Christian-Albrechts-Universität zu Kiel

Advanced Process Mining

Prof. Dr. Agnes Koschmider

Lecture 10: Summary

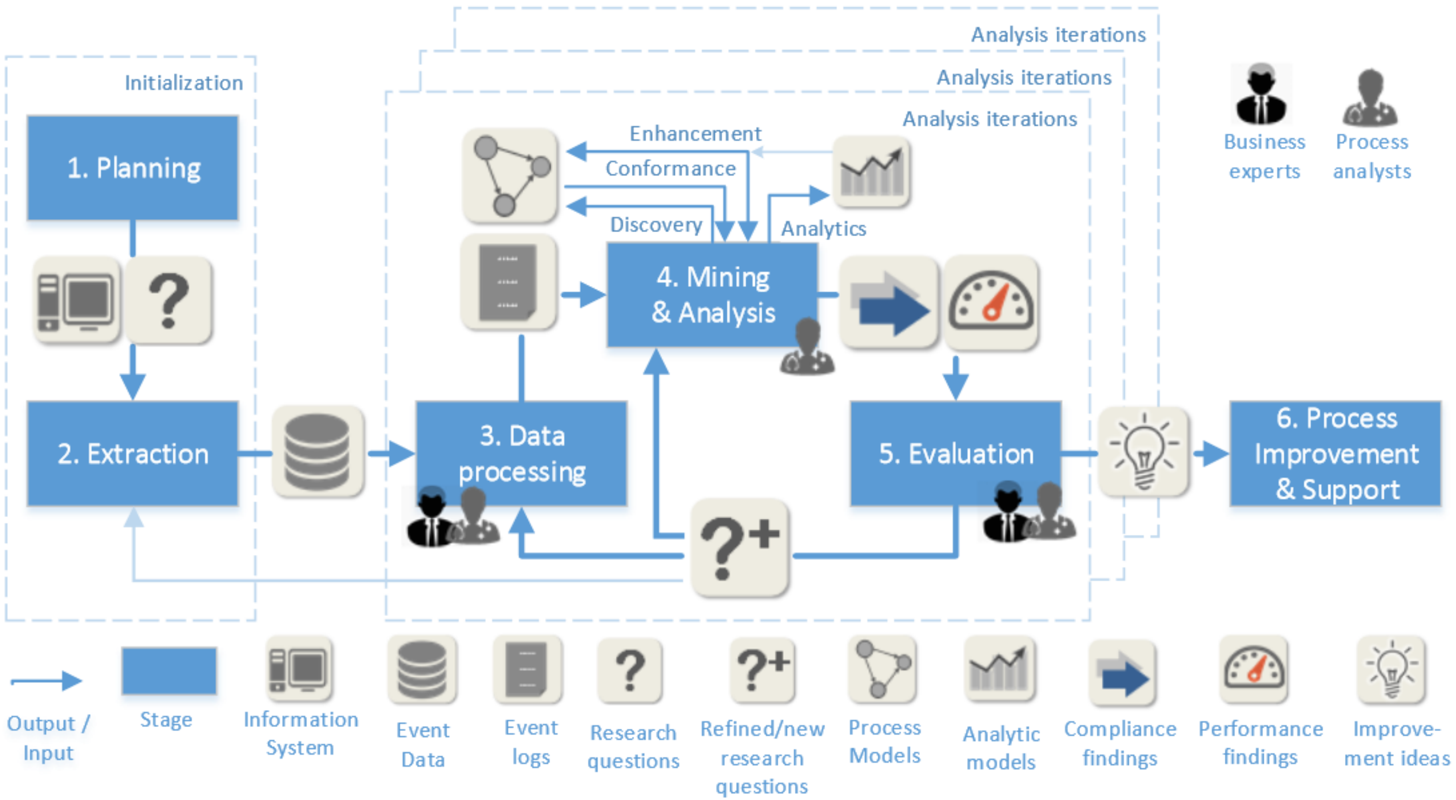
Process Analytics

Christian-Albrechts-Universität zu Kiel

Questionnaire

Process Mining Project Methodology

Christian-Albrechts-Universität zu Kiel



M.L. Eck, van, X. Lu, S.J.J. Leemans, W.M.P. Aalst, van der:
PM2 : a Process Mining Project Methodology, CAiSE 2015, Springer

Process Mining
Directly-Follows Graph

Christian-Albrechts-Universität zu Kiel

- Use event data to show what people, machines, and organizations are really doing
- Provides novel insights that can be used identify and address performance and compliance problems
- Commercial tools resort to producing Directly-Follows Graphs (DFGs) based on event data

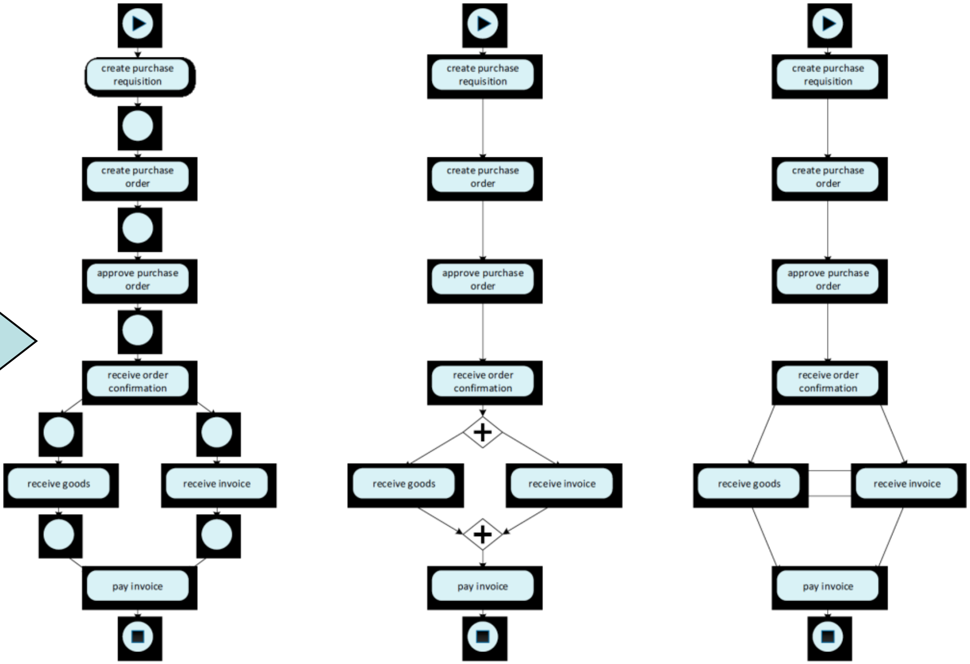
Limitations of the Directly-Follows Graph

Christian-Albrechts-Universität zu Kiel

» to tackle complexity DFGs are seamlessly simplified by removing nodes and edges based on frequency thresholds

» DFGs may be misleading

case id (here an order)	activity	timestamp	resource	costs	customer
...	...	...	...	...	..
2019-88201	create purchase requisition	25-07-2019:09.15	John	€20.20	9950
2019-88201	create purchase order	25-07-2019:09.35	Mary	€48.30	9950
2019-88201	approve purchase order	25-07-2019:09.55	Sue	€30.70	9950
2019-88202	create purchase requisition	25-07-2019:10.15	John	€28.20	9955
2019-88202	create purchase order	25-07-2019:10.25	Mary	€29.30	9955
2019-88202	approve purchase order	25-07-2019:10.40	Sue	€37.60	9955
2019-88201	receive order confirmation	25-07-2019:11.50	Mary	€42.10	9950
2019-88201	receive goods	27-07-2019:09.35	Peter	€50.20	9950
2019-88202	receive order confirmation	27-07-2019:09.45	Mary	€42.30	9955
2019-88202	receive invoice	28-07-2019:10.10	Sue	€44.90	9955
2019-88201	receive invoice	28-07-2019:10.20	Sue	€30.80	9950
2019-88201	pay invoice	29-07-2019:11.05	Sue	€30.70	9950
2019-88202	receive goods	29-07-2019:11.35	Peter	€51.30	9955
2019-88202	pay invoice	29-07-2019:12.15	Sue	€29.20	9955
...	..	...	...	...	...



(a) Petri net (b) BPMN model (c) DFG

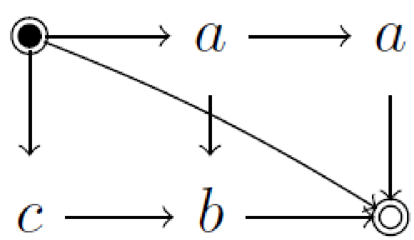
Soundness & DFG

» In the construction of DFG edges are filtered out based on a user chosen parameter to reduce the complexity of the model

→ this may lead to soundness issues

Definition (DFM soundness). *Let* (N, E) *be a DFM. Then, the DFM is* sound *if every node* $\in N$ *is on a path from start to end:*

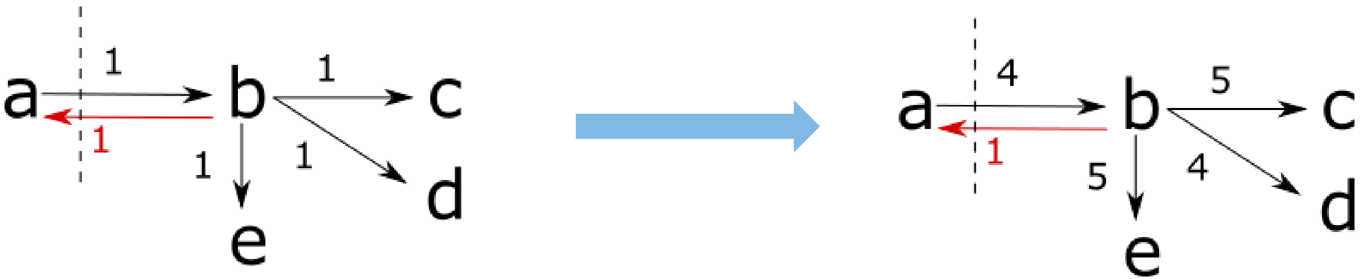
$$\forall_{x \in N} \exists_{a_1 \ldots a_n \in N} a_1 = start \wedge a_n = end \wedge \exists a_j = x \\ \wedge \forall_{1 \leq i < n} (a_i, a_{i+1}) \in E$$



- The DFM without any edges is not sound, as the start and end nodes are not on a path as requested by the definition.

Eventually-Follows-Graph

- Despite Heuristic-style Filtering infrequent edges might remain
- Use of the Eventually-Follows-Graph, which is the 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

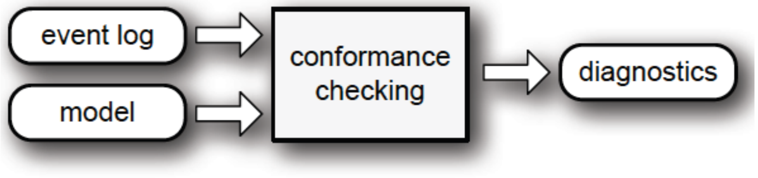


Conformance Checking

C | A | U
Christian-Albrechts-Universität zu Kiel

Conformance checking

- Detect discrepancies between process model and observed information
- Analyse deviations

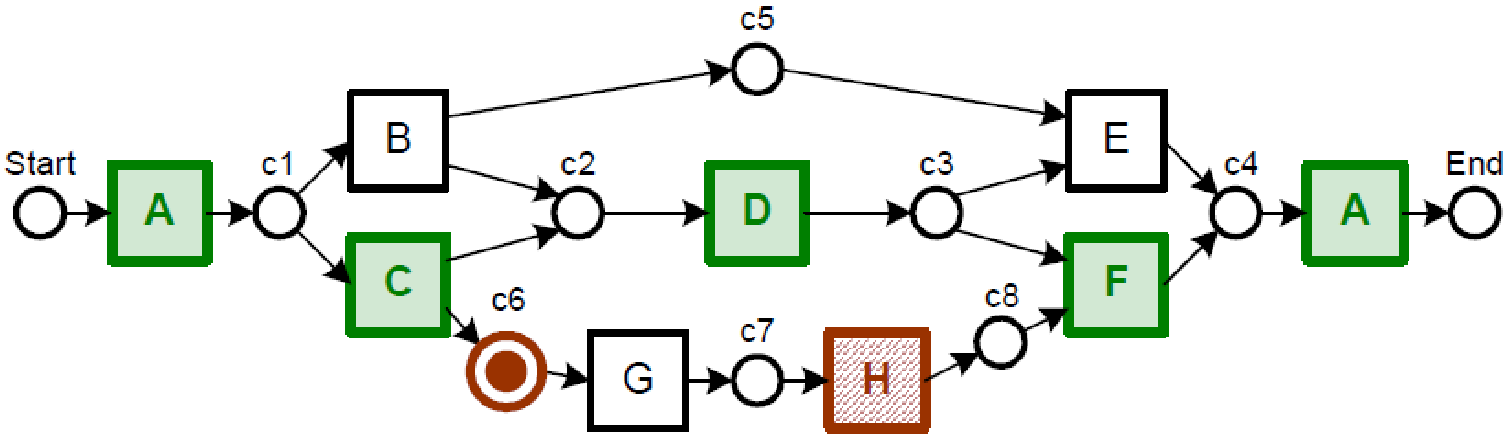


Perspectives

- Local feedback on deviations at the level of individual traces in the log
- Local feedback on deviations at the level of individual process model parts, e.g., activities
- Global feedback on overall conformance

Local Feedback in Trace

Christian-Albrechts-Universität zu Kiel



Per trace:

Transitions that are in line with model and those that are not

m = 1
r = 1
c = 8
p = 8

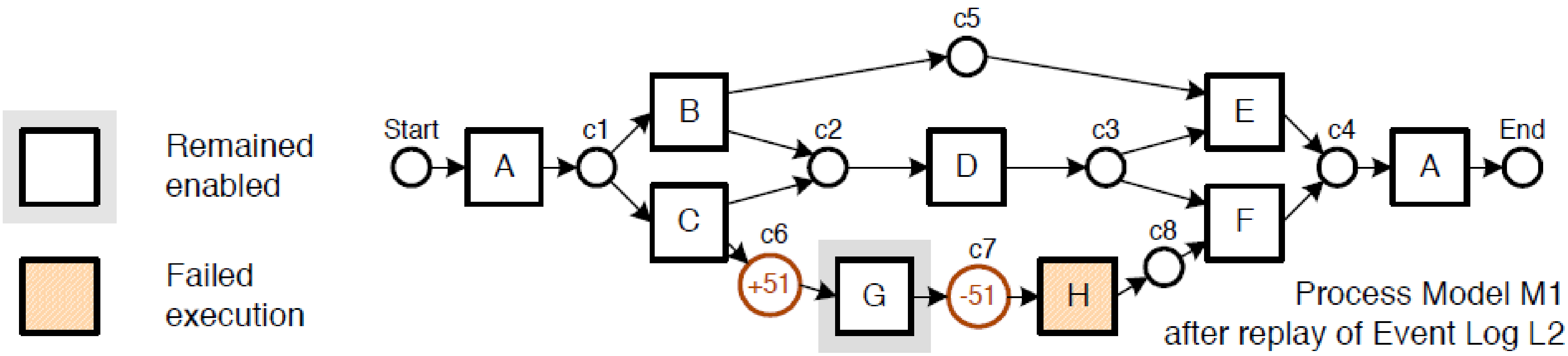
No. of Instances	Log Traces
1207	ABDEA
145	ACDGHFA
56	ACGDHFA
23	**ACHDFA**
28	ACDHFA

Local Feedback in Model

C | A | U
Christian-Albrechts-Universität zu Kiel

On the model level:

- Aggregated information about missing and remaining tokens
- Identification of “hotspots” of non-conformance
- Highlights major issues when replaying log and separates some from noise



Alignment-based Conformance Checking

Christian-Albrechts-Universität zu Kiel

Assessing the conformance of an event log with a model based on an alignment of activities and events

- Consider the set of activities of the model as a set of symbols
- Then, each execution sequence of the model is a sequence of symbols
- Each trace of the event log is also a sequence of symbols

An alignment between two sequences is established by

- Linking pairs of symbols in each sequence
- Such that the order between aligned symbols is preserved

The notion of an alignment allows for quantification of conformance and insights on non-conformance

Recall: Alignments

Trace: ADAB

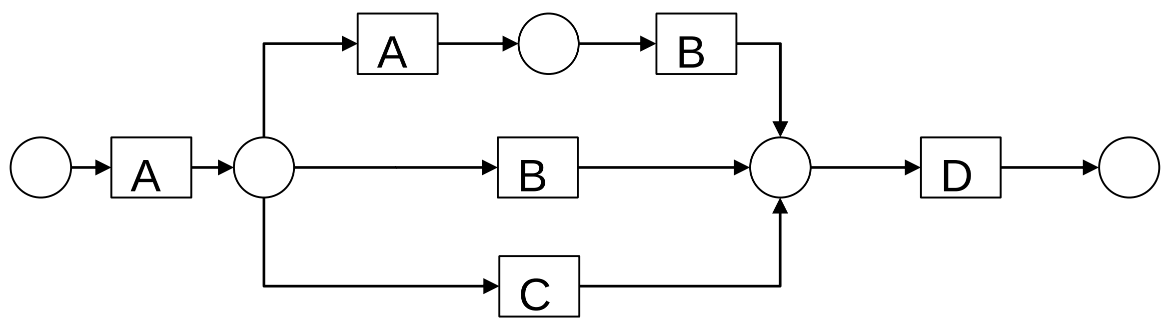
move in model: an event that should have been observed according to modeled behavior but missed in the trace

Alignment:

A	D	A	B	
A		A	B	D

synchronous move in the ideal alignment (“good move”)

move in log:
observed event not allowed by the modeled



Examples Again

Christian-Albrechts-Universität zu Kiel

$$\gamma_1 = \begin{array}{|c|c|c|c|c|} a & c & d & e & h \\ \hline a & c & d & e & h \end{array}$$

Cost: 0

$$\gamma_2 = \begin{array}{|c|c|c|c|c|c|c|} a & b & \perp & d & e & g & \perp \\ \hline a & \perp & c & d & e & \perp & h \end{array}$$

Cost: 4

$$\gamma_3 = \begin{array}{|c|c|c|c|c|c|c|c|} a & b & \perp & d & e & \perp & \perp & g \\ \hline \perp & a & c & d & \perp & e & h & \perp \end{array}$$

Cost: ∞

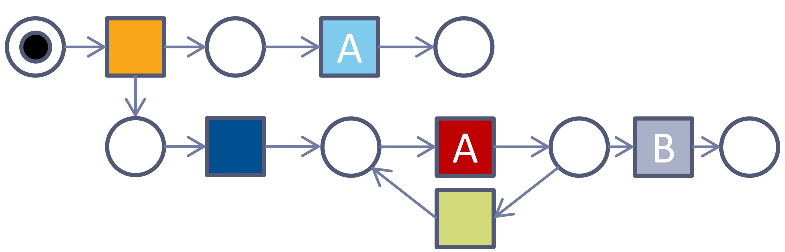
The Problem of Finding Optimal Alignments

The search space is a “product” of the statespace of the model and the trace

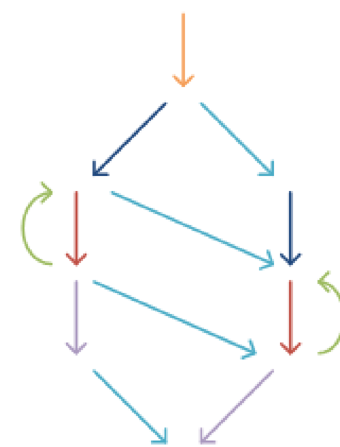
Each node is a combination of a state in the model and the executed events in the trace

Each arc is a move in model, move in log or move in both

Example:

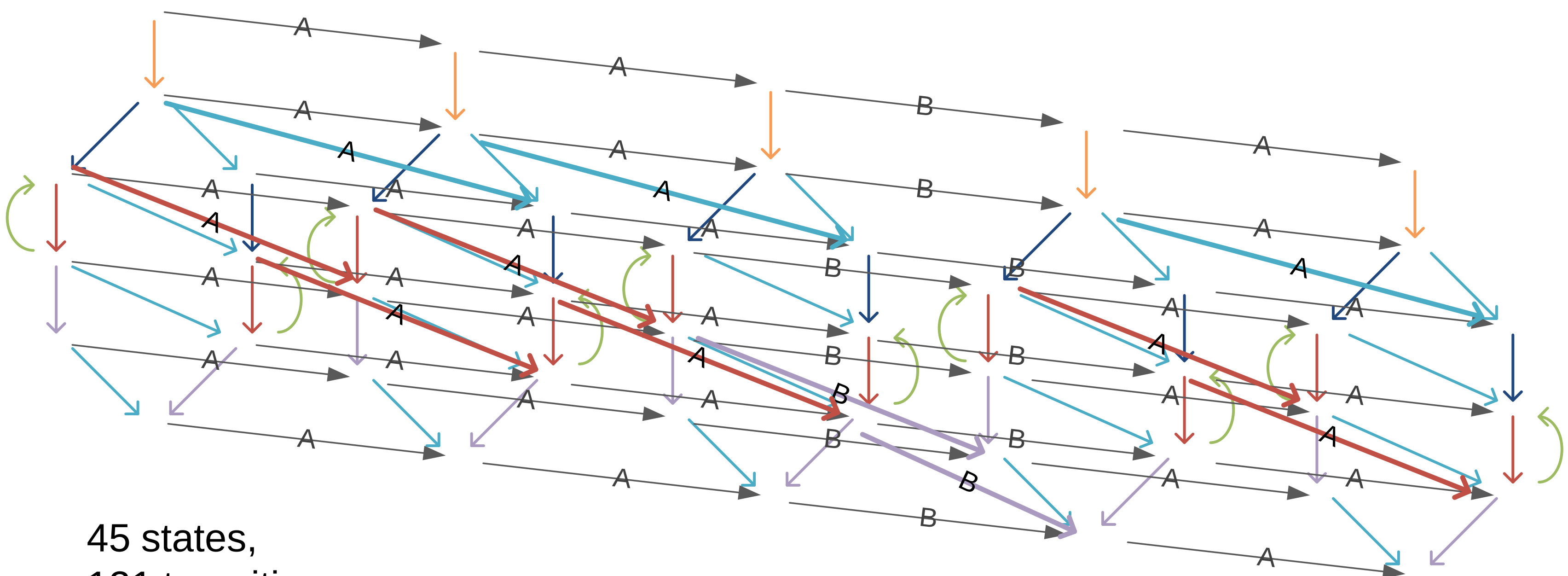


9 states,
13 transitions



Trace: < A, A, B, A >

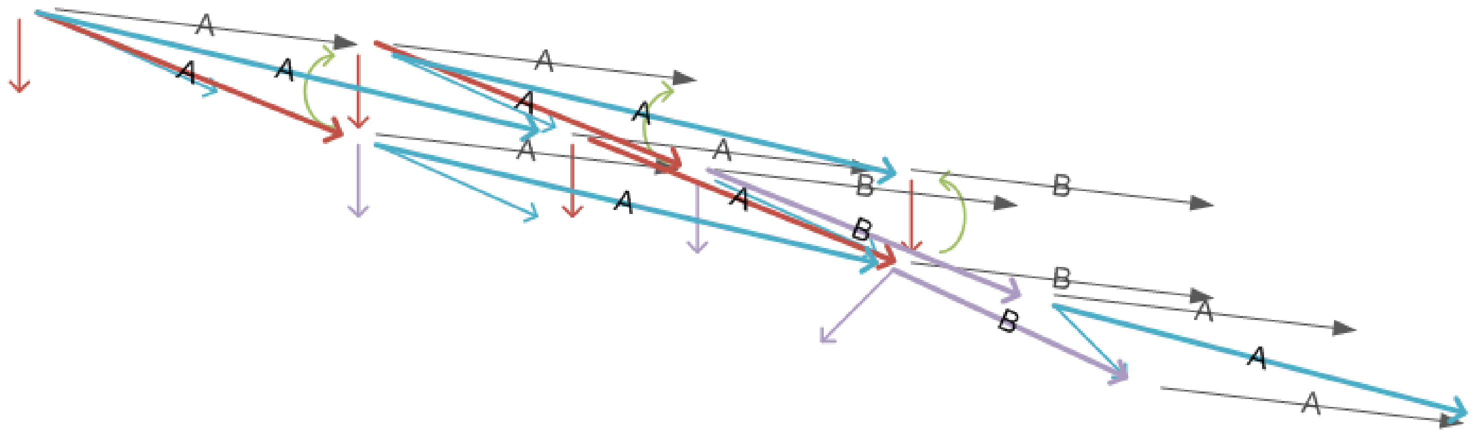
Alignment search space



45 states,
121 transitions

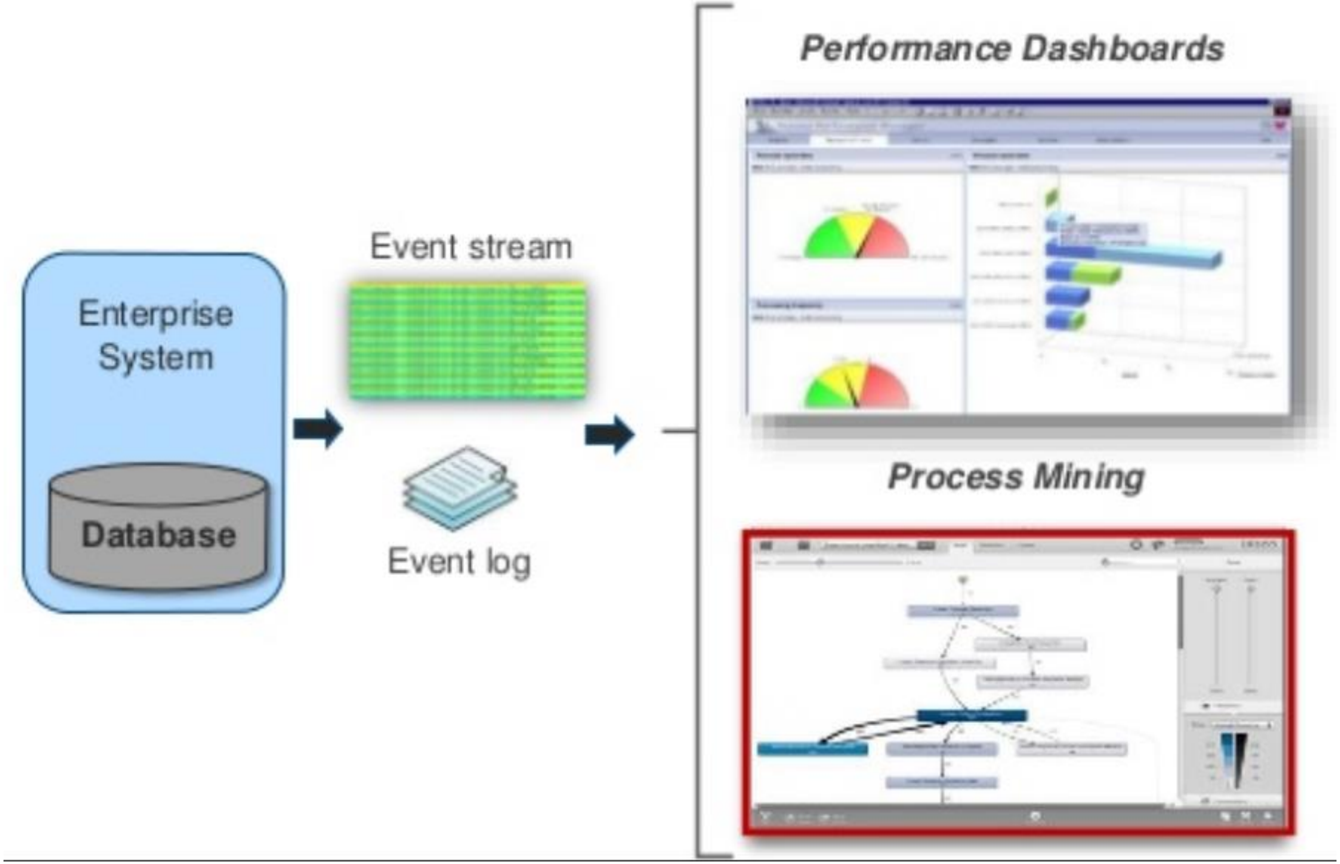
Find the shortest path from the top-left to the bottom right

A* by Example

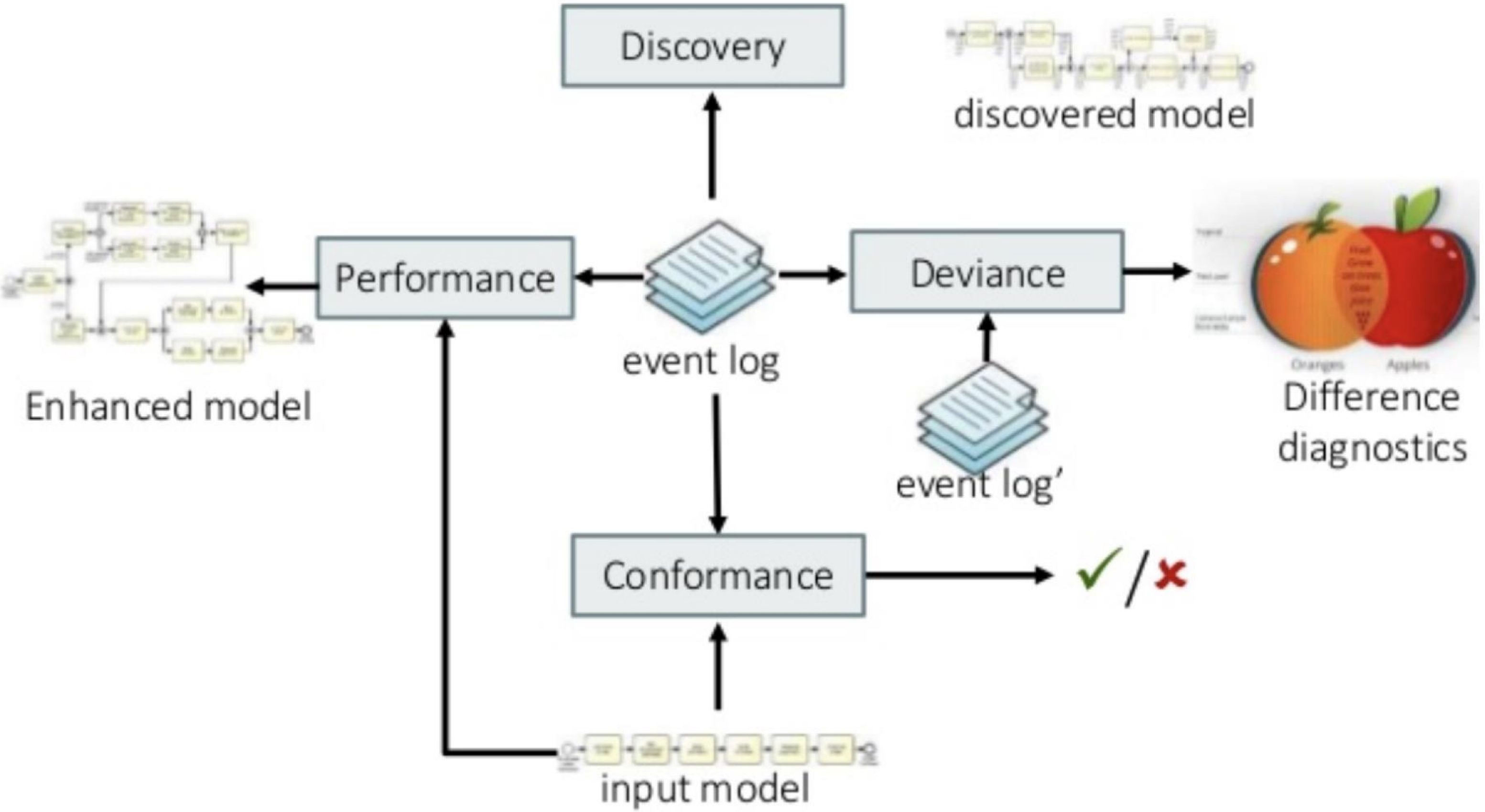


Simple estimator for our setting:
Length of remaining trace * minimal cost for each transition

Business Process Monitoring



Offline Process Mining



Deviance Mining via Sequence Classification

Christian-Albrechts-Universität zu Kiel

- Apply discriminative sequence mining methods to extract features characteristic of one class
- Build classification models (e.g. decision trees)
- Extract difference diagnostics from classification model

Rank	Rule
1	{(*Open*, 2}:*anomalous*
2	{(*Closed*, 2), (*Postponed*, 0), (*Finished*, 0)}:*anomalous*
3	{(*Reopen*, 2)}:*anomalous*
4	{(*Closed*, 1), (*Rejected*, 1), (*Reopen*, 0)}:*anomalous*
5	{(*Reopen*, *Closed*, 1)}:*anomalous*

Sequence encoding(1)

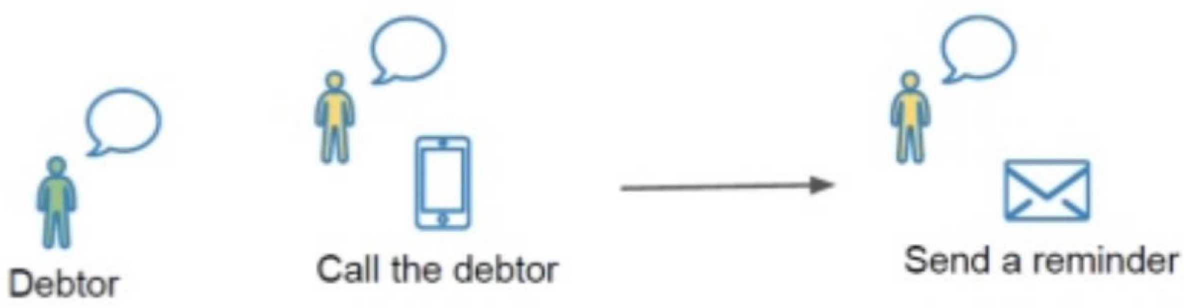
C | A | U
Christian-Albrechts-Universität zu Kiel

Case ID	Timestamp	Activity	Resource	Loan goal	Requested amt	Offered amt
C001	18-10-2016	Check completeness	Sue	Mortgage	100 000	-
C001	19-10-2016	Check credit history	Sue	Mortgage	100 000	-
C001	19-10-2016	Calculate risk score	Bob	Mortgage	100 000	-
C001	20-10-2016	Make offer	Mike	Mortgage	100 000	70 000
C001	25-10-2016	Make offer	Mike	Mortgage	100 000	80 000
C002	20-10-2016	Check completeness	Sue	Car	15 000	-
C002	20-10-2016	Check credit history	Sue	Car	15 000	-
C002	22-10-2016	Calculate risk score	Elsa	Car	15 000	-
C002	24-10-2016	Reject application	Elsa	Car	15 000	-
C003	02-11-2016	Check completeness	Maria	Mortgage	30 000	-
C003	04-11-2016	Ask for additional data	Maria	Mortgage	30 000	-
C003	10-11-2016	Check credit history	Maria	Mortgage	30 000	-
...	...	...	...	...	...	...

Feature vector x Target y

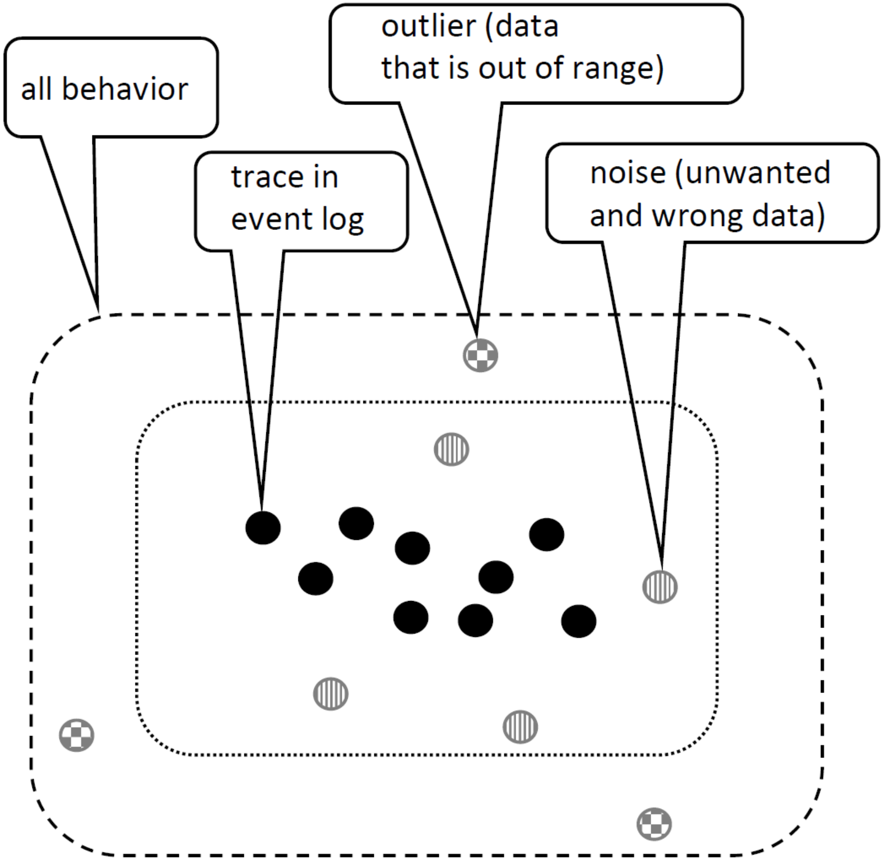
x^1	1	0	0	...	1	0	0	...	14	0	0	0	...	1
x^2	1	0	0	...	0	1	0	...	15	1	0	0	...	1
x^3	0	1	0	...	0	1	0	...	18	0	0	0	...	0
x^4	0	0	1	...	0	1	0	...	10	0	0	0	...	1
x^5	0	0	1	...	0	0	1	...	17	0	1	0	...	0

Predictive Monitoring with Unstructured Data



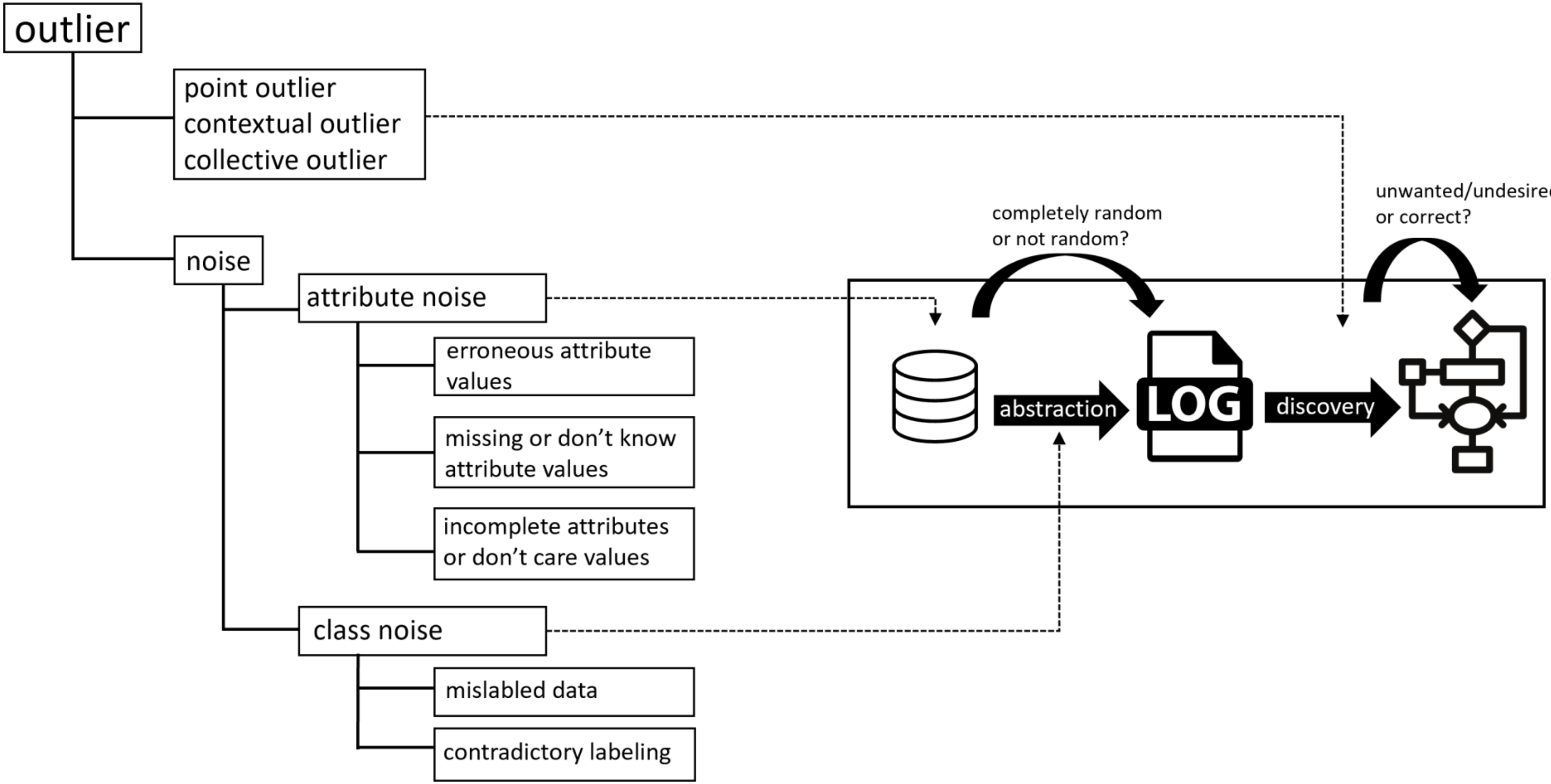
	Event1	Event2	Resource1	Resource2	Debtor	Summary1	Summary2
Trace1	Call the debtor	Send a reminder	Sue	Bob	Mark	?	?

Outlier vs. Noise



Classification of outliers and noise

Christian-Albrechts-Universität zu Kiel



Quality issues in event log attributes

Christian-Albrechts-Universität zu Kiel

Event Log Imperfection Patterns

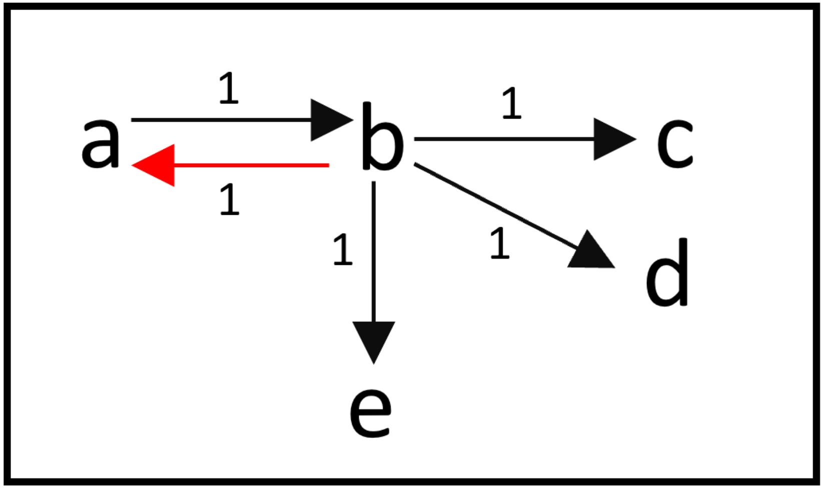
		Event Log Entities								
		case	event	relationship	case attrs.	position	activity name	timestamp	resource	event attrs.
Event Log Quality Issues	Missing data	I1	I2	I3	I4	I5	I6	I7	I8	I9
	Incorrect data	I10	I11	I12	I13	I14	I15	I16	I17	I18
	Imprecise data			I19	I20	I21	I22	I23	I24	I25
	Irrelevant data	I26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

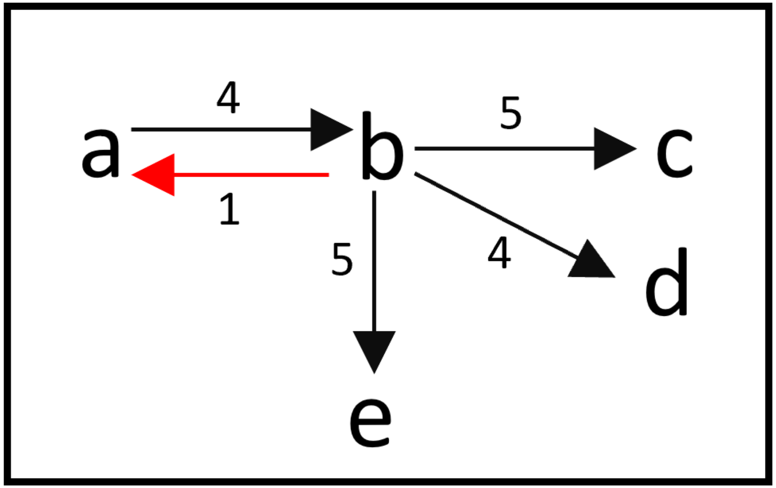
source: http://www.workflowpatterns.com/patterns/logimperfection/

Filtering Outliers within the Process Mining Algorithm

Christian-Albrechts-Universität zu Kiel

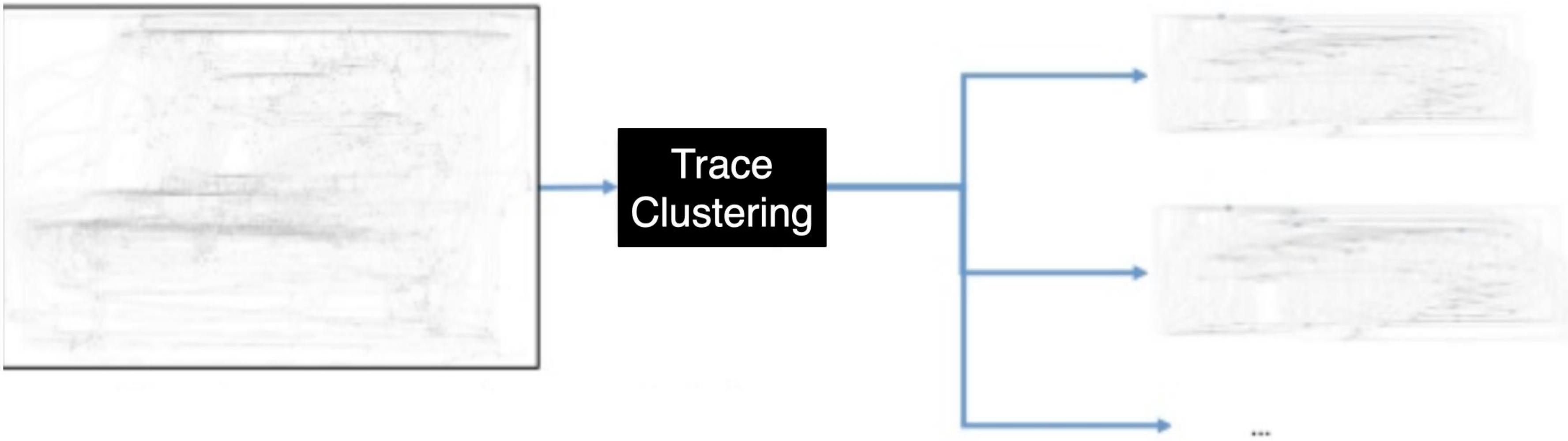
- the infrequent edge (b,a) can be filtered out





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

Trace Clustering



- Handle very complex event data
- Handle unknown number of clusters
- Incorporate and leverage domain knowledge

Clustering Methods – Distance Measures

C | A | U
Christian-Albrechts-Universität zu Kiel

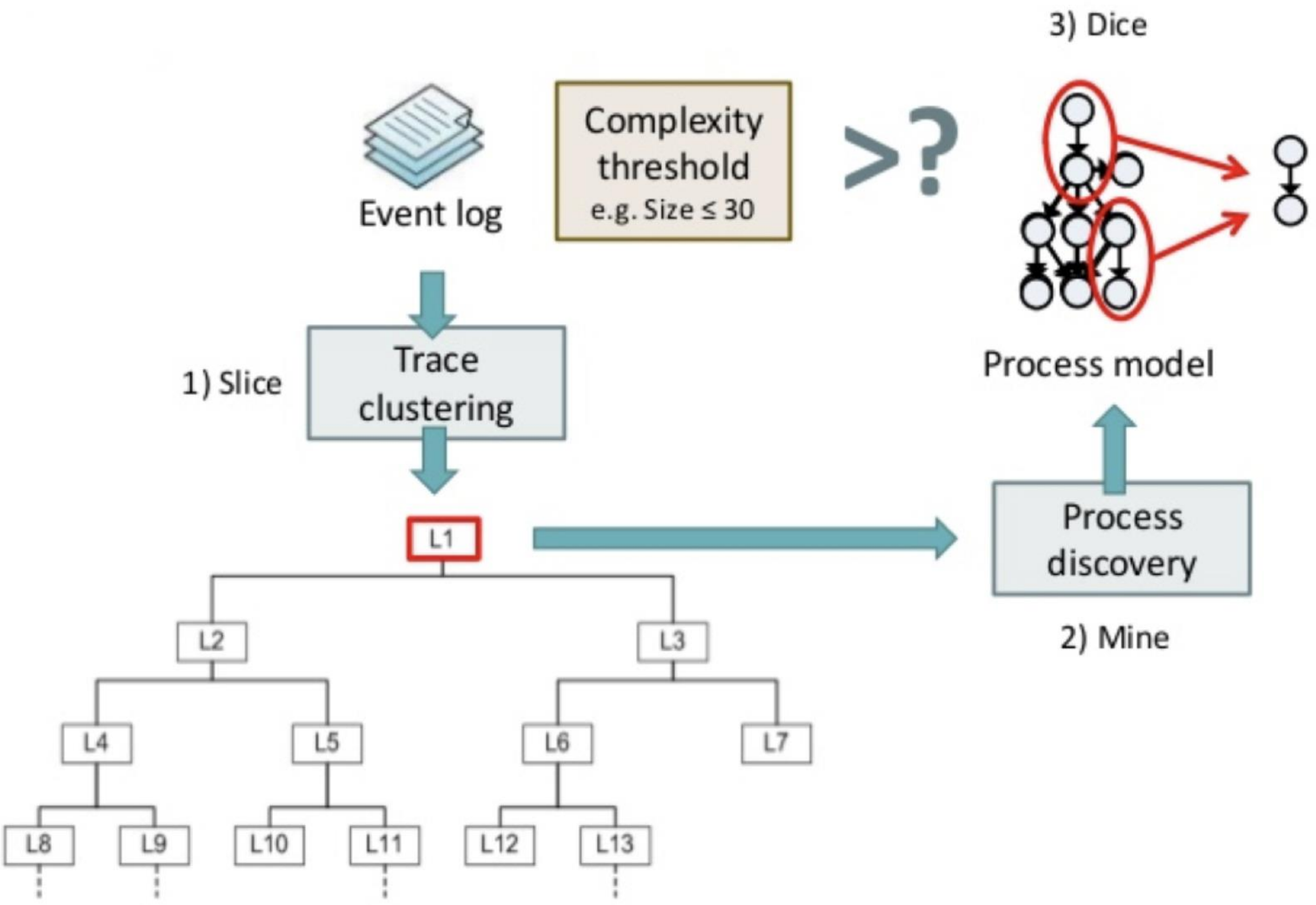
- Distance Measures
  →To calculate the similarity between cases

- Three distance measures
  - Euclidean distance(c_j,c_k) = $\sqrt{\sum_{l=1}^{n} |i_{jl} - i_{kl}|^2}$
  - Hamming distance(c_j,c_k) = $\sum_{l=1}^{n} \delta(i_{jl}, i_{kl})/n$

    where $\delta(x, y) = \begin{cases} 0 \text{ if } (x > 0 \ \wedge \ y > 0) \vee (x = y = 0) \\ 1 \text{ otherwise} \end{cases}$
  - Jaccard distance(c_j,c_k) = $1 - (\sum_{l=1}^{n} i_{jl} i_{kl}) / (\sum_{l=1}^{n} i_{jl}^2 + \sum_{l=1}^{n} i_{kl}^2 - \sum_{l=1}^{n} i_{jl} i_{kl})$

→*n*: the number of items extracted from the process log
→Case c_j: corresponds to the vector (i_{j1}, i_{j2}, … i_{jn})
→i_{jk}: the number of appearance of item *k* in the case *j*

Slice, Mine and Dice (1)



Multi-Perspective Alignments

Idea: Lift alignments from control-flow to multiple process perspectives

- "Step" in the alignment relates attribute values of a trace of an event log and of an execution sequence of a model to each other
- Attributes are: Activity, data, resources, time, …
- Assigning costs to steps, optimal alignments are defined as before

log trace	execution sequence
As @*Feb. 1, 12:31* (Res: <user>, Amount: €12,000)	*As* @*Feb. 1, 12:31* (Res: <user>, Amount: €12,000)
Aa @*Feb. 1, 12:32* (Res: John)	*Aa* @*Feb. 1, 12:32* (Res: John)
Fa @*Feb. 3, 09:00* (Res: John)	*Fa* @*Feb. 3, 09:00* (Res: John)

Multiple Trace Alignment

Christian-Albrechts-Universität zu Kiel

Construct alignment between traces of event log

- No longer a question of optimising the alignment cost for a single trace
- Global view: overall alignment cost should be minimal

Problem well-known in genomics

- Alignments of nucleic acid sequences
- Yet, also known to be an NP-complete problem

Various heuristic techniques to find multiple trace alignment that may be non-optimal

- Typically based on iterative approach
- Often based on hierarchical clustering

Identifying the Right Use Cases

Christian-Albrechts-Universität zu Kiel

- 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