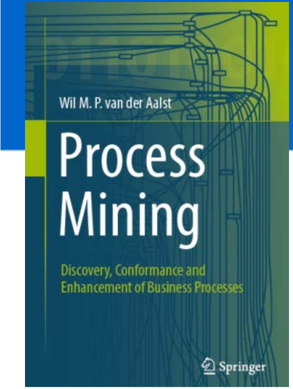
Process Mining: Data Science in Action

Guidelines for Logging

prof.dr.ir. Wil van der Aalst

www.processmining.org



Wil M. P. van der Aalst

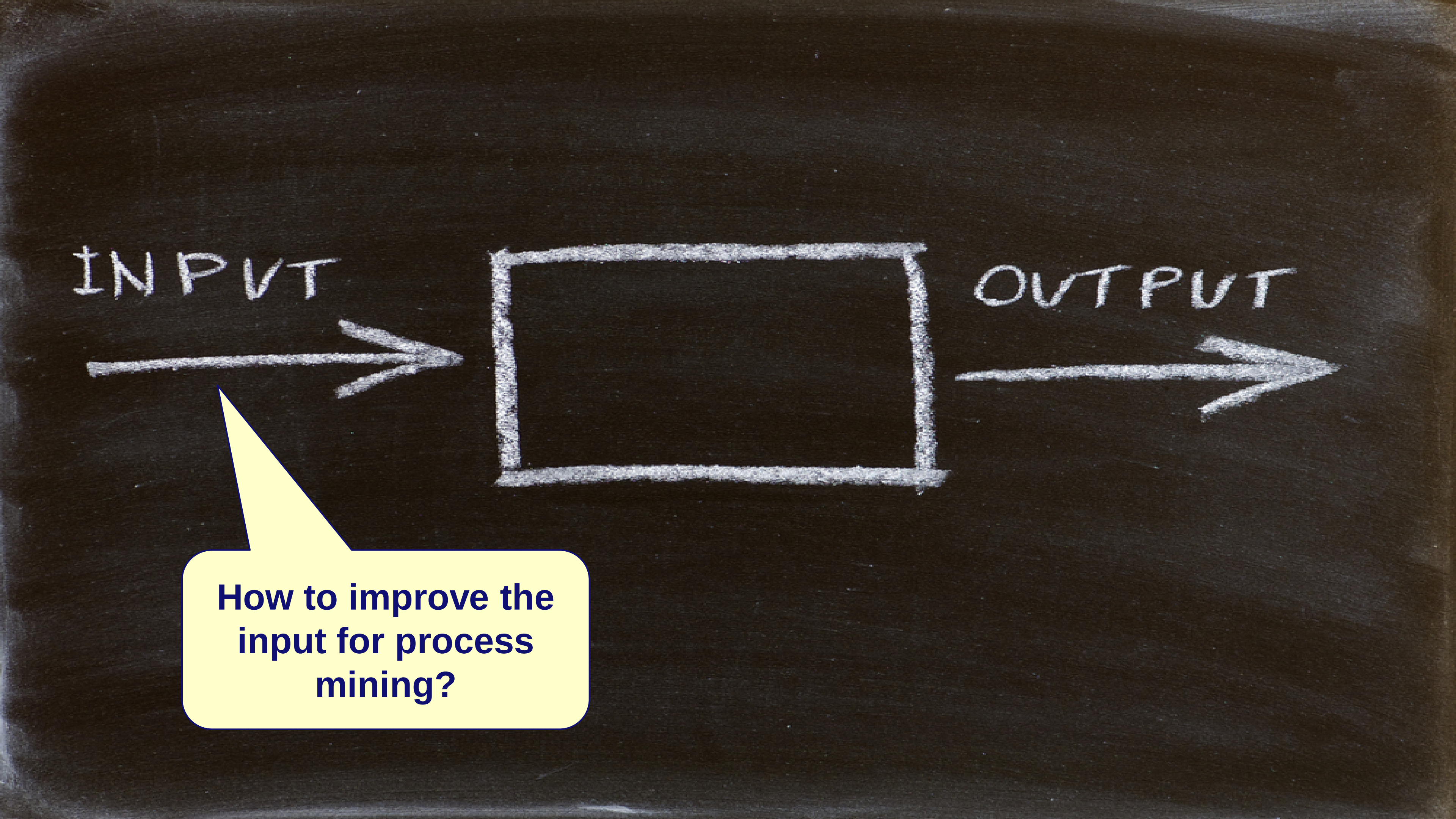
Process Mining

Discovery, Conformance and Enhancement of Business Processes

Springer

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts



How to improve the input for process mining?

	case	event	belongs to	c attribute	position	activity name	timestamp	resource	e attribute
missing data	In reality a case has been executed but it has not been recorded in the log	Events are missing within the trace although they occurred in reality.	Association between events and cases is lost (correlation problem)	Case attribute was not recorded.	Ordering of events in the trace is lost.	Activity names of events are missing.	Timestamps of events are missing.	Resources that executed an activity have not been recorded.	Event attribute was not recorded.
incorrect data	Some cases in the log belong to a different process.	Events that were not actually executed for some cases are logged	Association between events and cases are logged incorrectly.	Values corresponding to case attributes are logged incorrectly.	Order is mixed up.	Wrong activity names are recorded.	Incorrect timestamps.	Incorrect resource assigned to event.	Attributes of events are recorded incorrectly.
imprecise data			Difficult to correlate events to specific cases (too coarse).	Provided value is too coarse, e.g., city but no address.	For example concurrent events may have become been totally ordered.	Activity names are too coarse.	Days rather than minutes or seconds. Hence, precise order cannot be derived.	Just role or department is recorded.	Provided value is too coarse.
irrelevant data	Irrelevant cases are included and cannot be removed easily.	Events may be irrelevant and difficult to remove	**data quality problems**						

Terminology

- Events are things that happen and that are described by references and attributes.
- References have a reference name and an identifier that refers to some object (person, case, ticket, machine, room, etc.).
- Attributes have a name and a value, e.g., age=48 or time="28-6-2014 03:14:00".

Guidelines for Logging (G4L)

G4L-1

Reference and attribute names should have clear semantics, i.e., they should have the same meaning for all people involved in creating and analyzing event data. Different stakeholders should interpret event data in the same way.

G4L-2

There should be a structured and managed collection of reference and attribute names. Ideally, names are grouped hierarchically (like a taxonomy or ontology). A new reference or attribute name can only be added after there is consensus on its value and meaning.

G4L-3

References should be stable (e.g., identifiers should not be reused or rely on the context). For example, references should not be time, region, or language dependent. Some systems create different logs depending on the language settings. This is unnecessarily complicating analysis.

G4L-4

Attribute values should be as precise as possible. If the value does not have the desired precision, this should be indicated explicitly (e.g., through a qualifier).

For example, if for some events only the date is known but not the exact timestamp, then this should be stated explicitly.

G4L-5

Uncertainty with respect to the occurrence of the event or its references or attributes should be captured through appropriate qualifiers. For example, due to communication errors, some values may be less reliable than usual. Note that uncertainty is different from imprecision.

G4L-6

Events should be at least partially ordered. The ordering of events may be stored explicitly (e.g., using a list) or implicitly through an attribute denoting the event's timestamp. If the recording of timestamps is unreliable or imprecise, there may still be ways to order events based on observed causalities (e.g., usage of data).

G4L-7

If possible, also store transactional information about the event (start, complete, abort, schedule, assign, suspend, resume, withdraw, etc.). Having start and complete events allows for the computation of activity durations.

G4L-8

Perform regularly automated consistency and correctness checks to ensure the syntactical correctness of the event log. Check for missing references or attributes, and reference/attribute names not agreed upon. Event quality assurance is a continuous process (to avoid degradation of log quality over time).

G4L-9

Ensure comparability of event logs over time and different groups of cases or process variants. The logging itself should not change over time (without being reported). For comparative process mining, it is vital that the same logging principles are used.

G4L-10

Do not aggregate events in the event log used as input for the analysis process. Aggregation should be done during analysis and not before (since it cannot be undone). Event data should be as "raw" as possible.

G4L-11

Do not remove events and ensure provenance. Reproducibility is key for process mining. For example, do not remove a student from the database after he dropped out since this may lead to misleading analysis results. Also: concerts are not deleted - they are canceled, employees are not deleted - they are fired, etc.

G4L-12

Ensure privacy without losing meaningful correlations. Sensitive or private data should be removed as early as possible (i.e., before analysis). However, if possible, one should avoid removing correlations. Hashing can be a powerful tool in the trade-off between privacy and analysis.

Guidelines for Logging (G4L)

[G4L1] Reference and attribute names should have clear semantics.
[G4L2] There should be a structured and managed collection of reference and attribute names.
[G4L3] References should be stable.
[G4L4] Attribute values should be as precise as possible.
[G4L5] Uncertainty with respect to the occurrence of the event or its references or attributes should be indicated expelitly.
[G4L6] Events should be at least partially ordered.
[G4L7] If possible, also store transactional information about the event.
[G4L8] Perform regularly automated consistency and correctness checks.
[G4L9] Ensure comparability of event logs over time and different groups of cases or process variants.
[G4L10] Do not aggregate events in the event log used as input for the analysis process.
[G4L11] Do not remove events and ensure provenance.
[G4L12] Ensure privacy without losing meaningful correlations.

record

analyze

Avoiding GIGO

Part I: Preliminaries

Chapter 1 Introduction

Chapter 2 Process Modeling and Analysis

Chapter 3 Data Mining

Part II: From Event Logs to Process Models

Chapter 4 Getting the Data

Chapter 5 Process Discovery: An Introduction

Chapter 6 Advanced Process Discovery Techniques

Part III: Beyond Process Discovery

Chapter 7 Conformance Checking

Chapter 8 Mining Additional Perspectives

Chapter 9 Operational Support

Part IV: Putting Process Mining to Work

Chapter 10 Tool Support

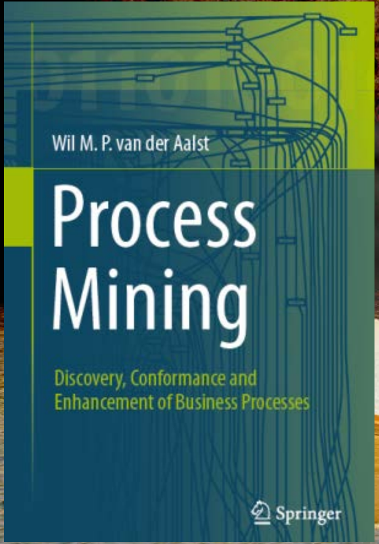
Chapter 11 Analyzing “Lasagna Processes”

Chapter 12 Analyzing “Spaghetti Processes”

Part V: Reflection

Chapter 13 Cartography and Navigation

Chapter 14 Epilogue

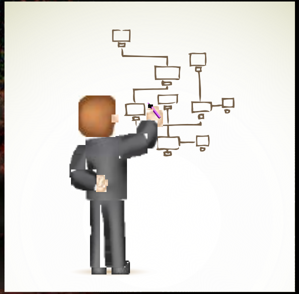


Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

Springer



W. van der Aalst. Extracting Event Data from Databases to Unleash Process Mining. BPM Center Report BPM-14-10, *BPMcenter.org*, 2014.