Spreadsheets for Business Process Management

Using process mining to deal with "events" rather than "numbers"

Wil M.P. van der Aalst

Preprint, to appear in Business Process Management Journal (BPMJ)
http://www.emeraldinsight.com/journal/bpmj
Paper received: 3 October 2016 / Paper accepted: 15 January 2017

Abstract

Purpose - Process mining provides a generic collection of techniques to turn event data into valuable insights, improvement ideas, predictions, and recommendations. This paper uses spreadsheets as a metaphor to introduce process mining as an essential tool for data scientists and business analysts. The goal is to illustrate that process mining can do with events what spreadsheets can do with numbers.

Design/methodology/approach - The paper discusses the main concepts in both spreadsheets and process mining. Using a concrete data set as a running example the different types of process mining are explained. Where spreadsheets work with numbers, process mining starts from event data with the aim to analyze processes.

Findings - Differences and commonalities between spreadsheets and process mining are described. Unlike process mining tools like ProM, spreadsheets programs cannot be used to discover processes, check compliance, analyze bottlenecks, animate event data, and provide operational process support. Pointers to existing process mining tools and their functionality are given.

Practical implications - Event logs and operational processes can be found everywhere and process mining techniques are not limited to specific application domains. Comparable to spreadsheet software widely used in finance,

Wil M.P. van der Aalst
Process and Data Science (PADS), Department of Computer Science, RWTH Aachen University, 52056 Aachen, Germany.
E-mail: wvdaalst@pads.rwth-aachen.de

production, sales, education, and sports, process mining software can be used in a broad range of organizations.

Originality/value - The paper provides an original view on process mining by relating it to spreadsheets. The value of spreadsheet-like technology tailored towards the analysis of behavior rather than numbers is illustrated by the over 20 commercial process mining tools available today and the growing adoption in a variety of application domains.

Keywords - Process mining, Business Process Management (BPM), Spreadsheets, Data science

1 Introduction

Spreadsheets are used everywhere. A spreadsheet is composed of cells organized in rows and columns. Some cells serve as input, other cells have values computed over a collection of other cells (e.g., taking the sum over an array of cells). *VisiCalc* was the "killer application" for the Apple II computer in 1979 and *Lotus 1-2-3* played a comparable role for the IBM PC in 1983. People were buying these computers in order to run spreadsheet software [Ceruzzi, 2003]: A nice example of the "tail" (*VisiCalc*/*Lotus 1-2-3*) wagging the "dog" (Apple-II/IBM PC). After decades of spectacular IT-developments, spreadsheet software can still be found on most computers (e.g. *Excel* is part of Microsoft's *Office*) and can be accessed online (e.g., *Google Sheets* as part of *Google Docs*). Spreadsheet software survived 50 years of IT-developments because spreadsheets are highly generic and valuable for many. The situations in which spreadsheets can be used in a meaningful way are almost endless [Jelen, 2005]. *Spreadsheets can be used to do anything with numbers.* Of course one needs to write dedicated programs if computations get complex or use database technology if data sets get large. However, for the purpose of this paper we assume that spreadsheets adequately deal with numerical data. We would like to argue that *process mining software enables users to do anything with events.* In this paper, we introduce process mining against the backdrop of spreadsheets.

Instead of *numbers* we consider discrete *events*, i.e., things that have happened and could be recorded. Events may take place inside a machine (e.g., an ATM or baggage handling system), inside an enterprise information system (e.g., a purchase decision or salary payment), inside a hospital (e.g., making an X-ray), inside a social network (e.g., sending a twitter message), inside a transportation system (e.g., checking in at an airport), etc. Events may be "life events", "machine events", or "organization events". The term *Internet of Events* (IoE), coined in [Aalst, 2014], refers to all event data available. The IoE is roughly composed of the Internet of Content (IoC), the Internet of People (IoP), Internet of Things (IoT), and Internet of Locations (IoL). These are overlapping, e.g., a tweet sent by a mobile phone from a particular location is

in the intersection of IoP and IoL. *Process mining aims to exploit event data in a meaningful way*, for example, to provide insights, identify bottlenecks, anticipate problems, record policy violations, recommend countermeasures, and streamline processes [Aalst, 2016].

Process mining should be in the toolbox of data scientists, business analysts, and others who need to analyze event data. Unfortunately, process mining is not yet a widely adopted technology. Surprisingly, the process perspective is absent in the majority of Big Data initiatives and data science curricula. We argue that event data should be used to improve *end-to-end* processes: It is not sufficient to consider "numbers" and isolated activities. Data science approaches tend to be process agonistic whereas Business Process Management (BPM) approaches tend to be model-driven without considering the "evidence" hidden in the data [Aalst, 2013].

Developments in BPM have resulted in a well-established set of principles, methods and tools that combine knowledge from information technology, management sciences and industrial engineering for the purpose of improving business processes [Weske, 2007, Aalst, 2013, Dumas et al., 2013]. BPM can be viewed as a continuation of the Workflow Management (WFM) wave in the the nineties. The maturity of WFM/BPM is partly reflected by a range of books:

- [Jablonski & Bussler, 1996] (first comprehensive WFM book focusing on the different workflow perspectives and the MOBILE language),
- [Leymann & Roller, 1999] (book on production WFM systems closely related to IBM's workflow products),
- [Aalst et al., 2000] (edited book that served as the basis for the BPM conference series),
- [Aalst & Hee, 2004] (most cited WFM book; a Petri net-based approach is used to model, analyze and enact workflow processes),
- [Muehlen, 2004] (book relating WFM systems to operational performance),
- [Dumas et al., 2005] (edited book on process-aware information systems),
- [Smith & Fingar, 2006] (visionary book linking management perspectives to the pi calculus),
- [Weske, 2007] (book presenting the foundations of BPM, including different languages and architectures),
- [Hofstede et al., 2010] (book based on YAWL and the workflow patterns),
- [Brocke & Rosemann, 2010, Brocke & Rosemann, 2014] (handbooks on Business Process Management),
- [Becker et al., 2011] (book on the design of process-oriented organizations),
- [Reichert & Weber, 2012] (book on supporting flexibility in process-aware information systems), and
- [Dumas et al., 2013] (tutorial-style book covering the whole BPM lifecycle).

As mentioned, WFM/BPM approaches tend to be model-driven. Notable exceptions are the process mining approaches developed over the last decade [Aalst, 2016].

Process mining can be seen as a means to bridge the gap between data science and classical process management (WFM/BPM) [Aalst, 2013]. By framing process mining as a spreadsheet-like technology for event data, we hope to increase awareness in the information systems community.

The remainder of this paper is organized as follows. Section 2 introduces a concrete data set which will be used as a running example. By using an easy-to-understand business setting to introduce both spreadsheets and process mining, we can explain their differences and commonalities. Section 3 summarizes the basic concepts used by spreadsheet software like *Excel* and also describes the relevance of spreadsheets in a historical context. Section 4 demonstrates that process mining technology can be positioned as spreadsheets to analyze dynamic behavior rather than numbers. Process mining techniques such as process discovery and conformance checking are illustrated using the running example. Section 5 concludes the paper.

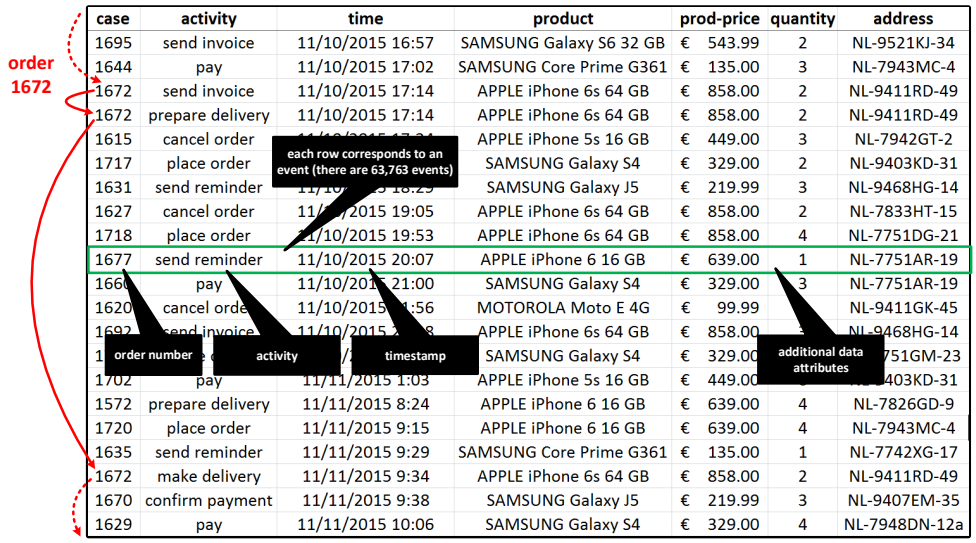
2 Running Example

As an example, let us consider the process of handling customer orders. Customers can order phones via the website of a telecom company. The customer first places an order. Multiple phones of the same type can be ordered at the same time. The customer is expected to pay before the phones are delivered. An invoice is sent to the customer, but the customer can also pay before receiving the invoice. If the customer does not pay in time, a reminder is sent. This is only done after sending the invoice. If the customer does not pay after two reminders, the order is canceled. If the customer pays, the order's delivery is prepared, followed by the actual delivery and a conformation of payment (in any order).

Figure 1 shows some event data recorded for our order handling process. Each row corresponds to an event, i.e., the execution of an activity for a particular order. The highlighted row refers to the sending of a reminder for order 1677 on 11/10/2015. There may be multiple rows (i.e., events) related to the same order. For example, the small fragment shows three events related to order 1672 (see red lines). Order 1672 consists of six events in total. This is close to the average number of events per order (6.38).

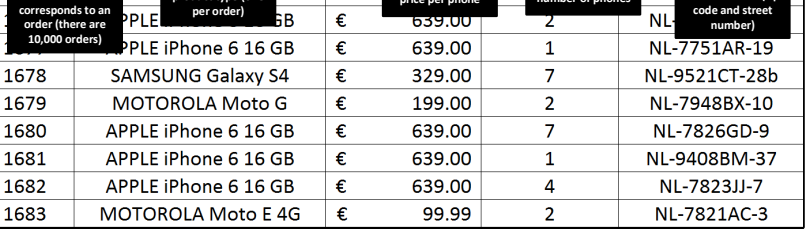
Whereas Figure 1 shows the "raw" events, Figure 2 shows more high-level data with precisely one row per order. For example, all events related to order 1672 are "collapsed" into a single row. There are 10,000 orders. Per order we can see the quantity, number of phones ordered, and a zip code with street number (plus possible suffix) uniquely identifying an address in the Netherlands.

The data sets shown in Figure 1 and Figure 2 will be used to introduce process mining techniques and to relate these to spreadsheet-based analysis.



case	activity	time	product	prod-price	quantity	address
1695	send invoice	11/10/2015 16:57	SAMSUNG Galaxy S6 32 GB	€ 543.99	2	NL-9521KJ-34
1644	pay	11/10/2015 17:02	SAMSUNG Core Prime G361	€ 135.00	3	NL-7943MC-4
1672	send invoice	11/10/2015 17:14	APPLE iPhone 6s 64 GB	€ 858.00	2	NL-9411RD-49
1672	prepare delivery	11/10/2015 17:14	APPLE iPhone 6s 64 GB	€ 858.00	2	NL-9411RD-49
1615	cancel order	[illegible]	APPLE iPhone 5s 16 GB	€ 449.00	3	NL-7942GT-2
1717	place order	[illegible]	SAMSUNG Galaxy S4	€ 329.00	2	NL-9403KD-31
1631	send reminder	[illegible]	SAMSUNG Galaxy J5	€ 219.99	3	NL-9468HG-14
1627	cancel order	11/[illegible]/2015 19:05	APPLE iPhone 6s 64 GB	€ 858.00	2	NL-7833HT-15
1718	place order	[illegible]/10/2015 19:53	APPLE iPhone 6s 64 GB	€ 858.00	4	NL-7751DG-21
1677	send reminder	11/10/2015 20:07	APPLE iPhone 6 16 GB	€ 639.00	1	NL-7751AR-19
1660	pay	11/10/201[illegible] 21:00	SAMSUNG Galaxy S4	€ 329.00	3	NL-7751AR-19
1620	cancel orde	11/10/2015 [illegible]:56	MOTOROLA Moto E 4G	€ 99.99		NL-9411GK-45
1692	send invoice	11/10/2015 2[illegible]8	APPLE iPhone 6s 64 GB	€ 858.00	3	NL-9468HG-14
1[illegible]	[illegible]	[illegible]	SAMSUNG Galaxy S4	€ 329.00		[illegible]751GM-23
1702	pay	11/11/2015 1:03	APPLE iPhone 5s 16 GB	€ 449.00		[illegible]403KD-31
1572	prepare delivery	11/11/2015 8:24	APPLE iPhone 6 16 GB	€ 639.00	4	NL-7826GD-9
1720	place order	11/11/2015 9:15	APPLE iPhone 6 16 GB	€ 639.00	4	NL-7943MC-4
1635	send reminder	11/11/2015 9:29	SAMSUNG Core Prime G361	€ 135.00	1	NL-7742XG-17
1672	make delivery	11/11/2015 9:34	APPLE iPhone 6s 64 GB	€ 858.00	2	NL-9411RD-49
1670	confirm payment	11/11/2015 9:38	SAMSUNG Galaxy J5	€ 219.99	3	NL-9407EM-35
1629	pay	11/11/2015 10:06	SAMSUNG Galaxy S4	€ 329.00	4	NL-7948DN-12a

Fig. 1 Small fragment of a larger event log holding 63,763 events.



case	product	prod-price	quantity	address
1669	SAMSUNG Galaxy S4	€ 329.00	2	NL-7942GT-2
1670	SAMSUNG Galaxy J5	€ 219.99	3	NL-9407EM-35
1671	APPLE iPhone 5s 16 GB	€ 449.00	4	NL-7944RD-8
1672	APPLE iPhone 6s 64 GB	€ 858.00	2	NL-9411RD-49
167[illegible]	APPLE iPh[illegible]e 5s 16 GB	€ [illegible]49.00	4	NL-78[illegible]7AC-13
1674	SAMSUNG Ga[illegible]y S6 32 GB	€ 5[illegible].99	5	NL-946[illegible]G-14
[illegible]	[illegible]PLE [illegible]GB	€ [illegible]	[illegible]	NL-[illegible]
[illegible]	[illegible]PPLE [illegible]GB	€ 639.00	2	NL-[illegible]
[illegible]	[illegible]PPLE iPhone 6 16 GB	€ 639.00	1	NL-7751AR-19
1678	SAMSUNG Galaxy S4	€ 329.00	7	NL-9521CT-28b
1679	MOTOROLA Moto G	€ 199.00	2	NL-7948BX-10
1680	APPLE iPhone 6 16 GB	€ 639.00	7	NL-7826GD-9
1681	APPLE iPhone 6 16 GB	€ 639.00	1	NL-9408BM-37
1682	APPLE iPhone 6 16 GB	€ 639.00	4	NL-7823JJ-7
1683	MOTOROLA Moto E 4G	€ 99.99	2	NL-7821AC-3

Fig. 2 Fragment of a data set with one row per order (10,000 orders in total).

3 Spreadsheets: History and Concepts

Most organizations use spreadsheets in financial planning, budgeting, work distribution, etc. Hence, it is interesting to view process mining against the backdrop of this widely used technology.

3.1 History

Richard Mattessich pioneered computerized spreadsheets in the early 1960-ties. Mattessich realized that doing repeated "what-if" analyses by hand is not productive. He described the basic principles (computations on cells in a matrix) of today's spreadsheets in [Mattessich, 1964] and provided some initial Fortran IV code written by his assistants Tom Schneider and Paul Zitlau. The ideas were not widely adopted because few organizations owned computers. Rene Pardo and Remy Landau created in 1969 the *LANPAR* (LANguage for Programming Arrays at Random) electronic spreadsheet already allowing for forward references and natural order recalculation (handling cells that depend on one another). Again the market did not seem ready for spreadsheet software.

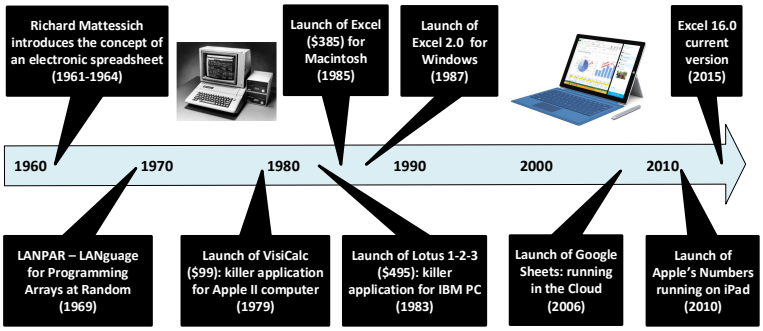


Fig. 3 Timeline showing a few milestones in the history of spreadsheet technology (1960-2015).

The first widely used spreadsheet program was *VisiCalc* ("Visible Calculator") developed by Dan Bricklin and Bob Frankston, founders of Software Arts (later named VisiCorp). VisiCalc was released in 1979 for the Apple II computer. It is generally considered as Apple II's "killer application", because numerous organizations purchased the Apple II computer just to be able to use *VisiCalc*. In the years that followed the software was ported to other platforms including the Apple III, IBM PC, Commodore PET, and Atari. In the same period *SuperCalc* (1980) and *Multiplan* (1982) were released following the success of *VisiCalc*.

Lotus Development Corporation was founded in 1982 by Mitch Kapor and Jonathan Sachs. They developed *Lotus 1-2-3*, named after the three ways the product could be used: as a spreadsheet, as a graphics package, and as a database manager. When *Lotus 1-2-3* was launched in 1983, *VisiCalc* sales dropped dramatically. *Lotus 1-2-3* took full advantage of IBM PC's capabilities and better supported data handling and charting. What *VisiCalc* was for Apple II, *Lotus 1-2-3* was for IBM PC. For the second time, a spreadsheet program generated a tremendous growth in computer sales [Rakovic et al., 2014].

Lotus 1-2-3 dominated the spreadsheet market until 1992. The dominance ended with the uptake of Microsoft Windows.

Microsoft's *Excel* was released in 1985. Microsoft originally sold the spreadsheet program *Multiplan*, but replaced it by *Excel* in an attempt to compete with *Lotus 1-2-3*. The software was first released for the Macintosh computer in 1985. Microsoft released *Excel 2.0* in 1987 which included a run-time version of MS Windows. Five years later, *Excel* was market leader and became immensely popular as an integral part of the Microsoft's *Office* suite. Borland's *Quattro* which was released in 1988 competed together with *Lotus 1-2-3* against *Excel*, but could not sustain a reasonable market share. *Excel* has dominated the spreadsheet market over the last 25 years. In 2015, the 16th release of *Excel* became available.

Online cloud-based spreadsheets such as *Google Sheets* (part of *Google Docs* since 2006) provide spreadsheet functionality in a web browser. *Numbers* is a spreadsheet application developed by Apple available on iPhones, iPads (iOS), and Macs (OS X). Dozens of other spreadsheet apps are available via Google Play or Apple's App Store.

Figure 3 summarizes 55 years of spreadsheet history. The key point is that spreadsheets have been one of the primary reasons to use computers in business environments.

3.2 Basic Concepts

In a spreadsheet (sometimes called worksheet), data and formulas are arranged over cells grouped in rows and columns. In *Excel* multiple worksheets can be combined into a workbook. Here, we only consider the spreadsheet depicted in Figure 4.

In a spreadsheet, each row is represented by a number and each column is represented by a letter. Cell **A1** is the cell where the first row (**1**) and column (**A**) meet. Cell **D9968** in Figure 4 has value 4 indicating that 4 iPhones were ordered. A cell may have a concrete value or may be computed using an expression operating on any number of cell values.

In Figure 4, row 1 is a header row containing column names. Rows 2 until 10,001 and columns **A** until **E** contain the data values already explained in Figure 2. Row **F** has 10,000 cells whose values are computed using the values in columns **D** and **C**. The expression associated to a cell may use a range of arithmetic operations (add, subtract, multiply, etc.) and predefined functions (e.g., taking the sum over an array of cells). *Excel* provides hundreds of functions including statistical functions, math and trigonometry functions, financial functions, and logical functions. The value of cell **I9969** was obtained by taking the sum over all values in row **F**: The total value of all orders summed up to €14,028,176.20.

Figure 4 also shows a so-called *pivot table* automatically summarizing the data. The pivot table shows the sales per type of phone, both in term of items and revenue. The pie chart shows that the "APPLE iPhone 6 16 GB" was sold

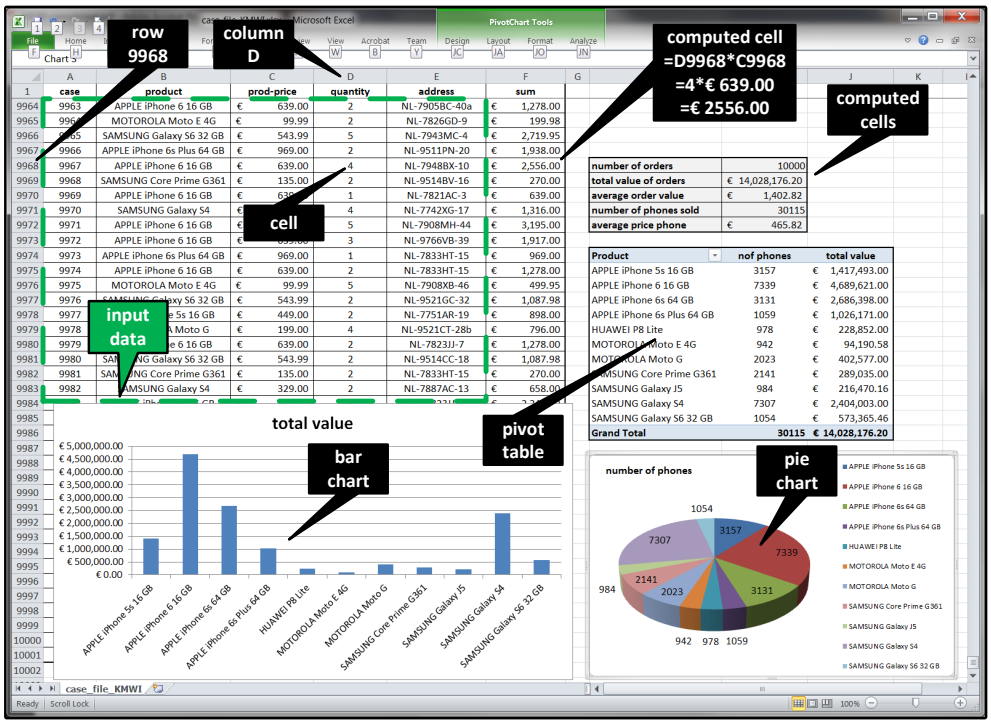


Fig. 4 Example spreadsheet analyzing the sales per product.

most (7339 phones). The bar chart shows the distribution in terms of revenue. The "APPLE iPhone 6s Plus 64 GB" ranks 5th although only 1059 phones were sold.

3.3 Analyzing Event Data?

Although spreadsheet software is very generic and offers many functions, programs like *Excel* are not suitable for analyzing event data. In Section 3.2 we analyzed the data of Figure 2 using simple operations such as multiplication, division, counting, and summation. When analyzing dynamic behavior such operations are not suitable. Consider for example the event data in Figure 1. We can count the number of events per case using a pivot table. However, spreadsheet software cannot be used to analyze bottlenecks and deviations. *The process notion is completely missing in spreadsheets.* Processes cannot be captured in numerical data and operations like summation.

4 Process Mining: Spreadsheets For Dynamic Behavior

As argued in the previous section, spreadsheet software can be used to do anything with numbers. However, spreadsheets cannot capture processes and cannot handle event data well. *Therefore, we propose process mining as a spreadsheet-like technology for processes starting from events.*

4.1 Event Logs

Starting point for any process mining effort is a collection of *events* commonly referred to as an *event log* (although events can also be stored in a database). Each event is characterized by:

- a *case* (also called *process instance*), e.g., an order number, a patient id, or a business trip,
- an *activity*, e.g., "evaluate request" or "inform customer",
- a *timestamp*, e.g., "2015-11-23T06:38:50+00:00",
- additional (optional) *attributes* such as the *resource* executing the corresponding event, the *type* of event (e.g., start, complete, schedule, abort), the *location* of the event, or the *costs* of an event.

All events corresponding to a case (i.e. process instance) form a *trace*. The order of events in a trace is determined by the timestamps. If we focus on activity names only, we can represent the trace corresponding to order 1672 by the sequence: *place order*, *pay*, *send invoice*, *prepare delivery*, *make delivery*, *confirm payment*. An event log is a collection of events that can be grouped into traces. Dedicated formats such as *XES* (`www.xes-standard.org`) and *MXML* exist to store events data in an unambiguous manner.

Event logs can be used for a wide variety of process mining techniques. Figure 1 shows an event log. The first three columns correspond to the mandatory attributes (case, activity, and timestamp). Cases correspond to orders in this example.

An event log provides a view on reality. Just like a workbook in *Excel* may hold multiple worksheets, we may consider multiple processes or multiple views on the same process. Sometimes multiple case notions are possible providing different views on the same event data. However, for simplicity, we consider only one, relatively simple, event log (like the one in Figure 1) as input for process mining here.

Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). The interest in process mining is rising. This is reflected by the availability of commercial tools like *Disco* (Fluxicon), *Celonis Process Mining* (Celonis), *ProcessGold Enterprise Platform* (ProcessGold), *ARIS PPM* (Software AG), *QPR ProcessAnalyzer* (QPR), *SNP Business Process Analysis* (SNP AG), *minit* (Gradient ECM), *myInvenio* (Cognitive Technology), *Perceptive Processing Mining* (Lexmark), etc. (see Section 4.8). In the academic world, *ProM* is

the de-facto standard (`www.processmining.org`) and research groups all over the world have contributed to the hundreds of *ProM* plug-ins available. *All analysis results depicted in this paper were obtained using ProM.*

4.2 Exploring Event Data

Starting from an event log like the one in Figure 1, we can explore the set of events. Simple descriptive statistics can be applied to the event log, e.g., the average flow time of cases or the percentage of cases completed within one week. Univariate statistical analysis focuses on a single variable like flow time, including its central tendency (including the mean, median, and mode) and dispersion (including the range and quantiles of the data-set, and measures of spread such as the variance and standard deviation). Bivariate statistical analysis focuses on the relationship between variables, e.g., correlation. However, to get a good feel for the behavior captured in the event log, one needs to look beyond basic descriptive statistics.

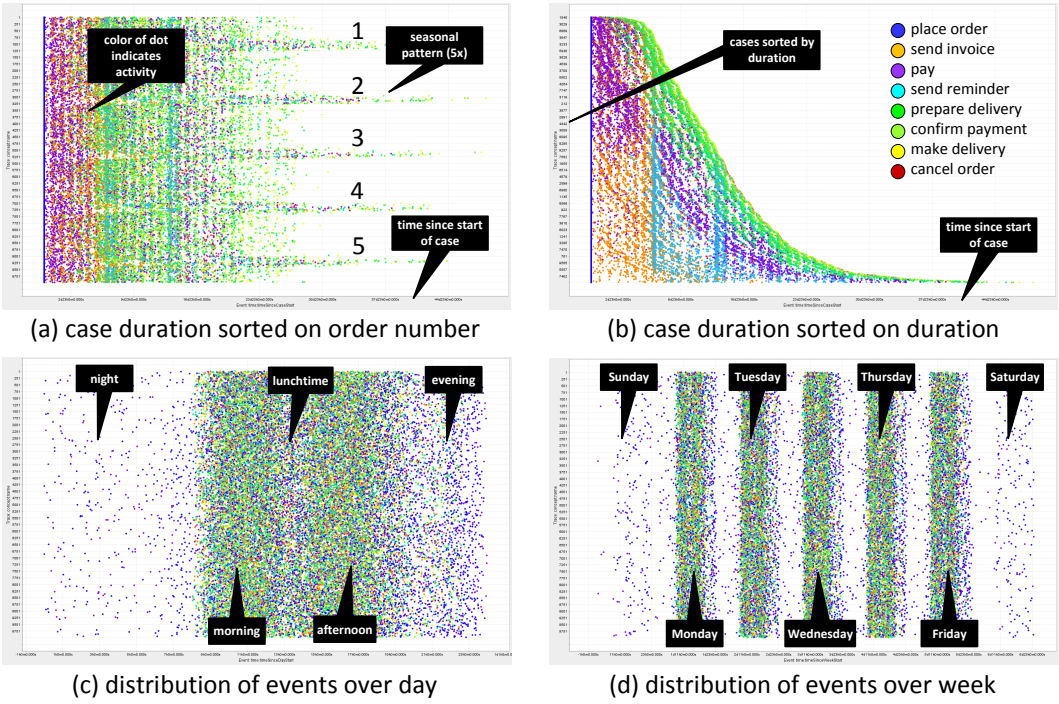


Fig. 5 Exploring the event log using *ProM*'s dotted charts.

Figure 5 shows four so-called "dotted charts" for the data set shown in Figure 1. Each of the four charts shows 63,763 dots arranged over 10,000 rows. The color of the dot refers to the corresponding activity. See the legend in

Figure 5(b) for the mapping, e.g., the dark blue dot refers to activity *place order*. In all four diagrams, the X-axis refers to a temporal property of the event and the Y-axis refers to the corresponding case (i.e., customer order). In Figure 5(a) the time since the start of the case is used for the X-axis. All orders start with a blue dot at time zero indicating that cases start with activity *place order*. The colored bands show that activities tend to happen in certain periods, e.g., the first reminder (if any) is typically sent after a week. One can also see clearly seasonal patterns; at certain periods flow times are considerably longer. Figure 5(a) shows five such periods. In Figure 5(b) the cases are sorted based on their flow time. The top cases take the least time to completion; the bottom cases take the longest. Again one can see clear patterns. For example, cases that take longer have multiple reminders. Figure 5(c) shows the distribution of events over the day. Most activities take place during office hours. One can also note the effect of lunch breaks. During the night we only see blue and purple dots indicating the placing of orders and payments. These activities are done by customers not bound to office hours. Figure 5(d) shows the distribution of events over the week. Again we can clearly notice that, apart from placing of orders and making payments, most activities take place during office hours and not during weekends.

Figure 5 provides insights that get lost if events are aggregated into numbers. Unlike spreadsheets, process mining treats concepts such as *case* (X-axis), *time* (Y-axis), and *activity* (color dot) as first-class citizens during analysis.

4.3 Process Discovery

Most of process mining research focused on the *discovery of process models from event data* [Aalst, 2016]. The process model should be able to capture causalities, choices, concurrency, and loops. Process discovery is a notoriously difficult problem because event logs are often far from complete and there are at least four competing quality dimensions: (1) *fitness*, (2) *simplicity*, (3) *precision*, and (4) *generalization*. A model with good *fitness* allows for most of the behavior seen in the event log. A model has a perfect fitness if all traces in the log can be replayed by the model from beginning to end. The *simplest* model that can explain the behavior seen in the log is the best model. This principle is known as Occam's Razor. Fitness and simplicity alone are not sufficient to judge the quality of a discovered process model. For example, it is very easy to construct an extremely simple process model that is able to replay all traces in an event log (but also any other event log referring to the same set of activities). Similarly, it is undesirable to have a model that only allows for the exact behavior seen in the event log. Remember that the log contains only example behavior and that many traces that are possible may not have been observed yet. A model is *precise* if it does not allow for "too much" behavior. A model that is not precise is "underfitting", i.e., the model allows for behaviors very different from what was seen in the log. At the same time, the model should generalize and not restrict behavior to just the examples

seen in the log. A model that does not *generalize* is "overfitting". Overfitting means that an overly specific model is generated whereas it is obvious that the log only holds example behavior (i.e., the model explains the particular sample log, but there is a high probability that the model is unable to explain the next batch of cases).

The discussion above shows that process discovery needs to deal with various trade-offs. Therefore, most process discovery algorithms have parameters to influence the result. Hence, different models can be created based on the questions at hand.

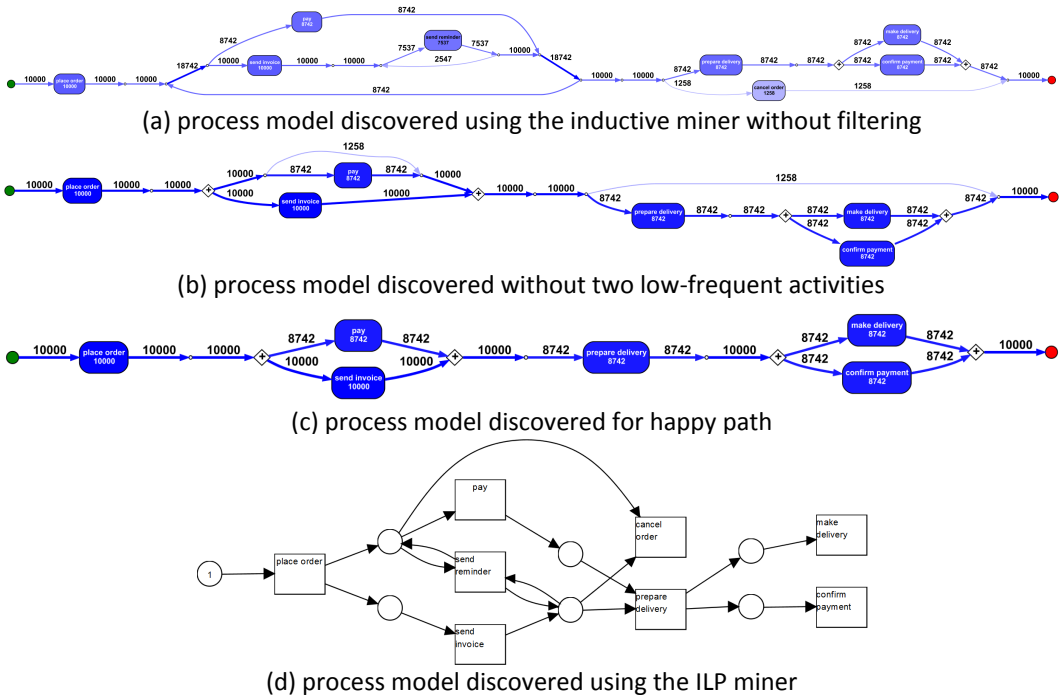


Fig. 6 Four process models automatically discovered using *ProM*'s *Inductive Miner* and *ILP Miner*.

Over the last decade there have been tremendous advances in automated process discovery. Figure 6 shows four process models discovered for the data set consisting of 63,763 events related to 10,000 orders. The first three models have been discovered using the *Inductive Miner* [Leemans et al., 2014, Leemans et al., 2015] and the last one was discovered using the *ILP Miner* [Werf et al., 2010, Zelst et al., 2015]. These models could have been automatically converted to BPMN models [Dumas et al., 2013] or other notations like UML activity diagrams, statecharts, EPCs, and the like. However, to see some of the important subtleties, we keep the native representation used by these process discovery techniques. (For example, a straightforward mapping of the

Petri net in Figure 6(d) to a BPMN model having precisely the same behavior is impossible.)

Figure 6(a) shows a perfectly fitting process model showing all eight activities. Each case starts with the placement of an order and ends with a cancellation, a delivery, or a confirmation of payment. The diamond shaped "+" nodes correspond to AND-splits/joins. All other splits/joins are of type XOR. Figure 6(b) shows a perfectly fitting process model after automatically removing the two least frequent activities. Note that the placement of an order is always followed by the sending of an invoice and sometimes by a payment. For 1,258 orders there was no payment as shown by the number on the arc bypassing activity *pay*. Figure 6(c) shows another automatically discovered process model, but now the *Inductive Miner* was asked to uncover the "happy path" (i.e., the most frequent behavior). In this idealized model all customer pay (either before or after receiving the invoice), there are no cancelations, the order is always delivered, and payment is always confirmed.

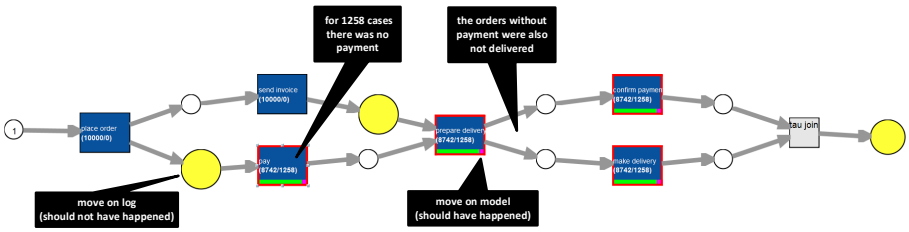
Figure 6(a) is perfectly fitting but not very precise. Using the *ILP Miner* we discovered the Petri net shown in Figure 6(d). Using Petri nets we can express things missing in the earlier diagrams. For example, Figure 6(d) shows that cancelation only takes place after sending the invoice and missing payment. If the customer pays before cancelation, the order is eventually delivered. Moreover, reminders are only sent after sending the invoice and before payment.

Each of the four models could be discovered in a few seconds on a normal laptop. Note that the discovered process model is not the end-goal of process mining: It is the backbone for further analysis!

4.4 Checking Compliance

The second type of process mining is *conformance checking* [Aalst, 2016]. 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 process model used as input may be hand-made or discovered. To check compliance often a normative hand-crafted model is used. However, to find exceptional cases, one can also use a discovered process model showing the mainstream behavior. It is also possible to "repair" process models based on event data.

To illustrate the kind of results conformance checking may deliver, consider Figure 7. The original event log with 63,763 events is replayed on a process model that describes the "happy flow", i.e., the path followed by orders that are paid in time and not canceled. The model is represented as a Petri net in Figure 7(a), but is from a behavioral point of view identical to Figure 6(c) (i.e., the model discovered based on the most frequent behavior). The replay results show that there are 1,258 cases for which a payment and delivery were both missing.



(a) model view showing where model does not fit reality

(b) log view showing where reality does not fit model

Fig. 7 Conformance checking as a means to show deviations between observed and modeled behavior.

The diagnostics in Figure 7 are based on so-called *alignments*, i.e., traces in the event log are mapped onto nearest paths in the model [Aalst et al., 2012]. Basically, there are two types of deviations:

- *Move on model*: An activity was supposed to happen according to the model but did not happen in reality, i.e., the corresponding *event was missing* in the event log. Such deviations are indicated in purple.
- *Move on log*: An activity happened in reality but was not supposed to happen at this stage according to the model, i.e., there is an event in the log that *was not allowed* at that point in time. Such deviations are indicated in yellow.

Figure 7(a) shows a model-based view with conformance diagnostics. The small purple lines at the bottom of the four highlighted activities show the moves on model. For example, activity prepare delivery was skipped 1,258 times. The yellow places correspond to states where activities happened in reality, but were not allowed according to the model. Figure 7(b) shows a log-based conformance view. Again the colors indicate deviations.

Using conformance checking one can analyze the severity of the different types of deviations. It is also possible to select cases having a specific type of deviation and automatically see what differentiates them from conforming cases. In this way, we can learn about the root causes of non-conforming behavior.

4.5 Analyzing Performance

Using the notion of alignments, we can replay any event log on the corresponding model even when there are deviations. Recall that each event in the log has a timestamp (third column in Figure 1). While replaying the event log we can take into account these timestamps and measure the time spent in-between activities. This way we can analyze waiting times. If logs have both start and complete events for activities, we can also measure the duration of such activities. If event logs also have resource information, we can detect over/under-utilization of resources. Hence, while replaying we can get all information needed for performance analysis.

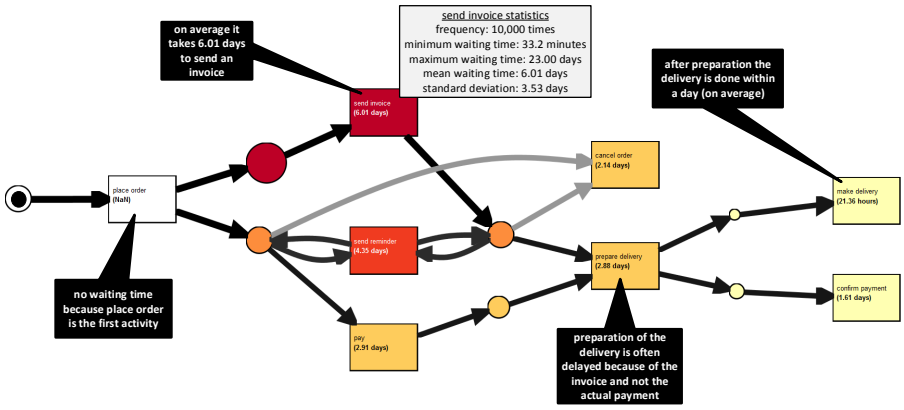


Fig. 8 Average waiting times for activities computed by replaying the whole event log.

All 63,763 events in Figure 1 are complete events. Therefore, we can only analyze the times in-between activities. Figure 8 shows the mean waiting times for activities using the model discovered by the *ILP Miner* (cf. Figure 6(d)). Next to the mean we can show the minimum, maximum, median, standard deviation, variance, etc. The main bottleneck in the process seems to be the sending of invoices. It is also possible to select cases taking longer than some normative time and see what differentiates them from the other cases. This allows us to diagnose bottlenecks and generate ideas for process improvement.

4.6 Process Animation

Replaying the event log using alignments can be used to generate *animations* of the process. These are computed based on both the model and event data. Instead of showing a diagram like in Figure 8, we can show a "process movie". Figure 9 shows snapshots of an animation created using a model discovered by the *Inductive Miner*. Figure 9(a) shows the status of the overall process (without activity *send reminder*) at a particular point in time. The moving

yellow dots refer to orders recorded in the event log. Figure 9(b) zooms-in on the last part of the model. Figure 9(c) shows the queues for the *pay* and *send invoice* activities.

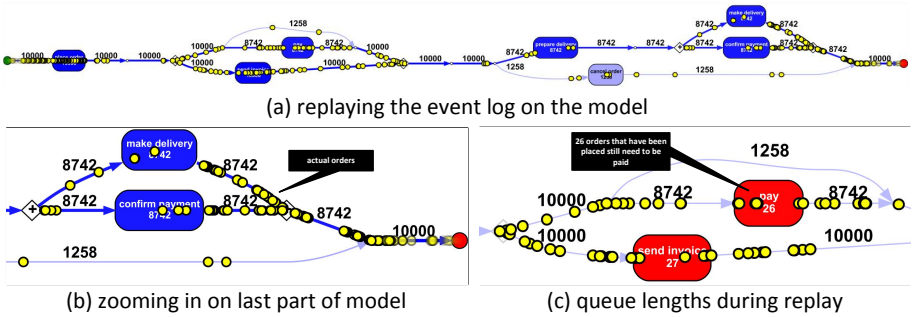


Fig. 9 Animations created using the event in Figure 1 and a discovered process model.

Process animations (like the one shown in Figure 9) help to build consensus in process improvement projects. In most reengineering projects, some stakeholders tend to question numerical arguments or data quality to avoid painful conclusions. However, objectively visualizing the developments in a process (process animation) with the ability to drill down to individual cases, leaves no room for biased interpretations. This helps to shortcut discussions and take the actions needed.

4.7 Operational Support

Thus far we only discussed process mining in a *offline* setting. This helps to understand and improve compliance and performance issues. However, process mining can also be applied in an *online* setting [Aalst, 2016]. We would like to predict delays, warn for risks, and recommend counter measures. Compare this to the weather forecast. We are less interested in historic weather data if these cannot be used to predict today's or tomorrow's weather. Sometimes delays or risks are partly unavoidable; however, it is valuable to predict them at a point in time where stakeholders can still influence the process.

Most process mining techniques can be employed for operational support, i.e., influencing running processes on-the-fly rather than redesigning them [Aalst, 2016]. For example, cases that have not completed yet can be replayed and combined with historic information. Consider for example Figure 9(c), showing queue lengths at a particular point in time. Such information can also be provided at runtime. Compare this to the use of Doppler radar to locate precipitation, calculate its motion, and estimate its type (e.g., rain, snow, or hail).

Stochastic process models with probabilities and delay distributions discovered from event data can be used to predict the trajectory of a running

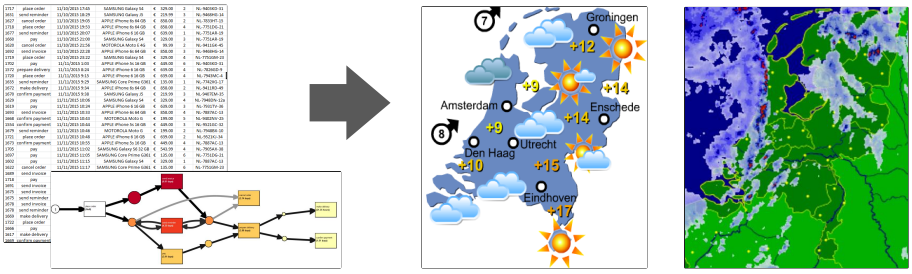


Fig. 10 Operational support using process mining: event data and process models (a) are used to create a "process forecast" (b) and a "process radar" (c).

case or a group of cases (like the weather radar). Moreover, process models can be continuously revised based on the latest event data. Figure 10 aims to convey the relationship between process analytics and weather information. Operational support is challenging—just like predicting the weather—and only provides reliable results if the process's behavior is indeed predictable.

4.8 Tool Support

The successful application of process mining relies on good tool support. ProM is the leading open-source process mining tool. The lion's share of academic research is conducted by using and extending ProM (and related variants such as RapidProM). Many of the commercial process mining tools are based on ideas first developed in the context of ProM. Table 1 shows an overview of some of the current tools. The functionality of these tools is summarized in Table 2. Note that the process mining field is developing rapidly, so the information is likely to be outdated soon. However, the two tables provide a snapshot of the current tools and their capabilities.

Most tools support XES (`www.xes-standard.org`), the official IEEE standard for exchanging event data. All tools support process discovery and performance analysis, i.e., all can automatically create a process model highlighting the bottlenecks in the process. There is limited support for conformance checking. Scalability issues (e.g., computing alignments may be too time consuming) and informal semantics (e.g., not being able to distinguish between AND-joins and XOR-joins) are some of the hurdles commercial vendors are facing. Note that in Table 2 the comparison of two process model graphs is not considered as a way to support compliance checking (e.g., myInvenio supports this). In our view, compliance checking requires replaying observed behavior on a model that has clear semantics. Most vendors support animation, but operational support (e.g., recommending the next activity to be executed or predicting future bottlenecks) is rarely supported.

Short name	Full name of tool	Version	Vendor	Webpage
Celonis	Celonis Process Mining	4	Celonis GmbH	www.celonis.de
Disco	Disco	1.9.5	Fluxicon	www.fluxicon.com
Fujitsu	Interstage Business Process Manager Analytics	12.2	Fujitsu Ltd	www.fujitsu.com
Icaro	Icaro EVERFlow	1	Icaro Tech	www.icarotech.com
Icris	Icris Process Mining Factory	1	Icris	www.processminingfactory.com
LANA	LANA Process Mining	1	Lana Labs	www.lana-labs.com
Minit	Minit	1	Gradient ECM	www.minitlabs.com
myInvenio	myInvenio	1	Cognitive Technology	www.my-invenio.com
Perceptive	Perceptive Process Mining	2.7	Lexmark	www.lexmark.com
ProcessGold	ProcessGold Enterprise Platform	8	Processgold International B.V.	www.processgold.com
ProM	ProM	6.6	Open Source hosted at TU/e	www.promtools.org
ProM Lite	ProM Lite	1.1	Open Source hosted at TU/e	www.promtools.org
QPR	QPR ProcessAnalyzer	2015.5	QPR	www.qpr.com
RapidProM	RapidProM	4.0.0	Open Source hosted at TU/e	www.rapidprom.org
Rialto	Rialto Process	1.5	Exeura	www.exeura.eu
Signavio	Signavio Process Intelligence	2016	Signavio GmbH	www.signavio.com
SNP	SNP Business Process Analysis	15.27	SNP Schneider-Neureither Partner AG	www.snp-bpa.com
PPM	webMethods Process Performance Manager	9.9	Software AG	www.softwareag.com
Worksoft	Worksoft Analyze and Process Mining for SAP	2016	Worksoft, Inc.	www.worksoft.com

Table 1 Overview of available process mining tools (not intended to be incomplete)

Name of tool	XES support	Process discovery	Compliance checking	Performance analysis	Process animation	Operational support
Celonis	Yes	Yes	Yes	Yes	Yes	No
Disco	Yes	Yes	No	Yes	Yes	No
Fujitsu	No	Yes	No	Yes	No	No
Icaro	No	Yes	No	Yes	No	No
Icris	Yes	Yes	No	Yes	No	No
LANA	Yes	Yes	Yes	Yes	No	No
Minit	Yes	Yes	No	Yes	Yes	No
myInvenio	Yes	Yes	No	Yes	Yes	No
Perceptive	No	Yes	No	Yes	Yes	No
ProcessGold	Yes	Yes	No	Yes	Yes	No
ProM	Yes	Yes	No	Yes	Yes	No
ProM Lite	Yes	Yes	No	Yes	Yes	No
QPR	Yes	Yes	No	Yes	Yes	No
RapidProM	Yes	Yes	No	Yes	Yes	No
Rialto	Yes	Yes	No	Yes	No	No
Signavio	No	No	Yes	Yes	No	No
SNP	Yes	Yes	No	Yes	No	No
PPM	No	Yes	No	Yes	No	No
Worksoft	No	Yes	No	Yes	No	No

Table 2 Process mining tasks supported by tool. Disclaimer: This table is based on the information currently available. The functionality of tools is changing rapidly, so please consult the vendor for the most recent information.

As mentioned, tables 1 and 2 merely provide a snapshot. However, they illustrate the emergence of a new class of tools able to analyze event data in a truly generic manner.

5 Conclusion

Just like spreadsheet software, process mining aims to provide a generic approach not restricted to a particular application domain. Whereas spreadsheets focus on *numbers*, process mining focuses on *events*. There have been some attempts to extend spreadsheets with process mining capabilities. For example, QPR's *ProcessAnalyzer* can be deployed as an *Excel* add-in. However, processes and events are very different from bar/pie charts and numbers. Process models and concepts related to cases, events, activities, timestamps, and resources need to be treated as first-class citizens during analysis. Data mining tools and spreadsheet programs take as input any tabular data without distinguishing between these key concepts. As a result, such tools tend to be process-agnostic.

5.1 Comparison of Concepts

Table 3 summarizes some of the main concepts in spreadsheets and process mining. The event notion does not exist in spreadsheets. Spreadsheets can produce a variety of charts, but cannot discover a process model from event data. The input for process mining is an *event log* that consists of *events* grouped in *cases*. Each case (also called process instance) is described by a sequence of events. Events may have any number of *attributes*. Each event refers to an *activity* and has a *timestamp*. An event may also refer to a *resource* (person, machine, software component, etc.) and carry transactional information (start, complete, suspend, etc.). Based on event data a *process model* can be discovered showing bottlenecks, mainstream behavior, exceptional execution paths, etc. A process model can also be given as input to conduct conformance checking or to enrich or repair process models. Any type of process model can be used as long as it can be related to sequences of events. Table 3 shows that discovered models, social networks, compliance diagnostics, predictions, and recommendations are possible outputs of process mining activities. The table also shows that the concepts are as generic as the concepts one can find in a spreadsheet.

5.2 Challenges

Still we can learn from spreadsheets and improve the accessibility of process mining. The direct manipulation of data combined with a large repertoire of functions is very powerful. Moreover, spreadsheets implicity encode analysis

	Spreadsheet	Process mining
Input	Worksheet	Event log
	Cell	Event
	Row	Case (process instance)
	Column	Activity
		Timestamp
		Resource
		Type (start, complete, abort, etc.)
		Normative process model
Output	Bar charts, pie charts, area charts, radar charts, etc.	Discovered process models (control-flow and possibly other perspectives)
	Pivot tables	Social networks
	Sums, averages, standard deviations, etc.	Deviations (e.g., alignments)
		Bottlenecks
		Process-aware predictions and recommendations

Table 3 Summary of the main concepts in spreadsheets and process mining. Concepts such as case, event, activity, timestamp, and resource do not exist in spreadsheets.

workflows. Intermediate results stored in cells can be used as input for subsequent analysis steps. In this context we would like to refer to *RapidProM* [Mans et al., 2014] which supports process mining workflows in a visual manner.

The spectacular growth of event data provides many opportunities for automated process discovery based on *facts*. Event logs can be replayed on process models to check conformance and analyze bottlenecks. However, still missing are reliable techniques to *automatically* improve operational processes. Existing process mining techniques can be used to diagnose problems, but the transition from "as-is" to "to-be" models is not yet supported adequately.

Since the first industrial revolution, productivity has been increasing because of technical innovations, improvements in the organization of work, and the use of information technology. Frederick Taylor (1856-1915) introduced the initial principles of scientific management. In his book "The Principles of Scientific Management" he proposed to standardize best practices and suggested techniques for the elimination of waste and inefficiencies [Taylor, 1919]. These ideas have matured and approaches have been developed over the last century. Business Process Management (BPM) follows the same tradition. However, the abundance of (event) data is changing the BPM landscape rapidly. Today, we are witnessing the fourth industrial revolution ("Industrie 4.0"). Operations management, and in particular operations research, is a branch of management science heavily relying on modeling. Here a variety of mathematical models ranging from linear programming and project planning to queueing models, Markov chains, and simulation are used. These models often focus on a particular decision (at run-time or at design-time) rather than the process *as a whole*. The "holy grail" of scientific management has been to automatically

improve operational processes, i.e., to observe a process as it is unfolding and us this to provide clear and reliable suggestions for improvement. Although the practical value of evidence-based automated process optimization is evident, it has only been realized for rather specific operational decisions. However, the omnipresence of event data and the availability of reliable and fast process mining techniques make it possible to discover *faithful* control-flow models and to align reality with these discovered models. This creates new opportunities for scientific management.

The focus of future process mining research should be on *automatically improving processes* by changing the underlying process models or by better controlling existing ones. *How to do this?*

- Starting point should be the discovered as-is models. These models and the event data can be used for comparative process mining. Given multiple variants of the same process, the same process in different periods, or different types of cases within the same process, we can discover characteristic commonalities and differences while exploiting the underlying event data. This provides novel diagnostic information aiming at better understanding the factors influencing performance.
- The as-is model can also be used for predictive analytics, e.g., predicting the remaining flow time for a running case or recommending a suitable resource at run-time.
- It is also possible to combine the as-is model with so-called change constraints. Here also domain knowledge is used to determine the "degrees of freedom" in redesign. To automatically suggest improved process designs, as-is models, event data, change constraints, and goals are used as input. The resulting (hopefully) improved to-be process models can be evaluated using a combination of real event data and simulated event data.

The overall approach envisioned supports a data-driven approach to automatically improve process performance. This goes far beyond existing approaches that only support "what-if" analysis and require experts to model the process.

In conclusion, we promoted process mining as a generic technology on the interface between data science and Business Process Management (BPM). We hope that process mining will become the "tail wagging the dog" (with the dog being Big Data initiatives) and play a role comparable to spreadsheets. This may seem unrealistic, but there is a clear need to bridge the gap between data science and process management. Process mining provides the glue connecting both worlds, but there is room for improvement. As indicated, the challenge is to move from diagnostics to semi-automated process improvement. Process mining comes in three principal flavors: descriptive, predictive and prescriptive. The focus has been on descriptive analytics. Now it is time to focus on predictive and prescriptive analytics. Process mining tools like ProM already support techniques like prediction. However, process mining for prescriptive analytics is still a rather unexplored territory in BPM.

References

[Aalst, 2013] Aalst, W.M.P. van der 2013. Business Process Management: A Comprehensive Survey. ISRN Software Engineering, pages 1–37. doi:10.1155/2013/507984.

[Aalst, 2014] Aalst, W.M.P. van der 2014. Data Scientist: The Engineer of the Future. In Mertins, K., F. Benaben, R. Poler, & J. Bourrieres (eds), Proceedings of the I-ESA Conference, volume 7 of *Enterprise Interoperability*, pages 13–28. Springer-Verlag, Berlin.

[Aalst, 2016] Aalst, W.M.P. van der 2016. Process Mining: Data Science in Action. Springer-Verlag, Berlin.

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

[Aalst et al., 2000] Aalst, W.M.P. van der, J. Desel, & A. Oberweis (eds) 2000. Business Process Management: Models, Techniques, and Empirical Studies, volume 1806 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin.

[Aalst & Hee, 2004] Aalst, W.M.P. van der, & K.M. van Hee 2004. Workflow Management: Models, Methods, and Systems. MIT press, Cambridge, MA.

[Becker et al., 2011] Becker, J., M. Kugeler, & M. Rosemann (eds) 2011. Process Management: A Guide for the Design of Business Processes, International Handbooks on Information Systems. Springer-Verlag, Berlin.

[Brocke & Rosemann, 2010] Brocke, J. vom, & M. Rosemann (eds) 2010. Handbook on Business Process Management, International Handbooks on Information Systems. Springer-Verlag, Berlin.

[Brocke & Rosemann, 2014] Brocke, J. vom, & M. Rosemann (eds) 2014. Handbook on Business Process Management 1: Introduction, Methods, and Information Systems, International Handbooks on Information Systems. Springer-Verlag, Berlin.

[Ceruzzi, 2003] Ceruzzi, P.E. 2003. A History of Modern Computing. MIT Press.

[Dumas et al., 2005] Dumas, M., W.M.P. van der Aalst, & A.H.M. ter Hofstede 2005. Process-Aware Information Systems: Bridging People and Software through Process Technology. Wiley & Sons.

[Dumas et al., 2013] Dumas, M., M. La Rosa, J. Mendling, & H. Reijers 2013. Fundamentals of Business Process Management. Springer-Verlag, Berlin.

[Hofstede et al., 2010] Hofstede, A.H.M. ter, W.M.P. van der Aalst, M. Adams, & N. Russell 2010. Modern Business Process Automation: YAWL and its Support Environment. Springer-Verlag, Berlin.

[Jablonski & Bussler, 1996] Jablonski, S., & C. Bussler 1996. Workflow Management: Modeling Concepts, Architecture, and Implementation. International Thomson Computer Press, London, UK.

[Jelen, 2005] Jelen, B. 2005. The Spreadsheet at 25: 25 Amazing Excel Examples that Evolved from the Invention that Changed the World. Holy Macro! Books.

[Leemans et al., 2014] Leemans, S.J.J., D. Fahland, & W.M.P. van der Aalst 2014. Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour. In Lohmann, N., M. Song, & P. Wohed (eds), Business Process Management Workshops, International Workshop on Business Process Intelligence (BPI 2013), volume 171 of *Lecture Notes in Business Information Processing*, pages 66–78. Springer-Verlag, Berlin.

[Leemans et al., 2015] Leemans, S.J.J., D. Fahland, & W.M.P. van der Aalst 2015. Scalable Process Discovery with Guarantees. In Gaaloul, K., R. Schmidt, S. Nurcan, S. Guerreiro, & Q. Ma (eds), Enterprise, Business-Process and Information Systems Modeling (BPMDS 2015), volume 214 of *Lecture Notes in Business Information Processing*, pages 85–101. Springer-Verlag, Berlin.

[Leymann & Roller, 1999] Leymann, F., & D. Roller 1999. Production Workflow: Concepts and Techniques. Prentice-Hall PTR, Upper Saddle River, New Jersey, USA.

[Mans et al., 2014] Mans, R., W.M.P. van der Aalst, & E. Verbeek 2014. Supporting Process Mining Workflows with RapidProM. In Limonad, L., & B. Weber (eds), Business Process Management Demo Sessions (BPMD 2014), volume 1295 of *CEUR Workshop Proceedings*, pages 56–60. CEUR-WS.org.

[Mattessich, 1964] Mattessich, R. 1964. Simulation of the Firm Through a Budget Computer Program. Homewood, R.D. Irwin.

[Muehlen, 2004] Muehlen, M. zur 2004. Workflow-based Process Controlling: Foundation, Design and Application of workflow-driven Process Information Systems. Logos, Berlin.

[Rakovic et al., 2014] Rakovic, L., M. Sakal, & V. Pavlicevic 2014. Spreadsheets - How It Started. International Scientific Journal of Management Information Systems, 9(4):9–14.

[Reichert & Weber, 2012] Reichert, M., & B. Weber 2012. Enabling Flexibility in Process-Aware Information Systems: Challenges, Methods, Technologies. Springer-Verlag, Berlin.

[Smith & Fingar, 2006] Smith, H., & P. Fingar 2006. Business Process Management: The Third Wave. Meghan Kiffer Press.

[Taylor, 1919] Taylor, F.W. 1919. The Principles of Scientific Management. Harper and Bothers Publishers, New York.

[Werf et al., 2010] Werf, J.M.E.M. van der, B.F. van Dongen, C.A.J. Hurkens, & A. Serebrenik 2010. Process Discovery using Integer Linear Programming. Fundamenta Informaticae, 94:387–412.

[Weske, 2007] Weske, M. 2007. Business Process Management: Concepts, Languages, Architectures. Springer-Verlag, Berlin.

[Zelst et al., 2015] Zelst, S.J. van, B.F. van Dongen, & W.M.P. van der Aalst 2015. ILP-Based Process Discovery Using Hybrid Regions. In Proceedings of the International Workshop on Algorithms and Theories for the Analysis of Event Data (ATAED 2015), volume 1371 of *CEUR Workshop Proceedings*, pages 47–61. CEUR-WS.org.