# Process Mining: a Recent Framework for Extracting a Model from Event Logs

Luís Santos, Universidade do Minho/UNU-EGOV, Portugal, luispsantos@sapo.pt

## Abstract

Business Process Management (BPM) is a well-known discipline, with roots in previous theories related with optimizing management and improving businesses results. One can trace BPM back to the beginning of this century, although it was in more recent years when it gained a special focus of attention. Usually, traditional BPM approaches start from top and analyse the organization according some known rules from its structure or from the type of business. Process Mining (PM) is a completely different approach, since it aims to extract knowledge from event logs, which are widely present in many of today's organizations. PM uses specialized data-mining algorithms, trying to uncover patterns and trends in these logs, and it is an alternative approach where formal process specification is not easily obtainable or is not cost-effective. This paper makes a literature review of major works issued about this theme.

**Keywords:** BPM; Process Mining; Automated process discovery

## 1. INTRODUCTION

One of the trends in Business Process Management (BPM) is a relatively new area, Process Mining (PM). In opposition to the traditional process definition, processes, within this technique, are not human-defined, but they are derived from specialized data-mining algorithms, with the intention of achieving to identify patterns that are not easily recognizable by human eyes (and minds).

Process mining is also known as *Automated Business Process Discovery* (ABPD) (Gartner, n.d.).

This paper aims to conduct a literature review, identifying major works regarding this interesting approach of BPM, based on raw data (event logs), which are not process-aware.

The departure point is the huge amount of data available and generated by digital systems, always increasing, and the information that can be extracted – nor in the perspective of traditional data-mining, but seeking to extract models, as we will see. Information systems (IS) from all organizations collect millions or even billions of event-logs, often associated with date-time stamps, human operators, costs, status of some objects or facts, which "hide" patterns that, by using appropriate techniques, could unleash processes.


*ISSN 2183-489X*
*DOI http://dx.doi.org/10.18803/capsi.v17.00-00*

We will discuss the origins and development of process mining in the context of the BPM movement, and we will stress the contributions of major authors and works.

Key concepts of the field will be presented and discussed, while the methodology will explain how this review was conducted.

The next step will be how a process mining can be accomplished, meaning that techniques and tools are highlighted.

The final issue is the set of challenges that process mining is facing today.

## 2. THE FOUNDATIONS OF PROCESS MINING

Since the Industrial Revolution, and mostly in the 20th century, many authors made contributions to improving management. Taylor, Ford, Fayol or Weber are often between the most cited (Heames & Breland, 2010, p. 433), each of them focusing on different aspects of how to improve business results and in the pursuit of economic efficiency.

In the last few decades, a process-centric approach rose as an evolution of previous ones, like Quality Management (QM), Management Resource Planning (MRP), Management of Information Systems (MIS) and many others. One could say that BPM was driven by three major traditions or areas: management, quality and information technology (Kohlborn, Müller, Pöeppelbuß, & Röglinger, 2014, p. 1).

Processes within organizations are often derived from the analysis of the business, its external stimuli and the structure of the organization. In other words, there is usually a more or less formal way to describe those processes.

The rapid dissemination and ubiquity of computers and other digital equipments made available huge amounts of data. We can think about the millions of records of product selling that are stored each day in the system of supermarket chain, or about the GPS source of position of the vehicles of a logistics provider, or about the quality parameters of each chip produced in a factory.

PM's approach is based in these event logs, be it because it's difficult to obtain a formal description of some business processes, or, if one exists, because its quality can be challenged and questioned, or even because the processes are changing too rapidly to be possible to analyse: "Due to the omnipresence of the Internet and its standards, information systems change on-the-fly" (van der Aalst, ter Hofstede, & Weske, 2003, p. 3).

According to Aufare and Zimányi (2013, p. 57), *process mining*, "a term recently coined, aims to bridge the gap between BI [Business Intelligence] and BPM by combining event data and process models".

Process Mining was early defined as "extracting models from logs" (van der Aalst et al., 2003, p. 6).

Among the pioneers of PM there is Wil van der Aaalst, a Dutch computer science professor of the Eindhoven University of Technology. Aaalst shifted his main interest from workflow models to PM, and wrote with Weijters one of the first papers using the expression process mining (Weijters & Aalst, 2001).

PM is a research discipline that aims

> to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in today's (information) systems. […]. Process mining includes (automated) process discovery (i.e., extracting process models from an event log), conformance checking (i.e., monitoring deviations by comparing model and log), social network/organizational mining, automated construction of simulation models, model extension, model repair, case prediction, and history-based recommendations.
>
> (W. v. d. Aalst et al., 2012, p. 170)

A few years ago, Professor Aalst led a group of some tenths of academics and professionals who, in the context of *IEEE Task Force on Process Mining*, published a Process Mining Manifesto (W. v. d. Aalst et al., 2012), defining principles and listing challenges, seeking to promote PM. This *Manifesto* should be considered one milestone and a key work regarding PM. Aalst is also the author of a major and key work on PM, *Process Mining: Discovery, Conformance and Enhancement of Business Processes* (2011). This book, the first regarding process mining, links two traditionally separated areas, BPM and BI, explaining and discussing *process discovery* (and other types of the method, as we explore ahead) from event data.

## 3. METHODOLOGY

In this work, the "framework" (Aalst, 2014, p. 5) we chose for the review is centred on the concepts related and involved with PM. We also briefly discuss some type of PM and available software tools.

The literature review was conducted doing a comprehensive search in search engines and scientific databases.

Initial researches were made in secondary sources (Coutinho, 2014, p. 61) in order to form a broader view on the subject and to gather some primary sources. Among those sources were generic search engines, like Google, library repositories (from national and foreign universities, including RepositoriUM) and encyclopaedias (like Wikipedia, Britannica and Larousse).

Indirect researches also took place, picking references from the most relevant previous works we've found, based either in the author or in the title. After identifying key authors and their major

works, further indirect research was done using these works' references, and trying to address different approaches, and also aiming to identify ongoing controversies.

In Table 1 we can see the number of references found using the keywords "process mining" and "business process discovery":

| Keywords / DB/engines | "process mining" | "business process discovery" |
|---|---|---|
| Google | 377 000 | 53 600 |
| Scholar | 13 500 | 366 |
| B-on | 3 914 | 139 |
| WoC | 1 012 | 16 |
| IEEE | 62 900 | 5 |
| Sage | 3 493 | 653 |
| Scopus | 1 851 | 34 |
| Science Direct | 952 | 9 |

Table 1 - Number of references found in scientific databases and search engines (2017.02.14)

A closer look, however, showed that not all those references were truly related with PM – the accuracy depends on the internal algorithm of each database/search engine. Part of the references was very specific, related with the use of process mining in particular industries or economic activities, like logistics or healthcare. Later on, the search was widen adding variants or restrictions, or searching by keywords from their lexical and semantical fields, or using related concepts. We also found that the great majority of the references was published in the last decade (B-on, 3548 from 2007 to 2017, 90% of the total). After reading papers' abstracts, and sometimes also the introduction and conclusions, about forty works were selected, by using criteria like the author, the number of citations in other papers or books, the themes and the focus of the work.

Being a quite new area, PM needs some clarification of concepts and techniques, and also to set the boundaries of the field.

The relative youth of the area, as we have seen, and the limited number of researchers still involved, often leads, with no surprise, of cross-citations, and almost there is no paper or book about process mining where the name of Professor Aalst would not be referred, and where also he frequently is co-author. We'll not be an exception, for the first part…

## 4. CONCEPTS IN THE LITERATURE

The basics concepts of PM were first established by Aalst and then by the cited *Process Mining Manifesto*. The concepts and definitions in several works we read are alike, and often redirect to the *Manifesto* or the writings of van der Aalst.

For instance, Reijers says that "Process mining is an approach to infer what a business process looks like from traces that are left behind in all kinds of information systems when executing that process" (Dumas, Rosa, Mendling, & Reijers, 2013, p. 178), and points out that "process mining is a much less subjective means to discover" the processes (*idem*, p. 179).

Table 2 resumes some of the key concepts related with PM, corresponding to basic types of the method.

| | | | |
|---|---|---|---|
| **Process Mining** | Techniques, tools, and methods to discover, monitor and improve real processes […]by extracting knowledge from event logs commonly available in today's (information) systems.<br>(W. v. d. Aalst et al., 2012, p. 19) | The goal of process mining is to extract information about processes from transaction logs.<br>(Dustdar, Hoffmann, & van der Aalst, 2005, p. 133) | The practice of business process mining attempts to reconstruct complete process models from data logs containing real process execution data.<br>(Tiwari, Turner, & Majeed, 2008) |
| **Process Discovery** | Based on an event log a process model is learned. […] identifying process patterns in collections of events.<br>(W. v. d. Aalst et al., 2012, p. 19) | *Process discovery* is defined as the act of gathering information about an existing process and organizing it in terms of an as-is process model.<br>(Dumas, Rosa, Mendling, & Reijers, 2013, p. 155) | |
| **Conformance Checking** | Analyzing whether reality, as recorded in a log, conforms to the model and vice versa. The goal is to detect discrepancies and to measure their severity.<br>(W. v. d. Aalst et al., 2012, p. 18) | Modeled and observed behavior are analyzed to determine if a model denies a faithful representation of the behavior observed at the log.<br>(Munoz-Gama, 2014, p. 7) | Conformance validation can be used to check process rules and improve processes within any organization.<br>(Caldeira & Abreu, 2016, p. 256) |
| **Enhancement** | A process model is extended or improved using information extracted from some log.<br>(W. v. d. Aalst et al., 2012, p. 18) | […] finding the best possible route while ensuring validity and reliability.<br>(Okoye, Tawil, Naeem, & Lamine, 2015, p. 364) | |
| **Event log** | Collection of events used as input for process mining<br>(W. v. d. Aalst et al., 2012, p. 18) | […] data generated from the execution of the process.<br>(Dumas, Rosa, Mendling, & Reijers, 2013, p. 353) | […] a log recording the execution of activities in some business processes.<br>(Song, Günther, & van der Aalst, 2009) |

Table 2 - Concepts associated with Process Mining

One recent and interesting development is the emerging of a *predictive model* approach of PM. Traditionally, PM's focus was past data, and tools were developed to discover business process models from past events. With predictive models, the goal is "develop a *predictive process modelling technique* based on *process mining* and *grammatical inference* that *accurately* predicts future behaviour of business processes and provides comprehensible results" (Breuker, Matzner, Delfmann, & Becker, 2016, p. 1010).

The purpose is to be able, for instance, to "warn decision makers about undesirable events that are likely to happen in the future" (Breuker et al., 2016, p. 1009).

## 5. TECHNIQUES AND TOOLS

Authors like Aalst stress the importance of the event logs, since processes will be derived from those logs: "We would even like to claim that *logging should be first class citizen* for any system that is used to support business processes!" (2006, p. 3), and, again, the *Manifesto* emphasizes that "The quality of a process mining result heavily depends on the input" (W. v. d. Aalst et al., 2012, p. 8).

The *Manifesto* refers some examples of events generating logs that could be submitted to process mining, like "withdrawal of cash from an ATM, a doctor adjusting an X-ray machine, a citizen applying for a driver license, the submission of a tax declaration, and the receipt of an e-ticket number by a traveller" (W. v. d. Aalst et al., 2012, p. 15).

As we said before, there are three basic types of process mining approach, as we can see in Figure 1.

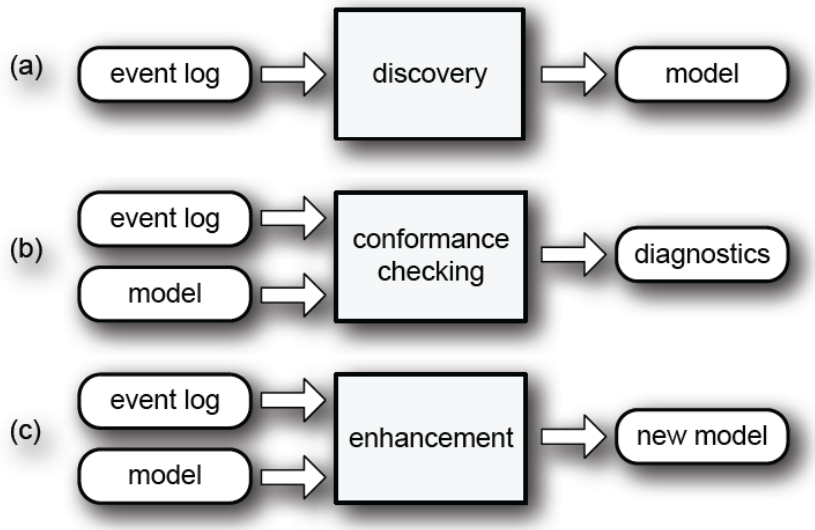


Figure 1 - The three basic types of Process Mining explained in terms of input and output; (a) discovery, (b) conformance checking, and (c) enhancement. (W. v. d. Aalst et al., 2012, p. 5)

When one is going through *Discovery*, there is no a-priori model, the purpose is to find the processes, perhaps "hidden" within the data, and although process mining should not be reduced do the discovery of control-flow.

The nest phase, *Conformance checking*, does already have an a-priori model, resulting from the previous *discovery*.

The *Enhancement* phase, sometimes referred also as *extension*, also has a model from the second phase, and, as the name indicates, will try to improve it.

Aalst (2006, p. 5) identifies 5 dimensions in PM techniques (the presence of a-priori model, the functional/non-functional properties, the number of process instances involved, the period of time considered and the result type), and argues that they are all orthogonal, meaning that they can vary independently; so, he concludes, there will be, at least, 180 different classes of PM techniques!

A few years ago, Turner, Tiwari, Olaiya, and Xu (2012, p. 502) conducted a research on software tools capable of "constructing a process model or a flow-chart", and found only a few: Futura Reflect, Fluxicon, Comprehend, ARIS PPM, BPM one, Iontas Focus Suite, Fujitsu APDS (idem, p. 503). They concluded that, although 85% of the commercial business process software claimed to address the discovery phase, only 40% claim to deal with the data noise.

Another important perspective is brought to us by Vera-Baquero, Colomo-Palacios and Molloy (2016, p. 793): some processes change so rapidly than there is no time to a "normal" way of reassessing, reformulating and reengineering the process architecture within an organization. That is also an opportunity there for process mining:

> The combination of BI and business activity monitoring (BAM) technologies may provide mechanisms to infer knowledge about business performance, but these are not sufficient for answering most of the demanding questions of today's business users. There currently exists an increasing demand for more advanced analytics such as root cause analysis of performance issues, predictive analysis and the ability to perform ''what-if" type simulations. These features are powerful assets for analysts, expanding their knowledge beyond the limits of what current platforms typically offer.
>
> (Vera-Baquero et al., 2016, p. 794)

The authors also underline the limitations of these platforms, since they "are normally business domain specific and have not been sufficiently process-aware to support the needs of process improvement type activities".

Recent works have been published about testing and evaluation of the process mining algorithms. Without this validation, it is not possible to know if the "unleashed processes" are indeed relevant. Practitioners and people involved in Business Process Management (BPM) field consider Business Process Model and Notation (BPMN) is a de-facto standard (today in version 2.0).

Mitsyuk, Shugurov, Kalenkova, & van der Aalst (2017) proposed what they call "propose a formal token-based executable BPMN semantics" and a tool that allows simulation of process models and compare the generated data with "real" one.

## 6. CHALLENGES

The PM approach is not without potential problems and faces several challenges to become effective.

One important point in favour of PM is the incredible growth of data, collected, as we have already seen, from an explosive number of digital devices, combined with the fast drop of storage prices. Some years ago, most of that data would have been discarded, because of lack of space or the costs involved, but now it can be (and it mostly is) kept, raising, as we well know, problems of privacy – for instance, on December 2016, the European Court of Justice (Joined Cases C-203/15 & C-698/15) delivered a judgment striking down laws regarding communications data retention for long periods (as much as 12 months, in some countries) as inconsistent with the Charter of Fundamental Rights of the European Union. But not all the data is human-identifiable and most process mining applications are not targeting a specific person. The purpose is to find *trends*, like what is happening when millions of citizens buy tickets for flying.

As it was pointed out, data quality is a problem, due to several causes: misfiling by humans, when they are part of the process, bad input protection in the software, different data formats, data sources not sharing the same identifiers, incomplete data, noise in the data, different types of granularity – all this is already known from areas like *data-warehousing* and *data-mining*, and the underlying process of ETL (Extract, Transform, Load). Attention should be payed to the *maturity* of event logs and they should be treated as "first-class citizens": they need to be trustworthy, complete reliable.

The quality of the event logs lead the Manifesto (van der Aalst et al., 2012, p.179) to enunciate 6 Guiding Principles (GP):

| **GUIDING PRINCIPLES** | |
|---|---|
| GP1 | Event Data Should Be Treated as First-Class Citizens |
| GP2 | Log Extraction Should Be Driven by Questions |
| GP3 | Concurrency, Choice and Other Basic Control-Flow Constructs Should Be Supported |
| GP4 | Events Should Be Related to Model Elements |
| GP5 | Models Should Be Treated as Purposeful Abstractions of Reality |
| GP6 | Process Mining Should Be a Continuous Process |

Table 3 - *Manifesto's* Guiding Principles for PM (van der Aalst et al., 2012, p.179)

What is the meaning of those principles? Let's take a closer look, for instance, to GP1. Anyone who is acquainted with large databases, especially if they are manually filled by humans, has noticed incompleteness, inconsistency and several other types of errors. To treat event data as *first-class citizens* stresses the importance of the implementation of procedures to increase data quality, like assuring that the event really happened, preventing the insertion of wrong data, or leaving unfilled important fields, or different ways for registering similar data. Privacy and security concerns should also be addressed.

The *Manifesto* (van der Aalst et al., 2012, p. 185) list 11 challenges to PM, and the above paragraph regarding data quality is only about the first. The complete set is in Table 4.

| **CHALLENGES IN PM** | |
|---|---|
| C1 | Finding, Merging, and Cleaning Event Data |
| C2 | Dealing with Complex Event Logs Having Diverse Characteristics |
| C3 | Creating Representative Benchmarks |
| C4 | Dealing with Concept Drift |
| C5 | Improving the Representational Bias Used for Process Discovery |
| C6 | Balancing between Quality Criteria Such as Fitness, Simplicity, Precision, and Generalization |
| C7 | Cross-Organizational Mining |
| C8 | Providing Operational Support |
| C9 | Combining Process Mining with other Types of Analysis |
| C10 | Improving Usability for Non-experts |
| C11 | Improving Understandability for Non-experts |

Table 4 - Manifesto's challenges in PM (2012, p. 185)

It is out of the scope of this work to be exhaustive about all these challenges, but one can count among them with the diversity of event logs, the drifting of concepts (meaning the changing of the process itself while being mined), and the balancing of quality criteria, the usability and understandability for non-experts.

## 7. CONCLUSIONS

With this paper it was possible to gather interesting information of one perhaps less known approach of BPM, which leads to the discovery and refining a model from today widely present event logs. These logs are generated in everyday life from digital artefacts, including, but not restricted, to computers and more complex information systems.

We described the origins and development of process mining in the context of the BPM movement, and we pointed out concepts and major authors, as well of their work. A comparison of

the understanding of these concepts among several authors is made, even if it recognized the overall convergence of them, as we concluded.

The methodology stressed the importance of key concepts and how this review was conducted.

Techniques and tools were described and, based in the contributions of the literature, we concluded that process mining is facing great challenges, with almost 200 techniques to explore and a just a small set of tools commercially available.

In spite of the relative youth of this area, there is an increasing number of papers, books and other works being issued.

For future work, it would be important to do some systematic area classification within process mining, and also to deepen the comparative analysis regarding the results obtained by all these techniques. The last, but not the least, *predictive models* deserve a close monitoring, being one of the newest and promising approaches of PM.

## REFERENCES

Aufaure, M. A., & Zimányi, E. (2013). *Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures*: Springer Berlin Heidelberg.

Breuker, D., Matzner, M., Delfmann, P., & Becker, J. (2016). Comprehensible Predictive Models for Business Processes. *MIS Quarterly, 40*(4), 1009-1034.

Caldeira, J., & Abreu, F. B. (2016). *Software Development Process Mining: Discovery, Conformance Checking and Enhancement*. Paper presented at the 2016 10th International Conference on the Quality of Information and Communications Technology (QUATIC).

Coutinho, C. P. (2014). *Metodologia de Investigação em Ciências Sociais: Teoria e Prática* (2ª ed.). Coimbra: Almedina.

Dumas, M., Rosa, M. L., Mendling, J., & Reijers, H. (2013). *Fundamentals of Business Process Management*: Springer-Verlag Berlin Heidelberg.

Dustdar, S., Hoffmann, T., & van der Aalst, W. (2005). Mining of ad-hoc business processes with TeamLog. *Data & Knowledge Engineering, 55*, 129–158. doi:10.1016/j.datak.2005.02.002

Gartner (n.d.). *IT Glossary*. Retrieved 2017.03.08 from http://www.gartner.com/it-glossary/automated-business-process-discovery-abpd

Heames, J. T., & Breland, J. W. (2010). Management pioneer contributors: 30-year review. *Journal of Management History, 16*(4), 427-436. doi:10.1108/17511341011073915

Kohlborn, T., Müller, O., Pöeppelbuß, J., & Röglinger, M. (2014). New frontiers in business process management (BPM). *Business Process Management Journal, 20*(4). doi:10.1108/BPMJ-02-2014-0015

Munoz-Gama, J. (2014). *Conformance checking and diagnosis in process mining*. (PhD thesis), Barcelona.

Mitsyuk, A. A., Shugurov, I. S., Kalenkova, A. A., & van der Aalst, W. M. P. (2017). *Generating event logs for high-level process models. Simulation Modelling Practice and Theory, 74*, 1-16. doi:http://dx.doi.org/ 10.1016/j.simpat.2017.01.003

Okoye, K., Tawil, A. R. H., Naeem, U., & Lamine, E. (2015). *Semantic Process Mining Towards Discovery and Enhancement of Learning Model Analysis.* Paper presented at the 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems.

Song, M., Günther, C. W., & van der Aalst, W. M. P. (2009). Trace Clustering in Process Mining. In D. Ardagna, M. Mecella, & J. Yang (Eds.), *Business Process Management Workshops: BPM 2008 International Workshops, Milano, Italy, September 1-4, 2008. Revised Papers* (pp. 109-120). Berlin, Heidelberg: Springer Berlin Heidelberg.

Tiwari, A., Turner, C. J., & Majeed, B. (2008). A review of business process mining: state-of-the-art and future trendsnull. *Business Process Management Journal, 14*(1), 5-22. doi:10.1108/14637150810849373

Turner, C. J., Tiwari, A., Olaiya, R., & Xu, Y. (2012). Process mining: From theory to practice. *Business Process Management Journal, 18*(3), 493-512. doi:10.1108/14637151211232669

van der Aalst, W. P. (2006). Process Mining and Monitoring Processes and Services: Workshop Report. In F. Leymann, W. Reisig, S. R. Thatte, & W. v. d. Aalst (Eds.), *The Role of Business Processes in Service Oriented Architectures* (Vol. 06291-834). Schloss Dagstuhl, Germany: Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

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

van der Aalst, W. M. P. (2014). Process Mining in the Large: A Tutorial. In E. Zimányi (Ed.), *Business Intelligence: Third European Summer School, eBISS 2013, Dagstuhl Castle, Germany, July 7-12, 2013, Tutorial Lectures* (pp. 33-76). Cham: Springer International Publishing.

van der Aalst, W. P., Adriansyah, A., Medeiros, A. K. A. d., Arcieri, F., Baier, T., Blickle, T., . . . Wynn, M. (2012). Process Mining Manifesto. In F. Daniel, K. Barkaoui, & S. Dustdar (Eds.), *Business Process Management Workshops* (Vol. 99, pp. 169-194). Campus des Cézeaux, Clermont-Ferrand: Springer Berlin Heidelberg.

van der Aalst, W. P., ter Hofstede, A. M., & Weske, M. (2003). Business Process Management: A Survey. In W. P. van der Aalst & M. Weske (Eds.), *Business Process Management* (Vol. 2678, pp. 1-12): Springer Berlin Heidelberg.

Vera-Baquero, A., Colomo-Palacios, R., & Molloy, O. (2016). Real-time business activity monitoring and analysis of process performance on big-data domains. *Telematics and Informatics*, 33, 793-807. doi:10.1016/j.tele.2015.12.005

Weijters, A. J. M. M., & Aalst, W. M. P. v. d. (2001). *Process Mining: Discovering Workflow Models from Event-Based Data*. Paper presented at the 13th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC 2001), BNVKI, Maastricht. http://www.processmining.org/blogs/pub2001/process_mining_ discovering_workflow_models_from_event-based_data.