Development of the Process Mining Discipline

(prof.dr.ir. Wil van der Aalst, RWTH Aachen University)

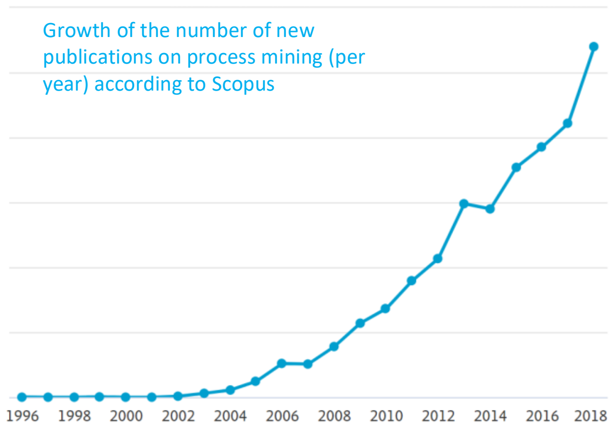
It is exciting to see the spectacular developments in process mining since I started to work on this in the late 1990-ties. Many of the techniques we developed 15-20 years ago have become standard functionality in today's process mining tools. Therefore, it is good to view current and future developments in this historical context.

This chapter starts with a brief summary of the history of process mining showing how ideas from academia got adopted in commercial tools. This provides the basis to talk about the expanding scope of process mining, both in terms of applications and in terms of functionalities supported. Despite the rapid development of the process mining discipline, there are still several challenges. Some of these challenges are new, but there are also several challenges that have been around for a while and still need to be addressed urgently. This requires the concerted action of process mining users, technology providers, and scientists.

Adoption of traditional process mining techniques

Process mining started in the late nineties when I had a sabbatical and was working for one year at the University of Colorado in Boulder (USA). Before, I was mostly focusing on concurrency theory, discrete event simulation, and workflow management. We had built our own simulation engines (e.g., ExSpect) and workflow management systems. Although our research was well-received and influential, I was disappointed by the average quality of process models and the impact process models had on reality. In both simulation studies and workflow implementations, the real processes often turned out to be very different from what was modeled by the people involved. As a result, workflow and simulation projects often failed. Therefore, I decided to focus on the analysis of processes through event data [1]. Around the turn of the century, we developed the first process discovery algorithms [2]. The Alpha algorithm was the first algorithm able to learn concurrent process models from event data and still provide formal guarantees. However, at the time, little event data were available and the assumptions made by the first algorithms were unrealistic. People working on data mining and machine learning were (and perhaps still are) not interested in process analysis. Therefore, it was not easy to convince other researchers to work on this. Nevertheless, for me, it was crystal clear that process mining would become a crucial ingredient of any process management or process improvement initiative.

In the period that followed, I stopped working on the traditional business process management topics and fully focused on process mining. It is interesting to see that concepts such as conformance checking, organizational process mining, decision mining, token animation, time prediction, etc. were already developed and implemented 15 years ago [2]. These capabilities are still considered to be cutting-edge and not supported by most of the commercial process mining tools.



- 1999 Start of process mining research at TU/e
- 2000-2002 Development of the first process mining algorithms (Alpha and Heuristic miner)
- 2004 Release of the first version of the ProM framework
- 2004-2006 Pioneering work on token-based conformance checking, organization mining, decision mining, etc.
- 2007 First process mining company (Futura PI)
- 2010 Pioneering work on alignment-based conformance checking and operational support
- 2009-2011 Founding of Fluxicon, Celonis, and ProcessGold
- 2011 First process mining book
- 2014 Coursera process mining MOOC
- 2016 “Process mining data science in action” book
- 2018 First Market Guide for Process Mining by Gartner
- 2018 Over 30 software vendors offer process mining tools
- 2018 Celonis becomes a Unicorn
- 2019 First international process mining conference takes place in Aachen (ICPM 2019)

Figure 1: Summary of the history of process mining also showing the growth of scientific papers on the topic.

Figure 1 illustrates the development of the field. On the one hand, the graph shows the growth of scientific process mining literature. Each year, a growing number of process mining paper is published in journals and presented at conferences. On the other hand, the right-hand side of the figure also mentions a few milestones illustrating the uptake in industry. The first process mining company (Futura PI) was founded in 2007 by one of my students (Peter van de Brand). The software was later integrated into the tools of Pallas Athena and Perceptive Software. A few years later, Fluxicon, Celonis, and ProcessGold were founded. Concurrently, the first process mining books appeared and the first online course for process mining was created (followed by over 120.000 participants). However, until 2015 the practical adoption of process mining in industry was limited. Only in recent years, the actual usage accelerated [3, 4]. This is illustrated by the growing number of process mining vendors. Currently, there are over 30 process mining vendors (e.g., Celonis, Disco, ProcessGold, myInvenio, PAFnow, Minit, QPR, Mehrwerk, Puzzledata, LanaLabs, StereoLogic, Everflow, TimelinePI, Signavio, and Logpickr). In 2018, the first International Conference on Process Mining (ICPM) was organized illustrating the growing maturity of the field. Moreover, as this book shows there are many exciting applications in organizations such as Siemens, BMW, Uber, Springer, ABB, Bosch, Bayer, Telekom, etc.

Although some process mining vendors have added conformance checking techniques and more advanced discovery techniques like the inductive mining approach, the basis of most commercial process mining tools is still the *Directly-Follows Graph* (DFG). This was actually the graph that served as input for the classical Alpha algorithm twenty years ago. The DFG can also be viewed as a traditional transition system or a Markov chain (when adding probabilities). A DFG is a graph with nodes that correspond to activities and directed edges that correspond to directly-follows relationships [2, 5]. The frequency on an arc connecting activity X to activity Y shows how often X is directly followed by Y for a specific case (i.e., process instance). Similarly, the arc can be annotated with time information to show bottlenecks. Using frequencies it is possible to seamlessly simplify such process models. It is also possible to animate the cases using tokens moving along the directed arcs. This is all easy to understand and highly scalable. Therefore, this basic functionality is present in all of today's process mining tools.

Expanding the scope of process mining

Over the last two decades, the scope of process mining expanded in different ways. First of all, process mining grew out of academia into industry. Also the number of application domains expanded [6]. Traditionally, applications were limited to financial or administrative processes. The Order-to-Cash (O2C) and Purchase-to-Pay (P2P) processes are obvious candidates to apply process mining. However, nowadays process mining is also applied in healthcare, logistics, production, customs, transportation, user-interface design, security, trading, energy systems, smart homes, airports, etc. This makes perfect sense since event data and processes are not limited to specific application domains. Finally, there is also a clear expansion in the capabilities of process mining tools. Initially, the focus was exclusively on process discovery based on historic data. However, the scope of process mining expanded far beyond this as is illustrated by the four trends depicted in Figure 2 and discussed next.

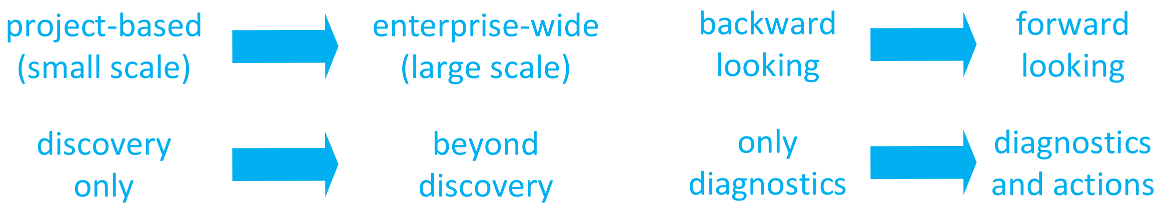


Figure 2: Four trends showing that the scope of process mining is expanding.

The scope of process mining expanded from a tool for a data scientist used in an improvement project to the *enterprise-wide*, *continuous* application of process mining. Process mining tools only supporting the construction of DFGs with frequencies and times tend to be used in smaller projects only. These projects often have a limited scope and duration. As a result, few people use the results and there is no support for continuous improvement. Given the investments needed for data preparation, it is often better to apply process mining at an enterprise-wide scale with many people using the results on a daily basis. It does not make sense to see process mining as a one-time activity. However, the enterprise-wide, continuous application of process mining requires substantial resources in terms of computing power and data management (e.g., to extract the data and convert these into the right format). Moreover, to lower the threshold for a larger process mining community within an organization, one needs to create customized dashboards. Therefore, process mining needs to be supported by higher-level management to realize the scale at which it is most effective.

Initially, process mining efforts focused on process discovery [2]. However, over time it has become clear that *process discovery is just the starting point* to process improvement. One can witness an uptake in conformance checking and performance analysis techniques. Moreover, process mining is often combined with data mining and machine learning techniques to find root causes for inefficiencies and deviations. Although process discovery will remain important, attention is shifting to the steps following discovery using optimization, machine learning, and simulation.

A third trend is the shift in focus *from backward looking to forward looking*. Traditional process mining techniques start from historic data. This can be used to diagnose conformance and compliance problems. However, organizations are often more interested in what is happing now or what is going to happen next. Backward-looking process mining can be used to fundamentally improve processes, but provides little support for the day-to-day management of processes. Therefore, event data need to be updated continuously and process mining techniques need to be able to analyze cases that are still running. This is needed to control and influence the running process instances. Some of the commercial process mining tools provide excellent capabilities to show the current state of the process. A next step is the application of more forward-looking techniques able to predict what will happen to individual cases and where bottlenecks are likely to develop. Techniques for operational support (i.e., detecting compliance and performance problems at runtime, predicting such problems, and recommending actions) have been around for more than a decade. However, their quality still leaves much to be desired. The problem is that cases (i.e., process instances) highly influence each other when competing for resources. Moreover, there may be a range of contextual factors influencing processes. By simply applying existing machine learning and data mining techniques, one cannot get any reliable predictions. Hence, additional work is needed.

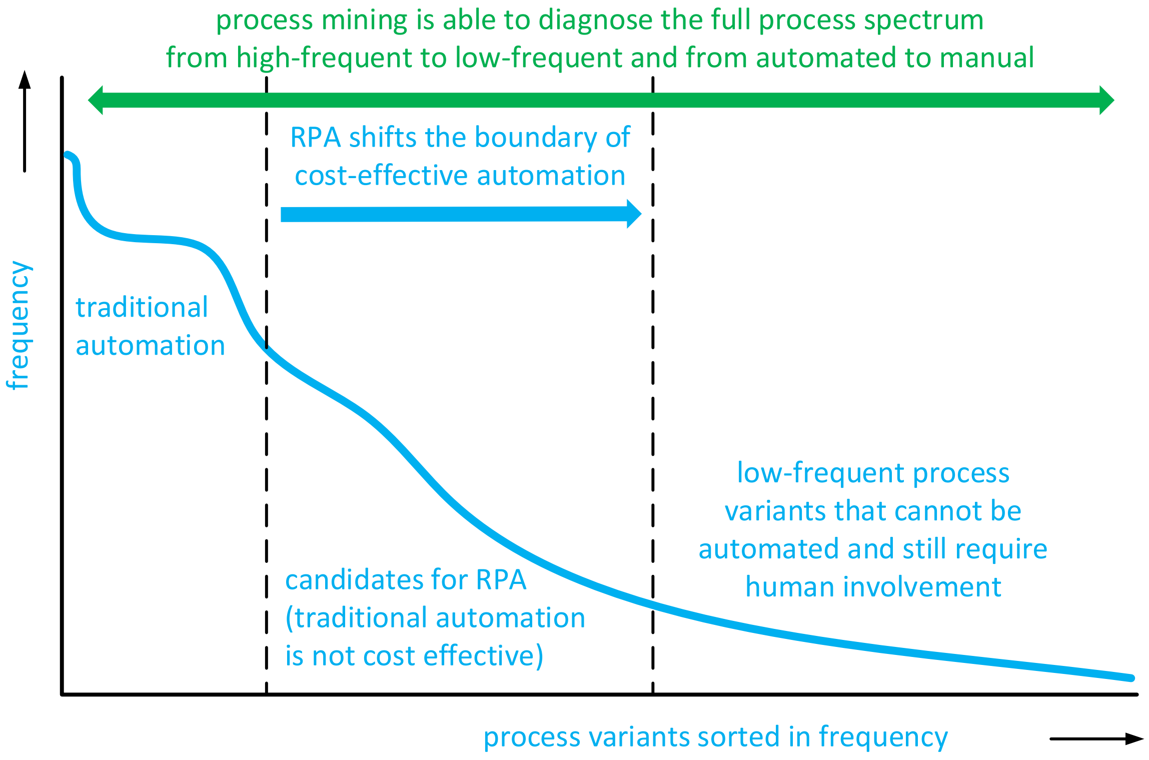

process mining is able to diagnose the full process spectrum
from high-frequent to low-frequent and from automated to manual
RPA shifts the boundary of
cost-effective automation
frequency
traditional
automation
low-frequent process
variants that cannot be
automated and still require
human involvement
candidates for RPA
(traditional automation
is not cost effective)
process variants sorted in frequency

Figure 3: Process mining can be used to identify candidates for RPA and monitor any mixture of automated/non-automated frequent/infrequent process variants.

The fourth trend is the increased focus on actually improving the process. Process mining tends to focus on diagnostics and not on the *interventions* that should follow. Insights generated by process mining should be actionable. Therefore, process mining is increasingly combined with *Robotic Process Automation (RPA)*. Process mining can be used to identify manual tasks that can be automated and monitor software robots. This allows for the automation of processes for which traditional workflow automation would be too expensive. Figure 3 shows the spectrum of process variants. High-frequent variants are candidates for automation, but lower frequent variants cannot be automated in a cost-effective manner. RPA helps to shift the boundary where (partial) automation is still cost-effective. Interestingly, process mining can be used before and after automation for any mixture of process variants (automated or not). However, RPA is just one of many ways to turn process mining diagnostics into actions. Here, I would also like to coin the term *Robotic Process Management (RPM)* to refer to automatic process interventions. Unlike RPA, RPM is not automating steps in the operational process. RPM translates process mining diagnostics into management actions. For example, when a bottleneck emerges RPM may take actions such as alerting the manager, informing the affected customers, assigning more workers, etc. Another example would be to prioritize targeted auditing activities when a significant increase in process deviations occurs. These examples show that the diagnostics provided by process mining are just the starting point.

An inconvenient truth

Despite the rapid developments in process mining, many of the original challenges remain [2]. Although there has been considerable progress in process mining research, commercial tools tend to not use the state-of-the-art due to pragmatic reasons such as speed and simplicity. Commercial software tends to make "short-cuts" that seem harmless at first, but inevitably lead to problems at a later stage.

The first inconvenient truth is that *filtered Directly-Follows Graphs (DFGs) have well-known problems*. DFGs cannot express concurrency and filtering them may provide misleading results. Yet, the default discovery techniques used by commercial tools are all based on filtered DFGs.

To illustrate the problem consider a fictive purchasing process consisting of six steps: *place order*, *receive invoice*, *receive goods*, *pay order*, and *close*. In this idealized process, all five activities are performed for all procurement orders. The process always starts with activity *place order* and ends with activity *close*. However, the three middle activities are executed in any order. For example, in rare cases the organization pays before receiving the invoice and goods. Hence, there are six different process variants. In our event log there is information about 2000 orders, i.e., in total there are 10000 events. The most frequent variant is ⟨*place order*, *receive invoice*, *receive goods*, *pay order*, *close*⟩ which occurs 857 times. The least frequent variant is ⟨*place order*, *pay order*, *receive goods*, *receive invoice*, *close*⟩ which occurs only 4 times.

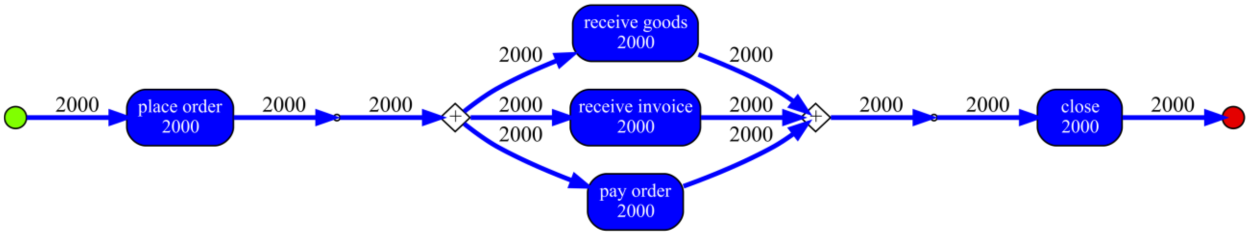


Figure 4: Process model discovered by ProM's Inductive Miner. Note that all activities occur once per procurement order.

Figure 4 shows the process model discovered by ProM's Inductive Miner [2]. The three unordered activities in the middle are preceded by an AND-split and are followed by an AND-join. The model correctly shows that all activities happen precisely 2000 times. Applying other basic process discovery algorithms like the Alpha algorithm and region-based techniques yield the same compact and correct process model (but then often directly expressed in terms of Petri nets).

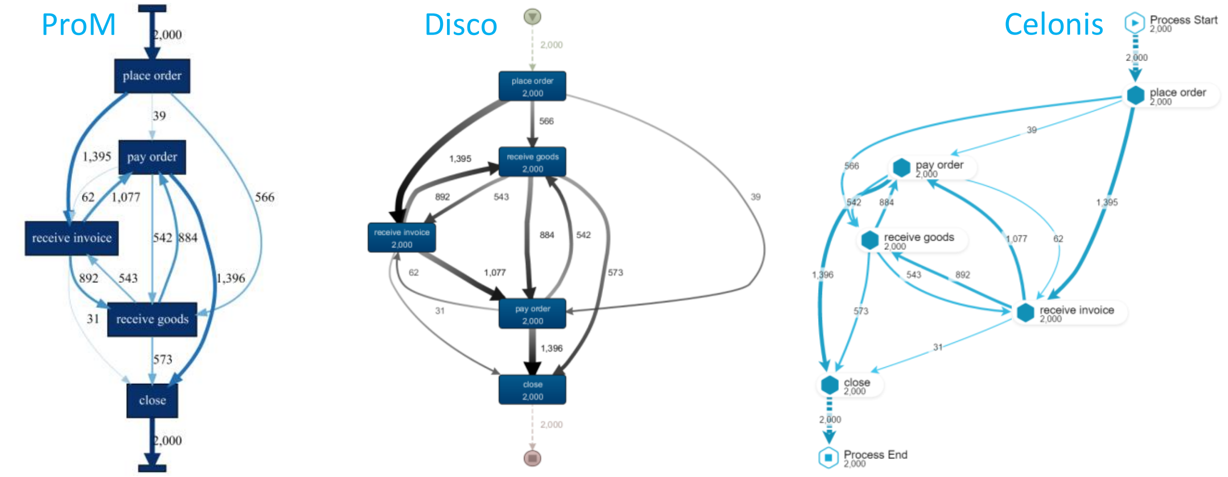


Figure 5: Three identical DFGs returned by ProM, Disco, and Celonis.

Let us now look at the corresponding Directly-Follows Graphs (DFG) used by most commercial process mining tools [2]. Figure 5 shows three DFGs generated by ProM, Disco, and Celonis. Apart from layout differences, these three models are identical. Surprisingly, the DFGs suggest that there are loops in the process although each activity is executed precisely once for each order. The process also seems more complex. To simplify such DFGs, all process mining tools can leave out infrequent paths to simplify the model. However, this may lead to highly misleading results, e.g., frequencies no long add up and averages are based on unclear fragments of behavior.

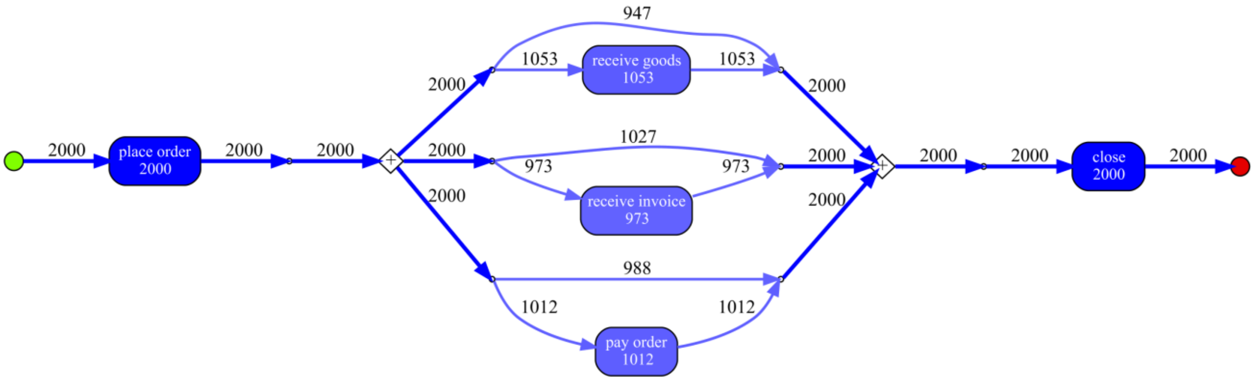


Figure 6: Process model discovered by ProM's Inductive Miner for the event logs where the middle three activities can be skipped.

To further illustrate the problem, we now consider a variant of the order process where each of the middle activities is skipped with 50% probability. This means that for approximately 50%·50%·50%=12.5% of cases only the activities *place order* and *close* are performed. Figure 6 shows the process model discovered by ProM's Inductive Miner clearly showing that the three middle activities can be skipped. For example, the process model shows that for 988 of the 2000 orders there was no payment. Figure 7 shows the corresponding Petri net model without frequencies.

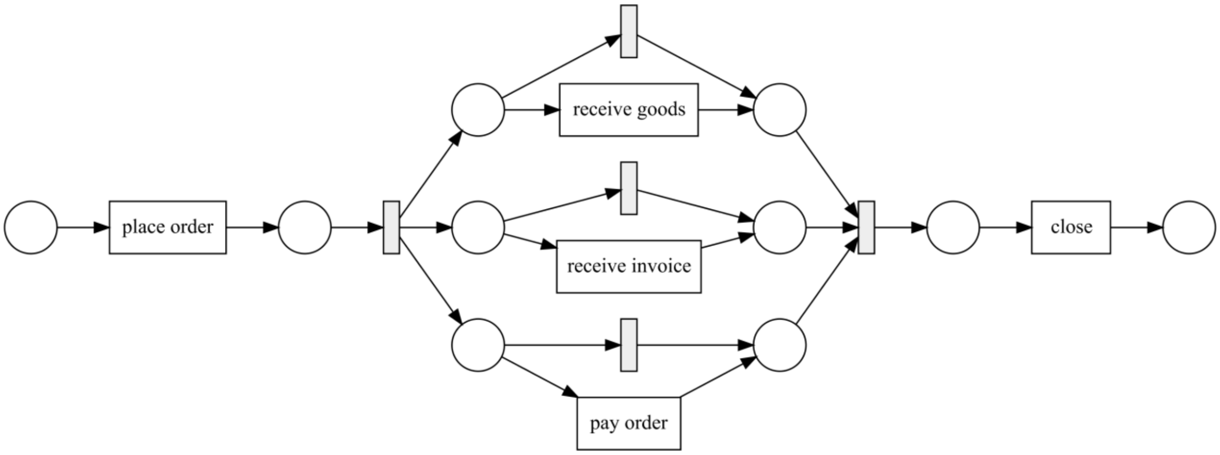


Figure 7: Petri net discovered by ProM showing that each of the three middle activities can be skipped.

As before, we can also discover the DFG for this second event log. The result is shown in Figure 8. Again ProM, Disco, and Celonis generate identical DFGs. We can see that for 236 orders, all three middle activities are skipped. Again we see loops that do not exist and the underlying structure of the process clearly depicted in Figures 6 and 7 is invisible in the three DFGs.

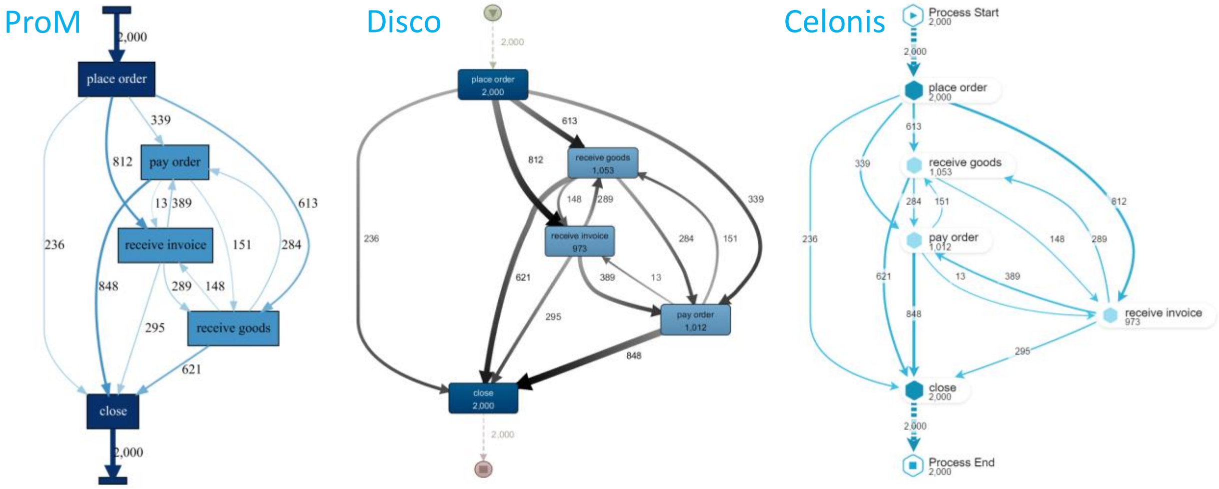


Figure 8: Three identical DFGs for the process where the three middle activities can be skipped.

As mentioned DFGs can be seamlessly simplified by leaving out infrequent activities and arcs. Of the three middle activities, activity *receive goods* is most frequent. Hence, this activity remains when we filter the model to retain the three most frequent activities. Activities *place order* and *close* occur 2000 times, and activity *receive goods* occurs 1053 times. Figure 9 shows the filtered DFGs generated by Disco and Celonis. Now there are some surprising differences. These differences illustrate that filtered DFGs can be very misleading.

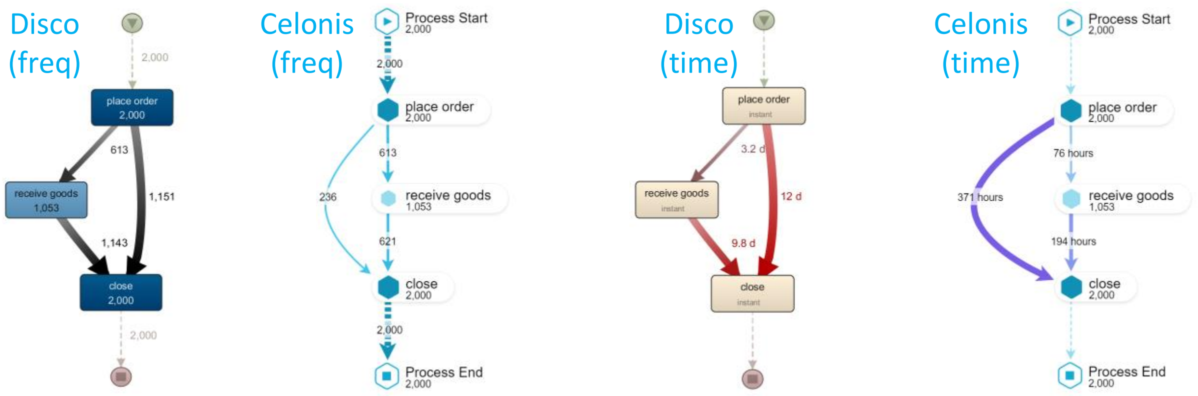


Figure 9: Filtered DFGs generated by Disco and Celonis showing frequencies (left) and time (right).

Note that Figure 8 and Figure 9 are based on the same event data and the frequencies of activities are the same in all DFGs shown, e.g., activity *receive goods* occurs 1053 times in all DFGs. However, the information on the arcs is very different. Let us first focus on the connection between activity *place order* and activity *close*. Activity *place order* directly followed activity *close* in 236 cases. This was correctly shown in Figure 8. However, Celonis reports the same number (i.e., 236) when the other activities have been removed (Figure 9). However, this can of course no longer be the case. After removing *receive invoice* and *pay order* there are more cases where *place order* is directly followed activity *close*. Also, Disco reports an incorrect number (i.e., 1151). After removing the two activities, there are 947 cases where *place order* is directly followed by activity *close* and the average time between these activities for these cases is 13.8 days. Surprisingly, Disco reports 12 days and Celonis reports 15.5 days. Hence, Celonis and Disco report different frequencies and times and both fail to show correct values for frequencies and times.

Another example is the connection between activity *receive goods* and activity *close*. If we project the log onto the three remaining activities, we can see that there are 1053 cases where activity *receive goods* directly follows activity *close* and this takes on average 8.5 days. Disco reports a frequency of 1143 (too high) and an average time of 9.8 days (too high). Celonis reports a frequency of 621 (too low) and an average time of 8 days (too low).

Moreover, the diagrams in Figure 9 are internally inconsistent. The frequencies involved in split and join should add up to 2000, but we can witness 613+1151 and 236+613 for the split and 1143+1151 and 236+621 for the join. These inconsistencies will confuse any user that would really like to understand the process. Figure 10 shows that this is not necessary.

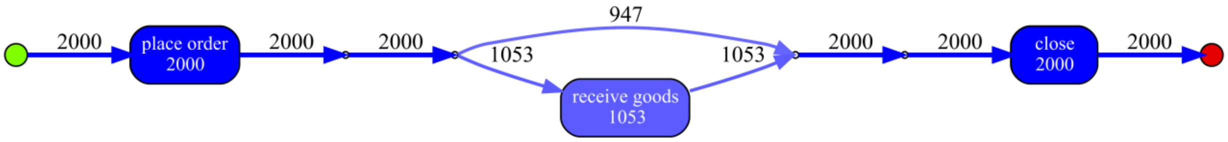


Figure 10: Process model discovered by ProM's Inductive Miner showing the correct frequencies after abstracting from the two lower frequent activities.

These small examples illustrate an inconvenient truth. Simplistic discovery techniques providing just a DFG with sliders to simplify matters are not adequately capturing the underlying process. The frequencies and times reported are wrong (or at best misleading) and when activities are not executed in a fixed order there will always be loops in the model even when these do not exist in reality. These problems are not specific for Disco or Celonis. Almost all commercial process mining tools make shortcuts to ensure good performance and provide similar results. Although this is a known problem that has been reported repeatedly over the last decade, vendors are reluctant to address it. There are two main reasons: *simplicity* and *performance*. Petri nets, BPMN models, UML diagrams, etc. are considered to be too complicated for the user. However, the price to pay for this simplicity is the presence of spaghetti-like diagrams with many non-existing loops. To ensure good performance, filtering of the DFG is done on the graph rather than on the original data. This explains the incorrect frequencies and times in Figure 9.

Another inconvenient truth is the limited support for *conformance checking* [2]. Although conformance checking is considered to be important from a practical point is view, it is still not very well support and rarely used. Conformance checking approaches ranging from token-based relay [7] to alignments [8] have been developed over the past two decades. Several vendors try to support conformance checking by comparing discovered DFGs with normative DFGs derived from hand-made process models. This, of course, does not work. Compare for example the DFGs in Figure 5 with the DFGs in Figure 8. The only difference is the connection between activity *place order* and activity *close*. However, it is very difficult to see that in the second data sets all possible subsets of these three activities where skipped.

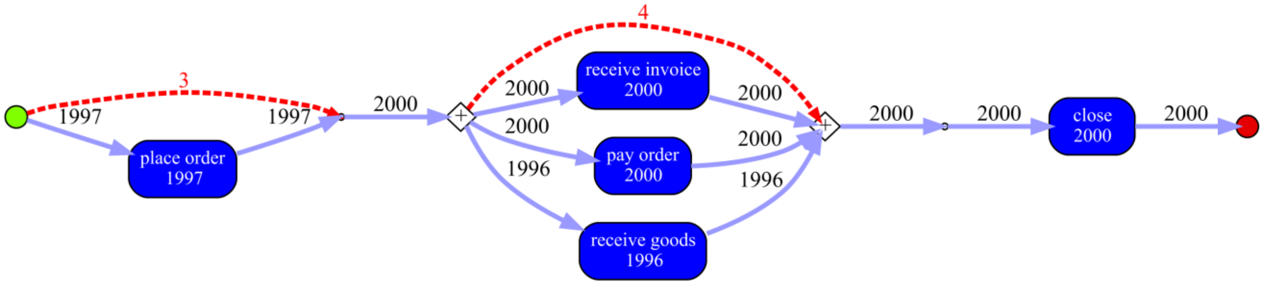


Figure 11: Conformance checking applied to an event log with seven deviating cases: Three cases skipped activity place order and four cases skipped receive goods.

To illustrate the kind of diagnostics one would expect, we refer to Figure 11. These are diagnostics returned by ProM given an event log where we modified seven cases in such a way that all seven are non-compliant. The red arcs show the deviations. Activity *place order* is skipped three times. Activity *receive goods* is skipped four times. Using ProM one can easily drill down on the deviations. The red arcs separate the deviations from the model and one can select any arc to see the corresponding non-conforming procurement orders. Figure 12 shows the normative BPMN model next two the DFGs generated by Disco and Celonis. Based on these DFGs it remains partly unclear what the deviations are. It is possible to identify the three cases where activity *place order* is skipped. However, based on the arcs it is impossible to see that activity *receive goods* was skipped four times. The numbers on the arcs are not revealing this.

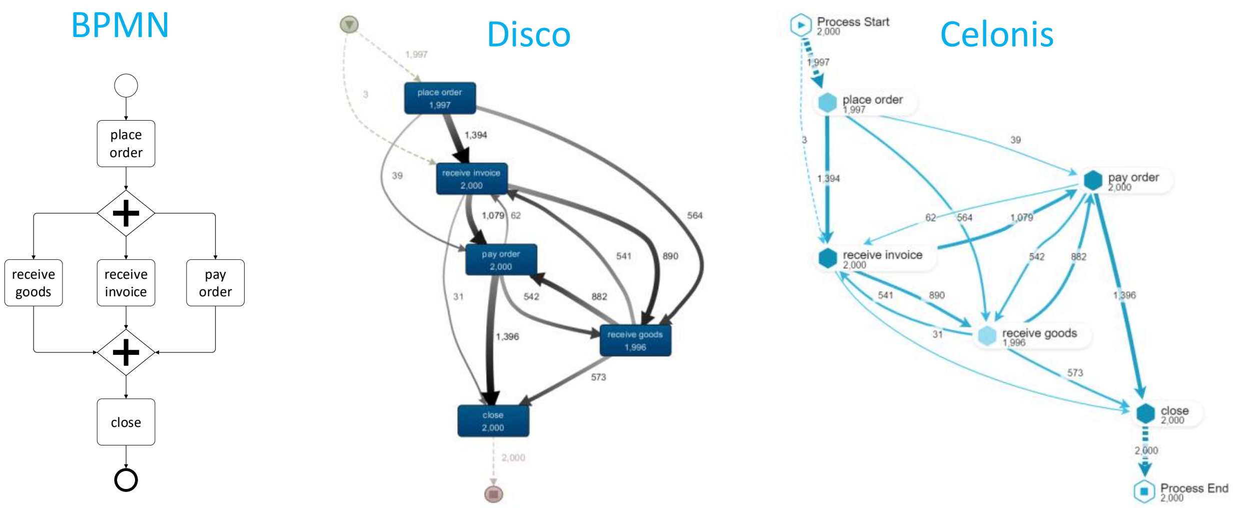


Figure 12: The normative BPMN model and the DFGs generated by Disco and Celonis for the event log with seven deviating cases.

Some of the commercial software tools have added conformance checking capabilities in recent years, e.g., Celonis supports a variant of token-based replay. However, the usability, quality of diagnostics, and scalability of existing software tools leave much to be desired.

Also from a scientific point of view, process discovery and conformance checking are still challenging. These two foundational process mining problems have not yet been solved satisfactorily and there is still a lot of room for improvement [2]. However, the current state-of-the-art techniques provide already partial solutions that perform well on real-life data sets. The hope and expectation is that commercial systems will adopt these techniques when users get more critical and expect more precise and fully correct diagnostics.

Novel challenges

As discussed, process discovery and conformance checking are still challenging and one can expect further improvements in the coming years. However, next to these core process mining tasks, novel process mining capabilities have been identified (see Figure 13). These provide new scientific and practical challenges. Here we briefly mention a few.

- ***Challenge: Bridging the gap between process modeling and process mining.*** Many organizations use tools for modeling processes. With the uptake of process mining, it becomes clear that these models do not correspond to reality. Although such idealized models are valuable, the gap between discovered and hand-made models needs to be bridged [9]. A promising approach is the use of so-called hybrid process models that have a backbone formally describing the parts of the process that are clear and stable, and less rigid data-driven annotations to show the things that are less clear. Hybrid process models allow for formal reasoning, but also reveal information that cannot be captured using mainstream formal models because the behavior is too complex or there is not enough data to justify a firm conclusion [9]. Next to combining formal modeling constructs (precisely describing the possible behaviors) and informal modeling constructs (data-based annotations not allowing for any form of formal reasoning), there is also the need to deal with multiple abstraction levels. Modeled processes tend to be at a higher level of abstraction than discovered process models. One high-level activity can correspond to many low-level events at different levels of granularity. It is not easy to bridge the gap between process modeling and process mining. However, both need to be integrated and supported in a seamless manner. A convergence of process modeling and process mining tools is needed and also inevitable.

- ***Challenge: Incorporating stochastic information in process models to improve conformance checking and prediction.*** Frequencies of activities and process variants are essential for process mining. A highly frequent process variant (i.e., many cases having the same trace) should be incorporated in the corresponding process model. This is less important for a process variant that occurs only once. One can view frequencies as estimates for probabilities, e.g., if 150 cases end with activity *reject* and 50 cases end with activity *accept*, then the data suggests that there is a 75% probability of rejection and a 25% probability of acceptance. Such information is typically not incorporated in process models. For example, the normative BPMN model may have a gateway modeling the choice between activity *reject* and activity *accept*, but typically the probability is not indicated. Obviously, such information is essential for predictive process mining. However, the same information is vital for conformance checking. Typically, four conformance dimensions are identified: (1) *recall*: the discovered model should allow for the behavior seen in the event log (avoiding "non-fitting" behavior), (2) *precision*: the discovered model should not allow for behavior completely unrelated to what was seen in the event log (avoiding "underfitting"), (3) *generalization*: the discovered model should generalize the example behavior seen in the event log (avoiding "overfitting"), and (4) *simplicity*: the discovered model should not be unnecessarily complex. The first two are most relevant for comparing observed and modeled behavior (the other two relate more to the completeness and redundancy of event data and the understandability of the process model). Recall (often called fitness) is typically well-understood. Although there are many precision notions, it turns out to be problematic to define precision for process models without probabilities. Adding infrequent behavior to a model may significantly lower traditional precision notions. This seems counterintuitive. Moreover, from a practical point of view, a process may be no longer be compliant if the distribution over the various paths in the model dramatically changes. These observations suggest that probabilities need to be added to process models to allow for both prediction and all forms of comparison (including conformance checking). Moreover, adding stochastics to process models also enables the interplay between process mining and simulation [10]. Currently, simulation is rarely used for process management. However, the uptake of process mining may lead to a revival of business process simulation.
- ***Challenge: Process mining for multiple processes using different case notions.*** Traditional process mining techniques assume that each event refers to one case and that each case refers to one process. In reality, this is more complex [5]. There may be different intertwined processes and one event may be related to different cases (convergence) and, for a given case, there may be multiple instances of the same activity within a case (divergence). To create a traditional process model, the event data need to be "flattened". There are typically multiple choices possible, leading to different views and process models that are disconnected. Consider the classical example where the information system holds information about customer orders, products, payments, packages, and deliveries scattered over multiple tables. *Object-centric*

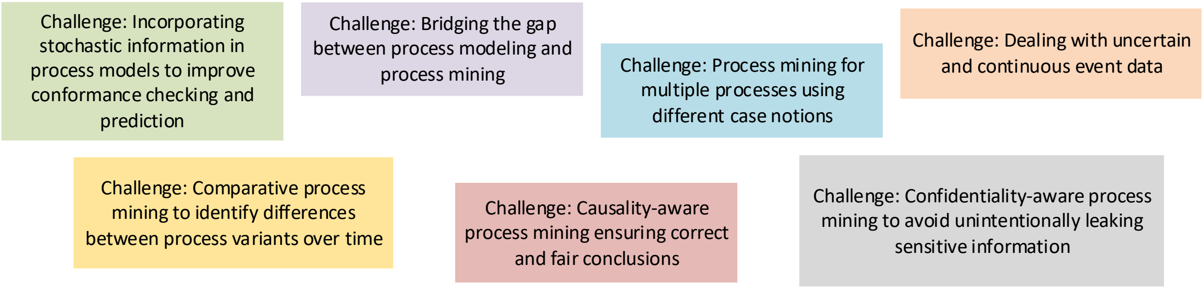
process mining relaxes the traditional assumption that each event refers to one case [5]. An event may refer to any number of business objects and using novel process mining techniques one can discover one integrated process model showing *multiple perspectives*.

- ***Challenge: Dealing with uncertain and continuous event data.*** The starting point for any process mining effort is a collection of events. Normally, we assume that events are discrete and certain, i.e., we assume that each event reported has actually happened and that its attributes are correct. However, due the expanding scope of process mining other types of event data are encountered. Events may be uncertain, e.g., the time may not be known exactly, the activity is not certain, and there may be multiple candidate case identifiers. Consider for example sensor data that needs to be discretized or text data that needs to be preprocessed. In such settings we use classifiers that have a maximal accuracy of for example 80%. By lowering thresholds we may get more "false positives" (e.g., an activity that did not really happen was added to the event log) and by increasing thresholds we get more "false negatives". The attributes of events may also have continuous variables that determine the actual meaning of the event. For example, a blood test may provide several measurements relevant to the treatment process. Another example is the event log of a wind turbine showing information about wind speeds, wind direction, voltage, etc. Such information can be used to derive an event "shut down turbine because of strong winds". Future process mining tools and approaches will need to be able to deal better with uncertain event data and measurements that are continuous in nature.
- ***Challenge: Comparative process mining to identify differences between process variants over time.*** Processes change over time and the same process may be performed at different locations. Hence, it is valuable to compare the corresponding process variants [11]. The relative performance often provides more insights than the absolute values. For example, what were the main differences between January and February or what are the differences between the Berlin office and the Amsterdam office? Comparing different process variants is not so easy. Compare for example the DFGs in Figure 5 to the DFGs in Figure 8. It is not so easy to spot relevant differences. Next to changes in the process structure, also frequencies and times may change. Techniques for comparative process mining aim to address this [11]. This includes techniques to support the visual comparison of process variants and machine learning techniques using process-centric features [12].
- ***Challenge: Causality-aware process mining ensuring correct and fair conclusions.*** Process mining techniques can be used to quickly uncover performance and compliance problems. Based on bottleneck analysis and conformance checking results, we can annotate event data to expose desirable and undesirable "situations"(i.e., good and bad choices, cases, routes, etc.). Moreover, using feature extraction it is possible to turn such situations into supervised learning problems and use data mining and machine learning techniques to uncover root causes for performance and compliance problems. Such a combination of techniques yields a powerful approach to automatically diagnose process-related problems. However, correlation does not imply causation. Unfortunately, these terms that are mostly misunderstood and often used interchangeably. For example, ice cream sales may strongly correlate with burglary. However, this does not imply that eating ice cream causes theft. There is third variable (the weather) that is influencing both ice cream sales and burglary. When delays in a process correlate with deviations, then this does not imply that one causes the other. Therefore, explicit causal models are required to guide root-cause analysis in process mining. In a causal model relations between processes features can be

made explicit using a mixture of domain knowledge and statistical evidence. Similar techniques can be used to avoid unfair or even discriminating conclusions. For example, it is pointless to blame the workers that are overloaded for delays. Also the most experienced workers often take the most difficult cases possibly leading to unfair conclusions if one only considers the data.

- ***Challenge: Confidentiality-aware process mining to avoid unintentionally leaking sensitive information.*** Event data are potentially very sensitive. A few timestamped events are often enough to identify a customer or employee. Even when one removes explicit timestamps, the order of activities may already be enough for identification. Preserving confidentiality is, therefore, a primary concern in process mining. Current research aims at dedicated anonymization and encryption techniques. For example, one does not need to store complete cases to generate DFGs. It is sufficient to store direct succession relations without correlating all events belonging to a case. The introduction of the EU General Data Protection Regulation (GDPR) illustrates the growing importance of data privacy. Therefore, the next generation of process mining tools will need to support confidentiality preserving techniques. Confidentiality-aware process mining is part of the broader domain of *Responsible Data Science* (RDS) focusing on *fairness*, *accuracy*, *confidentiality*, and *transparency*. All four RDS aspects are relevant for process mining. Not addressing these concerns may slow down the adoption of process mining.

The above list of challenges is far from complete. Process mining is a relatively young, but also broad, discipline. It is interesting to compare the above list with the eleven challenges in the Process Mining Manifesto written in 2011 [13]. This shows that the field developed rapidly.



Challenge: Incorporating stochastic information in process models to improve conformance checking and prediction

Challenge: Bridging the gap between process modeling and process mining

Challenge: Process mining for multiple processes using different case notions

Challenge: Dealing with uncertain and continuous event data

Challenge: Comparative process mining to identify differences between process variants over time

Challenge: Causality-aware process mining ensuring correct and fair conclusions

Challenge: Confidentiality-aware process mining to avoid unintentionally leaking sensitive information

Figure 13: Some of the challenges getting more attention in research and thus showing the anticipated development of the process mining discipline.

Process hygiene

Most of the challenges mentioned in this chapter require the concerted action of process mining users, technology providers, and scientists. A collaborative effort is needed to make process mining "the new normal". Process mining should be as normal as personal hygiene and not require a business case. Activities such as brushing your teeth, washing your hands after going to the toilet, and changing clothes do not require a business case. Process mining can be seen as the means to ensure *Process Hygiene* (PH) or *Business Process Hygiene* (BPH). Objectively monitoring and analyzing key processes is important for the overall health and well-being of an organization. Unfortunately, managers, auditors, and accountants often still use medieval practices. Financial reporting frameworks such as the nation-specific GAAP (Generally Accepted Accounting Principles) standards still depend on the notion of materiality. As a result sampling suffices. Given the availability of data and our ability to analyze processes this is remarkable. Process Hygiene (PH) should not require a business case. Not using process mining is a sign of self-neglect showing an inability or unwillingness to manage processes properly. Hence, *not* using process mining should require a justification and not the other way around. Using data quality and privacy concerns as reasons to not conduct process mining should be considered as poor hygiene leading to "smelly processes".

References

1. Aalst, W.M.P.v.d., *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. 2011: Springer-Verlag, Berlin.
2. Aalst, W.M.P.v.d., *Process Mining: Data Science in Action*. 2016: Springer-Verlag, Berlin.
3. Kerremans, M., *Gartner Market Guide for Process Mining, Research Note G00387812*. 2019.
4. Kerremans, M., *Gartner Market Guide for Process Mining, Research Note G00353970*. 2018.
5. Aalst, W.M.P.v.d. *Object-Centric Process Mining: Dealing With Divergence and Convergence in Event Data*. in *Software Engineering and Formal Methods (SEFM 2019)*. 2019. Springer-Verlag, Berlin.
6. Koplowitz, R., et al., *Process Mining: Your Compass For Digital Transformation: The Customer Journey Is The Destination*. 2019.
7. Rozinat, A. and W.M.P.v.d. Aalst, *Conformance Checking of Processes Based on Monitoring Real Behavior.* Information Systems, 2008. **33**(1): p. 64-95.
8. Carmona, J., et al., *Conformance Checking: Relating Processes and Models*. 2018: Springer-Verlag, Berlin.
9. Aalst, W.M.P.v.d., et al. *Learning Hybrid Process Models From Events: Process Discovery Without Faking Confidence*. in *International Conference on Business Process Management (BPM 2017)*. 2017. Springer-Verlag, Berlin.
10. Aalst, W.M.P.v.d. *Process Mining and Simulation: A Match Made in Heaven!* in *Computer Simulation Conference (SummerSim 2018)*. 2018. ACM Press.
11. Aalst, W.M.P.v.d. *Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining*. in *Asia Pacific Conference on Business Process Management (AP-BPM 2013)*. 2013. Springer-Verlag, Berlin.
12. Bolt, A., M. de Leoni, and W.M.P.v.d. Aalst, *Process Variant Comparison: Using Event Logs to Detect Differences in Behavior and Business Rules.* Information Systems, 2018. **74**(1): p. 53-66.
13. IEEE Task Force on Process Mining. *Process Mining Manifesto*. in *Business Process Management Workshops*. 2012. Springer-Verlag, Berlin.