Before we start mining any data, we need to define the problem we want to solve and, most importantly, gather the right data to help us find the solution. If we don’t have the right data, we need to get it. If data mining is not properly approached, there is the possibility of “garbage in—garbage out”. To be effective in data mining, you will typically follow a four-step process:

Sequence for Determining Necessary Data. Wrong: Catalog everything you have, and decide what data is important. Right: Work backward from the solution, define the problem explicitly, and map out the data needed to populate the investigation and models.

—James Taylor with Neil Raden, authors, Smart (Enough) Systems

The Data Mining Process

Before we start mining any data, we need to define the problem we want to solve and, most importantly, gather the right data to help us find the solution. If we don’t have the right data, we need to get it. If data mining is not properly approached, there is the possibility of “garbage in—garbage out”. To be effective in data mining, you will typically follow a four-step process:
Defining the Business Problem

This is the most important step. In this step, a domain expert determines how to translate an abstract business objective such as “How can I sell more of my product to customers?” into a more tangible and useful data mining problem statement such as “Which customers are most likely to purchase product A?” To build a model that predicts who is most likely to buy product A, we first must acquire data that describes the customers who have purchased product A in the past. Then we can begin to prepare the data for mining.

Gathering and Preparing the Data

Now we take a closer look at our data and determine what additional data may be necessary to properly address our business problem. We often begin by working with a reasonable sample of the data. For example, we might examine several hundred of the many thousands, or even millions, of cases by looking at statistical summaries and histograms. We may perform some data transformations to attempt to tease the hidden information closer to the surface for mining. For example, we might transform a “Date_of_Birth” field into an “AGE” field, and we might derive new field such as “No_Times_Amt_Exceeds_N” from existing fields. The power of SQL simplifies this process.

Model Building and Evaluation

Now we are ready to build models that sift through the data to discover patterns. Generally, we will build several models, each one using different mining parameters, before we find the best or most useful model(s).

Knowledge Deployment

Once ODM has built a model that models relationships found in the data, we will deploy it so that users, such as managers, call center representatives, and executives, can apply it to find new insights and generate predictions. ODM’s embedded data mining algorithms eliminate any need to move (rewrite) the models to the data in the database or to extract huge volumes of unscored records for scoring using a predictive model that resides outside of the database. Oracle Data Mining provides the ideal platform for building and deploying advanced business intelligence applications.
The data mining process involves a series of steps to define a business problem, gather and prepare the data, build and evaluate mining models, and apply the models and disseminate the new information.

Exadata and Oracle Data Mining

Oracle Exadata is a family of high performance storage software and hardware products that can improve data warehouse query performance by a factor of 10X or more. Oracle Data Mining scoring functions in Oracle Database 11g Release 2 score in the storage layer and permit very large data sets to be mined very quickly, thus further increasing the competitive advantage already gained from Oracle's in-database analytics.

```sql
SELECT cust_id
FROM customers
WHERE region = 'US'
AND prediction_probability(churnmod, 'Y' USING *) > 0.8;
```
A Wide Range of Cutting Edge Algorithms

Oracle Data Mining provides a broad suite of data mining techniques and algorithms to solve many types of business and technical problems:

<table>
<thead>
<tr>
<th>TECHNIQUE</th>
<th>APPLICABILITY</th>
<th>ALGORITHMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Classification techniques use historical data to build models that can be used to classify new data and make predictions about class membership (e.g. 0 or 1) or class value (numerical value).</td>
<td>Logistic Regression (GLM)—classic statistical technique inside the Oracle Database. Supports thousands of input variables, text and transactional data. Naive Bayes—Fast, simple, commonly applicable. Support Vector Machine—Cutting edge algorithm that supports binary and multi-class problems. SVMs also excel at handling shallow, yet wide, and nested data problems e.g. transaction data or gene expression data analysis. Supports text mining use cases. Decision Tree—Popular algorithm useful for many classification problems that can help explain the model’s logic using human-readable “If... Then...” rules.</td>
</tr>
<tr>
<td>Regression</td>
<td>Technique for predicting a continuous numerical outcome such as customer lifetime value, house value, process yield rates.</td>
<td>Multiple Regression (GLM)—classic statistical technique available inside the Oracle Database Supports thousands of input variables, text and transactional data. Support Vector Machine — Cutting edge algorithm that supports regression problems. Supports text mining and transactional data use cases.</td>
</tr>
<tr>
<td>Attribute Importance</td>
<td>Ranks attributes according to strength of relationship with target attribute, for example finding factors associated with people who respond to an offer.</td>
<td>Minimum Description Length—Considers each attribute as a simple predictive model of the target class. Attribute Importance algorithm finds the attributes that have the most influence on a target attribute.</td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>Identifies unusual or</td>
<td>One-Class Support Vector</td>
</tr>
<tr>
<td>Technique</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>Useful for exploring data and finding natural groupings within the data. Members of a cluster are more like each other than they are like members of a different cluster. Common examples include finding new customer segments, and life sciences discovery.</td>
<td></td>
</tr>
<tr>
<td><strong>Association</strong></td>
<td>Finds rules associated with frequently co-occurring items, used for market basket analysis, cross-sell, and root cause analysis. Useful for product bundling, in-store placement, and defect analysis.</td>
<td></td>
</tr>
<tr>
<td><strong>Feature Extraction</strong></td>
<td>Produces new attributes as linear combination of existing attributes. Applicable for text data, latent semantic analysis, data compression, data decomposition and projection, and pattern recognition.</td>
<td></td>
</tr>
<tr>
<td><strong>Machine</strong></td>
<td>Unsupervised learning technique that trains on “normal cases” to build a model. Then when applied, it flags unusual cases with the probability that they are not “normal”.</td>
<td></td>
</tr>
<tr>
<td><strong>Enhanced K-Means</strong></td>
<td>Supports text mining, hierarchical clustering, distance based.</td>
<td></td>
</tr>
<tr>
<td><strong>Orthogonal Partitioning</strong></td>
<td>Hierarchical clustering, density based.</td>
<td></td>
</tr>
<tr>
<td><strong>Apriori</strong></td>
<td>Industry standard for market basket analysis.</td>
<td></td>
</tr>
<tr>
<td><strong>Non-negative Matrix Factorization (NMF)</strong></td>
<td>Creates new attributes that represent the same information using fewer attributes</td>
<td></td>
</tr>
</tbody>
</table>

**Supervised Learning Algorithms**

Most data mining algorithms can be separated into supervised learning and unsupervised learning data mining techniques. Supervised learning requires the data analyst to identify a target attribute or dependent variable with examples of the possible classes (e.g., 0/1, Yes/No, High/Med, Low, etc.). The supervised-learning technique then sifts through data trying to find patterns and relationships among the independent attributes (predictors) that can help separate the different classes of the dependent attribute.
For example, let’s say that we want to build a predictive model that can help our Marketing and Sales departments focus on people who are most likely interested in purchasing a new car. The target attribute will be a column that designates whether each customer has purchased a car—for example, a “1” for yes and a “0” for no. The supervised data mining algorithm sifts through the data searching for patterns and builds a data mining model that captures the relationships found in the data. Typically, for supervised learning, the data is separated into two parts — one for model training and another hold out sample for model testing and model evaluation. Because we already know the outcome — who purchased a car and who hasn’t — we can apply our ODM predictive model to our hold out sample to evaluate the model’s accuracy and make decisions about the usefulness of the model. ODM models with acceptable prediction capability can have high economic value. Binary and multi-class classification problems represent a majority of common business challenges addressed through Oracle Data Mining, including database marketing, response and sales offers, fraud detection, profitability prediction, customer profiling, credit rating, churn anticipation, inventory requirements, failure anticipation, and many others. Oracle Data Mining also provides utilities for evaluating models in terms of model accuracy and “lift” — or the incremental advantage of the predictive model over the naïve guess.

**Naïve Bayes**

Naïve Bayes (NB) is a supervised-learning technique for classification and prediction supported by Oracle Data Mining. The Naïve Bayes algorithm is based on conditional probabilities. It uses Bayes’ Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows.

Bayes' Theorem:  \( \text{Prob}(B \text{ given } A) = \frac{\text{Prob}(A \text{ and } B)}{\text{Prob}(A)} \)

To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone.

After ODM builds a NB model, the model can be used to make predictions. Application developers can integrate ODM models to classify and predict for a variety of purposes, such as:

- Identify customers likely to purchase a certain product or to respond to a marketing campaign
- Identify customers most likely to spend greater than $3,000
- Identify customers likely to churn

NB affords fast model building and scoring and can be used for both binary and multi-class classification problems. NB cross-validation, supported as an optional way to run NB, permits the user to test model accuracy on the same data that was used to build the model, rather than building the model on one portion of the data and testing it on a different portion. Not having to hold aside a portion of the data for testing is especially useful if the amount of build data is relatively small.
Decision Trees

Oracle Data Mining supports the popular Classification Tree algorithm. The ODM Decision Tree model contains complete information about each node, including Confidence, Support, and Splitting Criterion. The full Rule for each node can be displayed, and in addition, a surrogate attribute is supplied for each node, to be used as a substitute when applying the model to a case with missing values.

Support Vector Machines

ODM’s Support Vector Machines (SVM) algorithm supports binary and multi-class classification, prediction, and regression models, that is, prediction of a continuous target attribute. SVMs are particularly good at discovering patterns hidden in problems that have a very large number of independent attributes, yet have only a very limited number of data records or observations.

SVM models can be used to analyze genomic data with only 100 patients who have thousands of gene expression measurements for each patient. SVMs can build models that predict disease treatment outcome based on genetic profiles.

Generalized Linear Models (Logistic and Multiple Regression)

ODM 11g Release 2 adds support for the multipurpose classical statistical algorithm, Generalized Linear Models (GLM). ODM supports, as two mining functions: classification (binary Logistic Regression) and regression (Multivariate Linear Regression). GLM is a parametric modeling technique. Parametric models make assumptions about the distribution of the data. When the
assumptions are met, parametric models can be more efficient than non-parametric models. Oracle Data Mining’s GLM implementation provides extensive model quality diagnostics and predictions with confidence bounds.

Oracle Data Mining supports ridge regression for both regression and classification mining functions. ODM’s GLM automatically uses ridge if it detects singularity (exact multicollinearity) in the data. ODM supports GLM with the added capability to handle many hundreds to thousands of input attributes. Traditional external statistical software packages typically are limited to 10-30 input attributes.

Attribute Importance
Oracle Data Mining’s Attribute Importance algorithm helps to identify the attributes that have the greatest influence on a target attribute. Often, knowing which attributes are most influential helps you to better understand and manage your business and can help simplify modeling activities. Additionally, these attributes can indicate the types of data that you may wish to add to your data to augment your models.

Attribute Importance can be used to find the process attributes most relevant to predicting the quality of a manufactured part, the factors associated with churn, or the genes most likely related to being involved in the treatment of a particular disease.

Unsupervised Learning Algorithms
In unsupervised learning, the user does not specify a target attribute for the algorithm. Unsupervised learning techniques, such as associations and clustering algorithms, make no assumptions about a target field. Instead, they allow the data mining algorithm to find associations and clusters in the data independent of any a priori defined business objective.

Clustering
Oracle Data Mining provides two algorithms, Enhanced k-Means and Orthogonal Partitioning Clustering (O-Cluster), for identifying naturally occurring groupings within a data population. ODM’s Enhanced k-Means (EKM) and O-Cluster algorithms support identifying naturally occurring groupings within the data population. ODM’s EKM algorithm supports hierarchical clusters, handles numeric and categorical attributes and will cut the population into the user specified number of clusters.

ODM’s O-Cluster algorithm handles both numeric and categorical attributes and will automatically select the best cluster definitions. In both cases, ODM provides cluster detail information, cluster rules, cluster centroid values, and can be used to “score” a population on their cluster membership. For example, Enhanced k-Means Clustering can be used to find new customer segments or to reveal subgroups within a diseased population.
Association Rules (Market Basket Analysis)

ODM’s Association Rules (AR) finds co-occurring items or events within the data. Often called “market basket analysis”, AR counts the number of combinations of every possible pair, triplet, quadruplet, etc., of items to find patterns. Association Rules represent the findings in the form of antecedents and consequents. An AR rule, among many rules found, might be “Given Antecedents Milk, Bread, and Jelly, then Consequent Butter is also expected with Confidence 78% and Support 12%. Translated in simpler English, this means that if you find a market basket having the first three items, there is a strong chance (78% confidence) that you will also find the fourth item and this combination is found in 12% of all the market baskets studied. The associations or “rules” thus discovered are useful in designing special promotions, product bundles, and store displays.

AR can be used to find which manufactured parts and equipment settings are associated with failure events, what patient and drug attributes are associated with which outcomes or which items or products is a person who has purchased item A most likely to buy?

Anomaly Detection

Release 2 of Oracle Data Mining 10g introduced support for a new mining application—anomaly detection, that is, the detection of “rare cases” when very few or even no examples of the rare case are available. Oracle Data Mining can “classify” data into “normal” and “abnormal” even if only one class is known. ODM uses a special case of the Support Vector Machines algorithm to create a model of known cases. When the model is applied to the general population, cases that don’t fit the profile are flagged as anomalies (that is, abnormal or suspicious). ODM’s anomaly detection algorithm is extremely powerful in finding truly rare occurrences when you have a lot of data but need to find needles in the haystacks.

Feature Extraction

ODM’s Nonnegative Matrix Factorization (NMF) is useful for reducing a large dataset into representative attributes. Similar in high level concept to Principal Components Analysis (PCA), but able to handle much larger amounts of attributes and create new features in an additive nature, NMF is a powerful, cutting-edge data mining algorithm that can be used for a variety of use cases.

NMF can be used to reduce large amounts of data, e.g., text data, into smaller, more sparse representations that reduce the dimensionality of the data, i.e., the same information can be preserved using far fewer variables. The output of NMF models can be analyzed using supervised learning techniques such as SVMs or unsupervised learning techniques such as clustering techniques. Oracle Data Mining uses NMF and SVM algorithms to mine unstructured text data.
Text Mining and Unstructured Data

Oracle Data Mining provides a single unified analytic server platform capable of mining both structured, that is, data organized in rows and columns, and unstructured data. ODM can mine unstructured data, that is, “text” as a text attribute that can be combined with other structured data, for example, age, height, and weight to build classification, prediction, and clustering models. ODM could add, for example, a physician’s notes to the structured “clinical” data to extract more information and build better data mining models.

This ability to combine structured data with unstructured data opens new opportunities for mining data. For example, law enforcement personnel can build models that predict criminal behavior based on age, number of previous offenses, income, and so forth, and combine a police officer’s notes about the person to build more accurate models that take advantage of all available information.

Additionally, ODM’s ability to mine unstructured data is used within Oracle Text to classify and cluster text documents stored on the Database, e.g. Medline. Oracle Data Mining’s NMF and SVM models can be used with Oracle Text to build advanced document classification and clustering models.