IMPLEMENTATION PROCESS FOR AUTOMATED DATA ANALYSIS IN MINERAL EXTRACTION COMPANIES

by

RICHARD A. LEACH

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2015

Acknowledgements

I want to express my gratitude to all those in the Industrial, Manufacturing, and Systems Engineering Department at The University of Texas at Arlington having an impact on my educational career. I want to thank my supervising professor Dr. John W. Priest for his guidance, understanding, and patience during the completion of this dissertation. I want to thank my doctoral committee members Dr. Donald H. Liles and Dr. Brian L. Huff and graduate advisor Dr. Sheik N. Imrhan for their interest and encouragement in my doctoral progress. For everything all of you have done for me over the years, I thank you. Finally, I want to thank all of the faculty, staff, and students I have had contact with in the Industrial, Manufacturing, and Systems Engineering Department. You have all made a positive impact and been responsible for a very memorable and rewarding experience in my life.

I want to express my gratitude to my family. My parents, B. H. and Syble U. Leach, never missed an opportunity to instill a desire for education in me. Their attitude always supported an understanding I could accomplish anything with the required effort. My daughter, McKenna, and son, Cameron, were the inspiration and source of my continuing determination to complete my engineering degrees. Finally, my wife, Dr. Karla N. Leach, has provided the support and understanding necessary for the completion of my dissertation. Without her, I would not have had the time needed to make the following doctoral work possible.

November 11, 2015

Abstract

IMPLEMENTATION PROCESS FOR AUTOMATED DATA ANALYSIS
IN MINERAL EXTRACTION COMPANIES

Richard A. Leach, PhD

The University of Texas at Arlington, 2015

Supervising Professor: John W. Priest

The need to determine knowledge from increasing amounts of information and raw data is a current and ongoing goal [1]. As technology continues to advance in the ability to collect and save more and more data, companies today must assimilate and understand this information as a valuable resource [2]. Companies in marketing and sales have found the use of a data warehouse and automated data analysis to be valuable resources [2]. This automated analysis of data has also been implemented in disciplines including scientific and medical research [2]. However, there continue to be other lower tech industrial areas of activity such as mineral extraction that have the capacity to generate and save large amounts of data that do not use automated knowledge discovery. Wyoming-based mineral extraction companies are collecting and storing an increasing number of data points. While the data and information exists and is available, there may be lost opportunity for insight into company activities. This may result in the loss of valuable improved efficiency. This research project provides detailed insights for an accepted comprehensive implementation process to explore automated data analysis in low-tech mineral extraction companies without previous experience using automated data analysis tools.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

Introduction

Many activities have had and continue to experience a dramatic increase in the ability to collect data, including scientific research and business marketing [3]. At the same time, storage capacity and access available in databases has become inexpensive [2]. Data collected in databases has increased in the number of records and the attributes associated [2]. Data is currently collected and stored in databases or data warehouses from point-of-sale information, financial transactions, telephone calls, and many other daily activities [2]. Manual statistical analysis and interpretation has been the accepted method of using data to determine knowledge [2]. While manual statistical methods continue to be valid for many small data applications, as the data increases in records and attributes, the size makes manual statistical methods impractical [1]. Large data sources can cause manual statistical methods to become slow, expensive, and subjective [2]. There are limitations in the use of traditional statistical analysis [2]. These limitations must be overcome in order to discover useful information and interesting patterns and allow decision makers to advance business objectives or organizational goals [2]. Automated data analysis can manage the data that an analyst must consider and present the results in a manner desirable for decisions [4]. While automated data analysis cannot replace the analyst's creative ability to understand data in determining knowledge, the amount of data available requires a discovery process using automated analysis to prevent the available data from being uninvestigated [2].

This research did not find the use of automated data analysis in Wyoming mineral extraction companies. These companies are engaged in activities not typically associated with the use of these discovery tools. They are experiencing the same increase in collection and storage of data as companies in marketing or scientific

research. Without the use of automated data analysis, these companies have the same opportunity for increased knowledge about company activities that can be lost. These companies also experience a common need when considering the use of advanced data discovery tools. Standardized process models have been proposed in knowledge discovery literature and used to provide the desired common framework for companies engaged in ongoing automated data analysis. These Wyoming mineral extraction companies require an implementation process for initial use of automated data analysis. This research produces an implementation process that provides initial requirements and incorporates necessary information within the process for these Wyoming mineral extraction companies wanting to consider the initiation of automated data analysis.

Chapter 2

Problem

Automated data analysis is a generic description used in this research for the desired result from the use of developed technologies found in knowledge discovery and data mining. Between 2012 and 2014, 16 Wyoming companies were contacted concerning their use of automated data analysis. These initial discussions were conducted to assess the existence of company data sources producing sufficient records or fields that could benefit from automated data analysis and current application of this technology. The Wyoming companies contacted were involved in various activities, including the following:

Coal surface mining

Shale natural gas exploration and production

Oil and natural gas production

Natural gas processing

Trona mining

Manufacturing using trona

Manufacturing using phosphates for food or agribusiness

Renewable energy services, and

Utility companies

Initial discussions confirmed the existence of an increased data collection and storage by these companies. The data sources did exist, but the use of automated data analysis did not. One example is the ability of these companies to collect and store electricity usage data in such detail, including the number of sources and frequency, which was not possible until recently. Electricity use is a major cost of operations. Automated data analysis could find unknown patterns in electricity use leading to savings

in this expense. These companies have a developed use of traditional statistical data analysis for decisions. Unfortunately, the application of automated data analysis has not been implemented by any of the companies contacted. However, all of the companies contacted were very interested in the use of automated data analysis and the possibility of determining unknown patterns in their existing data. This information could give direction to future data collection and storage. These companies have the authority to take action regarding their data analysis direction and the authority to expend resources in their determined direction. The lack of current experience with automated data analysis resulted in the same concern by the companies contacted. A brief explanation of automated data analysis was followed by the companies desire to understand the requirements needed for project initiation. What would the company require and experience during implementation of automated data analysis? These companies needed an implementation process with sufficient detail including critical issues of this technology for their industry.

In order to determine if an implementation process existed, a subset of the companies engaging in similar activities was selected. The majority of the Wyoming companies contacted were engaged in activities related to mineral extraction. Mineral extraction is a comprehensive term used to encompass traditional mining and all other operations that extract a valuable commodity from the earth. Mineral extraction companies were selected for a literature search. A literature search of companies in a similar business was conducted to determine any previous use of automated data analysis. A literature search of mineral extraction or mining with a variety of terms associated with automated data analysis did not indicate previous use. In addition, while process models for ongoing data discovery projects exist, these companies need different information and presentation as an introduction to a comprehensive

implementation process for automated data analysis. This research conducted a literature search of automated data analysis including knowledge discovery and data mining terminology. This literature search established the background, concepts, components, process models, and critical issues contained in the literature. This information was used to develop an implementation process for Wyoming mineral extraction companies. This implementation process includes the information and detail these companies need in order to consider the use of automated data analysis.

Chapter 3

Literature

3.1 Knowledge Discovery, Decision Support, OLAP and Data Mining

Knowledge Discovery in Databases (KDD) is the activity that develops tools for automated data analysis of the increasing collection and storage of data in databases of research, business, and manufacturing institutions [2]. Knowledge Discovery in Databases is an answer to the problem of having more data than can be analyzed with manual techniques [5]. Knowledge Discovery and Data Mining (KDDM) is the process of knowledge discovery applied to a data source [6]. Decision Sciences is an approach to making decisions that includes Decision Support (DS) [7]. Decision Support is now being associated with data mining and has previously used principles from disciplines including operations research, decision analysis, and decision support systems [7]. In the future, Decision Support is expected to receive contributions from advances in data warehouses, integration with data mining, developments in modeling and computing, and networking [7]. Decision support systems are concerned with improving effectiveness, not the efficiency of decisions [8,9]. Companies may use different decision support systems for different types of decisions [8]. A decision support system uses a particular decision process and set of methods, techniques, and approaches developed to produce the required decision for a specific objective [8].

Online Analytical Processing (OLAP) tools for business applications use a deductive approach for advanced data analysis and decision support [8,10]. Data mining is not superior to Online Analytical Processing and Online Analytical Processing is not superior to data mining [8]. The deductive approach of Online Analytical Processing is limited by the effectiveness of the analyst to acquire the most valuable information, trends, and patterns from the data selected [8]. Online Analytical Processing requires the

analyst to ask the right question to get the desired answer [8,11]. Data mining uses an inductive approach to analyzing data that allows the answers to more investigative questions, enables the solution of different problems, and acquires different information than Online Analytical Processing [8,10].

3.2 Knowledge Discovery Evolution

The Knowledge Discovery Process (KDP) is considered to be the same as Knowledge Discovery in Databases, a process that attempts to reveal new information within data [12]. The Knowledge Discovery in Databases process attempts to accomplish the goal of new information by reducing large data sources into manageable data that can be presented in a report, modeling the process that produced the data, or a predictive model for use in evaluation of new data [5]. Knowledge Discovery in Databases has the capability to identify valid, novel, useful, and understandable patterns in data [2,5]. The process uses data mining methods as the means to discover unknown patterns and extraction of pattern knowledge [5]. Data mining methods are the means to accomplish the goals of the knowledge discovery process, but only require an estimated 15% to 25% of the entire effort as one step in the process [4,13]. Data mining is defined as the discovery of patterns or models from observed data [4]. Patterns discovered should be determined with a degree of certainty to be valid for observed data, novel-producing new information for the user, potentially useful-producing answers to a desired objective, and understandable-producing results in a form useful to decision makers [2,5,14]. Data mining is considered a valuable activity [2,14]. The process requires some search or inference, as results are not achieved by the application of a computation using the values contained in the data [2,14]. From the first Knowledge Discovery in Databases workshop in 1989, meetings have been held to investigate the problems inherent in knowledge extraction from large databases involving the many steps in the

process, including data access and manipulation to fundamental mathematical and statistical inference, search, and reasoning [2]. Knowledge Discovery in Databases is advancing the goal of new pattern or knowledge discovery from large data sets through continuing evolution of contributions from research fields such as machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, data visualization, and high performance computing [2,5,14]. Knowledge Discovery and Data Mining includes, in the knowledge discovery process, the storage and access of data, the development of algorithms that are efficient and useful in analyzing massive data sets, the interpretation and presentation of results in an understandable manner, and how to continue the process of discovery by analysts using data [2,5,6]. Knowledge Discovery in Databases is a process involving a number of steps with differences depending on the process model, but agreement the steps are repeated in multiple iterations [14]. Representative process steps include business understanding, data understanding, data preparation, data mining, results evaluation, and knowledge deployment [14]. Knowledge discovery is understood to not strictly follow a step-by-step process, because any step is iterative and interactive and may require the return to an earlier step in a number of feedback loops [4,5]. Knowledge discovery from data requires statistical inference of general patterns using a representative sample from a larger population, providing a means for estimating the uncertainty of the determined characteristics [5]. This leads to the expectation that quantitative measures exist for evaluating the discovered patterns [2]. Knowledge Discovery in Databases is particularly interested in discovering understandable patterns that can be interpreted as providing useful or interesting knowledge [2,5]. The Knowledge Discovery in Databases process usually involves more search in model discovery and investigate larger data sets with more variables than usually considered by statistics [14]. Knowledge Discovery in

Databases was initially advanced by existing work concerned with inferring models from data [2]. The data mining step in Knowledge Discovery in Databases uses methods from statistical pattern recognition, applied statistics, machine learning, and neural networks to discover patterns from data [2]. Existing knowledge is an important set of expectations or beliefs about the problem being investigated [15]. Pattern discovery methods may not give appropriate consideration of the prior knowledge available from decision makers allowing the discovery of obvious or irrelevant patterns [15]. Researchers and analysts have observed that many existing tools discover a large number of valid but obvious or irrelevant patterns causing Knowledge Discovery in Databases effort to concentrate on the validity of results and invest less effort in the discovery of novel and useful results [15]. Finally, Knowledge Discovery in Databases has the same judgment for projects as those for other applications of advanced technology including the potential impact on a business objective, the absence of an easier alternative solution, and organizational support for using the technology [14].

3.3 Knowledge Discovery Structure

The first workshop on Knowledge Discovery in Databases in 1989 considered the need for a Knowledge Discovery and Data Mining process model in order to establish and use a common approach to Knowledge Discovery and Data Mining projects [6]. A standard process model provides direction during the planning and execution of knowledge discovery projects [6]. The concept of a standard process model establishes the generally accepted content and steps to follow in a knowledge discovery project [12]. Chapter 2 of *Data Mining a Knowledge Discovery Approach* lists the following five reasons for a standard process model [12]:

1. The end product must be useful for the user/owner of the data….
2. A well-defined KDP model should have a logical, cohesive, well-thought-out structure and approach that can be presented to

decision-makers who may have difficulty understanding the need, value, and mechanics behind a KDP….
3. Knowledge discovery projects require a significant project management effort that needs to be grounded in a solid framework….
4. Knowledge discovery should follow the example of other engineering disciplines that already have established models….
5. There is a widely recognized need for standardization of the KDP. (Cios, K.J.; Pedrycz, W.; Swiniarski, R.W.; Kurgan, L., 2007) [12]

The book *Advances in Knowledge Discovery and Data Mining* released in 1996 offered an initial attempt at a standard process model using input from researchers and data analysts performing knowledge discovery projects, defining three main tasks of model selection and execution, data analysis, and output generation [6]. The first task was given additional detail with the division into data segmentation, model selection, and parameter selection [6]. The second task was given additional detail with the division into model specification, model fitting, model evaluation, and model refinement [6]. The third task was given additional detail with the division into generation of reports and development of implementing and monitoring of the obtained results into the original problem objective [6]. An example of a model with nine steps in chapter 2 of *Data Mining a Knowledge Discovery Approach* follows [12]:

1. Developing and understanding the application domain. This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge.
2. Creating a target data set. Here, the data miner selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset.
3. Data cleaning and preprocessing. This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes.
4. Data reduction and projection. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.

5. Choosing the data mining task. Here, the data miner matches the goals defined in Step 1 with a particular DM method, such as classification, regression, clustering, etc.
6. Choosing the data mining algorithm. The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.
7. Data mining. This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc.
8. Interpreting mined patterns. Here, the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models.
9. Consolidating discovered knowledge. The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge. (Cios, K.J.; Pedrycz, W.; Swiniarski, R.W.; Kurgan, L., 2007) [12]

3.4 Knowledge Discovery Process Models

Knowledge discovery process models contain similar steps, including problem understanding, data mining, and evaluation of discovered knowledge, however, there are important differences [12]. These differences cause the models to have the potential to perform better or worse depending on the application problem and desired objective [12]. A five-step model developed for industrial activities does not include a data-understanding step included in otherwise similar six-step models [12]. This five-step model includes business objectives determination, data preparation, data mining, results analysis, and knowledge assimilation [1,16]. This five-step model continues with the following explanation for each step. Business objectives determination clearly identifies the business problem to be investigated [1]. Data preparation involves data selection, preprocessing, and transformation [1]. Data mining involves model or algorithm selection and application [1]. Results analysis evaluates whether any results are novel or interesting [1]. Finally, knowledge assimilation determines the possible deployment using the new knowledge [1]. An eight-step model developed for academic activities has

considerable detail of required actions in the early steps of a knowledge discovery project, but does not include a step concerned with deploying the discovered knowledge [6,12]. A nine-step model developed for academic activities delays the steps involving data mining and algorithm decisions until later in the process [6]. The previous models have steps involving data mining and algorithm decisions earlier in the process before preprocessing of the data allowing the data to be correctly prepared for the data mining step, avoiding the need to repeat some of the earlier steps [6].

The six-step Cross-Industry Standard Process for Data Mining (CRISP-DM) is an industrial model developed by a number of European companies providing industrial perspective on all development issues with academic and governmental support [1,12]. The Cross-Industry Standard Process for Data Mining was developed using European Commission funding and included data mining providers and early users of data mining technology [17,18]. This model offers considerable detail for each step that has very helpful documentation and has been used in a number of industrial applications [12]. A summary of the six steps of the CRISP-DM model in chapter 2 of *Data Mining a Knowledge Discovery Approach* follows [12]:

1. *Business understanding*. This step focuses on the understanding of objectives and requirements from a business perspective. It also converts these into a DM problem definition, and designs a preliminary project plan to achieve the objectives. It is further broken into several substeps, namely,
   a. determination of business objectives,
   b. assessment of the situation,
   c. determination of DM goals, and
   d. generation of a project plan.
2. *Data understanding.* This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, and detection of interesting data subsets. Data understanding is further broken down into
   a. collection of initial data,
   b. description of data,
   c. exploration of data, and
   d. verification of data quality.

3. *Data preparation*. This step covers all activities needed to construct the final data set, which constitutes the data that will be fed into DM tool(s) in the next step. It includes table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
   a. selection of data,
   b. cleansing of data,
   c. construction of data,
   d. integration of data, and
   e. formatting of data substeps.
4. *Modeling*. At this point, various modeling techniques are selected and applied. Modeling usually involves the use of several methods for the same DM problem type and the calibration of their parameters to optimal values. Since some methods may require a specific format for input data, often reiteration into the previous step is necessary. This step is subdivided into
   a. selection of modeling technique(s),
   b. generation of test design,
   c. creation of models, and
   d. assessment of generated models.
5. *Evaluation*. After one or more models have been built that have high quality from a data analysis perspective, the model is evaluated from a business objective perspective. A review of the steps executed to construct the model is also performed. A key objective is to determine whether any important business issues have not been sufficiently considered. At the end of this phase, a decision about the use of the DM results should be reached. The key substeps in this step include
   a. evaluation of the results,
   b. process review, and
   c. determination of the next step.
6. *Deployment*. Now, the discovered knowledge must be organized and presented in a way that the customer can use. Depending on the requirements, this step can be as simple as generating a report or as complex as implementing a repeatable KDP. This step is further divided into
   a. plan deployment,
   b. plan monitoring and maintenance,
   c. generation of final report, and
   d. review of the process substeps. (Cios, K.J.; Pedrycz, W.; Swiniarski, R.W.; Kurgan, L., 2007) [12]

The standard process model for knowledge discovery allows advance understanding of the required steps and the actions required during each step in a data mining project, increasing the opportunity for successful effort [17]. A very important understanding of the data mining step of the knowledge discovery process is the

common need to perform repeated iterative application of specific data mining methods [5]. Data mining uses a search method that will usually require evaluation of hypotheses, evaluation of the search results, and appropriate use of the results [19]. Understanding of search methods is not accomplished by statistical methods, but statistics has considerable value in the evaluation of hypotheses in the conducting a search, in evaluating the results of a search, and in understanding the appropriate uses of the results [19]. The understanding of issues covered above will assist in the correct use of data mining methods with valid results [2,5]. A data mining project that does not correctly apply statistics to the problem should not be attempted [2,5].

The following table provides a comparison of six process models, including the type of activity each was developed for, which lists two were intended for industrial applications [6]. The comparison of steps for each process model indicates the considerable variations [6]. Implementation information is not a listed step in any of the process models.

Table 3-1 Comparison of Knowledge Discovery and Data Mining Models

Model	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*	Generic model
Area	Academic	Industrial	Academic	Industrial	Academic	N/A
No of steps	9	5	8	6	6	6
Refs	(Fayyad *et al.*, 1996d)	(Cabena *et al.*, 1998)	(Anand & Buchner, 1998)	(Shearer, 2000)	(Cios *et al.*, 2000)	N/A
Steps	1 Developing and Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identification	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
			2 Problem Specification			
	2 Creating a Target Data Set	2 Data Preparation	3 Data Prospecting	2 Data Understanding	2 Understanding the Data	2 Data Understanding
			4 Domain Knowledge Elicitation			
	3 Data Cleaning and Preprocessing		5 Methodology Identification	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	4 Data Reduction and Projection		6 Data Preprocessing			
	5 Choosing the DM Task					
	6 Choosing the DM Algorithm					
	7 DM	3 DM	7 Pattern Discovery	4 Modeling	4 DM	4 DM
	8 Interpreting Mined Patterns	4 Domain Knowledge Elicitation	8 Knowledge Post-processing	5 Evaluation	5 Evaluation of the Discovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge		6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deployment

Source: Data from Kurgan, L. A.; Musilek, P., table 1 [6]

The next table provides a more detailed comparison of the activities contained in the steps of the six process models [6]. This comparison continues to indicate the variation of the process models [6].

Table 3-2 Individual Steps of Knowledge Discovery and Data Mining Models

Model	Generic	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*
Steps	STEP 1. Application Domain Understanding	1 Learning goals of the end-user and relevant prior knowledge	1 Understanding the business problem and defining business objectives, which are later redefined into DM goals	1 Identification of human resources and their roles 2 Partitioning of the project into smaller tasks that can be solved using a particular DM method	1 Understanding of business objectives and requirements, which are converted into a DM problem definition	1 Defining project goals, identifying key people, learning current solutions and domain terminology, translation of project goals into DM goals, and selection of DM methods for Step 4
	STEP 2 Data Understanding	2. Selection of a subset of variables and sampling of the data to be used in later steps	2 Identification of internal and external data sources, selection of subset of data relevant to a given DM task. It also includes verifying and improving data quality, such as noise and missing data. Determination of DM methods that will be used in the next step and transformation of the data into analytical model required by selected DM methods	3 Analysis of accessibility and availability of data, selection of relevant attributes and a storage model 4 Elicitation of the project domain knowledge	2 Identification of data quality problems, data exploration, and selection of interesting data subsets	2 Collecting the data, verification of data completeness, redundancy, missing values, plausibility, and usefulness of the data with respect to the DM goals
Steps	STEP 3 Data Preparation and Identification of DM Technology	3 Preprocessing of noise, outliers, missing values, etc, and accounting for time sequence information 4 Selection of useful attributes by dimension reduction and transformation, development of invariant data representation 5 Goals from Step 1 are matched with a particular DM method, i.e. classification, regression, etc. 6 Selection of particular data model(s), method(s), and method's parameters		5 Selection of the most appropriate DM method, or a combination of DM methods 6 Preprocessing of the data, including removal of outliers, dealing with missing and noisy data, dimensionality reduction, data quantization, transformation and coding, and resolution of heterogeneity issues	3 Preparation of the final dataset, which will be fed into DM tool(s), and includes data and attribute selection, cleaning, construction of new attributes, and data transformations	3 Preprocessing via sampling, correlation and significance tests, cleaning, feature selection and extraction, derivation of new attributes, and data summarization. The end result is a data set that meets specific input requirements for the selected DM methods
	STEP 4 Data Mining	7 Generation of knowledge (patterns) from data, for example classification rules, regression model, etc.	3 Application of the selected DM methods to the prepared data	7 Automated pattern discovery from the preprocessed data	4 Calibration and application of DM methods to the prepared data	4 Application of the selected DM methods to the prepared data, and testing of the generated knowledge

Source: Data from Kurgan, L. A.; Musilek, P., table 2 [6]

Table 3-2—*Continued*

Model	Generic	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*
Steps	STEP 5 Evaluation	8 Interpretation of the model(s) based on visualization of the model(s) and the data based on the model(s)	4 Interpretation and analysis of DM results; usually visualization technique(s) are used	8 Filtering out trivial and obsolete patterns, validation and visualization of the discovered knowledge	5 Evaluation of the generated knowledge from the business perspective	5 Interpretation of the results, assessing impact, novelty and interestingness of the discovered knowledge. Revisiting the process to identify which alternative actions could have been taken to improve the results
	STEP 6 Knowledge Consolidation and Deployment	9 Incorporation of the discovered knowledge into a final system, creation of documentation and reports, checking and resolving potential conflicts with previously held knowledge	5 Presentation of the generated knowledge in a business-oriented way, formulation of how the knowledge can be exploited, and incorporation of the knowledge into organization's systems		6 Presentation of the discovered knowledge in a customer-oriented way. Performing deployment, monitoring, maintenance, and writing final report	6 Deployment of the discovered knowledge. Creation of a plan to monitor the implementation of the discovered knowledge, documenting the project, extending the application area from the current to other possible domains
Notes	Unified set of steps. Each step's scope can be inferred from the corresponding steps of other models	Significant iterations by looping between any two steps are possible, but no details are given. This model became a cornerstone for the future models, and is currently the most cited model in the scientific literature	The first business-oriented model, which is easy to comprehend by the layman. Emphasizes iterative nature of the model, but no details are given. Authors note that DM is often performed together with the Step 5	Provides a detailed breakdown of the initial steps. Emphasizes iterative nature of the model, where experts examine the knowledge after the last step and may decide to refine and rerun part or the entire process. Lacks step where the discovered knowledge is applied	Uses easy to understand vocabulary, and has good documentation. Divides steps into sub-steps that provide all necessary details. Acknowledges strong iterative nature of the process, but without details	Emphasizes and explicitly describes iterative and interactive aspects of the process*

* The specific feedback loops described in Cios *et al.* model include (Cios & Kurgan, 2005).

- From Step 2 to Step 1: execution of this loop is triggered by the need for additional domain knowledge to improve data understanding.
- From Step 3 to Step 2: execution of this loop is triggered by the need for additional or more specific information about the data to guide choice of specific data preprocessing.
- From Step 4 to Step 1: the loop is performed if results generated by selected DM methods are not satisfactory and modification of project's goals is required.
- From Step 4 to Step 2: the most common reason is poor understanding of the data, which results in incorrect selection of DM method(s) and its subsequent failure (e.g. data was misclassified as continuous and discretized in Understanding the Data step).
- From Step 4 to Step 3: the loop is motivated by the need to improve data preparation; this is often caused by specific requirements of used DM method, which may have been unknown during Step 3.
- From Step 5 to Step 1: the most common cause is invalidity of the discovered knowledge; there are several possible reasons including misunderstanding or misinterpretation of the domain, incorrect design or misunderstanding of problem restrictions, requirements, or goals. In these cases the entire KDDM process needs to be repeated.
- From Step 5 to Step 4: this loop is executed when the discovered knowledge is not novel, interesting, or useful; the least expensive solution is to choose a different DM tool and repeat the DM step.

The importance of these feedback mechanisms has been confirmed by several research application of the model (Cios *et al.*, 2000; Sacha *et al.*, 2000; Kurgan *et al.*, 2001; Maruster *et al.*, 2002; Kurgan *et al.*, 2005). Introduction and detailed description of these mechanisms and their triggers is important as it increases awareness and helps the user of the process to avoid similar problems by deploying appropriate countermeasures.

Source: Data from Kurgan, L. A.; Musilek, P., table 2 [6]

The next table provides information on five of the process models development, use, and drawbacks [6]. The development of three of the process models did not have industry involvement [6]. Experience with data mining is required by four of the process models [6]. Indication of feedback loops is neglected by four of the process models [6].

Table 3-3 Comparison of Development and Use of Knowledge Discovery and Data Mining Models

KDDM model	1	2	3				4	5	6	7	8	9	10
			a	b	c	d							
Fayyad *et al.* (9 step)	1996	83	8	2	10	10	Not listed	Medicine, engineering, production, e-business, software	0	Yes MineSet™	No Web site, description based on research papers	4 Requires background in DM	— Prepared data may not be suitable for the tool of choice, and thus unnecessary loop back previous steps may be required —limited discussion of feedback loops
Cabena *et al.* (5 step)	1998	2	0	1	1	1	Not listed	Marketing and sales	2 (one company)	No	No Web site, description based on a book	2 Requires some knowledge of DM terminology	—Omits the data understanding step —Limited discussion of feedback loops
Anand & Buchner (8 step)	1998	2	1	1	2	0	Not listed	Marketing and sales	0	No	No Web site, description based on research papers	4 Requires background in DM	—Too detailed breakdown of steps in the early phases of the KDDM process —Does not accommodate for a step that is concerned with putting the discovered knowledge to work —Limited discussion of feedback loops
CRISP-DM	2000	4	5	6	11	9	46%	Medicine, engineering, marketing and sales, environment	5 (consortium of companies)	Yes Clementine™	Has Web site, description based on research and white papers	1 Easy to understand, in lay words	—limited discussion of feedback loops
Cios *et al.* (6 step)	2000	4	21	0	21	16	Not listed	Medicine, software	0	No	No Web site, description based on research papers	3 Requires knowledge of DM terminology	—Popularized and geared towards research applications

(1 – year when the model was introduced, 2 – total number of citations per year, 3 – number of applications of the model: 3a – in academia, 3b – in industry, 3c – total number of applications, 3d – total number of applications with applications by model authors discounted, 4 – average KDnuggets poll results, 5 – application areas, 6 – industry involvement (0 none – 5 strong), 7 – software tool support, 8 – documentation, 9 – ease of use (0 novice – 5 expert), 10 – main drawbacks)

Source: Data from Kurgan, L. A.; Musilek, P., table 4 [6]

The following figure compares estimates of the relative time needed for a list of six steps in process models [6].

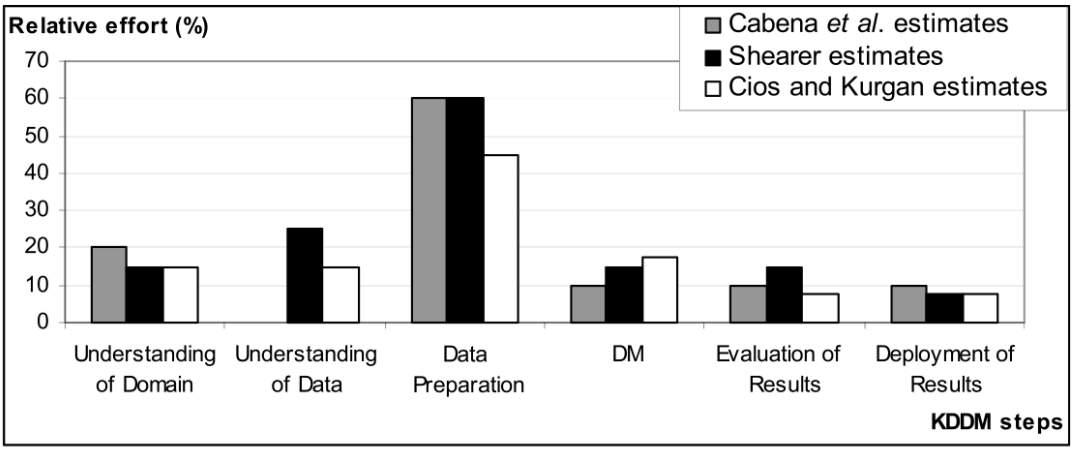

Figure 3-1 Relative Effort for Knowledge Discovery and Data Mining Steps. [6]

The following figure is a comparison of proposed and actual stages in a process model example with duration times [1].

Comparison of proposed vs. actual stages.

Proposed Project Stages	Planned Project Stages	Planned Duration (in days)	Actual Project Stages	Actual Duration (in days)	Actual Days as % of Total	# of Site Visits
Business Objectives Determination	Business Objectives Determination	1	Business Objectives Determination	1	5%	1
Data Preparation	Data Preparation	6	Data Preparation	6	30%	1
Data Mining	Interactive Data Mining and Results Analysis	3	*Data Audit*	1	5%	1
Results Analysis	Results Synthesis and Presentation	3	Interactive Data Mining and Results Analysis	3	15%	3
Knowledge Assimilation			*Back End Data Mining*	6	30%	2
			Results Synthesis and Presentation	3	15%	1

Figure 3-2 Comparison of Proposed and Actual Stages in an Example Knowledge Discovery and Data Mining Project. [1]

The following figure documents a representative timeline for a process model project [1].

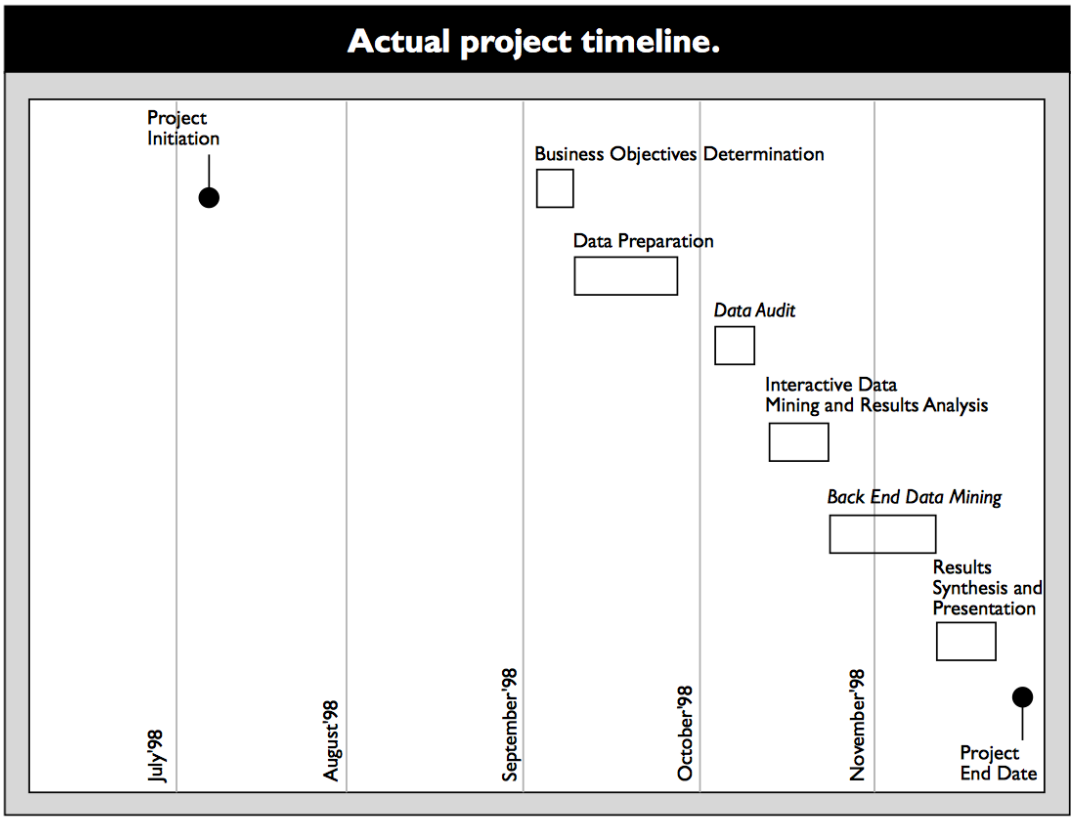

Figure 3-3 Actual Process Model Project Timeline. [1]

3.5 Knowledge Discovery Goals

Decisions made concerning how a system will be used determine the knowledge discovery goals [14]. Knowledge discovery goals are initially divided into verification, a system verifying the hypothesis of the user, and discovery, a system finding new patterns without user direction [5,14]. The discovery goal can be considered as having the two categories of prediction: a system predicting how the investigated function will act in the future, and description, where the system discovers patterns and presents the results to the user in an understandable form [5,14]. Data mining usually has the primary goals of

prediction and description [5]. Prediction uses a known sample to predict the properties of an unknown sample, where the assumption is made that the two samples originate from the same distribution of data [19]. Prediction has the same goals of accuracy and uncertainty as estimation, and the validity of the results is measured by the variance of the predictor [19]. In the use of data mining, prediction and description can have considerable difference in importance [5]. Data attributes need to have relevance to the discovery task: if the available attributes do not contain the required information, no amount of data will allow prediction [14]. Another consideration is the existence of noise, a condition of patterns that are difficult to discover unless a large number of cases can reduce noise and produce aggregate patterns [14]. Prior knowledge may be the most important input to consider [14]. The prior knowledge about important factors, relationships, known patterns, and the user-determined objective provides valuable direction in the data set investigation [14].

3.6 Data Mining Focus

Knowledge Discovery in Databases has concentrated most effort on the step of data mining [5]. Data mining is defined as the analysis and extraction of data patterns from databases discovering new and valuable information in the patterns and rules from data relationships [1]. Data mining observes data to determine patterns and infer knowledge from fitted models [2,5]. The knowledge discovery process includes the requirement of subjective user judgment to decide if models contain useful knowledge [2,5]. Data mining is a very important step in knowledge discovery, producing the extraction of knowledge from data, while the balance of the knowledge discovery process has a number of other actions to accomplish [20]. However, without the other steps such as data preparation, data selection, data cleaning, prior knowledge consideration, and proper interpretation of the results, the knowledge discovery process may fail to obtain

useful knowledge from the data [2,5,14]. The knowledge discovery process uses the database resulting from any required selection, preprocessing, subsampling, and transformations [5]. The knowledge discovery process applies data mining methods or algorithms to the database with the intent to discover patterns and evaluate the discovered patterns for the determination of patterns that contain new knowledge [5]. The data mining step is limited by acceptable constraints in the search for patterns in the data [2,5]. Data dredging is the application of data mining methods in a manner leading to discovery of meaningless patterns [2,5,14]. Patterns occurring may be numerous from a data set and because pattern searches are usually involved with large data sets, data dredging is avoided with computational constraints restricting the data that can be explored by a data mining algorithm [2]. The data mining step of knowledge discovery finds patterns from data using techniques from machine learning, pattern recognition, and statistics [5]. The literature of statistics, pattern recognition, machine learning, and databases contain descriptions of many data mining algorithms [2,5]. Data mining goals can be predictive, focusing on accuracy, or descriptive, focusing on understanding the data generating process, or a combination of these goals [2]. Data mining is a multidisciplinary field using concepts from a variety of disciplines including artificial intelligence, database theory, data visualization, marketing, mathematics, operations research, pattern recognition, and statistics [1]. Statistics considers databases uncontrolled sources of samples causing data mining results to be interesting and, often difficult, inference determinations [19]. However, the value of data mining is the knowledge that can be inferred from the data and put to use [2]. Data mining results are beneficial when deployed by a business strategy to advance a business objective [1]. Data mining uses principles from a number of disciplines including statistics that uses the hypothesis approach to focus on inferring patterns or models from data [1]. Unlike

statistics, data mining uses a discovery directed approach without a stated hypothesis for the problem being investigated [1]. Clustering, association, and predictive modeling are three predominant data mining approaches that demonstrate the difference from statistics [1]. Clustering divides data into subsets using common characteristics, association algorithms uncover rules, and predictive modeling uses a number of algorithms such as binary decision tree, logistic regression, and standard linear regression [1]. Data mining discovery approaches can be considered as supervised or unsupervised [8,21]. Supervised discovery has the goal to predict the value of an outcome based on a number of input measures, while unsupervised discovery has the goal to describe associations and patterns among a set of input measures, where no outcome measure exists [8]. Data mining accepts the data may determine the problem definition and lead to discovery of previously unknown but interesting patterns unlike other data analysis techniques [8]. Data mining methods contrast with statistical methods in the use of model representations containing more information, increasing the possibility of information discovery, trends, and patterns that make the results understandable to the knowledge discovery process users [8].

3.7 Data Mining Approaches

Data mining can be used through the two approaches of software tools or application systems [8]. The first approach of software tools accomplishes data mining projects by the application of data mining software tools [8,22,23,24]. The use of data mining software tools requires considerable experience in data mining methods, databases, and statistics [8,23]. Data mining software tools offer a variety of methods and parameters that the user must understand in order to use them effectively [8,23]. The use of the data mining software tools approach has disadvantages in the number of specialists required for a project, the necessary understanding of the software tools, and

in the acceptable use of results and models [8,25,26]. These disadvantages result in the need for the second approach of data mining application system [8,27]. The data mining application systems approach offers the ability to develop decision support systems that use data mining methods without requiring business users to have experience in data mining [8,27]. This approach allows business users and other decision makers to have data mining models presented in a user-understandable form [8,27].

3.8 Human-centric and Data-centric Models

Human-centric models focus on the interactive involvement of a data analyst during the process, while the data-centric models focus on the iterative and interactive data analysis tasks [6]. Knowledge discovery may best result from a knowledge analyst using a set of intelligent, visual, and perceptual tools for data analysis that go far beyond the existing statistical tools and significantly enhance the human capabilities for data analysis [28]. The time needed to discover knowledge is not a problem [28]. Useful knowledge discovered should be introduced into the database using integrity rules, semantic optimization rules, or functional dependencies [28]. Users need assistance in analysis if many independent variables exist and require consideration to model the system correctly [4]. The need for an analyst, in the loop and tools to allow analyst guidance of the rule discovery process, has been emphasized because the decision of useful or interesting patterns is, in many knowledge discovery projects, dependent on the application [29]. The decision maker must have alternative answers to consider, understanding possible alternative conditions, understanding the probabilities of possible conditions, and understanding of the relative advantages for each possible answer in each possible condition [19]. Decision rules determine the alternative to select using all the information available to the decision maker [19].

3.9 Model Fitting

Model fitting uses statistical or logical approaches [5]. Statistical model fitting allows for uncontrolled search in the model, whereas logical model fitting is completely controlled [5]. A pattern is a description of a subset of the data contained in a database or a model applicable to the subset [5]. Pattern discovery is fitting a model to data, determining structure in data, or making a description of a set of data [5]. Significant dependencies among variables are discovered by dependency models [5]. There are two levels of dependency models, the structural level of the model determines variables having dependence and the quantitative level of the model determines the relative importance of the dependencies [5]. Machine learning provides a model of data or makes predictions of data from the goal of a trained learning machine [30]. Data mining goals of prediction and description are accomplished using the following example data mining methods [14]:

1. **Classification:** learning a function that maps (classifies) a data item into one of several predefined classes.
2. **Regression:** learning a function which maps a data item to a real-valued prediction variable and the discovery of functional relationships between variables.
3. **Clustering:** identifying a finite set of categories or clusters to describe the data. Closely related to clustering is the method of *probability density estimation,* which consists of techniques for estimating from data the joint multi-variate probability density function of all of the variables/fields in the database.
4. **Summarization:** finding a compact description for a subset of data, e.g., the derivation of summary or association rules and the use of multivariate visualization techniques.
5. **Dependency Modeling:** finding a model which describes significant *dependencies* between variables (e.g., learning of belief networks).
6. **Change and Deviation Detection:** discovering the most significant changes in the data from previously measured or normative values. (Fayyad, U.; Piatetski-Shapiro, G.; Smyth, P., 1996) [14]

Three primary components of a data mining algorithm follow [14]:

1. **Model Representation:** the language used to describe discoverable patterns. If the representation is too limited, then no amount of training time or examples will produce an accurate model for the data. It is important that a data analyst fully comprehend the representational assumptions which may be inherent in a particular method. It is equally important that an algorithm designer clearly state which representational assumptions are being made by a particular algorithm. Note that more powerful representational power for models increases the danger of over fitting the training data resulting in reduced prediction accuracy on unseen data.
2. **Model Evaluation Criteria:** quantitative statements (or "fit functions") of how well a particular pattern (a model and its parameters) meet the goals of the KDD process. For example, predictive models are often judged by the empirical prediction accuracy on some test set. Descriptive models can be evaluated along the dimensions of predictive accuracy, novelty, utility, and understandability of the fitted model.
3. **Search Method:** consists of two components: *parameter search* and *model search*. Once the model representation (or family of representations) and the model evaluation criteria are fixed, then the data mining problem has been reduced to purely an optimization task: find the parameters/models from the selected family which optimize the evaluation criteria. In *parameter search*, the algorithm must search for the parameters which optimize the model evaluation criteria given observed data and a fixed model representation. *Model search* occurs as a loop over the parameter search method: the model representation is changed so that a family of models are considered. (Fayyad, U.; Piatetski-Shapiro, G.; Smyth, P., 1996) [14]

Knowledge Discovery in Databases considers the need for modeling algorithms with capacity to deal with large and noisy data sets [5]. Model representations include decision trees, linear models, nonlinear models, example models, dependency models, and relational variable models [2,14]. Machine learning methods include classification and regression trees that use data to develop prediction models [31]. The models are developed by dividing the data and fitting a prediction model within each division, allowing the results to be represented as a decision tree [31]. Classification trees are used for dependent variables that occur in specific data sets having prediction error measured in terms of the loss for incorrect classification [31]. Regression trees are used

for dependent variables that are in continuous or discrete data having prediction error measured by the sum of squares difference between the observed and predicted data [31]. Model selection decisions are a tradeoff between the most accurate or complex model that provides the best fit and the reliability of the fit with less ability of decision makers to understand the model [31]. Researchers are usually interested in complex models, while business users are usually more interested in less complex models for the reasons above [2,32,33]. The following is a discussion of constraints in the probability distribution [19]:

> In data mining contexts, the constraints are typically either supplied by human experts or automatically inferred from the database. For example, regression assumes a particular functional form for relating variables or, in the case of logistic regression, relating the values of some variables to the probabilities of other variables; but constraints are implicit in any prediction method that uses a database to adjust or estimate the parameters used in prediction. Other forms of constraint may include independence, conditional independence, and higher-order conditions on correlations (e.g., tetrad constraints). On average, a prediction method guaranteeing satisfaction of the constraints realized in the probability distribution is more accurate and has a smaller variance than a prediction method that does not. Finding the appropriate constraints to be satisfied is the most difficult issue in this sort of prediction. As with estimation, prediction can be improved by model averaging, provided the probabilities of the alternative assumptions imposed by the model are available.
>
> Another sort of prediction involves interventions that alter the probability distribution-as in predicting the values (or probabilities) of variables under a change in manufacturing procedures or changes in economic or medical treatment policies. Making accurate predictions of this kind requires some knowledge of the relevant causal structure and is generally quite different from prediction without intervention, although the same caveats about uncertainty and model averaging apply. (Glymour, C.; Madigan, D.; Pregibon, D.; Smyth, P., 1996) [19]

The success of a Knowledge Discovery in Databases process is influenced by model selection decisions [2]. Different models have different attributes that are considered more or less desirable by researchers from different fields [34]. Statistical researchers may decide to use models with less performance in order to insure the ability to interpret the model and determine meaning from the model [34]. Model preference is discussed in the following [19]:

> The evidence provided by data should lead us to prefer some models and hypotheses to others and to be indifferent about still other models. A *score* is any rule that maps models and data to numbers whose numerical ordering corresponds to a preference ordering over the space of models, given the data. For such reasons, scoring rules are often an attractive alternative to tests…..Typical rules assign to a model a value determined by the likelihood function associated with the model, the number of parameters, or dimension, of the model, and the data…..
>
> There is a notion of consistency appropriate to scoring rules; in the large sample limit, the true model should almost surely be among those receiving maximal scores….The probability (*p*) values assigned to statistics in hypothesis tests of models are scores, but it does not seem to be known whether and under what conditions they form a consistent set of scores. There are also uncertainties associated with scores, since two different samples of the same size from the same distribution can yield not only different numerical values for the same model but even different orderings of models. (Glymour, C.; Madigan, D.; Pregibon, D.; Smyth, P., 1996) [19]

3.10 Association Rules

Patterns must be evaluated for usefulness, an important and difficult necessity [30]. Association rules are useful analysis and prediction tools, derived from normal occurring situations in databases [29]. Unexpected association rules depart from the user's prior knowledge and interesting association rules affect the user's prior knowledge [15]. Informative patterns may be confused with incorrect conditions in a database [30]. Association rule learning systems have beneficial characteristics including the ease of user understanding and are superior to decision tree learning systems for a number of problems [35]. A disadvantage of rule learning systems is that as sample size increases or has noisy data, the results are less beneficial [35]. Techniques in modern rule learning systems originate from decision tree learning systems that use overfit and reduction on noisy data [35]. An initial hypothesis results in a complex tree that is reduced with the goal of improving errors in problems having noisy data [35]. Statistics defines underfit or bias as a tradeoff with overfit or variance [34]. Large data sets can reduce model overfit using subsets retained for model testing where results on test subsets decline signaling the need to end model determination [34]. Small data sets may require the use of all the

data for learning, needing new data to refine the fit [34]. Discovering all rules is more important than verifying a specific rule, verifying rules may cause unexpected rules and different conditions to be misunderstood [29].

3.11 Algorithms

Useful models describing data have been and continue to be an important method of advancing knowledge [34]. Researchers developing algorithms to extract rules explaining database characteristics are from a growing number of disciplines [34]. Search algorithms include either parameter search of an existing model or a model search in a model space [2]. Data mining algorithms use different levels of involvement of the model, the preference criteria, and the search algorithm [2]. The model is developed from the requirements for the function of the model, the representation method for the model, and characteristics derived from the data [2]. The preference criteria is the reasoning involved in considering one model superior over others, usually a fit determination of the model to a data set, a model not having overfit, or a model unconstrained by the data set [2]. The search algorithm is a specific algorithm searching for specific models and conditions with a preference criteria in a data set [2]. Databases and the development of data warehouses have created the need for data mining algorithms [5]:

> Database techniques for gaining efficient data access, grouping, and ordering operations when accessing data, and optimizing queries constitute the basics for scaling algorithms to larger data sets. Most data mining algorithms from statistics, pattern recognition, and machine learning assume data are in the main memory and pay no attention to how the algorithm breaks down if only limited views of the data are possible.
>
> A related field evolving from databases is *data warehousing*, which refers to the popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support. Data warehousing helps set the stage for KDD in two important ways: (1) data cleaning and (2) data access.
>
> **Data cleaning:** As organizations are forced to think about a unified logical view of the wide variety of data and databases they possess, they have to address the issues of mapping data to a single naming convention, uniformly

representing and handling missing data, and handling noise and errors when possible.

Data access: Uniform and well-defined methods must be created for accessing the data and providing access paths to data that were historically difficult to get to (for example, stored offline). (Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P., 1996) [5]

A data mining algorithm has three primary components [5]:

> One can identify three primary components in any data-mining algorithm: (1) model representation, (2) model evaluation, and (3) search….
>
> *Model representation* is the language used to describe discoverable patterns. If the representation is too limited, then no amount of training time or examples can produce an accurate model for the data….
>
> *Model-evaluation criteria* are quantitative statements (or *fit functions*) of how well a particular pattern (a model and its parameters) meets the goals of the KDD process….
>
> *Search method* consists of two components: (1) parameter search and (2) model search. Once the model representation (or family of representations) and the model-evaluation criteria are fixed, then the data-mining problem has been reduced to purely an optimization task: Find the parameters and models from the selected family that optimize the evaluation criteria. In parameter search, the algorithm must search for the parameters that optimize the model-evaluation criteria given observed data and a fixed model representation. Model search occurs as a loop over the parameter-search method: The model representation is changed so that a family of models is considered….
>
> Decision trees and rules that use univariate splits have a simple representational form, making the inferred model relatively easy for the user to comprehend….
>
> To a large extent, they depend on likelihood-based model-evaluation methods, with varying degrees of sophistication in terms of penalizing model complexity….
>
> These methods consist of a family of techniques for prediction that fit linear and nonlinear combinations of basis functions (sigmoids, splines, polynomials) to combinations of the input variables….
>
> The representation is simple: Use representative examples from the database to approximate a model; that is, predictions on new examples are derived from the properties of similar examples in the model whose prediction is known….
>
> Graphic models specify probabilistic dependencies...In its simplest form, the model specifies which variables are directly dependent on each other….
>
> Although decision trees and rules have a representation restricted to propositional logic, *relational learning* (also known as *inductive logic programming*) uses the more flexible pattern language of first-order logic….
>
> Understanding data mining and model induction at this component level clarifies the behavior of any data mining algorithm and makes it easier for the user to understand its overall contribution and applicability to the KDD process….
>
> The practical criteria for KDD projects are similar to those for other applications of advanced technology and include the potential impact of an application, the absence of simpler alternative solutions, and strong

organizational support for using technology. (Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P., 1996) [5]

Algorithms are not equal when applied to different problems, each algorithm is better or worse than other algorithms when applied to a specific problem [2]. An overall superior data mining algorithm does not exist, the choice of an algorithm for a specific problem requires experience and expertise [2]. Understanding the problem to be investigated can be a better use of time and resources than the search for an optimal data mining algorithm method [2].

3.12 Interestingness

Data mining projects provide value when discovering significant or interesting unknown patterns in the number of discovered patterns during the data investigation [36]. Novelty, utility, relevance, and statistical significance are considered to determine how interesting a discovered pattern is for the user of the knowledge discovery system [36]. The determination of interestingness is also affected by the value of answer deployment in the context of the considered business objective [36]. Interestingness can have a component of understandable discovered patterns, a simple pattern may have more value than a complex pattern [2].

3.13 Inference

Inference has qualities including: estimation, consistency, uncertainty, assumptions, robustness, and model averaging [19,37]. Inference is often involved in the estimation of a value either determining the domain of a data set or the characteristics of an object contained in a data set [19,37]. An estimator has the quality of consistency, when the available data increases to an upper limit, the estimator approaches the correct value of the estimated feature [19,37]. Inference contains uncertainty when estimating values of a limited sample, this uncertainty arises as a probability distribution of estimates if using same size samples obtained by same methods [19,37]. Statistics quantifies the

uncertainty when patterns are inferred from a specific data sample obtained from a larger data set [2]. Statistics contains uncertainty measures and the means to quantify the uncertainty for estimators [19,37]. Estimates are determined using a set of assumptions or model, however, the assumptions may not be correct [19,37]. If the assumptions or model is incorrect, the estimates should be incorrect, however, this is not certain [19,37]. Comparing current values in the data with previous values, change and deviation detection reveals the significant data changes [5]. Hypothesis testing usually does not contain any useful information [19]. Theory formation, hypothesis of new rules and unknown activity, separating needed information from entire data, and determination of hypotheses that need considerable problem information represents analysis with a greater opportunity for success [3].

3.14 Causation

Statistics provided the understanding of causation [19]. Inference of causation can contain many errors when using uncontrolled convenience samples [19]. The first important source of error is latent variables, a variable that is not captured in the data but changes and affects variables captured in the data [19]. Data mining should always consider the existence of latent variables since causal inference in the data can be determined where none exists [19]. The second important source of error is sample selection bias, creating causal inference without valid basis [19]. The third important source of error is model equivalence, data mining methods may select from equivalents indicating causal activity to users that does not exist [19]. All of the above sources of error are the reason regression methods provide inconsistent answers to data mining problems [19].

3.15 Data Reduction

Massive data sets need automated reduction to a manageable size for analysis [3]. Analysis of manageable data sets allow researchers to concentrate on activities machines are incapable of accomplishing [3]. These include creative data analysis, theory, hypothesis, and insight development [3]. The problem inherent in data reduction is valuable data required to understand unknown phenomena may not be available in the analysis [3]. Data reduction performs passes over the raw data using cataloging, classification, segmentation, and partitioning of data [3]. Data reduction is usually divisible into smaller independent problems [3]. A reduced data set can be analyzed using manual or statistical methods [3]. While domain constraints narrow the search, constraints may reduce the discovery of unexpected solutions [28]. Data relevant to a specific problem may be relatively small to a very large data set [19]. Data of important characteristics is required for a proper data set [28]. Missing data can result in the inability to compute values, incorrect data can cause incorrect results unknown to the analysts [38]. Data cleaning methods are the processes of identifying and correcting incomplete or incorrect information in databases [38].

3.16 Cautions

Data mining literature contains the following cautions [19]. Data mining estimation and search procedures should be consistent when compared to existing beliefs concerning conditions [19]. Uncertainty is a condition to reveal and use [19]. Take advantage of model averaging using search error calibration [19]. Conditioning and intervening should not be confused: the error probabilities of hypothesis tests are not the same as error probabilities of search procedures [19].

Chapter 4

Results

The literature search of automated data analysis including knowledge discovery and data mining terminology documented in the previous section provided the necessary information on background, concepts, components, and process models. In addition, the literature search resulted in the needed understanding of the critical decisions that may be encountered during the implementation of automated data analysis.

The standardized process models have been compared for best alternatives. Examples and feedback loops have been added to the derived steps. The questions and concerns of the Wyoming mineral extraction companies contacted are answered with an initiation requirements step. This implementation process has information and detail needed by these companies considering the use of automated data analysis. Critical issues listed after the prescribed steps provide highlighting of important decision making for these companies. While discussions with Wyoming mineral extraction companies have provided the major influence in development of the following implementation process, companies in other industries can have the same benefit from the information.

4.1 Implementation Process

Step 1. Initiation Requirements [18,20,39]

- Consider needed resources [18,20,39]

An initial automated data analysis project starts with understanding the company resources that will be required [18,20,39]. This includes recognizing the employees needed to support the project, employees needed on the project team, data access required for the team, time needed to meet and direct the project, time spent in each phase of the project, and expected duration of the project [18,20,39].

Time requirements are estimated as follows:

Relative percentages of effort estimated in literature for representative steps

1. Understanding of Business Objectives		15-20%
2. Understanding of Data		15-25%
3. Data Preparation		45-60%
4. Data Mining		10-18%
5. Evaluation of Results		8-15%
6. Deployment of Results		8-10% [6]

Time and percentages from an actual project

1. Business Objectives Determination	1 day	5%
2. Data Preparation	6 days	30%
3. Data Audit	1 day	5%
4. Interactive Data Mining and Results Analysis	3 days	15%
5. Back-end Data Mining	6 days	30%
6. Results Synthesis and Presentation	3 days	15% [1]

Duration of time from actual project timeline

Time from project initiation to project steps	2 months
Total time for project steps	3 months [1]

- Understand expertise sources [18]

Team members directing the project will represent the functions in the company the project will consider [18]. Employees familiar with current data collection and analysis will be needed on the project team [18]. Decisions are made about the use of data analysis experts [18].

- Evaluate data representation options [18]

A data analysis project outcome is enhanced by representation in a form that company

management has previously used to make decisions [18].

- Set realistic expectations [18,20]

It is important that initial projects are not promoted as having capabilities that may be desirable, but unknown [18,20]. The outcome of automated data analysis may uncover valuable unknown patterns, but may also be inconclusive [18,20].

- Realize number of iterations [18,20]

The project team will need to accept that a number of decisions can result in the requirement to return to a previous step [18,20]. These iterations are covered in the following steps and may have to be repeated more than once as the project proceeds [18,20].

- Understand project result of previous knowledge [14]

A common result of the project is the verification of previously known information [14]. Verification may be a goal of automated data analysis [14]. Other projects may consider this result as failure [14]. The decision should be made to report or not report previous knowledge by the project team as to avoid reducing confidence in the project [14].

- Accept outcome of no new knowledge [15]

No new knowledge may be discovered at the conclusion of the project [15]. This should not be considered failure [15]. New knowledge may not exist or the number of decisions required could have taken the project in an unsuccessful direction [15]. As the project team determines where unknown patterns and information do not exist there may be success in a different direction [15].

- Document actions and terminology [18]

A necessary and highly important activity of a successful automated data analysis project is the documentation of all actions, decisions, and alternatives made during each step [18]. The definition of automated data analysis terminology is documented [18].

Step 2. Business Understanding [2,16,18,20,39]

- Understand business objectives [2,16,18,20,39]

The project team must reach an understanding of the core company business objectives in order to expend effort in a direction with desired improvement potential [2,16,18,20,39].

- Determine factors impacting objectives [18]

The core company business objectives are usually affected by various factors [18]. These factors need to be identified and the details of each should be determined to understand the specific attributes of the factors [18]. This identification includes departments and key employees affected by the business objectives [18].

- Prioritize impact of factors [18]

The identified factors have relative impact on the business objectives [18]. The project team considers the best alternative for inclusion in a data analysis project [18].

- Determine data availability for factors [18]

The project team determines the data sources containing information concerning the business objective factors being considered [18]. Relevant and sufficient data sources are necessary to proceed [18]. This may require decisions until a business objective factor with the needed data source is determined [18].

- Make initial selection of data mining project [18]

The data mining component of the data analysis project is defined [18]. This requires a determination of assumptions, resources, and limitations inherent in the data mining project [18]. Data mining cost is compared to business objective improvement savings [18].

- Define data mining project goals [2,16,18,20]

Business objective goals are transformed into data mining goals [2,16,18,20]. Data mining goals will express the business objective goals in terms of information that may be

contained in the data analyzed [2,16,18,20].

- Determine previous knowledge [18]

Company activity experts are requested to list existing knowledge about the impact of the factors on the business objectives in the data mining project [18]. Previous knowledge can indicate a direction with potential and reduce effort in a direction without potential [18].

- Set tasks to accomplish data mining project [18,39]

The project team sets the steps necessary for the data mining project implementation [18,39]. This includes a determination of time, resources, analysis tools, and evaluation methods required for each step [18,39]. The data mining project tasks are evaluated after each step and adjustments are made to the tasks as needed [18,39].

Step 3. Data Understanding [18,20]

- Assess data mining project data availability and accessibility [18,39]

The selected data mining project data is acquired [18,39]. A record is made of the location and other attributes of the data for future reference [18,39]. These attributes include method of storage, records and fields, terminology, and definitions of the data [18,39].

- Identify data quality for completeness, redundancy, missing values, and plausibility [16,18,20]

The need for data preparation may become evident [16,18,20].

- Select a subset of data with sufficient records or fields [2,18]

A subset of project data may be useful at this stage if dealing with an unmanageable number of data items for understanding [2,18].

- Decide data usefulness for data mining project [16,18,20]

The relevance, quantity, and quality of the data are assessed for ability to progress with

the selected project [16,18,20].

- If data or data understanding is insufficient, return to Business Understanding to select different data mining project [18,20]

This iteration is needed when data understanding uncovers a problem using the current selected data for the proposed purpose [18,20].

Step 4. Data Preparation [2,16,18,20,39]

- Use data mining goals to select data mining modeling [2,18,20,39]

Data mining goals are transformed into the required data mining modeling [2,18,20,39]. Data mining models are selected for ability to produce the intended project goals and present the outcome in a desired form [2,18,20,39].

Data mining model examples include:

Decision tree classifiers [40]

Cluster analysis [40]

Association rules [40]

Link analysis [40]

Classification and regression trees [40]

- Accomplish improvement of data for acceptability in data mining modeling [2,18,20,39]

The data to be used or eliminated are determined from the data available [2,18,20,39]. Decisions concerning any data cleaning or constructed data including derived attributes, generated records, merged data, or formatting changes are made [2,18,20,39].

- If specific data knowledge is needed for data preparation, return to Data Understanding [20]

This iteration is needed when decisions concerning Data Preparation indicate Data Understanding should be improved [20].

Step 5. Data Mining [2,16,18,20,39]

- Select data mining tool or tools producing desired modeling [2,16,18,20,39,40]

Modeling determinations are used to select data mining tools [2,16,18,20,39].

Data mining tool or algorithm examples for previous data mining model examples include:

C4.5 - decision tree classifiers, flowchart, supervised learning, continuous and discrete [40]

k-means - cluster analysis, sensitive to outliers and initial choice of centroids, unsupervised [40]

Apriori - association rules applied to large data numbers, correlations and relations, unsupervised [40]

PageRank - link analysis, relative importance in a network, unsupervised [40]

CART - classification and regression trees, decision trees return class, regression trees return number, supervised [40]

- Develop the model or models from the data [2,16,18,20,39]

The data mining tool may require the data set to have specific attributes [2,16,18,20,39]. The data set considered ready for use in the data mining tool should be separated into train and test sets [2,16,18,20,39]. The model is built using the train set then quality and validity are determined by the test set [2,16,18,20,39]. After the data and model are considered acceptable, the selected data set is used in the data mining tool to produce a model or models 2,16,18,20,39].

- Assess the produced model or models [18,20]

Using company knowledge, business improvement direction, and data mining goals, the produced model or models are evaluated for accuracy and quality [18,20].

- Refine iterations of a model or models until the results are considered optimal [18]

Time spent refining a model or models will achieve improvement [18].

- If data mining is not satisfactory, return to Business Understanding for goal modification [20]

Data mining goals may need adjustment for satisfactory modeling [20].

- If the data was not correctly characterized, return to Data Understanding [20]

Data usefulness may need improvement for satisfactory modeling [20].

- If model development may be improved, return to Data Preparation [20]

Decisions made concerning data acceptability may require review [18,20].

Step 6. Results Evaluation and Back-end Data Mining [2,16,18,20,39]

- Interpret the results assessing impact, novelty, and interestingness of the discovered knowledge [2,16,18,20,39]

Model or models are approved as relevant to business objectives [2,16,18,20,39]. Evaluation is made of new information that may be contained in the model or models [2,16,18,20,39].

- Identify possible problems in data mining application [18,20]

Determine if the model or models were properly built, correct data attributes used and future availability of data attributes [18,20].

- If evaluation unsatisfactory, return to Business Understanding for selection of different data mining project [18,20]

The project team may decide the data mining project is unacceptable for deployment and a new project is required [18,20].

- If evaluation determines the data mining project does not produce new knowledge, return to Data Mining for different modeling tool [18,20]

A different modeling tool may reveal knowledge not evident with the modeling tool applied [18,20].

Step 7. Knowledge Presentation [18]

- Present results to company management in documentation and reports determined to maximize understanding of results and potential impact [16,18]

The data mining results are presented to the company decision makers for decisions on possible use [16,18].

- Explore conflicts with existing knowledge [2,18]

Determine explanation for conflicts of project results with previous knowledge prior to deployment [2,18].

Step 8. Knowledge Deployment [16,18,20]

- Determine plan for use and implementation of knowledge [16,18,20]

A plan of scope and steps is developed for introduction of data mining project results into company activities [16,18,20].

- Assess exploitation of discovered knowledge [18]

Introduction of data mining project results is evaluated for improvement of plan [18].

- Monitor and review knowledge implementation [18]

Implementation of data mining project results is monitored using a specific plan and reviewed for correct application maintenance [18,20].

4.2 Critical Issues

Knowledge Discovery Goals [5,14]

- The two major goals of knowledge discovery are verification and discovery [5,14]. Discovery is further divided into prediction and description [5,14]. Considerations include relevance of attributes, high noise, and prior knowledge [14].

Data Mining Focus [1,2,5,8,14]

- The application of data mining tools should not become the focus of the automated data analysis project [2,5,14].
- The data mining step is valuable when used to achieve a specific business objective [1].
- Data mining can provide understandable presentations of results for company decisions [8].

Data Mining Approaches [8,23,27]

- Data mining software tool approach requires more expertise in data mining methods, databases, and statistics [8,23].
- Data mining application systems approach is more understandable by company users and decision makers [8,27].

Human-centric and Data-centric Models [6]

- Human-centric models use a data analyst during the data analysis project to enhance discovery through intelligent, visual, and perceptional tools [6].
- Data-centric models use a more iterative and interactive approach to the data analysis project [6].

Model Fitting [2,5,14,31,32,33]

- Two mathematical approaches are used in model fitting [5].

- Statistical model fitting allows for nondeterministic effects [5].
- Logical model fitting is entirely deterministic [5].
- Common model functions in data mining projects are classification, regression, clustering, summarization, dependency modeling, link analysis, sequence analysis, change, and deviation detection [2,5,14]. Primary components of a data mining algorithm are model representation, model evaluation criteria, and search method [14].
- Complex models may provide a closer fit of the data [31].
- Simpler models may provide less fit of the data, but better understanding [2,32,33].

Association Rules [15,29,30,34,35]

- Defining informative patterns is a difficult but important problem [30].
- Association rules are a set of database regularities [29].
- Rules are compared to beliefs and unexpectedness of a rule is interesting if it affects company beliefs of the project team [15].
- Rule sets are easy for the project team to understand, but may not scale easily on noisy data sets [35].
- The tradeoff between model underfit and overfit contained in statistical investigation is a consideration [34].

Algorithms [2]

- Algorithms have been designed to perform a parameter search of a model or a model search over model space [2].
- Data mining algorithms consider the relative use of the model, preference criterion, and search algorithm [2].
- The model considers the function of the model, the representational

form, and characteristics of the data [2].

- Preference criterion is the reason for considering one model acceptable over others [2].
- Search algorithm is an algorithm used to find specific models using performance criterion in data [2].

Interestingness [36]

- The data analysis project team determines the discovered patterns that have the most potential for impact [36].
- Impact is affected by novelty, utility, relevance, and statistical significance of the detected pattern in relation to the estimated benefit for the business objective [36].

Inference [19,37]

- Inference considers important features of model determination [19,37].
- These important features include estimation, consistency, uncertainty, assumptions, robustness, and model averaging [19,37].

Causation [19]

- Data mining project results have the same difficulties as statistics in assigning causation [19].

Data Reduction [3]

- Creating manageable data for analysis is an understandable goal for automated data analysis projects [3].
- The use of automated data analysis to reduce data should consider the loss of new knowledge if the data is not completely understood [3].

Cautions [19]

- Estimation and search procedures must be consistent with conditions

thought to exist in applications [19].

- Uncertainty should be revealed and used [19].
- Errors of search are calibrated for honesty and advantages of model averaging [19].
- Conditioning is not intervening [19].

Chapter 5

Conclusion

Companies in industries not usually associated with the use of automated data analysis are experiencing an increase in data collection and storage. Current articles on the Internet of Industrial Things by Rahul Vijayaraghavan contain evidence of this change.

The first article (Vijayaraghavan, 2015) states:

> The Internet of Industrial Things (IoIT), a concept derived from the ability to connect assets, business processes, and people across the enterprise, has transformed the plant floor into a state of hyper-connectivity. However, with greater connectivity comes a drastic increase in the inflow of data. In addition to traditional structured data, the manufacturing sector is facing a spike in semi-structured and unstructured data from sensors, machines, Web, and social media. In the current setup, it is therefore imperative for end users to append new data management tools to store and process the large influx of data as well as utilize state-of-the-art analytical platforms to derive actionable insights for core operations in the facility. [41]

The second article (Vijayaraghavan, 2015) states:

> A new report anticipates a dramatic jump in the need for data analysis as the Industrial Internet of Things continues to expand.
>
> The analysis by consulting firm Frost & Sullivan expects demand for predictive solutions from connected devices to increase at a compound annual growth rate of 57 percent between 2014 and 2021.
>
> The study said that the manufacturing, technology, automotive, aerospace, life sciences, and food and beverage sectors accounted for 14 percent of global stored data, and that companies in those industries are experiencing spikes in unstructured data from equipment, sensors and social media.
>
> In order to address the influx of data and derive information that will be critical to making decisions, analysts said, companies will need to add new management tools. The systems should also be customizable and easily accessible and intuitive to a wide range of employees. [42]

This research did find increasing data collection and storage in Wyoming mineral extraction companies. This research did not find these same companies using advanced data discovery methods. Increased knowledge about the company activities can be lost without the use of automated data analysis. A common need exists for these companies

when considering the use of advanced data discovery tools. These Wyoming mineral extraction companies require an implementation process for initial use of automated data analysis. The absence of a comprehensive implementation process is delaying the use of automated data analysis by companies in Wyoming. This research introduces an implementation process that provides initial requirements and adds necessary information within the process for these Wyoming mineral extraction companies wanting to consider the initiation of automated data analysis.

Chapter 6

Summary

Sixteen Wyoming companies were contacted concerning their use of automated data analysis. These companies confirmed the existence of an increasing collection and storage of data. These companies were not using automated data analysis. All of the companies contacted were very interested in the use of automated data analysis and the possible determination of unknown patterns in their existing data. A brief discussion of automated data analysis was followed with company questions about the requirements needed for project initiation. Wyoming mineral extraction companies needed a comprehensive implementation process for automated data analysis. The results would advance the use of the technology by these Wyoming companies. A literature search for automated data analysis, including knowledge discovery and data mining terminology, was conducted. This literature search documented the background, concepts, components, process models, and critical issues contained in the literature. A comprehensive implementation process for automated data analysis was developed for Wyoming mineral extraction companies. The following implementation process overview includes information and detail needed by these companies considering the use of automated data analysis.

6.1 Implementation Process Overview

Step 1. Initiation Requirements [18,20,39]

- Consider needed resources [18,20,39]
- Understand expertise sources [18]
- Evaluate data representation options [18]
- Set realistic expectations [18,20]
- Realize number of iterations [18,20]

- Understand project result of previous knowledge [14]
- Accept outcome of no new knowledge [15]
- Document actions and terminology [18]

Step 2. Business Understanding [2,16,18,20,39]

- Understand business objectives [2,16,18,20,39]
- Determine factors impacting objectives [18]
- Prioritize impact of factors [18]
- Determine data availability for factors [18]
- Make initial selection of data mining project [18]
- Define data mining project goals [2,16,18,20]
- Determine previous knowledge [18]
- Set tasks to accomplish data mining project [18,39]

Step 3. Data Understanding [18,20]

- Assess data mining project data availability and accessibility [18,39]
- Identify data quality for completeness, redundancy, missing values, and plausibility [16,18,20]
- Select a subset of data with sufficient records or fields [2,18]
- Decide data usefulness for data mining project [16,18,20]
- If data or data understanding insufficient, return to Business Understanding to select different data mining project [18,20]

Step 4. Data Preparation [2,16,18,20,39]

- Use data mining goals to select data mining modeling [2,18,20,39]
- Accomplish improvement of data for acceptability in data mining modeling [2,18,20,39]

- If specific data knowledge is needed for data preparation, return to Data Understanding [20]

Step 5. Data Mining [2,16,18,20,39]

- Select data mining tool or tools producing desired modeling [2,16,18,20,39]
- Develop the model or models from the data [2,16,18,20,39]
- Assess the produced model or models [18,20]
- Refine iterations of a model or models until the results are considered optimal [18]
- If data mining is not satisfactory, return to Business Understanding for goal modification [20]
- If the data was not correctly characterized, return to Data Understanding [20]
- If model development may be improved, return to Data Preparation [20]

Step 6. Results Evaluation and Back-end Data Mining [2,16,18,20,39]

- Interpret the results assessing impact, novelty, and interestingness of the discovered knowledge [2,16,18,20,39]
- Identify possible problems in data mining application [18,20]
- If evaluation unsatisfactory, return to Business Understanding for selection of different data mining project [18,20]
- If evaluation determines the data mining project does not produce new knowledge, return to Data Mining for different modeling tool [18,20]

Step 7. Knowledge Presentation [18]

- Present results to company management in documentation and

reports determined to maximize understanding of results and potential impact [16,18]

- Explore conflicts with existing knowledge [2,18]

Step 8. Knowledge Deployment [16,18,20]

- Determine plan for use and implementation of knowledge [16,18,20]
- Assess exploitation of discovered knowledge [18]
- Monitor and review knowledge implementation [18]

6.2 Critical Issues Overview

Knowledge Discovery Goals [5,14]

Data Mining Focus [1,2,5,8,14]

Data Mining Approaches [8,23,27]

Human-centric and Data-centric Models [6]

Model Fitting [2,5,14,31,32,33]

Association Rules [15,29,30,34,35]

Algorithms [2]

Interestingness [36]

Inference [19,37]

Causation [19]

Data Reduction [3]

Cautions [19]

References

[1] K. Hirji, “Exploring data mining implementation,” *Commun. ACM,* vol. 44, no. 7, pp. 87-93, 2001.

[2] U. Fayyad et al., “The KDD process for extracting useful knowledge from volumes of data,” *Commun. ACM*, vol. 39, no. 11, pp. 27-34, 1996.

[3] U. M. Fayyad et al., “KDD for science data analysis: Issues and examples,” in *Proc. 2nd. Int. Conf. on Knowledge Discovery and Data Mining (KDD-96),* Menlo Park, CA, 1996.

[4] M. Goebel and L. Gruenwald, “A survey of data mining knowledge discovery software tools,” *SIGKDD Explorations,* vol. 1, no. 1, pp. 20-33, 1999.

[5] U. Fayyad et al., “From data mining to knowledge discovery in databases,” *AI Mag.,* vol. 17, no. 3, pp. 37-54, 1996.

[6] L. A. Kurgan and P. Musilek, “A survey of knowledge discovery and data mining process models,” *The Knowledge Eng. Review,* vol. 21, no. 1, pp. 1-24, 2006.

[7] M. Bohanec, “What is Decision Support?” in *Proc. of Inform. Soc. IS- 2001: Data Mining and Decision Support in Action,* Slovenia, 2001, pp. 86-89.

[8] M. Kukar and R. Rupnik, “Decision support system to support decision processes with data mining,” *J. of Informational and Organizational Sciences*, vol. 31, no. 1, pp. 217-232, 2007.

[9] M. Bohanec, “Decision support,” in *Data Mining and Decision Support: Integration and Collaboration,* M. Bohanec et al., Eds., Dordrecht, The Netherlands: Kluwer Academic, 2003, pp. 23-35.

[10] JSR-73 Expert Group. “Java Specification Request 73: Java Data Mining (JDM),” Java Community Process, 2004.

[11] N. Lavrac and M. Grobelnik, “Data mining,” in *Data Mining and Decision Support: Integration and Collaboration*, M. Bohanec et al., Eds., Dordrecht, The Netherlands: Kluwer Academic, 2003, pp. 3-14.

[12] K. J. Cios et al., “The knowledge discovery process,” in *Data mining: A Knowledge Discovery Approach,* Springer, NY, 2007, pp. 9-24.

[13] R. Brachman and T. Anand, “The process of knowledge discovery in databases: A human-centered approach,” in *Advances in Knowledge Discovery and Data Mining,* U. Fayad et al., Eds., Cambridge, MA: MIT Press, 1996, pp. 37-58.

[14] U. Fayyad et al., “Knowledge discovery and data mining: Towards a unifying framework,” in *KDD-96 Proc.,* Menlo Park, CA, 1996, pp. 82-88.

[15] B. Padmanabhan and A. Tuzhilin, “A belief-driven method for discovering unexpected patterns,” in *KDD-98 Proc.,* Menlo Park, CA, 1998, pp. 94-100.

[16] P. Cabena et al., *Discovering data mining from concept to implementation,* Englewood Cliffs, NJ: Prentice Hall, 1998.

[17] C. Clifton and B. Thuraisingham, ”Emerging standards for data mining,” *Computer Standards and Interfaces,* vol. 23, pp. 187-193, 2001.

[18] P. Chapman et al., “CRISP-DM (1.0),” SPSS, 1999.

[19] C. Glymour et al., “Statistical inference and data mining,” *Commun. ACM,* vol. 39, no. 11, pp. 35-41, 1996.

[20] K. Cios and L. Kurgan, “Trends in data mining and knowledge discovery,” in *Advanced Techniques in Knowledge Discovery and Data Mining,* N. Pal and L. Jain, Eds., Springer, NY, 2005, pp. 1-26.

[21] E. Frank and I. H. Witten, “Filtering algorithms,” in *Data Mining: Practical Machine Learning Tools and Techniques,* San Francisco, CA: Morgan Kaufmann, 2005, pp. 393-402.

[22] R. Agrawal and G. Psaila, “Active data mining,” in *KDD-95 Proc.,* Menlo Park, CA, 1995, pp. 3-8.

[23] R. Kohavi and M. Sahami, “KDD-99 panel rep.: Data mining into vertical solutions,” *SIGKDD Explorations,* vol. 1, no. 2, pp. 55-58, 2000.

[24] M. Holsheimer, “Data mining by business users: integrating data mining in business process,” in *Proc. Int. Conference on Knowledge Discovery and Data Mining KDD-99,* 1999, pp. 266-291.

[25] R. Cooley et al., “Web usage mining: Discovery and applications of usage patterns from web data,” *SIGKDD Explorations,* vol. 1, no. 2, pp. 12-23, 2000.

[26] R. Rupnik and M. Kukar, “Data mining and decision support: An integrative approach,” in *Decision Support Syst.*, C. S. Jao, Ed., Vukovar, Croatia: InTech, 2010, pp. 63-86.

[27] C. Aggarwal, “Towards effective and interpretable data mining by visual interaction,” *SIGKDD Explorations,* vol. 3, no. 2, pp. 11-22, 2002.

[28] G. Piatetsky-Shapiro, “Knowledge discovery in real databases: A rep. on the IJCAI-89 workshop,” *AI Mag.,* vol. 11, no. 5, pp. 68-70, 1991.

[29] R. Agrawal et al., “Fast discovery of association rules,” in *Advances in Knowledge Discovery and Data Mining,* U. Fayyad et al., Eds., Menlo Park, CA: AAAI Press, 1996, pp. 307-328.

[30] O. Guyon et al., “Discovering informative patterns and data cleaning,” in *Workshop on Knowledge Discovery in Databases (AAAI Tech. Rep. WS-94-03),* 1994, pp. 145-156.

[31] W-Y. Loh, “Classification and regression trees,” *WIREs Data Mining Knowledge Discovery*, vol. 1, pp. 14-23, 2011.

[32] G. Wu. (2013). *Why more data and simple algorithms beat complex analytics models* [online]. Available: http://data-informed.com/why-more-data-and-simple-algorithms-beat-complex-analytics-models

[33] T. Anand and R. J. Brachman, "The process of knowledge discovery in databases: A first sketch," in *Workshop on Knowledge Discovery in Databases. (AAAI Tech. Rep. WS-94-03),* 1994, pp. 1-11.

[34] J. Elder and D. Pregibon, "A statistical perspective on KDD," in *KDD-95 Proc.: Advances in Knowledge Discovery and Data Mining,* U. Fayyad et al., Eds., Cambridge, MA: AAAI Press, 1996, pp. 87-93.

[35] W. Cohen, "Fast effective rule induction," in *Proc. Twelfth Int. Conference on Mach. Learning*," Lake Tahoe, CA, 1995, pp. 115-123.

[36] G. Piatetsky-Shapiro and C. Matheus, "The interestingness of deviations," in *Workshop on Knowledge Discovery in Databases (AAAI Tech. Rep. WS-94-03),* 1994, pp. 25-36.

[37] C. Chatfield, "Model uncertainty: Data mining and statistical inference," *J. Royal Statistical Soc.. Series A (Stat. in Soc.),* vol. 158, no. 3, pp. 419-466, 1995.

[38] R. Kerber et al., "Using recon for data cleaning," in *KDD-95 Proc.* Menlo Park, CA, 1995, pp. 282-287.

[39] S. Anand and A. Buchner, *Decision support using data mining*, London, United Kingdom: Financial Times Management.

[40] J. Ghosh et al., "Top 10 algorithms in data mining," *Knowledge and Inform. Syst.*, vol. 14, no. 1, pp. 1-37, 2008.

[41] Frost & Sullivan. (2015, August). *Manufacturing Domain Will Witness an Upsurge in Advanced Mach. Learning Solutions* [online]*.* Available:

http://ww2.frost.com/news/press-releases/manufacturing-domain-will-witness-upsurge-advanced-machine-learning-solutions

[42] A. Szal. (2015). *Rep.: IoT to generate increase in demand for predictive data* [online]. Available: http://www.impomag.com/news/2015/08/report-iot-generate-increase-demand-predictive-data

Biographical Information

Richard Allen Leach was born in Fort Worth, Texas on June 18, 1949. He received a Bachelor of Arts in Government from The University of Texas at Austin in 1974. He continued his education receiving a Bachelor of Science in Industrial Engineering in 1993 and a Master of Science in Industrial Engineering in 1997 from The University of Texas at Arlington. Richard was last employed as the Intellectual Property Manager for The University of Texas at Arlington. He is currently self-employed as a consultant for organizations and companies in the southwest area of Wyoming.