No Knowledge Without Processes

Process Mining as a Tool to Find Out What People and Organizations Really Do

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

TU/e Technische Universiteit Eindhoven University of Technology

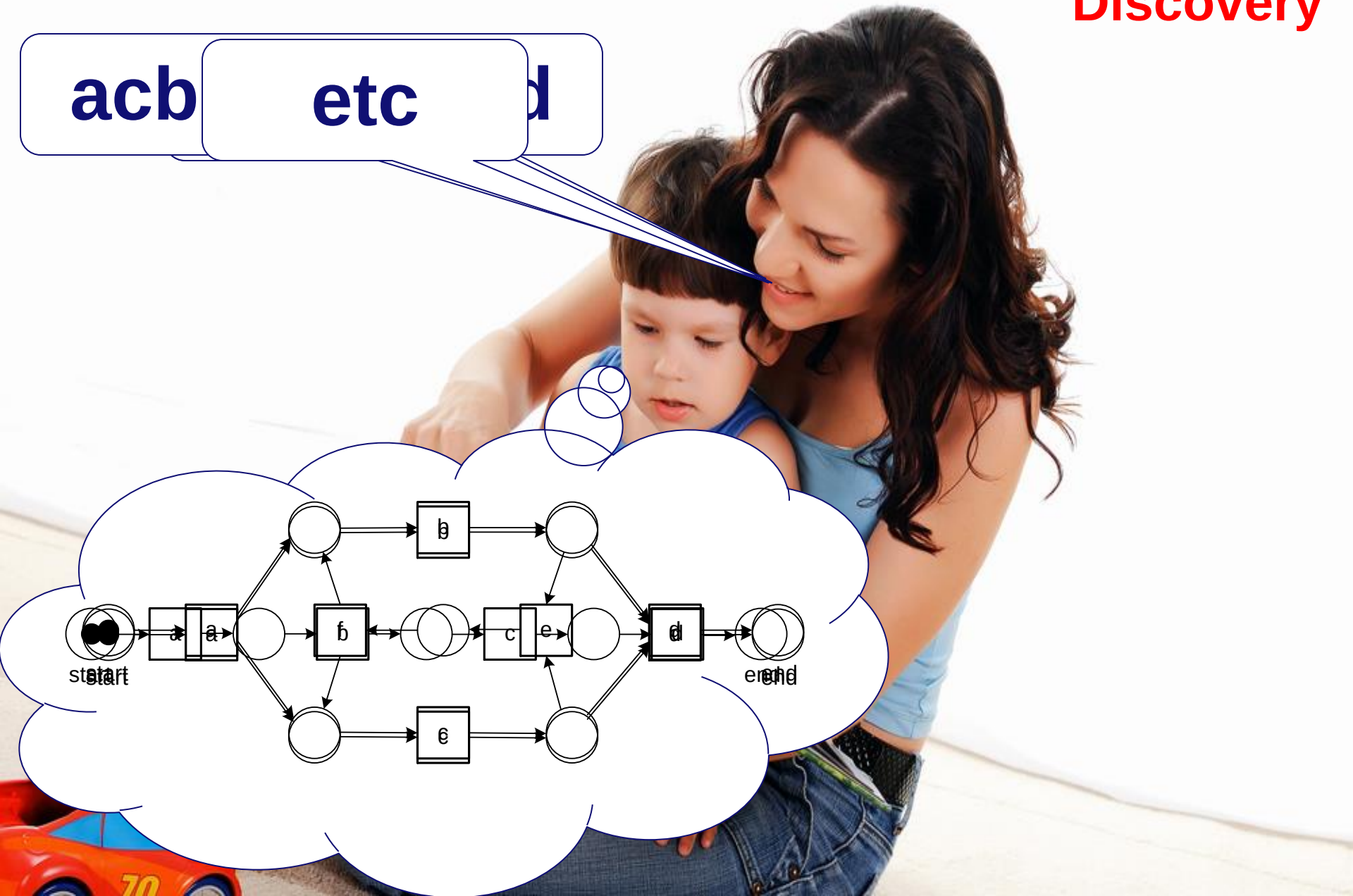
Where innovation starts


Technische Universiteit
Eindhoven
University of Technology
Data Science Center Eindhoven
process mining
intro

Process discovery

www.olifantenpaadjes.nl

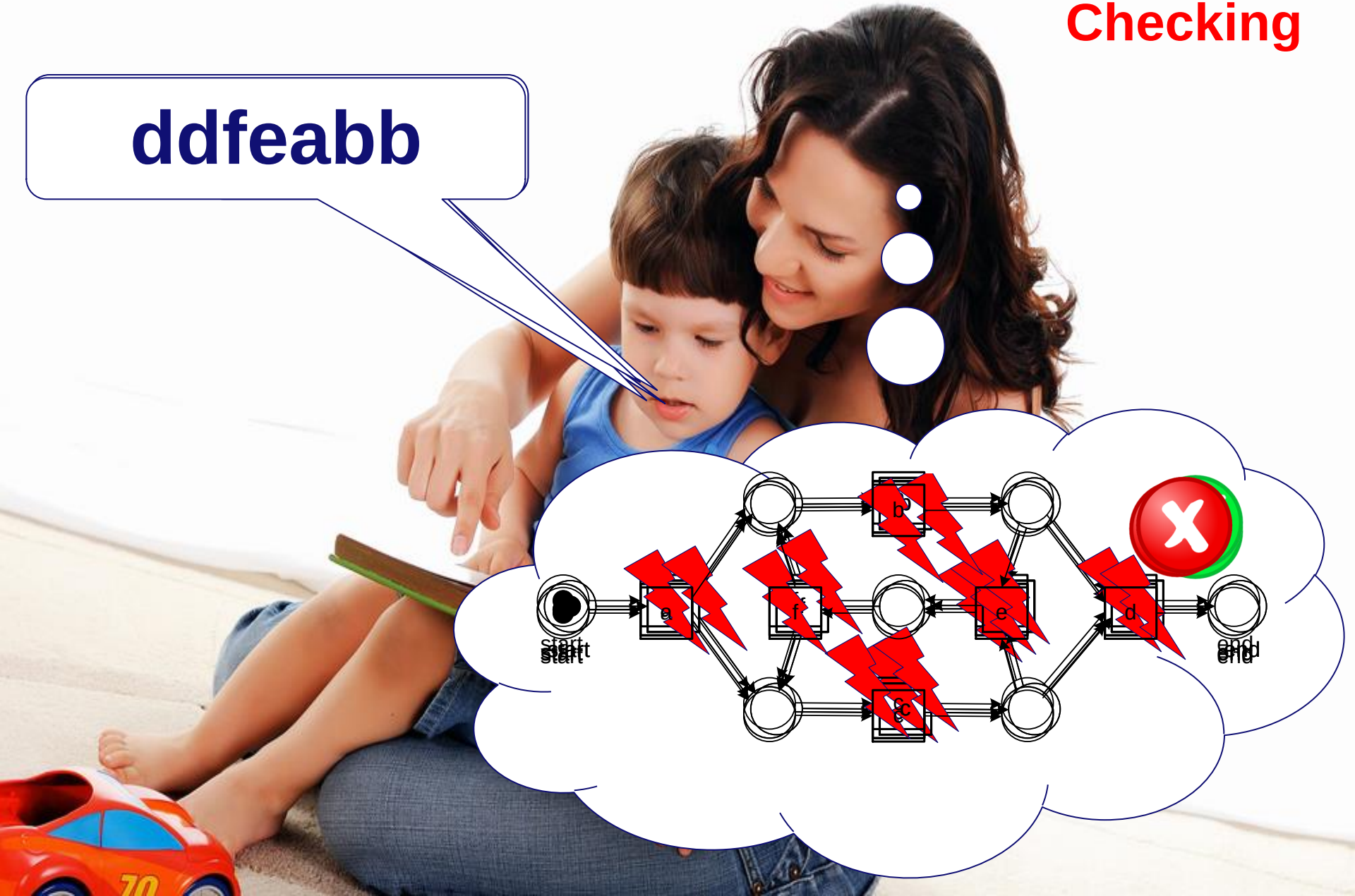
Process Discovery



Conformance checking

Conformance Checking

ddfeabb

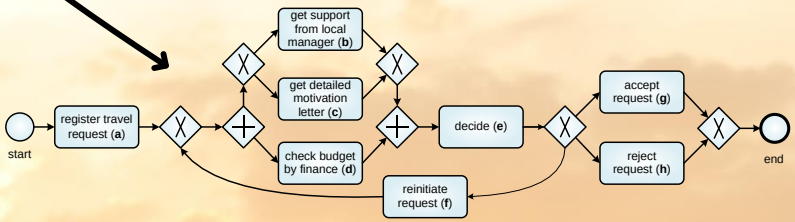


Case	Activity	Timestamp	Resource
432	register travel request (a)	18-3-2014:9.15	John
432	get support from local manager (b)	18-3-2014:9.25	Mary
432	check budget by finance (d)	19-3-2014:8.55	John
432	decide (e)	19-3-2014:9.36	Sue
432	accept request (g)	19-3-2014:9.48	Mary

Play-In

Replay

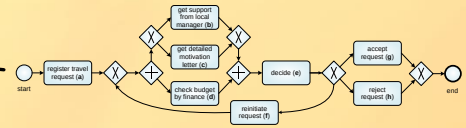
Play-Out



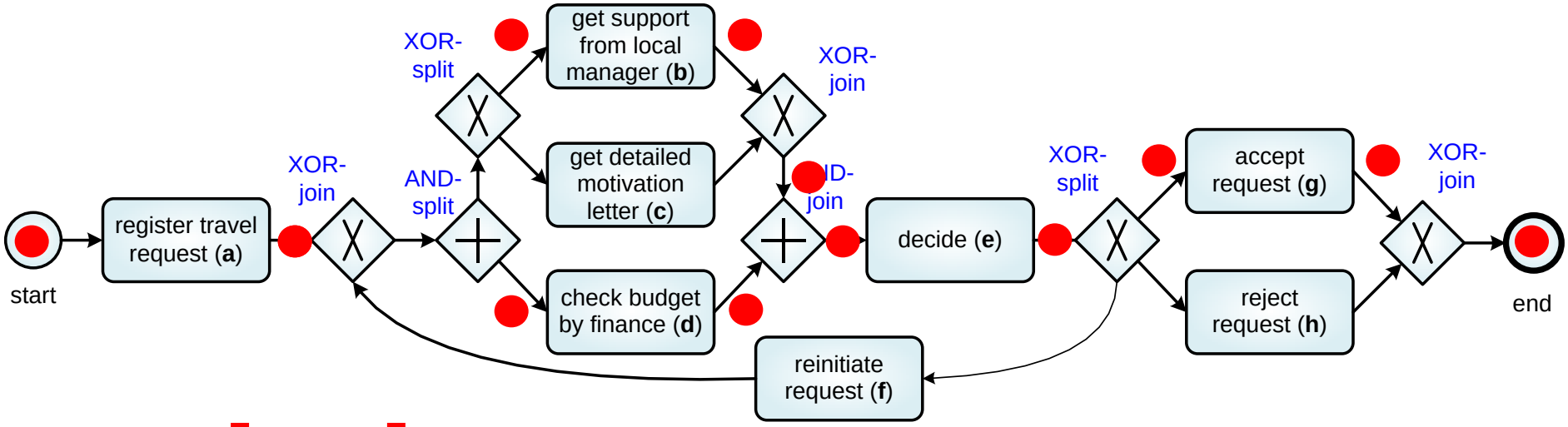
Let's play

Play-Out

Case	Activity	Timestamp	Resource
432	register travel request (a)	18-3-2014:9.15	John
432	get support from local manager (b)	18-3-2014:9.25	Mary
432	check budget by finance (d)	19-3-2014:8.55	John
432	decide (e)	19-3-2014:9.36	Sue
432	accept request (g)	19-3-2014:9.48	Mary



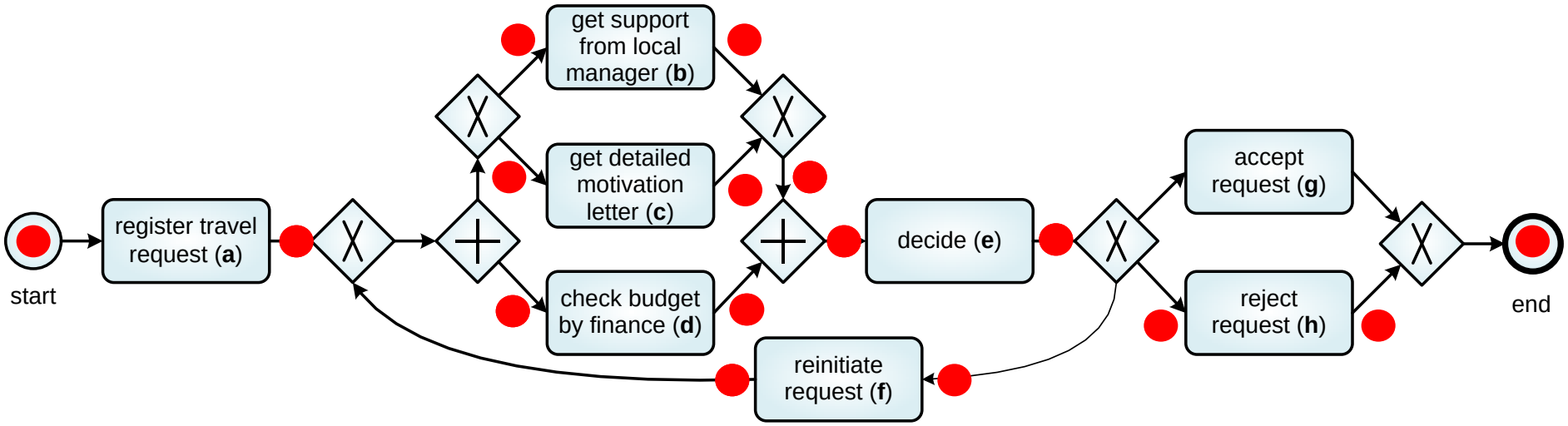
Play Out: A possible scenario



a b d e g

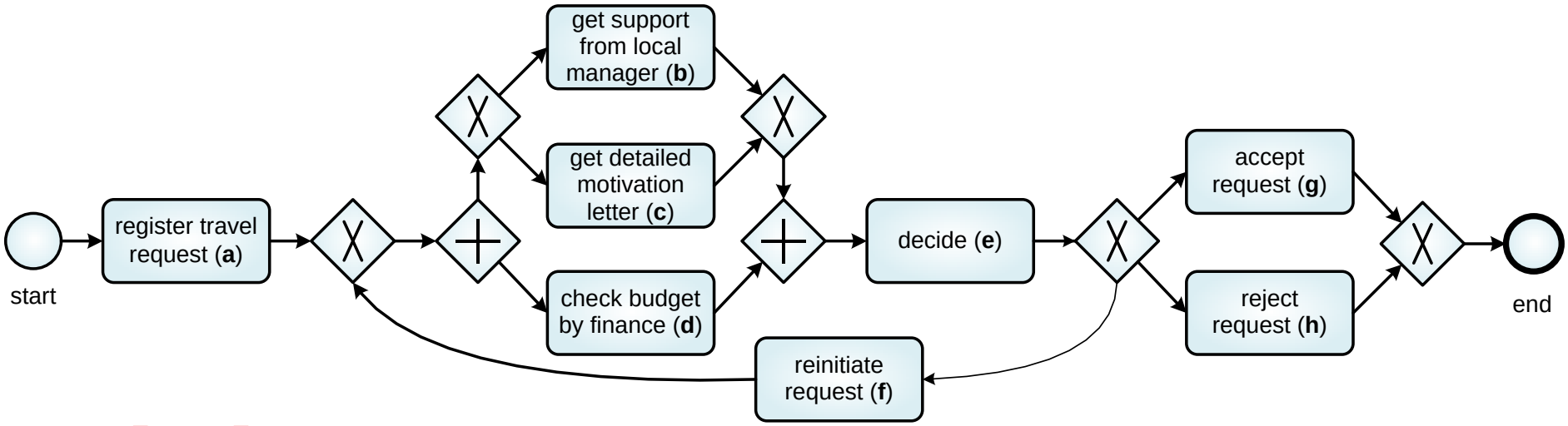
Case	Activity	Timestamp	Resource
432	register travel request (a)	18-3-2014:9.15	John
432	get support from local manager (b)	18-3-2014:9.25	Mary
432	check budget by finance (d)	19-3-2014:8.55	John
432	decide (e)	19-3-2014:9.36	Sue
432	accept request (g)	19-3-2014:9.48	Mary

Play Out: Another scenario



a d c e f b d e h

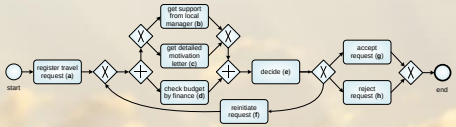
Play Out: Process model allows for many more scenarios



abdeg
adcefcdefbdefbdeg
abcefbdeg
acdefcdefbdeg
abdeg
acdefcdefbdeh
acbefbdeg
adceh
adcefcdefbdefbdeg

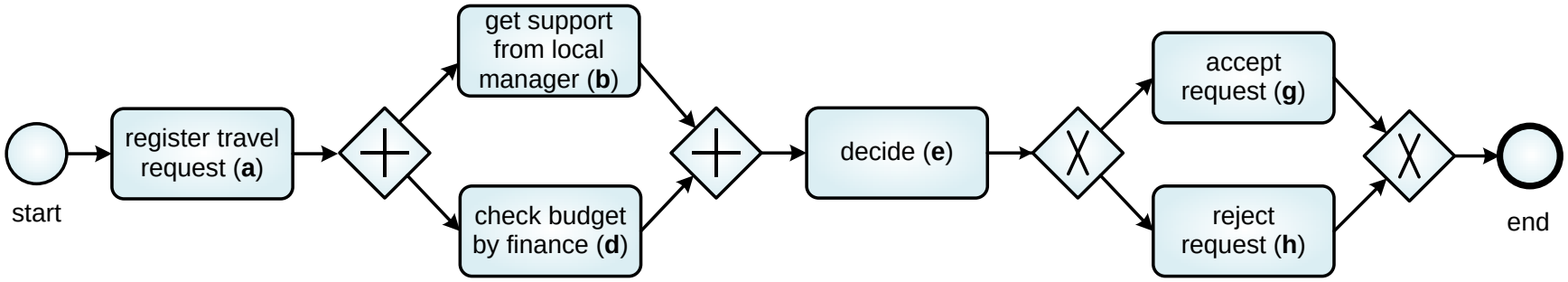
Play-In

Case	Activity	Timestamp	Resource
432	register travel request (a)	18-3-2014:9.15	John
432	get support from local manager (b)	18-3-2014:9.25	Mary
432	check budget by finance (d)	19-3-2014:8.55	John
432	decide (e)	19-3-2014:9.36	Sue
432	accept request (g)	19-3-2014:9.48	Mary



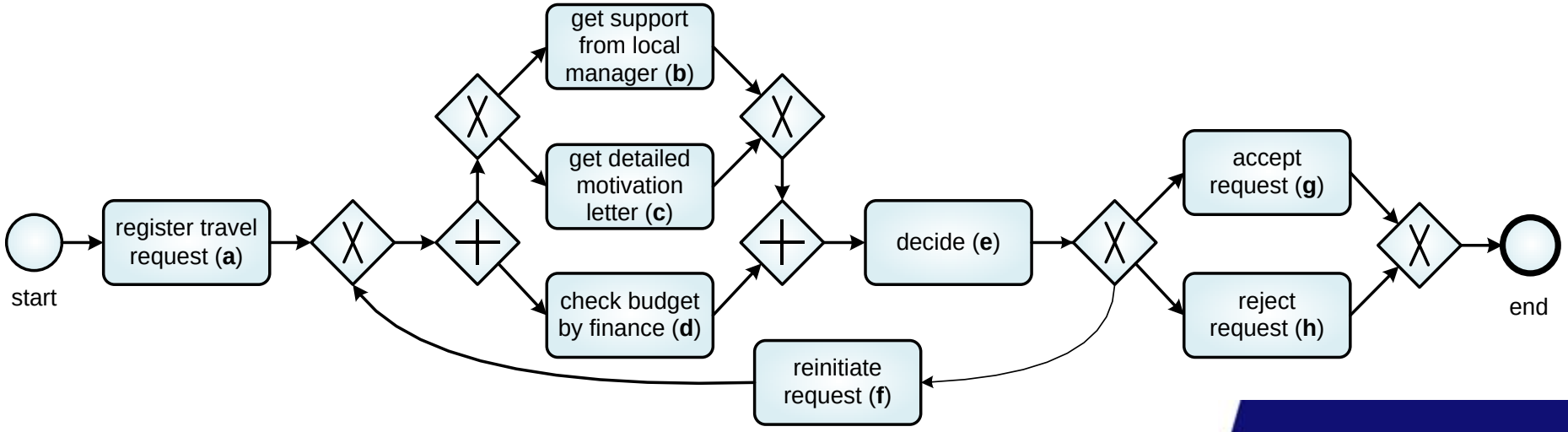
Play In:
Simple process allowing for 4 traces

abdeg adbeg adbeg
abdeh abdeg abdeh
abdeh abdeh adbeh
adbeh abdeh adbeh adbeh



Play In:
Process allowing for more traces

abdeg adcefcdefbdefbdeg adcefcdefbdefbdeg adbeh acbefbdeg adbeh abcefbdefcdefbdefbdeg adcefcdefbdefbdefbdeg acbefbdeh adcefcdefbdefbdeh adceh adbeh adcefcdefbdeh

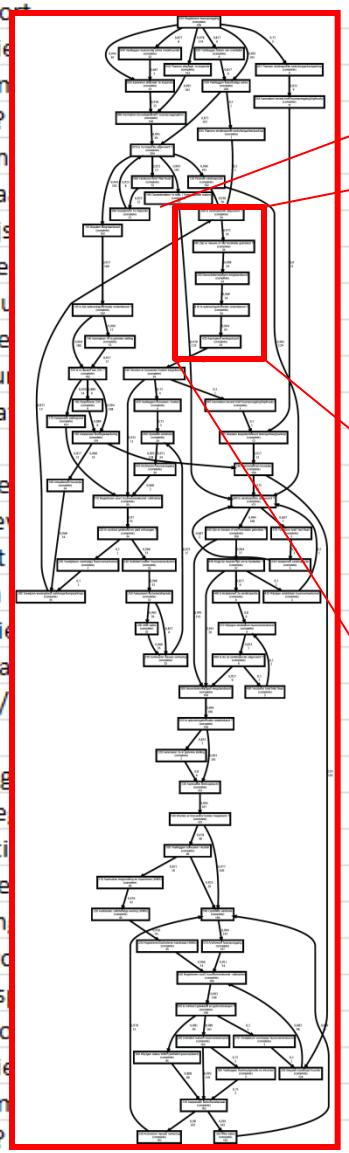


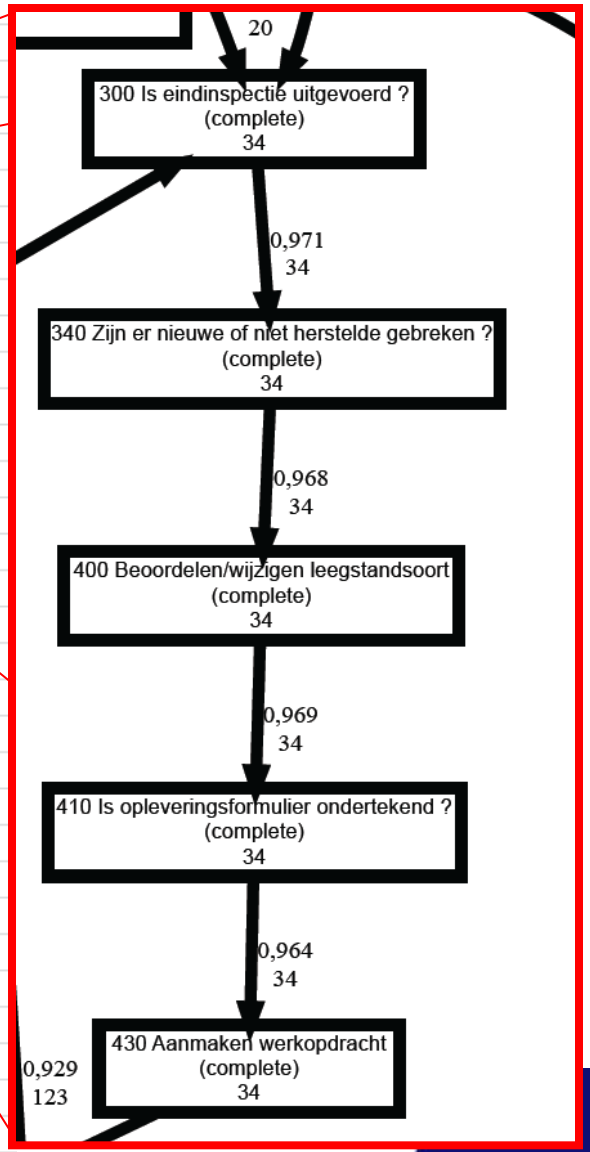
No modeling needed!

Example Process Discovery

(Dutch housing agency, 208 cases, 5987 events)

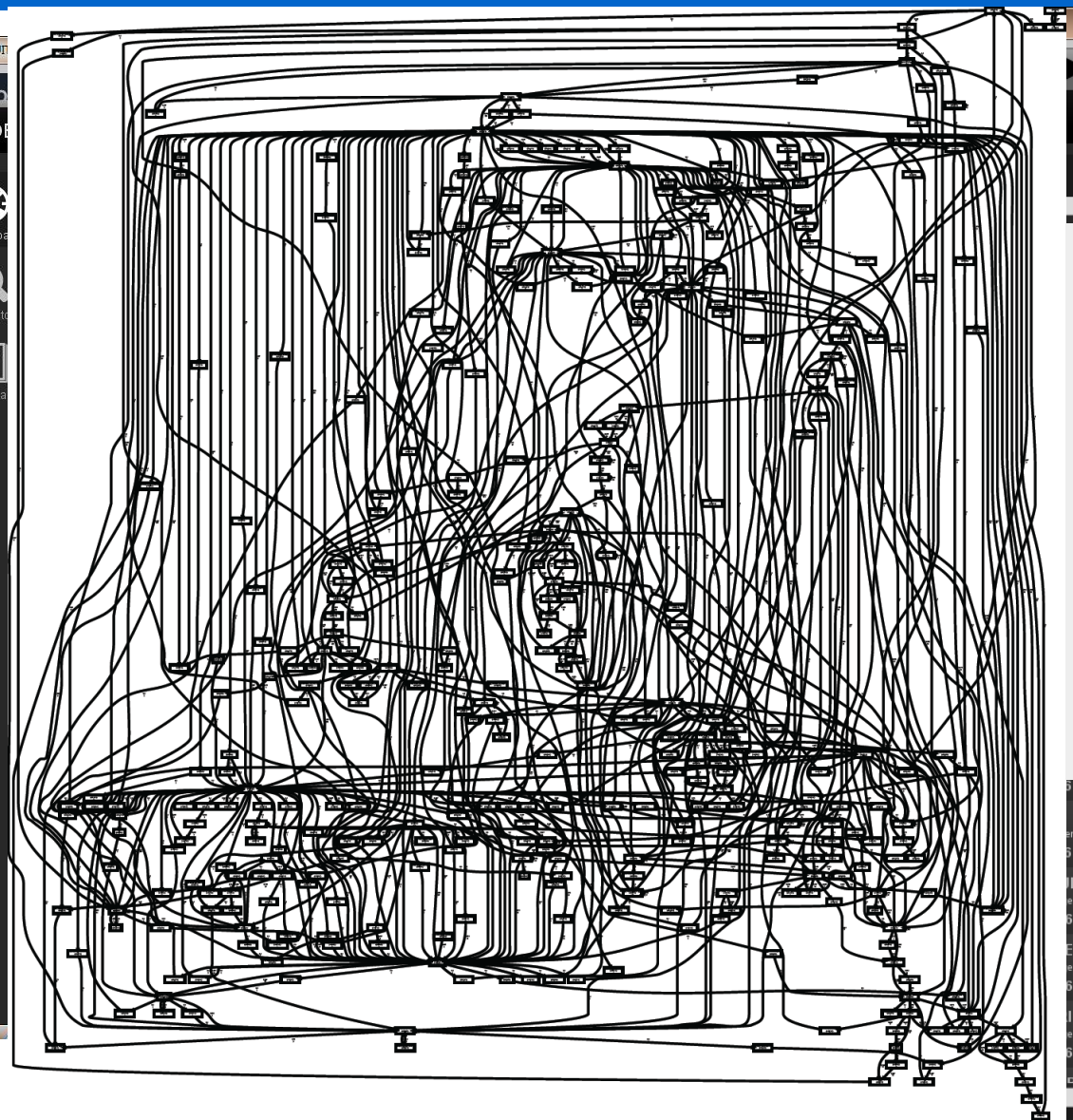
117315	110 Bepalen leegstandsoort	16.05.2007 14:06:23
117315	120 Plannen eindinspectie	16.05.2007 14:36:01
117315	130 Is het opleveringsform	23.05.2007 09:41:40
117315	150 Is er sprake van ZAV ?	23.05.2007 09:41:51
117315	170 Aanpassen plattegron	23.05.2007 11:57:18
117315	180 Aanpassen woningwa	23.05.2007 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 Huu	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	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 ins	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

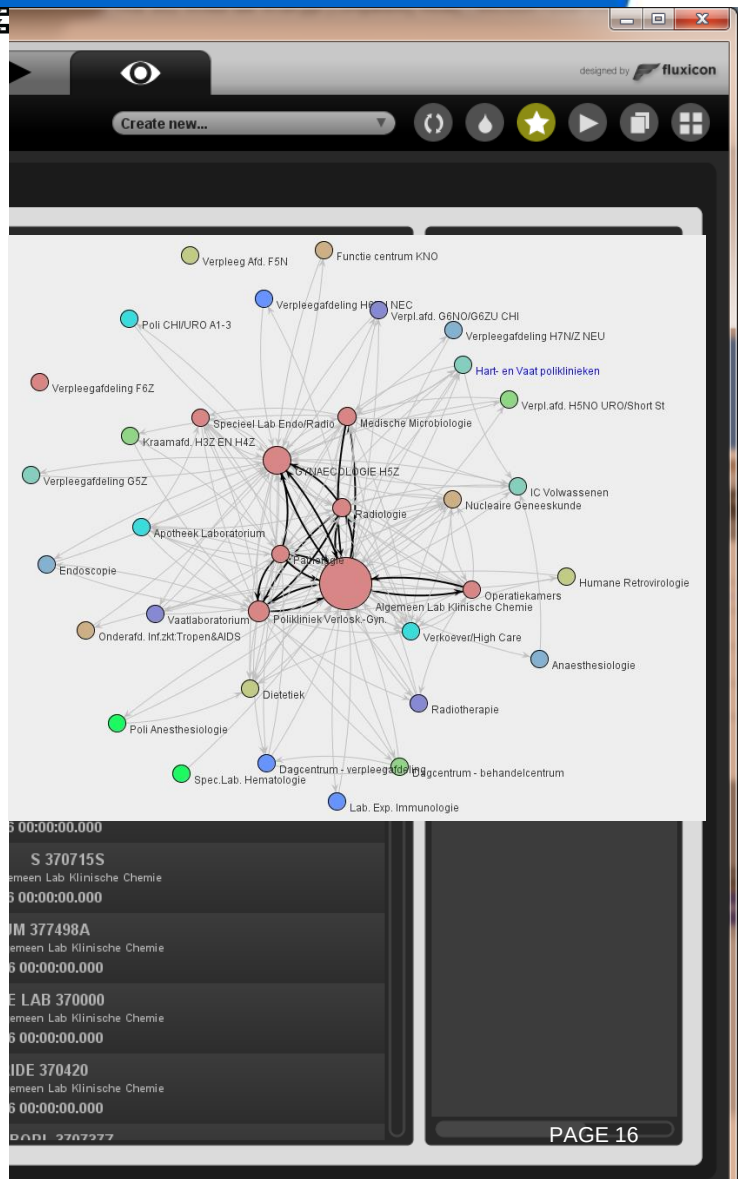




Example process discovery for hospital

(627 gynecological oncology patients, 24331 events)





Process discovery algorithms
(small selection)

automata-based learning

distributed genetic mining

heuristic mining

language-based regions

genetic mining

state-based regions

stochastic task graphs

LTL mining

ETM genetic algorithm

Inductive Miner (infrequent)

fuzzy mining

neural networks

mining block structures

hidden Markov models

α algorithm

conformal process graph

multi-phase mining

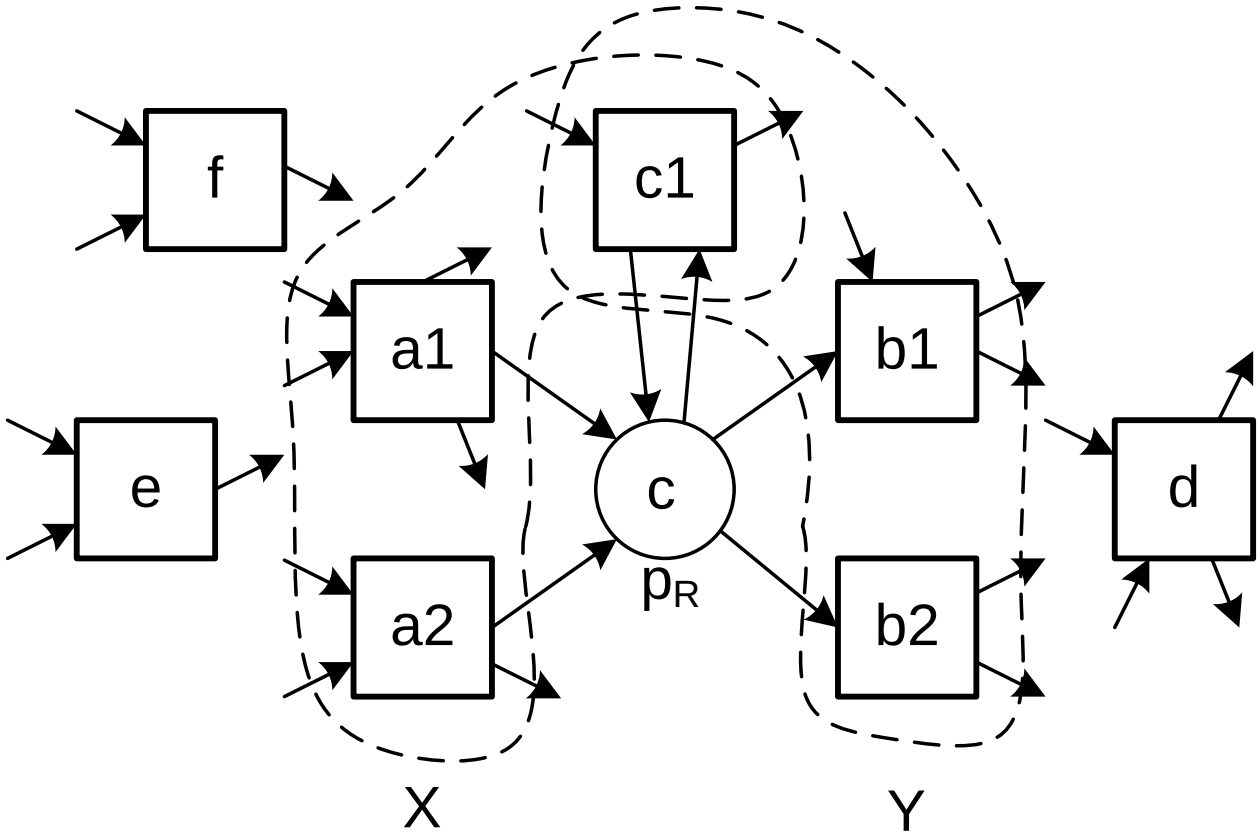
partial-order based mining

α# algorithm

ILP mining

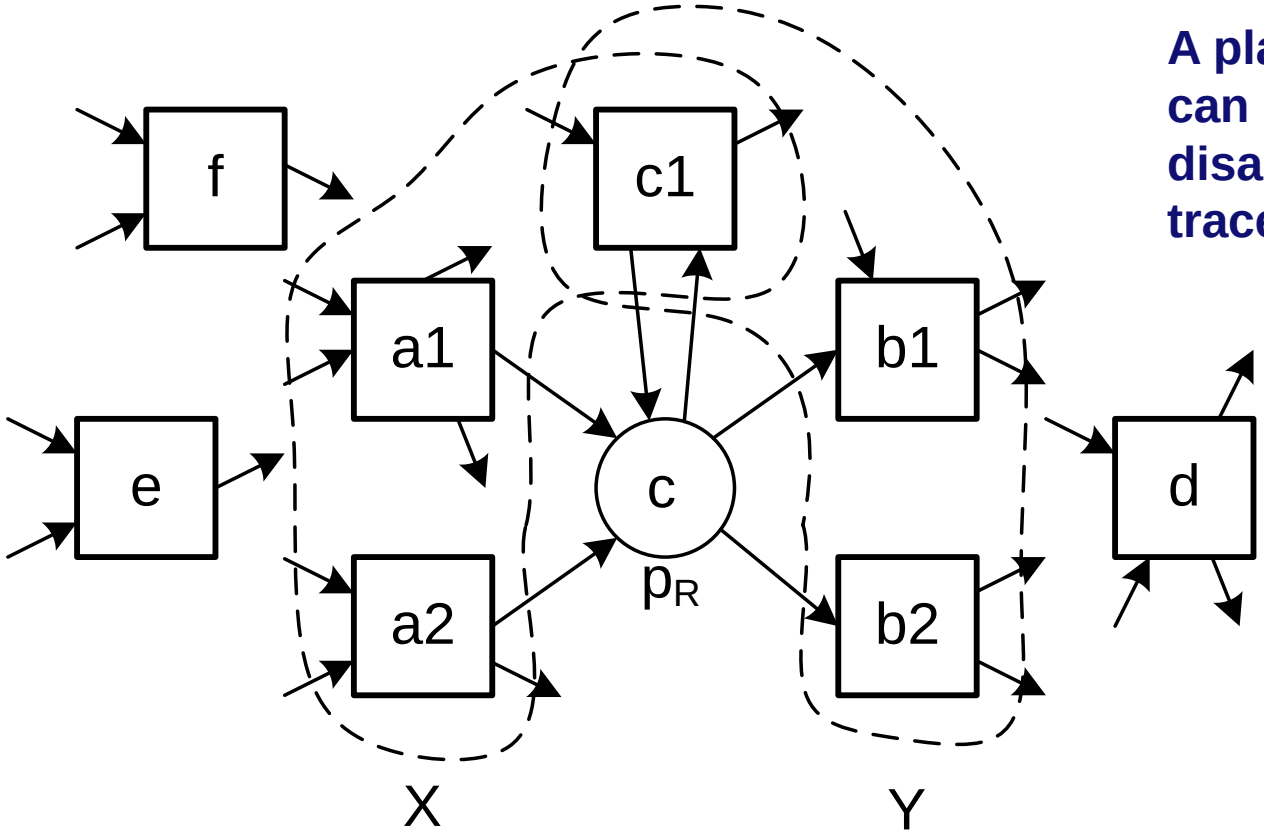
α++ algorithm

Language based regions



Region R = (X,Y,c) corresponding to place p_R: X = {a1,a2,c1} = transitions producing a token for p_R, Y = {b1,b2,c1} = transitions consuming a token from p_R, and c is the initial marking of p_R.

Basic idea: enough tokens should be present when consuming



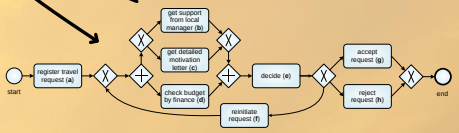
A place is feasible if it can be added without disabling any of the traces in the event log.

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

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

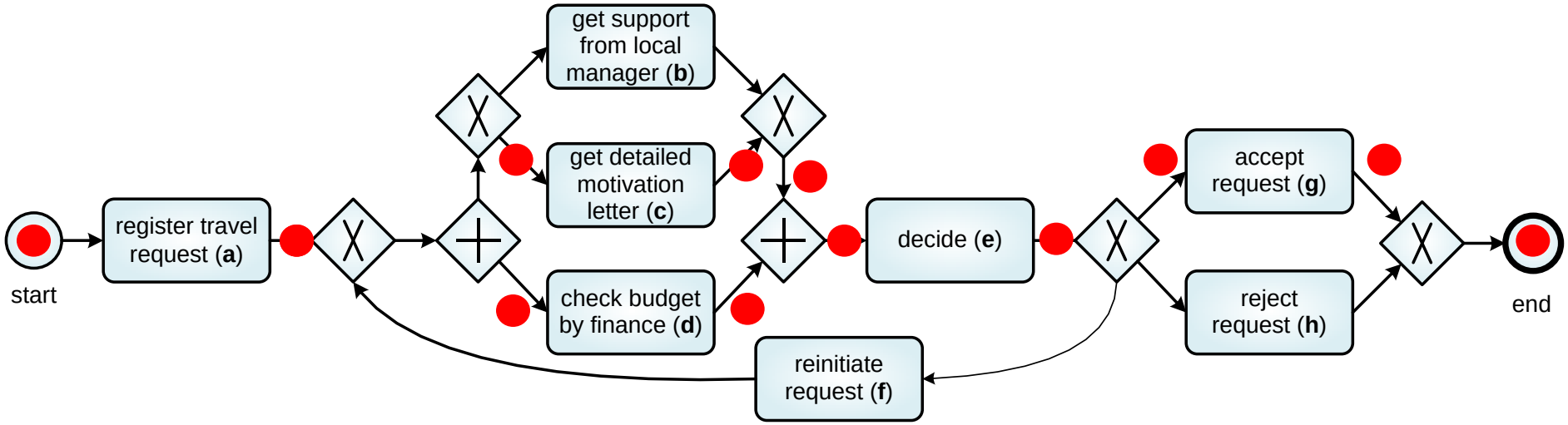
Replay

Case	Activity	Timestamp	Resource
432	register travel request (a)	18-3-2014:9.15	John
432	get support from local manager (b)	18-3-2014:9.25	Mary
432	check budget by finance (d)	19-3-2014:8.55	John
432	decide (e)	19-3-2014:9.36	Sue
432	accept request (g)	19-3-2014:9.48	Mary



Replay

a c d e g

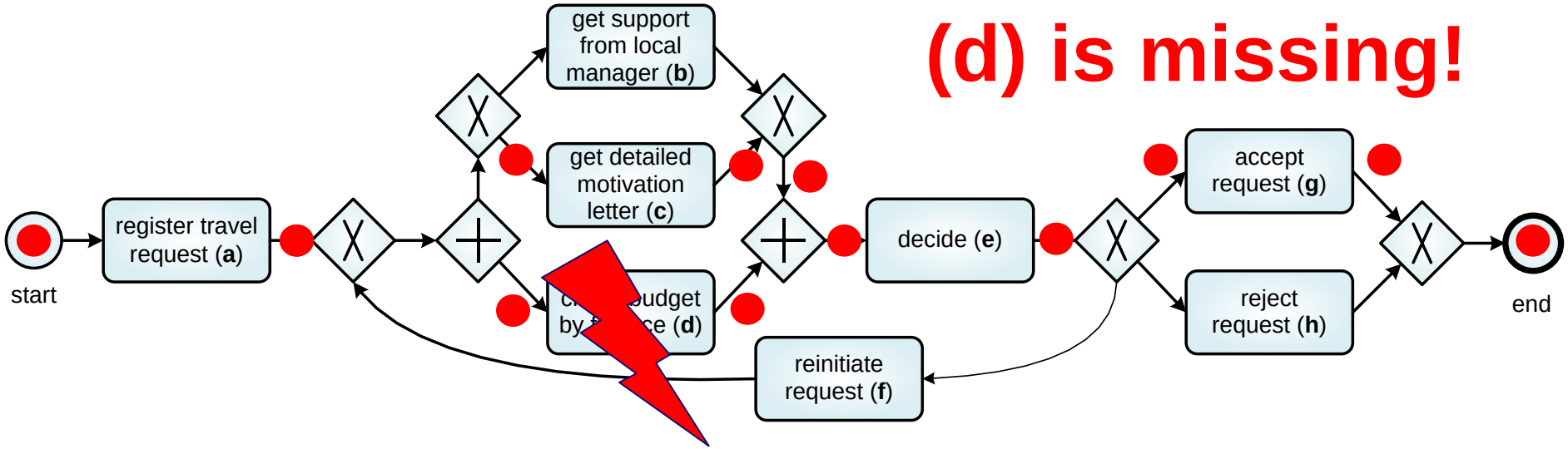


Replay

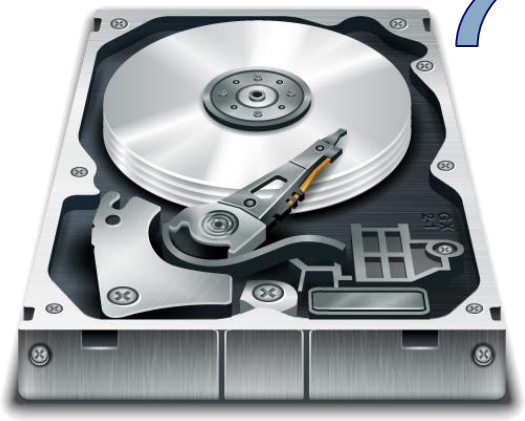
a c e g

?

check budget (d) is missing!

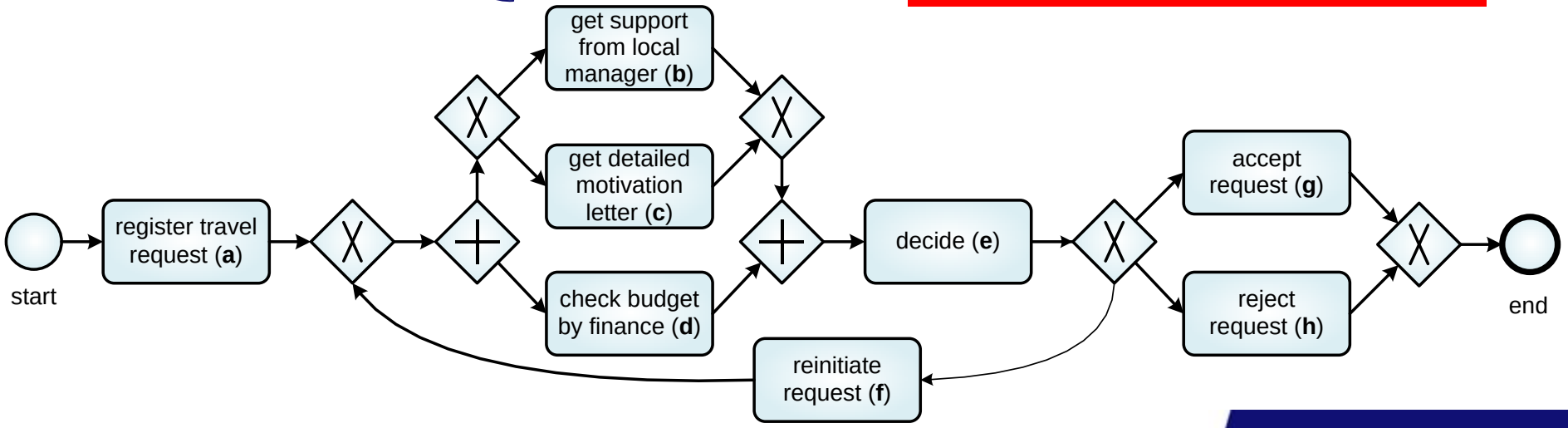


Alignments: Relating reality and model



a	c	»	e	g
a	c	d	e	g

check budget (d) did not happen but should have according to the model

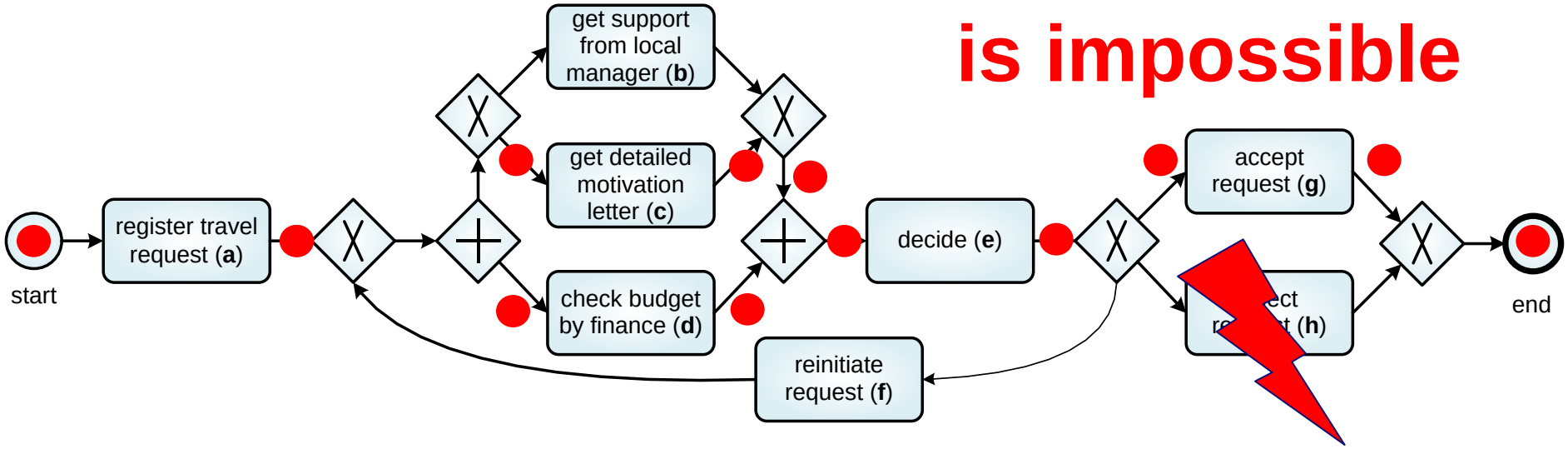


Replay

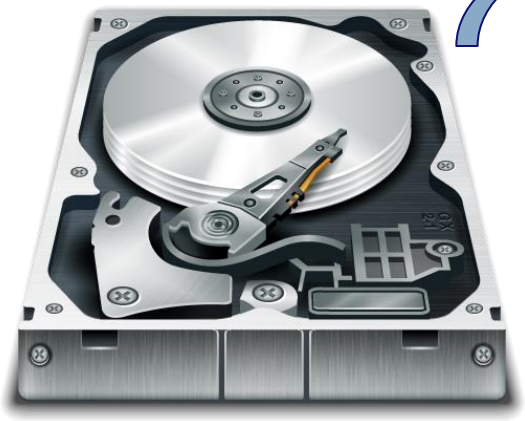
a c h d e g

?

reject request (h) is impossible

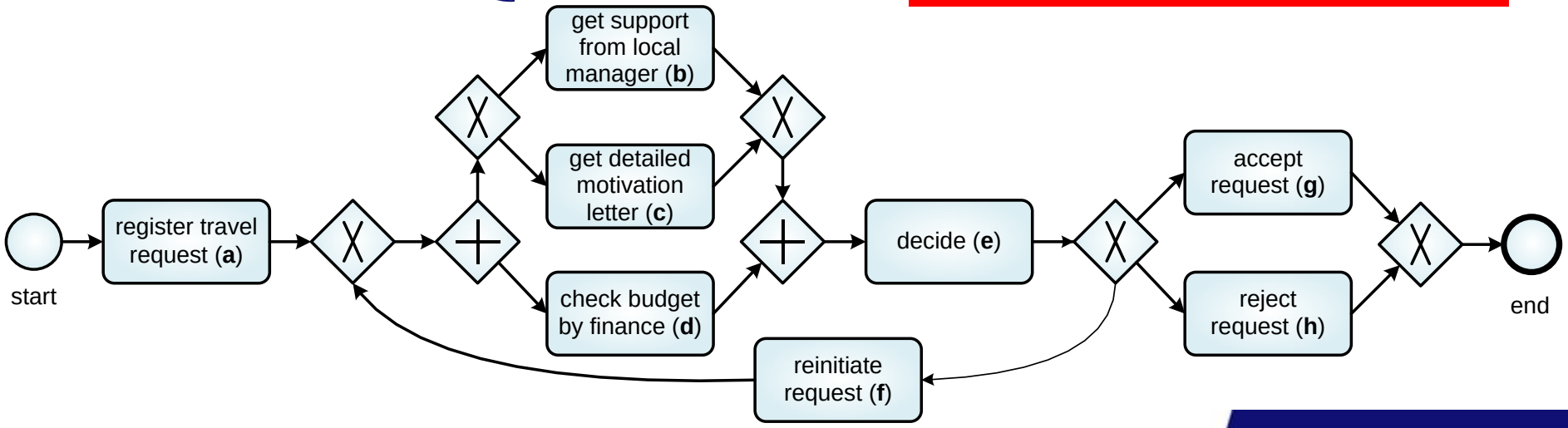


Alignments: Relating reality and model

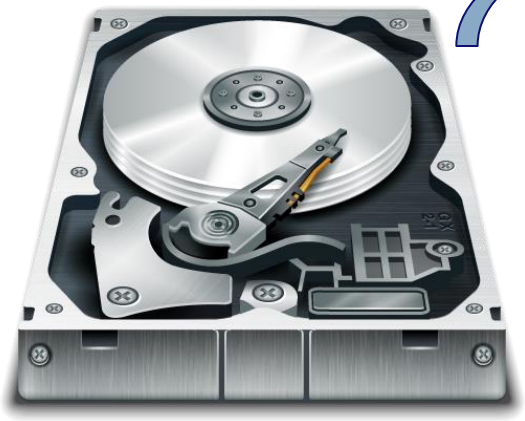


a	c	h	d	e	g
a	c	»	d	e	g

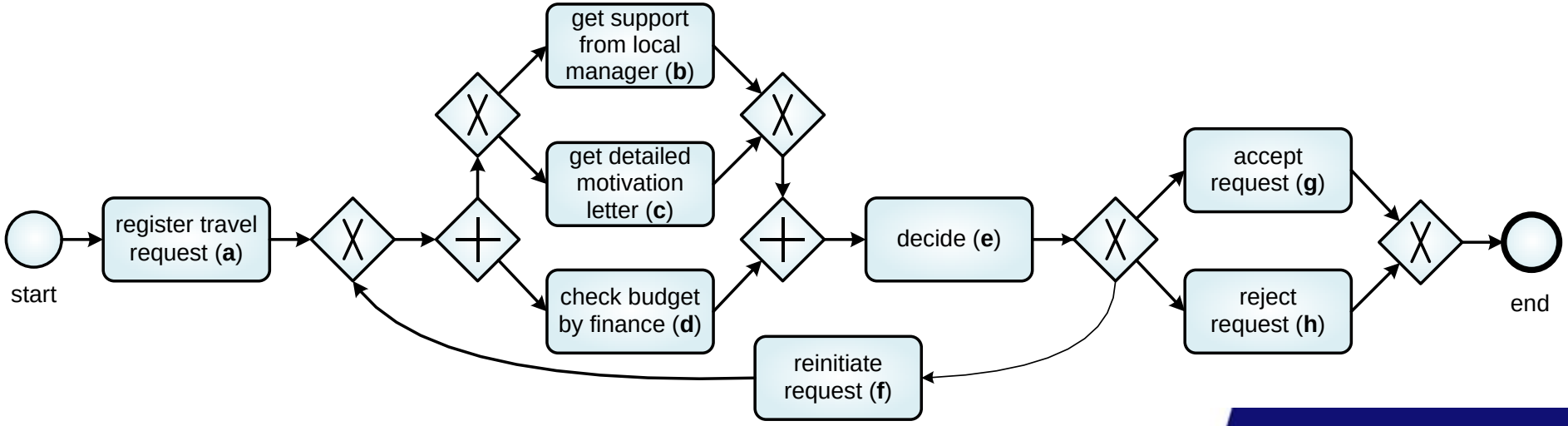
reject request (h) happened but could not happen according to the model



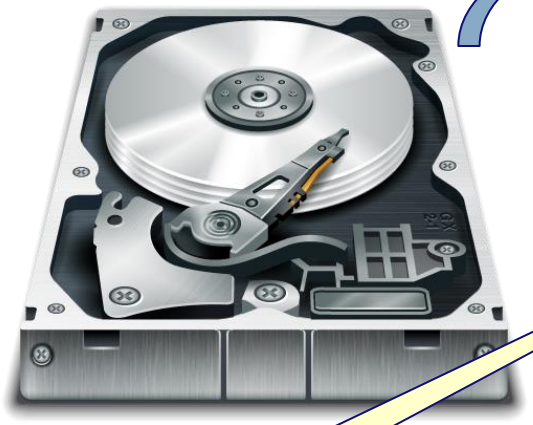
Any trace in reality can be related to a path in the model



a	c	e	f	d	d	b	c	e	h



Any trace in reality can be related to a path in the model



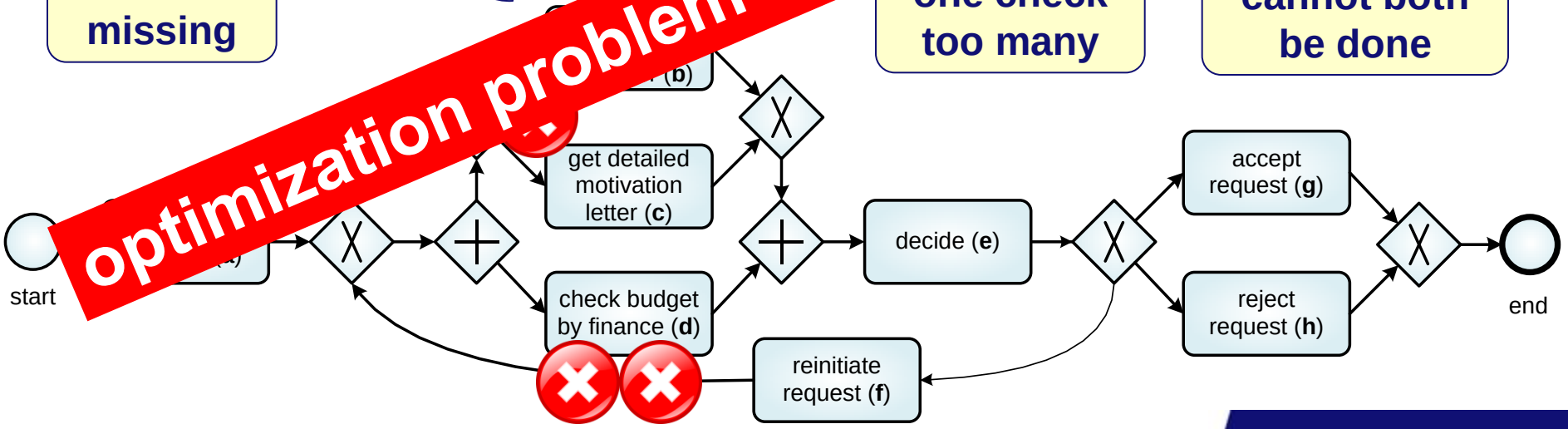
a	c	»	e	f	d	d	b	c	e	h
a	c	d	e	f	d			»	e	h

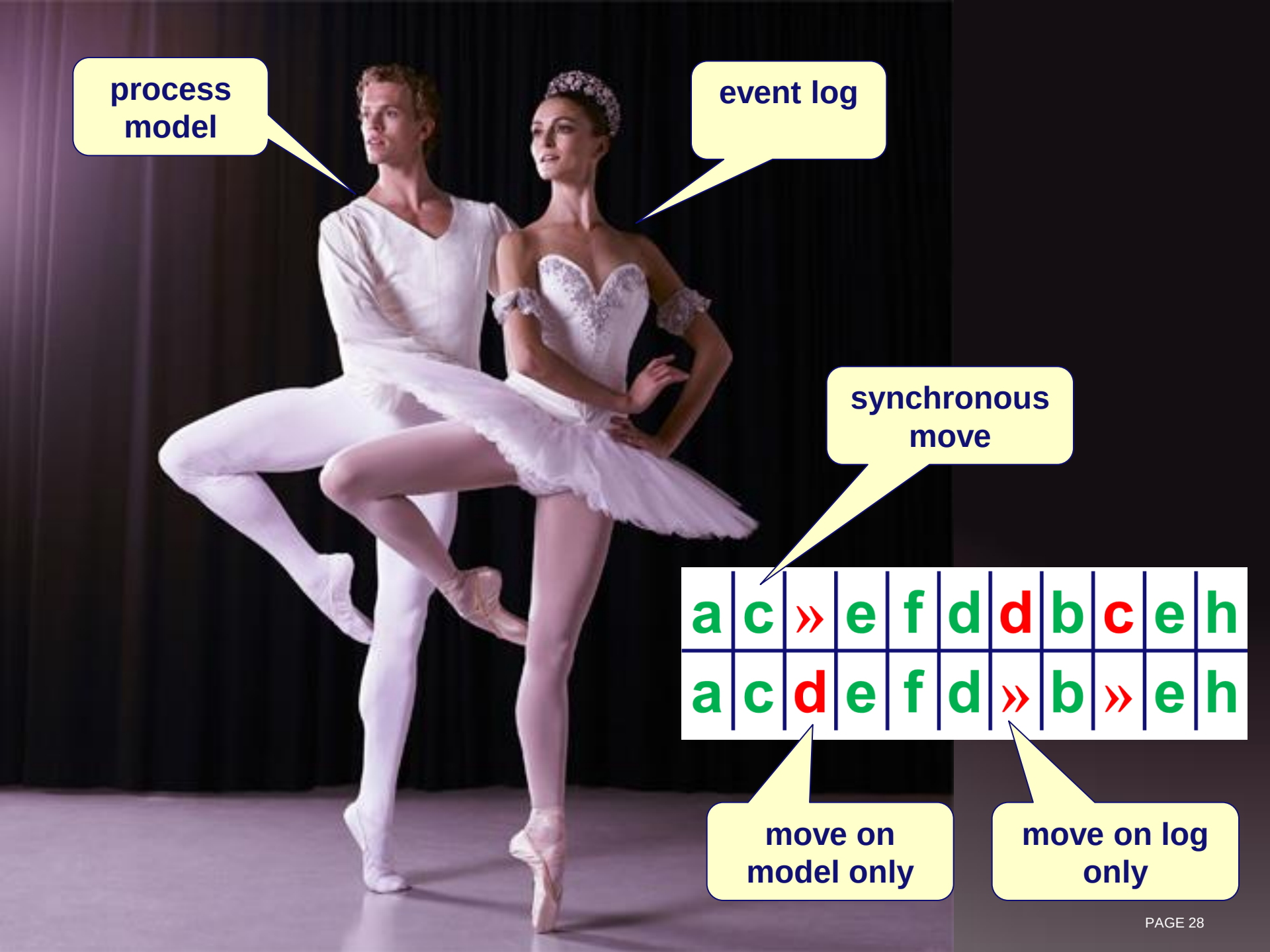
check is missing

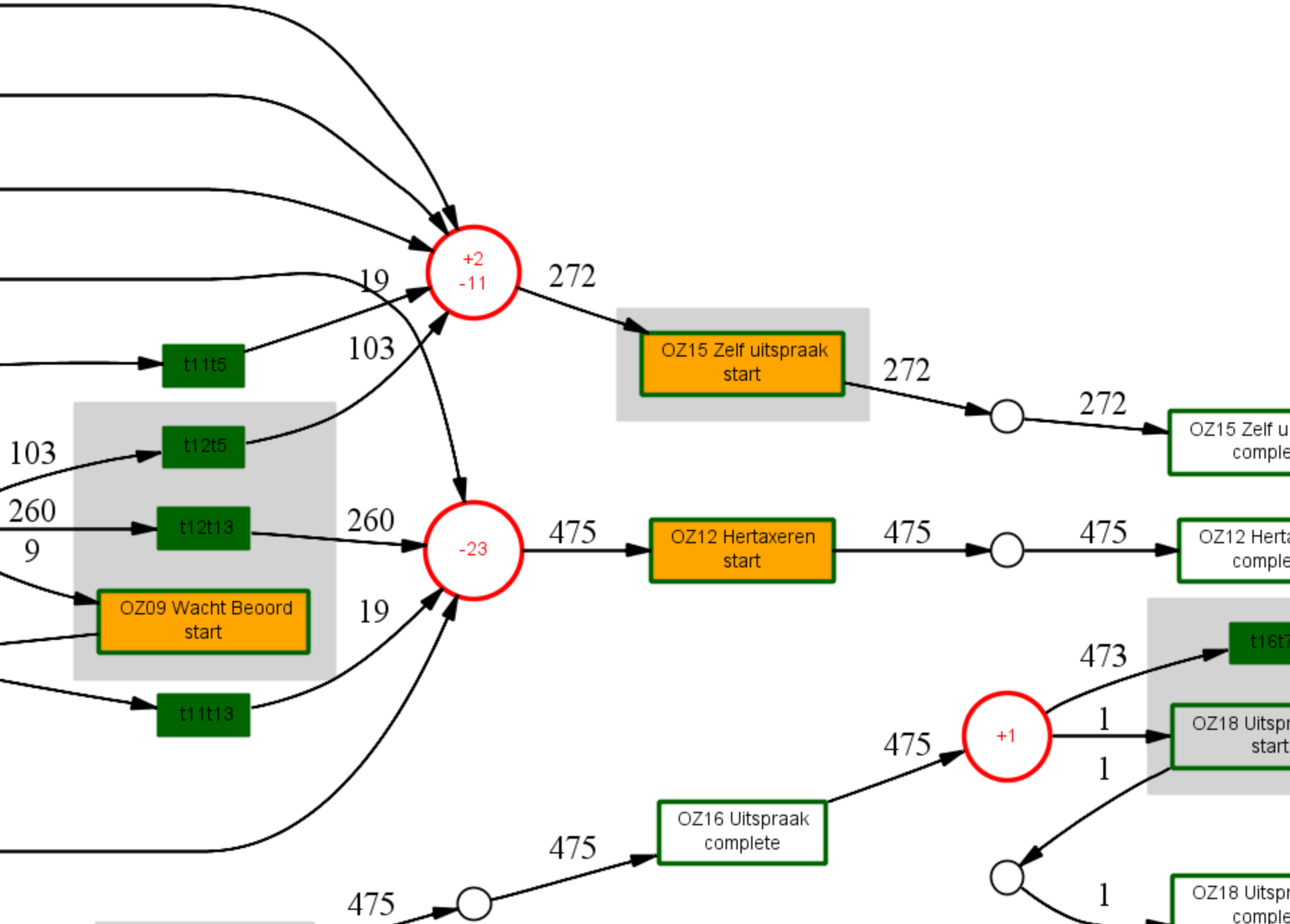
one check too many

cannot both be done

optimization problem using a cost function

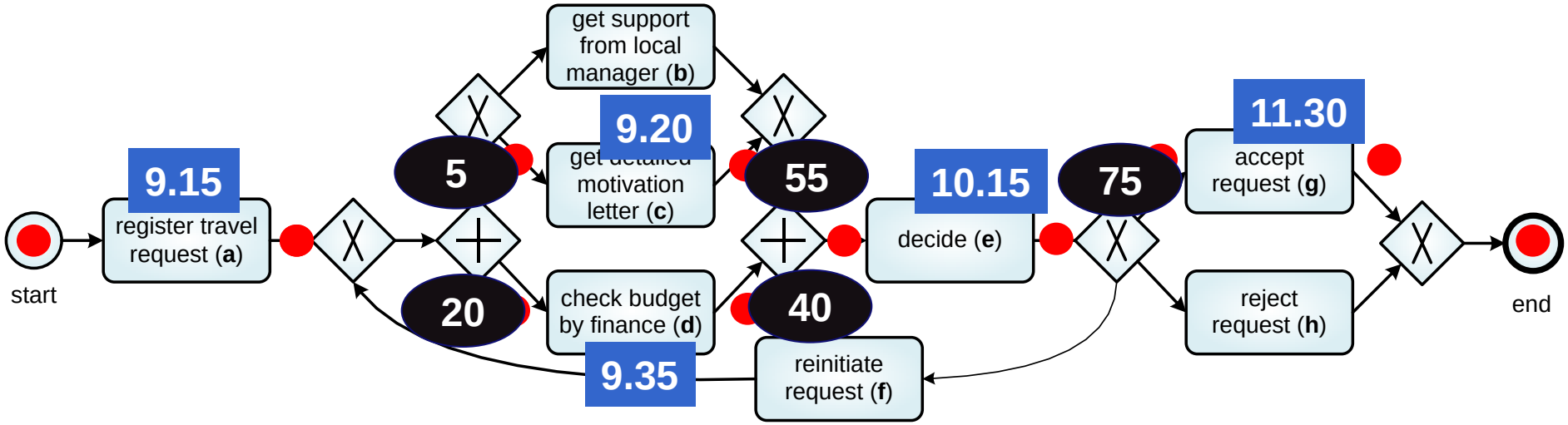



process model
event log
synchronous move
a c » e f d d b c e h
a c d e f d » b » e h
move on model only
move on log only


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

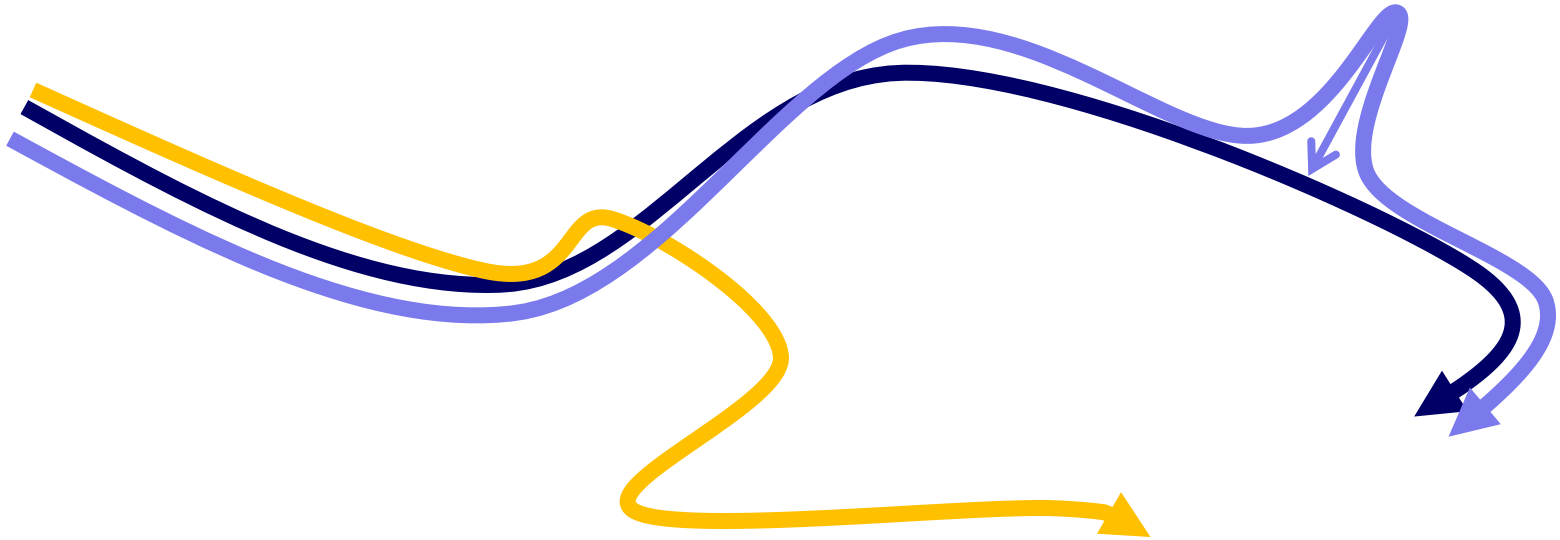
Replay with timestamps

$a^{9.15}$ $c^{9.20}$ $d^{9.35}$ $e^{10.15}$ $g^{11.30}$



Replay with timestamps for many traces

Alignments are essential!



- conformance checking to diagnose deviations
- squeezing reality into the model to do model-based analysis

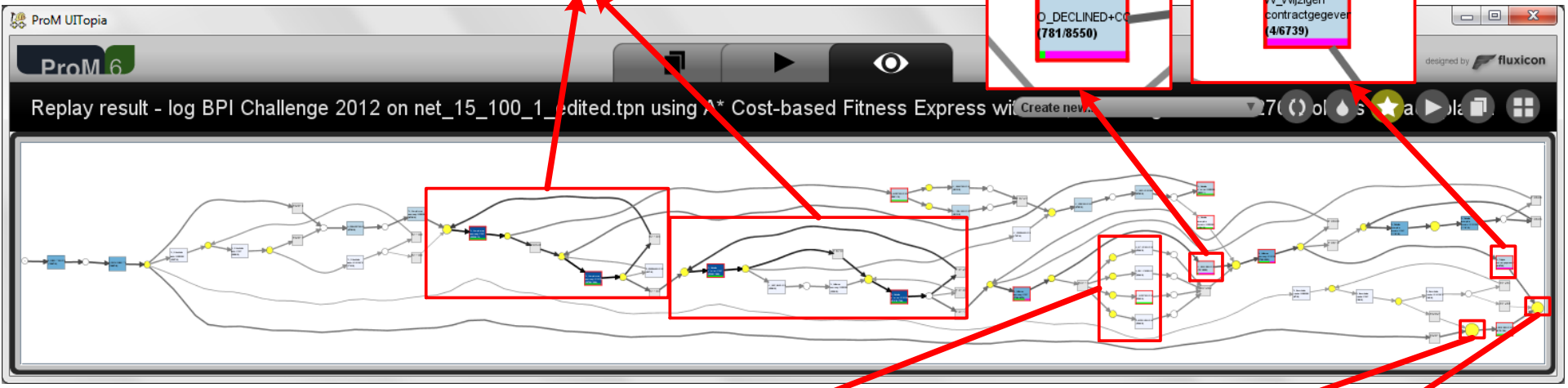
a	c	»	e	f	d	d	b	c	e	h
a	c	d	e	f	d	»	b	»	e	h

Example: BPI Challenge 2012

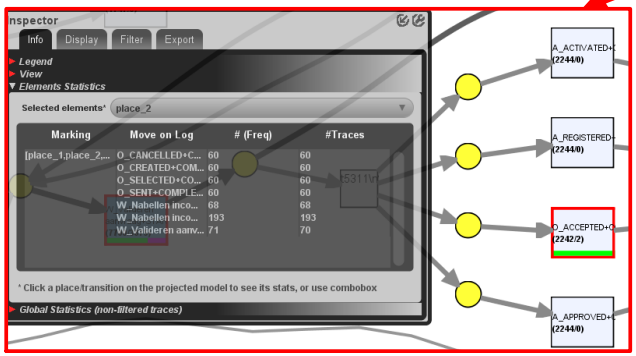
(Dutch financial institute, doi:10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f)

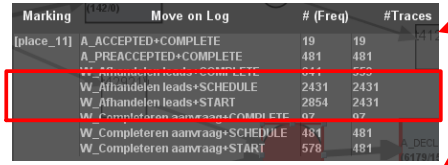
Loops of “W_Completeren aanvraag” and “W_Nabellen offertes” are often performed

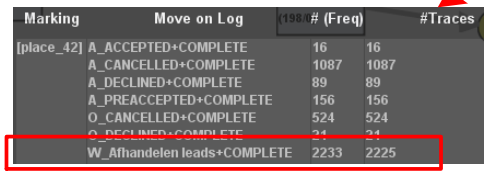
“O_DECLINED” and “W_Wijzigen contractgegevens” are often skipped



Many moves on log of “O_CANCELLED”, ”O_CREATED”, ”O_SELECTED”, “O_SENT” occurred with the same frequency value (i.e. 60) before parallel branch







Many moves on log of “W_Afhandelen leads” (> 2200 times) occurred in the end of traces

Work of Arya Adriansyah (Replay project)

Loops of “W_Completeren aanvraag” and
n offertes” are often performed

“O_DECLINED” and “W_Wijzigen
contractgegevens” are often skipped

Many moves o
“O_CANC
”O_CRE
”O_SELE
“O_SENT” o
with th
frequency va
60) before

Synchronous moves of
“Completeren aanvraag”

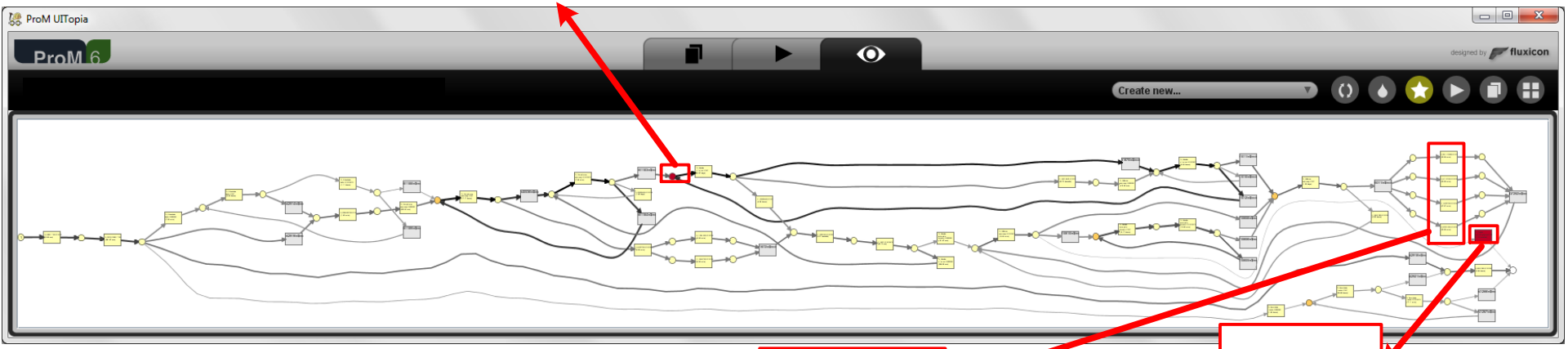
Move on log of “Completeren aanvraag”

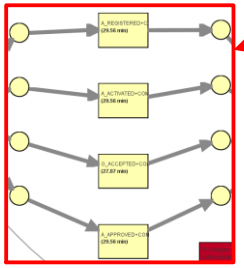
Move on log of “O_CANCELLED” and “A_CANCELLED”

Moves on model towards end of traces

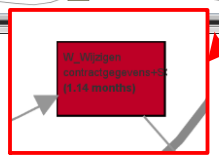
Property	Min.	Max.	Avg.	Std. Dev	Freq.
Waiting time	0.00 ms	29.78 days	2.83 days	3.30 days	24,229
Synchronization time	0.00 ms	0.00 ms	0.00 ms	0.00 ms	24,229
Sojourn time	0.00 ms	29.78 days	2.83 days	3.30 days	24,229

The average waiting time for the input place of "W_Nabellen offertes+START" is very long (2.83 days) compares to the average waiting time of other places





"O_ACCEPTED" has average sojourn time of 27.07 minutes, while "A_REGISTERED", "A_ACTIVATED", and "A_APPROVED" have average sojourn time of 29.56 minutes



Property	Min.	Max.	Avg.	Std. Dev	Freq.
Throughput time	0.00 ms	0.00 ms	0.00 ms	0.00 ms	4
Waiting time	1.55 hours	3.43 months	1.14 months	1.55 months	4
Sojourn time	1.55 hours	3.43 months	1.14 months	1.55 months	4
#Unique cases ...	4				

Activity "W_Wijzigen contractgegevens" is the bottleneck, but it occured rarely (only 4 times)

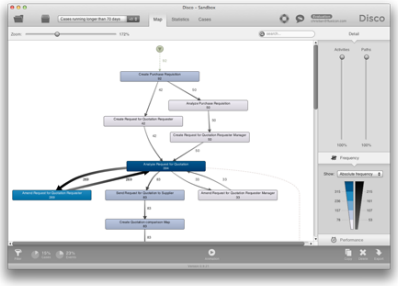
Software

ProM
process mining workbench

600+ plug-ins available covering the whole process mining spectrum

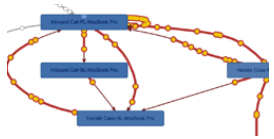
Disco
by Fluxicon.

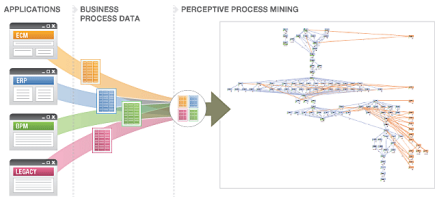
perceptivesoftware
a Lexmark company



QPR
Quality. Processes. Results.

celonis
process mining





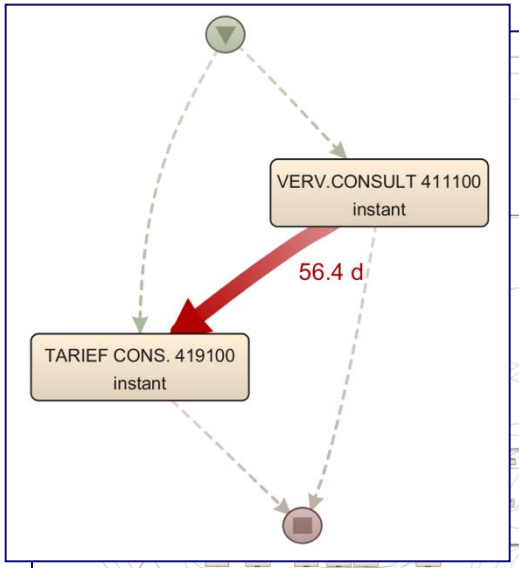
XMPRO
GET BETTER AT GETTING WORK DONE

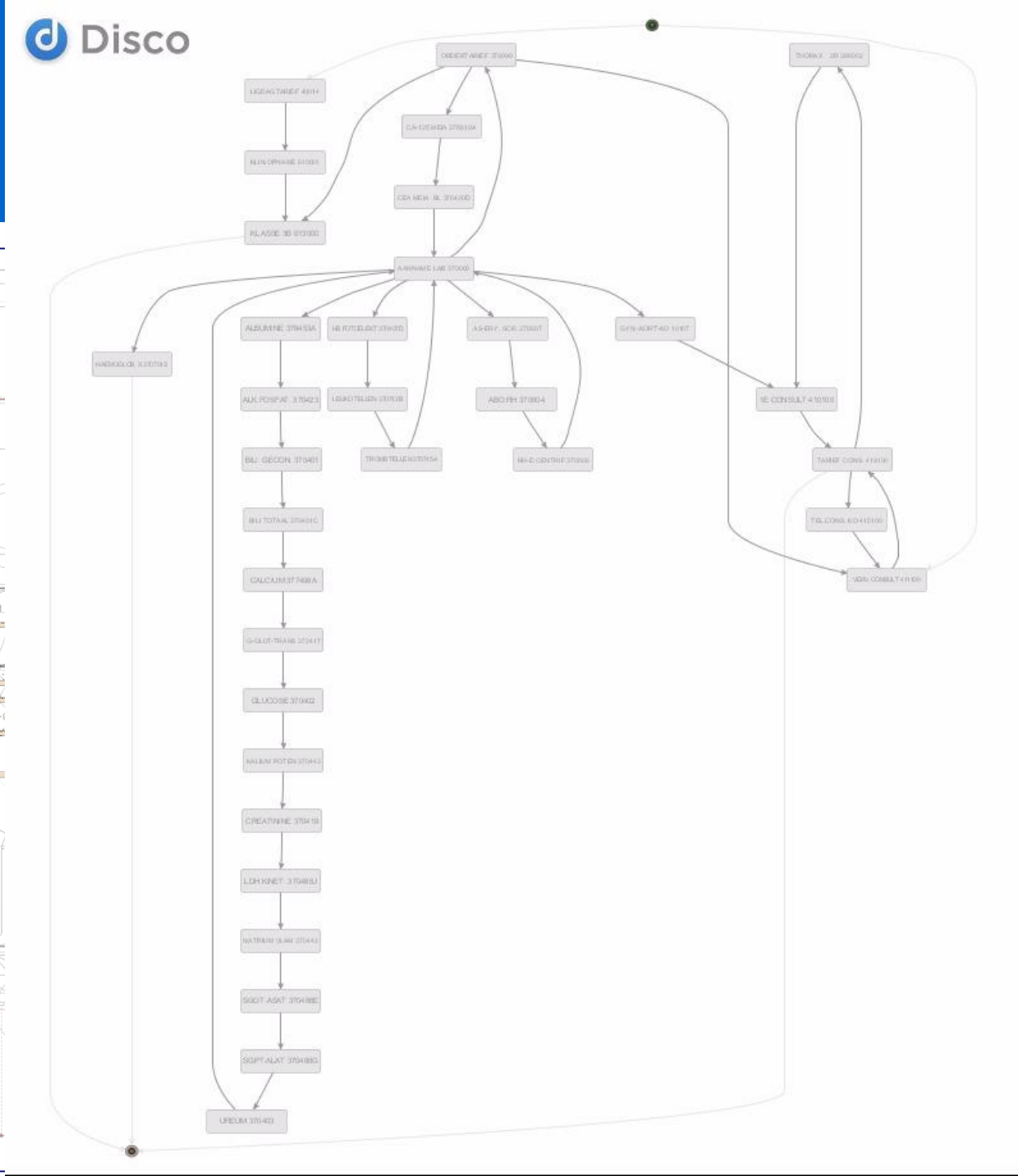
StereoLogic

ARIS
Process Performance Manager

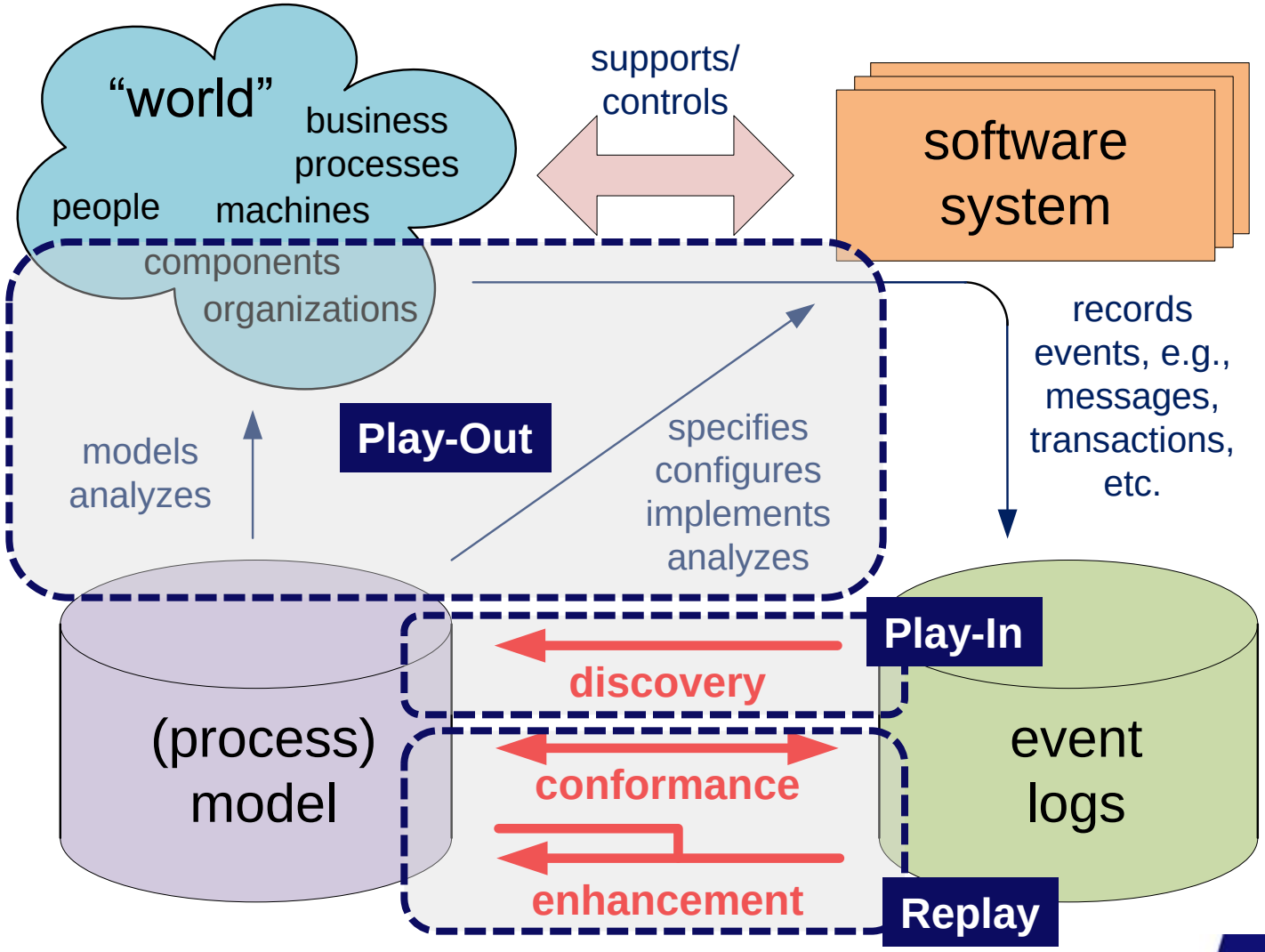
FUJITSU

Disco


VERV.CONSULT 411100
instant
56.4 d
TARIEF CONS. 419100
instant


Disco

Overview: Role of process models




Technische Universiteit
Eindhoven
University of Technology
Data Science Center Eindhoven
decomposed/distributed
process mining

Big data

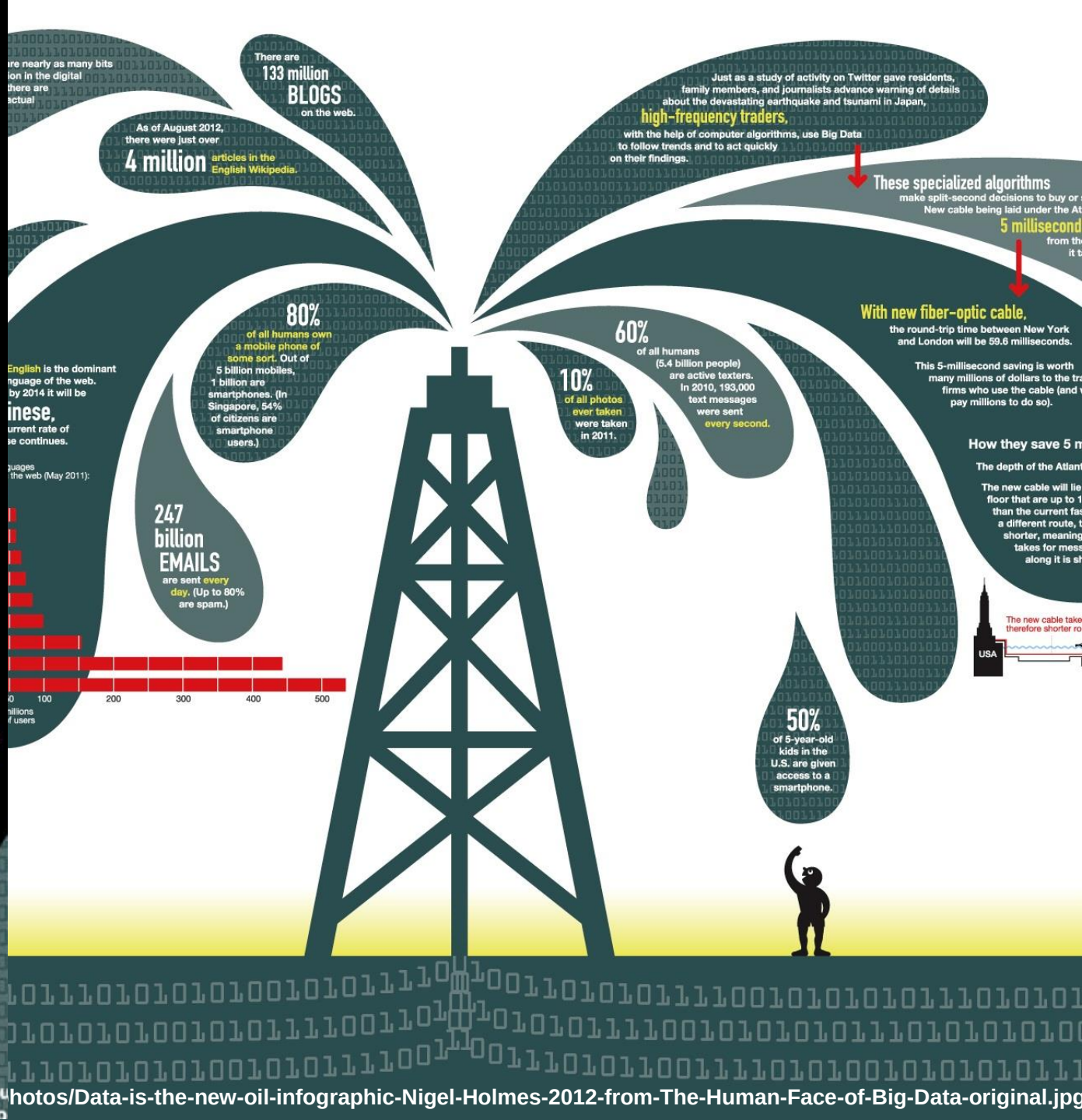
"DATA IS THE NEW OIL."

From the beginning of recorded time until 2003, we created **5 exabytes** (5 billion gigabytes) **of data.**

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

Every hou
create eno



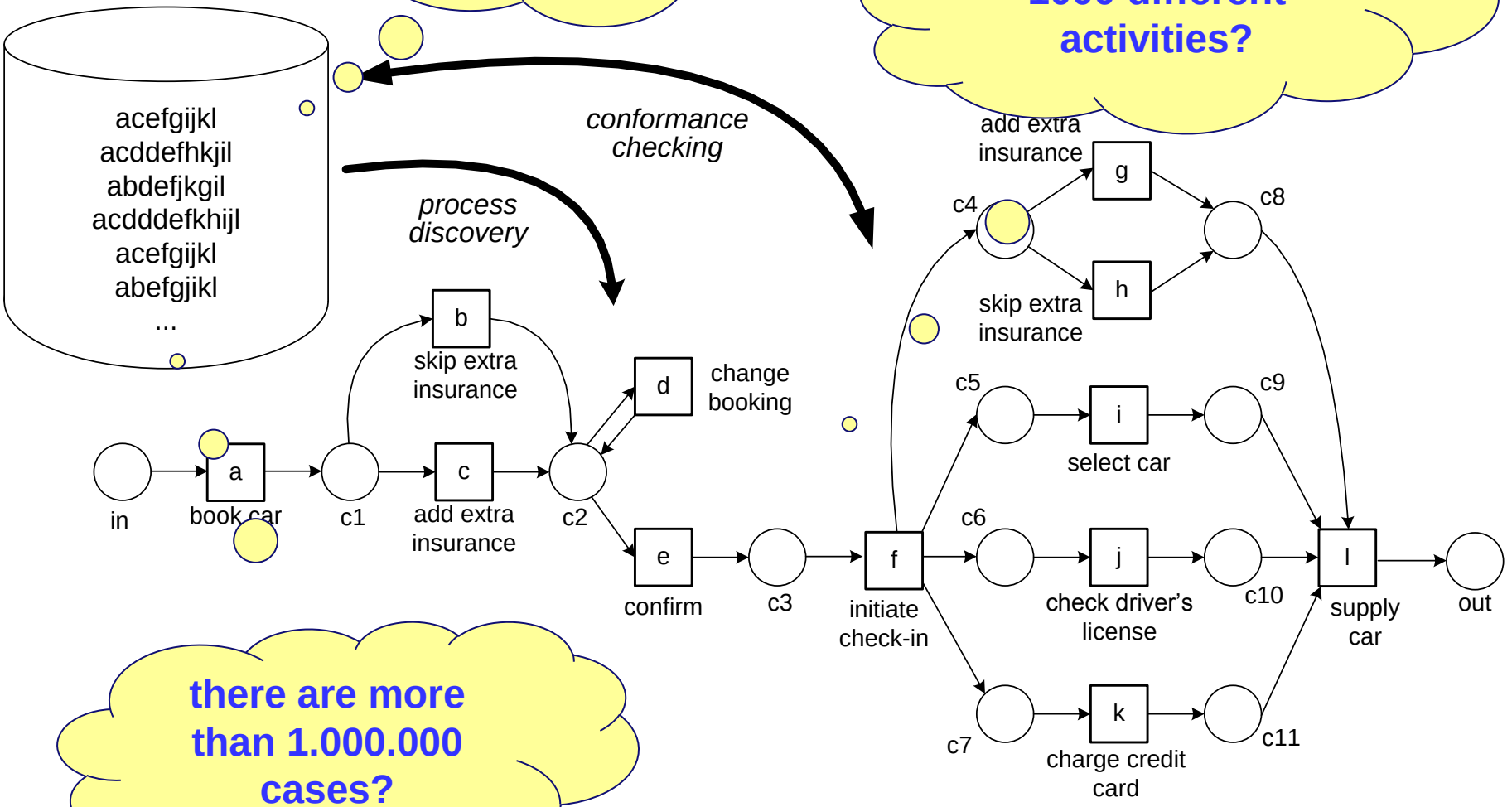
photos/Data-is-the-new-oil-infographic-Nigel-Holmes-2012-from-The-Human-Face-of-Big-Data-original.jpg

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?



Decompose event log!

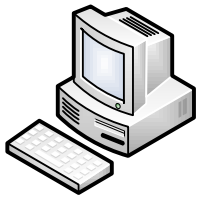
vertical or horizontal

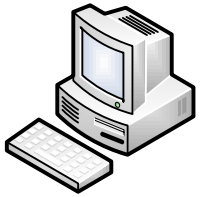
sets of cases

sets of activities

Vertical distribution: Split cases

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

Horizontal distribution

sets of activities

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

{a,b,e,f,g}

abeg
abefbeg
abeg
abefbeg
abefbeg
abefbeg
abeg
abeg
abefbefbeg
abeg
abefbefbeg
abefbeg
abeg
abeg
abefbeg
abeg

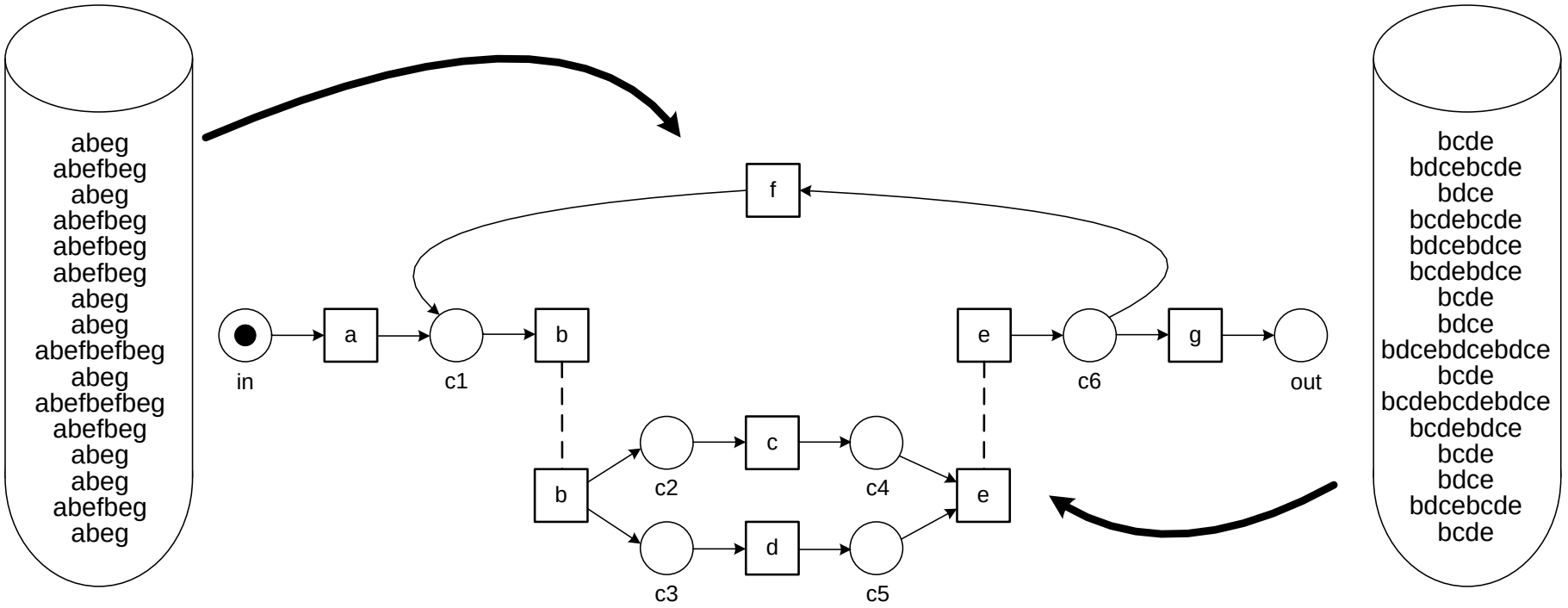
{b,c,d,e}

bcde
bdcebcde
bdce
bcdebcde
bdcebdce
bcdebdce
bcde
bdce
bdcebdcebdce
bcde
bcdebcdebdce
bcdebdce
bcde
bdce
bdcebcde
bcde

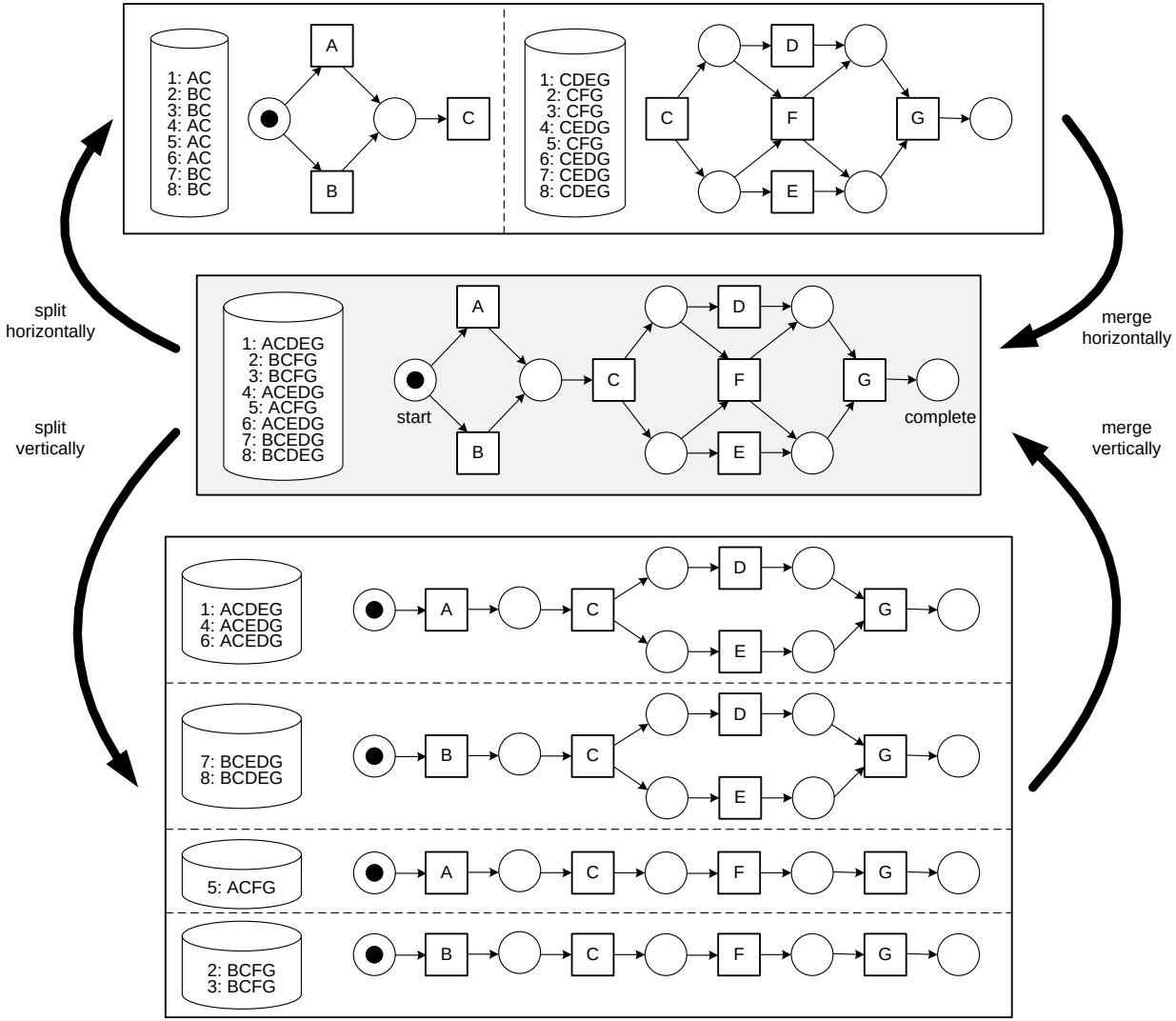
Horizontal distribution: The key idea

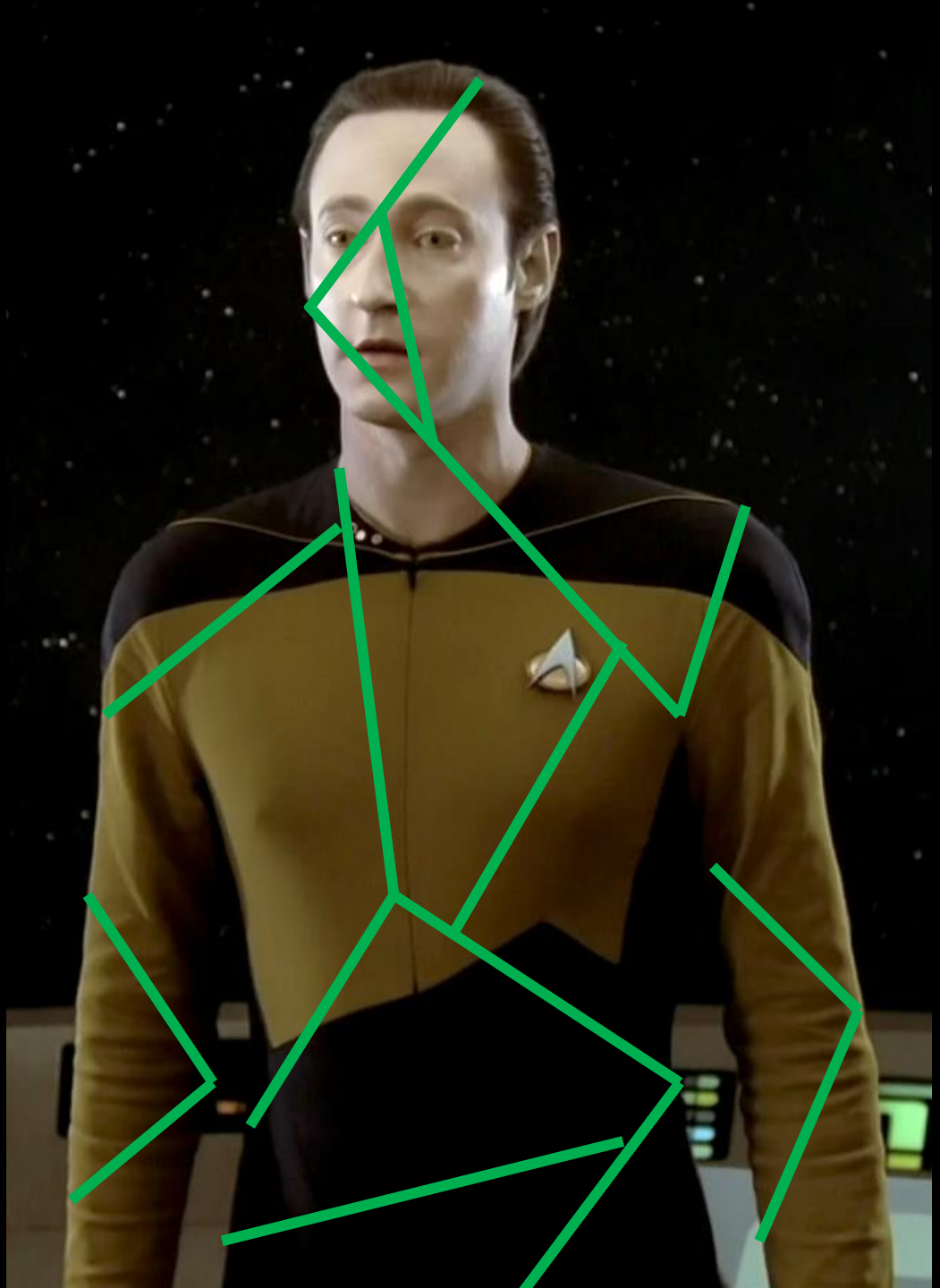
projected on {a,b,e,f,g}

projected on {b,c,d,e}

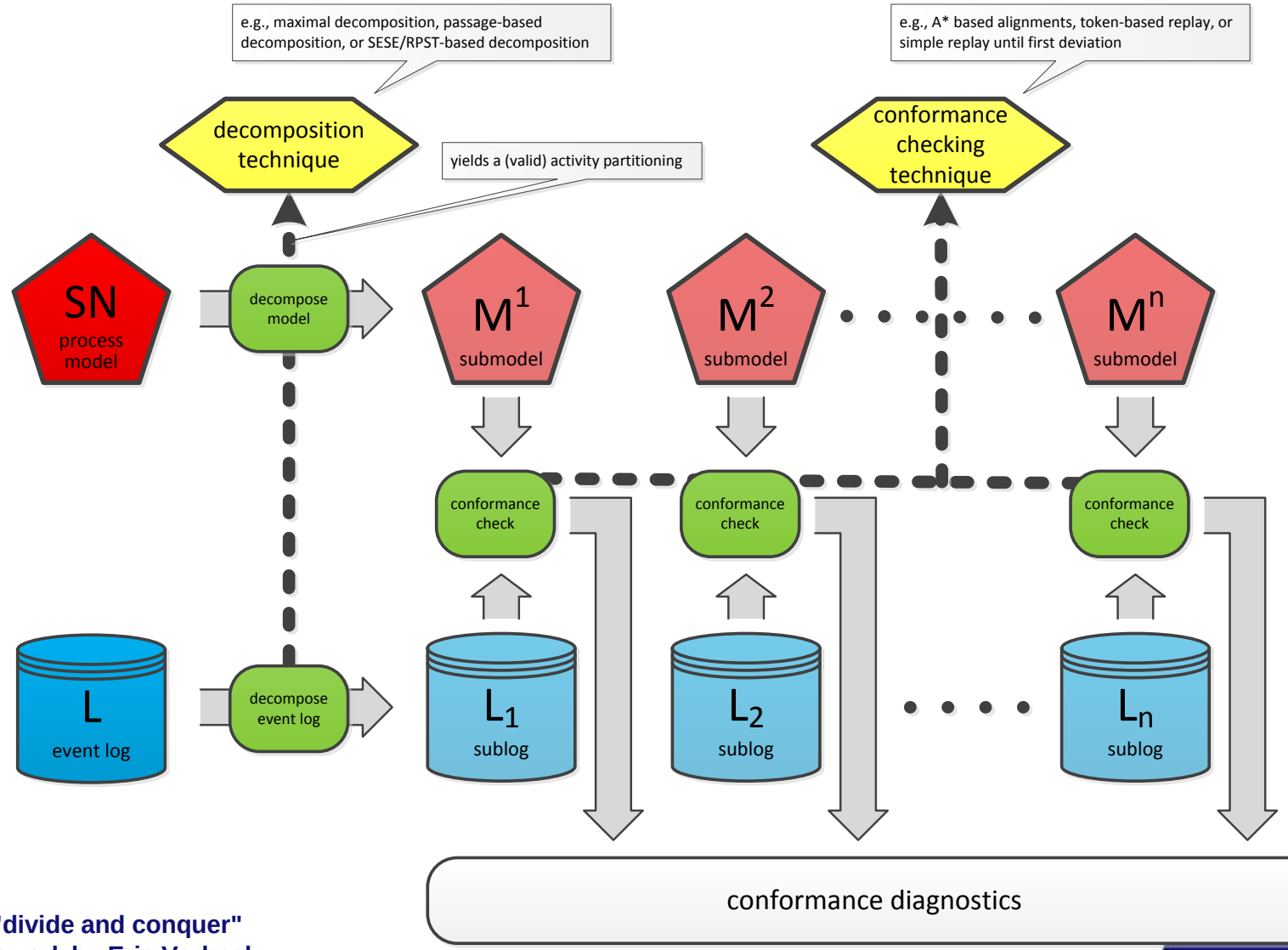


Two foundational ways of spitting event data: horizontal or vertical



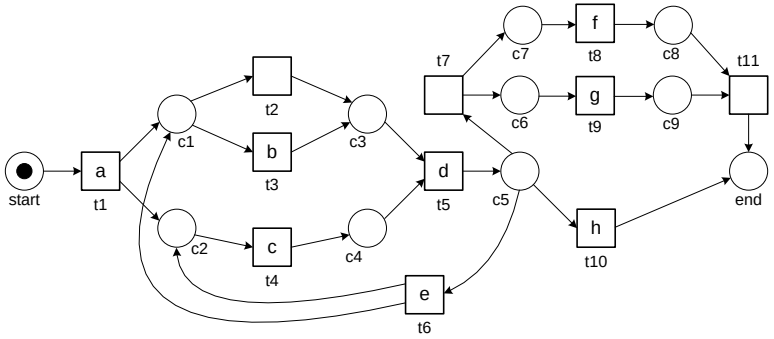


Decomposing Conformance Checking

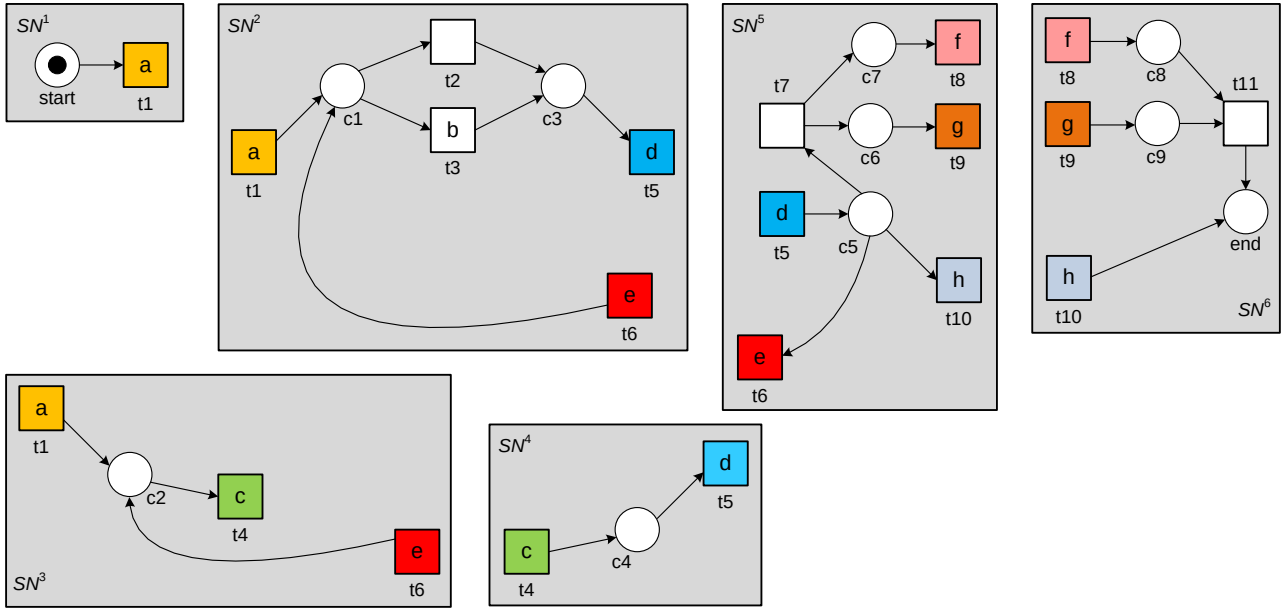


See "divide and conquer" framework by Eric Verbeek.

Example of a valid decomposition

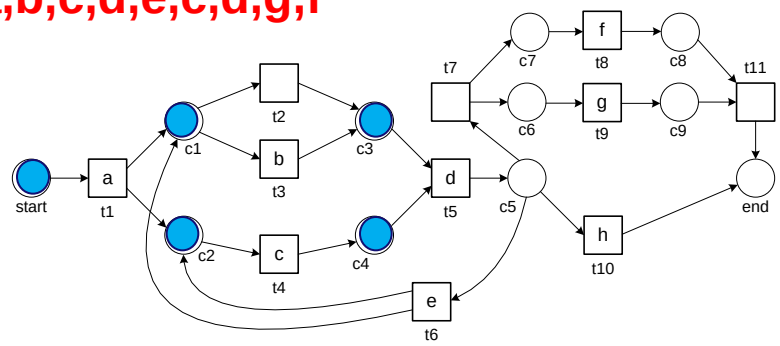


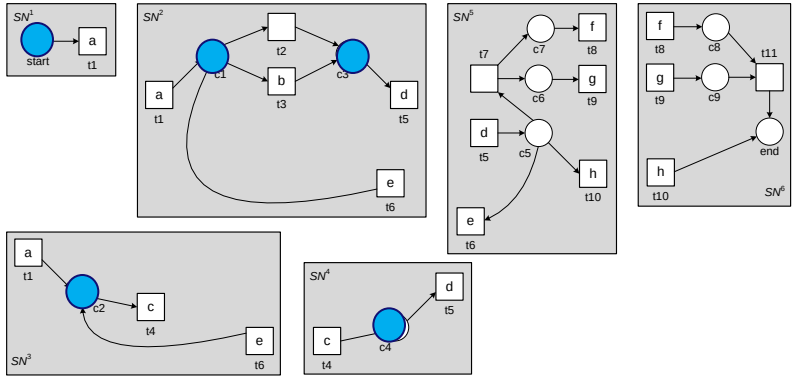
Log can be split in the same way!



Example of an alignment for observed trace a,b,c,d,e,c,d,g,f

a,b,c,d,e,c,d,g,f





Etc.

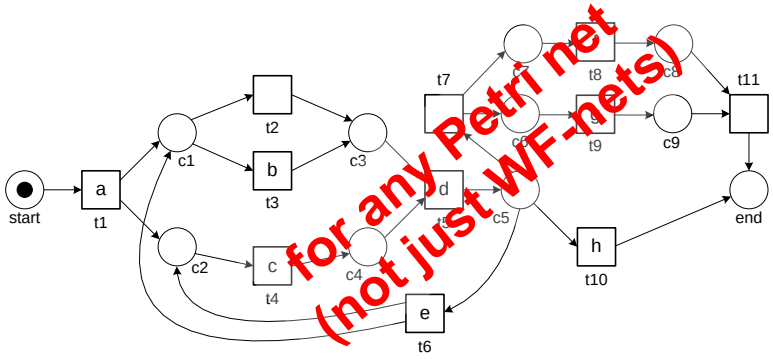
$$\gamma_3 = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|} \hline 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \hline a & b & c & d & e & c & \gg & d & \gg & g & f & \gg \\ \hline a & b & c & d & e & c & \tau & d & \tau & g & f & \tau \\ t1 & t3 & t4 & t5 & t6 & t4 & t2 & t5 & t7 & t9 & t8 & t11 \\ \hline \end{array}$$

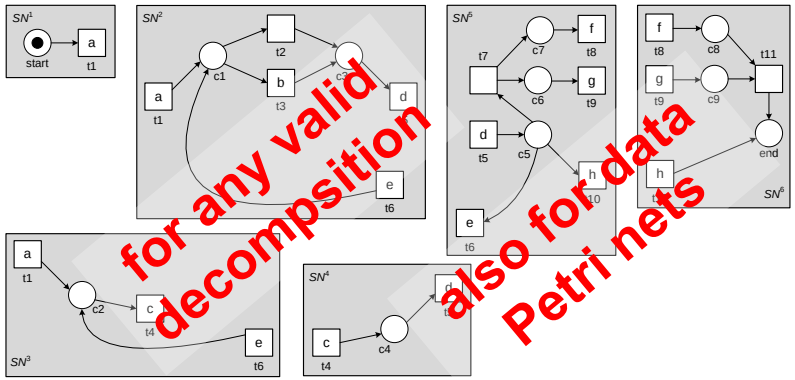
$$\gamma_3^1 = \begin{array}{|c|} \hline 1 \\ \hline a \\ \hline a \\ t1 \\ \hline \end{array} \qquad \gamma_3^2 = \begin{array}{|c|c|c|c|c|c|} \hline 1 & 2 & 4 & 5 & 7 & 8 \\ \hline a & b & d & e & \gg & d \\ \hline a & b & d & e & \tau & d \\ t1 & t3 & t5 & t6 & t2 & t5 \\ \hline \end{array} \qquad \gamma_3^3 = \begin{array}{|c|c|c|c|} \hline 1 & 3 & 5 & 6 \\ \hline a & c & e & c \\ \hline a & c & e & c \\ t1 & t4 & t6 & t4 \\ \hline \end{array}$$

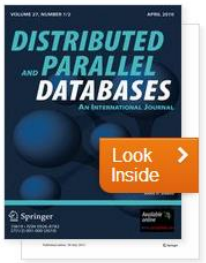
$$\gamma_3^4 = \begin{array}{|c|c|c|c|} \hline 3 & 4 & 6 & 8 \\ \hline c & d & c & d \\ \hline c & d & c & d \\ t4 & t5 & t4 & t5 \\ \hline \end{array} \qquad \gamma_3^5 = \begin{array}{|c|c|c|c|c|c|} \hline 4 & 5 & 8 & 9 & 10 & 11 \\ \hline d & e & d & \gg & g & f \\ \hline d & e & d & \tau & g & f \\ t5 & t6 & t5 & t7 & t9 & t8 \\ \hline \end{array} \qquad \gamma_3^6 = \begin{array}{|c|c|c|} \hline 10 & 11 & 12 \\ \hline g & f & \gg \\ \hline g & f & \tau \\ t9 & t8 & t11 \\ \hline \end{array}$$

Conformance checking can be decomposed !!!

- **General result for any valid decomposition: Any event log or trace is perfectly fitting the overall model if and only if it is also fitting all the individual fragments**



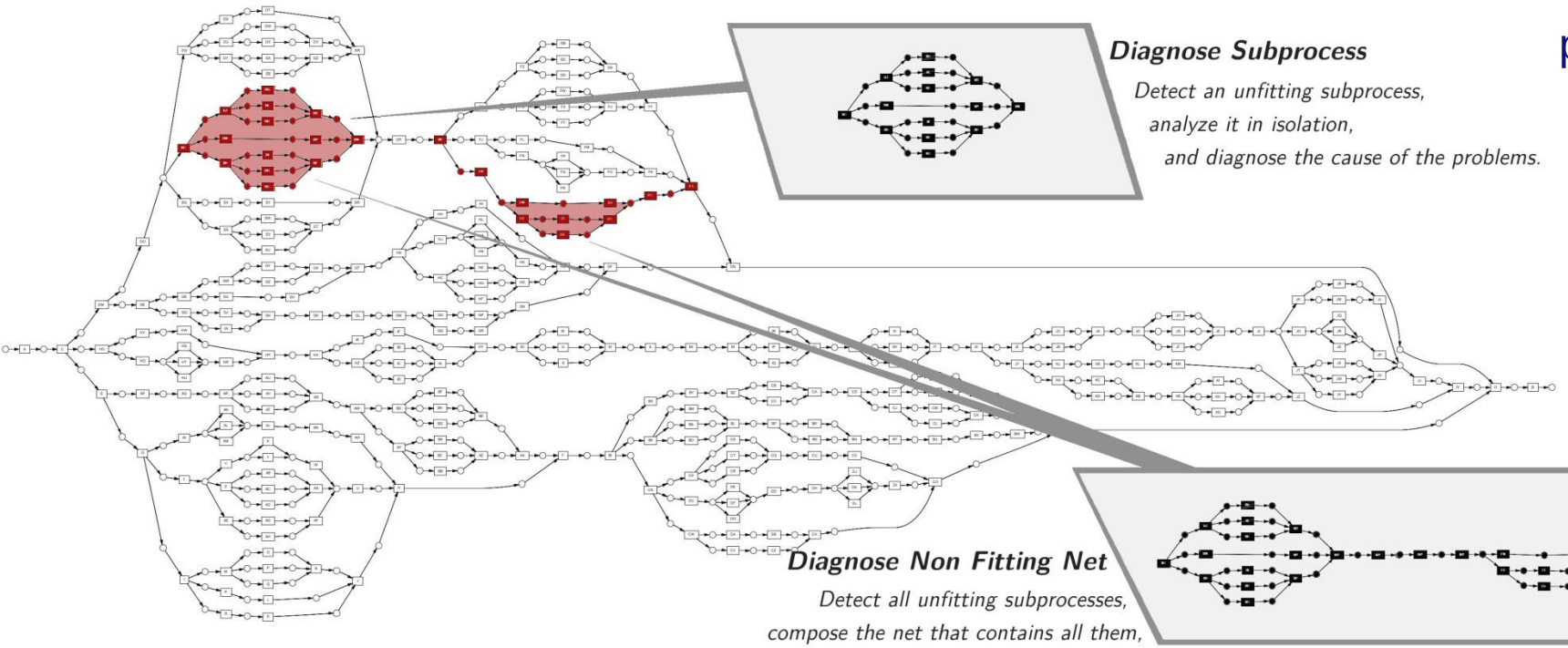


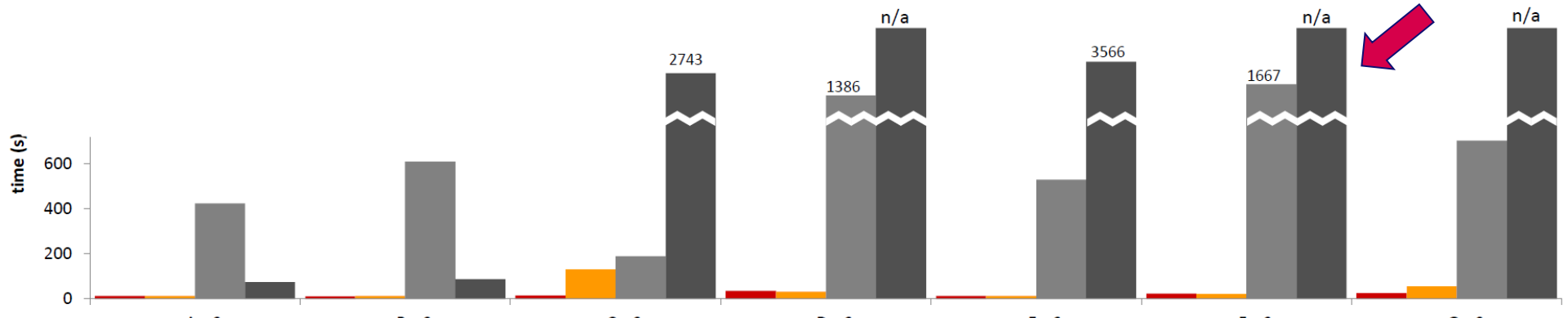


Wil van der Aalst, Decomposing Petri nets for process mining: A generic approach. Distributed and Parallel Databases, Volume 31, Issue 4, pp 471-507, 2013

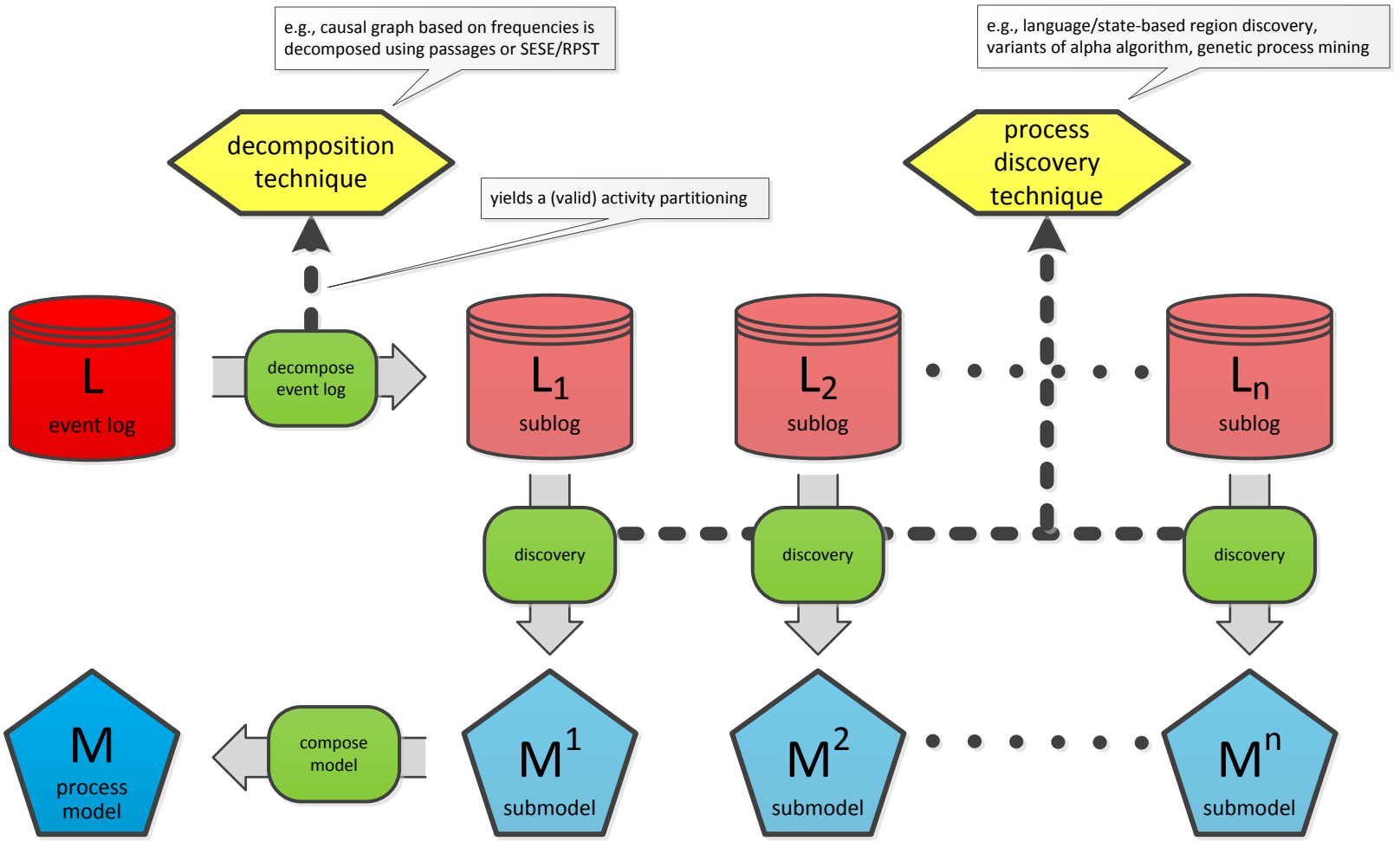
Example

(work with Jorge Munoz-Gama and Josep Carmona)





Decomposing Process Discovery



See "divide and conquer" framework by Eric Verbeek.

conclusion


Process mining: mediating between modeled and observed behavior
Decomposition as a way to deal with "Big" process mining tasks
Many challenges, e.g., process discovery
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

ProM
process mining workbench

Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

www.processmining.org

www.win.tue.nl/ieeetfpm/

Springer

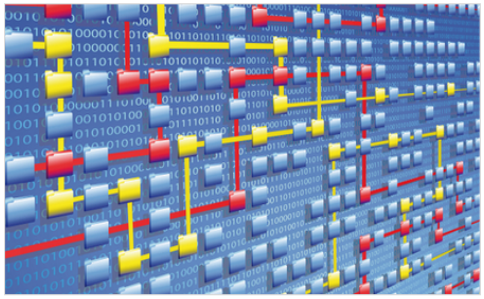
Process Mining: Data Science in Action

https://www.coursera.org/course/procmin

TU/e Technische Universiteit Eindhoven University of Technology

Process Mining: Data science in Action

Process mining is the missing link between model-based process analysis and data-oriented analysis techniques. Through concrete data sets and easy to use software the course provides data science knowledge that can be applied directly to analyze and improve processes in a variety of domains.



First Massive Open Online Course (MOOC) on Process Mining

About the Course

Data science is the profession of the future, because organizations that are unable to use (big) data in a smart way will not survive. It is not sufficient to focus on data storage and data analysis. The data scientist also needs to relate data to process analysis. **Process mining bridges the gap between traditional model-based process analysis (e.g., simulation and other business process management techniques) and data-centric analysis techniques such as machine learning and data mining.** Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). This technology has become available only recently, but it can be applied to any type of operational processes (organizations and systems). Example applications include: analyzing treatment processes in hospitals, improving customer service processes in a multinational, understanding the browsing behavior of customers using a booking site, analyzing failures of a baggage handling system, and improving the user interface of an X-ray machine. All of these applications have in common that dynamic behavior needs to be related to process models. Hence, we refer to this as "data science in action".

The course explains the key analysis techniques in process mining. Participants will learn various process discovery algorithms. These can be used to automatically learn process models from raw event data. Various other process analysis techniques that use event data will be presented. Moreover, the course will provide **easy-to-use software**, **real-life data sets**, and **practical skills** to **directly apply the theory** in

Sessions

Nov 12th 2014 - Dec 24th 2014

Starts in 3 months

Eligible for

Statement of Accomplishment

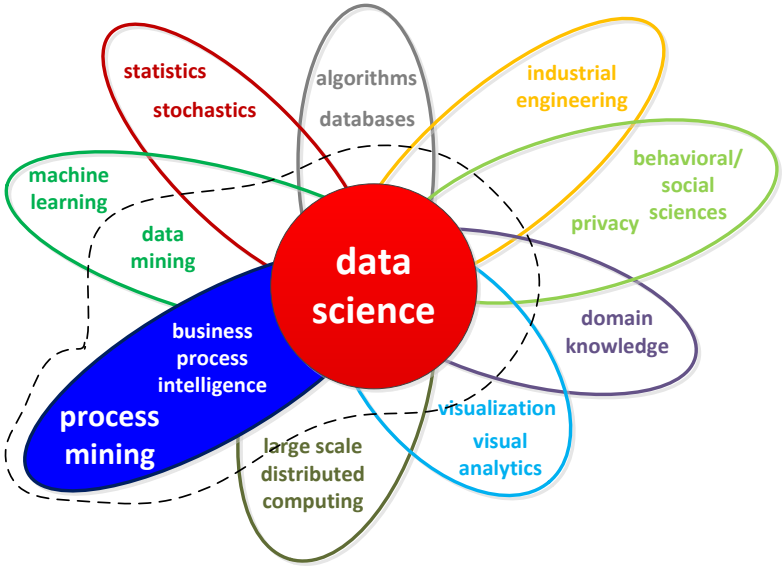
Course at a Glance

- 6 weeks of study
- 4-6 hours of work / week
- English
- English subtitles

Instructors

Wil van der Aalst
Eindhoven University of Technology

Categories



coursera

TU/e Technische Universiteit Eindhoven University of Technology

Data Science Center Eindhoven