Data & Knowledge Engineering 117 (2018) 359–372

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

# Business-driven data analytics: A conceptual modeling framework

Soroosh Nalchigar *, Eric Yu

*Department of Computer Science, University of Toronto, Toronto, Canada*

ARTICLE INFO

*Keywords:*
Conceptual modeling
Data analytics
Machine learning
Business analytics
Goal-oriented requirements engineering
Enterprise modeling

ABSTRACT

The effective development of advanced data analytics solutions requires tackling challenges such as eliciting analytical requirements, designing the machine learning solution, and ensuring the alignment between analytics initiatives and business strategies, among others. The use of conceptual modeling methods and techniques is seen to be of considerable value in overcoming such challenges. This paper proposes a modeling framework (including a set of metamodels and a set of design catalogues) for requirements analysis and design of data analytics systems. It consists of three complementary modeling views: business view, analytics design view, and data preparation view. These views are linked together to connect enterprise strategies to analytics algorithms and to data preparation activities. The framework includes a set of design catalogues that codify and represent an organized body of business analytics design knowledge. As the first attempt to validate the framework, three real-world data analytics case studies are used to illustrate the expressiveness and usability of the framework. Findings suggest that the framework provides an adequate set of concepts to support the design and implementation of analytics solutions.

## 1. Introduction

The effective design and implementation of advanced data analytics solutions has proven to be difficult. Despite the hype around data analytics, many organizations struggle to find how to use analytics to derive value and gain competitive advantage [1].

*Requirements elicitation for data analytics systems* is a complex task [2,3]. Analytics requirements are often vague and incomplete at the early phases of projects. While business stakeholders often have a clear understanding of their strategic goals (e.g., improve marketing campaigns, reduce inventory levels), they are not clear on how analytics can help them achieve those goals. This is, to a great extent, due to the huge conceptual distance between business strategies, decision processes and organizational performance on one hand, and the implementation of analytics systems in terms of databases, preprocessing activities, and machine learning algorithms on the other hand. Previous researches report that the leading barrier to using analytics techniques is the lack of understanding of how to use analytics and unlock its value to improve the business [4,5].

Moreover, *designing analytics solutions* involves making critical design decisions, taking into account quality requirements (i.e., softgoals) and tradeoffs [6]. Analytics requirements, once discovered and elaborated, must lead to solution design and implementations that includes choice of algorithms, machine learning experimentations, and data preparation activities. A large number of machine learning and data mining algorithms exist and new ones are being developed continuously. During analytics projects, one needs to make design choices such as: what are potential algorithms that can address the problem at hand? What criteria should be considered to evaluate those algorithms? What/how data should be prepared to be used by algorithms? These decisions have important implications in several aspects of the eventual analytics solution, such as scalability, understandability, tolerance to noisy data and missing values.

* Corresponding author. Department of Computer Science, University of Toronto, 40 St George Street, Toronto, Ontario M5S 2E4, Canada.
*E-mail addresses:* soroosh@cs.toronto.edu (S. Nalchigar), eric@cs.toronto.edu (E. Yu).

http://dx.doi.org/10.1016/j.datak.2018.04.006

Available online 16 April 2018

On the other hand, *aligning analytics with business strategies* is critical for achieving value through analytics [4,7]. Lack of this alignment can result in unclear expectations of how analytics contribute to business strategies, lack of executive sponsorship, and analytics project failures. Such alignment requires determining what objectives the business is trying to accomplish with the analytics initiatives, how to allocate resources, and what kinds of data assets to focus on [8]. It is important for organizations to discover, justify, and establish why there is a need for the organization to conduct analytics initiatives. Towards this end, discovering the business goals and translating them into analytics goals is a critical step [9,10].

Another challenge in realizing value from analytics is *a shortage of talent with deep expertise in statistics and machine learning* [5,11]. The rapid advance of the machine learning domain, as well as in platforms for handling large datasets add to this challenge by making the design of analytics solutions more difficult. Effective use of advanced analytics requires executives and stakeholders to know what machine learning can do and how to use insights derived from data to manage the enterprise [12].

Conceptual modeling as a field of study is concerned with defining formal and suitable forms of higher abstraction of the application domain aiming to support effective and efficient development of information systems [13]. The use of conceptual models is seen to be of considerable value in overcoming the challenges mentioned above, and hence in achieving effective design and implementation of advanced analytics solutions [14]. Conceptual modeling techniques can be used to systematically reveal use cases for big data analytics towards achieving strategic goals. They can be used by businesses to understand where to start, what data to focus on and how to derive business values through analytics initiatives. They can be used to design analytics solutions and to support algorithm selection. In addition, they can provide analysis support for ensuring the alignment and understanding the impact of analytics to business. The models inherently capture the design of analytics solutions for addressing business needs, and hence encode the otherwise implicit knowledge on how to use analytics. Importantly, conceptual models can potentially facilitate the communication between stakeholders and analytics developers (i.e., data scientists) that is critical for project success - by bridging the gap between business strategies, machine learning algorithms and data assets.

This paper presents a conceptual modeling framework for developing advanced business analytics guided by business objectives. The framework includes three complementary modeling views. (i) The *Business View* represents an enterprise in terms of strategies, actors, decisions, analytics questions, and required insights. This view is used to systematically elicit and clarify analytics requirements and to inform the types of analytics that the user needs. (ii) The *Analytics Design View* represents the core design of an analytics system in terms of analytical goals, machine learning algorithms, quality requirements (i.e., softgoals), and performance metrics. This view identifies design tradeoffs, captures the experiments (to be) performed with a range of algorithms, and supports algorithm selection. (iii) The *Data Preparation View* represents data preparation processes in terms of mechanisms, algorithms and preparation tasks. This view expresses the structure and content of data sources and the design of data preparation tasks. The three views are used together to link enterprise strategies to analytics algorithms and data stores and preparation activities.

An important component of the framework is a set of design catalogues. The catalogues codify, organize, and express analytics design knowledge and expertise in terms of conceptual models. They provide generic and repeatable designs and solutions for commonly known and recurring analytics problems. Three kinds of catalogues are distinguished in the framework, each kind corresponding to a modeling view: (i) the *Business Questions Catalogue* provides a categorization and a large sample of business questions that can be addressed by analytics solutions. It provides knowledge on what types of business questions are answerable by what types of analytics techniques. (ii) The *Algorithms Catalogue* provides machine learning know-how to support design of the solution in the second modeling view. It provides knowledge on when to use what algorithm(s) and how to compare and evaluate them. (iii) The *Data Preparation Catalogue* provides data preprocessing know-how to support the design of preparation workflows in the third modeling view. It provides knowledge on how to prepare and clean raw datasets.

This paper is organized as follows. Section 2 presents an illustration of the proposed framework in a real analytics project. Section 3 introduces primitive concepts and presents metamodels. Section 4 offers three kinds of analytics design catalogues. Section 5 discusses findings from applying the framework in three data analytics case studies. Section 6 reviews related work and highlights the contribution. Section 7 concludes the paper with directions for future work and discussion of threats to validity.

## 2. An illustration

We illustrate the framework using a project aimed at developing an analytics system to predict upcoming software system outages. The company has around 300 globally accessible software applications hosted in its data centers across the world. Software system outages are costly and predicting them can enable preventive maintenance activities.

Fig. 1 illustrates the *Business View* for the software outage prediction project. The purpose of this view is to represent the analytics needs of an organization and to ensure that those needs are driven by organizational strategies and the decisions that the organization faces. This view models the business analytics requirements in terms of its *Strategic Goals*, *Indicators*, *Actors*, *Decision Goals*, *Question Goals*, and *Insights*.

The model in Fig. 1 shows that, at the very top level, the **Executive board** of the company aims to **Maximize stakeholder value**, and several indicators such as **Return on equity (ROE)** and **Profit margin** are used to evaluate the achievement of this business goal. In order to **Maximize stakeholder value**, the company aims to **Improve asset efficiency**, **Improve revenue growth**, and **Improve operating margin**, shown via the **AND** decomposition links. At lower levels, the model shows that **Improve maintenance of IT systems** is another goal of the company, where **Mean time between failures** and **Uptime (%)** are corresponding indicators. Business goals are refined into lower level goals and eventually into decision goals. Decision on software outage prevention is an example of a decision goal. The model indicates that in order to **Prevent software outages**, the corresponding actor needs to decide on how to prevent a software from failing. **Decision on software repair** is another example of a decision goal. The models shows that in order to

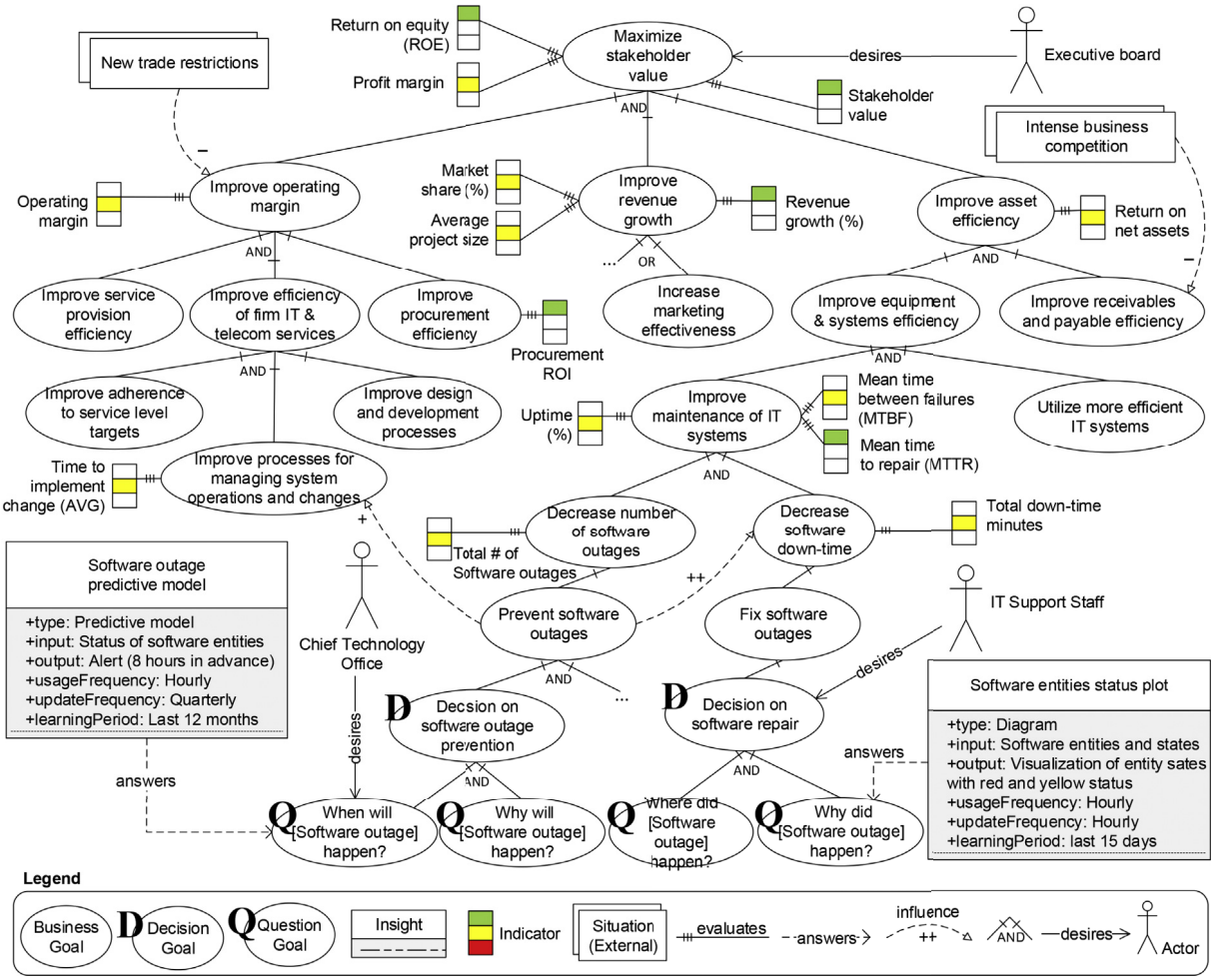


**Fig. 1.** Business View for the software outage prediction project (partial). This model is constructed based on interviews with domain experts, review of reporting dashboards and metrics in place, supplemented with some assumptions.

**Fix software outage**, the **IT Support Staff** need to decide how to repair the software outage. Decision goals are further decomposed into question goals, representing what it is it that the actor needs to know to make the decision. **When will [Software outage] happen?** is an example of a question goal. The model depicts that in order to make the **Decision on software outage prevention**, the **Chief Technology Office** needs to know when a software outage will happen in the near future. **Why did [Software outage] happen?** is another example of a question goal. The models shows that in order to make the **Decision on software repair** one needs to know why an outage happened.

Question goals are answered by insights. **Software outage predictive model** is an example of an insight to be generated by the intended analytics solution. It is a **Predictive model** that, in runtime, receives data on **Status of software entities** as input and generates **Alerts (8 h in advance)** before an upcoming outage. Such a predictive model would be used on an **Hourly** basis and it is retrained every **Quarter** (to be updated on new outage patterns) on a dataset of **Last 12 months**. More examples of each modeling concept can be found in Fig. 1.

By modeling *Decision Goals*, this view represents the decision areas that need support from analytics insights. It ensures the link between analytics, organizational decision processes, and strategic goals. This concept also facilitates linking and turning analytics-driven insights into actions, resulting from decisions. Through the *Question Goals*, the framework captures the business needs that the analytics solution is intended to address. The Business Questions Catalogue (introduced in Section 4) can be used while performing modeling activities in this view. Eliciting the questions at the early phases of analytics results in performing the right analysis for the right user. Later in the analytics process and once the findings are generated, the questions can also facilitate the process of interpreting and framing the findings. By modeling *Insights*, this view represents the knowledge that is extracted from data for answering the questions. The insight elements link the Business View to the Analytics Design View.

Fig. 2 illustrates the *Analytics Design View* for the software outage prediction project. The purpose of this view is to represent design alternatives for the analytics system, show tradeoffs and support algorithm selection. This view expresses the design of an analytics

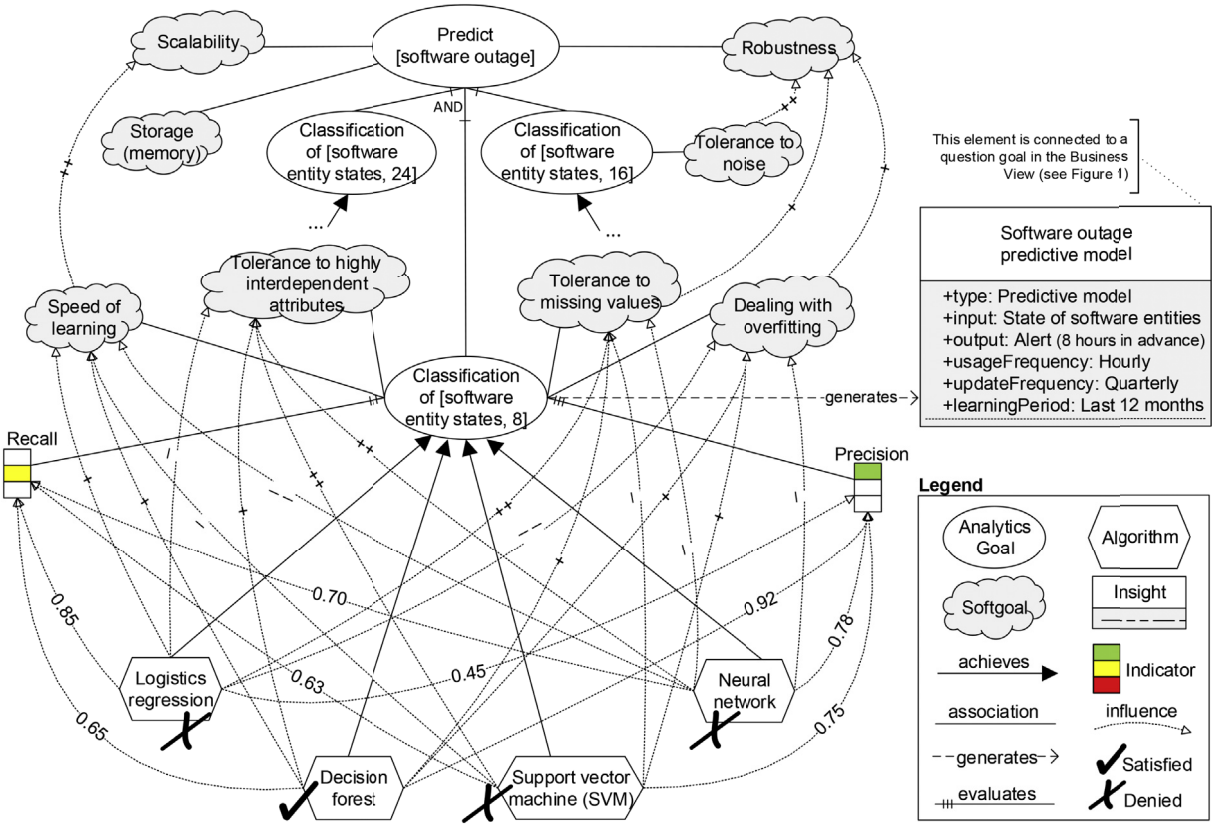


**Fig. 2.** Analytics Design View for software outage prediction project (partial).

solution in terms of *Analytics Goals*, (machine learning) *Algorithms*, *Softgoals*, *Influences*, and *Indicators*.

In Fig. 2, **Predict software outage** is an example of an analytics goal. To achieve this goal, the system needs to achieve the **Classification of software entity states** goals. In this project, there were three classification goals that together addressed prediction periods of 8, 16, 24 h. Each of those classification goals **generates** a different instance of the insight element. The model shows that **Logistics regression, Decision forest, Support vector machine (SVM)**, and **Neural networks** are alternative algorithms that perform classification. Moreover, the model represents the contributions from algorithms towards indicators and softgoals, i.e., how well the analytics goals are achieved. For example, the influence link from **Decision Forest** algorithm towards **Precision** means that during experiments, the algorithm resulted in the value of **0.92** for the **Precision** measure. Also, this algorithm has **some (+)** contribution towards the **Speed of learning**. In addition, the model shows the influences among softgoals. For example, it shows that **Speed of learning** (to train the classifier) has a **make (++)** influence on Scalability which itself is a softgoal requirement for the **Predict software outage** goal. By capturing these, the view describes tradeoffs and supports algorithm selection during the design of analytics systems. Systematic reasoning procedures (such as those in Ref. [15]) can be applied here to analyze satisfiability of softgoals during early phases of design and shorten the required machine learning experiments. In the first case study, the **Precision** indicator had the highest priority[1] which justified the choice of **Decision forest** for the corresponding classification goal. The algorithms catalogue (introduced in Section 4) assists users in this modeling view and supports the designing process. This view is linked to the Business View via the *generates* link from analytics goals to insights.

Fig. 3 illustrates the *Data Preparation View* for the software outage prediction project. The purpose of this view is to support the design and documentation of data preparation workflows. This view models the data preparation aspect of the analytics in terms of *Mechanisms*, *Algorithms*, *Preparation Tasks*, *Data Flows,* and *Entities*.

The model in Fig. 3 shows the content and structure of data sources. The company has a cross-platform data center management system that logs computer systems operations. The model shows that a software **Application** is related to many **Assets** (e.g., web servers) and each asset in turn can have many **Managed Entities**. A managed entity is an object of interest (e.g., a disk) within an asset whose health and status is being monitored. The **State** data captures the status of the managed entities at every millisecond using some

[1] In classification tasks, a low precision score indicates a large number of false positives (i.e., false alerts) by the predictive model. In this project, the stakeholder (Chief Technology Office) required the analytics team to come up with a solution with lowest possible false positives and hence priority was given to the precision indicator.

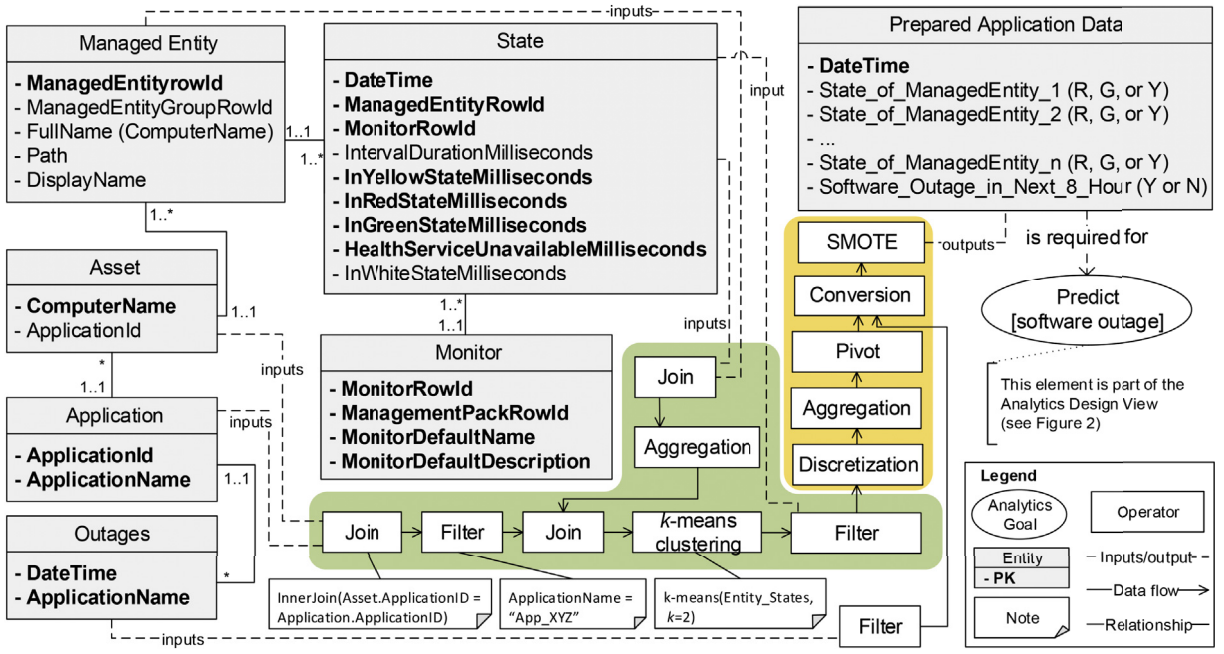


**Fig. 3.** Data Preparation View for software failure prediction project (partial).

**Monitors**.

Given such a schema, the model also shows the sequence of data preparation mechanisms and algorithms that were executed in this project. **Join, Filter**, and **Aggregation** are examples of mechanisms used for data preparation. **SMOTE**[2] is an example of an algorithm that was used in this case for data preparation purposes. A set of mechanisms and algorithms together form a data preparation task. In Fig. 3, the green-shaded area shows a **Data numerosity reduction task**. This task is responsible for removing managed entities whose **State** data is not showing any meaningful relationship with software outage. The ***k*-means clustering** is an example of an algorithm which, in this case, performs the main part of the data reduction task. The yellow-shaded area shows an example of a **Data transformation task** that changes the shape of the data thorough **Discretization** and **Pivot** mechanisms such that it is ready for mining by classification algorithms.

The main outcome of the workflows is the **Prepared Application Data** table that **is required for** the analytics goal of **Predict software outage**. The data attribute **Software Outage in Next 8 Hour (Y or N)** is the binary target attribute (class label) to be used for training and testing the algorithms. The data preparation catalogue (introduced in Section 4) assists users in this modeling view and supports the design of the data preparation workflows. This view is connected to the Analytics Design View through the *is required for* links from entities to analytics goals.

## 3. Metamodels

This section presents metamodels that capture the semantics and formal relationships of primitive concepts in the three modeling views. The rationale for designing the framework in terms of the three views is to enable its users to separate and comprehend elements in several areas of concerns, each one focusing on a specific aspect of the analytics system [16]. Each modeling view captures analytics application domain form a different perspective, which requires a different source of knowledge and information to construct the artifacts, leading to different kinds of usages, analyses and users. The three modeling views, while having different focuses and serving different purposes, are linked to each other and bridge the gap between strategic goals, machine learning algorithms, and data tables.

### 3.1. Business view

This view aims to (i) support the elicitation and clarification of analytics requirements, (ii) enable analysis of those requirements (e.g., decomposition, refinement, prioritization), and (iii) ensure the alignment of analytics requirements to business strategies. The main modeling elements are *Business Goals, Actors, Decision Goals, Question Goals, Insights, Indicators, Influences, and Situations*. Fig. 4 shows metamodel of the Business View in terms of a UML class diagram. Concepts in the gray-shaded area are adopted from the Business

[2] SMOTE (Synthetic Minority Over-sampling Technique) is an algorithm used for oversampling the imbalanced datasets (given that software outage instances were rare in the training dataset) for classification purposes.

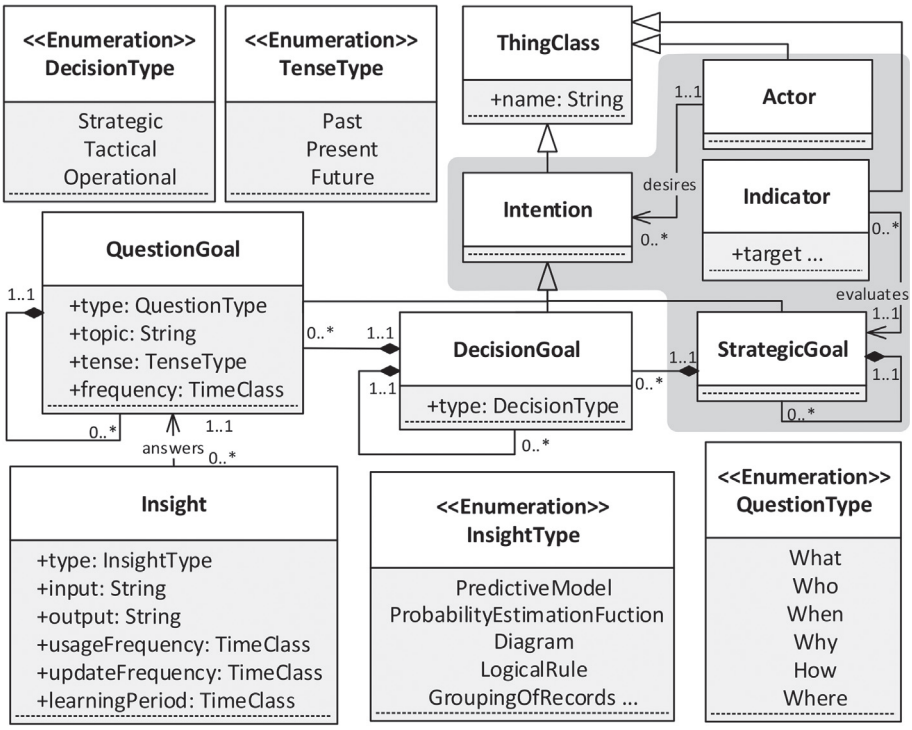


**Fig. 4.** Part of the metamodel for the Business View.

Intelligence Model (BIM) [17,18]. Here we explain concepts that are added to extend BIM.

**Decision Goals.** This concept represents intention of an actor for taking actions towards achieving strategic goals. Strategic goals can be decomposed into one or more decision goals. Modeling decision goals in this view supports discovering what decisions in the context need support from analytic initiatives that is perceived as critical during analytics projects [8,19,20].

**Question Goals.** This concept represents the desire of an actor for understanding or knowing something that is required for making decisions (i.e., achieving decisions goals). It captures the "needs-to-know" of an actor. Decision goals are decomposed into one or more questions. Questions can be refined into one or more questions. Modeling question goals in this view helps in formulating and confirming the right business questions which is critical for performing the right analytics and eventually project success [3,7,21].

Question goals are analyzed into a *Type* and *Topic* as in the NFR framework [22], and also *Tense* (see the metamodel in Fig. 4). The question type denotes the question phrase (e.g., **When** in Fig. 1), while the question topic denotes the subject and focus of the (intended) analysis (e.g. **[Software outage]** ) in Fig. 1). The question tense captures the time horizon that a question goal addresses. Elicitation of question type and tense together allows specifying what kinds of analytics and machine learning algorithms are required as part of the intended system. Moreover, identification of topic allows specifying what kind of data (or what parts of which databases) the intended analytics system will use for mining. In addition, as shown in Fig. 4, question goals are specified in terms of their *Frequency*. This attribute captures time scales and frequencies that the corresponding question is being raised. High frequency analytics question have more potential to be embedded into automated analytics systems and tools [23].

**Insights**.[3] This concept represents a structured (machine) understandable pattern (i.e., relationship among data), that is extracted from data by applying analytics algorithms. It represents a piece of information/knowledge that (partially) answers a question goal, and thereafter facilitates decision making and contributes to strategic goals. Insights are modeled in terms of *Type*, *Input*, *Output*, *Usage Frequency*, *Update Frequency*, and *Learning Period*. The type attribute specifies the actual outcome of the (machine learning) algorithm [24]: *Predictive Model*, *Probability Estimation Function*, *Logical Rule* (e.g., association rules), *Groupings of Records* (e.g., clusters), and *Diagrams* (e.g., trees, correlation heatmap, plots). By defining the type, the project team reveals the (group of) analytics techniques that are applicable for the problem at hand. Input, output, usage frequency, update frequency, and learning period attributes further characterize how the insight would be used and updated at runtime (see examples in Fig. 1). This concept connects to question goals through the *answers* links. It represents the immediate output of the data analytics activities. Multiple insights can be connected to a question goal.

[3] The term insight here refers to a (machine learning) model or representation, mined from data, that serves as a function in runtime for answering questions.

### 3.2. Analytics design view

This view aims to (i) capture alternate approaches for the analytics problem at hand, (ii) facilitate design of (machine learning) experiments and making trade-offs among quality requirements, and (iii) support algorithm selection and monitoring their performance over time. The main modeling elements are *Analytics Goals*, *Algorithms*, *Softgoals*, *Influences*, and *Indicators*.

**Analytics Goals.** This concept (see the metamodel in Fig. 5) represents the top-goal of the data analytics system, i.e., to extract insight from data. There are three types of analytics goals. *Prediction Goal* represents an intention to predict value of a target data attribute (i.e., label attribute) by using other existing attributes in the dataset. It shows the desire to find the relationship between the target feature and other existing features in the dataset. Two subtypes of this concept are *Classification* (predicts categorical values) and *Numeric Prediction*. *Description Goal* represents an intention to summarize and describe the dataset and includes two subtypes: *Clustering* and *Pattern Discovery*. *Prescription Goal* represents an intention to find the optimal alternative among a set of potential alternatives. *Optimization* and *Simulation* are subtypes of prescription goals.

**Algorithms.** This concept represents a procedure or task that addresses an analytics goal. An algorithm is a set of steps that are necessary for an analytics goal to be achieved. It is a way through which insight is extracted from data in order to satisfy an analytics goal. This concept is connected to analytics goal through the *achieves* link, representing a means-ends relationship.

**Indicators and Softgoals.** Indicators [17] are numeric metrics that measure performance with regard to some goal (analytics goal in this modeling view). Softgoals [25] capture qualities that should sufficiently hold when performing analytics (see examples in Fig. 2). Algorithms connect to indicators and softgoals through the *influence* links. Influence links that are directed towards an indicator, can be labeled with the corresponding numeric value (resulting from experiments). Contributions that are directed towards qualities can range from positive to negative, following *i** guidelines [25].

Analytics projects involve experimenting with different algorithms. During design time, indicators and softgoals represent criteria to be considered for comparison and evaluation of alternative algorithms that perform the analytics task at hand. They can be used to reduce the domain of experiments. During runtime they can be used for monitoring the performance of the running analytics system. The Analytics Design View is linked to Business View via the *generates* links (see Figs. 2 and 5).

### 3.3. Data preparation view

This view aims to (i) support share and reuse of prepared data assets, (ii) increase data awareness among analytics users, and (iii) speed up the data understanding phase by providing a reference for developers on data preparation activities. The main modeling elements are *Mechanisms*, *Algorithms*, *Preparation Tasks*, *Data Flows,* and *Entities*.

**Data Preparation Tasks.** This concept (see the metamodel in Fig. 6) represents the general task of preparing the data that is required for achieving some analytics goal. A data preparation task consists of one or more *Operator(s)*. It has four subtypes [26]: *Data Reduction* task generates a data set that is smaller in size than the input data set and yet produces the same analytical results (i.e., serves the same analytical goals). *Data Numerosity Reduction* (see an example in Fig. 3) and *Data Dimensionality Reduction* are two types of data reduction tasks. Data Cleaning represents the tasks that remove errors from the input dataset and also treat missing values in it. *Clean Missing Value* and *Clean Noisy Attribute* are subtypes of this concept. *Data Transformation* transforms the shape of data in a way that is more appropriate for analytics algorithms to mine and find patterns (see an example of this concept in Fig. 3). *Data Normalization* and *Data Discretization* are subtypes of this concept. *Data Integration* merges data from different data sources.

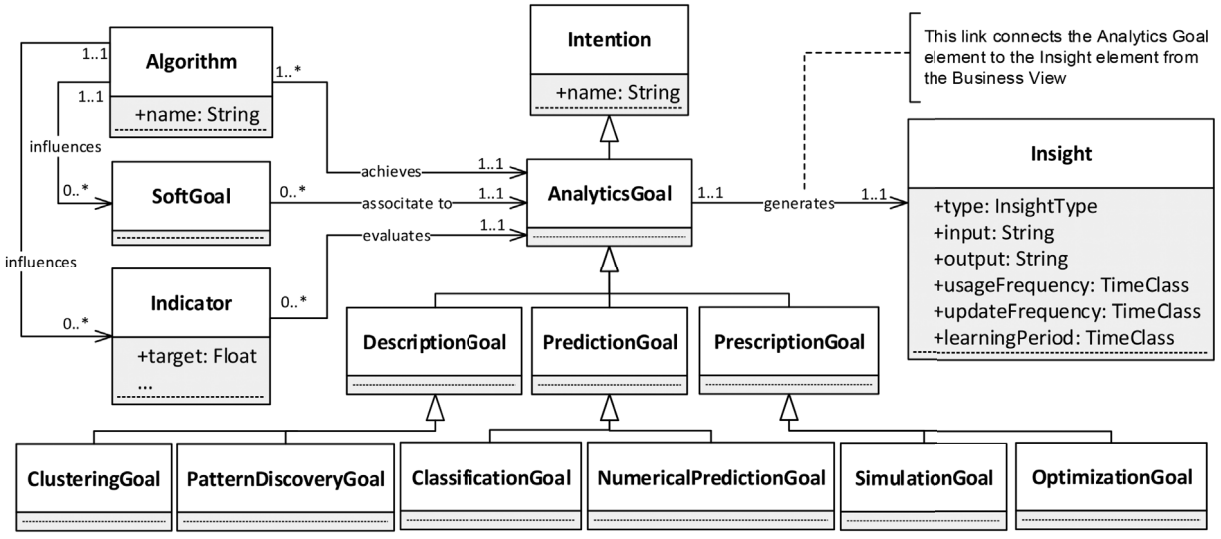


**Fig. 5.** Part of the metamodel for the Analytics Design View.

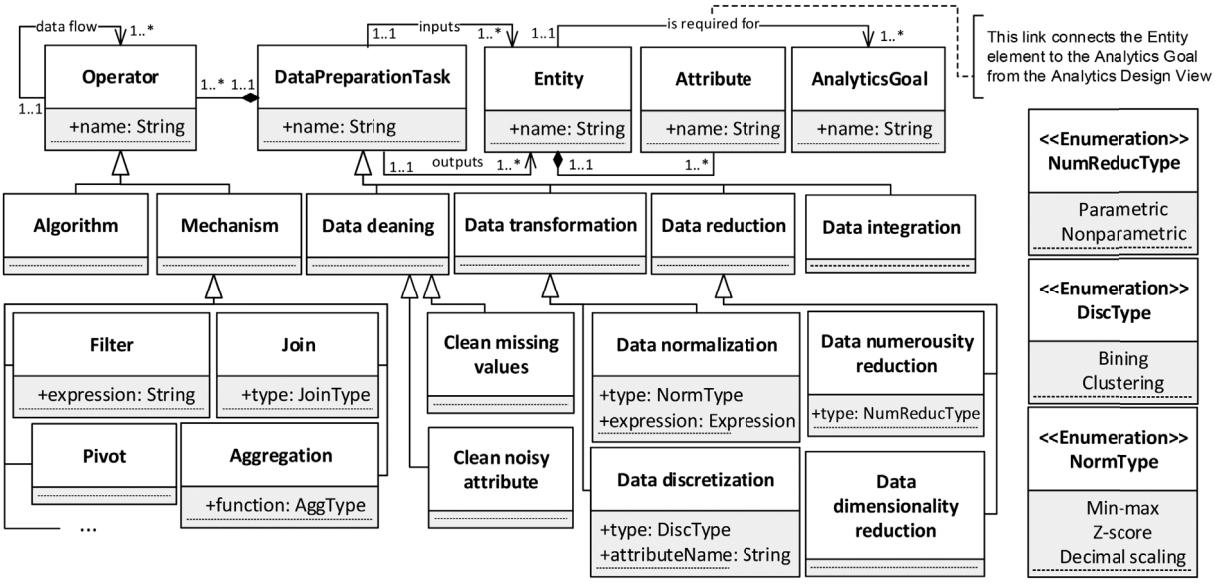


**Fig. 6.** Part of the metamodel for the Data Preparation View.

**Operator.** It represents an atomic activity that performs (part of) a data preparation task. Operators are linked by *Data Flows* to represent the sequence and dependency. There are two types of operators. *Mechanism* represents fundamental data preparation operations such as *Join* and *Filter* [27,28]. *Algorithm* is identical with algorithm in the previous view. In the data preparation view, this concept captures situations where machine learning algorithms are used for preparing data, and not for performing the actual analytics task.

The Data Preparation View is connected to the Analytics Design View via the *is required for* links (see Figs. 3 and 6).

## 4. Cataloguing analytics design knowledge

The proposed framework includes three kinds of design catalogues. The catalogues bring relevant analytics knowledge to the attention of the project team for use and re-use during the design and development process. They encode proven solution designs to common and recurring analytics problems in business domains. They provide an organized body of analytics design knowledge, accumulated from surveys (e.g., [29,30]), textbooks (e.g., [26,31]), formal ontologies (e.g., [24,32]), and project experiences.

**Business Questions Catalogue.** This catalogue represents knowledge about the types of question goals, and their associated analytics types. It categorizes question goals based on their *Type* and *Tense* (see Section 3.1) and associates each category with relevant analytics goal(s). Table 1 presents the high level schema of the catalogue. This catalogue is populated with a wide collection of instances for each category of questions goals. For example, the question goal of **Who (customers) will terminate the subscription?** belongs to the **Who** and **Future** category in Table 1 **(Who will be involved in it?)**, and can be addressed by **Prediction** type of analytics techniques. As another example, the question goal of **When will software outage happen?** from Fig. 1, belongs to the **When will it happen?** category in Table 1. This catalogue can be used by analytics team and stakeholders during the modeling activities of Business View. It can facilitate the elicitation of analytics requirements (i.e., needs to know) by suggesting and refining question goals. It also guides users to the kinds of analytics (descriptive, predictive, or prescriptive) that can address their needs.

The examples in Table 1 are presented selectively and belong to different industries and domains. In future, such catalogues can become domain specific per industry serving as reference models for performing analytics.

**Algorithms Catalogue.** This catalogue systematically organizes machine learning algorithms that are available for addressing different types of analytics goals. The catalogue provides existing metrics to be taken into account while comparing and evaluating performances of different algorithms. It also presents critical softgoals that need to be taken into account while developing analytics solutions. In addition, it encodes the knowledge on how each algorithm is known to perform with regard to different softgoals (influence links). A portion of this catalogue is illustrated in Fig. 7. As an example, it shows that ***k*-Nearest neighbor** is an algorithm that performs **Classification** and its performance can be evaluated using the **Recall** and **Precision** metrics, among others. Through the influence links, the catalogue provides the knowledge that this algorithm has **some (+)** influence on the softgoal **Dealing with overfitting**.

The *Context* semantics from Ref. [33] are used to associate context with machine learning algorithms. In this way, the catalogue represents when certain machine learning algorithms are known to perform well based on a collection of previous evidences and experiments in the literature or relevant sources. This can guide the decision on which algorithms are more appropriate for the analytics goal and shorten the experimentation phase of the projects. In Fig. 7, the context **C1** shows that the **Classification** goal is activated when

**Table 1**
The structure of the Business Questions Catalogue. For each pair of question Type and Tense, the general form of questions under that category is given along with two illustrative examples. Each instance is mapped to its relevant Analytics Goal.

| Question Type | Question Tense | | |
|---|---|---|---|
| | **Past** | **Present** | **Future** |
| **What** | **What happened?** | **What is happening?** | **What will happen?** |
| | What was total sales amount for last year? *<br>What promotion channel had highest click rate? * | What products are often purchased together? *<br>What is the optimal order size for each product? ➳ | What will be return to equity (ROE) in next quarter? ★<br>What will be the next purchase of a given user? ★ |
| **Who** | **Who was involved in it?** | **Who is involved in it?** | **Who will be involved in it?** |
| | Who were the most impulsive buyers? *<br>Who were promoted in last fiscal year? * | Who are the most active online users? *<br>Who (employees) are top performers? * | Who (customers) will terminate the subscription? ★<br>Who will be clicking on the marketing email link? ★ |
| **When** | **When did it happen?** | **Is it happening now?** | **When will it happen?** |
| | When (season) did we have maximum sales? *<br>When most employees left the firm? * | Is the current credit card transaction a fraud? ★<br>Is the current user review a negative sentiment? ★ | When will software outage happen? ★<br>When will each user most probably open the app? ★ |
| **Where** | **Where did it happen?** | **Where is it happening?** | **Where will it happen?** |
| | Where (warehouses) did we have minimum waste amount? *Where (store locations) had maximum sale? * | Where (geospatial points) have the most rainfall? *Where (city areas) are similar in housing values? * | Where (province) will have maximum online visit? ★Where will most likely each customer group shop? ★ |
| **Why** | **Why did it happen?** | **Why is it happening?** | **Why will it happen?** |
| | Why store sales were below the target?<br>Why did the marketing campaign perform well? | Why is there a decreasing trend in website traffic? *Why visit session ends after a certain click? * | Why will a given product be of interest to a customer? ★<br>Why will a certain customer churn? ★ |
| **How** | **How did it happen?** | **How is it happening?** | **How will it happen?** |
| | How was the overall ratings in user reviews? *<br>How effective was the new website map? * | How frequent on average a user visit the website? *<br>How is the new promotion impacting total sales? * | How long will the current visit rate Continue to grow? ★<br>How will new supply policy impact product levels? ➳ |

Symbols *, ★, and ➳ refer to the analytics goals of Description, Prediction, and Prescription respectively.

**Target attribute type (the value to be predicted) is categorical**. On the other hand, **C4** shows that **ARIMA**[4] can be used for **Numeric prediction**, when **there is a long and stable time series dataset**. Due to space limitations, not all the contexts are given in Fig. 7. Also, the metamodel for this catalogue is not discussed in this paper.

**Data Preparation Techniques Catalogue.** This catalogue captures knowledge on available methods for different types of data preparation tasks. It makes use of the same modeling elements as in the algorithm catalogue. As shown in Fig. 8, **Using median** is a method for **Cleaning missing values** when the corresponding **Attribute has a skewed distribution**. Also, **Data normalization** is a type of **Data transformation** that can be performed **When using a distance-based mining algorithm (e.g., k-nearest neighbor)**. Analytics development team can browse through this catalogue, decide on data preparation techniques and design data preparation workflows.

## 5. Case studies

The proposed framework has been applied to three analytics cases. As our first attempt to validate the framework, our main focus in these three case studies was to examine the expressiveness aspect of the framework. In particular, we were interested to see if the framework can represent analytical requirements and solution designs in some real world contexts. The first two case studies were reconstructions of completed projects. The third case study was an application of the framework to an on-going analytics project. These cases together can be seen as demonstration and evaluation activities in a design science research cycle [34].

The first project focused on using advanced analytics to predict upcoming software system outages (see Section 2 for instances of models). The second project focused on finance analytics. The purpose of this project was to predict an upcoming event regarding financial metrics in the company's network. The third project focused on search engine analytics. The purpose of this project was to use analytics to provide query suggestions to online users within the company's internal search system.

In Section 2, we used the first case study for illustrating the modeling views, examples of constructs in each view, and their expressiveness. Similarly, the three modeling views were instantiated for the second case study and instances of models were presented to and understood by several stakeholders (after the project was completed). Our main finding from these cases is that the modeling

[4] Autoregressive Integrated Moving Average (ARIMA).

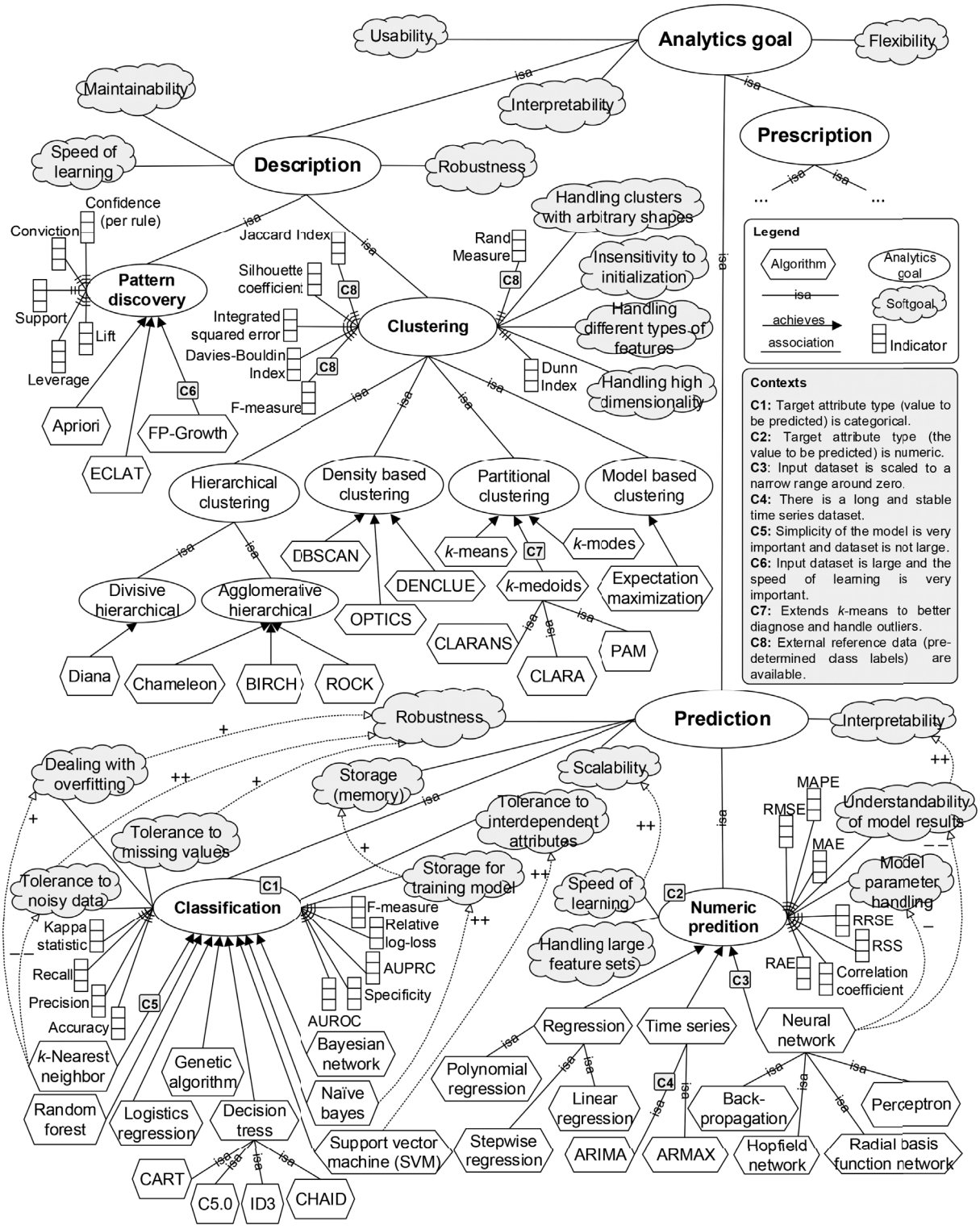


**Fig. 7.** A portion of the Algorithms Catalogue. Softgoals and their influences are shown selectively to keep the model readable.

views together provide an adequate set of concepts for connecting strategic goals to analytics algorithms and data preparation activities. We observed that the framework can be used for representing analytics requirements, can show design tradeoffs and support algorithm selection, can capture data preparation activities, and can represent the alignment between analytics systems and business strategies.

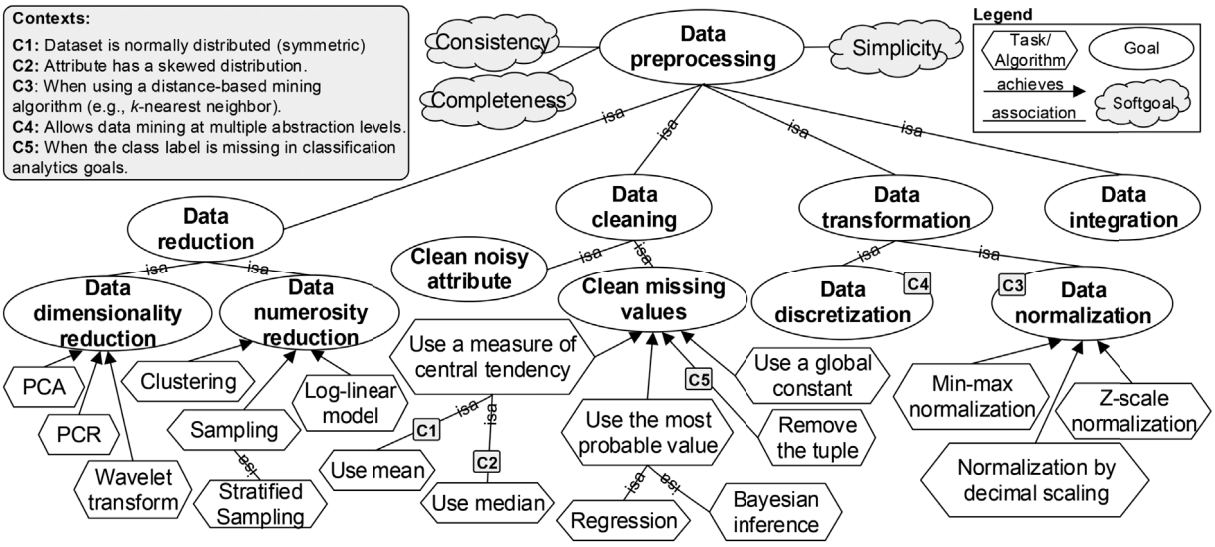


**Fig. 8.** A portion of Data Preparation Catalogue. Not all contexts are shown here.

The limitation in these two case studies was that the modeling framework was applied after the original projects had completed. Nevertheless, validation of the models by the original stakeholders suggested that the framework has the potential to guide project activities.

In the third case, we were able to confirm the ability of the framework to guide an analytics project. Fig. 9 shows a fragment of the business view model that was constructed in collaboration with stakeholders, at the requirements elicitation phase of the project. While at the beginning the focus of the project was broad and imprecise (to use analytics for improving users' search experience), the models effectively helped the team to narrow down the scope and reach an agreement about the "to-be" analytics system (to use analytics to provide query suggestions). The identification of decisions goals and refining them to questions goals, during discussions and meetings, helped the analytics team and stakeholders to clarify and agree on analytics requirements through an incremental process. We presented models to stakeholders and asked if instances of such models can communicate the requirements and be understood by them. We observed that stakeholders were able to understand the content of the graphical model and were able to work with analytics team to elaborate on the models and to prioritize the requirements (see the gray-shaded area in Fig. 9). In addition, the models raised effective discussions during meetings and resulted in removing some question goals and adding others. These suggest that the framework can enhance the communication between business domain experts and data scientists (who develop analytics systems). Models from data analytics design view were constructed and updated during the project, mostly by the project manager and data scientists. The softgoals (most importantly Scalability) were used for making design decisions.

## 6. Related work

Several distinct bodies of work aim to support various aspects of advanced analytics projects. Process models have been proposed to guide data mining and knowledge discovery projects. They prescribe the sequence of tasks for conducting data mining projects. The work by Fayyad et al. [35] is often considered as the first reported data mining process model. The CRISP-DM model [9] is often mentioned as the most used and the *de facto* standard process model. Reference [36] provides a survey and a comparison of data mining and knowledge discovery process models. These process models do not provide systematics modeling and analysis support for requirements elicitation, elaboration, and design of business-driven analytics solutions. Unlike the framework proposed in this paper, they do not guide domain users step by step from business goals to uncover relevant decisions, then to generate analytics questions and solution design.

A variety of data science tools and platforms (both commercial and open-source) exist that support users in performing advanced analytics, e.g., IBM Watson Analytics, Microsoft Azure Machine Learning, RapidMiner, SAS, Weka. Although these platforms automate and facilitate data preparation and experimentation with algorithms, they do not provide any support for understanding and characterizing the business domain (e.g., stakeholders and their objectives, decisions, and indicators) and its interaction with the intended analytics solution. Moreover, explicit representation and consideration of quality requirements is not supported in such tools.

Another approach taken to support data mining is to establish formal ontologies. Reference [37] proposes ontologies to facilitate algorithm selection and the design of data mining workflows. The ontology in Ref. [32] formally represents data mining experiments to enable meta-learning. The work in Ref. [38] proposes an ontology-based system that explores design space and assists users in composing data mining processes. Reference [24] proposes a general ontology to unify data mining research. A review of such ontologies has influenced the development of the Analytics Design View and the Algorithms catalogue of the proposed framework (e.g., algorithm and indicators concepts in the framework also exist in many of these ontologies). However, similar to data mining process models and tools,

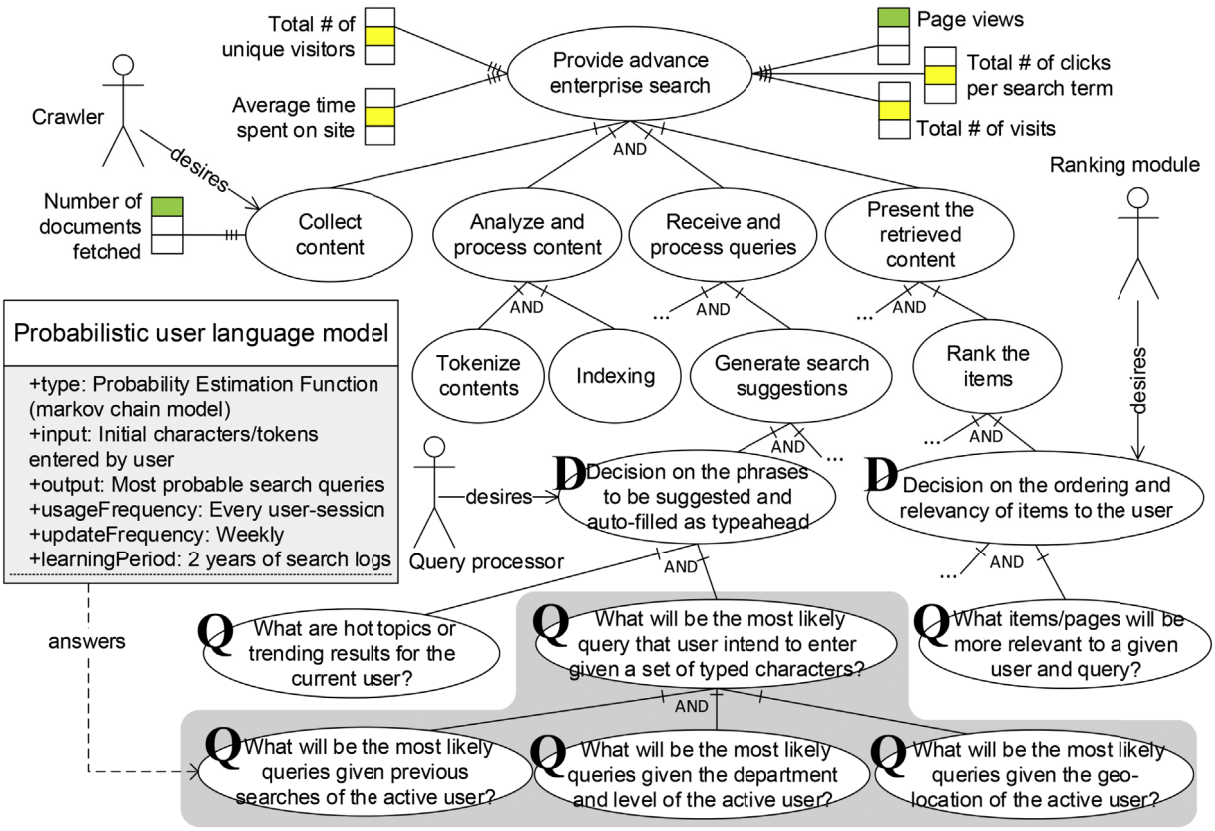


**Fig. 9.** A fragment of the Business View model for the third case study. Question goals in the gray-shaded area were perceived as more critical during the project scoping activities. See Fig. 1 for legend.

these ontologies do not provide any support for modeling and analyzing the business strategies and requirements aspects of analytics solutions.

Conceptual modeling techniques have been used to support various tasks or aspects related to business analytics. In the rest of this section, we briefly review related work in each of the relevant subareas and highlight the contribution of the framework.

The Business Intelligence Model (BIM) [18] represents a business in terms of strategic goals, processes, performance indicators, influences, and situations. Aiming at bridging the gap between business and data, it is developed based on well-established concepts in the business community as well as conceptual modeling techniques [39]. BIM supports a wide range of automated reasoning and business analysis techniques [17,40]. The case study in Ref. [41] reports that the language can facilitate design and development of BI solutions through a hybrid data mart design approach. Authors in Ref. [42] extend the BIM metamodel with new concepts to support modeling and reasoning on business plans. The work in Ref. [43] extends BIM to enable stress testing of business strategies. While BIM can support the design and development of BI systems, it does not support the modeling of advanced analytics and machine learning solutions. The framework in this paper uses concepts from BIM (gray-shaded area in Fig. 4) but extends them with new constructs (decision goals, question goals, and insights) to capture analytics requirements and link them to analytics design and then further to data preparation tasks.

Another group of related work are those that propose goal-oriented conceptual modeling approaches for requirements engineering of data warehouses. Reference [44] proposes a goal-oriented, Tropos-based methodology for requirements analysis in data warehouses and applies that in a real-world case study. The work in Ref. [45] proposes a goal-oriented, model-driven approach for development of data warehouses. The approach uses model-driven architecture (QTV model transformation) to arrive at a conceptual multi-dimensional design for data warehouse. Authors in Refs. [46,47] propose a Goal-Decision-Information model for analyzing data warehouse requirements. They provide a decision requirements metamodel [48] and use informational scenarios [49] for data warehouse requirements acquisition. The concept of decision goals in the Business View is similar to that of the proposals in Refs. [45,46]. The framework in this paper is different in that it focuses on extracting insights and applying machine learning algorithm on the datasets that are sourced from data warehouses in business domains. It supports the development of predictive and prescriptive types of analytics systems, in addition to descriptive ones.

In relation to the Data Preparation View of the framework, conceptual modeling approaches to ETL (Extract-Transform-Load) have been proposed. The work in Ref. [50] presents a metamodel and notation for modeling ETL processes in the early stages of data warehouse projects. In Ref. [27] authors propose a UML-based approach for conceptual modeling of ETL processes. They define a set of

common ETL activities (e.g., aggregation, filter, join) in terms of stereotyped classes and use UML dependencies to link them together. Reference [51] defines a model-driven architecture approach to transform ETL conceptual models to code. In Ref. [28] a BPMN-based modeling approach for ETL processes is presented. The work in Ref. [52] transforms BPMN models of ETL processes to the Business Process Execution Language (BPEL). While these works provide many of the modeling constructs in the Data Prep View (e.g., mechanism from Ref. [27]), the proposed framework connects these constructs to machine learning and organizational aspects of analytics solutions through the other modeling views.

Advanced analytics has also been an important and growing area of research within the information systems research community [53]. Related to this paper are those studies that aim to understand how organizations can create value through the use of analytics. For example, the work in Ref. [19] advocates the need for understanding the relationship between business analytics, decision making processes, and organizational performance. It argues that the first-order effects of analytics would be on organizational decisions through which the improvements in performance could be achieved. Authors in Refs. [54,55] identify the pathways through which business analytics contribute to business value. They present the process model of analyze-insight-decision-action through which an organization creates value from analytics. These studies motivate the need for capturing the relationship between business objectives, decision processes and analytical questions as part of the Business View. They provide theoretical support for the modeling constructs in the Business View.

To our knowledge, there is no existing framework that provides modeling support that connects business goals to advanced analytics system design through to data preparation, as proposed in this paper. Earlier works by the authors provided a condensed overview of the proposed framework [56] and its potential benefits [57].

## 7. Conclusions and future work

Effective development of advanced analytics systems has proven to be challenging for many business organizations [1]. Challenges include difficulties in determining the right analytics needs, selecting the right analytics algorithms, aligning analytics initiatives with high-level business objectives, and shortage of machine learning expertise in organizations. To overcome these challenges, the paper offered a conceptual modeling framework that includes three complementary modeling views along with a set of design catalogues.

The framework has been tested in three case studies. The case studies were used to illustrate the framework and to perform a preliminary validation of its expressiveness and usability. Findings suggest that the proposed framework can support the design and implementation of analytics solutions. One limitation of these case studies was that they were all conducted with the same analytics team within a single company. We are currently engaged with two other industrial partners to further validate and improve the framework, towards completing other pieces of the design science research approach. This includes a definition and operationalization of requirements for the three modeling views (as design artefacts) and evaluation of the framework against such requirements in real-world context. We are conducting empirical studies with users who are not the researchers. Usage, comprehensibility and learning curve of the modeling views can be examined for different types of roles (from business decision makers to data scientists) that are typically involved in analytics projects. These studies can lead to the definition of a model-based methodology, as part of the framework, for developing analytics systems. We are also developing guidelines and examples on how modeling views can be constructed and where catalogues can be helpful in such processes. We expect the contents of the analytics catalogues to be extended and validated in real usage settings in future work. We also plan to develop tools to support the framework.

## References

[1] S. Ransbotham, D. Kiron, P.K. Prentice, Beyond the hype: the hard work behind analytics success, MIT Sloan Manag. Rev. 57 (3) (2016).
[2] S. Viaene, A. Van den Bunder, The secrets to managing business analytics projects, MIT Sloan Manag. Rev. 53 (1) (2011) 65–69.
[3] E. Kandogan, A. Balakrishnan, E.M. Haber, J.S. Pierce, From data to insight: work practices of analysts in the enterprise, IEEE Comput. Graph. Appl. 34 (5) (2014) 42–50.
[4] S. LaValle, M. Hopkins, E. Lesser, R. Shockley, N. K, Analytics: the new path to value, MIT Sloan Manag. Rev. (2010) https://www-935.ibm.com/services/uk/gbs/pdf/Analytics_The_new_path_to_value.pdf.
[5] J. Manyika, et al., Big Data: the Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011.
[6] M. Luca, J. Kleinberg, S. Mullainathan, Algorithms need managers, too, Harv. Bus. Rev. 94 (2016) 96–101.
[7] R. Kohavi, L. Mason, R. Parekh, Z. Zheng, Lessons and Challenges From Mining Retail e-Commerce Data, Mach. Learn. 57 (2004) 83–113.
[8] T.H. Davenport, J.G. Harris, W. David, A.L. Jacobson, Data to knowledge to results: building an analytic capability, Calif. Manag. Rev. 43 (2) (2001) 117–138.
[9] P. Chapman, et al., CRISP-dm 1.0 Step-by-step Data Mining Guide, SPSS Inc., 2000.
[10] R. Kohavi, N.J. Rothleder, E. Simoudis, Emerging trends in business analytics, Commun. ACM 45 (8) (2002) 45–48.
[11] A. McAfee, E. Brynjolfsson, T.H. Davenport, D. Patil, D. Barton, Big data: the management revolution, Harv. Bus. Rev. 90 (10) (2012) 61–67.
[12] M. Yeomans, What every manager should know about machine learning, Harv. Bus. Rev. 93 (7) (2015).
[13] V.C. Storey, J.C. Trujillo, S.W. Liddle, Research on conceptual modeling: themes, topics, and introduction to the special issue, Data Knowl. Eng. 98 (2015) 1–7.
[14] V.C. Storey, I.-Y. Song, Big data technologies and Management: what conceptual modeling can do, Data Knowl. Eng. 108 (2017) 50–67.
[15] J. Horkoff, E. Yu, Comparison and evaluation of goal-oriented satisfaction analysis techniques, Requir. Eng. 18 (3) (2013) 199–222.
[16] ISO/IEC/IEEE Systems and software engineering – architecture description, in: ISO/IEC/IEEE 42010:2011(E), Dec. 2011, pp. 1–46.
[17] J. Horkoff, et al., Strategic business modeling: representation and reasoning, Software Syst. Model 13 (3) (2014) 1015–1041.
[18] L. Jiang, D. Barone, D. Amyot, J. Mylopoulos, Strategic models for business intelligence, in: ER 2011, vol. 6998, 2011, pp. 429–439.
[19] R. Sharma, S. Mithas, A. Kankanhalli, Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations, Eur. J. Inf. Syst. 23 (4) (2014) 433–441.
[20] T.H. Davenport, Business Intelligence and Organizational Decisions, Organizational Applications of Business Intelligence Management: Emerging Trends, 2012, p. 1.
[21] L. Fahey, Exploring 'analytics' to make better decisions–The questions executives need to ask, Strat. Leader. 37 (5) (2009) 12–18.
[22] L. Chung, B.A. Nixon, E. Yu, J. Mylopoulos, Non-functional Requirements in Software Engineering, Springer Science & Business Media, 2012.
[23] T. Menzies, T. Zimmermann, Software analytics: so what? IEEE Softw. 30 (4) (2013) 31–37.

[24] P. Panov, L.N. Soldatova, S. Džeroski, Towards an ontology of data mining investigations, in: Discovery Science, vol. 5808, 2009, pp. 257–271.
[25] E. Yu, Modelling strategic relationships for process reengineering, Soci. Model. Requir. Eng. 11 (2011) 2011.
[26] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Elsevier, 2012.
[27] J. Trujillo, S. Luján-Mora, A UML based approach for modeling ETL processes in data warehouses, in: ER 2003, vol. 2813, 2003, pp. 307–320.
[28] Z. El Akkaoui, J.-N. Mazón, A. Vaisman, E. Zimányi, BPMN-based conceptual modeling of ETL processes, in: DaWaK 2012, vol. 7448, 2012, pp. 1–14.
[29] S. Kotsiantis, Supervised machine learning: a review of classification techniques, Informatica 31 (3) (2007).
[30] I. Bose, R.K. Mahapatra, Business data mining—a machine learning perspective, Inf. Manag. 39 (3) (2001) 211–225.
[31] T.H. Davenport, J.G. Harris, R. Morison, Analytics at Work: Smarter Decisions, Better Results, Harvard Business Press, 2010.
[32] J. Vanschoren, H. Blockeel, B. Pfahringer, G. Holmes, Experiment databases - a new way to share, organize and learn from experiments, Mach. Learn. 87 (2) (2012) 127–158.
[33] R. Ali, F. Dalpiaz, P. Giorgini, A goal-based framework for contextual requirements modeling and analysis, Requir. Eng. 15 (4) (2010) 439–458.
[34] K. Peffers, T. Tuunanen, M.A. Rothenberger, S. Chatterjee, A design science research methodology for information systems research, J. Manag. Inf. Syst. 24 (3) (2007) 45–77.
[35] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Mag. 17 (3) (1996) 37–54.
[36] G. Mariscal, Ó. Marbán, C. Fernández, A survey of data mining and knowledge discovery process models and methodologies, Knowl. Eng. Rev. 25 (02) (2010) 137–166.
[37] C.M. Keet, A. Lawrynowicz, C. d'Amato, M. Hilario, Modeling issues and choices in the data mining OPtimization ontology, in: OWLED 2013, Montpellier, France, 2013.
[38] A. Bernstein, F. Provost, S. Hill, Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification, IEEE Trans. Knowl. Data Eng. 17 (4) (2005) 503–518.
[39] D. Barone, E. Yu, J. Won, L. Jiang, J. Mylopoulos, Enterprise modeling for business intelligence, in: IFIP Working Conference on the Practice of Enterprise Modeling, 2010, pp. 31–45.
[40] D. Barone, L. Jiang, D. Amyot, J. Mylopoulos, Composite indicators for business intelligence, in: ER 2011, vol. 6998, 2011, pp. 448–458.
[41] D. Barone, T. Topaloglou, J. Mylopoulos, Business intelligence modeling in action: a hospital case study, in: CAiSE 2012, vol. 7328, 2012, pp. 502–517.
[42] F. Francesconi, F. Dalpiaz, J. Mylopoulos, TBIM: a language for modeling and reasoning about business plans, in: ER 2013, vol. 8217, 2013, pp. 33–46.
[43] A. Maté, J. Trujillo, J. Mylopoulos, Stress testing strategic goals with SWOT analysis, in: ER 2015, vol. 9381, 2015, pp. 65–78.
[44] P. Giorgini, S. Rizzi, M. Garzetti, GRAnD: a goal-oriented approach to requirement analysis in data warehouses, Decis. Support Syst. 45 (1) (2008) 4–21.
[45] J.-N. Mazón, J. Pardillo, J. Trujillo, A model-driven goal-oriented requirement engineering approach for data warehouses, in: ER Workhops 2007, vol. 4802, 2007, pp. 255–264.
[46] N. Prakash, A. Gosain, An approach to engineering the requirements of data warehouses, Requir. Eng. 13 (1) (2008) 49–72.
[47] N. Prakash, A. Gosain, Requirements driven data warehouse development, in: CAiSE Short Paper Proceedings, 2003, pp. 13–17.
[48] N. Prakash, D. Prakash, D. Gupta, Decisions and decision requirements for data warehouse systems, in: CAiSE Forum, vol. 72, 2010, pp. 92–107.
[49] N. Prakash, Y. Singh, A. Gosain, Informational scenarios for data warehouse requirements elicitation, in: ER 2004, vol. 3288, 2004, pp. 205–216.
[50] P. Vassiliadis, A. Simitsis, S. Skiadopoulos, Conceptual modeling for ETL processes, in: DOLAP 2002, ACM, 2002, pp. 14–21.
[51] L. Muñoz, J.-N. Mazón, J. Trujillo, Automatic generation of ETL processes from conceptual models, in: DOLAP 2009, ACM, 2009, pp. 33–40.
[52] Z. El Akkaoui, E. Zimányi, Defining ETL workflows using BPMN and BPEL, in: DOLAP '09, 2009, pp. 41–48.
[53] H. Chen, R.H. Chiang, V.C. Storey, Business intelligence and analytics: from big data to big impact, MIS Q. 36 (4) (2012) 1165–1188.
[54] P.B. Seddon, D. Constantinidis, T. Tamm, H. Dod, How does business analytics contribute to business value? Inf. Syst. J. 27 (3) (2017) 237–269.
[55] T. Tamm, P. Seddon, G. Shanks, Pathways to value from business analytics, in: Presented at the Thirty Fourth International Conference on Information Systems, 2013.
[56] S. Nalchigar, E. Yu, R. Ramani, A conceptual modeling framework for business analytics, in: ER 2016, 2016, pp. 35–49.
[57] S. Nalchigar, E. Yu, Conceptual modeling for business analytics: a framework and potential benefits, in: 2017 IEEE 19th Conference on Business Informatics (CBI), vol. 01, 2017, pp. 369–378.

**Soroosh Nalchigar** is a Ph.D. candidate at the Department of Computer Science, University of Toronto, Canada. His research examines how machine learning and advanced analytics techniques can be used more effectively in organizations to support decision making. His areas of interest include business analytics, requirements engineering, conceptual modeling, information systems design, and operations research. Prior to his current studies, he received a M.Sc. in Data Mining and Knowledge Management from the University of Pierre and Marie Curie (Paris 6), and a M.Sc. and a Ph.D. in Management (Information Systems) from University of Tehran. His publications have appeared in *Expert Systems with Applications* and *Applied Mathematical Modelling*.

**Eric S. K. Yu** is Professor at the University of Toronto, Canada. He has research interests in information systems modeling and design, requirements engineering, software engineering, knowledge management, enterprise modeling, and adaptive enterprise architecture based on data analytics. He was originator of the i* modeling framework, which brings social and organizational modeling into information and software systems analysis and design. Books published include: Social Modeling for Requirements Engineering (MIT Press, 2011); Conceptual Modeling: Foundations and Applications (Springer, 2009); and Non-Functional Requirements in Software Engineering (Springer, 2000). He is series co-editor for the MIT Press book series Information Systems. He was Program Co-chair for the 27th and 33rd Int. Conference on Conceptual Modeling (ER 2008, 2014).