"MINE YOUR OWN BUSINESS": USING PROCESS MINING TO TURN BIG DATA INTO REAL VALUE

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

Abstract

Like most IT-related phenomena, also the growth of event data complies with Moore's Law. Similar to the number of transistors on chips, the capacity of hard disks, and the computing power of computers, the digital universe is growing exponentially and roughly doubling every 2 years. Although this is not a new phenomenon, suddenly many organizations realize that increasing amounts of "Big Data" (in the broadest sense of the word) need to be used intelligently in order to compete with other organizations in terms of efficiency, speed and service. However, the goal is not to collect as much data as possible. The real challenge is to turn event data into valuable insights. Only process mining techniques directly relate event data to end-to-end business processes. Existing business process modeling approaches generating piles of process models are typically disconnected from the real processes and information systems. Data-oriented analysis techniques (e.g., data mining and machines learning) typically focus on simple classification, clustering, regression, or rule-learning problems. This keynote paper provides pointers to recent developments in process mining thereby clearly showing that process mining provides a natural link between processes and data on the one hand and performance and compliance on the other hand.

Keywords: Process Mining, Process Discovery, Conformance Checking, Business Process Management.

1 Big Data as the Fuel for Mining Your Own Business

Recently, process mining emerged as a new scientific discipline on the interface between process models and event data (Van der Aalst, 2011). On the one hand, conventional Business Process Management (BPM) and Workflow Management (WfM) approaches and tools are mostly model-driven with little consideration for event data. On the other hand, Data Mining (DM), Business Intelligence (BI), and Machine Learning (ML) focus on data without considering end-to-end process models, cf. (Mitchell, 1997) and (Hand, Mannila, and Smyth, 2001). Process mining aims to bridge the gap between BPM and WfM on the one hand and DM, BI, and ML on the other hand. Here, the challenge is to turn torrents of event data ("Big Data") into valuable insights related to performance and compliance. Fortunately, process mining results can be used to identify and understand bottlenecks, inefficiencies, deviations, and risks. Process mining helps organizations to "mine their own business", i.e., they are enabled to discover, monitor and improve real processes by extracting knowledge from event logs.

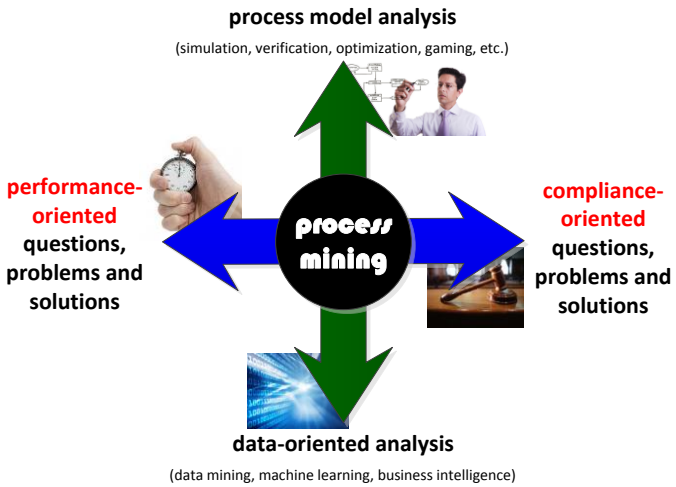


Figure 1. Process mining can be positioned as the missing link between process model analysis and data-oriented analysis. Process mining is not limited to automated process discovery based on event data: it can be used to answer a wide variety of performance and compliance questions in a unified and integrated manner.

As shown by Hilbert and Lopez (2011), our increasing capabilities to process and store data are undeniable. This will change the way operational processes can be analyzed and improved. This can be illustrated as follows. Consider a typical 1 TB hard disk purchased in 2010. The disk can store 10^{12} bytes (i.e., one Terabyte). According to IDC, the entire "Digital Universe" was 1.2 Zettabyte (1.2×10^{21} bytes) at that time. This estimate taken from IDC's annual report "The Digital Universe Decade: Are You Ready?" published in May 2010. Hence, the 1 TB disk needs to grow $2^{30.16} = 1.2 \times 10^{21} / 10^{12}$ times. Based on the average growth rate of hard disks over the last decades and an extrapolation of Moore's law, we assume that hard disks indeed double every 1.56 years (like in the past 40 years). This implies that in $30.16 \times 1.56 = 47.05$ years a standard hard disk may contain the whole "Digital Universe" of 2010. This includes the entire internet, all computer files, transaction logs, movies, photos, music, books, databases, a scientific data, etc. This simple calculation exemplifies the increasing relevance of data for process analysis by simply assuming a continuing growth of event data in the next decennia. It is obvious to see that business processes will generate more and more

event data that can be used for analysis. Detailed transaction data and sensor data (cf. RFID tags) will enable new process mining applications replacing traditional analysis based on hand-made models (Van der Aalst, 2011).

Since the McKinsey report "Big Data: The Next Frontier for Innovation, Competition, and Productivity" (Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh, and Byers, 2011), the term "Big Data" appeared on the radar of all larger organizations. Consultants, software suppliers, and IT specialists have high-jacked the term and all define "Big Data" in a different manner. In scientific computing, large scale experiments like the discovery of the Higgs-particle by CERN's Large Hadron Collider (LHC) are considered as primary examples of Big Data. The four detectors of the LHC-particle collider produce in the order of fifteen petabytes (15×10^{15} bytes) per year, i.e., the equivalent of three million dvd's. However, Facebook, Google, and YouTube are managing even larger data sets. The desire to manage huge datasets has resulted in key technologies such as the Hadoop programming framework (inspired by Google's MapReduce). Data is the fuel for new analysis techniques and people like to brag about the volume of data being stored and analyzed. In fact, sometimes people lose track of the original objectives. If data is the "fuel" of analysis, it cannot be the goal to consume as much data as possible. Instead, the focus should be on the efficient and effective use of data (mileage and speed).

2 Elephant Trails in Big Data

The starting point for process mining is not just any data, but *event* data (IEEE Task Force on Process Mining, 2012). Data should refer to discrete events that happened in reality. A collection of related events is referred to as 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. It is important to note that an event log contains only *example behavior*, i.e., we cannot assume that all possible runs have been observed. In fact, an event log often contains only a fraction of the possible behavior (Van der Aalst, 2011). Often event logs store additional information about events and these additional data attributes may be used during analysis. For example, many 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 viewed as "Olifantenpaadjes". This is the Dutch word for "elephant trails" commonly known as *desire lines*. Desire lines refer to tracks worn across grassy spaces - where people naturally walk - regardless of formal pathways. A desire line emerges through erosion caused by footsteps of humans (or animals) and the width and degree of erosion of the path indicates how frequently the path is used. Typically, the desire line follows the shortest or most convenient path between two points. Moreover, as the path emerges more people are encouraged to use it, thus stimulating further erosion. Dwight Eisenhower is often mentioned as one of the persons using this emerging group behavior. Before becoming the 34th president of the United States, he was the president of Columbia University. When he was asked how the university should arrange the sidewalks to best interconnect the campus buildings, he suggested letting the grass grow between buildings and delay the creation of sidewalks. After some time the desire lines revealed themselves. The places where the grass was most worn by people's footsteps were turned into sidewalks.

The *digital desire lines* recorded in event logs may be very different from formal procedures or expected behavior (i.e., the "sidewalks" in processes). As more events are recorded, it becomes possible to determine desire lines in organizations, systems, and products. Besides visualizing such desire lines, we can also investigate how these desire lines change over time, characterize the people following a particular desire line, etc. Desire lines may reveal behaviors that are "undesirable" (unsafe, inefficient, unfair, etc.) and used for auditing and compliance purposes (Van der Aalst, Van Hee, Van

der Werf, and Verdonk, 2010). Uncovering such phenomena is a prerequisite for process and product improvement. Process mining can be used to redesign procedures and systems ("reconstructing the formal pathways"), to recommend people taking the right path ("adding signposts were needed"), or to build in safeguards ("building fences to avoid dangerous situations").

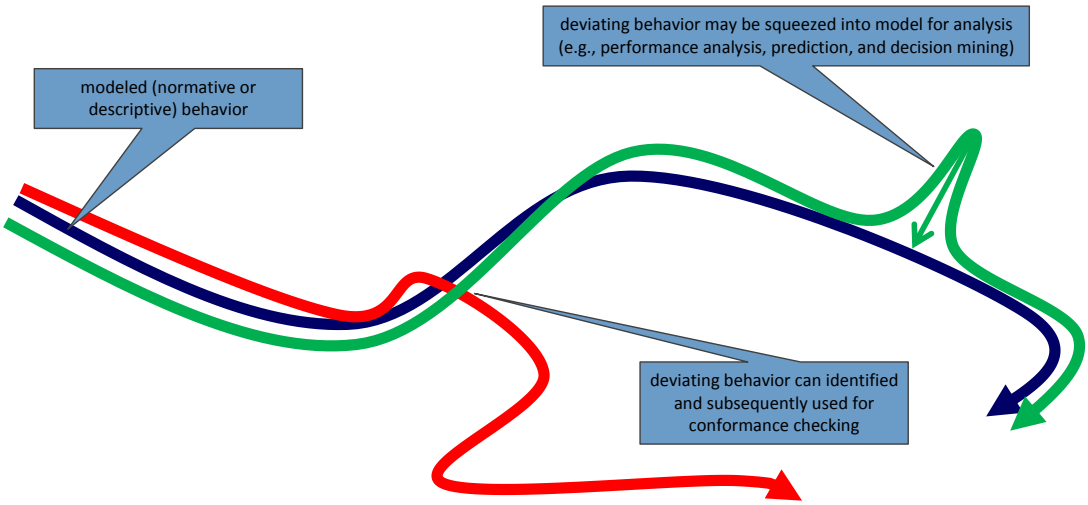
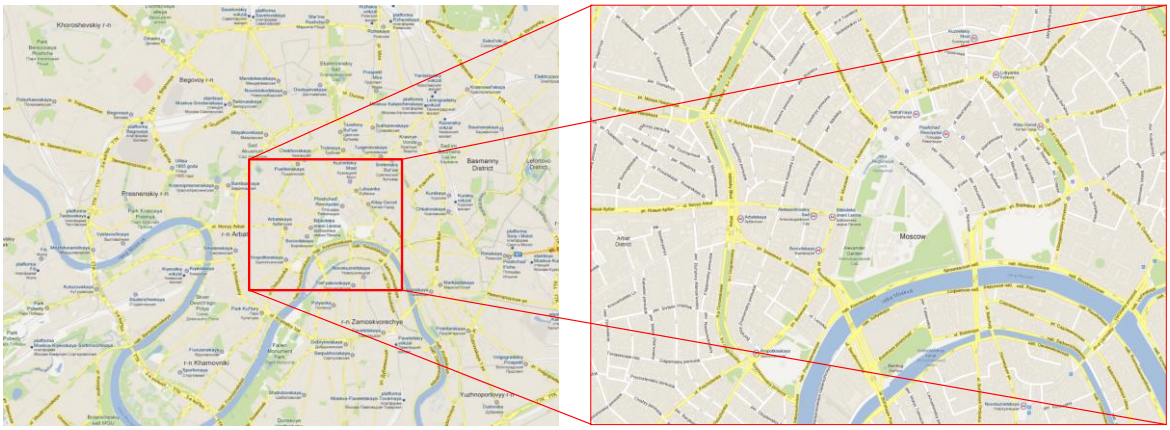


Figure 2. Process mining aligns observed and modeled behavior: "moves" seen in reality are related to "moves" in the model (if possible).

One of the key contributions of process mining is its *ability to relate observed and modeled behavior at the event level*, i.e., traces observed in reality (process instances in event log) are aligned with traces allowed by the model (complete runs of the model). As shown in Figure 2 it is useful to *align* both even when model and reality disagree. First of all, it is useful to highlight where and why there are discrepancies between observed and modeled behavior. Second, deviating traces need to be "squeezed" into the model for subsequent analysis, e.g., performance analysis or predicting remaining flow times. The latter is essential in case of non-conformance (Van der Aalst, Adriansyah, and Van Dongen 2012). Without aligning model and event log, subsequent analysis is impossible or biased towards conforming cases.

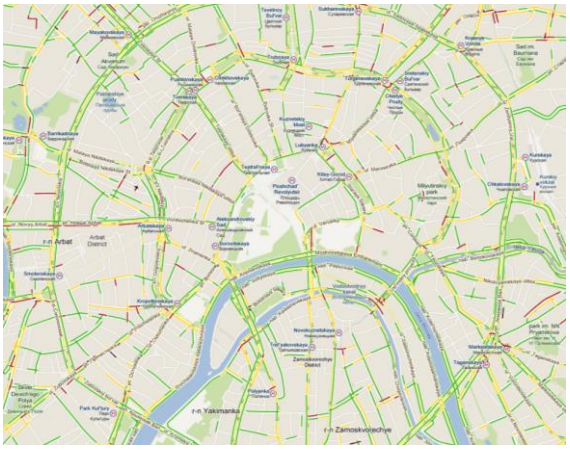


(a) map of Moscow (b) zooming in on the center of Moscow

(c) satellite view of center

(d) subway map of Moscow

(e) photos projected on map



(f) traffic jams projected on map

Figure 3. Process models should be viewed as maps (like in Google Maps). Typically, there are multiple useful maps showing the same physical reality. Moreover, it should be possible to seamlessly zoom-in and project information (e.g., traffic congestion) onto maps.

The desire line metaphor suggests that we can view process models as *maps*. Often modelers aim to create one "perfect process model" that needs to serve all possible purposes. However, a process model is merely a view on the real process. Depending on the questions that need to be answered,

different views may be needed. There may be highway maps, subway maps, city maps, bicycle maps, boating maps, and hiking maps covering (parts of) the same area. Some elements may not be shown while other elements are emphasized. Some maps may show a larger area with less detail (cf. Fig. 3a) whereas other maps show a smaller area with more details (cf. Fig. 3b). When using the subway another map is desired (cf. Fig. 3d). It is also possible to map current or historic information on maps. For example, Fig. 3f shows the traffic jams in the center of Moscow on a Monday morning in April 2013. These examples illustrate that depending on the intended purpose (discussion, bottleneck analysis, auditing, simulation, etc.), different process models are needed.

3 Use Cases Related to Process Mining

To conclude this keynote paper, we discuss the main BPM use cases related to process mining. In (Van der Aalst, 2013) *twenty use cases* are used to structure the BPM discipline and to show "how, where, and when" BPM techniques can be used. These are summarized in Fig. 4. *Models* are depicted as pentagons marked with the letter **M**. A model may be descriptive (**D**), normative (**N**), and/or executable (**E**). A "**D|N|E**" tag inside a pentagon means that the corresponding model is descriptive, normative, or executable. *Configurable models* are depicted as pentagons marked with **CM**. *Event data* (e.g., an event log) are denoted by a disk symbol (cylinder shape) marked with the letter **E**. *Information systems* used to support processes at runtime are depicted as squares with rounded corners and marked with the letter **S**. *Diagnostic information* is denoted by a star shape marked with the letter **D**. We distinguish between *conformance-related diagnostics* (star shape marked with **CD**) and *performance-related diagnostics* (star shape marked with **PD**). The twenty atomic use cases can be chained together in so-called *composite* use cases. These composite cases can be used to describe realistic BPM scenarios.

In (Van der Aalst, 2013), BPM literature is analyzed to see trends in terms of the twenty use cases, e.g., topics that are getting more and more attention. Here we only mention the use cases most related to process mining.

- Use case *Log Event Data* (LogED) refers to the recording of event data, often referred to as event logs. Such event logs are used as input for various process mining techniques. XES (extensible event stream), the successor of MXML (mining XML format), is a standard format for storing event logs (www.xes-standard.org).
- Use case *Discover Model from Event Data* (DiscM) refers to the automated generation of a process model using process mining techniques. Examples of discovery techniques are the alpha algorithm (Van der Aalst, Weijters, and Maruster, 2004), language-based regions (Werf, Van Dongen, Hurkens, and Serebrenik, 2010), and state-based regions (Carmona, Cortadella, and Kishinevsky, 2008). Note that classical synthesis approaches (Darondeau, 2004) need to be adapted since the event log only contains examples.
- Use case *Check Conformance Using Event Data* (ConfED) refers to all kinds of analysis aiming at uncovering discrepancies between modeled and observed behavior. Conformance checking may be done for auditing purposes, e.g., to uncover fraud or malpractices. Token-based (Rozinat and Van der Aalst, 2008) and alignment-based (Van der Aalst, Adriansyah, and Van Dongen, 2012) techniques replay the event log to identify non-conformance (Weerdt, De Backer, Vanthienen, and Baesens, 2011).
- Use case *Analyze Performance Using Event Data* (PerfED) refers to the combined use of models and timed event data. By replaying an event log with timestamps on a model, one can measure delays, e.g., the time in-between two subsequent activities. The results of timed replay can be used to highlight bottlenecks. Moreover, the gathered timing information can be used for simulation or prediction techniques (Rozinat, Mans, Song, and Van der Aalst, 2009).

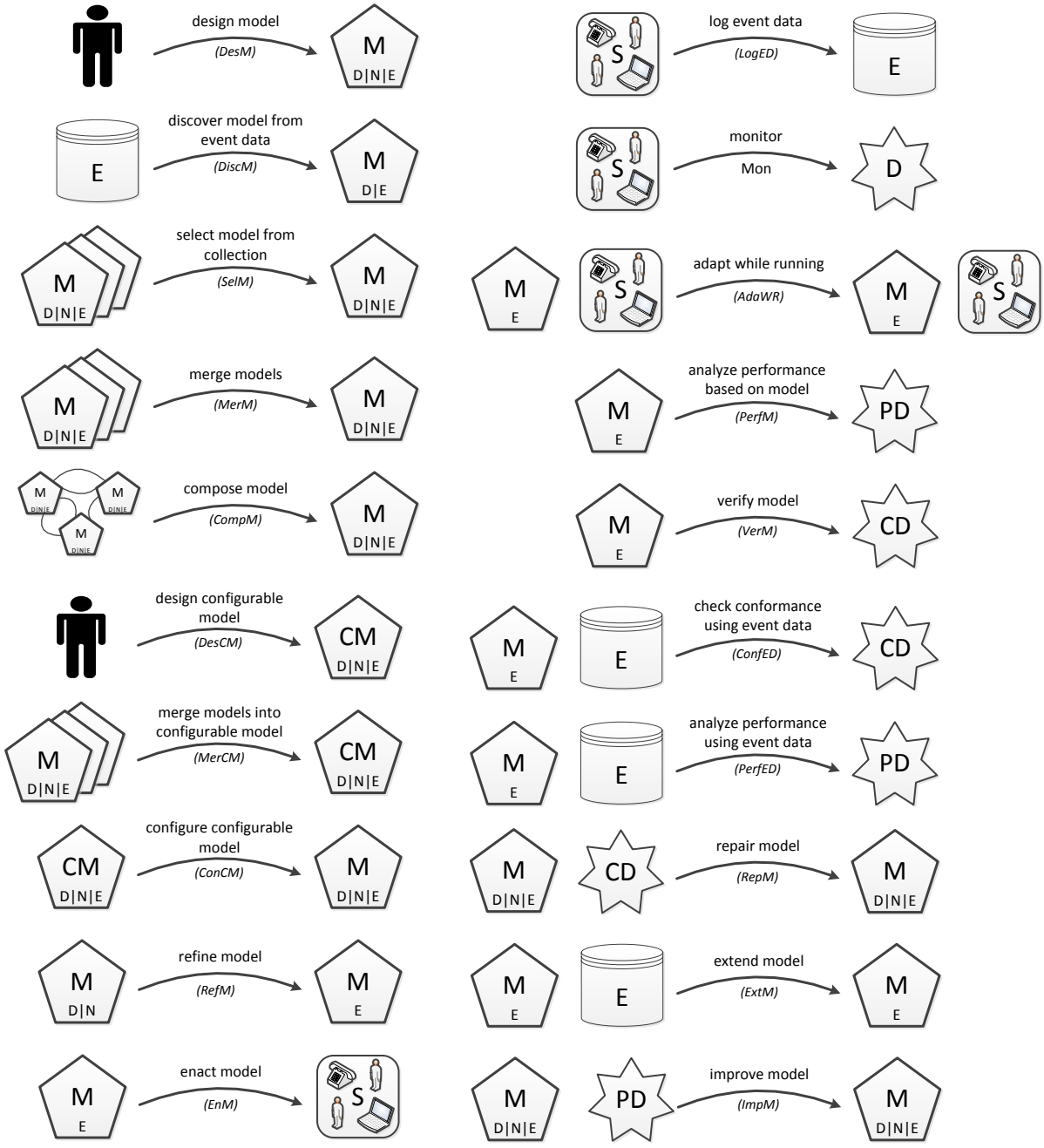


Figure 4. Twenty BPM use cases (Van der Aalst, 2013). Use cases Log Event Data (LogED), Discover Model from Event Data (DiscM), Check Conformance Using Event Data (ConfED), Analyze Performance Using Event Data (PerfED), Repair Model (RepM), Extend Model (ExtM), Improve Model (ImpM) are most related to process mining.

- Use case *Repair Model* (RepM) uses the diagnostics provided by use case ConfED to adapt the model such that it better matches reality. On the one hand, a process model should correspond to the observed behavior. On the other hand, there may be other forces influencing the desired target model, e.g., a reference model, desired normative behavior, and domain knowledge.
- Event logs refer to activities being executed and events may be annotated with additional information such as the person/resource executing or initiating the activity, the timestamp of

the event, or data elements recorded with the event. Use case *Extend Model* (ExtM) refers to the use of such additional information to enrich the process model. For example, timestamps of events may be used to add delay distributions to the model. Data elements may be used to infer decision rules that can be added to the model. Resource information can be used to attach roles to activities in the model (Rozinat, Wynn, Van der Aalst, Ter Hofstede, Fidge, 2009).

- Use case *Improve Model* (ImpM) uses the performance related diagnostics obtained through use case PerfED. ImpM is used to generate alternative process models aiming at process improvements, e.g., to reduce costs or response times. These models can be used to do ``what-if'' analysis. Note that unlike RepM the focus ImpM is on improving the process itself.

4 Min(d) Your Own Business

The phrase "Mind your own business" is a common English saying suggesting people to focus on their own affairs rather than prying into the lives of others. In this keynote paper, the phrase is used to encourage the reader to apply process mining techniques to the event data that can be found for any operational process. The torrents of event data available in most organizations enable *evidence-based Business Process Management* (ebBPM). We predict that there will be a remarkable shift from pure model-driven or questionnaire-driven approaches to data-driven process analysis as we are able to monitor and reconstruct the real business processes using event data. See (Van der Aalst, 2011) for techniques supporting this shift. Note that the current version of ProM holds over 550 plug-ins. Each plug-in provides some analysis capability, e.g., discovering a Petri net from event data or animating historic data on a fuzzy model.

References

Carmona, J., Cortadella, J., and Kishinevsky, M. (2008). A Region-Based Algorithm for Discovering Petri Nets from Event Logs. In Business Process Management (BPM2008). 358-373.

Darondeau, P. (2004). Unbounded Petri Net Synthesis. In Lectures on Concurrency and Petri Nets, J. Desel, W. Reisig, and G. Rozenberg, Eds. Lecture Notes in Computer Science Series, vol. 3098. Springer-Verlag, Berlin, 413-438.

Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M., and Shan, M. (2004). Business Process Intelligence. Computers in Industry 53, 3, 321-343.

Hand, D., Mannila, H., and Smyth, P. (2001). Principles of Data Mining. MIT press, Cambridge, MA.

Hilbert, M. and Lopez, P. (2011) The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025), 60-65.

IEEE Task Force on Process Mining (2012). Process Mining Manifesto. In F. Daniel, K. Barkaoui, and S. Dustdar, editors, Business Process Management Workshops, volume 99 of Lecture Notes in Business Information Processing, pages 169-194. Springer-Verlag, Berlin.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute.

Mitchell, T. (1997). Machine Learning. McGraw-Hill, New York.

Rozinat, A. and Van der Aalst, W.M.P. (2008). Conformance Checking of Processes Based on Monitoring Real Behavior. Information Systems, 33(1), 64-95.

Rozinat, A., Wynn, M., Van der Aalst, W.M.P., Ter Hofstede, A., Fidge, C. (2009). Workflow Simulation for Operational Decision Support. Data and Knowledge Engineering, 68(9), 834-850.

Rozinat, A., Mans, R., Song, M., and Van der Aalst, W.M.P. (2009). Discovering Simulation Models. Information Systems, 34(3), 305-327.

Weerdt, J., M. De Backer, Vanthienen, J., and Baesens, B. (2011). A Robust F-measure for Evaluating Discovered Process Models. In IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2011), N. Chawla, I. King, and A. Sperduti, Eds. IEEE, Paris, France, 148-155.

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

Van der Aalst, W.M.P. (2013). Business Process Management: A Comprehensive Survey. ISRN Software Engineering, doi:10.1155/2013/507984, 1-37.

Van der Aalst, W.M.P. , Adriansyah, A., and Van Dongen, B. (2012). Replaying History on Process Models for Conformance Checking and Performance Analysis. WIREs Data Mining and Knowledge Discovery, 2(2), 182-192.

Van der Aalst, W.M.P., Van Hee, K.M., Van der Werf, J.M. and Verdonk, M. (2010). Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor. IEEE Computer, 43(3):90-93.

Van der Aalst, W., Weijters, A., and Maruster, L. (2004). Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering 16(9), 1128-1142.

Werf, J., Van Dongen, B., Hurkens, C., and Serebrenik, A. (2010). Process Discovery using Integer Linear Programming. Fundamenta Informaticae 94, 387-412.