CS 4072 - Topics in CS
Process Mining

Lecture # 20

May 17, 2022

Spring 2022

FAST - NUCES, CFD Campus

Dr. Rabia Maqsood

rabia.maqsood@nu.edu.pk

Today's Topics

- Conformance Checking
  - Quick recap: Naïve approach & causal footprints
  - Token replay

- Project discussion

Approaches for Conformance Checking

Model and Log Fitness

Naïve approach

A naïve approach towards conformance checking would be to simply count the fraction of cases that can be “parsed completely” (i.e., the proportion of cases corresponding to firing sequences leading from [*start*] to [*end*]).

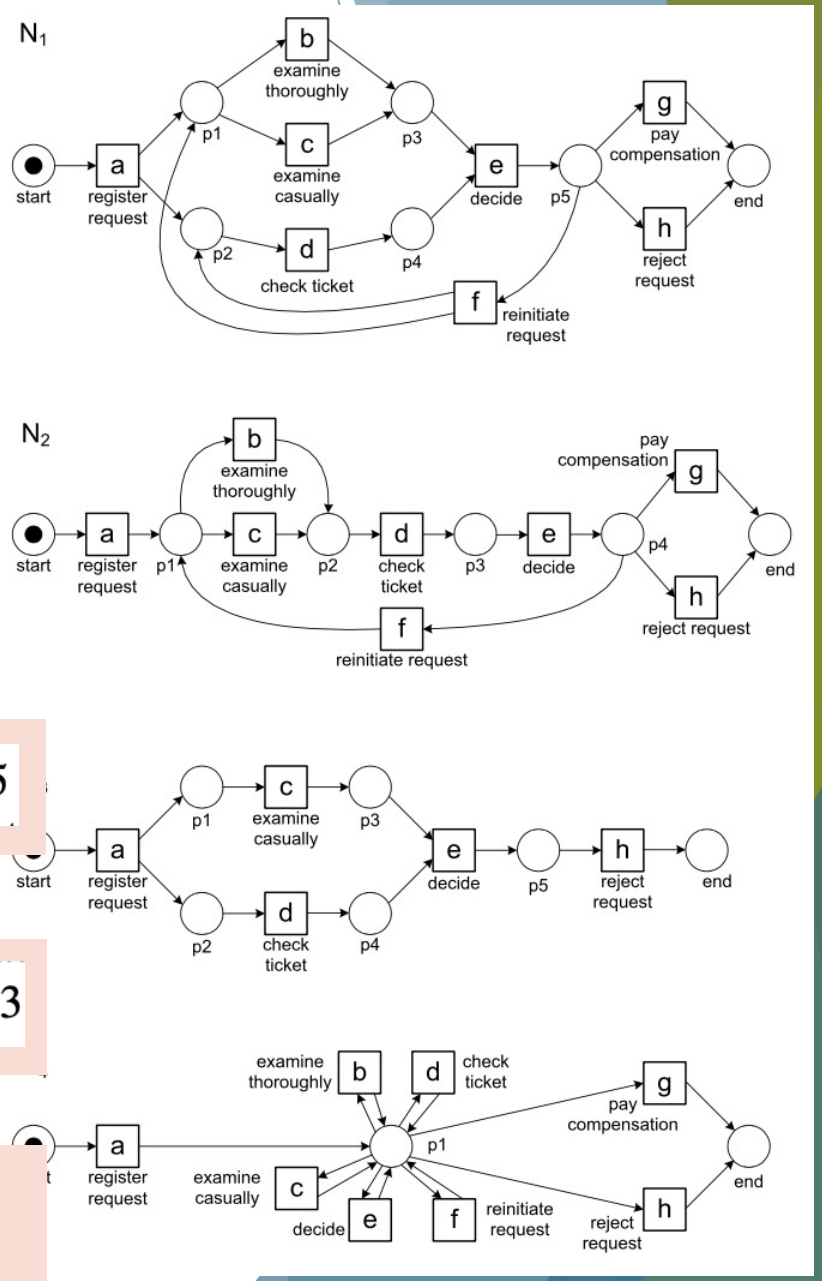
Frequency	Reference	Trace
455	σ_1	$\langle a, c, d, e, h \rangle$
191	σ_2	$\langle a, b, d, e, g \rangle$
177	σ_3	$\langle a, d, c, e, h \rangle$
144	σ_4	$\langle a, b, d, e, h \rangle$
111	σ_5	$\langle a, c, d, e, g \rangle$
82	σ_6	$\langle a, d, c, e, g \rangle$
56	σ_7	$\langle a, d, b, e, h \rangle$
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}	$\langle a, d, c, e, f, d, b, e, f, c, d, e, f, d, b, \ldots$

Fitness of N1 is $\frac{1391}{1391} = 1$

Fitness of N2 is $\frac{948}{1391} = 0.6815$

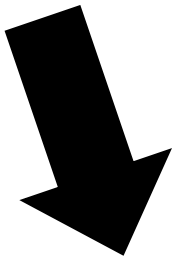
Fitness of N3 is $\frac{632}{1391} = 0.4543$

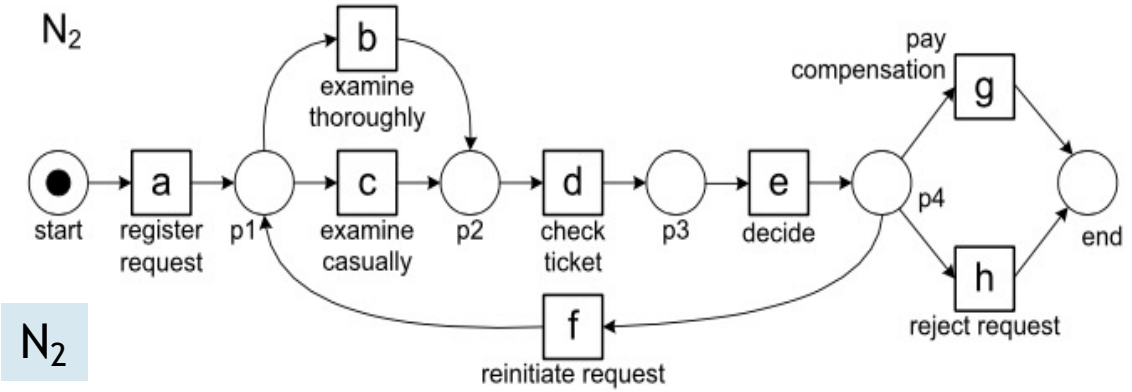
Fitness of N4 is $\frac{1391}{1391} = 1$



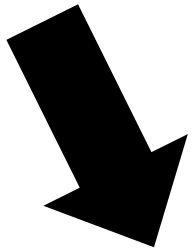
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

L_{full}





N_2



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

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

Identify and count the no. of mismatches

Quantifying the differences

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

(x:y where x is in log and y in N_2)

$$1 - \frac{12}{64} = 0.8125$$

Count the fitness as above.

Token-based Replay

Conformance checking & fitness measure

Token-based Replay

- In the Naïve approach, we stopped replaying a trace once we encounter a problem.
- An alternative could be to compute the fitness at events level rather than complete traces.
- In token-based reply, we continue replaying the trace on the model but record all situations where a transition is forced to fire without being enabled, i.e., we count all missing tokens.
- Additionally, we also record the events that remain at the end.

Token-based Replay

- Following four counters will be maintained at each step:
  - **p** (produced tokens),
  - **c** (consumed tokens),
  - **m** (missing tokens), and
  - **r** (remaining tokens)

Fitness measure in Token-based replay

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

p (produced tokens)
c (consumed tokens)
m (missing tokens)
consumed while not there
r (remaining tokens)
produced but not consumed

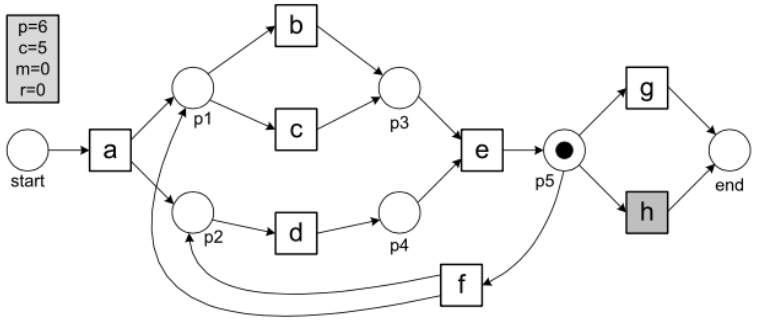
Fraction of missing tokens relative to the number of consumed tokes

Equals to 1 if there are no missing tokens

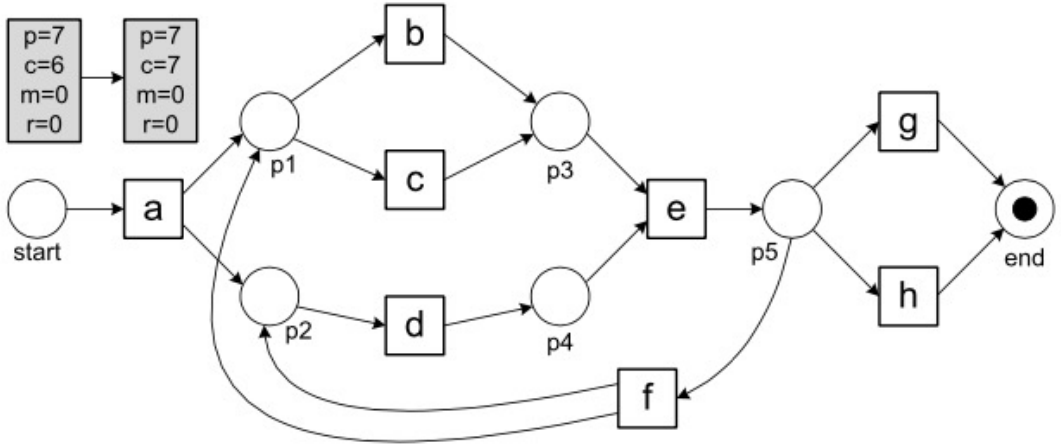
Fraction of remaining tokens relative to the number of produced tokes

Equals to 1 if there are no remaining tokens

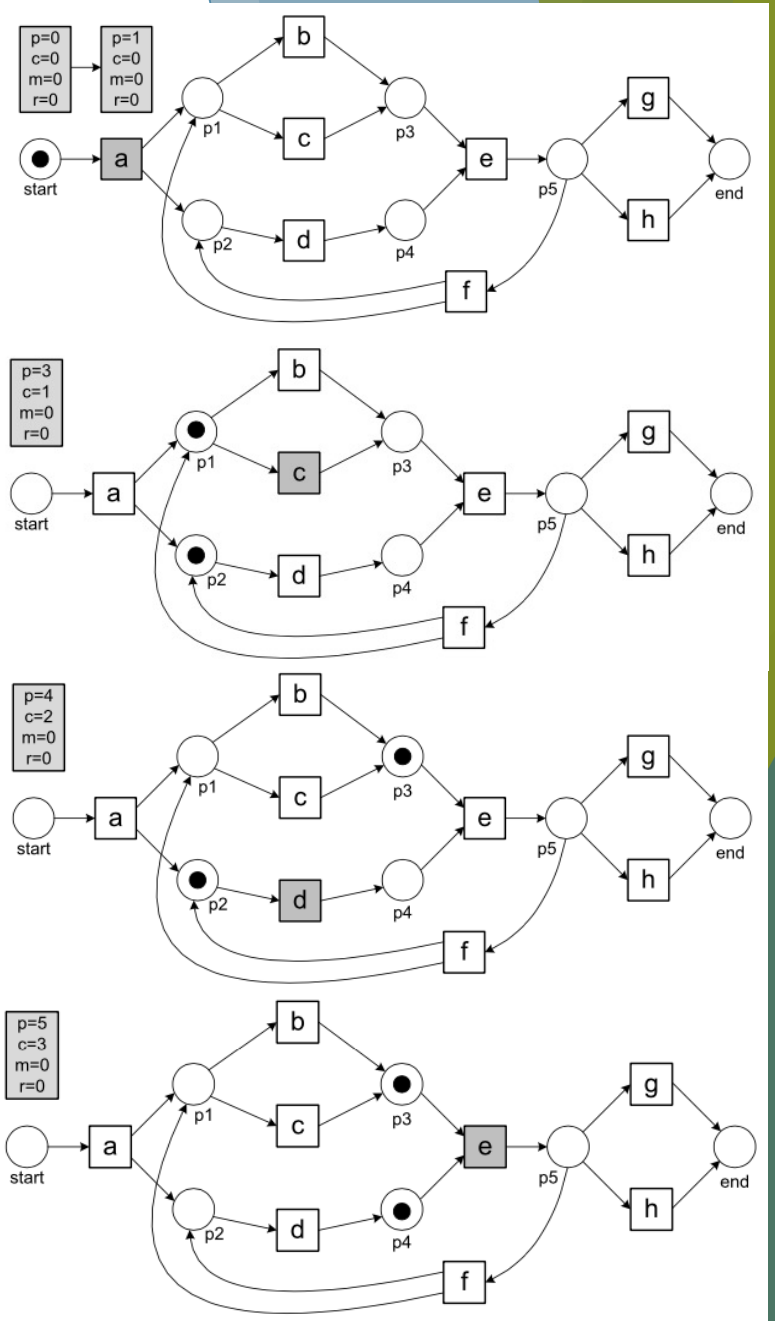
Example 1

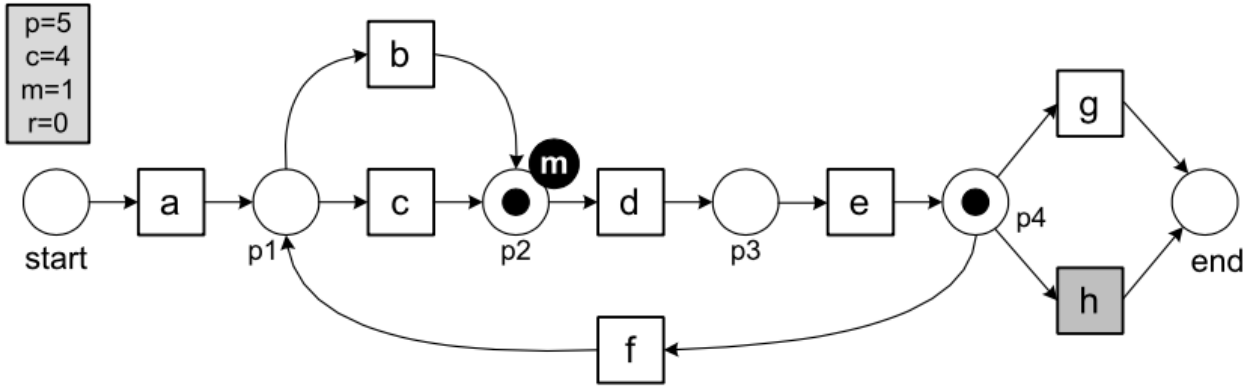


Let the trace $\sigma_1 = \langle a,c,d,e,h \rangle$

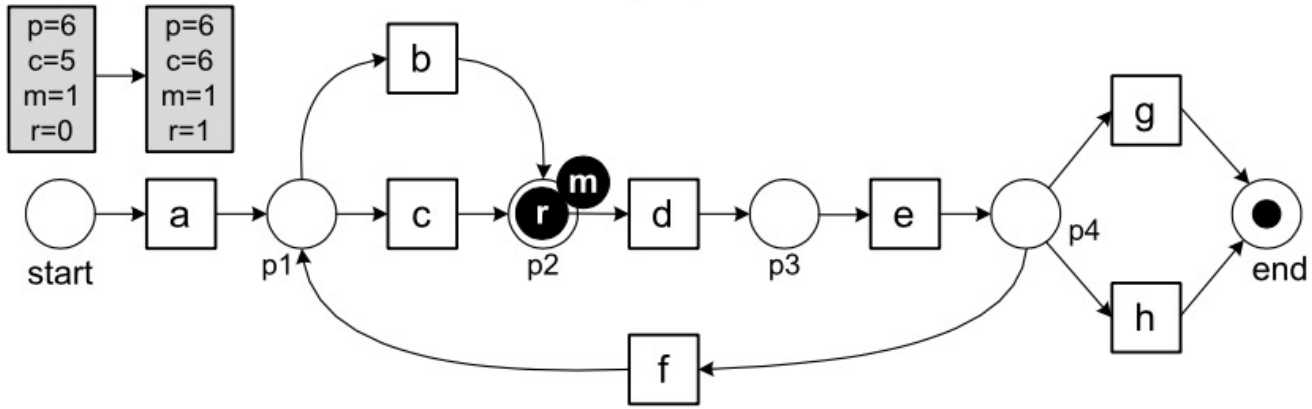


$$fitness(\sigma_1, N_1) = \frac{1}{2}(1 - \frac{0}{7}) + \frac{1}{2}(1 - \frac{0}{7}) = 1$$

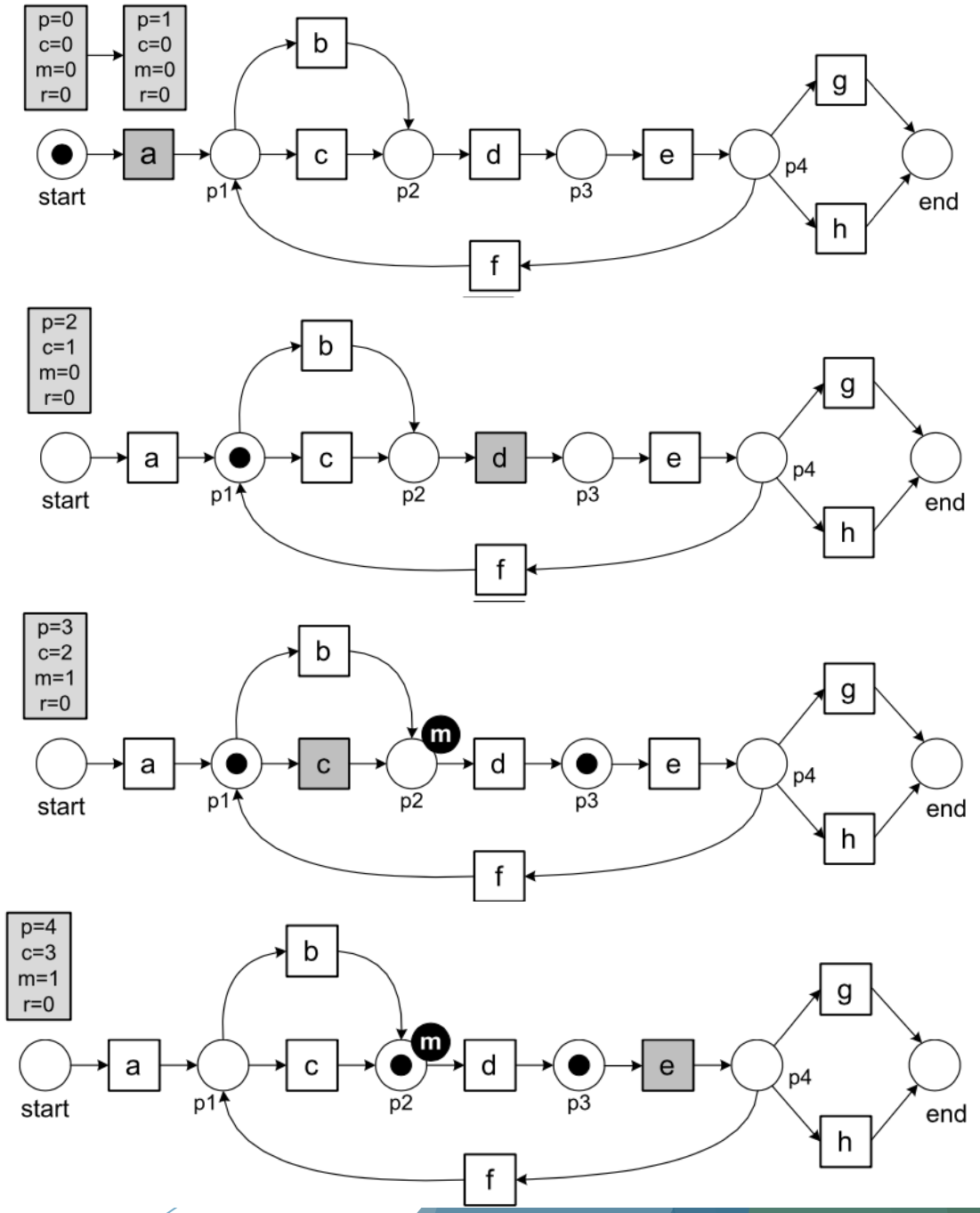




Let the trace $\sigma_3 = \langle a,d,c,e,h \rangle$ and, WF-net N_2:

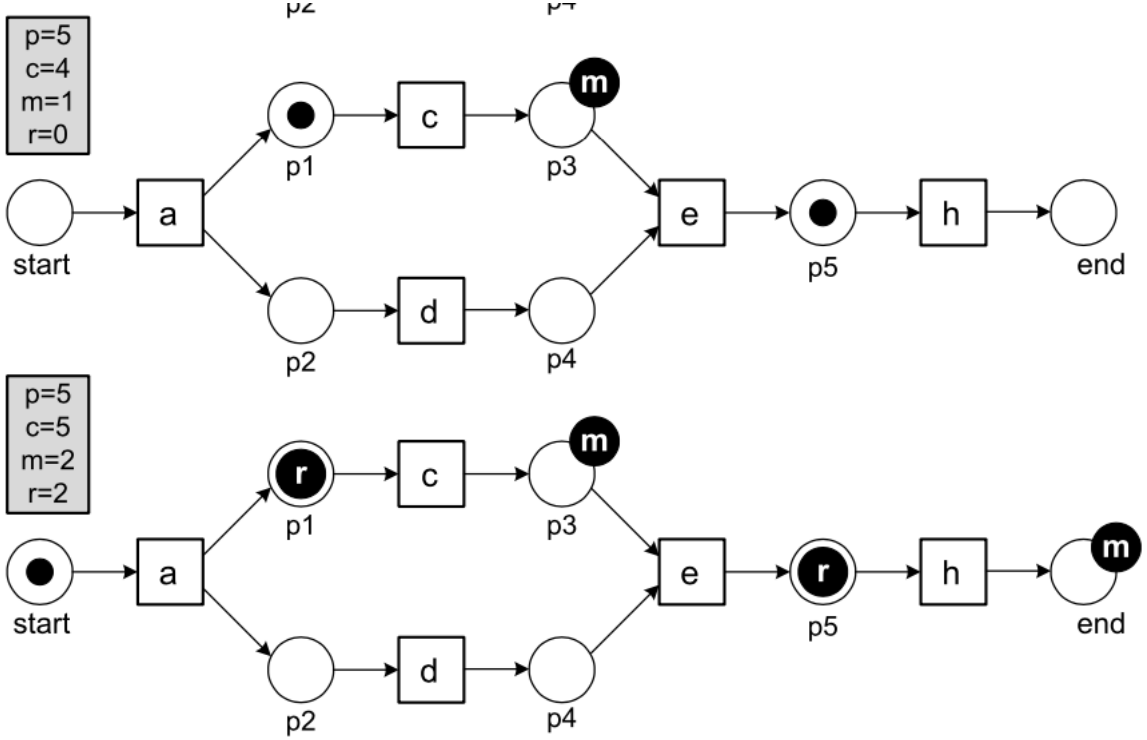


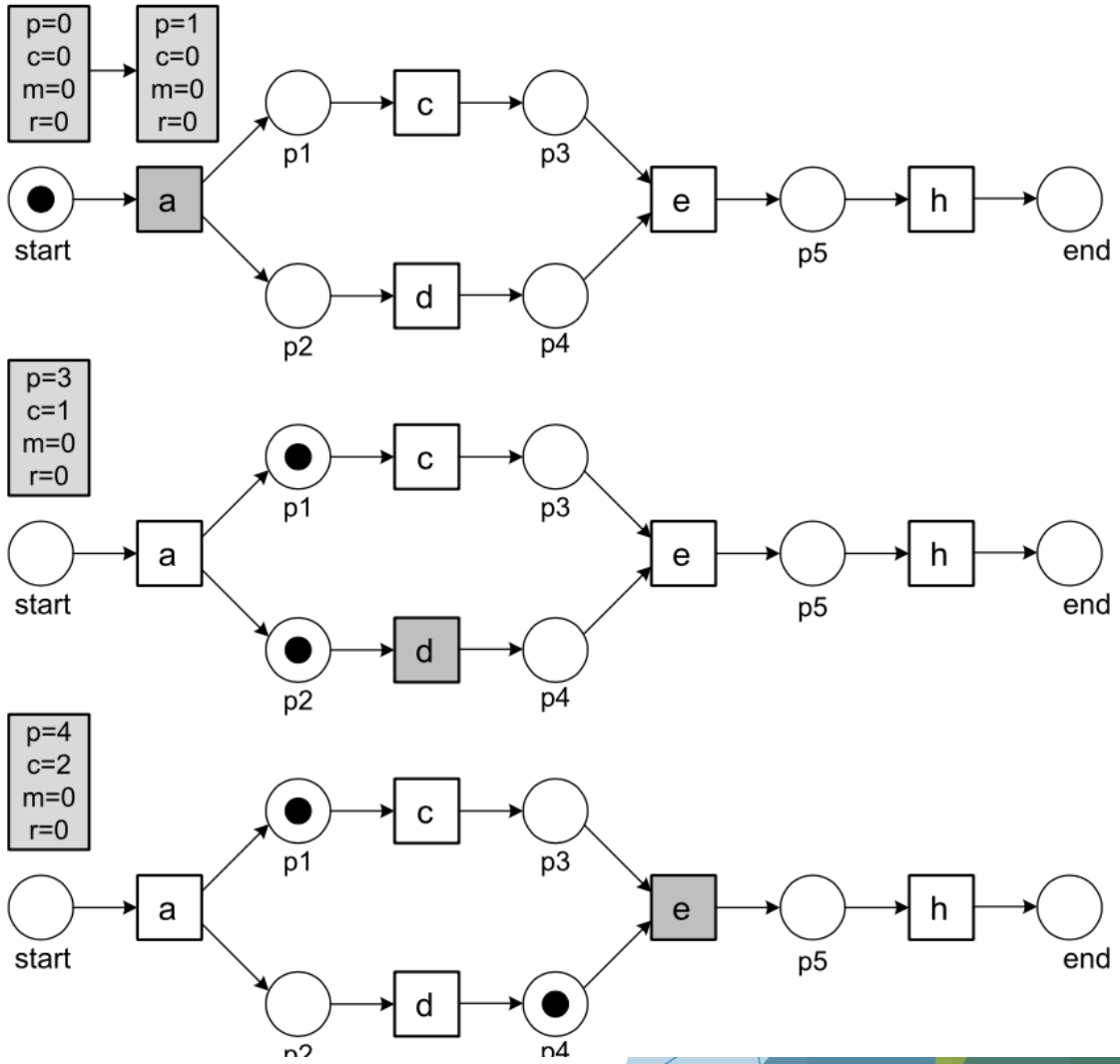
$$fitness(\sigma_3, N_2) = \frac{1}{2}\left(1 - \frac{1}{6}\right) + \frac{1}{2}\left(1 - \frac{1}{6}\right) = 0.8333$$



Example 3: solution

Let the trace $\sigma_2' = \langle a,d,e \rangle$
and, WF-net N_3:





$$fitness(\sigma_2, N_3) = \frac{1}{2}\left(1 - \frac{2}{5}\right) + \frac{1}{2}\left(1 - \frac{2}{5}\right) = 0.6$$

Fitness measure at the Log level

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

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

Fitness measure at the Log level

Missing tokens

Remaining tokens

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

Consumed tokens

Produced tokens

Note that all the measures are summation of the product of trace frequency

Fitness measure at the Log level

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}	$\langle a, c$
14	σ_{11}	$\langle a, c$
11	σ_{12}	$\langle a, c$
9	σ_{13}	$\langle a, d$
8	σ_{14}	$\langle a, d$
5	σ_{15}	$\langle a, d$
3	σ_{16}	$\langle a, c$
2	σ_{17}	$\langle a, d$
2	σ_{18}	$\langle a, d$
1	σ_{19}	$\langle a, d$
1	σ_{20}	$\langle a, d$
1	σ_{21}	$\langle a, d$

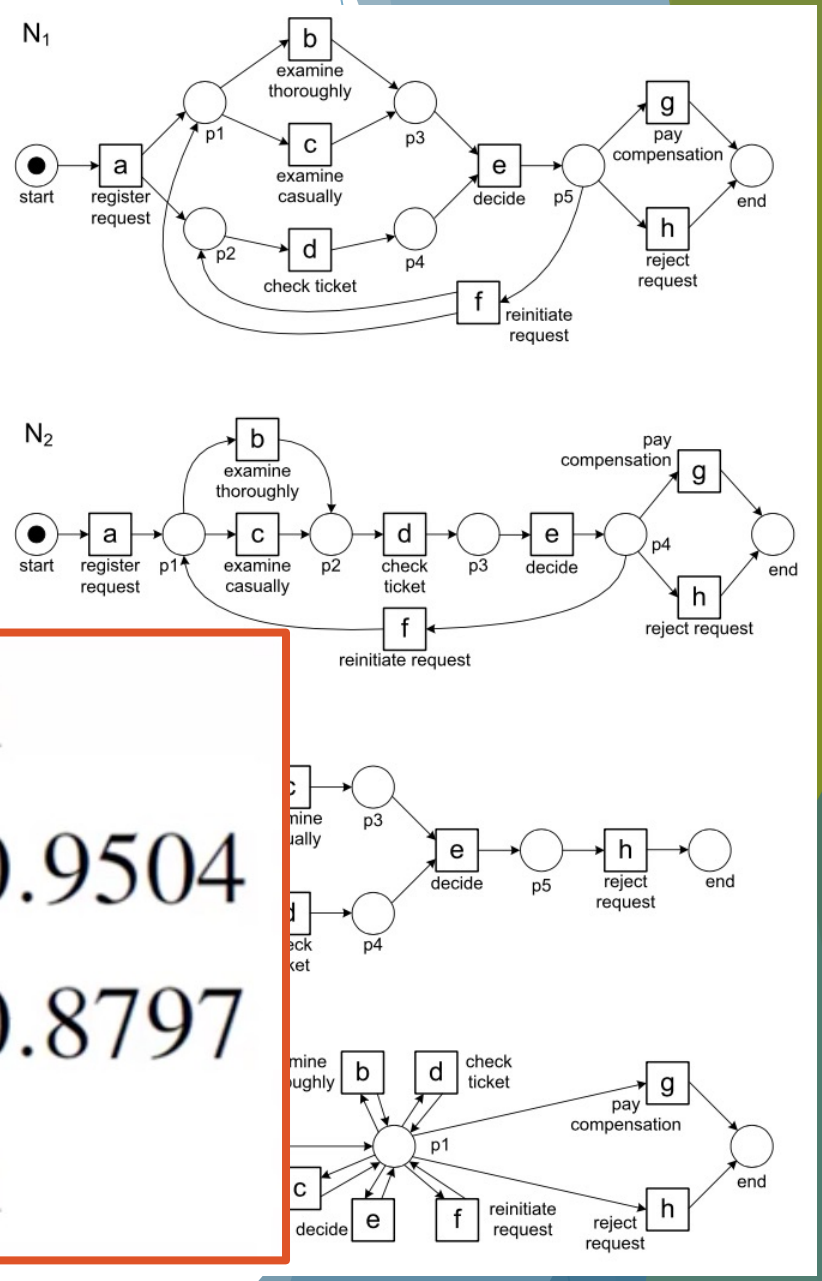
?

$$fitness(L_{full}, N_1) = 1$$

$$fitness(L_{full}, N_2) = 0.9504$$

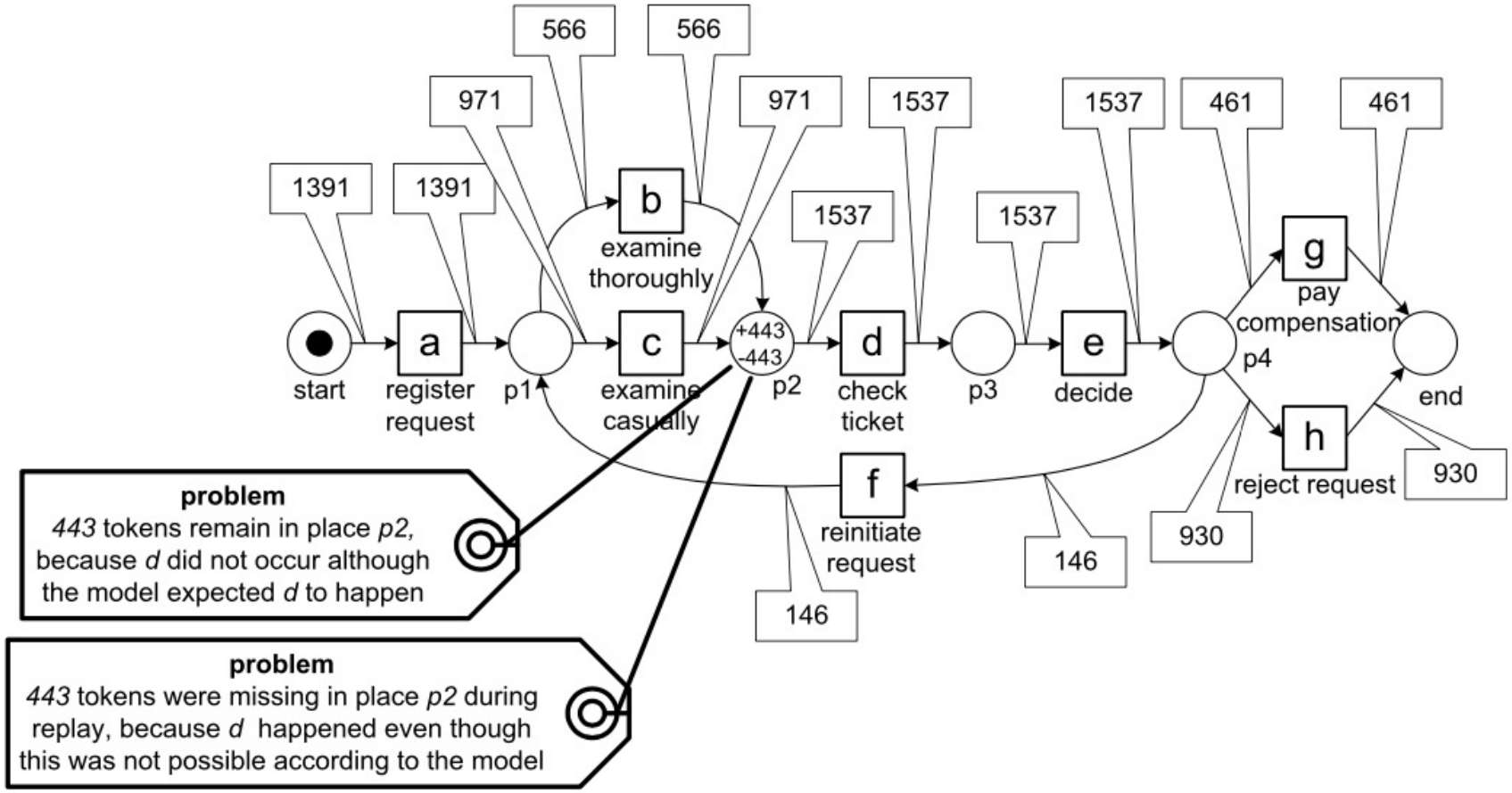
$$fitness(L_{full}, N_3) = 0.8797$$

$$fitness(L_{full}, N_4) = 1$$

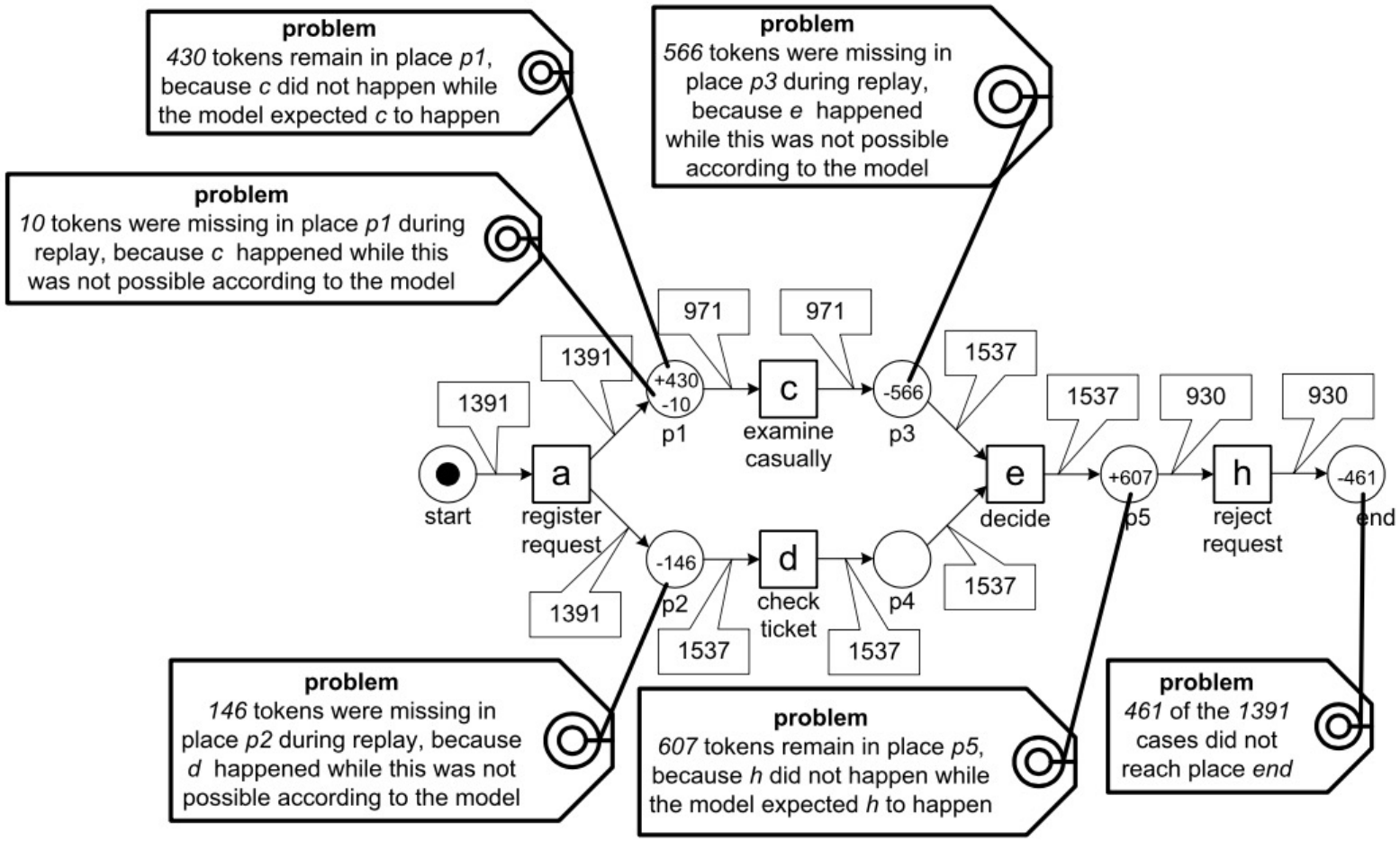


Usefulness of Conformance Checking

Diagnostics: example

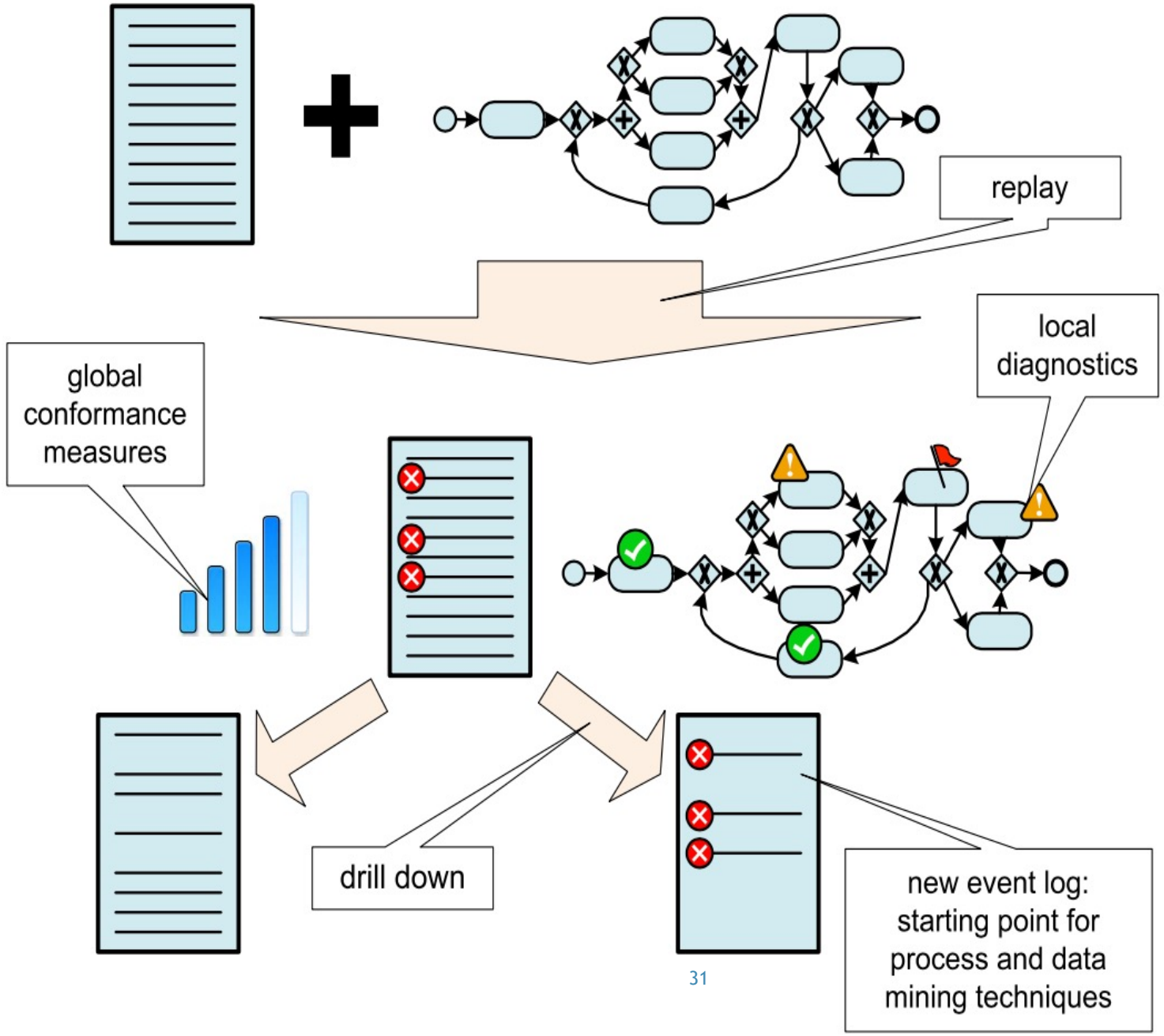


Diagnostics: another example



Conformance Checking

- Further analysis can be performed on fitting and non-fitting cases (for example using machine learning models)



- Token based replay conformance is available in the ProM 5.2
- Prom 6 supports advanced variations of conformance checking (i.e., alignments)

Reading Material

- Chapter 8: Aalst