Data-driven Management of Interconnected Business Processes

Contributions to Predictive and Prescriptive Process Mining

Dissertation

zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

der Rechts- und Wirtschaftswissenschaftlichen Fakultät

der Universität Bayreuth

Vorgelegt

von

Wolfgang Kratsch

aus

Augsburg

Dekan:	Prof. Dr. Jörg Schlüchtermann
Erstberichterstatter:	Prof. Dr. Maximilian Röglinger
Zweitberichterstatter:	Prof. Dr. Torsten Eymann
Tag der mündlichen Prüfung:	15.12.2020

PhD

Finally Done

Just the Beginning

Much More to Discover

Curiosity

Ich möchte mich bei meiner Familie, meinen Freunden und meiner Frau Alexandra bedanken. Ohne Eure Unterstützung in allen Lebenslagen wäre ich bei diesem Marathon sicherlich nicht ins Ziel gekommen.

Mein besonderer Dank gilt auch meinem Doktorvater Prof. Maximilian Röglinger. Du hast es geschafft, die wissenschaftliche Neugier in mir nachhaltig zu wecken.

Abstract

Business process management (BPM) is an accepted paradigm of organizational design to orchestrate distributed work involving various activities, resources, and actors, connecting the physical and digital world. While traditional research in BPM focused on process models and model-based information systems (e.g., workflow management systems), the focus has recently shifted toward data-driven methods such as process mining. Process mining strives to discover, monitor, and improve business processes by extracting knowledge from process (or event) logs. As process mining has evolved into one of the most active streams in BPM, numerous approaches have been proposed in the last decade, and various commercial vendors transferred these methods into practice, substantially facilitating event data analysis. However, there are still manifold unsolved challenges that hinder the adoption and usage of process mining at the enterprise level. First, finding, extracting, and preprocessing relevant event data remains challenging. Second, most process mining approaches operate on a single-process level, making it hard to apply process mining to multiple interconnected processes. Third, process managers strongly require forward-directed operational support, but most process mining approaches provide only descriptive ex-post insights. Driven by these challenges, this thesis contributes to the existing body of knowledge related to data-driven management of interconnected business processes. By proposing methods that enhance and automate the extraction of event logs from typical sources (research paper #1) and exploiting novel sources containing process-relevant information (research papers #2 and #3), this thesis contributes to the first challenge of finding, extracting, and preprocessing relevant event data. Regarding the second challenge to apply process mining to a multi-process perspective, this thesis proposes approaches for log-driven prioritization of interconnected business processes (research papers #4 and #5). As the proposed process prioritization methods build on predicting processes' future performance, they also contribute to the third challenge of providing forward-directed operational support for process managers. Providing accurate predictions leveraging the increasing volume of available data is key to develop predictive and prescriptive process mining approaches. Consequently, the thesis also elaborates on how predictive process monitoring can benefit from the promising trend of deep learning (research paper #6).

Table of Contents

I. Introduction and Motivation[1]

Business process management (BPM) is an accepted paradigm of organizational design and a source of corporate performance (Dumas et al. 2018). Due to substantial progress in process identification, analysis, implementation, and improvement (Recker and Mendling 2016; Vanwersch et al. 2016), BPM receives constant attention from industry (Harmon 2020). In times of market consolidation and increasing competition, operational excellence (i.e., continuously optimizing an organization's processes in terms of effectiveness and efficiency) is key to staying competitive. While traditional research in BPM focused on process models and model-based information systems (e.g., workflow management systems), recently, the focus has shifted to data-driven methods such as process mining (Diba et al. 2020). In contrast to model-driven BPM, process mining uses execution data in the form of events arising during process enactment, which may be exploited in several ways (van der Aalst 2016). Process mining strives to discover, monitor, and improve processes by extracting knowledge from event logs available in information systems (van der Aalst et al. 2011a). The most commonly applied use case in process mining is discovering as-is process models that also serve as a starting point for more detailed analysis (van der Aalst 2020). Based on the mined as-is-process, the use case of conformance checking helps to point out deviations from normative, predefined process models and actual process enactments (e.g., unintended handover of tasks, skipped activities, missed performance goals). As process mining analyzes information on an event-level, it also helps evaluate the actual process performance (e.g., measuring cycle times, interruptions, exceptions). In sum, process mining can help to ensure process hygiene, constituting a fundamental requirement to achieve operational excellence (van der Aalst 2020).

As process mining is one of the most active streams in BPM, numerous approaches have been proposed in the last decade, and various commercial vendors transferred these methods into practice, substantially facilitating event data analysis (Viner et al. 2020). At the tip of the iceberg, Celonis expanded in only seven years from start-up to a unicorn, indicating the enormous cross-industry business potential of process mining (Browne 2019). By 2023, Markets and Markets predicts a market potential of 1.42 billion US$ for process mining technologies (Research and Markets 2020). However, there are still numerous unsolved challenges that hinder the further adoption and usage of process mining at the enterprise level (vom Brocke et al. 2020). First, finding, extracting, and preprocessing relevant event data is still challenging and requires a significant amount

[1] This Section is partly comprised of content taken from the research papers included in this thesis. To improve the readability of the text, I omit the standard labeling of these citations.

of time in a process mining project and, thus, remains a bottleneck without providing appropriate support (Li et al. 2015). Second, most process mining approaches operate on a single-process level, but organizations are confronted with a process network covering hundreds of interdependent processes (vom Brocke et al. 2020). Third, process managers strongly require forward-directed operational support, but most process mining approaches provide only descriptive ex-post insights, e.g., discovered models or performance analysis of a past period (van der Aalst 2020). Since these challenges mainly drive this doctoral thesis, they will be discussed in detail below.

First, finding, extracting, and preprocessing relevant event data is still challenging. This is most frequently due to the lack of domain knowledge about the process, the distributed storage of required data in different databases and tables, and the requirement of advanced data engineering skills (Li et al. 2015). Most recent process mining approaches assume high-quality event logs without describing how such logs can be extracted from process-aware (PAIS) and particularly non-process-aware information systems (non-PAIS) (Suriadi et al. 2017, Wynn et al. 2017). In case of solely relying on process-aware information systems (PAIS) that directly output minable event logs, the risks of neglecting process-relevant information arise, and so-called blind spots can occur. For instance, if processes contain activities enacted by physical resources or software bots that are not directly connected to PAIS, details of these enactments cannot be explored using classical PAIS-based event logs. Due to increasingly digitized organizations, a growing part of the available data is highly unstructured (e.g., text, video, or audio files) and requires the application of novel concepts (van der Aalst 2020). To sum up, although process mining approaches significantly matured in the last decade, the step of data extraction is still too weekly supported and often results in bottlenecks that negatively affect the quality of process mining analysis.

Second, most process mining approaches operate on a single-process level. However, process mining currently evolves from project-based single-process analysis to an enterprise-wide ongoing task (van der Aalst 2020). Thus, methods for scaling process mining approaches on an enterprise-level are needed (vom Brocke et al. 2020). One of the most challenging topics relies on applying process mining methods that operate mostly on a single-process-perspective to enterprise-wide process networks, frequently covering hundreds of highly interconnected processes. Typically, process mining initiatives consume substantial resources, such as computing resources, but also expensive experts such as process owners or business analysts. Event-data-driven process prioritization approaches considering process interdependencies can be the missing part of the puzzle to ensure allocating these scarce resources to the most critical and central processes.

Third, process managers strongly require forward-directed operational support. Traditionally, process mining approaches focused on historical data for backward-looking, descriptive process mining (e.g., discovering process models). Backward-looking, descriptive process mining is an excellent starting point to improve processes, however, process managers need operational in their forward-looking day-to-day business (van der Aalst 2020). As exemplary forward-looking predictive process mining use cases, predictive and prescriptive process monitoring are growing in importance (Maggi et al. 2014). Predicting the behavior, performance, and outcomes of process instances help organizations act proactively in fast-changing environments. By combining process predictions with the decision area from normative process data (e.g., performance thresholds), prescriptive process mining approaches are able to trigger actions autonomously, e.g., by scheduling improvement projects (van der Aalst 2020). The increasing volume of data (i.e., event records and event properties) offers new opportunities and poses great challenges for predictive monitoring methods. As most approaches are still based on classical machine learning (ML) techniques such as decision trees (Evermann et al. 2016), their performance heavily depends on manual feature engineering in low-level feature representations (Goodfellow et al. 2016). Deep learning (DL) has proven its potential to exploit sensible and robust predictions based on nearly unprocessed, low-level input data in diverse applications (e.g., autonomous driving). Also, from a BPM perspective, DL promises to leverage the rapidly increasing volume of event data for predictive purposes. However, the rare use of DL, especially for outcome-oriented predictive process monitoring, reflects a lack of understanding about when the use of DL is sensible.

Visualized in Figure 1, BPM strives for connecting the real-world – physical in nature – with the digital world enabling value co-creation between human beings and machines (i.e., physical machines or software systems). The physical world consists of actors interacting with physical resources. Commonly, actors and resources are orchestrated through processes that relate to PAIS, creating a digital footprint (i.e., events) of each performed process activity. As the digital world's central element, the event log can be seen as a digital twin of the actual processes. Physical actors might also interact with non-PAIS or perform manual activities that are not connected with the digital world and, consequently, are not covered by PAIS-generated logs.

Inspired by the three challenges introduced above, this cumulative doctoral thesis consists of six research papers. Research paper #1 (RP#1) to research paper #3 (RP#3) are mainly related to the first challenge of finding, merging, and extracting event data from various data sources. Research paper #4 (RP#4) and research paper #5 (RP#5) cover data-driven process prioritization, helping to focus on the most critical processes in interdependent process networks and can be assigned to

prescriptive process mining. Finally, research paper #6 (RP#6) strives for providing guidelines on sensible usage of DL in predictive process monitoring and thus features predictive process mining.

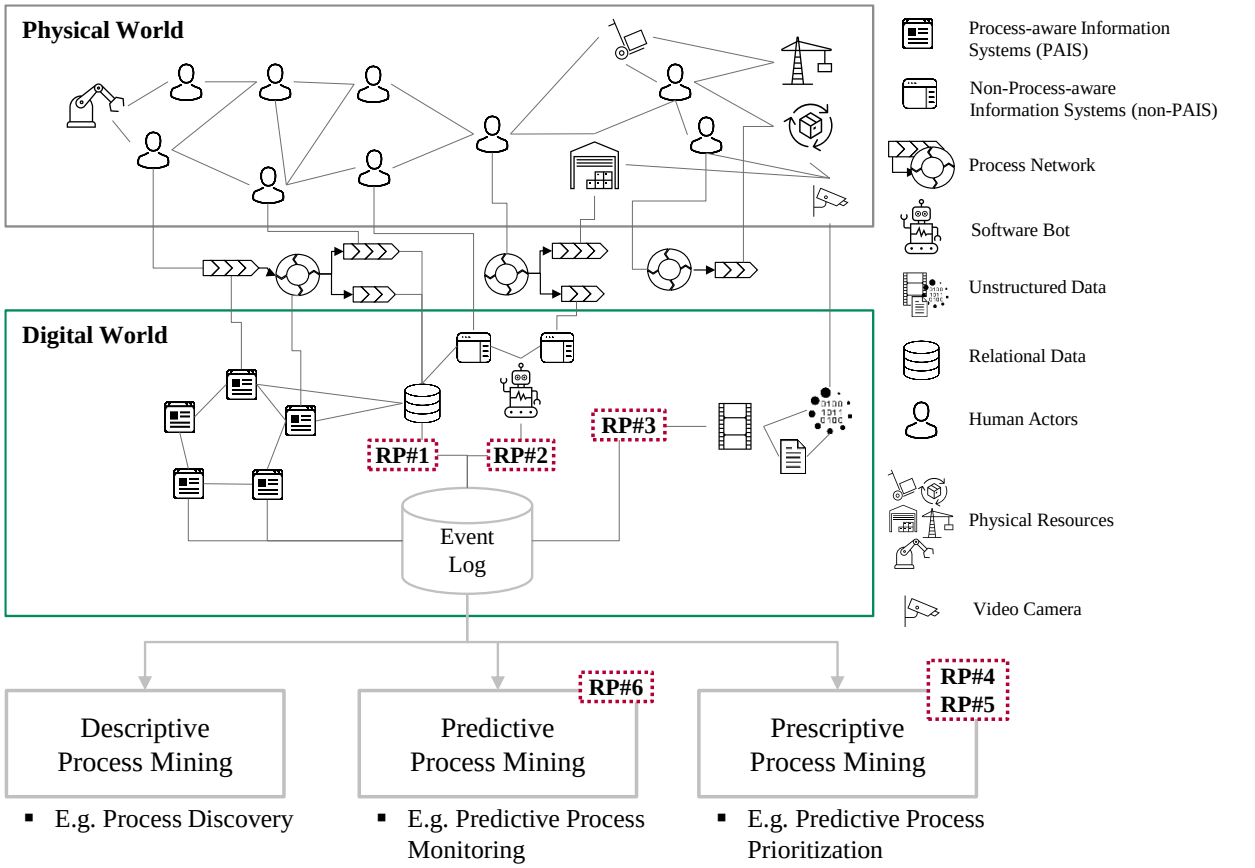


Figure 1: Assignment of individual Research Papers to forward-directed Process Mining

After introducing essential foundations and definitions of this thesis in Section II, Section III follows the structure outlined in Figure 1. First, the thesis addresses the log extraction of structured data by proposing a method supporting the quality-informed event log extraction from commonly used relational databases and a method merging process logs and bot logs stemming from robotic process automation to allow for integrated analysis (Section 1.1, including RP#1 and RP#2). Furthermore, the issue of blind spots due to physical processes running isolated from the digital world is addressed by developing a reference architecture to use unstructured video data for process mining (Section 1.2, including RP#3). All three approaches address the first challenge of finding, extracting, and preprocessing relevant event data. Second, to address the challenge that most process mining approaches operate on a single-process level, the thesis proposes two methods that use log data to prioritize processes and schedule in-depth analysis or improvement projects (Section 2.1, including RP#4 and RP#5). The proposed methods help bring the focus of process mining from a

single process perspective to an ongoing task at the enterprise level. Third, the challenge that process managers strongly require forward-directed operational support is addressed by systematically comparing DL with classical ML approaches using different publicly available data sets (Section 2.2, including RP#6). Finally, Section III summarizes the key insights and provides indications for future research. In addition to the publication bibliography in Section IV, Section V reports on additional information on all research papers (V.1), my individual contribution (V.2), and the research papers themselves (V.3 – V.8).

II. Foundations and Definitions[2]

1 Business Process Management and Process Mining

Business Process Management (BPM) is the art and science of overseeing how work is performed to ensure consistent outcomes and take advantage of improvement opportunities (Dumas et al. 2018). BPM combines knowledge from information technology and management sciences (van der Aalst 2013). By connecting the physical world with the digital world, BPM strives to coordinate value co-creation and information flow between human beings and machines (i.e., physical machines or information systems). In practice, hardly any process is executed in isolation. Instead, processes are organized in independent process networks (Lehnert et al. 2018). Hence, understanding process dependencies is key for decision-makers (Dijkman et al. 2016).

BPM activities are commonly organized along lifecycle phases, such as identification, discovery, analysis, improvement, implementation, monitoring, and controlling (Dumas et al. 2018). Required capabilities are structured six so-called core elements of BPM, namely, Strategic Alignment, Governance, Methods, Information Technology (IT), People, and Culture (Rosemann and vom Brocke 2015). With the increasing availability of data, novel capabilities such as process data analytics or advanced process automation have emerged (Kerpedzhiev et al. 2020). Thereby, big data analysis is considered one of the most promising technologies for BPM (Beverungen et al. 2020). Thus, whereas the origin of BPM relied on model-driven approaches, methods centered around process data become increasingly important.

As expanding data-driven research stream in BPM, process mining strives to discover, monitor, and improve processes by extracting knowledge from process logs (also referred to as event logs) commonly available in information systems such as enterprise resource planning (ERP) or customer relationship management (CRM) systems (van der Aalst et al. 2011a). Process logs record series of process-related events, with each event referring to a distinct task in a process instance. Process logs store standard attributes such as event names, performing resources, timestamps but also additional information about events and their context (van der Aalst 2014; vom Brocke et al. 2016). Initially, the process mining manifesto defined three use cases, namely process discovery (generation of as-is models), conformance checking (comparing as-is against to-be models), and

[2] This Section is partly comprised of content taken from the research papers included in this thesis. To improve the readability of the text, I omit the standard labeling of these citations.

model enhancement (enriching existing models through log insights) (van der Aalst et al. 2011a). Figure 2 shows the general framework on process mining that was refined in 2016.

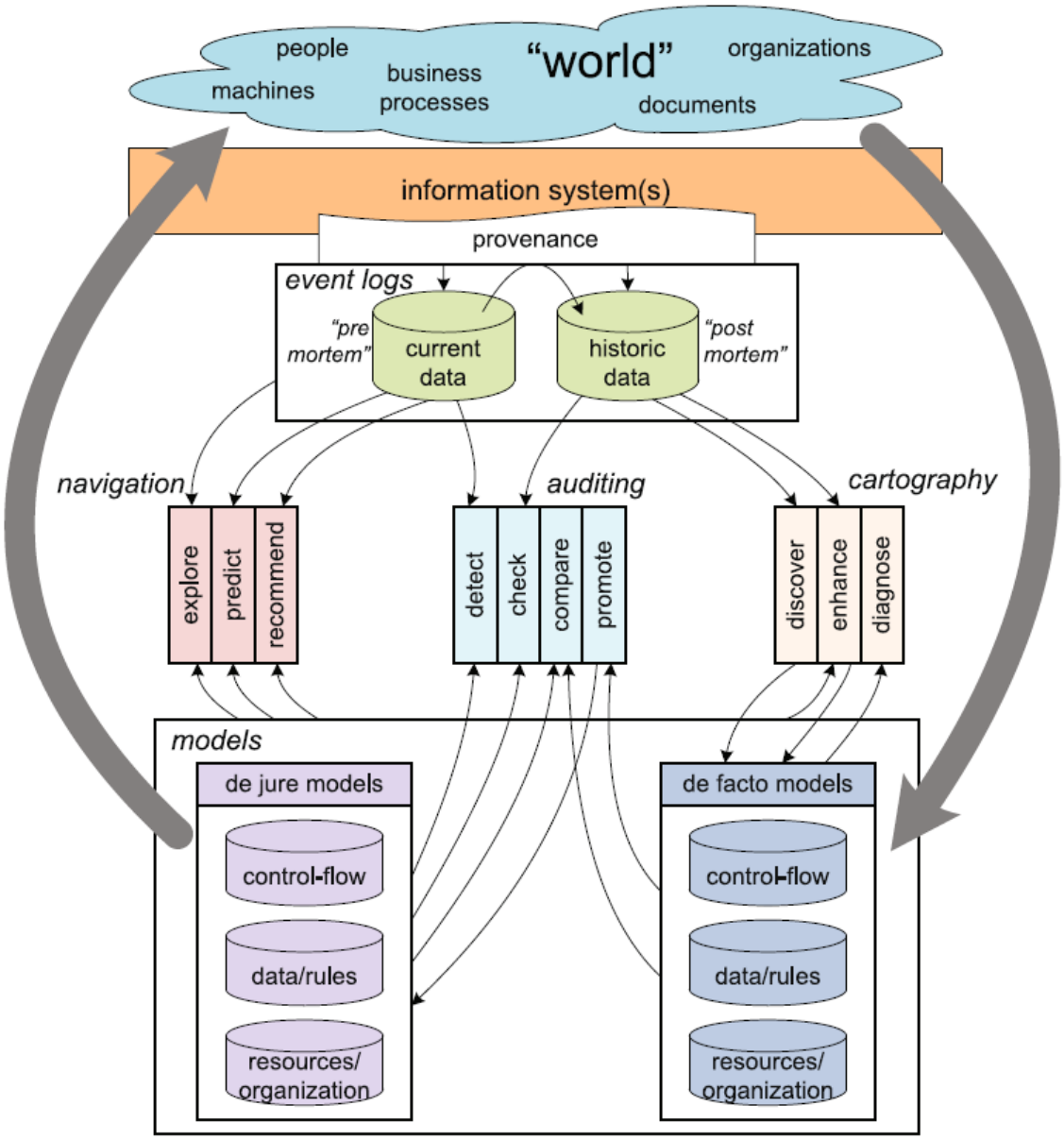


Figure 2: Process Mining Framework (van der Aalst 2016)

Generally, process mining use cases can be structured around required data sources (i.e., current or historical event data as well as normative or as-is process models) (van der Aalst 2016). Consequently, all use cases taking historical data can be subsumed as offline use cases. These offline use cases mainly relate to *cartography* (i.e., use cases working with process models or maps) and *auditing* (i.e., use cases that check whether business processes are executed within certain boundaries set by stakeholders). Some auditing use cases also take normative data (i.e., *de jure* models)

to explore deviance in historical cases. All offline approaches are descriptive, i.e., they analyze and visualize what happened in the past without looking into the future (van der Aalst 2016).

In recent years, the scope of process mining has evolved from mainly backward-looking analysis to forward-looking decision support (van der Aalst 2020). Thus, online use cases are increasingly becoming the focus of research. Such use cases are based on historical data and, in some cases, on partial trace information from ongoing cases, i.e., *pre-mortem* data. *Navigation*, as well as *auditing* use cases, take both data sources as input to infer statements about future process behavior. For example, predictive process monitoring applies predictive models to correlate extracted features from partial trace information with historical traces in real-time (Marquez-Chamorro et al. 2018). Predictive process monitoring approaches differ in terms of applied methods and, more importantly, the target of prediction (Di Francescomarino et al. 2018). Such targets can be the remaining cycle time of an ongoing case (van der Aalst et al. 2011b), the outcome (or an anomaly) of a case (Kratsch et al. 2020b), or the next action that will take place in further case processing (Schönig et al. 2018). The latter prediction task reveals that there is a fluid transition between *prediction* and *prescription*. By predicting the next action and identifying the decision area from normative process data (e.g., the *de jure* process model), one can set up models predicting each possible action's outcome and recommend the most favorable one (van der Aalst 2016). When such recommending models are combined with other automating process technologies – e.g., robotic process automation – the transition to prescriptive process analytics (e.g., models that determine what will happen) is fulfilled (van der Aalst 2020).

2 Machine Learning, Deep Learning, and Computer Vision

As forward-directed use cases, predictive and prescriptive process mining mainly base on techniques stemming from research areas around artificial intelligence (AI), namely machine learning (ML), deep learning (DL), and computer vision (CV). Since these AI-related terms have recently been incorrectly used as synonyms, this Section aims to distinguish them from each other clearly.

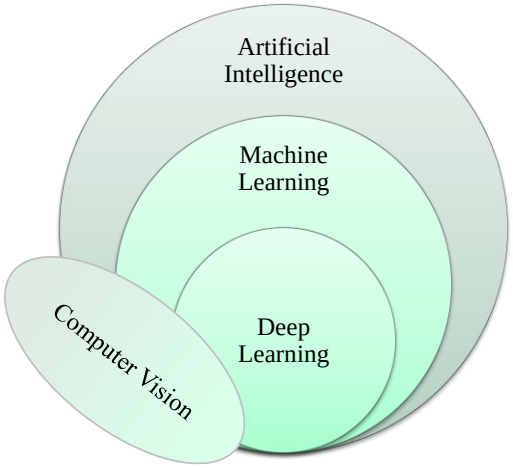


Figure 3: Venn diagram distinguishing AI-related terms

As depicted by Figure 3, AI is the umbrella term for automatically solving intellectual tasks usually performed by humans (Chollet 2018). ML is a subfield of artificial intelligence that uses real-world knowledge to make human-like decisions without defined rules (Goodfellow et al. 2016). DL adds the capability to automatically learn representations of data with multiple abstraction levels, enabling exploiting unstructured data (LeCun et al. 2015). Drawing on all of these concepts, CV is a specific use case to exploit visual data (Szeliski 2011).

ML uses statistical methods to learn structural patterns in typically large datasets in a (semi-) automated manner (Witten et al. 2017). Typical use cases of ML are classification, clustering, regression, and anomaly detection (Alpaydin 2020). Moreover, ML techniques can be divided into supervised and unsupervised learning (Alpaydin 2020). Supervised learning takes historical data that has already been classified by an external source and uses it for reproducing classifiers (Alpaydin 2020). By contrast, unsupervised learning algorithms process input data to gain insights by themselves (Alpaydin 2020). Whereas unsupervised learning is often used to group similar cases with an unclear definition of classes (e.g., for anomaly detection), supervised learning is appropriate when classifying cases according to predefined classes. Commonly applied ML techniques in supervised learning are, for example, random forests (Breiman 2001), logistic regressions (Hosmer et al. 2013), support vector machines (Cortes and Vapnik 1995), and shallow neural

networks (Haykin 2009). The performance of classical ML approaches is highly dependent on the representation of input data (Goodfellow et al. 2016). This aspect of preprocessing is also known as feature engineering (Witten et al. 2017).

DL reduces manual feature engineering effort by applying the divide-and-conquer principle, which introduces representations that are themselves expressed in terms of simpler representations (Goodfellow et al. 2016). DL is a relatively new term for the application of deep neural networks (DNN) (Witten et al. 2017). Until 2006, DNN were generally thought to be too difficult to train (Goodfellow et al. 2016). Innovations in algorithms and hardware have enabled the application of DL in productive services, for example, in recognition of a photo's location without geographical data in Google photos or video recommendations on YouTube (Schmidhuber 2015; Weyand et al. 2016; Covington et al. 2016).

CV applies mathematical techniques to visual data (e.g., images and videos), striving to achieve or even surpass human-like perceptual interpretation capabilities (Szeliski 2011; Microsoft Research 2019; Prince 2012). To date, CV has enabled several real-world use cases, including self-driving cars (Huval et al. 2015), facial recognition (Masi et al. 2018), and the analysis of medical images for healthcare (Gao et al. 2018). Figure 4 provides an overview of the most commonly applied CV capabilities (Voulodimos et al. 2018).

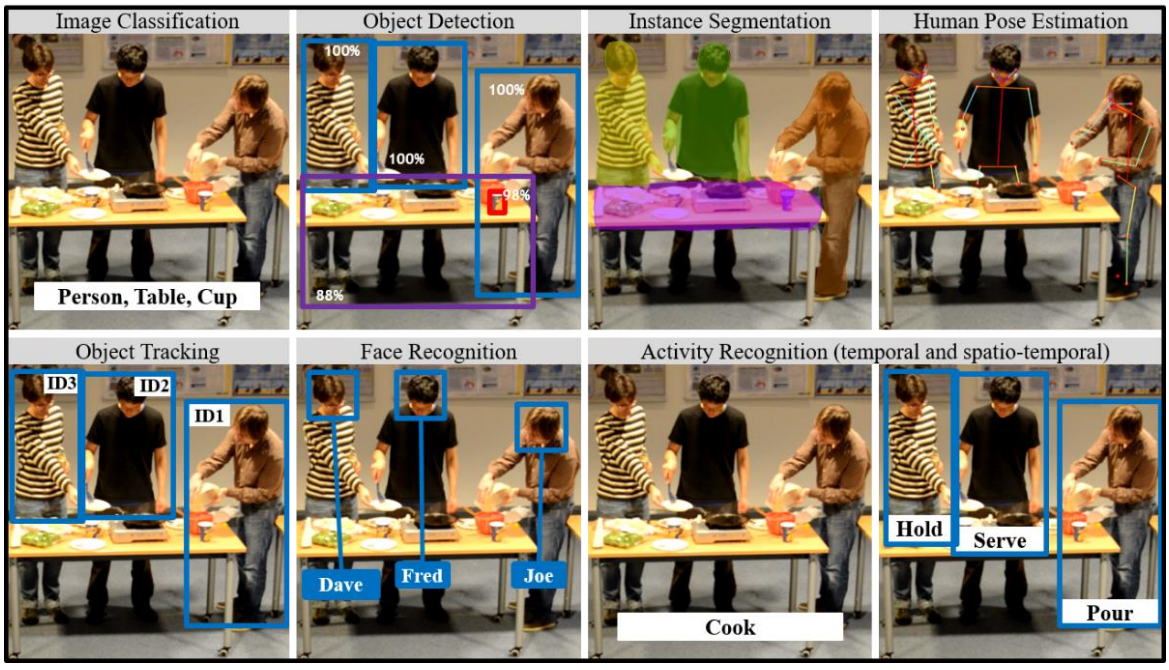


Figure 4: Illustration of commonly applied computer vision capabilities (Lee et al. 2015)

Image classification predicts probabilities for the occurrence of certain object classes (e.g., people) in images (Guo et al. 2016). Object detection aims to find instances of object classes and further

localizes their positions (Voulodimos et al. 2018). Instance segmentation additionally distinguishes between instances of the same object classes (Garcia-Garcia et al. 2018). Human pose estimation addresses the problem of localizing human body parts or anatomical key points (e.g., elbow, wrist) in images (Sun et al. 2019). Object tracking aims to find instances of object classes and further localize their positions (Voulodimos et al. 2018). Face recognition (re-)identifies individuals by their faces, which can be a challenging endeavor, e.g., due to head rotation, facial expression, or aging (Masi et al. 2018). As super-capability, activity recognition takes the output of other CV capabilities to identify the activities and actions of at least one person in an unfamiliar sequence of image frames (i.e., video) (Aggarwal and Ryoo 2011).

Most CV capabilities are based on algorithms and techniques that increasingly draw on DL. The improvements in hardware, the availability of large labeled datasets, and algorithmic advances have enabled the rise of DL in the CV area (Deng and Yu 2014). A substantial advantage of DL methods is the automation of feature engineering (Deng 2018). This gives DL methods successful generalization capabilities when provided with large labeled datasets (Kong and Fu 2018; Herath et al. 2017). On the other hand, training deep neural networks "from scratch" (i.e., with randomly initialized parameters) on smaller datasets may prove difficult due to the massive number of model parameters that have to be updated (Yim et al. 2017). To solve this issue, transfer learning (i.e., fine-tuning) is an approach that reuses the lower layers of on large datasets pre-trained neural networks. This is sensible, as the lower layers of deep neural networks only contain very unspecific information (e.g., accumulation of edges and shapes representing a human face) that is helpful in most domains (Chollet 2018). Thus, transfer learning helps to reach high predictive performance on small datasets with reasonable training effort (Shin et al. 2016).

III. Overview and Context of the Research Papers[3]

1 Finding, Merging, and Cleaning Event Data

1.1 Extraction of Structured Event Data

Finding, merging, and cleaning event data is the first challenge initially defined in the process mining manifesto (van der Aalst et al. 2011a). Although process mining research above-average addressed data extraction compared to other challenges, it is still one of the most challenging and time-consuming steps in process mining projects (Li et al. 2015). To lower the barriers for non-data-engineers to extract appropriate event logs, research paper #1 presents RDB2Log, a semi-automated, quality-informed approach to event log generation from relational databases. RDB2Log takes a relational database as input and generates an assessment of its data quality based on common data quality dimensions. By offering this data quality assessment, RDB2Log supports mapping data columns to event log attributes and generating an appropriate event log. The artifact proposed in this research paper #1 is envisioned as a step towards a process data quality lifecycle: systematic detection, repair, and tracking of data quality issues. By providing a graphical interface that helps users extract high-quality event logs, research paper #1 also strives to improve usability for non-experts.

Figure 5 shows the architectural overview of RDB2Log. The RDB2Log exploits database constraints and data quality assessments to support the semi-automated extraction of event records. To do so, database relationship assessments are performed (i.e., by evaluating primary key to foreign key relationships). Furthermore, RDB2Log uses a metrics-based concept to assess the data quality of event attribute candidates. Using this automated quality assessment, the user can decide which attributes are mapped to which events. Thus, the mappings are (i) table by table , (ii) user-selected, and (iii) quality-informed. That is, each database column is assessed against several quality dimensions for each event log attribute role assignment while taking into account acceptance criteria based on one or more quality dimensions and threshold values.

[3] This Section is partly comprised of content taken from the research papers included in this thesis. To improve the readability of the text, I omit the standard labeling of these citations.

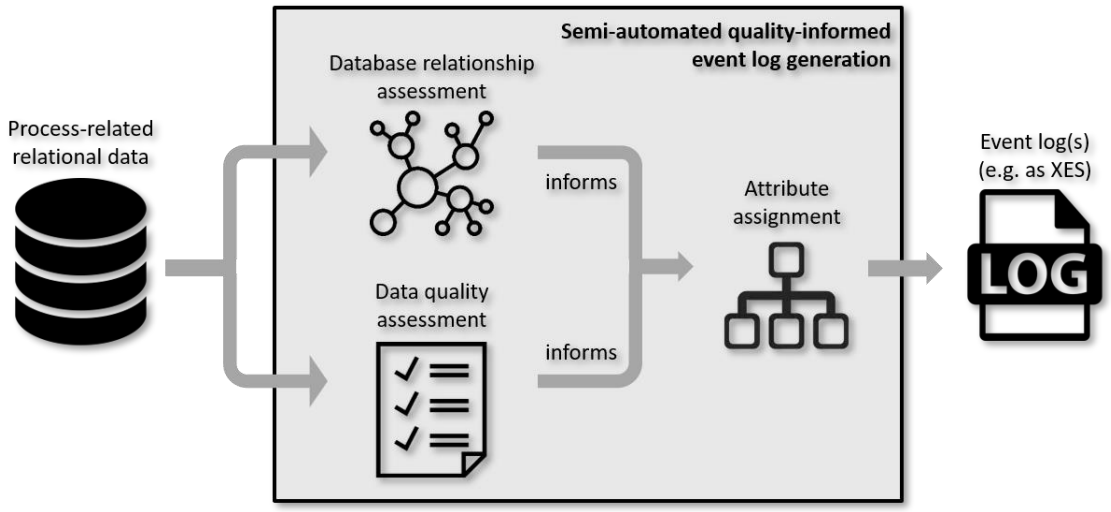


Figure 5: RDB2Log - Quality-informed event log generation (Andrews et al. 2020)

Regarding data quality assessment, RDB2Log incorporates 12 well-established data quality dimensions that have been deemed quantifiable and also relevant for process mining. To retain clarity, research paper #1 focuses on three dimensions that are exceptionally relevant for process mining: precision, uniqueness, and completeness. The operationalization of the dimensions is not specific to RDB2Log and can be modified. Most metrics can be computed on a data column level and can therefore inform quality dimensions on an attribute level as well as at an overall log level. Others contribute only to log level assessment.

To evaluate the RDB2Log, research paper #1 applied the DSR evaluation framework proposed by Sonnenberg and vom Brocke (2012), striving for ex-ante and ex-post evaluation. Regarding ex-ante evaluation, design objectives from existing knowledge have been derived. The artifact's design specification was discussed against competing artifacts, and its understandability and real-world fidelity have been challenged with process mining experts from industry and academia. For ex-post evaluation, RDB2Log was implemented as a software prototype and applied to two data sets in a laboratory setting (Figure 6). To provide a naturalistic setting, research paper #6 also applies the prototype to real-world data of a medium-sized manufacturing company and reports on the results as well as discussions with relevant stakeholders.

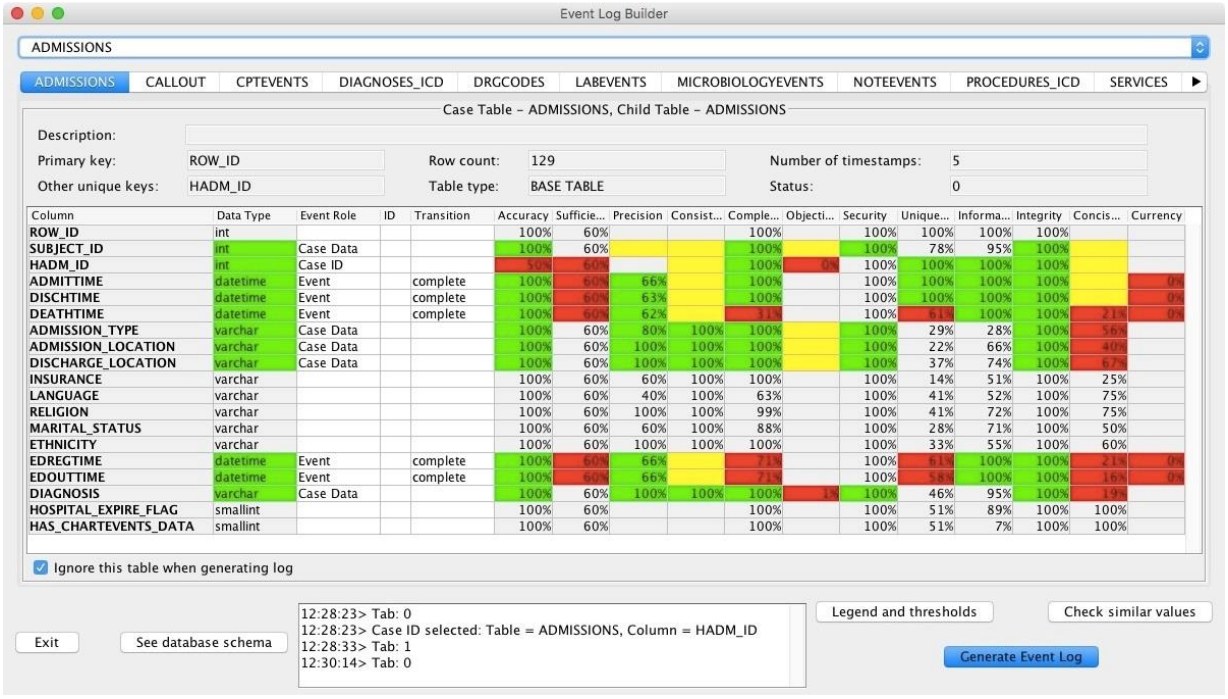


Figure 6: Example screenshot of the event log attribute selection using RDB2Log (Andrews et al. 2020)

While traditional information systems strongly rely on relational data, upcoming process technologies such as robotic process automation (RPA) often operate using event streams without storing these events to relational databases. Consequently, with the increasing handover of process tasks to software robots, blind spots in traditional PAIS are growing. Furthermore, due to a growing number of software bots, interdependencies in process networks become increasingly important. Hence, to obtain an end-to-end process perspective, process mining must consider log data stemming from RPA, namely bot logs.

Therefore, research paper #2 proposes an approach enabling integrated analysis using bot and process logs that provides new insights into bot-human interaction. An integrated analysis of bot and process data can also show the effects of bots on business processes and explore how exceptions are handled. Joint data analysis of bot and process data might also benefit the redesign of bots used in business processes. As a central artifact, research paper #2 proposes an integrated conceptual data model specifying the relations between bots and business processes. Based on this data model, it is possible to merge bot logs and process logs, allowing for integrated analysis. Figure 7 shows an overview of the proposed approach in research paper #2.

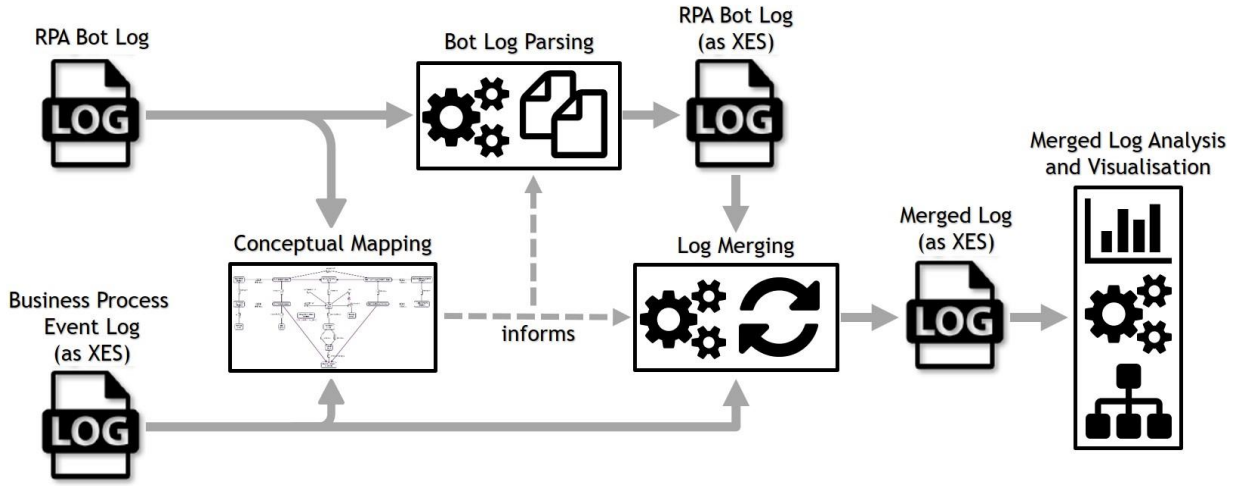


Figure 7: Bot Log Mining Approach (Egger et al. 2020)

A merged log provides opportunities for a more detailed analysis of the underlying processes. By creating new measures and visualizations for process mining that use a merged log as input, it is possible to provide useful information on the underlying partly automated processes as output. There are many possibilities for new measures. However, research paper #2 showcases two exemplary measures, namely *Exception Time Impact* (ETI) and *Relative Fails* (RF), illustrating the underlying approach's concept.

Figure 8 visualizes how ETI can be applied to indicate for every activity how much longer a trace takes, on average, to complete if the activity fails compared to if the activity does not fail. The colors refer to whether the activity was always executed by bots (blue), always manually (green), or both (yellow). The darker the color, the higher is the relative failure rate. Take, for instance, the activity "finish editing": When this activity fails, on average, it takes 70 days longer to end the whole process compared to when "finish editing" did not fail, which indicates that "finish editing" seems to be correlated with a longer remaining duration in the process. Furthermore, when a bot executes "finish editing", the bot activity "check spelling" seems to be correlated with a longer remaining duration, since if this bot activity fails, on average, it takes 92 days longer to end the whole process. This information is useful for bot redesign: stakeholders could consider revising "check spelling" in the bot process. Without merging bot data to process data, this insight would have been lost behind a blind spot.

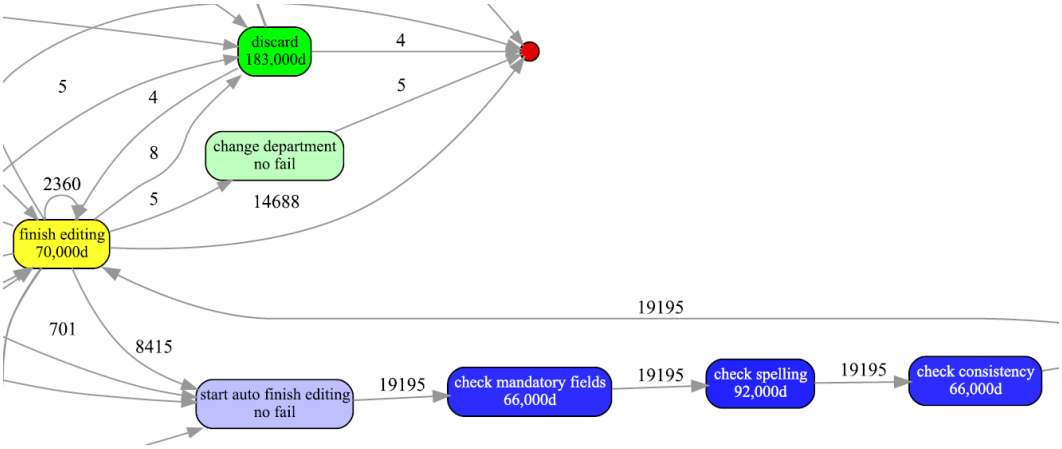


Figure 8: Directly follows graph enriched with information from the merged log

1.2 Extraction of Unstructured Event Data

As described in Section III.1.1, current process mining applications highly rely on structured business data, often gathered from process-aware information systems (PAIS) or other enterprise information systems (e.g., ERP, CRM). However, according to *Forbes* and *CIO* magazine, 80 to 90 percent of available data is unstructured, i.e., data without a functional, retrievable data scheme (Davis 2019; Marr 2019). Handling uncertain, continuous, and unstructured data adds another dimension to the challenge of finding, merging, and cleaning event data that requires the application of novel concepts (van der Aalst 2020). Moreover, unstructured data is increasing much more rapidly than structured data (Marr 2019). Consequently, most process mining analyses only exploit 10 to 20 percent of the available data. Although it only uses a small percentage of available data, applying process mining to highly digitalized processes (e.g., ERP processes) is already a mature practice, yet still only targets a small proportion of existing business processes. On the other hand, there are hardly digitized processes that may contain numerous manual activities. Since manual activities are not usually tracked in PAIS, it is rarely possible to mine such processes using existing approaches. The results are undiscoverable blind spots.

However, in many cases, vast amounts of unstructured data (e.g., media files or text documents) related to these blind spots are available. Consequently, all of the academic experts questioned in a recent Delphi study stated that BPM should prioritize the exploration of unstructured data (Kerpedzhiev et al. 2020). Initial approaches propose techniques to make unstructured data usable for process mining. Some of these techniques apply natural language processing (NLP) to text documents (van der Aa et al. 2018). However, sensor-based approaches cannot be scaled for use

in broader contexts as measured values are dependent on the deployment location. Furthermore, full equipment with sensors appears to be an unrealistic scenario having broad system boundaries or open systems, e.g., when external actors are included. In contrast, NLP-based approaches are much easier to generalize but – just like structured log data – describe only activities performed within information systems (e.g., mail systems). Video data (e.g., from surveillance cameras) bears the potential to make processes that partly run away from information systems (i.e., blind spots) more observable.

Initial technically-driven approaches support the use of video data for specific use cases (e.g., object detection and activity recognition) in highly specific contexts, e.g., production and logistics, often in laboratory settings (Reining et al. 2019). Most recent CV approaches build on DL techniques that have led to technological breakthroughs in the course of their productive application (e.g., Tesla's autopilot (Tesla 2020) and optical football-tracking of Track160 (2020)). These examples suggest that DL-enabled CV could be the key to extracting, piece-by-piece, structured information (e.g., a traffic sign or the position of a football player) from a vast amount of unstructured data (e.g., eight high definition camera streams, in the case of Tesla). Having extracted structured features and their temporal contexts, existing approaches (e.g., the use of distance measures to calculate collision potential with other cars or off-side positions of football players) can process and analyze this information efficiently. Transferred to process mining, events and actors extracted from video data could feed into structured event logs, to which the various existing process mining approaches can be applied. Thus, using video data as a basis for process mining approaches could help to reduce blind spots.

Research paper #3 proposes a Video Mining reference architecture (Figure 9), consisting of the three subsystem layers *Data Preprocessor*, *Information Extractor*, and *Event Processor*. Since the reference architecture is configurable, optional components are indicated by dotted frames and different instantiation variants are highlighted in color. By producing an event log and offering an event notification service, the RA connects to various BPM applications to support diverse process mining use cases. The Data Preprocessor serves as an interface to the input data. The Information Extractor receives these frames to perform different CV capabilities that hierarchically extract meaningful information. Taking the extracted low-level events as input, *the Event Processor* applies event generalization and abstraction concepts to output high-level business events.

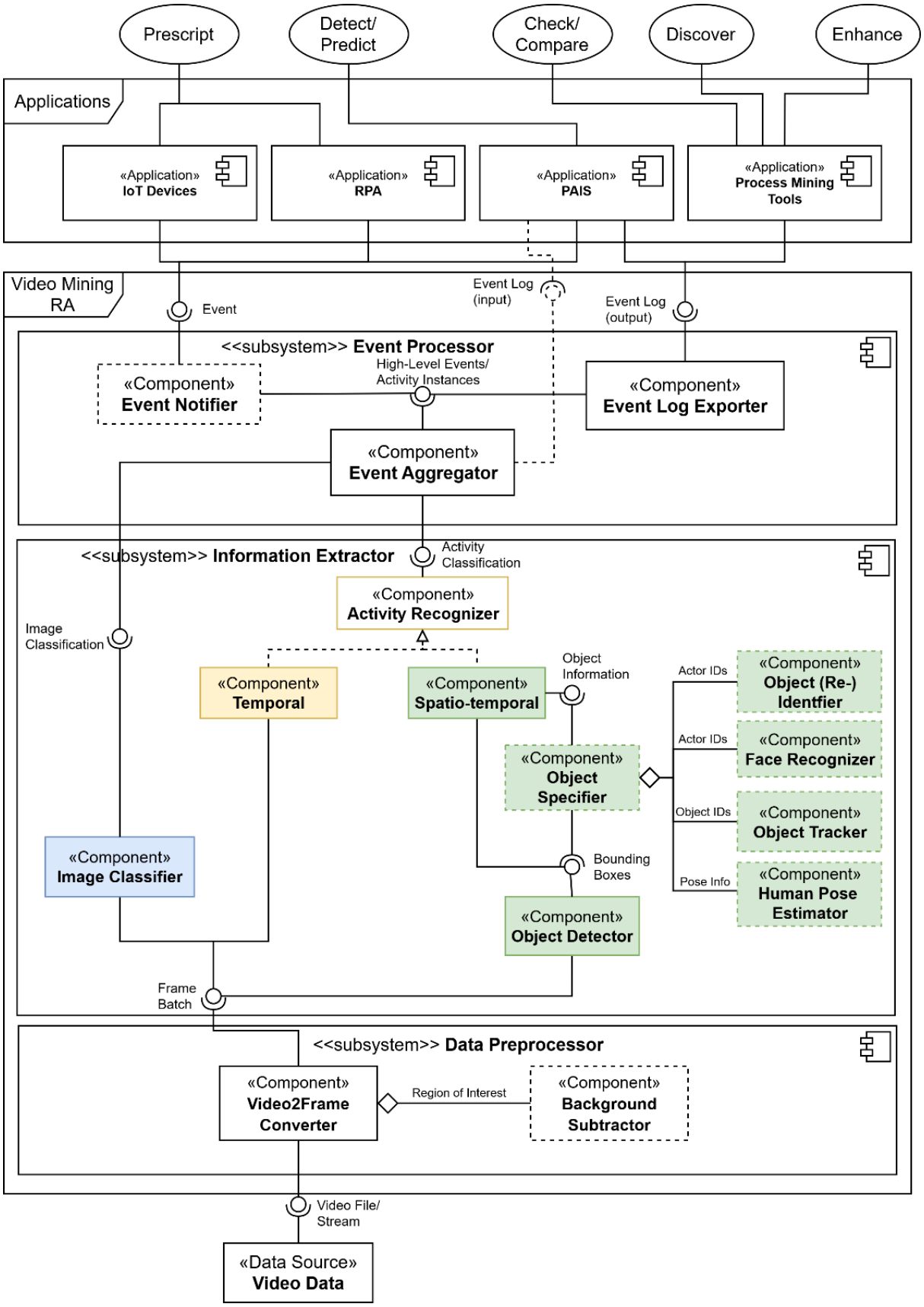


Figure 9: Video Process Mining Reference Architecture (Kratsch et al. 2020a)

Figure 10 visualizes the output of the prototype used to evaluate the proposed Video Mining Reference Architecture. Using several CV capabilities, the prototype is able to identify different actors and track activities they are performing. On the right side of Figure 10, the predicted event data is visualized using a directly follows graph in Disco. The three tokens represent the three actors, and it is evident that Actor 2 (the upper token) is behind and still finishes stirring the dough, whereas Actors 1 and 3 are already pouring.

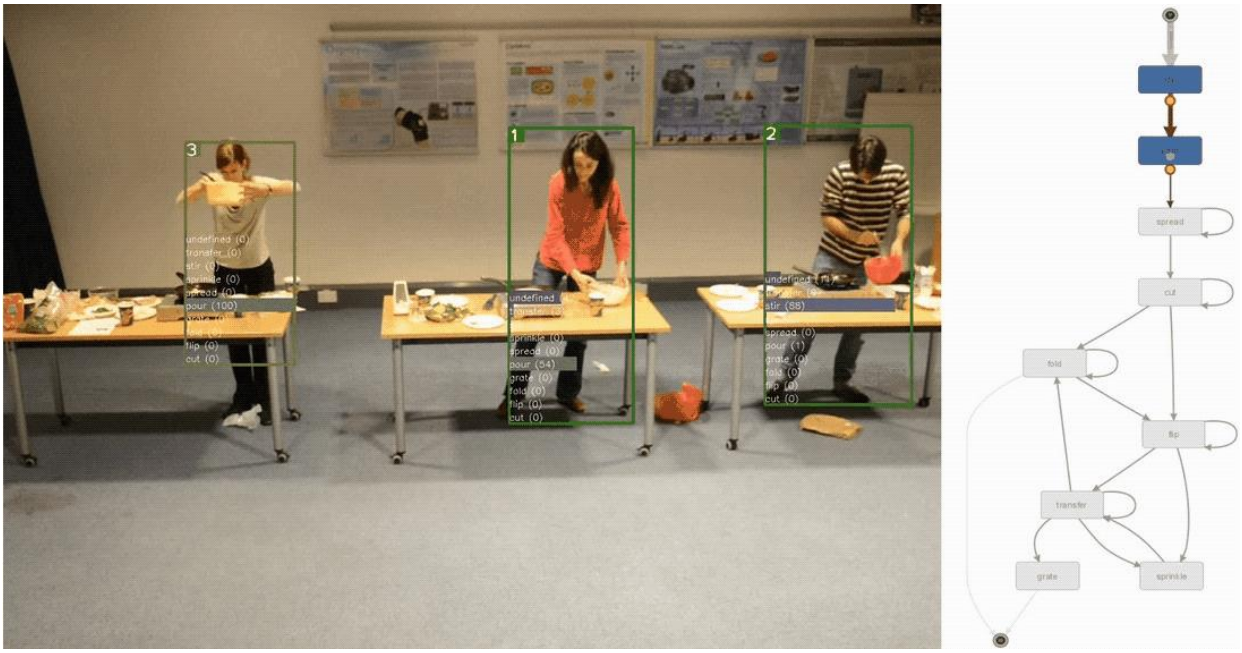


Figure 10: Evaluation video showing the instantiated Video Mining Reference Architecture in action

To conclude Section III.1, research papers #1 to #3 address the challenge of finding, extracting and preprocessing relevant event data as an essential prerequisite to successfully applying process mining techniques. Research paper #1 guides the quality-informed extraction of event logs from relational databases, where most event data of PAIS systems such as SAP ERP is stored. Research paper #2 and #3 aim to exploit novel data sources for process mining purposes. Extending the step of event data extraction to bot logs (i.e., research paper #2) is a promising approach to make human-bot interaction explorable and support the reasonable and sustainable automation of business processes. By providing an initial idea how video data can be leveraged for process mining purposes, research paper #3 strives to exploit valuable process-relevant information beyond structured data sources bearing the potential to broaden the coverage of process mining analysis substantially.

2 Novel Approaches for Predictive and Prescriptive Process Monitoring

2.1 Prescriptive Prioritization of Interdependent Processes

Regarding process prioritization, the BPM literature offers multiple approaches (Bandara et al. 2015). Extant approaches can be split into performance- and non-performance-based approaches. Non-performance-based approaches prioritize processes using criteria such as urgency, strategic importance, or difficulty of improvement (Hanafizadeh and Osouli 2011). Performance-based approaches prioritize processes by quantifying their actual and target performance, deriving their need for improvement, and ranking them (Leyer et al. 2015). When multiple processes must be arranged and orchestrated, structural process dependencies arise (e.g., core processes use support processes) (Dijkman et al., 2016). Besides structural dependencies, processes are subject to stochastic dependencies (Letmathe et al. 2013). Process logs may not only include data about tasks, paths, and task performance but also about structural and stochastic process dependencies (Wen et al. 2006). Logging the events related to multiple processes, relevant information such as the distribution of process costs, the frequency of core processing using support processes, or the autocorrelation of process instances can be mined. However, most performance-based process prioritization approaches use expert opinions instead of log data for performance-based process prioritization.

To decide which processes should be in focus of process mining initiatives, process prioritization can be applied. Research paper #4 address this gap by proposing the Data-driven Process Prioritization approach (D2P2), leveraging performance, and dependency data from process logs to determine the risky performance of all involved processes. Thereby, the D2P2 accounts for structural dependencies (e.g., processes that use other processes) and stochastic dependencies (e.g., instances that affect other instances of the same process). Based on the dependency-adjusted risky process performance, the D2P2 predicts when each process is likely to violate predefined performance thresholds and schedules it for in-depth analysis to future planning periods. Process analysts can then check whether the process under consideration requires improvement. Basing on event log data, the D2P2's output is more reliable and detailed than other process prioritization approaches.

The D2P2 prioritizes processes by leveraging performance data (i.e., process cash flows) and dependency data (e.g., how often processes use other processes) from process logs. As shown in Figure 11, the D2P2 includes three steps: (1) extraction of the involved processes' dependency-adjusted risky performance, (2) prediction of these processes' risky future performance, and (3) scheduling of the involved processes for in-depth analysis.

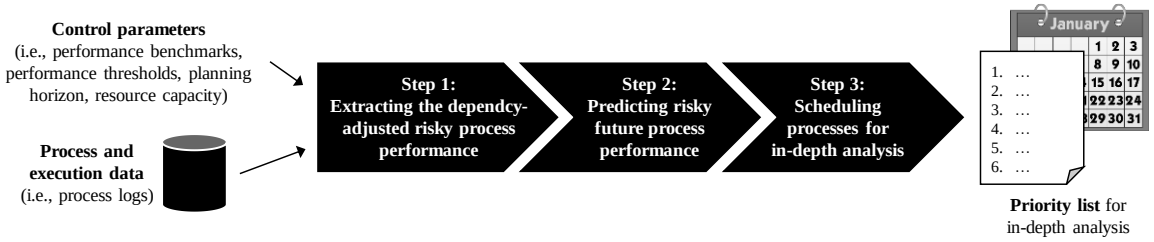


Figure 11: Overview of the D2P2 approach (Kratsch et al. 2017)

Two essential concepts of the D2P2 are process networks and process performance variants, particularly for determining the dependency-adjusted risky process performance in Step 1. Figure 12 on the left shows an exemplary process network, serving as a running example. In this example, process P1 uses P2 but can also be executed stand-alone. Thus, there is a directed edge from P1 to P2, representing a use dependency, and a self-directed edge for P1 capturing stand-alone executions.

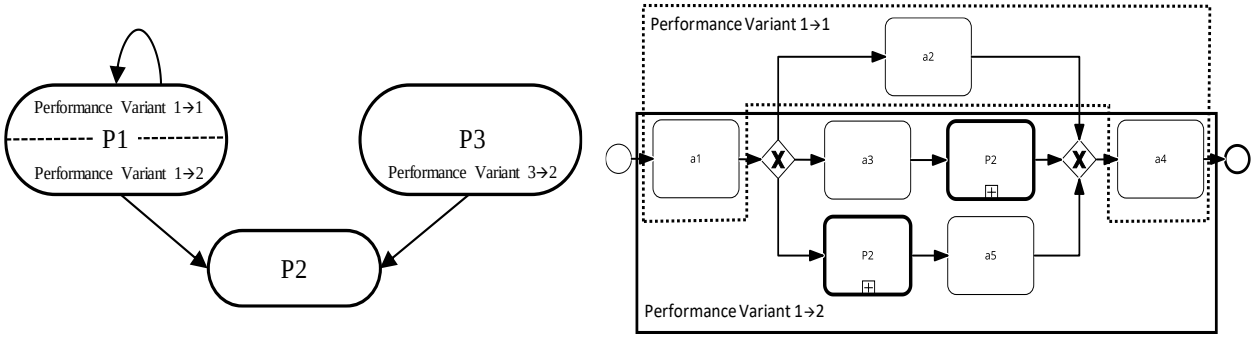


Figure 12: Exemplary process network (left) and its performance variants (right) (Kratsch et al. 2017)

Leveraging this information contained in process networks, each process' performance can be decomposed into performance variants basing on structural dependencies of process variants. For instance, Figure 10 on the right illustrates the performance variants of process P1 from our running example. P1 has two performance variants, i.e., 1→2 to capture the use dependency with P2 and 1→1 as P1 can be executed stand-alone. Performance variant 1→2 includes a common (i.e., a1 and a4) and a variant-specific part (i.e., a3, a5, and P2), which splits into an exclusive part (i.e., a3 and a5) and a part caused by using P2. To extract the dependency-adjusted performance variants out of process logs, the D2P2 builds on multi-variate regression analysis. Multi-variate regression analysis is commonly used to determine a functional relationship (dependency) between a dependent variable (i.e., a known process performance or part of it) and multiple independent variables (i.e., other parts of the process performance) (Freedman 2009).

As for Step 2, D2P2 predicts the processes' future dependency-adjusted risky performance. With process managers prioritizing processes by comparing their actual and target performance, the

D2P2 assesses over- and under-performance. As performance differences of individual instances are too fine-grained for process prioritization, the D2P2 is able to aggregate the performance difference of all instances. The aggregated difference is the D2P2's central indicator for determining when to schedule a process for an in-depth analysis. The aggregated difference is uncertain and may take any value. As a sum of random variables, the aggregated difference's value range is cone-shaped (Figure 13), i.e., its value range is small in the near future and continuously broadens in the more distant future.

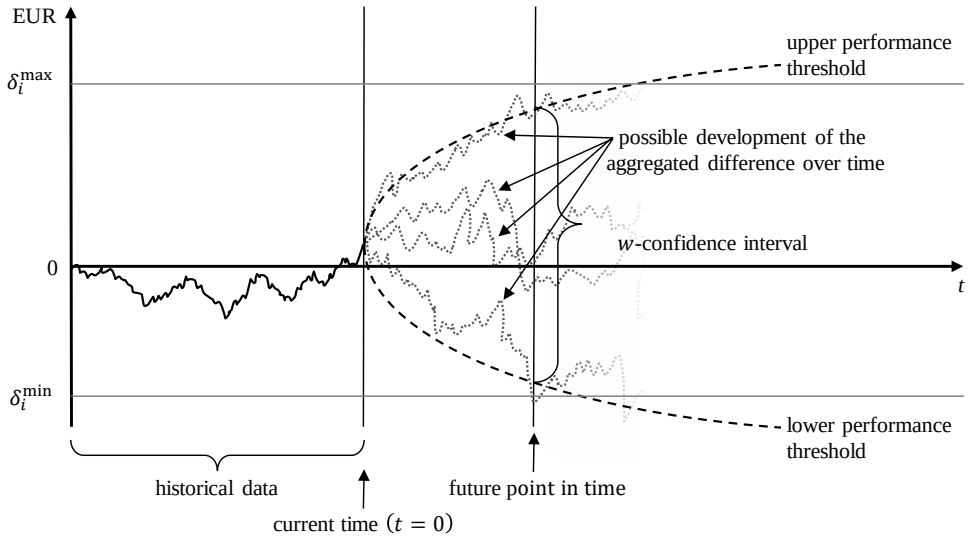


Figure 13: Cone-shaped structure of the predicted aggregated performance difference (Kratsch et al. 2017)

In Step 3, the D2P2 schedules the involved processes for an in-depth analysis based on their absolute aggregated under- or over-performance determined in Step 2. Thereby, a mixed-integer linear program seeks an assignment of in-depth analyses to planning periods that minimizes the opportunity costs for lost improvement potential in case of lower threshold violations and untapped opportunities in upper threshold violations. Applying the instantiated D2P2 to real-world data yields superior results compared to competing artifacts. Thus, by neglecting process dependencies and risky future process performance, process prioritization decisions are biased.

While D2P2 ends up providing a prioritized list of process candidates for an in-depth analysis, research paper #5 expands the scope of process prioritization to schedule improvement projects providing an even more prescriptive support. To do so, research paper #5 proposes the PMP2 drawing on the main concepts of D2P2 and extends an economic decision model optimizing the assignment of improvement project alternatives. By combining Markov reward models (MRM) and normative analytical modeling, PMP2 helps organizations determine business process im-

provement roadmaps (i.e., sequential implementation of improvement projects on business processes), which maximize an organization's long-term firm value while catering for process dependencies and interactions among projects. Thereby, PMP2 takes a multi-period, multi-process, and multi-project perspective. The PMP2 considers dependencies between processes and improvement projects and thus schedules improvement projects to optimize an organization's long-term firm value. Table 1 shows one of the evaluation scenarios indicating that the project dimension can significantly impact process prioritization. For this scenario, we assume different modification factors representing varying impacts. Analyzing the project impact solely, improving the cash flows of Process 3 has by far the highest effect (reduction by 40%). Even if factoring in the two-fold effect of lead time reductions and thus opting for that, the decision-maker would prioritize process 3, as a 20% reduction can be achieved. However, as shown in Table 1, conducting a 10% cash flow reduction of Process 1 is superior to all other projects. This outlines the importance of analyzing improvement projects and the underlying process network in an integrated manner, as independent analysis yields inferior results.

Table 1: Results of an exemplary scenario analysis (Bitomsky et al. 2019)

Basic Input - Trans. Rates		Dwelling time (hrs)	Transition Prob.
$\lambda_{1,2}$	0.04		0.4
$\lambda_{1,3}$	0.06		0.6
$\lambda_{1,4}$	0		0
		10	
$\lambda_{2,1}$	0.2		1
$\lambda_{2,3}$	0		0
$\lambda_{2,3}$	0		0
		5	
$\lambda_{3,1}$	0		0
$\lambda_{3,2}$	0		0
$\lambda_{3,4}$	0.5		1
		2	

Basic Input - Cash Flows			
cf_1	cf_2	cf_3	PV
-150	-100	-100	5000

Output			
Process	**Modification target**	**Modification factor**	**Relative change in %**
1	Cash flow	0.9	12.693
2	Cash flow	0.7	3.903
3	Cash flow	0.6	3.716
1	Transition rates	1.025	4.79
2	Transition rates	1.15	3.784
3	Transition rates	1.2	3.241

Roadmap		
T	***Project***	***Process***
1	1	1
2	2	1
3	1	1
4	1	2
5	2	3

Summarizing, process prioritization based on event log data can focus on the most central processes when scaling process mining initiatives to an enterprise level, thus addressing the second challenge that process mining approaches operate on a single-process level. Attributing to prescriptive process mining, the proposed approaches in research papers #4 and #5 also account for the third challenge of providing forward-directed operational support to process managers.

2.2 Using Deep Learning for Predictive Process Monitoring

When it comes to forward-directed process mining, predictive monitoring represents one of the most critical capabilities. In the end, prescriptive process mining approaches (such as prescriptive process prioritization as proposed in Section 2.1) are also based on predictive monitoring combined with predefined decision rules. Various predictive process monitoring approaches use machine learning (ML) techniques as, in contrast to rule-based monitoring techniques, there is no need to rely on subjective expert-defined decision rules (Kang et al. 2012). Moreover, the increasing availability of data lowers the barriers to the use of ML. Although the popularity of deep learning (DL) has increased in predictive process monitoring, most works still use classical ML techniques such as decision trees, random forests (RF), or support vector machines (SVM) (Evermann et al. 2016). However, a drawback of such techniques is that their performance heavily depends on manual feature engineering in case of low-level feature representations (Goodfellow et al. 2016). From a BPM perspective, DL promises to leverage process data for predictive purposes. However, the rare use of DL, especially for outcome-oriented predictive process monitoring, reflects a lack of understanding about when the use of DL is sensible.

Research paper #6 addresses this research gap by extensively comparing the performance of different ML (i.e., Random Forests and Support Vector Machines) and DL (i.e., simple feedforward Deep Neural Networks and Long Short Term Memory Networks) techniques for a diverse set of five publicly available logs in terms of established evaluation metrics (i.e., Accuracy, F-Score, and ROC AUC). To provide generalizable results, research paper #6 combines data-to-description and description-to-theory strategies (Yin 1994). Also referred to as Level-1 inference (Yin 1994), data-to-description generalization takes empirical data as input, condensed into higher-level yet still empirical observations or descriptions. This strategy also covers the well-known statistical sample-to-population generalization. Description-to-theory generalization, which is also referred to as analytical generalization or Level-2 inference (Yin 1994), aims at inferring theoretical statements in the form of propositions, i.e., "variables and the relationships among them" (Lee and Baskerville 2003, p. 236), from empirical observations or descriptions. As for Level-1 inference, research paper #6 analyzed the performance of the selected techniques per event log in terms of evaluation metrics and related statistical measures (i.e., mean and standard deviation). As for Level-2 inference, research paper #6 identified relationships between the techniques' performance across the logs and related these cross-log observations to the log properties.

To allow for Level-2 inference, it was necessary to develop a framework ensuring the purposeful sampling of event logs. Event logs can be classified according to their properties in terms of a data and a control-flow perspective. Figure 14 shows the used event logs' classification according to the control flow perspective, whereas Figure 15 illustrates the data perspective.

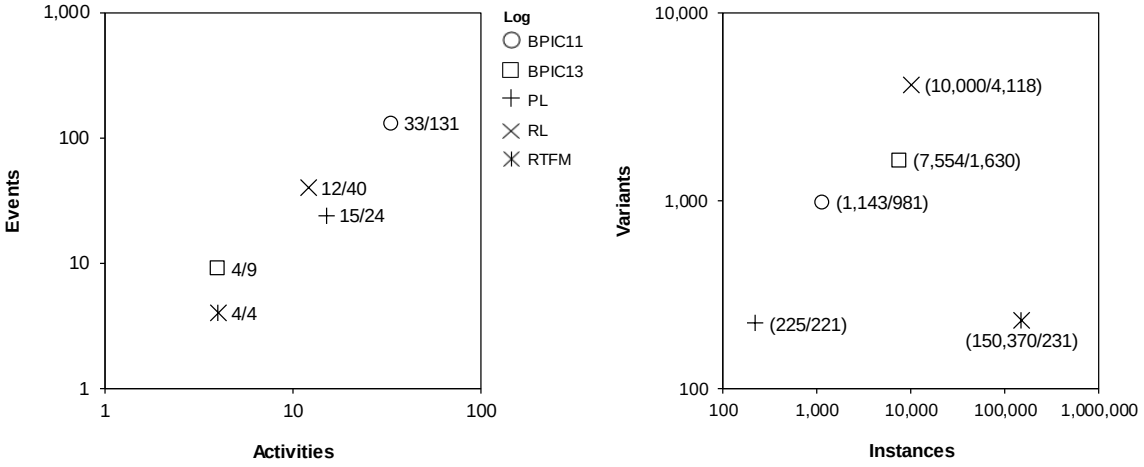


Figure 14: Control flow perspective of log classification: events-to-activity ratio (left, a), and variants-to-instances ratio (right, b) (logarithmic scales) (Kratsch et al. 2020b)

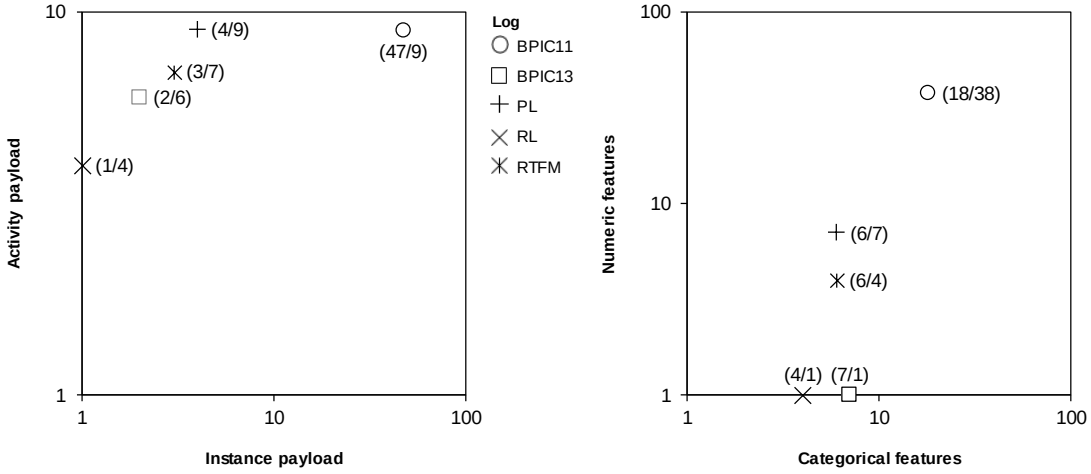
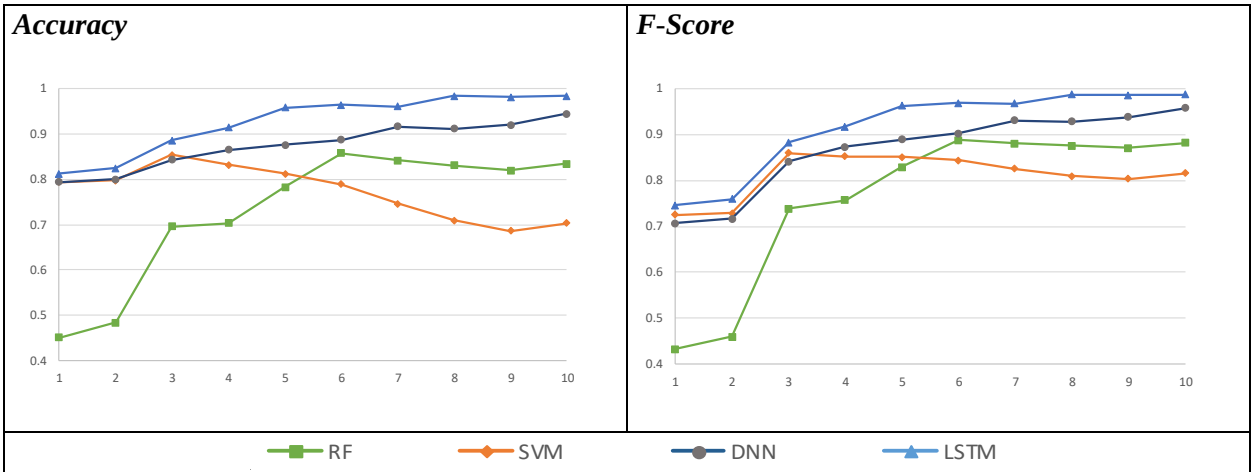


Figure 15: Data perspective of log classification: activity-to-instance ratio (left, a) and numeric-to-categorical ratio (right, b) of payload data (logarithmic scales) (Kratsch et al. 2020b)

Table 2 on the top shows classifiers' performance depending on the runtime of a process instance for an exemplary log (BPIC13). In this case, predictions are more uncertain at an early processing stage of a process instance than in later processing stages. Regarding Accuracy and F-Score, the DL techniques show higher overall accuracy and a lower standard deviation. Compared to DNN, LSTM shows a substantial dominance, especially in later prediction time points. Concerning the classical techniques, SVM shows advantages in earlier prediction time points, whereas RF yields

better results after the sixth activity. All classifiers deliver good results regarding the ROC AUC. The DL classifiers outperform the classical ML classifiers. However, DNN only slightly outperforms SVM, while RF falls behind. In general, DL techniques show higher temporal stability than RF and SVM. The performance advantage regarding the accuracy and the F-Score is especially high for earlier prediction time points. On the bottom, Table 3 reports overall prediction points aggregated performance measures emphasizing the overall outperformance of DL techniques, especially LSTM, for this exemplary event log.

Table 2: Performance analysis for an exemplary event log (BPIC13) (Kratsch et al. 2020b)



	Accuracy		*F-Score*		*ROC AUC*	
	Mean	***Std. Dev.***	***Mean***	***Std. Dev.***	***Mean***	***Std. Dev.***
RF	0.7299	0.1419	0.7616	0.1658	0.7807	0.1644
SVM	0.7725	0.0580	0.8118	**0.0485**	0.8482	**0.0423**
DNN	0.8761	**0.0511**	0.8688	0.0894	0.8697	0.0600
LSTM	**0.9269**	0.0620	**0.9166**	0.0881	**0.9226**	0.0688

Based on the analysis of the individual logs, the following observations can be made about the performance of the classifiers across the logs (i.e., Level-2 inference):

O1: *DL classifiers generally outperform classical ML classifiers regarding accuracy and F-Score.*

O2: *DL classifiers substantially outperform classical ML classifiers regarding accuracy and F-Score for logs with a high variant-to-instance ratio.*

O3: *DL classifiers substantially outperform classical ML classifiers regarding ROC AUC for logs with a high event-to-activity ratio and imbalanced class labels.*

O4: *LSTM substantially outperforms DNN regarding accuracy and F-Score for logs featuring a high activity-to-instance payload ratio.*

Based on these observations across all event logs, three specific propositions can be inferred:

- First, the outperformance of DL techniques is particularly strong for logs with a high variant-to-instance ratio (i.e., many non-standard cases).
- Second, DL techniques perform more stably in case of imbalanced target variables, especially for logs with a high event-to-activity ratio (i.e., many loops in the control flow).
- Third, logs with a high activity-to-instance payload ratio (i.e., input data is predominantly generated at runtime) call for the application of LSTM.

In sum, research paper #6 shows that DL can help inferring more reliable predictions out of increasing volumes of data. However, some cases provide more favorable application environments for DL than others. To allocate scarce resources to these cases, research paper #6 provides guidelines when the application of DL is sensible.

To conclude Section III.2, research papers #4 to #6 offer predictive and prescriptive process mining approaches and contribute to the challenge of providing process managers with forward-directed operational support. Besides, research papers #4 and #5 provide approaches to selecting a process network's most central processes for process mining initiatives and thus help bring process mining to an enterprise level.

IV. Summary and Future Research[4]

1 Summary

Recently, the focus in BPM has shifted from model-based to data-driven methods. Consequently, process mining, i.e., the data-driven analysis of event data, is one of the most active streams in BPM. Numerous approaches have been proposed in the last decade, and various commercial vendors transferred these methods into practice. However, there are still unsolved challenges that hinder the further adoption and usage of process mining at the enterprise level. First, finding, extracting, and preprocessing relevant event data is still challenging. Second, most process mining approaches operate on a single-process level, making it hard to apply process mining being confronted with a multitude of processes. Third, process managers strongly require forward-directed operational support, but most process mining approaches provide only descriptive ex-post insights.

Addressing the first challenge of finding, extracting, and preprocessing relevant event data, Section III.1 proposes approaches for supporting process miners in extracting appropriate event logs and exploiting novel data sources that may contain valuable process and context information. Section III.1.1 focuses the extraction of event logs out of structured data. Research paper #1 presents the RDB2Log supporting a quality-informed, semi-automated data extraction out of relational databases. The contribution of RDB2Log is twofold. First, it helps non-technical users join associated events stored in multiple databases and tables. Second, by integrating data quality assessments in the early stage of data processing, the data quality of extracted event logs can be improved. Research paper #2 strives to integrate bot logs stemming from robotic process automation into process mining analysis. Therefore, research paper #2 proposes an approach to merge bot and process logs and proposes exemplary measures that benefit from a merged log. Section III.1.2 enters the world of unstructured data by proposing the Video Mining Reference Architecture, comprising a construction plan for developing solutions that use video data for process mining purposes. Thus, research paper #3 pushes process mining barriers to explore weakly digitized activities and further context information not contained in event logs based on structured process data.

Contributing to the second challenge, the single-process level of most process mining approaches, Section III.2.1 explores the log-based prioritization of interdependent processes. Research paper

[4] This Section is partly comprised of content taken from the research papers included in this thesis. To improve the readability of the text, I omit the standard labeling of these citations.

#4 proposes the D2P2, an approach that uses log data to mine structural and stochastic dependencies among processes and predicts processes' dependency-adjusted performance. These predictions are used to schedule an in-depth analysis when performance thresholds are violated, which justifies assigning the D2P2 to prescriptive process mining. Research paper #5 directly connects to that research developing the PMP2 that directly prescript process improvement projects (e.g., process mining projects) to degenerating processes.

Section III.2.2 more deeply explores the third challenge of providing operational, forward-directed support to process managers. Research paper #6 addresses this challenge by extensively comparing the performance of different ML and DL techniques for a diverse set of five public available logs. In a nutshell, the observations led to conclude that the application of DL is specifically promising when it comes to variant-rich processes producing a vast amount of data during runtime.

2 Future Research

As usual in research, this thesis' results are subject to specific assumptions leading to limitations that may be relaxed future research. While all individual research papers justify the assumption made and already address respective limitations (see Appendix VI.3-IV.8), this Section focuses on meta-findings across the six research papers that provide ideas for future research to further advance data-driven management of interconnected business processes.

Related to the first challenge of finding, merging, and cleaning event data, the thesis introduces approaches facilitating the extraction of appropriate event logs, exploiting novel data sources shedding light on existing blind spots. Future research should focus on further automate the process of data extraction. Constituting one of the largest bottlenecks in process mining, the barriers for non-experts to extract and prepare data for process mining analysis should be reduced. As one of the most important drivers for appropriate process mining results, future research should consider data quality aspects in the very early stages of process mining, namely the data identification and extraction. First, future research should develop interactive approaches to enabling users to more informed decisions. In the second step, future research could draw on innovative methods such as constructive ML (e.g., generative adversarial networks) to create approaches that automatically repair data quality issues, such as missing attributes or events. Using historical event traces as input, these methods can infer partial traces that might fill existing gaps. Also, unstructured data could be consulted as "second truth" to verify structured event data directly stemming from PAIS in critical application areas requiring a four-eye principle. As indicated by research paper #6, unstructured data can also be a source for exploring weakly digitized processes. Future research

should consider additional sources of structured and unstructured data. Lastly, as most event logs contain one single process, the extraction of multi-process logs, including dependencies among processes, should be further explored.

In light of the second challenge, most process mining approaches still operate on a single-process level, this thesis research points to how process mining can enhance traditional BPM methods, such as process prioritization and data-driven scheduling of improvement projects. Following this research avenue, event-log-driven insights could serve as a foundation for several BPM activities (e.g., process improvement, process redesign, or process implementation), calling for a structured end-to-end integration of process mining into the BPM lifecycle. To achieve the goal of supporting BPM on an enterprise level approaches capable of integrating novel technologies (e.g., software robots) and managing inter-organizational processes are required.

Regarding the third challenge, process managers strongly require forward-directed operational support, the thesis elaborates on how predictive monitoring approaches can benefit from applying DL techniques. However, future research must explore the intersection between process mining, on the one hand, as well as operations management and decision analysis on the other even more intensively. A promising direction also relies on connecting with other emerging process technologies, e.g., using predictive monitoring methods as smarter input for automated steering of software bots or smart devices. Further, predictive monitoring techniques that consider processes' context are required, as organizations' processes are not a closed system, and individual context factors may affect process behavior (van der Aalst 2020). This closes the circle to the first challenge of finding and merging various data sources containing relevant context information about the process instance under consideration. By extending the focus of data extraction beyond the typically sourced systems (e.g., ERP, CRM), these context factors can be illuminated. However, this requires applying novel concepts (e.g., CV) to extract and abstract structured features out of vast amounts of data, as exemplified in research paper #3.

In sum, the thesis contributes to the existing body of knowledge on data-driven management of interconnected business processes. I hope this thesis provides a basis for applying process mining in a forward-looking view and, thus, supports researchers and practitioners on the journey of converting project-based and isolated process mining initiatives to an ongoing supplement to the core of traditional BPM methods.

V. References

Aggarwal JK, Ryoo MS (2011) Human Activity Analysis: A Review. ACM Computing Surveys 43(3):1–43. doi:10.1145/1922649.1922653.

Alpaydin E (2020) Introduction to machine learning. MIT Press, Cambridge, Massachusetts.

Andrews R, van Dun C, Wynn MT, Kratsch W, Röglinger M, ter Hofstede A (2020) Quality-informed semi-automated event log generation for process mining. Decision Support Systems 132:113265. doi:10.1016/j.dss.2020.113265.

Bandara W, Guillemain A, Coogans P (2015) Prioritizing Process Improvement: An Example from the Australian Financial Services Sector. In: vom Brocke J, Rosemann M (eds) Handbook on Business Process Management 2. Springer, Berlin, Heidelberg, pp 289–307.

Beverungen D, et al. (2020) Seven Paradoxes of Business Process Management in a Hyper-Connected World. Business & Information Systems Engineering. doi:10.1007/s12599-020-00646-z.

Bitomsky L, Huhn J, Kratsch W, Röglinger M (2019) Process Meets Project Prioritization – A Decision Model for Developing Process Improve-ment Roadmaps. In: ECIS 2019 Proceedings.

Breiman L (2001) Random Forests. Machine Learning 45(1):5–32. doi:10.1023/A:1010933404324.

Browne R (2019) How three friends turned a college project into a $2.5 billion software unicorn. CNBC. https://www.cnbc.com/2019/11/21/celonis-raises-290m-series-c-funding-round-at-2point5b-valuation.html. Accessed 2020-11-02.

Chollet F (2018) Deep learning with Python. Manning, Shelter Island, NY.

Cortes C, Vapnik V (1995) Support-Vector Networks. Machine Learning 20(3):273–297. doi:10.1023/A:1022627411411.

Covington P, Adams J, Sargin E (2016) Deep Neural Networks for YouTube Recommendations. In: ACM RECSYS 2016 Proceedings.

Davis D (2019) AI Unleashes the Power of Unstructured Data. https://www.cio.com/article/3406806/ai-unleashes-the-power-of-unstructured-data.html. Accessed 2020-06-14.

Deng L (2018) Artificial Intelligence in the Rising Wave of Deep Learning: The historical path and future outlook. IEEE Signal Processing Magazine 35(1):180-177. doi:10.1109/MSP.2017.2762725.

Deng L, Yu D (2014) Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing 7(3-4):197–387. doi:10.1561/2000000039.

Di Francescomarino C, Ghidini C, Maggi FM, Milani F (2018) Predictive Process Monitoring Methods: Which One Suits Me Best? In: BPM 2018 Proceedings, pp 462–479.

Diba K, Batoulis K, Weidlich M, Weske M (2020) Extraction, correlation, and abstraction of event data for process mining. WIREs Data Mining and Knowledge Discovery 10(3). doi:10.1002/widm.1346.

Dijkman R, Vanderfeesten I, Reijers HA (2016) Business process architectures: Overview, comparison and framework. Enterprise Information Systems 10(2):129–158. doi:10.1080/17517575.2014.928951.

Dumas M, La Rosa M, Mendling J, Reijers HA (2018) Fundamentals of Business Process Management. Springer, Berlin, Heidelberg.

Egger A, ter Hofstede AHM, Kratsch W, Leemans SJJ, Röglinger M, Wynn MT (2020) Bot Log Mining: Using Logs from Robotic Process Automation for Process Mining. In: ER 2020 Proceedings, pp 51–61.

Evermann J, Rehse J-R, Fettke P (2016) A Deep Learning Approach for Predicting Process Behaviour at Runtime. In: PARISE 2016 Proceedings.

Freedman D (2009) Statistical models: Theory and practice. Cambridge Univ. Press, Cambridge.

Gao J, Yang Y, Lin P, Park DS (2018) Computer Vision in Healthcare Applications. Journal of healthcare engineering 2018:5157020. doi:10.1155/2018/5157020.

Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Martinez-Gonzalez P, Garcia-Rodriguez J (2018) A survey on deep learning techniques for image and video semantic segmentation. Applied Soft Computing 70:41–65.

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge, Massachusetts, London, England.

Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS (2016) Deep learning for visual understanding: A review. Neurocomputing 187:27–48. doi:10.1016/j.neucom.2015.09.116.

Hanafizadeh P, Osouli E (2011) Process selection in re-engineering by measuring degree of change. Business Process Management Journal 17(2):284–310. doi:10.1108/14637151111122356.

Harmon P (2020) The state of business process management 2020. http://www.bptrends.com/bpt/wp-content/uploads/2020-BPM-Survey.pdf. Accessed 2020-11-09.

Haykin SS (2009) Neural networks and learning machines: A comprehensive foundation. Pearson, New York.

Herath S, Harandi M, Porikli F (2017) Going deeper into action recognition: A survey. Image and Vision Computing 60:4–21. doi:10.1016/j.imavis.2017.01.010.

Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression. Wiley, Hoboken N.J.

Huval B, Wang T, Tandon S, Kiske J, Song W, Pazhayampallil J, Andriluka M, Rajpurkar P, Migimatsu T, Cheng-Yue R, Mujica F, Coates A, Ng AY (2015) An Empirical Evaluation of Deep Learning on Highway Driving.

Kerpedzhiev GD, König UM, Röglinger M, Rosemann M (2020) An Exploration into Future Business Process Management Capabilities in View of Digitalization. Business & Information Systems Engineering. doi:10.1007/s12599-020-00637-0.

Kang B, Kim D, Kang S-H (2012) Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction. Expert Systems with Applications 39(5):6061–6068.

Kong Y, Fu Y (2018) Human Action Recognition and Prediction: A Survey. ArXiv, abs/1806.11230. https://arxiv.org/pdf/1806.11230 (Preprint).

Kratsch W, König F, Röglinger M (2020a) Shedding Light on Blind Spots: Developing a Reference Architecture to Leverage Video Data for Process Mining. https://arxiv.org/pdf/2010.11289 (Preprint).

Kratsch W, Manderscheid J, Reißner D, Röglinger M (2017) Data-driven Process Prioritization in Process Networks. Decision Support Systems 100:27–40. doi:10.1016/j.dss.2017.02.011.

Kratsch W, Manderscheid J, Röglinger M, Seyfried J (2020b) Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction. Business & Information Systems Engineering. doi:10.1007/s12599-020-00645-0.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. doi:10.1038/nature14539.

Lee AS, Baskerville RL (2003) Generalizing Generalizability in Information Systems Research. Information Systems Research 14(3):221–243. doi:10.1287/isre.14.3.221.16560.

Lee K, Ognibene D, Chang HJ, Kim T-K, Demiris Y (2015) STARE: Spatio-Temporal Attention Relocation for Multiple Structured Activities Detection. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 24(12):5916–5927. doi:10.1109/TIP.2015.2487837.

Lehnert M, Röglinger M, Seyfried J (2018) Prioritization of Interconnected Processes. Business & Information Systems Engineering 60(2):95–114. doi:10.1007/s12599-017-0490-4.

Letmathe P, Petersen L, Schweitzer M (2013) Capacity management under uncertainty with inter-process, intra-process and demand interdependencies in high-flexibility environments. OR Spectrum 35(1):191–219. doi:10.1007/s00291-012-0285-4.

Leyer M, Heckl D, Moormann J (2015) Process Performance Measurement. In: vom Brocke J, Rosemann M (eds) Handbook on Business Process Management 2. Springer, Berlin, Heidelberg, pp 227–241.

Li J, Wang HJ, Bai X (2015) An intelligent approach to data extraction and task identification for process mining. Information Systems Frontiers 17(6):1195–1208. doi:10.1007/s10796-015-9564-3.

Marquez-Chamorro AE, Resinas M, Ruiz-Cortes A (2018) Predictive Monitoring of Business Processes: A Survey. IEEE Transactions on Services Computing 11(6):962–977. doi:10.1109/TSC.2017.2772256.

Marr B (2019) What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone. https://www.forbes.com/sites/bernardmarr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/#266bcb1b15f6. Accessed 2020-06-14.

Masi I, Wu Y, Hassner T, Natarajan P (2018) Deep Face Recognition: A Survey. In: 31st Conference on Graphics, Patterns and Images. SIBGRAPI 2018 : proceedings : 29 October-1 November 2018, Foz do Iguaçu, Brazil. Conference Publishing Services, IEEE Computer Society, Los Alamitos, CA, pp 471–478.

Microsoft Research (2019) Visual Computing - Microsoft Research. https://www.microsoft.com/en-us/research/group/visual-computing/. Accessed 2020-08-05.

Prince SJD (2012) Computer Vision: Models, Learning, and Inference. Cambridge University Press, New York, NY.

Recker J, Mendling J (2016) The State of the Art of Business Process Management Research as Published in the BPM Conference. Business & Information Systems Engineering 58(1):55–72. doi:10.1007/s12599-015-0411-3.

Reining C, Niemann F, Moya Rueda F, Fink GA, Hompel M ten (2019) Human Activity Recognition for Production and Logistics—A Systematic Literature Review. Information 10(8):245. doi:10.3390/info10080245.

Research and Markets (2020) Process Analytics Market by Process Mining Type (Process Discovery, Process Conformance & Process Enhancement), Deployment Type, Organization Size, Application (Business Process, It Process, & Customer Interaction) & Region - Global Forecast to 2023. https://www.researchandmarkets.com/reports/4576970/process-analytics-market-by-process-mining-type. Accessed 2020-06-29.

Rosemann M, vom Brocke J (2015) The Six Core Elements of Business Process Management. In: vom Brocke J, Rosemann M (eds) Introduction, methods, and information systems. Springer, Berlin, pp 105–122.

Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Networks 61:85–117. doi:10.1016/j.neunet.2014.09.003.

Schönig S, Jasinski R, Ackermann L, Jablonski S (2018) Deep Learning Process Prediction with Discrete and Continuous Data Features. In: ENASE 2018 Proceedings, pp 314–319.

Shin H-C, et al. (2016) Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Transactions on Medical Imaging 35(5):1285–1298. doi:10.1109/TMI.2016.2528162.

Shrestha A, Cater-Steel A, Toleman M, Tan W-G (2015) A Method to Select IT Service Management Processes for Improvement. Journal of Information Technology Theory and Application 15(3). http://aisel.aisnet.org/jitta/vol15/iss3/3.

Sun K, Xiao B, Liu D, Wang J (2019) Deep High-Resolution Representation Learning for Human Pose Estimation. In: CVPR 2019 Proceedings, pp 5686–5696.

Szeliski R (2011) Computer Vision: Algorithms and Applications. Springer, London.

Tesla (2020) Autopilot. https://www.tesla.com/autopilot?redirect=no. Accessed 2020-09-17.

Track160 (2020). https://track160.com/. Accessed 2020-09-17.

van der Aa H, Carmona J, Leopold H, Mendling J, Padró L (2018) Challenges and Opportunities of Applying Natural Language Processing in Business Process Management. In: COLING 2018 Proceedings, pp 2791–2801.

van der Aalst W, et al. (2011a) Process Mining Manifesto. In: BPM 2011 Workshops Proceedings, pp 169–194.

van der Aalst W (ed) (2016) Process Mining: Data Science in Action. Springer Berlin Heidelberg, Berlin, Heidelberg.

van der Aalst W (2020) Academic View: Development of the Process Mining Discipline. In: Reinkemeyer L (ed) Process mining in action. Principles, use cases and outlook. Springer International Publishing, Cham, pp 181–196.

van der Aalst WMP (2013) Business Process Management: A Comprehensive Survey. ISRN Software Engineering 2013(1):1–37. doi:10.1155/2013/507984.

van der Aalst WMP (2014) Data Scientist: The Engineer of the Future. In: Mertins K, Bénaben F, Poler R, Bourrières J-P (eds) Enterprise Interoperability VI. Interoperability for Agility, Resilience and Plasticity of Collaborations. Springer International Publishing, Cham, s.l., pp 13–26.

van der Aalst WMP, Schonenberg MH, Song M (2011b) Time prediction based on process mining. Information Systems 36(2):450–475. doi:10.1016/j.is.2010.09.001.

Vanwersch RJB, et al. (2016) A Critical Evaluation and Framework of Business Process Improvement Methods. Business & Information Systems Engineering 58(1):43–53. doi:10.1007/s12599-015-0417-x.

Viner D, Stierle M, Matzner M (2020) A Process Mining Software Comparison. In: ICPM 2020 Proceedings.

vom Brocke J, Jans M, Mendling J, Reijers HA (2020) Process Mining at the Enterprise Level: Call for Papers, Issue 5/2021. Business & Information Systems Engineering 62(2):185–187. doi:10.1007/s12599-020-00630-7.

vom Brocke J, Zelt S, Schmiedel T (2016) On the role of context in business process management. International Journal of Information Management 36(3):486–495. doi:10.1016/j.ijinfomgt.2015.10.002.

Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep Learning for Computer Vision: A Brief Review. Computational intelligence and neuroscience 2018. doi:10.1155/2018/7068349.

Wen L, Wang J, Sun J (2006) Detecting Implicit Dependencies Between Tasks from Event Logs. In: Asia-Pacific Web Conference 2006 Proceedings, pp 591–603.

Weyand T, Kostrikov I, Philbin J (2016) PlaNet - Photo Geolocation with Convolutional Neural Networks. In: ECCV 2016 Proceedings, pp 37–55.

Witten IH, Frank E, Hall MA, Pal CJ (2017) Data mining: Practical machine learning tools and techniques. Morgan Kaufmann/Elsevier, Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo.

Yim J, Joo D, Bae J, Kim J (2017) A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In: CVPR 2017 Proceedings, pp 7130–7138.

Yin RK (1994). Case study research: design and methods (2nd ed). Sage, Thousand Oaks, Calif.

VI. Appendix

1 Index of Research Papers

Research paper #1: Quality-Informed Semi-Automated Event Log Generation for Process Mining

Andrews R, van Dun C, Wynn M, Kratsch W, Röglinger M, ter Hofstede A (2020) Quality-Informed Semi-Automated Event Log Generation for Process Mining. In: *Decision Support Systems, 132, 113265.*

Research paper #2: Bot Log Mining: Using Logs from Robotic Process Automation for Process Mining

Egger A, ter Hofstede A, Kratsch W, Leemans S, Röglinger M, Wynn M (2020) Bot Log Mining: Using Logs from Robotic Process Automation for Process Mining. In: *Proceedings of the 39th International Conference on Conceptual Modelling (ER Conference 2020), Vienna, Austria (Short Paper).*

Research paper #3: Shedding Light on Blind Spots – Developing a Reference Architecture to systematically use Video Data for Process Mining

Kratsch W, König F, Röglinger M (2020) Shed Light on Blind Spots – Developing a Reference Architecture to systematically use Video Data for Process Mining. *Submitted Working Paper*

Research paper #4: Data-driven Process Prioritization in Process Networks

Kratsch W, Manderscheid J, Reißner D, Röglinger M (2017) Data-driven process prioritization in process networks. In: *Decision Support Systems, 100, 27-40.*

Research paper #5: Process Meets Project Prioritization – A Decision Model for Developing Process Improvement Roadmaps

Bitomsky L, Huhn J, Kratsch W, Röglinger M (2019) Process Meets Project Prioritization – A Decision Model for Developing Process Improvement Roadmaps. In: *Proceedings of the 27th European Conference on Information Systems (ECIS 2019), Stockholm & Uppsala, Sweden.*

Research paper #6: Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction

Kratsch W, Manderscheid J, Röglinger M, Seyfried J (2020) Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction. In: *Business & Information Systems Engineering.*

2 Individual Contribution to the Included Research Papers

This cumulative thesis comprises six research papers building the main body of this work. All included research papers were written in teams with multiple co-authors. Thus, this Section is to detail respective project settings and my individual contribution to each research paper.

Research paper #1 (Andrews et al. 2020) was written with five co-authors – three of whom work at another international research institution. The team jointly conceptualized and elaborated on the article's content. Together, we developed an approach supporting semi-automated, quality informed event log extraction from relational databases (RDB2Log). Supporting the real-world evaluation, I was primarily responsible for designing a questionnaire assessing the selected evaluation criteria and implement it as an online survey. Furthermore, together with one of the co-authors, I applied RDB2Log to real data from a German electronics manufacturing services company and conducted an evaluation workshop with internal process experts. Throughout, I was substantially involved in all parts of the project.

Research paper #2 (Egger et al. 2020) was written with five co-authors – three of whom work at another international research institution. All co-authors jointly developed an approach to use logs stemming from robotic process automation for process mining. I was involved in conceptualizing, developing, and reworking text sections throughout the article. Overall, the authors made equal contributions to the paper's content, and I was substantially involved in each part of the project.

Research paper #3 (Kratsch et al. 2020a) was written with two co-authors. As the leading author, I had a main role in ideating the research topic using video data for process mining purposes. Furthermore, I mainly developed the Video Mining Reference Architecture as the primary artifact of the research paper, whereas one of my co-authors instantiated the Video Mining Reference Architecture as a software prototype. Moreover, I was primarily responsible for integrating the research paper in existing process mining research. Additionally, I was in the lead in writing and finalizing the paper to get it ready for submission. Although the research paper represents, to a large extent, my work, the two co-authors were involved in all parts of the project and helped to advance our contribution.

Research paper #4 (Kratsch et al. 2017) was developed together with three co-authors. Based on an initial idea provided by one of the co-authors, the team jointly conceptualized and elaborated on the paper's content. Together, we developed an approach to prioritizing business processes for in-depth analysis based on their simulated future performance. Personally, I had the key role in conceptualizing and implementing the mixed-integer linear program, optimizing the performance of an interdependent

process network. Whereas one of the co-authors mainly developed the statistical forecasting model, I was again in the lead to merge the forecasting and the optimization models into one software prototype. I also took the main responsibility for revising the paper to get it finally accepted.

Research paper #5 (Bitomsky et al. 2019) was developed with a team of three co-authors. Based on my idea to extend the process prioritization approach of research paper #4 to the project level, the team jointly conceptualized and elaborated the paper's content. Together, we developed an approach that schedules process improvement projects based on predicted future process performance. I was involved in conceptualizing, developing, and reworking text sections throughout the article. Overall, I was involved in each part of the project.

Research paper #6 (Kratsch et al. 2020b) was written with three co-authors. All co-authors jointly performed a structured comparison of traditional ML and DL approaches for predictive process monitoring. I mainly conceptualized the study design and implemented the machine and DL models. Furthermore, striving for conceptual completeness, I developed a framework to select datasets and ML algorithms purposefully. I also supported the data preprocessing, which was mainly undertaken by one of the other co-authors. Besides, I took responsibility for revising the paper for resubmission. In sum, I had a central role in each part of the project.

3 Research Paper #1: Quality-Informed Semi-Automated Event Log Generation for Process Mining

Authors: Andrews R, van Dun C, Wynn M, Kratsch W, Röglinger M, ter Hofstede A

Published in: Decision Support Systems, 2020, *132, 113265*

Abstract: Process mining, as any form of data analysis, relies heavily on the quality of input data to generate accurate and reliable results. A fit-for-purpose event log nearly always requires time-consuming, manual preprocessing to extract events from source data, with data quality de-pendent on the analyst's domain knowledge and skills. Despite much being written about data quality in general, a generalizable framework for analyzing event data quality issues when extracting logs for process mining remains unrealized. Following the DSR paradigm, we present RDB2Log, a quality-aware, semi-automated approach for extracting event logs from relational data. We validated RDB2Log's design against design objectives extracted from literature and competing artifacts, evaluated its design and performance with process mining experts, implemented a prototype with a defined set of quality metrics, and applied it in laboratory settings and in a real-world case study. The evaluation shows that RDB2Log is understandable, of relevance in current research, and supports process mining in practice.

Keywords: Process Mining, Data Quality, Event Log, Log Extraction

4 Research Paper #2: Bot Log Mining: Using Logs from Robotic Process Automation for Process Mining

Authors: Egger A, ter Hofstede A, Kratsch W, Leemans S, Röglinger M, Wynn M

Published in: ER Conference Proceedings, 2020

Abstract: Robotic Process Automation (RPA) is an emerging technology for automating tasks using bots that can mimic human actions on computer systems. Most existing research focuses on the earlier phases of RPA implementations, e.g. the discovery of tasks that are suitable for automation. To detect exceptions and explore opportunities for bot and process redesign, historical data from RPA-enabled processes in the form of bot logs or process logs can be utilized. However, the isolated use of bot logs or process logs provides only limited insights and not a good understanding of an overall process. Therefore, we develop an approach that merges bot logs with process logs for process mining. A merged log enables an integrated view on the role and effects of bots in an RPA-enabled process. We first develop an integrated data model describing the structure and relation of bots and business processes. We then specify and instantiate a 'bot log parser' translating bot logs of three leading RPA vendors into the XES format. Further, we develop the 'log merger' functionality that merges bot logs with logs of the underlying business processes. We further introduce process mining measures allowing the analysis of a merged log. We evaluate the proposed approach on real-world and artificial bot and process logs.

Keywords: Robotic Process Automation, Process Mining, Business Process Management

5 Research Paper #3: Shedding Light on Blind Spots: Developing a Reference Architecture to Leverage Video Data for Process Mining

Authors: Kratsch W, König F, Röglinger M

Submitted Working Paper

Extended Abstract

Big data analytics is one of the most promising technology enablers for business process management (BPM) [1]. As an exemplary domain-specific big data technology, process mining strives to discover, monitor, and improve processes by extracting knowledge from event logs commonly available in information systems [2]. In recent years, process mining has evolved into one of the most active and fast-growing research streams in BPM. The first international conference on process mining (ICPM), which took place in Aachen in 2019, underlined the scientific relevance of the subject [3]. In practice, Celonis' super-fast expansion from start-up to unicorn in only seven years indicates the enormous cross-industry business potential of process mining. By 2023, *Markets and Markets* predicts a market potential of 1.42 billion US$ for process mining technologies [4]. Current process mining applications are highly reliant on structured business data, often gathered from process-aware information systems (PAIS) or other enterprise information systems (e.g., ERP, CRM). However, according to *Forbes* and *CIO* magazine, 80 to 90 percent of available data is unstructured, i.e., data without a functional, retrievable data scheme [5,6]. Moreover, unstructured data is increasing much more rapidly than structured data [6]. Consequently, most process mining analyses only exploit 10 to 20 percent of the available data. Video data (e.g., from surveillance cameras) has the potential to make processes that partly run away from information systems (i.e., blind spots) more observable. Thus, our research question is as follows: How can video data be systematically exploited to support process mining?

Here, we propose the Video Mining Reference Architecture (RA) supporting the extraction of structured information from unstructured video data, as well as the transformation of structured information into a format suitable for process mining use cases. As the central research artifact, the Video Mining RA facilitates the use-case-driven implementation and integration of computer vision capabilities into process mining architectures. By instantiating the Video Mining RA for exemplary process mining use cases, we also provide operational support for the practical implementation of such an architecture and demonstrate which computer vision capabilities are suitable

for which process mining contexts. Our results also show that an exemplary software prototype instantiation of the proposed reference architecture is capable of automatically extracting most of the process-relevant events from unstructured video data.

References

[1] D. Beverungen, J.C.A.M. Buijs, J. Becker, C. Di Ciccio, W.M.P. van der Aalst, C. Bartelheimer, J. vom Brocke, M. Comuzzi, K. Kraume, H. Leopold, M. Matzner, J. Mendling, N. Ogonek, T. Post, M. Resinas, K. Revoredo, A. del-Río-Ortega, M. La Rosa, F.M. Santoro, A. Solti, M. Song, A. Stein, M. Stierle, V. Wolf, Seven Paradoxes of Business Process Management in a Hyper-Connected World, Bus Inf Syst Eng (2020). https://doi.org/10.1007/s12599-020-00646-z.

[2] W. van der Aalst et al., Process Mining Manifesto, in: F. Daniel, K. Barkaoui, S. Dustdar (Eds.), Business Process Management Workshops, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 169–194.

[3] 2019 International Conference on Process Mining: ICPM 2019 proceedings Aachen, Germany, 24-26 June 2019, IEEE Computer Society, Conference Publishing Services, Los Alamitos, California, Washington, Tokyo, 2019.

[4] R.a.M. ltd, Process Analytics Market by Process Mining Type (Process Discovery, Process Conformance & Process Enhancement), Deployment Type, Organization Size, Application (Business Process, It Process, & Customer Interaction) & Region - Global Forecast to 2023, 2020. https://www.researchandmarkets.com/reports/4576970/process-analytics-market-by-process-mining-type (accessed 29 June 2020).

[5] D. Davis, AI Unleashes the Power of Unstructured Data, 2019. https://www.cio.com/article/3406806/ai-unleashes-the-power-of-unstructured-data.html (accessed 14 June 2020).

[6] B. Marr, What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone, 2019. https://www.forbes.com/sites/bernard-marr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/#266bcb1b15f6 (accessed 14 June 2020).

6 Research Paper #4: Data-Driven Process Prioritization In Process Networks

Authors: Kratsch W, Manderscheid J, Reißner D, Röglinger M

Published in: Decision Support Systems, 2017, *100, 27-40.*

Abstract: Business process management (BPM) is an essential paradigm of organizational design and a source of corporate performance. The most value-creating activity of BPM is process improvement. With effective process prioritization being a critical success factor for process improvement, we propose the Data-Driven Process Prioritization (D2P2) approach. By ad-dressing the weaknesses of extant process prioritization approaches, the D2P2 accounts for structural and stochastic process dependencies and leverages log data. The D2P2 returns a priority list that indicates in which future periods the processes from a process network should undergo the next in-depth analysis to check whether they actually require improvement. The D2P2 contributes to the prescriptive knowledge on process prioritization and process decision-making. As for evaluation, we discussed the D2P2's design specification against theory-backed design objectives and competing artefacts. We also instantiated the D2P2 as a software prototype and applied the prototype to a real-world scenario based on the 2012 BPI Challenge log.

Keywords: Business Process Management, Process Prioritization, Process Improvement, Business Process Architecture, Process Logs

7 Research Paper #5: Process Meets Project Prioritization – A Decision Model for Developing Process Improvement Roadmaps

Authors: Bitomsky L, Huhn J, Kratsch W, Röglinger M

Published in: ECIS 2019 Proceedings, 2019

Abstract: Improving business processes is a key success factor for organizations and, at the same time, a major challenge for decision makers. For process improvement to be successful, effective prioritization is essential. Despite the existence of approaches for the prioritization of process improvement projects or business processes, prescriptive research at the intersection of both re-search streams is missing. Existing approaches do not simultaneously prioritize business processes and improvement projects. Hence, scarce corporate funds may be misallocated. To address this research gap, we propose the PMP2, an economic decision model that assists organizations in the identification of business process improvement (BPI) roadmaps. Based on stochastic processes and simulation, the decision model maps different improvement projects to individual business processes within a process network. Thereby, it caters for process dependencies and basic interactions among projects. Drawing from the principles of value-based management, the decision model determines the process improvement roadmap with the highest contribution to the long-term firm value. To evaluate the PMP2, we instantiated it as a software proto-type and performed different scenario analyses based on synthetic data. The results highlight the importance of prioritizing business processes and improvement projects in an integrated manner.

Keywords: Business Process Management, Business Process Improvement, Process Prioritization, Process Dependencies, Network Analysis

8 Research Paper #6: Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction

Authors: Kratsch W, Manderscheid J, Röglinger M, Seyfried J

Published in: Business Information & Systems Engineering, 2020

Abstract: Predictive process monitoring aims at forecasting the behavior, performance, and outcomes of business processes at runtime. It helps identify problems before they occur and re-allocate re-sources before they are wasted. Although deep learning (DL) has yielded breakthroughs, most existing approaches build on classical machine learning (ML) techniques, particularly when it comes to outcome-oriented predictive process monitoring. This circumstance reflects a lack of understanding about which event log properties facilitate the use of DL techniques. To address this gap, the authors compared the performance of DL (i.e., simple feedforward Deep Neural Networks and Long Short Term Memory Networks) and ML techniques (i.e., Random Forests and Support Vector Machines) based on five publicly available event logs. It could be observed that DL generally outperforms classical ML techniques. Moreover, three specific propositions could be inferred from further observations: First, the outperformance of DL techniques is particularly strong for logs with a high variant-to-instance ratio (i.e., many non-standard cases). Second, DL techniques perform more stably in case of imbalanced target variables, especially for logs with a high event-to-activity ratio (i.e., many loops in the control flow). Third, logs with a high activity-to-instance payload ratio (i.e., input data is predominantly generated at runtime) call for the application of Long Short Term Memory Networks. Due to the purposive sampling of event logs and techniques, these findings also hold for logs outside this study.

Keywords: Predictive Process Monitoring, Business Process Management, Outcome Prediction, Deep Learning, Machine Learning