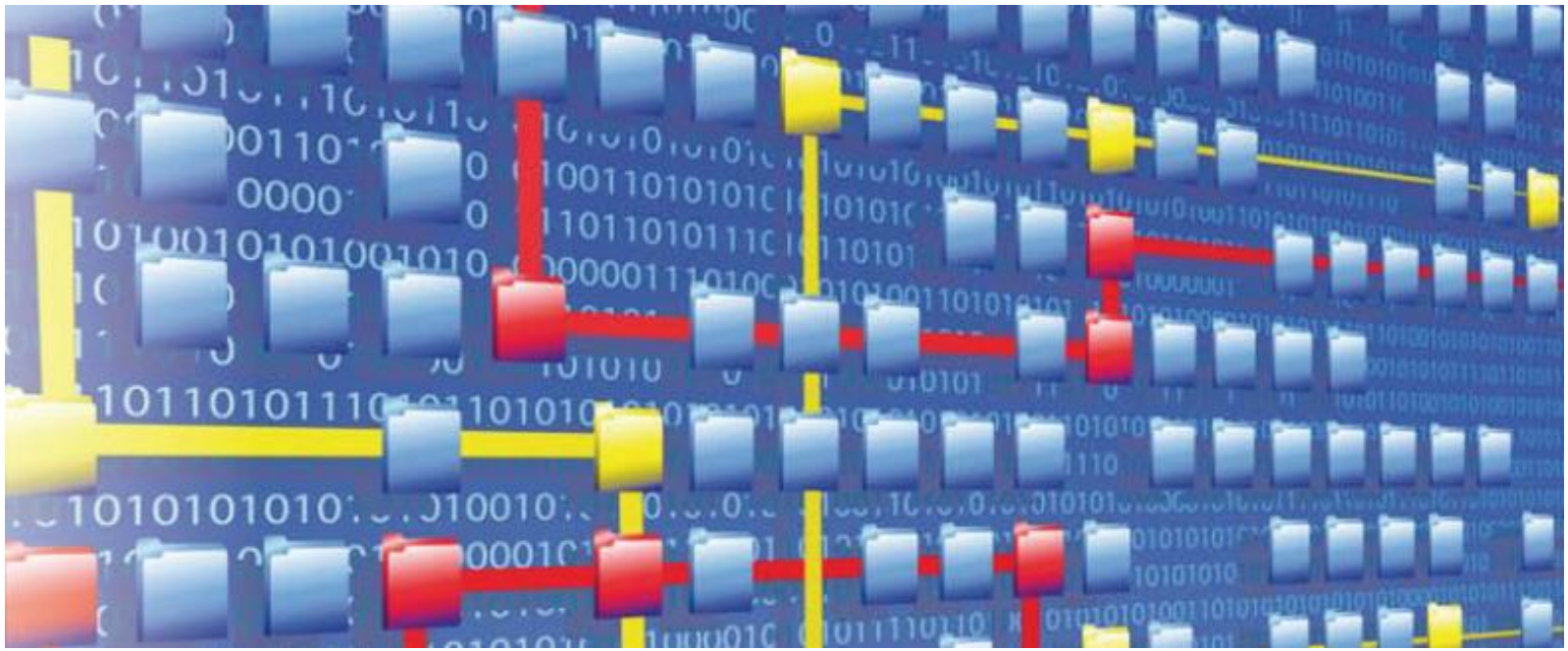


Process Mining

I.A.Lomazova

National Research University Higher School of Economics,
Laboratory for Process-Aware Information Systems

Slides use materials by W. van der Aalst and Process Mining Manifesto

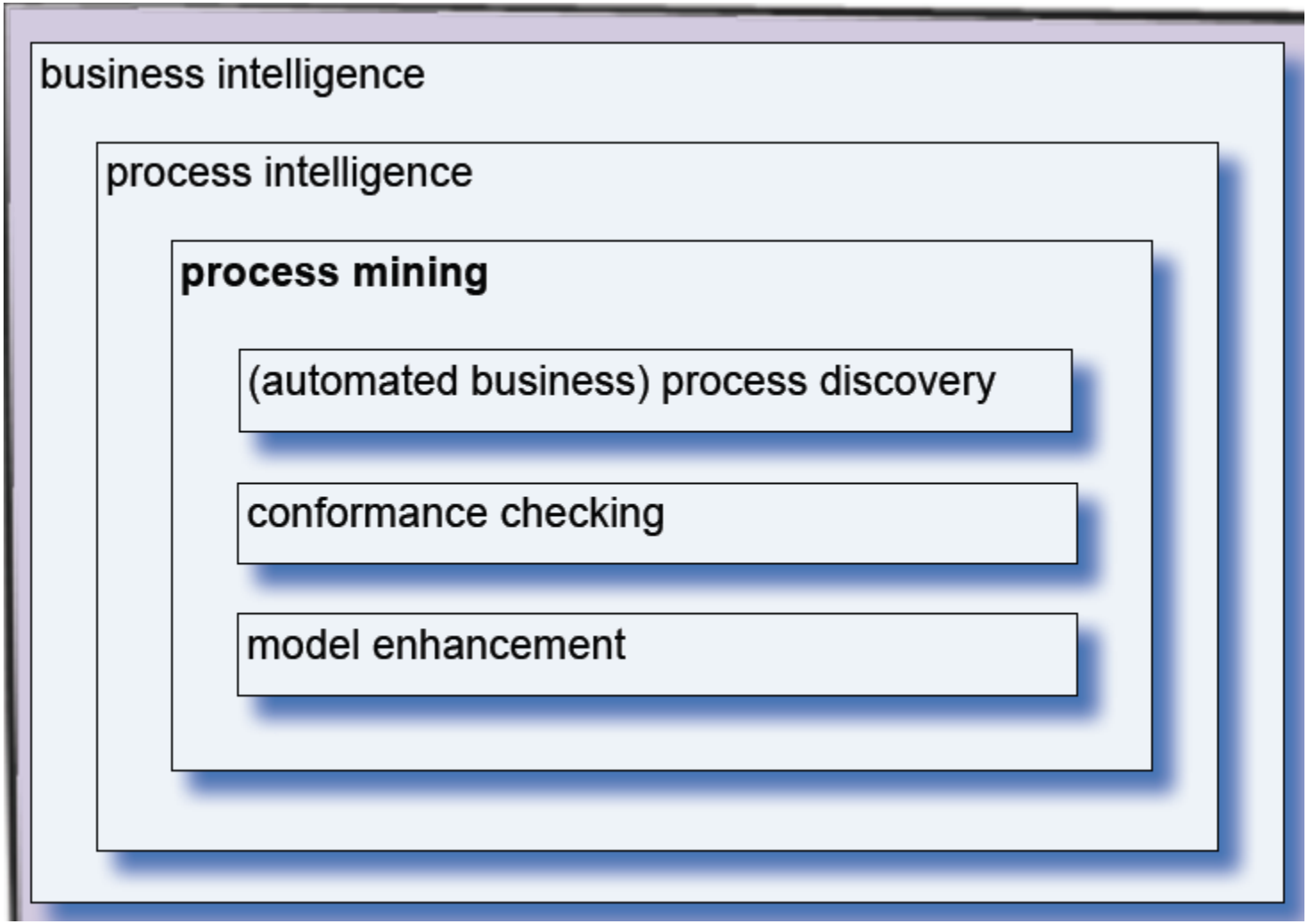
It is not just about technology …


Process Mining =
Event Data + Processes
Data Mining + Process Analysis
Machine Learning + Formal Methods

- **Process Mining:** techniques, tools, and methods to
  - ***discover***,
  - ***monitor***, and
  - ***improve***

  real processes (i.e., not assumed processes) by extracting knowledge from event logs commonly available in today's (information) systems.

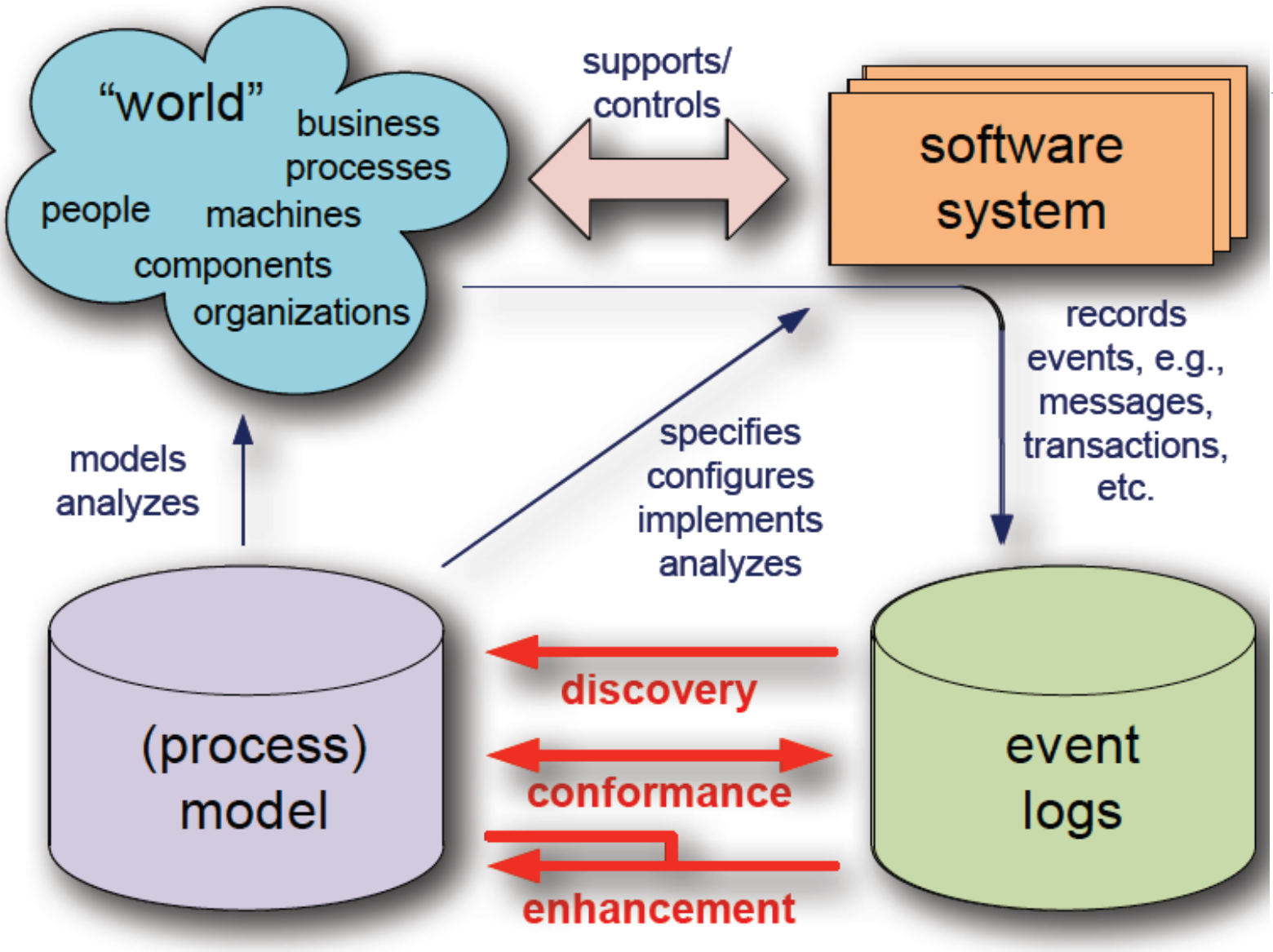
Relating the different terms

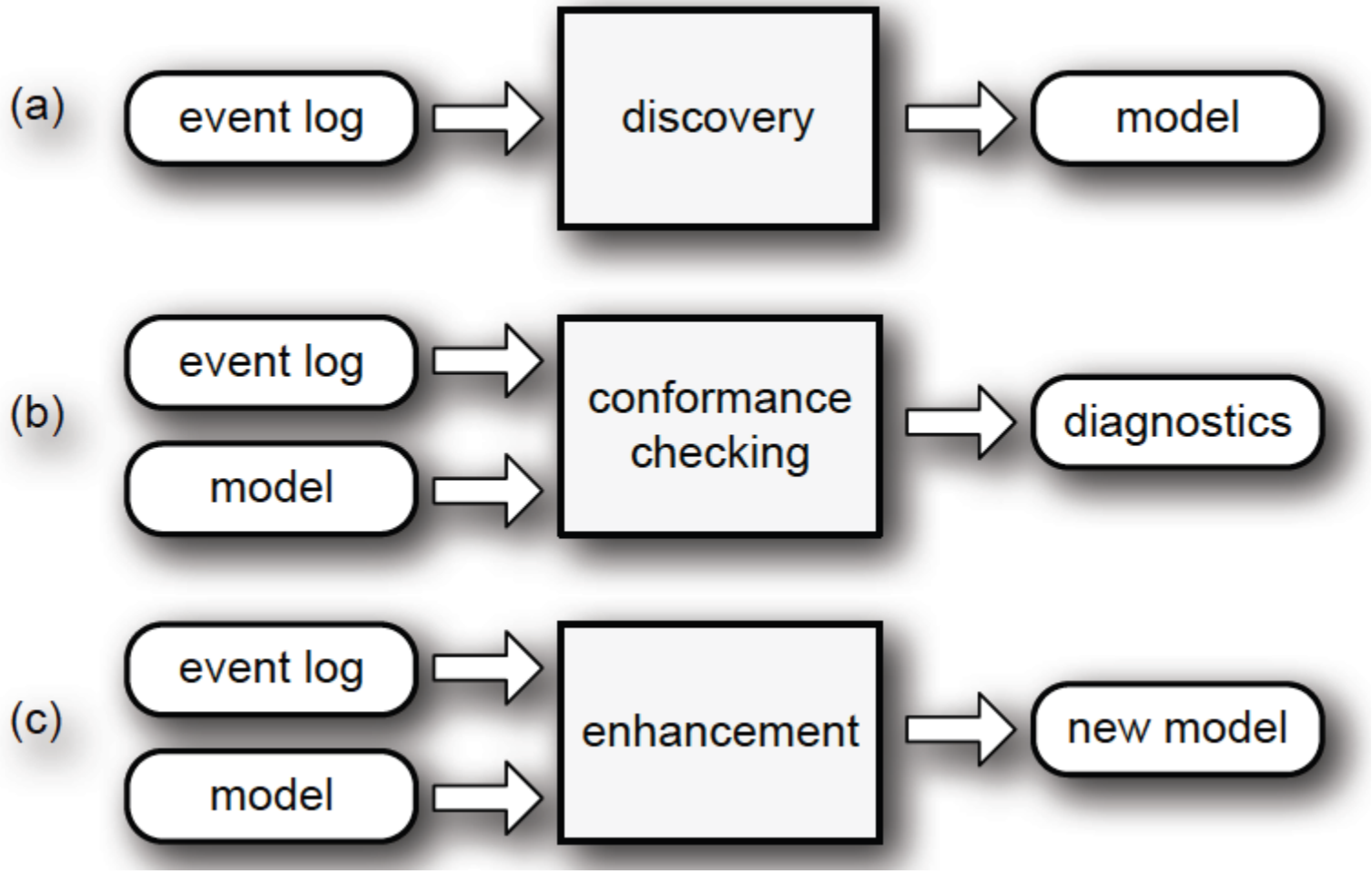



Also applies to cloud computing!
Web-services!
Processes!!
Dealing with variability
Process variants/ configuration

New opportunities!

Cross-organizational process mining!!


"world"
business processes
people
machines
components
organizations
supports/ controls
software system
records events, e.g., messages, transactions, etc.
specifies configures implements analyzes
models analyzes
(process) model
discovery
conformance
enhancement
event logs


(a)
event log
discovery
model
(b)
event log
model
conformance checking
diagnostics
(c)
event log
model
enhancement
new model

Starting point: event log

case id	event id	properties				
		timestamp	activity	resource	cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	register request			
	35654522	30-12-2010:15.06	examine casually			
	35654524	30-12-2010:16.34	check ticket			
	35654525	06-01-2011:09.18	decide			
	35654526	06-01-2011:12.18	reinitiate request			
	35654527	06-01-2011:13.06	examine thoroughly			
	35654530	08-01-2011:11.43	check ticket			
	35654531	09-01-2011:09.55	decide			
	35654533	15-01-2011:10.45	pay compensation			
4	35654641	06-01-2011:15.02	register request			
	35654643	07-01-2011:12.06	check ticket			
	35654644	08-01-2011:14.43	examine thoroughly			
	35654645	09-01-2011:12.02	decide			
	35654647	12-01-2011:15.44	reject request			
5	35654711	06-01-2011:09.02	register request			
	35654712	07-01-2011:10.16	examine casually			
	35654714	08-01-2011:11.22	check ticket			
	35654715	10-01-2011:13.28	decide			
	35654716	11-01-2011:16.18	reinitiate request			
	35654718	14-01-2011:14.33	check ticket			
	35654719	16-01-2011:15.50	examine casually			
	35654720	19-01-2011:11.18	decide	Sara	200	...
	35654721	20-01-2011:12.48	reinitiate request	Sara	200	...
	35654722	21-01-2011:09.06	examine casually	Sue	400	...
	35654724	21-01-2011:11.34	check ticket	Pete	100	...
	35654725	23-01-2011:13.12	decide	Sara	200	...
	35654726	24-01-2011:14.56	reject request	Mike	200	...
6	35654871	06-01-2011:15.02	register request	Mike	50	...
	35654873	06-01-2011:16.06	examine casually	Ellen	400	...
	35654874	07-01-2011:16.22	check ticket	Mike	100	...
	35654875	07-01-2011:16.52	decide	Sara	200	...
	35654877	16-01-2011:11.47	pay compensation	Mike	200	...
...	...	...	...	...	...	...

case id	event id	properties				
		timestamp	activity	resource	cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...

XES, MXML, SA-MXML, CSV, etc.

Simplified event log

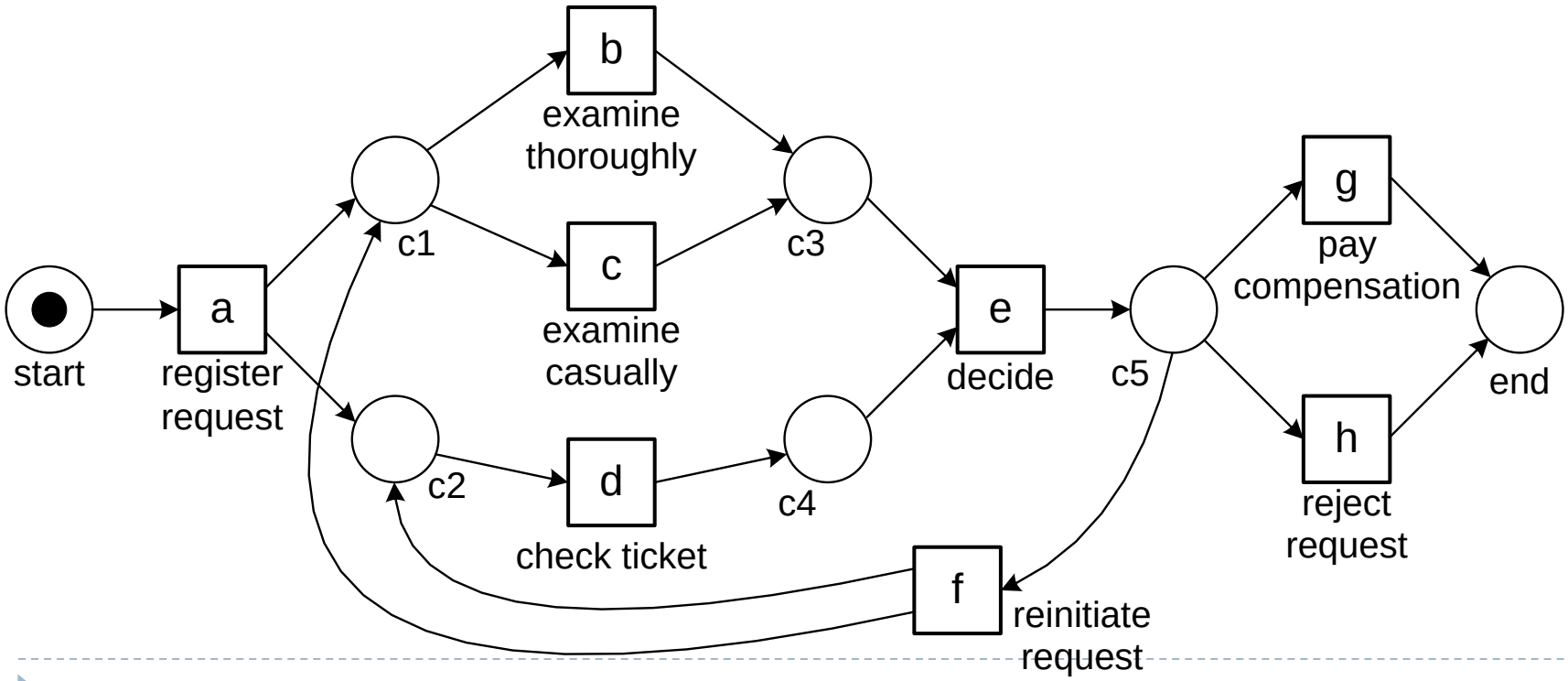
case id	event id	properties				
		timestamp	activity	resource		
1	35654423	30-12-2010:11.02	register request	Pete		
	35654424	31-12-2010:10.06	examine thoroughly	Sue		
	35654425	05-01-2011:15.12	check ticket	Mike		
	35654426	06-01-2011:11.18	decide	Sara		
	35654427	07-01-2011:14.24	reject request	Pete		
2	35654483	30-12-2010:11.32	register request	Mike		
	35654485	30-12-2010:12.12	check ticket	Mike		
	35654487	30-12-2010:14.16	examine casually	Pete		
	35654488	05-01-2011:11.22	decide	Sara		
	35654489	08-01-2011:12.05	pay compensation	Ellen		
3	35654521	30-12-2010:14.32	register request	Pete		
	35654522	30-12-2010:15.06	examine casually	Mike		
	35654524	30-12-2010:16.34	check ticket	Ellen		
	35654525	06-01-2011:09.18	decide	Sara		
	35654526	06-01-2011:12.18	reinitiate request	Sara		
	35654527	06-01-2011:13.06	examine thoroughly	Sean		
	35654530	08-01-2011:11.43	check ticket	Pete		
	35654531	09-01-2011:09.55	decide	Sara		
	35654533	15-01-2011:10.45	pay compensation	Ellen		
4	35654641	06-01-2011:15.02	register request	Pete	50	…
	35654643	07-01-2011:12.06	check ticket	Mike	100	…
	35654644	08-01-2011:14.43	examine thoroughly	Sean	400	…
	35654645	09-01-2011:12.02	decide	Sara	200	…
	35654647	12-01-2011:15.44	reject request	Ellen	200	…
5	35654711	06-01-2011:09.02	register request	Ellen	50	…
	35654712	07-01-2011:10.16	examine casually	Mike	400	…
	35654714	08-01-2011:11.22	check ticket	Pete	100	…
	35654715	10-01-2011:13.28	decide	Sara	200	…
	35654716	11-01-2011:16.18	reinitiate request	Sara	200	…
	35654718	14-01-2011:14.33	check ticket	Ellen	100	…
	35654719	16-01-2011:15.50	examine casually	Mike	400	…
	35654720	19-01-2011:11.18	decide	Sara	200	…
	35654721	20-01-2011:12.48	reinitiate request	Sara	200	…
	35654722	21-01-2011:09.06	examine casually	Sue	400	…
	35654724	21-01-2011:11.34	check ticket	Pete	100	…
	35654725	23-01-2011:13.12	decide	Sara	200	…
	35654726	24-01-2011:14.56	reject request	Mike	200	…
6	35654871	06-01-2011:15.02	register request	Mike	50	…
	35654873	06-01-2011:16.06	examine casually	Ellen	400	…
	35654874	07-01-2011:16.22	check ticket	Mike	100	…
	35654875	07-01-2011:16.52	decide	Sara	200	…
	35654877	16-01-2011:11.47	pay compensation	Mike	200	…
…	…	…	…	…	…	…

case id	trace
1	$\langle a, b, d, e, h \rangle$
2	$\langle a, d, c, e, g \rangle$
3	$\langle a, c, d, e, f, b, d, e, g \rangle$
4	$\langle a, d, b, e, h \rangle$
5	$\langle a, c, d, e, f, d, c, e, f, c, d, e, h \rangle$
6	$\langle a, c, d, e, g \rangle$
…	…

a = register request,
b = examine thoroughly,
c = examine casually,
d = check ticket,
e = decide,
f = reinitiate request,
g = pay compensation,
and h = reject request

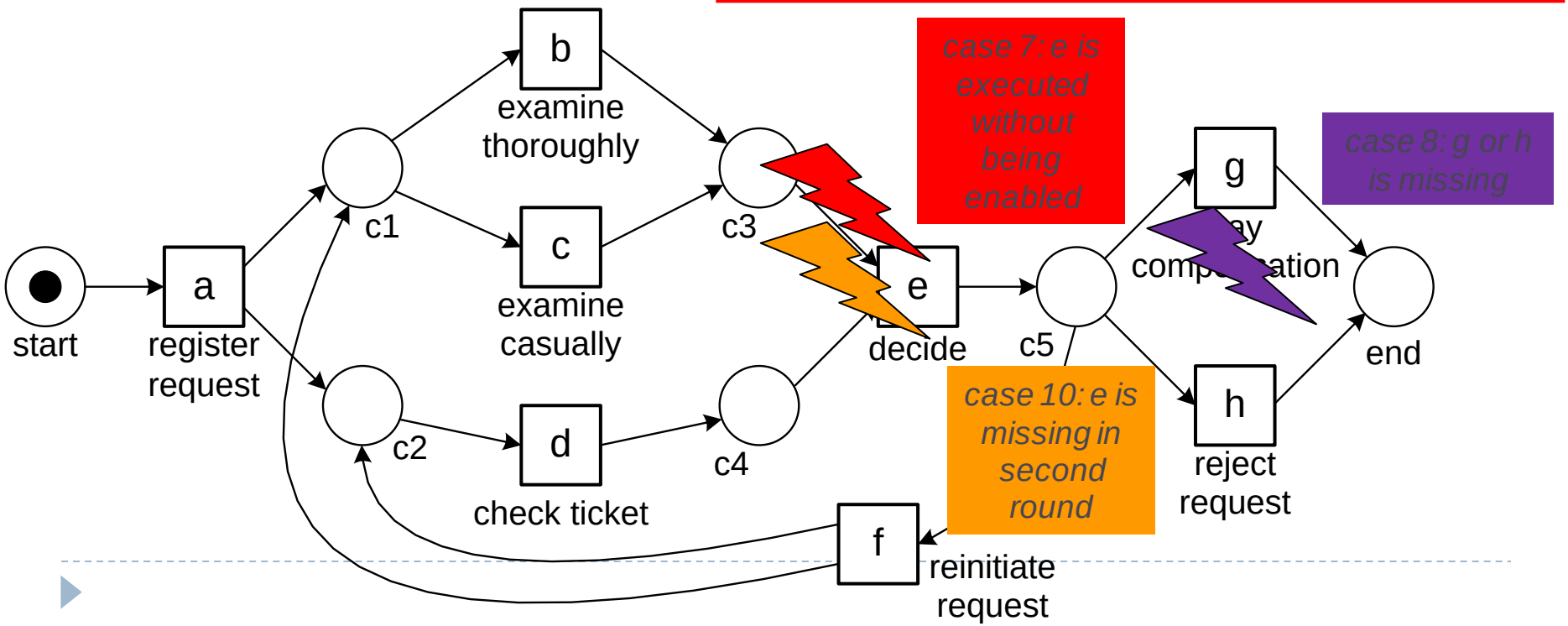
Process discovery

case id	trace
1	$\langle a, b, d, e, h \rangle$
2	$\langle a, d, c, e, g \rangle$
3	$\langle a, c, d, e, f, b, d, e, g \rangle$
4	$\langle a, d, b, e, h \rangle$
5	$\langle a, c, d, e, f, d, c, e, f, c, d, e, h \rangle$
6	$\langle a, c, d, e, g \rangle$
...	...

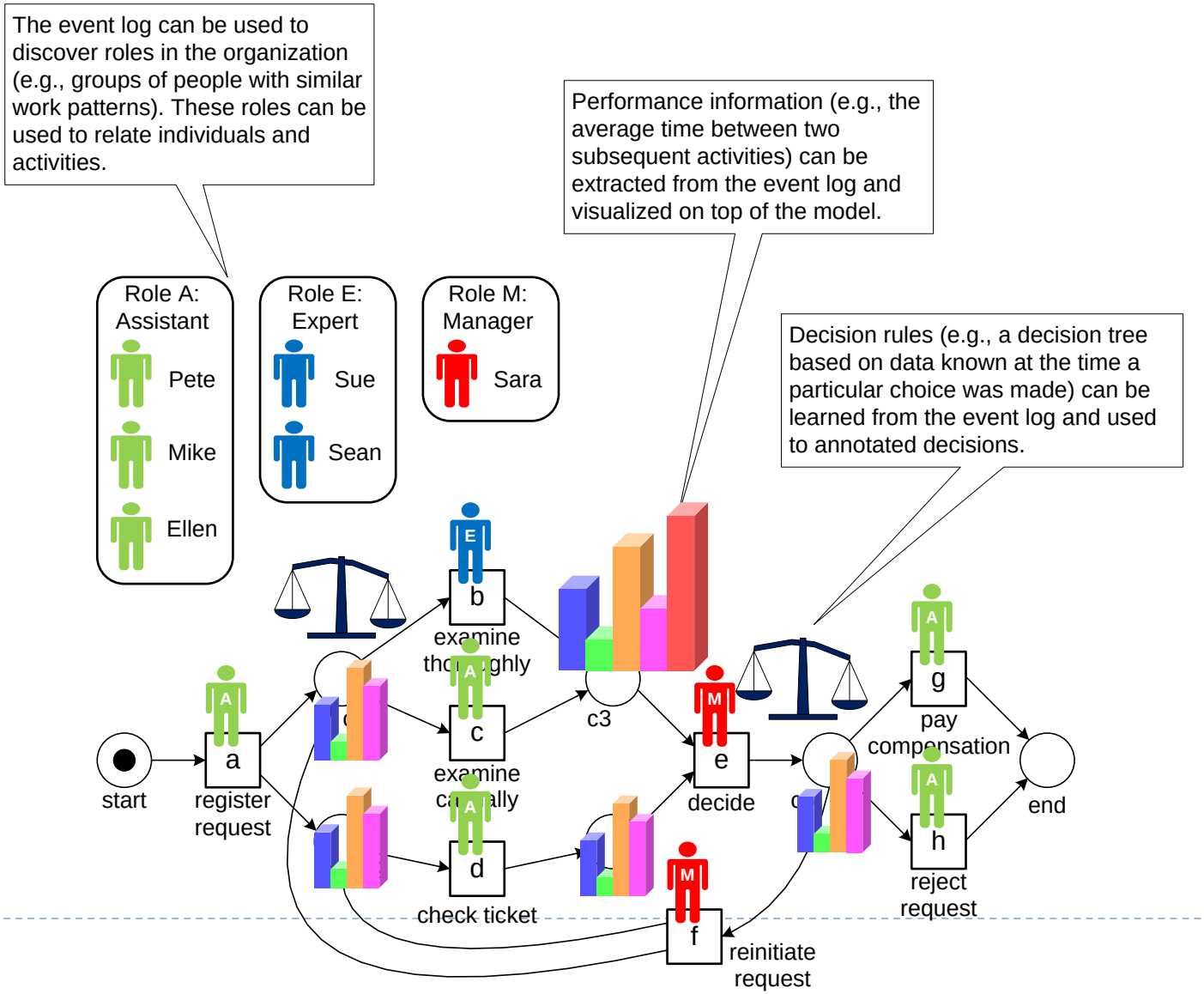


Conformance checking

case id	trace
1	$\langle a,b,d,e,h \rangle$
2	$\langle a,d,c,e,g \rangle$
3	$\langle a,c,d,e,f,b,d,e,g \rangle$
4	$\langle a,d,b,e,h \rangle$
5	$\langle a,c,d,e,f,d,c,e,f,c,d,e,h \rangle$
6	$\langle a,c,d,e,g \rangle$
7	$\langle \mathbf{a},\mathbf{b},\mathbf{e},\mathbf{g} \rangle$
8	$\langle \mathbf{a},\mathbf{b},\mathbf{d},\mathbf{e} \rangle$
9	$\langle a,d,c,e,f,d,c,e,f,b,d,e,h \rangle$
10	$\langle \mathbf{a},\mathbf{c},\mathbf{d},\mathbf{e},\mathbf{f},\mathbf{b},\mathbf{d},\mathbf{g} \rangle$



Extension: Adding perspectives to model based on event log



All supported by …

- 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

IEEE Task Force on Process Mining

www.win.tue.nl/ieeetfpm/

IEEE Task Force on Process Mining is established in the context of the Data Mining Technical Committee (DMTC) of the Computational Intelligence Society (CIS) of the Institute of Electrical and Electronic Engineers, Inc. (IEEE).

The goal of this – to promote the research, development, education and understanding of process mining.

Process Mining Manifesto

Originally published in “Business Process Management Workshops 2011, Lecture Notes in Business Information Processing, Vol. 99, Springer-Verlag, 2011, and has been translated into various languages.

Guiding Principles

- GP1: Event Data Should Be Treated as First-Class Citizens

Starting point: collections of events --- event logs (*database tables, message logs, mail archives, transaction logs, and other data sources*)

GP1: Event Data Should Be Treated as First-Class Citizens

Quality of event data:

trustworthy, i.e., it should be safe to assume that the recorded events actually happened and that the attributes of events are correct;

well-defined **semantics;**

safe (privacy and security) ;

Level	Characterization	Examples
★★★★★	Highest level: the event log is of excellent quality (i.e., trustworthy and complete) and events are well-defined. Events are recorded in an automatic, systematic, reliable, and safe manner. Privacy and security considerations are addressed adequately. Moreover, the events recorded (and all of their attributes) have clear semantics. This implies the existence of one or more ontologies. Events and their attributes point to this ontology.	Semantically annotated logs of BPM systems.
★★★★	Events are recorded automatically and in a systematic and reliable manner, i.e., logs are trustworthy and complete. Unlike the systems operating at level ★★★, notions such as process instance (case) and activity are supported in an explicit manner.	Events logs of traditional BPM/ workflow systems.
★★★	Events are recorded automatically, but no systematic approach is followed to record events. However, unlike logs at level ★★, there is some level of guarantee that the events recorded match reality (i.e., the event log is trustworthy but not necessarily complete). Consider, for example, the events recorded by an ERP system. Although events need to be extracted from a variety of tables, the information can be assumed to be correct (e.g., it is safe to assume that a payment recorded by the ERP actually exists and vice versa).	Tables in ERP systems, event logs of CRM systems, transaction logs of messaging systems, event logs of high-tech systems, etc.
★★	Events are recorded automatically, i.e., as a by-product of some information system. Coverage varies, i.e., no systematic approach is followed to decide which events are recorded. Moreover, it is possible to bypass the information system. Hence, events may be missing or not recorded properly.	Event logs of document and product management systems, error logs of embedded systems, worksheets of service engineers, etc.
★	Lowest level: event logs are of poor quality. Recorded events may not correspond to reality and events may be missing. Event logs for which events are recorded by hand typically have such characteristics.	Trails left in paper documents routed through the organization ("yellow notes"), paper-based medical records, etc.

Table 1: Maturity levels for event logs.

GP2: Log Extraction Should Be Driven by Questions

Without concrete questions it is very difficult to extract meaningful event data.

- *Given a database with event data related to orders, order lines, and deliveries, there are different process models that can be discovered.*
- *One can extract data with the goal to describe the life-cycle of individual orders. However, it is also possible to extract data with the goal to discover the life-cycle of individual order lines or the life-cycle of individual deliveries.*

GP3: Concurrency, Choice and Other Basic Control-Flow Constructs Should be Supported

Basic workflow constructs (patterns) supported by all mainstream languages are

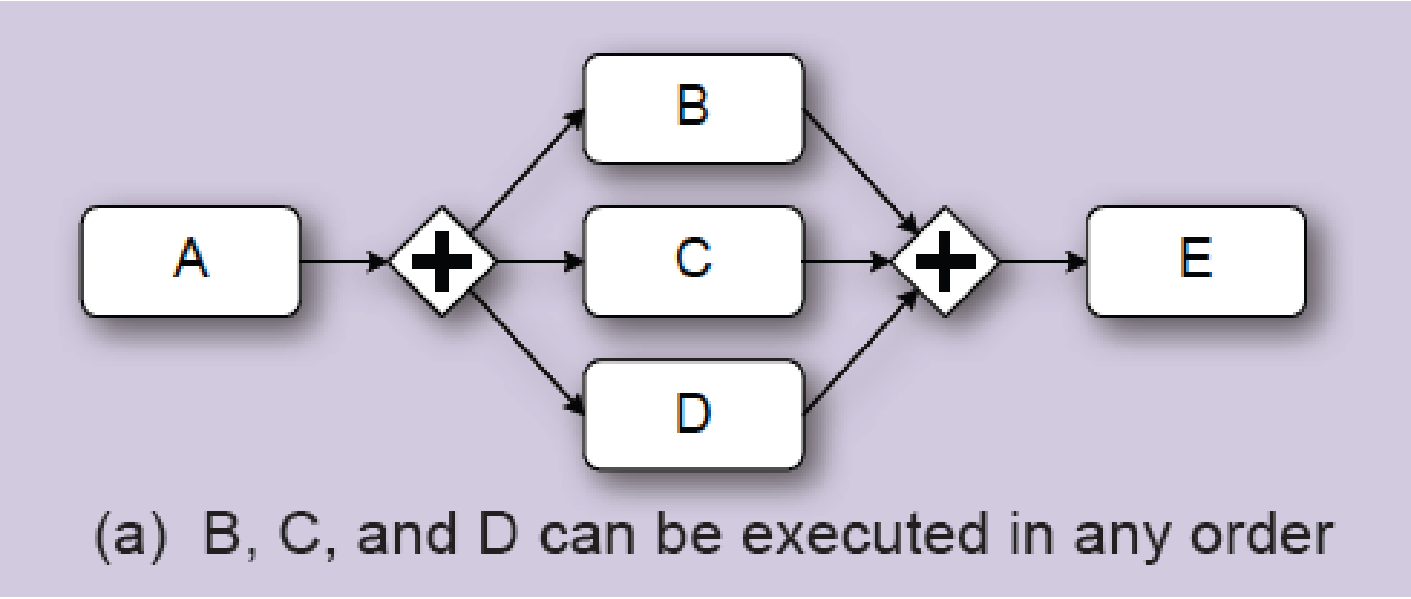
- *sequence,*
- *parallel routing (AND-splits/joins),*
- *choice (XOR-splits/joins), and*
- *loops.*

Consider an event log

$L=\{\langle A, B, C, D, E\rangle, \langle A, B, D, C, E\rangle, \langle A, C, B, D, E\rangle, \langle A, C, D, B, E\rangle, \langle A, D, B, C, E\rangle, \langle A, D, C, B, E\rangle\}$.

L contains cases that start with A and end with E.

Activities B, C, and D occur in any order in-between A and E.



(a) B, C, and D can be executed in any order

GP4: Events Should Be Related to Model Elements

Discovered process models may cover various perspectives:

- organizational perspective,
- time perspective,
- data perspective,
- etc.

Conformance checking and enhancement heavily rely on the relationship between elements in the model and events in the log.

GP5: Models Should Be Treated as Purposeful Abstractions of Reality

The model should emphasize the things relevant for a particular type of user.

- *a manager may want to see a coarse informal process model focusing on costs whereas a process analyst may want to see a detailed process model focusing on deviations from the normal flow.*

Different stakeholders view a process at different levels:

- *strategic level (long-term effects, event data over a longer period),*
- *tactical level (medium-term effects and recent data),*
- *operational level (immediate effects, event data related to running cases).*

GP6: Process Mining Should Be a Continuous Process

- Processes change while they are being analyzed.
- Process mining tools can help end users

(a) by navigating through processes,

(b) by projecting dynamic information onto process maps (e.g., showing "traffic jams" in business processes),

(c) by providing predictions regarding running cases (e.g., estimating the "arrival time“ of a case that is delayed).

Challenges

C1: Finding, Merging, and Cleaning Event Data

- Data may be distributed over a variety of sources.
- Event data are often "object centric“ rather than "process centric".
- Event data may be incomplete.
- An event log may contain outliers, i.e., exceptional behavior (noise).
- Logs may contain events at different levels of granularity.
- Events occur in a particular context.

C2: Dealing with Complex Event Logs Having Diverse Characteristics

- Efforts are needed to improve performance and scalability.
- Dealing with incompleteness by using an "open world assumption": *the fact that something did not happen does not mean that it cannot happen*.
- Using a trial-and-error approach to see whether an event log is suitable for process mining.

C3: Creating Representative Benchmarks

Process mining is an emerging technology. This explains why good benchmarks are still missing.

Standards proposed for process modeling are rather complicated.

Some initial word is done:

- There are various metrics for measuring the quality of process mining results (fitness, simplicity, precision, and generalization).
- Several event logs are publicly available (cf. www.processmining.org).

C4: Dealing with Concept Drift

- The term *concept drift* refers to the situation in which the process is changing while being analyzed.
- Processes may change due to periodic/seasonal changes.

Possible solution: splitting the event log into smaller logs and analyzing the "footprints" of the smaller logs.

C5: Improving the Representational Bias Used for Process Discovery

- Processes that cannot be represented by the target language cannot be discovered.
- This so-called "*representational bias*" used during the discovery process should be a conscious choice and should not be (only) driven by the preferred graphical representation.

C6: Balancing Between Quality Criteria such as Fitness, Simplicity, Precision, and Generalization

C7: Cross-Organizational Mining

- The overall process may be cut into parts and distributed over organizations that need to cooperate to successfully complete cases.
- Analyzing the event log within one of these organizations involved is insufficient.
- We may also consider the setting where different organizations are essentially executing the same process while sharing experiences, knowledge, or a common infrastructure.

C8: Providing Operational Support

- Today many data sources are updated in (near) real-time and sufficient computing power is available to analyze events when they occur.
- Therefore, process mining should not be restricted to off-line analysis and can also be used for online operational support.
- Three operational support activities:
  - detect,
  - predict,
  - recommend.

C9: Combining Process Mining With Other Types of Analysis

- linear programming,
- Project planning,
- queuing models,
- Markov chains,
- Simulation,
- Data mining,
- visual analytics,
- …

C10: Improving Usability for Non-Experts

C11: Improving Understandability for Non-Experts

Authors

Wil van der Aalst
Arya Adriansyah
Ana Karla Alves de Medeiros
Franco Arcieri
Thomas Baier
Tobias Blickle
Jagadeesh Chandra Bose
Peter van den Brand
Ronald Brandtjen
Joos Buijs
Andrea Burattin
Josep Carmona
Malu Castellanos
Jan Claes
Jonathan Cook
Nicola Costantini
Francisco Curbera
Ernesto Damiani
Massimiliano de Leoni
Pavlos Delias
Boudewijn van Dongen
Marlon Dumas
Schahram Dustdar
Dirk Fahland
Diogo R. Ferreira
Walid Gaaloul
Frank van Geffen
Sukriti Goel
Christian Günther
Antonella Guzzo
Paul Harmon
Arthur ter Hofstede
John Hoogland
Jon Espen Ingvaldsen
Koki Kato
Rudolf Kuhn
Akhil Kumar
Marcello La Rosa
Fabrizio Maggi
Donato Malerba
Ronny Mans
Alberto Manuel
Martin McCreesh
Paola Mello
Jan Mendling
Marco Montali
Hamid Motahari Nezhad
Michael zur Muehlen
Jorge Munoz-Gama
Luigi Pontieri
Joel Ribeiro
Anne Rozinat
Hugo Seguel Pérez
Ricardo Seguel Pérez
Marcos Sepúlveda
Jim Sinur
Pnina Soffer
Minseok Song
Alessandro Sperduti
Giovanni Stilo
Casper Stoel
Keith Swenson
Maurizio Talamo
Wei Tan
Chris Turner
Jan Vanthienen
George Varvaressos
Eric Verbeek
Marc Verdonk
Roberto Vigo
Jianmin Wang
Barbara Weber
Matthias Weidlich
Ton Weijters
Lijie Wen
Michael Westergaard
Moe Wynn

Process Mining Book

www.processmining.org/book/

W.M.P. van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin, 2011.