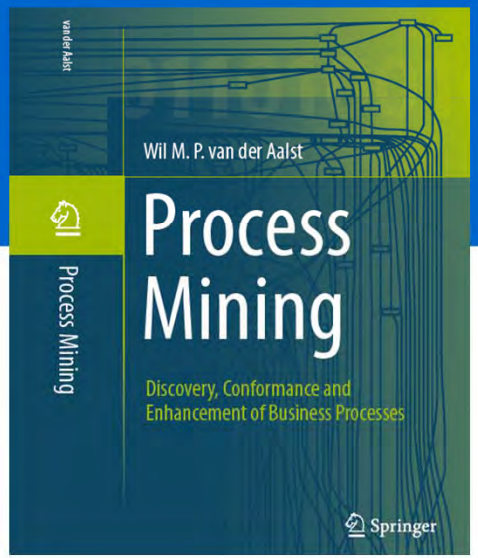
Distributed Process Discovery and Conformance Checking

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

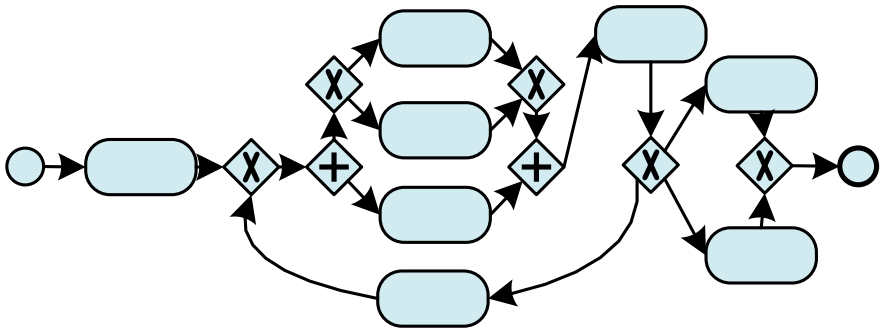


TU/e Technische Universiteit **Eindhoven** University of Technology

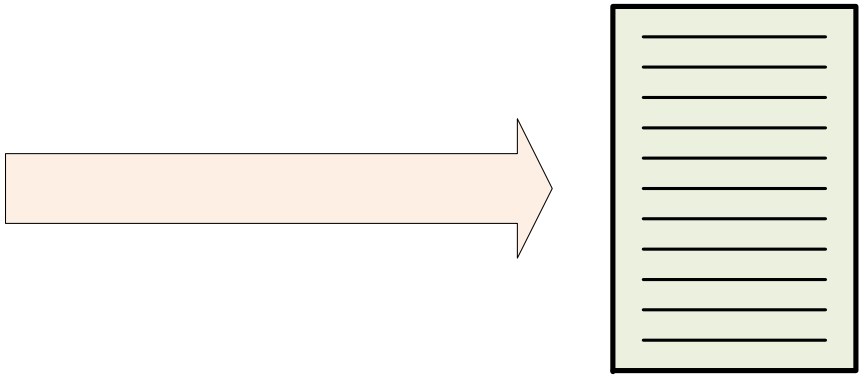
Where innovation starts

On the different roles of (process) models …

Play-Out

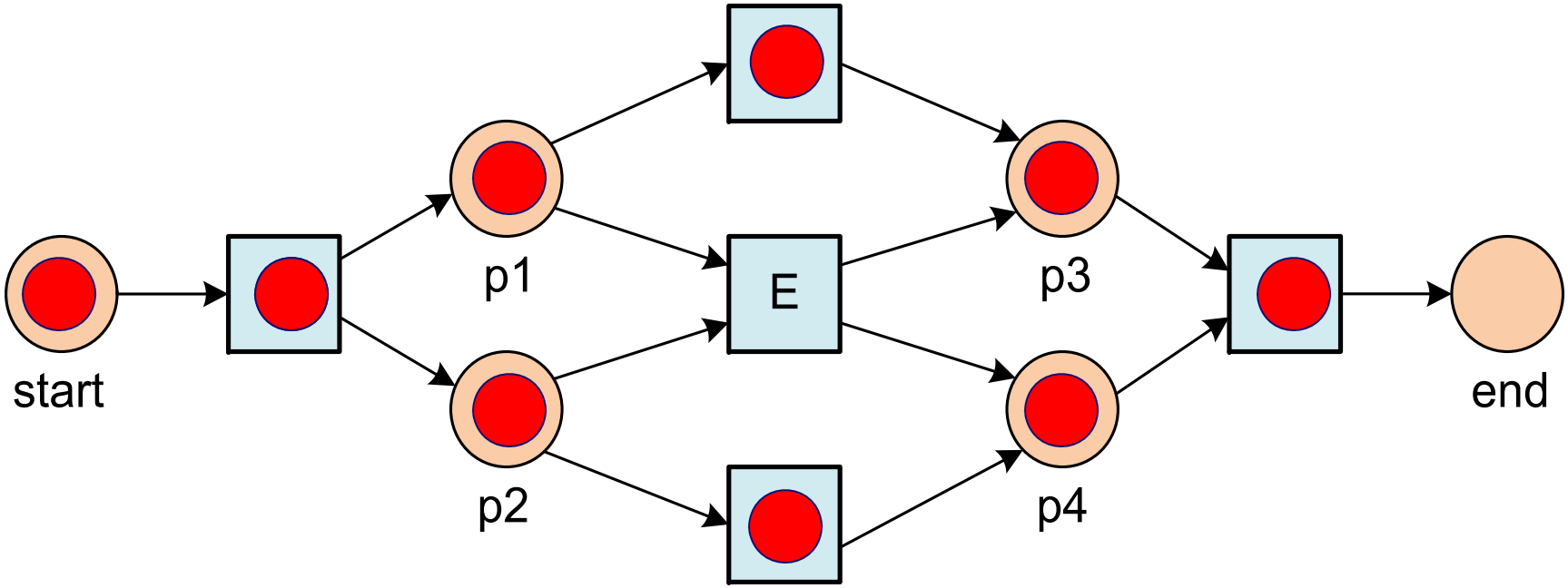


process model



event log

Play-Out (Classical use of models)

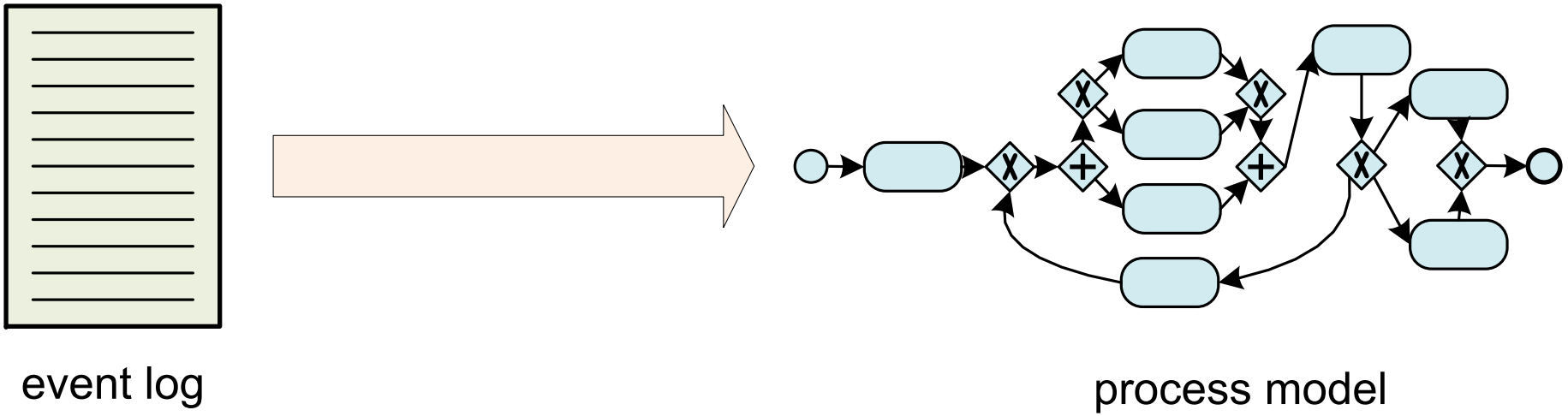


A B C D A E D A E D

A B C D A C B D

A C B D A E D A C B D

Play-In



event log

process model

Play-In

A B C D

A E D

A E D

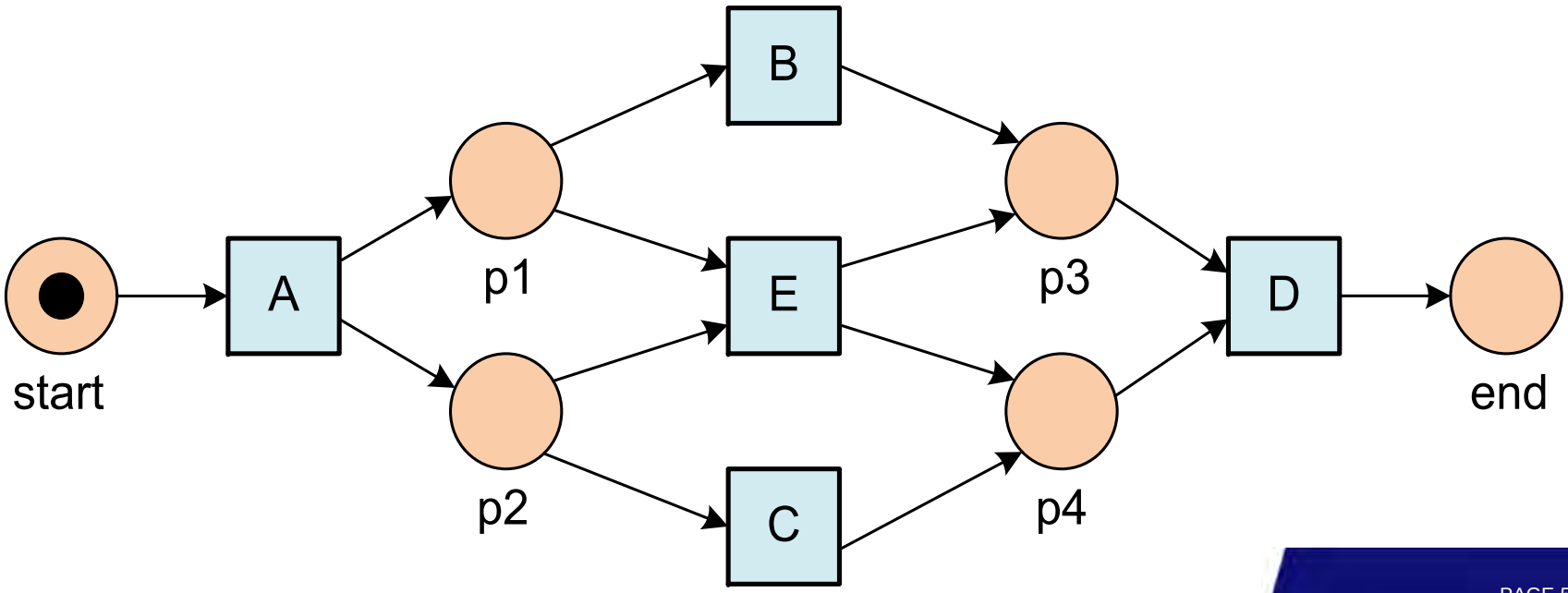
A B C D

A C B D

A C B D

A E D

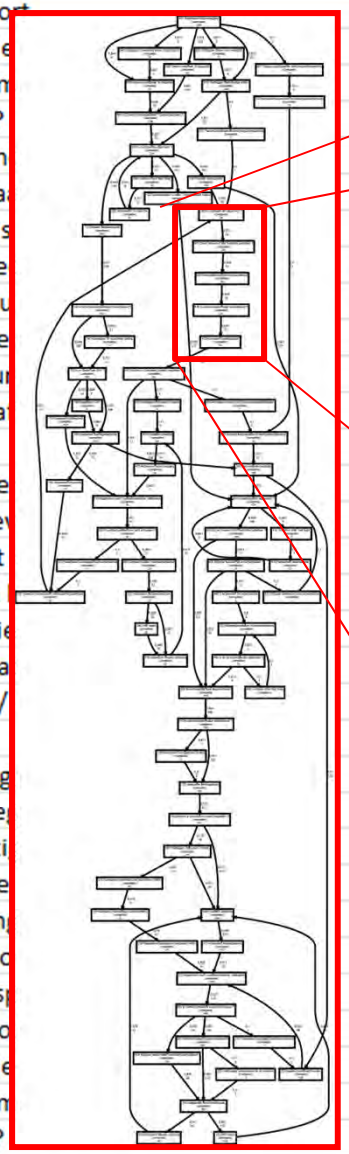
A C B D

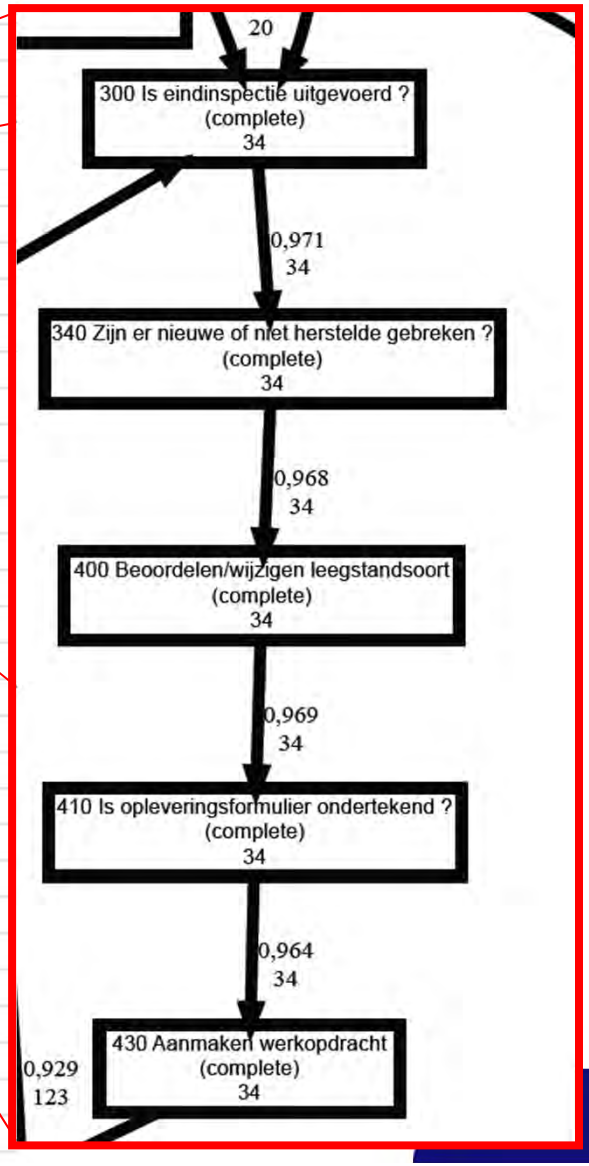


Example Process Discovery

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

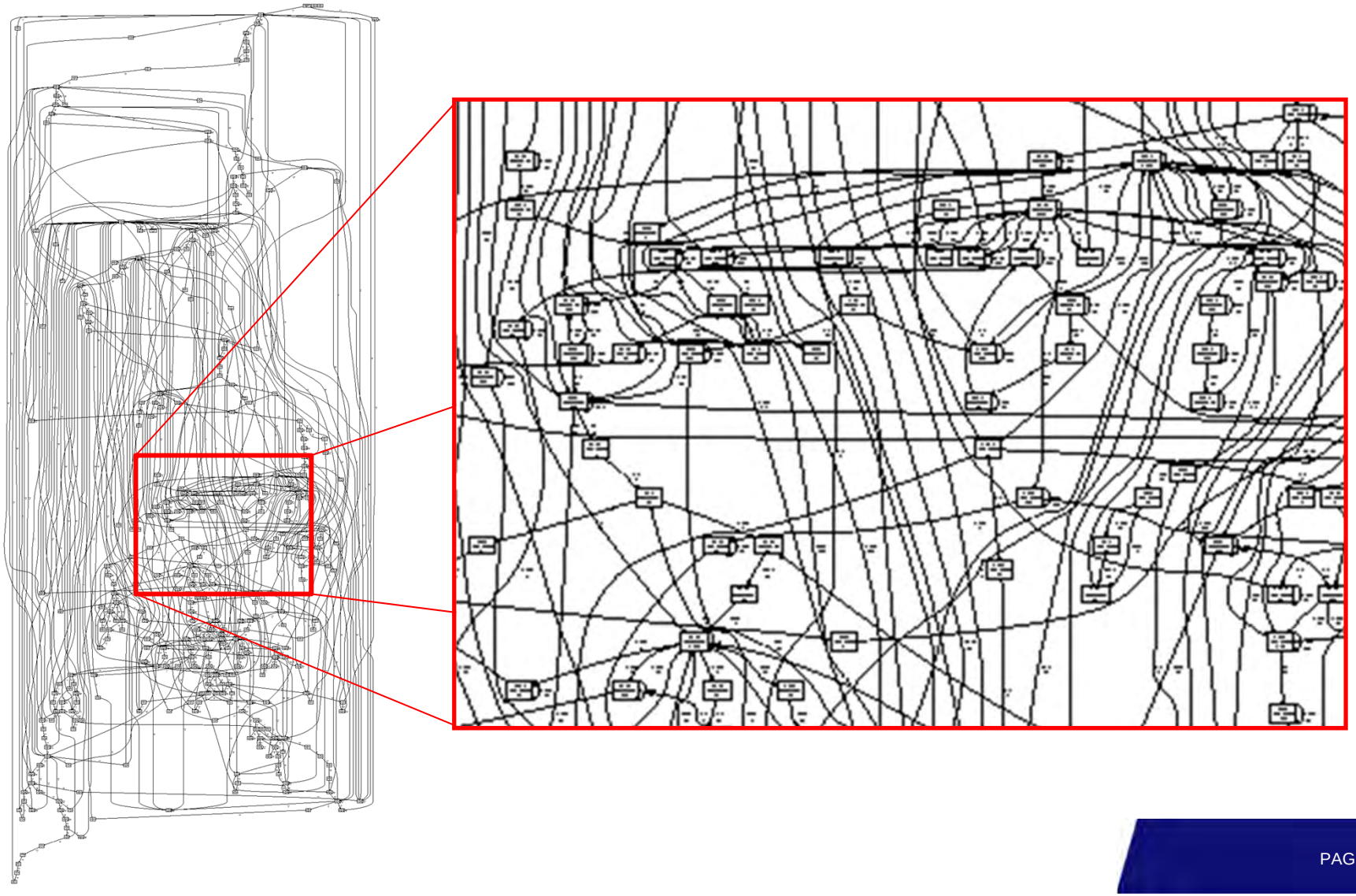
117315	110 Bepalen leegstandsoo	16.05.2007 14:06:23
117315	120 Plannen eindinspectie	16.05.2007 14:36:01
117315	130 Is het opleveringsform	23.05.2007 09:41:40
117315	150 Is er sprake van ZAV ?	23.05.2007 09:41:51
117315	170 Aanpassen plattegron	23.05.2007 11:57:18
117315	180 Aanpassen woningwa	23.05.2007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





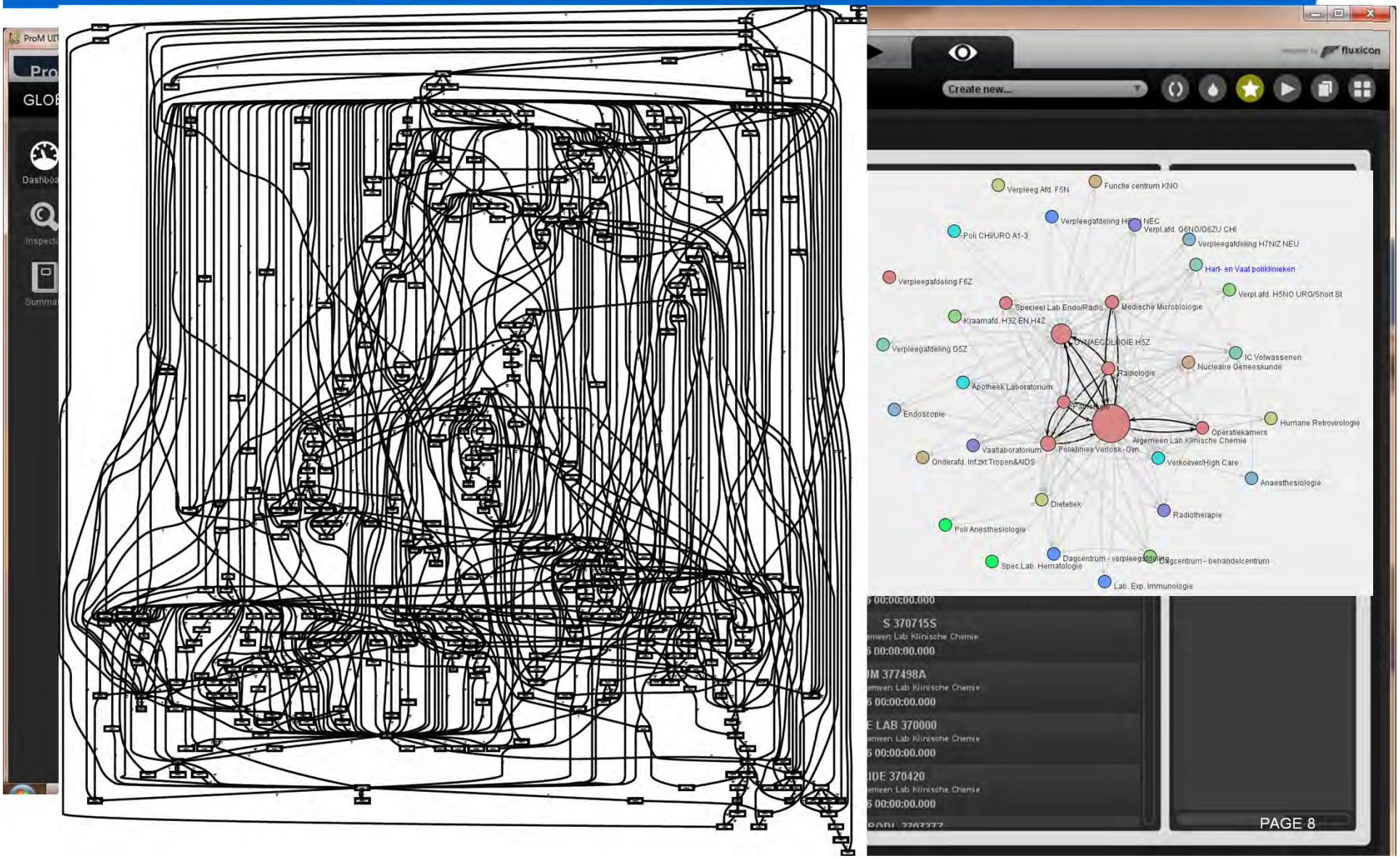
Example Process Discovery

(ASML, test process lithography systems, 154966 events)

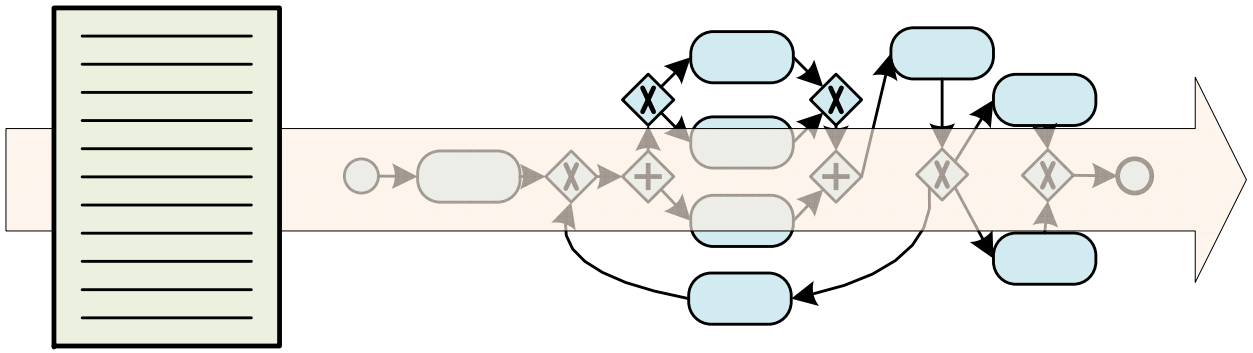


Example Process Discovery

(AMC, 627 gynecological oncology patients, 24331 events)



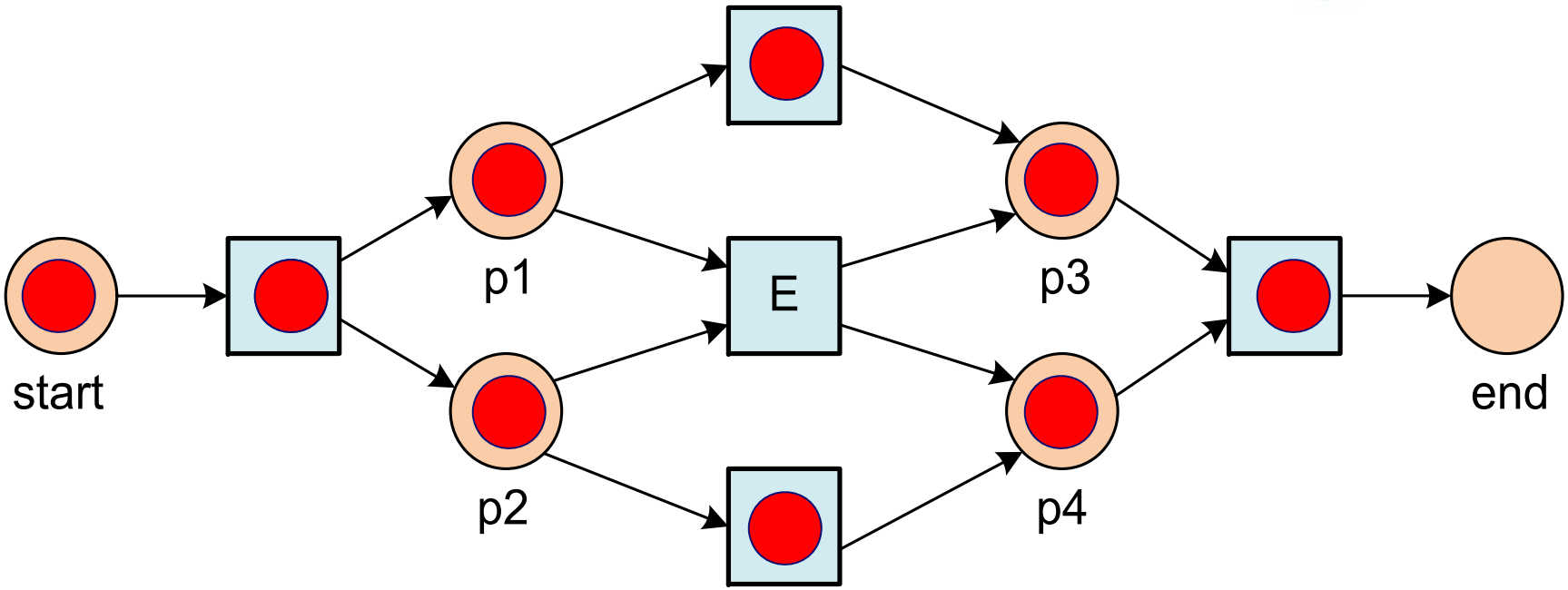
Replay



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

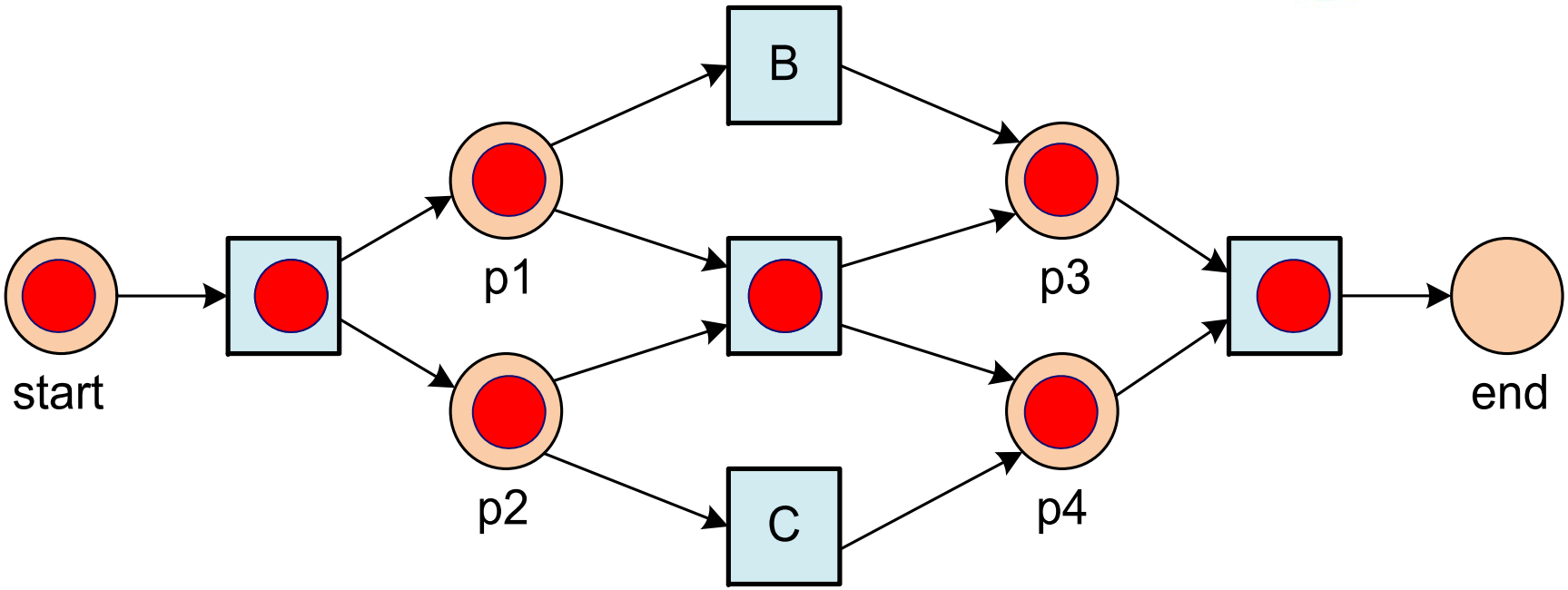
Replay

A B C D

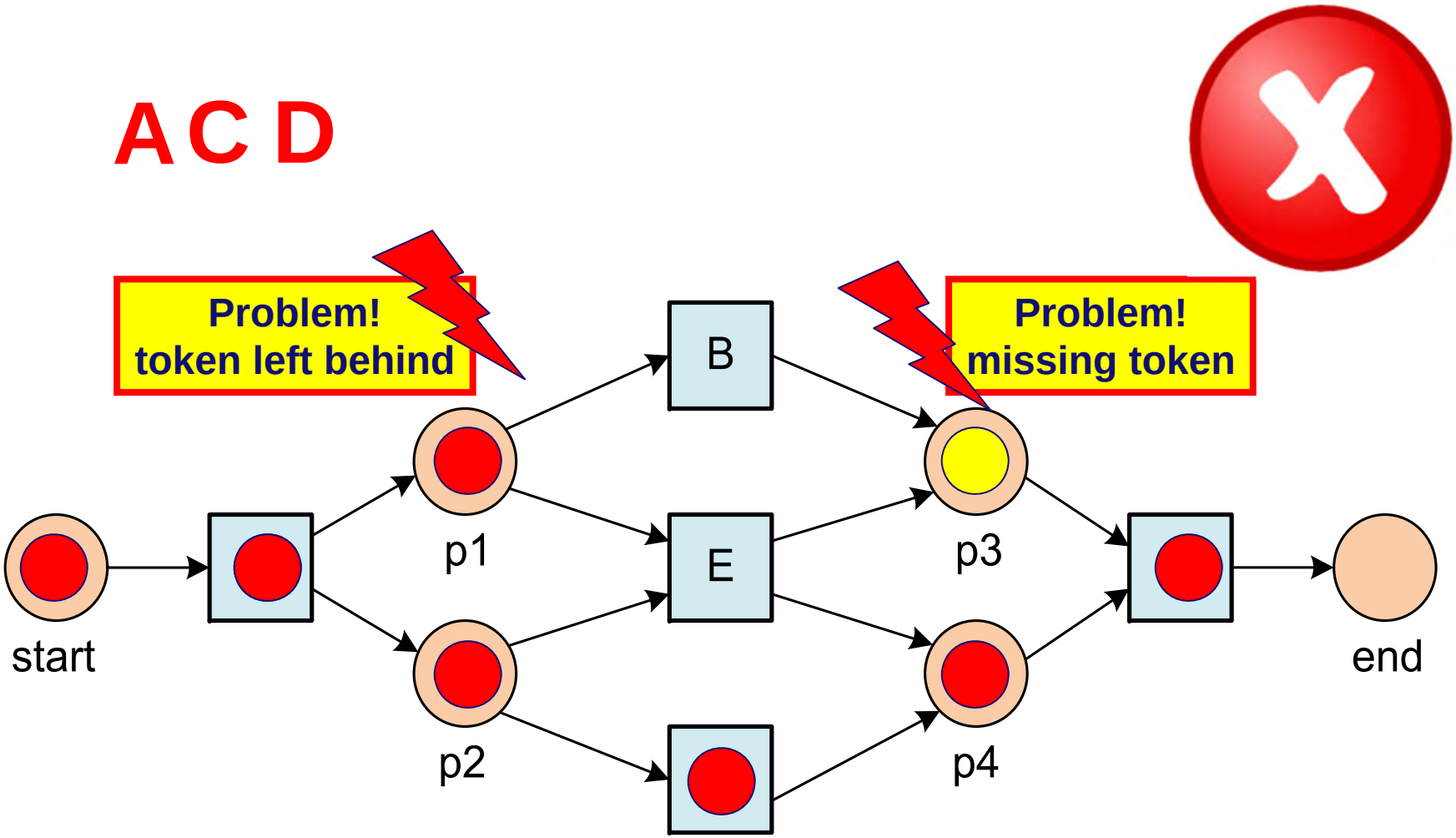


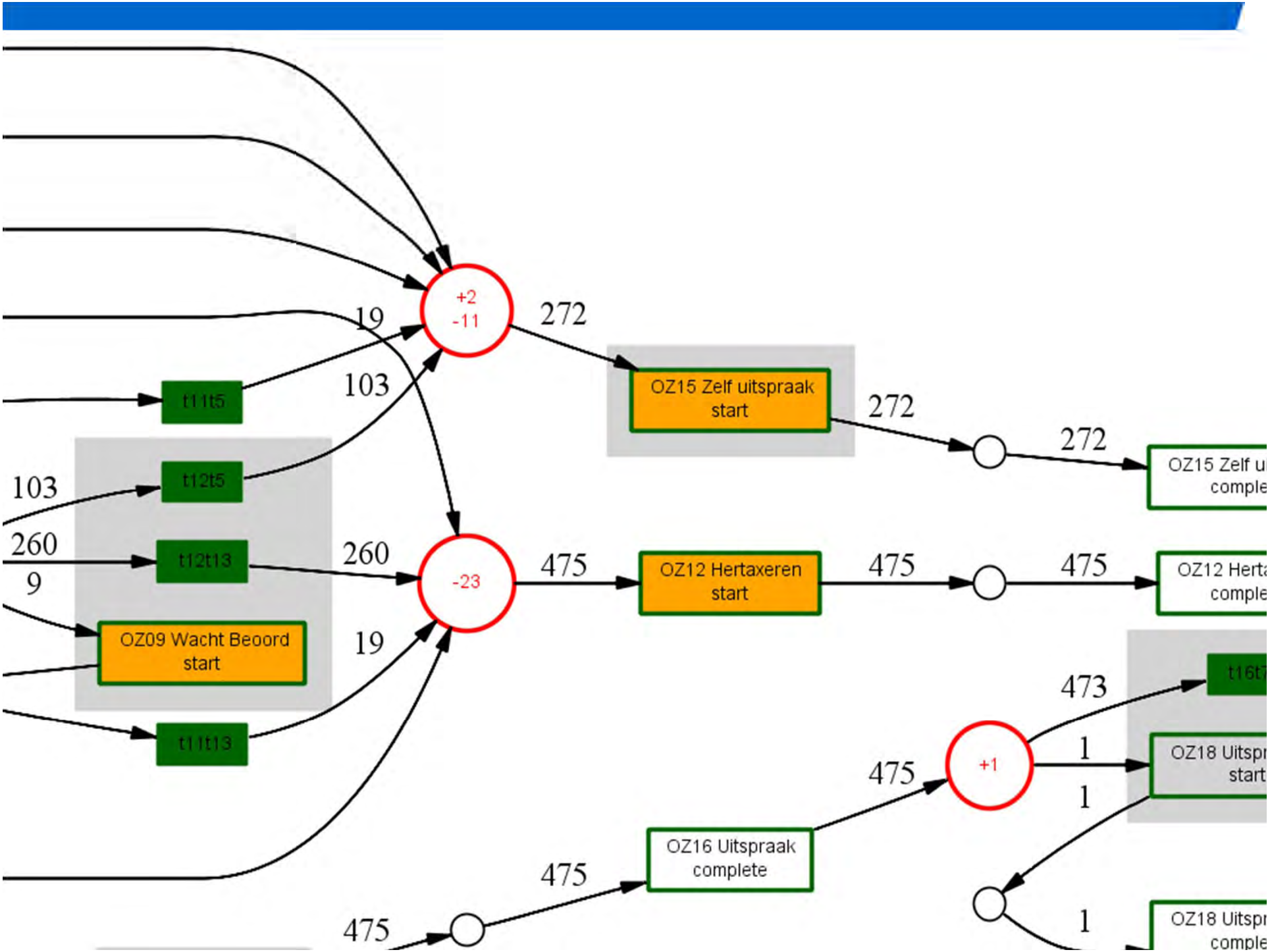
Replay

A E D



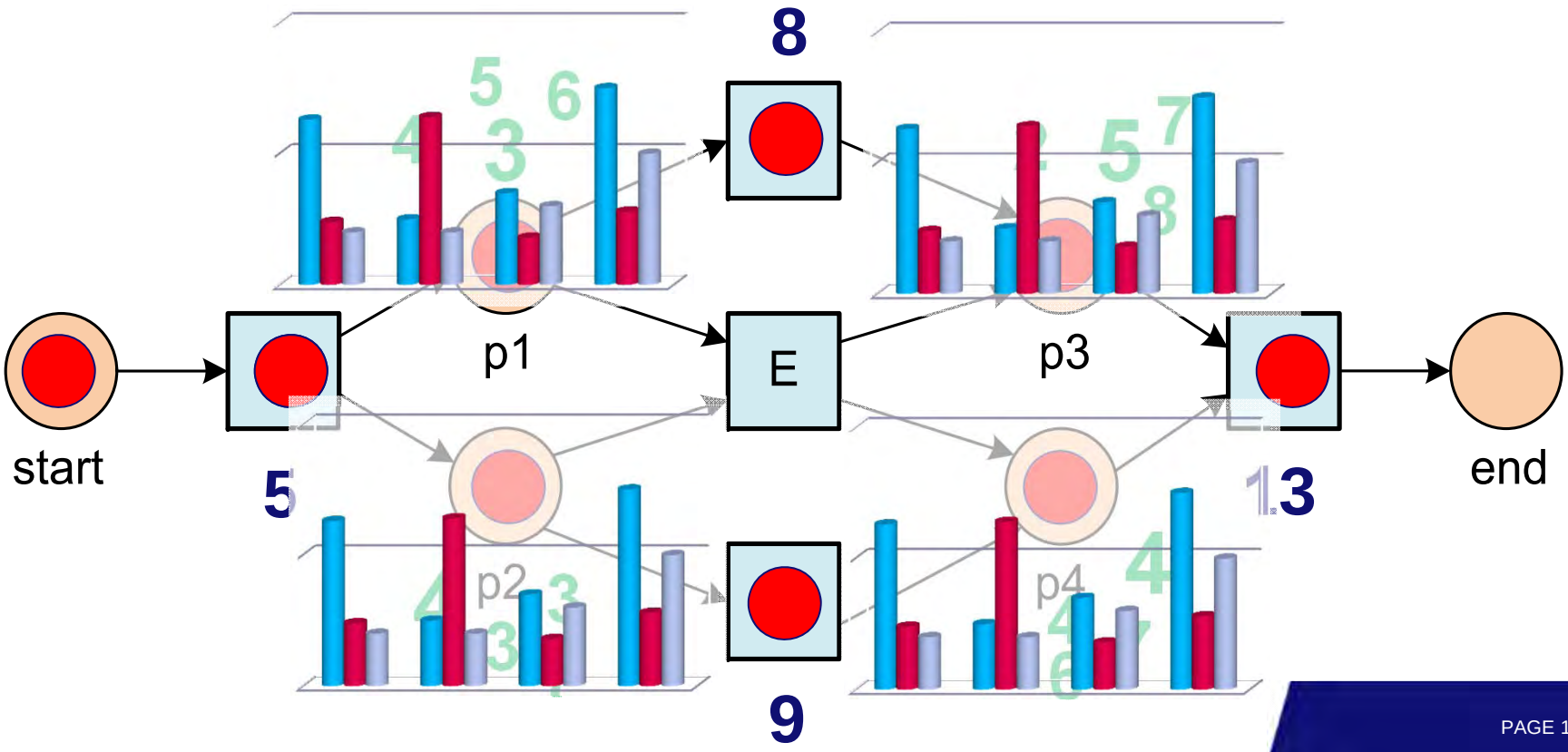
Replay can detect problems




+2
-11
19
103
272
t11t5
OZ15 Zelf uitspraak start
272
272
OZ15 Zelf u
comple
103
t12t5
260
t12t13
260
-23
475
OZ12 Hertaxeren start
475
475
OZ12 Herta
comple
9
OZ09 Wacht Beoord start
19
t11t13
473
t16t7
1
OZ18 Uitspr
start
475
+1
1
OZ16 Uitspraak complete
475
475
1
OZ18 Uitspr
comple

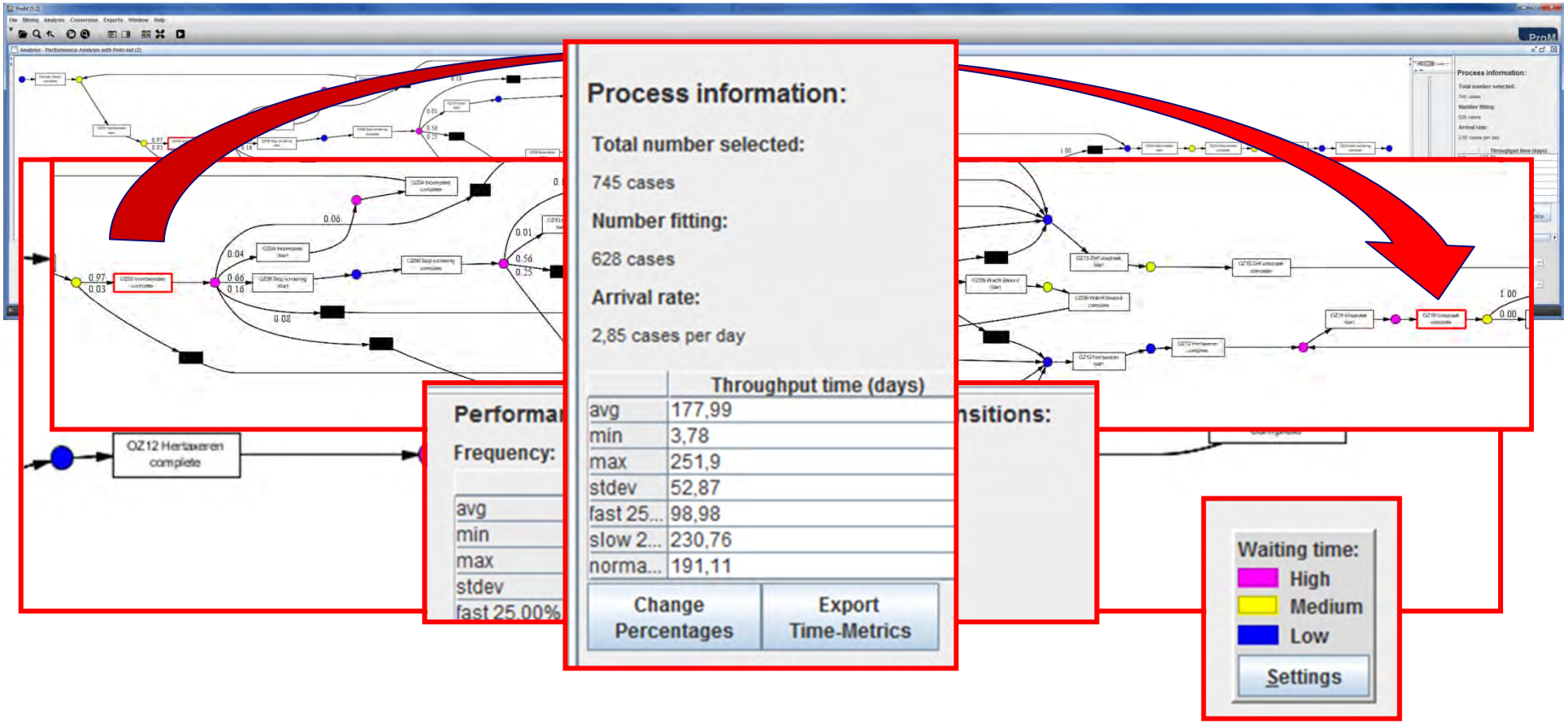
Replay can extract timing information

$A^{5}B^{8}C^{9}D^{13}$



Performance Analysis Using Replay

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



Big Data

Big Data

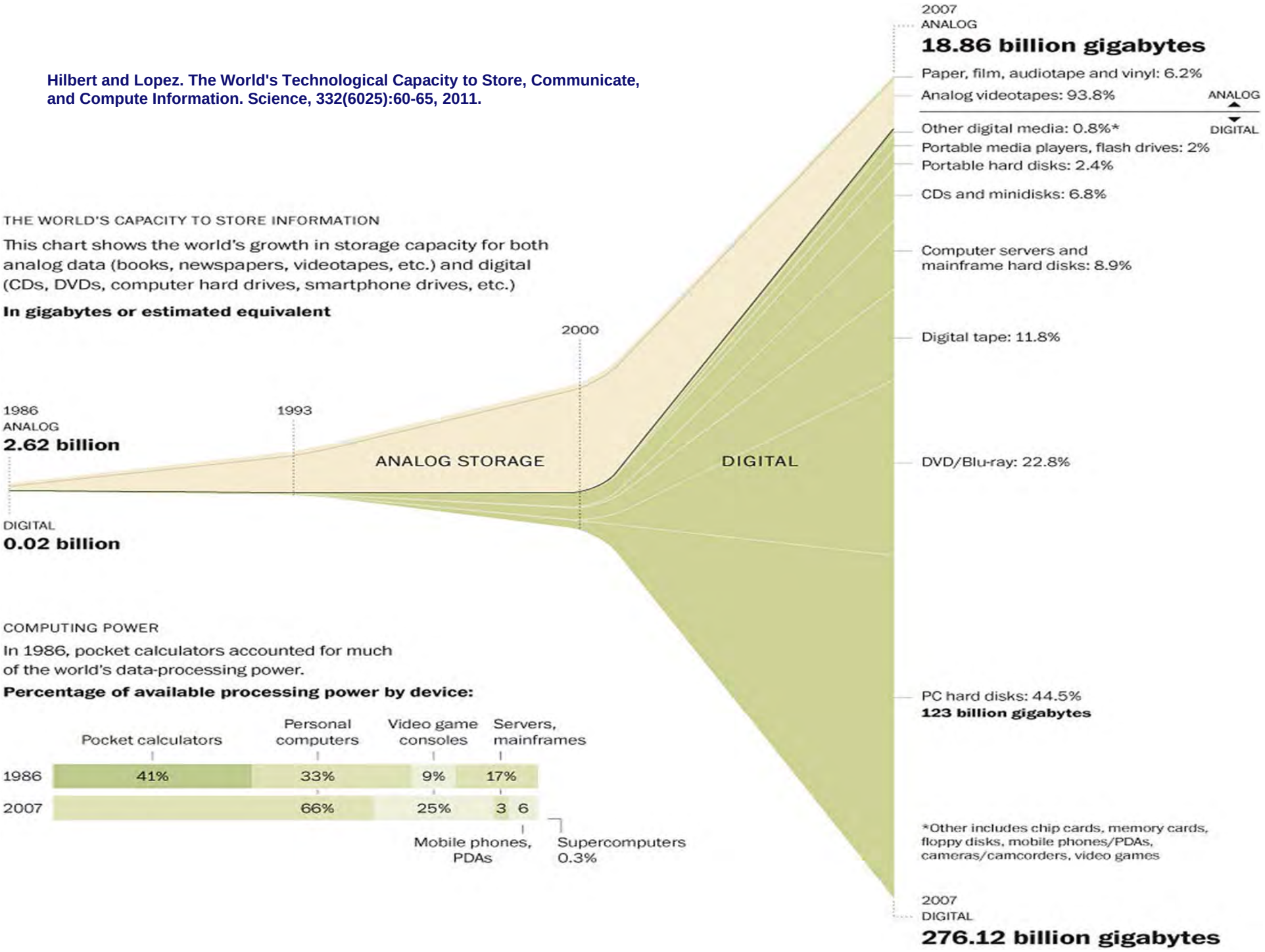
"All of the world's music can be stored on a $600 disk drive."

"Enterprises globally stored more than 7 exabytes of new data on disk drives in 2010, while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks."

"Indeed, we are generating so much data today that it is physically impossible to store it all. Health care providers, for instance, discard 90 percent of the data that they generate."

Source: "Big Data: The Next Frontier for Innovation, Competition, and Productivity" McKinsey Global Institute, 2011.

Hilbert and Lopez. The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025):60-65, 2011.

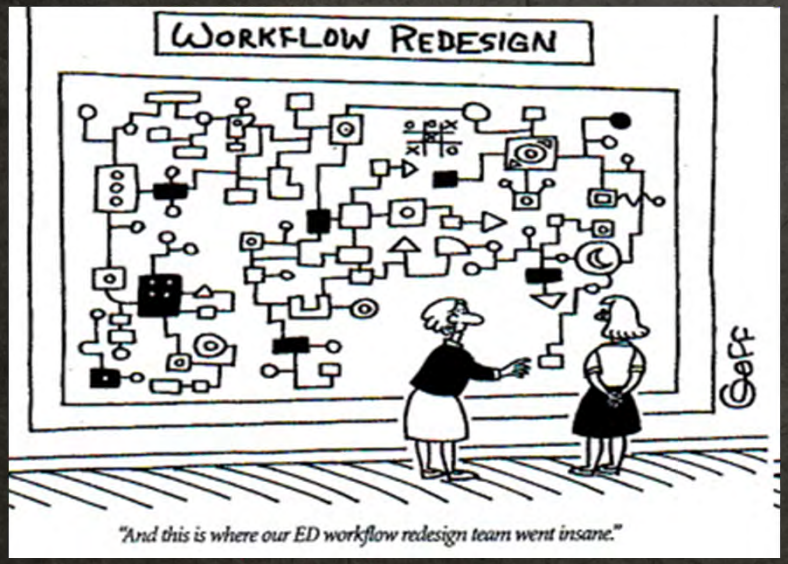



www.olifantenpaadjes.nl
01001100110101010
0100110101010
60


0100110101010101010

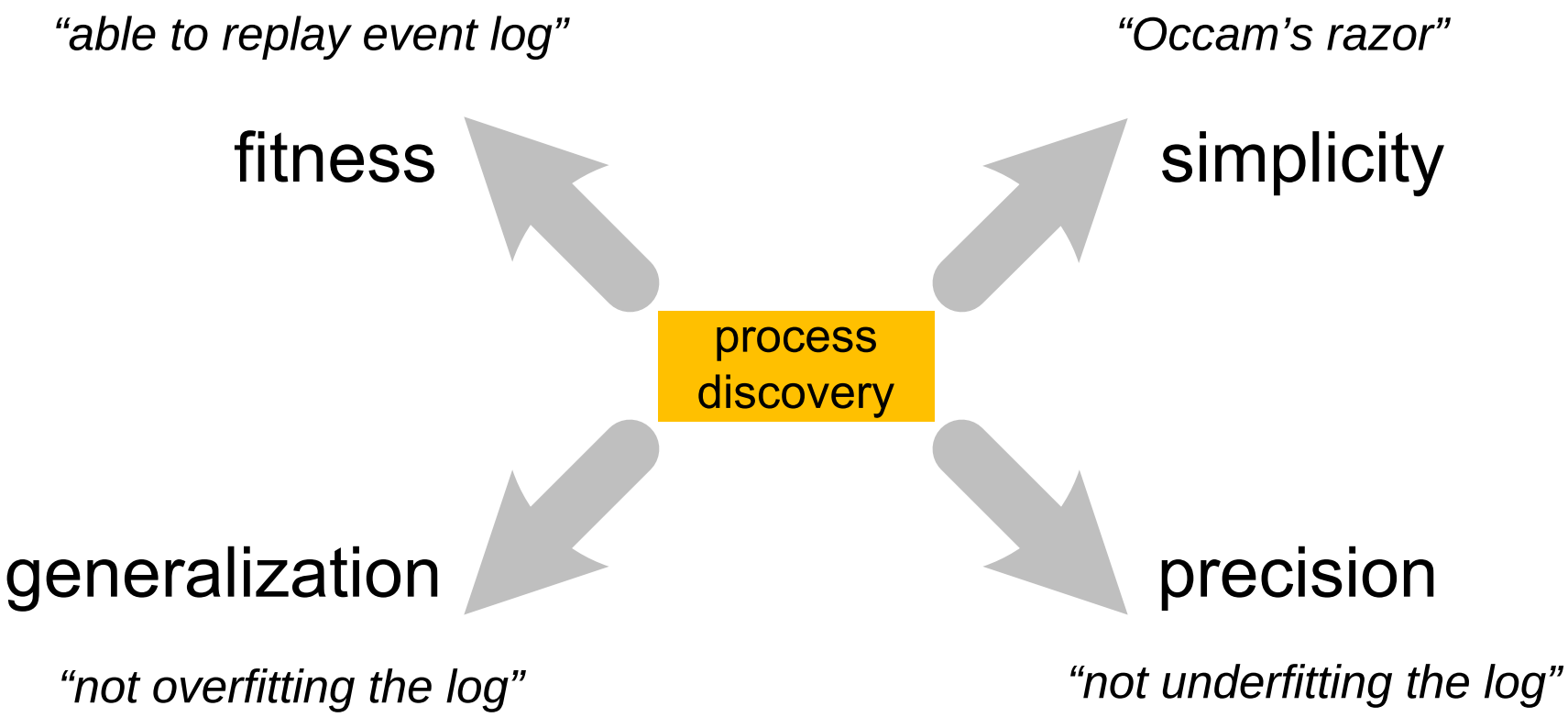

0100110011010101010

Evidence-Based Computer Science

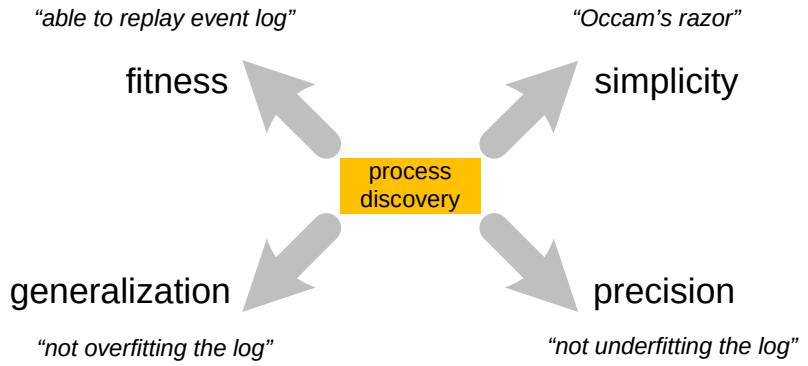


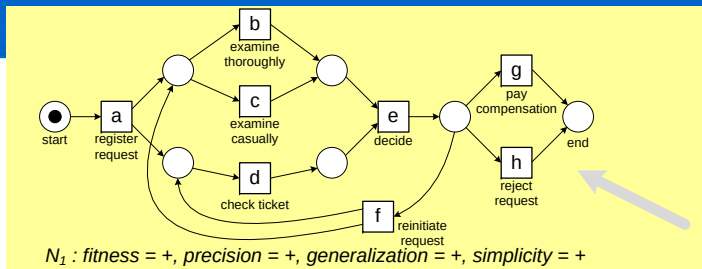
How good is my model?

Four Competing Quality Criteria

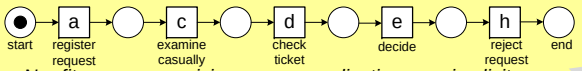


Example: one log four models

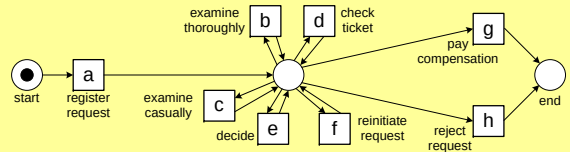




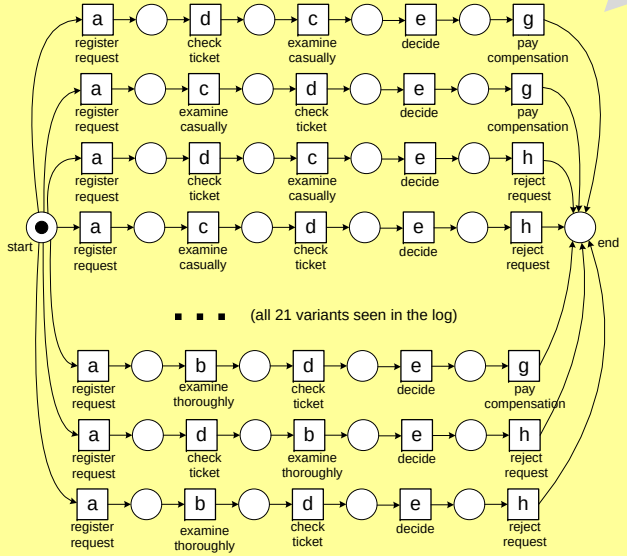
N_1 : *fitness* = +, *precision* = +, *generalization* = +, *simplicity* = +



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



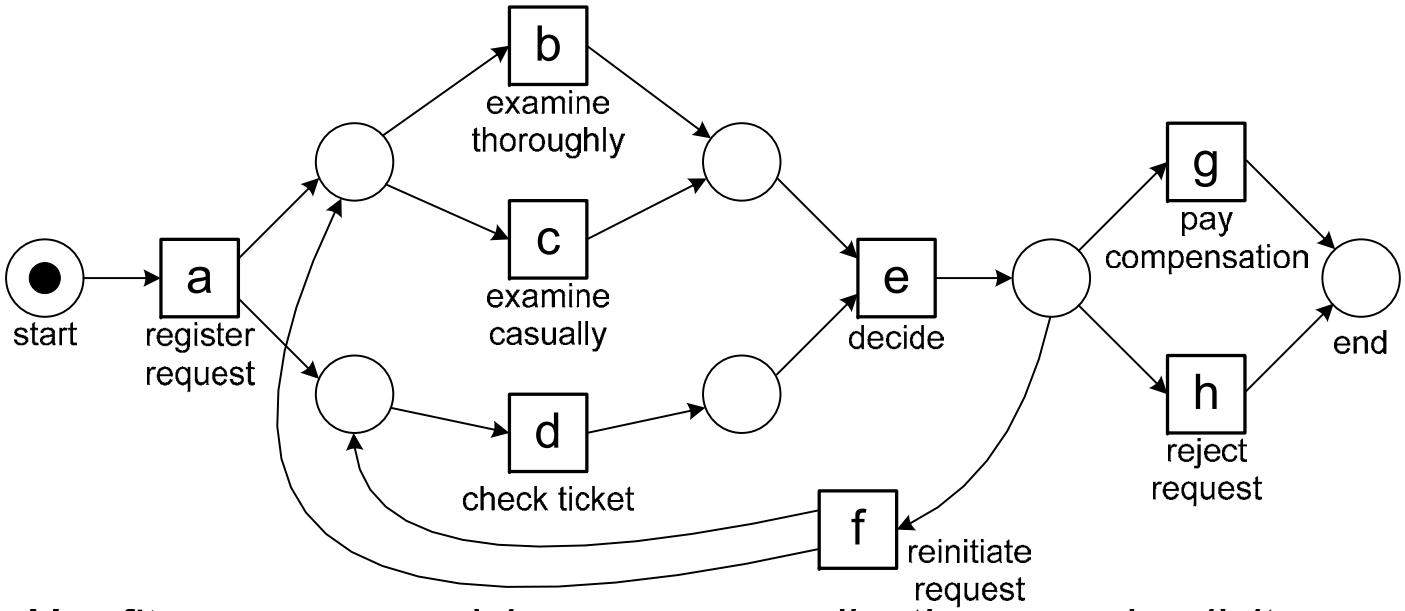
N_3 : *fitness* = +, *precision* = -, *generalization* = +, *simplicity* = +



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

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

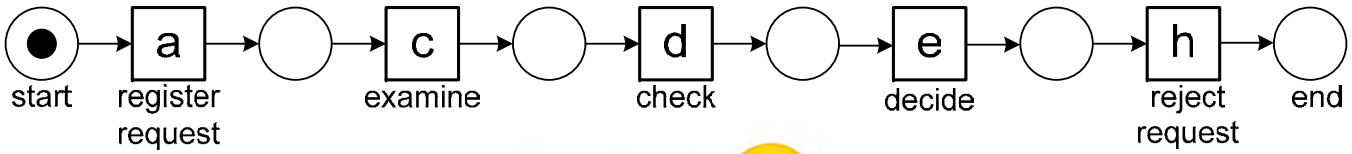
Model N_1



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

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

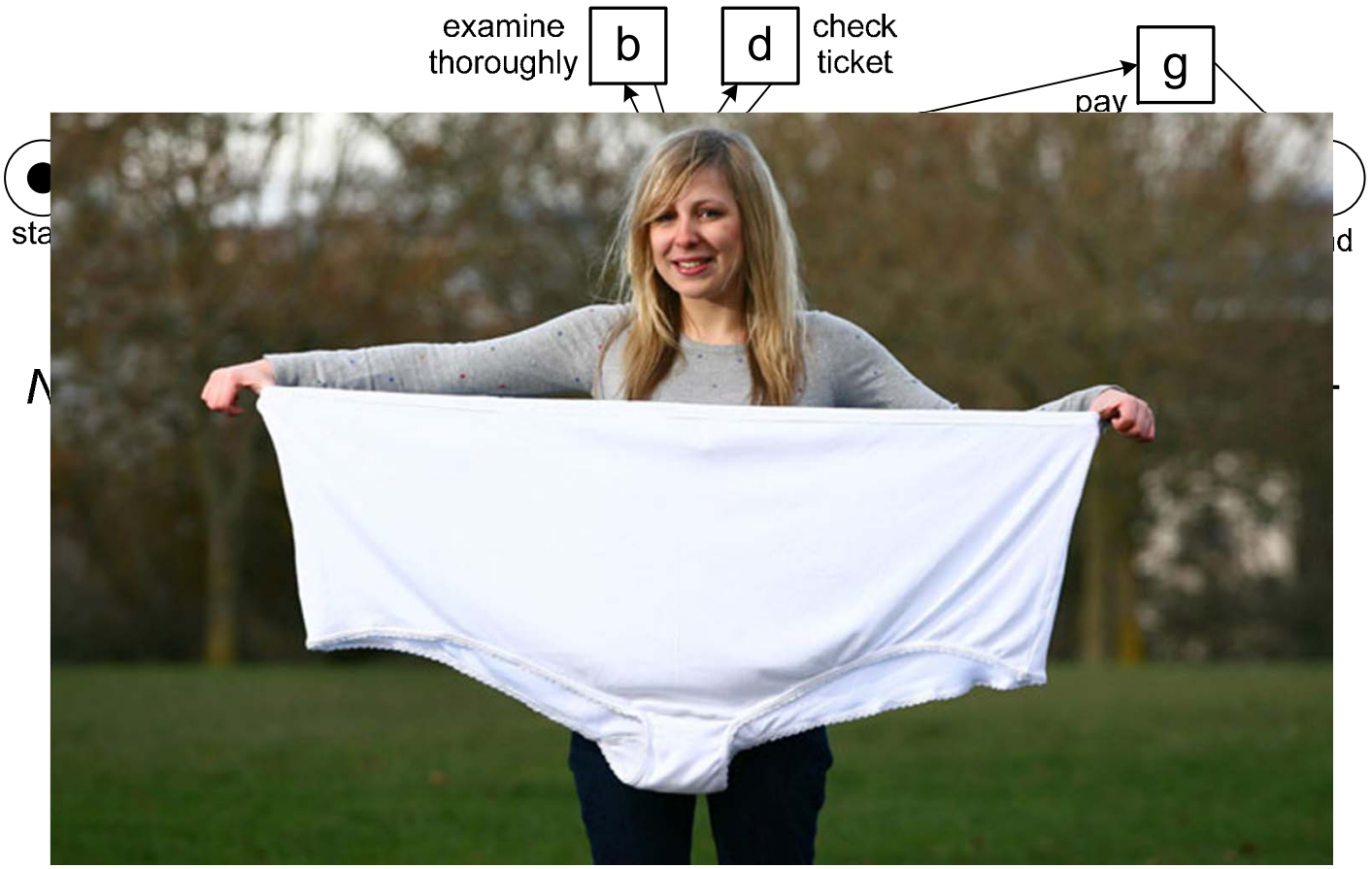
Model N_2



N_2 : *fitness* = , *simplicity* = +

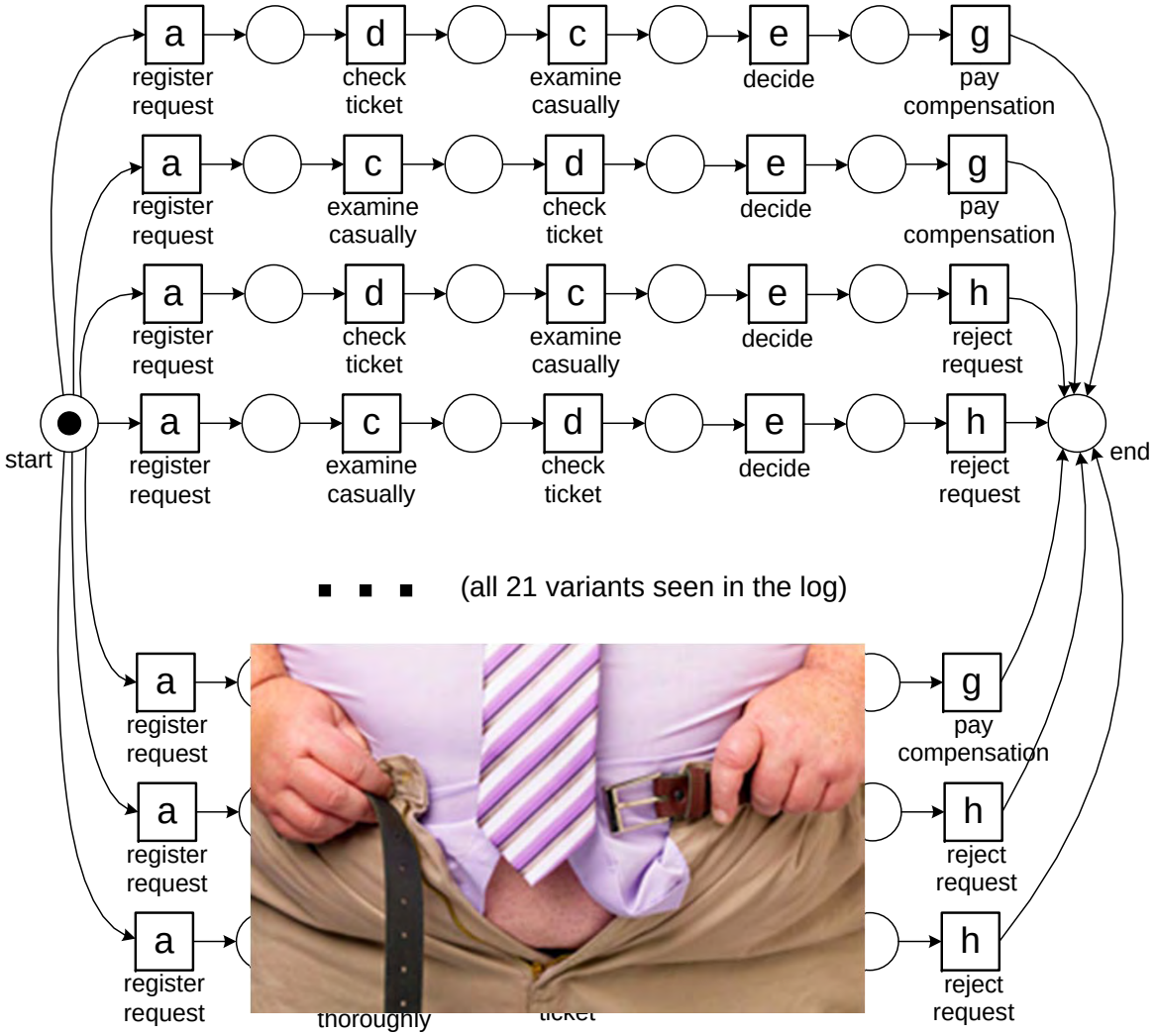
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

Model N_3



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

Model N_4



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

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

Process Discovery

Process Discovery (small selection)

automata-based learning

distributed genetic mining

language-based regions

heuristic mining

genetic mining

state-based regions

stochastic task graphs

LTL mining

neural networks

fuzzy mining

hidden Markov models

mining block structures

α algorithm

conformal process graph

multi-phase mining

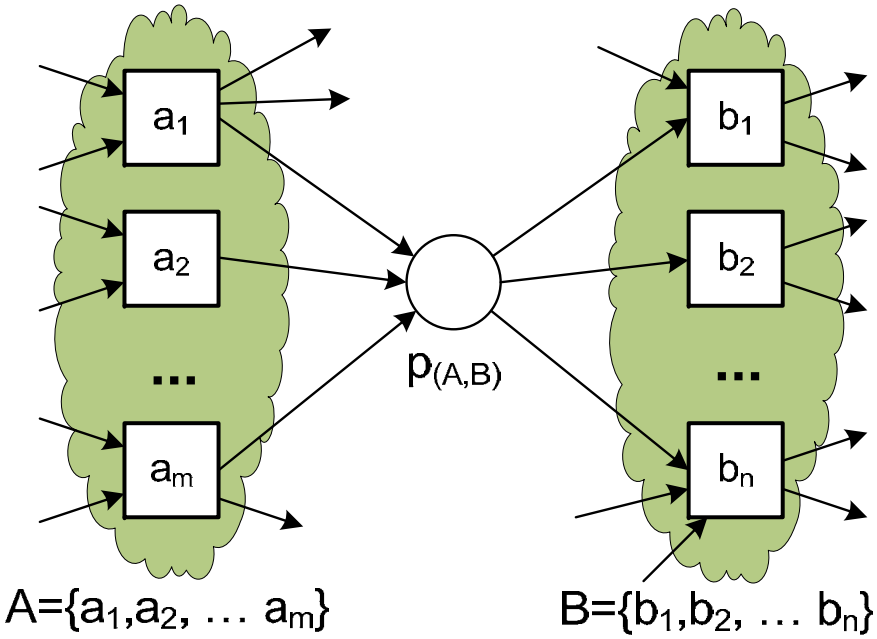
partial-order based mining

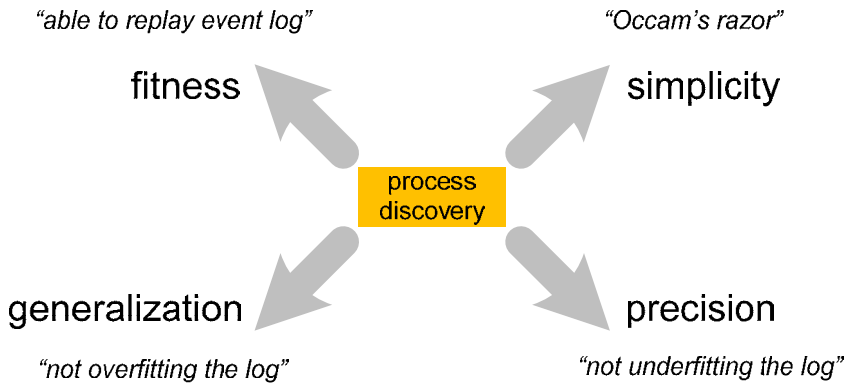
α# algorithm

ILP mining

α++ algorithm

Petri net view:
Just discover the places …

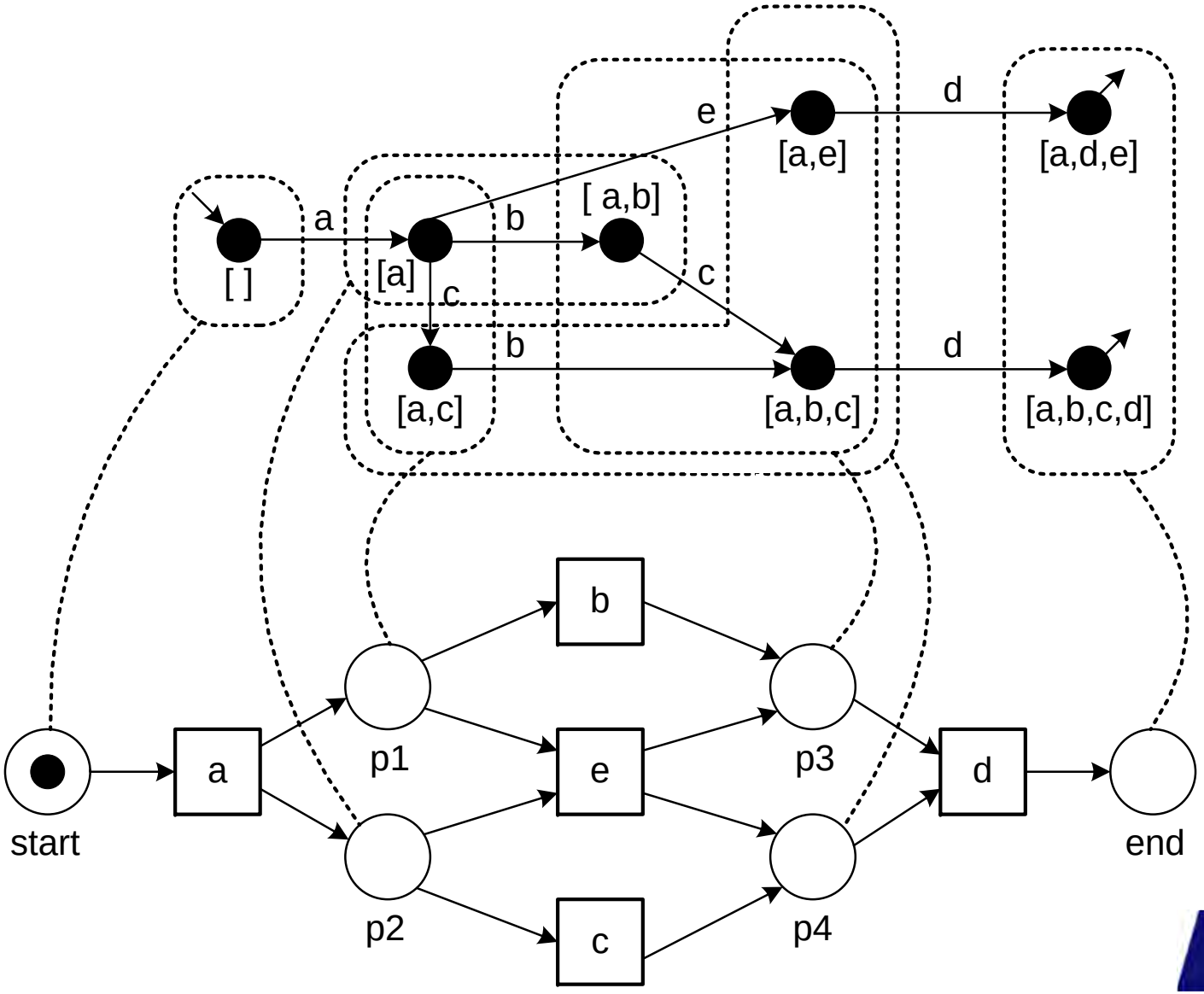




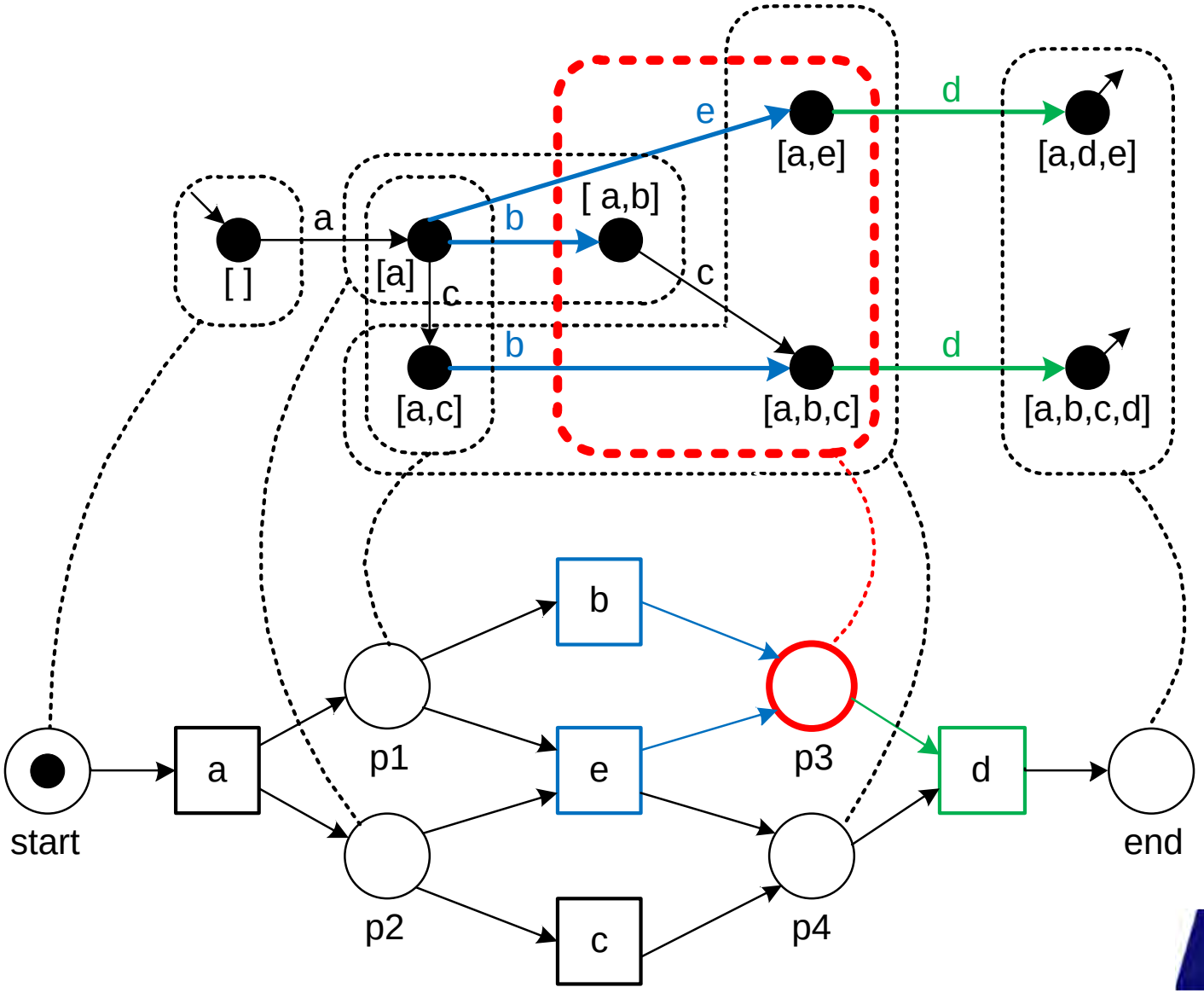
Adding a place limits behavior:

- **overfitting ≈ adding too many places**
- **underfitting ≈ adding too few places**

Example: Process Discovery Using State-Based Regions

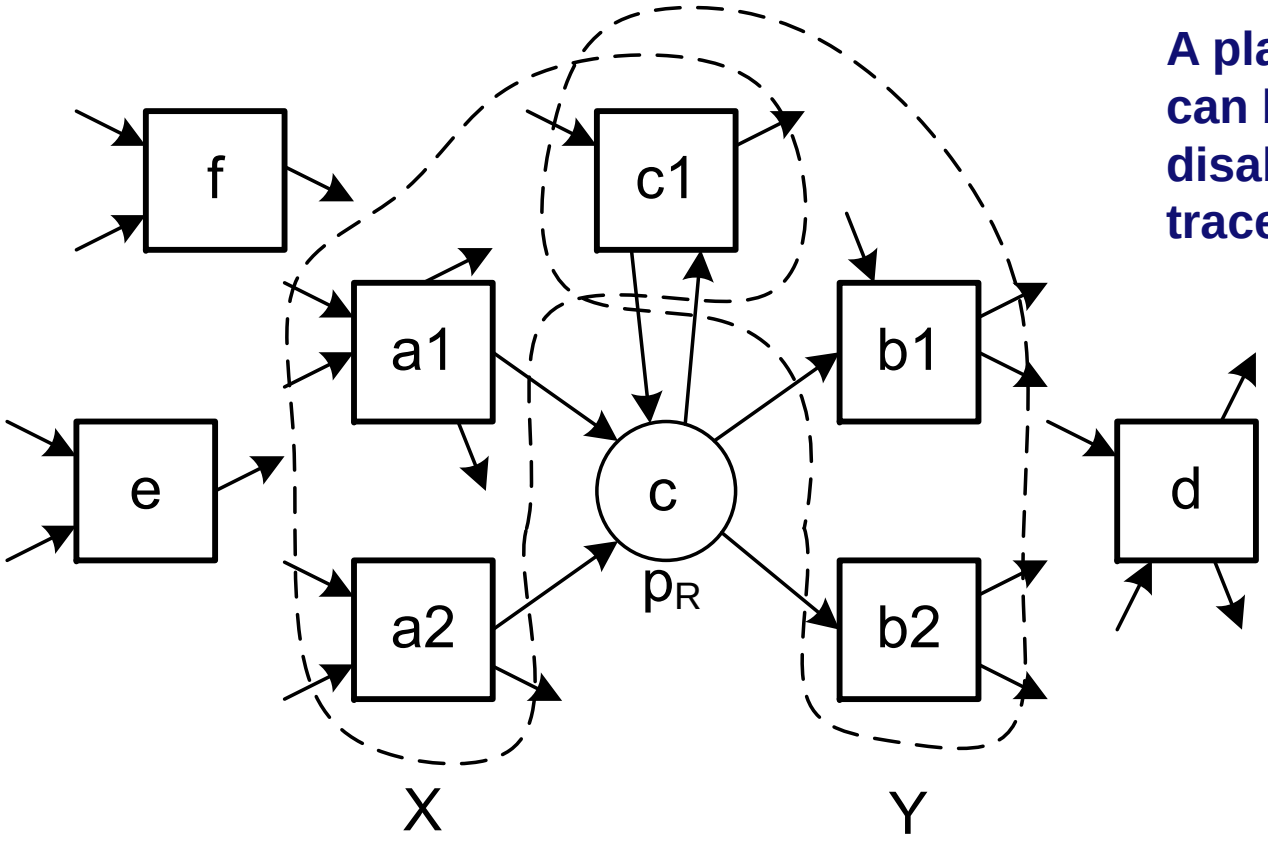


Example of Region



enter: b,e
leave: d
do-not-cross: a,c

Example: Process Discovery Using Language-Based Regions



A place is feasible if it can be added without disabling any of the traces in the event log.

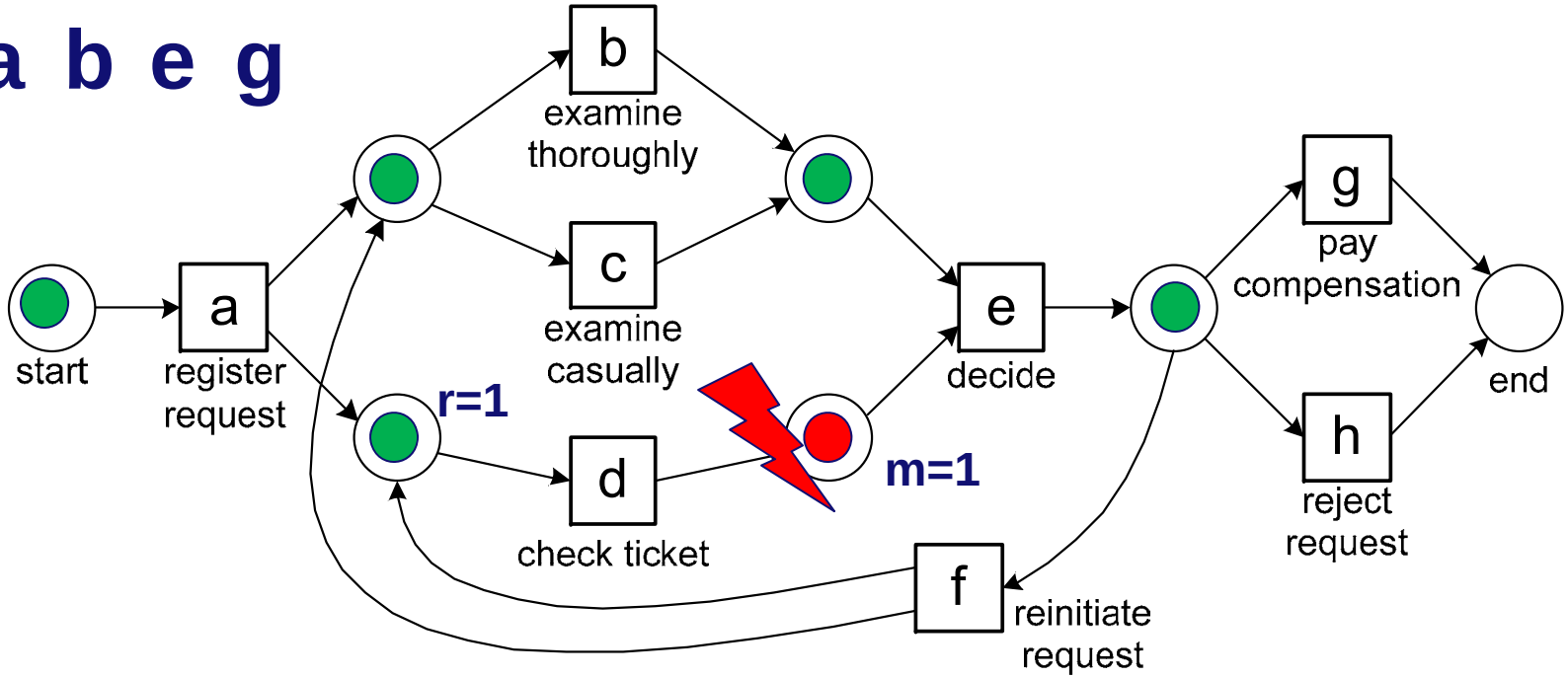
for any $\sigma \in L, k \in \{1, \ldots, |\sigma|\}, \sigma_1 = hd^{k-1}(\sigma), a = \sigma(k), \sigma_2 = hd^{k}(\sigma) = \sigma_1 \oplus a$:

$$c + \sum_{t \in X} \partial_{multiset}(\sigma_1)(t) - \sum_{t \in Y} \partial_{multiset}(\sigma_2)(t) \geq 0.$$

Conformance Checking

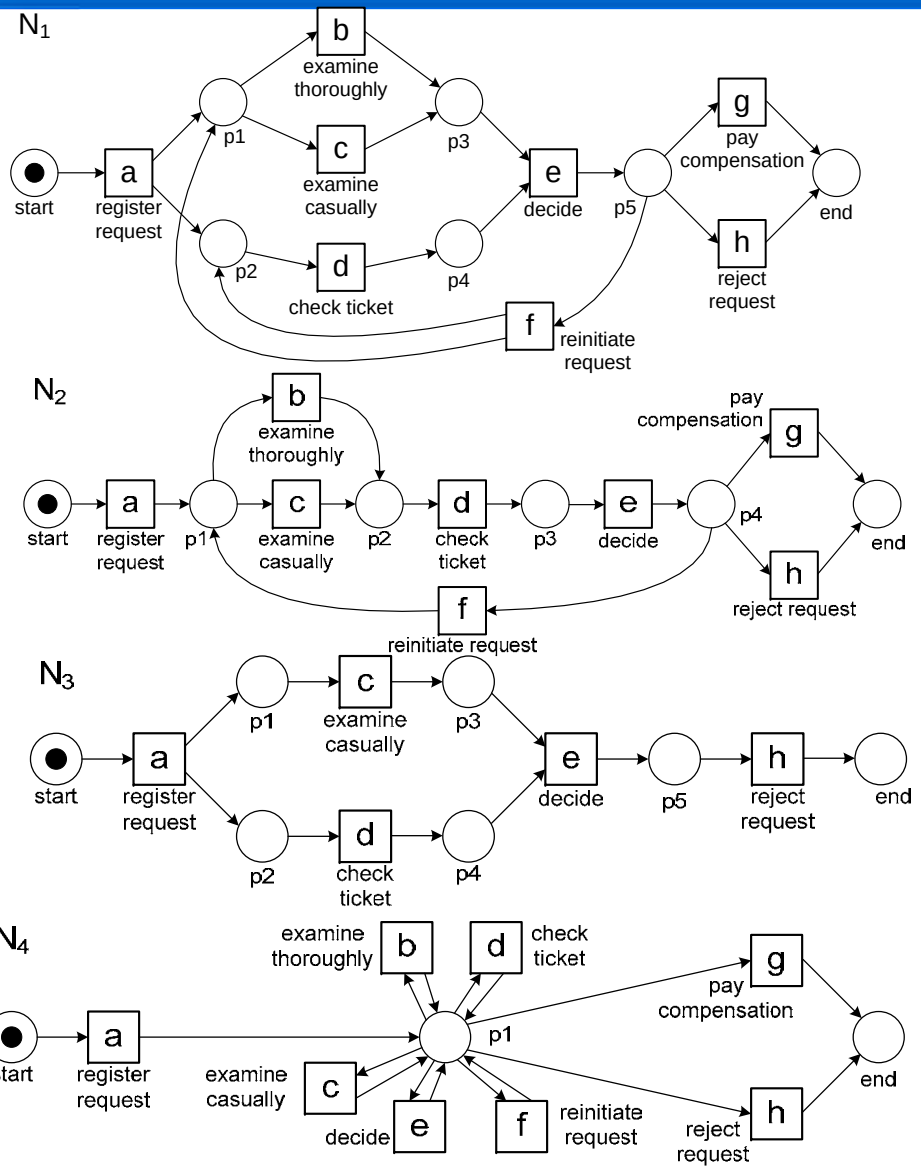
Replaying trace “abeg”

a b e g



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

Can be lifted to log level



#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

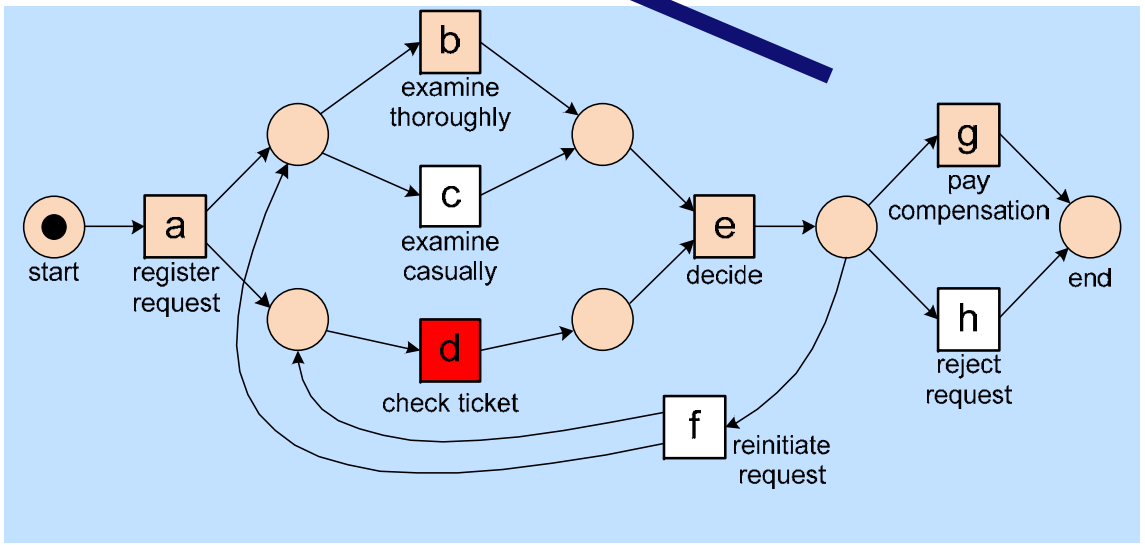
$$fitness(L_{full}, N_1) = 1$$
$$fitness(L_{full}, N_2) = 0.9504$$
$$fitness(L_{full}, N_3) = 0.8797$$
$$fitness(L_{full}, N_4) = 1$$

From “playing the token game” to optimal alignments …

observed trace: “abeg”

a	b	»	e	g
a	b	d	e	g

move in model only

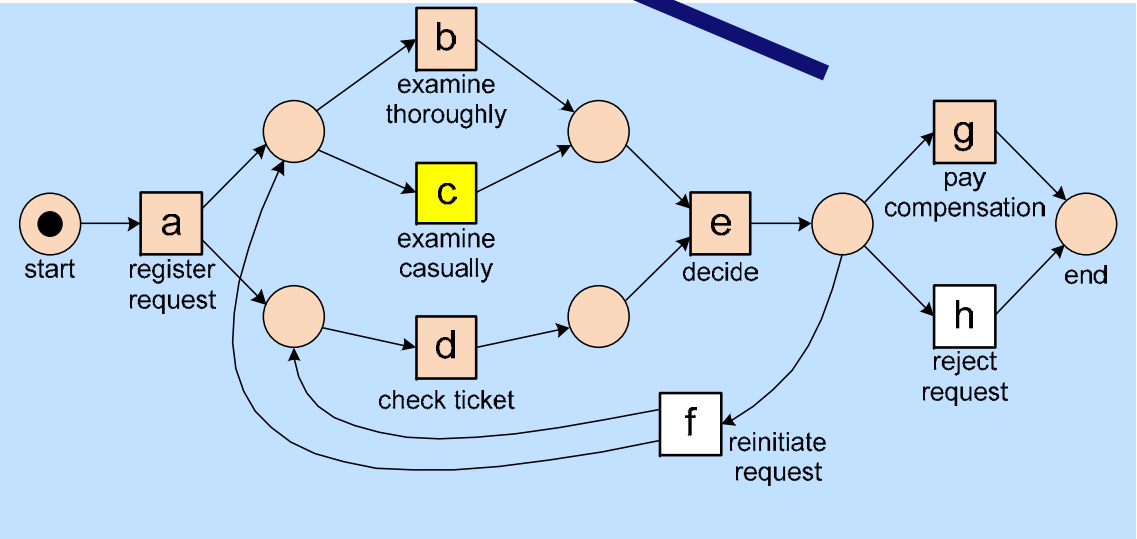


Another alignment

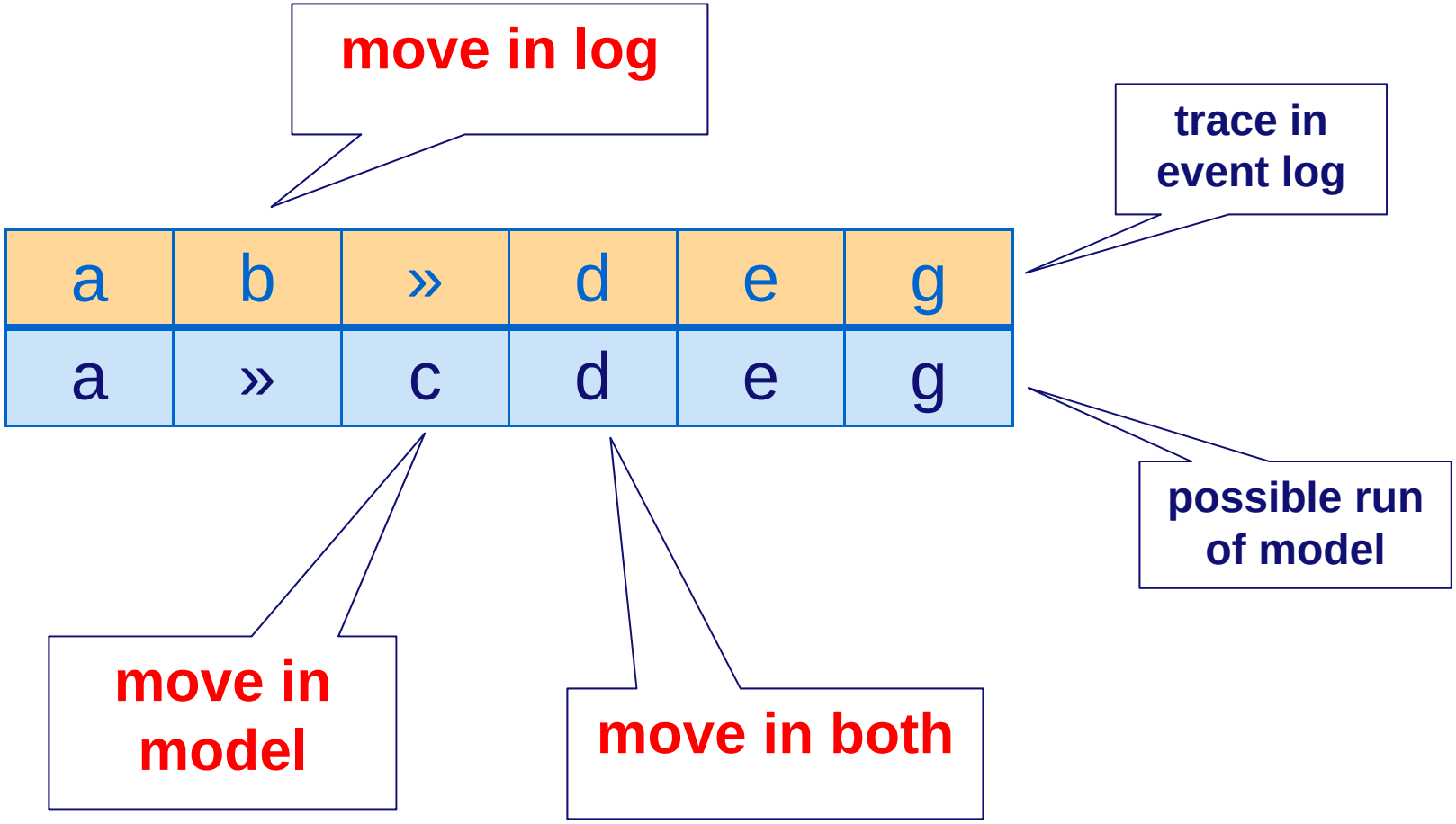
observed trace: “abcdeg”

a	b	c	d	e	g
a	b	»	d	e	g

move in log only

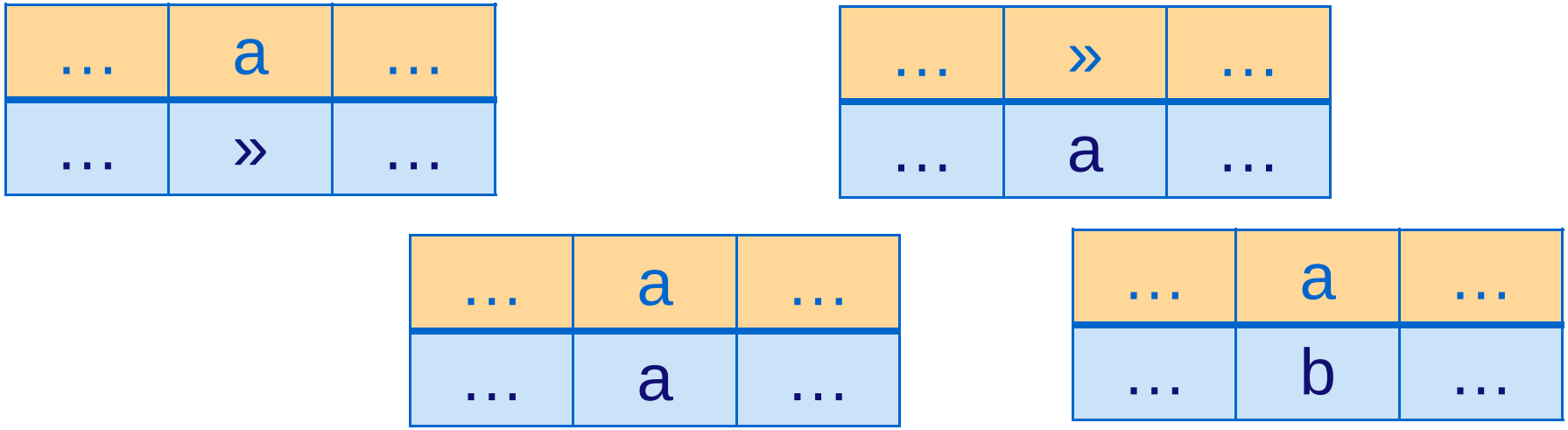


Moves in an alignment



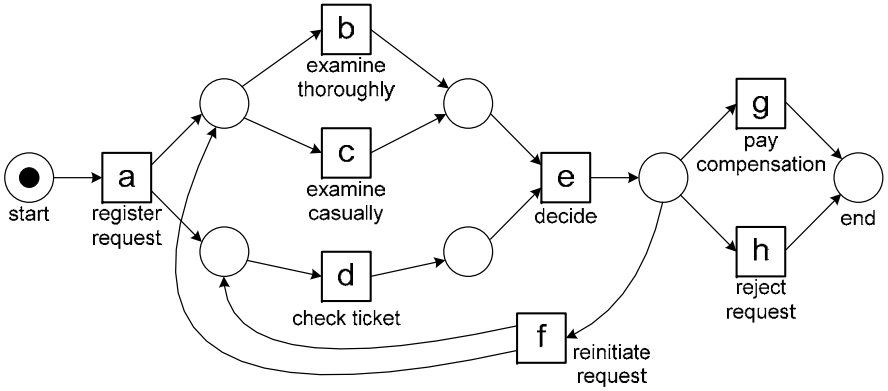
Optimal alignment describes modeled behavior closest to observed behavior

Moves have costs



- **Standard cost function:**
  - c(x,») = 1
  - c(»,y) = 1
  - c(x,y) = 0, if x=y
  - c(x,y) = ∞, if x≠y

Non-fitting trace: abefdeg



abefdeg

a	b	»	e	f	d	»	e	g
a	b	d	e	f	d	b	e	g

2

a	b	e	f	d	e	g
a	b	»	»	d	e	g

2

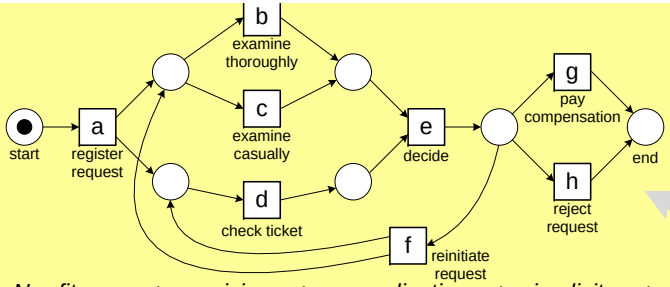
Any cost structure is possible

...	send-letter(John,2 weeks, $400)	...
...	send-email(Sue,3 weeks,$500)	...

- **Similar activities (more similarity implies lower costs).**
- **Resource conformance (done by someone that does not have the specified role).**
- **Data conformance (path is not possible for this customer).**
- **Time conformance (missed the legal deadline)**

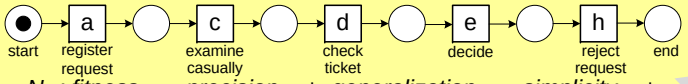
Fitness

1.0



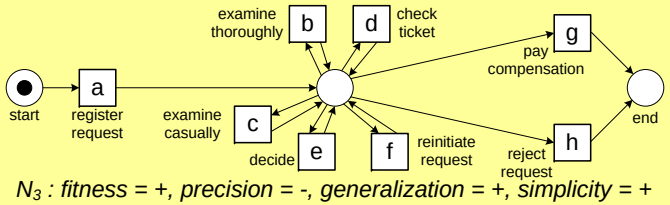
N_1 : *fitness* = +, *precision* = +, *generalization* = +, *simplicity* = +

0.8



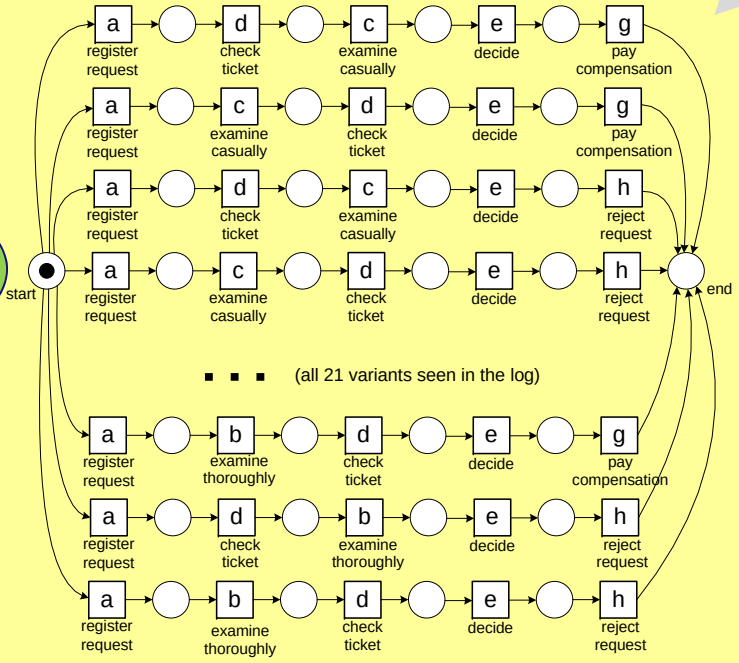
N_2 : *fitness* = -, *precision* = +, *generalization* = -, *simplicity* = +

1.0



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

1.0



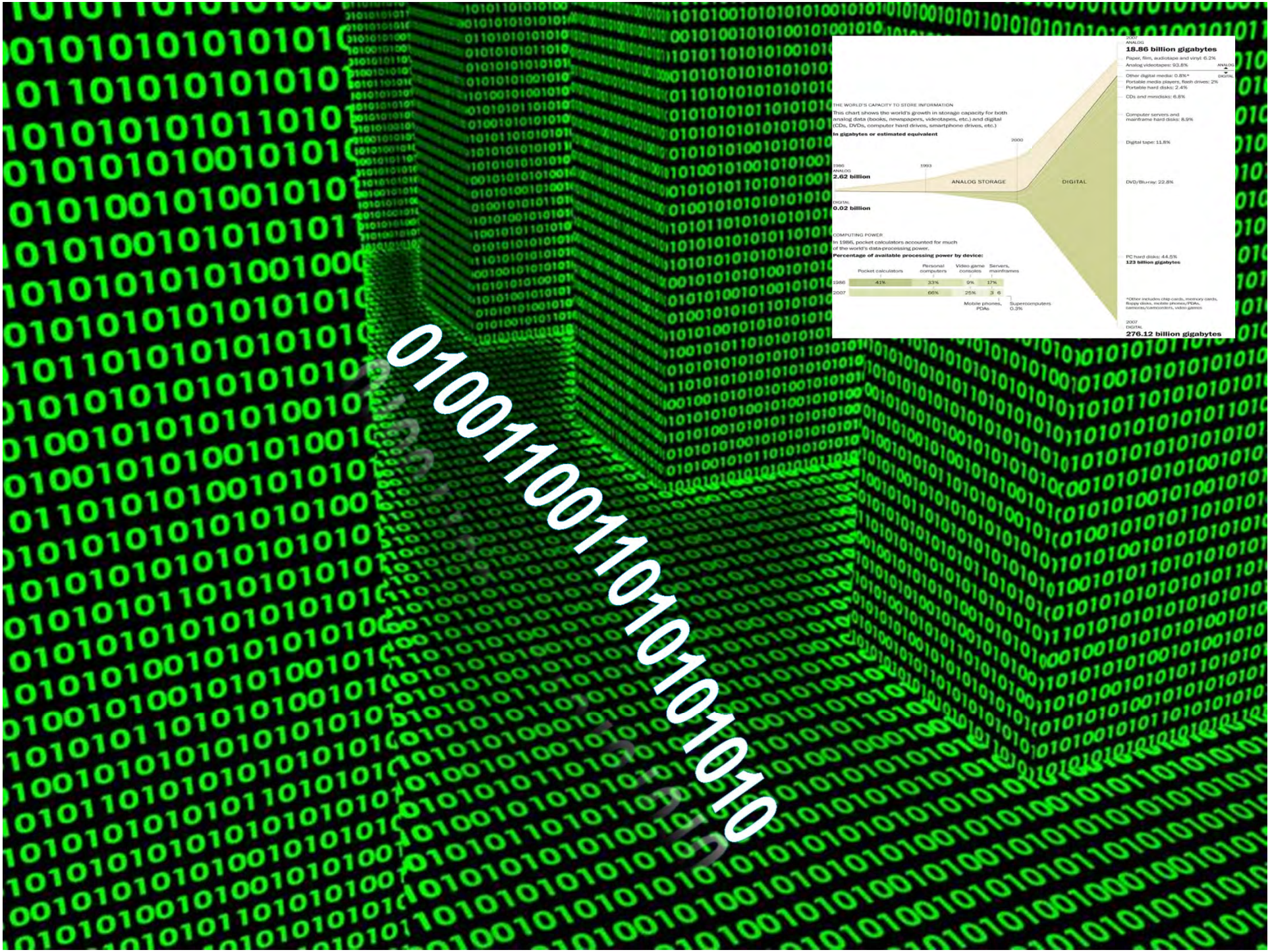
N_4 : *fitness* = +, *precision* = +, *generalization* = -, *simplicity* = -

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

Our A* algorithm exploits the Petri net marking equation and uses other “tricks” to prune the search space.

Aligned event log is starting point for other types of analysis.

Distributing/Decomposing “Big Data” Process Mining Problems

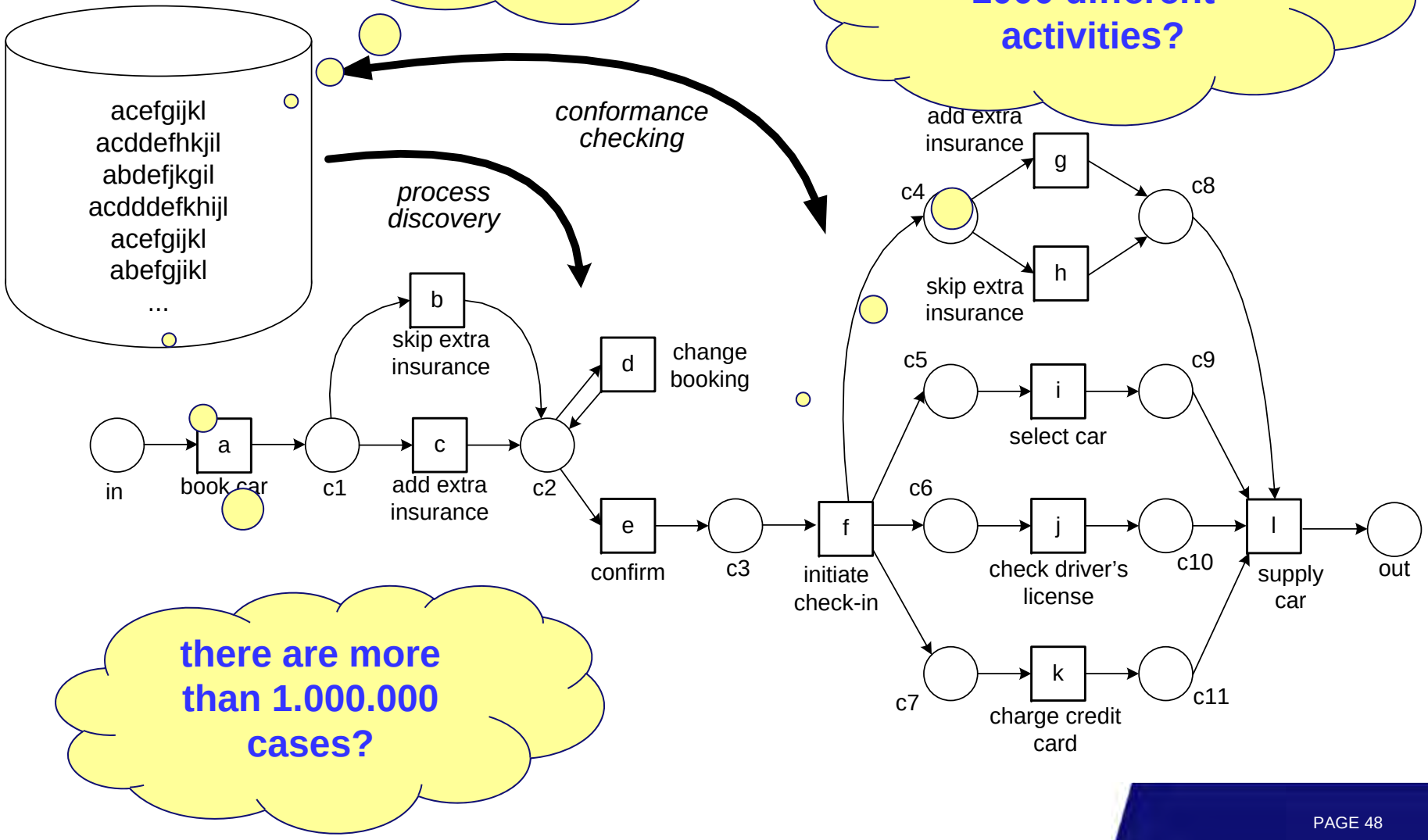

THE WORLD'S CAPACITY TO STORE INFORMATION
This chart shows the world's growth in storage capacity for both analog data (books, newspapers, videotapes, etc.) and digital (CDs, DVDs, computer hard drives, smartphone drives, etc.)
In gigabytes or estimated equivalent
1986
ANALOG
2.62 billion
DIGITAL
0.02 billion
1993
2000
ANALOG STORAGE
DIGITAL
2007
ANALOG
18.86 billion gigabytes
Paper, film, audiotape and vinyl: 6.2%
Analog videotapes: 93.8%
ANALOG
DIGITAL
Other digital media: 0.8%*
Portable media players, flash drives: 2%
Portable hard disks: 2.4%
CDs and minidisks: 6.8%
Computer servers and mainframe hard disks: 8.9%
Digital tape: 11.8%
DVD/Blu-ray: 22.8%
PC hard disks: 44.5%
123 billion gigabytes
*Other includes chip cards, memory cards, floppy disks, mobile phones/PDAs, cameras/camcorders, video games
2007
DIGITAL
276.12 billion gigabytes
COMPUTING POWER
In 1986, pocket calculators accounted for much of the world's data-processing power.
Percentage of available processing power by device:
Pocket calculators
Personal computers
Video game consoles
Servers, mainframes
1986
41%
33%
9%
17%
2007
66%
25%
3
6
Mobile phones, PDAs
Supercomputers 0.3%

What if?

there are more than 100.000.000 events?

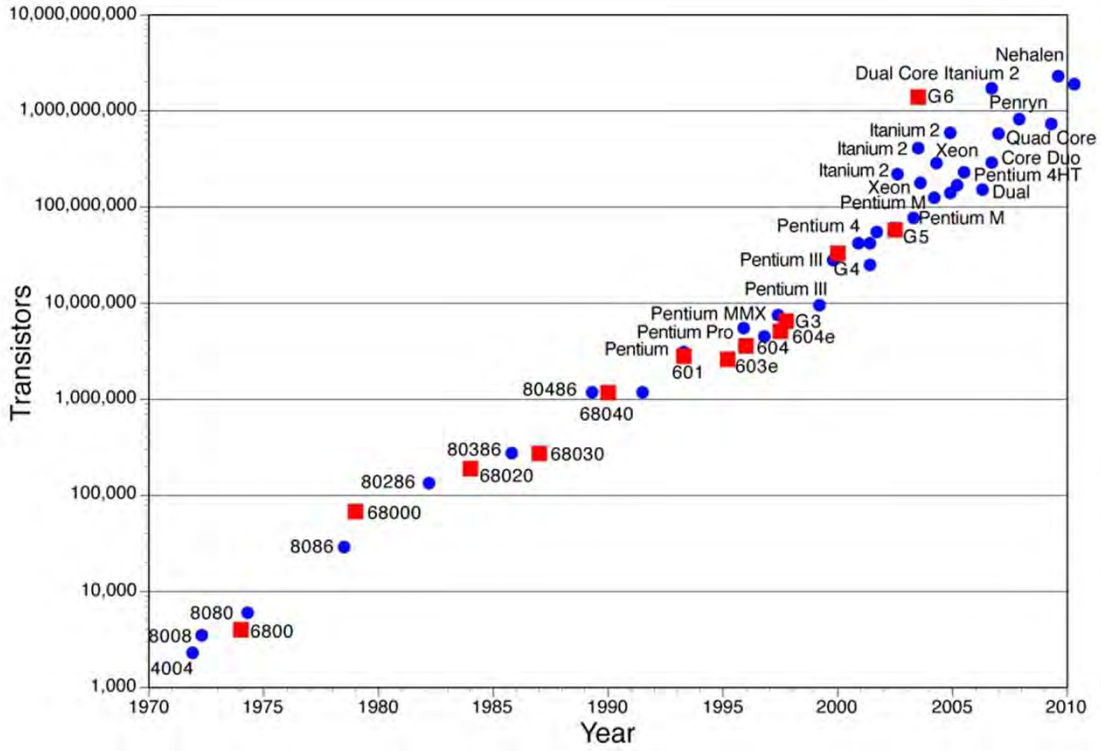
there are more than 1000 different activities?

there are more than 1.000.000 cases?

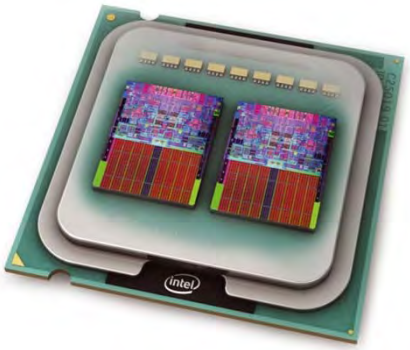


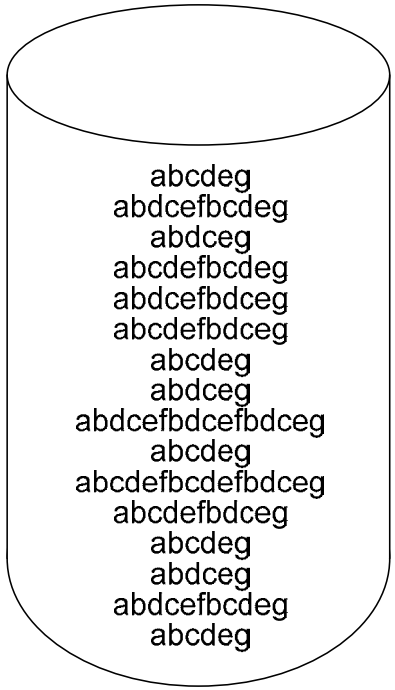
Distributed computing

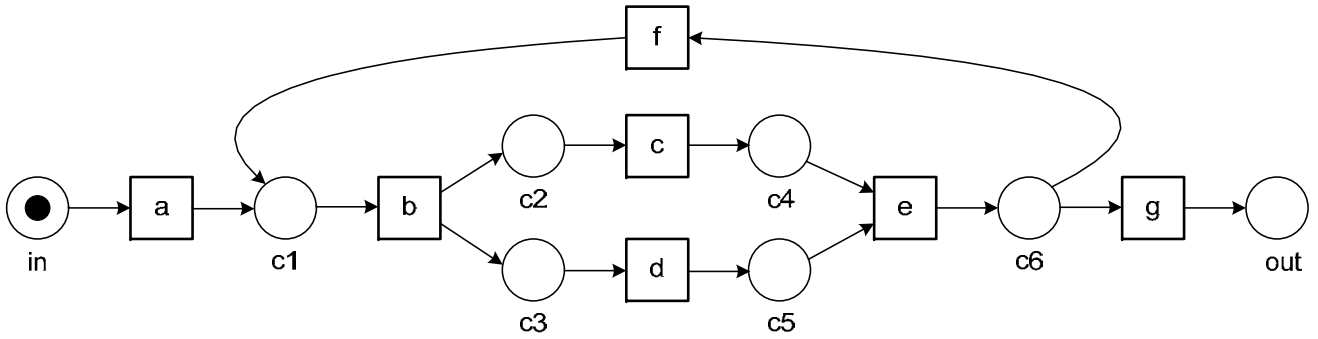
- multicore CPU
- manycore GPU
- cluster computing
- grid computing
- cloud computing
- …



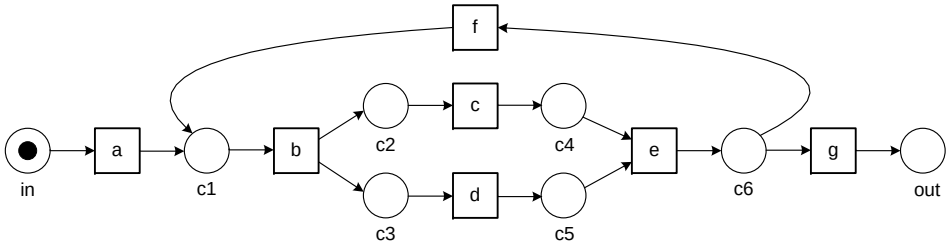
How to distribute process discovery?

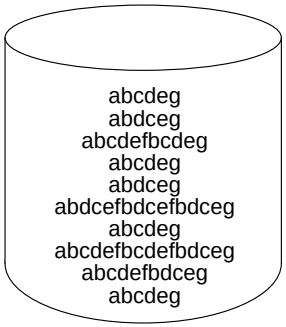


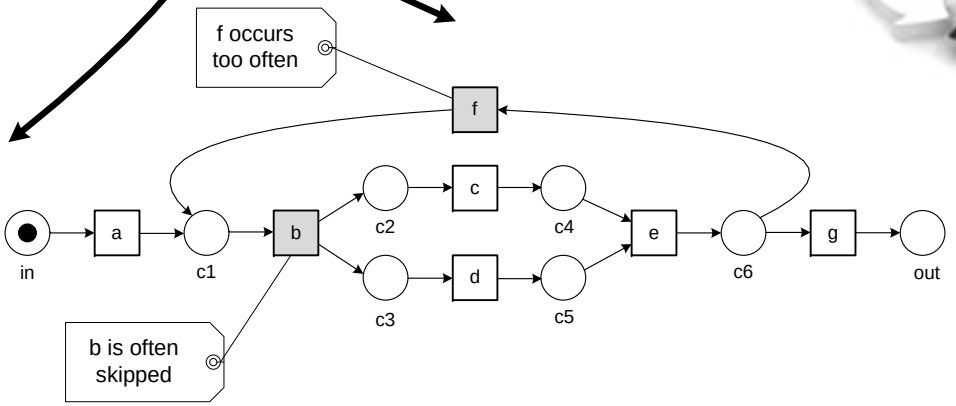




How to distribute conformance checking?





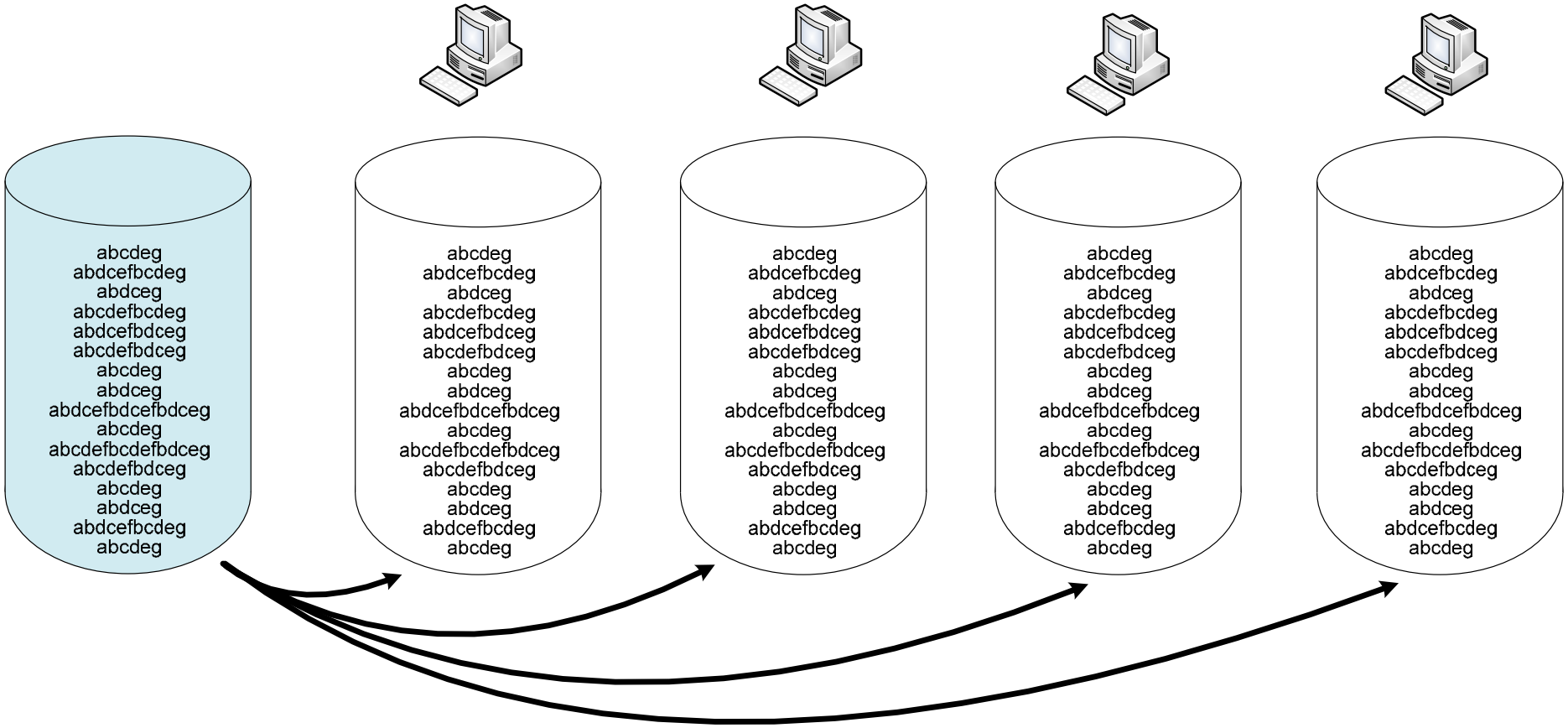


Classification based on partitioning of event log: vertical and horizontal

sets of cases

sets of activities

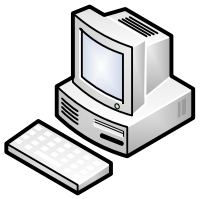
Replication: Same event log on all computing nodes

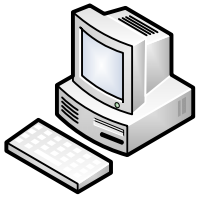


Only makes sense if random elements, e.g., genetic process mining.

Vertical distribution I: Split cases arbitrarily

sets of cases



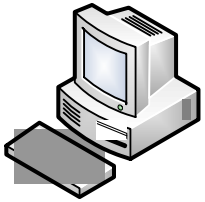


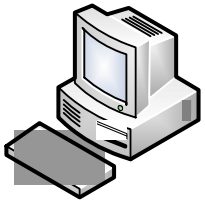
abcdeg
abdcefbcdeg
abdceg
abcdefbcdeg
abdcefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbdcefbdceg
abcdeg
abcdefbcdefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbcdeg
abcdeg

abcdeg
abdcefbcdeg
abdceg
abcdefbcdeg
abdcefbdceg
abcdefbdceg
abcdeg
abdceg

abdcefbdcefbdceg
abcdeg
abcdefbcdefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbcdeg
abcdeg

Vertical distribution II: Split cases based on a specific feature





abcdeg
abdcefbcdeg
abdceg
abcdefbcdeg
abdcefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbdcefbdceg
abcdeg
abcdefbcdefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbcdeg
abcdeg

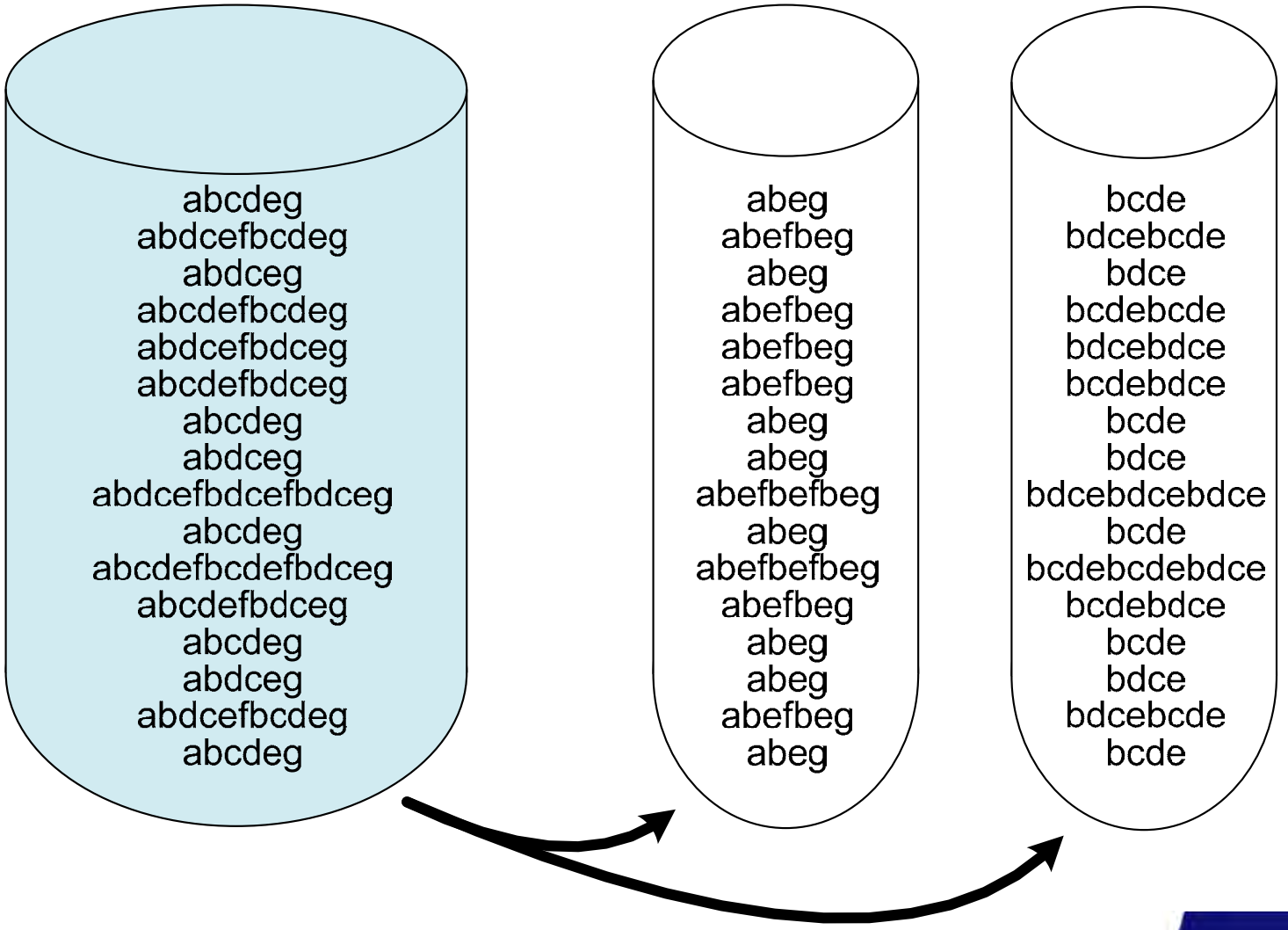
abcdeg
abdceg
abcdeg
abdceg
abcdeg
abcdeg
abdceg
abcdeg

abdcefbcdeg
abcdefbcdeg
abdcefbdceg
abcdefbdceg
abcdefbdceg
abdcefbcdeg

abdcefbdcefbdceg
abcdefbcdefbdceg

Horizontal distribution

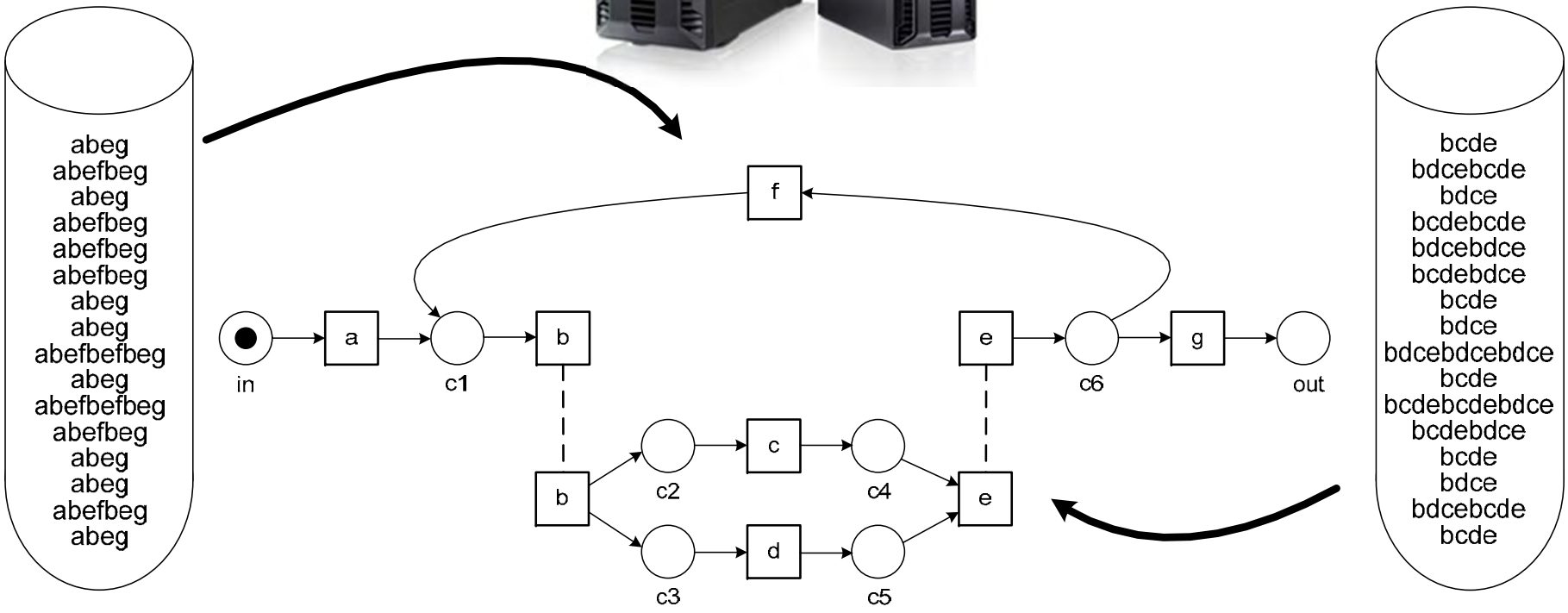
sets of activities



Horizontal distribution: The key idea

projected on a,b,e,f,g

projected on b,c,d,e



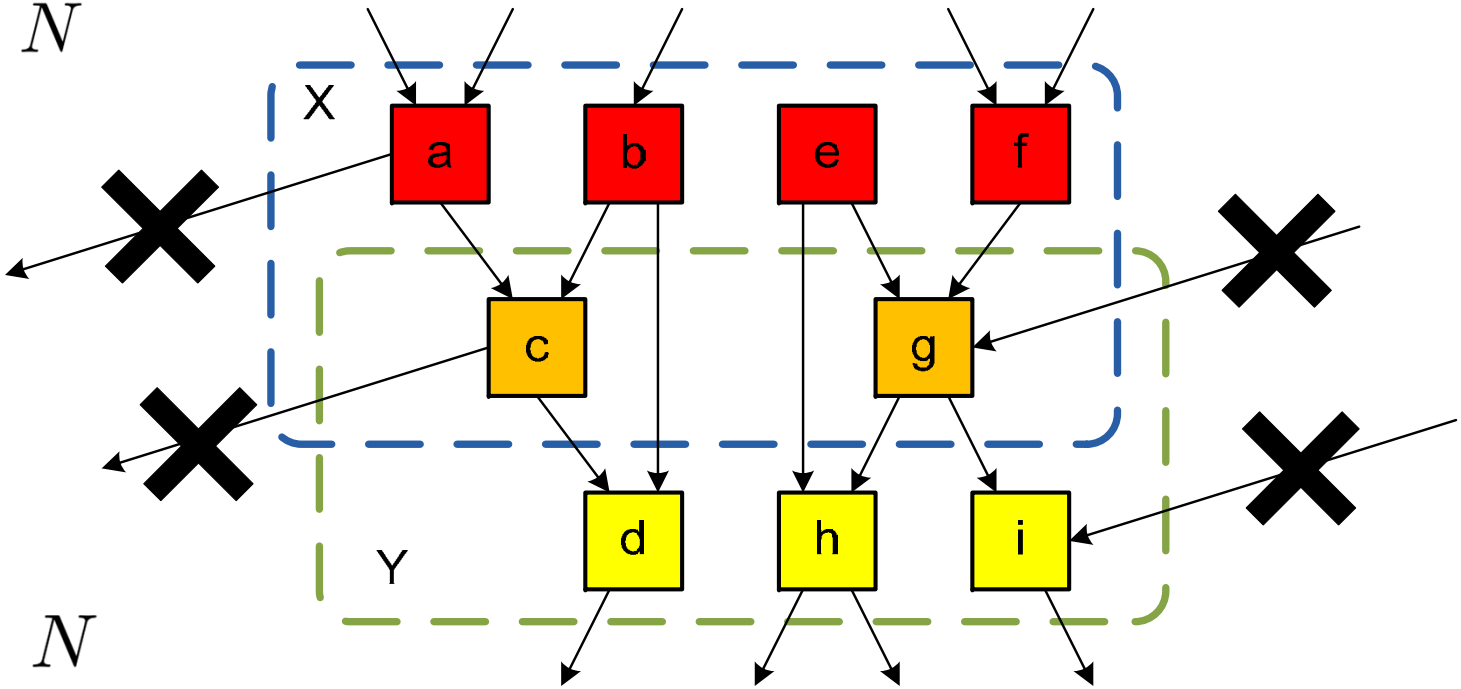
Passages for Horizontal Distribution

Passages

Passage *P=(X,Y)*

causal dependency:
may trigger or enable

$$\emptyset \neq X \subseteq N$$

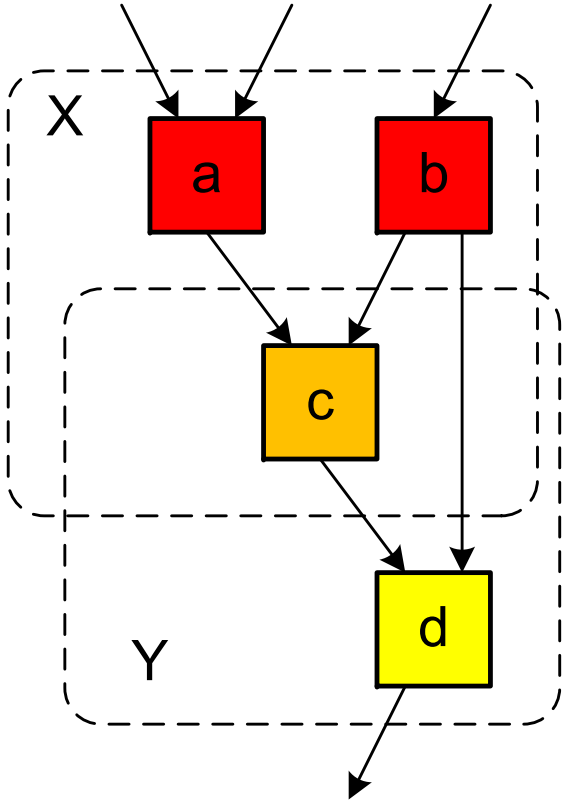


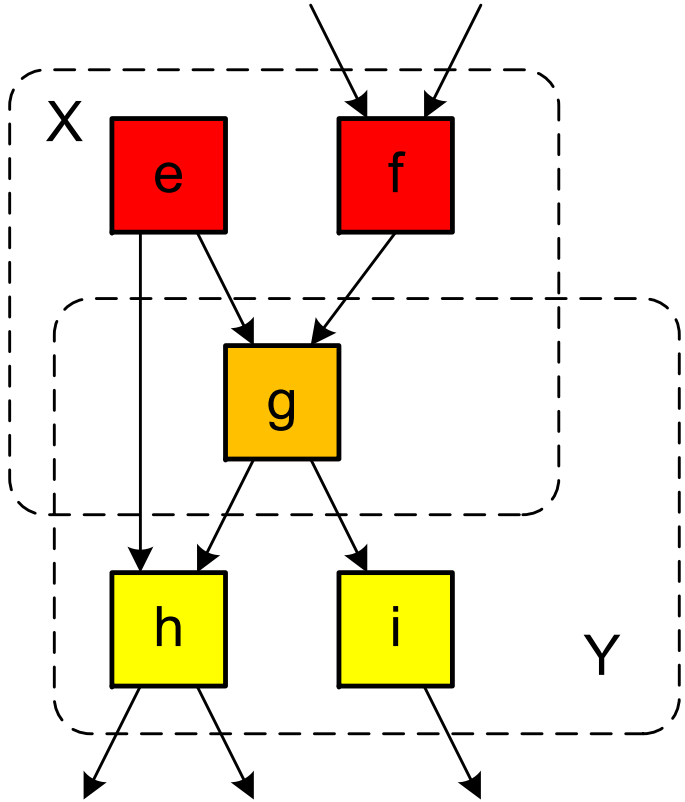
$$\emptyset \neq Y \subseteq N$$

$$X \overset{G}{\bullet} = Y$$

$$X = \overset{G}{\bullet} Y$$

Minimal passages



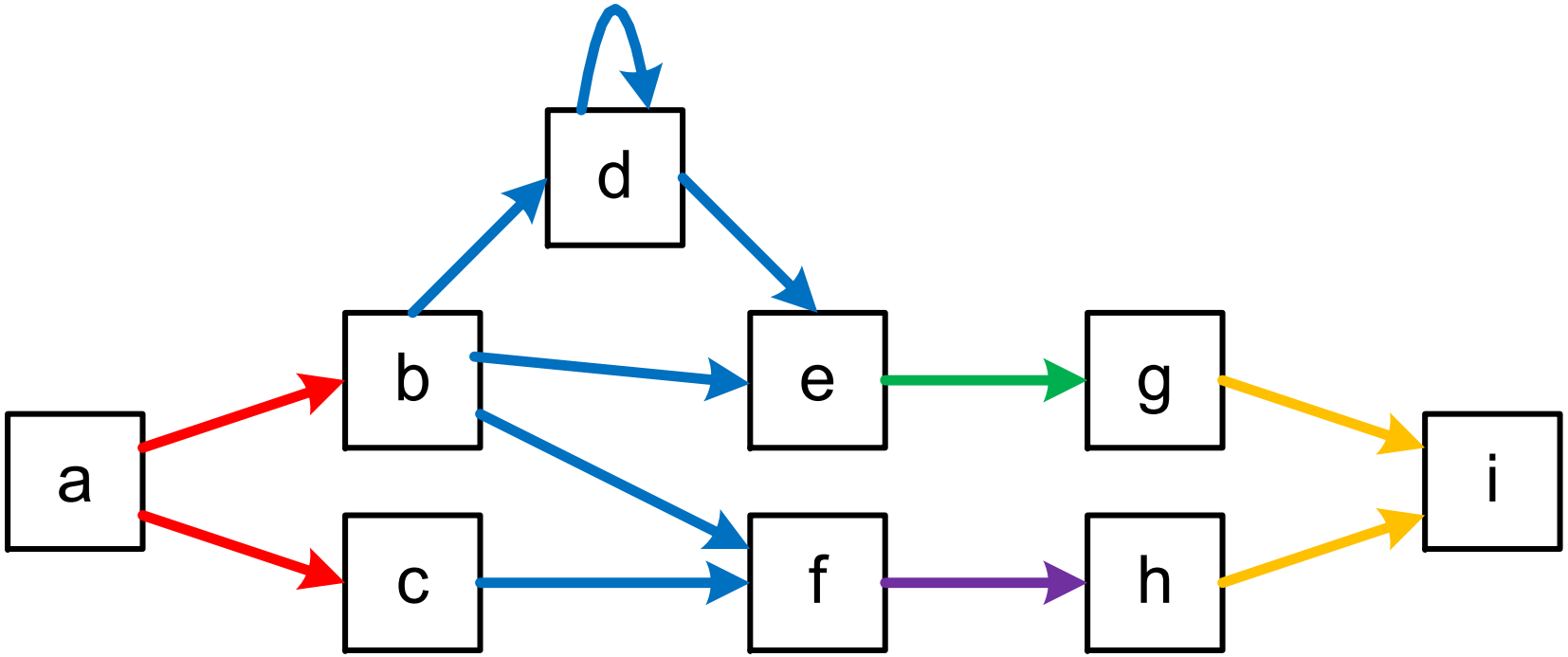


$$X \overset{G}{\bullet} = Y$$

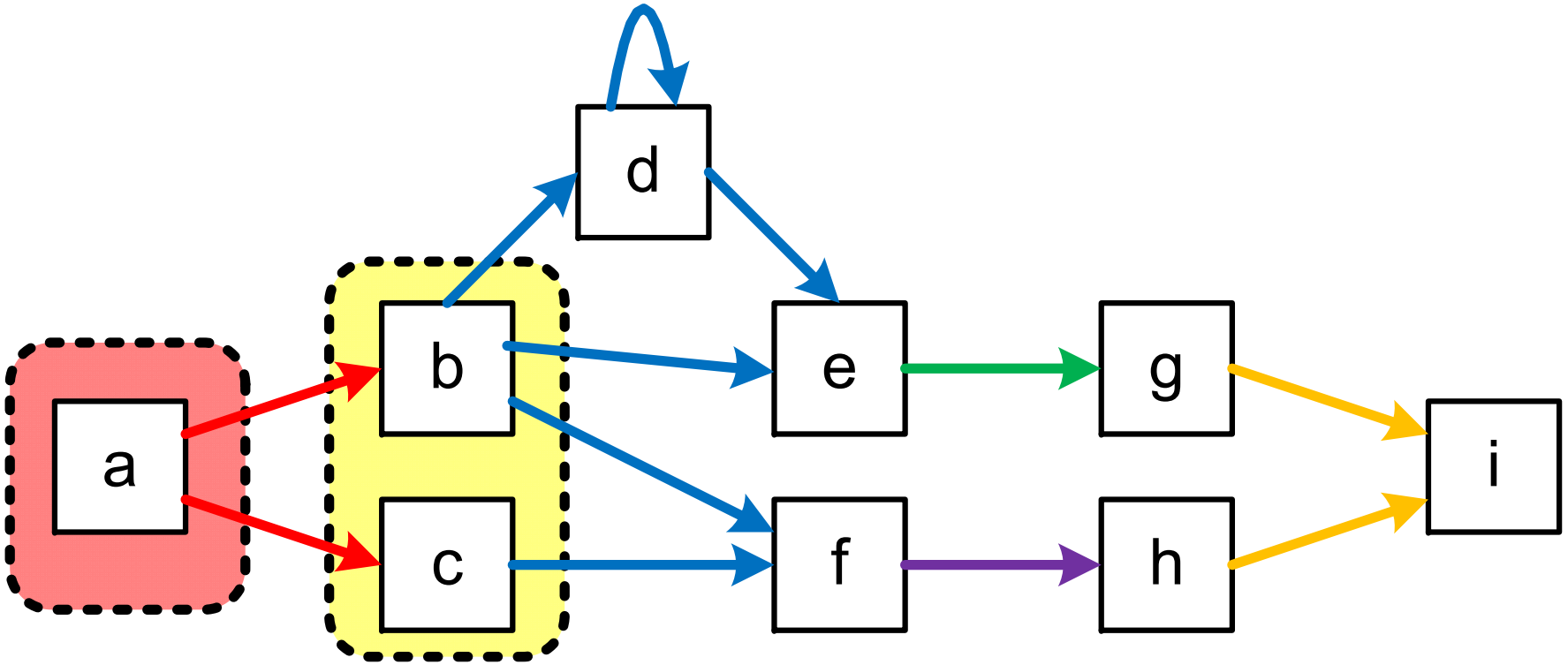
$$X = \overset{G}{\bullet} Y$$

a passage is minimal if it does not contain smaller passages

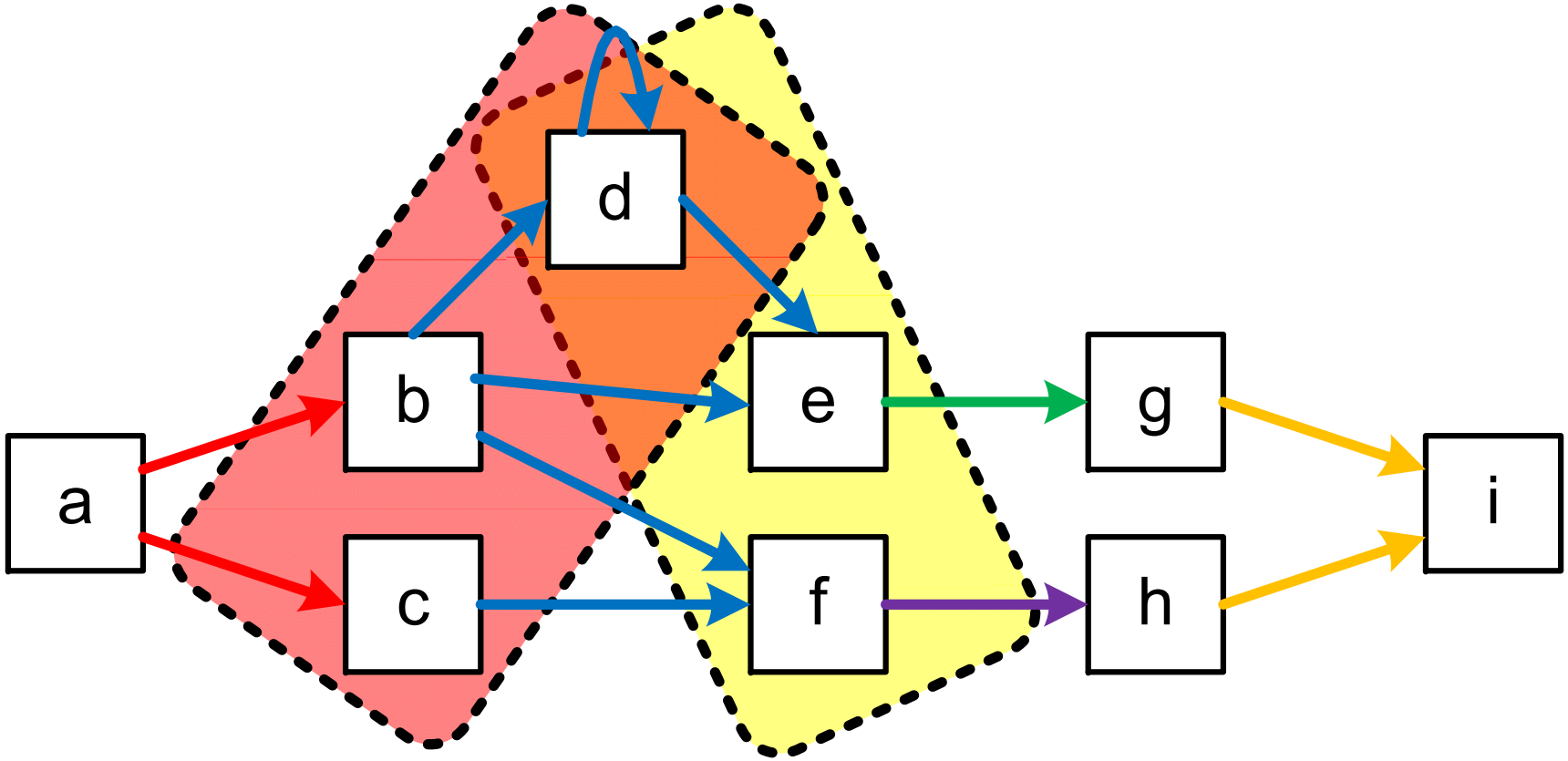
Passages define an equivalence relation on the edges in the graph



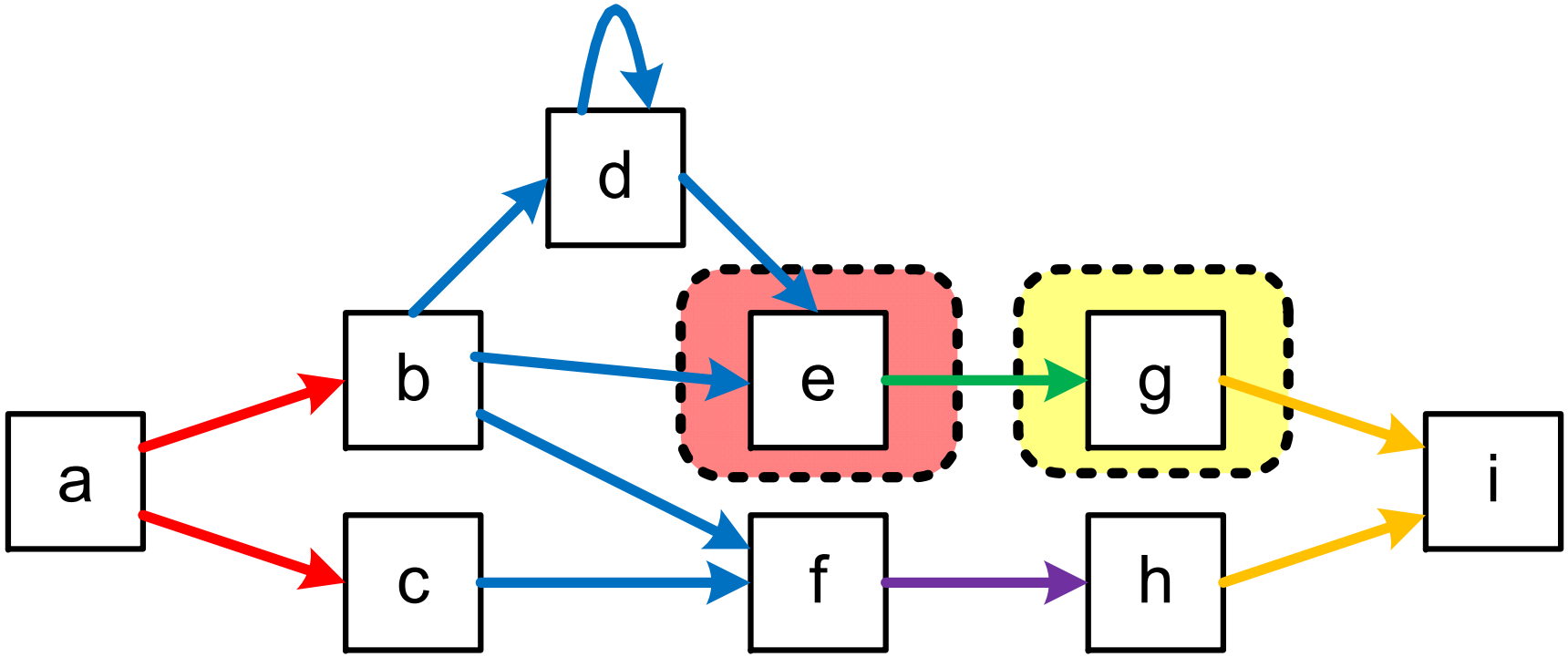
Minimal passage 1: ({a},{b,c})



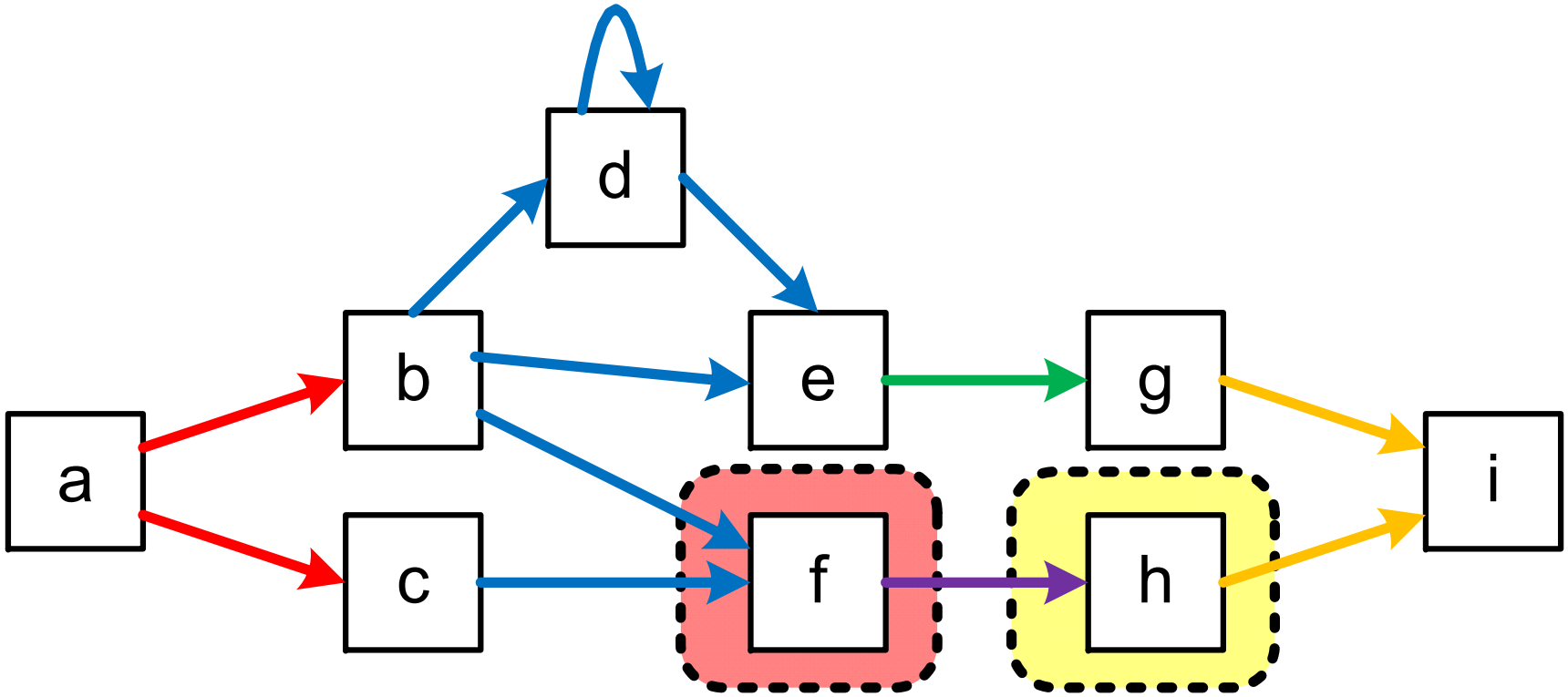
Minimal passage 2: ({b,c,d},{d,e,f})



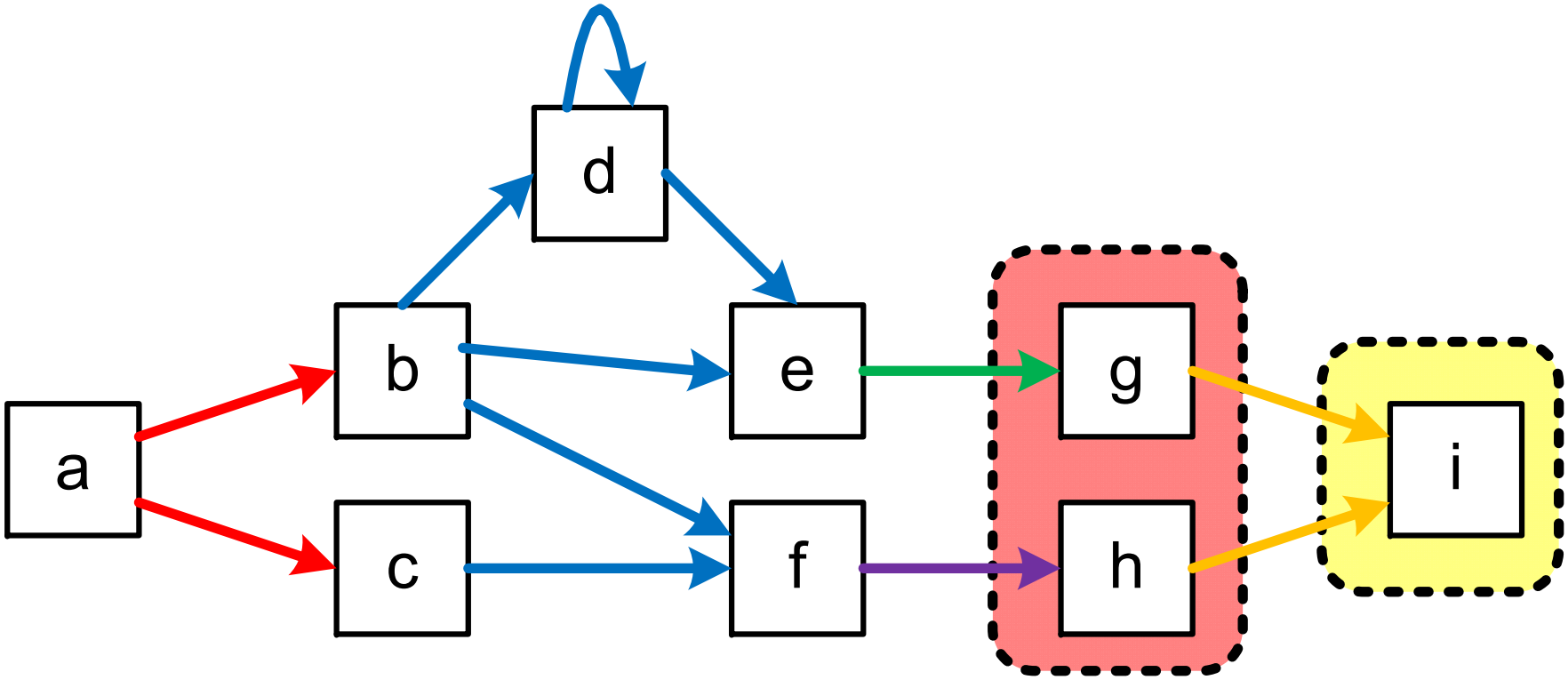
Minimal passage 3: ({e},{g})



Minimal passage 4: ({f},{h})

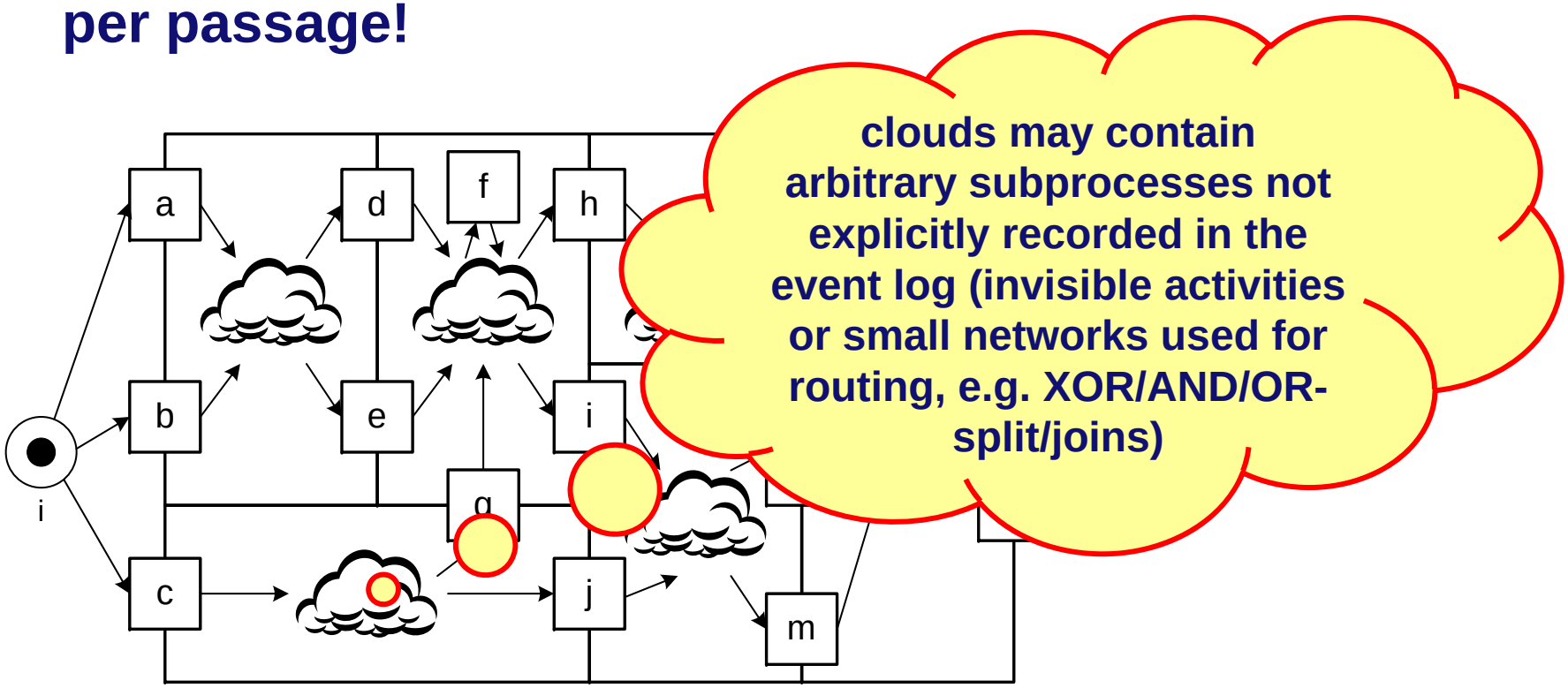


Minimal passage 5: ({g,h},{i})



So What?

- **Any process model can be partitioned in minimal passages.**
- **Discovery and conformance checking can be done per passage!**

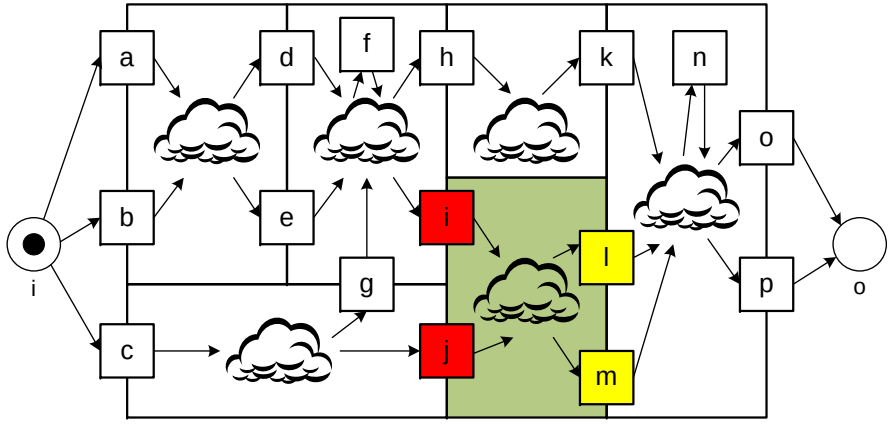


Example result for Petri nets

Theorem 1 (Main Theorem). *Let $L \in \mathcal{B}(A^*)$ be an event log and let $WF = (PN, in, T_i, out, T_o)$ be a WF-net with $PN = (P, T, F, T_v)$.*
L is perfectly fitting system net $SN = (PN, [in], [out])$ if and only if

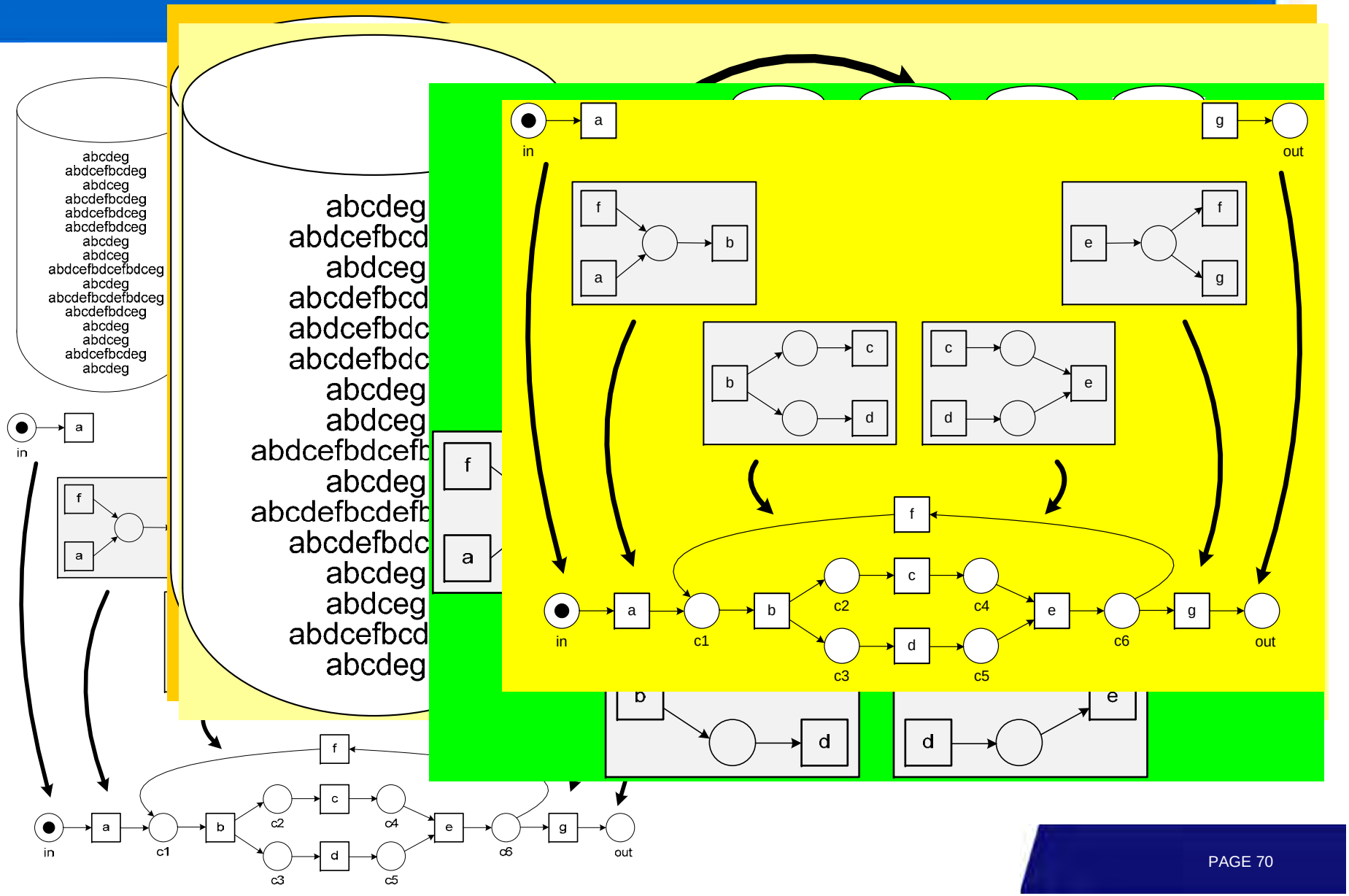
- *for any $\langle a_1, a_2, \ldots a_k \rangle \in L$: $a_1 \in T_i$ and $a_k \in T_o$, and*
- *for any $(X, Y) \in pas_{min}(skel(PN))$: $L \restriction_{X \cup Y}$ is perfectly fitting $SN^{(X,Y)} = (PN^{(X,Y)}, [\,], [\,])$.*

"The event log fits all passages if and only if the event log fits the whole model."

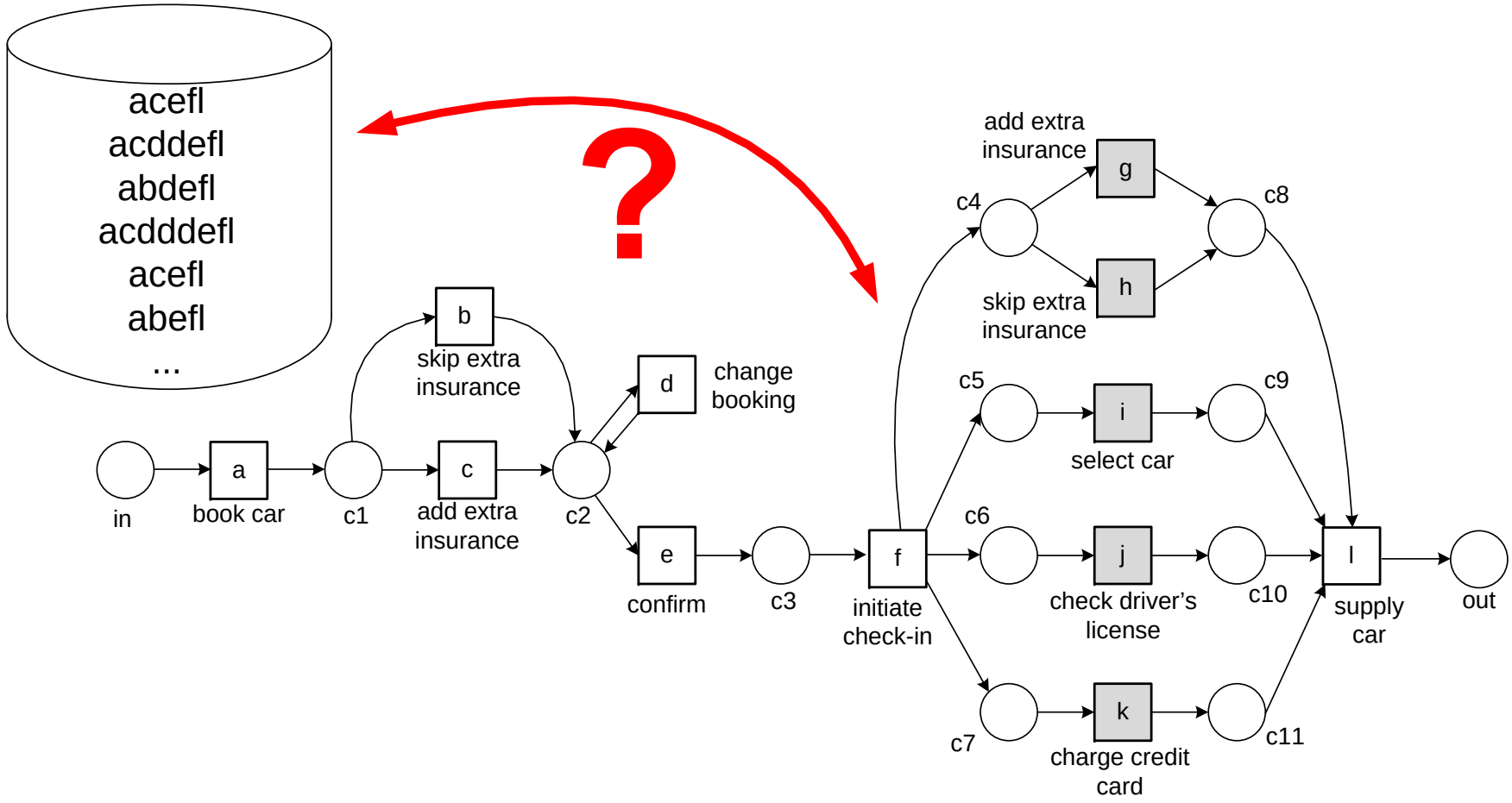


Key insight: interface transitions controlled by event log

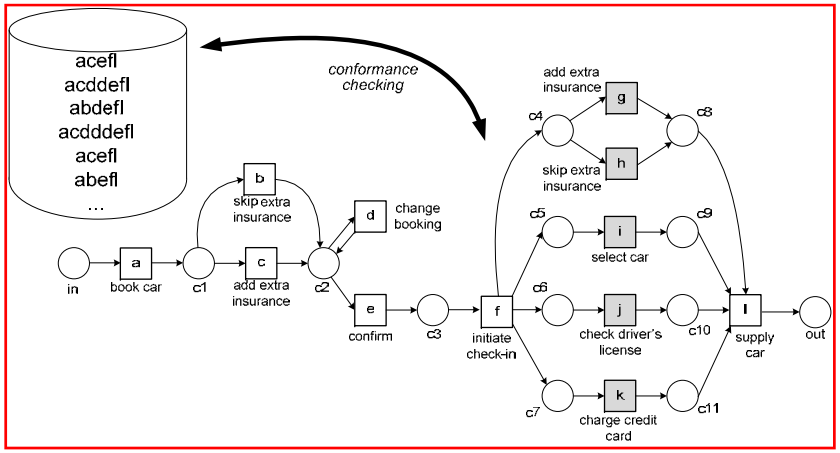
Discovery example

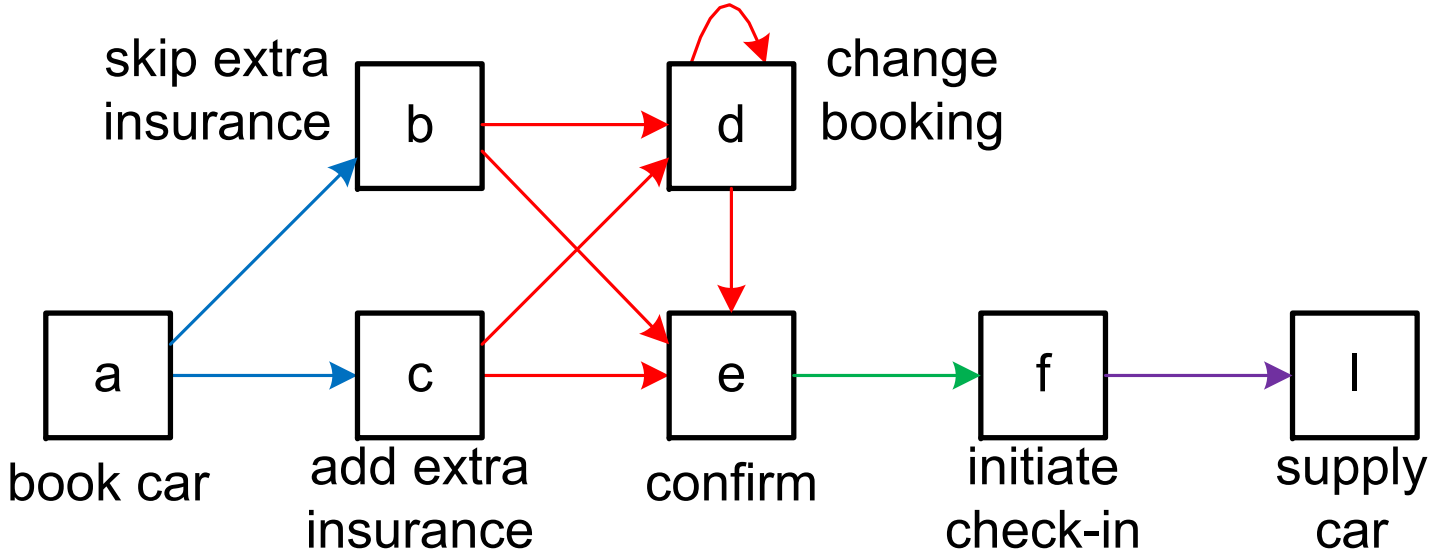


Conformance checking

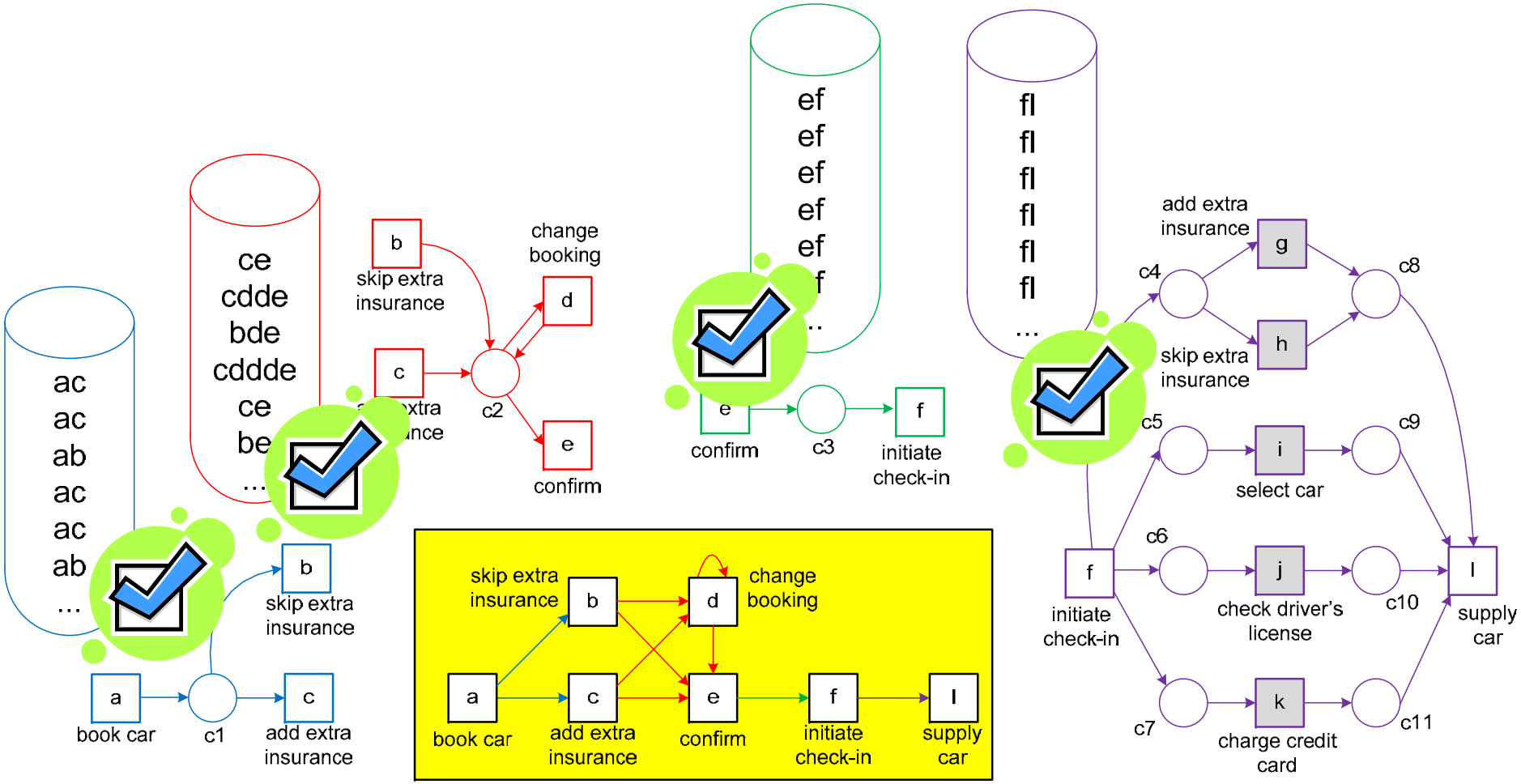


Create Skeleton

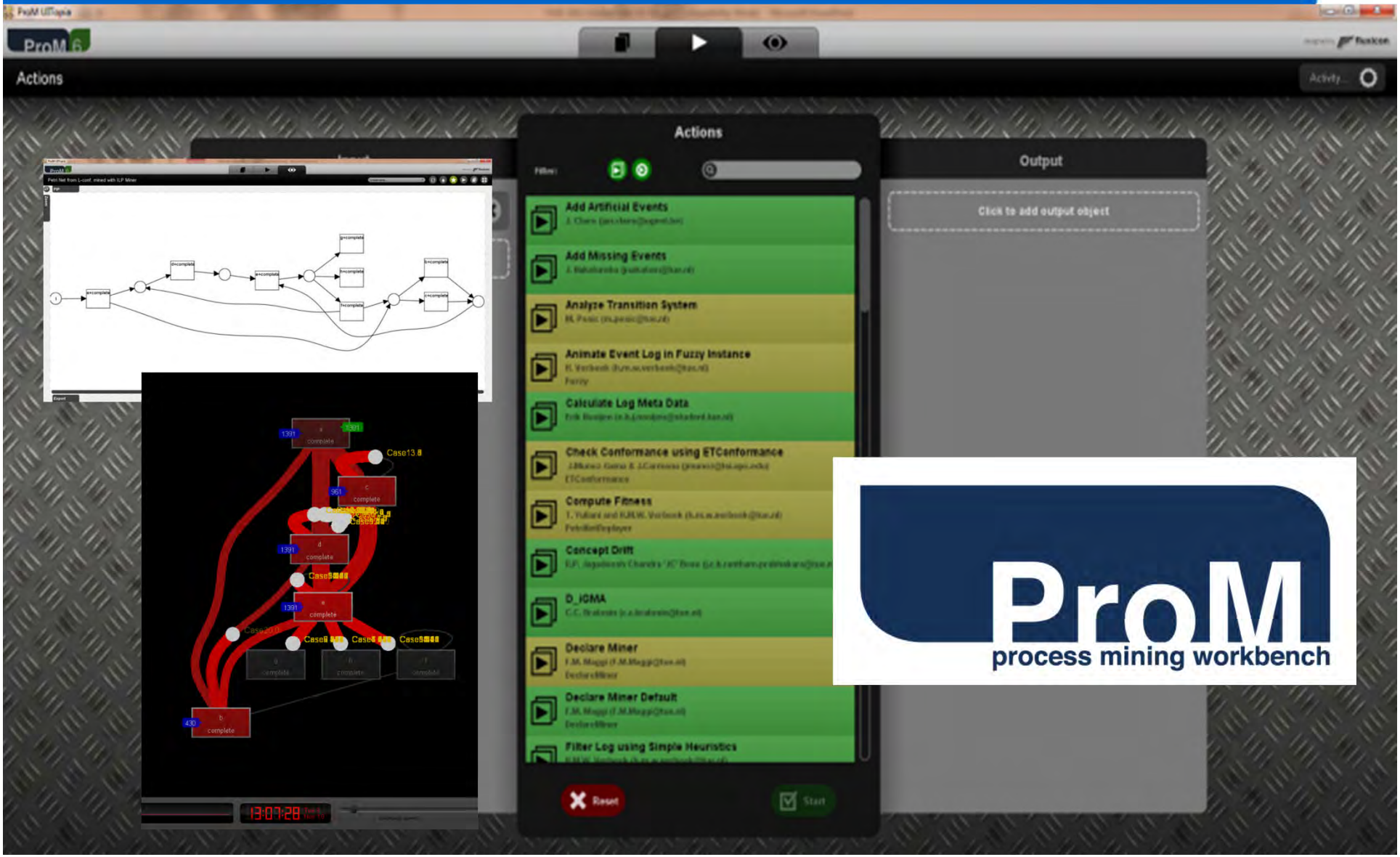




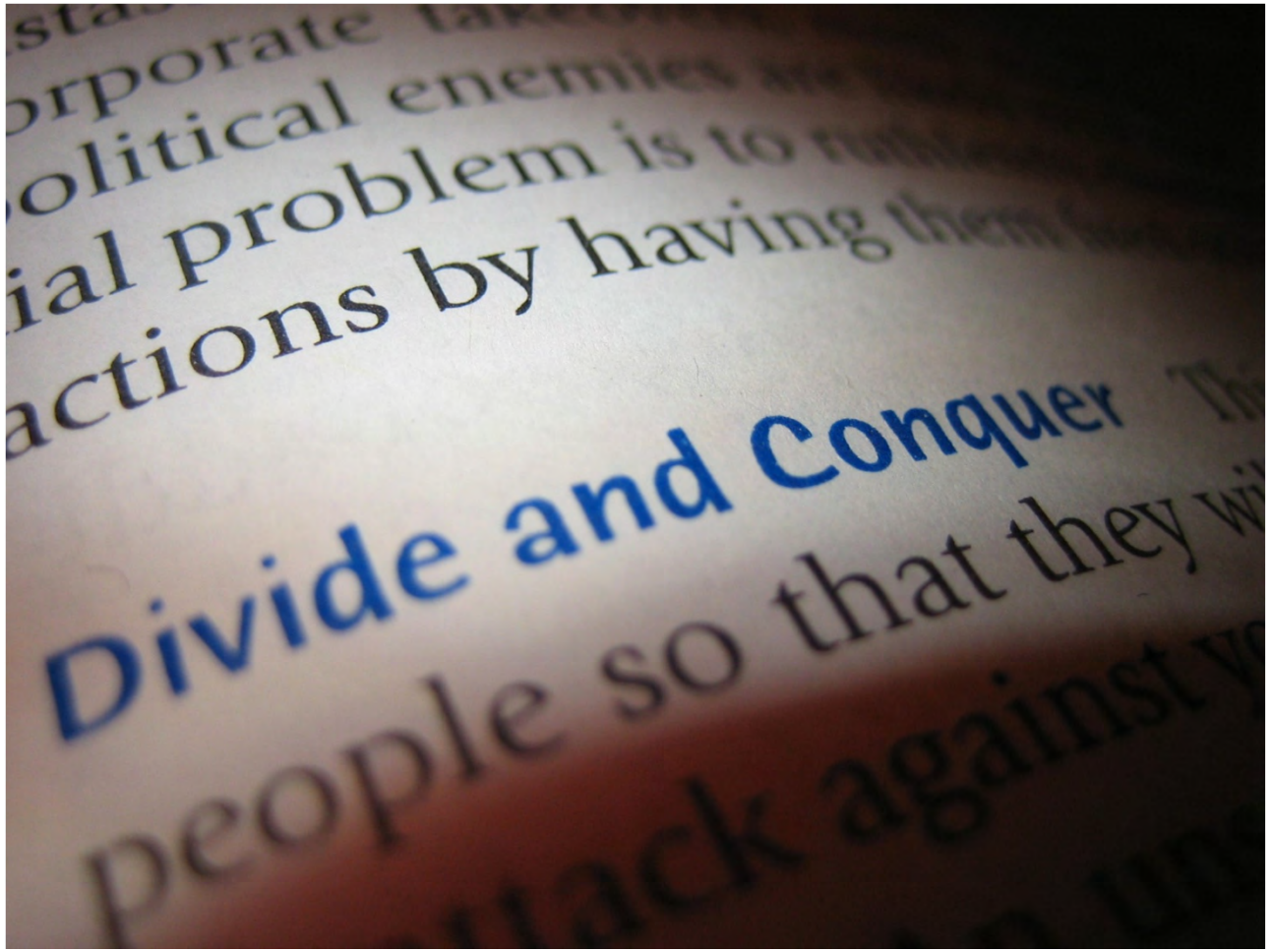
Net fragments per passage



Initial implementation in ProM

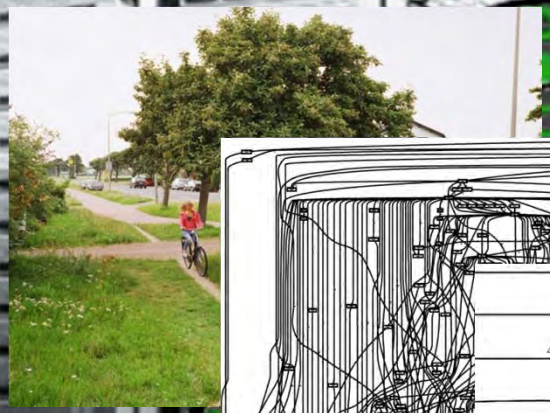


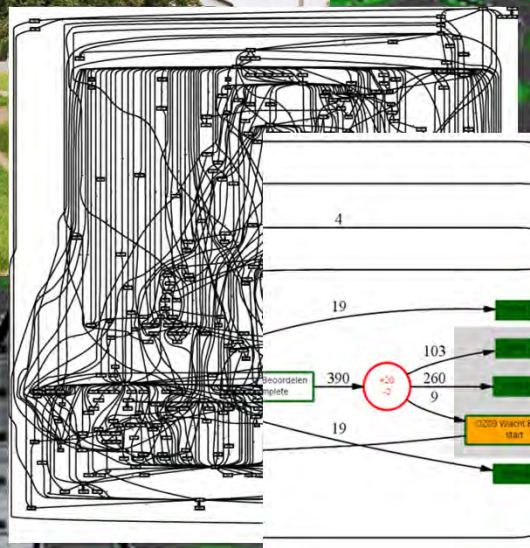
Super linear speedups possible (even when using a single computer decomposition helps)

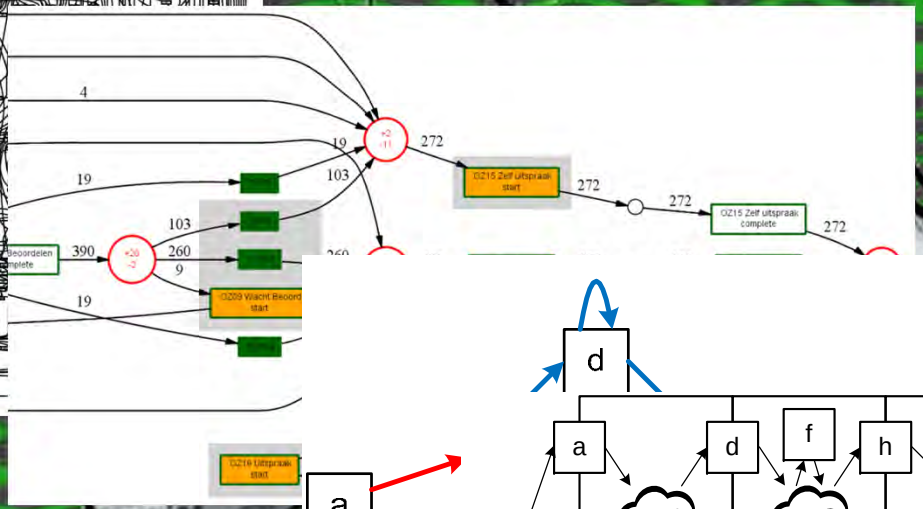


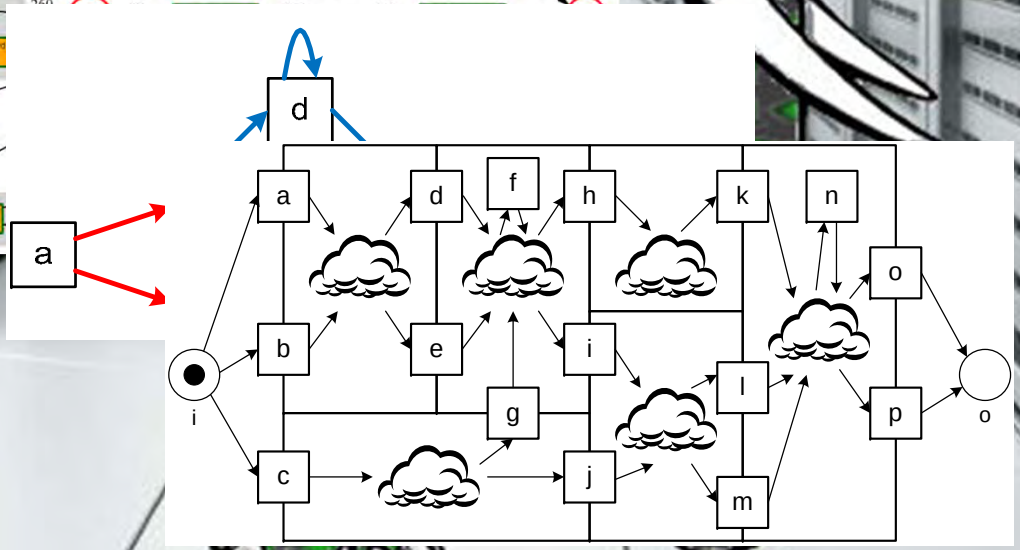
Conclusion

Conclusion









"Big Data"

Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

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

Wil van der Aalst delivers the first book on process mining. It aims to be self-contained while covering the entire process mining spectrum from process discovery to operational support. In Part I, the author provides the basics of business process modeling and data mining necessary to understand the remainder of the book. Part II focuses on process discovery as the most important process mining task. Part III moves beyond discovering the control flow of processes and highlights conformance checking, and organizational and time perspectives. Part IV guides the reader in successfully applying process mining in practice, including an introduction to the widely used open-source tool ProM. Finally, Part V takes a step back, reflecting on the material presented and the key open challenges.

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

Features and Benefits:

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

Computer Science

ISBN 978-3-642-19344-6

springer.com

van der Aalst

Process Mining

Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

www.processmining.org

www.win.tue.nl/ieeetfpm/

Springer