Communications of the Association for Information Systems

Volume 51 Article 12

2022

Generating Research Questions from Digital Trace Data: A Machine-Learning Method for Discovering Patterns in a Dynamic Environment

Henrik Kallio
Aalto University School of Economics, henrik.kallio@aalto.fi

Pekka Malo
Aalto University School of Business

Timo Lainema
University of Turku

Johanna Bragge
Aalto University School of Business

Tomi Seppälä
University of Eastern Finland, Business School

See next page for additional authors

Recommended Citation

Kallio, H., Malo, P., Lainema, T., Bragge, J., Seppälä, T., & Penttinen, E. (2022). Generating Research Questions from Digital Trace Data: A Machine-Learning Method for Discovering Patterns in a Dynamic Environment. Communications of the Association for Information Systems, 51, pp-pp. Retrieved from https://aisel.aisnet.org/cais/vol51/iss1/12

Generating Research Questions from Digital Trace Data: A Machine-Learning Method for Discovering Patterns in a Dynamic Environment

Authors

Henrik Kallio, Pekka Malo, Timo Lainema, Johanna Bragge, Tomi Seppälä, and Esko Penttinen

Communications of the Association for Information Systems

Generating Research Questions from Digital Trace Data: A Machine-Learning Method for Discovering Patterns in a Dynamic Environment

Henrik Kallio
Department of Information and Service Management
Aalto University School of Business
henrik.kallio@aalto.fi

Pekka Malo
Department of Information and Service Management
Aalto University School of Business

Timo Lainema
Center for Collaborative Research, Turku School of Economics, University of Turku

Johanna Bragge
Department of Information and Service Management,
Aalto University School of Business

Tomi Seppälä
Department of Information and Service Management,
Aalto University School of Business; University of Eastern Finland, Business School

Esko Penttinen
Department of Information and Service Management,
Aalto University School of Business

Please cite this article as: Kallio, Henrik; Malo, Pekka; Lainema, Timo; Bragge, Johanna; Seppälä, Tomi; Penttinen, Esko: Generating Research Questions from Digital Trace Data: A Machine-Learning Method for Discovering Patterns in a Dynamic Environment, *Communications of the Association for Information Systems* (forthcoming), In Press*.*

Communications of the Association for Information Systems

Research Paper ISSN: 1529-3181

Generating Research Questions from Digital Trace Data: A Machine-Learning Method for Discovering Patterns in a Dynamic Environment

Henrik Kallio
Department of Information and Service Management
Aalto University School of Business
henrik.kallio@aalto.fi

Pekka Malo
Department of Information and Service Management
Aalto University School of Business

Timo Lainema
Center for Collaborative Research, Turku School of Economics, University of Turku

Johanna Bragge
Department of Information and Service Management, Aalto University School of Business

Tomi Seppälä
Department of Information and Service Management, Aalto University School of Business; University of Eastern Finland, Business School

Esko Penttinen
Department of Information and Service Management, Aalto University School of Business

Abstract:

Digital trace data derived from organizations' information systems represent a wealth of possibilities in analyzing decision-making processes and organizational performance. While data-mining methods have advanced considerably over recent years, organizational process research has rarely *analyzed this type of trace data with the objective of better understanding organizations' decision-making processes. However,* accurately tracking decision-making actions via digital trace data can produce numerous applications that represent new and unexplored opportunities for IS research.

The paper presents a novel method developed to combine quantitative process mining approaches with a variance perspective. Its viability is demonstrated by looking at teams' decision patterns from a dynamic business-simulation game. This exploratory data-driven method represents a promising starting point for translating complex raw process data into interesting research questions connected with dynamic decision-making environments.

Keywords: Digital Trace Data, Process Data, Process Mining, Data-Driven, Problematization, Machine Learning, Business-Simulation Game, Dynamic Decision-Making.

[Department statements, if appropriate, will be added by the editors. Teaching cases and panel reports will have a statement, which is also added by the editors.]

[Note: this page has no footnotes.]

This manuscript underwent [editorial/peer] review. It was received xx/xx/20xx and was with the authors for XX months for XX revisions. [firstname lastname] served as Associate Editor.] **or** The Associate Editor chose to remain anonymous.]

1 Introduction

Modern information systems often record well-structured data, the creation of which takes place in a continuous, time-dependent manner. Digital technologies produce an unprecedented volume of digital traces as byproducts to fulfilling their primary purpose. A trace can be seen as recorded evidence that something has occurred. Howison, Wiggins, and Crowston (2011) define **digital trace data** as "records of activity (trace data) undertaken through an online information system" (p. 769). These digital traces can be connected to form comprehensive views of both individuals' and group behavior, with the "potential to transform our understanding of individuals, organizations, and societies" (Lazer et al., 2009).

Trace data have several noteworthy characteristics. Firstly, they are a byproduct of activities rather than a product of a research instrument designed for this end. Also, trace data are event-based data (not summarized data) and are produced through, and stored by, an information system. In addition, because events occur over a span of time, trace data are of a longitudinal nature (Howison, Wiggins, & Crowston, 2011). Further, digital trace data typically consist of timestamped logged sequential events (Pentland, Recker, Wolf, & Wyner, 2020). Because of the timestamps, intervals between trace-data events can be captured as well. Hence, it becomes possible to analyze digital trace data to interrelate events chronologically and over time (Yoo, 2010), or to reconstruct processes composed of events. A typical process thus examined might be the order-delivery process of a business organization, comprising decisions and actions (events), triggered by the system's users and the system itself.

As research has shown, inventive use of digital trace data can lead to innovations not anticipated by those who developed the information system in the first place (Yoo, Boland, Lyytinen, & Majchrzak, 2012). One stream of research in this vein is the recently established field of computational social science. Its practitioners collect and analyze digital traces of human behavior in a discreet way, with high precision and granularity (Stier, Breuer, Siegers, & Thorson, 2020). Much discussion has ensued, especially around how digital trace data can afford the analysis of online social networks (Agarwal, Gupta, & Kraut, 2008; Karanasios, Thakker, Lau, Allen, Dimitrova, & Norman, 2013). Examples include studying social-media data for assessing public opinion (Driscoll & Walker, 2014; Jungherr, Schoen, Posegga, & Jürgens, 2017), predicting voting behavior (Bach et al., 2021), researching international youth protests related to climate change (Boulianne, Lalancette, & Ilkiw, 2020), and tracking population movements over time (Fiorio, Zagheni, Abel, Hill, Pestre, Letouzé, & Cai, 2021).

The social-media context is not the only setting in which digital trace data are collected. Enterprise information systems record digital trace data in vast quantities. Organizations' information systems track both automatically generated events—for instance, an enterprise resource planning (ERP) system may trigger purchase orders for components when the inventory level falls below a set reorder point—and human-generated actions, as is the case when a sales-department representative checks a customer order and passes it on for manufacturing. The timestamps allow sorting the events and identifying how discrete events may have triggered other events. Still, from surveying recent publications, Pentland, Vaast, and Wolf (2021) note that organizational process research does not display evidence of any studies of process mining, even though, in their view, process-mining-based techniques for analysis and interpretation of digital trace data can help bridge the research fields of organizational process theory and process mining.

The events giving rise to digital trace data typically occur as part of a larger sequence that constitutes a process, including routines, projects, workflows, and/or business processes (Bose & van der Aalst, 2009; Pentland et al., 2020). In an organizational context, digital trace data have only rarely been applied in process analysis research. One example comes from Pentland et al. (2021), who studied the process dynamics of health-care routines at four dermatology clinics. They found that all four cases exhibited process changes that the clinical staff could not explain—the personnel were unaware that any changes in the processes had even occurred.

An essential step in any research is the act of constructing and articulating a carefully formulated research question. Some traditional ways of taking this step are via gap-spotting from earlier research and problematization that entails questioning prior theories' underlying assumptions (Sandberg & Alvesson, 2011). Most of these techniques rely on scrutinizing the literature and theory, then drawing on them to develop research questions. While this is likely to remain the main route for generating research questions that hold potential to stimulate significant theoretical advances, we argue that recent technological advancements in domains such as social media and enterprise systems—coupled with abundant digital trace data—offer new and interesting avenues whereby researchers can develop research questions from data.

To unleash the power of trace data, IS researchers can benefit from inductively generating novel theory from trace data of all forms (Berente, Seidel, & Safadi, 2019). Barton and Court (2012), who highlight the importance of advanced analytics for modern enterprise information systems, argue that advanced analytics operations are likely to become a core element of companies' performance-improvement efforts. Developing analytics tools focused on such improvement should be integral to their work. For example, decision-support systems based on algorithmic learning and artificial intelligence could aid organizations.

Trace data and their applications tie in also with real-time information processing and decision-making in dynamic contexts. The notion of dynamic decision-making (DDM) refers to the need to make several interdependent decisions in an environment that changes both endogenously (via the decision-maker's actions) and exogenously (through actions the decision-maker cannot control) (Edwards, 1962). Nowadays, decision environments of this sort typify several important domains, including business (Gonzalez, 2005), where related capabilities are important. For instance, pressure to succeed with decision-making in increasingly short timeframes renders organizations more and more dependent on real-time information (Machado, Cunha, Pereira, & Oliveira, 2019). Hence, companies are moving toward real-time analytics based on trace data, as opposed to conventional usage of business-intelligence systems that operate with aggregated data (Hahn & Packowski, 2015). Tasks that demand decisions in real-time dynamic environments are only increasing in number; data must be processed and decisions made as soon as the data are available, or not long after (Tendick, Denby, & Ju, 2016).

The view we express with this paper is that the time component is central to value from digital trace data. Traditional data-based reporting does not fully exploit the data's temporal aspect or the possibility of considering the temporal structure of the events described in the data. Similarly, traditional statistical techniques encounter limitations when the data have to be analyzed in flow and in real time. The likelihood of organizations losing out on performance improvements while important phenomena remain hidden in the non-analyzed trace data is high. We demonstrated a useful approach for seeking precisely such patterns and relationships from huge masses of digital trace data or process data by looking at data collected from a business simulation. The decision tasks in such environments are similar to tasks carried out for managing real-world supply chains with their various temporal elements.

Applying multiple methods can support finding different aspects of the phenomenon studied. For instance, Miranda, Berente, Seidel, Safadi, & Burton-Jones, (2022) emphasize this aspect of applying rigor when one's purpose is to construct a computational theory. Our approach, combining process mining and variance modeling, meshes well with the DDM literature's emphasis on the importance of analyzing both the decision process and the outcome (Gonzalez, Lerch, & Lebiere, 2003). With process mining we are able to study how decision processes progress over time, and variance modeling lets us examine their links to outcomes from decision-making. This two-pronged approach offers new ways to study DDM and generate research questions.

2 Setting the Scene: Data Mining

Data mining is frequently used to extract nontrivial, implicit, unknown, and potentially useful information from data (Witten & Frank, 2005). The task is to use machine learning (ML) and statistical methods to find patterns in large datasets from database systems. Data mining is one step in the process of knowledge discovery in databases (KDD). In KDD, a well-established approach among data-mining researchers, the objective is to uncover valuable knowledge via data, and data-mining methods facilitate discovering patterns (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The five steps characterizing the established KDD process are 1) selection, 2) preprocessing, 3) transformation, 4) data mining, and 5) interpretation or evaluation of patterns that develops knowledge (Fayyad et al., 1996). One may apply awareness of these patterns in such activities as making decisions (see, for instance, Kusiak, 2002; Witten & Frank, 2005). A scientific data-mining process is iterative and interactive (Kamath, 2009). An inductive approach to analyzing the data exploits observations to generate patterns, which then inform the formation of hypotheses and theories (Goddard & Melville, 2004). Data mining is an inductive analysis technique that can assist greatly in such work as developing hypotheses from the data.

How, then, can researchers tap into the proliferation of event logs' and real-time technologies' application in organizations? Data-driven theory development represents a way forward. Discussing the potential of big data analytics in IS research, Müller, Junglas, vom Brocke, and Debortoli (2016) approached IS studies specifically from this angle. Naturally occurring user-generated big data can greatly benefit IS research because digital trace data of actual behavior are likely to be more reliable than surveys or data from self-reporting. Scholars have argued also that variance models explaining variation in a dependent variable could be combined with process models designed to analyze the temporal order of events (Sabherwal & Robey, 1993, 1995). This kind of combination holds particular potential to reveal new aspects of data and assist theorizing in IS research. Mining of process data is situated in the field of business-process management, which focuses primarily on creating, managing, and analyzing business processes (Breuker & Mantzner, 2014). However, bodies of process data in organizational contexts are notoriously difficult to analyze and preprocess.

Digital trace data can provide ample and fine-grained data about processes mediated or enabled by digital technologies (Pentland et al., 2020). The term “process data” is generally understood to refer to data on events that can aid in understanding how and why things evolve in a certain manner. Such data can have several units and levels of analysis, and they may be qualitative or quantitative. In addition, data of this nature can be described as “data of variable temporal embeddedness” and eclectic. With this combination of factors, both interpretation and analysis of process data prove especially challenging (see Langley, 1999). Computational theory discovery (CTD) offers another path, as suggested recently by Berente, Seidel, and Safadi (2019). The CTD method, a pattern-tracking approach favored by IS researchers, leads to scientific inventions via machine-learning algorithms used to track patterns from data (Berente et al., 2019). By serving as an inductive research approach (generating theories from data), it proves especially useful in developing insights from a host of phenomena by considering log files. Berente et al. (2019) state that hypothetico-deductive methods, which have thus far dominated scientific research, might come up against limitations of today’s computation technologies and may be somewhat a product of history.

Our approach, which complements those mentioned above, attests to how recently introduced quantitative data-mining techniques can, when applied in conjunction with variance models, afford generating research questions via digital trace data derived from processes. We propose an “exploratory technique” approach, a specific data-driven method that helps researchers translate complicated raw process data into interesting research questions as subjects for further validation. This approach employs perspectives from both CTD and KDD.

Data mining is essential to KDD. Importantly, the analysis involves both automated (ML-based) methods and application of human sense-making to the results. The aforementioned iterative and interactive nature of a scientific data-mining process naturally creates opportunities for each step to refine previous steps (Kamath, 2009). Additionally, data mining typically has a subjective flavor, with the analyst’s beliefs and prior knowledge of the phenomenon playing an important role in, for example, identifying interesting and useful patterns (Geng & Hamilton, 2006). Clearly, applying big data algorithms is not the only phase in the analysis: interpretation of the results produced by the algorithms is essential. Müller et al. (2016) describe the process as comparing algorithm-identified patterns/results in dialogue with the theoretical literature. The need for these new approaches stems from the increased complexity and volume of processes and their digital trace data, which render it highly challenging for humans to conceptualize the relevant events and detect patterns among them.

Below, we present a CTD-inspired framework for data-driven generation of research questions. Through this framework, we can 1) apply process-mining techniques to extract information from event-log files, 2) articulate appropriate use of modern regularized regression techniques for exploratory variance modeling, and 3) demonstrate how the problematization methodology developed by Alvesson and Sandberg (2011) can help researchers generate meaningful research questions.

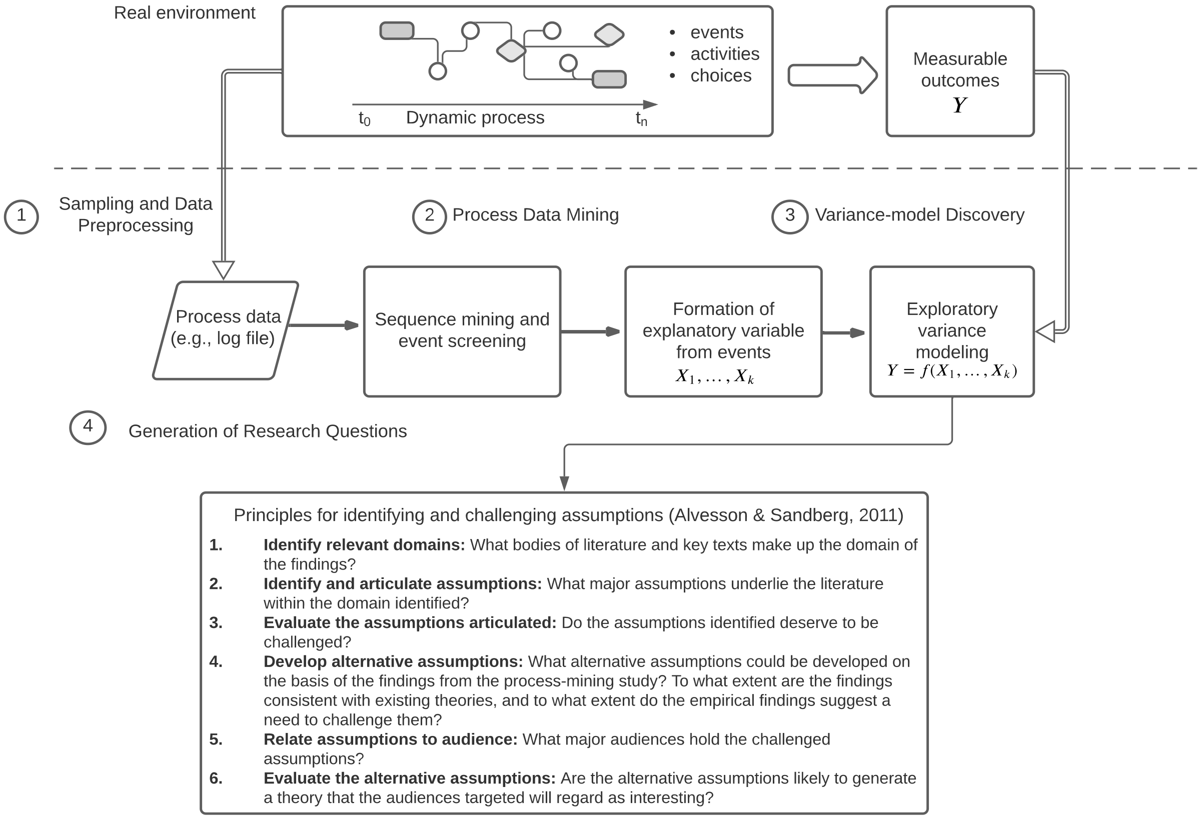
To demonstrate the framework in action, we applied it via a strictly exploratory technique in the context of a clock-driven simulation game wherein teams of business workers manage supply-chain-related tasks of simulated companies. Timely decisions are necessary for maintaining a balanced supply-chain process from procurement activities to delivery to customers. The game setting records behavior data in event logs that provide detailed real-time description of all decisions made by the participants. While our study does not permit us to form conclusions on how decision-makers act in dynamic decision-making environments or how these settings’ dynamic nature and time-urgency influence decision-making, this research is a step toward finding analysis methods with which to develop research that addresses such questions.

The remainder of the article is broken into two parts. Firstly, we present the proposed method for data-driven generation of research questions from digital trace data, specifically from process details. Then, we concretize it with a demonstration of the method’s application to make sense of dynamic decision-making behavior in a business context. The paper concludes with a discussion of the contributions of the suggested approach.

3 A Machine-Learning Approach to Generating Research Questions from Process Data

Scholars have discussed several strategies for making sense of process data (see Burton-Jones, McLean, & Monod, 2015; Langley, 1999; Pentland, 2013). Firstly, a researcher may consider the choice of concept types and relationships (e.g., a process, variance, or systems perspective). One can distinguish also among strategies that emphasize deductive (theory-driven) methods, abductive methods, and inductive (data-driven) ones. Thirdly, one can categorize strategies by the type of data used in the study (qualitative vs. quantitative). Finally, the researcher may find use in distinguishing between ostensive (abstract-pattern) and performative (specific-actions) aspects of process data (Pentland & Feldman, 2005). Though the options are often treated as separate or mutually exclusive (Markus & Robey, 1988; Mohr, 1982; Seddon, 1997; Van de Ven, 2007), scholars such as Pentland (1995), Langley (1999), de Guinea and Webster (2017), and Burton-Jones et al. (2015) advocate a hybrid approach and point out that the individual strategies for theorizing from process data should be seen not as strict competitors but as complementary lenses for sense-making that tend to differ, depending on the form of theory produced—none of which is intrinsically better or inherently worse.

The ML method we propose for data-driven process problematization entails four stages, presented in Figure 1, below: 1) sampling, 2) mining of process data, 3) variance-model mining, and 4) application of the framework proposed by Alvesson and Sandberg to generate research questions and conjectures from the data-mining tools' results. The following subsections discuss each step in detail.


Real environment
events
activities
choices
t$_0$
Dynamic process
t$_n$
Measurable outcomes
Y
1
Sampling and Data Preprocessing
2
Process Data Mining
3
Variance-model Discovery
Process data (e.g., log file)
Sequence mining and event screening
Formation of explanatory variable from events
$X_1, \ldots, X_k$
Exploratory variance modeling
$Y = f(X_1, \ldots, X_k)$
4
Generation of Research Questions

Principles for identifying and challenging assumptions (Alvesson & Sandberg, 2011)

1. **Identify relevant domains:** What bodies of literature and key texts make up the domain of the findings?
2. **Identify and articulate assumptions:** What major assumptions underlie the literature within the domain identified?
3. **Evaluate the assumptions articulated:** Do the assumptions identified deserve to be challenged?
4. **Develop alternative assumptions:** What alternative assumptions could be developed on the basis of the findings from the process-mining study? To what extent are the findings consistent with existing theories, and to what extent do the empirical findings suggest a need to challenge them?
5. **Relate assumptions to audience:** What major audiences hold the challenged assumptions?
6. **Evaluate the alternative assumptions:** Are the alternative assumptions likely to generate a theory that the audiences targeted will regard as interesting?

Figure 1. A machine-learning approach to generating research questions from process data.

3.1 Step 1: Sampling and Data Preprocessing

In broad terms, process data comprise events, activities, and choices, all ordered over time, and can be translated into stories about what has happened, who did what, and when (Langley, 1999; van der Aalst & Weijters, 2004). For specifically distinguishing between an event and a variable, we follow van der Aalst (2015) in defining an event as corresponding to the creation, deletion, or modification of objects and relations, which leaves a footprint by changing the underlying database and thereby affecting the state of an entity, represented by a variable. While process data traditionally are compiled and codified by humans, most of the process data analyzed today are generated automatically by business systems, such as ERP systems and online applications, then stored in machine-readable log files. Some companies have recently started mining the process data found in ERP systems to uncover, among other things, the root causes for costly deviations in their organizational processes (Lehto, 2020).

A typical process dataset can be assumed to contain at least the following details: 1) a case identifier (ID) at the level of the unit of analysis, such as the decision-making team in our context; 2) the task identifier, denoting a single step or activity in the process, in our study a single decision item; and 3) a timestamp, which indicates when particular activities took place (for instance, when some decision actually was made). Preparation of the data can encompass other items in addition—for example, the value of the activity and its geographical location. The dataset for process mining might encompass all data in a certain timeframe, including all activities recorded in the specified time period. The other option is to take into account all cases that were opened or completed within some fixed period.

Table 1 provides an example of process-log data. It contains information on two cases, with the corresponding "case identifiers" (e.g., unique labels for the decision-making teams involved). Together, these cases have three completed tasks (these might be three distinct decisions), with the attached task identifiers indicating that case 1 has only one task completed (two instances of task A) while case 2 performed two distinct tasks (tasks B and C). Importantly, many applications gather data on more attributes than the table illustrates, but one can extract sequence information from the timestamped data in even this simplified scenario.

Table 1. Imaginary Process-Log Data

Case identifier	Task identifier	Timestamp
case 1	task A	1
case 2	task B	1
case 1	task A	2
case 2	task C	2

One useful action in preprocessing of data, before the work proceeds to the actual process data mining, is to assess whether the process logs' data naturally fall into distinct temporal phases. These temporal categories often prove to be valuable elements in generating research questions via process data. A further action that can hold great value as a tool for meaningful preprocessing is polar selection. In the extreme case method, per Patton (1990), a case is chosen on the basis of its extreme value for the independent or dependent variable of interest (see Seawright & Gerring, 2008). Here, the selection of cases is not intended to be statistically representative of a population; the purpose is to maximize variance on the selected criterion dimension. This is an exploratory method that meshes well with the idea of data mining. One might choose cases with replication of a prior study in mind or to extend emerging theory, but the selection could equally well populate theoretical categories and supply examples of polar types (Eisenhardt, 1989). For instance, Pettigrew (1988) used polar selection when studying how a firm's ability to manage change in strategies and operations is linked to relative competitive performance. A polar selection method also supported our data analysis for studying differences between team performances.

3.1 2: Process Data Mining

Process data mining, process mining for short, is situated in the field of business process management, which focuses predominantly on creating, managing, and analyzing business processes (Breuker & Mantzner, 2014). Described by van der Aalst and Weijters (2004) as a method of distilling a structured process description from a set of real events, process mining is, in simple terms, analysis of business processes that is based on observed events from those processes. The starting point for process mining can be any database or any stream of events (e.g., van Zelst, van Dongen, & van der Aalst, 2018). The way in which events are stored and processed can take any of several forms, depending on the application. Process mining's purpose is to analyze the digital records (log files etc.) that, as Berente et al. (2019) emphasize, information technology almost always generates and explore the trends and patterns that emerge from the data—patterns and trends with potential to yield understanding of the underlying processes. Such analysis is growing in importance as rapid changes in modern organizational environments render more advanced ways of processing information crucial (Robey & Newman, 1996). According to van der Aalst and Weijters, the main idea behind process mining is a control-flow perspective, wherein the ordering of the tasks is central, enabling both the analysis of event ordering and the calculation of an event's processing time.

While many alternative process mining methods exist (cf. Mannhardt, de Leoni, Reijers, van der Aalst, & Toussaint, 2018; Tax, Sidorova, Haakma, & van der Aalst, 2018), sequential pattern mining, or sequence mining, is among the most commonly applied techniques (Breuker & Mantzner, 2014). A sequence is an ordered list of events, and the source database often covers several sequences, related to various objects (e.g., individual decision-makers or customers). To find patterns in the data, we study the frequency of certain sequences in the dataset. In the setting particular to our study, we defined the amount of support for any given sequence as the percentage of teams for which said sub-sequence featured in the decision-making events and actions was represented in the logs (Agrawal & Srikant, 1995). A specific sequence was deemed to have support if meeting certain minimal criteria. For instance, if the sequence database contained 10 case identifiers (for example, decision-makers), two of them displaying a sequence A-B, then the support for this sequence was quantified as 2/10, or 20 percent.

In cases such as ours, contiguous sequential pattern (CSP) mining is suitable: the analysis considers only those sequences in which the events in question are temporally adjacent (cf. Chen & Cook, 2007; Fournier-Viger, Lin, Uday, Koh, & Thomas, 2017). Finally, complex sequences are, in general, more useful and interesting. This does render sequence mining a relatively subjective task, in that the user's beliefs and existing knowledge of the associated phenomena play an important role in uncovering them (Geng & Hamilton, 2006). Notwithstanding its subjective nature, process mining can still be an immensely powerful tool for conceptualizing events and detecting patterns among them. Without the aid of sequence mining and other ML techniques, the path from the vast volume of raw data to meaningful events would be difficult—if not impossible.

At this point, it is reassuring to note that the outcome of sequential pattern mining does not differ across the various algorithms designed to discover sequential patterns (e.g., GSP, SPADE, PrefixSpan, SPAM, LAPIN, CM-SPAM, and CM-SPADE). All sequence-mining algorithms return the same set of sequences if run with the same parameters (e.g., with the minimum-support threshold chosen by the user) on the same database. While the competing algorithms do not differ in their output, they differ in efficiency, since their searches for patterns employ different strategies and data structures (cf. Fournier-Viger et al., 2017).

3.2 Step 3: Variance-Model Discovery

Although process mining is effective for identifying patterns of events and their occurrence over time, it does not reveal how the events may influence the state of an entity (a variable) or identify the effect of a contextual variable on events' evolution. To address research questions of this nature, process analysis needs to be reconciled with variance modeling. While process modeling focuses on temporal dependence of events, variance models represent endeavors to explain an outcome variable Y in terms of a set of independent variables. We found that the patterns of events detected through process-mining techniques can be used to construct suitable independent variables for variance modeling.

A key challenge in working with the sequences pinpointed via process mining is how to deal with the large number of alternative sequences detected by the algorithm. Commonly, a simple regression cannot be directly applied, since the number of explanatory variables readily grows too large for the time period considered. In pursuit of a parsimonious variance model, we advocate the use of regularized regression techniques to screen a subset of covariates, which are considered in the final model (Tibshirani, 1996). One option is to conduct regularized regression with a least absolute shrinkage and selection operator (LASSO). The technique accomplishes both variable selection and regularization to improve the prediction accuracy and interpretability of the model, by means of a constraint to the parameters in the model whereby some variables' regression coefficients shrink to zero. The variables whose coefficients vanish are not included in the final model. We used the LASSO technique to find our dataset's most important variables, which we then employed in our further analysis.

Taking the technique further, Efron, Hastie, Johnstone, and Tibshirani (2004) developed an algorithm based on least angle regression, wherein linear interpolation from the output sequence of LASSO solutions produces further LASSO solutions. In essence, the algorithm follows a stepwise procedure wherein a single addition or deletion of regression coefficients takes place in every step, and the LASSO mechanism selects a subset of regression coefficients for every shrinkage parameter that controls coefficient size and, thereby, the strength of regularization. When one employs this mechanism in generalized linear models (GLMs), the response variables can be non-normally distributed (Friedman, Hastie, & Tibshirani, 2010), so the operator is applicable also in logistic regression, wherein the response/dependent variable is binary (e.g., Feyerharm, 2015). The LASSO method imposes a fixed upper bound on the sum of the absolute values of the model's parameters, which one can do in a GLM setting by penalizing the negative log-likelihood via L1 normalization (Fonti, 2017). Friedman et al. (2010) have proposed an algorithm that applies coordinate descent to support least-squares solutions for linear regression models and iteratively reweighted least-squares solutions for logistic regression models with LASSO and elastic net penalties (Feyerharm, 2015).

3.3 Step 4: Generation of Research Questions

Gap-spotting and problematization are two ways to generate research questions (Alvesson & Sandberg, 2011). The dominant approach in management research is the former, identifying gaps in the literature that need to be filled. Here, the researchers strive for systematic contributions to scholarship but do not challenge the assumptions behind prior literature. Since it lacks problematization, gap-spotting is unlikely to increase the number of theories that have a large impact. Problematizing assumptions, on the other hand, constitutes the core of generating research questions. One can group the assumptions into five categories: 1) "in-house" ones, held by particular schools of thought; 2) "root metaphors," or broad images related to specific subjects; 3) "paradigm assumptions," with ontological, epistemological, and methodological aspects; 4) "ideology," or political, moral, and gender-related assumptions; and 5) "field assumptions," which are widely held assumptions across several schools of thought. Problematizing these is more likely to lead to new research questions and, ultimately, novel theories (Alvesson & Sandberg, 2011).

Problematization of assumptions follows the six core principles outlined by Alvesson and Sandberg (2011): In the first step, one must identify a domain of literature for assumption-challenging investigations. Next, assumptions underlying the chosen domain are identified and articulated. Step 3 is to evaluate those assumptions, with the most essential element being to assess the theoretical potential of challenging some specific assumption. For instance, challenging paradigm-related or field assumptions is more likely to lead to impactful theories, but assumptions of these types are trickier to identify and challenge successfully. Fourthly, grounds for alternative assumptions must be developed. At base, this step consists of precise formulation of alternative assumptions. In the fifth step, one considers the assumptions in relation to the audience: the assumptions being challenged have to be considered with reference to the groups holding them (typically comprising members of several audiences, perhaps differing somewhat in their assumptions). Finally, one must evaluate the alternative assumptions' grounds. The factor influencing the success of problematization most is how interesting the audience finds the alternative assumptions.

Davis (1971, as cited by Alvesson & Sandberg, 2011) presents three possible responses from the audience. "That's obvious!" is the first one, indicating that the alternative grounds are in line with the target audience's assumptions. Alternatively, the response might be "That's absurd!" This implies that the alternative grounds are at odds with every assumption held by the group targeted and, hence, are regarded as unbelievable. The third type of response is the ideal one: "That's interesting!" It is seen when the alternative grounds are consistent with some assumptions by the audience targeted and counter to others. Problematization may be particularly applicable and relevant for research in fields that display political domination and cognitive closure, with rootedness in a largely monolithic established tradition. Political domination refers to a situation wherein social-interest bias and/or political factors rather than sound ideas govern the production of knowledge. Cognitive closure is more likely in research fields colonized by some specific view of the world. Any debate with critical tones remains limited, and only a few contrary ideas challenge that particular view (Alvesson & Sandberg, 2011).

Figure 1, summarizing our approach, presents all four steps. Next, we offer a concrete example of it, from our collection of process-related digital trace data from the game, describing the team decision-making behavior. The nature of our approach, as an inductive way of applying process mining and variance-model discovery to seek patterns in the data and form research questions in light of the patterns observed, suggests that no theories or hypotheses are needed at the outset: the data analysis reveals the patterns from which we construct the questions. Such approaches have been described also as a bottom-up way of conducting research, in that observations are foundation stones in generating patterns (Goddard & Melville, 2004). In the quest to extract hidden decision-making patterns from our dataset, we followed the practices of both KDD (using data mining to comb the data for patterns in aims of revealing valuable knowledge) and the closely related CTD (applying ML to track the data's patterns in pursuit of scientific innovation), as referred to above.

4 Empirical Application: Decisions in a Business-Simulation Game

We are now ready to anchor the data-driven procedure diagrammed in Figure 1 empirically, illustrating it with the real-world case we examined. While dynamic decision-making environments are common in many areas of business, little is known about how decisions unfold in such environments and about any variations in decision-making patterns related to success. We used a simulation application that generates real-time log data describing the decision-making in the context of the game. Similarly to ERP systems, the simulations acted as information systems programmed to collect log data covering all actions by the decision-makers. The empirical environment in our research setting is unusual in applying clock-driven simulation (addressed below) rather than the much more commonplace turn-based type of processing. Clock-driven conditions provided for exceptionally rich, continuously recorded, and timestamped data. Our study exploited several sources of trace data: besides the decisions themselves, the details of the actions connected with those decisions, such as which reports were viewed in the lead-up to the decisions.

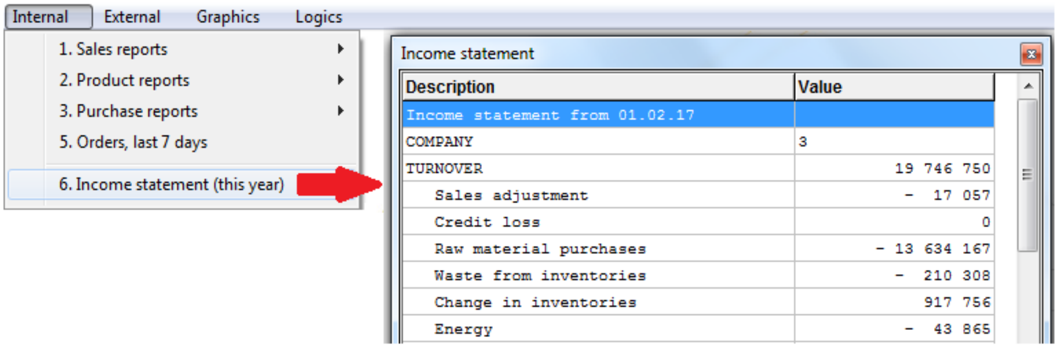
4.1 Illustration of Step 1: Sampling and Data Preprocessing

The source of our data was RealGame (Lainema, 2003), which operates in a real-time manner driven by the simulation's internal clock. Those engaging with the simulation made decisions as the simulation clock advanced, and the game engine noted each participant's actions in log tables. Thus, the system recorded each decision by each decision-making team as digital trace data during the simulation runs. We took these as our study's events, and each was categorized as either a "window activation" or a "decision." Activations consist of opening a report, non-interactive graphical window, or decision-making window, and decisions in this setting take place through activating a window, entering the decision (normally a numerical figure for a decision-making variable), and finally confirming the decision. The data from the simulation enabled us to experiment with the ML approach and test whether research questions could be generated from trace data.

Our data comprised 23 training sessions from the game's use at a large manufacturing organization in Finland. The training modules were part of the company's middle-management development program, completed in 2008-2011. All told, 407 individual employees participated in these sessions, once each. At the beginning of the simulation, the participants were assigned to teams of 2-3 people, for 144 teams in total, with preference being given to three-person teams: three-person teams accounted for more than 80 percent of the teams (115), and the remaining 26 teams (18%) had two members each. The team size was dictated partly by practical factors; three people cluster easily around a computer screen and can fluently and effectively refer to the information displayed there. Furthermore, our prior experience of the simulation was that it proves sufficiently challenging for a group of three people but no more. In the game setting, each team was responsible for running its own company. The teams, playing manufacturers of two kinds of product (mountain bikes and road bikes) competed against each other in common supply and end-product markets. The simulation environment was customized to some extent for the players' employer, especially with regard to delays in the supply-chain processes between suppliers and end customers. For data-mining purposes, we had access to log-file information about the decisions and actions during the gaming from 141 teams; for three teams, this information was missing. These decisions and actions formed the set of events analyzed.

For the game, which represents the supply chain's functioning, participants were in charge of the related procurement, inventory, manufacturing, sales, and delivery decisions in the simulated manufacturing firm. Furthermore, the players could affect their market success by making decisions on marketing and product development. The aforementioned novelty of the simulation's internal clock lay in advancing at a steady pace independent of participant actions, in one-hour cycles. Events at the simulated companies were visible to the participants accordingly, reflecting decision-making's character as a dynamic, time-bound task characterized by the need for informed decisions in conditions of temporal urgency. For example, components must be purchased well in advance if they are to arrive in time for production, which proceeds hour by hour in line with the bill of materials for the products. Likewise, customer orders need to be delivered via methods that are fast enough to honor the promised delivery times. In the game, participants' ability to manage the supply chain of the company is assessed in terms of several key performance indicators, such as gross margin, profit, delivery accuracy, order backlog, and average production costs.

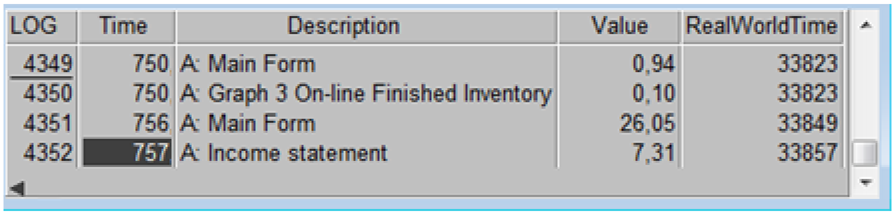
In the case of menu-item selection (see Figure 2), the real-world time of the activation is saved in memory, and the user moving away from the selected window (i.e., focus moving elsewhere in the game application) triggers recording a log line (trace data) with the window selected ("Income statement" in the figure), the activation time (in both simulation-internal and real-world time), and how long the selected window was active.



Description	Value
Income statement from 01.02.17	
COMPANY	3
TURNOVER	19 746 750
Sales adjustment	- 17 057
Credit loss	0
Raw material purchases	- 13 634 167
Waste from inventories	- 210 308
Change in inventories	917 756
Energy	- 43 865

Figure 2. An example of menu-item selection and activation: the user selects the income-statement window from the menu, and that window appears on the screen.

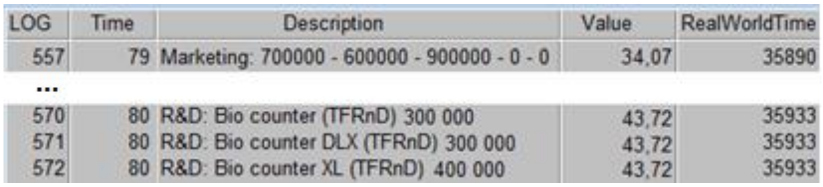
Figure 3, below, illustrates how this kind of activation is logged as trace data. The last line of the log file in the image is an example from an "Income statement" window activation. The log's first column indicates the line number, a serial number for the action (in this case, 4352 refers to the 4,352nd action recorded in the log file for this company). In our data, this serial number is the task identifier and is unique among the data from any one decision-making team. The "Time" column is for the action time in the simulation's internal time (the 757th hour since the simulation year began (here, February 1 at 1pm in internal time), while the final column, "RealWorldTime," gives the real-world time (seconds from the start of the day; 33857, or 33,857 seconds, = 9 hours, 24 min., 17 sec.). Both values are recorded at the end of the action. Finally, the "Description" column indicates the target of the decision and "Value" the activation length (how long the window was active—in this case, 7.31 seconds).



LOG	Time	Description	Value	RealWorldTime
4349	750	A: Main Form	0,94	33823
4350	750	A: Graph 3 On-line Finished Inventory	0,10	33823
4351	756	A: Main Form	26,05	33849
4352	757	A: Income statement	7,31	33857

Figure 3. An example of how simulation actions are recorded in the log file.

Figure 4 provides additional examples of how decisions are recorded in the log file.



LOG	Time	Description	Value	RealWorldTime
557	79	Marketing: 700000 - 600000 - 900000 - 0 - 0	34,07	35890
...				
570	80	R&D: Bio counter (TFRnD) 300 000	43,72	35933
571	80	R&D: Bio counter DLX (TFRnD) 300 000	43,72	35933
572	80	R&D: Bio counter XL (TFRnD) 400 000	43,72	35933

Figure 4. Examples of decisions recorded in the log file.

The first row in Figure 4 denotes a marketing decision. The user has chosen to make a monthly marketing investment of 700,000 euros in the first market region, 600,000 euros in the second, and so on. The final three rows shown indicate the monthly development budget dedicated to quality development for the various end products manufactured by the simulated company. Post-simulation, we divided the 141 teams into two sub-groups on the basis of success in the simulation as measured by the simulated company's business result (that is, "Profit" from the various simulated companies' income statements).

For our research context, we used polar selection to identify the team performing best and the team performing worst from each available simulation session, in terms of the teams' business results. The session-level highest- and lowest-profit performer (instead of, for example, the 10 teams performing best and worst overall from all 141 data-producing teams) were used. Not all data from all of the various sessions could be pooled, because each session has its own dynamics. For example, a session might feature the teams starting a price war, such that the profitability of all companies in that session suffers. The resulting company profiles could differ greatly—even for the teams performing relatively well in the respective session—from those yielded by a simulation session without price wars and thus exert a confounding influence.

We split the body of log data into two parts by the phase in the simulation: the first day of the game vs. the second day. This allowed us to study the strategies (patterns, in the parlance of data mining) with regard to two distinct time bands among the highest- and the lowest-performance teams. Furthermore, we were able to examine whether the decision-making patterns differed between the two stretches of time. A few log items were omitted from analysis because they lacked a timestamp. Additionally, log data for one team's sessions on the second day were missing. Therefore, our data encompassed 46 teams for the first day and 45 for the second. The log files contained 88 distinctly labeled actions or decisions, which were aggregated for analysis purposes under 30 labels (see the appendix). After our aggregation and splitting operations, we had 17,087 rows for the data-mining analysis for the well-performing teams from the first day and 17,011 from the second day, with the corresponding totals for the low-performance teams being 15,742 and 15,217, respectively. We could see from these numbers that the teams performing best visited and activated approximately 10 percent more decision-making windows and reports in both time periods than the worst-performing teams.

4.2 Illustration of Step 2: Process Data Mining

For the second step in applying the data-driven framework, we used sequence mining to identify potentially interesting decision-making patterns from the data. Typically, researchers apply constraints related to minimum support and sequence length, to reduce the number of patterns to a more manageable level in aims of finding highly interesting patterns via the data mining (Fournier-Viger et al., 2017). Limiting the search space is a subjective task, and one way of approaching this is to set the minimum-support threshold in line with user expectations for the phenomenon studied (Geng & Hamilton, 2006). This allows the algorithm to operate more efficiently: uninteresting patterns are removed from estimation. In sequential pattern mining, the task is to find all sequences exhibiting high frequency in the database of full sequences. If a sequence satisfies the user-specified minimum-support condition, it is considered a frequent sequential pattern. The analysis assumes that these patterns are of interest to the user and potentially useful for understanding the decision-making process (Fournier-Viger et al., 2017).

We set a minimum-support criterion of five percent for sequences to be included in our analysis. We used a maximum sequence length of 7 (and a minimum of 2). According to Geng and Hamilton (2006), generality is one essential measure of interestingness in pattern mining. That is, a pattern is considered more interesting if able to characterize larger chunks of information in the dataset, which implies greater support. Therefore, when a sequence was a sub-sequence of a longer one that met the minimum-support criterion of five percent, our analysis considered only the longer sequence (though its length still had to be no greater than 7, since choosing overly complicated sequences renders interpretation excessively difficult). For the first and second day of the game, 50 patterns satisfying the given criteria were selected for further phases of analysis. This also guaranteed that the patterns found would be actionable in the following phases, because it enables a reasonable pattern-selection strategy (Geng & Hamilton, 2006). We took the appearance of individual sequential patterns as independent binary variables, indicating the sequence's presence in or absence from the log data produced by the decision-making team in question.

4.3 Illustration of Step 3: Variance-Model Discovery

Our method applied the LASSO approach with binary response to find the set of independent variables, which is used to estimate the final model. We took the sequences as the independent variables for the LASSO analysis, and the division of teams between high and low performers was the dependent variable. The technique applied binary coding for the independent variables: we assigned the value 1 when the team displayed support for a particular decision-making pattern and 0 otherwise. The sequences from the regularized logistic regression selection served as our independent variables, and we selected as the dependent variable a binary variable indicating whether the given decision-making team is among the high vs. the low performers.

The variables shown in Table 2 and Table 3 are the decision-making patterns uncovered via sequence mining and then selected through regularized logistic regression analysis separately for the first and second day of the game. The patterns constitute sequences from the log files. The first column in these tables presents our label and interpretation for each pattern, and the second column shows which events form that pattern. For example, pattern x2 in Table 2 is formed from successive decision-maker activities of taking a look at production lines/shifts and changing these; returning to the production window; and, after this, checking inventory levels, returning to a production-line view, and changing the production shifts again.

The interpretations summarized in tables 2 and 3 are based on the order of the window activations and on the decisions that the participants made while interacting with those windows. For example, independent variable *x*1 is based on the content and decision-making possibilities presented in the views that are part of the following sequence of events: The window for sales offers shows the participants the sales terms of their offers, including the sales prices, lot sizes, payment term, and promised delivery time. After having checked this information, they move to the production window, from which they may, if they wish, manage work shifts and capacity investments. In this sequence, however, participants browse only “Production line” information, including production capacity and the bills of materials for the various products. After this, they activate the “Inventory” window and are shown the stock levels for components, semi-finished goods, and end products. Then, the participants return to the “Production line” window. This sequence was typical for decision-makers who were seeking an understanding of the relationships between the core activities in the simulation supply chain—sales offers, inventories, and production. In the case of independent variable *x*1, the sequence does not include decisions, just gathering of information. For the other independent variables (shown in tables 2 and 3), we developed similar interpretations for the sequence of individual windows activated by the participants, in terms of the window content and the decision-making types and alternatives.

Two patterns emerging from the second day’s activity represent roughly similar production-line activities. We combined these to form a single variable, *x*7, for “tuning production output (elementary)”: if either of these binary-coded variables has a value of 1, then the value for the combined variable is 1, and it is 0 otherwise. We took these variables/patterns as input to the next part of the analysis, as independent variables (as the tables show).

Table 2. Regularized Logistic Regression Selection Results for Day 1 of Simulation Gaming, with the Dependent Variable Being Simulation-Gaming Success

Day 1	
Independent variable (sequence)	**Pattern: Events in the sequence**
*x*1: Balancing the supply chain (balancing customer demand, production-line output, and raw-material supply) Interpretation: This is the core process of the simulation game—the participants work to understand the various elements in the supply chain and balance them with each other.	Sales offers -> Production line -> Inventory -> Production line
*x*2: Tuning production output (tuning the production-line work shifts’ response to the raw-material volumes in stock) Interpretation: Participants attempt to balance the production-line shifts (output) to respond to the supply of raw materials available. This is especially important early in the simulation, when participants have not yet created purchase routines based on customer demand.	Shifts -> Production decisions -> Production line -> Inventory -> Shifts -> Production decisions -> Production line
*x*3: Investment considerations (considering the need for bigger monetary investments in marketing and information systems, which are support functions for the operations functions) Interpretation: The decision-makers reflect on investments in marketing and information systems. They are aware that, to be able to serve the customers properly, they must automate the basic supply-chain operations and increase visibility to customers.	Marketing investments -> Information-system investments

*x*4: Sales offers in light of inventory (adjusting sales-offer terms and analyzing the inventory situation) Interpretation: The decision-makers adjust sales offers' terms to increase sales / support their sales strategy, then check stock levels again and assess how the inventory supports their sales decisions.	Offer decisions -> Sales offers -> Offer decisions -> Sales offers -> Offer decisions -> Sales offers -> Inventory
*x*5: Inventory and raw-material purchases (analyzing inventory levels and adjusting purchases of raw materials) Interpretation: The decision-makers assess their inventory levels, then purchase raw materials. The fact that participants go back and forth between inventory and purchases so many times in succession indicates that they may not be entirely confident in their ability to buy the right amount. This is probably related to lack of familiarity with the products, their sales volumes, and the bills of materials.	Inventory -> Raw-material purchases -> Inventory -> Raw-material purchases -> Inventory -> Raw-material purchases -> Decisions -> Raw-material purchases
*x*6: Research and development decisions (making decisions on research and development) Interpretation: The decision-makers study changes in the markets, and they react by adjusting the inputs from research and development.	Research and development decisions -> Report

Table 3: Regularized Logistic Regression Selection Results for Day 2 of Simulation Gaming, with the Dependent Variable Being Simulation-Gaming Success

Day 2	
Independent variable (sequence)	**Pattern: Events in the sequence**
*x*7: Tuning production output (elementary) Interpretation: Participants try to balance the production-line shifts, but this is not connected to the raw-material supply available.	Shifts -> Production decision -> Shifts -> Production decision -> Production line -> Shifts -> Production decision or Shifts -> Production decisions -> Shifts -> Production decisions -> Shifts -> Production decisions -> Production line
*x*8: Balancing production and sales (balancing production-line output with sales, with efforts to slow down or increase sales by adjusting the sales offers) Interpretation: Participants understand how to affect customer demand and are actively balancing production output and customer demand. For this reason, they change sales offers, to either slow down or increase sales.	Production decision -> Production line -> Sales offers -> Production line -> Sales offers
*x*9: Evaluating inbound flows and production (analyzing inbound material flows and assessing how they influence production-line output) Interpretation: The decision-makers analyze inbound material flows and try to ascertain how the inbound flows affect the production process.	Forthcoming changes -> Inventory -> Forthcoming changes -> Production line
*x*10: Shift decisions (managing production-line shift decisions) Interpretation: The decision-makers adjust production lines' work shifts (3 shifts are available) and try to achieve balance in shift use between/among those production lines that depend on each other.	Shifts -> Production decision -> Shifts -> Production decision -> Shifts -> Production decision

	-> Shifts

Our next step was to conduct a logistic regression analysis separately for the two time periods (day 1 and day 2), using only the independent variables identified in the regularized logistic regression analysis (see tables 2 and 3), and do the same with the dependent variable, comparing the best- and worst-performing teams as before. We present the results of the logistic regression in Table 4, which shows the coefficients, significance levels, and odds ratios for the independent variables.

Table 4. Results of Logistic Regression Analysis, with the Dependent Variable Being Simulation-Gaming Success

Independent variable	**Coefficient**	**Significance**	**Odds ratio**
Day 1			
Intercept	-0.68	0.449	
*x*1: Balancing the supply chain	1.33	0.058	14.43
*x*2: Tuning production output	-0.62	0.301	0.29
*x*3: Investment considerations	-1.67	0.027	0.04
*x*4: Sales offers in light of inventory	1.09	0.028	8.89
*x*5: Inventory and raw-material purchases	-1.67	0.063	0.04
*x*6: Research and development decisions	1.37	0.040	15.61
Model log-likelihood ratio	23.15		
Day 2			
Intercept	2.11	0.019	
*x*7: Tuning production output (elementary)	-1.26	0.008	0.08
*x*8: Balancing production and sales	1.34	0.039	14.58
*x*9: Evaluating inbound flows and production	1.36	0.034	15.15
*x*10: Shift decisions	-0.21	0.64	0.65
Model log-likelihood ratio	25.68		

Since the choice of independent variables involves a data-mining procedure, such a study is necessarily exploratory in nature. Therefore, the results of the logistic regression analysis are interpreted as indications of possible relationships between the independent and dependent variables that may point to interesting research questions for testing against additional data, from simulation- or real-world-based data.

In our case, the probability modeled is that of the team belonging to the set of well-performing ones. For the first-day model, the log likelihood ratio is 23.15 (p = 0.001), indicating an acceptable model. The Hosmer-Lemeshow goodness-of-fit statistics indicate an appropriate fit (p = 0.846), while Nagelkerke's R^2 is 0.527, which is considered acceptable. The results indicate that investment considerations (p = 0.027), sales offers in light of inventory (p = 0.028), and research and development decisions (p = 0.040) are potential predictors of a team's success on the first day of playing. The negative sign of the coefficient for investment considerations suggests that this pattern has an adverse effect on performance.

The estimated odds ratio is 8.89 (p = 0.028) for teams that applied the decision-making pattern labeled "sales offers in light of inventory" in their simulation-based gaming as opposed to those not using it. This means that when a team followed this pattern in its decision-making, the odds of that team being in the high-performance set are 8.89 times greater than the odds of it being among the poor performers. Our interpretation is that, when following this sequence, teams were actively balancing their sales offers and, thereby, their outbound material flow with their end-product stock levels. Thus, the teams carefully tracked whether they had sufficient goods for filling customer orders, and they handled cases of product shortages by adjusting the flow of incoming orders through adjustment of the sales terms.

We calculated the estimated odds ratio to be 15.61 (p = 0.040) for teams that employed the "research and development decisions" decision-making pattern relative to those not employing it. Accordingly, for a team using this pattern in its decision-making, the likelihood of belonging to the set of high performers is 15.61 times greater than that of belonging to the low-performance set. Our interpretation for this variable is that those teams displaying better performance examined the research and development options more carefully. When combined with the studying of reports, this may point to considering changes in the markets and then reacting by altering the input levels for research and development in response. In investigating opportunities to develop new products, as a route to expanding the simulated company's markets, the teams were probably trying to create a competitive advantage. This could be interpreted as an indication of active decision-making and strategic thinking.

For the variable named "balancing the supply chain," the estimated odds ratio is 14.43 (p = 0.058) for teams that followed this pattern as compared to those that did not; that is, where this pattern was followed in a team's decision-making, the odds of said team belonging to the set of high performers are 14.43 times greater than its odds of being one of the poor performers. The variable represents attempts to comprehend the functioning and dynamics of the holistic supply chain: trying to balance customer demand, production-line output, and supplies of raw materials. These teams were actively studying the logic of the supply chain, which is the most central of the processes in the game's simulation setting.

The estimated odds ratio is 0.04 (p = 0.027) for teams that operated in line with the pattern denoted as "investment considerations" in the gaming versus those teams not following it. A team that followed this pattern in its decision-making would show 25 times greater odds of belonging to the low-performance set than of being one of the high performers. Our interpretation for this variable is that the teams were hesitant to invest in information systems that might have aided in making the supply chain more efficient and streamlined (the game made two information-system investments available, so repeatedly coming back to this decision tells us that the team was not able to pick one of the systems with confidence). Also, they seem to have been unsure as to what level of marketing investments suffices and repeatedly returned to the options for changing their marketing investment. Perhaps the teams showed hesitance to make investments because of worries about the investments creating too great a risk to their profitability, or they may have been unable to understand the payoff of the investment.

Also worth mentioning are the issues highlighted by means of the "inventory and raw-material purchases" variable, which is statistically significant at a 10 percent level of risk (p = 0.063). The odds ratio is 0.04 (one would expect a team with this pattern to be 25 times as likely to belong to the set of poor performers than to be among the teams doing especially well). The pattern involves reactively checking inventory levels and changing the company's raw-material purchases accordingly. It betrays potential problems: an apparent lack of proper command of the fundamental cause-effect relations between these purchases and management of production output. In the game setting, this may have stemmed from unfamiliarity with the end products of the simulated company, its sales volumes, and its material costs.

Table 4 also shows the results for the second day of gaming, for which the model's log-likelihood ratio is 25.68 (p < 0.0001), indicating acceptable modeling. The Hosmer-Lemeshow test indicates good fit (p = 0.799), and the Nagelkerke's R^2 value is 0.580, suggesting that our model is suitable. Three variables seem to show an association with the response variable: "tuning production output (elementary)" (p = 0.008), "balancing production and sales" (p = 0.039), and "evaluating inbound flows and production" (p = 0.034).

We calculated an estimated odds ratio of 15.15 (p = 0.034) for teams that used the "evaluating inbound flows and production" pattern in the simulation versus those not doing so (that is, where this pattern was used in a team's decision-making, the odds of the relevant team belonging to the high-performance set are 15.15 times greater than those of it being a poor performer). This sequence represents a fine-tuned balancing process: decision-makers analyzing inbound material flows and trying to assess how those flows affect the production process. The sequence suggests that the team achieved control over the overall dynamics of the main supply-chain process. Another variable with a positive coefficient is "balancing production and sales." For teams following the corresponding pattern rather than not using it, the estimated odds ratio is 14.58 (p = 0.039), or, where this pattern was used in the decision-making, the odds of the team in question being one of the high performers are 14.58 times greater than those of it showing poor performance. This sequence, in turn, indicates decision-makers' active balancing of production output against customer demand, as opposed to passive waiting for production and demand to meet. Participants displaying this pattern understood how to influence customer demand, and they altered the simulated company's sales offers to either slow down or increase sales. In other words, these teams conducted fine-granularity analysis of interactions between sales decisions and production. The two sequences together characterize a fine-tuned process, which is possible because the overall supply chain is already balanced. Both variables express orientation toward active strategies.

Finally, for the “tuning production output (elementary)” pattern, the estimated odds ratio for teams using it relative to not using it in the game is 0.08 (p = 0.008). If this pattern was used in a team’s decision-making, one would expect it to be 12.5 times as likely to be in the low-performance set than to be among the high performers. This variable points to a team still struggling with the fundamental process of the simulation—a process that the high-performance teams had already started mastering on the first day. Mastery of this process, which entails maintaining balance among customer demand, production-line output, and raw-material supply, requires at least basic-level understanding of the production dynamics in the simulation setting. An alternative explanation is that the teams struggling in this regard lacked the capacity to look beyond this fundamental process: the basics required all their mental capacity, leaving none for expanding into new decision areas.

4.4 Illustration of Step 4: Generation of Research Questions

The patterns shown by the teams performing well on the first day of playing the game indicate more comprehensive use of information related to purchases, inventory, and sales offers—the fundamental activities of the company in the simulation. The high-performance teams directed attention to analyzing how inbound material flows affect the output of the production line. They also appeared to spend more time studying the reports on the company’s situation and on investing in research and development activities, both of which may be viewed as signs of more strategic thinking about how competitive advantage might be achieved. On the second day, in contrast, production-line activities related to material stocks no longer played so large a role among these teams. Over the second day of play, well-performing teams developed a more comprehensive view of the overall supply-chain process by carefully analyzing and modifying it. These teams also examined the interactions between production and sales offers. Hence, the teams showing the best performance demonstrated an ability to cope with the dynamics of the overall supply chain. These results highlight that those decisions related to active management of the inventory, supply chain, and distribution of sales offers are what distinguish teams in this category from others. Furthermore, the results highlight the importance of analyzing the logic of production: how inbound material flows affect production lines’ output. From the patterns detected, the well-performing teams were evidently more inclined to develop a certain structure to the decision-making: their focus initially was on learning the basics of the interdependent issues by actively balancing and studying various functions of the business, and then they started fine-tuning their decision-making related to management of the supply chain. The active strategy for sales-related decisions on both days of gaming was pronounced among successful teams, while reactive behavior (addressing issues only as they arise) was suboptimal for decision-making in this context. Proceeding from the foregoing discussion, we offer three key observations:

> Observation 1: The teams producing the best business results appear to utilize information differently from the least successful ones.
>
> Observation 2: Well-performing teams might demonstrate abilities related to multitasking and flexibility.
>
> Observation 3: Well-performing teams would be more likely to adopt an active strategy in their decision-making.

The remainder of this section is devoted to evaluating our findings, for revealing the extent to which they are consistent with preexisting theory and how much these new empirical findings may suggest a need to challenge it.

The information-use patterns revealed by our study suggest that success in decision-making demands a comprehensive approach, at least in the complexity-filled context considered here. This is in line with Ashby’s notion of requisite variety (1958). One possible explanation is that successful teams took a more analytical and deliberative approach. In complex and dynamic decision-making contexts, the ability to exploit information from various sources and functions is likely to lead to better decisions. For instance, when finding a positive correlation between exploration and performance in a simulation-based game, Edelstein (2013) identified the cause as a decision-making strategy that relies largely on extensive search and analytical reasoning.

Structural knowledge is an important factor in dynamic decision-making (Tremblay et al., 2017). Even in our simulation-based setting, the rules for conducting business could be rather complex, as they are in real life, so seeking additional information is vital for successful decisions. The decision-making patterns found in our log data revealed the logic followed by the decision-makers and how they perceived the causal relations of the business dynamics. The key differences stem from the fact that deliberative decisions are likely to be grounded in logical reasoning and cognitive effort, which need information as a propellant. In the game's dynamic and complex decision-making environment, the participants had to devote cognitive effort to ascertaining how the system works and how its various parts are linked to one another. The decision-making patterns of successful teams point to an aim of understanding the causal connections of the environment. With causal relations among sales, production, and inventory management being important in the game, well-performing teams studied these, while poor performers focused mostly on production-related tasks. This finding demonstrates that, relative to the teams showing the worst performance, those doing well took in the business dynamics more completely.

This is an important result, visible from both days of gaming: successful teams actively followed and applied information from many distinct business functions, while low-performance teams dedicated their attention mainly to production-related activities. This finding, evident especially from the logistic regression analysis (in which multiple activities related to production were statistically significant), echoes Güss, Tuason, and Orduña's (2015) conclusion that a fixed way of making decisions is suboptimal. Our results are also in line with research identifying a positive association between performance and sales activities (Edelstein, 2013; Güss, Strohschneider, & Halcour, 2000). Also, Güss et al.'s recent work (2015) found that success in real-time decision-making was correlated with active strategies that entail staying focused and demonstrating intention. In our setting, successful teams' active strategy visible for sales-related activities—which, being a key aspect of the simulation, require special attention—on both days suggests that those teams faring well indeed formed an intentional, focused strategy that accounts for the specific requirements for making decisions in the simulation. Moreover, the well-performing teams were able to cope with both sales and production activities, a factor that emerged as important in the earlier study by Güss et al. also (2000). Our research supports earlier work related to information too: Edelstein (2013) found that success in simulations is related to one's ability to pinpoint those information and variables that have the greatest influence on results. In Edelstein's work, the high-performance teams were able to uncover relevant information early in the simulation and did not get bogged down in one solution only.

An important question arises from our observation that, throughout the simulation-based training, successful teams maintained a wide focus, encompassing numerous business functions and processes, and actively sought information:

> New research question: *Why were the poorly performing teams unable or unwilling to change their decision-making patterns?*

Even though they must have been aware that the results were not good, the stragglers adhered to clearly unprofitable patterns. This topic merits further study and highlights our approach's potential in generating research questions.

The results demonstrate that our analytical procedure can uncover new research questions from the data. In previous studies related to decision-making, scholars have devoted most of their attention to studying static contexts. We can see a connection with examining the grounds for underlying assumptions as discussed above. The change in pace of business contexts necessitates various dynamic analysis capabilities if one is to achieve a fundamental understanding of how, for example, decision-makers behave under time pressure in dynamic environments and what kinds of differences the behaviors display from decision-making patterns analyzed in more traditional, static contexts. While research in static contexts is certainly valuable, it does not reveal the role of time in the behavior in a realistic manner.

We would propose as grounds for an alternative assumption that decision-making behavior in dynamic environments may even differ quite significantly from what has been identified in static contexts. Evidence from prior studies already attests to the importance of studying such elements of the decision-making environment as causal relationships and which information is valuable for success. Because the real-world environments involve time pressure in complex situations, traditional analysis and theoretical underpinnings might not be appropriate for the analysis. In addition, data gathered from these kinds of environments (big data content) demand modern advanced analytics solutions based on ML and artificial intelligence. Additionally, the inductive approach to conducting research can be fruitful for building theories in less-studied areas. The ideas of computation-based theory discovery may prove especially beneficial in the DDM context, as this paper illustrates. From the perspective of theory, our results are intriguing and contribute to scholarship. Our demonstration of the approach here highlights the value of multitasking and active strategies in dynamic decision-making conditions. In addition, it pinpoints the link between more strategic thinking and success. Organizations can derive practical benefit from that finding but also benefit significantly from the analysis framework presented. For instance, our approach might suit testing the development of "learning machines" for management. These systems could be trained with organizations' historical data for purposes of illuminating how specific kinds of decision-making behavior affect the business outcome. In addition, analysis tools of the sort presented here could reveal more general principles articulating how particular kinds of information serve particular aspects of decision-making. Perhaps our work's greatest contribution is best summarized thus: the patterns behind our results emerged from decision-making log files and would have remained opaque to the tools and methods of static studies.

5 Limitations

With regard to limitations, we stress that the proposed approach is explorative: inferences are based on data rather than predefined hypotheses. Especially for probing new research domains, the concept meshes well with the aim of finding tentative results as seeds for further work. Sequence mining is a heuristic method, which means that it offers several means of handling and constraining the search space. Also, the selection of independent variables is machine-driven with regularized logistic regression. This reflects the core idea of data mining: to reveal unseen and useful patterns from data. Our analysis of patterns was based on groups of players, the dynamics of which are notoriously challenging to identify. Our work with the data could not reveal many aspects of the teams' internal collaboration, such as whether one member of the group took control. Clearly, richer evaluation in such settings requires additional data, qualitative material especially, which were not available in this study.

It is clear that our dataset is not identical to what modern business systems produce. Any business simulation is inevitably a simplification of real-world enterprise systems. The digital trace data from real enterprise systems are more complex in volume, scope, and detail. Still, this should not pose an obstacle: tools developed in a simpler environment can be refined and later deployed for real-world settings. Likewise, it ought to be obvious that the method from our study should be verified in different contexts. For example, it would be extremely interesting to compare decision-making patterns between industries and examine how the associated patterns vary.

In addition, it would be worth looking at our data from different angles and seeking alternative methods for uncovering key decision-making patterns from trace data. In particular, patterns may await discovery in events and decisions that are separated from each other by large spans of time. How can we recognize these pattern pieces and put the puzzle together across time? This clearly is an outstanding challenge with trace data from real-world decision and implementation processes, which may extend for years.

Finally, while it may be obvious that digital trace data seldom shed light on “softer” aspects of decision-making---such as the nature of relationships between decision-makers---enterprise systems lack data on other pertinent elements too, perhaps most importantly the external environment that inevitably affects decisions. Modern organizations are tightly attached to suppliers, customers, and collaboration partners, with the nature and detail of these relationships not being reflected in the trace data. Analyzing digital trace data without taking into account the external environment as part of the web renders the analysis vulnerable to incorrect conclusions. For example, it may inaccurately present the management decision-making as questionable, even erroneous.

6 Conclusions

We can point to two interrelated contributions of our work. Our framework for data-driven generation of research questions from process data is one of them, combining the benefits of a process perspective and a variance perspective for process-data analysis. Specifically, the framework explains how researchers can use modern process-mining methods in combination with regularized regression techniques to generate meaningful research questions from large quantities of machine-readable process data. Our second contribution lies in our demonstration of how the proposed framework can operate in uncovering broader decision-making patterns from log-file data in a real-world empirical application. Our study showcases this approach's potential in finding meaningful and interesting process patterns from decision-making log data. Yet, at the same time, following the proposed methodology does not by itself guarantee a successful problematization outcome.

We find the framework similar to the problematization methodology proposed by Alvesson and Sandberg: it is a tool that can encourage researchers to go beyond the classical logic of gap-spotting. The objective is to proceed not from a set of *a priori* assumptions but from letting the dataset speak for itself and thereby generate research questions that may lead to development of more interesting theories. However, since this is an exploratory approach, one should not regard the framework as a replacement for more established approaches. Any research question arrived at via the proposed procedure should be subjected to a proper validation study. As with any data-mining or machine-learning approach, there is always a risk of overfitting: “If you torture the data long enough, it will confess to anything,” in the words of Nobel Laureate economist Ronald Coase. In line with the principles of machine learning, the best way to protect against overfitting is to carry out a carefully planned follow-up experiment to test the conjectures with an independent dataset not considered previously. A model that has overfitted one dataset is unlikely to show good performance for another sample. Conversely, if the findings yielded by means of the exploratory procedure are indeed valid, they should remain so when evaluated against a new sample of process data. It cannot be stressed enough that, while the framework presented here holds great potential to enrich scholarship, it is only a start, a proposed method for taking the initial step toward theorizing from digital trace data of processes. As Langley (1999) notes, we should not be shy about taking ideas from data and seeing whether they can be attached to theoretical perspectives in novel ways.

References

Agarwal, R., Gupta, A. K., & Kraut, R. (2008). Editorial overview—the interplay between digital and social networks. *Information Systems Research*, *19*(3), 243-252.

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the International Conference on Data Engineering* (pp. 3-14). Washington, DC: IEEE Computer Society.

Alvesson, M., & Sandberg, J. (2011). Generating research questions through problematization. *Academy of Management Review*, *36*(2), 247-271.

Ashby, W. R. (1958). Requisite variety and its implications for the control of complex systems. *Cybernetica*, *1*, 83-99.

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting voting behavior using digital trace data. *Social Science Computer Review*, *39*(5), 862-883.

Barton, D., & Court, D. (2012). Making advanced analytics work for you. Harvard Business Review, *90*, 79-83.

Berente, N., Seidel, S., & Safadi, H. (2019). Research commentary: Data-driven computationally-intensive theory development. *Information Systems Research*, *30*(1), 50-64.

Boulianne, S., Lalancette, M., & Ilkiw, D. (2020). “School strike 4 climate”: Social media and the international youth protest on climate change. *Media and Communication*, *8*(2), 208-218.

Bose, R. J. C., & van der Aalst, W. M. (2009). Context aware trace clustering: Towards improving process mining results. In *Proceedings of the 2009 Society for Industrial and Applied Mathematics (SIAM) International Conference on Data Mining* (pp. 401-412).

Breuker, D., & Mantzner, M. (2014). Performances of business processes and organizational routines: Similar research problems, different research methods—a literature review. In *ECIS 2014 proceedings: 22nd European Conference on Information Systems*.

Burton-Jones, A., McLean, E. R., & Monod, E. (2015). Theoretical perspectives in IS research: From variance and process to conceptual latitude and conceptual fit. *European Journal of Information Systems*, *24*(6), 664-679.

Chen, J., & Cook, T. (2007). Mining contiguous sequential patterns from Web logs. In *WWW ’07: Proceedings of the 16th International Conference on World Wide Web* (pp. 1177-1178).

Davis, M. S. (1971). That’s interesting! Towards a phenomenology of sociology and a sociology of phenomenology. *Philosophy of Social Sciences*, *1*, 309-344.

de Guinea, A., & Webster, J. (2017). Combining variance and process in information systems research: Hybrid approaches. *Information and Organization*, *27*(3), 144-162.

Driscoll, K., & Walker, S. (2014). Big data, big questions| working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication*, *8*, Article 20.

Edelstein, H. (2013). *Success and failure of experts and novices in a complex and dynamic business simulation* (Publication No. 447) [Graduate thesis, University of North Florida].

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. Human Factors, *4*, 59-73.

Efron, B., Hastie, T. J., Johnstone, I. M., & Tibshirani, R. (2004). Least angle regression (with discussion). *Annals of Statistics*, *32*, 407-499.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*(4), 532-550.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. J. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37-54.

Feyerharm, R. (2015). *Lasso regularization for generalized linear models in Base SAS® using cyclical coordinate descent* (paper 3297-2015). SAS. Retrieved from https://support.sas.com/resources/papers/proceedings15/3297-2015.pdf

Fiorio, L., Zagheni, E., Abel, G., Hill, J., Pestre, G., Letouzé, E., & Cai, J. (2021). Analyzing the effect of time in migration measurement using georeferenced digital trace data. *Demography*, *58*(1), 51-74.

Fonti, V. (2017). *Feature selection using LASSO* (research paper). Vrije Universiteit Amsterdam. Retrieved from https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf

Fournier-Viger, P., Lin, C.-W., Uday, R., Koh, Y. S., & Thomas, R. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition*, *1*, 54-77.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 122-156.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys*, *38*(3), Article 9.

Goddard, W., & Melville, S. (2004). *Research methodology: An introduction* (2nd ed.). Blackwell.

Gonzalez, C. (2005). Decision support for real-time, dynamic decision-making tasks. *Organizational Behavior and Human Decision Process*, *96*, 142-154.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591-635.

Güss, D., Strohschneider, S., & Halcour, D. (2000). *Strategies in complex and dynamic decision making: Cross-cultural analyses between India and Germany* (unpublished manuscript). Bamberg, Germany: Institut für Theoretische Psychologie, Otto-Friedrich Universität.

Güss, C. D., Tuason, M. T., & Orduña, L. V. (2015). Strategies, tactics, and errors in dynamic decision making in an Asian sample. *Journal of Dynamic Decision Making*, *1*, Article 3.

Hahn, G. J., & Packowski, J. (2015). A perspective on applications of in-memory analytics in supply chain management. *Decision Support Systems*, *76*, 45-52.

Jones, M. (2019). What we talk about when we talk about (Big) Data. *Journal of Strategic Information Systems*, *28*(1), 3-16.

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social Science Computer Review*, *35*(3), 336-356.

Kamath, C. (2009). *Scientific data mining: A practical perspective*. Society for Industrial and Applied Mathematics (SIAM).

Karanasios, S., Thakker, D., Lau, L., Allen, D., Dimitrova, V., & Norman, A. (2013). Making sense of digital traces: An activity theory driven ontological approach. *Journal of the American Society for Information Science and Technology*, *64*(12), 2452-2467.

Kusiak, A. (2002). Data mining and decision making. In *Proceedings of SPIE 4730, Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV*.

Lainema, T. (2003). *Enhancing organizational business process perception: Experiences from constructing and applying a dynamic business simulation game*. Publications of the Turku School of Economics and Business Administration, Series Ae-5:2003. Finland.

Langley, A. (1999). Strategies for theorizing from process data. *Academy of Management Review*, *24*(4), 691-710.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). *Science*, *323*(5915), 721-723.

Lehto, T. (2020). Real-time process mining with QPR ProcessAnalyzer 2020.1. *QPR Blog*. Retrieved from https://www.qpr.com/blog/real-time-process-mining-with-qpr-processanalyzer-2020.1

Machado, G. V., Cunha, Í., Pereira, A. C., & Oliveira, L. B. (2019). DOD-ETL: Distributed on-demand ETL for near real-time business intelligence. *Journal of Internet Services and Applications*, *10*(1), Article 21.

Mannhardt, F., de Leoni, M., Reijers, H., van der Aalst, W. & Toussaint, P. (2018). Guided Process Discovery—a pattern-based approach. *Information Systems*, *76*, 1-18.

Markus, M. L., & Robey, D. (1988). Information technology and organizational change: Causal structure in theory and research. *Management Science*, *34*(5), 583-598.

Miranda, S., Berente, N., Seidel, S., Safadi, H., & Burton-Jones, A. (2022). Computationally intensive theory construction: A primer for authors and reviewers. *MIS Quarterly*, *46*(2), iii-xviii.

Mohr, L. B. (1982). *Explaining organizational behavior*. San Francisco: Jossey-Bass.

Müller, O., Junglas, I., vom Brocke, J., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: Challenges, promises and guidelines. *European Journal of Information Systems*, *25*(4), 289-302.

Patton, M. (1990). *Qualitative evaluation and research methods*. SAGE.

Pentland, B. (1995). Grammatical models of organizational processes. *Organization Science*, *6*(5), 541-556.

Pentland, B. (2013). Desperately seeking structures: Grammars of action in information systems research. *ACM SIGMIS Database*, *44*(2), 7-18.

Pentland, B., & Feldman, M. (2005). Organizational routines as a unit of analysis. *Industrial and Corporate Change*, *14*(5), 793-815.

Pentland, B. T., Recker, J., Wolf, J. R., & Wyner, G. (2020). Bringing context inside process research with digital trace data. *Journal of the Association for Information Systems*, *21*(5), Article 5.

Pentland, B. T., Vaast, E., & Wolf, J. R. (2021). Theorizing process dynamics with directed graphs: A diachronic analysis of digital trace data. *MIS Quarterly*, *45*(2), 967-983.

Pettigrew, A. (1988). Longitudinal field research on change: Theory and practice. Paper presented at the National Science Foundation Conference on Longitudinal Research Methods in Organizations, Austin, TX.

Robey, D., & Newman, M. (1996). Sequential patterns in information systems development: An application of a social process model. *ACM Transactions on Information Systems*, *14*(1), 30-63.

Sabherwal, R., & Robey, D. (1993). An empirical taxonomy of implementation processes based on sequences of events in information system development. *Organization Science*, *4*(4), 548-576.

Sabherwal, R., & Robey, D. (1995). Reconciling variance and process strategies for studying information system development. *Information Systems Research*, *6*(4), 303-327.

Sandberg, J., & Alvesson, M. (2011). Ways of constructing research questions: Gap-spotting or problematization? *Organization*, *18*(1), 23-44.

Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly*, *61*, 294-308.

Seddon, P. B. (1997). A respecification and extension of the DeLone and McLean model of IS success. *Information Systems Research*, *8*(3), 240-253.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, 38(5), 503-516.

Tax, N., Sidorova, N., Haakma, R., & van der Aalst, W. M. (2016). Event abstraction for process mining using supervised learning techniques. In *Proceedings of [the] SAI Intelligent Systems Conference* (pp. 251-269). Cham, Switzerland: Springer.

Tendick, P. H., Denby, L., & Ju, W. H. (2016). Statistical methods for complex event processing and real time decision making. *Wiley Interdisciplinary Reviews: Computational Statistics*, *8*(1), 5-26.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B, *58*, 267-288.

Tremblay, S., Gagnon, J.-F., Lafond, D., Hodgetts, H. M., Doiron, M., & Jeuniaux, P. (2017). A cognitive prosthesis for complex decision-making. *Applied Ergonomics*, *58*, 349-360.

Van de Ven, A. (2007). *Engaged scholarship: A guide for organizational and social research*. New York: Oxford University Press.

Van der Aalst, W. (2015). Extracting event data from databases to unleash process mining. In J. vom Brocke & T. Schmiedel (Eds.), *BPM—driving innovation in a digital world* (pp. 105-128). Cham, Switzerland: Springer.

Van der Aalst, W., & Weijters, A. (2004). Process mining: A research agenda. *Computers in Industry*, *53*(3), 231-244.

Van Zelst, S. J., van Dongen, B. F., & van der Aalst, W. M. P. (2018). Event stream-based process discovery using abstract representations. *Knowledge Information Systems*, *54*, 407-435.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77-84.

Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques, San Francisco, Elsevier.

Yoo, Y. (2010). Computing in everyday life: A call for research on experiential computing. *MIS Quarterly*, 34(2), 213-231.

Yoo, Y., Boland, R. J., Jr., Lyytinen, K., & Majchrzak, A. (2012). Organizing for innovation in the digitized world. *Organization Science*, *23*(5), 1398-1408.

Appendix

Table A1. Aggregate Percentages for Each Type of Decision and Action during the Simulation

	Best teams on day 1, %	Worst teams on day 1, %	Best teams on day 2, %	Worst teams on day 2, %
Decisions				
Automation decisions	0.15	0.13	0.04	0.07
Production decisions	6.27	7.24	7.71	11.63
Research and development decisions	0.83	0.93	0.45	0.68
Marketing decisions	0.64	0.51	0.34	0.64
Offer decisions	3.36	3.95	2.56	2.37
Production-capacity investment decisions	0.08	0.04	0.29	0.19
Raw-material purchase decisions	3.63	3.16	2.88	1.72
Subcontracting decisions	0.00	0.02	0.01	0.31
Actions				
Delivery methods	1.70	2.41	0.91	1.37
Delivery reports	1.23	0.85	1.46	1.26
Automation for raw materials	0.43	0.85	0.88	1.26
Financial information and cash flows	5.43	5.80	2.98	2.85
Financial information and profit	2.83	1.91	2.85	1.94
Forthcoming changes	5.25	4.97	7.17	5.15
Graphs related to production	1.49	1.79	1.08	0.89
Information-systems investments	1.08	0.89	0.24	0.64
Inventory	10.53	10.74	11.60	10.12
Marketing and sales	2.91	2.22	2.67	2.75
Marketing investments	1.76	1.83	0.82	1.61
Orders	4.06	1.49	7.44	4.43
Companies’ production quality	2.21	1.62	1.75	1.08
Production	1.05	1.15	2.75	1.47
Production lines	10.91	11.68	13.73	13.43
Purchasing of raw materials	7.27	7.45	5.66	4.76
Reporting	2.67	2.19	2.56	2.07
Research and development	2.80	3.01	1.24	1.61
Sales offers	11.93	13.52	10.36	11.21
Selection of marketing reports	2.17	1.63	2.77	2.64
Shifts	4.90	5.46	4.71	9.27
Subcontracting deliveries	0.43	0.56	0.11	0.62

About the Authors

Henrik Kallio holds a PhD in Management Science from the Aalto University School of Business, Finland. His research interests include machine learning, data mining, and decision-making. His research has published, among others, in the *Journal of Information Technology Theory and Application, Journal of Financial Services Marketing, and International Journal of Information Technology and Decision Making*.

Pekka Malo is a tenured associate professor of statistics at Aalto University School of Business. His research has published in leading journals in operations research, information science, and artificial intelligence. Pekka is considered as one of the pioneers in the development of evolutionary optimization algorithms for solving challenging bilevel programming problems. His research interests include business analytics, computational statistics, machine learning, optimization and evolutionary computation, and their applications to marketing, finance, and healthcare.

Timo Lainema holds a PhD in Economics and Business Administration from Turku School of Economics, Finland. His research interests are learning through simulation gaming, conceptual change, knowledge sharing in virtual working contexts, and dynamic decision making in time intensive environments. He holds an Adjunct Professorship in Education in University of Turku, and in Information Research in University of Tampere. He has applied simulation games in university education, executive training and in in-house management training programs, both in Europe, North-America and Asia. Presently he is CEO in RealGame Business Training, a company which expertise is business simulation training.

Johanna Bragge holds a PhD in Management Science from the Helsinki School of Economics and works as Principal University Lecturer of Information Systems Science at Aalto University School of Business. She acts also as the Associate Programme Director of the Bachelor's Programme in Business and Economics at Aalto University. Her research interests include research profiling with text-mining and visualization tools, e-collaboration, and service co-creation. Her research has been published, among others, in the *Journal of the Association for Information Systems*, *Research Policy*, *Expert Systems with Applications*, *Journal of Business Research*, *Futures*, and *Group Decision and Negotiation*.

Tomi Seppälä is a tenured professor of Statistical Methods and Data Analytics at the University of Eastern Finland Business School, a Senior Fellow at Aalto University Business School and Adjunct Professor at Lappeenranta University of Technology Business School. He has a Ph.D. degree from Purdue University in Management Science and Quantitative Methods and a M.Sc. Degree in Mathematics from the University of Helsinki. He has also studied Communication and Journalism at the University of Helsinki and has experience as a TV reporter and journalist. He has earlier worked as a Professor in Business Mathematics, Statistics, Management Science and Finance at Helsinki School of Economics and Turku School of Economics. He has also lectured at many other Universities, involving Purdue University. He has taught courses in Statistics, Mathematics, Management Science/Operations Research, Logistics, and Finance. His research involves Statistical Methods in such areas as Decision Making, Operations Management, Logistics, International Business, Finance and Accounting, Marketing, Economics, Medicine, and Education.

Esko Penttinen is Associate Professor (Tenured) in Information Systems at Aalto University School of Business, Finland. Penttinen is an avid student of the coordination of work between humans and machines, organizational implementation of artificial intelligence, and governance issues related to outsourcing and virtual organizing. Penttinen leads the Real-Time Economy Competence Center at Aalto University and acts as the chairman of XBRL Finland. His main practical expertise lies in the development, assimilation, and economic implications of inter-organizational information systems, focusing on application areas such as electronic financial systems, government reporting, and electronic invoicing. Penttinen's research has appeared in leading IS outlets such as *MIS Quarterly, Journal of the Association for Information Systems, Information Systems Journal, Journal of Information Technology, Information Systems Frontiers, International Journal of Electronic Commerce*, and *Electronic Markets*.