Process Mining and its context

Barbara Re

Process Mining
MSc in Computer Science (LM-18)
University of Camerino

Summary

1. Motivations and Context
2. Process Mining

Summary

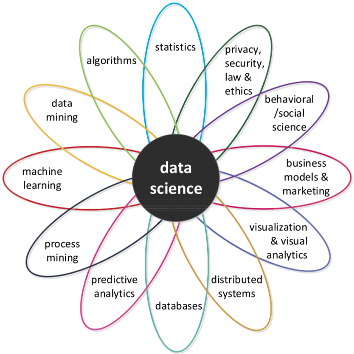
Data Explosion

Data are collected about anything, at any time, and at any place

One of the main challenges of todays organizations is to extract information and value from data stored in their information systems

Scenario

In 2020 Data digitally stored will account to 44 ZB ($1ZB = 2^{70} \approx 10^{21}$B). Most of the data stored in the digital universe is unstructured, and organizations have problems dealing with such large quantities of data



Increasing relevance of Information systems

The importance of information systems is not only reflected by the spectacular growth of data, but also by the role that these systems play in today's business processes as the digital universe and the physical universe are becoming more and more aligned

- The "state of a bank" is mainly determined by the data stored in the bank's information system
- The "real" state of a warehouse is the one in the managing information system, and not the one of the physical world

Internet of Events

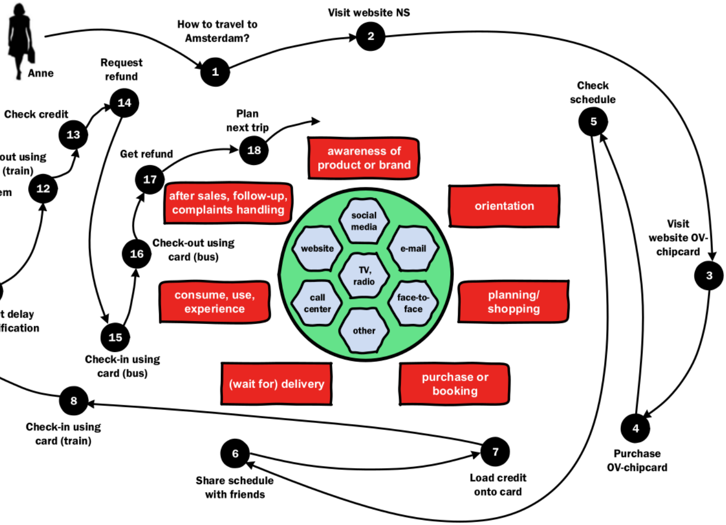
The spectacular growth of the digital universe makes it possible to record, derive, and analyze events

Events may be life events, machine events, or organization events

The term Internet of Events (IoE) refers to all event data available

- The Internet of Content (IoC), i.e., all information created by humans to increase knowledge on particular subjects. The IoC includes traditional web pages, articles, encyclopedia like Wikipedia, YouTube, e-books, newsfeeds, etc.
- The Internet of People (IoP), i.e., all data related to social interaction. The IoP includes e-mail, Facebook, Twitter, forums, LinkedIn, etc.
- The Internet of Things (IoT), i.e., all physical objects connected to the network. The IoT includes all things that have a unique id and a presence in an Internet-like structure.
- The Internet of Locations (IoL) which refers to all data that have a geographical or geospatial dimension. With the uptake of mobile devices (e.g., smartphones) more and more events have location or movement attributes.

Digitization of life and events



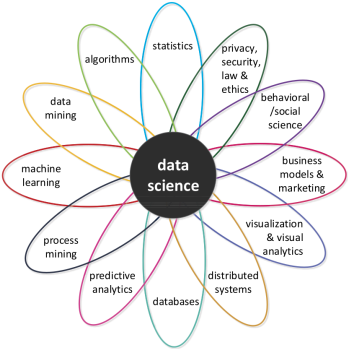
Archetypal customer journey stages

1. Awareness of product or brand: the customer needs to be aware of the product and/or brand to start a customer journey.
2. Orientation: the customer is interested in a product, possibly of a particular brand.
3. Planning/shopping: the customer may decide to purchase a product or service. This requires planning and/or shopping, e.g., browsing websites for the best offer.
4. Purchase or booking. If the customer is satisfied with a particular offering, the product is bought or the service (e.g., flight or hotel) is booked.
5. (Wait for) delivery: this is the stage after purchasing the product or booking the service, but before the actual delivery.
6. Consume, use, experience: the product or service is used. While using the product or service, a multitude of events may be generated. The recorded event data can be used to understand the actual use of the product by the customer.
7. After sales, follow-up, complaints handling: this is the stage that follows the actual use of the product or service. At this seventh stage, new add-on products may be offered (e.g., air filters).

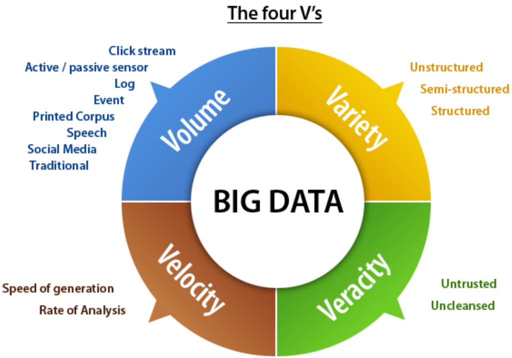
The ingredients contributing to data science

Not a linear process - establishing relationships between events, is one of the key challenges in data science (Event correlation)

Data science is an amalgamation of different partially overlapping (sub)disciplines



Four V's of Big Data



Data Science

Definition

Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes:

- data extraction
- data preparation
- data exploration
- data transformation
- storage and retrieval
- computing infrastructures
- various types of mining and learning
- presentation of explanations and predictions
- exploitation of results taking into account ethical, social, legal, and business aspects

Questions for data scientists

A data scientist can answer a variety of data-driven questions. These can be grouped into the following four main categories:

- Reporting – What happened?
- Diagnosis – Why did it happen?
- Prediction – What will happen?
- Recommendation – What is the best that can happen?

Process mining as the bridge between data science and process science

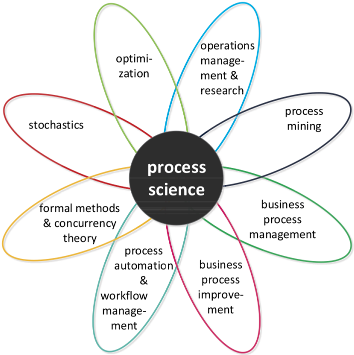
Summary

1. Motivations and Context
2. Process Mining

Process Mining

Process Science

Process science is an umbrella term for the broader discipline that combines knowledge from information technology and knowledge from management sciences to improve and run operational processes



Goals

The goal of process mining is to use event data to extract process-related information, e.g., to automatically discover a process model by observing events recorded by some enterprise system

Models and Reality

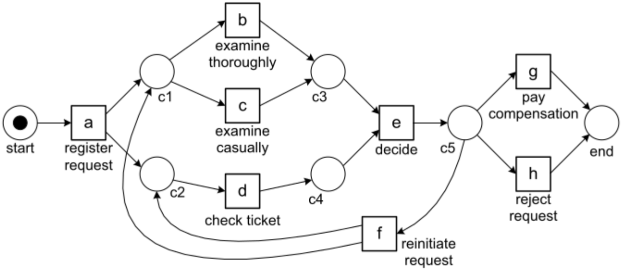
Models are abstractions and languages are needed to express them and many different notations to express models and run related activities:

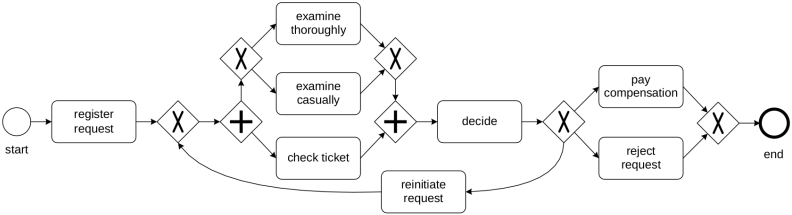
- Formal vs. Informal Notations
- PN, BPMN, UML Activity, EPC, . . .

But

- Executable models may be used to force people to work in a particular manner
- However, most models are not well-aligned (or time passing get misaligned) with reality
- Most hand-made models are disconnected from reality and provide only an idealized view on the processes at hand: "paper tigers"

Example – PN vs. BPMN





PAIS

Process-Aware Information Systems

Software systems that support processes and not just isolated activities (e.g. ERP (Enterprise Resource Planning) systems (SAP, Oracle, etc.), BPM (Business Process Management) systems (Pegasystems, Bizagi, Appian, IBM BPM, etc.), WFM (Workflow Management) systems, CRM (Customer Relationship Management) systems, rule-based systems, call center software, high-end middleware (WebSphere), etc.)

There is a process notion present in the software (e.g., the completion of one activity triggers another activity) and that the information system is aware of the processes it supports (e.g., collecting information about flow times).

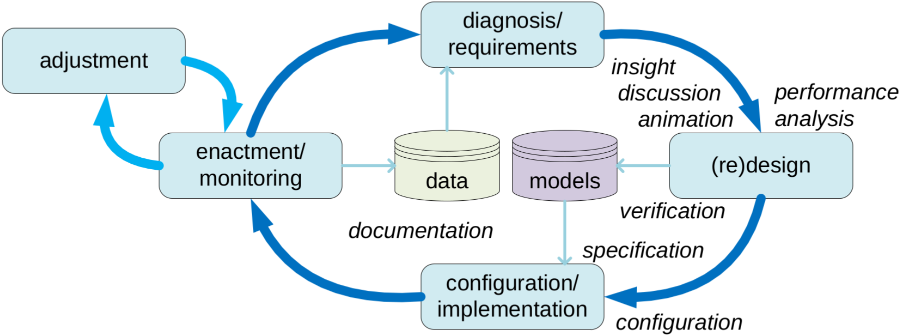
A particular class of PAISs is formed by generic systems that are driven by explicit process models, changing the model corresponds (in theory) to automatically changing the process.

What are process models used for?

Process Models are defined and used for several reasons:

- insight: while making a model, the modeler is triggered to view the process from various angles
- discussion: the stakeholders use models to structure discussions
- documentation: processes are documented for instructing people or certification purposes (cf. ISO 9000 quality management)
- verification: process models are analyzed to find errors in systems or procedures (e.g., potential deadlocks)
- performance analysis: techniques like simulation can be used to understand the factors influencing response times, service levels, etc.
- animation: models enable end users to “play out” different scenarios and thus provide feedback to the designer
- specification: models can be used to describe a PAIS before it is implemented and can hence serve as a “contract” between the developer and the end user/management
- configuration: models can be used to configure a system

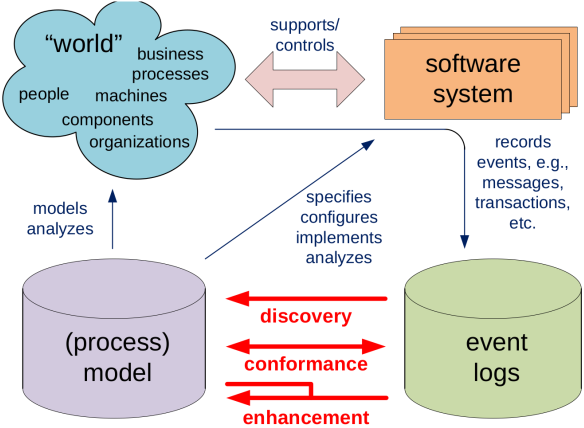
BPM life-cycle



Process Mining - flavours and ingredients

Opportunity

Given (a) the interest in process models, (b) the abundance of event data, and (c) the limited quality of hand-made models, it seems worthwhile to relate event data to process models



Perspectives

- The control-flow perspective focuses on the control-flow, i.e., the ordering of activities
- The organizational perspective focuses on information about resources hidden in the log, i.e., which actors (e.g., people, systems, roles, and departments) are involved and how are they related
- The case perspective focuses on properties of cases, e.g., cases can also be characterized by the values of the corresponding data elements
- The time perspective is concerned with the timing and frequency of events

Starting point: the 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			
	35654488	05-01-2011:11.22	decide			
	35654489	08-01-2011:12.05	pay compensation			
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	...

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.

Starting point: data preparation and transformation

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

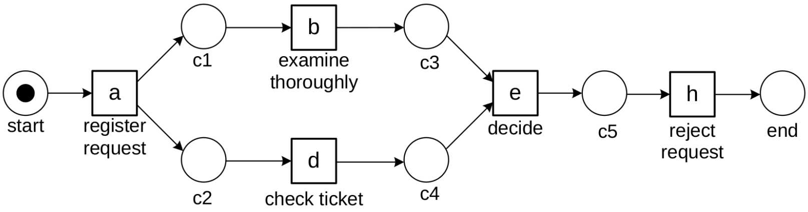
From events log to process models

The problem of automatically infering a model from observed data is an old one

- In the formal language area is referred as Grammar inference
- We do not reinvent the wheel: the α-algorithm has been the starting point for many other techniques

Example

$\mathcal{L} = \{\langle a, b, d, e, h\rangle, \langle a, d, b, e, h\rangle\}$



More traces

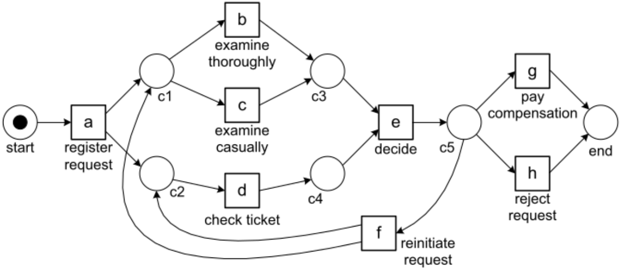
case id	trace
1	$\langle a,b,d,e,h \rangle$
2	$\langle a,d,c,e,g \rangle$
3	$\langle a,c,d,e,f,b,d,e,g \rangle$
4	$\langle a,d,b,e,h \rangle$
5	$\langle a,c,d,e,f,d,c,e,f,c,d,e,h \rangle$
6	$\langle a,c,d,e,g \rangle$
...	...

Does the previous model fits wrt the Event log?

More traces

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	$\langle a,c,d,e,g \rangle$
...	...

Does the previous model fits wrt the Event log?



Mining techniques phenomenons

Let's consider additional traces that could be observed: $\langle a, b, e, g \rangle$, $\langle a, d, c, e, f, d, c, e, f, b, d, e, h \rangle$, $\langle a, c, d, e, f, b, d, g \rangle$ are them permitted by the model?

How can we judge the quality of the model then?

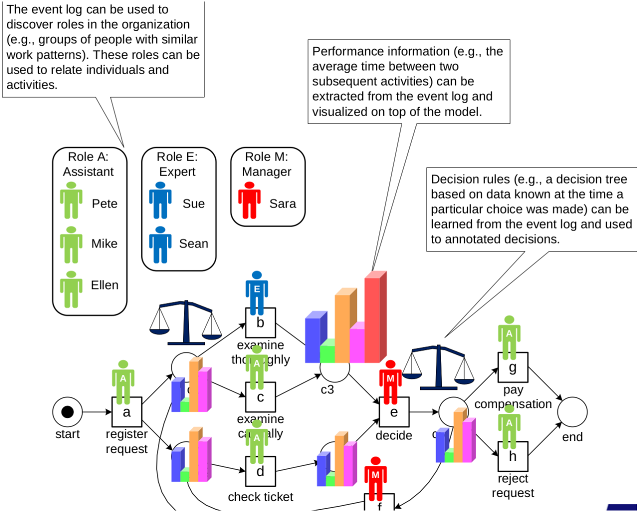
Mining techniques phenomenons

Let's consider additional traces that could be observed: $\langle a, b, e, g\rangle$, $\langle a, d, c, e, f, d, c, e, f, b, d, e, h\rangle$, $\langle a, c, d, e, f, b, d, g\rangle$ are them permitted by the model?

How can we judge the quality of the model then?

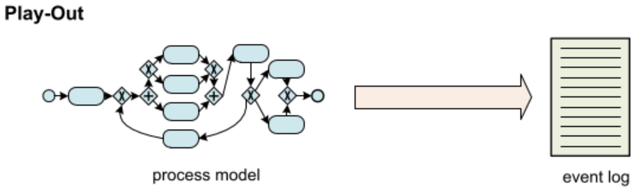
- Overfitting - It means that the generated model is too specific and only admits behaviour similar to that observed
- Underfitting - It means that the generated model is too general which also accepts behaviours that are probably unrelated to the observed one

Extensions


The event log can be used to discover roles in the organization (e.g., groups of people with similar work patterns). These roles can be used to relate individuals and activities.
Performance information (e.g., the average time between two subsequent activities) can be extracted from the event log and visualized on top of the model.
Decision rules (e.g., a decision tree based on data known at the time a particular choice was made) can be learned from the event log and used to annotated decisions.
Role A: Assistant
Pete
Mike
Ellen
Role E: Expert
Sue
Sean
Role M: Manager
Sara
start
register request
examine thoroughly
examine casually
check ticket
c3
decide
pay compensation
reject request
end

Play-out

Key elements of process mining is the emphasis on establishing a strong relation between a process model and the "reality" captured in the form of an event log

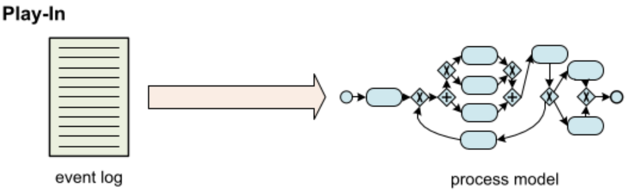


Play-out

Given a model it is possible to generate behaviour, the traces are obtained by repeatedly "playing the token game"

- Simulation tools also use a Play-Out engine to conduct experiments
- Classical verification approaches using exhaustive state-space analysis can be seen as Play-Out methods

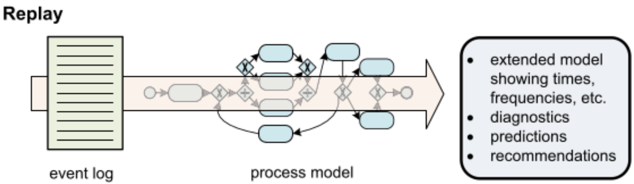
Play-in



Play-in

Example behaviour is taken as input and the goal is to construct a model, play-In is often referred to as inference

Replay



Replay

Uses an event log and a process model as input, and the event log is “re- played” on top of the process model

An event log may be replayed for different purposes:

- Conformance checking
- Extending the model with frequencies and temporal information
- Constructing predictive models
- Operational support