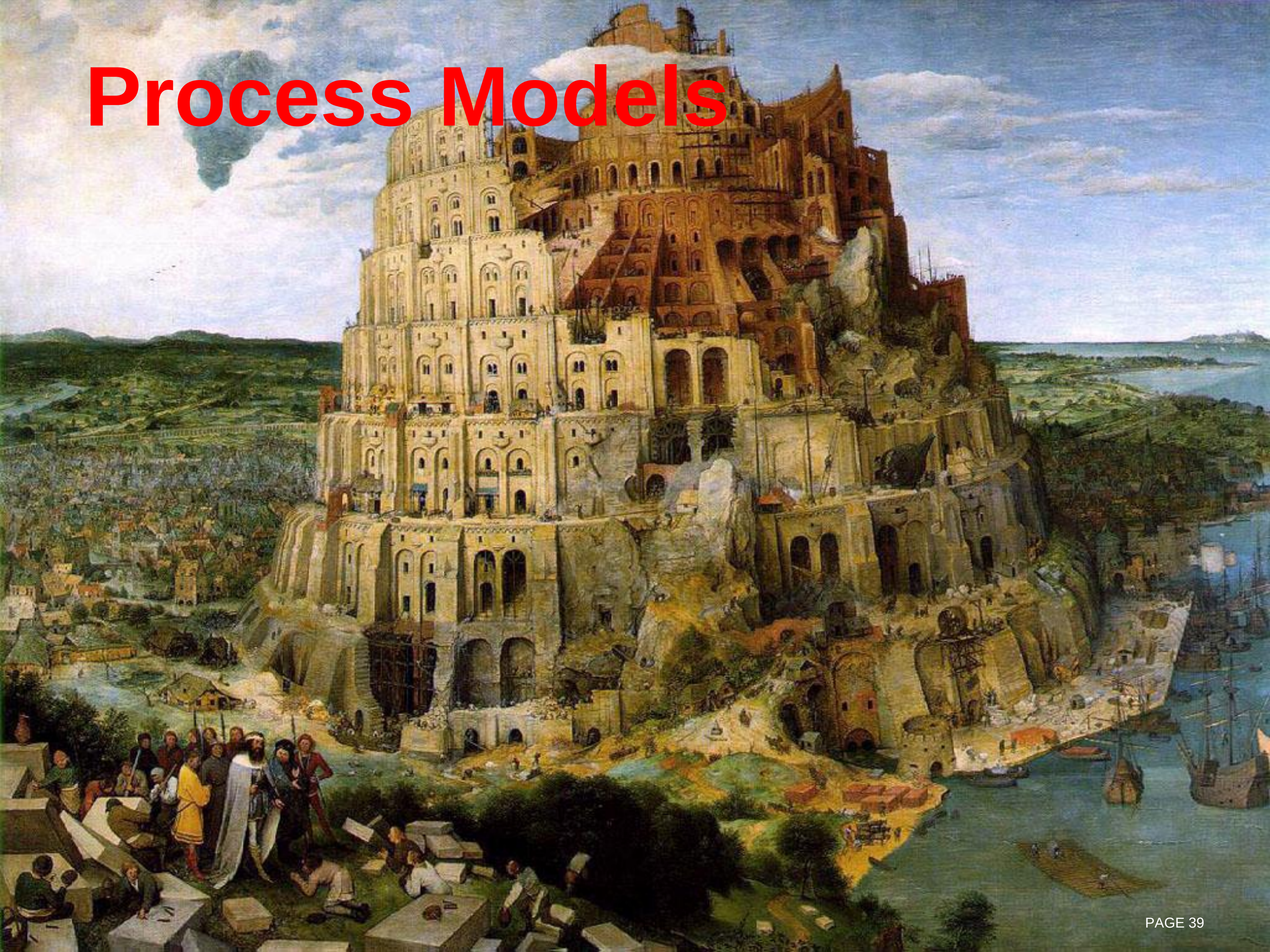
- **Municipalities** (e.g., Alkmaar, Heusden, Harderwijk, etc.)
- **Government agencies** (e.g., Rijkswaterstaat, Centraal Justitieel Incasso Bureau, Justice department)
- **Insurance related agencies** (e.g., UWV)
- **Banks** (e.g., ING Bank)
- **Hospitals** (e.g., AMC hospital, Catharina hospital)
- **Multinationals** (e.g., DSM, Deloitte)
- **High-tech system manufacturers and their customers** (e.g., Philips Healthcare, ASML, Ricoh, Thales)
- **Media companies** (e.g. Winkwaves)
- **...**

All supported by …

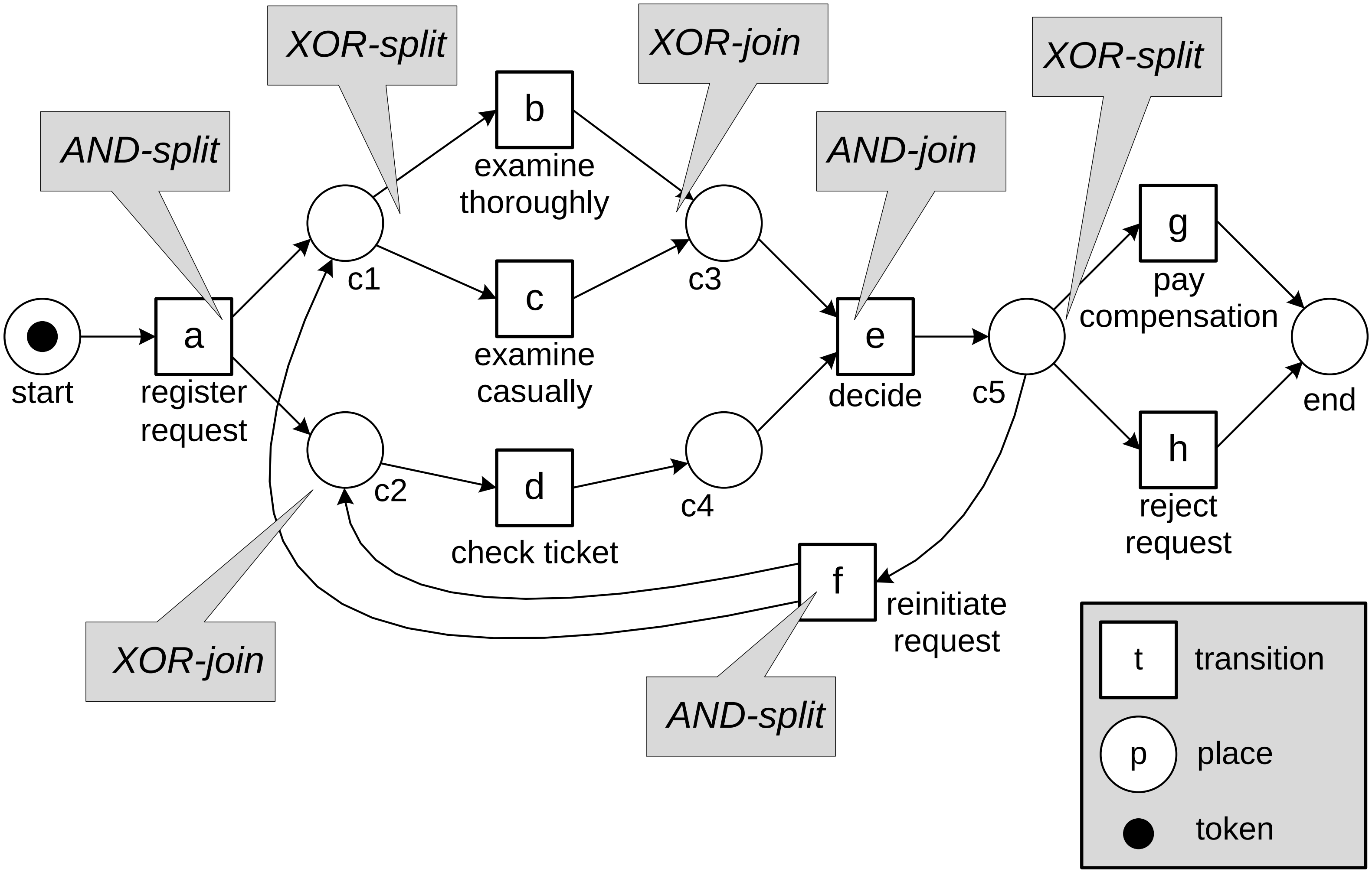
ProM
process mining workbench

- Open-source (L-GPL), cf. www.processmining.org
- Plug-in architecture
- Plug-ins cover the whole process mining spectrum and also support classical forms of process analysis

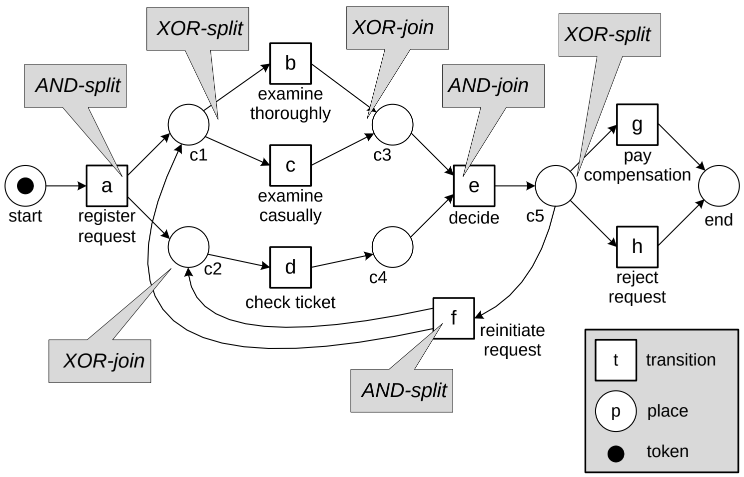
Process Models

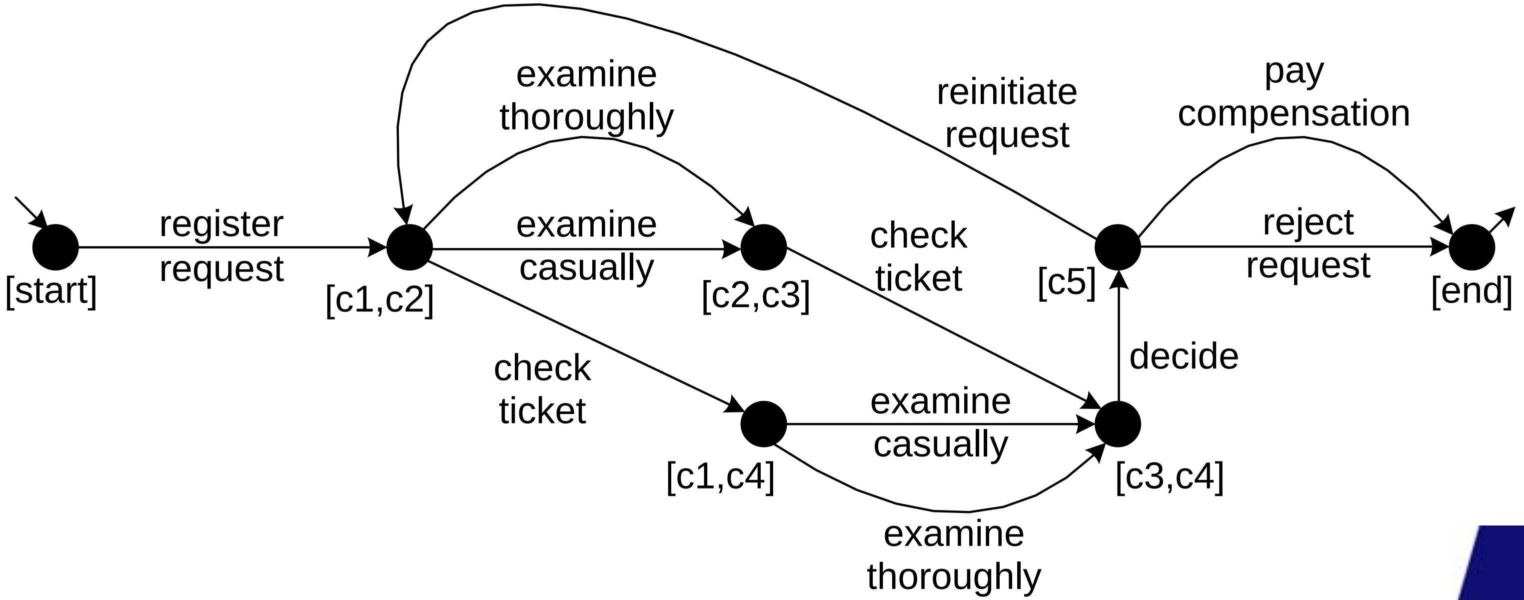


Petri nets

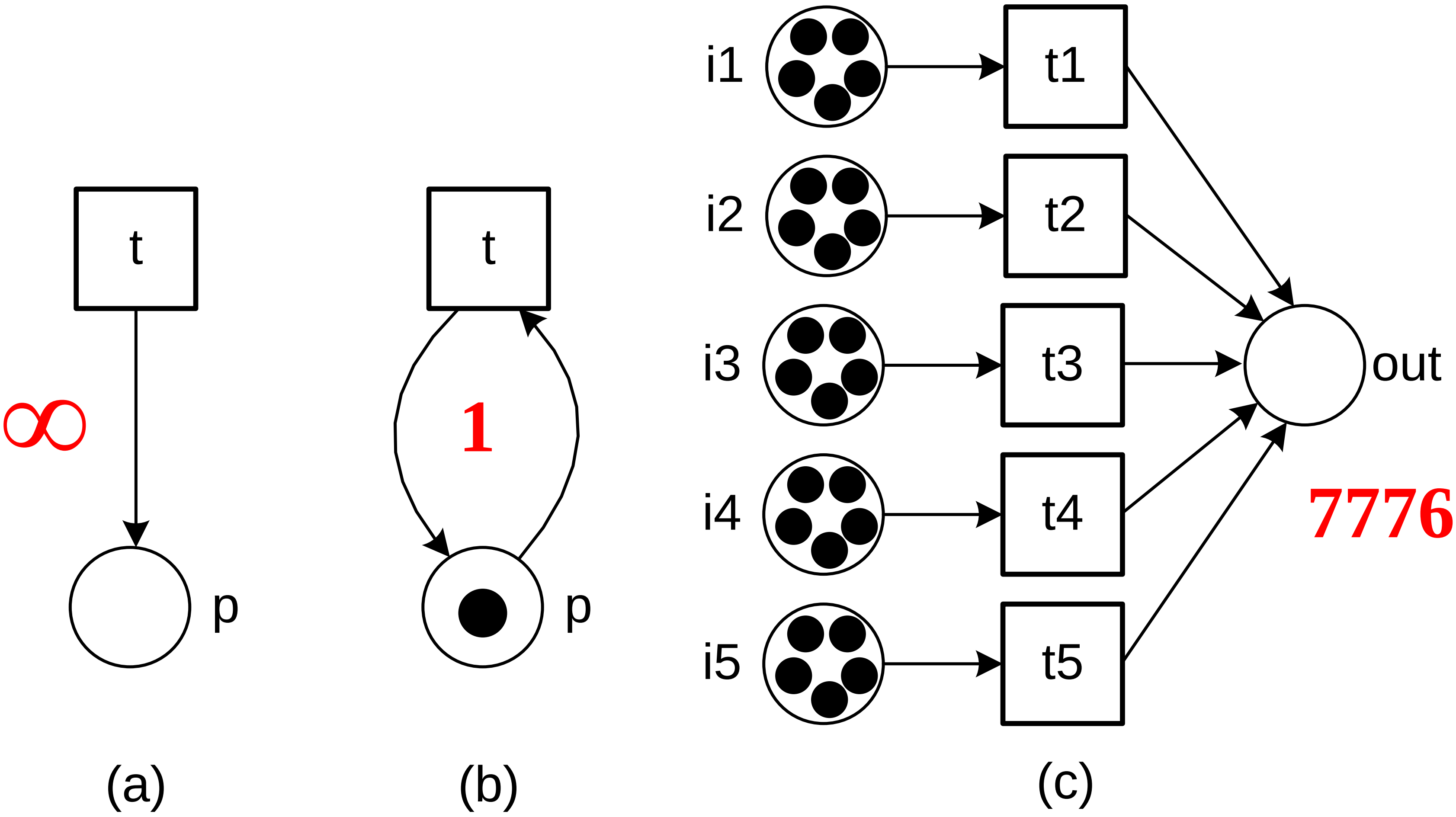

XOR-split
XOR-join
AND-split
AND-join
XOR-split
b
examine thoroughly
c
examine casually
d
check ticket
a
register request
start
c1
c2
c3
c4
e
decide
c5
g
pay compensation
h
reject request
end
f
reinitiate request
XOR-join
AND-split
t
transition
p
place
token

Reachability graph

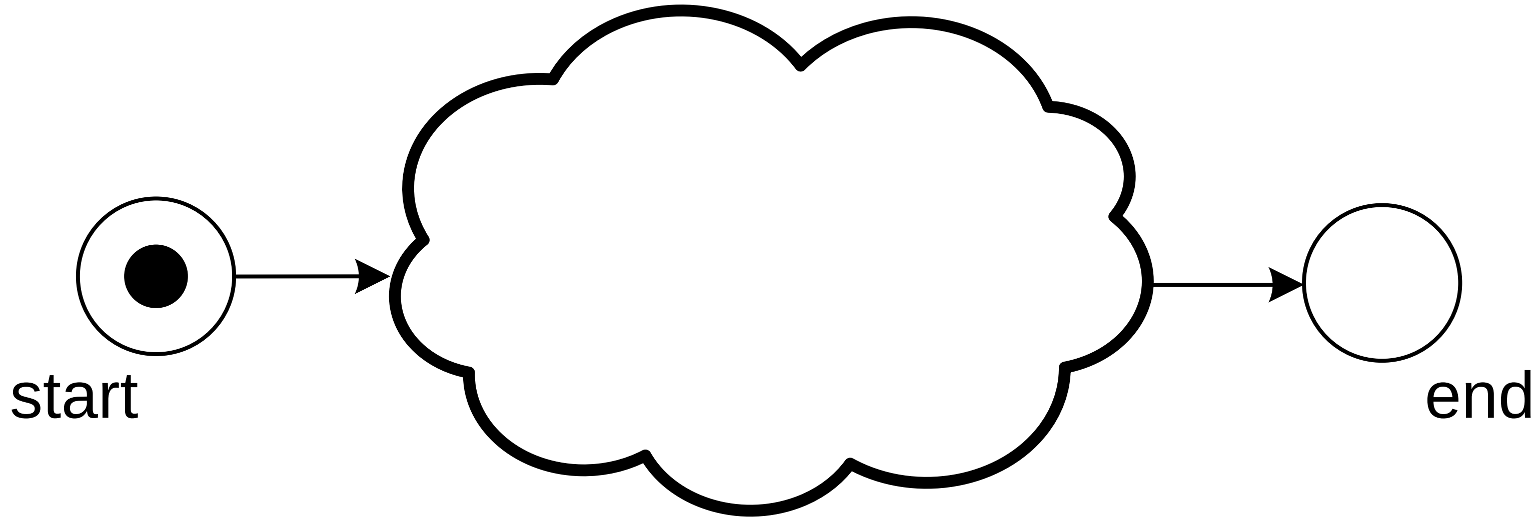




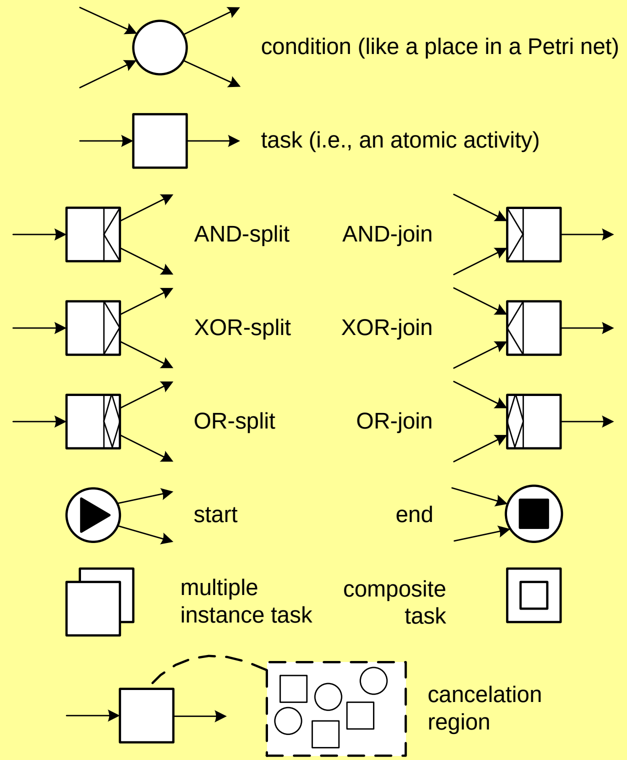
How many states?

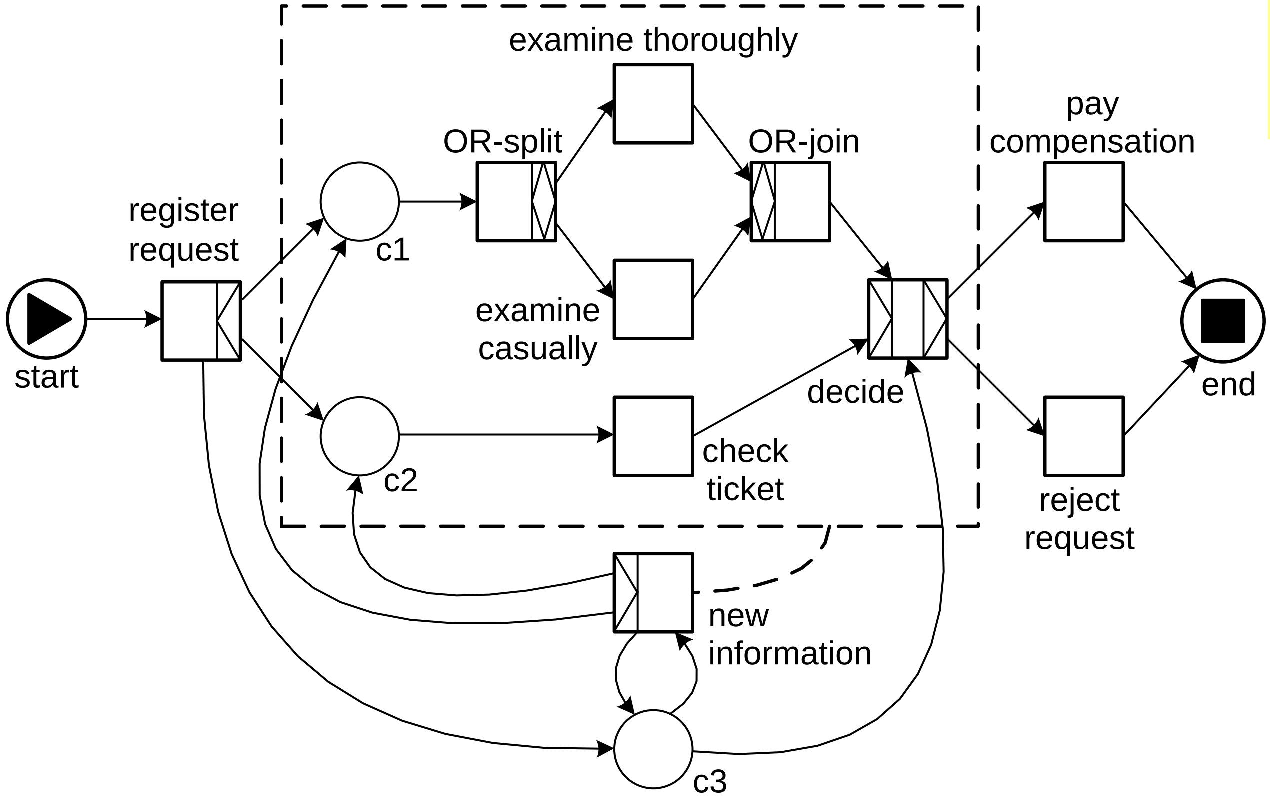


WF-nets and soundness

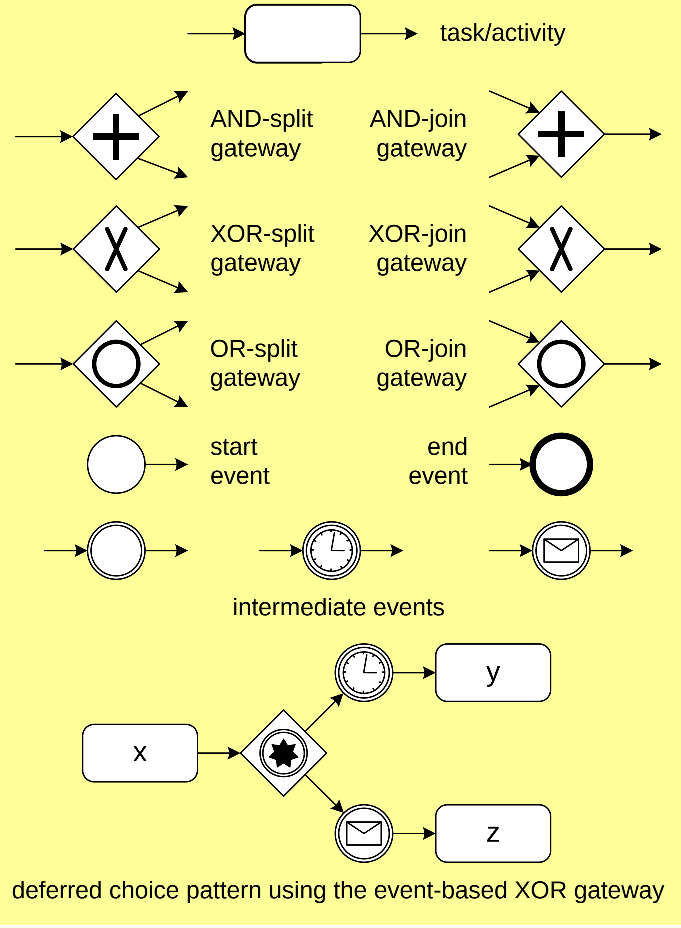


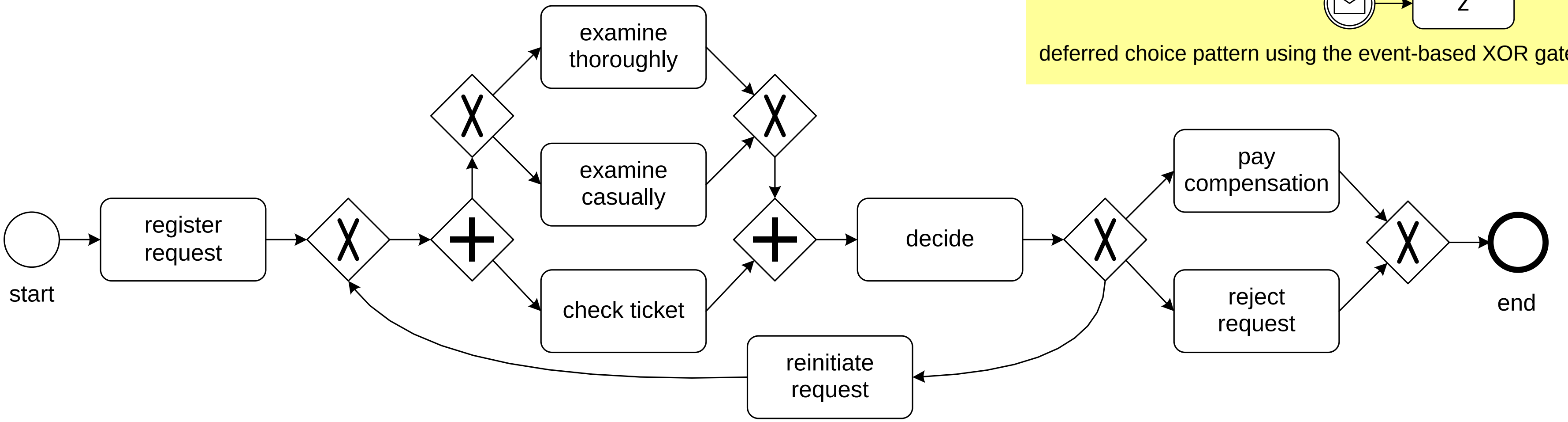
YAWL


condition (like a place in a Petri net)
task (i.e., an atomic activity)
AND-split
AND-join
XOR-split
XOR-join
OR-split
OR-join
start
end
multiple instance task
composite task
cancelation region

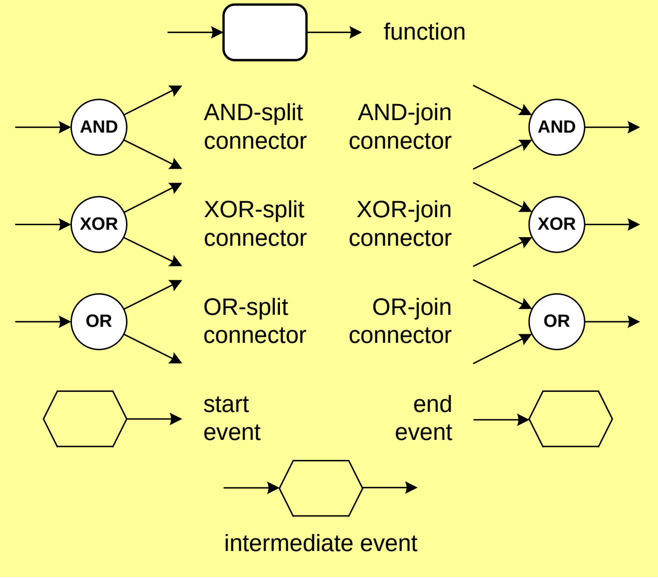

examine thoroughly
OR-split
OR-join
pay compensation
register request
c1
examine casually
decide
start
end
check ticket
c2
reject request
new information
c3

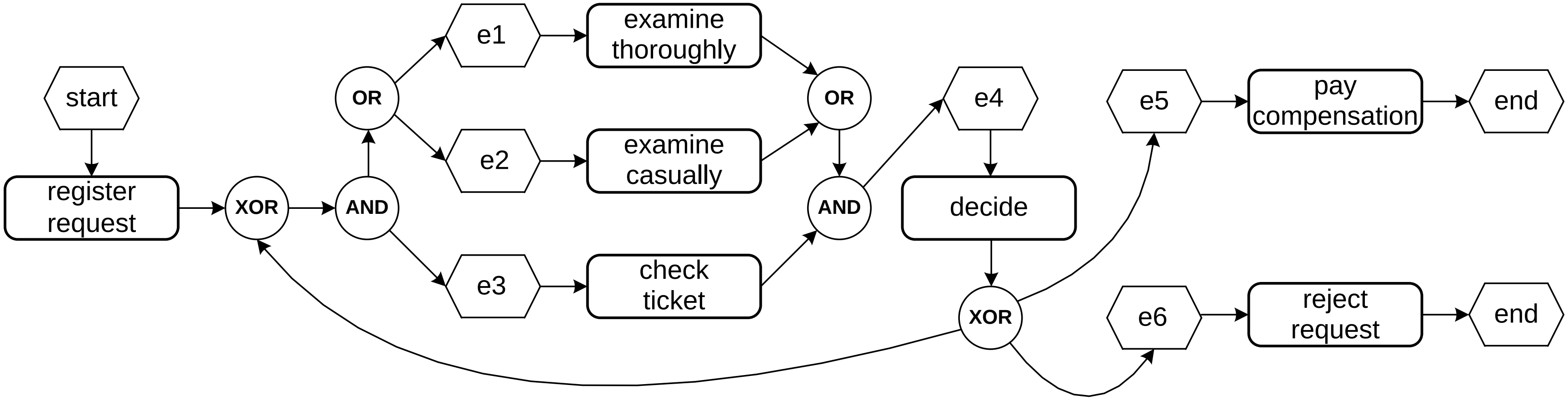
BPMN


task/activity
AND-split gateway
AND-join gateway
XOR-split gateway
XOR-join gateway
OR-split gateway
OR-join gateway
start event
end event
intermediate events
x
y
z
deferred choice pattern using the event-based XOR gateway


start
register request
examine thoroughly
examine casually
check ticket
decide
pay compensation
reject request
reinitiate request
end

Event-Driven Process Chains (EPCs)



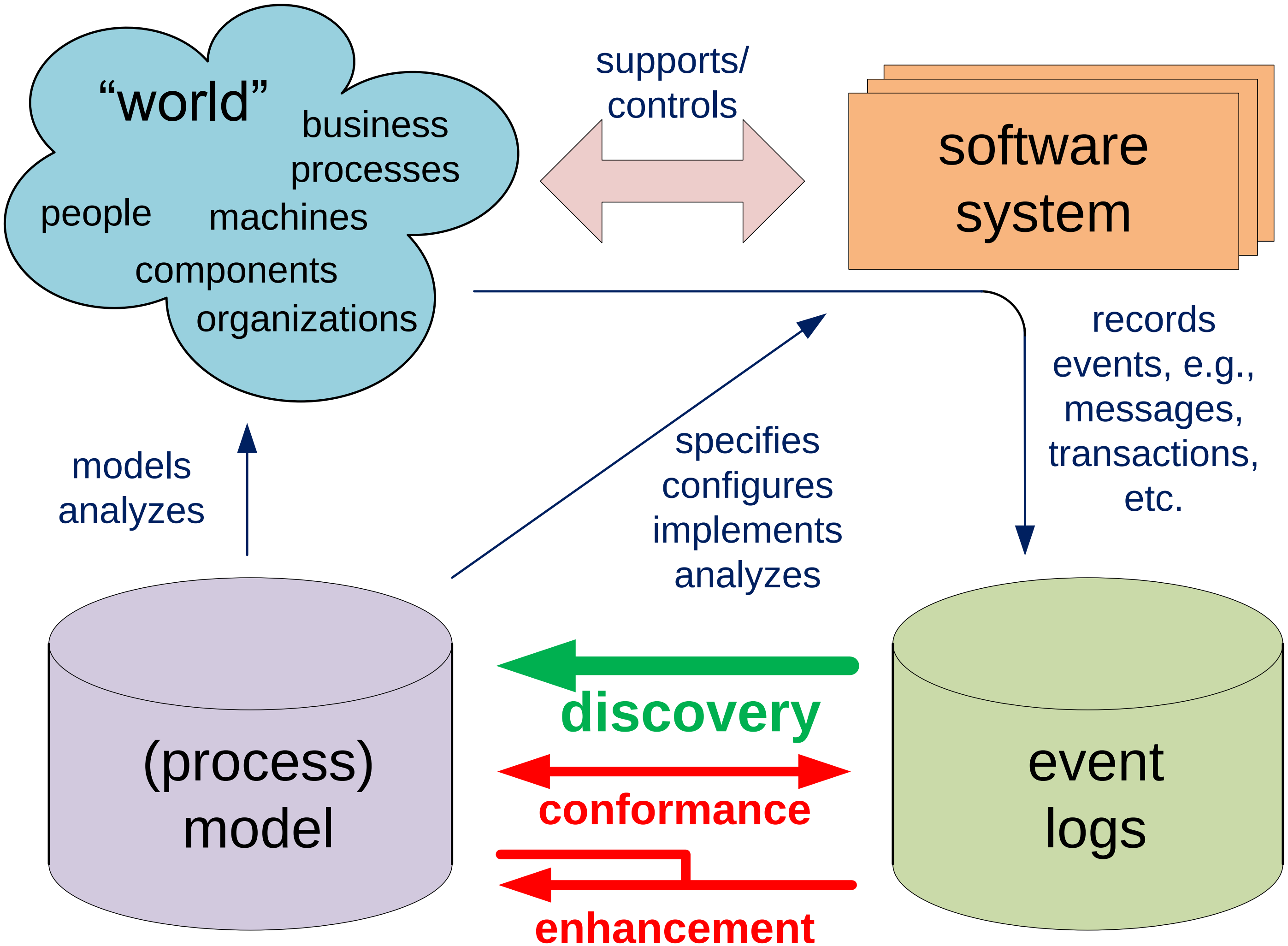


Representational bias matters, but …

…the visual representation is less relevant!

The representational bias determinates the search space, but can be decoupled from the visualization.

Process discovery



Process Discovery Techniques (small selection)

distributed genetic mining

automata-based learning

language-based regions

heuristic mining

state-based regions

partial-order based mining

genetic mining

LTL mining

pattern-based mining

neural networks

stochastic task graphs

fuzzy mining

mining block structures

hidden Markov models

α algorithm

multi-phase mining

conformal process graph

α# algorithm

ILP mining

α++ algorithm

Language identification in the limit (Mark Gold 1967)

A language is learnable in the limit if there exists a perfect child that generates only finitely many hypotheses.

Language identification in the limit by E Mark Gold, Information and Control, 10(5):447–474, 1967.

Learning is not easy …

- Even simple languag
  regular languag
  learnable i
- M
  with or
  tive examples,
  cles, etc.

Classical approaches (before 1998) did not consider concurrency and definitely not end-to-end business process models.

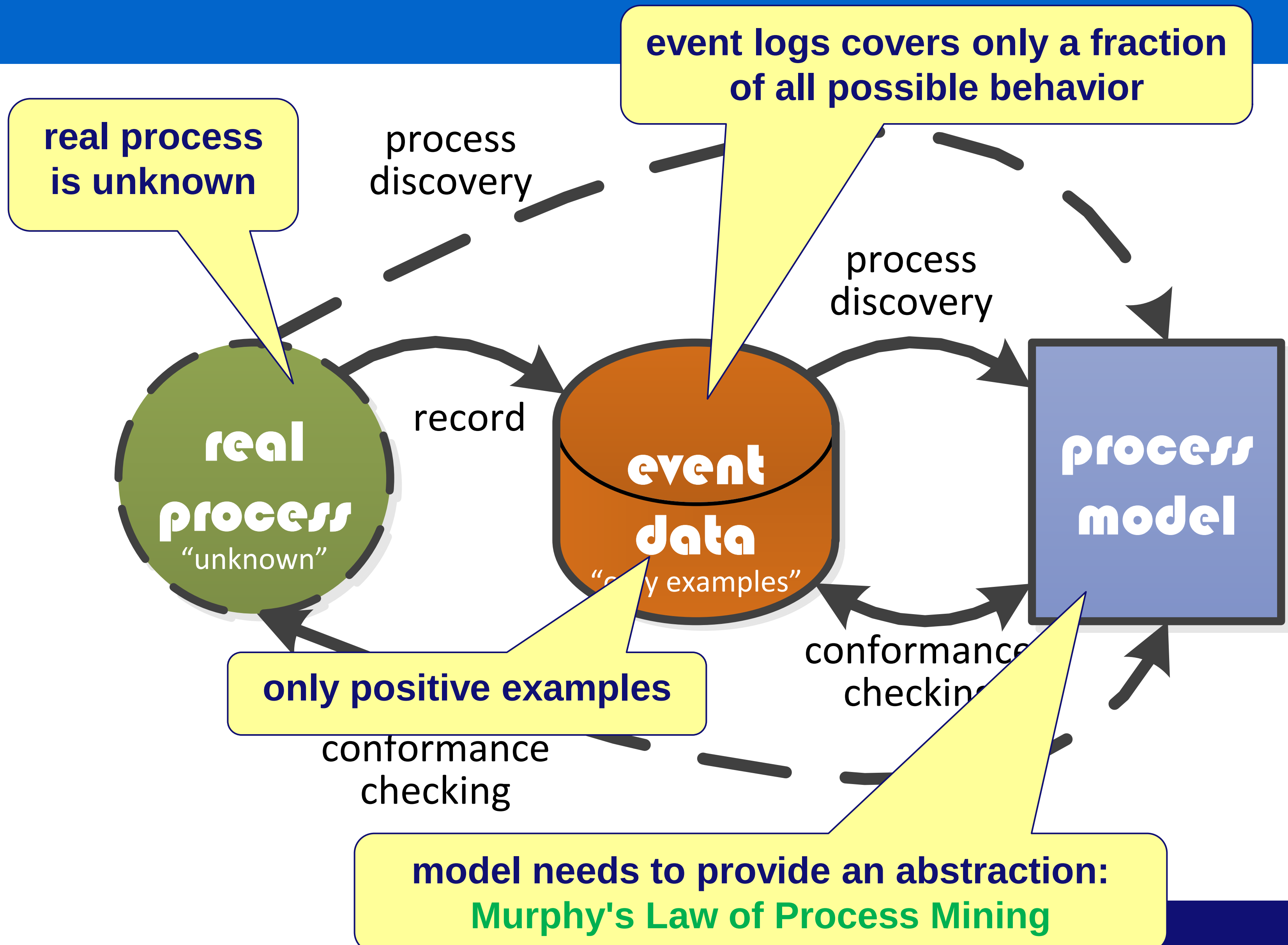
tence ≅ trace in event log

language ≅ process model

Why is process discovery such a difficult problem?

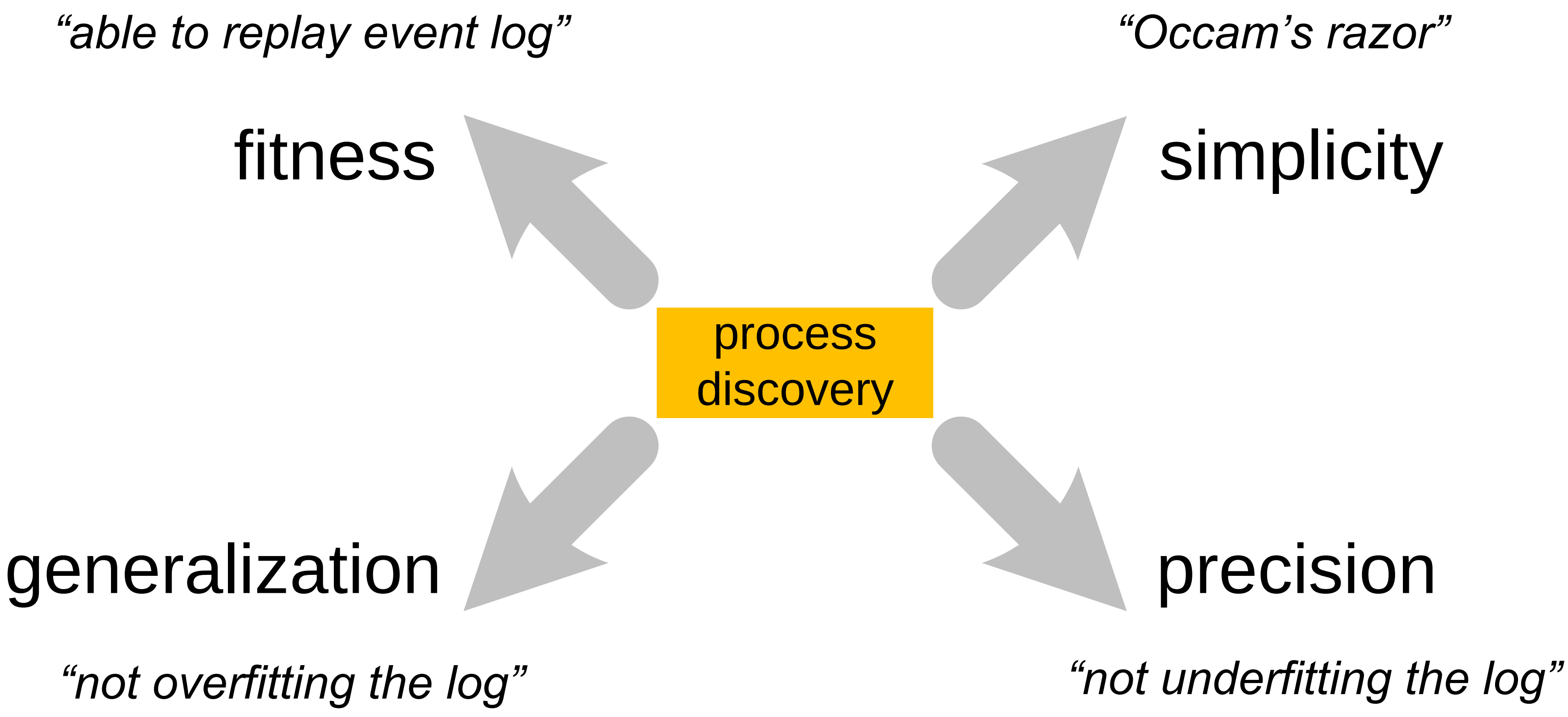
- There are no negative examples (i.e., a log shows what has happened but does not show what could not happen).
- Due to concurrency, loops, and choices the search space has a complex structure and the log typically contains only a fraction of all possible behaviors.
- There is no clear relation between the size of a model and its behavior (i.e., a smaller model may generate more or less behavior although classical analysis and evaluation methods typically assume some monotonicity property).

Problem

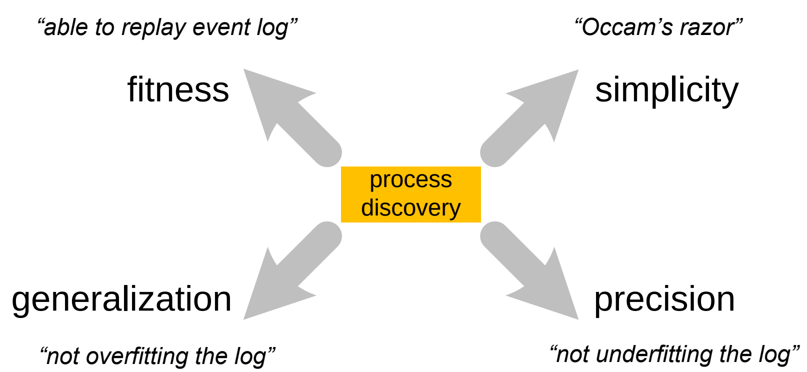


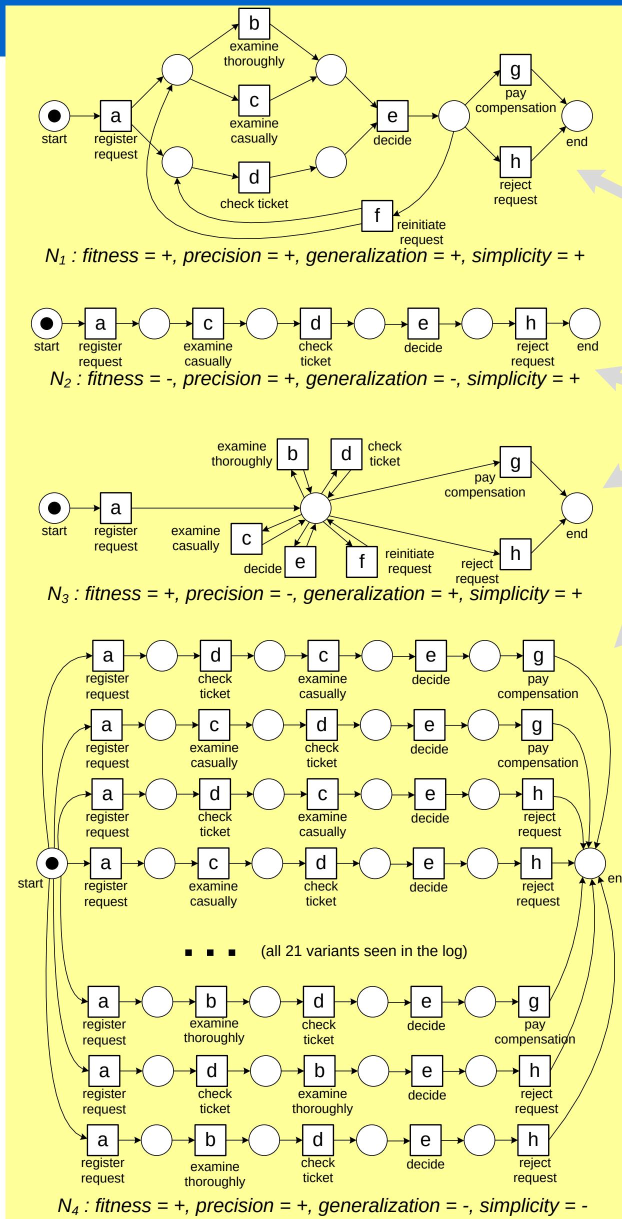
real process is unknown
event logs covers only a fraction of all possible behavior
process discovery
process discovery
record
real process
“unknown”
event data
process model
only positive examples
conformance checking
model needs to provide an abstraction:
Murphy's Law of Process Mining

Challenge: four competing quality criteria



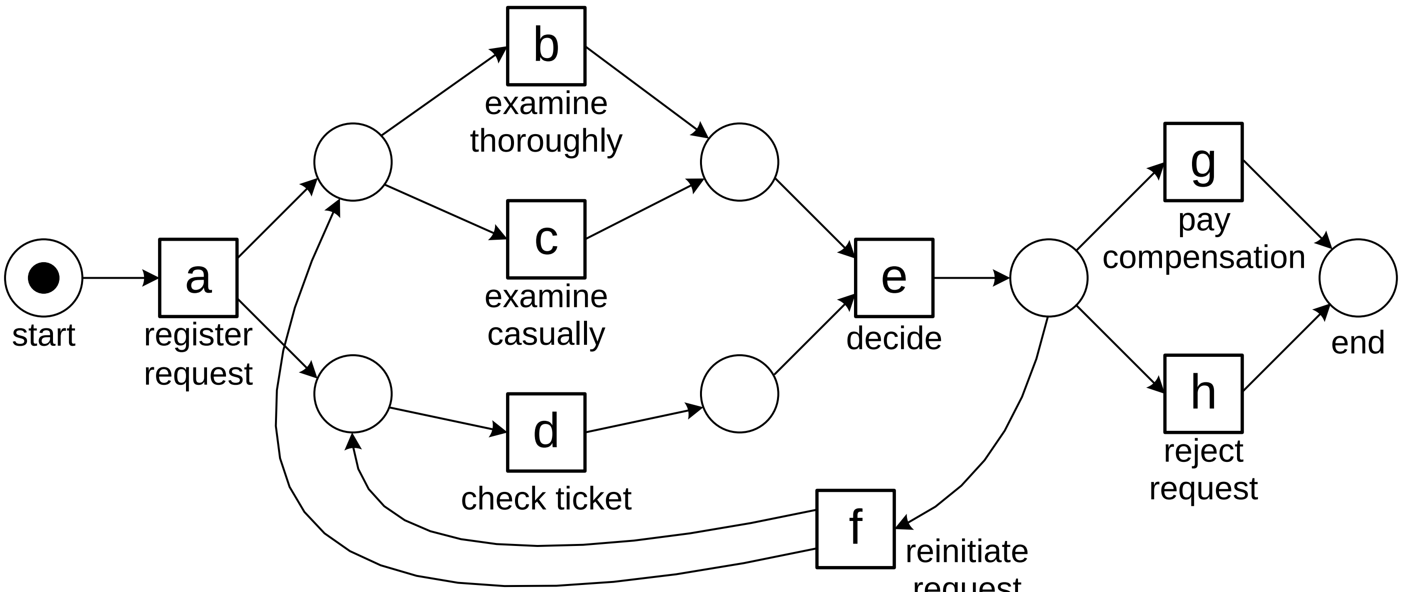
Example: one log four models





#	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	

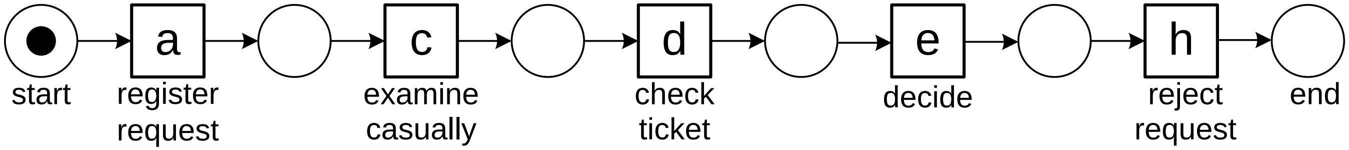
Model N_1



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	

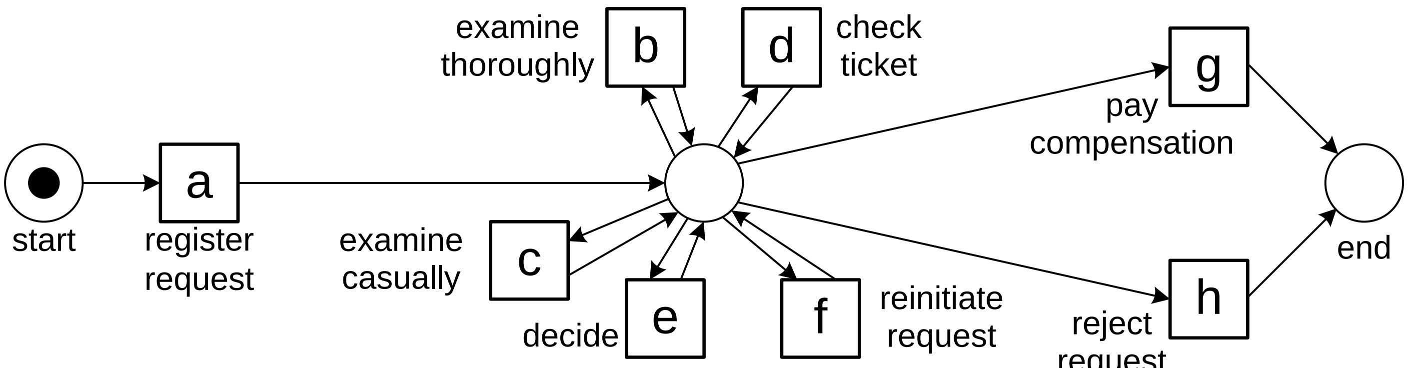
Model N_2



N_2 : *fitness* = -, *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	

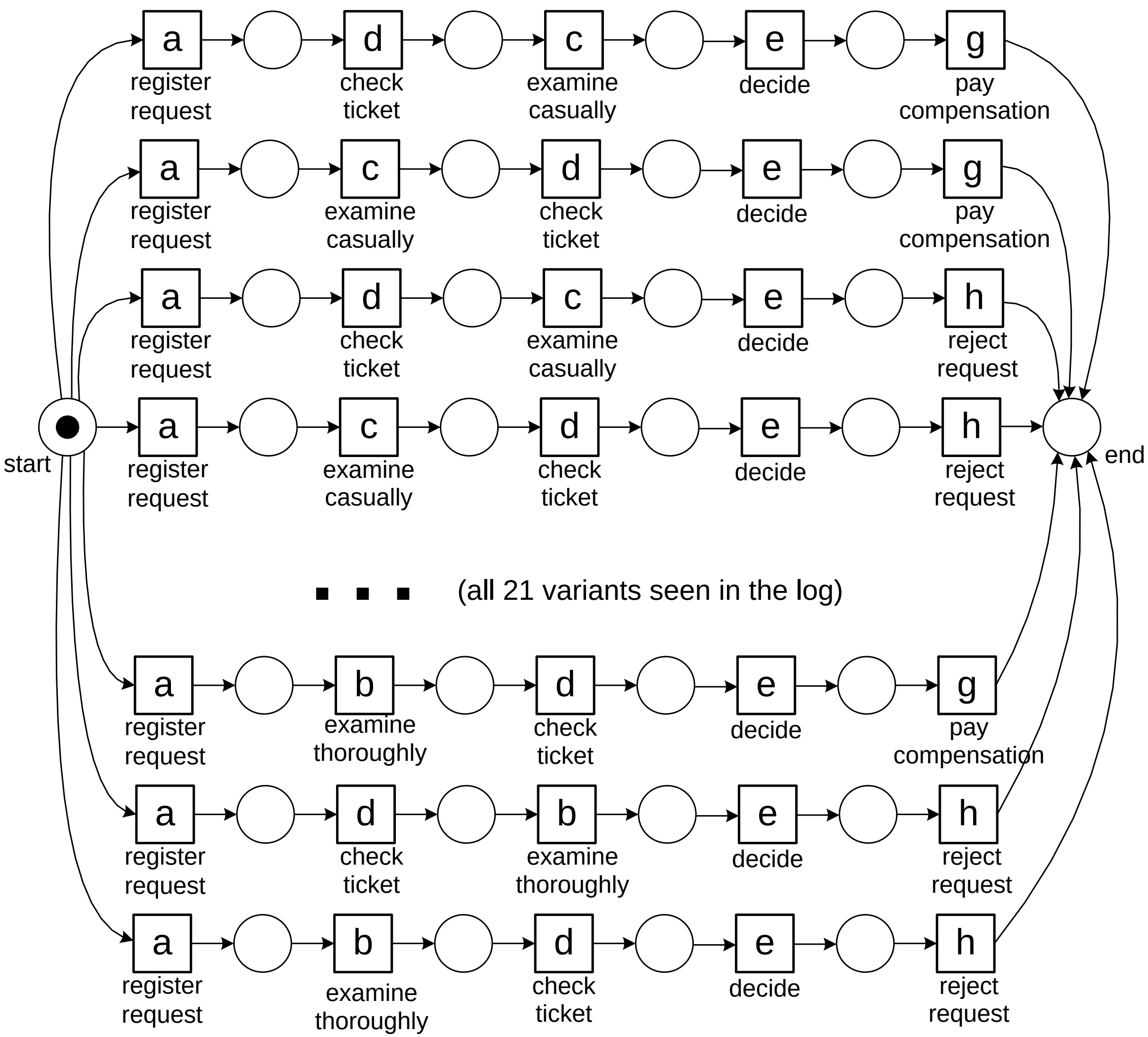
Model N_3



N_3 : 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	

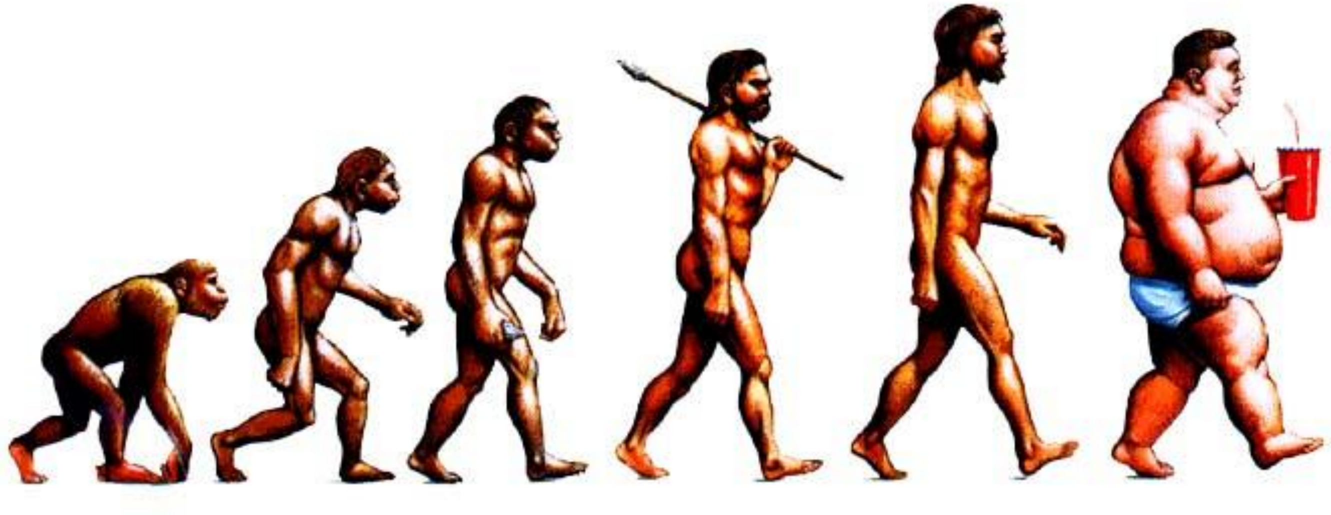
Model N_4



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

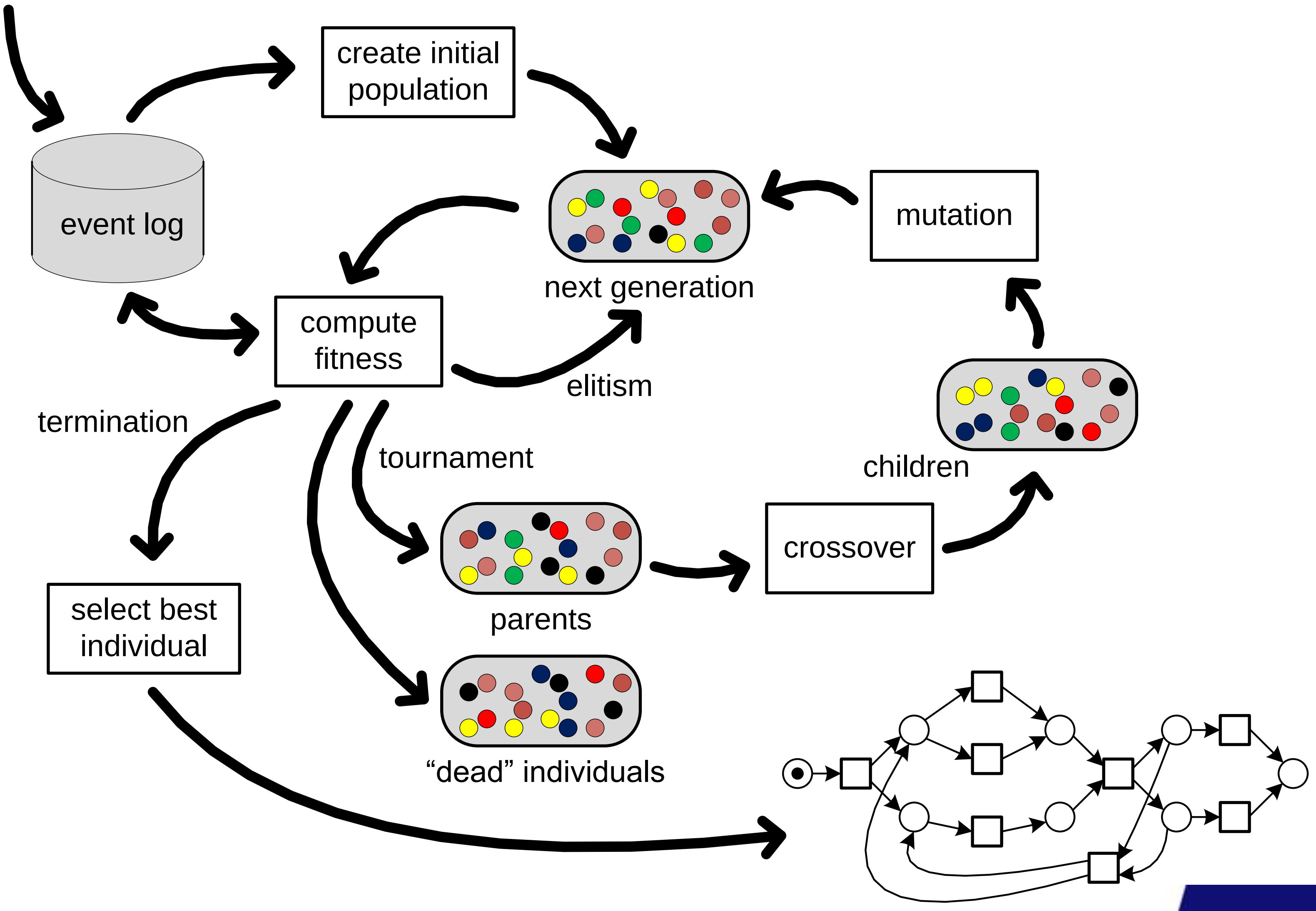
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

Example of a process discovery technique: Genetic Mining

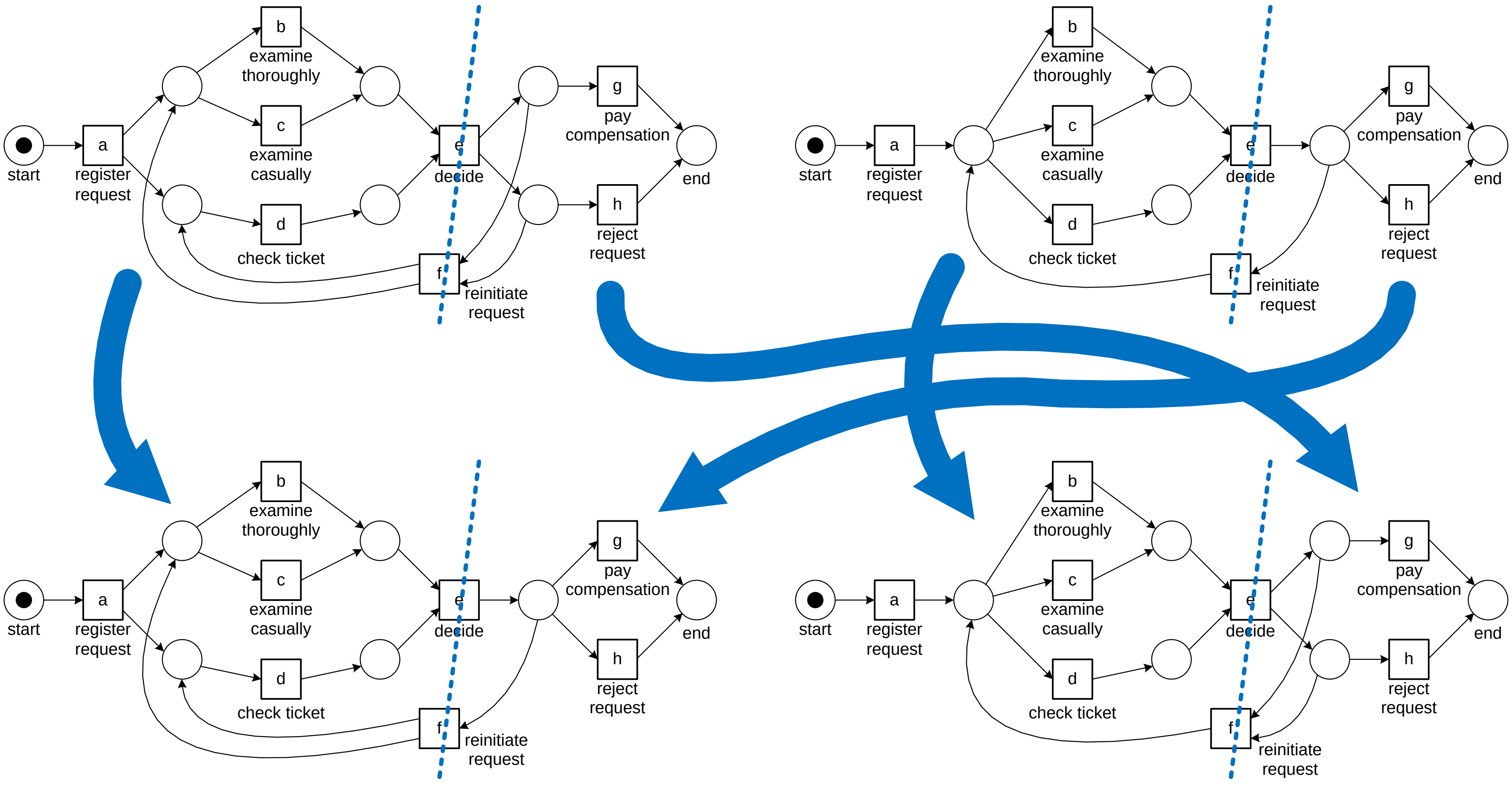


- **Characteristics**
  - requires a lot of computing power, but can be distributed easily,
  - can deal with noise, infrequent behavior, duplicate tasks, invisible tasks,
  - allows for incremental improvement and combinations with other approaches (heuristics post-optimization, etc.).

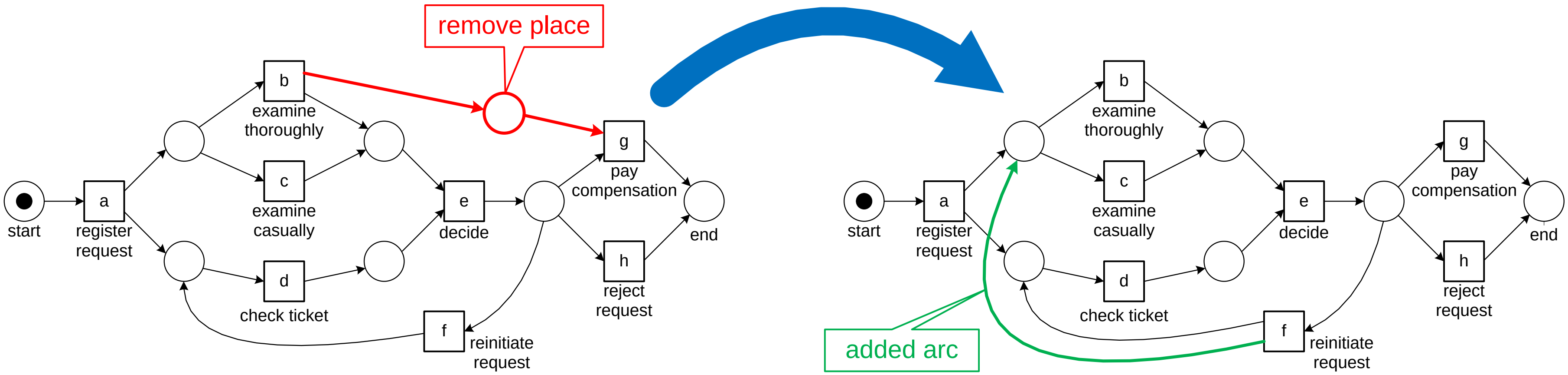
Genetic process mining: Overview



Example: crossover



Example: mutation



Process discovery algorithms
(small selection)

distributed genetic mining

automata-based learning

language-based regions

heuristic mining

state-based regions

genetic mining

LTL mining

stochastic task graphs

neural networks

fuzzy mining

hidden Markov models

mining block structures

α algorithm

conformal process graph

multi-phase mining

partial-order based mining

α# algorithm

ILP mining

α++ algorithm

The α algorithm

>,→,||,# relations

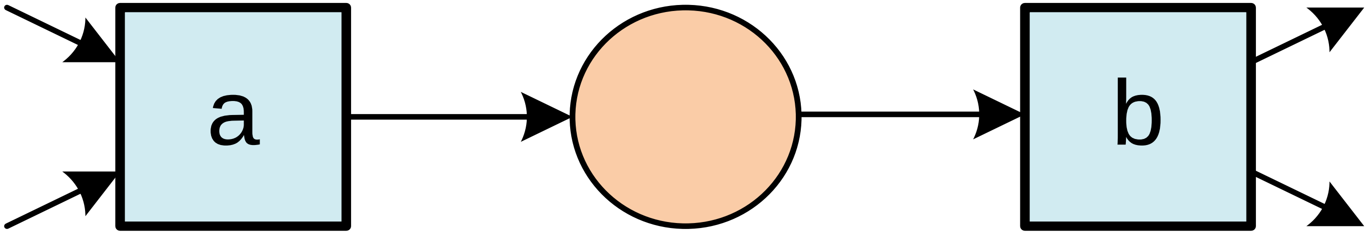
- Direct succession: x>y iff for some case x is directly followed by y.
- Causality: x→y iff x>y and not y>x.
- Parallel: x||y iff x>y and y>x
- Choice: x#y iff not x>y and not y>x.

$$L_1 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^2, \langle a,e,d\rangle]$$

abcd
acbd
aed

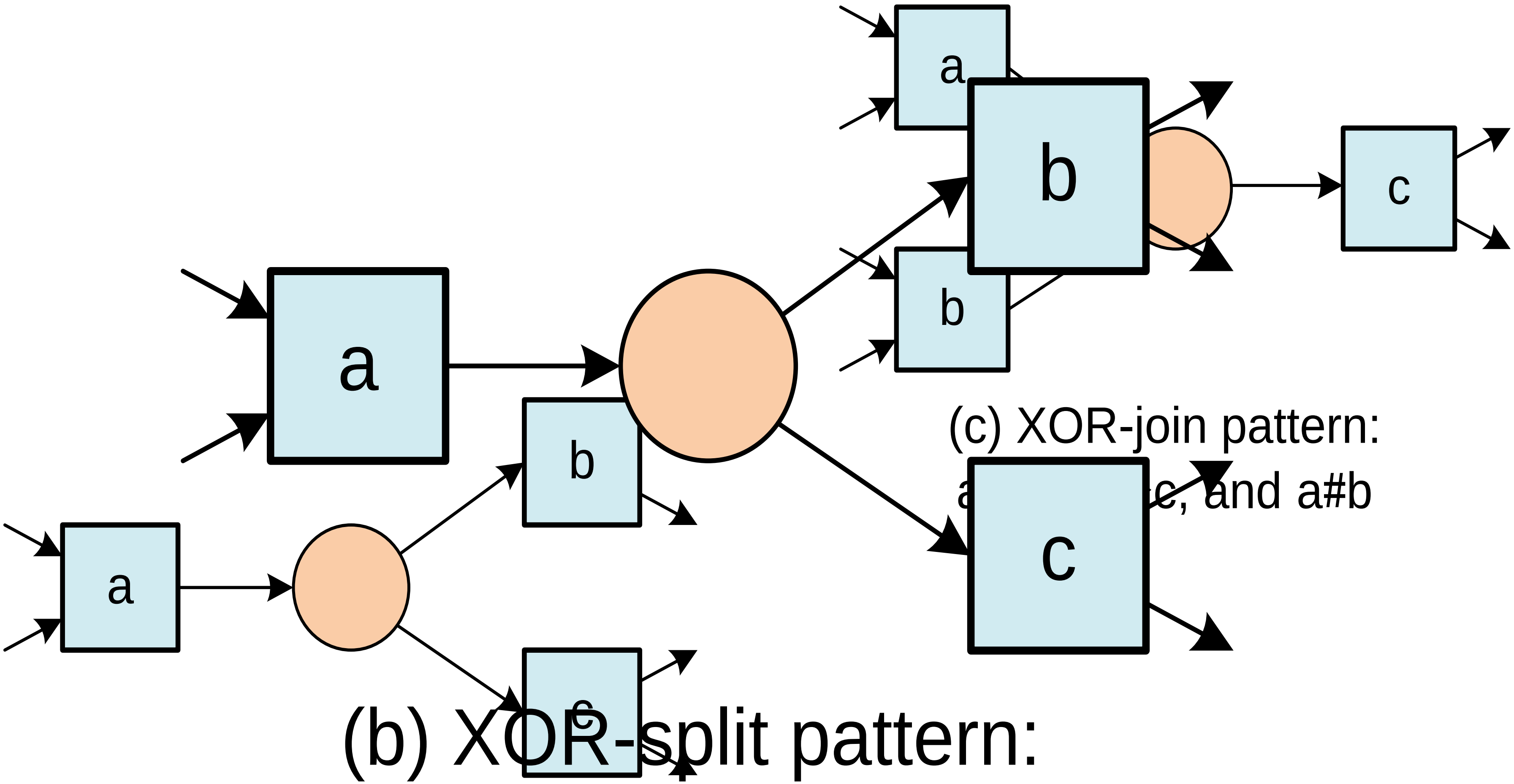
a>b a>c a>e b>c b>d c>b c>d e>d	a→b a→c a→e b→d c→d e→d	b||c c||b	b#e e#b c#e a#d …

Basic idea used by α Algorithm (1)



(a) sequence pattern: a→b

Basic idea used by α Algorithm (2)

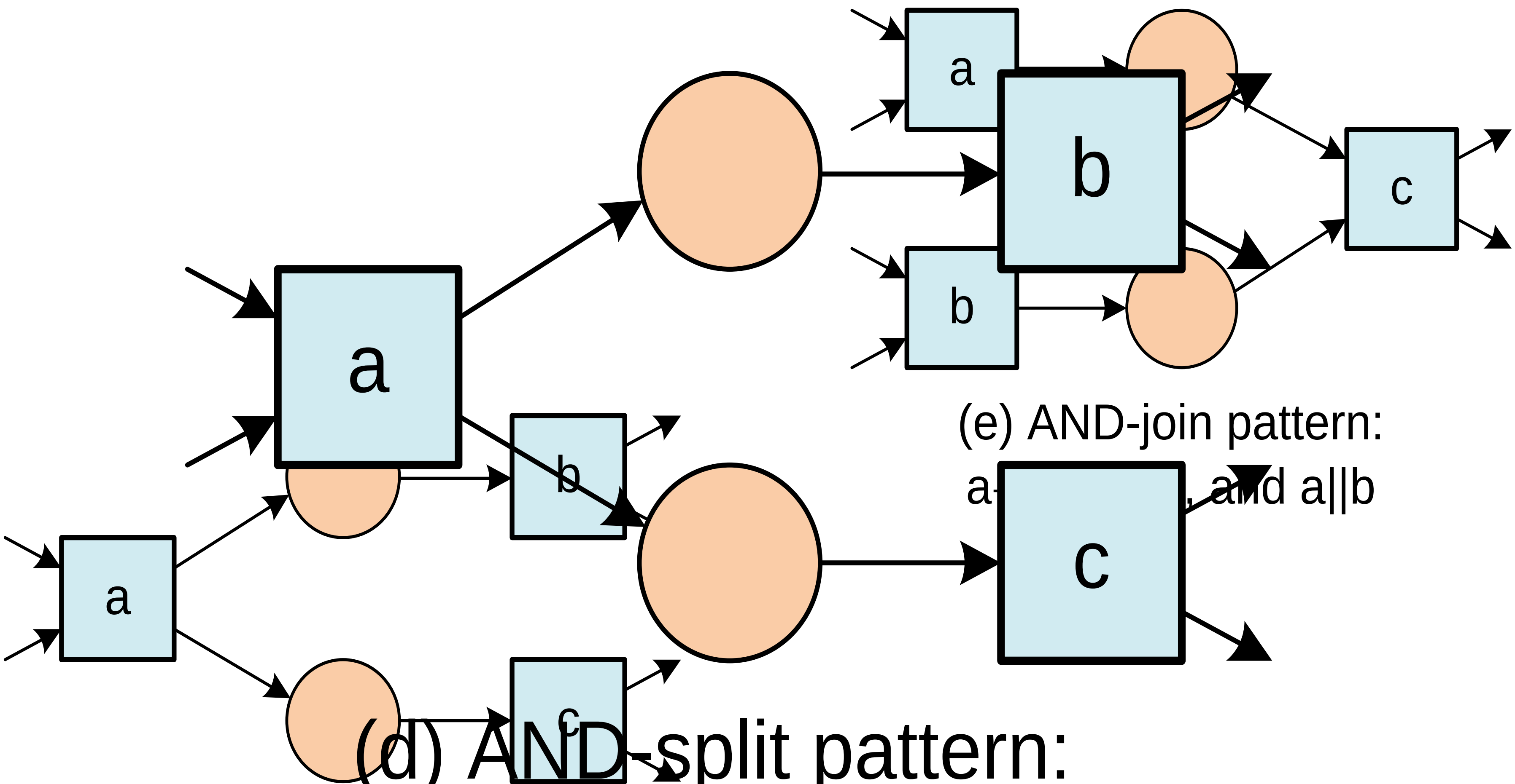


(c) XOR-join pattern:
a c, and a#b

(b) XOR-split pattern:
a→b, a→c, and b#c

(b) XOR-split pattern:
a→b, a→c, and b#c

Basic idea used by α Algorithm (3)



(e) AND-join pattern:
a[illegible], and a||b

(d) AND-split pattern:
a→b; a→c, and b||c

(d) AND-split pattern:
a→b, a→c, and b||c'

Example Revisited

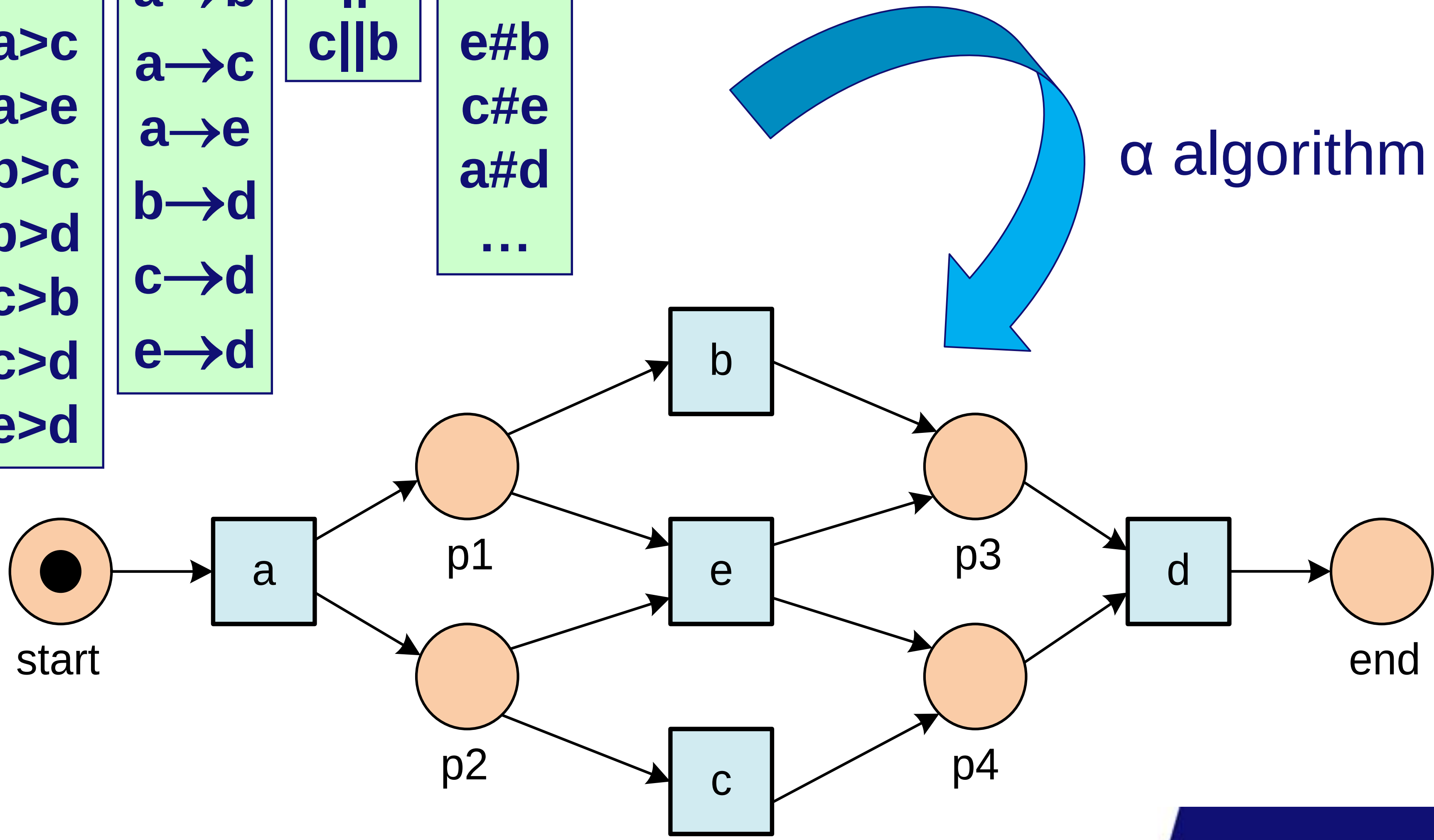
$$L_1 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^2, \langle a,e,d\rangle]$$

a>b
a>c
a>e
b>c
b>d
c>b
c>d
e>d

a→b
a→c
a→e
b→d
c→d
e→d

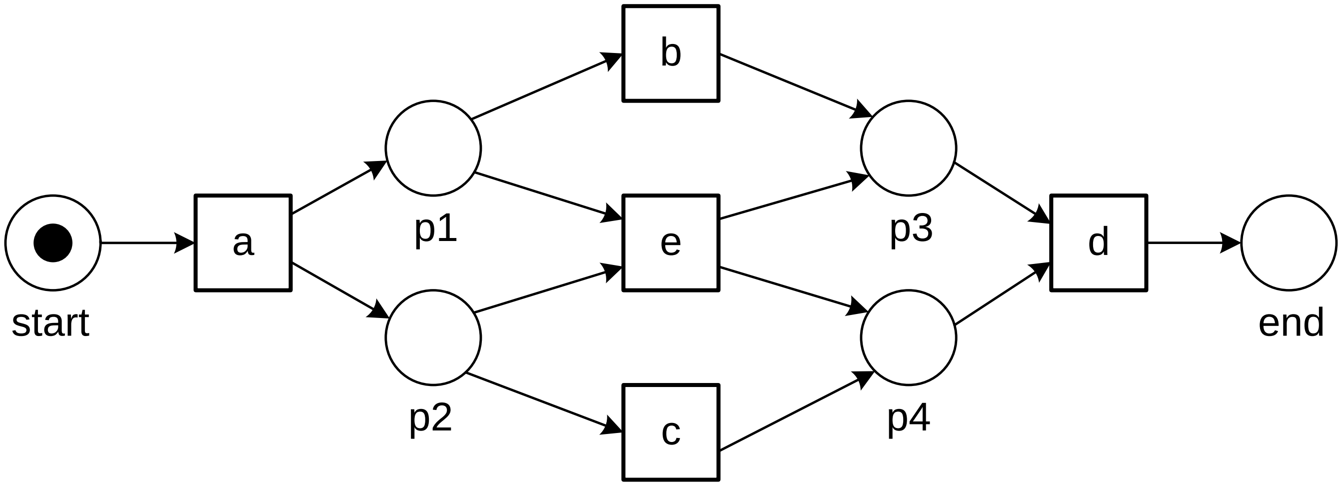
b||c
c||b

b#e
e#b
c#e
a#d
…



Footprint of L_1

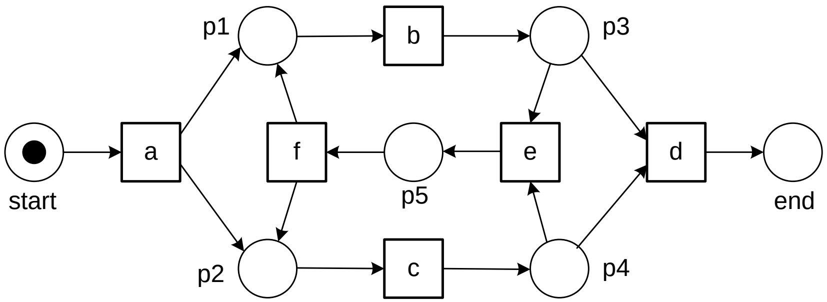
$$L_1 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^2, \langle a,e,d\rangle]$$



	a	b	c	d	e
a	$\#_{L_1}$	$\rightarrow_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$
b	$\leftarrow_{L_1}$	$\#_{L_1}$	$\Vert_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
c	$\leftarrow_{L_1}$	$\Vert_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
d	$\#_{L_1}$	$\leftarrow_{L_1}$	$\leftarrow_{L_1}$	$\#_{L_1}$	$\leftarrow_{L_1}$
e	$\leftarrow_{L_1}$	$\#_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$

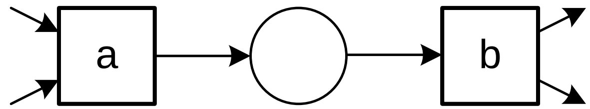
Footprint of L_2

$$L_2 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^4, \langle a,b,c,e,f,b,c,d\rangle^2, \langle a,b,c,e,f,c,b,d\rangle, \langle a,c,b,e,f,b,c,d\rangle^2, \langle a,c,b,e,f,b,c,e,f,c,b,d\rangle]$$

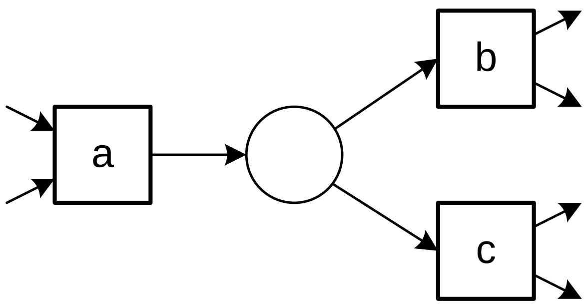


	a	b	c	d	e	f
a	$\#$	$\rightarrow$	$\rightarrow$	$\#$	$\#$	$\#$
b	$\leftarrow$	$\#$	$\parallel$	$\rightarrow$	$\rightarrow$	$\leftarrow$
c	$\leftarrow$	$\parallel$	$\#$	$\rightarrow$	$\rightarrow$	$\leftarrow$
d	$\#$	$\leftarrow$	$\leftarrow$	$\#$	$\#$	$\#$
e	$\#$	$\leftarrow$	$\leftarrow$	$\#$	$\#$	$\rightarrow$
f	$\#$	$\rightarrow$	$\rightarrow$	$\#$	$\leftarrow$	$\#$

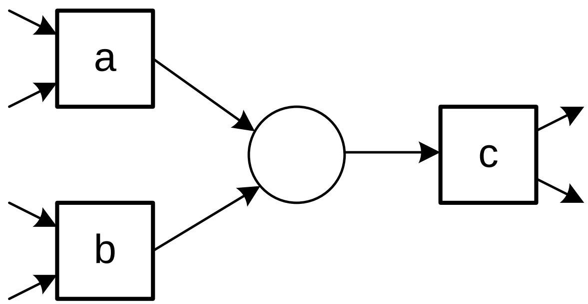
Simple patterns



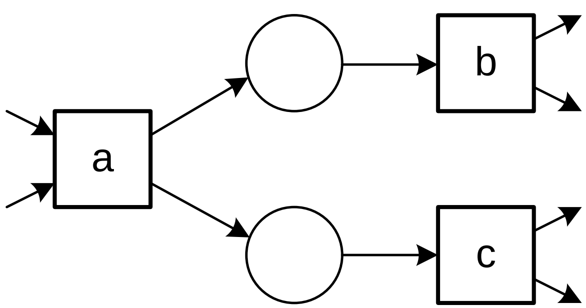
(a) sequence pattern: a→b



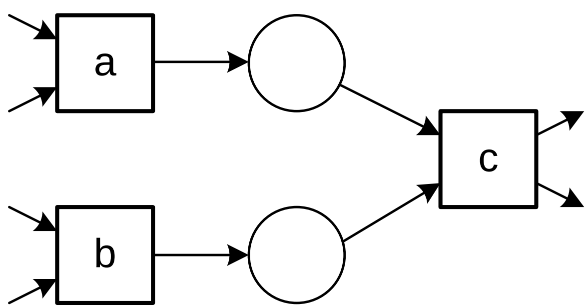
(b) XOR-split pattern:
a→b, a→c, and b#c



(c) XOR-join pattern:
a→c, b→c, and a#b



(d) AND-split pattern:
a→b, a→c, and b||c



(e) AND-join pattern:
a→c, b→c, and a||b

Algorithm

Let L be an event log over T. $\alpha(L)$ is defined as follows.

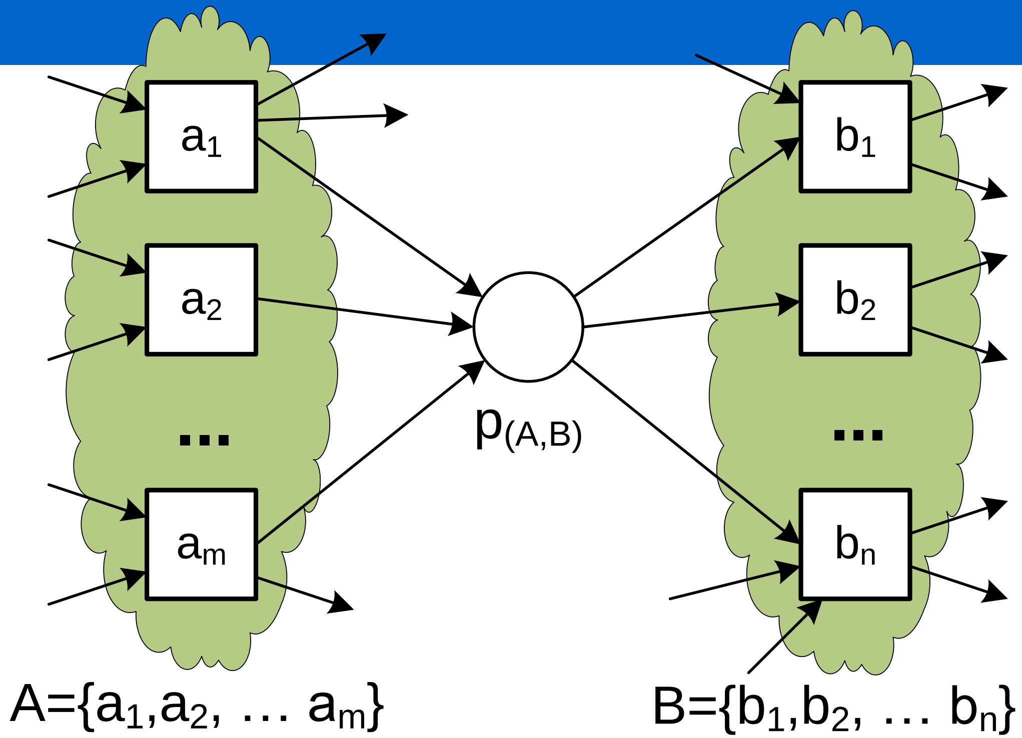
1. $T_L = \{ t \in T \mid \exists_{\sigma \in L}\, t \in \sigma \}$,
2. $T_I = \{ t \in T \mid \exists_{\sigma \in L}\, t = \mathit{first}(\sigma) \}$,
3. $T_O = \{ t \in T \mid \exists_{\sigma \in L}\, t = \mathit{last}(\sigma) \}$,
4. $X_L = \{ (A,B) \mid A \subseteq T_L \wedge A \neq \emptyset \wedge B \subseteq T_L \wedge B \neq \emptyset \wedge \forall_{a \in A} \forall_{b \in B}\, a \rightarrow_L b \wedge \forall_{a1,a2 \in A}\, a_1 \#_L a_2 \wedge \forall_{b1,b2 \in B}\, b_1 \#_L b_2 \}$,
5. $Y_L = \{ (A,B) \in X_L \mid \forall_{(A',B') \in X_L} A \subseteq A' \wedge B \subseteq B' \Rightarrow (A,B) = (A',B') \}$,
6. $P_L = \{ p_{(A,B)} \mid (A,B) \in Y_L \} \cup \{i_L, o_L\}$,
7. $F_L = \{ (a, p_{(A,B)}) \mid (A,B) \in Y_L \wedge a \in A \} \cup \{ (p_{(A,B)}, b) \mid (A,B) \in Y_L \wedge b \in B \} \cup \{ (i_L, t) \mid t \in T_I \} \cup \{ (t, o_L) \mid t \in T_O \}$, and
8. $\alpha(L) = (P_L, T_L, F_L)$.

The α–algorithm

Let L be an event log over T. Then, $\alpha(L)$ is defined as follows:

1. $T_L = \{ t \in T \mid \exists_{\sigma \in L}\, t \in \sigma \}$,
   Each activity in L corresponds to a transition in $\alpha(L)$.
2. $T_I = \{ t \in T \mid \exists_{\sigma \in L}\, t = \mathit{first}(\sigma) \}$
   Fix the set of start activities – that is, the first elements of each trace: $\langle t_1, \ldots, t_n \rangle, \ldots, \langle t'_1, \ldots, t'_m \rangle$
3. $T_O = \{ t \in T \mid \exists_{\sigma \in L}\, t = \mathit{last}(\sigma) \}$
   Fix the set of end activities – that is, elements that appear last at a trace : $\langle t_1, \ldots, t_n \rangle, \ldots, \langle t'_1, \ldots, t'_m \rangle$

Intuition next steps: Find places



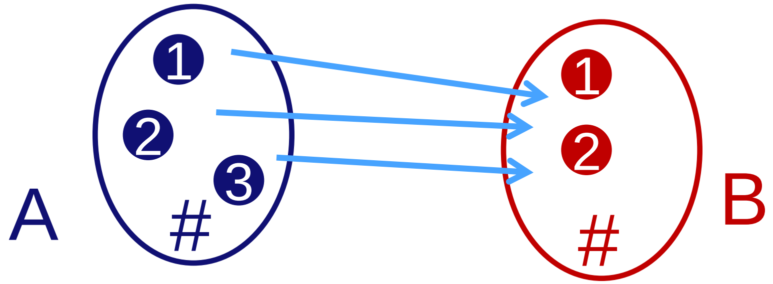
Step 4: Calculate pairs (A, B)

Step 5: Delete nonmaximal pairs (A, B)

Step 6: Determine places $p_{(A, B)}$ from pairs (A, B)

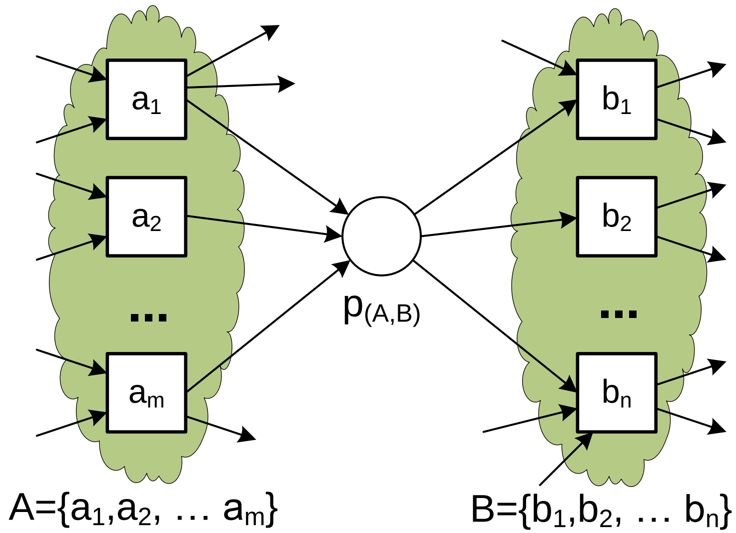
The α–algorithm (cont.)

4. $X_L = \{ (A,B) \mid A \subseteq T_L \wedge A \neq \emptyset \wedge B \subseteq T_L \wedge B \neq \emptyset$
$\wedge \forall_{a \in A} \forall_{b \in B}\; a \rightarrow_L b$
$\wedge \forall_{a1,a2 \in A}\; a_1 \#_L a_2$
$\wedge \forall_{b1,b2 \in B}\; b_1 \#_L b_2 \}$,



Find pairs (A, B) of sets of activities such that every element $a \in A$ and every element $b \in B$ are causally related (i.e., $a \rightarrow_L b$), all elements in A are independent ($a_1 \#_L a_2$), and all elements in B are independent ($b_1 \#_L b_2$).

Places as footprints

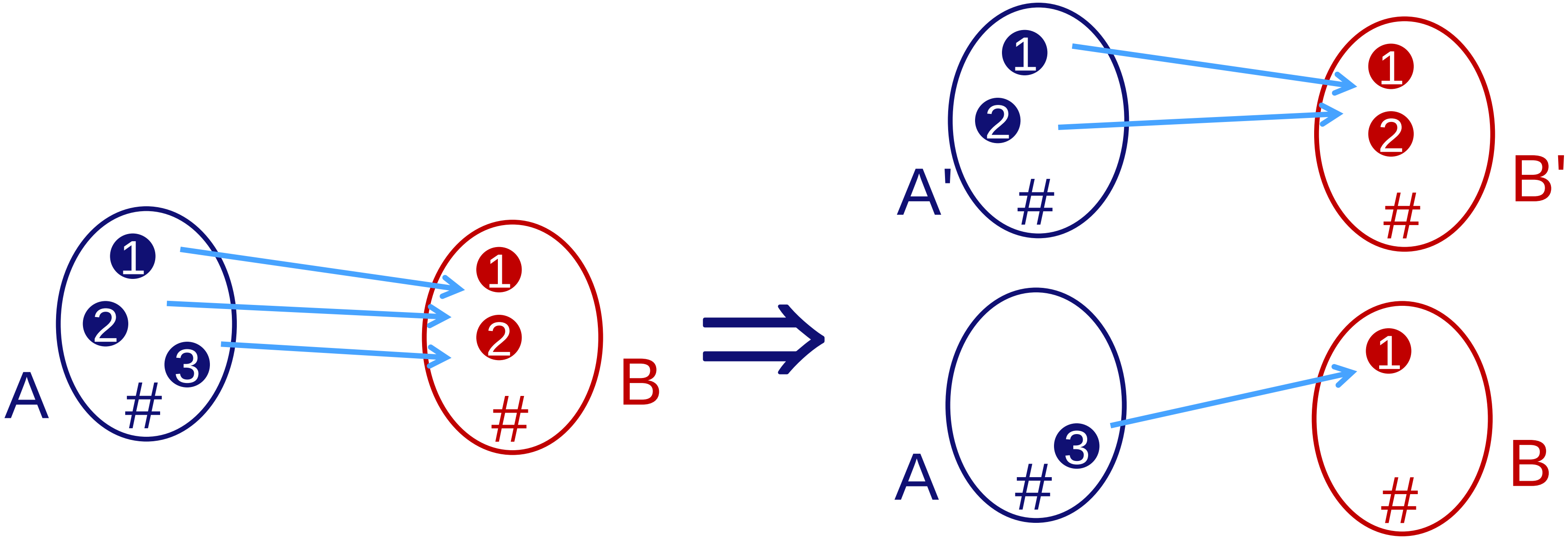


	a_1	a_2	$\ldots$	a_m	b_1	b_2	$\ldots$	b_n
a_1	#	#	$\ldots$	#	$\rightarrow$	$\rightarrow$	$\ldots$	$\rightarrow$
a_2	#	#	$\ldots$	#	$\rightarrow$	$\rightarrow$	$\ldots$	$\rightarrow$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
a_m	#	#	$\ldots$	#	$\rightarrow$	$\rightarrow$	$\ldots$	$\rightarrow$
b_1	$\leftarrow$	$\leftarrow$	$\ldots$	$\leftarrow$	#	#	$\ldots$	#
b_2	$\leftarrow$	$\leftarrow$	$\ldots$	$\leftarrow$	#	#	$\ldots$	#
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
b_n	$\leftarrow$	$\leftarrow$	$\ldots$	$\leftarrow$	#	#	$\ldots$	#

The α–algorithm (cont.)

5. $Y_L = \{ (A,B) \in X_L \mid \forall_{(A',B') \in X_L} A \subseteq A' \wedge B \subseteq B' \Rightarrow (A,B) = (A',B') \}$

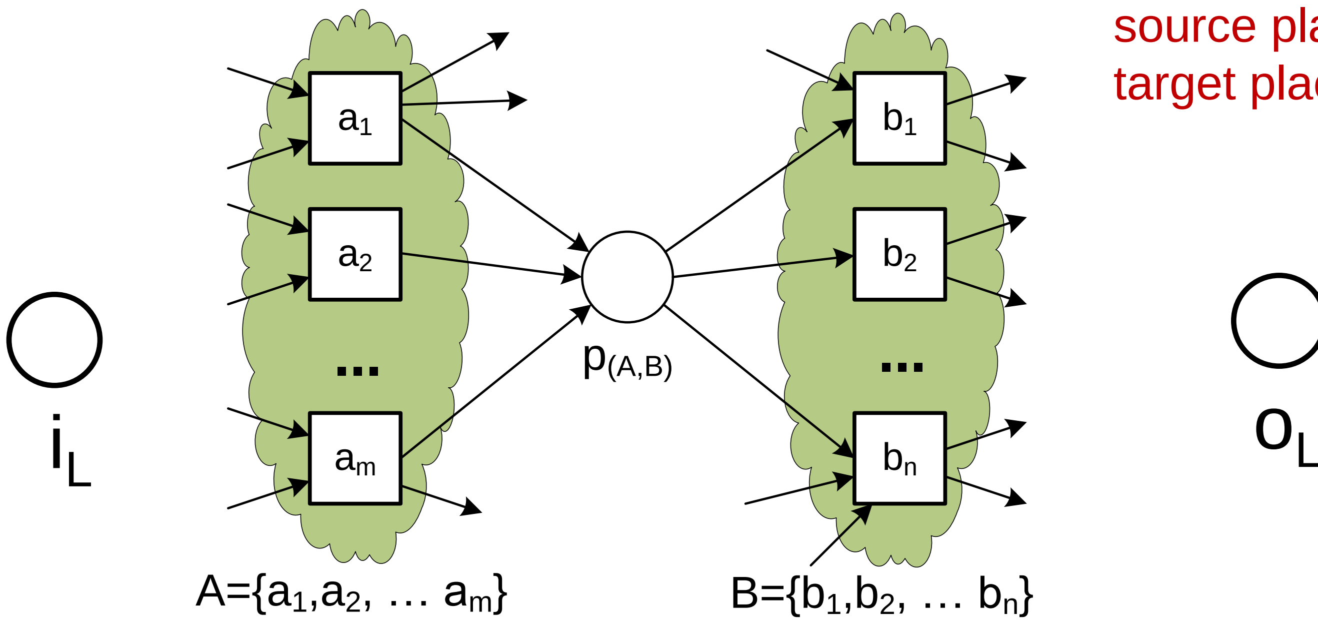
Delete from set X_L all pairs (A, B) that are not maximal!



The α–algorithm (cont.)

6. $P_L = \{ p_{(A,B)} \mid (A,B) \in Y_L \} \cup \{i_L, o_L\}$,

Determine the place set: Each element (A, B) of Y_L is a place. To ensure the workflow structure, add a source place i_L and a target place o_L



The α–algorithm (cont.)

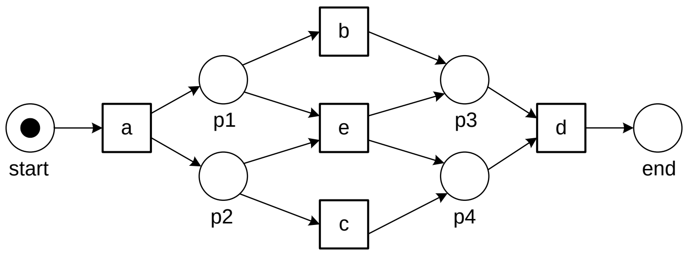
7. $F_L = \{ (a,p_{(A,B)}) \mid (A,B) \in Y_L \wedge a \in A \}$
   $\cup \{ (p_{(A,B)},b) \mid (A,B) \in Y_L \wedge b \in B \}$
   $\cup \{ (i_L,t) \mid t \in T_I \} \cup \{ (t,o_L) \mid t \in T_O \}$

   Determine the flow relation: Connect each place $p_{(A,B)}$ with each element a of its set A of source transitions and with each element of its set B of target transitions. In addition, draw an arc from the source place i_L to each start transition $t \in T_I$ and an arc from each end transition $t \in T_O$ to the sink place o_L.

8. $\alpha(L) = (P_L, T_L, F_L)$

$$L_1 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^2, \langle a,e,d\rangle]$$

	a	b	c	d	e
a	$\#_{L_1}$	$\rightarrow_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$
b	$\leftarrow_{L_1}$	$\#_{L_1}$	$\parallel_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
c	$\leftarrow_{L_1}$	$\parallel_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
d	$\#_{L_1}$	$\leftarrow_{L_1}$	$\leftarrow_{L_1}$	$\#_{L_1}$	$\leftarrow_{L_1}$
e	$\leftarrow_{L_1}$	$\#_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$



$$X_{L_1} = \{(\{a\},\{b\}),(\{a\},\{c\}),(\{a\},\{e\}),(\{a\},\{b,e\}),(\{a\},\{c,e\}),(\{b\},\{d\}),(\{c\},\{d\}),(\{e\},\{d\}),(\{b,e\},\{d\}),(\{c,e\},\{d\})\}$$

$$Y_{L_1} = \{(\{a\},\{b,e\}),(\{a\},\{c,e\}),(\{b,e\},\{d\}),(\{c,e\},\{d\})\}$$

Another event log L_3

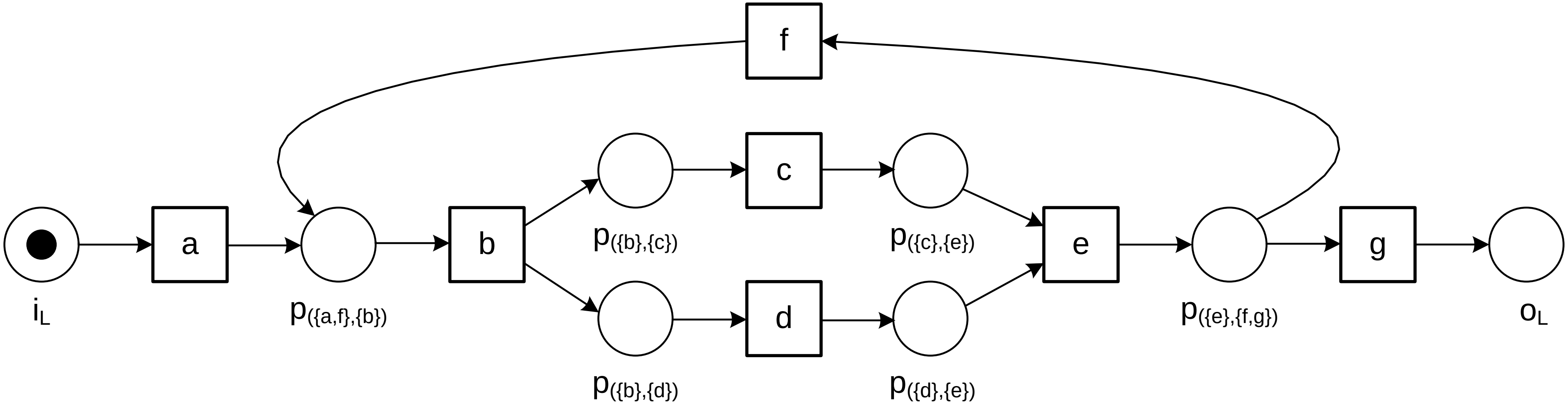
$$L_3 = [\langle a,b,c,d,e,f,b,d,c,e,g\rangle, \\ \langle a,b,d,c,e,g\rangle^2, \\ \langle a,b,c,d,e,f,b,c,d,e,f,b,d,c,e,g\rangle]$$

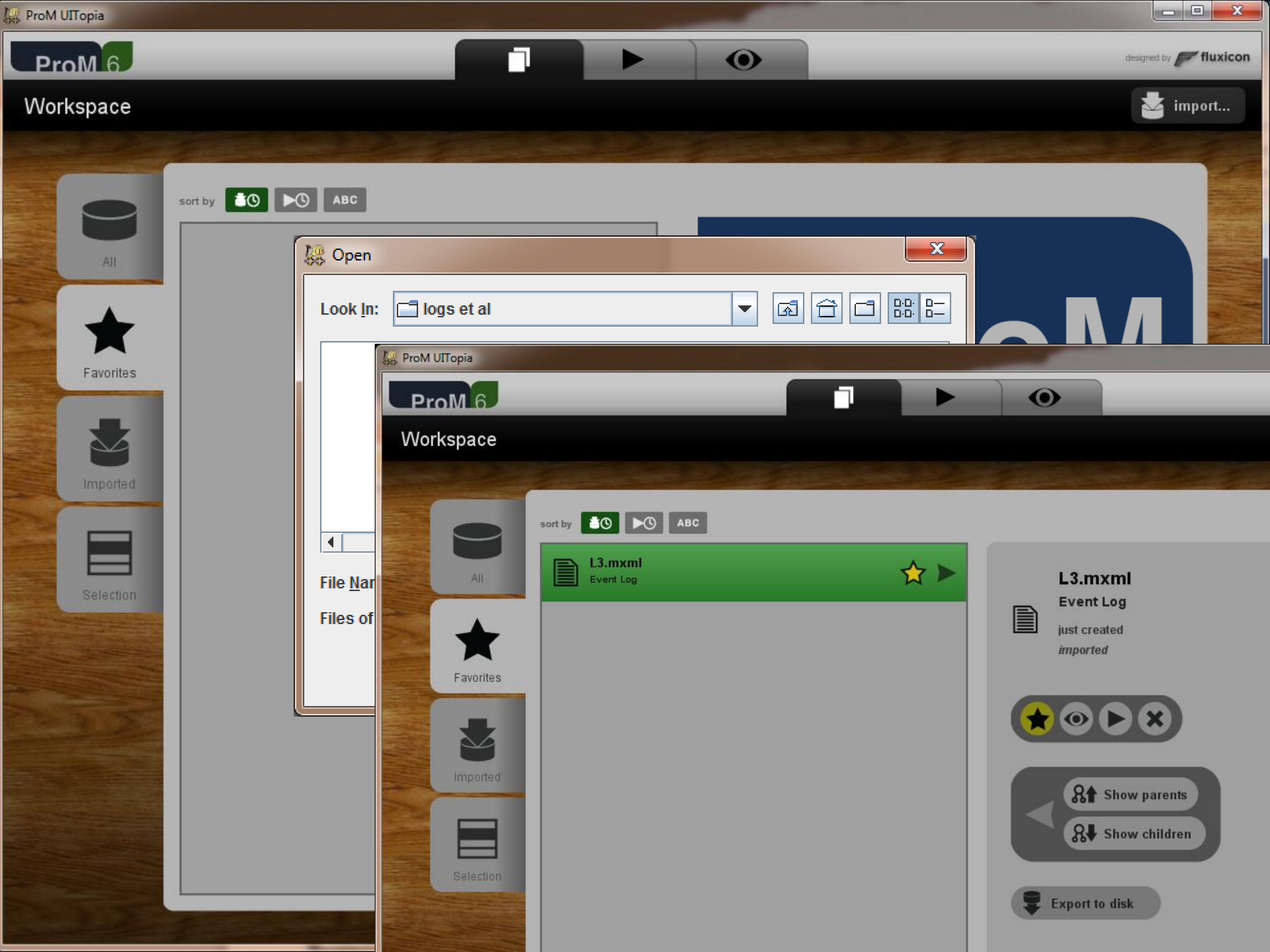
	a	b	c	d	e	f	g
a	$\#$	$\rightarrow$	$\#$	$\#$	$\#$	$\#$	$\#$
b	$\leftarrow$	$\#$	$\rightarrow$	$\rightarrow$	$\#$	$\leftarrow$	$\#$
c	$\#$	$\leftarrow$	$\#$	$\parallel$	$\rightarrow$	$\#$	$\#$
d	$\#$	$\leftarrow$	$\parallel$	$\#$	$\rightarrow$	$\#$	$\#$
e	$\#$	$\#$	$\leftarrow$	$\leftarrow$	$\#$	$\rightarrow$	$\rightarrow$
f	$\#$	$\rightarrow$	$\#$	$\#$	$\leftarrow$	$\#$	$\#$
g	$\#$	$\#$	$\#$	$\#$	$\leftarrow$	$\#$	$\#$

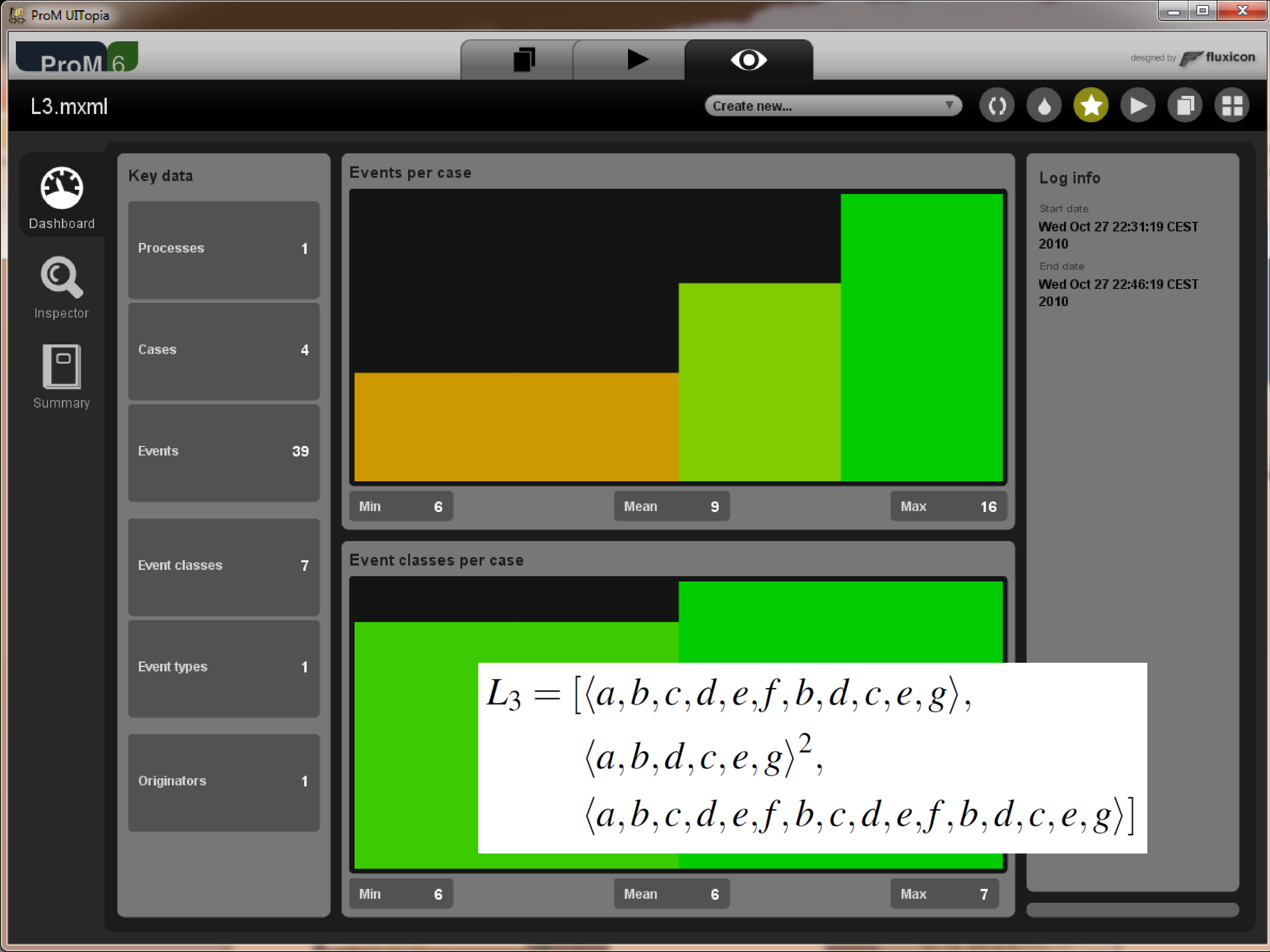
Model for L_3

	a	b	c	d	e	f	g
a	#	→	#	#	#	#	#
b	←	#	→	→	#	←	#
c	#	←	#	‖	→	#	#
d	#	←	‖	#	→	#	#
e	#	#	←	←	#	→	→
f	#	→	#	#	←	#	#
g	#	#	#	#	←	#	#

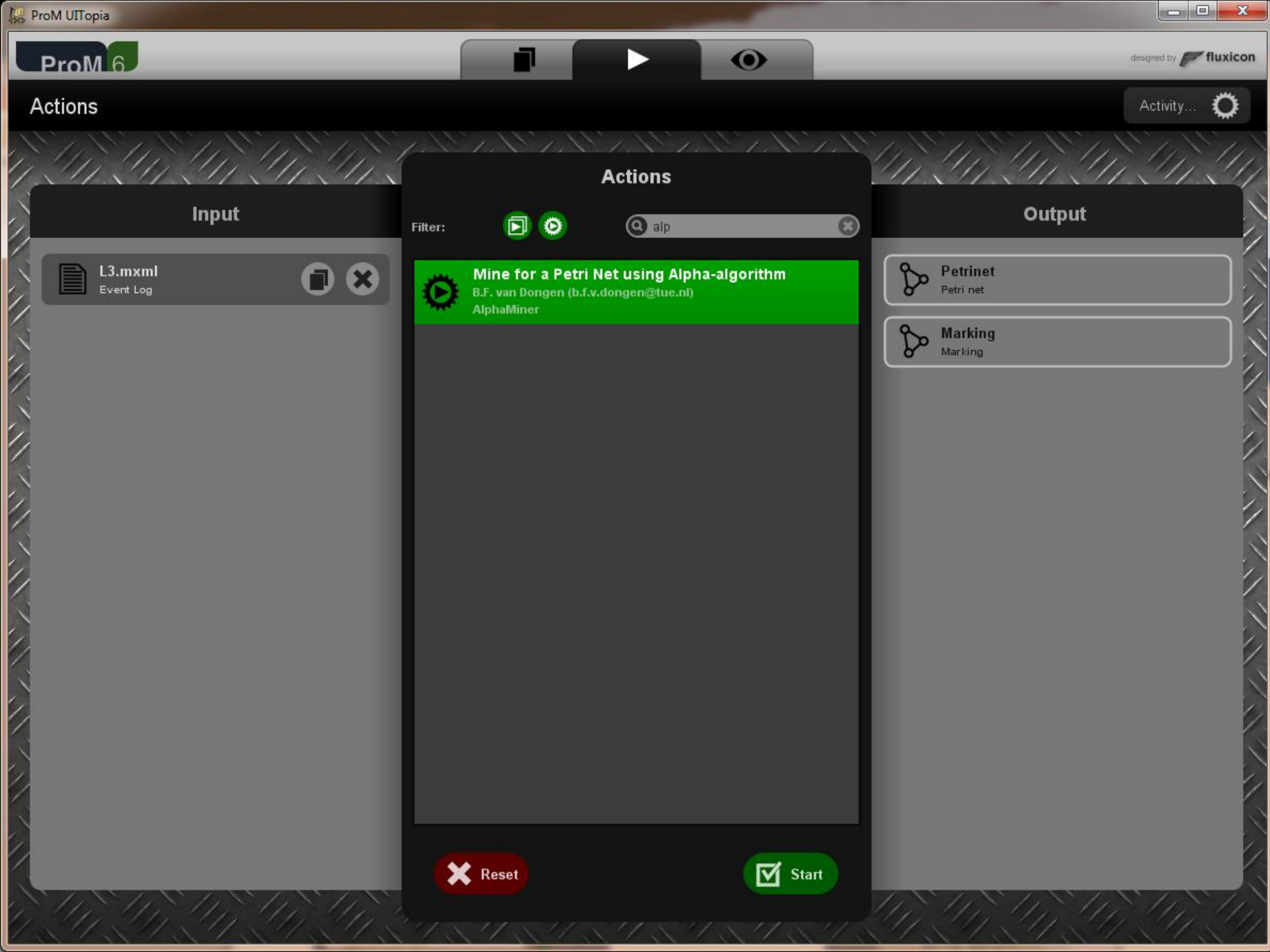
$$L_3 = [\langle a,b,c,d,e,f,b,d,c,e,g\rangle,$$
$$\langle a,b,d,c,e,g\rangle^2,$$
$$\langle a,b,c,d,e,f,b,c,d,e,f,b,d,c,e,g\rangle]$$

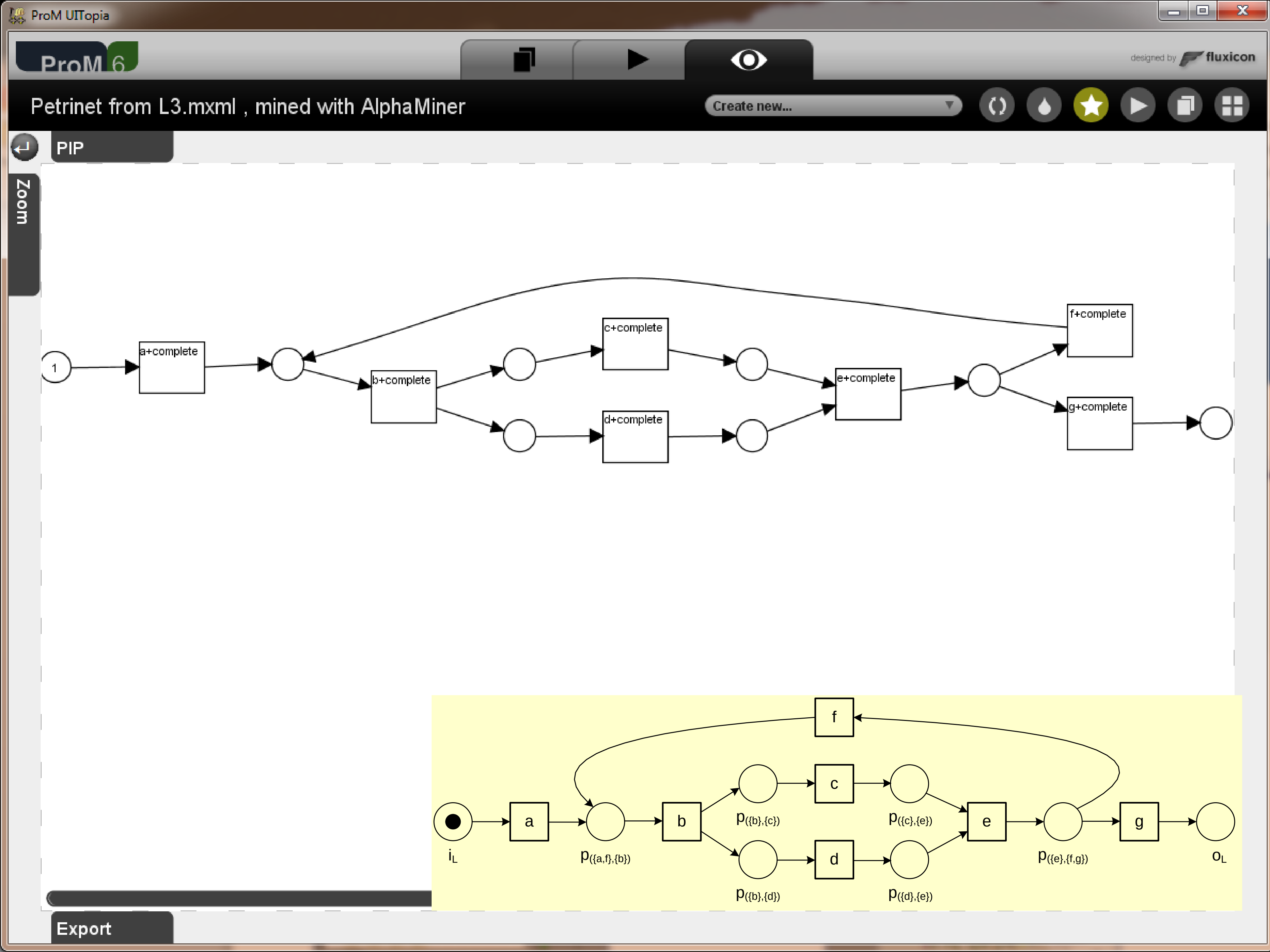



ProM UITopia
ProM 6
designed by fluxicon
Workspace
import...
sort by
ABC
All
Favorites
Imported
Selection
Open
Look In:
logs et al
File Nar
Files of
ProM UITopia
ProM 6
Workspace
sort by
ABC
L3.mxml
Event Log
All
Favorites
Imported
Selection
L3.mxml
Event Log
just created
imported
Show parents
Show children
Export to disk



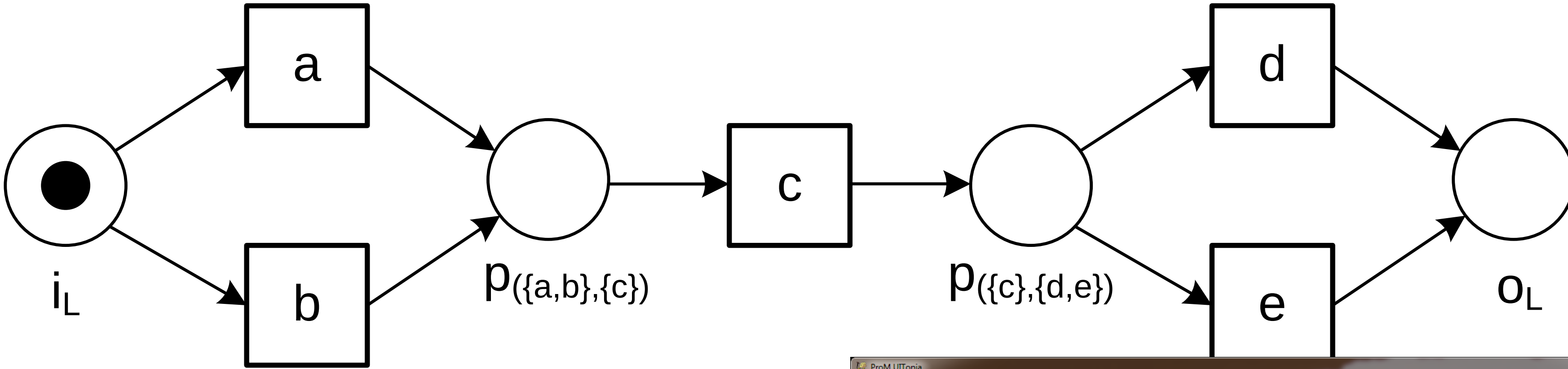
$$L_3 = [\langle a,b,c,d,e,f,b,d,c,e,g \rangle, \langle a,b,d,c,e,g \rangle^2, \langle a,b,c,d,e,f,b,c,d,e,f,b,d,c,e,g \rangle]$$

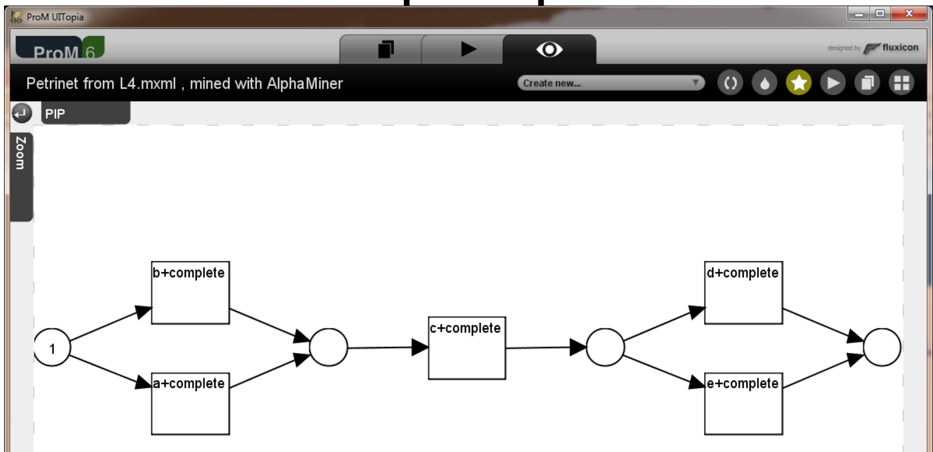

ProM UITopia
ProM 6
designed by fluxicon
Actions
Activity...
Actions
Input
Output
Filter:
alp
L3.mxml
Event Log
Mine for a Petri Net using Alpha-algorithm
B.F. van Dongen (b.f.v.dongen@tue.nl)
AlphaMiner
Petrinet
Petri net
Marking
Marking
Reset
Start


ProM UITopia
ProM 6
designed by fluxicon
Petrinet from L3.mxml , mined with AlphaMiner
Create new...
PIP
Zoom
1
a+complete
b+complete
c+complete
d+complete
e+complete
f+complete
g+complete
f
a
b
c
d
e
g
i_L
$p_{(\{a,f\},\{b\})}$
$p_{(\{b\},\{c\})}$
$p_{(\{b\},\{d\})}$
$p_{(\{c\},\{e\})}$
$p_{(\{d\},\{e\})}$
$p_{(\{e\},\{f,g\})}$
o_L
Export

Another event log L_4

$$L_4 = [\langle a,c,d\rangle^{45}, \langle b,c,d\rangle^{42}, \langle a,c,e\rangle^{38}, \langle b,c,e\rangle^{22}]$$





Event log L_5

$$L_5 = [\langle a,b,e,f\rangle^2, \langle a,b,e,c,d,b,f\rangle^3, \langle a,b,c,e,d,b,f\rangle^2, \langle a,b,c,d,e,b,f\rangle^4, \langle a,e,b,c,d,b,f\rangle^3]$$

	a	b	c	d	e	f
a	$\#$	$\rightarrow$	$\#$	$\#$	$\rightarrow$	$\#$
b	$\leftarrow$	$\#$	$\rightarrow$	$\leftarrow$	$\|$	$\rightarrow$
c	$\#$	$\leftarrow$	$\#$	$\rightarrow$	$\|$	$\#$
d	$\#$	$\rightarrow$	$\leftarrow$	$\#$	$\|$	$\#$
e	$\leftarrow$	$\|$	$\|$	$\|$	$\#$	$\rightarrow$
f	$\#$	$\leftarrow$	$\#$	$\#$	$\leftarrow$	$\#$

$$T_L = \{a,b,c,d,e,f\}$$

$$T_I = \{a\}$$

$$T_I = \{f\}$$

$$X_L = \{(\{a\},\{b\}),(\{a\},\{e\}),(\{b\},\{c\}),(\{b\},\{f\}),(\{c\},\{d\}),$$
$$(\{d\},\{b\}),(\{e\},\{f\}),(\{a,d\},\{b\}),(\{b\},\{c,f\})\}$$

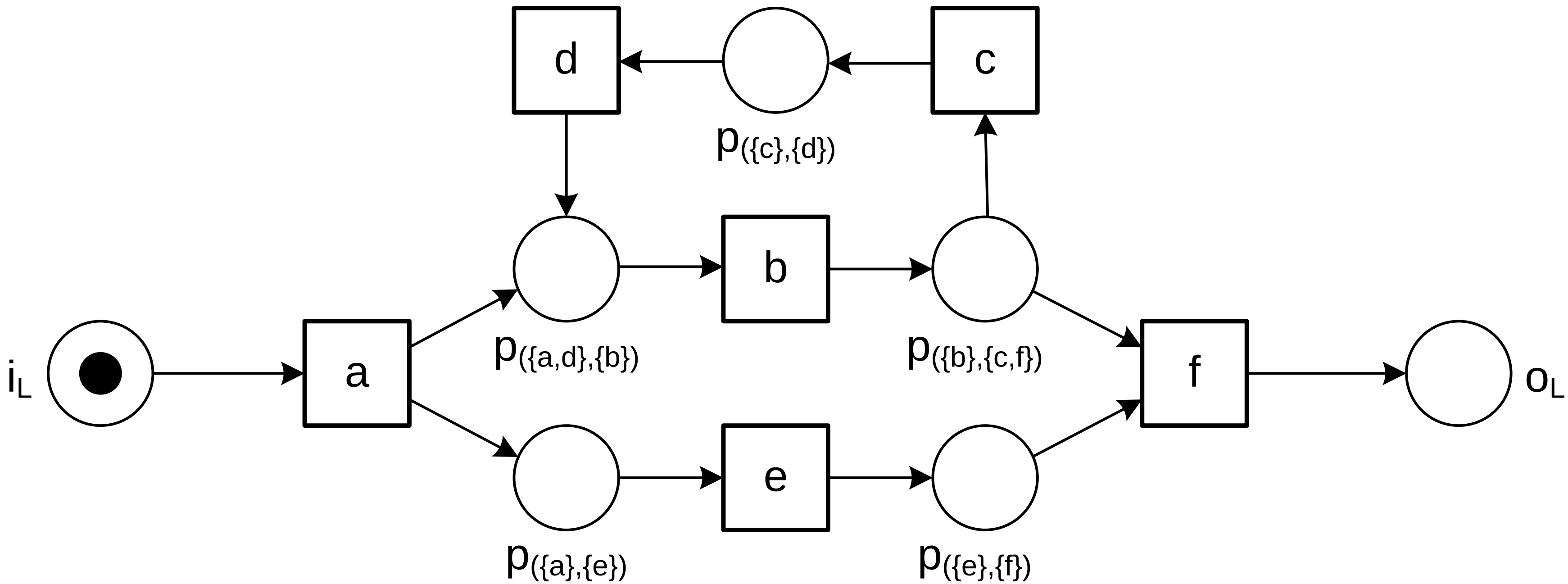
$$Y_L = \{(\{a\},\{e\}),(\{c\},\{d\}),(\{e\},\{f\}),(\{a,d\},\{b\}),(\{b\},\{c,f\})\}$$

$$P_L = \{p_{(\{a\},\{e\})},p_{(\{c\},\{d\})},p_{(\{e\},\{f\})},p_{(\{a,d\},\{b\})},p_{(\{b\},\{c,f\})},i_L,o_L\}$$

$$F_L = \{(a,p_{(\{a\},\{e\})}),(p_{(\{a\},\{e\})},e),(c,p_{(\{c\},\{d\})}),(p_{(\{c\},\{d\})},d),$$
$$(e,p_{(\{e\},\{f\})}),(p_{(\{e\},\{f\})},f),(a,p_{(\{a,d\},\{b\})}),(d,p_{(\{a,d\},\{b\})}),$$
$$(p_{(\{a,d\},\{b\})},b),(b,p_{(\{b\},\{c,f\})}),(p_{(\{b\},\{c,f\})},c),(p_{(\{b\},\{c,f\})},f),$$
$$(i_L,a),(f,o_L)\}$$

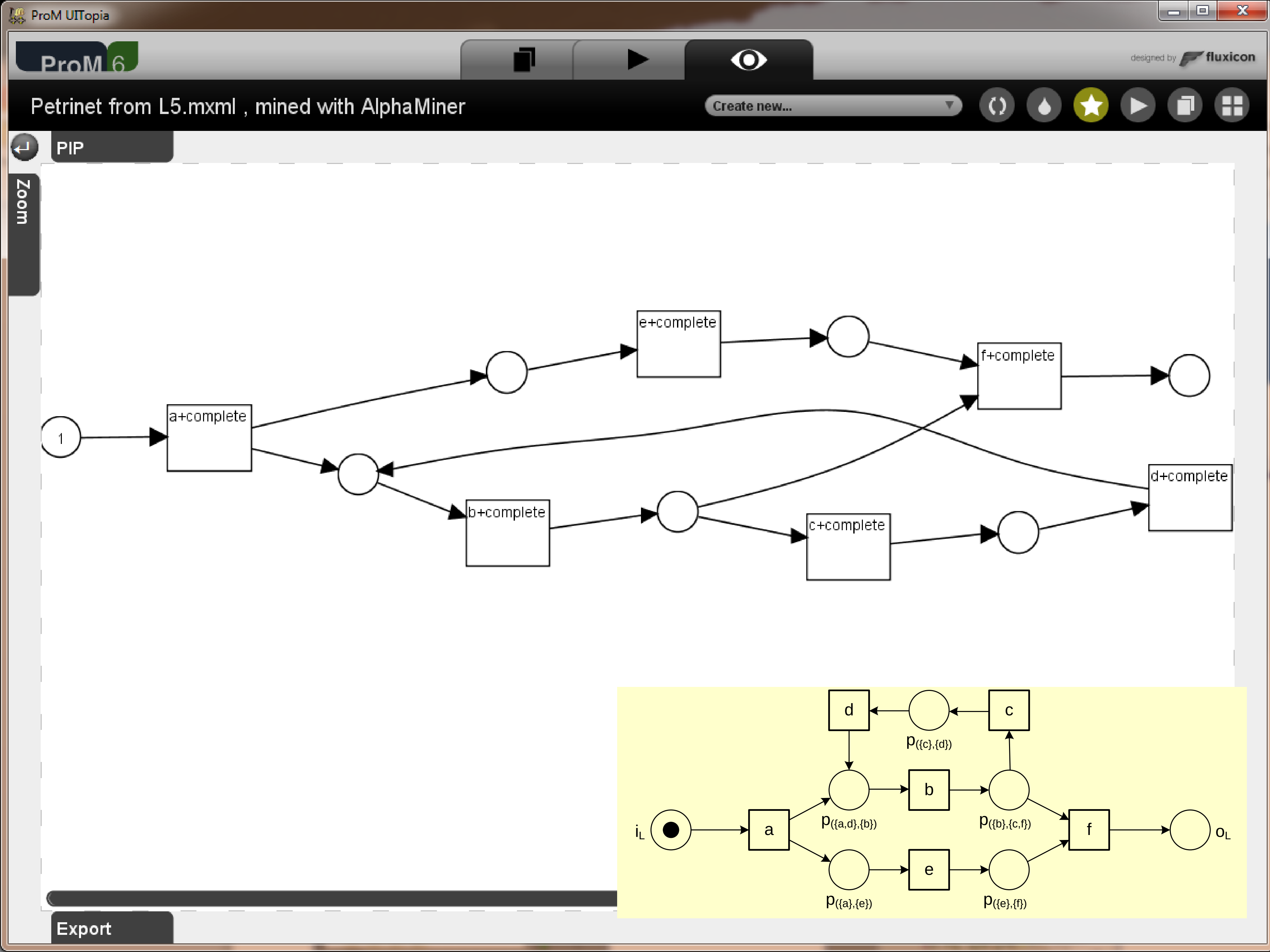
$$\alpha(L) = (P_L,T_L,F_L)$$

Discovered model



$$X_L = \{(\{a\},\{b\}),(\{a\},\{e\}),(\{b\},\{c\}),(\{b\},\{f\}),(\{c\},\{d\}),$$
$$(\{d\},\{b\}),(\{e\},\{f\}),(\{a,d\},\{b\}),(\{b\},\{c,f\})\}$$

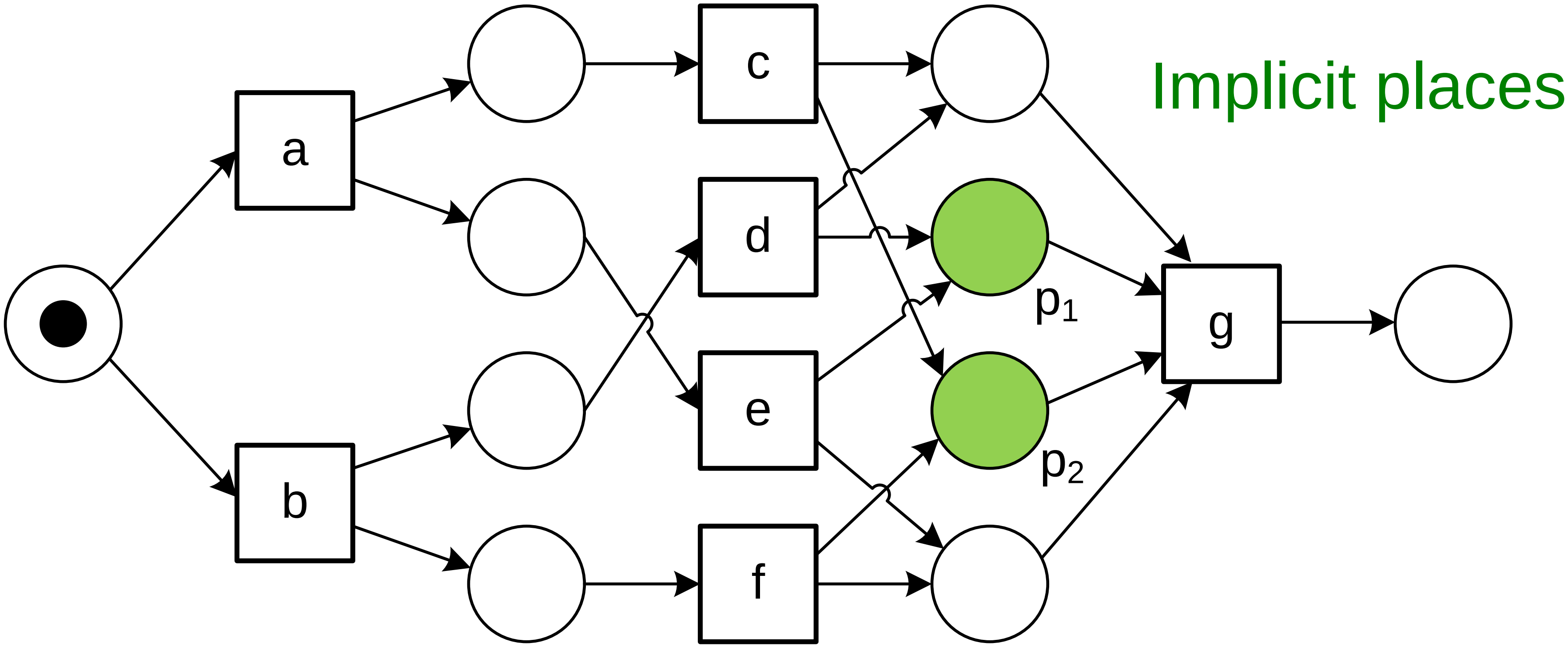
$$Y_L = \{(\{a\},\{e\}),(\{c\},\{d\}),(\{e\},\{f\}),(\{a,d\},\{b\}),(\{b\},\{c,f\})\}$$



ProM UITopia
ProM 6
designed by fluxicon
Petrinet from L5.mxml , mined with AlphaMiner
Create new...
PIP
Zoom
Export
1
a+complete
b+complete
c+complete
d+complete
e+complete
f+complete
d
c
b
a
e
f
$p_{(\{c\},\{d\})}$
$p_{(\{a,d\},\{b\})}$
$p_{(\{b\},\{c,f\})}$
$p_{(\{a\},\{e\})}$
$p_{(\{e\},\{f\})}$
i_L
o_L

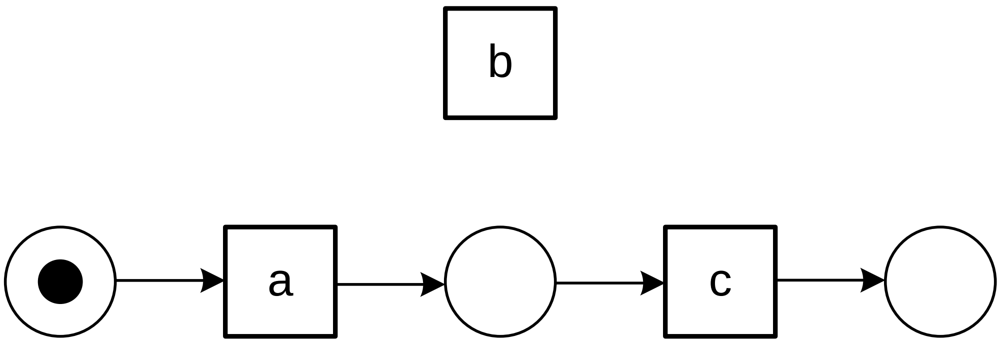
Limitation of α algorithm: Implicit places

$$L_6 = [\langle a,c,e,g\rangle^2, \langle a,e,c,g\rangle^3, \langle b,d,f,g\rangle^2, \langle b,f,d,g\rangle^4]$$



Limitation of α algorithm: Loops of length 1

$$L_7 = [\langle a,c\rangle^2, \langle a,b,c\rangle^3, \langle a,b,b,c\rangle^2, \langle a,b,b,b,b,c\rangle^1]$$

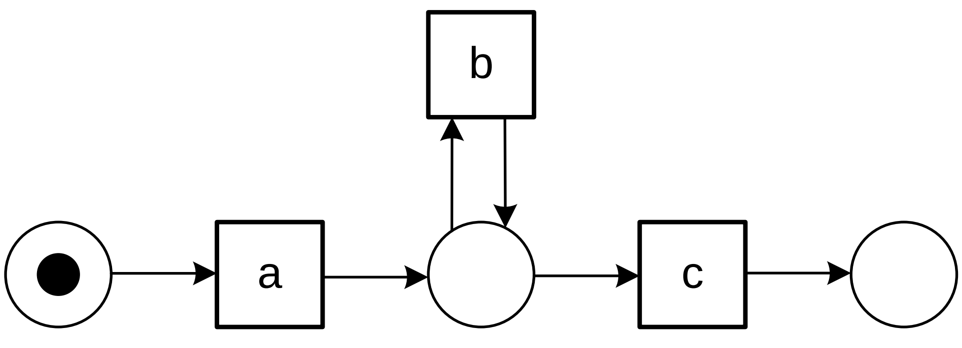


a>b
a>c
b>b
b>c

a→b
a→c
b→c

b||b

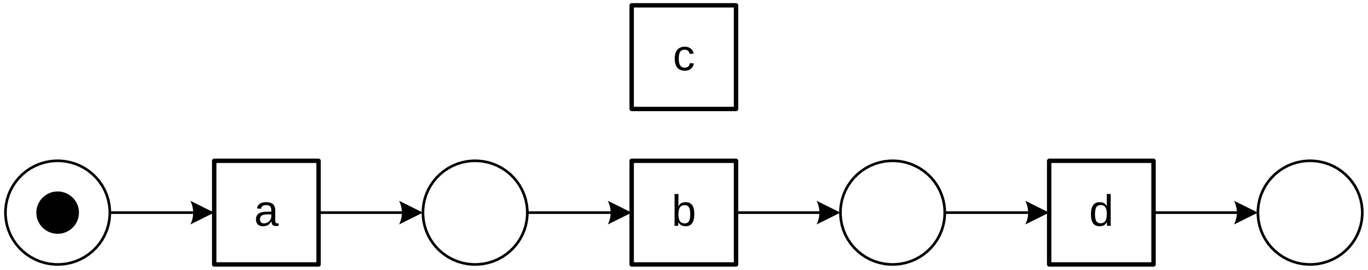
a#a
c#c
…



desired model

Limitation of α algorithm: Loops of length 2

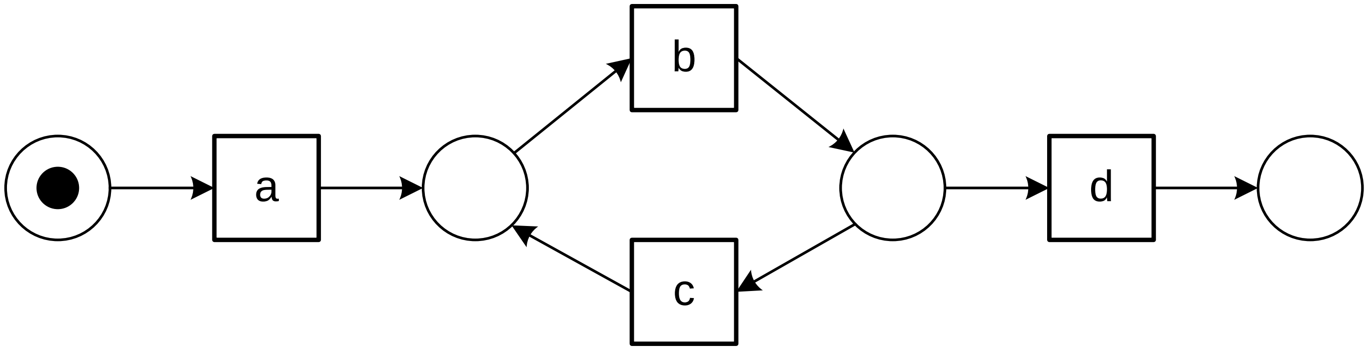
$$L_8 = [\langle a,b,d\rangle^3, \langle a,b,c,b,d\rangle^2, \langle a,b,c,b,c,b,d\rangle]$$



a>b
b>c
b>d
c>b

a→b
b→d

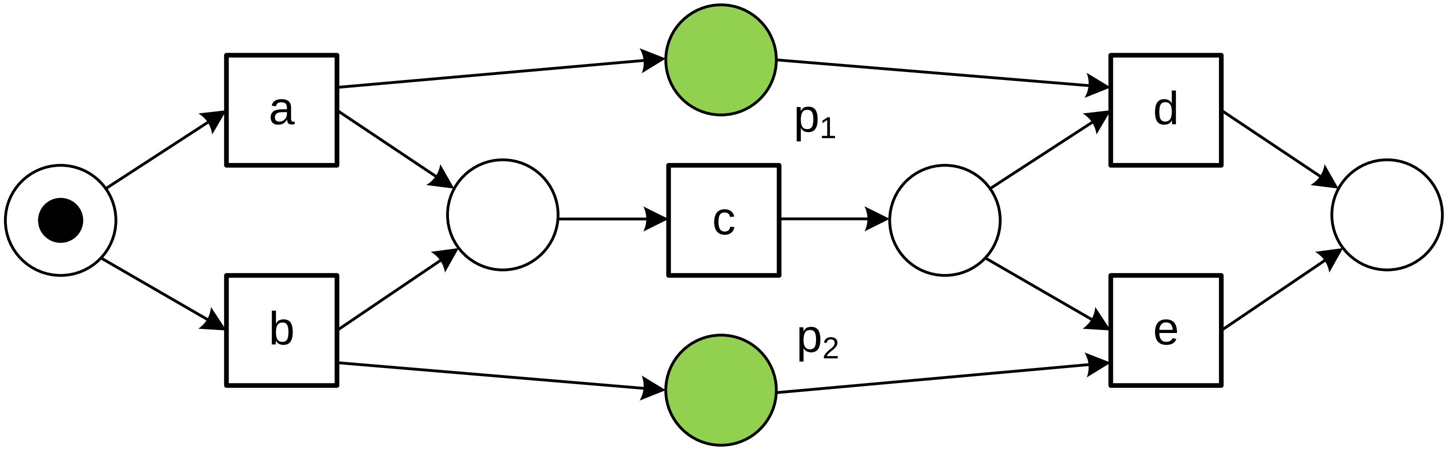
b||c



desired model

Limitation of α algorithm: Nonlocal dependencies

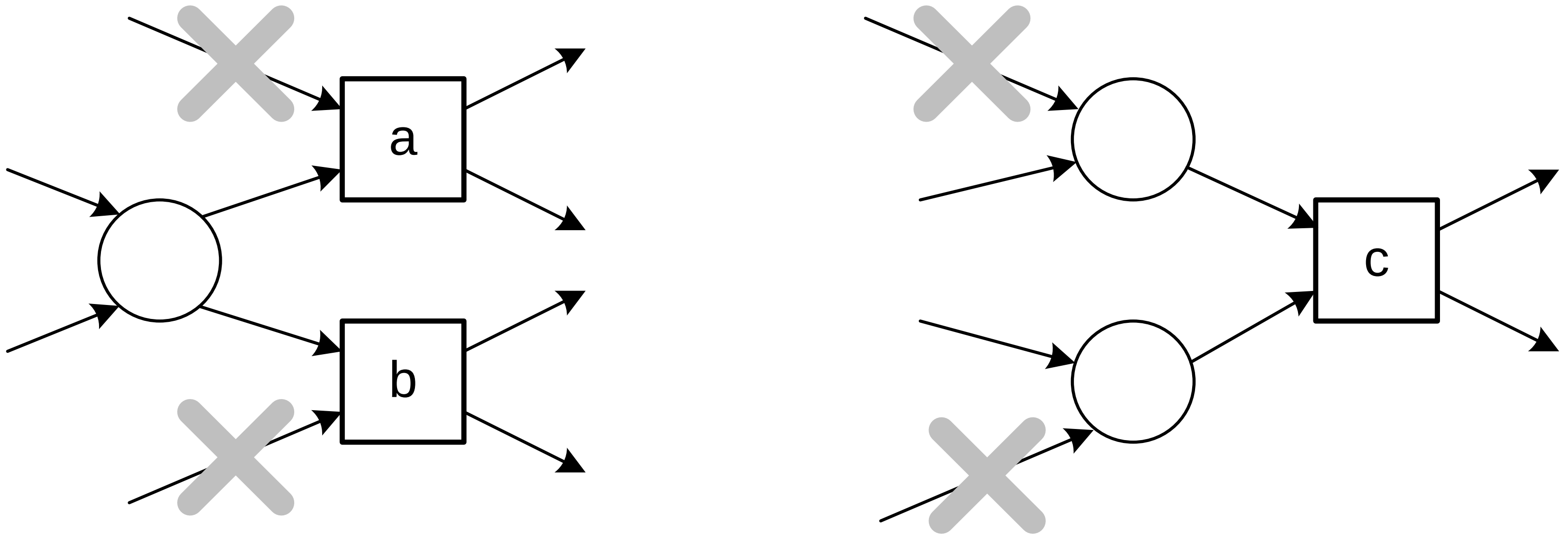
$$L_9 = [\langle a,c,d\rangle^{45}, \langle b,c,e\rangle^{42}]$$



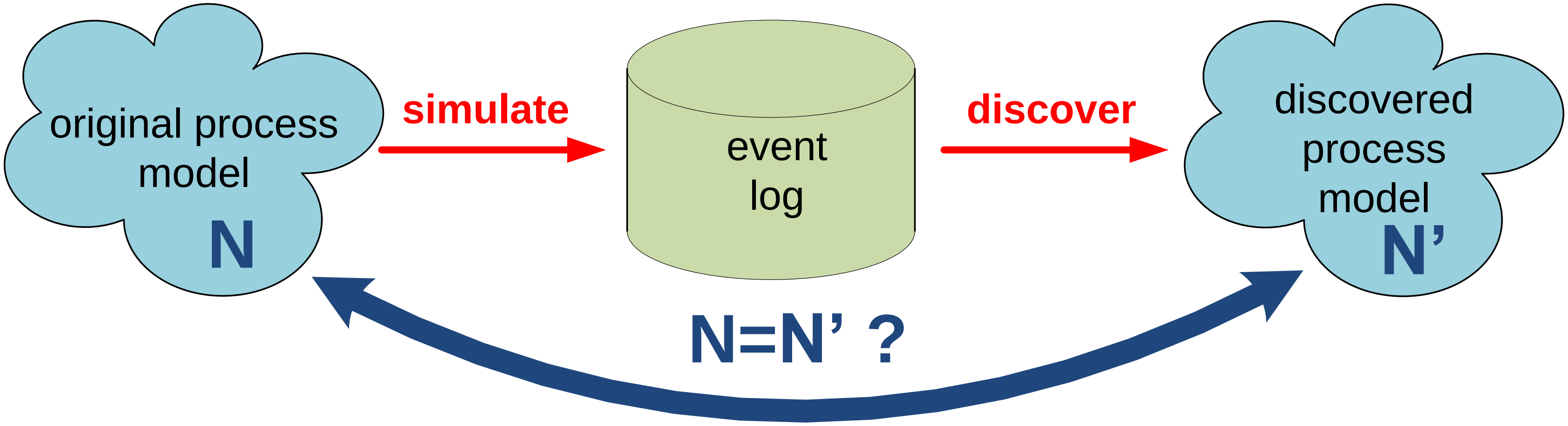
Not discovered!

$$L_4 = [\langle a,c,d\rangle^{45}, \langle b,c,d\rangle^{42}, \langle a,c,e\rangle^{38}, \langle b,c,e\rangle^{22}]$$

Difficult constructs for α algorithm



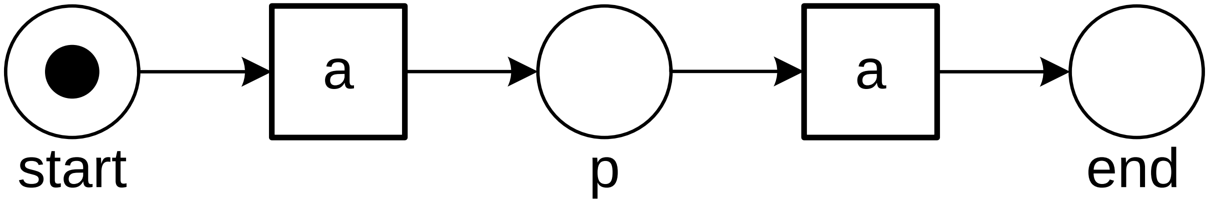
Rediscovering process models



The rediscovery problem:
Is the discovered model N′ equivalent to the original model N?

Challenge: Finding the right representational bias

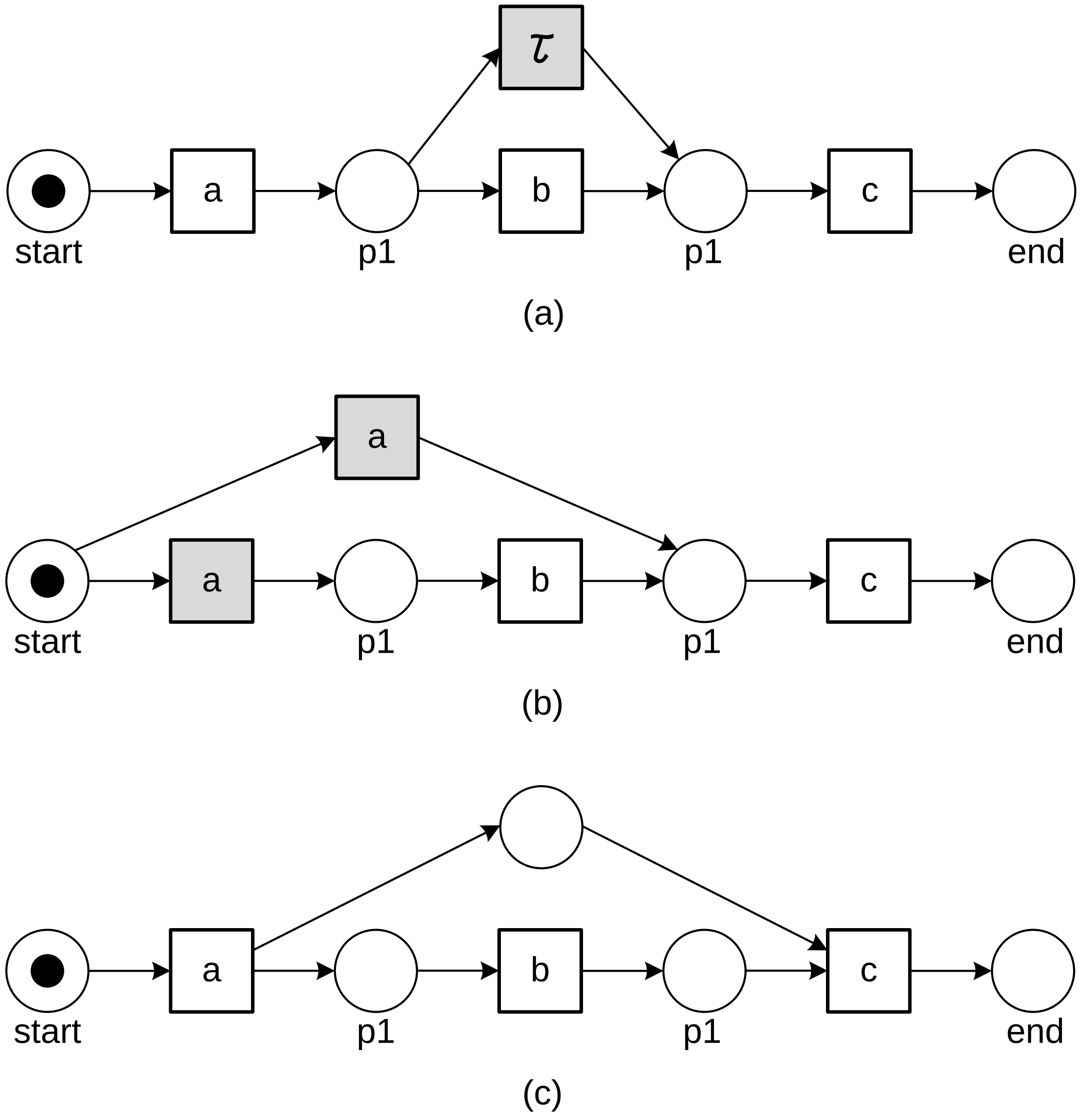
$$L_{10} = [\langle a, a \rangle^{55}]$$



There is no WF-net with unique visible labels that exhibits this behavior.

Another example

$$L_{11} = [\langle a,b,c\rangle^{20}, \langle a,c\rangle^{30}]$$



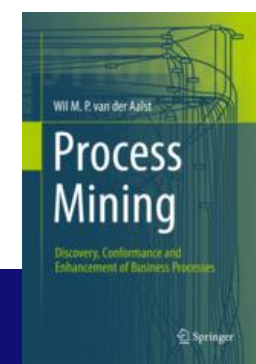
There is no WF-net with unique visible labels that exhibits this behavior.

Challenge: noise and incompleteness

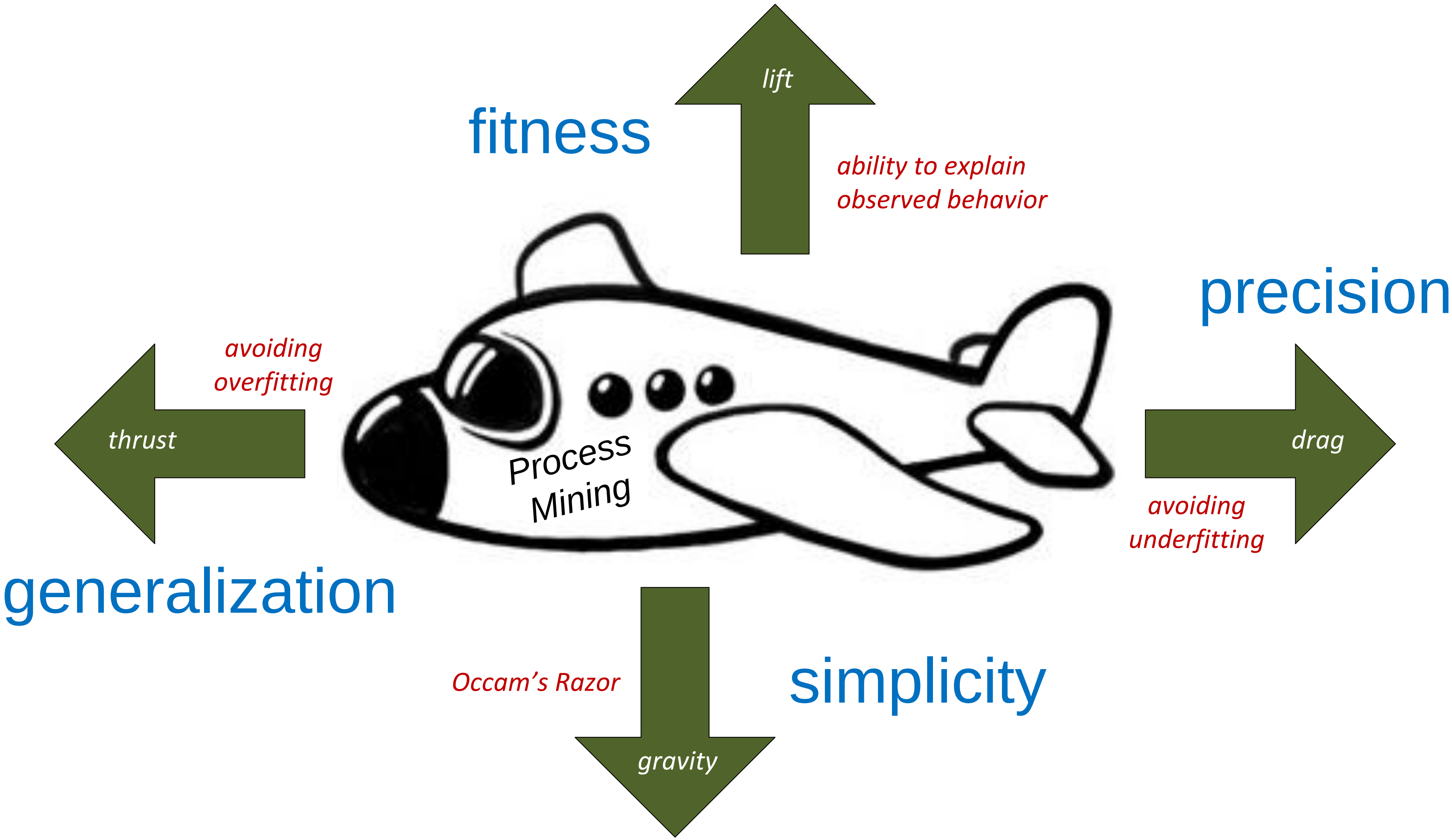
- To discover a suitable process model it is assumed that the event log contains a representative sample of behavior.
- Two related phenomena:
  - Noise: the event log contains rare and infrequent behavior not representative for the typical behavior of the process.
  - Incompleteness: the event log contains too few events to be able to discover some of the underlying control-flow structures.

More on incompleteness

To illustrate the relevance of completeness, consider a process consisting of 10 activities that can be executed in parallel and a corresponding log that contains information about 10,000 cases. The total number of possible interleavings in the model with 10 concurrent activities is $10! = 3{,}628{,}800$. Hence, it is impossible that each interleaving is present in the log as there are fewer cases (10,000) than potential traces (3,628,800). Even if there are 3,628,800 cases in the log, it is extremely unlikely that all possible variations are present. To motivate this consider the following analogy. In a group of 365 people it is very unlikely that everyone has a different birthdate. The probability is $365!/365^{365} \approx 1.454955 \times 10^{-157} \approx 0$, i.e., incredibly small. The number of atoms in the universe is often estimated to be approximately 10^{79} [129].

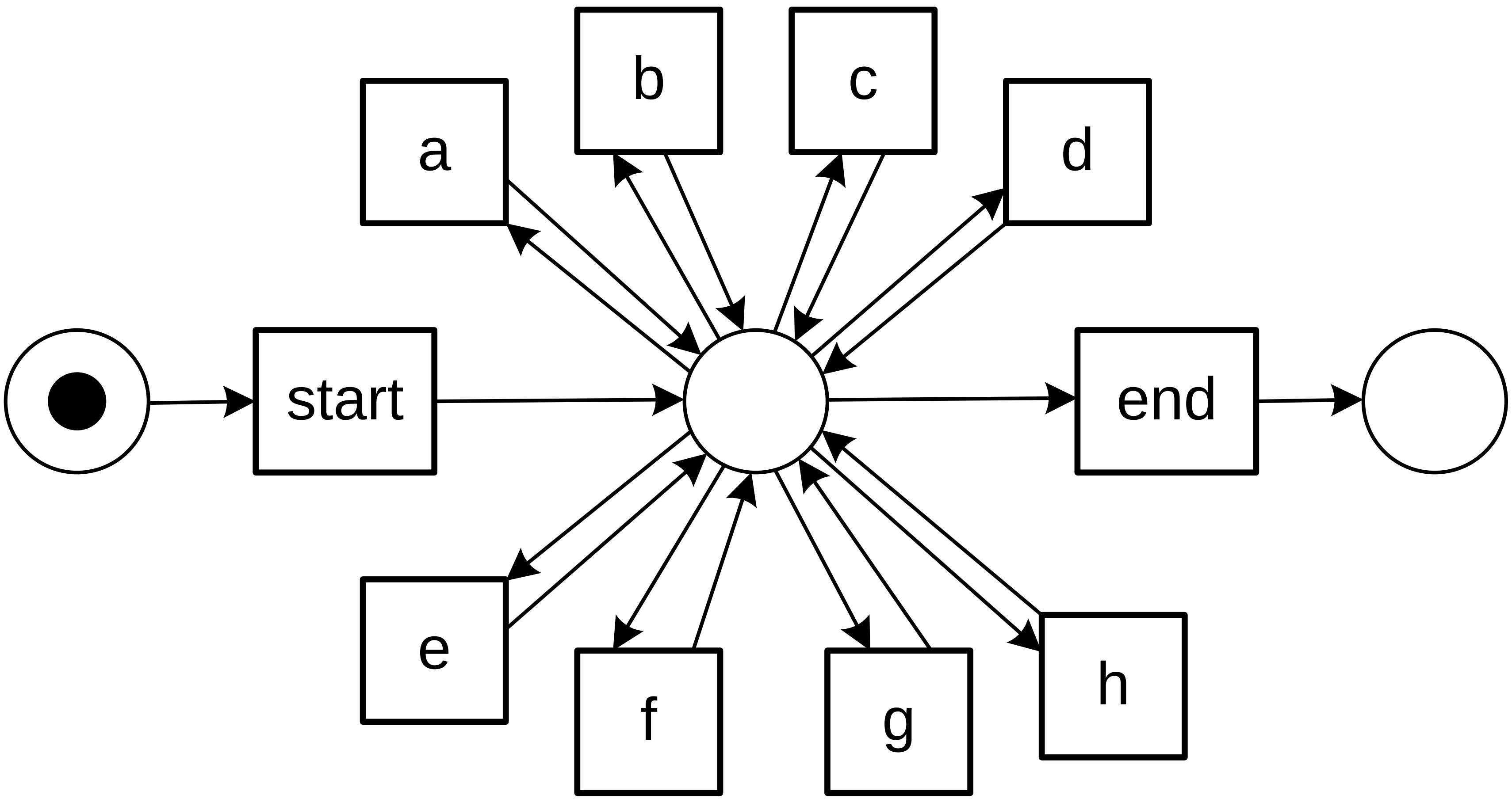


Balance four forces



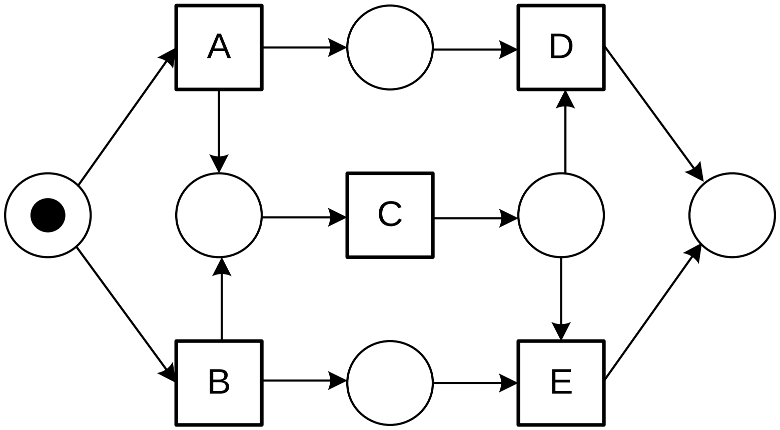
Challenge: Balancing Between Underfitting and Overfitting

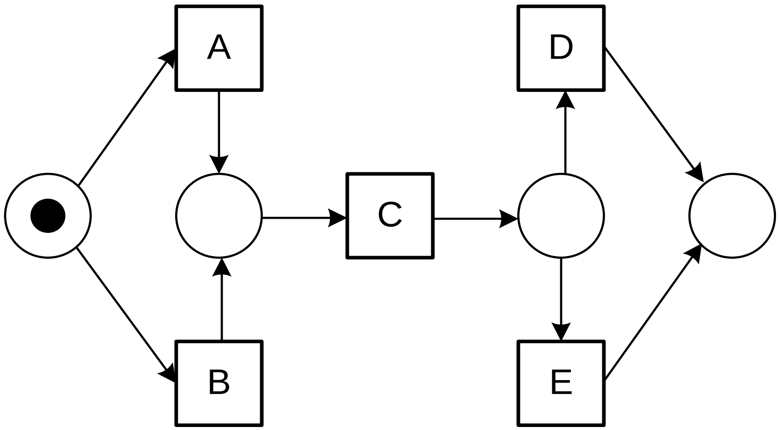
Flower model



What is the best model?

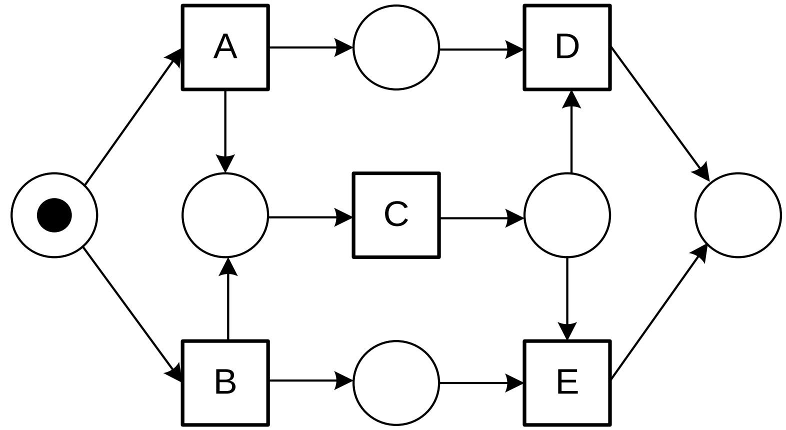
ACD	**99**
ACE	**0**
BCE	**85**
BCD	**0**

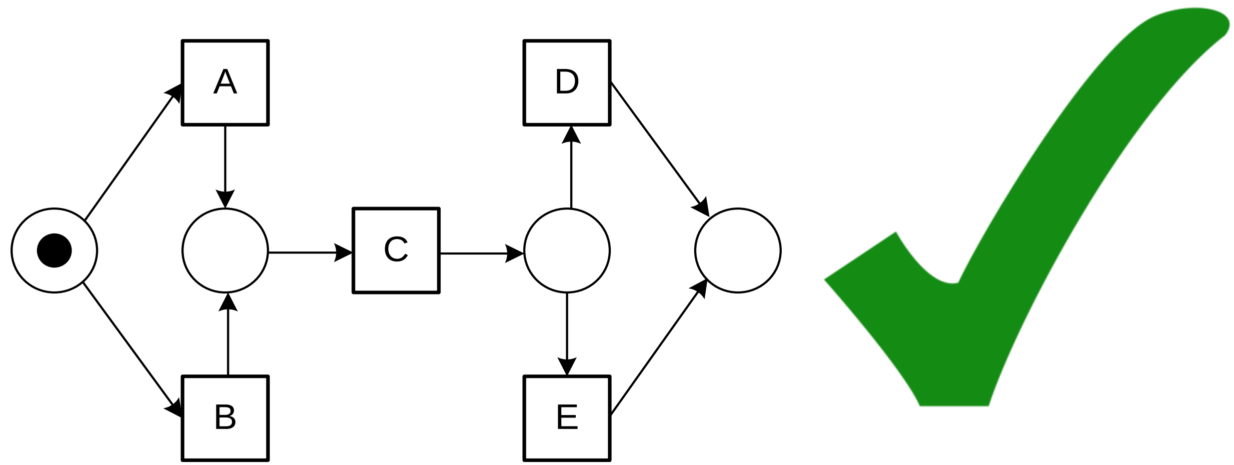




What is the best model?

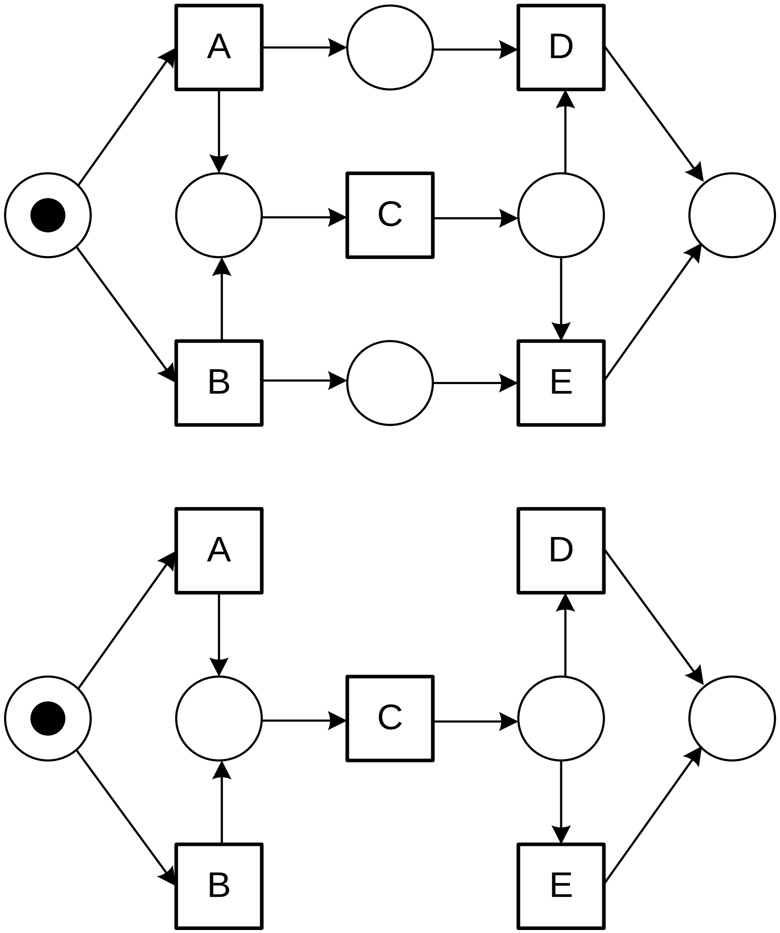
ACD	**99**
ACE	**88**
BCE	**85**
BCD	**78**





What is the best model?

ACD	**99**
ACE	**2**
BCE	**85**
BCD	**3**



α algorithm is just a starting point …

distributed genetic mining

automata-based learning

language-based regions

heuristic mining

state-based regions

partial-order based mining

genetic mining

LTL mining

pattern-based mining

neural networks

stochastic task graphs

fuzzy mining

mining block structures

hidden Markov models

α algorithm

multi-phase mining

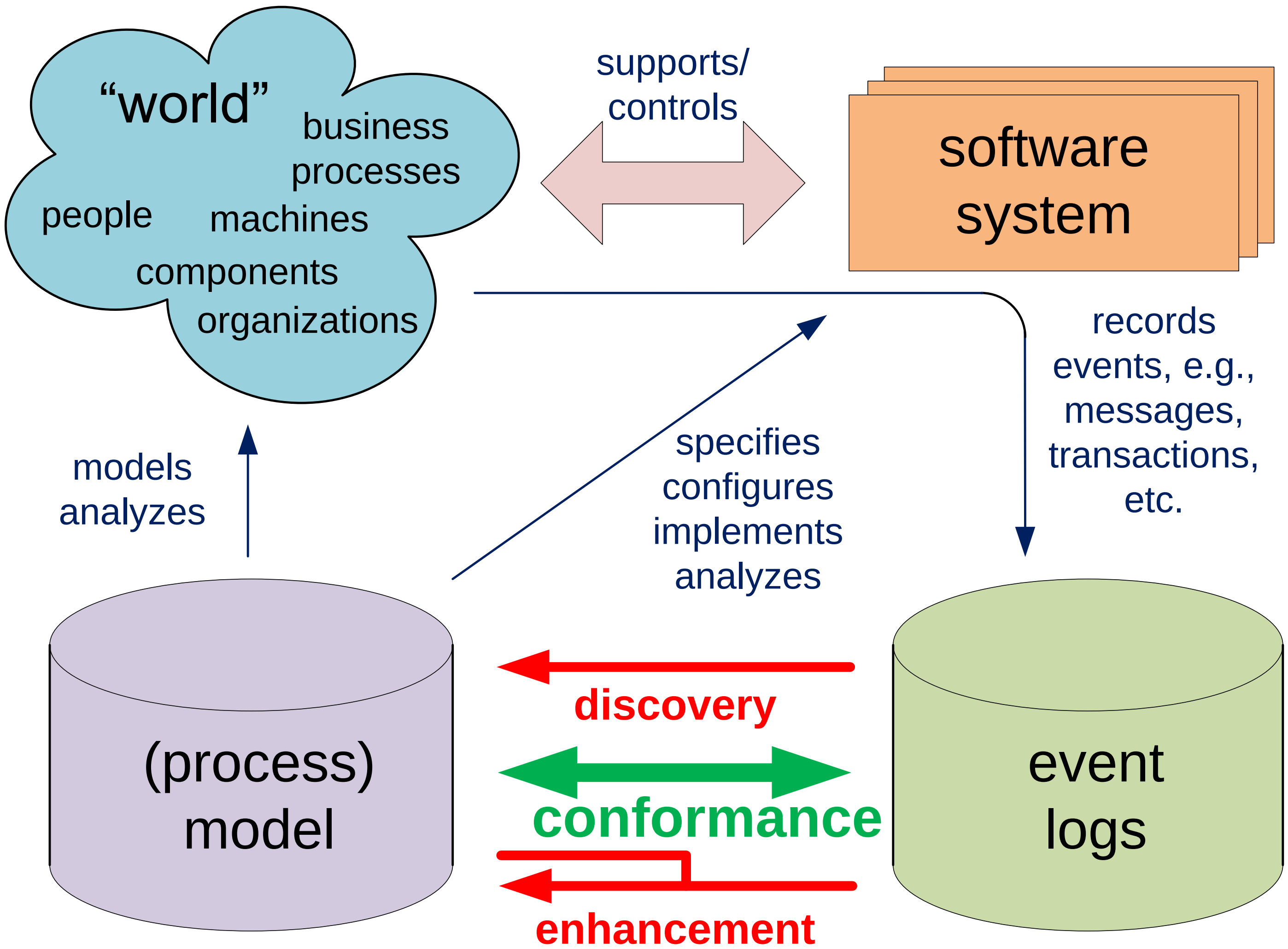
conformal process graph

α# algorithm

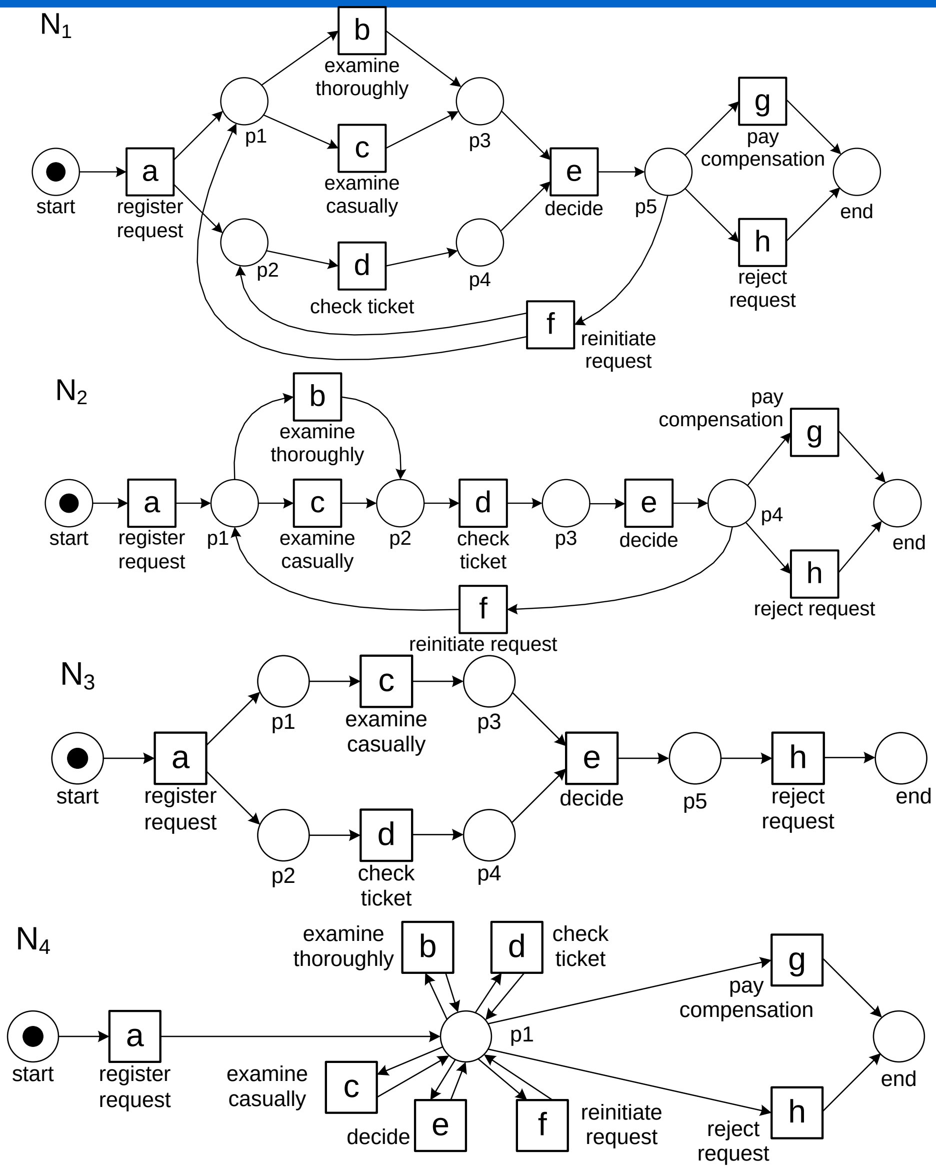
ILP mining

α++ algorithm

Conformance checking



Four models, one log

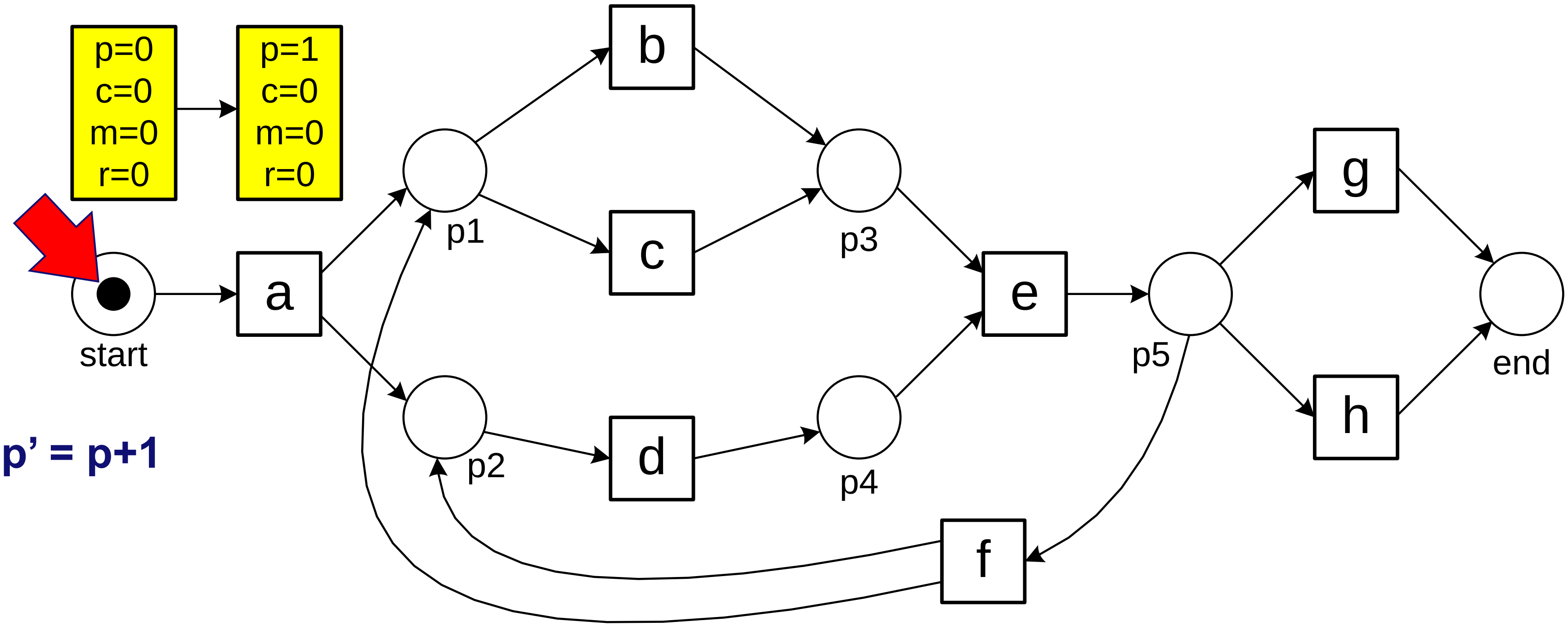


frequency	reference	trace
455	σ_1	$\langle a,c,d,e,h\rangle$
191	σ_2	$\langle a,b,d,e,g\rangle$
177	σ_3	$\langle a,d,c,e,h\rangle$
144	σ_4	$\langle a,b,d,e,h\rangle$
111	σ_5	$\langle a,c,d,e,g\rangle$
82	σ_6	$\langle a,d,c,e,g\rangle$
56	σ_7	$\langle a,d,b,e,h\rangle$
47	σ_8	$\langle a,c,d,e,f,d,b,e,h\rangle$
38	σ_9	$\langle a,d,b,e,g\rangle$
33	σ_{10}	$\langle a,c,d,e,f,b,d,e,h\rangle$
14	σ_{11}	$\langle a,c,d,e,f,b,d,e,g\rangle$
11	σ_{12}	$\langle a,c,d,e,f,d,b,e,g\rangle$
9	σ_{13}	$\langle a,d,c,e,f,c,d,e,h\rangle$
8	σ_{14}	$\langle a,d,c,e,f,d,b,e,h\rangle$
5	σ_{15}	$\langle a,d,c,e,f,b,d,e,g\rangle$
3	σ_{16}	$\langle a,c,d,e,f,b,d,e,f,d,b,e,g\rangle$
2	σ_{17}	$\langle a,d,c,e,f,d,b,e,g\rangle$
2	σ_{18}	$\langle a,d,c,e,f,b,d,e,f,b,d,e,g\rangle$
1	σ_{19}	$\langle a,d,c,e,f,d,b,e,f,b,d,e,h\rangle$
1	σ_{20}	$\langle a,d,b,e,f,b,d,e,f,d,b,e,g\rangle$
1	σ_{21}	$\langle a,d,c,e,f,d,b,e,f,c,d,e,f,d,b,e,g\rangle$

frequency	reference	trace
455	σ_1	$\langle a,c,d,e,h \rangle$
191	σ_2	$\langle a,b,d,e,g \rangle$
177	σ_3	$\langle a,d,c,e,h \rangle$
144	σ_4	$\langle a,b,d,e,h \rangle$
111	σ_5	$\langle a,c,d,e,g \rangle$
82	σ_6	$\langle a,d,c,e,g \rangle$
56	σ_7	$\langle a,d,b,e,h \rangle$
47	σ_8	$\langle a,c,d,e,f,d,b,e,h \rangle$
38	σ_9	$\langle a,d,b,e,g \rangle$
33	σ_{10}	$\langle a,c,d,e,f,b,d,e,h \rangle$
14	σ_{11}	$\langle a,c,d,e,f,b,d,e,g \rangle$
11	σ_{12}	$\langle a,c,d,e,f,d,b,e,g \rangle$
9	σ_{13}	$\langle a,d,c,e,f,c,d,e,h \rangle$
8	σ_{14}	$\langle a,d,c,e,f,d,b,e,h \rangle$

Replaying (1/7) σ_1 on N_1

$$\sigma_1 = \langle a, c, d, e, h \rangle$$

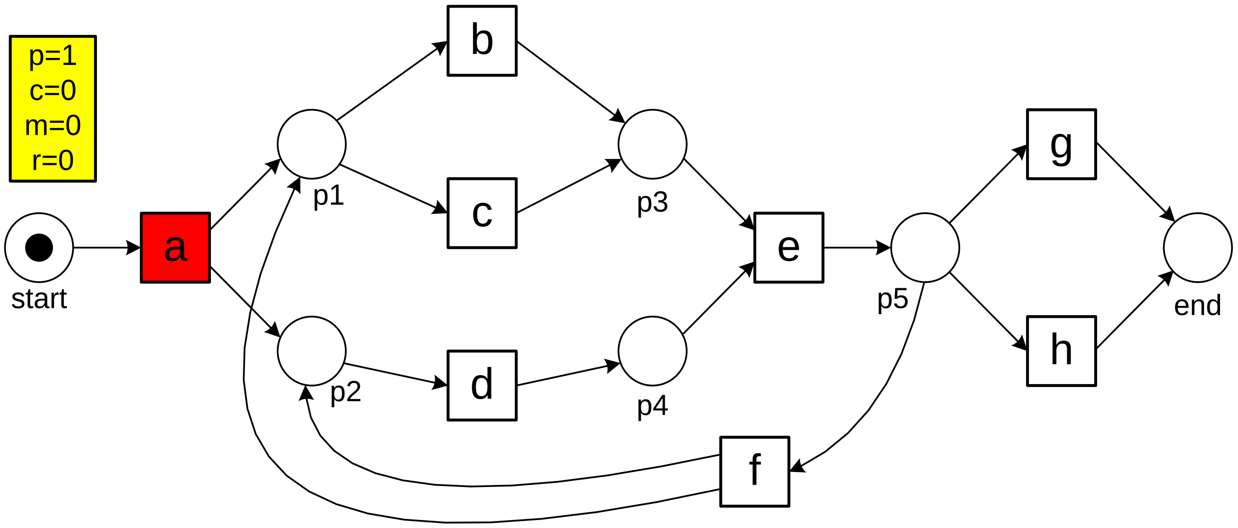


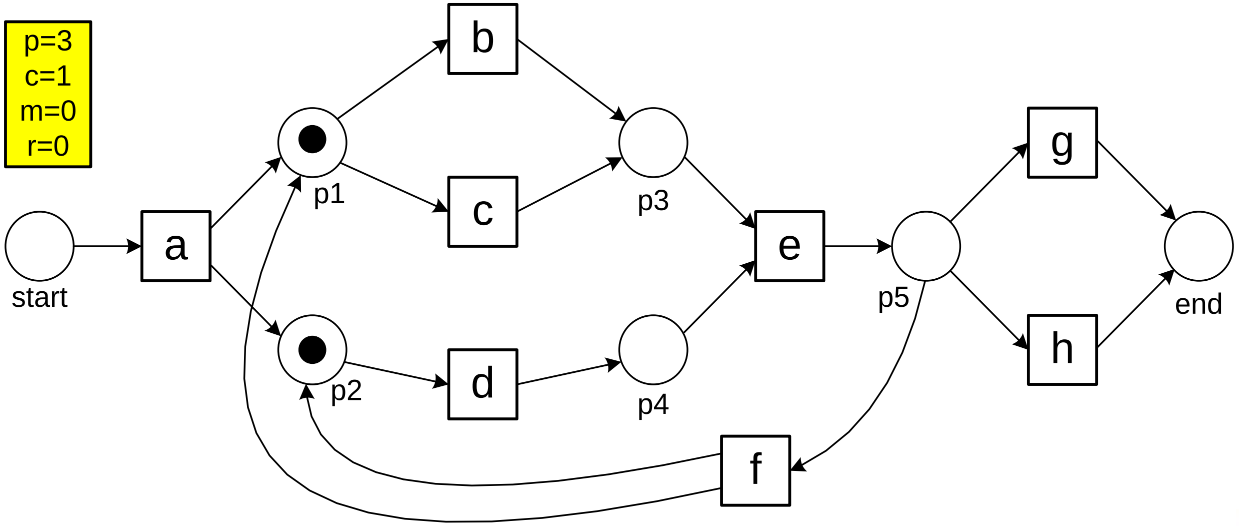
p' = p+1

p = produced
c = consumed
m = missing ≤ c
r = remaining ≤ p

Replaying (2/7)
σ_1 on N_1

$$\sigma_1 = \langle a, c, d, e, h \rangle$$

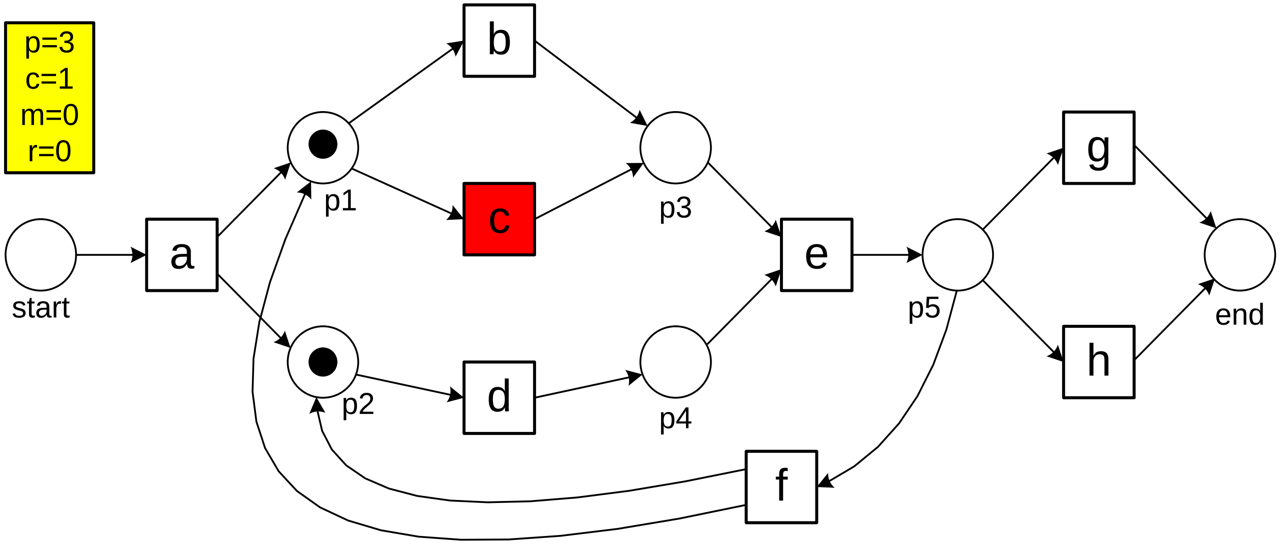




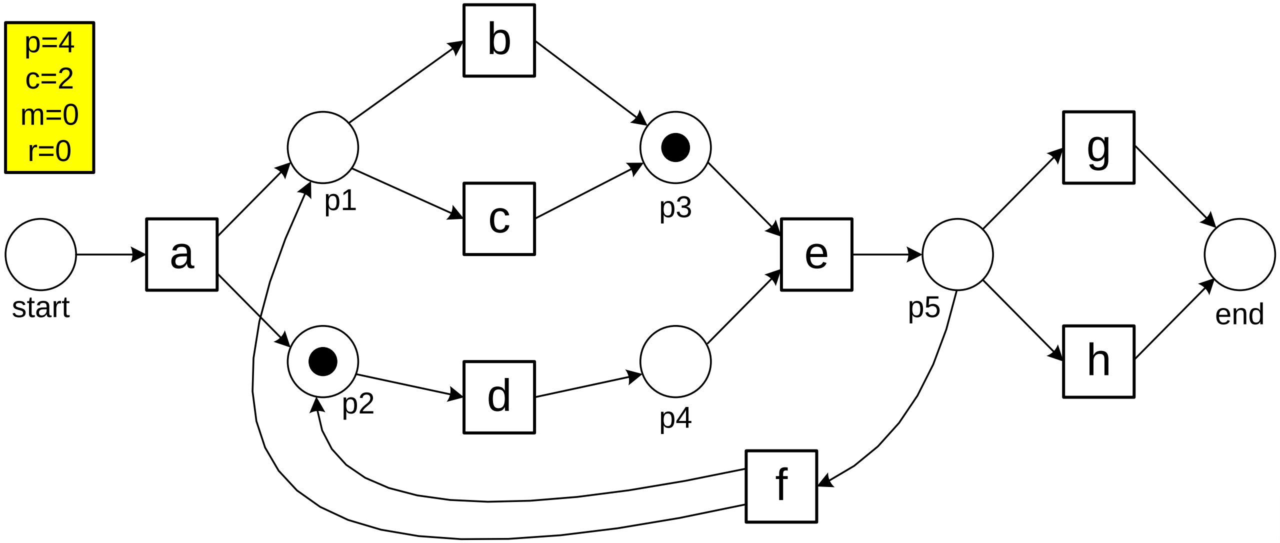
p' = p+2
c' = c+1

Replaying (3/7)
σ_1 on N_1

$$\sigma_1 = \langle a, c, d, e, h \rangle$$

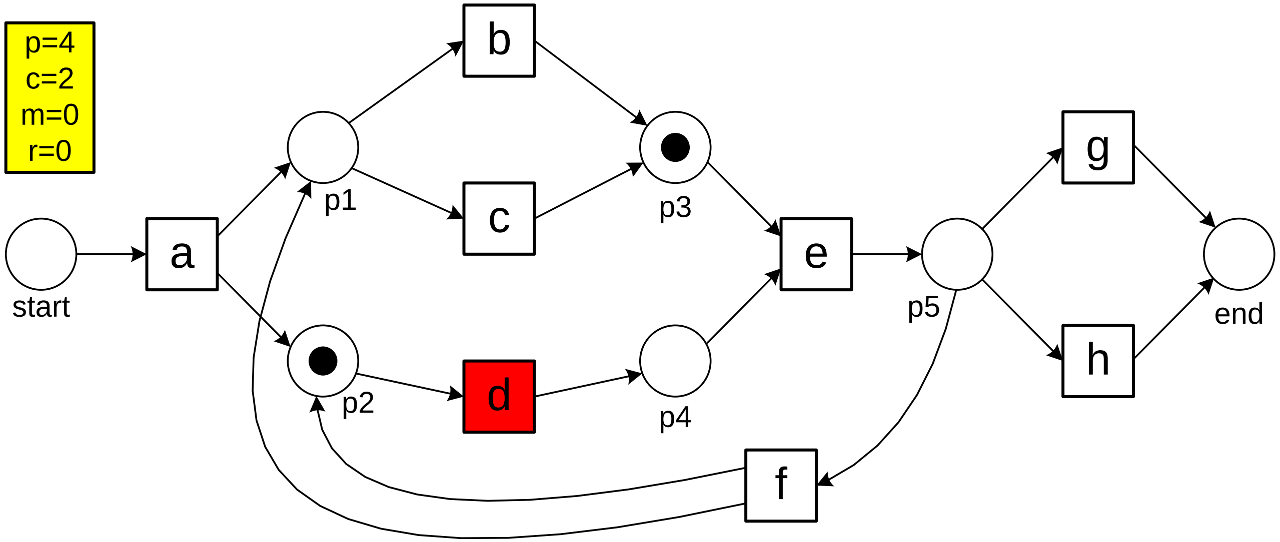


p' = p+1
c' = c+1

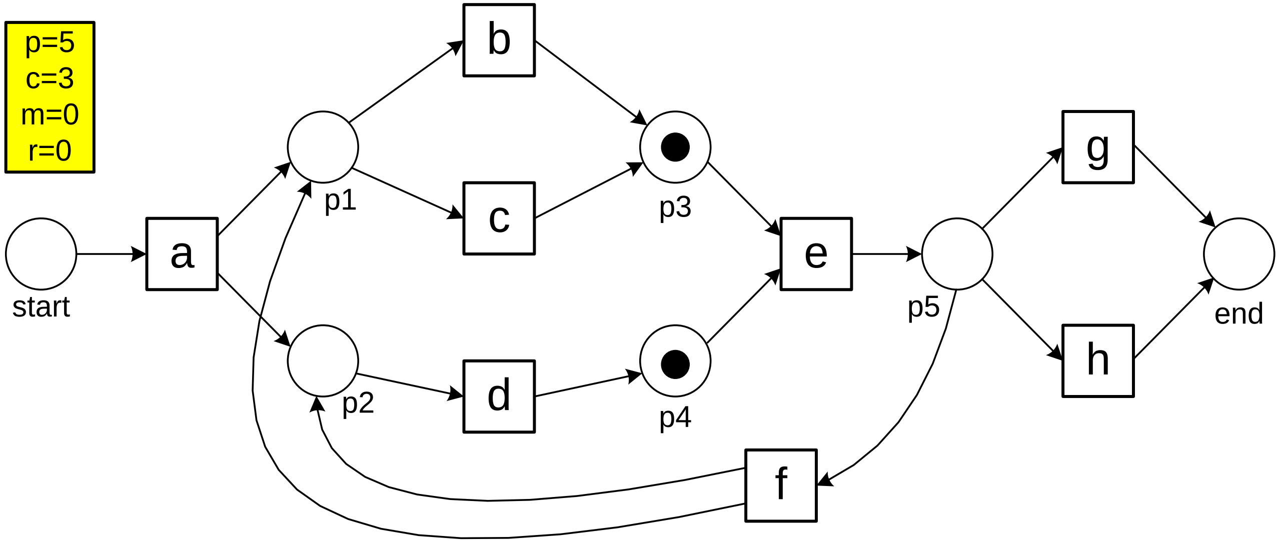


Replaying (4/7)
σ_1 on N_1

$\sigma_1 = \langle a, c, d, e, h \rangle$

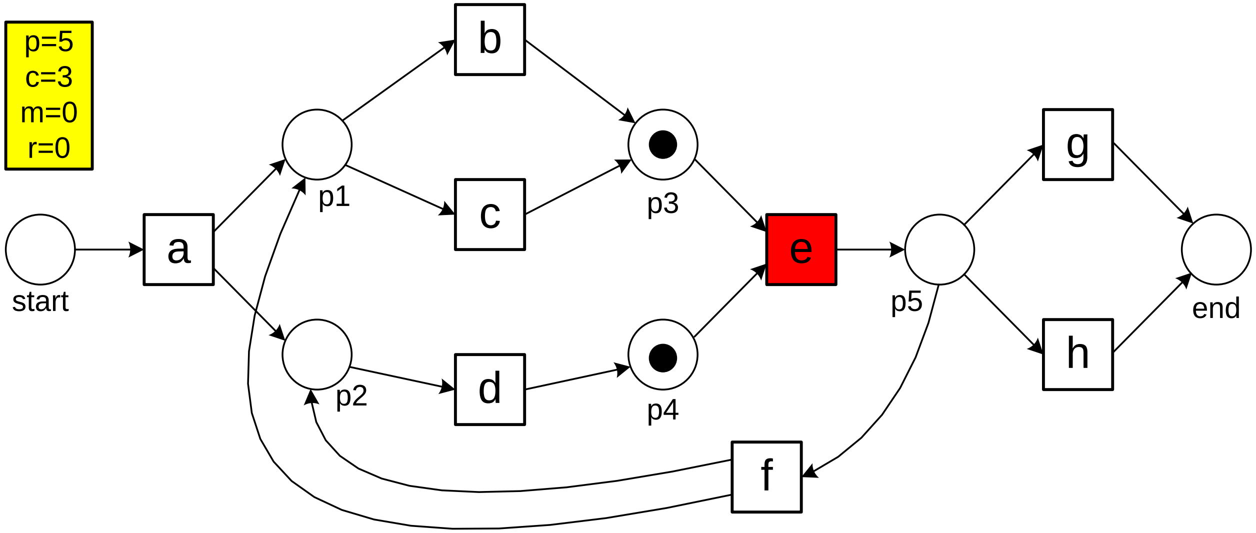


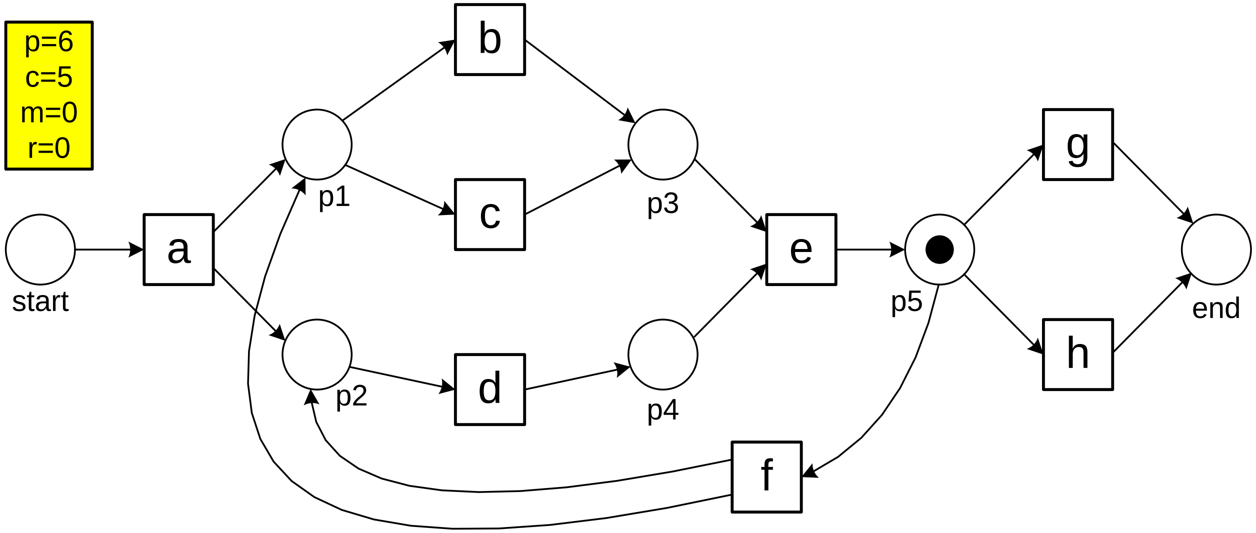
$p' = p+1$
$c' = c+1$



Replaying (5/7) σ_1 on N_1

$$\sigma_1 = \langle a, c, d, e, h \rangle$$

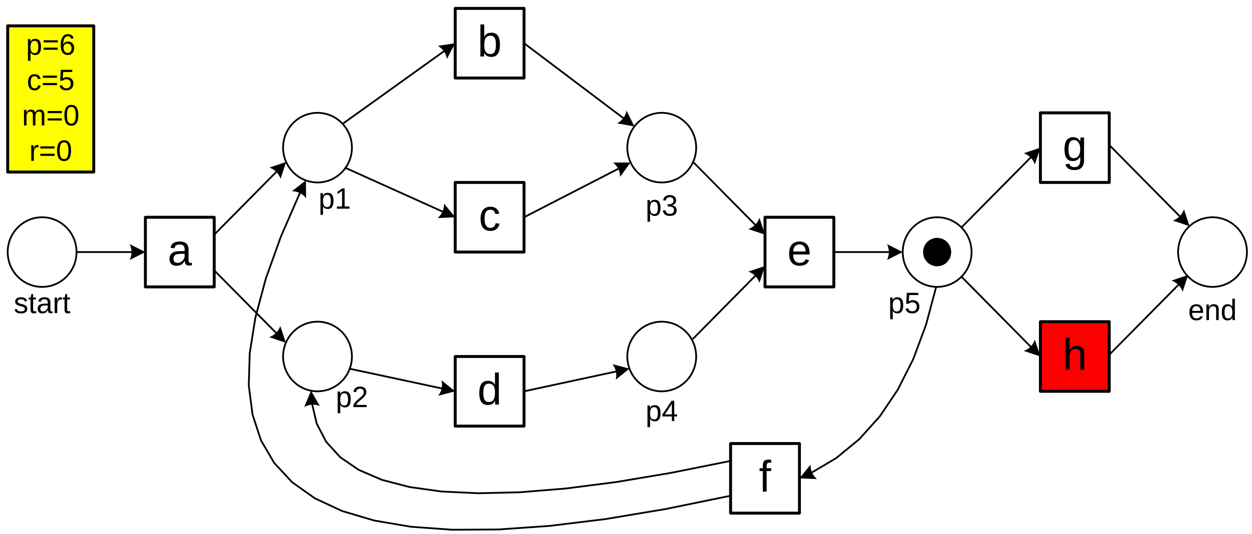


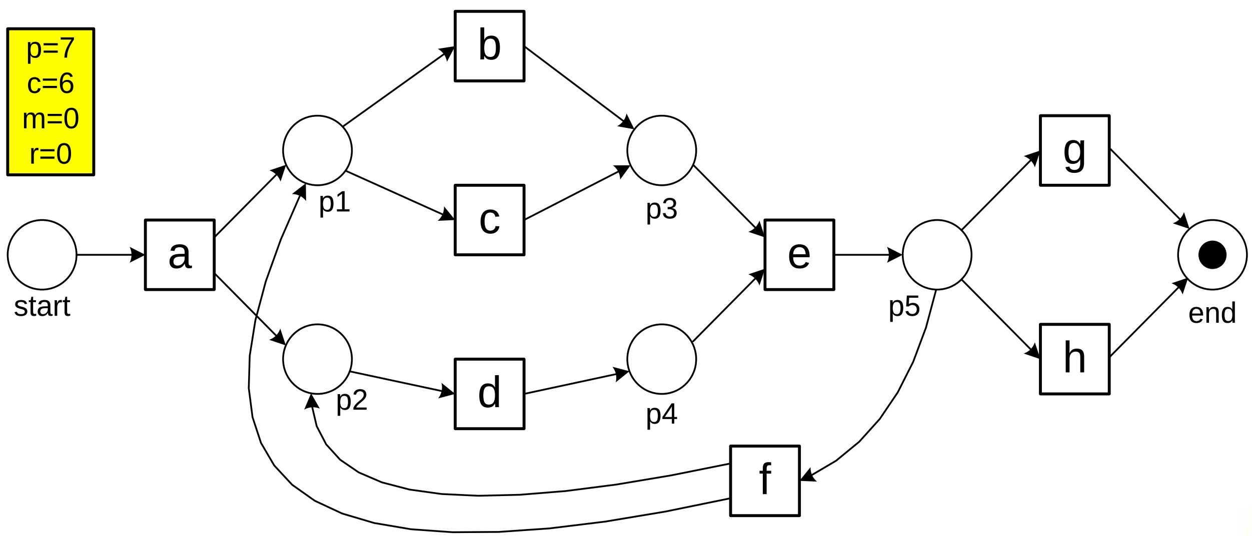


p' = p+1
c' = c+2

Replaying (6/7)
σ_1 on N_1

$$\sigma_1 = \langle a, c, d, e, h \rangle$$

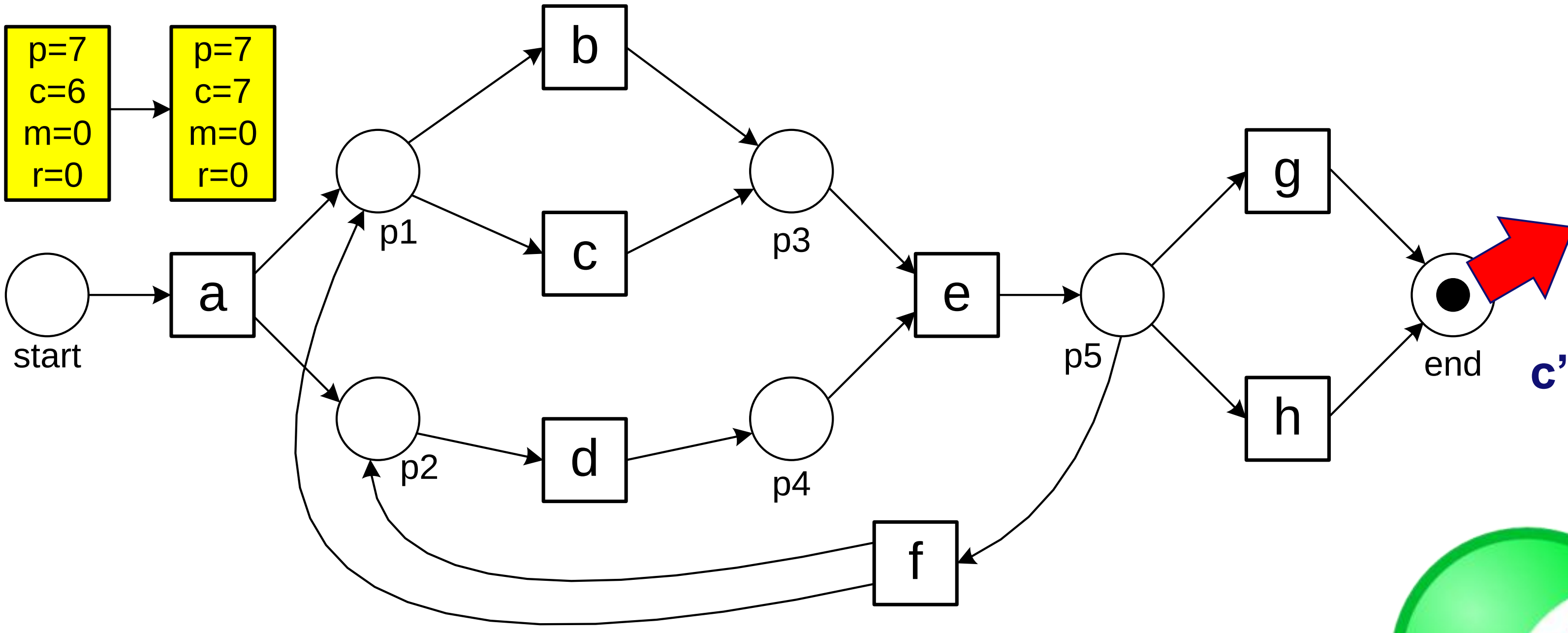




p' = p+1
c' = c+1

Replaying (7/7)
σ_1 on N_1

$\sigma_1 = \langle a, c, d, e, h \rangle$

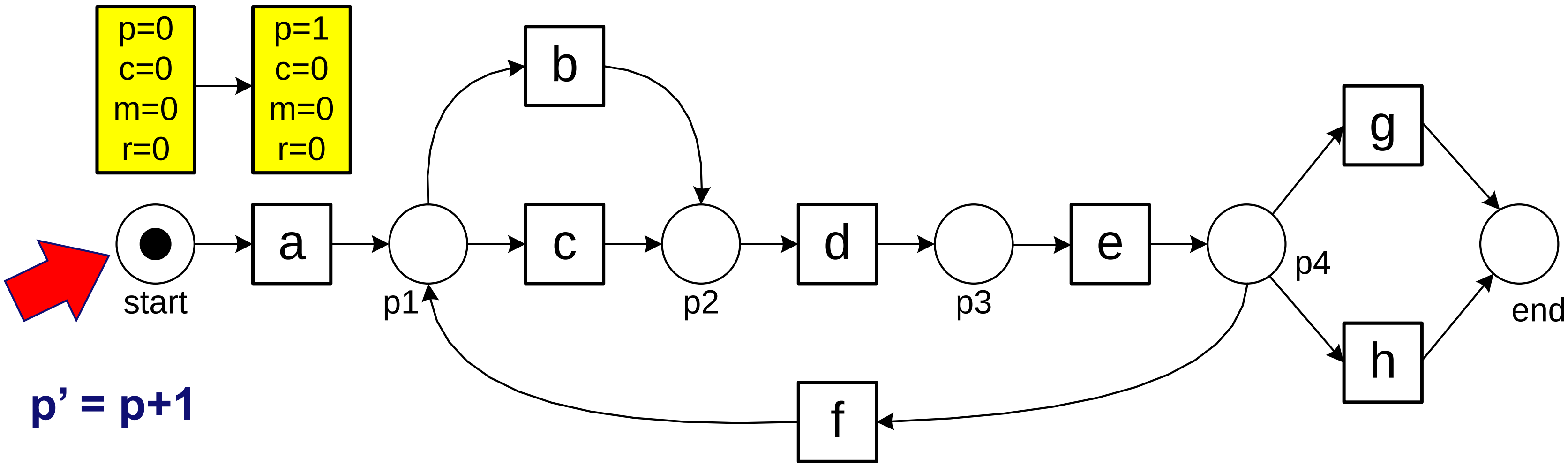


c' = c+1

m = 0
r = 0
no problems encountered

Replaying (1/7) σ_3 on N_2

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

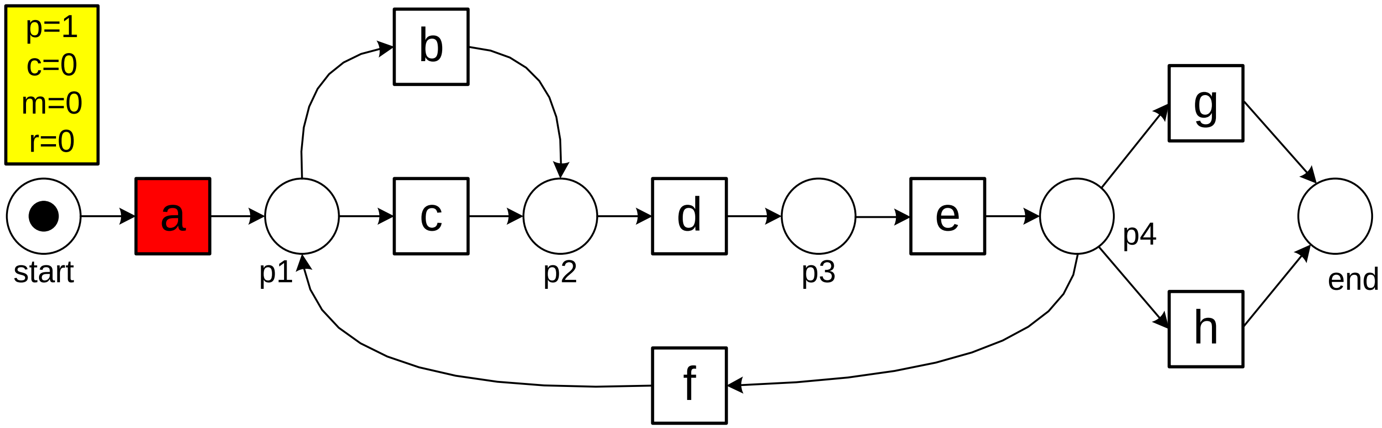


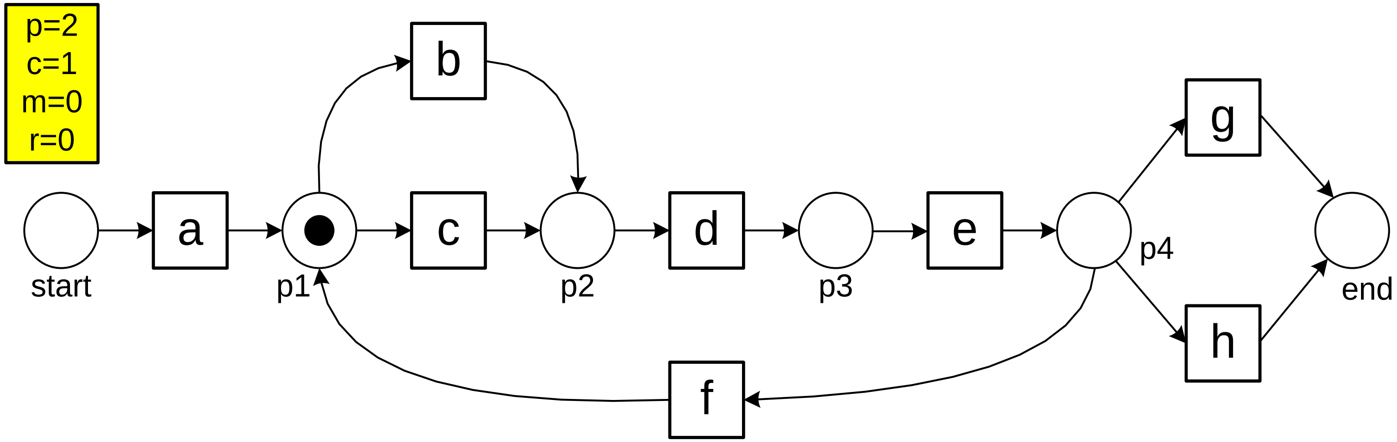
p' = p+1

p = produced
c = consumed
m = missing ≤ c
r = remaining ≤ p

Replaying (2/7)
σ_3 on N_2

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

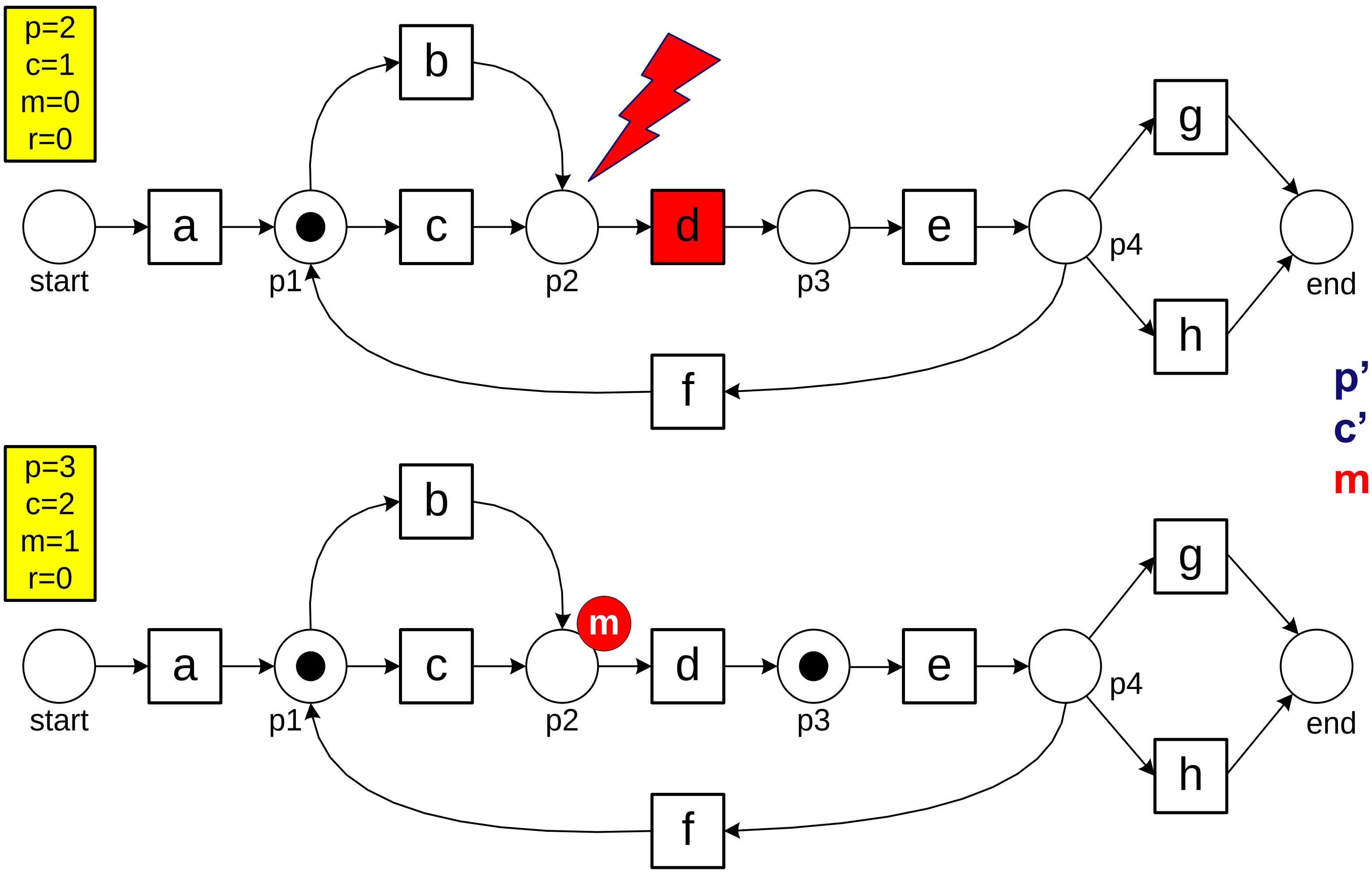




p' = p+1
c' = c+1

Replaying (3/7) σ_3 on N_2

$$\sigma_3 = \langle a, d, c, e, h \rangle$$



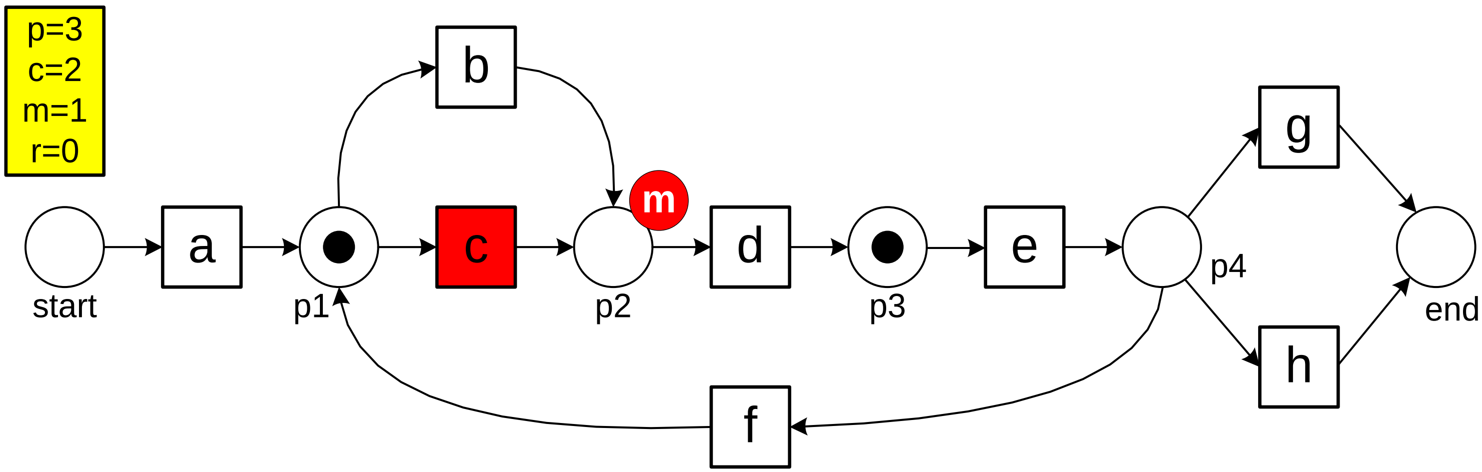
$p' = p+1$

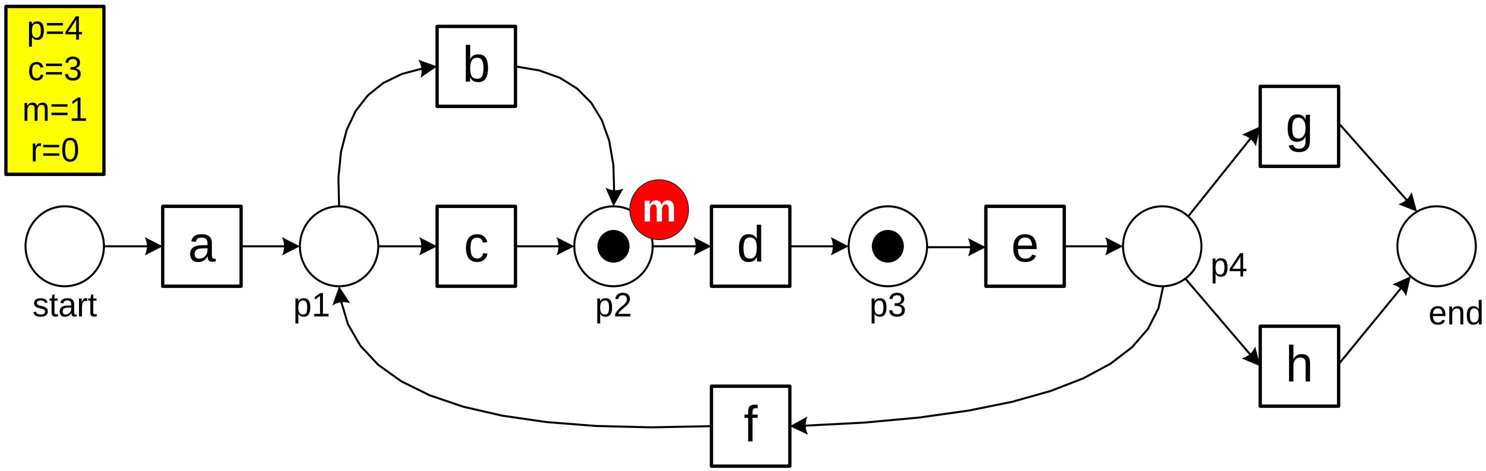
$c' = c+1$

$m' = m+1$

Replaying (4/7)
σ_3 on N_2

$$\sigma_3 = \langle a, d, c, e, h \rangle$$



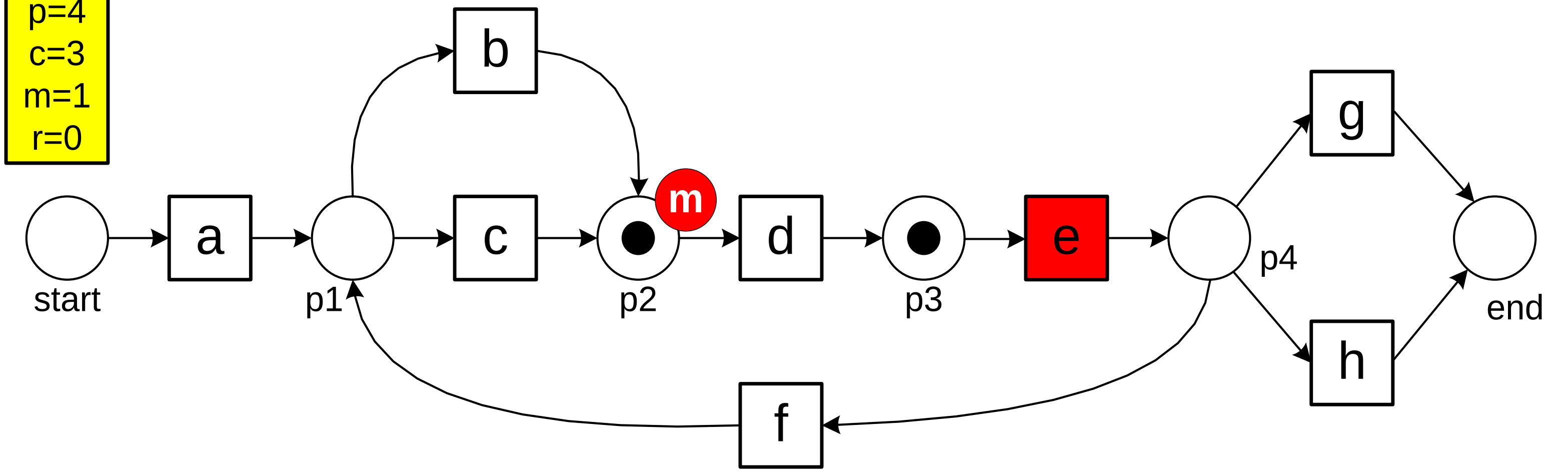


p' = p+1
c' = c+1

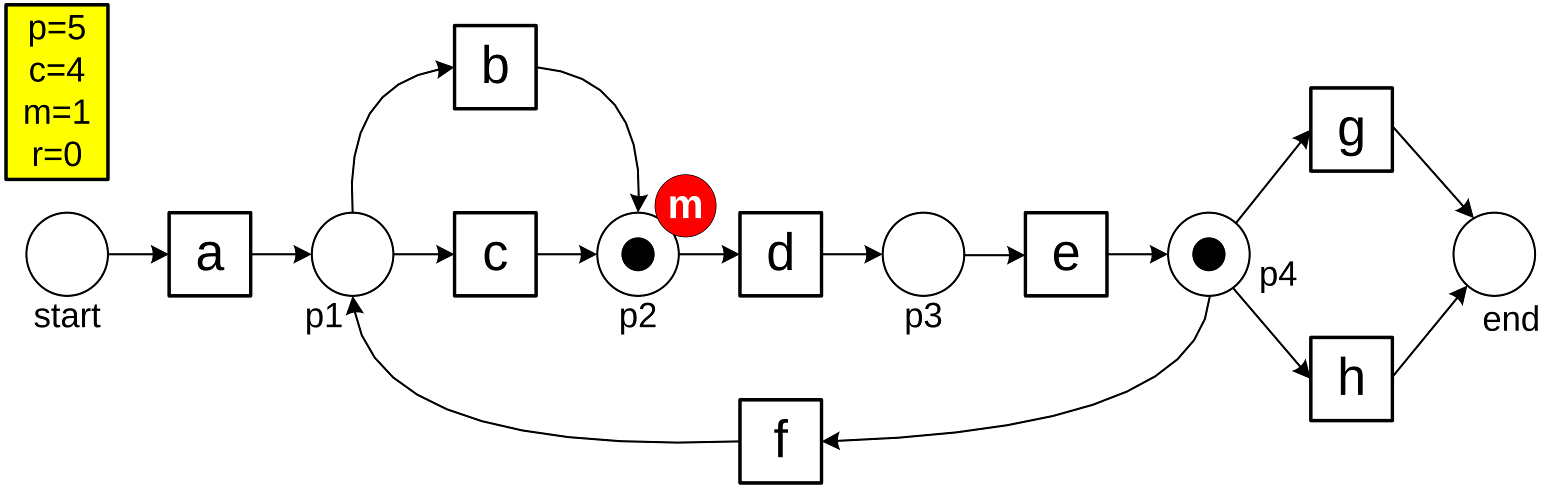
Replaying (5/7)
σ_3 on N_2

$\sigma_3 = \langle a, d, c, e, h \rangle$

p=4
c=3
m=1
r=0



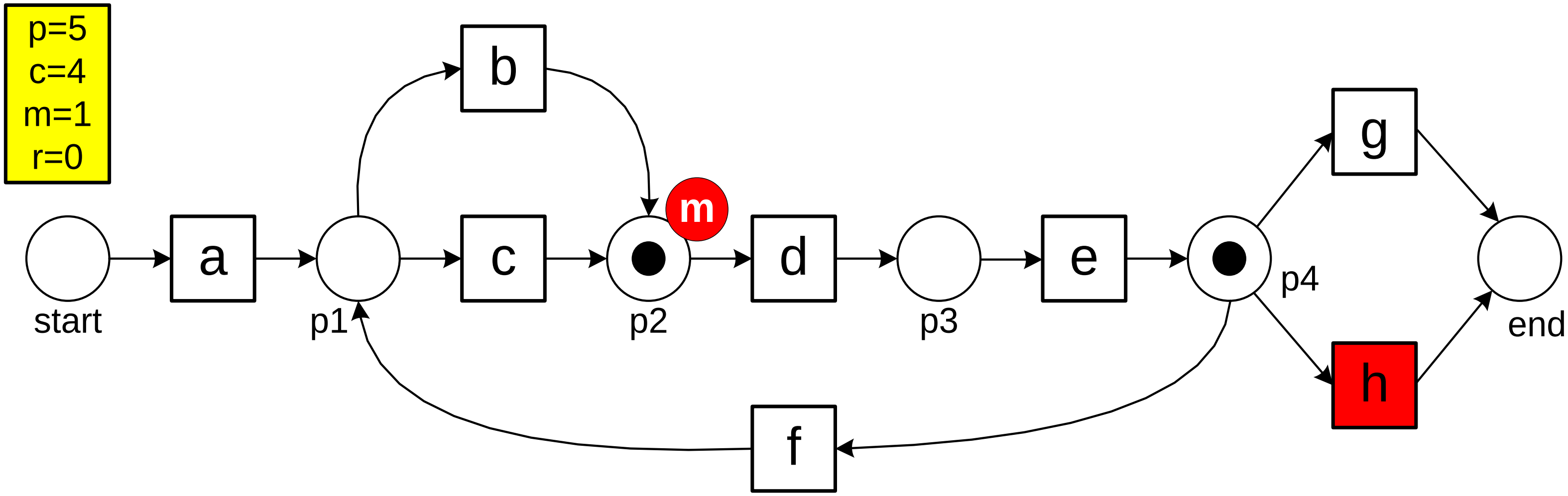
p=5
c=4
m=1
r=0

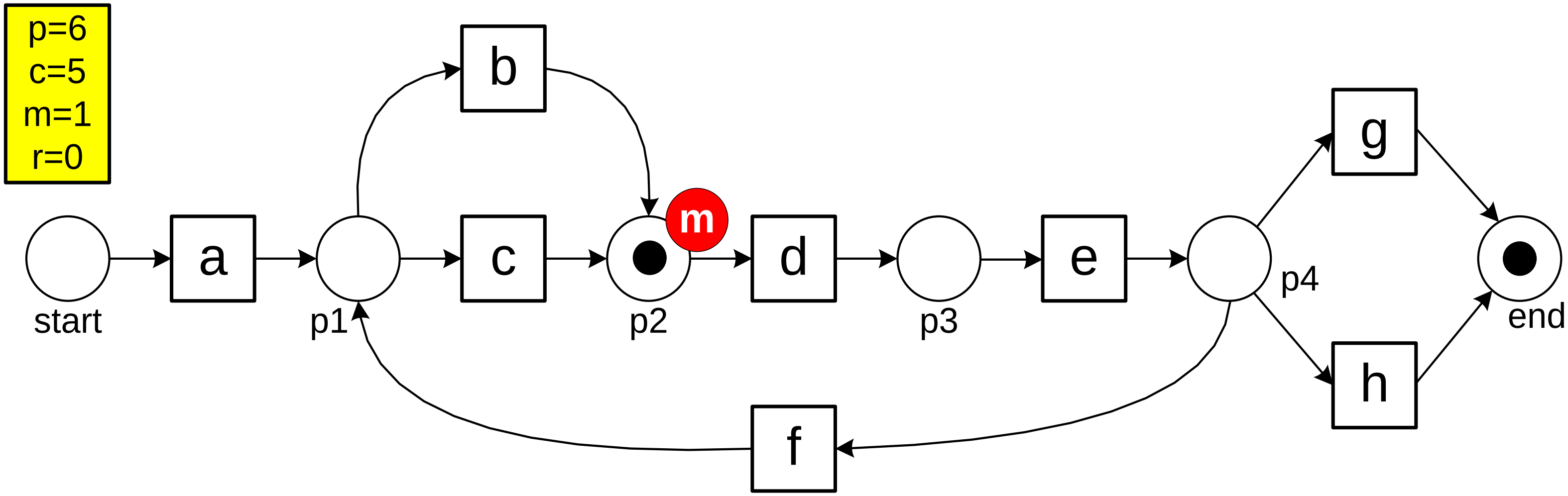


p' = p+1
c' = c+1

Replaying (6/7) σ_3 on N_2

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

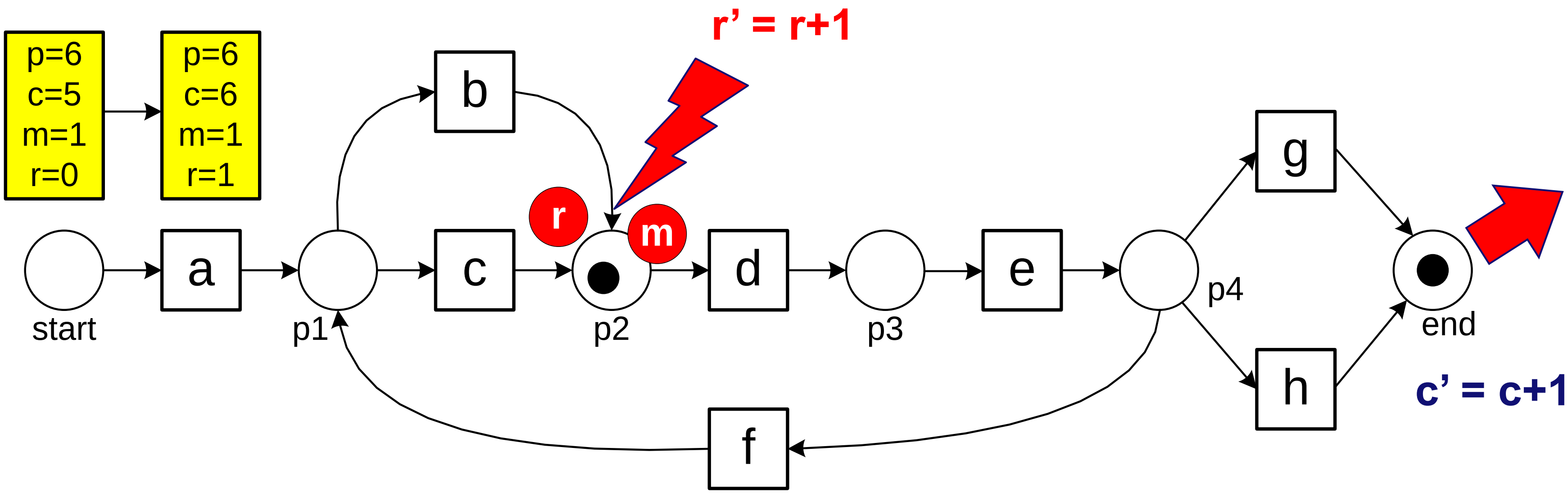




p' = p+1
c' = c+1

Replaying (7/7) σ_3 on N_2

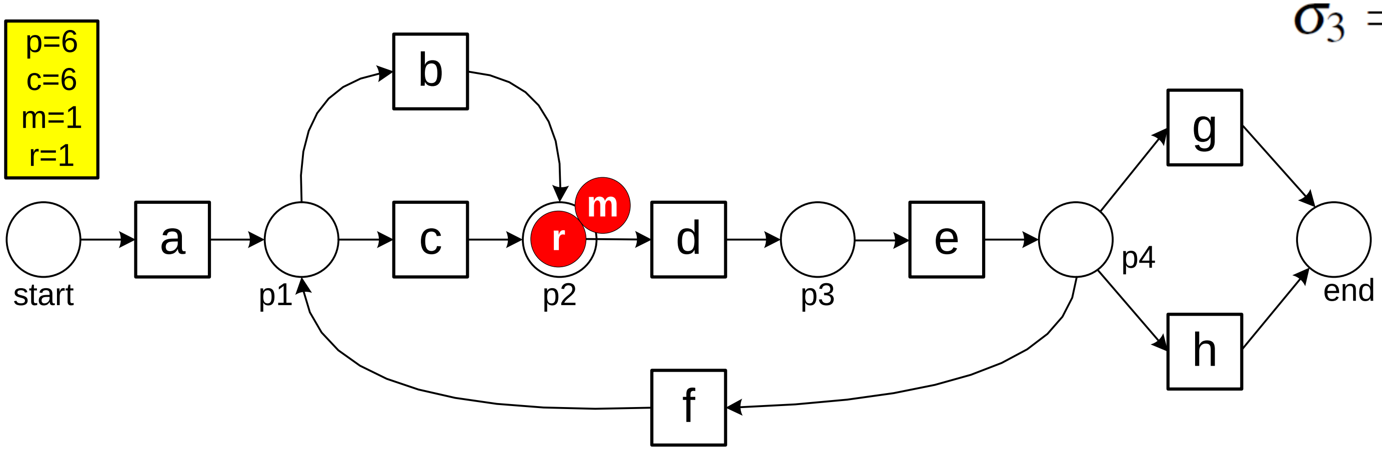
$$\sigma_3 = \langle a, d, c, e, h \rangle$$



Problems:

- **m = 1 :** **d was forced to occur without being enabled**
- **r = 1 :** **output of c was not used by d**

Computing fitness at trace level



$$\sigma_3 = \langle a, d, c, e, h \rangle$$

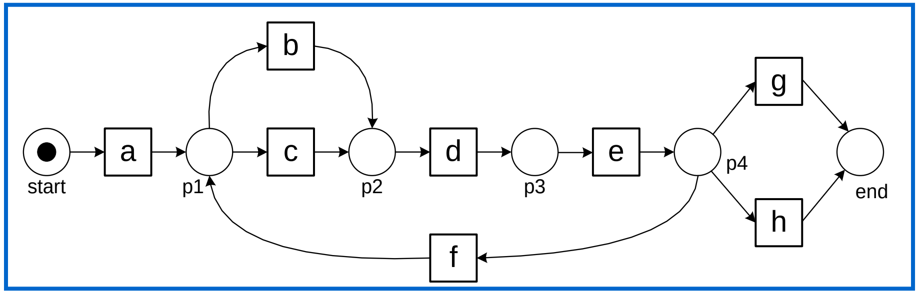
$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{m}{c}\right) + \frac{1}{2}\left(1 - \frac{r}{p}\right)$$

$$fitness(\sigma_3, N_2) = \frac{1}{2}\left(1 - \frac{1}{6}\right) + \frac{1}{2}\left(1 - \frac{1}{6}\right) = 0.8333$$

Computing fitness at the log level

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

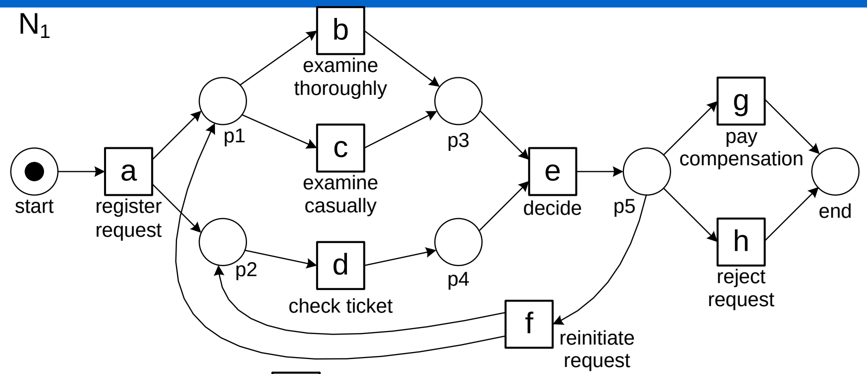
frequency	reference	trace
455	σ_1	$\langle a,c,d,e,h \rangle$
191	σ_2	$\langle a,b,d,e,g \rangle$
177	σ_3	$\langle a,d,c,e,h \rangle$
144	σ_4	$\langle a,b,d,e,h \rangle$
111	σ_5	$\langle a,c,d,e,g \rangle$
82	σ_6	$\langle a,d,c,e,g \rangle$
56	σ_7	$\langle a,d,b,e,h \rangle$
47	σ_8	$\langle a,c,d,e,f,d,b,e,h \rangle$
38	σ_9	$\langle a,d,b,e,g \rangle$
33	σ_{10}	$\langle a,c,d,e,f,b,d,e,h \rangle$
14	σ_{11}	$\langle a,c,d,e,f,b,d,e,g \rangle$
11	σ_{12}	$\langle a,c,d,e,f,d,b,e,g \rangle$
9	σ_{13}	$\langle a,d,c,e,f,c,d,e,h \rangle$
8	σ_{14}	$\langle a,d,c,e,f,d,b,e,h \rangle$
5	σ_{15}	$\langle a,d,c,e,f,b,d,e,g \rangle$
3	σ_{16}	$\langle a,c,d,e,f,b,d,e,f,d,b,e,g \rangle$
2	σ_{17}	$\langle a,d,c,e,f,d,b,e,g \rangle$
2	σ_{18}	$\langle a,d,c,e,f,b,d,e,f,b,d,e,g \rangle$
1	σ_{19}	$\langle a,d,c,e,f,d,b,e,f,b,d,e,h \rangle$
1	σ_{20}	$\langle a,d,b,e,f,b,d,e,f,d,b,e,g \rangle$
1	σ_{21}	$\langle a,d,c,e,f,d,b,e,f,c,d,e,f,d,b,e,g \rangle$



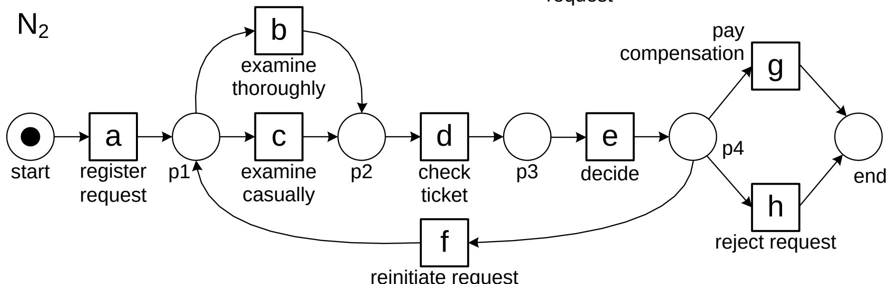
Example values

frequency	reference	trace
455	σ_1	$\langle a,c,d,e,h \rangle$
191	σ_2	$\langle a,b,d,e,g \rangle$
177	σ_3	$\langle a,d,c,e,h \rangle$
144	σ_4	$\langle a,b,d,e,h \rangle$
111	σ_5	$\langle a,c,d,e,g \rangle$
82	σ_6	$\langle a,d,c,e,g \rangle$
56	σ_7	$\langle a,d,b,e,h \rangle$
47	σ_8	$\langle a,c,d,e,f,d,b,e,h \rangle$
38	σ_9	$\langle a,d,b,e,g \rangle$
33	σ_{10}	$\langle a,c,d,e,f,b,d,e,h \rangle$
14	σ_{11}	$\langle a,c,d,e,f,b,d,e,g \rangle$
11	σ_{12}	$\langle a,c,d,e,f,d,b,e,g \rangle$
9	σ_{13}	$\langle a,d,c,e,f,c,d,e,h \rangle$
8	σ_{14}	$\langle a,d,c,e,f,d,b,e,h \rangle$
5	σ_{15}	$\langle a,d,c,e,f,b,d,e,g \rangle$
3	σ_{16}	$\langle a,c,d,e,f,b,d,e,f,d,b,e,g \rangle$
2	σ_{17}	$\langle a,d,c,e,f,d,b,e,g \rangle$
2	σ_{18}	$\langle a,d,c,e,f,b,d,e,f,b,d,e,g \rangle$
1	σ_{19}	$\langle a,d,c,e,f,d,b,e,f,b,d,e,h \rangle$
1	σ_{20}	$\langle a,d,b,e,f,b,d,e,f,d,b,e,g \rangle$
1	σ_{21}	$\langle a,d,c,e,f,d,b,e,f,c,d,e,f,d,b,e,g \rangle$

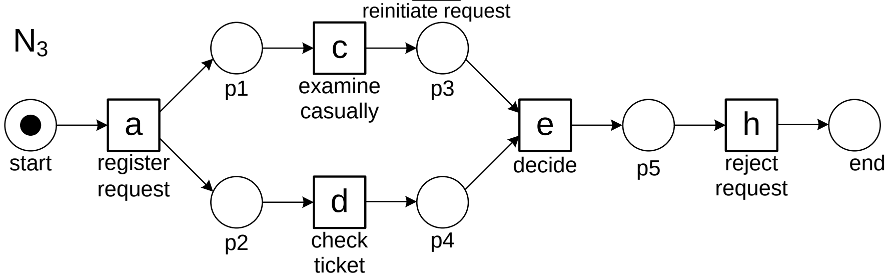
N_1



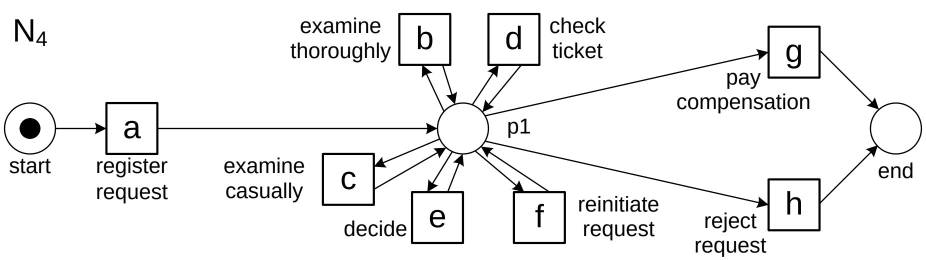
N_2



N_3



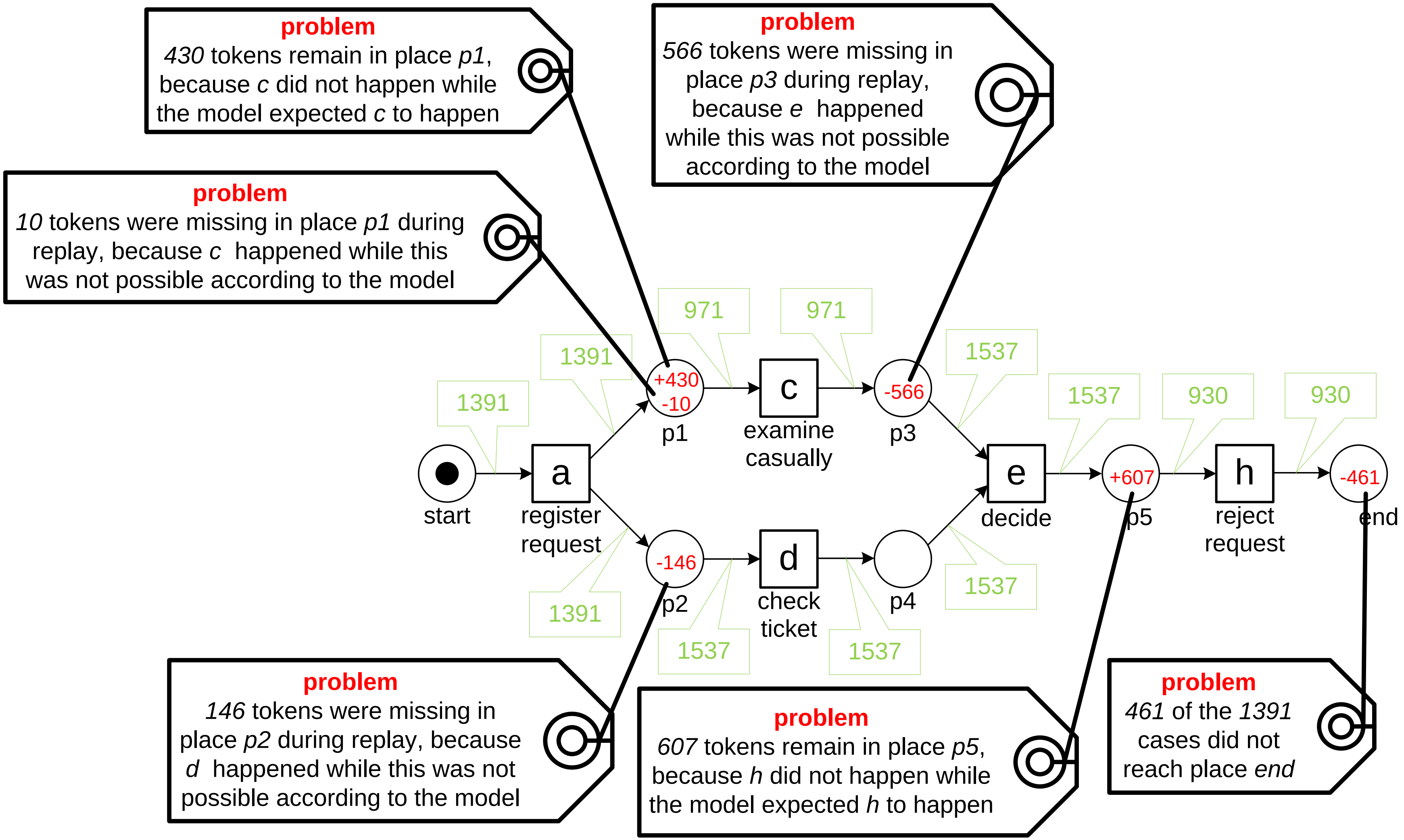
N_4



$$fitness(L_{full}, N_1) = 1$$
$$fitness(L_{full}, N_2) = 0.9504$$
$$fitness(L_{full}, N_3) = 0.8797$$
$$fitness(L_{full}, N_4) = 1$$

Diagnostics

$$(fitness(L_{full}, N_3) = 0.8797)$$



Challenges related to conformance checking

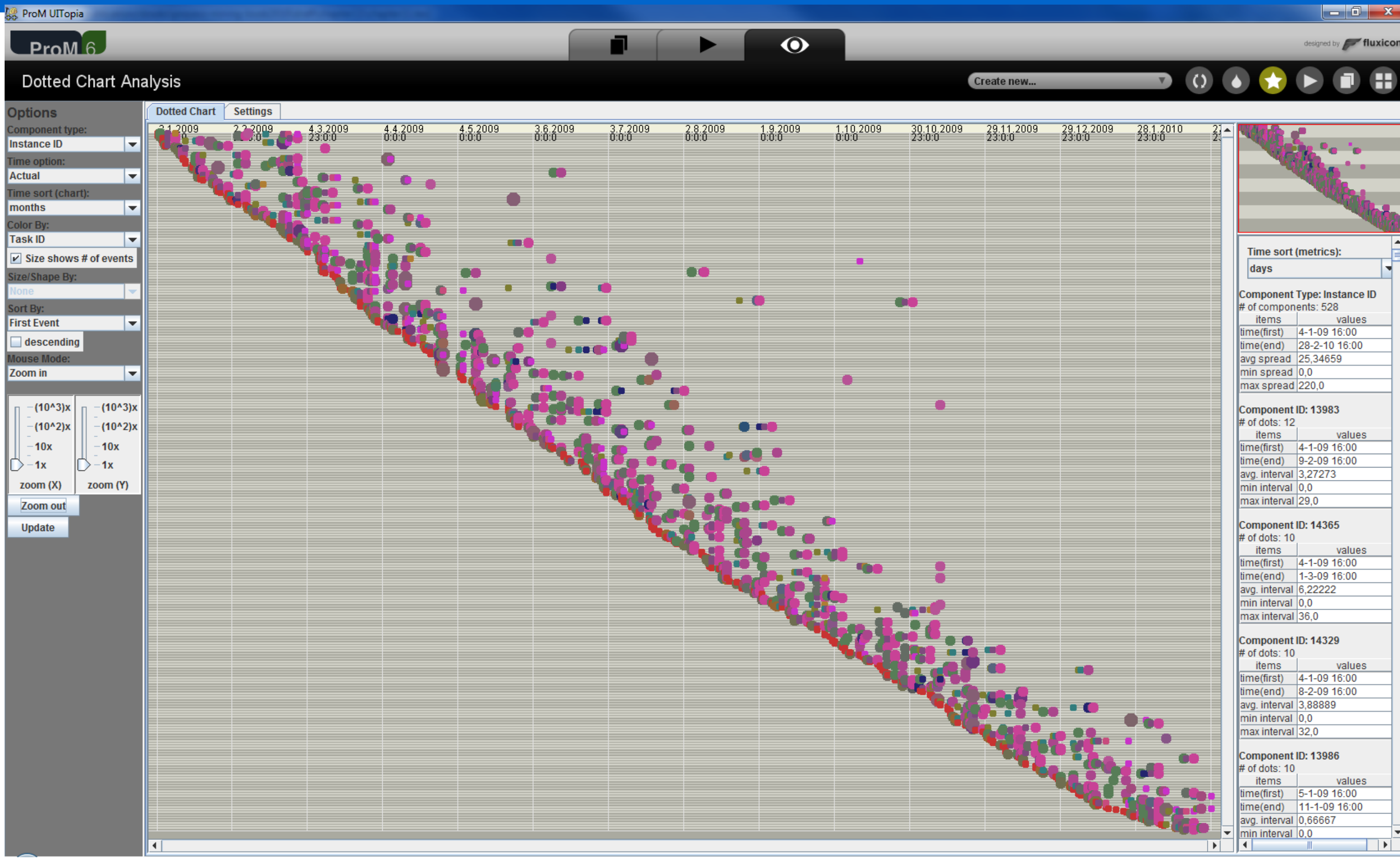
- Not as simple as it seems!
- In case of duplicate tasks (two transition with the same label) or silent tasks (τ labeled transitions), multiple paths need to be considered (state space analysis, heuristics, or optimization).
- More general formulation of the problem with costs associated to skipping/inserting particular tasks, see ProM latest conformance checker (A* algorithm).
- Computing the most likely alignment is needed for other types of process mining (time analysis, measuring precision, social network analysis, etc.).

How can process mining help?

- Detect bottlenecks
- Detect deviations
- Performance measurement
- Suggest improvements
- Decision support (e.g., recommendation and prediction)

- Provide mirror
- Highlight important problems
- Avoid ICT failures
- Avoid management by PowerPoint
- From “politics” to “analytics”

Example of a Lasagna process: WMO process of a Dutch municipality

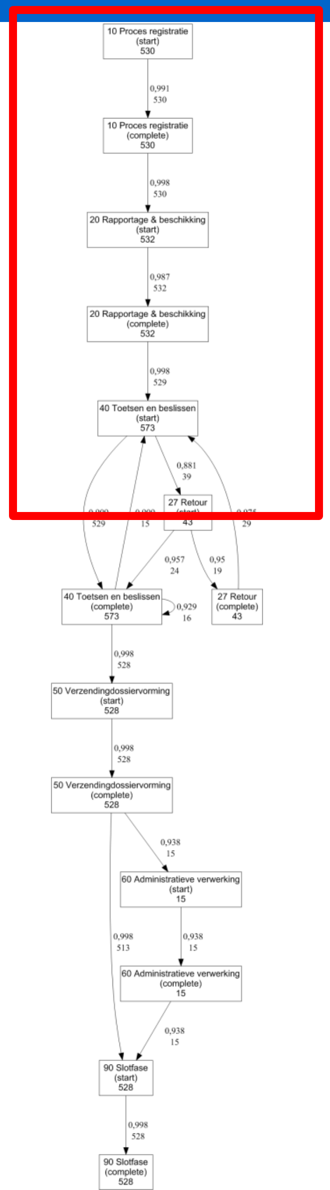


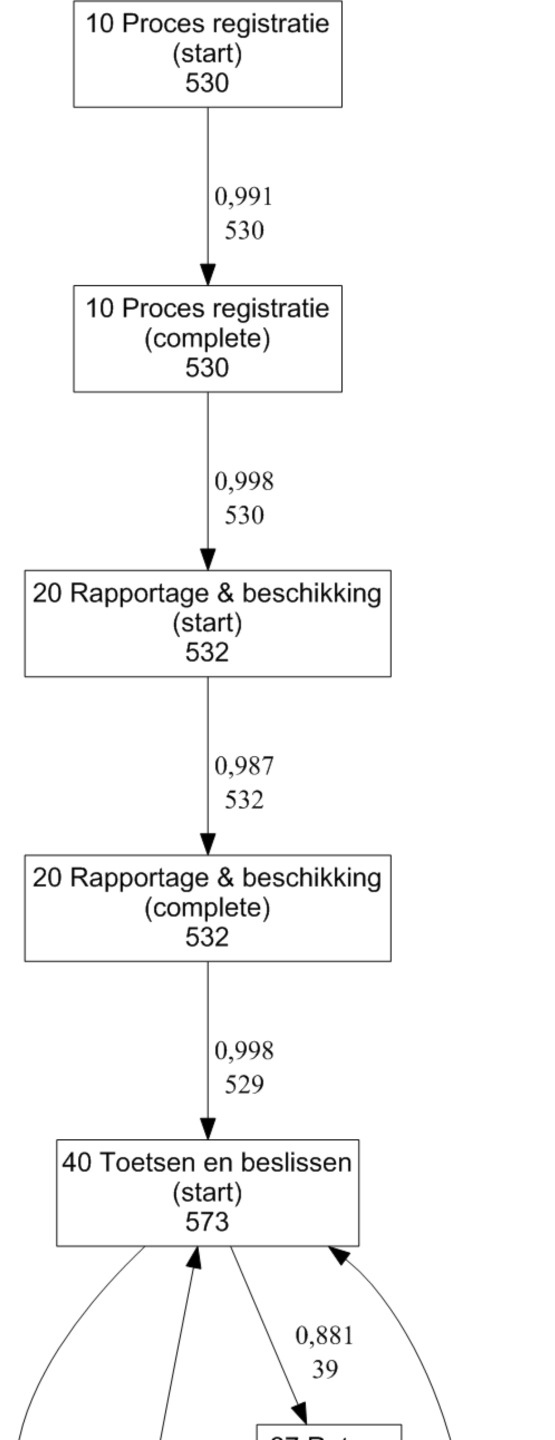
Each line corresponds to one of the 528 requests that were handled in the period from 4-1-2009 until 28-2-2010. In total there are 5498 events represented as dots. The mean time needed to handled a case is approximately 25 days.

WMO process
(Wet Maatschappelijke Ondersteuning)

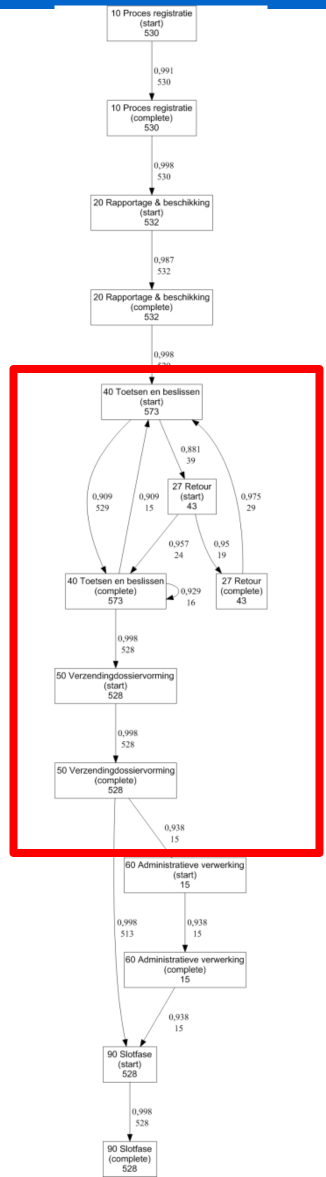
- **WMO refers to the social support act that came into force in The Netherlands on January 1st, 2007.**
- **The aim of this act is to assist people with disabilities and impairments. Under the act, local authorities are required to give support to those who need it, e.g., household help, providing wheelchairs and scootmobiles, and adaptations to homes.**
- **There are different processes for the different kinds of help. We focus on the process for handling requests for household help.**
- **In a period of about one year, 528 requests for household WMO support were received.**
- **These 528 requests generated 5498 events.**

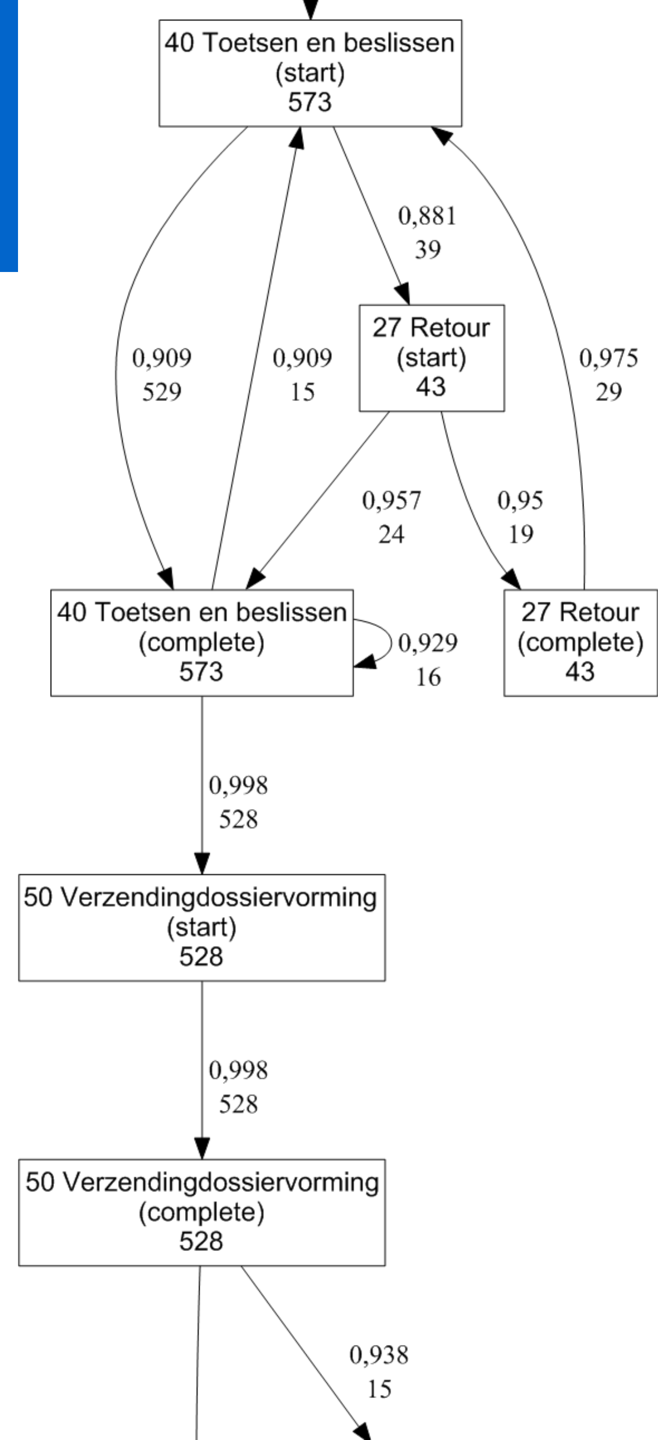
C-net discovered using heuristic miner (1/3)



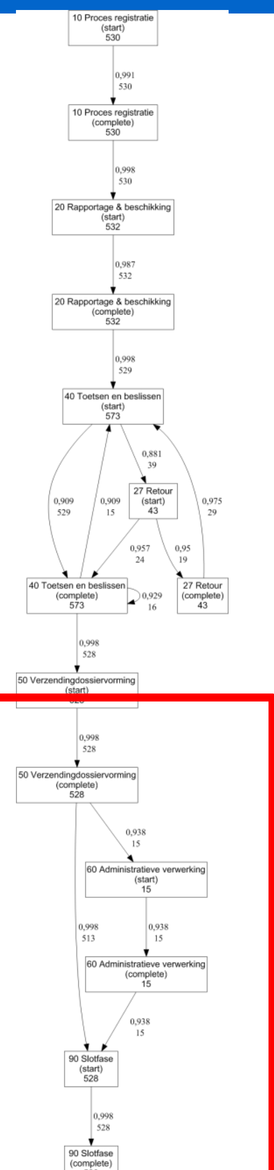


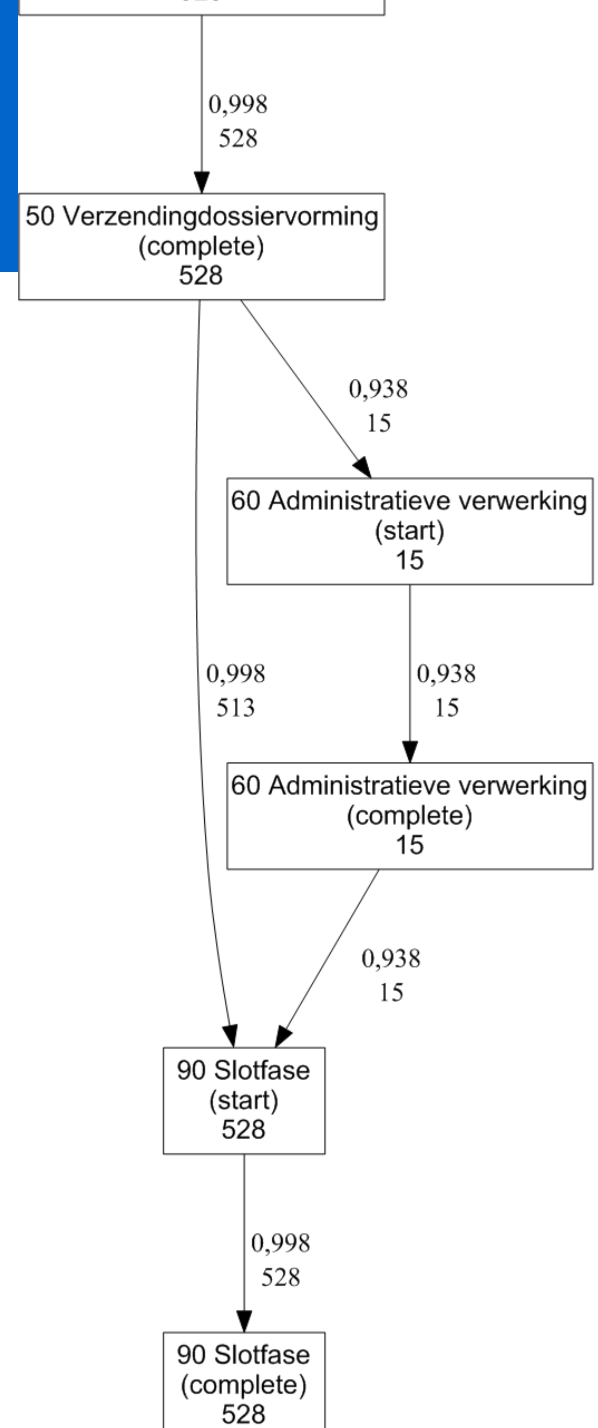
C-net discovered using heuristic miner (2/3)



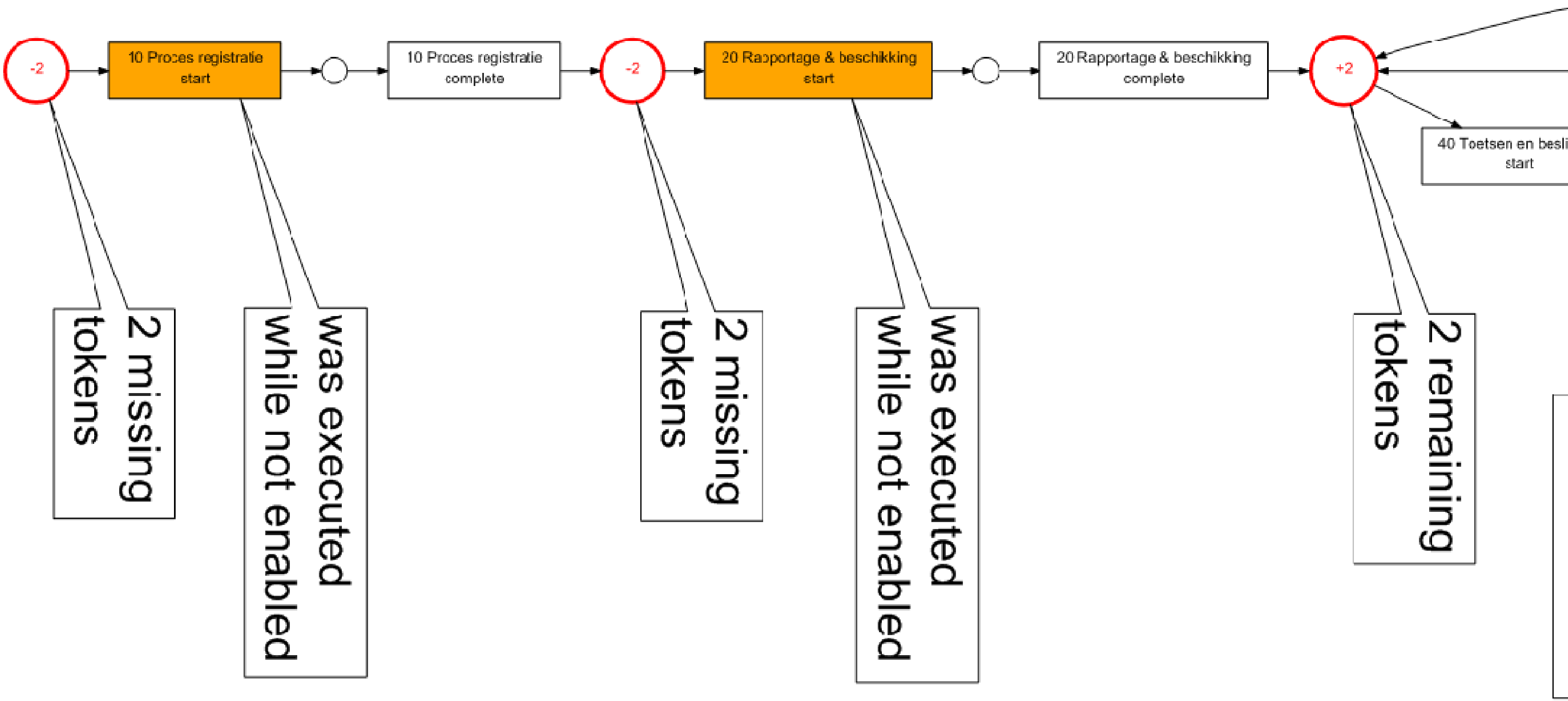


C-net discovered using heuristic miner (3/3)

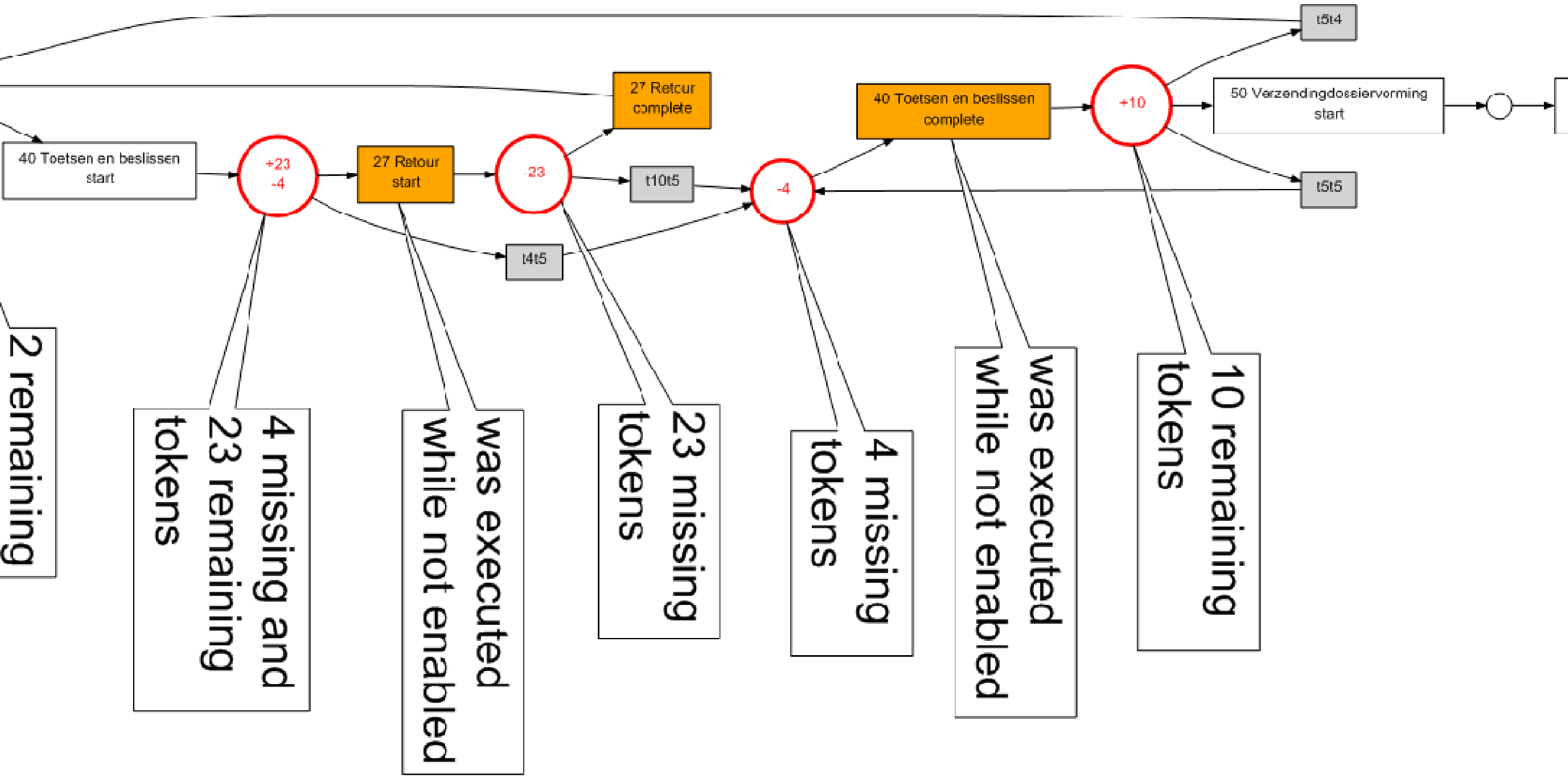




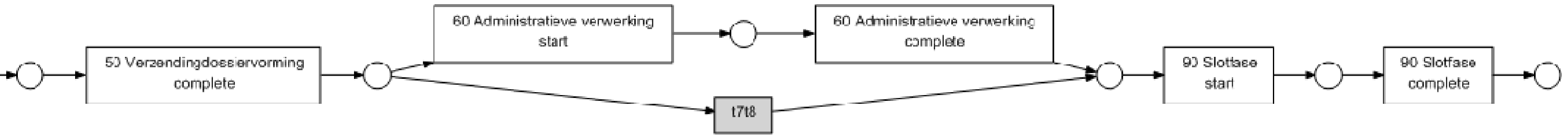
Conformance check WMO process (1/3)



Conformance check WMO process (2/3)

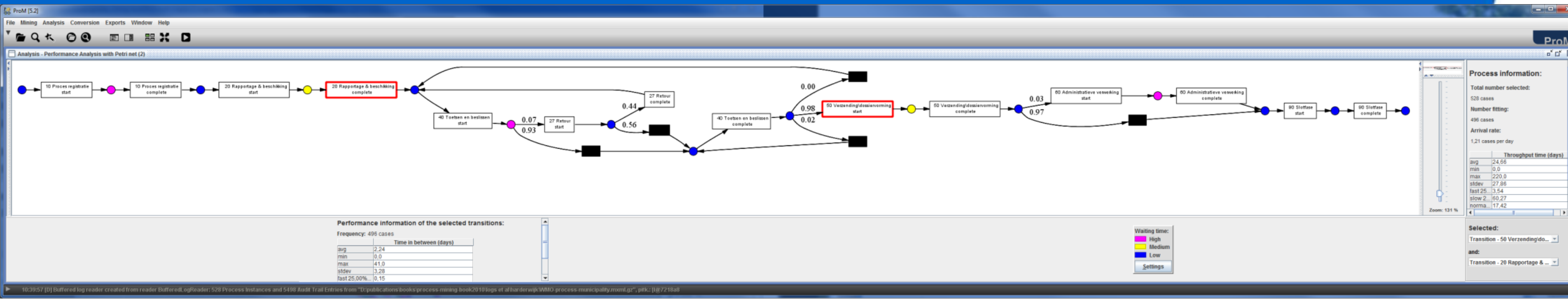


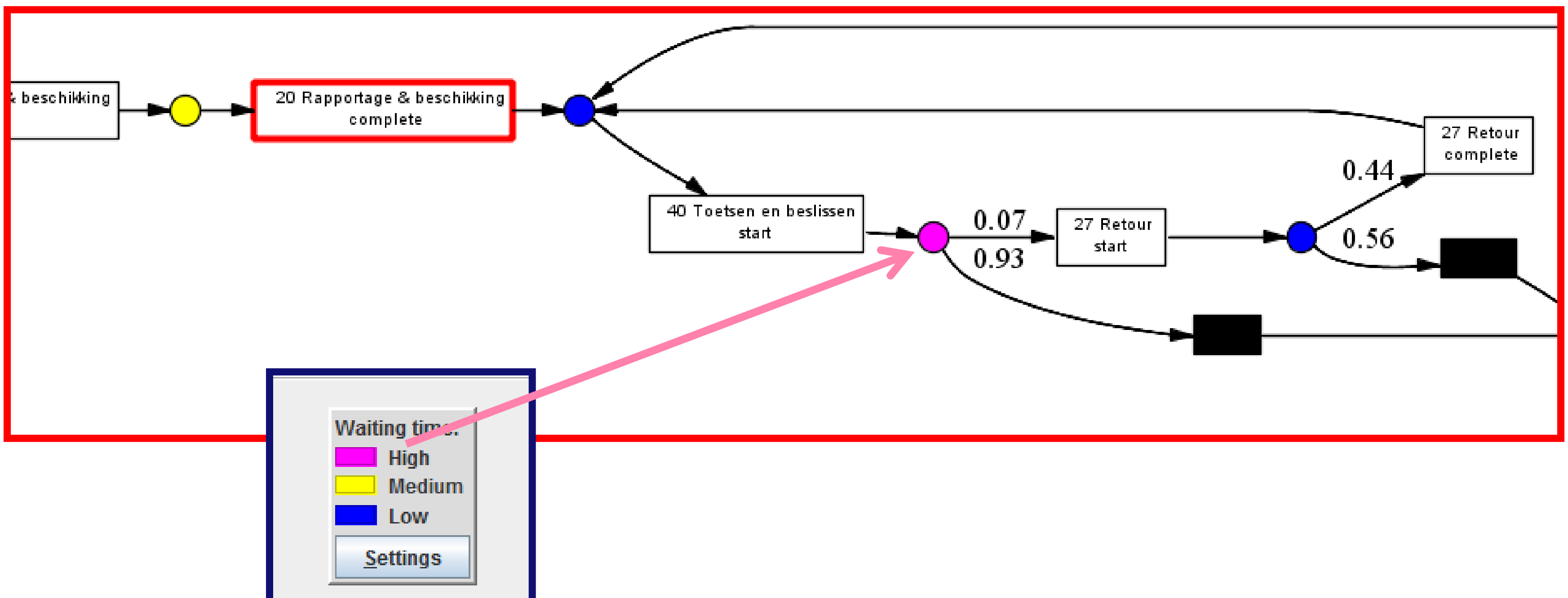
Conformance check WMO process (3/3)



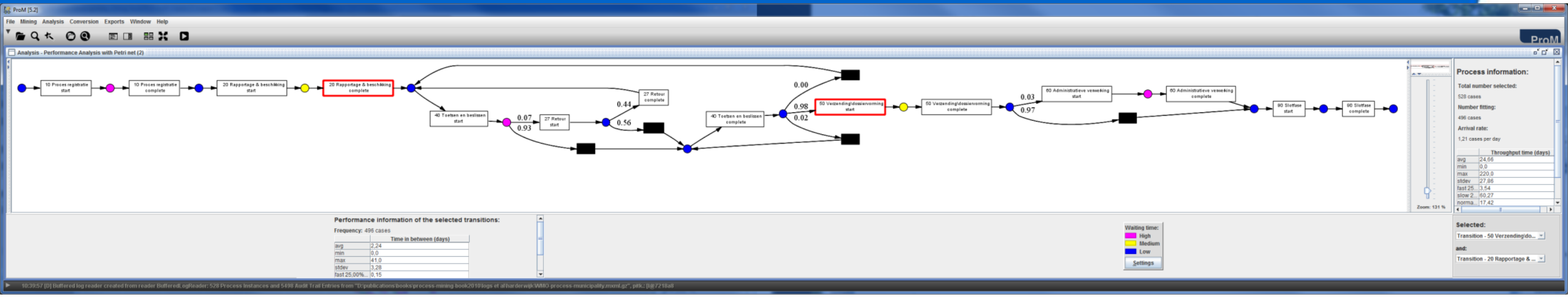
The fitness of the discovered process is 0.99521667. Of the 528 cases, 496 cases fit perfectly whereas for 32 cases there are missing or remaining tokens.

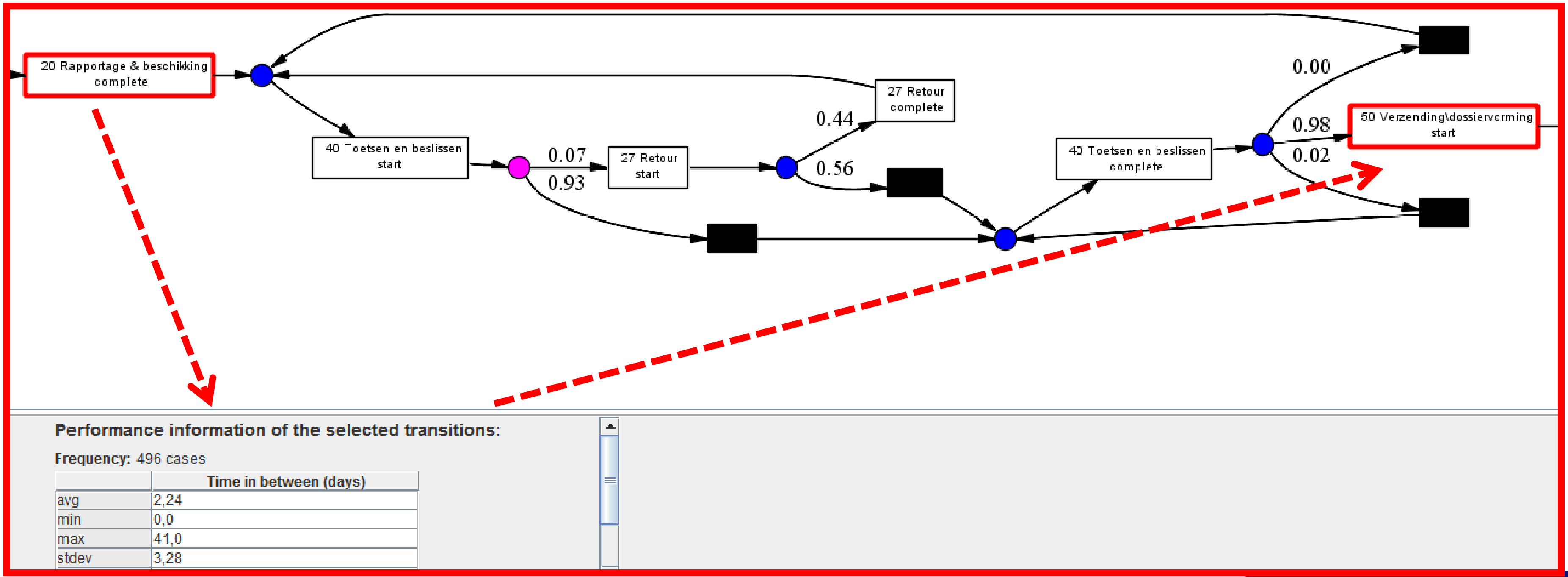
Bottleneck analysis WMO process (1/3)



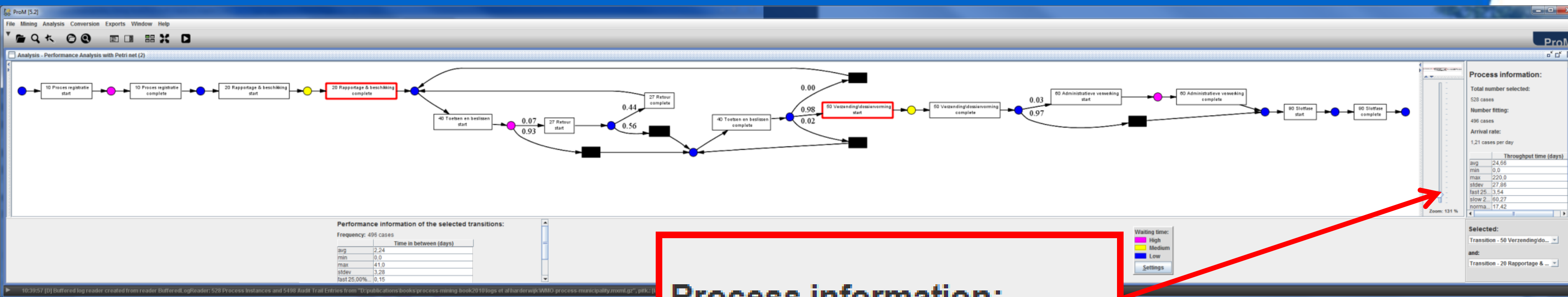


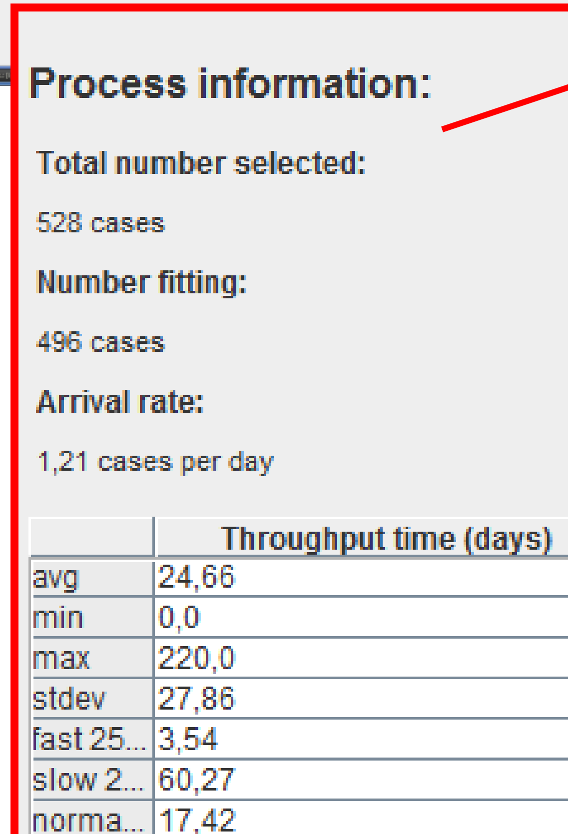
Bottleneck analysis WMO process (2/3)





Bottleneck analysis WMO process (3/3)




Process information:

Total number selected:

528 cases

Number fitting:

496 cases

Arrival rate:

1,21 cases per day

	Throughput time (days)
avg	24,66
min	0,0
max	220,0
stdev	27,86
fast 25...	3,54
slow 2...	60,27
norma...	17,42

flow time of approx. 25 days with a standard deviation of approx. 28

Two additional Lasagna processes

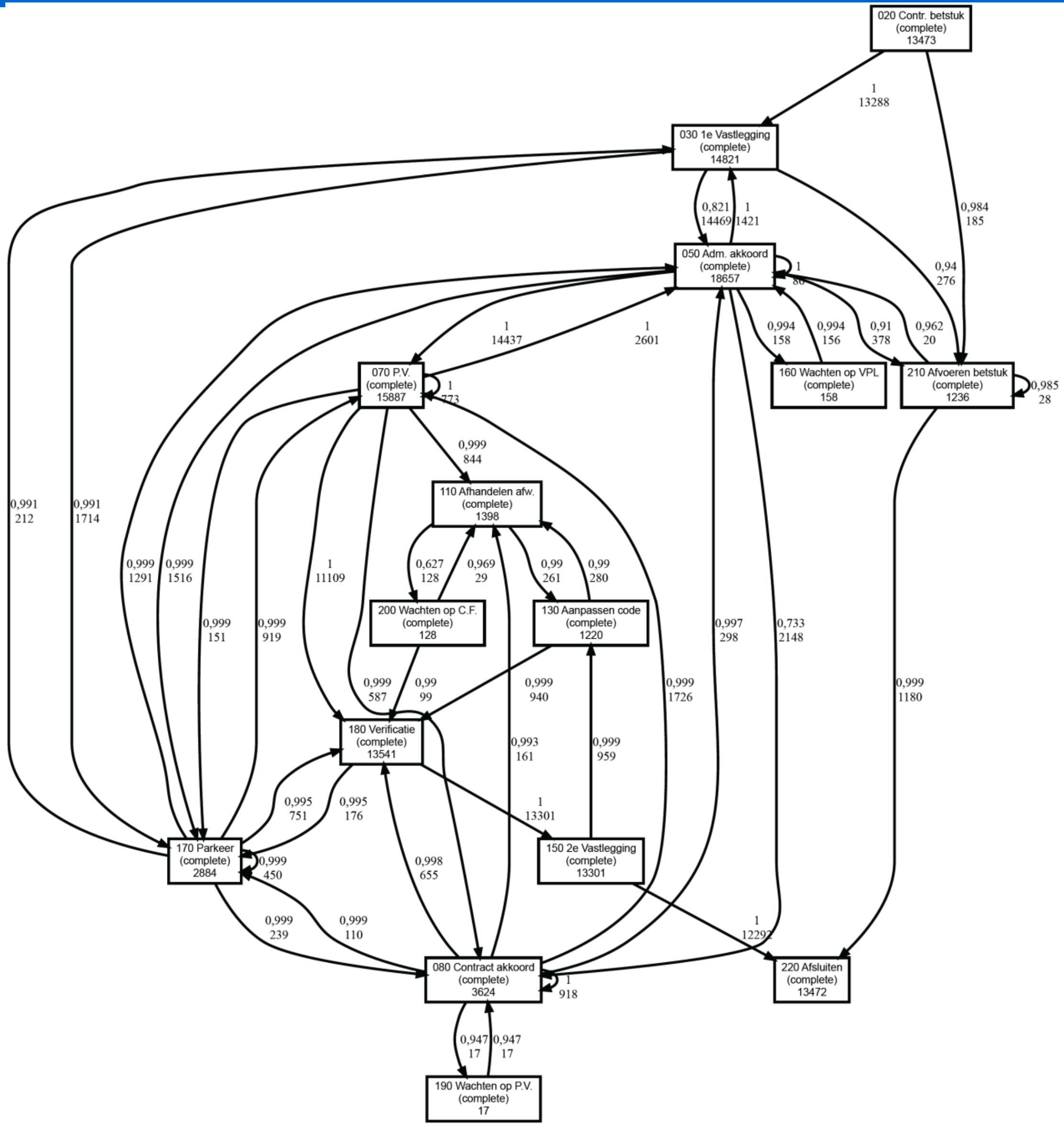
RWS (“Rijkswaterstaat”) process

WOZ (“Waardering Onroerende Zaken”) process

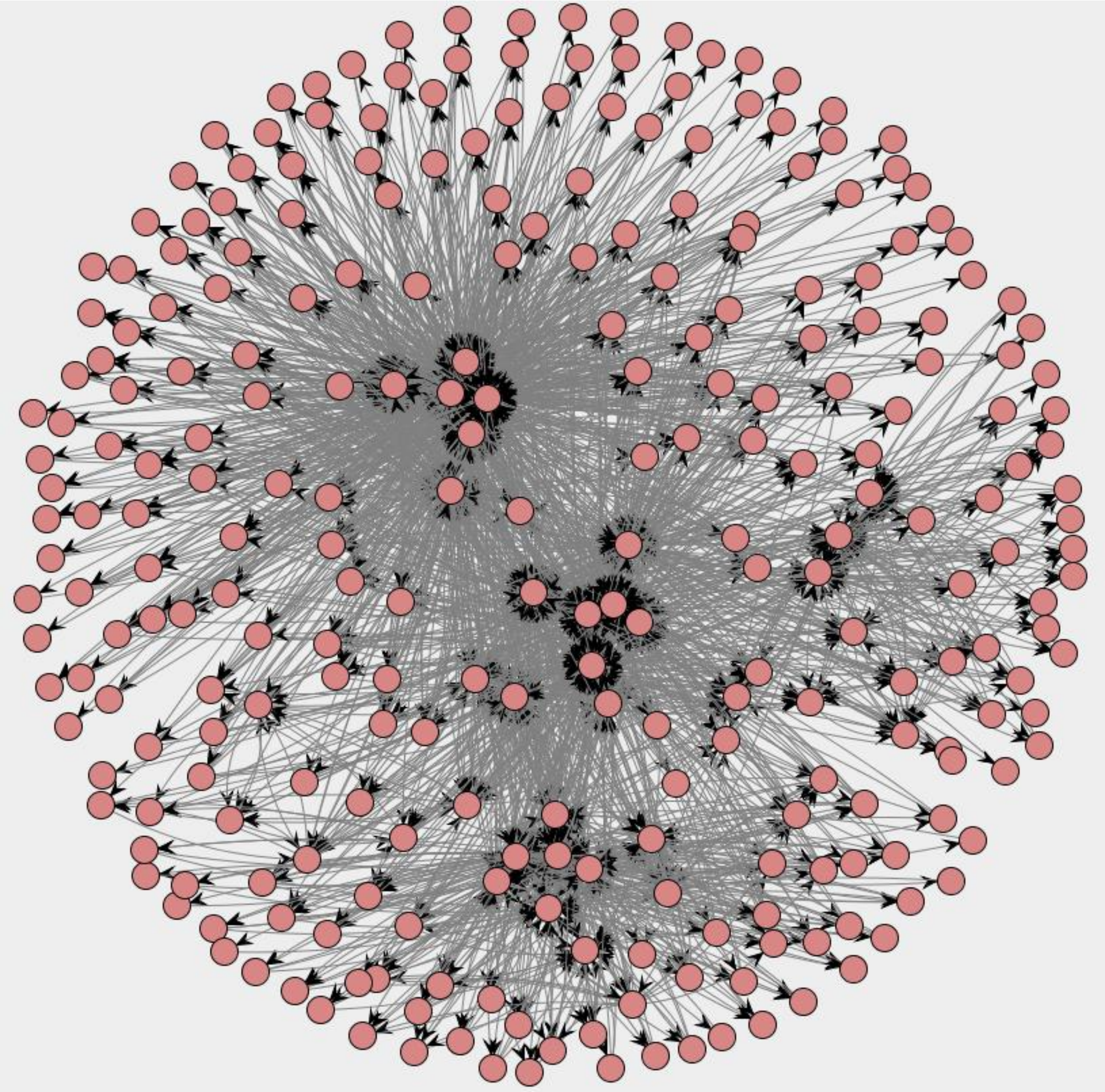
RWS Process

- The Dutch national public works department, called "Rijkswaterstaat" (RWS), has twelve provincial offices. We analyzed the handling of invoices in one of these offices.
- The office employs about 1,000 civil servants and is primarily responsible for the construction and maintenance of the road and water infrastructure in its province.
- To perform its functions, the RWS office subcontracts various parties such as road construction companies, cleaning companies, and environmental bureaus. Also, it purchases services and products to support its construction, maintenance, and administrative activities.

C-net discovered using heuristic miner

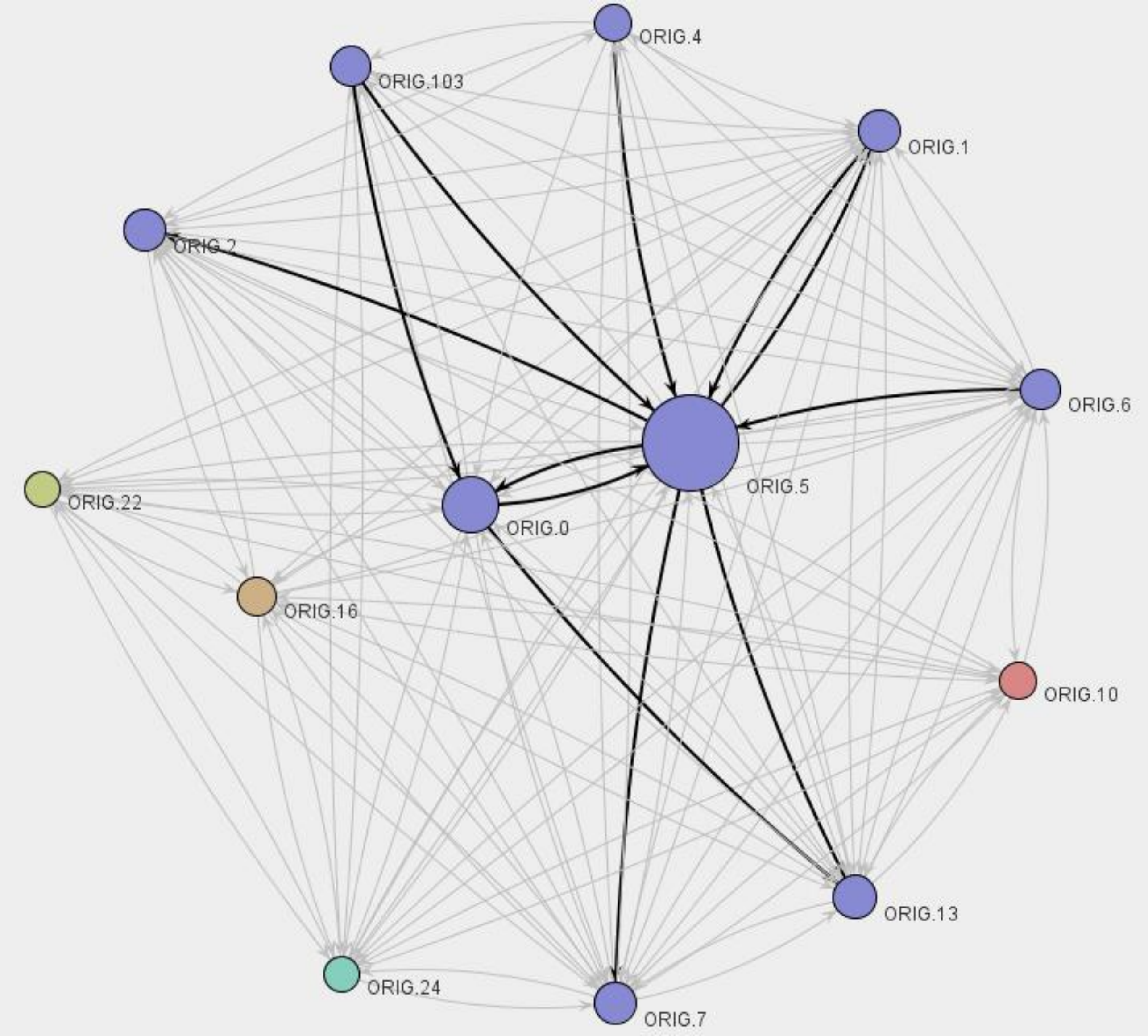


Social network constructed based on handovers of work



Each of the 271 nodes corresponds to a civil servant. Two civil servants are connected if one executed an activity causally following an activity executed by the other civil servant

Social network consisting of civil servants that executed more than 2000 activities in a 9 month period.

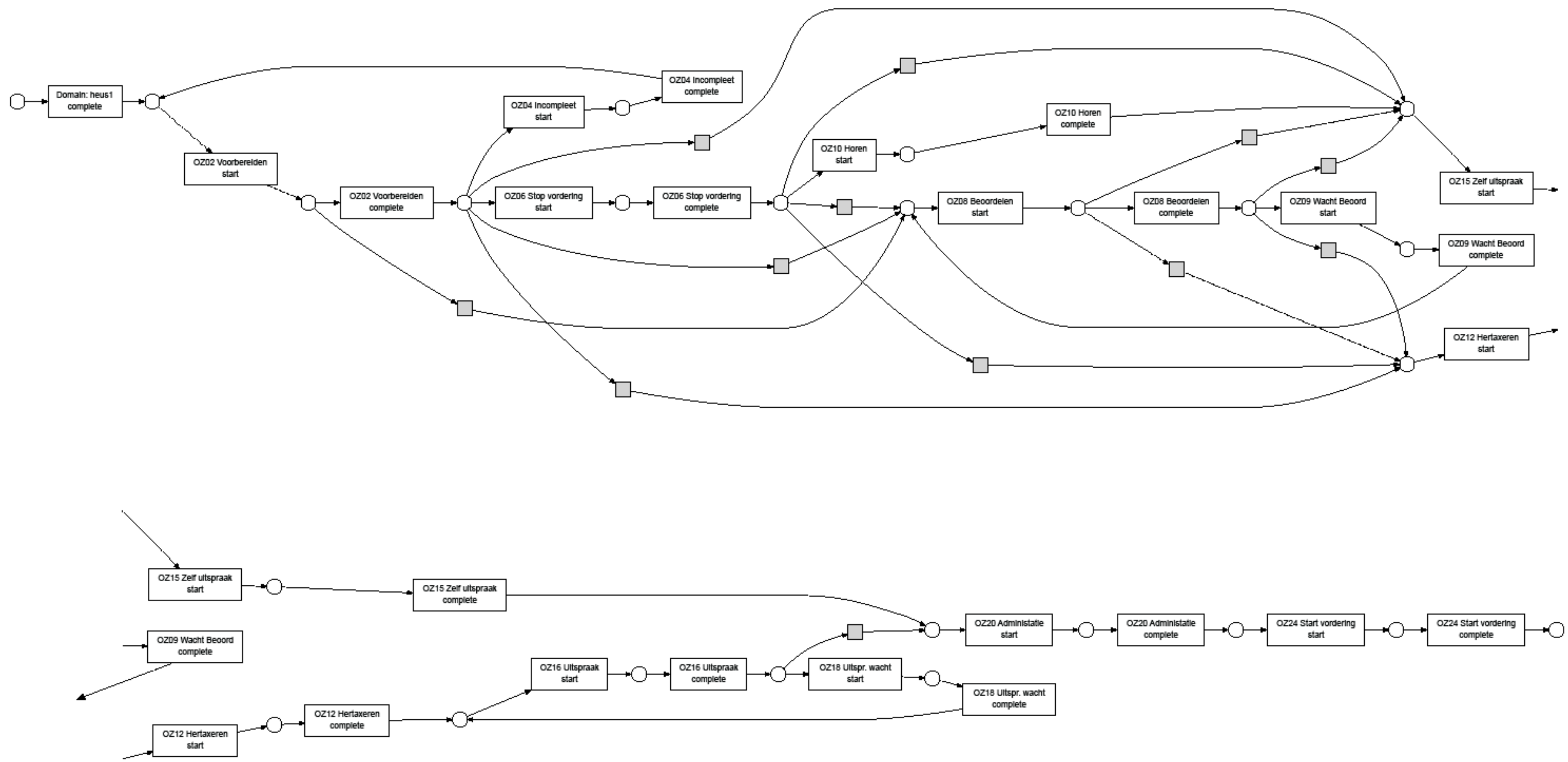


The darker arcs indicate the strongest relationships in the social network. Nodes having the same color belong to the same clique.

WOZ process

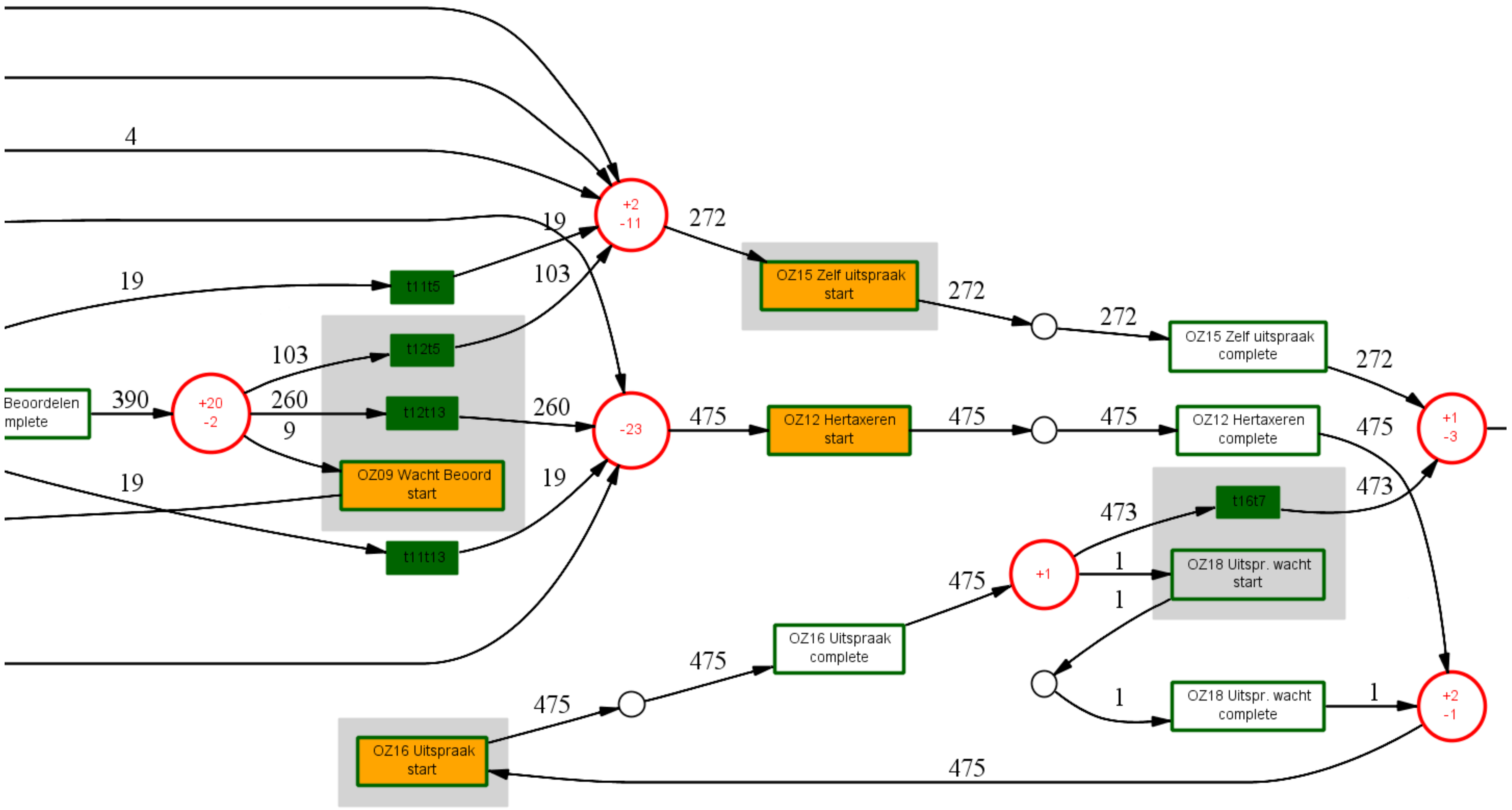
- Event log containing information about 745 objections against the so-called WOZ ("Waardering Onroerende Zaken") valuation.
- Dutch municipalities need to estimate the value of houses and apartments. The WOZ value is used as a basis for determining the real-estate property tax.
- The higher the WOZ value, the more tax the owner needs to pay. Therefore, there are many objections (i.e., appeals) of citizens that assert that the WOZ value is too high.
- "WOZ process" discovered for another municipality (i.e., different from the one for which we analyzed the WMO process).

Discovered process model

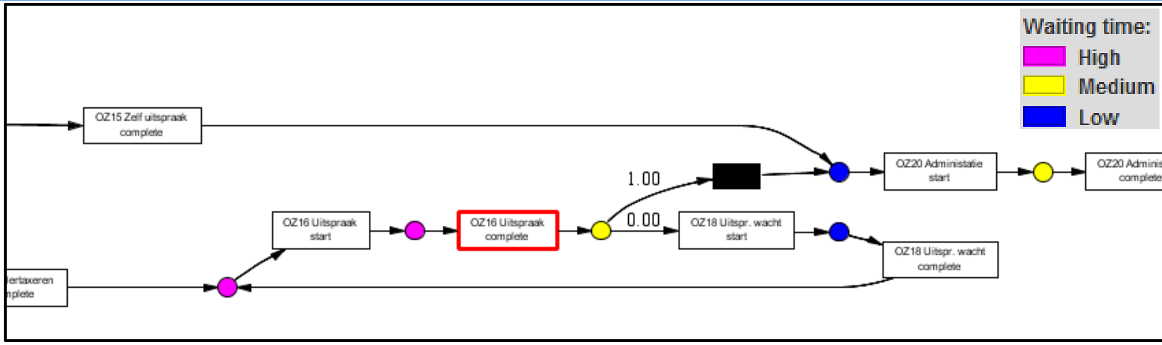


The log contains events related to 745 objections against the so-called WOZ valuation. These 745 objections generated 9583 events. There are 13 activities. For 12 of these activities both start and complete events are recorded. Hence, the WF-net has 25 transitions.

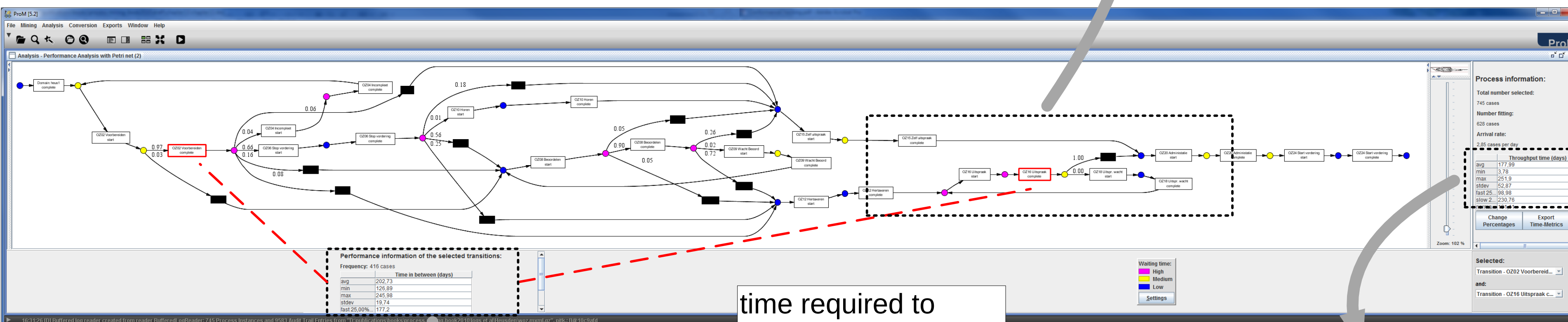
Conformance checker:
(fitness is 0.98876214)



Performance analysis



bottleneck detection: places are colored based on average durations



time required to move from one activity to another

Performance information of the selected transitions:

Frequency: 416 cases

	Time in between (days)
avg	202,73
min	126,89
max	245,98
stdev	19,74
fast 25,00%...	177,2

information on total flow time

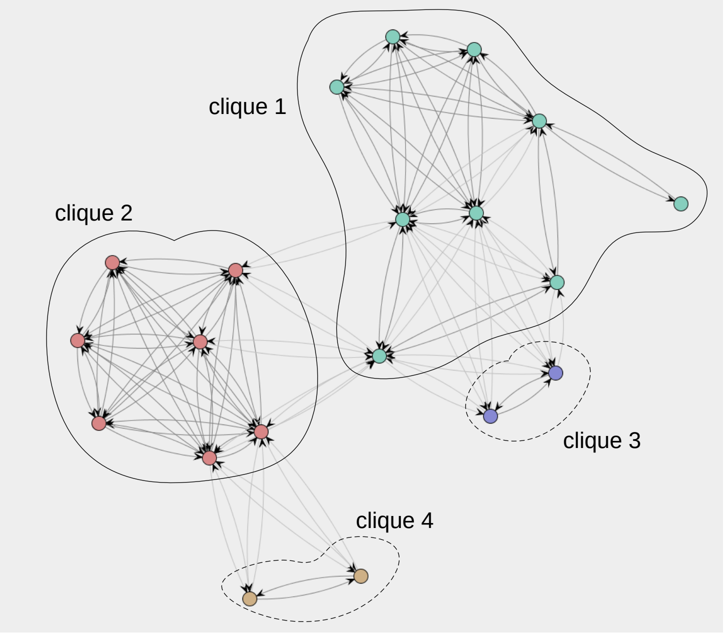
Arrival rate:

2,85 cases per day

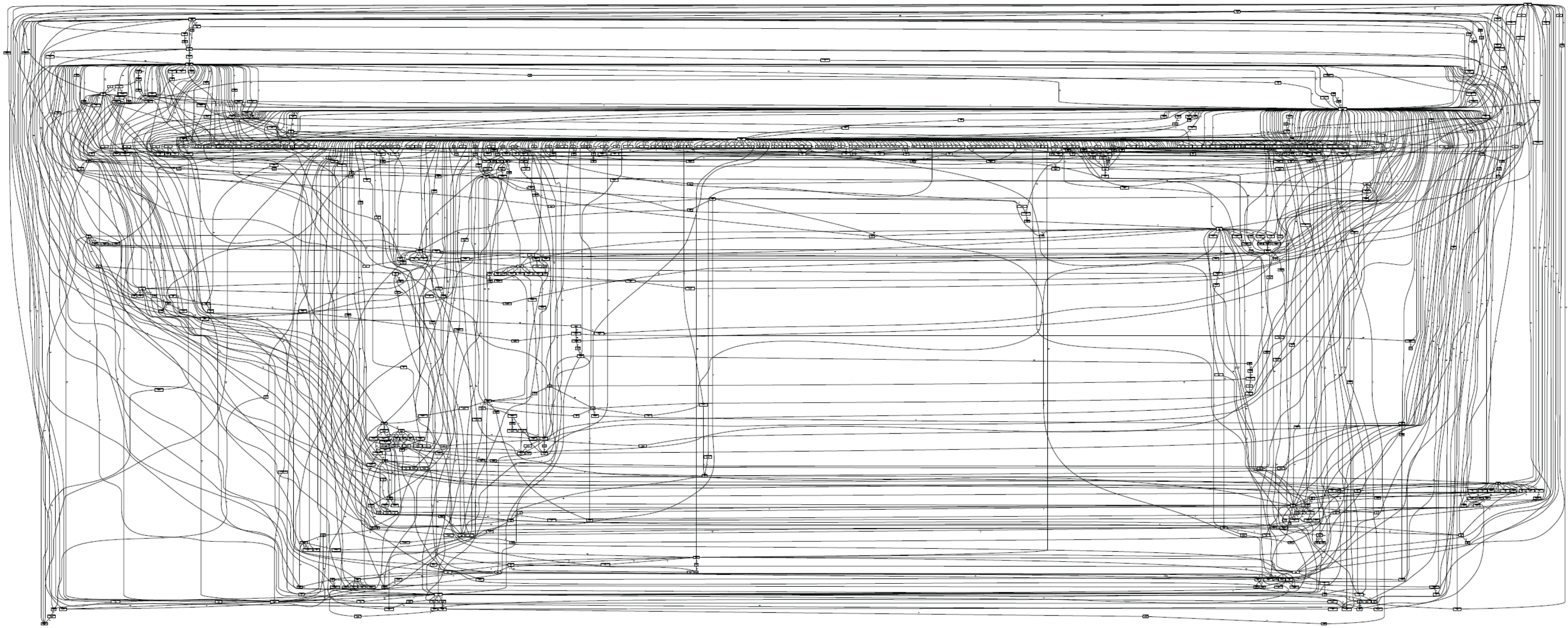
	Throughput time (days)
avg	177,99
min	3,78
max	251,9
stdev	52,87
fast 25...	98,98
slow 2...	230,76
norma...	191,11

Resource-activity matrix (four groups discovered)

user	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
user 1	0	0	51	0	0	0	0	0	0	0	0	0	0
user 2	1	2	0	0	2	0	0	0	0	38	0	69	0
user 3	0	9	0	0	0	0	0	0	0	0	0	0	0
user 4	2	0	0	0	0	0	0	0	0	0	0	0	0
user 5	117	0	4	0	3	0	0	0	0	1	0	20	6
user 6	172	6	14	0	7	3	0	0	1	2	0	48	53
user 7	1	41	8	14	275	8	8	865	55	180	0	128	5
user 8	2	868	7	6	105	0	0	79	266	441	0	844	3
user 9	90	0	2	0	1	2	0	0	1	2	0	27	28
user 10	0	0	0	899	0	0	0	0	0	0	0	0	1019
user 11	336	1	3	1	4	2	0	0	0	1	0	18	23
user 12	1	645	13	21	419	3	0	3	217	281	1	334	9
user 13	0	1	0	0	0	0	0	0	0	0	0	0	0
user 14	0	0	0	0	0	0	0	0	0	1	0	0	0
user 15	0	0	0	0	0	0	0	2	2	0	0	2	0
user 16	1	3	3	2	1	0	0	1	2	3	1	0	0
user 17	0	4	0	0	0	0	0	0	0	0	0	0	0
user 18	9	0	0	0	0	0	0	0	0	0	0	0	0
user 19	13	1	0	0	1	0	0	0	0	0	0	4	0
user 20	0	0	0	21	0	0	0	0	0	0	0	0	258

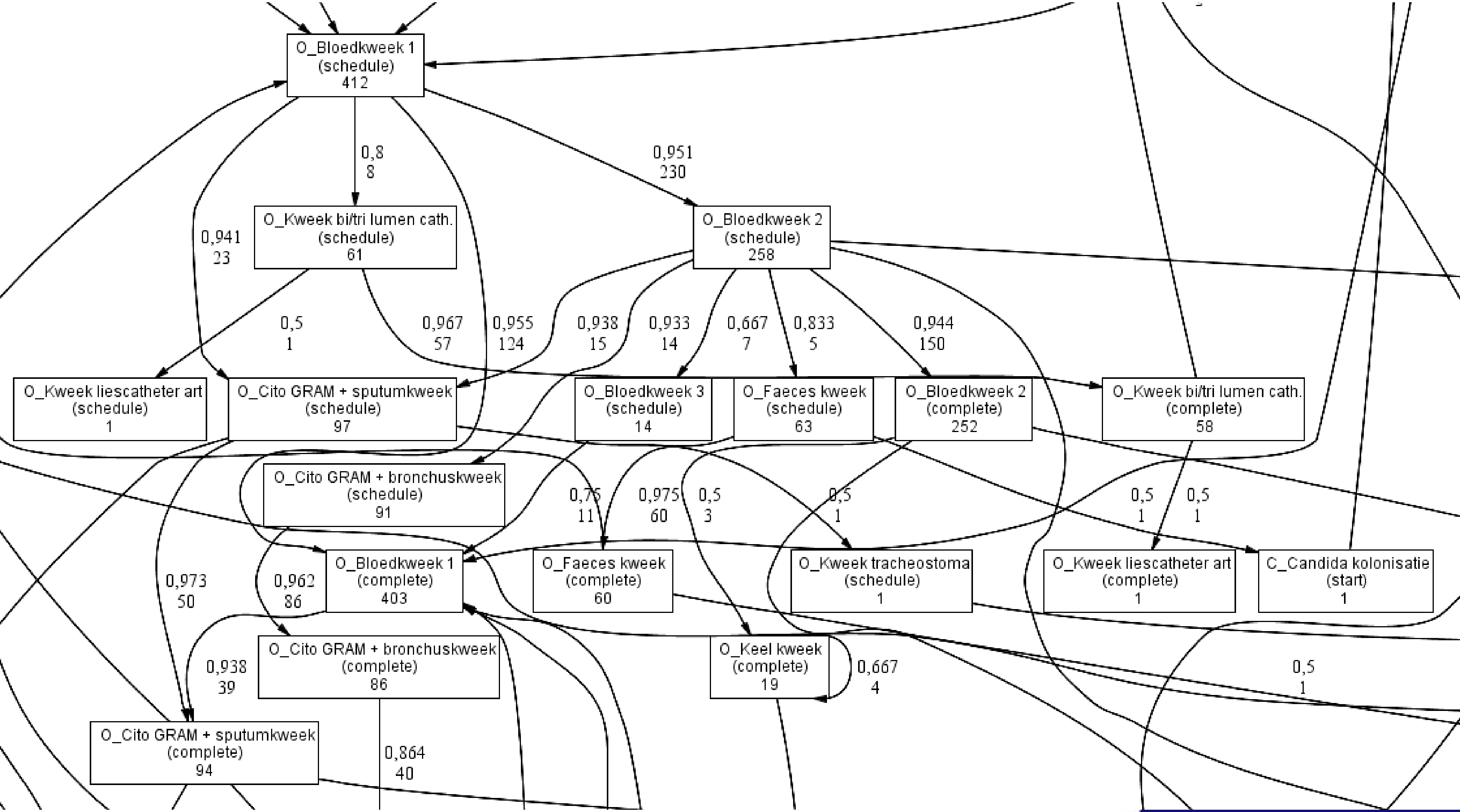


Example of a Spaghetti process

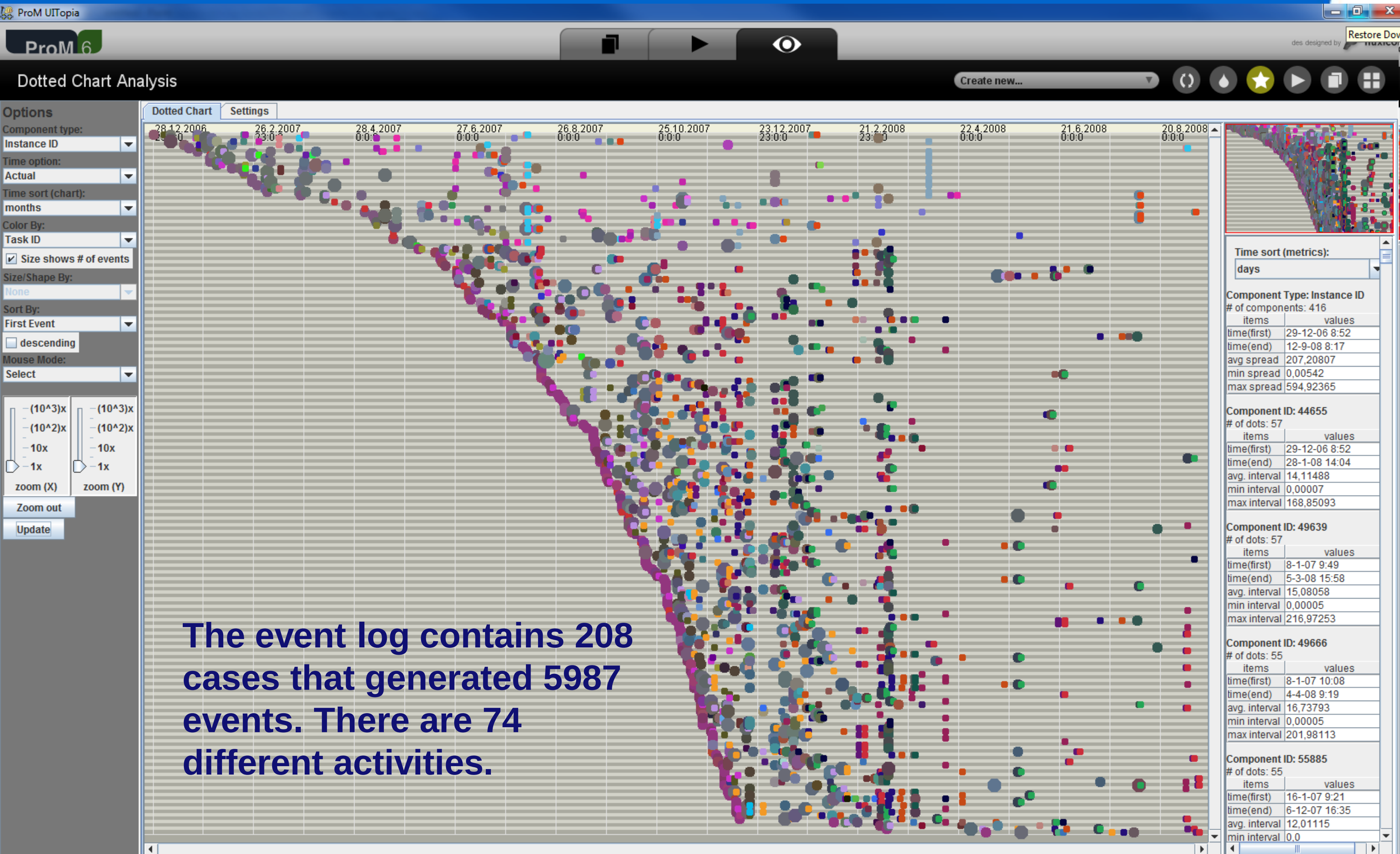


Spaghetti process describing the diagnosis and treatment of 2765 patients in a Dutch hospital. The process model was constructed based on an event log containing 114,592 events. There are 619 different activities (taking event types into account) executed by 266 different individuals (doctors, nurses, etc.).

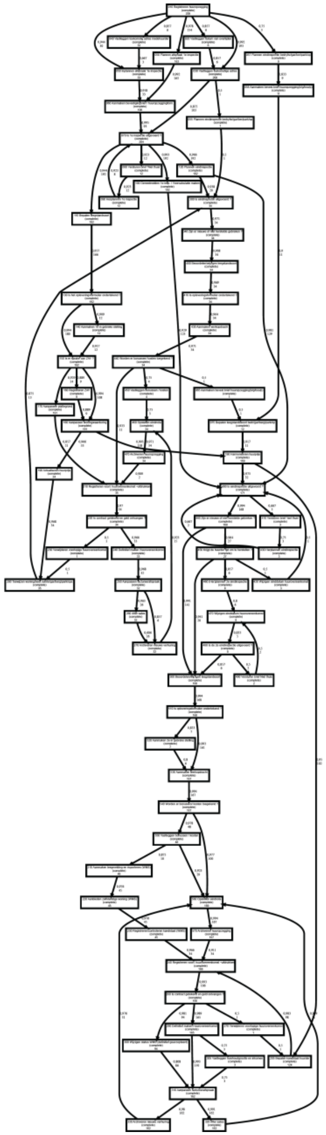
Fragment
18 activities of the 619 activities (2.9%)

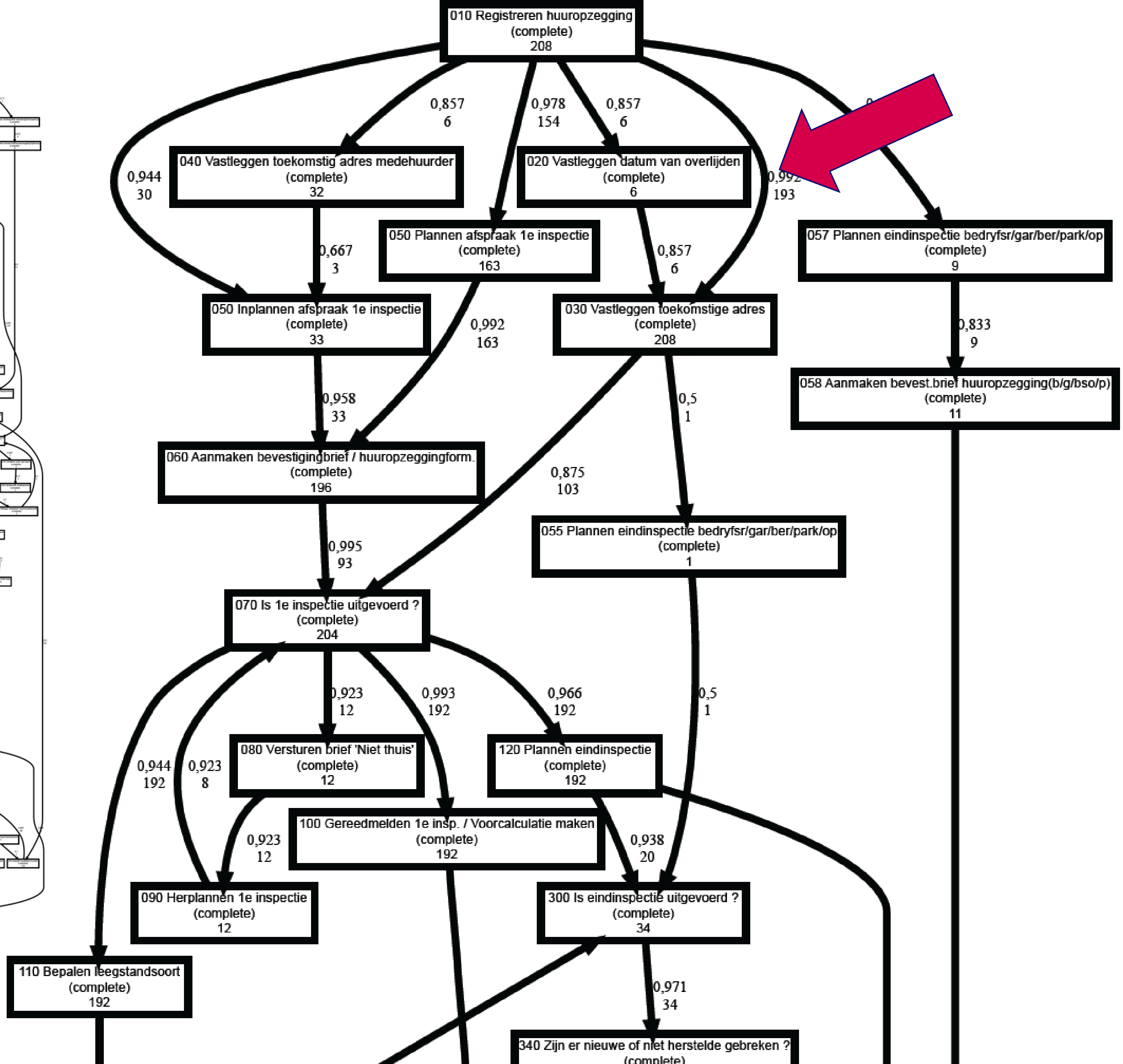


Another example
(event log of Dutch housing agency)



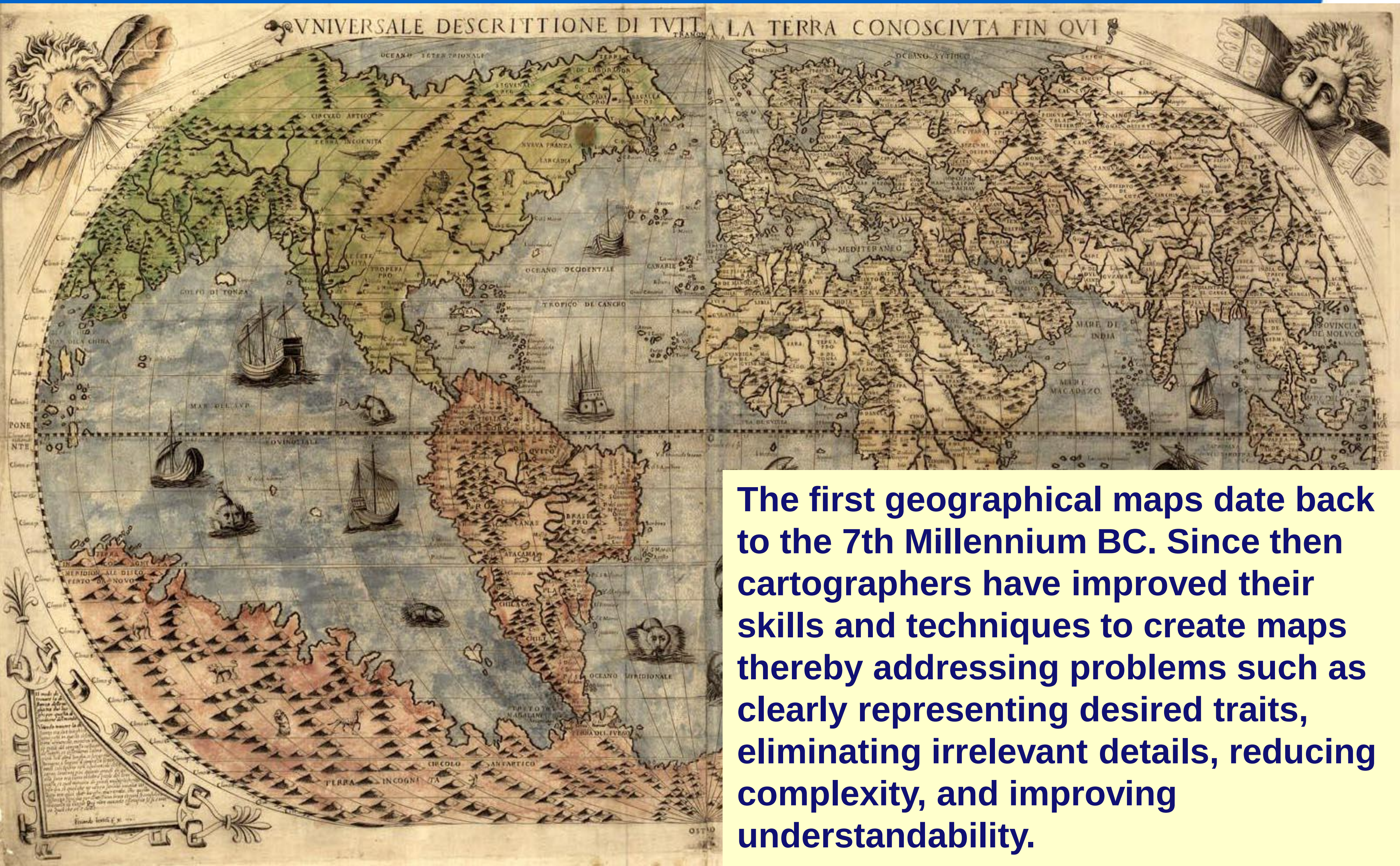
The event log contains 208 cases that generated 5987 events. There are 74 different activities.



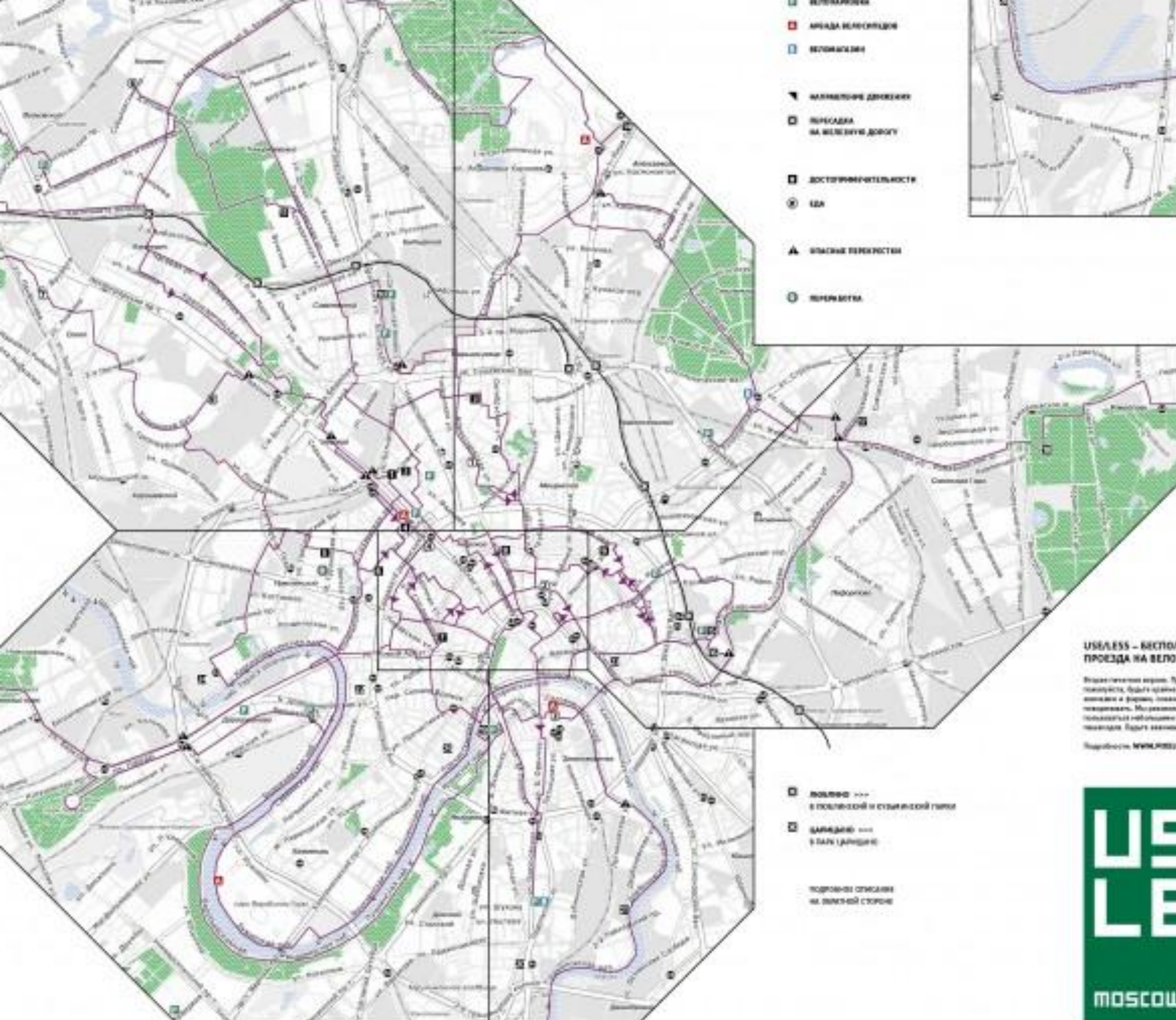


010 Registreren huuropzegging
(complete)
208
0,857
6
0,978
154
0,857
6
0,944
30
040 Vastleggen toekomstig adres medehuurder
(complete)
32
020 Vastleggen datum van overlijden
(complete)
6
0,992
193
057 Plannen eindinspectie bedryfsr/gar/ber/park/op
(complete)
9
0,667
3
050 Plannen afspraak 1e inspectie
(complete)
163
0,857
6
050 Inplannen afspraak 1e inspectie
(complete)
33
0,992
163
030 Vastleggen toekomstige adres
(complete)
208
0,833
9
058 Aanmaken bevest.brief huuropzegging(b/g/bso/p)
(complete)
11
0,958
33
0,5
1
060 Aanmaken bevestigingbrief / huuropzeggingform.
(complete)
196
0,875
103
055 Plannen eindinspectie bedryfsr/gar/ber/park/op
(complete)
1
0,995
93
070 Is 1e inspectie uitgevoerd ?
(complete)
204
0,923
12
0,993
192
0,966
192
0,5
1
0,944
192
0,923
8
080 Versturen brief 'Niet thuis'
(complete)
12
120 Plannen eindinspectie
(complete)
192
100 Gereedmelden 1e insp. / Voorcalculatie maken
(complete)
192
0,923
12
0,938
20
090 Herplannen 1e inspectie
(complete)
12
300 Is eindinspectie uitgevoerd ?
(complete)
34
110 Bepalen leegstandsoort
(complete)
192
0,971
34
340 Zijn er nieuwe of niet herstelde gebreken ?
(complete)

Business process maps



The first geographical maps date back to the 7th Millennium BC. Since then cartographers have improved their skills and techniques to create maps thereby addressing problems such as clearly representing desired traits, eliminating irrelevant details, reducing complexity, and improving understandability.



МАРШРУТ В КОЛОМЕНСКОЕ

АРЕНДА ВЕЛОСИПЕДОВ

ВЕЛОМАГАЗИНЫ

НАПРАВЛЕНИЕ ДВИЖЕНИЯ

ПЕРЕСАДКА НА ЖЕЛЕЗНУЮ ДОРОГУ

ДОСТОПРИМЕЧАТЕЛЬНОСТИ

ЕДА

ОПАСНЫЕ ПЕРЕКРЕСТКИ

ПЕРЕРАБОТКА

ЛЮБЛИНО >>>
В ЛЮБЛИНСКИЙ И КУЗЬМИНСКИЙ ПАРКИ

ЦАРИЦЫНО >>>
В ПАРК ЦАРИЦЫНО

ПОДРОБНОЕ ОПИСАНИЕ НА ОБРАТНОЙ СТОРОНЕ

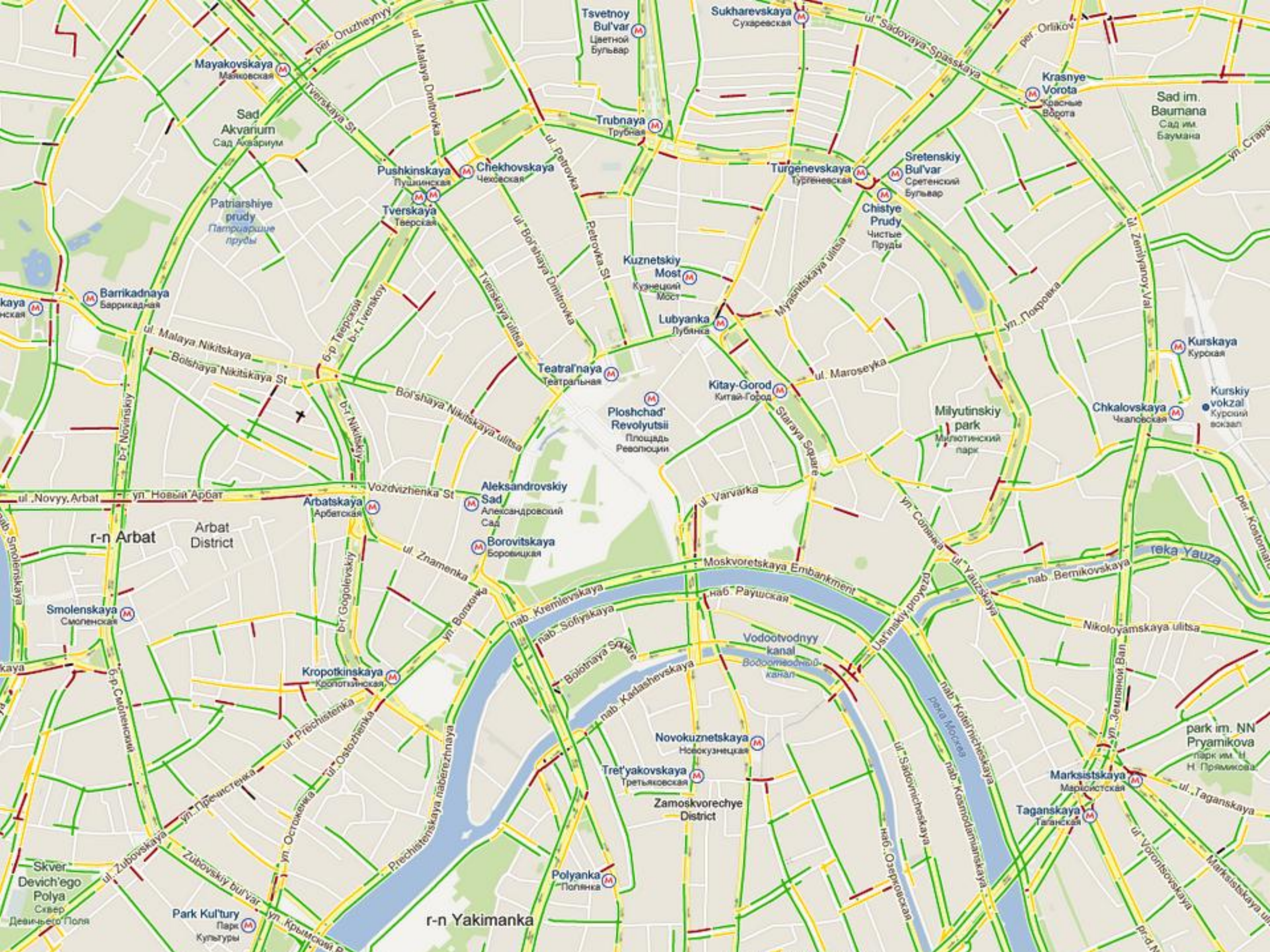
USE/LESS – БЕСПОЛЕЗНАЯ СХЕМА ПРОЕЗДА НА ВЕЛОСИПЕДЕ ПО МОСКВЕ

[illegible]

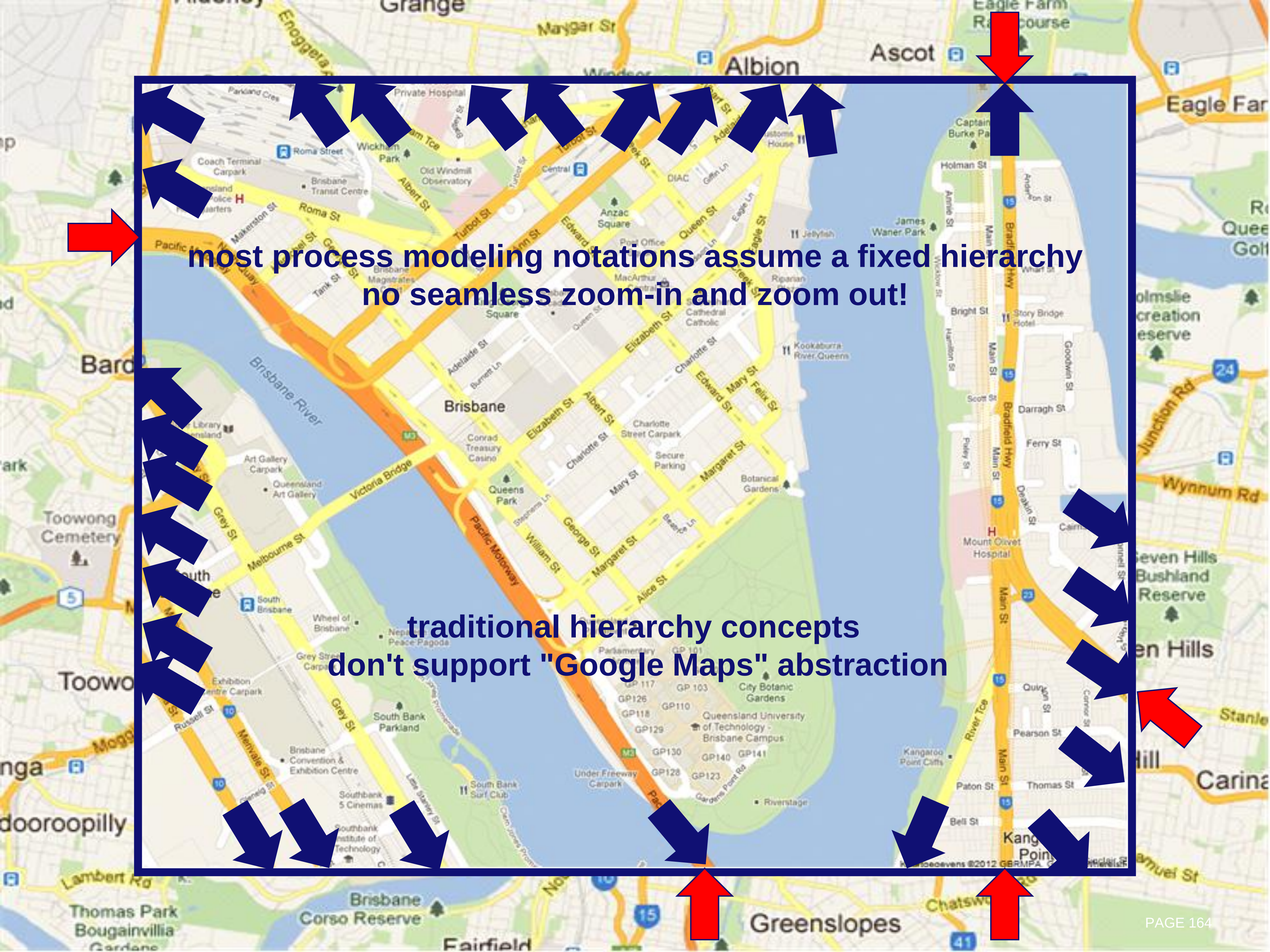
Подробности: [illegible]

USE LESS

MOSCOW

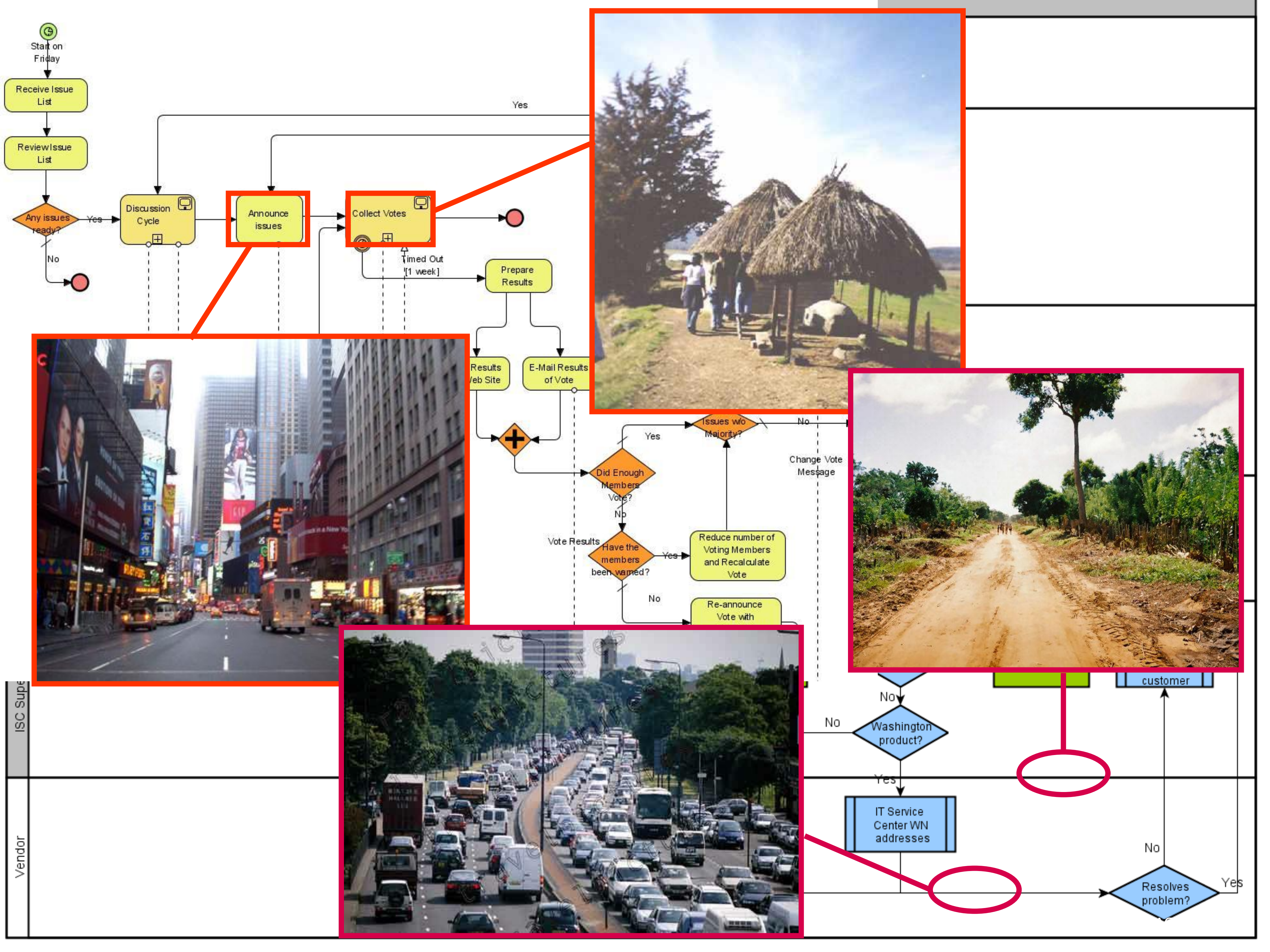


Tsvetnoy Bul'var
Цветной Бульвар
Sukharevskaya
Сухаревская
ul. Sadovaya Spasskaya
per. Orlikov
per. Oruzheynyy
Mayakovskaya
Маяковская
ul. Malaya Dmitrovka
Tverskaya St
Krasnye Vorota
Красные Ворота
Sad im. Baumana
Сад им. Баумана
Sad Akvarium
Сад Аквариум
Trubnaya
Трубная
ul. Petrovka
Pushkinskaya
Пушкинская
Chekhovskaya
Чеховская
Turgenevskaya
Тургеневская
Sretenskiy Bul'var
Сретенский Бульвар
Chistye Prudy
Чистые Пруды
Tverskaya
Тверская
Patriarshiye prudy
Патриаршие пруды
ul. Bol'shaya Dmitrovka
Petrovka St
Kuznetskiy Most
Кузнецкий Мост
Myasnitskaya ulitsa
ul. Zemlyanoy Val
Barrikadnaya
Баррикадная
Tverskaya ulitsa
Lubyanka
Лубянка
ul. Покровка
Kurskaya
Курская
ul. Malaya Nikitskaya
Bolshaya Nikitskaya St
Teatral'naya
Театральная
Kitay-Gorod
Китай-Город
ul. Maroseyka
Kurskiy vokzal
Курский вокзал
Chkalovskaya
Чкаловская
Bol'shaya Nikitskaya ulitsa
Ploshchad' Revolyutsii
Площадь Революции
Staraya Square
Milyutinskiy park
Милютинский парк
b-r Novinskiy
b-r Nikitskiy
ul. Novyy Arbat
ул. Новый Арбат
Vozdvizhenka St
Aleksandrovskiy Sad
Александровский Сад
Arbatskaya
Арбатская
ul. Varvarka
ul. Solyanka
r-n Arbat
Arbat District
Borovitskaya
Боровицкая
ul. Znamenka
Moskvoretskaya Embankment
reka Yauza
nab. Bernikovskaya
ul. Yauzskaya
Ust'inskiy proyezd
nab. Kremlevskaya
наб. Раушская
nab. Sofiyskaya
Nikoloyamskaya ulitsa
Smolenskaya
Смоленская
b-r Gogolevskiy
ul. Volkhonka
Vodootvodnyy kanal
Водоотводный канал
Bolotnaya Square
Kropotkinskaya
Кропоткинская
nab. Kadashevskaya
ul. Prechistenka
ul. Ostozhenka
nab. Kotel'nicheskaya
park im. NN Pryamikova
парк им. Н. Н. Прямикова
Novokuznetskaya
Новокузнецкая
ul. Sadovnicheskaya
nab. Kosmodamianskaya
Tret'yakovskaya
Третьяковская
Marksistskaya
Марксистская
ul. Taganskaya
Zamoskvorechye District
Taganskaya
Таганская
Prechistenskaya naberezhnaya
ул. Пречистенка
ul. Zubovskaya
ul. Ostozhenka
наб. Озерковская
ul. Voronlsovskaya
Marksistskaya ulitsa
Skver Devich'ego Polya
Сквер Девичьего Поля
Zubovskiy bul'var
Polyanka
Полянка
Park Kul'tury
Парк Культуры
r-n Yakimanka
ул. Крымский



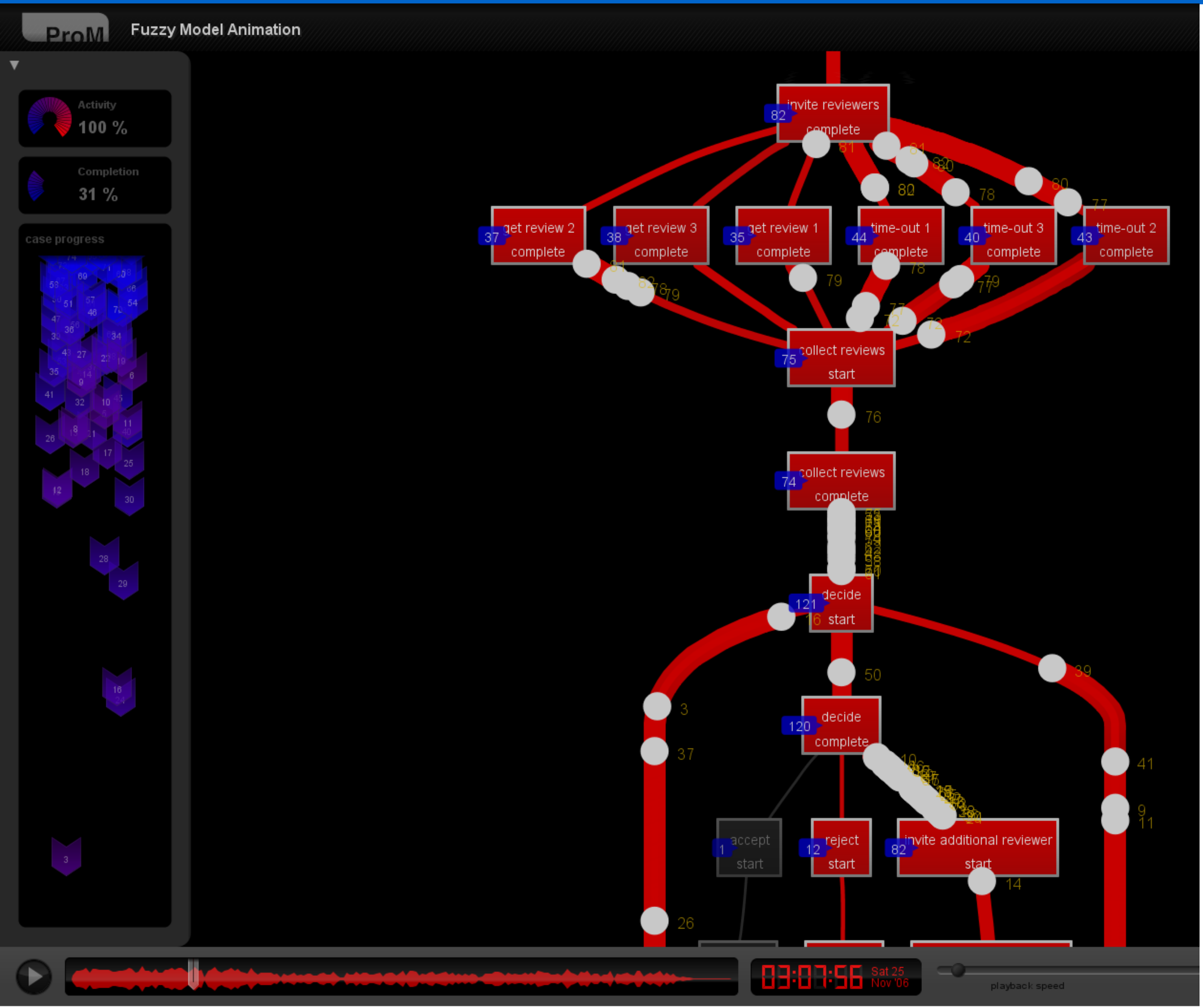
most process modeling notations assume a fixed hierarchy
no seamless zoom-in and zoom out!

traditional hierarchy concepts
don't support "Google Maps" abstraction



Start on Friday
Receive Issue List
Review Issue List
Any issues ready?
Yes
No
Discussion Cycle
Announce issues
Collect Votes
Yes
Timed Out [1 week]
Prepare Results
E-Mail Results of Vote
Issues w/o Majority?
No
Yes
Change Vote Message
Did Enough Members Vote?
No
Vote Results
Have the members been warned?
Yes
Reduce number of Voting Members and Recalculate Vote
No
Re-announce Vote with
customer
No
No
Washington product?
Yes
IT Service Center WN addresses
Vendor
No
Resolves problem?
Yes

Business process movies



Navigation

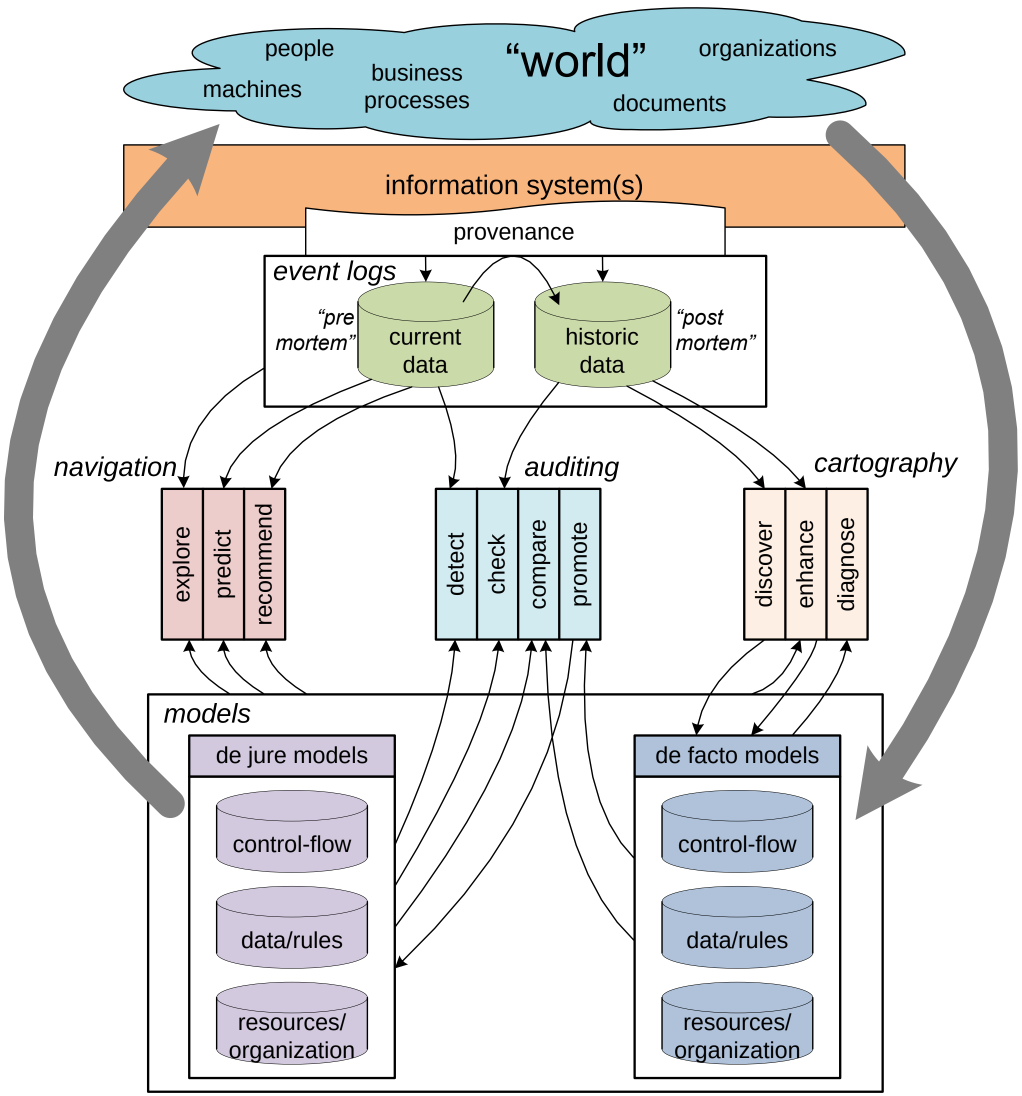
- Whereas a TomTom device is continuously showing the expected arrival time, users of today's information systems are often left clueless about likely outcomes of the cases they are working on.
- Car navigation systems provide directions and guidance without controlling the driver. The driver is still in control, but, given a goal (e.g. to get from A to B as fast as possible), the navigation system recommends the next action to be taken.
- Operational support provides TomTom functionality for business processes.

Recommend: How to get home ASAP? Take a left turn!

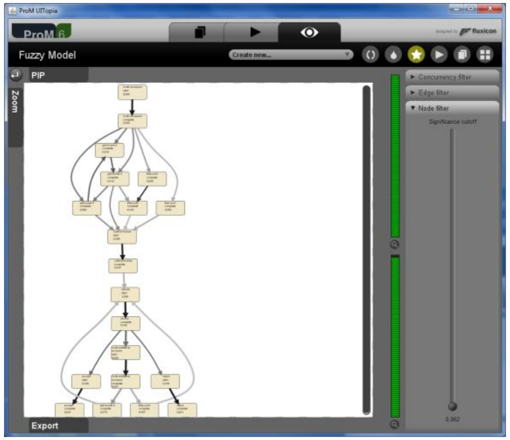
Detect: You drive too fast!

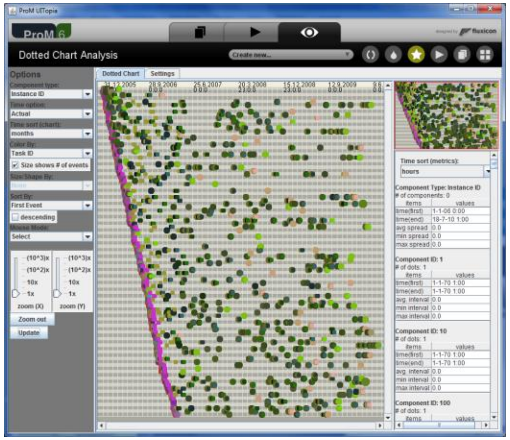
Predict: When will I be home? At 11.26!

Relating the process mining framework to cartography and navigation



600+ plug-ins available covering the whole process mining spectrum

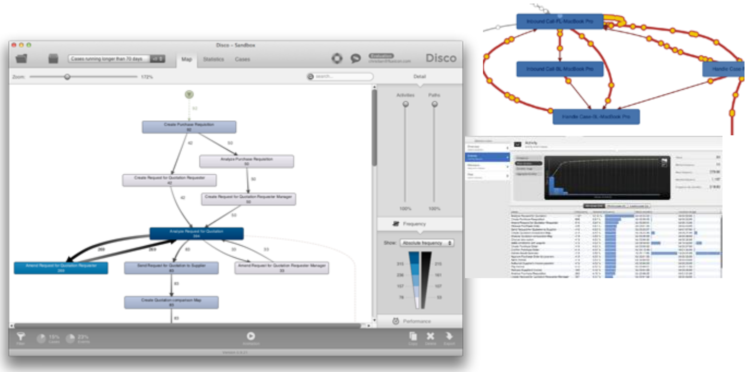


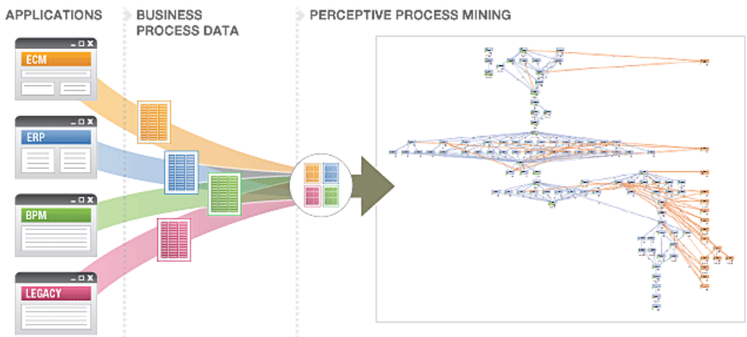


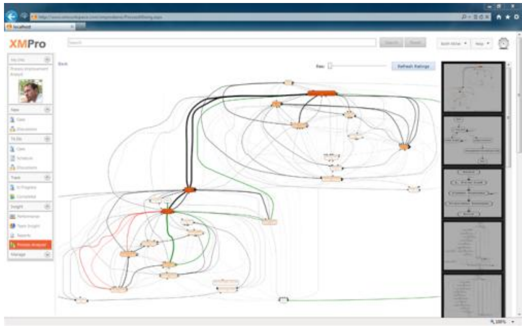
Download from: www.processmining.org

Commercial Alternatives

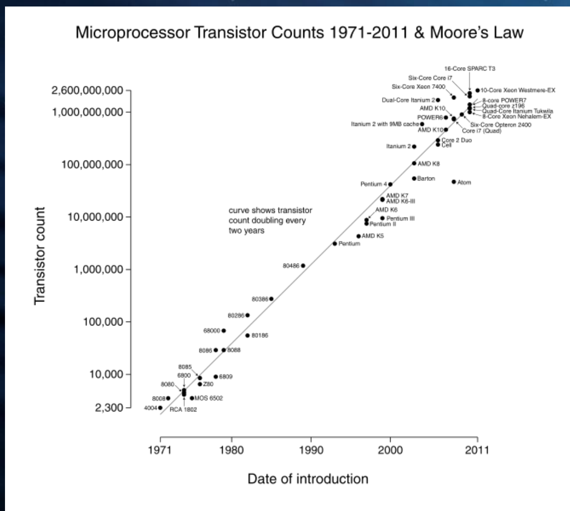
- Disco (Fluxicon)
- Perceptive Process Mining (before Futura Reflect and BPM|one)
- ARIS Process Performance Manager
- QPR ProcessAnalyzer
- Interstage Process Discovery (Fujitsu)
- Discovery Analyst (StereoLOGIC)
- XMAnalyzer (XMPro)
- …

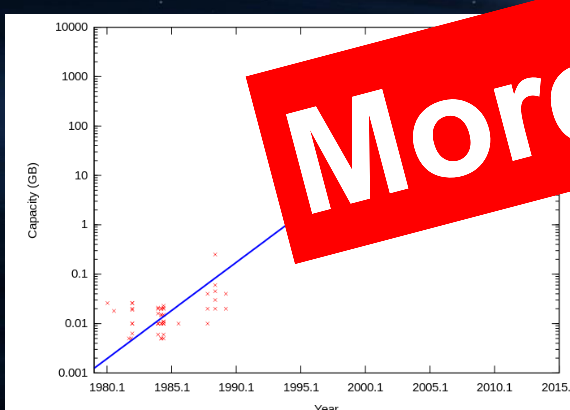


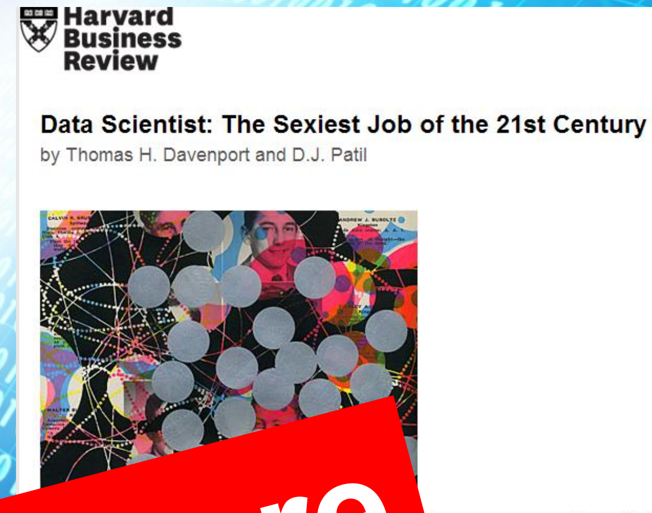




The Sexiest Job of the 21st century (thanks to Moore's Law)

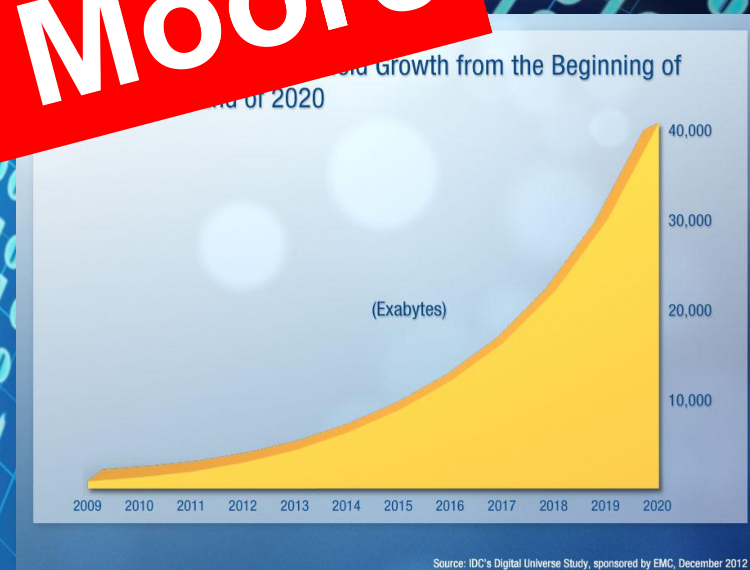





Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil



More from Moore

1965

Conclusion

- **Process leave traces in event logs. So, if you are interested in processes, use them!**
- **Process mining: challenging and highly relevant.**
- **Process discovery challenge**
  - **balancing between different objectives**
  - **only example behavior**
- **Conformance checking challenge**
  - **finding the most likely trace**
  - **dealing with silent/duplicate steps**
- **Relation to Google Maps and TomTom**
- **Eldorado for exciting research!**

Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

More and more information about business processes is recorded by information systems in the form of so-called "event logs". Despite the omnipresence of such data, most organizations diagnose problems based on fiction rather than facts. Process mining is an emerging discipline based on process model-driven approaches and data mining. It not only allows organizations to fully benefit from the information stored in their systems, but it can also be used to check the conformance of processes, detect bottlenecks, and predict execution problems.

Wil van der Aalst delivers the first book on process mining. It aims to be self-contained while covering the entire process mining spectrum from process discovery to operational support. In Part I, the author provides the basics of business process modeling 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.

Computer Science

ISBN 978-3-642-19344-6

▸ springer.com

van der Aalst

Process Mining

Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

www.processmining.org

www.win.tue.nl/ieeetfpm/

Springer