

KNOWLEDGE DISCOVERY AND APPLICATION

A WHITE PAPER

LAURENCE JACOBS, PH.D.
CHIEF TECHNOLOGY OFFICER

kdlabs

Flurstrasse 32
CH-8066 Zürich
Switzerland

Phone +41 1 405 2100
Fax +41 1 405 2101

Contents

1.0	PRELIMINARIES	3
1.1	Examples	5
2.0	THE TECHNOLOGY	7
2.1	Assumptions and Constraints	7
3.0	AREAS OF APPLICATION	8
4.0	PROJECT STRUCTURE	9
	A phased approach to KDA	9
5.0	RANGE OF APPLICATIONS	10
	Examples of KDA implementations	10
	Some Profitable areas of KDA	11
	Range of applications of KDA	12
	<i>Action Categories</i> of KDA	13
6.0	A SUCCESSFUL PROJECT MODEL	14
6.1	Components	14
6.1.1	Phase I: Rapid Value Assessment	15
6.1.2	Phase II: Rapid Value Projection	16
6.1.3	Phase III: Rapid Value Delivery	17
7.0	CRITICAL SUCCESS FACTORS	19
7.1	Business problem	19
7.2	Data	19
7.3	Channel infrastructure	20
7.4	Management structures	20

1.0 Preliminaries

Knowledge Discovery and Application (KDA) is defined by a collection of methodologies and mathematical algorithms that have the aim of discovering and characterizing trends and patterns in datasets.

To understand why this can be important in profiling and predicting behavioral trends (in, for example, a particular customer), it is instructive to think of a dataset as a geometric object. An example should make this clear. Consider the specific case of a customer dataset containing a large number of records, one per customer, each of which contains M attributes that summarize all information known about each customer. These attributes might include static (or nearly static) characteristics, such as customer age, profession, income, marital status, and so on. Other characteristics may refer to the relationship of the owner of the dataset (a company, for example), and the customer. Examples of this could be a list of the products that the customer has purchased from the company, measures of how the customer uses these products, and revenue associated with this usage. With this example in mind, imagine that each record represents a point in M dimensions, with the value of each attribute representing one of the M coordinates of the point. It is, of course, somewhat difficult to picture this object in general, but it suffices to do so when M is less or equal to three.

The fundamental hypothesis in KDA (sometimes called the *Proximity Principle*) is the following. Two records that are very close to each other in $M-N$ of their coordinates (attributes) are likely to also be very close in the remaining N attributes. This becomes intuitively more and more plausible as M becomes large for fixed N (in other words, as the number of similarities between the two records far exceeds the number of unknowns).

As an example, think of two customers of a bank, Mr. A, and Mr. B. Both A and B live in the same town (two blocks from each other), have the same profession (mechanical engineers), work in the same field (manufacturing), are of similar ages (37 and 38), and both have been married for 8 years. Both Mr. A and Mr. B are customers of the same bank, own and use almost the same set of bank services and products, and have very similar incomes. However, for the last six months, Mr. A has been using the bank's internet facility to do most of his banking online, while Mr. B continues to use the bank's branch office near his home to do all of his banking. Mr. A and Mr. B are, according to the information collected by the bank, nearly identical in all but one of their attributes ($N=1$ in this case). Additionally, from the perspective of the bank, Mr. B is likely to be less

profitable than Mr. A because the costs associated with serving Mr. B are higher than those spent in serving Mr. A. The KDA hypothesis in this case is that Mr. B is very likely to become an online banking customer, if he is approached intelligently by the bank.

To complete the argument above we must quantify the *likelihood* that the common characteristics of Mr. A and Mr. B are, in fact, *typical* of online banking users. To do this, we must look at the neighborhood surrounding the two points corresponding to Mr. A and Mr. B, and measure the relative frequencies of online banking users and non-users in this neighborhood. An estimate of the confidence of the prediction that Mr. B will become an online banking customer is proportional to the ratio of users to non-users in this neighborhood.

The above is an example of prediction by classification (that is, the conclusion regarding Mr. B follows from the discovery that Mr. B is a member of a subclass of customers; specifically those whose profiles indicate a strong affinity to use internet banking, like Mr. A). There are several other application areas of Knowledge Discovery, but classification is both the most widely used, and, in general, should suffice as an example to describe the basis of KDA methods.

Continuing with the above example, an application of the discovered knowledge for the bank would be to run a marketing campaign to induce its customers to use their internet banking facilities. The process would work as follows. The first step is to determine whether there are clusters (blobs in the M-1 dimensional space) of customer groups characterized by usage of the target product (internet banking in this case). If such clusters are discovered, the next step would be to target all those customers in each of the affinity clusters who do not, at present, use internet banking. The resulting list of potential internet customers would then be shipped to a campaign management system in the bank for transmission to the prospects. An additional step would be to customize the campaign itself using the information that characterizes the prospect clusters. As an example, one might find that one of the affinity clusters is partly determined by an age range, say 22 to 29 years of age. A customized approach to that cluster might be designed to address that particular age group. Different prospect groups would be similarly addressed using different messages inspired by the properties of their cluster.

Notice that no further hypotheses enter into the determination of likely prospects in the example above; the only assumption we seem to have made is the *Proximity Principle* described above. This is both the greatest strength of KDA, as well as its greatest weakness: KDA generates knowledge *inductively*. The strength comes from the fact that the rules discovered are likely to include rules which were not known before; the weakness is related to the fact that the only rules that the method will discover are those that pertain to the data that is available. For example,

suppose that the product being marketed using KDA methods has a very strong age-related specificity, but the customer data used to build the KDA models does not include customer age as an attribute. In this case, the knowledge about age will not be discovered (although, in some cases, this may appear indirectly through a combination of several other attributes, but the example should suffice to make the point). Knowledge Discovery methods will only produce new knowledge if the underlying data supports it. By the same token, given the right data, the main strength of KD methods is that they produce results that are *specific* to the population that generated the data used in modeling. This has as a natural consequence that the models generated using KD methods are likely to be more accurate than general models when applied to the population in question.

A natural question that would arise at this point is whether it might be possible to improve on a model generated using KD methods by somehow combining the discovered knowledge with existing, domain-specific, known facts. This is most definitely the case, and successful use of KDA often includes this type of knowledge combination.

1.1 Examples

The marketing example above is but one of the applications of prediction by classification. There are many others. Another important example is the determination of the risk of default on credit. General assumptions on credit-worthiness are not very accurate, but are generally the method of choice of lending institutions today. KDA methods will almost always lead to more accurate determination of credit risk, with the caveats described above. This is one of the richest application potentials of KD methods today, and an increasing number of lending institutions are moving in this direction at present, while others, quietly, have been doing already it for a few years.

Another important application is the discovery of *unmarked* clusters (generally known as *a-priori clustering*). The idea is quite similar to the one we have described above, except that in this case there is no *target* attribute. In the previous two examples, specific attributes (internet banking and credit risk) are used to discover the clusters; in the case of a-priori clustering, all data attributes play the same role in principle. The guiding principle here is still the Proximity Principle, but in this case it is applied to all attributes. For this reason, this application is often referred to as *unsupervised learning*. The value of this application is the potential discovery of *natural* segments in the data (for example, customer subgroups) that might lead to improvements in revenue-generation. This is also important in that it may help to point out interesting directions for subsequent supervised modeling in the individual discovered clusters.

The last example we mention here is what we call *Complete Cross-Sell* (sometimes also called the *next-best product* application). This application leads to a relative score of all products for all customers. For any given customer, it results in a list of all products the customer might be likely to buy, ordered by the predicted likelihood of each one of the products. This application has immediate and highly valuable uses with channels such as a call center or direct internet commerce.

2.0 The Technology

The concepts introduced above are very simple and intuitive; as is often the case, and is particularly true with KDA, most of the complexity is in the technical details of the necessary tools. These details, however, have been mostly solved, and are now generally available in commercial products.

A number of commercial packages exist in the market today, the best ones offer an array of algorithms and tools that represent the current state of the art in the field. Some of the most popular tools today are, *C&RT* (Classification and Regression Trees), *ANN* (Artificial Neural Networks, of which there are many variants), Bayesian classifiers, and *SOM* (Self-Organizing Maps).

2.1 Assumptions and Constraints

While the Proximity Principle is independent of the target application and of the details of the data used in its application, there are three basic assumptions that must be true for KD to be successful. We list these below.

1. Patterns (clusters) must exist in the M-dimensional space
2. The tools and algorithms used to detect these patterns should lead to good approximations to the cluster characteristics
3. There should be enough data to validate the statistical conclusions which follow from the application of these tools and algorithms

3.0 Areas of Application

KDA methods are not restricted to any particular industry. As we mention in some more detail below, for each abstract application of KDA there are specific examples in almost all industries. The language and some other details change from one industry to another, as does the detailed value of any particular KD application. The only constraints of applicability are those we have discussed already.

In most industries today, identification, control, and management of customer attrition represents the richest application of KD technologies. Attrition, and its complement, Retention, are some of the most complex problems in which KD technologies are used today. Addressing and solving the attrition problem requires a comprehensive development of a complete customer-centric system. This system includes most of the known Knowledge Discovery applications.

The ultimate goal of any well-designed KDA project is the development of a customer-centric picture in all the business areas of an enterprise. The economic benefits associated with correct customer segmentation and related marketing and customer-care activities are generally very large. Retention and grooming of valuable customers, careful optimization of risks, and other targets of KDA, when used correctly, represent a very large return on investment.

Short-term, successful uses of Knowledge Discovery (often called *Quick Wins*) are important, however, since they provide the impetus necessary to embark on a new business direction, or to introduce new and unfamiliar technologies into the business processes in a company. However, it is the strategic goals of KDA, which should be kept in mind, because it is these aspects of the technology that have the most profound economic impact. Customer Loyalty and its management should be at the forefront of the thinking of any business leader today. A good choice for a Quick Win is therefore one that both generates a measurable return in a short time, and fits into the strategic goals of a long-term KDA project.

4.0 Project Structure

For essentially practical reasons, most successful implementations of KDA projects are globally divided into three general phases. In very general terms, the processes involved in each of these phases can be summarized as follows.

Phase	Main Activities	Duration
1	Proof of Concept Quick Wins Technology and Process Introduction	1-6 months
2	Development of Production Systems Business Methodology Integration Systematic Exploration of Business cases Programming Strategic Goals Channel Definition and Selection	3-12 months
3	Integration of Systems and Processes Process Automatization Channel Integration and Automatization Application Development and Roll-out	6-18 months

A phased approach to KDA

It should be understood that these are very general concepts, and that (important) details vary from industry to industry, and from case to case in a given industry. As a general rule, however, the table above captures most of the essential steps required for a successful implementation of a KDA project.

5.0 Range of Applications

Data-rich industries, such as Telecommunications, Retail, Direct Sales and Financial industries (including Insurance) have been some of the most successful pioneers in the implementation of KDA into their core activities. The main reasons for this are competition, market dynamics, and dimension of customer base. A solution of the attrition problem in any of these industries is generally more than sufficient to justify the costs associated with the introduction of new technologies and new business practices. In some cases, an improvement of 1-3% in retention alone is easily sufficient to justify the expenditures and risks associated with the development and implementation of KDA.

KDA has been used successfully in a wide range of business areas. Some of the most common examples can be found in the following industries:

- Banking
- Insurance
- Telecommunications
- Pharmaceuticals
- Retail
- Direct Sales
- Subscription Services

It should be stressed that, in most cases, a given application (e.g., *Cross-sell*) has applications in all of the above industries. The data and the language may change from industry to industry, but the substance of the application is portable across industries.

Some of the most common examples of applications and the ranges of expected results are shown in the following table.

Business Area	Application
Customer Loyalty	Attrition
Marketing	Cross-sell, Up-sell
Credit	Risk assessment
New Product Development	Product Bundles
Profitability	Segmentation
Risk Assessment	Credit Risk

Examples of KDA implementations

The expected ranges of improved profitability, for example, depend on factors that vary greatly from one application to another and from one industry to another, and cannot be estimated in general. However, fairly simple analyses based on rough costs and benefits associated with each customer segment or each product or service lead to interesting numbers in every case shown here.

A maximally successful implementation of KDA would require an understanding of the scales of costs and benefits described above. It is clear, for example, that an abstract solution of the attrition problem (one that does not take profit data into consideration, for example) makes very little sense from a practical perspective. An actual estimation of risk should include not only the probability of attrition, but also the magnitude of the capital at risk associated with the possible loss of a given customer. In general, though, as enterprises begin to realize the potential benefits of KDA and related technologies, these issues are confronted rapidly and efficiently to maximize financial returns.

Some specific areas of application most commonly used in Marketing and Product Development in data-rich industries are shown in the table below

Type	Examples
New Product Marketing	Compatible Subsets Natural affinities Basket analysis Product gaps Compatible tariff structures
Customer Segmentation	Product or service groups Tariff groups Profit-based groups Demographic groups

Some Profitable areas of KDA

Typical examples of business problems that fit into the categories of either *Quick Wins* or *Strategic Problems* are given below. Of course, it is to be understood that most problems that fall into the *Quick Win* category have strategic value; the classification merely addresses the time scales associated with return of value.

Type	Examples
Quick Wins	Cross-sell Up-sell One-to-one Marketing Customer Segmentation Profitability Analysis Risk Assessment Customer Profiling
Strategic Problems	Risk Management New Product Development Campaign Design Attrition and Retention Product-gap Analysis Wallet-share

Range of applications of KDA

Viewed from a different perspective, profitable business cases in data-rich industries fall into *Action* categories as follows.

Factor	Components
Profitability	Product Mix Product Bundling Product Gap Wallet Share Product-Customer Mismatch
Cluster Discovery and Analysis	Customer-Advisor choice Natural Segmentation Cluster Profiling Specific Goal Subclusters
A-priori Segmentation	Characteristics Main Drivers
Attrition and Retention	Segment Correlation Main Drivers
Marketing	Segment Orientation Campaign Customization Basket Analysis New Product Definition

Action Categories of KDA

6.0 A Successful Project Model

In this section we present an example of the structure of a successful KDA project based on our experience in the industry.

kd|labs offers a service package called the *Rapid Value System*[®] (RVS). This package comprises a series of iterative service offerings that aim at demonstrating the business value that can be identified in the client's existing data. The implementation and delivery of these services requires individuals with the highest skills and experience and with proven track records in KDA.

The RVS offering is built around a proven three-phase methodology. While the approach is designed to be a proof of concept, it delivers measurable value at all stages, with well-defined exit points. These exit points are also thresholds for investment limitation and help make the decision on whether or not to make further investment in the project. Given enough financial information, at each exit point in RVS a quantitatively accurate measure of return on investment for the project is made up to that point, making a go/no-go decision simple. The main goal of our methodology is to provide the greatest value in the shortest possible time.

Each succeeding phase in this approach delivers incremental value, and sets the stage for the next phase in such a way that the overall technical risk is minimized.

The only assumptions in RVS concern the properties and structure of the data available for analysis, which is specified concretely before an engagement, and which is verified in the very early stage of the project.

6.1 Components

The *Rapid Value System*, is designed around three phases:

1. Rapid Value Assessment
2. Rapid Value Projection
3. Rapid Value Delivery

The first two phases are generally fixed, while the third offers a series of options.

6.1.1 Phase I: Rapid Value Assessment

This phase consists of a series of activities aimed at determining the value of the data for the general purposes of Knowledge Discovery and Application. This phase comprises three items, with each item successively exploring various aspects of the dataset more deeply. The most important step in this phase is a first analysis of natural clusters in the data, together with the generation of approximate rules describing such clusters. This step is essential for a full eventual business characterization of the client's customer base.

Using existing financial data, the natural clusters discovered during this phase will lead to the definition of natural business segments as well as to the identification of the main differences between profitable and unprofitable customer segments. A differential analysis of the cluster identifiers will then lead to the discovery of the main drivers that can be used to move customers into the most profitable business segments.

Item 1 Data Quality Assessment

This first step aims to determine very general aspects of data quality and its usability in KDA. It consists of a series of sub-steps:

1. Data cleaning
2. Data consistency analysis
3. Exploration of possible attribute transformations
4. General statistical profile of the dataset

Item 2 Initial Knowledge Content Assessment

Once general data quality has been measured and quantified, the next step is to look for specific attributes corresponding to products or services and to their level of predictability. This step has the following components:

1. Initial customer segmentation
2. Initial profit (or revenue) cross-correlation with natural segments
3. Exploration of predictability of interesting products or services
4. Search for unknown information leakers
5. Feature selection cross tables

Item 3 Basic Cross-sell Potential Assessment

This analysis consists of a first derivation of association statistics and rules between product and service pairs in the client's offering portfolio. The

analysis generates a first view of those products and/or services that have a natural affinity with each other, leading to potentially useful cross-sell models.

Deliverables

The main deliverables in this phase are numerical scores, which quantify the output of the three items described above, as well as the results of initial customer segmentation and basic cross-sell tables.

6.1.2 Phase II: Rapid Value Projection

Given the results of the preceding phase, and given adequate financial data (described below), during this phase we develop and measure the likely business impact of the full project by producing accurate financial projections for all Phase III options. Specifically, during this phase we will estimate the following financially relevant metrics:

1. Lift
2. Margin
3. ROI
4. ROI *Sweet Spot*

These metrics will be computed for those applications that were determined to be most promising during Phase I, and will thus allow a clear choice of options for Phase III by concretely specifying the likely financial impact of each.

Derivation and testing of models for targeted marketing

On the basis of collaborative business analysis by **kd|labs** and client staff, a small number of products or services (three or four) are chosen to be used in marketing campaigns. Preliminary financial projections characterizing the chosen products will precede the full development of marketing models (clearly only those products studied during Phase II should be considered here). Our task here will be to build models that characterize the profiles associated with *typical* users of the chosen products.

These models will be applied to the entire customer dataset and, for each of the chosen products, a probabilistic score will be derived for each customer in the dataset. This score expresses the unscaled probability that each customer will wish to acquire the given product if it were to be offered.

Finally, a set of scored lists derived from our models will be combined through a boosting process to generate a production list of potential users of each of the chosen products. These scored lists will be used in marketing campaigns to be defined and designed in collaboration with the client.

The usual testing method used at this stage of a KDA project is known as a *paper audit*. In simple terms, this refers to an analysis done by testing our models against unseen data (a small sample of the analysis dataset) kept by the client. The models test successfully if the statistics measured using unseen data are statistically consistent with our predictions.

6.1.3 Phase III: Rapid Value Delivery

There are several broad option categories (non-exclusive) for a follow-up to Phase II. We describe some of these choices below as examples.

Credit Risk

For each credit instrument of interest to the client, Phase III will proceed as follows. An initial, or *baseline* measure of risk assessment will be made by an inductive determination of current decision rules (grant) and the direct economic consequences over a given period (one year is usually sufficient). Once the baseline has been determined, for each credit instrument of interest, we will derive a full inductive model of risk assessment.

Full direct economic value of the inductive model over the baseline will then be determined by simple differential analysis.

Credit Risk: Deliverables

There are four families of deliverables for this sub-phase. For each credit instrument chosen by the client we will deliver,

1. Baseline Model
2. Full Inductive Model
3. Scoring of all active customers according to the full inductive model
4. Full differential risk analysis (per customer and per credit instrument)

The models can be delivered either in direct human-readable language, or in any of a number of computer languages (such as C, C++, Java, or AWK). In a possible eventual production phase, given an adequate database structure, models can also be delivered (and installed) as stored procedures.

Targeted Marketing

The ultimate goal of this option is to develop targeted marketing models for products chosen by the client

A series of steps are presented below which, if carried out, can greatly improve the business value of the main goal of this sub-phase. Some of the proposed analyses can be considered optional at this point, but we strongly recommend that they be done.

Targeted Marketing: Distribution channel affinity analysis

An analysis of possible natural affinities between customers and channels will be performed for eventual use in customized marketing and service offerings to the client's customer base. Assuming that this information exists in the analysis dataset, it is likely that channel affinities will surface as independent predictors of other customer behavior, such as, for example, those associated with an existing (or new) product or service. For future analyses, such as those leading to attrition and retention models, channel affinity will also probably play an important role.

Targeted Marketing: A-priori clustering analysis

An analysis of natural clusters in the data, together with the generation of approximate rules describing such clusters will be performed. This step is essential for a full eventual business characterization of the client's customer base.

Using existing financial data, the natural clusters discovered during this phase will lead to the definition of natural business segments as well as to the identification of the main differences between profitable and unprofitable customer segments. A differential analysis of the cluster identifiers will then lead to the discovery of the main drivers that can be used to move customers in the most profitable directions.

Targeted Marketing: Deliverables

1. Full Inductive Model
2. Scoring of all prospects
3. Analysis of customer response to the campaign

7.0 Critical Success Factors

A number of critical success factors have already been mentioned in the preceding sections. Here, to conclude, we summarize those which appear to us to be most important.

Below we describe some of the main requirements associated with these components.

7.1 Business problem

A good business problem to attack using KDA should have two main characteristics: (1) Its solution should be technically feasible, and (2), its solution should lead to clearly measurable benefits for the customer. The following issues should be considered in determining whether a candidate problem meets these conditions,

1. Business language description
2. Range of expected economic benefits associated with its solution
3. Relationship to other business cases
4. Causal dependencies with other business problems
5. Relative technical difficulty
6. Expected solution time
7. Relative potential value

7.2 Data

As should be clear from our discussion, adequate data is at the core of a successful use of KDA methods. It is generally not possible to give general quantifiable metrics in this case, but the following issues are essential in determining whether the data available for the solution of a given problem has the necessary structure and characteristics. In the most important case of customer-centric applications, such as those exemplified in this document, the most important are the following,

1. Clearly identifiable target identifier (generally speaking, the data should identify customers)
2. It should include enough target information to differentiate customers

3. There should be enough data to make the results of KD methods statistically significant

7.3 Channel infrastructure

It is necessary to have the vehicle to use the results of KDA (for example, a marketing function), as well as the means to measure the response and effectiveness of the application of KD (for example, identifiable customer response to a marketing campaign).

7.4 Management structures

A successful implementation of KDA is generally the result of a close collaboration between the technical and business components of an enterprise, typically between the IT and Marketing departments.