Process Mining - 02269

Lecture 2

Events logs

Andrea Burattin

Slides based on material from Matthias Weidlich and Wil van der Aalst

The Context

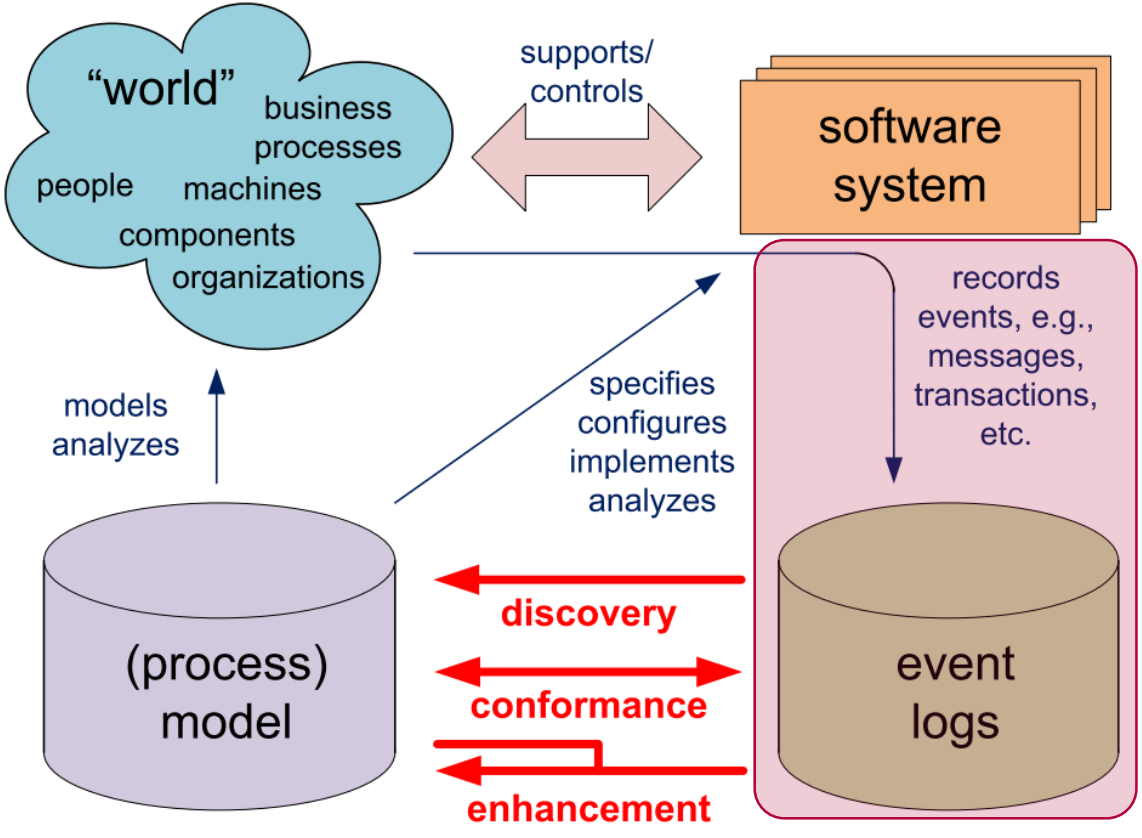


Image source: W. van der Aalst, "Process Mining", 2nd ed, Springer 2016.

Logs as an Information Source

- Logs contain information to answer questions
  - When have process instances been executed?
  - How many instances have been executed?
  - Have there been recurring patterns in the executions of activities?
  - Is it possible to construct process models based on the log data?
  - Which sequences of activities have been executed very frequently?
  - Does a process model contain execution sequences that have never been executed?
- Logs are the basis for evidence-based answers to these questions
  - Not biased by human perception of how a process is conducted
  - Not biased by fragmentation of process knowledge
  - Yet, assuming high data quality

Log Entries

- Example log entries
  - *Check of invoice with number 4567 finished on 12.11.2010 at 9:19:57*
  - *StoreCustomerData(“Müller”, c1987, “Bad Bentheim”) executed on 12.11.2010 at 9:22:24*
  - *Invoice sent for invoice number 4567 finished on 12.11.2010 at 9:23:18*
  - *Inserted data (c1987, PromoMailing) into customer database on 12.11.2010 at 9:24:10*
  - *StoreCustomerData(”Miller”, c1988, “Osnabrück”) executed on 12.11.2010 at 9:26:08*
  - *Check of invoice with number 4568 finished on 12.11.2010 at 9:26:38*

From heterogeneous data sources to process mining

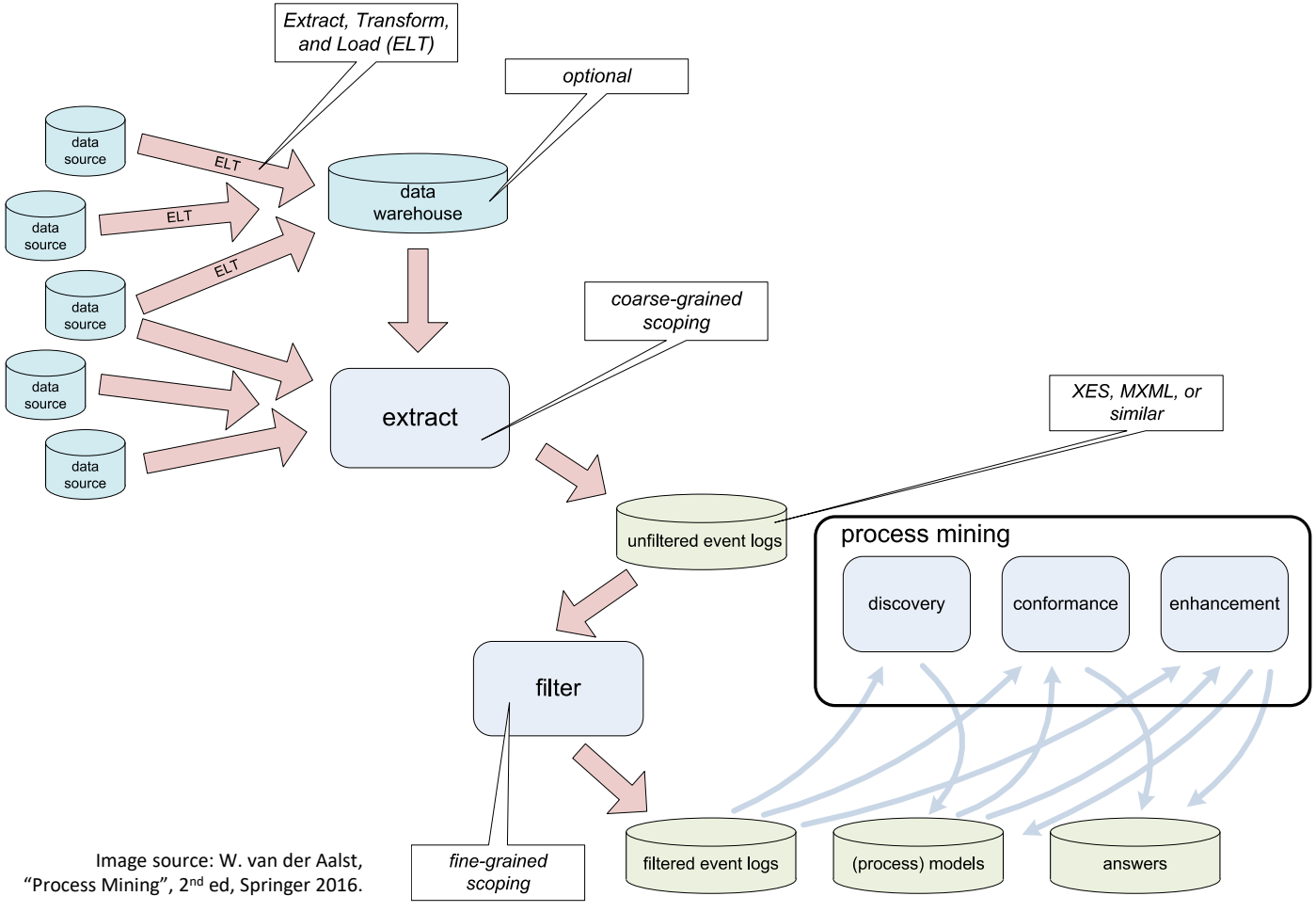
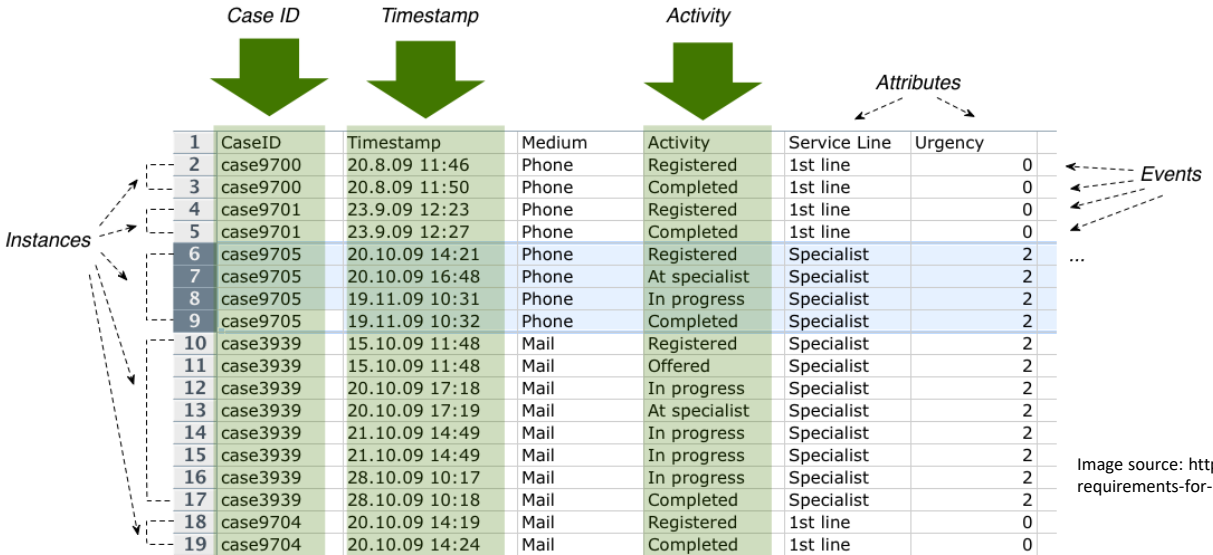


Image source: W. van der Aalst, "Process Mining", 2nd ed, Springer 2016.

Example of event log

- A process consists of **cases**
- A case consists of **events** such that each event relates to precisely one case
- Events within a case are **ordered**
- Events can have **attributes**
  - Examples of typical attribute names are **activity**, **time**, costs, and resource



	CaseID	Timestamp	Medium	Activity	Service Line	Urgency
2	case9700	20.8.09 11:46	Phone	Registered	1st line	0
3	case9700	20.8.09 11:50	Phone	Completed	1st line	0
4	case9701	23.9.09 12:23	Phone	Registered	1st line	0
5	case9701	23.9.09 12:27	Phone	Completed	1st line	0
6	case9705	20.10.09 14:21	Phone	Registered	Specialist	2
7	case9705	20.10.09 16:48	Phone	At specialist	Specialist	2
8	case9705	19.11.09 10:31	Phone	In progress	Specialist	2
9	case9705	19.11.09 10:32	Phone	Completed	Specialist	2
10	case3939	15.10.09 11:48	Mail	Registered	Specialist	2
11	case3939	15.10.09 11:48	Mail	Offered	Specialist	2
12	case3939	20.10.09 17:18	Mail	In progress	Specialist	2
13	case3939	20.10.09 17:19	Mail	At specialist	Specialist	2
14	case3939	21.10.09 14:49	Mail	In progress	Specialist	2
15	case3939	21.10.09 14:49	Mail	In progress	Specialist	2
16	case3939	28.10.09 10:17	Mail	In progress	Specialist	2
17	case3939	28.10.09 10:18	Mail	Completed	Specialist	2
18	case9704	20.10.09 14:19	Mail	Registered	1st line	0
19	case9704	20.10.09 14:24	Mail	Completed	1st line	0

Image source: https://fluxicon.com/blog/2012/02/data-requirements-for-process-mining/

Tree structure of an event log

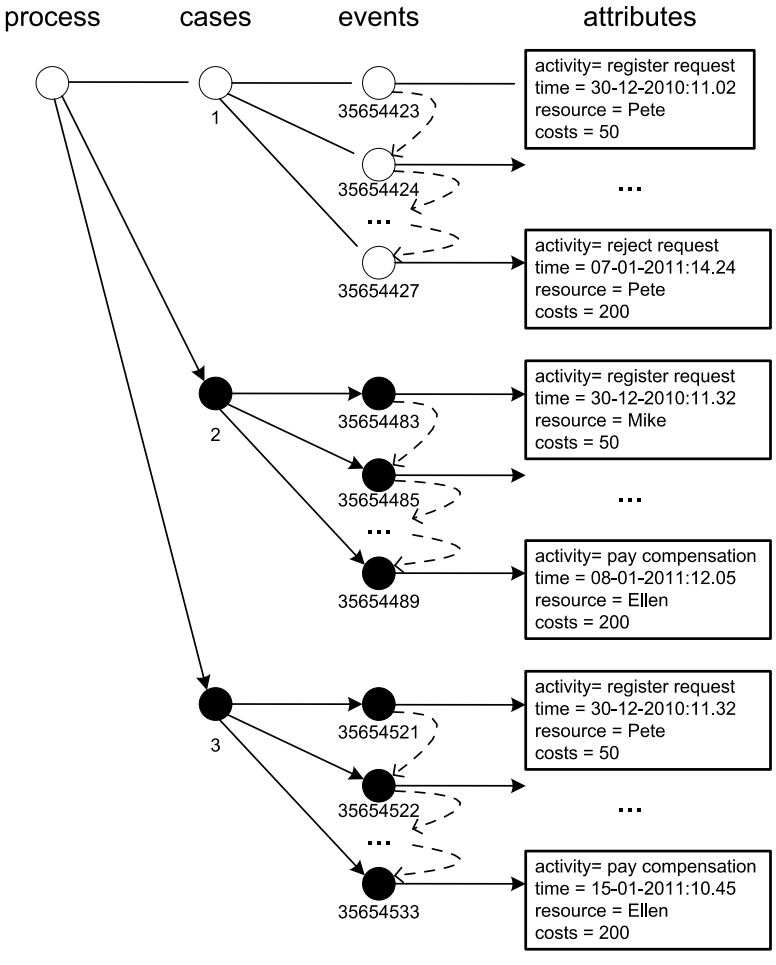


Image source: W. van der Aalst, "Process Mining", 2nd ed, Springer 2016.

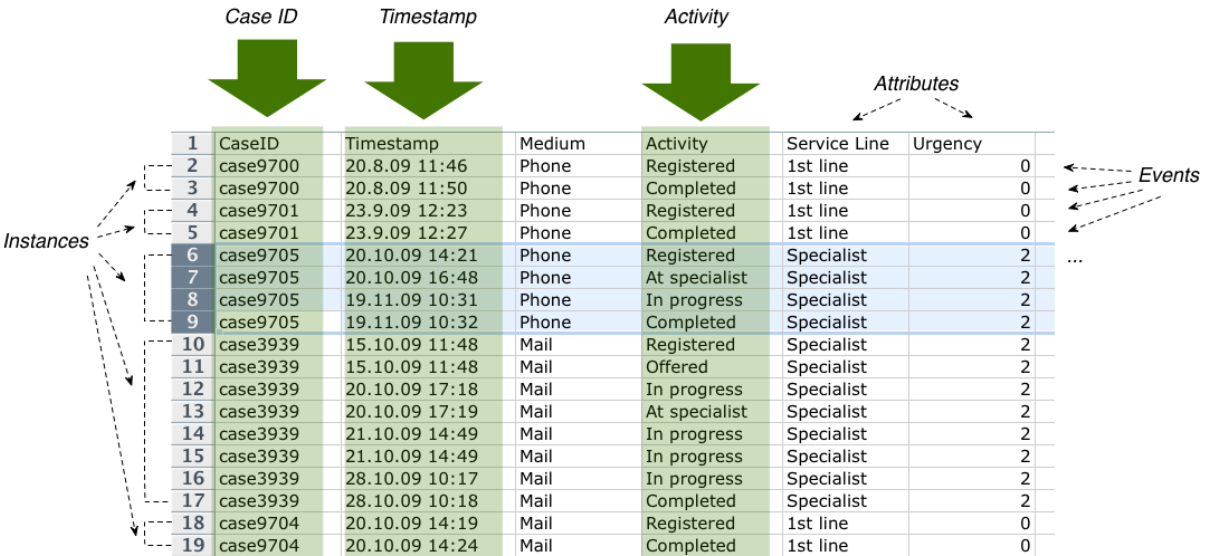
Event log terminology

- We assume the presence of an **event log**
- An event log is a collection of **cases**
- A case is a *trace* (or *sequence*) of **events**
- Each event refers to a case (case id), an activity and a point in time
  - As seen, events can have many more attributes

Event data might come from any source and format

- A database system (e.g., patients in a hospital)
- Transaction logs (e.g., a trading system)
- ERP systems (e.g., Oracle, SAP)
- API to social media/websites (e.g., Twitter or Facebook)

- CSV files
- Spreadsheet
- …



	CaseID	Timestamp	Medium	Activity	Service Line	Urgency
2	case9700	20.8.09 11:46	Phone	Registered	1st line	0
3	case9700	20.8.09 11:50	Phone	Completed	1st line	0
4	case9701	23.9.09 12:23	Phone	Registered	1st line	0
5	case9701	23.9.09 12:27	Phone	Completed	1st line	0
6	case9705	20.10.09 14:21	Phone	Registered	Specialist	2
7	case9705	20.10.09 16:48	Phone	At specialist	Specialist	2
8	case9705	19.11.09 10:31	Phone	In progress	Specialist	2
9	case9705	19.11.09 10:32	Phone	Completed	Specialist	2
10	case3939	15.10.09 11:48	Mail	Registered	Specialist	2
11	case3939	15.10.09 11:48	Mail	Offered	Specialist	2
12	case3939	20.10.09 17:18	Mail	In progress	Specialist	2
13	case3939	20.10.09 17:19	Mail	At specialist	Specialist	2
14	case3939	21.10.09 14:49	Mail	In progress	Specialist	2
15	case3939	21.10.09 14:49	Mail	In progress	Specialist	2
16	case3939	28.10.09 10:17	Mail	In progress	Specialist	2
17	case3939	28.10.09 10:18	Mail	Completed	Specialist	2
18	case9704	20.10.09 14:19	Mail	Registered	1st line	0
19	case9704	20.10.09 14:24	Mail	Completed	1st line	0

Notions of a case

- The definition of a process instance is not always rigid or defined *a priori*

- Example scenario: e-mail as an event
- What is a possible mapping of an inbox to an event log? (i.e., which field is the activity name, the case id, etc)
- An e-mail has:
  - A sender (“from”)
  - A set of recipients (“to”)
  - A subject
  - A timestamp
  - A body
  - Other attributes…

Notion of case for emails

- One of the possible mappings
  - A sender (“from”) ↔ resource, activity name
  - A set of recipients (“to”) ↔ other attributes
  - A subject ↔ case id
  - A timestamp ↔ timestamp
  - A body ↔ other attributes
  - Other attributes…

- Other mappings might be meaningful as well… it depends on the context and on the questions we are answering

Notion of case for student data

- An event is an exam attempt by a student. It contains
  - Student id
  - Student gender
  - Student nationality
  - Couse
  - Exam data
  - Mark

- What is the case id and the activity name?

Standard transactional life-cycle of activities

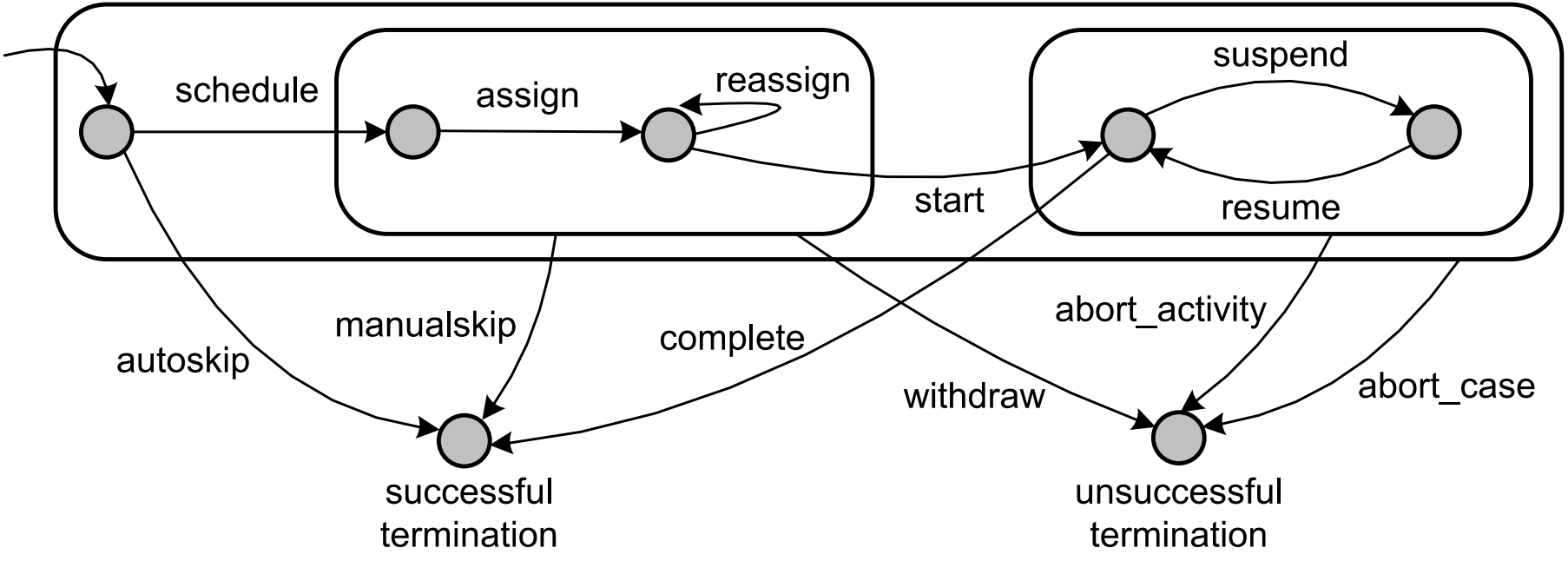
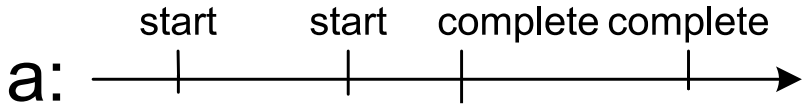


Image source: W. van der Aalst, "Process Mining", 2nd ed, Springer 2016.

Overlapping activity instances

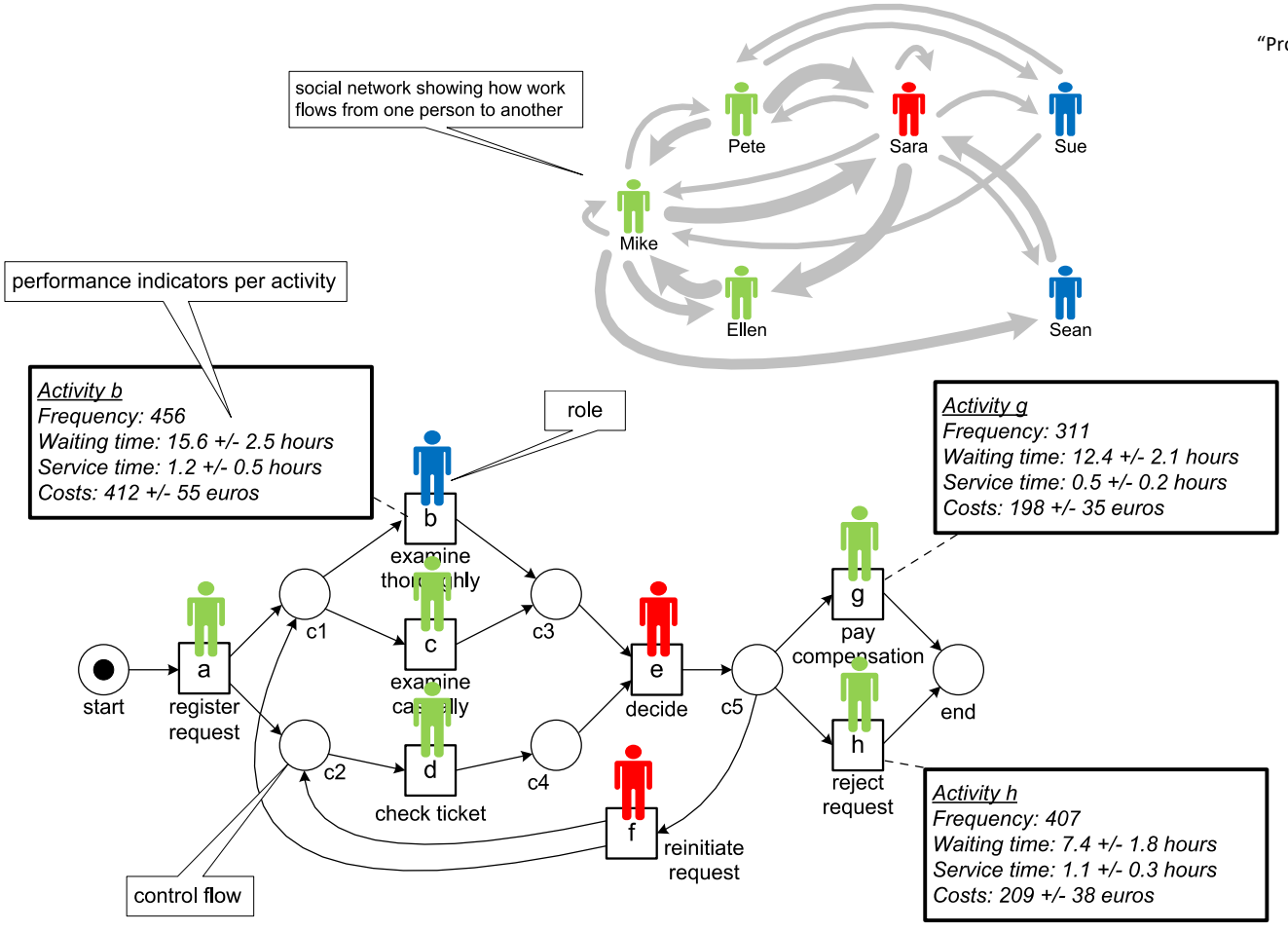
- Not only it is necessary to correlate events to process instance, but events might require a “secondary correlation”, i.e., correlate two events to the same activity



- Solutions: add more information or use heuristics (e.g., first-in-first-out order)
- See also Allen, J. “*Maintaining knowledge about temporal intervals*”. Communications of the ACM. 26 (11): 832–843, 1983. https://doi.org/10.1145%2F182.358434

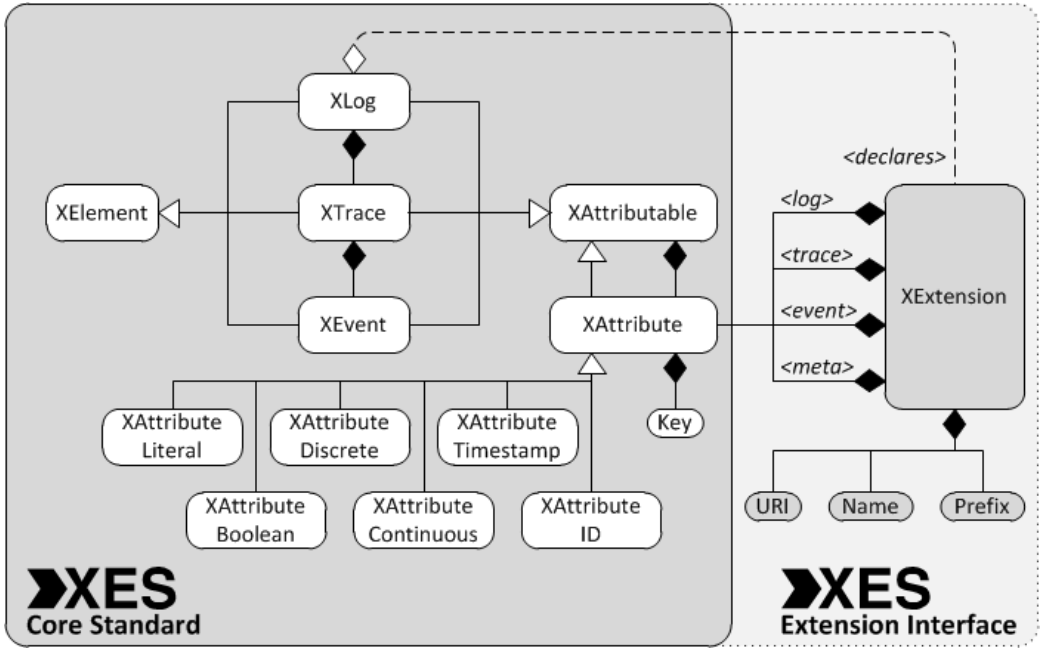
Possible uses of the attributes

Image source: W. van der Aalst, "Process Mining", 2nd ed, Springer 2016.



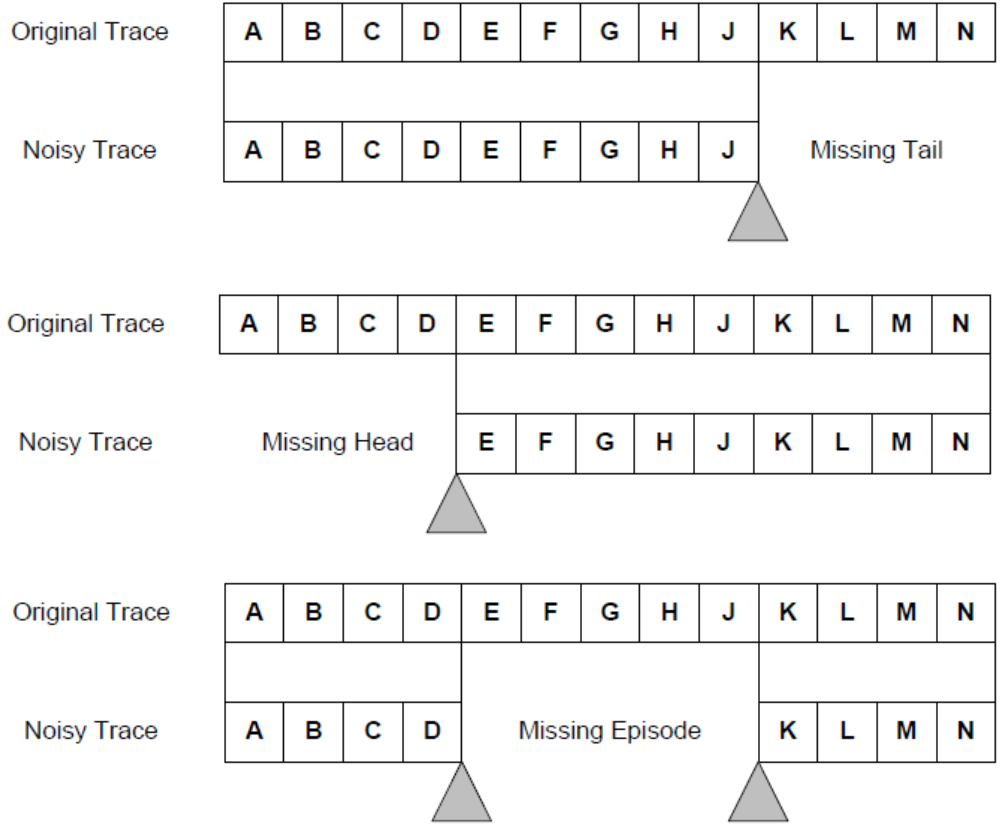
XES (eXtensible Event Stream)

- IEEE Standard, https://standards.ieee.org/standard/1849-2016.html
- Supported by most commercial and open source vendors
- There are possibilities to convert from CSV to XES and vice versa
- XML syntax with OpenXES library open source
- More info https://xes-standard.org/

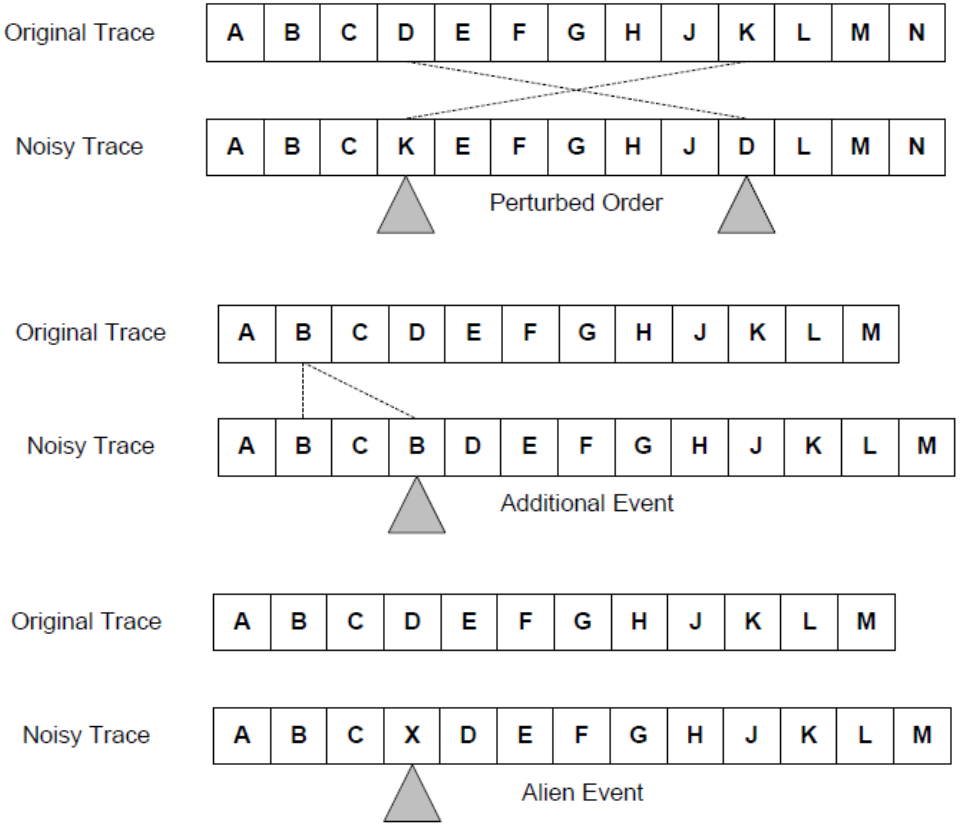


Types of Noise

- Logging was temporarily not available



Types of Noise cont.



Consequences of Noise

- Massive impact on discovery, conformance, and enhancement techniques – we will get back to this
- Already an issue in the construction of event logs

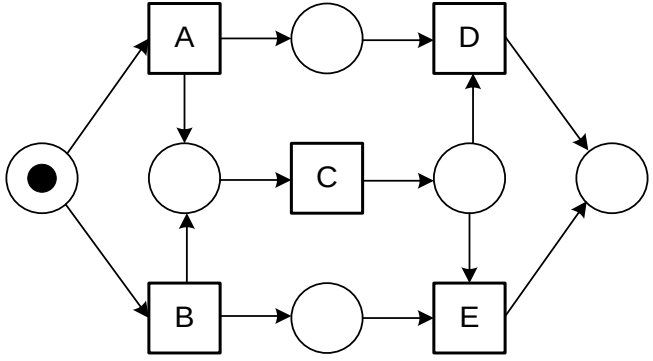
- Major issue: noise is close to impossible to characterise without domain knowledge

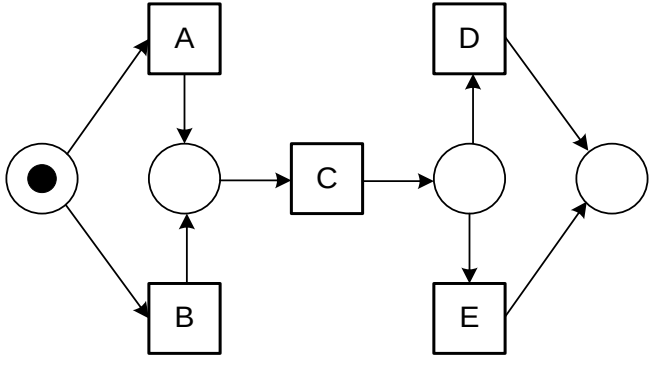
Noise Example

ACD	99
ACE	0
BCE	85
BCD	0

Noise Example (cont.)

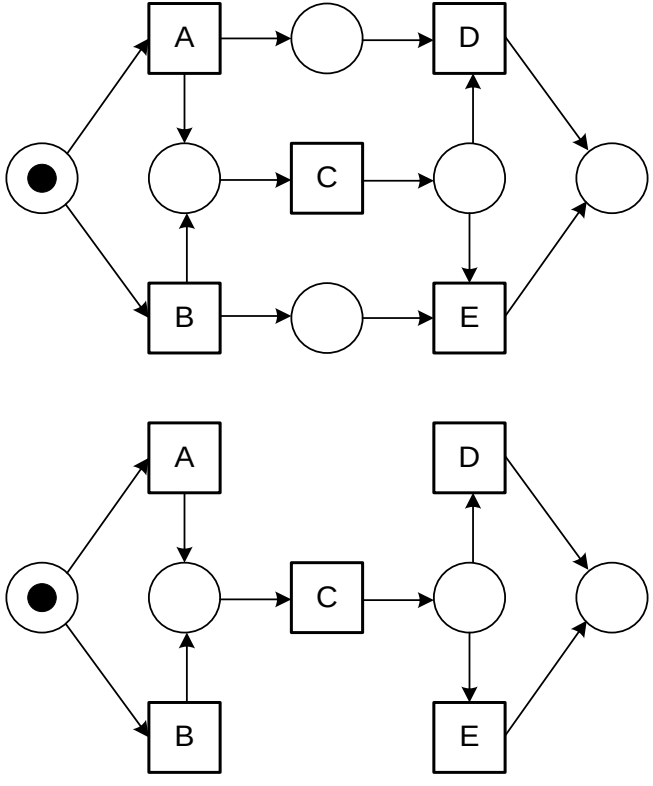
ACD	99
ACE	88
BCE	85
BCD	78





Noise Example (cont.)

ACD	99
ACE	2
BCE	85
BCD	3



?

Log-based Noise Handling

- Rely on frequency analysis to identify noise in event log

- Assumption: noise is rare
  - Very infrequent traces can be considered noise
  - Traces that contain very infrequent transitions can be considered noise
  - Operationalization based on standard data mining techniques – association rules mining

- Again, this assumption may be wrong!

Practical Considerations

- Event logs take various different forms and instantiations
- Differences in semantics, e.g., related to
  - Timestamps
  - Total vs. partial order
- Difference in quality, e.g., related to
  - Completeness
  - Noise-level
  - Data richness
- Technical alignment by means of standards
- But: semantic alignment a major issue