

Unleashing Operational Process Mining

Edited by

Rafael Accorsi¹, Ernesto Damiani², and Wil van der Aalst³

1 Universität Freiburg, DE, rafael.accorsi@iig.uni-freiburg.de

2 Università degli Studi di Milano – Crema, IT, ernesto.damiani@unimi.it

3 Eindhoven University of Technology, NL, w.m.p.v.d.aalst@tue.nl

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13481 “Unleashing Operational Process Mining”. Process mining is a young research discipline connecting computational intelligence and data mining on the one hand and process modeling and analysis on the other hand. The goal of process mining is to discover, monitor, diagnose and improve real processes by extracting knowledge from event logs readily available in today’s information systems. Process mining bridges the gap between data mining and business process modeling and analysis. The seminar that took place November 2013 was the first in its kind. About 50 process mining experts joined forces to discuss the main process mining challenges and present cutting edge results. This report aims to describe the presentations, discussions, and findings.

Seminar 25.–29. November, 2013 – www.dagstuhl.de/13481

1998 ACM Subject Classification H.2.8 Database Applications (Data mining), H.2.1 Logical Design (Data models), K.6 Management of Computing and Information Systems

Keywords and phrases Process mining, Big data, Conformance checking

Digital Object Identifier 10.4230/DagRep.3.11.154

1 Executive Summary

Rafael Accorsi

Ernesto Damiani

Wil van der Aalst

License © Creative Commons BY 3.0 Unported license
© Rafael Accorsi, Ernesto Damiani, and Wil van der Aalst

Society shifted from being predominantly “analog” to “digital” in just a few years. This has had an incredible impact on the way we do business and communicate. Gartner uses the phrase “The Nexus of Forces” to refer to the convergence and mutual reinforcement of four interdependent trends: social, mobile, cloud, and information. The term “Big Data” is often used to refer to the incredible growth of data in recent years. However, the ultimate goal is not to collect more data, but to turn data into real value. This means that data should be used to improve existing products, processes and services, or enable new ones.

Event data are the most important source of information. Events may take place inside a machine (e.g., an X-ray machine or baggage handling system), inside an enterprise information system (e.g., an order placed by a customer), inside a hospital (e.g., the analysis of a blood sample), inside a social network (e.g., exchanging e-mails or twitter messages), inside a transportation system (e.g., checking in, buying a ticket, or passing through a toll booth), etc.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

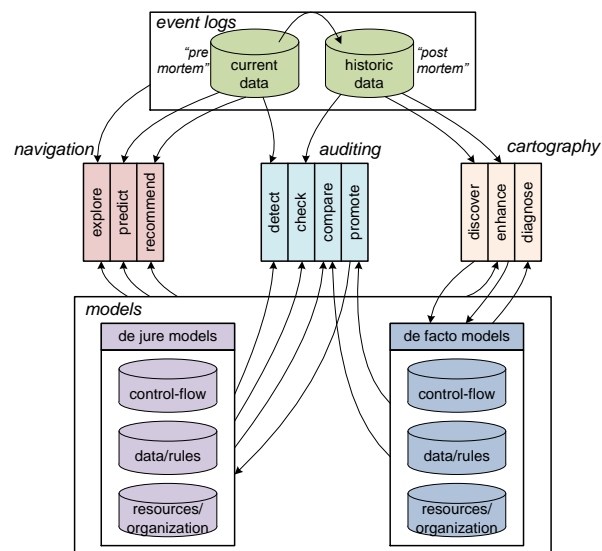
Unleashing Operational Process Mining, *Dagstuhl Reports*, Vol. 3, Issue 11, pp. 154–192

Editors: Rafael Accorsi, Ernesto Damiani, and Wil van der Aalst



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Overview of the different process mining tasks (taken from “Process Mining: Discovery, Conformance and Enhancement of Business Processes”).

Process mining aims to *discover, monitor and improve real processes by extracting knowledge from event logs* readily available in today’s information systems¹. 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 can be used to conduct three types of process mining. 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. 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 thereby 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.

Process mining algorithms have been implemented in various academic and commercial

¹ Process Mining: Discovery, Conformance and Enhancement of Business Processes by W. M. P. van der Aalst, Springer Verlag, 2011 (ISBN 978-3-642-19344-6).

systems. The corresponding tools are being increasingly relevant in industry and have proven to be essential means to meet business goals. ProM is the de facto standard platform for process mining in the academic world. Examples of commercial tools are Disco (Fluxicon), Perceptive Process Mining (before Futura Reflect and BPM|one), QPR ProcessAnalyzer, ARIS Process Performance Manager, Celonis Discovery, Interstage Process Discovery (Fujitsu), Discovery Analyst (StereoLOGIC), and XMAalyzer (XMPro). Representatives of ProM community and the first three commercial vendors participated in Dagstuhl Seminar 13481 “Unleashing Operational Process Mining”.

The Dagstuhl Seminar was co-organized with the *IEEE Task Force on Process Mining* (see <http://www.win.tue.nl/ieeetfpm/>). The goal of this Task Force is to promote the research, development, education and understanding of process mining. Sixty organizations and over one hundred experts have joined forces in the IEEE Task Force on Process Mining.

Next to some introductory talks (e.g., an overview of the process mining field by Wil van der Aalst), 31 talks were given by the participants. The talks covered the entire process mining spectrum, including:

- from theory to applications,
- from methodological to tool-oriented,
- from data quality to new analysis techniques,
- from big data to semi-structured data,
- from discovery to conformance,
- from health-care to security, and
- from off-line to online.

The abstracts of all talks are included in this report.

It was remarkable to see that all participants (including the academics) were very motivated to solve real-life problems and considered increasing the adoption of process mining as one of the key priorities, thereby justifying the title and spirit of the seminar, namely “Unleashing the Power of Process Mining”. This does not imply that there are not many foundational research challenges. For example, the increasing amounts of event data are creating many new challenges and new questions have emerged. Such issues were discussed both during the sessions and on informal meetings during the breaks and at the evening.

Half of the program was devoted to discussions on a set of predefined themes. These topics were extracted based on a questionnaire filled out by all participants before the seminar.

1. Process mining of multi-perspective models (Chair: Akhil Kumar)
2. Data quality and data preparation (Chair: Frank van Geffen)
3. Process discovery: Playing with the representational bias (Josep Carmona)
4. Evaluation of process mining algorithms: benchmark data sets and conformance metrics (Chair: Boudewijn van Dongen)
5. Advanced topics in process discovery: on-the-fly and distributed process discovery (Chair: Alessandro Sperduti)
6. Process mining and Big Data (Chair: Marcello Leida)
7. Process mining in Healthcare (Chair: Pnina Soffer)
8. Security and privacy issues in large process data sets (Chair: Simon Foley, replacing Günter Müller)
9. Conformance checking for security, compliance and auditing (Chair: Massimiliano De Leoni, replacing Marco Montali)
10. How to sell process mining? (Chair: Anne Rozinat)

11. What is the ideal tool for an expert user? (Chair: Benoit Depaire)
12. What is the ideal tool for a casual business user? (Chair: Teemu Lehto)

Summaries of all discussions are included in this report. The chairs did an excellent job in guiding the discussions. After the each discussion participants had a better understanding of the challenges that process mining is facing. This definitely include many research challenges, but also challenges related to boosting the adoption of process mining in industry.

The social program was rich and vivid, including an excursion to Trier's Christmas market, a night walk to ruins, table football, table tennis, and late night discussions.

Next to this report, a tangible output of the seminar is a special issue of IEEE Transactions on Services Computing based on the seminar. This special issue has the title "Processes Meet Big Data" and will be based on contributions from participants of this seminar (also open to others). This special issue of IEEE Transaction on Service-Oriented Computing is intended to create an international forum for presenting innovative developments of process monitoring, analysis and mining over service-oriented architectures, aimed at handling "big logs" and use them effectively for discovery, dash-boarding and mining. The ultimate objective is to identify the promising research avenues, report the main results and promote the visibility and relevance of this area.

Overall, the seminar was very successful. Most participants encouraged the organizers to organize another Dagstuhl Seminar on process mining. Several suggestions were given for such a future seminar, e.g., providing event logs for competitions and complementary types of analysis before or during the seminar. These recommendations were subject of the discussion sessions, whose summaries can be found below.

2 Table of Contents

Executive Summary

<i>Rafael Accorsi, Ernesto Damiani, and Wil van der Aalst</i>	154
---	-----

Overview of Talks

Knowledge and Business Intelligence Technologies in Cross-Enterprise Environments (KITE.it) <i>Antonio Caforio</i>	161
To Unleash, or not to Unleash, that is the Question! <i>Josep Carmona</i>	161
Mining Collaboration in Business Process <i>Paolo Ceravolo</i>	161
Agile Process Mining? <i>Jonathan Cook</i>	162
Sesar Lab Activities <i>Ernesto Damiani</i>	163
Building Better Simulation Models with Process Mining <i>Benoit Depaire</i>	163
Mining the Unknown Security Frontier from System Logs <i>Simon N. Foley</i>	163
Beyond Tasks and Gateways: Towards Rich BPMN Process Mining <i>Luciano Garcia-Banuelos</i>	164
Working with BPMN in ProM <i>Anna A. Kalenkova</i>	164
Benchmarking Process Mining Algorithms on Noisy Data: Does Log Sanitization Help? <i>Akhil Kumar</i>	165
Analytics for Case Management and other Semi-Structured Environments <i>Geetika T. Lakshmanan</i>	165
Continuous Data-driven Business Process Improvement for SAP Order To Cash process <i>Teemu Lehto</i>	165
QPR ProcessAnalyzer (Tool Demo) <i>Teemu Lehto</i>	166
Big Data Techniques for Process Monitoring <i>Marcello Leida</i>	166
Process Mining and BigData <i>Marcello Leida</i>	166
On the Suitability of Process Mining to Produce Current-State Role-based Access Control Models <i>Maria Leitner</i>	167

Declarative Process Mining with ProM <i>Fabrizio Maria Maggi</i>	167
Scientific Workflows within the Process Mining Domain <i>Ronny S. Mans</i>	168
Conformance Analysis of Inventory Processes using Process Mining <i>Zbigniew Paszkiewicz</i>	168
Predictive Security Analysis@Runtime – Lessons Learnt from Adaptation to Industrial Scenarios <i>Roland Rieke</i>	169
Data Collection, Integration, and Cleaning for Process Mining: Reflections on Some Projects <i>Stefanie Rinderle-Ma</i>	169
Disco (Tool Demo) <i>Anne Rozinat</i>	170
DPMine/P Complex Experiment Model Markup Language as Applied to ProM <i>Sergey A. Shershakov</i>	170
A Process Mining-based Analysis of Intentional Noncompliance <i>Pnina Soffer</i>	171
Outpatient Process Analysis with Process Mining: A Case Study <i>Minseok Song</i>	171
PROMPT: Process Mining for Business Process Improvement <i>Alessandro Sperduti</i>	171
SecSy: Security-aware Synthesis of Process Event Logs <i>Thomas Stocker</i>	172
ProM 6.3 (Tool Demo) <i>Eric Verbeek</i>	172
Process Mining in China - Recent Work on Event Quality <i>Jianmin Wang</i>	172
Turning Event Logs into Process Movies: Animating What Has Really Happened <i>Massimiliano de Leoni</i>	173
Introducing Process Mining at Rabobank <i>Frank van Geffen</i>	173
Discussion Sessions	
Discussion Session 1: Process Mining of Multi-Perspective Models	174
Discussion Session 2: Data Quality and Data Preparation	175
Discussion Session 3: Process Discovery: Playing with the Representational Bias	177
Discussion Session 4: Evaluation of Process Mining Algorithms: Benchmark Data Sets and Conformance Metrics	178
Discussion Session 5: Advanced Topics in Process Discovery: On-the-fly and Distributed Process Discovery	179
Discussion Session 6: Process Mining and Big Data	181

160 13481 – Unleashing Operational Process Mining

Discussion Session 7: Process Mining in Health-care 183

Discussion Session 8: Security and Privacy Issues in Large Process Data Sets . . . 184

Discussion Session 9: Conformance Checking for Security, Compliance and Auditing 185

Discussion Session 10: How to Sell Process Mining? 187

Discussion Session 11: What is the Ideal Tool for an Expert User? 188

Discussion Session 12: What is the Ideal Tool for a Casual Business User? 189

Participants 192

3 Overview of Talks

3.1 Knowledge and Business Intelligence Technologies in Cross-Enterprise Environments (KITE.it)

Antonio Caforio (University of Salento, IT)

License © Creative Commons BY 3.0 Unported license
© Antonio Caforio

Main reference L. Fischer (Ed.), “Delivering Competitive Advantage through BPM – Real-World Business Process Management,” Excellence in Practice Series, ISBN 978-0-9849764-5-4, Future Strategies Inc., 2012.

URL <http://futstrat.com/books/CompetitiveAdvantage.php>

In this talk Antonio Caforio described the work ongoing in an Italian industrial research project, Knowledge and business Intelligence Technologies in cross-enterprise Environments (KITE.it), that aims to support the creation and management of processes in the Value Networks (VN). The main project outcomes are methodologies and platforms to enable the alignment of processes with the organizations’ goals in the VN and the measurement of the VN effectiveness. A focus will be made on the Aeroengine Maintenance, Repair and Overhaul (MRO) and its main Overhaul process to understand how process mining can help improve the management of this process.

3.2 To Unleash, or not to Unleash, that is the Question!

Josep Carmona (UPC – Barcelona, ES)

License © Creative Commons BY 3.0 Unported license
© Josep Carmona

In this talk Josep Carmona introduced two very different approaches for unleashing (or not) process mining, that are being developed in my group. The first one, based on the use of portfolio-based algorithm selection techniques, is devoted to guide the application of process mining algorithms by using a recommender system. The second one, totally opposed to the first, aims at providing a process-oriented computing environment for the exploration and creation of process mining algorithms. These two approaches are meant to cover a wide variety of process mining practices and, together with existing frameworks, offer a new perspective to the field.

3.3 Mining Collaboration in Business Process

Paolo Ceravolo (University of Milan, IT)

License © Creative Commons BY 3.0 Unported license
© Paolo Ceravolo

Main reference F. Frati, I. Seeber, “CoPrA: A Tool For Coding and Measuring Communication In Teams,” in Proc. of the 7th IEEE Int’l Conf. on Digital Ecosystems and Technologies (DEST’13), pp. 43–48, IEEE, 2013.


URL <http://dx.doi.org/10.1109/DEST.2013.6611327>

Observing the evolution of several research programs focusing on collaborating communities, we encounter a call for diachronic analysis. It follows that Process Mining can contribute in refining and enriching the next generation of these studies. In whatever way, this implies to understand the research questions that are driving the analysis on collaborative process

to then identify the challenges for evolving Process Mining techniques. In this talk, Paolo Ceravolo looked back on the evolution of some area related to collaborative process and pointing out open issue and interesting research directions.

3.4 Agile Process Mining?

Jonathan Cook (New Mexico State University, US)

License  Creative Commons BY 3.0 Unported license
© Jonathan Cook

Joint work of Cook, Jonathan; Bani-Hani, Imad

Agile software development methodologies are in some sense a reaction to overly prescriptive development processes, and to a large extent desired to throw out a strict process model and move the team collaboration and project management off the computer and make it human centric. Thus we see management tools such as white boards with sticky notes being used to manage the project. Can such a process be mined effectively? We argue yes, because agile methodologies still embody practices that should result in some regular, observable patterns of behavior or at least constraints over what activities take place when. For some examples, test driven design should show itself in the activity of creating a test case (or more) before a feature is implemented; time-boxed iteration should show itself in very regular release tagging; continuous integration should show itself in regular feature merges into the main build; and refactoring should show itself in particular code edit patterns or commit messages. Knowing the process (or practices) that a project team is performing will help them in assessing their own agility and potentially show them areas where they could improve. Beyond the set of recommended practices, agile processes should be flexible; thus very closed-form control-flow process mining algorithms should probably not work well on an agile process (and if they do, it may not be very agile). Open model mining algorithms such as DeclareMiner, which infers particular LTL rule patterns, should be much more suitable for agile process mining. Other aspects of process mining, such as role mining, organizational structure mining, and social network mining may help in agile process mining; for example, an agile team may have a goal to be as interactive and collaborative as possible, sharing duties equally, but in practice they may slip into very specific roles without realizing it. Finally, before pursuing agile process mining, we are creating qualitative methods (e.g., a questionnaire) for measuring the agility of a process, and have defined five dimensions over which to measure agility: team (level of interaction and collaboration practices); customer (level of customer involvement); iteration (level of true iterative practices); testing (level of test-centric practices); and design (level of ongoing refactoring effort). In this talk, Jon Cook asked 3–6 questions in each area and then create a radial chart showing the level of agility for each dimension and giving a visualization of overall agility.

3.5 Sesar Lab Activities

Ernesto Damiani (Università degli Studi di Milano – Crema, IT)

License  Creative Commons BY 3.0 Unported license
© Ernesto Damiani

Secure Service-oriented Architectures Research (SESAR) Lab within the Computer Science Department of the Università degli Studi di Milano. The research activities are mainly focused on the following subjects ranging from Service-oriented Architectures to Knowledge Management over Open Source Development Paradigms and Security. The staff is composed by full time Professors, Researchers, Post-Docs, PhDs and Research Collaborators. The research activities are carried out in collaboration with Italian and European partners, within national and international research projects and agreements with enterprises. The Lab offers to University students the chance to carry out degree theses, stages for the acknowledgment of university credits, and the opportunity to participate to the activities carried out in each research project, achieving experiences for the future work activities.

3.6 Building Better Simulation Models with Process Mining

Benoit Depaire (Hasselt University – Diepenbeek, BE)

License  Creative Commons BY 3.0 Unported license
© Benoit Depaire

Joint work of Depaire, Benoit; Martin, Niels; Caris, An

Simulation models are a useful tool to run complex what-if scenarios and to make informed decisions. However, these simulation models are often constructed in a highly subjective way (through interviews, documents how processes should be executed and guesstimates on simulation parameters). In this talk Benoit Depaire argued that process mining holds the tools and potential to construct more reliable simulation models. We presented a SWOT analysis of business process simulation based on the current state of the art in literature, presented a framework how simulation and process mining could be linked together and identified different challenges where the process mining community should focus on.

3.7 Mining the Unknown Security Frontier from System Logs

Simon N. Foley (University College Cork, IE)

License  Creative Commons BY 3.0 Unported license
© Simon N. Foley

Joint work of Foley, Simon N.; Pieczul, Olgierd

The scale and complexity of modern computer systems has meant that it is becoming increasingly difficult and expensive to formulate effective security policies and to deploy efficacious security controls. As a consequence, security compliance tends to focus on those activities perceived to be critical, with an assumption that the other activities, known or unknown, are not significant. However, often it is these side-activities that can lead to a security compromise of the system. While security controls provide monitoring and enforcement of the critical activities related to the security policy, effectively, little is known about the nature of the other activities. Our preliminary results show that such activities can

be modeled and do exist in real-world systems. In this talk, Simon Foley demonstrated how process mining techniques can be explored in practice to discover and check for perturbations to these activities in system logs.

3.8 Beyond Tasks and Gateways: Towards Rich BPMN Process Mining


Luciano Garcia-Banuelos (University of Tartu, EE)

License  Creative Commons BY 3.0 Unported license
© Luciano Garcia-Banuelos

During the last decade, process mining techniques have reached a certain level of sophistication and maturity, evidenced by the availability of a range of functional academic prototypes and commercial tools in the field. In parallel to these developments, BPMN has emerged as a widely adopted standard for modeling and analyzing business processes. BPMN offers a rich set of constructs for modeling business processes in a structured way, including sub-processes with interrupting and non-interrupting boundary events and multi-instance activities as well as a comprehensive set of event types. Surprisingly though, the bulk of research in process mining in general, and automated process model discovery in particular, has focused on the problem of discovering process models consisting purely of tasks and control-flow dependencies (in essence: tasks and some types of gateways). In this talk, Luciano Garcia-Banuelos presented his initial work on automated discovery of rich BPMN process models, meaning process models that make use of the BPMN notation beyond its “task and gateways” subset. He discussed initial achievements and key challenges he had identified so far.

3.9 Working with BPMN in ProM


Anna A. Kalenkova (NRU Higher School of Economics – Moscow, RU)

License  Creative Commons BY 3.0 Unported license
© Anna A. Kalenkova

ProM is a tool for implementing and integrating process mining algorithms within a standard environment. ProM plugins support plenty of different process model formats, among them are Petri nets, transition systems, casual nets, fuzzy models and others which are widely used by researchers. But at the same time it might be rather difficult for an inexperienced user (or for an external customer) to estimate the result of applying process mining techniques and understand the semantics of process models. This indicates that there is a need for ProM to support commonly known process modeling standards also. BPMN (Business Process Modeling Notation) is a process modeling and executing notation understandable by a wide audience of analytics and developers. Representing process models in this standard way will give an ability to bridge the gap between ProM and variety of process modeling tools. Also BPMN gives a holistic view on the process model: BPMN diagrams could be enhanced with roles, interactions, timers, conformance/performance info, etc. In her talk, Anna Kalenkova gave an overview of ProM functionality related to BPMN. Import/export capabilities and internal BPMN meta-model were discussed. Also the plugins which implement conversions from different formalisms to BPMN and vice-versa were considered.

3.10 Benchmarking Process Mining Algorithms on Noisy Data: Does Log Sanitization Help?


Akhil Kumar (Pennsylvania State University, US)

License  Creative Commons BY 3.0 Unported license
© Akhil Kumar

Akhil Kumar proposed a technique to sanitize noisy logs by first building a classifier on a subset of the log and applying the classifier rules to remove noisy traces from the log. The technique is evaluated on synthetic logs from six benchmark models of increasing complexity on both behavioral and structural recall and precision metrics. The results show that mined models thus produced from sanitized log are superior on the evaluation metrics. They show better fidelity to the reference models and are more compact. The rule based approach generalizes to any noise pattern. The rules can be explained and modified.

3.11 Analytics for Case Management and other Semi-Structured Environments

Geetika T. Lakshmanan (IBM TJ Watson Research Center – Cambridge, US)

License  Creative Commons BY 3.0 Unported license
© Geetika T. Lakshmanan

There is considerable scope for both predictive and descriptive analytics in case management and other semi-structured environments. Predictive analytics could provide guidance to a case worker handling a current case instance on the likelihood of a future task occurrence or attribute value. By training a classifier such as a decision tree on a set of completed case execution traces, the classifier can be used to make predictions about the likelihood of occurrence of a task execution or predict the value of a continuous variable in the case such as time for a currently running case instance. Descriptive analytics could be applied to provide insight about correlations and patterns derived from historically completed instances of a case. In order to be applied in a real world setting, these analytics require solving an array of challenges. In addition to providing easily consumable results, these analytics have to be highly confident of the predictions and correlations they compute. In her talk, Geetika Lakshmanan provided an overview of the challenges of applying predictive and descriptive analytics to case management and other real world settings.

3.12 Continuous Data-driven Business Process Improvement for SAP Order To Cash process


Teemu Lehto (QPR Software – Helsinki, FI)

License  Creative Commons BY 3.0 Unported license
© Teemu Lehto

This talk was a business-driven case study for sharing experiences (1) SAP is the leading ERP system globally measured by revenue. (2) SAP creates great quality records for process mining purposes. (3) Order to cash is a critical importance business process for organizations. (4) Order to cash is not as systematic or optimized as one could think of.

3.13 QPR ProcessAnalyzer (Tool Demo)

Teemu Lehto (QPR Software – Helsinki, FI)

License  Creative Commons BY 3.0 Unported license
© Teemu Lehto

This talk is a demo of a commercial Process Mining tool QPR ProcessAnalyzer.

3.14 Big Data Techniques for Process Monitoring

Marcello Leida (Khalifa University – Abu Dhabi, AE)

License  Creative Commons BY 3.0 Unported license
© Marcello Leida

Joint work of Leida, Marcello, Andrej Chu, Basim Majeed

Main reference M. Leida, A. Chu, “Distributed SPARQL Query Answering over RDF Data Streams,” in Proc. of 2013 IEEE Int’l Congress on Big Data (BigData’13), pp. 369–378, IEEE, 2013.

URL <http://dx.doi.org/10.1109/BigData.Congress.2013.56>

Modern Business process analysis requires an extremely flexible data model and a platform able to minimize response times as much as possible. In order to efficiently analyze a large amount of data, this talk illustrated novel technologies that rely on an improved data model supported by a grid infrastructure, allowing storing the data in-memory across many grid nodes and distributing the workload, avoiding the bottleneck represented by constantly querying a traditional database. Both process data and domain knowledge are represented using standard metadata formats: process logs are stored as RDF triples referring to company specific activities. The data collected by the process log monitor is translated to a continuous flow of triples that capture the status of the processes. This continuous flow of information can be accessed through the SPARQL query language used to extract and analyze process execution data. Although the query engine has been developed as part of a Business Process Monitoring platform, it is a general purpose engine that can be used in any system that requires scalable analysis of semantic data. The system presented has some unique features such as grid-based infrastructure, extreme scalability, efficient real-time query answering and an on the fly access control layer that were presented in detail during the talk.

3.15 Process Mining and BigData

Marcello Leida (Khalifa University – Abu Dhabi, AE)

License  Creative Commons BY 3.0 Unported license
© Marcello Leida

This discussion session focused on the various big data technologies and how can they be applied to process mining area.

3.16 On the Suitability of Process Mining to Produce Current-State Role-based Access Control Models

Maria Leitner (Universität Wien, AT)

License © Creative Commons BY 3.0 Unported license
© Maria Leitner

Joint work of Leitner, Maria; Rinderle-Ma, Stefanie

Main reference M. Leitner, “Delta analysis of role-based access control models,” in Proc. of the 14th Int’l Conf. on Computer Aided Systems Theory (EUROCAST’13), Part I, LNCS, Vol. 8111, pp. 507–514, Springer, 2013.

URL http://dx.doi.org/10.1007/978-3-642-53856-8_64

Role-based access control (RBAC) is the de facto standard for access control in process-aware information systems. With existing techniques in organizational mining, we can adapt these to derive not only organizational models but also RBAC models. In a case study, we evaluated role derivation, role hierarchy mining, organizational mining, and staff assignment mining on the suitability to derive RBAC models. We compared the derived models to the original and evaluated the results with quantitative measures. Furthermore, we adapted delta analysis to the RBAC domain to investigate the similarity of RBAC models and to analyze differences between the models. As an example, we analyzed the structural similarity using error correcting graph matching. With this approach, we can not only identify RBAC misconfiguration but also detect violations of the original RBAC policy.

References

- 1 Maria Leitner, Anne Baumgrass, Sigrid Schefer-Wenzl, Stefanie Rinderle-Ma, and Mark Strembeck: A Case Study on the Suitability of Process Mining to Produce Current-State RBAC Models. Business Process Management Workshops. LNBIP. Springer, pp. 719–724 (2013)
- 2 Maria Leitner: Delta Analysis of Role-based Access Control Models: Proceedings of the 14th International Conference on Computer Aided Systems Theory (EUROCAST 2013). LNCS. Springer, pp. 507–514 (2013) (in press)

3.17 Declarative Process Mining with ProM

Fabrizio Maria Maggi (University of Tartu, EE)

License © Creative Commons BY 3.0 Unported license
© Fabrizio Maria Maggi

The increasing availability of event data recorded by contemporary information systems makes process mining a valuable instrument to improve and support business processes. Starting point for process mining is an event log. Typically, three types of process mining can be distinguished: (a) process discovery (learning a model from example traces in an event log), (b) conformance checking (comparing the observed behavior in the event log with the modeled behavior), and (c) model enhancement (extending models based on additional information in the event logs, e.g., to highlight bottlenecks). Existing process mining techniques mainly use procedural process modeling languages for describing the business processes under examination. However, these languages are suitable to be used in stable environment where process executions are highly predictable. In turbulent environments, where process executions involve multiple alternatives, process models tend to be complex and difficult to understand. In this talk, Fabrizio Maggi introduced a new family of process mining techniques based on declarative languages. These techniques are very suitable to be

used for analyzing less structured business processes working in environments where high flexibility is required. These techniques have been implemented in the process mining tool ProM and range from process discovery to models repair and extension, to offline and online conformance checking.

3.18 Scientific Workflows within the Process Mining Domain

Ronny S. Mans (Eindhoven University of Technology, NL)

License  Creative Commons BY 3.0 Unported license
© Ronny S. Mans

Within the process mining domain there is currently no support for the construction and execution of a workflow which describes all analysis steps and their order, i.e. a scientific workflow. In the tool demo we demonstrated how we have integrated the scientific workflow management system RapidMiner with the process mining framework ProM 6. That is, several interesting workflows, consisting of multiple process mining tasks, will be constructed and executed.

3.19 Conformance Analysis of Inventory Processes using Process Mining

Zbigniew Paszkiewicz (Poznan University of Economics, PL)

License  Creative Commons BY 3.0 Unported license
© Zbigniew Paszkiewicz

Case study: Conformance analysis of inventory processes using process mining Production companies monitor deviations from the assumed procedures to satisfy quality requirements. In his talk, Zbigniew Paszkiewicz showed how process mining contributes to quality management efforts by analysis event logs about inventory operations registered in a warehouse management system. The analyzed company has pointed six aspects to be scrutinized: 1. conforming to model: inventory process instances must follow a pre-defined de jure model; 2. First In First Out policy: products that were produced first must be shipped first; the FIFO rule must be satisfied within particular product families; 3. quality assurance: all the pallets before being shipped to a client must be checked by the quality department; 4. process performance: a particular pallet cannot be stored in the warehouse for more than fourteen days; additional constraints concern the execution time of particular activities related to pallet management; 5. pallet damage handling: a pallet in disrepair must be transported to a special storage area; all the storekeepers are responsible for handling damaged pallets in this way; 6. work distribution: all the shifts should perform an equal amount of work; storekeepers taking pallets from production lines should not be involved in their shipping from the warehouse, and vice-versa. Conformance checking analysis has been performed with both ProM and commercial tools. Unwanted and repeatable parts of inventory processes in their business contexts have been identified with our novel RMV method. Unwanted and repeatable parts are represented as activity patterns which encompass the definition of activities and social relations among process participants. Preliminary results confirm that the RMV method provides useful insights about collaboration among process participants.

3.20 Predictive Security Analysis@Runtime – Lessons Learnt from Adaptation to Industrial Scenarios

Roland Rieke (Fraunhofer SIT – Darmstadt, DE)

License © Creative Commons BY 3.0 Unported license
© Roland Rieke

Joint work of Rieke, Roland; Repp, Jürgen; Zhdanova, Maria; Eichler, Jörn

Main reference R. Rieke, J. Repp, M. Zhdanova, J. Eichler, “Monitoring Security Compliance of Critical Processes,” in Proc. of the 22th Euromicro Int’l Conf. on Parallel, Distributed and Network-Based Processing (PDP’14), IEEE CS, to appear.

The Internet today provides the environment for novel applications and processes which may evolve way beyond pre-planned scope and purpose. Security analysis is growing in complexity with the increase in functionality, connectivity, and dynamics of current electronic business processes. Technical processes within critical infrastructures also have to cope with these developments. To tackle the complexity of the security analysis, the application of models is becoming standard practice. However, model-based support for security analysis is not only needed in pre-operational phases but also during process execution, in order to provide situational security awareness at runtime. This talk given by Roland Rieke presented an approach to support model-based evaluation of the security status of process instances. In particular, challenges with respect to the assessment whether instances of processes violate security policies or might violate them in the near future were addressed. The approach is based on operational formal models derived from process specifications and security compliance models derived from high-level security and safety goals. Events from process instances executed by the observed system are filtered for their relevance to the analysis and then mapped to the model of the originating process instance. The applicability of the approach is exemplified utilizing processes from several industrial scenarios. Lessons learnt from the adaptation of the method to the scenarios are addressed. In particular, event model abstraction, process instance identification, semi-automatic model mining, and cross process instance reasoning is discussed. Furthermore, the need for a method to derive measurement requirements from security and dependability goals is motivated and a meta model aiming at an integrated security strategy management is presented.

3.21 Data Collection, Integration, and Cleaning for Process Mining: Reflections on Some Projects

Stefanie Rinderle-Ma (Universität Wien, AT)

License © Creative Commons BY 3.0 Unported license
© Stefanie Rinderle-Ma

Joint work of Rinderle-Ma, Stefanie; Ly, Linh Thao; Mangler, Jürgen; Indiono, Conrad; Dunkl, Reinhold; Kriglstein, Simone; Wallner, Günter; Binder, Michael; Dorda, Wolfgang; Duftschmid, Georg; Fröschl, Karl Anton; Gall, Walter; Grossmann, Wilfried; Harmankaya, Kaan; Hronsky, Milan; Rinner, Christoph; Weber, Stefanie

Main reference L. Thao Ly, C. Indiono, J. Mangler, S. Rinderle-Ma, “Data Transformation and Semantic Log Purging for Process Mining,” in Proc. of the 24th Int’l Conf. on Advanced Information Systems Engineering (CAiSE’12), LNCS, Vol. 7328, pp. 238–253, Springer, 2012.

URL http://dx.doi.org/10.1007/978-3-642-31095-9_16/

In this talk, Stefanie Rinderle-Ma highlighted some process mining challenges from her own projects. The first project is EBMC2 which is a joint work between University of Vienna and Medical University of Vienna on patient treatment processes in skin cancer. The goal of the project is to discover the actual treatment processes and compare them with skin cancer

guidelines in order to analyze possible deviations. Though several data sources are available several data integration and quality problems occur, e.g., with respect to activity granularity and time. The second project is on higher education processes (HEP) where we tried to mine reference processes based on semantic log purging. Finally, some results on visualizing process difference graph including instance traffic are presented.

References

- 1 Michael Binder, Wolfgang Dorda, Georg Duftschmid, Reinhold Dunkl, Karl Anton Fröschl, Walter Gall, Wilfried Grossmann, Kaan Harmankaya, Milan Hronsky, Stefanie Rinderle-Ma, Christoph Rinner, Stefanie Weber: On Analyzing Process Compliance in Skin Cancer Treatment: An Experience Report from the Evidence-Based Medical Compliance Cluster (EBMC2). Int'l Conf on Advanced Information Systems Engineering (CAiSE'12), pp. 398–413 (2012)
- 2 Linh Thao Ly, Conrad Indiono, Jürgen Mangler, Stefanie Rinderle-Ma: Data Transformation and Semantic Log Purging for Process Mining. Int'l Conf on Advanced Information Systems Engineering (CAiSE'12), pp. 238–253 (2012)
- 3 Simone Kriglstein, Günter Wallner, Stefanie Rinderle-Ma: A Visualization Approach for Difference Analysis of Process Models and Instance Traffic. Int'l Conf on Business Process Management (BPM'13), pp. 219–226 (2013)

3.22 Disco (Tool Demo)


Anne Rozinat (*Fluxicon Process Laboratories, NL*)

License  Creative Commons BY 3.0 Unported license
© Anne Rozinat

This talk presented Disco, a professional tool for process mining practitioners.

3.23 DPMine/P Complex Experiment Model Markup Language as Applied to ProM


Sergey A. Shershakov (*NRU Higher School of Economics – Moscow, RU*)

License  Creative Commons BY 3.0 Unported license
© Sergey A. Shershakov

In his talk, Sergey Shershakov considered DPMine/P, a new language for modeling domain-specific Process Mining experiments, and tool support for this language based on ProM platform. The language under development aims at the unification of the separate phases of an experiment into a single sequence, that is an experiment model, support of looping constructs and other execution threads controls, provision of a clear but flexible (and, what is important, expandable) semantics. DPMine/P language is considered at the level of ProM tool as a set of plug-ins and data objects (which are the input and output data for the plug-ins). A description of some modules and examples of their use is provided.

3.24 A Process Mining-based Analysis of Intentional Noncompliance

Pnina Soffer (University of Haifa, IL)

License  Creative Commons BY 3.0 Unported license
© Pnina Soffer

Business process workarounds are specific forms of incompliant behavior, where employees intentionally decide to deviate from the required procedures although they are aware of them. Detecting and understanding the workarounds performed can guide organizations in redesigning and improving their processes and support systems. In this talk, Pnina Soffer presents her work on building specific types of workarounds found in practice, and defining corresponding log patterns for detecting them by process mining. Pnina analyzed logs of 5 real-life processes and find correlations between the frequency of specific workaround types and properties of the processes and of specific activities. The analysis results promote the understanding of workaround situations and sources.

3.25 Outpatient Process Analysis with Process Mining: A Case Study

Minseok Song (UNIST, KR)

License  Creative Commons BY 3.0 Unported license
© Minseok Song

In the talk of Minseok Song, a case study with a real life log from a hospital in Korea is explained. Based on the outpatients, event log in the hospital, he derived the process model and compared it with the standard model in the hospital. In addition, he conducted performance analysis to make a simulation model and analyzed the process patterns according to patient types. According to the result of comparing the event log and their standard process model, the matching rate was as 89.01%. That is, they relatively well understood workflows of outpatients and the process was well-managed by the hospital. Using the performance analysis result, he generated the simulation model. The simulation shows that the 10% increase of patients makes the largest change in consultation waiting time. Thus, he recommended less than 10% of increase. He extracted the process models and analyzed the process patterns according to patient types. The most frequent pattern of each patient type was discovered. The patterns are used to build a smart guidance app in the ubiquitous healthcare system in the hospital. As a future work, he will analyze more processes such as call clinical pathways, payment processes, etc.

3.26 PROMPT: Process Mining for Business Process Improvement

Alessandro Sperduti (University of Padova, IT)

License  Creative Commons BY 3.0 Unported license
© Alessandro Sperduti

This talk presented the PROMPT project and some of the results achieved by the Italian partners. Specifically, I present the basic ideas underpinning: a software for importing data from target information systems; a role mining algorithm; an approach for automatic selection of values for discovery algorithms parameters; a family of algorithms for on-the-fly process discovery. Work in progress is outlined as well.

3.27 SecSy: Security-aware Synthesis of Process Event Logs


Thomas Stocker (Universität Freiburg, DE)

License  Creative Commons BY 3.0 Unported license
© Thomas Stocker

One difficulty at developing mechanisms for business process security monitoring and auditing is the lack of representative, controllably generated test runs to serve as an evaluation basis. SecSy tries to fill this gap by providing tool support for event log synthesis. The novelty is that it considers the activity of an “attacker” able to purposefully infringe security and compliance requirements or simply manipulate the process’ control and data flow, thereby creating deviations of the intended process model. The resulting logs can be readily replayed on a reference monitor, or serve as input for auditing tools based upon, e.g., process mining.

3.28 ProM 6.3 (Tool Demo)

Eric Verbeek (Eindhoven University of Technology, NL)

License  Creative Commons BY 3.0 Unported license
© Eric Verbeek

Process mining has emerged as a way to analyze business processes based on event logs. These events logs need to be extracted from operational systems and can subsequently be used to discover or check the conformance of processes. ProM is a widely used tool for process mining. Earlier versions of ProM were distributed under the CPL license, required a GUI to run, and came with all functionality in a single bundle. As a result, it was not possible to run a mining algorithm from, say, a command line prompt, and we had problems using third-party libraries that came with a conflicting license. ProM 6 overcomes these problems, and ProM 6.3 is the latest version in this line of ProM releases. ProM 6.3 can be downloaded from <http://www.promtools.org/prom6>.

3.29 Process Mining in China - Recent Work on Event Quality

Jianmin Wang (Tsinghua University Beijing, CN)

License  Creative Commons BY 3.0 Unported license
© Jianmin Wang

Business process management has been used in Chinese enterprises widely in last 10 years. Investigating the accumulated event logs will enhance their competition capacities. However, the event quality is often not good enough. In this talk, Jianmin Wang introduced his recent work on event quality. 1) He studied the efficient techniques for recovering missing events. Advanced indexing and pruning techniques based on Petri net unfolding theories are developed to improve the recovery efficiency. 2) A generic pattern based matching framework was proposed, which is compatible with the existing structure based techniques. To improve the matching efficiency, he devised several bounds of matching scores for pruning. 3) An algorithm of mining the non-free choice structure from the dirty log with missing event was also introduced. Finally, the academic research groups in China are emulated and future research directions of our group are presented.

3.30 Turning Event Logs into Process Movies: Animating What Has Really Happened

Massimiliano de Leoni (University of Padova, IT)

License  Creative Commons BY 3.0 Unported license
© Massimiliano de Leoni

Today's information systems log vast amount of data which contains information about the actual execution of business processes. The analysis of this data can provide a solid starting point for business process improvement. This is the realm of process mining, an area which has provided a repertoire of many analysis techniques. Despite the impressive capabilities of existing process mining algorithms, dealing with the abundance of data recorded by contemporary systems and devices remains a challenge. Of particular importance is the capability to guide the meaningful interpretation of this "ocean" of data by process analysts. To this end, insights from the field of visual analytics can be leveraged. The talk discussed an approach where process states are reconstructed from event logs and visualised in succession, leading to an animated history of a process. This approach is customizable in how a process state, partially defined through a collection of activity instances, is visualized: one can select a map and specify a projection of activity instances on this map based on their properties. The approach is implemented as plug-in for process-mining framework ProM. The talk will also show the application to a case study with one of Australia's largest insurance companies: Suncorp.

3.31 Introducing Process Mining at Rabobank

Frank van Geffen (Rabobank – Utrecht, NL)

License  Creative Commons BY 3.0 Unported license
© Frank van Geffen

Our challenge is to match the customer needs of tomorrow. The speed and complexity of today's changes require a different approach to process improvement. Process mining, or automated business process discovery, is a bpm technique that helps in gaining insight in how processes are actually performed, how systems are used and how people work together. Through the explosive growth of data and significant advances in analysis and visualization technology it's possible to unlock valuable process information by analyzing transaction data. The use of automated business process discovery techniques yield new valuable insights. Process analysis done this way becomes fact based, full, for real and fast. Frank van Geffen told about his experience with introducing this new technology at Rabobank. Based on his practical experience, he stated the specific value of this technique for Rabobank. Besides the successes, Frank will also share the pitfalls he encountered and what measures can be taken to circumvent these obstacles.

4 Discussion Sessions

The seminar comprised 12 discussion sessions. Figure 2 depicts their organization and the chairs of each sessions. The following provides a summary of these discussion sessions, as reported by the discussion chairs.

Process mining of multi-perspective models (Chair: Akhil Kumar) Process discovery: Playing with the representational bias (Josep Carmona)	Data quality and data preparation (Chair: Frank van Geffen) Evaluation of process mining algorithms: benchmark data sets and conformance metrics (Boudewijn van Dongen)
Advanced topics in process discovery: on-the- fly and distributed process discovery (Alessandro Sperduti) Process mining in Healthcare (Pnina Soffer)	Process mining and Big Data (Marcello Leida) Security and privacy issues in large process data sets (Simon Foley)
Conformance checking for security, compliance and auditing (Massimiliano de Leoni) What is the ideal tool for an expert user? (Benoit Depaire)	How to sell process mining? (Anne Rozinat) What is the ideal tool for a casual business user? (Teemu Lehto)

■ **Figure 2** Overview of discussions sessions.

4.1 Discussion Session 1: Process Mining of Multi-Perspective Models

The motivation for this group was to discuss whether it would help to explore multiple perspectives in developing process mining algorithms. The group started out by identifying multiple perspectives in addition to the control flow perspective, i.e.:

- Data
- Resource/role/organizational
- Inter-process communication
- Time, costs, risks, energy consumption
- State of a process
- Performance
- Context

The discussion mainly centered on the data and organizational perspectives. We summarize the main issues discussed.

The complexity problem

To deal with the added complexity of multiple perspectives, one could start from a control flow model and enhance it with data related conditions at choice nodes and roles associated with tasks in the model. This may not always work because in some examples of BPMN model discovery it leads to a “spaghetti model.” Yet it was felt that each new perspective potentially adds value. Two approaches discussed are: 1) Treat each perspective as a layer of an onion, where the order of layers would be situation and need dependent; 2) Analyze each perspective separately, and integrate them. However, it was noted that a clean separation may not always be possible.

Representation Problem

While different perspectives can be used for filtering, clustering and alignment, how do we visualize them? An appeal of the control flow model lies in the ease of its visual representation as a graph. Some aspects can be added to this model by means of conditions and rules at gateway nodes, and association of roles with tasks. However, security constraints like binding and separation of duties would be hard to show. There is also need to avoid clutter and give users an ability to select what perspectives they wish to see and to zoom in-out, etc. as is done with maps. Further, multidimensional information could be displayed in 2-D by pairwise cartography.

Log issues

To perform multi-perspective mining the log must include additional data beyond events and timestamps. Thus, the need for additional data in logs was discussed. It was also noted that data can help to discover causality relationships and thus lead to inference of control flow also. The limits of our analysis capability are naturally limited by the information provided by a user in the log. The group felt it would also be interesting to think about extending the XES standard to include event data.

Conclusions

This discussion group sees value in research on multi-perspective process mining. It sees research challenges in the representation and complexity problems. It is felt that multiple perspectives can be analyzed serially or separately based on user questions and data availability. Finally, a need is perceived for user-definable interfaces that allow selection of perspectives, zoom feature, etc. and for extending the XES standard to represent event data in a log.

4.2 Discussion Session 2: Data Quality and Data Preparation

The purpose of the discussion was to gain insights into practical data quality and preparation challenges experienced by the participants. To this end, we first collected typical data challenges from the group and then presented our own challenges, which we had prepared. Frank discussed some challenges based on two real examples from the Rabobank to make it concrete. To summarize the data quality problems/issues regarding the *data quality*, we used an existing framework² to categorize the challenges we had collected in the group and before:

It became apparent during the discussions that missing events and timestamp problems were the most frequently issues mentioned. The discussion session discussed further data quality issues, which for the sake of space will not be described in detail here.

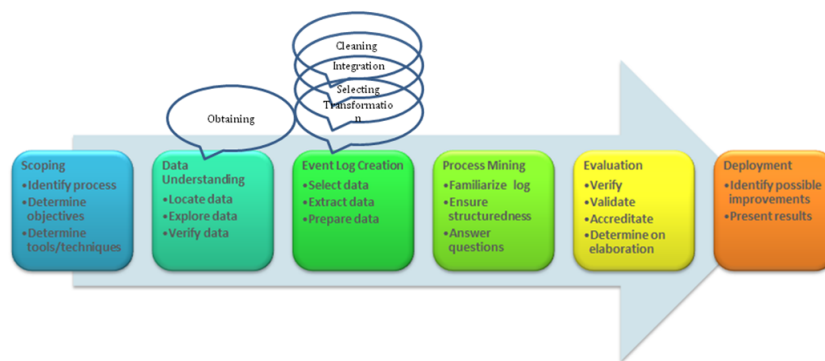
Turning to the *data preparation*, we added the data preparation phase “Obtaining” before “Cleaning”, “Integration”, “Selection”, and “Transformation” as a result of the discussion. To position the discussed activities in the context of a process mining project, we used an existing lifecycle model³ that also illustrates the iterative nature of data preparation, validation, data cleaning, etc. This is depicted in Fig. 4.

² R. P. Jagadeesh Chandra Bose, Ronny S. Mans, Wil M. P. van der Aalst. Wanna Improve Process Mining Results? It's High Time We Consider Data Quality Issues Seriously. BPM Center Report BPM-13-02, BPMcenter.org, 2013

³ T. van der Heijden, Process Mining Project Methodology: Developing a General Approach to Apply Process Mining in Practice. Master Thesis, 2012

	case	event	belongs to	c attribute	position	activity name	timestamp	resource	e attribute
<i>Missing Data</i>	I	V	III			I	II		I
<i>Incorrect Data</i>			I				III		I
<i>Imprecise Data</i>				I			II	I	I
<i>Irrelevant Data</i>		III							
	I		II		III	IV			V

■ **Figure 3** Data quality categories.



■ **Figure 4** Lifecycle for data preparation.

The complete list of collected data preparation challenges is the following:

- *Obtaining*: Diverse data sources; Very large datasets; Transportation of data; Same tables for different processes; What's the right process / activity scope?; Lack / mismatch between technical (data) documentation and reality; Access to different domain roles / function for understanding.
- *Cleaning*: Validation following individual cases (verifying data quality); Removing duplicates; Correcting date-timestamps; Action codes translation to real names (human-readable, URL to semantic action); What's `case_id` pointing at? (customers / products / documents / complaints / combinations.); Sampling (understanding data); Server times vs. Local times.
- *Integration*: Merging data sources; How to deal with large blobs full of free text?; Large file size, long waiting time; Usage of different separator `<quotes>`; Which merging sequence / order of steps? (when to merge which data / first merge then process view or visa versa / automation).
- *Selection*: Connecting multiple case IDs to follow end-to-end process; What amount of data is required? (as much as possible, prize, anonymizing, decision criteria); Does the dataset need to be enriched?; Sampling (criteria?).
- *Transformation*: Formatting: activities in columns, you lose loops and assume a pre-specified process is followed; Server times vs. Local times; Which environment is receiving the prepared data? (Disco sets other standards than Rapid miner, Tableau or Click view for example); Working with / without weekends and night-shifts (adjusting timestamps to match the organizations opening and closing times); Multiple `case_id`'s (`customer_id`, `product_id`, `document_id`, `update_id`, `session_id`, `authorization_id`).

The discussion then elicited the following questions and action points:

- *Questions*
 - How to teach checking and communicating data requirements for process mining?
 - Are there tools to help with the understanding of the data properties?
 - What are the consequences of the different data quality levels from the manifesto on certain process analysis questions?
 - When to use sampling and according to which criteria?
 - Could there be a tool to automatically guide through the data preparation process?
 - Organize seminars together with the database community?
- *Actions*
 - Share data sets online.
 - Share best practices.
 - BPI Challenge: compare and summarize the submissions to gain insight in the different approaches

4.3 Discussion Session 3: Process Discovery: Playing with the Representational Bias

The goal of this discussion was to identify the main challenges the representational bias brings into the process discovery discipline. One of the initial discussion run through the proper notion, specially what characterizes a discovery algorithm in terms of the implicit bias it has.

Right after discussing the general notion, the group has identified two different levels from which the representational bias can be considered:

- the *logical* level: where aspects like the semantics of the model (imperative/declarative), the patterns to be represented, and even the ability to transform the derived models is an important issue.
- the *user* level: where quality metrics related to the user must be taken into account, e.g., truthfulness, readability, multi-perspective, are examples of this.

Both of these levels are by themselves challenging, and it is agreed that very few work has been done into guiding process discovery algorithms for satisfying them.

At the *logical* level, there are well-known examples of patterns that differentiate discovery algorithms: expressive power, concurrency, skip/duplicate activities, non-free-choiceness, hierarchy, loops or cancellation are typical examples of patterns that not all algorithms have. On the other hand, process discovery is harden by the presence of other problems like noise, incompleteness, concept drift (it seems is less stringent in practice). In addition, the granularity of event information has been identified as a problematic issue, but also the selection of parameters given a particular process discovery technique. One promising direction has been identified, which may alleviate some of the problems before: define log features that can help into making decisions and transformations into the log for improving the discovery.

At the *user* level, an issue which is important is the current situation of process discovery algorithms: do the current users know their bias? the discussion group has identified as a challenge the user explanation of each algorithm's bias. Apart from that, other factors like execution time available for discovery, expected truthfulness of the model, or desired

readability are crucial factors that the user may want to determine when using an algorithm. Again, very few techniques possess the aforementioned abilities.

General challenges have been identified, which listed below:

1. How to control the representational bias of process discovery algorithms? the conclusion of the discussion group was that it seems an important aspect, although few techniques offer it. One example is the ability of certain algorithms, like the inductive miner or the miners implementing the theory of regions, to focus the search of a model to certain quality criteria established a priori.
2. Meta-discovery: both at the logical and at the user level, it seems relevant to decide generically the representation. An example now comes handy: declarative models are known to be good in “turbulent” scenarios, while imperative models may better fit structured scenarios. One can use domain knowledge, log features or the like to decide it. The same can apply at a more concrete level, like what particular formalism may be better for the user, e.g., BPMN or Petri nets.
3. Industrial bias: the fact that industry is mostly considering standards like BPMN does not mean that process discovery algorithms should only aim at discovering these models. It is better to concentrate on the identification of patterns that may then be translated to the visual representation in terms of a particular formalism. Also, the group has identified the importance of having transformations between formalisms, even in the presence of precision losses or similar inaccuracies.

As starting point for further actions, the group has created a Dropbox folder where related papers will be collected in order to iterate over the literature and find synergies. As future work, it may be possible to trigger some collaborations in different dimensions (writing a report, joint efforts, and the like).

4.4 Discussion Session 4: Evaluation of Process Mining Algorithms: Benchmark Data Sets and Conformance Metrics

In this discussion session, we considered the maturity of the process mining community. While the research on process mining is maturing, the need arises for a clear benchmarking methodology, such that (a) researchers can *objectively compare* their results and performance against other researchers results and (b) researchers can *easily exchange* comparison results. The methodology should be language independent and community accepted.

During the discussion, we established that a model, when drawn by a process modeling specialist, is always created for a particular purpose. The purpose of a model should always be considered by process mining researchers in their evaluations, i.e. comparing process mining techniques should only be done for those techniques that serve the same purpose. In many papers today, comparisons are made without looking at the purpose of the models, thus leading to false comparisons. During our discussion, we identified several purposes, such as:

- Prediction, i.e. answering “what if?” questions,
- Happy flow discovery, i.e. visualizing the main process flow,
- Perfect representation, i.e. models that very accurately show what behavior was observed,
- Performance analysis, i.e. models that provide insights into the performance of an “as-is” situation, and
- Deviation discovery, i.e. where models explicitly show deviations from reference models, business rules, etc.

Depending on the purpose of a model, several aspects of the model may be more or less important, hence when evaluating the quality of a process model against an event log, several dimensions should be considered. During the discussion, we identified the following dimensions for which we believe language independent metrics should be developed:

- Replay fitness, i.e. the fraction of the observed behavior that fits a model,
- Precision, a measure for the amount of behavior allowed by a model, but not observed,
- Simplicity, which quantifies the understandability of the model given the behavior it expresses,
- Generalization, which quantifies to what extent the model generalizes from the observed behavior, and
- Level of decomposition, hierarchical or otherwise, quantifying how “flat” a model is.

When comparing process mining techniques, it is important to realize that optimality can often not be reached in all dimensions, i.e. models may be Pareto optimal. During the discussion, we came to the conclusion that there is currently no clear methodology for evaluating process mining techniques. Therefore, a full methodology will be proposed in a paper to be written by several participants of the Dagstuhl seminar.

To conclude, some clear points were made that should be considered already today when comparing process mining techniques to existing work:

- Compare new techniques with *all* existing techniques serving the *same purpose*,
- Compare new techniques against *many, randomly generated datasets*,
- Compare new techniques on *public, real-life datasets* available in the 3TU datacenter and
- Always *publish synthetic data* in the 3TU datacenter and preferably publish real life data too.

4.5 Discussion Session 5: Advanced Topics in Process Discovery: On-the-fly and Distributed Process Discovery

The goal of this discussion was to identify the main challenges posed by on-the-fly and distributed process discovery.

On-the-fly process discovery requires the compliance to the typical stream processing constraints: i) since it is impossible to store the complete stream, only a finite memory budget is allowed; ii) backtracking over a data stream is not feasible, so discovery algorithms are required to make only one pass over data, taking bounded time per event; iii) it is important to quickly adapt the process model to cope with evolving processes (concept drift); iv) the approach must deal with variable system conditions, such as fluctuating stream rates. Some on-the-fly discovery algorithms able to generate control-flow models and DECLARE models have already be defined. The success of these algorithms has been evaluated using traditional metrics defined for off-line process discovery. Thus it was debated how to define a proper evaluation measure for stream discovery tasks. It was agreed that the use of some of the data from the stream to evaluate fitness, precision, and other already defined measures can be considered a satisfactory solution, especially if these measures are then integrated over time. After some discussion, the addition of the social and data perspectives was recognized as an important first challenge. Considering on-the-fly both social and data perspectives is not obvious since the process model may change over time (concept drift). It was suggested that, concerning the data perspective, a possible solution could be to define a stability index over the control-flow and when the control-flow is stable, learn rules for choice points. When this problem is considered from a declarative point of view, a critical issue is whether the

concept of “activation of a constraint” is still “valid” in a stream setting. The result of the discussion on this issue has been that such concept seems to make sense as long as single events (disregarding the trace they belong to) are considered, while it seems not to be meaningful when considering an event within a single trace. Different declarative discovery algorithms can be devised according to whether event-based or trace-based focus is adopted.

Another important challenge that was discussed is how can discovery algorithms which are not based on simple statistics, as the already proposed algorithms, be extended to cope with the stream scenario constraints. The discussion suggested two kind of answers. A first suggested general approach has been to face a single constraint at time, so to evolve versions of the algorithm that eventually will be able to cope with all of them. A more specific approach could be to use a model update strategy that works only on the parts of the model that are affected by the current event (as already suggested by some authors in similar scenarios); however, it is not clear that the constraint on computational time will be satisfied. In addition, a potential problem of this latter approach is that concept drift, e.g. in seasonal processes, may not affect fitness while seriously affecting precision. A suggested way to cope with this issue is to adopt strategies to recognize which parts of the current process model is not used anymore and then remove them.

The discussion then focused on what should be visualized as output by these type of discovery algorithms. It was observed that it is not sufficient to output only the current model. It is more informative to display a model where not recently used parts are identified by a “cold” color, while most recently used parts are identified by a “hot” color. This allows for a comprehensive summary of historical behaviors. Moreover, it could be nice to generate a “movie” showing the evolution of the model in time. A grand challenge would be to mine the model evolution to extract a summary of how the model has evolved in time.

The discussion then turned on the usefulness of on-the-fly discovery algorithms in practical applications. There was a general agreement on the fact that first of all, companies are more and more adopting information systems able to produce and to process streams of data; secondly, this kind of algorithms are anytime algorithms which can be used under user defined time and storage constraints: user does not want to wait hours to discover that she/he selected the wrong data for process discovery. Thus, there is a positive side-effect in designing stream discovery algorithms, since this will allow the user to significantly shorten the exploration of event logs.

A final discussed issue concerning on-the-fly process discovery was whether GPUs can be used to speed-up computation for more demanding discovery algorithms. The discussion did lead to three outcomes. First of all, it was observed that first results on computing fitness for traces in a log are negative, mainly because it is very time consuming to transfer data from RAM to GPU global RAM (GRAM); moreover alignment is done in a sequential fashion. Evolution of GPU architectures and a smarter way to perform alignment could improve the situation in the near future. Secondly, computation of fitness could be distributed over many CPUs, while GPU computing can be used for other computations which are more suited for GPUs architecture. Finally, one solution to the above problems could be to study whether computation in discovery algorithms can be cast in a mathematical form which is amenable to fast GPU computation, such as matricial computation.

The second main argument of discussion was the possibility to distribute computation of process discovery (and conformance checking) algorithms over many CPUs. In fact, recently it was observed that Petri nets can be decomposed (under specific mild constraints) into small parts, so to allow distribution of computation. An interesting observation is that such decomposition can be inherited by any log generated by the target Petri net. This

allows to define a distributed discovery algorithm where the log is first decomposed into several small parts; these parts are then used to discover corresponding process models; the discovered process models are then glued together and eventually the resulting process model is simplified to obtain the final process model.

An important challenge is how to partition the log so to guarantee that the “right” process is discovered. It was observed that there are two possible ways to partition the log. A trivial one is to partition the log horizontally, i.e. a different subset of traces is assigned to each CPU core. An alternative way is to split it vertically, i.e. split each trace in several pieces and distribute them among the CPUs cores.

It is not clear, however, how to do the vertical partitioning in an optimal way, i.e. by reducing the computational effort while obtaining the correct model. A final raised question was whether conformance checking can take advantage by distributed computation. The answer was affirmative, Trivially, each CPU core has a copy of the model and checks one trace; results are finally aggregated.

In summary, on-the-fly and distributed process discovery constitute very useful techniques which pose several computational challenges. Promising research lines to successfully face these computational challenges, however, emerged from the discussion, and there is concrete hope that very soon new and more efficient and effective process discovery algorithms will be devised.

4.6 Discussion Session 6: Process Mining and Big Data

The idea of this session was to discuss the relation between (a) Big Data, (b) Big Data technologies, and (c) process mining. We now live in a time where the amount of data created daily goes easily beyond the processing capabilities of nowadays systems. Nevertheless the strategic importance of the knowledge hidden in such data, for effective decision making is paramount. The ability of organizations, governments but also individuals to collect information in a plethora of different systems/formats has largely overwhelmed the ability to extract useful knowledge from it⁴; not to cite the attempt to integrate such knowledge with relevant information available outside organizational boundaries. The rapidly growing data sets with event data provide opportunities and also challenges.

The session started by introducing to the audience the term Big Data as “the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.” (Wikipedia). The group discussed about the need for a big data approach in process mining and the availability and existence of so called big data event logs. A part of the group was arguing on one hand that there is not really a urgent need for applying big data techniques to Process Mining since the research community is still focusing on solving other issues and that anyway big data sets are not easily accessible; on the other hand the fact that enterprises, governmental organizations and the likes are storing increasing amount of data and the process mining approaches and algorithms need to be adapted to this situation in order to be effective in the real world.

⁴ Gross, B. M. (1964). *The managing of organizations: the administrative struggle*. New York: Free Press of Glencoe.

The group identified the possibility for the research community to access large event logs from no profit organizations, governmental institutions, supply chain processes where data need to be shared. The first action point of the discussion was set to identify big data logs and made them publicly available for the Process Mining research community. Then the discussion moved on the three Vs (Volume, Velocity, and Variety) which define the dimension of BigData set in relation to Process Mining. We focused on measuring the event logs in relation to the three Vs: the group identified that big Volume in logs can mean a big number of process execution traces and/or big number of events per trace and/or big number of attributes per process/event. Big Velocity means in Process Mining that the logs needs to be processed before a given time, the rate between the incoming logs and the consumed ones need to be constant and this is valid also for the process mining algorithms. Finally we discussed about what Variety means in process logs; one obvious observation was on the fact that logs can have multiple formats and the systems should be able to deal with this, but a less obvious comment was on the fact that logs can have multiple points of view and by changing the identifier of the process case the process can be seen from a completely new perspective. Also the group moved on discussing if going full data makes sense: the trend is nowadays is a “throw in all” approach but this needs to be carefully done by analyzing costs versus benefits of this trend. Moreover it was pointed out that with big data in the picture it becomes paramount to help the user to “find the needle in the haystack” and so local or partial mining/visualization techniques may become necessary in the future. Then, after a short introduction to the main big data technologies the discussion focused on what technologies can be relevant to Process Mining: depending on the problem to solve Map Reduce can be used or not but it needs to be carefully planned because forcing a map reduce approach can easily degrade performances.

Map Reduce has been used in some cases for preprocessing the logs for correlation however the Map Reduce framework imposes some relevant constraints on the way the conformance checking or process discovery algorithm access the log data. The particular Data Partitioning step required for distributed process mining is the main reason why Map Reduce cannot be easily used for the generic approach in distributed process discovery and conformance checking. The group identified the fact that map reduce can be used for some simple Process Mining algorithm such as the Alpha algorithm⁵, other more complex algorithm, especially the ones sharing global states cannot be easily implemented in map reduce and therefore a shared memory approach (memcached, grid computing, GPUs) is advisable. Problems like concept drift on streams of events can be solved using a distributed stream processing approach (such as Storm). The discussion then moved on presenting a set of research works on distributed mining⁶ that can be considered the actual state of the art. Some approaches⁷ use distributed computing to speed up Process Mining algorithms such as the genetic process mining, however the log is replicated across the nodes and therefore this approach is not possible is the logs cannot be stored entirely in one machine. Therefore the group focused on the fact that the partitioning of the Logs is of extreme importance for effective distributed process mining. Moreover horizontal partitioning technique provides some additional benefits

⁵ A paper on this aspect titled “Big Data meets Process Mining. Implementation of Alpha algorithm in Map Reduce” will be published at EE track ACM-SAC 2014)

⁶ W. M. P. van der Aalst. Decomposing Petri nets for process mining: A generic approach. *Distributed and Parallel Databases* 31(4):471-507, 2013.

⁷ C. Bratosin, N. Sidorova, and W. M. P. van der Aalst. Distributed Genetic Process Mining. *IEEE World Congress on Computational Intelligence*, pp. 1951–1958. IEEE, 2010)

like data compression. Also Process Cubes⁸ can benefit from a distributed approach in order to speed up the slicing, dicing, drilling down and rolling up of process traces and distribute the mining of separate set of process traces. Finally the discussion ended by introducing the concept of open data sets representing all sort of data (weather, biological, traffic records, ...) which nowadays are publicly available; some Data Mining tools such as rapid miner started providing plug ins in order to use this type of data in the mining process, but this sort of data have never been used for process mining therefore it may be interesting to see if this data can provide benefits especially in the analytical aspect.

4.7 Discussion Session 7: Process Mining in Health-care

Health care is considered an interesting and promising domain for process mining application, due to its challenging processes, where significant impact can be made. The discussion took as a starting point two different views of a medical process:

- *Clinical view*: actions done for affecting the current physical state of patients. The emphasis of this view is on curing patients, improving their life expectancy and quality, and being able to predict outcomes of treatment. Treatment processes should comply with clinical guidelines.
- *Logistic / administrative view*: execution of the medical process using resources over time, spending and gaining money. This view emphasizes KPI and resource optimization, while meeting standards and constraints. It also addresses scheduling, costing, billing, and mitigating legal risks.

Process mining research has so far mainly addressed the logistic view. The Frequently asked questions of process mining in health care⁹ have relevance for both views, but their essence is at the logistic view. For the clinical view, data as well as control flow and data perspectives should be emphasized.

Current process mining approaches are capable of meeting most of the needs of the logistic view. Hence, this view poses an opportunity for the process mining community to make a significant impact and show good results.

The clinical view has so far received less attention. The challenges it raises are many fold. First, it requires addressing the data perspective, so in addition to considering the actions, their outcomes should also be addressed (e.g., the result of the X-ray). Second, understanding the data requires domain knowledge, hence collaboration with physicians is needed. Third, compliance should be assessed with respect to medical knowledge (clinical guidelines), whose representation requires expressiveness beyond that of business process models (e.g., temporal constraints). Finally, mining results need to be visualized in a way which captures all the relevant aspects and is meaningful to domain experts. Of current output forms, the output of declarative decision mining can be suitable, especially if transformed to natural language representation.

Nevertheless, important results can emerge from mining the clinical view. Specifically, these results can provide improved decision support for physicians. Furthermore, conformance

⁸ W. M. P. van der Aalst. Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining. Volume 159 of Lecture Notes in Business Information Processing, pp. 1–22, Springer 2013.

⁹ R. S. Mans et al. “Process mining in healthcare: Data challenges when answering frequently posed questions.” Process Support and Knowledge Representation in Health Care. Vol. 7738 of Lecture Notes in Computer Science, pp. 140–153. Springer Berlin Heidelberg, 2013.

checking is of importance since different decisions might be taken by different physicians in similar situations. Using current technology requires much data preparation, including preprocessing and cleaning (e.g., combining activities of the same type), dynamic labeling, and tagging events during execution.

Case Study on a Treatment of “Urinary Tract Infection”. Considering a given process model, two groups (for the clinical and the logistic view) discussed the questions to be addressed and the desired results of mining.

Logistic view: The analysis process should include three different phases: (1) initial information gathering from experts: the problem, a normative process model, KPIs. (2) from detailed questions, detailed analysis can be performed, and (3) the results can lead to further exploration. The specific questions would depend on the purpose of analysis and on the stakeholder it should serve (e.g., hospitals seek to maximize throughput and minimize costs). Benchmark data and relevant standards will be needed. Example questions may include: (1) what is the cost of every test, which tests and what is their order. Analysis can contribute to final decision (skip tests, reordering); (2) how long for each step; (3) are there reworks (tests that are repeated); (4) are things that are not recorded not necessarily happening; (5) are actions done in batches; (6) what is the average time between tests / scheduling constraints / how long are patients waiting; (7) are there differences in treatment path among physicians.

Clinical view: This group devoted less attention to the details of the specific case study, and discussed clinical view mining requirements in general. Tooling – representation should be dynamic and interactive, capable of handling spaghetti-like processes, and allow switching between views (e.g., showing simple flowchart and projecting other data). New views might also be needed. The role of context is crucial. Context includes patient data attributes (e.g., age), treatment history, and status of other running instances. To support this, mining can relate to feature selection for extracting context. Then all mining can relate to context: treatment pathways, treatment outcomes (with respect to context-dependent goals), forecasting and operational support (e.g., possible consequences of treatment options). Decision support should provide recommendations by a case-based system.

4.8 Discussion Session 8: Security and Privacy Issues in Large Process Data Sets

This session considered security and privacy issues from two perspectives. Firstly, how process mining techniques may help to secure systems and, secondly, the security and privacy issues that arise as a consequence of process mining.

Enterprise system security can be characterized in terms of the security *controls* that are required to mitigate the *threats* to the objectives of its business *processes*. Threats can range from failures in business processes to more infrastructure-level vulnerabilities such as those cataloged by vulnerability databases. Process Mining can help with threat identification by generating reference models of normal behavior against which anomalous behavior in logs can be detected and explained (conformance checking).

The application of Process Mining to the configuration and selection of security controls was discussed. Organizational mining can help in the discovery of RBAC configurations and it was suggested that process discovery could help in the discovery of behaviors used for the configuration of task-based security policies. While a discovered process can be used for subsequent conformance-checking, an alternative viewpoint is whether it is possible to also

use this process to generate/recommend security controls that enforce conformant behavior. For example, using a discovered document workflow to help deploy checksum-based controls in a document handling system. A related question is how Process Mining could be used to explore whether the current security controls are resilient to changes in the process, and vice-versa.

Audit procedures test the efficacy of security controls and it was suggested that Process Mining could provide a basis for a more complete check on control efficacy. Security controls and their audit procedures are not necessarily integrated into business processes. For example, a procedure that regularly searches the file system for stray plain-text credit card numbers operates independently of credit-card based transaction processes. A research challenge is how security controls, along with their audit procedures, can be correlated with discovered business processes in order to provide more threat-aware conformance checking.

Notwithstanding the conventional integrity, availability, authenticity, non-repudiation and confidentiality challenges surrounding process and log data, Process Mining introduces particular assurance issues. Log data can come from different sources with varying degrees of assurance and trustworthiness. One question is how these relationships might be securely managed and how they might be reflected, not just in the original log-data, but also surfaced into any discovered process. Would such a scheme require a single security authority with jurisdiction over all log data and sources, or can a more decentralized approach be taken? The latter may be useful if organizations share log data. For example, organizations merging or aligning their interdependent processes in a supply-chain. For these federated logs, what are their security and privacy requirements and how might they be implemented?

Different users may have different views on different process logs and process mining should preserve these view restrictions. A challenge is the extent to which Process Mining can be carried out on views alone rather than on the full log-data. An advantage of the former is that any (security) failure in mining does not expose data outside the view, while the latter provides more precision but requires assurance in the process mining software. A further challenge is how log data can be reliably de-identified/anonymized. Differential-privacy based techniques may be useful in implementing privacy aware views: a discovered process should not reveal data previously de-identified in the log.

Lastly, a recent Semantic study on the cost of data breaches identified human factors and business process failures as a significant contributory factor. Given that Process Mining helps provide deeper understanding and control of business processes it would be worthwhile investigating its application to identifying process weaknesses that may lead to data breach.

4.9 Discussion Session 9: Conformance Checking for Security, Compliance and Auditing

A large part of the discussion was concerned with coming to an agreement of the terminology. The discussion was very active but, unfortunately, it has been impossible to agree on the terminology. In particular, it was clear to identify two different schools of thought:

- A first school differentiated between a-priori and a-posteriori verification. Compliance checking is concerned with verify the TO-BE model against norms, regulations, security constraints. Therefore, compliance checking is strictly related to analysis and verification of an executable model, which should adhere to constraints imposed by laws and regulations. In many settings, it is not strictly enforced that the actual execution follows this model. Here, conformance checking comes to play. Conformance checking is about verifying

whether the actual executions, as recorded in the event log, follow the same constraints. In this case, there is not one single model but, rather, multiple models that are built ad-hoc for the sole purpose of rules' checking.

- A second school sees conformance checking as a mean to check compliance. Conformance Checking takes a regulatory model and an event log and highlights the non-conformances. In this sense, there is some common point with the view at point 1. The serious difference is with respect compliance checking, which is still about verifying the behavior observed in the event log. Compliance checking is concerned with a number of norms, regulations that are converted in a number of regulatory models. Using the conformance-checking means, each of the regulatory models is verified against the event log at disposal. In this second school, auditing is an umbrella under which compliance checking is placed, along with “static” conformance verification. The latter refers to an executable model that is wanted to check the adherence with laws and regulations.

During the discussion, many issues raised up about the languages. In the last decade, several languages have been standardized. The discussion's participants agreed on the fact that there are several standards but not one largely-recognized standard.

Some languages are characterized by a precise syntax and semantics but a bit complex to use for average process analysts. Some are more user-friendly but at the cost of less accuracy in the semantics. This raised another important point: the usability of the languages. The language complexity can be problematic if process analysts do not have a sufficient background in Mathematics. Therefore, the syntax should be kept simple and intuitive for moderately-skilled process analysts but, at the same time, should have a precise and well-defined semantics to not lead to multiple interpretations. Furthermore, decidability and complexity of the algorithms that are used for conformance/compliance checking put a bound on the expressiveness of the languages used to define the conformance/compliance rules.

Once a language and compliance/conformance algorithms are chosen, it is possible to check for compliance and conformance. Of course, a pure YES/NO answer is not enough, i.e. the executions of single process instances are compliant/conformant with rules and regulations. The discussion participants acknowledge the important of pinpointing deviations and their root-causes. Moreover, in many settings, deviations do not occur in isolation: a deviation may cause a cascading effect, which can lead to further deviations. Hence, it is important to relate to each other when a relation may exist.

A few other important thoughts were discussed during the session. Firstly, in continuous auditing, it is important to be equipped with decision support that guides participants to not violate compliance constraints at run time. Using event logs, machine-learning techniques can be used to discover the common patterns that lead to problems and, hence, the decision support system can suggest execution paths that avoid them. Problems of applying runtime reasoning, e.g. Supervisory Control Theory, are rarely applicable in this context as the problem is inherently hard and, often, become undecidable. Secondly, in many settings, primarily in Security, the conformance/compliance of the execution of process instances cannot be checked in isolations. The compliance of an instance may depend on other instances of the same process or, even, on instances of different processes. This also highlights the importance of contextual information, which should be incorporated into the analysis.

As a conclusion, every discussion participant agreed on the fact that checking compliance/conformance is an important topic in the field of business process management and auditing. The main issue seems that, unfortunately, there is not a large consensus on what checking compliance/conformance actually means. An effort should be made in order to make sure that there exists no different wording for the same concepts. The discussion was

very animated and live. This certifies that checking compliance/conformance is certainly a topic that will attract much attention in the future. As a matter of fact, many of the existing techniques are still in a development stage and there is a lot of room for future improvements.

4.10 Discussion Session 10: How to Sell Process Mining?

Goal: Since everything in selling is about understanding the customer, the purpose of the discussion was to first gain a more nuanced view about the different types of possible target customers for process mining.

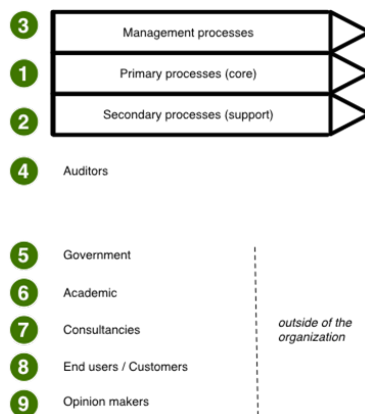
Approach: We collected a broad range of target customers from the group and then discussed three profiles in more detail. Afterwards, the results were put together by Frank van Geffen and Anne Rozinat.

Results: An operational manager of a business process is confronted with different (sometimes conflicting) goals, as depicted in Fig. 5.



■ **Figure 5** Conflicting goals.

We used the value chain-model of Porter to categorize the various business functions we had collected and added the other categories outside of the organization. The value-chain model is depicted in Fig. 6.



■ **Figure 6** Porter's value-chain model.

The following target customers for process mining were mentioned in these categories:

1. Operational manager, Sales department, Customer satisfaction representative, Process managers, Requirements engineer, Operational people, Process manager / Department head, Knowledge worker, Software product development, Supply management

2. Business consultants, CIO, Quality and process improvement department, IT managers, Process management department, Business analyst, Development and analyst people, Knowledge worker
3. Higher managers, Business controllers, CEO
4. Financial auditors, operational auditors, IT auditors, Crisis and fraud people
5. European commission
6. Other scientists
7. Consultancy firms
8. End user
9. Analyst firms, such as Gartner and Forrester.

We used the following three profiles to paint an as detailed picture of the respective customer as possible:

- A Business analyst at a hospital
- An End user (customer) of a service organization like the Rabobank
- An Auditor at a manufacturing company like Boeing

For each of these profiles we then tried to answer these questions: “How do they spend their day?”, “Which processes are in their sphere?” “What are their challenges?”, and “How can process mining help?”

4.11 Discussion Session 11: What is the Ideal Tool for an Expert User?

The idea of this session was to discuss the requirements and ideas for realizing process mining tools aiming at experts. The discussion focusses on the following three goals: identification of expert user types, identification of functional and non-functional requirements and suggestions how the academic community could support the development of expert tools.

Since “expert user” is a rather vague and general term, the discussion started with the identification of different expert user types. Two dimensions were proposed to identify different experts, i.e. the problem and user dimension. The problem dimension divides process mining problems along a range from well structured, standardized and repeating problems to unique, ill-structured and more generic problems. As for the user dimension, a distinction should be made between developers of process mining algorithms and actual users of process mining algorithms. Based on these two dimensions, three expert users could be identified.

The *Algorithm Developer* creates new process mining algorithms for generic process mining problems. Typically, this type of expert user can be found in academia. The *Data Scientist* can be found in a business environment and solves ill-structured process-related business questions by means of data and the use of existing process mining algorithms. The *Tool integrator* is also a business user, but in contrast to the *Data Scientist*, deals with structured and reoccurring process mining problems. They typically develop tool chains of process algorithm tools, business intelligence tools and enterprise systems to generate management dashboards.

Next, the discussion continued with the identification of non-functional requirements, which are the tool-usage specific properties. The *Algorithm Developer* requires a tool that allows reuse of existing algorithms, the ability to modify existing algorithms, proper documentation with instructions on how to develop with the expert tool, a system which

encourages or even enforces proper documentation of newly developed algorithms and a comprehensive overview of all available algorithms.

The *Data Scientist* requires a tool that provides scripting functionality and an easy integration with third party applications. The tool should allow the expert user to easily import and export data from and to a wide range of formats and should provide a flexible environment to manipulate data. Finally, the tool should offer algorithms that are robust and fast, well documented from a user's perspective and which can be tuned by means of parameters.

The Tool Integrator requires a tool that provides solid interconnectivity opportunities with various enterprise systems. If the tool aims to generate the dashboard directly, it should provide a flexible and customizable user interface. Furthermore, it would be a great asset if the tool already provided and supported predefined workflows for standard process mining problems.

Next, the discussion focused on the functional requirements which refers to currently unavailable algorithms which will become increasingly important in the near future. The discussion group identified the need for process mining algorithms which allows for: distributed process mining, data stream process mining, predictive process mining, multi-perspective process mining, direct comparison of processes at the process model level, process simulation, interactive visualization.

Finally, the meeting agreed on the importance that advances in process mining research gets integrated in expert tools and the need for the scientific community to contribute to the development of expert user tools. However, the concern was raised that currently there are little to no incentives for academia to actively contribute to the development of expert tools. For example, the current log loader of ProM has problems with loading big data, which hinders the further development of process mining algorithms for big data. While the community would clearly benefit from the development of a new log loader, it is in no individual's academic interest to spend a lot of time on this. To solve this catch-22, it was suggested that the research community agrees to assign budget in future (European) research projects to the development of much required, but academically non-interesting features and components of tools for algorithm developers.

4.12 Discussion Session 12: What is the Ideal Tool for a Casual Business User?

The idea of this session was to discuss the requirements and key success factors of a process mining tool tailored to business users. To this end, the following topics were discussed:

- What are the essential functionalities of a process mining tool for the casual business user?
- What is the functionality in existing tools that is most useful?
- What is missing?
- Should tools provide operational support (on-the-fly discovery, prediction, checking, recommendations)?
- How to visualize results?

In essence, the discussion could be capture as follows: vendors sell *features* while customers see the *benefits*. Given this line of discussion, Table 1 depicts the relationship – worked out during the discussion – between different types of uses and the corresponding benefits. Table 2 establishes the relationship between the benefits and the expected functionality.

■ **Table 1** Relationship between type of user and expected benefits.

User	Benefit
Process Owner	Verify if employees meet the rules she has in mind
Manager of document flow department.	Show if the operations meet regulation and SLA: case must be closed in 30 days
Scientist / researcher in hospital	Interested in finding compliance to medical guidelines
M&A director/team, HR director	Want to understand processes in order to integrate companies.
Auditor	Auditor wants to find unknown patters. Auditor knows what rules should be followed and wants to test if those rules are followed.
Operational manager of customer service process	Be proactive and improve processes
Process Innovation department/team reports to CEO	Optimize the performance of whole corporation
Security Auditor	Find evidence of fraud in transaction log
Process Owner / director of service delivery	Improve operational performance

■ **Table 2** Relationship between benefit and expected functionality.

User + Benefit	Functionality
Process Owner: Verify if employees meet the rules she had in mind	Compliance/conformance checking: list of cases that violate the tules
M&A director/team, HR director: Want to understand processes in order to integrate companies.	Use process mining discovery, show workflow, organization perspective, data flow diagrams.
Operational manager of customer service process: Be proactive and improve processes	1. deeper understanding of process...tool should give metrics , New measure values should come on weekly bases automatically. Consultant does the initial work and then business user can get what she wants by herself. 2. Operative management wants to get graphs and dashboards . And then predictive analysis ...imagine a metric in a process...if we keep going like this then I will meet my objectives ...
Auditor:	Model checking feature like 4-eye principle. Feed in business requirements and then show cases that meet the requirements and those which do not.
Security Auditor: Finding evidence of fraud in transaction log	They want to have simulation/what-if that comes up with possible changes in the processes and suggests the best mitigation action.

The discussion session ended up the following insights: An ideal tool...

- ...gets source data easily and with high quality. If the data does not come then there is no continuous benefits.
- ...actively supports the user and shows only relevant options and functions.
- ...supports interactive navigation during discovery phase.
- ...is configured by the consultant and used by the business user.
- ...has certain set of flexibility for Business End Users. They do not want to call always consultant to make changes.

Currently the users are typically the early adopters who are willing to use a complex tool. Typical business managers do not have time to play with the tool.

Participants

- Rafael Accorsi
Universität Freiburg, DE
- Antonio Caforio
University of Salento, IT
- Josep Carmona
UPC – Barcelona, ES
- Paolo Ceravolo
University of Milan, IT
- Jan Claes
Ghent University, BE
- Jonathan Cook
New Mexico State University, US
- Ernesto Damiani
Università degli Studi di Milano –
Crema, IT
- Massimiliano de Leoni
University of Padova, IT & TU
Eindhoven, NL
- Benoit Depaire
Hasselt Univ. – Diepenbeek, BE
- Simon N. Foley
University College Cork, IE
- Luciano Garcia-Banuelos
University of Tartu, EE
- Christian Günther
Fluxicon Process Lab., NL
- Anna A. Kalenkova
NRU Higher School of
Economics – Moscow, RU
- Akhil Kumar
Pennsylvania State Univ., US
- Geetika T. Lakshmanan
IBM TJ Watson Res. Center –
Cambridge, US
- Teemu Lehto
QPR Software – Helsinki, FI
- Marcello Leida
Khalifa Univ. – Abu Dhabi, AE
- Maria Leitner
Universität Wien, AT
- Fabrizio Maria Maggi
University of Tartu, EE
- Ronny S. Mans
TU Eindhoven, NL
- Alexey A. Mitsyuk
NRU Higher School of
Economics – Moscow, RU
- Jorge Munoz-Gama
UPC – Barcelona, ES
- Zbigniew Paszkiewicz
Poznan Univ. of Economics, PL
- Alessandro Provetti
Università di Messina, IT
- Hajo A. Reijers
TU Eindhoven, NL
- Joel Tiago Ribeiro
UPC – Barcelona, ES
- Roland Rieke
Fraunhofer SIT – Darmstadt, DE
- Stefanie Rinderle-Ma
Universität Wien, AT
- Anne Rozinat
Fluxicon Process Lab., NL
- Ricardo Seguel
Excellentia BPM, CL
- Marcos Sepulveda Fernandez
Pontificia Universidad Catolica
de Chile, CL
- Sergey A. Shershakov
NRU Higher School of
Economics – Moscow, RU
- Pnina Soffer
University of Haifa, IL
- Minseok Song
UNIST, KR
- Alessandro Sperduti
University of Padova, IT
- Thomas Stocker
Universität Freiburg, DE
- Wil van der Aalst
TU Eindhoven, NL
- Boudewijn van Dongen
TU Eindhoven, NL
- Frank van Geffen
Rabobank – Utrecht, NL
- Eric Verbeek
TU Eindhoven, NL
- Thomas Vogelgesang
Universität Oldenburg, DE
- Jianmin Wang
Tsinghua University Beijing, CN
- Barbara Weber
Universität Innsbruck, AT
- Ton Weijters
TU Eindhoven, NL

