No Knowledge Without Processes
Process Mining as a Tool to Find Out What People and Organizations Really Do

Wil M.P. van der Aalst
Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands
w.m.p.v.d.aalst@tue.nl

Keywords: Process Mining, Process Discovery, Conformance Checking, Data Mining, Business Intelligence

Abstract: In recent years, process mining emerged as a new and exciting collection of analysis approaches. Process mining combines process models and event data in various novel ways. As a result, one can find out what people and organizations really do. For example, process models can be automatically discovered from event data. Compliance can be checked by confronting models with event data. Bottlenecks can be uncovered by replaying timed events on discovered or normative models. Hence, process mining can be used to identify and understand bottlenecks, inefficiencies, deviations, and risks. Despite the many successful applications of process mining, few people are aware of the recent advances in process mining. One of the main reasons is that process mining is not part of existing (a) data mining, (b) machine learning, (c) business intelligence, (d) process modeling, and (e) simulation approaches and tools. For example, conventional "data miners" use a very broad definition of data mining, but at the same time focus on a limited set of classical problems unrelated to process models (e.g., decision tree learning, regression, pattern mining, and clustering). None of the classical data mining tools supports process mining techniques such as process discovery, conformance checking, and bottleneck analysis. This keynote paper briefly summarizes the differences between process mining and more established analysis and modeling approaches. Moreover, the paper emphasis the need to extract process-related knowledge.

1 INTRODUCTION

The current attention for *Big Data* and *Data Science* is driven by the increasing volume and value of data. Here we focus on *event* data, i.e., information on things that happen in organizations, machines, systems, and people's lives. In (Aalst, 2014b) the term *Internet of Events* (IoE) was coined. The IoE is composed of:

- The *Internet of Content* (IoC): 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): all data related to social interaction. The IoP includes e-mail, facebook, twitter, forums, LinkedIn, etc.
- The *Internet of Things* (IoT): 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. Things may have an internet connection or be tagged using Radio-Frequency Identification (RFID), Near Field Communication (NFC), etc.
- The *Internet of Locations* (IoL): refers to all data that have a spatial dimension. With the uptake of mobile devices (e.g., smartphones) more and more events have geospatial attributes.

Process mining aims to exploit the IoE to learn things related to the behavior of people, organizations, machines, and systems (Aalst, 2011). The starting point for process mining is an *event log*. Each event in such a log refers to an *activity* (i.e., a well-defined step in some process) and is related to a particular *case* (i.e., a *process instance*). The events belonging to a case are *ordered* and can be seen as one "run" of the process. Event logs may store additional information about events. In fact, whenever possible, process-mining techniques use extra information such as the *resource* (i.e., person or device) executing or initiating the activity, the *timestamp* of the event, or *data elements* recorded with the event (e.g., the size of an order).

Event logs are used to conduct four types of process mining (see (Aalst, 2011) for details):

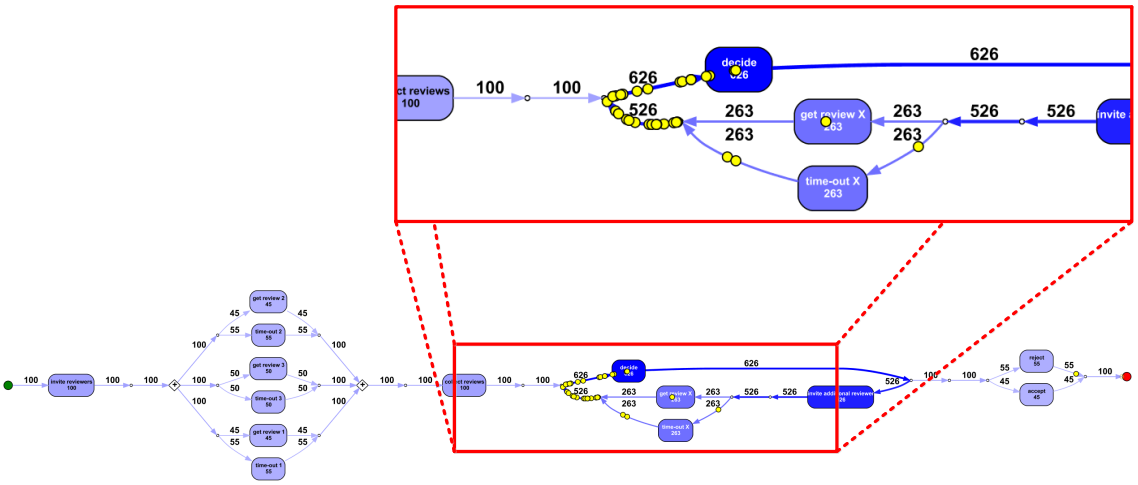

Figure 1: A process model discovered by ProM's inductive miner (Leemans et al., 2014). Real cases (see yellow dots) are replayed on the model to reveal bottlenecks and deviations.

- The first type of process mining is *discovery*. A discovery technique takes an event log and produces a model without using any a priori information (see Fig. 1). Process discovery is the most prominent process-mining technique. For many organizations it is surprising to see that existing techniques are indeed able to discover real processes merely based on example behaviors stored in event logs.
- The second type of process mining is *conformance*. Here, an existing process model is compared with an event log of the same process. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa.
- The third type of process mining is *enhancement*. Here, the idea is to extend or improve an existing process model by directly using information about the actual process recorded in some event log. Whereas conformance checking measures the alignment between model and reality, this third type of process mining aims at changing or extending the a priori model. For instance, by using timestamps in the event log one can extend the model to show bottlenecks, service levels, and throughput times.
- The fourth type of process mining is *operational support*. The key difference with the former three types is that analysis is not done off-line, but used to influence the running process and its cases in some way. Based on process models, either discovered through process mining or (partly) made by hand, one can check, predict, or recommend activities for running cases in an online setting. For example, based on the discovered model one can predict that a particular case will be late and propose counter-measures.

It is important to realize that the above types of analysis are only possible due to the combination of data and processes. Next to "data scientists" there is a need for "process scientists" that understand that knowledge discovery should not shy away from the complexities involving dynamic behavior. Understanding the processes at hand is key when analyzing systems, organizations, or human behavior.

2 HOW PROCESS MINING IS DIFFERENT FROM ...?

This section positions process mining versus other approaches such as data mining, machine learning, business intelligence, process modeling, and simulation (see Fig. 2). It is partly inspired by some of the blog postings on Fluxicon's website that compare process mining with other approaches, see (Fluxicon, 2014).

2.1 Process Mining Versus Data Mining

Data mining techniques can be divided into supervised and unsupervised learning. For *supervised learning* one needs labeled data (i.e., there is a response variable that labels each instance) and the goal is to explain this response variable (also called the dependent variable) in terms of predictor variables (also called independent variables). *Classification techniques* (e.g., decision tree learning) typically assume a categorical response variable (or the

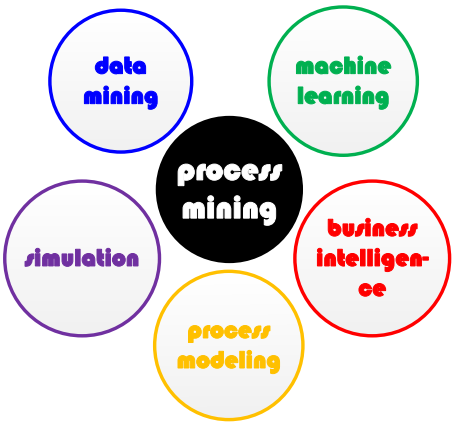

Figure 2: How process mining is different from ....

response variable is made categorical) and the goal is to classify instances based on the predictor variables. *Regression techniques* assume a numerical response variable. The goal is to find a function that fits the data with the least error. *Unsupervised learning* techniques assume unlabeled data, i.e., the variables are not split into response and predictor variables. Examples include *clustering* (e.g., k-means clustering and agglomerative hierarchical clustering) and *pattern discovery* (e.g., association rules). Although both process mining and data mining start from data, data mining techniques are typically *not* process-centric and do *not* focus on event data. For data mining techniques the rows (instances) and columns (variables) can mean anything. For process mining techniques, we assume event data where events refer to process instances and activities. Moreover, the events are ordered and we are interested in end-to-end processes rather than local patterns. End-to-end process models and concurrency are essential for process mining. Moreover, topics such as process discovery, conformance checking, and bottleneck analysis are not addressed by traditional data mining techniques and tools.

Data mining and process mining are complementary approaches that can strengthen each other. For example, consider the combined approach presented in (Leoni and Aalst, 2013) that allows for decision mining and performance prediction *in a process context*.

2.2 Process Mining Versus Machine Learning

Process discovery is the most visible form of process mining. Given an event log, a process model is discovered. Different notations are used as a *representational bias*: various types Petri nets (place-transition nets, workflow nets, colored nets, etc.), Business Process Model and Notation (BPMN) diagrams, Event-driven Process Chains (EPCs), UML Activity Diagrams, etc. Some process discovery approaches use transition systems as an intermediate format. Learning a process model can be viewed as a machine learning problem. However, process mining extends far beyond process discovery. Moreover, classical machine learning approaches such as learning hidden Markov models and language identification are significantly different from process discovery approaches such as the Alpha algorithm, the ILP miner, the inductive miner, the heuristic miner, and the various genetic process mining algorithms (Aalst, 2011). The well-known Baum-Welch algorithm is an Expectation-Maximization (EM) algorithm that, given a set of observation sequences, derives a hidden Markov model with a given number of states that maximizes the probability of producing a collection of traces (Alpaydin, 2010). Unlike more recent process mining approaches the resulting models are sequential, i.e., concurrency and other higher level control-flow constructs cannot be discovered. Many inductive inference problems have been studied since Gold's 1967 paper "Language Identification in the Limit" (Gold, 1967). The Myhill-Nerode theorem can be used to minimize the transition system for regular languages, but cannot cope with noise and concurrency. The setting for such approaches is very different from process mining. Typically, no concurrency is considered and the logs are assumed to be complete. However, event logs, by definition, only contain example behavior. This makes process mining also very different from classical synthesis approaches for formal models, e.g., the Theory of Regions for Petri nets (Ehrenfeucht and Rozenberg, 1989; Cortadella et al., 1998).

2.3 Process Mining Versus Business Intelligence

Boris Evelson of Forrester Research defines *Business Intelligence* (BI) as "a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision making". Although this definition does not exclude a process focus, BI methodologies and tools are typically not process-aware. BI products tend to focus on fancy-looking dashboards and rather simple reports, rather than a deeper analysis of the data collected. Moreover, like data mining, BI is not tailored towards the analysis of event data.

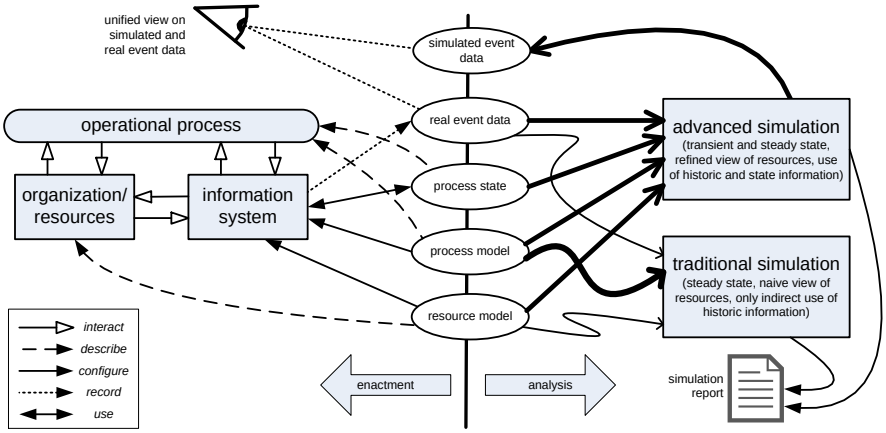

Figure 3: Process mining enables non-traditional ways of using simulation. Reality and simulated alternatives can both be viewed using process mining, thus making comparison easy. Moreover, operational support can realized using a combination of historic and current data (Aalst, 2014a).

2.4 Process Mining Versus Process Modeling

Process mining is "evidence-based", i.e., based on observed behavior a model is discovered and evaluated. Process modeling approaches do not ensure a close correspondence between reality and model. When modeling processes the designer tends to concentrate on the "normal" or "desirable" behavior. For example, the model may only cover 80% of the cases assuming that these are most representative. Typically this is not the case as the other 20% may cause 80% of the problems. Abstraction is fine, but simplified models should be based on evidence and deviations still need to be considered. The reasons for such oversimplifications are manifold (Aalst, 2011). The designer and management may not be aware of the many deviations that take place. Moreover, the perception of people may be biased depending on their role in the organization. Hand-made models tend to be subjective, and often there is a tendency to make things too simple just for the sake of understandability. Process mining, in particular the recently developed alignment-based approaches (Aalst et al., 2012), ensures a close correspondence between observed and modeled behavior. Moreover, event logs can be used to "breathe life" into otherwise static process models (Aalst, 2011).

2.5 Process Mining Versus Simulation

Process mining is based on facts, simulation results are based on simulation models rather than event data. The value of a simulation highly depends on the quality of the model. Process mining is all about understanding the current "as-is" processes. Simulation is more about playing out alternative "to-be" scenarios. The simulation model first aims to mimic the current process and is then modified to estimate the effects of changes (e.g., changes to the ordering of activities, adding resources, or new priority rules). Simulation can greatly benefit from process mining because process mining can provide better initial models.

Many organizations have purchased simulation tools, but these are rarely used for two obvious reasons: (a) it takes too much effort to build reliable simulation models and (b) people often do not believe simulation results because these are not based on reality (the model can be tweaked to produce any result). As discussed in (Aalst, 2014a), process mining can help to overcome these problems (see Fig. 3).

3 CONCLUSION

As argued in this paper, process mining is different from data mining, machine learning, business intelligence, process modeling, and simulation. Although process mining is often associated with these complementary approaches, existing tools for data mining, machine learning, business intelligence, process modeling, and simulation do not include techniques for process discovery, conformance checking, etc. Figure 4 shows the unique positioning of process mining (Aalst, 2011).

In recent years, academics working on *Business Process Management* (BPM) have embraced process mining as a new and exciting technology. BPM is a discipline involving any combination of modeling, automation, execution, control, measurement and op-

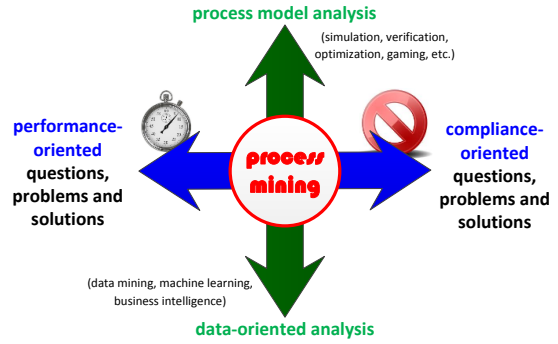

Figure 4: Positioning process mining.

timization of business activity flows, in support of enterprise goals, spanning systems, employees, customers and partners within and beyond the enterprise boundaries (Swenson, 2014). At academic BPM conferences a considerable portion of papers is using or proposing new process mining techniques. Also process mining tools are readily available. Unfortunately, the main BPM vendors are lagging behind and many end-users are unaware of this more data-centric approach to process analysis.

Process mining provides the glue between data & process (linking event data to process models), business & IT (the evidence-based nature creates commitment and makes both groups speak the same language), BI & BPM, performance & compliance (deviations and bottlenecks are analyzed using the same logs and tools), runtime & design time (process mining can be applied on historic data and on running cases), etc. This new glue provides a valuable set of tools for a new profession: the "process scientist". Process mining helps the process scientist to find out what organizations and people really do and use these insights to improve things.

The fact that process mining problems can be decomposed (Aalst, 2013b; Aalst, 2013a) also makes it feasible to analyze "Big Event Data". The more exploratory types of process mining are often linear in the size of the log. More precise techniques (e.g., for conformance checking) require more computing time, but can be distributed easily.

REFERENCES

Aalst, W. van der (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer-Verlag, Berlin.

Aalst, W. van der (2013a). A General Divide and Conquer Approach for Process Mining. In *Federated Conference on Computer Science and Information Systems (FedCSIS 2013)*, pages 1–10. IEEE Computer Society.

Aalst, W. van der (2013b). Decomposing Petri Nets for Process Mining: A Generic Approach. *Distributed and Parallel Databases*, 31(4):471–507.

Aalst, W. van der (2014a). Business Process Simulation Survival Guide. In Brocke, J. and Rosemann, M., editors, *Handbook on Business Process Management*, International Handbooks on Information Systems, Springer-Verlag, Berlin.

Aalst, W. van der (2014b). Data Scientist: The Engineer of the Future. In *Proceedings of the I-ESA Conference*, volume 7 of *Enterprise Interoperability*, pages 13–28. Springer-Verlag, Berlin.

Aalst, W. van der, Adriansyah, A., and Dongen, B. van (2012). Replaying History on Process Models for Conformance Checking and Performance Analysis. *WIREs Data Mining and Knowledge Discovery*, 2(2):182–192.

Alpaydin, E. (2010). *Introduction to Machine Learning*. MIT press, Cambridge, MA.

Cortadella, J., Kishinevsky, M., Lavagno, L., and Yakovlev, A. (1998). Deriving Petri Nets from Finite Transition Systems. *IEEE Transactions on Computers*, 47(8):859–882.

Ehrenfeucht, A. and Rozenberg, G. (1989). Partial (Set) 2-Structures - Part 1 and Part 2. *Acta Informatica*, 27(4):315–368.

Fluxicon (2014). Flux Capacitor: How is Process Mining Different `http://fluxicon.com/blog/`.

Gold, E. (1967). Language Identification in the Limit. *Information and Control*, 10(5):447–474.

Leemans, S., Fahland, D., and Aalst, W. van der (2014). Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour. In *Business Process Intelligence (BPI 2014)*, volume 171 of *Lecture Notes in Business Information Processing*, pages 66–78. Springer-Verlag, Berlin.

Leoni, M. and Aalst, W. van der (2013). Data-Aware Process Mining: Discovering Decisions in Processes Using Alignments. In Shin, S. and Maldonado, J., editors, *ACM Symposium on Applied Computing (SAC 2013)*, pages 1454–1461. ACM Press.

Swenson, K. (2014). One Common Definition for BPM. `social-biz.org`.