Enhancing discrete-event simulation with big data analytics: a review

Abstract

This article presents a literature review of the use of the OR technique of discrete-event simulation (DES) in conjunction with the big data analytics (BDA) approaches of data mining, machine learning, data farming, visual analytics and process mining. The two areas are quite distinct. DES represents a mature OR tool using a graphical interface to produce an industry strength process modelling capability. The review reflects this and covers commercial off-the-shelf DES software used in an organisational setting. On the contrary the analytics techniques considered are in the domain of the data scientist and usually involve coding of algorithms to provide outputs derived from big data. Despite this divergence the review identifies a small but emerging literature of use-cases and from this a framework is derived a DES development methodology that incorporates the use of these analytics techniques. The review finds scope for two new categories of simulation and analytics use: an enhanced capability for DES from the use of BDA at the main stages of the DES methodology as well as the use of DES in a data farming role to drive BDA techniques.

Keywords: discrete-event simulation; analytics; big data; OR; literature review

Introduction

Analytics is built upon various approaches to data-driven analysis and is defined by Liberatore and Luo (2010) as the process of transforming data into actions through analysis and insights in the context of organisational decision making and problem solving. Robinson et al. (2010) and Lustig et al. (2010) provide an original classification of analytics into *descriptive analytics* – a set of technologies and processes that use data to understand and analyze business performance, *predictive analytics* – the extensive use of data and mathematical techniques to uncover explanatory and predictive models of business performance representing the inherit relationship between data inputs and outputs/outcomes and *prescriptive analytics* – a set of mathematical techniques that

computationally determine a set of high-value alternative actions or decisions given a complex set of objectives, requirements, and constraints, with the goal of improving business performance. Royston (2013) states that there is a clear and considerable mutual advantage in pulling analytics and OR more explicitly together, not least that it should strengthen links to real-world concerns. Ranyard et al. (2015) state that "the case for Business Analytics is that it includes a novel tool set that has only partially been absorbed into OR, whilst expanding the range of applications, e.g. in credit risk & scoring and on-line marketing. The result is an expansion of OR's opportunities." Discrete Event Simulation (DES) (Banks et al., 2000) is arguably the most popular OR simulation technique (Jahangirian et al., 2010) and the most used OR technique in practice (Brailsford, 2014). Law (2015) defines discrete-event simulation as concerning "the modelling of a system as it evolves over time by a representation in which the state variables change instantaneously at separate points in time. These points in time are the ones at which an event occurs, where an event is defined as an instantaneous occurrence that may change the state of the system." Leemis and Park (2006) note that "the word simulation is used not only to characterise the computational model (computer program) but also the computational process of using the discrete-event simulation model to generate output statistical data and thereby analyse system performance." Robinson (2014) describes three options for developing discrete-event models of spreadsheets, programming languages and specialist simulation software. In this investigation the focus is on specialist simulation software, defined here as commercial off-the-shelf software (COTS) which provides a relatively fast and easy model development for practitioners in an organisational setting. Hlupic (2000) reported that the majority (55.5%) of industrial users employ simulators (COTS), 22% employ simulation

languages and the remaining users employ ad hoc programs in a general purpose language or spreadsheets.

In general, both analytics and OR are concerned with the collection and analysis of data in order to find patterns for possible explanations or for testing a hypothesis. The distinction lies in that in analytics the role of models is often subsidiary whilst in OR models play an essential role and all other steps are oriented toward assisting the model-building process as well as to test the validity of the model (Barcelo, 2015). Thus whilst analytics is data-driven and has a focus on data and outputs and may have little knowledge of underlying processes, simulation is model-driven with a deep knowledge about processes. Thus it is claimed that simulation provides a more detailed and flexible way to evaluate potential process changes (Miller et al., 2013). Relating simulation to analytics it can be considered for use for descriptive, predictive and prescriptive outcomes (Greasley, 2019). For descriptive purposes simulation can be used in trace mode in which an historical data set is used to replicate past performance. This can be used to provide 'as-is' metrics or to check for conformance. Simulation's main role is for its predictive capabilities with its ability to project future scenarios. Simulation can also be used for prescriptive purposes. This can be undertaken by repeated experimentation to find a best solution either manually or by using optimisation software.

In order to investigate the relationship between analytics and OR further, a literature review is presented of the current use of DES as a modelling tool applied in the context of analytics. Powell and Mustafee (2017) argue that big data analytics can enhance simulation studies, but do not go into detail as to how. Business analytics is concerned primarily with the context in which techniques from OR and data science are

deployed (Hindle and Vidgen, 2018) and the review will help establish how the analytics toolkit can be applied to one of the most used techniques in OR. It will also help establish the relevance to DES in achieving one of the main aims of analytics: to derive decisions and actions (Davenport and Harris, 2017).

Addressing these questions is important as Mortenson et al. (2015) state that with other academic and practitioner communities engaging with analytics and increasing research in these areas, OR is in danger of being left behind; thus OR should follow the original conception of the discipline to use the most relevant methods available to solve business problems. This article contributes to the body of knowledge by providing a framework that assists in the integration of DES with analytics techniques by identifying the relationship between big data analytics techniques such as data mining and process mining and the main steps in a DES methodology.

In the next section the scope of the review around the categories of analytics techniques considered is outlined; the methodology for the literature review is presented in the section that follows. Summary results of the review are then provided with DES and analytics software identified. The paper goes on to outline the application of the analytics techniques in the context of a DES study methodology, and present a framework for the integration of DES and analytics tools. The final sections summarise the outcomes of the study, offer a research agenda and present the conclusions of the study.

Defining the categories of big data analytics techniques covered in the review

There are many analysis techniques that can be considered analytics techniques. Davenport and Harris (2017) list techniques for internal processes such as activity-based costing and multiple regression analysis; and techniques for external processes including econometric modelling, time series experiments and yield management.

Members of the OR community would consider some of these to be OR techniques. To avoid joining an as-yet unresolved discussion, the scope of this article is confined to what is often referred to as big data analytics (BDA) (De Reyck et al., 2017), which relates to the main analytic techniques which are used for analysis on large-scale datasets termed 'big data'. Big data analytics covers techniques such as association rule mining, decision trees, support vector machines and neural networks that undertake functions such as optimisation, classification, association and clustering (Nguyen et al., 2018). Articles that undertook big data analytics in conjunction with DES were found and categorised under the approaches of data mining, machine learning, process mining, visual analytics and data farming. These categories were chosen based upon the keywords used by the authors in the research found in this search. A brief description follows of each of these main approaches to big data analytics and the techniques associated with them. More details of the use of big data analytics techniques are in Dasgupta (2018), Evans (2017) and Foreman (2014) and in the individual review papers.

Data Mining

The distinction between data mining and machine learning is far from clear in the literature. This study uses the definitions of Davenport and Harris (2017) who define data mining in terms of identifying patterns in complex and ill-defined data sets. Particular data mining techniques include the following: *identifying associations* involves establishing relationships about items that occur at a particular point in time; *identifying sequences* involves showing the sequence in which actions occur (e.g. click-stream analysis of a web site); *classification* involves analysing historical data into patterns to predict future behaviour (e.g. identifying groups of web site users who display similar visitor patterns); and *clustering* involves finding groups of facts that

were previously unknown (e.g. identifying new market segments of customers or detecting e-commerce fraud). Data mining techniques found in the review include Self-Organising Maps (SOM) (Kohonen, 1995) which provide a visual map of data dependencies and Flexible Pattern Mining (FPM) (Bandaru et al., 2017) which aims at extracting patterns of rules within a given data set.

Machine Learning

According to Davenport and Harris (2017), machine learning describes technologies that can learn from data over time. Machine learning may use data mining techniques such as classification and clustering to manipulate data, but is distinguished by the use of algorithms that can learn from data, and therefore can build decision models that try to emulate regularities from training data in order to make predictions (Bishop, 2006).

The machine learning techniques found in the review are defined as follows (Dasgupta, 2018). Association rules mining (ARM) uses a rules-based approach to finding relationships between variables in a dataset. Decision trees (DT) generate rules that derive the likelihood of a certain outcome based on the likelihood of the preceding outcome. In general, decision trees are typically constructed similarly to a flowchart. Decision trees belong to a class of algorithms that are often known as CART (Classification and Regression Trees). Random Forest Decision Trees are an extension of the decision tree model, where many trees are developed independently and each "votes" for the tree that gives the best classification of outcomes. Support vector machines (SVM) are a class of machine learning algorithm that are used to classify data into one or another category using a concept called hyperplanes. k-Nearest Neighbours (k-NN) is a classification algorithm that attempts to find similarity based on closeness. Neural networks or Artificial Neural Networks (ANN) are a network of connected layers of (artificial) neurons. These mimic neurons in the human brain, that "fire"

(produce an output) when their stimulus (input) reaches a certain threshold. Only the network's overall input and output layers are "visible"; the others are hidden. Neural networks with three or more hidden layers are generally known as deep neural networks or deep learning systems. Naïve Bayes Classifier (NBC) is a supervised machine learning technique which employs a training set for classification. For more details on the techniques mentioned here, and references to some of the earliest work, see Frias-Martinez et al. (2006).

Process Mining

The concept of process mining is to use factual data to obtain an objective view on how processes are really executed (Mans et al., 2013). Process mining uses event data, recorded in an event log, which at a minimum contains information regarding the *case* (such as patient or order), the *activity* (what happened) and the *time* that the activity happened. The chronological ordering of events for a particular case yields a *trace*. A trace is similar to a simulation run in that it is only one example of possibly many different behaviours (van der Aalst, 2016). Process mining can be used to generate a type of simulation termed 'short-term simulation' (Rozinat et al., 2009b). Here historical data is projected forward and the simulation runs from the current state with the focus of the analysis on the transient behaviour. Process mining applications may use fuzzy mining which is a process discovery analytics technique that views process models as if they are geographic maps (van der Aalst, 2016).

Visual Analytics

The basic idea of visual analytics is to present large-scale data in some visual form, allowing the human to get insight into the data, draw conclusions, and interact with the data to confirm or disregard those conclusions (Feldkamp et al., 2015). Soban et al.

(2016) characterise visual analytics as particularly suited to exploring and understanding a particular data set with no preconceived notions of the expected outcome.

Data Farming

Data farming is purposeful data generation from any model evaluated computationally, including simulation models (Lucas et al., 2015). The machine learning literature often refers to data generated in this way as synthetic data (Patki et al., 2016). In this role Sanchez (2015) outlines the use of simulation to provide capabilities in data farming by generating large data sets. Here large-scale simulation experiments can be initiated by varying many input variables, examining many different scenarios or both.

Method

Current reviews of the use of DES typically are based within an application domain such as manufacturing (Negahban and Smith, 2014), focus on a specific application such as healthcare (Gul and Guneri, 2015) or a specific research issue such as behavioural modelling (Greasley and Owen, 2018). In order to provide an overview of the nature and scale of articles presenting the use of DES in the context of big data analytics and owing to the lack of any current reviews of this nature, a literature search was performed using the Scopus, ScienceDirect, Google Scholar, Emerald Insight and Web of Science databases. The review covers the period from 2006 which is associated with the popularisation of the use of BDA open source software such as Hadoop, R and Python (Davenport, 2017) and the publication of the seminal article 'Competing on Analytics' (Davenport, 2006). The review period runs until November 2018. The identified keyword terms *simulat** in combination with either *discrete-event* or *discrete event* were searched (by full-text if the database allowed) to identify studies in the

domain of discrete-event simulation. These keywords were combined with the keyword terms *visual analytics, data farming, data mining, machine learning, process mining, data analytics, big data analytics, business analytics*. Papers were filtered for relevance by title and then at abstract and finally at article level. This review is focused on the practical use of DES in conjunction with analytics in an organisational setting. Thus articles in domains such as disease outbreaks (Budgaga et al., 2016) are not considered and only implementations using what could be termed commercial off-the-shelf (COTS) DES software are considered (as defined in the INFORMS software survey at www.orms-today.org). Articles must contain an actual implementation of the DES model, providing details of the DES software and analytics software employed. Examples of excluded articles are Arroyo et al. (2010) that uses agent-based simulation rather than DES and Opçin et al. (2017) which does not employ a COTS DES software. The review of titles, abstracts and articles was undertaken by the authors of this article between June and November 2018. The review follows the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) guidelines (Moher et al., 2009) and the procedure is shown in Figure 1.

[TAKE IN FIGURE 1 ABOUT HERE]

The inclusion and exclusion criteria are shown in Table 1. The criteria for exclusion (numbered 1 to 5) are shown in brackets in Figure 1 and can be cross-referenced with the criteria numbered in Table 1. The literature review found 18 articles that met the criteria filtered from an original search count of 2883. These articles are listed in Table 2.

[TAKE IN TABLE 1 ABOUT HERE]

[TAKE IN TABLE 2 ABOUT HERE]

Results

Table 2 lists the 18 articles that were identified. This low number is probably due to the infancy of the area and the strict inclusion criteria demanding actual implementation of big data analytics techniques in conjunction with DES. These criteria were designed to ensure actual implementation details such as software and analytical techniques could be assessed. Articles are categorised by their area of analytics with data farming identified as the category for those articles that use DES to generate data.

In terms of a statistical analysis 18 articles were identified over the years 2006-2018 with 15 of these articles published since 2014 (Figure 2). The infancy of the area is emphasised by the relatively high proportion of conference papers found in the review (11 papers) compared to journal articles (7 papers). Of the 11 conference papers found, 7 of these were presented at the Winter Simulation Conference (WSC) organised by INFORMS (https://connect.informs.org/simulation/conferences/wsc-conferences). Two of the articles are in the Journal of the Operational Research Society and 2 in the Journal of Simulation.

[TAKE IN FIGURE 2 ABOUT HERE]

One issue when using a combination of OR techniques and analytics is the background of the practitioners themselves. An overview of this issue is provided by Harris and Mehrotra (2014) who conducted a survey of analytics professionals to see how they viewed their work in the organisation. About one-third of participants viewed themselves as data scientists with a computer science background, using tools such as R and Python that access and manipulate big data on distributed servers such as Hadoop. The remaining participants viewed themselves as analysts with an OR background working mainly with numeric data, using statistical and modelling tools to report, predict and optimise. For this review the background of the authors for the 18 articles

was categorised using the affiliation information supplied with the papers. The results are that for the 58 academic authors identified the majority of authors (45) are from Information Systems (26) and Engineering departments (19). Of the 13 authors from Business and Management Schools, only 7 are associated with articles from the post-2013 period. Although based on a small sample these results could indicate that the majority of recent research in analytics in relation to DES is being developed from a data science rather than an operational research perspective. A lack of interdisciplinary research teams is evident in that all articles have author teams assigned exclusively to one of the three discipline areas. One article has a further 2 people from industrial organisations cited bringing the total author count to 60.

Discussion

Well known examples of DES methodologies include Law (2015: 67) and Robinson (2014: 64). Here, in order to relate DES methodology to the big data analytics techniques found in the review articles, an adaption of the methodology of Greasley (2004) is used and shown in Figure 3. This has been chosen because it provides a useful correspondence between the stages of process mapping with process mining and modelling input data with data mining and machine learning applications.

[TAKE IN FIGURE 3 ABOUT HERE]

From the review it became apparent that DES is used in two main ways in conjunction with big data analytics. Figure 4 shows the use of DES to drive the BDA of machine learning and visual analytics. The analytics techniques are then used to facilitate stages of the DES methodology (Figure 3).

[TAKE IN FIGURE 4 ABOUT HERE]

Figure 5 shows the use of the BDA techniques of process mining, machine learning and data mining to facilitate stages of the DES methodology (Figure 3).

[TAKE IN FIGURE 5 ABOUT HERE]

In order to investigate how big data analytics techniques can improve the capability of DES the review articles are now thematised around the main stages of the DES methodology presented in Figure 3 and from the perspectives shown in Figures 4 and 5. The relationships found are then presented as a framework for the use of analytics in DES.

Data Collection

Data quality and availability are two of the challenging issues in many simulation projects. Inefficient data collection has been identified as one of the serious barriers to developing and deploying useful models within an appropriate timeframe and within budget (Onggo and Hill, 2014). In a study of data readiness of SMEs for DES modelling, Ivers et al. (2014) found that 88% of companies indicated that some or most of their data was collected manually with only 12% having fully automated collection by an IT system. Furthermore only 40% of data was held on centralised/integrated IT systems with 44% on local PCs and 16% held on paper. Volovoi (2016) outlines how DES is currently mainly an "offline" activity where the collection and processing of input data creates a major bottleneck in the modelling process. Volovoi (2016) further states that big data changes this balance by providing abundant data for input modelling and shifting the bottleneck to the modelling stage but may well require specific preparation in terms of infrastructure and data compatibility. An example of infrastructure requirements is provided by Kuo et al. (2015) who propose as a

preparation to the use of a DES model the installation of RFID enabled devices for data collection. In terms of data quality, data generated from sensors may also require a cleaning or pre-processing stage. Zhou et al. (2014) outlines the following procedures for preparing event log data which are automatically executed using Matlab. They consist of removing typos such as misspelt or joined up words, removing outliers such as out of range numeric values and replacing missing values with approximations. Marshall et al. (2015) address some of the practical considerations when integrating the use of big data with simulation models, such as gaining access to big data, data cleaning and privacy and security issues.

Real-time DES applications imply the need for interoperability between simulation software and software applications to provide automated data collection. In a review of data exchange standards Barlas and Heavey (2016) find that the only standard originating from the area of DES, CMSD (Core Manufacturing Simulation Data) is the most implemented standard for data exchange between simulation and other software applications. The standard is incorporated into COTS DES software such as Arena and ProModel and an example of the use of DES and CMSD is provided in Byrne et al. (2015). An example of a real-time DES application is provided by Celik et al. (2010). Here the aim is to incorporate real-time dynamic data into an executing DES model in order to facilitate short term decisions in a semiconductor manufacturing supply chain. Specifically the application provides a dynamic preventative maintenance schedule based on a DES prediction derived from real-time information obtained from machine sensor data.

A number of articles show how the need for historical big data can be avoided by the use of simulation to generate synthetic data, referred to as data farming (Lucas et al., 2015). An advantage of data farming is that the data generated is under the control

of the modeller with the amount generated solely dependent on the experimental setup and on the performance measures of interest, and can be adjusted through intelligent design of simulation experiments (Feldkamp et al., 2015). An example of the use of data farming is provided by Feldkamp et al. (2016) who present a case study of truck haulage in a gold mining facility. One aspect of the analysis is to investigate the relationship between haulage cost per ton and productivity. This was achieved by simulating 262,144 design points and using cluster analysis to group the results of the simulation runs.

In terms of the use of simulation as a generator of big data, none of the articles actually specifies the size of the data files generated. Traditionally big data is defined as a volume of data over 1 terabyte (Kumar, 2017). Whether these applications meet this criteria or not, what can be seen is the need for new data architectures and processing techniques that simulation practitioners may now be required to use. Feldkamp et al. (2017a) provides an example of the infrastructure used at these high volume data levels:

0.5 million simulation runs are performed using the COTS Siemens Plant Simulation (DES) parallelized on 10 machines

Output data is streamed in small blocks of files to a dedicated Apache Hadoop Distributed File System (HDFS) server

Data is then clustered using the k-means package from the Apache Spark Computing Framework

Data is then used to train a decision tree model using the Hoeffding Tree Implementation on the Massive Online Analysis software framework (MOA) (Bifet et al., 2010).

Process Mapping

In terms of the DES methodology the review finds that articles showing the use of process mining are relevant in the process mapping stage (Figure 5). This stage involves defining the scope and level of detail of the simulation, and using a diagramming tool such as a process map or activity cycle diagram to define the process flow of the model. Lamine et al. (2015) use process mining to build a conceptual model of an emergency call service that is then used to build a DES. The process mining was undertaken using the Fluxicon DISCO software tool (https://fluxicon.com/disco/) in order to discover the control flow of the incoming call regulation process. Zhou et al. (2014) also use the DISCO tool for process mining of an outpatient clinic. Fuzzy mining is used to generate a process map and k-means clustering is used to group patient types. Abohamad et al. (2017) use process mining to identify the workflows of patients in an emergency department of a hospital. Here a real-time patient tracking information system generated a data set with 229,971 event logs representing 40,777 patients. Process mining software was used to generate process flow models from this data which were subsequently used to develop a DES model. The authors state that the use of the process mining technique uncovered a large number of unique control-flows and identifying these process flows would have been an impossible task using traditional information sources for conceptual modelling.

Modelling Input Data

The review finds that data mining is relevant to the modelling input data stage. In a study of a hospital emergency department, Ceglowski et al., (2007) used data mining to group patients by similarity of treatment using a non-parametric method called SOM

(Kohonen, 1995) which is similar to a k-means clustering method (Kennedy et al., 1998) and undertaken using the Viscovery SOMine software tool. The analysis is abstracted to a level of how patient and treatment differences affect queue time, rather than modelling the physical movement of patients.

The remaining articles in this section are concerned with the use of machine learning. Bergmann et al. (2014) outline the identification of job dispatching rules with production data being used to train an artificial neural network (ANN). This application was developed in Bergmann et al. (2017; 2015) to include an assessment of various methods such as classification, decision trees and neural networks. The integration of these methods, including ANN, is considered as doable in every DES which supports external library interfaces and the studies are seen as the first step in achieving automatic simulation model generation. Priore et al. (2018) use simulation to generate training and test sets which are used for a variety of machine learning techniques in scheduling a flexible manufacturing system (FMS). The simulation is used to randomly generate 1100 combinations of 7 control attributes (such as work-in-progress and mean utilisation of the FMS). The simulation is then used to compare the scheduling performance of the trained machine learning based algorithms and further traditional scheduling rules such as SPT (shortest process time).

Glowacka et al. (2009) use association rule mining (ARM) to generate decision rules for patient no-shows in a healthcare service. The ARM method generates a number of rules and a subset of these were embedded as conditional and probability statements in the DES model. The authors state that when establishing the nature of the association between variables, the use of a rule-based approach such as ARM has advantages over a linear regression approach in that the variables (model factors) do not need to be traded off against each other and the rule-based model is easy to explain to

practising managers.

Gyulai et al. (2014) use simulation in conjunction with the random forest tree-based machine learning technique. The aim is to assign products to assembly lines in a way that minimises the overall cost of production. The article states that the main limitation of using random forest tree techniques for regression is that the regression cannot be applied beyond the ranges of the training dataset.

Building the Model

Dynamic-Data-Driven Application Systems (DDDAS) (Darema, 2004; Fujimoto et al., 2018) create online, adaptable data-driven models that change their specification in real-time in response to the event log data. This review has identified an example of a DDDAS using the Arena DES software in Celik et al. (2010). Here sensors installed in machines obtain data from the real system and transmit it to the simulation through a web server. From this data, algorithms embedded in the DES using the VBA facilities of Arena run to generate control tasks such as data filtering of abnormal behaviour, limiting data use due to availability of computational resources and providing a prediction of future performance. It is clear that the generation of real-time adaptable data-driven DES models poses particular challenges, with models needing to readjust on-the-fly and consistently perform validation, analysis and optimisation (Adra, 2016). Distributed simulation architectures are often needed to provide the speed of execution required (Taylor, 2018) and there is a need for an architecture for the interaction between the physical and simulated system (Onggo et al., 2018). These concepts can be considered within the related area of the use of simulation to provide a Digital Twin. A Digital Twin can be defined as an integrated simulation of a complex product/system

that, through physical models and sensor updates, mirrors the life of its corresponding twin (Negri et al., 2019).

Experimentation and Analysis

In terms of experimentation the main method employed in DES is to perform multiple replications of the simulation and construct confidence intervals of metrics of interest. Additional statistical tests such as t-tests may also be undertaken (Law, 2014) in which scenarios are compared based on variable input parameters.

Kibira et al. (2015) use association techniques of data mining to discover simulation input parameters that have a significant impact on the performance metric of energy consumption in a manufacturing plant. The authors call for standards in the areas of data collection, data representation, model composition and system integration in order to implement their framework for analytics and simulation optimisation. Uriarte et al. (2017) show the use of a multi-objective optimisation technique to find optimal solutions from simulation experiments. The use of the data mining technique of flexible pattern mining (FPM) (Bandaru et al., 2017) provides specific knowledge about the solutions generated by the optimisation stage. Aqlan et al. (2017) use a traditional simulation methodology to develop a model of a high-end server fabrication process. The model reports on a number of performance measures including cycle time and defective work. The defect parameters obtained from the simulation, such as product number and root cause for the defect, are written to an Excel spreadsheet. The spreadsheet then serves as an input data file for a neural network (ANN) model which predicts the defect solution (such as scrap, repair or return to supplier) and the corresponding confidence value of the prediction. The neural network has previously been trained using data collected on defect parameters. The authors intend to develop

the model to operate in real-time and provide decision support to failure analysis workers.

Visual Analytics are also relevant in the experimentation stage: four articles from the same team cover this area. Feldkamp et al. (2015) conduct a design of experiments analysis using the Plant Simulation software. For this large scale simulation experimental design it was found that a full factorial design would generate too many experiments, so a nearly orthogonal Latin hypercube (NOLH) sampling method was employed, with the number of experiments reduced to 491,160. This data was stored on a MongoDB noSQL database which provided the flexibility required to adapt to dataset modifications. Clustering methods were then used to explore simulation output data by treating each object in a cluster as a single simulation run allocated on selected parameter results. For example, in a two dimensional analysis the variables cycle time and throughput time may be used. Once the clusters are mapped out visually, analysts can investigate which input settings led to the corresponding systems performance measures that define this cluster. This example is developed in Feldkamp et al. (2016) who recognise a challenge in this type of analysis in terms of the speed and flexibility of the software required. In this study a nearly balanced nearly orthogonal hypercube design (Vieira et al., 2011) was used to generate 262,144 design points. The DES model was implemented using the SLX software (Strassburger, 2015) which is known for its speed of execution (Henriksen, 1999). Data is written to a MongoDB noSQL database and experiment design and data mining performed by MatLab. Feldkamp et al. (2017a) analyse simulation data with online stream-based data mining algorithms that work incrementally and allow data mining while simulation experiments are running. It was found that valid assumptions about the underlying system could be made even when only 10% of the experiments were completed. Feldkamp et al. (2017b)

investigate a manufacturing system's robustness against variance in the product mix. A visual analytics investigation is undertaken based on a binary decision tree that maps the relationship between simulation input factors and output factors.

These articles show that Visual Analytics has the potential to provide a useful additional tool when interpreting simulation output data. It is particular relevant to big data applications in that the visual method provides a way of synthesising large amounts of data and helps to reveal patterns and relationships between variables that might otherwise be hidden or difficult to find. However, the identification of relationships using visual inspection may be less precise and more open to interpretation than traditional approaches (Feldkamp et al., 2015). It will also require the training in and use of new analytics software and analysis methods by simulation practitioners.

A Methodology for the use of BDA in DES

The review articles present a number of DES methodologies for the use of the specific big data analytics techniques outlined in the individual articles. These include Kibira et al. (2015) who link the collection of raw data and information from the simulation conceptual development to generate analytics to assist the simulation build and optimisation stage. Lamine et al. (2015) and Abohamad et al. (2017) present methodologies in which process mining generates a process map that can be used as the basis for an as-is simulation model. Feldkamp et al. (2015; 2016) present a methodology for the use of DES for data farming which then through a process of data mining and visual analytics leads to knowledge discovery. Uriarte et al. (2017) present an approach to decision making in healthcare that combines DES, simulation-based multi-objective optimization (SMO) and data mining. Priore et al. (2018) present a framework for the use of DES to generate training and test examples for a machine learning algorithm. Further articles that outline a simulation methodology that employs BDA include

Onggo et al. (2018) that present the components of a symbiotic simulation system incorporating machine learning and Taylor (2019) that presents a workflow for distributed simulation in operational research that employs BDA techniques to analyse large-scale simulation output.

However these articles do not present a methodology that shows the relationship between each stage of the DES methodology and all the main categories of BDA techniques. Building on the DES methodology presented in Figure 3 and from an analysis of the linkages between BDA and stages in the DES methodology found in the review presented in Figures 4 and 5, a methodology for the use of BDA in DES is presented in Figure 6.

[TAKE IN FIGURE 6 ABOUT HERE]

In Figure 6 the direction of the data flow between the big data analytics techniques and the DES methodology stages is indicated by the flow line arrows. Firstly at the data collection stage Figure 6 shows that there may now be a requirement to collect big data instead of or supplemental to the traditional simulation data collection methods. The process mapping stage can now be facilitated by the techniques of process mining. The modelling input data, building the model and experimentation and analysis stages can be facilitated by data mining or machine learning techniques, and the experimentation and analysis stage can also be facilitated by visual analytics. The experimentation stage can be used to facilitate data farming to in turn generate big data as an alternative to collection from real system data sources. The methodology also incorporates the use of simulation to generate synthetic data to train and test machine learning algorithms for use in an analytics application rather than use in a subsequent simulation study. From Figure 6 it can be seen that big data analytics has the potential for impact at all stages of the DES methodology. There now follows a discussion of the

theoretical and practical implications of the relationship between DES and big data analytics techniques.

Summary

This section presents the theoretical and practical contributions in the context of OR and analytics and provides a research agenda around the challenges in enhancing DES with big data analytics derived from the review.

Theoretical Contribution

In general terms the review has provided exemplars of the combined use of the model-driven technique of DES with data-driven analytics techniques of data mining, machine learning, process mining, visual analytics and data farming. The relationship between these techniques identified in the review in terms of the nature of the data that is driving each category is presented in Figure 7.

[TAKE IN FIGURE 7 ABOUT HERE]

The categories in Figure 7 cover data-driven analytics techniques that use raw data to learn from the past to represent a *selected reality* based on the variables and observations included; and model-driven simulation techniques that use sampled data from the past to represent a *simplified reality*. The predictive capabilities of both of these approaches are limited by the transient nature of organisational processes. No matter how large the dataset used in a data-driven approach it may not describe a future behaviour owing to changes in the system causing that behaviour. This will occur at least until the new behaviour has been incorporated into the data provided to the learning algorithms. For model-driven approaches no matter how large the model we may not incorporate a future behaviour owing to the simplified representation of the

model, at least until we have recoded the model to incorporate the cause of that behaviour.

The outcome of this review has been to identify exemplars in two further categories, also shown in Figure 7. Data-driven simulation that uses data from analytics to drive simulation to provide a *digital reality*; and model-driven analytics that use data from simulation to drive analytics techniques to provide a *farmed reality*.

In terms of data-driven simulation applications these are demonstrated in a number of articles to facilitate a *digital reality*. These applications allow big data processed through process mining, data mining and machine learning techniques to advance DES process mapping, modelling input data, building the model and experimentation (figure 5). The use of data-driven tools to provide model building capabilities and thus enable reconfiguration of the simulation model to reflect the actual state of a system is a particularly important advance represented by the use of applications such as Digital Twins. This is termed *digital reality* as the approach is used to construct a real-time digital replica of a physical object. Thus the review shows the potential contribution of data-driven analytics techniques to all the main stages of the DES methodology, but DES practitioners need to take into account the limitations of the data-driven approach in terms of the use of historical data to represent future system behaviour.

In terms of model-driven analytics a number of articles use DES to create a *farmed reality* based on simulated data. From the review it is clear that there is a role for DES in training and testing machine learning algorithms that may be used in analytic applications or for subsequent use in simulation studies for input modelling and experimentation. Also large scale experiments can be observed using visual analytics (figure 4). This is termed a *farmed reality* in reference to the term data farming which

refers to the use of a simulation model to generate synthetic data. Here the limitation is based around the use of a sampled dataset that is a simplification of the raw data generated by the real system.

Practical Contribution

The review provides exemplars for practitioners in the use of big data analytics with DES software. DES practitioners in the OR domain typically combine the technical knowledge required to undertake DES such as model building and statistical methods with an understanding of an application domain such as manufacturing or healthcare. In a business setting Vidgen et al. (2017) found that analytics was undertaken by teams consisting of data scientists with data, statistical and IT skills, business analysts with deep domain knowledge and IT professionals to develop data products.

However, outside OR one often finds a much less charitable perspective on the role of business analysts, and by implication OR specialists generally. A recent article in a widely-read technology magazine compared the difference between a data scientist and a business analyst to that between a medical researcher and a lab technician (https://searchenterpriseai.techtarget.com/answer/Data-scientist-vs-business-analyst-Whats-the-difference). This may be related to perceptions of coding abilities. Although many experienced simulation practitioners began their simulation careers coding models in simulation languages such as SIMAN and using languages such as FORTRAN for file processing, in the light of the development of drag and drop interfaces in such tools as Arena, recent users may find it a particular challenge to adapt to the need for coding when developing a machine learning algorithm in Matlab, R or Python. This could be a reason why this review found that the majority of authors come from a data scientist background based in Information Systems and Engineering departments. This coding issue, potentially affecting wider credibility and respectability, applies to all analytics

techniques, not just big data analytics, and to many other areas of OR as well as DES. One way of addressing this issue may be to emphasise the need for training of DES practitioners in data science techniques (Marshall et al., 2016) and the adoption of a multi-disciplinary approach to research and training in the OR community (Taylor, 2015; Mortenson et al., 2015). An increase in interdisciplinary collaborative work between OR specialists and data scientists might serve to improve the perceptions each has of the other, but it is a salutary lesson to note that the review found no examples of such a collaboration.

Finally it may be that clients for these tools may lack knowledge of OR techniques or even statistics. Thus it may require approaches such as Visual Analytics to be used in collaboration with experts from academic institutions or private companies. Uriate et al. (2017) emphasise that whilst this can open the door to fruitful collaboration between research and practice it is important that the results of a project are adapted and presented in a way that meets the needs and backgrounds of the decision makers as much as possible.

Further Work and Research Agenda

The analysis leading to the framework for using DES with analytics raises a number of issues that together form a research agenda.

Data Interoperability

Simulation in conjunction with big data analytics techniques will require data interoperability which may be achieved through the use of data exchange standards. Barlas and Heavey (2016) discuss the high learning curve needed to implement data exchange standards correctly, and suggest as a potential future direction of research that more guidelines and tutorials be developed for the use of data exchange formats such as

CMSD.

Using Simulation to Train and Test Machine Learning Algorithms

DES in data farming mode can be used to train and test machine learning algorithms without any further use in the DES study (Gyulai et al., 2014; Priore et al. 2018). Data farming provides a repeatable and reliable environment in which the performance of machine learning algorithms can be compared and these studies may even include the simulation of poor quality data (Bergmann et al., 2015). Data farming may also be useful because even if real data is available it may not be available in the quantity required. For example in a factory environment of any complexity when testing a robot's performance it is unlikely that a dataset is available that is large enough to contain every possible combination of actions a robot may take. Data farming is also particularly relevant to the machine learning technique of reinforcement learning (Sutton and Barto, 2018) which involves an autonomous agent which learns to interact with its environment via trial and error. Deploying an untrained real system in a trial and error approach may be dangerous and so simulation provides a platform in which the agent can interact with its environment safely. Thus there is the possibility of further research in the use of DES for training and testing algorithms for current systems and for systems that do not currently exist.

Embedding Machine Learning Algorithms in DES

Bergmann et al. (2017) address the issue of how the decision rules derived by machine learning techniques can be used during the simulation run. The approaches suggested are to either call an external machine learning tool from the simulation system, or transfer or (re-)implement the decision model into the simulation system using its modelling and programming facilities. Bergmann et al. (2014) implement the first

approach using an interface between the simulation and the Matlab Neural-Network Toolbox. The second approach is used by Bergmann et al. (2017) who translate a decision tree into nested "if- statements" that can be coded into the model. Further research is needed to evaluate the use of approaches to embed machine learning algorithms in DES.

Continuous Machine Learning in DES

It should be noted all the examples found in the review of the use of machine learning use previously trained algorithms (trained using big data or data farming) to facilitate stand-alone DES models. Furthermore in the article by Celik et al. (2010) which uses DES in a real-time mode an embedded algorithm is used to detect abnormal sensor measurements. Here the algorithm is recalibrated at runtime in response to the data stream, but is described as static as its structure does not change dynamically during the simulation run (Celik et al., 2010). However Bergmann et al. (2014) do suggest that an ANN could be constantly trained in online simulation mode. Further studies are required to investigate the use of continuous (constantly trained) machine learning during simulation run-time for both stand-alone (offline) and real-time (online) applications.

Digital Twins

Although significant challenges are apparent in developing Digital Twin applications they offer the promise of extending the use of simulation from traditional stand-alone system design applications to simulation as a core functionality of systems by means of seamless assistance across the entire lifecycle from design, engineering, operations to service (Boschert and Rosen, 2016). Tao et al. (2019) present a review of the recent rapid growth of Digital Twin applications in industry.

One requirement for a Digital Twin is the ability for real-time model adaption which is considered under the term Dynamic-Data-Driven Application Systems (DDDAS). In the model building stage an example of a DDDAS was identified implemented using the Arena COTS DES (Celik et al., 2010). The implementation of adaptable data-driven models can be achieved through the use of a data-driven simulation approach (Goodall et al, 2019). This is primarily achieved by the definition of generic model objects with key data passed into the simulation from external files (Smith et al., 2018).

Apart from real-time model adaption, a simulation requirement for a Digital Twin is to provide an architecture for the interaction between the physical and simulated system which is considered under the term Symbiotic Simulation System (SSS) (Onggo et al., 2018). An SSS architecture proposes the use of simulation with BDA and data streaming technology.

In addition there is a need for an architecture to enable fast simulation execution speed when enabling a Digital Twin and this is considered under the term Distributed Simulation (DS) which uses parallel and distributed computing techniques and multiple computers to allow the processing of large-scale big simulations and the processing of associated outputs (Taylor, 2019). In terms of implementation using a COTS DES, Jain et al. (2017) state that the High Level Architecture (HLA) (Kuhl et al., 1999) standard for distributed simulation, updated for web services support (IEEE, 2010) and the standard for COTS Simulation Package Interoperability (SISO, 2010) are developments that have significantly facilitated the use of distributed simulation arrangements. Examples of cloud platforms that can facilitate rapid simulation execution include COTS DES packages such as Simio (https://www.simio.com/software/simio-portal.php) which uses the Microsoft Azure platform and Anylogic

(https://www.anylogic.com/features/cloud/) which uses the Amazon Web Services platform. Taylor et al. (2009) discuss interoperability between models using identical COTS simulation packages and between models using different COTS simulation packages.

Further applied case studies are required to evaluate the use of DES and BDA in Digital Twin implementations.

Simulation Experimental Design

In terms of the experimentation stage, analytics techniques which can be incorporated into DES methodology include data mining and visual analytics. Here one example of data mining is concerned with the analysis of the results of a DES optimisation (Uriarte, 2017). Over the last decade there has been substantial growth in the use of optimisation in DES (Hoad et al., 2015); further research in this area will assist in leveraging the capability of big data analytics in developing and refining optimisation methods for DES. Examples of visual analytics found by the review (Feldkamp et al., 2015; 2016; 2017a; 2017b) point to this being of most relevance to large scale simulation experimentation studies. This finding supports previous studies that have called for further research into the use of visualisation techniques across the breadth of OR/MS methods (Mortenson et al., 2015).

Generation of the Process Map using Process Mining

The link between the DES process mapping stage and process mining is direct and the capability of process mining to generate representative process maps for DES is shown in the review (Abohamad, 2017; Lamine, 2015; Zhou, 2014). Process mining offers the promise of fast construction of representations of complex processes incorporating

activities that may not be captured by traditional manual development of the DES process map (Lamine et al., 2015). However process mining does not generally generate a usable process map directly from the event logs but uses a variety of analytics techniques such as inductive mining for abstraction, dealing with issues such as noisy and incomplete data. Also current approaches to event log abstraction try to abstract events in an automated way that does not capture the required domain knowledge to fit business activities (Baier et al., 2014). Thus process mining does not necessarily generate process maps that are accurate and in the correct form for a DES study. Abohamad et al. (2017) suggest that process maps derived using process mining should be cross-checked and validated prior to developing simulation models using information obtained from interviews and process documentation. Thus there are research questions around the validation of DES models when using the data-driven abstraction methods of process mining.

Input Modelling using Machine Learning

Rabe and Scheidler (2014) propose the merging of data mining with standard simulation input modelling in order to increase the accuracy of DES input. Here an addition to input modelling in simulation methodology is presented in terms of the use of data mining (Ceglowski et al., 2007) and machine learning (Glowacka et al., 2009; Bergmann et al., 2014; 2015; 2017) to generate decision rules. The main issue for simulation methodology here is ensuring model validity. The logic of rules developed using techniques such as ARM can be inspected, but those using black-box analytics techniques (such as ANN or other deep learning approaches) cannot.

Integrating BDA capabilities into a COTS DES

The integration of BDA capabilities into a COTS DES would help facilitate the

increased use of the combination in future applications by reducing the technical expertise required to interface DES and BDA. A barrier to this may be the number of BDA libraries available and the need to carefully match the BDA library to the particular needs of the simulation study. In terms of current approaches to combining DES and BDA the main approach is to use the library-based application programming interfaces (APIs) provided in COTS DES packages. For example the review found the use of the C interface of the Tecnomatix Plant Simulation to access the functions of MatLab (Bergmann et al., 2017). The COTS DES Simio offers Visual C# user extensions in areas such user defined model selection rules. AnyLogic offers Java user extensions that can make use of Java-based libraries such as Deeplearning4j (https://deeplearning4j.org/). This approach does require coding ability so there is further work in embedding BDA capabilities using the current facilities of COTS DES software packages.

Limitations

In terms of limitations this article is based on a literature review method; although every effort was made to include all publications relevant to the topic of enhancing DES with big data analytics, some articles may not have been captured, especially those written in languages other than English. Furthermore, the process of evaluation and interpretation of the articles is reliant on the academic judgement of the author team.

Conclusion

This article provides an examination of the use of DES in the context of the main areas of big data analytics: data mining, machine learning, process mining, visual analytics and data farming. It is clear that the use of these techniques can lead to benefits for DES. The use of process mining offers the promise of providing a means of capturing

complex processes which have formerly been simplified out of the model. In addition the use of data mining and machine learning can supplement the input modelling, model building and experimentation stages of the DES methodology. Furthermore DES can be used to generate synthetic data for training and testing machine learning algorithms and for data visualisation studies. These combinations represent two new categories of using simulation and analytics of data-driven simulation creating a digital reality and model-driven analytics creating a farmed reality. Achieving these benefits requires progress on a research agenda around the integration of data-driven and model-driven methods that ensures valid DES models. It also requires DES practitioners with an operational research background to extend their capabilities into the areas of the data scientist - or to team up with data scientists - to avail themselves of these opportunities.

References

Abohamad, W., Ramy, A., & Arisha, A. (2017). A Hybrid Process-Mining Approach for Simulation Modeling, In W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, & E. Page, (Eds.) *Proceedings of the 2017 Winter Simulation Conference* (pp. 1527-1538). IEEE

Adra, A. (2016). Realtime predictive and prescriptive analytics with real-time data and simulation. In T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, & S. E. Chick (Eds.) *Proceedings of the 2016 Winter Simulation Conference* (pp. 3646-3651). IEEE.

Aqlan, F., Ramakrishnan, S., & Shamsan, A. (2017). Integrating data analytics and simulation for defect management in manufacturing environments, *Proceedings of the 2017 Winter Simulation Conference* (pp. 3940-3951). IEEE

Arroyo, J., Hassan, S., Gutiérrez, C., Pavón, J. (2010) Re-thinking simulation: a methodological approach for the application of data mining in agent-based modelling, *Comput Math Organ Theory*, 16, pp. 416-435.

Baier, T., Mendling, J., & Weske, M. (2014). Bridging abstraction layers in process mining, *Information Systems*, 46, pp. 123-139.

Bandaru, S., Ng, A.H.C., Deb, K. (2017) Data mining methods for knowledge discovery in multi-objective optimization: Part B – new developments and applications, *Expert Syst. Appl.*, 70, pp. 119-138.

Banks, J., Carson, J.S., Nelson B.L., & Nicol. D.M. (2000). *Discrete-Event System Simulation* 3rd ed. Prentice-Hall, Inc: Upper Saddle River, New Jersey.

Barcelo, J. (2015). Analytics and the art of modelling, *Intl. Trans. In Op. Res*., *22*(3), pp. 429-471.

Barlas, P. & Heavey, C. (2016). Automation of input data to discrete-event simulation for manufacturing: a review, *International Journal of Modeling, Simulation and Scientific Computing*, 7(1), pp. 1630001(1)-1630001(27).

Bergmann, S., Feldkamp, N., & Strassburger, S. (2015). Approximation of dispatching rules for manufacturing simulation using data mining methods. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, & M. D. Rossetti (Eds.) *Proceedings of the 2015 Winter Simulation Conference* (pp. 2329-2340). IEEE.

Bergmann, S., Feldkamp, N., & Strassburger, S. (2017). Emulation of control strategies through machine learning in manufacturing simulations, *Journal of Simulation*, *11*(1), pp. 38-50.

Bergmann, S., Stelzer, S., & Strassburger, S. (2014). On the use of artificial neural networks in simulation-based manufacturing control, *Journal of Simulation*, *8*(1), pp. 76-90.

Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B. (2010) MOA: Massive Online Analysis, *The Journal of Machine Learning Research*, 11, pp. 1601-1604.

Bishop, C.M. (ed) (2006). *Pattern Recognition and Machine Learning: Information Science and Statistics*. New York: Springer.

Brailsford, S. (2014). Theoretical comparison of discrete-event simulation and system dynamics. In S. Brailsford, L. Churilov, & B. Dangerfield (Eds.), *Discrete-Event Simulation and System Dynamics for Management Decision Making* (pp. 105-124). Chichester: John Wiley and Sons Ltd.

Budgaga, W., Malensek, M., Pallickara, S., Harvey, N., Breidt, F.J., Pallickara, S. (2016) Predictive analytics using statistical, learning and ensemble methods to support real-time exploration of discrete-event simulations, *Future Generation Computer Systems*, 56, pp. 360-374.

Byrne, J., Liston, P., Ferreira, D.C., Byrne, P.J. (2015) Cloud Based Capture and Representation for Simulation in Small and Medium Enterprises, *Proceedings of the 2015 Winter Simulation Conference*, IEEE, pp. 2195-2206.

Ceglowski, R., Churilov, L., & Wasserthiel, J. (2007). Combining Data Mining and Discrete Event Simulation for a value-added view of a hospital emergency department, *Journal of the Operational Research Society*, *58*(2), pp. 246-254.

Celik, N., Lee, S., Vasudevan, K., & Son, Y-J. (2010). DDDAS-based multi-fidelity simulation framework for supply chain systems, *IIE Transactions*, *42*(5), pp. 325-341.

Darema, F. (2004). Dynamic Data Driven Applications Systems: A New Paradigm for Application Simulations and Measurements, In *Proceedings of the International Conference on Computational Science* (pp. 662-669), Berlin, Heidelberg: Springer.

Dasgupta, N. (2018). *Practical Big Data Analytics* Birmingham, UK: Packt Publishing.

Davenport, T.H. & Harris, J. G. (2017). *Competing on Analytics: The New Science of Winning.* Boston: Harvard Business School Publishing Corporation.

De Reyck, B., Fragkos, I., Grushka-Cockayne, Y., Lichtendahl, C., Guerin, H., & Kritzer, A. (2017). Vungle Inc. improves monetization using big data analytics. *Interfaces, 47*(5), pp. 454-466. doi:https://doi.org/10.1287/inte.2017.0903

Feldkamp,N., Bergmann, S., & Strassburger, S. (2015). Visual Analytics of Manufacturing Simulation. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, & M. D. Rossetti (Eds.) *Proceedings of the 2015 Winter Simulation Conference* (pp. 779-790). IEEE.

Feldkamp, N., Bergmann, S., & Strassburger, S. (2017a). Online Analysis of Simulation Data with Stream-based Data Mining, *Proceedings of the 2017 ACM SIGSIM-PADS Conference* (pp. 241-248). New York: ACM.

Feldkamp, N., Bergmann, S., Strassburger, S., & Schulze, T. (2016). 'Knowledge Discovery in Simulation Data: A Case Study of a Gold Mining Facility'. In T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, & S. E. Chick (Eds.) *Proceedings of the 2016 Winter Simulation Conference* (pp. 1607-1618). IEEE.

Feldkamp, N., Bergmann, S., Strassburger, S., & Schulze, T. (2017b). Knowledge Discovery and Robustness Analysis in Manufacturing Simulations. In W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, & E. Page,

(Eds.) *Proceedings of the 2017 Winter Simulation Conference* (pp. 3952-3963). IEEE.

Frias-Martinez, E., Magoulas, G., Chen, S., & Macredie, R. (2006). Automated user modeling for personalized digital libraries. *International Journal of Information Management, 26*(3), pp. 234-248. doi: 10.1016/j.ijinfomgt.2006.02.006

Fujimoto, R., Barjis, J., Blasch, E., Cai, W., Jin, D., Lee, S., Son, Y-J. (2018) Dynamic Data Application Systems: Research Challenges and Opportunities, *Proceedings of the 2018 Winter Simulation Conference,* (pp. 664-678). IEEE.

Glowacka, K.J., Henry, R.M., & May J.H. (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling, *Journal of the Operational Research Society*, *60*(8), pp. 1056-1068.

Goodall, P., Sharpe, R., West, A. (2019) A data-driven simulation to support remanufacturing operations, *Computers in Industry*, 105, pp. 48-60.Greasley, A. (2004). *Simulation Modelling for Business.* Aldershot, UK: Ashgate Publishing Ltd.

Greasley, A. (2019). *Simulating Business Processes for Descriptive, Predictive and Prescriptive Analytics*. De Gruyter.

Greasley, A., & Owen, C. (2018). Modelling people's behaviour using discrete-event simulation: A review, *International Journal of Operations and Production Management*, *38*(5), pp. 1228-1244.

Gul, M., & Guneri, A.F. (2015). "A comprehensive review of emergency department simulation applications for normal and disaster conditions", *Computers and Industrial Engineering*, *83*, pp. 327-344.

Gyulai, D., Kádár, B., & Monostori, L. (2014). Capacity planning and resource allocation in assembly systems consisting of dedicated and reconfigurable lines, *Procedia CIRP*, *25*, pp. 185-191.

Henriksen, J.O. (1999). SLX – The X is for eXtensibility. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.) *Proceedings of the 1999 Winter Simulation Conference* (pp. 167-175). IEEE.

Hindle, G.A., & Vidgen, R. (2018). Developing a business analytics methodology: A case study in the foodbank sector, *European Journal of Operational Research*, *268*(3), pp. 836-851.

Hlupic, V. (2000). Simulation Software: An Operational Research Society Survey of Academic and Industrial Users, *Proceedings of the 2000 Winter Simulation Conference* (pp. 1676-1683). IEEE

Hoad, K., Monks, T., & O'Brien, F. (2015). The use of search experimentation in discrete-event simulation practice, *Journal of the Operational Research Society*, *66*(7), pp. 1155-1168.

IEEE (2010). *1516.1-2010 IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) – Federate Interface Specification*, NJ: Institute of Electrical and Electronics Engineers.

Ivers, A.M., Byrne, J. & Byrne, P.J. (2016) Analysis of SME data readiness: a simulation perspective, *Journal of Small Business and Enterprise Development*, 23(1), pp. 163-188.

Jahangirian, M., Eldabi, T., Naseer, A., Stergioulas, L.K., & Young, T. (2010). Simulation in manufacturing and business: a review, *European Journal of Operational Research*, *203*(1), pp. 1-13.

Jain, S., Shao, G., Shin, S-J. (2017) Manufacturing data analytics using a virtual factory representation, *International Journal of Production Research*, 55(18), pp. 5450-5464.

Kennedy, R.L., Lee, Y., Van Roy, B., Reed, C.D., Lippman, R.D. (1998) *Solving Data Mining Problems through Pattern Recognition*, Prentice Hall: Englewood Cliffs, NJ.

Kibira, D., Hatim, Q., & Kumara, S. (2015). Integrating data analytics and simulation methods to support manufacturing decision making. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, & M. D. Rossetti (Eds.) *Proceedings of the 2015 Winter Simulation Conference* (pp. 2100-2111). IEEE.

Kim, B.O., Kang, B.G., Choi, S.H., & Kim, T.G. (2017). Data modelling versus simulation modelling in the big data era: case study of a greenhouse control system, *Simulation: Transactions of the Society for Modeling and Simulation International*, *93*(7), pp. 579-594.

Kohonen, T. (1995). *Self-organizing Maps.* Berlin: Springer.

Kritzinger, W., Karner, M., Traar, G., Henjes, J., Sihn, W. (2018) Digital Twin in manufacturing: A categorical literature review and classification, *IFAC PapersOnLine*, pp. 1016-1022.

Kuhl, F., Weatherly, R., Dahmann, J. (1999) *Creating Computer Simulations: An Introduction to the High Level Architecture*, NJ: Prentice Hall.

Kumar, U.D. (2017) Business Analytics: The Science of Data-Driven Decision Making, New Delhi: Wiley.

Kuo, Y-H., Leung, J.M.Y., Tsoi, K.K.F., Meng, H.M., & Graham, C.A. (2015). Embracing Big Data for Simulation Modelling of Emergency Department Processes and Activities, *Proceedings of the 2015 IEEE International Congress on Big Data*. (pp. 313-316). Washington, DC: IEEE Computer Society.

Lamine, E., Fontanili, F., Di Mascolo, M., & Pinguad, H. (2015). Improving the Management of an Emergency Call Service by Combining Process Mining and Discrete Event Simulation Approaches. In L. M. Camarinha-Matos, F. Bénaben, W. Picard (Eds.) *Proceedings of the 16th Working Conference on Virtual Enterprises (PROVE)*, *Albi, France* (pp. 527-538). Berlin: Springer.

Law, A.M. (2015). *Simulation Modeling and Analysis*, 5th Edition, New York: McGraw-Hill Education.

Leemis, L.M. and Park, S.K. (2006) Discrete-Event Simulation: A First Course, New Jersey: Pearson Education.

Liberatore, M.J., & Luo, W. (2010). The Analytics Movement: Implications for Operations Research, *Interfaces*, *40*(4), pp. 313-324.

Lucas, T.W., Kelton, W.D., Sanchez, P.J., Sanchez, S.M., & Anderson, B.L. (2015). Changing the Paradigm: Simulation, Now a Method of First Resort, *Naval Research Logistics*, *62*(4), pp. 293-303.

Lustig, I., Dietrich, B., Johnson, C., & Dziekan, C. (2010). The Analytics Journey. Analytics Magazine, (November/December), (pp. 11–18). Retrieved from http://analytics-magazine.org/the-analytics-journey/

Mans, R., Reijers, H., Wismeijer, D., & van Genuchten, M. (2013). A process-oriented methodology for evaluating the impact of IT: A proposal and an application in healthcare, *Information Systems*, *38*(8), pp. 1097-1115.

Marshall, D.A., Burgos-Liz, L., Pasupathy, K.S., Padula, W.V., Ijzerman, M.J., Wong, P.K., Higashi, M.K., Engbers, J., Wiebe, S., Crown, W., Osgood, N.D. (2016) Transforming Healthcare Delivery: Integrating Dynamic Simulation Modelling and Big Data in Health Economics and Outcomes Research, *PharmoEconomics*, 34, pp. 115-126.

Miller, J.A., Cotterell, M.E., & Buckley, S.J. (2013). Supporting a modeling continuum in scalation: From predictive analytics to simulation modeling. R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, & M. E. Kuhl (Eds.) *Proceedings of the 2013 Winter Simulation Conference* (pp. 1191-1202). IEEE.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., & The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement, *PLoS Medicine*, *6*(7), pp. 1-6.

Mortenson, M.J., Doherty, N.F., & Robinson, S. (2015). Operational Research from Taylorism to Terabytes: a research agenda for the analytics age, *European Journal of Operational Research*, *241*(3), pp. 583-595.

Nagahban, A., & Smith, J.S. (2014). Simulation for manufacturing system design and operation: Literature Review and Analysis, *Journal of Manufacturing Systems*, *33*(2), pp. 241-261.

Negri, E., Fumagalli, L., Cimino, C., Macchi, M. (2019) FMU-supported simulation for CPS Digital Twin, International Conference on Changeable, Agile, Reconfigurable and Virtual Production, *Procedia Manufacturing*, 28, pp. 201-206.

Nguyen, T., Zhou, L., Spiegler, V., Ieromonachou, P., Lin, Y. (2018). Big data analytics in supply chain management: A state-of-the-art literature review, *Computers and Operations Research*, 98, pp. 254-264.

Onggo, B.S., Mustafee, N., Juan, A.A., Molloy, O., Smart, A. (2018) Symbiotic Simulation System: Hybrid systems model meets big data analytics, *Proceedings of the 2018 Winter Simulation Conference*, IEEE, pp. 1358-1369.

Opçin, A.E., Buss, A.H., Lucas, T.W., Sánchez, P.J. (2017) Modeling Anti-Air Warfare with Discrete-Event Simulation and Analyzing Naval Convoy Operations, *Proceedings of the 2017 Winter Simulation Conference*, IEEE, pp. 4048-4057.

Patki, N., Wedge, R., Veeramacheneni, K. (2016) The synthetic data vault, *Proceedings of the 3rd IEEE International Conference on Data Science and Advanced Analytics*, IEEE, pp. 399-410.

Powell, J. H., & Mustafee, N. (2017). Widening requirements capture with soft methods: an investigation of hybrid M&S studies in health care. *Journal of the Operational Research Society, 68*(10), pp. 1211-1222. doi:https://doi.org/10.1057/s41274-016-0147-6

Priore, P., Ponte, B., Puente, J., & Gómez, A. (2018) Learning-based scheduling of flexible manufacturing systems using ensemble methods, *Computers & Industrial Engineering*, 126, pp. 282-291.

Rabe, M., & Scheidler, A.A. (2014) An approach for increasing the level of accuracy in supply chain simulation by using patterns on input data, *Proceedings of the 2014 Winter Simulation Conference*, IEEE, pp. 1897-1906.

Ranyard, J.C., Fildes, R., & Hu, T-I. (2015). Reassessing the scope of OR practice: The influences of problem structuring methods and the analytics movement, *European Journal of Operational Research*, *245*(1), pp. 1-13.

Robinson, A., Levis, J., & Bennett, G. (2010). INFORMS to officially join analytics movement. *OR/MS Today, 37*(5), pp. 59.

Robinson, S. (2014). *Simulation: The Practice of Model Development and Use*, Second Edition, New York: Palgrave Macmillan.

Royston, G. (2013). Operational Research for the Real World: big questions from a small island, *Journal of the Operational Research Society*, *64*(6), pp. 793-804.

Rozinat, A., de Jong, I.S.M., Gunther, C.W., & van der Aalst, W.M.P. (2009a). Process mining applied to the test process of wafer scanners in ASML, *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, *39*(4), pp. 474-479.

Rozinat, A., Wynn, M., van der Aalst, W.M.P., ter Hofstede, A.H.M., & Fidge, C. (2009b). Workflow simulation for operational decision support, *Data and Knowledge Engineering*, *68*(9), pp. 834-850.

Sanchez, S.M. (2015). 'Simulation experiments: Better data, not just big data'. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, & M. D. Rossetti (Eds.) *Proceedings of the 2015 Winter Simulation Conference* (pp. 800-811). IEEE.

SISO (2010) SISO-STD-006-2010, *Standard for Commercial Off-The-Shelf (COTS) Simulation Package Interoperability (CSPI) Reference Models*. Orlando, FL: Simulation Interoperability Standards Organisation.

Smith, J.S., Sturrock, D.T., Kelton, W.D. (2018). *Simio and Simulation: Modeling, Analysis, Applications*, 5th Edition, Simio LLC.

Soban, D., Thornhill, D., Salunkhe, S., & Long, A. (2016). Visual analytics as an enabler for manufacturing process decision-making, *Procedia CIRP*, *56*, pp. 209-214.

Strassburger, S. (2015). HLA-based optimistic synchronisation with SLX. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, & M. D. Rossetti (Eds.) *Proceedings of the 2015 Winter Simulation Conference* (pp. 1717-1728). IEEE.

Sutton, R.S. and Barto, A.G. (2018). *Reinforcement Learning: An Introduction*, Second Edition, MIT Press.

Tao, F., Zhang, H., Liu, A. Nee, A.Y.C. (2019) Digital Twin in Industry: State-of-the-Art, *IEEE Transactions on Industrial Informatics*, 15 (4), pp. 2405-2415.

Taylor, S.J.E. (2015). The impact of big data on M&S: Do we need to get "big"? In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, & M. D. Rossetti (Eds.) *Proceedings of the 2015 Winter Simulation Conference* (pp. 3085-). IEEE.

Taylor, S.J.E. (2019). Distributed simulation: state-of-the-art and potential for operational research, *European Journal of Operational Research*, 273, pp. 1-19.

Taylor, S.J.E., Mustafee, N., Turner, S.J., Pan, K., & Strassburger, S. (2009). Commercial-Off-The-Shelf Simulation Package Interoperability: Issues and Futures, *Proceedings of the 2009 Winter Simulation Conference*, IEEE, 203-215.

Uriarte, A.G., Zúñiga, E.R., Moris, M.U., & Ng, A.H.C. (2017). How can decision makers be supported in the improvement of an emergency department? A simulation, optimization and data mining approach, *Operations Research for Health Care*, 15, pp. 102-122.

van der Aalst, W. (2016). *Process Mining: Data Science in Action*, Second Edition Berlin: Springer-Verlag.

van der Aalst, W. (2018). Process mining and simulation: A match made in heaven!, *Proceedings of the SummerSim-SCSC18 Conference*, July 9-12, Bordeaux, France, SCS, 39-50.

Vidgen, R., Shaw, S., & Grant, D.B. (2017). Management challenges in creating value from business analytics, *European Journal of Operational Research*, *261*(2), pp. 626-639.

Vieira, H., Sanchez, K., Kienitz, K.H., & Belderrain, M.C.N. (2011). Improved efficient, nearly orthogonal, nearly balanced mixed designs. In S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, & M. Fu (Eds.) *Proceedings of the 2011 Winter Simulation Conference* (pp. 3600-3611). IEEE.

Volovoi, V. (2016) Simulation of Maintenance Processes in the Big Data Era, *Proceedings of the 2016 Winter Simulation Conference*, IEEE, pp. 1872-1883.

Zhou, Z., Wang, Y., & Li, L. (2014). Process mining based modeling and analysis of workflows in clinical care – a case study in a Chicago outpatient clinic. In *Proceedings of the 11th IEEE International Conference on Networking, Sensing and Control* (pp. 590-595). IEEE.

Table 1. Inclusion and exclusion criteria for the studies reviewed.

	Inclusion Criteria	Exclusion Criteria
Population	Discrete-Event Simulation Visual Analytics Data Mining Process Mining Data Farming Machine Learning Data Analytics Big Data Analytics Business Analytics	1. *Irrelevance of content* • Other simulation methods such as agent-based simulation, system dynamics and Petri nets. • Non organisational issues such as medical applications or semiconductor design • Not an article, for example a conference program description 2. *Duplication of content* • Duplication across search databases • Duplication across definitions
Outcomes	Evidence of DES model build and results Evidence of use of analytics technique	3. *Discussion of DES and analytics but no simulation application* 4. *No evidence of use of analytics technique*
Study Design	Papers must be in English and accessible	5. *Papers not in English or not accessible*

Table 2. Examples of the use of Discrete Event Simulation with Big Data Analytics

Authors	Title	DES Software	Analytics Software (Analytics Technique)	Dataset Source (Dataset Size)
MACHINE LEARNING				
Priore et al. (2018)	Learning-based scheduling of flexible manufacturing systems using ensemble methods	Witness	RapidMiner (DT, SVM, ANN, k-NN)	**DATA FARMING** (1100 combinations of 7 control attributes)
Aqlan et al. (2017)	Integrating Data Analytics and Simulation for Defect Management in Manufacturing Environments	Arena	IBM SPSS Modeler (ANN)	Real- Not Stated (5173 defects)
Bergmann et al. (2017)	Emulation of control strategies through machine learning in manufacturing simulations	Tecnomatix Plant Simulation	MatLab (CART, NBC, k-NN, SVM, ANN)	**DATA FARMING** (5 decision rules x 10000 decisions x 4 data quality Scenarios)
Bergmann et al. (2015)	Approximation of dispatching rules for manufacturing simulation using data mining methods	Tecnomatix Plant Simulation	MatLab (CART, NBC, k-NN, SVM)	**DATA FARMING** (5 decision rules x 10000 decisions x 4 data quality Scenarios)
Bergmann et al. (2014)	On the use of artificial neural networks in simulation-based manufacturing control	Tecnomatix Plant Simulation	MatLab (ANN)	**DATA FARMING** (not stated)
Gyulai et al. (2014)	Capacity planning and resource allocation in assembly systems	Tecnomatix Plant Simulation	R (Random Forest DT)	**DATA FARMING** (not stated)

	consisting of dedicated and reconfigurable lines			
Celik et al. (2010)	DDDAS-based multi-fidelity simulation framework for supply chain systems	Arena	VBA (Bayesian Belief Network)	Real – machine sensors (real-time system)
Glowacka et al. (2009)	A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling	SimProcess	SPSS Clementine 10 (ARM)	Real – clinic records and barcodes (1809 patient visits)
DATA MINING				
Ceglowski et al. (2007)	Combining data mining and discrete-event simulation for a value-added view of a hospital emergency department	Simul8	Viscovery SOMine (SOM)	Real – patient records (56906 patients)
Kibira et al. (2015)	Integrating Data Analytics and Simulation Methods to Support Manufacturing Decision Making	Arena	Not Stated (Association Techniques)	Real- sensors, RFID. (not stated)
Uriarte et al. (2017)	How can decision makers be supported in the improvement of an emergency department? A simulation, optimisation and data mining approach	FlexSim	Not Stated (FPM)	Real – patient records (50,000 Patient visits)
VISUAL ANALYTICS				
Feldkamp et al. (2017a)	Online analysis of simulation data with stream-based data mining	Tecnomatix Plant Simulation	Massive Online Analysis Software Framework (DT)	**DATA FARMING** (0.5 million runs)
Feldkamp et al. (2017b)	Knowledge discovery and robustness analysis in manufacturing simulations	Tecnomatix Plant Simulation	Massive Online Analysis Software Framework (DT)	**DATA FARMING** (102,400 runs)
Feldkamp et al. (2016)	Knowledge discovery in simulation	Wolverine SLX	MatLab	**DATA FARMING**

	data: A case study of a gold mining facility		(Visual Data Mining)	(262,144 runs)
Feldkamp et al. (2015)	Visual analytics of manufacturing simulation data	Tecnomatix Plant Simulation	MatLab (Visual Data Mining)	**DATA FARMING** (491,160 runs)
PROCESS MINING				
Abohamad et al. (2017)	A hybrid process-mining approach for simulation modelling	Anylogic	ProM, Disco (Fuzzy Mining)	Real – Patient records (229,971 events)
Lamine et al. (2015)	Improving the management of an emergency call service by combining process mining and discrete event simulation approaches	Witness	Disco (Fuzzy Mining)	Real – emergency calls (not stated)
Zhou et al. (2014)	Process Mining Based Modeling and Analysis of Workflows in Clinical Care – A Case Study in an Chicago Outpatient Clinic	ProModel	ProM, Disco (Fuzzy Mining)	Real – Patient records (20,000 cases)

List of Figure captions

Figure 1. PRISMA flow diagram illustrating the literature review procedure.

Figure 2. Publication by Year.

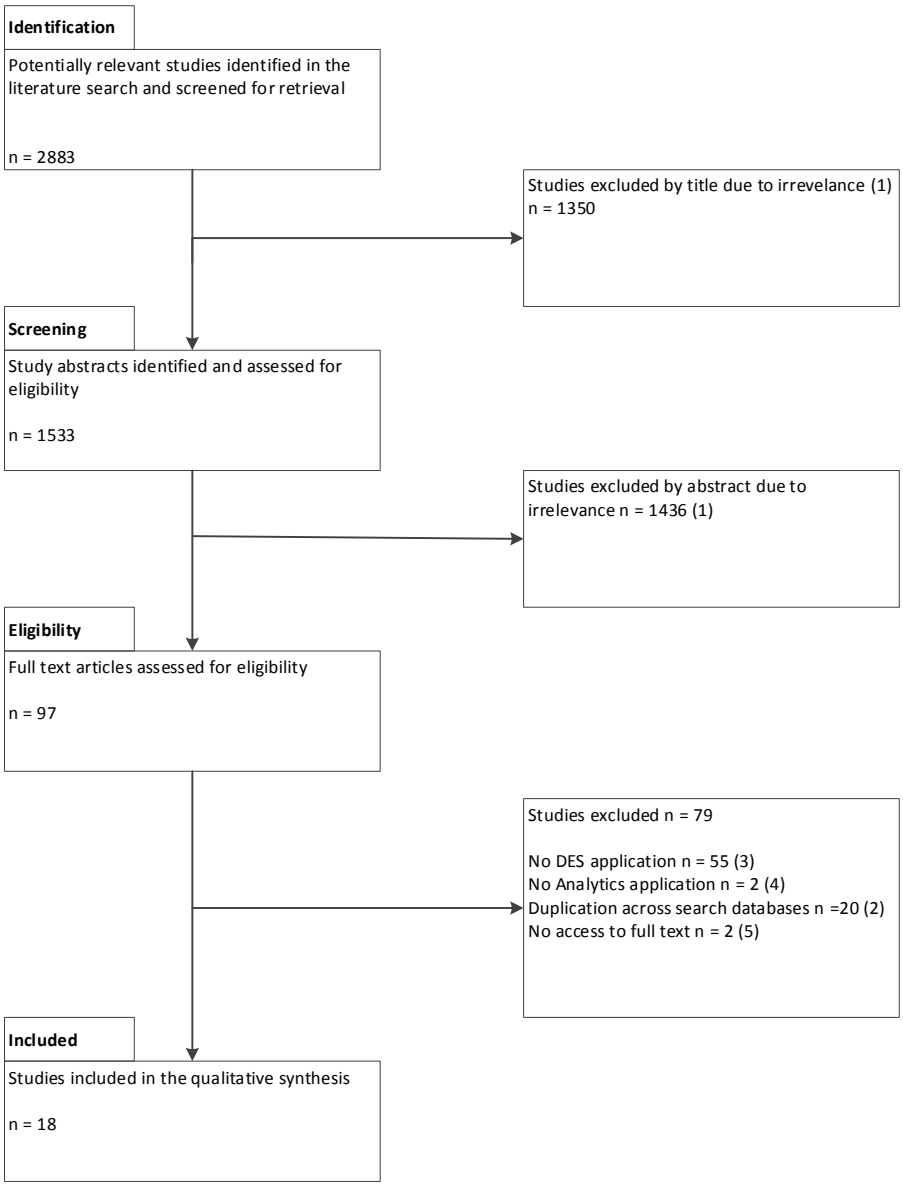
Figure 3. DES methodology

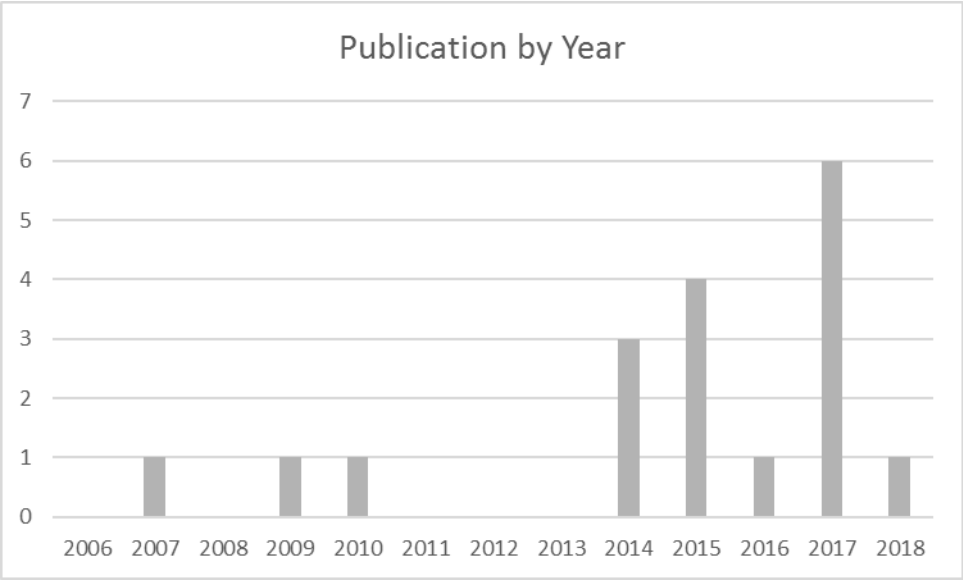
Figure 4. DES to facilitate BDA

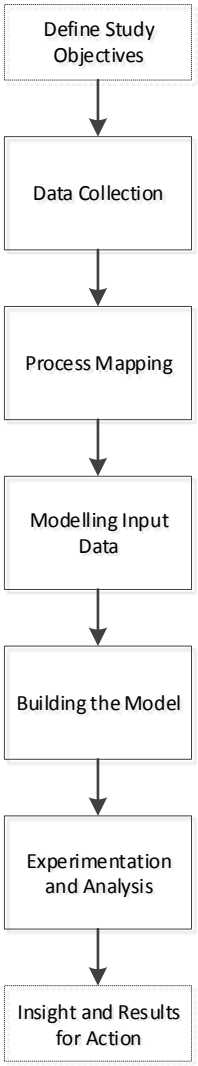
Figure 5. BDA to facilitate DES

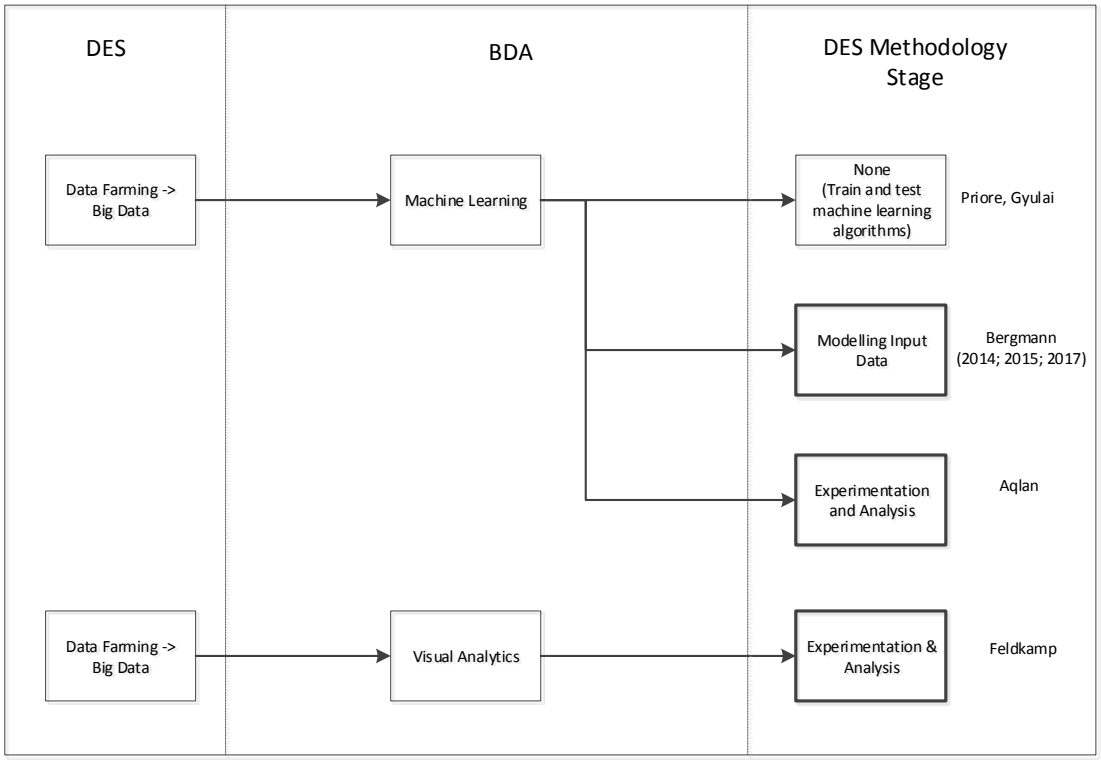
Figure 6. A methodology for the use of BDA in DES

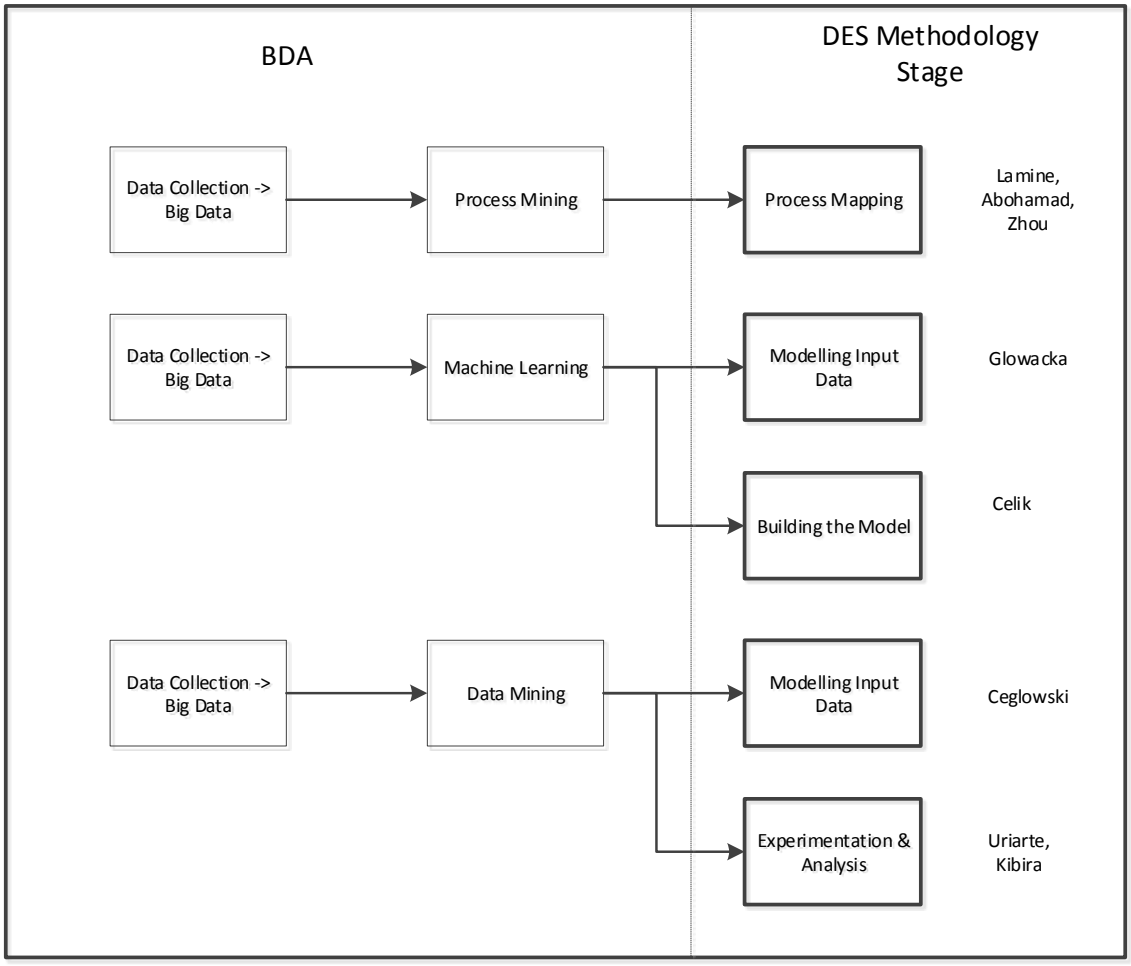
Figure 7. The relationship between simulation and analytics


Identification
Potentially relevant studies identified in the literature search and screened for retrieval
n = 2883
Studies excluded by title due to irrevelance (1)
n = 1350
Screening
Study abstracts identified and assessed for eligibility
n = 1533
Studies excluded by abstract due to irrelevance n = 1436 (1)
Eligibility
Full text articles assessed for eligibility
n = 97
Studies excluded n = 79
No DES application n = 55 (3)
No Analytics application n = 2 (4)
Duplication across search databases n =20 (2)
No access to full text n = 2 (5)
Included
Studies included in the qualitative synthesis
n = 18

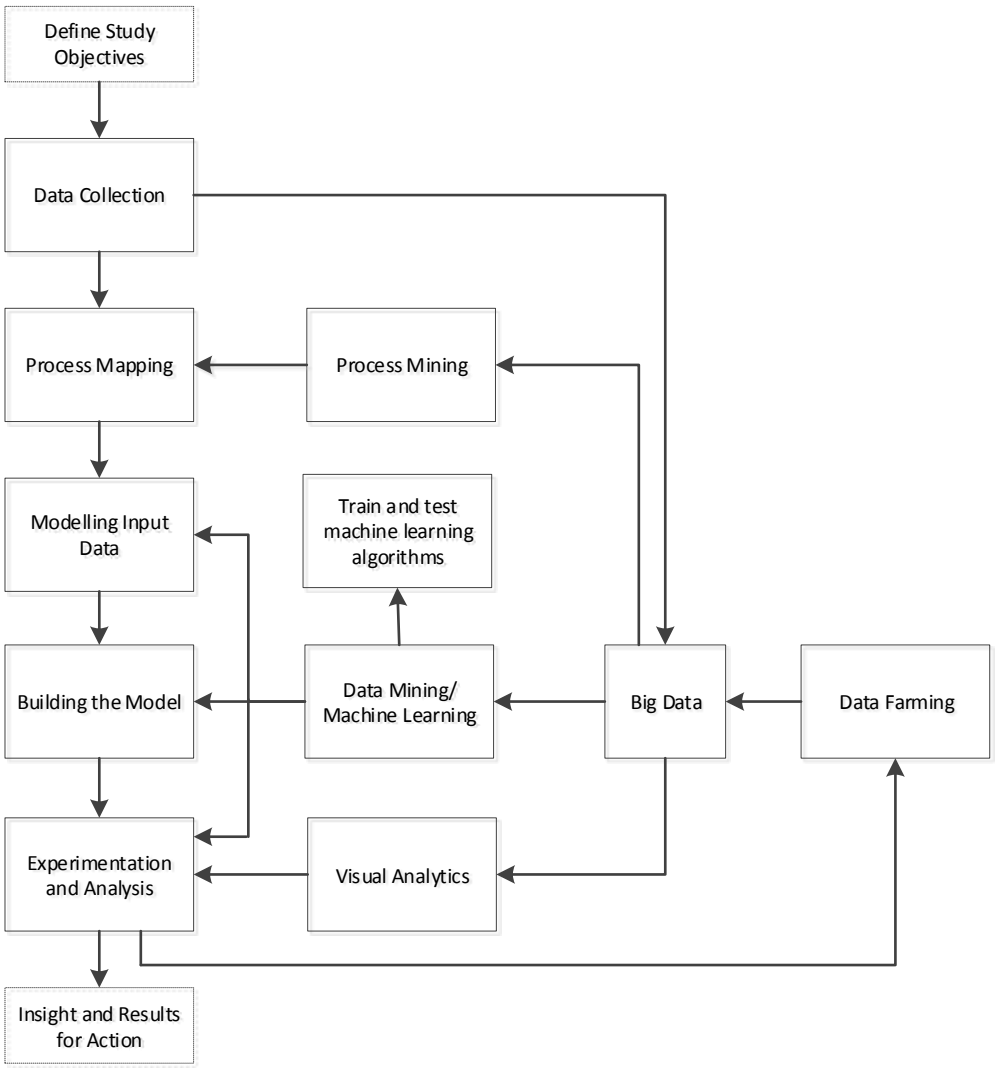

Publication by Year
7
6
5
4
3
2
1
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

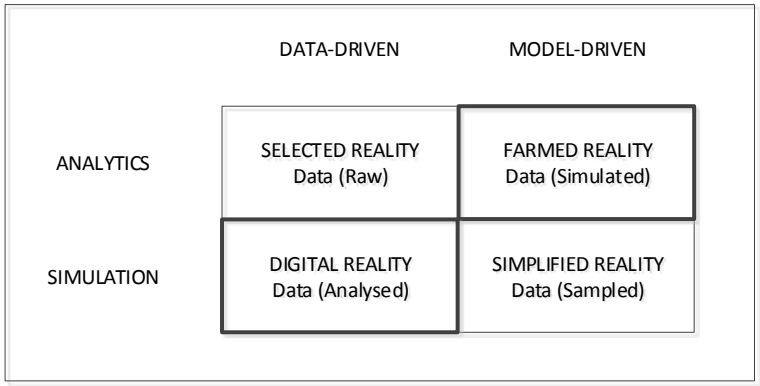

Define Study Objectives
Data Collection
Process Mapping
Modelling Input Data
Building the Model
Experimentation and Analysis
Insight and Results for Action


DES
BDA
DES Methodology Stage
Data Farming -> Big Data
Machine Learning
None (Train and test machine learning algorithms)
Priore, Gyulai
Modelling Input Data
Bergmann (2014; 2015; 2017)
Experimentation and Analysis
Aqlan
Data Farming -> Big Data
Visual Analytics
Experimentation & Analysis
Feldkamp


BDA
DES Methodology Stage
Data Collection -> Big Data
Process Mining
Process Mapping
Lamine, Abohamad, Zhou
Data Collection -> Big Data
Machine Learning
Modelling Input Data
Glowacka
Building the Model
Celik
Data Collection -> Big Data
Data Mining
Modelling Input Data
Ceglowski
Experimentation & Analysis
Uriarte, Kibira

DES Methodology Big Data Analytics




DATA-DRIVEN
MODEL-DRIVEN
ANALYTICS
SELECTED REALITY
Data (Raw)
FARMED REALITY
Data (Simulated)
SIMULATION
DIGITAL REALITY
Data (Analysed)
SIMPLIFIED REALITY
Data (Sampled)