Process Mining: Spreadsheet-Like Technology for Processes

Wil van der Aalst

Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands,
w.m.p.v.d.aalst@tue.nl

Abstract

Spreadsheets can be viewed as a success story. Since the late seventies spreadsheet programs have been installed on the majority of computers and play a role comparable to text editors and databases management systems. Spreadsheets can be used to do anything with *numbers*, but are unable to handle *process models* and *event data*. Event logs and operational processes can be found everywhere. Recent breakthroughs in process mining resulted in novel techniques to discover the real processes, to detect deviations from normative process models, and to analyze bottlenecks and waste. Comparable to spreadsheet programs like *Excel* which are widely used in finance, production, sales, education, sports, process mining software can be used in a broad range of organizations. Whereas spreadsheets work with numbers, process mining starts from event data with the aim to analyze processes. This *keynote paper* uses spreadsheets as an analogy to make the case for process mining as an essential tool for data scientists and business analysts.

1 Spreadsheets: Handling Numbers

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).

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.

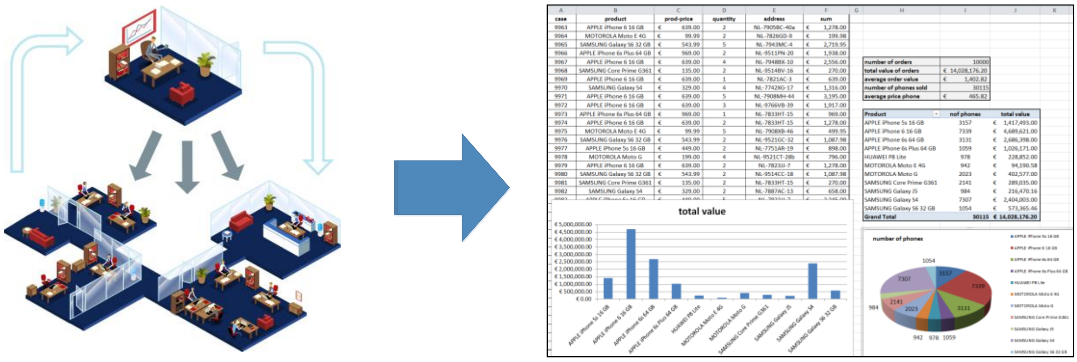
The first widely used spreadsheet program was *VisiCalc* ("Visible Calculator") developed by Dan Bricklin and Bob Frankston, founders of Software Arts (later named VisiCorp). VisiCalc was released in 1979 for the Apple II computer. It is generally considered as Apple II's "killer application", because 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.



case	product	prod-price	quantity	address	sum
9963	APPLE iPhone 6 16 GB	€ 639.00	2	NL-7905BC-40a	€ 1,278.00
9964	MOTOROLA Moto E 4G	€ 99.99	2	NL-7826GD-9	€ 199.98
9965	SAMSUNG Galaxy S6 32 GB	€ 543.99	5	NL-7943MC-4	€ 2,719.95
9966	APPLE iPhone 6s Plus 64 GB	€ 969.00	2	NL-9511PN-20	€ 1,938.00
9967	APPLE iPhone 6 16 GB	€ 639.00	4	NL-7948BX-10	€ 2,556.00
9968	SAMSUNG Core Prime G361	€ 135.00	2	NL-9514BV-16	€ 270.00
9969	APPLE iPhone 6 16 GB	€ 639.00	1	NL-7821AC-3	€ 639.00
9970	SAMSUNG Galaxy S4	€ 329.00	4	NL-7742XG-17	€ 1,316.00
9971	APPLE iPhone 6 16 GB	€ 639.00	5	NL-7908MH-44	€ 3,195.00
9972	APPLE iPhone 6 16 GB	€ 639.00	3	NL-9766VB-39	€ 1,917.00
9973	APPLE iPhone 6s Plus 64 GB	€ 969.00	1	NL-7833HT-15	€ 969.00
9974	APPLE iPhone 6 16 GB	€ 639.00	2	NL-7833HT-15	€ 1,278.00
9975	MOTOROLA Moto E 4G	€ 99.99	5	NL-7908XB-46	€ 499.95
9976	SAMSUNG Galaxy S6 32 GB	€ 543.99	2	NL-9521GC-32	€ 1,087.98
9977	APPLE iPhone 5s 16 GB	€ 449.00	2	NL-7751AR-19	€ 898.00
9978	MOTOROLA Moto G	€ 199.00	4	NL-9521CT-28b	€ 796.00
9979	APPLE iPhone 6 16 GB	€ 639.00	2	NL-7823JJ-7	€ 1,278.00
9980	SAMSUNG Galaxy S6 32 GB	€ 543.99	2	NL-9514CC-18	€ 1,087.98
9981	SAMSUNG Core Prime G361	€ 135.00	2	NL-7833HT-15	€ 270.00
9982	SAMSUNG Galaxy S4	€ 329.00	2	NL-7887AC-13	€ 658.00

number of orders	10000
total value of orders	€ 14,028,176.20
average order value	€ 1,402.82
number of phones sold	30115
average price phone	€ 465.82

Product	nof phones	total value
APPLE iPhone 5s 16 GB	3157	€ 1,417,493.00
APPLE iPhone 6 16 GB	7339	€ 4,689,621.00
APPLE iPhone 6s 64 GB	3131	€ 2,686,398.00
APPLE iPhone 6s Plus 64 GB	1059	€ 1,026,171.00
HUAWEI P8 Lite	978	€ 228,852.00
MOTOROLA Moto E 4G	942	€ 94,190.58
MOTOROLA Moto G	2023	€ 402,577.00
SAMSUNG Core Prime G361	2141	€ 289,035.00
SAMSUNG Galaxy J5	984	€ 216,470.16
SAMSUNG Galaxy S4	7307	€ 2,404,003.00
SAMSUNG Galaxy S6 32 GB	1054	€ 573,365.46
Grand Total	30115	€ 14,028,176.20

Figure 1: Reality is reduced to numbers in a spreadsheet: Concepts such as cases, events, activities, resources, etc. are missing and process models are not supported.

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 keynote 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.

2 Process Mining: Handling Events

Instead of numbers process mining starts from *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 (Van der 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 (Van der Aalst, 2011).

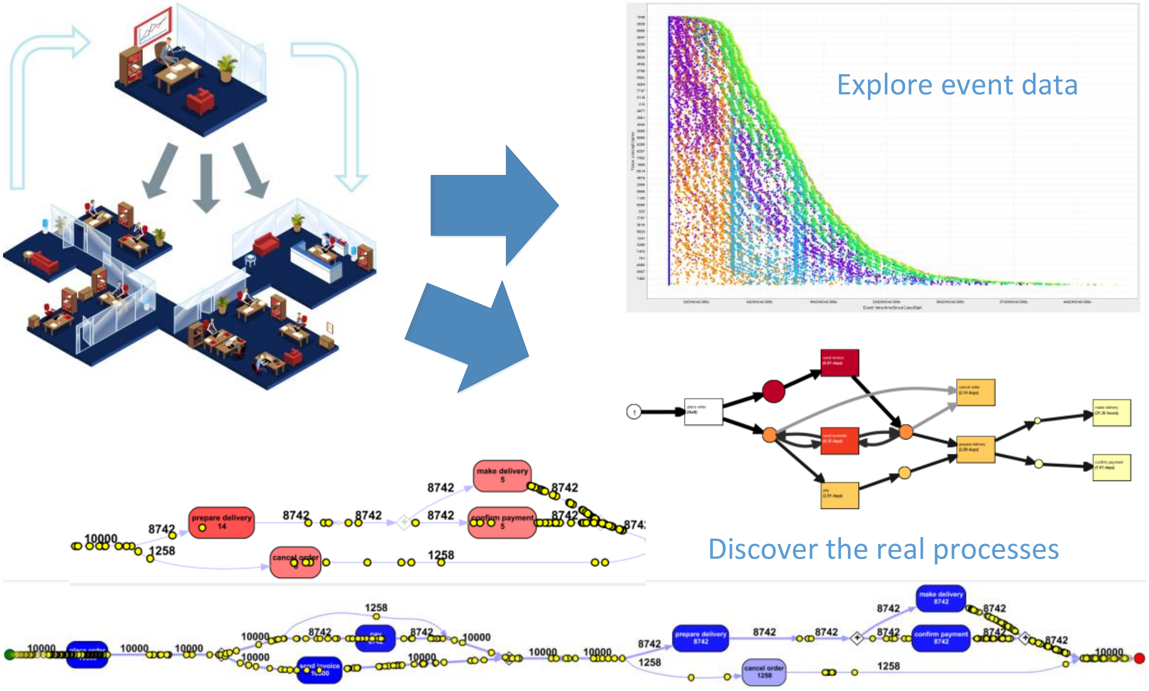


Figure 2: Process mining can be used to discover the real processes, detect deviations, predict delays and risks, and diagnose bottlenecks and waste. Concepts such as cases, events, activities, resources, etc. are natively supported and process models showing bottlenecks, risks, costs, etc. can be shown.

Process mining should be in the toolbox of data scientists, business analysts, and others that 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 process management approaches tend to be model-driven without considering the "evidence" hidden in the data. Process mining can be seen as a means to bridge the gap between data science and process management. *By positioning process mining as a spreadsheet-like technology for event data, we hope to increase awareness in the WirtschaftsInformatik (WI) / Business & Information Systems Engineering (BISE) community.*

3 Outlook

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 such as 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*.

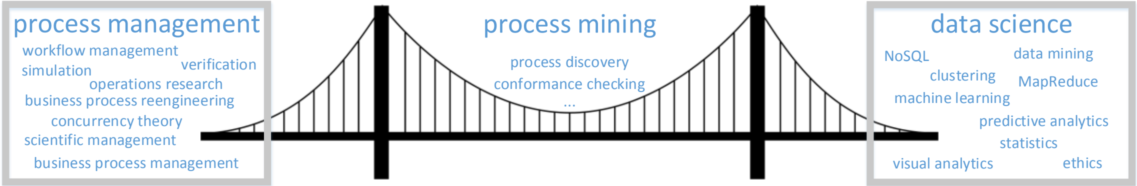


Figure 3: Process mining as the missing link between process management (BPM, WFM, BPR, etc.) and data science (data mining, statistics, etc.).

In this paper, we promoted process mining as a generic technology on the interface between data science and process management. 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 techniques connecting both worlds.

References

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

Van der Aalst, W. (2014). Data Scientist: The Engineer of the Future. In Mertins, K., Benaben, F., Poler, R., and Bourrieres, J., editors, Proceedings of the I-ESA Conference, volume 7 of Enterprise Interoperability, pages 13-28. Springer-Verlag, Berlin.

Van der Aalst, W. (2015). Spreadsheets for Business Process Management: How to deal with "events" rather than "numbers"? (Report, 18 pages, available upon request).

Ceruzzi, P. (2003). A History of Modern Computing. MIT Press.

Jelen, B. (2005). The Spreadsheet at 25: 25 Amazing Excel Examples that Evolved from the Invention that Changed the World.

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

Rakovic, L., Sakal, M., and Pavlicevic, V. (2014). Spreadsheets - How It Started. International Scientific Journal of Management Information Systems, 9(4):9-14.