

Conformance Checking (1/2)

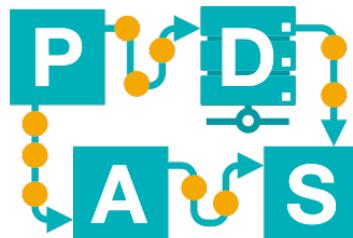
Lecture 12

prof.dr.ir. Wil van der Aalst

www.vdaalst.com @wvdaalst

www.pads.rwth-aachen.de

BPI-L12

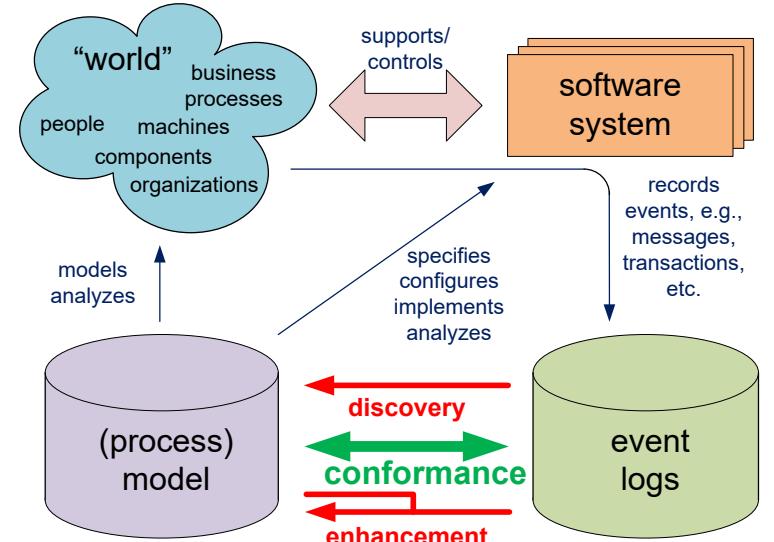


Chair of Process
and Data Science

RWTHAACHEN
UNIVERSITY

Conformance checking approaches

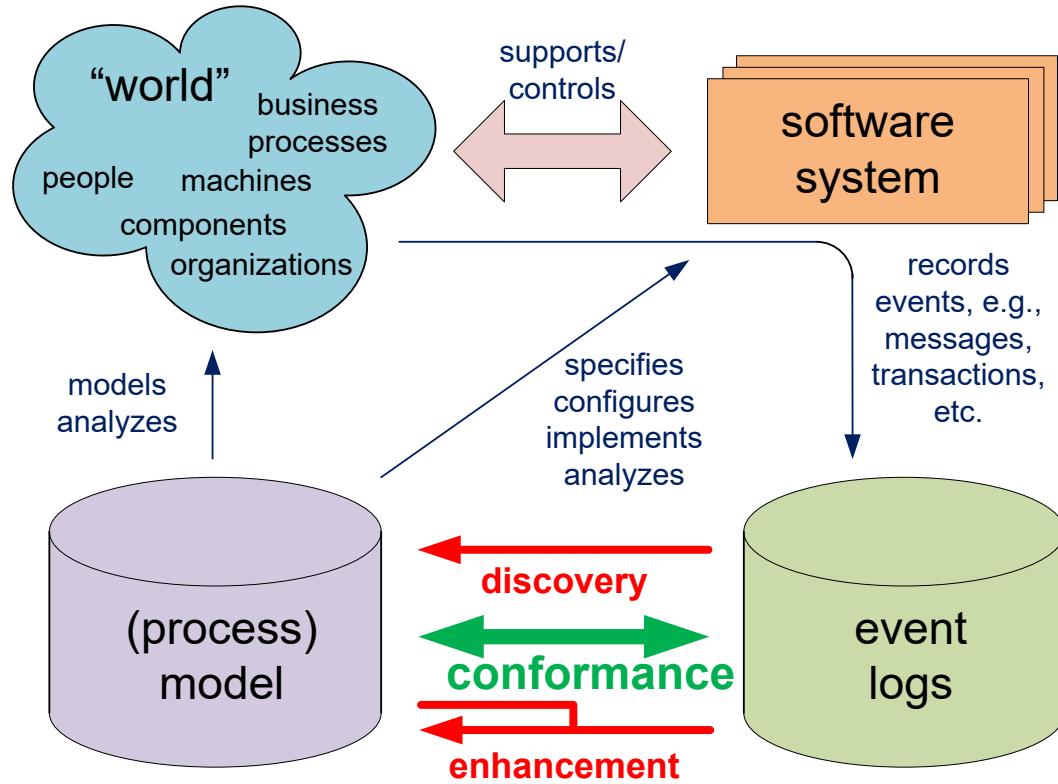
1. Conformance checking using **causal footprints**.
2. Conformance checking based on **token-based replay**.
3. **Alignment-based conformance checking.**



Conformance Checking



Conformance checking



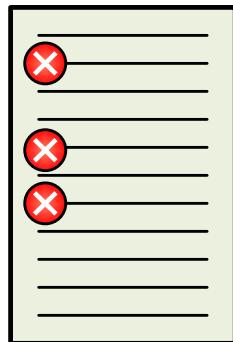
Three main use cases

- **Compliance checking (for auditing, fraud detection, etc.)**
 - Audits are performed to ascertain the validity and reliability of information about organizations and their associated processes.
 - This is done to check whether business processes are executed within certain boundaries set by managers, governments, and other stakeholders.
- **Evaluating process discovery results / algorithms**
 - Comparing discovered process models with the data used to learn the model or with unseen test data.
 - Evaluating a model or an algorithm (see k-fold cross validation).
- **Conformance to specification (software, services, etc.).**



Positive or negative deviants?

“Breaking the glass” may save lives!



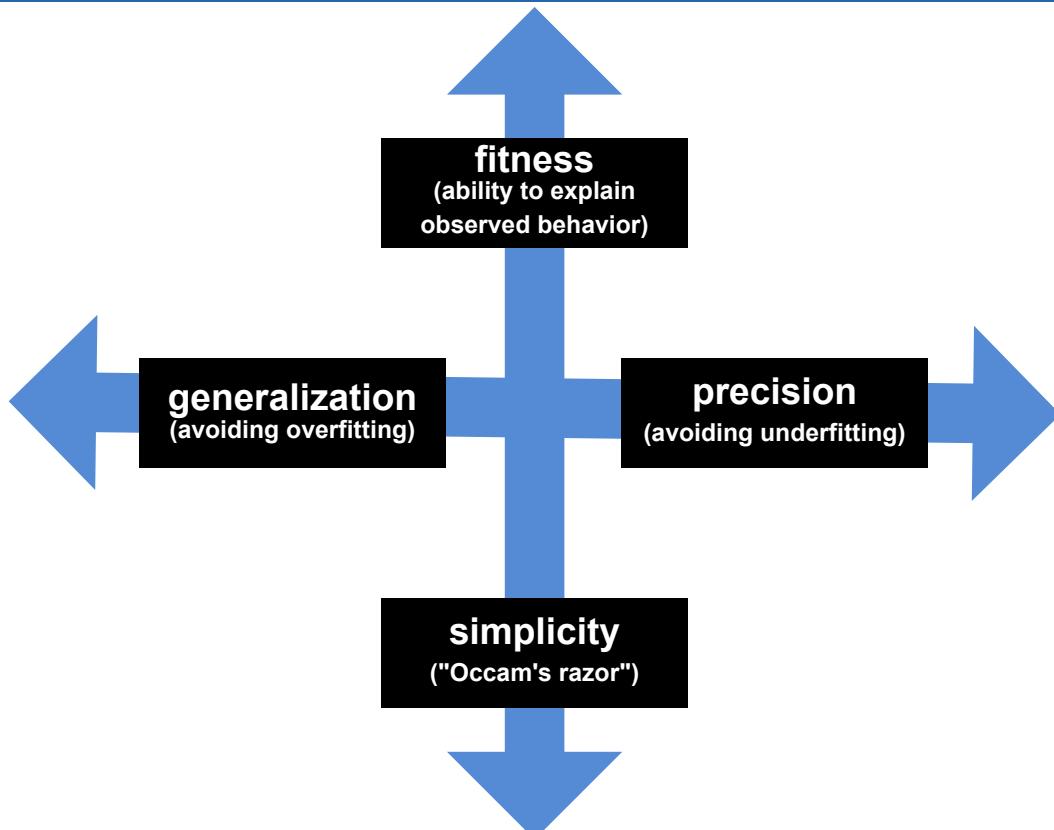
event log

Is the model wrong or is t

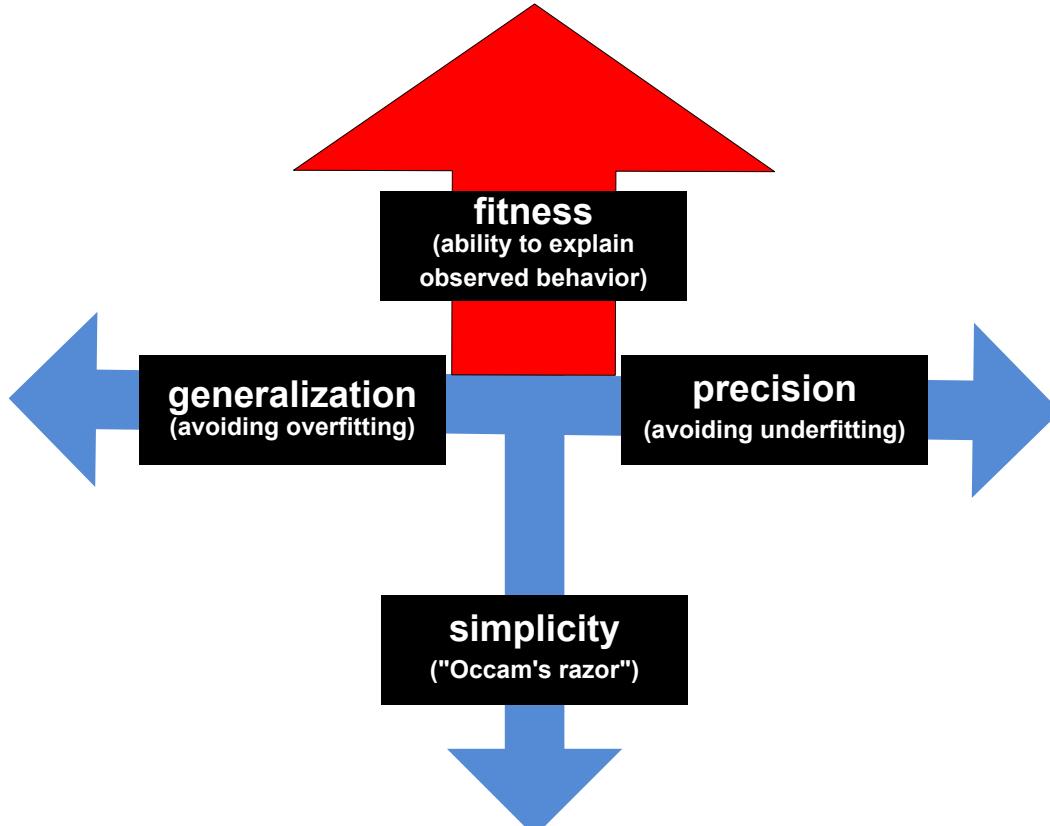


learning from
positive deviants

Four dimensions to compare log and model



Replay fitness is dominant



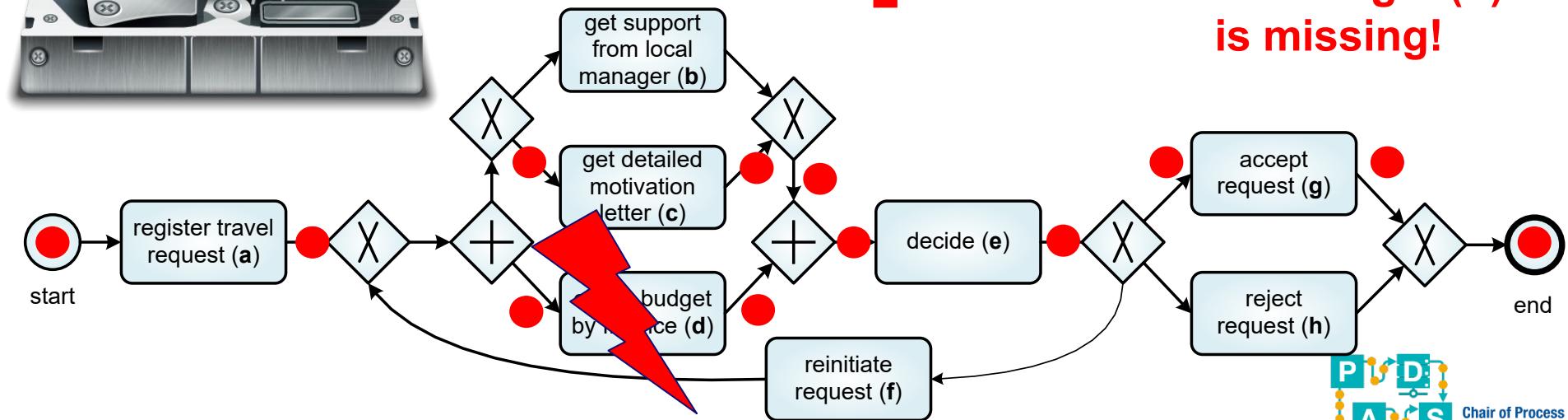
Replay example used before



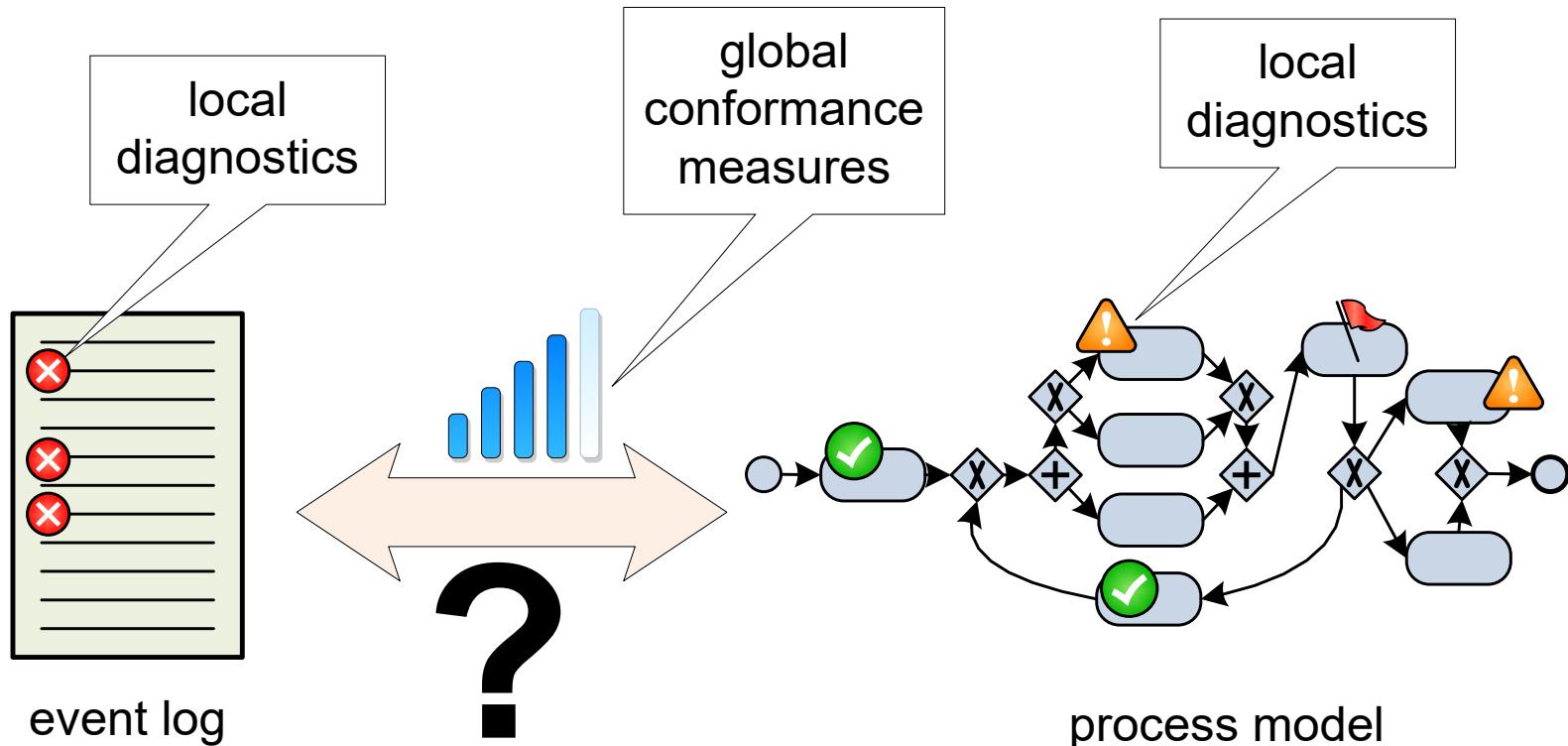
a c e g
?



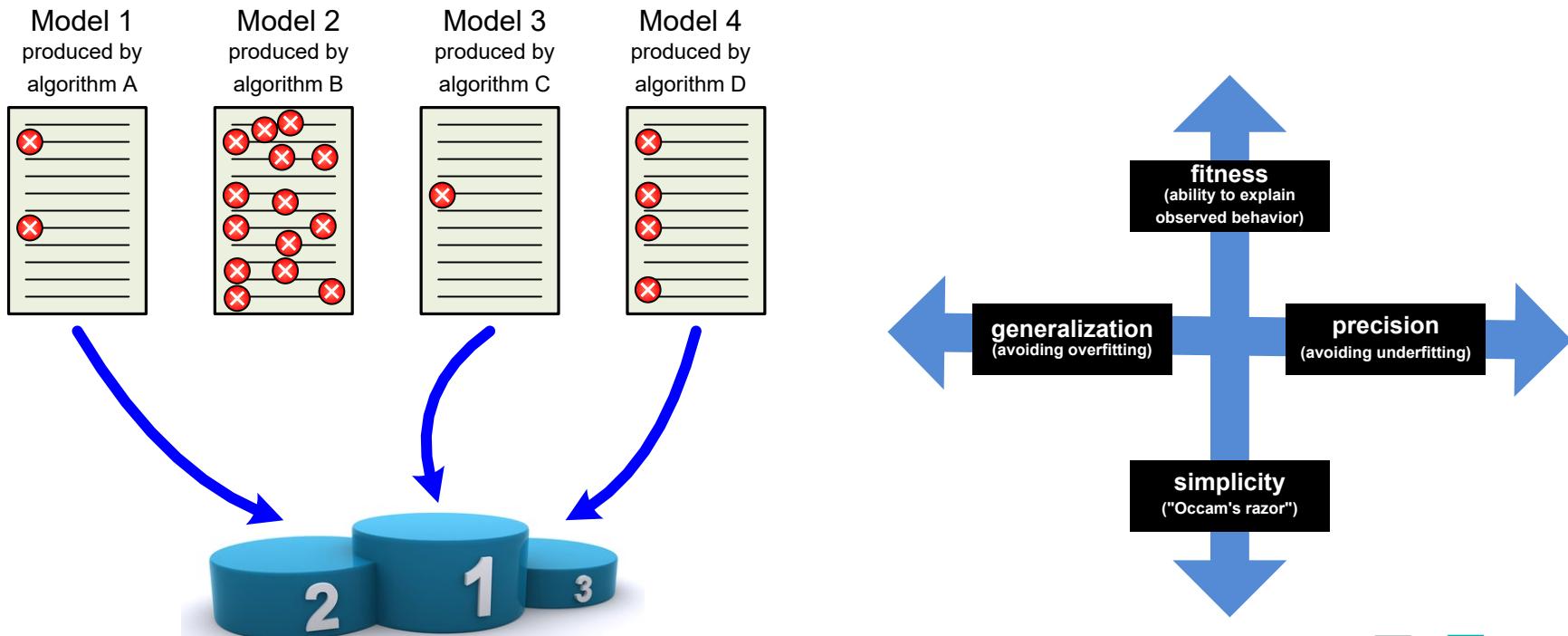
check budget (d)
is missing!



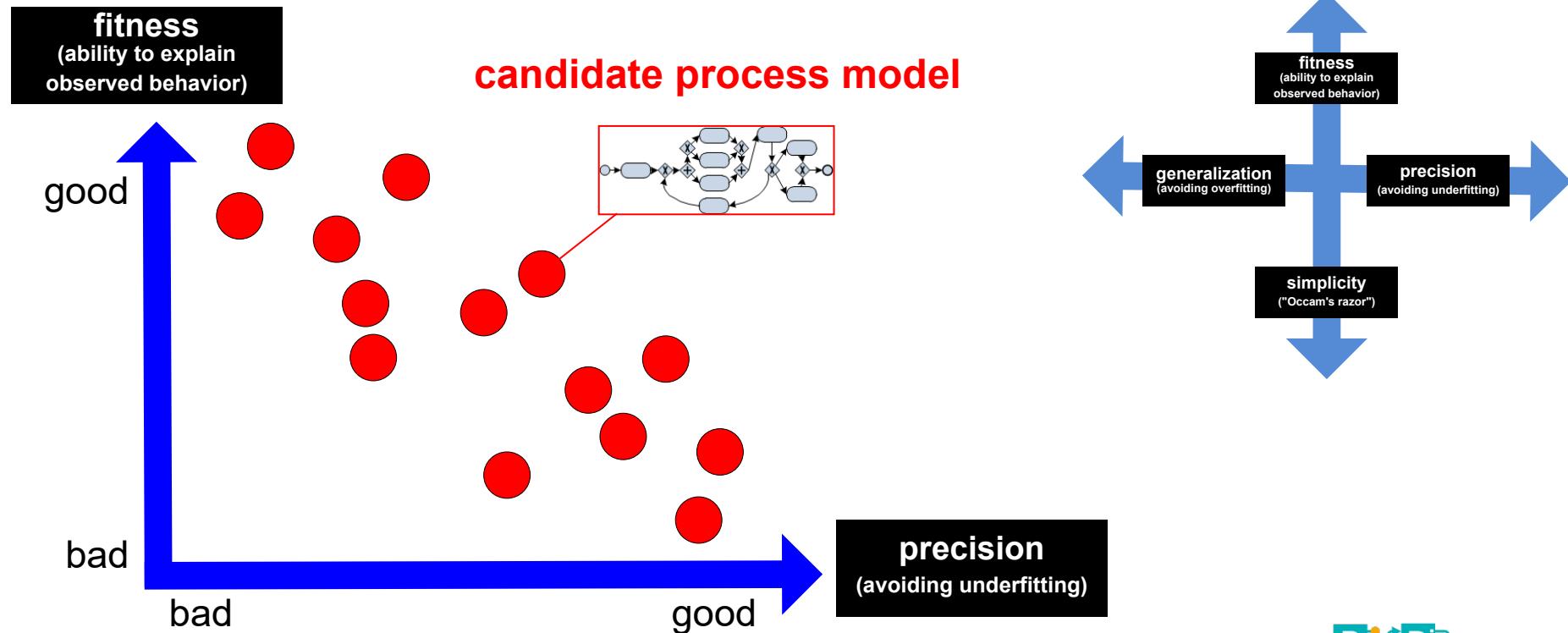
Conformance diagnostics and measures



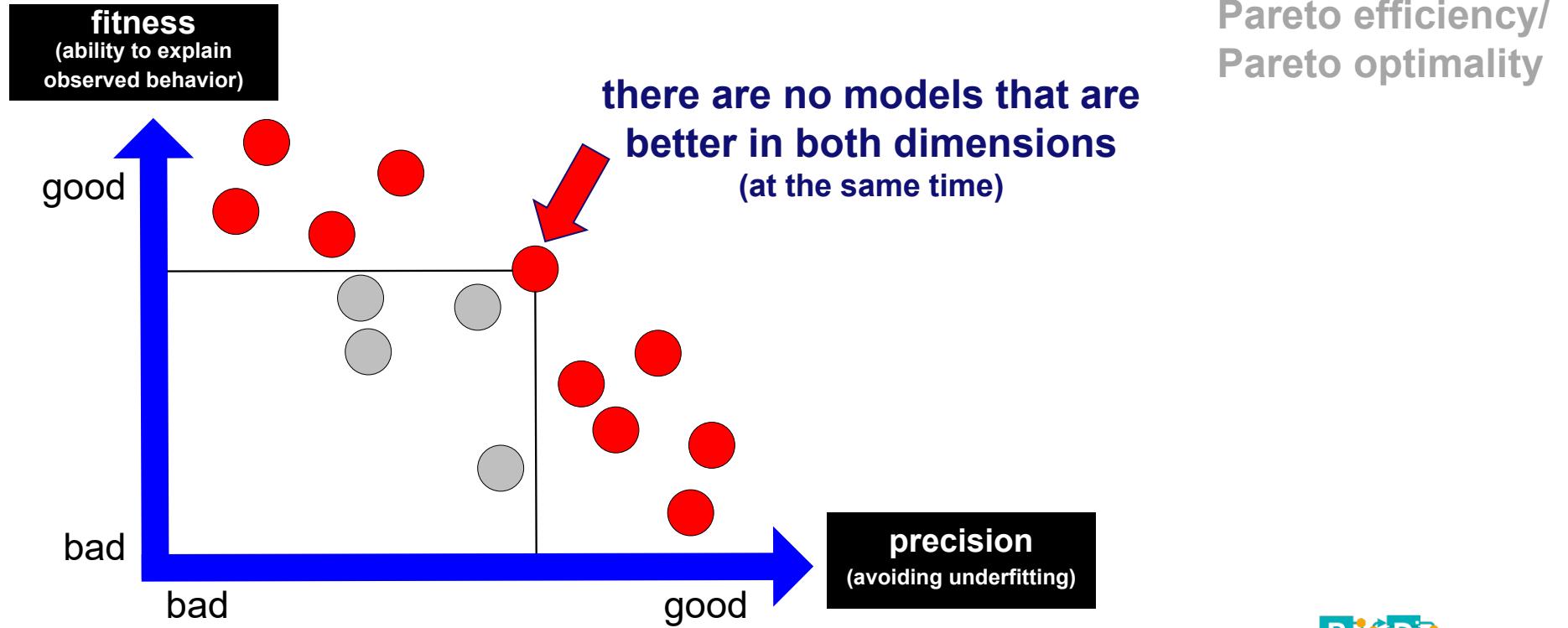
Evaluating process discovery algorithms



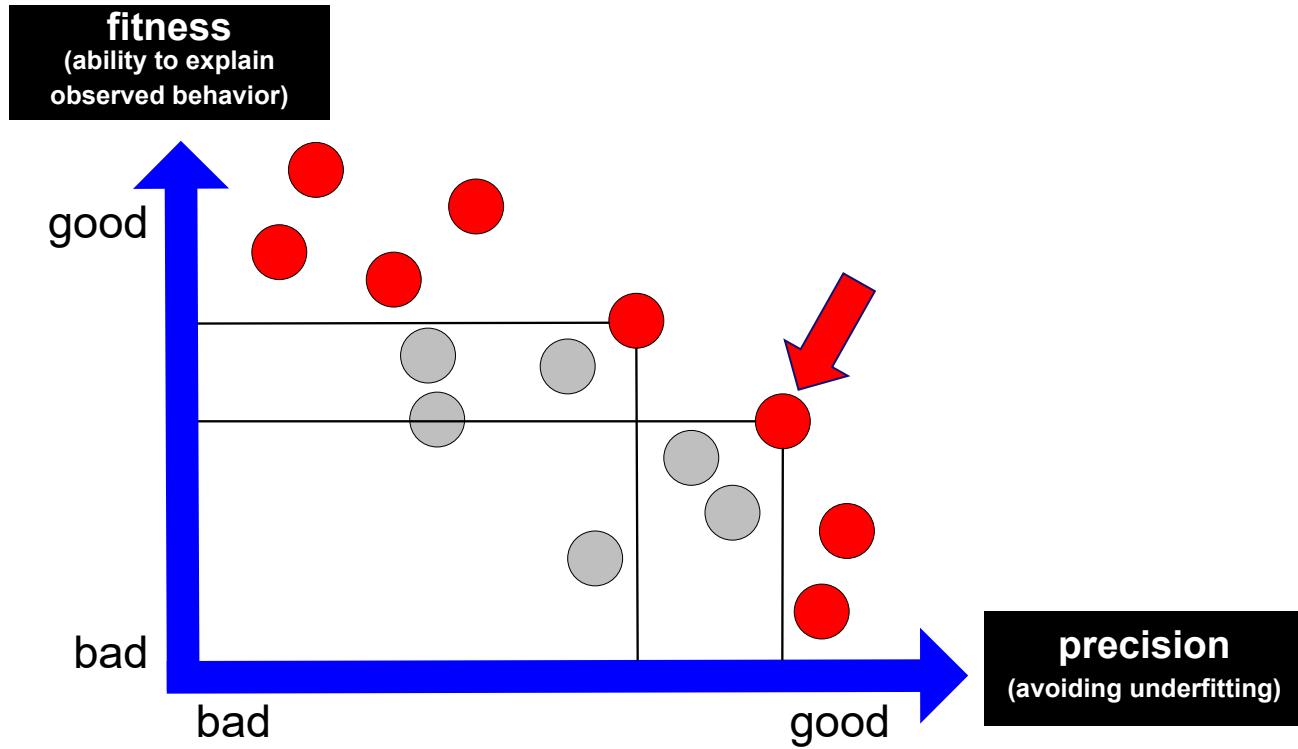
Pareto front considering two dimensions



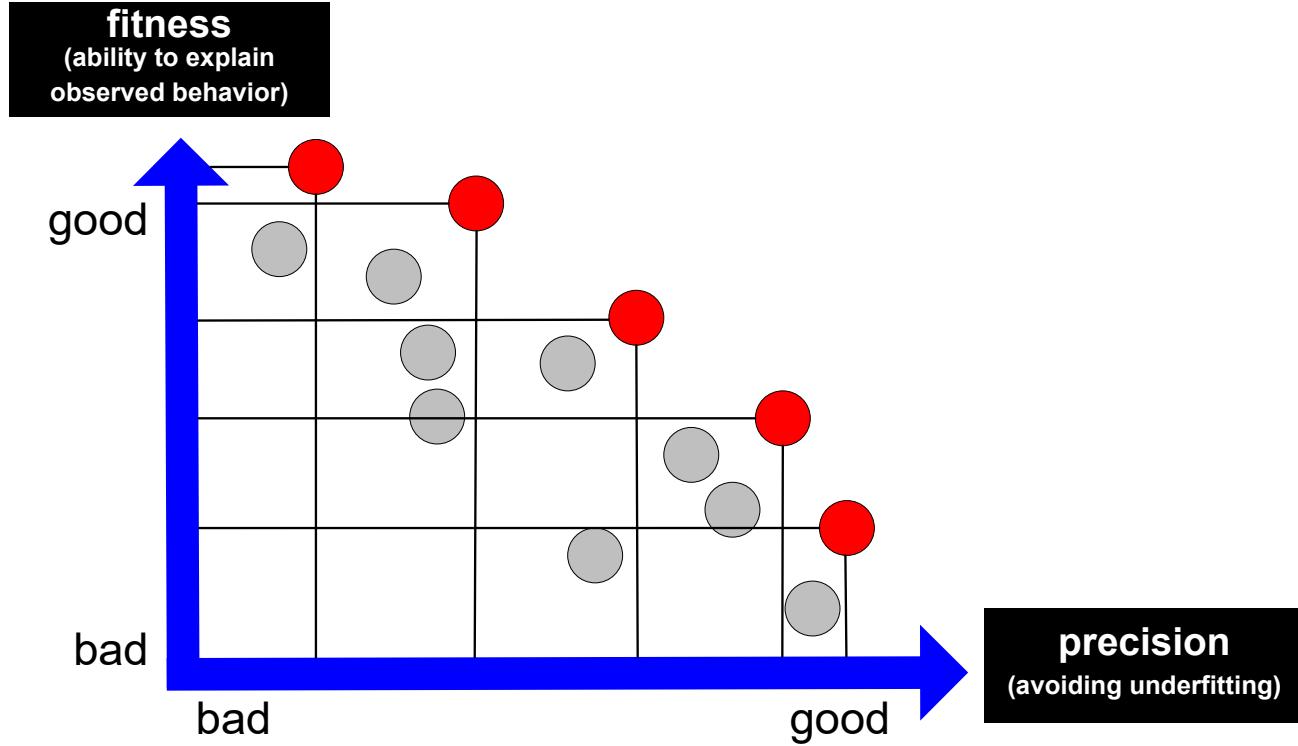
Pareto front considering two dimensions



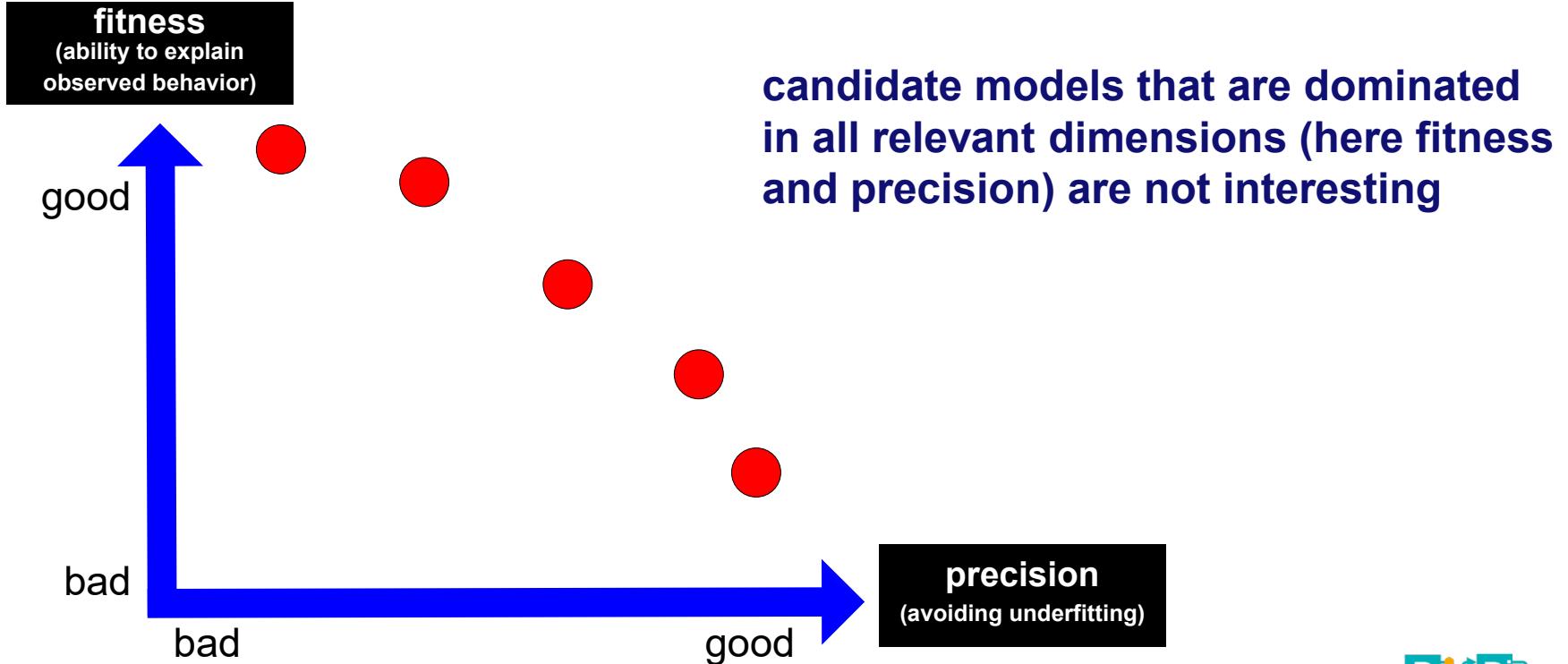
Pareto front considering two dimensions



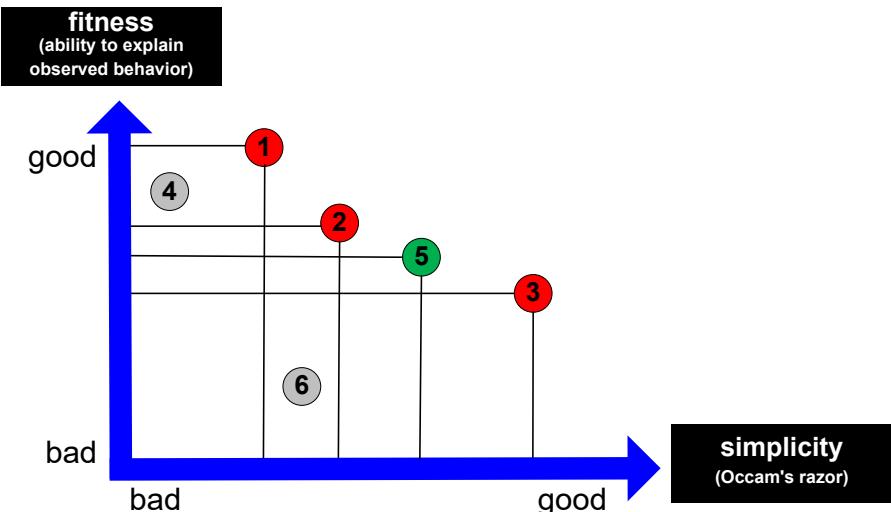
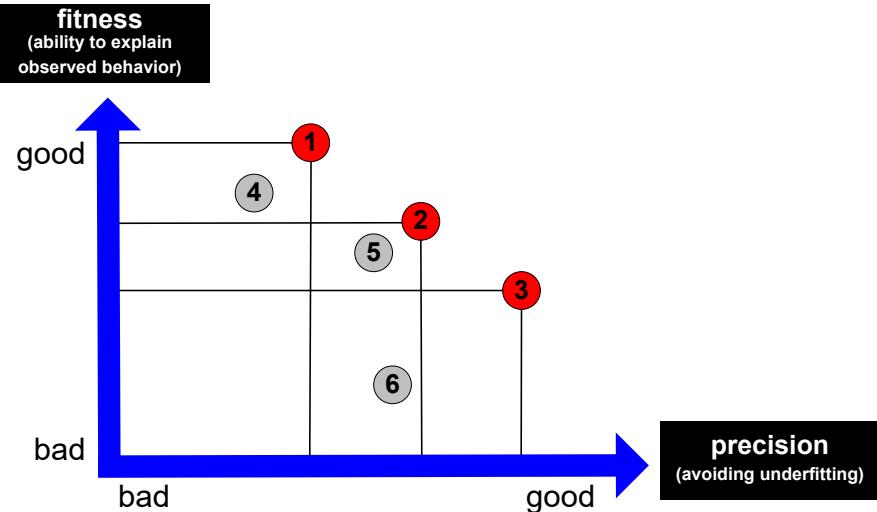
Pareto front considering two dimensions



Pareto front considering two dimensions

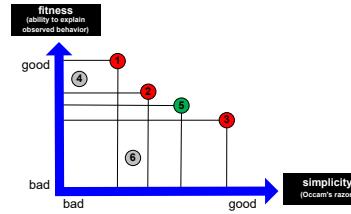
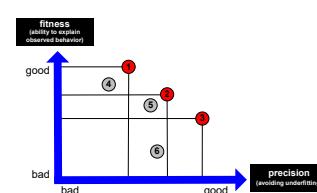
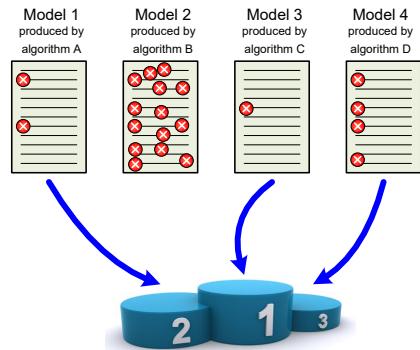


Adding another dimension: Simplicity



There is no model that (at the same time) has a better fitness, precision, and simplicity than model 5, i.e., it is not dominated.

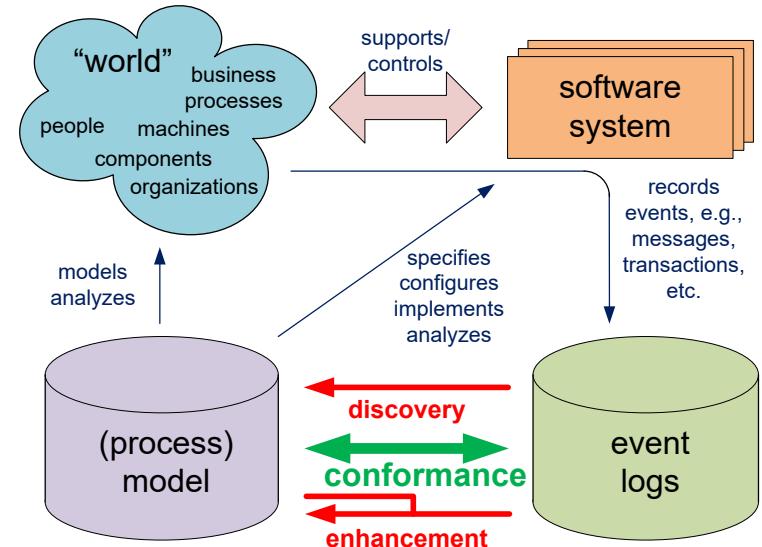
Comparing discovered models is not easy



- There is no such thing as "the best model".
 - Moreover, models may serve different purposes (see maps).
 - Initially, we focus on quantifying (replay) fitness (a.k.a. recall).

Conformance checking approaches

1. Conformance checking using **causal footprints**.
2. Conformance checking based on **token-based replay**.
3. **Alignment-based conformance checking.**



Focusing just on control-flow

- An event log is a finite multiset of traces:
$$L = [\langle a, b, c, d \rangle^6, \langle a, b, c, d \rangle^4]$$
- A model describes a (possibly infinite) set of traces:
$$M = \{ \langle a, b, c, d \rangle, \langle a, b, c, d \rangle \}$$
- As indicated before: Simple precision and recall measures do not work:
 - Loops lead to a recall of 0
 - Traces may be almost fitting
 - Log is only a sample and typically very incomplete
- Idea: Create a common finite abstraction, e.g., a DFG or causal footprint (next), dealing with incompleteness and loops.



Causal footprints



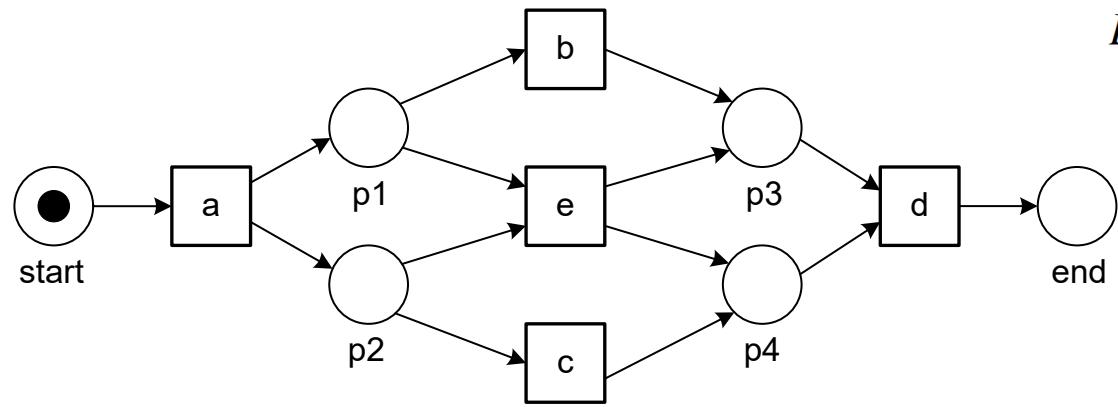
Footprint of L_1

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

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

- 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.**

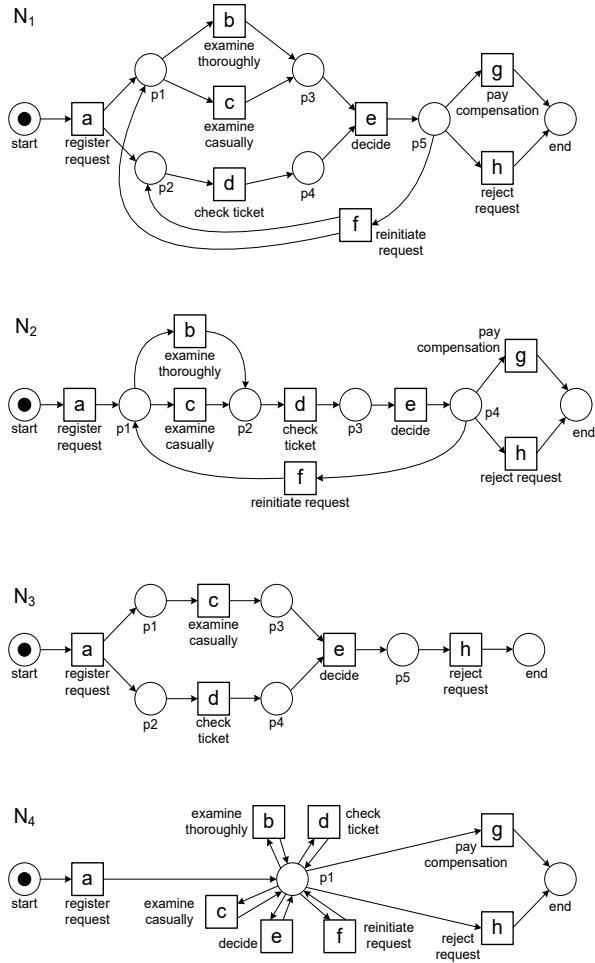
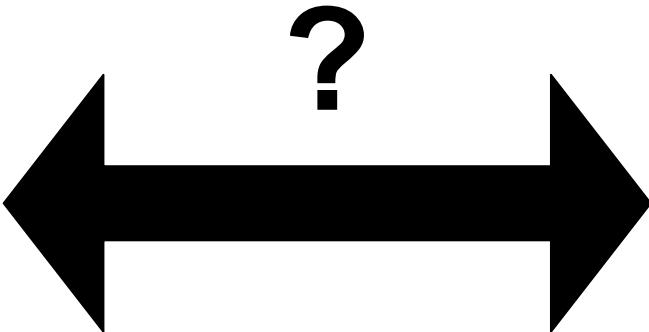
Discovered model has the same footprint



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

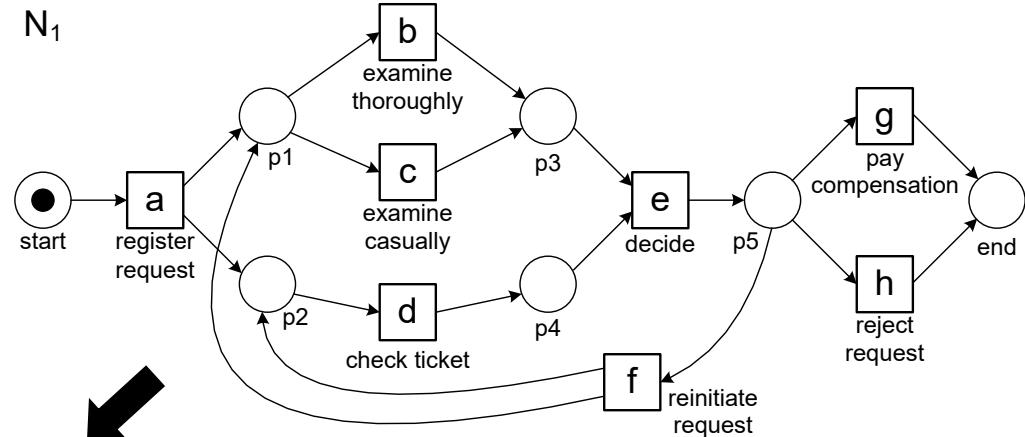
	a	b	c	d	e
a	$\#_{L_1}$	\rightarrow_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}
b	\leftarrow_{L_1}	$\#_{L_1}$	\parallel_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$
c	\leftarrow_{L_1}	\parallel_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}	$\#_{L_1}$
d	$\#_{L_1}$	\leftarrow_{L_1}	\leftarrow_{L_1}	$\#_{L_1}$	\leftarrow_{L_1}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

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdfbeh
38	adbeg
33	acdefbdeh
14	acdefbddeg
11	acdefdbeg
9	adcefcdbeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbddefbddeg
1	adcefdbefbdbeh
1	adbefbddefdbeg
1	adcefdbefcdefdbeg
1391	



#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbhe
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdbeh
11	acdefbdieg
9	adcefdeh
8	adcefdeh
5	adcfbdieg
3	acdefbdiegfbeg
2	adcfbdieg
2	adcfbdiegfbeg
1	adcfbdiegfbdeh
1	adbfbdiegfbeg
1	adcfdbefcdefdbeg
1	adcfdbefcdefdbeg

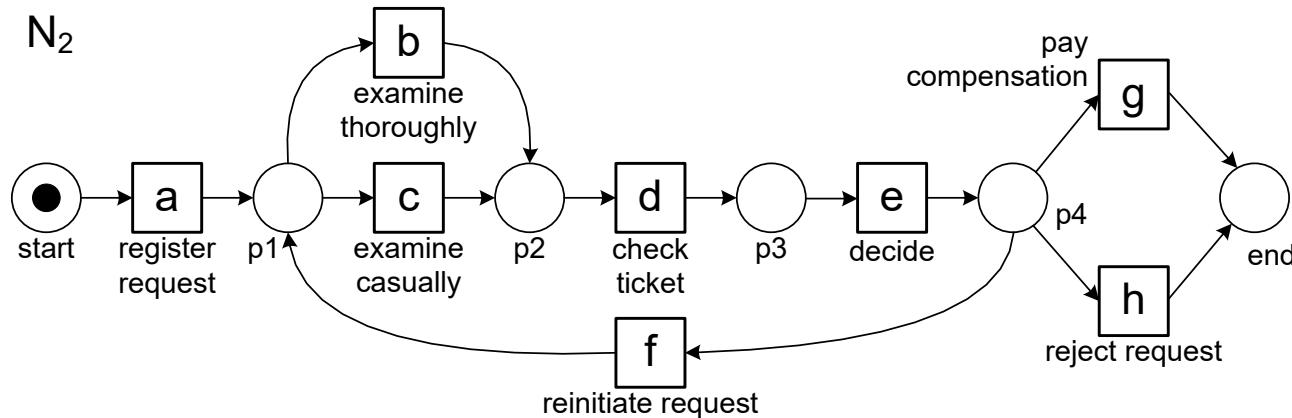
L_{full} and N_1



	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
<i>a</i>	#	→						#
<i>b</i>	←	#						#
<i>c</i>	←	#						#
<i>d</i>	←				"	"	"	#
<i>e</i>	#	←	←	←	#	→	→	→
<i>f</i>	#	→	→	→	→	#	#	#
<i>g</i>								#
<i>h</i>	#	#	#	#	←	#	#	#

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

footprints of log and model coincide



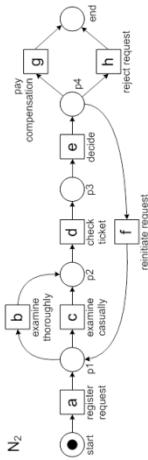
	a	b	c	d	e	f	g	h
a	#	\rightarrow	\rightarrow	#	#	#	#	#
b	\leftarrow	#	#	\rightarrow	#	\leftarrow	#	#
c	\leftarrow	#	#	\rightarrow	#	\leftarrow	#	#
d	#	\leftarrow	\leftarrow	#	\rightarrow	#	#	#
e	#	#	#	\leftarrow	#	\rightarrow	\rightarrow	\rightarrow
f	#	\rightarrow	\rightarrow	#	\leftarrow	#	#	#
g	#	#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

L_{full}

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdg
11	acdefdbeg
9	adcefdeh
8	adcefbeh
5	adcefbdg
3	adcefbdedefbeg
2	adcefdbeg
2	adcefbddefbdg
1	adcefdbefbdbeh
1	adbefbddefd beg
1	adcefbcfdefd beg
1391	

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

1391



N_2

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

Quantifying the differences

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

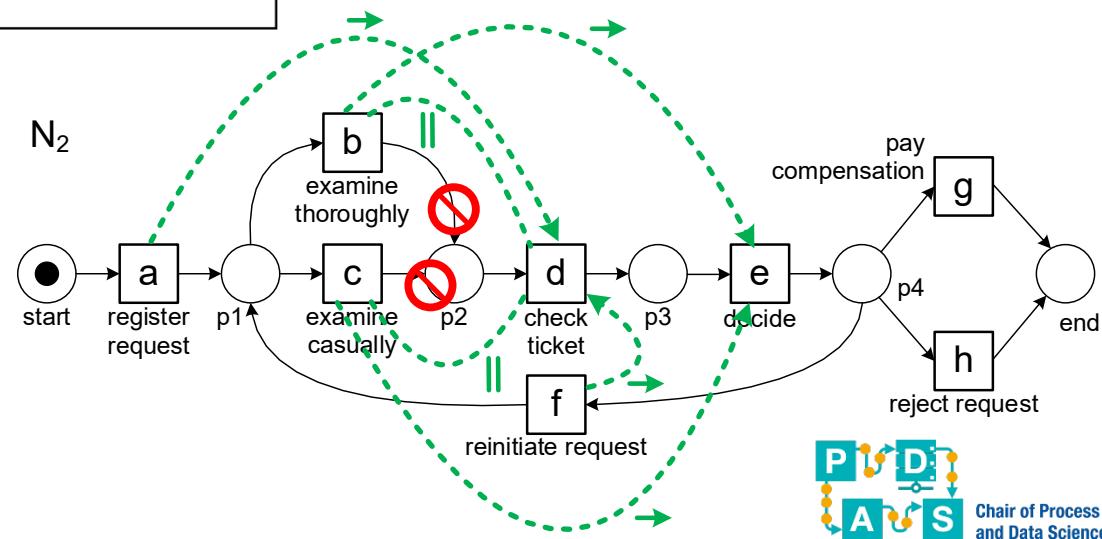
(x:y where x is in log and y in N₂)

footprint-based conformance

Diagnostics

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
<i>a</i>						$\rightarrow: \#$		
<i>b</i>					$\parallel : \rightarrow$	$\rightarrow: \#$		
<i>c</i>					$\parallel : \rightarrow$	$\rightarrow: \#$		
<i>d</i>	$\leftarrow: \#$	$\parallel : \leftarrow$	$\parallel : \leftarrow$				$\leftarrow: \#$	
<i>e</i>		$\leftarrow: \#$	$\leftarrow: \#$					
<i>f</i>						$\rightarrow: \#$		
<i>g</i>								
<i>h</i>								

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

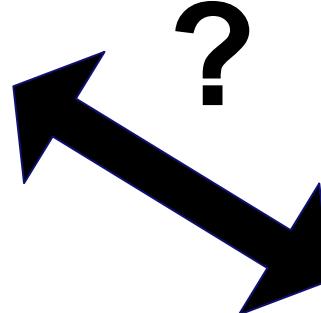


Question

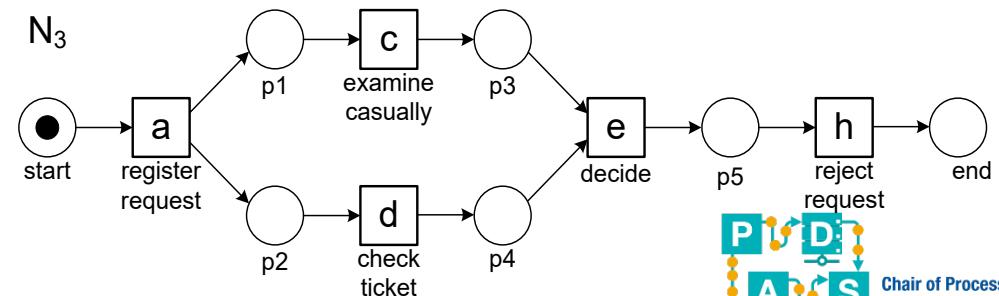
Estimate footprint-based conformance

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdeffbdeh
14	acdefbddeg
11	acdefdbeg
9	adcefcdbeh
8	adcefdbbeh
5	adcefbddeg
3	acdefbfdfdbeg
2	adcefdbeg
2	adcefbddefbddeg
1	adcefdbefbdbeh
1	adbefbfdfdbeg
1	adcefdbefcdfdbeg
1391	

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



Estimate the fraction of matching cells in footprint matrices

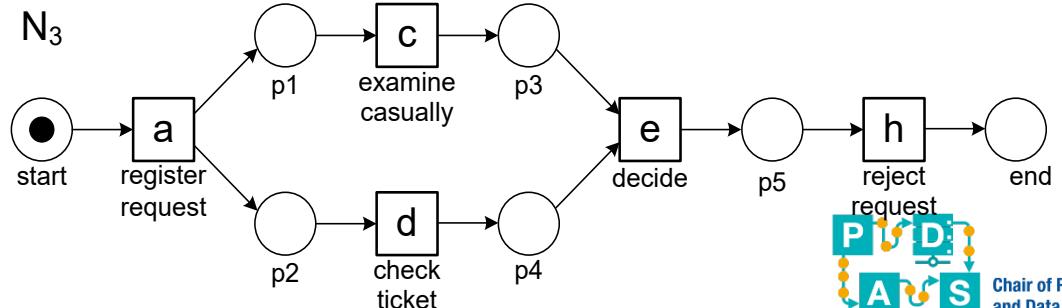


Answer

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
<i>a</i>	#	#	→	→	#	#	#	#
<i>b</i>	#	#	#	#	#	#	#	#
<i>c</i>	←	#	#		→	#	#	#
<i>d</i>	←	#		#	→	#	#	#
<i>e</i>	#	#	←	←	#	#	#	→
<i>f</i>	#	#	#	#	#	#	#	#
<i>g</i>	#	#	#	#	#	#	#	#
<i>h</i>	#	#	#	#	←	#	#	#

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

footprint-based conformance



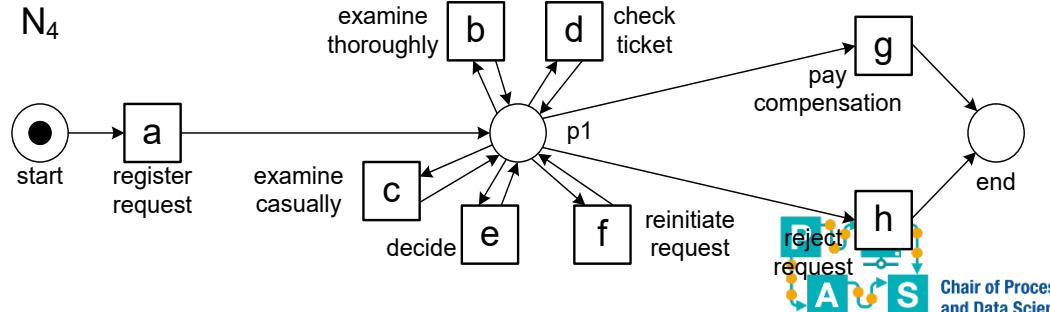
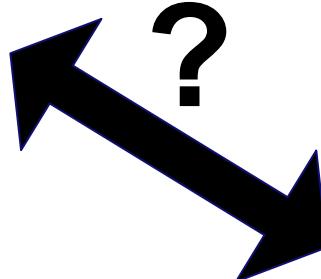
Question

Estimate footprint-based conformance

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdbeh
14	acdefbddeg
11	acdefdbeg
9	adcefcdbeh
8	adcefdbbeh
5	adcefbddeg
3	acdefbdfdfbeg
2	adcefdbeg
2	adcefbddefbddeg
1	adcefdbefbdbeh
1	adbefbddefdbeg
1	adcefdbeccdefdbeg
1391	

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

Estimate the fraction of matching cells in footprint matrices



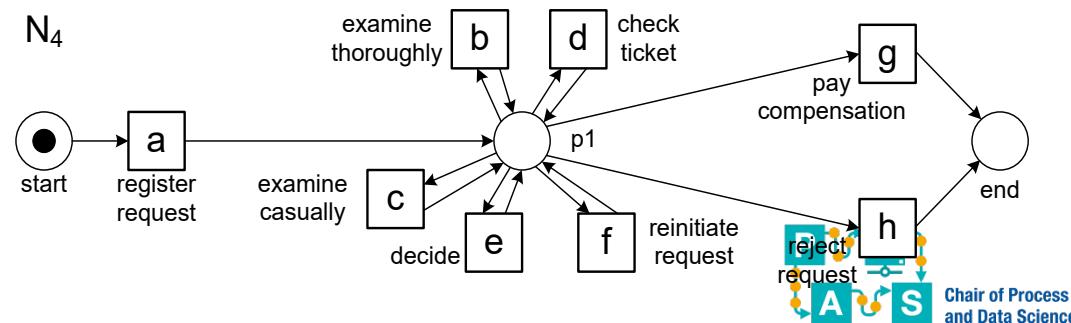
Answer

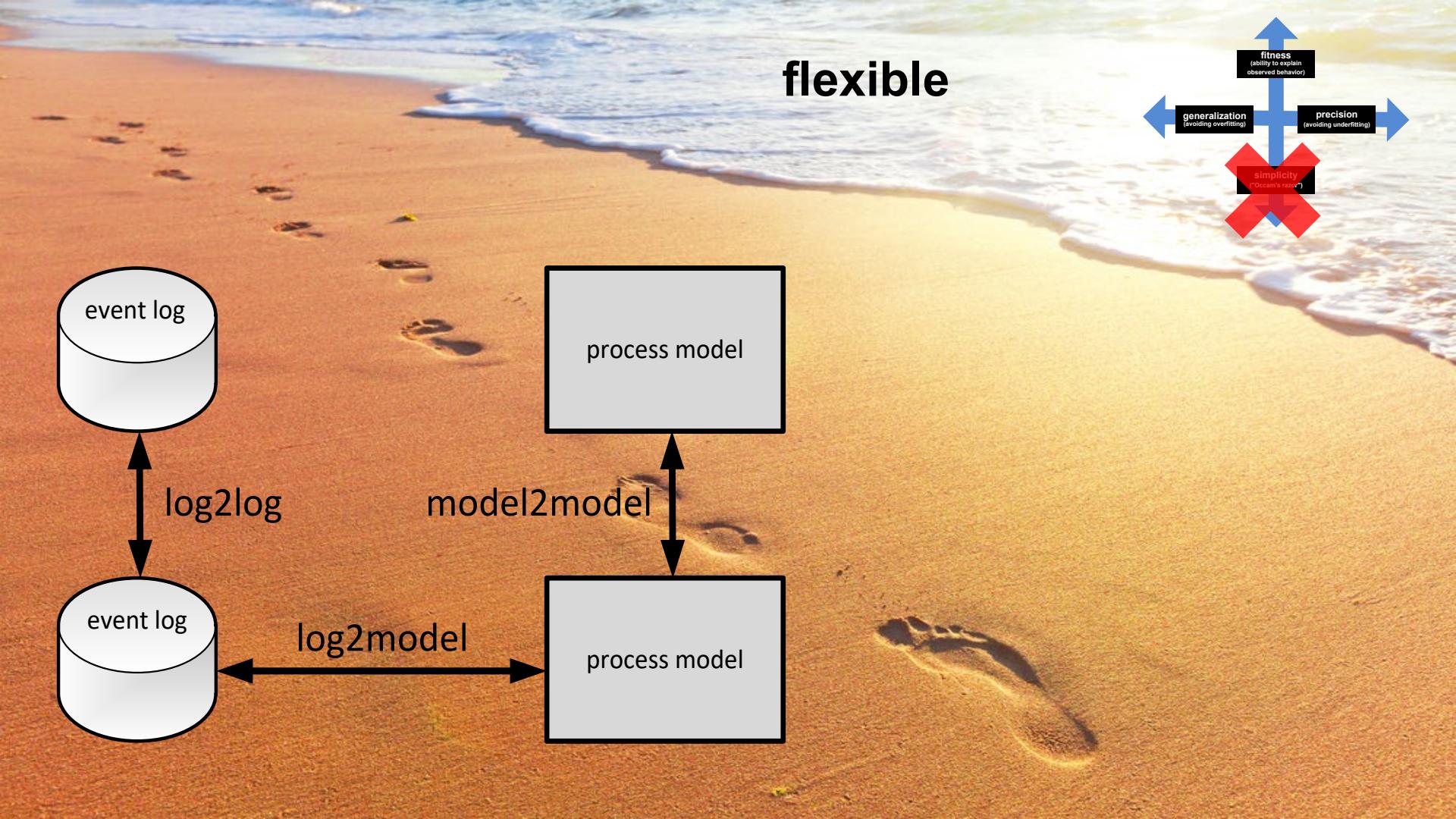
	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
<i>a</i>	#	→	→	→	#	#	#	#
<i>b</i>	←					←	#	#
<i>c</i>	←					←	#	#
<i>d</i>	←					←	#	#
<i>e</i>	#	←	←	←			→	→
<i>f</i>	#						#	#
<i>g</i>	#	←	←	←	←	←	#	#
<i>h</i>	#	←	←	←	←	←	#	#

Color
 • Log
 • Model

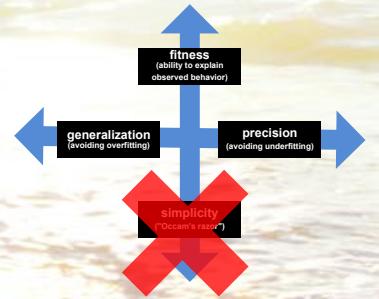
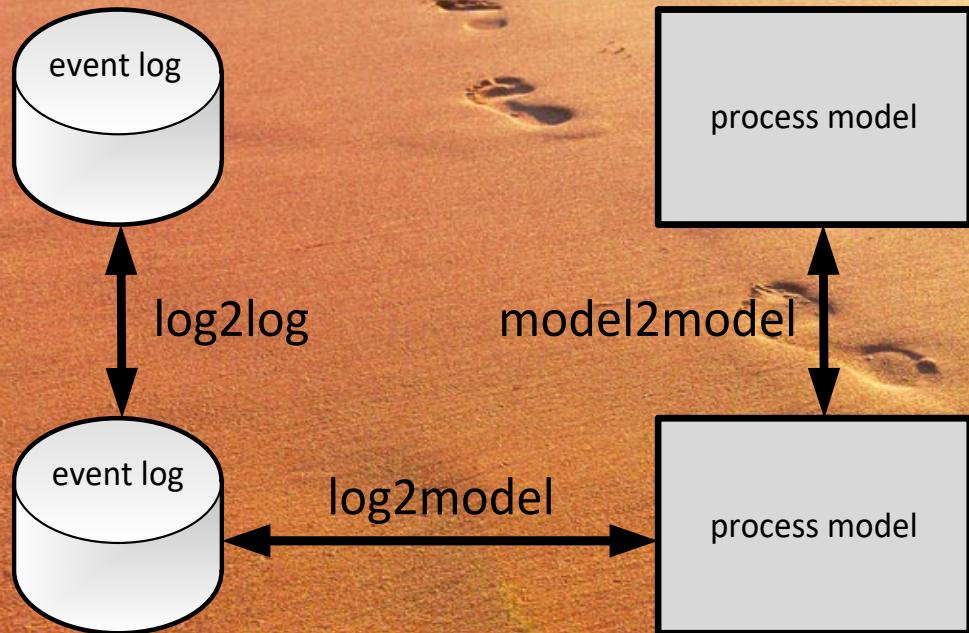
$$1 - \frac{45}{64} = 0.296875$$

footprint-based conformance



The background of the diagram is a photograph of a sandy beach meeting the ocean at the water's edge. Several sets of footprints are visible in the sand, leading towards the water.

flexible



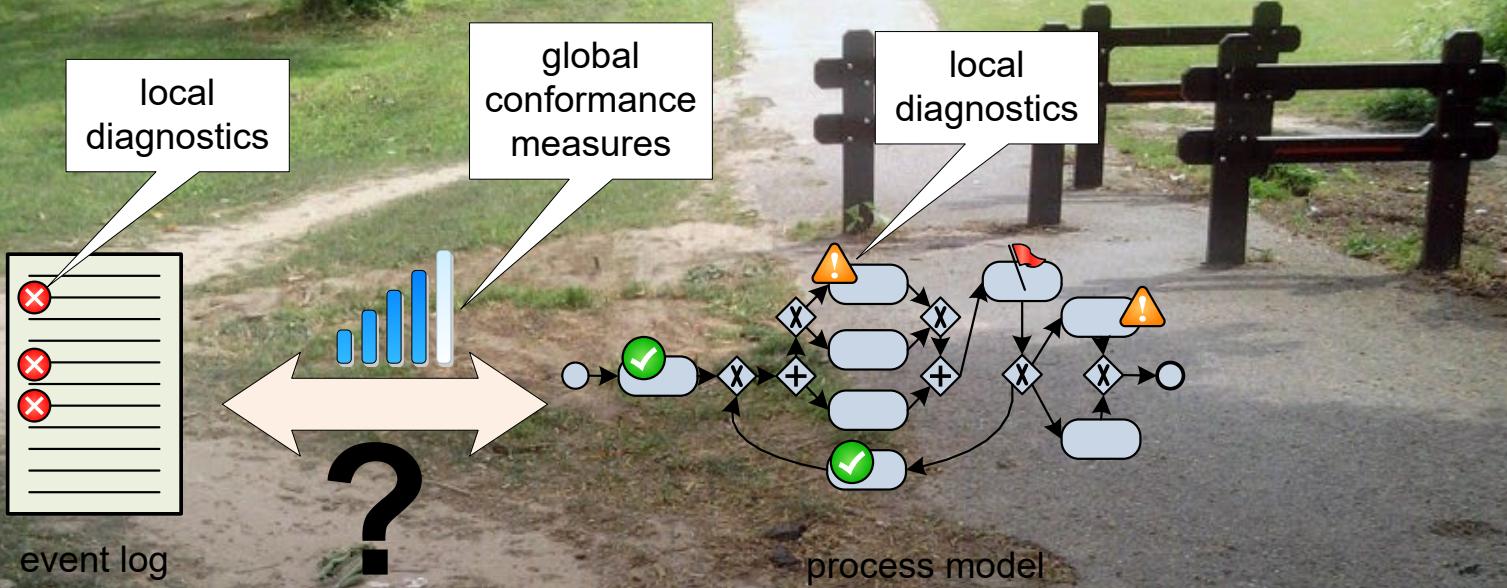
Limitations

- Frequencies are not used.
- Behavior is only considered indirectly (directly follows relation).
- Aims to capture fitness, precision and generalization in a single metric.

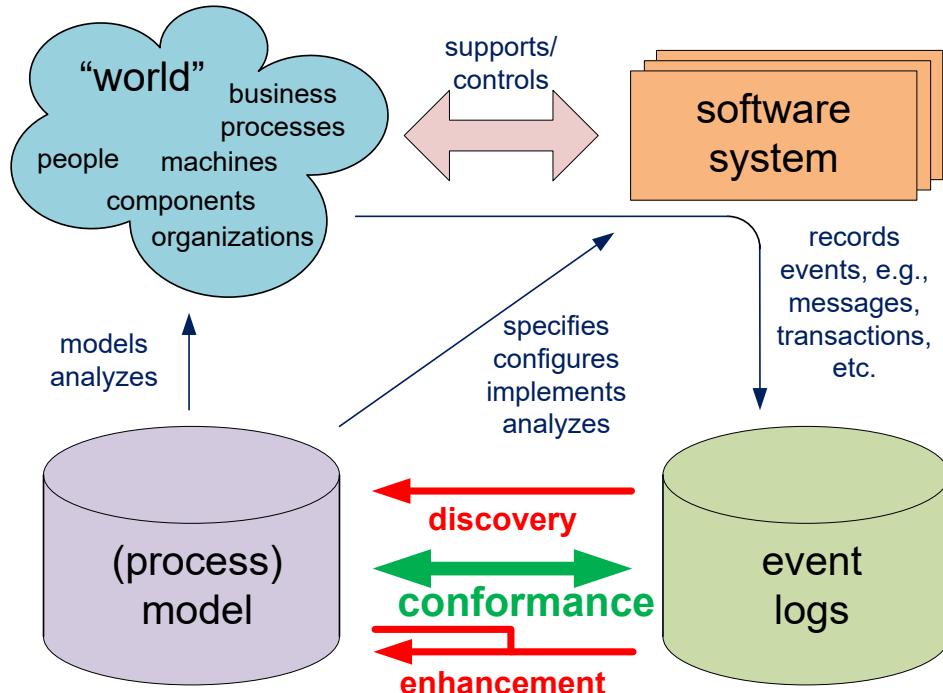
Next: conformance checking using token-based replay

Conformance checking using token-based replay



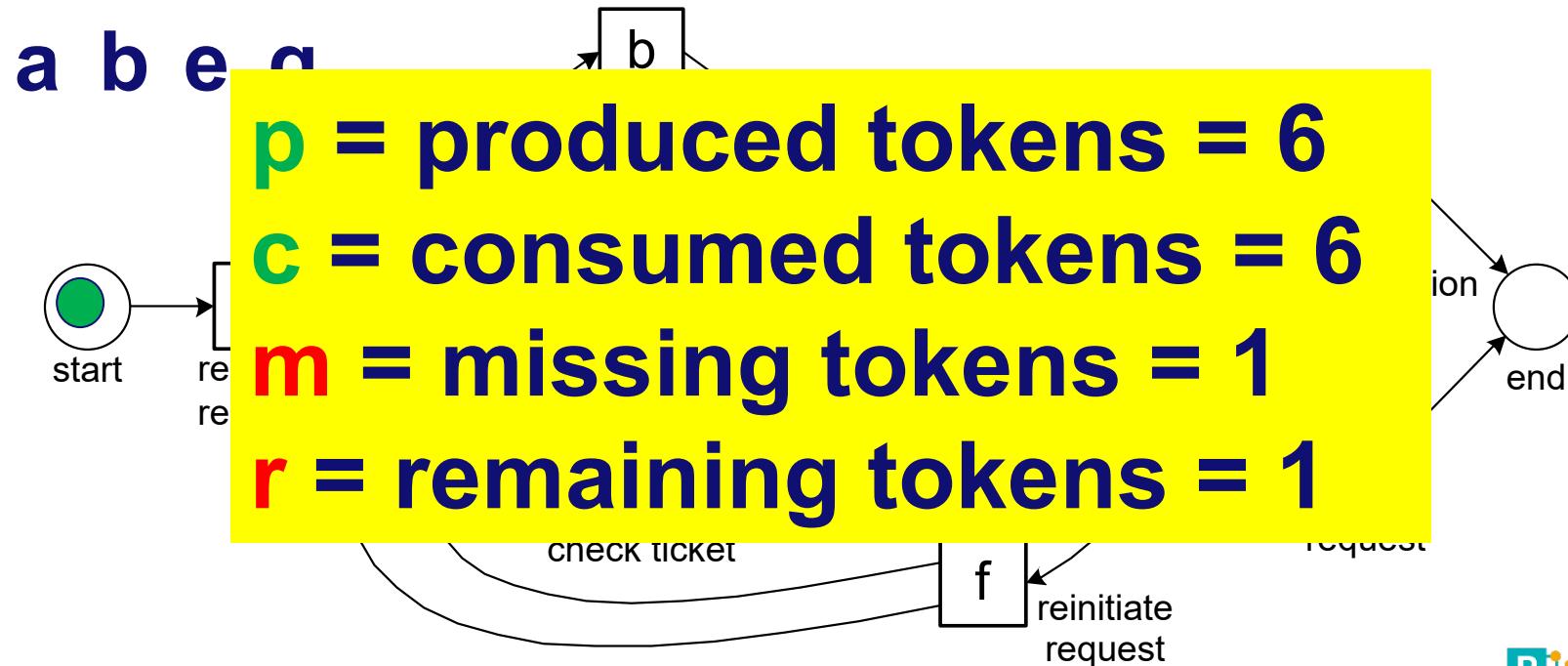


Conformance checking



1. Conformance checking using causal footprints.
2. Conformance checking based on **token-based replay**.
3. Alignment-based conformance checking.

Counting tokens while replaying



Quantifying fitness at the trace level

$$fitness(\sigma, N) = \frac{1}{2} \left(1 - \frac{1}{6}\right) + \frac{1}{2} \left(1 - \frac{1}{6}\right) = 0.83333$$

p = produced tokens = 6

c = consumed tokens = 6

m = missing tokens = 1

r = remaining tokens = 1

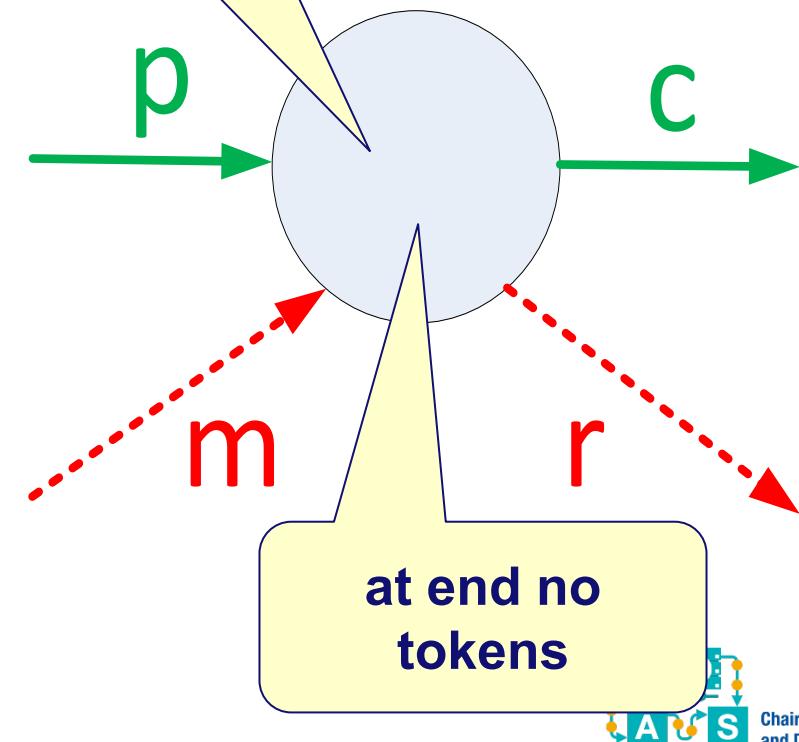


Approach (1/3)

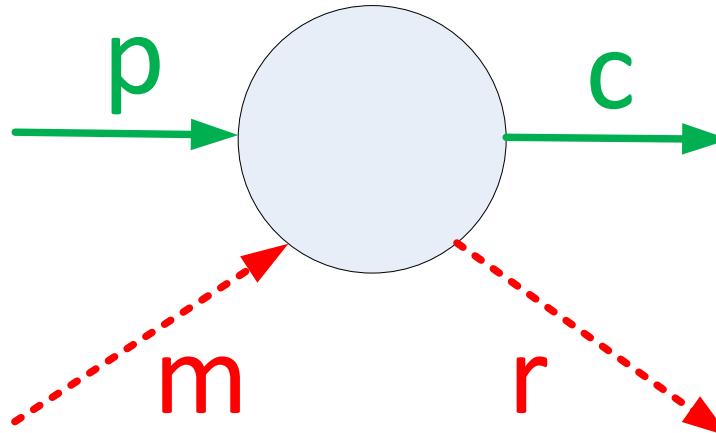
while running
 $p+m-c$ tokens

Use four counters:

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

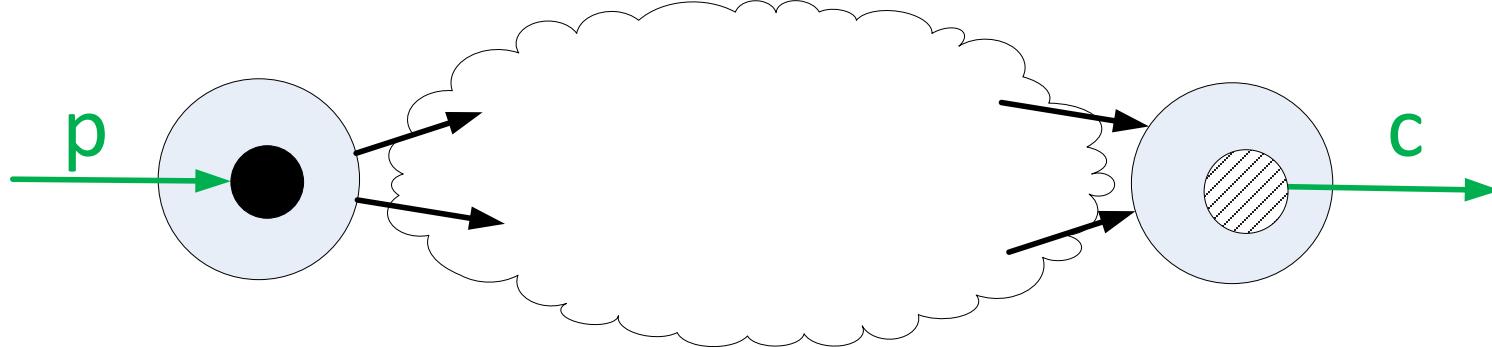


Approach (2/3)



- Invariants
 - At any time: $p+m \geq c \geq m$ (also per place)
 - At the end: $r = p + m - c$ (also per place)

Approach (3/3)

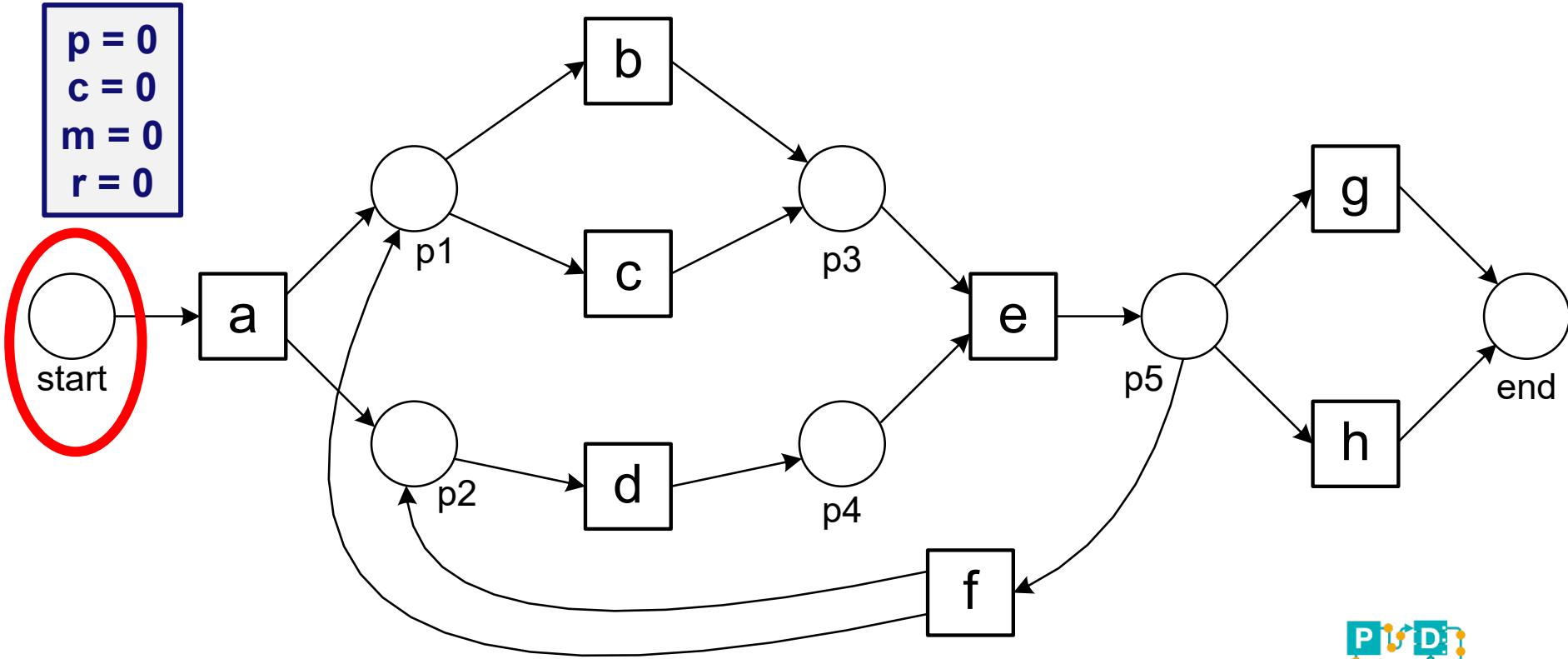


Initialization and finalization:

- In the beginning a token is **produced** for the source place: $p = 1$.
- At the end a token is **consumed** from the sink place (also if not there): $c' = c + 1$.

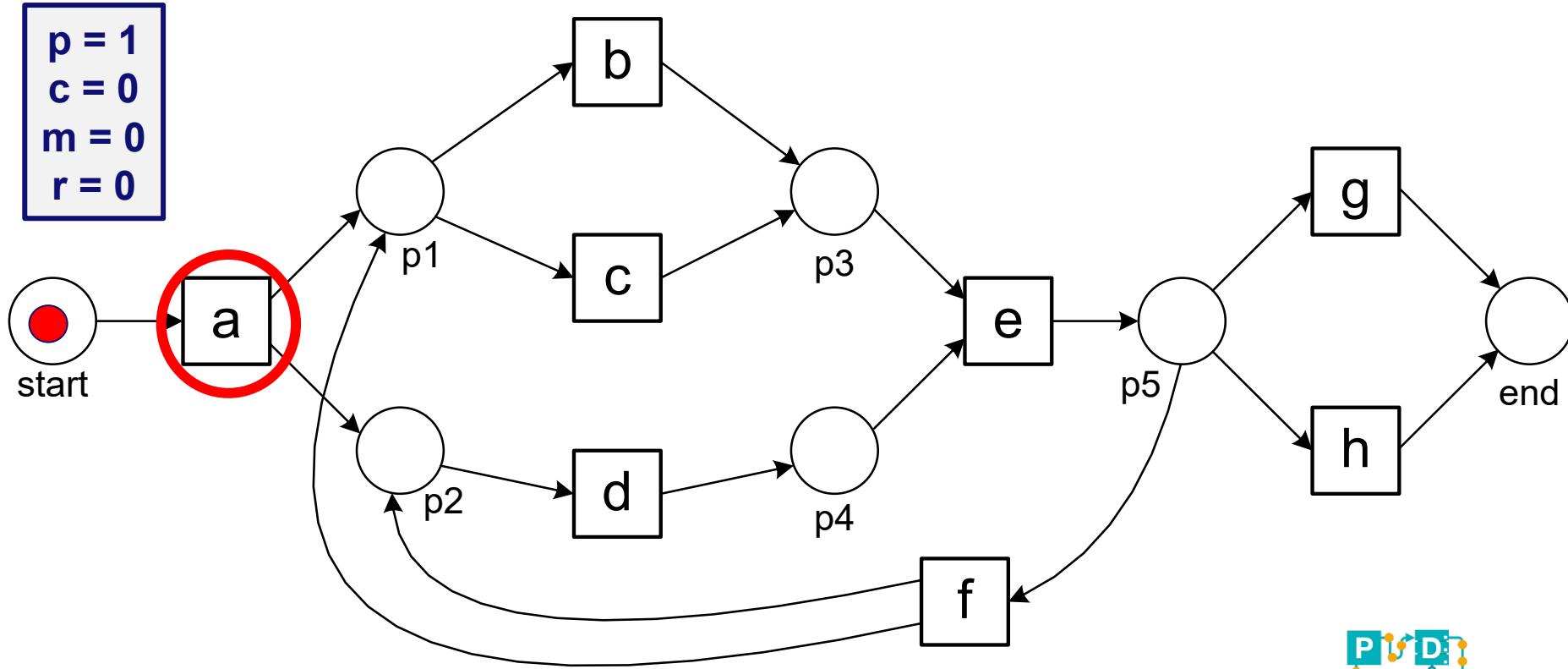
Replaying

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



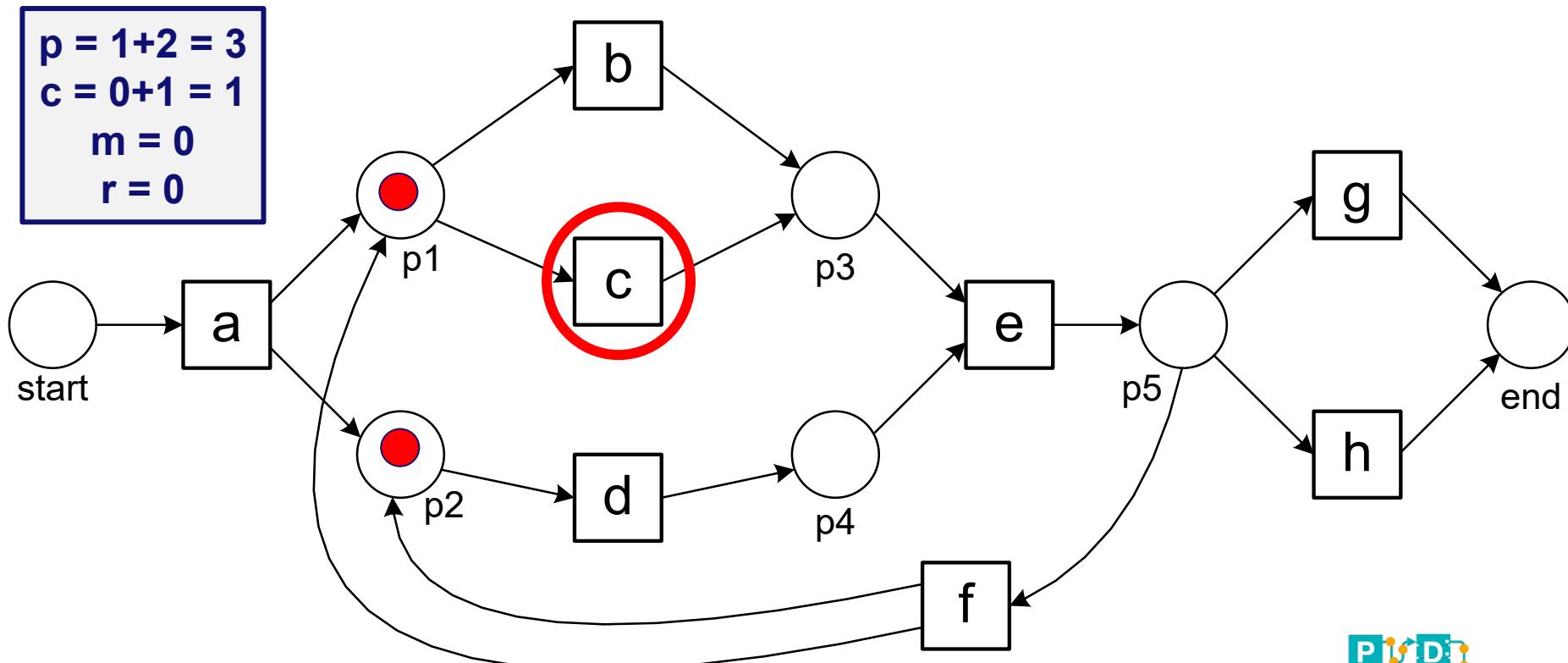
Replaying

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



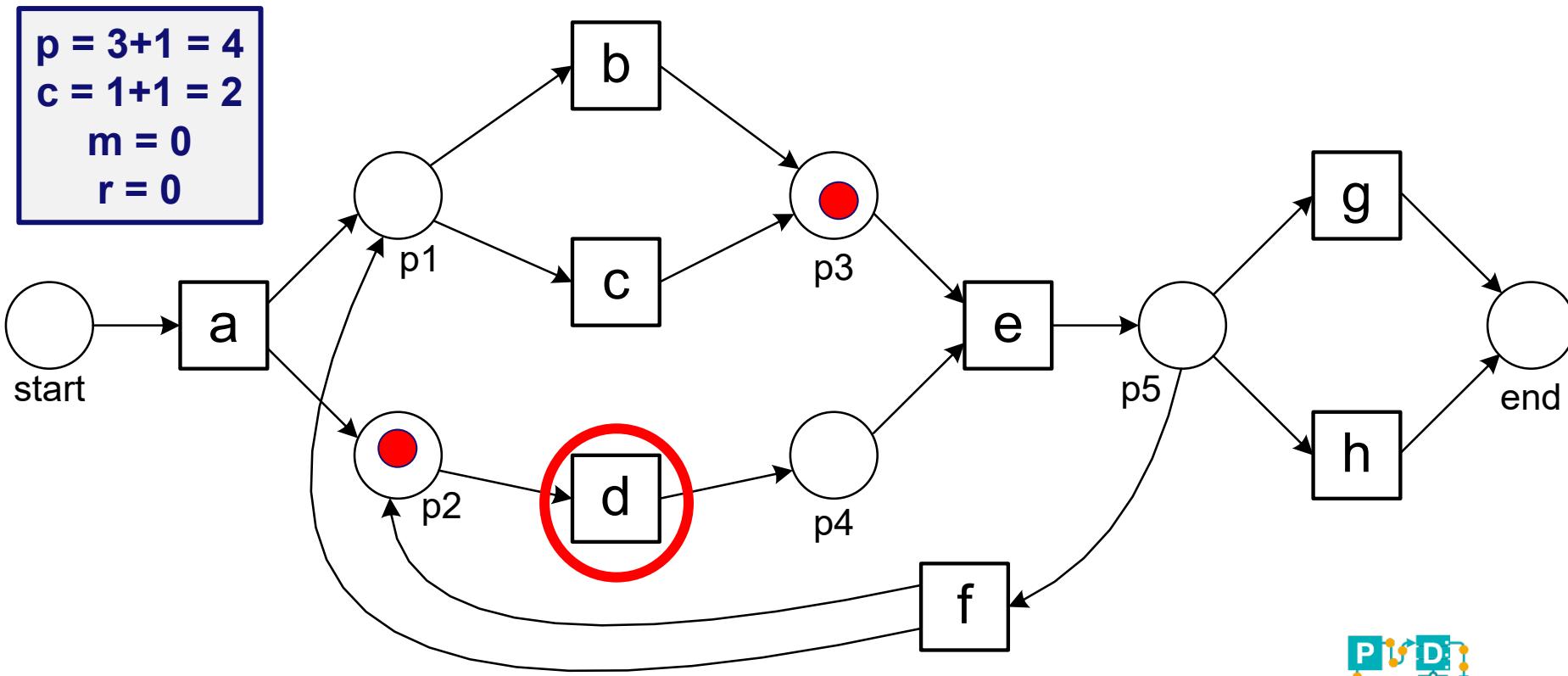
Replaying

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



Replaying

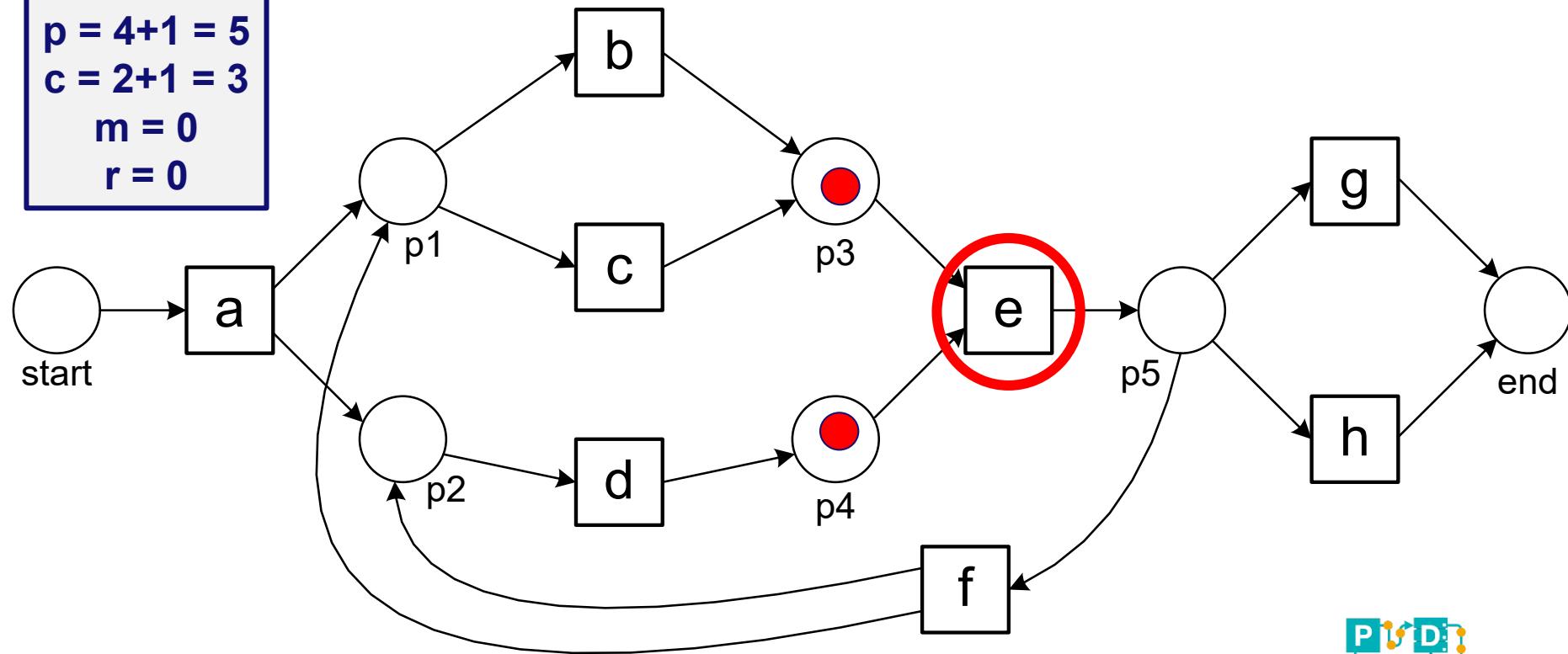
$$\sigma_1 = \langle a, c, d, e, h \rangle$$



Replaying

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

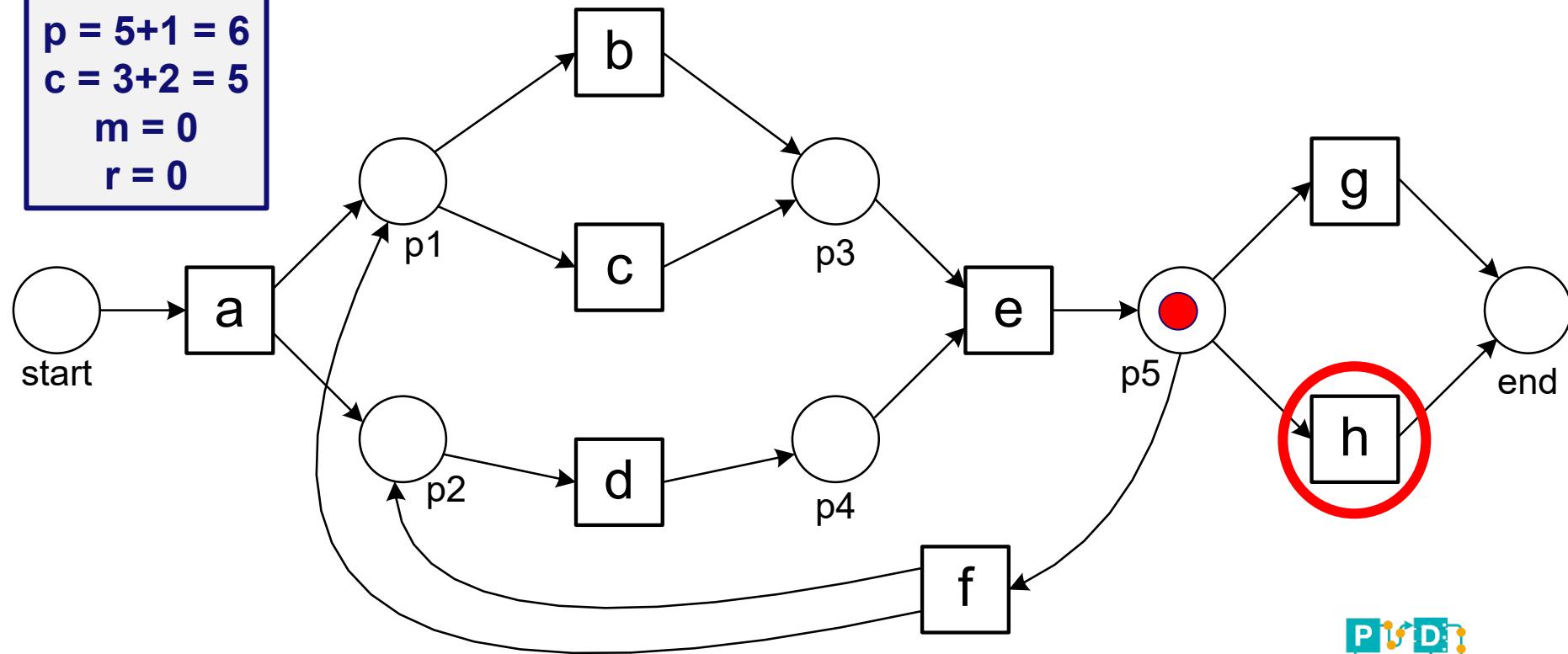
$p = 4+1 = 5$
 $c = 2+1 = 3$
 $m = 0$
 $r = 0$



Replaying

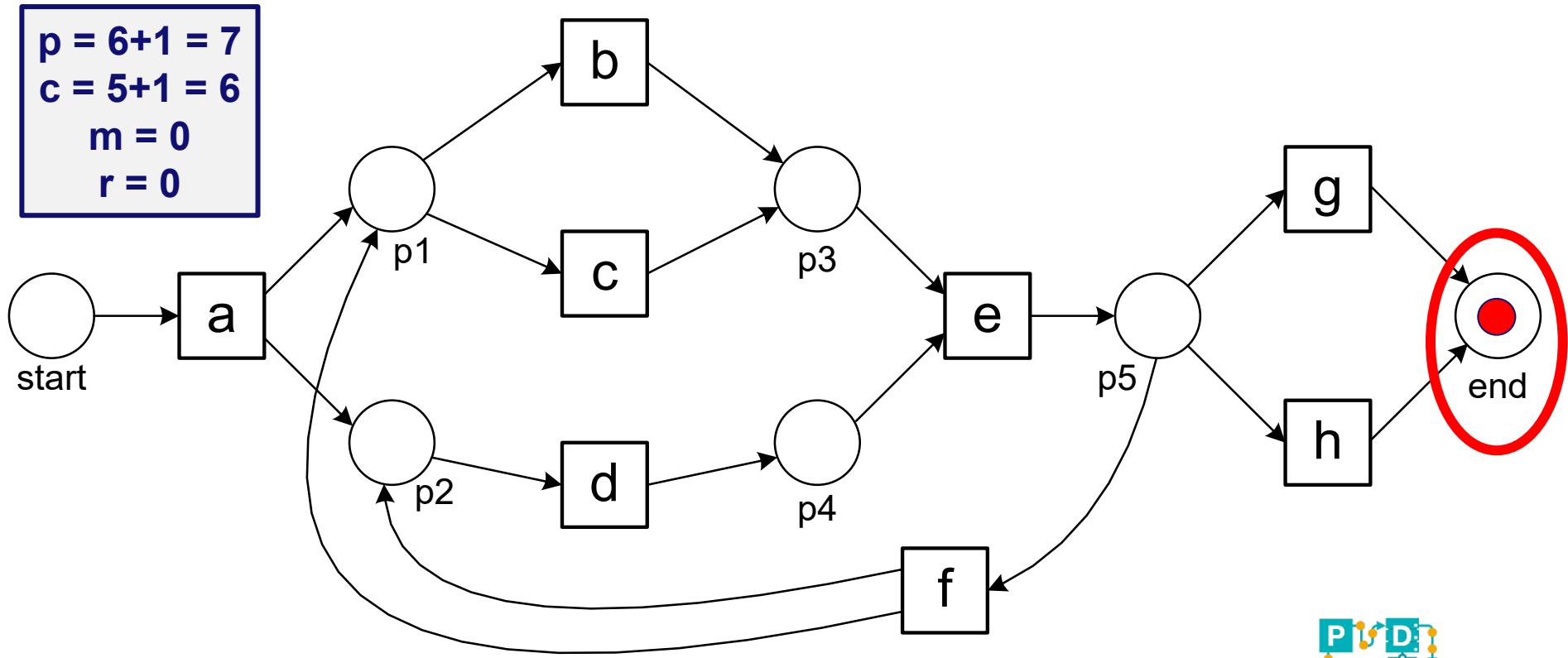
$$\sigma_1 = \langle a, c, d, e, h \rangle$$

$p = 5+1 = 6$
 $c = 3+2 = 5$
 $m = 0$
 $r = 0$



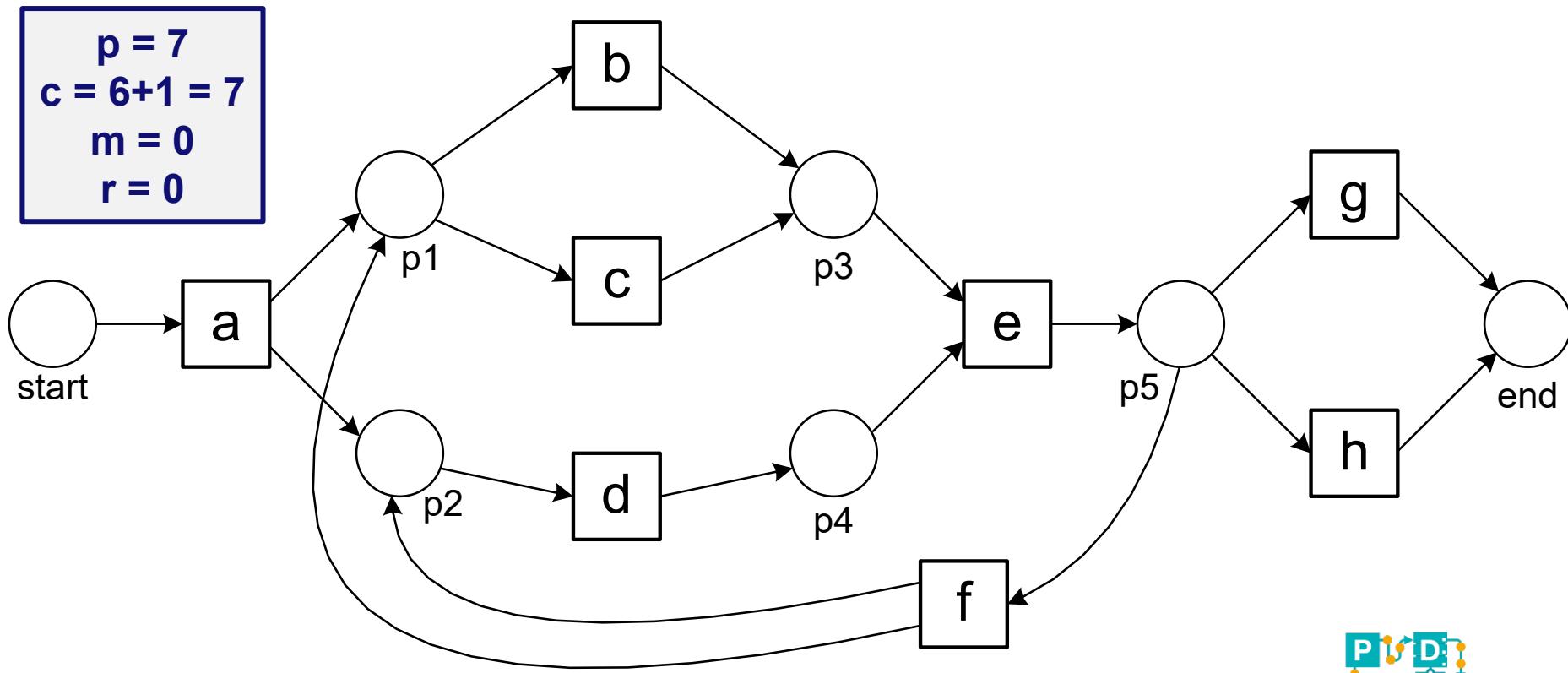
Replaying

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



Replaying

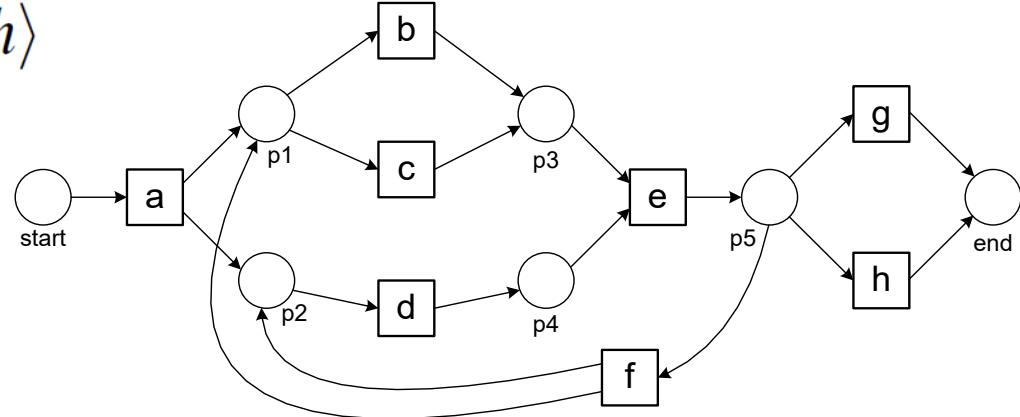
$$\sigma_1 = \langle a, c, d, e, h \rangle$$



Quantifying fitness at the trace level

p = 7
c = 7
m = 0
r = 0

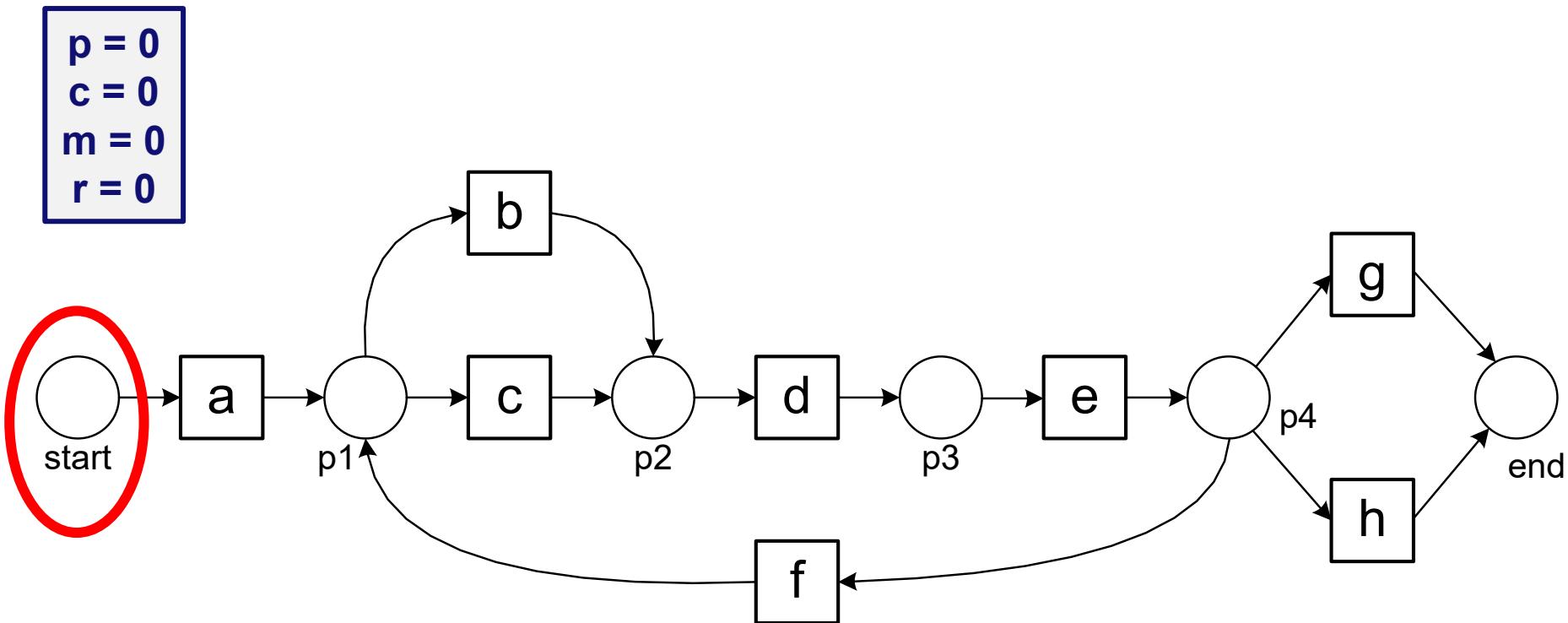
$$\sigma_1 = \langle a, c, d, e, h \rangle$$



$$fitness(\sigma, N) = \frac{1}{2} \left(1 - \frac{0}{7} \right) + \frac{1}{2} \left(1 - \frac{0}{7} \right) = 1$$

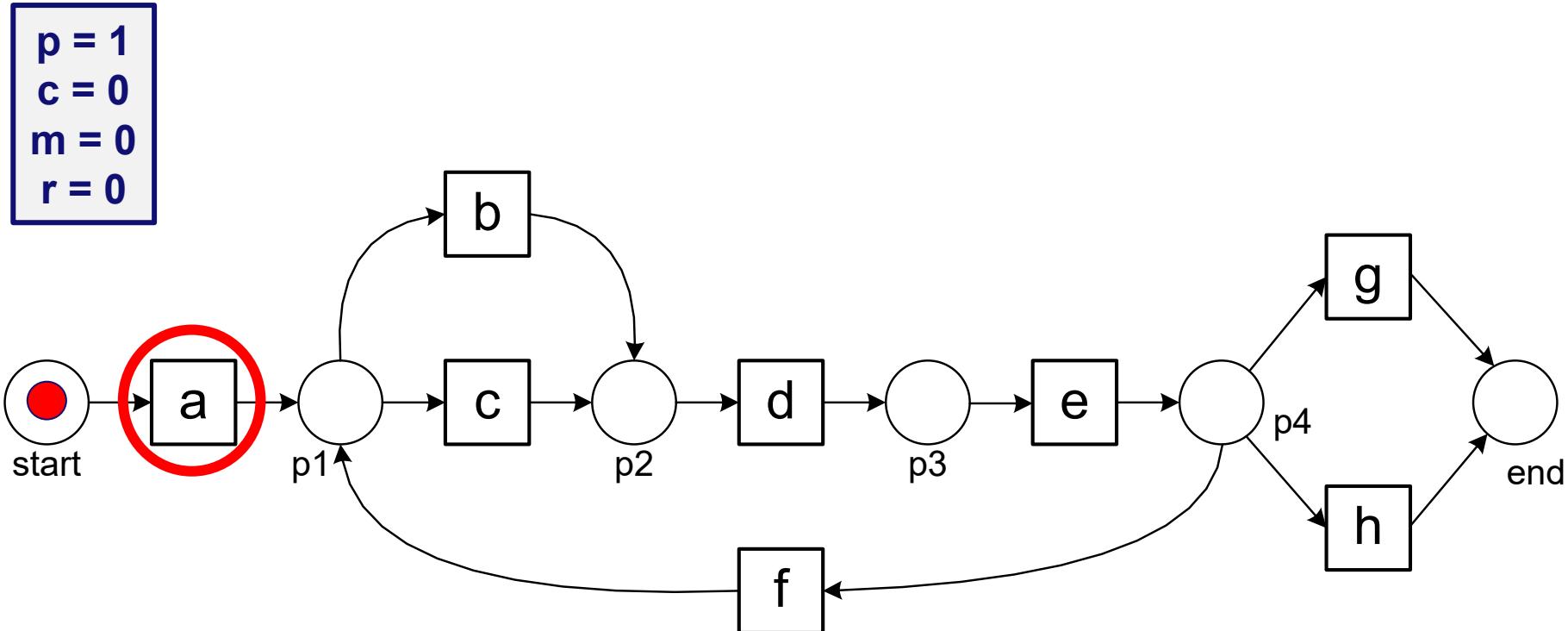
Replaying

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



Replaying

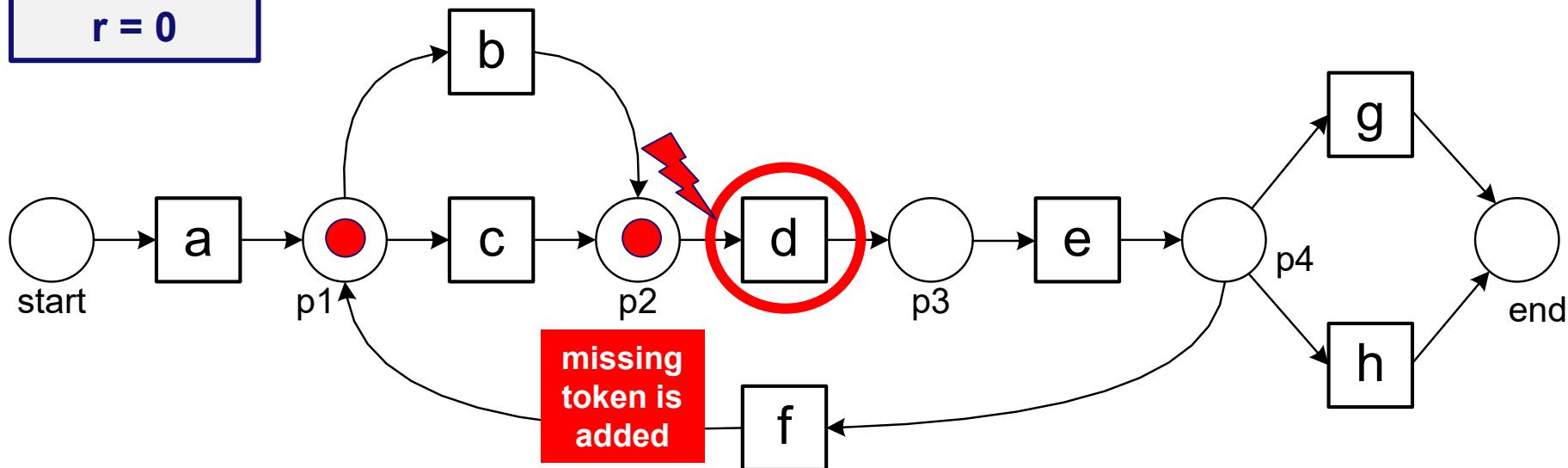
$$\sigma_3 = \langle a, d, c, e, h \rangle$$



Replaying

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

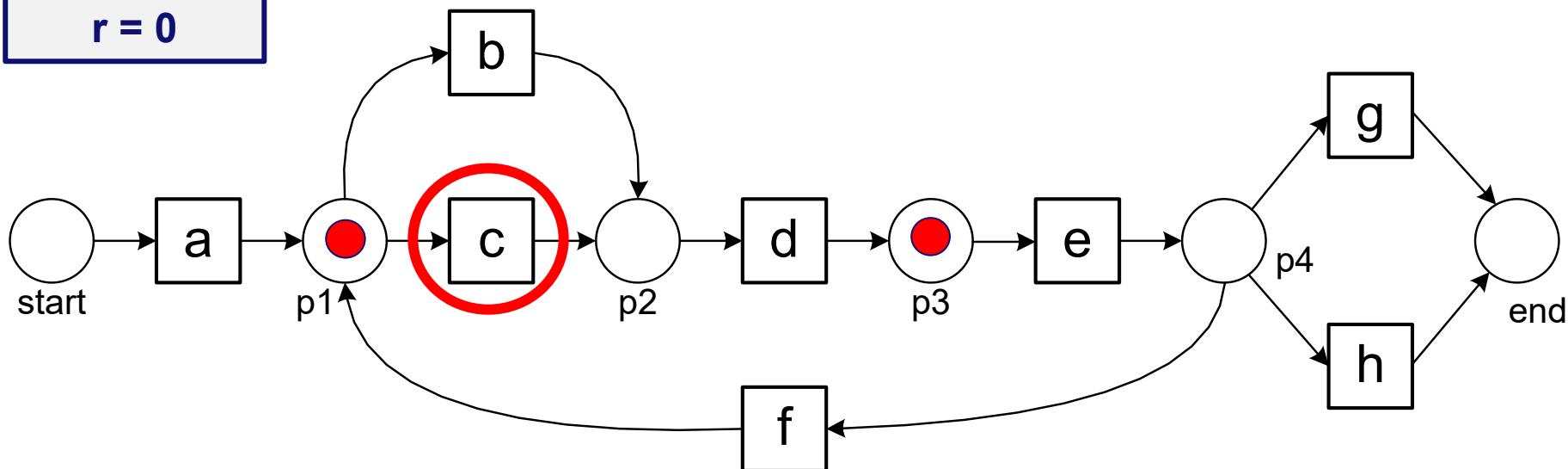
$p = 1+1 = 2$
 $c = 0+1 = 1$
 $m = 0$
 $r = 0$



Replaying

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

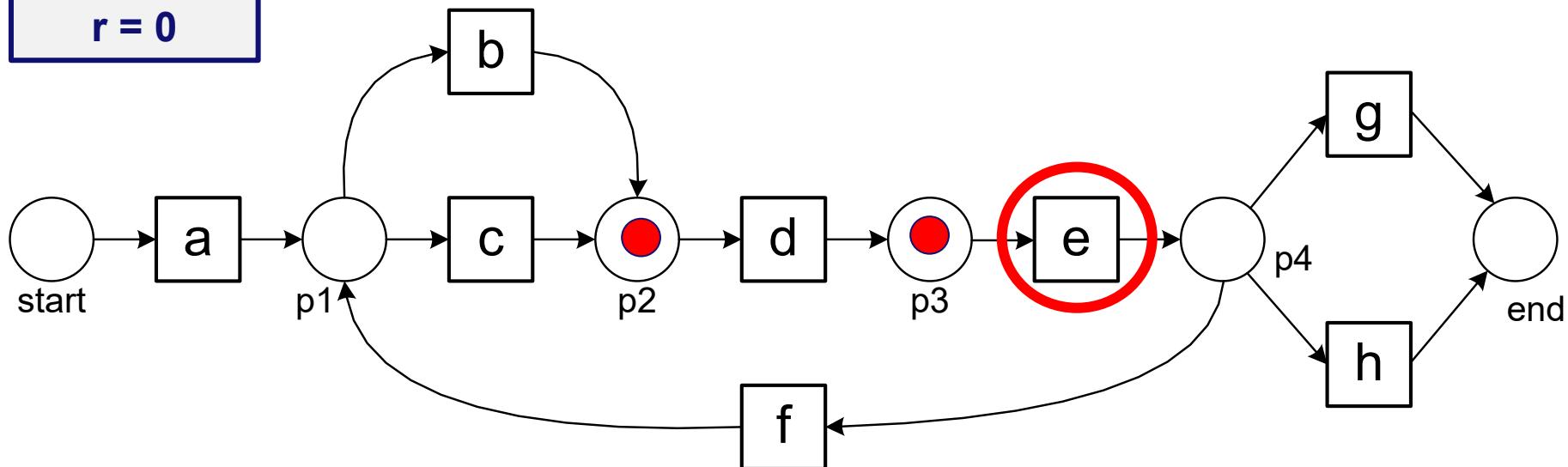
$$\begin{aligned} p &= 2+1 = 3 \\ c &= 1+1 = 2 \\ m &= 0+1 = 1 \\ r &= 0 \end{aligned}$$



Replaying

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

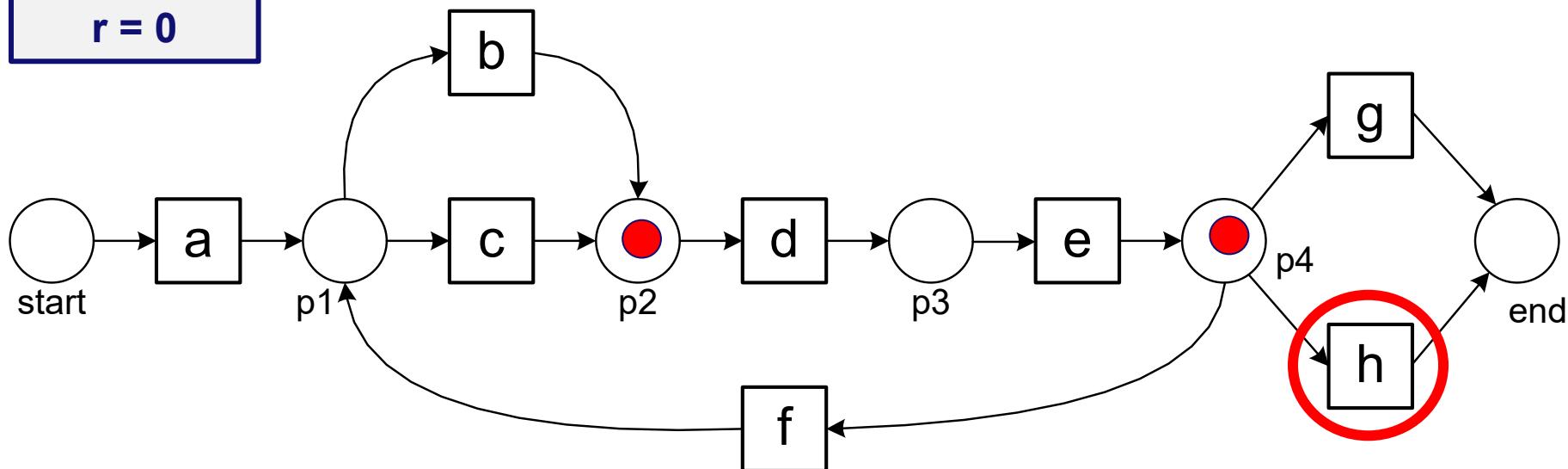
$$\begin{aligned} p &= 3+1 = 4 \\ c &= 2+1 = 3 \\ m &= 1 \\ r &= 0 \end{aligned}$$



Replaying

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

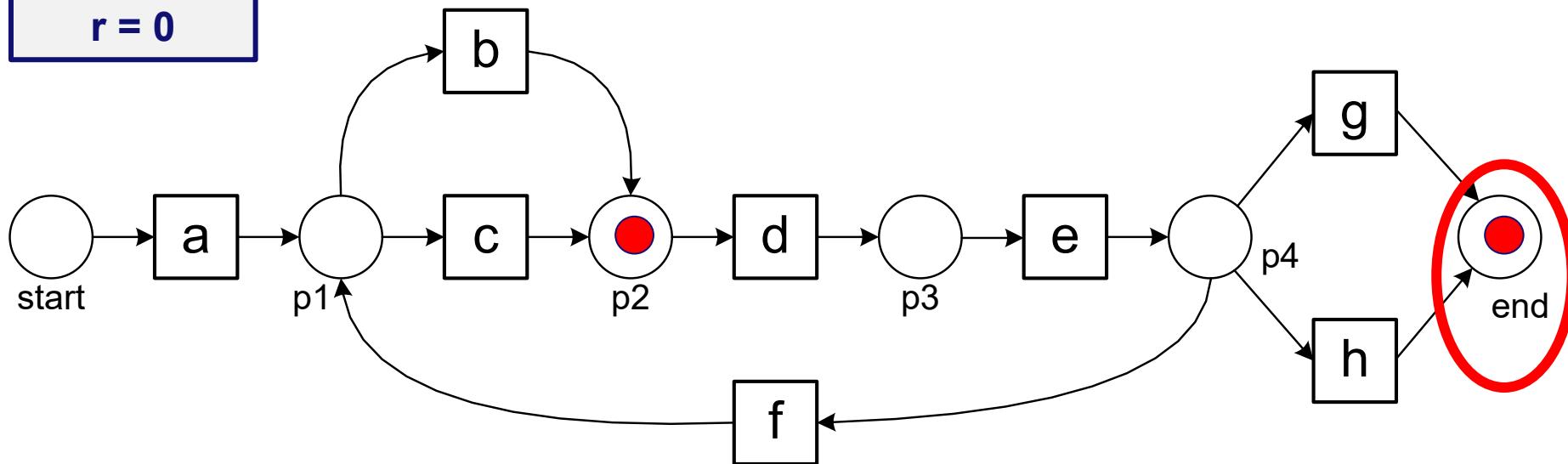
$$\begin{aligned} p &= 4+1 = 5 \\ c &= 3+1 = 4 \\ m &= 1 \\ r &= 0 \end{aligned}$$



Replaying

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

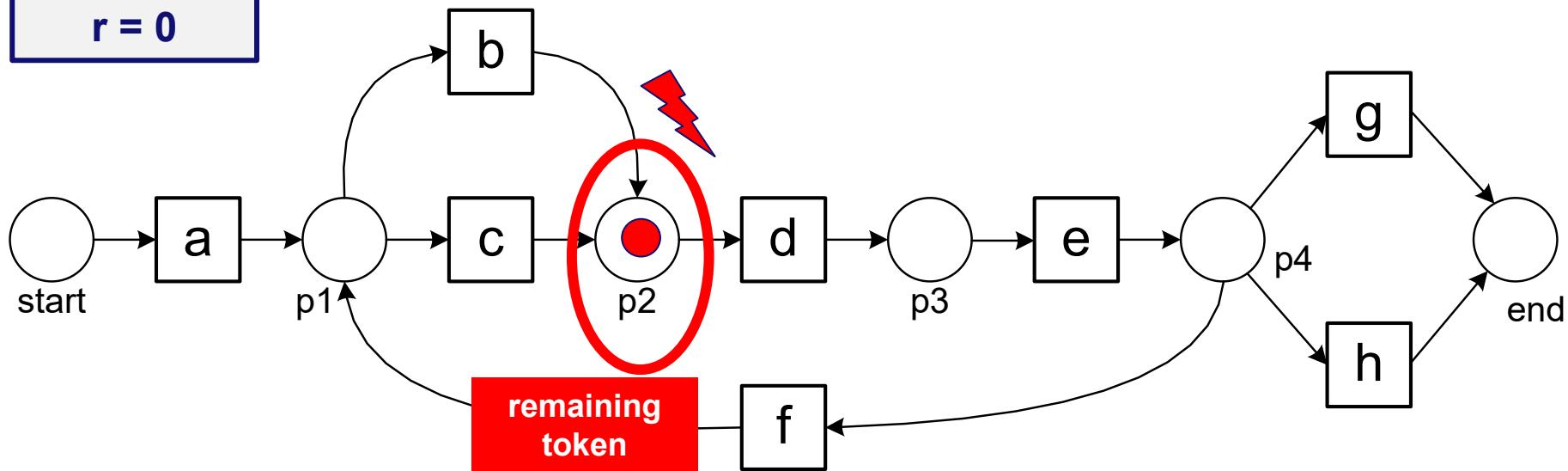
$$\begin{aligned} p &= 5+1 = 6 \\ c &= 4+1 = 5 \\ m &= 1 \\ r &= 0 \end{aligned}$$



Replaying

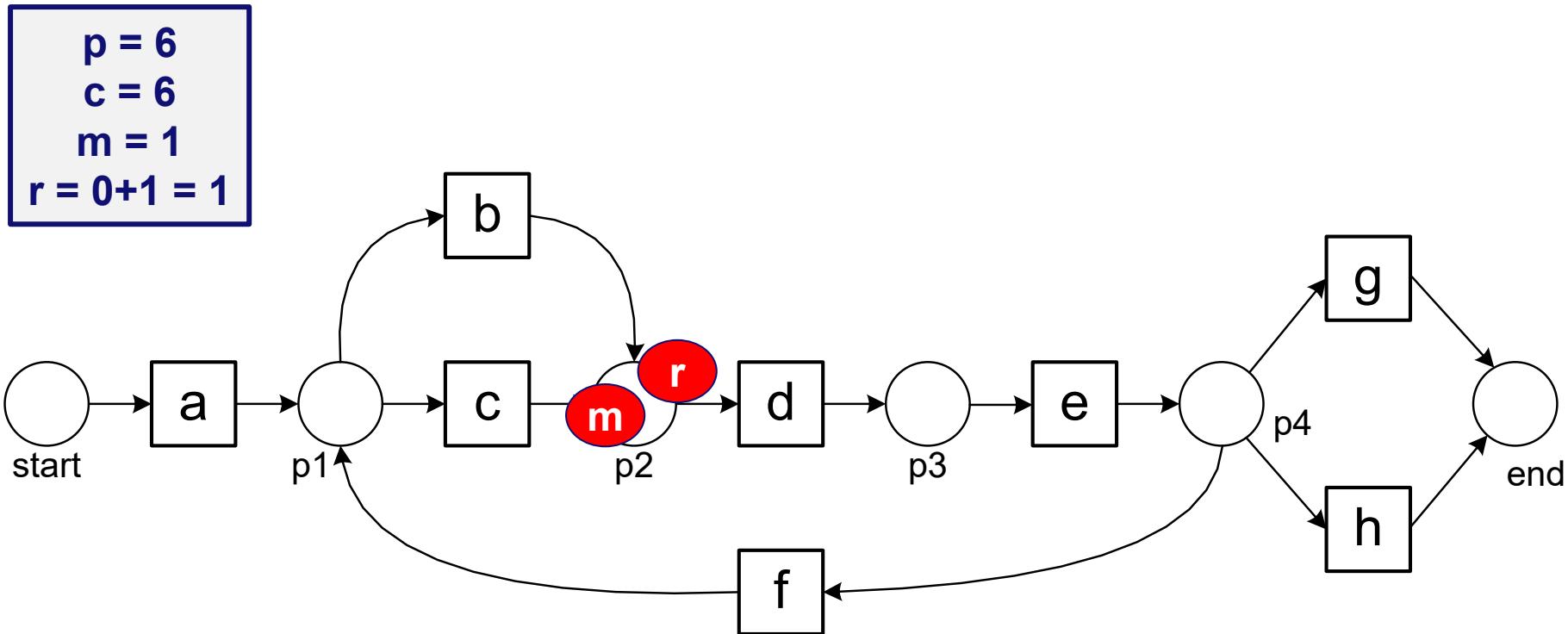
$$\sigma_3 = \langle a, d, c, e, h \rangle$$

$p = 6$
 $c = 5+1 = 6$
 $m = 1$
 $r = 0$



Replaying

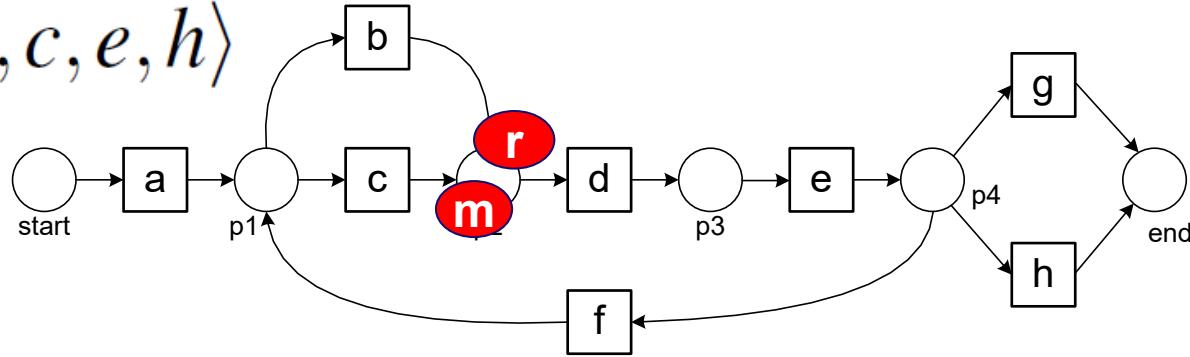
$$\sigma_3 = \langle a, d, c, e, h \rangle$$



Quantifying fitness at the trace level

p = 6
c = 6
m = 1
r = 1

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



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

Fitness at the log level

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

missing tokens

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

consumed tokens

remaining tokens

produced tokens

Looks scar
just needs t
sums of p, c, m, and r
over the multiset of
traces in de

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



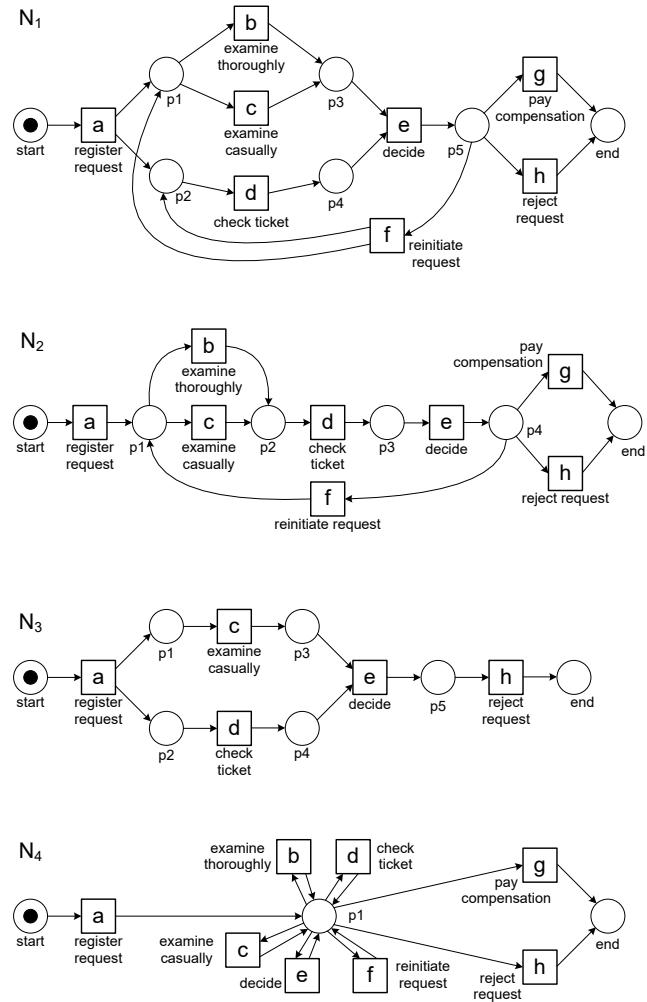
$$fitness(L, N) = \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}} \right) + \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}} \right)$$

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

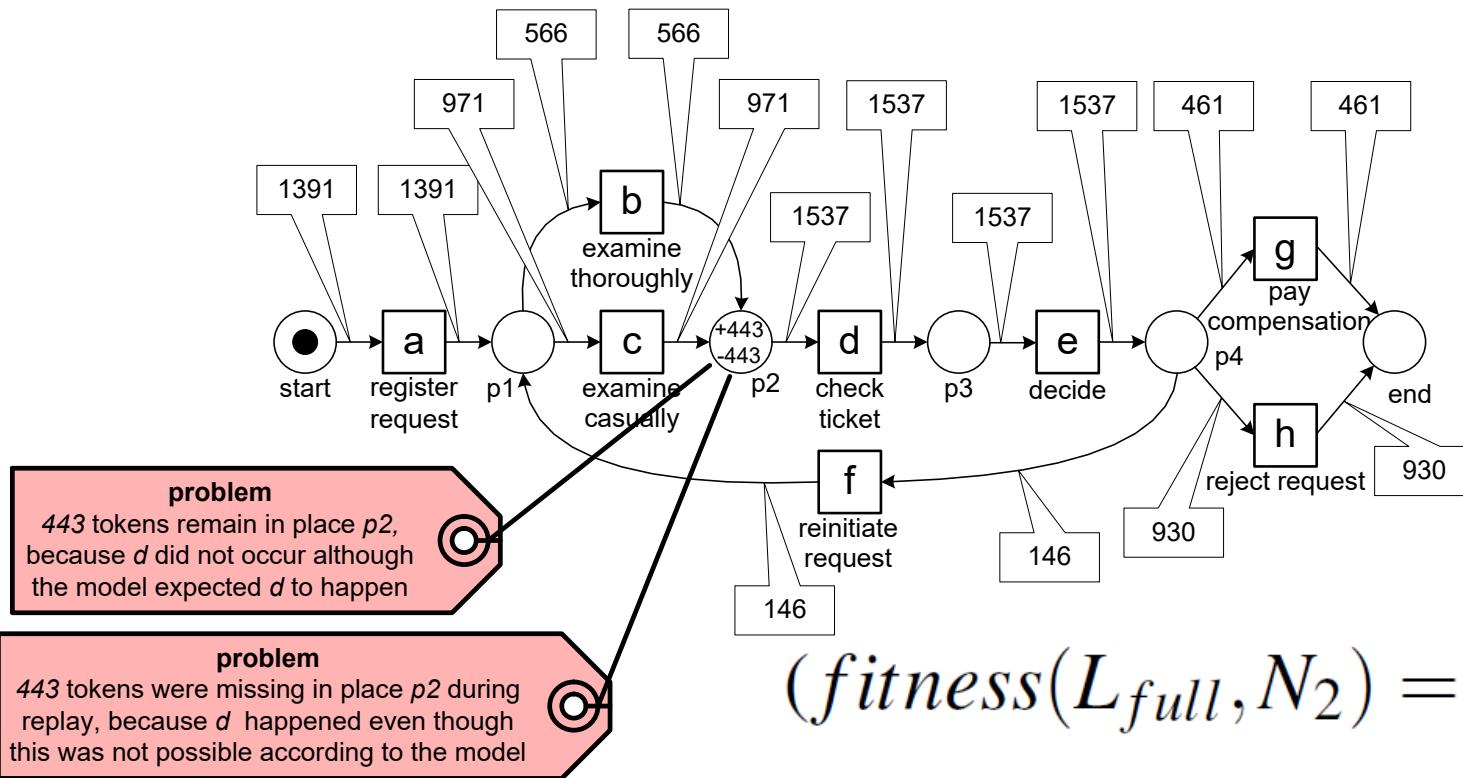
$$fitness(L_{full}, N_2) = 0.9504$$

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

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



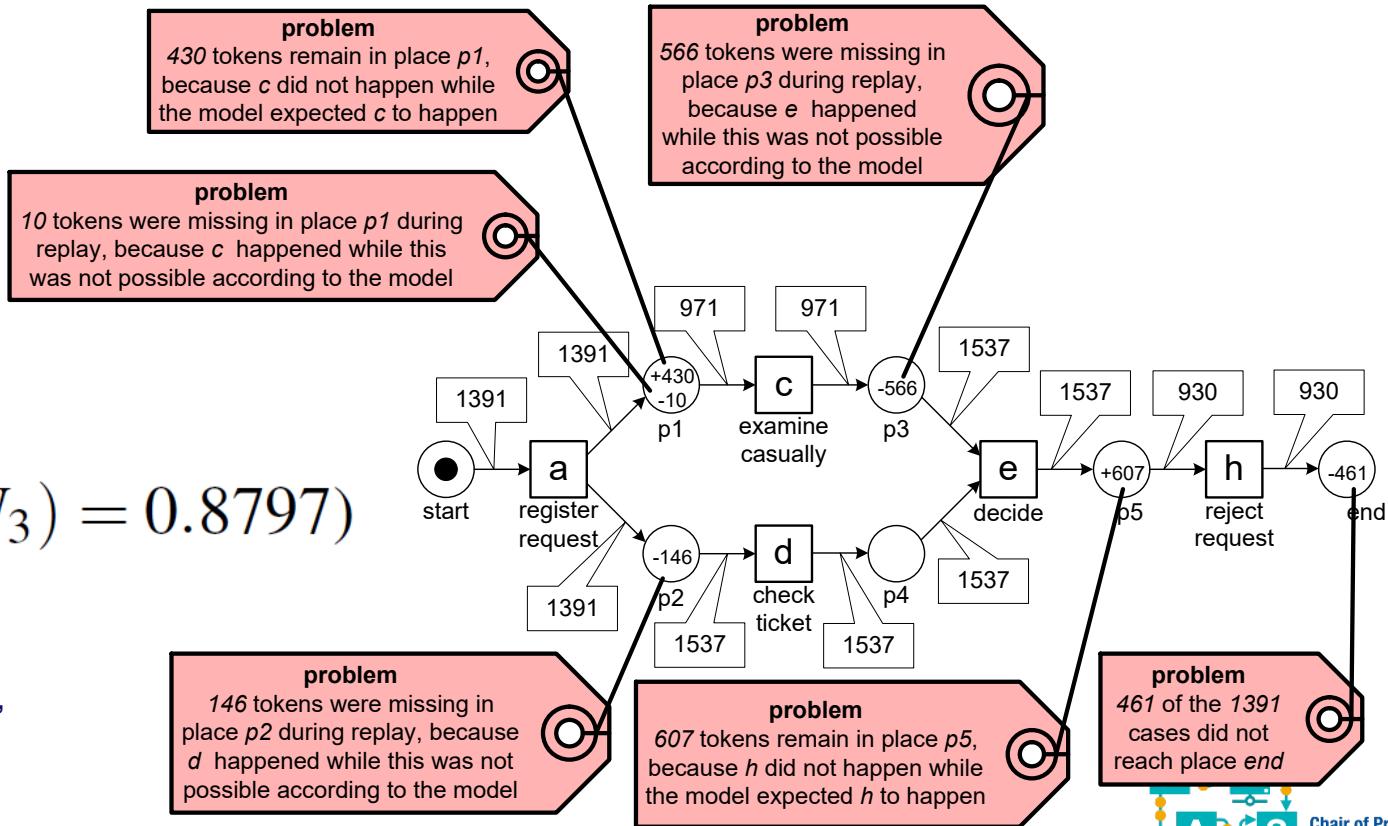
Diagnostics



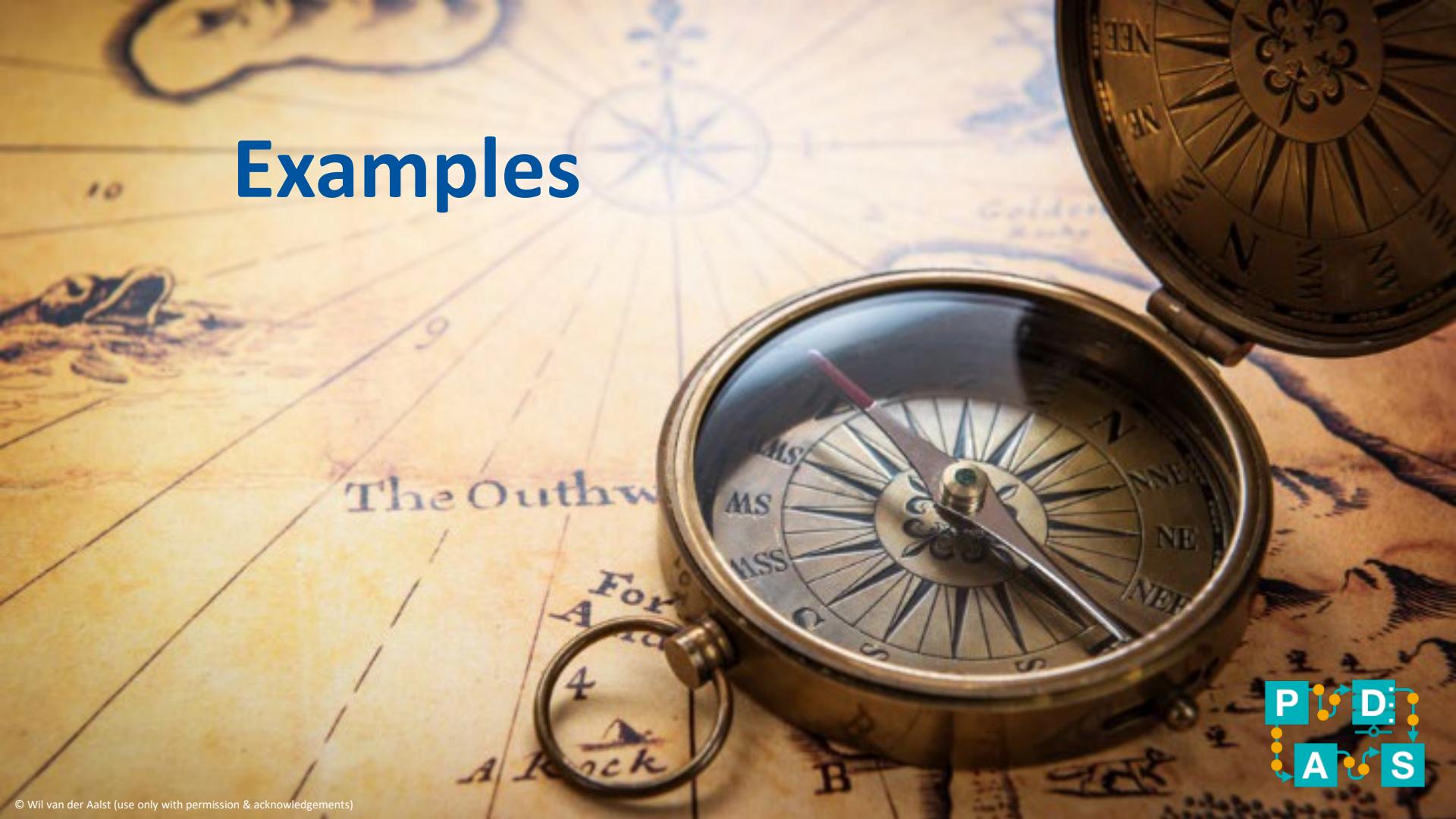
Diagnostics

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

Remark: If event log and model consider different sets of activities, this should be addressed first.
Activities in log, but not in model, are simply ignored.



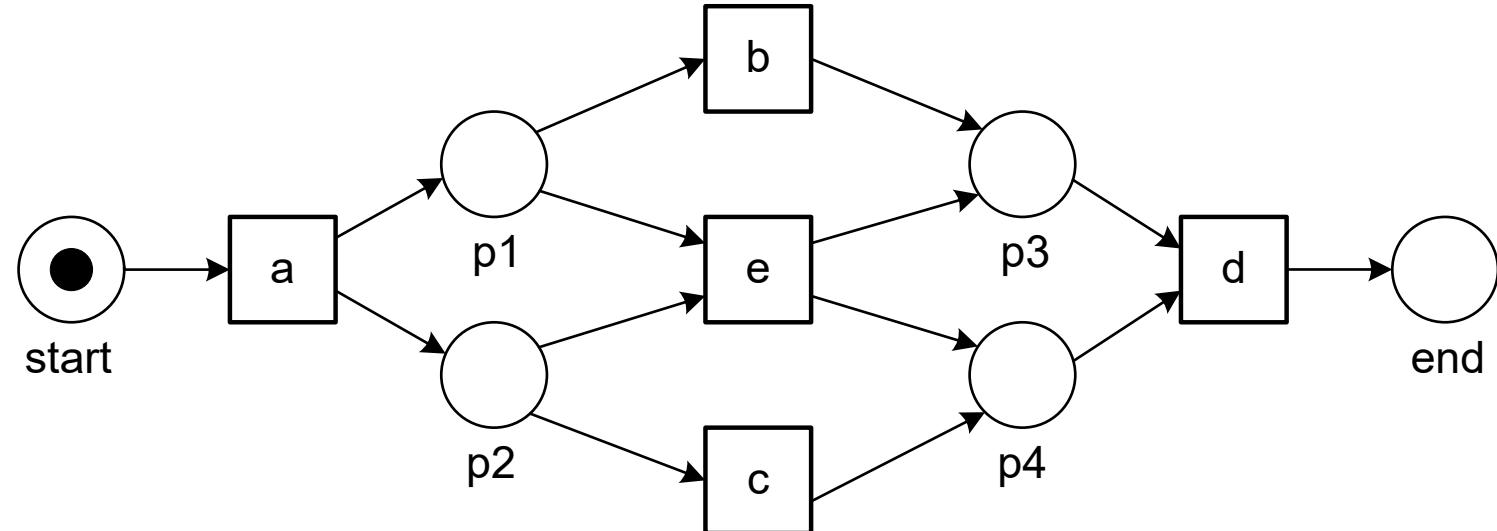
Examples



Question (may take some time)

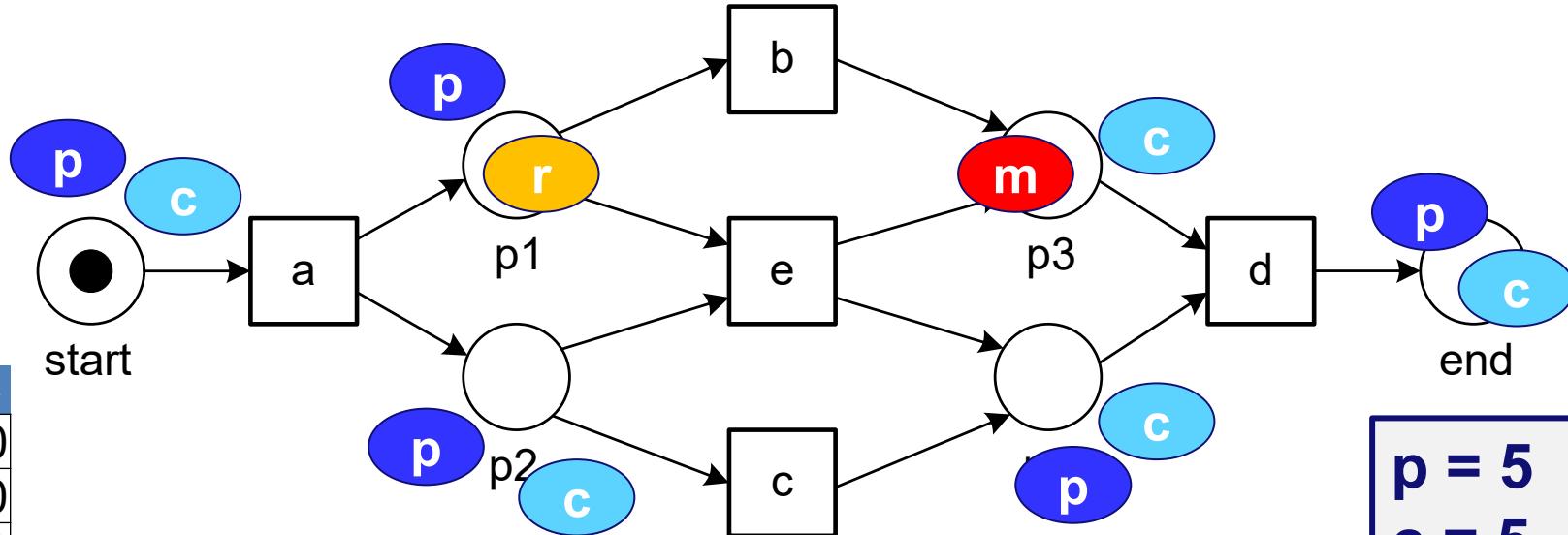
Compute fitness using missing and remaining tokens

trace	frequency
abcd	10
acbd	10
aed	10
abd	2
acd	1
ad	1
abbd	1



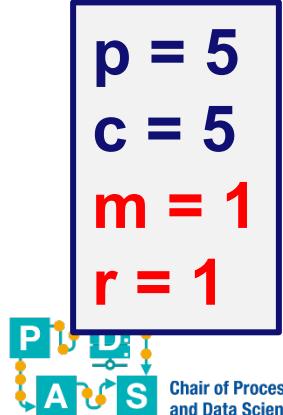
- Consider the event log containing 35 cases.
- What is the fitness?

Let us pick one trace: acd



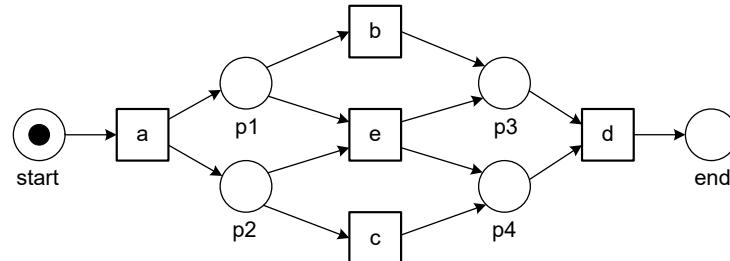
trace	frequency
abcd	10
acbd	10
aed	10
abd	2
acd	1
ad	1
abbd	1

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



Fitness = 0.9658

trace	frequency	produced tokens (p)	remaining tokens (r)	consumed tokens (c)	missing tokens (m)	produced tokens (pII)	remaining tokens (rII)	consumed tokens (cII)	missing tokens (mII)
abcd	10	6	0	6	0	60	0	60	0
acbd	10	6	0	6	0	60	0	60	0
aed	10	6	0	6	0	60	0	60	0
abd	2	5	1	5	1	10	2	10	2
acd	1	5	1	5	1	5	1	5	1
ad	1	4	2	4	2	4	2	4	2
abbd	1	6	2	6	2	6	2	6	2



205	7	205	7
sum p	sum r	sum c	sum m

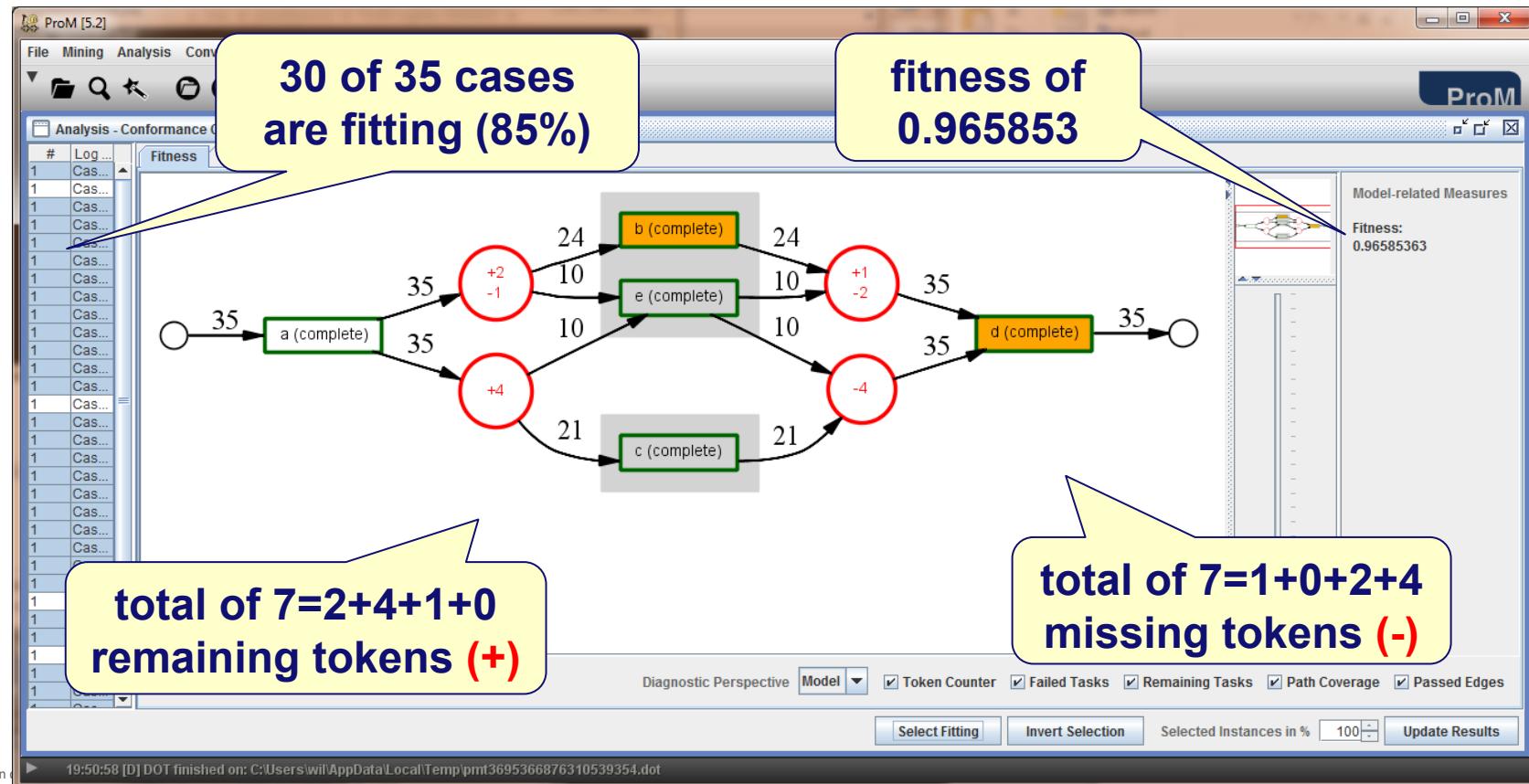
fitness	0.965853659
---------	-------------

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

ProM 5.2 output

Note that PM4Py and Celonis support variants of token-based replay.

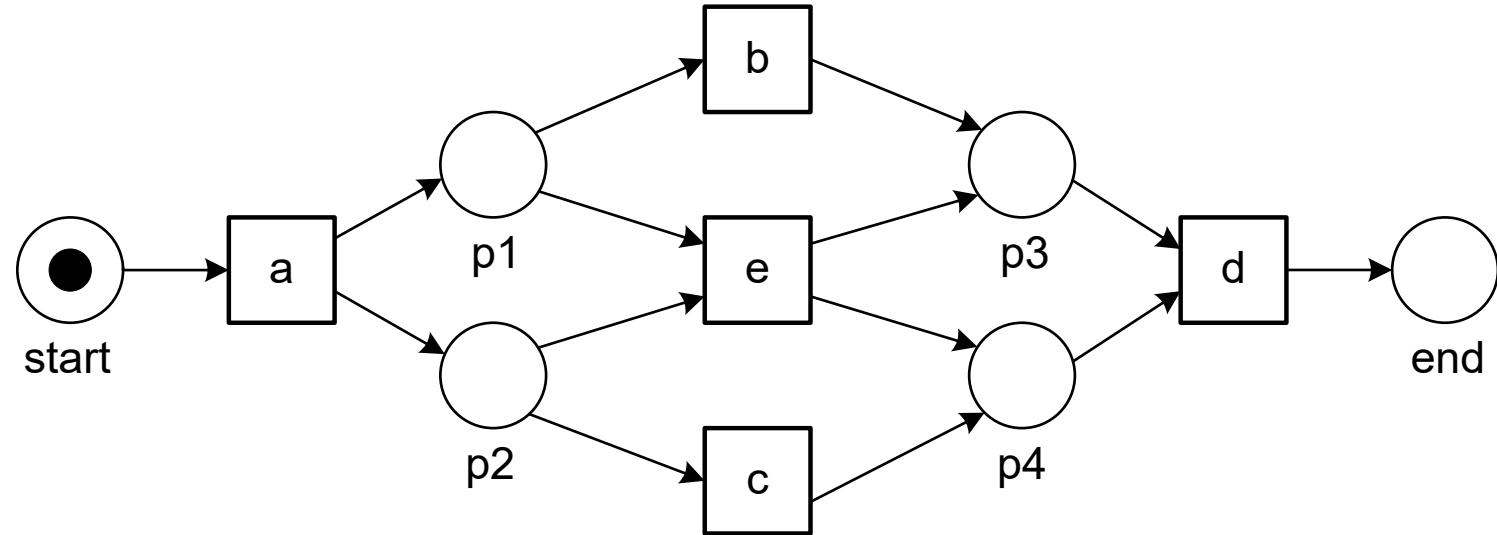
(ProM 6 only supports more advanced conformance checking techniques like alignments)



Question

Compute fitness using missing and remaining tokens

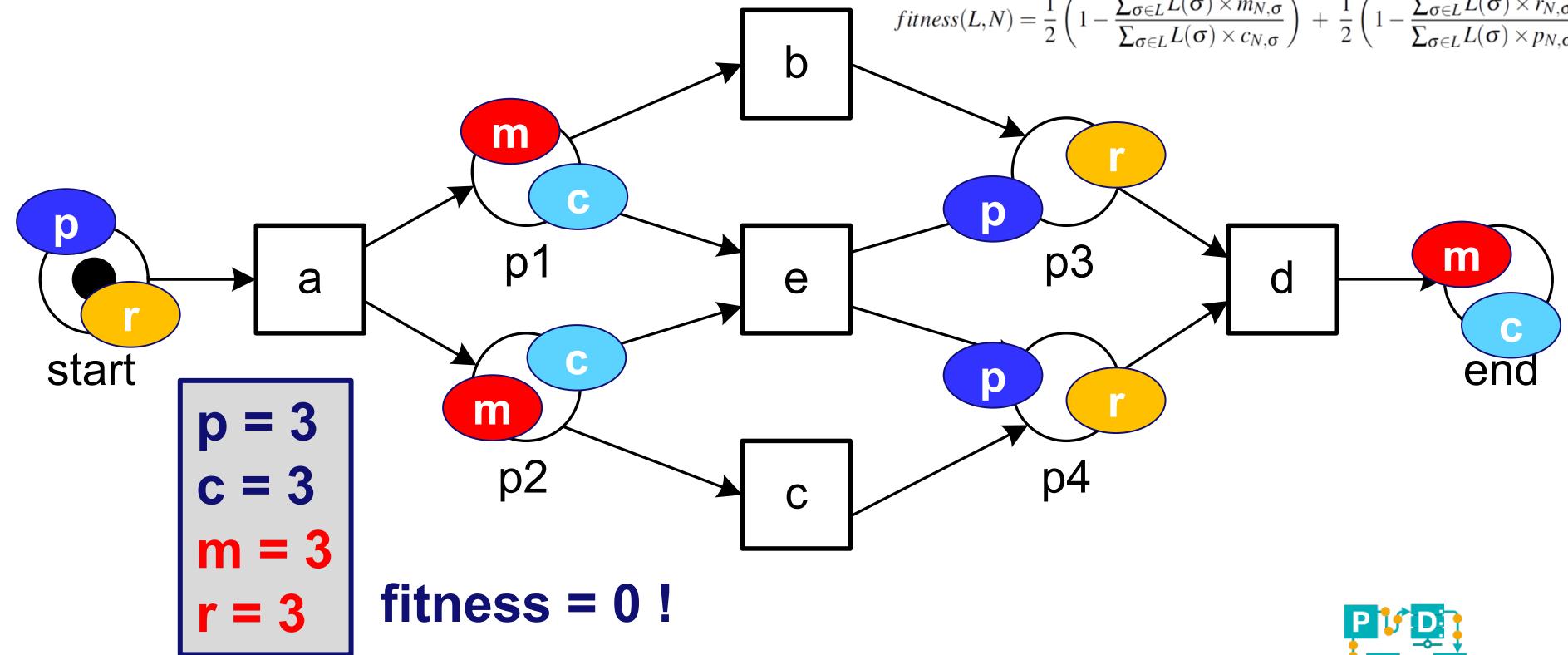
trace	frequency
e	1



- Consider the event log containing just one case: $L = [\langle e \rangle]$.
- What is the fitness (using token-based replay)?

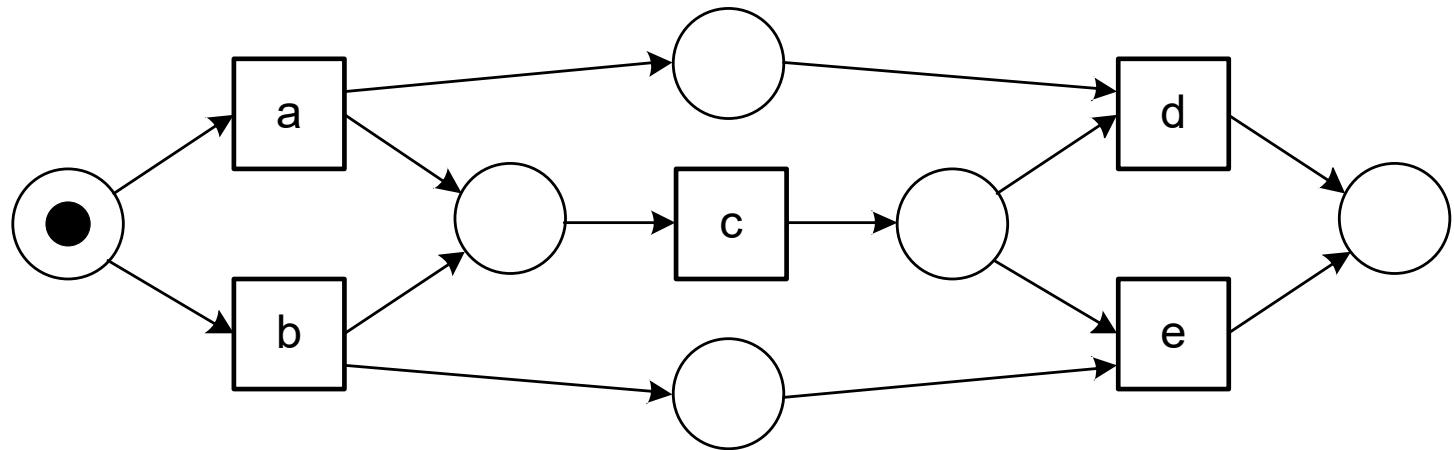
Answer obtained by replaying $\langle e \rangle$

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



Another example

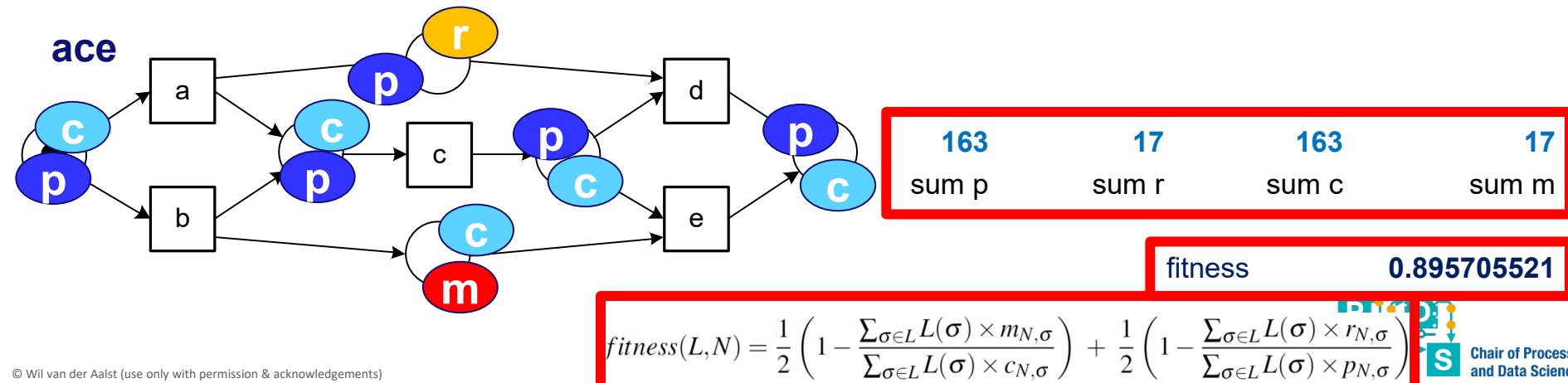
trace	frequency
acd	10
bce	10
ace	5
bcd	5
dca	1
abd	1
d	1



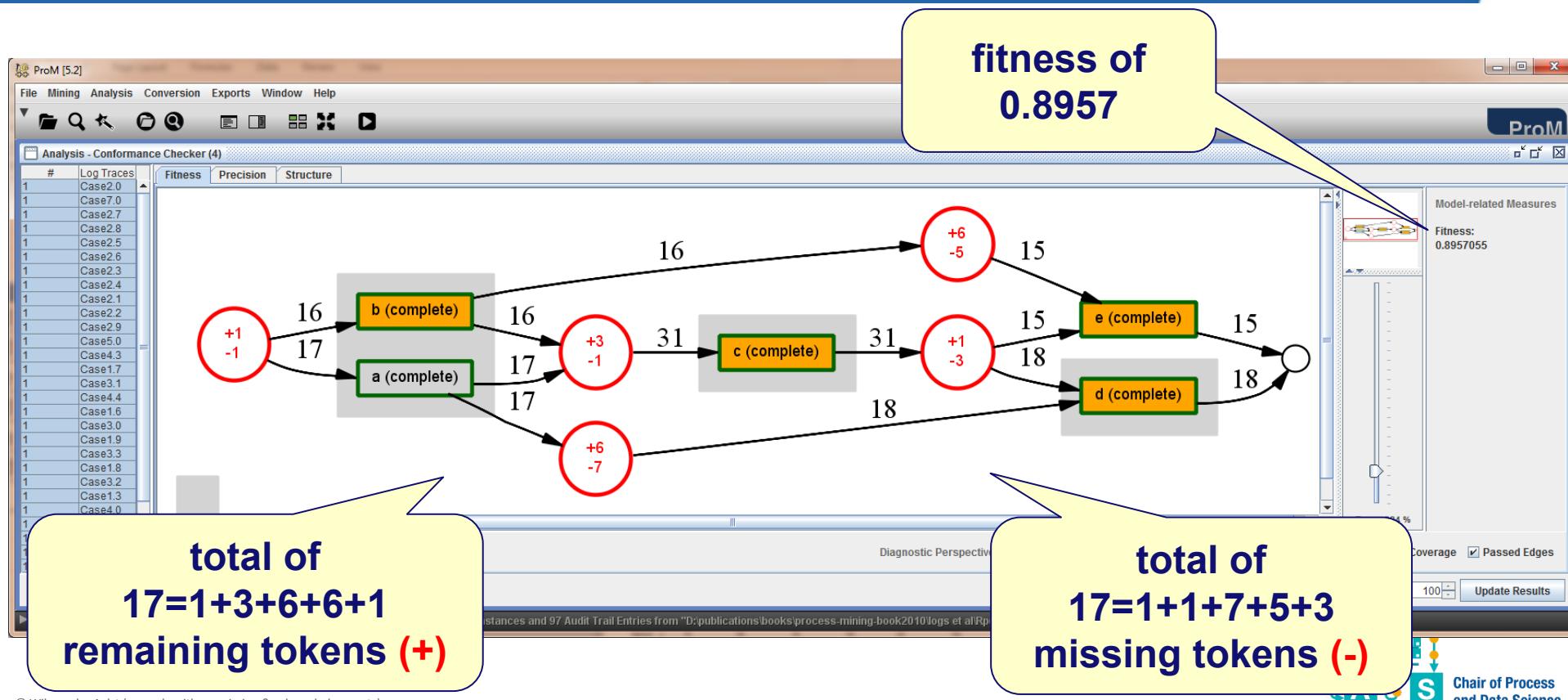
- Consider the event log containing 33 cases.
- What is the fitness?

Fitness = 0.895705521

trace	frequency	produced tokens (p)	remaining tokens (r)	consumed tokens (c)	missing tokens (m)	produced tokens (all)	remaining tokens (all)	consumed tokens (all)	missing tokens (all)
acd	10	5	0	5	0	50	0	50	0
bce	10	5	0	5	0	50	0	50	0
ace	5	5	1	5	1	25	5	25	5
bcd	5	5	1	5	1	25	5	25	5
dca	1	5	3	5	3	5	3	5	3
abd	1	6	3	5	2	6	3	5	2
d	1	2	1	3	2	2	1	3	2



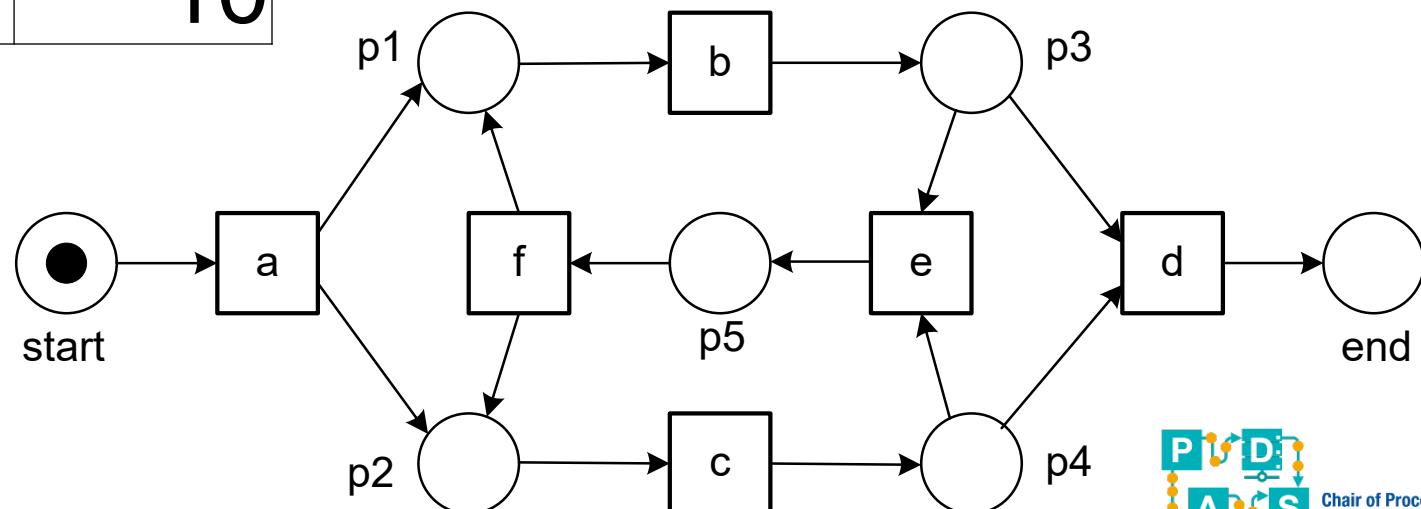
ProM 5.2 diagnostics



Another example

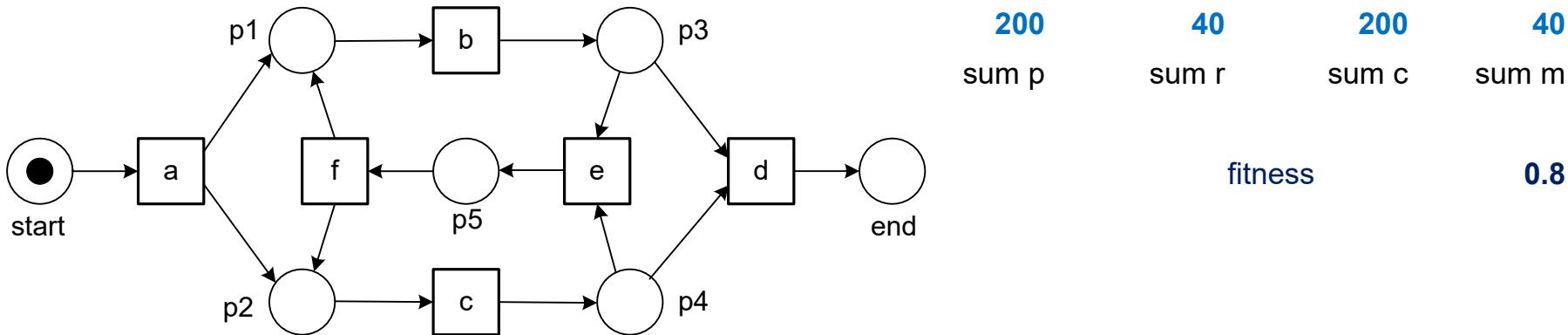
trace	frequency
abefcd	10
abbefcccd	10

- Consider the event log containing 20 cases.
- What is the fitness?



Fitness = 0.8

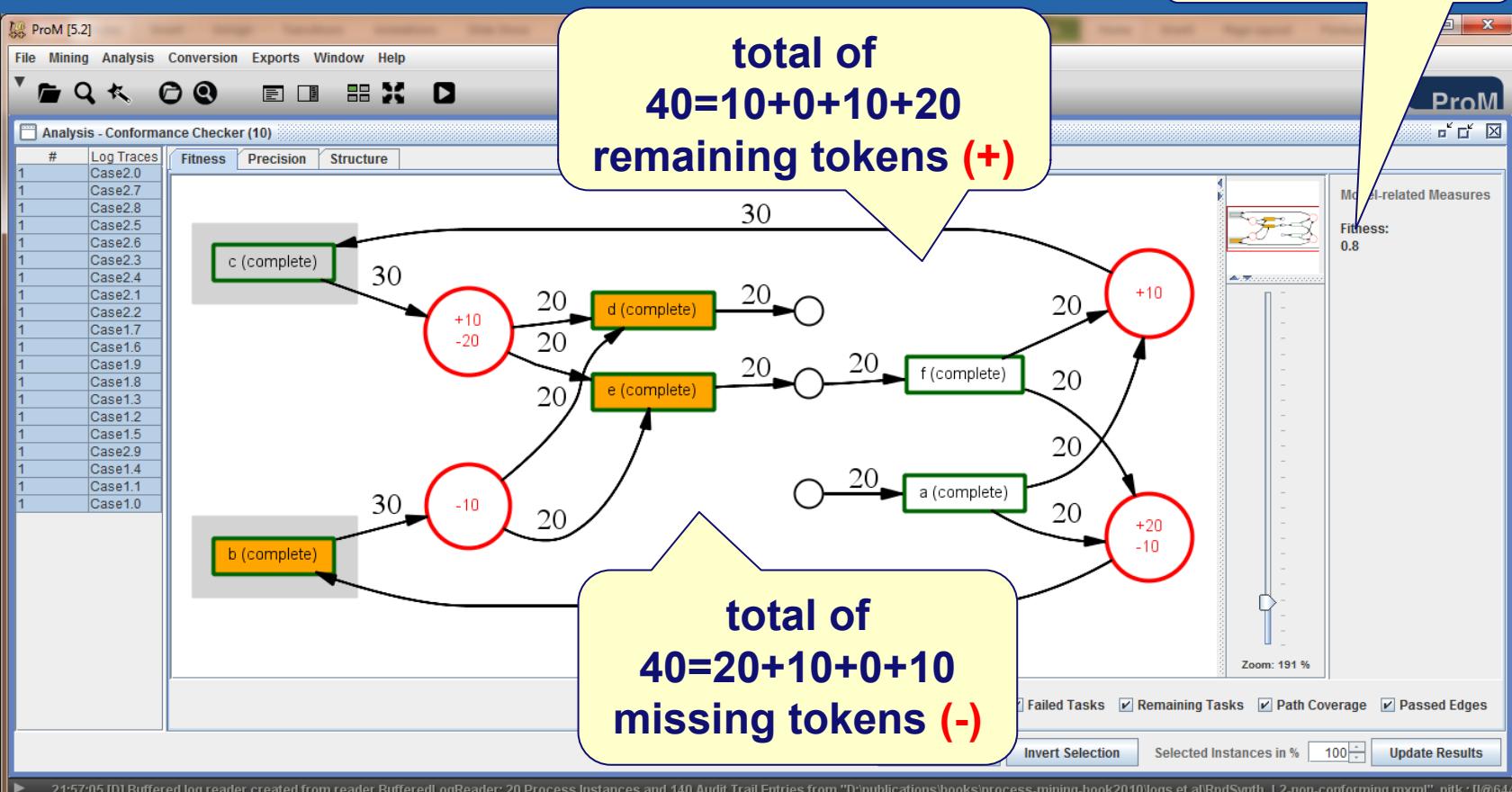
trace	frequency	produced tokens (p)	remaining tokens (r)	consumed tokens (c)	missing tokens (m)	produced tokens (all)	remaining tokens (all)	consumed tokens (all)	missing tokens (all)
abefcd	10	9	2	9	2	90	20	90	20
abbefcccd	10	11	2	11	2	110	20	110	20



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

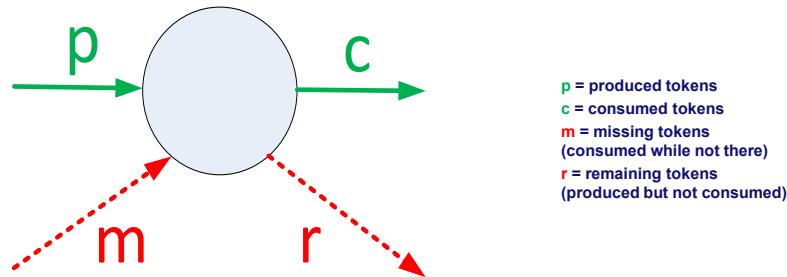
ProM 5.2 diagnostics

fitness of 0.8



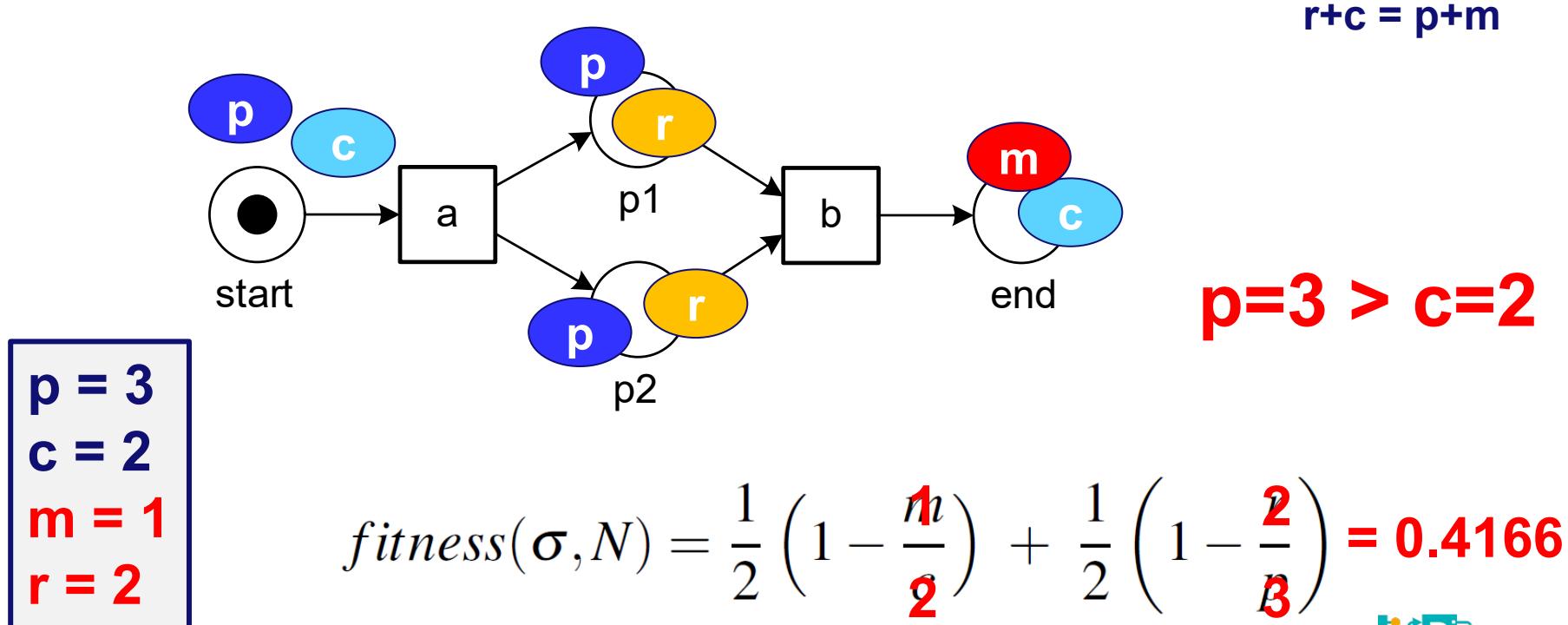
Thus far $c=p$ and $m=r$. Always?

- Provide, if possible, a log and model such that $c \neq p$ and $m \neq r$ at end.

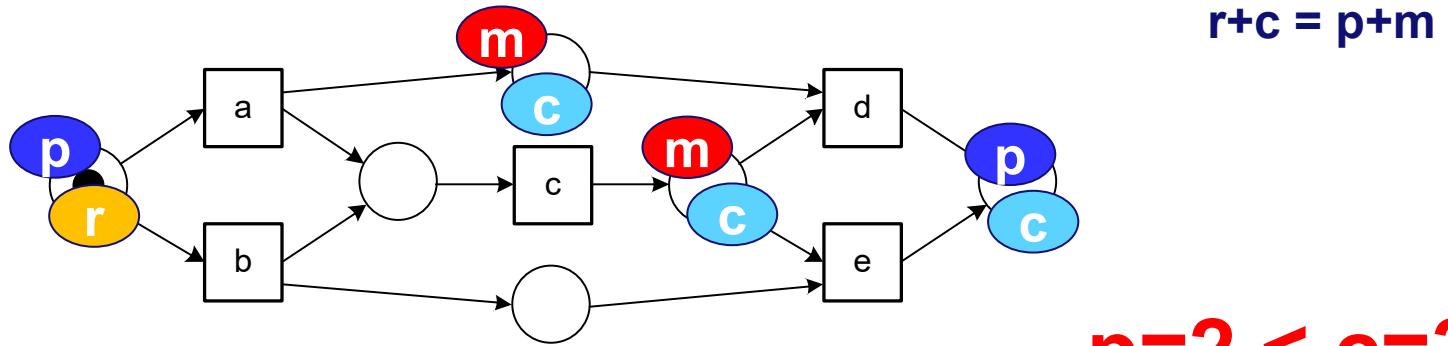


Hint: Recall that $r = p+m-c$ (i.e., $p+m=c+r$) at end.

Example: Model below and event log [$\langle a \rangle$]



Example: Model below and event log [$\langle d \rangle$]



$$\begin{aligned} p &= 2 \\ c &= 3 \\ m &= 2 \\ r &= 1 \end{aligned}$$

$$fitness(\sigma, N) = \frac{1}{2} \left(1 - \frac{2}{3} \right) + \frac{1}{2} \left(1 - \frac{1}{2} \right) = 0.4166$$

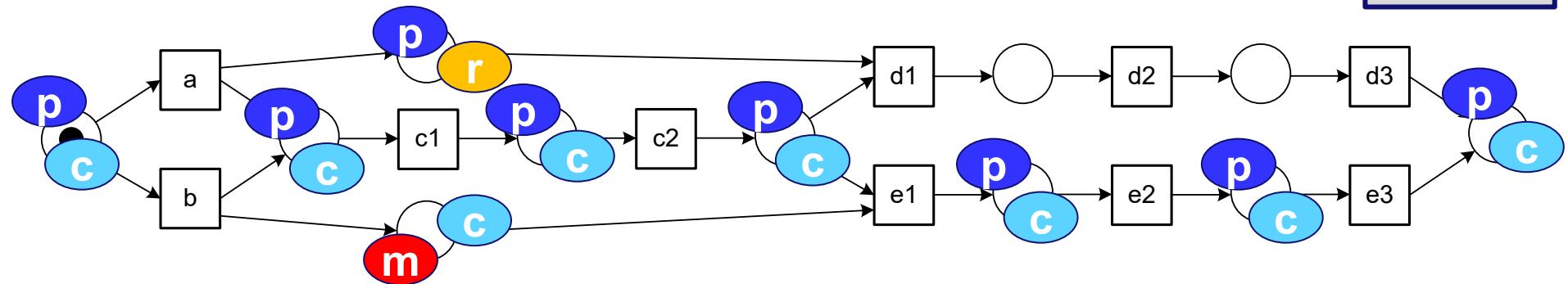
Limitations

- Basic replay approach assumes **visible & uniquely labeled transitions**.
- ProM implementation uses **heuristics** to deal with silent transitions and multiple transitions having the same label.
- Conformance values sometimes **too optimistic** due to "token flooding".
- Local decision making may lead to misleading results.

Local decision making is not enough ...

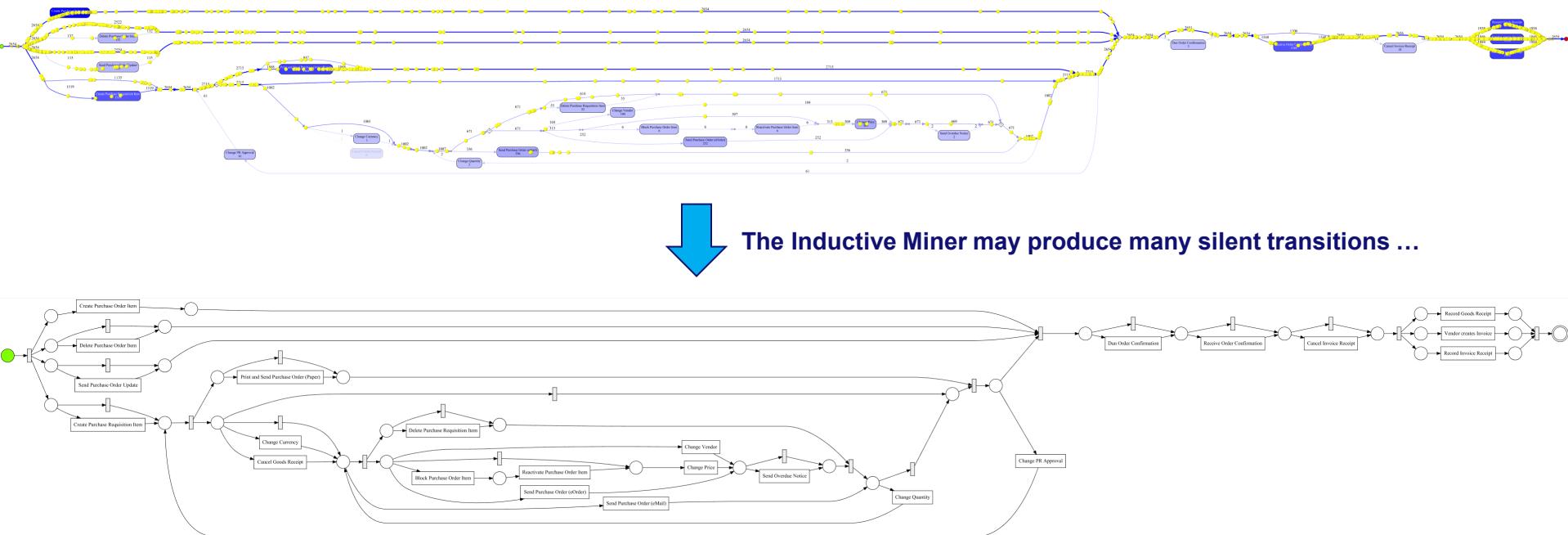
$p = 8$
 $c = 8$
 $m = 1$
 $r = 1$
 $f = 0.875$

$\langle a, c1, c2, e1, e2, e3 \rangle$

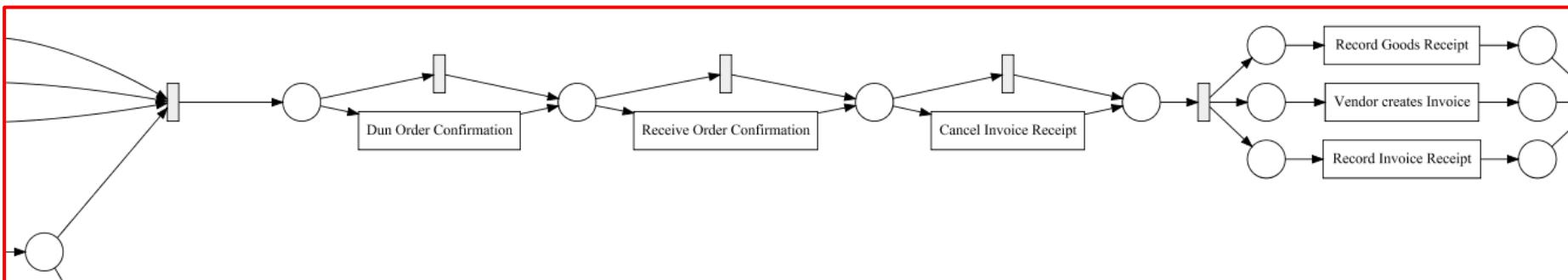
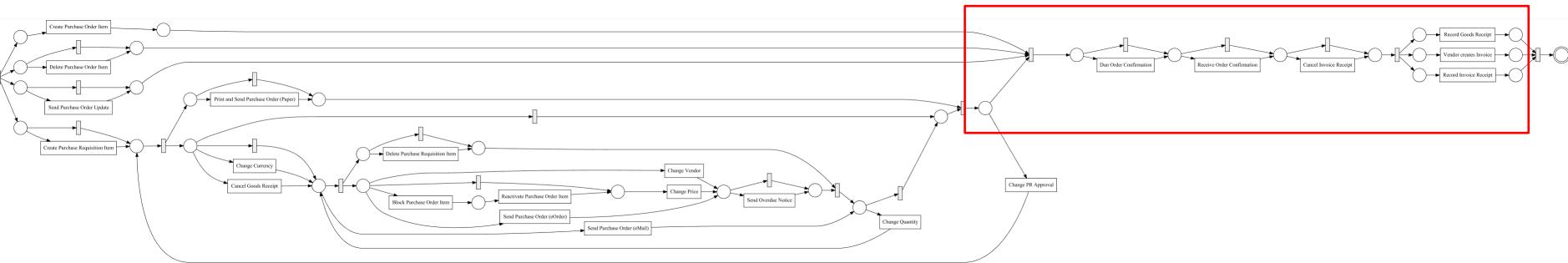


- Replay technique does **not** provide a corresponding path through the model (vital for conformance/performance analysis and other diagnostics).
- We would like to see the "closest path", i.e., $\langle b, c1, c2, e1, e2, e3 \rangle$.

Challenge: Silent and duplicate transitions

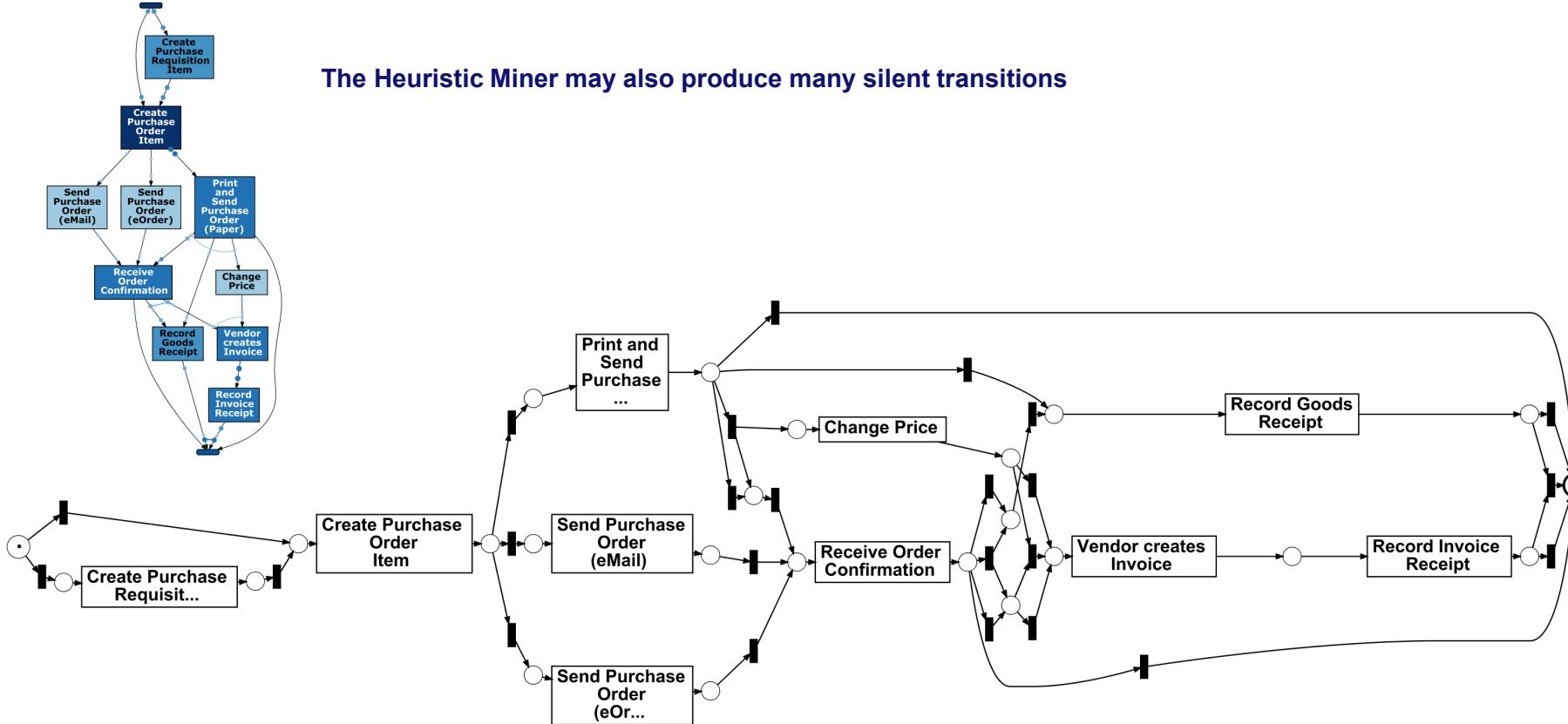


Challenge: Silent and duplicate transitions



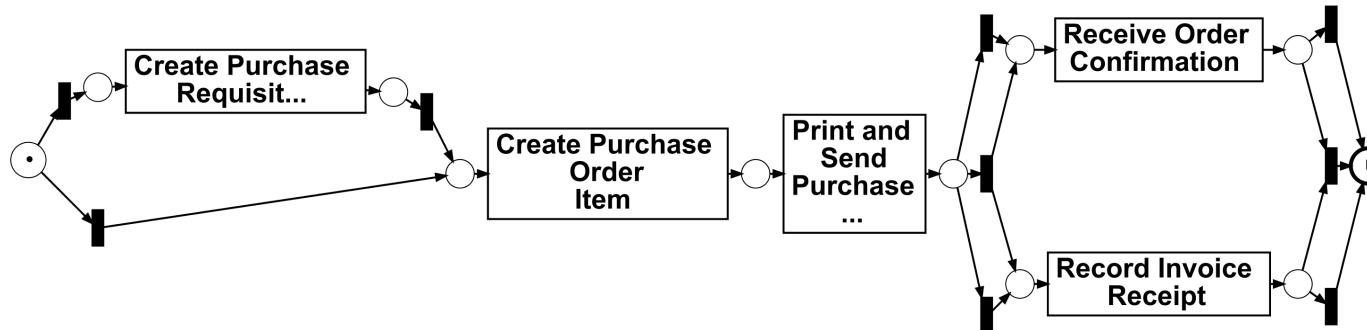
Challenge: Silent and duplicate transitions

The Heuristic Miner may also produce many silent transitions

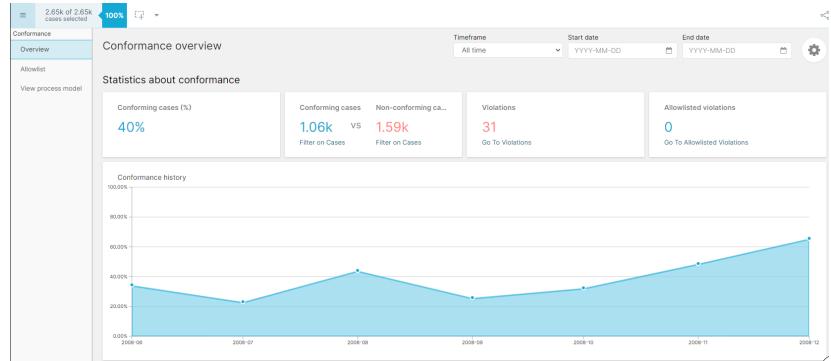
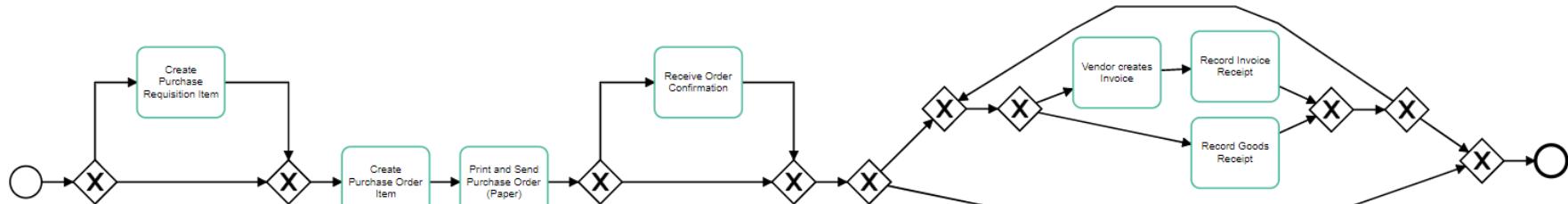


Challenge: Silent and duplicate transitions

- Some form of **state-space exploration** is needed to decide which silent transitions to fire.
- If a trace is **non-fitting**, the set of possibilities grows further.



Celonis's classical conformance checker uses a variant of token-based replay



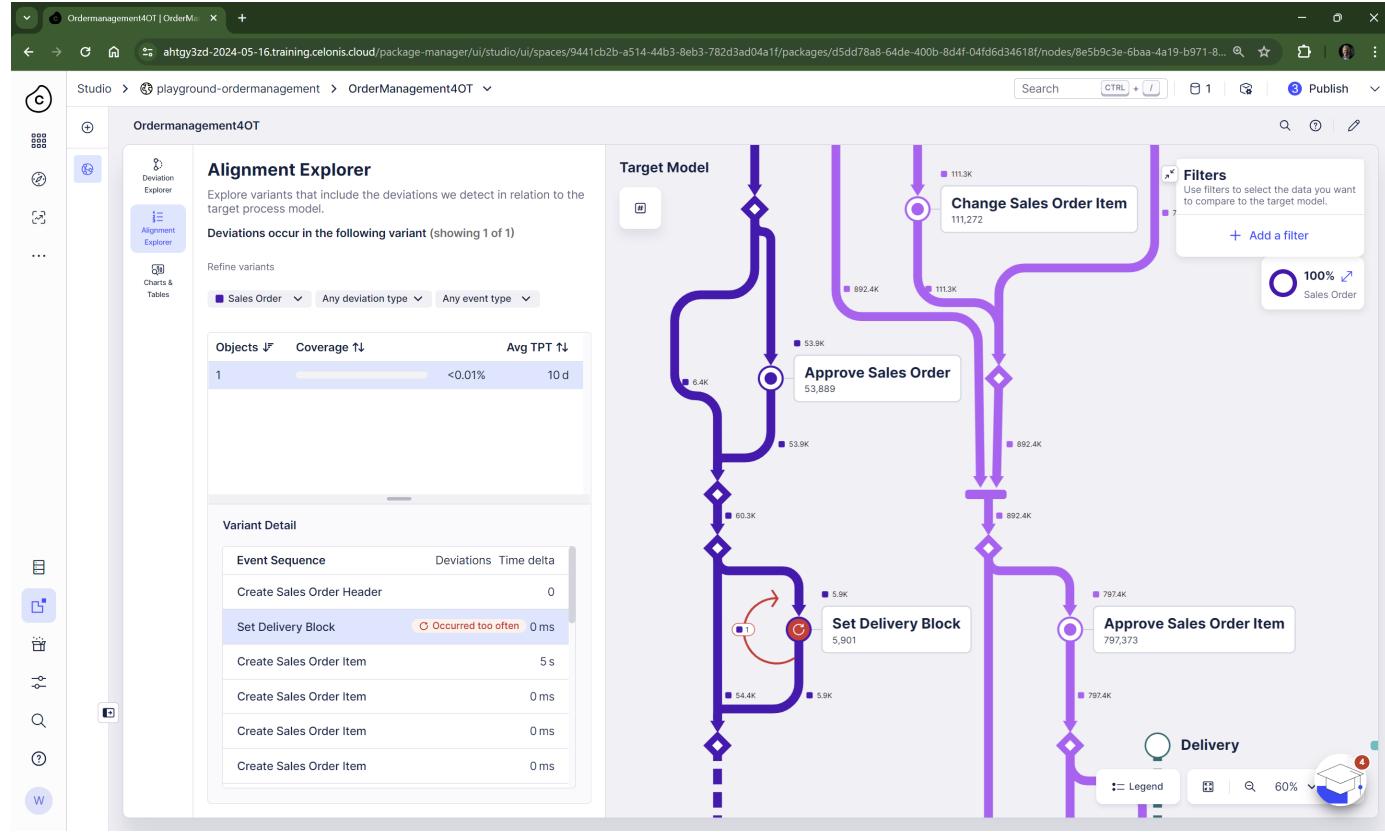
Violations

23% of cases	Change Price is an undesired activity View cases in... Effect on throughput time Effect on steps per case 16 Days longer + 2.6 Steps per case
14% of cases	Create Purchase Order Item is followed by Receive Order Confirmation View cases in... Effect on throughput time Effect on steps per case 9 Days longer + 1.6 Steps per case
13% of cases	Send Purchase Order (eMail) is an undesired activity View cases in... Effect on throughput time Effect on steps per case 4 Days longer + 0.8 Steps per case



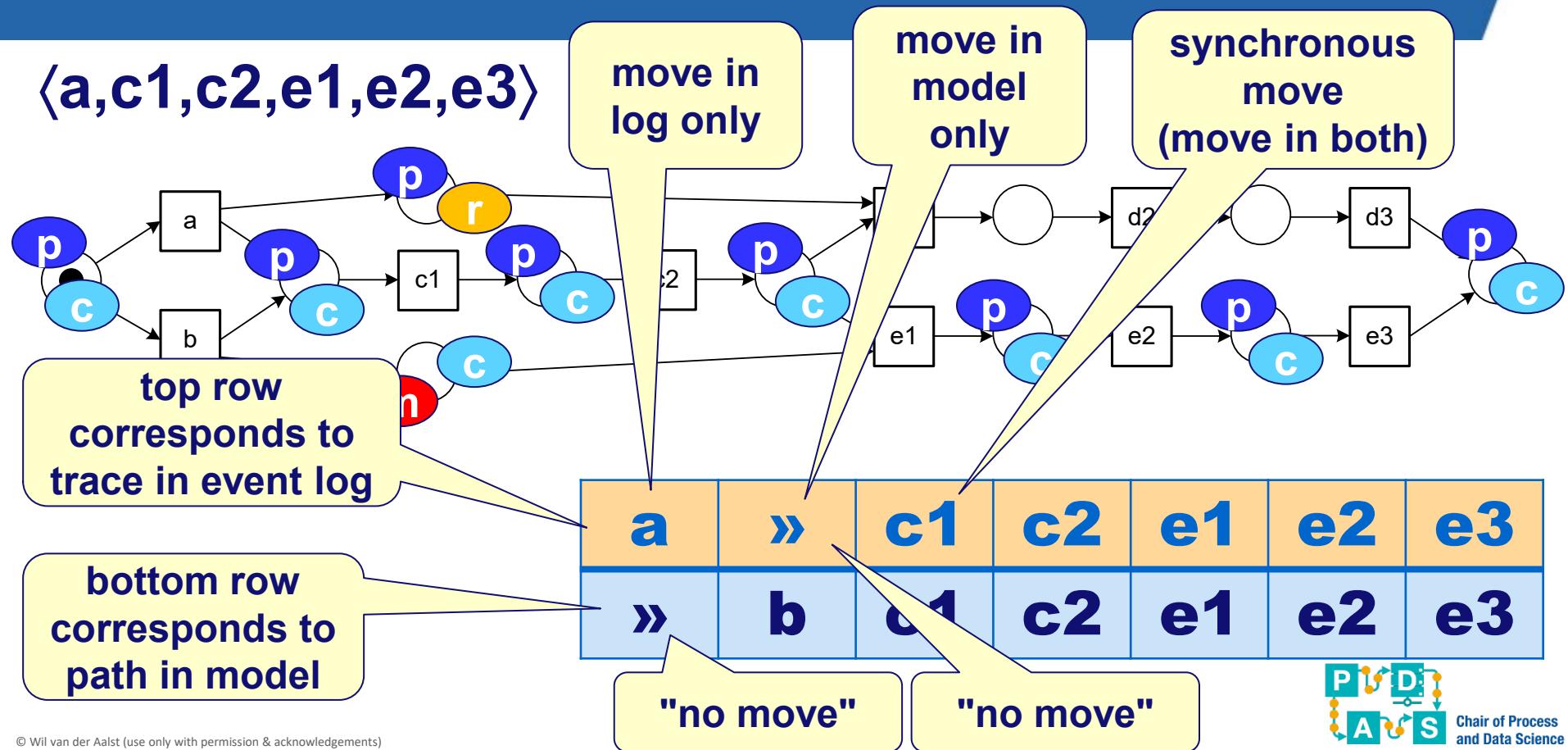
Chair of Process
and Data Science

Newer capabilities of Celonis (e.g. PAM) use alignments



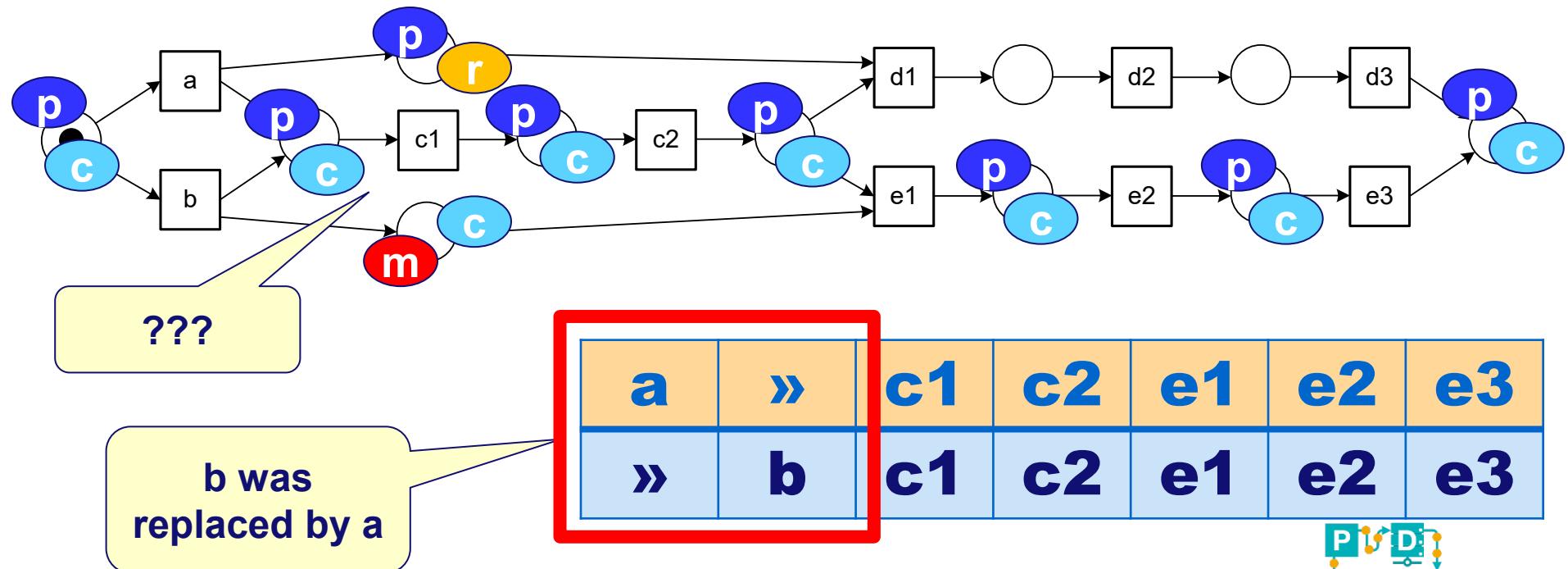
Next: alignments

$\langle a, c1, c2, e1, e2, e3 \rangle$

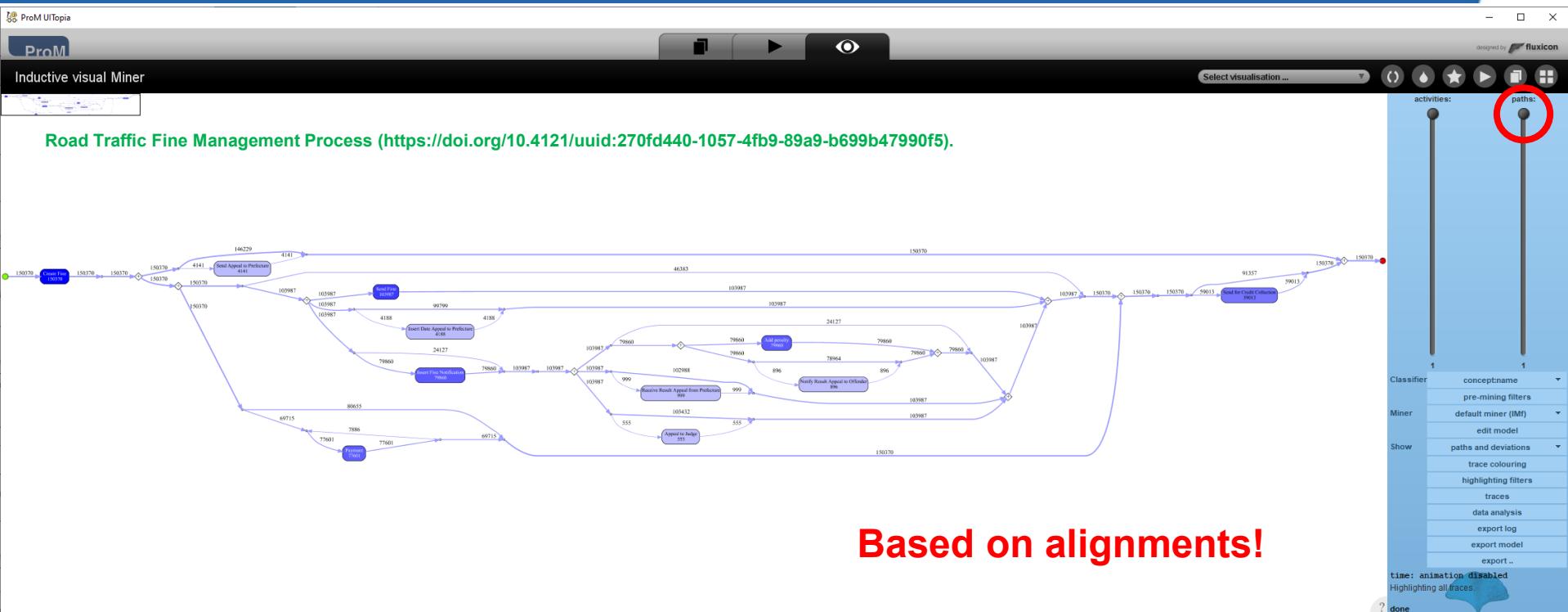


Alignments provide better diagnostics

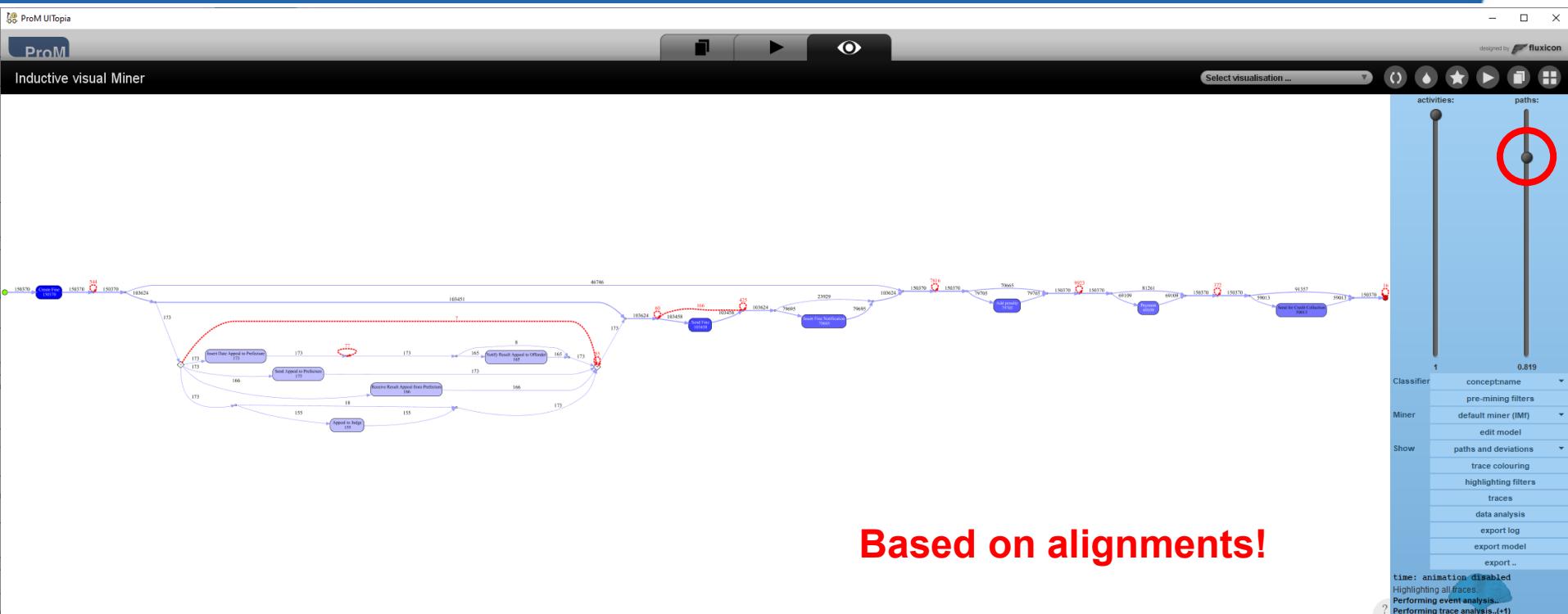
$\langle a, c1, c2, e1, e2, e3 \rangle$



Example: Inductive Visual Miner

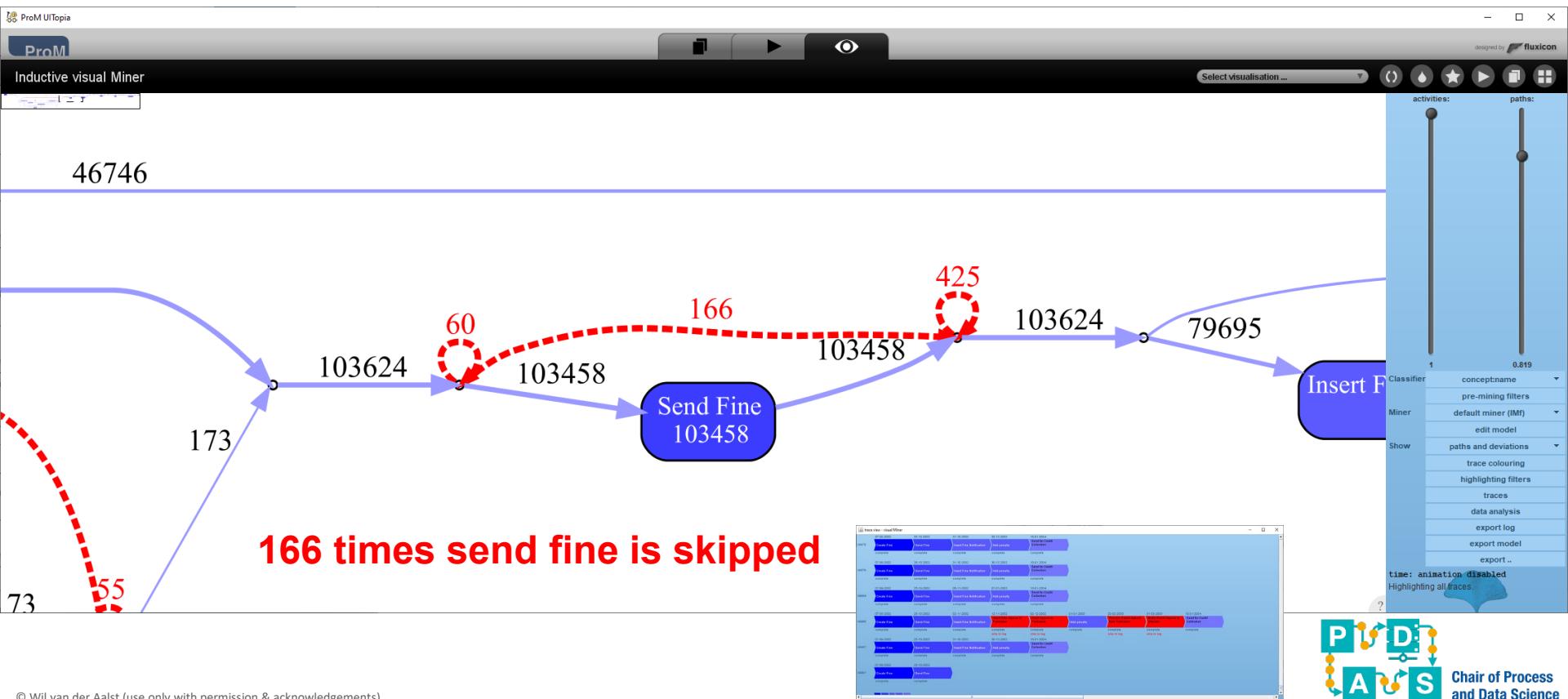


Example: Inductive Visual Miner

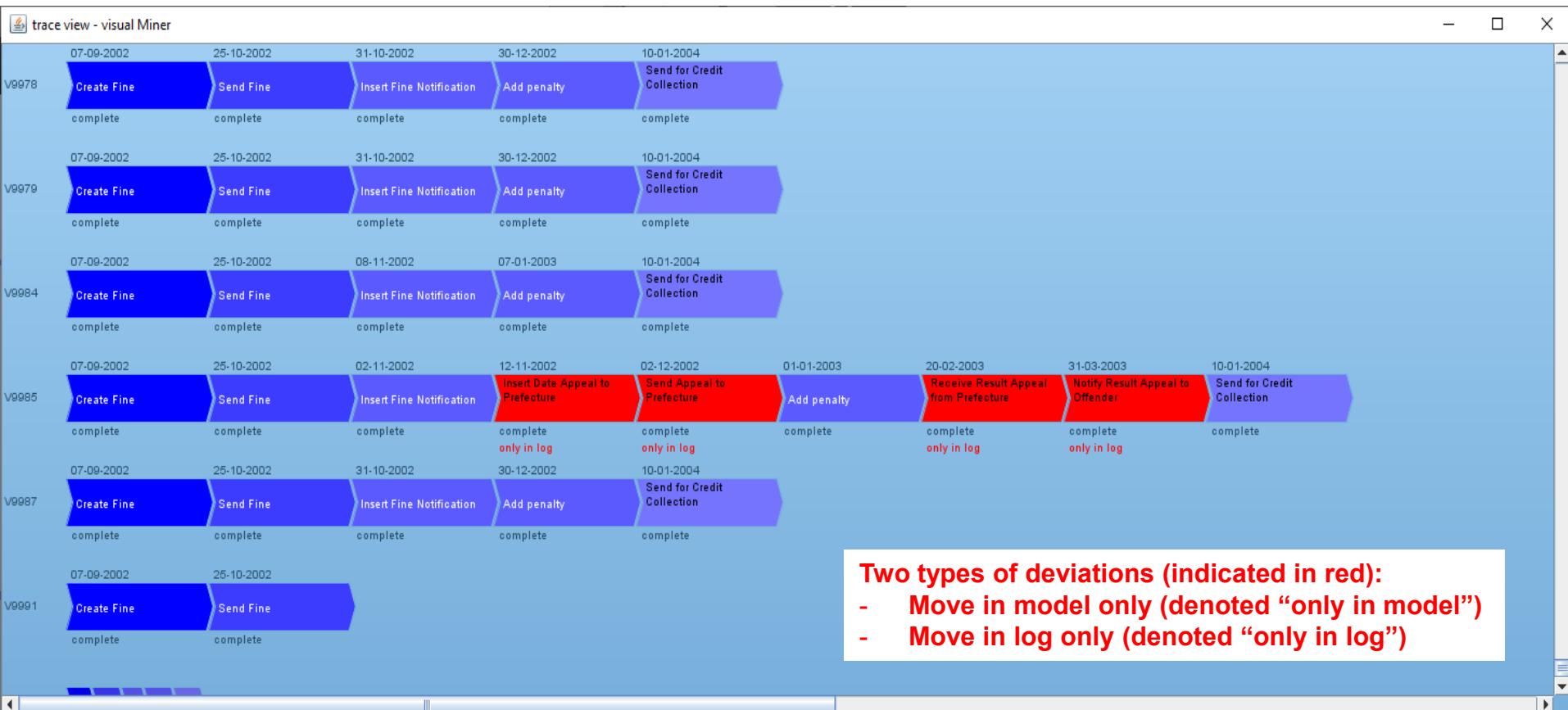


Based on alignments!

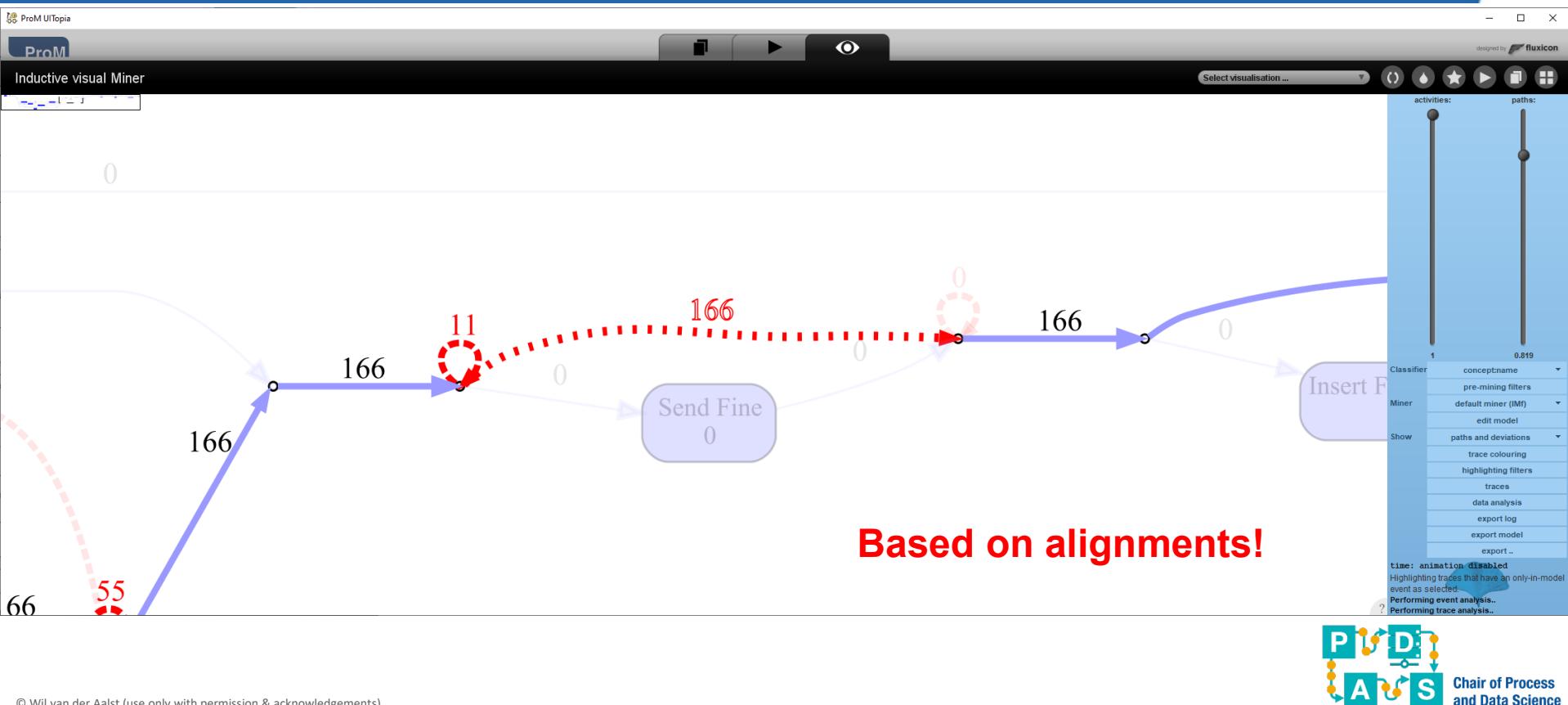
Example: Inductive Visual Miner



Example: Inductive Visual Miner



Example: Inductive Visual Miner



Example: Inductive Visual Miner

ProM UItopia designed by fluxicon

Inductive visual Miner

Select visualisation ...

activities: paths:

Classifier conceptname

Miner pre-mining filters

Show default miner (IMF) edit model paths and deviations trace colouring highlighting filters traces data analysis export log export model export ..

time: animation disabled

Highlighting traces that have an only-in-model event as selected.

Performing event analysis..

Performing trace analysis..

The screenshot shows the ProM UItopia interface with the 'Inductive visual Miner' tab selected. On the left, there is a process flow diagram with various nodes and transitions, some of which are highlighted in red. In the center, a list of 166 traces is displayed, each represented as a horizontal timeline of events. The traces are color-coded, and some specific events are highlighted in red. On the right, there are several filter and configuration options, including dropdown menus for 'Classifier' (conceptname), 'Miner' (pre-mining filters, default miner (IMF)), and 'Show' (edit model, paths and deviations, trace colouring, highlighting filters, traces, data analysis, export log, export model, export ..). Below these are buttons for 'time: animation disabled', 'Highlighting traces that have an only-in-model event as selected.', 'Performing event analysis..', and 'Performing trace analysis..'. At the bottom right, there is a logo for the Chair of Process and Data Science.

166 traces that have this deviation

Example: Inductive Visual Miner



Part I: Introduction

Chapter 1
Data Science in Action

Chapter 2
Process Mining:
The Missing Link

Part II: Preliminaries

Chapter 3
Process Modeling
and Analysis

Chapter 4
Data Mining

Part III: From Event Logs to Process Models

Chapter 5
Getting the Data

Chapter 6
Process Discovery:
An Introduction

Chapter 7
Advanced Process
Discovery Techniques

Part IV: Beyond Process Discovery

Chapter 8
Conformance
Checking

Chapter 9
Mining Additional
Perspectives

Chapter 10
Operational Support

Part V: Putting Process Mining to Work

Chapter 11
Process Mining
Software

Chapter 12
Process Mining in the
Large

Chapter 13
Analyzing “Lasagna
Processes”

Chapter 14
Analyzing “Spaghetti
Processes”

Part VI: Reflection

Chapter 15
Cartography and
Navigation

Chapter 16
Epilogue



ID	Topic	Date	Date	Place
	Lecture 1 Introduction to Process Mining	08.04.24	Monday	AH V
	Lecture 2 Data Science: Supervised Learning	09.04.24	Tuesday	AH V
	<i>Exercise 1 Tool Introduction</i>	09.04.24	Tuesday	AH III
	Lecture 3 Data Science: Unsupervised Learning and Evaluation	15.04.24	Monday	AH V
	Lecture 4 Introduction to Process Discovery	16.04.24	Tuesday	AH V
	<i>Exercise 2 Data Mining</i>	16.04.24	Tuesday	AH III
	Lecture 5 Alpha Algorithm 1	22.04.24	Monday	AH V
	Lecture 6 Alpha Algorithm 2	23.04.24	Tuesday	AH V
	<i>Exercise 3 Petri Nets</i>	23.04.24	Tuesday	AH III
	Lecture 7 Model Quality Representation	29.04.24	Monday	AH V
	Lecture 8 Heuristic Mining	30.04.24	Tuesday	AH V
	<i>Exercise 4 Alpha Miner</i>	30.04.24	Tuesday	AH III
	Lecture 9 Region-Based Mining	06.05.24	Monday	AH V
	<i>Exercise 5 Heuristic Mining and Region-Based Mining</i>	07.05.24	Tuesday	AH III
	Lecture 10 Inductive Mining	13.05.24	Monday	AH V
	Lecture 11 Event Data and Exploration	14.05.24	Tuesday	AH V
	<i>Exercise 6 Inductive Mining</i>	14.05.24	Tuesday	AH III
	Lecture 12 Conformance Checking 1	27.05.24	Monday	AH V
	Lecture 13 Conformance Checking 2	28.05.24	Tuesday	AH V
	<i>Q&A Session Assignment Part I</i>	28.05.24	Tuesday	AH III
	Deadline Assignment Part I	02.06.24	Sunday	
	<i>Exercise 7 Footprint and Token-Based Replay (Exercise)</i>	03.06.24	Monday	AH V
	<i>Exercise 8 Alignments (Exercise)</i>	04.06.24	Tuesday	AH V
	Lecture 14 Decision Mining	10.06.24	Monday	AH V
	<i>Lecture 15 Celonis Guest Lecture</i>	11.06.24	Tuesday	AH V
	<i>Exercise 9 Decision Mining</i>	11.06.24	Tuesday	AH III
	Lecture 16 Performance Analysis and Organizational Mining	17.06.24	Monday	AH V
	<i>Exercise 10 Performance Analysis (Exercise)</i>	18.06.24	Tuesday	AH V
	<i>Exercise 11 Organizational Mining</i>	18.06.24	Tuesday	AH III
	<i>Exercise 12 Celonis Case Study</i>	24.06.24	Monday	AH V
	Lecture 17 Operational Support and Process Mining Applications	01.07.24	Monday	AH V
	Lecture 18 Distributed, Streaming, and Comparative Process Mining	02.07.24	Tuesday	AH V
	<i>Exercise 13 Operational Process Mining</i>	02.07.24	Tuesday	AH III
	Lecture 19 Closing	08.07.24	Monday	AH V
	<i>Q&A Session Assignment Part II</i>	09.07.24	Tuesday	AH III
	Deadline Assignment Part II	14.07.24	Sunday	
	<i>Q&A Session Exam</i>	16.07.24	Tuesday	AH III



Conformance Checking (2/2)

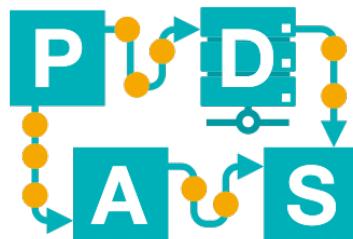
Lecture 13

prof.dr.ir. Wil van der Aalst

www.vdaalst.com @wvdaalst

www.pads.rwth-aachen.de

BPI-L13



Chair of Process
and Data Science

RWTHAACHEN
UNIVERSITY

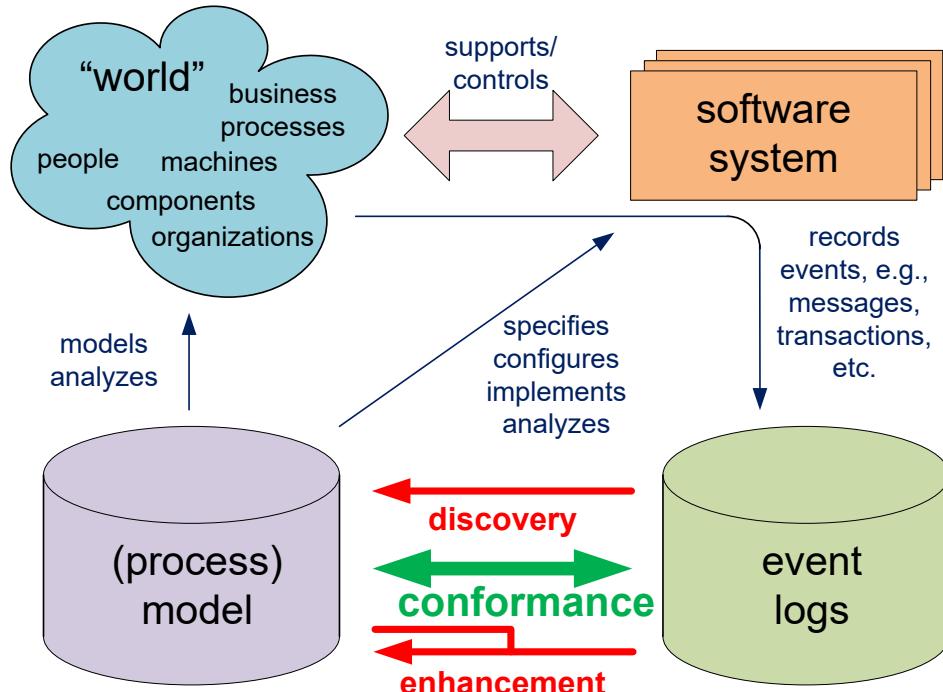
Outline

- A few end-to-end examples using token-based replay.
- Alignment-based conformance checking.
- Tool support for conformance checking.
- Applications of process mining.

Token-based replay revisited



Conformance checking



1. Conformance checking using causal footprints.
2. Conformance checking based on **token-based replay**.
3. Alignment-based conformance checking.

Last lecture: Token-based replay

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdfbeh
38	adbeg
33	acdefbdeh
14	acdefbddeg
11	acdefdbeg
9	adcefcdgeh
8	adcefdbeh
5	adcefbddeg
3	adcefbdedbeg
2	adcefdbeg
2	adcefbdedbdeg
1	adcefdbefbdbeh
1	adbefbdedbeg
1	adcefdbefcdefdbeg
1391	

missing tokens

?

consumed tokens

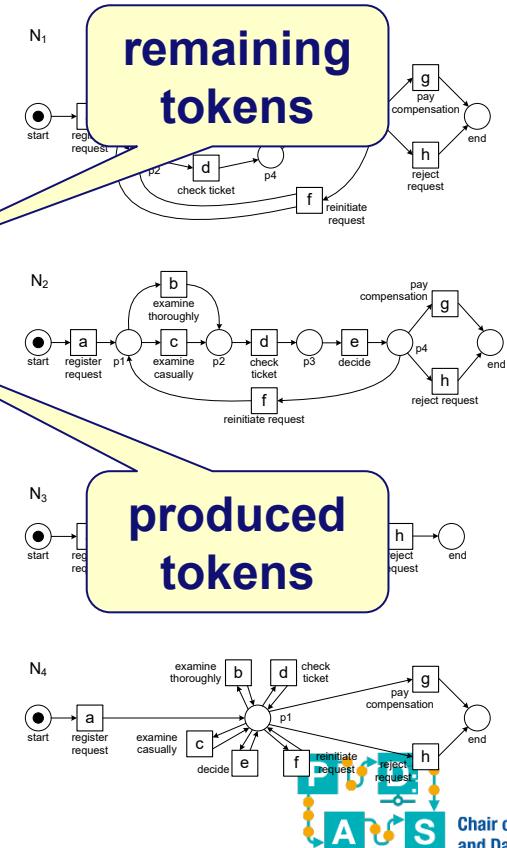
$$fitness(L, N) = \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}} \right) + \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}} \right)$$

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

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

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

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



End-to-end examples



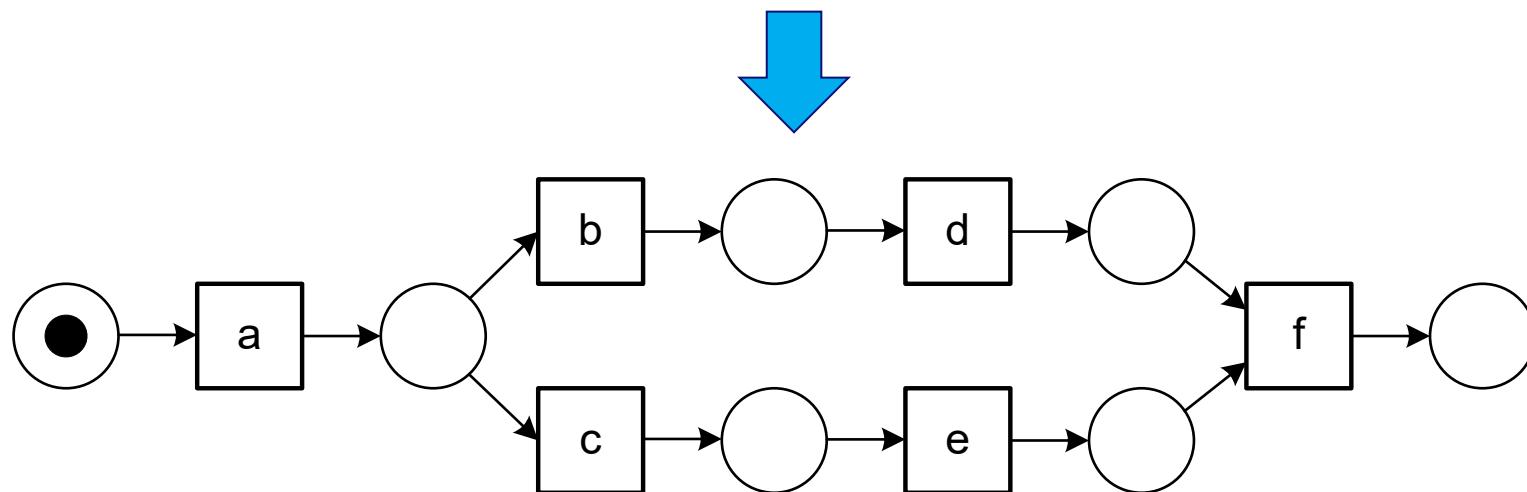
Question

$$L = [\langle a, b, d, e, f \rangle^{10}, \langle a, c, e, d, f \rangle^{10}]$$

- Consider the above event log.
- Give the model that the Alpha algorithm generates.
- Compute fitness using missing and remaining tokens.
- Comment on the findings.

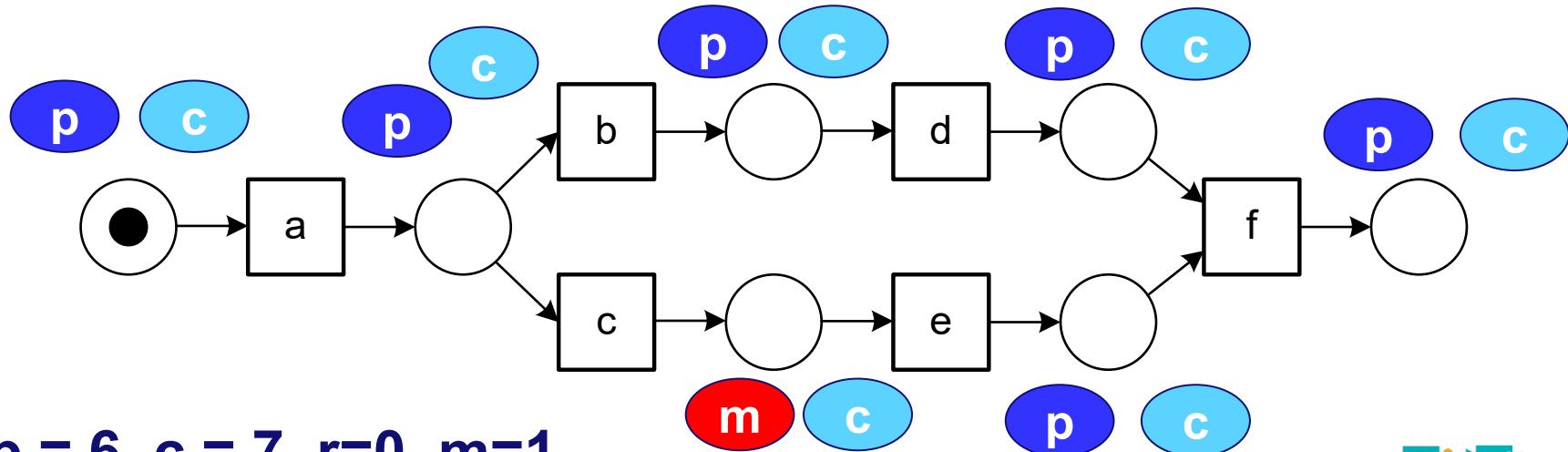
Model generated by the Alpha Algorithm

$$L = [\langle a, b, d, e, f \rangle^{10}, \langle a, c, e, d, f \rangle^{10}]$$



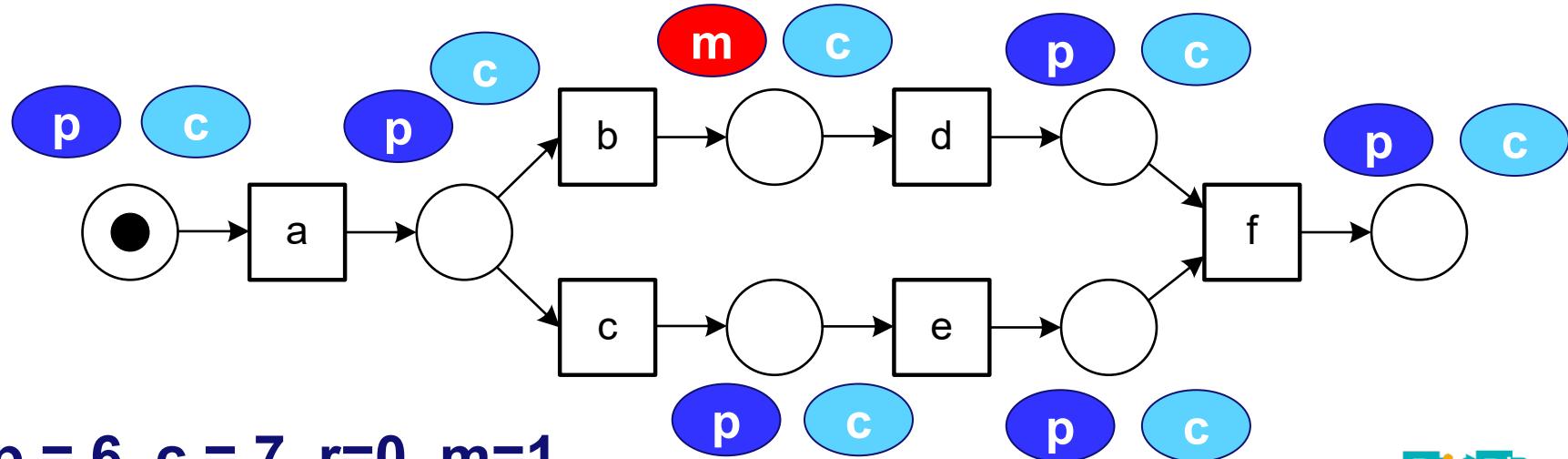
Trace abdef

$$L = [\langle a, b, d, e, f \rangle^{10}, \langle a, c, e, d, f \rangle^{10}]$$



Trace acedf

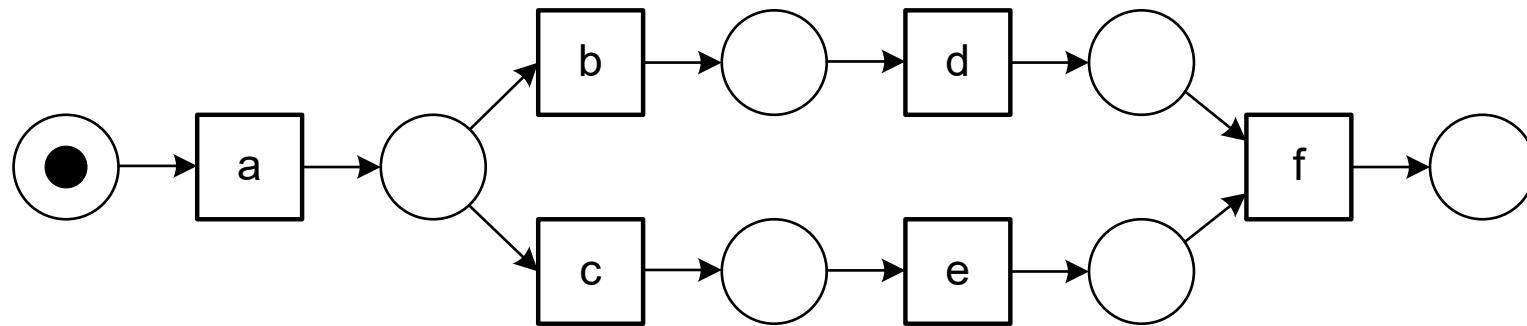
$$L = [\langle a, b, d, e, f \rangle^{10}, \boxed{\langle a, c, e, d, f \rangle^{10}}]$$



Overall fitness

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

$$L = [\langle a, b, d, e, f \rangle^{10}, \langle a, c, e, d, f \rangle^{10}]$$

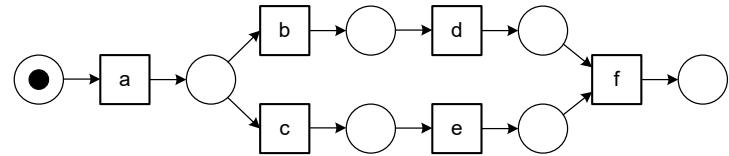


$$p = 2 * 10 * 6 = 120, c = 2 * 10 * 7 = 140, r = 2 * 10 * 0 = 0, m = 2 * 10 * 1 = 20$$

$$\frac{1}{2} \left(1 - \frac{20}{140} \right) + \frac{1}{2} \left(1 - \frac{0}{120} \right) = \frac{13}{14} \approx 0.93$$

Model is not sound!

$$L = [\langle a, b, d, e, f \rangle^{10}, \langle a, c, e, d, f \rangle^{10}]$$



$$\frac{1}{2} \left(1 - \frac{20}{140} \right) + \frac{1}{2} \left(1 - \frac{0}{120} \right) = \frac{13}{14} \approx 0.93$$

- The model is not sound. Actually there is no firing sequence leading to the target marking!
- How to interpret the result? Therefore, we typically require “relaxed soundness”, i.e., there is at least one firing sequence leading to the target marking.

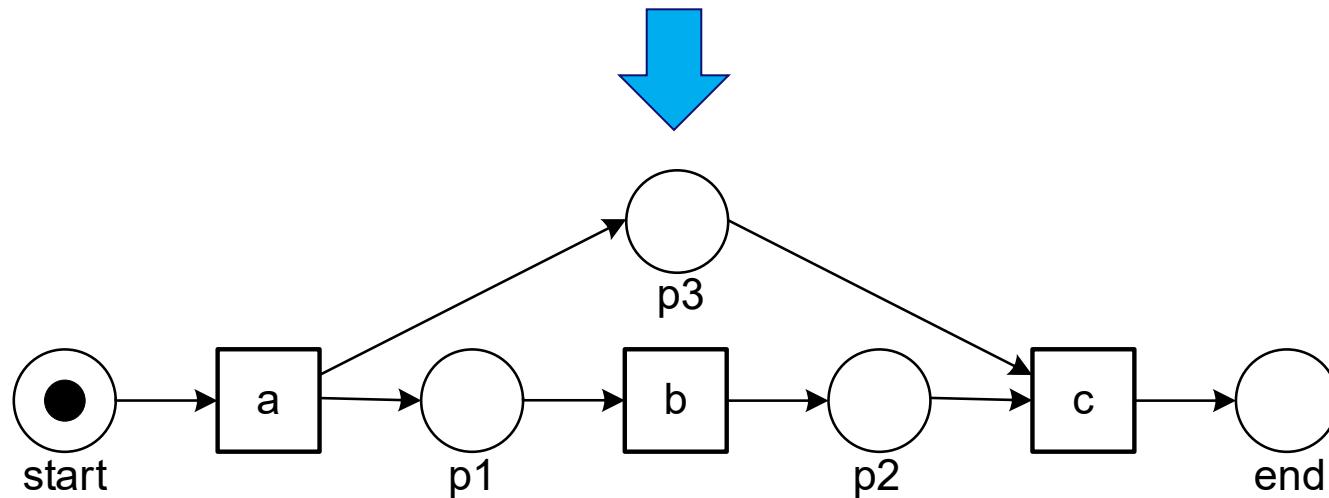
Question

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

- Consider the above event log.
- Give the model that the Alpha algorithm generates
- Compute fitness using missing and remaining tokens.
- Comment on the findings.

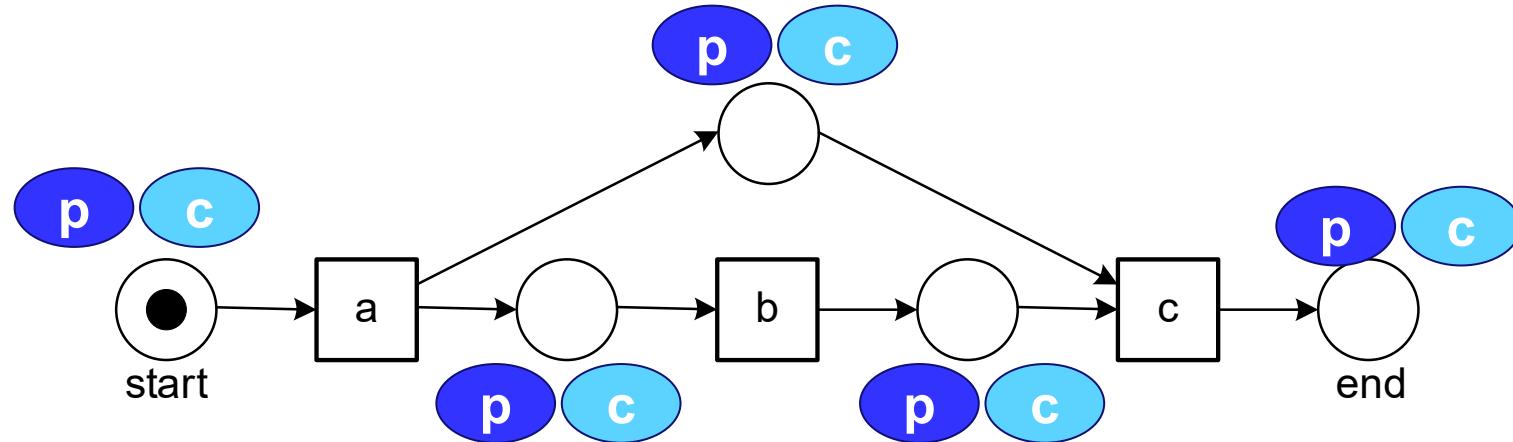
Model generated by the Alpha Algorithm

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



Trace abc

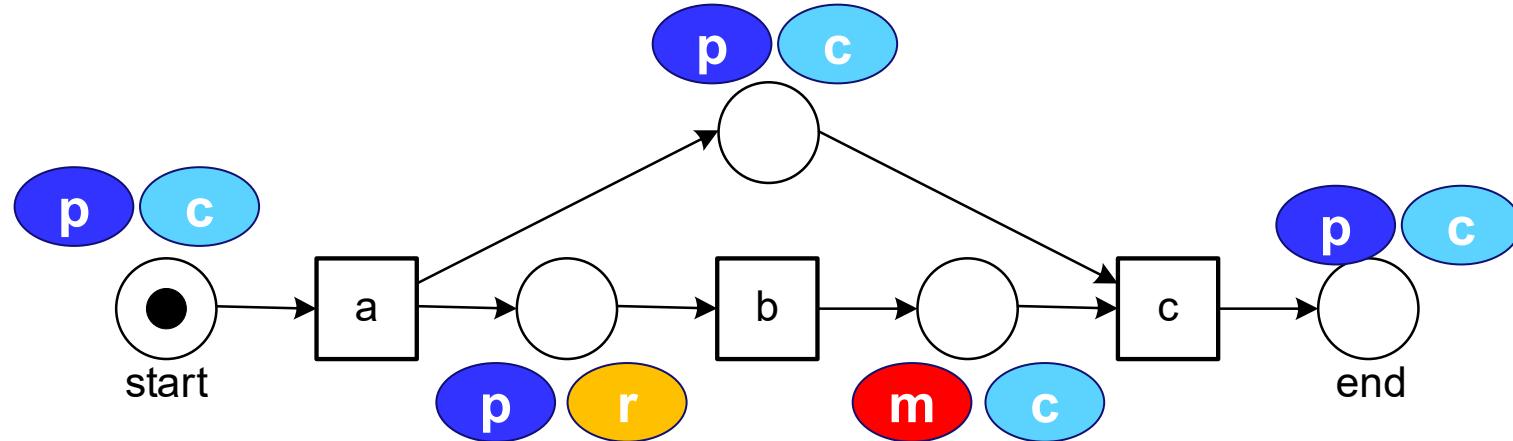
$$L_{11} = [\langle a, b, c \rangle^{20}, \langle a, c \rangle^{30}]$$



$p = 5, c = 5, r=0, m=0$

Trace ac

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



$p = 4, c = 4, r=1, m=1$

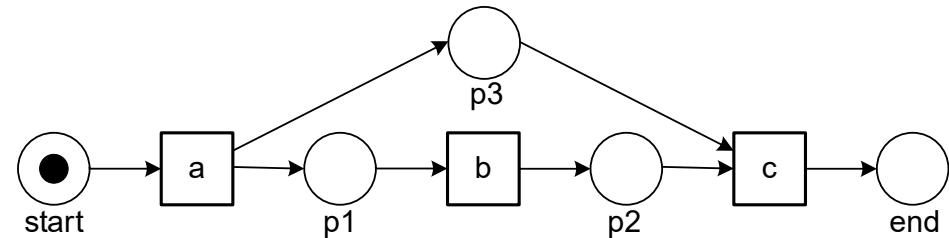
Overall fitness

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

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

p = 5, c = 5, r=0, m=0

p = 4, c = 4, r=1, m=1

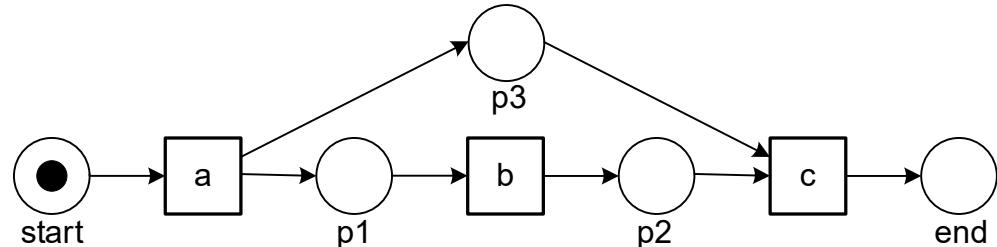


$$\begin{aligned} p &= 20*5+30*4 = 220, \quad c = 20*5+30*4=220, \\ r &= 20*0+30*1=30, \quad m=20*0+30*1=30 \end{aligned}$$

$$\frac{1}{2} \left(1 - \frac{30}{220} \right) + \frac{1}{2} \left(1 - \frac{30}{220} \right) = \frac{19}{22} \approx 0.86$$

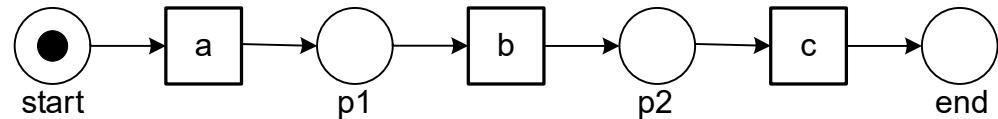
Redundant places impact fitness (1/2)

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



$$\frac{1}{2} \left(1 - \frac{30}{220} \right) + \frac{1}{2} \left(1 - \frac{30}{220} \right) = \frac{19}{22} \approx 0.86$$

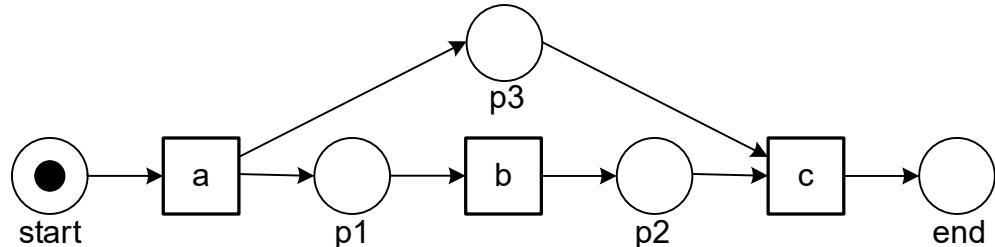
$$L_{11} = [\langle a, b, c \rangle^{20}, \langle a, c \rangle^{30}]$$



$$\frac{1}{2} \left(1 - \frac{30}{170} \right) + \frac{1}{2} \left(1 - \frac{30}{170} \right) = \frac{14}{17} \approx 0.82$$

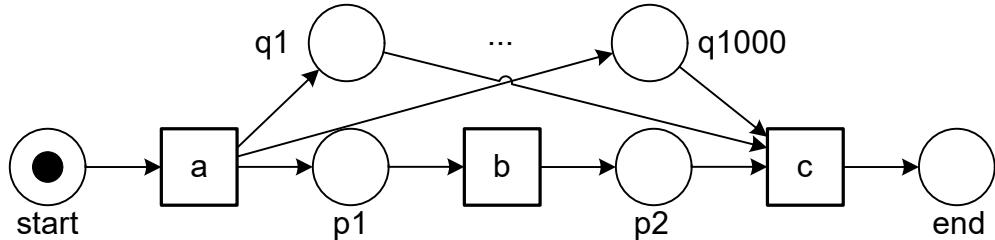
Redundant places impact fitness (2/2)

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



$$\frac{1}{2} \left(1 - \frac{30}{220} \right) + \frac{1}{2} \left(1 - \frac{30}{220} \right) = \frac{19}{22} \approx 0.86$$

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

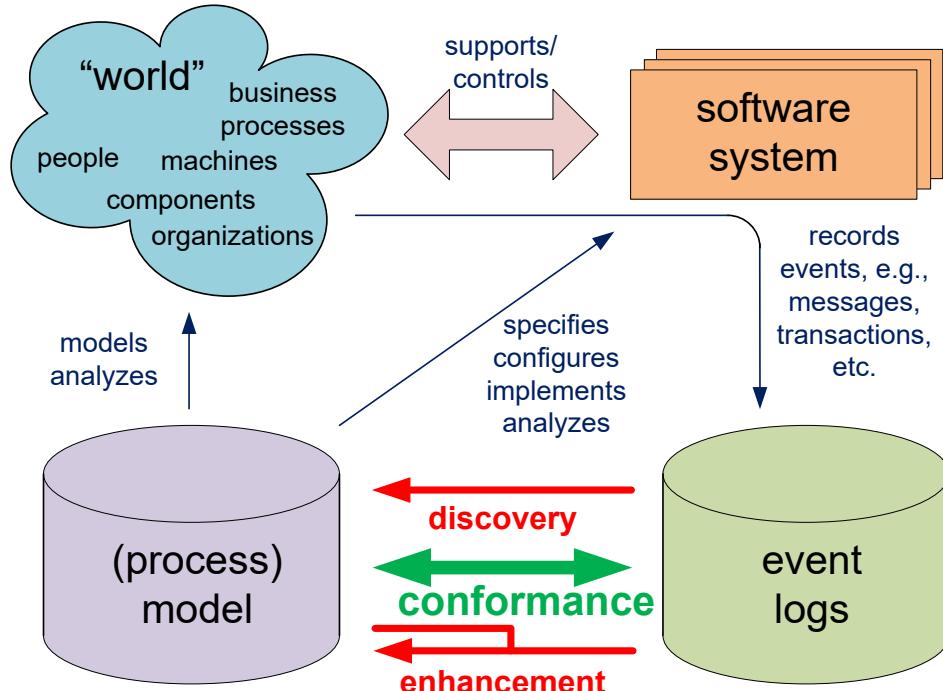


$$\frac{1}{2} \left(1 - \frac{30}{50170} \right) + \frac{1}{2} \left(1 - \frac{30}{50170} \right) = \frac{5014}{5017} \approx 0.999$$

Aligning Observed and Modeled Behavior



Conformance checking

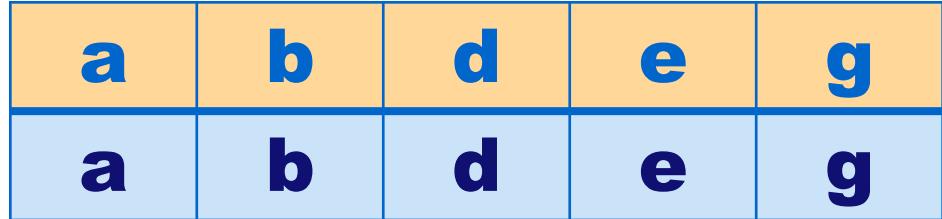


1. Conformance checking using causal footprints.
2. Conformance checking based on token-based replay.
3. Alignment-based conformance checking.

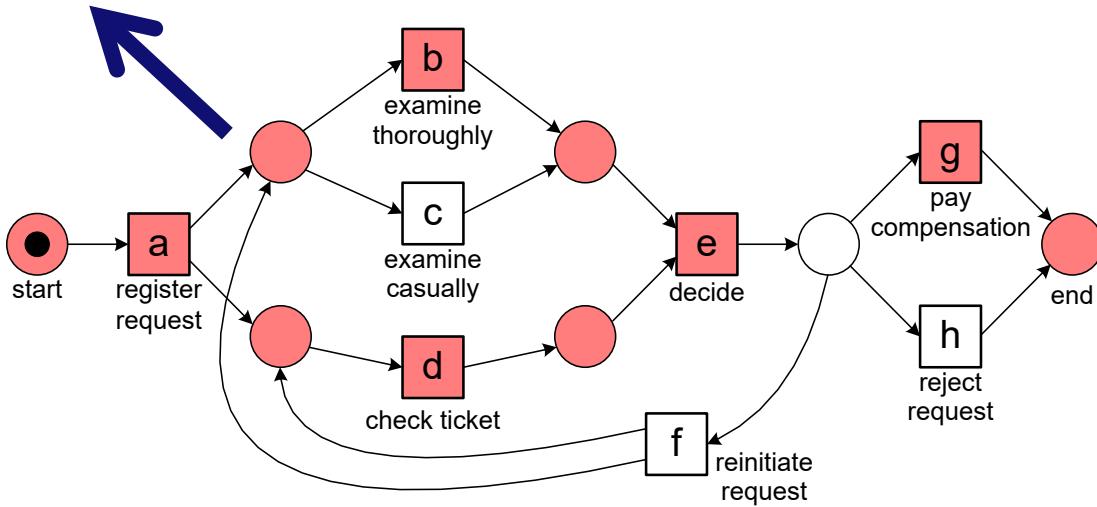
Requirements

- Conformance checking should **not** impose restrictions on the process notation (e.g., silent transitions and two transitions with the same label should be possible).
- Two **semantically equivalent** models should have the same conformance value.
- Should provide a "**closest matching path**" through the process model for any trace in the event log.
 - Also required for **performance analysis!**
 - **Beyond** the analysis of replay fitness (advanced diagnostics, precision, generalization, etc.).

Alignments



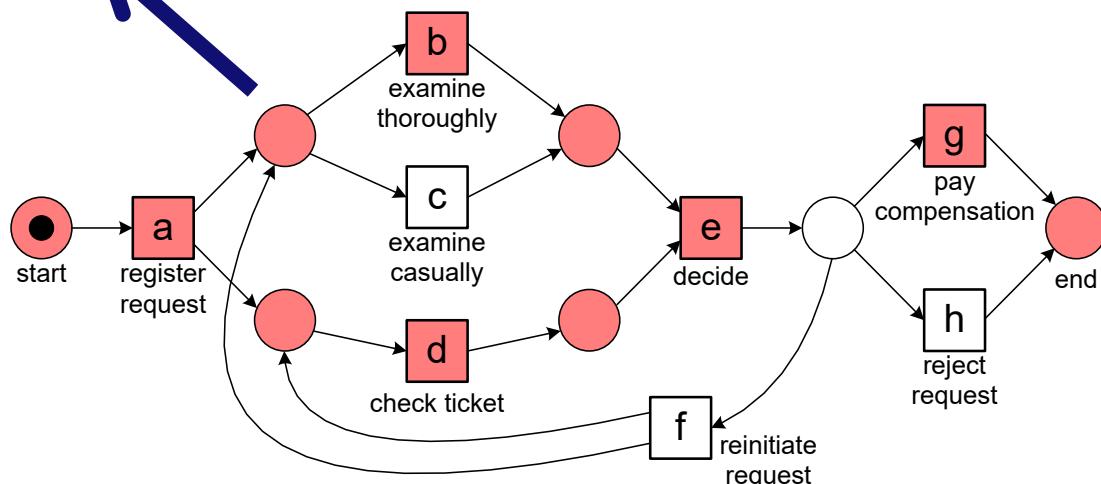
$\langle a, b, d, e, g \rangle$



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

Alignments

a	»	d	e	g	h
a	b	d	e	g	»



Terminology

alignment
(sequence of moves)

move in log only

move in model

move in both

a			d	e	g
	c	d	e	g	

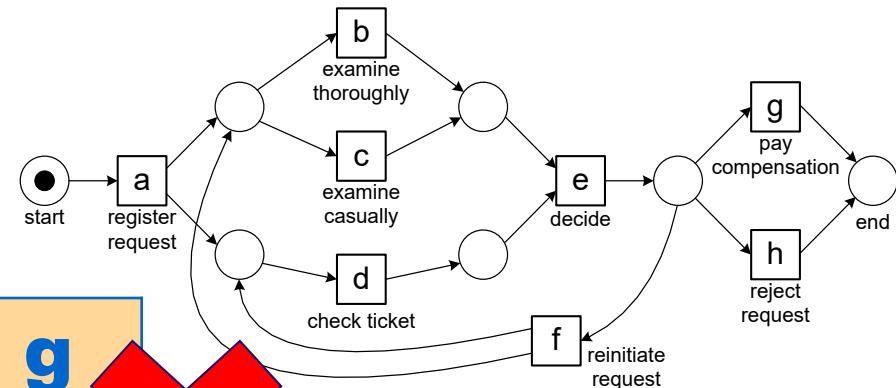
- Independent of process model notation!
- Projection on top row (remove "no moves") corresponds to the **run** in the event log.
 - Projection on bottom row (remove "no moves") corresponds to a **run of the model**.

Optimal alignment for $\langle a,b,d,e,g \rangle$

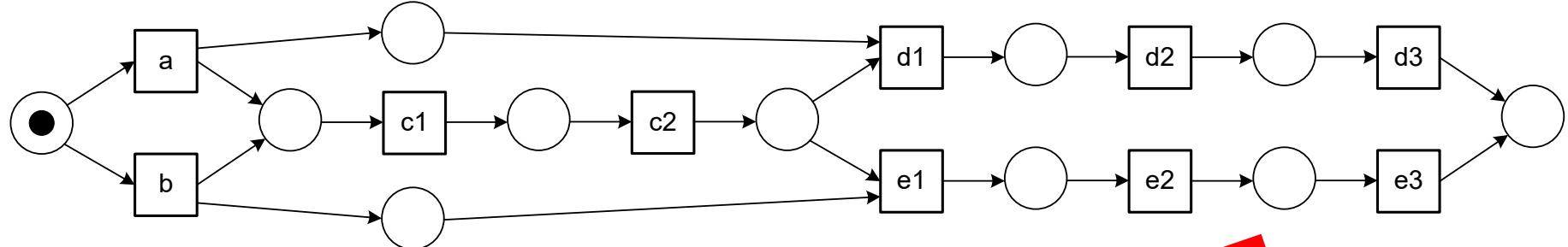
a	b	d	e	g
a	b	d	e	g

a	b	»	d	e	g
a	»	c	d	e	g

a	b	d	e	g	»	»	»	»	»	»
»	»	»	»	»	a	c	d	e	g	»



Optimal alignment for $\langle a, c_1, c_2, e_1, e_2, e_3 \rangle$?

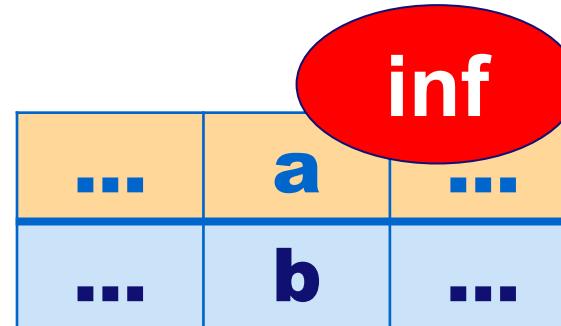
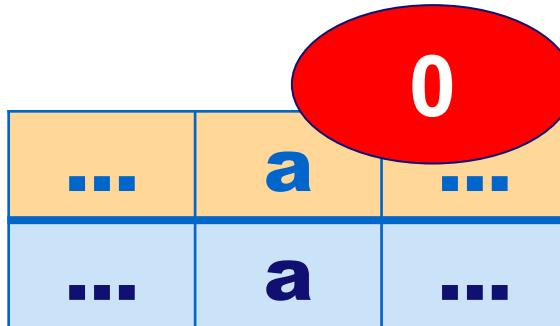
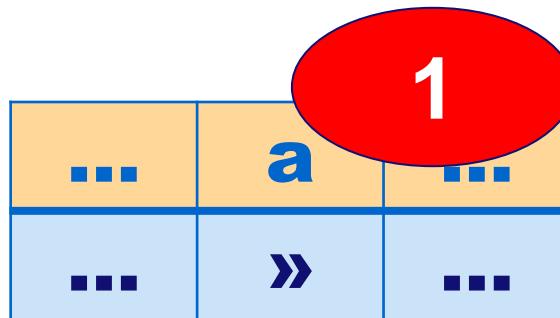


a	»	c1	c2	e1	e2	e3
»	b	c1	c2	e1		

a	c1	»	»	e1	e2	e3
a	c	d1	d2	d3	»	»

Depends on cost function!

Standard cost function count »'s in alignment



Using the standard cost function

optimal

there is no other alignment
that has lower costs

a	b	d	e	g
a	b	d	e	g

0

0

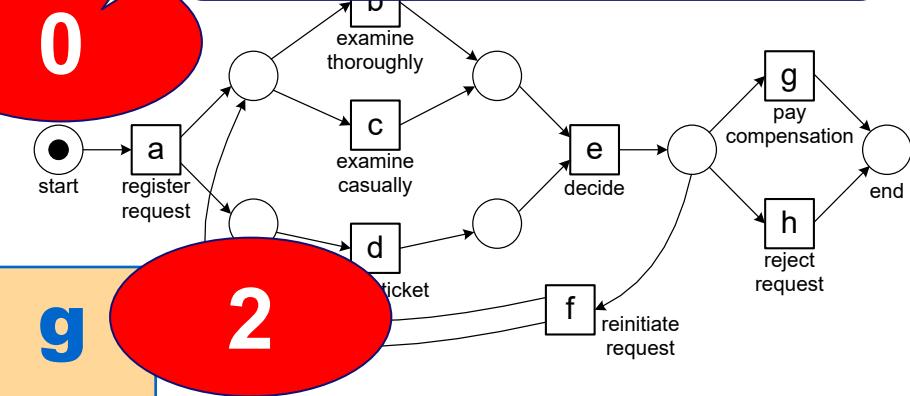
0

0

0

a	b	»	d	e	g
a	»	c	d	e	g

2



a	b	d	e	g	»	»	»	»	»	»
»	»	»	»	»	a	c	d	e	g	D

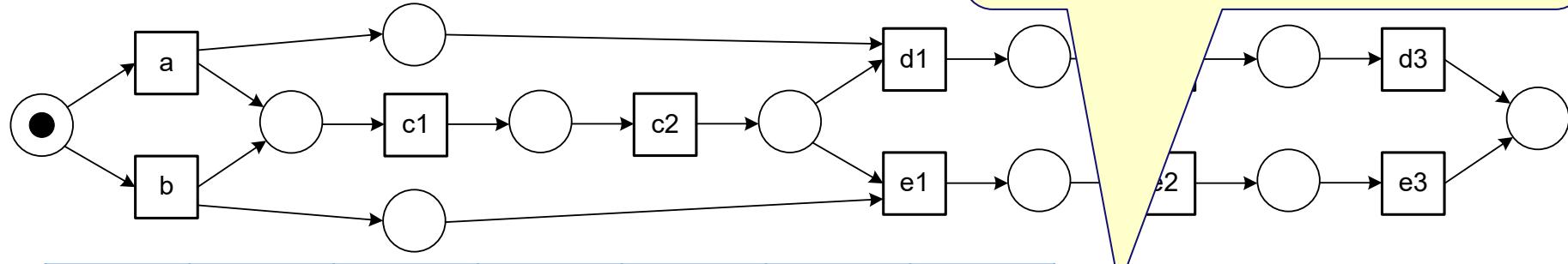
10

Using the standard cost function

trace in log: $\langle a, c1, c2, e1, e2, e3 \rangle$

optimal

there is no other alignment
that has lower costs



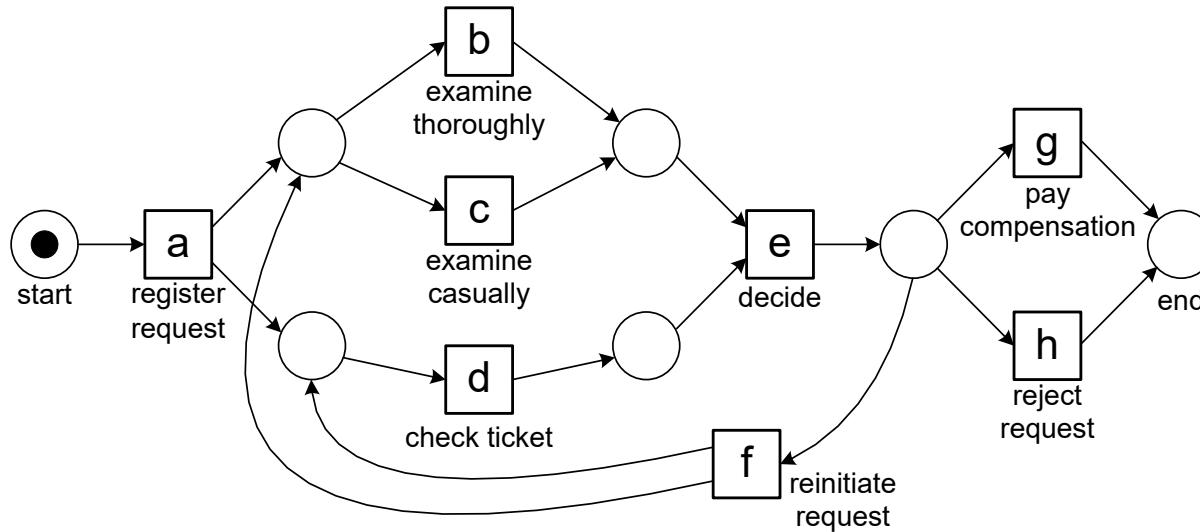
a	»	c1	c2	e1	e2	e3
»	b	c1	c2	e1	e2	e3

a	c1	c2	»	»	»	e1	e2	e3
a	c1	c2	d1	d2	d3	»	»	»

6

Optimal alignment for $\langle a,b,e,f,d,e,g \rangle$ (1/2)

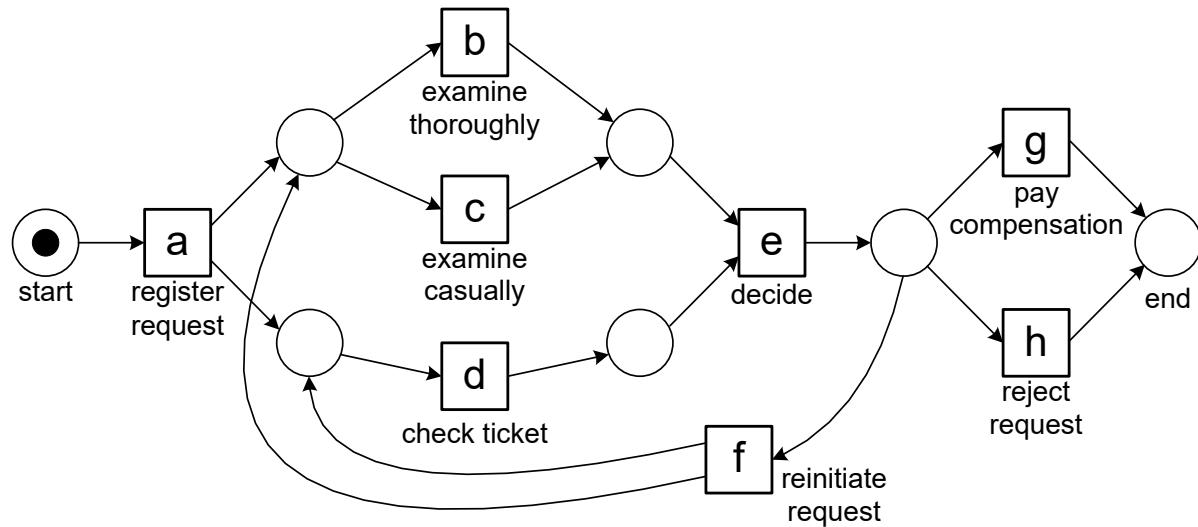
a	b	e	f	d	e	g	
a	b	»	»	d	e	g	2



loop is not taken: e and f in event log are discarded

Optimal alignment for $\langle a, b, e, f, d, e, g \rangle$ (2/2)

a	b	»	e	f	d	»	e	g	2
a	b	d	e	f	d	b	e	g	



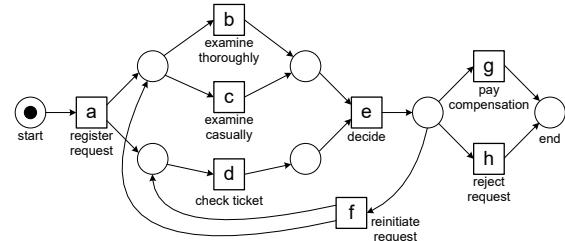
loop is taken:
d and b are
missing in
event log

Not one unique optimal alignment

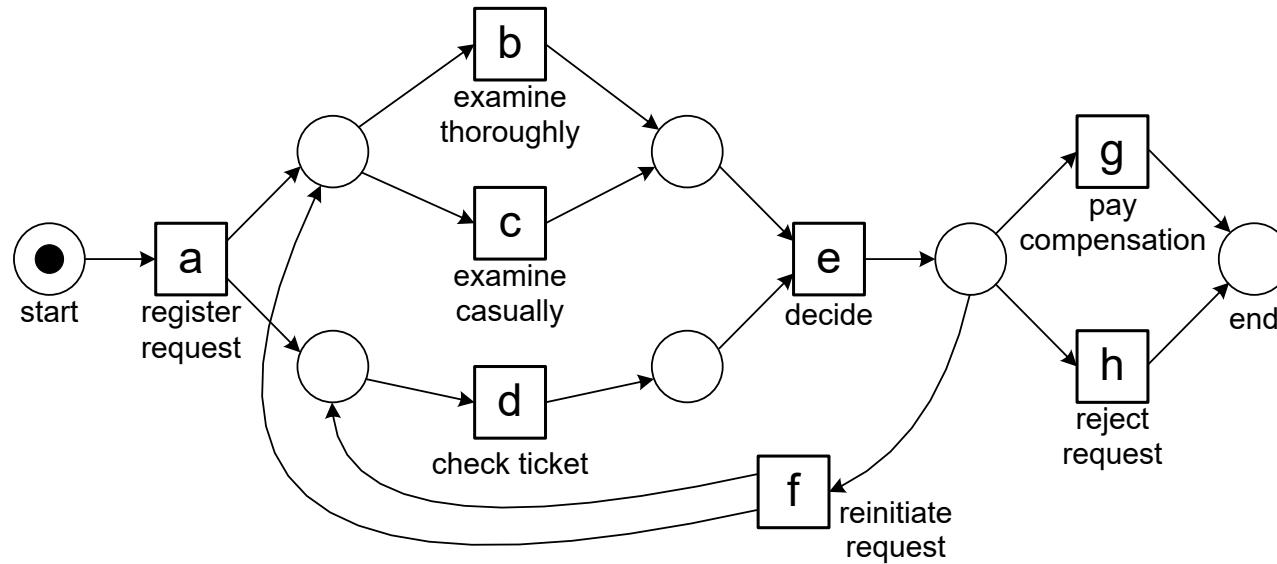
a	b	»	e	f	d	»	e	g	2
a	b	d	e	f	d	b	e	g	

a	b	e	f	d	e	g	2
a	b	»	»	d	e	g	

...



Question: How many optimal alignments are there for $\langle a,b,e,f,d,e,g \rangle$?



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

Answer: 9

$1 + (2 \times 2) + (2 \times 2) = 9$ optimal alignments having cost 2

a	b	e	f	d	e	g
a	b	»	»	d	e	g

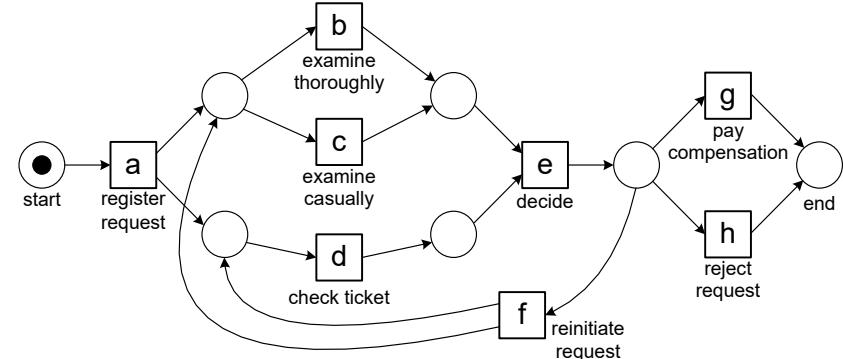
1x

move in model can be reordered in concurrent part



4x

a	b	»	e	f	d	»	e	g
a	b	d	e	f	d	b	e	g



4x

a	b	»	e	f	d	»	e	g
a	b	d	e	f	d	c	e	g

Any cost structure is possible

...	send-letter(John,4 weeks, \$400)	...
...	send-email(Sue,3 weeks,\$500)	...

Any cost structure is possible

...	send-letter(John,4 weeks, \$400)	...
...	send-email(Sue,3 weeks,\$500)	...

similar activities (lower costs for related activities)

Any cost structure is possible

...	send-letter(John,4 weeks, \$400)	...
...	send-email(Sue,3 weeks,\$50)	...

**resource-related conformance costs
(done by someone that does or does not have
the specified role)**



Any cost structure is possible

...	send-letter(John,4 weeks, \$400)	...
...	send-email(Sue,3 weeks,\$500)	...

**time-related conformance costs
(activity should happen within a preset deadline)**



Any cost structure is possible

...	send-letter(John,4 weeks, \$400)	...
...	send-email(Sue,3 weeks,\$500)	...

**data-related conformance costs
(routing condition is violated, e.g., path
only for more valuable orders)**

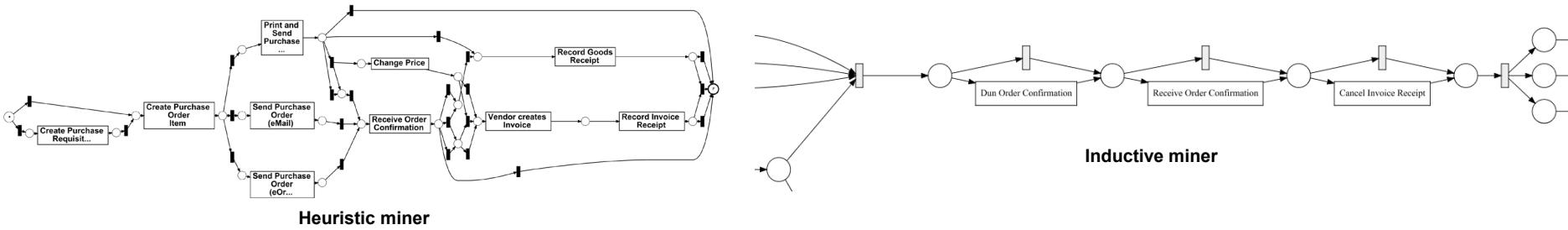


Any cost structure is possible

...	send-letter(John,4 weeks, \$400)	...
...	send-email(Sue,3 weeks,\$500)	...

risk-related conformance costs, context-dependent conformance costs, ...

Side note: Silent transitions

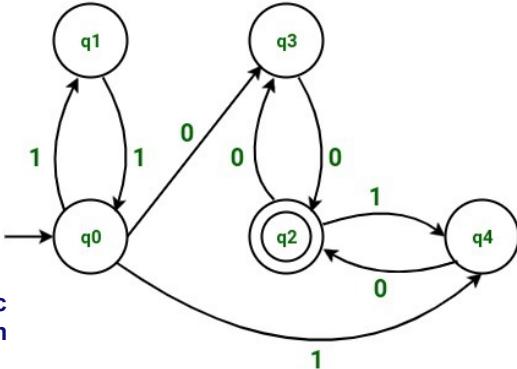
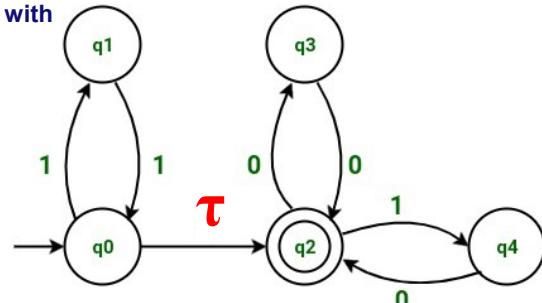


- Two options:
 - Remove these transitions when considering the model's behavior (Model reduction: any Nondeterministic Finite Automaton (NFA) with silent activities, can be translated into an NFA without silent activities, which, in turn, can be translated into a Deterministic Finite Automaton (DFA).)
 - The corresponding moves on model have cost 0 (or a very small epsilon in case of loops)

Example Transformations

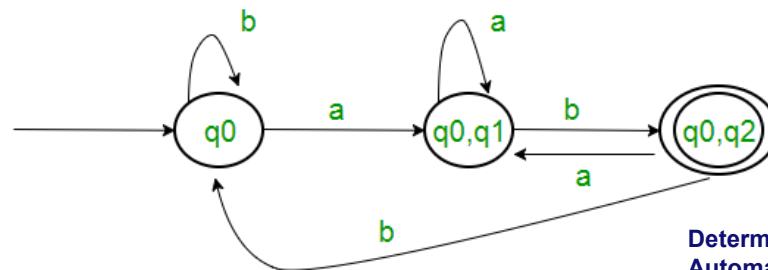
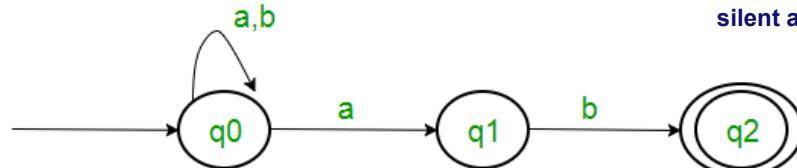
(outside scope course)

Nondeterministic Finite Automaton (NFA) with silent activities



Nondeterministic Finite Automaton
NFA without
silent activities

Nondeterministic Finite Automaton
NFA without
silent activities



Deterministic Finite Automaton (DFA)



Chair of Process
and Data Science

Examples taken from <https://www.geeksforgeeks.org/>

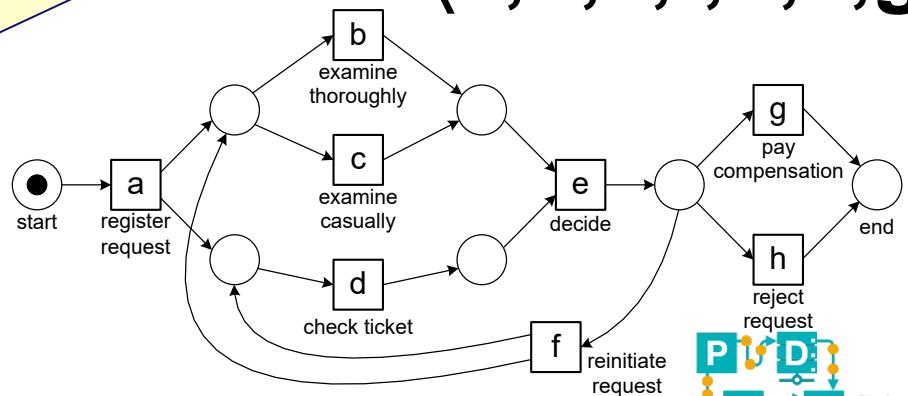
Computing fitness

Computing fitness
all events cause a move in log only
model.

$$1 - \frac{2}{7+5} = 0.833$$

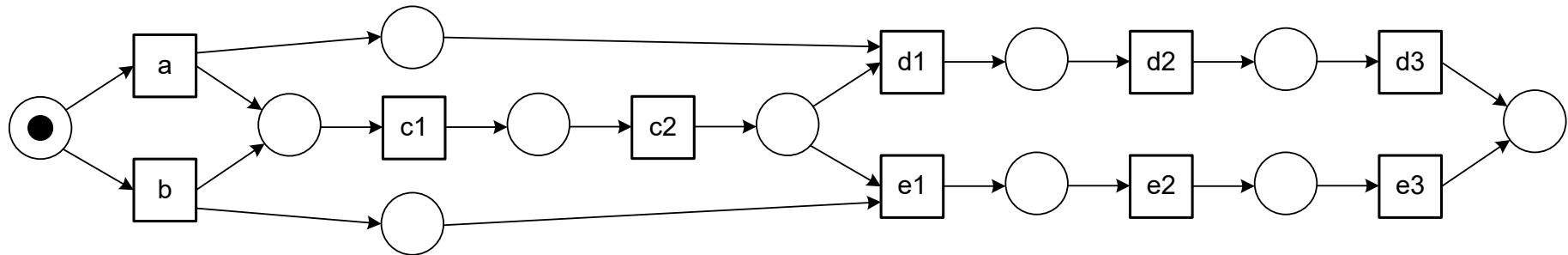
an optimal alignment with a
shortest path from initial state to final state

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



Question: Compute alignment-based fitness

$\langle a, c1, c2, e1, e2, e3 \rangle$

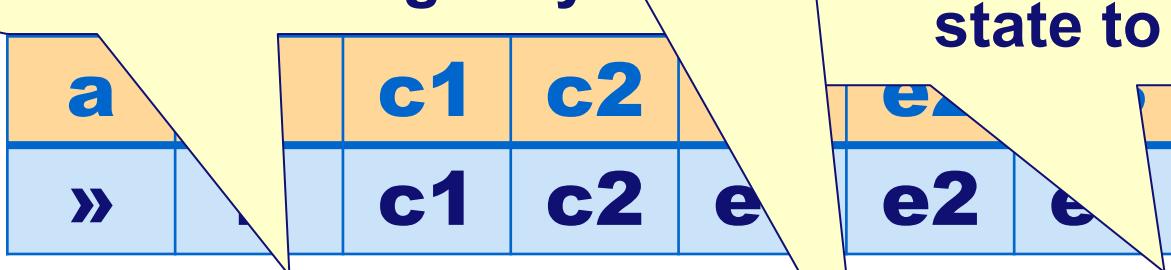


Answer

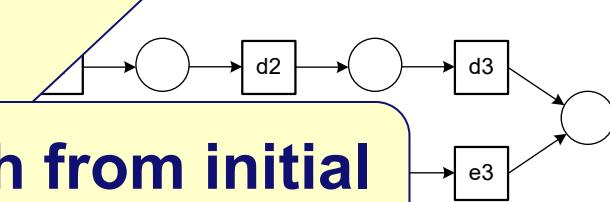
cost of optimal alignment = 2

$\langle a, c1, c2, e1, e2, e3 \rangle$

all even worst-case scenario
moves bring only



shortest path from initial state to final state

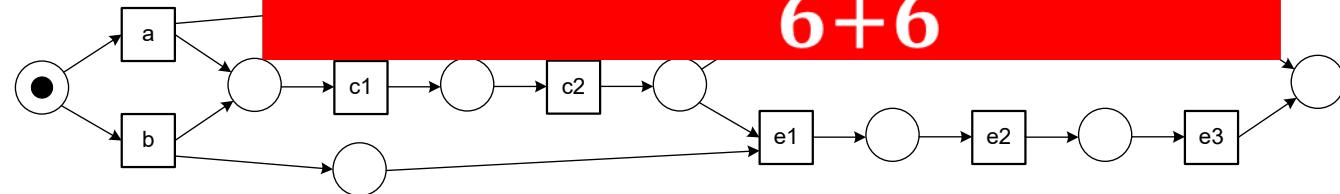


a	c1	c2	e1	e2	e3	»	»	»	»	»	»
»	»	»	»	»	»	a	c1	c2	d1	d2	d3

Answer

$\langle a, c1, c2, e1, e2, e3 \rangle$

$$\text{fitness} = 1 - \frac{2}{6+6} = 0.833$$

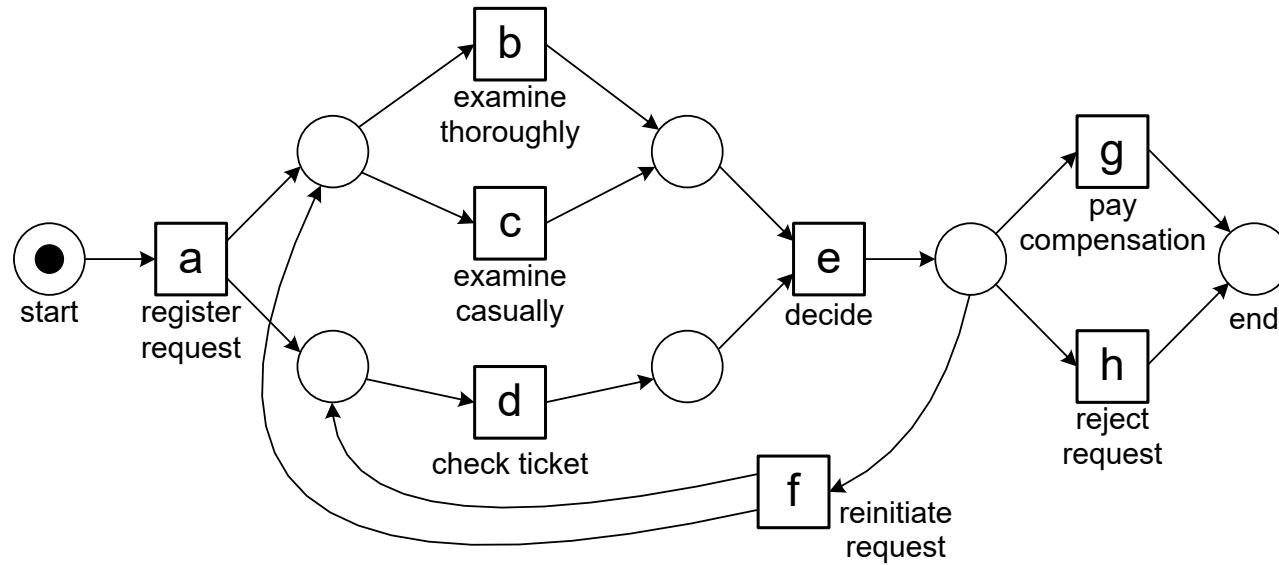


a	»	c1	c2	e1	e2	e3
»	b	c1	c2	e1	e2	e3

a	c1	c2	e1	e2	e3	»	»	»	»	»	»
»	»	»	»	»	»	a	c1	c2	d1	d2	d3

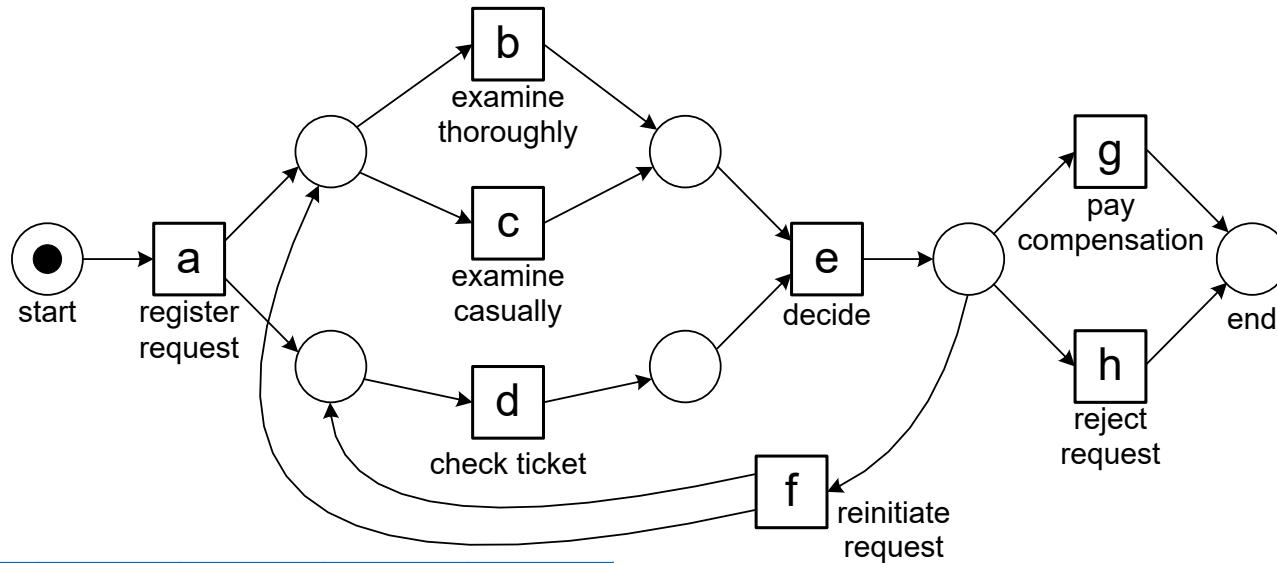
6

Another example: What is the alignment-based fitness of the trace $\langle b, c \rangle$?



$\langle b, c \rangle$

What is the alignment-based fitness of the trace $\langle b,c \rangle$?

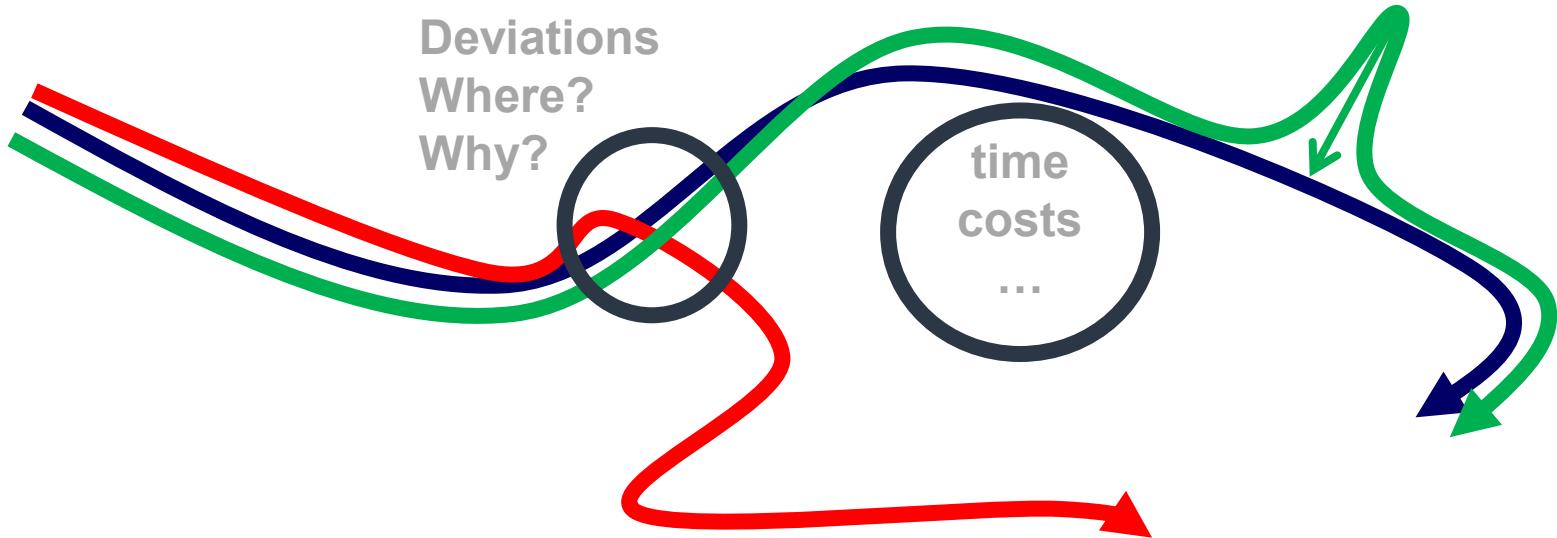


»	b	c	»	»	»
a	b	»	d	e	h

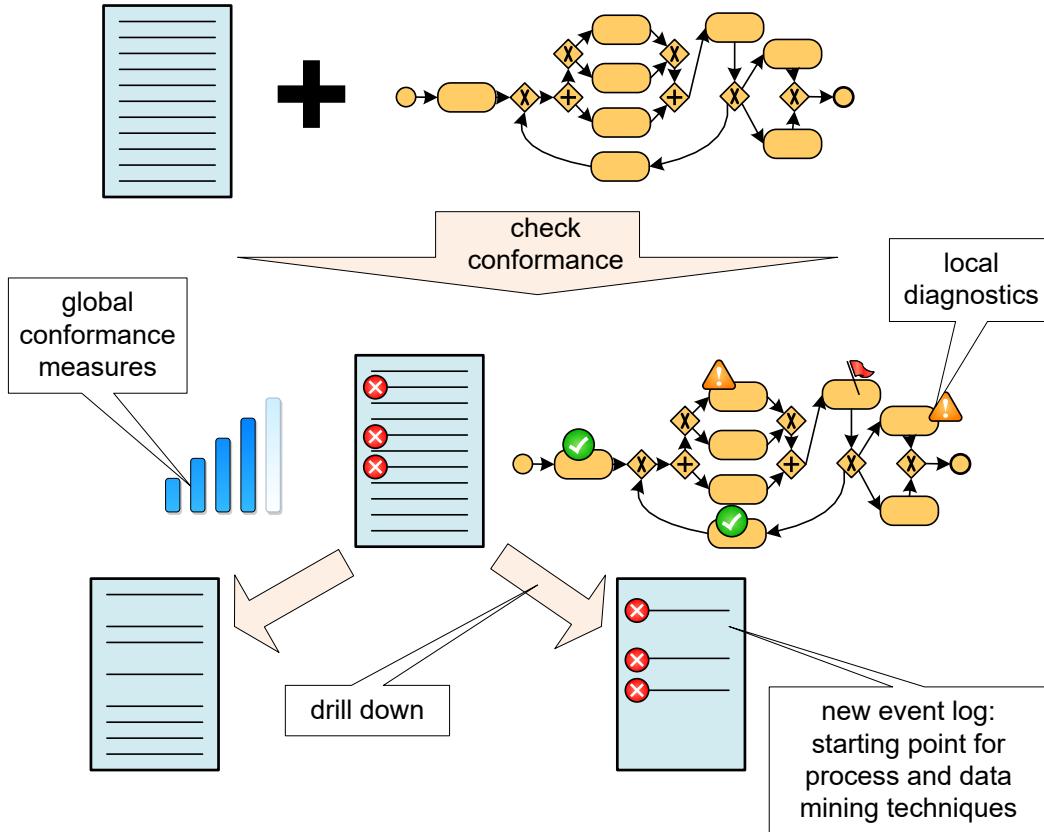
$$1 - \frac{5}{2 + 5} = \frac{2}{7} = 0.286$$

Advantages of aligning log and model

- Observed behavior is directly related to modeled behavior.
- Very flexible (any cost structure).
- Detailed diagnostics.
- After aligning log and model, other quality dimensions can be investigated (separation of concerns).



Drilling down



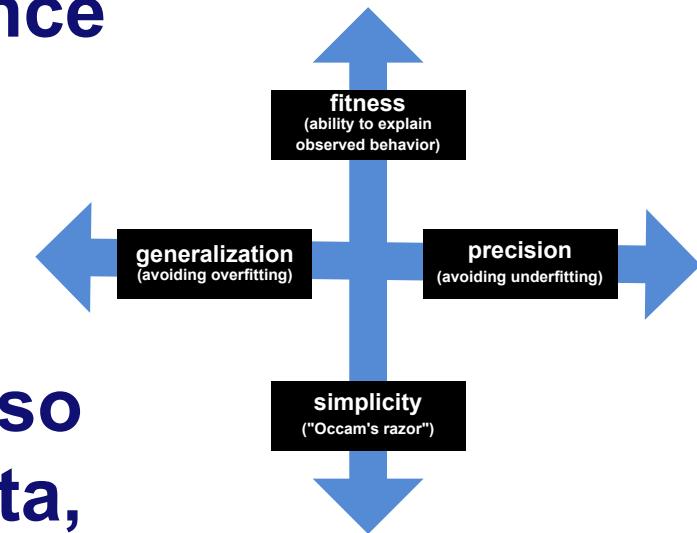
Example approach

- Create event log containing deviating (or non-deviating) cases.
- Apply process discovery to new log.
- Compare process models.

Later more on comparative process mining.

Beyond fitness and control-flow

- There are also solid conformance measures for **precision**, **generalization**, and **simplicity**.
- Multiple definitions possible.
- Conformance checking may also include **other perspectives** (data, resources, time, cost, etc.).
- Example: **data-aware alignments**.



Example: Precision (1/2)

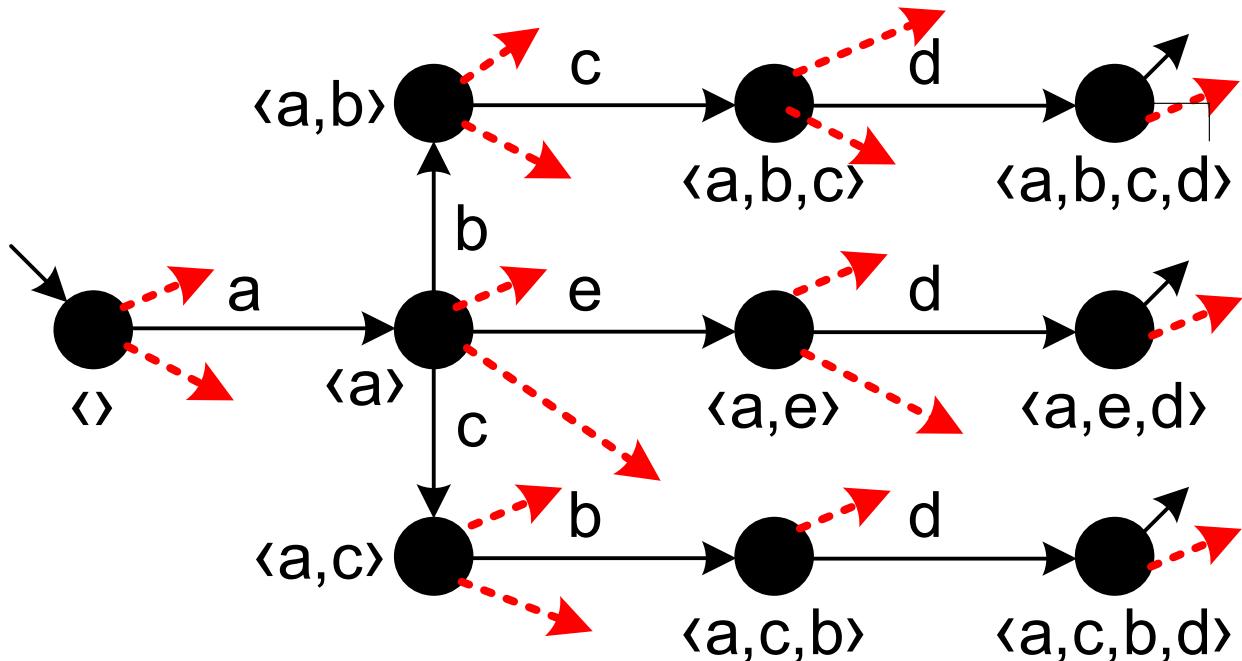
- **Naïve approach:** Compute the fraction of modeled behavior actually observed in the event log.
- **Problems:**
 - If “behavior” = “trace”, then the model with loops has a precision of 0 (because there are ∞ -many traces).
 - Sensitive to the size of the event log. Precision gets better when a longer period is taken.
 - Recall that an event log contains example behavior and often many unique traces.



Example: Precision (2/2)

- Many smarter approaches, e.g., using **escaping edges**.
- Assume the event log has been aligned, i.e., the remaining events are synchronous moves or model moves.
- Build a **prefix automaton** (see lecture on region-based mining) based on the aligned event log.
- Extend the prefix automaton with **escaping edges**, i.e., situation where the model allows for more behavior.
- Quantify such escaping edges.

Prefix Automaton with escaping edges



W.M.P. van der Aalst, A. Adriansyah, and B. van Dongen. Replay History on Process Models for Conformance Checking and Performance Analysis. WIREs Data Mining and Knowledge Discovery, 2(2):182-192, 2012.
A. Adriansyah, J. Munoz-Gama, J. Carmona, B.F. van Dongen, and W.M.P. van der Aalst. Measuring Precision of Modeled Behavior. Information Systems and e-Business Management, 13(1):37-67, 2015.

Red arcs indicate that while replaying the model could do more than what was observed in the event log. It is possible to qualify precision by taking into account how often a node is visited and what the fraction of escaping edges is.

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



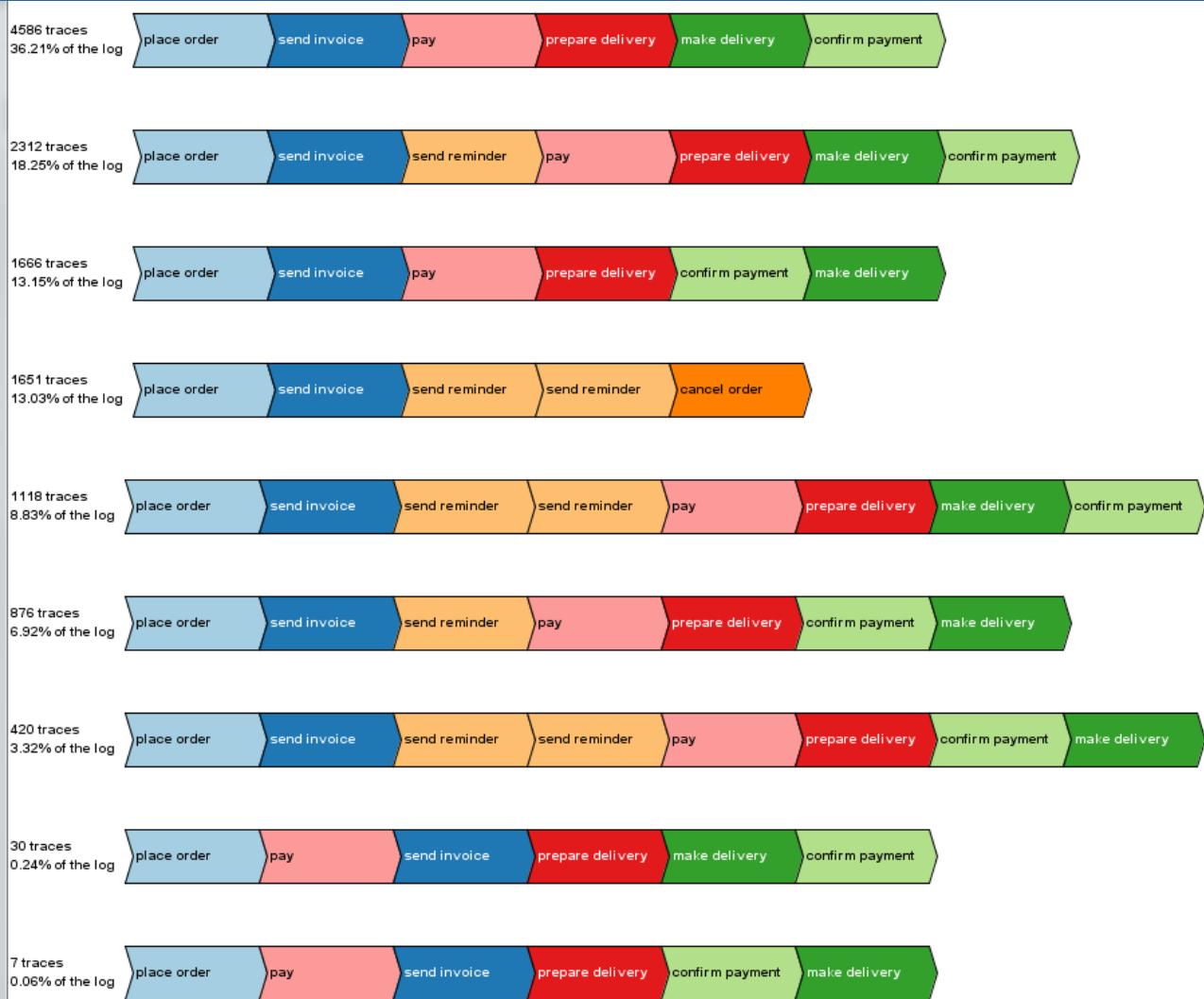
Tooling



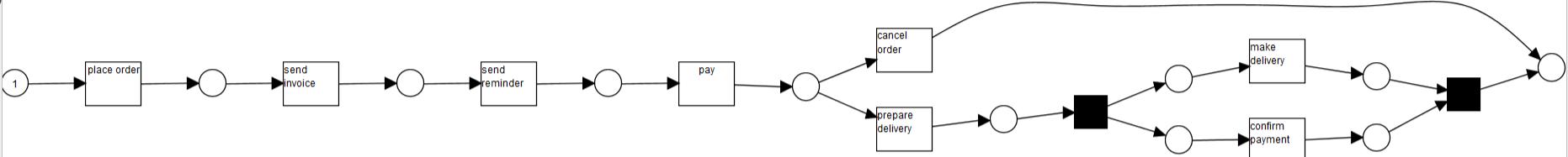
ProM: Load Event Log



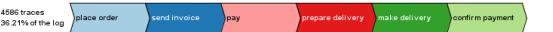
Event log



- **12,666 cases**
- **80,609 events**
- **8 unique activities**



No send reminder



No send reminder



Two send reminders



Two send reminders



Two send reminders

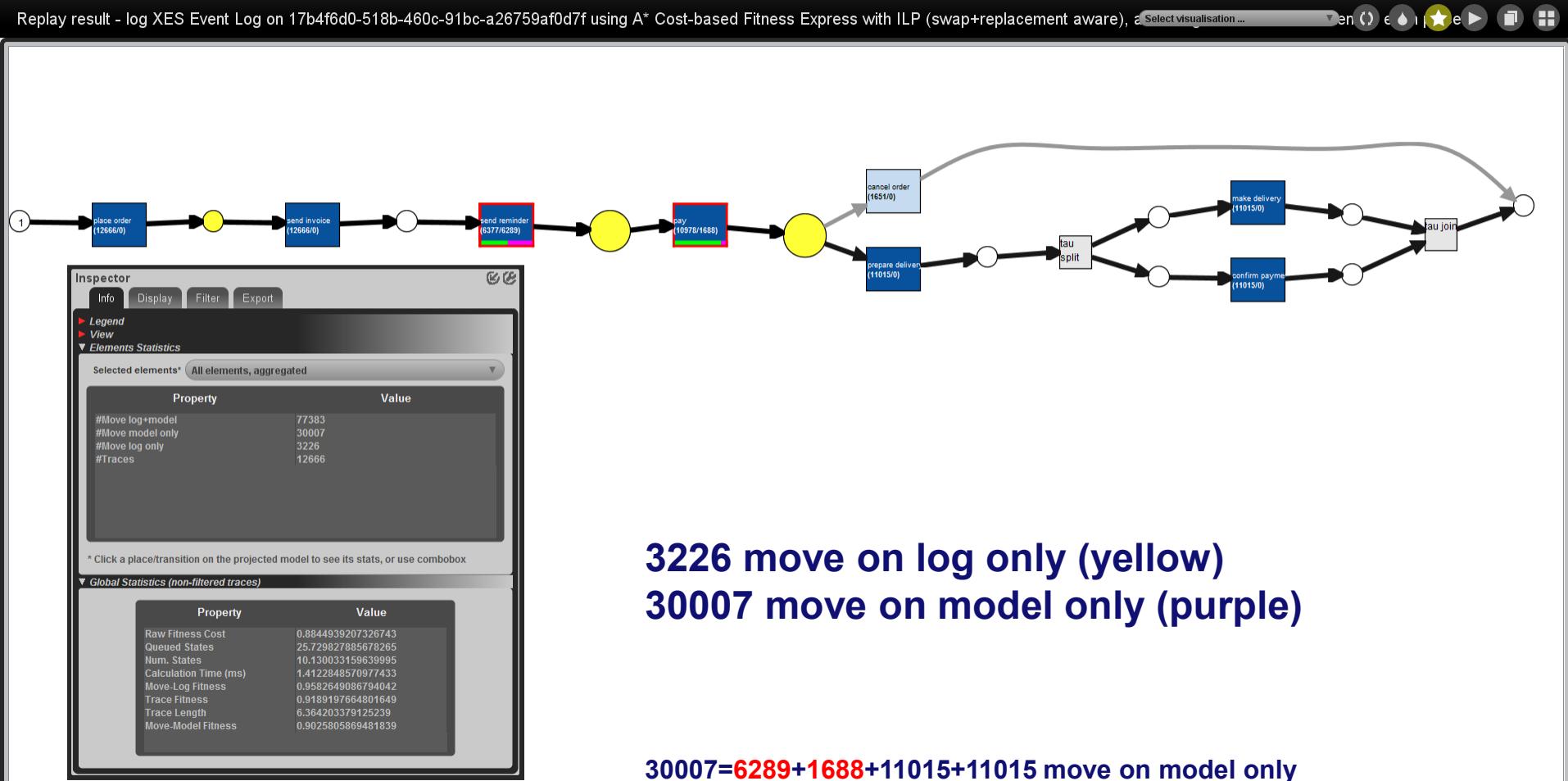


No send reminder and pay before send invoice

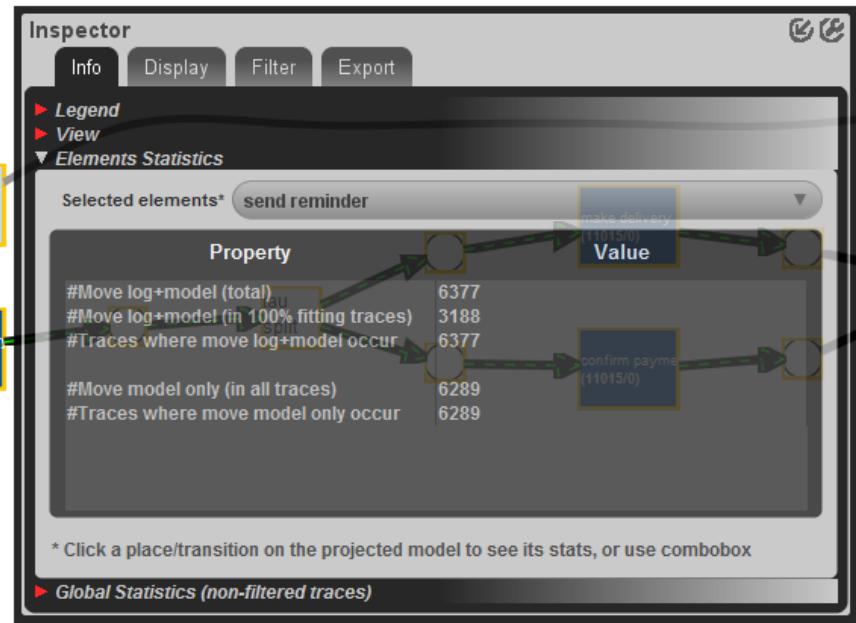
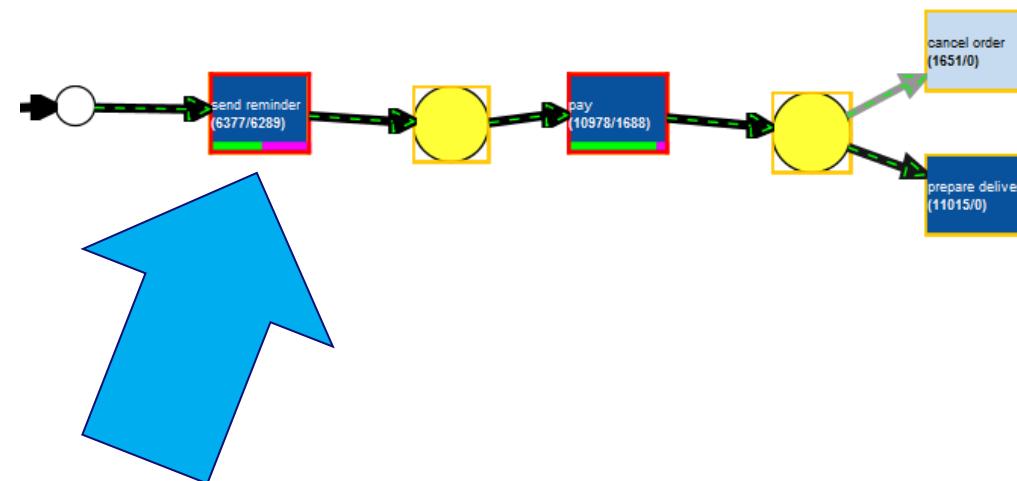


No send reminder and pay before send invoice

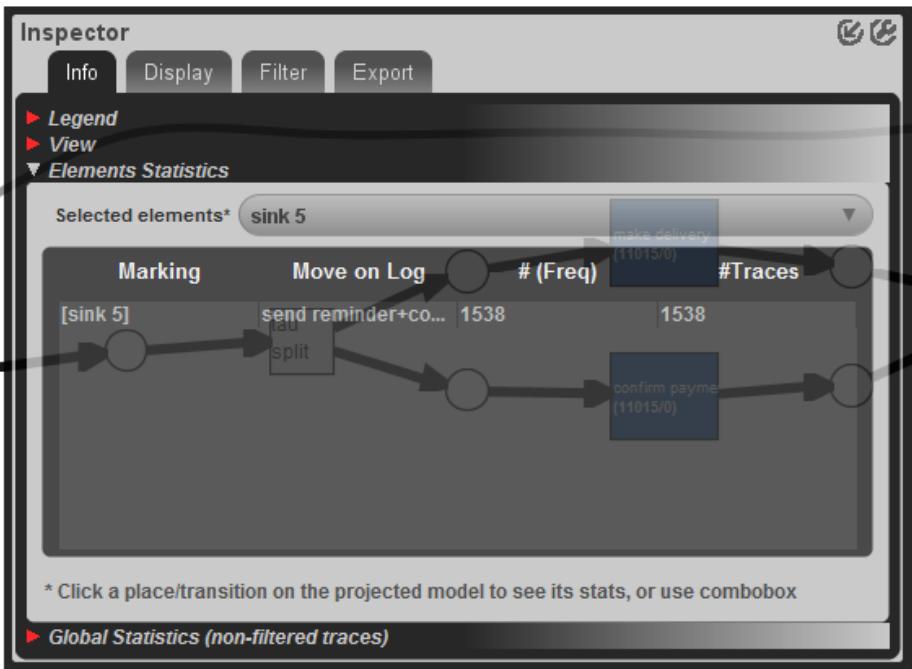
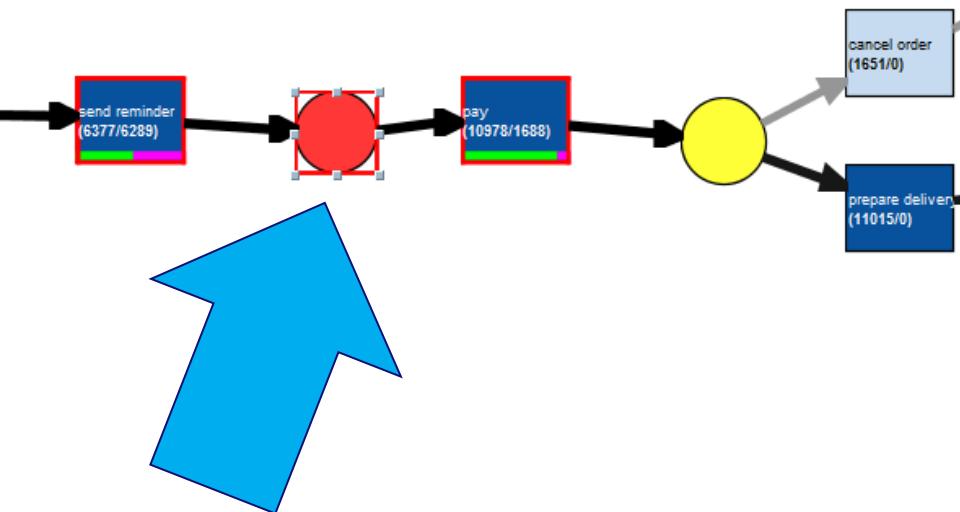




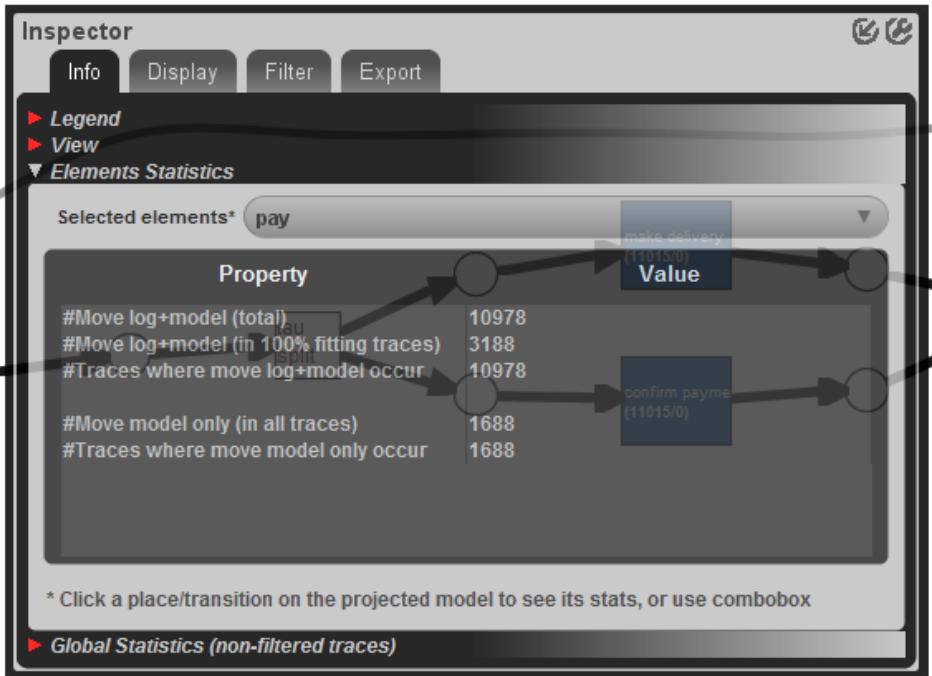
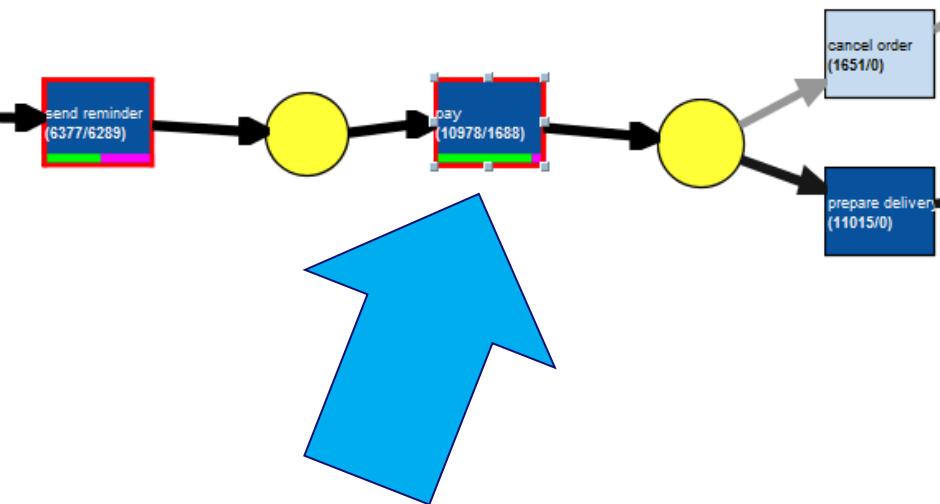
Send reminder is often skipped (6289 times)



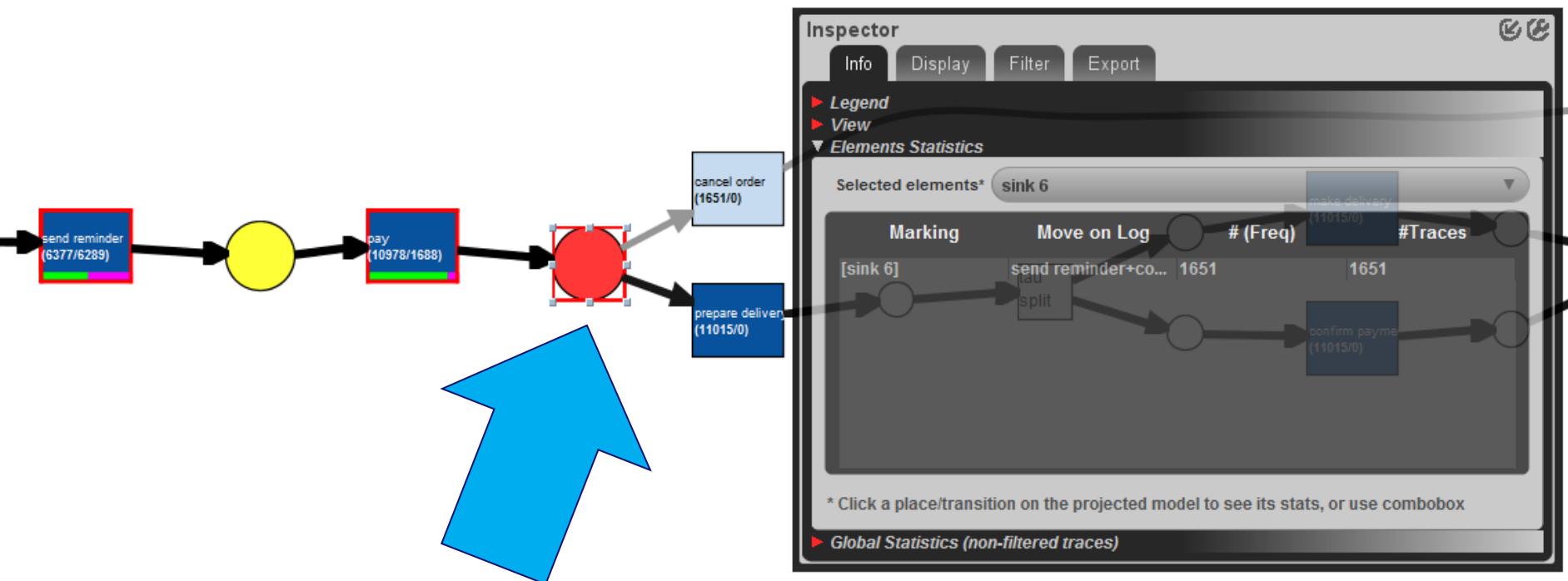
Additional send reminder (1538 times in this place)



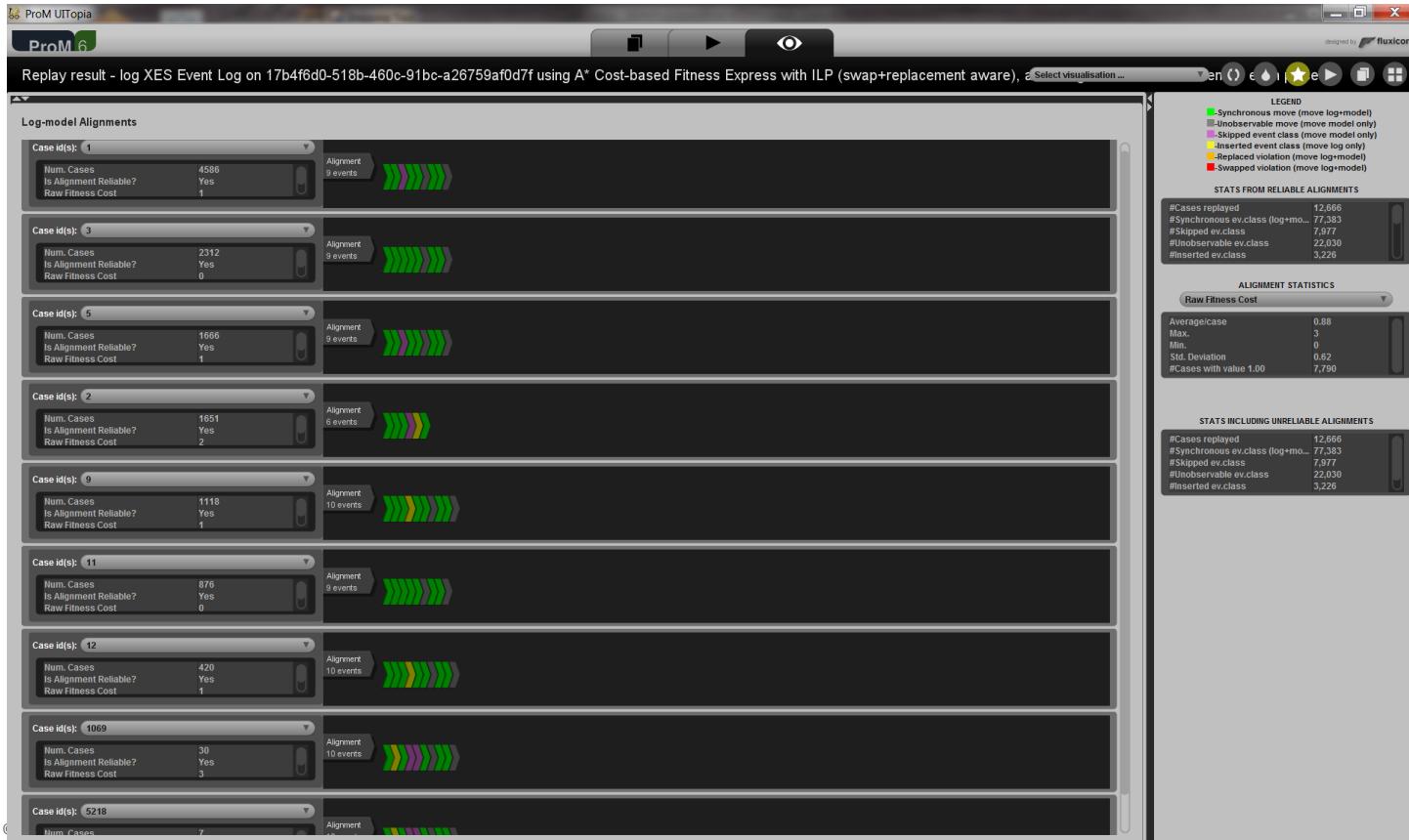
Pay is sometimes skipped (1688 times)



Additional send reminder (1651 times in this place)



Log view



synchronous move

move on model
(required activity was skipped in event log)

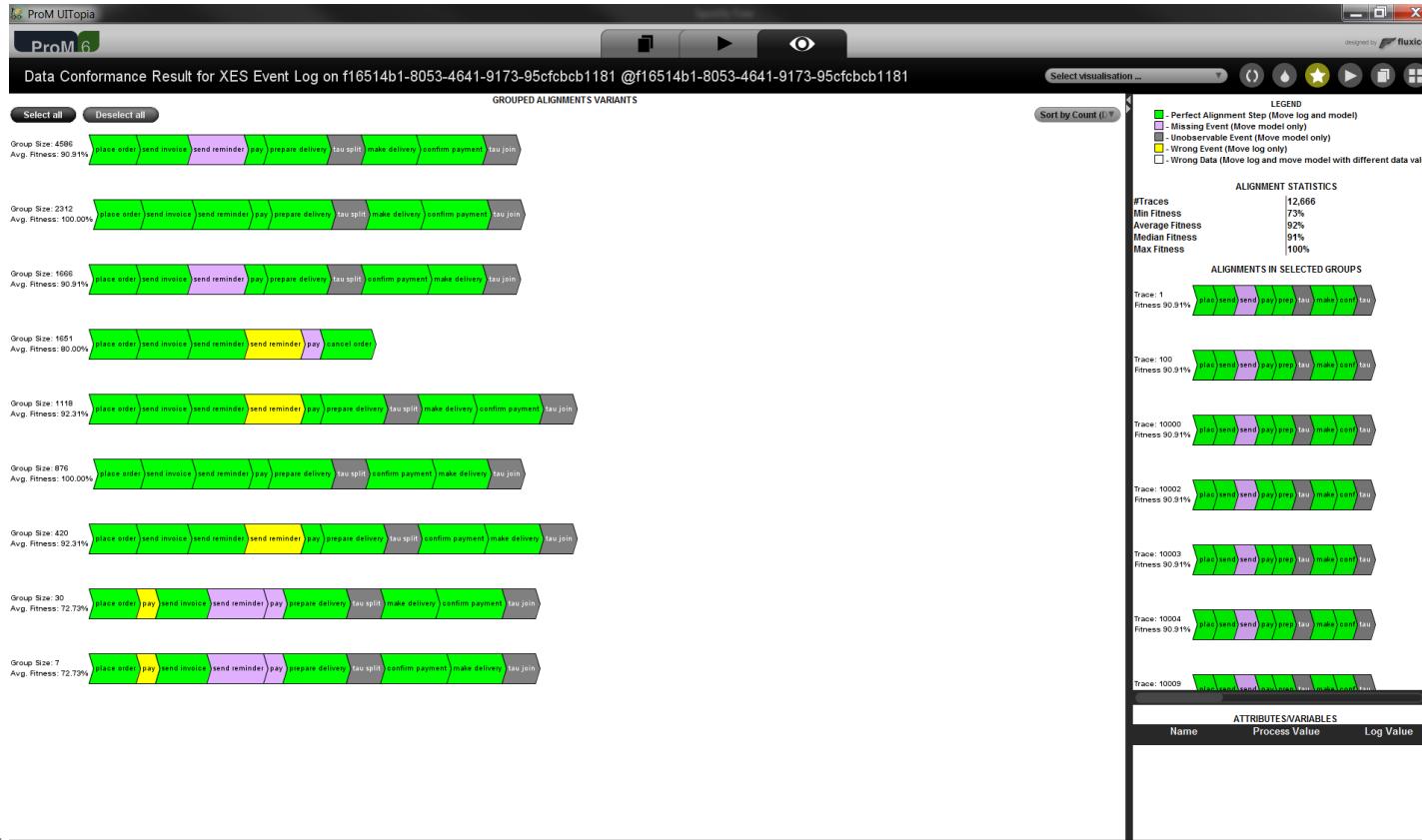
move on model
(silent transition in model was fired)

move on log
(activity in log was not possible in model)



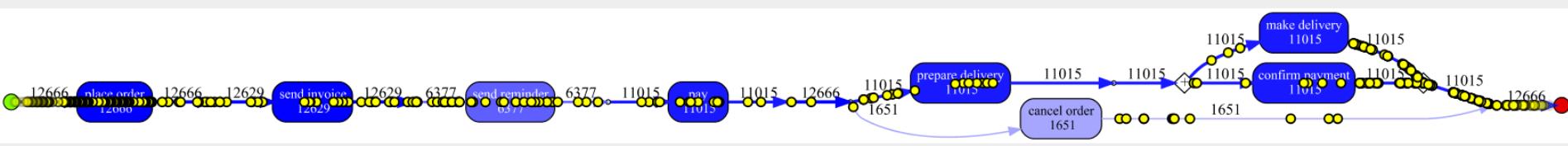
Chair of Process
and Data Science

Many more plug-ins exploring conformance (based on alignments or not)

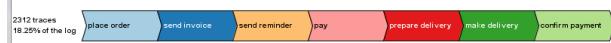


Chair of Process
and Data Science

Inductive miner



No send reminder



No send reminder



Two send reminders



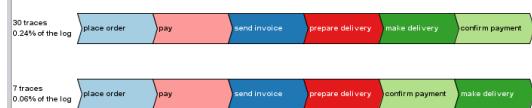
Two send reminders



Two send reminders

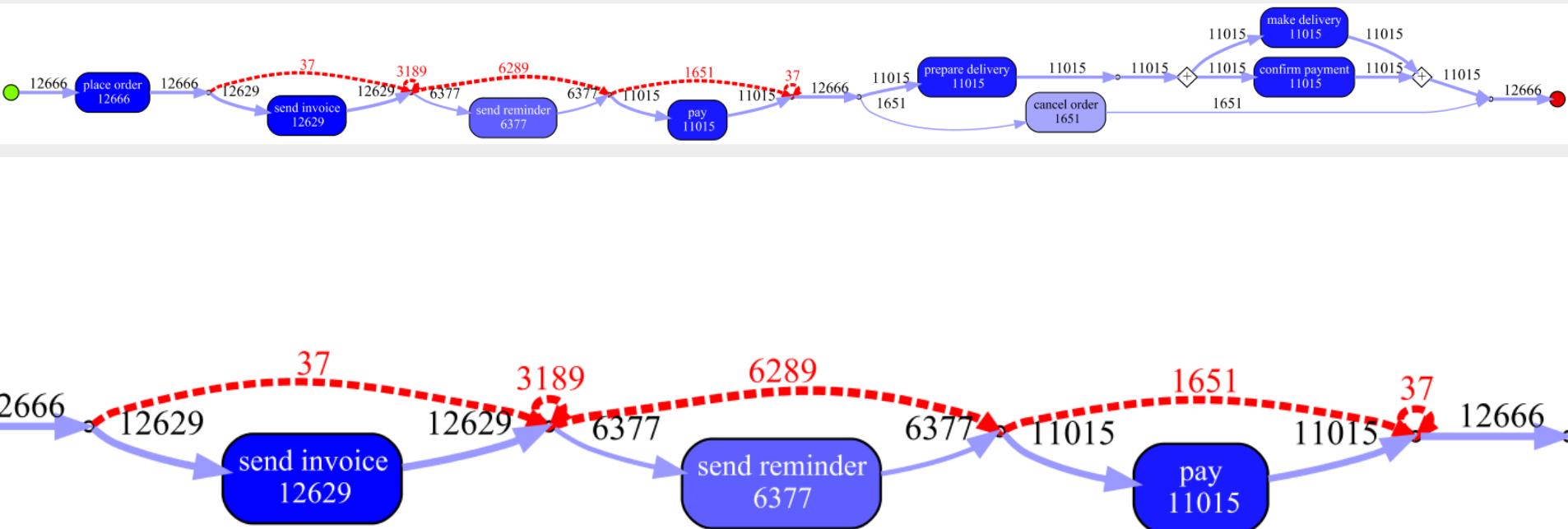


No send reminder and pay before send invoice



No send reminder and pay before send invoice

Deviations in inductive miner



Note that some numbers are different than before because of different alignments.

Drill down on deviations

ProM UItopia

Inductive visual Miner

ProM 6

10001 pla sen sen sen pay can

10005 pla sen sen sen pay can

10014 pla sen sen sen pay can

10019 pla sen sen sen pay can

10020 pla sen sen sen pay can

10027 pla sen sen sen pay can

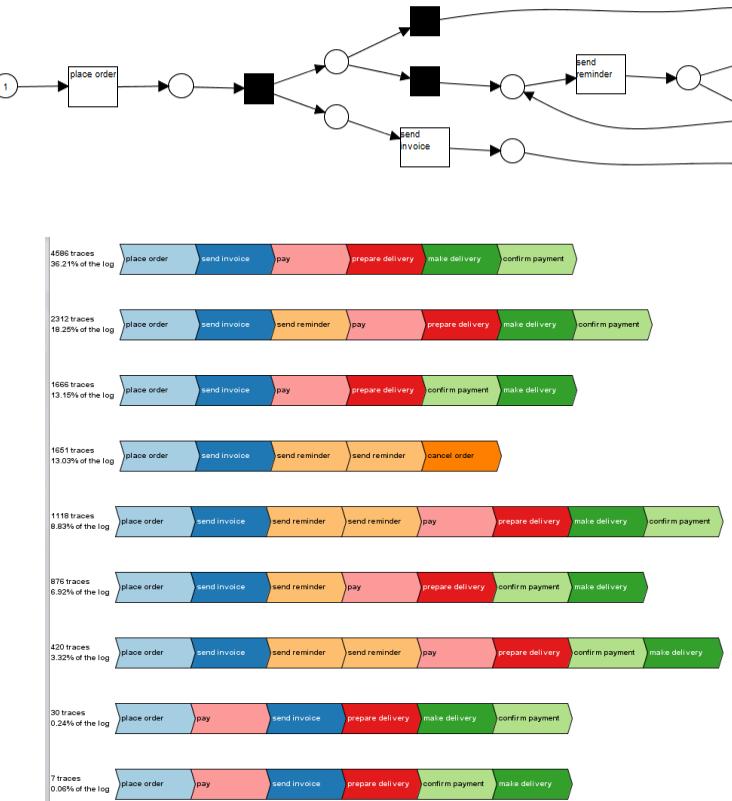
Pay is skipped 1651 times and all of these cases had an extra send reminder (these are the cases that were cancelled).

synchronous move

move on model
(required activity was skipped in event log)

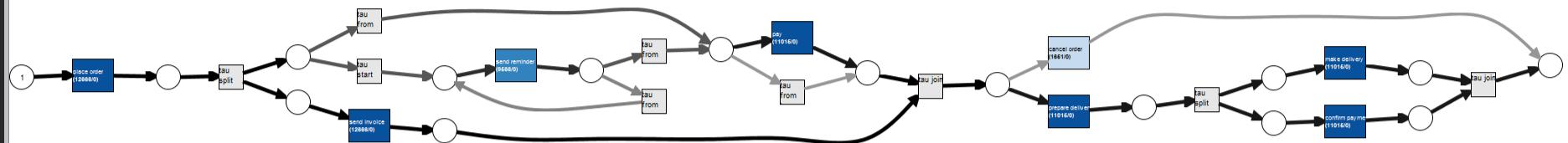
move on log
(activity in log was not possible in model)

Same event log and different (underfitting) model



Perfect replay fitness (can reproduce all traces), but underfitting (i.e., not precise enough).

Replay result - log XES Event Log on ad14ea2b-9e58-4d4c-b840-02be6fb3dff3 using A* Cost-based Fitness Express with ILP (swap+replacement aware), [Select visualisation ...](#)



Inspector

- Info
- Display
- Filter
- Export

Legend

View

Elements Statistics

Property	Value
#Move log+model	80609
#Move model only	71245
#Move log only	0
#Traces	12666

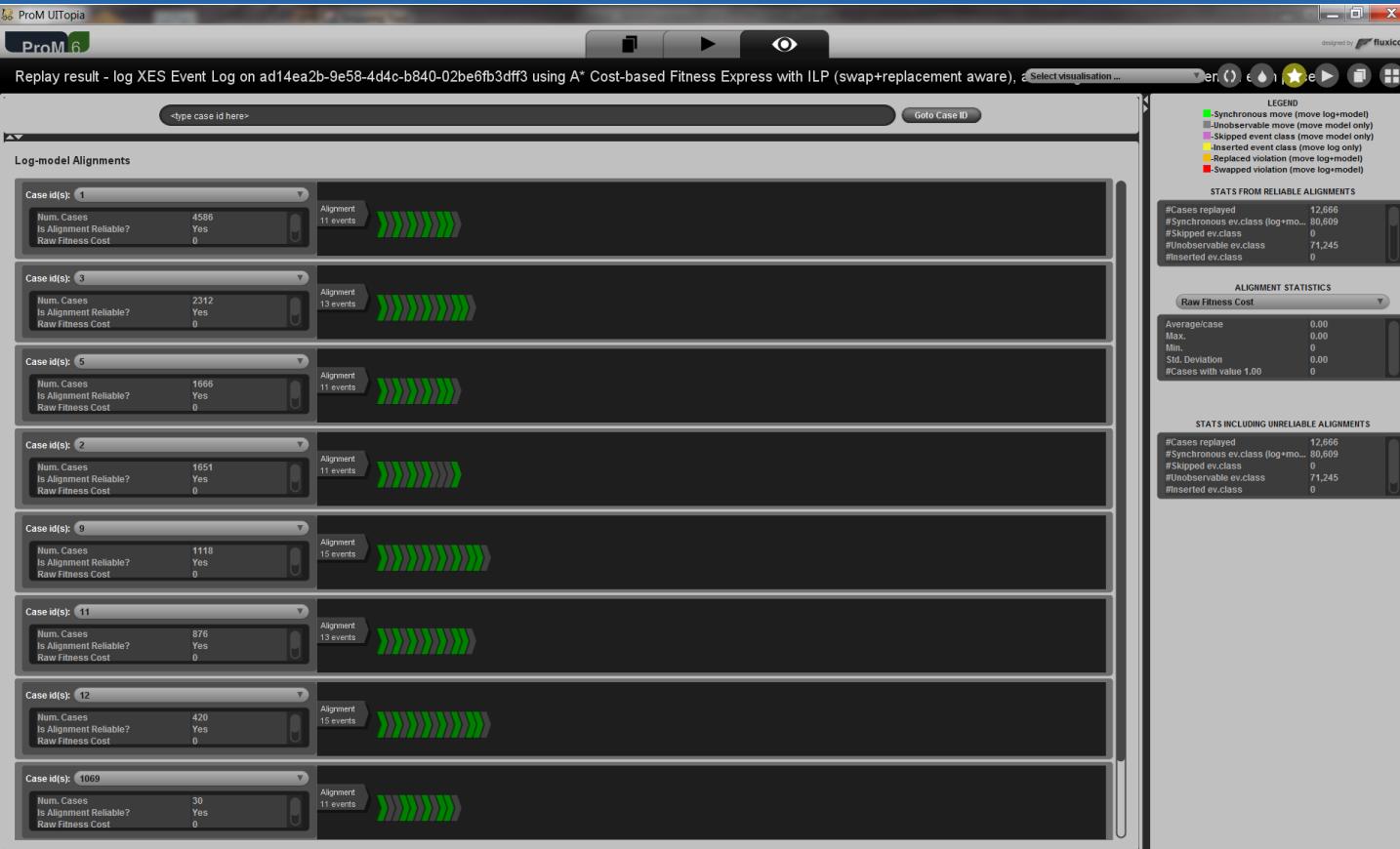
* Click a place/transition on the projected model to see its stats, or use combobox

Global Statistics (non-filtered traces)

Property	Value
Raw Fitness Cost	0.0
Queued States	45.91386388757287
Num. States	15.896652455392376
Calculation Time (ms)	8.713721774830239
Move-Log Fitness	1.0
Trace Fitness	1.0
Trace Length	6.364203379125239
Move-Model Fitness	1.0

No problems, only moves on model for silent steps.

Log view



Conformance checking in Celonis

- Input = BPMN model
- Output = List of violations
- Internally: A variant of token-based replay

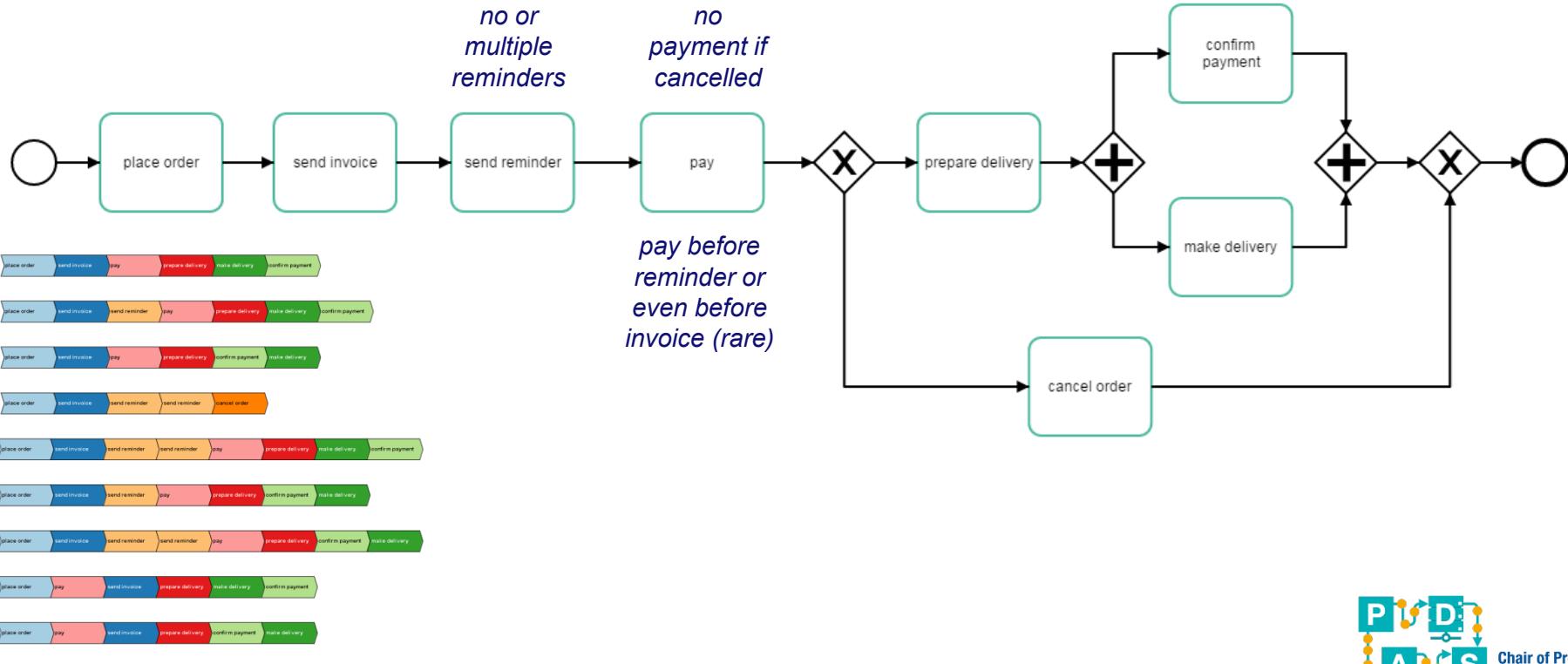
The screenshot shows the Celonis Platform interface with the 'Automation' tab selected. On the left, a sidebar lists various analysis packages: BPI 2022, Create Package, Order Handling 12666, counter example, Pizza Analysis, XXXXX, Performance Problems, Purchase-to-Pay, Purchase-to-Pay Analysis, Vesta-Cancel-Apartment, Vesta-Cancel-Apartment-Analysis, and Order Handling Analysis. The 'Order Handling 12666' package is currently active. In the center, a dashboard displays a progress bar showing 12.7k of 12.7k cases selected at 100%. Below the progress bar are four cards: 'New Sheet' (A new sheet waiting to be built), 'Process AI' (Detect and analyze deviations from the most common path), 'Process Overview' (Get the main insights on your process), and 'Process Explorer' (Analyze and understand your process). To the right of these cards is a large blue arrow pointing upwards towards the 'Conformance' card. The 'Conformance' card has the subtext: 'Compare the real process to your target process.'

This screenshot shows the same Celonis interface as above, but with a blue arrow highlighting the 'Conformance' card. Below the cards, a detailed BPMN diagram of an order handling process is shown. The process starts with a 'place order' activity, followed by a decision diamond ('pay'). From the 'pay' diamond, the process can lead to either a 'send invoice' or a 'cancel order' activity. Both of these activities then lead to another decision diamond ('send reminder'). From this diamond, the process can lead to either a 'prepare delivery' or another 'cancel order' activity. Finally, both paths converge at a third decision diamond ('rate delivery'), which leads back to the initial 'place order' activity.



Chair of Process
and Data Science

BPMN model (hand-made)



13k of 13k cases selected

100%



EDIT

Reset selections

Conformance
Overview

Whitelist

Edit process model

KPIs

Conformance overview

Timeframe

All time

From

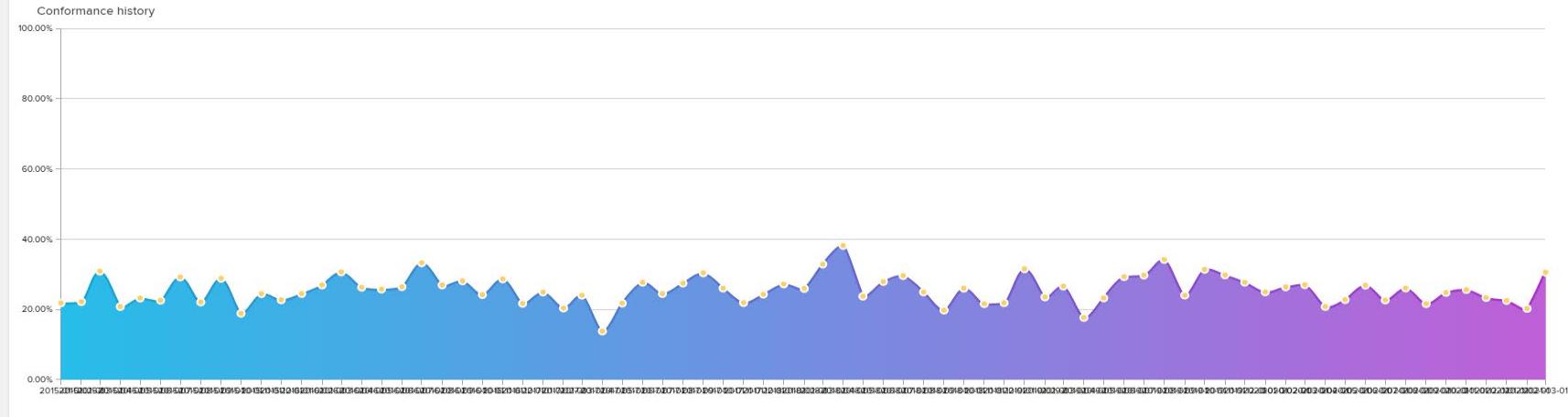
2015-01-04

To

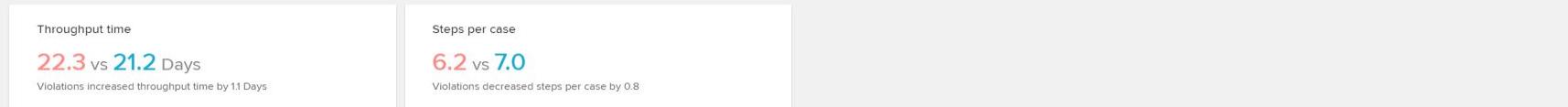
2021-04-27



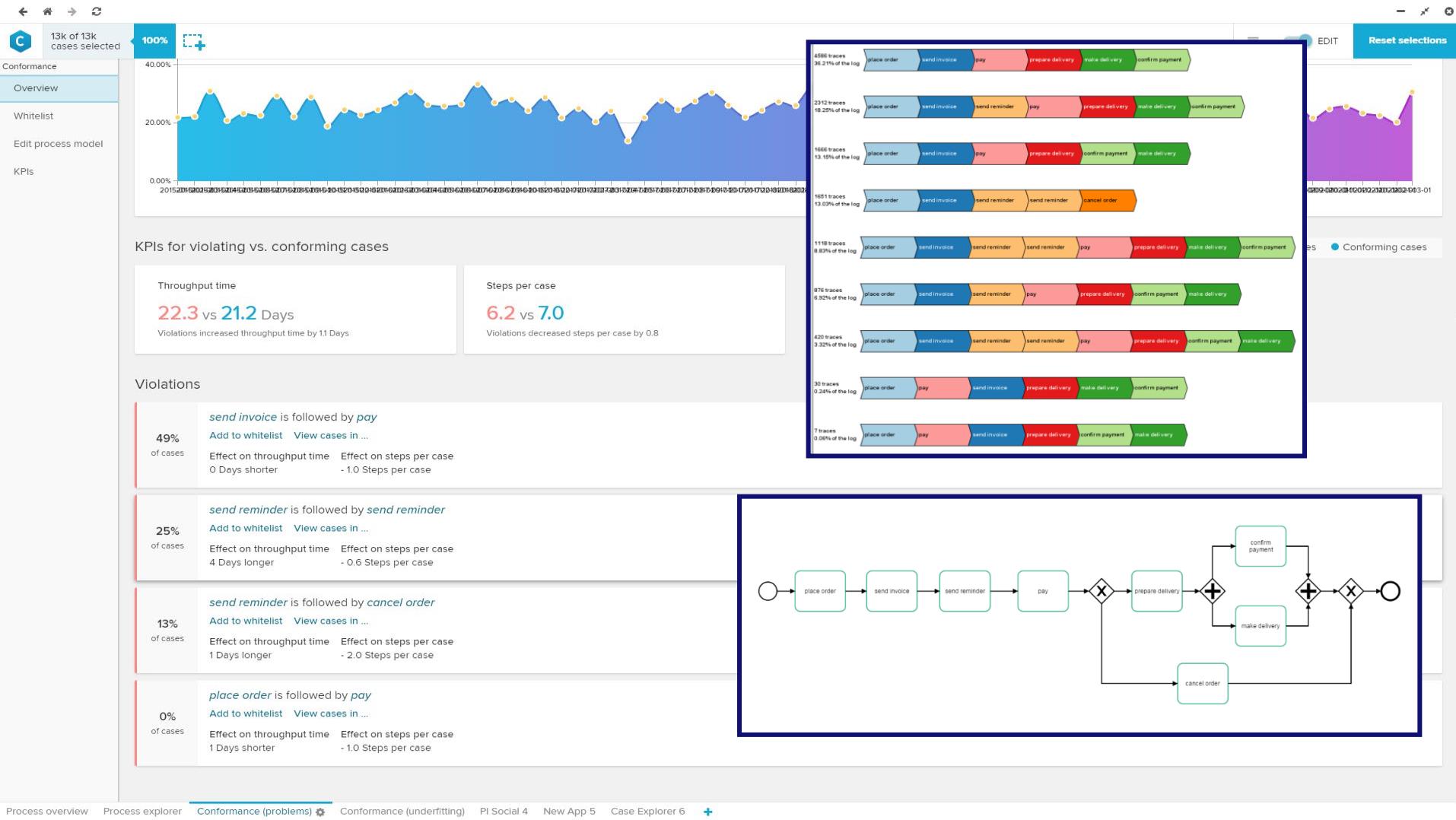
Statistics about conformance

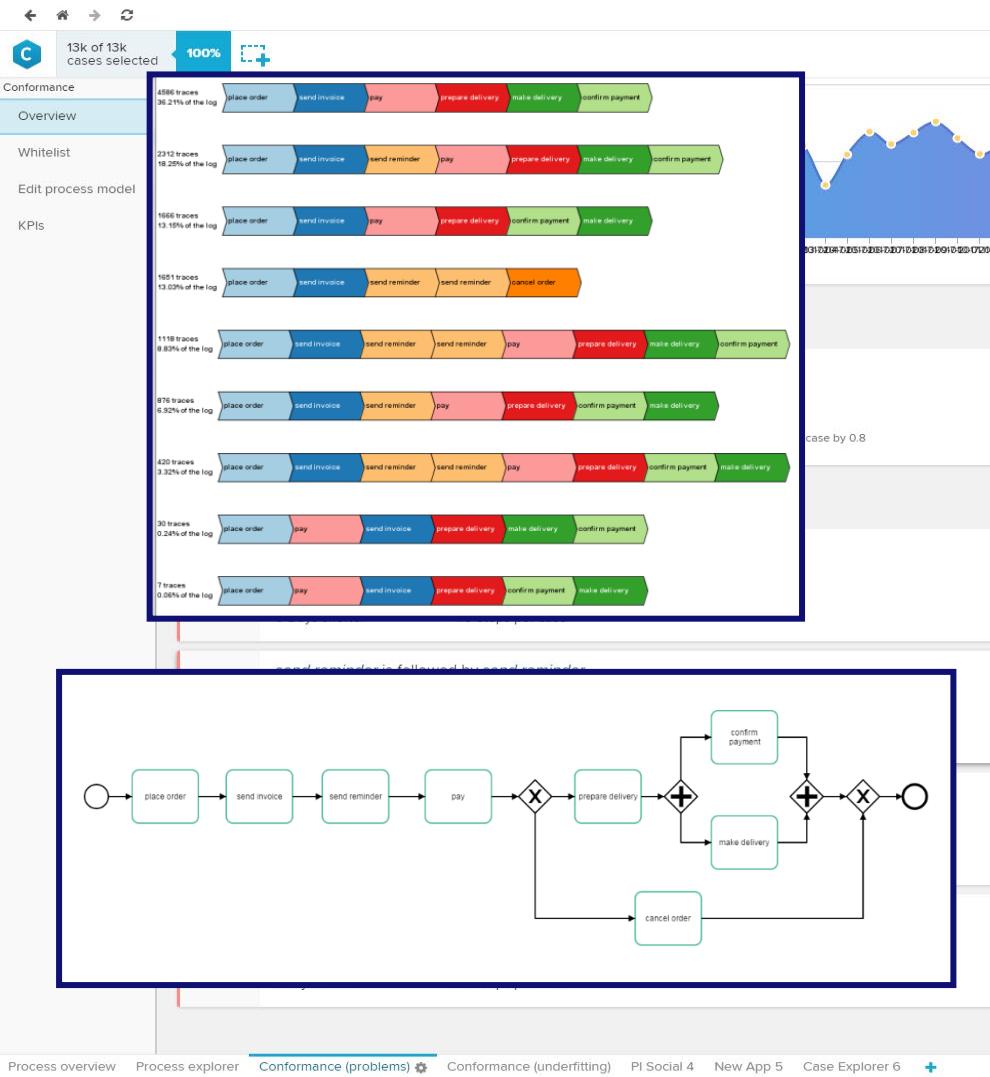


KPIs for violating vs. conforming cases

● Violating cases ● Conforming cases

Violations





Violations

send invoice is followed by *pay*

49%
of cases

Add to whitelist [View cases in ...](#)
Effect on throughput time 0 Days shorter
Effect on steps per case - 1.0 Steps per case

send reminder is followed by *send reminder*

25%
of cases

Add to whitelist [View cases in ...](#)
Effect on throughput time 4 Days longer
Effect on steps per case - 0.6 Steps per case

send reminder is followed by *cancel order*

13%
of cases

Add to whitelist [View cases in ...](#)
Effect on throughput time 1 Days longer
Effect on steps per case - 2.0 Steps per case

place order is followed by *pay*

0%
of cases

Add to whitelist [View cases in ...](#)
Effect on throughput time 1 Days shorter
Effect on steps per case - 1.0 Steps per case

Drill-down on most rare violation (37 cases)

37 of 13k cases selected 0% Violation: place order → pay 1

EDIT Reset selections

Activities

Connections

```
graph TD; Start((Process Start)) -- 37 --> PlaceOrder[place order]; PlaceOrder -- 37 --> Pay[pay]; Pay -- 37 --> SendInvoice[send invoice]; SendInvoice -- 37 --> PrepareDelivery[prepare delivery]; PrepareDelivery -- 31 --> MakeDelivery[make delivery]; MakeDelivery -- 6 --> ProcessEnd((Process End)); PlaceOrder -- 37 --> Pay; Pay -- 6 --> ConfirmPayment[confirm payment]; ConfirmPayment -- 9 --> ProcessEnd; ConfirmPayment -- 31 --> MakeDelivery; MakeDelivery -- 6 --> ProcessEnd; MakeDelivery -- 28 --> ConfirmPayment;
```

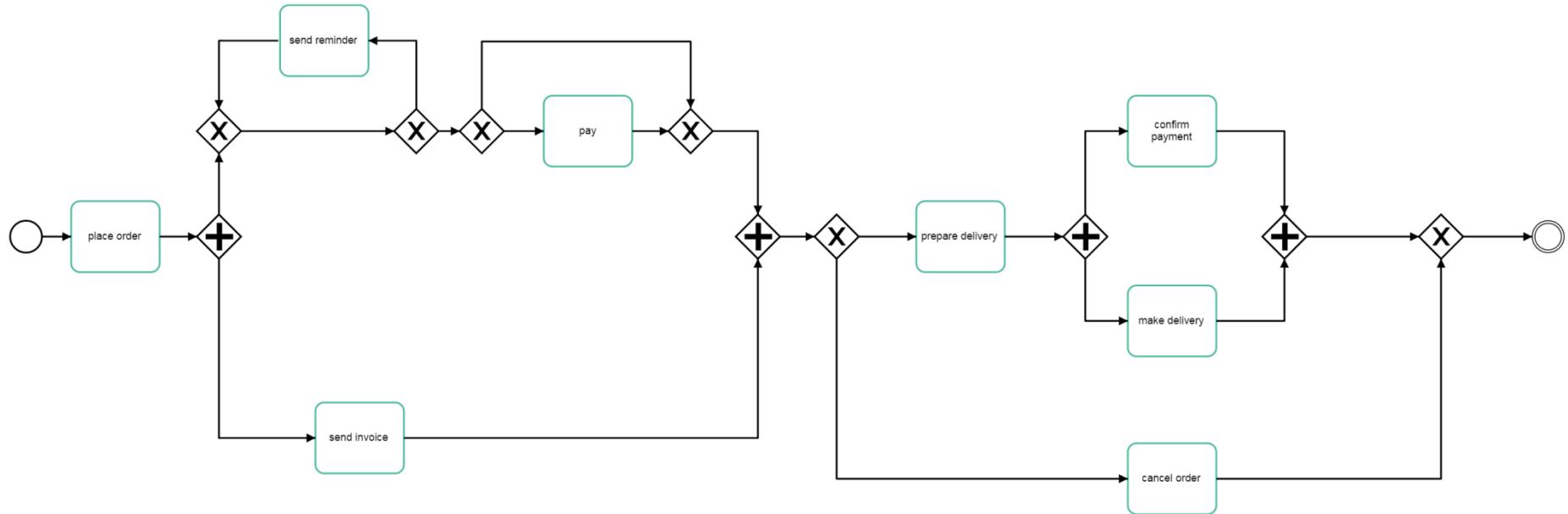
place order is followed by pay
Add to whitelist View cases in ...

Effect on throughput time 1 Days shorter

Effect on steps per case - 1.0 Steps per case

“Almost” perfectly fitting model

(not expecting any deviations)



Conformance

Overviews

Whitelists

Edit process

KPIs

place order ▾
Tue, Aug 4, 2015 12:15 PM -14d

pay ▾
Mon, Aug 17, 2015 9:31 AM -1d

prepare delivery ▾
Tue, Aug 18, 2015 4:33 PM 0

case	1133
end time	Tue, Aug 18, 2015 4:41 PM
resource	Sophia
product	SAMSUNG Galaxy S4
prod-price	3290
quantity	5
address	NL-7943MC-4

send invoice ▾
Tue, Aug 18, 2015 4:33 PM 0

case	1133
end time	Tue, Aug 18, 2015 4:41 PM
resource	Lily
product	SAMSUNG Galaxy S4
prod-price	3290
quantity	5
address	NL-7943MC-4

make delivery ▾
Wed, Aug 19, 2015 4:41 PM +1d

confirm payment ▾
Thu, Aug 20, 2015 3:47 PM +2d

Timeframe
All time From 2015-01-04 To 2021-04-27

g cases /S 4.00

Violations 1 found in process model

Whitelisted violations 0 configured in whitelist

4 cases where prepare delivery and send invoice have exactly the same timestamp

Violations

0.0

100% of cases

misleading diagnostics (send invoice is put after prepare delivery)

pay is followed by **prepare delivery**

Add to whitelist [View cases in ...](#)

Effect on throughput time 18 Days longer

Effect on steps per case + 6.0 Steps per case

pay is followed by prepare delivery

Process overview Process explorer Conformance (problems) Conformance (underfitting) PI Social 4 New App 5 Case Exp

Process Adherence Management (PAM): Discover Object-Centric BPMN

The screenshot shows the Celonis Model Miner interface within a web browser window titled "junk | OrderManagement4OT.pml". The main area displays a BPMN-like process flow for Order Management, featuring various objects like Sales Order, Sales Order Item, Delivery Item, and Customer. The flow consists of several parallel tracks and decision points, primarily colored purple and teal. A large blue banner at the top right of the process area reads "Uses the Inductive Mining algorithm". On the left side, there's a sidebar with various icons and sections such as "Mine a model that is closest to your target process", "Select object types", "Set the data scope", "Exclude incomplete cases", and "Filter object attributes". The bottom right corner of the process area has a graduation cap icon with a red '1'.

Studio > playground-ordermanagement > OrderManagement4OT

Search CTRL + / 1 Publish

junk Model Miner

Mine a model that is closest to your target process

Build the process model by selecting variants that are close to your target process. In the next step you will be able to edit this model to create your final model.

Select object types Select variants

10 (of 418) most frequent variants selected

Custom

Set the data scope

Exclude incomplete cases

Focus on complete cases to eliminate deviations caused by flows that are still in progress. Define complete cases by selecting valid start and end events for object types.

Filter object attributes

If your process model has a specific scope (e.g. region or product), please set your filters here.

Sales Order

Create Sales Order Header

Create Sales Order Item

Approve Sales Order

Set Delivery Block

Approve Sales Order Item

Delivery

Create Delivery Header

Create Delivery Item

Post Goods Issue

Sign Proof Of Delivery

Create Customer

Change Sales Order Item

Delivery Item

Legend

Save & Continue Cancel

Uses the Inductive Mining algorithm

Process Adherence Management (PAM): Check Conformance

Ordermanagement4OT | OrderManagement4OT

Studio > playground-ordermanagement > OrderManagement4OT

Search CTRL + F 1 Publish

Deviation Explorer

Conformance Rate: Sales Order 99.99%, Delivery 99.57%

Actual behavior (4)

- Violated exclusive gateway: Create Customer Invoice, 2.81% of objects, TPT impact -7 days, Sales Order Item
- Violated exclusive gateway: Post Goods Issue, 1.55% of objects, TPT impact -24 hours, Delivery Item
- Violated exclusive gateway: Sign Proof Of Delivery, 0.43% of objects, TPT impact -11 days, Delivery
- Occurred too often: Set Delivery Block, less than 0.01% of objects, TPT impact +9 days, Sales Order

Target Model

Filters

Use filters to select the data you want to compare to the target model.

+ Add a filter

100% Sales Order

Computes alignments

Legend: Sales Order (blue), Sales Order Item (purple), Delivery (green), Delivery Item (yellow)

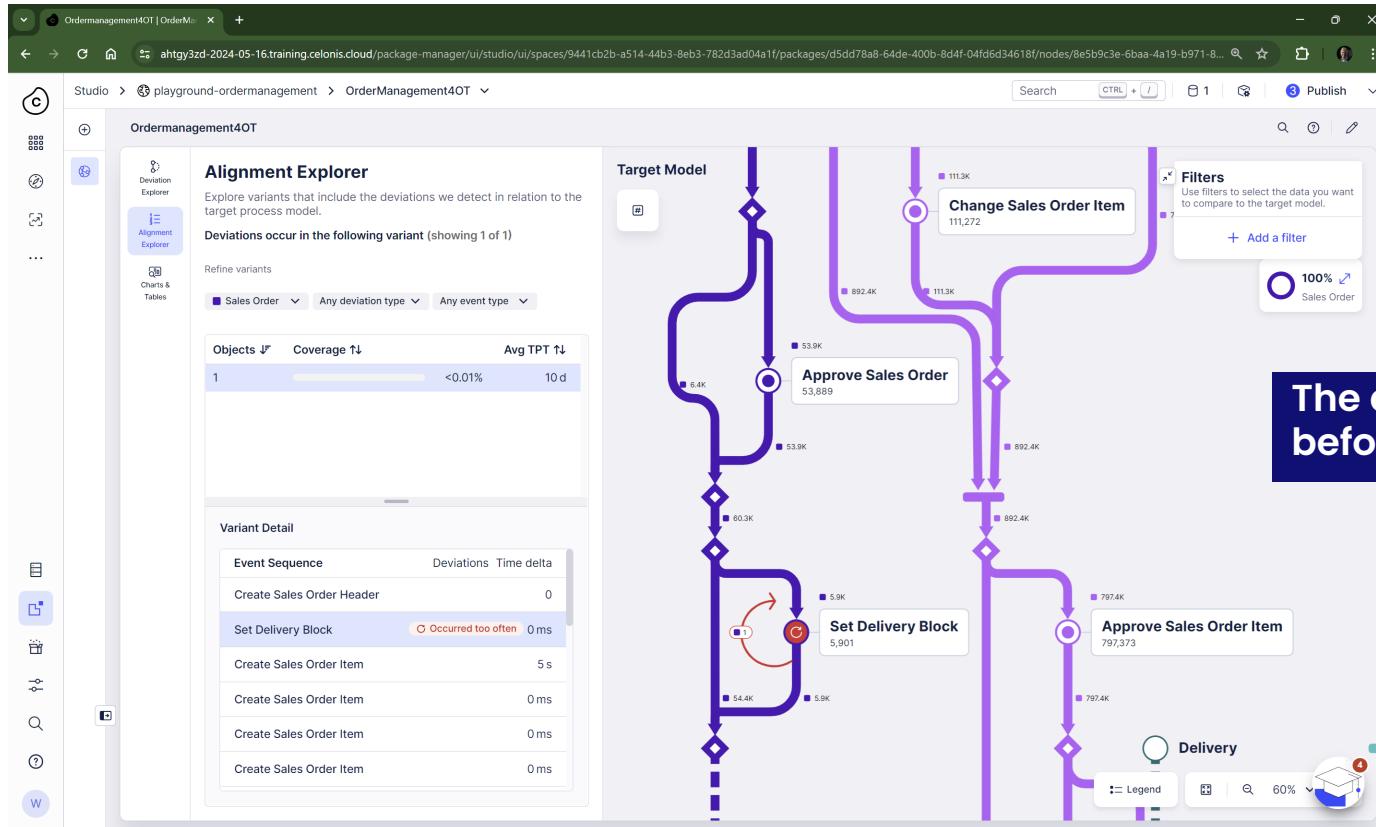
50%

© Wil van der Aalst (use only with permission & acknowledgements)



Chair of Process
and Data Science

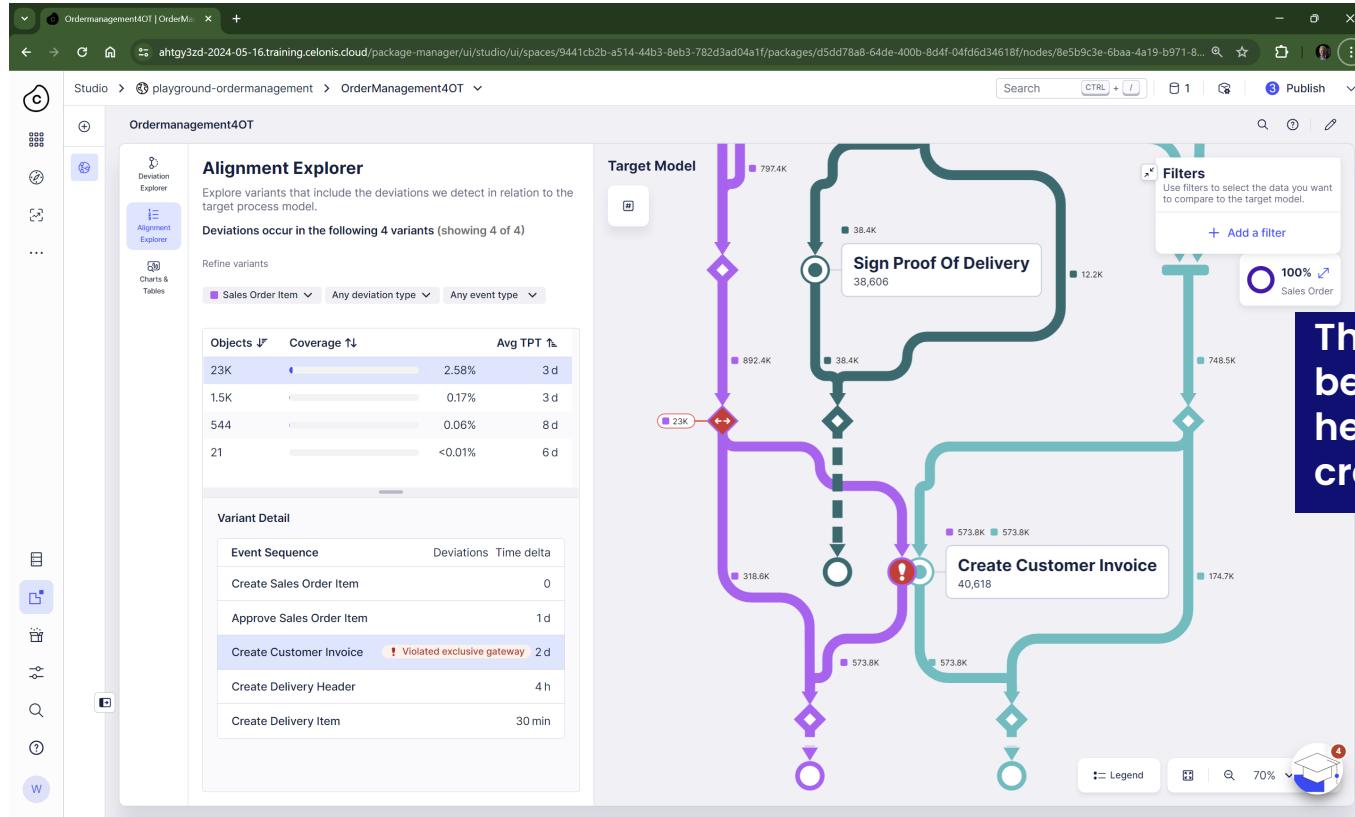
Process Adherence Management (PAM): Show Alignments



The delivery block is set before items are created



Process Adherence Management (PAM): Show Alignments



Process Adherence Management (PAM): Analyze Performance

The screenshot shows the OrderManagement4OT studio interface with the following components:

- Deviation Explorer:** Displays conformance rates for Sales Order (99.99%) and Delivery (99.57%). It also lists actual behaviors: "Violated exclusive gateway Create Customer Invoice" (2.81% objects) and "Violated exclusive gateway Post Goods Issue" (1.55% objects).
- Target Model:** A process flow diagram showing the target model for the Order Management process.
- Actual behavior:** A detailed process flow diagram showing the actual behavior compared to the target model, with nodes like "Create Sales Order Header", "Create Customer Invoice", "Delivery Item", and "Sales Order Item".
- Throughput Time:** A detailed analysis of throughput time for "Create Sales Order Header" and "Create Customer Invoice". It shows average and median times (avg. 14.62 Days, med. 12 Days), a histogram of times from 2 to 67 days, and a count of 40,618 Sales Order objects selected.

Compute the time between the creation of the order and the delivery of all items in the order



Demo



Applications of Process Mining



Process Mining Is Used Everywhere

Technology



Financial Services & Insurance



Life Sciences & Chemicals



Consumer & Retail



Manufacturing



Telecommunications & Media



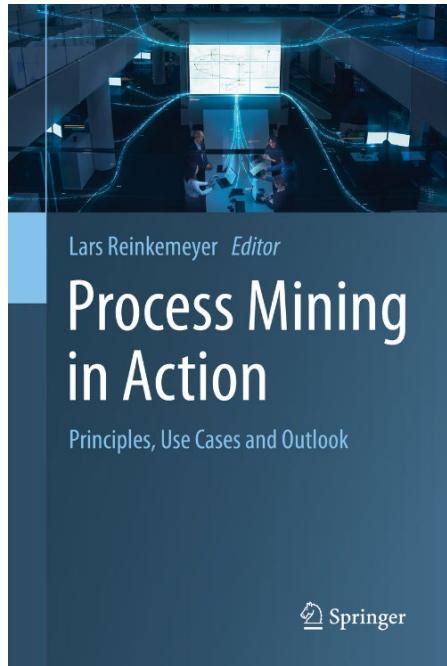
Energy & Utilities



Oil & Gas



Some Case Studies



Part II Best Practice Use Cases

- 9 **Siemens: Driving Global Change with the Digital Fit Rate in Order2Cash** 49
Gia-Thi Nguyen
- 10 **Uber: Process Mining to Optimize Customer Experience and Business Performance** 59
Martin Rowson
- 11 **BMW: Process Mining @ Production** 65
Patrick Lechner
- 12 **Siemens: Process Mining for Operational Efficiency in Purchase2Pay** 75
Khaled El-Wafi
- 13 **athenahealth: Process Mining for Service Integrity in Healthcare** 97
Corey Balint, Zach Taylor, and Emily James
- 14 **EDP Comercial: Sales and Service Digitization** 109
Ricardo Henrique
- 15 **ABB: From Mining Processes Towards Driving Processes** 119
Heymen Jansen
- 16 **Bosch: Process Mining—A Corporate Consulting Perspective** 129
Christian Buhrmann
- 17 **Schukat: Process Mining Enables Schukat Electronic to Reinvent Itself** 135
Georg Schukat
- 18 **Siemens Healthineers: Process Mining as an Innovation Driver in Product Management** 143
Jutta Reindler
- 19 **Bayer: Process Mining Supports Digital Transformation in Internal Audit** 159
Arno Boenner
- 20 **Tekom: Process Mining in Shared Services** 169
Gerrit Lillig



Chair of Process
and Data Science

Example: Siemens Order-to-Cash (O2C)

Siemens: Driving Global Change with the Digital Fit Rate in Order2Cash

Gia-Thi Nguyen

“Using out-of-the-box reports from the Process Mining tool, the results included an increase in automation by 24% and a reduction in manual rework by 11%.”

“Over 70 million sales order items across data from 90 countries with over 1.5 million process variants.”

# Sales Order Items	70,286,004	Statistical Transactional Value in EUR	232,797,668,141
# Activities	411,462,971	# Process Variants	1,511,644
Digital FIT Rate	2.18	SAP Systems	28
Automation Rate	63%	# Countries	90
Rework Rate	36%	# AREs	255
eBiz Rate All-in	64%	# Customers	257,236
Total Cycle Time (average)	48 Days	# Materials	1,728,677



Chair of Process
and Data Science

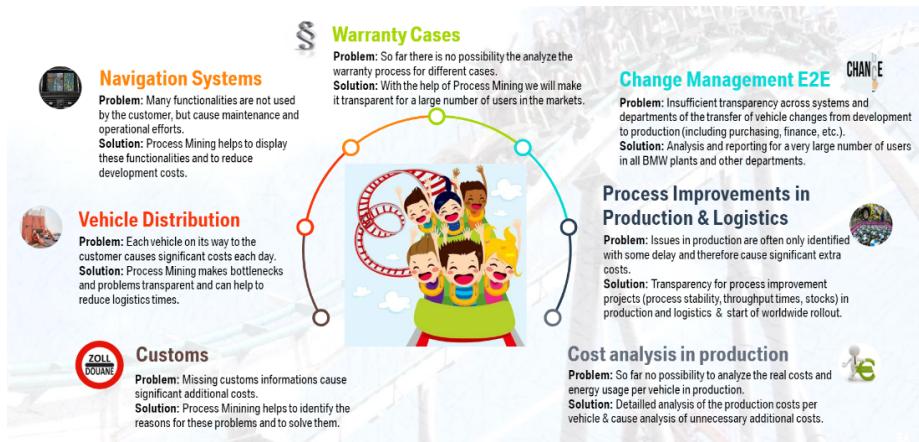
Example: BMW

BMW: Process Mining @ Production

11

Bringing Innovation to Production Processes and Beyond

Patrick Lechner



More than **850 registered users** on BMW Celonis Process Mining Infrastructure

49 of BMW's 50 data models are in use

More than **10 terabyte of raw relational data** is handled on the Process Mining Databases

More than **6.4 million process variants** are currently being analyzed within the BMW Process Mining Infrastructure

More than **500 million events** are analyzed on the BMW Celonis Process Mining Infrastructure

More than **30 million cases** are analyzed on the BMW Celonis Process Mining Infrastructure

Between **400 – 600 analytical views** are provided **each day** for the Process Mining Users of BMW

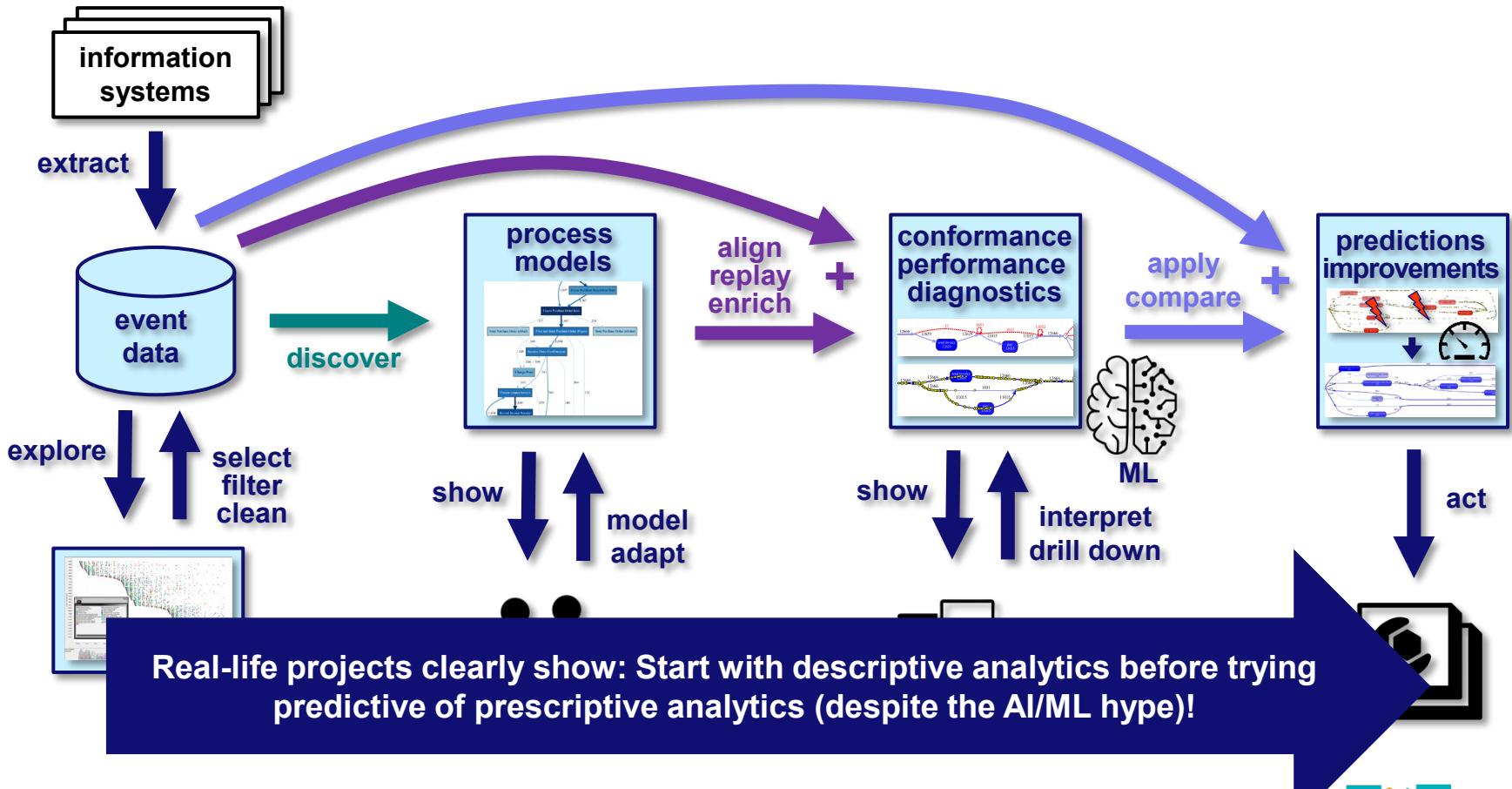
600 unique analyses are currently available for the 49 data models



Chair of Process
and Data Science

Example: Lufthansa





Part I: Introduction

Chapter 1
Data Science in Action

Chapter 2
Process Mining:
The Missing Link

Part II: Preliminaries

Chapter 3
Process Modeling
and Analysis

Chapter 4
Data Mining

Part III: From Event Logs to Process Models

Chapter 5
Getting the Data

Chapter 6
Process Discovery:
An Introduction

Chapter 7
Advanced Process
Discovery Techniques

Part IV: Beyond Process Discovery

Chapter 8
Conformance
Checking

Chapter 9
Mining Additional
Perspectives

Chapter 10
Operational Support

Part V: Putting Process Mining to Work

Chapter 11
Process Mining
Software

Chapter 12
Process Mining in the
Large

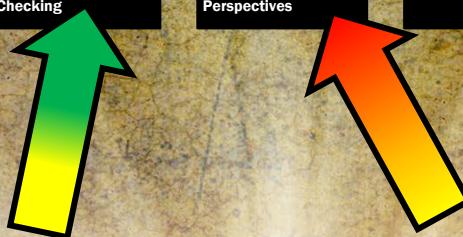
Chapter 13
Analyzing “Lasagna
Processes”

Chapter 14
Analyzing “Spaghetti
Processes”

Part VI: Reflection

Chapter 15
Cartography and
Navigation

Chapter 16
Epilogue



ID	Topic	Date	Date	Place
	Lecture 1 Introduction to Process Mining	08.04.24	Monday	AH V
	Lecture 2 Data Science: Supervised Learning	09.04.24	Tuesday	AH V
	<i>Exercise 1 Tool Introduction</i>	09.04.24	Tuesday	AH III
	Lecture 3 Data Science: Unsupervised Learning and Evaluation	15.04.24	Monday	AH V
	Lecture 4 Introduction to Process Discovery	16.04.24	Tuesday	AH V
	<i>Exercise 2 Data Mining</i>	16.04.24	Tuesday	AH III
	Lecture 5 Alpha Algorithm 1	22.04.24	Monday	AH V
	Lecture 6 Alpha Algorithm 2	23.04.24	Tuesday	AH V
	<i>Exercise 3 Petri Nets</i>	23.04.24	Tuesday	AH III
	Lecture 7 Model Quality Representation	29.04.24	Monday	AH V
	Lecture 8 Heuristic Mining	30.04.24	Tuesday	AH V
	<i>Exercise 4 Alpha Miner</i>	30.04.24	Tuesday	AH III
	Lecture 9 Region-Based Mining	06.05.24	Monday	AH V
	<i>Exercise 5 Heuristic Mining and Region-Based Mining</i>	07.05.24	Tuesday	AH III
	Lecture 10 Inductive Mining	13.05.24	Monday	AH V
	Lecture 11 Event Data and Exploration	14.05.24	Tuesday	AH V
	<i>Exercise 6 Inductive Mining</i>	14.05.24	Tuesday	AH III
	Lecture 12 Conformance Checking 1	27.05.24	Monday	AH V
	Lecture 13 Conformance Checking 2	28.05.24	Tuesday	AH V
	<i>Q&A Session Assignment Part I</i>	28.05.24	Tuesday	AH III
	Deadline Assignment Part I	02.06.24	Sunday	
	<i>Exercise 7 Footprint and Token-Based Replay (Exercise)</i>	03.06.24	Monday	AH V
	<i>Exercise 8 Alignments (Exercise)</i>	04.06.24	Tuesday	AH V
	Lecture 14 Decision Mining	10.06.24	Monday	AH V
	<i>Lecture 15 Celonis Guest Lecture</i>	11.06.24	Tuesday	AH V
	<i>Exercise 9 Decision Mining</i>	11.06.24	Tuesday	AH III
	Lecture 16 Performance Analysis and Organizational Mining	17.06.24	Monday	AH V
	<i>Exercise 10 Performance Analysis (Exercise)</i>	18.06.24	Tuesday	AH V
	<i>Exercise 11 Organizational Mining</i>	18.06.24	Tuesday	AH III
	<i>Exercise 12 Celonis Case Study</i>	24.06.24	Monday	AH V
	Lecture 17 Operational Support and Process Mining Applications	01.07.24	Monday	AH V
	Lecture 18 Distributed, Streaming, and Comparative Process Mining	02.07.24	Tuesday	AH V
	<i>Exercise 13 Operational Process Mining</i>	02.07.24	Tuesday	AH III
	Lecture 19 Closing	08.07.24	Monday	AH V
	<i>Q&A Session Assignment Part II</i>	09.07.24	Tuesday	AH III
	Deadline Assignment Part II	14.07.24	Sunday	
	<i>Q&A Session Exam</i>	16.07.24	Tuesday	AH III



Decision Mining

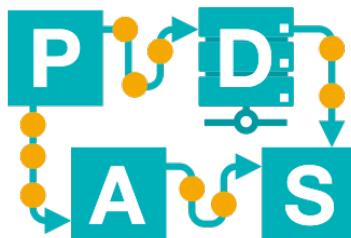
Lecture 14

prof.dr.ir. Wil van der Aalst

www.vdaalst.com @wvdaalst

www.pads.rwth-aachen.de

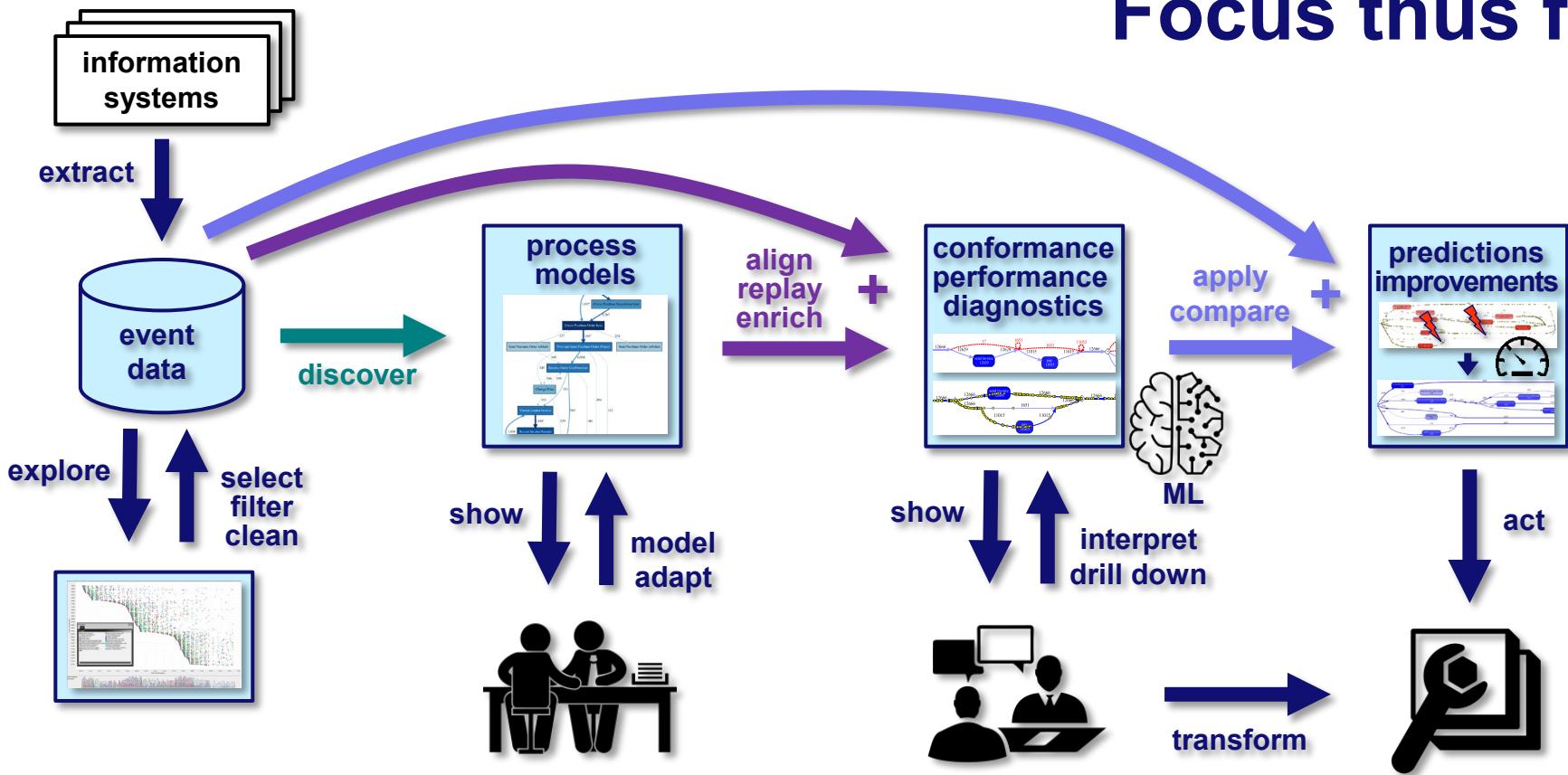
BPI-L14



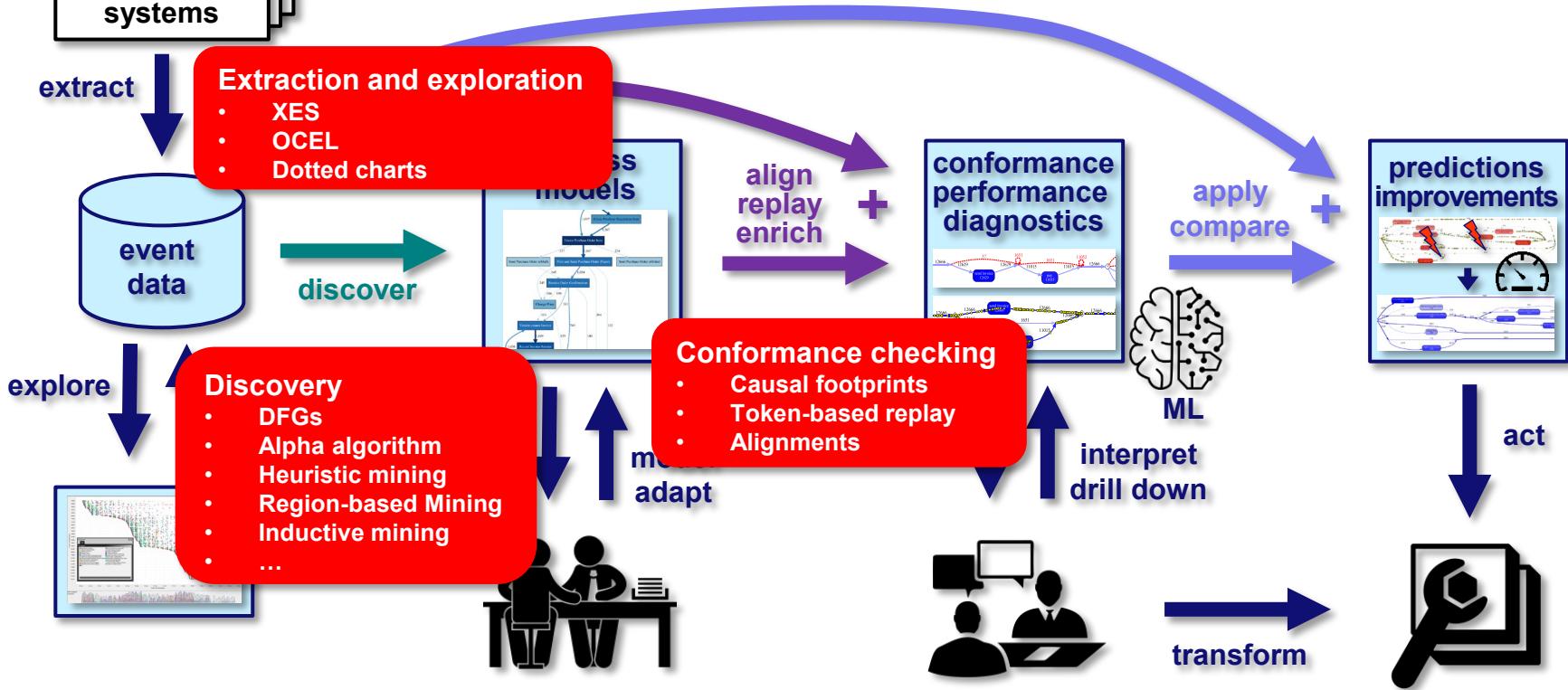
Chair of Process
and Data Science

RWTHAACHEN
UNIVERSITY

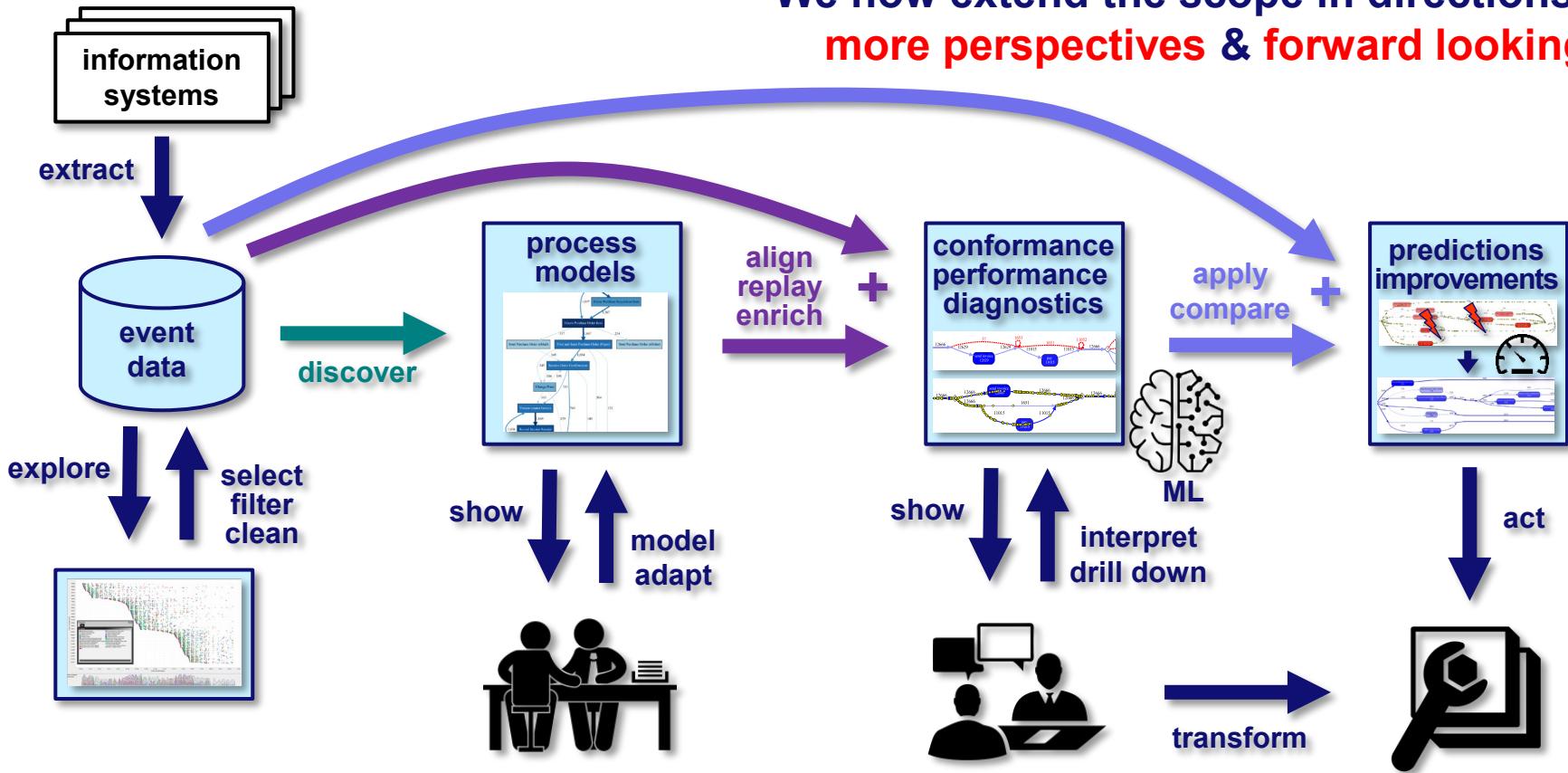
Focus thus far



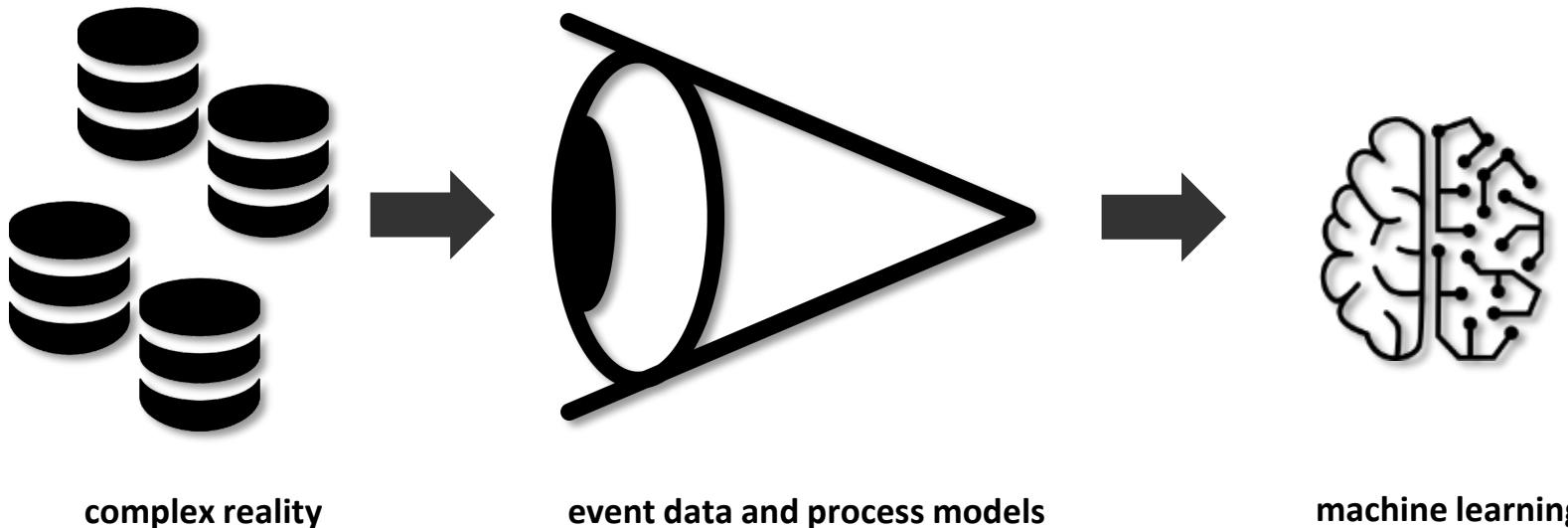
Focus thus far



We now extend the scope in directions: more perspectives & forward looking



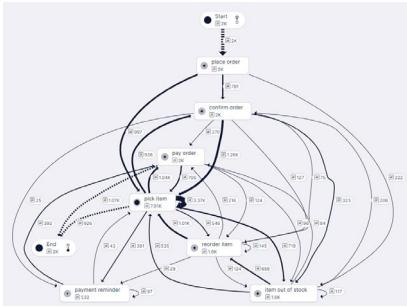
Creating Situation Tables Using PQL



Creating Situation Tables Using PQL



ORACLE
Epic
servicenow

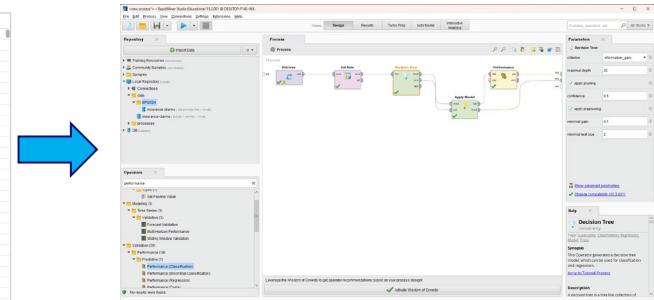


PQL

situation table

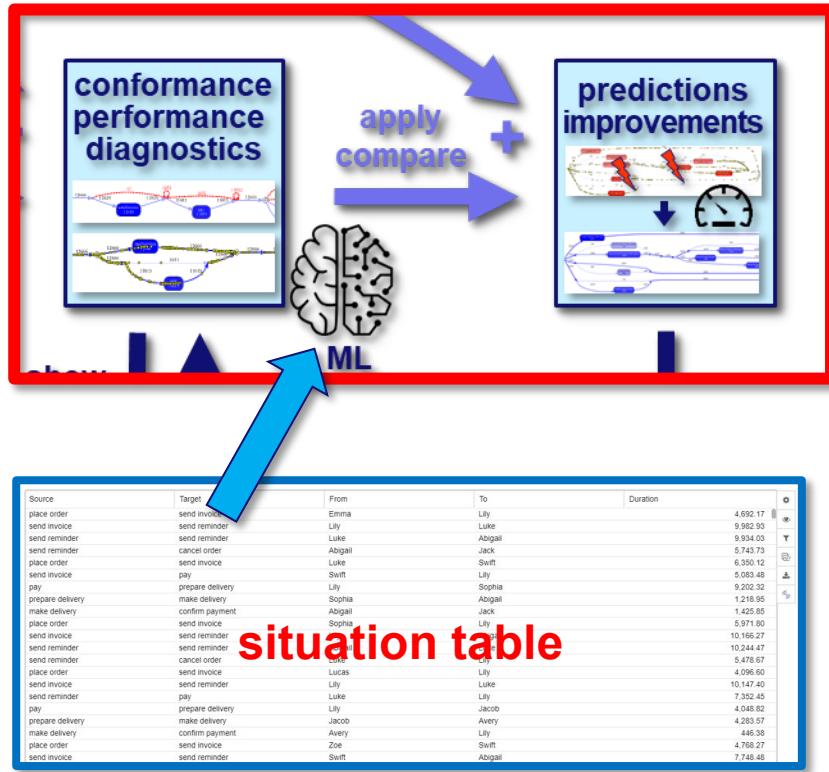
CASE	PRODUCT	ADDRESS	first resource	Total throughput time in h.	decision
1	SAMSUNG Galaxy J5	Munich	Carlo	239	pay
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	201	pay
3	APPLE iPhone 6s 16 GB	New York	Sophia	200	cancel order
4	MOTOROLA Moto E 4G	New York	Sophia	498	cancel order
5	SAMSUNG Core Prime G361	Aachen	Isabella	741	pay
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay
8	APPLE iPhone 5s 16 GB	Aachen	Speedy	200	pay
9	APPLE iPhone 5s 16 GB	Amsterdam	Speedy	412	pay
10	HUAWEI P8 Lite	Munich	Emma	415	pay
11	MOTOROLA Moto G	Aachen	Jacob	508	pay
12	HUAWEI P8 Lite	Munich	Speedy	469	pay
13	HUAWEI P8 Lite	Aachen	Jacob	439	pay
14	SAMSUNG Core Prime G361	Munich	Sophia	331	pay
15	SAMSUNG Galaxy S4	Aachen	Sophia	200	pay
16	SAMSUNG Galaxy S4	Aachen	Olivia	200	pay
17	SAMSUNG Galaxy S4	Munich	Luke	200	pay
18	SAMSUNG Galaxy S4	Munich	Lucas	200	pay
19	SAMSUNG Galaxy S4	New York	Luke	200	pay
20	APPLE iPhone 6s 64 GB	Munich	Sophia	200	cancel order
21	APPLE iPhone 6s 64 GB	Aachen	Jacob	200	cancel order
22	SAMSUNG Galaxy S4	Aachen	Speedy	200	cancel order
23	MOTOROLA Moto G	New York	Emma	448	cancel order
24	MOTOROLA Moto G	Aachen	Speedy	1363	pay
25	APPLE iPhone 5s 16 GB	Munich	Sophia	177	pay
26	APPLE iPhone 5s 16 GB	Aachen	Jacob	200	cancel order
27	MOTOROLA Moto G	Amsterdam	Aiden	987	pay
28	SAMSUNG Galaxy S4	New York	Aiden	500	cancel order

Celonis



RapidMiner

Creating Situation Tables Using PQL



- A **situation table** is a two-dimensional table.
 - Each row is an instance.
 - Each column is a variable.
 - There may be a split into a response variable and predictor variables (for supervised learning).
- Five types of situation tables:
 - **Case-based situation table:** Each row (instance) corresponds to a case with variables.
 - **Event-based situation table:** Each row (instance) corresponds to an event.
 - **Resource-based situation table:** Each row (instance) corresponds to a resource.
 - **Event-pair-based situation table:** Each row (instance) corresponds to a pair of events.
 - **Aggregate situation tables:** Each row (instance) corresponds to a combination of cases and/or events.



Creating Situation Tables Using PQL



Creating Situation Tables with PQL

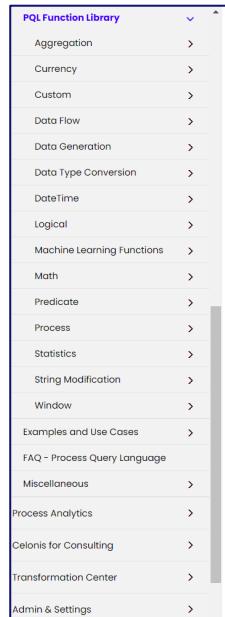
The starting point for every data mining and machine learning technique are **data**. The input data used for many data mining and machine learning techniques typically have a tabular form. In the data table, each row is an instance and each column is a variable. Instances correspond to individuals, entities, cases, objects, or records. Variables are often referred to as attributes, features, or data elements. In this course, we briefly cover data science topics related to supervised learning (e.g., Decision Trees) and unsupervised learning (e.g., Clustering). RapidMiner is a tool which provides a rich body of data science methods and algorithms to select and apply to your data. We use RapidMiner to demonstrate how algorithms such as Decision tree mining and K-means clustering can be applied to real data.

The data describing a process come in the form of an event log. The event log is the starting point for process mining techniques and it contains information regarding activities, cases, resources, and all other entities or objects that are involved in a process. Using data mining and machine learning one can gain process insights that go beyond process discovery and conformance checking. For instance, we may want to analyze which case attributes influence the path cases take in a process, how resources hand over work to each other, whether busy resources cause delays in particular process parts, and so on. These questions can be translated into concrete data science problems and thus, the existing techniques can be applied to help answering them. That translation is, however, not trivial. Given a particular process related question, its transformation into a data science problem requires defining the instances and the variables for the problem at hand. This is not necessarily the event log, where typically instances refer to events and variables refer to event attributes. It is only with the proper input data that the output of any data mining and machine learning technique yields correct and valid insights into the data. This is where the Celonis Process Query Language (PQL) comes to help. Using PQL, we can use the provided event data to generate any kind of data table we need depending on the question at hand (see Figure 1).

In this course, we learn how for process mining tasks related to decision mining, performance analysis and organizational mining, we can generate proper input data on which existing data mining and machine learning techniques can be applied. This input data tables are called **situation tables**. More specifically, in the course we learn how to generate five types of situation tables:

- **Case-based situation table:** Each row (instance) corresponds to a case.
- **Event-based situation table:** Each row (instance) corresponds to an event.
- **Resource-based situation table:** Each row (instance) corresponds to a resource.
- **Event-pair-based situation table:** Each row (instance) corresponds to a pair of events.
- **Aggregate situation table:** Each row (instance) corresponds to a combination of cases and/or events.

See document “Creating Situation Tables with PQL”



Celonis Product Documentation / PQL - Process Query Language / PQL Function Library

PQL Function Library

Description

PQL provides a wide variety of functions and operators that can be used within a query.

This sections contains all available functions and operators. In contrast to operators, functions obey a strict syntax of listing function parameters - especially the more complex ones like **CASE WHEN**.

Operator Precedence

If an expression contains more than one operator, the operators are evaluated in order of operator precedence. To influence the evaluation order you can use parentheses which have the highest operator precedence and are evaluated first. Operators with the same operator precedence are evaluated from left to right.

Expressions

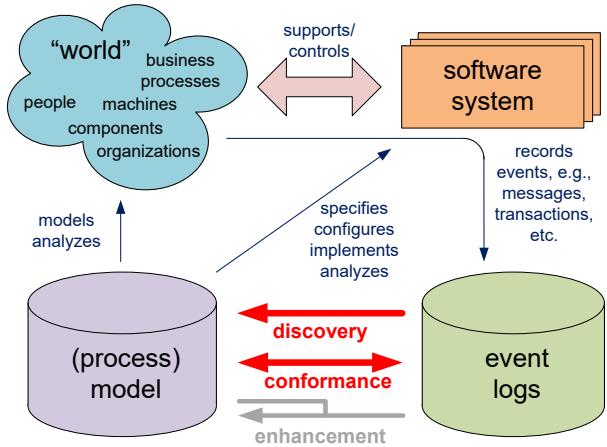
Precedence	Operator	Operation
Highest	()	parentheses
	+ , -	unary positive and negative operator
	* , / , %	multiplication, division, modulo
	, , -	addition, subtraction
		concatenation
Lowest	=, !=, <, >, <=, >=	equal, not equal, less than, less than or equal, greater than, greater than or equal

Reference to look up specific operators
<https://docs.celonis.com/en/pql-function-library.html>

Beyond control-flow



Focus thus far

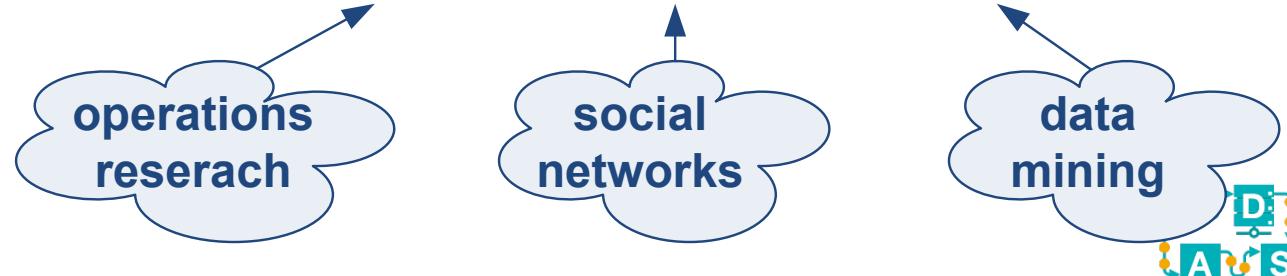


control-flow only	
discovery L → M	✓
conformance L+M → D	✓



Bigger picture

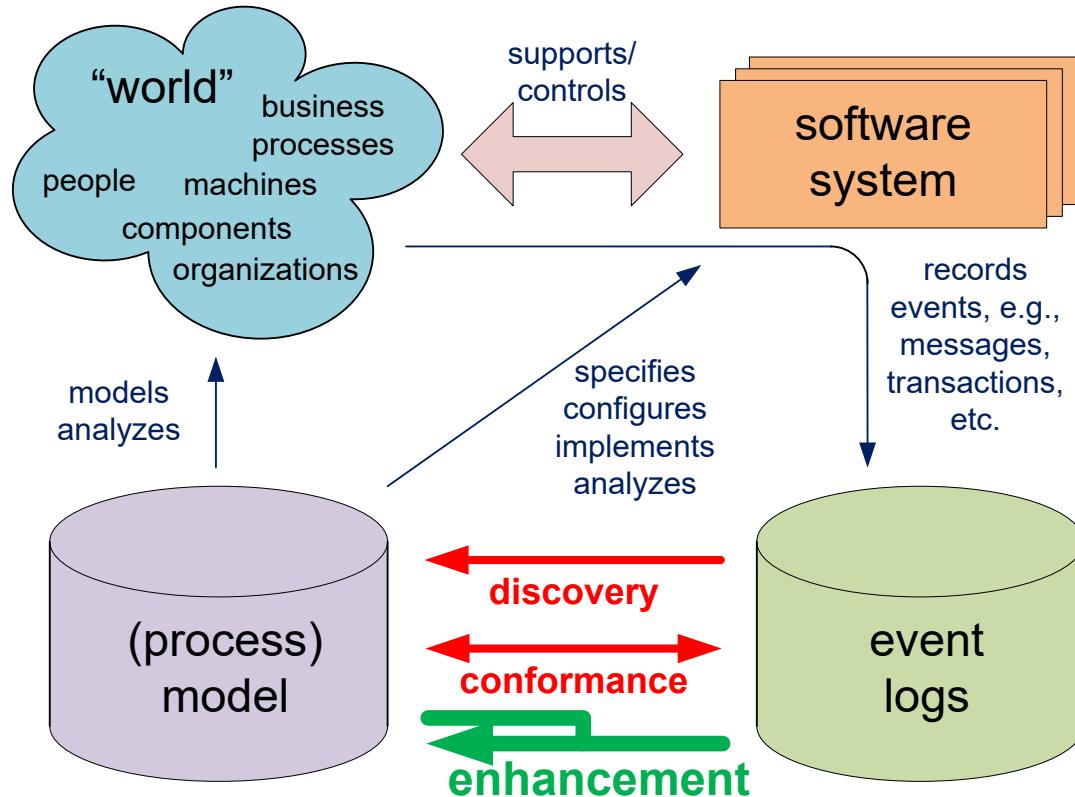
control-flow only	control-flow and ...				
	time	resources	data	
discovery 	✓	✗	✗	✗	✗
conformance 	✓	✗	✗	✗	✗
enhancement 	✗	✗	✗	✗	✗



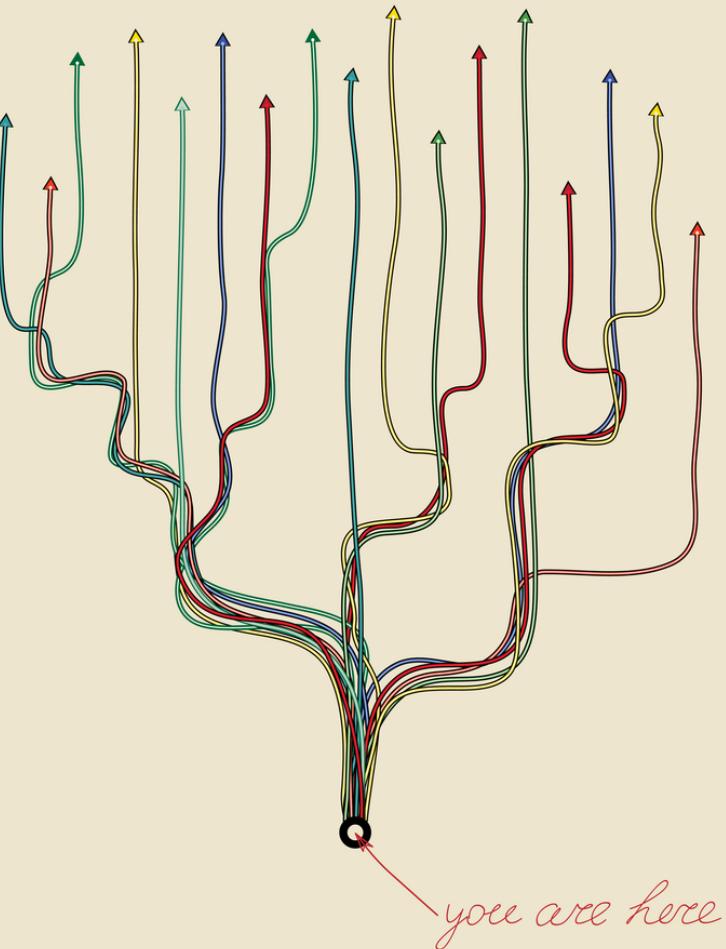
Decision mining



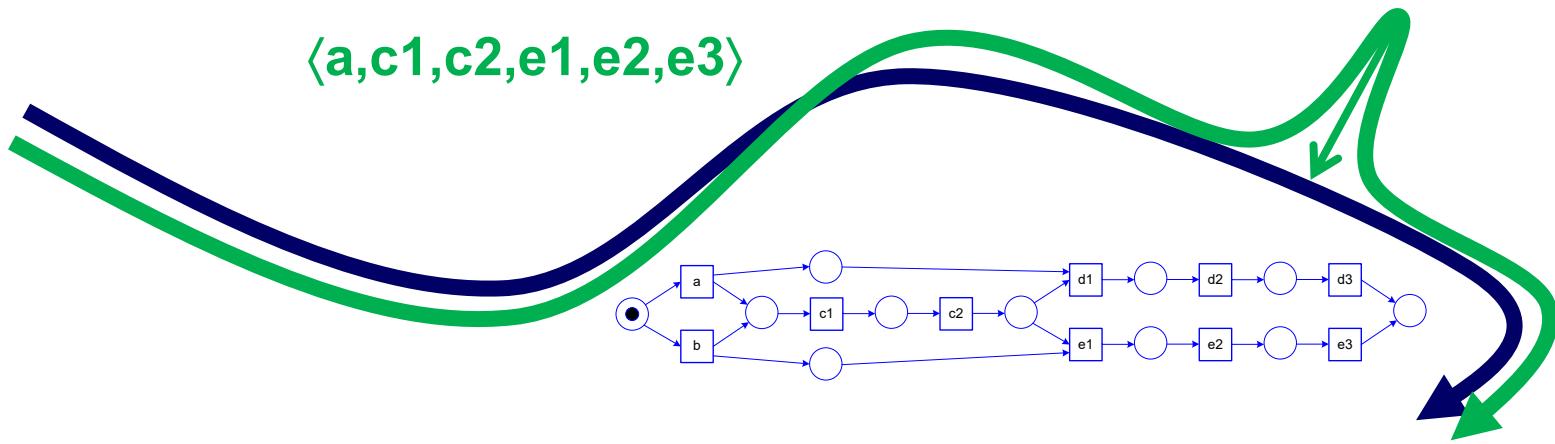
Enhancement: Extension and Repair



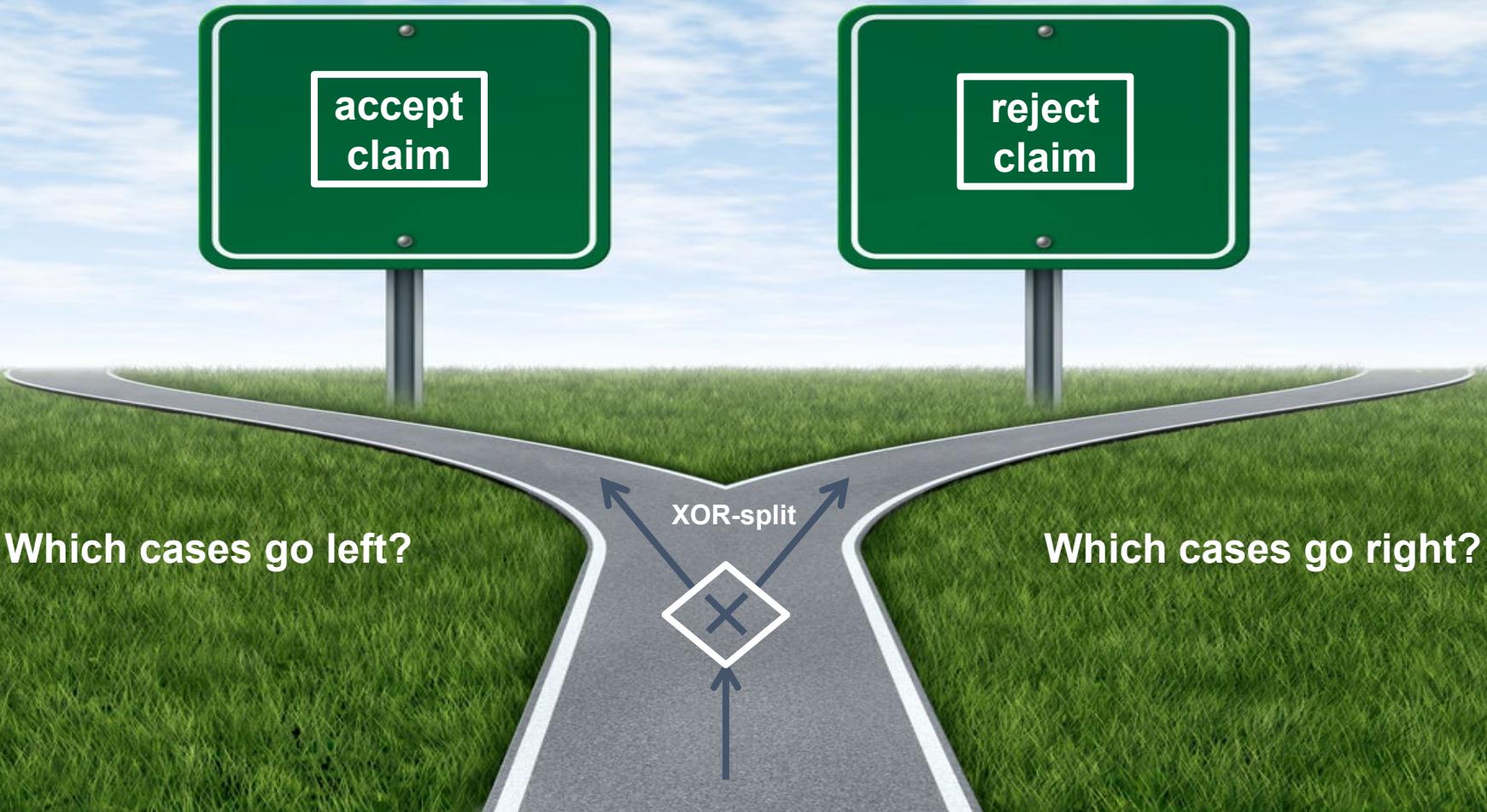
Mining Decision Points



- **Input:**
 - event log
 - process model
- **Assumption: Log and model have been aligned.**
 - Mapping of activity names in log and model.
 - Every trace can be related to a path through the model.

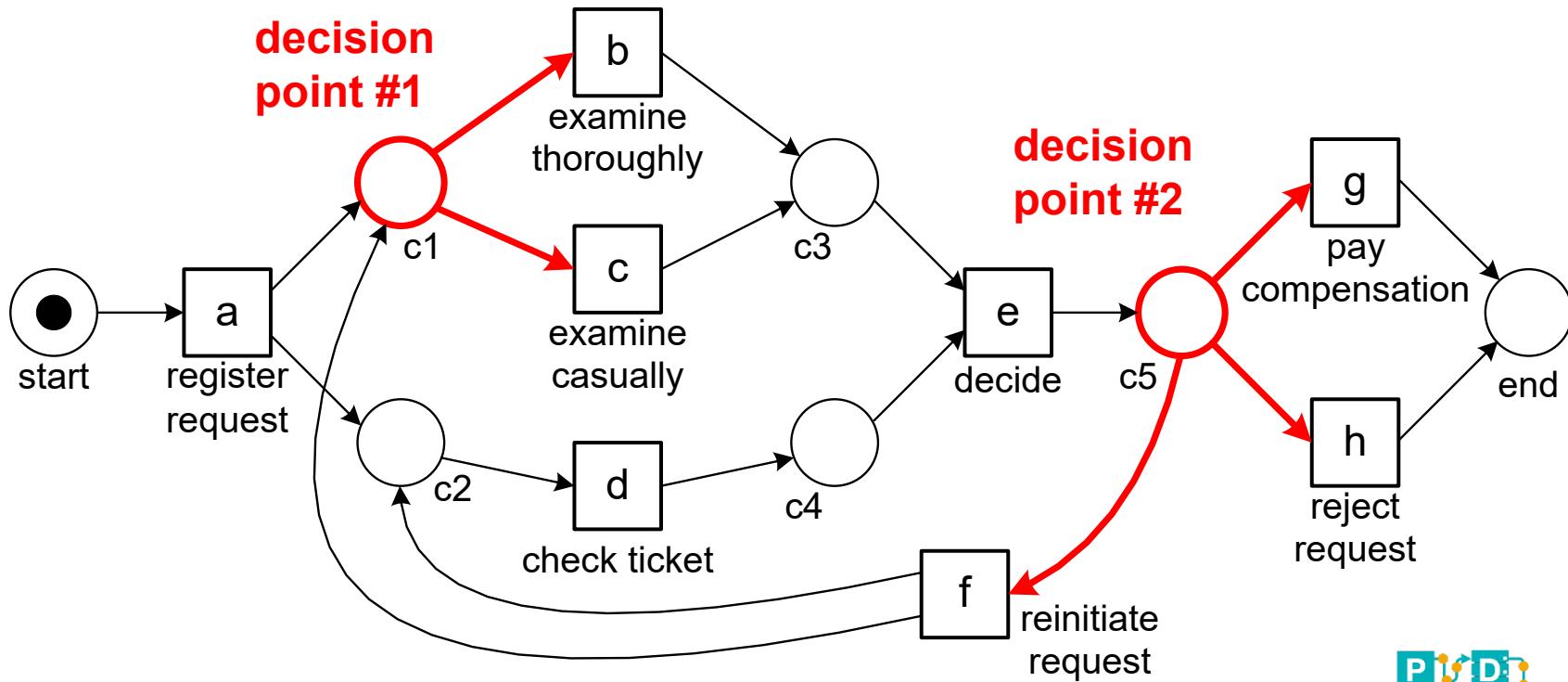


a	»	c1	c2	e1	e2	e3
»	b	c1	c2	e1	e2	e3

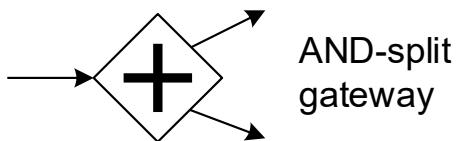




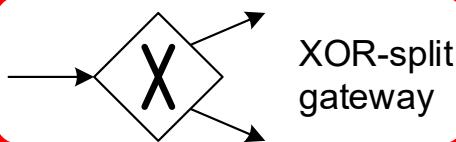
Places with multiple output arcs form decision points



Decision points in BPMN



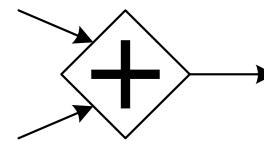
AND-split gateway



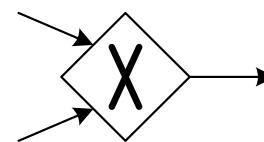
XOR-split gateway



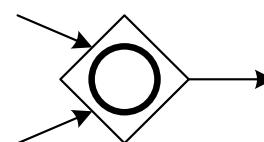
OR-split gateway



AND-join gateway



XOR-join gateway



OR-join gateway



**process
mining**

**machine
learning**

Remember: Classification using decision trees

gender	age	smoker	car brand	claim
female	47	yes	Volvo	no
male	31	no	Alfa Romeo	yes
male	59	no	Alfa Romeo	yes
male	28	no	Fiat	no
male	44	no	BMW	no
female	27	no	Fiat	no
male	29	no	Subaru	no
male	44	yes	Subaru	yes
male	39	no	BMW	no
male	35	?	Subaru	yes

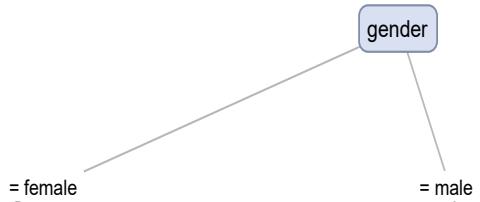
- **Response variable (dependent variable): claim (yes/no).**
- **Predictor variables (independent variables): gender, age, smoker, car brand.**

Goal: explain response variable in terms of relevant predictor variables.

Resulting decision tree

female drivers
don't claim
insurance

no



male Alfa Romeo
drivers claim
insurance

= Alfa Romeo = BMW = Fiat = Subaru = Volkswagen = Volvo

yes

no

no

yes

no

male Volvo drivers
younger than 25
claim insurance

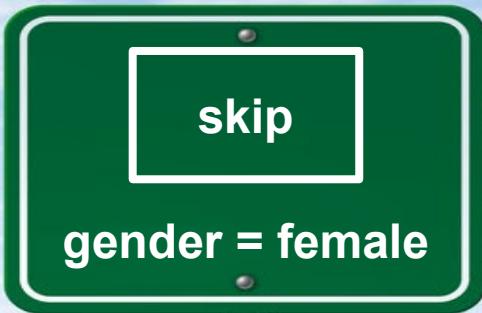
no

yes

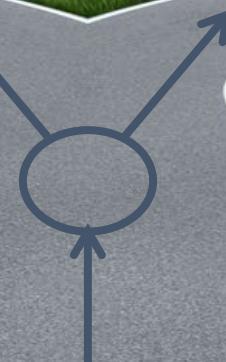
> 25.500

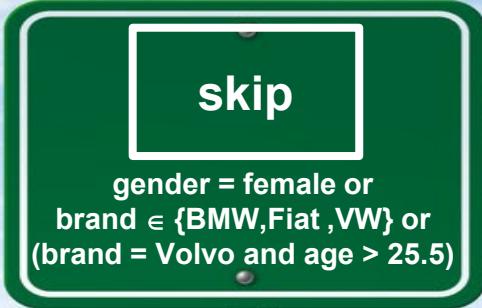
≤ 25.500

age

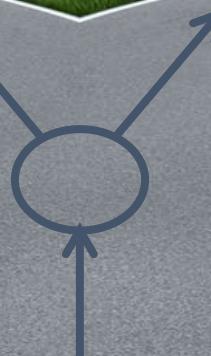


transition guard
derived from
decision tree

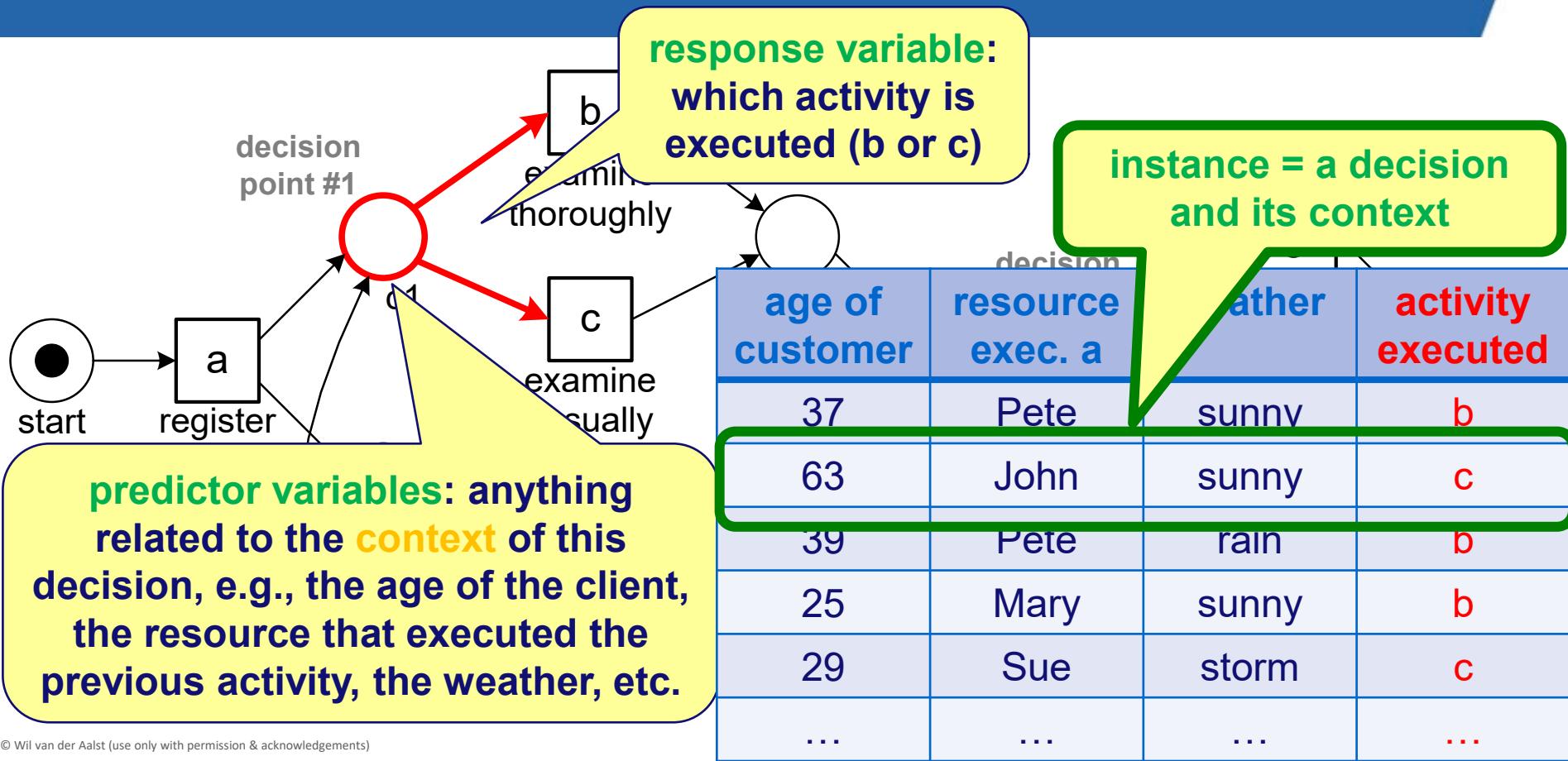




**transition guard
derived from
decision tree**



Creating a classification problem



Learning an XOR-split

type

gold

silver

gold

silver

silver

silver

gold

silver

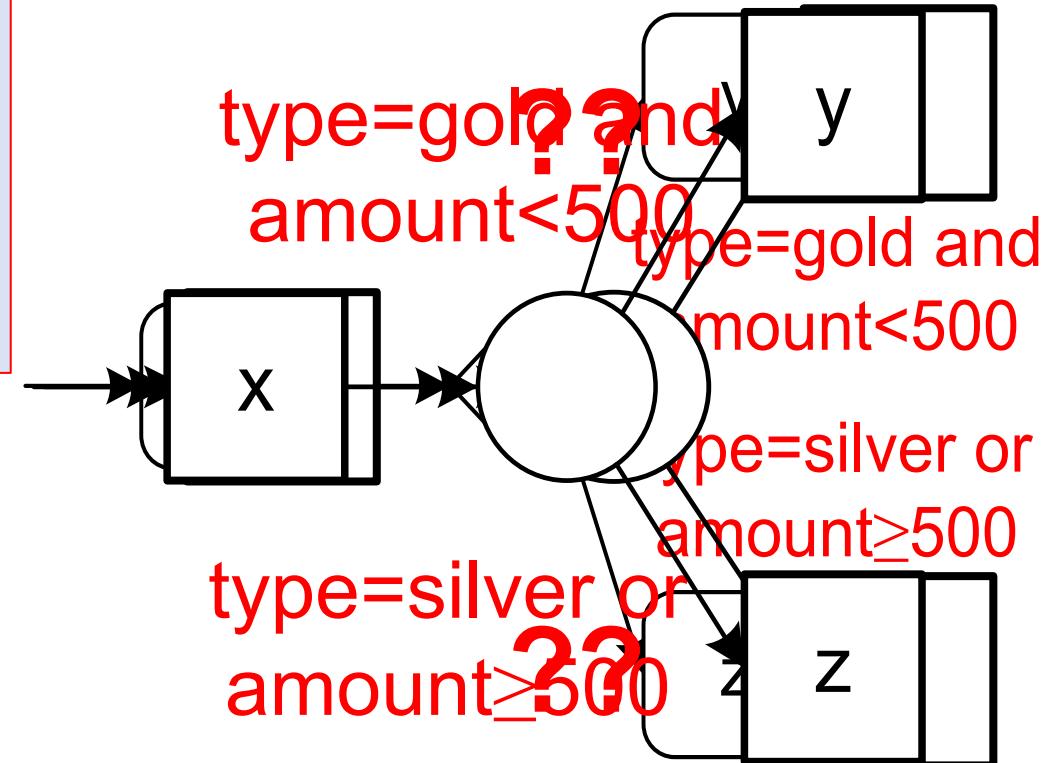
predictor variables:

- **type = silver**
- **region = south**
- **amount = 687.70**

response variable

- **activity = z**

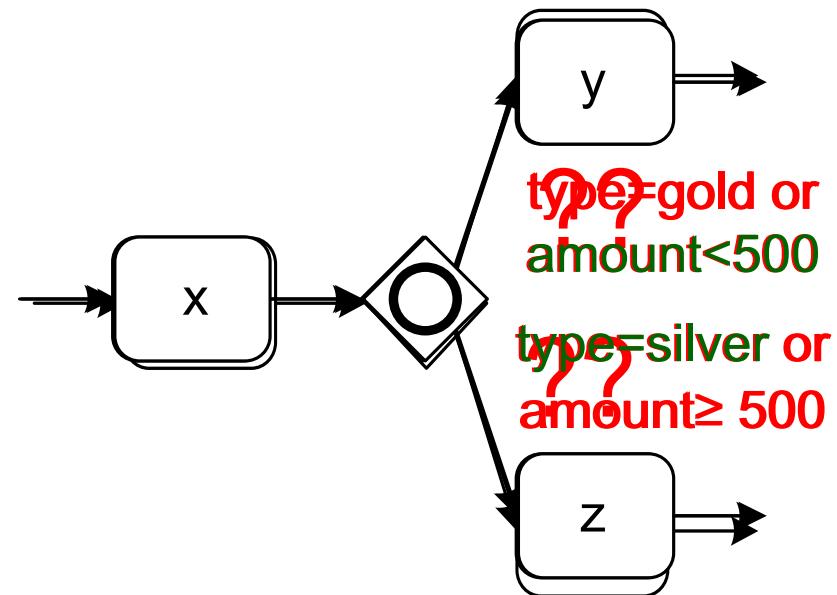
silver	south	673.70	z
gold	west	413.50	y
silver	south	687.70	z
gold	south	987.30	z
silver	north	378.80	z
gold	south	314.50	y
silver	north	537.70	z
silver	west	158.70	z
gold	east	344.50	y
...



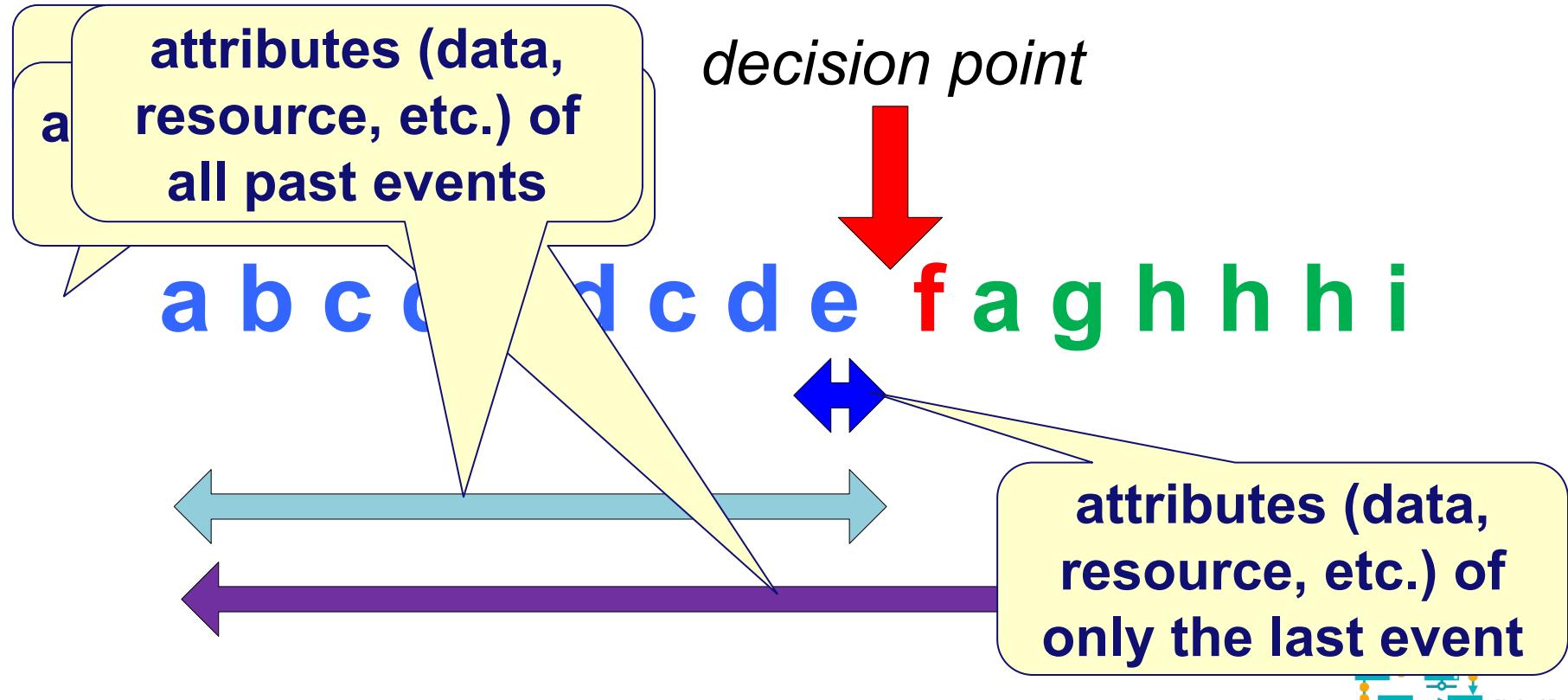
For any process notation, e.g., learning a BPMN XOR-split gateway.

Learning an OR-split

type	region	amount	activity
gold	south	987.30	y and z
silver			y and z
gold			y
silver			z
silver			y
silver			z
silver			y and z
gold	north	488.50	just y
silver	west	443.20	y and z
silver	south	673.70	just z
...

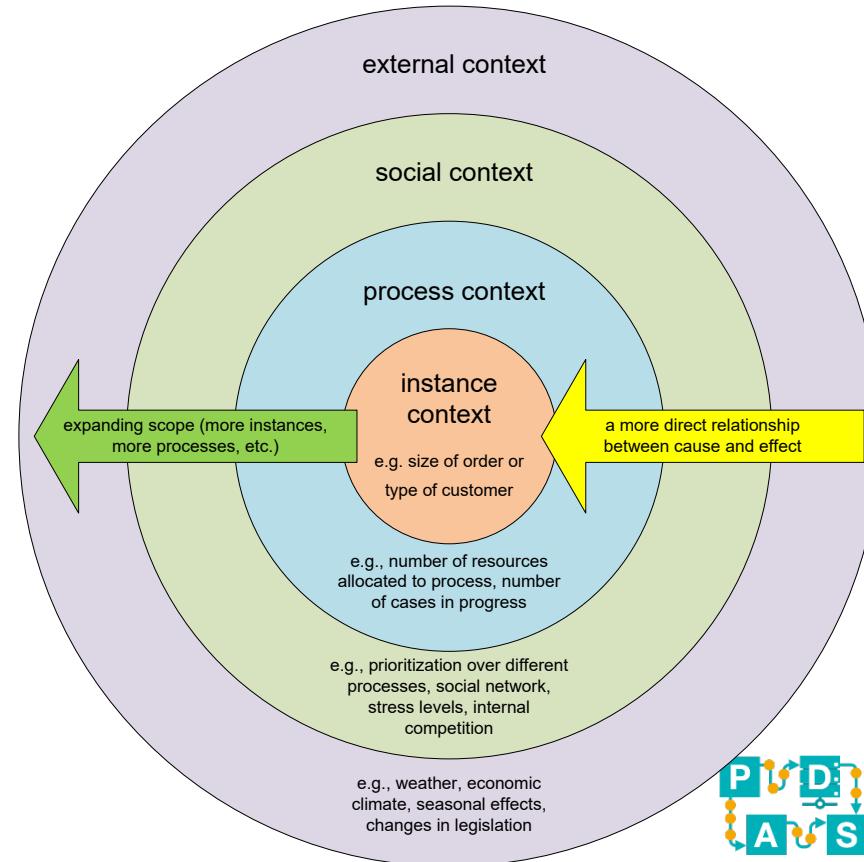


Where do the predictor variables come from?



Predictor variables may also be based on the **context** of the process instance

- Number of cases running (e.g. skip check if busy).
- Number of resources present.
- Workload of resource.
- Day of the week.
- Weather.



curse of dimensionality

more variables,
more combinations,
data gets sparser
(less instances per
combination),
danger of overfitting,

...





remaining
flow time
service levels

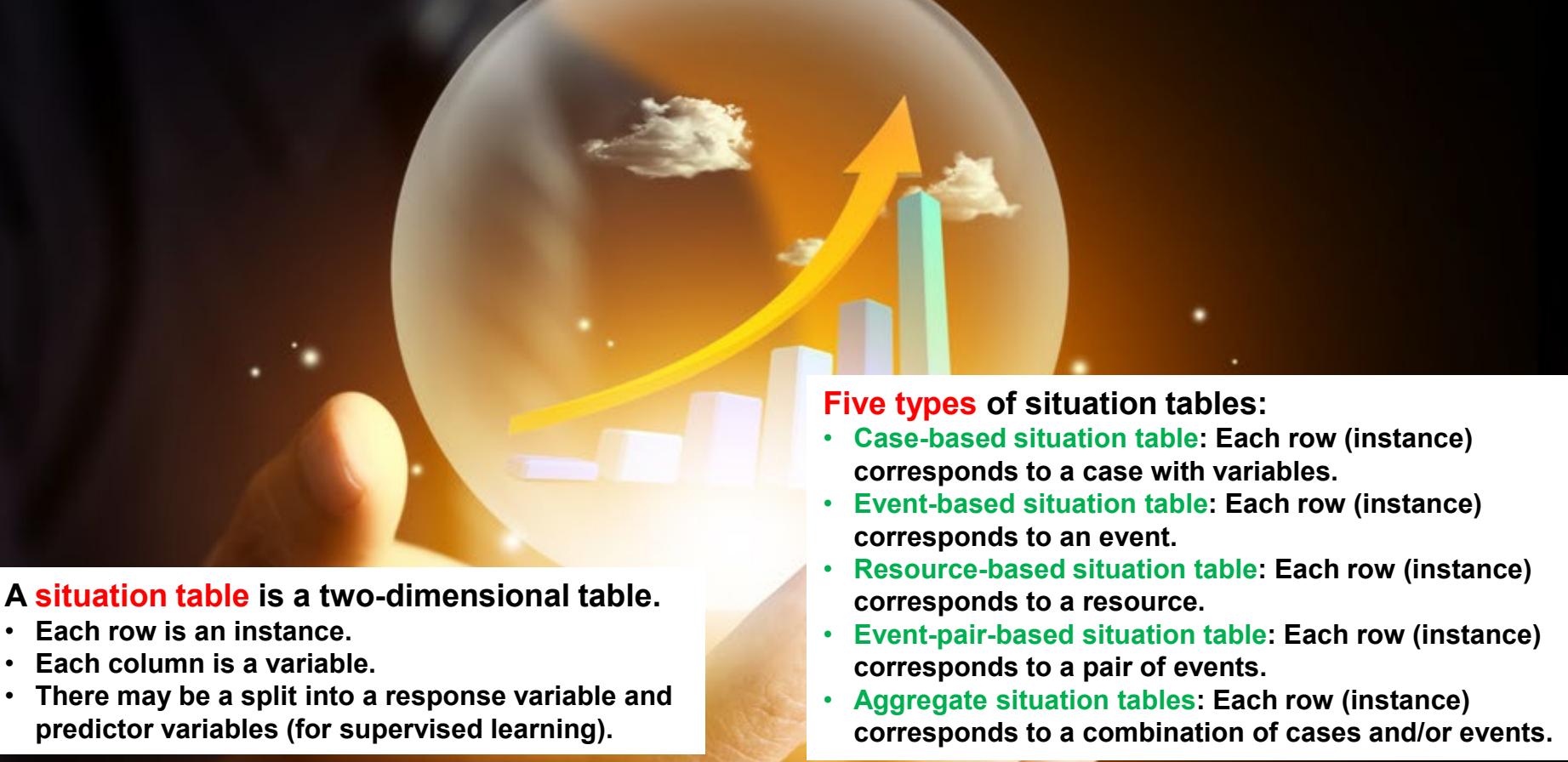
costs
incidents
risks
fraud

"next activity" is
just one of many
possible response
variables !!!

Later:

operational support

predictive analytics for processes

A hand holds a magnifying glass over a 3D bar chart. The chart consists of several bars of different heights and colors (blue, green, pink) with a yellow arrow pointing upwards from behind them. The background is a dark orange gradient with small white stars and clouds.

A **situation table** is a two-dimensional table.

- Each row is an instance.
- Each column is a variable.
- There may be a split into a response variable and predictor variables (for supervised learning).

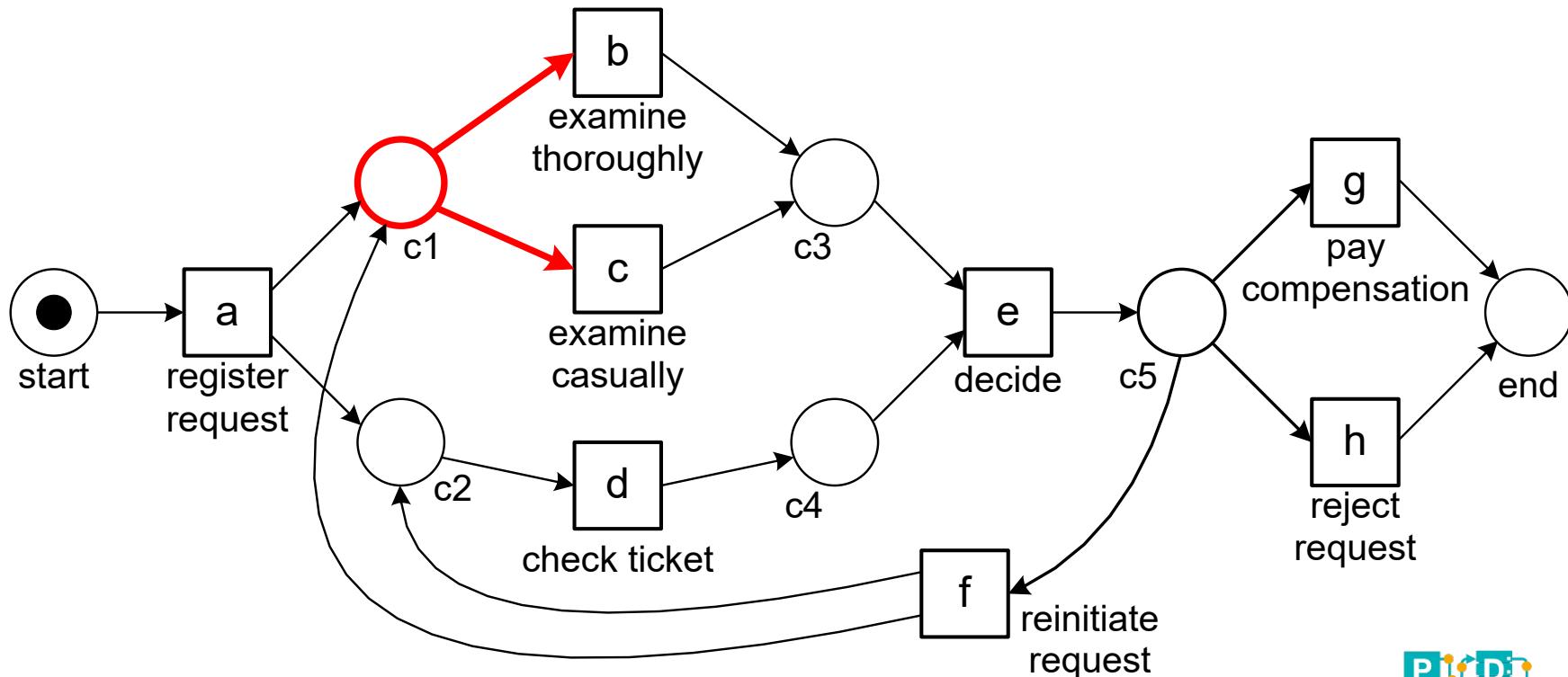
Five types of situation tables:

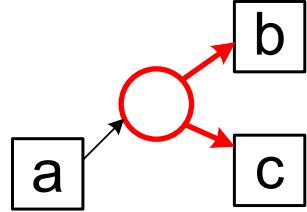
- **Case-based situation table:** Each row (instance) corresponds to a case with variables.
- **Event-based situation table:** Each row (instance) corresponds to an event.
- **Resource-based situation table:** Each row (instance) corresponds to a resource.
- **Event-pair-based situation table:** Each row (instance) corresponds to a pair of events.
- **Aggregate situation tables:** Each row (instance) corresponds to a combination of cases and/or events.

Discovering guards: Concept



Create guards for transitions b and c

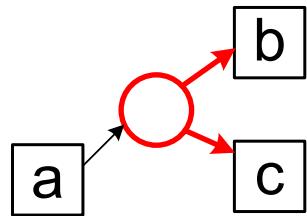




response variable:
choice between b
and c

predictor variables:
attributes
resource,
customer, and
amount of a
(assumption)

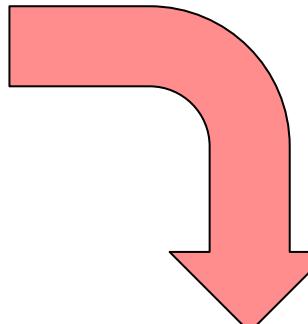
case	activity	resource	time	customer	amount
1	a	John	8.11	silver	500
2	a	Mary	8.12	gold	800
2	d	Sue	8.32	gold	800
1	b	John	9.12	silver	500
3	a	John	9.45	silver	300
3	b	Mary	9.56	silver	300
1	d	John	9.45	silver	500
2	c	Mary	9.56	gold	800
3	d	Mary	10.43	silver	300
4	a	John	11.34	gold	850
4	c	John	11.57	gold	850
...



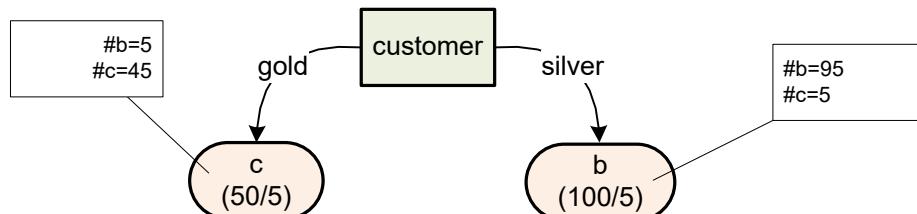
case	activity	resource	time	customer	amount
1	a	John	8.11	silver	500
2	a	Mary	8.12	gold	800
2	d	Sue	8.32	gold	800
1	b	John	9.12	silver	500
3	a	John	9.45	silver	300
2	b	Mary	9.56	silver	300

case	resource executing a	customer	amount	class
1	John	silver	500	b
2	Mary	gold	800	c
3	John	silver	300	b
4	John	gold	850	c
...

resource executing a	customer	amount	class
John	silver	500	b
Mary	gold	800	c
John	silver	300	b
John	gold	850	c
...

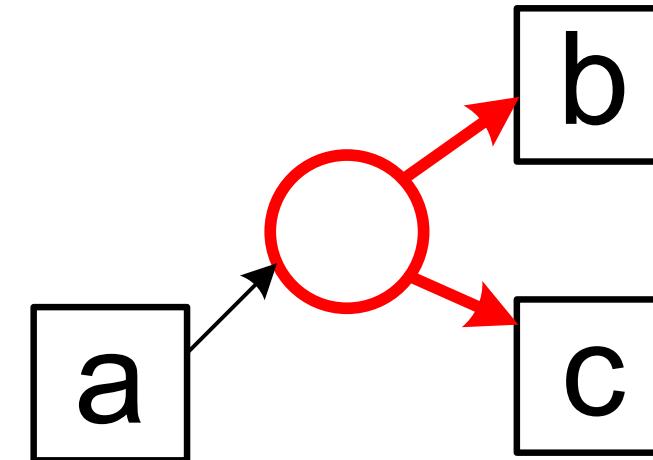
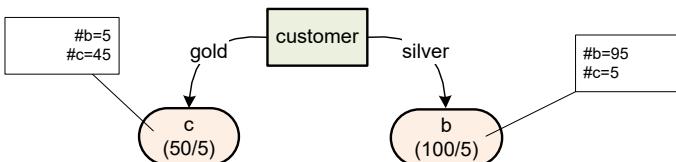


split on attribute
customer



Add guards

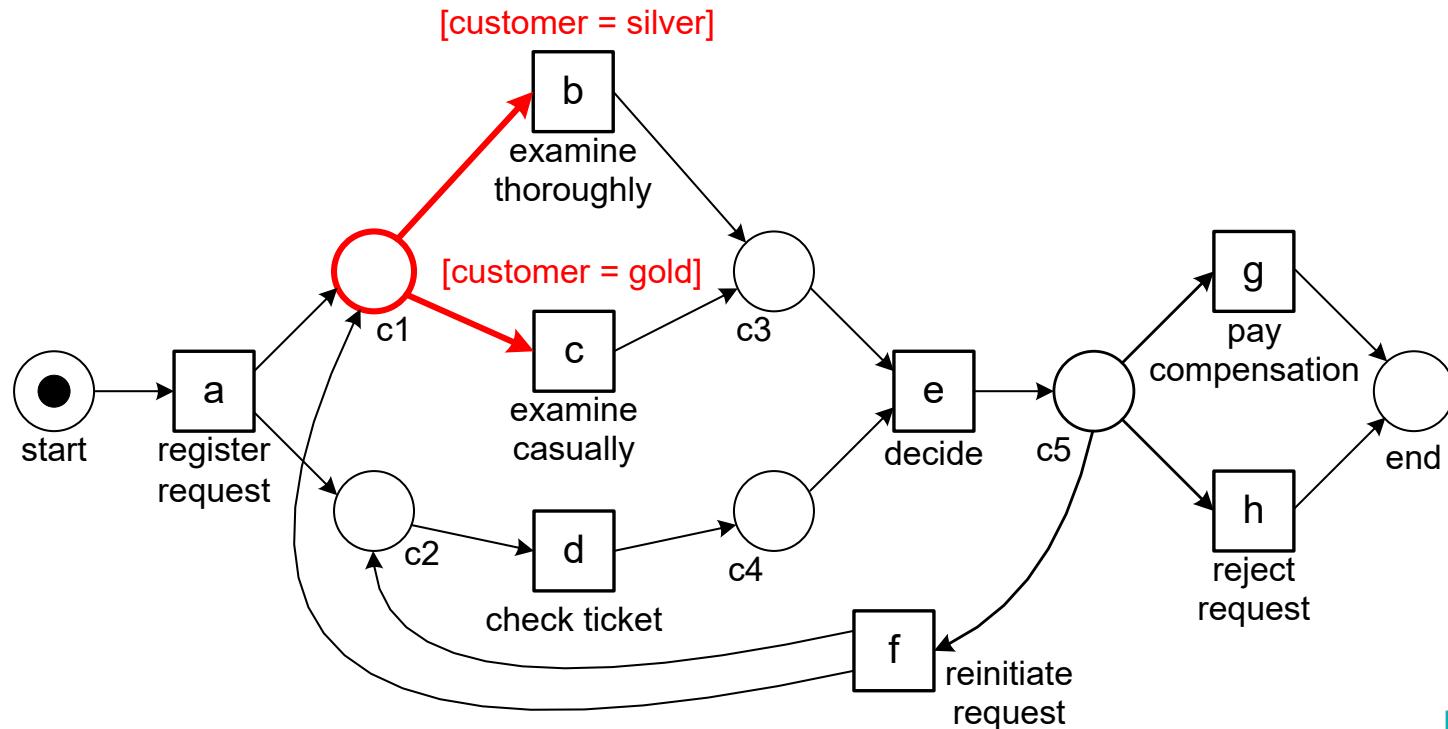
case	resource executing a	customer	amount	class
1	John	silver	500	b
2	Mary	gold	800	c
3	John	silver	300	b
4	John	gold	850	c
...



[customer = silver]

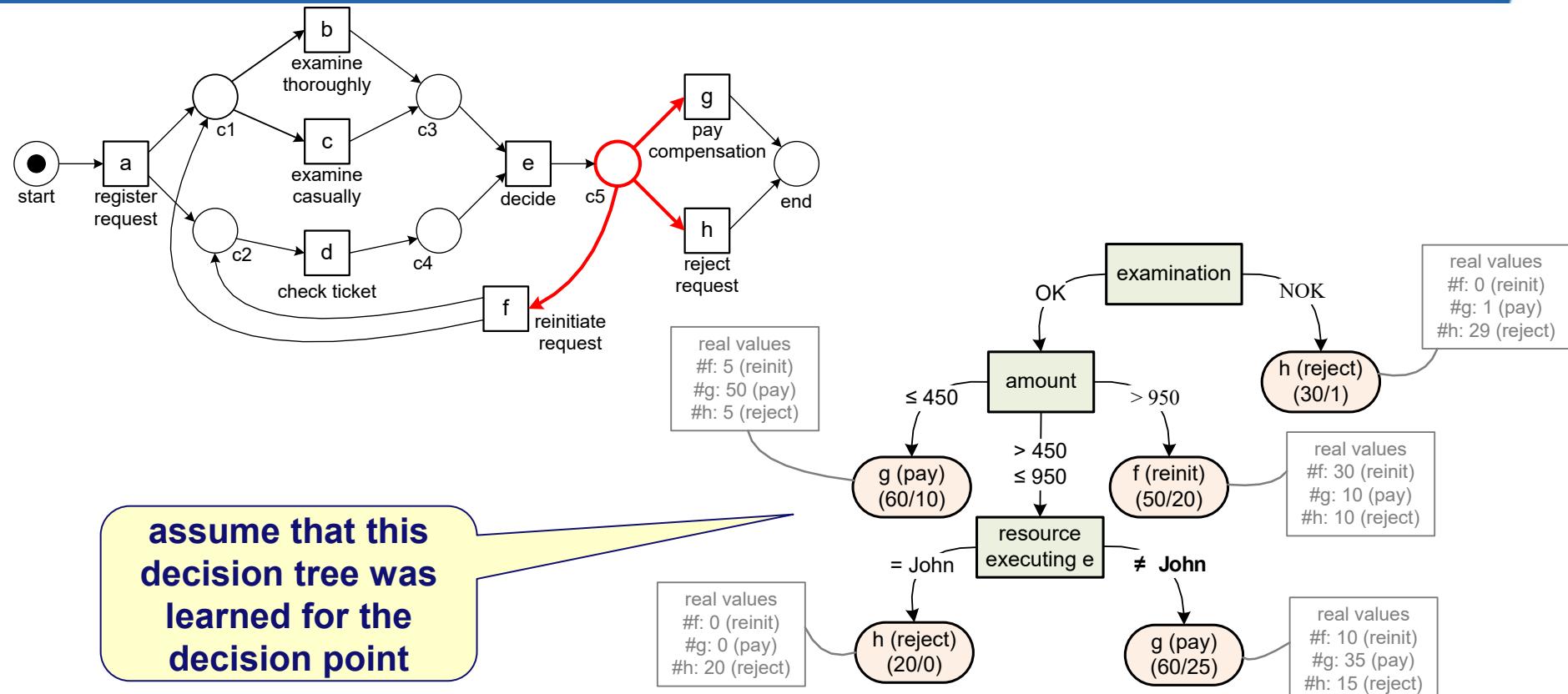
[customer = gold]

Data-aware process model



Question:

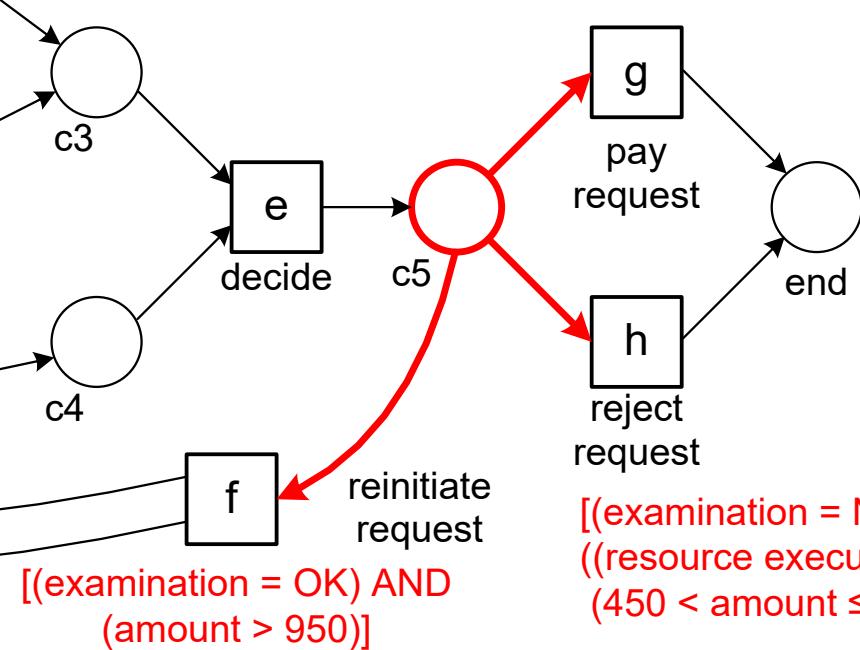
Create guards based on decision tree



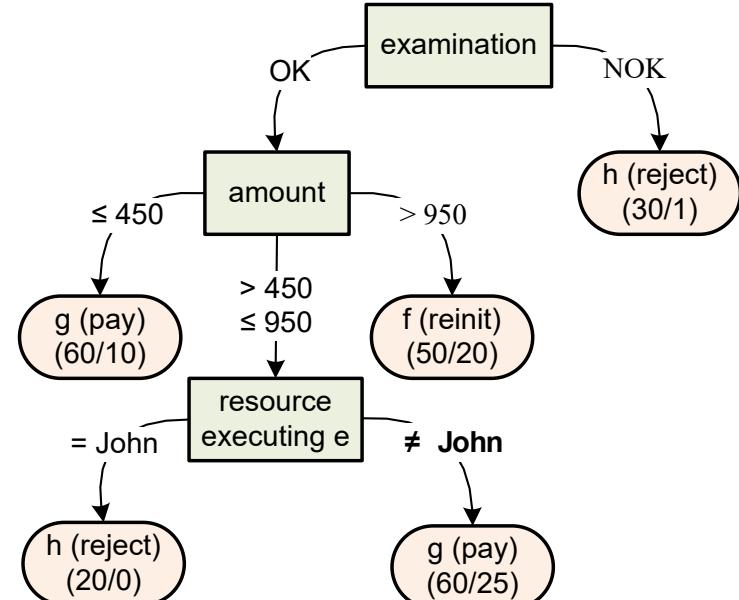
assume that this decision tree was learned for the decision point

Answer

`[(examination = OK) AND ((amount <= 450) OR
((amount ≤ 950) AND (resource executing e ≠ John)))]`

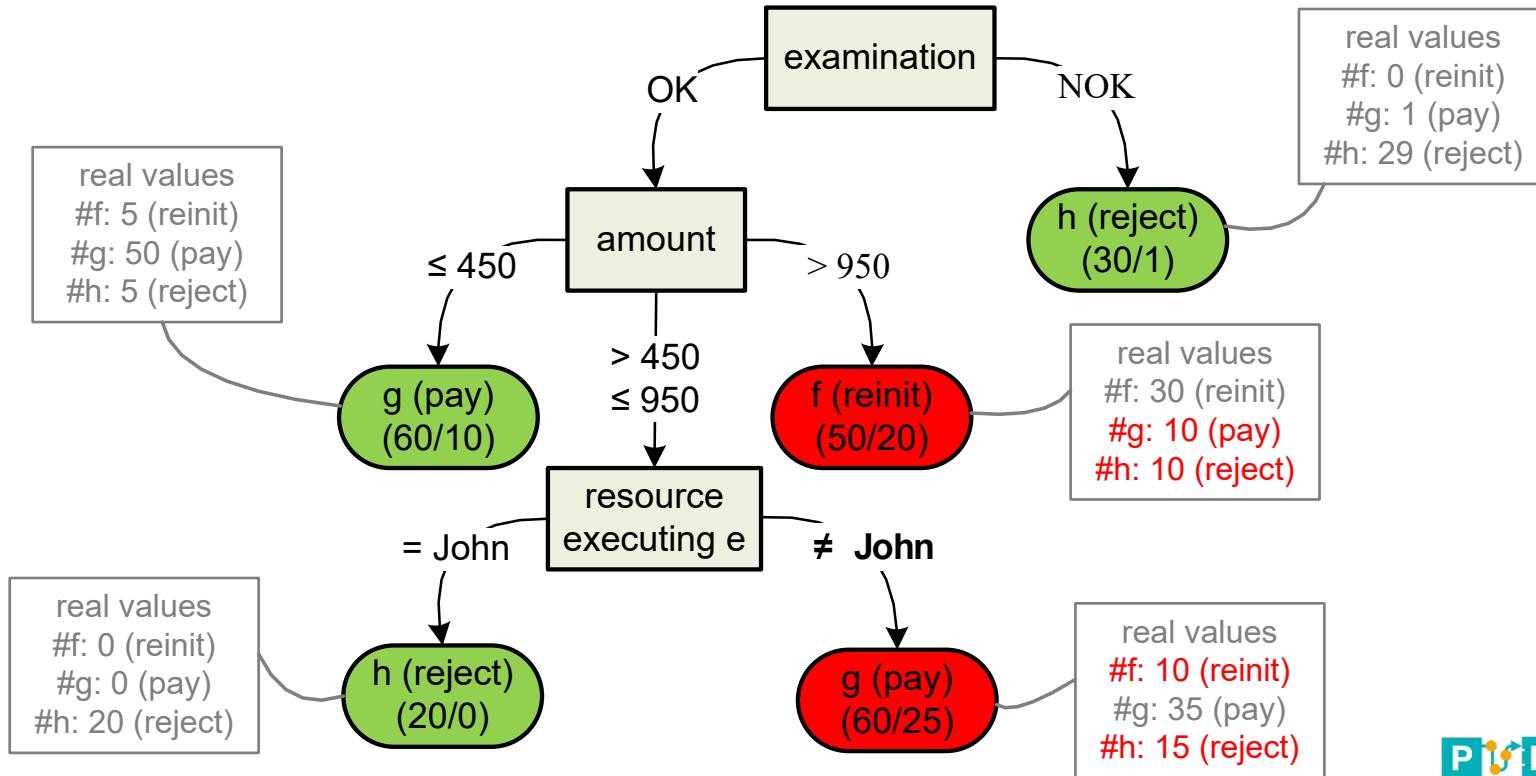


$[(\text{examination} = \text{NOK}) \text{ OR } ((\text{resource executing e} = \text{John}) \text{ AND } (450 < \text{amount} \leq 950))]$

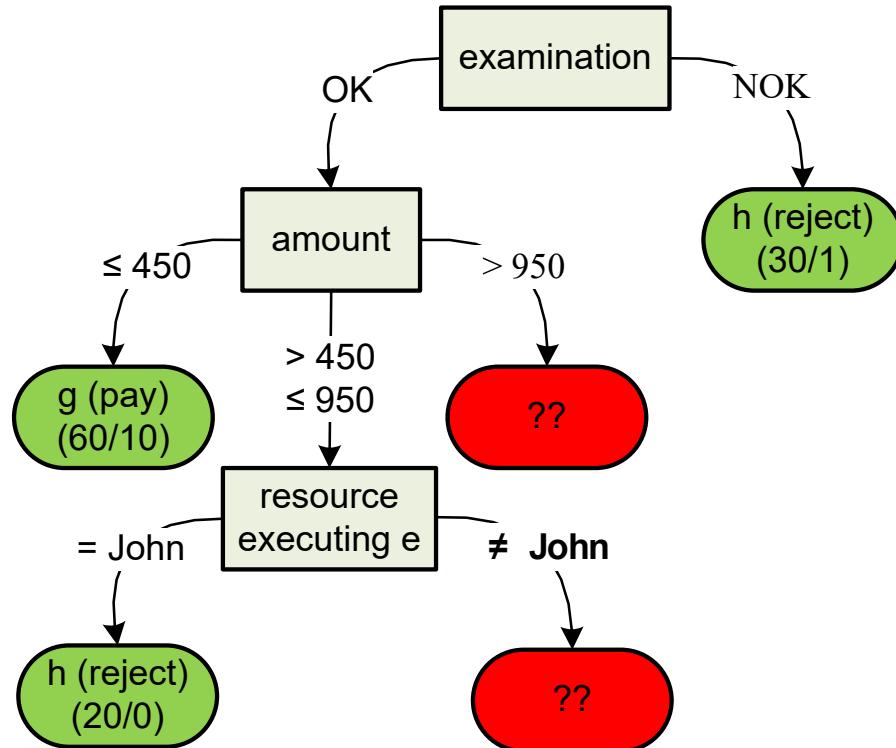


Dealing with uncertainty

(red = no consensus)



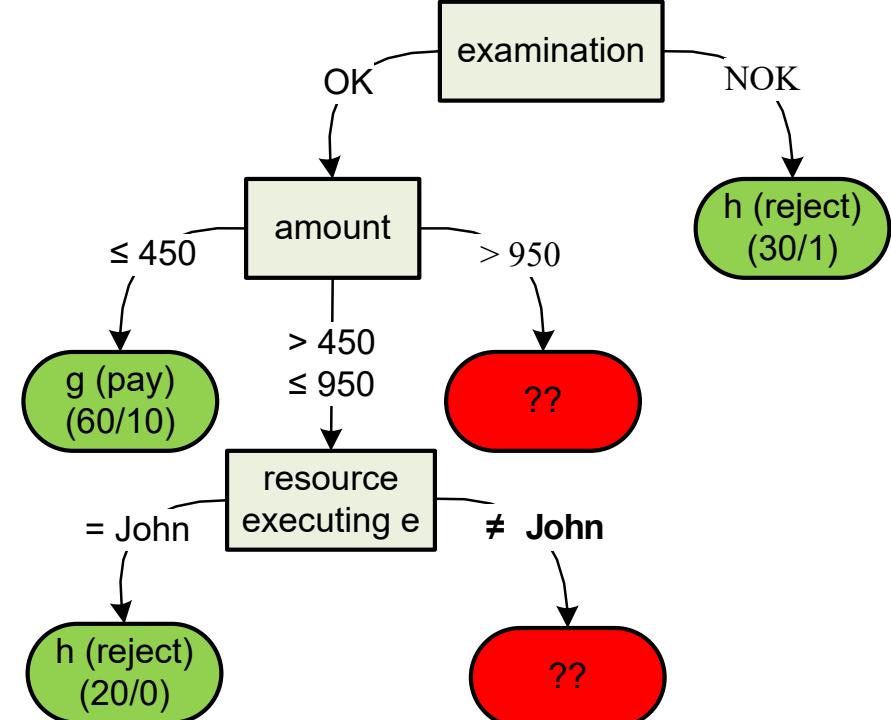
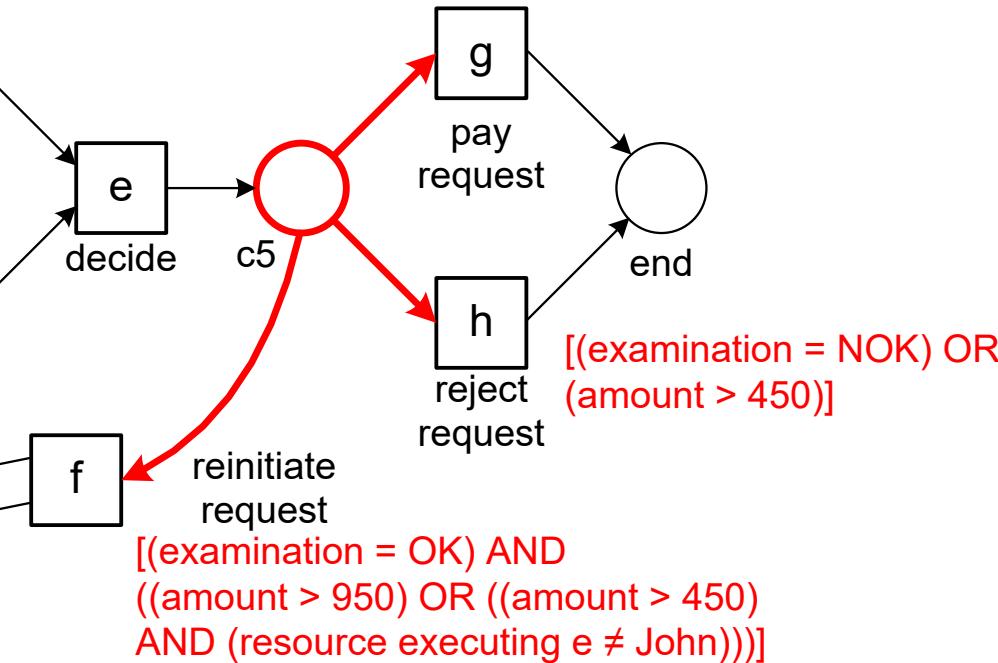
Non-deterministic guards



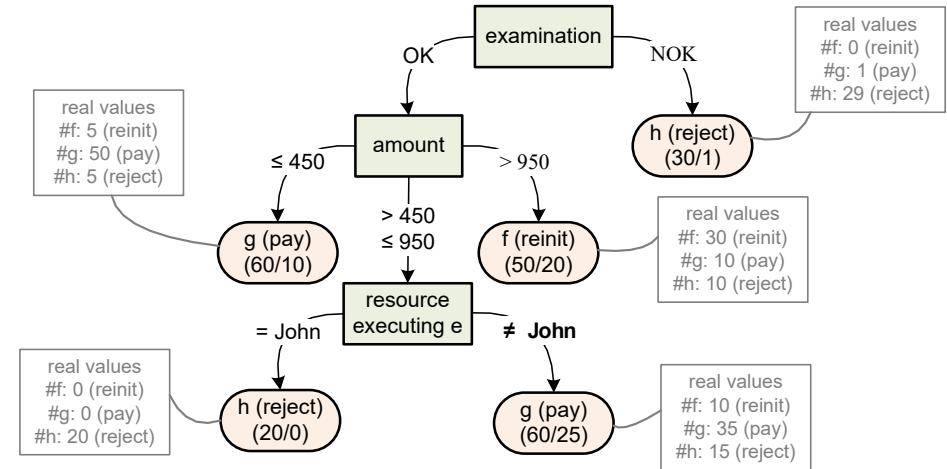
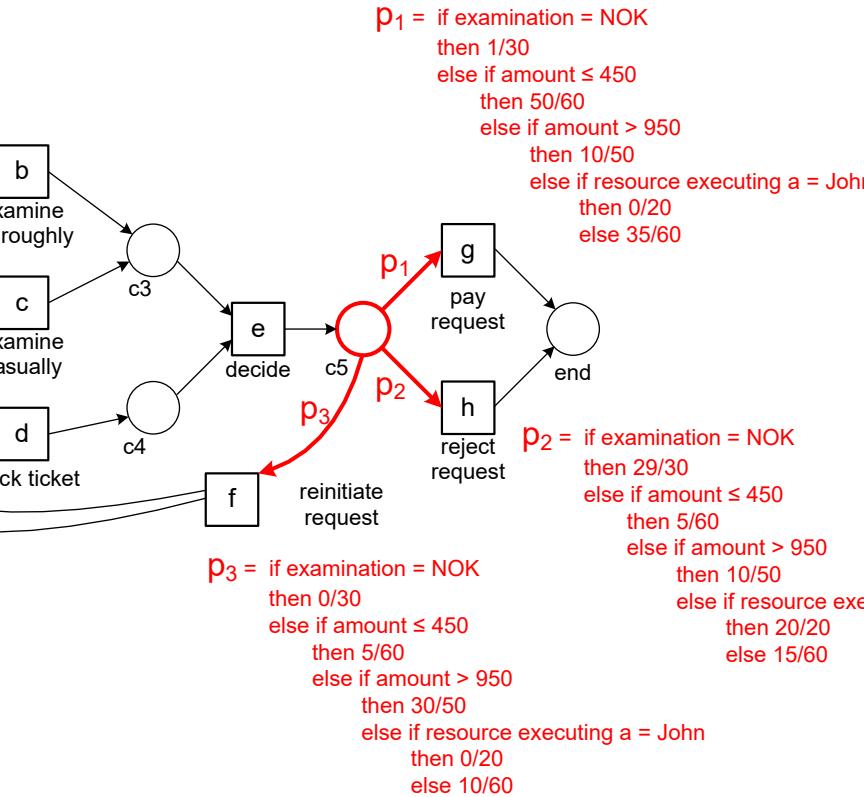
Non-deterministic guards

(conditions are weaker and overlapping)

$[(\text{examination} = \text{OK}) \text{ AND } ((\text{amount} \leq 450) \text{ OR } (\text{amount} > 950) \text{ OR } (\text{resource executing } e \neq \text{John}))]$



Data-dependent probabilities



Data-dependent probabilities rather than guards

$p_1 = \text{if examination} = \text{NOK}$

then $1/30$

else if amount ≤ 450

then $50/60$

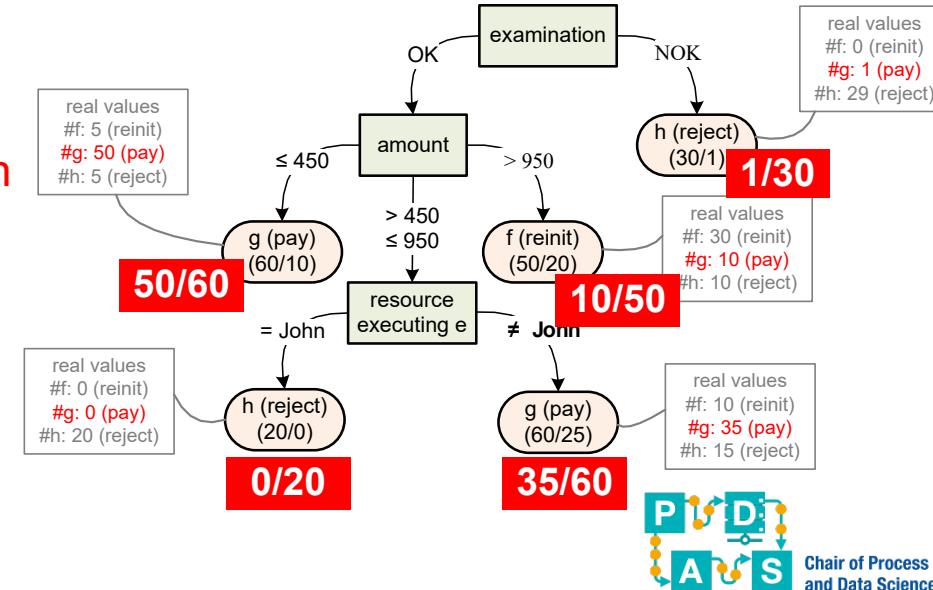
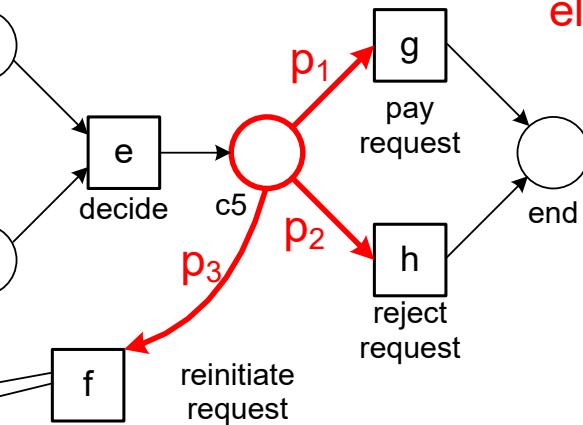
else if amount > 950

then $10/50$

else if resource executing a = John

then $0/20$

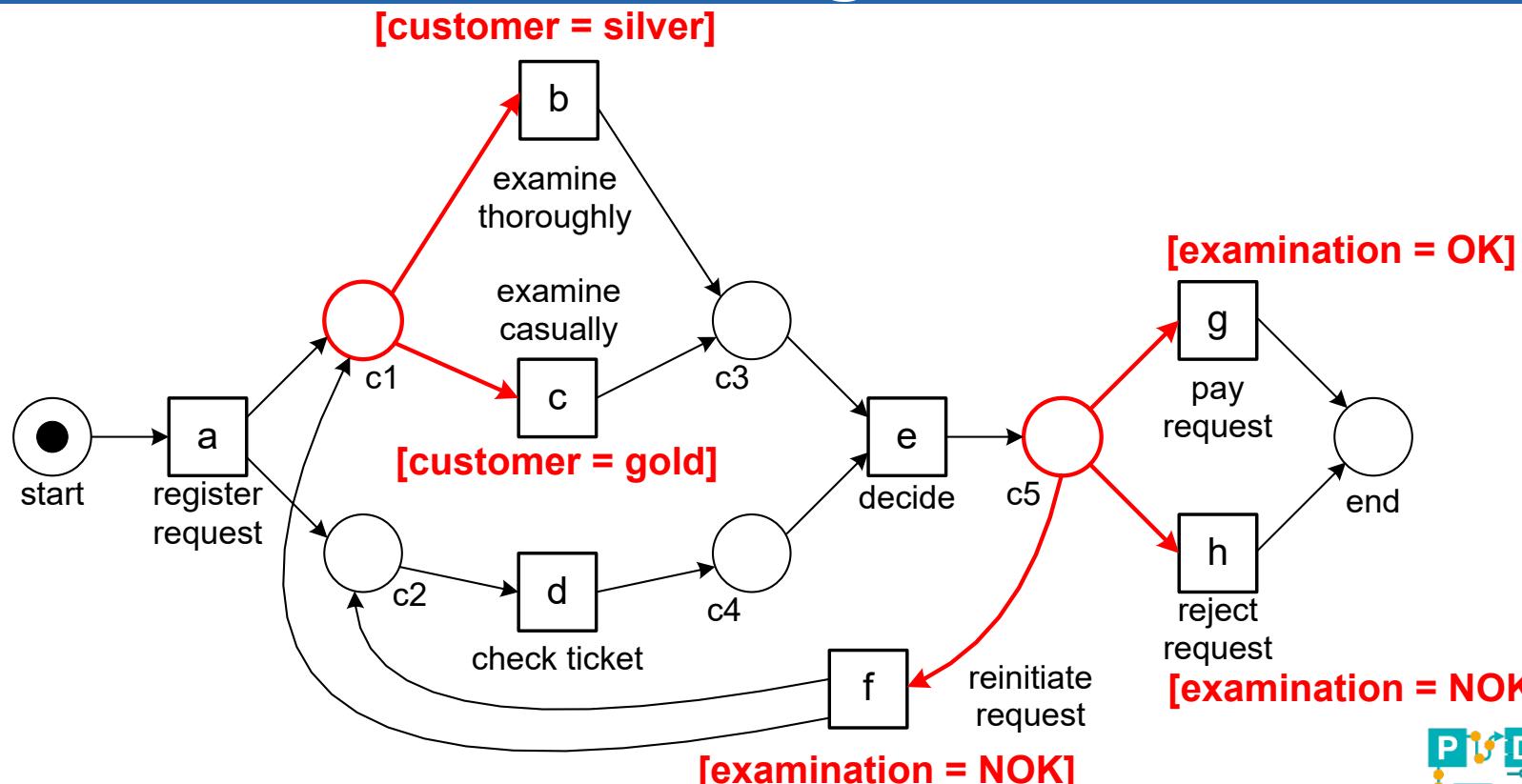
else $35/60$



descriptive ≠ prescriptive

The guards discovered are describing what has happened rather than what should have happened.

Data-aware Petri net can also be used for conformance checking!



Situation tables and PQL can be used for conformance checking

CASE	PRODUCT	ADDRESS	pay after placing order	function separation	
1	SAMSUNG Galaxy J5	Munich	ok	ok	
2	APPLE iPhone 6s 64 GB	Amsterdam	ok	ok	
3	APPLE iPhone 5s 16 GB	New York	violation	violation	
4	MOTOROLA Moto E 4G		violation	ok	
5	SAMSUNG Core Prime ...		ok	ok	
6	SAMSUNG Galaxy S4 16 GB	Amsterdam	ok	ok	
7	HUAWEI P8 Lite	Munich	ok	violation	
11	MOTOROLA Moto G	Munich	ok	ok	
12	APPLE iPhone 5s 16 GB	Aachen	ok	ok	
13	HUAWEI P8 Lite	Munich	ok	ok	
14	SAMSUNG Core Prime ...	Munich	ok	ok	
15	SAMSUNG Galaxy S4	Aachen	ok	violation	
16	SAMSUNG Galaxy S4	Aachen	ok	ok	
17	SAMSUNG Galaxy S4	Munich	ok	ok	
18	SAMSUNG Galaxy S4	Munich	ok	ok	
19	SAMSUNG Galaxy S4	New York	violation	violation	
20	APPLE iPhone 6s 64 GB	Munich	ok	ok	
21	APPLE iPhone 6s 64 GB	Aachen	ok	ok	
22	SAMSUNG Galaxy S4	Aachen	ok	ok	
23	MOTOROLA Moto G	New York	violation	violation	
24	MOTOROLA Moto G	Aachen	ok	ok	

Will be detailed later!

Syntax	Meaning	Example
..	Case contains the activity	'A'
^	Case starts with the activity	^'Scan Invoice'
\$	Case ends with the activity	'B'\$
>>	Activities directly follow	'A'>>'B'
	Logical OR	'A' 'B'
()	Group of activities	('A' 'B') >> ('C' >> 'D')
*	0 or more occurrences	('A' >> 'B')*
+	1 or more occurrences	('A' >> 'B')+
?	0 or 1 occurrences	('A' >> 'B')?
{<from>, <to>}	Between <from> and <to> occurrences	('A' >> 'B')(1, 3)
**	Any activity matches	'A' >> '**' >> 'B'
ANY	Any activity matches	'A' >> (ANY)+ >> 'B'
-	Any activity matches	'A' >> - >> 'C'
LIKE "%_%"	Activities that contain string	'A' >> LIKE "% Invoice%"
[! :]	Set of activities of which one needs to match	'A' >> [! 'B', 'C'] >> 'C'
[! :]	Set of activities of which none matches	'A' >> [! 'B'] >> 'C'
AS	Gives an alias to a regex	('A' >> (ANY)*) AS sequence, sequence >> 'B'

Any business rule can be turned into a query!

```
CASE WHEN MATCH_PROCESS_REGEX ( "events"."ACTIVITY", ('place order'>> ('*'*)* >> 'pay'))=1 THEN 'ok' ELSE 'violation' END
```

When placing an order, it needs to be paid eventually.

```
CASE WHEN PU_COUNT_DISTINCT ( "cases", "events"."RESOURCE" ) < 5 THEN 'violation' ELSE 'ok' END
```

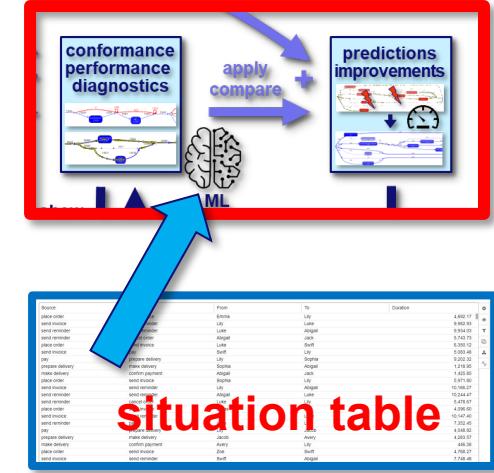
At least five people should be involved in each case.



What kind of situation table ?

There are five types of situation tables

- **Case-based situation table:** Each row (instance) corresponds to a case with variables.
 - **Event-based situation table:** Each row (instance) corresponds to an event.
 - **Resource-based situation table:** Each row (instance) corresponds to a resource.
 - **Event-pair-based situation table:** Each row (instance) corresponds to a pair of events.
 - Aggregate situation tables: Each row (instance) corresponds to a combination of cases and/or events.



The rows can be referred to as instances or “situations”

Semantics x-based situation table

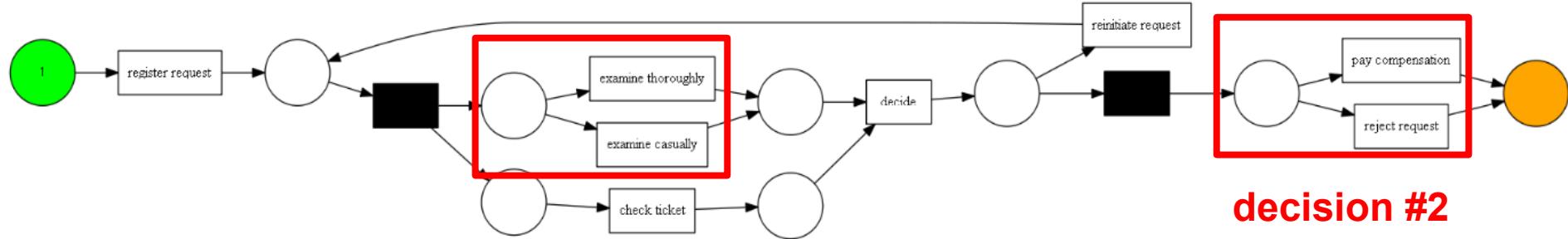
(x = event, case, resource, or event-pair)

- In an **x-based situation table**, each row refers to precisely one **x** and there cannot be multiple rows referring to the same **x**.
- The mapping from rows to **x-s**, is **functional**, **total**, and **injective**, but does not need to be **surjective**.
- For example:
 - Each row in an event-based table refers to one event and there cannot be multiple rows (instances/situations) that refer to the same event.
 - Each row in event-pair-based table refers to one pair of events and there cannot be multiple rows (instances/situations) that refer to the same event pair.



Event-based situation table or case-based situation table?

decision #1

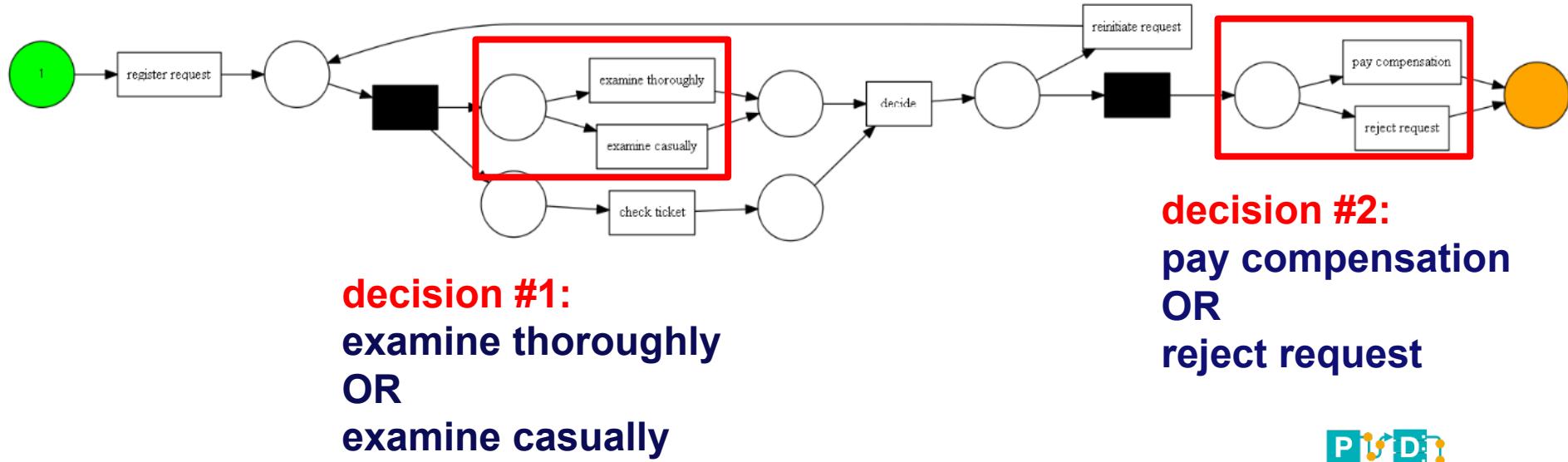


decision #2

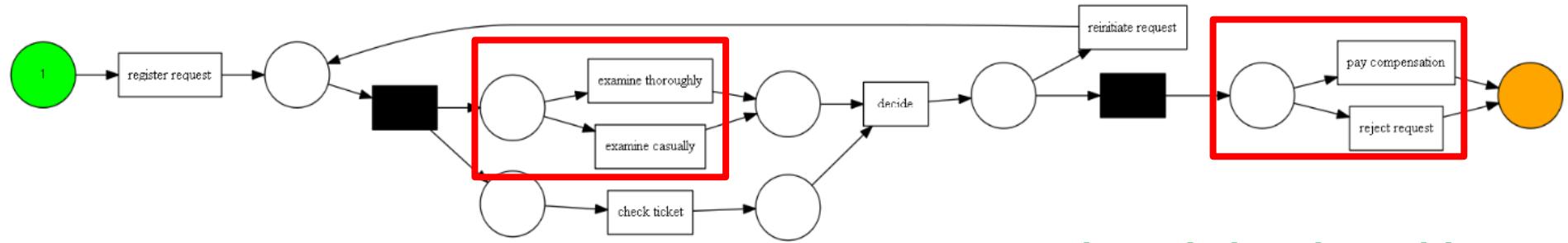
Recall the mapping from rows in the x-based situation table to x-s is functional, total, and injective, but does not need to be surjective.

It depends ...

Decision point #1: possibly multiple decisions for same case.
Decision point #2: only one decision for one case.



It depends ...



event-based
situation table

case-based situation table or
event-based situation table



Chair of Process
and Data Science

Decision mining using Celonis and RapidMiner



Basic Scheme

table with events (called an activity table in Celonis)



table with cases (called a case table in Celonis)

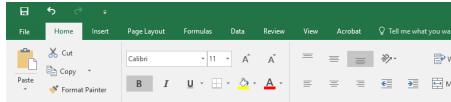


Process Query Language (PQL)



CASE	PRODUCT	ADDRESS	first resource	Total throughput time in h...	decision
1	SAMSUNG Galaxy J5	Munich	Caleb	239 pay	
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	201 pay	
3	APPLE iPhone 5s 16 GB	New York	Sophia	503 cancel order	
4	MOTOROLA Moto E 4G	New York	Sophia	498 cancel order	
5	SAMSUNG Core Prime G361	Aachen	Isabella	741 pay	
6	SAMSUNG Galaxy S4	Munich	Emma	406 pay	
7	MOTOROLA Moto G	Amsterdam	Lucas	598 pay	
8	APPLE iPhone 6 16 GB	Amsterdam	Sophia	209 pay	
9	APPLE iPhone 5s 16 GB	Munich	Aiden	412 pay	
10	HUAWEI P8 Lite	Amsterdam	Speedy	415 pay	
11	MOTOROLA Moto G	Munich	Emma	508 pay	
12	APPLE iPhone 5s 16 GB	Aachen	Jacob	480 pay	
13	HUAWEI P8 Lite	Munich	Speedy	409 pay	
14	SAMSUNG Core Prime G361	Munich	Sophia	331 pay	
15	SAMSUNG Galaxy S4	Aachen	Sophia		Export Data (XLSX)
16	SAMSUNG Galaxy S4	Aachen	Sophia		Export Data (CSV)
17	SAMSUNG Galaxy S4	Munich	Luke		Export Cases (XLSX)
18	SAMSUNG Galaxy S4	Munich	Lucas		Export Cases (CSV)
19	SAMSUNG Galaxy S4	New York	Luke		
20	APPLE iPhone 6s 64 GB	Munich	Sophia	359 pay	
21	APPLE iPhone 6s 64 GB	Aachen	Jacob	448 cancel order	
22	SAMSUNG Galaxy S4	Aachen	Aiden	1363 pay	
23	MOTOROLA Moto G	New York	Emma	177 pay	
24	MOTOROLA Moto G	Aachen	Speedy	505 cancel order	
25	APPLE iPhone 5s 16 GB	Munich	Sophia	987 pay	
26	HUAWEI P8 Lite	Aachen	Zoe	505 cancel order	
27	MOTOROLA Moto G	Aachen	Aiden	505 cancel order	
28	SAMSUNG Galaxy S4	New York	Aiden		

situation table



Clipboard		Font	Alignment				
		A	B	C	D	E	F
1	CASE	PRODUCT	ADDRESS	first resource	Total throughput	time	decision
2	1	SAMSUNG Galaxy S5	Munich	Caleb	239	pay	
3	2	APPLE iPhone 6s 16GB	Amsterdam	Lucas	209	pay	cancel order
4	3	APPLE iPhone 6s 16GB	New York	Sophia	303	paid	cancel order
5	4	MOTOROLA Moto E4	New York	Sophia	498	cancel order	
6	5	SAMSUNG Core Prime G361	Aachen	Isabella	741	pay	
7	6	SAMSUNG Galaxy S4	Munich	Emma	406	pay	
8	7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay	
9	8	APPLE iPhone 5s 16GB	Amsterdam	Sophia	209	pay	
10	9	APPLE iPhone 5s 16GB	Munich	Aiden	412	pay	
11	10	HUAWEI P8 Lite	Amsterdam	Speedy	415	pay	
12	11	MOTOROLA Moto G	Munich	Emma	508	pay	
13	12	APPLE iPhone 5s 16GB	Aachen	Jacob	480	pay	
14	13	HUAWEI P8 Lite	Munich	Speedy	469	pay	
15	14	SAMSUNG Core Prime G361	Munich	Sophia	331	pay	
16	15	SAMSUNG Galaxy S4	Aachen	Sophia	186	pay	
17	16	SAMSUNG Galaxy S4	Aachen	Olivia	378	pay	
18	17	SAMSUNG Galaxy S4	Munich	Luke	343	pay	
19	18	SAMSUNG Galaxy S4	Munich	Lucas	337	pay	
20	19	SAMSUNG Galaxy S4	New York	Luke	462	cancel order	
21	20	APPLE iPhone 5s 16GB	New York	Speedy	459	pay	
22	21	APPLE iPhone 5s 16GB	New York	Speedy	459	pay	
23	22	SAMSUNG Galaxy S4	New York	Speedy	459	pay	
24	23	MOTOROLA Moto G	New York	Speedy	459	pay	
25	24	MOTOROLA Moto G	New York	Speedy	459	pay	
26	25	APPLE iPhone 5s 16GB	Munich	Speedy	186	cancel order	
27	26	APPLE iPhone 5s 16GB	Munich	Speedy	177	cancel order	
28	27	APPLE iPhone 5s 16GB	Munich	Speedy	177	cancel order	

situation table



Challenges

- You need to load an **activity** and **case table**.
(Note: if you just upload events there is a trivial case table.)
- The Celonis Process Query Language (PQL) has over 200 operators, is very powerful and fast (>50 times faster than SQL), but is not easy to learn.
- Here PQL is introduced only by example.



Creating Situation Tables with PQL

The starting point for every data mining and machine learning technique are **data**. The input data used for many data mining and machine learning techniques typically have a tabular form. In the data table, each row is associated with a single entity, such as a process instance, a customer, an individual, an item, a document, objects, or records. Variables are often referred to as attributes, features, or data elements. In this course, we briefly cover data science topics related to supervised learning (e.g., Decision Trees) and unsupervised learning (e.g., Clustering). RapidMiner is a tool which provides a rich body of data science methods and algorithms to select and apply to the data. We use RapidMiner to demonstrate how algorithms such as Decision tree mining and K-means clustering can be applied.

The data describing a process come in the form of an event log. The event log is the starting point for process mining techniques and it contains information regarding activities, cases, resources, and all other entities or objects that are involved in a process. Using data mining and machine learning one can gain process insights that go beyond process discovery and conformance checking. For instance, we may want to analyze which case attributes influence the path cases take in a process, how resources handle cases, to which extent a workflow requires specific resources, and so on. These questions can be formulated into concrete data science problems and thus, the existing techniques can be applied to help answering them. That translation is, however, not trivial. Given a particular process related question, its transformation into a data science problem requires defining what data needs to be analyzed and how. In process mining, the basic unit of the event log is where typically instances refer to events and attributes to event attributes. It is only with the process input data that the output of any data mining and machine learning technique yields correct and valid insights into the data. This is where the Celonis Process Query Language (PQL) comes to help. Using PQL, we can use the provided event data to generate any kind of data table we need depending on the question at hand (see Figure 3).

In this course, we learn how for process mining tasks related to decision mining, performance analysis and organizational mining, we can generate situation tables from the event log. These data mining and machine learning techniques can then be applied to the generated data tables. More specifically, in this section we learn how to generate five types of situation tables:

- **Case-based situation table:** Each row (instance) corresponds to a case.
- **Event-based situation table:** Each row (instance) corresponds to an event.
- **Resource-based situation table:** Each row (instance) corresponds to a resource.
- **Event-pair-based situation table:** Each row (instance) corresponds to a pair of events.
- **Aggregate situation table:** Each row (instance) corresponds to a combination of cases and/or events.

<https://docs.celonis.com/en/pql-function-library.html>

To create and export OLAP tables

This analysis is empty.
Get started by adding a component to your analysis.

Add component

Process Overview 1 Product Overview 2 Case Overview 3 OLAP Table 4 New Sheet 5

New component

PROCESS ANALYSIS COMPONENTS

- Process Explorer
- Variant Explorer
- Throughput Time Search
- Activity Explorer

MACHINE LEARNING COMPONENTS

- Run ML Notebook

CHARTS AND TABLES

- OLAP Table
- Column Chart
- Pie Chart
- Donut Chart
- Line Chart

Component options

General options

Table title: OLAP Table

Component type: OLAP Table

DIMENSIONS

- CASE
- PRODUCT
- ADDRESS

KPIs

SORTING

- CASE

ADVANCED OPTIONS

- Distinct values

Done

10.0k of 10.0k cases selected 100%

COMPONENT + Edit PREVIEW

Order-Two-Table (Draft)

10.0k of 10.0k cases selected 100%

COMPONENT + Edit PREVIEW

Order-Two-Table-Analysis

Last edited 2 hours ago by Wil van der Aalst

Version Control Last data load 2h

Analysis settings

Saved formulas

Load script

Process explorer KPIs

Variables

Keyboard shortcuts

Activate LiveReload

10.0k of 10.0k cases selected 100%

COMPONENT + Edit PREVIEW

Order-Two-Table-Analysis

Last edited 2 hours ago by Wil van der Aalst

Version Control Last data load 2h

Analysis settings

Saved formulas

Load script

Process explorer KPIs

Variables

Keyboard shortcuts

Activate LiveReload

10.0k of 10.0k cases selected 100%

COMPONENT + Edit PREVIEW

Order-Two-Table-Analysis

Last edited 2 hours ago by Wil van der Aalst

Version Control Last data load 2h

Analysis settings

Saved formulas

Load script

Process explorer KPIs

Variables

Keyboard shortcuts

Activate LiveReload

If you do not change this setting, you cannot export the situation table.

CASE	PRODUCT	ADDR...	first res...	Total throughput...	decision
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay
2	APPLE iPhone 6s 64 GB	Amster...	Lucas	201	pay
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	cancel order
4	MOTOROLA Moto E 4G	New York	Sophia	498	cancel order
5	SAMSUNG Core Prime	Aachen	Isabella	741	pay
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay
7	MOTOROLA Moto G	Amster...	Lucas	598	pay
8	APPLE iPhone 6 16 GB	Amster...	Sophia	209	pay
9	APPLE iPhone 6s 16 GB	Munich	Aiden	412	pay
10	HUAWEI P8 Lite	Amster...	Speedy	415	pay
11	MOTOROLA Moto G	Munich	Emma	508	pay
12	APPLE iPhone 6s 16 GB	Aachen	Jacob	480	pay
13	HUAWEI P8 Lite	Munich	Speedy	409	pay
14	SAMSUNG Core Prime	Munich	Sophia	331	pay
15	SAMSUNG Galaxy S4	Aachen	Sophia	185	pay
16	SAMSUNG Galaxy S4	Aachen	Olivia	378	pay
17	SAMSUNG Galaxy S4	Munich	Luke	343	pay
18	SAMSUNG Galaxy S4	Munich	Lucas	337	pay
19	SAMSUNG Galaxy S4	New York	Luke	462	cancel order
20	APPLE iPhone 6s 64 GB	Munich	Sophia	211	pay
21	APPLE iPhone 6s 16 GB	Aachen	Speedy	353	pay
22	SAMSUNG Galaxy S4	Aachen	Aiden	359	pay
23	MOTOROLA Moto G	New York	Emma	448	cancel order
24	MOTOROLA Moto G	Aachen	Speedy	1363	pay
25	APPLE iPhone 5s 16 GB	Munich	Sophia	177	pay
26	HUAWEI P8 Lite	Aachen	Zoe	502	cancel order
27	MOTOROLA Moto G	Amster...	Aiden	987	pay
28	SAMSUNG Galaxy S4	New York	Aiden	503	cancel order

Example event log

 case_table.csv

5/27/2022 5:44 PM

Microsoft Excel Comma Separated Values File

400 KB

 event_table.csv

5/27/2022 5:44 PM

Microsoft Excel Comma Separated Values File

3,784 KB

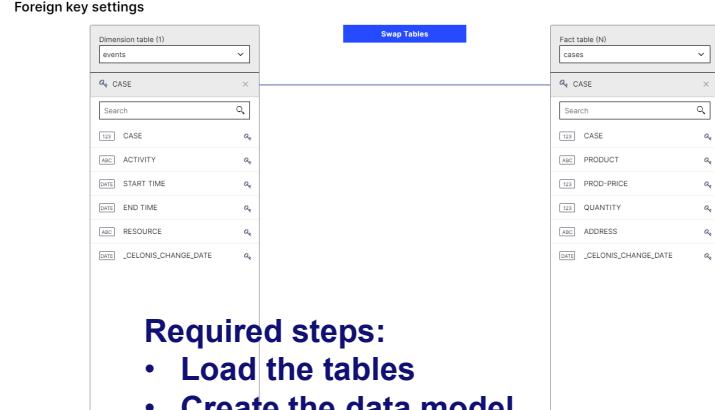
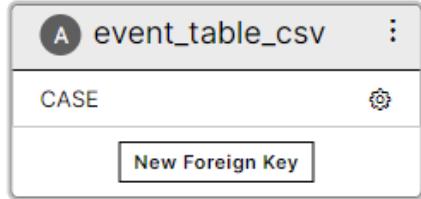
	A	B	C	D	E
1	case	product	prod-price	quantity	address
2	1	SAMSUNG Galaxy J5	219.99	3	Munich
3	2	APPLE iPhone 6s 64 GB	858	2	Amsterdam
4	3	APPLE iPhone 5s 16 GB	449	6	New York
5	4	MOTOROLA Moto E 4G	99.99	1	New York
6	5	SAMSUNG Core Prime G361	135	6	Aachen
7	6	SAMSUNG Galaxy S4	329	3	Munich
8	7	MOTOROLA Moto G	199	4	Amsterdam
9	8	APPLE iPhone 6 16 GB	639	6	Amsterdam
10	9	APPLE iPhone 5s 16 GB	449	5	Munich
11	10	HUAWEI P8 Lite	234	1	Amsterdam
12	11	MOTOROLA Moto G	199	2	Munich
13	12	APPLE iPhone 5s 16 GB	449	2	Aachen
14	13	HUAWEI P8 Lite	234	4	Munich
15	14	SAMSUNG Core Prime G361	135	3	Munich
16	15	SAMSUNG Galaxy S4	329	1	Aachen
17	16	SAMSUNG Galaxy S4	329	1	Aachen
18	17	SAMSUNG Galaxy S4	329	3	Munich
19	18				Munich
20	19				New York
21	20				Munich
22	21	APPLE iPhone 6s 64 GB	858	3	Aachen
23	22	SAMSUNG Galaxy S4	329	3	Aachen
24	23	MOTOROLA Moto G	199	1	New York
25	24	MOTOROLA Moto G	199	2	Aachen

case_table.csv

	A	B	C	D	E
1	case	activity	start time	end time	resource
2	1	place order	2015-01-05 09:00:07	2015-01-05 09:16:33	Caleb
3	2	place order	2015-01-05 10:18:21	2015-01-05 10:28:31	Lucas
4	3	place order	2015-01-05 11:54:49	2015-01-05 12:09:34	Sophia
5	4	place order	2015-01-05 14:07:45	2015-01-05 14:20:42	Sophia
6	5	place order	2015-01-05 15:33:38	2015-01-05 16:08:49	Isabella
7	6	place order	2015-01-05 17:25:23	2015-01-05 17:29:38	Emma
8	7	place order	2015-01-05 19:08:53	2015-01-05 19:21:26	Lucas
9	8	place order	2015-01-05 21:54:00	2015-01-05 21:59:36	Sophia
10	9	place order	2015-01-06 07:25:13	2015-01-06 07:31:06	Aiden
11	10	place order	2015-01-06 10:09:51	2015-01-07 05:31:14	Speedy
12	11	place order	2015-01-06 11:37:49	2015-01-06 11:41:29	Emma
13	12	place order	2015-01-06 13:33:45	2015-01-06 13:38:40	Jacob
14	13	place order	2015-01-06 15:25:38	2015-01-07 04:38:22	Speedy
15	14	place order	2015-01-06 17:09:23	2015-01-06 17:19:22	Sophia
16	15	place order	2015-01-06 18:36:53	2015-01-06 18:46:17	Sophia
17	16	place order	2015-01-06 21:26:54	2015-01-06 21:32:44	Olivia
18	17	place order	2015-01-07 04:42:36	2015-01-07 05:06:08	Luke
19	18				:05
20	19				Lucas
21	20				:36
22	21				:48
23	22				Sophia
24	23				Alexander
25	24				Lily
					Jacob
					Emily

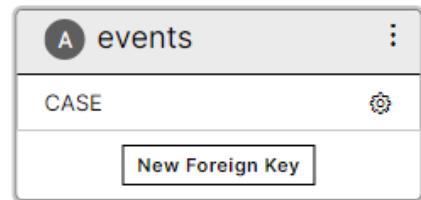
event_table.csv

Load into Celonis and create data model

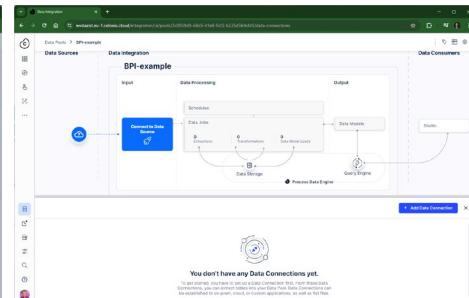
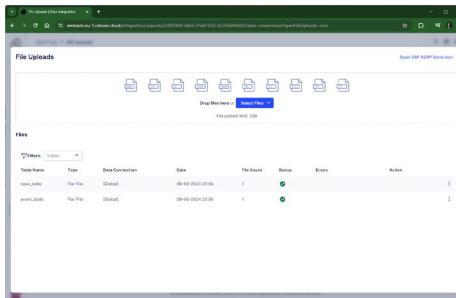
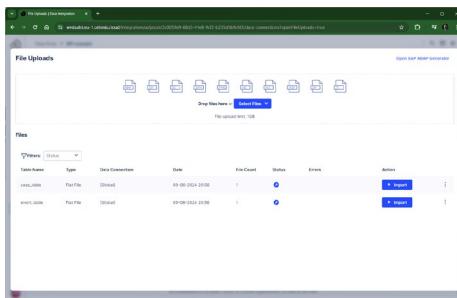
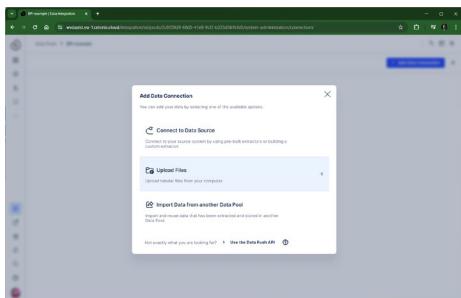
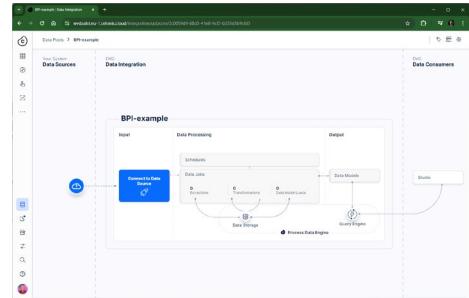
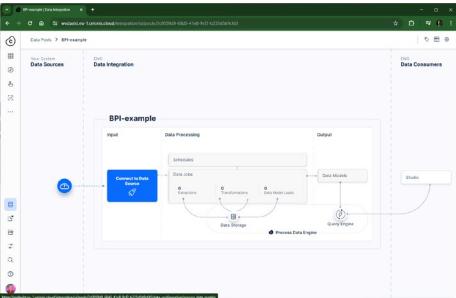
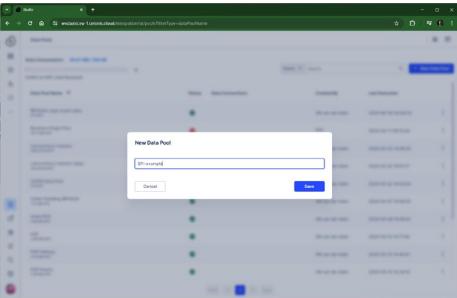
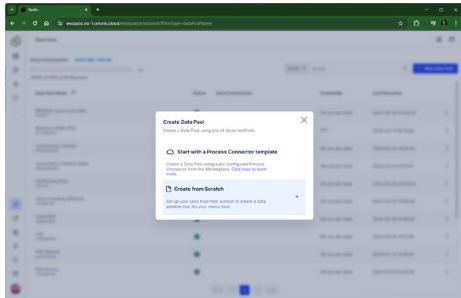


Required steps:

- Load the tables
- Create the data model
- Indicate activity and case table
- Relate activity and case table
- Create aliases (optional)
- Populate the data model



Load into Celonis and create data model



Load into Celonis and create data model

The collage consists of six screenshots illustrating the steps to load data into Celonis and create a data model:

- Data Models -> Add Data Model:** A modal window titled "Add Data Model" is shown, prompting the user to enter a name for the new data model.
- Data Models -> Activity Table:** A screen showing the "Activity Table" configuration step, where tables like "case_table" and "event_table" are listed under "Available items".
- Data Models -> Process:** A screen showing the "Process" configuration step, listing "Available items" such as "Order" and "Order Line".
- Activity Table Configuration:** A detailed view of the "Activity Table" configuration, showing the "TABLE NAME" as "event_table" and the note "Select event table to set activity table".
- Process Configuration:** A detailed view of the "Process" configuration, showing the "CASE_ID" column highlighted as a candidate key.
- Foreign key settings:** A configuration screen for defining foreign keys between tables like "case_table", "event_table", "customer", "product", "address", and "order_line".

Load into Celonis and create data model

The screenshots show the following steps in the Celonis Data Modeler:

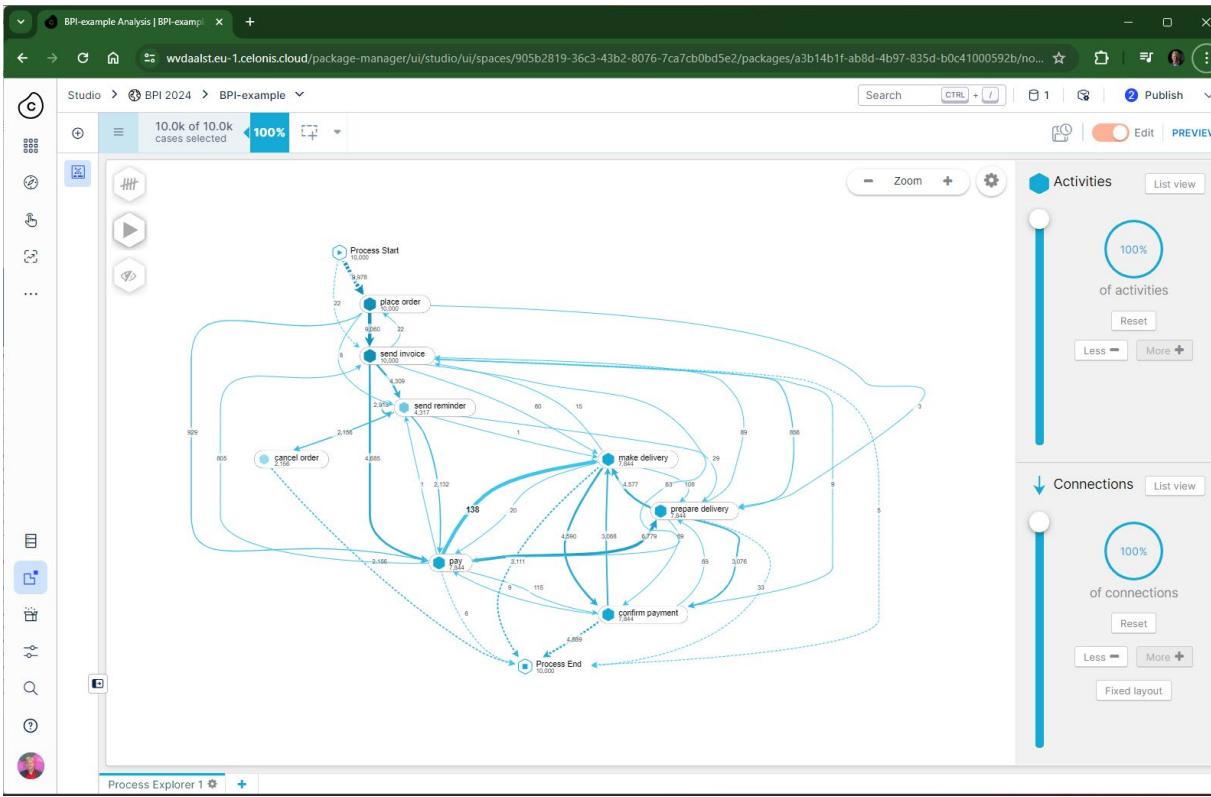
- Defining a new table: `event_table` is selected as the base table.
- Assigning a case table: `CASE` is selected as the case table.
- Setting primary keys: Primary keys are assigned to `event_table` and `CASE`.
- Table settings: A dialog shows the schema as `event_table`, name as `event_TABLE`, and table type as `FACT`. It also includes a note about giving a descriptive alternative name.
- Design key settings: Dimension tables are selected for `event_table` and fact tables for `CASE`.
- Data Modeler interface: Shows the Data Modeler workspace with tables `event_TABLE` and `CASE`.
- No live Data Model: A message indicates no live Data Model is present.
- Load in progress: A progress bar shows the status of the data load process.

Load into Celonis and create data model

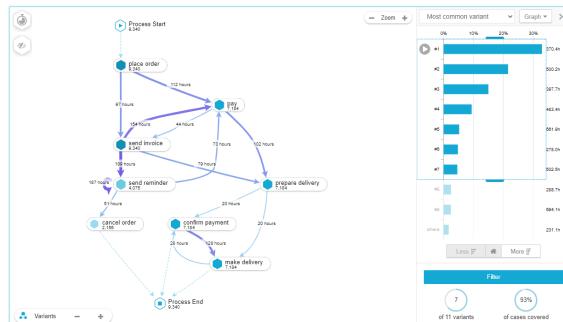
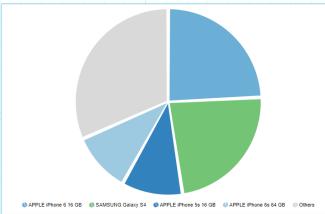
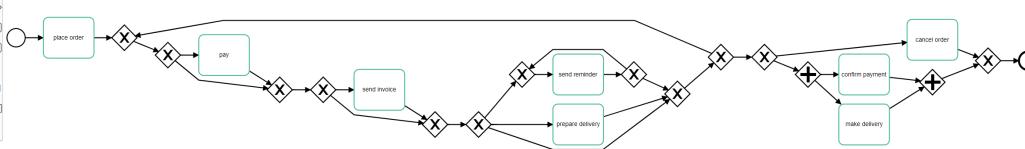
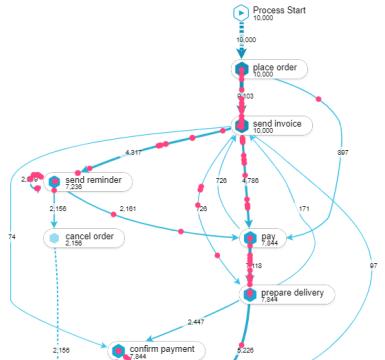
The collage illustrates the workflow from data loading to process analysis and modeling in the Celonis Platform:

- Top Left:** Data Integration interface showing a successful load of the "BPI-example" dataset.
- Top Middle:** Data Integration diagram illustrating the flow from Data Sources (SAP, Oracle, MySQL) through Data Processing (Data Jobs, Data Connectors, Data Mappers, Data Streamer, Data Transformer, Data Sink) to Data Consumers (Data Model, Data Engine).
- Top Right:** Process Mining interface showing a dashboard for "BPI 2024" with various process metrics and a "Create package" dialog.
- Bottom Left:** Process Mining interface showing a navigation menu and a search bar.
- Bottom Middle Left:** Process Mining interface showing a "New sheet" creation dialog with options like New Sheet, Process A, Process Overview, Process Explorer, Conformance, Social, and Case Explorer.
- Bottom Middle Right:** Process Explorer view showing a complex process graph with many nodes and connections.
- Bottom Right:** Process Explorer view showing a simplified process graph with fewer nodes and connections.

In Studio: Create a package and add an analysis (demonstrated before)



Here you can do many types of analysis



New Sheet: A new sheet waiting to be built.

Process AI: Detect and analyze deviations from the most common path.

Process Overview: Get the main insights on your process.

Process Explorer: Analyze and understand your process.

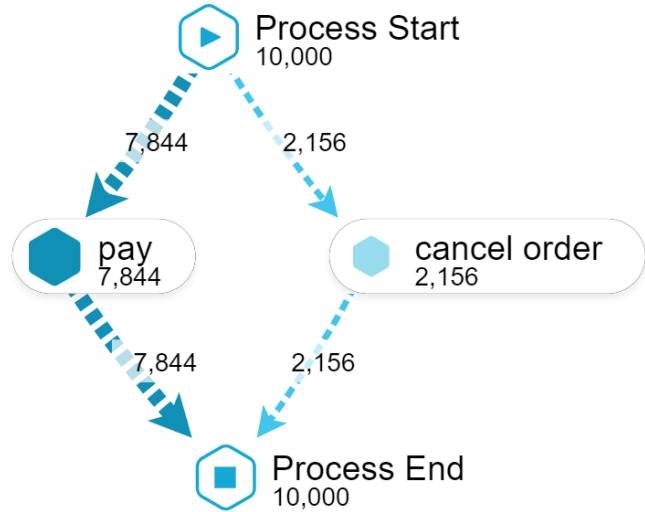
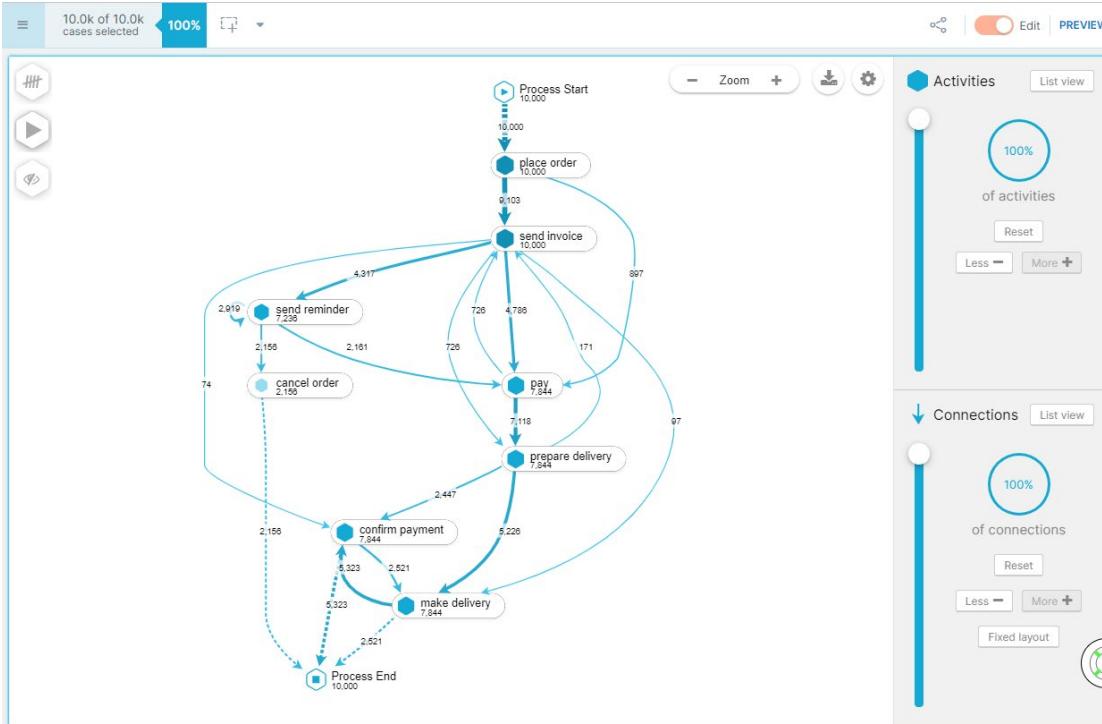
Conformance: Compare the real process to your target process.

Social: Understand how your team is working.

Case Explorer: Inspect individual cases.



Cases get paid or cancelled



Since there is precisely one payment or one cancellation per case, we can go for a case-based situation table.

Goal: Create situation table (Scenario 1)

The screenshot shows a software interface for creating a situation table. On the left, there is a preview window displaying a table with columns: CASE, PRODUCT, ADDRESS, first resource, total throughput ti..., and decision. The table contains 28 rows of data. On the right, a 'Component options' dialog is open, showing settings for an 'OLAP Table'. Under 'DIMENSIONS', the columns CASE, PRODUCT, and ADDRESS are listed. Under 'KPIs', there is a single entry. Under 'SORTING', the column CASE is selected. At the bottom, there is an 'ADVANCED OPTIONS' section.

CASE	PRODUCT	ADDRESS	first resource	total throughput ti...	decision
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	201	pay
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	cancel order
4	MOTOROLA Moto E 4G	New York	Sophia	498	cancel order
5	SAMSUNG Core Prime G361	Aachen	Isabella	741	pay
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay
8	APPLE iPhone 6 16 GB	Amsterdam	Sophia	209	pay
9	APPLE iPhone 5s 16 GB	Munich	Aiden	412	pay
10	HUAWEI P8 Lite	Amsterdam	Speedy	415	pay
11	MOTOROLA Moto G	Munich	Emma	508	pay
12	APPLE iPhone 5s 16 GB	Aachen	Jacob	480	pay
13	HUAWEI P8 Lite	Munich	Speedy	409	pay
14	SAMSUNG Core Prime G361	Munich	Sophia	331	pay
15	SAMSUNG Galaxy S4	Aachen	Sophia	186	pay
16	SAMSUNG Galaxy S4	Aachen	Olivia	378	pay
17	SAMSUNG Galaxy S4	Munich	Luke	343	pay
18	SAMSUNG Galaxy S4	Munich	Lucas	337	pay
19	SAMSUNG Galaxy S4	New York	Luke	462	cancel order
20	APPLE iPhone 6s 64 GB	Munich	Sophia	211	pay
21	APPLE iPhone 6s 64 GB	Aachen	Jacob	356	pay
22	SAMSUNG Galaxy S4	Aachen	Aiden	359	pay
23	MOTOROLA Moto G	New York	Emma	446	cancel order
24	MOTOROLA Moto G	Aachen	Speedy	1363	pay
25	APPLE iPhone 5s 16 GB	Munich	Sophia	177	pay
26	HUAWEI P8 Lite	Aachen	Zoe	505	cancel order
27	MOTOROLA Moto G	Amsterdam	Aiden	987	pay
28	SAMSUNG Galaxy S4	New York	Aiden	505	cancel order

- Each row (i.e., instance/situation) refers to one case.
- Each column refers to variable.
- The response variable has value pay or cancel order.
- The predictor variables list properties of the case that may influence the response variable.



Three case attributes

10.0k of 10.0k cases selected **100%**

Component options

General options

Table title:

Component type: OLAP Table

DIMENSIONS

- CASE
- PRODUCT
- ADDRESS

KPIs

SORTING

CASE

"cases"."CASE"

"cases"."PRODUCT"

"cases"."ADDRESS"

CASE	PRODUCT	ADDRESS	first resource	total throughput ti...	decision
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	201	pay
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	cancel order
4	MOTOROLA Moto E 4G	New York	Sophia	498	cancel order
5	SAMSUNG Core Prime G361	Aachen	Isabella	741	pay
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay
8	APPLE iPhone 6 16 GB	Amsterdam	Sophia	209	pay
9	APPLE iPhone 5s 16 GB	Munich	Aiden	412	pay
10	HUAWEI P8 Lite	Amsterdam	Speedy	415	pay
11	MOTOROLA Moto G	Munich	Emma	508	pay
12	APPLE iPhone 6s 64 GB	Aachen	Jacob	480	pay
13	SAMSUNG Galaxy S4	Munich	Speedy	409	pay
14	SAMSUNG Core Prime G361	Munich	Sophia	331	pay
15	SAMSUNG Galaxy S4	Aachen	Sophia	186	pay
16	SAMSUNG Galaxy S4	Munich	Olivia	378	pay
17	SAMSUNG Galaxy S4	Munich	Luke	343	pay
18	SAMSUNG Galaxy S4	Munich	Lucas	337	pay
19	SAMSUNG Galaxy S4	New York	Luke	462	cancel order
20	APPLE iPhone 6s 64 GB	New York		211	pay
21	APPLE iPhone 5s 16 GB	Aachen		356	pay
22	SAMSUNG Galaxy S4	Aachen	Aiden	359	pay
23	MOTOROLA Moto G	New York	Emma	448	cancel order

Example: Adding the case identifier

The screenshot shows a process mining tool interface with the following components:

- Top Bar:** PRODUCT, eye icon, edit icon (circled in red), list icon, close icon, three dots icon.
- Left Sidebar:** PRODUCT, Formatting, Standard (no format), Format, Units. Below this is a list of tables:
 - All
 - Standard Process KPI
 - cases
 - events
 - CASE - cases (highlighted with a red arrow)
 - PRODUCT - cases
 - PROD-PRICE - cases
 - QUANTITY - cases
 - ADDRESS - cases
 - _CELONIS_CHANGE_DATE - cases
 - CASE - events
 - ACTIVITY - events
 - START TIME - events
 - END TIME - events
- Editor Area:** A large black area labeled "EDITOR" containing the formula "1 \"cases\".\"CASE\"". A red arrow points from the sidebar entry "CASE - cases" to this formula.
- Code Editor:** A smaller window titled "EDITOR" showing the same formula "1 \"cases\".\"CASE\"".
- Preview Area:** A table titled "CASE - cases" showing the following data:

CASE	PRODUCT	ADDRESS	first resource	total throughput ti...	decision
1	1	Munich	Caleb	239	pay
2	2	Amsterdam	Lucas	201	pay
3	3	New York	Sophia	503	cancel order
- Bottom Right:** Chair of Process and Data Science logo (with letters P, D, A, S) and text "Chair of Process and Data Science".

Adding the first resource working on a case

PU_FIRST ("cases", "events"."RESOURCE")

CASE	PRODUCT	ADDRESS	first resource	total throughput time in ...	decision	
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay	
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	201	pay	
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	cancel order	
4	MOTOROLA Moto E 4G	New York	Sophia	498	cancel order	
5	SAMSUNG Core Prime G361	Aachen	Isabella	741	pay	
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay	
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay	
8	APPLE iPhone 6 16 GB	Amsterdam	Sophia	209	pay	
9	APPLE iPhone 5s 16 GB	Munich	Aidan	412	pay	

 Edit Formula

first resource
Formatting
Standard (no format)

Format
Units

Tables
Search KPIs..

All

Standard Process KPI

cases

events

Total throughput time in days

Ratio of cases flowing through an activity

Ratio of cases with a certain process flow

Average events per case

EDITOR

```
1 PU_FIRST ( "cases", "events"."RESOURCE" )
```

PU_FIRST
Description
 Returns the first element of the specified source column for each element in the given target table. An **order by** expression can be set to define the order that should be used to determine the first element.
PU_FIRST can be applied on any data type. The data type of the result is the same as the input column data type.

Syntax

```
PU_FIRST ( target_table, source_table.column [ , filter_expression ] [ , ORDER BY source_table.column [ asc | desc ] ] )
```

- target_table**: The table to which the aggregation result should be pulled. This can be:
 - a table from the data model. It needs to be, directly or indirectly, connected to the source_table, and there must be a IN relationship between the target_table and the source_table. Further documentation about relationships can be found in [Join functionality](#).
 - DOMAIN_TABLE or CONSTANT (see [Pull Up Aggregation Table Options](#)).
- source_table.column**: The column which should be aggregated for every row of the target_table.
- filter_expression** (optional): An optional filter expression to specify which values of the source_table.column should be taken into account for the aggregation.
- ORDER BY** (optional): Elements of the specified column are used to determine the first element. **asc** or **desc** can be used to use ascending or descending ordering. If the order direction is not specified, the ascending (**asc**) order is used. Using **PU_FIRST** with descending order is equivalent to using **PU_LAST** with ascending order.



Pull Up (PU) functions in PQL

⊕	PU_AVG
⊖	PU_COUNT
⌚	PU_COUNT_DISTINCT
⌚	PU_FIRST
⌚	PU_LAST
⌚	PU_MAX
⌚	PU_MEDIAN
⌚	PU_MIN
⌚	PU_MODE
⌚	PU_PRODUCT
⌚	PU_QUANTILE
⌚	PU_STDEV

- **PU_X (target_table, source_table.column)**
- The goal is to add a column to the target table.
- The result is computed over the corresponding rows in the source table.
- There needs to be a 1:N relationship between the target table and the source table.
- Typical pattern: **target_table = case table (cases)**, **source_table = activity table (events)**, and **column is some event attribute**.



Examples

- **PU_FIRST:** Returns the first element of the specified source column for each element in the given target table.
- **PU_LAST:** Returns the last element of the specified source column for each element of the given target table.
- **PU_COUNT:** Calculates the number of elements in the specified source column for each element in the given target table.
- **PU_AVG:** Calculates the average of the specified source column for each element in the given target table.
- **PU_MAX:** Calculates the maximum of the specified source column for each element in the given target table.
- **PU_COUNT_DISTINCT:** Calculates the number of distinct elements in the specified source column for each element in the given target table.
- Etc. See documentation.

Example PU function applications

Case	PU_FIRST Resource	PU_LAST Resource	PU_LAST Activity	PU_MIN Time	PU_MAX Time	Duration (hours)
1025	Lucas	Aubrey	make delivery	Thu Jul 16 2015 13:17:09	Wed Jul 29 2015 14:02:34	313
1026	Aiden	Lily	confirm payment	Thu Jul 16 2015 14:52:26	Wed Jul 29 2015 11:59:38	309
1027	Luke	Lily	confirm payment	Thu Jul 16 2015 17:02:59	Tue Aug 4 2015 09:59:03	448
1028	Aiden	Lily	cancel order	Thu Jul 16 2015 19:24:05	Fri Aug 28 2015 15:18:57	1029
1029	PU_MAX ("cases", REMAP_TIMESTAMPS("events"."START TIME",HOURS)) - PU_MIN ("cases", REMAP_TIMESTAMPS("events"."START TIME",HOURS))					
1030						
1031						
1032	Sophia	Kaylee	make delivery	Fri Jul 17 2015 11:25:00	Mon Aug 10 2015 17:50:22	582
1033	Lucas	Emily	confirm payment	Fri Jul 17 2015 13:08:07	Mon Jul 27 2015 15:23:53	242
1034	Sophia	Lily	confirm payment	Fri Jul 17 2015 14:34:21	Mon Aug 3 2015 19:15:15	413
1035	Caleb	Rush	make delivery	Fri Jul 17 2015 16:26:49	Thu Jul 30 2015 12:19:17	308
1036	Olivia	Charlotte	confirm payment	Fri Jul 17 2015 18:11:45	Fri Aug 21 2015 19:11:07	841
1037	Lucas	Jack	confirm payment	Fri Jul 17 2015 20:11:30	Tue Aug 4 2015 08:43:43	420
1038	Luke	Rush	make delivery	Fri Jul 17 2015 22:59:14	Wed Aug 26 2015 14:59:26	952
1039	Aiden	Emily	confirm payment	Sat Jul 18 2015 21:17:59	Fri Aug 7 2015 13:07:53	472

"cases"."CASE" 

 PU_FIRST ("cases", "events"."RESOURCE") 

 PU_LAST ("cases", "events"."RESOURCE") 

 PU_MIN ("cases", "events"."START TIME") 

 PU_MAX ("cases", "events"."START TIME") 

 PU_LAST ("cases", "events"."ACTIVITY") 



Adding the total throughput time in hours

CALC_THROUGHPUT(ALL_OCCURRENCE['Process Start'] TO ALL_OCCURRENCE['Process End'], REMAP_TIMESTAMPS("events"."START TIME", HOURS))

CASE	1	PRODUCT	ADDRESS	first resource	total throughput time in ...	decision	
1	SAMSUNG Galaxy J5	Munich	Caleb		239	pay	
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas		201	pay	
3	APPLE iPhone 5s 16 GB	New York	Sophia		503	cancel order	
4	MOTOROLA Moto E 4G	New York	Sophia		498	cancel order	
5	SAMSUNG Core Prime G361	Aachen	Isabella		741	pay	
6	SAMSUNG Galaxy S4	Munich	Emma		406	pay	
7	MOTOROLA Moto G	Amsterdam	Lucas		598	pay	
8	APPLE iPhone 6 16 GB	Amsterdam	Sophia		209	pay	
9	APPLE iPhone 5s 16 GB	Munich	Aidan		410	pay	



 Edit Formula

total throughput time in hours   Rounded number (#,###)  Format  Units 

Tables 

All	Search KPIs..
Standard Process KPI	Total throughput time in days
cases	Ratio of cases flowing through an activity
events	Ratio of cases with a certain process flow
	Average events per case

EDITOR

```
1 ✓ CALC_THROUGHPUT (
2   ALL_OCCURRENCE [ 'Process Start' ]
3   TO
4   ALL_OCCURRENCE [ 'Process End' ],
5   REMAP_TIMESTAMPS ( "events"."START TIME", HOURS )
6 )
```

The **REMAP_TIMESTAMPS** function counts the number of passed time units (here hours) for given dates since the epoch year (1970-01-01 00:00:00.000).



Chair of Process
and Data Science

CALC_THROUGHPUT(ALL_OCCURRENCE['Process Start'] TO ALL_OCCURRENCE['Process End'], REMAP_TIMESTAMPS("events"."START TIME", HOURS))

CALC_THROUGHPUT

Description

Throughput is used to calculate, for each case, the time between two activities. From which activity the calculated throughput time should start and at which it should end can be configured through range specifiers.

Syntax

```
CALC_THROUGHPUT ( begin_rangeSpecifier TO end_rangeSpecifier, timestamps [, activityTable.stringColumn ] )
```

- **begin_rangeSpecifier:** FIRST_OCCURRENCE['activity'] | LAST_OCCURRENCE['activity'] | CASE_START | ALL_OCCURRENCE['']
 - FIRST_OCCURRENCE['activity']: Throughput time starts at the first occurrence of the specified activity type.
 - LAST_OCCURRENCE['activity']: Throughput time starts at the last occurrence of the specified activity type.
 - CASE_START: Throughput time starts at the first activity of the case.
 - ALL_OCCURRENCE['']: Has the same meaning as CASE_START. The string parameter is ignored, but has to be specified.
- **end_rangeSpecifier:** FIRST_OCCURRENCE['activity'] | LAST_OCCURRENCE['activity'] | CASE_END | ALL_OCCURRENCE['']
 - FIRST_OCCURRENCE['activity']: Throughput time ends at the first occurrence of the specified activity type.
 - LAST_OCCURRENCE['activity']: Throughput time ends at the last occurrence of the specified activity type.
 - CASE_END: Throughput time ends at the last activity of the case.
 - ALL_OCCURRENCE['']: Has the same meaning as CASE_END. The string parameter is ignored, but has to be specified.
- **timestamps:** Integer column of an activity table, often REMAP_TIMESTAMPS is used to convert a TIMESTAMP column.

https://docs.celonis.com/en/calc_throughput.html

REMAP_TIMESTAMPS

Description

The REMAP_TIMESTAMPS function counts the number of passed time units for given dates since the epoch year (1970-01-01 00:00:00.000). The timestamps for which to calculate the passed time and also the time unit to use, are given as a parameter to the function call. Additionally, the user can specify a CALENDAR configuration which allows to restrict the dates considered in the calculations. For example, using the WEEKDAY_CALENDAR allows to only consider certain valid weekdays in the calculations.

Syntax

```
REMAP_TIMESTAMPS ( table.Column, timeUnit [, calendarSpecification ] [, calendarIdColumn ] )
```

- **column:** The column containing the timestamps to be remapped.
- **timeUnit:** The time unit to map the calculation to. One of DAYS, HOURS, MINUTES, SECONDS or MILLISECONDS
- **calendarSpecification:** One of WEEKDAY_CALENDAR, FACTORY_CALENDAR, WORKDAY_CALENDAR, or INTERSECT.
- **calendarIdColumn:** Column to create a mapping between the respective activities and their used calendar specification. This is mandatory when using multiple calendar specifications. For more details, please take a look at the respective documentation of the [DateTime Calendar](#).

https://docs.celonis.com/en/remap_timestamps.html



Many ways to achieve the same result

CASE	Duration (hours) Alt1	Duration (hours) Alt2	Duration (hours) Alt3	Duration (days) Alt1	Duration (days) Alt2	Duration (days) Alt3
1025	313	313	313	13	13	13
1026	309	309	309	13	13	13
1027	448	448	448	19	19	19
1028	1029	1029	1029	43	43	43
1029	784	784	784	33	33	33
1030	352	352	352	14	14	14
1031	75	75	75	31	31	31
1032	582	582	582	24	24	24
1033	242	242	242	10	10	10
1034	413	413	413	17	17	17
1035	308	308	308	13	13	13

"cases"."CASE"

CALC_THROUGHPUT (CASE_START TO CASE_END, REMAP_TIMESTAMPS("events"."START TIME", HOURS))

CALC_THROUGHPUT(ALL_OCCURRENCE['Process Start'] TO ALL_OCCURRENCE['Process End'], REMAP_TIMESTAMPS("events"."START TIME", HOURS))

PU_MAX ("cases", REMAP_TIMESTAMPS("events"."START TIME",HOURS)) - PU_MIN ("cases", REMAP_TIMESTAMPS("events"."START TIME",HOURS))

CALC_THROUGHPUT (CASE_START TO CASE_END, REMAP_TIMESTAMPS("events"."START TIME", DAYS))

CALC_THROUGHPUT(ALL_OCCURRENCE['Process Start'] TO ALL_OCCURRENCE['Process End'], REMAP_TIMESTAMPS("events"."START TIME", DAYS))

PU_MAX ("cases", REMAP_TIMESTAMPS("events"."START TIME",DAYS)) - PU_MIN ("cases", REMAP_TIMESTAMPS("events"."START TIME",DAYS))

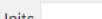
Adding the response variable

CASE WHEN PROCESS EQUALS 'pay' THEN 'pay' WHEN PROCESS EQUALS 'cancel order' THEN 'cancel order' END

CASE	PRODUCT	ADDRESS	first resource	total throughput time in ...	decision	
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay	
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	201	pay	
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	cancel order	
4	MOTOROLA Moto E 4G	New York	Sophia	498	cancel order	
5	SAMSUNG Core Prime G361	Aachen	Isabella	741	pay	
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay	
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay	
8	APPLE iPhone 6 16 GB	Amsterdam	Sophia	209	pay	
9	APPLE iPhone 5s 16 GB	Munich	Aidan	442	pay	

 Edit Formula



decision  Formatting Standard (no format)  Format  Units 

Tables  EDITOR

All	Total throughput time in days
Standard Process KPI	Ratio of cases flowing through an activity
cases	Ratio of cases with a certain process flow
events	Average events per case

Search KPIs..

```
1 CASE
2 WHEN PROCESS EQUALS 'pay' THEN 'pay'
3 WHEN PROCESS EQUALS 'cancel order' THEN 'cancel order'
4 END
```



Many ways to achieve the same result

(for this particular example)

```
CASE WHEN PROCESS EQUALS 'pay' THEN 'pay' WHEN PROCESS EQUALS 'cancel  
order' THEN 'cancel order' END
```

```
CASE WHEN PROCESS EQUALS 'pay' THEN 'pay' ELSE 'cancel order' END
```

```
CASE WHEN MATCH_PROCESS_REGEX ( "events"."ACTIVITY" , 'pay' ) = 1 THEN 'pay'  
WHEN MATCH_PROCESS_REGEX ( "events"."ACTIVITY" , 'cancel order' ) = 1 THEN  
'cancel order' END
```

```
CASE WHEN MATCH_PROCESS_REGEX ( "events"."ACTIVITY" , 'pay' ) = 1 THEN 'pay'  
ELSE 'cancel order' END
```

<https://docs.celonis.com/en/case-when.html>

<https://docs.celonis.com/en/process-equals.html>

https://docs.celonis.com/en/match_process_regex.html

MATCH_PROCESS_REGEX returns an INT value for each case which is 1 if the variant matches the pattern or 0 if it does not match.

Name PROCESS EQUALS is misleading here (activity needs to appear in variant), but the operator has many more options.



MATCH_PROCESS_REGEX

- MATCH_PROCESS_REGEX (activity_table.string_column, regular_expression)

Syntax	Meaning	Example
''	Case contains the activity	'A'
^	Case starts with the activity	^ 'Scan Invoice'
\$	Case ends with the activity	'B' \$
>>	Activities directly follow	'A' >> 'B'
	Logical OR	'A' 'B'
()	Group of activities	('A' 'B') >> ('C' >> 'D')
*	0 or more occurrences	(A' >> 'B')*
+	1 or more occurrences	(A' >> 'B')+
?	0 or 1 occurrences	(A' >> 'B')?
{<from>, <to>}	Between <from> and <to> occurrences	(A' >> 'B'){1, 3}
'*''	Any activity matches	'A' >> ('*' + >> 'B'
ANY	Any activity matches	'A' >> (ANY)+ >> 'B'
.	Any activity matches	'A' >> . >> 'C'
LIKE "%...%"	Activities that contain string	'A' >> LIKE '% Invoice%'
[,]	Set of activities of which one needs to match	'A' >> ['B', 'D', 'E'] >> 'C'
[! ,]	Set of activities of which none matches	'A' >> [! 'B'] >> 'C'
AS	Gives an alias to a regex	(A' >> (ANY)* AS sequence, sequence >> 'B'

MATCH_PROCESS_REGEX matches the variants of a process based on a regular expression. The regular expression defines a pattern over the activities of the variant. It returns an INT value which is 1 if the variant matches the pattern or 0 if it does not match.

Example: What are the cases where the customer pays immediately?

MATCH_PROCESS_REGEX ("events"."ACTIVITY", 'place order'>>'pay')



Export OLAP Table

(e.g., for RapidMiner)

CASE	PRODUCT	ADDRESS	first resource	Total throughput	decision
1	SAMSUNG Gal...	Munich	Caleb	239	pay
2	APPLE iPhone ...	Amsterdam	Lucas	201	pay
3	APPLE iPhone ...	New York	Sophia	503	cancel order
4	MOTOROLA M...	New York	Sophia	498	cancel order
5	SAMSUNG Cor...	Aachen	Isabella	741	pay
6	SAMSUNG Gal...	Munich	Emma	406	pay
7	MOTOROLA M...	Amsterdam	Lucas		
8	APPLE iPhone ...	Amsterdam	Sophia		
9	APPLE iPhone ...	Munich	Aiden		
10	HUAWEI P8 Lite	Amsterdam	Speedy		
11	MOTOROLA M...	Munich	Emma		
12	APPLE iPhone ...	Aachen	Jacob		
13	HUAWEI P8 Lite	Munich	Speedy		
14	SAMSUNG Cor...	Munich	Sophia		
15	SAMSUNG Gal...	Aachen	Sophia		
16	SAMSUNG Gal...	Aachen	Olivia		
17	SAMSUNG Gal...	Munich	Luke		
18	SAMSUNG Gal...	Munich	Lucas		
19	SAMSUNG Gal...	New York	Luke	462	cancel order
20	APPLE iPhone ...	Munich	Sophia	211	pay
21	APPLE iPhone ...	Aachen	Jacob	356	pay
22	SAMSUNG Gal...	Aachen	Aiden	359	pay

Analysis settings

General settings Saved formulas

General settings

Allow excel and csv export of analysis components.

Raw data export limit

20000

Allow BPMN export of the Process Explorer and Variant Explorer.

- Settings
- Component filter
- Export Data (XLSX)
- Export ...
- Export Data (CSV)
- Layers...
- Export Cases (XLSX)
- Copy component
- Export Cases (CSV)
- Delete
- Copy query



Chair of Process
and Data Science

Goal: Create situation table (Scenario 2)

- Case-based situation table: Each row (instance) corresponds to a case with variables.
- Event-based situation table: Each row (instance) corresponds to an event.
- Resource-based situation table: Each row (instance) corresponds to a resource.
- Event-pair-based situation table: Each row (instance) corresponds to a pair of events.
- Aggregate situation tables: Each row (instance) corresponds to a combination of cases and/or events.



CASE	PRODUCT	ADDRESS	first resource	total throughput tim...	decision
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay
2	APPLE iPhone 6s 64...	Amsterdam	Lucas	201	pay
3	APPLE iPhone 12 128...	New York	Sophia	503	cancel order
4	MOTOROLA Moto E ...	New York	Sophia	498	cancel order
5	SAMSUNG Core Pri...	Aachen	Isabella	741	pay
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay
8	APPLE iPhone 6 16 ...	Amsterdam	Sophia	209	pay
9	APPLE iPhone 5s 16...	Munich	Aiden	412	pay
10	HUAWEI P8 Lite	Amsterdam	Speedy	415	pay
11	MOTOROLA Moto G	Munich	Emma	508	pay
12	APPLE iPhone 5s 16...	Aachen	Jacob	480	pay
13	HUAWEI P8 Lite	Munich	Speedy	409	pay
14	SAMSUNG Core Pri...	Munich	Sophia	331	pay
15	SAMSUNG Galaxy S4	Aachen	Sophia	186	pay

The previous solution only works if for each case the decision to pay or cancel is made precisely once!



Creating an event-based situation table

- Select the events using a filter.
- Add variables.

FILTER "events"."ACTIVITY" IN ('pay', 'cancel order')

ACTIVITY	CASE
place order	1
send invoice	1
pay	1
prepare delivery	1
make delivery	1
confirm payment	1
place order	2
place order	2
pay	2
send invoice	2
prepare delivery	2
confirm payment	2
make delivery	2
place order	3
send invoice	3
send reminder	3
send reminder	3
cancel order	3

"events"."ACTIVITY"

"events"."CASE"

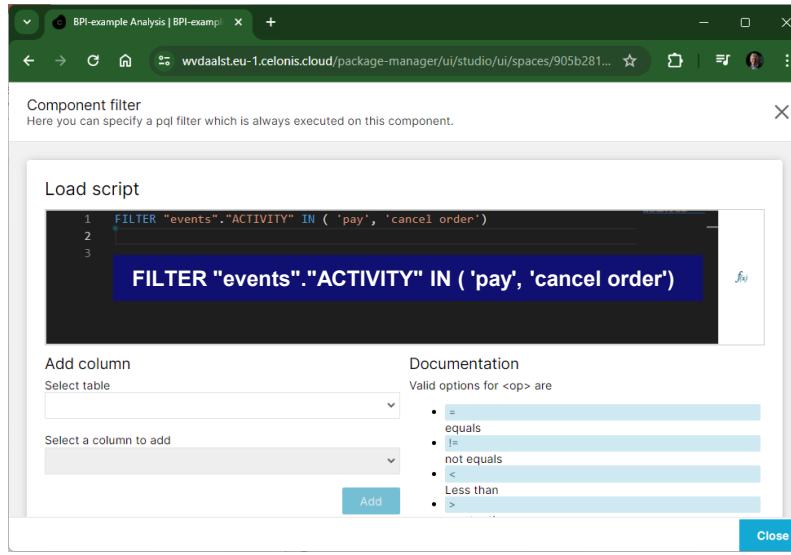
ACTIVITY	CASE
pay	2
send invoice	2
prepare delivery	2
confirm payment	2
make delivery	2
place order	3
send invoice	3
send reminder	3
cancel order	3
place order	4
send invoice	5
send reminder	5
send reminder	5



Chair of Process
and Data Science

Creating an event-based situation table

- Select the events using a filter.
- Add variables.



The screenshot shows the 'PREVIEW' tab of a situation table in Celonis Studio. The table has two columns: 'Event' and 'Row ID'. The 'Event' column shows repeated entries for 'pay' and 'cancel order'. The 'Row ID' column shows integers from 8 to 26. The top of the table displays the statistics: '10.0k of 10.0k cases selected' and '100%'. The table includes standard data manipulation icons like copy, paste, and delete.

Event	Row ID
pay	8
pay	9
pay	10
pay	11
pay	12
pay	13
pay	14
pay	15
pay	16
pay	17
pay	18
cancel order	19
pay	20
pay	21
pay	22
cancel order	23
pay	24
pay	25
cancel order	26

Access Load Script: Two ways

The screenshot shows the Process Modeler interface with two main panels. On the left, a sidebar menu is open, showing options like 'Order-Two-Table', 'Analysis settings', 'Saved formulas', 'Load script' (which is expanded), 'Process explorer' (with a red arrow pointing to it), 'Variables', 'Keyboard shortcuts', and 'Activate LiveReload'. On the right, a table titled 'ACTIVITY' and 'CASE' is displayed, listing 15 rows of activity types (pay, cancel order) and their corresponding case numbers (1 to 15). A vertical toolbar on the right side of the table includes icons for gear, eye, filter, download, and refresh, with a red arrow pointing to the filter icon, which is labeled 'Component filter'.

ACTIVITY	CASE
pay	1
pay	2
cancel order	3
cancel order	4
pay	5
pay	6
pay	7
pay	8
pay	9
pay	10
pay	11
pay	12
pay	13
pay	14
pay	15

It is a global setting! Note that it does not influence the process explorer.



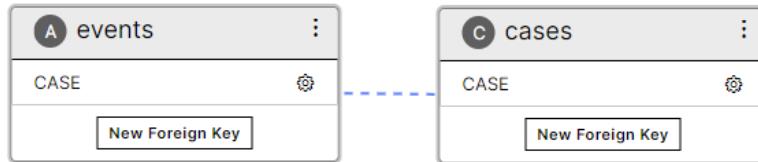
Filtering

Load script

```
1 FILTER "events"."ACTIVITY" IN ( 'pay', 'cancel order')
```

Filter can use any event or case attribute (note the implicit joins).

ACTIVITY	CASE	
pay	1	
pay	2	
cancel order	3	
cancel order	4	
pay	5	
pay	6	
pay	7	
pay	8	
pay	9	
pay	10	
pay	11	
pay	12	
pay	13	



Alternative filters

Load script Select events not conducted by Wil.

```
1 FILTER "events"."RESOURCE" != 'Wil'
```

Load script Select events conducted by Wil or Pete.

```
1 FILTER "events"."RESOURCE" in ('Wil', 'Pete')
```

Load script Select events belonging to cases where the customer paid immediately.

```
1 FILTER MATCH_PROCESS_REGEX ( "events"."ACTIVITY", 'place order'>>'pay') = 1
```

Load script Select events belonging to iPhone orders.

```
1 FILTER "cases"."PRODUCT" LIKE 'iPhone'
```

Adding variables after setting

FILTER "events"."ACTIVITY" IN ('pay', 'cancel order')

CASE	PRODUCT	ADDRESS	first resource	total throughput tim...	decision	
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay	
2	APPLE iPhone 6s 64...	Amsterdam	Lucas	201	pay	
3	APPLE iPhone 5s 16...	New York	Sophia	503	cancel order	
4	MOTOROLA Moto E ...	New York	Sophia	498	cancel order	
5	SAMSUNG Core Pri...	Aachen	Isabella	741	pay	
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay	
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay	
8	APPLE iPhone 6 16 ...	Amsterdam	Sophia	209	pay	
9	APPLE iPhone 5s 16...	Munich	Aiden	412	pay	
10	HUAWEI P8 Lite	Amsterdam	Speedy	415	pay	
11	MOTOROLA Moto G	Munich	Emma	508	pay	
12	APPLE iPhone 5s 16...	Aachen	Jacob	480	pay	
13	HUAWEI P8 Lite	Munich	Speedy	409	pay	
14	SAMSUNG Core Pri...	Munich	Sophia	331	pay	
15	SAMSUNG Galaxy S4	Aachen	Sophia	186	pay	

DIMENSIONS	Custom dimension +
CASE	
PRODUCT	
ADDRESS	
first resource	
total throughput time in hours	
decision	



Similar to before, but now each instance is an event and not a case!

CASE	PRODUCT	ADDRESS	first resource	total throughput time in hours	decision
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay
2	APPLE iPhone 6s 64...	Amsterdam	Lucas	201	pay
3	APPLE iPhone 5s 16...	New York	Sophia	503	cancel order
4	MOTOROLA Moto E ...	New York	Sophia	498	cancel order
5	SAMSUNG Core Pri...	Aachen	Isabella	741	pay
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay
8	APPLE iPhone 6 16 ...	Amsterdam	Sophia	209	pay
9	APPLE iPhone 5s 16...	Munich	Aiden	412	pay
10	HUAWEI P8 Lite	Amsterdam	Speedy	415	pay
11	MOTOROLA Moto G	Munich	Emma	508	pay
12	APPLE iPhone 5s 16...	Aachen	Jacob	480	pay
13	HUAWEI P6 Lite	Munich	Speedy	409	pay
14	SAMSUNG Core Pri...	Munich	Sophia	331	pay
15	SAMSUNG Galaxy S4	Aachen	Sophia	186	pay

Same results as before because once per case



Let's change the filter

Load script

```
1 FILTER "events"."ACTIVITY" IN ( 'pay', 'cancel order', 'send reminder')
```



CASE	PRODUCT	ADDRESS	first resource	total throughput time in h...	decision	
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay	
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	201	pay	
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	send reminder	
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	send reminder	
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	cancel order	
4	MOTOROLA Moto E 4G	New York	Sophia	498	send reminder	
4	MOTOROLA Moto E 4G	New York	Sophia	498	send reminder	
4	MOTOROLA Moto E 4G	New York	Sophia	498	cancel order	
5	SAMSUNG Core Prime G...	Aachen	Isabella	741	send reminder	
5	SAMSUNG Core Prime G...	Aachen	Isabella	741	send reminder	
5	SAMSUNG Core Prime G...	Aachen	Isabella	741	pay	
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay	
7	MOTOROLA Moto G	Amsterdam	Lucas	598	send reminder	
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay	
8	APPLE iPhone 6 16 GB	Amsterdam	Sophia	209	pay	

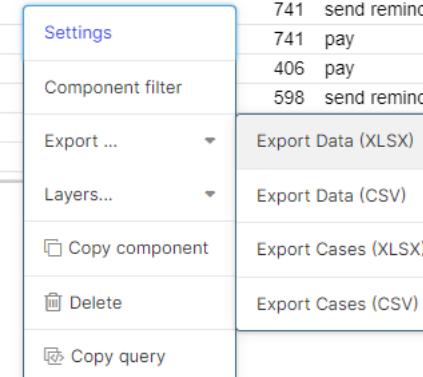
Now there may be multiple instances per case!

Remember: At any stage we can export the situation table and analyze with RapidMiner (or any other tool)

CASE	PRODUCT	ADDRESS	first resource	total throughput time in h...	decision
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	201	pay
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	send reminder
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	send reminder
3	APPLE iPhone 5s 16 GB	New York	Sophia	503	cancel order
4	MOTOROLA Moto E 4G	New York	Sophia	498	send reminder
4	MOTOROLA Moto E 4G	New York	Sophia	498	send reminder
4	MOTOROLA Moto E 4G	New York	Sophia	498	cancel order
5	SAMSUNG Core Prime G...	Aachen	Isabella	741	send reminder
5	SAMSUNG Core Prime G...	Aachen	Isabella	741	send reminder
5	SAMSUNG Core Prime G...	Aachen	Isabella	741	pay
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay
7	MOTOROLA Moto G	Amsterdam	Lucas	598	send reminder
7	MOTOROLA Moto G	Amsterdam	Lucas		
8	APPLE iPhone 6 16 GB	Amsterdam	Sophia		



Remove before or after export. In Celonis use the “eye” symbol in list of dimensions (right)



Analyzing the situation table in RapidMiner

RapidMiner Studio Free 9.10.008 @ DESKTOP-WVDAALST

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep More Find data, operators... etc All Studio

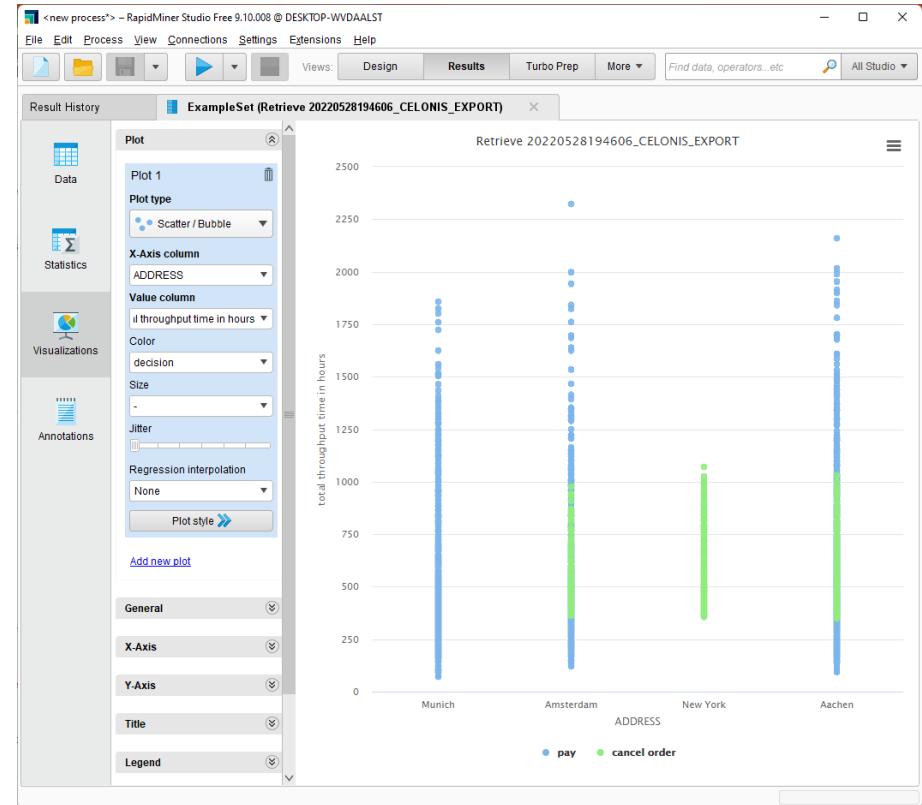
Result History ExampleSet (//Local Repository/data/20220528194606_CELONIS_EXPORT) ExampleSet (Retrieve 20220528194606_CELONIS_EXPORT) Tree (Decision Tree)

Data Statistics Visualizations Annotations

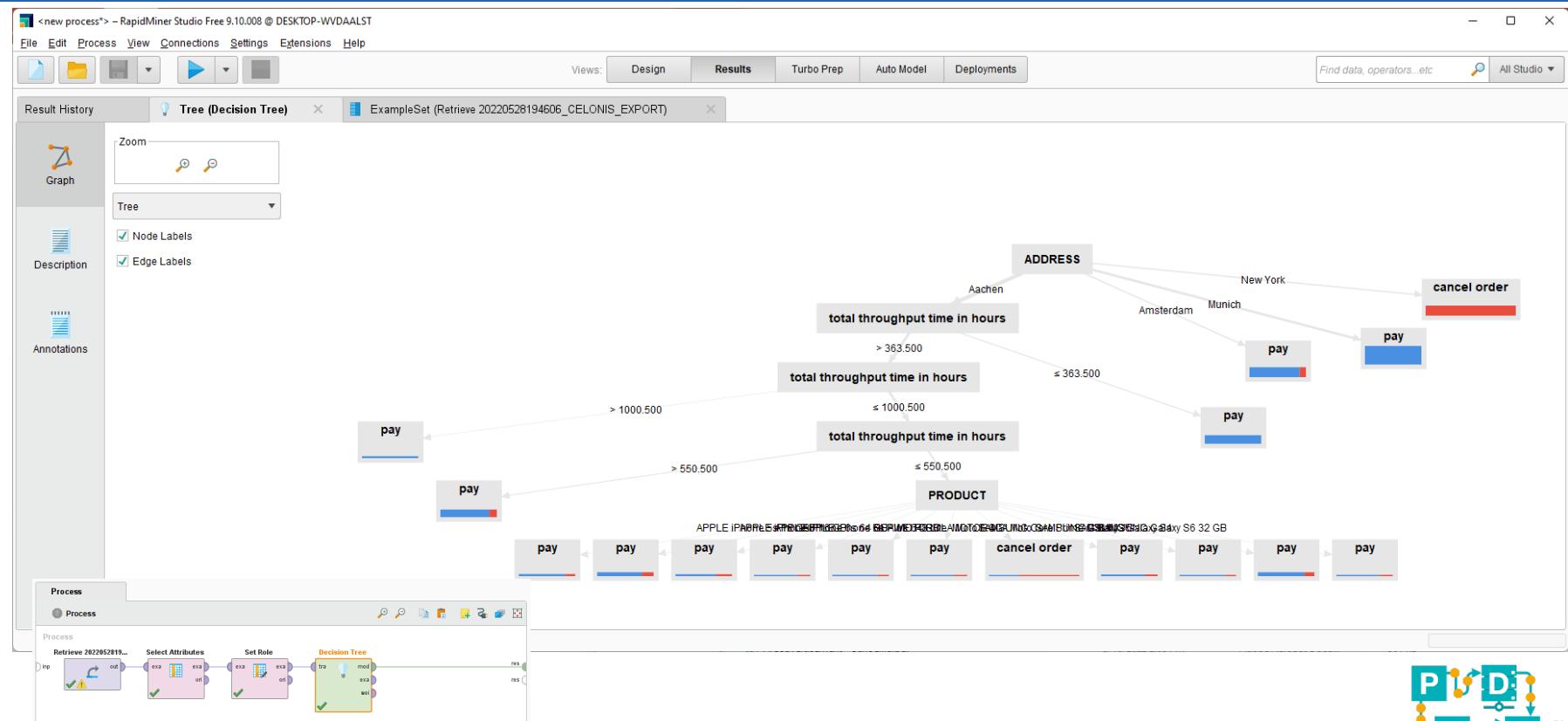
Open in Turbo Prep Auto Model Filter (10,000 / 10,000 examples): all

Row No.	CASE	PRODUCT	ADDRESS	first resource	total through...	decision
1	1	SAMSUNG G...	Munich	Caleb	239	pay
2	2	APPLE iPhone...	Amsterdam	Lucas	201	pay
3	3	APPLE iPhone...	New York	Sophia	503	cancel order
4	4	MOTOROLA...	New York	Sophia	498	cancel order
5	5	SAMSUNG C...	Aachen	Isabella	741	pay
6	6	SAMSUNG G...	Munich	Emma	406	pay
7	7	MOTOROLA...	Amsterdam	Lucas	598	pay
8	8	APPLE iPhone...	Amsterdam	Sophia	209	pay
9	9	APPLE iPhone...	Munich	Aiden	412	pay
10	10	HUAWEI P8 L...	Amsterdam	Speedy	415	pay
11	11	MOTOROLA...	Munich	Emma	508	pay
12	12	APPLE iPhone...	Aachen	Jacob	480	pay
13	13	HUAWEI P8 L...	Munich	Speedy	409	pay
14	14	SAMSUNG C...	Munich	Sophia	331	pay
15	15	SAMSUNG G...	Aachen	Sophia	186	pay
16	16	SAMSUNG G...	Aachen	Olivia	378	pay
17	17	SAMSUNG G...	Munich	Luke	343	pay
18	18	SAMSUNG G...	Munich	Lucas	337	pay
19	19	SAMSUNG G...	New York	Luke	462	cancel order
20	20	APPLE iPhone...	Munich	Sophia	211	pay
21	21	APPLE iPhone...	Aachen	Jacob	356	pay

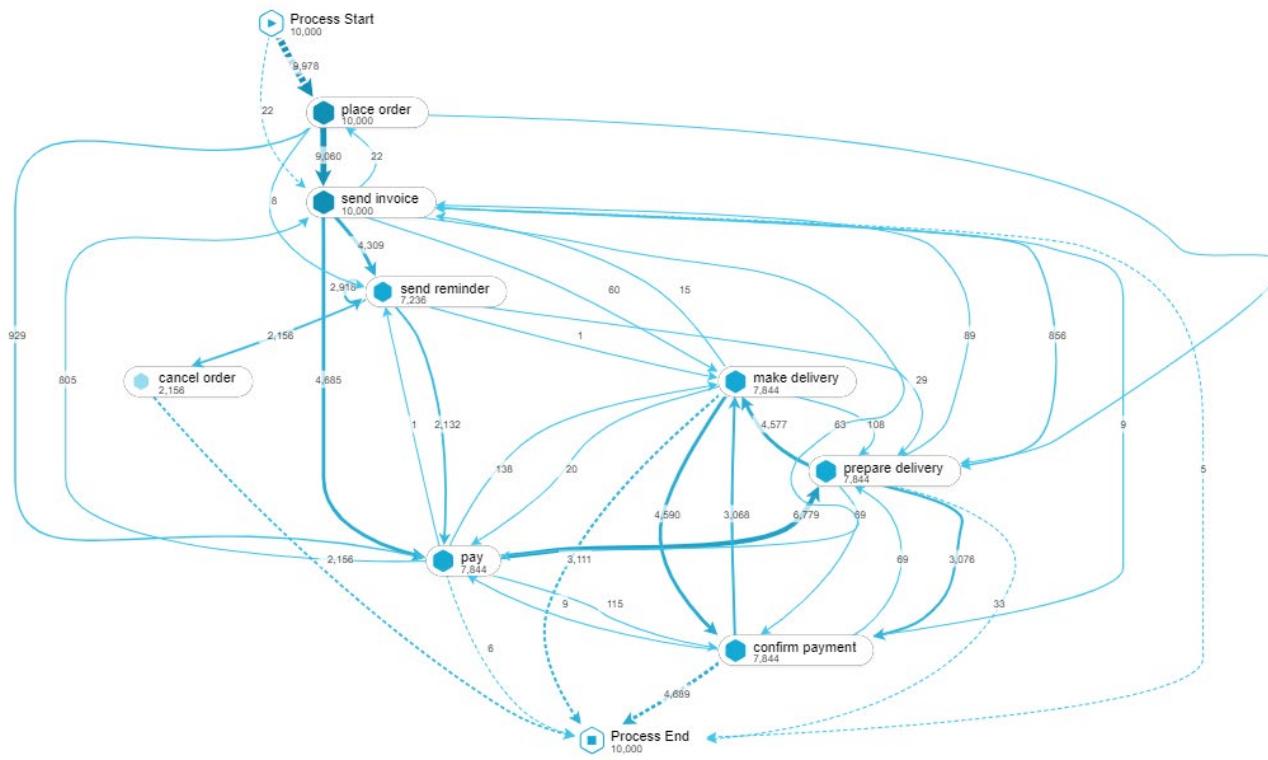
ExampleSet (10,000 examples, 0 special attributes, 6 regular attributes)



Analyzing the situation table in RapidMiner



Illustrating PQL: Creating a DFG



SOURCE	TARGET	COUNT
confirm payment	make delivery	3068
confirm payment	pay	9
confirm payment	prepare delivery	65
confirm payment	send invoice	9
make delivery	confirm payment	4590
make delivery	pay	20
make delivery	prepare delivery	108
make delivery	send invoice	15
pay	confirm payment	115
pay	make delivery	138
pay	prepare delivery	6779
pay	send invoice	805
pay	send reminder	1
place order	pay	929
place order	prepare delivery	3
place order	send invoice	9060
place order	send reminder	8
prepare delivery	confirm payment	3076
prepare delivery	make delivery	4577
prepare delivery	pay	69
prepare delivery	send invoice	89
send invoice	confirm payment	63
send invoice	make delivery	60
send invoice	pay	4685
send invoice	place order	22
send invoice	prepare delivery	856
send invoice	send reminder	4309
send reminder	cancel order	2156
send reminder	make delivery	1
send reminder	pay	2132
send reminder	prepare delivery	29
send reminder	send reminder	2918

Illustrating PQL: Creating a DFG

The screenshot shows the Celonis Studio interface with a 'DFG' tab selected. A component options dialog is open, titled 'Component options'. It contains several sections: 'General options' (Table title: 'events-based-situation table', Component type: 'OLAP Table'), 'DIMENSIONS' (SOURCE and TARGET dropdowns), 'KPIs' (COUNT dropdown), 'COUNT' (SOURCE dropdown), 'SORTING' (Add button), and 'ADVANCED OPTIONS' (checkboxes). On the left, there's a sidebar with various icons and a main area showing a table with 10.0k of 10.0k cases selected. The table has columns 'SOURCE', 'TARGET', and 'COUNT', listing various business activities like 'confirm payment', 'make delivery', etc., with their respective counts.

SOURCE("events"."ACTIVITY")

TARGET("events"."ACTIVITY")

COUNT(SOURCE("events"."ACTIVITY"))

SOURCE	TARGET	COUNT
confirm payment	make delivery	3068
confirm payment	pay	9
confirm payment	prepare delivery	69
confirm payment	send invoice	9
make delivery	confirm payment	4590
make delivery	pay	20
make delivery	prepare delivery	108
make delivery	send invoice	15
pay	confirm payment	115
pay	make delivery	138
pay	prepare delivery	6779
pay	send invoice	805
pay	send reminder	1
place order	pay	929
place order	prepare delivery	3
place order	send invoice	9060
place order	send reminder	8
prepare delivery	confirm payment	3076
prepare delivery	make delivery	4577
prepare delivery	pay	69
prepare delivery	send invoice	89
send invoice	confirm payment	63
send invoice	make delivery	60
send invoice	pay	4685
send invoice	place order	22
send invoice	prepare delivery	856
send invoice	send reminder	4309
send reminder	cancel order	2156
send reminder	make delivery	1
send reminder	pay	2132
send reminder	prepare delivery	29



Process Query Language (PQL)



Just the start: Will be continued (>200 operators) no need to memorize them but important for assignment and understanding of the link with other types of analysis.

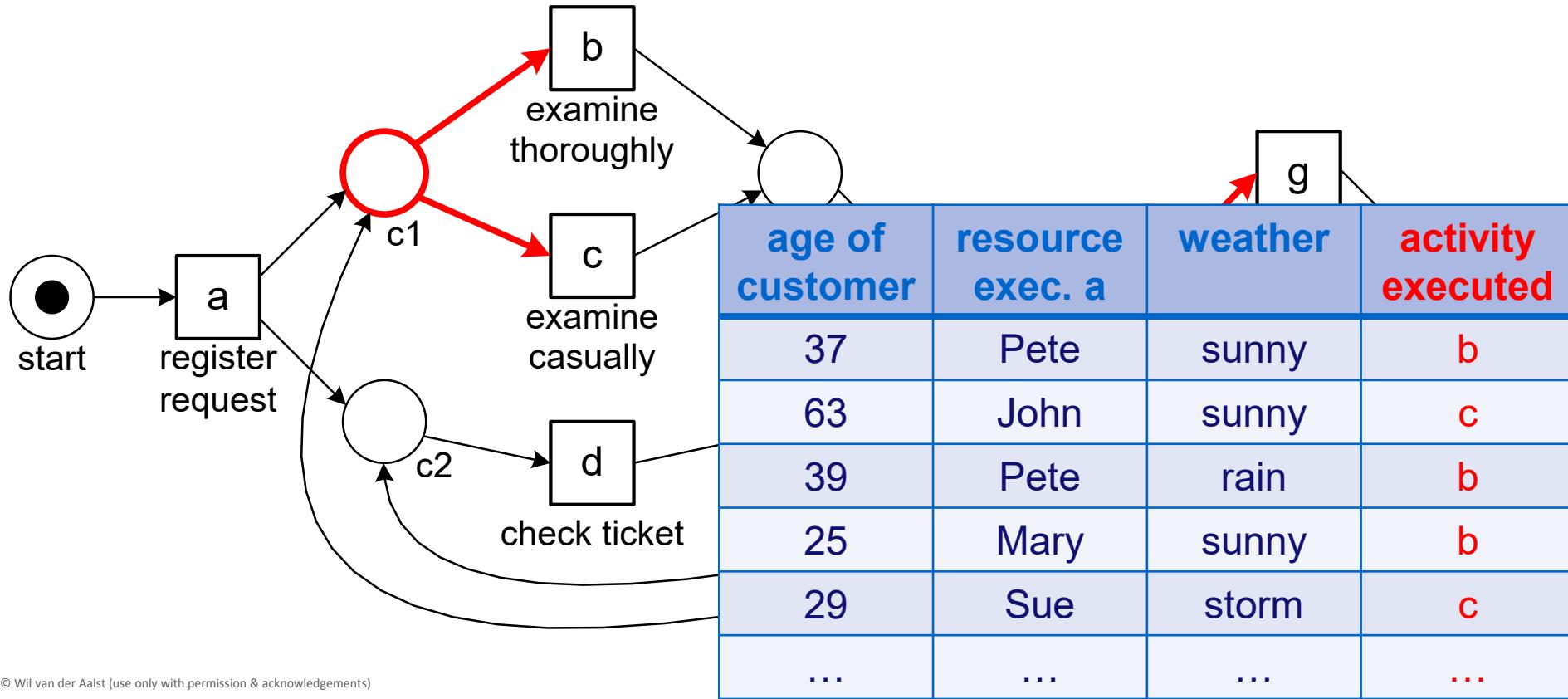
Summary



What should you be able to do?

- Given an event log with additional data attributes:
 - Learn a process model using some process discovery technique (model could also be given).
 - Identify decision points.
 - Create a case-based or event-based situation table per decision point.
 - Learn a decision tree for each decision point.
 - Add guards to the model.
- This may be a bit tedious. However, it is needed to understand the interaction between process mining and data mining / machine learning.

Recall: Decision mining is just an example of an ML/DM problem



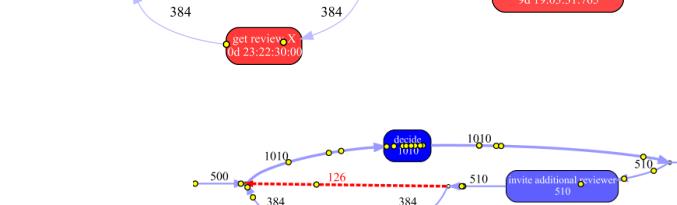
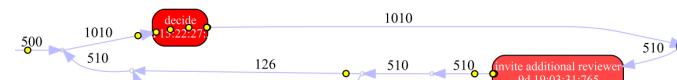
Situation table to analyze any problem

attrib. 1	attrib. 2	attrib. 3	...	attrib. n	outcome
37	Pete	sunny		2.12	OK
63	John	sunny		3.11	NOK
39	Pete	rain		2.66	NOK
25	Mary	sunny		5.22	OK
29	Sue	storm		3.34	NOK
...

- Each row corresponds to a situation.
- Situations may refer to events or cases.

Situation table to analyze any problem

- **Decision (as before)**
- **Outcome (e.g. case attribute or final activity)**
- **Bottlenecks**
 - Waiting time before activity
 - Service time activity
 - Time between two milestones
- **Compliance problems**
 - Missing/remaining tokens versus consumed/produced
 - Move on model/log versus synchronous moves



Part I: Introduction

Chapter 1
Data Science in Action

Chapter 2
Process Mining:
The Missing Link

Part II: Preliminaries

Chapter 3
Process Modeling
and Analysis

Chapter 4
Data Mining

Part III: From Event Logs to Process Models

Chapter 5
Getting the Data

Chapter 6
Process Discovery:
An Introduction

Chapter 7
Advanced Process
Discovery Techniques

Part IV: Beyond Process Discovery

Chapter 8
Conformance
Checking

Chapter 9
Mining Additional
Perspectives

Chapter 10
Operational Support

Part V: Putting Process Mining to Work

Chapter 11
Process Mining
Software

Chapter 12
Process Mining in the
Large

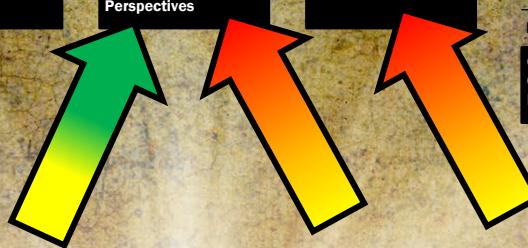
Chapter 13
Analyzing “Lasagna
Processes”

Chapter 14
Analyzing “Spaghetti
Processes”

Part VI: Reflection

Chapter 15
Cartography and
Navigation

Chapter 16
Epilogue



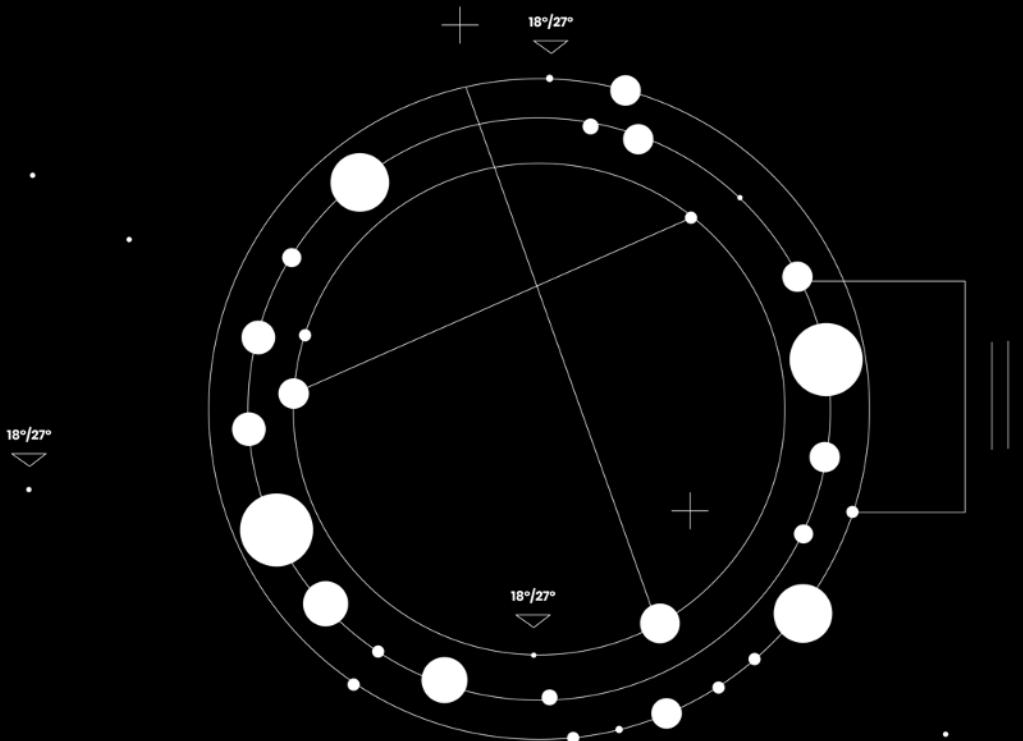
ID	Topic	Date	Date	Place
	Lecture 1 Introduction to Process Mining	08.04.24	Monday	AH V
	Lecture 2 Data Science: Supervised Learning	09.04.24	Tuesday	AH V
	<i>Exercise 1 Tool Introduction</i>	09.04.24	Tuesday	AH III
	Lecture 3 Data Science: Unsupervised Learning and Evaluation	15.04.24	Monday	AH V
	Lecture 4 Introduction to Process Discovery	16.04.24	Tuesday	AH V
	<i>Exercise 2 Data Mining</i>	16.04.24	Tuesday	AH III
	Lecture 5 Alpha Algorithm 1	22.04.24	Monday	AH V
	Lecture 6 Alpha Algorithm 2	23.04.24	Tuesday	AH V
	<i>Exercise 3 Petri Nets</i>	23.04.24	Tuesday	AH III
	Lecture 7 Model Quality Representation	29.04.24	Monday	AH V
	Lecture 8 Heuristic Mining	30.04.24	Tuesday	AH V
	<i>Exercise 4 Alpha Miner</i>	30.04.24	Tuesday	AH III
	Lecture 9 Region-Based Mining	06.05.24	Monday	AH V
	<i>Exercise 5 Heuristic Mining and Region-Based Mining</i>	07.05.24	Tuesday	AH III
	Lecture 10 Inductive Mining	13.05.24	Monday	AH V
	Lecture 11 Event Data and Exploration	14.05.24	Tuesday	AH V
	<i>Exercise 6 Inductive Mining</i>	14.05.24	Tuesday	AH III
	Lecture 12 Conformance Checking 1	27.05.24	Monday	AH V
	Lecture 13 Conformance Checking 2	28.05.24	Tuesday	AH V
	<i>Q&A Session Assignment Part I</i>	28.05.24	Tuesday	AH III
	Deadline Assignment Part I	02.06.24	Sunday	
	<i>Exercise 7 Footprint and Token-Based Replay (Exercise)</i>	03.06.24	Monday	AH V
	<i>Exercise 8 Alignments (Exercise)</i>	04.06.24	Tuesday	AH V
	Lecture 14 Decision Mining	10.06.24	Monday	AH V
	<i>Lecture 15 Celonis Guest Lecture</i>	11.06.24	Tuesday	AH V
	<i>Exercise 9 Decision Mining</i>	11.06.24	Tuesday	AH III
	Lecture 16 Performance Analysis and Organizational Mining	17.06.24	Monday	AH V
	<i>Exercise 10 Performance Analysis (Exercise)</i>	18.06.24	Tuesday	AH V
	<i>Exercise 11 Organizational Mining</i>	18.06.24	Tuesday	AH III
	<i>Exercise 12 Celonis Case Study</i>	24.06.24	Monday	AH V
	Lecture 17 Operational Support and Process Mining Applications	01.07.24	Monday	AH V
	Lecture 18 Distributed, Streaming, and Comparative Process Mining	02.07.24	Tuesday	AH V
	<i>Exercise 13 Operational Process Mining</i>	02.07.24	Tuesday	AH III
	Lecture 19 Closing	08.07.24	Monday	AH V
	<i>Q&A Session Assignment Part II</i>	09.07.24	Tuesday	AH III
	Deadline Assignment Part II	14.07.24	Sunday	
	<i>Q&A Session Exam</i>	16.07.24	Tuesday	AH III



Celonis Guest Lecture

Process Mining in Application

Angela-Sophia Gebert
Global Head of Academic Alliance



Today's agenda

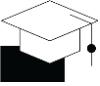
01 Celonis DNA

02 Use Case Study: Procurement

03 Getting the Data

04 A new Kind of Data – OCPM

The academic DNA of Celonis



1990s

-Process Mining gets discovered under the name of Workflow Mining



2011

-Alex, Basti und Martin discover Process Mining as part of a student project



2019-2024

-Celonis wins the German President's Award for Future Proof Innovation
-current market evaluation: 13bn USD



Today's agenda

01 Celonis DNA

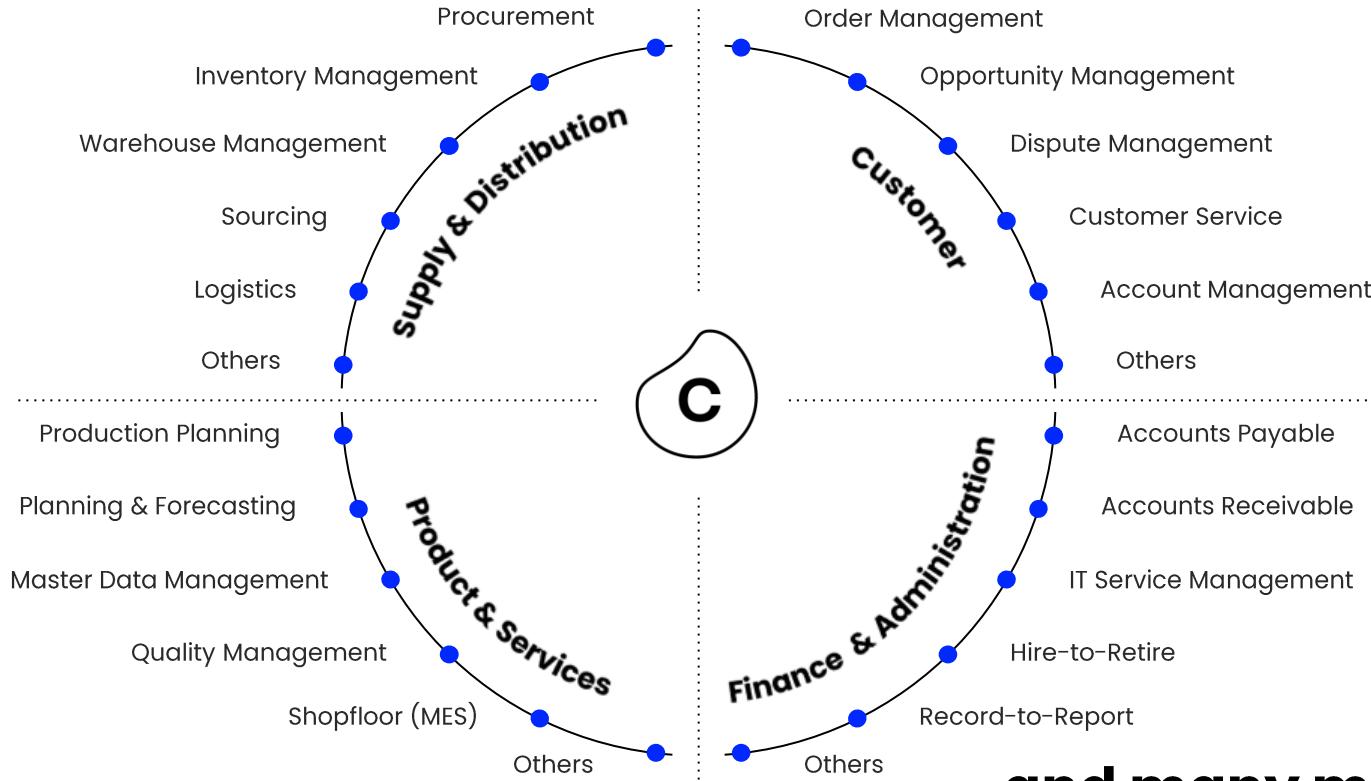
02 Use Case Study: Procurement

03 Getting the Data

04 A new Kind of Data – OCPM

Looking at typical processes.

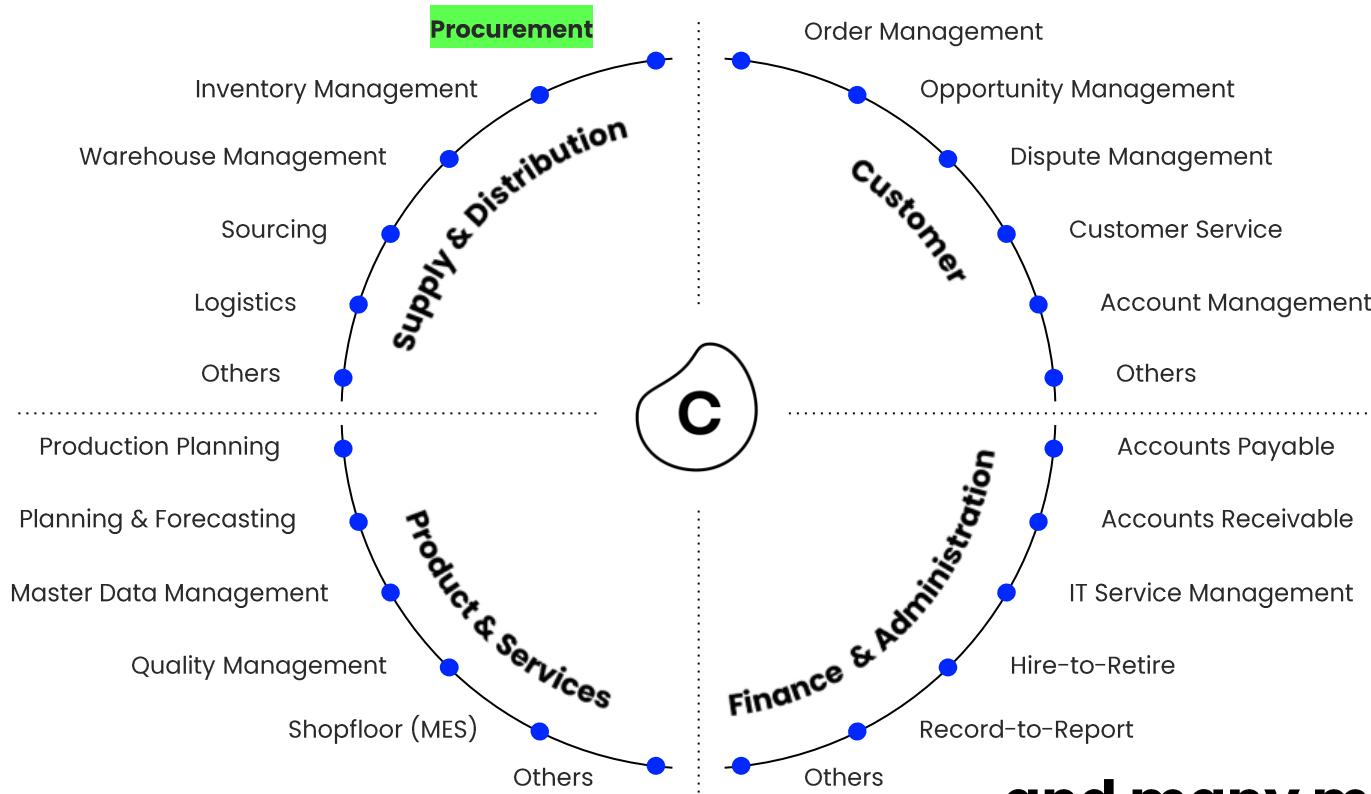
C



and many more...

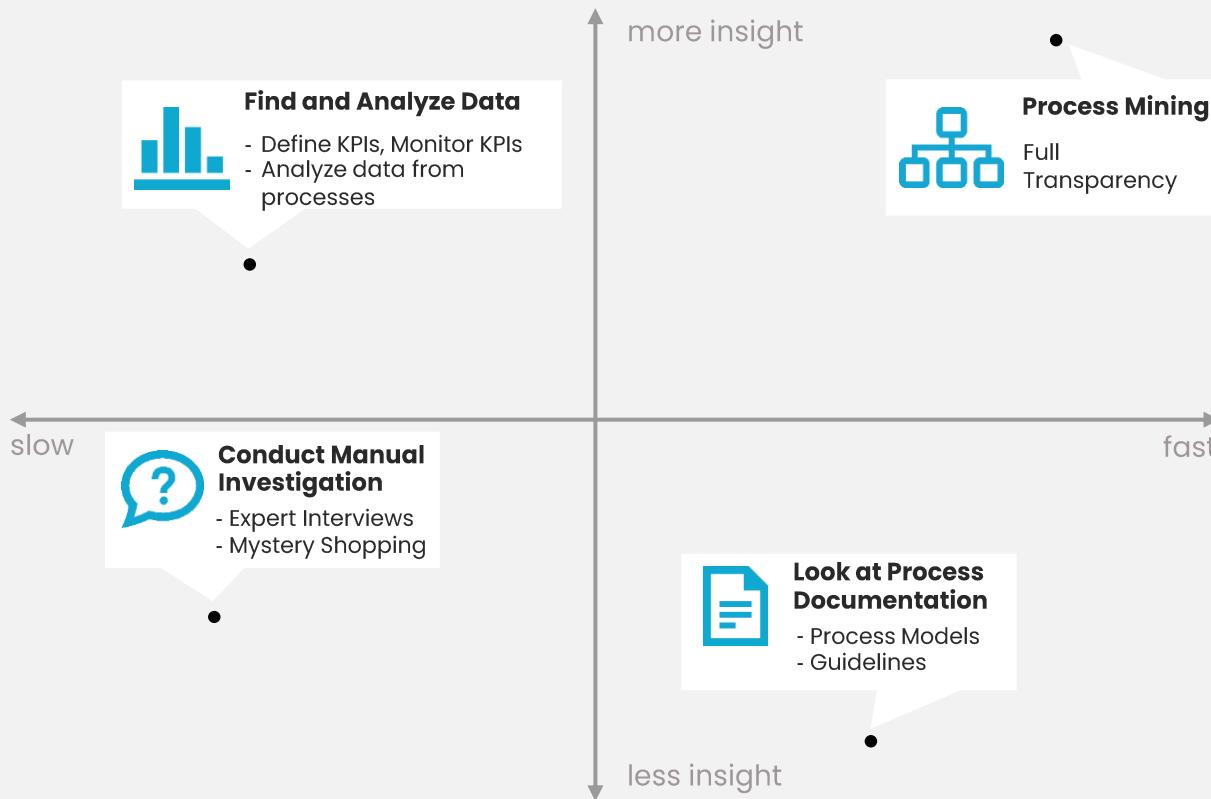
Looking at typical processes.

C



and many more...

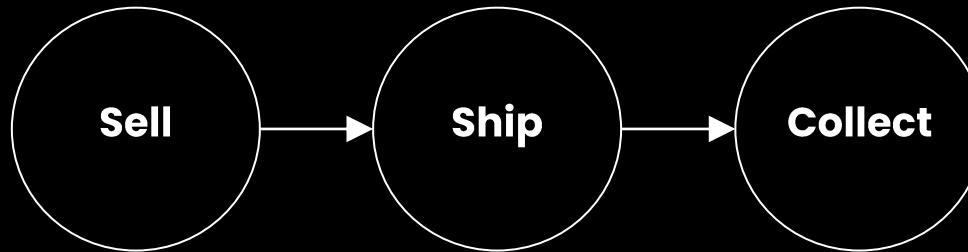
“How do I get Insights into Processes?”



**More transparency in business processes:
Switch on the lights!**



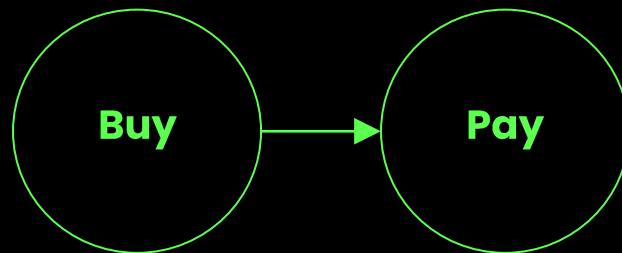
Every company sells, collects, buys, and pays



Opportunity
Management

Order Management
Inventory Management

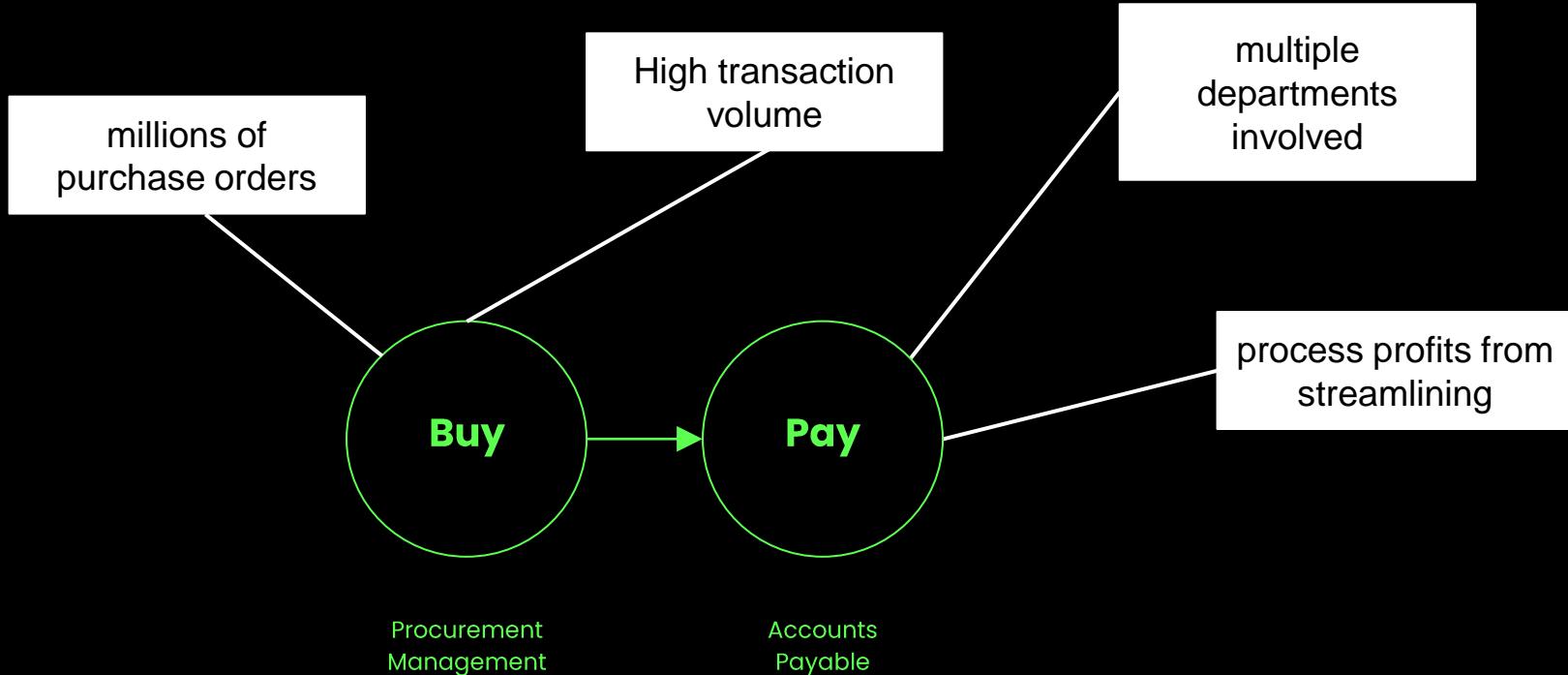
Accounts
Receivable



Procurement
Management

Accounts
Payable

Why Procurement can be a challenge.



The business impact of a good P2P process is huge.

Good Procurement Process

1

Streamlined,
reducing manual touchpoints

2

cost savings through
bundled orders

3

discounts with existing
vendors & supplier reliability

4

invoices paid at the ideal
point in time

Bad Procurement Process

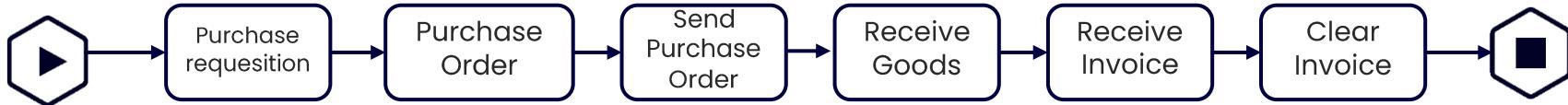
manual rework rates, price
changes, quantity
changes

maverick buying

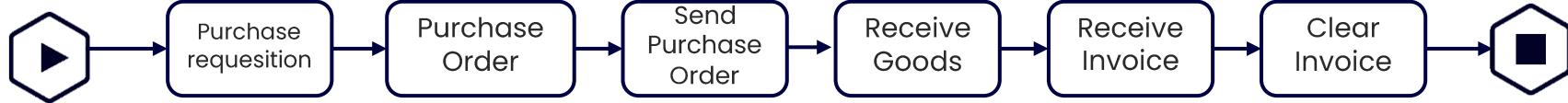
no vendor tracking, late
deliveries

lost working capital due
invoices being paid early or
early deliveries

Buying from your suppliers should be straightforward.



The Procurement process runs through one or several ERP systems in the background.



Process runs through IT landscape on backend:



ORACLE

sage

Microsoft

infor

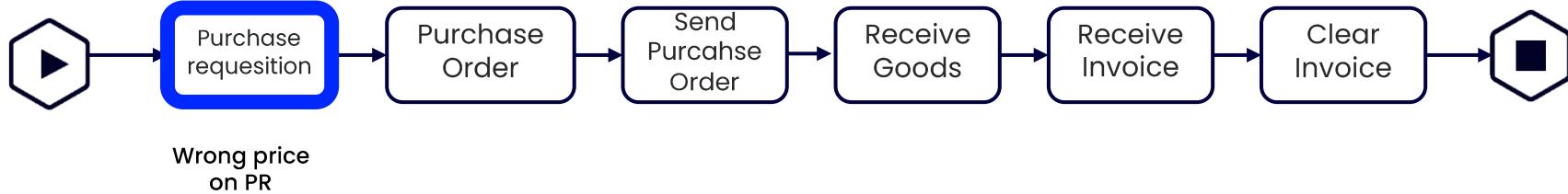
ABAS

IFS

PSI

STEP AHEAD

Complexity stops most companies from executing at their full potential.



Process runs through IT landscape on backend:



ORACLE

sage

Microsoft

infor

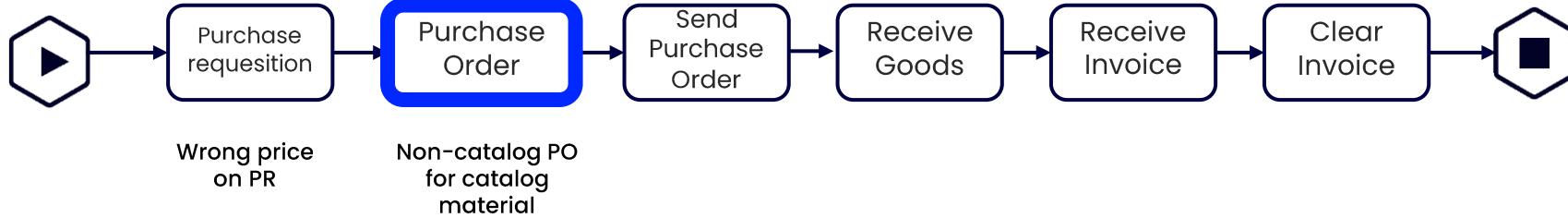
ABAS

IFS

PSI

STEP/AHEAD

Complexity stops most companies from executing at their full potential.



Process runs through IT landscape on backend:



ORACLE

sage

Microsoft

infor

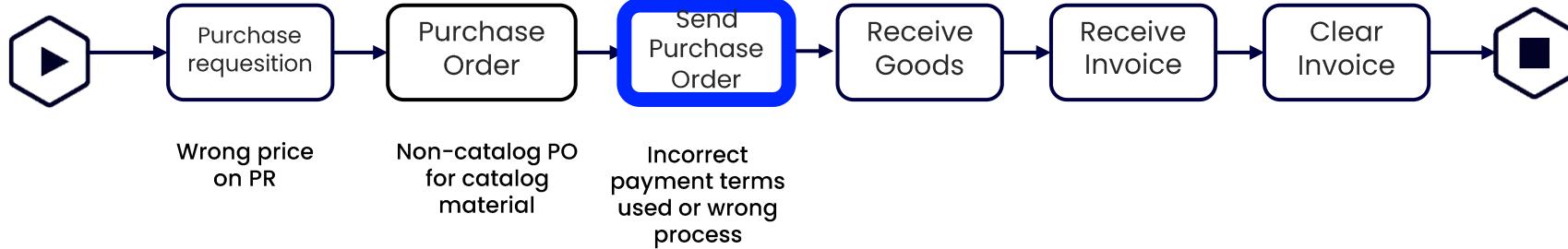
ABAS

IFS

PSI

STEP AHEAD

Complexity stops most companies from executing at their full potential.



Process runs through IT landscape on backend:

ORACLE®

SAP

sage

Microsoft

infor

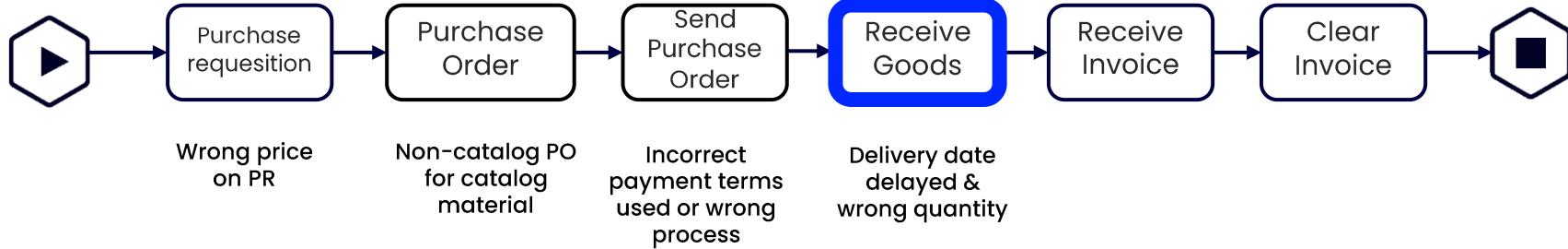
ABAS

IFS

PSI

STEP/AHEAD

Complexity stops most companies from executing at their full potential.



Process runs through IT landscape on backend:

ORACLE

SAP

sage

Microsoft

infor

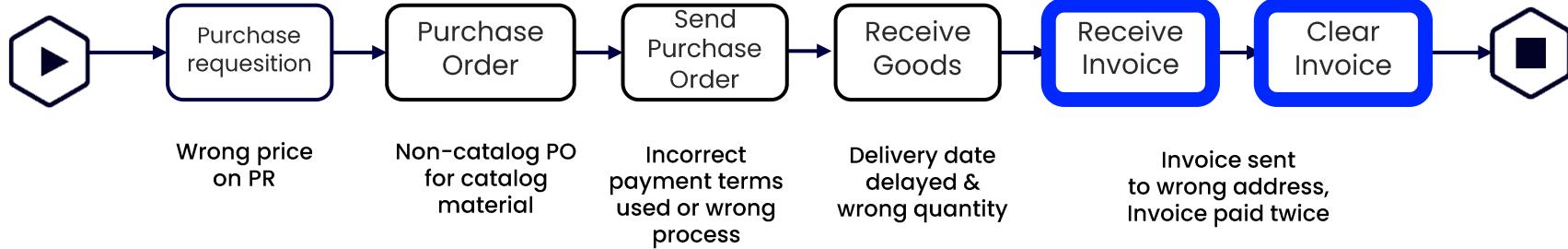
ABAS

IFS

PSI

STEP/AHEAD

Complexity stops most companies from executing at their full potential.



Process runs through IT landscape on backend:

ORACLE®

SAP

sage

Microsoft

infor

ABAS

IFS

PSI

STEP/AHEAD

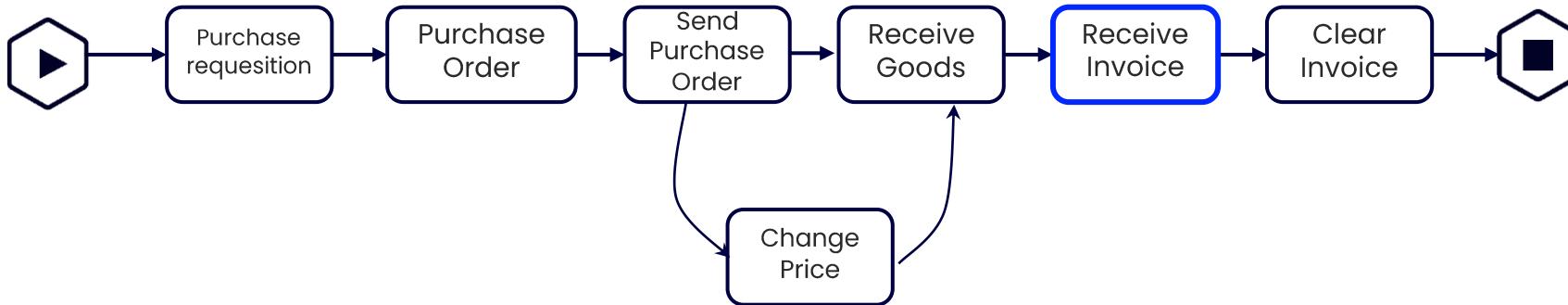
Zoom In: Price Changes – A common Problem

C

Price Changes explained:

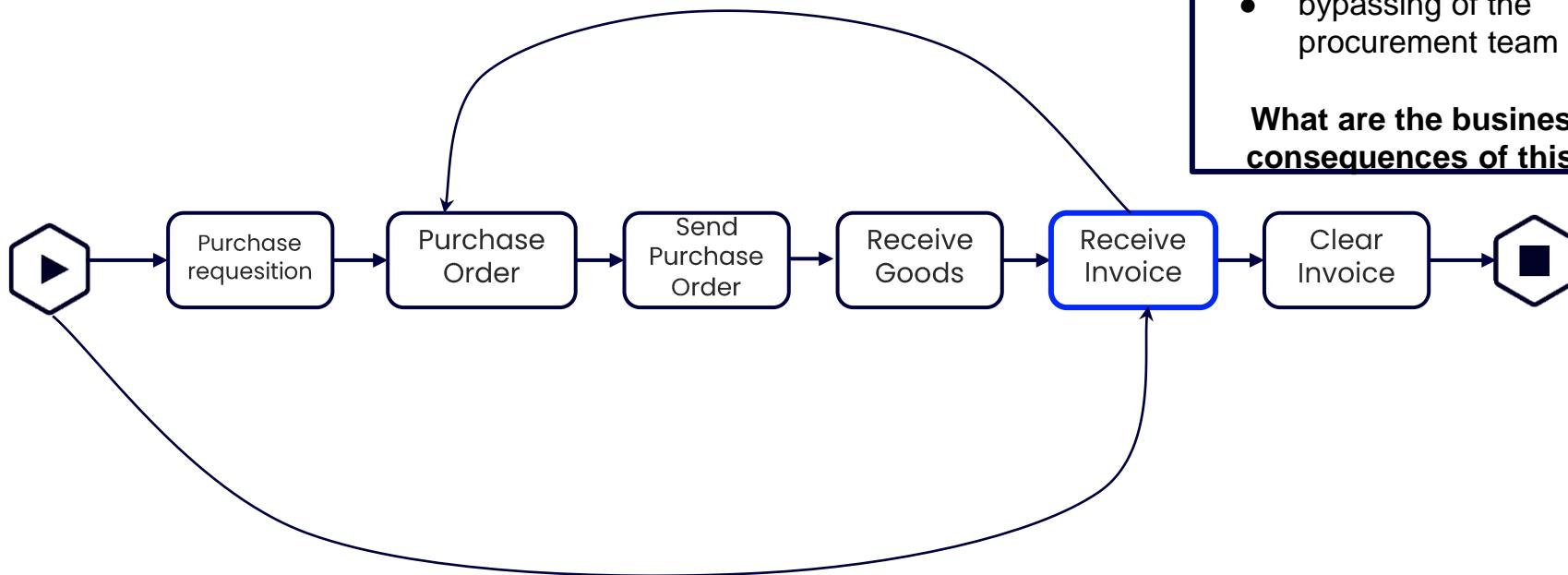
- prices get fixed when the PO is created
- manual change of prices after that point

What are the business consequences of this?



Zoom In: Maverick Buying – A common Problem

c

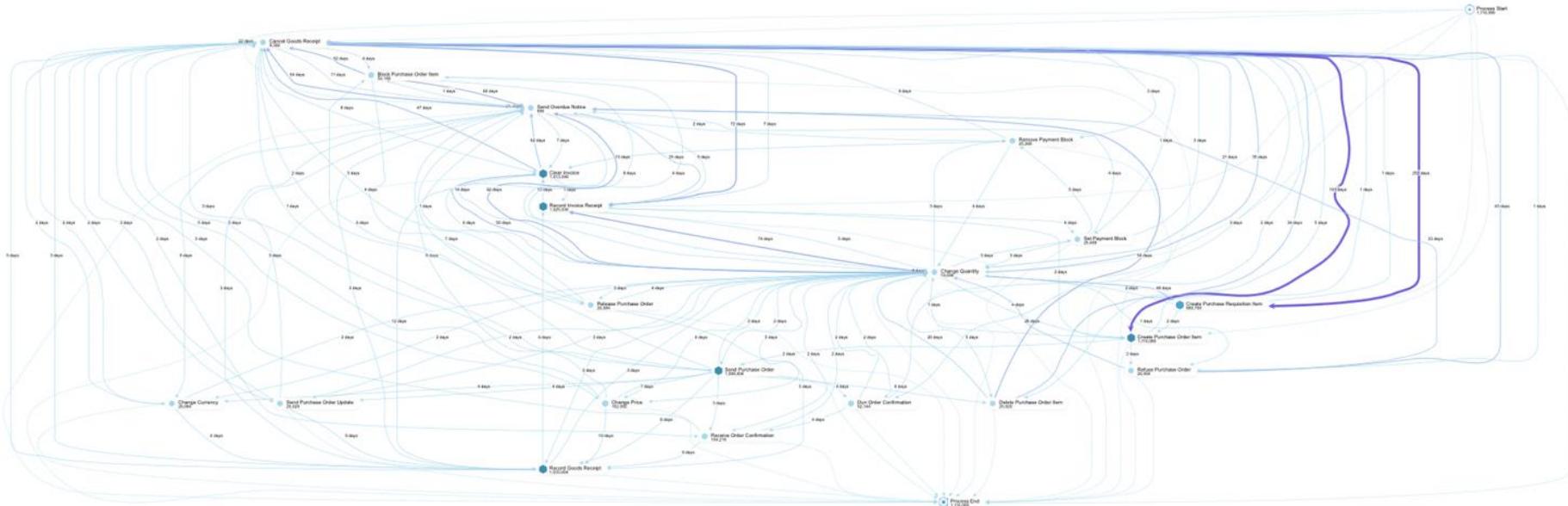


Maverick Buying explained:

- bypassing of the procurement team

What are the business consequences of this?

These execution gaps have serious consequences for the business.



Today's agenda

01 Celonis DNA

02 Use Case Study: Procurement

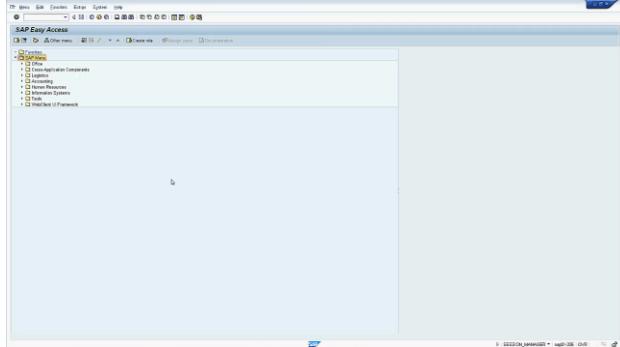
03 Getting the Data

04 A new Kind of Data – OCPM

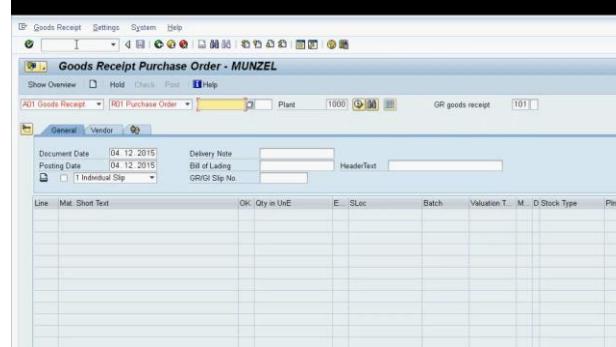
This is where the data comes from.



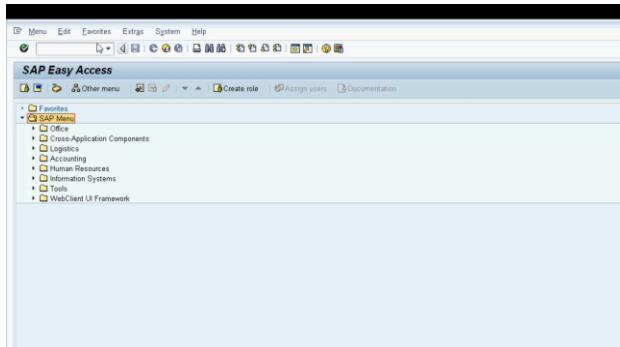
1. Create PO item



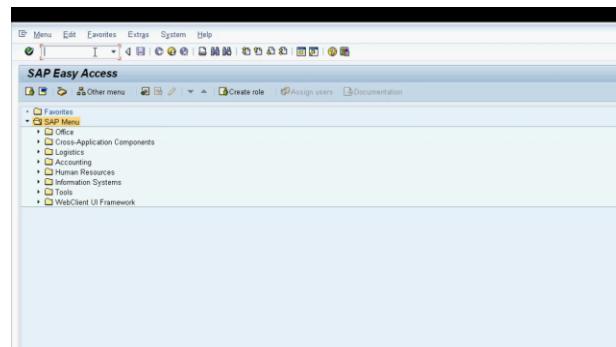
2. Receive Goods



3. Receive Invoice

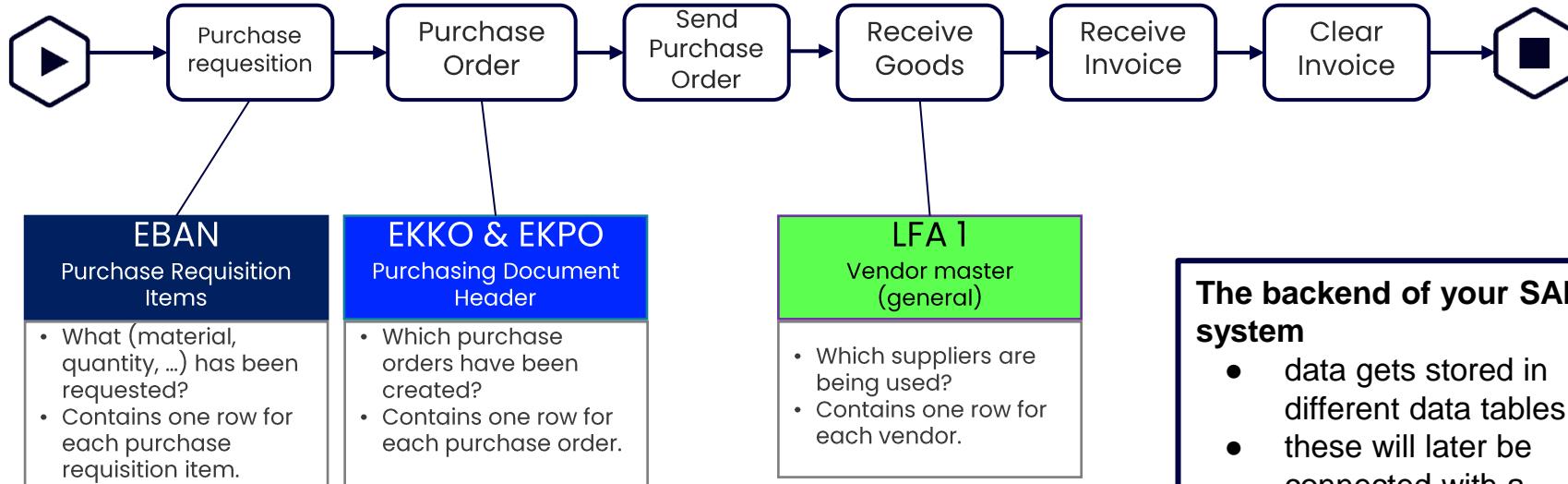


4. Pay Invoice



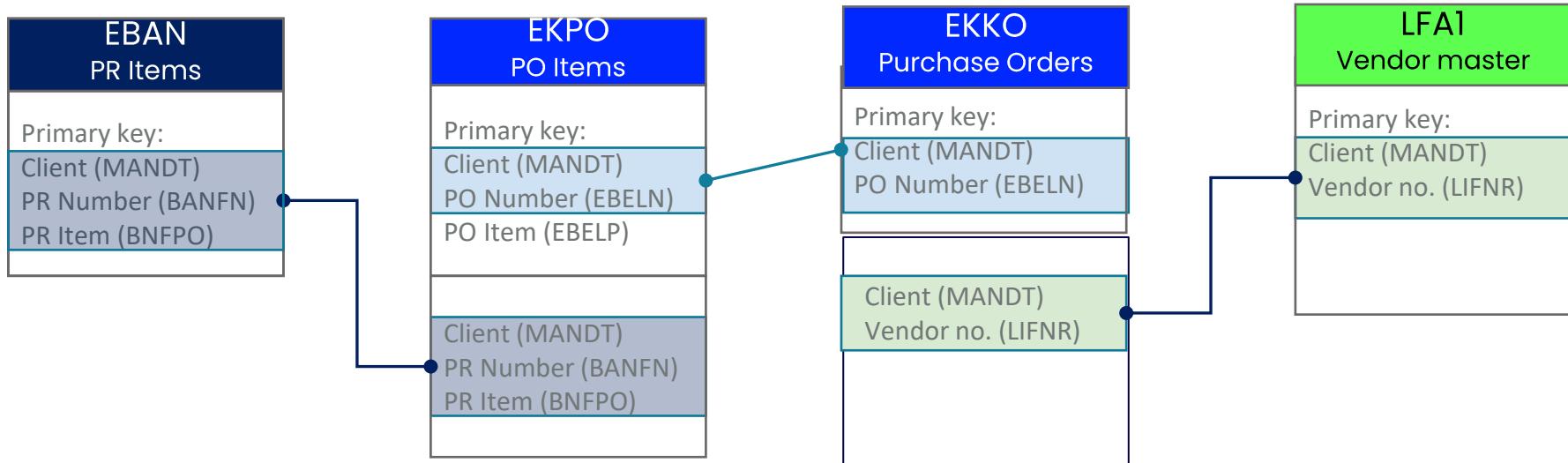
This is where the data comes from.

C



This is where the data comes from.

C



In reality this would look like this.

C

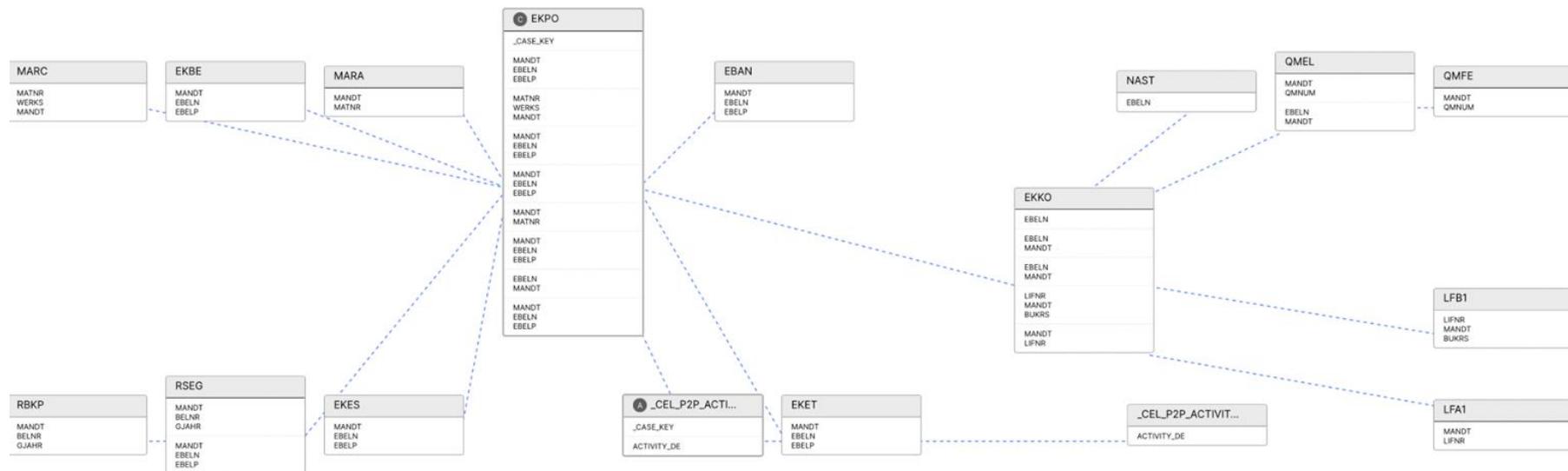
X

Data Model: SAP ECC - P2P Data Model - EN [in use]

Data Model

Data Model Loads

Switch Data Model



C

X

The extracted data would look like this.

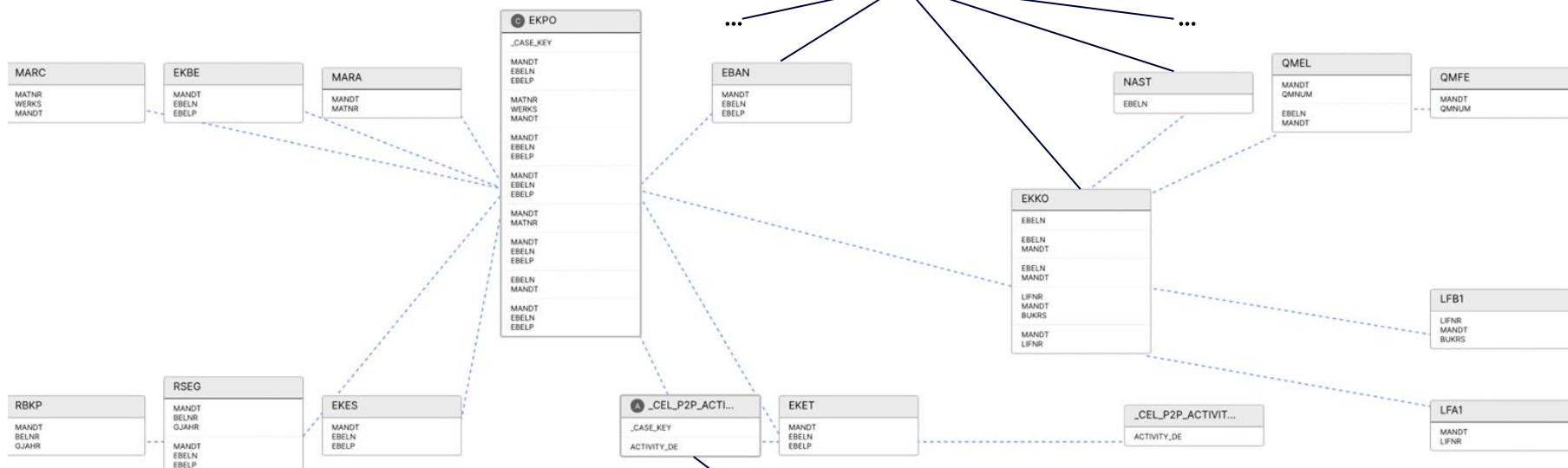
Data Model: SAP ECC - P2P Data Model - EN [in use]

Data Model

Data Model Loads

Switch Data Model

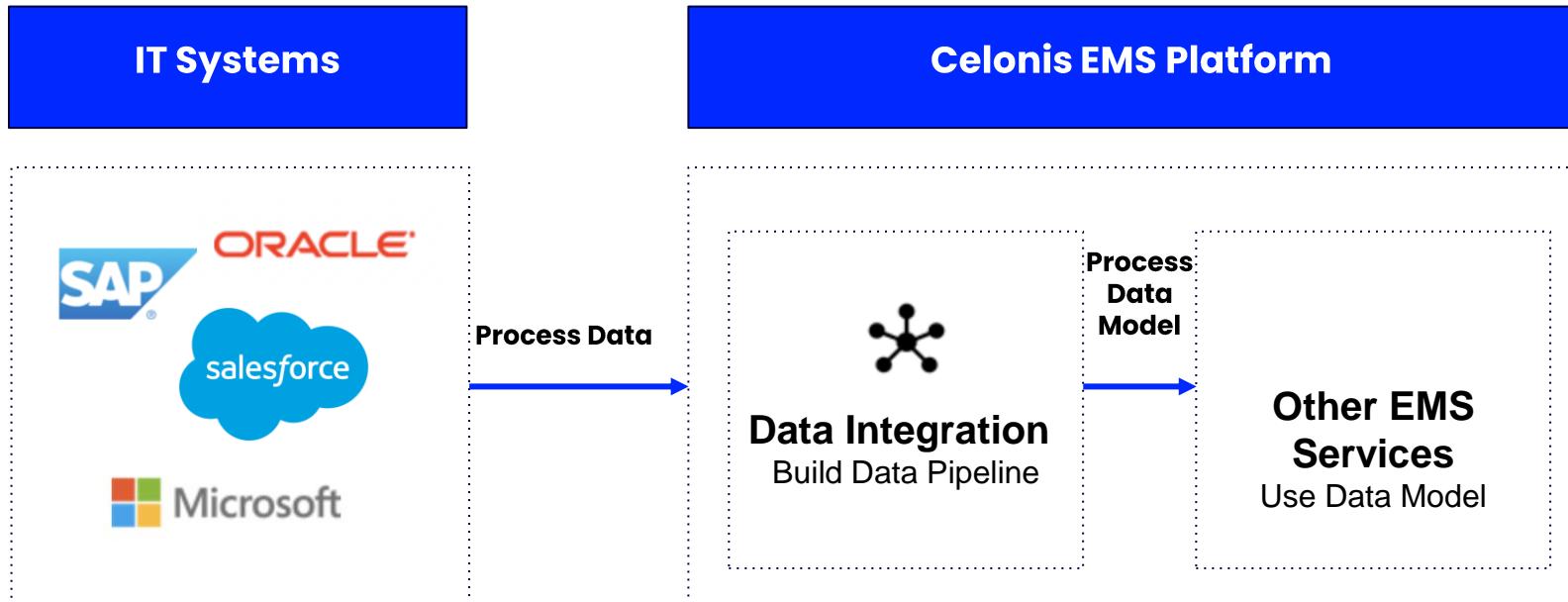
Additional data added through joins



Extracted events

This is how the data gets loaded into the EMS.

C



Real-time Data vs Scheduled Loads

C

Source Systems



ORACLE®



Real-time

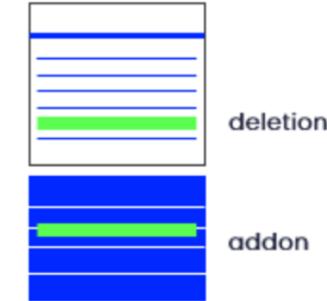
change tracking

VS

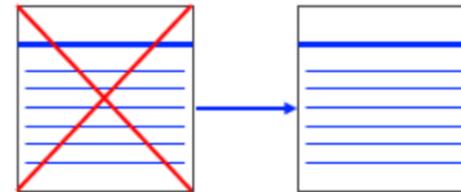
scheduled loads

EMS Data Integration

incremental changes
replicated in tables



Tables
deleted and
recreated

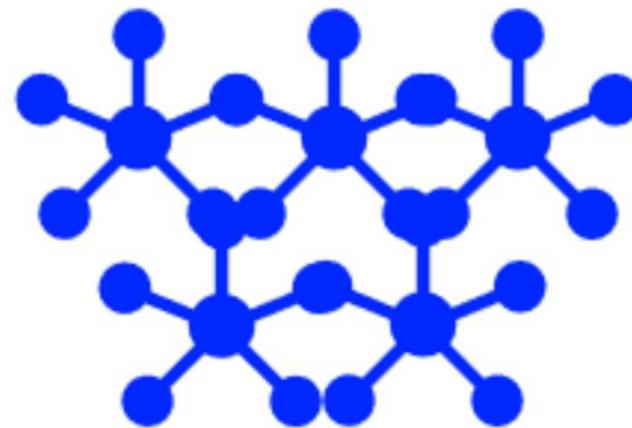


Operational vs Analytical Data Model

Operational Data Model



Analytical Data Model



Restricted scope with faster load,
only most recent cases, for day-to-day business

Full scope but less frequent & slower loads, for analysis

Today's agenda

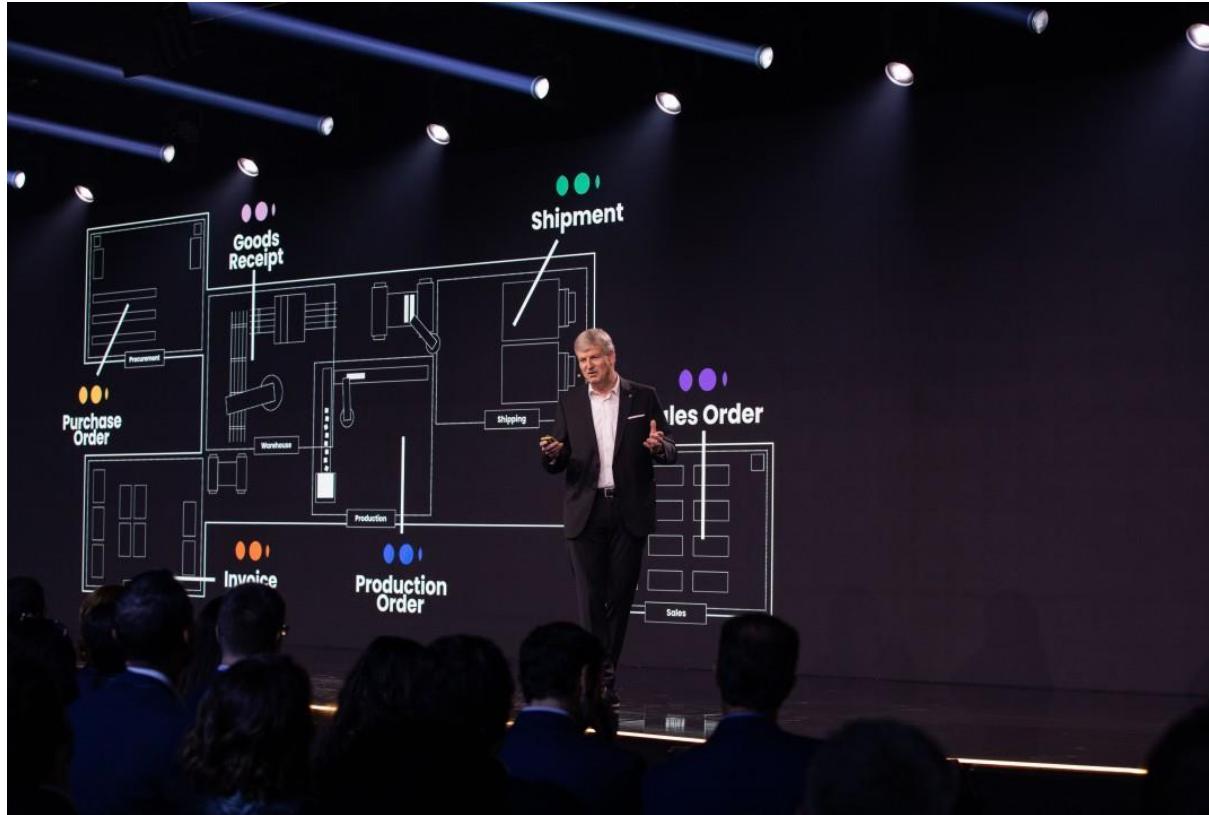
01 Celonis DNA

02 Use Case Study: Procurement

03 Getting the Data

04 A new Kind of Data – OCPM

What's new?





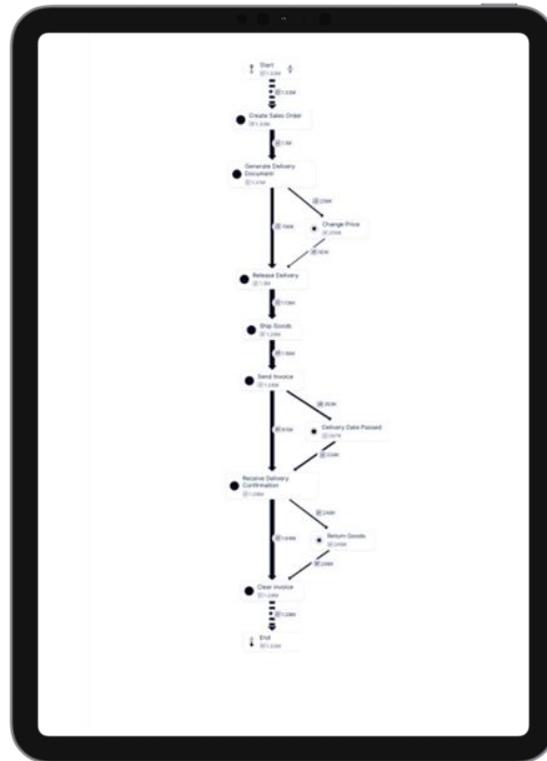
Limitations of Case Centric Process Mining

Restricted perspective

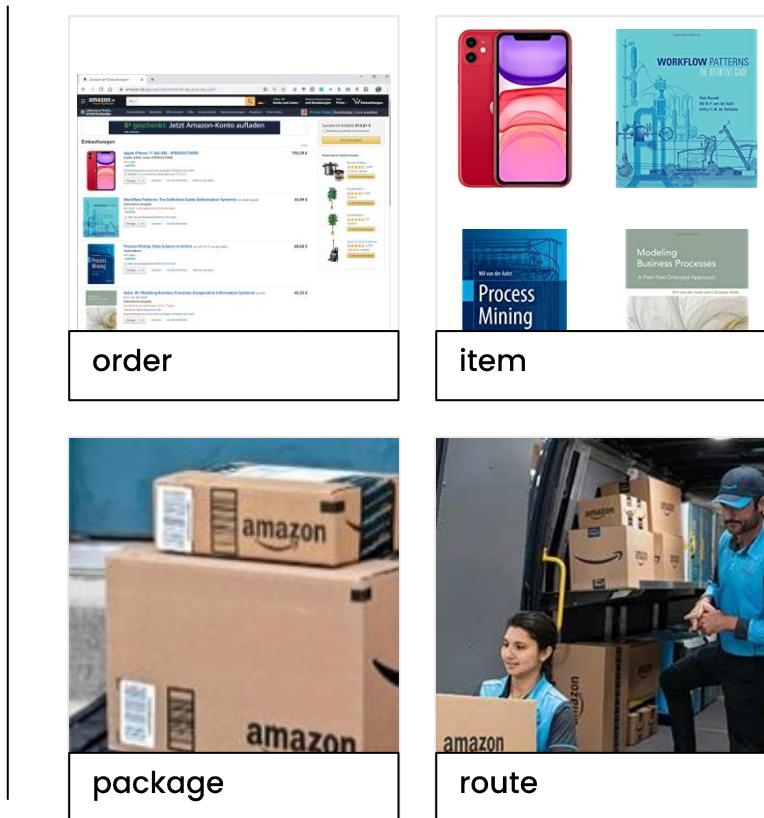
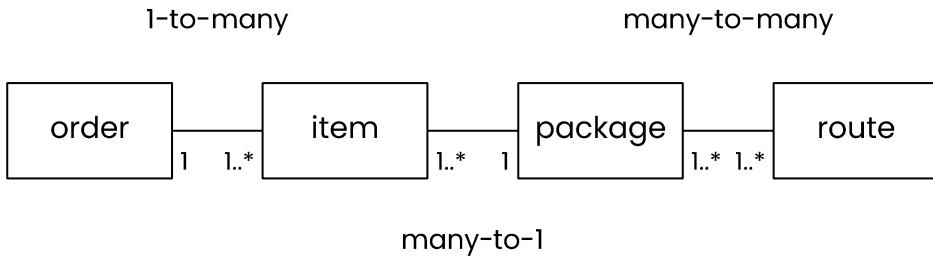
Incomplete information

Effort to extract and transform data

3D reality is squeezed to a 2D model



Example: Limitation of Restricted Perspective



Limitation of 2D Event Logs & Models

Deficiency

Events in the original event log that have no corresponding events in the flattened event log may **unintentionally disappear** from the data set.

Convergence

Events referring to multiple objects of the selected type are replicated, possibly leading to **unintentional duplication**.

Divergence

Two events referring to two **different objects** of a type not selected as the case notion many be considered to be **causally related** but are not.



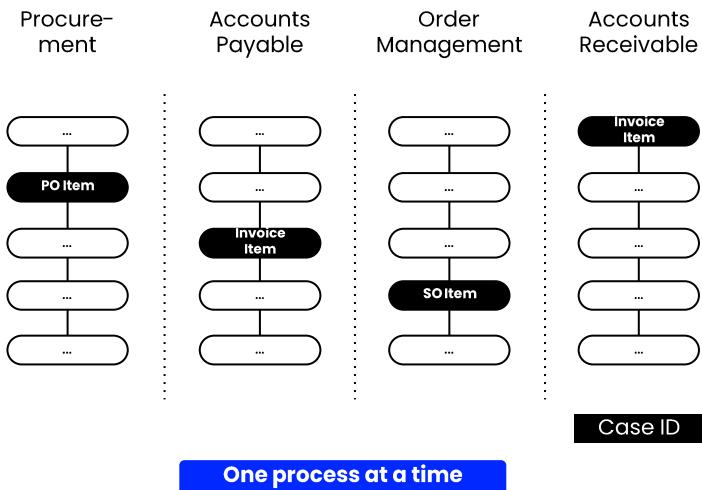
Object-centric process mining vs. traditional process mining



Case-centric (traditional) process mining

Traditional process mining lines up all the cases and events that make up your processes behind a predetermined case ID (eg. PO item, invoice item), forcing a narrow and limited view of a process.

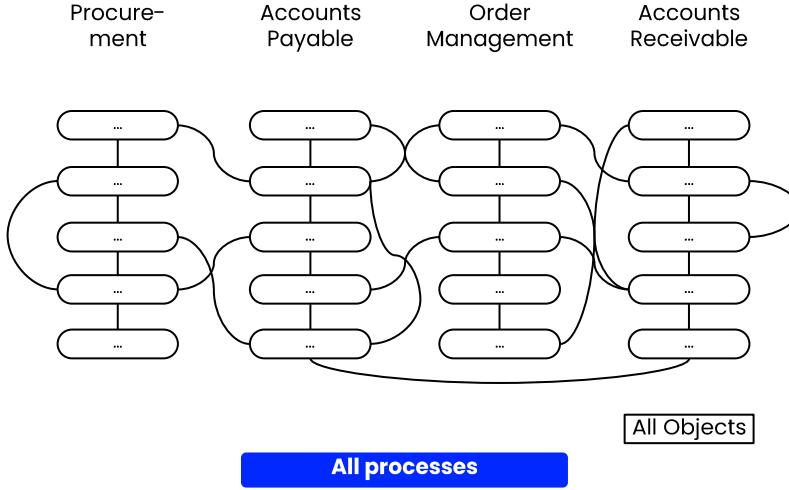
A new data model is needed for every case ID to gain more insights.



Object-centric process mining

Object-centric process mining captures object life cycles without the imposition of a case ID, enabling the accurate analysis of processes as they truly run.

One scalable data model is needed for full visualization of the complex process reality.

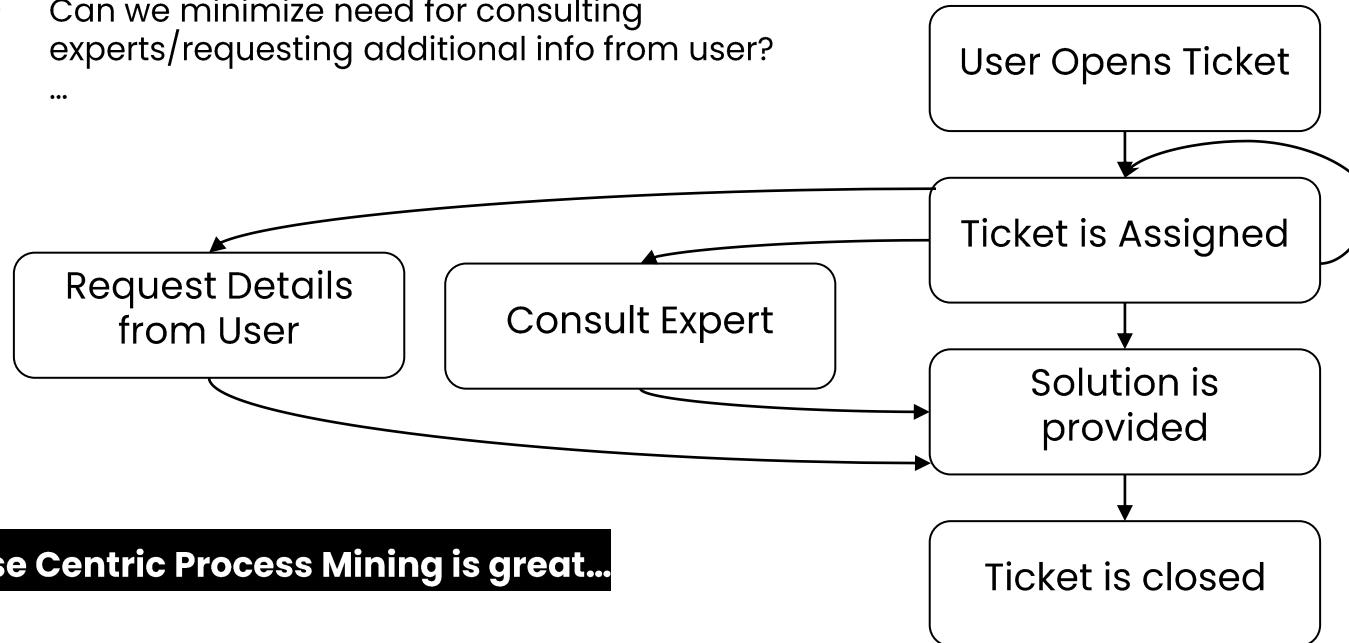


Case Centric Process Mining in Customer Support & Operations



Analyzing Flow of Ticket Process (Case = Ticket)

- How long does it take to close a ticket?
 - Could we optimize initial assignment?
 - Can we minimize need for consulting experts/requesting additional info from user?
- ...



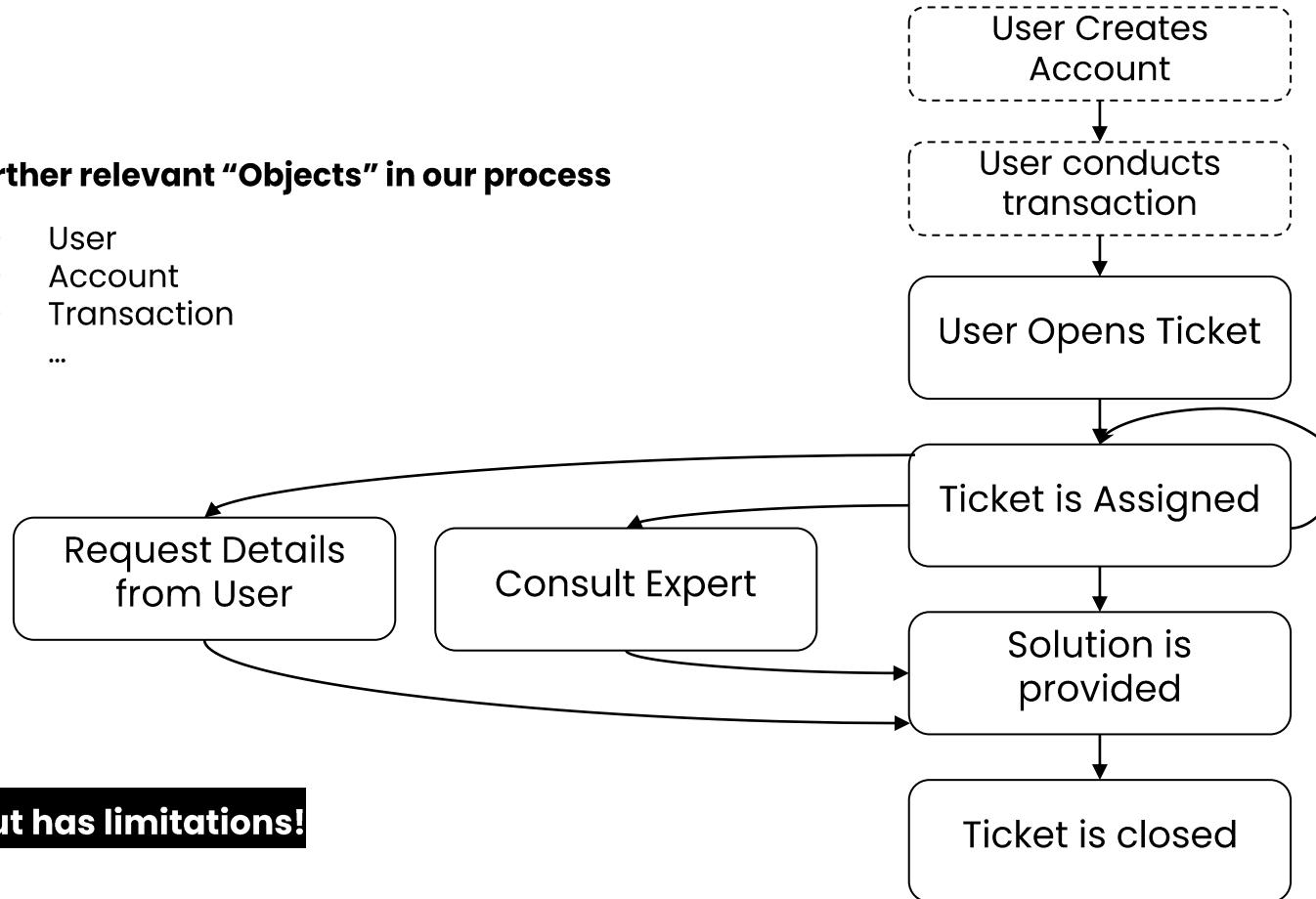
Case Centric Process Mining is great...



Extending the Analysis towards End-to-End Understanding

Further relevant “Objects” in our process

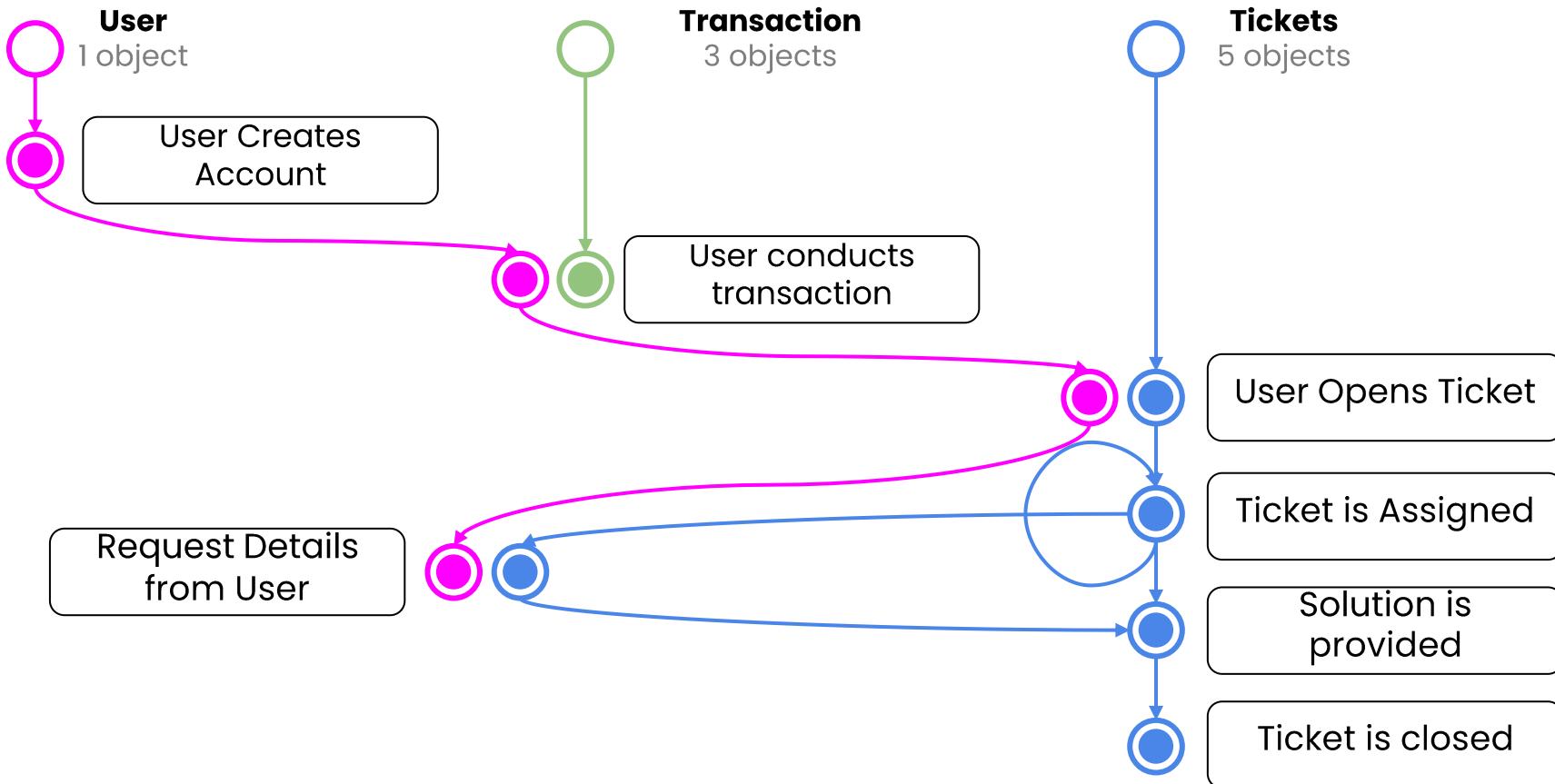
- User
- Account
- Transaction
- ...



...but has limitations!



How would this look in OCPM?



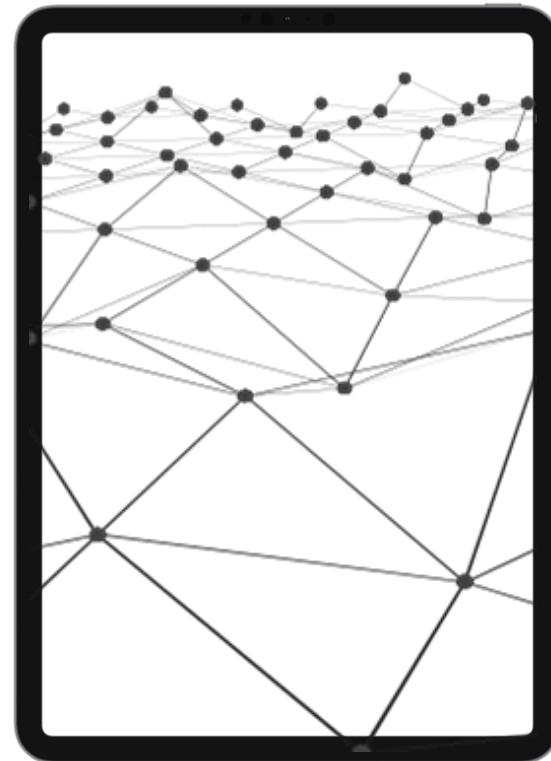
Benefits of Object Centric Process Mining



Data extraction is only done once

Interactions between objects are captured

3-dimensional event logs & process models



Agenda

01

Business Processes across the Value Chain

What is a Business Process?

What different business processes do we know?

02

Business Processes across the Value Chain

What is a Business Process?

What different business processes do we know?

03

Zoom In - Process Mining for Procurement

P2P Characteristics

Business Impact

04

Zoom In - Celonis in Action (Demo)

P2P from Discover to Enhancement

05

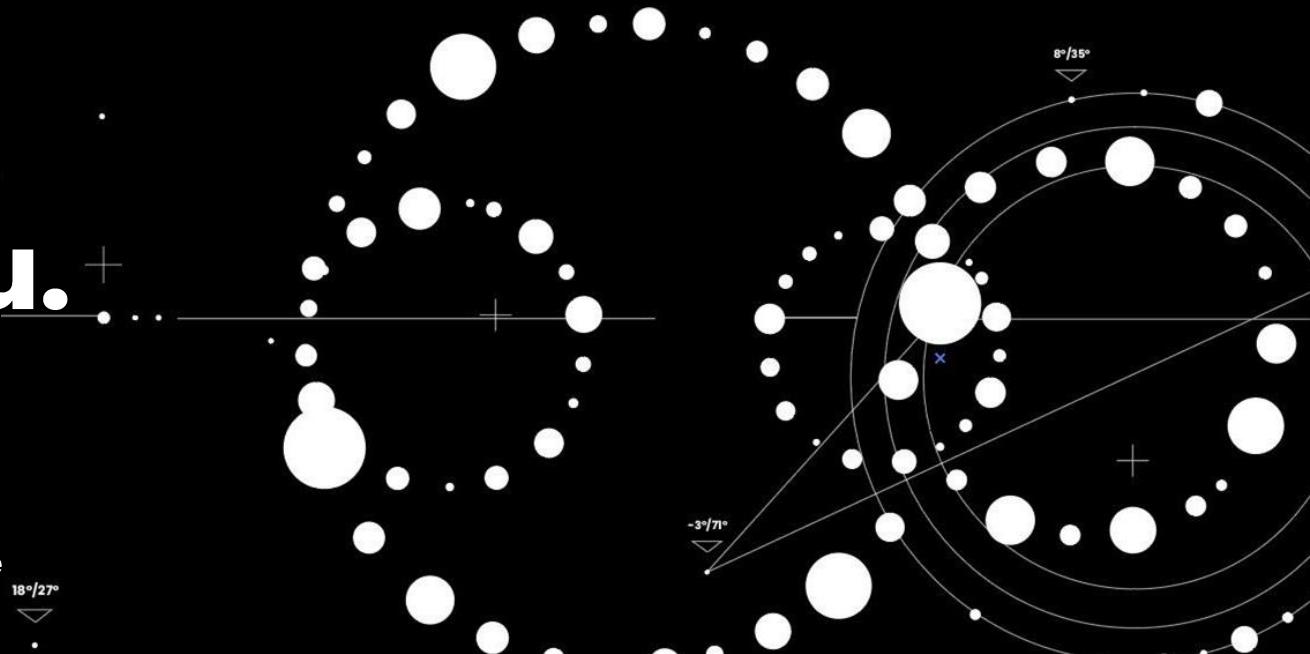
Process Enhancement – P2P

The move from Process Mining to EMS

Process Enhancement in Application

Thank you.

Angela-Sophia Gebert
Global Head of Academic Alliance



Performance Analysis and Organizational Mining

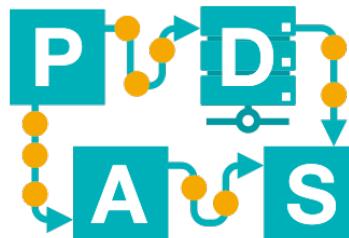
Lecture 16

prof.dr.ir. Wil van der Aalst

www.vdaalst.com @wvdaalst

www.pads.rwth-aachen.de

BPI-L16



Chair of Process
and Data Science

RWTHAACHEN
UNIVERSITY

After control-flow and data-flow ...

- Performance analysis.
- Social networks.
- Organizational mining.
- Creating situation tables in Celonis using PQL to analyze the above perspectives (a recurring theme).
- Integrating the different perspectives.

About the reminder

Assignment Part 2 - Published
von Lü, Lukas - Dienstag, 11. Juni 2024, 15:32

Dear students,

As you are aware (if not, please refer to the Study Guide), this course features a **mandatory assignment**, divided into two parts. It's essential to secure at least 50% of the total points for the entire assignment to be eligible for the exam. This means you need to aim for a minimum of 100 out of the 200 possible points in the assignments.

We have just published the second part of the assignment. It can be found in the corresponding section.

For the assignment, you are required to join one of the predefined groups. Each part of the assignment must be submitted by **a group of 2-3 students** (individual submissions are not permitted). The group forming for the second part of the assignment is open until **17.06.24** (refer to 'Groups for Assignment Part 2'). You have the freedom to join and switch between these groups during this period, however, please stick to the following:

1. You can use the corresponding forum (see [Sozializing / Group Finding](#)) to search for group members.
2. You are responsible for the group you are working with: Please check early, if your group members are reliable. You cannot contact us asking us to remove students from your group a few days before the assignment submission deadline.
3. **If you cannot find any group members, please enter an empty group before 17.06.24.** We will merge such single person groups into groups of 2-3 students shortly after the group forming deadline. If you are not in one group, you cannot participate.

Please do not randomly enter groups of other students you did not contact before. We will remove you from such groups if the other group members complain immediately after the group forming deadline. If you cannot find group members, enter an empty group.

Kind regards,

BPI Team

**Group forming until today!
Two instruction slots tomorrow!
Use the opportunities to learn PQL!**

Lecture 14 Decision Mining	10.06.24	Monday	AH V
<i>Lecture 15 Celonis Guest Lecture</i>	11.06.24	Tuesday	AH V
<i>Exercise 9 Decision Mining</i>	11.06.24	Tuesday	AH III
Lecture 16 Performance Analysis and Organizational Mining	17.06.24	Monday	AH V
<i>Exercise 10 Performance Analysis (Exercise)</i>	18.06.24	Tuesday	AH V
<i>Exercise 11 Organizational Mining</i>	18.06.24	Tuesday	AH III
<i>Exercise 12 Celonis Case Study</i>	24.06.24	Monday	AH V
Lecture 17 Operational Support and Process Mining Applications	01.07.24	Monday	AH V
Lecture 18 Distributed, Streaming, and Comparative Process Mining	02.07.24	Tuesday	AH V
<i>Exercise 13 Operational Process Mining</i>	02.07.24	Tuesday	AH III
Lecture 19 Closing	08.07.24	Monday	AH V
Q&A Session Assignment Part II	09.07.24	Tuesday	AH III
Deadline Assignment Part II	14.07.24	Sunday	
<i>Q&A Session Exam</i>	16.07.24	Tuesday	AH III

Survey Results: Thank You!

24S-Business Process Intelligence | 24S-12.25109 (2024)

Lernperson: Herr Univ.-Prof. Dr. Ir. Willibordus Martinus Pancratius van der Aalst

Lehrveranstaltungstyp: Lecture
Erfasste Fragebögen: 32



Globalwerte

Globalindikator

Die abgebildeten Globalwerteindikatoren (z.B. Konzept der Lehrveranstaltung, Vermittlung und Verhalten) beziehen sich auf den eingegebenen Antworten zu den Statistiken (Bspf zu ... nicht zu dem jeweiligen Fragebogen).

Zur Berechnung werden die einzelnen Auswertungsgegebnisse der Statistiken verwendet und nicht die bereit berechneten Mittelwerte der Befragten. Fragen, welche die erforderliche Mindestanzahl nicht erreichen, gehen nicht in die Berechnung ein.

Die Fragen „Ich bewerte das Konzept / ...“ und „Ich gebe die Leistungswertigkeit der Übungskomponente ...“ nehmen keinen Einfluss auf diese Indikatoren.

Konzept der Vorlesung / Lecture Concept

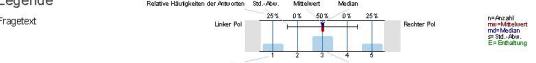
Konzept der Übung / Exercise Course Concept

Vermittlung und Verhalten - Vorlesung / Instruction and Behavior - Lecture

Vermittlung und Verhalten - Übung / Instruction and Behavior - Exercise Course

Auswertungsteil der geschlossenen Fragen

Legende



Konzept der Vorlesung / Lecture Concept



24S-Business Process Intelligence | 24S-12.25109 (2024)

Es werden Zusammenfassungen an sinnvollen Stellen gemacht. / Lecture material is summarized at appropriate intervals.

Der Schwierigkeitsgrad ist ... /

The degree of difficulty is ...

Ich bewerte das Konzept der Vorlesung mit ... /

I would evaluate the lecture concept as ...

Konzept der Übung / Exercise Course Concept

Die Lernziele der Übung sind definiert. / The learning goals of the exercise course are defined.

Die Übung hat eine klar erkennbare Struktur. / The exercise course is well structured.

Die zur Verfügung gestellten Materialien sind hilfreich. / The materials provided are helpful.

Vorlesung und Übung sind aufeinander abgestimmt. / The lecture and exercise course correspond to each other.

Die Übung hilft mir die Lehrinhalte des Moduls zu verstehen. / The exercise course contributes to a better understanding of module content.

Die Übungsaufgaben sind verständlich gestellt. / The exercise tasks posed in the exercise course are understandable.

Die vorgesetzten Übungsaufgaben werden innerhalb der Übungsdauer bearbeitet. / The planned tasks are worked out during the exercise course.

24S-Business Process Intelligence | 24S-12.25109 (2024)

Es werden Zusammenfassungen an sinnvollen Stellen gemacht. / Lecture material is summarized at appropriate intervals.

Der Schwierigkeitsgrad ist ... /

The degree of difficulty is ...

Ich bewerte das Konzept der Übung mit ... /

I would evaluate the exercise course concept as ...

24S-Business Process Intelligence | 24S-12.25109 (2024)

Falls Sie Ihre Lösung abgeben konnten: Wurde diese nachvollziehbar korrigiert? /
Was ist die Lösung von Ihnen korrigiert? Was ist die Lösung von Ihnen korrigiert?

Es wird keine Auswertung angezeigt, da die Anzahl der Antworten zu gering ist (<5).

Der Schwierigkeitsgrad ist ... /

The degree of difficulty is ...

zu leicht / too easy

angemessen / appropriate

zu schwer / too difficult

Ich bewerte das Konzept der Übung mit ... /

I would evaluate the exercise course concept as ...

so gut / very good

gut / good

befriedigend / satisfactory

ausreichend / sufficient

mangelhaft / poor

Vermittlung und Verhalten - Vorlesung / Instruction and Behavior - Lecture

erklärt den Stoff verständlich. /

... explains the subject matter clearly.

stimmt zu / strongly agree: 62.9% 37.1% 2.6% 0% 0%

stimmt nicht zu / strongly disagree: 37.1% 62.9% 97.4% 100%

geht auf Verständnisfragen ein. /

... is willing to answer questions.

stimmt zu / strongly agree: 62.9% 37.1% 2.6% 0% 0%

stimmt nicht zu / strongly disagree: 37.1% 62.9% 97.4% 100%

berücksichtigt unterschiedliche Kenntnisse der Studierenden. /

... considers students' different levels of knowledge.

stimmt zu / strongly agree: 62.9% 37.1% 2.6% 0% 0%

stimmt nicht zu / strongly disagree: 37.1% 62.9% 97.4% 100%

schafft es, mich für den Vorlesungstoff zu begeistern. /

... engages my interest in the topic.

stimmt zu / strongly agree: 23.9% 44.4% 22.2% 3.6% 0%

stimmt nicht zu / strongly disagree: 76.1% 55.6% 77.8% 96.4% 100%

... ist gut vorbereitet. /

... is well prepared.

stimmt zu / strongly agree: 12.9% 21.6% 0% 76.1% 0%

stimmt nicht zu / strongly disagree: 87.1% 78.4% 100% 0% 100%

... ist außerhalb der Vorlesung ansprechbar. /

... is available outside of the lecture.

stimmt zu / strongly agree: 1.0% 1.2% 0% 97.8% 0%

stimmt nicht zu / strongly disagree: 98.9% 98.8% 100% 0% 100%

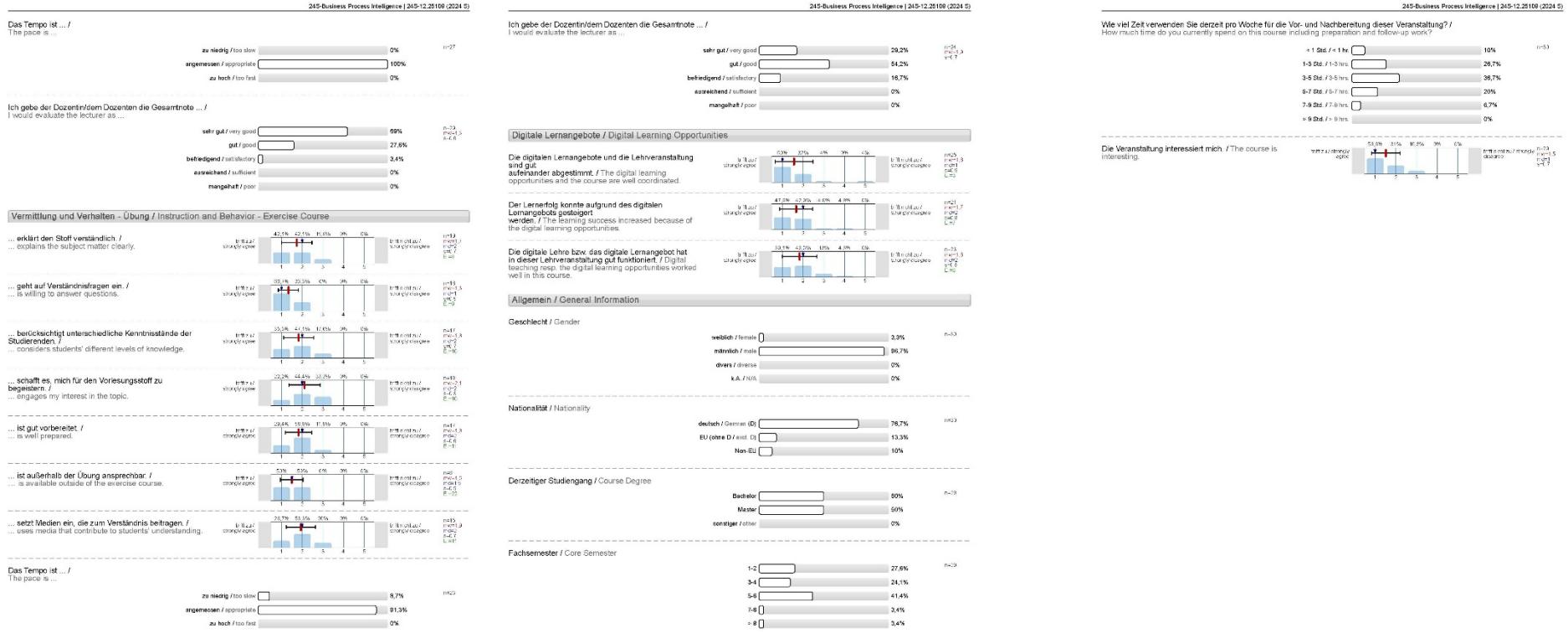
... setzt Medien ein, die zum Verständnis beitragen. /

... uses media that contribute to students' understanding.

stimmt zu / strongly agree: 54.3% 33.3% 11.0% 0% 0%

stimmt nicht zu / strongly disagree: 45.7% 66.7% 88.9% 100% 100%

Survey Results: Thank You!



Open comments and suggestions

- **Comments about lectures are positive.**
- **Positive about the link between theory and practice.**
- **Negative comments mostly about assignments (unclarity, changes) and overlap of timeslots.**
- **Some would like to see more technical details.**

Questions to you:

- **Why are not more students at the lecture?**
- **What would you like to see differently?**



Chair of Process
and Data Science

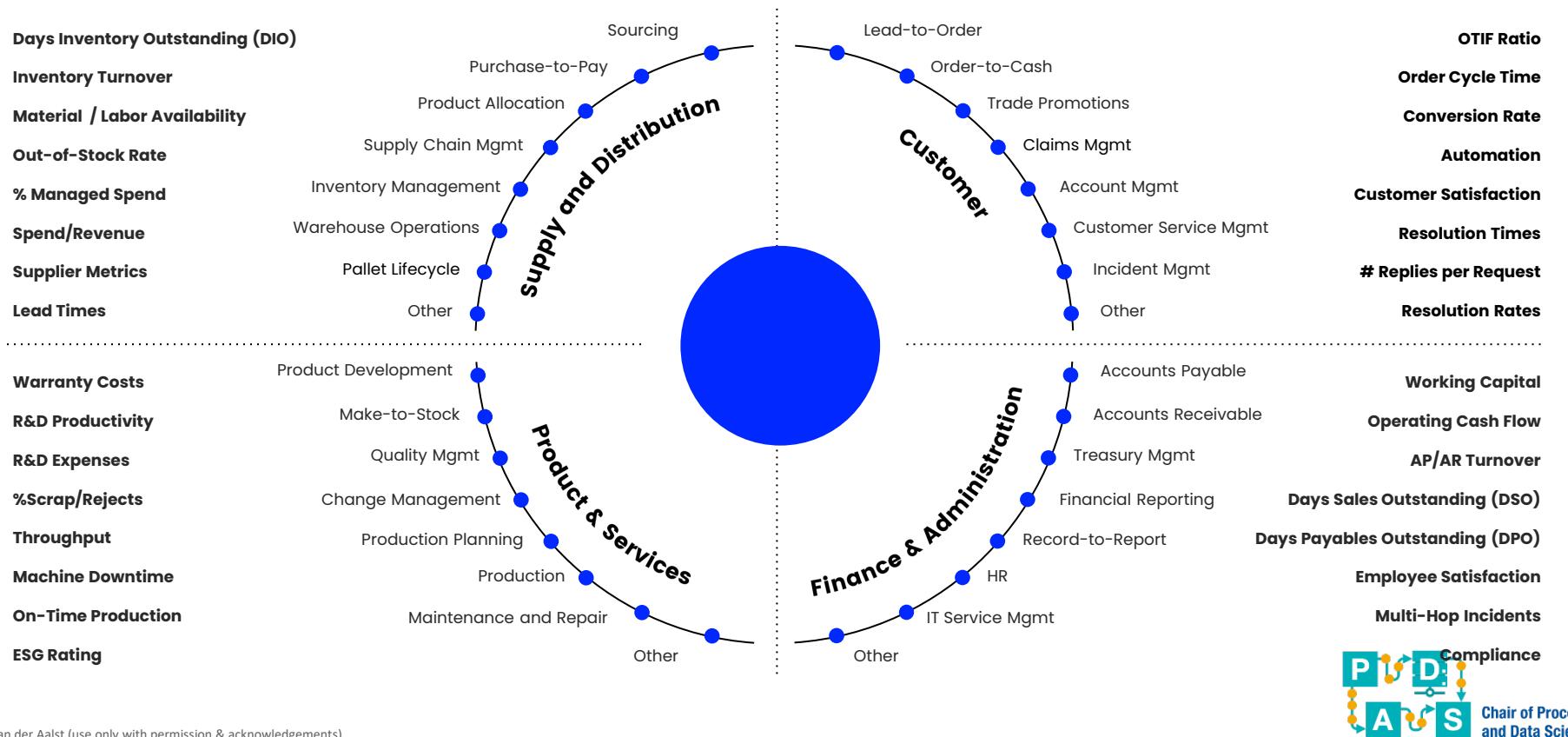
Performance analysis





**process mining is most interesting for processes
having performance problems**

Many Processes – Many Problems



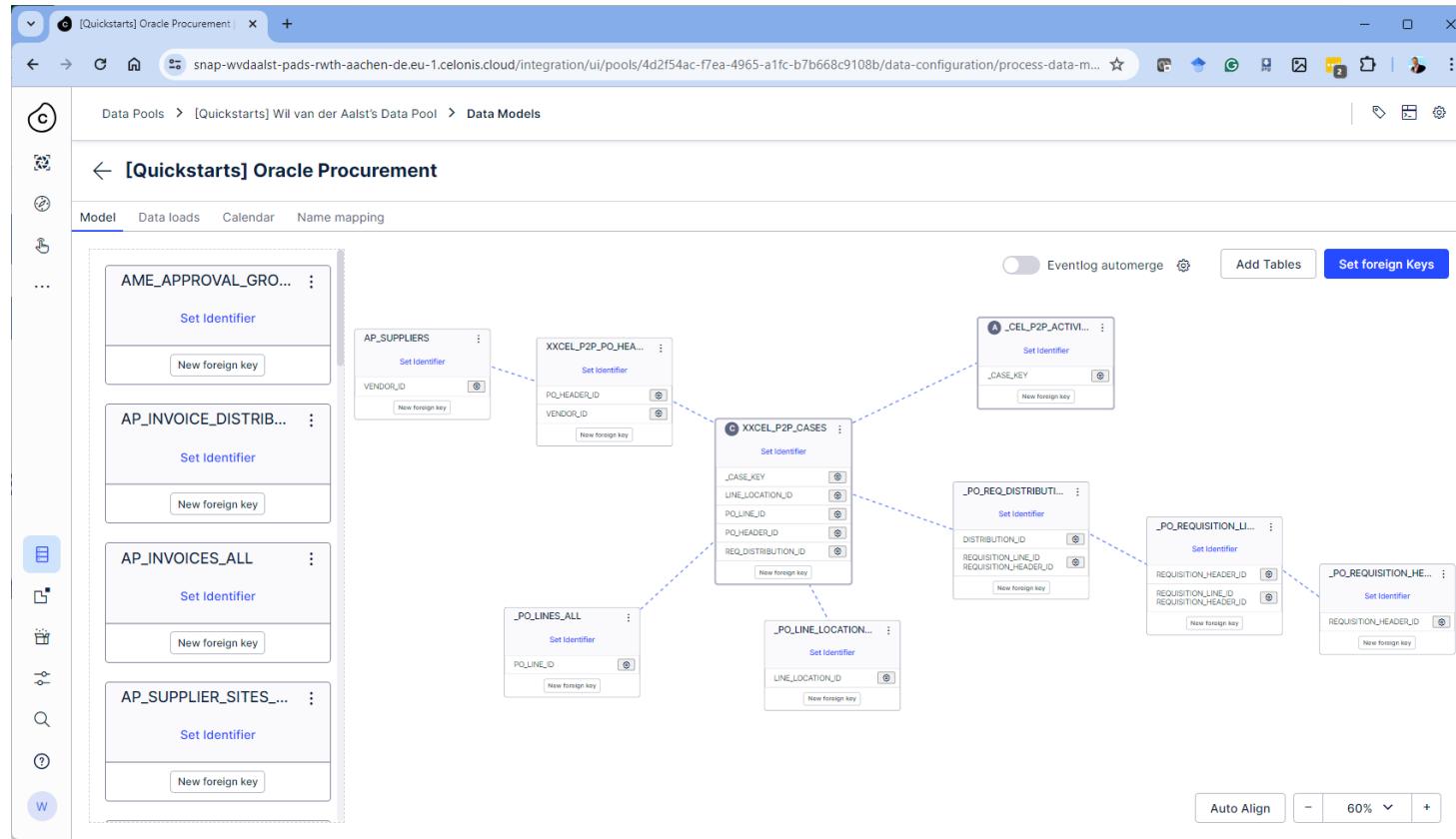
Example Key Performance Indicators (KPIs)

- **Touchless Rate or Automation Rate:** Percentage of orders automatically handled
- **OTIF (On Time In Full) Rate:** Percentage of orders delivered on the date agreed (on time) and with the quantities expected (in full).
- **Order Change Rate:** Percentage of orders requiring a change (e.g., a price change).
- **Maverick Buying Rate :** Percentage of orders placed without the involvement of the purchasing department.

Example Key Performance Indicators (KPIs)

- **Churn Rate:** Percentage of customers lost in a particular time period.
- **Duplicate Payment Rate:** Percentage of invoices paid multiple times.
- **Return Rate:** Percentage of ordered items returned to supplier.
- **Segregation of Duties (SoD) Violation Rate:** Percentage of cases violating the 4-eye principle.

Example: Oracle Procurement Data Model



Chair of Process
and Data Science

Oracle Procurement KPIs

The screenshot shows a comparison between two KPI definitions in a Celonis Cloud Studio interface.

Left Panel (KPIs List):

- Records: 29
- KPIs: 35
- Augmented Attributes: 0
- Filters: 0
- Triggers: 0
- Actions: 0
- Flags: 0
- Variables: 0
- Event Logs: 1
- Custom Objects: 0

Middle Panel (KPIs Table):

Name	ID	PQL Formula	Source
Count Table Po Req Distrib...	COUNT_TABLE_U95_PO_RE...	COUNT_TABLE("PO_REQ_DI...")	
Count Table Po Requisitio...	COUNT_TABLE__PO_REQUISI...	COUNT_TABLE("PO_REQUISI...")	
Count Table Po Requisitio...	COUNT_TABLE_U95_PO_RE...	COUNT_TABLE("PO_REQUISI...")	
Count Table Po Requisitio...	COUNT_TABLE__PO_REQUISI...	COUNT_TABLE("PO_REQUISI...")	
Count Table Rcv Shipmen...	COUNT_TABLE__RCV_SHIPM...	COUNT_TABLE("RCV_SHIPME...")	
Count Table Rcv Transact...	COUNT_TABLE__RCV_TRANS...	COUNT_TABLE("RCV_TRANSACT...")	
Count Table Xcel P2p C...	COUNT_TABLE__XXCEL_P2P...	COUNT_TABLE("XXCEL_P2P...")	
Count Table Xcel P2p P...	COUNT_TABLE__XXCEL_P2P...	COUNT_TABLE("XXCEL_P2P...")	
Filtered Count	FILTERED_COUNT	COUNT(CASE WHEN {p1} TH...	
Number Of Process Vari...	NUMBER_OF_PROCESS_VARI...	COUNT(DISTINCT SHORTEN...	
Process Variants Cel P2p ...	PROCESS_VARIANTS__CEL_P...	SHORTENED(VARIANT("CEL...")	
Ratio	RATIO	AVG(CASE WHEN {p1} THEN ...	
Total Throughput Time In...	TOTAL_THROUGHPUT_TIME...	AVG(CALC_THROUGHPUT(C...	

Right Panel (KPI Detail View):

- Total Throughput Time In Days ...** (Same as Base | Autogenerated | ⋮ | X)
- Please be aware that any change could impact other usages of this object
- Display Name:** Total Throughput Time In Days Cel P2p Activities Eventtime
- Short Display Name:** (empty)
- Description:** (empty)
- Internal Note:** (empty)
- Allow other analysts to edit KPI from extension Knowledge Model:**
- PQL Formula:**

```
AVG(CALC_THROUGHPUT(CASE_START TO CASE_END, REMAP_TIMESTAMPS("_CEL_P2P_ACTIVITIES"."EVENTTIME", DAYS)))
```
- Parameters:** (empty)
- Value:** Total Throughput Time In Days Cel P2p Activities Eventtime
0.060923948463428086

A large blue arrow points from the "Total Throughput Time In..." row in the middle panel to the "PQL Formula" field in the right panel.



Oracle Procurement Using a KPI

The screenshot shows the Oracle Data View interface. At the top, there's a navigation bar with 'Studio > BPI 2024 > Oracle Data'. Below it is a toolbar with various icons. The main area has a large blue header 'KPI' followed by a large numerical value '0.060923948463428086'. To the left of this is a process flow diagram with nodes like 'Start Case', 'Create PO Requisition', 'Approve Purchase Order', and 'End Case'. A blue arrow points from the KPI value towards a pie chart on the right. The pie chart is divided into four segments: 'Enter Invoice' (18.13%), 'Submit... (16.76%)', 'Approve Purchase Order (16.72%)', and 'Create Purchase Order (16.72%)'. Below the pie chart is a 'KPI' table with one row:

KPI	Value
0.060923948463428086	

On the left side, there's a 'PQL Editor' window with a search bar and a list of data sources. The list includes: Cel P2p Activities, Po Line Locations All, Po Lines All, Po Req Distributions All, Po Requisition Headers All, Po Requisition Lines All, Ame Approval Groups, Ap Invoice Distributions A, Ap Invoices All, Ap Supplier Sites All, Ap Suppliers, Cur Symbol, Qi Daily Rates, Gi Ledgers, Po Action History, Po Distributions All, Po Distributions Archive A, Po Headers All, and Po Line Locations All.

Let focus on time and probabilities



bottlenecks?

where?

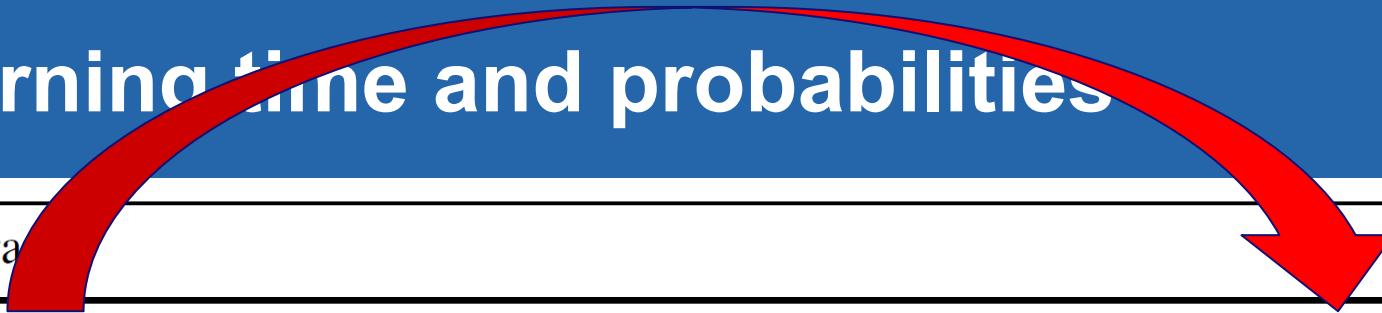
why?

Event data with timestamps and transactional information

a[12,19] b[25,32] d[26,33] e[35,40] e[50,54]

case id	trace
1	$\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26}, b_{complete}^{32}, d_{complete}^{33}, e_{start}^{35}, e_{complete}^{40}, h_{start}^{50}, h_{complete}^{54} \rangle$
2	$\langle a_{start}^{17}, a_{complete}^{23}, d_{start}^{28}, c_{start}^{30}, d_{complete}^{32}, c_{complete}^{38}, e_{start}^{50}, e_{complete}^{59}, g_{start}^{70}, g_{complete}^{73} \rangle$
3	$\langle a_{start}^{25}, a_{complete}^{30}, c_{start}^{32}, c_{complete}^{35}, d_{start}^{35}, d_{complete}^{40}, e_{start}^{45}, e_{complete}^{50}, f_{start}^{50}, f_{complete}^{55}, b_{start}^{60}, d_{start}^{62}, b_{complete}^{65}, d_{complete}^{67}, e_{start}^{80}, e_{complete}^{87}, g_{start}^{90}, g_{complete}^{98} \rangle$
...	...

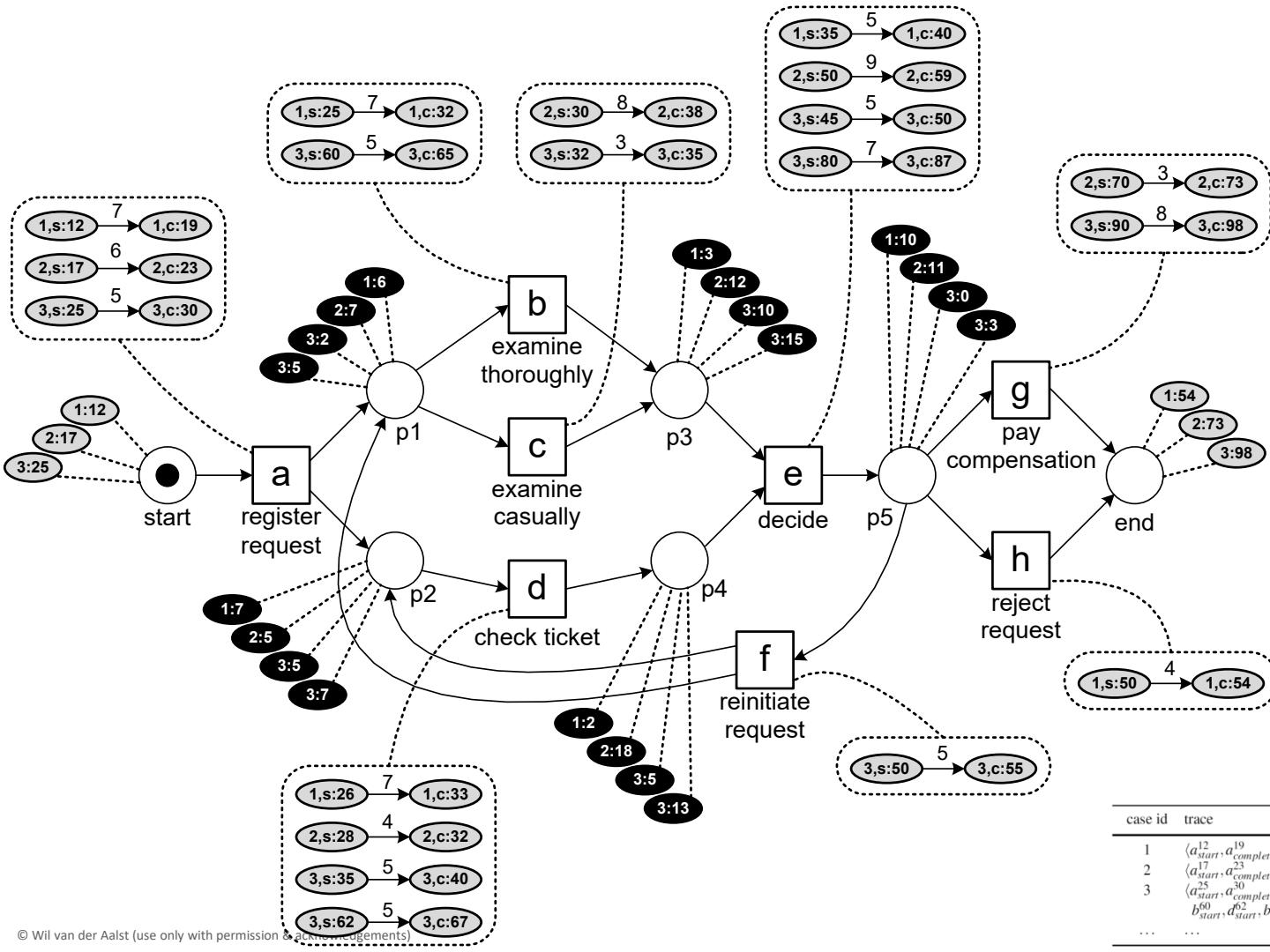
Learning time and probabilities



case id	tra
1	$\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26}, b_{complete}^{32}, d_{complete}^{33}, e_{start}^{35}, e_{complete}^{40}, h_{start}^{50}, h_{complete}^{54} \rangle$
2	$\langle a_{start}^{17}, a_{complete}^{23}, d_{start}^{28}, c_{start}^{30}, d_{complete}^{32}, c_{complete}^{38}, e_{start}^{50}, e_{complete}^{59}, g_{start}^{70}, g_{complete}^{73} \rangle$
3	$\langle a_{start}^{25}, a_{complete}^{30}, c_{start}^{32}, c_{complete}^{35}, d_{start}^{35}, d_{complete}^{40}, e_{start}^{45}, e_{complete}^{50}, f_{start}^{50}, f_{complete}^{55}, b_{start}^{60}, d_{start}^{62}, b_{complete}^{65}, d_{complete}^{67}, e_{start}^{80}, e_{complete}^{87}, g_{start}^{90}, g_{complete}^{98} \rangle$
...	...

- **case 1 starts at time 12 and ends at time 54**
- **case 2 starts at time 17 and ends at time 73**
- **case 3 starts at time 25 and ends at time 98**

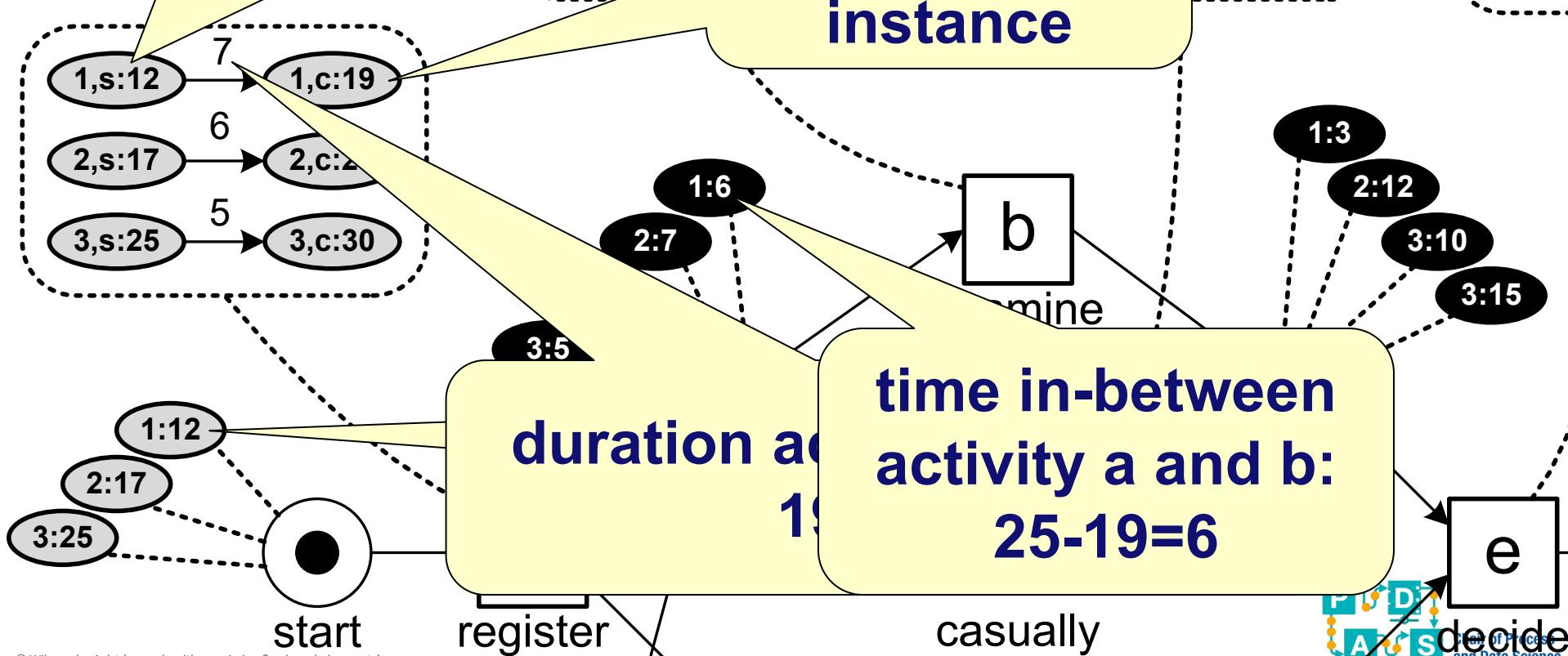
replaying the first three cases

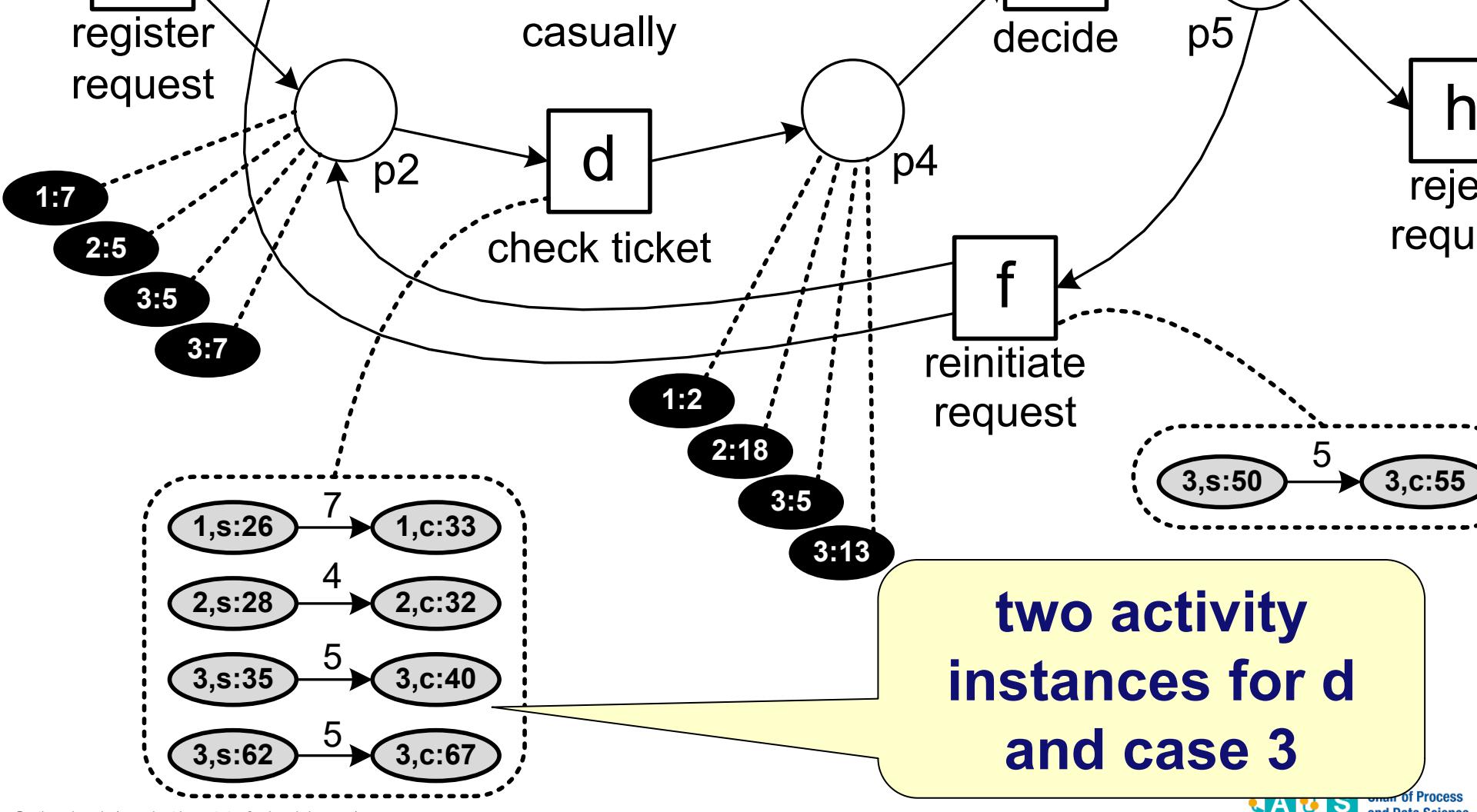


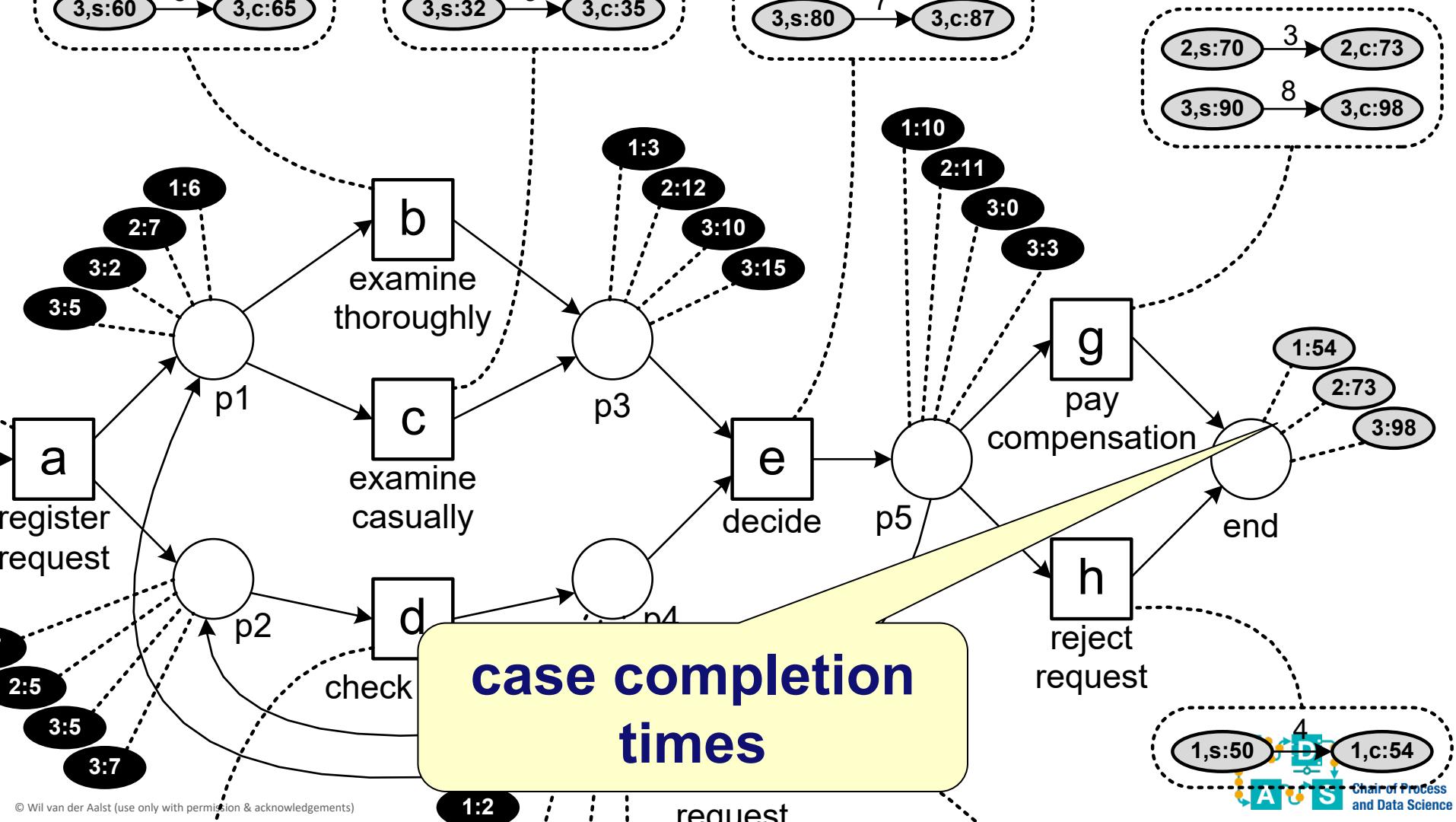
case id	trace
1	$\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26}, b_{complete}^{32}, d_{complete}^{33}, e_{start}^{35}, e_{complete}^{40}, h_{start}^{50}, h_{complete}^{54} \rangle$
2	$\langle a_{start}^{17}, a_{complete}^{23}, d_{start}^{28}, c_{start}^{30}, d_{complete}^{32}, c_{complete}^{38}, e_{start}^{45}, e_{complete}^{50}, g_{start}^{59}, g_{complete}^{73} \rangle$
3	$\langle a_{start}^{25}, a_{complete}^{30}, c_{start}^{32}, c_{complete}^{35}, d_{start}^{35}, d_{complete}^{40}, e_{start}^{45}, e_{complete}^{50}, f_{start}^{55}, b_{start}^{60}, d_{start}^{62}, b_{complete}^{67}, d_{complete}^{70}, e_{start}^{87}, e_{complete}^{90}, g_{start}^{98} \rangle$
...	...

start time activity instance

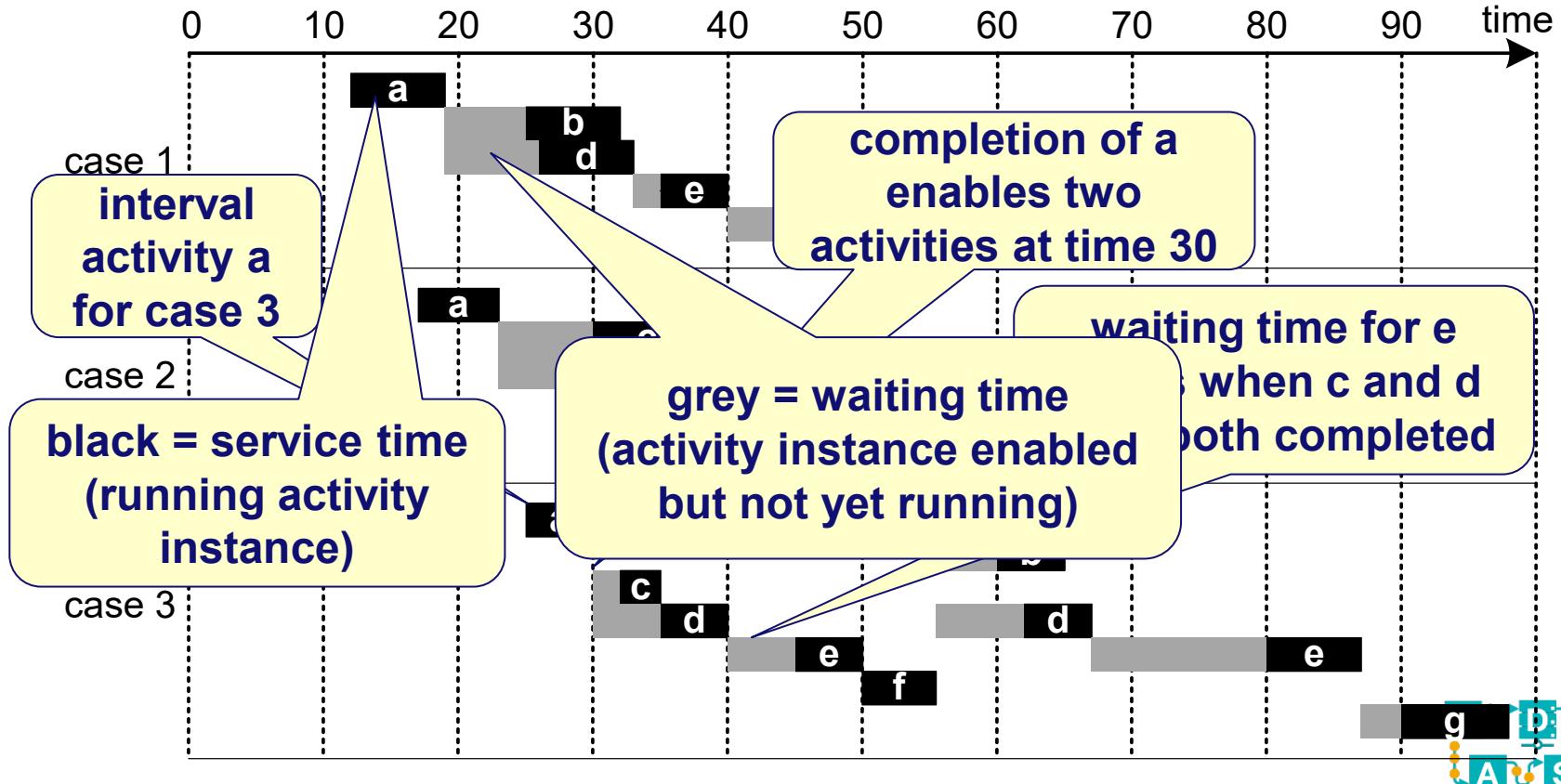
completion time activity instance



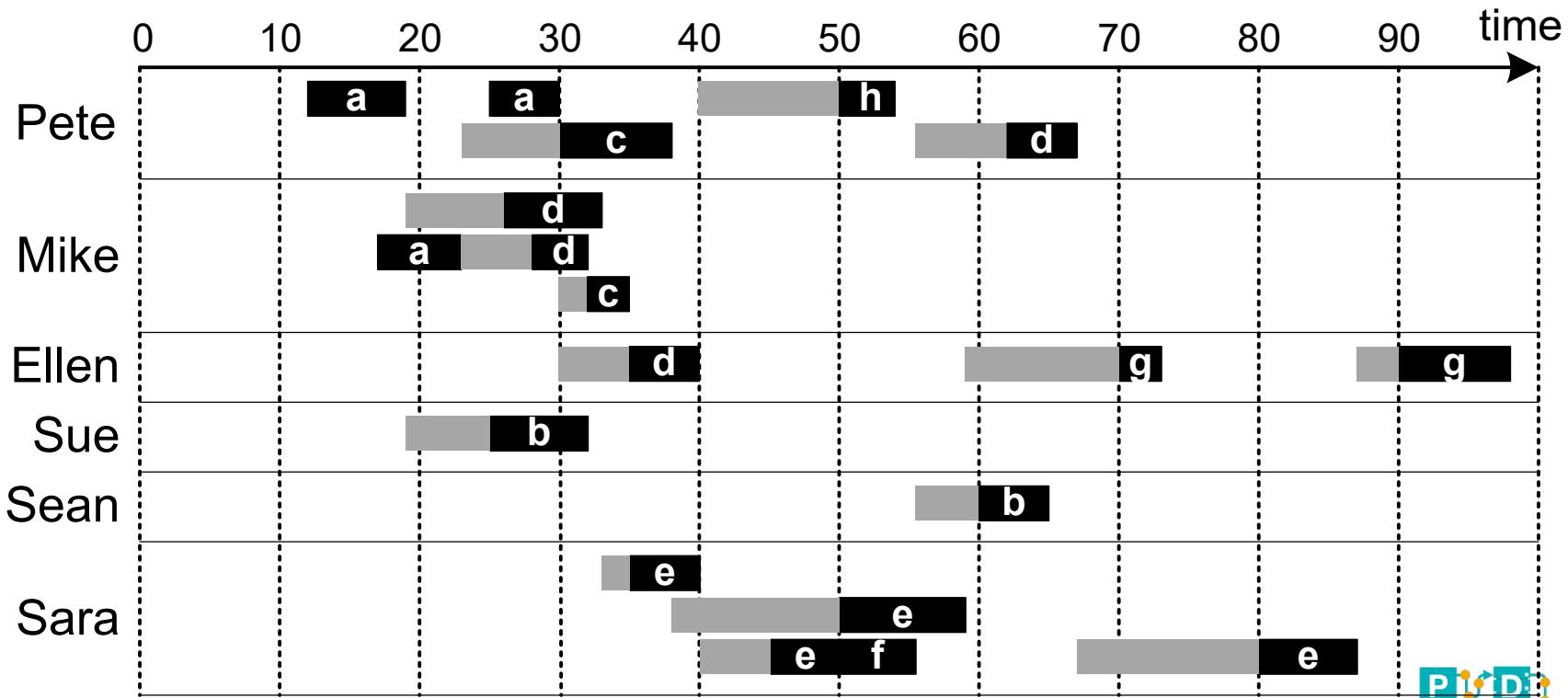




Another view on the timed replay of the first three cases



Timed replay projected onto resources

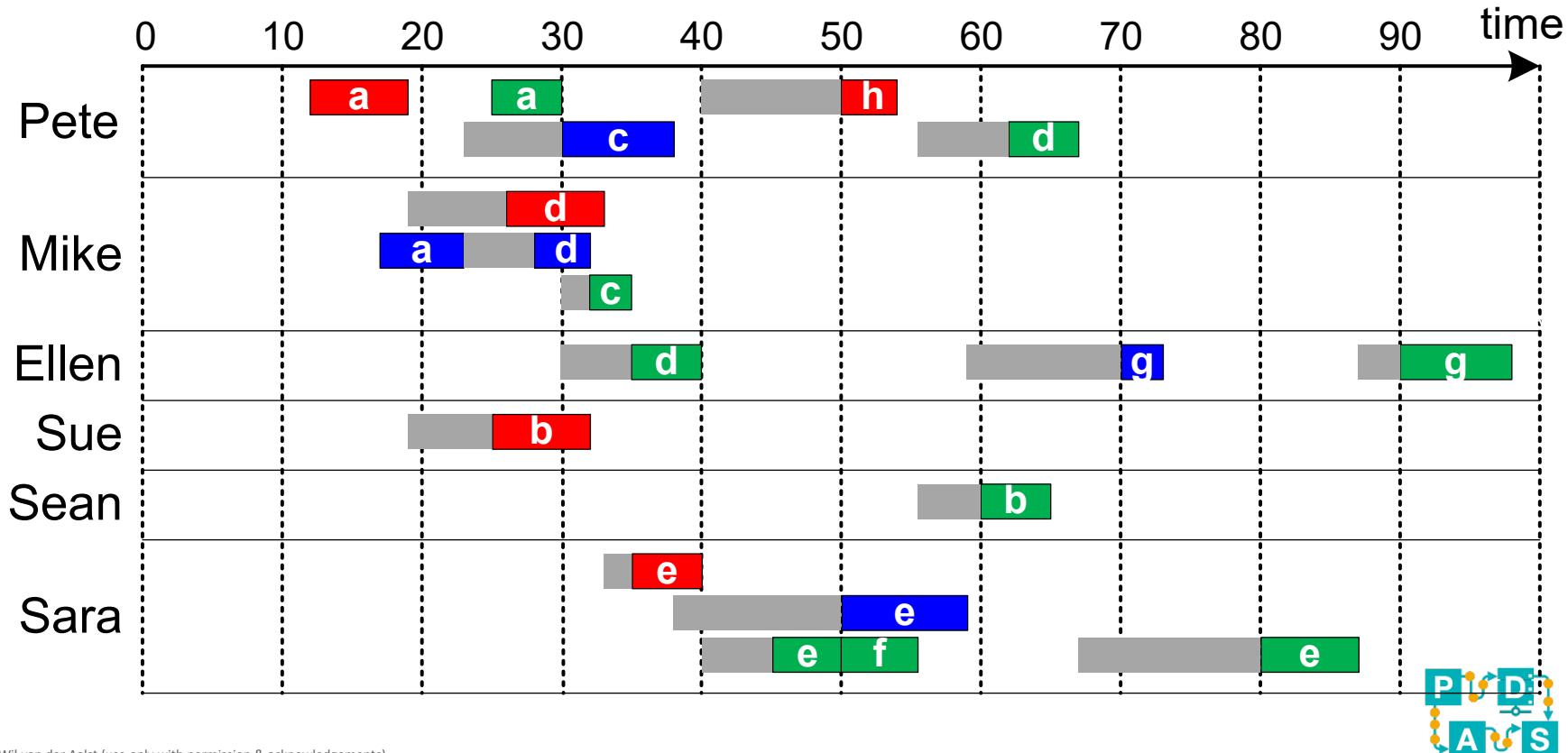


Timed replay projected onto resources (activities colored by case)

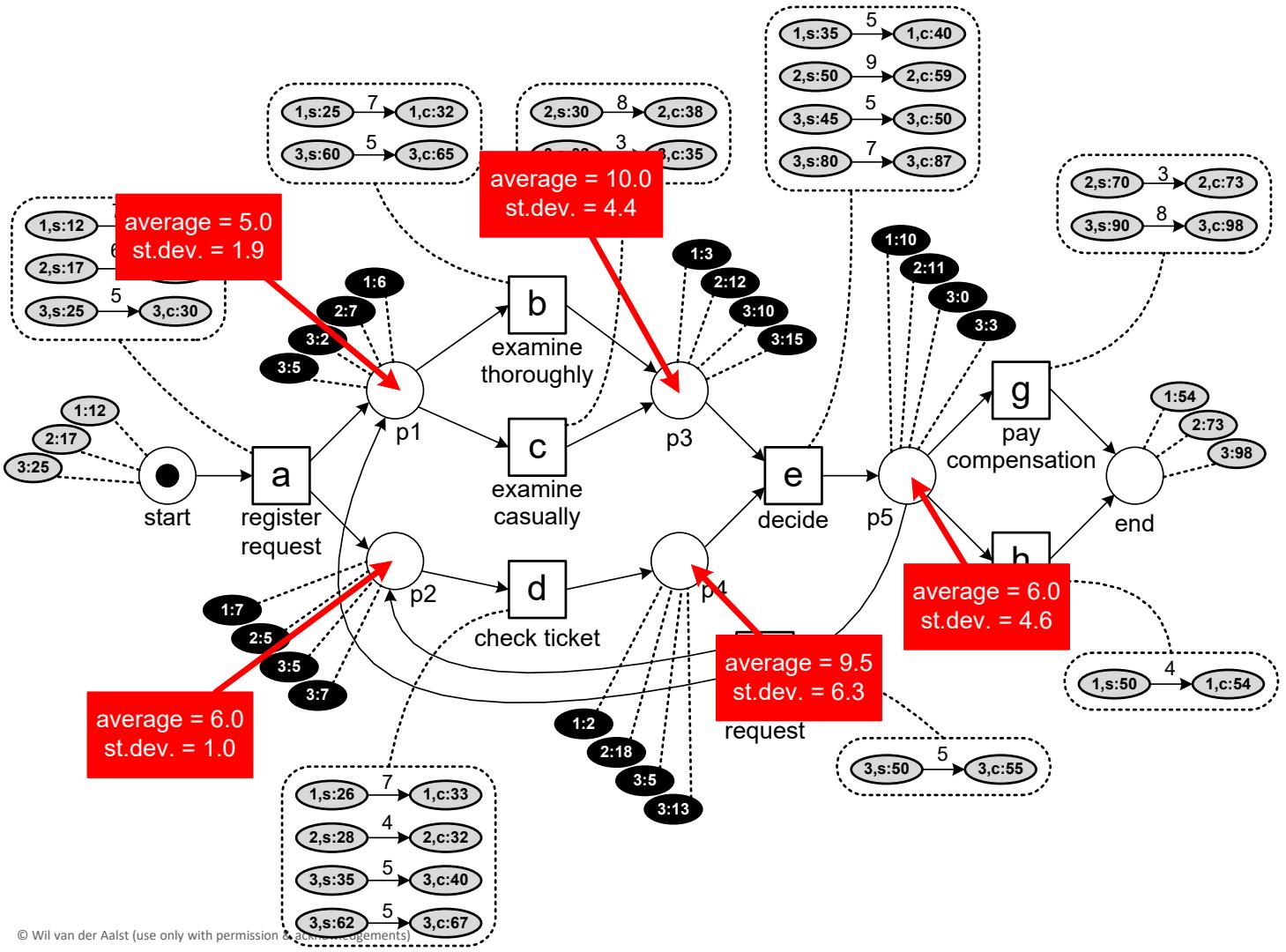
1

2

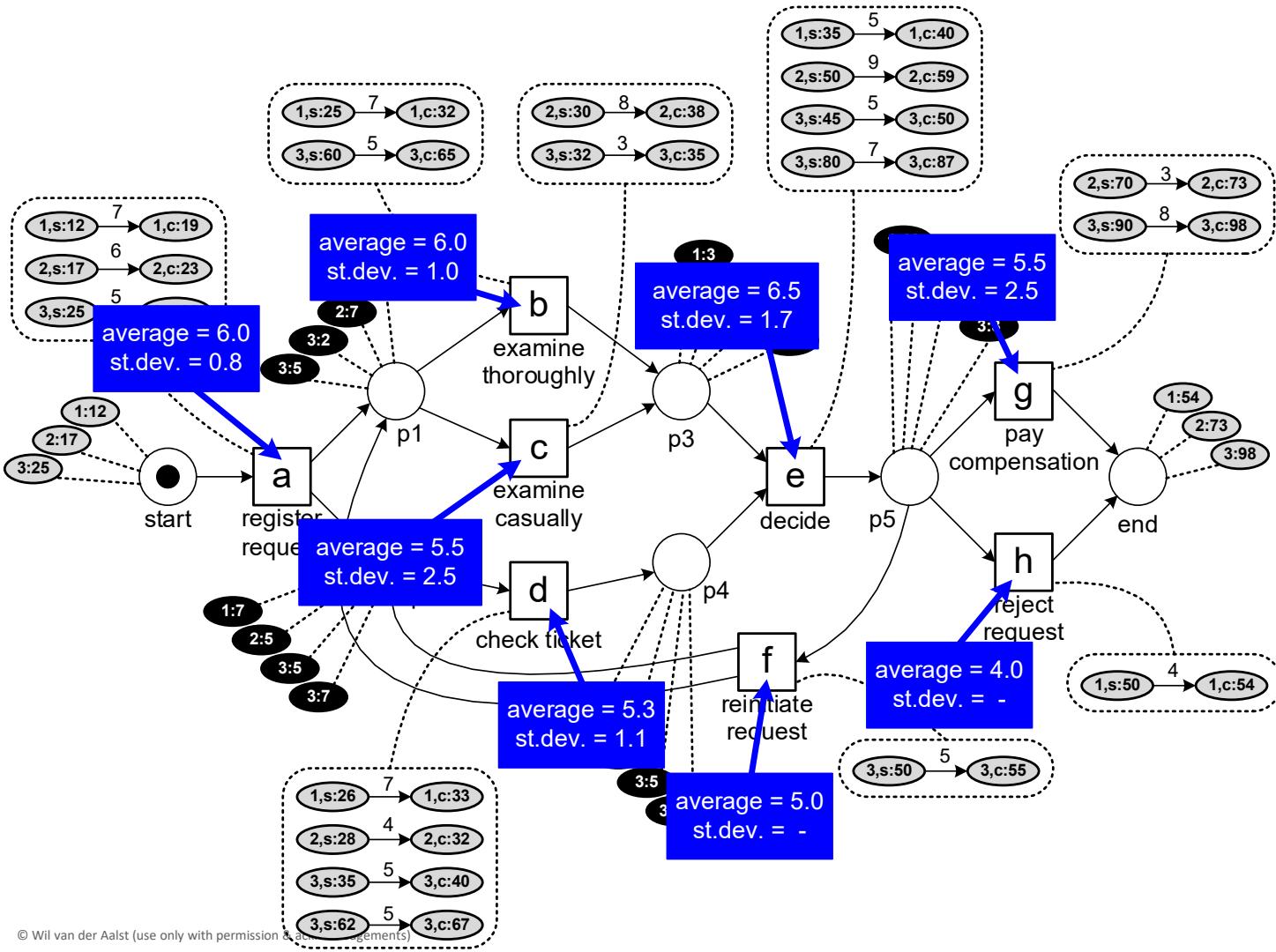
3



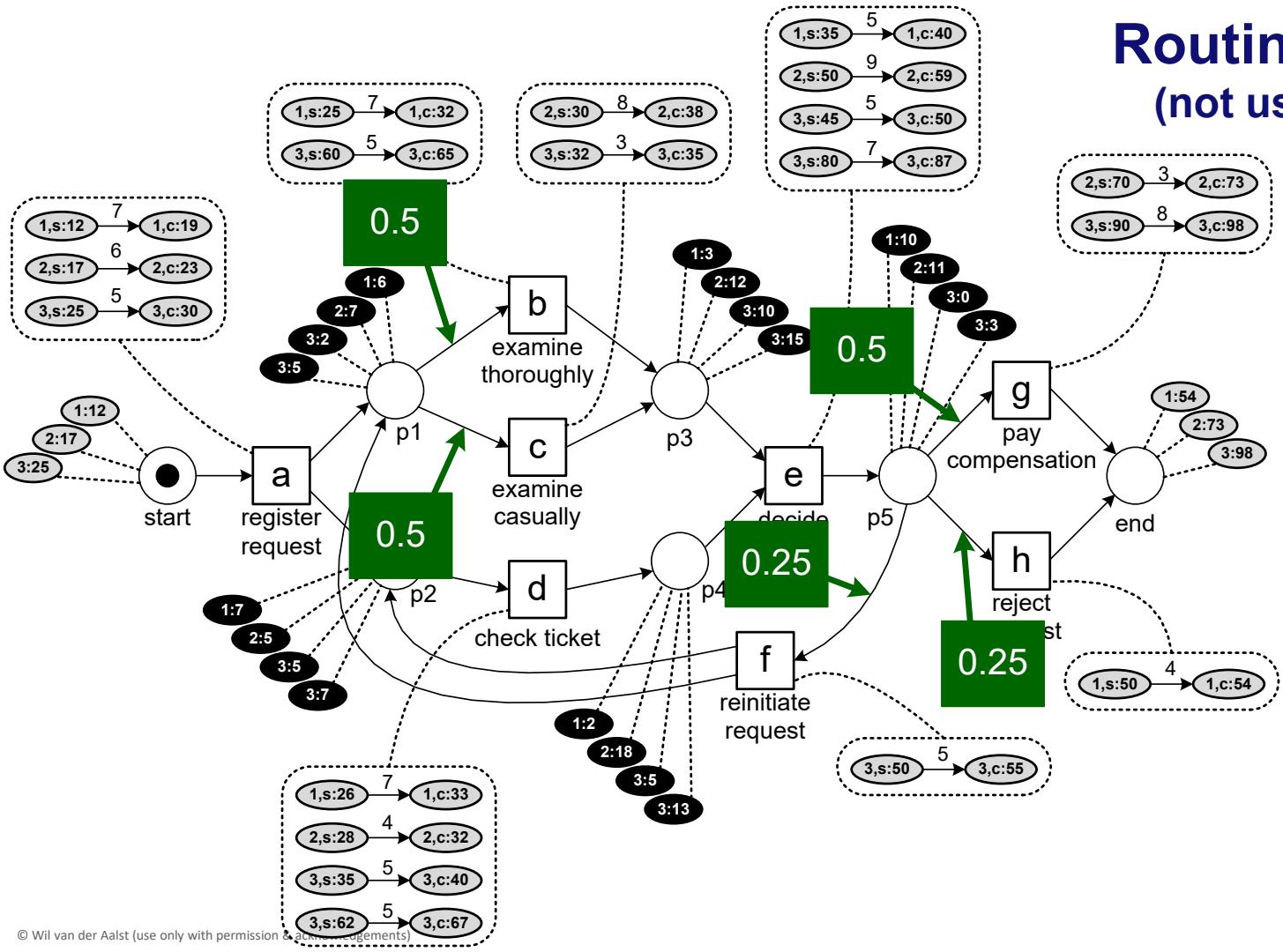
Waiting times (average and standard deviation)



Service times (average and standard deviation)



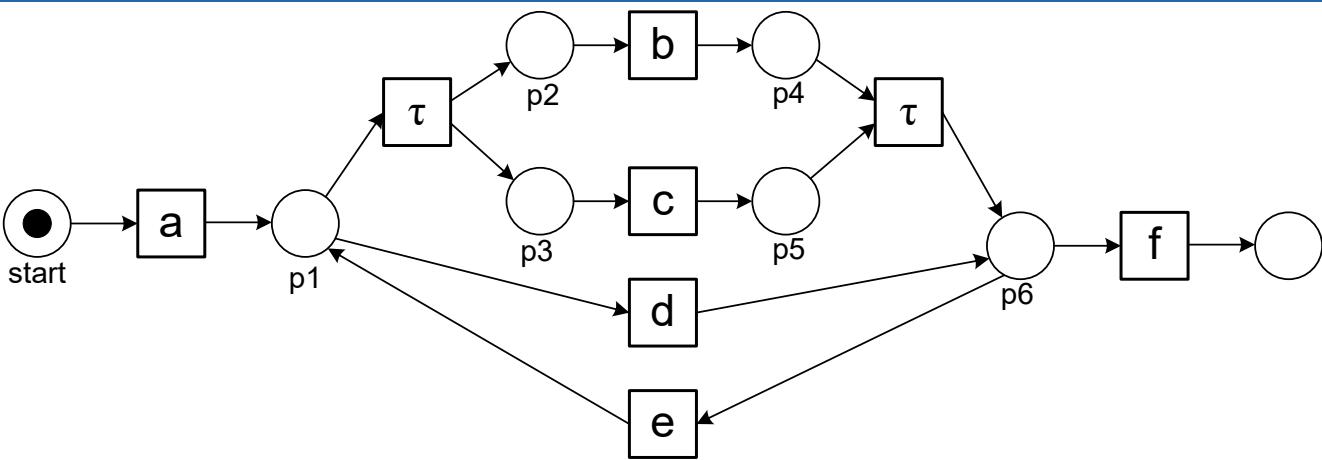
Routing probabilities (not using data attributes)



Another event log

case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete
1	c	complete	18	Sue
3	a	start	20	Pete
2	b	start	22	Mary

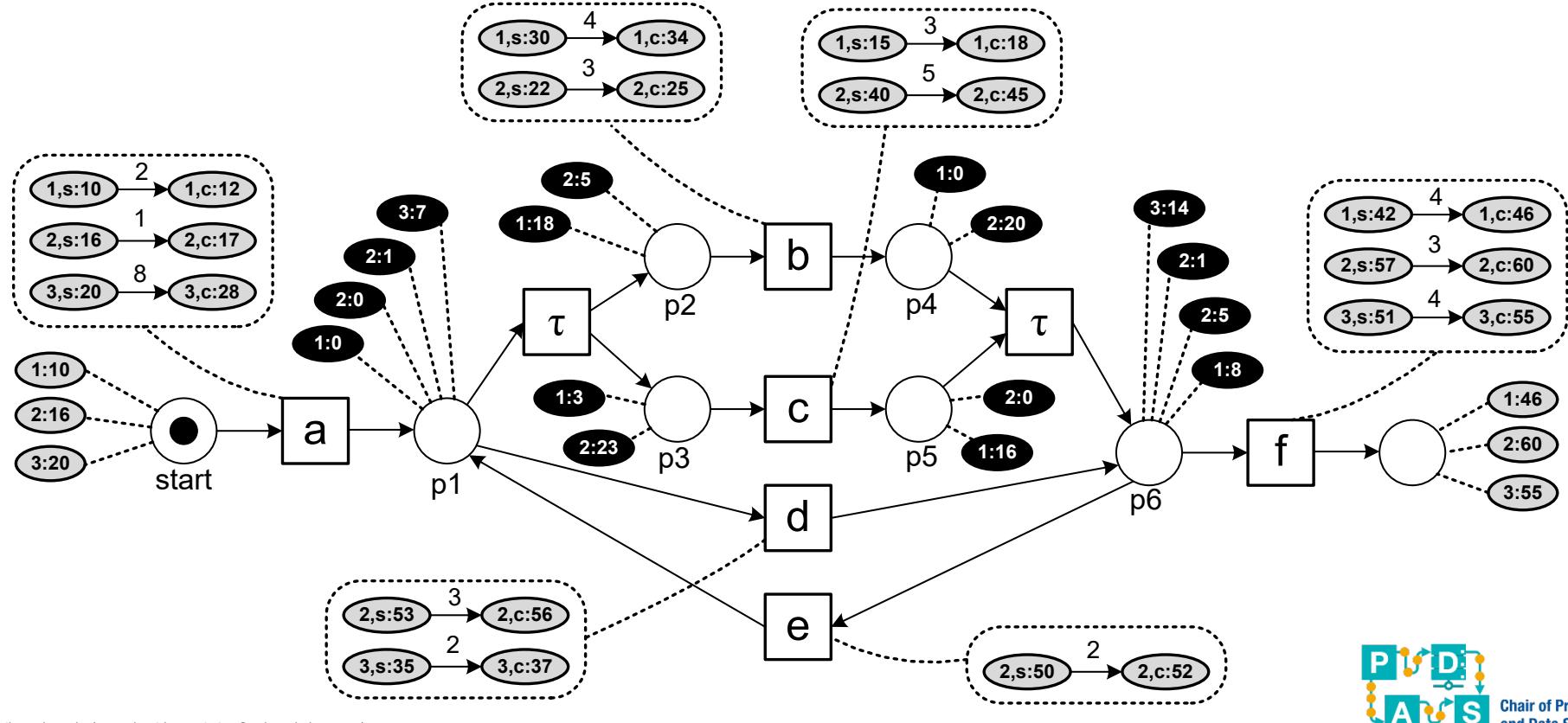
The corresponding model



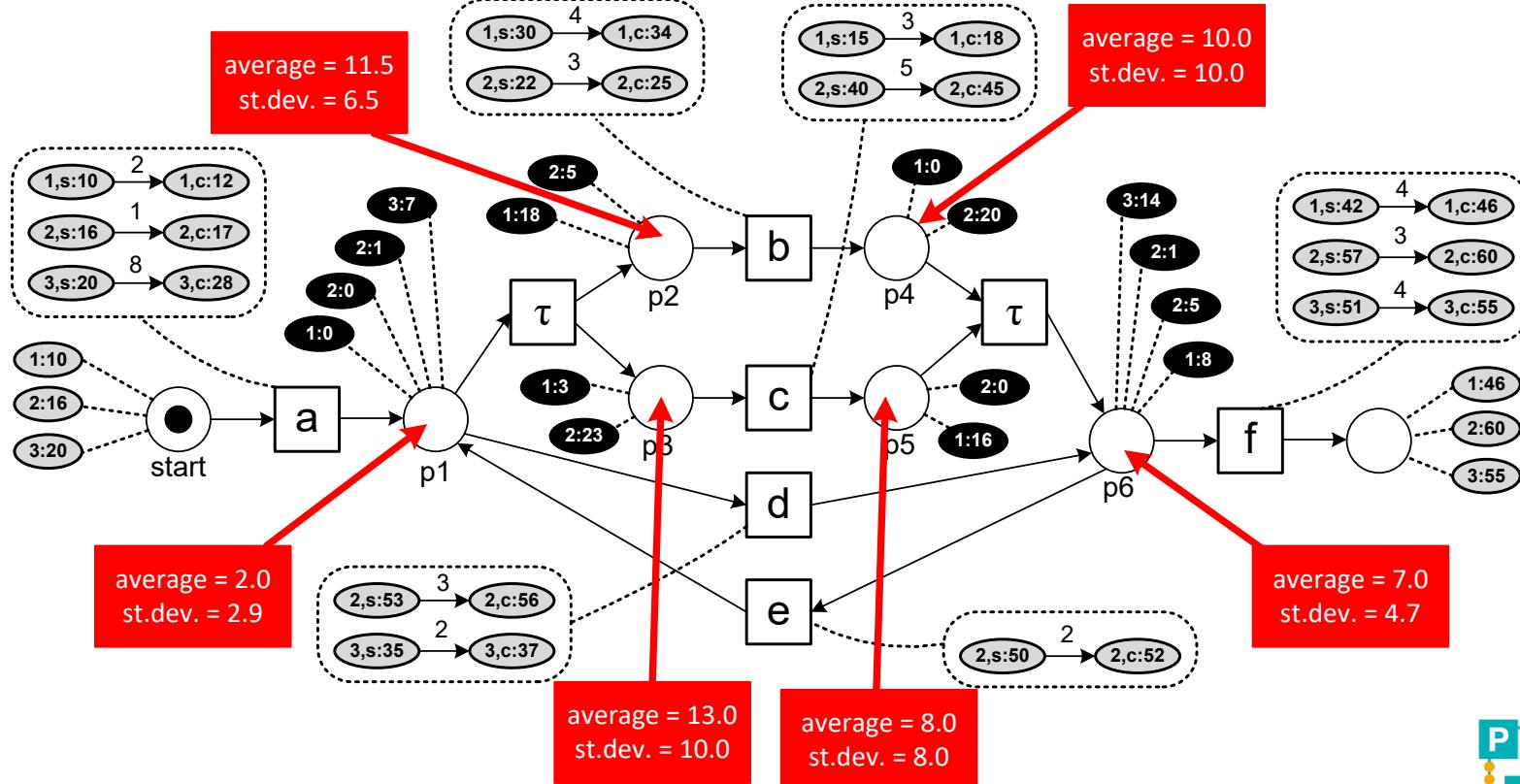
Let's estimate service times,
waiting times, and routing
probabilities

case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete
1	c	complete	18	Sue
3	a	start	20	Pete
2	b	start	22	Mary
2	b	complete	25	Mary
3	a	complete	28	Pete
1	b	start	30	Mary
1	b	complete	34	Mary
3	d	start	35	Mary
3	d	complete	37	Mary
2	c	start	40	Sue
1	f	start	42	Carol
2	c	complete	45	Sue
1	f	complete	46	Carol
2	e	start	50	Kirsten
3	f	start	51	Carol
2	e	complete	52	Kirsten
2	d	start	53	Mary
3	f	complete	55	Carol
2	d	complete	56	Mary
2	f	start	57	Carol
2	f	complete	60	Carol

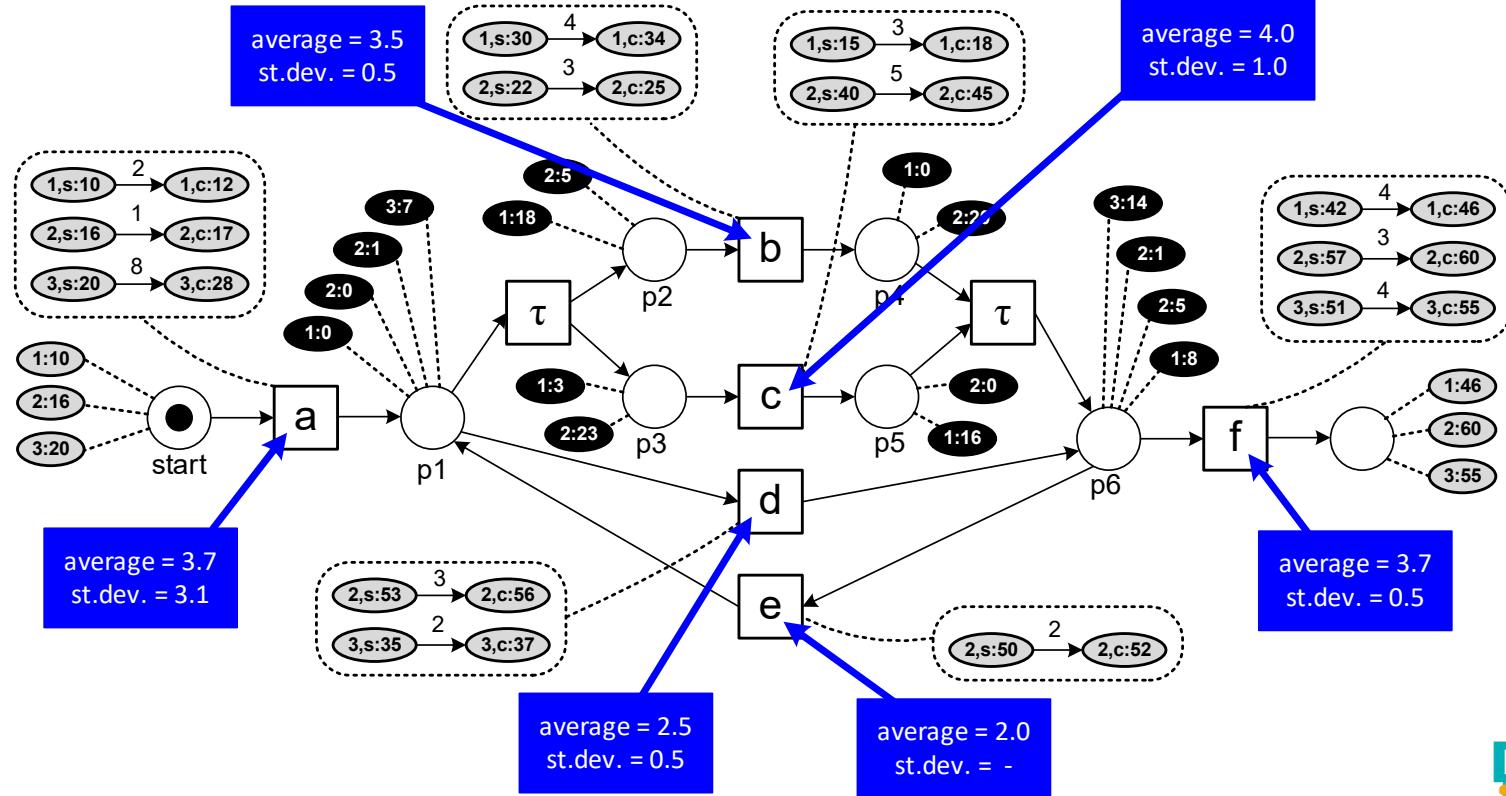
Times recorded during replay



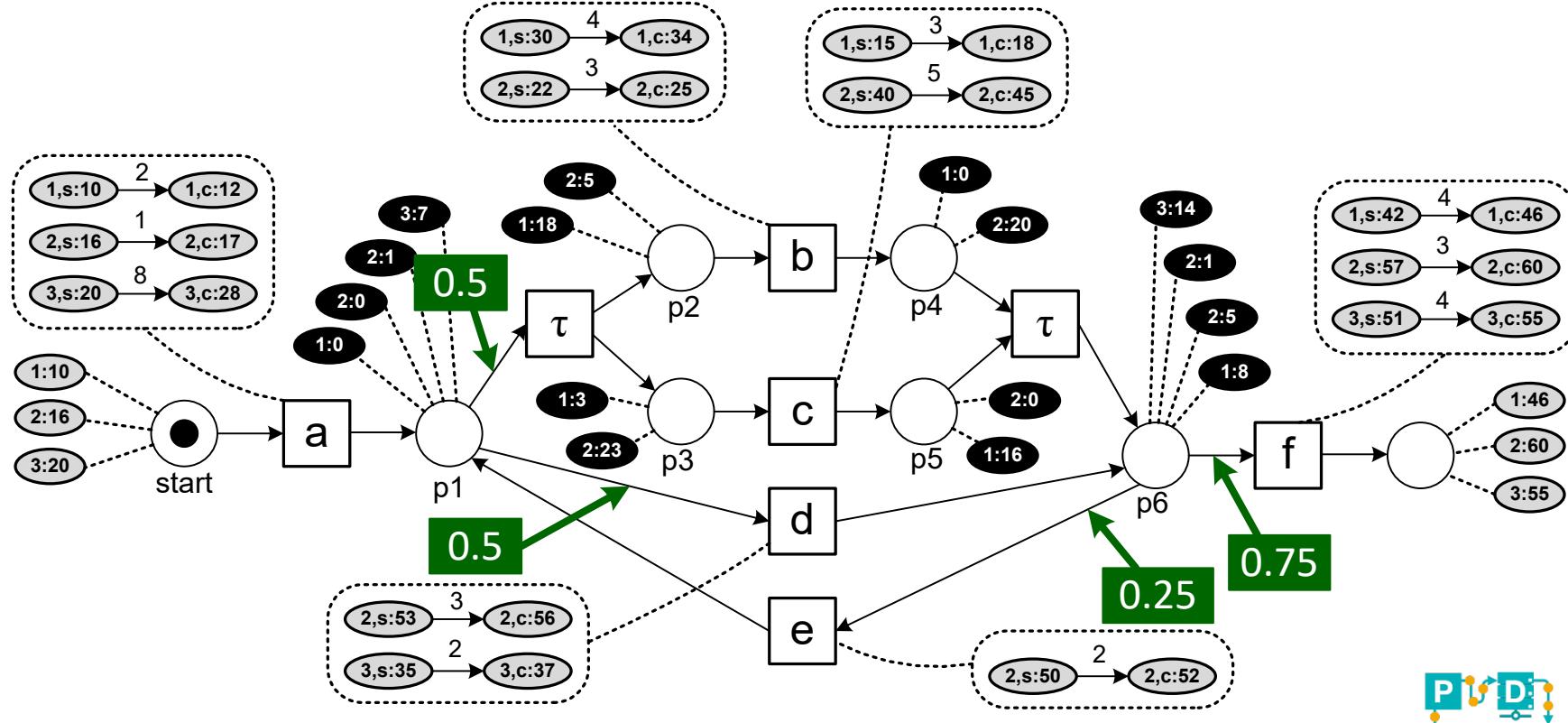
Waiting times



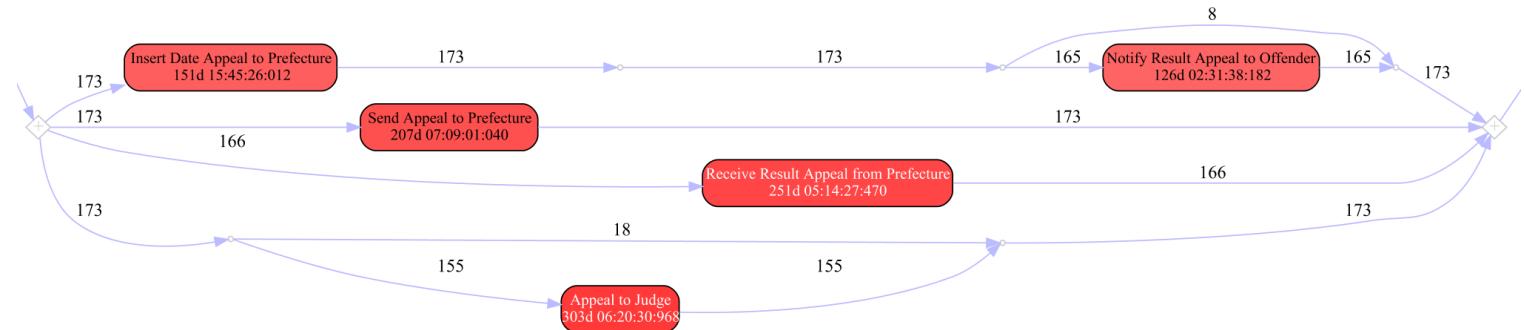
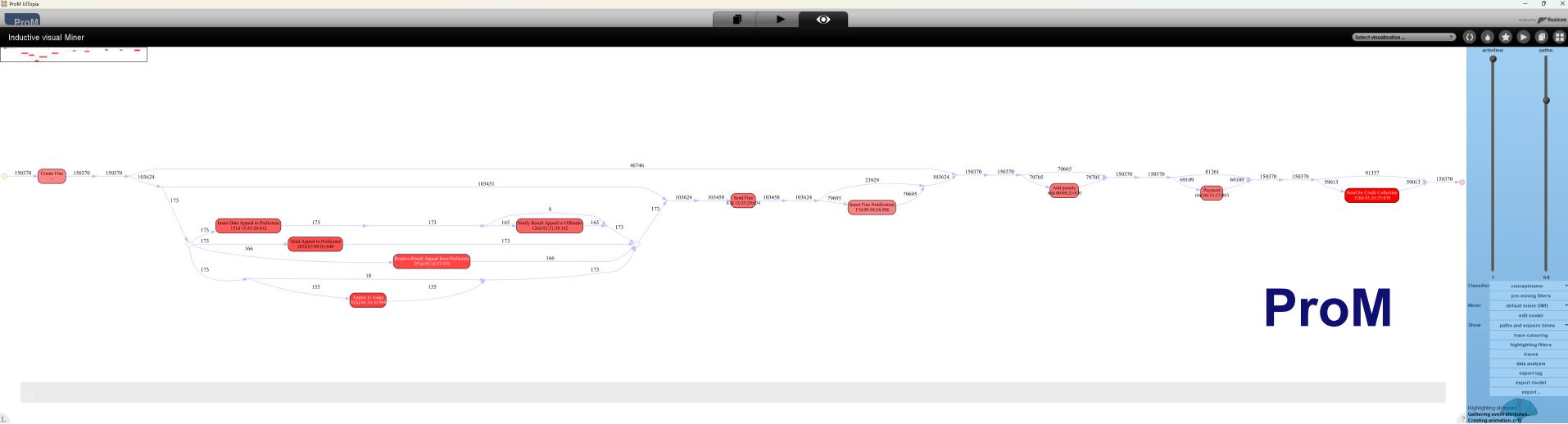
Service times



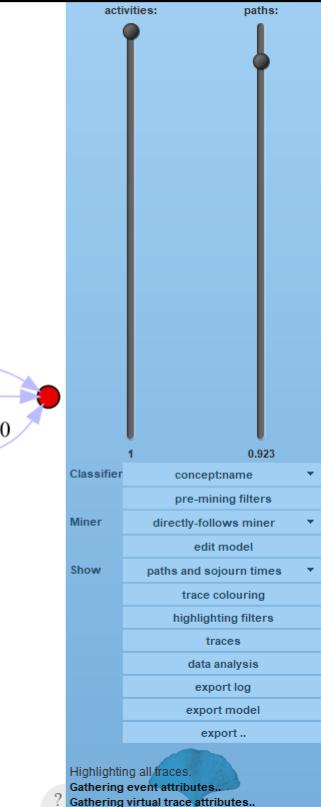
Routing probabilities



we have seen this before ...

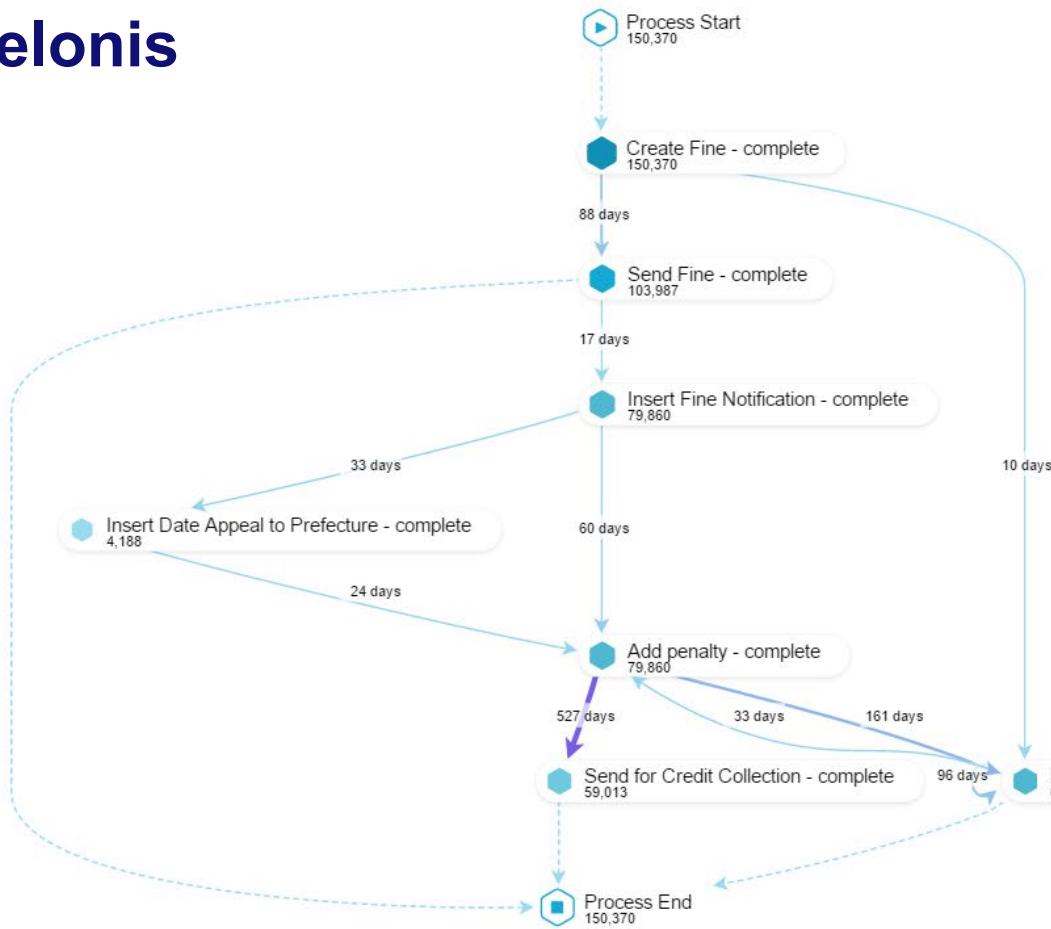


Chair of Process
and Data Science



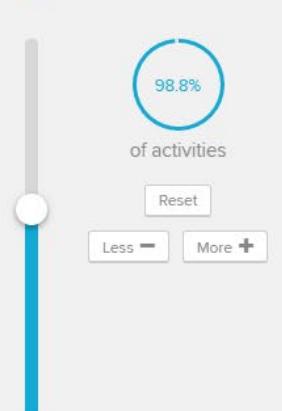


Celonis

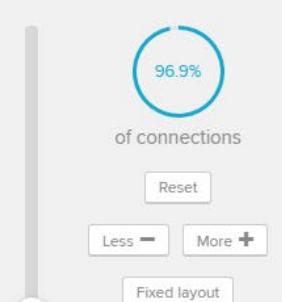


- Zoom +

Activities List view

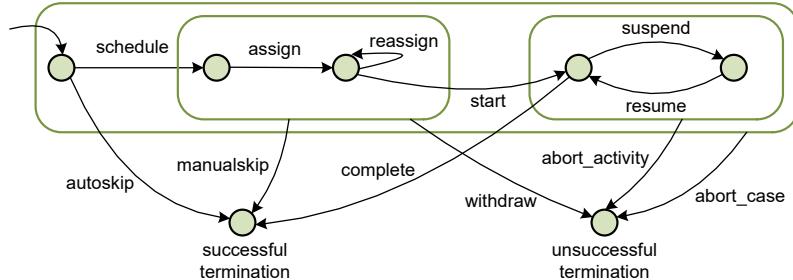
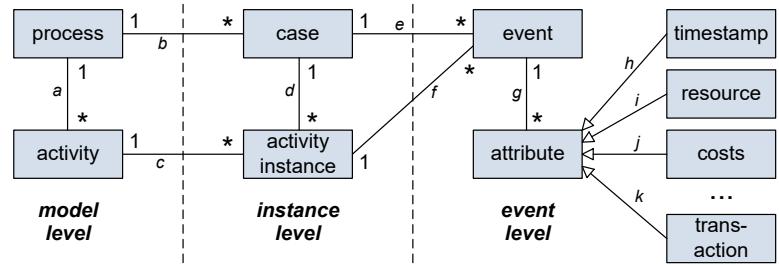


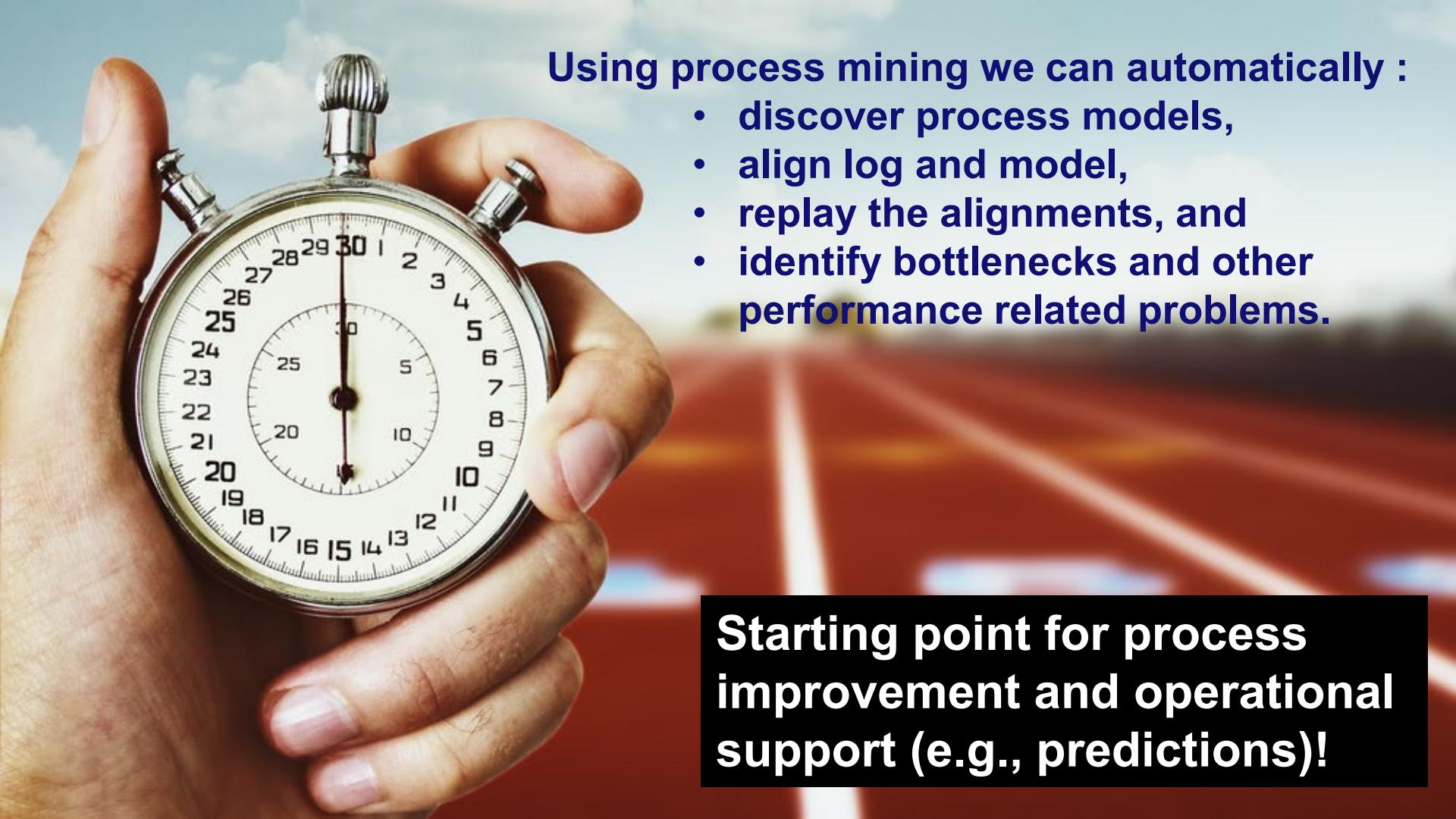
Connections List view



Note on start/complete events

- Events are **atomic**.
- Recall that XES defines a **transactional model**.
- Rarely actively used by commercial systems.
- Some support start/complete only, others use just one type.
- Two options in Celonis
 - Use just the start events (or just the complete events) and add duration as an attribute.
 - Use both events using different names, e.g., A-start and A-complete.





Using process mining we can automatically :

- discover process models,
- align log and model,
- replay the alignments, and
- identify bottlenecks and other performance related problems.

Starting point for process improvement and operational support (e.g., predictions)!

Targeted performance analysis using Celonis

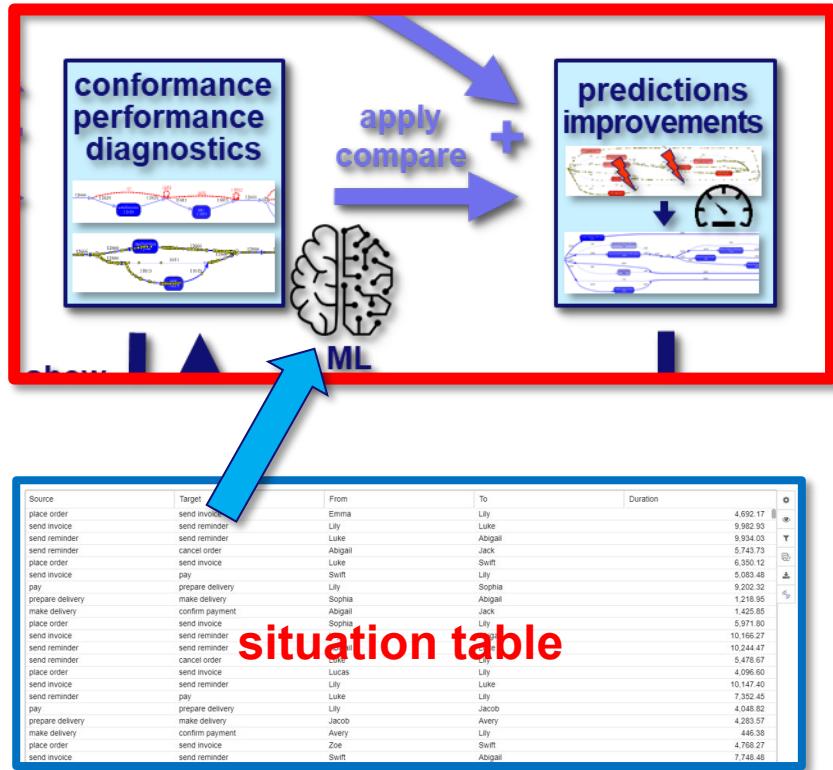


Analyzing performance using situation tables

- **Step 1:** Determine the **response variable** (e.g., throughput time, time elapsed between specific activities).
- **Step 2:** Determine the **type of situation table** (throughput time is a property of cases → case-based table, elapsed time is a property of event pairs → event-pair-based table).
- **Step 3:** Determine **predictor variables**.
- **Step 4:** Analyze the **situation table** (e.g., using RapidMiner).



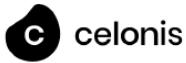
Recall: Situation Tables



- A **situation table** is a two-dimensional table.
 - Each row is an instance.
 - Each column is a variable.
 - There may be a split into a response variable and predictor variables (for supervised learning).
- Five types of situation tables:
 - **Case-based situation table:** Each row (instance) corresponds to a case with variables.
 - **Event-based situation table:** Each row (instance) corresponds to an event.
 - **Resource-based situation table:** Each row (instance) corresponds to a resource.
 - **Event-pair-based situation table:** Each row (instance) corresponds to a pair of events.
 - **Aggregate situation tables:** Each row (instance) corresponds to a combination of cases and/or events.



Creating Situation Tables Using PQL



Creating Situation Tables with PQL

The starting point for every data mining and machine learning technique are *data*. The input data used for many data mining and machine learning techniques typically have a tabular form. In the data table, each row is an instance and each column is a variable. Instances correspond to individuals, entities, cases, objects, or records. Variables are often referred to as attributes, features, or data elements. In this course, we briefly cover data science topics related to supervised learning (e.g., Decision Trees) and unsupervised learning (e.g., Clustering). RapidMiner is a tool which provides a rich body of data science methods and algorithms to select and apply to your data. We use RapidMiner to demonstrate how algorithms such as Decision tree mining and K-means clustering can be applied to real data.

The data describing a process come in the form of an event log. The event log is the starting point for process mining techniques and it contains information regarding activities, cases, resources, and all other entities or objects that are involved in a process. Using data mining and machine learning one can gain process insights that go beyond process discovery and conformance checking. For instance, we may want to analyze which case attributes influence the path cases take in a process, how resources hand over work to each other, whether busy resources cause delays in particular process parts, and so on. These questions can be translated into concrete data science problems and thus, the existing techniques can be applied to help answering them. That translation is, however, not trivial. Given a particular process related question, its transformation into a data science problem requires defining the instances and the variables for the problem at hand. This is not necessarily the event log, where typically instances refer to events and variables refer to event attributes. It is only with the proper input data that the output of any data mining and machine learning technique yields correct and valid insights into the data. This is where the Celonis Process Query Language (PQL) comes to help. Using PQL, we can use the provided event data to generate any kind of data table we need depending on the question at hand (see Figure 1).

In this course, we learn how for process mining tasks related to decision mining, performance analysis and organizational mining, we can generate the proper input data on which existing data mining and machine learning techniques can be applied. This input data tables are called *situation tables*. More specifically, in the course we learn how to generate five types of situation tables:

- *Case-based situation table*: Each row (instance) corresponds to a case.
- *Event-based situation table*: Each row (instance) corresponds to an event.
- *Resource-based situation table*: Each row (instance) corresponds to a resource.
- *Event-pair-based situation table*: Each row (instance) corresponds to a pair of events.
- *Aggregate situation table*: Each row (instance) corresponds to a combination of cases and/or events.

- Celonis uses a snowflake schema.
- Tables are being joined implicitly.
- Order of execution:
 - a) Joins and regular PQL functions (not aggregations). The common table is defined after joining the required tables.
 - b) Filters are applied (if there are filters defined).
 - c) Standard aggregations (AVG, COUNT, SUM, etc.).
- If the result is different from what you expect, it is probably due to the ordering.



Chair of Process
and Data Science

Case-based situation table

CASE	1	PRODUCT	ADDRESS	first resource	decision	Steps	Distinct reso...	total throughput time ...	
1	SAMSUNG Galaxy J5	Munich	Caleb	pay		6	5	239	
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	pay		6	5	201	
3	APPLE iPhone 5s 16 GB	New York	Sophia	cancel order		5	4	503	
4	MOTOROLA Moto E 4G	New York	Sophia	cancel order		5	5	498	
5	SAMSUNG Core Prime G361	Aachen	Isabella	pay		8	5	741	
6	SAMSUNG Galaxy S4	Munich	Emma	pay		6	5	406	
7	MOTOROLA Moto G	Amsterdam	Lucas	pay		7	6	598	
8	APPLE iPhone 6 16 GB	Amsterdam	Sophia	pay		6	5	209	
9	APPLE iPhone 5s 16 GB	Munich	Aiden	pay		6	4	412	
10	HUAWEI P8 Lite	Amsterdam	Speedy	pay		7	7	415	
11	MOTOROLA Moto G	Munich	Emma	pay		6	6	508	
12	APPLE iPhone 5s 16 GB	Aachen	Jacob	pay		6	6	480	
13	HUAWEI P8 Lite	Munich	Speedy	pay		6	5	409	
14	SAMSUNG Core Prime G361	Munich	Sophia	pay		6	6	331	
15	SAMSUNG Galaxy S4	Aachen	Sophia	pay		6	4	186	
16	SAMSUNG Galaxy S4	Aachen	Olivia	pay		7	7	378	
17	SAMSUNG Galaxy S4	Munich	Luke	pay		6	5	343	
18	SAMSUNG Galaxy S4	Munich	Lucas	pay		6	6	337	
19	SAMSUNG Galaxy S4	New York	Luke	cancel order		5	3	462	
20	APPLE iPhone 6s 64 GB	Munich	Sophia	pay		6	6	211	
21	APPLE iPhone 6s 64 GB	Aachen	Jacob	pay		6	6	356	
22	SAMSUNG Galaxy S4	Aachen	Aiden	pay		7	7	359	
23	MOTOROLA Moto G	New York	Emma	cancel order		5	4	448	
24	MOTOROLA Moto G	Aachen	Speedy	pay		6	5	1,363	
25	APPLE iPhone 5s 16 GB	Munich	Sophia	pay		6	4	177	
26	HUAWEI P8 Lite	Aachen	Zoe	cancel order		5	4	505	
27	MOTOROLA Moto G	Amsterdam	Aiden	pay		6	5	987	
28	SAMSUNG Galaxy S4	New York	Aiden	cancel order		5	4	505	

Case-based situation table

DIMENSIONS Custom dimension +

- CASE
- PRODUCT
- ADDRESS
- first resource
- decision
- Steps
- Distinct resources
- total throughput time in hours

"cases"."PRODUCT"

PU_FIRST ("cases", "events"."RESOURCE")

CASE WHEN PROCESS EQUALS 'pay' THEN 'pay' WHEN
PROCESS EQUALS 'cancel order' THEN 'cancel order' END

PU_COUNT ("cases", "events"."ACTIVITY")

PU_COUNT_DISTINCT ("cases", "events"."RESOURCE")

CALC_THROUGHPUT(ALL_OCCURRENCE['Process Start']
TO ALL_OCCURRENCE['Process End'],
REMAP_TIMESTAMPS("events"."END TIME", HOURS))

CASE	PRODUCT	ADDRESS	first resource	decision	Steps	Distinct reso	total throughput time ..
1	SAMSUNG Galaxy J5	Munich	Caleb	pay	6	5	239
2	APPLE iPhone 6s 64 GB	Amsterdam	Lucas	pay	6	5	201
3	APPLE iPhone 5s 16 GB	New York	Sophia	cancel order	5	4	503
4	APPLE iPhone 6s 16 GB	New York	Sophia	cancel order	5	4	500
5	SAMSUNG Core Prime G361	Aachen	Isabella	pay	6	5	243
6	SAMSUNG Galaxy S4	Munich	Emma	pay	6	5	241
7	MOTOROLA Moto G	Amsterdam	Sophia	pay	7	6	598
8	APPLE iPhone 6s 16 GB	Amsterdam	Sophia	pay	6	5	259
9	APPLE iPhone 5s 16 GB	Munich	Aiden	pay	6	4	412
10	HUAWEI P8 Lite	Amsterdam	Speedy	pay	7	7	415
11	MOTOROLA Moto G	Munich	Emma	pay	6	6	508
12	APPLE iPhone 5s 16 GB	Aachen	Jacob	pay	6	6	480
13	APPLE iPhone 6s 16 GB	Munich	Speedy	pay	6	5	409
14	SAMSUNG Core Prime G361	Munich	Sophia	pay	6	6	351
15	SAMSUNG Galaxy S4	Aachen	Sophia	pay	6	4	198
16	SAMSUNG Galaxy S4	Aachen	Olivia	pay	7	7	378
17	SAMSUNG Galaxy S4	Munich	Luke	pay	6	5	343
18	SAMSUNG Galaxy S4	Munich	Lucas	pay	6	5	337
19	SAMSUNG Galaxy S4	New York	Luke	cancel order	5	5	462
20	APPLE iPhone 6s 64 GB	Munich	Sophia	pay	6	6	211
21	APPLE iPhone 5s 16 GB	Aachen	Jacob	pay	6	6	356
22	APPLE iPhone 6s 16 GB	Aachen	Aiden	pay	7	7	359
23	MOTOROLA Moto G	New York	Emma	cancel order	5	4	448
24	MOTOROLA Moto G	Aachen	Speedy	pay	6	5	1363
25	APPLE iPhone 5s 16 GB	Munich	Sophia	pay	6	4	177
26	HUAWEI P8 Lite	Aachen	Zoe	cancel order	5	4	505
27	MOTOROLA Moto G	Amsterdam	Aiden	pay	6	5	987
28	SAMSUNG Galaxy S4	New York	Aiden	cancel order	5	4	505

Setting the threshold (50% fast, 50% slow)

CA...	PRODUCT	ADDRESS	first resource	decision	Steps	Distinct...	total throughp...	Speed
1	SAMSUNG Galaxy J5	Munich	Caleb	pay	6	5	239	fast
2	APPLE iPhone 6s 6...	Amsterdam	Lucas	pay	6	5	201	fast
3	APPLE iPhone 5s 1...	New York	Sophia	cancel order	5	4	503	slow
4	MOTOROLA Moto E...	New York	Sophia	cancel order	5	5	498	slow
5	SAMSUNG Core Pri...	Aachen	Isabella	pay	8	5	741	slow
6	SAMSUNG Galaxy S4	Munich	Emma	pay	6	5	406	slow
7	MOTOROLA Moto G	Amsterdam	Lucas	pay	7	6	598	slow
8	APPLE iPhone 6 16 ...	Amsterdam	Sophia	pay	6	5	209	fast
9	APPLE iPhone 6s 1...	Munich	Aiden	pay	6	4	412	slow
10	HUAWEI P8 Lite	Amsterdam	Speedy	pay	7	7	415	slow
11	MOTOROLA Moto G	Munich	Emma	pay	6	6	508	slow
12	APPLE iPhone 5s 1...	Aachen	Jacob	pay	6	6	480	slow
13	HUAWEI P8 Lite	Munich	Speedy	pay	6	5	409	slow
14	SAMSUNG Core Pri...	Munich	Sophia	pay	6	6	331	fast
15	SAMSUNG Galaxy S4	Aachen	Sophia	pay	6	4	186	fast
16	SAMSUNG Galaxy S4	Aachen	Olivia	pay	7	7	378	fast
17	SAMSUNG Galaxy S4	Munich	Luke	pay	6	5	343	fast
18	SAMSUNG Galaxy S4	Munich	Lucas	pay	6	6	337	fast
19	SAMSUNG Galaxy S4	New York	Luke	cancel order	5	3	462	slow
20	APPLE iPhone 6s 6...	Munich	Sophia	pay	6	6	211	fast
21	APPLE iPhone 6s 6...	Aachen	Jacob	pay	6	6	356	fast
22	SAMSUNG Galaxy S4	Aachen	Aiden	pay	7	7	359	fast
23	MOTOROLA Moto G	New York	Emma	cancel order	5	4	448	slow
24	MOTOROLA Moto G	Aachen	Speedy	pay	6	5	1,363	slow
25	APPLE iPhone 5s 1...	Munich	Sophia	pay	6	4	177	fast
26	HUAWEI P8 Lite	Aachen	Zoe	cancel order	5	4	505	slow
27	MOTOROLA Moto G	Amsterdam	Aiden	pay	6	5	987	slow
28	SAMSUNG Galaxy S4	New York	Aiden	cancel order	5	4	505	slow

50% Quantile	
454	

```

1 ✓ QUANTILE (
2 ✓ CALC_THROUGHPUT (
3   ALL_OCCURRENCE [ 'Process Start' ]
4   TO
5   ALL_OCCURRENCE [ 'Process End' ],
6   REMAP_TIMESTAMPS ( "events"."START TIME" , HOURS )
7   ) ,
8   0.5
9 )

```

```

1 ✓ CASE
2 ✓ WHEN
3 ✓ CALC_THROUGHPUT (
4   ALL_OCCURRENCE [ 'Process Start' ]
5   TO
6   ALL_OCCURRENCE [ 'Process End' ],
7   REMAP_TIMESTAMPS ( "events"."START TIME" , HOURS )
8   )
9   <
10   454
11 ✓ THEN
12   'fast'
13 ✓ ELSE
14   'slow'
15 END

```

Creating an event-pair-based situation table

- **Event-pair-based situation table:** Each row (instance) corresponds to a pair of events.
- Many ways to do this. Here we use both:
 - **REMAP_VALUES** (REMAP_VALUES allows you to map values of a column of type STRING, https://docs.celonis.com/en/remap_values.html)
 - **SOURCE – TARGET** (SOURCE and TARGET functions provide a way to combine values from two different rows of the activity table into the same row, e.g. for calculating the throughput time between consecutive events inside a case. <https://docs.celonis.com/en/source---target.html>)

REMAP_VALUES

- REMAP_VALUES can be used to add a column of type STRING with modified values.
- REMAP_VALUES (table.column, [old_value1, new_value1] , ... [, others_value]).
- Examples:
 - REMAP_VALUES ("events"."ACTIVITY", ['send reminder','PAY NOW'], ['pay','THANKS'])
 - REMAP_VALUES ("events"."ACTIVITY",['place order','START'],['pay','END'], NULL)
 - REMAP_VALUES ("events"."RESOURCE" ,['Caleb','male'], ['Emily','female'], ['Lily','female'], ['Lucas','male'], 'unknown')
 - REMAP_VALUES ("events"."RESOURCE", ['Caleb','Team A'], ['Emily','Team A'], ['Lily','Team B'], NULL)

If not explicitly mapped, the old value remains (without an explicit default value at the end) or is replaced by optional default value.



Example: REMAP_VALUES

1

2

3

4

CASE	ACTIVITY	RESOURCE	START TIME	Remap 1	Remap 2	Remap 3	Remap 4
1	place order	Caleb	Mon Jan 5 20...	place order	START	male	Team A
1	send invoice	Emily	Thu Jan 8 201...	send invoice	-	female	Team A
1	pay	Lily	Fri Jan 9 2015...	THANKS	END	female	Team B
1	prepare delivery	Lucas	Wed Jan 14 2...	prepare delivery	-	male	-
1	make delivery	Aubrey	Wed Jan 14 2...	make delivery	-	unknown	-
1	confirm payment	Lily	Thu Jan 15 20...	confirm payment	-	female	Team B
2	place order	Lucas	Mon Jan 5 20...	place order	START	male	-
2	pay	Lily	Fri Jan 9 2015...	THANKS	END	female	Team B
2	send invoice	Jack	Mon Jan 12 2...	send invoice	-	unknown	-
2	prepare delivery	Aiden	Tue Jan 13 20...	prepare delivery	-	unknown	-
2	confirm payment	Lily	Tue Jan 13 20...	confirm payment	-	female	Team B
2	make delivery	Michael	Tue Jan 13 20...	make delivery	-	unknown	-
3	place order	Sophia	Mon Jan 5 20...	place order	START	unknown	-
3	send invoice	Lily	Fri Jan 9 2015...	send invoice	-	female	Team B
3	send reminder	Abigail	Fri Jan 16 201...	PAY NOW	-	unknown	-
3	send reminder	Abigail	Fri Jan 23 201...	PAY NOW	-	unknown	-

1. REMAP_VALUES ("events"."ACTIVITY", ['send reminder','PAY NOW'], ['pay','THANKS'])
2. REMAP_VALUES ("events"."ACTIVITY",['place order','START'],['pay','END'], NULL)
3. REMAP_VALUES ("events"."RESOURCE",['Caleb','male'], ['Emily','female'], ['Lily','female'], ['Lucas','male'], 'unknown')
4. REMAP_VALUES ("events"."RESOURCE",['Caleb','Team A'], ['Emily','Team A'], ['Lily','Team B'], NULL)

SOURCE - TARGET

- You can create one row for pairs of events using:
 - SOURCE ("events"."ACTIVITY")
 - TARGET ("events"."ACTIVITY")
- And add computations such as
 - HOURS_BETWEEN(SOURCE("events"."START TIME"),TARGET("events"."START TIME"))
- SOURCE and TARGET belong together and have many configuration options (see <https://docs.celonis.com/en/source---target.html>)
- A temporary table is created!

source	target	duration
place order	send invoice	74.77
send invoice	pay	30.28
pay	prepare delivery	116.03
prepare delivery	make delivery	4.01
make delivery	confirm payment	14.60
place order	pay	97.83
pay	send invoice	71.48
send invoice	prepare delivery	23.54
prepare delivery	confirm payment	5.71
confirm payment	make delivery	2.85
place order	send invoice	94.10
send invoice	send reminder	169.16
send reminder	send reminder	171.25
send reminder	cancel order	68.01
place order	send invoice	47.62

Common Table: <Temporary table: SOUR...



Example: SOURCE - TARGET

1 2 3 4 5 6 7 8

source act	target act	source res	target res	source case	target case	duration hours	duration days
place order	send invoice	Caleb	Emily	1	1	74.77	3.12
send invoice	pay	Emily	Lily	1	1	30.28	1.26
pay	prepare delivery	Lily	Lucas	1	1	116.03	4.83
prepare delivery	make delivery	Lucas	Aubrey	1	1	4.01	0.17
make delivery	confirm payment	Aubrey	Lily	1	1	14.60	0.61
place order	pay	Lucas	Lily	2	2	97.83	4.08
pay	send invoice	Lily	Jack	2	2	71.48	2.98
send invoice	prepare delivery	Jack	Aiden	2	2	23.54	0.98
prepare delivery	confirm payment	Aiden	Lily	2	2	5.71	0.24
confirm payment	make delivery	Lily	Michael				
place order	send invoice	Sophia	Lily				
send invoice	send reminder	Lily	Abigail				
send reminder	send reminder	Abigail	Abigail				
send reminder	cancel order	Abigail	Swift				
place order	send invoice	Sophia	Alexander	4	4	47.62	1.98

1. SOURCE ("events"."ACTIVITY")
2. TARGET ("events"."ACTIVITY")
3. SOURCE ("events"."RESOURCE")
4. TARGET ("events"."RESOURCE")



Example: SOURCE - TARGET

1 2 3 4 5 6 7 8

source act	target act	source res	target res	source case	target case	duration hours	duration days
place order	send invoice	Caleb	Emily	1	1	74.77	3.12
send invoice	pay	Emily	Lily	1	1	30.28	1.26
pay	prepare delivery	Lily	Lucas	1	1	116.03	4.83
prepare delivery	make delivery	Lucas	Aubrey	1	1	4.01	0.17
make delivery	confirm payment	Aubrey	Lily	1	1	14.60	0.61
place order	pay	Lucas	Lily	2	2	97.83	4.08
pay	send invoice	Lily	Jack	2	2	71.48	2.98
send invoice	prepare delivery	Jack	Aiden	2	2	23.54	0.98
prepare delivery	confirm payment	Aiden	Lily	2	2	5.71	0.24
confirm payment	make delivery	Lily	Michael	2	2	2.85	0.12
place order	send invoice	Sophia	Lily	3	3	94.10	3.92
send invoice	send reminder	Lily	Abigail	3	3	169.16	7.05
send reminder	send reminder	Abigail	Abigail	3	3	171.25	7.14
							2.83
							1.98

5. SOURCE ("events"."CASE")
6. TARGET ("events"."CASE")
7. HOURS_BETWEEN (SOURCE ("events"."START TIME") , TARGET ("events"."START TIME"))
8. DAYS_BETWEEN (SOURCE ("events"."START TIME") , TARGET ("events"."START TIME"))



How to measure the time between two specific activities in the process?

- Create an **event-pair-based situation table** using **REMAP_VALUES**, **SOURCE**, **TARGET**.
- The **SOURCE** operator may use an **activity_table.filter_column**, i.e., an **optional filter column to skip certain events**.
- Events with a **NULL** value in the related entry of the **filter column** **are ignored!** (**Important to understand!**)
- Usually, the filter column is created using **REMAP_VALUES** setting values to **NULL**.



Two approaches

1. Explicitly map activities to NULL to remove them.
2. Map all activities to NULL except a few selected ones (value does not matter).

1

```
SOURCE ( "events"."ACTIVITY" ,  
REMAP_VALUES ("events"."ACTIVITY" ,  
[ 'send invoice' , NULL ] ,  
[ 'send reminder' , NULL ] ,  
[ 'prepare delivery' , NULL ] ,  
[ 'make delivery' , NULL ] ,  
[ 'confirm payment' , NULL ] ) )
```

2

```
SOURCE ( "events"."ACTIVITY" ,  
REMAP_VALUES ( "events"."ACTIVITY" ,  
[ 'place order' , '' ] ,  
[ 'pay' , '' ] ,  
[ 'cancel order' , '' ] ,  
NULL ) )
```

Small print: SOURCE and TARGET belong together. Not really specific for SOURCE, it is applied to all events before. Can also be applied to resource or other attributes.



Chair of Process
and Data Science

Same result

source	target	duration
place order	pay	105.05
place order	pay	97.83
place order	cancel order	502.52
place order	cancel order	498.26
place order	pay	577.36
place order	pay	361.49
place order	pay	333.97
place order	pay	157.62
place order	pay	202.10
place order	pay	334.90
place order	pay	199.52
place order	pay	358.70
place order	TARGET ("events"."ACTIVITY")	HOURS_BETWEEN(SOURCE("events"."START TIME"),TARGET("events"."START TIME"))
place order	pay	48.49

1

```
SOURCE ( "events"."ACTIVITY" ,
REMAP_VALUES ( "events"."ACTIVITY" ,
[ 'send invoice' , NULL ] ,
[ 'send reminder' , NULL ] ,
[ 'prepare delivery' , NULL ] ,
[ 'make delivery' , NULL ] ,
[ 'confirm payment' , NULL ] ))
```

Specify what needs to be removed.

source	target	duration
place order	pay	105.05
place order	pay	97.83
place order	cancel order	502.52
place order	cancel order	498.26
place order	pay	577.36
place order	pay	361.49
place order	pay	333.97
place order	pay	157.62
place order	pay	202.10
place order	pay	334.90
place order	pay	199.52
place order	pay	358.70
place order	TARGET ("events"."ACTIVITY")	HOURS_BETWEEN(SOURCE("events"."START TIME"),TARGET("events"."START TIME"))
place order	pay	48.49

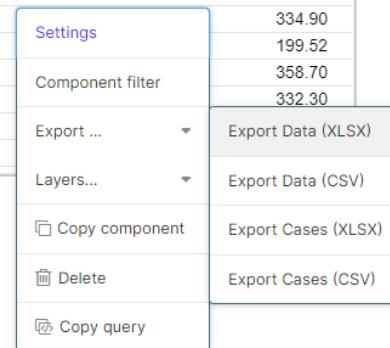
2

```
SOURCE ( "events"."ACTIVITY" ,
REMAP_VALUES ( "events"."ACTIVITY" ,
[ 'place order' , '' ] ,
[ 'pay' , '' ] ,
[ 'cancel order' , '' ] ,
NULL ))
```

Specify what needs to stay.

Event-pair-based situation table can be exported and analyzed

source act	target act	source res	target res	duration
place order	pay	Caleb	Lily	105.05
place order	pay	Lucas	Lily	97.83
place order	cancel order	Sophia	Swift	502.52
place order	cancel order	Sophia	Lily	498.26
place order	pay	Isabella	Lily	577.36
place order	pay	Emma	Lily	361.49
place order	pay	Lucas	Jack	333.97
place order	pay	Sophia	Swift	157.62
place order	pay	Aiden	Jack	202.10
place order	pay	Speedy	Emily	334.90
place order	pay	Emma	Jack	199.52
place order	pay	Jacob	Jack	358.70
place order	pay	Speedy	Lily	332.30
place order	pay	Sophia	Caleb	
place order	pay	Sophia	Lily	



	A	B	C	D	E
1	source act	target act	source res	target res	duration
2	place order	pay	Caleb	Lily	105.051111
3	place order	pay	Lucas	Lily	97.8277778
4	place order	cancel order	Sophia	Swift	502.5213889
5	place order	cancel order	Sophia	Lily	498.2647222
6	place order	pay	Isabella	Lily	577.3558333
7	place order	pay	Emma	Lily	361.4863889
8	place order	pay	Lucas	Jack	333.9661111
9	place order	pay	Sophia	Swift	157.6213889
10	place order	pay	Aiden	Jack	202.0994444
11	place order	pay	Speedy	Emily	334.8986111
12	place order	pay	Emma	Jack	199.5180556
13	place order	pay	Jacob	Jack	358.7025
14	place order	pay	Speedy	Lily	332.2955556
15	place order	pay	Sophia	Caleb	140.2055556
16	place order	pay	Sophia	Lily	48.488050556



```
SOURCE (
  "events"."ACTIVITY",
  REMAP_VALUES (
    "events"."ACTIVITY",
    [ 'place order', '' ],
    [ 'pay', '' ],
    [ 'cancel order', '' ],
    NULL )
)
```

```
TARGET ( "events"."ACTIVITY" )
```

```
SOURCE ( "events"."RESOURCE" )
```

```
TARGET ( "events"."RESOURCE" )
```

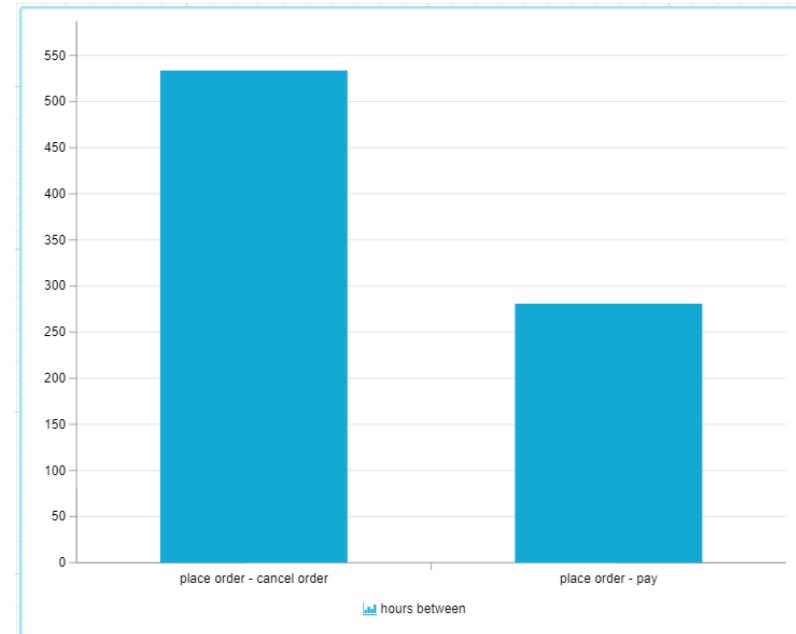
```
HOURS_BETWEEN(SOURCE("events"."START TIME"),TARGET("events"."START TIME"))
```

Computations inside Celonis

source	target	hours between
place order	pay	105
place order	pay	98
place order	cancel order	503
place order	cancel order	498
place order	pay	577
place order	pay	361
place order	pay	334
place order	pay	158
place order	pay	202
place order	pay	335
place order	pay	200
place order	pay	359
place order	pay	332
place order	pay	140
place order	pay	48
place order	pay	330
place order	pay	64
	pay	176
	cancel order	462
	TARGET ("events"."ACTIVITY")	151
	pay	190
	pay	440
	cancel order	

SOURCE ("events"."ACTIVITY" ,
 REMAP_VALUES ("events"."ACTIVITY" ,
 ['place order' , '] ,
 ['pay' , '] ,
 ['cancel order' , '] ,
 NULL))
 place order

HOURS_BETWEEN(SOURCE("events"."START TIME"),TARGET("events"."START TIME"))



AVG(HOURS_BETWEEN (SOURCE("events"."START TIME"),TARGET("events"."START TIME"))))



Computations inside Celonis

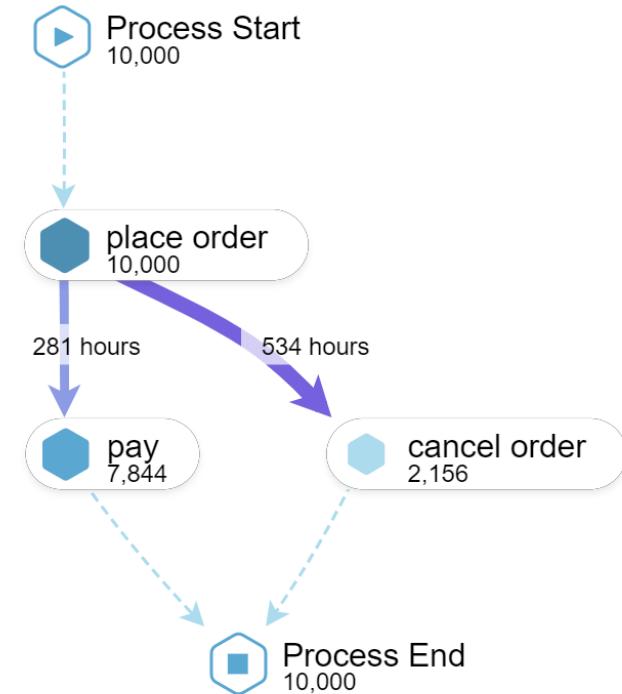
source	target	average duration
place order	cancel order	533.65
place order	pay	280.87

```
SOURCE ("events"."ACTIVITY",  
REMAP_VALUES (  
"events"."ACTIVITY",  
[ 'place order' , " ] ,  
[ 'pay' , " ] ,  
[ 'cancel order' , " ] ,  
NULL ))
```

TARGET ("events".
"ACTIVITY")

AVG (HOURS_BETWEEN (
SOURCE("events"."START TIME"),
TARGET("events"."START TIME")))

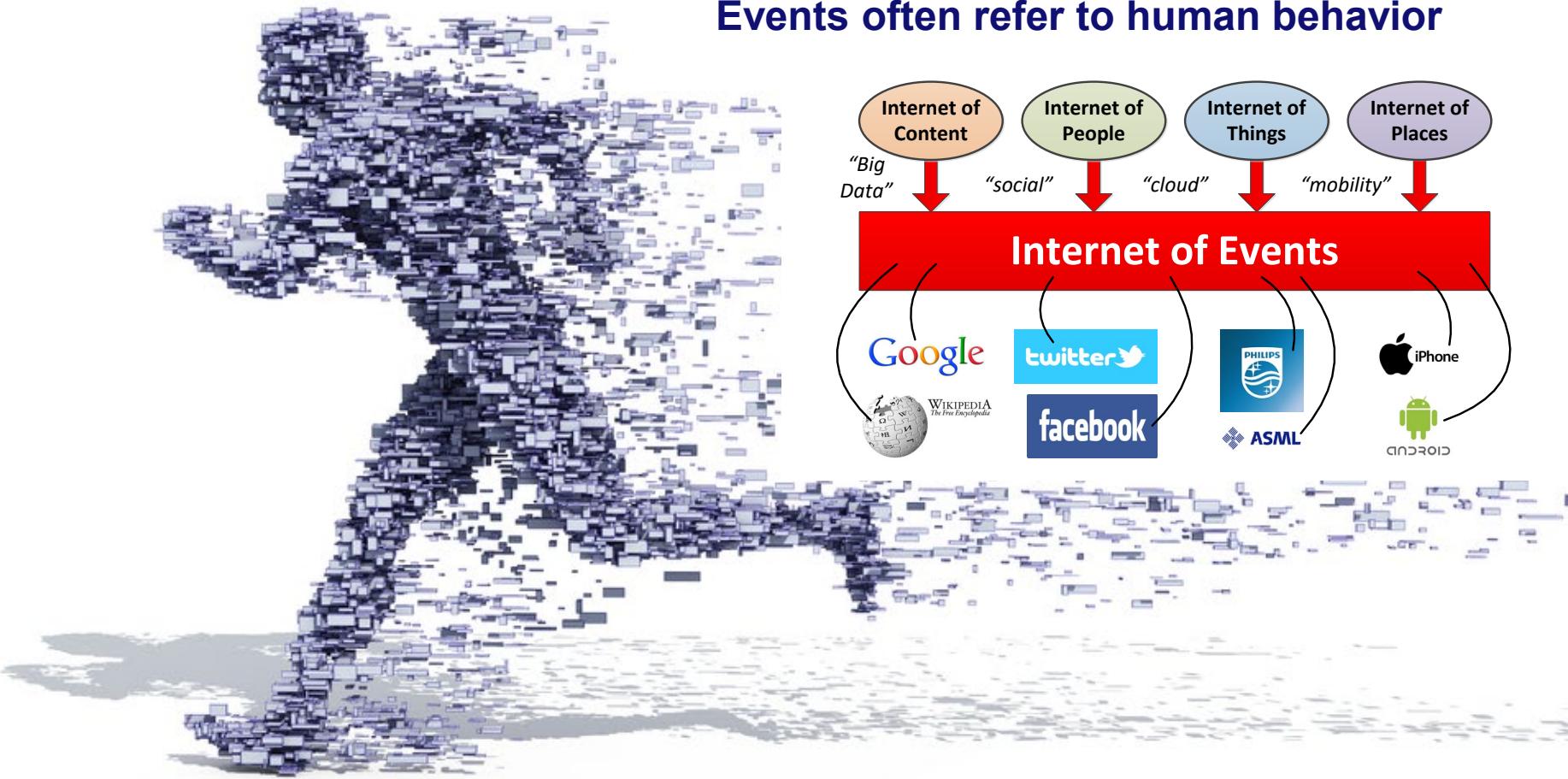
KPI



Mining Social Networks



Events often refer to human behavior



Events having a resource attribute

case id trace

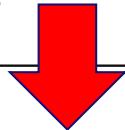
- | | |
|-----|---|
| 1 | $\langle a^{Pete}, b^{Sue}, d^{Mike}, e^{Sara}, h^{Pete} \rangle$ |
| 2 | $\langle a^{Mike}, d^{Mike}, c^{Pete}, e^{Sara}, g^{Ellen} \rangle$ |
| 3 | $\langle a^{Pete}, c^{Mike}, d^{Ellen}, e^{Sara}, f^{Sara}, b^{Sean}, d^{Pete}, e^{Sara}, g^{Ellen} \rangle$ |
| 4 | $\langle a^{Pete}, d^{Mike}, b^{Sean}, e^{Sara}, h^{Ellen} \rangle$ |
| 5 | $\langle a^{Ellen}, c^{Mike}, d^{Pete}, e^{Sara}, f^{Sara}, d^{Ellen}, c^{Mike}, e^{Sara}, f^{Sara}, b^{Sue}, d^{Pete}, e^{Sara}, h^{Mike} \rangle$ |
| 6 | $\langle a^{Mike}, c^{Ellen}, d^{Mike}, e^{Sara}, g^{Mike} \rangle$ |
| ... | ... |
-

($a = \text{register request}$, $b = \text{examine thoroughly}$, $c = \text{examine casually}$, $d = \text{check ticket}$, $e = \text{decide}$, $f = \text{reinitiate request}$, $g = \text{pay compensation}$, and $h = \text{reject request}$)

Resource-activity matrix

(mean number of times a resource performs an activity per case)

case id	trace
1	$\langle a^{Pete}, b^{Sue}, d^{Mike}, e^{Sara}, h^{Pete} \rangle$
2	$\langle a^{Mike}, d^{Mike}, c^{Pete}, e^{Sara}, g^{Ellen} \rangle$
3	$\langle a^{Pete}, c^{Mike}, d^{Ellen}, e^{Sara}, f^{Sara}, b^{Sean}, d^{Pete}, e^{Sara}, g^{Ellen} \rangle$
4	$\langle a^{Pete}, d^{Mike}, b^{Sean}, e^{Sara}, h^{Ellen} \rangle$
5	$\langle a^{Ellen}, c^{Mike}, d^{Pete}, e^{Sara}, f^{Sara}, d^{Ellen}, c^{Mike}, e^{Sara}, f^{Sara}, b^{Sue}, d^{Pete}, e^{Sara}, h^{Mike} \rangle$
6	$\langle a^{Mike}, c^{Ellen}, d^{Mike}, e^{Sara}, g^{Mike} \rangle$
...	...



Note small error in example (Pete and Mike are swapped), concept should be clear.

	a	b	c	d	e	f	g	h
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0

Resource-activity matrix

(mean number of times a resource performs an activity per case)

In 30% of the cases, a is executed by Pete, 50% is executed by Mike, and 20% is executed by Ellen.

Activities e and f matrix provides basic "who is doing what".

Pete

a

0.3

b

0

c

0.345

d

0.69

e

0

f

0

g

0.135

h

0.165

Mike

a

0.5

b

0

c

0.575

d

0.15

e

0

f

0

g

0.225

h

0.275

Ellen

a

0.2

b

0

c

0.23

d

0.11

e

0

f

0

g

0.09

h

0.11

Sue

a

0

b

0.46

c

0

d

0

e

0

f

0

g

0

h

0

Sean

a

0

b

0.69

c

0

d

0

e

0

f

0

g

0

h

0

Sara

a

0

b

0

c

0

d

0

e

2.3

f

1.3

g

0

h

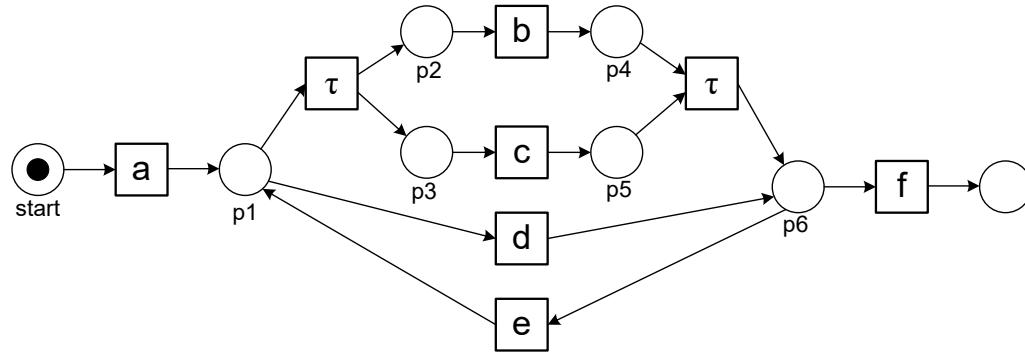
0

Question: Create resource activity matrix

case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete
1	c	complete	18	Sue
3	a	start	20	Pete
2	b	start	22	Mary
2	b	complete	25	Mary
3	a	complete	28	Pete

Question: Create resource activity matrix

Process context:



may take some time ...

case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete
1	c	complete	18	Sue
3	a	start	20	Pete
2	b	start	22	Mary
2	b	complete	25	Mary
3	a	complete	28	Pete
1	b	start	30	Mary
1	b	complete	34	Mary
3	d	start	35	Mary
3	d	complete	37	Mary
2	c	start	40	Sue
1	f	start	42	Carol
2	c	complete	45	Sue
1	f	complete	46	Carol
2	e	start	50	Kirsten
3	f	start	51	Carol
2	e	complete	52	Kirsten
2	d	start	53	Mary
3	f	complete	55	Carol
2	d	complete	56	Mary
2	f	start	57	Carol
2	f	complete	60	Carol



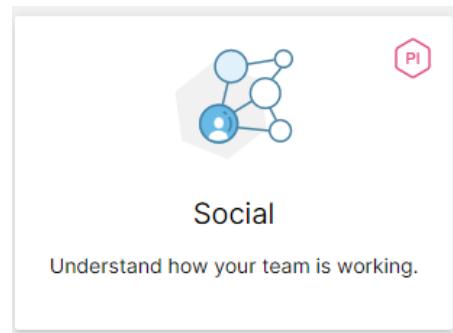
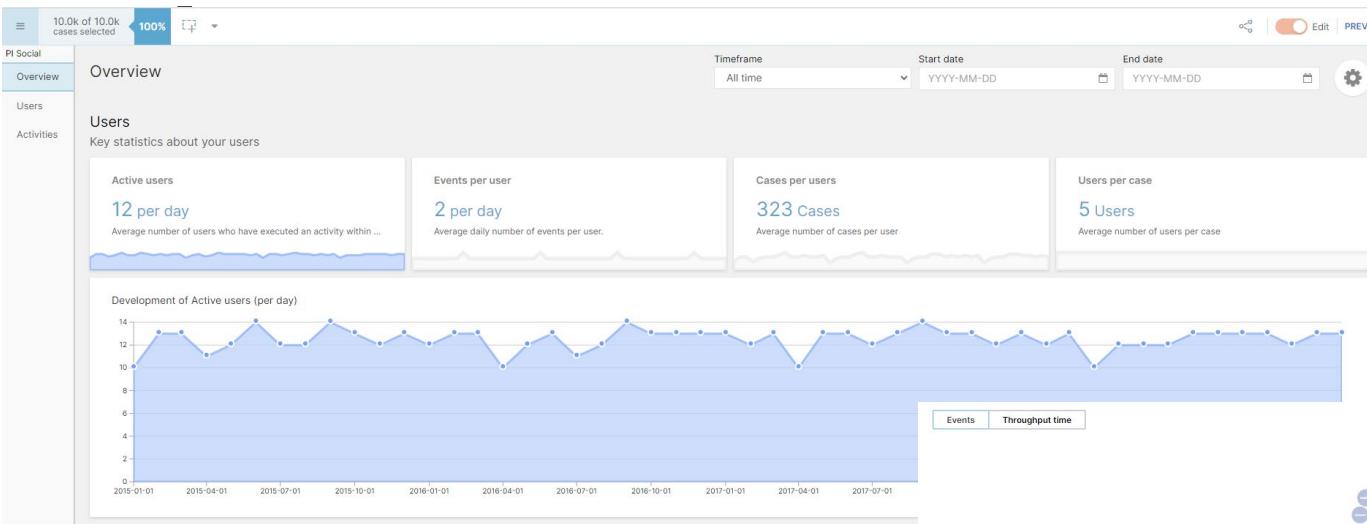
Answer: Resource activity matrix

mean number of times a resource performs an activity per case

	a	b	c	d	e	f
Pete	1.00	0.00	0.00	0.00	0.00	0.00
Mary	0.00	0.67	0.00	0.67	0.00	0.00
Sue	0.00	0.00	0.67	0.00	0.00	0.00
Kirsten	0.00	0.00	0.00	0.00	0.33	0.00
Carol	0.00	0.00	0.00	0.00	0.00	1.00

case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete
1	c	complete	18	Sue
3	a	start	20	Pete
2	b	start	22	Mary
2	b	complete	25	Mary
3	a	complete	28	Pete
1	b	start	30	Mary
1	b	complete	34	Mary
3	d	start	35	Mary
3	d	complete	37	Mary
2	c	start	40	Sue
1	f	start	42	Carol
2	c	complete	45	Sue
1	f	complete	46	Carol
2	e	start	50	Kirsten
3	f	start	51	Carol
2	e	complete	52	Kirsten
2	d	start	53	Mary
3	f	complete	55	Carol
2	d	complete	56	Mary
2	f	start	57	Carol
2	f	complete	60	Carol

Some predefined views in Celonis



Some predefined views in Celonis

PI Social Overview 10.0k of 10.0k cases selected 100%  PRE

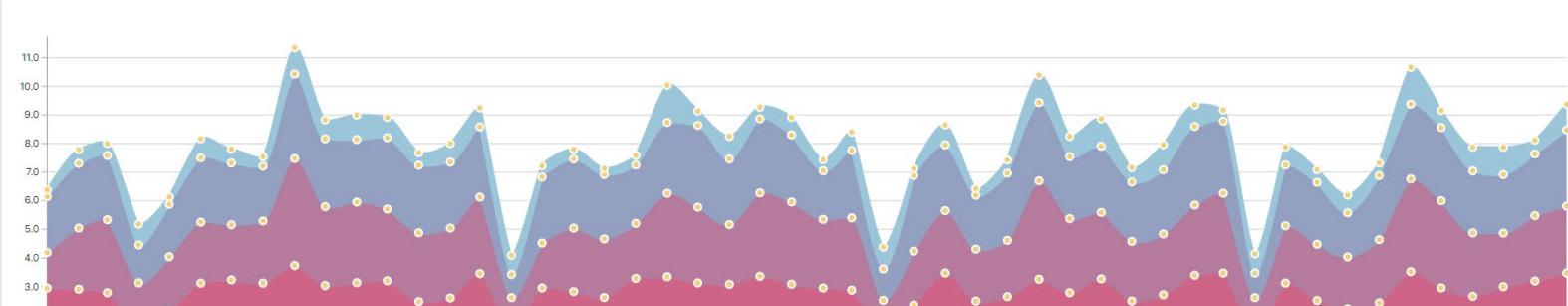
Select user(s) (31)
Search or select a user from the chart to view their tasks and performance

Lily

Performance of *Lily*
[Ignore user](#) [View cases in...](#)

Events 8 per day Average number of events per day.	Activities 3 per day Average number of different activities...	Throughput time 480.1 Hours Average end to end throughput time ...	Last active -2y Last active on 2019-12-30 14:50	Cases come from Sophia, Abigail and Aiden In 38%	Cases go to Abigail, Sophia and Luke In 41%
---	---	---	--	---	--

Lily's most frequently performed activities



Creating a resource activity matrix (OLAP table)

RESOURCE	ACTIVITY	Activities count
Abigail	make delivery	592
Abigail	send reminder	4,347
Aiden	place order	2,087
Aiden	prepare delivery	1,659
Alexander	cancel order	120
Alexander	confirm payment	397
Alexander	pay	417
Alexander	send invoice	557
Aubrey	make delivery	1,584
Avery	make delivery	769
Caleb	cancel order	68
Caleb	confirm payment	244
Caleb	pay	258
Caleb	place order	180
Caleb	prepare delivery	130
Caleb	send invoice	327

"events"."RESOURCE"

"events"."ACTIVITY"

COUNT("events"."ACTIVITY")

Component type OLAP Table

DIMENSIONS Add

RESOURCES

ACTIVITIES

KPIs Add

Activities count

Not normalized: Divide by the number of cases (here 10.000) to get the mean number of times a resource performs an activity per case.



Chair of Process
and Data Science

Creating a resource activity matrix (Pivot table)

Pivot: ACTIVITY KPI: Activities count
Rows: RESOURCE

	cancel order	confirm payment	make delivery	pay	place order	prepare delivery	send invoice	send reminder	
Abigail				592					4,347
Aiden						2,087	1,659		
Alexander	120	397			417			557	
Aubrey				1,584					
Avery				769					
Caleb	68	244			258	180	130	327	
Charlotte	11	30			40			38	
Chloe	10	56			53			71	
Ella				1,311					
Emily	195	649			711			817	
Emma						1,013	809		
Harper				321					
Isabella						127	104		
Jack	494	1,959			1,888			2,384	
Jacob						256	180		
James	11	46			39			68	
Kaylee				545					
Layla				431					
Lily	849	3,186			3,055			3,988	

Configure

Table title

DIMENSIONS

- ACTIVITY
- RESOURCE

KPIs

Add

Activities count

Pivot Dimension

Selected KPI

ADVANCED OPTIONS

- Show KPI summary
- Component is not filtered with selections
- Disable Selections

?

"events"."RESOURCE"

"events"."ACTIVITY"

COUNT("events"."ACTIVITY")



Chair of Process
and Data Science

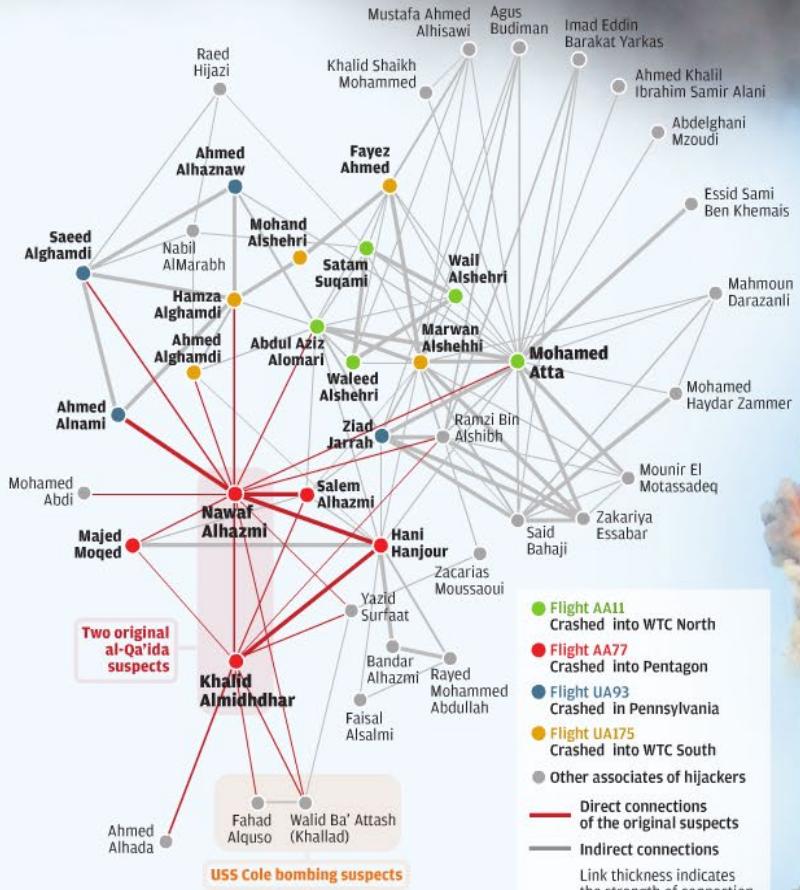
social networks

How two names and a sheaf of newspaper cuttings revealed the 9/11 team

This social network of the 19 hijackers behind the 9/11 attacks in the United States, and their associates, was drawn up at the end of 2001. Valdis Krebs, a commercial consultant in network analysis, started with newspaper reports of the two original terrorist suspects, Nawaf Alhazmi and Khalid Almihdhar. He then plotted the position of the other hijackers and associates. His analysis highlighted the central role played by Mohamed Atta. It also shows the close associations between the “Hamburg cell” that Atta set up, as well as the close links with the two original suspects – critical information that may have helped to avert an attack had it been known.

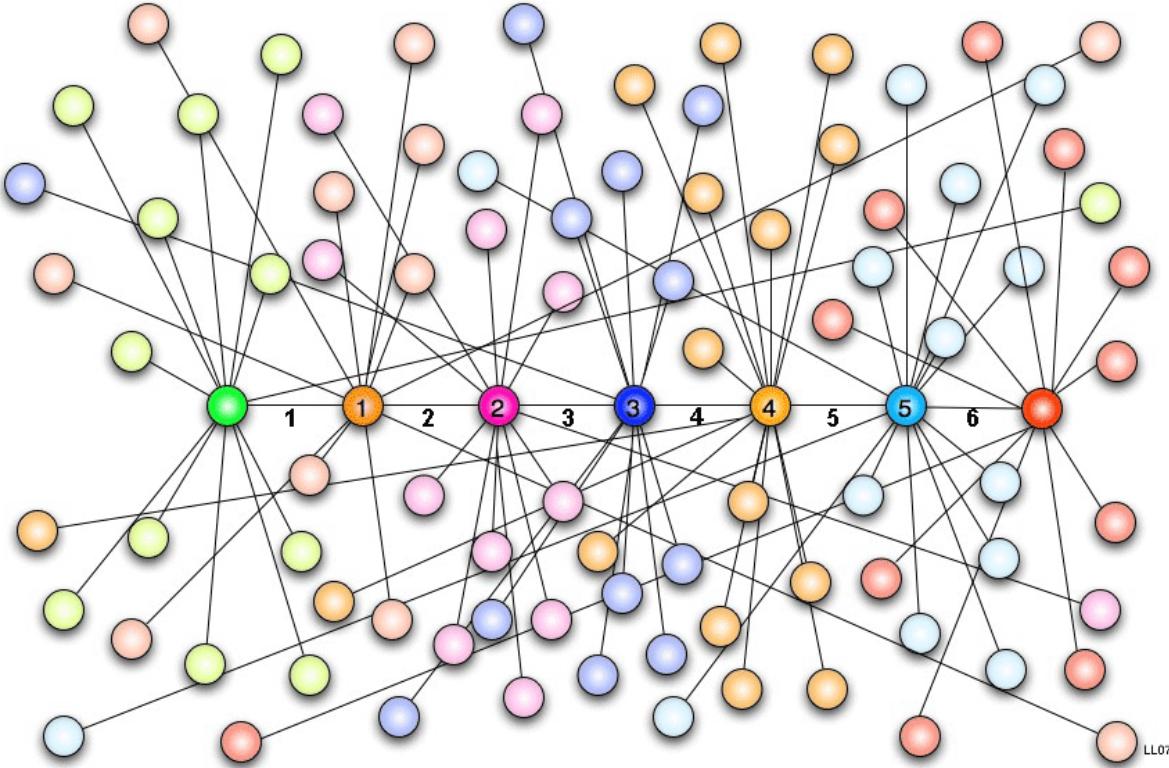


Emergency services attend the scene after Flight AA77 crashes into the Pentagon



The second plane, Flight UA175, crashes into the South Tower of the World Trade Centre

Six degrees of separation (e.g. facebook links)



A chain of "a friend of a friend" links connect any two persons in a maximum of six steps.

Posed by Frigyes Karinthy in 1929 and popularized through John Guare's "Six Degrees of Separation" (play/movie) and the "Six Degrees of Kevin Bacon" game.

FROM JORDAN TO THE QUEEN (AND OTHER CLOSE ENCOUNTERS)



Jordan the model had a year-long affair with...

Dane Bowers the singer who dueted with...

Victoria Beckham who is good friends with...

Elton John who went to the stag do of his chum...

Prince Andrew whose mother is, of course...

The Queen



Prince Charles was enchanted to meet...

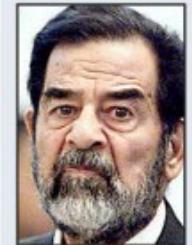
Beyoncé who was in Austin Powers 3 with...

Mike Myers who also co-starred with...

Liz Hurley who hung out at the Oscars with...

Pamela Anderson, who posed for...

Hugh Hefner, in his magazine Playboy



Saddam Hussein met in Baghdad in 1990 with...

Sir Edward Heath who sat on Tory benches with...

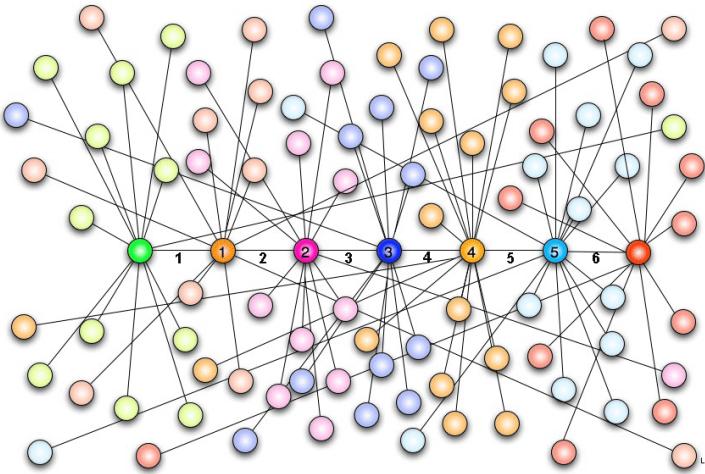
Liam Fox MP who was once close to...

Natalie Imbruglia who starred in Neighbours with...

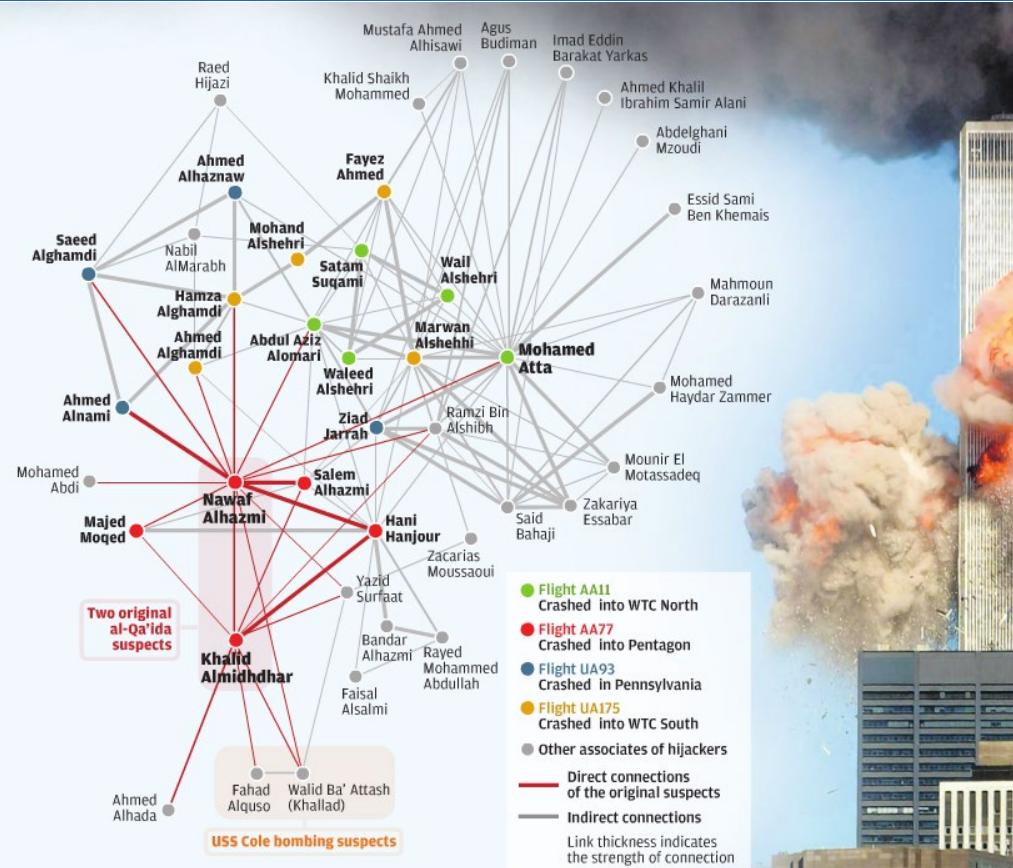
Anne Charleston, who had earlier worked with...

Kylie Minogue

Paths in social networks

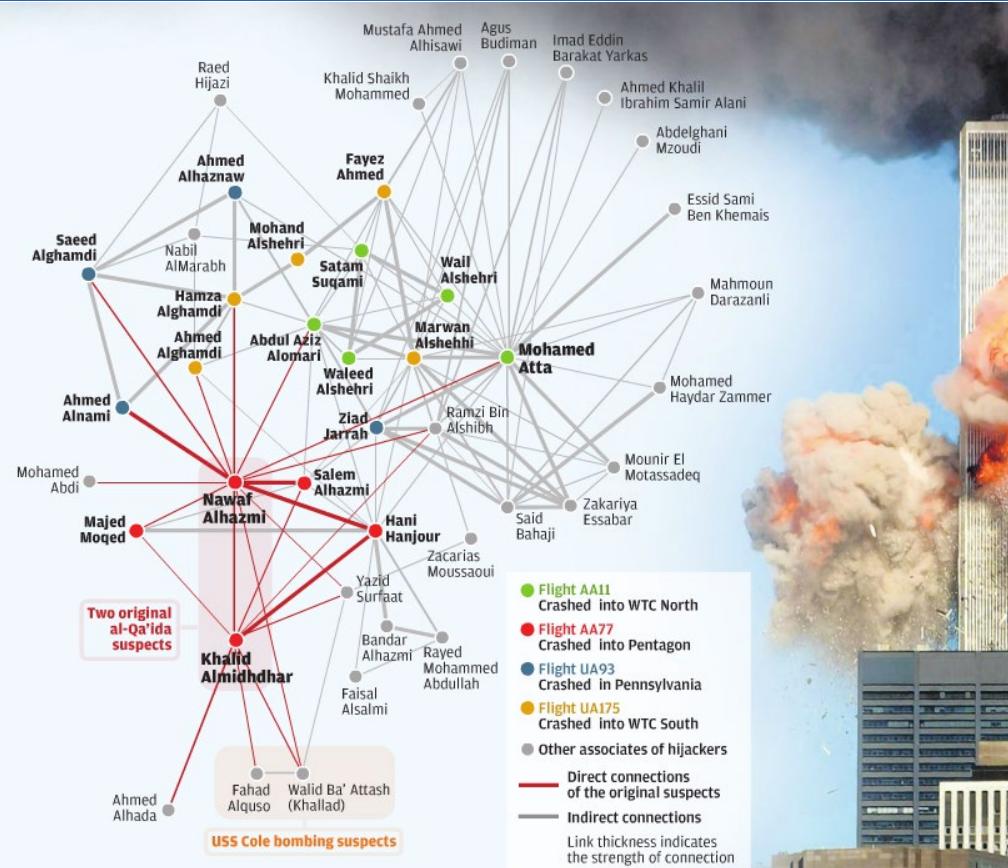


Social network analysis



- **Sociometry:** present data on interpersonal relationships in graph or matrix form.
- **Jacob Levy Moreno** used such techniques in the 1930s to better assign students to residential cottages.

Social network analysis



- Arcs: weights or (inverted) distance.
- Metrics to denote importance:
 - centrality,
 - closeness,
 - betweenness.
 - ...
- Identification of cliques.

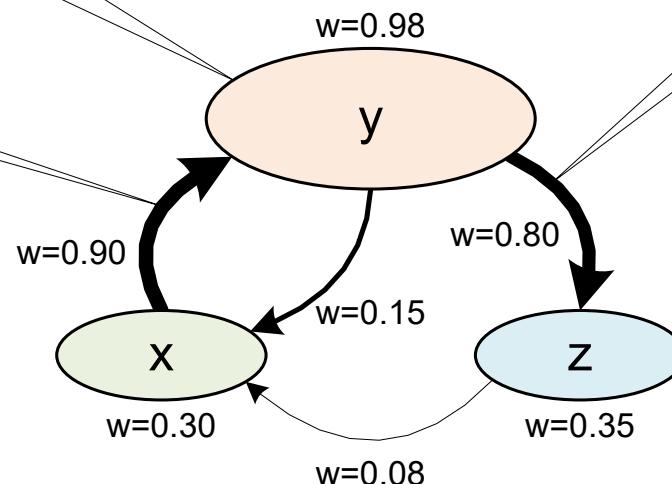
Social network

organizational entity (resource, person, role, department, etc.)

the thickness of the arc indicates the weight of the relationship

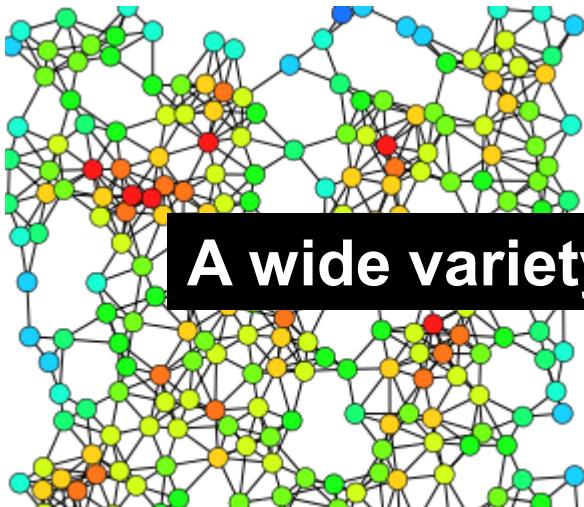
relationship

the size of the oval indicates the weight of the entity

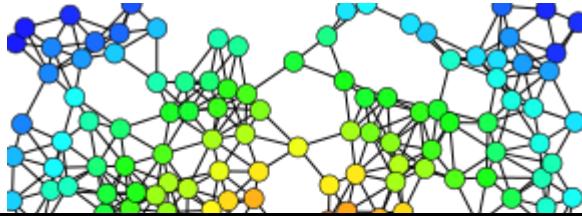


Importance of nodes in a social network

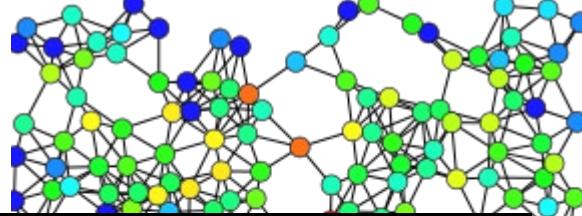
(figures by Claudio Rocchini, cc BY-SA 3.0)



degree centrality:
number of connections
a particular node has



closeness centrality:
1 divided by the sum of
all shortest paths to a
particular node



betweenness centrality:
fraction of shortest paths
between any two nodes
passing a particular node

Handover of work matrix

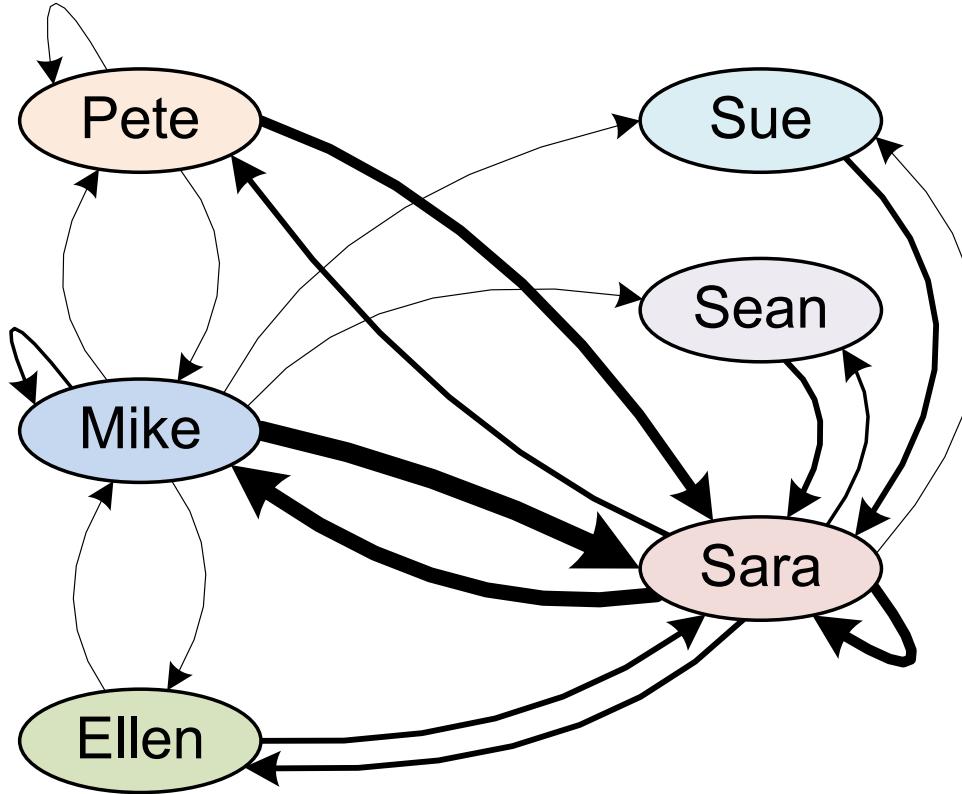
case id t
1
2
3
4
5
6
...

On average Pete hands over work to himself 0.135 times per case.

On average Sara hands over work to Mike 1.475 times per case.

	Pete	Mike	Ellen	Sue	Sean	Sara
Pete	0.135	0.225	0	0.06	0.09	1.035
Mike	0.225	0.375	0	0.1	0.15	1.725
Ellen	0.09	0.15	0.06	0.04	0.06	0.69
Sue	0	0	0	0	0	0.46
Sean	0	0	0	0	0	0.69
Sara	0.885	1.475	0.59	0.26	0.39	1.3

Social network based on handover of work (threshold of 0.1)



In this figure only the thickness of the arcs is based on frequencies. All nodes have the same size.

	Pete	Mike	Ellen	Sue	Sean	Sara
Pete	0.135	0.225	0.09	0.06	0.09	1.035
Mike	0.225	0.375	0.15	0.1	0.15	1.725
Ellen	0.09	0.15	0.06	0.04	0.06	0.69
Sue	0	0	0	0	0	0.46
Sean	0	0	0	0	0	0.69
Sara	0.885	1.475	0.59	0.26	0.39	1.3

Handover of work matrix in Celonis

(use again the SOURCE –TARGET construct presented before)

source	target	Activities count	IF 1
Sophia	Lily	1,566	
Abigail	Lily	1,209	
Lily	Abigail	1,136	
Aiden	Lily	1,054	
Abigail	Abigail	1,047	
Luke	Lily	1,012	
Sophia	Jack	926	
Lily	Sophia	919	
Lily	Lily	897	
Lily	Luke	876	
Abigail	Luke	719	
Luke	Abigail	711	
Lucas	Lily	701	
Jack	Abigail	697	
Abigail	Jack	671	
Luke	Jack	658	
Aiden	.Jack	634	

SOURCE("events"."RESOURCE")

TARGET("events"."RESOURCE")

Component type OLAP Table

DIMENSIONS

source					
target					

KPIs

Activities count					
------------------	--	--	--	--	--

COUNT(SOURCE("events"."ACTIVITY"))

Not normalized: Divide by the number of cases (here 10.000) to get the mean number of times a resource performs an activity per case.



Chair of Process
and Data Science

Handover of work matrix in Celonis (Pivot Table)

Pivot: source	KPI:	Activities count	Rows: target	Configure
Abigail	1047	92	131	
Aiden		93	10	
Alexander	157	23	53	33
Aubrey	230	27	35	3
Avery	82	15	18	
Caleb	98	17	33	15
Charlotte		1	1	33
Chloe		2	2	3
Ella		10	10	143
Emily		228	42	4
Emma		27	60	15
Harper		44	7	76
Isabella			11	11
Jack			4	13
Jacob			10	10
James			12	12
Keylee			13	13
Layla			10	10
Lily			7	7
Lucas			1130	57
Luke			711	60
Madelyn			25	25
Madison			28	28
Mia			10	10
Michael			2	2
Olivia			17	17
Rush			30	30
Sophia			210	210
Speedy			45	45
Swift			78	78
Zoe			22	22
Abigail	1047	92	131	401
Aiden		93	10	258
Alexander	157	23	53	9
Aubrey	230	27	35	22
Avery	82	15	18	7
Caleb	98	17	33	8
Charlotte		1	1	42
Chloe		2	2	1
Ella		10	10	1
Emily		228	42	1
Emma		27	60	1
Harper		44	7	1
Isabella			11	1
Jack			13	1
Jacob			10	1
James			12	1
Keylee			13	1
Layla			10	1
Lily			7	1
Lucas			1130	1
Luke			711	1
Madelyn			25	1
Madison			28	1
Mia			10	1
Michael			2	1
Olivia			17	1
Rush			30	1
Sophia			210	1
Speedy			45	1
Swift			78	1
Zoe			22	1

DIMENSIONS

Add

source			
target			

KPIs

Add

Activities count			
------------------	--	--	--

SOURCE("events"."RESOURCE")

TARGET("events"."RESOURCE")

COUNT(SOURCE("events"."ACTIVITY"))

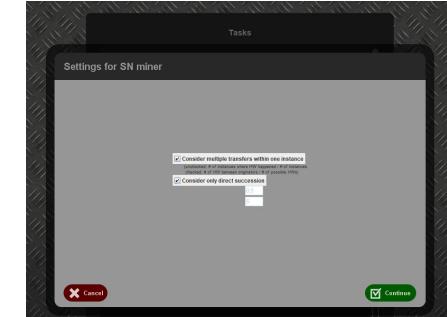
Not normalized: Divide by the number of cases (here 10.000) to get the mean number of times a resource performs an activity per case.

Pivot: source KPI: Activities count
Rows: target

Abigail	Aiden	Alexander
Abigail	1047	92
Aiden		93
Alexander	157	136
Aubrey		230
Avery		27
Caleb		84
Charlotte		15
Chloe		22
Ella		10
Emily		17
Emma		23
Harper		27
Isabella		30
Jack		33
Jacob		35
James		36
Keylee		37
Layla		38
Lily		39
Lucas		40
Luke		41
Madelyn		42
Madison		43
Mia		44
Michael		45
Olivia		46
Rush		47
Sophia		48
Speedy		49
Swift		50
Zoe		51

Many possible refinements

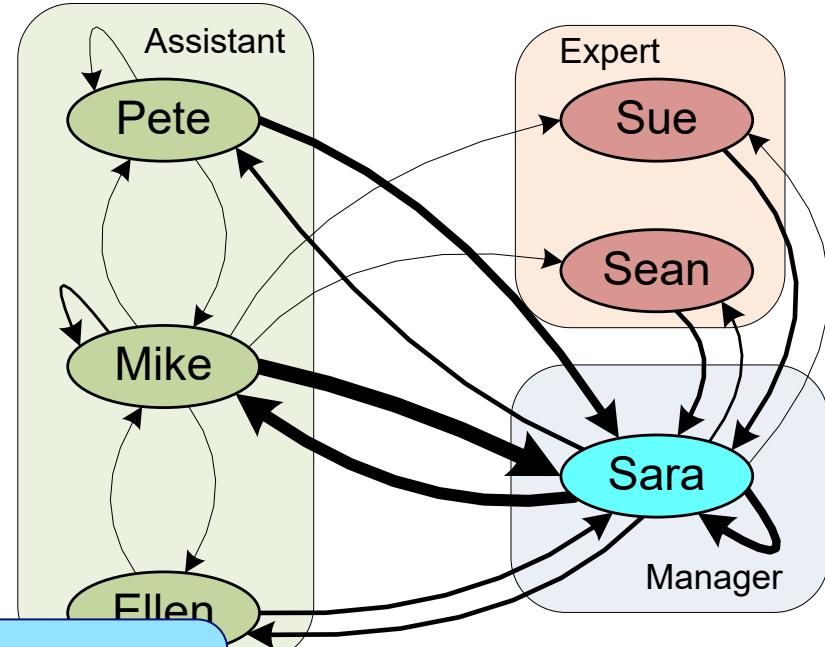
- Taking into account causality (i.e., process model) or not.
- Taking into account the number of transfer within a case or not.
- Taking into account distance.
- Using additional org. information.
- Etc.



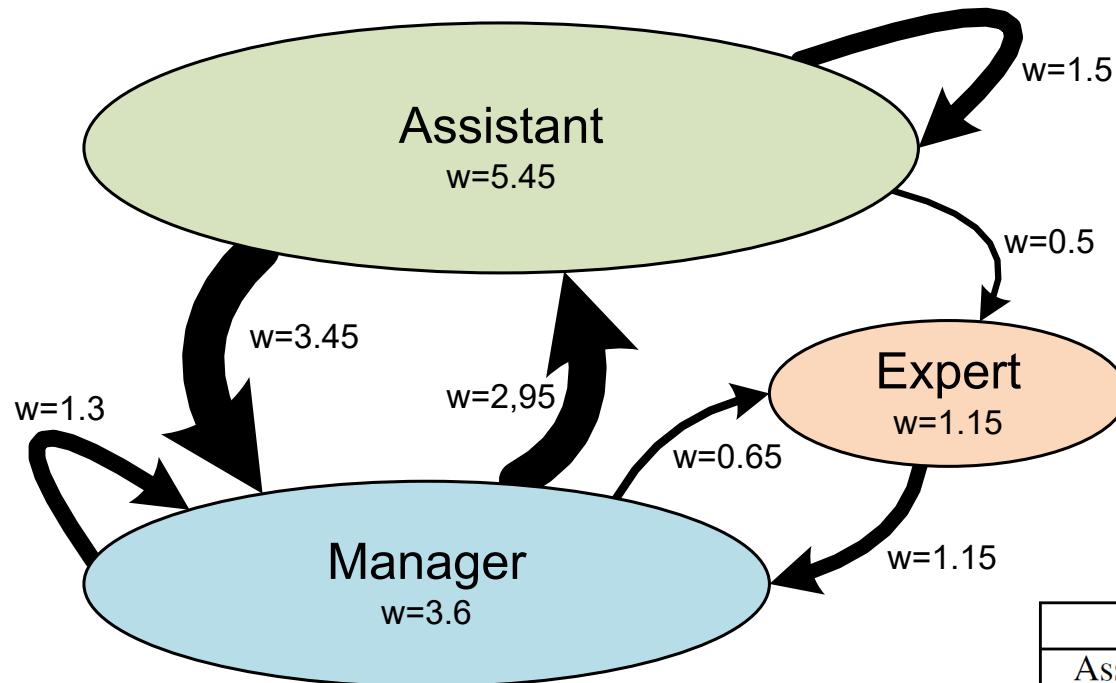
Sometimes we have explicit role or group information

- Information system can provide such information (like an **address book** or **directory**).
- It may also be recorded **with the event itself**.
- Let's assume three roles: **Assistant**, **Expert**, and **Manager**.

Later we will see that roles can also be learned from event data.



Handover of work at role level



In this figure also the size of each node is based on frequencies.

	Assistant	Expert	Manager
Assistant	1.5	0.5	3.45
Expert	0	0	1.15
Manager	2.95	0.65	1.3

Social network miner in ProM

The screenshot shows the ProM UI interface with a central workspace and a right-hand panel for actions.

Actions Panel:

- Filter: A dropdown menu currently set to "social".
- Icons for play, refresh, and search.
- Search bar with the term "social".
- Close button.

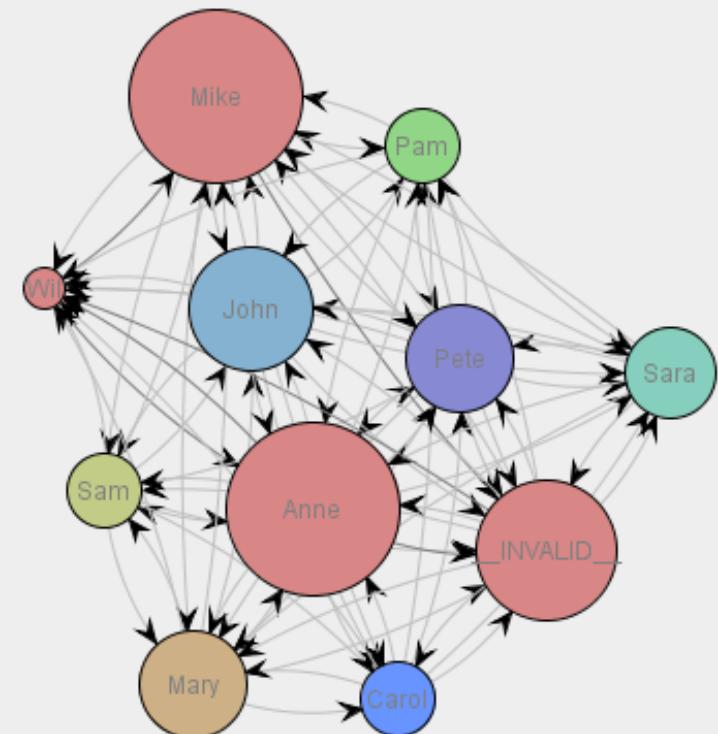
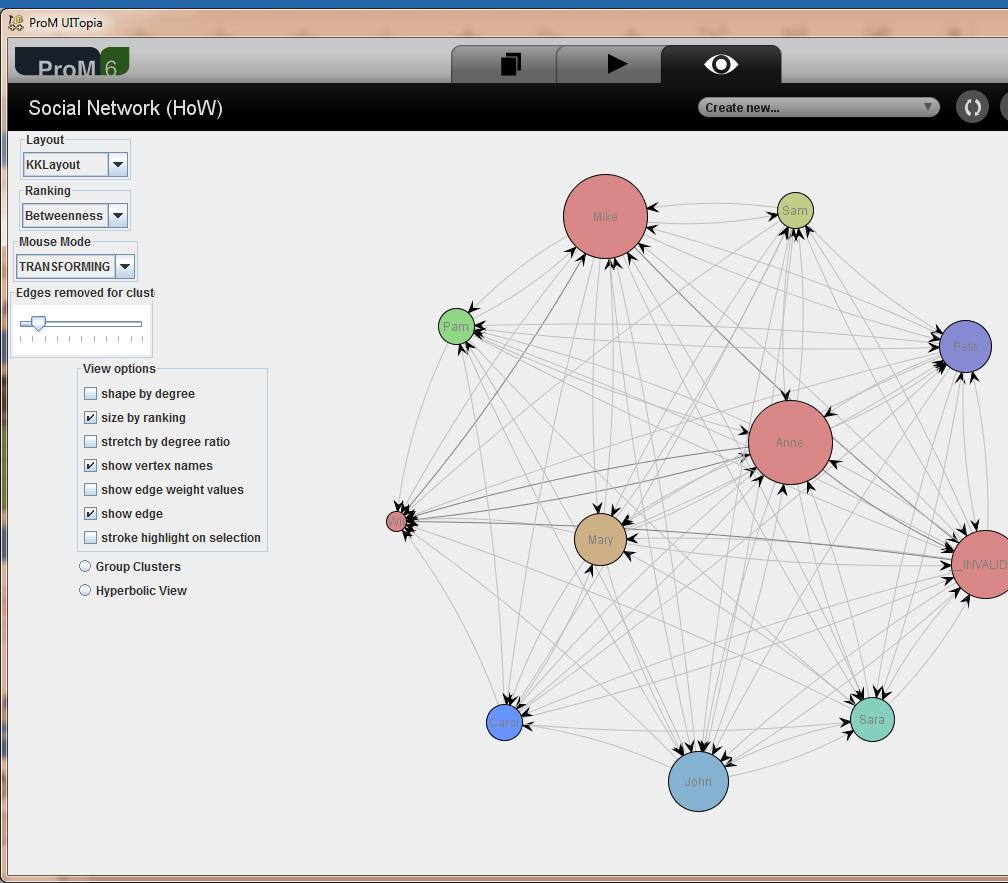
Results List:

- Mine for a Handover-of-Work Social Network by M. Song (m.song@unist.ac.kr)
- Mine for a Reassignment Social Network by M. Song (m.song@unist.ac.kr)
- Mine for a Similar-Task Social Network by M. Song (m.song@unist.ac.kr)
- Mine for a Subcontracting Social Network by M. Song (m.song@unist.ac.kr)
- Mine for a Working-Together Social Network by M. Song (m.song@unist.ac.kr)

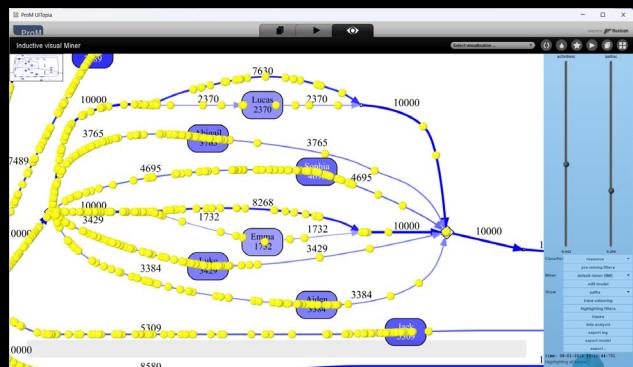
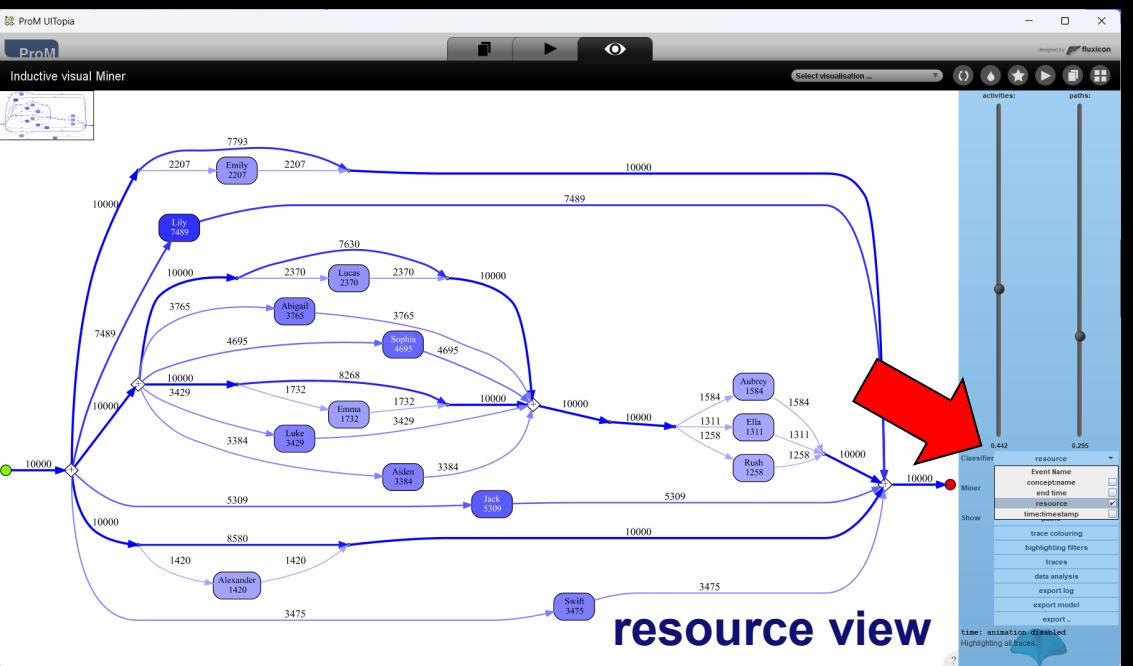
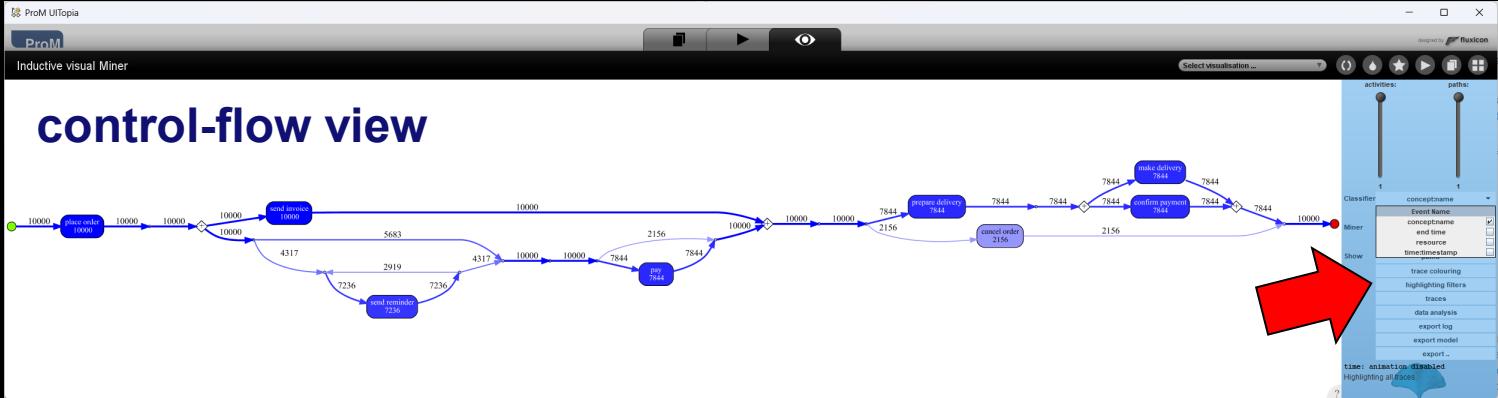
Central Workspace (Yellow Boxes):

- Handover of work (defined before).**
- People are "close" if they have a similar mix of activities.**
- People are "close" if they often work on the same cases.**

Social network based on hand-over of work

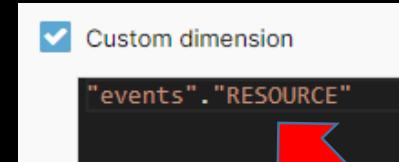
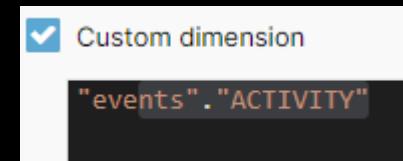
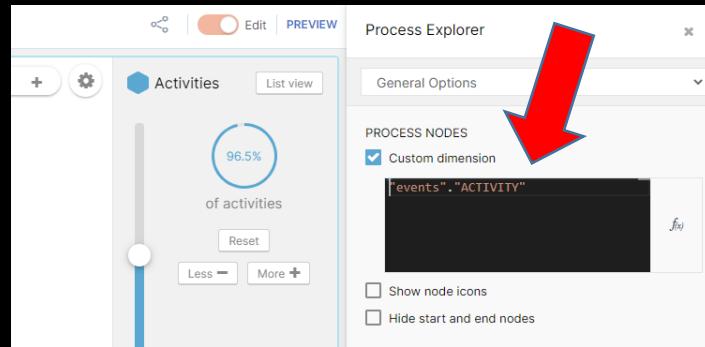
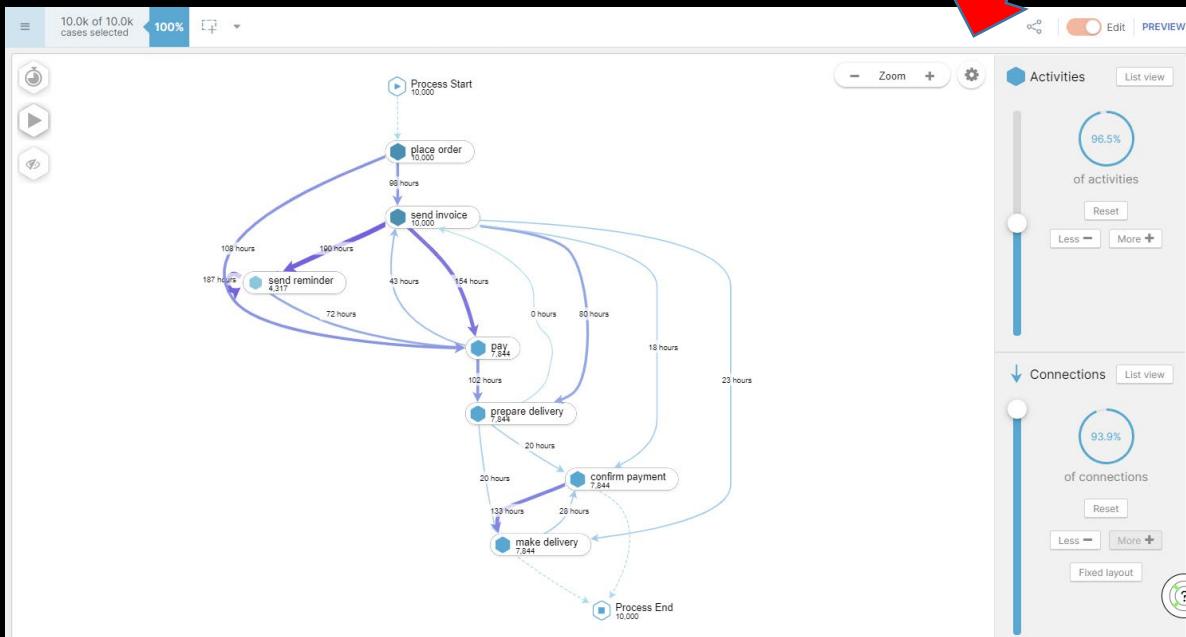


Changing the classifier from activity name to resource name.



Same can be done in Celonis!

activity = resource



10.0k of 10.0k
cases selected

100%



Edit PREVIEW



- Zoom +



Activities

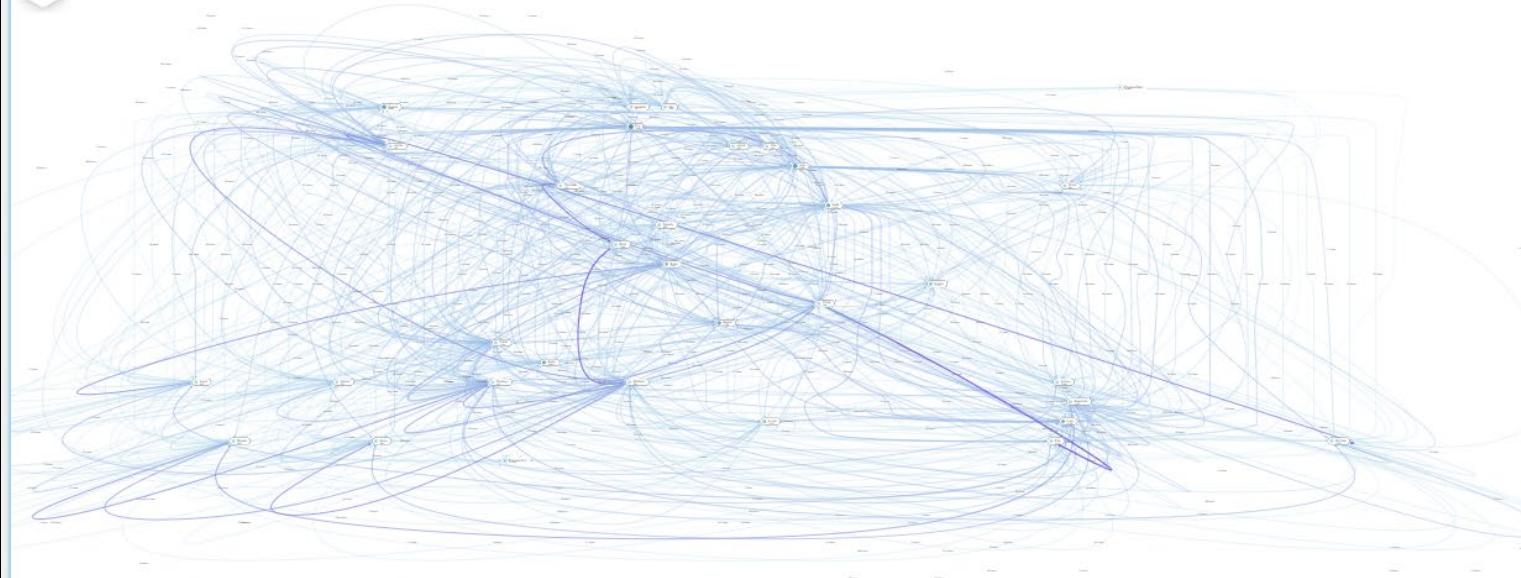
List view



of activities

Reset

Less - More +



Connections

List view



of connections

Reset

Less - More +

Fixed layout

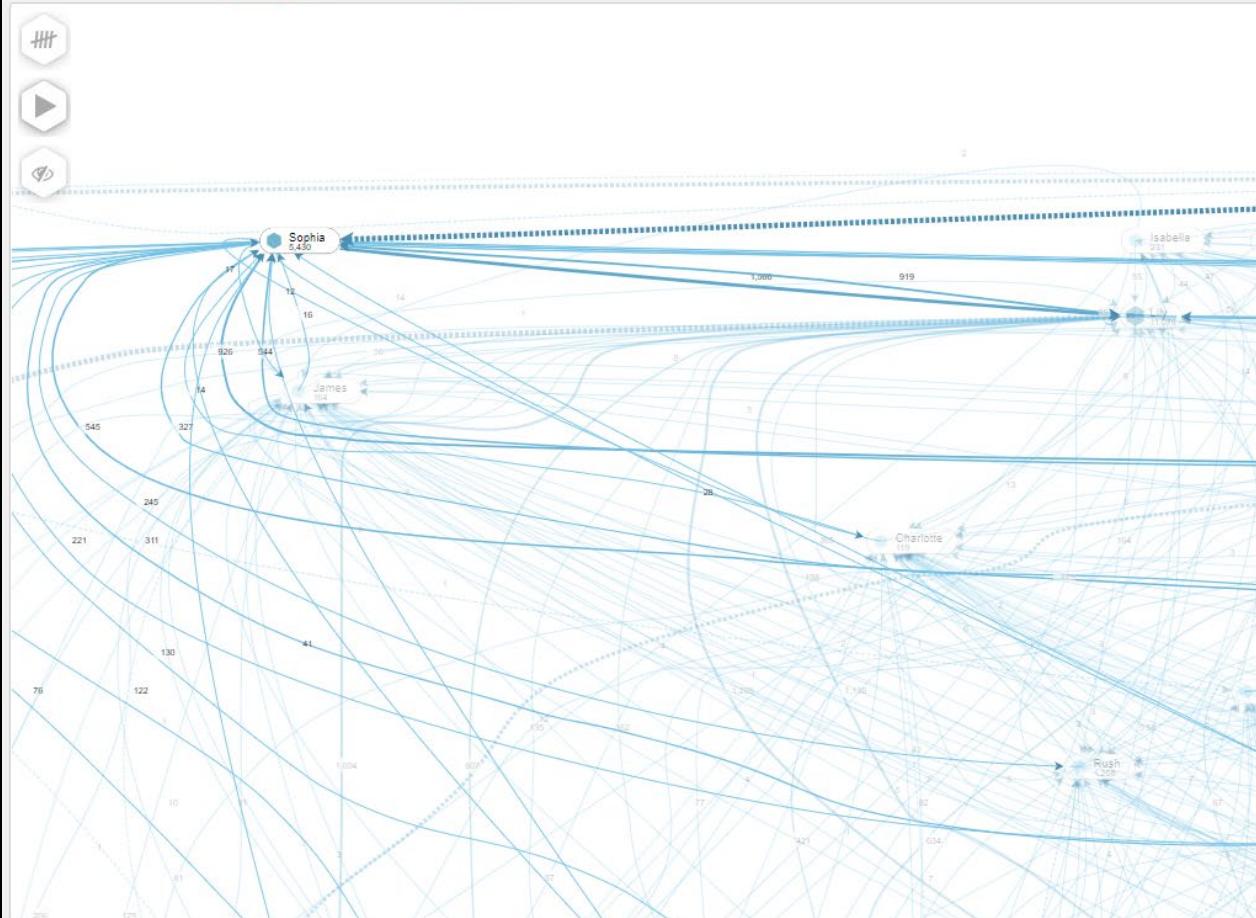


10.0k of 10.0k
cases selected

100%

Keep selection? 

Edit PREVIEW



Sophia

47%

Occurs in 47% of cases

Occurs on average 1.2 times per case

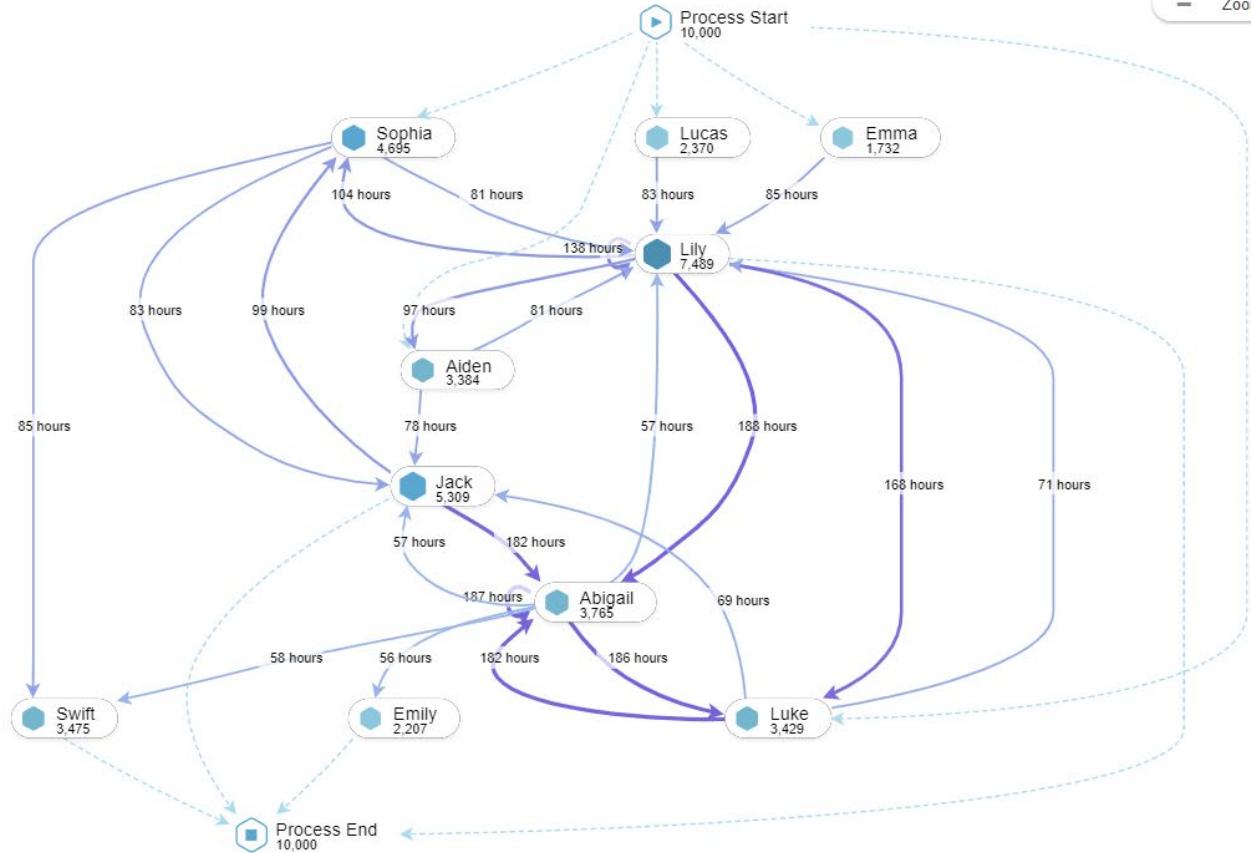
Select cases

with this activity

Cases come from

Cases go to

Activity	Activity Frequency
Lily	1,566
Jack	926
Swift	545
Aubrey	295
Ella	262
Rush	245
Michael	197
Emily	311
Avery	138
Alexander	21
Abigail	130



-

Zoom

+



Activities

List view

76.9%

of activities

Reset

Less - More +

Connections

List view

46.9%

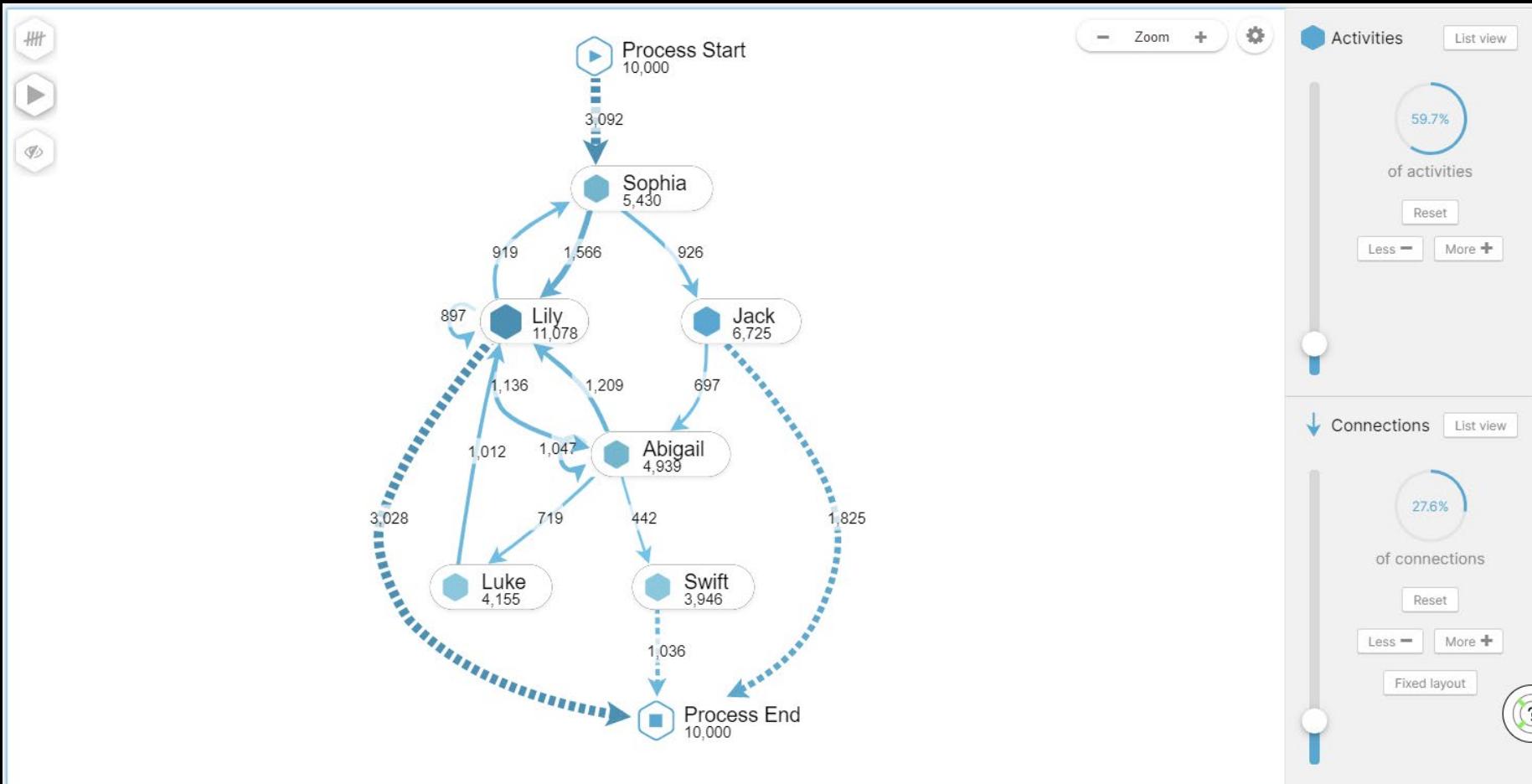
of connections

Reset

Less - More +

Fixed layout

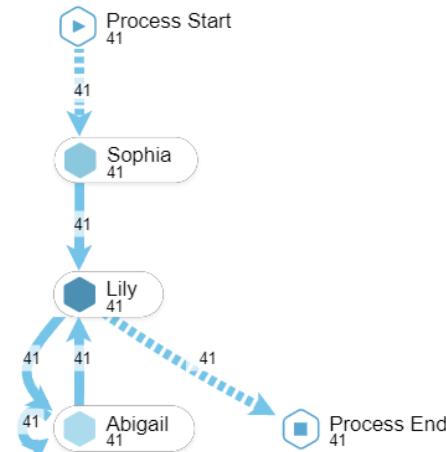
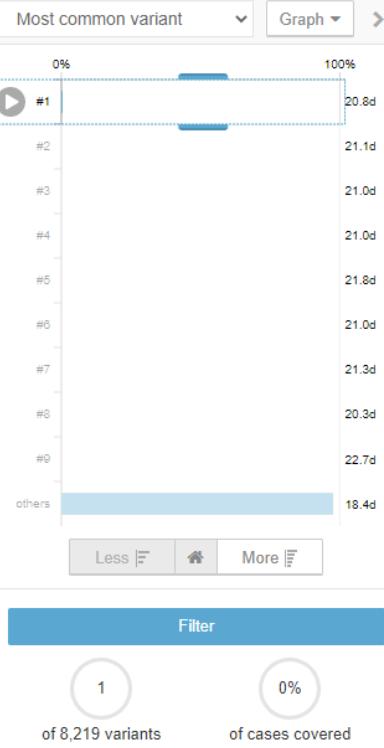




Custom dimension

"events"."RESOURCE"

- Zoom +



General Options

TITLE

PROCESS NODES

Custom dimension

"events"."RESOURCE"

?

BORDER OPTIONS

Show Border

Thickness

Style

Color

Opacity

100%

BACKGROUND OPTIONS

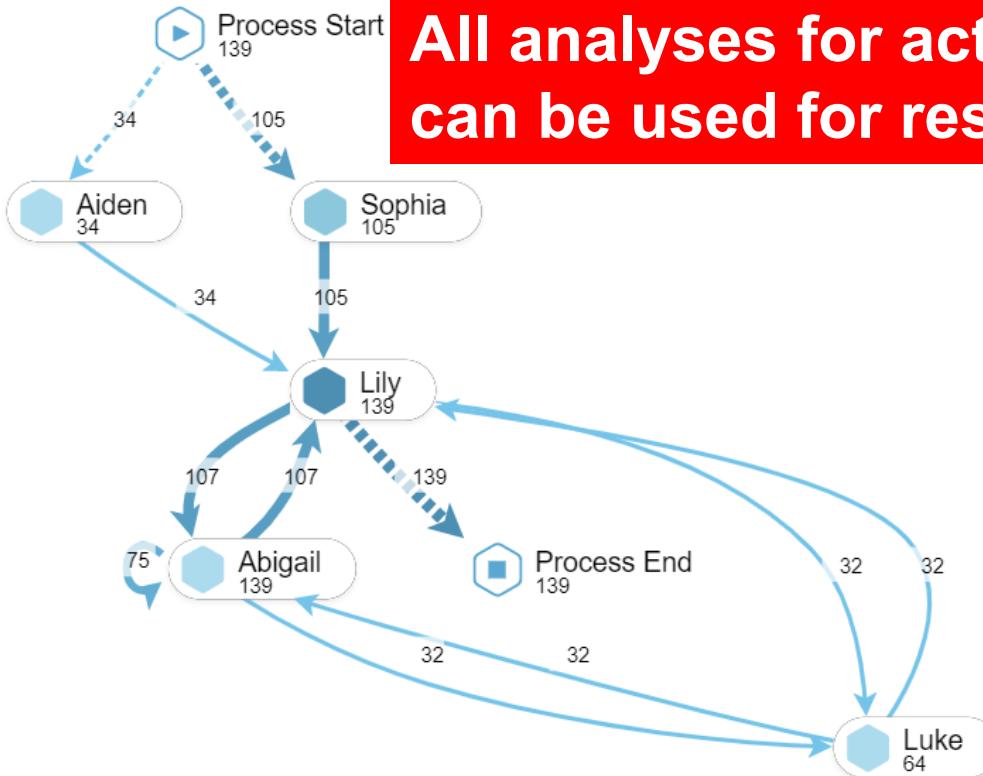
Show background

OTHER OPTIONS

Component is not filtered with selections



All analyses for activities can be used for resources!





- **Resource-activity matrix**
(Who is doing what?).
- **Handover of work matrix**
(How is work passed on?)
- Used to create **social networks** (one of many possibilities).
- Social network can be **analyzed** in many ways.
- Resources can be grouped.

Organizational Mining



Resource-activity matrix

(seen before)

case id trace

1	$\langle a^{Pete}, b^{Sue}, d^{Mike}, e^{Sara}, h^{Pete} \rangle$
2	$\langle a^{Mike}, d^{Mike}, c^{Pete}, e^{Sara}, g^{Ellen} \rangle$
3	$\langle a^{Pete}, c^{Mike}, d^{Ellen}, e^{Sara}, f^{Sara}, b^{Sean}, d^{Pete} \rangle$
4	$\langle a^{Pete}, b^{Mike}, c^{Sue}, d^{Sean}, e^{Ellen} \rangle$

Mean number of times a resource performs an activity per case.

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0

Question: Which resources are similar?

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0

Suppose that you need to make three groups with similar resources. What would these groups be?

Answer

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0

{ Pete, Mike, Ellen }

{ Sue, Sean }

{ Sara }

Distance based on resource-activity matrix

Standard notions of "distance" can be used e.g., Euclidian distance, Manhattan distance, Minkowski distance, and Pearson's correlation coefficient.

$$P_{Pete} = (0.3, 0, 0.345, 0.69, 0, 0, 0.135, 0.165)$$

$$P_{Mike} = (0.5, 0, 0.575, 1.15, 0, 0, 0.225, 0.275)$$

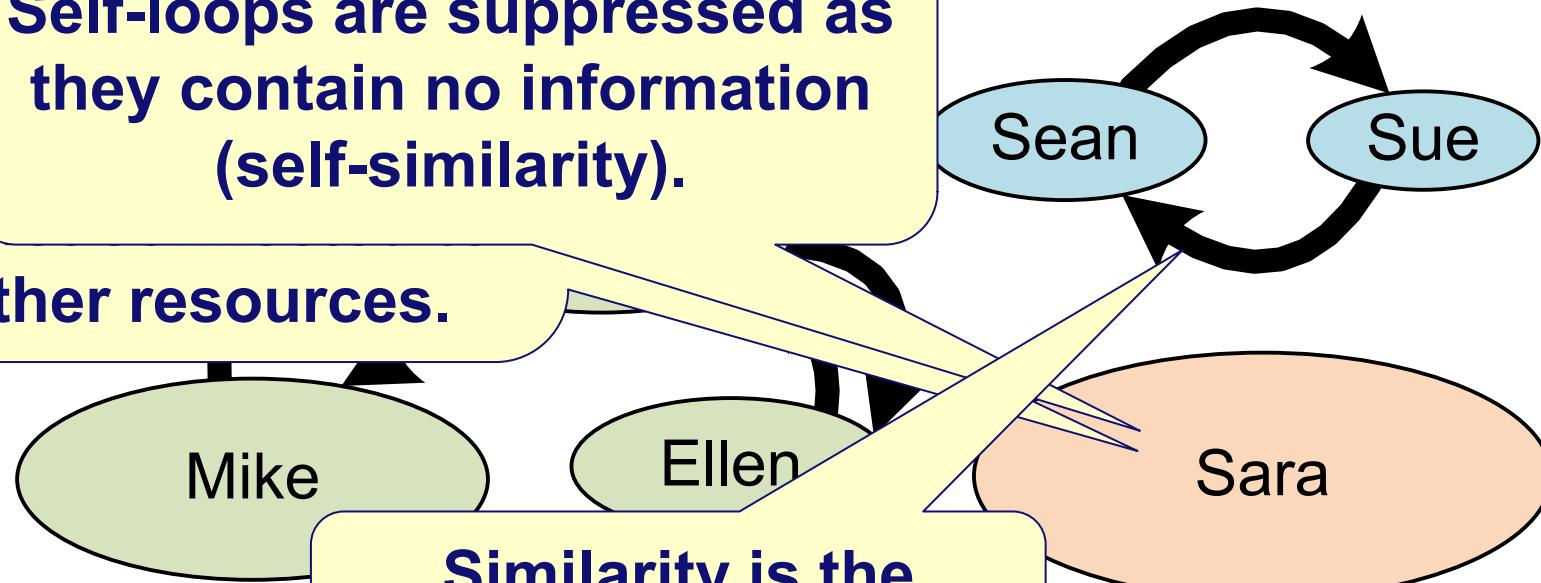
$$P_{Sara} = (0, 0, 0, 0, 2.3, 1.3, 0, 0)$$

Social network based on similarity of profiles

Resources
similarities
resources
and
is

Self-loops are suppressed as they contain no information (self-similarity).

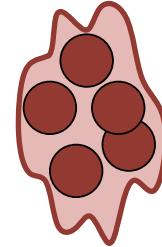
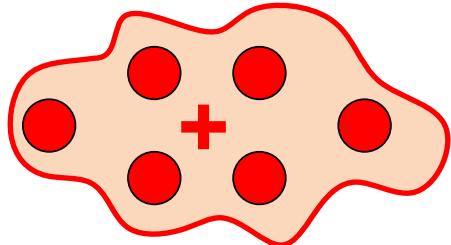
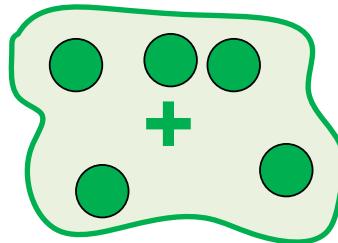
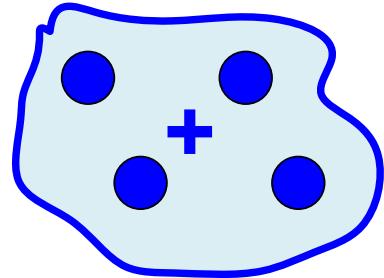
other resources.



Similarity is the inverse of distance.

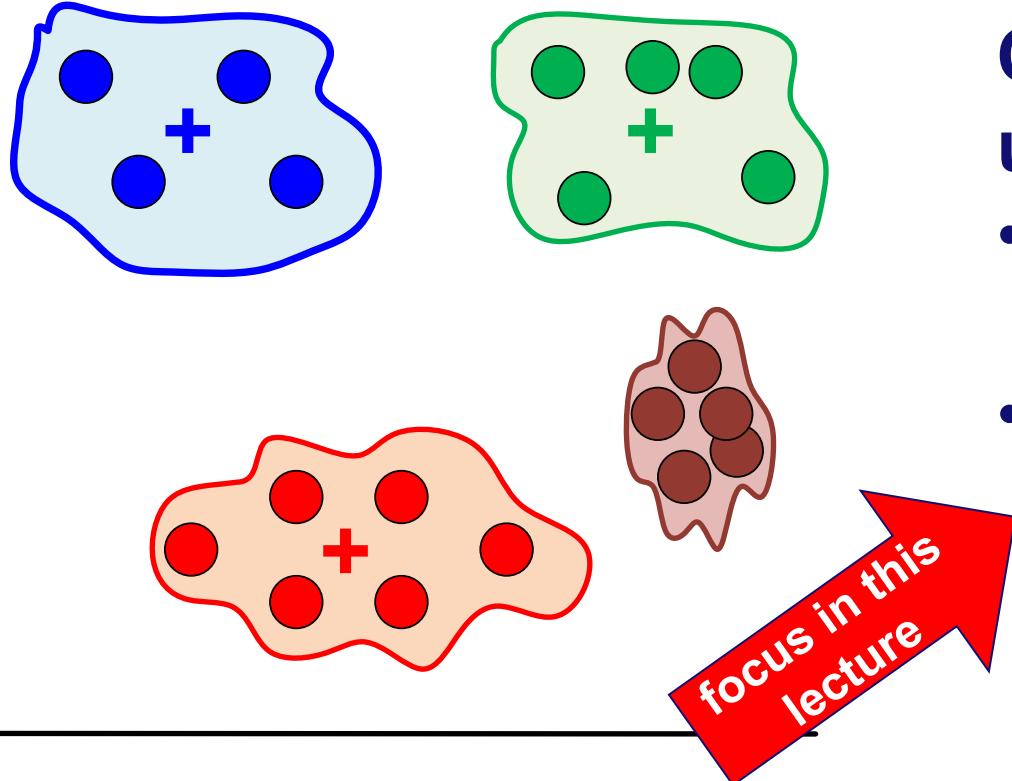
Related to clustering

(basic data mining technique discussed before)



- **k -means clustering**
- **agglomerative hierarchical clustering**

Clustering: cases and resources



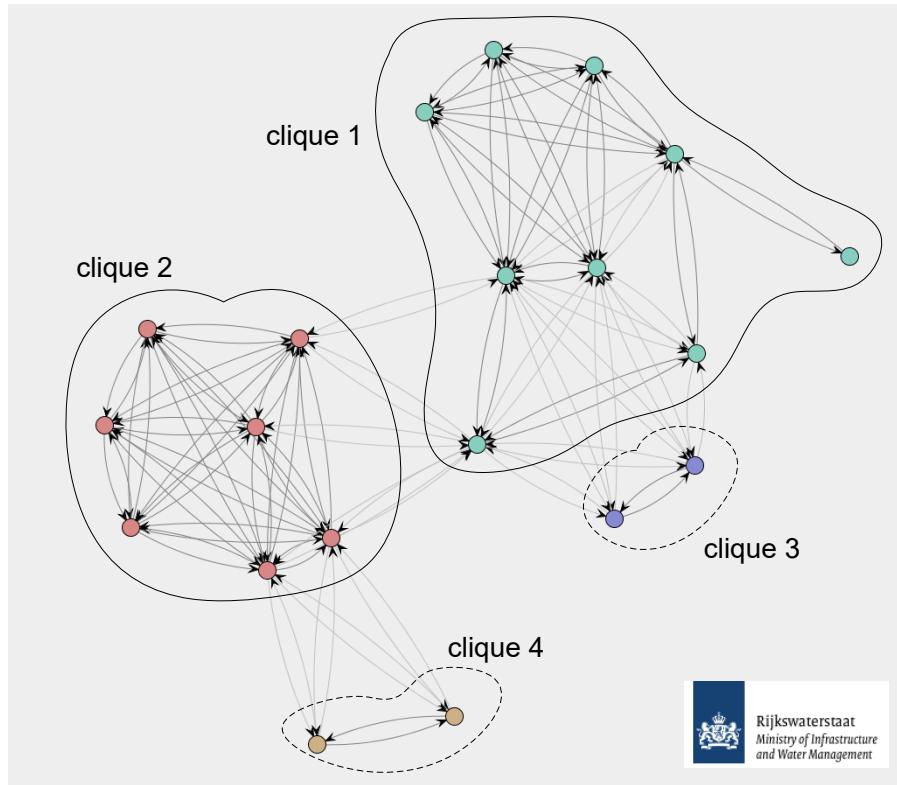
Clustering may be used for:

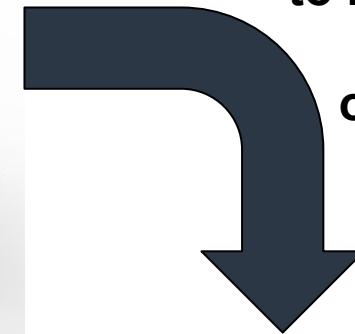
- grouping **cases** (process variants)
- grouping **resources** (identifying roles)

Real-life example: Roles found by ProM

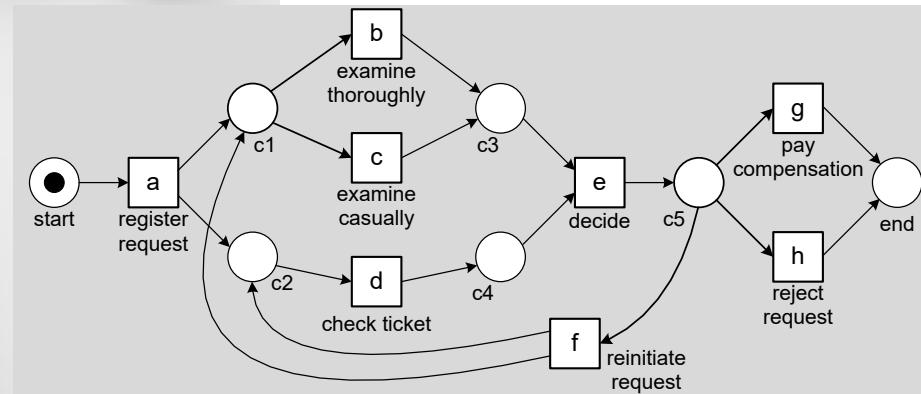
(not normalized per case)

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

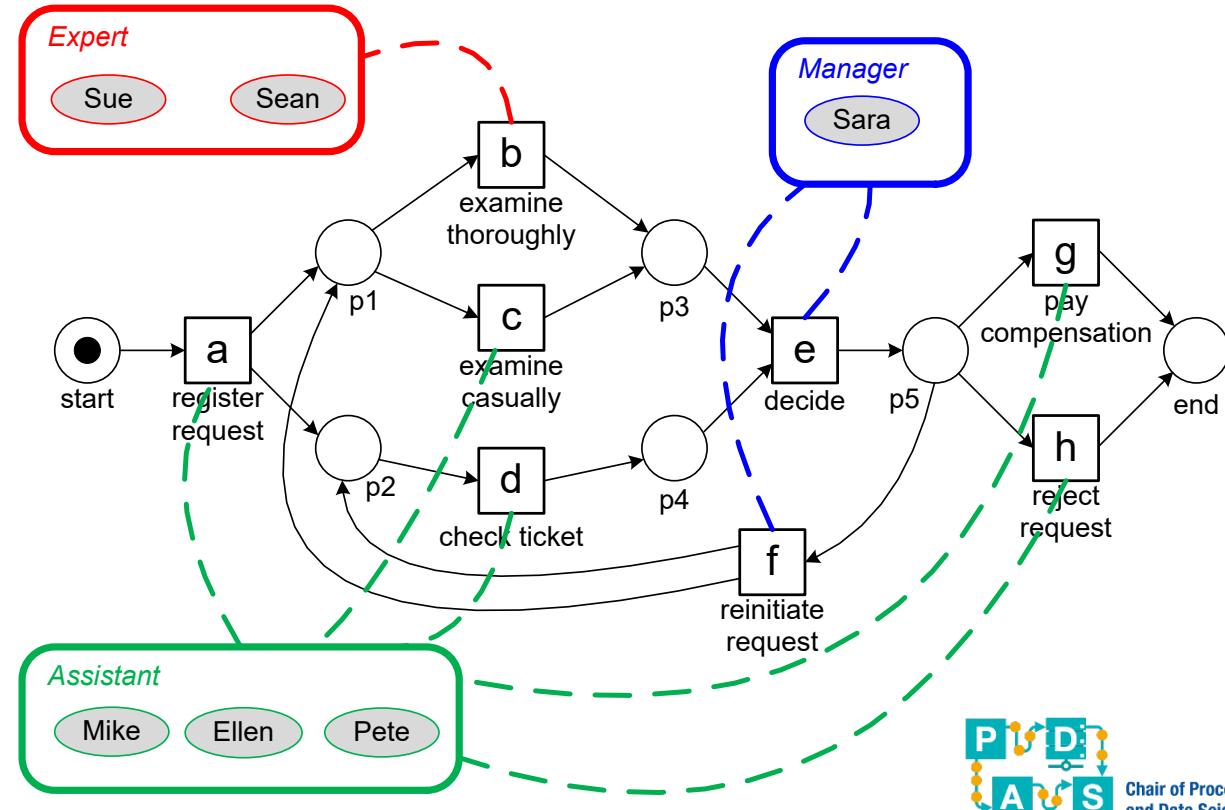
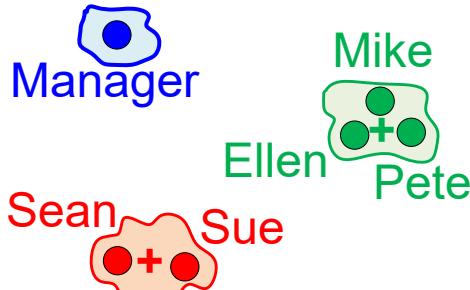




assign activities
to roles, groups,
or other
organizational
entities



Extending process models with the organizational perspective

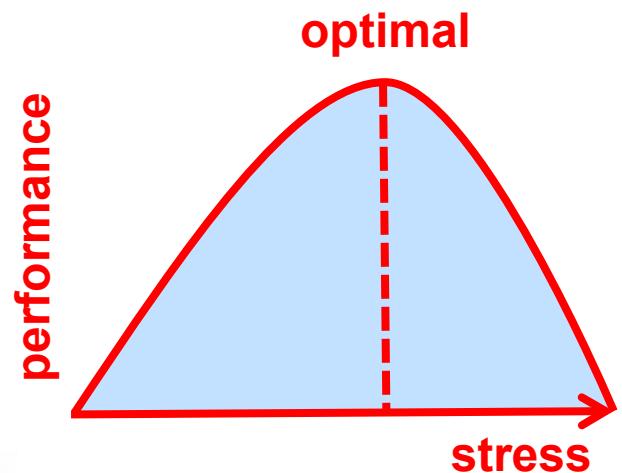


Learning more about resources

- When are resources **available**?
 - part-time, vacation, illness, ...
 - shared among different processes
- Which resources **collaborate well**?
- Which resources perform well on **specific activities**?
- ...



Yerkes-Dodson law of arousal



Resource-based situation table

- **Five types of situation tables:**

- **Case-based situation table:** Each row (instance) corresponds to a case with variables.
- **Event-based situation table:** Each row (instance) corresponds to an event.
- **Resource-based situation table:** Each row (instance) corresponds to a resource.
- **Event-pair-based situation table:** Each row (instance) corresponds to a pair of events.
- **Aggregate situation tables:** Each row (instance) corresponds to a combination of cases and/or events.

By now you should understand the basic mechanisms in PQL and see that you can compute any process feature.

RESOURCE	Case count	Activities count	Distinct addresses	Distinct activities
Abigail	3,765	4,939	4	2
Aiden	3,384	3,746	4	2
Alexander	1,420	1,491	4	4
Aubrey	1,584	1,584	3	1
Avery	769	769	3	1
Caleb	1,154	1,207	4	6
Charlotte	119	119	4	4
Chloe	189	190	4	4
Ella	1,311	1,311	3	1
Emily	2,207	2,372	4	4
Emma	1,732	1,822	4	2
Harper	321	321	3	1
Isabella	227	231	4	2
Jack	5,309	6,725	4	4
Jacob	429	436	4	2
James	163	164	4	4
Kaylee	545	545	3	1
Layla	431	431	3	1
Lily	7,489	11,078	4	4
Lucas	2,370	2,535	4	2
Luke	3,429	4,155	4	3
Madelyn	515	524	4	4
Madison	335	338	4	4
Mia	149	151	4	2
Michael	1,033	1,033	3	1
Olivia	640	655	4	2
Rush	1,258	1,258	3	1
Sophia	4,695	5,430	4	2
Speedy	899	917	4	2
Swift	3,475	3,946	4	4
Zoe	344	345	4	2

Resource-based situation table

10.0k of 10.0k cases selected 100% PREVIEW

RESOURCE	Case count	Activities count	Distinct addresses	Distinct activities
Abigail	3,765	4,939	4	2
Aiden	3,384	3,746	4	2
Alexander	1,420	1,491	4	4
Aubrey	1,584	1,584	3	1
Avery	769	769	3	1
Caleb	1,154	1,207	4	6
Charlotte	119	119	4	4
Chloe	189	190	4	4
Ella	1,311	1,311	3	1
Emily	2,207	2,372	4	4
Emma	1,732	1,822	4	2
Harper	321	321	3	1
Isabella	227	231	4	2
Jack	5,309	6,725	4	4
Jacob	429	436	4	2
James	163	164	4	4
Kaylee	545	545	3	1
Layla	431	431	3	1
Lily	7,489	11,078	4	4
Lucas	2,370	2,535	4	2
Luke	3,429	4,155	4	3
Madelyn	515	524	4	4
Madison	335	338	4	4
Mia	149	151	4	2
Michael	1,033	1,033	3	1
Olivia	640	655	4	2
Rush	1,258	1,258	3	1
Sophia	4,695	5,430	4	2
Speedy	899	917	4	2
Swift	3,475	3,946	4	4
Zoe	344	345	4	2

Component options

General options

Table title:

Component type: OLAP Table

DIMENSIONS

RESOURCE:

KPIs

Case count:

Activities count:

Distinct addresses:

SORTING

ADVANCED OPTIONS

Distinct values

"events"."RESOURCE"

COUNT_TABLE("cases")

COUNT("events"."ACTIVITY")

COUNT(DISTINCT "cases"."ADDRESS")

COUNT(DISTINCT "events"."ACTIVITY")



Chair of Process
and Data Science

All-inclusive process models are needed ...

Processes, people, roles, and other organizational entities are intertwined and cannot be viewed in isolation.



Combining Different Perspectives



**data
perspective**

**resource
perspective**



**time
perspective**

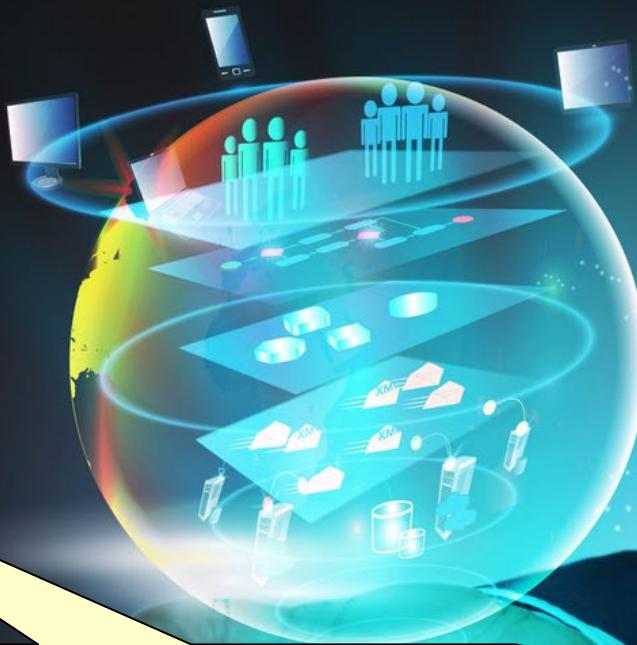
**control-flow
perspective**

A group of blind men heard that a strange animal, called an elephant, had been brought to the town, but none of them were aware of its shape and form. So, they decided to go and see it out, and when they found it they groped about it. The first person, whose hand landed on the trunk, said, "This being is like a thick snake". For another who understanding reached its ear, it seemed like a kind of fan. As for another person, whose hand was upon its leg, said, the elephant is a pillar like a tree-trunk. The blind man who placed his hand upon its side said the elephant, "is a wall". Another who felt its tail, described it as a rope. The last felt its tusk, stating the elephant is that which is hard, smooth and like a spear.

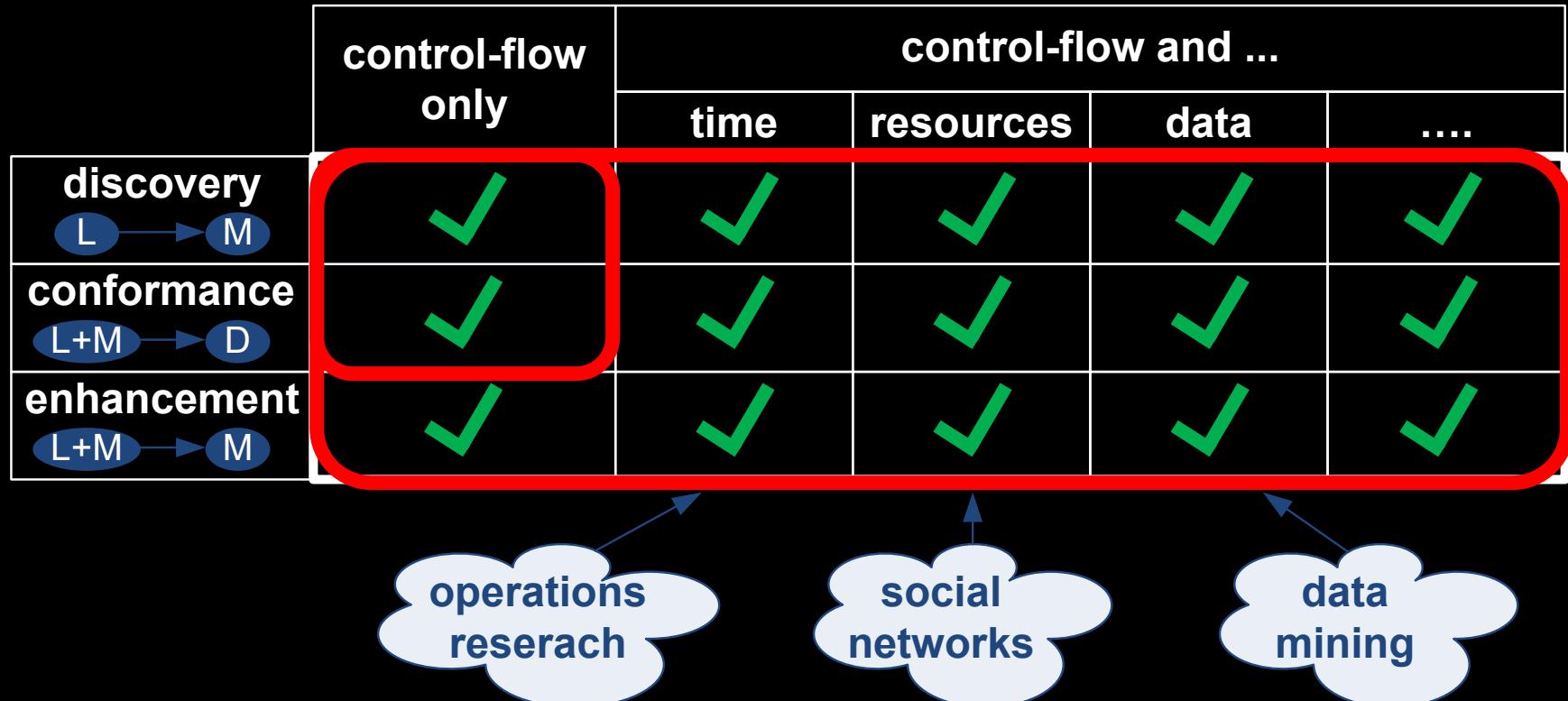
- process model
- control-flow
- data
- resources
- time
- ...

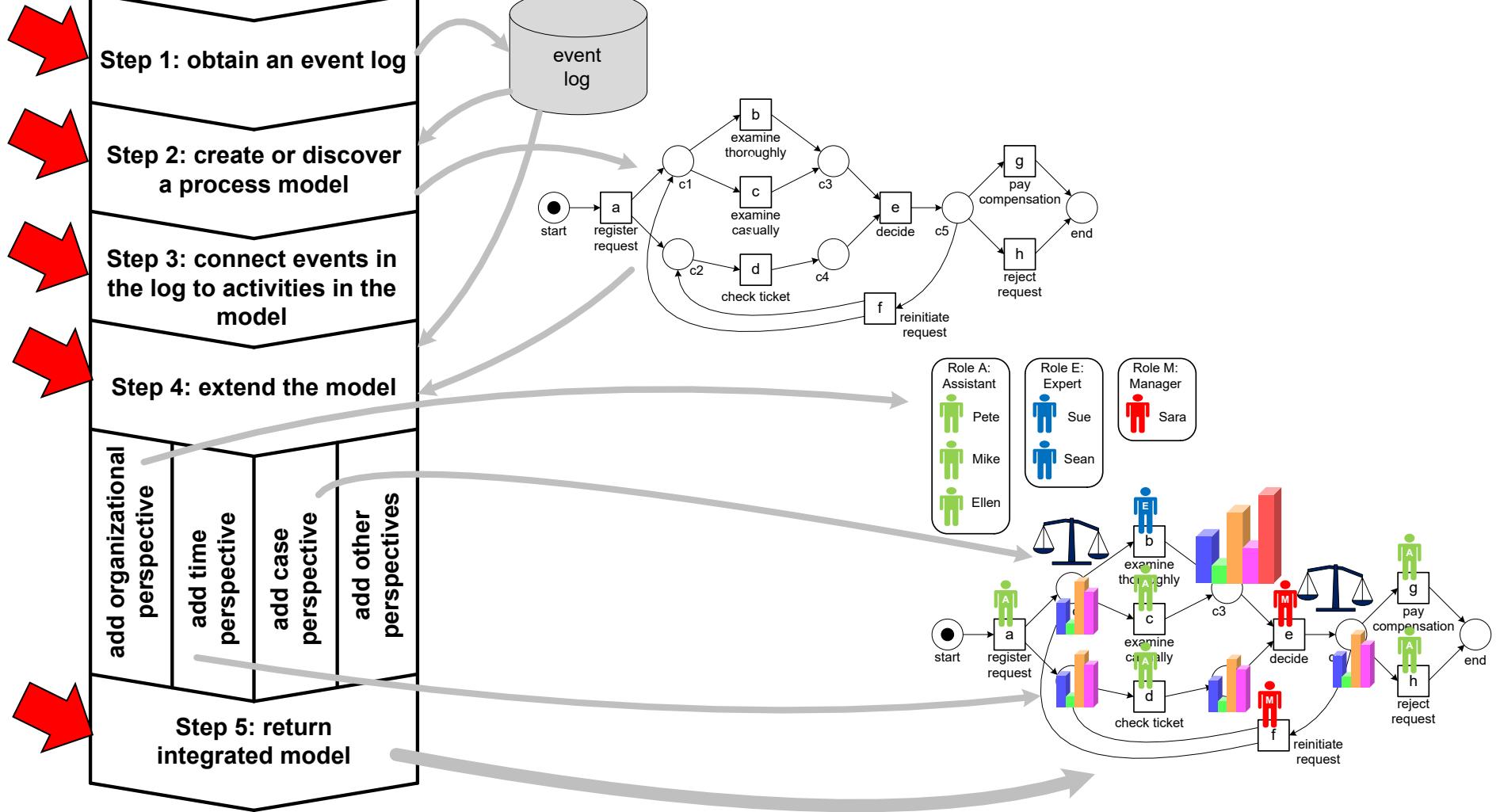


backbone

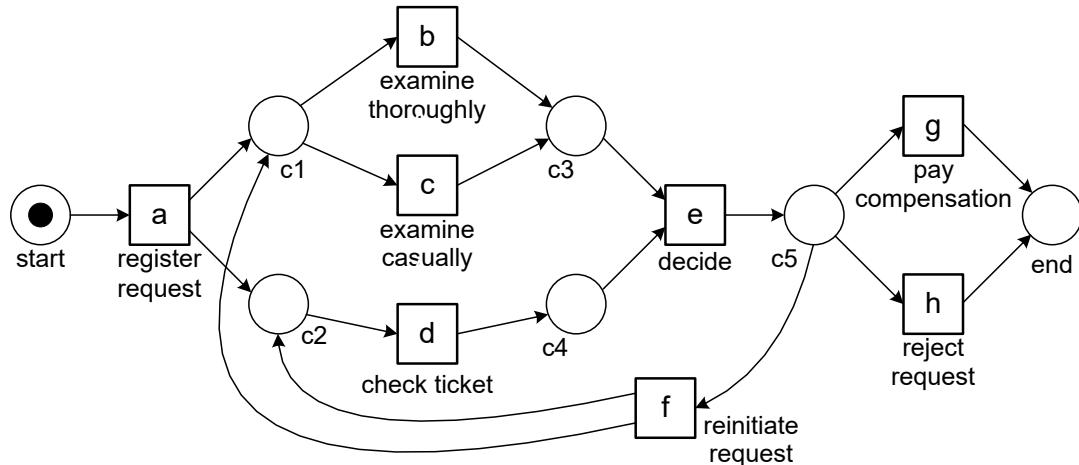


Bigger picture



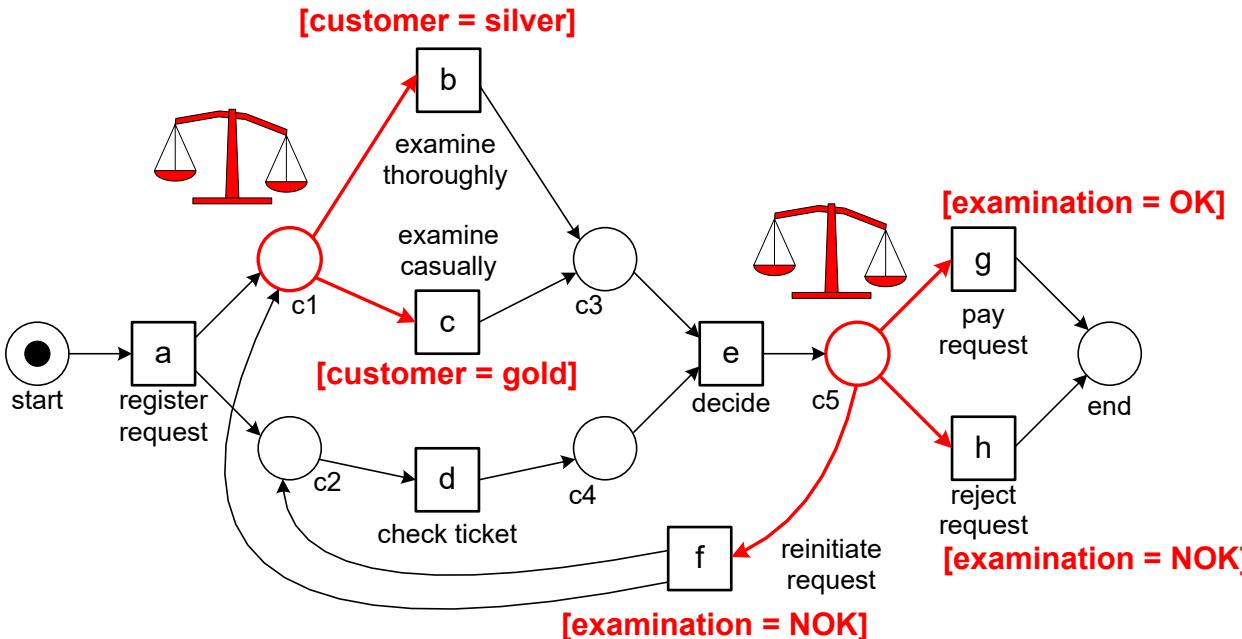


Starting point: Control-flow



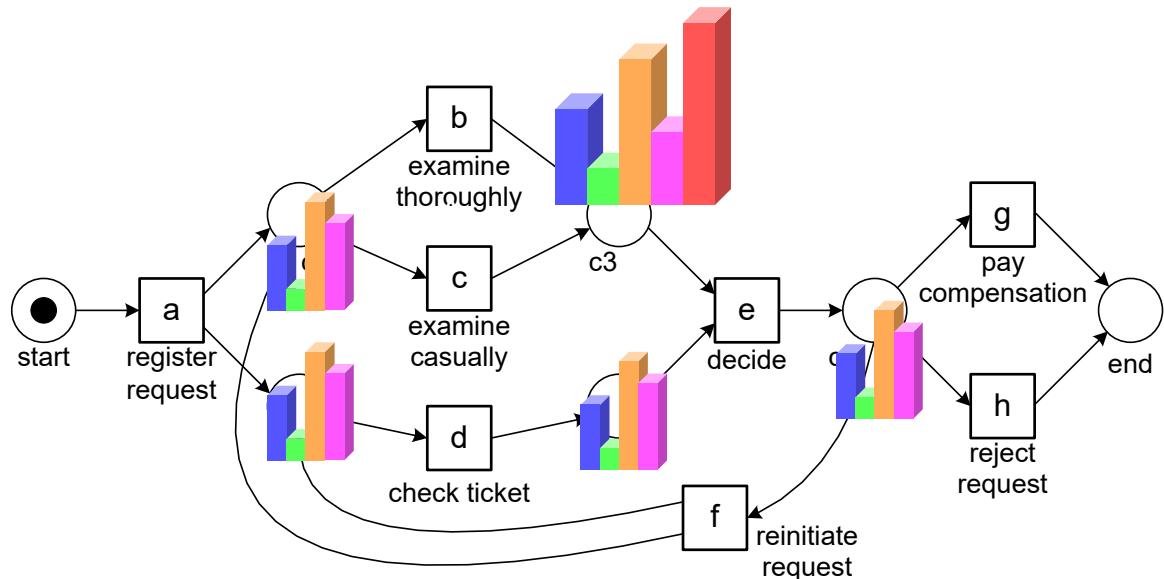
- Discovered or made by hand.
- Should be aligned with event log.

Adding the data perspective



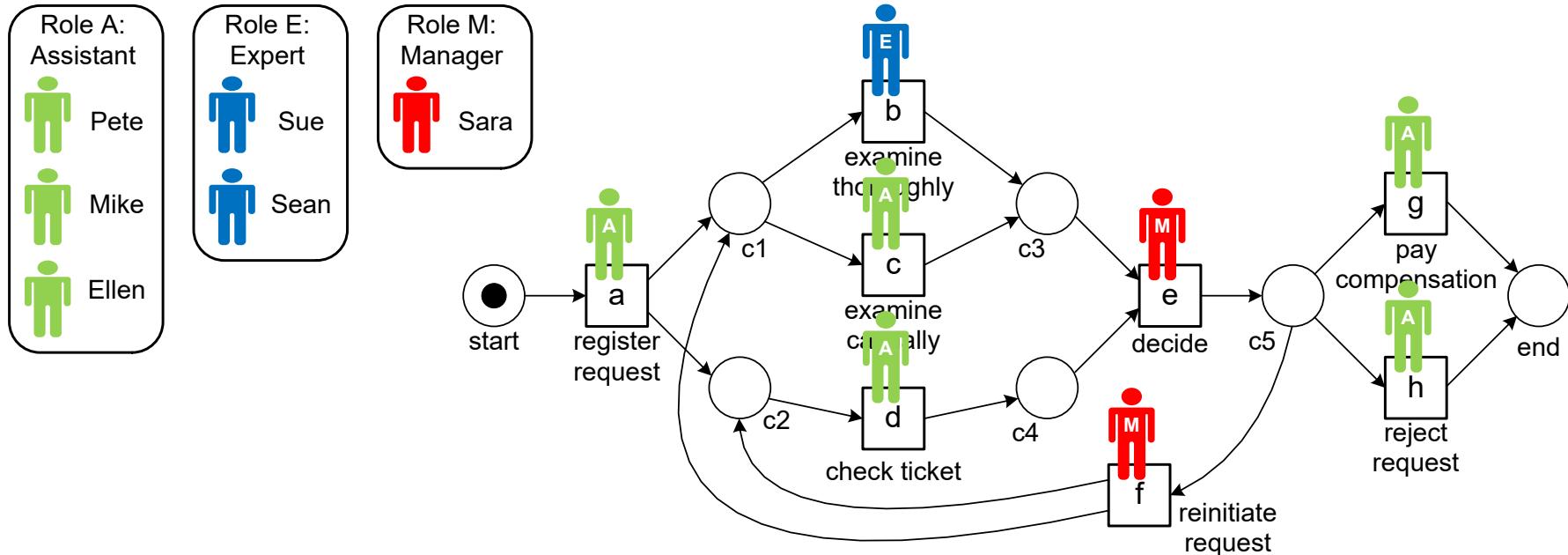
- **Decision tree learning can be used to create guards.**
- **Not shown:**
 - **variables**
 - **read & write arcs**

Adding the time perspective



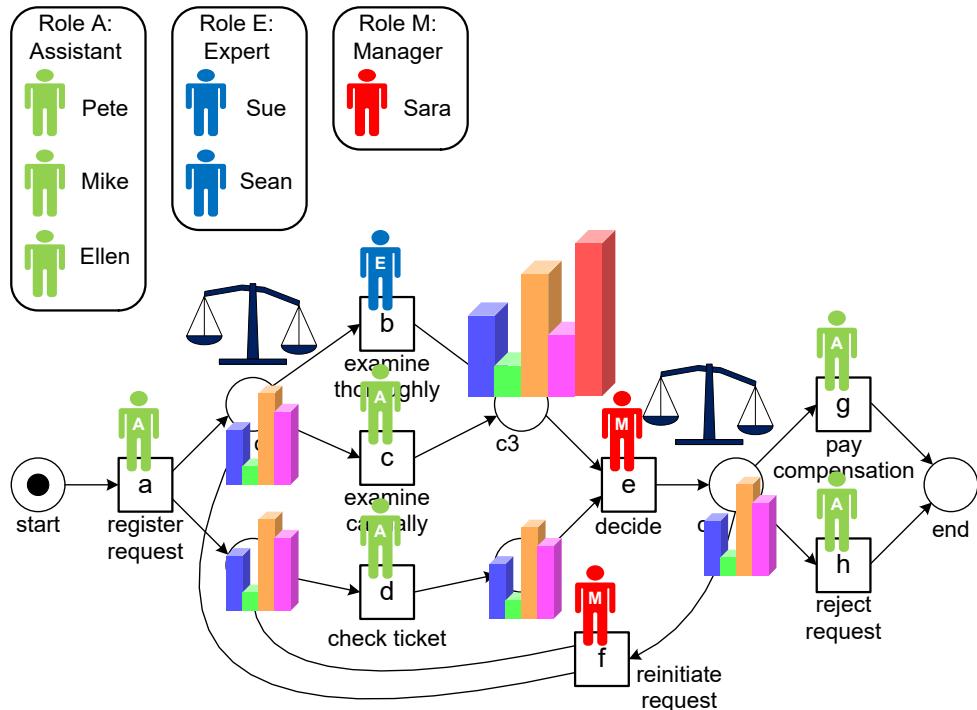
- Replay event log to compute **waiting times** and **service times** (distribution or just mean and variance).
- Also capture routing probabilities.

Adding the resource perspective



Roles are discovered automatically (e.g., clustering based on resource-activity matrix) or obtained from information system.

Integrated model



- **Input for**
 - (holistic) diagnosis
 - reengineering
 - operational support
- **Example: simulation.**
- **Be aware of limitations**
model (over- or under-fitting), descriptive (not normative), etc.

Part I: Introduction

Chapter 1
Data Science in Action

Chapter 2
Process Mining:
The Missing Link

Part II: Preliminaries

Chapter 3
Process Modeling
and Analysis

Chapter 4
Data Mining

Part III: From Event Logs to Process Models

Chapter 5
Getting the Data

Chapter 6
Process Discovery:
An Introduction

Chapter 7
Advanced Process
Discovery Techniques

Part IV: Beyond Process Discovery

Chapter 8
Conformance
Checking

Chapter 9
Mining Additional
Perspectives

Chapter 10
Operational Support

Part V: Putting Process Mining to Work

Chapter 11
Process Mining
Software

Chapter 12
Process Mining in the
Large

Chapter 13
Analyzing “Lasagna
Processes”

Chapter 14
Analyzing “Spaghetti
Processes”

Part VI: Reflection

Chapter 15
Cartography and
Navigation

Chapter 16
Epilogue



ID	Topic	Date	Date	Place
	Lecture 1 Introduction to Process Mining	08.04.24	Monday	AH V
	Lecture 2 Data Science: Supervised Learning	09.04.24	Tuesday	AH V
	<i>Exercise 1 Tool Introduction</i>	09.04.24	Tuesday	AH III
	Lecture 3 Data Science: Unsupervised Learning and Evaluation	15.04.24	Monday	AH V
	Lecture 4 Introduction to Process Discovery	16.04.24	Tuesday	AH V
	<i>Exercise 2 Data Mining</i>	16.04.24	Tuesday	AH III
	Lecture 5 Alpha Algorithm 1	22.04.24	Monday	AH V
	Lecture 6 Alpha Algorithm 2	23.04.24	Tuesday	AH V
	<i>Exercise 3 Petri Nets</i>	23.04.24	Tuesday	AH III
	Lecture 7 Model Quality Representation	29.04.24	Monday	AH V
	Lecture 8 Heuristic Mining	30.04.24	Tuesday	AH V
	<i>Exercise 4 Alpha Miner</i>	30.04.24	Tuesday	AH III
	Lecture 9 Region-Based Mining	06.05.24	Monday	AH V
	<i>Exercise 5 Heuristic Mining and Region-Based Mining</i>	07.05.24	Tuesday	AH III
	Lecture 10 Inductive Mining	13.05.24	Monday	AH V
	Lecture 11 Event Data and Exploration	14.05.24	Tuesday	AH V
	<i>Exercise 6 Inductive Mining</i>	14.05.24	Tuesday	AH III
	Lecture 12 Conformance Checking 1	27.05.24	Monday	AH V
	Lecture 13 Conformance Checking 2	28.05.24	Tuesday	AH V
	<i>Q&A Session Assignment Part I</i>	28.05.24	Tuesday	AH III
	Deadline Assignment Part I	02.06.24	Sunday	
	<i>Exercise 7 Footprint and Token-Based Replay (Exercise)</i>	03.06.24	Monday	AH V
	<i>Exercise 8 Alignments (Exercise)</i>	04.06.24	Tuesday	AH V
	Lecture 14 Decision Mining	10.06.24	Monday	AH V
	<i>Lecture 15 Celonis Guest Lecture</i>	11.06.24	Tuesday	AH V
	<i>Exercise 9 Decision Mining</i>	11.06.24	Tuesday	AH III
	Lecture 16 Performance Analysis and Organizational Mining	17.06.24	Monday	AH V
	<i>Exercise 10 Performance Analysis (Exercise)</i>	18.06.24	Tuesday	AH V
	<i>Exercise 11 Organizational Mining</i>	18.06.24	Tuesday	AH III
	<i>Exercise 12 Celonis Case Study</i>	24.06.24	Monday	AH V
	Lecture 17 Operational Support and Process Mining Applications	01.07.24	Monday	AH V
	Lecture 18 Distributed, Streaming, and Comparative Process Mining	02.07.24	Tuesday	AH V
	<i>Exercise 13 Operational Process Mining</i>	02.07.24	Tuesday	AH III
	Lecture 19 Closing	08.07.24	Monday	AH V
	<i>Q&A Session Assignment Part II</i>	09.07.24	Tuesday	AH III
	Deadline Assignment Part II	14.07.24	Sunday	
	<i>Q&A Session Exam</i>	16.07.24	Tuesday	AH III



Operational Support and Process Mining Applications

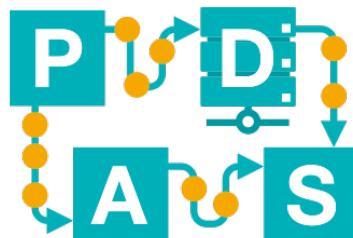
Lecture 17

prof.dr.ir. Wil van der Aalst

www.vdaalst.com @wvdaalst

www.pads.rwth-aachen.de

BPI-L17

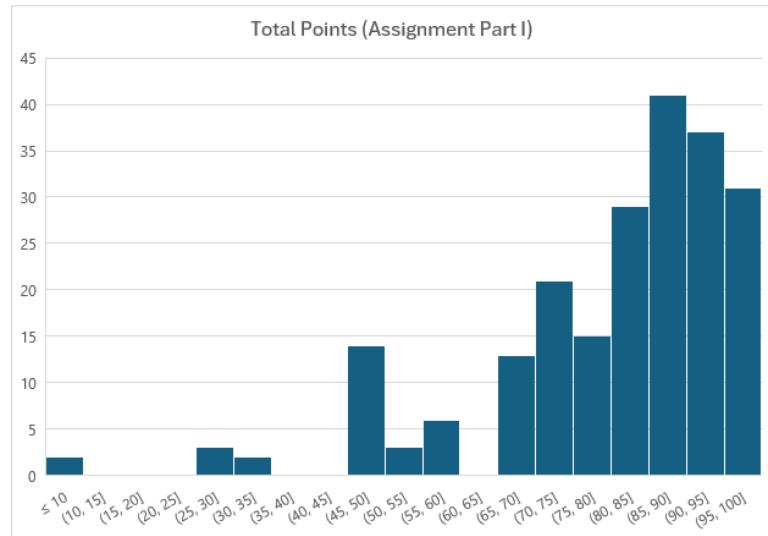
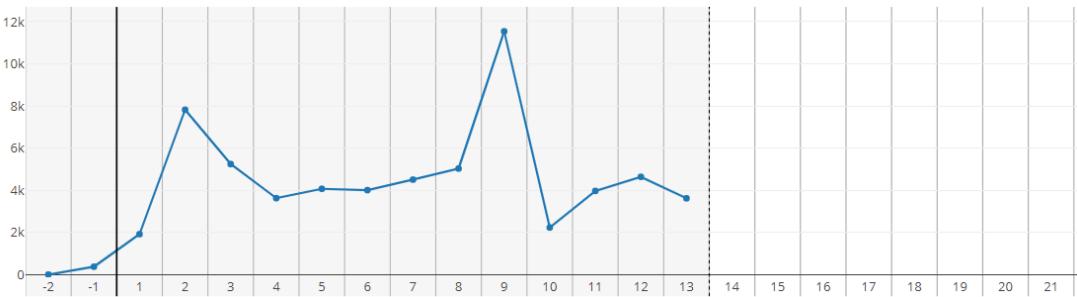


Chair of Process
and Data Science

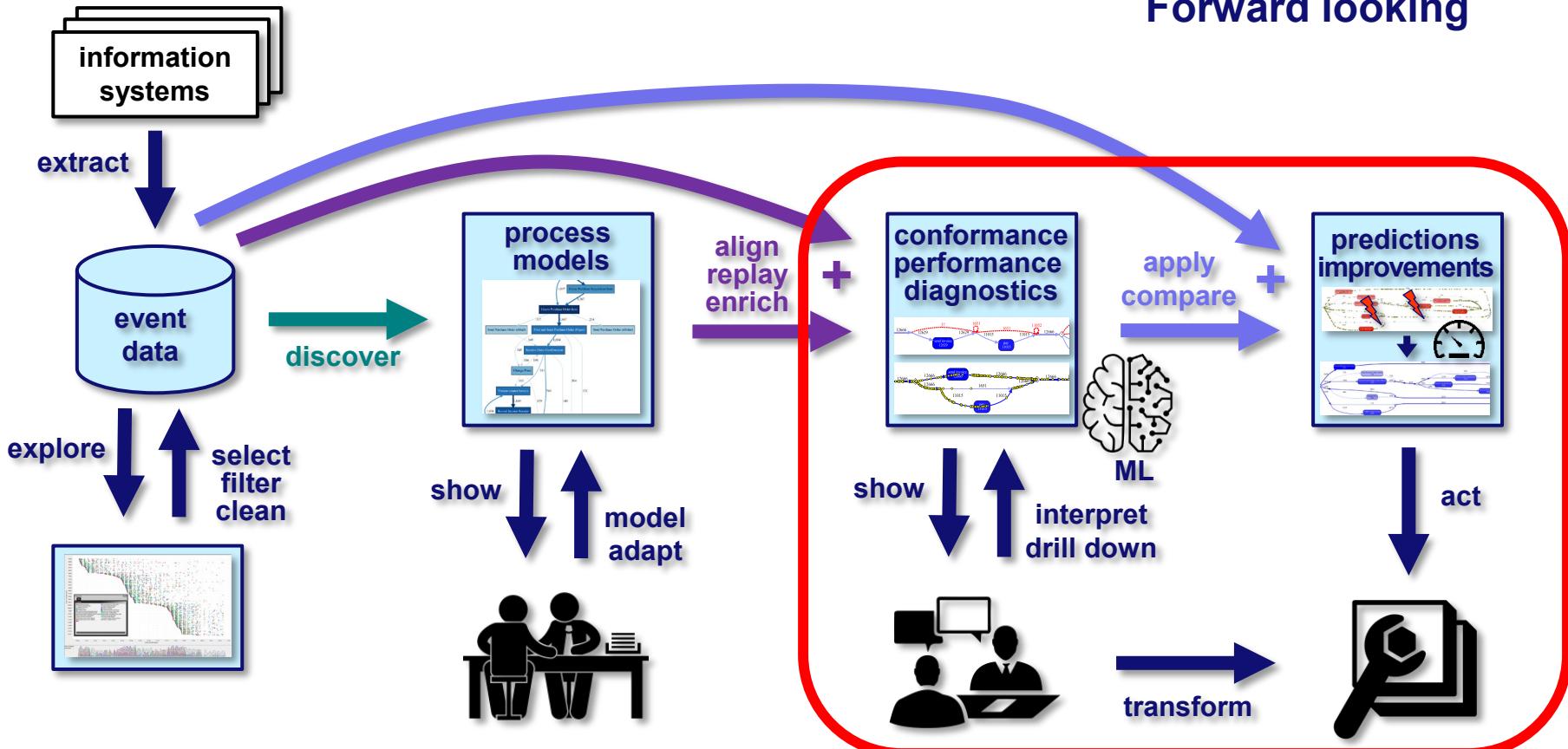
RWTHAACHEN
UNIVERSITY

One week to go ...

Lecture 17 Operational Support and Process Mining Applications	01.07.24	Monday	AH V
Lecture 18 Distributed, Streaming, and Comparative Process Mining	02.07.24	Tuesday	AH V
<i>Exercise 13 Operational Process Mining</i>	02.07.24	Tuesday	AH III
Lecture 19 Closing	08.07.24	Monday	AH V
Q&A Session Assignment Part II	09.07.24	Tuesday	AH III
Deadline Assignment Part II	14.07.24	Sunday	
Q&A Session Exam	16.07.24	Tuesday	AH III



Forward looking



Overview

- Refined process mining framework (overview of all process mining activities and different artifacts)
- Operational support
- Doing a process mining project
- Two types of processes:
 - Lasagna processes
 - Spaghetti processes



Refined Process Mining Framework





offline

Situation tables as a generic tool to link to
non-process-centric forms of analytics

also online

process discovery

(alpha miner, heuristic miner,
region-based miners, etc.)

conformance checking

(token-based, footprints, alignments, etc.)

organizational mining

bottleneck mining

decision point mining

operational support

prediction

reference model

artifact-centric mining

model repair

data quality

queue mining

...

mining of event streams

mining on partially ordered event data

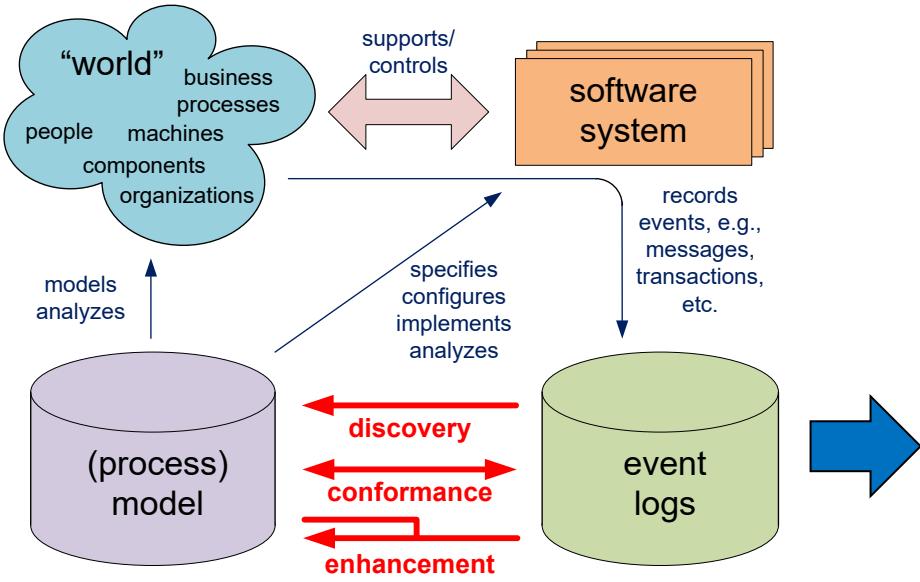
concept drift analysis

recommendation

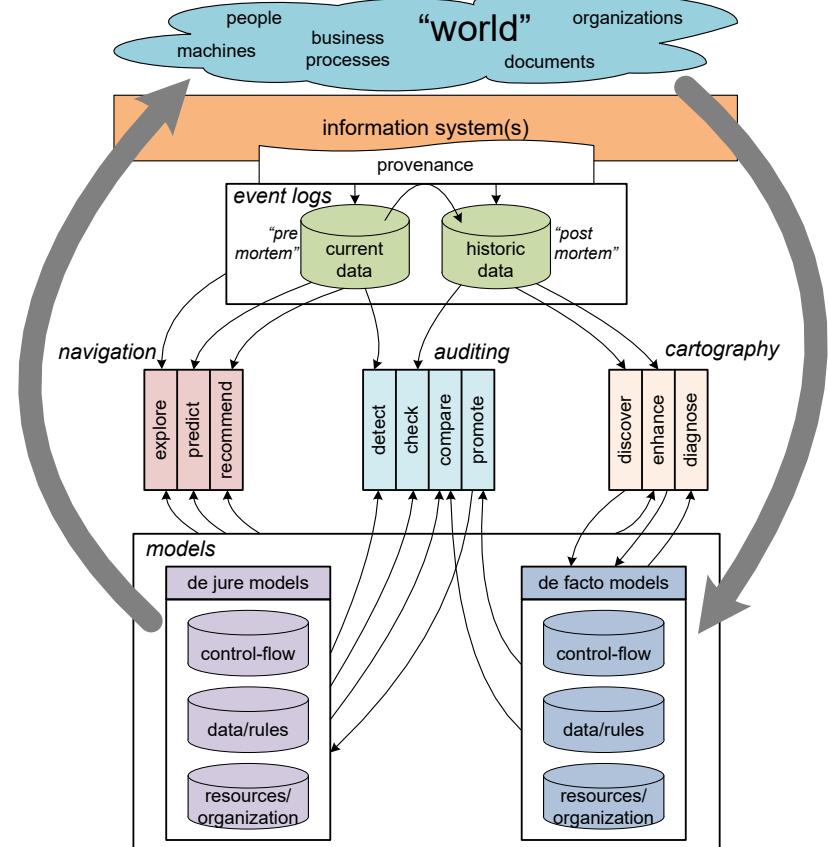
declarative mining

process configuration distributed

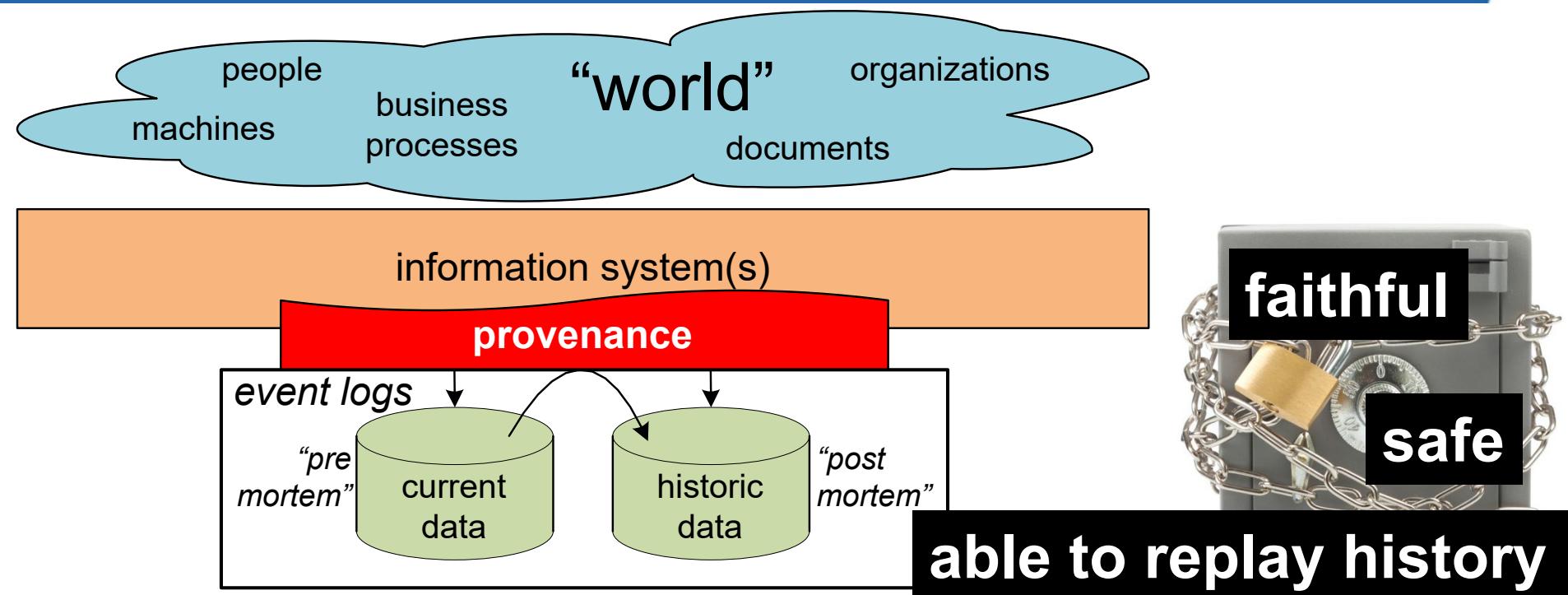
process mining



refined process mining framework

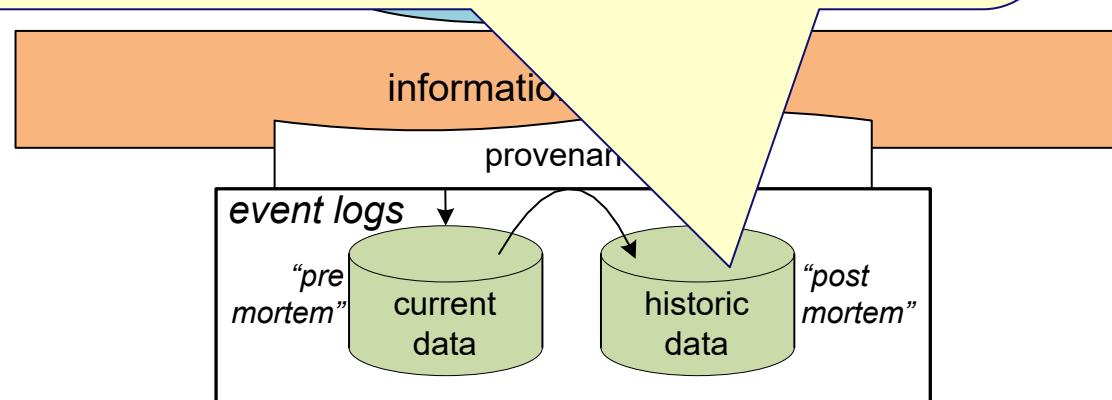


(Business) process provenance



Pre mortem and post mortem event data

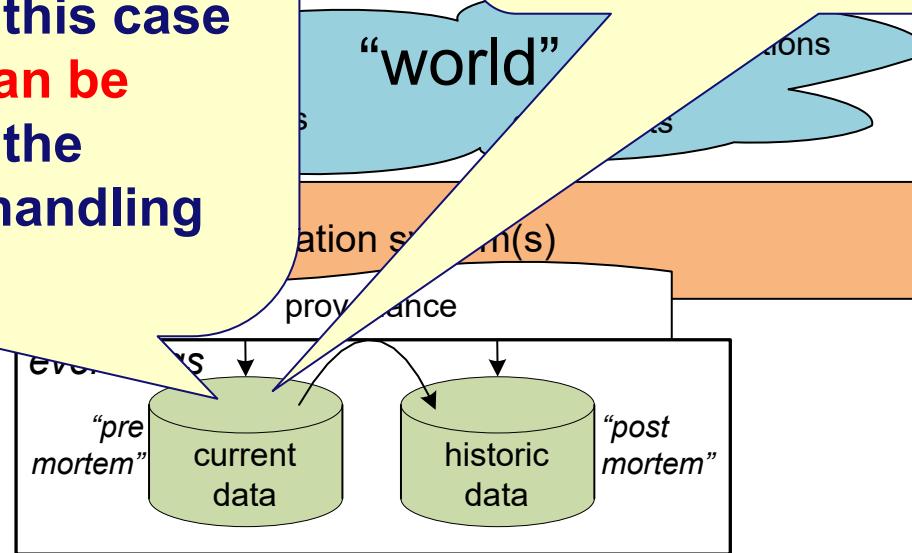
“Post mortem” event data refer to information about cases that have completed, i.e., these data can be used for process improvement and auditing, but **not for influencing** the cases they refer to.



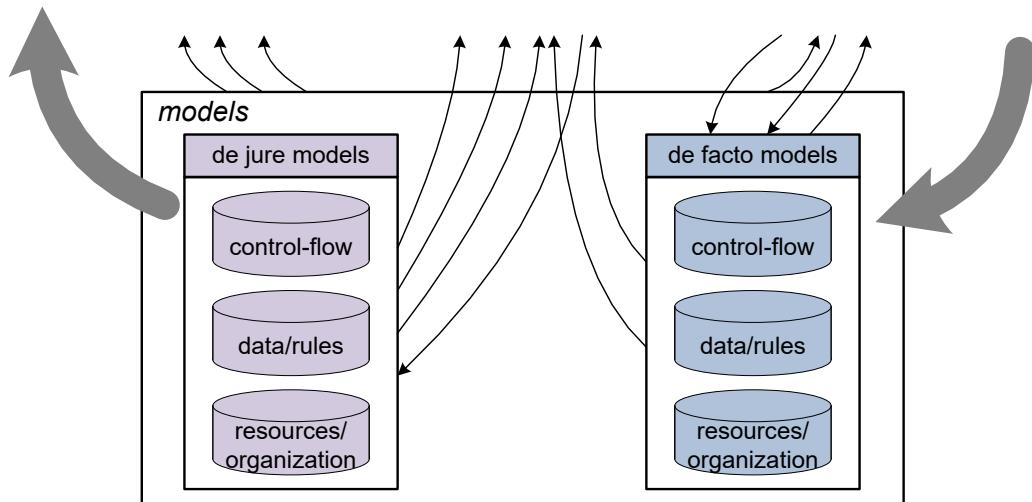
Pre mortem and post mortem event data

If a case is still running, i.e., the case is still “alive” (pre mortem), then it may be possible that information in the event log about this case (i.e., current data) can be exploited to ensure the correct or efficient handling of this case.

“Pre mortem” event data refer to cases that have not yet completed.

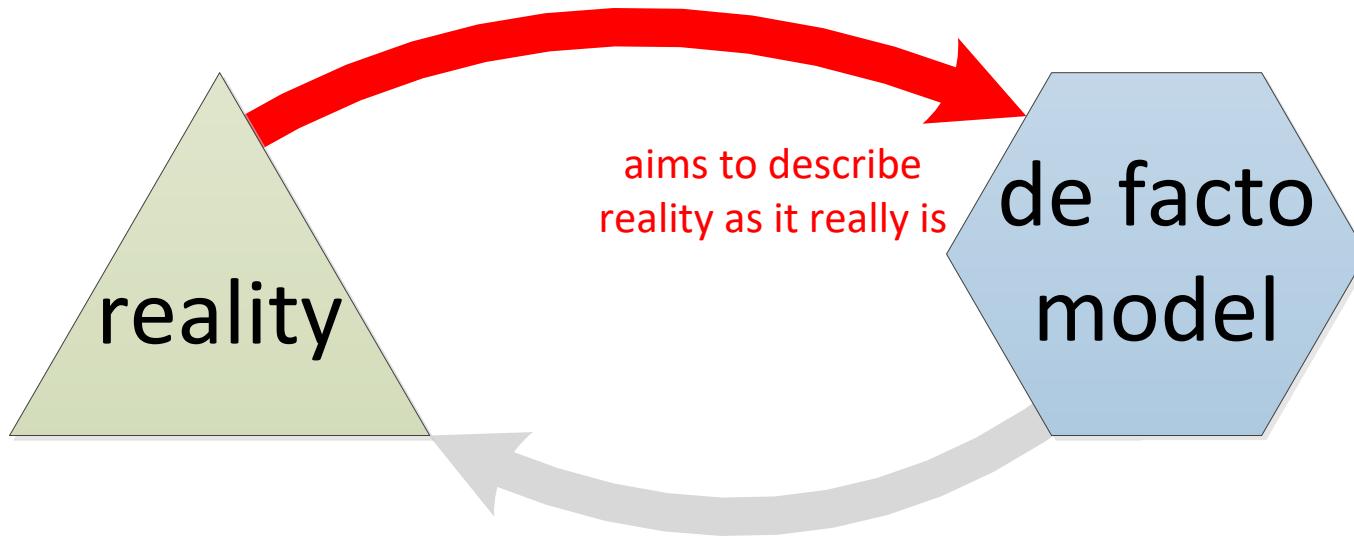


Different types of models



- Models may cover one or more perspectives
 - control-flow
 - data/rules
 - resources/org.
 - time
 - costs
 - ...
- “de jure” models and “de facto” models

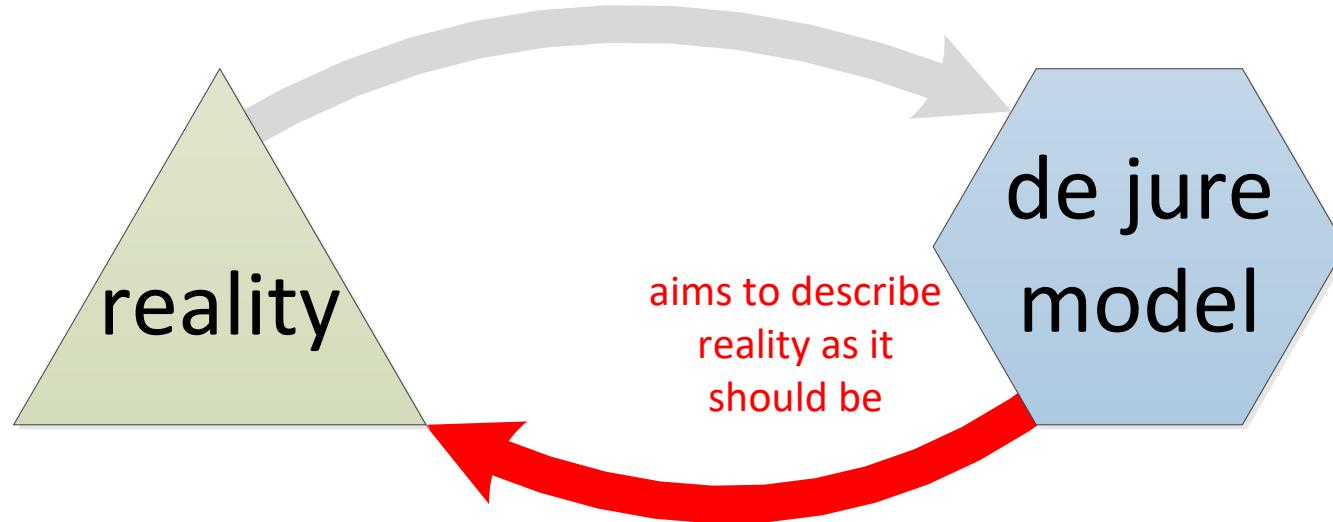
A **de facto** model is **descriptive**



- Its purpose is to describe the "as is" process, and not to steer or control reality. De facto models aim to capture reality.
- Insights may be used for reengineering, operational support, etc.

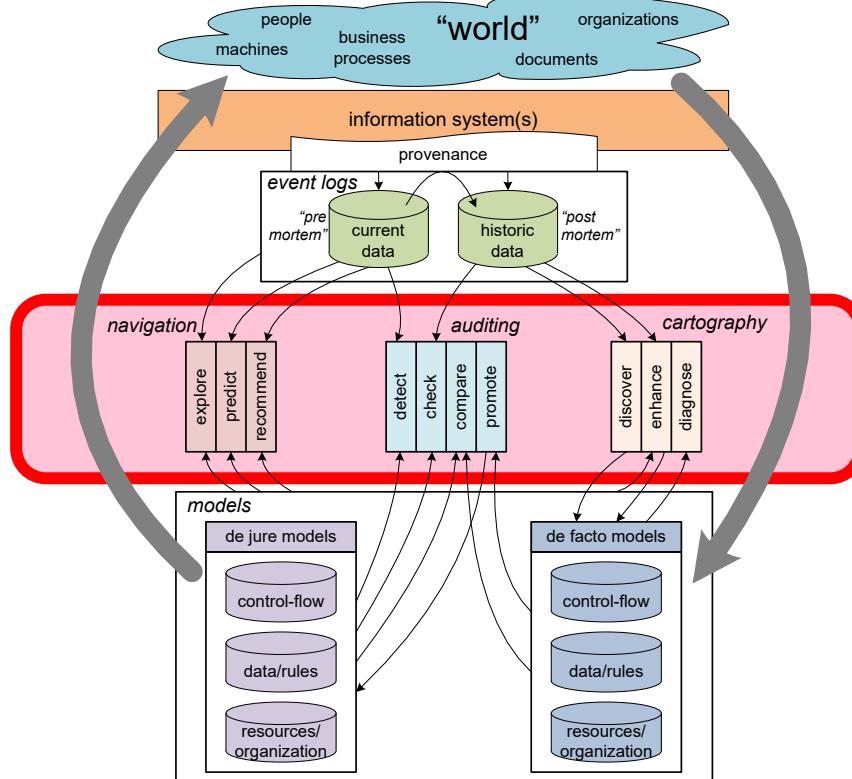
A de jure model is **normative**

It specifies how things should be done or handled



- For example, a process model used to configure a BPM system is normative and forces people to work in a particular way.
- In other situations, normative models may be ignored by workers ("wallpaper models").

Ten process mining activities



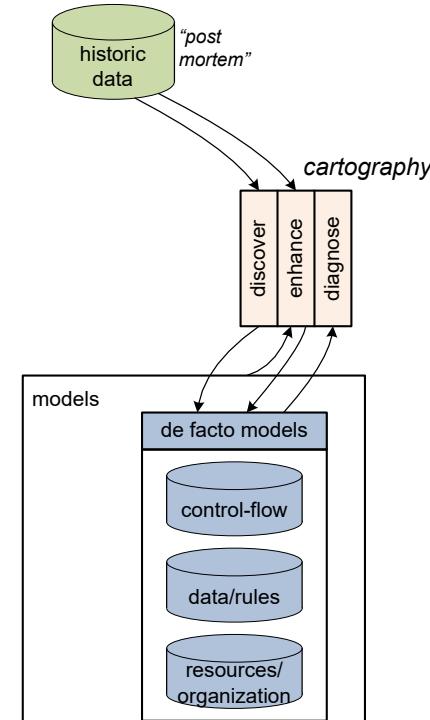
not intended
to be
complete



Chair of Process
and Data Science

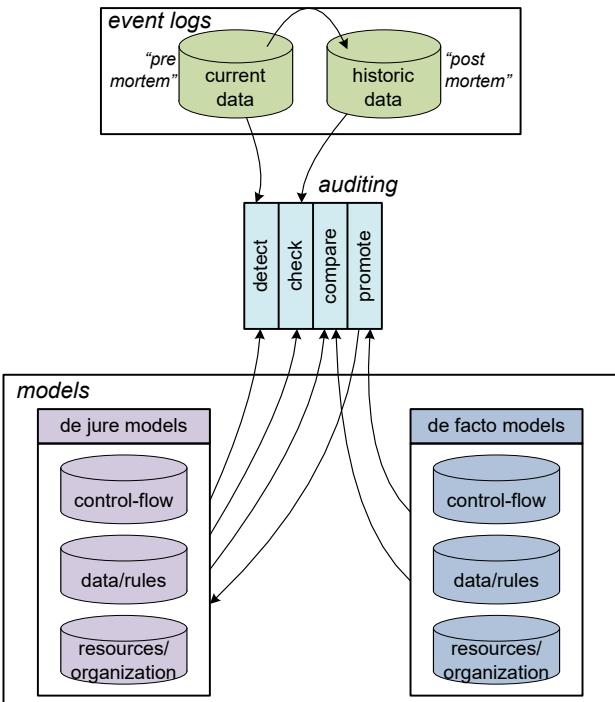
Cartography: Process models as maps

- **Discover.** This activity is concerned with the extraction of (process) models.
- **Enhance.** When existing process models (either discovered or hand-made) can be related to events logs, it is possible to enhance (extend and repair) these models.
- **Diagnose.** This activity does not directly use event logs and focuses on classical model-based analysis.



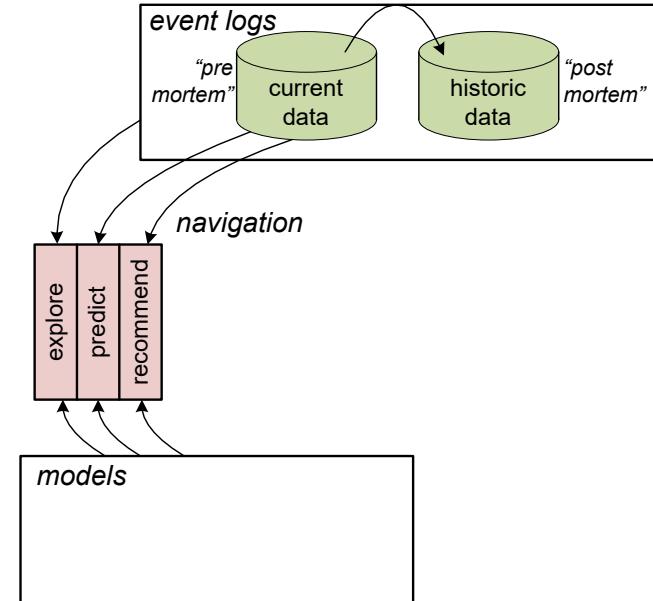
Auditing: Confronting model and reality

- **Detect.** Compares de jure models with current “pre mortem” data. The moment a predefined rule is violated, an alert is generated (**online**).
- **Check.** The goal of this activity is to pinpoint deviations and quantify the level of compliance (**offline**).
- **Compare.** De facto models can be compared with de jure models to see in what way reality deviates from what was planned or expected.
- **Promote.** Promote parts of the de facto model to a new de jure model.



Navigation: Supporting and guiding process execution

- **Explore.** The combination of event data and models can be used to explore business processes at run-time.
- **Predict.** By combining information about running cases with models, it is possible to make predictions about the future, e.g., the remaining flow time and the probability of success.
- **Recommend.** The information used for predicting the future can also be used to recommend suitable actions (e.g. to minimize costs or time).



Operational Support: Detect, Predict, and Recommend



A woman with long dark hair and a red, patterned dress with fringe and sequins is performing a belly dance pose. She is looking directly at the camera with a serious expression. Her hands are raised, showing intricate fingerwork. The background is a solid dark red.

Four generic data science questions

#1



What
happened?

#2

A woman with long dark hair, wearing a red and black patterned top and matching pants, is shown from the waist up. She has her hands raised, fingers spread, in a dramatic pose. A large white arrow points from the right side of the frame towards her head. The background is dark red.

Why did
it happen?

#3



What will
happen?

#4

A woman with long dark hair, wearing a red patterned dress and black fingerless gloves, stands with her arms outstretched. A white circle is positioned around her torso, containing the text.

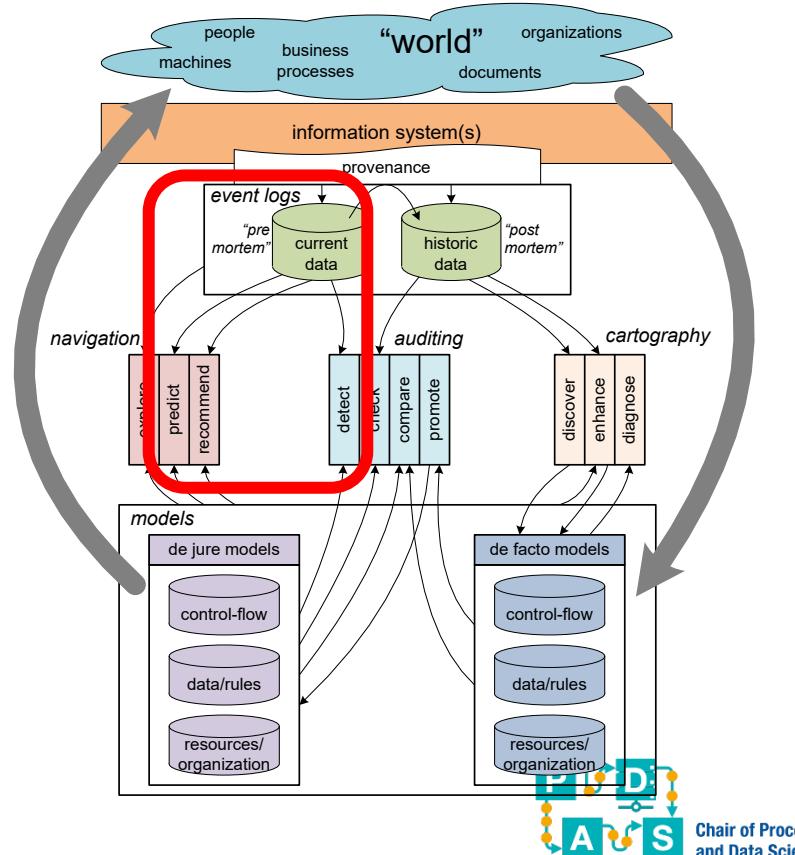
Action!

What is
the best that
can happen?

Operational support

Focus on pre mortem data

- **Detect**
 - Something is going wrong **now!**
 - This case is deviating **now!**
 - The deadline **just expired!**
- **Predict**
 - When **will** the case finish?
 - **Will** the case be rejected?
 - **Will** the case deviate?
- **Recommend**
 - Which activity **should** be executed?
 - Who **should** execute it?

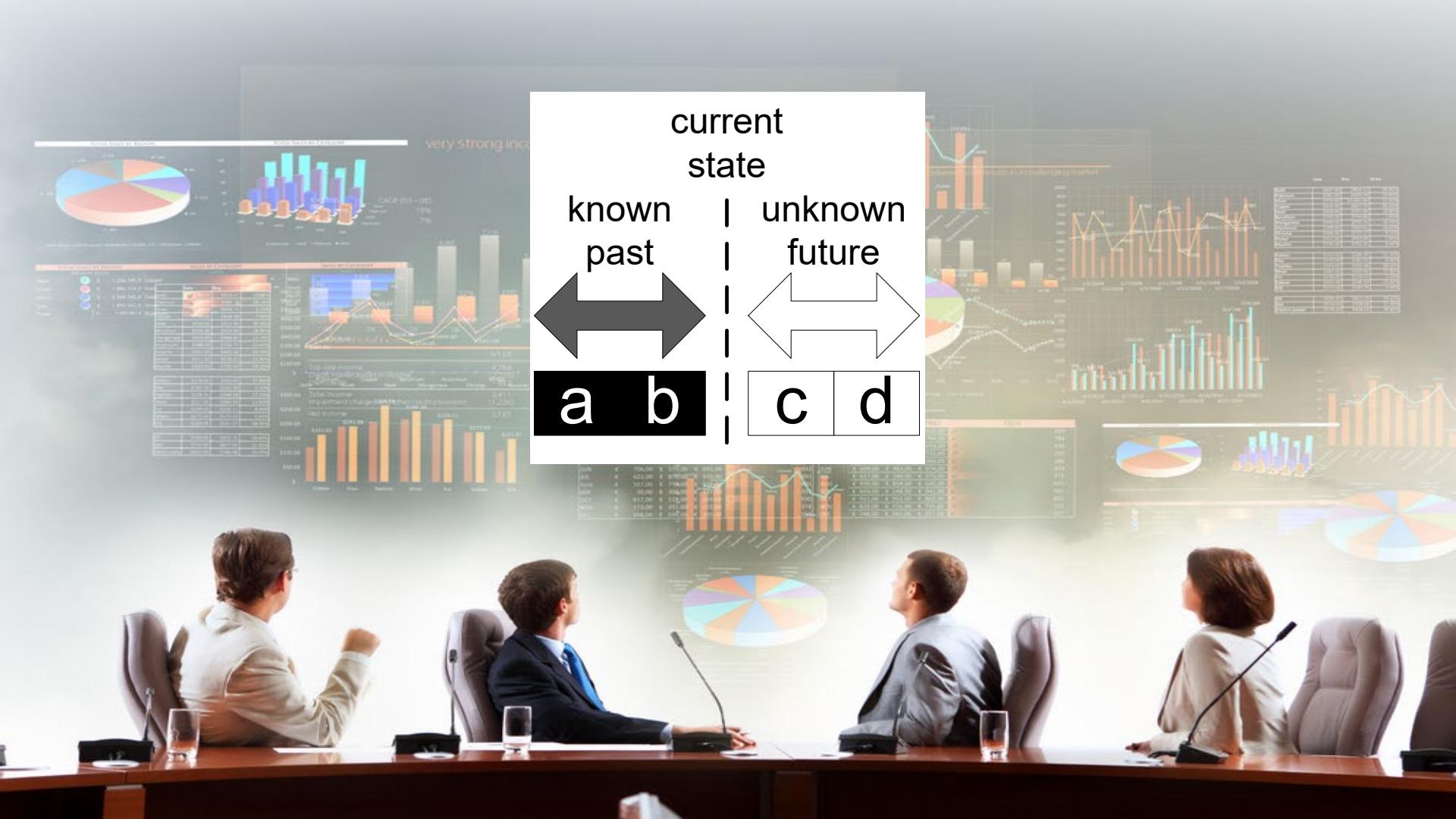


current
state

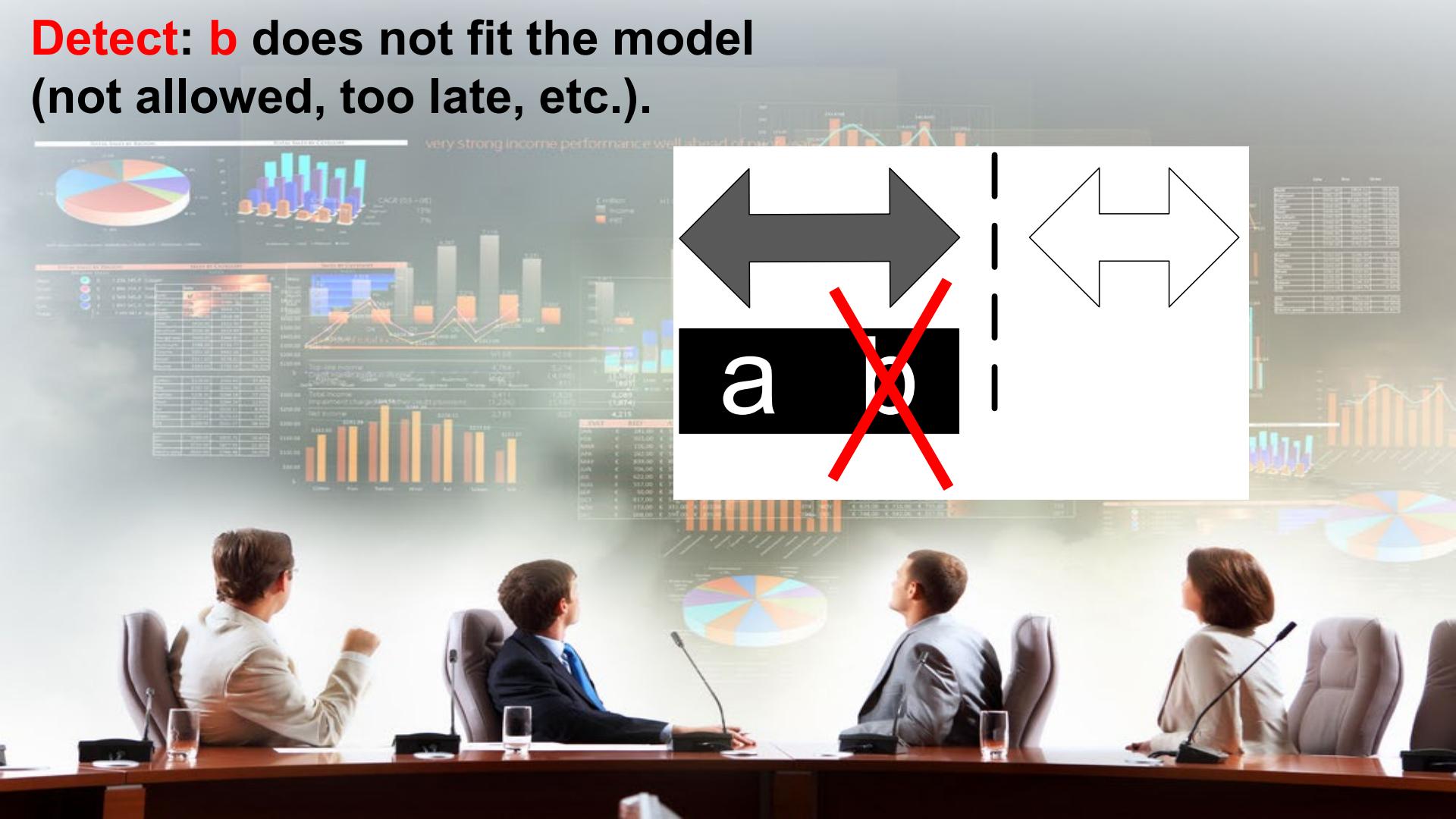
known | unknown
past future

a b

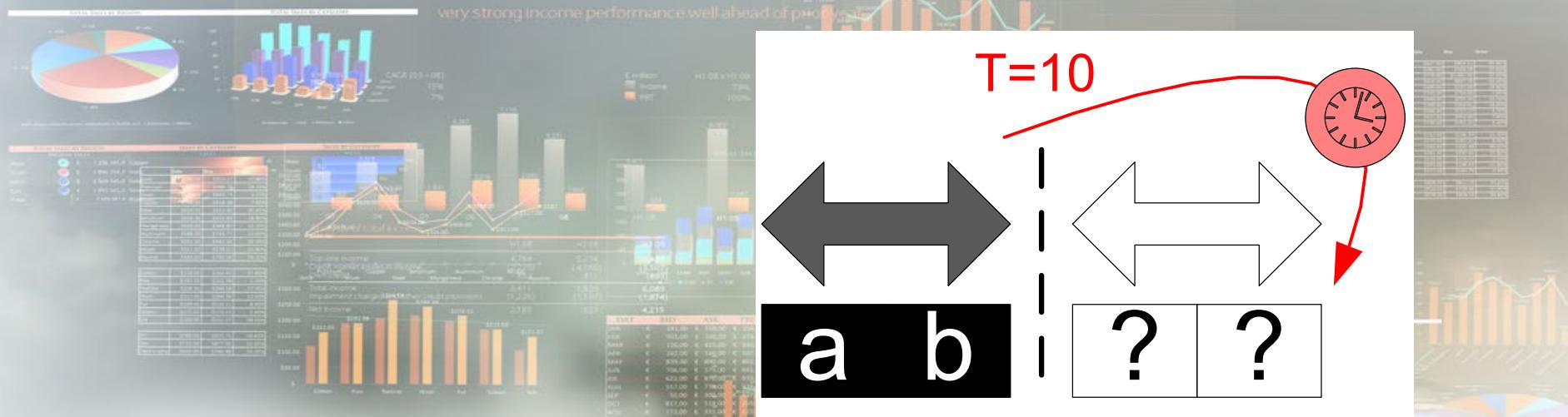
c d



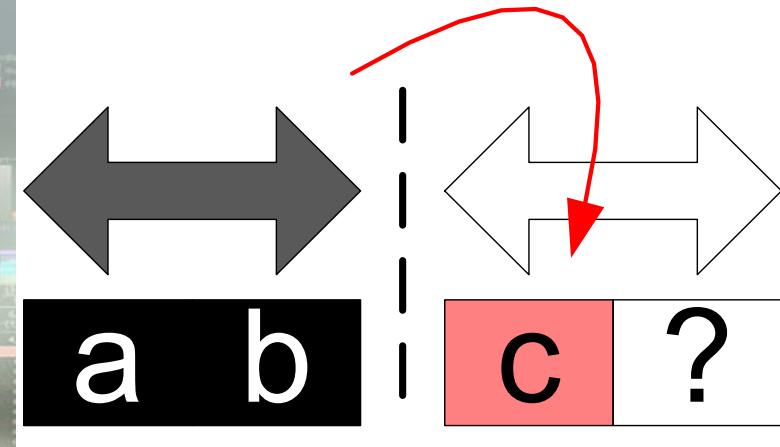
Detect: b does not fit the model (not allowed, too late, etc.).



Predict: some prediction is made about the future (e.g. completion date or outcome).

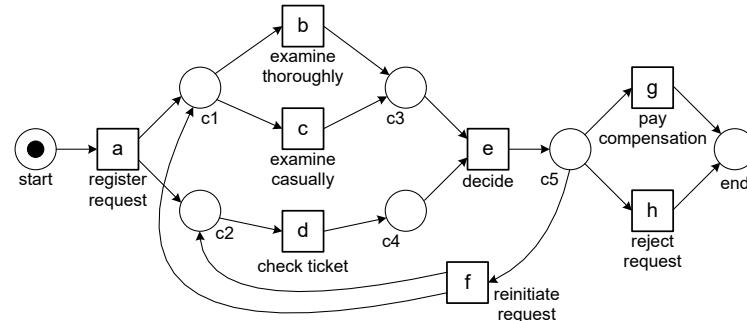


Recommend: based on past experiences **c** is recommended (e.g., to minimize costs).

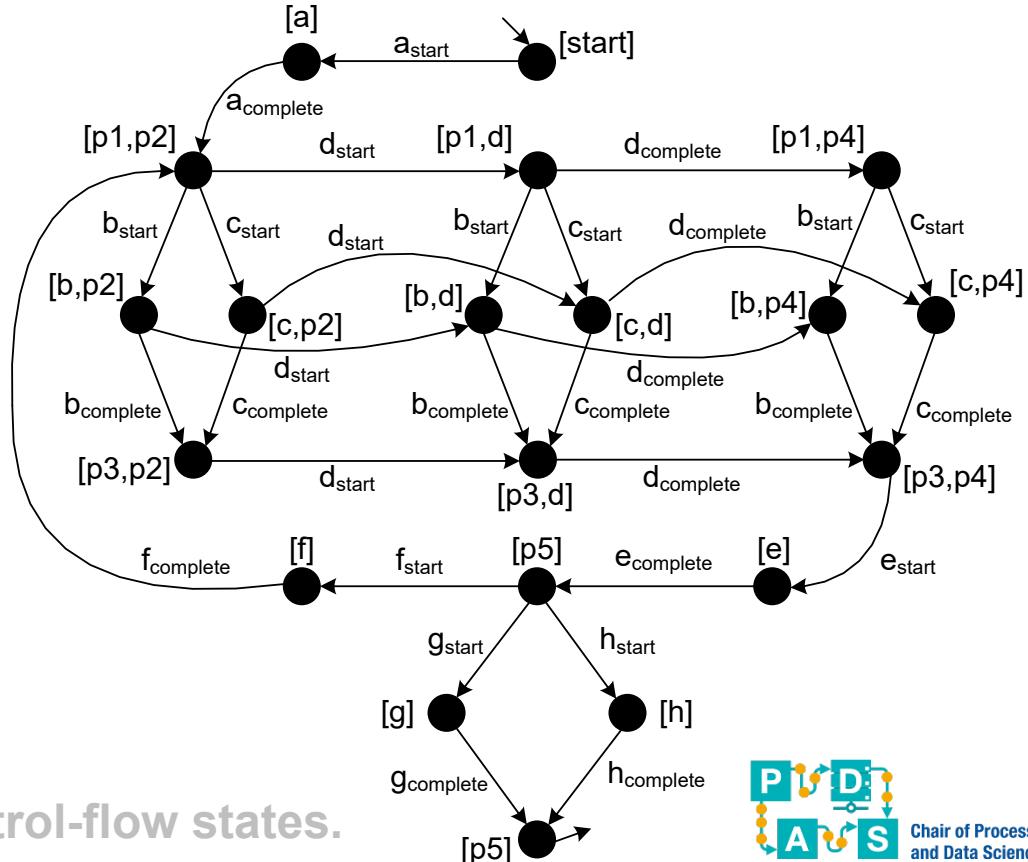
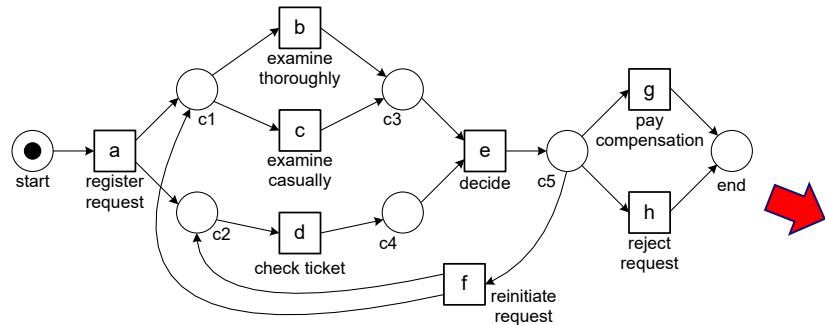


Running example for simplicity we focus on control-flow and time

case id	trace
1	$\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26}, b_{complete}^{32}, d_{complete}^{33}, e_{start}^{35}, e_{complete}^{40}, h_{start}^{50}, h_{complete}^{54} \rangle$
2	$\langle a_{start}^{17}, a_{complete}^{23}, d_{start}^{28}, c_{start}^{30}, d_{complete}^{32}, c_{complete}^{38}, e_{start}^{50}, e_{complete}^{59}, g_{start}^{70}, g_{complete}^{73} \rangle$
3	$\langle a_{start}^{25}, a_{complete}^{30}, c_{start}^{32}, c_{complete}^{35}, d_{start}^{35}, d_{complete}^{40}, e_{start}^{45}, e_{complete}^{50}, f_{start}^{50}, f_{complete}^{55}, b_{start}^{60}, d_{start}^{62}, b_{complete}^{65}, d_{complete}^{67}, e_{start}^{80}, e_{complete}^{87}, g_{start}^{90}, g_{complete}^{98} \rangle$
...	...

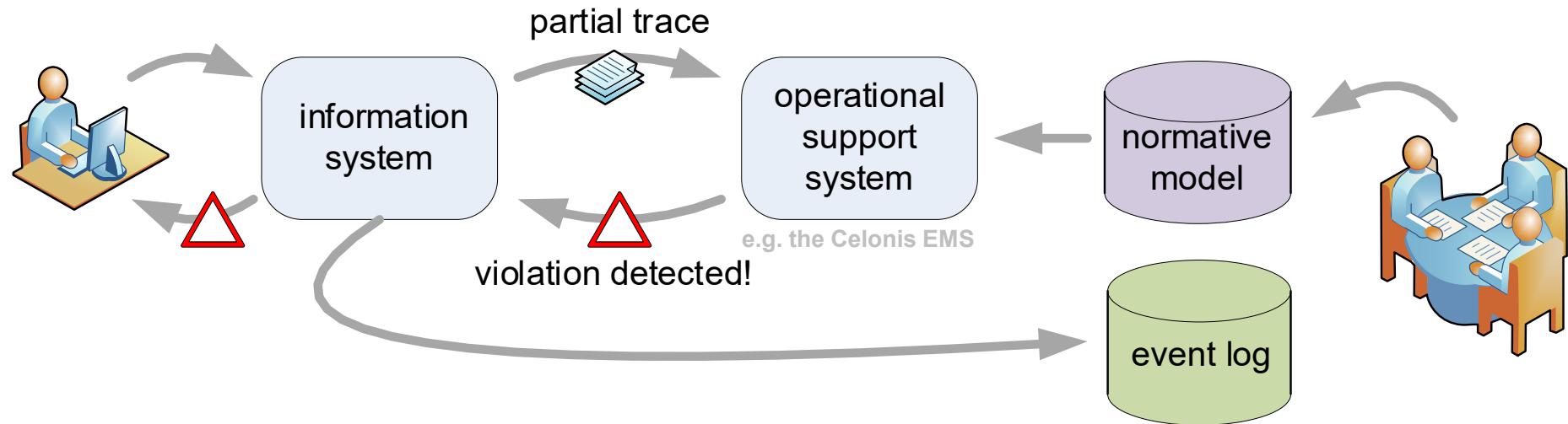


Transition system

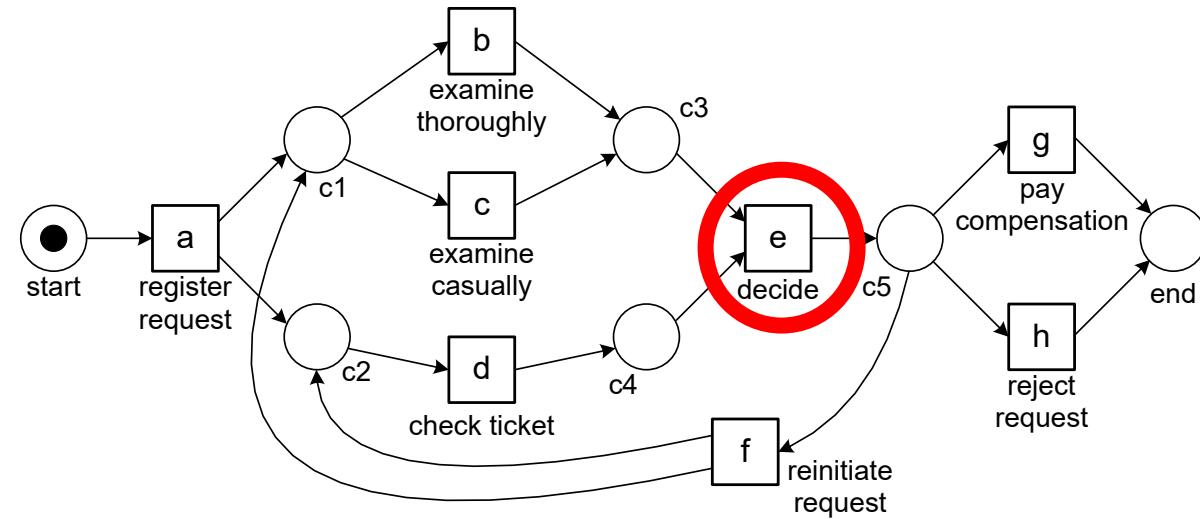


Lists all control-flow states.

Operational support: Detect



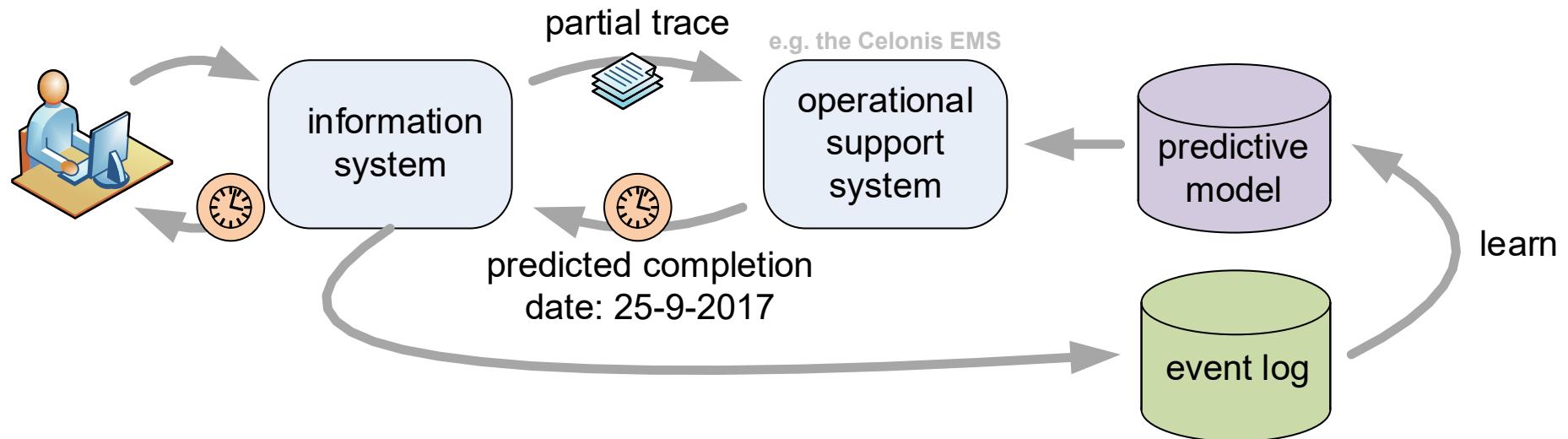
Example



alert @28!!!!

$$\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{27}, e_{start}^{28}, \dots \rangle$$

Operational support: Predict



Examples of predictions

The predicted remaining flow time for this case is 14 days.

The predicted probability of meeting the legal deadline is 0.72 for this case.

The predicted probability that person r will work on this case is 0.57.

The predicted probability that activity a will occur is 0.34.

The predicted total cost of this case is €4500.

The predicted probability that this case will be rejected is 0.67.

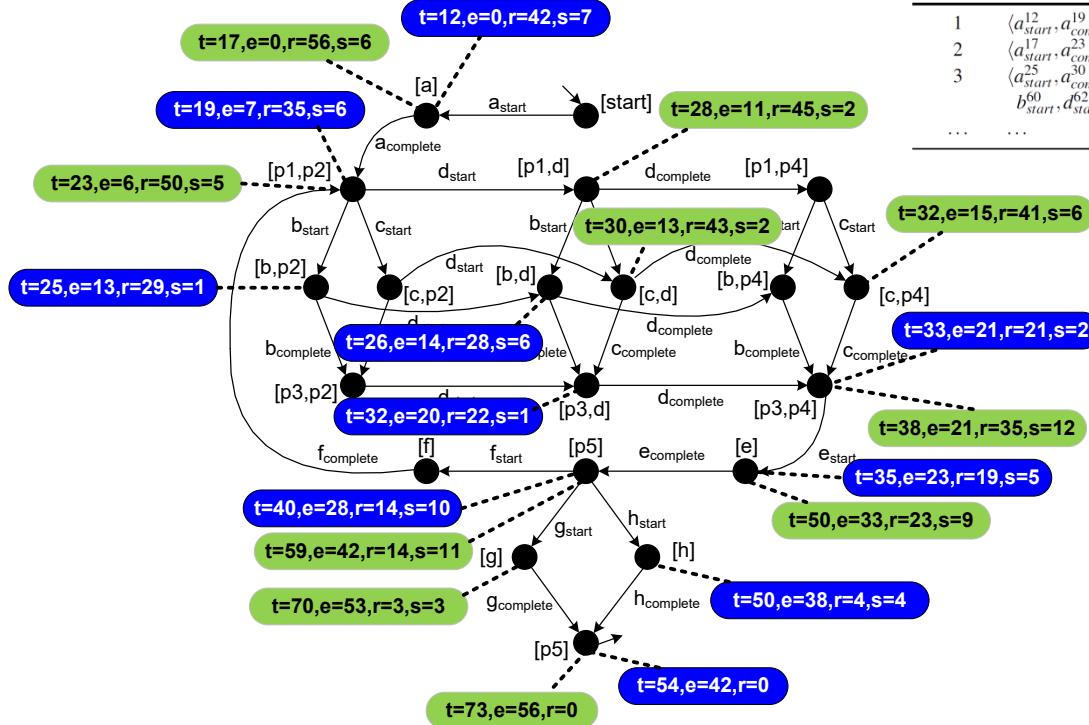


Annotated transition system

(based on replay with time, see previous lectures)

case 1

case 2



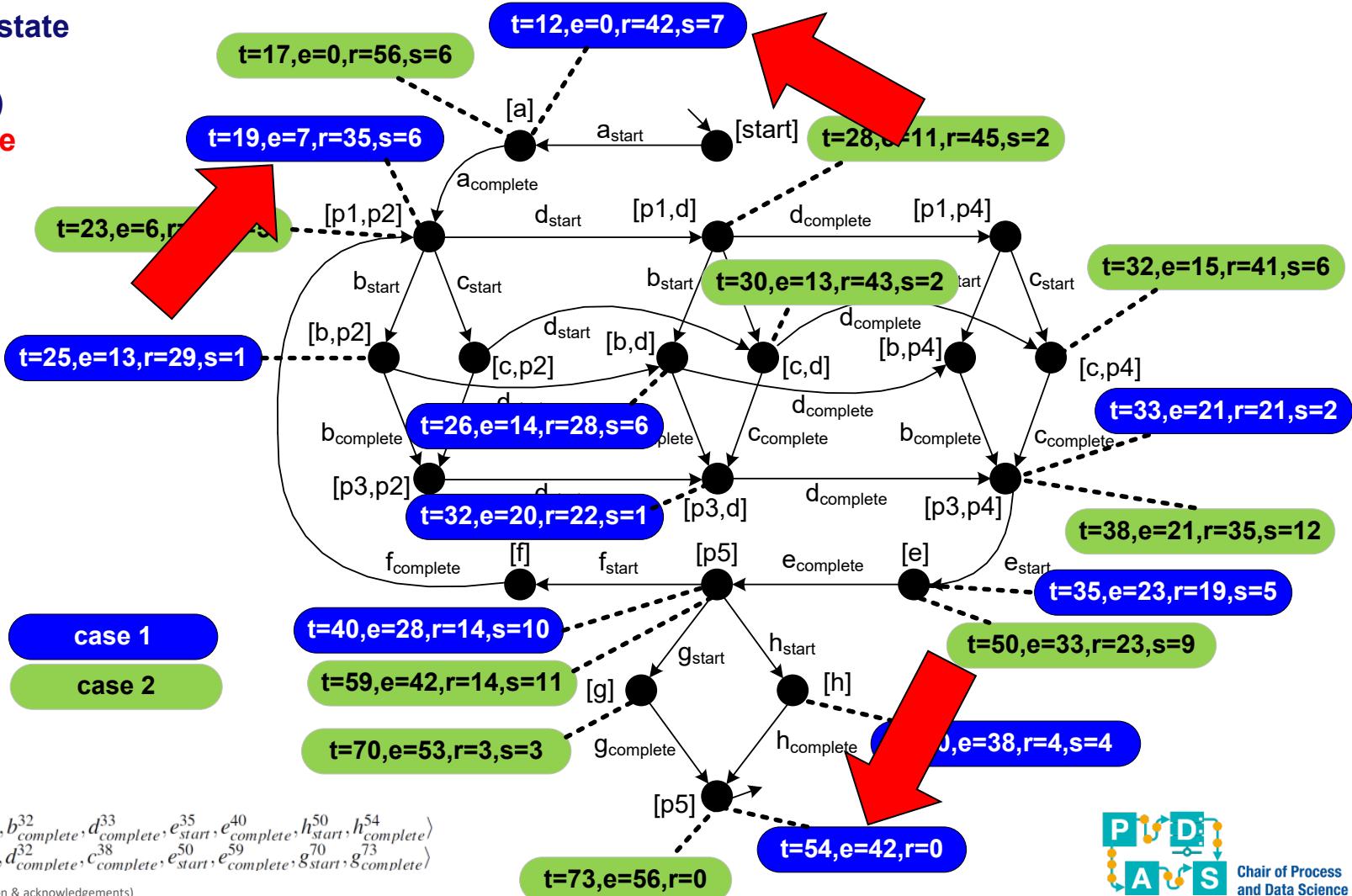
case id trace

1	$\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26}, b_{complete}^{32}, d_{complete}^{33}, e_{start}^{35}, e_{complete}^{40}, h_{start}^{50}, h_{complete}^{54} \rangle$
2	$\langle a_{start}^{17}, a_{complete}^{23}, d_{start}^{28}, c_{start}^{30}, d_{complete}^{32}, c_{complete}^{38}, e_{start}^{50}, e_{complete}^{59}, g_{start}^{70}, g_{complete}^{73} \rangle$
3	$\langle a_{start}^{25}, a_{complete}^{30}, c_{start}^{32}, c_{complete}^{35}, d_{start}^{35}, d_{complete}^{40}, e_{start}^{45}, e_{complete}^{50}, f_{start}^{50}, f_{complete}^{55}, b_{start}^{60}, b_{complete}^{65}, d_{start}^{67}, d_{complete}^{80}, e_{start}^{87}, e_{complete}^{90}, g_{start}^{98}, g_{complete}^{99} \rangle$
...	...

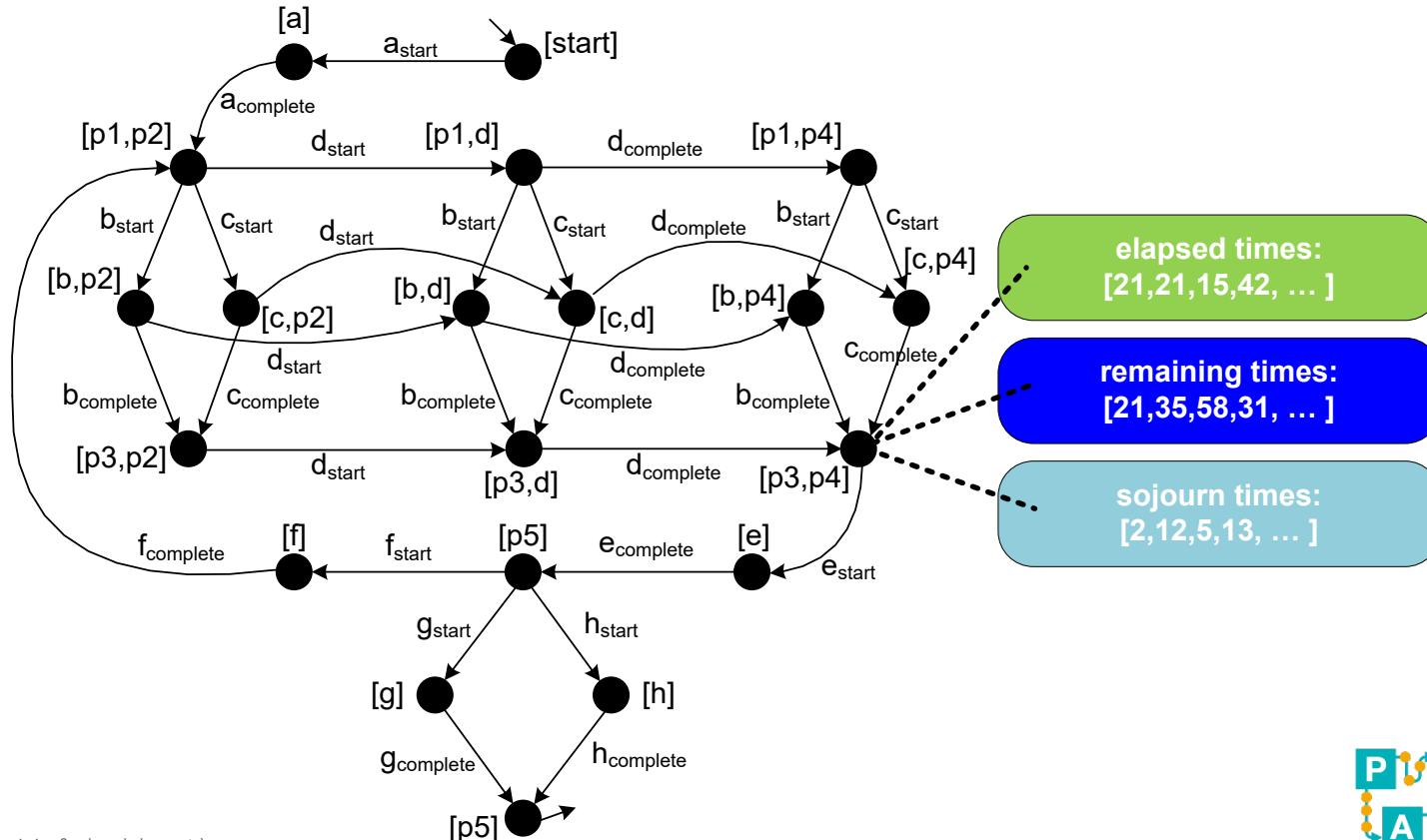
- **t = time entering state**
- **e = elapsed time (since first event)**
- **r = remaining time (until last event)**
- **s = sojourn time (time in state)**



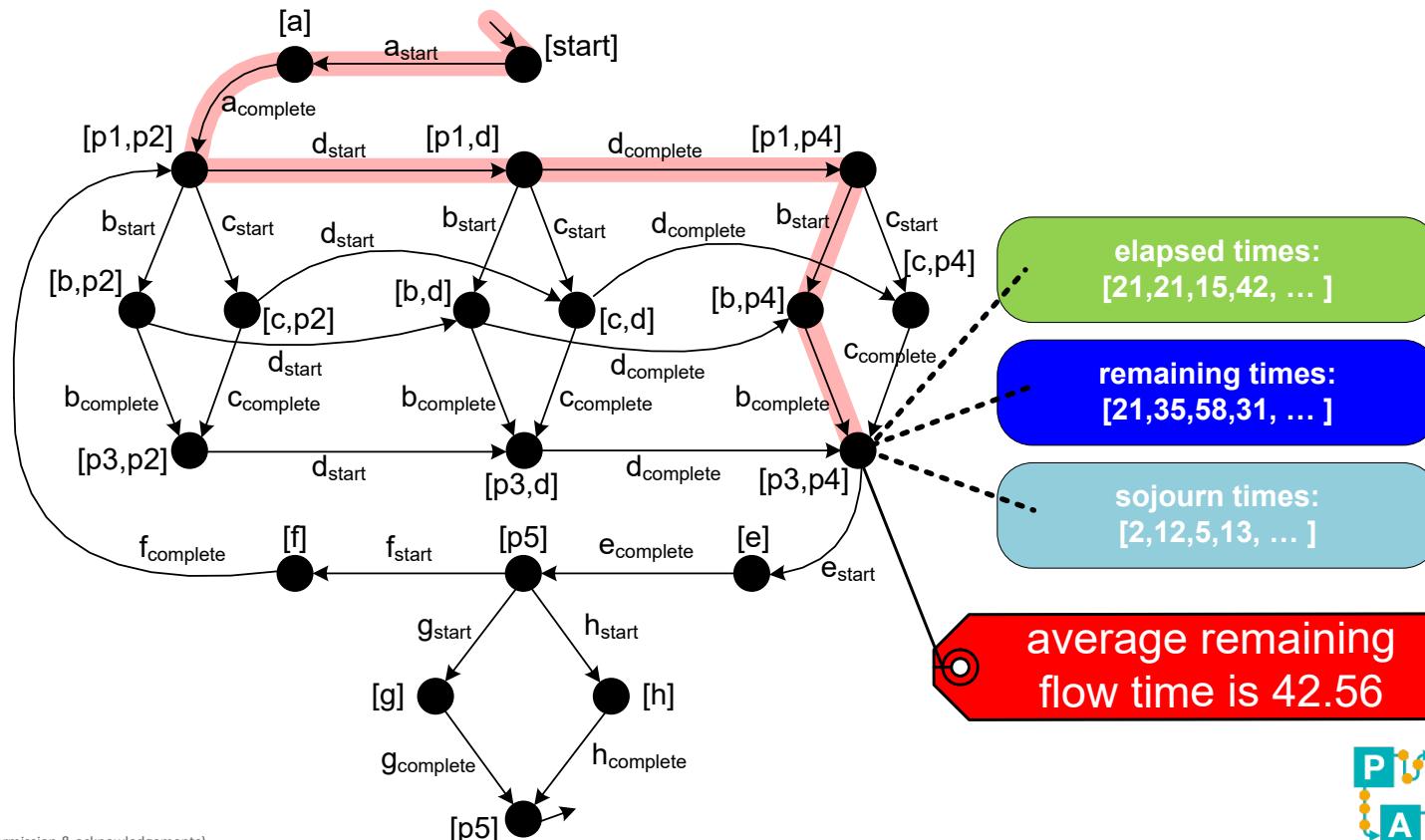
t = time entering state
e = elapsed time
 (since first event)
r = remaining time
 (until last event)
s = sojourn time
 (time in state)



Collect results per state

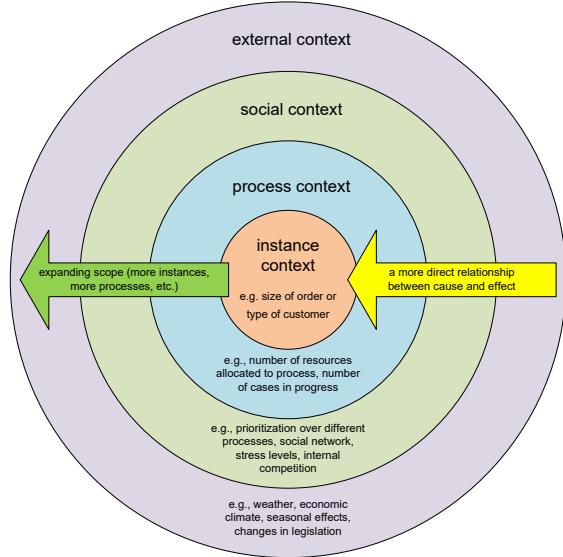


Predict based on current state

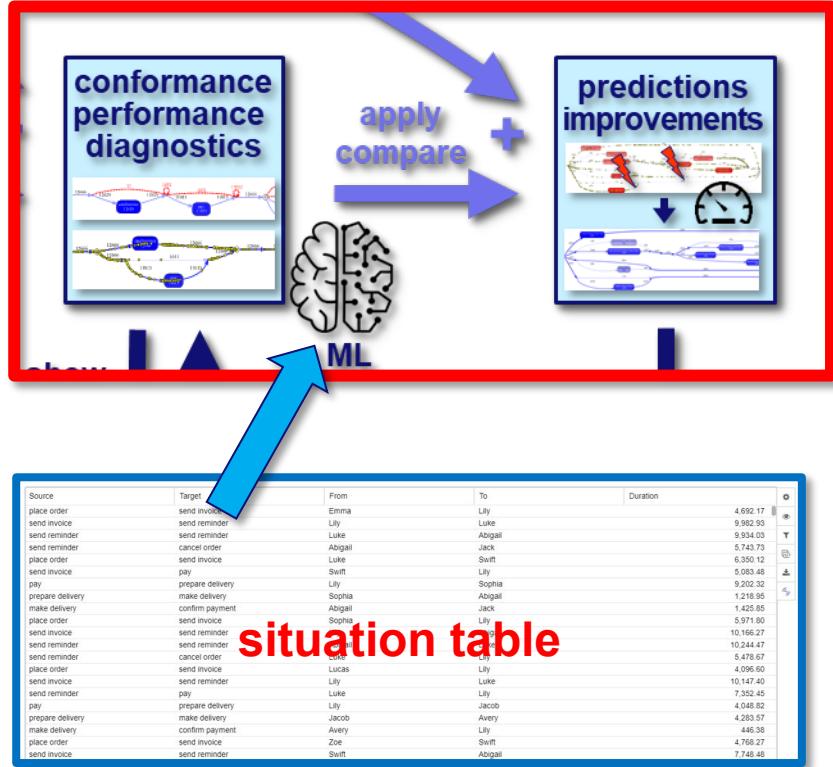




Use context information!



Recall: Situation Tables (e.g. in Celonis)



- A **situation table** is a two-dimensional table.
 - Each row is an instance.
 - Each column is a variable.
 - There may be a split into a response variable and predictor variables (for supervised learning).
- Five types of situation tables:
 - **Case-based situation table:** Each row (instance) corresponds to a case with variables.
 - **Event-based situation table:** Each row (instance) corresponds to an event.
 - **Resource-based situation table:** Each row (instance) corresponds to a resource.
 - **Event-pair-based situation table:** Each row (instance) corresponds to a pair of events.
 - **Aggregate situation tables:** Each row (instance) corresponds to a combination of cases and/or events.



Remaining flow time

We are interested in predicting the **remaining flow time** for running cases.

Answer the following questions:

- What kind of situation table? (case-based, event-based, resource-based, event-pair-based, or else)
- What is the response variable?
- What are predictor variables?

Remaining flow time

- **What kind of situation table? Event based**
(look up the last event for the case you want to predict)
- **What is the response variable? Time of last event for the case minus the time of current event.**
- **What are predictor variables? Many possibilities:**
 - Number of times resource R performed activities for the same case.
 - Was activity A executed?
 - Number of cases running.
 - Time since case start.
 - Monetary value of the case.
 - Supplier, location, weather, etc.

See the PQL commands shown before that allow you to add the variables.



Example

PRODUCT	QUANTITY	ADDRESS	Last activity	Pay immediately	Remaining time (hours)
SAMSUNG Galaxy J5	3	Munich	place order	0	239
SAMSUNG Galaxy J5	3	Munich	send invoice	0	165
SAMSUNG Galaxy J5	3	Munich	pay	0	134
SAMSUNG Galaxy J5	3	Munich	prepare delivery	0	18
SAMSUNG Galaxy J5	3	Munich	make delivery	0	14
SAMSUNG Galaxy J5	3	Munich	confirm payment	0	0
APPLE iPhone 6s 64 GB	2	Amsterdam	place order	1	201
APPLE iPhone 6s 64 GB	2	Amsterdam	pay	1	103
APPLE iPhone 6s 64 GB	2	Amsterdam	send invoice	1	32
APPLE iPhone 6s 64 GB	2	Amsterdam	prepare delivery	1	8
APPLE iPhone 6s 64 GB	2	Amsterdam	confirm payment	1	3
APPLE iPhone 6s 64 GB	2	Amsterdam	make delivery	1	0
APPLE iPhone 5s 16 GB	6	New York	place order	0	503
APPLE iPhone 5s 16 GB	6	New York	send invoice	0	408
APPLE iPhone 5s 16 GB	6	New York	send reminder	0	239
APPLE iPhone 5s 16 GB	6	New York	send reminder	0	68
APPLE iPhone 5s 16 GB	6	New York	cancel order	0	0
MOTOROLA Moto E 4G	1	New York	place order	0	498
MOTOROLA Moto E 4G	1	New York	send invoice	0	451
MOTOROLA Moto E 4G	1	New York	send reminder	0	280
MOTOROLA Moto E 4G	1	New York	send reminder	0	110
MOTOROLA Moto E 4G	1	New York	cancel order	0	0
SAMSUNG Core Prime G361	6	Aachen	place order	0	741
SAMSUNG Core Prime G361	6	Aachen	send invoice	0	574
SAMSUNG Core Prime G361	6	Aachen	send reminder	0	104

Example (see previous lectures!)

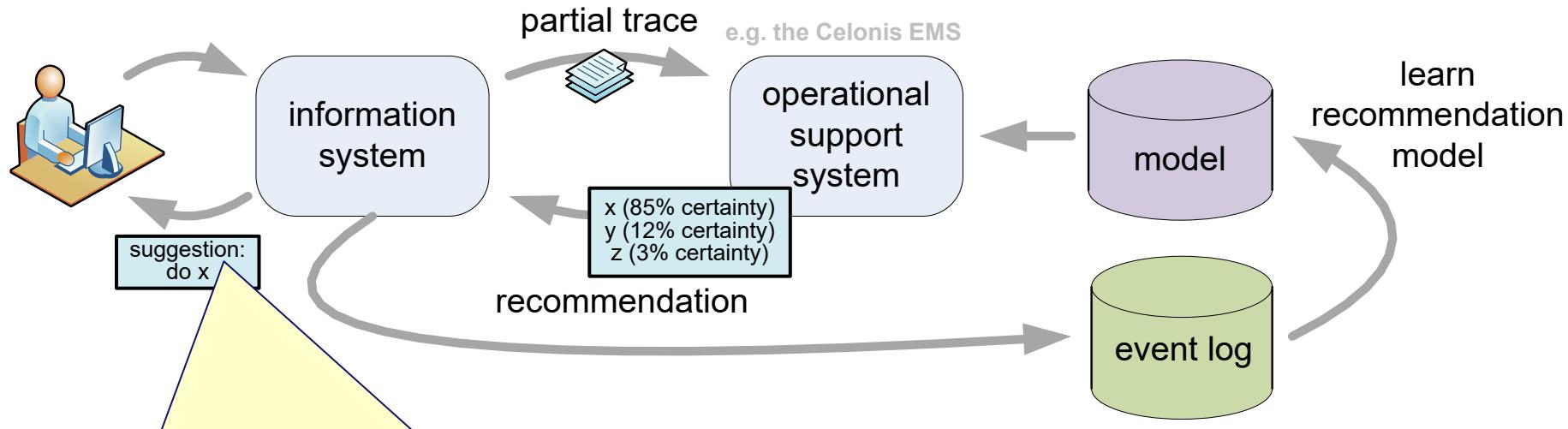
PRODUCT	QUANTITY	ADDRESS	Last activity	Pay immediately	Remaining time (hours)
SAMSUNG Galaxy J5	3	Munich	place order	0	239
SAMSUNG Galaxy J5	3	Munich	send invoice	0	165
SAMSUNG Galaxy J5	3	Munich	pay	0	134
SAMSUNG Galaxy J5	3	Munich	prepare	0	18
SAMSUNG Galaxy J5	3	Munich	make delivery	0	14
SAMSUNG Galaxy J5	3	Munich	confirm	0	0
APPLE iPhone 6s 64 GB	2	Amsterdam	place order	1	201
APPLE iPhone 6s 64 GB	2	Amsterdam	pay	1	103
APPLE iPhone 6s 64 GB	2	Amsterdam	send invoice	1	32
APPLE iPhone 6s 64 GB	2	Amsterdam	prepare	1	8
APPLE iPhone 6s 64 GB	2	Amsterdam	confirm	1	3
APPLE iPhone 6s 64 GB	2	Amsterdam	make delivery	1	0
APPLE iPhone 6s 16 GB	6	New York	place order	0	503
APPLE iPhone 6s 16 GB	6	New York	send invoice	0	408
APPLE iPhone 6s 16 GB	6	New York	send reminder	0	239
APPLE iPhone 6s 16 GB	6	New York	send reminder	0	68
APPLE iPhone 5s 16 GB	6	New York	cancel order	0	0
"cases"."PRODUCT" "cases"."QUANTITY" "cases"."ADDRESS" "events"."ACTIVITY" 0 0					
MATCH_PROCESS_REGEX ("events"."ACTIVITY", 'place order'>>'pay')					
MOTOROLA Moto E 4G	1	New York	send reminder	0	0

REMAP_TIMESTAMP(*PU_LAST("cases", "events"."START TIME")***,HOURS) -**
REMAP_TIMESTAMP("events"."START TIME",HOURS)

For example, predict the remaining flow time

PRODUCT	QUANTITY	ADDRESS	Last activity	Pay immediately	Remaining time (hours)
SAMSUNG Galaxy J5	3	Munich	place order	0	239
SAMSUNG Galaxy J5	3	Munich	send invoice	0	165
SAMSUNG Galaxy J5	3	Munich	pay	0	134
SAMSUNG Galaxy J5	3	Munich	prepare delivery	0	18
SAMSUNG Galaxy J5	3	Munich	make delivery	0	14
SAMSUNG Galaxy J5	3	Munich	confirm payment	0	0
APPLE iPhone 6s 64 GB	2	Amsterdam	place order	1	201
APPLE iPhone 6s 64 GB	2	Amsterdam	pay	1	103
APPLE iPhone 6s 64 GB	2	Amsterdam	send invoice	1	32
APPLE iPhone 6s 64 GB	2	Amsterdam	prepare delivery	1	8
APPLE iPhone 6s 64 GB	2	Amsterdam	confirm payment	1	3
APPLE iPhone 6s 64 GB	2	Amsterdam	make delivery	1	0
APPLE iPhone 5s 16 GB	6	New York	place order	0	503
APPLE iPhone 5s 16 GB	6	New York	send invoice	0	408
APPLE iPhone 5s 16 GB	6	New York	send reminder	0	239
APPLE iPhone 5s 16 GB	6	New York	send reminder	0	68
APPLE iPhone 5s 16 GB	6	New York	cancel order	0	0
MOTOROLA Moto E 4G	1	New York	place order	0	498
MOTOROLA Moto E 4G	1	New York	send invoice	0	451
MOTOROLA Moto E 4G	1	New York	send reminder	0	280
MOTOROLA Moto E 4G	1	New York	send reminder	0	110
MOTOROLA Moto E 4G	1	New York	cancel order	0	0
SAMSUNG Core Prime G361	6	Aachen	place order	0	741
SAMSUNG Core Prime G361	6	Aachen	send invoice	0	574
SAMSUNG Core Prime G361	6	Aachen	send reminder	0	104

Operational support: Recommend



Typical recommendations:

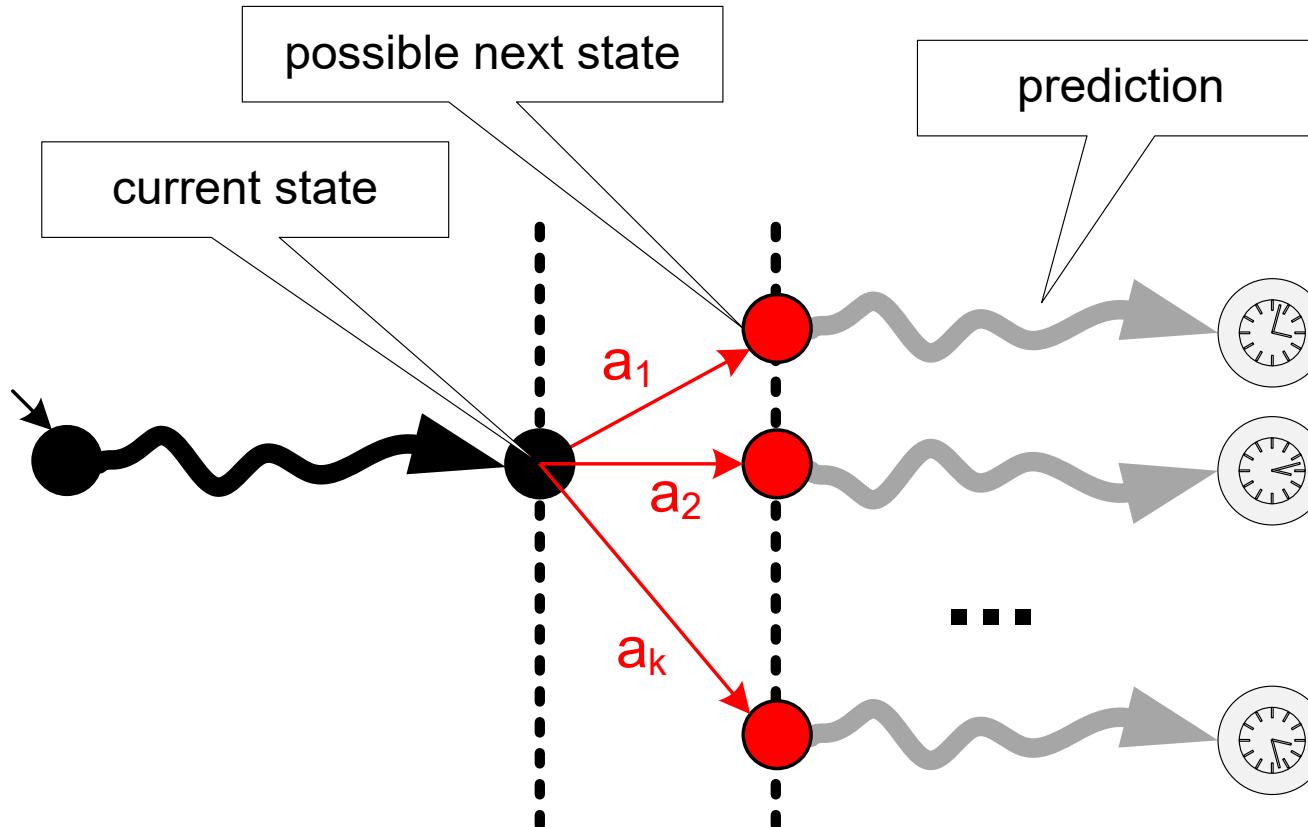
- Next activity (choice or ordering related).
- Suitable resource.

A recommendation is always given with respect to a specific goal

- Minimize the remaining flow time.
- Minimize the total costs.
- Maximize the fraction of cases handled within 4 weeks.
- Maximize the fraction of cases that is accepted.
- Minimize resource usage.



Relation between prediction and recommendation



ability to predict ⇒ ability to recommend



Intermezzo: Link to Simulation

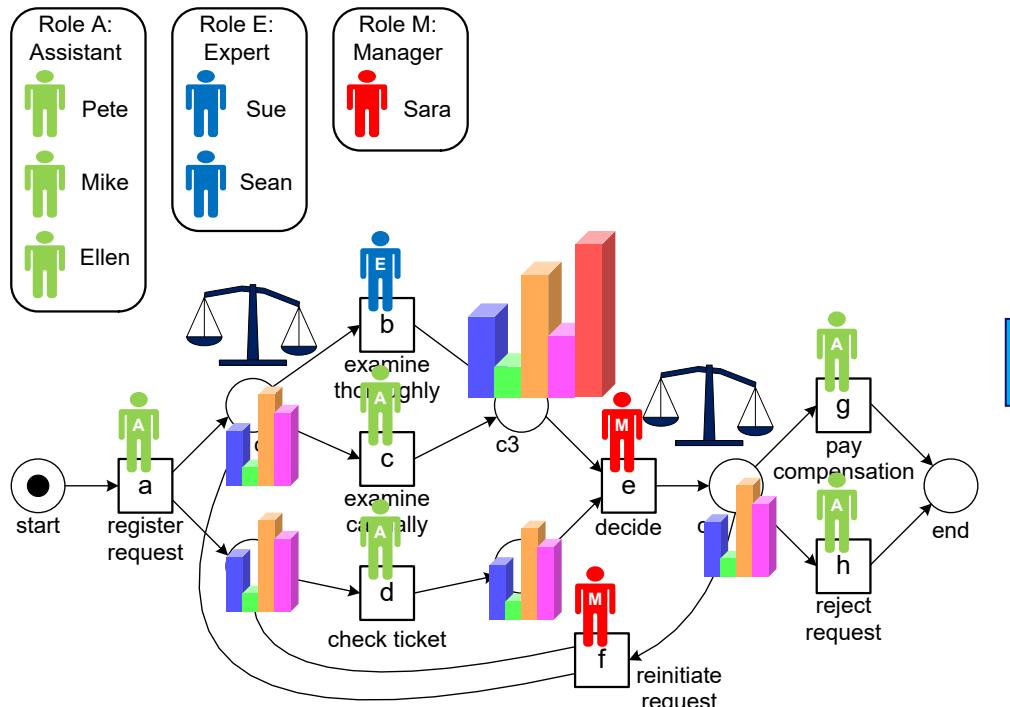
(also forward looking and hence related to OS)



Before: How to combine the different perspectives into "complete" process models?



Integrated model can be used for simulation



Tools: Arena, ProModel, FlexSim, Simul8, Witness, ExtendSim, Simio, PlantSimulation, AnyLogic, Simprocess, Automod, Micro Saint, Quest, CPN Tools, etc.



discrete event simulation



Chair of Process
and Data Science

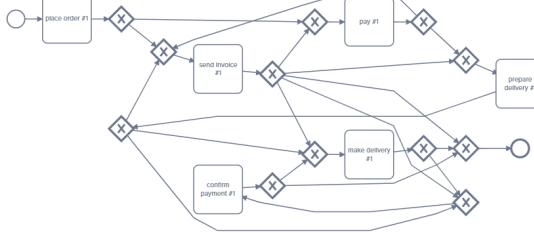
Simulation is also supported by Celonis

(out of scope course)

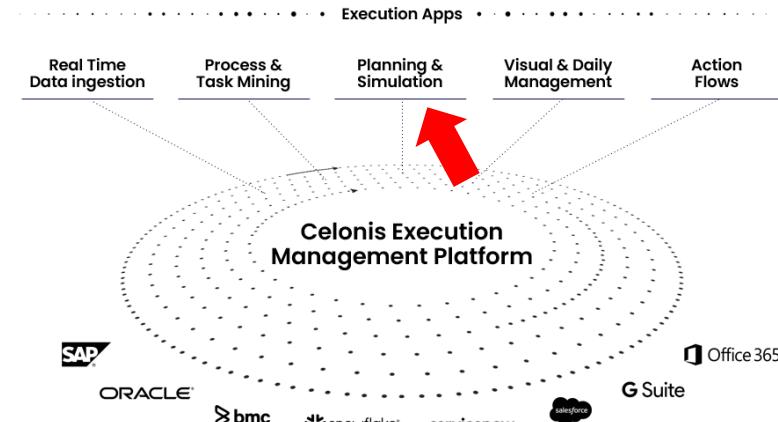
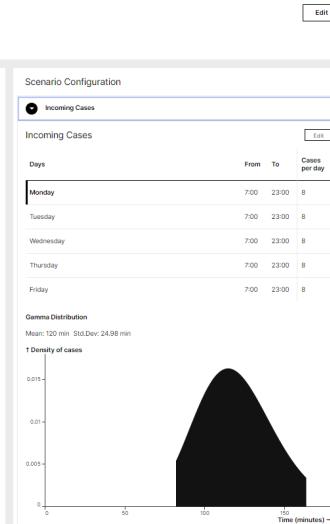
Order-Two-Table-Sim

Digital Twin was configured successfully!
To simulate a scenario, publish the package and head over to the Simulation Dashboard in Business Views.

Diagram



100%



Simulation Configuration Wizard



Today, many people call this a digital twin ...

Saturday
10 June 2023
20:20 BST

NEWS REPORTS CONFERENCES AWARDS ADVERTISE Log In SUBSCRIBE FREE

THE TECH CAPITAL

TOPICS MARKETS CAPITAL PREMIUM LEADERSHIP VIDEOS PODCASTS OPINION

TRENDING DIGITAL INFRASTRUCTURE FIBRE COLOCATION SUSTAINABILITY DIGITALBRIDGE

Digital Twin: The modern time machine for change processes

The idea of a digital twin is alluring in a virtual world, where all possible decisions can be evaluated - without causing damage or costs, writes Prof. Wil van der Aalst is Professor of Computer Science at RWTH Aachen University and Chief Scientist of Celonis.

Updated July 28, 2022 / Original July 26, 2022

By Wil van der Aalst

Professor of Computer Science, RWTH Aachen University

SHARE THIS ARTICLE

4,980

By Wil van der Aalst

Professor of Computer Science, RWTH Aachen University

SHARE THIS ARTICLE

4,980



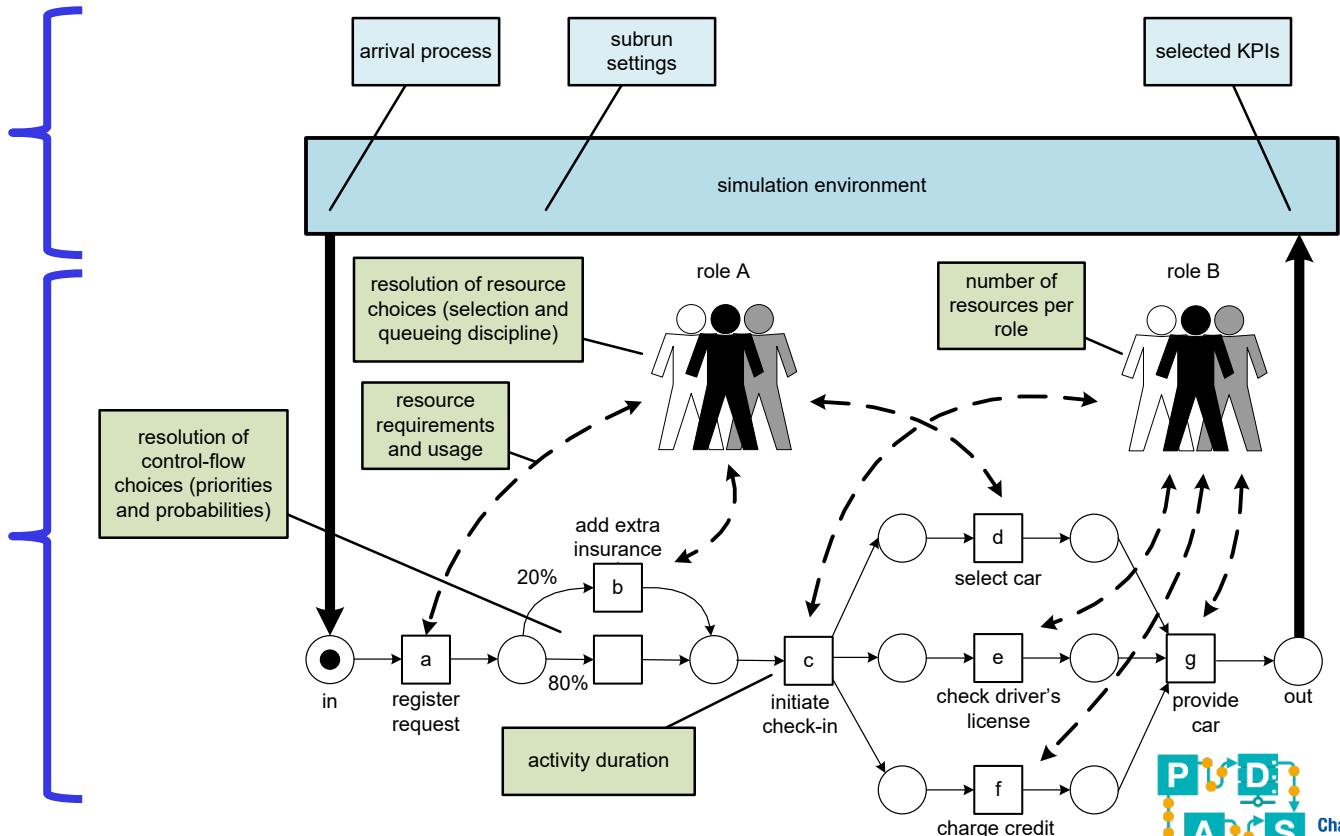
More on simulation

- Repeatedly "playing out a model" to better understand the modeled process.
- Can be used to explore different alternative designs or policies.
 - What is the effect of making the process more concurrent?
 - What is the effect of adding resources?
- Requires a model of the process and its environment.
- Input from process mining!!

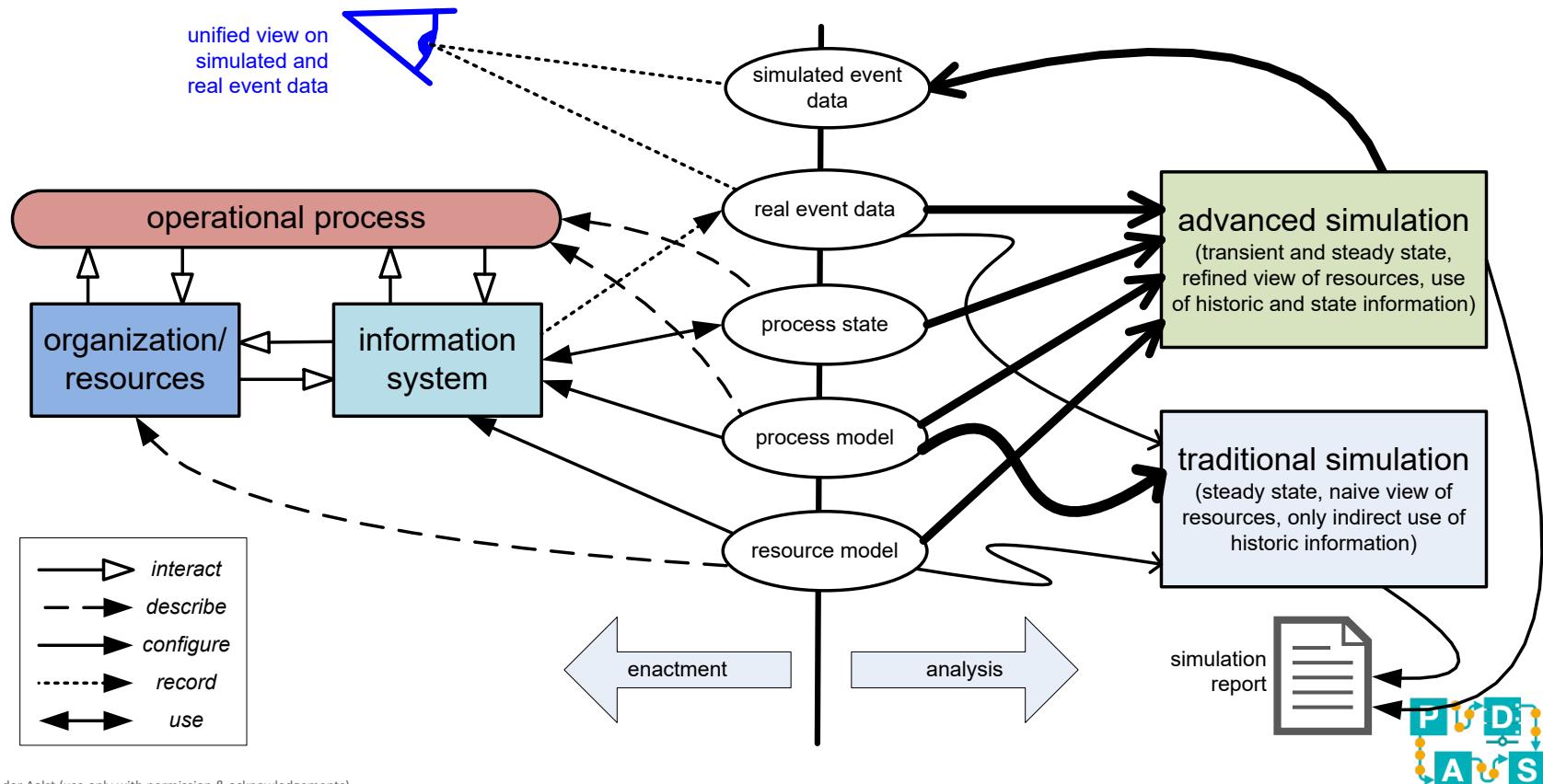
Ingredients

setting up
the
experiment

learn from
process
mining



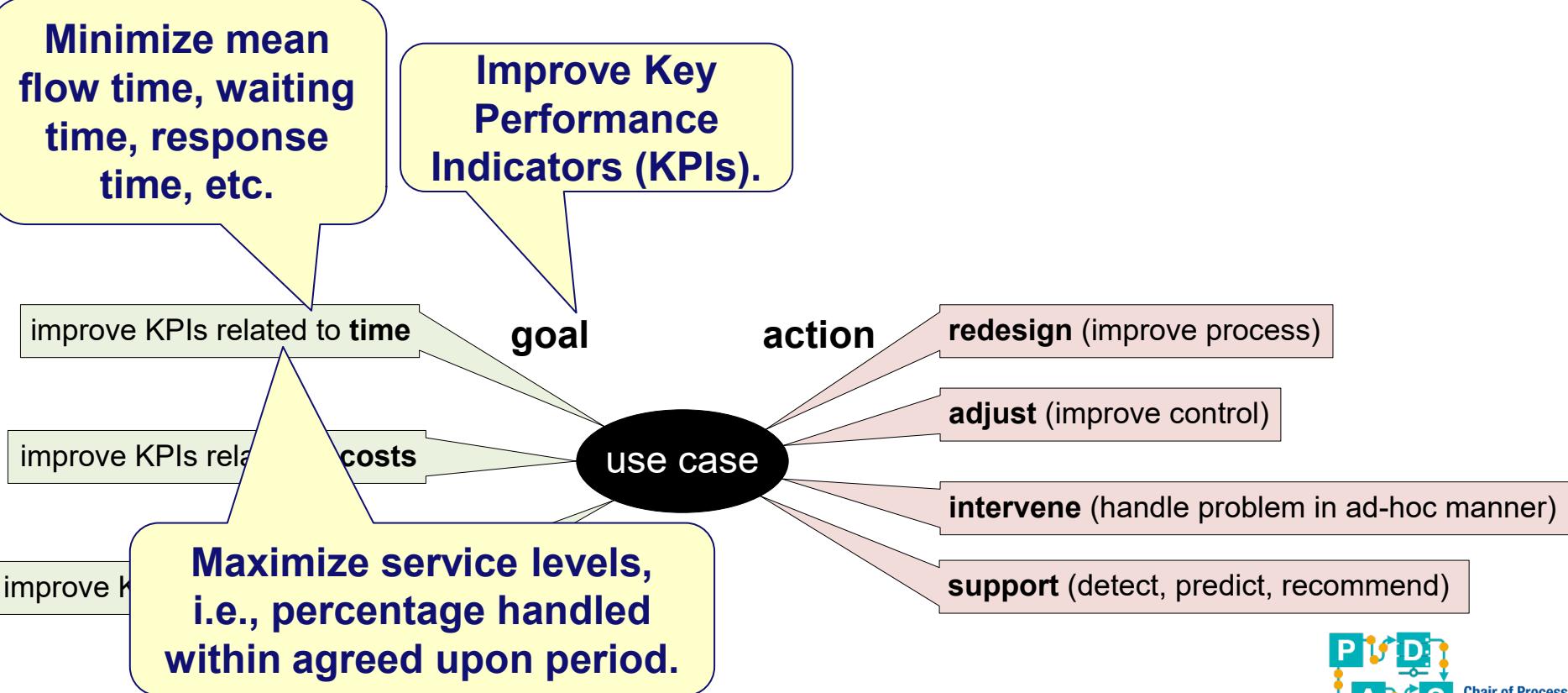
Beyond traditional simulation: Unified view and fast forward



How To Conduct a Process Mining Project?



Process mining use cases



Process mining use cases

Redesign: Structural changes to the process based on insights, e.g., making the process more concurrent or adding controls.

Adjust: Non structural (i.e., temporary) changes, e.g., adding more resources because of fluctuations in case volume.

improve KPIs related to time

goal

improve KPIs related to costs

use case

mining should be actionable
in a timely manner.

action

redesign

(improve process)

adjust (improve control)

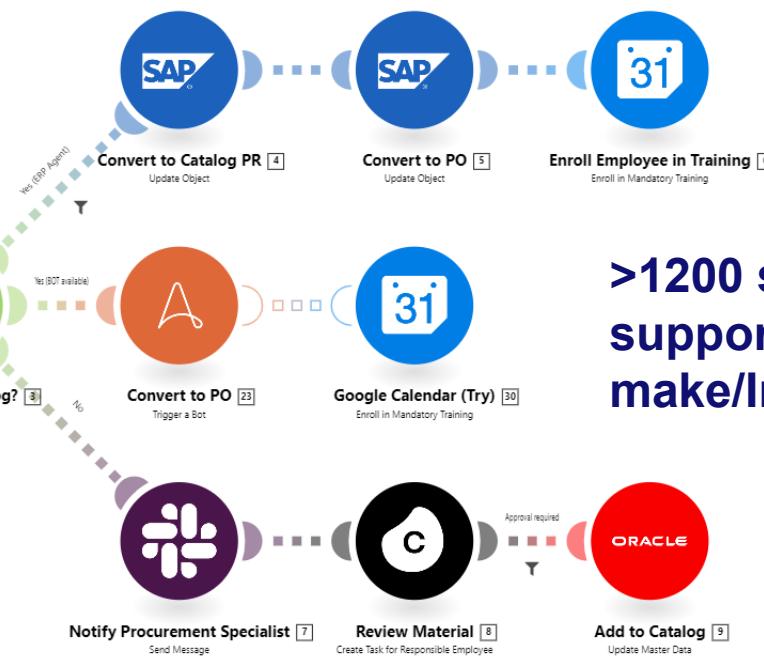
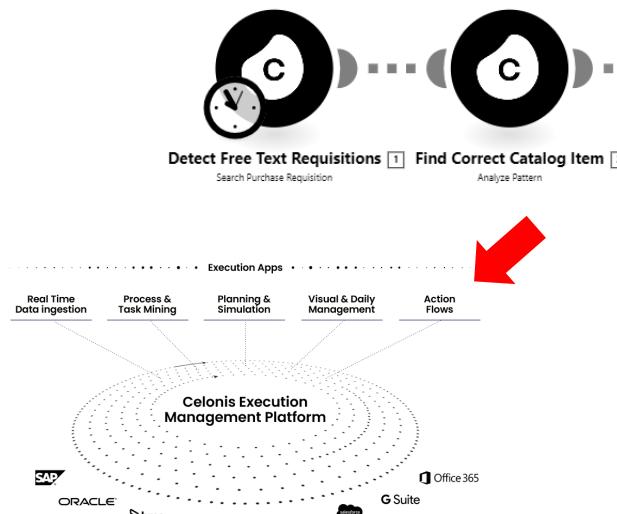
intervene (handle problem in ad-hoc manner)

support (detect, predict, recommend)

Support: Systematically using pre mortem event data, e.g., for recommending the activity most likely to minimize the flow time.

Example: The Celonis Action Flows

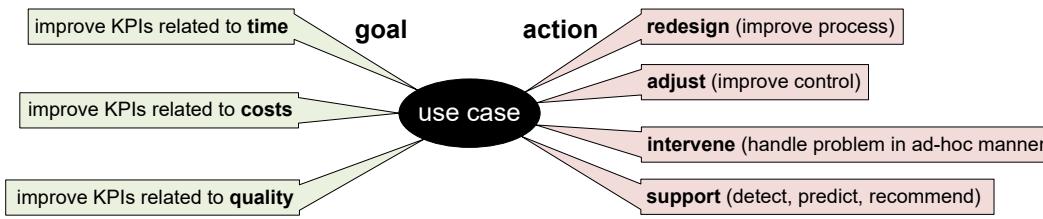
Triggered from process mining diagnostics



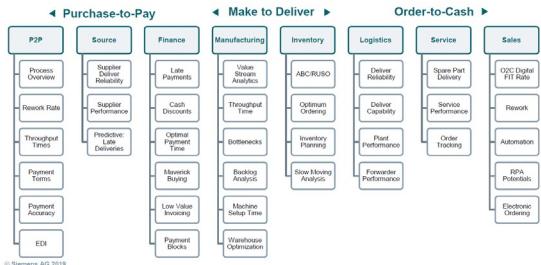
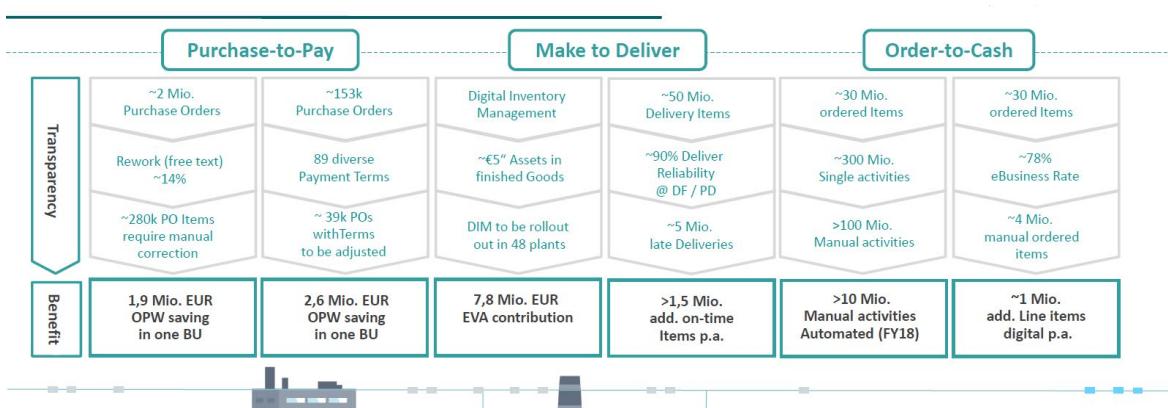
>1200 systems supported by make/Integromat

Process mining use cases

- Identification of bottlenecks to trigger a process redesign that reduces the overall flow time with 30%.
- Identification of compliance problems using conformance checking. Some of the compliance problems result in ad-hoc interventions whereas others lead to adjustments of the parameters used for work distribution.
- Harmonization of two processes after a merger based on a comparison of the actual processes. The goal of such a harmonization is to reduce costs.
- Predicting of the remaining flow time to improve customer service.
- Providing recommendations for resource allocation aiming at a more balanced utilization of workers.
- Identification of exceptional cases that generate too much additional work. By learning the profile of such cases, they can be handled separately to reduce the overall flow time.
- Visualization of the 10 most complicated or time consuming cases to identify potential risks.



Example: PM @ Siemens



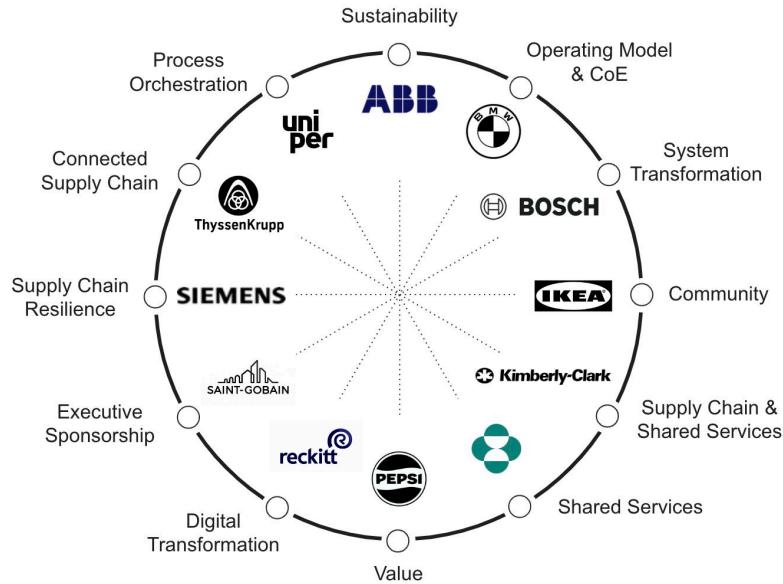
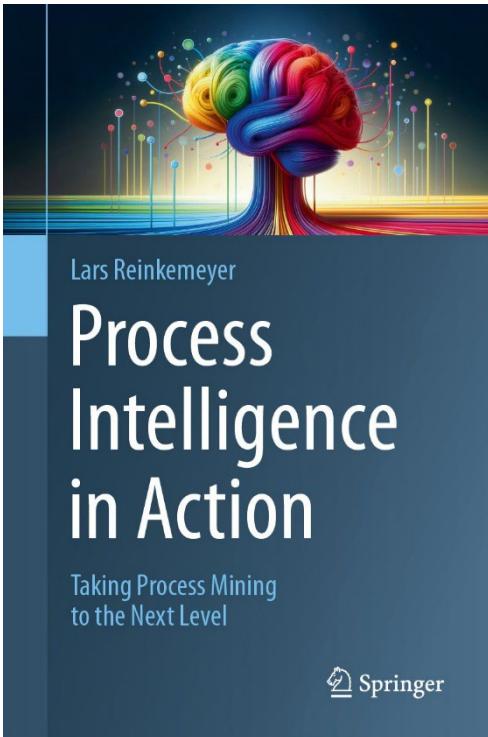
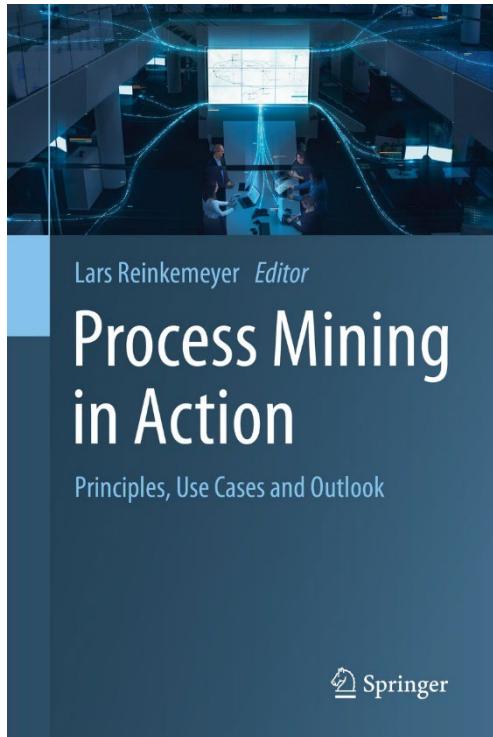
- **> 6000 active Celonis users (P2P, O2C, etc.)**
- **Millions of savings by reducing rework, process unification, etc.**
- **Improved reliability and responsiveness.**
- **At an amazing scale, e.g., Order to Cash (O2C) process with >30M cases, >300M events, and >900K variants.**



Chair of Process
and Data Science

Thanks to Lars Reinkemeyer (Siemens)

Books describing use cases



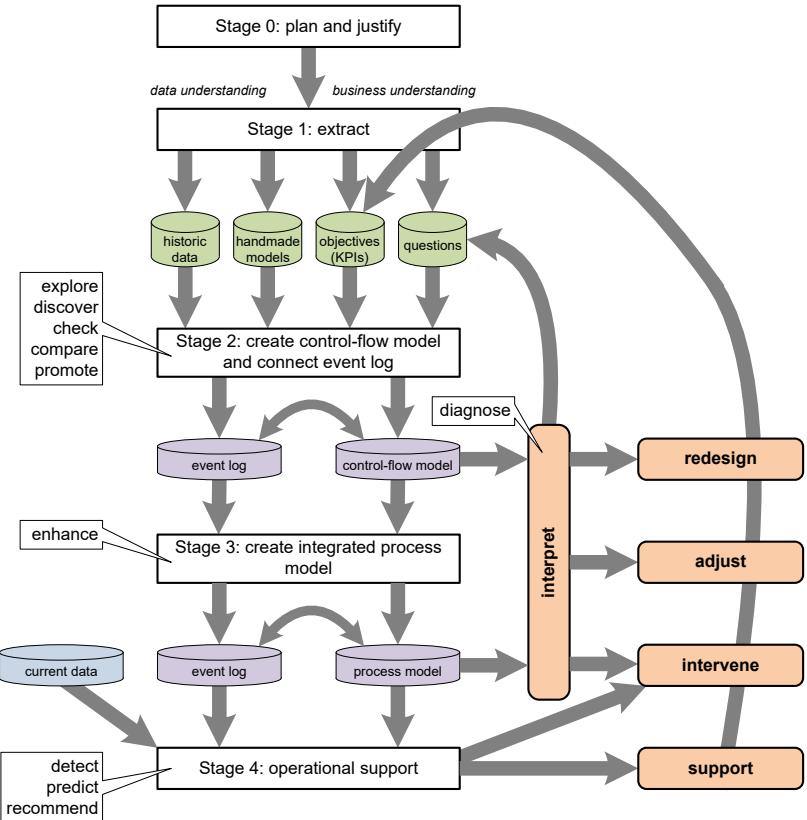
12 operational use-cases from ABB, BMW Group, Bosch, IKEA, Kimberly-Clark, Merck Group, PepsiCo, Reckitt, Saint-Gobain, Siemens, thyssenkrupp Rasselstein, Uniper

Book presentation July 10th at 17.00 https://celonis.zoom.us/webinar/register/WN_UoLW5GGRCWnt-98R7FvEw#/registration



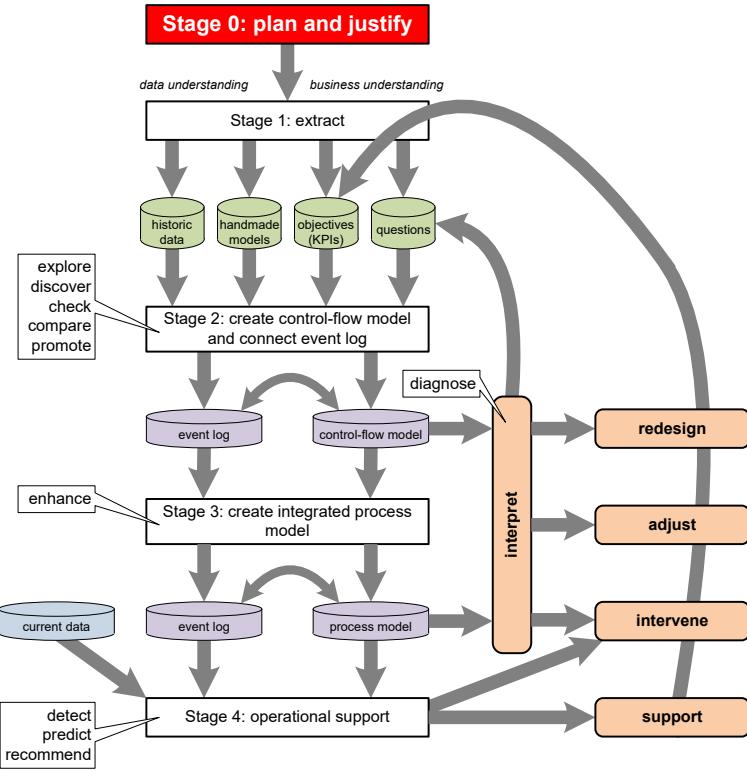
L* lifecycle model for process mining

L* lifecycle model for process mining



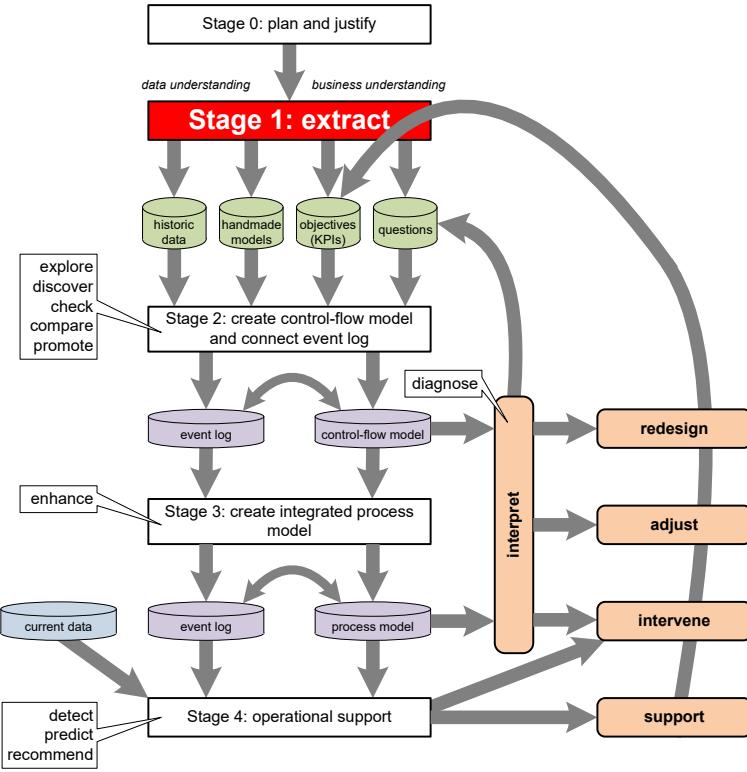
Describes the lifecycle of an idealized process mining project (assuming "Lasagna processes").

Stage 0: Plan and justify



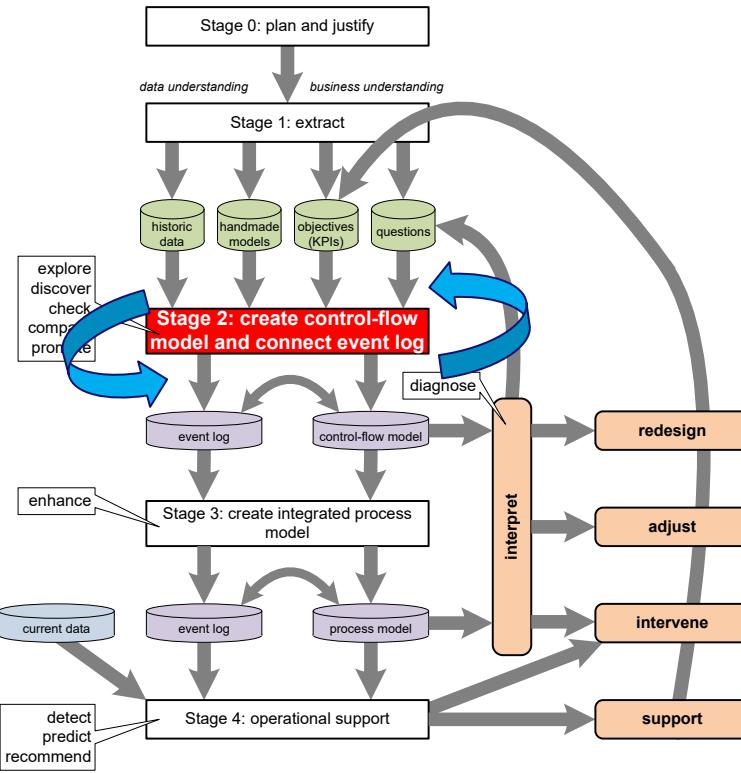
- **Three types of projects:**
 - data-driven ("curiosity" driven)
 - question-driven ("why?")
 - goal-driven (improve KPI)
- **Plan project.**
- **Justify planned activities ("business case").**

Stage 1: Extract



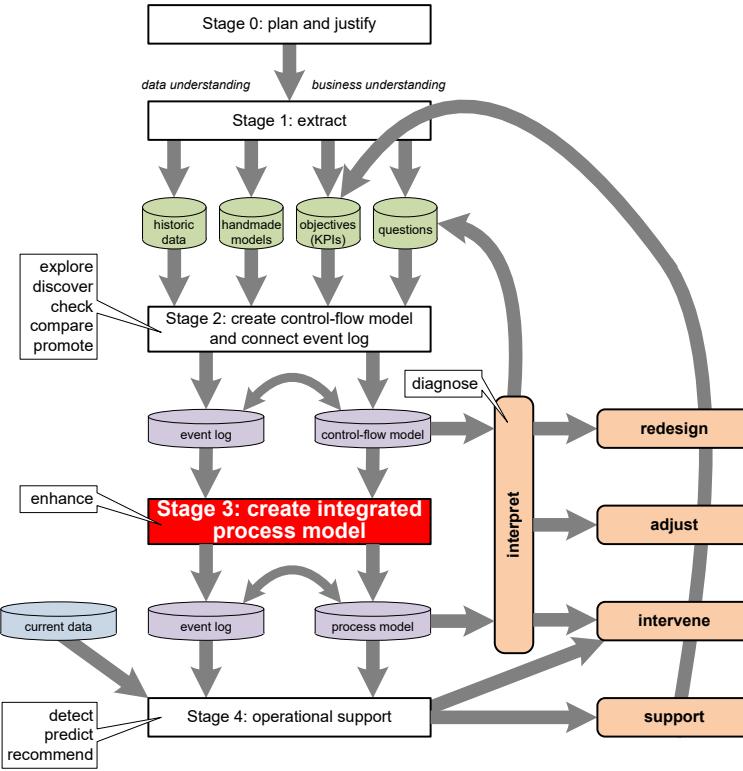
- **Locate, extract and transform event data (non-trivial, see previous lectures).**
- **Moreover, collect:**
 - **models and other artifacts,**
 - **objectives (KPIs), and**
 - **questions.**
- **Exploit existing (domain) knowledge!**

Stage 2: Create control-flow model and connect event log



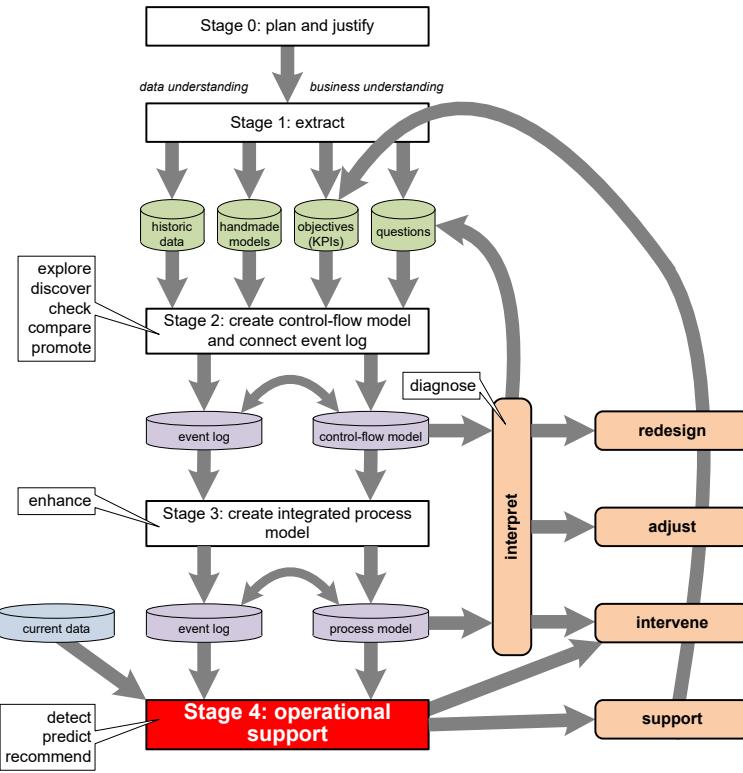
- Control-flow is the backbone of any process.
- Therefore, first create a suitable control-flow model well-connected to the available event data.
- Conformance checking and alignments are key!
- Iterative (like other stages).

Stage 3: Create integrated process model



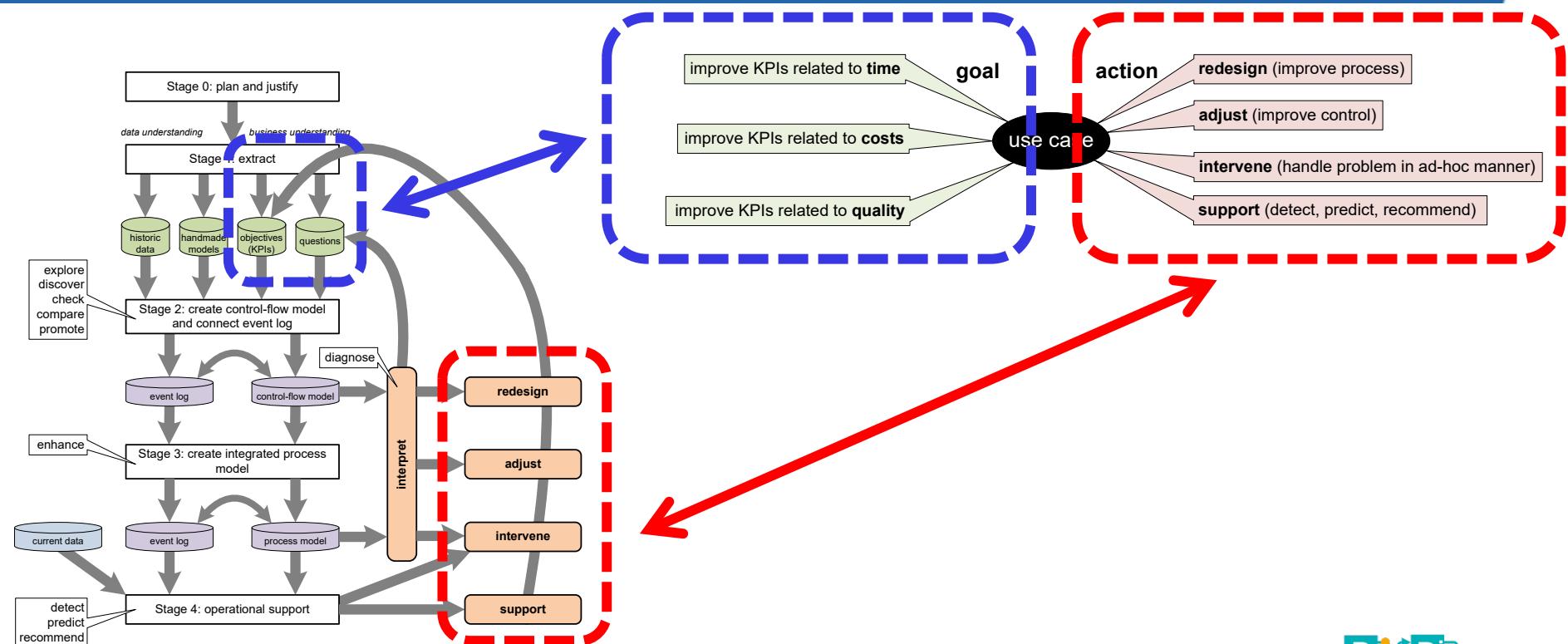
- **Replay event data on control-flow model to learn about the other perspectives (time, data, resources, ...).**
- **Merge into an overall model showing the different perspectives.**

Stage 4: Operational support

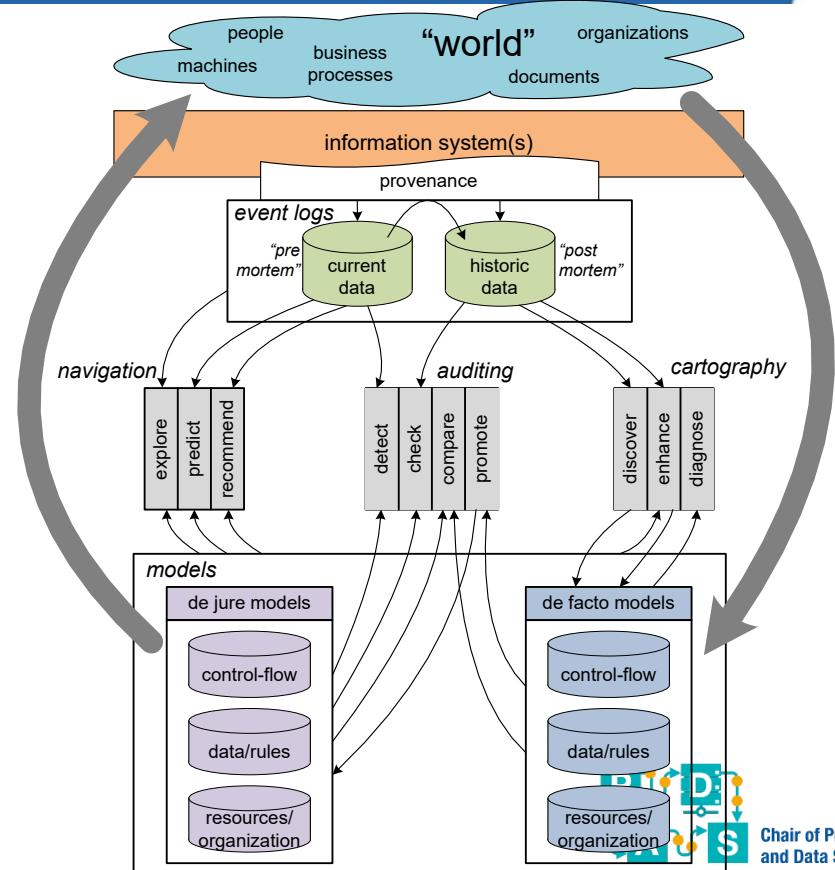
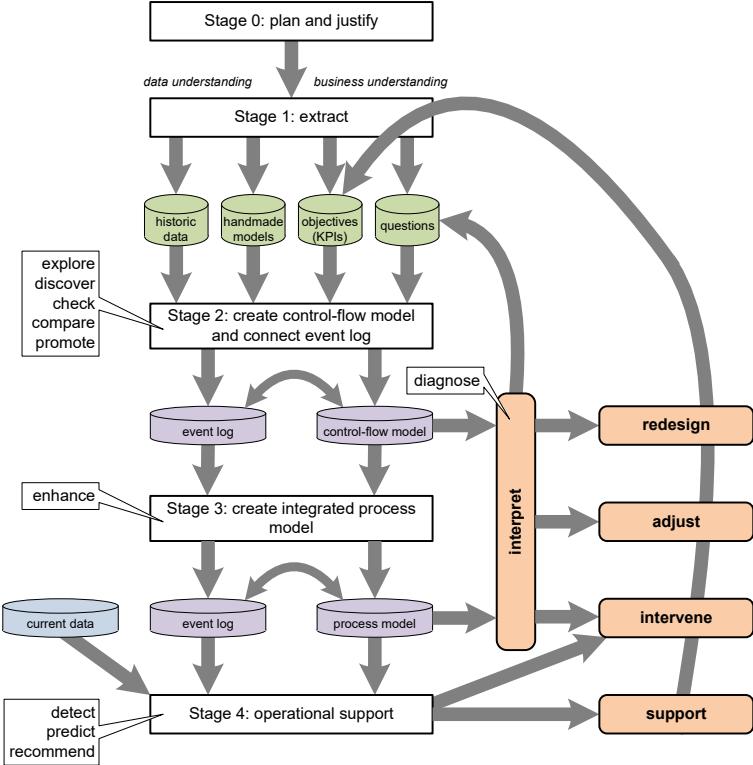


- Use current (pre-mortem) data for on-the-fly deviation detection, predictions, and recommendations.
- Only possible for Lasagna processes!

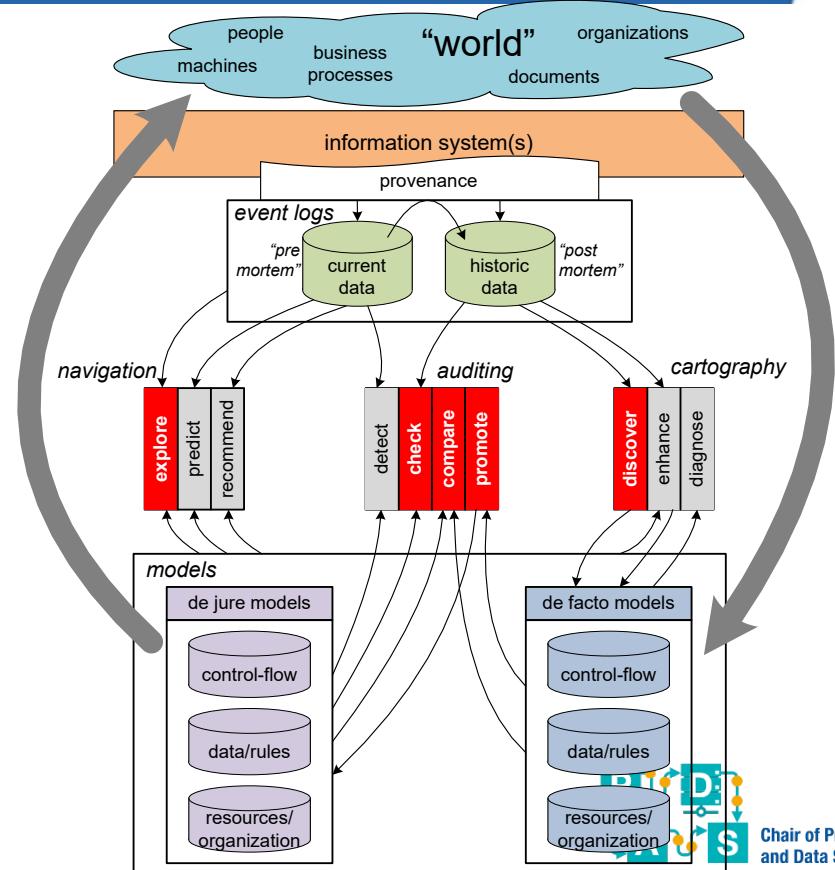
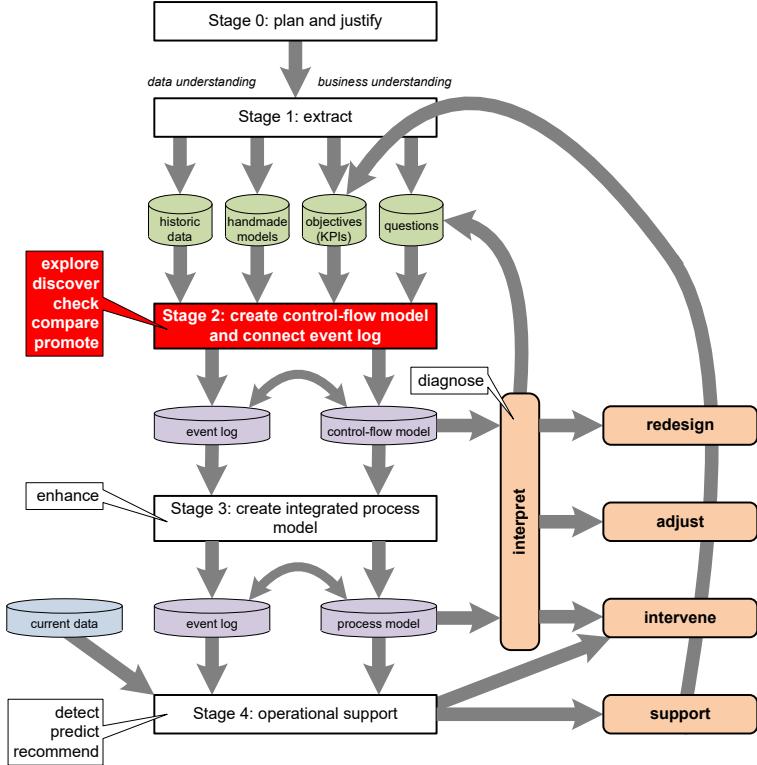
Relation to use cases



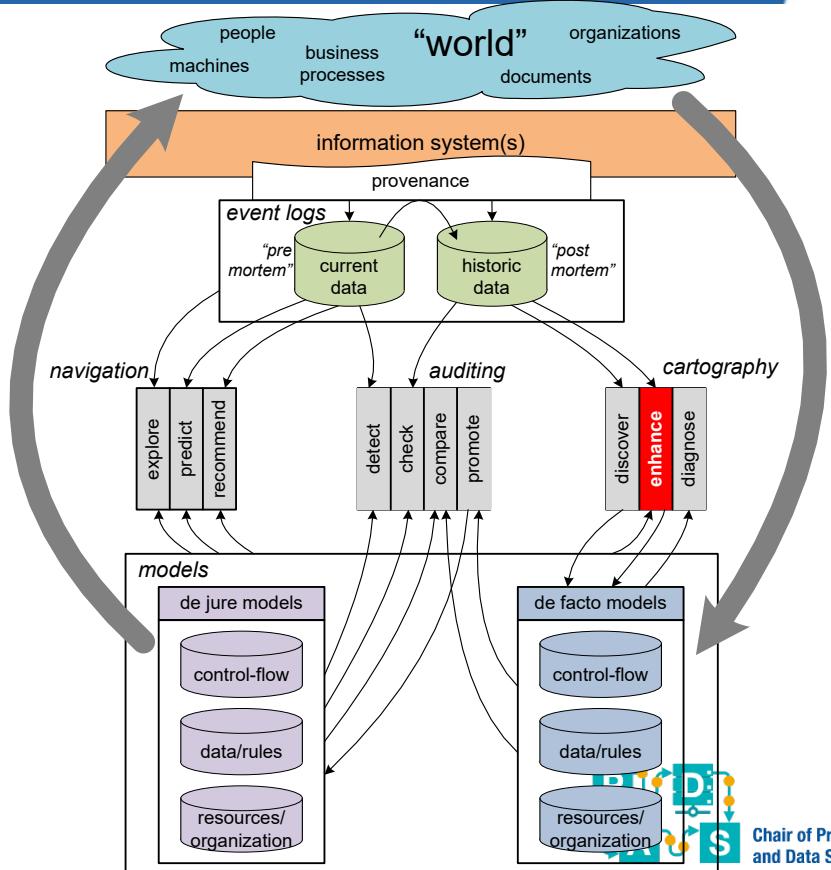
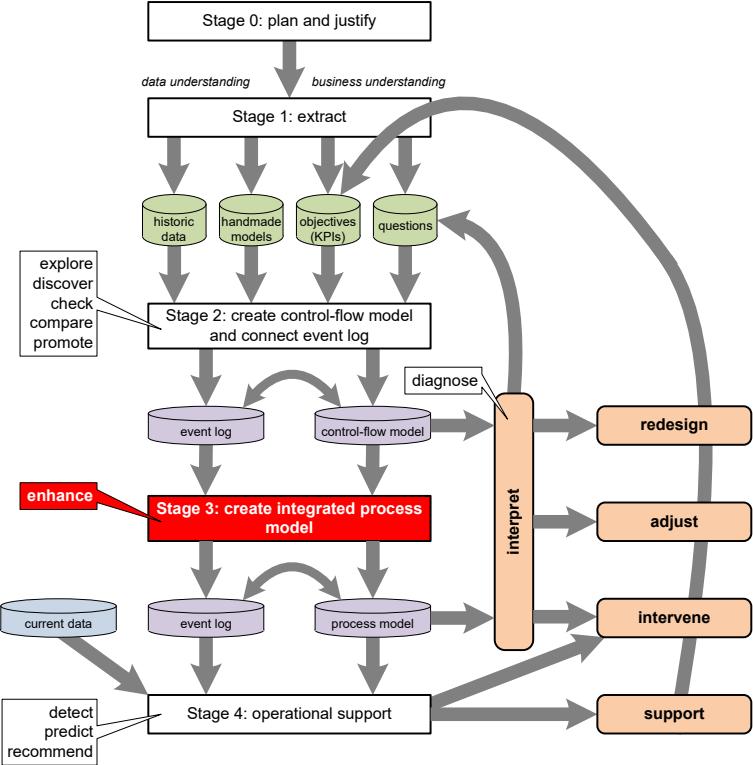
Linking L* to the refined process mining framework



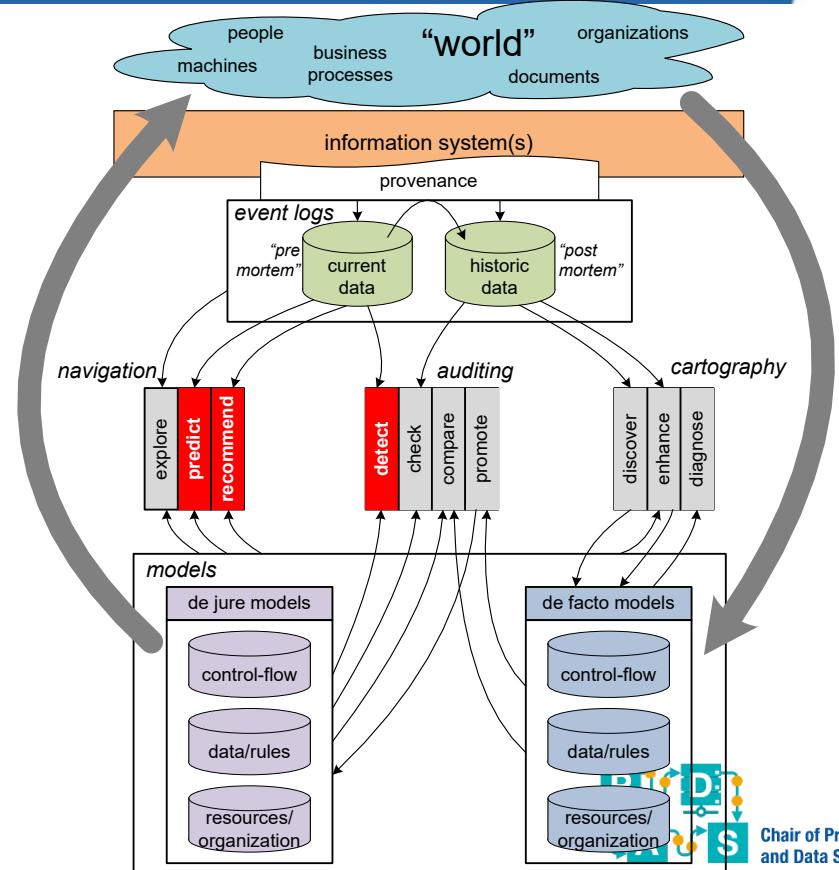
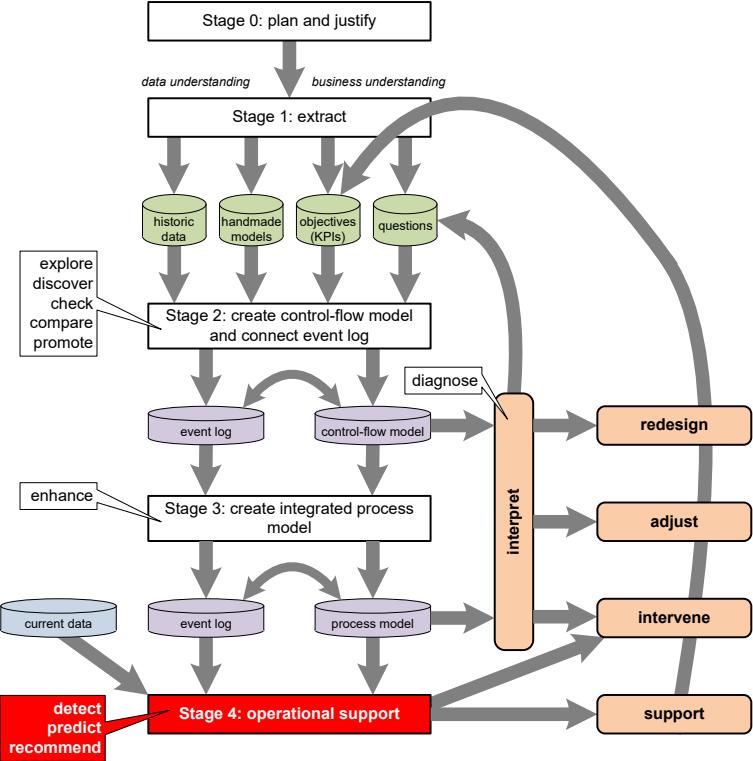
Linking L* to the refined process mining framework



Linking L* to the refined process mining framework

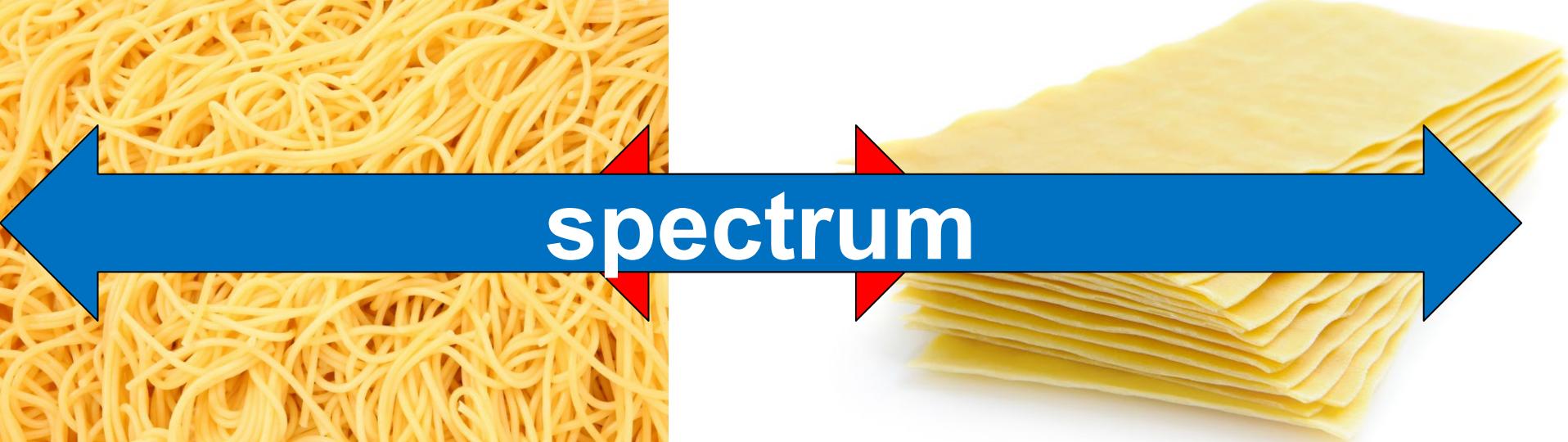


Linking L* to the refined process mining framework



Mining Lasagna Processes

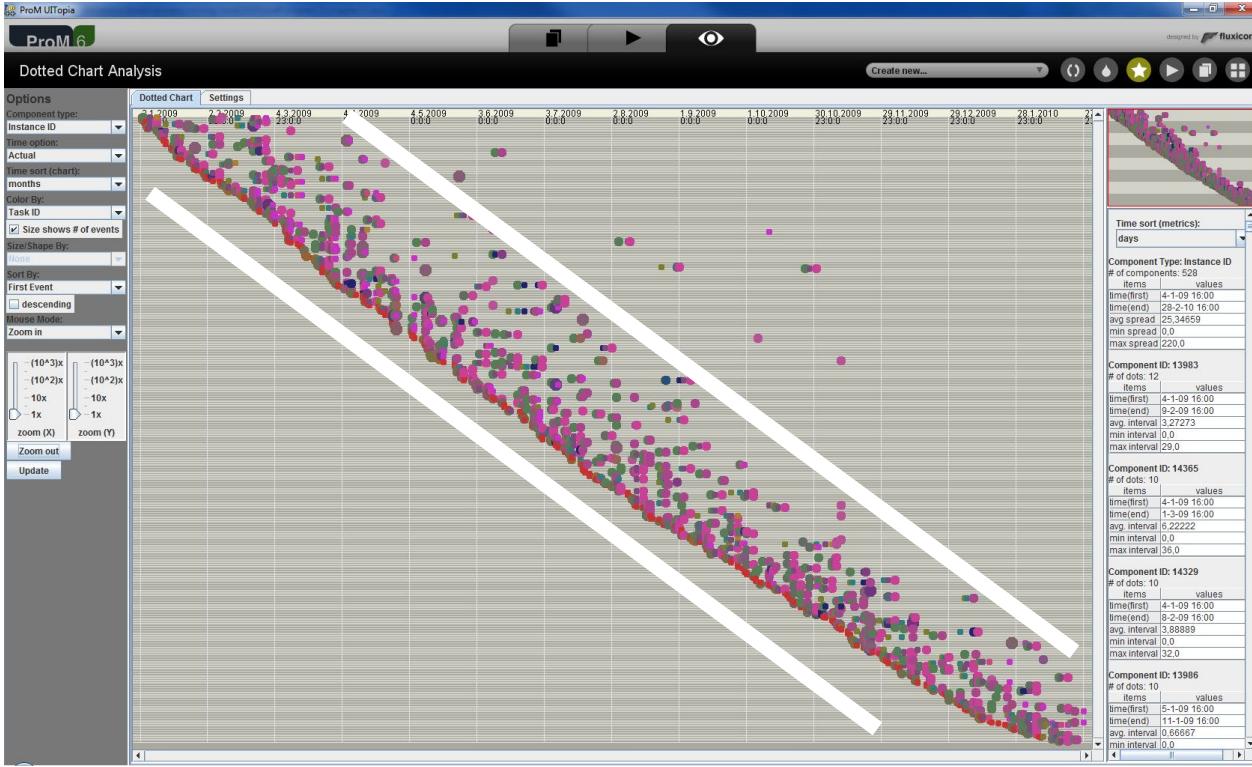




**unstructured
irregular
flexible
variable**

**structured
regular
controllable
repetitive**

Example of a Lasagna process



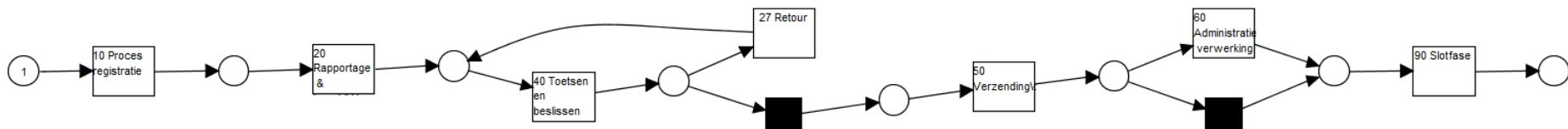
- WMO process of a Dutch municipality.
- 528 requests for household help.
- 5498 events.

Process model

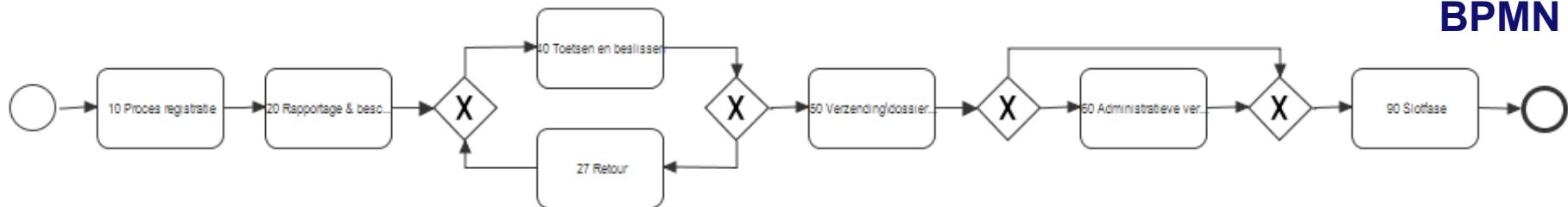
Inductive miner



Petri net



BPMN



Fitness is close to perfect

Replay result - log UNIFIED (filtered on simple heuristics) on 481bde02-d86d-4afa-90f4-bfb783897bd2 using A* Cost-based Fit! Create new... p+ () la n★av e, ss

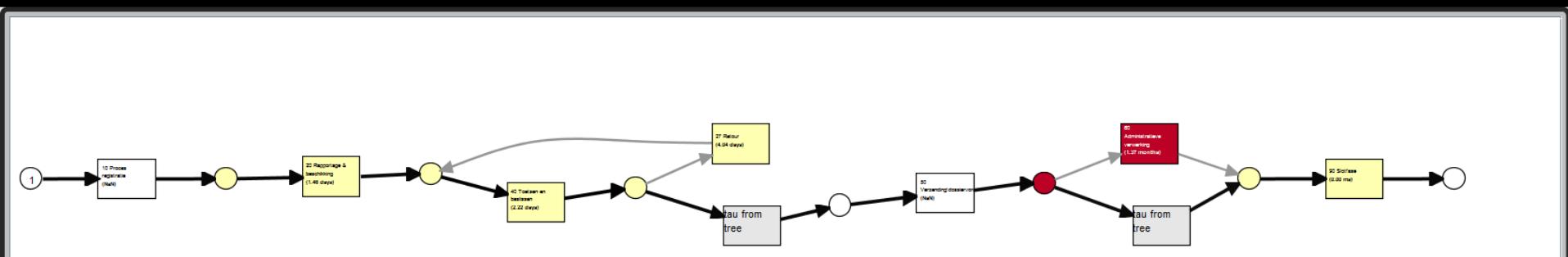
1 → [10 Proces registratie (62810)] → [20 Receptie beschikking (62810)] → [40 Toetsen en beslissen (62810)] → [27 Retour (2440)] → [50 Verzending doc (62810)] → [60 Administratieve Verwerking (62810)] → [90 Slaap (62810)] → []

Inspector

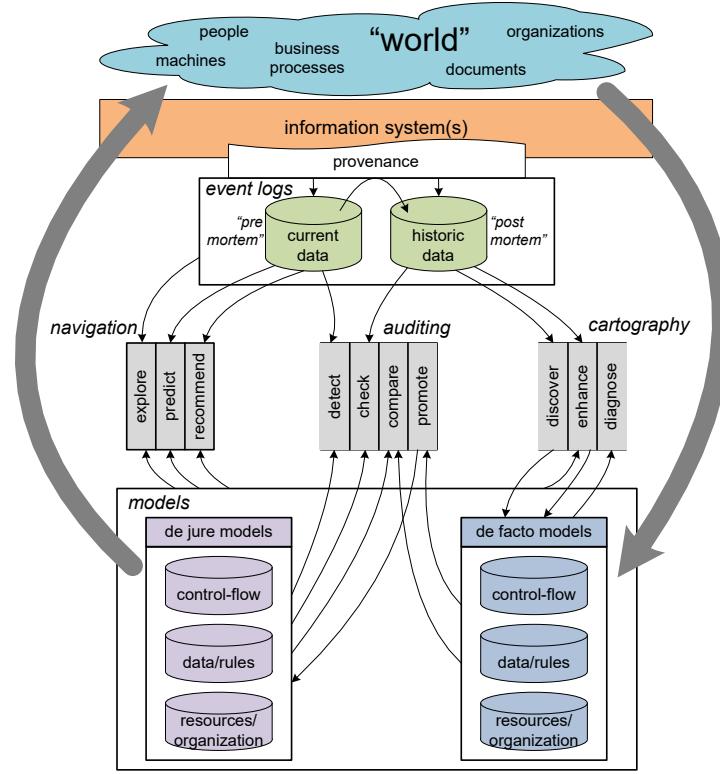
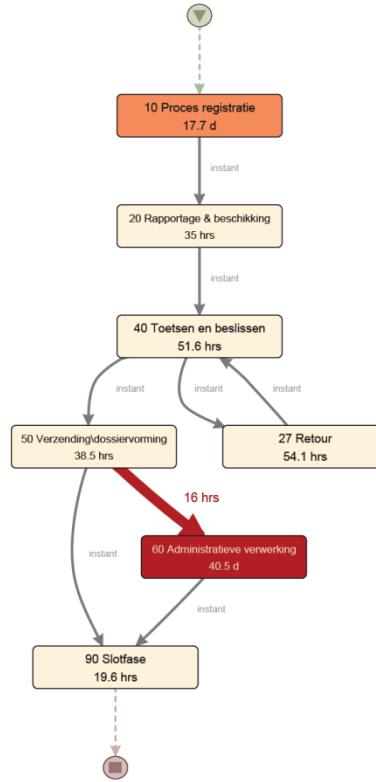
Property	Value
Raw Fitness Cost	0.087121212121215
Queued States	21.9393939393923
Num. States	8.5075757575756
Calculation Time (ms)	1.13825757575756
Move-Log Fitness	0.9877946127946128
Trace Fitness	0.9928300865800871
Trace Length	5.20643939393945
Move-Model Fitness	1.0

505 of the 528 requests are perfectly fitting

Performance analysis



All 10 process mining activities can be performed for Lasagna processes



Lasagna processes

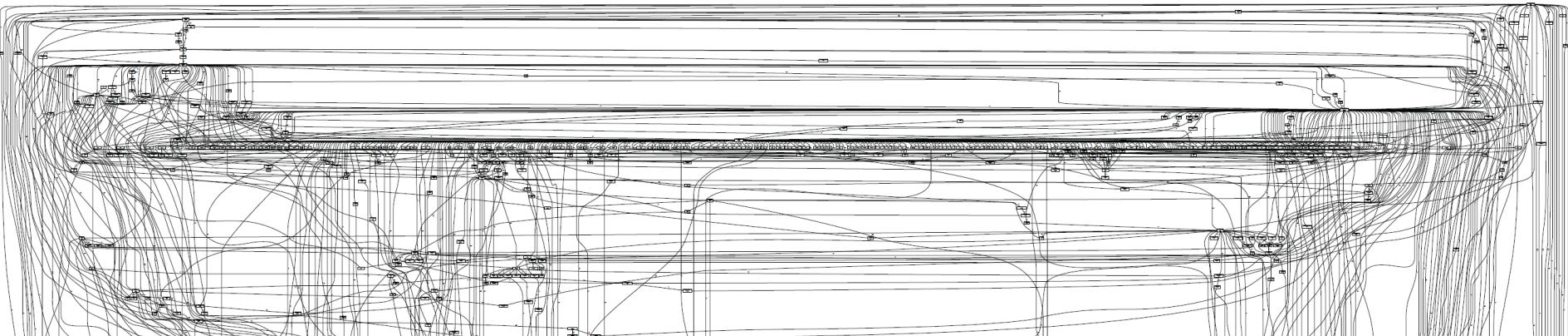
- Easy to discover, but it is less interesting to show the "real" process. (close to expectation)
- Whole process mining toolbox can be applied.
- Added value is predominantly in more advanced forms of process mining based on aligning log and model.



Mining Spaghetti Processes

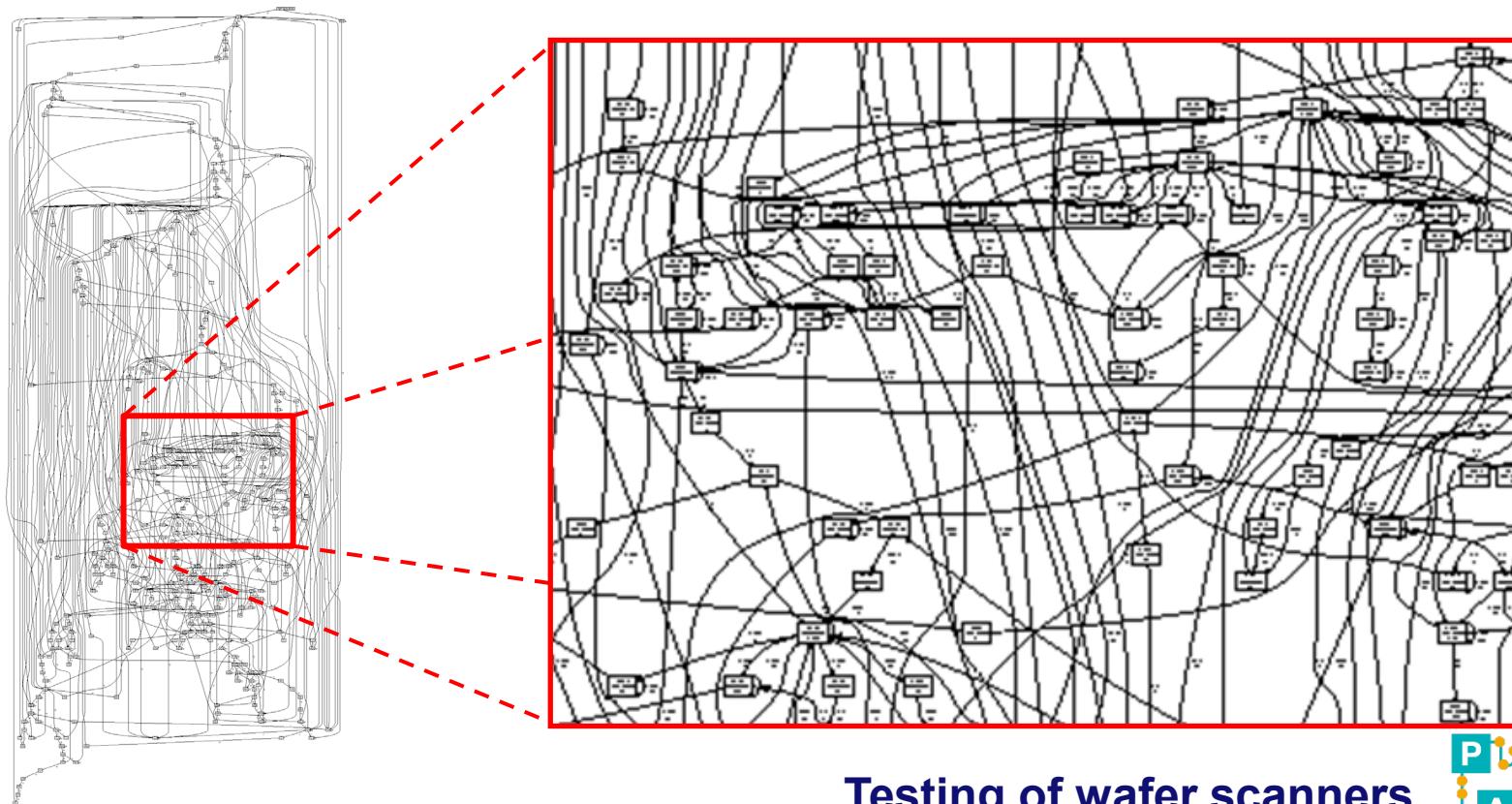


Example of a Spaghetti process



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.).

Another example of a Spaghetti process

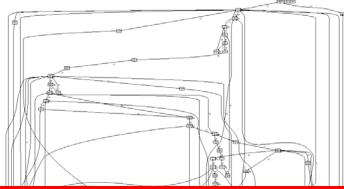


Testing of wafer scanners.

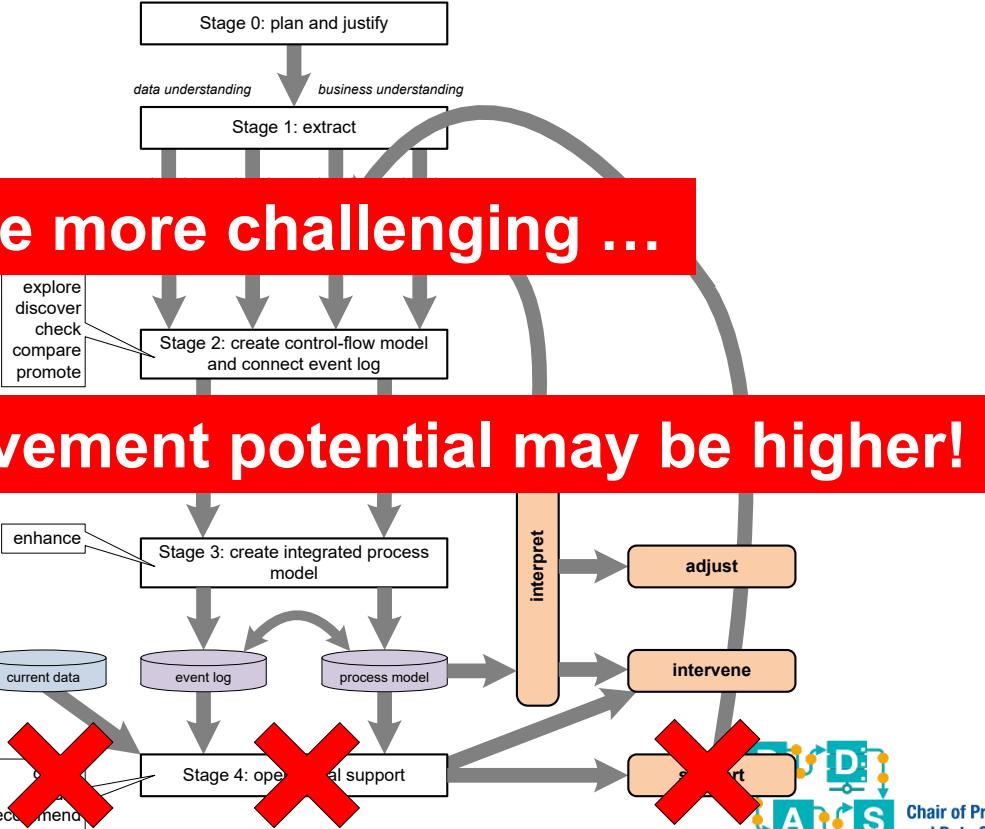
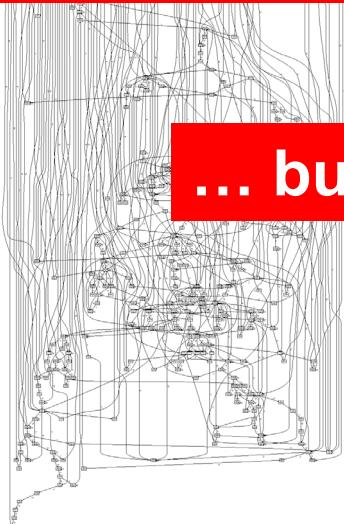


Chair of Process
and Data Science

The last stage(s) of the L* lifecycle model cannot be reached for Spaghetti processes



Earlier stages may also be more challenging ...



Don't be afraid of Spaghetti ...



Simplification

- Focus on a **subset of activities**:
 - only most frequent ones
 - selected region of process (e.g., backoffice)
- Focus on a **subset of cases**:
 - homogeneous groups of cases (clustering)
 - natural subclasses (e.g., gold customers)
- Focus on a **subset of paths**:
 - only most frequent ones

Variant Explorer | Business Views

1.33M of 1.33M cases selected 100%

Process Start
938,251

Timecard Submitted
938,251 | DelTek

Timecard Approved
938,251 | DelTek

Perform Pay Calculations
938,251 | PeopleSoft

Submit to Payroll Provider
938,251 | ADP

Submit Direct Deposit Payment
938,251 | ADP

Process End
938,251

Zoom

Most common variant

Graph

Payroll processing

Top 1 (most frequent variant)

Variant Explorer

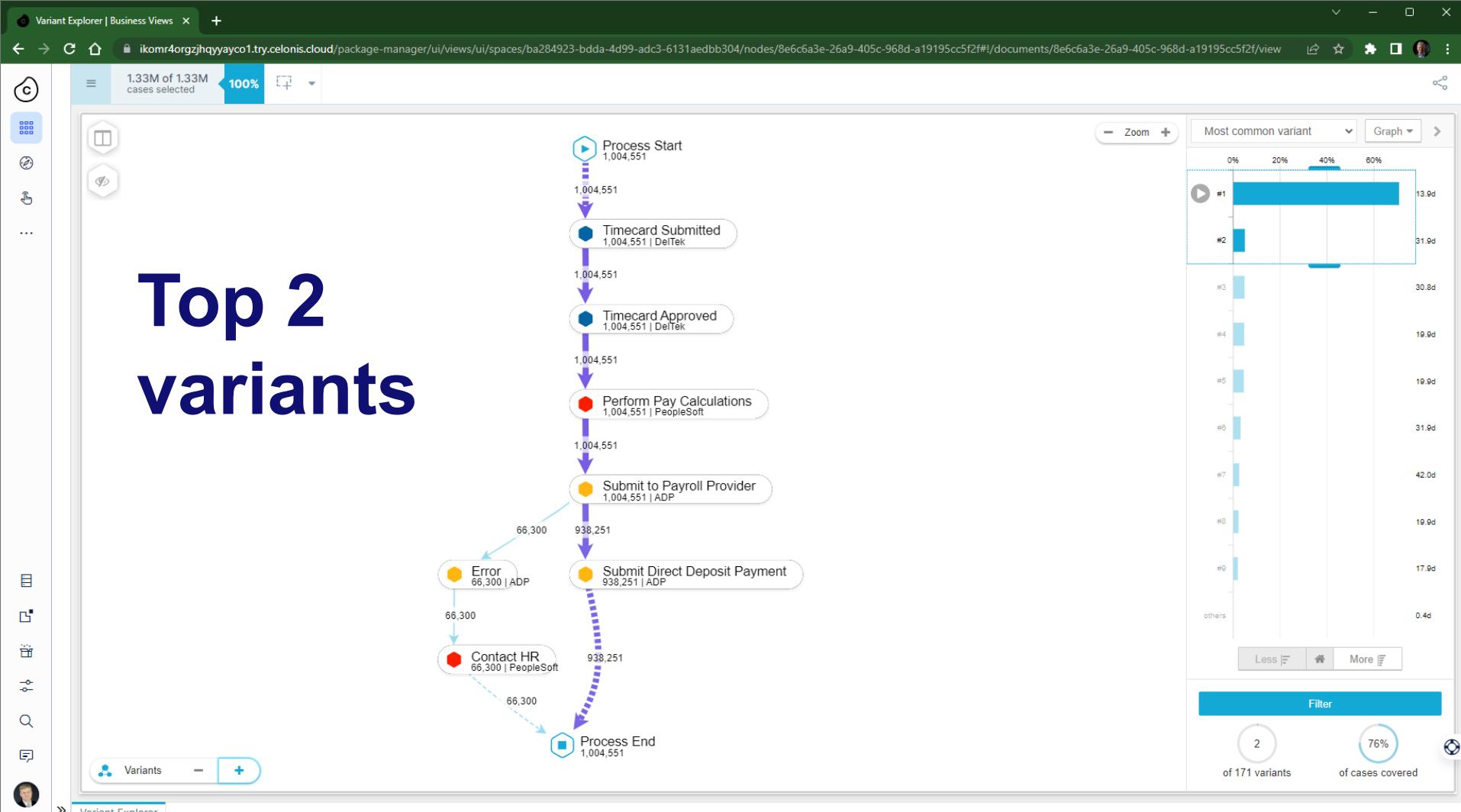
```
graph TD; Start((Process Start 938,251)) --> Submit1((Timecard Submitted 938,251 | DelTek)); Submit1 --> Approve1((Timecard Approved 938,251 | DelTek)); Approve1 --> Calc1((Perform Pay Calculations 938,251 | PeopleSoft)); Calc1 --> Submit2((Submit to Payroll Provider 938,251 | ADP)); Submit2 --> Submit3((Submit Direct Deposit Payment 938,251 | ADP)); Submit3 --> End((Process End 938,251))
```

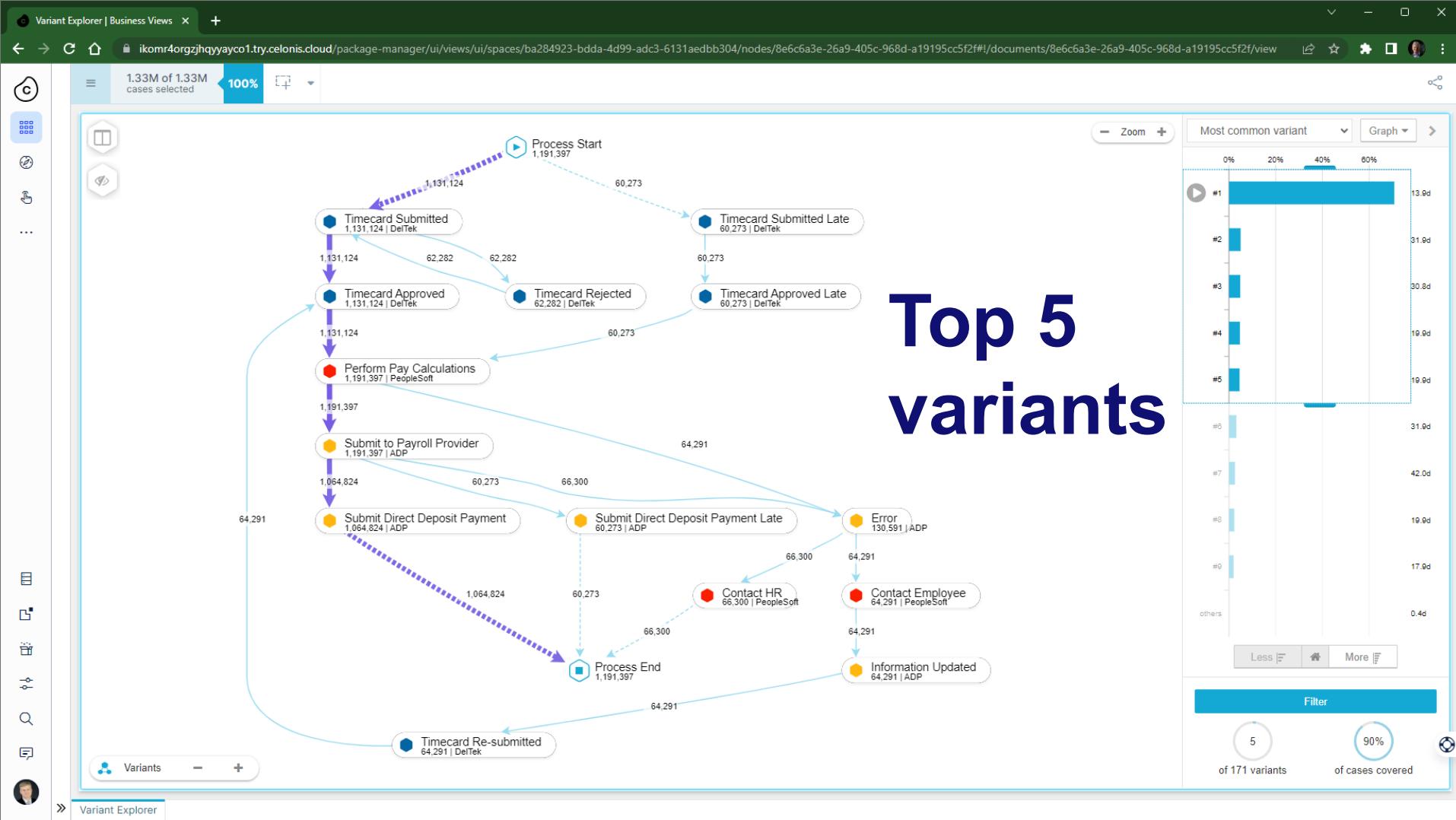
Variant	Percentage
#1	13.9d
#2	31.9d
#3	30.8d
#4	19.9d
#5	19.9d
#6	31.9d
#7	42.0d
#8	19.9d
#9	17.9d
others	0.4d

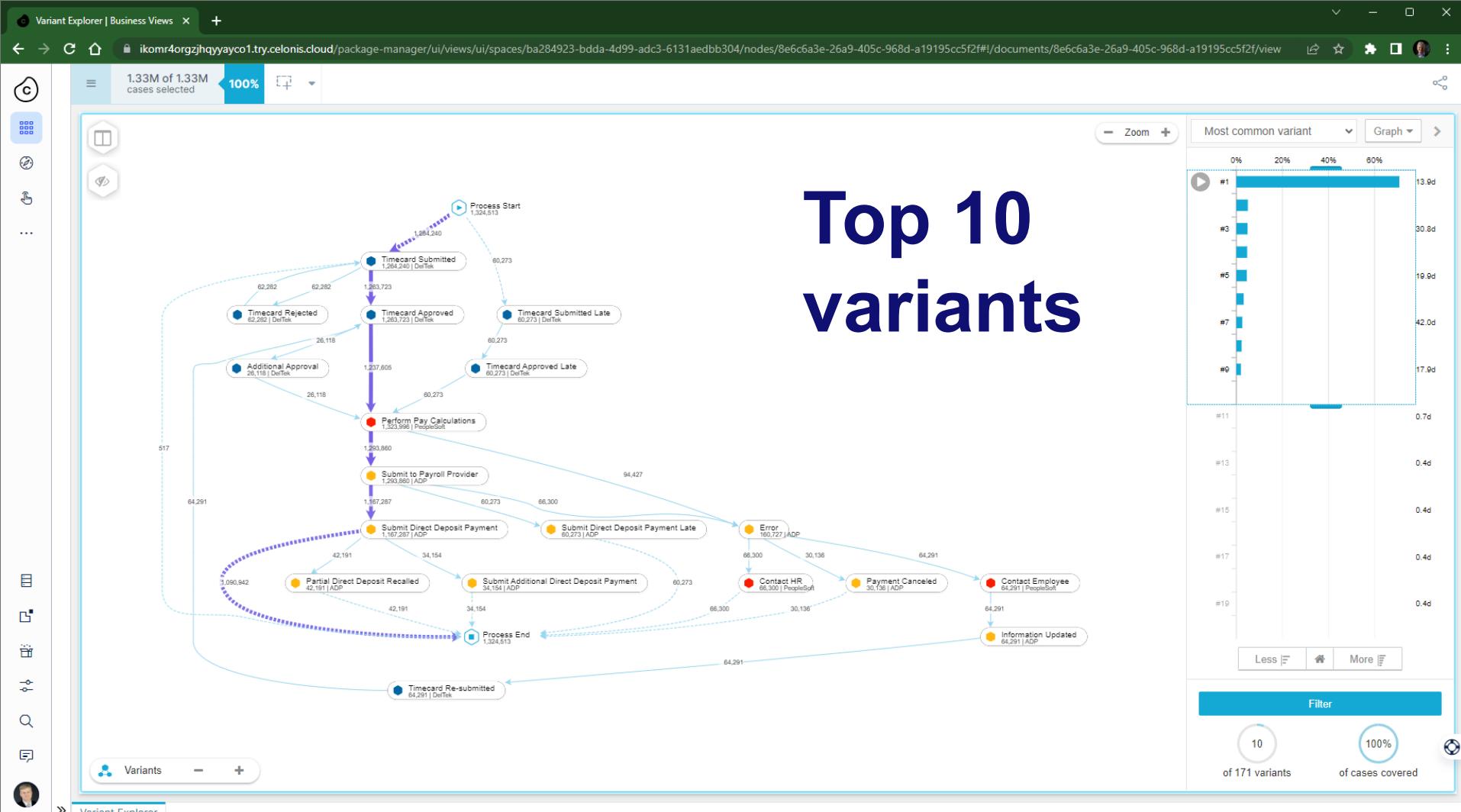
Filter

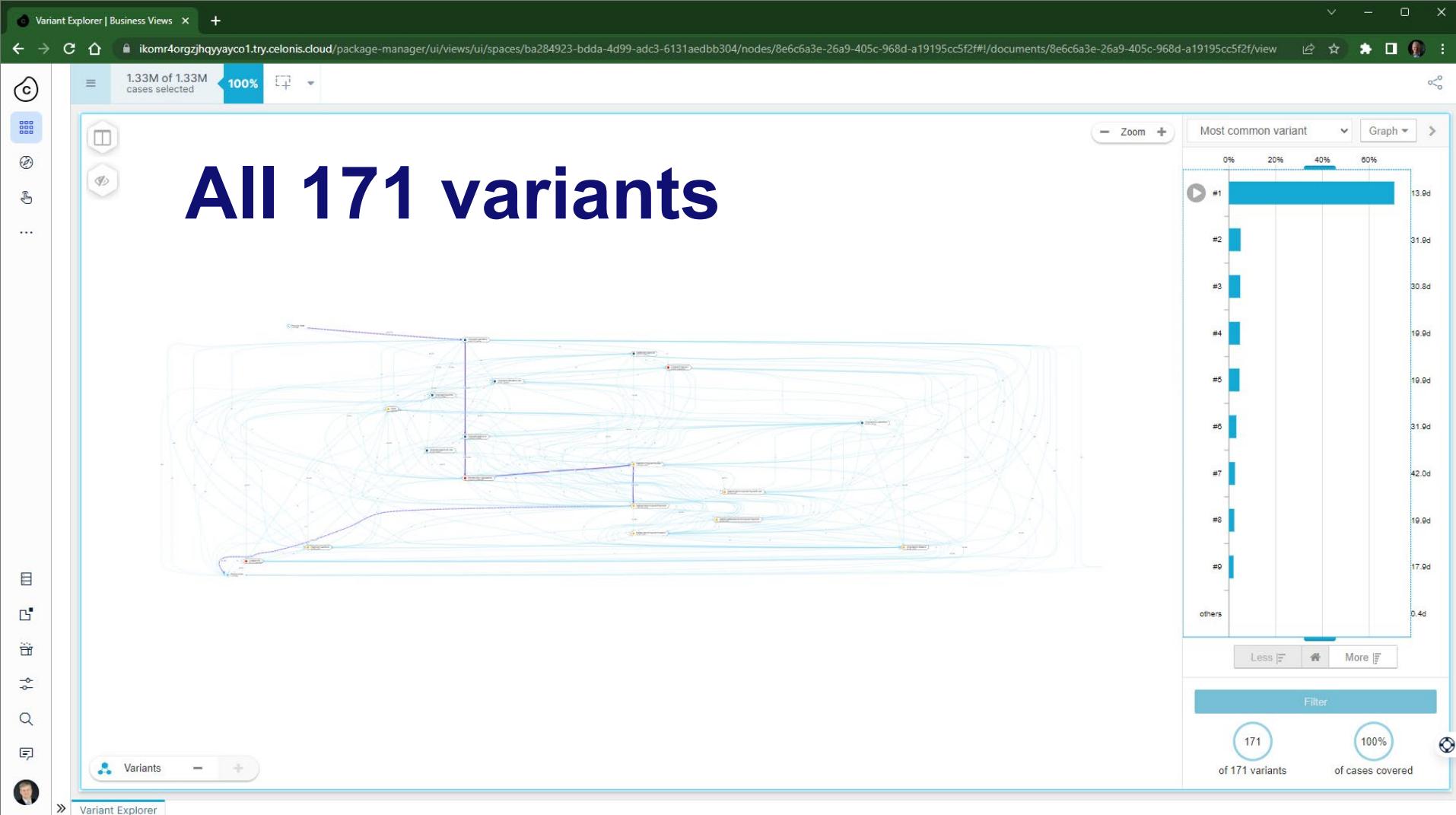
1 of 171 variants

71% of cases covered











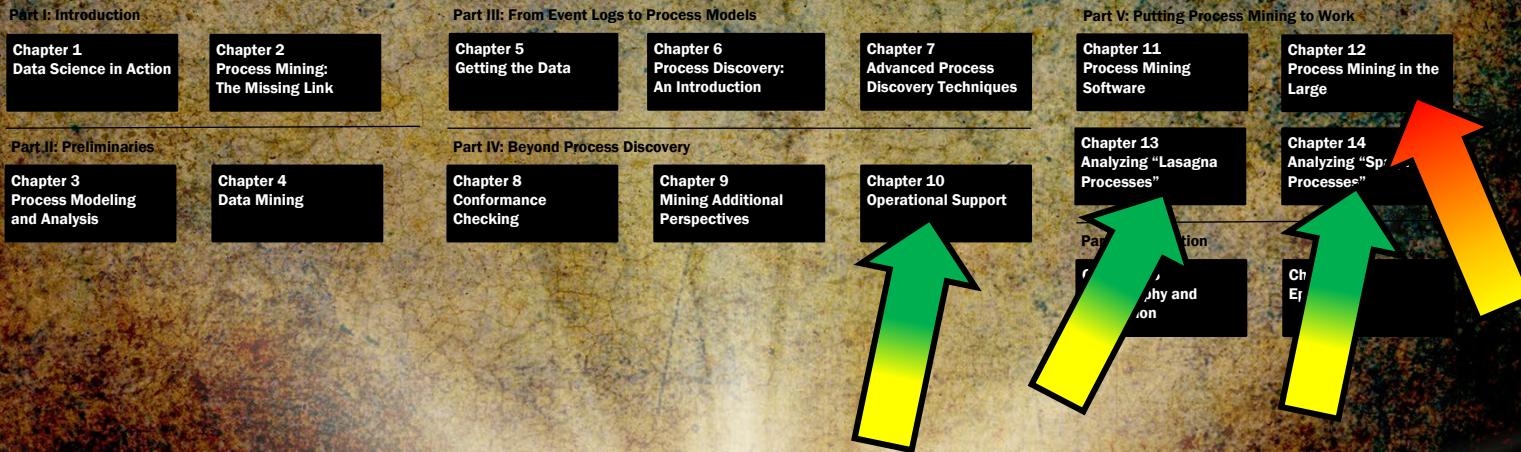
- Detect and remove bottlenecks
- Analyze root causes for deviations
- Comparative process mining
- Operational support (prediction, etc.)



- Process discovery: show mirror
- Highlight important problems
- From “management by PowerPoint” to “evidence-based” process improvement
- ...



Chair of Process
and Data Science



ID	Topic	Date	Date	Place
	Lecture 1 Introduction to Process Mining	08.04.24	Monday	AH V
	Lecture 2 Data Science: Supervised Learning	09.04.24	Tuesday	AH V
	<i>Exercise 1 Tool Introduction</i>	09.04.24	Tuesday	AH III
	Lecture 3 Data Science: Unsupervised Learning and Evaluation	15.04.24	Monday	AH V
	Lecture 4 Introduction to Process Discovery	16.04.24	Tuesday	AH V
	<i>Exercise 2 Data Mining</i>	16.04.24	Tuesday	AH III
	Lecture 5 Alpha Algorithm 1	22.04.24	Monday	AH V
	Lecture 6 Alpha Algorithm 2	23.04.24	Tuesday	AH V
	<i>Exercise 3 Petri Nets</i>	23.04.24	Tuesday	AH III
	Lecture 7 Model Quality Representation	29.04.24	Monday	AH V
	Lecture 8 Heuristic Mining	30.04.24	Tuesday	AH V
	<i>Exercise 4 Alpha Miner</i>	30.04.24	Tuesday	AH III
	Lecture 9 Region-Based Mining	06.05.24	Monday	AH V
	<i>Exercise 5 Heuristic Mining and Region-Based Mining</i>	07.05.24	Tuesday	AH III
	Lecture 10 Inductive Mining	13.05.24	Monday	AH V
	Lecture 11 Event Data and Exploration	14.05.24	Tuesday	AH V
	<i>Exercise 6 Inductive Mining</i>	14.05.24	Tuesday	AH III
	Lecture 12 Conformance Checking 1	27.05.24	Monday	AH V
	Lecture 13 Conformance Checking 2	28.05.24	Tuesday	AH V
	<i>Q&A Session Assignment Part I</i>	28.05.24	Tuesday	AH III
	Deadline Assignment Part I	02.06.24	Sunday	
	<i>Exercise 7 Footprint and Token-Based Replay (Exercise)</i>	03.06.24	Monday	AH V
	<i>Exercise 8 Alignments (Exercise)</i>	04.06.24	Tuesday	AH V
	Lecture 14 Decision Mining	10.06.24	Monday	AH V
	<i>Lecture 15 Celonis Guest Lecture</i>	11.06.24	Tuesday	AH V
	<i>Exercise 9 Decision Mining</i>	11.06.24	Tuesday	AH III
	Lecture 16 Performance Analysis and Organizational Mining	17.06.24	Monday	AH V
	<i>Exercise 10 Performance Analysis (Exercise)</i>	18.06.24	Tuesday	AH V
	<i>Exercise 11 Organizational Mining</i>	18.06.24	Tuesday	AH III
	<i>Exercise 12 Celonis Case Study</i>	24.06.24	Monday	AH V
	Lecture 17 Operational Support and Process Mining Applications	01.07.24	Monday	AH V
	Lecture 18 Distributed, Streaming, and Comparative Process Mining	02.07.24	Tuesday	AH V
	<i>Exercise 13 Operational Process Mining</i>	02.07.24	Tuesday	AH III
	Lecture 19 Closing	08.07.24	Monday	AH V
	<i>Q&A Session Assignment Part II</i>	09.07.24	Tuesday	AH III
	Deadline Assignment Part II	14.07.24	Sunday	
	<i>Q&A Session Exam</i>	16.07.24	Tuesday	AH III



Distributed, Streaming, and Comparative Process Mining

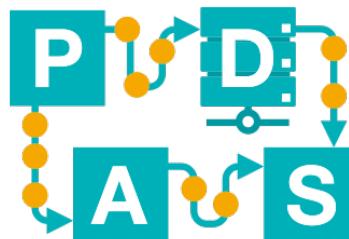
Lecture 18

prof.dr.ir. Wil van der Aalst

www.vdaalst.com @wvdaalst

www.pads.rwth-aachen.de

BPI-L18



Chair of Process
and Data Science

RWTHAACHEN
UNIVERSITY

Outline

- Dealing with **huge event logs**
- **Distributed/decomposed process mining**
- **Streaming process mining**
- **Comparative process mining**
- **Research challenges**





BIG DATA

VOLUME

DATA SIZE

VELOCITY

SPEED OF CHANGE

VARIETY

DIFFERENT FORMS
OF DATA SOURCES

VERACITY

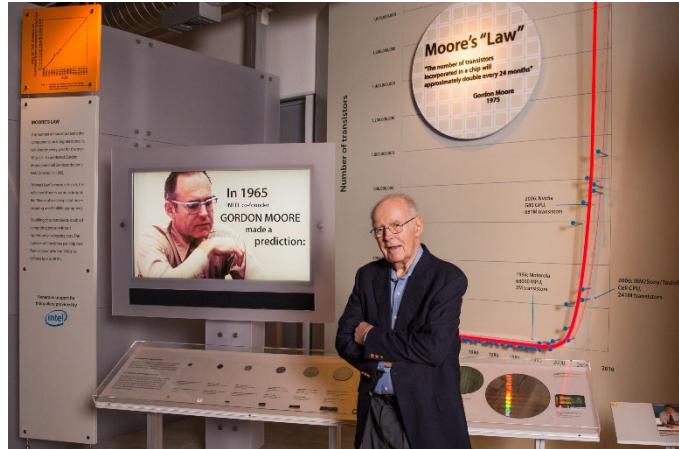
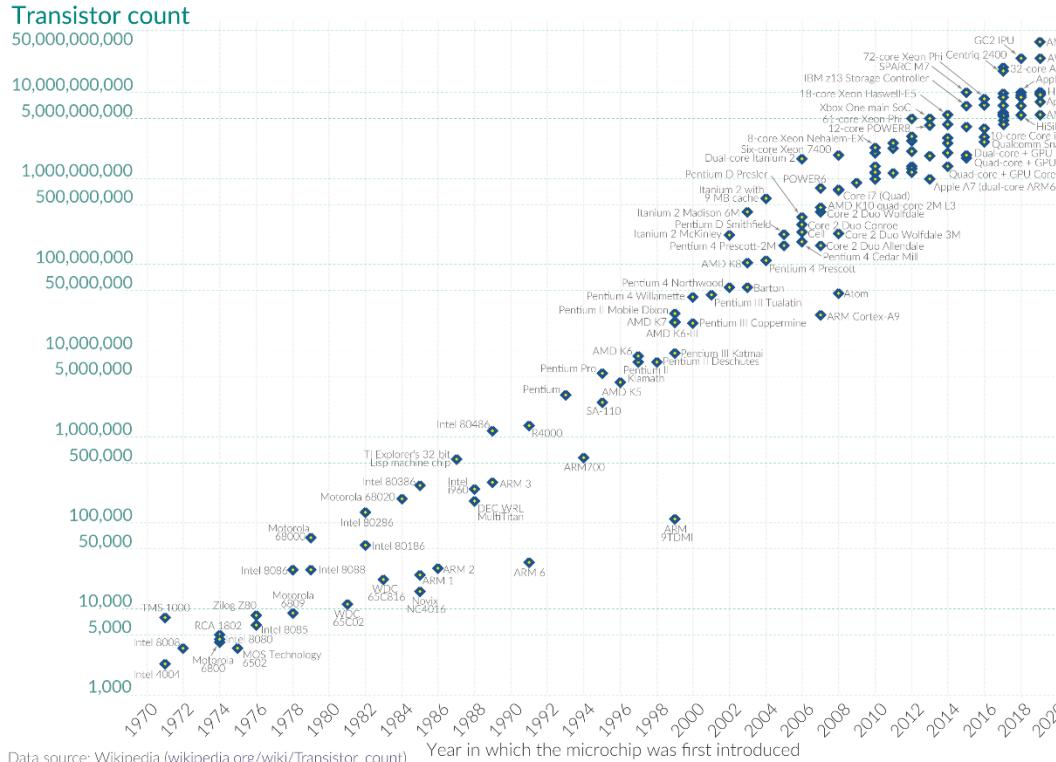
UNCERTAINTY OF
DATA

Moore's law

Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

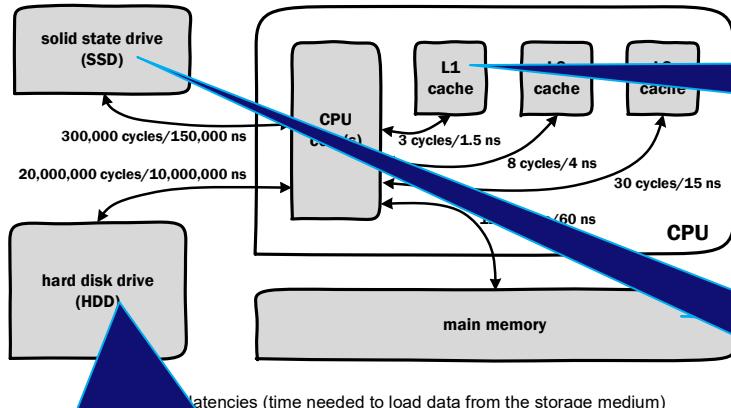
Our World
in Data



Gordon Moore(1929-2023)

“Distances” in computing

(numbers are from 2016, but principles did not change)



getting a
Nespresso from
the kitchen

getting a coffee from the
Starbucks around the
corner

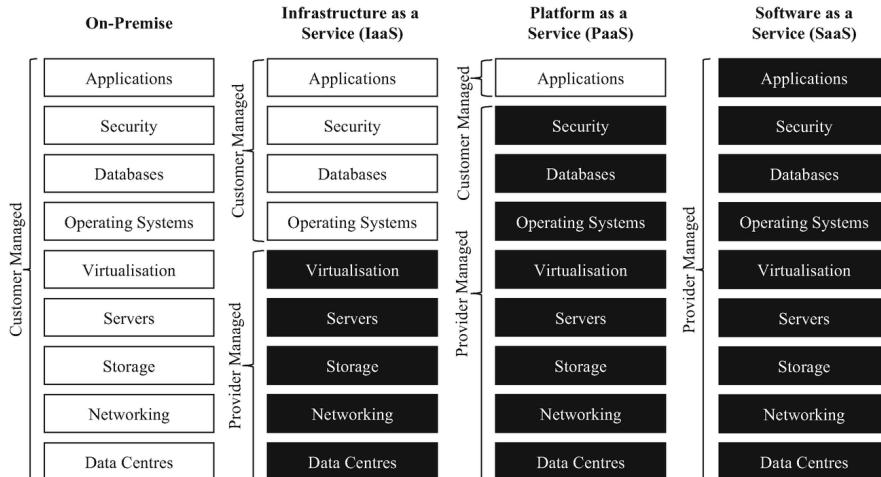
flying to Colombia, Ethiopia
and Kenya, process the beans
in Amsterdam, and then fly to
Rome

flying to Milano for a
coffee



Chair of Process
and Data Science

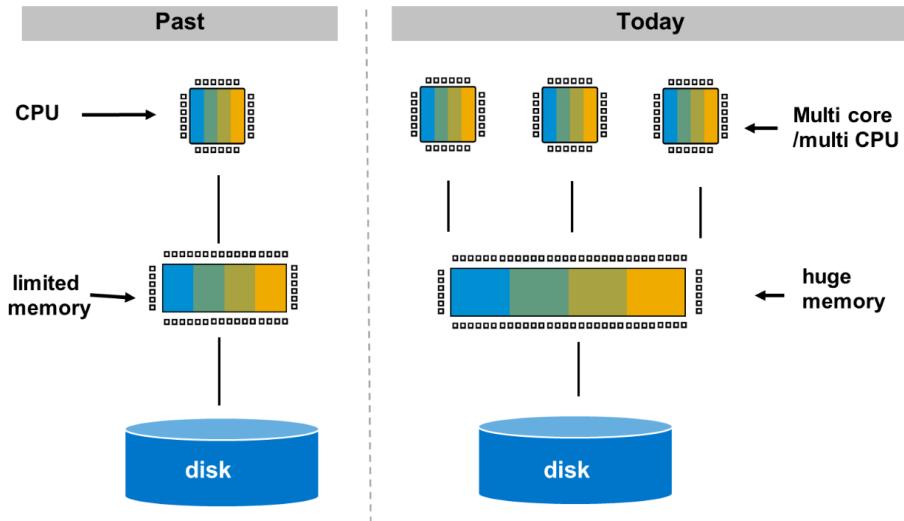
Moving to the cloud to ensure scalability



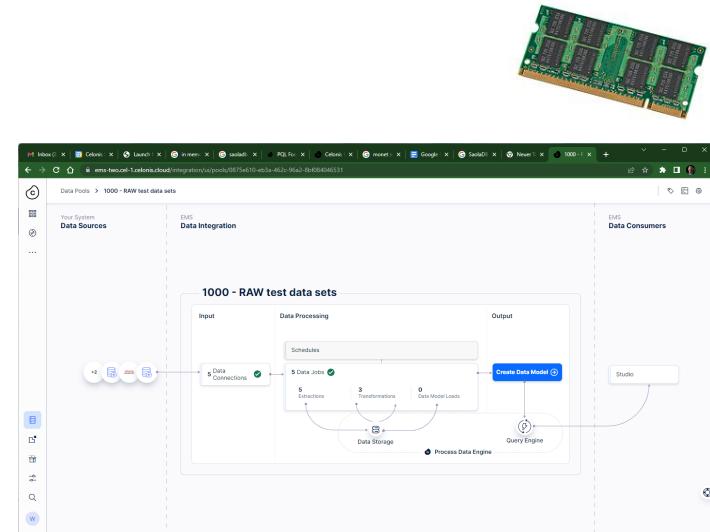
Pierangelo Rosati & Theo Lynn. Measuring the Business Value of Cloud Computing (2020),
https://doi.org/10.1007/978-3-030-43198-3_2

Example: SAP Hana and SaolaDB

SAP HANA is an in-memory, column-oriented, relational database management system



SaolaDB, the core of the Celonis engine, is a dedicated in-memory database, much faster than state-of-the-art databases (because it is tailored towards event data)

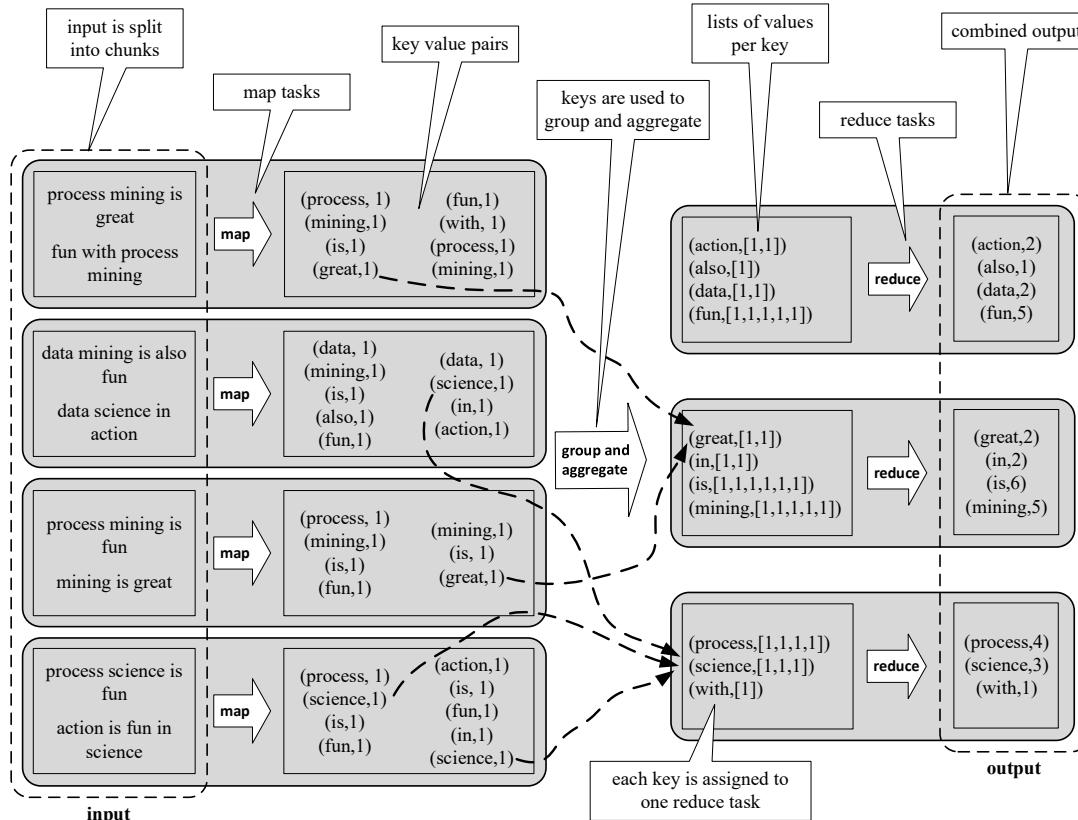




distributing the work

Example distribution: MapReduce

Later used to build a DFG



Characterizing event logs and streams



Event data

- In most cases, all relevant event data is stored in logs/tables: This is the setting we assumed thus far.
- However, the volume may be so large that events **cannot be stored** (or even when they can be stored there is no possibility to **do computations over them later**).
- Moreover, even when all event data can be stored, **certain types of analysis may be intractable or simply take too long**.
- **Questions:** How to store the data? How to characterize events? What if we cannot store the data?

Data warehouse versus data lake

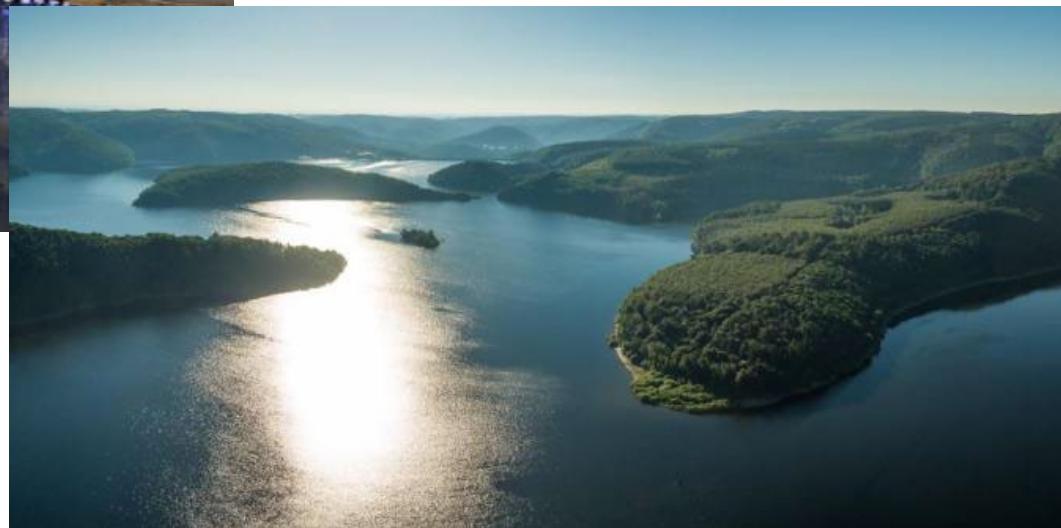


Data warehouse:

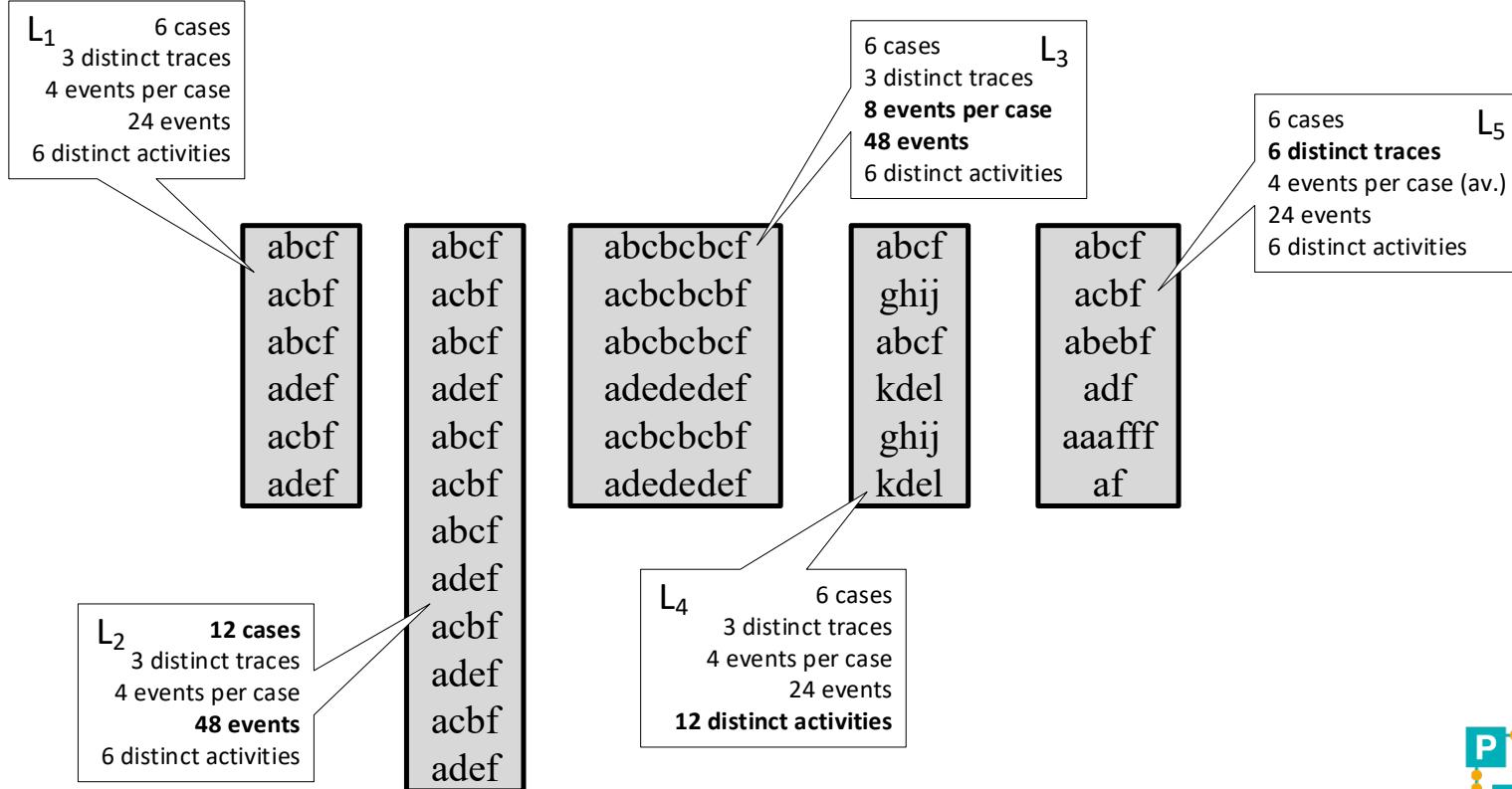
- Structured
- Cleaned first
- Schema predefined
- Inflexible

Data lake:

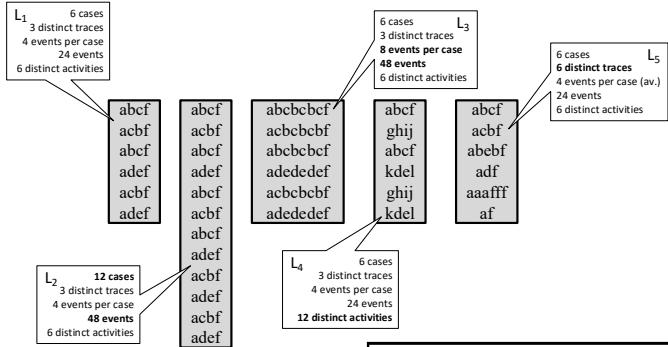
- Unstructured
- Cleaned later
- Schema on read
- Flexible



A few tiny event logs showing the diversity



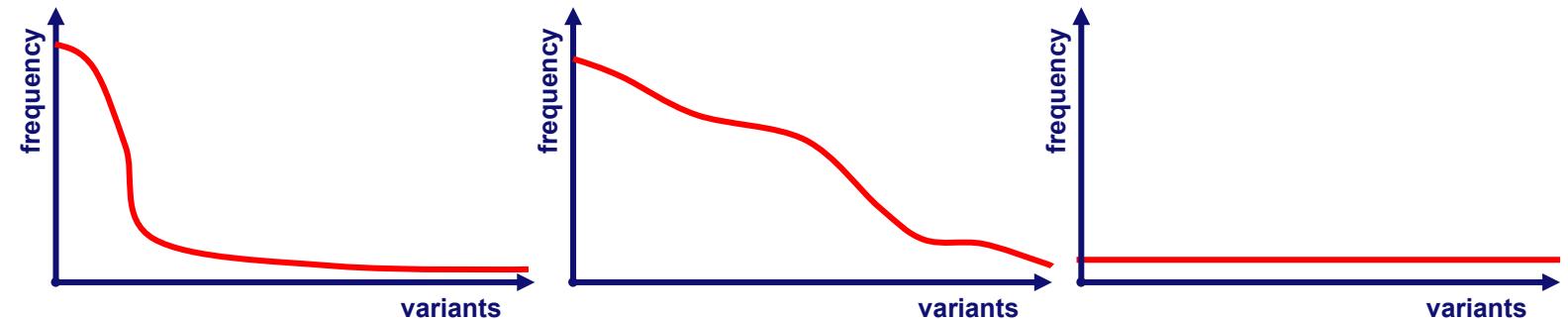
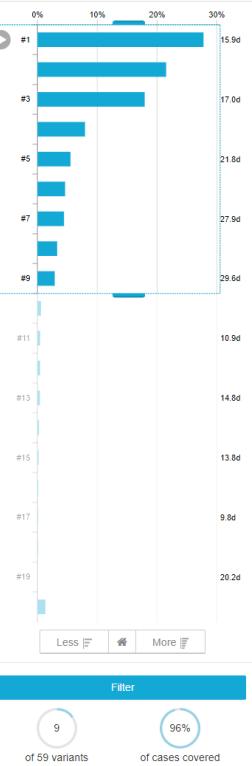
Different event log metrics



$$av_{sbnot}(L) = 1 - \frac{\sum_{\sigma_1, \sigma_2 \in L} L(\sigma_1) \times L(\sigma_2) \times \frac{|\partial_{set}(\sigma_1) \cap \partial_{set}(\sigma_2)|}{|\partial_{set}(\sigma_1) \cup \partial_{set}(\sigma_2)|}}{|L|^2}$$

event log metric	L_1	L_2	L_3	L_4	L_5
number of cases	$\#cases(L_i)$	6	12	6	6
average trace length of cases	$av_{tloc}(L_i)$	4	4	8	4
number of distinct activities	$\#acts(L_i)$	6	6	6	12
average number of dist. act. per case	$av_{dapc}(L_i)$	4	4	4	4
average set-based non-overlap of traces	$av_{sbnot}(L_i)$	0.296	0.296	0.296	0.667
number of distinct cases	$\#dc(L_i)$	3	3	3	3
number of events	$\#events(L_i)$	24	48	48	24
number of direct successions	$\#ds(L_i)$	9	9	10	9
number of start activities	$\#sa(L_i)$	1	1	1	3
number of end activities	$\#ea(L_i)$	1	1	1	1

Pareto distribution?



- “Zipability” (inverse of entropy)
- Minimum Description Length (MDL) principle

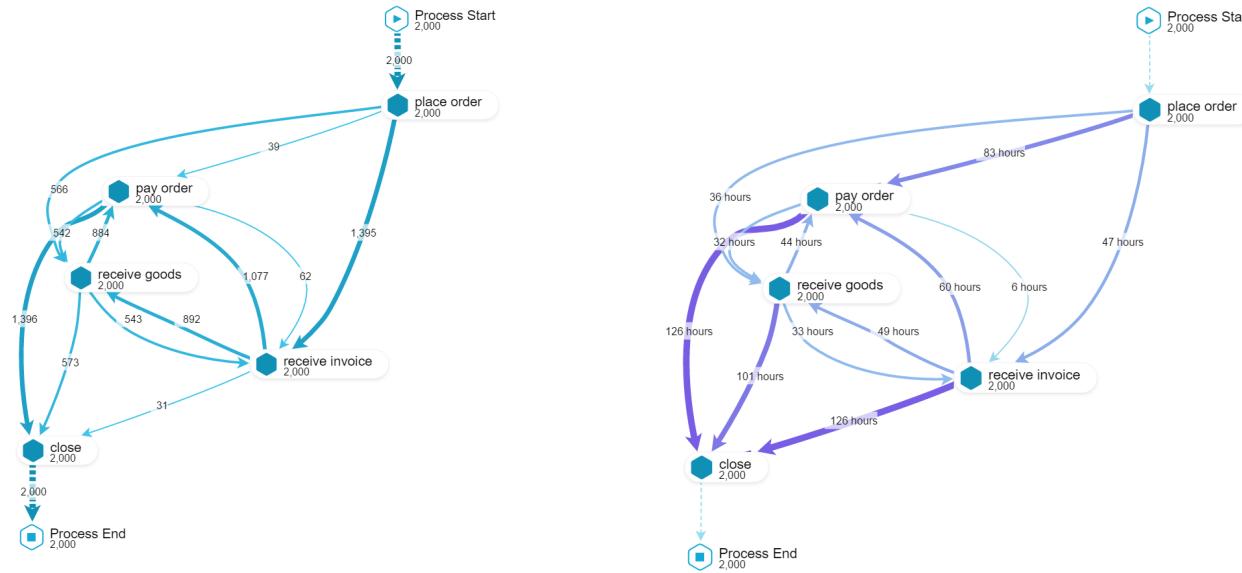
From previous lectures: How to simplify Spaghetti models?

- Project on subset of activities (<20).
- Group cases in to homogenous “clusters”.
- Sort variants by frequency.

Performance considerations for a few algorithms (informal)

Directly-Follows Graph (Discovery)

- One pass through the event data suffices
- Collect statistics (time/frequency) per pair of activities



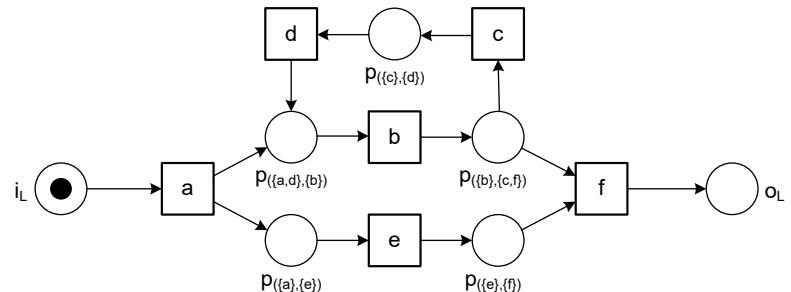
Alpha Algorithm (Discovery)

- One pass through the data suffices (like the DFG)
- The number of candidate places is $2^{|A|} \times 2^{|A|}$.
However, it is easy to rule out most candidates.

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

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

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



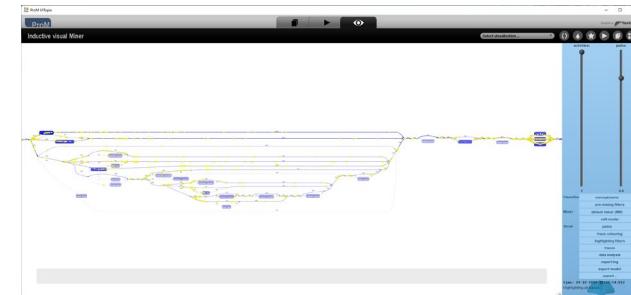
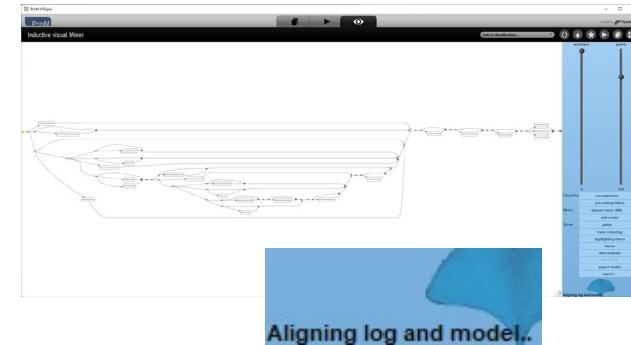
Inductive Mining (Discovery)

- The basic IM algorithm may need to **create many event logs** (recursive partitioning of activities). Approx. $2|A| - 1$ logs.*
- A **DFG is computed for each event log**, but event logs get smaller and more homogeneous.
- It is very effective to store **trace variants** instead of cases.
- There are also more efficient variants that work only on the DFG and do not create event logs (less guarantees).

Inductive Visual Miner: What takes time?

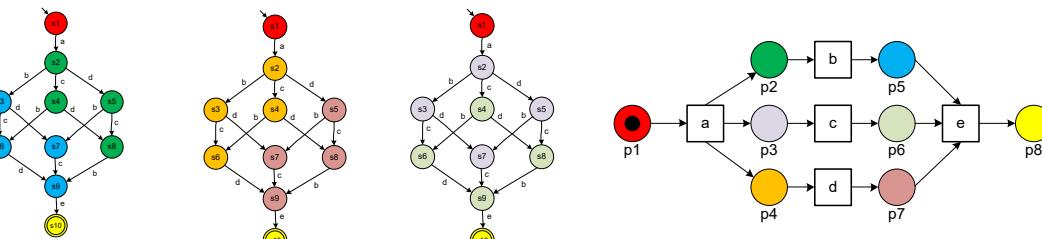
2654 cases, 16226 events, 24 activities

- Steps:
 - Preprocess the event log (<1 sec)
 - Create process tree (<1 sec)
 - Layout process tree (<1 sec)
 - **Compute alignments (180 sec)**
 - Add numbers in model (<1 sec)
 - Animate tokens (<1sec)

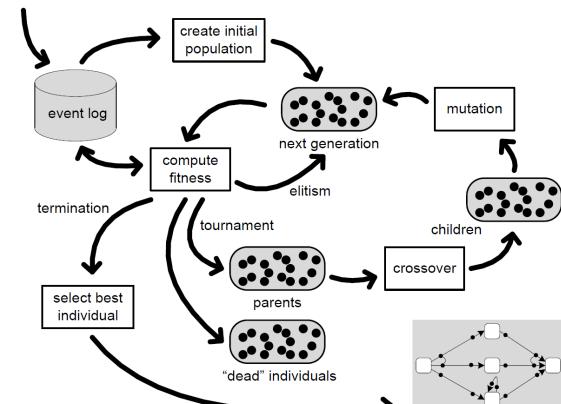


Other process discovery approaches

- Both state-based regions and language-based regions are very slow.
- Generic process discovery algorithms are also extremely slow.



$$c \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_a \\ x_b \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} y_a \\ y_b \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$



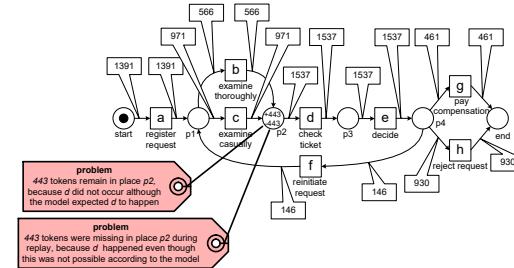
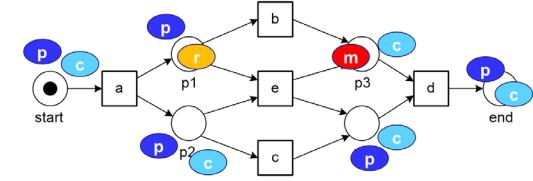
Comparing footprints (Conformance)

- Need to create a DFG for the event log and a reachability graph (or similar) for the process model.
- Hence, we need one pass through the event log and state space of the model (for specific models, e.g., free-choice nets and process trees this can be done faster).
- The actual comparison is fast.

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

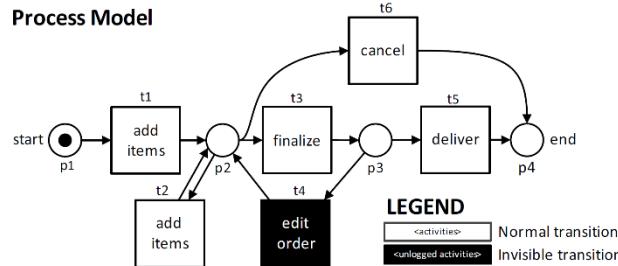
Token-Based Replay (Conformance)

- Need to go through the event log once and “play the token game” for each case.
- Efficient if there are no duplicate and silent activities.
- For duplicate and silent activities, state-space explorations are needed.
- Trace variants can be used to speed up the process, but timing information need to be collected per case.
- Recall that Celonis uses token-based replay (next to alignments).



Alignments (Conformance)

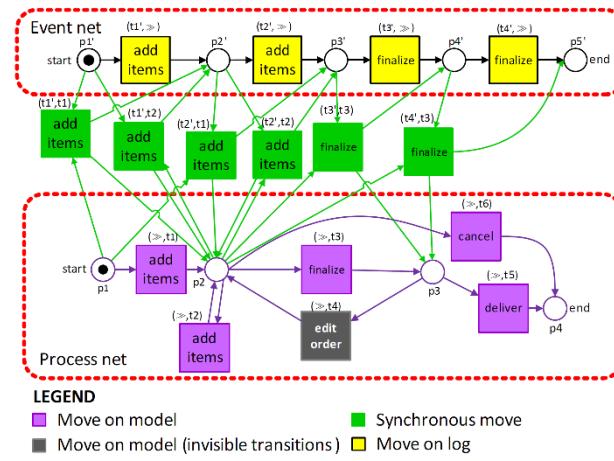
- By far the **most expensive approach**.
- Note that a process model typically has infinitely many possible traces (i.e., standard edit distance does not work).
- State-of-the-art techniques combine optimization (e.g., ILP, A*) with the Petri-net-based Marking Equation.



Trace : <add items, add items, finalize, finalize>

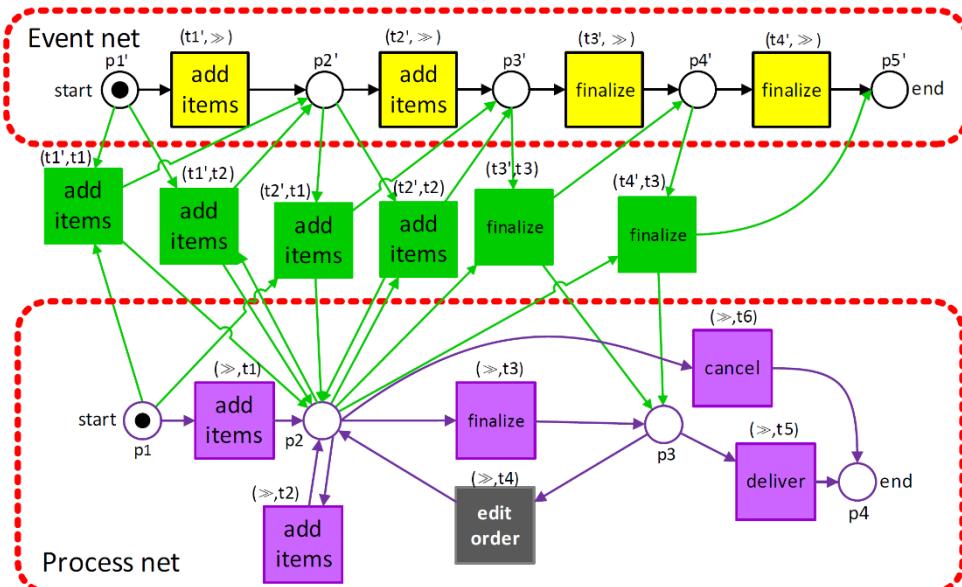
$$\gamma_1 = \begin{array}{|c|c|c|c|c|} \hline & [add\ items] & [add\ items] & [finalize] & [finalize] & \gg \\ \hline & t_1 & t_2 & t_3 & t_5 & \\ \hline \end{array}$$

$$\gamma_2 = \begin{array}{|c|c|c|c|c|c|} \hline & [add\ items] & [add\ items] & [finalize] & \gg & [finalize] & \gg \\ \hline & t_1 & t_2 & t_3 & t_4 & t_3 & t_5 \\ \hline \end{array}$$



Shortest path problem!

Product of a process model and an event net



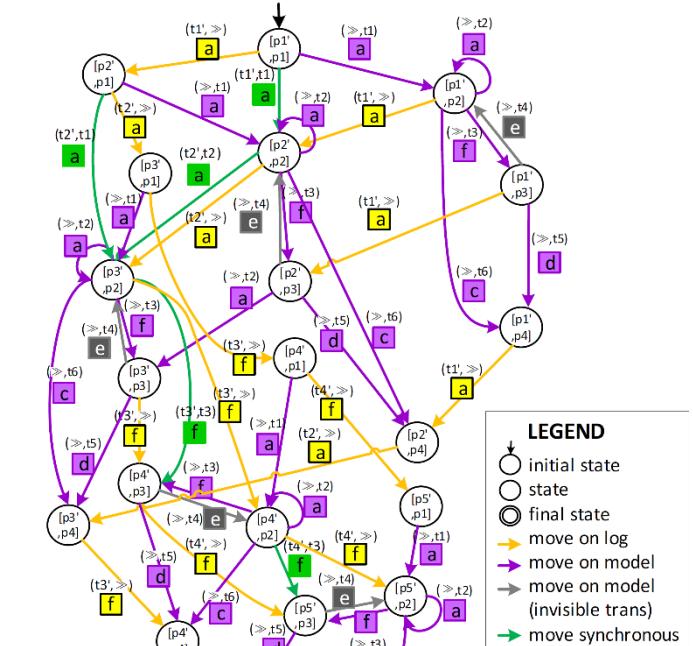
LEGEND

Move on model

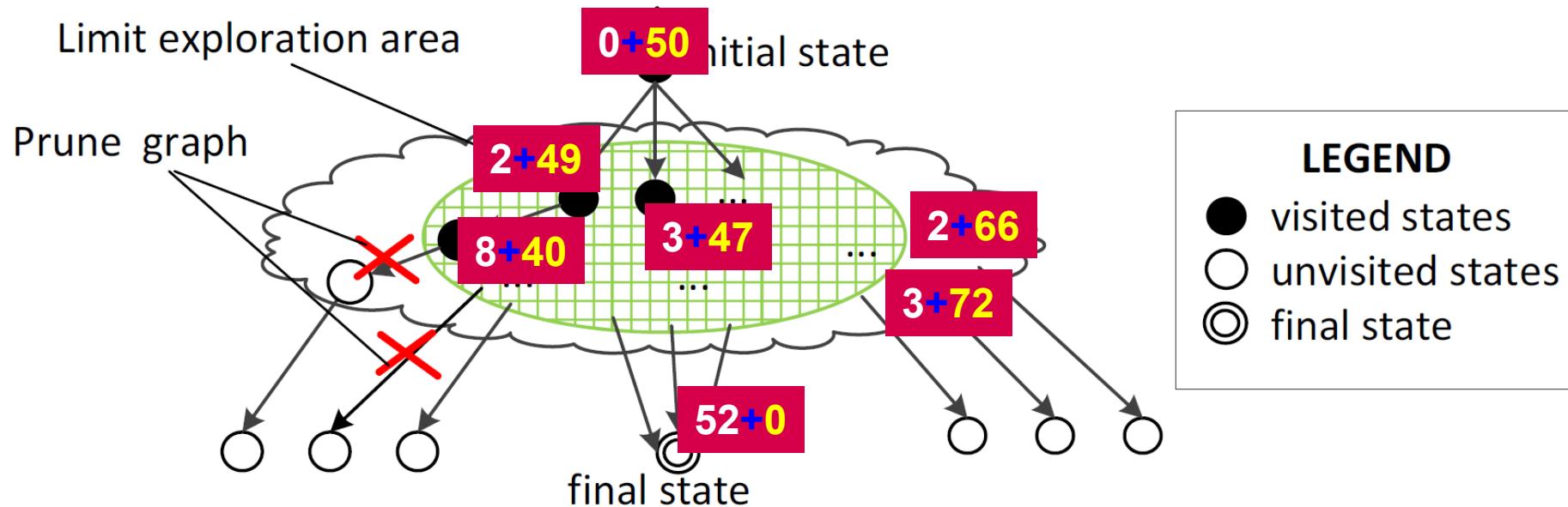
Move on model (invisible transitions)

Synchronous move

Move on log



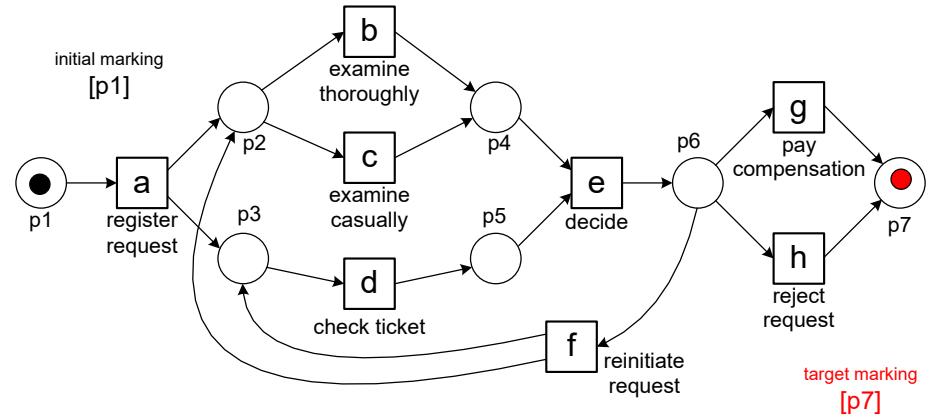
A* and Marking Equation



Use an underestimate based on the marking equation and apply the A* based approach. In an A* based state exploration, an underestimate is used to avoid exploring less promising candidates.

Marking Equation

(Just the idea)



$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Linear-algebraic characterization of all paths leading to the final marking that can be combined with cost information.

$$N = \begin{pmatrix} a & b & c & d & e & f & g & h \\ p1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ p2 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 \\ p3 & 1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ p4 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ p5 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ p6 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & -1 \\ p7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

- Provides a necessary requirement, but not a sufficient requirement.
- Can be used to prune state space.
- Can be used to provide an underestimate for the costs to reach the end.



Streaming process mining



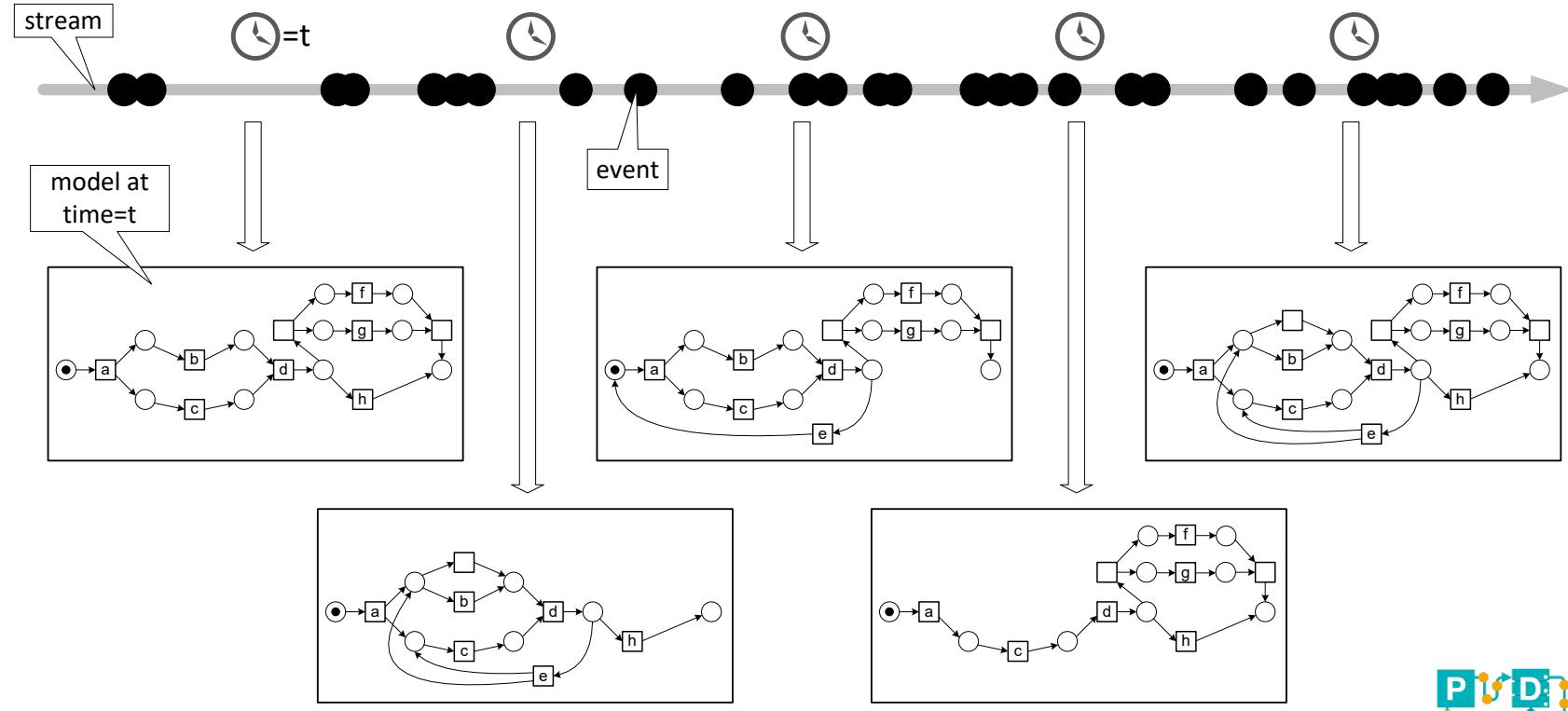
Drinking from a firehose



Fixed bounds on storage (cannot store everything) and (therefore) the need to handle the input as it arrives!

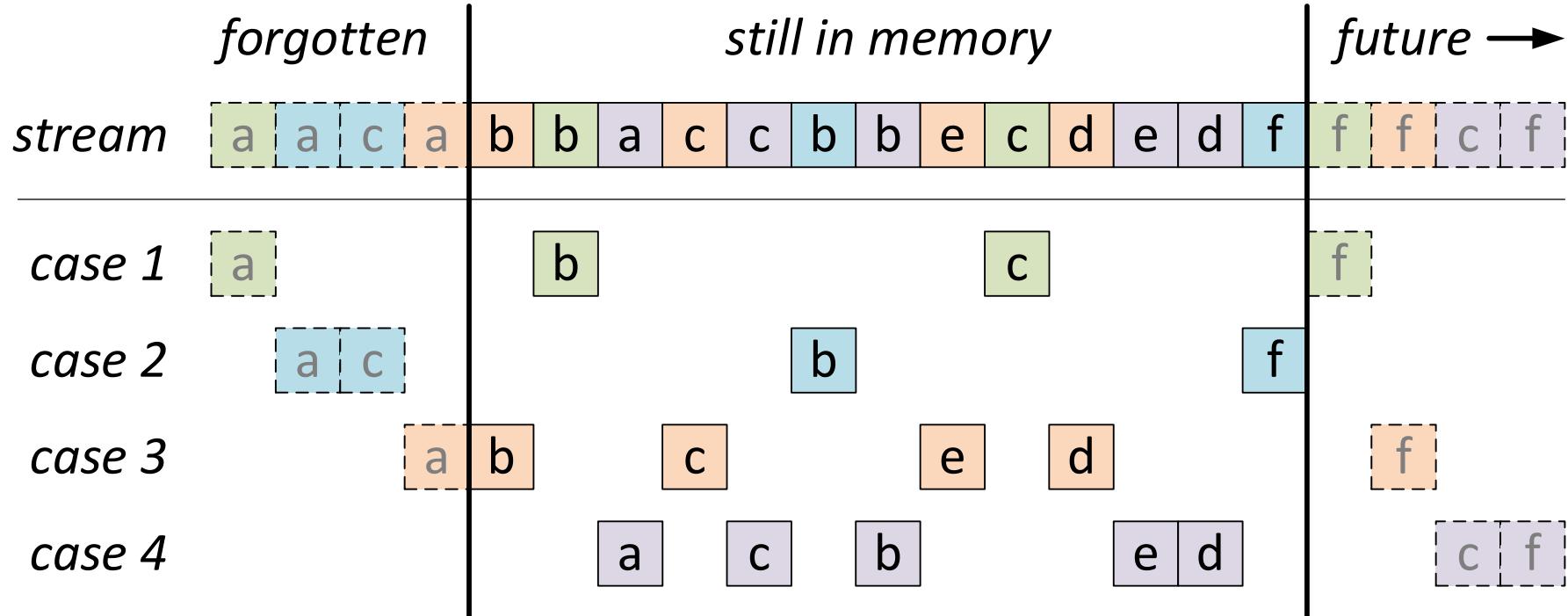
Event stream

Evolving models without storing events (related to concept drift)

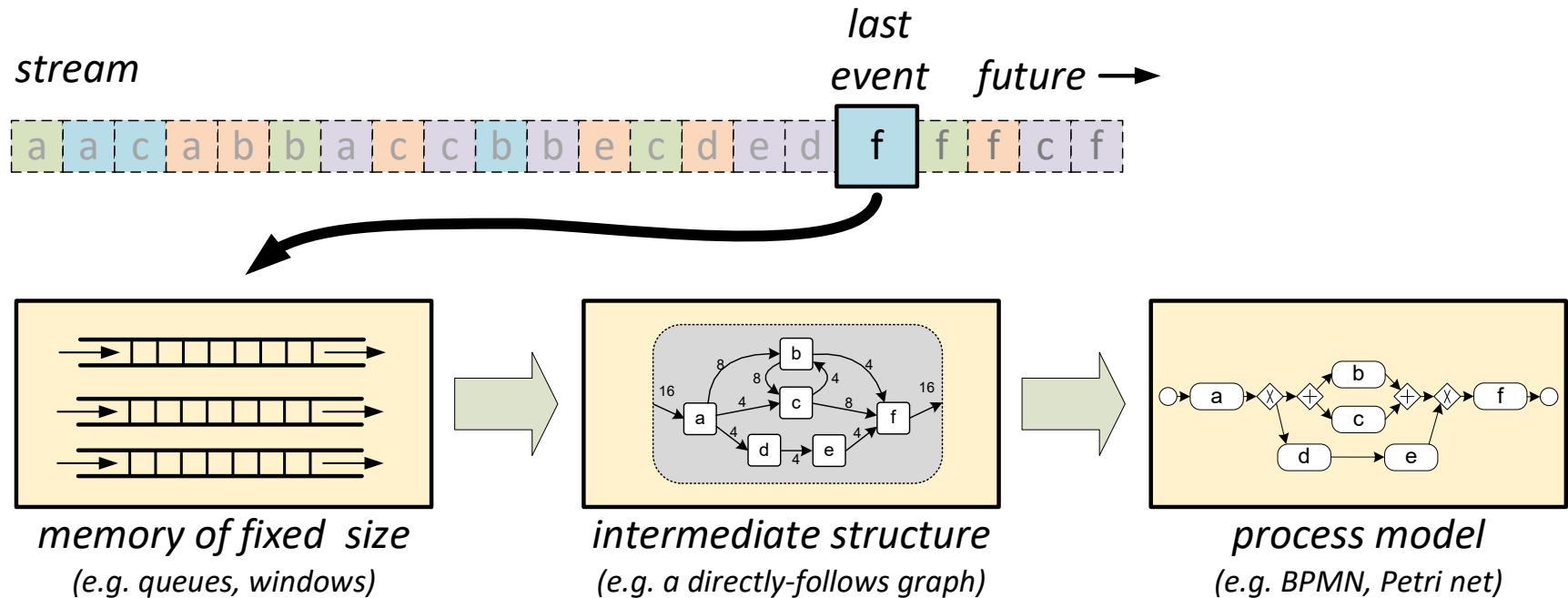


Using windows is not so easy

(one does not know when cases have ended and when there is no memory left deletions are unavoidable)

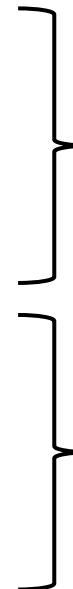


Architecture



Naïve DFG Streaming Algorithm

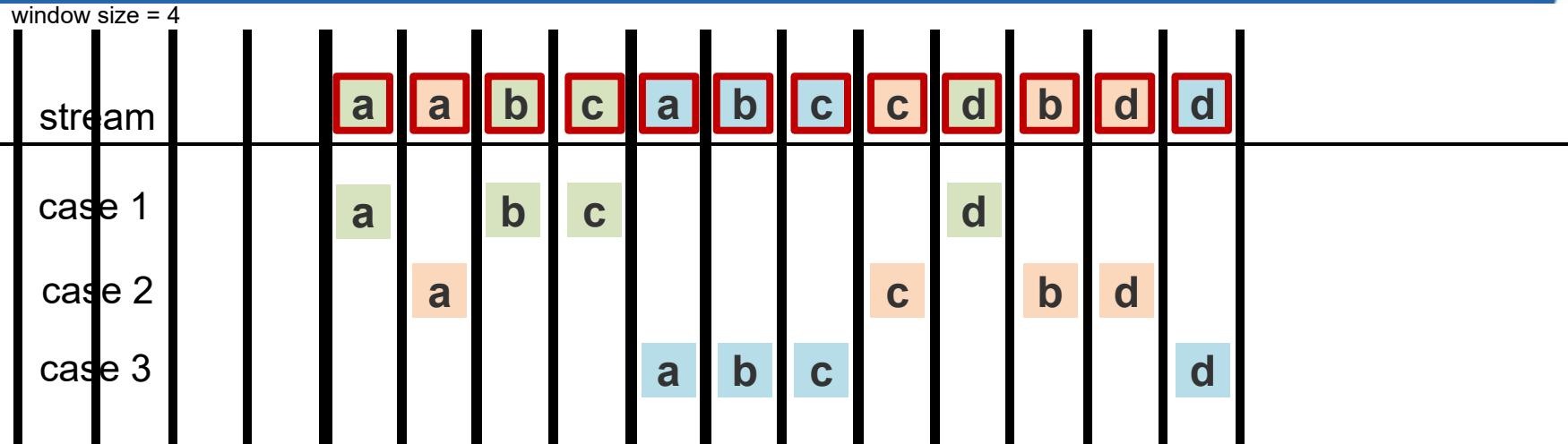
- New event e_{new} with activity a_{new} for caseId cid_{new} arrives
 - Before moving window
 - Determine event to remove (e_{remove})
 - If window size reached
 - $e_{remove} = \text{earliest event in memory with activity } a_{remove}$
 - If e_{remove} is the only event of its case in memory:
 - Add $(a_{remove}, \blacksquare)$ to DFG
 - Else: None
 - After moving window
 - Add e_{new} to memory
 - If e_{new} has predecessor event with activity a_{pre} in its case
 - Add (a_{pre}, a_{new}) to DFG
 - Else:
 - Add $(\triangleright, a_{new})$ to DFG
 - Update done



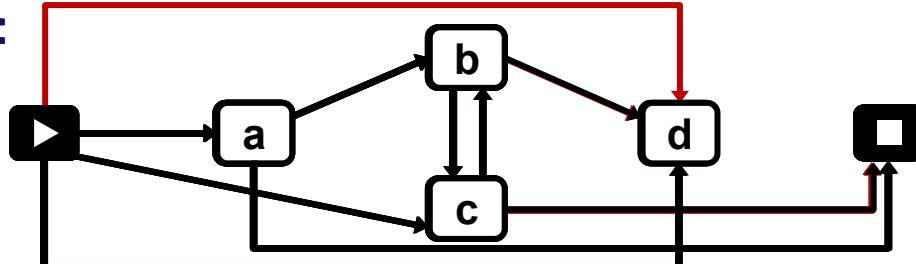
Check if case ends

Check if case starts or continues

Example: DFG Streaming Algorithm



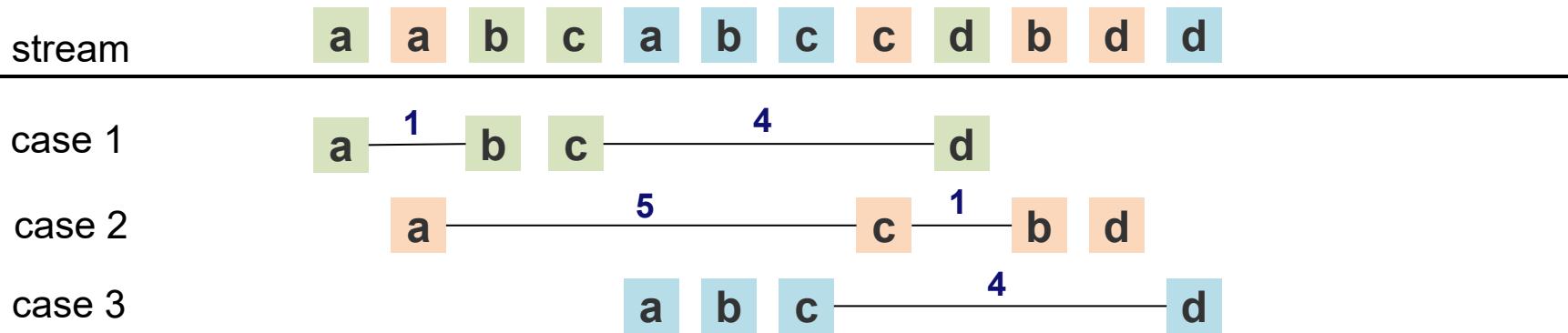
State after
Step 10:



Preserving Window Sizes

- Minimal **Multiset** Preserving Window Size
 - Minimal Window size so that for all cases all the directly follows relations are captured by the simple DFG streaming algorithm (for a given stream)
 - The last event recorded for a case is not removed unless it is the last event of the case. Hence, the DFG is correct.
- Minimal **Set** Preserving Window Size
 - Minimal window size so that all the directly follows relations are captured by the simple DFG streaming algorithm at least once (for a given stream)
 - The DFG may be incorrect and have **too many** connections (connections to start and end).

Example: Preserving Window Size



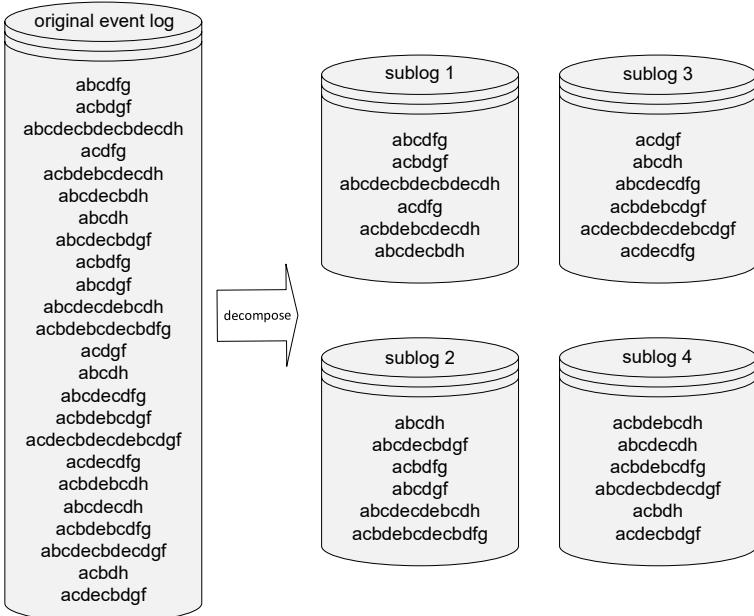
Compute distance between events of the same case.

Minimal multiset preserving window size = max distance + 2

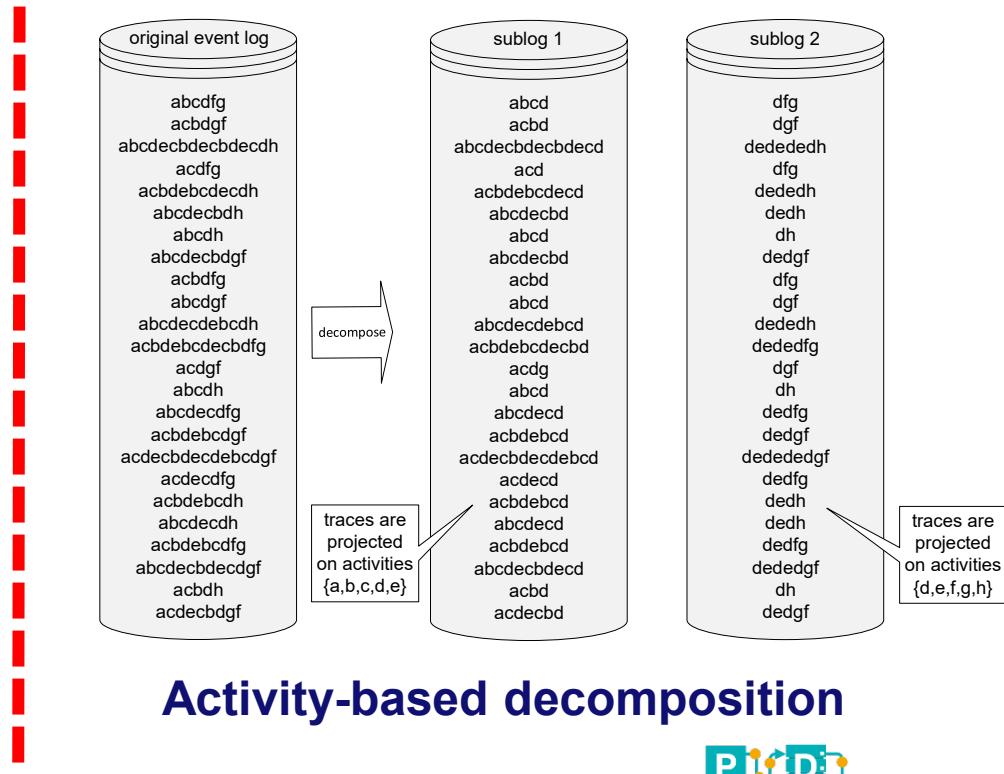
Decomposed/distributed process mining



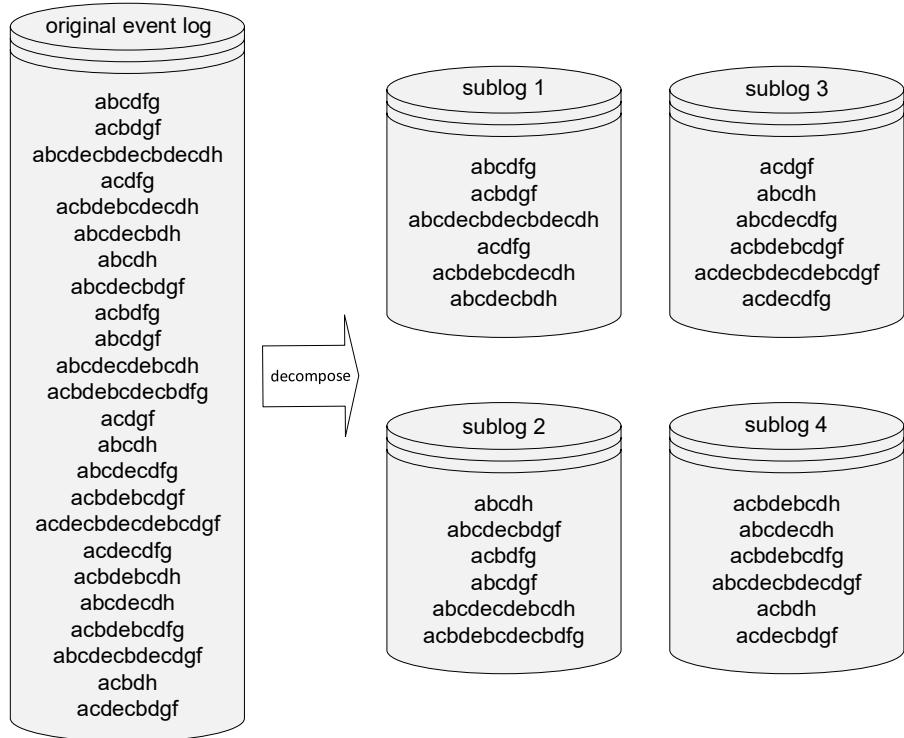
Two ways of decomposing



Case-based decomposition

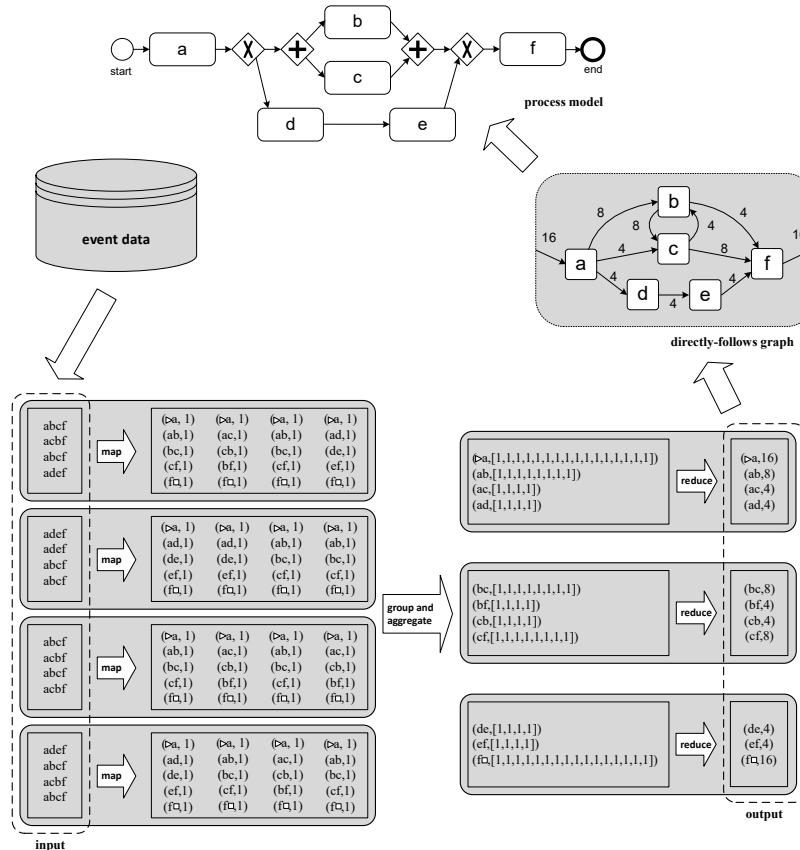


Case-based decomposition

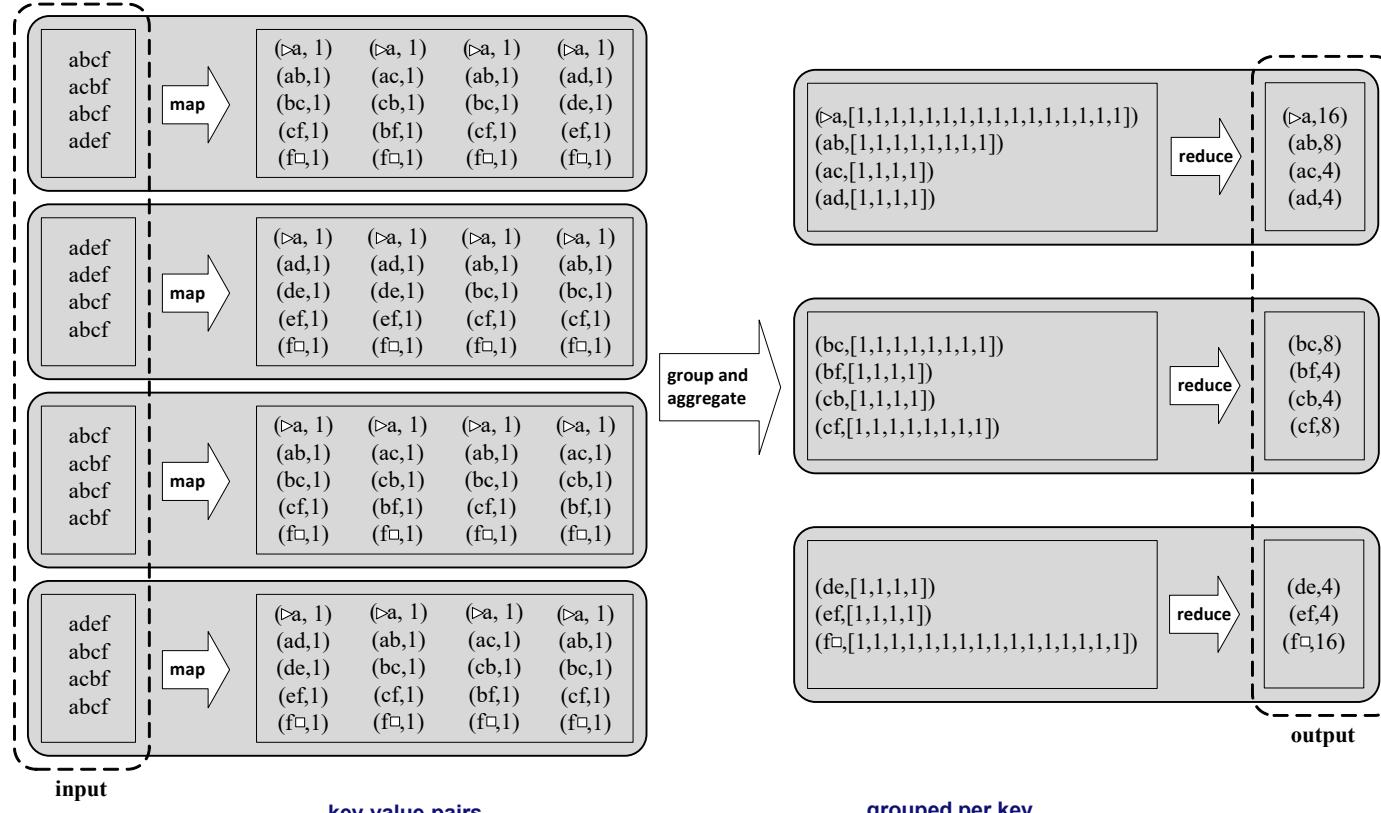


- Conformance checking can be decomposed / distributed in a trivial manner.
 - Discovery is also easy as long as it boils down to counting.

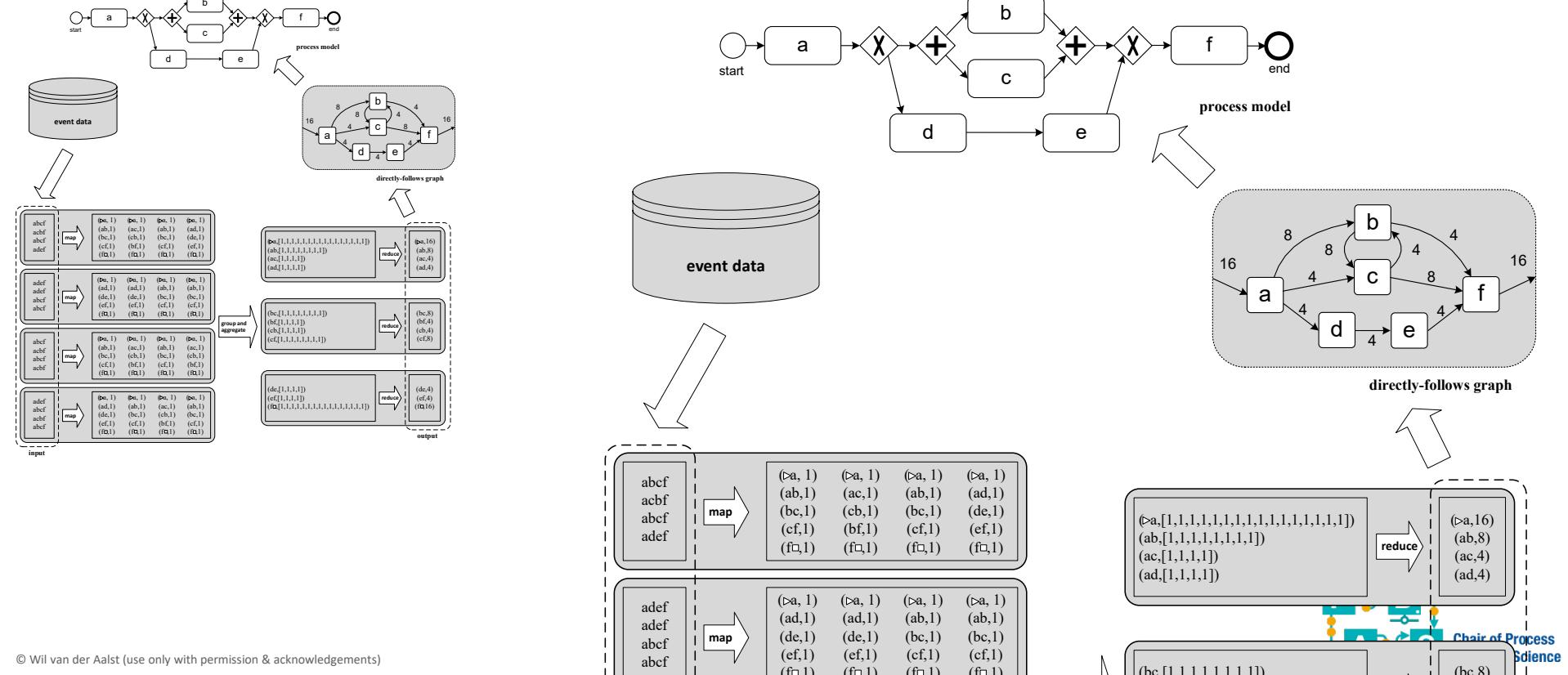
Case-based decomposition can be combined with MapReduce



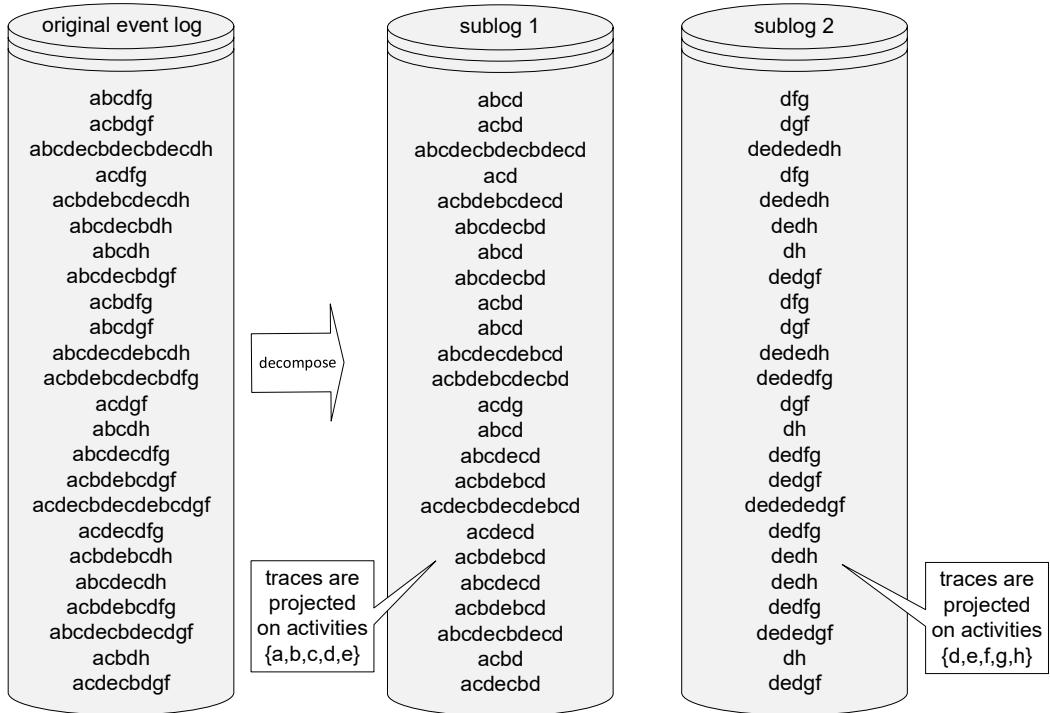
Can be combined with MapReduce



Can be combined with MapReduce

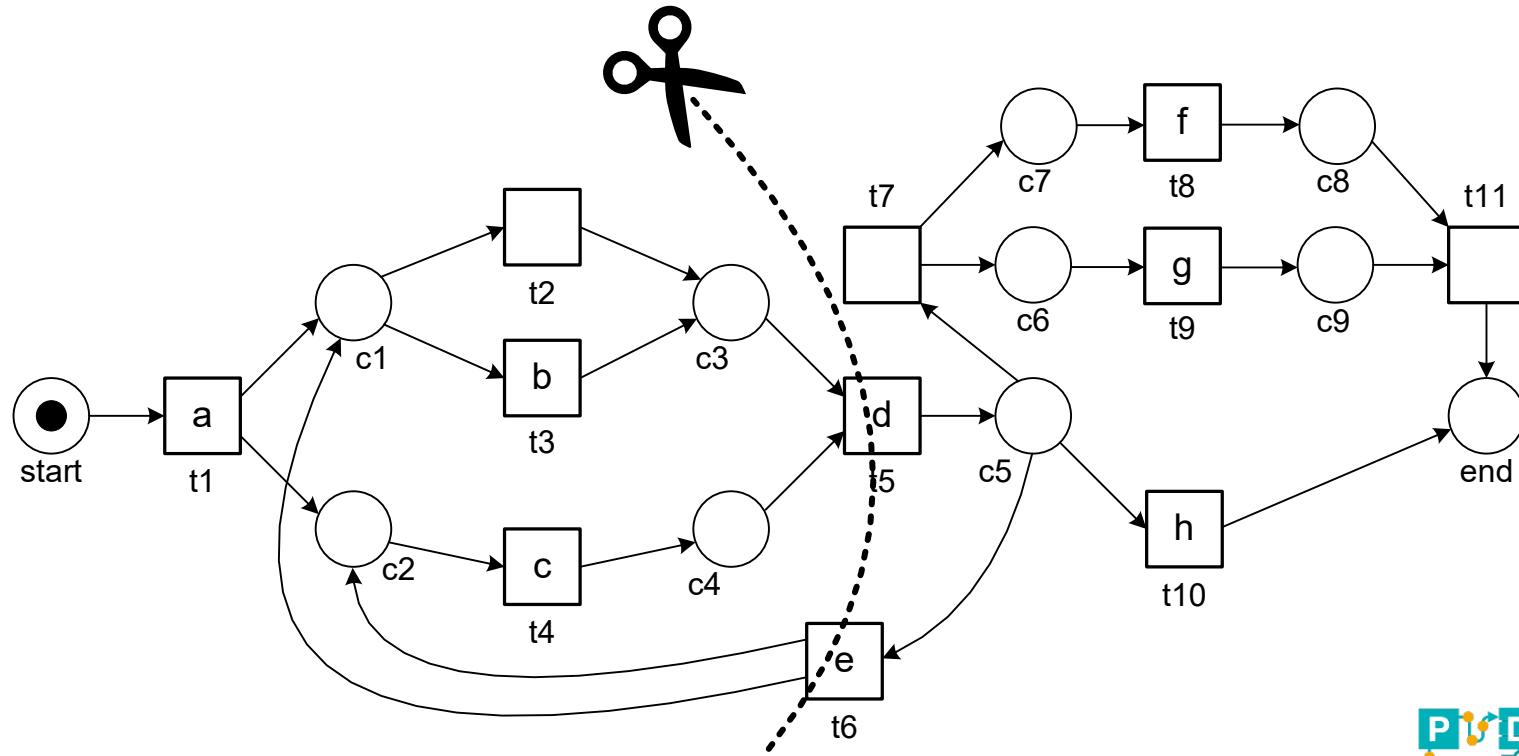


Activity-based decomposition

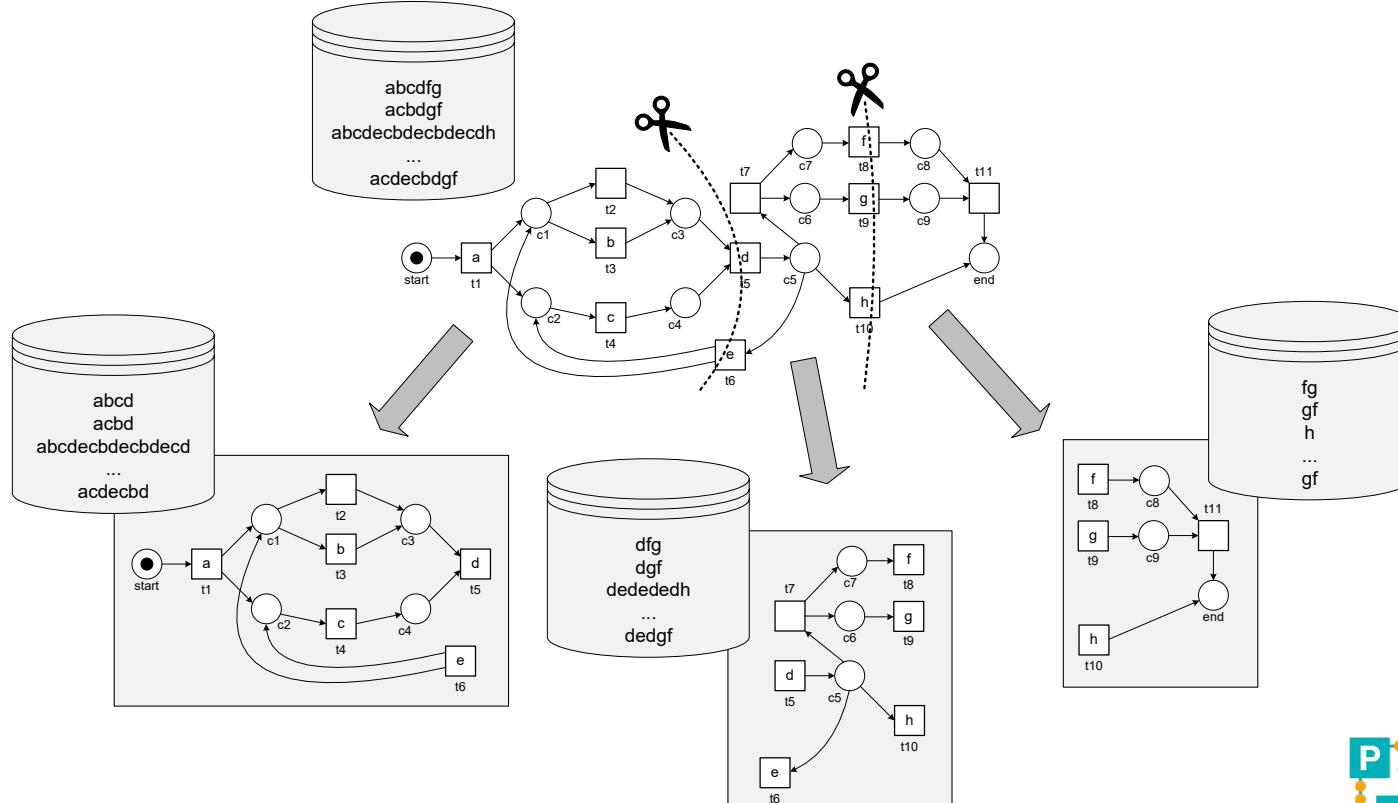


- Much more involved.
- But, many algorithms are linear in the size of the event log and exponential in the number of activities

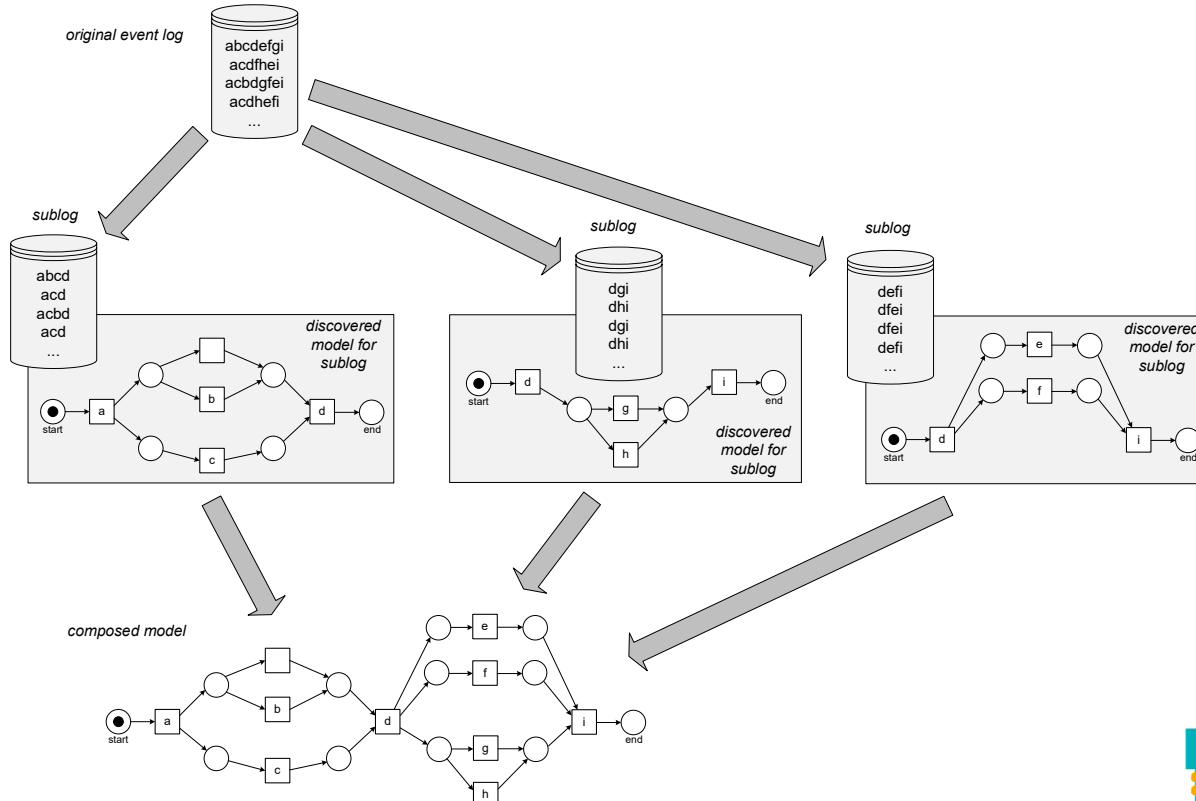
Activity-based decomposition



Conformance checking using activity-based decomposition



Discovery using activity-based decomposition



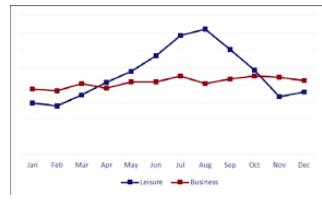
Comparative process mining



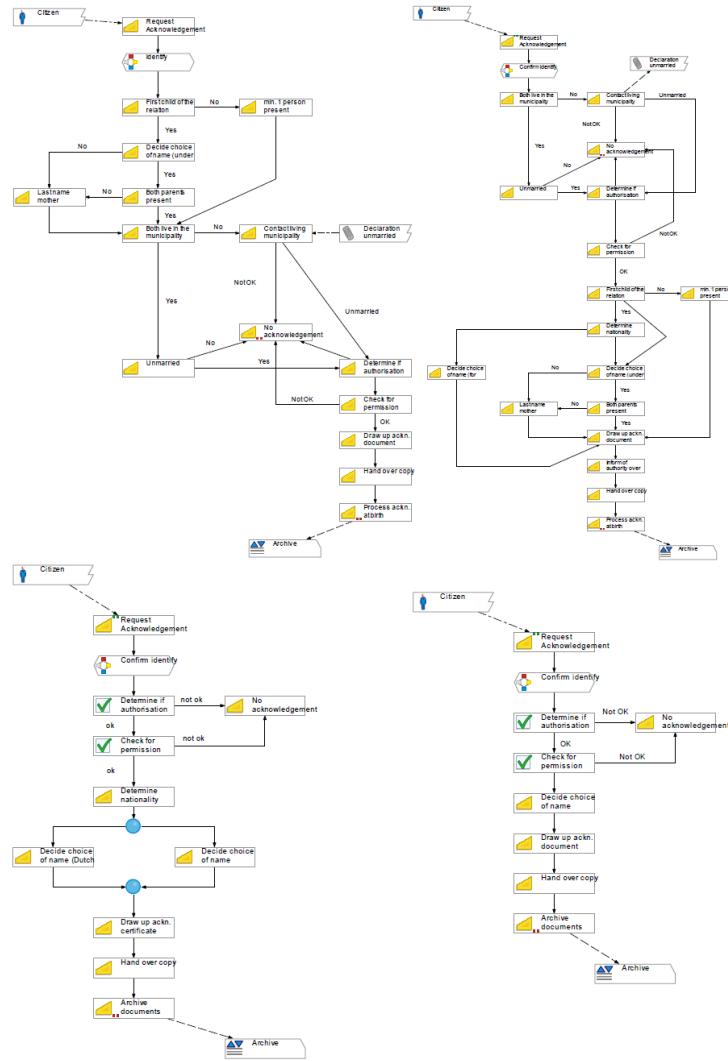
Comparative process mining

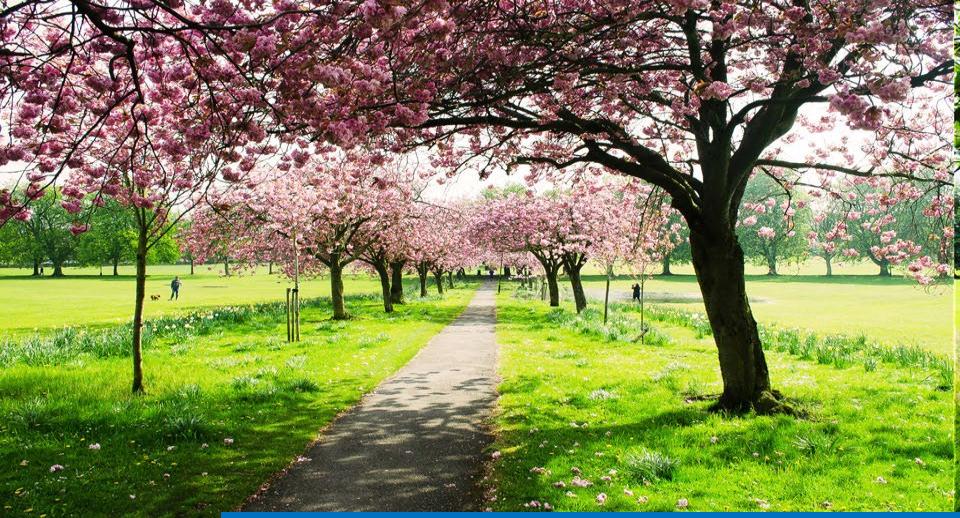


Hertz has 8,650 locations in 146 countries.

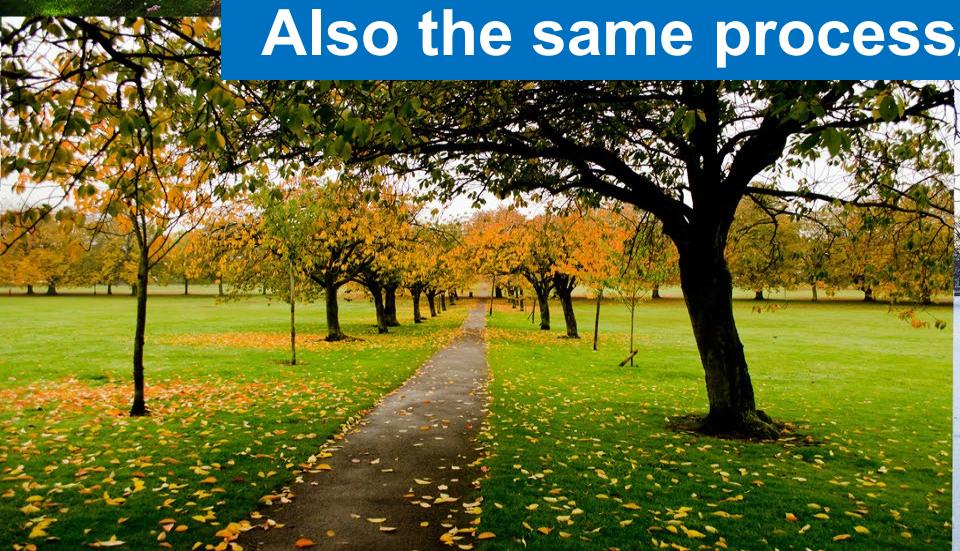


CoSeLoG





Also the same process/organization over time

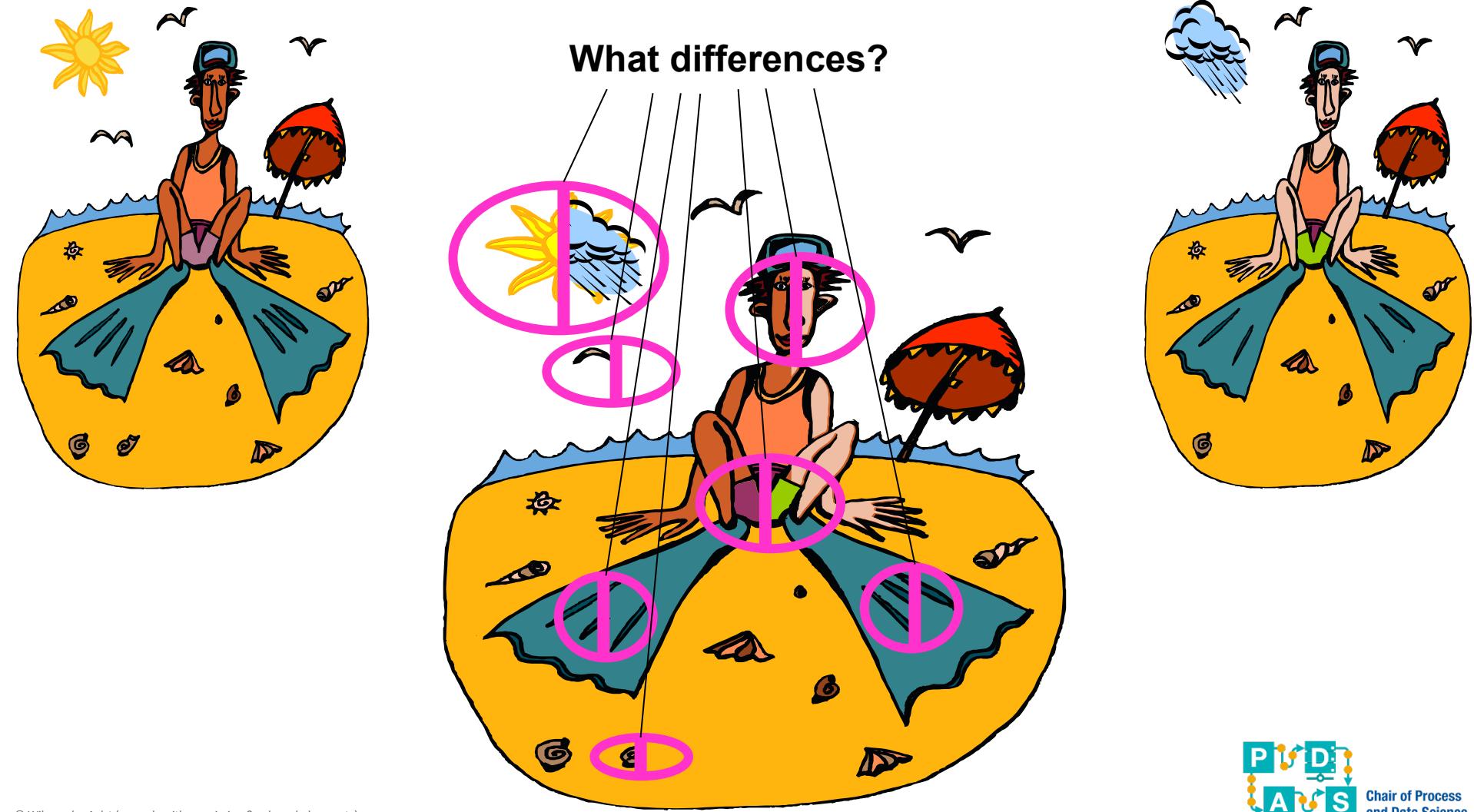




Find the seven differences

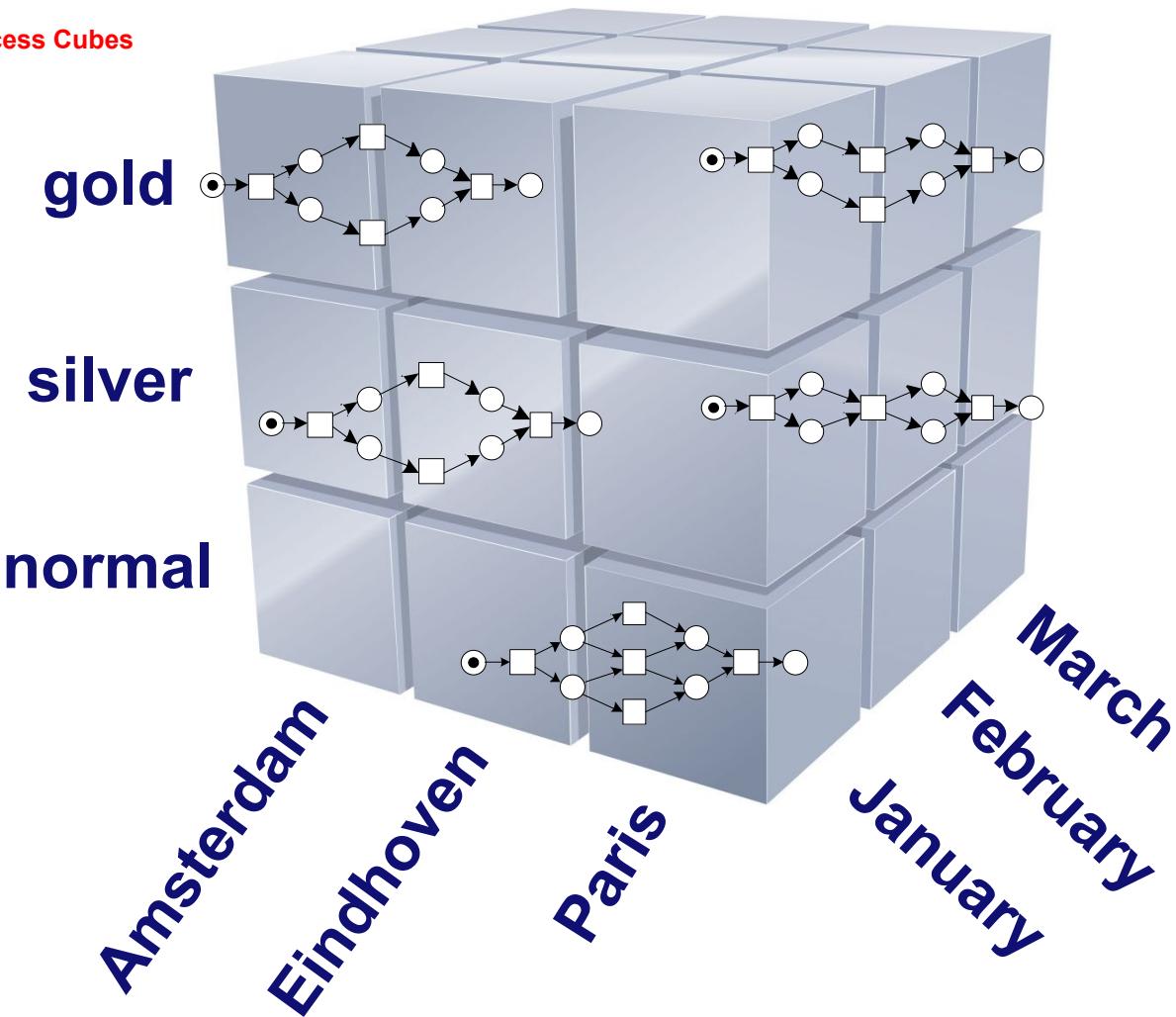


What differences?



Key questions

- What are the differences in terms of
 - control flow,
 - data flow,
 - resources,
 - time, etc.
- Are these differences good/bad?
- How to relate the differences (doing XXX leads to higher/lower YYY)?



Hertz has 8,650 rental locations and different types of customers.

large

medium

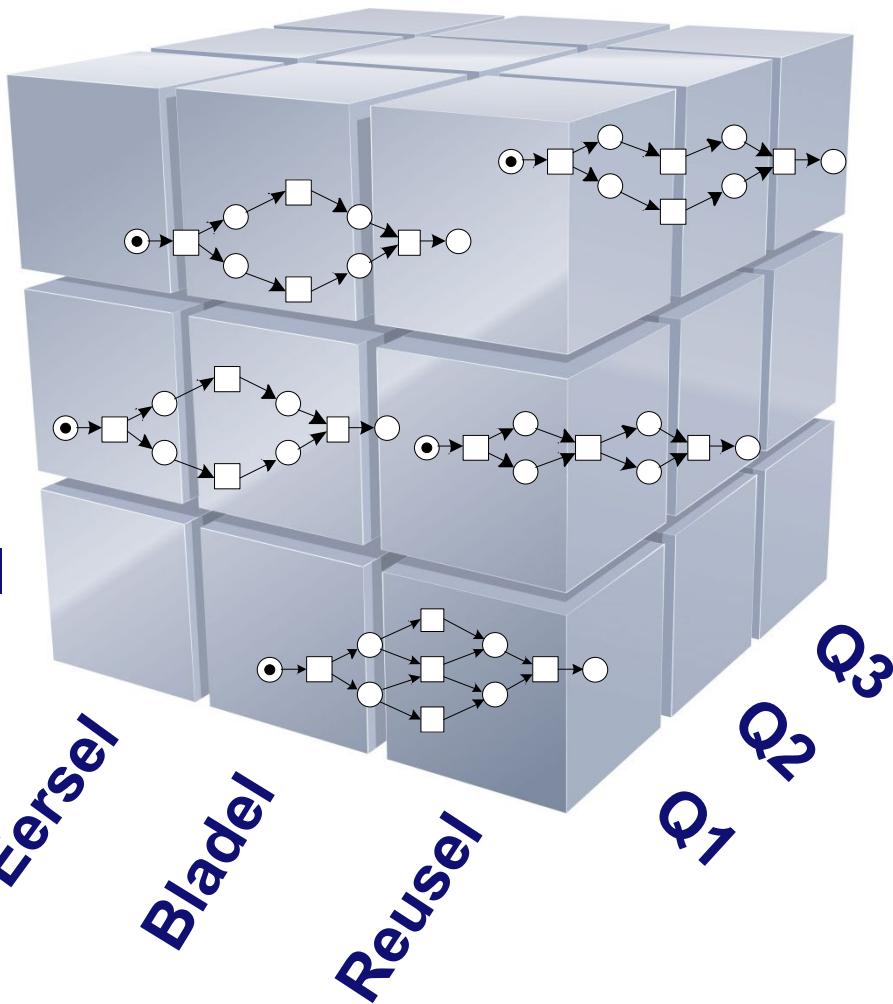
small

Eersel

Bladel

Reusel

Q_1 Q_2 Q_3



All Dutch municipalities are handing out building permits within the boundaries set by the Dutch law.

8-10

6-7

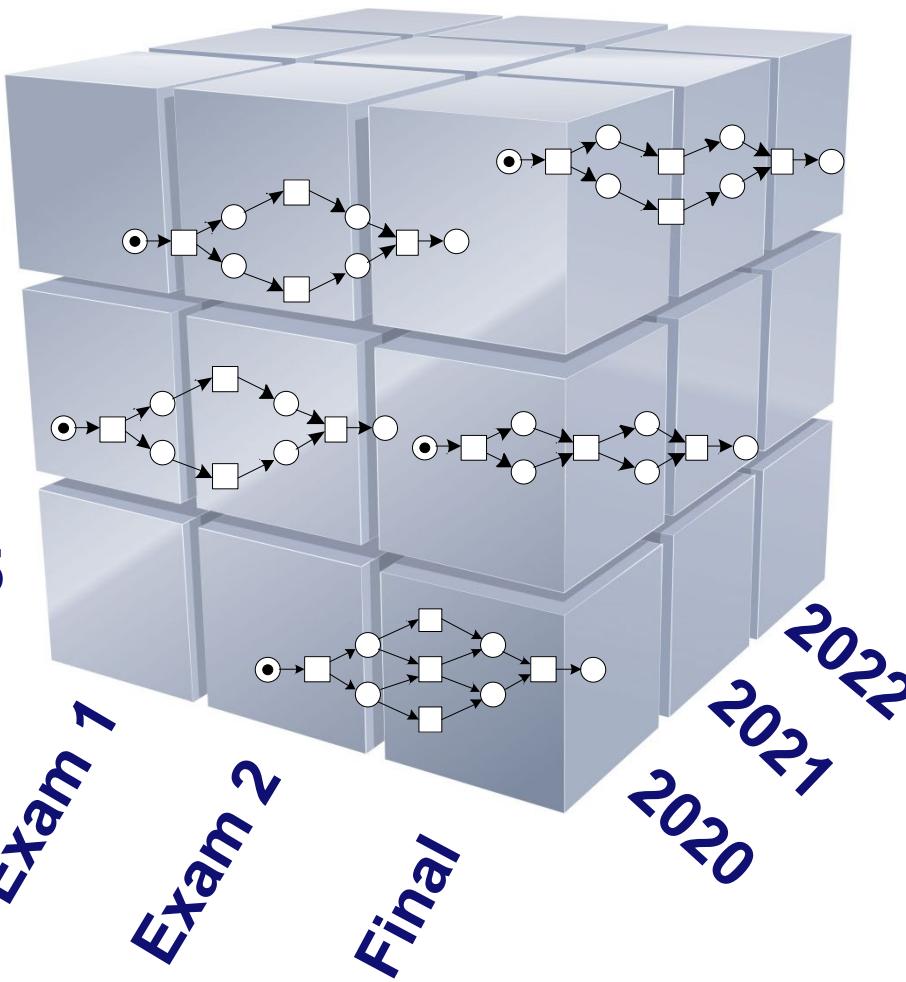
1-5

Exam 1

Exam 2

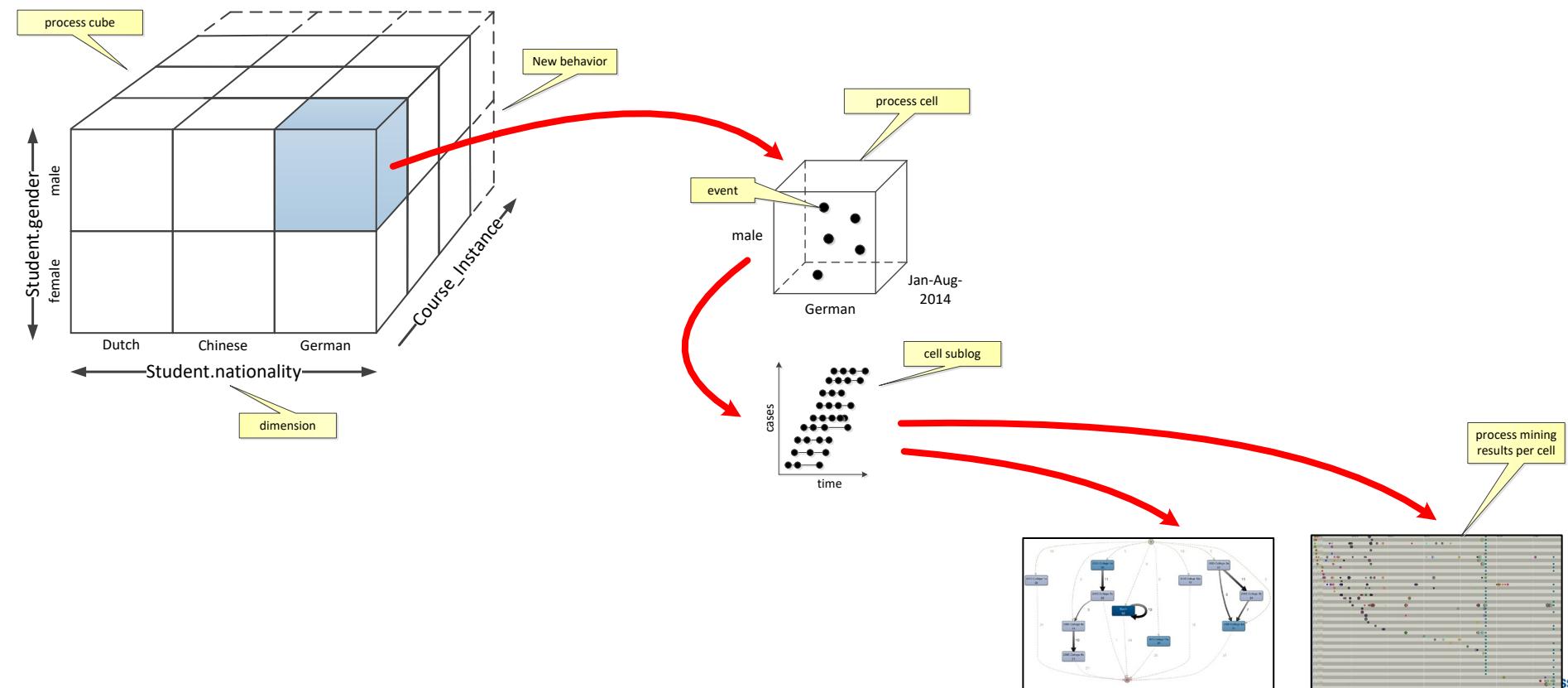
Final

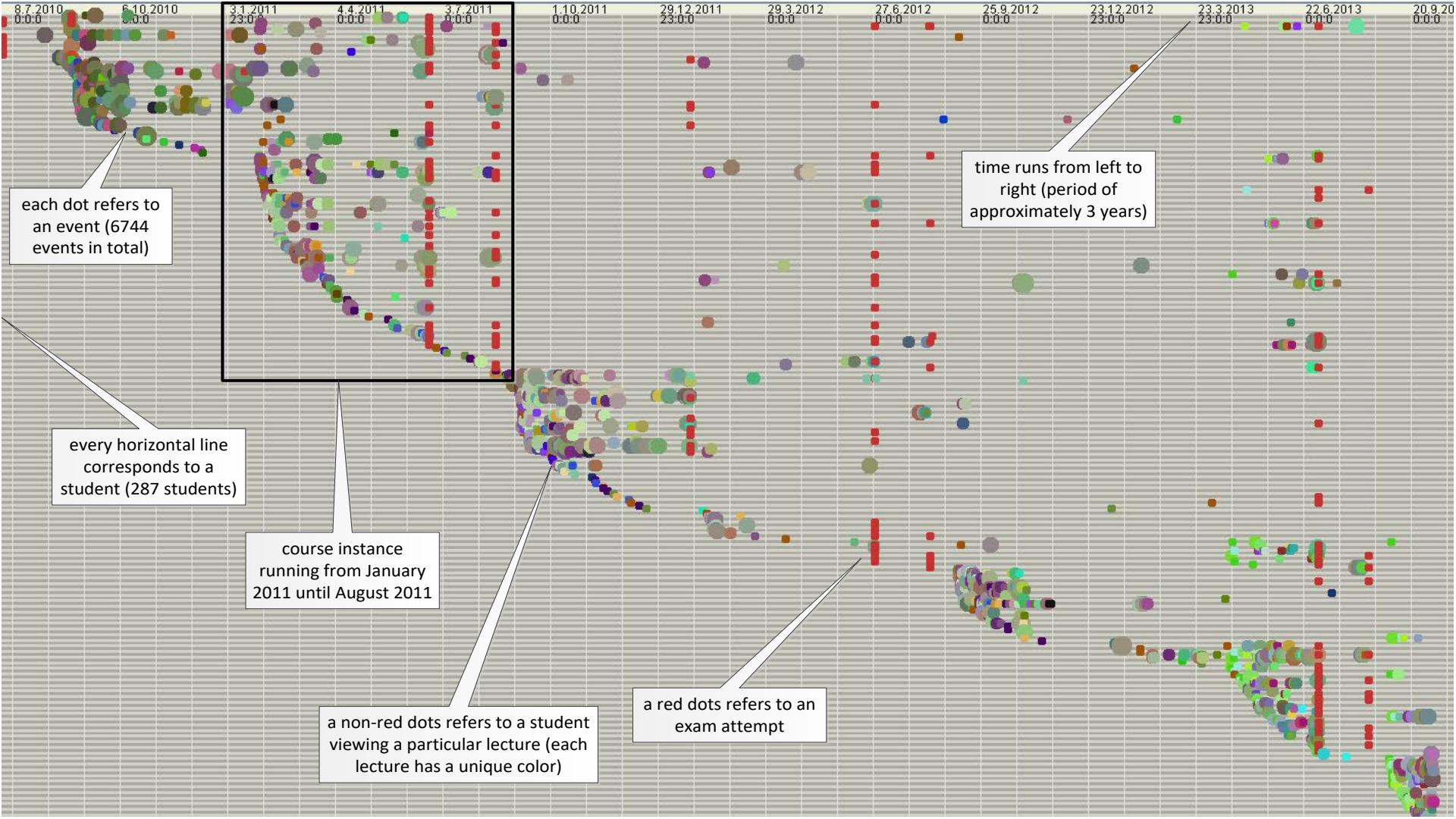
2022
2021
2020



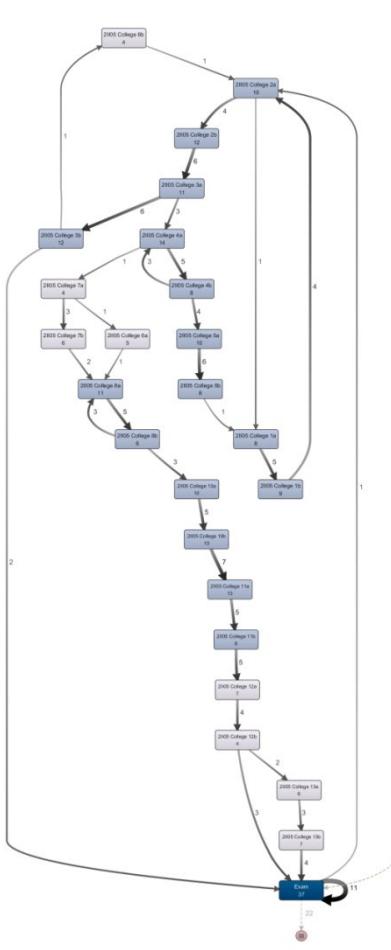
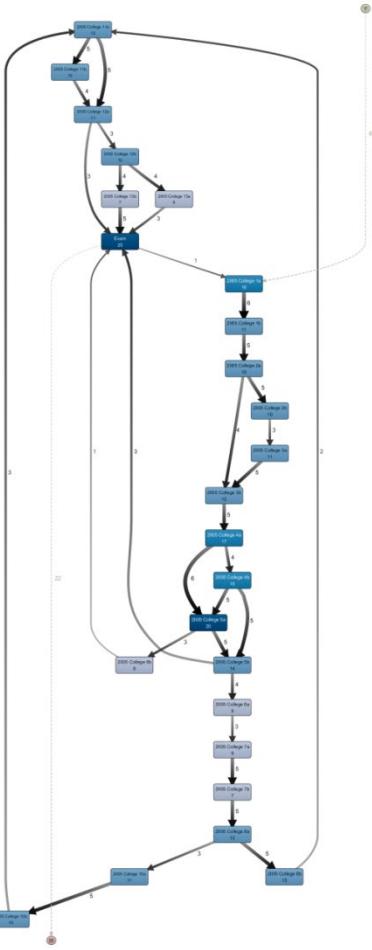
**Students
making
homework,
assignments
and exams.**

Process cube: Basic idea is simple





PASSED

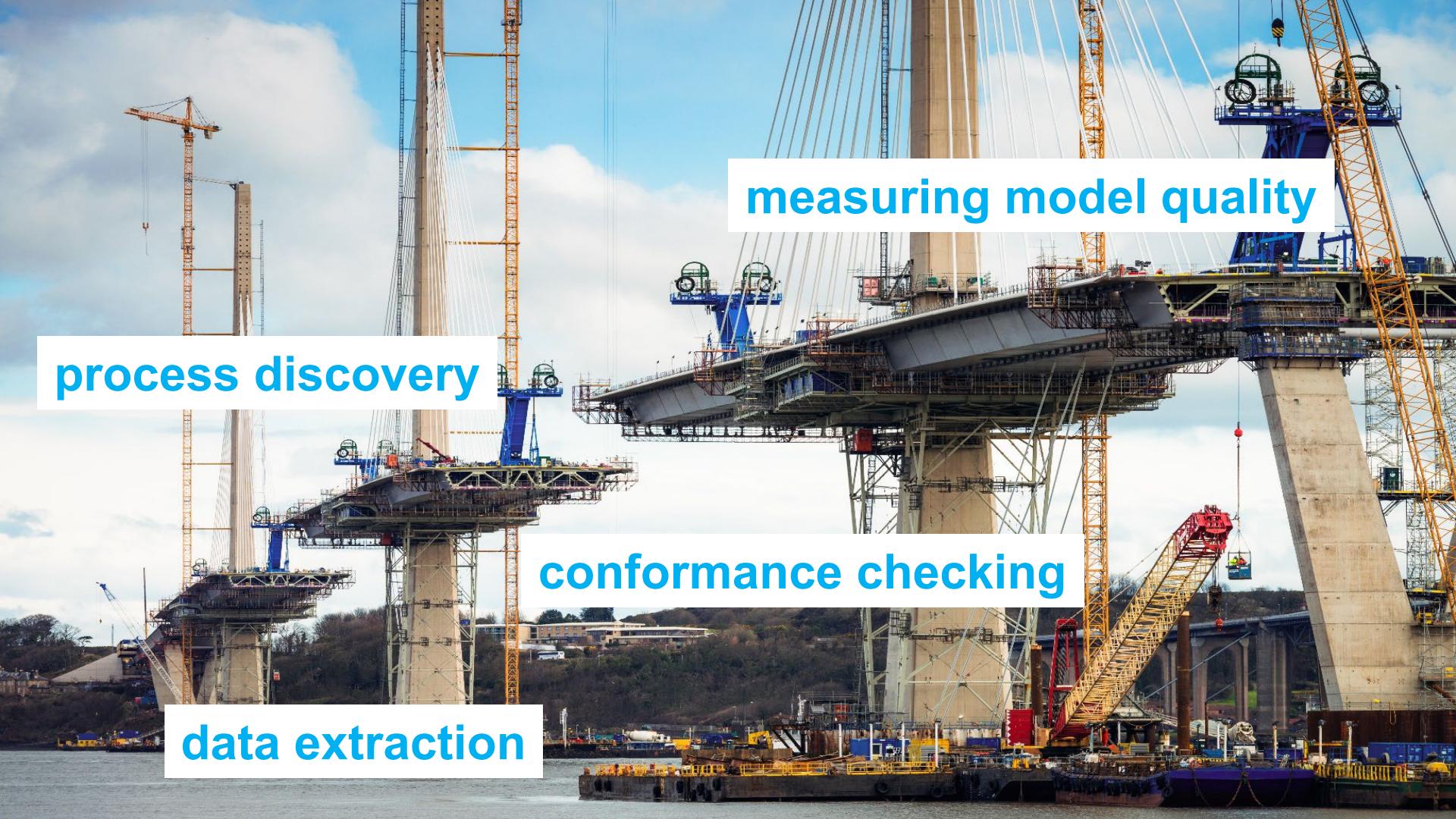


Research challenges





“Old” Challenges

A wide-angle photograph of a bridge under construction, likely a cable-stayed bridge, with several tall concrete piers and a complex network of white cables. Multiple orange construction cranes are positioned around the site. The bridge spans a body of water, with some industrial structures visible in the background.

process discovery

measuring model quality

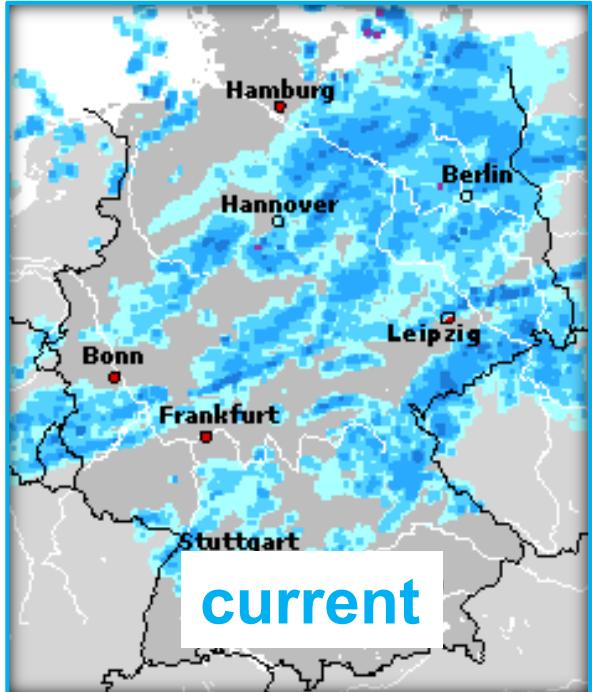
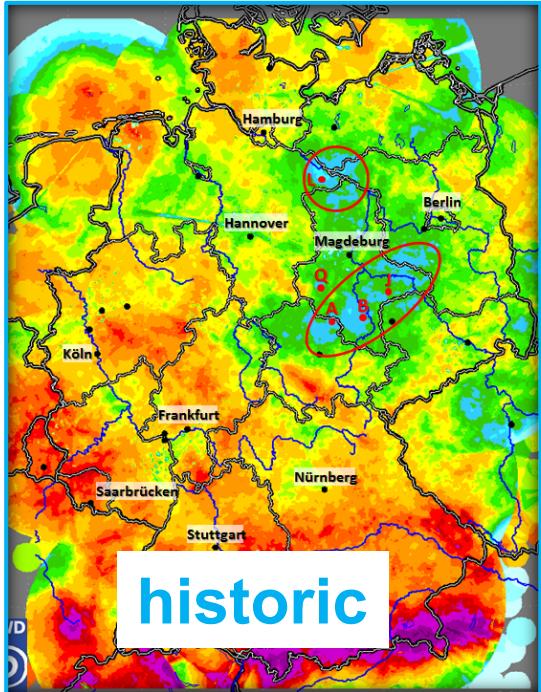
conformance checking

data extraction



**From backward looking
to forward looking**

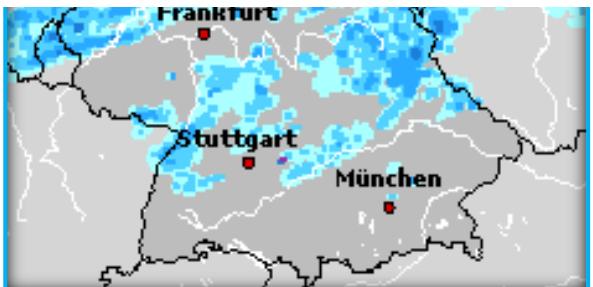
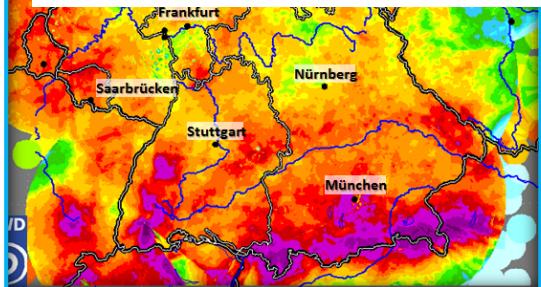
From backward to forward looking



From backward to forward looking

quality of predictions needs to improve

predictions need to be turned into actions



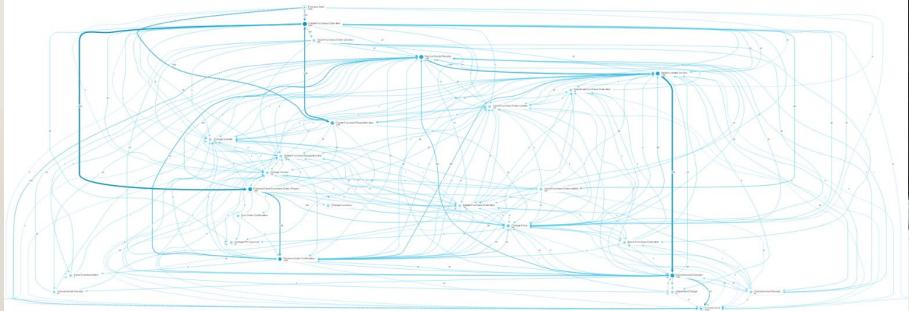
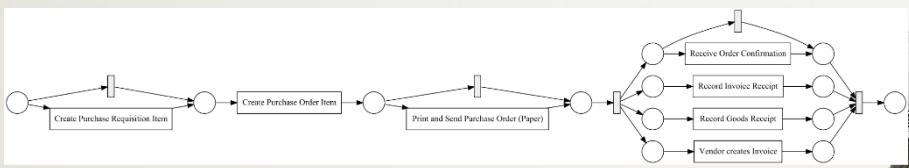
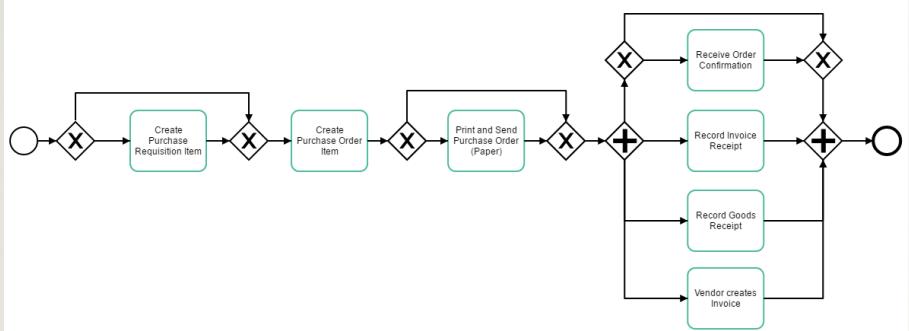


The background image is an aerial photograph of a complex multi-level highway interchange. The roads are filled with numerous cars, and the surrounding area includes residential buildings, commercial structures, and parking lots. A large blue arrow graphic is overlaid on the left side of the image, pointing upwards and to the right.

Context matters!



Better integration of
process discovery,
conformance checking,
and process modeling



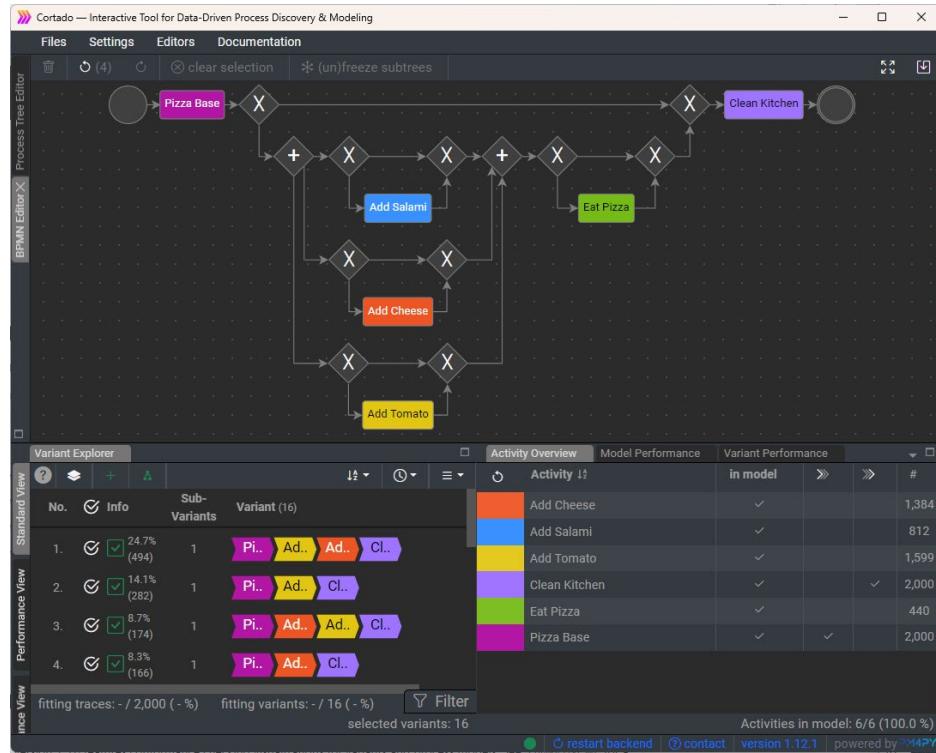
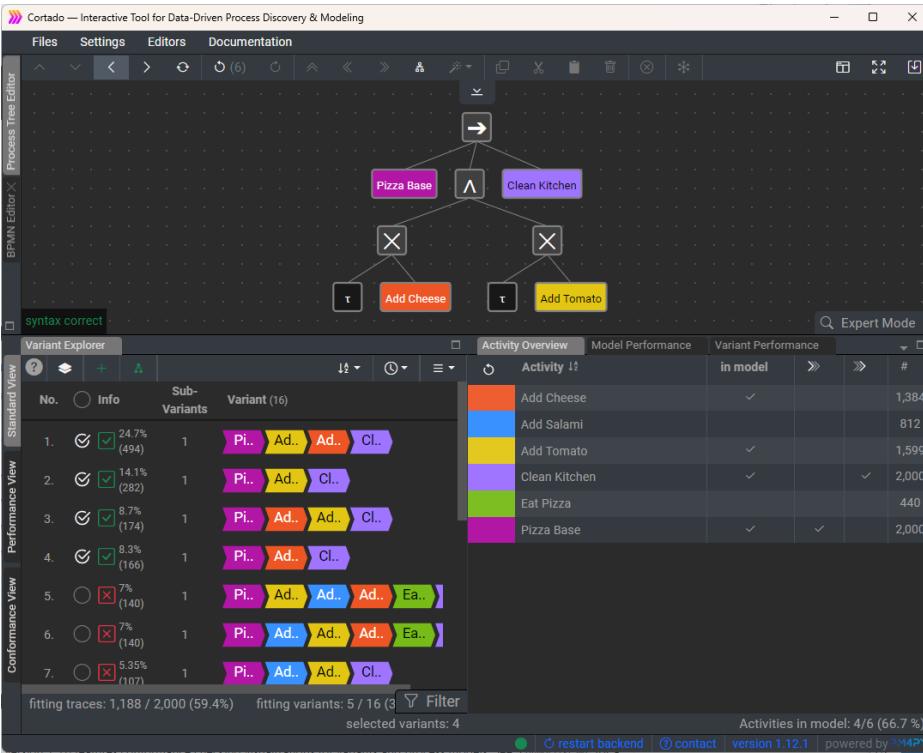
Hybrid process models are needed showing both the “sure/formal” and the “unsure/informal” parts.

Process modeling with “haptic feedback” based on data

Interactive process modeling still outperforms
fully automated discovery techniques.



Example: Cortado





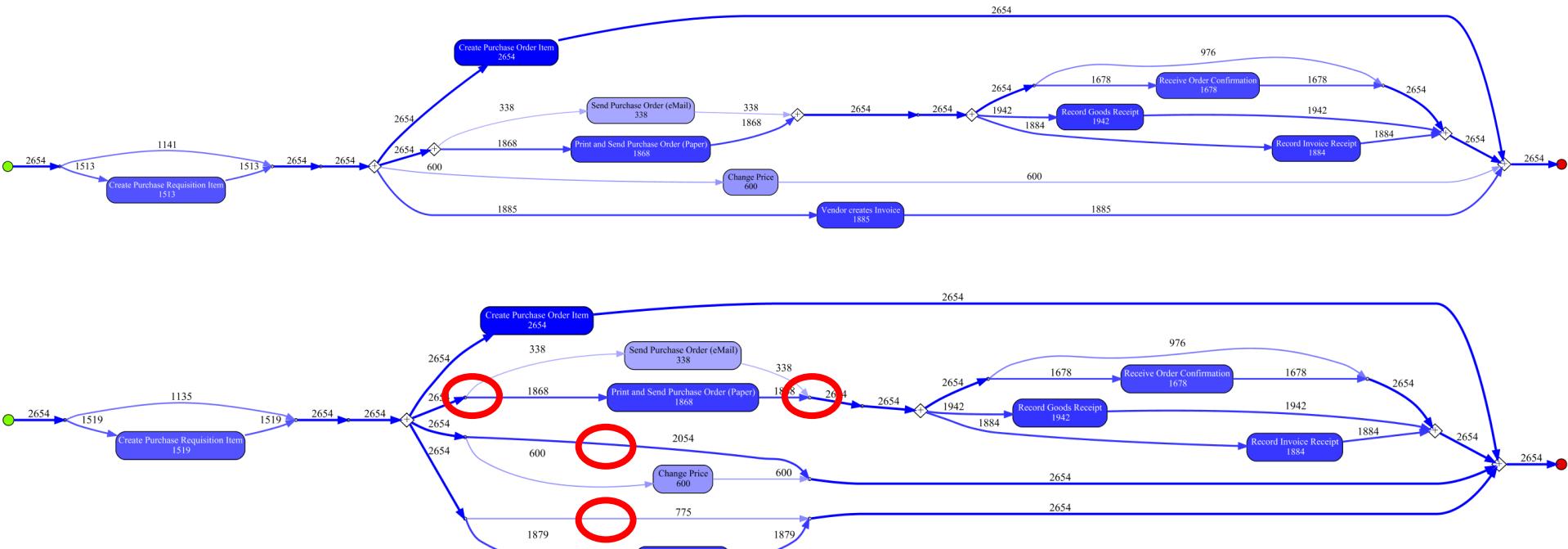
Better support for
comparative process
mining





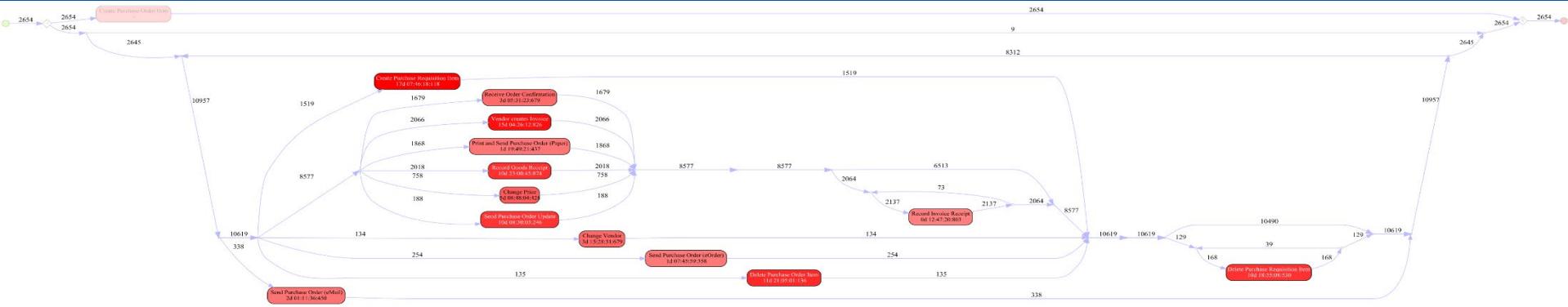
What are the differences?

(different periods, departments, customer groups)

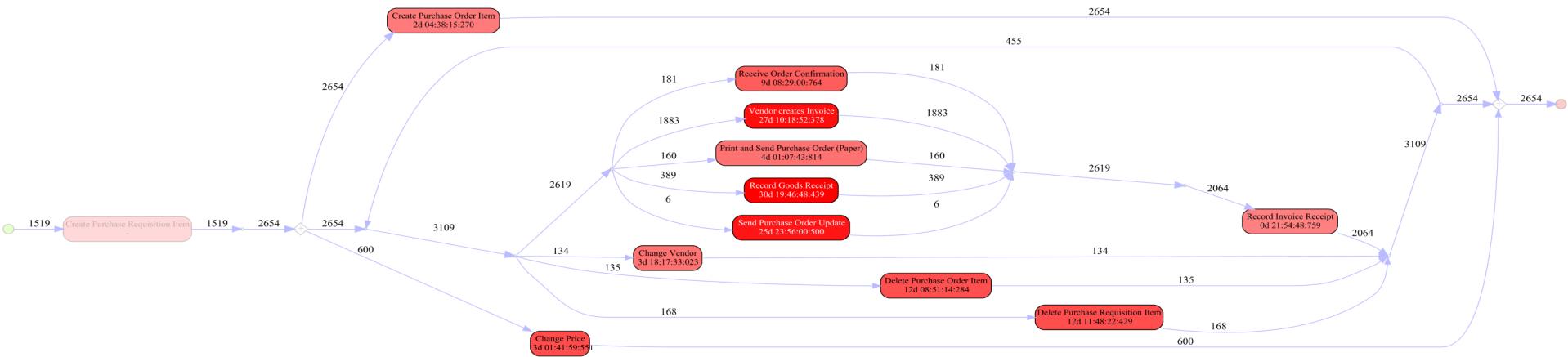


What are the differences?

(different periods, departments, customer groups)



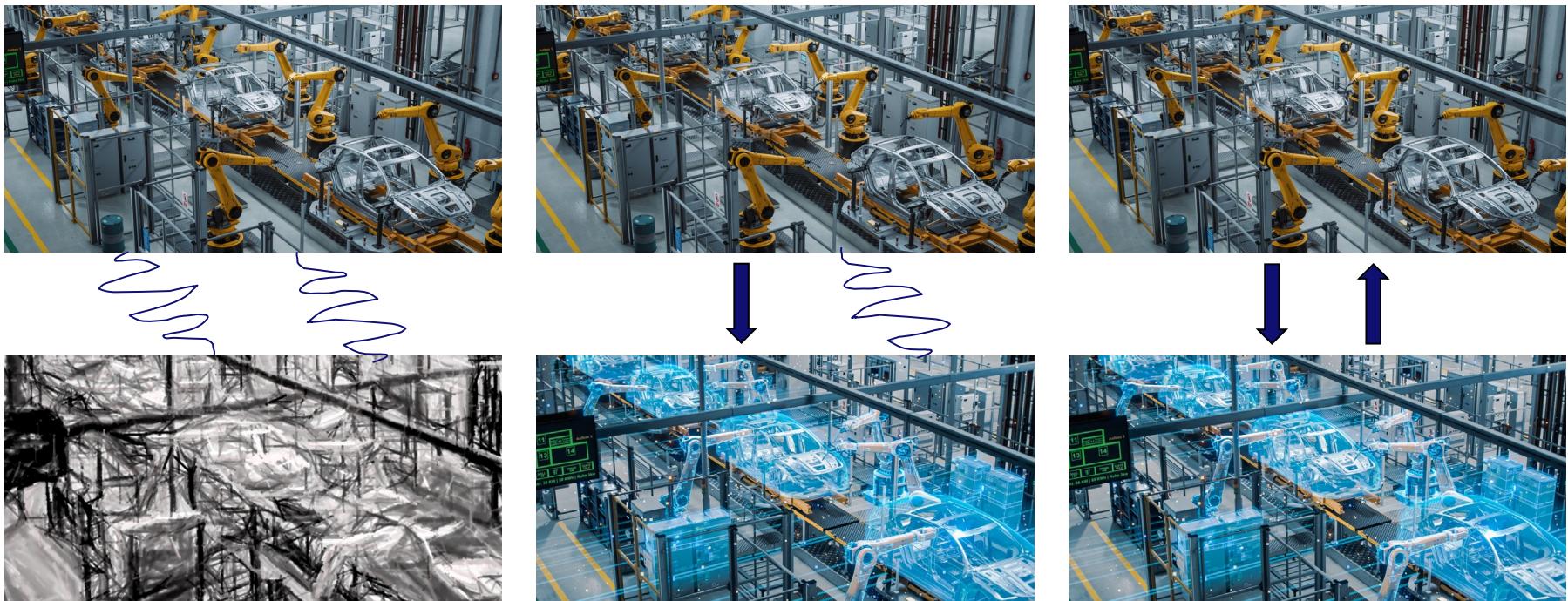
More difficult are differences in frequencies, times, etc.





Creating digital twins (e.g., simulation models)

Towards digital twins



Digital Model

Digital Shadow

Digital Twin



Chair of Process
and Data Science

The background of the slide features a blue-toned globe with a glowing binary code pattern (0s and 1s) covering the entire surface. A bright, glowing blue starburst or light effect is positioned at the top center of the globe, radiating outwards.

**Considering causality and
fairness to suggest more
meaningful improvements**

Process mining can be used to identify compliance and performance problems

If Wil works on a case, check activities are more likely to be skipped.



In this department, checks are performed after the legal deadline.

- mandatory activity is skipped
- activity is performed too late
- wrong order
- unauthorized resource

Process mining can be used to show the root causes of such problems, but...

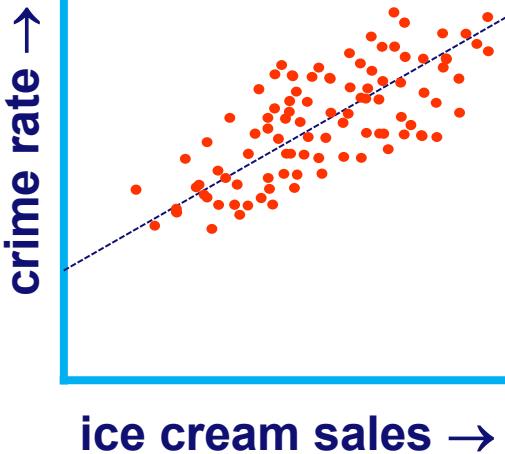
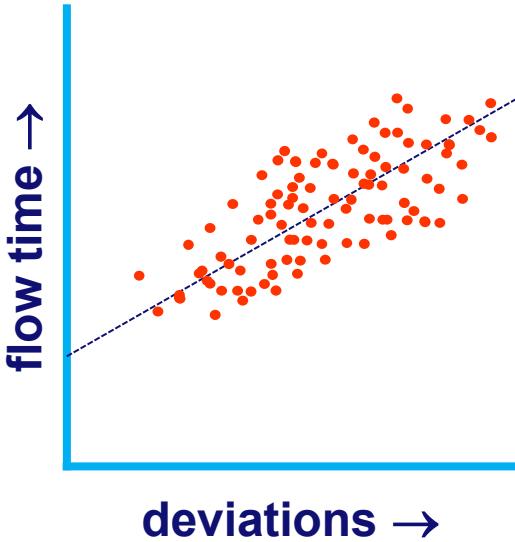
Cases for this supplier tend to have many price changes.



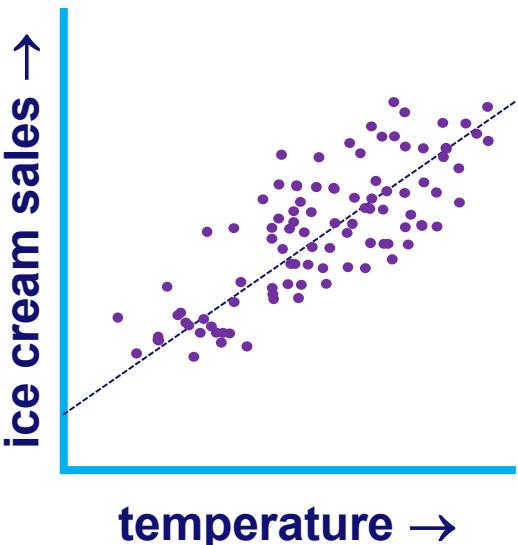
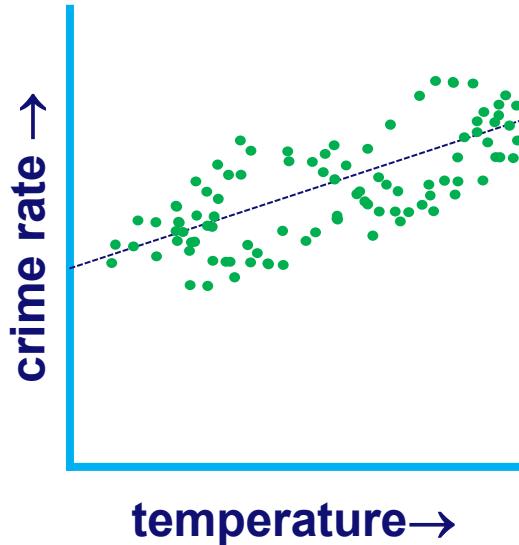
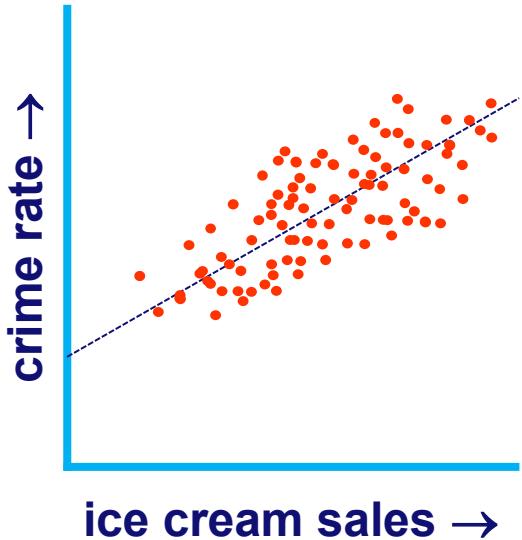
There are often delays in the back office on Fridays.

- bottlenecks and delays
- unnecessary rework
- waste
- overproduction

Root case analysis?



Correlation ≠ Causality

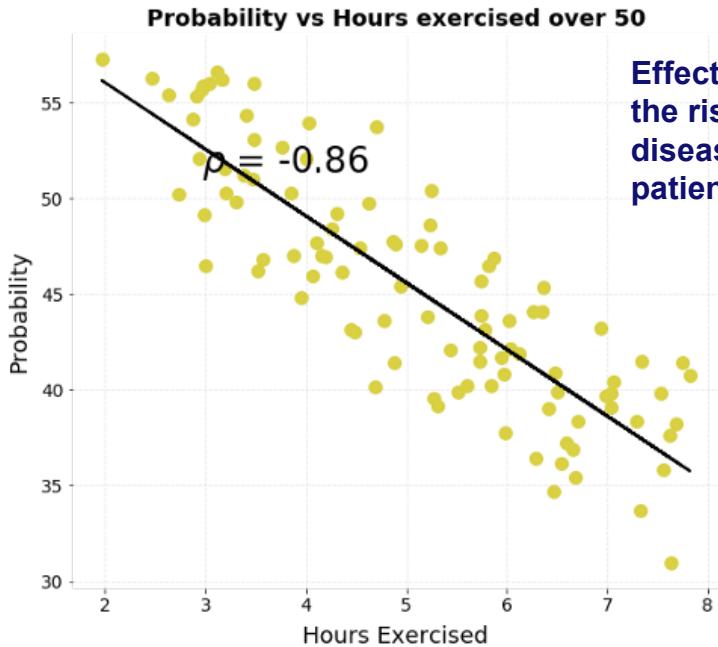
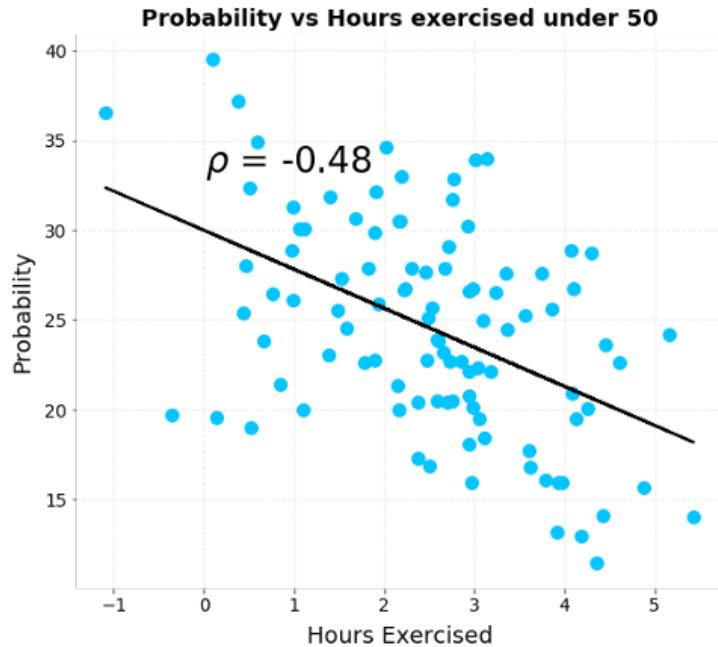


Simpson's paradox

RWTH	computer science		mathematics		all	
	get degree	drop out	get degree	drop out	get degree	drop out
female	80 (80%)	20 (20%)	400 (40%)	600 (60%)	480 (44%)	620 (56%)
male	700 (70%)	300 (30%)	30 (30%)	70 (70%)	730 (66%)	370 (34%)

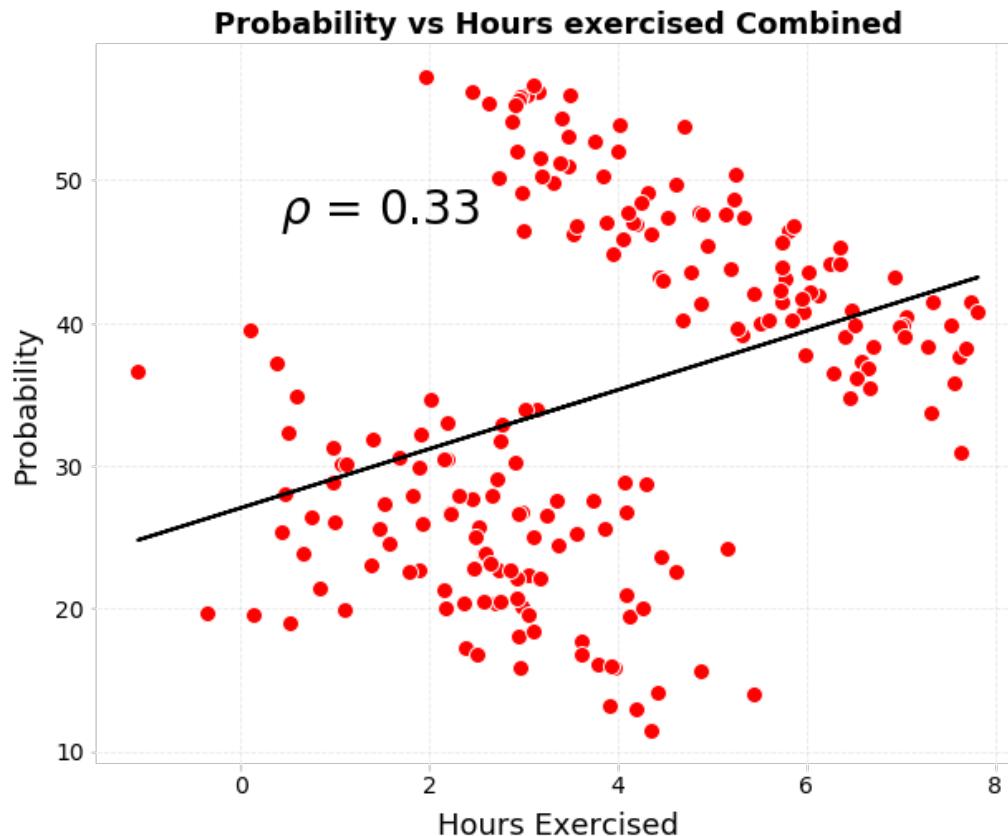
1100 females and 1100 males, 1100 CS students and 1100 math students

Simpson's paradox (it is good to exercise)



Effect of exercising on the risk of developing a disease for two sets of patients (below/over 50).

Simpson's paradox (exercising will kill you)

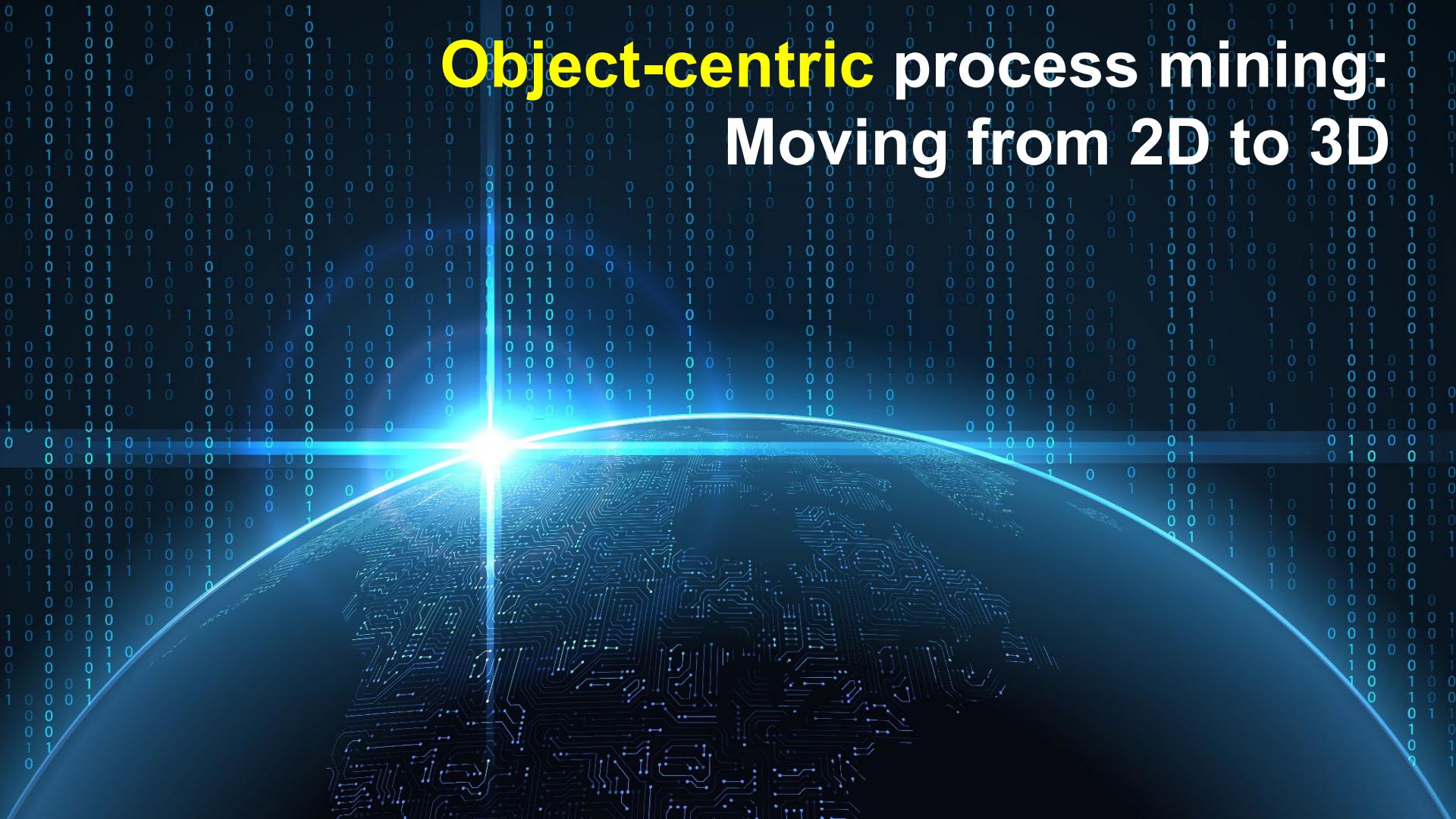


Simpson's Paradox: How to Prove Opposite Arguments with the Same Data by William Koehrsen towardsdatascience.com
© Wil van der Aalst (use only with permission & acknowledgements)

Fairness

- **Don't blame overloaded resources for causing bottlenecks.**
- **Don't blame the most experienced resources taking the most difficult cases for deviating.**
- **Discrimination-aware data/process mining aims to avoid such errors.**
- **Trade-off between fairness and accuracy.**

Object-centric process mining: Moving from 2D to 3D





Max Verstappen, pit stop 1.86 seconds, Russian GP 2020

ASTON MARTIN

Objects: 1 driver, 1 car, 4+4 tires, 20 pit crew members, 1 race, etc.



Remember: a classical event = case + activity + timestamp + ...



Traditional process mining is like following one object, e.g., one tire.

Convergence problem:

- Assume we have a high-level event “pitstop” involving 30+ objects.
- Taking “tire” as a case perspective, each pitstop occurs 8 times.



Divergence problem:

- Assume we consider low-level events like “stop”, 4x “remove tire”, 4x “mount tire”, “drive”, etc.
- Taking “car” as a case perspective, we can no longer see causalities or track individual tires.

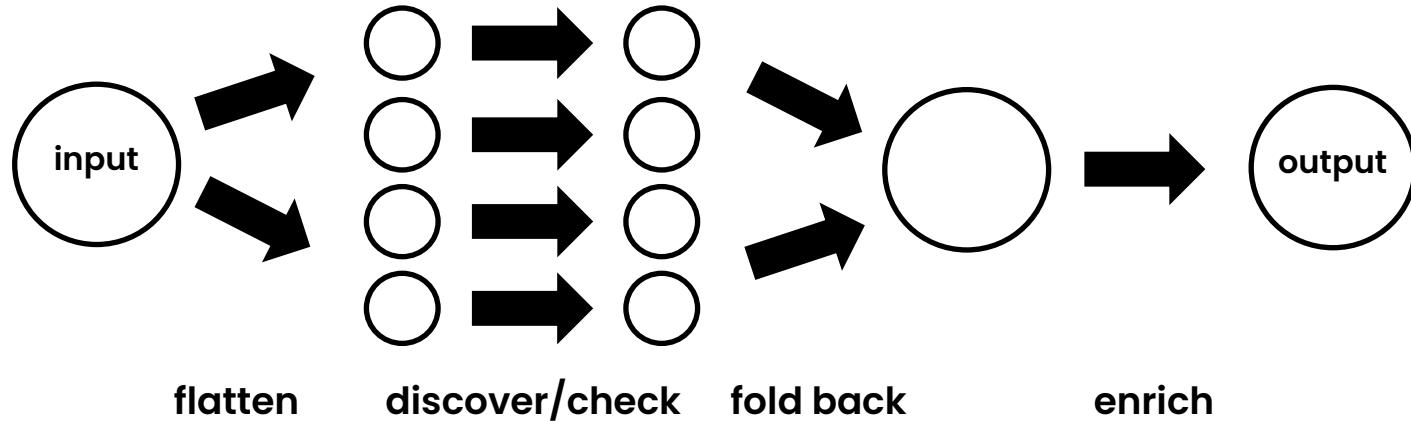


Solution: Event = objects + activity + timestamp + ...



Baseline Approach OCPM

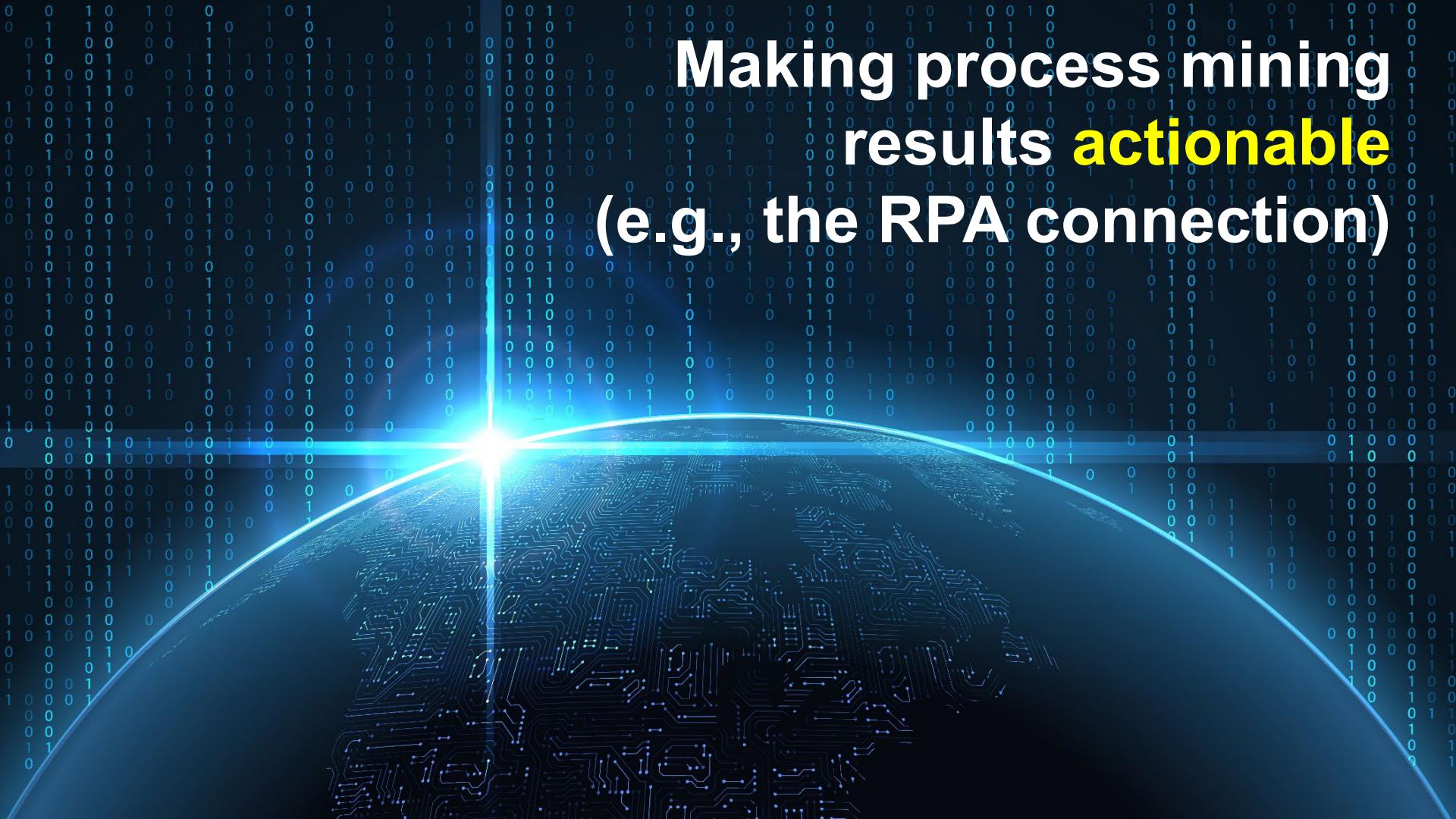
(Process Discovery and Conformance checking)



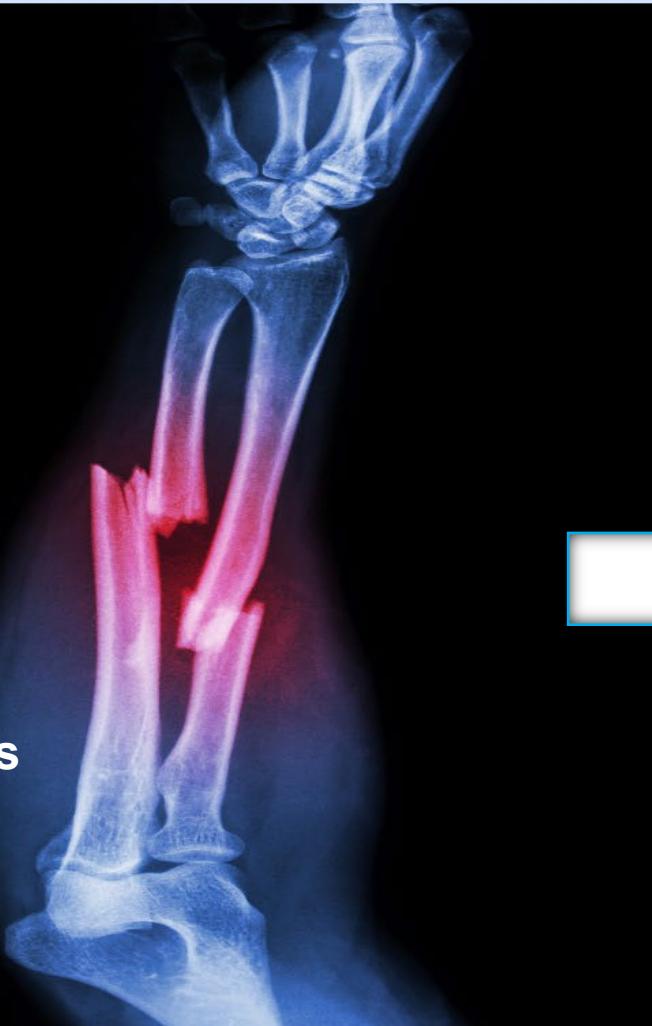
- Leverage existing case-centric process mining techniques: flatten, discover/check, fold back.
- To fold back, activities need to be unique, and objects need to agree on frequency.
- Approach applies to process discovery and conformance checking.
- Merged output can be enriched, e.g., annotations or checks related to cardinalities can be added.

Just the starting point: Many research challenges



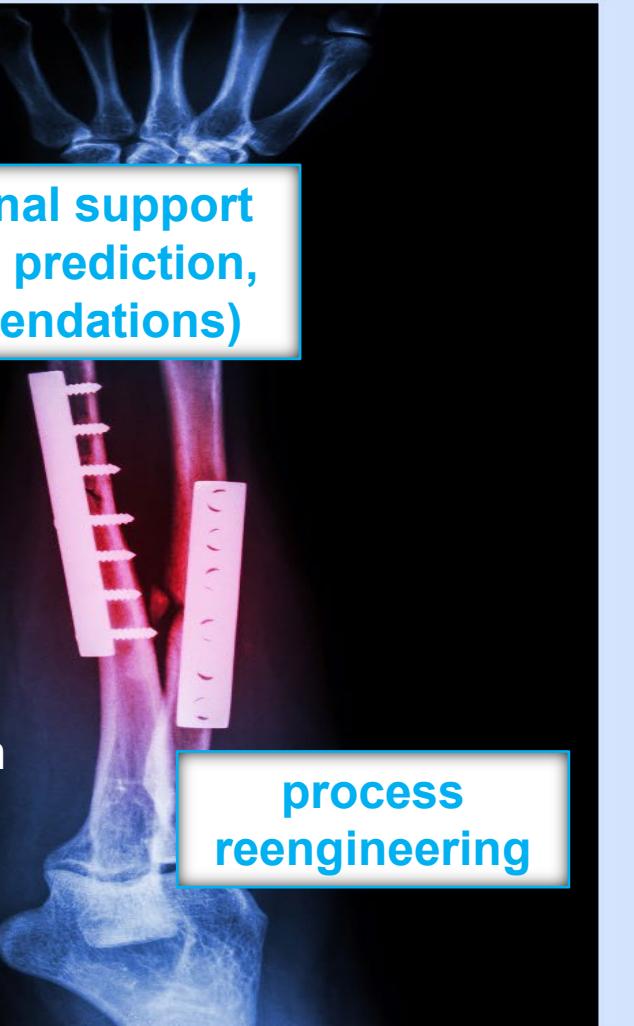
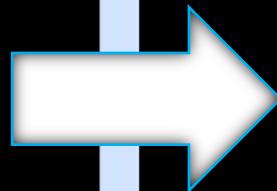


Making process mining results **actionable** (e.g., the RPA connection)



diagnosis

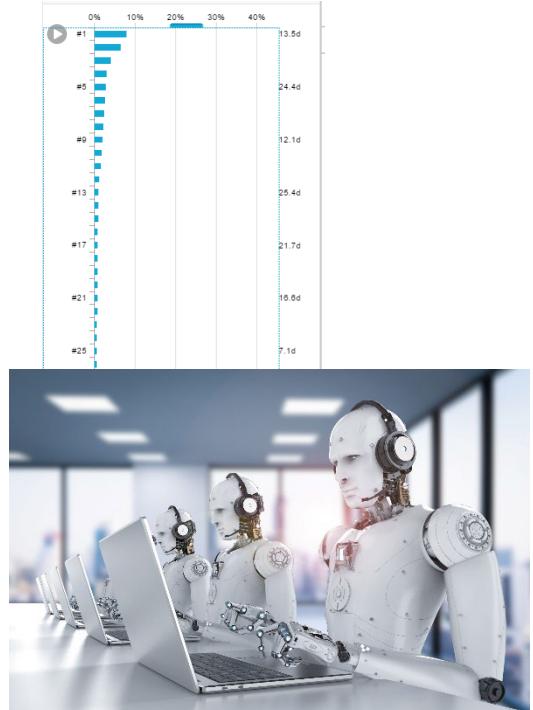
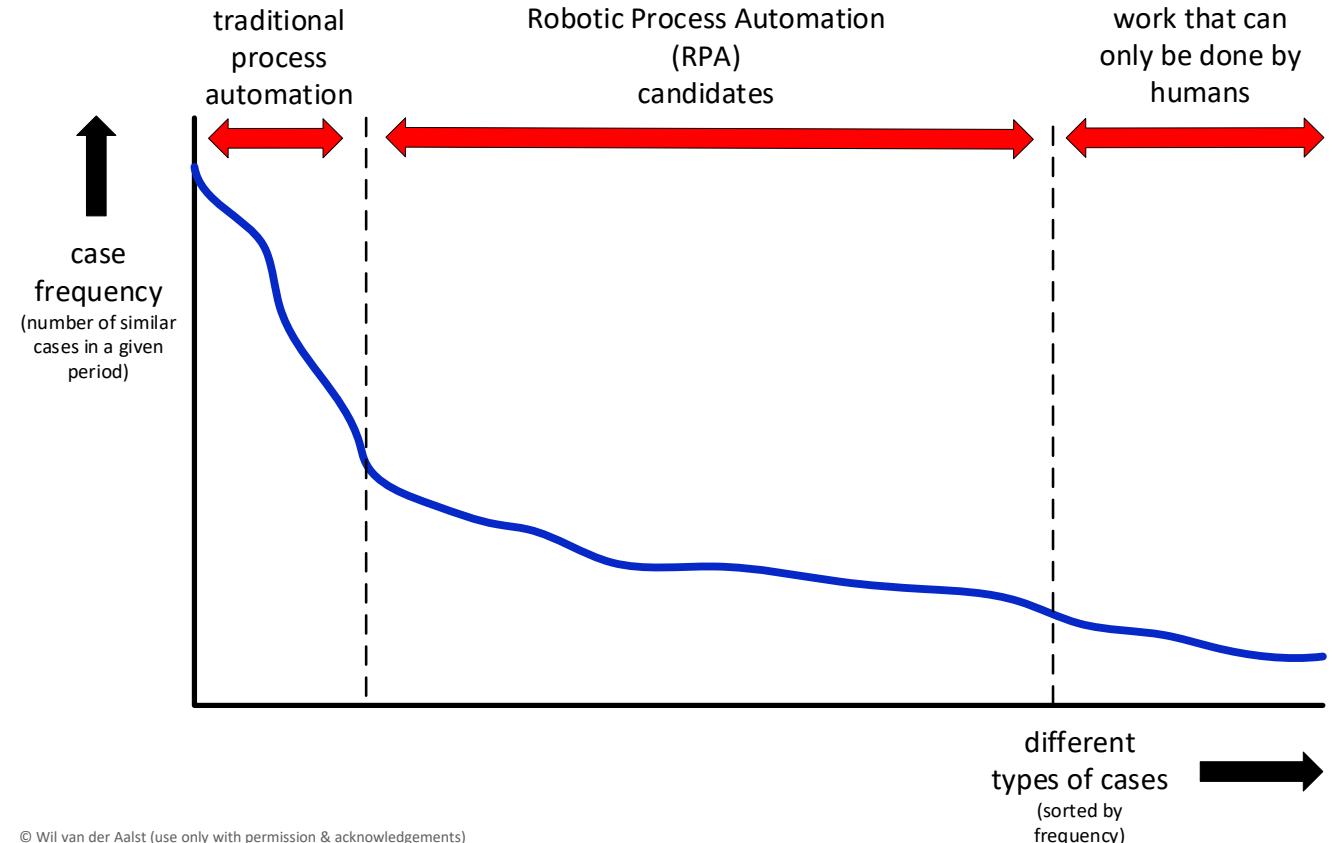
operational support
(alerting, prediction,
recommendations)



action

process
reengineering

Process-mining-informed automation (RPA)



Foundational models for operational processes



Plenty of “co-pilots”

Copilot OM | Order Management

co-pilot-assets.try.celonis.cloud/package-manager/ui/views/ui/spaces/325adbd1-f102-4c98-80d8-4a486cc09...

Paused

Apps > 0 - Copilot OM > Order Management

Inbox

Enhance Accuracy (1 improvements)

Copilot OM

Inspect

What is my on-time delivery rate?

Please select the correct option for 'On-Time Delivery Rate'

On-Time Delivery Rate (OTD) Late Delivery Rate # Late Deliveries

Early Delivery Rate Late Delivery Value [EUR] None of the above

Inspect this response

On-Time Delivery Rate (OTD)

The current On-Time Delivery Rate (OTD) for your sales order items.

On-Time Delivery Rate (OTD) 63.7%

85%

Number of on-time sales order items divided by the total considered number of sales order items. What is considered can be changed in the Validation Cockpit and is described in the about section.

Inspect this response

Analyze m... Find opport... Explore th... Build a vis...

If this is your first session, please choose a sample question from the prompts above to familiarize yourself with the Copilot.

Mistakes are possible with Process Copilot, please be sure to inspect responses.

Copilot OM | Order Management

co-pilot-assets.try.celonis.cloud/package-manager/ui/views/ui/spaces/325adbd1-f102-4c98-80d8-4a486cc09...

Paused

Apps > 0 - Copilot OM > Order Management

Inbox

Enhance Accuracy (1 improvements)

Copilot OM

Inspect

100% of cases 988.1K of 988.1K

Show the process model

Here is the process model for your business process.

View Expanded Process Explorer

Inspect this response

How was this calculated?

1. Data Collection

Searched for the relevant fields in the data using keywords such as "no touch" and "rate."

2. Data Processing

Found a pre-computed KPI in the data named "No Touch rate" with id "KPL_NO_TOUCH_RATE" which directly gives the no touch rate.

3. Identification of No Touch Rate

Retrieved the value of the KPI from the data.

4. No Touch Rate Calculation

The value retrieved was 0.933181, which was then converted to a percentage to get 93.31%. The no touch rate was directly available in the data and no further calculations were needed.

Filters

+ Add filter

View Code

Code

```
1  settings:
2   eventLogs:
3    - eventLog:
4      O2C_EVENTLOG
5        id:
6          9dabd7e5-6956-435d-b316-283
7          order: "100"
8          graphControls: sliders
9          expandable: false
```

View Expanded Process Explorer

Inspect this response

Show me what my process looks like | List all activities of my process

Analyze m... Find opport... Explore th... Build a vis...

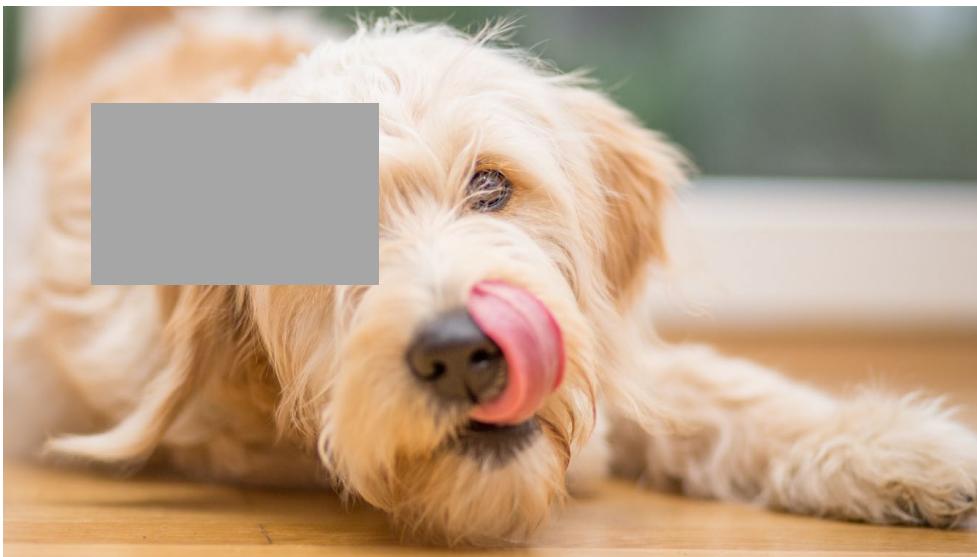
If this is your first session, please choose a sample question from the prompts above to familiarize yourself with the Copilot.

Mistakes are possible with Process Copilot, please be sure to inspect responses.

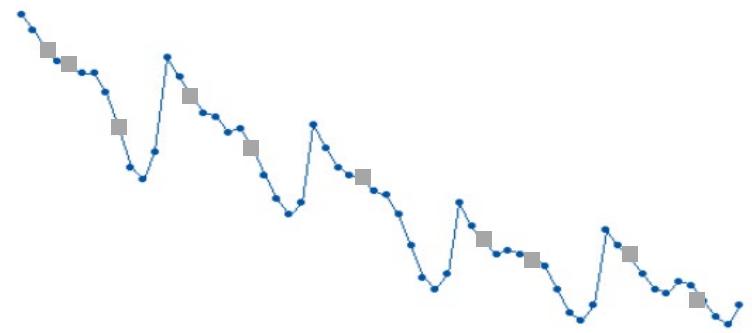
But ... ChatGPT & Co know nothing about your processes (unless you tell)

- The **prompt is crucial** and may contain (next to the question): actual data (e.g., events and objects), aggregated data (e.g., DFG, variants, times), or just metadata (e.g., event types and object types in combination with RAG).
- Generative AI **works best** if at least parts of the input or output have a **clear structure and semantics**.
- When you can generate SQL you can also generate PQL (Process Query Language), but it may have all of the known problems.
- The Celonis copilot exploits the structure and semantics of the process intelligence graph (objects, events, KPIs, etc.) and PQL.

Self-supervised learning



All roads lead to [REDACTED]. This is where you can [REDACTED] the best pizzas and [REDACTED] the best [REDACTED].



- Why not for event data?
- Challenges: limited data (but structured), words like “Rome” and images of a dog have an out-of-sample/universal meaning, a value or activity name not, things that are computationally hard (e.g., computing alignments) will not benefit from genAI.

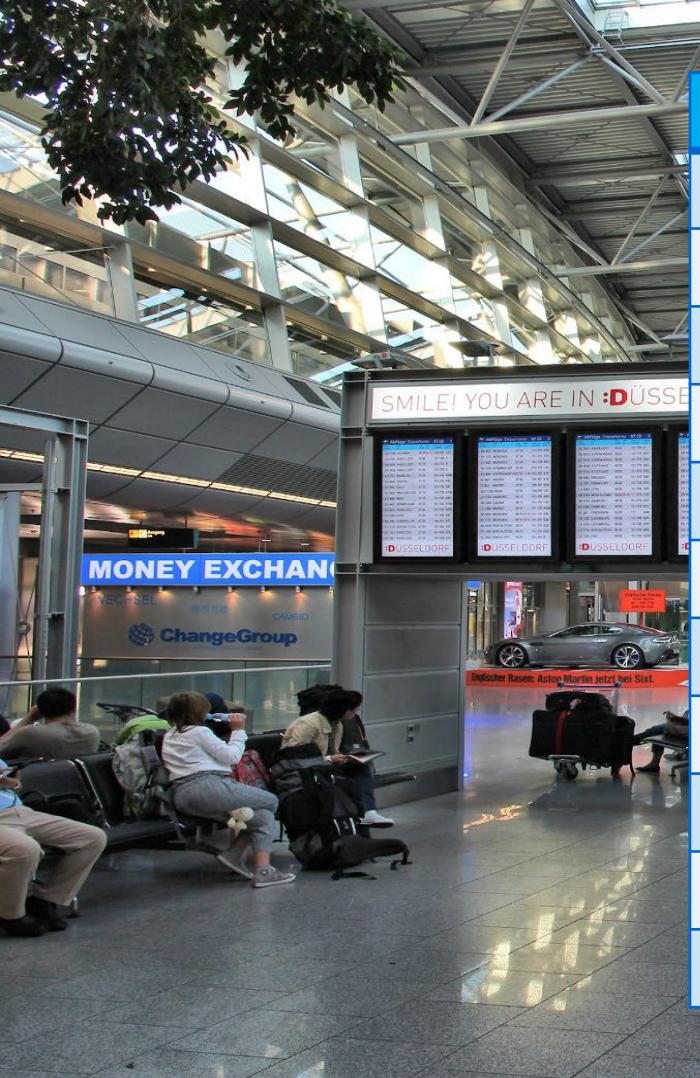
Ensuring confidentiality in process mining



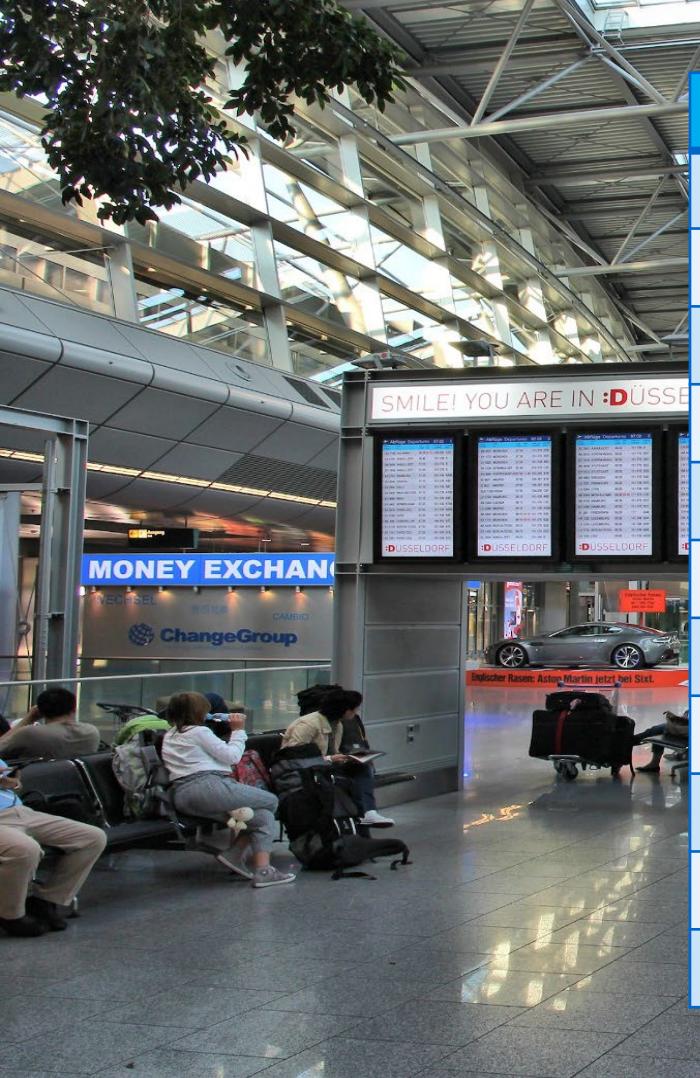


**Event data are
very sensitive!**

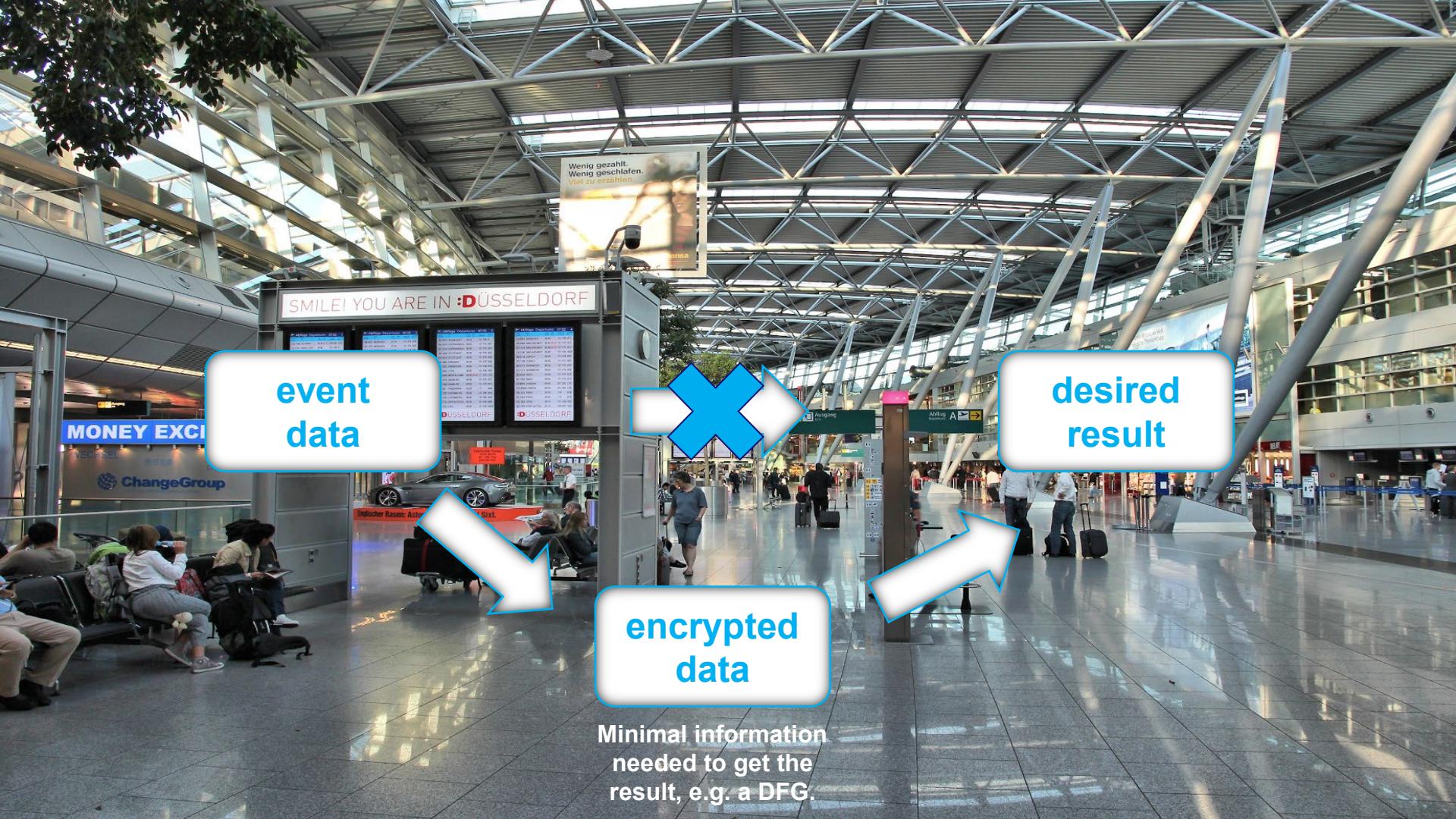
Name	Activity	Date	Time	From	To
WvdA	check-in	30-8-19	8:20	CGN	VIE
WvdA	drop-bag	1-9-19	9:20	CGN	VIE
WvdA	board	1-9-19	9:40	CGN	VIE
...
WvdA	check-in	15-9-19	22:58	DUS	OSL
WvdA	drop-bag	17-9-19	18:38	DUS	OSL
WvdA	board	17-9-19	19:10	DUS	OSL
...
WvdA	check-in	23-9-19	22:46	DUS	NRT
WvdA	drop-bag	24-9-19	16:18	DUS	NRT
WvdA	board	24-9-19	17:12	DUS	NRT



Name	Activity	Date	Time	From	To
#6@7	check-in	30-8-19	8:20	CGN	VIE
#6@7	drop-bag	1-9-19	9:20	CGN	VIE
#6@7	board	1-9-19	9:40	CGN	VIE
...
#6@7	check-in	15-9-19	22:58	DUS	OSL
#6@7	drop-bag	17-9-19	18:38	DUS	OSL
#6@7	board	17-9-19	19:10	DUS	OSL
...
#6@7	check-in	23-9-19	22:46	DUS	NRT
#6@7	drop-bag	24-9-19	16:18	DUS	NRT
#6@7	board	24-9-19	17:12	DUS	NRT



Name	Activity	Date	Time	From	To
#6@7	check-in	30-8-19			VIE
#6@7	drop-bag	1-9-19			VIE
#6@7	board	1-9-19			VIE
...
#6@7	check-in	15-9-19			OSL
#6@7	drop-bag	17-9-19			OSL
#6@7	board	17-9-19			OSL
...
#6@7	check-in	23-9-19			NRT
#6@7	drop-bag	24-9-19			NRT
#6@7	board	24-9-19			NRT



event
data

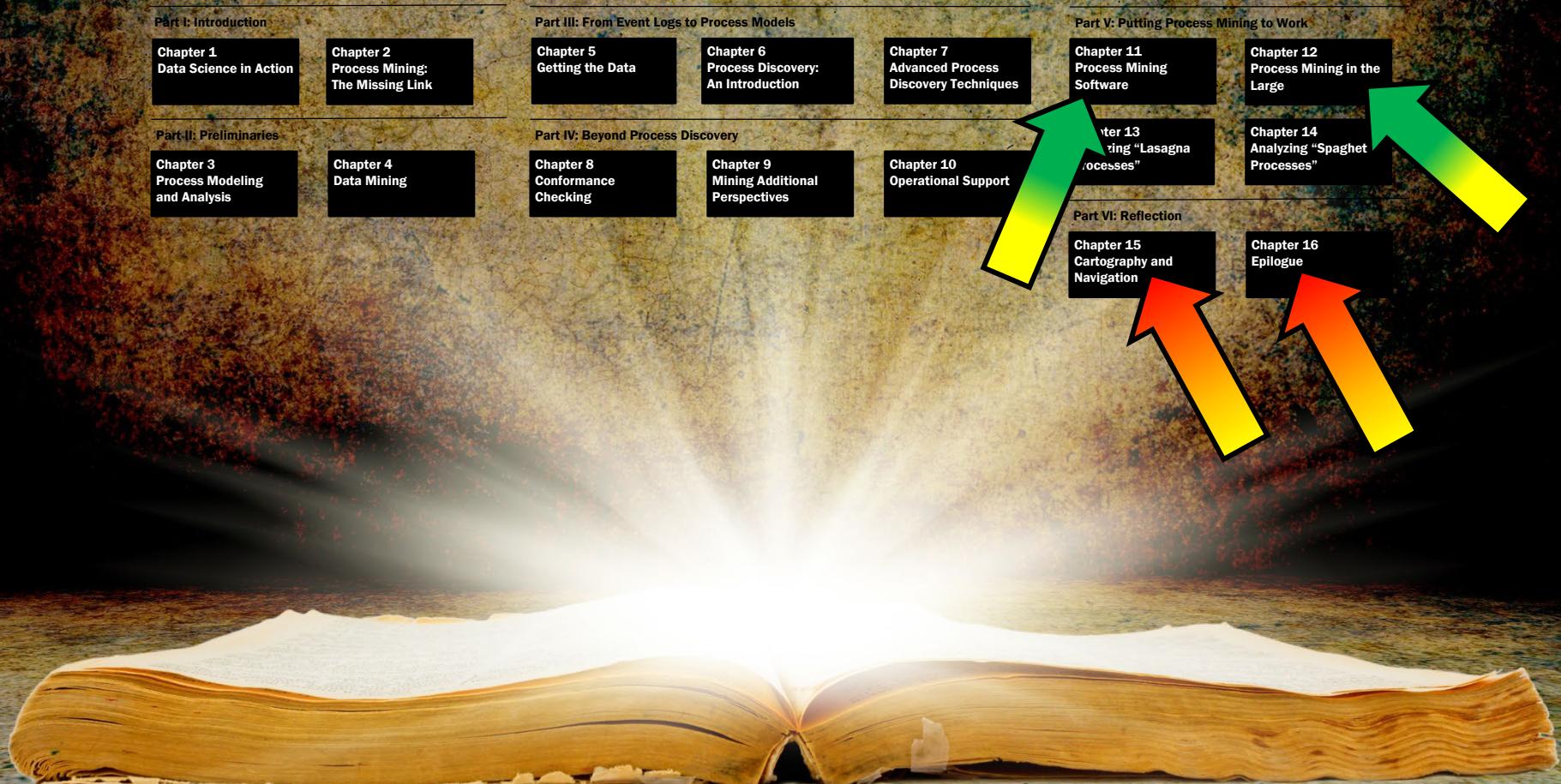
encrypted
data

desired
result

Minimal information
needed to get the
result, e.g. a DFG.

Conclusion

- Data sets can be **huge** and **hard to analyze**.
- Algorithms have **very different performance characteristics**.
- Process mining provides many **challenges**.
 - These have a huge practical relevance.
 - Interesting from a scientific point of view.
 - See PADS Bachelor and Master projects.



ID	Topic	Date	Date	Place
	Lecture 1 Introduction to Process Mining	08.04.24	Monday	AH V
	Lecture 2 Data Science: Supervised Learning	09.04.24	Tuesday	AH V
	<i>Exercise 1 Tool Introduction</i>	09.04.24	Tuesday	AH III
	Lecture 3 Data Science: Unsupervised Learning and Evaluation	15.04.24	Monday	AH V
	Lecture 4 Introduction to Process Discovery	16.04.24	Tuesday	AH V
	<i>Exercise 2 Data Mining</i>	16.04.24	Tuesday	AH III
	Lecture 5 Alpha Algorithm 1	22.04.24	Monday	AH V
	Lecture 6 Alpha Algorithm 2	23.04.24	Tuesday	AH V
	<i>Exercise 3 Petri Nets</i>	23.04.24	Tuesday	AH III
	Lecture 7 Model Quality Representation	29.04.24	Monday	AH V
	Lecture 8 Heuristic Mining	30.04.24	Tuesday	AH V
	<i>Exercise 4 Alpha Miner</i>	30.04.24	Tuesday	AH III
	Lecture 9 Region-Based Mining	06.05.24	Monday	AH V
	<i>Exercise 5 Heuristic Mining and Region-Based Mining</i>	07.05.24	Tuesday	AH III
	Lecture 10 Inductive Mining	13.05.24	Monday	AH V
	Lecture 11 Event Data and Exploration	14.05.24	Tuesday	AH V
	<i>Exercise 6 Inductive Mining</i>	14.05.24	Tuesday	AH III
	Lecture 12 Conformance Checking 1	27.05.24	Monday	AH V
	Lecture 13 Conformance Checking 2	28.05.24	Tuesday	AH V
	<i>Q&A Session Assignment Part I</i>	28.05.24	Tuesday	AH III
	Deadline Assignment Part I	02.06.24	Sunday	
	<i>Exercise 7 Footprint and Token-Based Replay (Exercise)</i>	03.06.24	Monday	AH V
	<i>Exercise 8 Alignments (Exercise)</i>	04.06.24	Tuesday	AH V
	Lecture 14 Decision Mining	10.06.24	Monday	AH V
	<i>Lecture 15 Celonis Guest Lecture</i>	11.06.24	Tuesday	AH V
	<i>Exercise 9 Decision Mining</i>	11.06.24	Tuesday	AH III
	Lecture 16 Performance Analysis and Organizational Mining	17.06.24	Monday	AH V
	<i>Exercise 10 Performance Analysis (Exercise)</i>	18.06.24	Tuesday	AH V
	<i>Exercise 11 Organizational Mining</i>	18.06.24	Tuesday	AH III
	<i>Exercise 12 Celonis Case Study</i>	24.06.24	Monday	AH V
	Lecture 17 Operational Support and Process Mining Applications	01.07.24	Monday	AH V
	Lecture 18 Distributed, Streaming, and Comparative Process Mining	02.07.24	Tuesday	AH V
	<i>Exercise 13 Operational Process Mining</i>	02.07.24	Tuesday	AH III
	Lecture 19 Closing	08.07.24	Monday	AH V
	<i>Q&A Session Assignment Part II</i>	09.07.24	Tuesday	AH III
	Deadline Assignment Part II	14.07.24	Sunday	
	<i>Q&A Session Exam</i>	16.07.24	Tuesday	AH III

