# Chapter 1
# Data Mining and Information Systems: Quo Vadis?

Robert Stahlbock, Stefan Lessmann, and Sven F. Crone

## 1.1 Introduction

Information and communication technology has been a steady source of innovations which have considerably impacted the way companies conduct business in the digital as well as the physical world. Today, information systems (IS) holistically support virtually all aspects of corporations and nonprofit institutions, along internal processes from purchasing and operations management toward sales, marketing, and eventually the customer (horizontally along the supply chain), from these operational functions toward finance, accounting, and upper management activities (vertically across the hierarchy) and externally to collaborate with external partners, suppliers, or customers. The holistic support of internal business processes and external relationships by means of IS has, in turn, led to the vast growth of internal and external data being stored and processed within corporate environments.

The progressive gathering of very large and heterogeneous data sets, accompanied by the increasing computational power and evolving database technology, summoned an increasing interest in data mining (DM) as a (novel) tool for discovering knowledge in data. In addition to technological advances, the success of DM – at least in corporate environments – can also be attributed to changes in the business environment. For example, increasing competition through the advent of electronic commerce and the removal barriers for new market entrants, more informed and thus

Robert Stahlbock
Institute of Information Systems, University of Hamburg, Von-Melle-Park 5, D-20146 Hamburg, Germany; Lecturer at the FOM University of Applied Sciences, Essen/Hamburg, Germany,
e-mail: stahlbock@econ.uni-hamburg.de

Stefan Lessmann
Institute of Information Systems, University of Hamburg, Von-Melle-Park 5, D-20146 Hamburg, Germany, e-mail: lessmann@econ.uni-hamburg.de

Sven F. Crone
Centre for Forecasting, Lancaster University Management School, Lancaster LA1 4YX, UK,
e-mail: sven.f.crone@crone.de

R. Stahlbock et al. (eds.), *Data Mining,* Annals of Information Systems 8,
DOI 10.1007/978-1-4419-1280-0_1, 

demanding customers as well as increasing saturation in many markets created a need for enhanced insight, understanding, and actionable plans that allow companies to systematically manage and deepen customer relationships (e.g., insurance companies identifying those individuals most likely to purchase additional policies, retailers seeking those customers most likely to respond to marketing activities, or banks determining the creditworthiness of new customers). The corresponding developments in the areas of corporate data warehousing, computer-aided planning, and decision support systems constitute some of the major topics in the discipline of IS.

As deriving knowledge from data has historically been a statistical endeavor [22], it is not surprising that size of data sets is emphasized as a constituting factor in many definitions of DM (see, e.g., [3, 7, 20, 24]). In particular, traditional tools for data analysis had not been designed to cope with vast amounts of data. Therefore, the size and structure of the data sets naturally determined the topics that emerged first, and early activities in DM research concentrated mainly on the development and advancement of highly scalable algorithms. Given this emphasis on methodological issues, many contributions to the advancement of DM were made by statistics, computer science, and machine learning, as well as database technologies. Examples include the well-known Apriori algorithm for mining associations and identifying frequent itemsets [1] and its many successors, procedures for solving clustering, regression, and time series problems, as well as paradigms like ensemble learning and kernel machines (see [52] for a recent survey regarding the top-10 DM methods). It is important to note that data set size does refer not only to the number of examples in a sample but also to the number of attributes being measured per case. Particularly applications in the medical sciences and the field of information retrieval naturally produce an extremely large number of measurements per case, and thus very high-dimensional data sets. Consequently, algorithms and induction principles were needed which overcome the curse of dimensionality (see, e.g., [25]) and facilitate processing data sets with many thousands of attributes, as well as data sets with a large number of instances at the same time. As an example, without the advancements in statistical learning [45–47], many applications like the analysis of gene expression data (see, e.g., [19]) or text classification (see, e.g., [27, 28]) would not have been possible. The particular impact of related disciplines – and efforts to develop DM as a discipline in its own right – may also be seen in the development of a distinct vocabulary within similar taxonomies; DM techniques are routinely categorized according to their primary objective into predictive and descriptive approaches (see, e.g., [10]), which mirror the established distinction of supervised and unsupervised methods in machine learning. We are not in a position to argue whether DM has become a discipline in its own right (see, e.g., the contributions by Hand [22, 21]). At least, DM is an interdisciplinary field with a vast and nonexclusive list of contributors (although many contributors to the field may not consider themselves "data miners" at all, and perceive their developments solely within the frame of their own established discipline).

The discipline of IS however, it seems, has failed to leave its mark and make substantial contributions to DM, despite its apparent relevance in the analytical support of corporate decisions. In accordance with the continuing growth of data, we are

able to observe an ever-increasing interest in corporate DM as an approach to analyze large and heterogeneous data sets for identifying hidden patterns and relationships, and eventually discerning actionable knowledge. Today, DM is ubiquitous and has even captured the attention of mainstream literature through best sellers (e.g., [2]) that thrive as much on the popularity of DM as on the potential knowledge one can obtain from conventional statistical data analysis. However, DM has remained focused on methodological topics that have captured the attention of the technical disciplines contributing to it and selected applications, routinely neglecting the decision context of the application or areas of potential research, such as the use of company internal data for DM activities. It appears that the DM community has primarily developed independently without any significant contributions from IS. The discipline of IS continues to serve as a mediator between management and computer science, driving the management of information at the interface of technological aspects and business decision making. While original contributions on methods, algorithms, and underlying database structure may rightfully develop elsewhere, IS can make substantial contributions in bringing together the managerial decision context and the technology at hand, bridging the gap between real-world applications and algorithmic theory.

Based on the framework provided in this brief review, this special issue seeks to explore the opportunities for innovative contributions at the interface of IS with DM. The chapters contained in this special issue embrace many of the facets of DM as well as challenging real-world applications, which, in turn, may motivate and facilitate the development of novel algorithms – or enhancements to established ones – in order to effectively address task-specific requirements. The special issue is organized into six sections in order to position the original research contributions within the field of DM it aims to contribute to: confirmatory data analysis (one chapter), knowledge discovery from supervised learning (three chapters), classification analysis (four chapters), hybrid DM procedures (four chapters), web mining (two chapters), and privacy-preserving DM (two chapters). We hope that the academic community as well as practitioners in the industry will find the 16 chapters of this volume interesting, informative, and useful.

## 1.2 Special Issues in Data Mining

### *1.2.1 Confirmatory Data Analysis*

In their seminal paper, Fayyad et al. [10] made a clear and concise distinction between DM and the encompassing process of knowledge discovery in data (KDD), whereas these terms are mainly used interchangeably in contemporary work. Still, the general objective of identifying novel, relevant, and actionable patterns in data (i.e., knowledge discovery) is emphasized in many, if not all, formal definitions of

DM. In contrast, techniques for confirmatory data analysis (that emphasize the reliable confirmation of preconceived ideas rather than the discovery of new ones) have received much less attention in DM and are rarely considered within the adjacent communities of machine learning and computer science. However, techniques such as structural equation modeling (SEM) that are employed to verify a theoretical model of cause and effect enjoy ongoing popularity not only in statistics and econometrics but also in marketing and information systems (with the most popular models being LISREL and AMOS). The most renowned example in this context is possibly the application of partial least squares (PLS) path modeling in Davis' famous technology acceptance model [9]. However, earlier applications of causal modeling predominantly employed relatively small data sets which were often collected from surveys.

Recently, the rapid and continuing growth of data storage paired with internet-based technologies to easily collect user information online facilitates the use of significantly larger volumes of data for SEM purposes. Since the underlying principles for induction and estimation of SEM are similar to those encountered in other DM applications, it is desirable to investigate the potential of DM techniques to aid SEM in more detail. In this sense, the work of Ringle et al. [41] serves as a first step to increase the awareness of SEM within the DM community. Ringle et al. introduce finite-mixture PLS as a state-of-the-art approach toward SEM and demonstrate its potential to overcome many of the limitations of ordinary PLS. The particular merit of their approach originates from the fact that the possible existence of subgroups within a data set is automatically taken into account by means of a latent class segmentation approach. Data clusters are formed, which are subsequently examined independently in order to avoid an estimation bias because of heterogeneity. This approach differs from conventional clustering techniques and exploits the hypothesized relationships within the causal model instead of finding segments by optimizing some distance measure of, e.g., intercluster heterogeneity. The possibility to incorporate ideas from neural networks or fuzzy clustering into this segmentation step has so far been largely unexplored and therefore represents a promising route toward future research at the interface of DM and confirmatory data analysis.

### *1.2.2 Knowledge Discovery from Supervised Learning*

The preeminent objective of DM – discovering novel and useful knowledge from data – is most naturally embodied in the unsupervised DM techniques and their corresponding algorithms for identifying frequent itemsets and clusters. In contrast, contributions in the field of supervised learning commonly emphasize principles and algorithms for constructing predictive models, e.g., for classification or regression, where the quality of a model is assessed in terms of predictive accuracy. However, a predictive model may also fulfill objectives concerned with "knowledge discovery" in a wider sense, if the model's underlying rules (i.e., the relationships discerned from data) are made interpretable and understandable to human decision makers.

Whereas a vast assortment of valid and reliable statistical indicators has been developed for assessing the accuracy of regression and classification models, an objective measurement of model comprehensibility remains elusive, and its justification a nontrivial undertaking. Martens and Baesens [36] review research activities to conceptualize comprehensibility and further extend these ideas by proposing a general framework for acceptable prediction models. Acceptability requires a third constraint besides accuracy and comprehensibility to be met. That is, a model must also be in line with domain knowledge, i.e., the user's belief. Martens and Baesens refer to such accordance as justifiability and propose techniques to measure this concept.

The interpretability of DM procedures, and classification models in particular, is also taken up by Le Bras et al. [31]. They focus on rule-based classifiers, which are commonly credited for being (relatively easily) comprehensible. However, their analysis emphasizes yet another important property that a prediction model has to fulfill in the context of knowledge discovery: its results (i.e., rules) have to be interesting. In this sense, the concept of interestingness complements Martens and Baesens [36] considerations on adequate and acceptable models. And although issues of measuring interestingness have enjoyed more attention in the past (see, e.g., Freitas [14], Liu et al. [34], and the recent survey by Geng and Hamilton [17]), designing respective measures remains as challenging as in the case of comprehensibility and justifiability. Drawing on the wealth of well-developed approaches in the field of association rule mining, Le Bras et al. consider so-called associative classifiers which consist of association rules whose consequent part is a class label. Two key statistics in association rule mining are support and confidence, which measure the number of cases (i.e., the database transactions in association rule mining) that contain a rule's antecedent and consequent parts and the number of cases that contain the consequent part among those containing the antecedent part, respectively. In that sense, support and confidence may be interpreted as measures of a rule's interestingness. In addition, these figures are of pivotal importance for the task of developing efficient rule induction algorithms. For the case of associative classification, it has been shown that the confidence measure possesses the so-called universal existential upward closure property, which facilitates a fast top-down derivation of classification rules. Le Bras et al. generalize this measure and provide necessary and sufficient conditions for the existence of this property. Furthermore, they demonstrate that several alternative measures of rule interestingness also exhibit general universal existential upward closure. This is important because the suitability of interestingness measures depends upon the specific requirements of an application domain. Therefore, the contribution of Le Bras et al. will allow users to select from a broad range of measures of a rule's interestingness, and to develop tailor-made ones, while maintaining the efficiency and feasibility of a rule mining algorithm.

The field of logic mining represents a special form of classification rule mining in the sense that the resulting models are expressed as logic formulas. As this type of model representation may again be seen as particularly easy to interpret, logic mining techniques represent an interesting candidate for knowledge discovery in general, and for resolving classification problems that require comprehensible

models in particular. A respective approach, namely the box-clustering technique, is considered by Felici et al. [11]. Box clustering offers the advantage that preprocessing activities to transform a data set into a logical form, as required by any logic mining technique, are performed implicitly. Although logic mining in general and box clustering in particular are appealing due to their inherent model comprehensibility, they also suffer from an important limitation: algorithms to construct a model from empirical data are less developed than for alternative classifiers. In particular, methodologies and best practices for avoiding the well-known problem of overfitting are very mature in the case of, e.g., support vector machines (SVMs) or artificial neural networks (ANNs). On the contrary, overfitting remains a key challenge in box clustering. To overcome this problem, Felici et al. propose a bi-criterion procedure to select the best box-clustering solution for a given classification problem and balance the two goals of having a predictive and at the same time simple model. Therefore, these procedures can be seen as an approach to implement the principles of statistical learning theory [46] in logic mining, providing potential advancements both in accuracy and in robustness for logic mining.

### *1.2.3 Classification Analysis*

In predictive DM, the area of classification analysis has received unrivalled attention – both within literature and in practice. Classification has proven its effectiveness to support decision making and to solve complex planning tasks in various real-world application domains, including credit scoring (see, e.g., Crook et al. [8]) and direct marketing (see, e.g., Bose and Xi [4]). The predominant popularity of developing novel classification algorithms in the DM community seems to be only surpassed by the (often marginal) extension of existing algorithms in fine-tuning them to a particular data set or problem at hand. Consequently, Hand reflects that much of the claimed progress in DM research may turn out to be only illusive [23]. This leads to his reasonable expectation that advances will be based rather upon progress in computer hardware with more powerful data storage and processing ability than on building fine-tuned models of ever-increasing complexity. However, Friedman argues in a recent paper [15] that the development of kernel methods (e.g., SVMs) and ensemble classifiers, which form predictions by aggregating multiple basic models, both within the field of machine learning and DM, has further "revitalized" research within this field. Those methods may be seen as promising approaches toward future research in classification.

A novel ensemble classifier is introduced by Lemmond et al. [32] who draw inspiration from Breiman's random forest algorithm [6] and construct a random forest of linear discriminant models. Compared to classification trees used in the original algorithm, the base classifiers of linear discriminant analysis perform multivariate splits and are capable of exhibiting a higher diversity, which constitute novel and promising properties. It is theorized that these features may allow the resulting ensemble to achieve an even higher accuracy than the original random forest.

Lemmond et al. consider examples of the field of signal detection and conduct several empirical experiments to confirm the validity of this hypothesis.

SVM classifiers are employed in the work of Özöğür-Akyüz et al. [40], who propose a new paradigm for using this popular algorithm more effectively and efficiently in practical applications. Contrary to ensemble classifiers, standard practice in using SVMs stipulates the use of a single suitable model selected from a candidate pool determined by the algorithm's parameters. Regardless of potential disadvantages of this explicit "model selection" with respect to the reliability and robustness of the results, this principle is particularly counterintuitive because, prior to selecting this single model, a large number of SVM classifiers have to be built in order to determine suitable parameter settings in the absence of a robust methodology in specifying SVMs for data sets with distinct properties. In other words, the prevailing approach to employ SVMs is to first construct a (large) number of models, then to discard all but one of them and use this one to generate predictions. The approach by Özöğür-Akyüz et al. proposes to keep all classifier candidates and select either a single "most suitable" SVM or a collection of suitable classifiers for each individual case that is to be predicted. This procedure achieves appealing results in terms of forecasting accuracy and also computational efficiency, and it serves to integrate the established solutions of ensembles (an aggregate model selection) and individual model selection. Moreover, the general idea of reusing classifiers constructed within model selection and integrating them to produce ensemble forecasts can be directly transferred to other algorithms such as ANNs and other wrapper-based approaches, and thus contributes considerably to the general understanding of how such procedures can/should be used effectively.

Irrespective of substantial methodological and algorithmic advancements, the task of specifying classification models capable of dealing with imbalanced class distributions remains a particular challenge. In many empirical classification problems (where the target variable to be predicted takes on a nominal scale) one target class in the database is heavily underrepresented. Whereas such minority groups are usually of key importance for the respective application (e.g., detecting anomalous behavior of credit card use or predicting the probability of a customer defaulting on a loan), algorithms that strive to maximize the number of correct classifications will always be biased toward the majority class and impair their predictive accuracy on the minority group (see, e.g., [26, 50]). This problem is also considered by Liu et al. [33] in the context of classification with naive Bayes and SVM classifiers. Two popular approaches to increase a classifier's sensitivity for examples of the minority class involve either resampling schemes to elevate their frequency, e.g., through duplication of instances or the creation of artificial examples, or cost sensitive learning, essentially making misclassification of minority examples more costly. Whereas both techniques have been used successfully in previous work, a clear understanding as to how and under what conditions an approach works is yet lacking. To overcome this shortcoming, Liu et al. examine the formal relationship between cost-sensitive learning and different forms of resampling, most notably both from a theoretical and from an empirical perspective.

Learning in the presence of class and/or cost imbalance is one example where classification on empirical data sets proves difficult. Markedly, it has been observed that some applications do not enable a high classification accuracy to be obtained per se. The study of Weiss [49] aims at shedding light on the origins of this artifact. In particular, small disjuncts are identified as one influential source of high error rates, providing the motivation to examine their influence on classifier learning in detail. The term disjunct refers to a part of a classification model, e.g., a single rule within a rule-based classifier or one leaf within a decision tree, whereby the size of a disjunct is defined as the number of training examples that it correctly classifies. Previous research suggests that small disjuncts are collectively responsible for many individual classification errors across algorithms. Weiss develops a novel metric, error concentration, that captures the degree to which this pattern occurs in a data set and provides a single number measurement. Using this measure, an exhaustive empirical study is conducted that investigates several factors relevant to classifier learning (e.g., training set size, noise, and imbalance) with respect to their impact on small disjuncts and error concentration in particular.

### *1.2.4 Hybrid Data Mining Procedures*

As a natural result to the predominant attention of classification algorithms in DM a myriad of hybrid algorithms have been explored for specific classification tasks, combining neuro, fuzzy genetic, and evolutionary approaches. But there are also promising innovations beyond the mere hybridization of an algorithm tailored to a specific task. In practical applications, DM techniques for classification, regression, or clustering are rarely used in isolation but in conjunction with other methods, e.g., to integrate the respective merits of complementary procedures while avoiding their demerits and, thereby, best meet the requirements of a specific application. This is particularly evident from the perception of DM within the process of knowledge discovery from databases [10], which exemplifies an iterative and modular combination of different algorithms. Although a purposive combination of different techniques may be particularly valuable beyond the singular optimization within each step of the KDD process, this has often been neglected in research. This special issue includes four examples of such hybrid approaches.

A joint use of supervised and unsupervised methods within the process of KDD is considered by Figueroa [12] and Karamitopoulos et al. [30]. Figueroa conducts a case study within the field of customer relationship management and develops an approach to estimate customer loyalty in a retailing setting. Loyalty has been identified as one of the key drivers of customer value, and the concepts of customer lifetime value have been firmly established beyond DM. Therefore, it may prove sensible to devote particular attention to the loyal customers and, e.g., target marketing campaigns for cross-/up-selling specifically to this subgroup. However, defining the concept of loyalty is, in itself, a nontrivial undertaking, especially in noncontractual settings where changes in customer behavior are difficult to identify. The task

is further complicated by the fact that a regular and frequent update of respective information is essential. Figueroa proposes a possible solution to address these challenges: supervised and unsupervised learning methods are integrated to first identify customer subgroups and loyalty labels. This facilitates a subsequent application of ANNs to score novel customers according to their (estimated) loyalty.

Unsupervised methods are commonly employed as a means of reducing the size of data sets prior to building a prediction model using supervised algorithms. A respective approach is discussed by Karamitopoulos et al. who consider the case of multivariate time series analysis for similarity detection. Large volumes of time series data are routinely collected by, e.g., motion capturing or video surveillance systems that record multiple measurements for a single object at the same time interval. This generates a matrix of observations (i.e., measurements $\times$ discrete time periods) for each object, whereas standard DM routines such as clustering or classification would require objects being represented by row vectors. As a simple data transformation would produce extremely high-dimensional data sets, it would thereby further complicate analysis of such time series data. To alleviate this difficulty, Karamitopoulos et al. suggest reducing data set size and dimensionality by means of principal component analysis (PCA). This well-explored statistical approach will generate a novel representation of the data, which consists of a vector of the $m$ largest eigenvalues (with $m$ being a user-defined parameter) and a matrix of respective eigenvectors of the original data set's covariance matrix. As Karamitopoulos et al. point out, if two multivariate time series are similar, their PCA representations will be similar as well. That is, the produced matrices will be close in some sense. Consequently, Karamitopoulos et al. design a novel similarity measure based upon a time series' PCA signature. The concept of measuring similarity is at the core of many time series DM tasks, including clustering, classification, novelty detection, motif, or rule discovery as well as segmentation or indexing. Thus, it ensures broad applicability of the proposed approach. The main difference from other methods is that the novel similarity measure does not require applying a computer-intensive PCA to a query object: resource-intensive computations are conducted only once to build up a database of PCA signatures, which allows the identification of a query object's most similar correspondent in the database quickly. The potential of this novel similarity measure is supported by evidence from empirical experimentation using nearest neighbor classification.

Another branch of hybridization by integrating different DM techniques is explored by Johansson et al. [29] and Gijsberts et al. [18], who employ algorithms from the field of meta-heuristics to construct predictive classification and regression models. Meta-heuristics can be characterized as general search procedures to solve complex optimization problems (see, e.g., Voß [48]). Within DM, they are routinely employed to select a subset of attributes for a predictive model (i.e., feature selection), to construct a model from empirical data (e.g., as in the case of rule-based classification) or to tune the (hyper-)parameters of a specific model to adapt it to a given data set. The latter case is considered by Gijsberts et al. who evaluate evolutionary strategies (ES) to parameterize a least-square support vector regression (SVR) model. Whereas this task is commonly approached by means of genetic algorithms,

ES may be seen as a more natural choice because they avoid a transformation of the continuous SVR parameters into a binary representation. In addition, Gijsberts et al. examine the potential of genetic programming (GP) for SVR model building. SVR belongs to the category of kernel methods that employ a kernel function to perform an implicit mapping of input data into a higher dimensional feature space in order to account for nonlinear patterns within data. Exploiting the mathematical properties of such kernel functions, Gijsberts et al. develop a second approach that utilizes GP to "learn" an appropriate kernel function in a data-driven manner.

A related approach is designed by Johansson et al. for classification, where they employ GP to optimize the parameters of a $k$-nearest neighbor (kNN) classifier, most importantly the number of neighbors (i.e., $k$) and the weight individual features receive within distance calculations. In their study, Johansson et al. encompass classifier ensembles, whereby a collection of base kNN models is produced utilizing the stochasticity of GP to ensure diversity among ensemble members. As the general robustness of kNN with respect to resampling (i.e., the prevailing approach to construct diverse base classifiers) has hindered an application of kNN within an ensemble context, the approach of employing GP is particularly appealing to overcome this obstacle. Furthermore, Johansson et al. show that the predictive performance of the GP–kNN hybrid can be further increased by partitioning the input space into subregions and optimizing $k$ and the feature weights locally within these regions. A large-scale empirical comparison across different 27 UCI data sets provides valid and reliable evidence of the efficacy of the proposed model.

### *1.2.5 Web Mining*

The preceding papers concentrate mainly on the methodological aspects of DM. Clearly, the relevance of sophisticated data analysis tools in general, and their advancements in particular, is given by their broad range of challenging applications in various domains well beyond that of business and management. One domain of particular importance for corporate decision making, information systems and DM alike, is the World Wide Web, which has provided a new set of challenges through novel applications and data to many disciplines. In the context of DM, the term web mining has been coined to refer to the three branches of website structure, website content, and website usage mining.

A novel approach to improve website usability is proposed by Geczy et al. [16]. They focus on knowledge portals in corporate intranets and develop a recommendation algorithm to assist a user's navigation by predicting which resource the user is ultimately interested in and provide direct access to this resource by making the respective link available. This concept improves upon traditional techniques that usually aim only at estimating the next page within a navigation path. Consequently, providing the opportunity to access a potentially desired resource in a more direct manner would help to save a user's time, computational resources of servers, and bandwidth of networks.

A second branch of web mining is concerned with analyzing website content, e.g., to automatically categorize websites into predefined groups or to judge a page's relevance for a given user query in information retrieval. As techniques for natural language processing have reached a mature stage, unstructured data (such as web pages) can be transformed into a machine-readable format to facilitate DM with relative ease. The opportunities arising thereof is exploited by Ryan and Hamel [42]. The internet is considered as a pool of opinions, where current topics are discussed and shared among users, whose aggregation may facilitate the generation of accurate forecasts. Their research aims at constructing a forecasting model on the basis of search engine query results in order to predict future events. The proposed techniques allow the internet to be used as one large prediction market and, as such, represent an innovative approach toward forecasting. Current and future developments within the scope of Web 2.0 (e.g., social networking, blogging) as well as the Semantic Web can be expected to further increase the potential of this idea. This idea, in turn, will require the development of supporting IS (e.g., for gathering query results, transforming text data into machine-readable formats, as well as aggregating and possibly weighting resulting information) for a successful development in the long run.

### *1.2.6 Privacy-Preserving Data Mining*

The availability of very large data sets of detailed customer-centric information, e.g., on the purchasing behavior of an individual consumer or detailed information on a surfer's web usage behavior, not only offers opportunities from a DM perspective but also summons serious concerns regarding data privacy. As a consequence, both the relevance of privacy issues in DM and the awareness thereof continuously increase. This is mirrored by the increasing research activities within the field of privacy-preserving DM. In particular, substantial work has been conducted to conceptualize different models of privacy and develop privacy-preserving data analysis procedures. Privacy models like $k$-anonymity require that, after deleting identifiers from a data set, tuples of attributes which may serve as so-called quasi-identifiers (e.g., age, zip code.) show identical values across at least $k$ data records. This prohibits a reidentification of instances and hence insures privacy. Achieving $k$-anonymity or extended variants may thus necessitate some transformation of the original attributes, whereby inherent information has to be sustained to the largest degree possible in order to not impede subsequent DM activities. Truta and Campan [44] review alternative privacy models and propose two novel algorithms for achieving privacy levels of extended $p$-sensitive $k$-anonymity. Both techniques compare favorably to the established *Incognito* algorithm in terms of three different performance metrics (i.e., discernibility, normalized average cluster size, and running time) within an empirical comparison. Furthermore, Truta and Campan propose new privacy models that allow decision makers to constrain the degree to which quasi-identifier attributes are generalized within data anonymization. These

models are more aligned with the needs of real-world application by enabling a user to control the trade-off between privacy on the one hand and specific DM objectives (e.g., forecasting accuracy and between-cluster heterogeneity.) on the other explicitly. One of these models is tailored to the specific requirements of privacy in social networks, which have experienced a rapid growth within the last years. Up to now, their proliferation has not been accompanied by sufficient efforts to maintain privacy of users as well as their network relationships. In this context, the novel model for $p$-sensitive $k$-anonymity social networks may be seen as a particularly important and timely contribution.

Employing the techniques described by Truta and Campan allows anonymization of a single data set, so that an identification of individual data records through quasi-identifier attributes becomes impossible. However, such precautions can be circumvented if multiple data sets are linked and related to each other. For example, a respective case has been reported within the scope of the *Netflix* competition. A large data set of movie ratings from anonymous users has been published within this challenge to develop and assess recommendation algorithms. However, it was shown that users could be reidentified by linking the anonymous rating data with some other sources [37], which indicates the risk of severely violating privacy through linking data sets. On the other hand, a strong desire exists to share data sets with collaborators and engage in joint DM activities, e.g., within the scope of supply chain management or medical diagnosis to support and improve decision making. To enable this, Mangasarian and Wild [35] develop an approach that facilitates a distributed use of data for a DM, but avoids actually sharing it between participating entities. Mangasarian and Wild exploit the particular characteristics of kernel methods and develop a privacy-preserving SVM classifier, which is shown to effectively overcome the alleged trade-off between privacy and accuracy. Short of a true trade-off between accuracy and privacy, the proposed technique not only preserves privacy but also achieves equivalent accuracy as a classifier that has access to all data.

## 1.3 Conclusion and Outlook

Quo vadis, IS and DM? IS have been a key originator of corporate data growth and remain to have a core interest in the advancement of sophisticated approaches to analytical decision support in management. Processes, systems, and techniques in this field are commonly referred to as business intelligence (BI) within the IS community, and DM is acknowledged as part of corporate BI. However, in comparison to other analytical approaches such as OLAP (online analytical processing) or data warehouses, it has received only limited attention. On the contrary, disciplines like statistics, computer sciences, machine learning, and, more recently, operational research (see, e.g., [39, 38, 13]) have been most influential, which explains the emphasis on methodological aspects in the DM domain. This focus is well justified when considering the ever-growing number of novel applications and respective

requirements DM methods have to fulfill. Continuously sustaining such compliance with application needs requires that research activities do not only focus on established direction like procedures for predictive and descriptive data analysis but are also geared toward concrete decision contexts. Very recently, this understanding gave rise to two novel streams in DM research, namely utility-based DM (UBDM) (see, e.g., [51]) and domain-driven DM (see, e.g., [53]). Both acknowledge the importance of novel algorithms, but stress that their development should be guided by real-world decision contexts and constraints. This is precisely the approach toward decision support that has always been prevalent within the IS community. Consequently, more research along this line is highly desirable and needed to systematically exploit the core competencies found in IS and DM, respectively, and further improve the support of managerial decision making. Noteworthy examples of how this may be achieved have recently appeared in leading IS journals [43, 5] and reemphasize the potential of research at the interface between these two fields.

To the understanding of the reviewers and editors, the chapters in this special issue have captured those essential aspects in a convincing and clear manner and provide interesting, original, and significant contributions to the advancement of both DM and IS in the context of decision making. Therefore, in some sense, they can be considered as building blocks of the road that shows at least one possible direction for the further development of DM and IS. Of course, it is far beyond our goals and means to suggest one beatific direction. However, for DM and IS may be a fruitful answer to "Where are you going?" could be "Wherever we will go, we should accompany each other."

**Acknowledgments** We would like to thank all authors who submitted their work for consideration to this focused issue. Their contributions made this special issue possible. We would like to thank especially the reviewers for their time and their thoughtful reviews. Finally, we would like to thank the two series editors, Ramesh Sharda and Stefan Voß, for their valuable advice and encouragement, and the editorial staff for their support in the production of this special issue (Hamburg, June 2009).

## References

1. Agrawal, R. and Srikant, R. Fast algorithms for mining association rules in large databases. In: Bocca, J. B., Jarke, M., and Zaniolo, C. (eds.), *Proc. of the 20th Intern. Conf. on Very Large Databases (VLDB'94)*, pp. 487–499, Santiago de Chile, Chile, 1994. Morgan Kaufmann.
2. Ayres, I. *Super Crunchers: Why Thinking-By-Numbers Is the New Way to Be Smart*. Bantam Dell, New York, 2007.
3. Berry, M. J. A. and Linoff, G. *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management*. Wiley, New York, 2nd ed., 2004.
4. Bose, I. and Xi, C. Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, 195(1):1–16, 2009.
5. Boylu, F., Aytug, H., and Köhler, G. J. Induction over strategic agents. *Information Systems Research*, forthcoming.
6. Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001.
7. Cabena, P., Hadjnian, P., Stadler, R., Verhees, J., and Zanasi, A. *Discovering Data Mining: From Concept to Implementation*. Prentice Hall, London, 1997.

8. Crook, J. N., Edelman, D. B., and Thomas, L. C. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3):1447–1465, 2007.
9. Davis, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340, 1989.
10. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery in databases: An overview. *AI Magazine*, 17(3):37–54, 1996.
11. Felici, G., Simeone, B., and Spinelli, V. Classification techniques and error control in logic mining. *Annals of Information Systems*, in this issue.
12. Figueroa, C. J. Predicting customer loyalty labels in a large retail database: A case study in Chile. *Annals of Information Systems*, in this issue.
13. Fildes, R., Nikolopoulos, K., Crone, S. F., and Syntetos, A. A. Forecasting and operational research: A review. *Journal of the Operational Research Society*, 59:1150–1172, 2006.
14. Freitas, A. On rule interestingness measures. *Knowledge-Based Systems*, 12(5–6):309–315, October 1999. URL `http://www.cs.kent.ac.uk/pubs/1999/1407`.
15. Friedman, J. H. Recent advances in predictive (machine) learning. *Journal of Classification*, 23(2):175–197, 2006.
16. Geczy, P., Izumi, N., Akaho, S., and Hasida, K. Behaviorally founded recommendation algorithm for browsing assistance systems. *Annals of Information Systems*, in this issue.
17. Geng, L. and Hamilton, H. J. Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38(3):Article No. 9, 2006.
18. Gijsberts, A., Metta, G., and Rothkrantz, L. Evolutionary optimization of least-squares support vector machines. *Annals of Information Systems*, in this issue.
19. Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
20. Han, J. and Kamber, M. *Data mining: Concepts and Techniques*. The Morgan Kaufmann series in data management systems. Morgan Kaufmann, San Francisco, 7th ed., 2004.
21. Hand, D. J. Data mining: Statistics and more? *American Statistician*, 52(2):112–118, 1998.
22. Hand, D. J. Statistics and data mining: Intersecting disciplines. *ACM SIGKDD Explorations Newsletter*, 1(1):16–19, 1999.
23. Hand, D. J. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.
24. Hand, D. J., Mannila, H., and Smyth, P. *Principles of Data Mining*. Adaptive computation and machine learning. MIT Press, Cambridge, London, 2001.
25. Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2002.
26. Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–450, 2002.
27. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In: Nedellec, C. and Rouveirol, C. (eds.), *Proc. of the 10th European Conf. on Machine Learning*, vol. 1398 of *Lecture Notes in Computer Science*, pp. 137–142, Chemnitz, Germany, 1998. Springer.
28. Joachims, T. Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C. J. C., and Smola, A. J. (eds.), *Advances in Kernel Methods: Support Vector Learning*, pp. 169–184. MIT Press, Cambridge, 1999.
29. Johansson, U., König, R., and Niklasson, L. Genetically evolved kNN ensembles. *Annals of Information Systems*, in this issue.
30. Karamitopoulos, L., Evangelidis, G., and Dervos, D. PCA-based time series similarity search. *Annals of Information Systems*, in this issue.
31. Le Bras, Y., Lenca, P., and Lallich, S. Mining interesting rules without support requirement: A general universal existential upward closure property. *Annals of Information Systems*, in this issue.
32. Lemmond, T. D., Chen, B. Y., Hatch, A. O., and Hanley, W. G. An extended study of the discriminant random forest. *Annals of Information Systems*, in this issue.

33. Liu, A., Martin, C., La Cour, B., and Ghosh, J. Effects of oversampling versus cost-sensitive learning for Bayesian and SVM classifiers. *Annals of Information Systems*, in this issue.
34. Liu, B., Hsu, W., Chen, S., and Ma, Y. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55, 2000.
35. Mangasarian, O. L. and Wild, E. W. Privacy-preserving random kernel classification of checkerboard partitioned data. *Annals of Information Systems*, in this issue.
36. Martens, D. and Baesens, B. Building acceptable classification models. *Annals of Information Systems*, in this issue.
37. Narayanan, A. and Shmatikov, V. How to break anonymity of the Netflix prize dataset, 2006. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0610105`.
38. Olafsson, S. Introduction to operations research and data mining. *Computers and Operations Research*, 33(11):3067–3069, 2006.
39. Olafsson, S., Li, X., and Wu, S. Operations research and data mining. *European Journal of Operational Research*, 187(3):1429–1448, 2008.
40. Özöğür-Akyüz, S., Hussain, Z., and Shawe-Taylor, J. Prediction with the SVM using test point margins. *Annals of Information Systems*, in this issue.
41. Ringle, C. M., Sarstedt, M., and Mooi, E. A. Repose-based segmentation using finite mixture partial least squares. *Annals of Information Systems*, in this issue.
42. Ryan, S. and Hamel, L. Using web text mining to predict future events: A test of the wisdom of crowds hypothesis. *Annals of Information Systems*, in this issue.
43. Saar-Tsechansky, M. and Provost, F. Decision-centric active learning of binary-outcome models. *Information Systems Research*, 18(1):4–22, 2007.
44. Truta, T. M. and Campan, A. Avoiding attribute disclosure with the (extended) p-sensitive k-anonymity model. *Annals of Information Systems*, in this issue.
45. Vapnik, V. N. *Estimation of Dependences Based on Empirical Data*. Springer, New York, 1982.
46. Vapnik, V. N. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
47. Vapnik, V. N. *Statistical Learning Theory*. Wiley, New York, 1998.
48. Voß, S. Meta-heuristics: The state of the art. In: Nareyek, A. (ed.), *Local Search for Planning and Scheduling*, vol. 2148 of *Lecture Notes in Artificial Intelligence*, pp. 1–23. Springer, Berlin, 2001.
49. Weiss, G. M. The impact of small disjuncts on classifier learning. *Annals of Information Systems*, in this issue.
50. Weiss, G. M. Mining with rarity: A unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, 2004.
51. Weiss, G. M., Zadrozny, B., and Saar-Tsechansky, M. Guest editorial: special issue on utility-based data mining. *Data Mining and Knowledge Discovery*, 17(2):129–135, 2008.
52. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z.-H., Steinbach, M., Hand, D., and Steinberg, D. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.
53. Yu, P. (ed.). *Proc. of the 2007 Intern. Workshop on Domain Driven Data Mining*. ACM, New York, 2007.