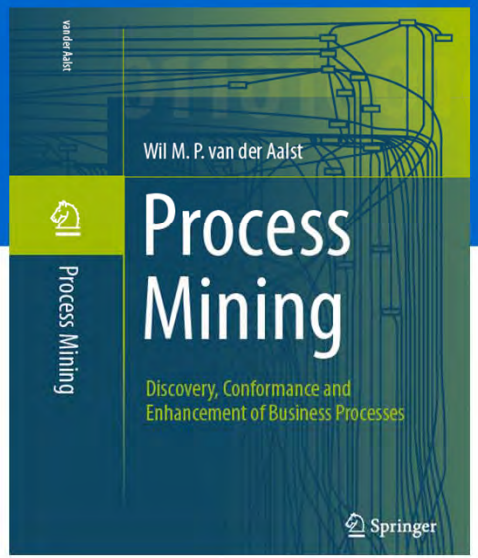
Process Mining: Auditing Based on Facts Rather than Fiction

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

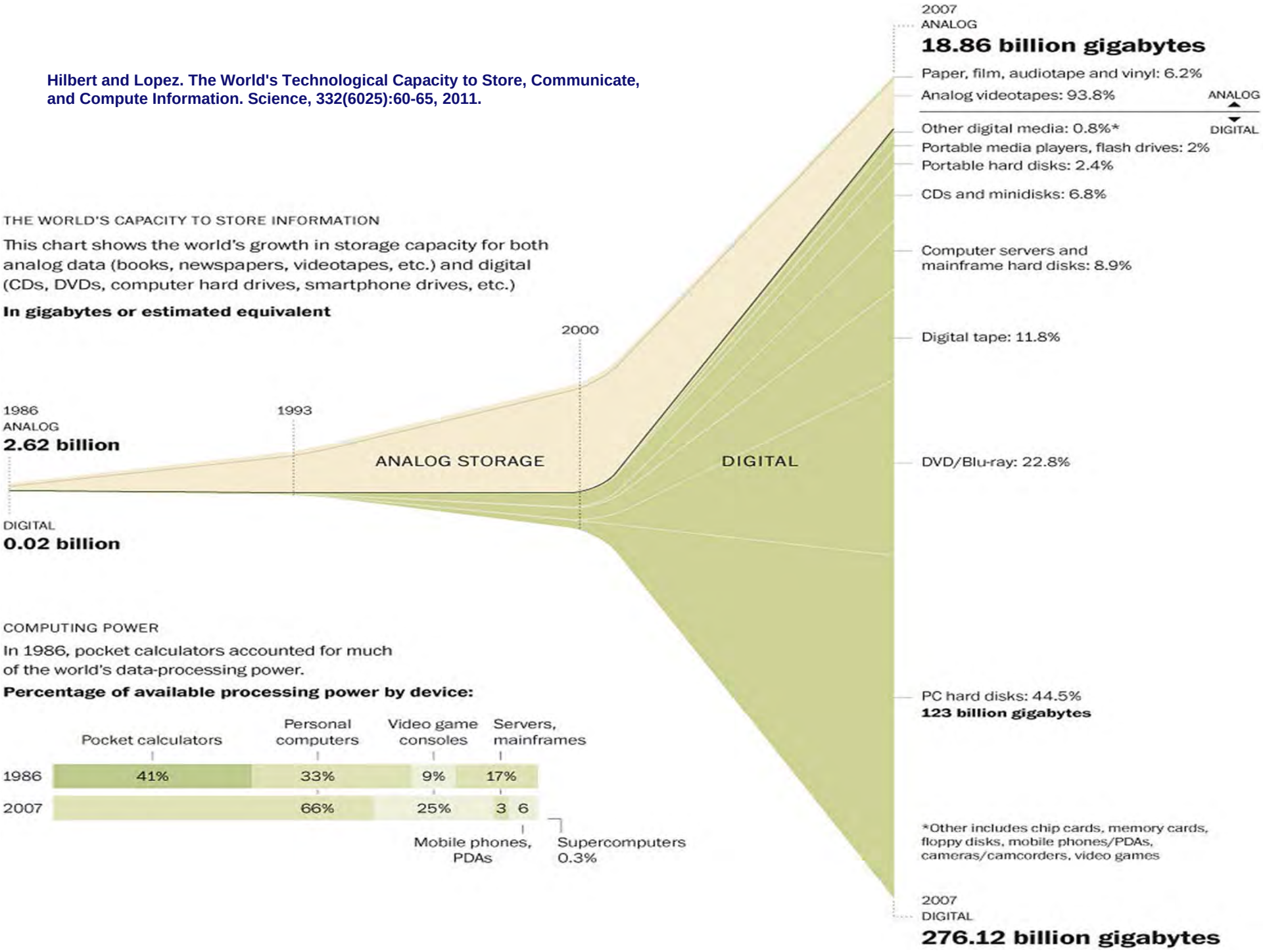


TU/e Technische Universiteit **Eindhoven** University of Technology

Where innovation starts

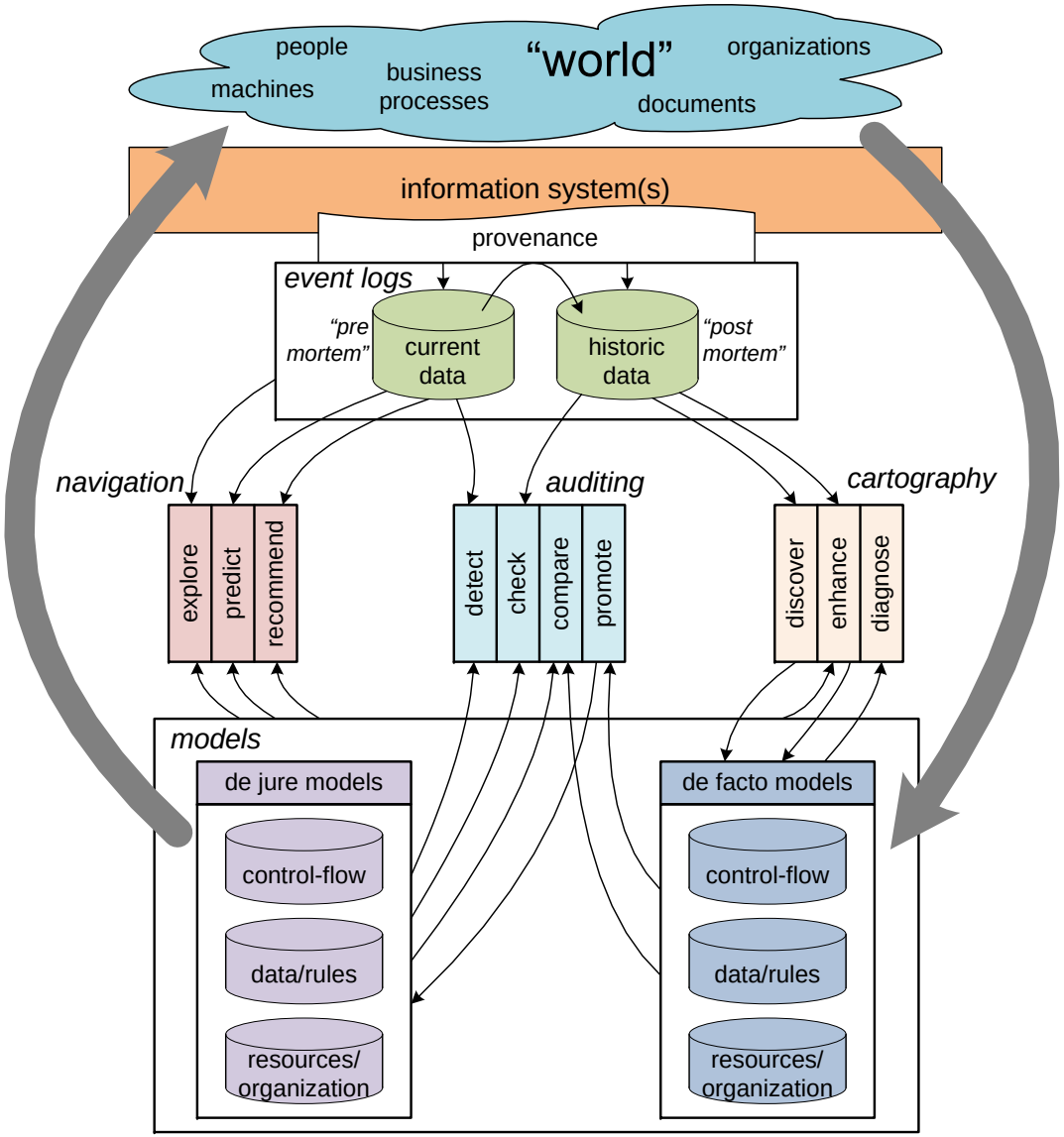

Big Data
"All of the world's music can be stored on a $600 disk drive."
"Enterprises globally stored more than 7 exabytes of new data on disk drives in 2010, while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks."
"Indeed, we are generating so much data today that 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.




Process Mining =
Event Data + Processes
Data Mining + Process Analysis

Process Mining: Overview



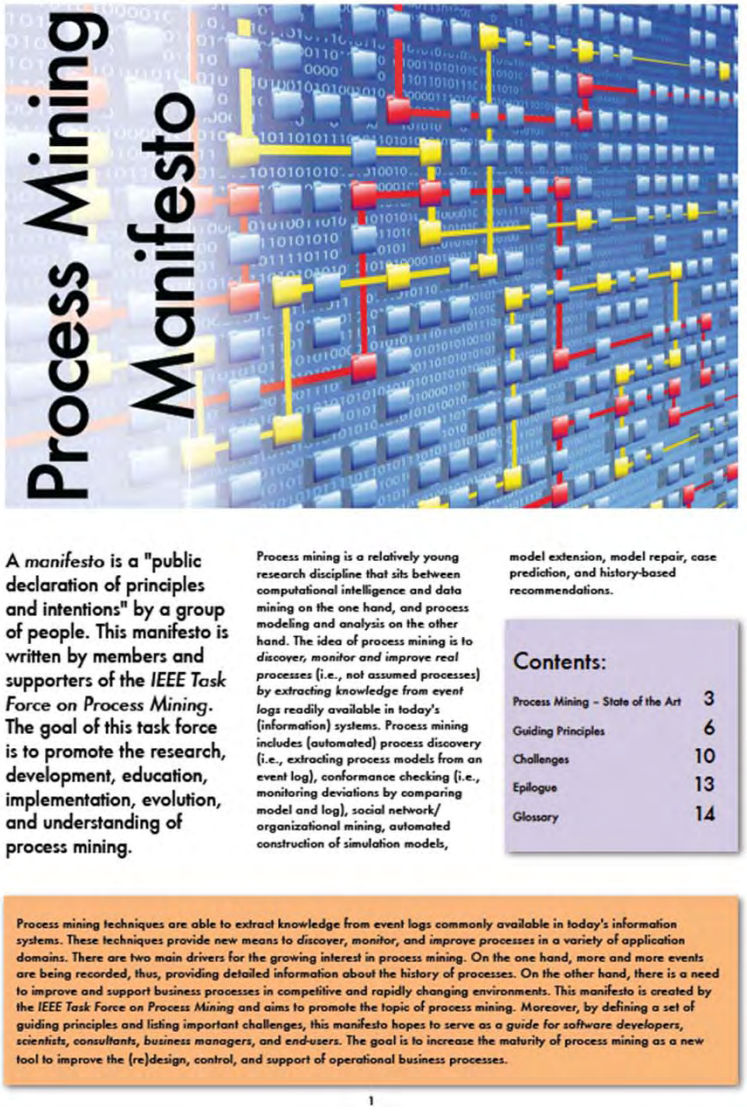
We applied ProM in >100 organizations

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

Hundreds to plug-ins available covering the whole spectrum

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

Process Mining Manifesto

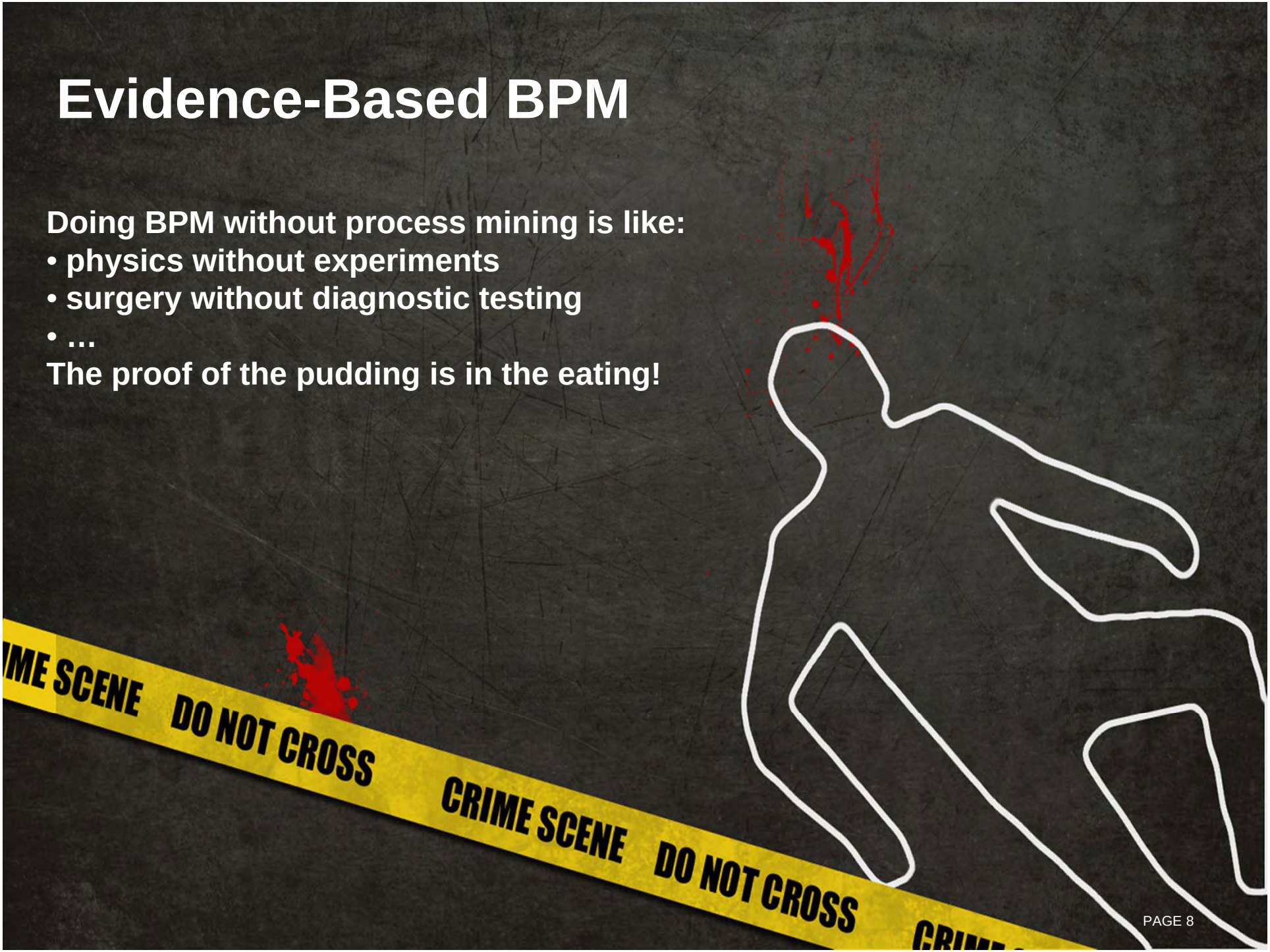


- **On 7 October 2011, the IEEE Task Force on Process Mining released the Process Mining Manifesto**
- **53 organizations support the manifesto**
- **77 process mining experts contributed to it**
- **Translated into Chinese, German, French, Spanish, Greek, Italian, Korean, Dutch, Portuguese, Turkish, and Japanese.**

Evidence-Based BPM

Doing BPM without process mining is like:
- **physics without experiments**
- **surgery without diagnostic testing**
- **…**

The proof of the pudding is in the eating!

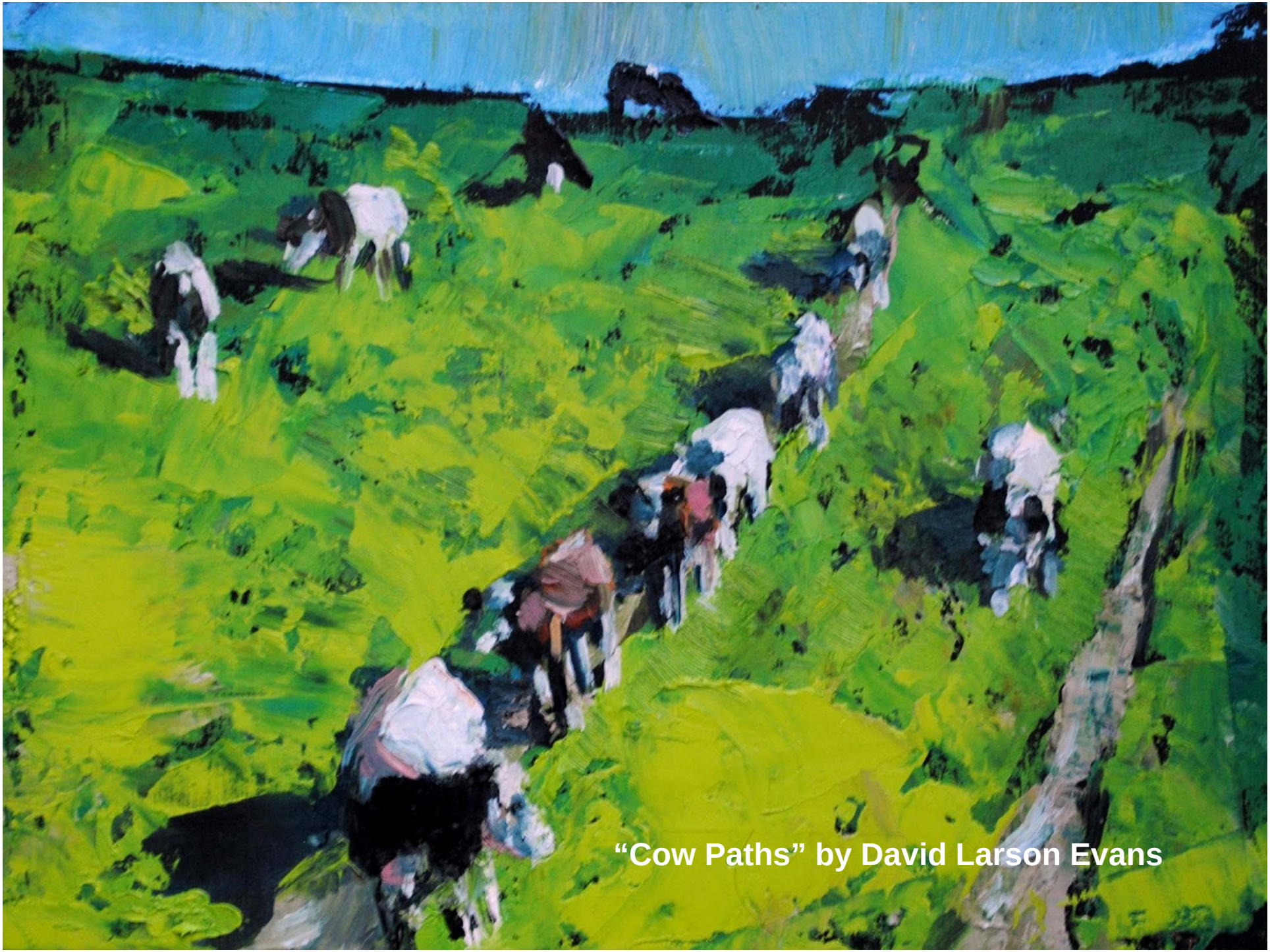


Desire Lines or Cow Paths?


www.olifantenpaadjes.nl
01001100110101010
0100110101010
60

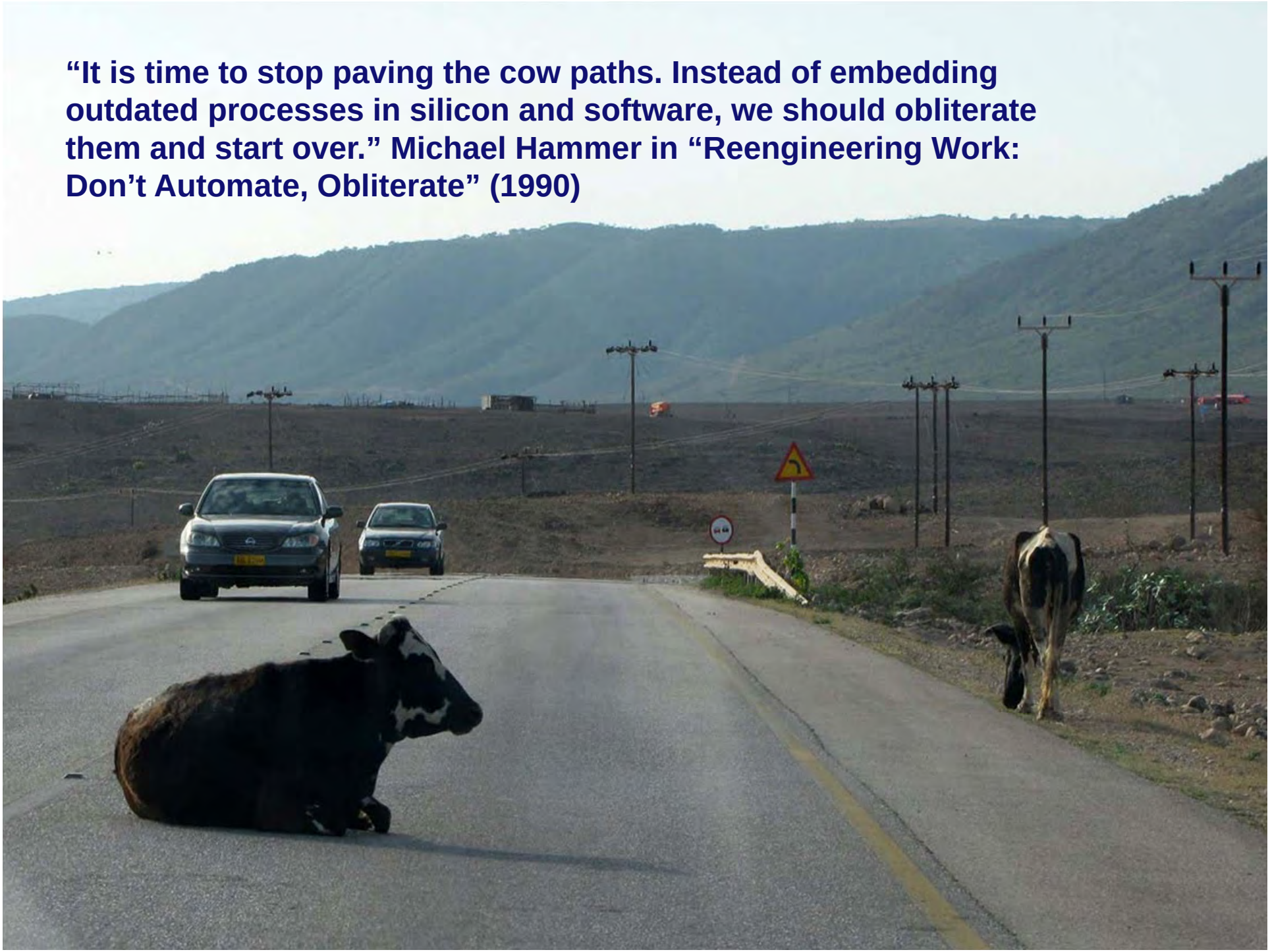

0100110101010101010


0100110011010101010


“Cow Paths” by David Larson Evans


BPR mantra: Don’t pave the cow paths

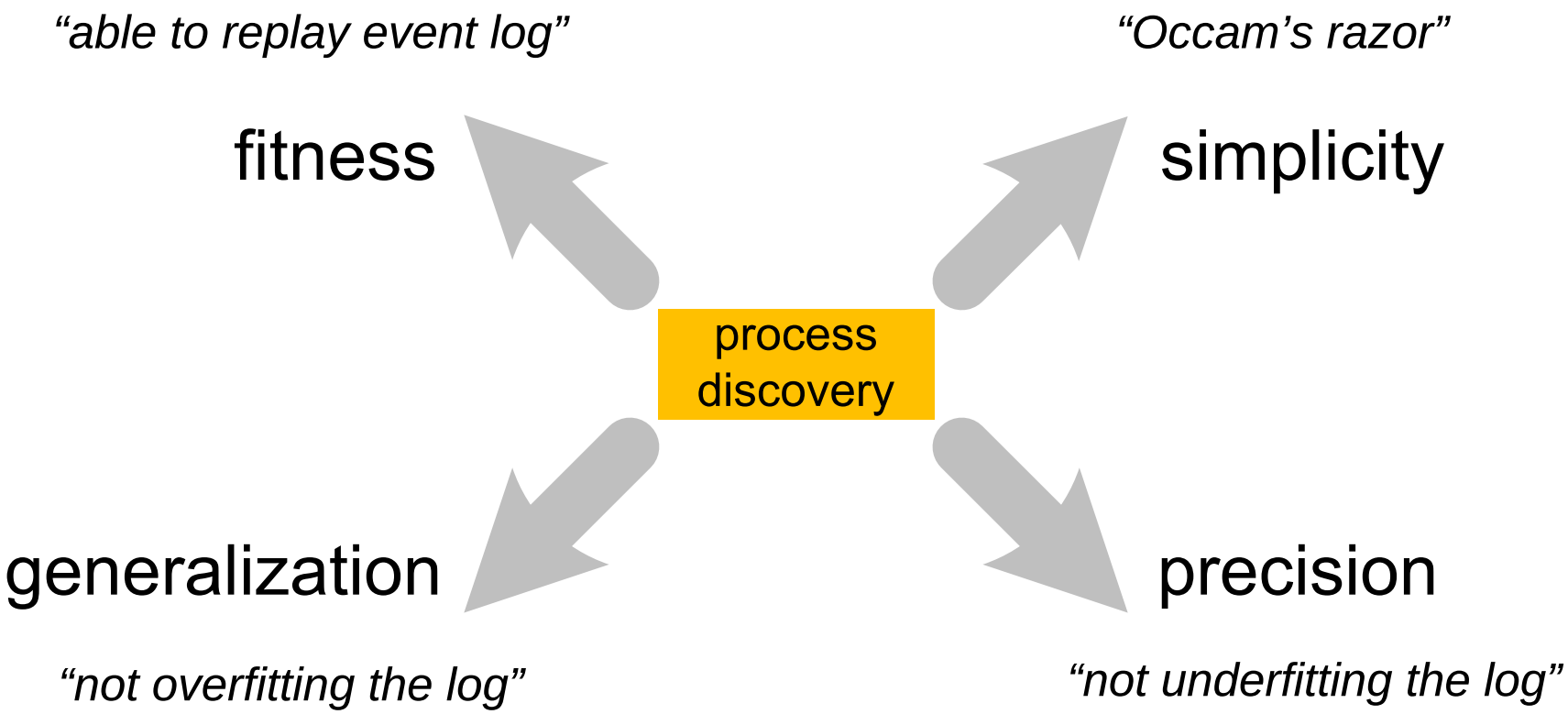
“It is time to stop paving the cow paths. Instead of embedding outdated processes in silicon and software, we should obliterate them and start over.” Michael Hammer in “Reengineering Work: Don’t Automate, Obliterate” (1990)



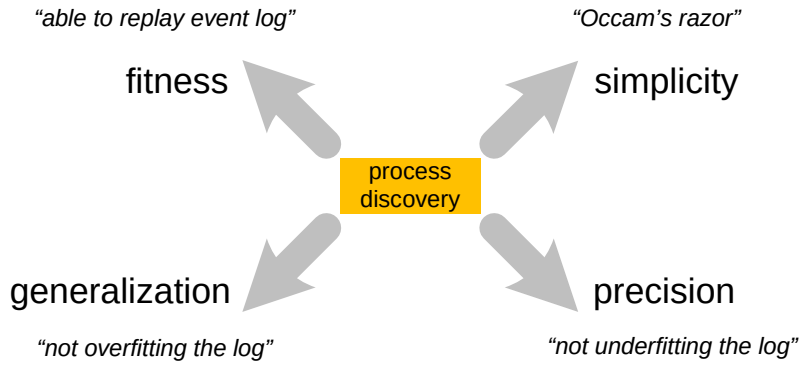

start
register request
examine thoroughly
examine casually
check ticket
decide
pay compensation
reject request
reinitiate request
end
Don’t pave new roads if you do not know the cow paths!
Don’t close your eyes for event data!

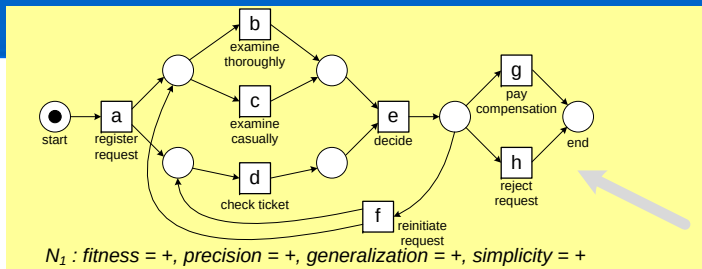
The Four Dimensions of Conformance Checking

Challenge: four competing quality criteria

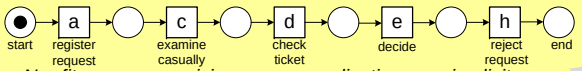


Example: one log four models

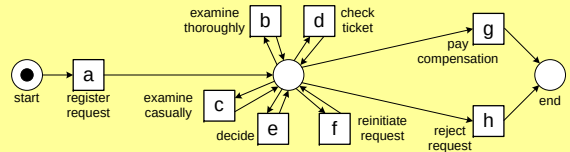




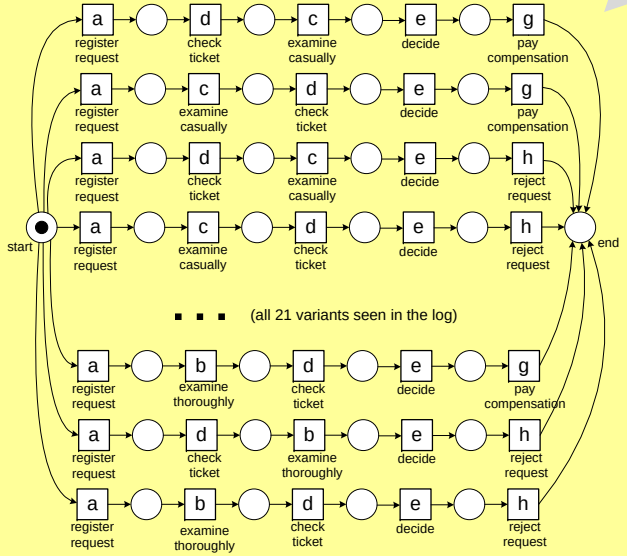
N_1 : *fitness* = +, *precision* = +, *generalization* = +, *simplicity* = +



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



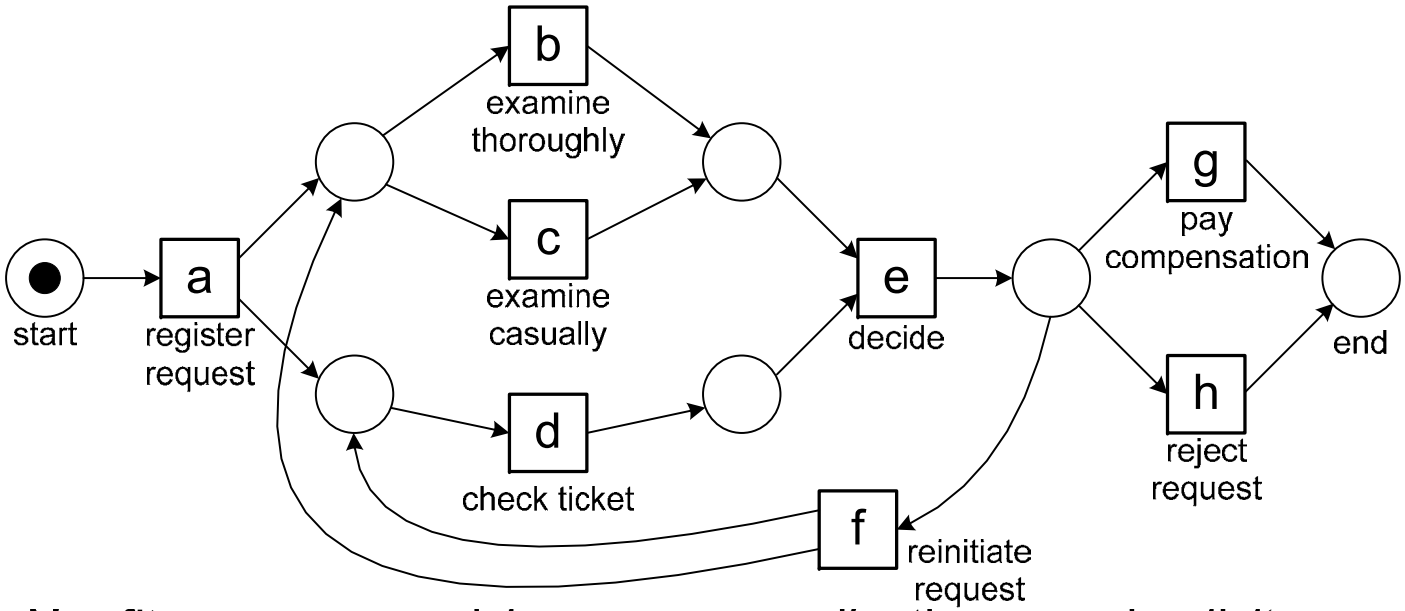
N_3 : *fitness* = +, *precision* = -, *generalization* = +, *simplicity* = +



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	

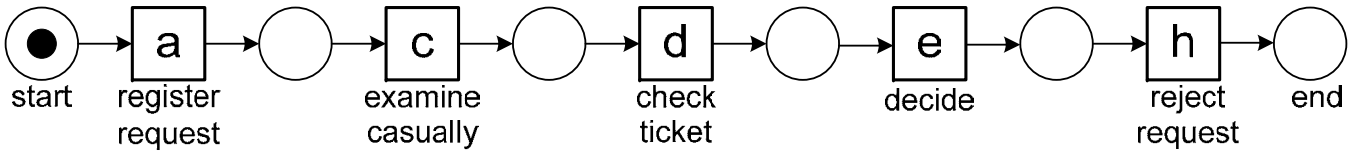
Model N_1



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	

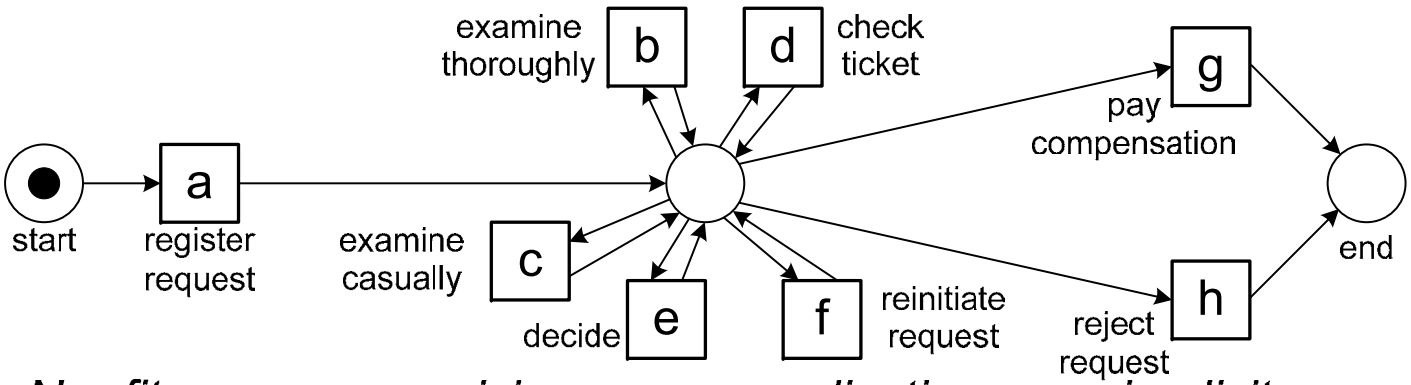
Model N_2



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

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

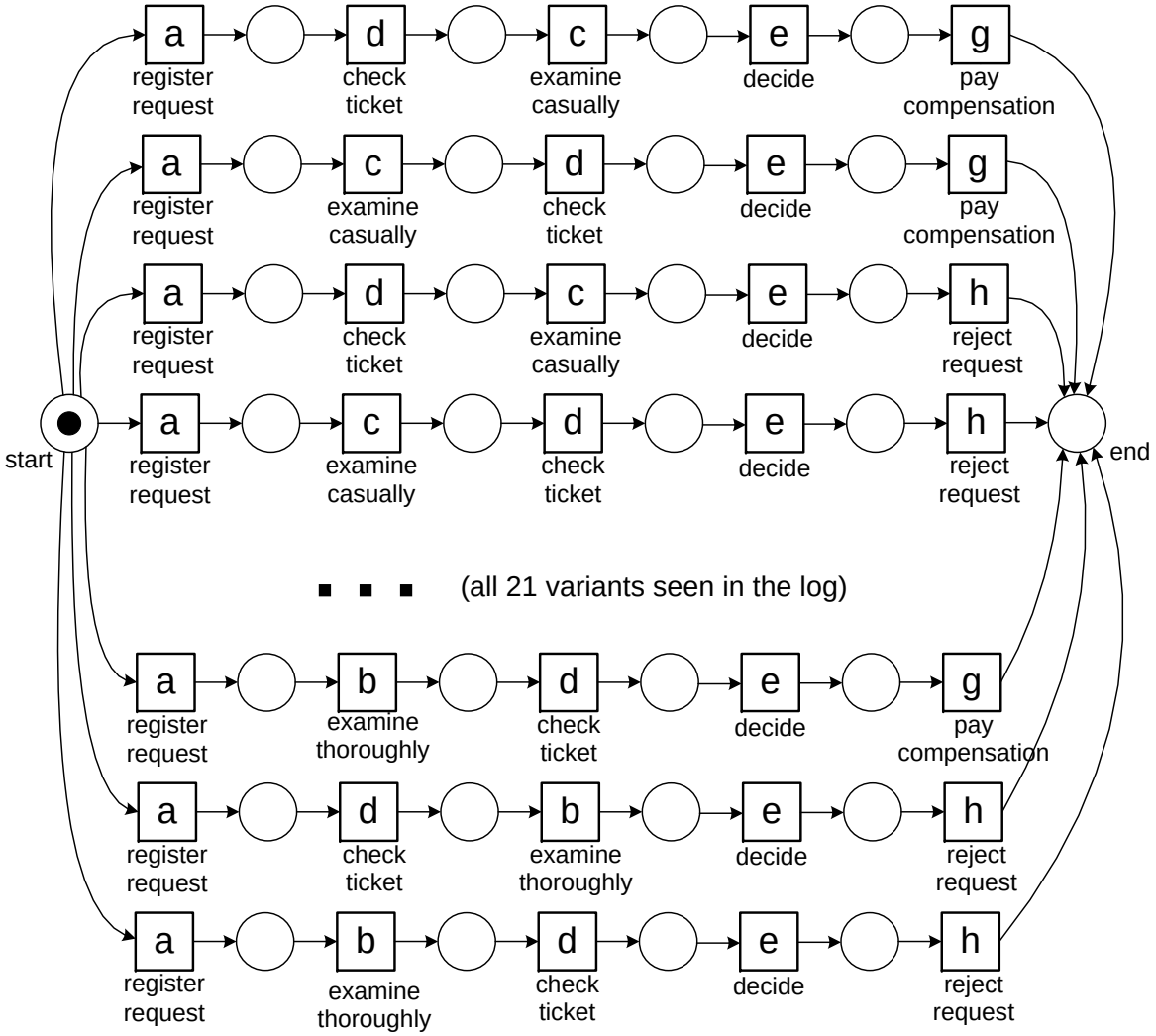
Model N_3



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

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

Model N_4



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

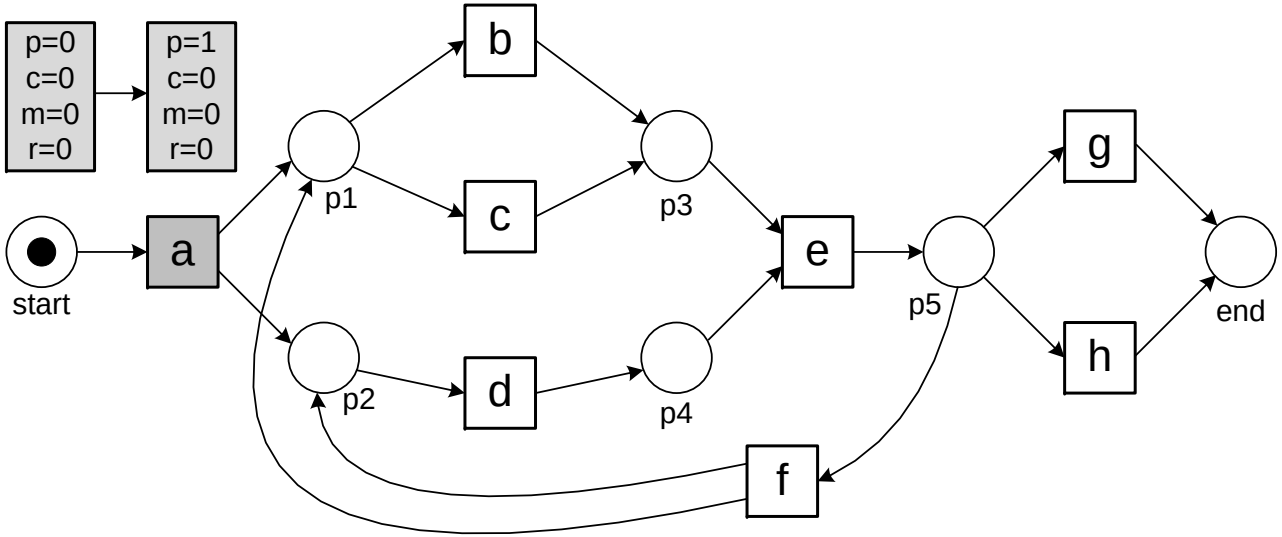
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

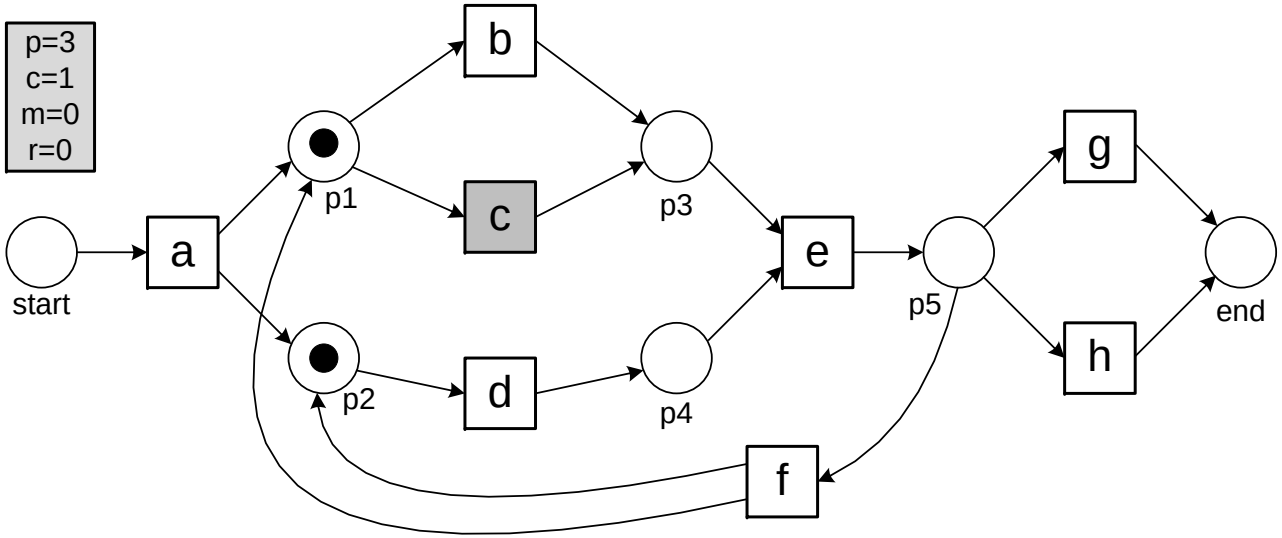
Conformance Checking by Playing the Token Game

Joint work with Anne Rozinat

Replaying (1/3)
σ_1 on N_1

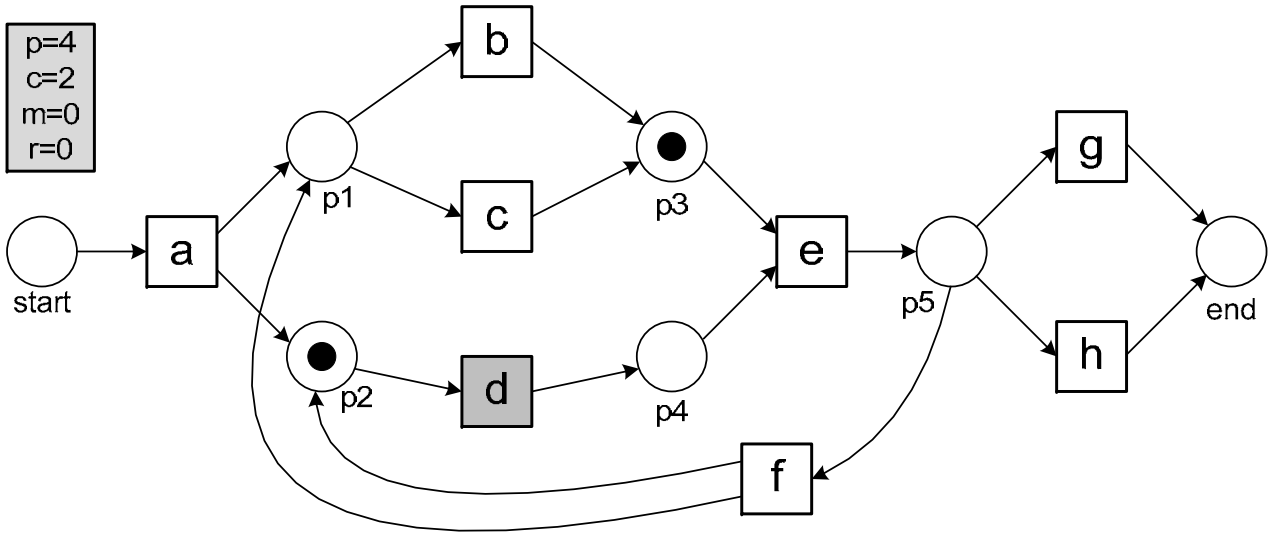
$$\sigma_1 = \langle a, c, d, e, h \rangle$$

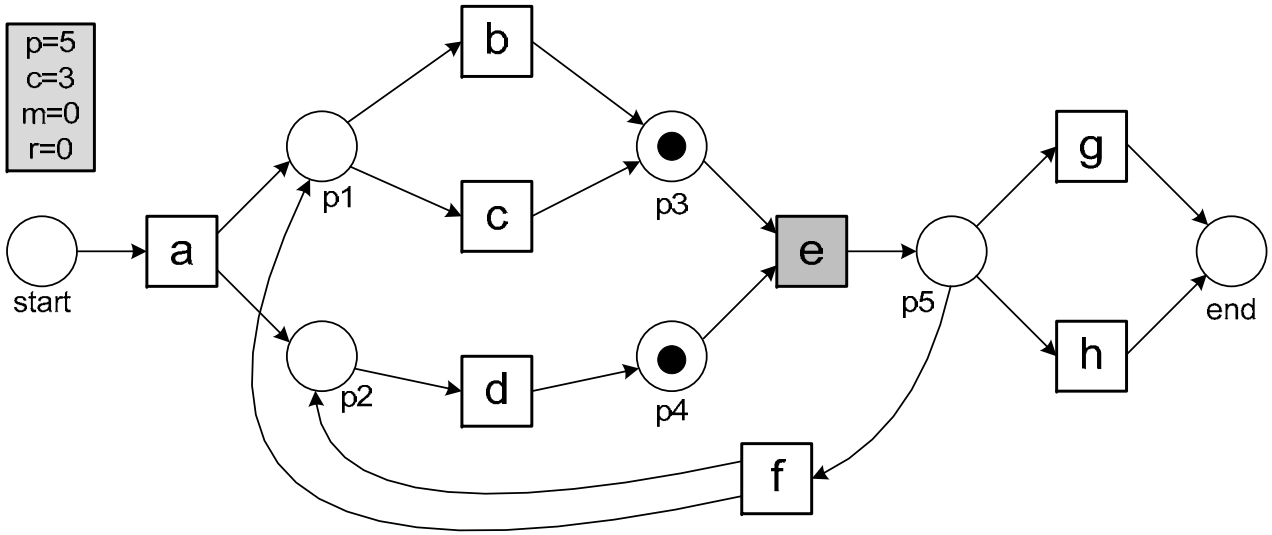




Replaying (2/3)

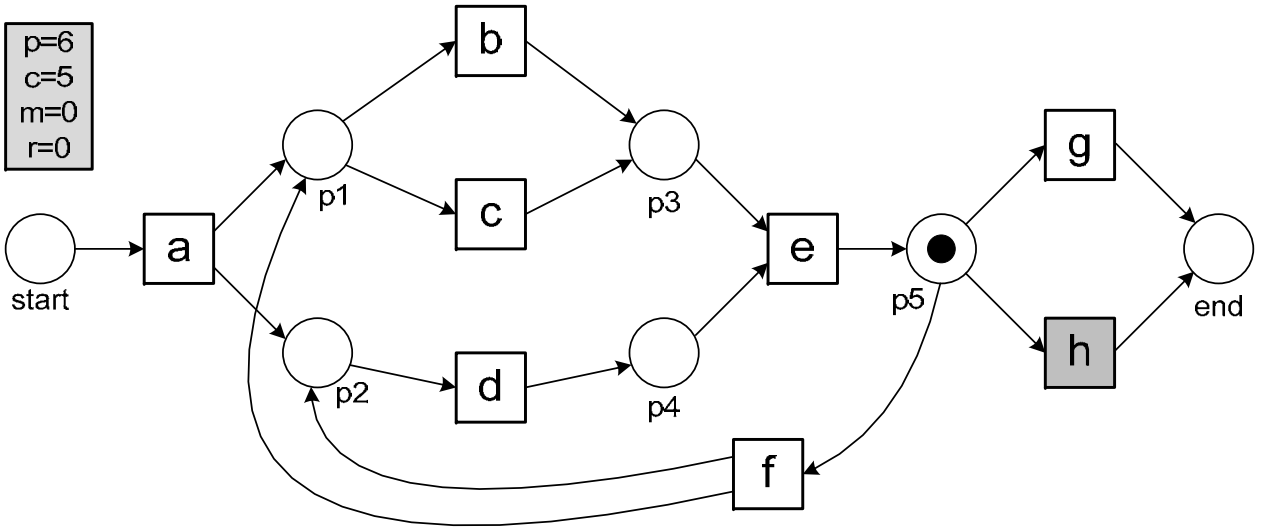
$$\sigma_1 = \langle a, c, d, e, h \rangle$$

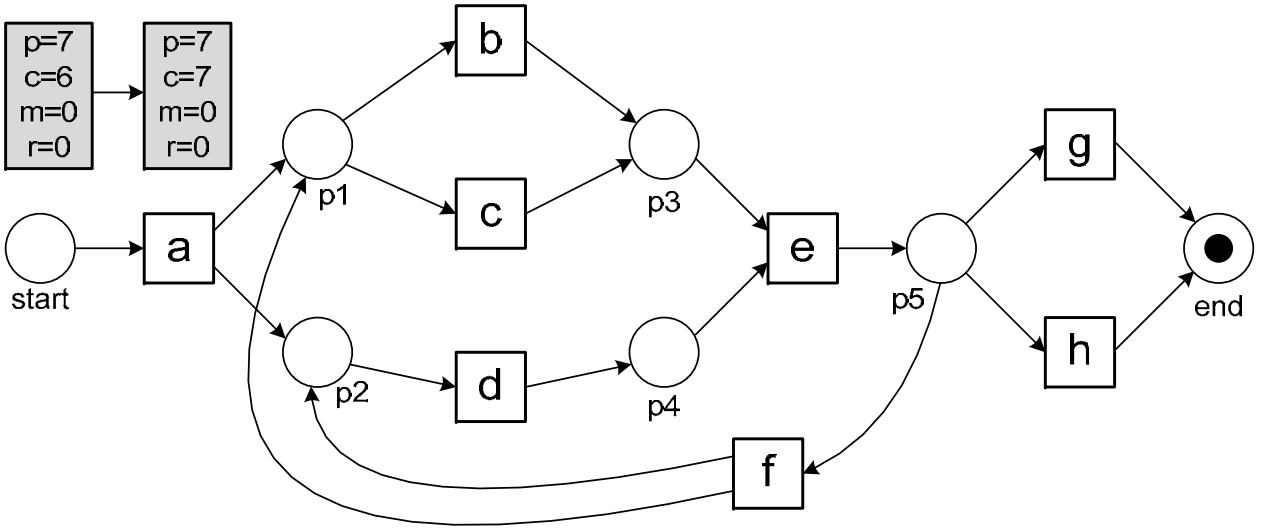




Replaying (3/3)

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

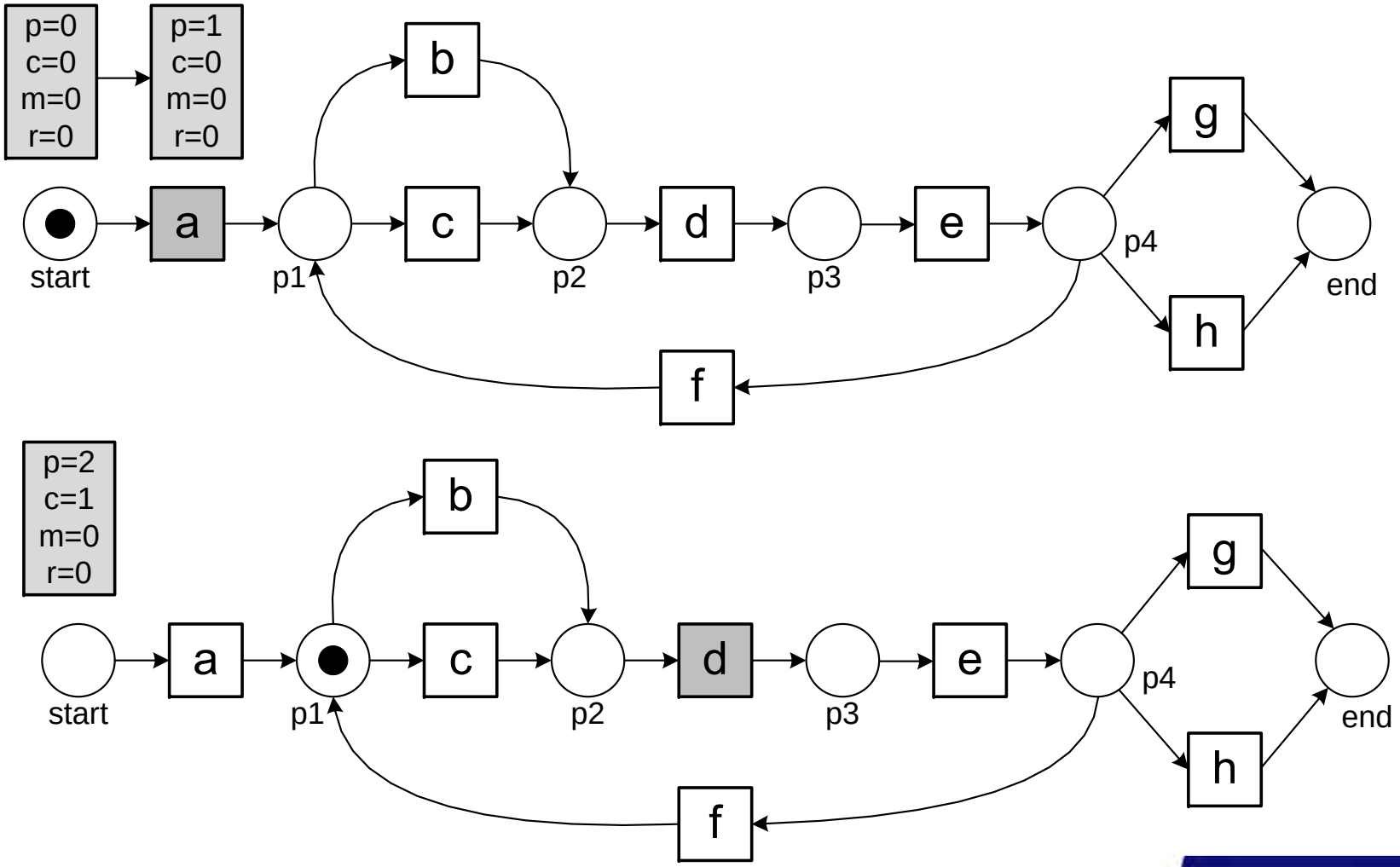




No problems found!

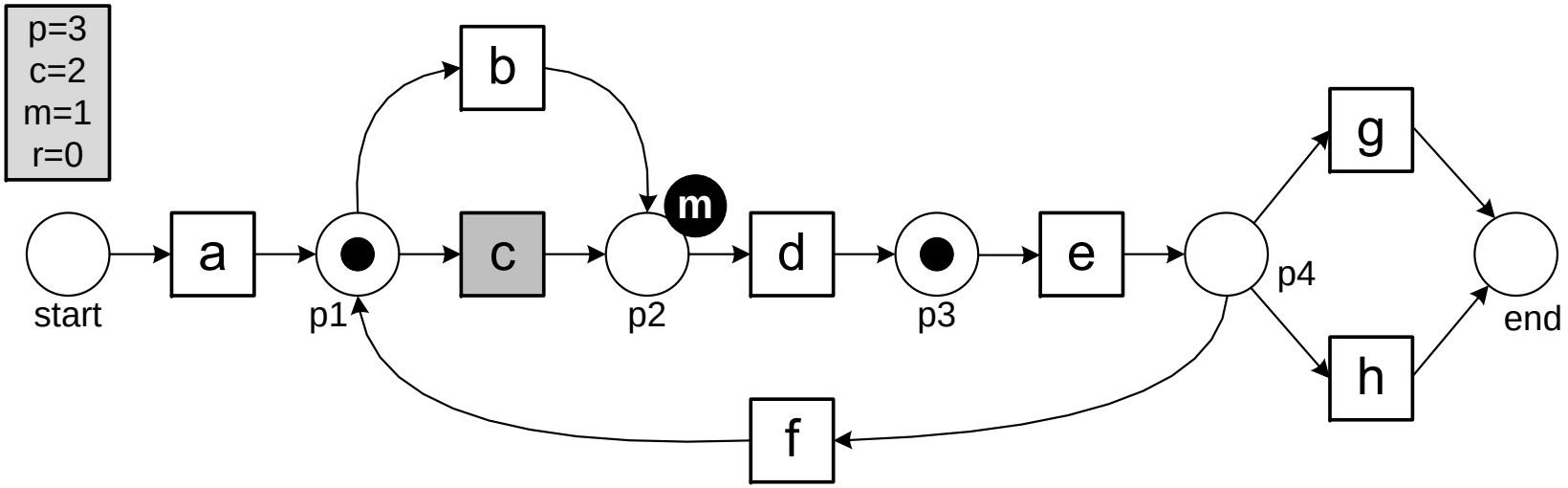
Replaying (1/3)
σ_3 on N_2

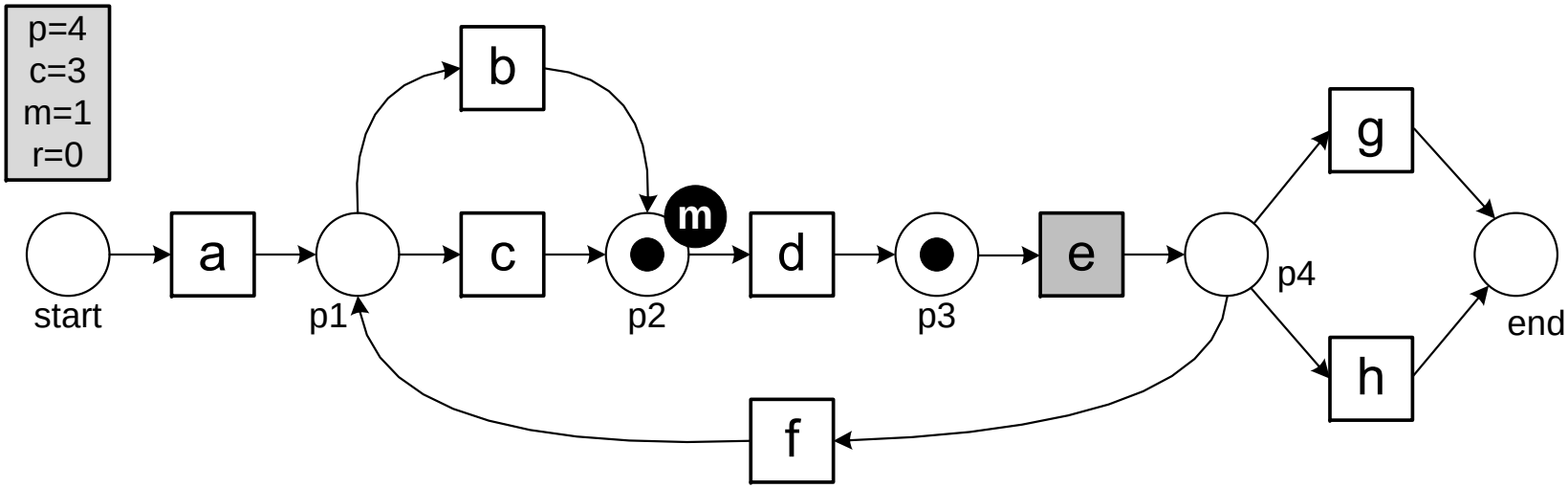
$$\sigma_3 = \langle a, d, c, e, h \rangle$$



Replaying (2/3)

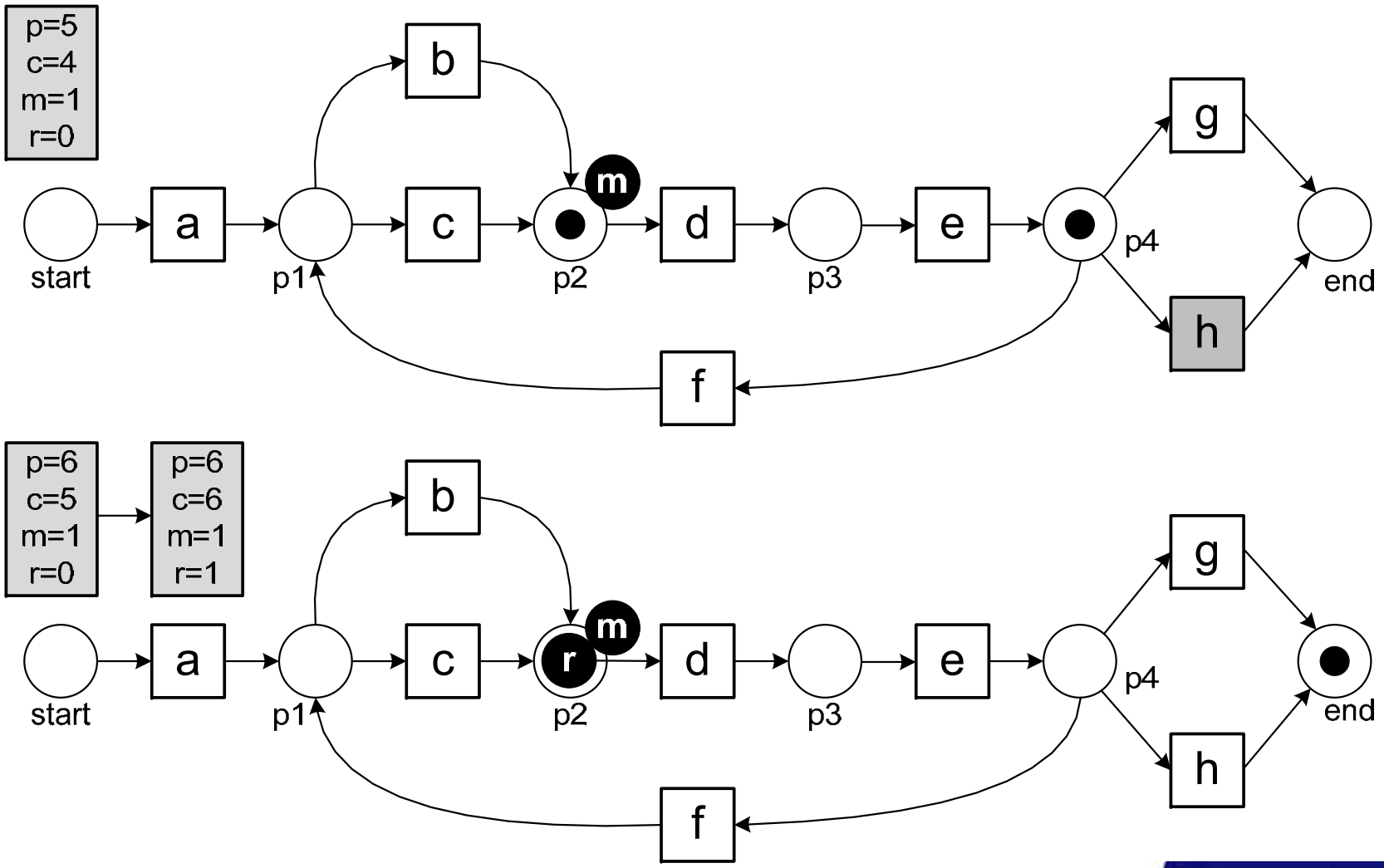
$$\sigma_3 = \langle a, d, c, e, h \rangle$$





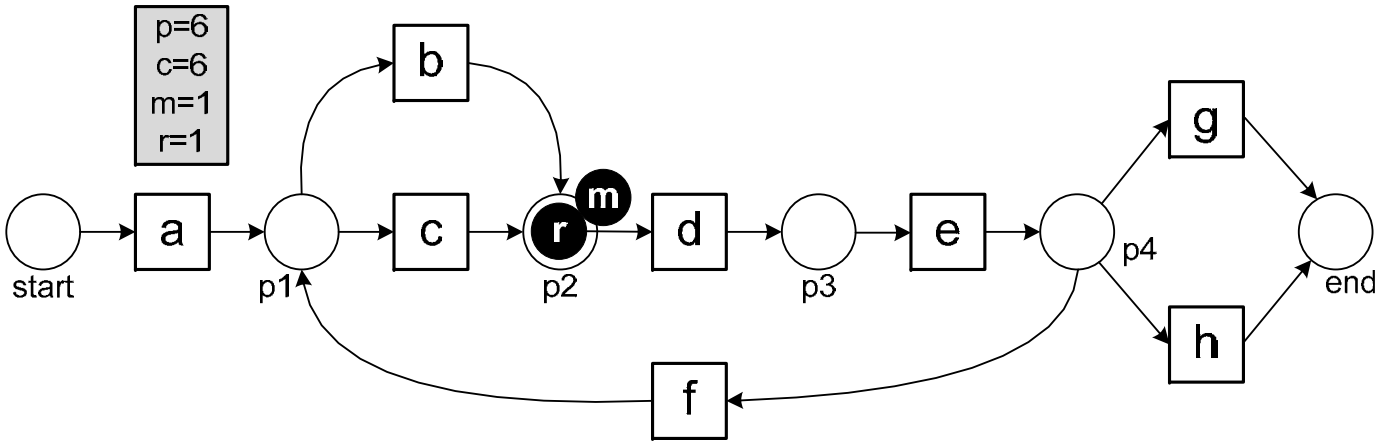
Replaying (3/3)

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



Problems encountered when replaying σ_3 on N_2

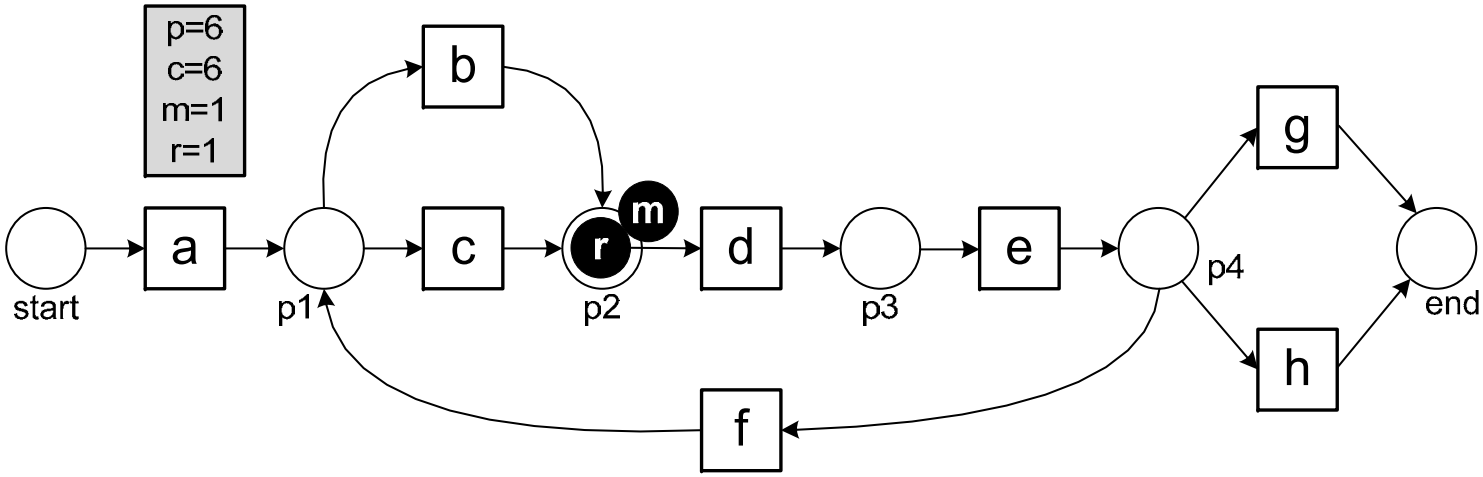
$$\sigma_3 = \langle a, d, c, e, h \rangle$$



- One missing token (of 6 consumed tokens)
- One remaining token (of 6 produced tokens)

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

Computing fitness at trace level



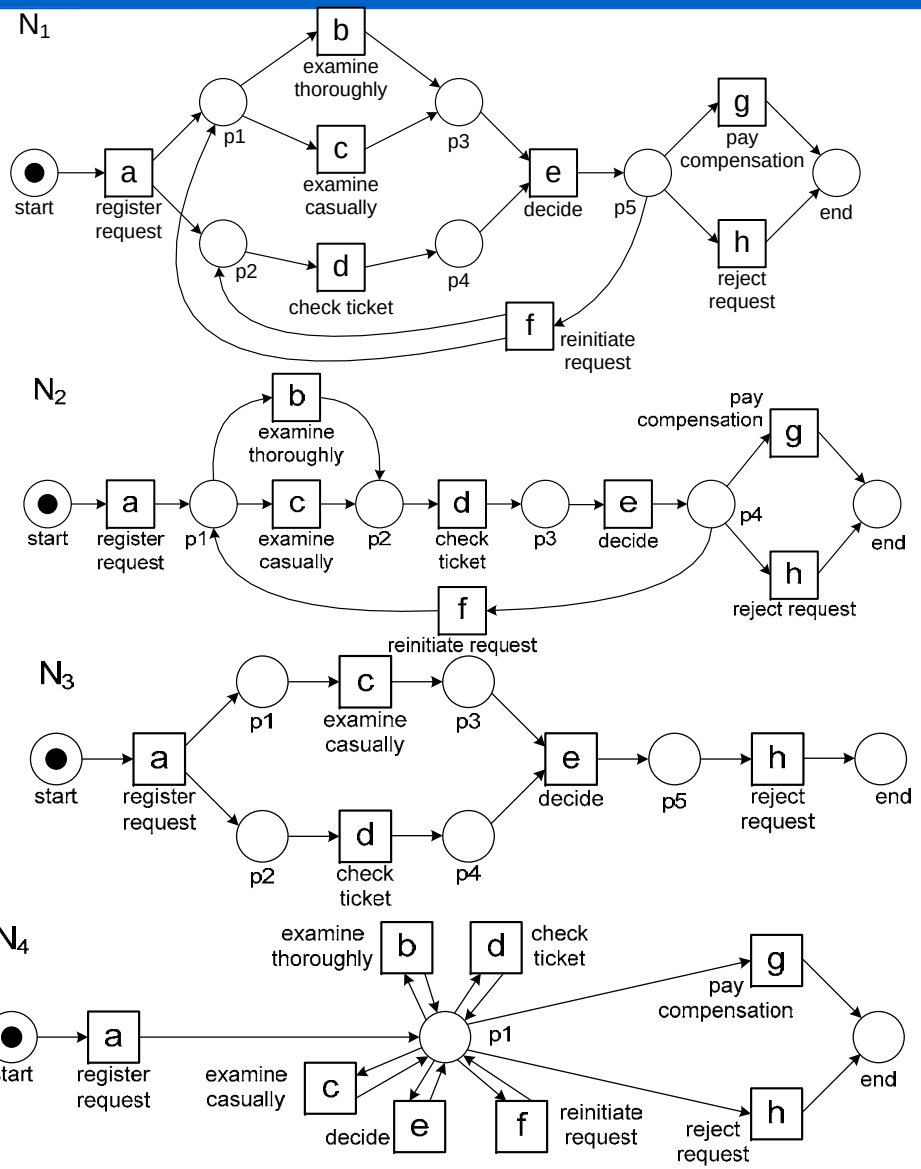
$$fitness(\sigma_3, N_2) = \frac{1}{2}\left(1 - \frac{1}{6}\right) + \frac{1}{2}\left(1 - \frac{1}{6}\right) = 0.8333$$

Computing fitness at the log level

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

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

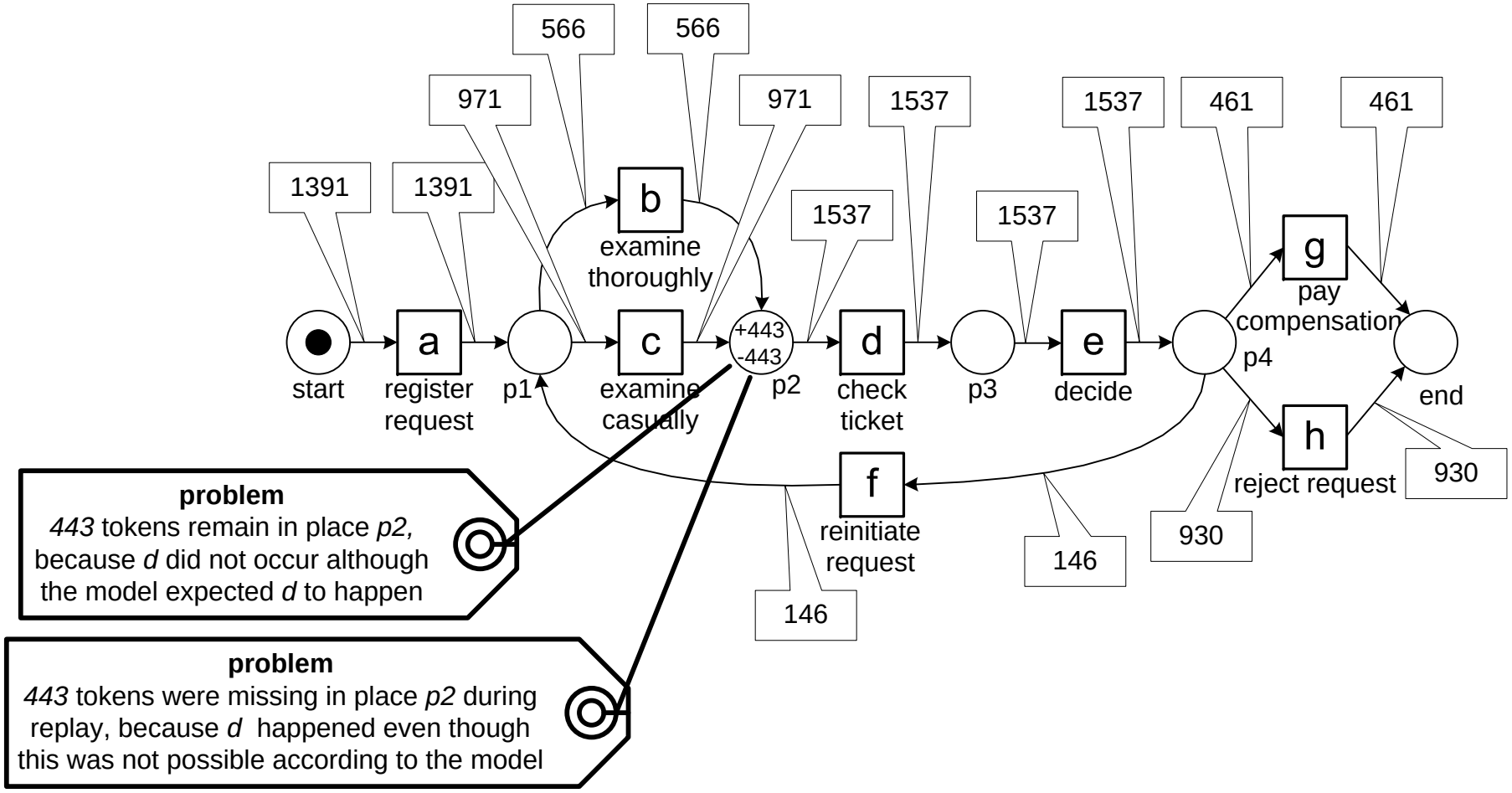
Example values



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

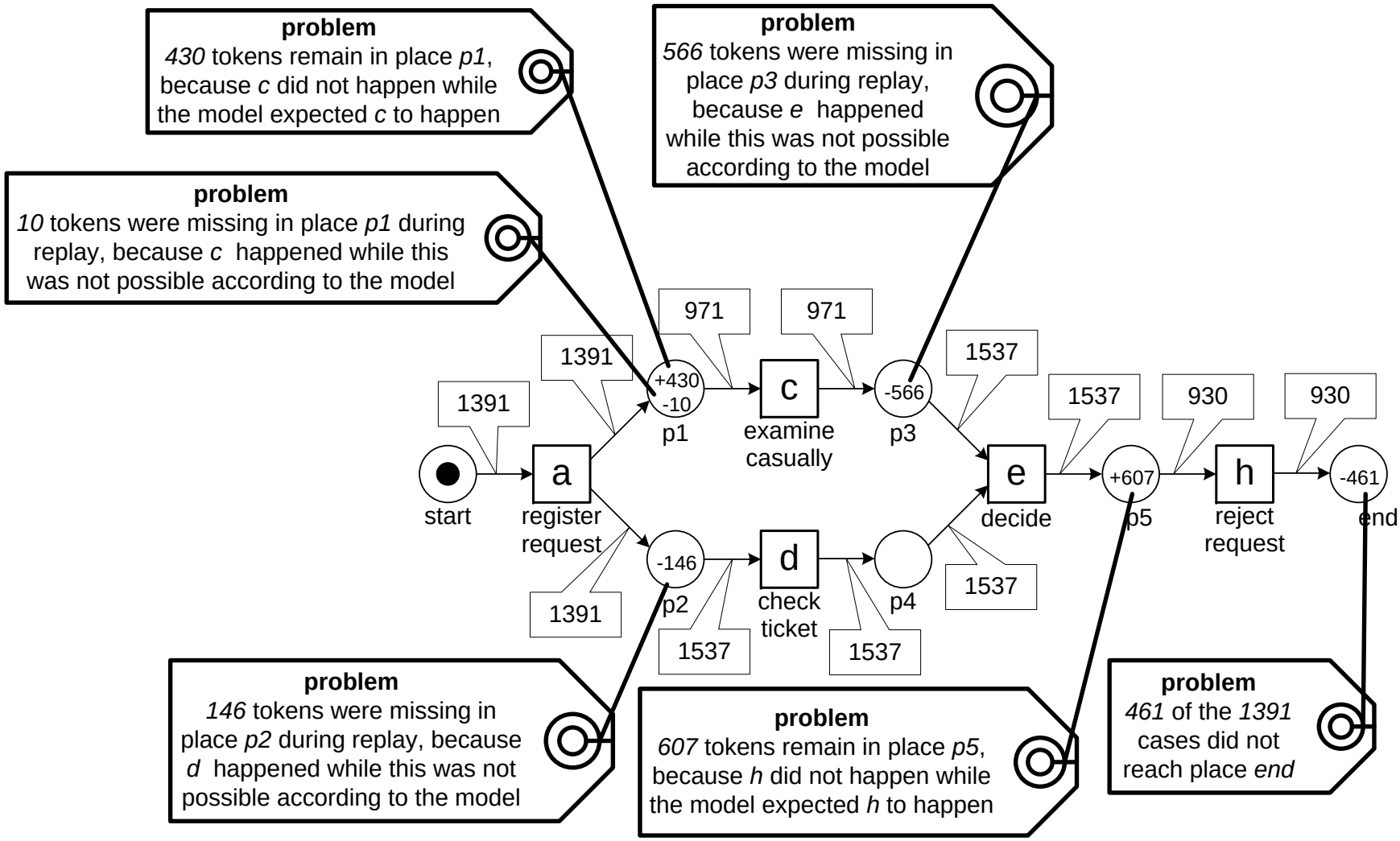
Diagnostics

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

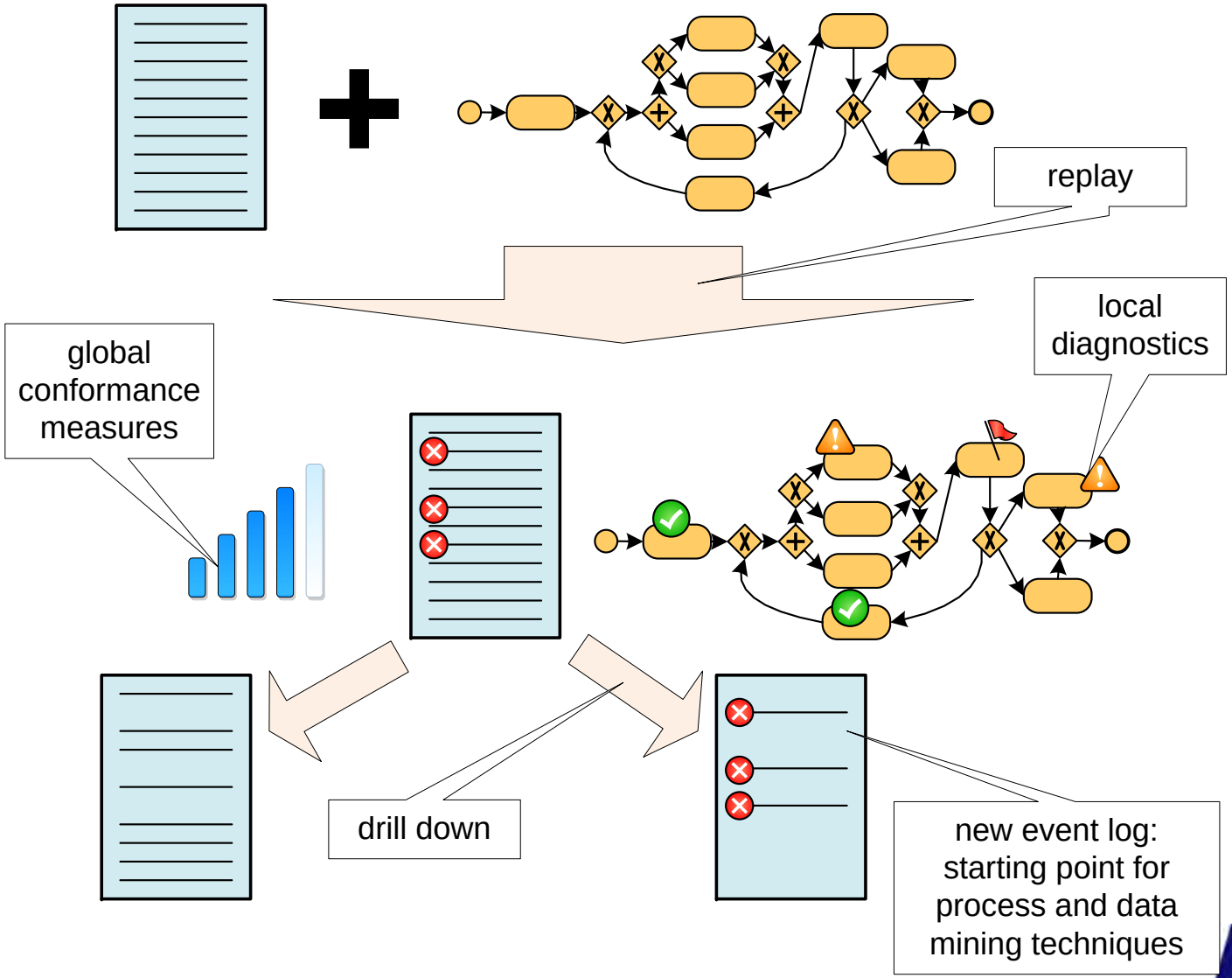


Diagnostics

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



Drilling down



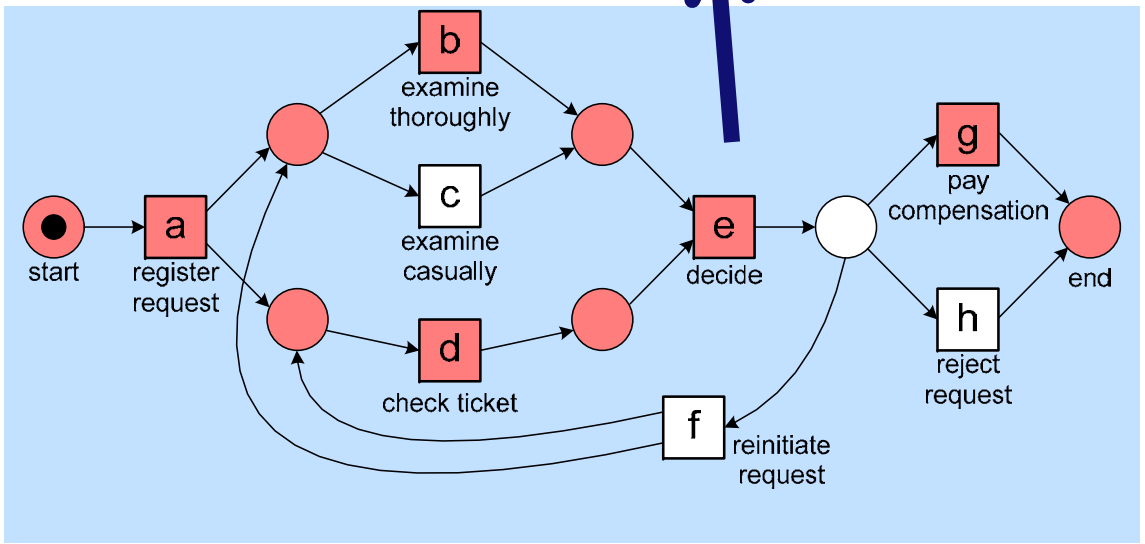
Conformance Checking Based on Alignments

Joint work with Arya Adriansyah and Boudewijn van Dongen (also see poster)

From “playing the token game” to optimal alignments …

191 times “abdeg”

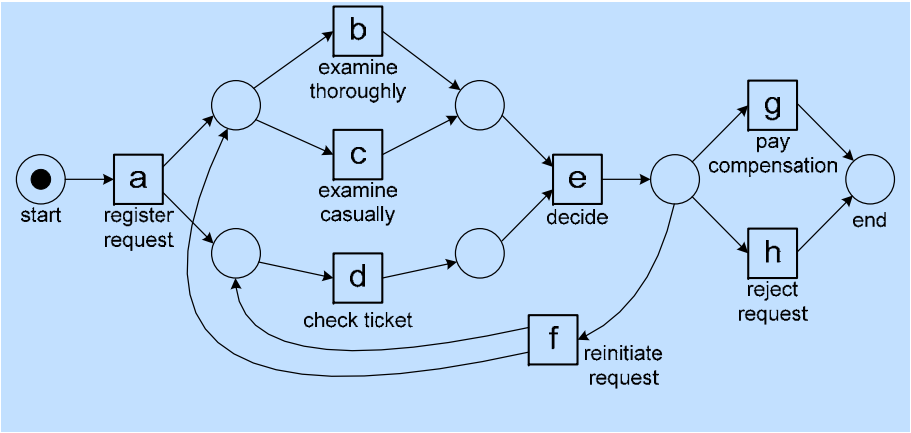
a	b	d	e	g
a	b	d	e	g



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

Example alignments

abdeg

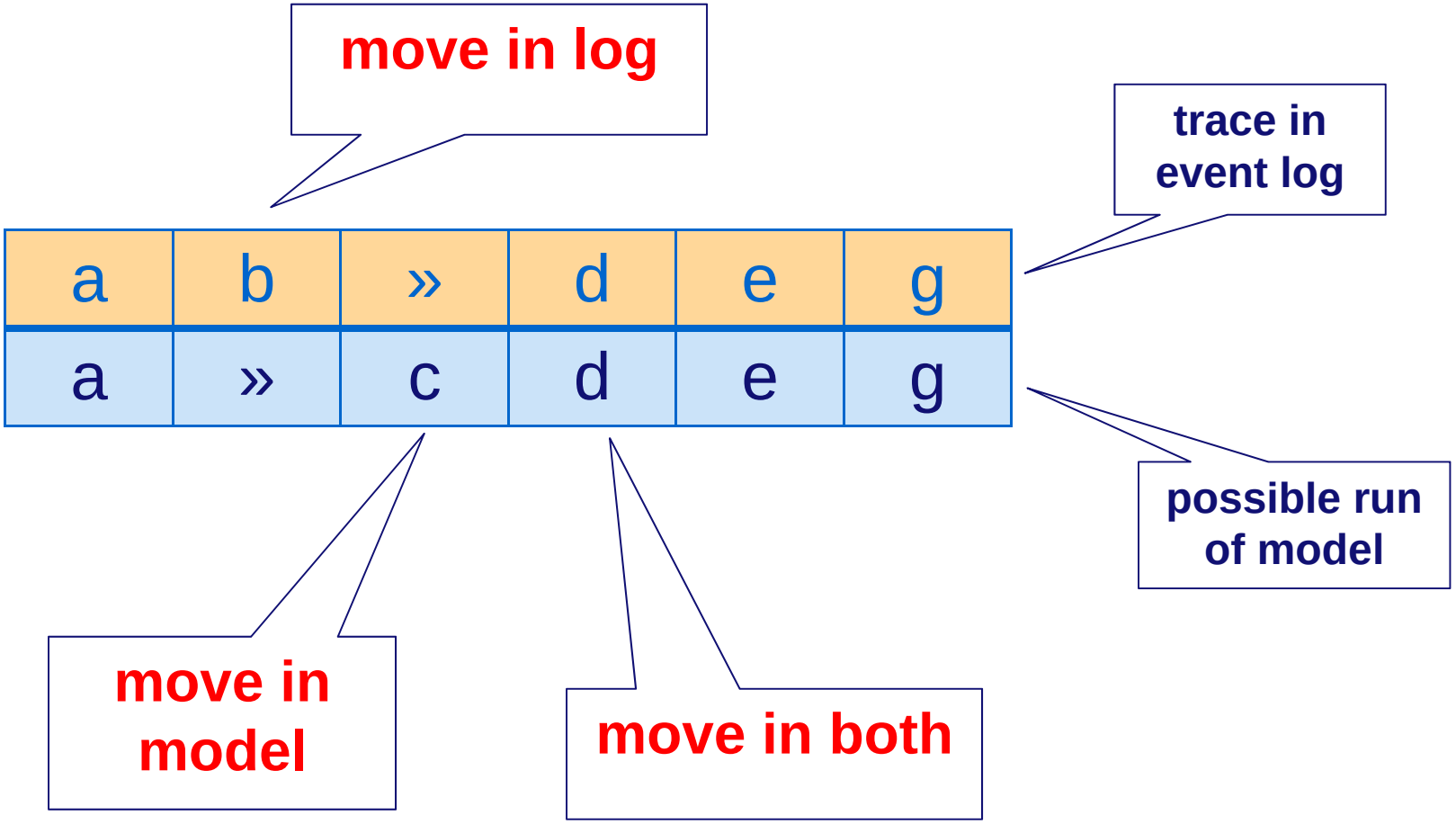


a	b	d	e	g
a	b	d	e	g

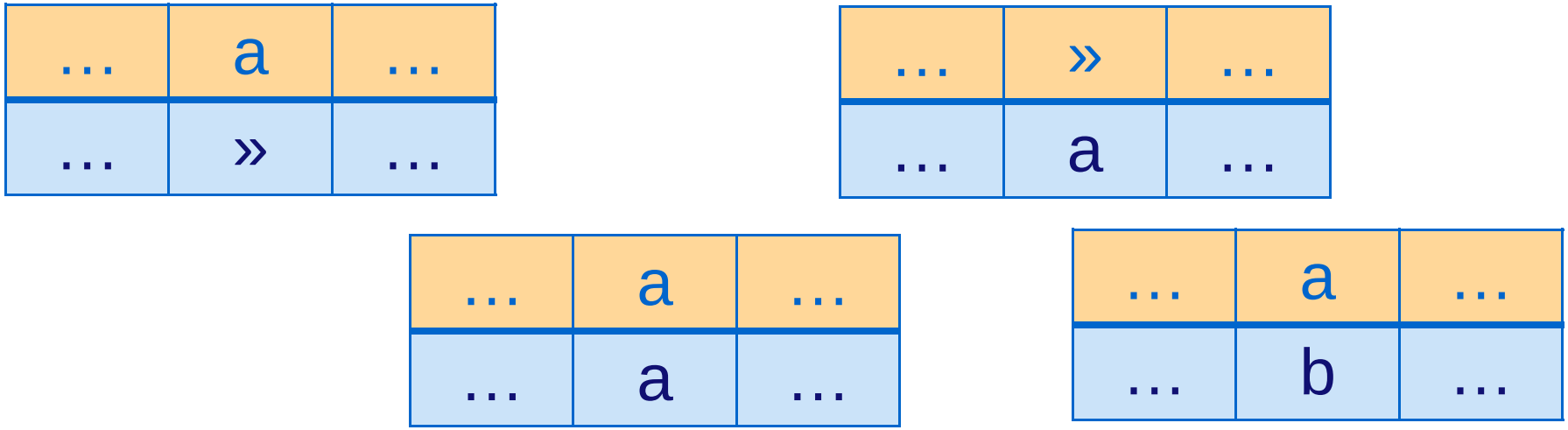
a	b	»	d	e	g
a	»	c	d	e	g

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

Moves in an alignment

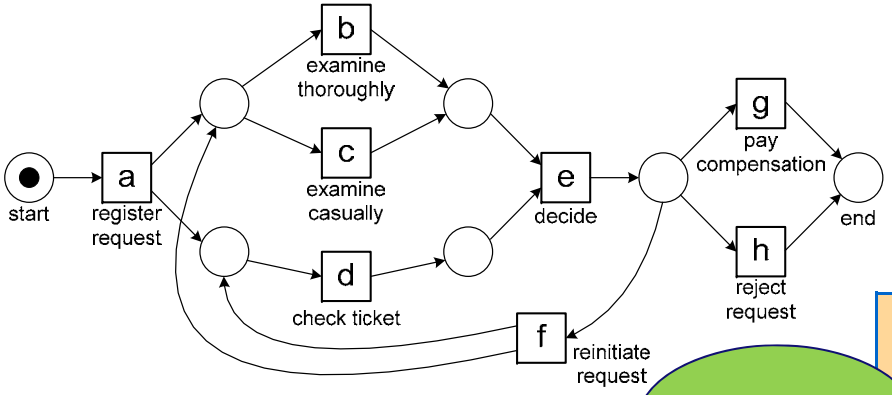


Moves have costs



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

Optimal alignment (smallest costs)



abdeg

a	b	d	e	g
a	b	d	e	g

optimal

0

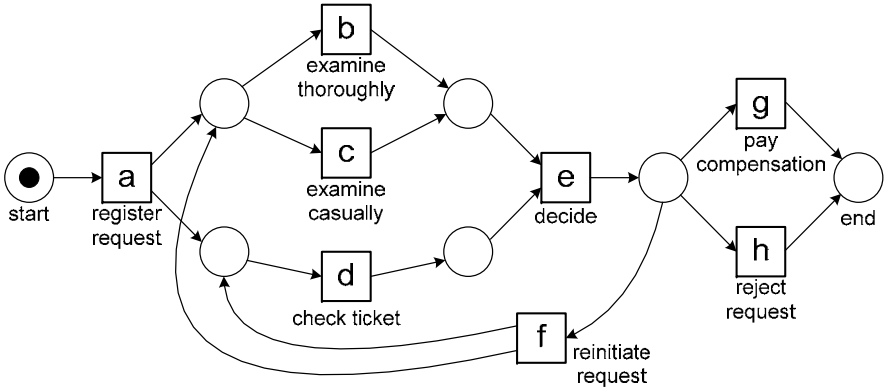
a	b	»	d	e	g
a	»	c	d	e	g

2

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

10

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

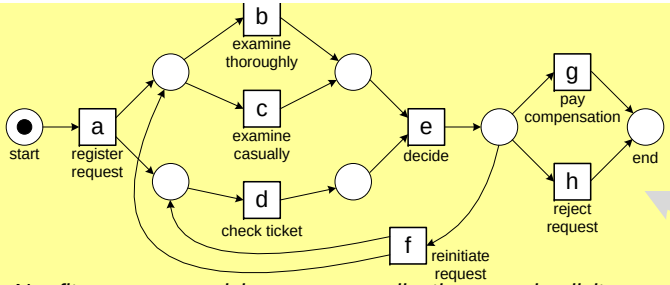
Any cost structure is possible

...	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).**
- **cf. cost/risk-aware BPM (costs = risk).**

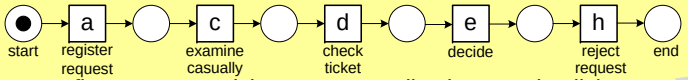
Fitness

1.0



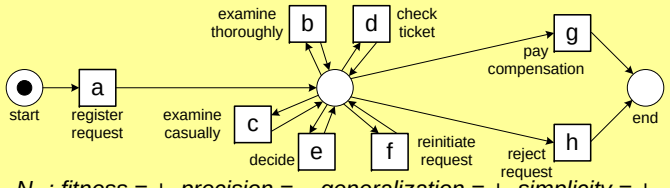
N_1 : *fitness* = +, *precision* = +, *generalization* = +, *simplicity* = +

0.8



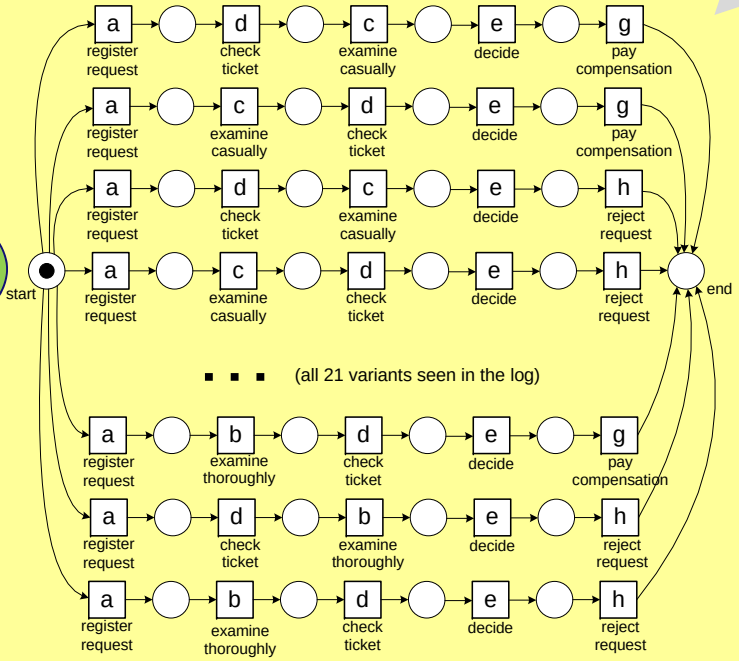
N_2 : *fitness* = -, *precision* = +, *generalization* = -, *simplicity* = +

1.0



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

1.0

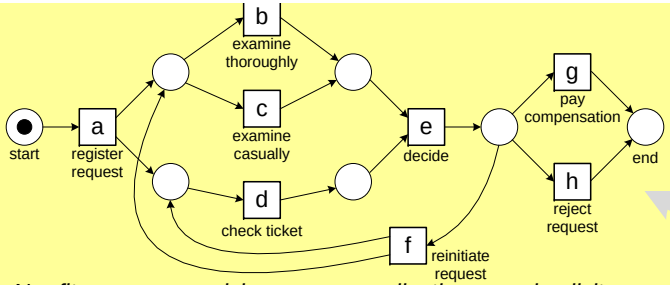


N_4 : *fitness* = +, *precision* = +, *generalization* = -, *simplicity* = -

#	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	

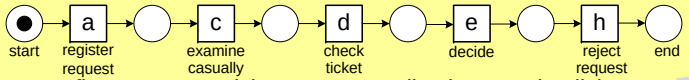
Precision

0.97



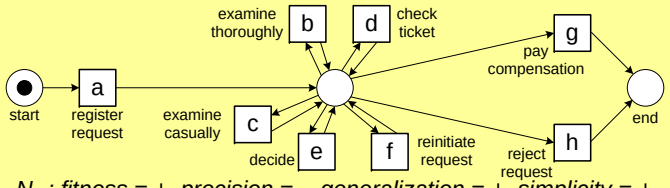
N_1 : *fitness* = +, *precision* = +, *generalization* = +, *simplicity* = +

1



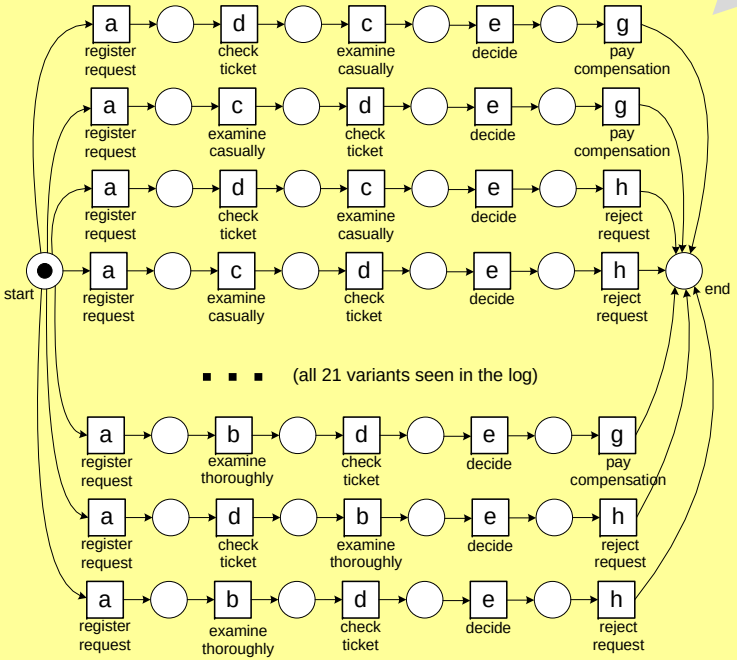
N_2 : *fitness* = -, *precision* = +, *generalization* = -, *simplicity* = +

0.41



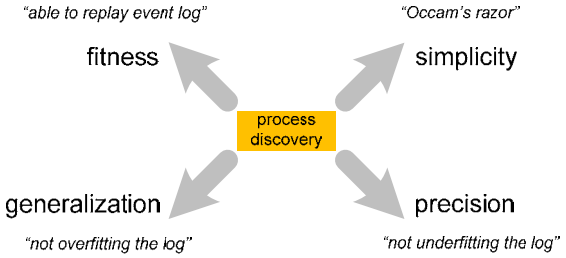
N_3 : *fitness* = +, *precision* = -, *generalization* = +, *simplicity* = +

1



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	



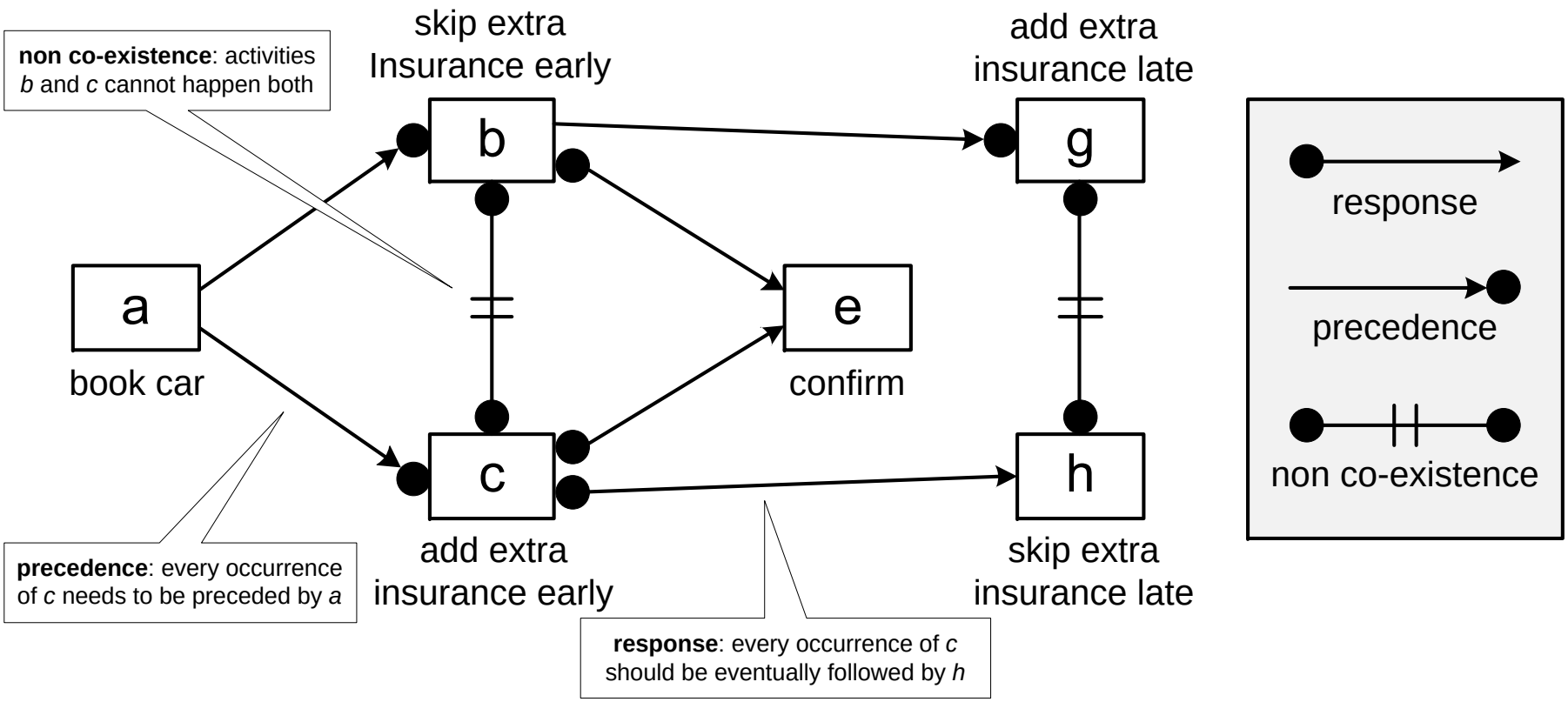
Advantages of Aligning Log and Model

- **Observed behavior is directly related to modeled behavior.**
- **Highly flexible (any cost structure).**
- **Detailed diagnostics.**
- **After aligning log and model, other quality dimensions can be investigated (separation of concerns).**
- **Efficiently implemented in ProM (see work of Arya Adriansyah).**

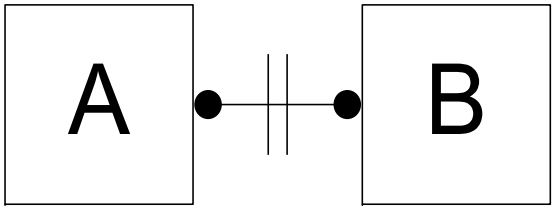
Using Declarative Languages

Joint work with Fabrizio Maggi, Michael Westergaard, Maja Pesic, et al.

The Declare language



Basic idea

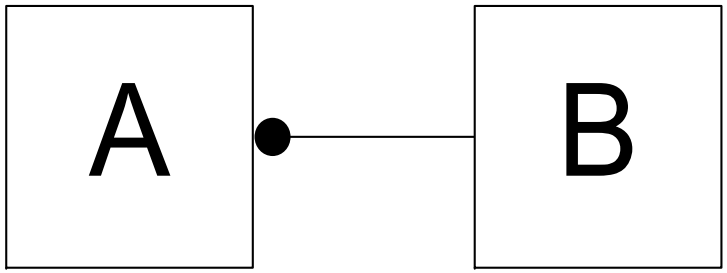


$$\Diamond(A) \Leftrightarrow \neg(\Diamond(B))$$

LTL semantics

name	notation	explanation
nexttime	$\bigcirc F$	F has to hold at the next state, e.g., *[A,F,B,C,D,E]*, *[A,F,F,F,F,F,B,C,D,E]*, *[F,F,F,F,A,B,C,D,E]*, etc.
eventually	$\Diamond F$	F has to hold eventually, e.g., *[F,A,B,C,D,E]*, *[A,B,C,F,D,E]*, *[ABFCDFEF]*, etc.
always	$\Box F$	F has to always hold, e.g., *[F,F,F,F,F,F]*.
until	$F \sqcup G$	G holds at the current state or at some future state, and F has to hold until G holds. When G holds F does not have to hold any more. Examples are *[G,A,B,C,D,E]*, *[F,G,A,B,C,D,E]*, *[F,F,F,F,G,A,B,C,D,E]*, *[F,F,F,F,G,A,B,G,F,C,D,E,F,G]*, etc.

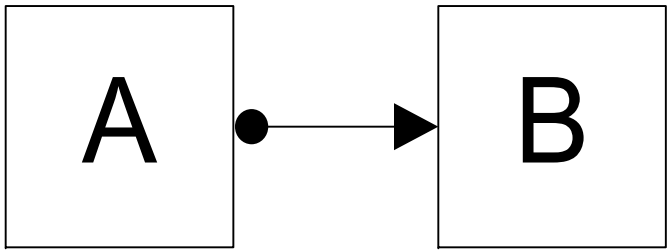
Example: *"existence response"*



$$\Diamond(A) \Rightarrow \Diamond(B)$$

- **OK:**
  - **[ ]**
  - **[A,B,C,D,E]**
  - **[A,A,A,C,D,E,B,B,B]**
  - **[B,B,A,A,C,D,E]**
  - **[B,C,D,E]**
- **NOK**
  - **[A]**
  - **[A,A,C,D,E]**

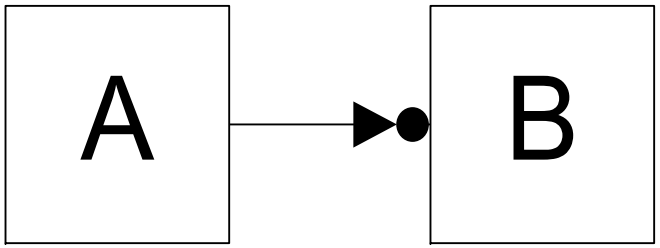
Example: *"response"*



$$\Box(A \Rightarrow \Diamond(B))$$

- **OK:**
  - **[ ]**
  - **[A,B,C,D,E]**
  - **[A,A,A,B,C,D,E]**
  - **[B,B,A,A,B,C,D,E]**
  - **[B,C,D,E]**
- **NOK**
  - **[A]**
  - **[B,B,B,B,A,A]**

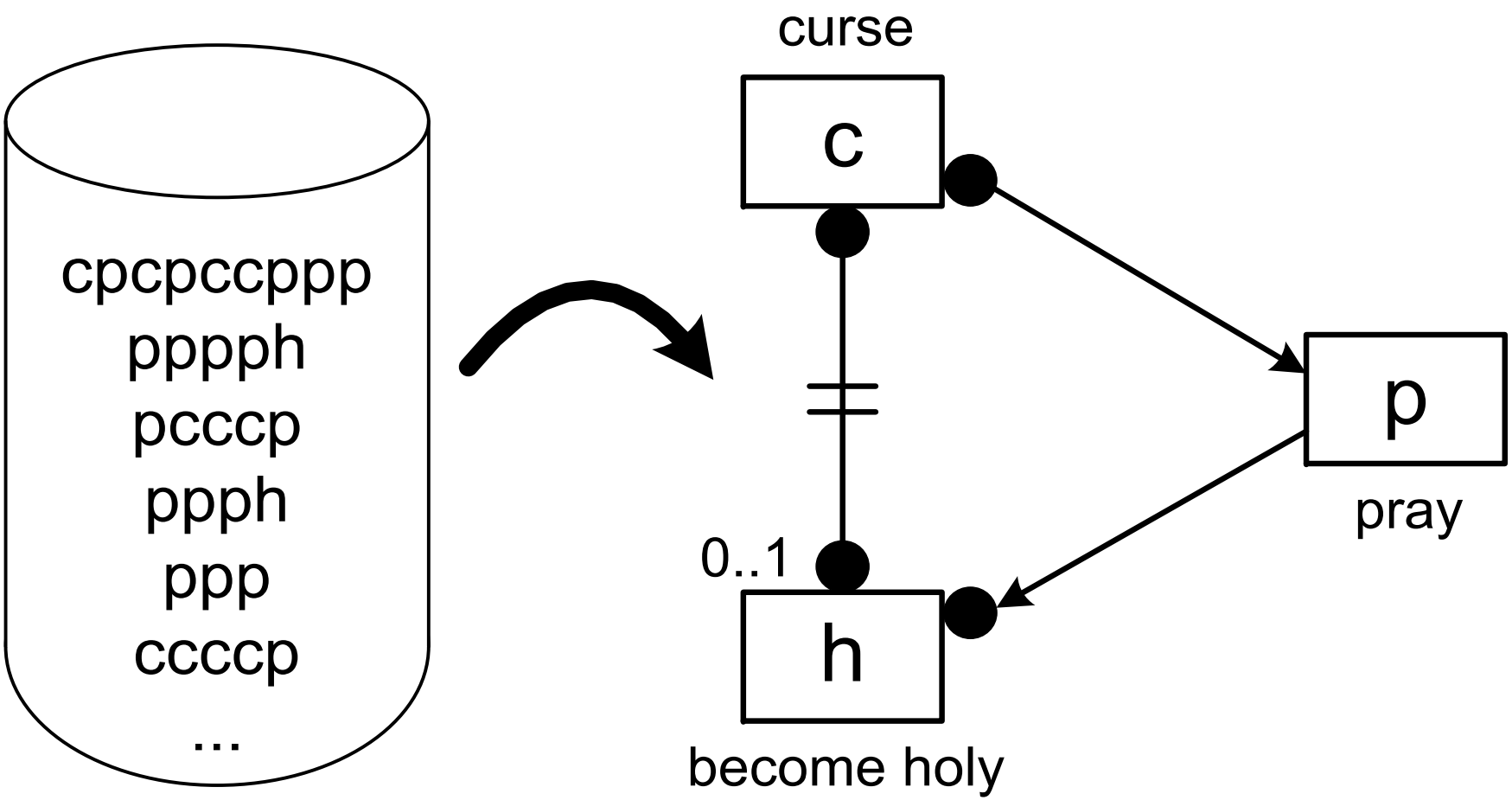
Example: *"precedence"*



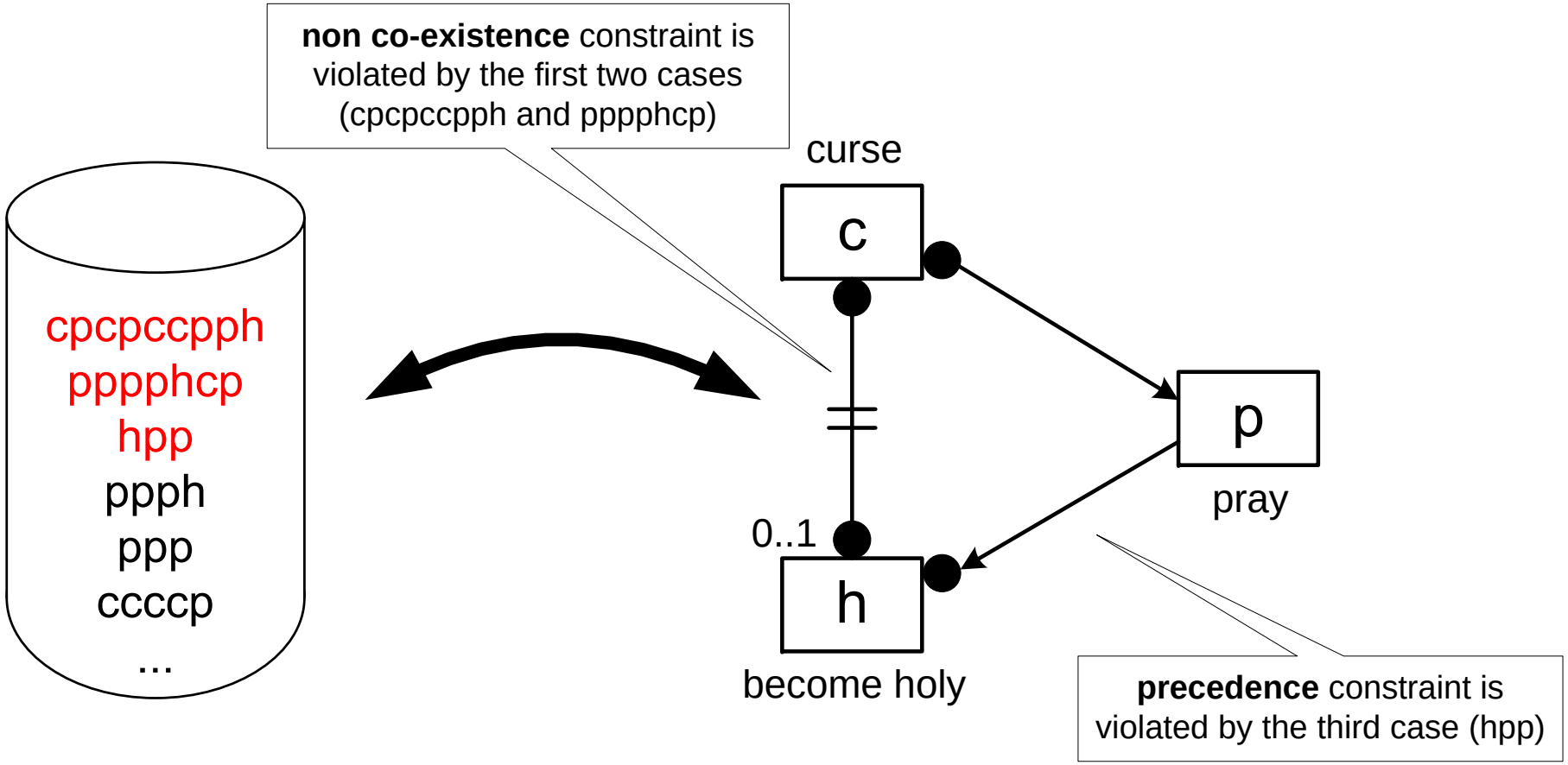
$$\Diamond(B) \Rightarrow ((\neg B) \sqcup A)$$

- **OK:**
  - **[ ]**
  - **[A,B,C,D,E]**
  - **[A,A,A,C,D,E,B,B,B]**
  - **[A,A,C,D,E]**
- **NOK**
  - **[B]**
  - **[B,A,C,D,E]**

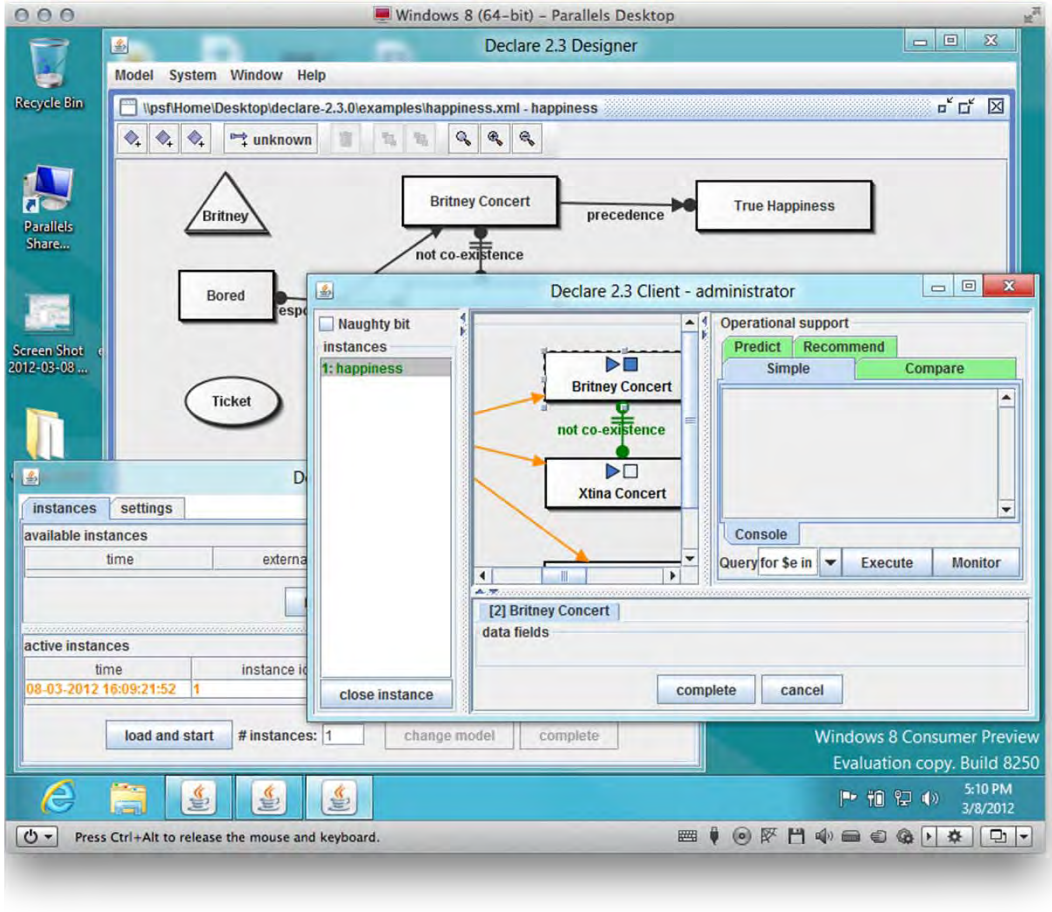
Discovering Declare models

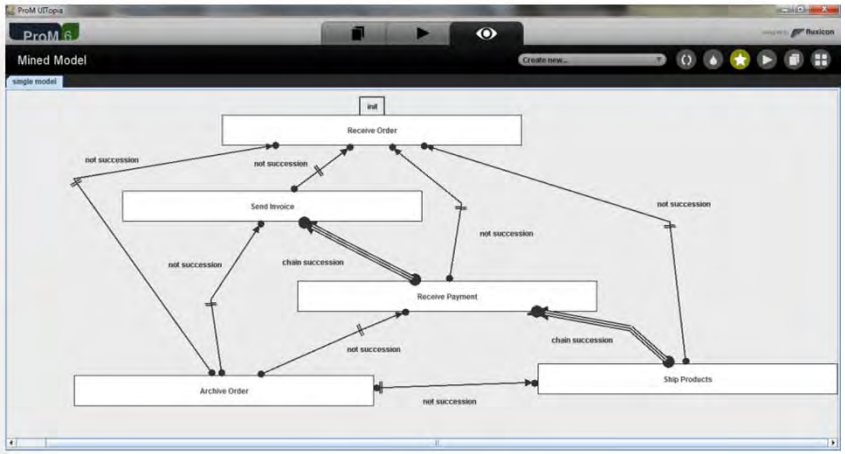


Conformance checking of Declare models



Tool Support

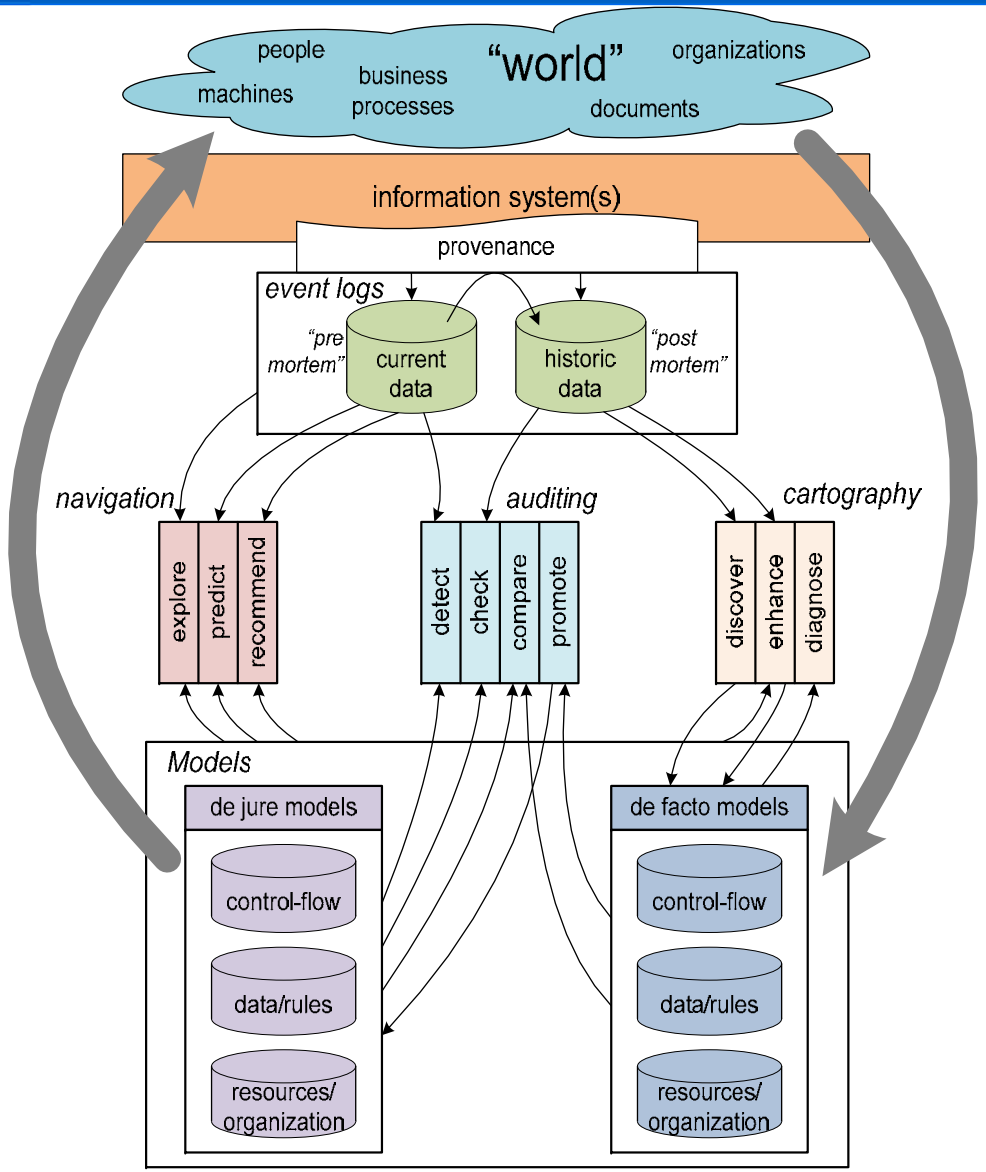




http://declare.sf.net

Conclusion

Overview


people
machines
business processes
"world"
organizations
documents
information system(s)
provenance
event logs
"pre mortem"
current data
historic data
"post mortem"
navigation
auditing
cartography
explore
predict
recommend
detect
check
compare
promote
discover
enhance
diagnose
Models
de jure models
control-flow
data/rules
resources/ organization
de facto models
control-flow
data/rules
resources/ organization

Learn More?

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

Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor

Wil M.P. van der Aalst, *Eindhoven University of Technology and Queensland University of Technology*
Kees M. van Hee and Jan Martijn van der Werf, *Eindhoven University of Technology*
Marc Verdonk, *Deloitte Netherlands and Eindhoven University of Technology*

Auditors can use process mining techniques to evaluate all events in a business process, and do so while it is still running.

Auditors validate information about organizations and their business processes. Reliable information is needed to determine whether these processes

coupled with process mining technology enable a new form of auditing that will dramatically change the role of auditors: Auditing 2.0.

PROCESS MINING

The systematic, reliable, and trustworthy recording of events, known as *business provenance*, is essential to

IEEE Computer, vol. 43, no. 3, pp. 90-93, Mar. 2010

Focus Article

Replaying history on process models for conformance checking and performance analysis

Wil van der Aalst, Arya Adriansyah and Boudewijn van Dongen

Process mining techniques use event data to *discover* process models, to *check the conformance* of predefined process models, and to *extend* such models with information about bottlenecks, decisions, and resource usage. These techniques are driven by observed events rather than hand-made models. Event logs are used to learn and enrich process models. By replaying history using the model, it is possible to establish a *precise relationship between events and model elements*. This relationship can be used to check conformance and to analyze performance. For example, it is possible to diagnose deviations from the modeled behavior. The severity of each deviation can be quantified. Moreover, the relationship established during replay and the timestamps in the event log can be combined to show bottlenecks. These examples illustrate the importance of maintaining a proper alignment between event log and process model. Therefore, we elaborate on the realization of such alignments and their application to conformance checking and performance analysis.

How to cite this article:
WIREs Data Mining Knowl Discov 2012, 2: 182–192 doi: 10.1002/widm.1045

More Information

IEEE Task Force on Process Mining

- ProM Software: prom.sourceforge.net
- Process mining: www.processmining.org
- ProM 5 series nightly builds: prom.win.tue.nl/tools/prom/nightly5/
- ProM 6 series nightly builds: prom.win.tue.nl/tools/prom/nightly/
- Converting logs (MXML-based) promimport.sourceforge.net
- XES: www.xes-standard.org and www.openxes.org
- Papers et al.: vdaalst.com
- IEEE Task Force on Process Mining: www.win.tue.nl/ieeetfpm/