C | A | U

Christian-Albrechts-Universität zu Kiel

Advanced Process Mining

Prof. Dr. Agnes Koschmider

Lecture 7: Event Log Quality

Process Analytics

Lecture Overview

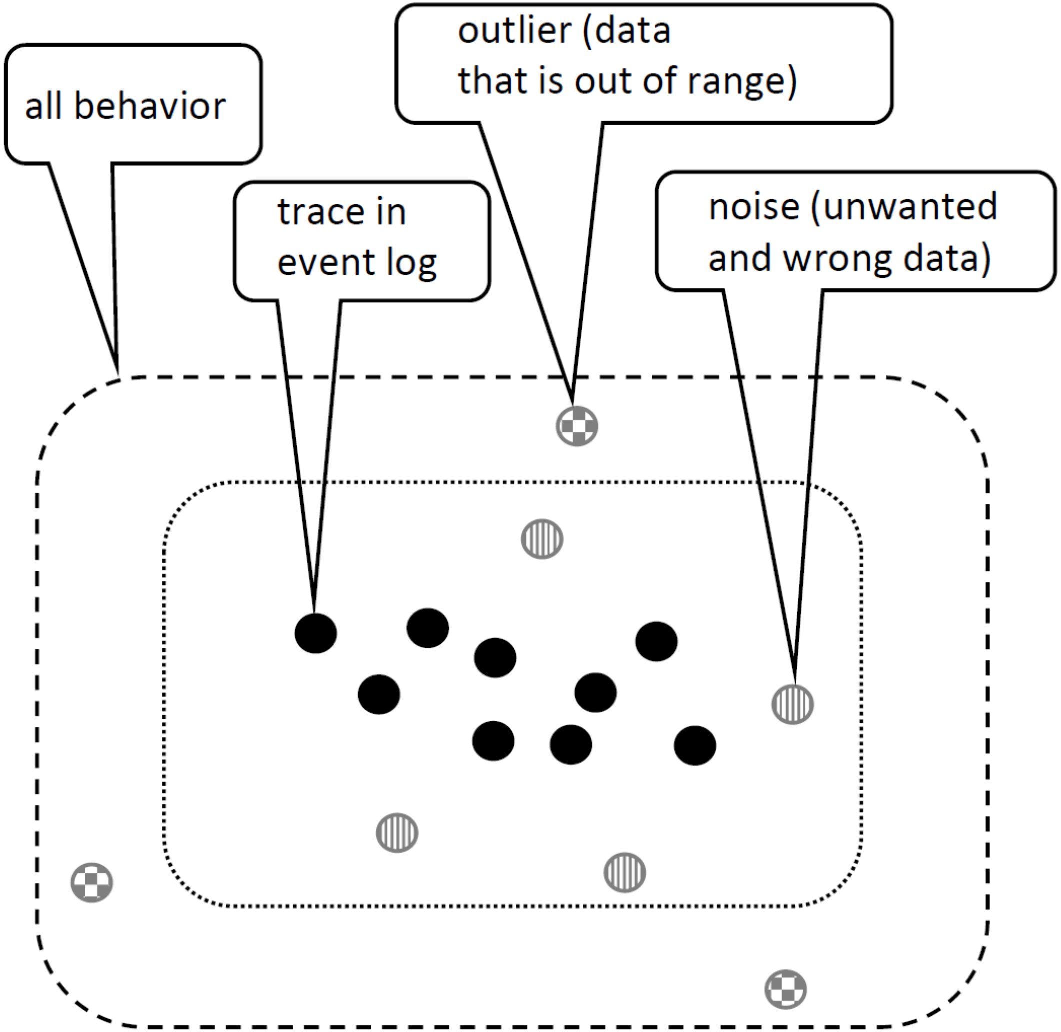
0	Organization and Introduction
I	Process Discovery
II	Process Conformance
III	Predictive Process Monitoring
▶ IV	Event Log Preparation
V	Practical Tasks

Event Log Quality

- quality of the data presented to process modeling algorithms is critical to the success of any process mining exercise

- Pre-processing (cleaning) event logs to address quality issues prior to conducting a process mining analysis is necessary, but time-consuming task

Outlier vs. Noise


all behavior
trace in event log
outlier (data that is out of range)
noise (unwanted and wrong data)

Outlier

- outlier detection is an essential task in process mining
- commonly termed ”anomalies”
- something that differs considerably from all or most other behavior in an event log
- includes divergent data
- negatively influence the usefulness of the discovered process model
- also refer to interesting and useful information about the underlying system

Noise

- refers to bad measurement in data caused by e.g., erroneous recording during process execution
- mistakes introduced into data
- any undesirable or unwanted value

Types of Outliers

point as outlier:

- in the context of an event log point outlier would be an activity, which significantly deviates from the rest of the traces
- L = { ⟨A, B, C, E, F⟩ , ⟨A, B, C, D, E, F⟩ , ⟨A, A, B, C, E, F⟩ }
- the third trace is the only trace containing two A’s where the second A is an outlier

Types of Outliers

context as outliers (conditional anomalies):

- is given if an observation is uncommon in a certain context but not unexpected in another context
- context dimensions in process mining: personal & social, task, environmental and spatial-temporal
- A contextual outlier could be a trace that deviates significantly based on a selected context.

Types of Outliers

subsequence as outliers:

- a subset of the trace deviating significantly from the whole trace
- even if the individual activities in the subset may not be outliers
- L={ ⟨A, B, C, E, F, G⟩ , ⟨A, B, C, D, E, F, G⟩ , ⟨A, B, C⟩ } the last trace is an outlier since there are very few traces with the same length

Noise: Attribute Noise

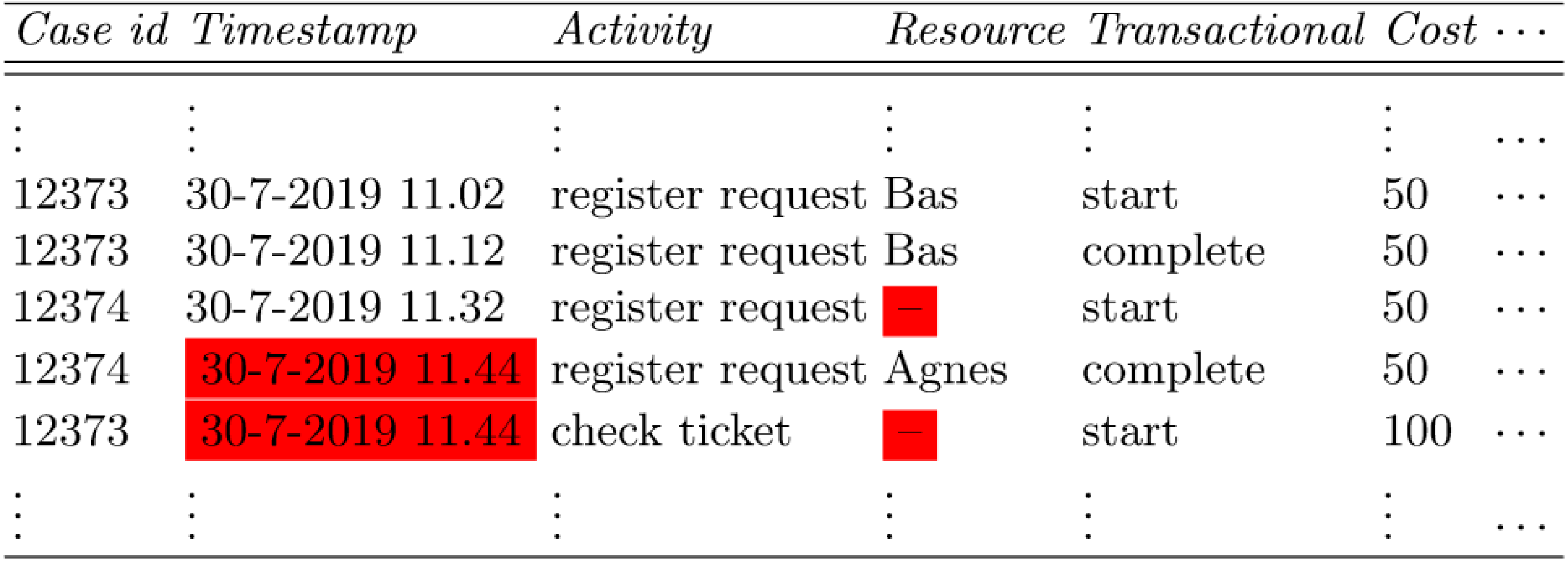
- arises when imprecision or an error is introduced to one or more attributes
- can be totally unpredictable i.e., random, or simply a low variation with respect to the correct value
- types: erroneous attribute values, missing or don’t know values and incomplete or don’t care values
- can arise at event, activity, trace of the log level

Noise: Attribute Noise

- Event might contain erroneous values due to a logging error that recorded identical timestamp for several events
- Missing values might arise due to e.g., faults in sensor devices
- Erroneous activity value might result from unknown attribute values
- Incomplete attributes might occur due to irregularities in sampling
- Trace noise occur when activities were not collected at all

Event Log with Attribute Noise

- Event log with attribute noise (highlighted in red) due to identical timestamps and information that was not recorded for resources.

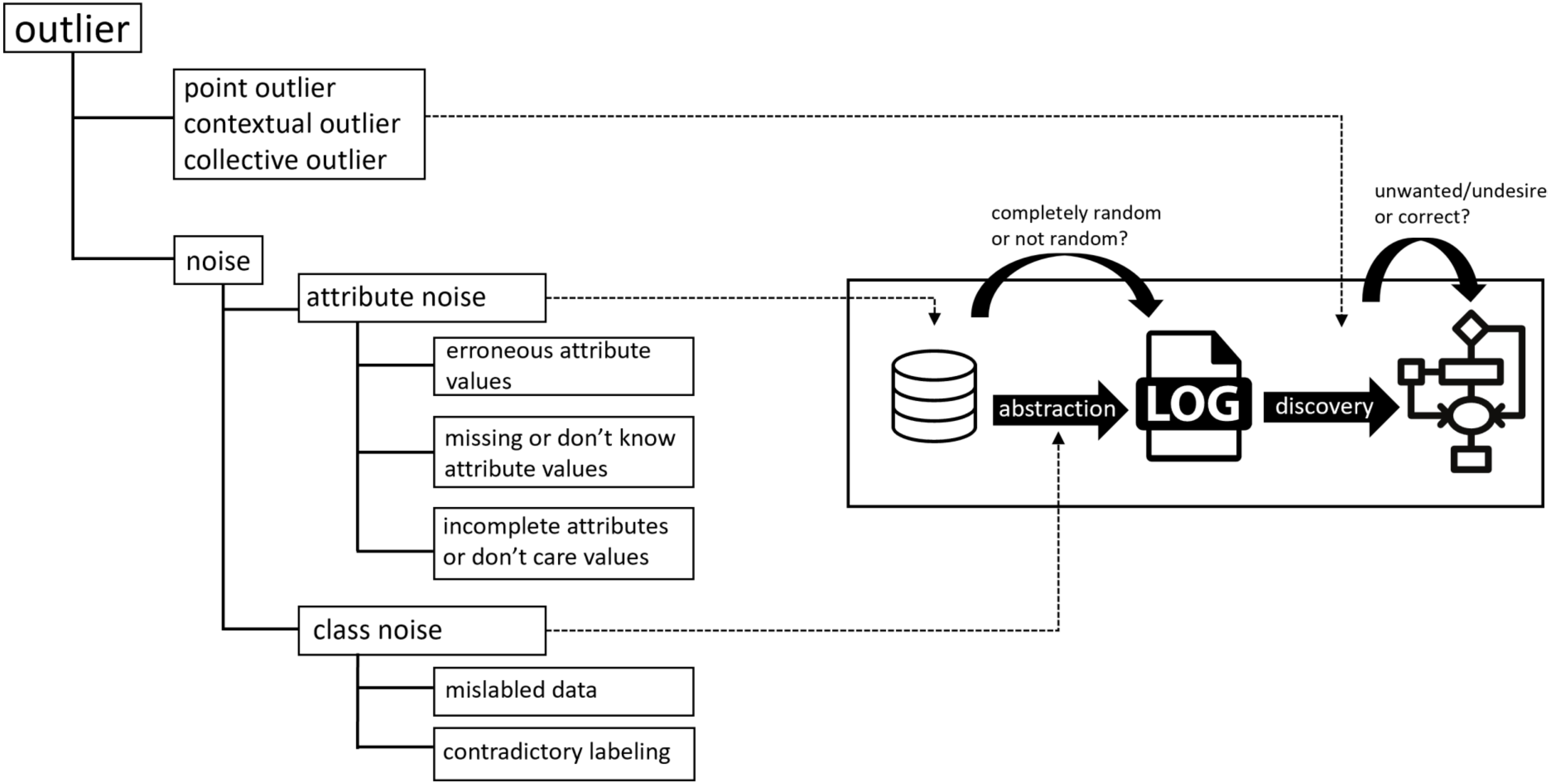


Case id	*Timestamp*	*Activity*	*Resource*	*Transactional*	*Cost*	···
⋮	⋮	⋮	⋮	⋮	⋮	...
12373	30-7-2019 11.02	register request	Bas	start	50	···
12373	30-7-2019 11.12	register request	Bas	complete	50	···
12374	30-7-2019 11.32	register request	–	start	50	···
12374	30-7-2019 11.44	register request	Agnes	complete	50	···
12373	30-7-2019 11.44	check ticket	–	start	100	···
⋮	⋮	⋮	⋮	⋮	⋮	...

Noise: Class Noise

- caused by contradictory labeling or mislabeled data
- activities were wrongly labeled during event-activity abstraction
- are contradictory labeled due to undetected homonyms or synonyms in the data set

Classification of outliers and noise



outlier

point outlier
contextual outlier
collective outlier

noise

attribute noise

erroneous attribute values

missing or don’t know attribute values

incomplete attributes or don’t care values

class noise

mislabled data

contradictory labeling

completely random or not random?

unwanted/undesired or correct?

abstraction

discovery

Techniques for Outlier Detection

1. Density Based Outlier Detection
2. Distance Based Outlier Detection
3. Clustering Based Outlier Detection
4. Partition Based Outlier Detection

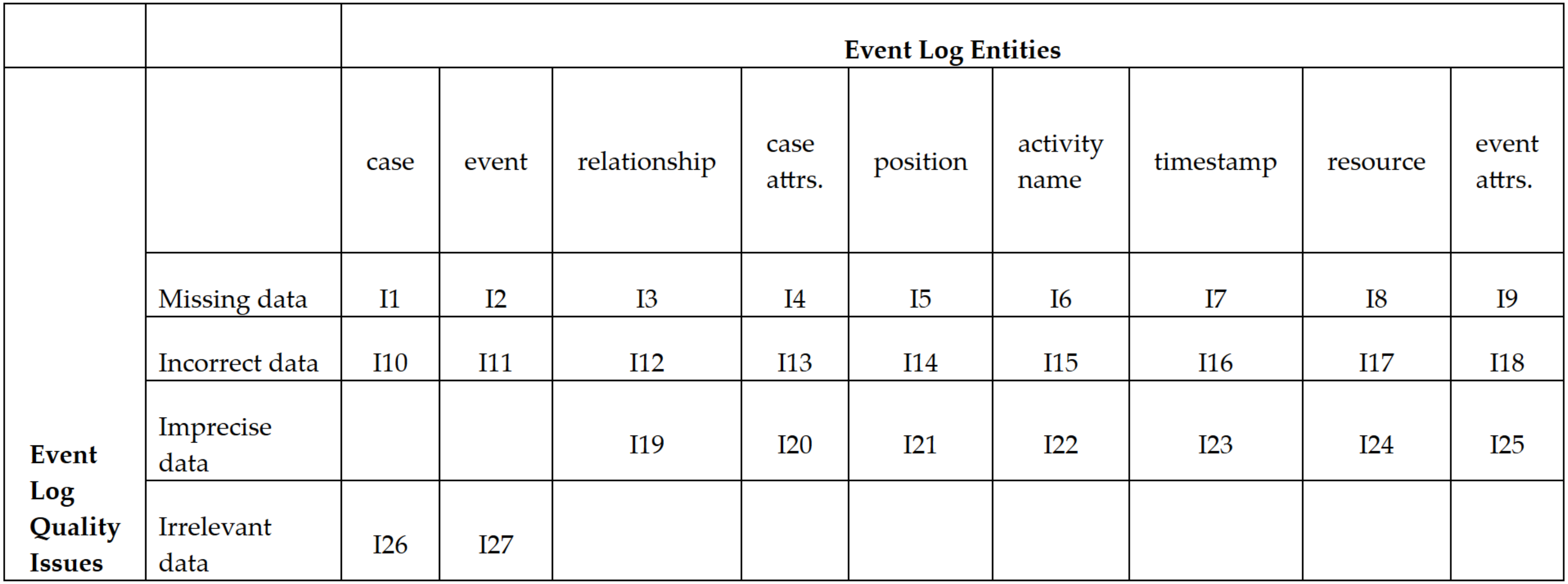
Quality of the Event Log

- **Missing Data:** different kinds of information can be missing in a log although it is mandatory
- **Incorrect Data:** data may be provided in a log, it may turn out that, based on context information, the data is logged incorrectly
- **Imprecise Data:** the logged entries are too coarse leading to a loss of precision
- **Irrelevant Data:** logged entries may be irrelevant as it is for analysis but another relevant entity may have to be derived/obtained (e.g., through filtering/aggregation) from the logged entities

source: http://www.workflowpatterns.com/patterns/logimperfection/

Quality issues in event log attributes

Event Log Imperfection Patterns



		Event Log Entities								
Event Log Quality Issues		case	event	relationship	case attrs.	position	activity name	timestamp	resource	event attrs.
	Missing data	I1	I2	I3	I4	I5	I6	I7	I8	I9
	Incorrect data	I10	I11	I12	I13	I14	I15	I16	I17	I18
	Imprecise data			I19	I20	I21	I22	I23	I24	I25
	Irrelevant data	I26	I27							

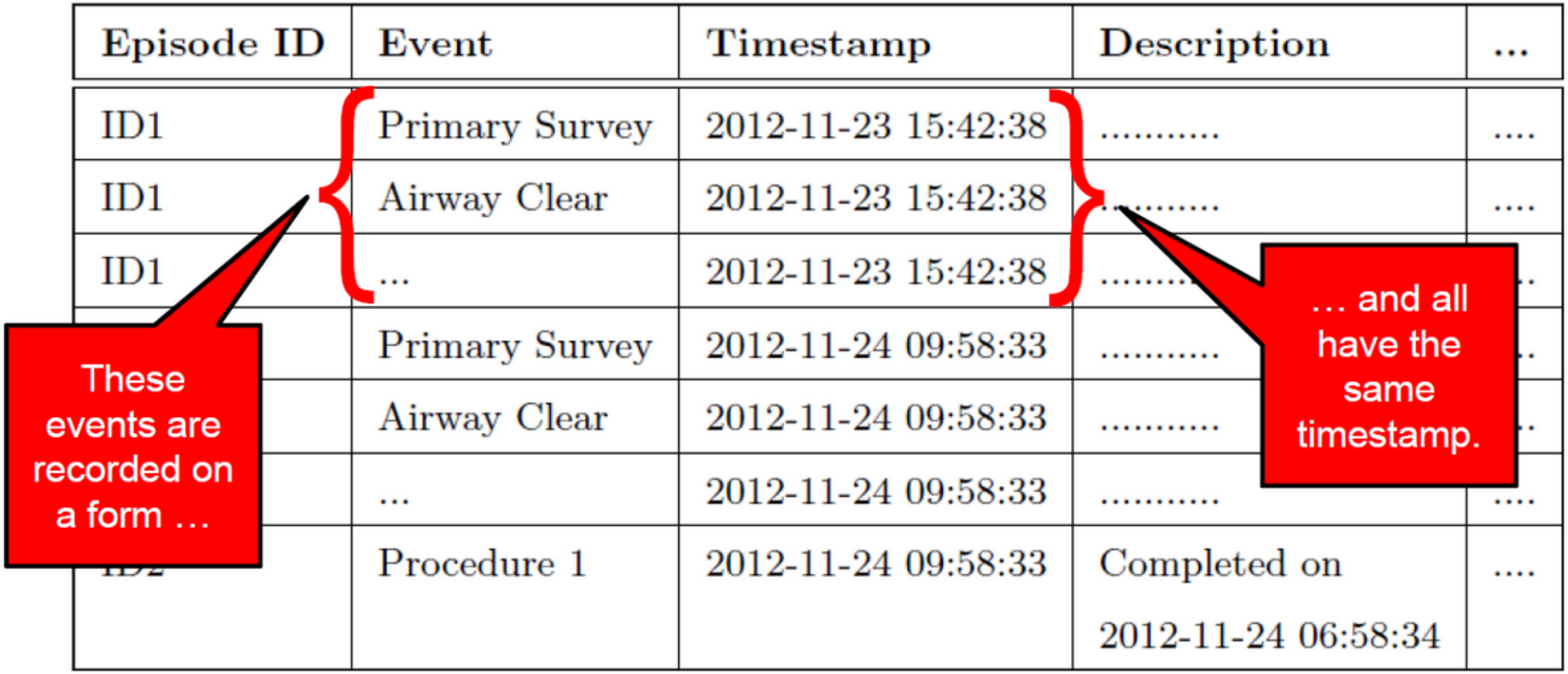
source: http://www.workflowpatterns.com/patterns/logimperfection/

Event Log Imperfection Patterns

1. Form-based Event Capture
2. Inadvertent Time Travel
3. Unanchored Event
4. Scattered Event
5. Elusive Case
6. Scattered Case
7. Collateral Events
8. Polluted Label
9. Distorted Label
10.Synonymous Labels
11.Homonymous Label

Form-based Event Capture

- I16 - Incorrect data: timestamp
- I27 - Irrelevant data: event



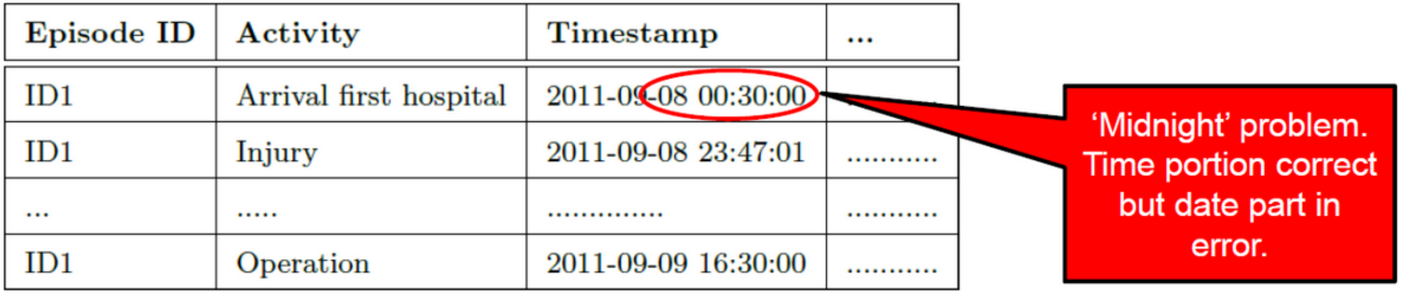
Episode ID	Event	Timestamp	Description	...
ID1	Primary Survey	2012-11-23 15:42:38	....	....
ID1	Airway Clear	2012-11-23 15:42:38	....	....
ID1	...	2012-11-23 15:42:38	....	....
[illegible]	Primary Survey	2012-11-24 09:58:33	....	....
[illegible]	Airway Clear	2012-11-24 09:58:33	....	....
[illegible]	...	2012-11-24 09:58:33	....	....
ID2	Procedure 1	2012-11-24 09:58:33	Completed on 2012-11-24 06:58:34	....

These events are recorded on a form …

… and all have the same timestamp.

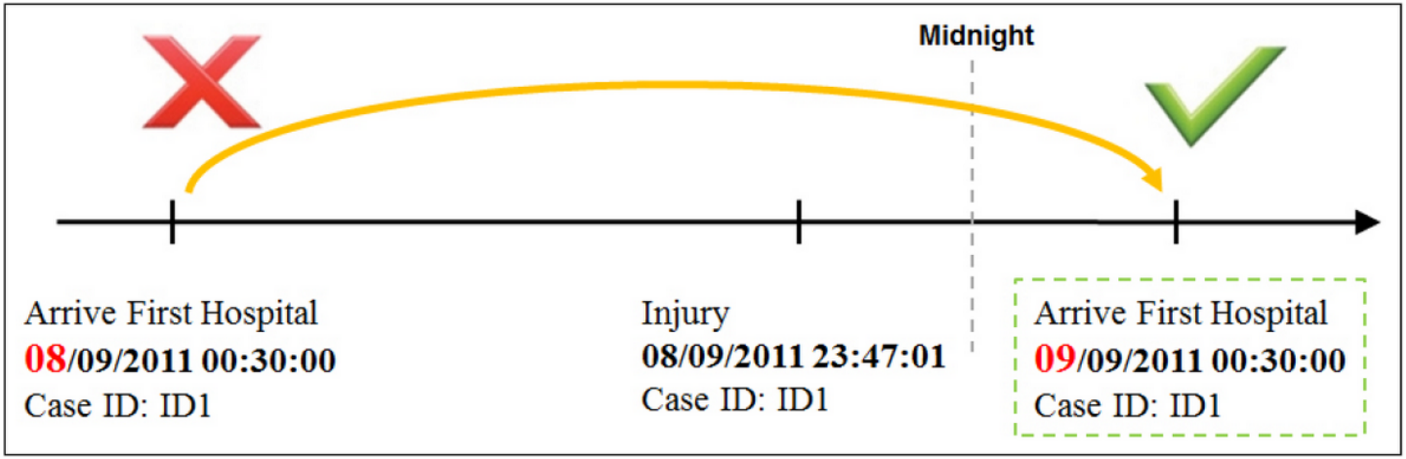
Inadvertent Time Travel

- I16 - Incorrect data: timestamp



Episode ID	Activity	Timestamp	...
ID1	Arrival first hospital	2011-09-08 00:30:00	....
ID1	Injury	2011-09-08 23:47:01	....
...	.....	....	....
ID1	Operation	2011-09-09 16:30:00	....

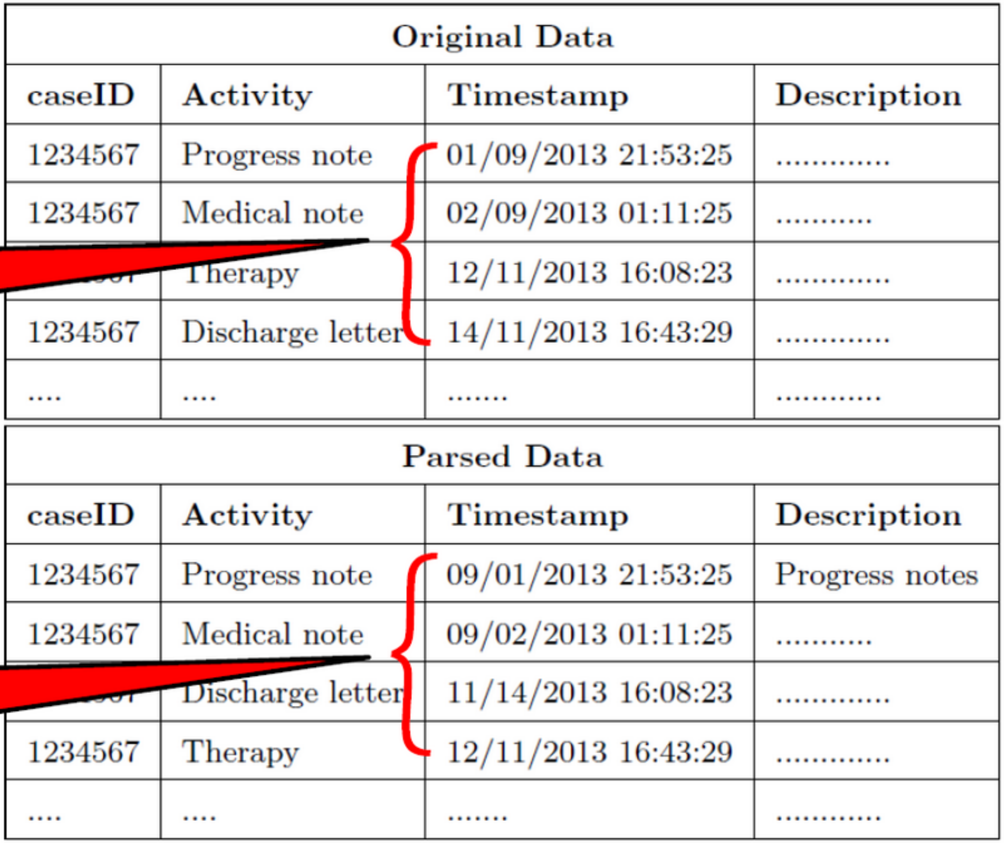
'Midnight' problem. Time portion correct but date part in error.



Unanchored Event

- I23 - Imprecise data: timestamp

Event timestamps in dd/mm/yyyy format are imported …



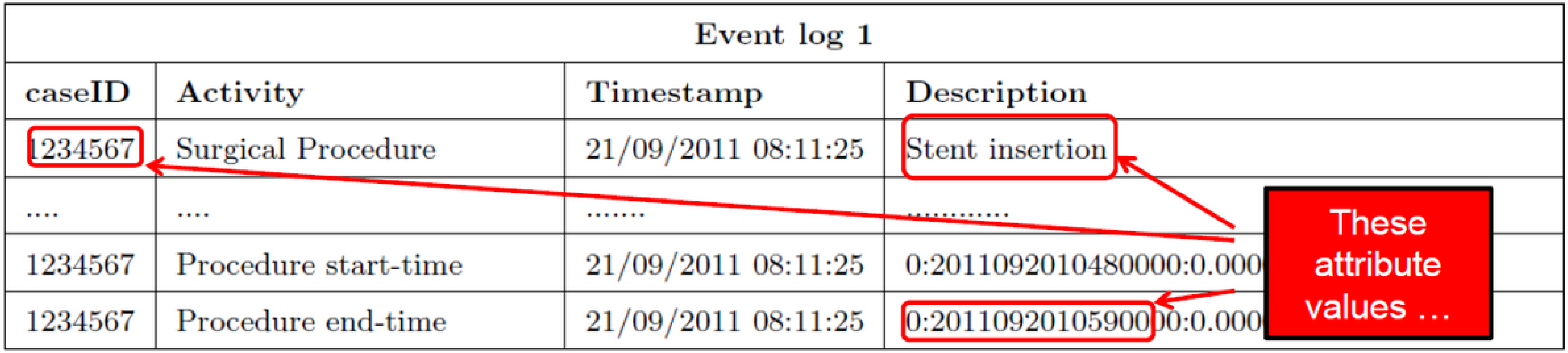
Original Data			
caseID	Activity	Timestamp	Description
1234567	Progress note	01/09/2013 21:53:25	....
1234567	Medical note	02/09/2013 01:11:25	....
1234567	Therapy	12/11/2013 16:08:23	....
1234567	Discharge letter	14/11/2013 16:43:29	....
....	....	....	....

… as mm/dd/yyyy format

Parsed Data			
caseID	Activity	Timestamp	Description
1234567	Progress note	09/01/2013 21:53:25	Progress notes
1234567	Medical note	09/02/2013 01:11:25	....
1234567	Discharge letter	11/14/2013 16:08:23	....
1234567	Therapy	12/11/2013 16:43:29	....
....	....	....	....

Scattered Event

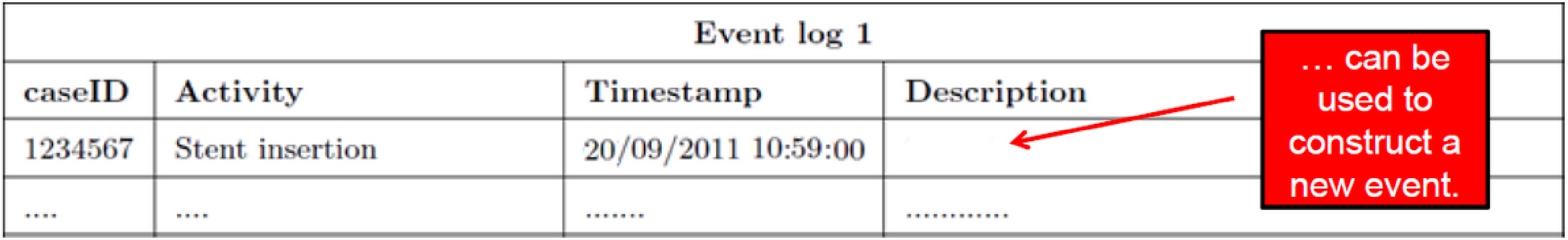
- I16 - Missing data: event



Event log 1

caseID	Activity	Timestamp	Description
1234567	Surgical Procedure	21/09/2011 08:11:25	Stent insertion
....	....	....	....
1234567	Procedure start-time	21/09/2011 08:11:25	0:2011092010480000:0.000
1234567	Procedure end-time	21/09/2011 08:11:25	0:2011092010590000:0.000

These attribute values …



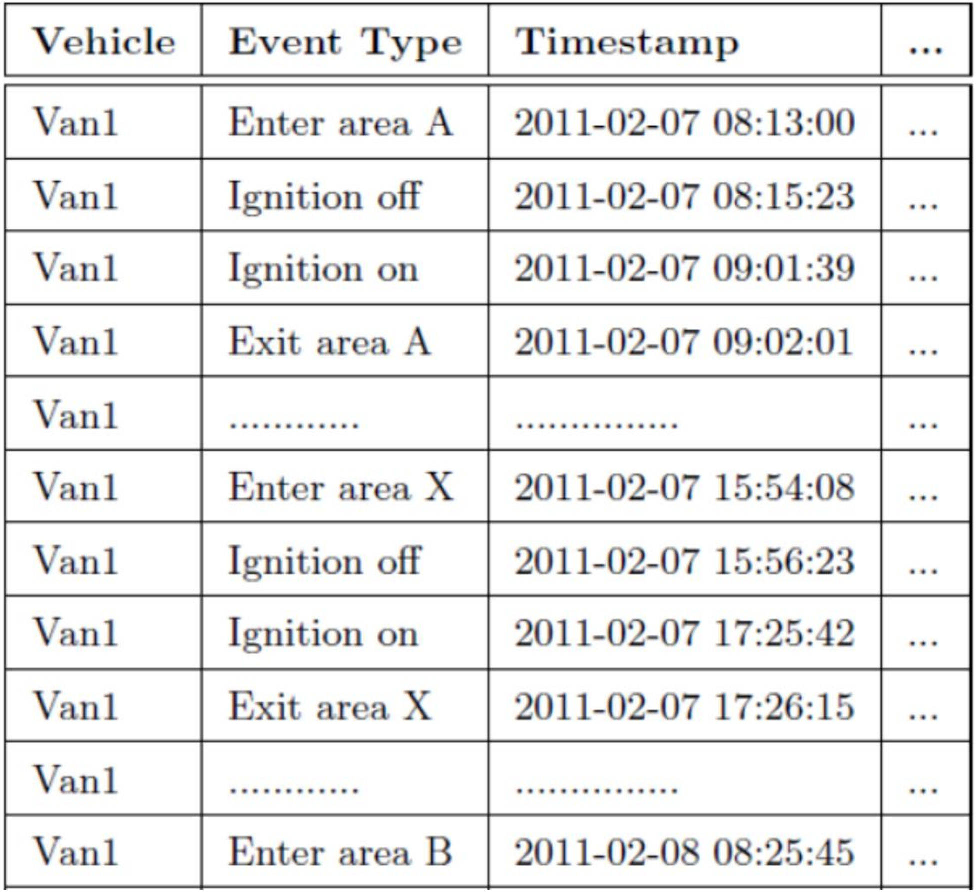
Event log 1

caseID	Activity	Timestamp	Description
1234567	Stent insertion	20/09/2011 10:59:00	
....	....	....	....

… can be used to construct a new event.

Elusive Case

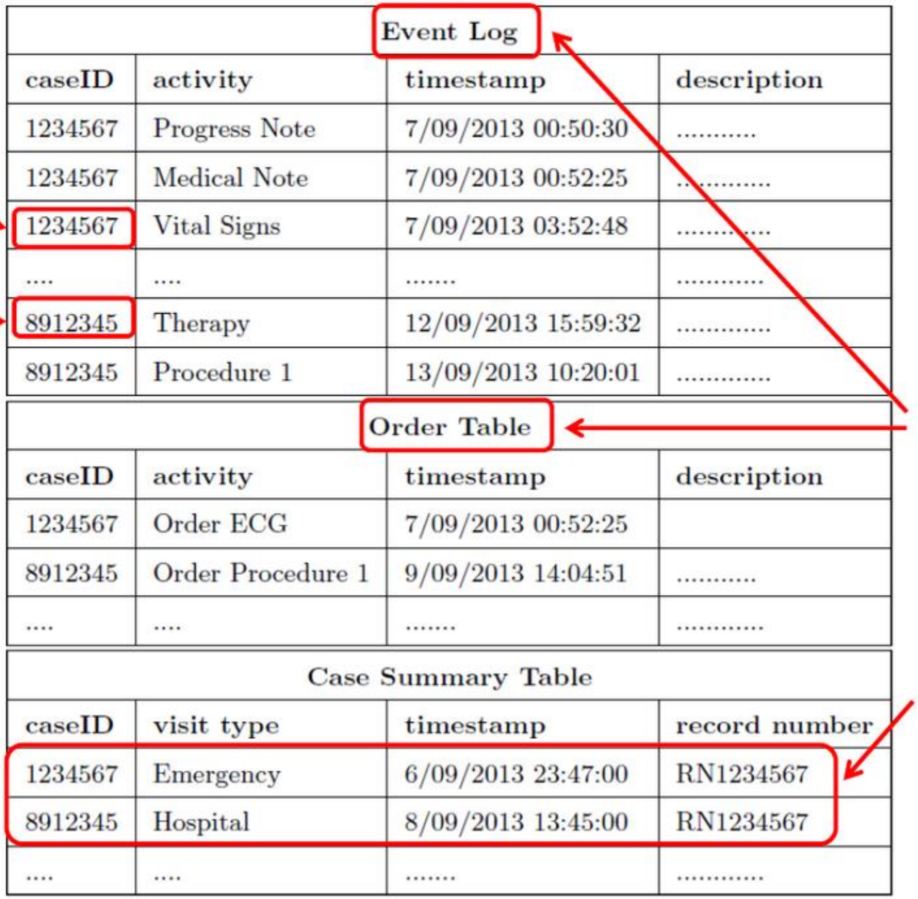
- I16 - Missing data: relationships



Vehicle	Event Type	Timestamp	...
Van1	Enter area A	2011-02-07 08:13:00	...
Van1	Ignition off	2011-02-07 08:15:23	...
Van1	Ignition on	2011-02-07 09:01:39	...
Van1	Exit area A	2011-02-07 09:02:01	...
Van1	....	....	...
Van1	Enter area X	2011-02-07 15:54:08	...
Van1	Ignition off	2011-02-07 15:56:23	...
Van1	Ignition on	2011-02-07 17:25:42	...
Van1	Exit area X	2011-02-07 17:26:15	...
Van1	....	....	...
Van1	Enter area B	2011-02-08 08:25:45	...

Scattered Case

- I12 - Incorrect data: relationship



Event Log			
caseID	activity	timestamp	description
1234567	Progress Note	7/09/2013 00:50:30	....
1234567	Medical Note	7/09/2013 00:52:25	....
1234567	Vital Signs	7/09/2013 03:52:48	....
....	....	....	....
8912345	Therapy	12/09/2013 15:59:32	....
8912345	Procedure 1	13/09/2013 10:20:01	....

Order Table			
caseID	activity	timestamp	description
1234567	Order ECG	7/09/2013 00:52:25	
8912345	Order Procedure 1	9/09/2013 14:04:51	....
....	....	....	....

Case Summary Table			
caseID	visit type	timestamp	record number
1234567	Emergency	6/09/2013 23:47:00	RN1234567
8912345	Hospital	8/09/2013 13:45:00	RN1234567
....	....	....	....

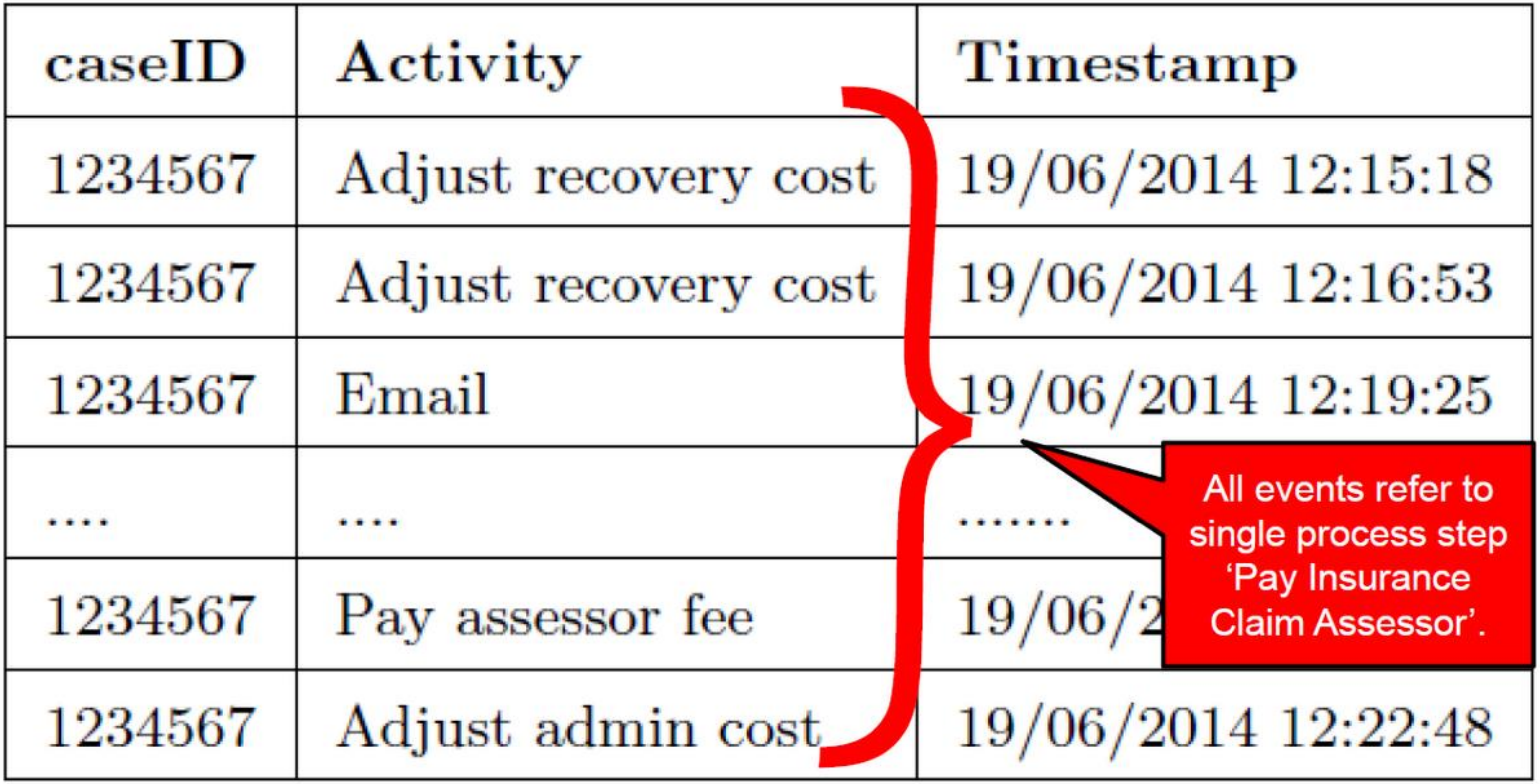
Apparently 2 distinct 'caseID's' …

… scattered across multiple sources …

… that can be linked by a common 'record number'.

Collateral Events

- I27 - Irrelevant data: event



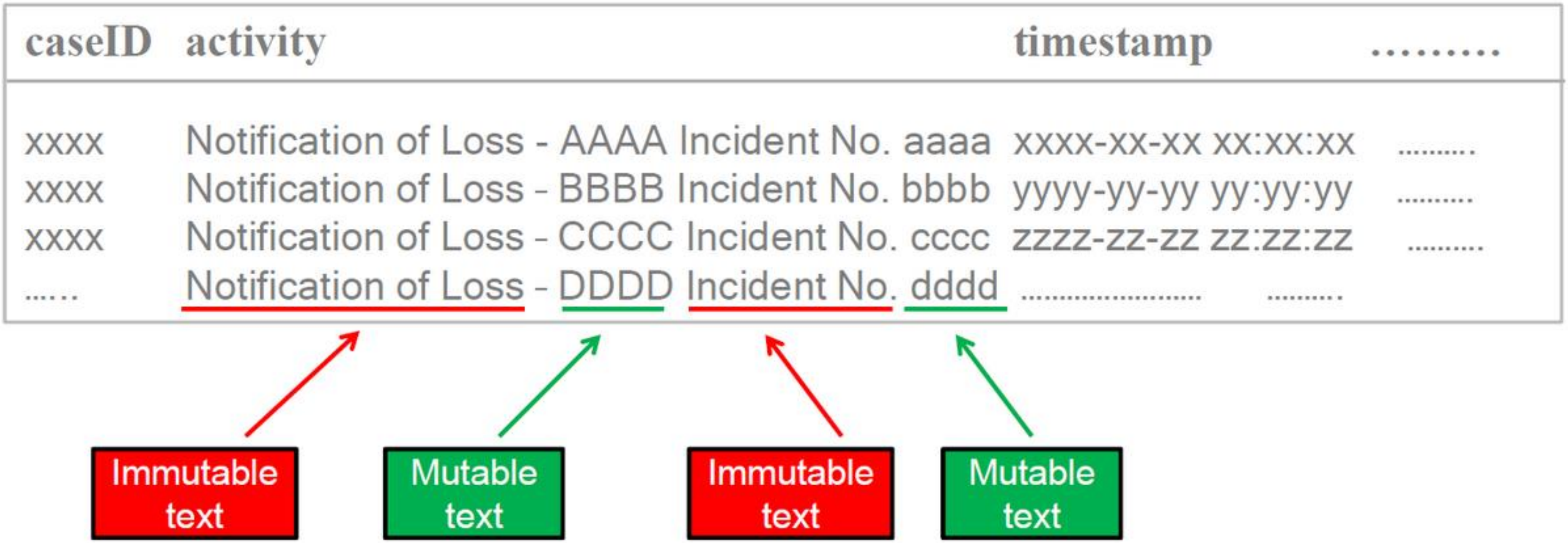
caseID	Activity	Timestamp
1234567	Adjust recovery cost	19/06/2014 12:15:18
1234567	Adjust recovery cost	19/06/2014 12:16:53
1234567	Email	19/06/2014 12:19:25
....	....	....
1234567	Pay assessor fee	19/06/2
1234567	Adjust admin cost	19/06/2014 12:22:48

All events refer to single process step 'Pay Insurance Claim Assessor'.

Polluted Label

Christian-Albrechts-Universität zu Kiel

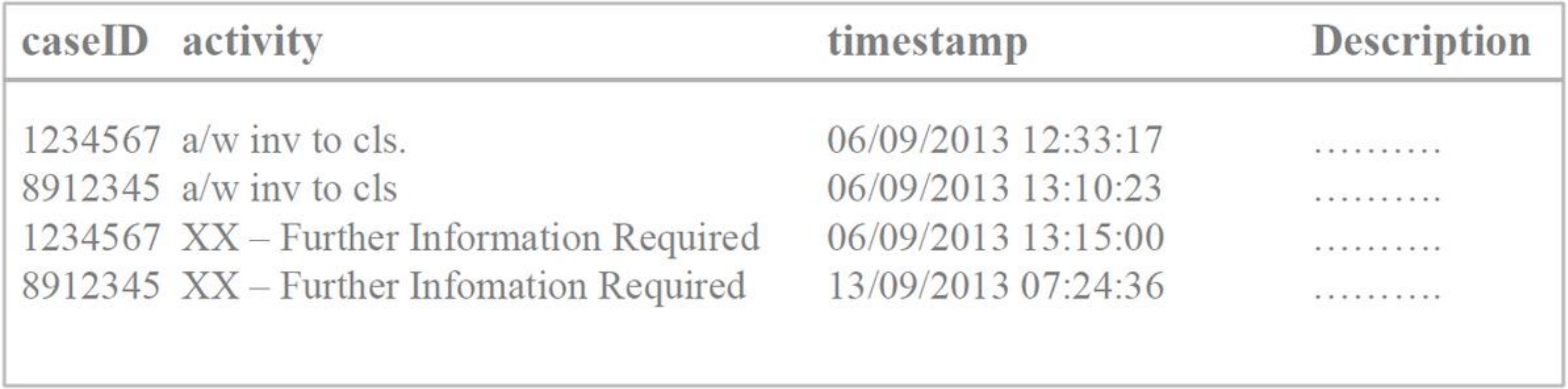
- I15 - Incorrect data: activity name
- I17 - Incorrect data: resource



caseID	activity	timestamp	....
xxxx	Notification of Loss - AAAA Incident No. aaaa	xxxx-xx-xx xx:xx:xx	....
xxxx	Notification of Loss - BBBB Incident No. bbbb	yyyy-yy-yy yy:yy:yy	....
xxxx	Notification of Loss - CCCC Incident No. cccc	zzzz-zz-zz zz:zz:zz	....
......	Notification of Loss - DDDD Incident No. dddd	....	....

Distorted Label

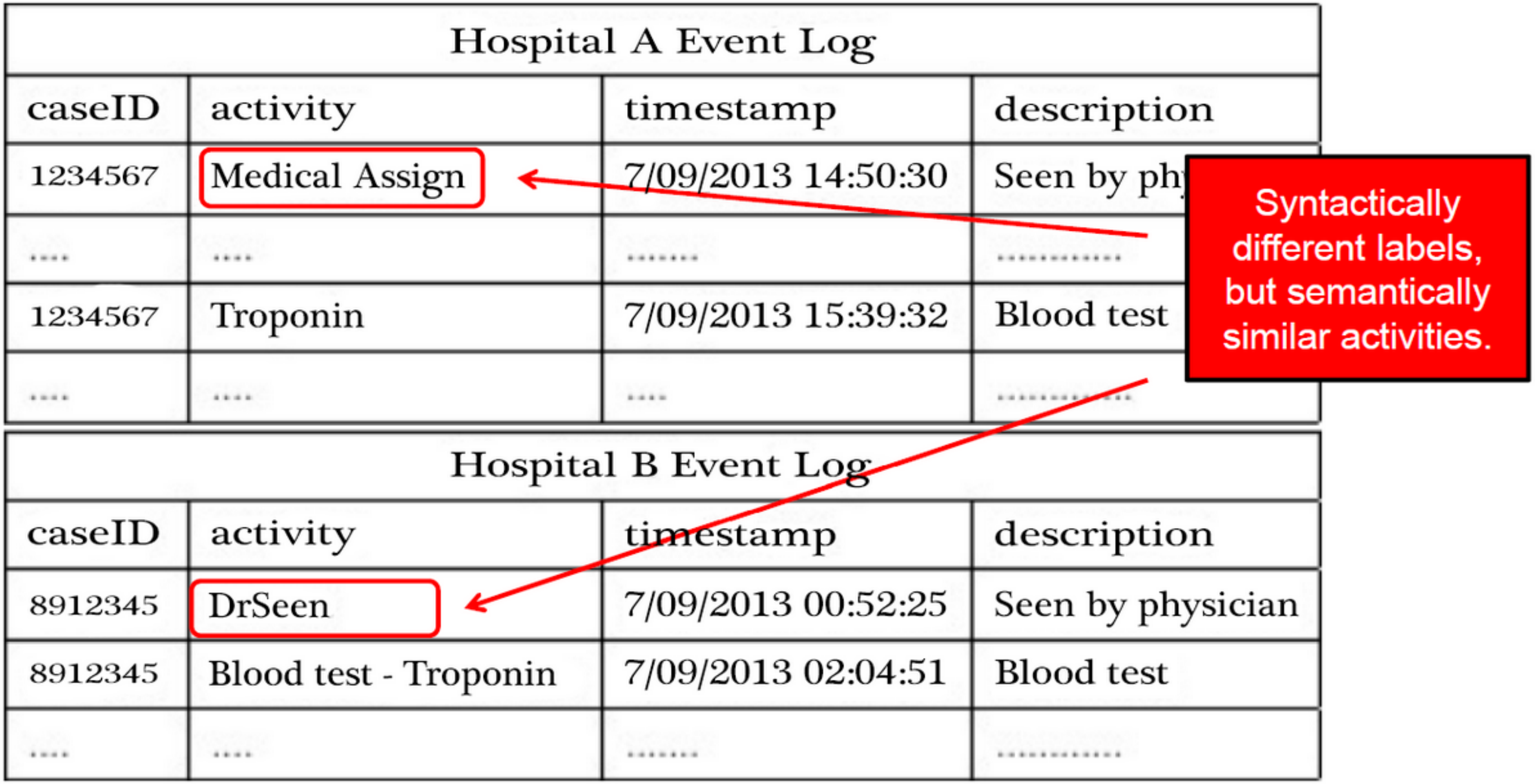
- I15 - Incorrect data: activity name



caseID	activity	timestamp	Description
1234567	a/w inv to cls.	06/09/2013 12:33:17	....
8912345	a/w inv to cls	06/09/2013 13:10:23	....
1234567	XX – Further Information Required	06/09/2013 13:15:00	....
8912345	XX – Further Infomation Required	13/09/2013 07:24:36	....

Synonymous Labels

- I22 - Imprecise data: event attributes



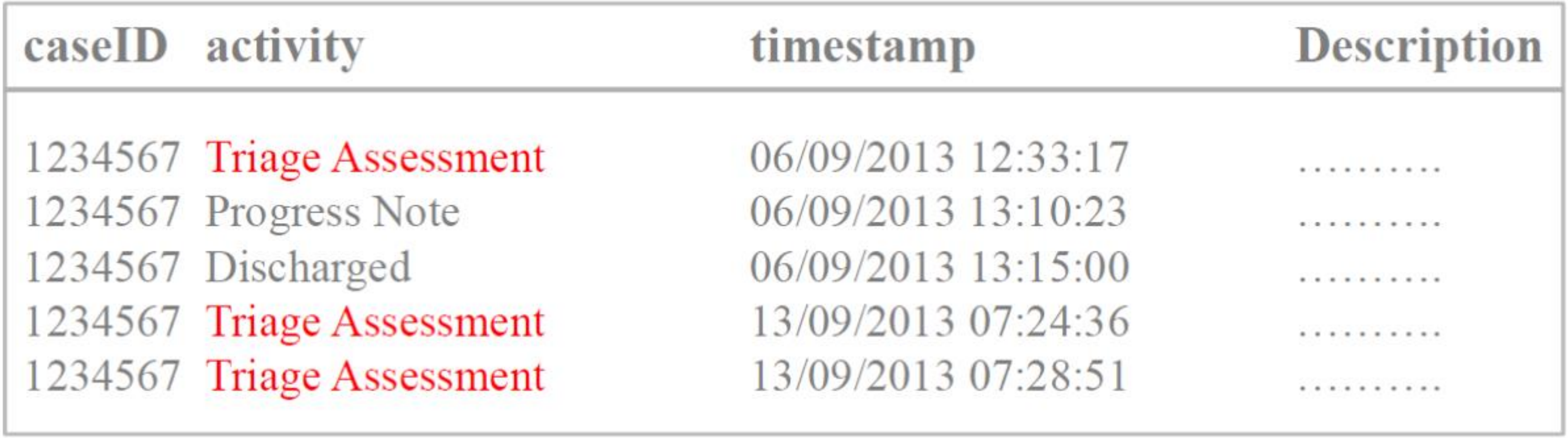
Hospital A Event Log			
caseID	activity	timestamp	description
1234567	Medical Assign	7/09/2013 14:50:30	Seen by ph
....	....	....	....
1234567	Troponin	7/09/2013 15:39:32	Blood test
....	....	....	....

Hospital B Event Log			
caseID	activity	timestamp	description
8912345	DrSeen	7/09/2013 00:52:25	Seen by physician
8912345	Blood test - Troponin	7/09/2013 02:04:51	Blood test
....	....	....	....

Syntactically different labels, but semantically similar activities.

Homonymous Label

- I2 - Imprecise data: activity name



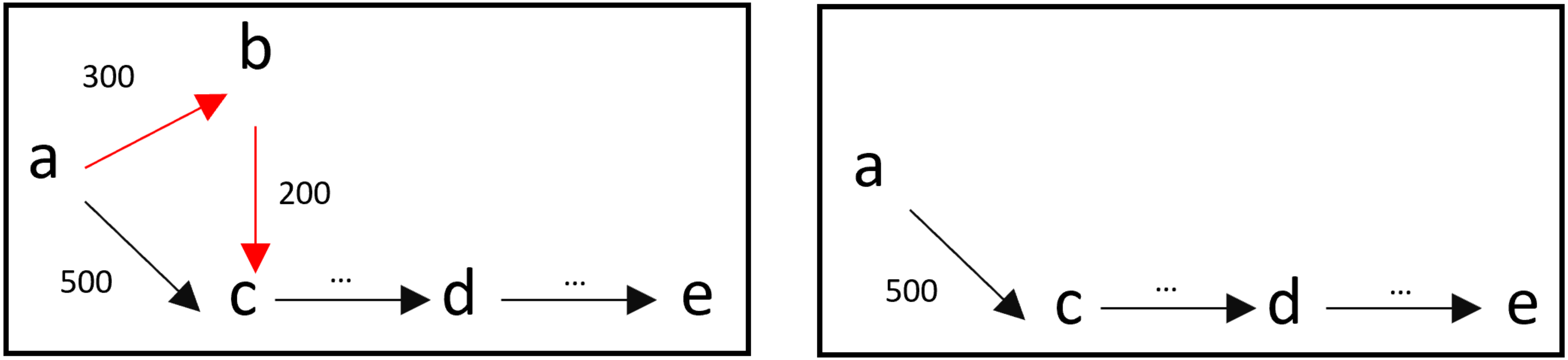
caseID	activity	timestamp	Description
1234567	Triage Assessment	06/09/2013 12:33:17	....
1234567	Progress Note	06/09/2013 13:10:23	....
1234567	Discharged	06/09/2013 13:15:00	....
1234567	Triage Assessment	13/09/2013 07:24:36	....
1234567	Triage Assessment	13/09/2013 07:28:51	....

Filtering Outliers within the Process Mining Algorithm

- infrequent behavior are paths that are taken infrequently, or traces that only differ by occurrence of infrequent activities
- L = [⟨a, c, d, e, b⟩ , ⟨a, b, a, e, d, c⟩ , ⟨a, e, c, b, d⟩ , ⟨a, d, b, c, e⟩
- second trace is the only trace containing two as where the second a is infrequent
- **in a directly-follows graph:** outgoing edges of a node having a frequency of less than k-times of the most frequent outgoing edge of the identical node

Filtering Outliers within the Process Mining Algorithm

- the DFG is filtered until it only contains most frequent edges or the mainstream behavior
- DFG may be misleading



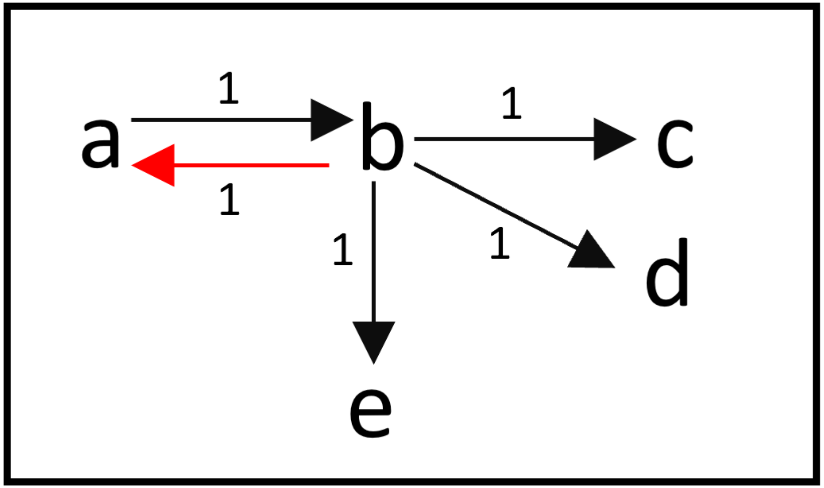
Left: DFG showing all activities and its frequencies.
Right: DFG filtered by one infrequent path resulting in incorrect DFG

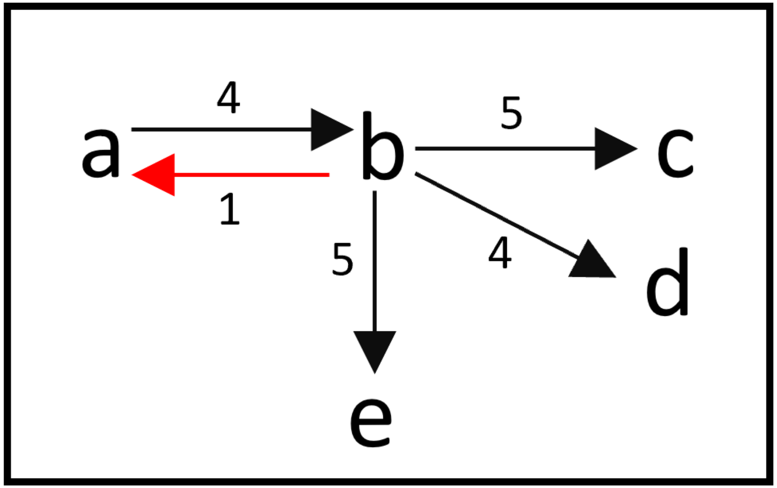
Eventually-follows Graph

- transitive closure of the directly-follows relation
- an edge (a,b) is present if and only if a is followed by b somewhere in the log

Filtering Outliers within the Process Mining Algorithm

- the infrequent edge (b,a) can be filtered out

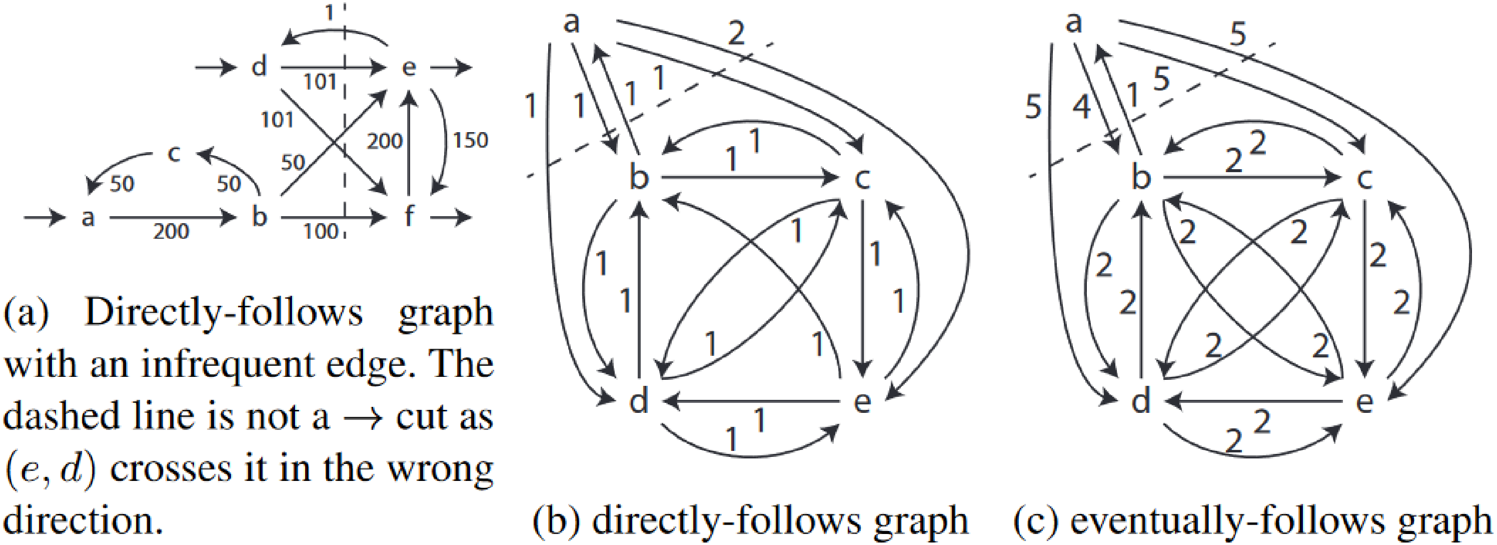




Left: DFG showing all activities and its frequencies for L2.
Right: Even-tually Follows Graph for L2

Inductive Miner Infrequent

- L2 = [⟨a,c,d,e,b⟩ , ⟨a,b,a,e,d,c⟩ , ⟨a,e,c,b,d⟩ , ⟨a,d,b,c,e⟩]



(a) Directly-follows graph with an infrequent edge. The dashed line is not a $\rightarrow$ cut as (e, d) crosses it in the wrong direction.

(b) directly-follows graph

(c) eventually-follows graph

source:https://www.win.tue.nl/~dfahland/publications/LeemansFA_2013_bpi.pdf