CS 4072 – Topics in CS
Process Mining

Lecture # 19

May 16, 2022

Spring 2022

FAST – NUCES, CFD Campus

Dr. Rabia Maqsood

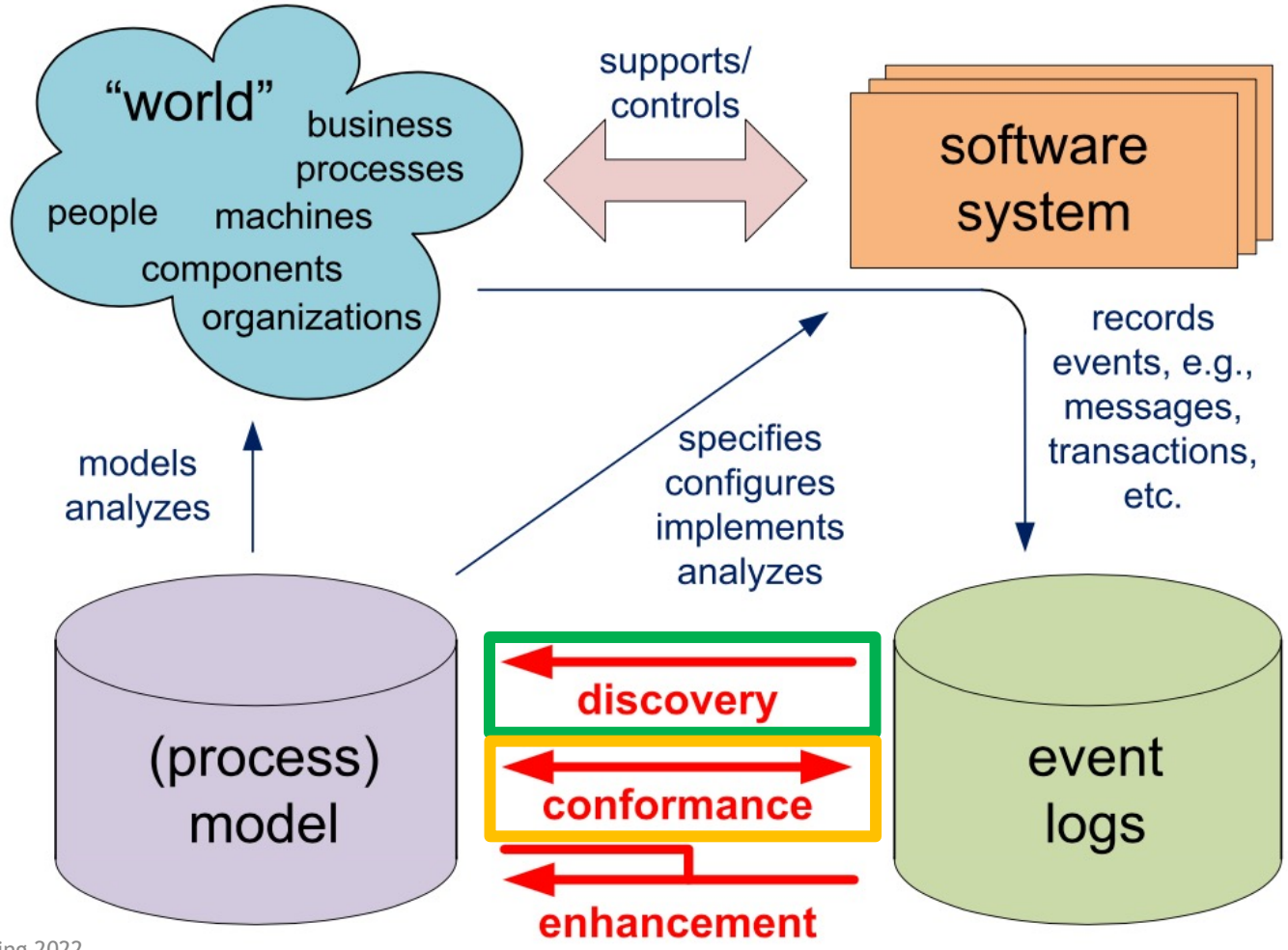
rabia.maqsood@nu.edu.pk

Today's Topics

- Conformance Checking
  - Naïve approach
  - Using causal footprints

- Project discussion

Process Mining Tasks



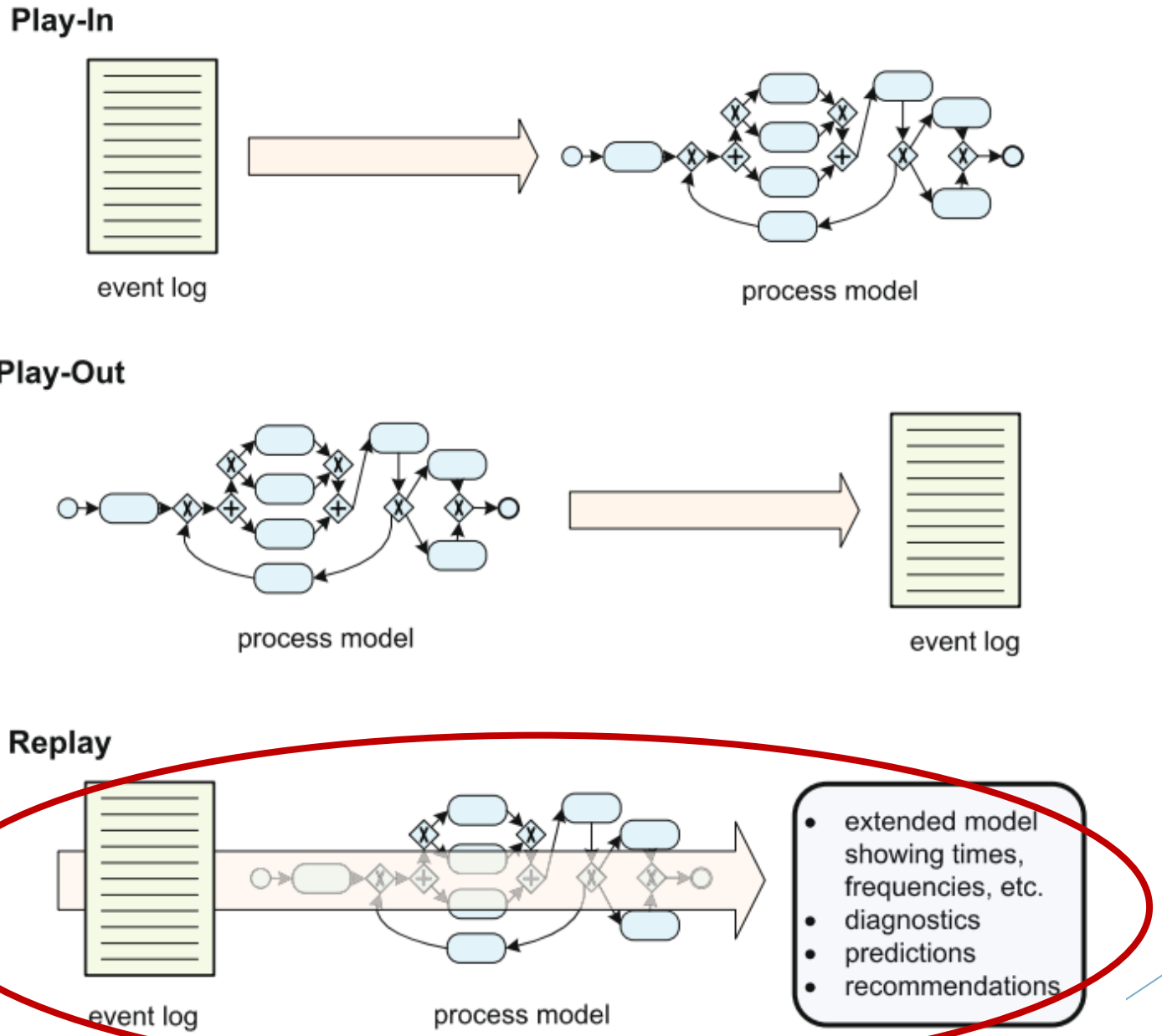


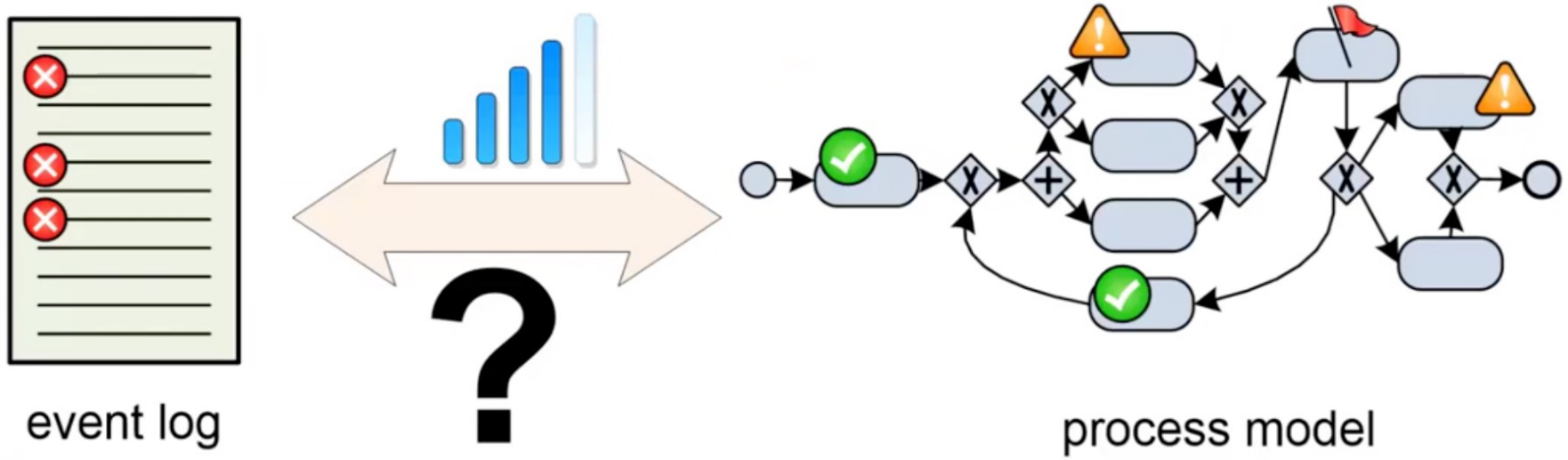
Fig. 1.8 Three ways of relating event logs (or other sources of information containing example behavior) and process models: *Play-in*, *Play-out*, and *Replay*

Conformance Checking: use cases

- Compliance checking (for auditing, fraud detection, etc.)
- Evaluating process discovery results/algorithms
- Conformance to specification (software, services, etc.)

The objective of conformance checking is to find commonalities and discrepancies.

Impact of deviations: Positive or Negative?

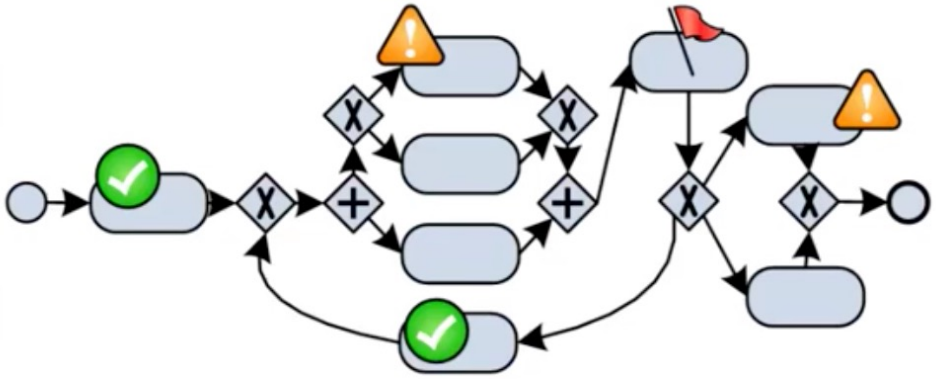


Is the model wrong or something went wrong in the log?

Impact of deviations: Positive or Negative?

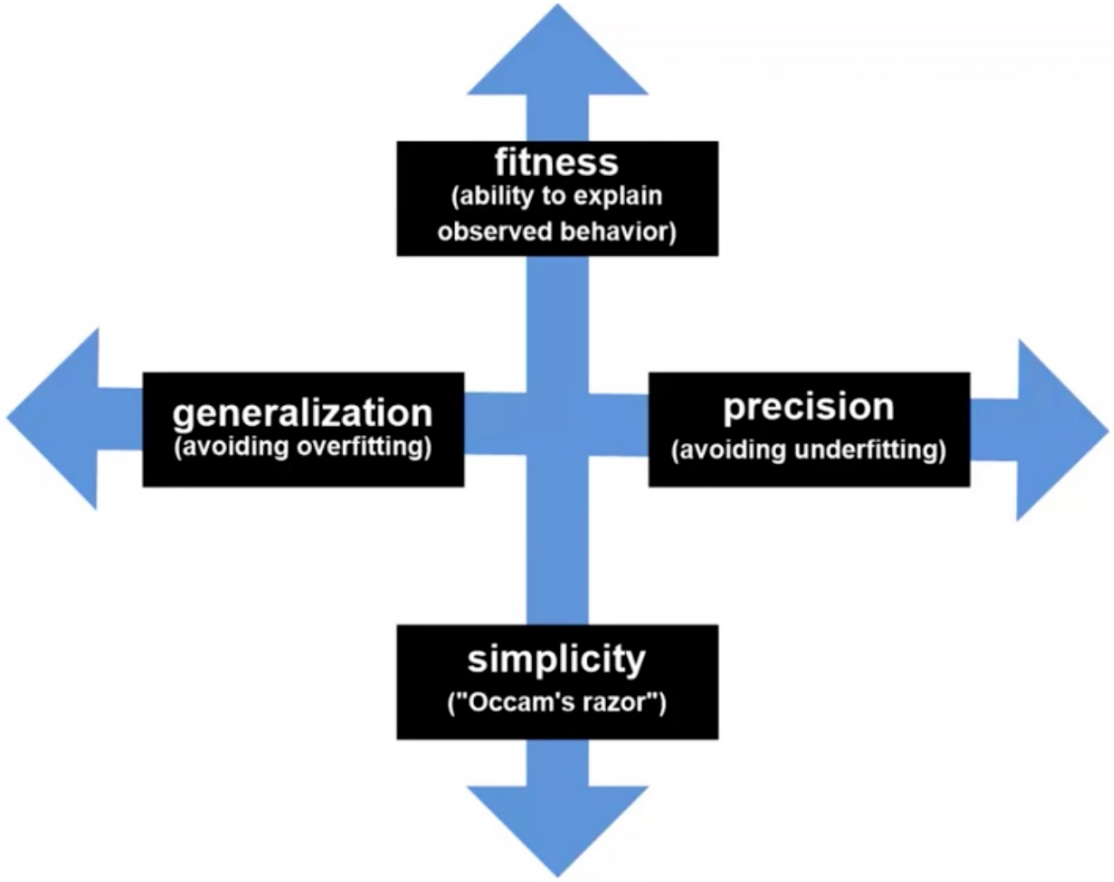
Breaking the glass may saves lives!

event log

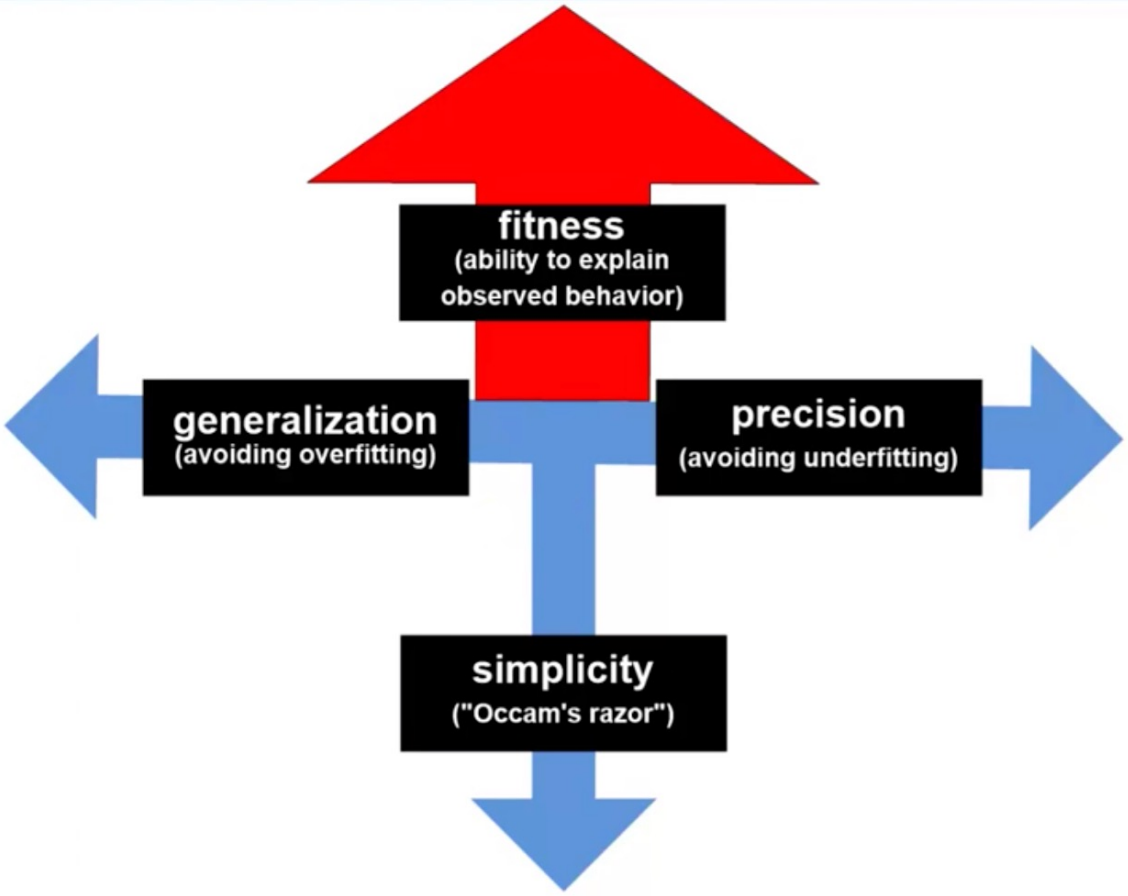


process model

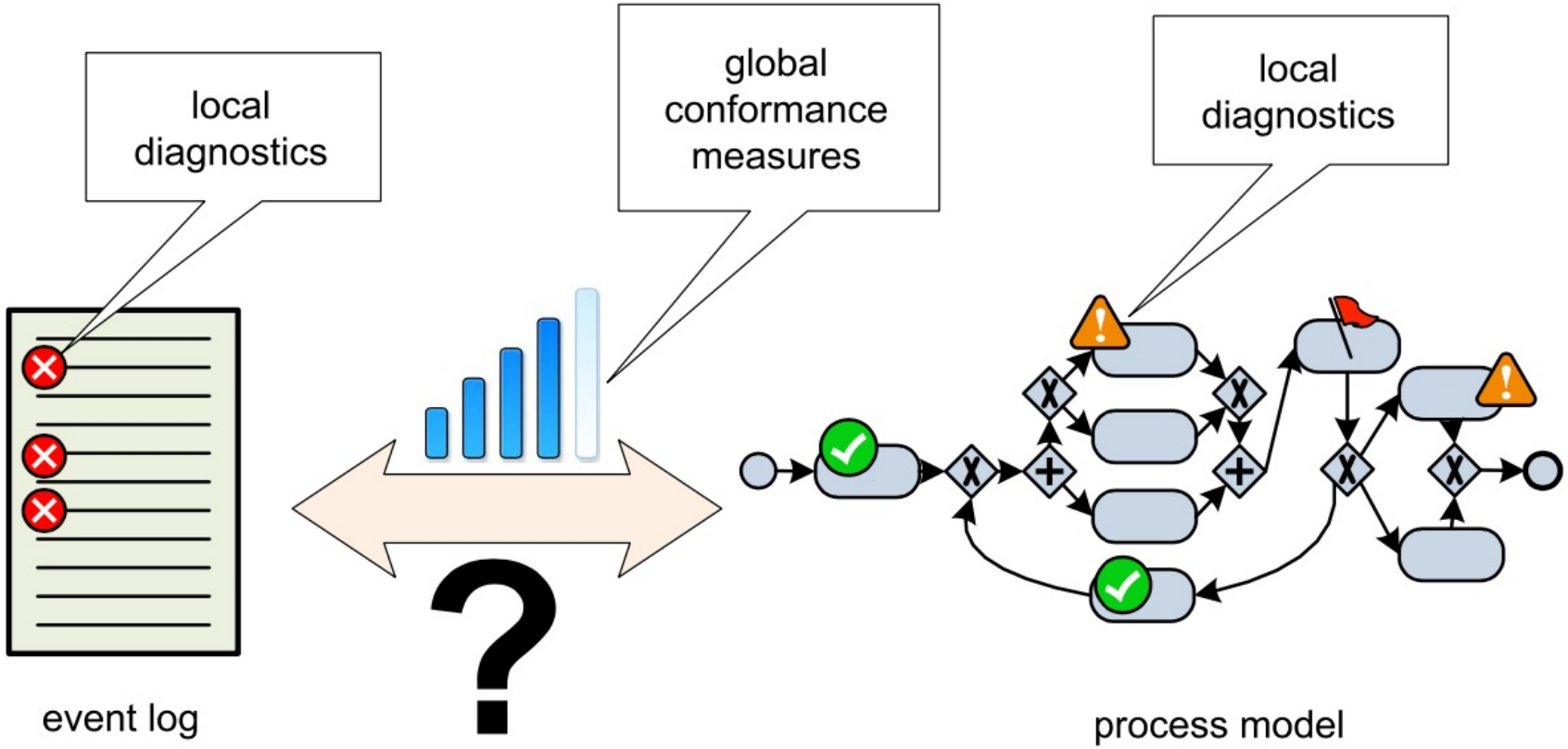
Four dimensions to compare the log and model



Replay fitness is dominant

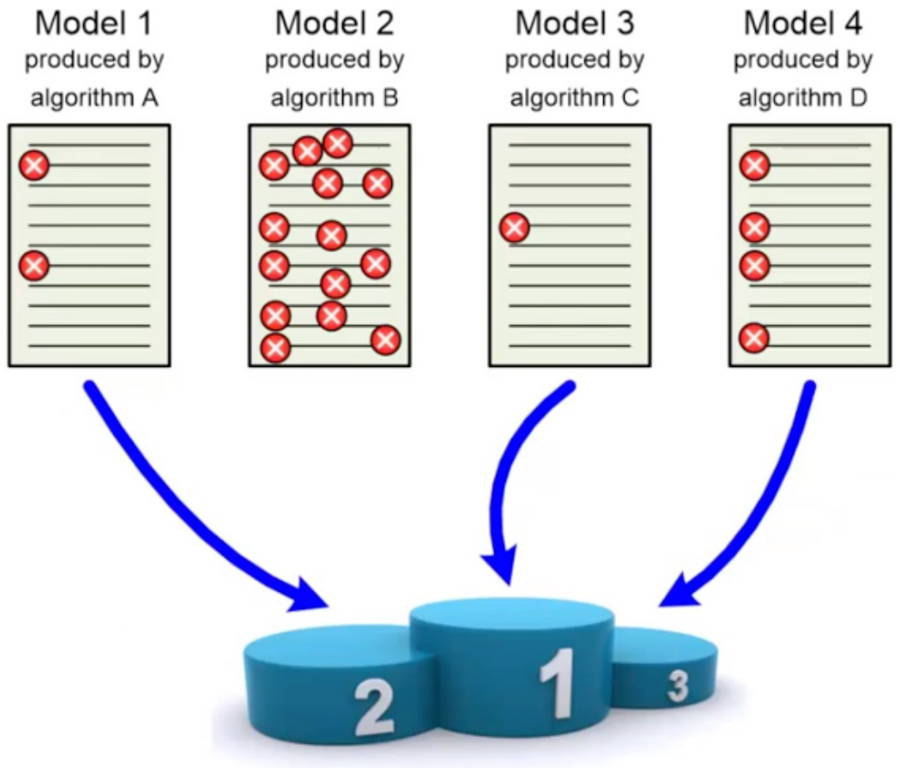


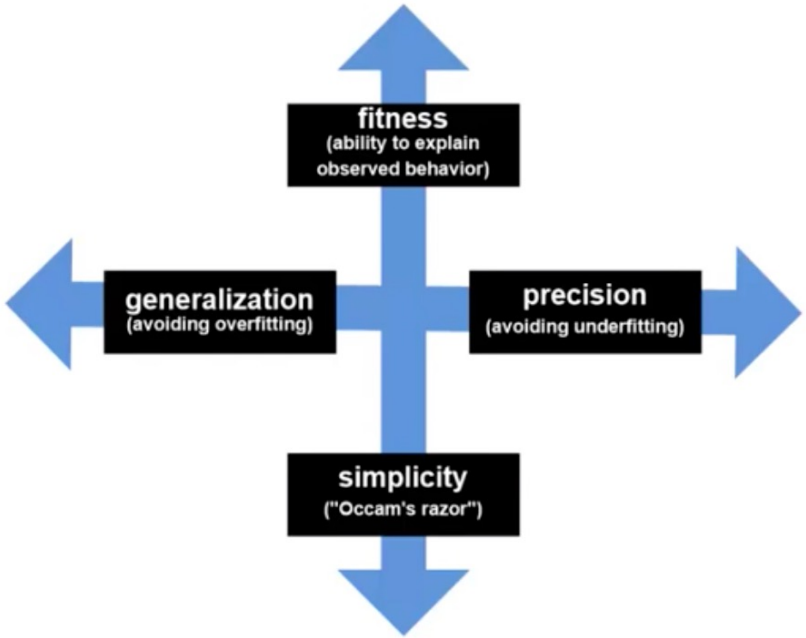
Conformance diagnostics and measures

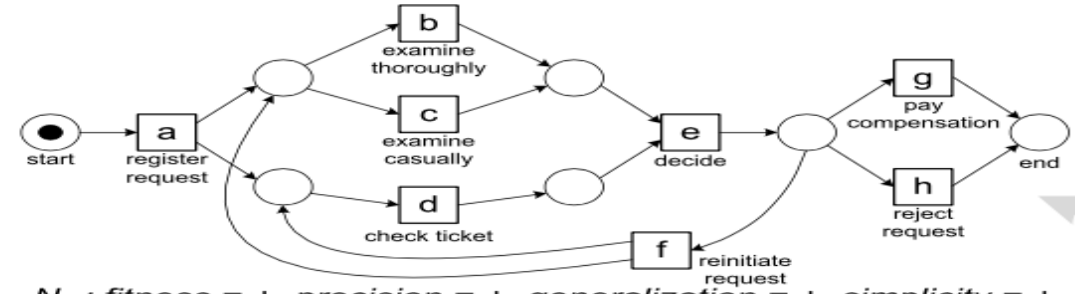


Global conformance measure example: 85% of the cases in the event log can be replayed by the model.
Local diagnostics example: activity x was executed 15 times although this was not allowed according to the model.

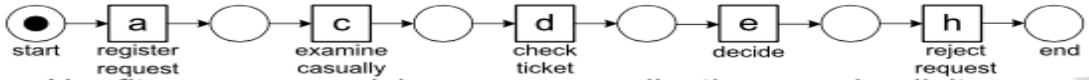
Evaluating Process Discovery Algorithms



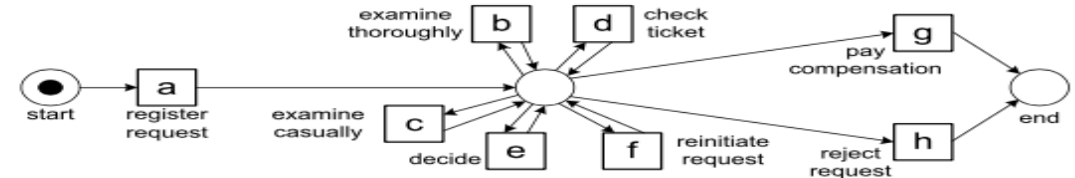




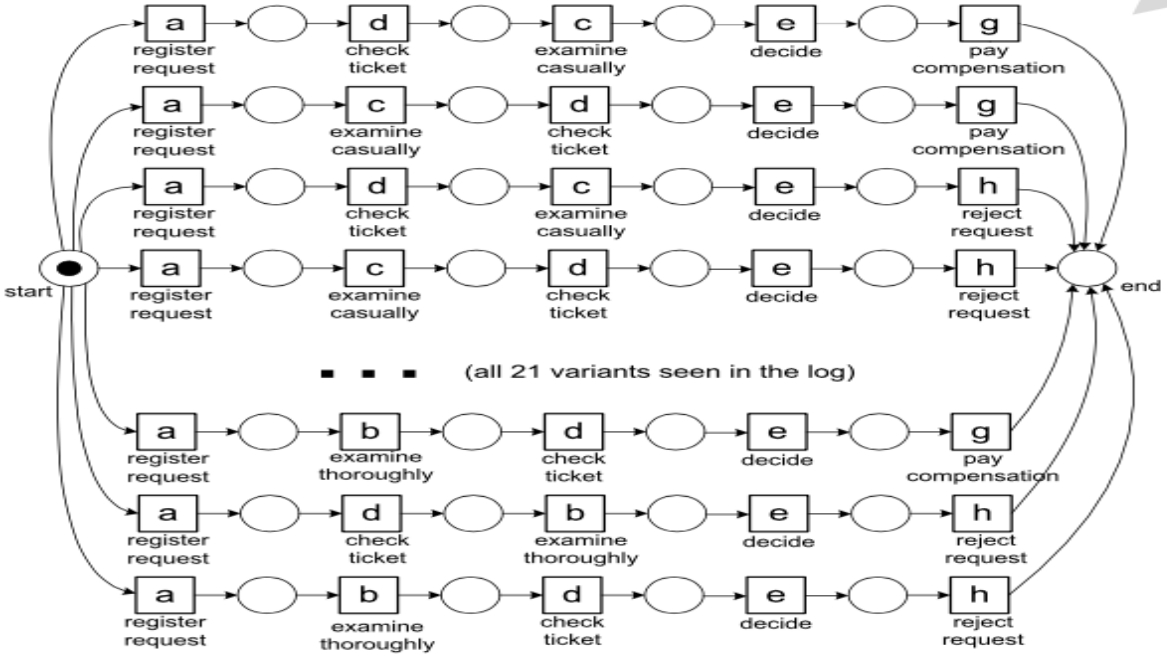
N_1 : *fitness* = +, *precision* = +, *generalization* = +, *simplicity* = +



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



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

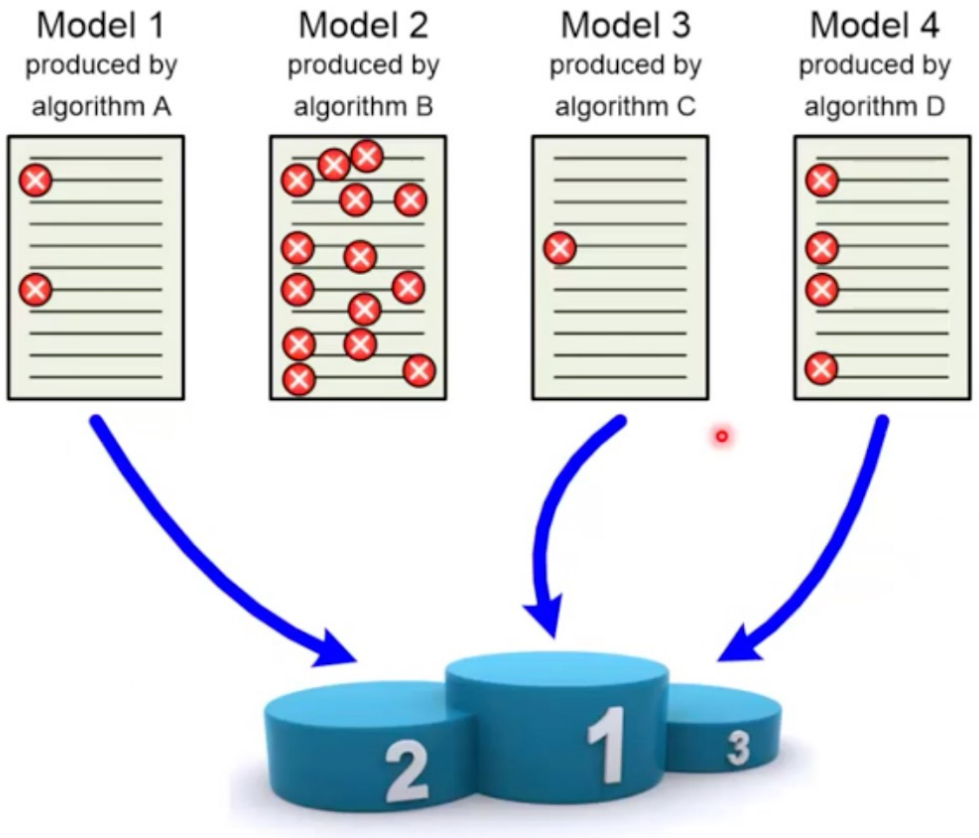


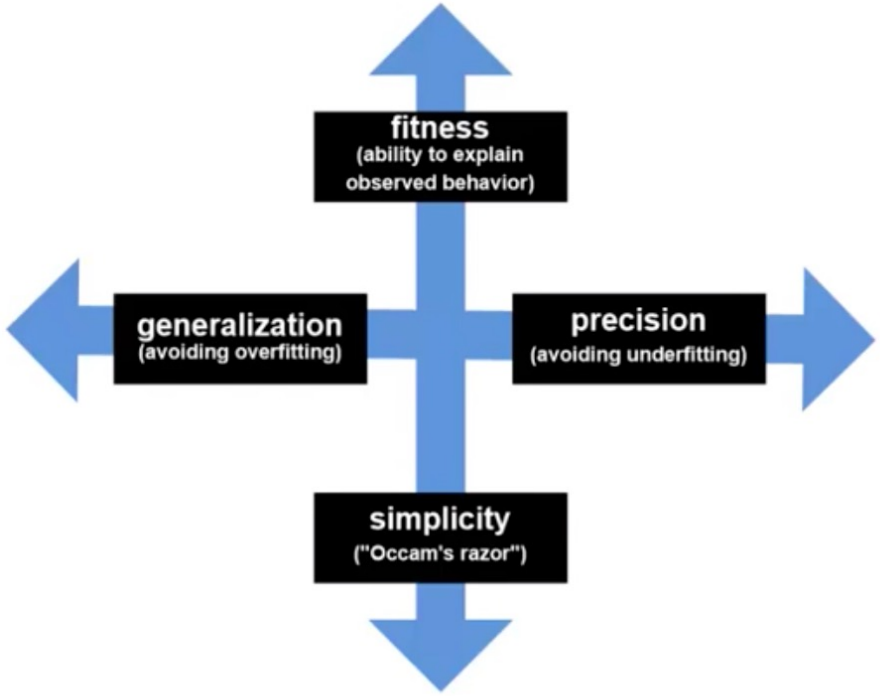
N_4 : *fitness* = +, *precision* = +, *generalization* = -, *simplicity* = -

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

Selection of a difference measurement criterion may vote for a different model.

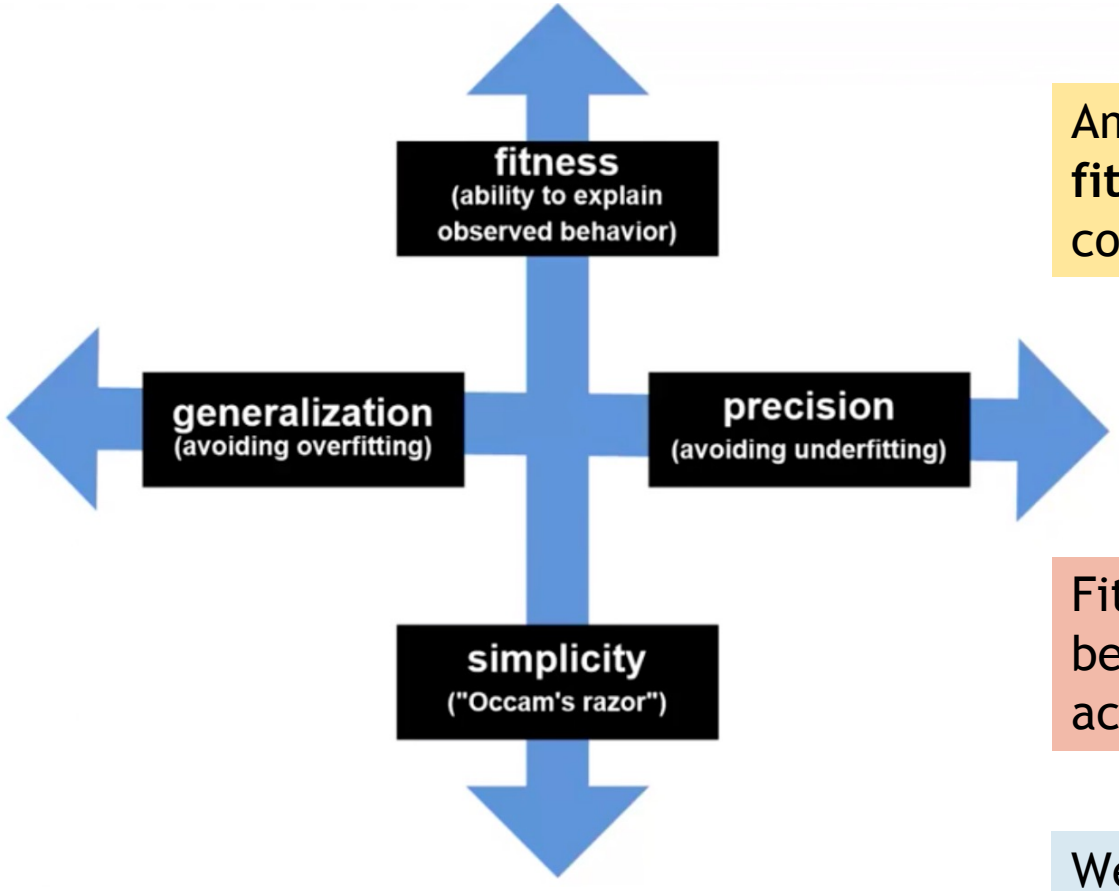
Evaluating Process Discovery Algorithms





There is no such thing as **"the best model"**

Four dimensions to compare the log and model



Among these four quality measures, **fitness** is most related to conformance.

Fitness measures “the proportion of behavior in the event log possible according to the model”.

We will quantitatively define the notion of fitness in a while.

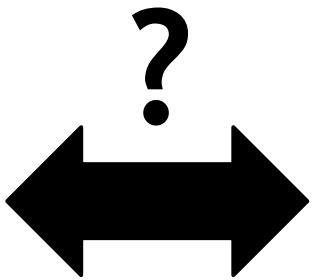
Approaches for Conformance Checking

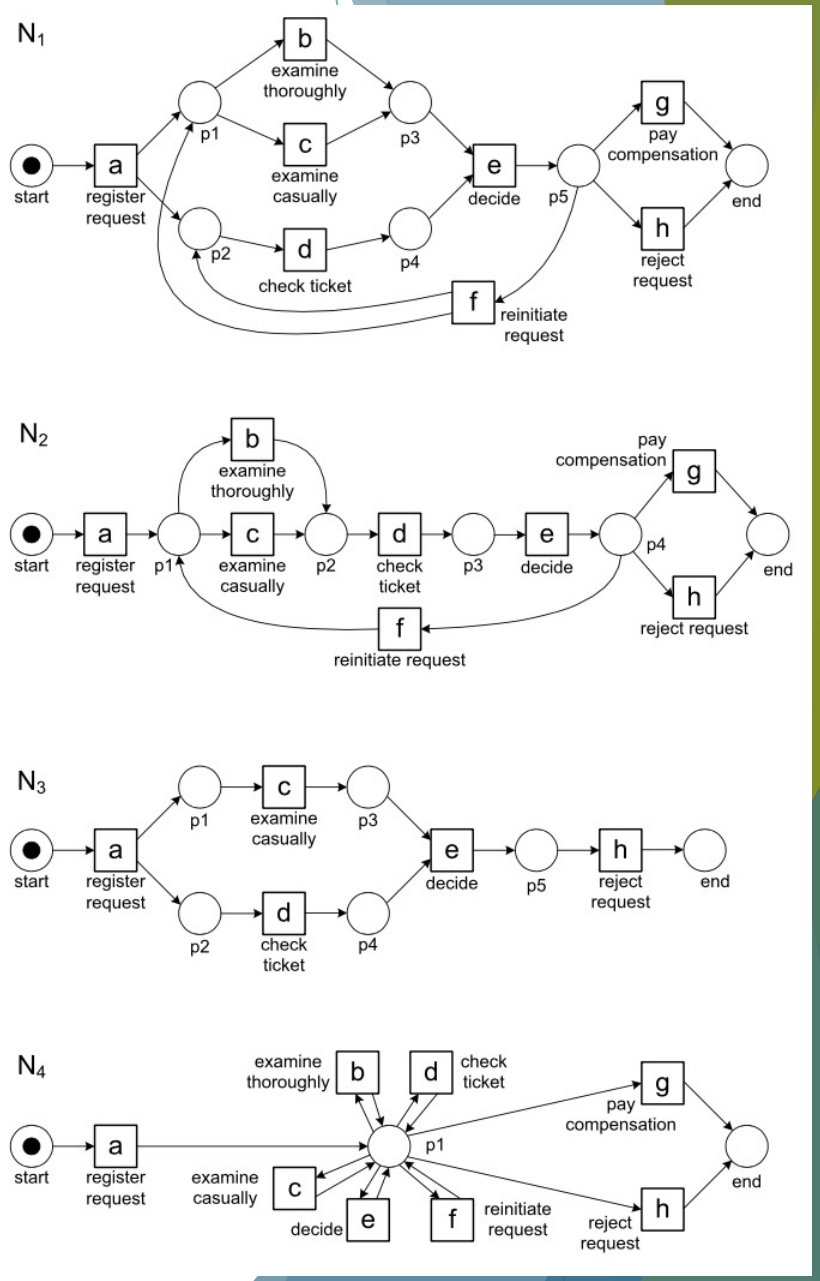
Model and Log Fitness

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





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*]).

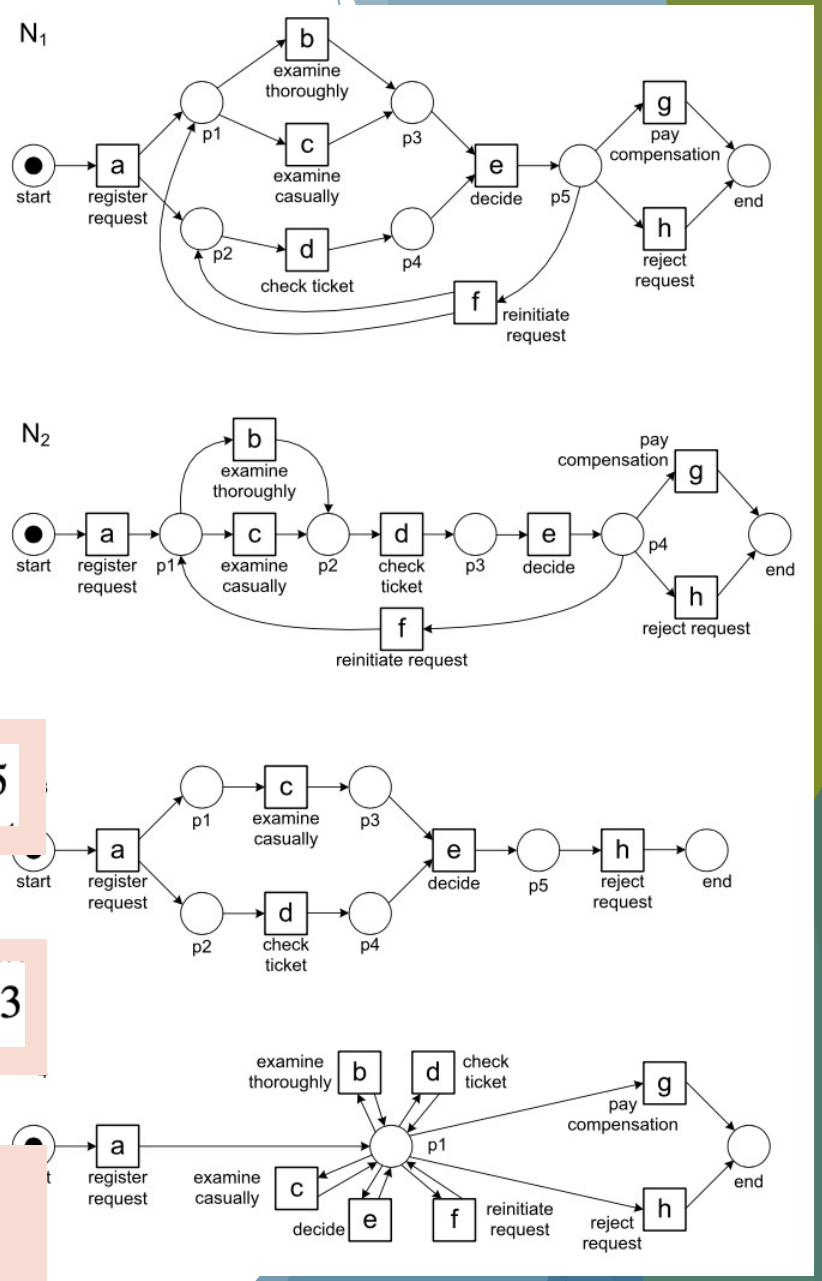
Frequency	Reference	Trace
455	σ_1	$\langle a, c, d, e, h\rangle$
191	σ_2	$\langle a, b, d, e, g\rangle$
177	σ_3	$\langle a, d, c, e, h\rangle$
144	σ_4	$\langle a, b, d, e, h\rangle$
111	σ_5	$\langle a, c, d, e, g\rangle$
82	σ_6	$\langle a, d, c, e, g\rangle$
56	σ_7	$\langle a, d, b, e, h\rangle$
47	σ_8	$\langle a, c, d, e, f, d, b, e, h\rangle$
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}	$\langle a, d, c, e, f, d, b, e, f, c, d, e, f, d, b,$

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

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

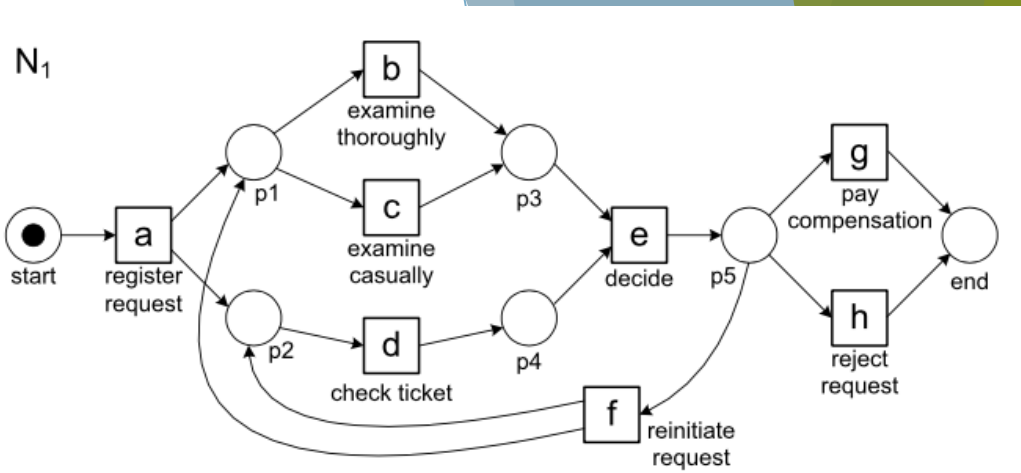
Fitness of N3 is $\frac{632}{1391} = 0.4543$

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



Causal footprint approach

#	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	



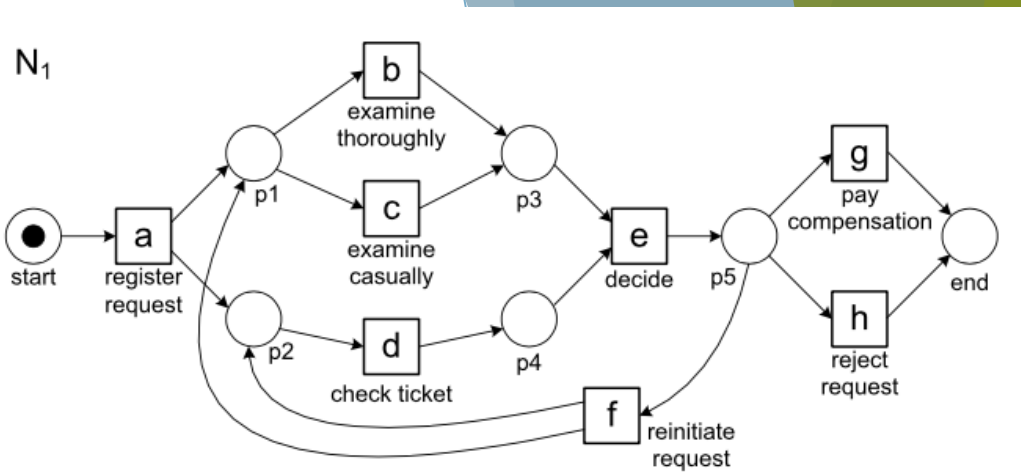
L_{full}

N_1

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

Both will yield the same footprint matrix

#	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_1

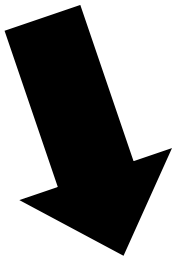
	a	b	c	d	e	f	g	h
a	#	→	→	→	#	#	#	#
b	←	#	#	‖	→	←	#	#
c	←							#
d	←							#
e	#							→
f	#							#
g	#	#	#	#	←	#	#	#
h	#	#	#	#	←	#	#	#

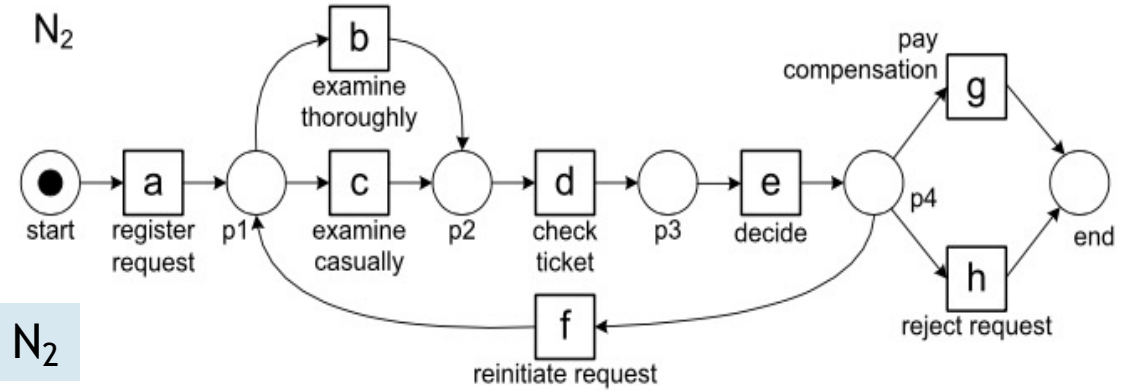
Footprint-based conformance = 1
(1 = perfect match)
(0 = worst match possible)

footprints of log and model coindice

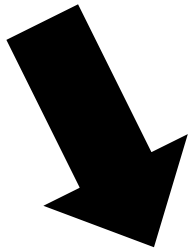
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

L_{full}





N_2



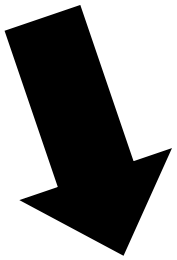
	a	b	c	d	e	f	g	h
a	#	→	→	→	#	#	#	#
b	←	#	#	‖	→	←	#	#
c	←	#	#	‖	→	←	#	#
d	←	‖	‖	#	→	←	#	#
e	#	←	←	←	#	→	→	→
f	#	→	→	→	←	#	#	#
g	#	#	#	#	←	#	#	#
h	#	#	#	#	←	#	#	#

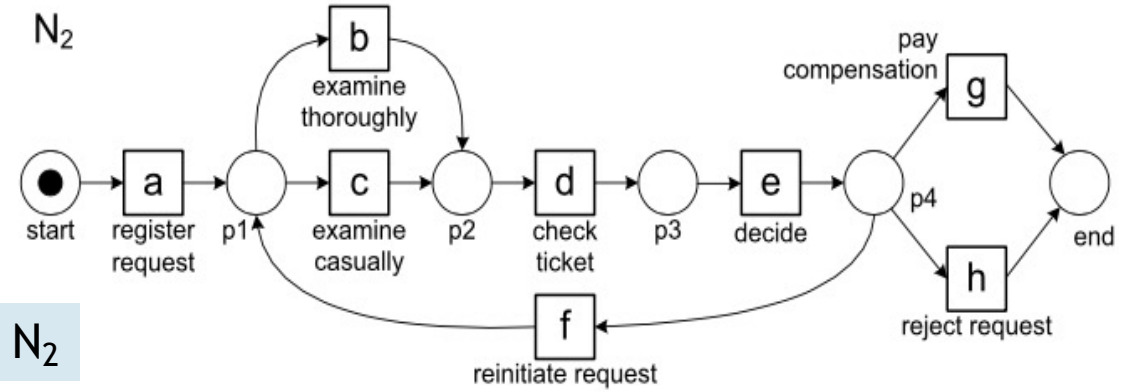
	a	b	c	d	e	f	g	h
a	#	→	→	#	#	#	#	#
b	←	#	#	→	#	←	#	#
c	←	#	#	→	#	←	#	#
d	#	←	←	#	→	#	#	#
e	#	#	#	←	#	→	→	→
f	#	→	→	#	←	#	#	#
g	#	#	#	#	←	#	#	#
h	#	#	#	#	←	#	#	#

Both produces different footprint matrices

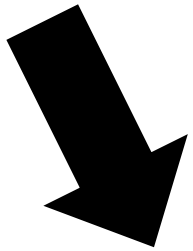
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

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

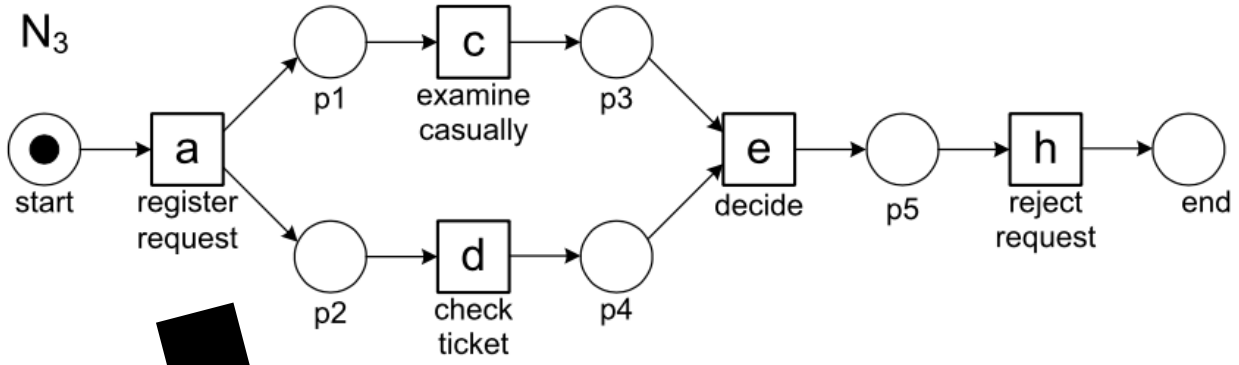
	a	b	c	d	e	f	g	h
a				$\rightarrow: \#$				
b				$\| : \rightarrow$	$\rightarrow: \#$			
c				$\| : \rightarrow$	$\rightarrow: \#$			
d	$\leftarrow: \#$	$\| : \leftarrow$	$\| : \leftarrow$			$\leftarrow: \#$		
e		$\leftarrow: \#$	$\leftarrow: \#$					
f				$\rightarrow: \#$				
g								
h								

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

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

Count the fitness as above.

Solution



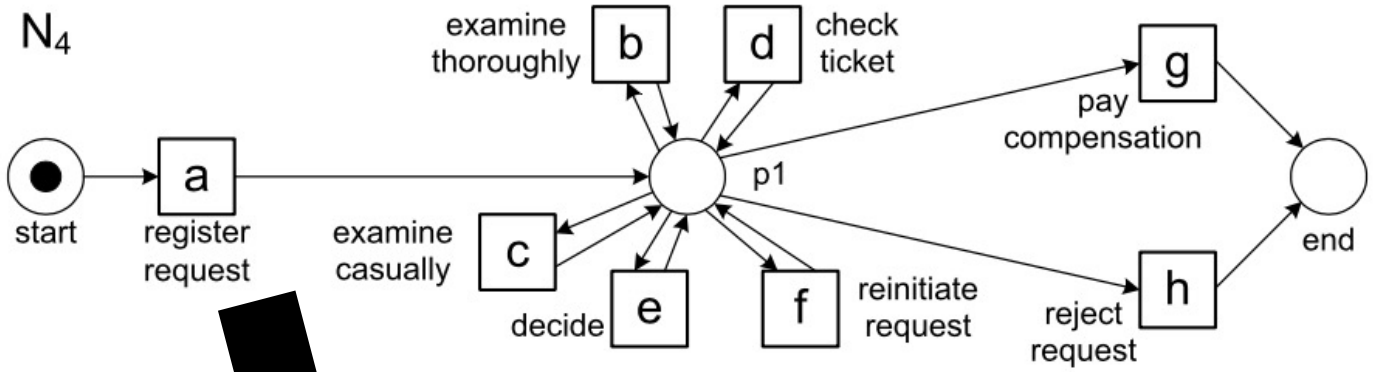
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

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

Color
- Log
- Model

$$1 - \frac{16}{64} = 0.75$$

Solution



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

#	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	

Color
- Log
- Model

$$1 - \frac{45}{64} = 0.296875$$

Project

2 weeks time (**Sunday, May 29, 2022**)

3 persons team

Marks: 10%

Teams' names should be placed on the Google Classroom today by 21.00

Reading Material

- Chapter 8: Aalst