



SPE 84441

## Essential Components of an Integrated Data Mining Tool for the Oil & Gas Industry, With an Example Application in the DJ Basin

Shahab D. Mohaghegh, SPE, West Virginia University

Copyright 2002, Society of Petroleum Engineers Inc.

This paper was prepared for presentation at the SPE Annual Technical Conference and Exhibition held in San Antonio, Texas, 29 September–2 October 2002.

This paper was selected for presentation by an SPE Program Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the Society of Petroleum Engineers and are subject to correction by the author(s). The material, as presented, does not necessarily reflect any position of the Society of Petroleum Engineers, its officers, or members. Papers presented at SPE meetings are subject to publication review by Editorial Committees of the Society of Petroleum Engineers. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of the Society of Petroleum Engineers is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of where and by whom the paper was presented. Write Librarian, SPE, P.O. Box 833836, Richardson, TX 75083-3836, U.S.A., fax 01-972-952-9435.

### Abstract

Data Mining seems to be the new buzz word. During the past several years many industries other than ours, have realized the potential benefits of Data Mining and have established sophisticated operations in order to implement this exciting technology in their respective organizations. Data Mining is not new. It has been around for many years. What is new about its current implementation is the incorporation of machine learning techniques. Oil & gas industry has become familiar with machine learning techniques since early 90s. Neural networks, genetic optimization and fuzzy logic have been used in numerous applications from well log interpretations to hydraulic fracturing optimization. Therefore, the new interest in Data Mining in our industry is not surprising. Our industry is at its peak state for benefiting from what Data mining has to offer, thanks to abundance of digital data.

A word of caution is in order, which has been the main motivation behind writing this paper. As with many other new tools and technologies, the term “Data Mining” can be, and is currently being, misused in several occasions. In this paper we attempt to answer questions such as, what is Data Mining. How can it be accomplished? What are the essential components of an integrated Data Mining process? What would be the benefits of such a process?

In addition to answering questions such as those mentioned above, this paper will provide a road map (a set of guidelines) for a successful Data Mining project. Finally the paper is concluded by applying the presented guidelines to a hydraulic fracturing data set in the DJ basin for a Data Mining study.

### Introduction

In the past two decades oil and gas companies have spent millions of dollars to collect digital data or to convert the existing data into digital form. This is due to the fact that they have realized the value of data and the potential it possesses in enhancing their operations. IT departments in larger oil and gas companies and major service companies and other vendors have developed sophisticated software tools that allow operators to organize their data, currently existing in different databases, into a cohesive data warehouse and make it available to engineers. Furthermore, more software applications have been developed to put all that information on the engineers and geologists finger tips so they can look at all sorts of data pertaining to a reservoir, a field or a well. Although these are absolutely essential for successful operation of a large company, it has created a new monster. There are far more data that can be effectively analyzed. Human brain, although being the most remarkable information processing entity, can only work simultaneously in so many dimensions and is incapable of processing very large volumes of data. Data mining and knowledge discovery, as an integrated process shown in Figure 1, can come to rescue in such occasions.

Data mining market size was about \$540 million in 2002 and is expected to grow to about \$1.5 billion in 2005<sup>1</sup>. Many industries have realized its value and are jumping on the band wagon of implementing and integrating it into their operations. Data mining uses machine learning, statistical and visualization techniques to discover and present knowledge in a form which is easily comprehensible to humans<sup>2</sup>.

By sifting through large volumes of data already available in corporate data warehouses and extracting patterns, information and knowledge from these databases, data mining allows managers to be proactive. It helps them become prospective in the companies operations rather than retrospective<sup>3</sup>. It must be noted that company managers are not the only beneficiaries of data mining. In the oil and gas industry there are many field related operations that can benefit from the tools and capabilities that data mining has to offer. One of the natural applications of data mining and knowledge discovery processes in drilling, reservoir and production operations

would be identification of best practices<sup>4</sup>. Data mining processes can have as many applications in our industry as our engineers can dream up. It can be applied to identification of new infill drilling locations, optimization of hydraulic fracturing results, candidate selection for stimulations using hydraulic fracturing versus chemical treatments in primary or enhanced production operations and/or storage fields, anticipation of well operation anomalies from real time downhole data, formation evaluation integrating log, and seismic data just to name a few.

There can be different approaches in using data mining processes. Many times the operation is exploratory in nature. One may just want to explore the potential of finding valuable information from an existing database. Other times data miner is after some particular objectives. This is when the process is more guided and has the potential to result in immediate benefits for the company. It is important that the goals and objectives of the project be identified in advance and some metrics for measuring its success be determined.

### Data Mining Classifications

Data mining has recently been enjoying a renewed interest from many different industries. The new interest in data mining has its roots in the large amount of digital data that is being collected and stored in databases and data warehouses. The data owners have spend an enormous amount of resources in the collection and preservation of the data. They would like to utilize this asset by turning data into information and ultimately into knowledge. Data mining has been defined as the major tool for knowledge discovery from the data. A more formal definition of data mining would be; "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data<sup>5</sup>".

The new interest in data mining may be attributed to the fact that the new set of processes that are called data mining are a super set of the processes that previously were known as data mining. The original data mining processes were summarized as a collection of statistical analysis. The new data mining processes include several machine learning techniques as well as statistical analysis. The addition of the recently popularized machine learning and intelligent processes such as artificial neural networks, genetic algorithms, fuzzy logic, and modified cluster analyses have considerably increased the capabilities and utilities offered by data mining.

Many authors have offered different classifications of the processes that are collectively known as data mining. The most appropriate of these definitions (one that suites our industry most appropriately) seems to be the one that identifies two classes of data mining processes. These are descriptive and predictive data mining. It is the belief of the author that descriptive data mining essentially is a subset of predictive data mining. In other words, in order to perform predictive data mining successfully, one, most probably, will have to perform a descriptive data mining first and then use the

information and the results of this process to complete the predictive data mining. Descriptive and predictive data mining share several common processes as shown in Figure 2.

### Descriptive Data Mining

Descriptive data mining is very useful for getting an initial understanding of the presented data. Descriptive data mining is an exploratory process and attempts to discover patterns and relationships between different features present in the database. During the descriptive data mining process the data miner must keep in mind that relevance is an important issue. In other words, the relationships discovered by the miner must be those that users would care about. During this process many non-obvious patterns may pop out that may be of interest to the data owners.

The tools used during the descriptive data mining process are usually consisted of different types of cluster analysis such as hierarchical clustering, k-mean clustering, and fuzzy c-mean clustering. Other popular descriptive data mining tools are rule induction techniques and self-organizing maps. Self-organizing maps utilize unsupervised neural networks such as Kohonene networks.

### Predictive Data Mining

As was mentioned previously, predictive data mining is a super set that should include descriptive data mining as part of its processes, or at least, that is how we would like to define it based on our past experience. During the predictive data mining the descriptive data mining processes are used as a prelude to development of a predictive model. The predictive model can then be used in order to answer questions and assist the data miner in identifying trends in the data. What is most interesting about predictive data mining that distinguishes it from the descriptive data mining is that it can identify the type of patterns that might not yet exist in the dataset but has the potential of developing.

Unlike the descriptive data mining that is an unsupervised process, predictive data mining is very much a supervised process. Predictive data mining not only discovers the present patterns and information in the data it attempts to solve problems. Through the existence of modeling processes in the analysis the predictive data mining can answer questions that cannot be answered by other techniques.

Tools that are used in the predictive data mining process include decision trees, neural networks, genetic algorithms and fuzzy systems. Decision trees are ideal for solving problems that can be dissected into a logical progression of events. The existing types of decision trees include Chi-Square Automatic Interaction Detection developed by Kass<sup>6</sup>, Classification and Regression Trees, CART originally developed by Breimam and Friedman and enhanced by Olshen and Stone<sup>7</sup>, and C4.5 developed by Quinlan<sup>8</sup>.

Neural networks include several types that can be used to solve different kinds of problems. The most popular neural network (actually learning algorithm) is the backpropagation. Most of the neural networks in our industry and most other industries use backpropagation algorithm. They are easy to use and understand. They are capable of solving many complex problems. Other neural networks that are also used to develop models based on historical data are radial basis functions, general regression neural networks and probabilistic neural networks. Nature of the data that is being studied as well as the complexity of the problem usually dictates the type of the neural network that should be used during the model building process. Genetic algorithms (as one of the many analytical tools known as evolutionary computation) is an intelligent search and optimization tool that has proven to be an indispensable tool for any data mining study.

Role of fuzzy logic in data mining and knowledge discovery analyses cannot be over emphasized. Fuzzy logic should form the data miner's basic platform and be used as the fundamental approach to studying complex problems. Fuzzy logic plays an important role in descriptive data mining as well as predictive data mining. If we agree that the data we deal with are instances of reality and nature, and that the complexity of reality and nature cannot sufficiently be explained using the binary system of belief, then the fuzzy logic becomes the most important tool in our data mining studies. For a more complete look at the role of neural networks, genetic algorithms and fuzzy logic please refer to author's articles in Journal of Petroleum Technology<sup>9-11</sup>. There are many books that provide a good background for those interested in this exciting technology<sup>12,13</sup>.

### Components of an Integrated Data Mining Tool

An integrated data mining tool must include the following components. Furthermore, these components must allow interaction such that the result of each component can be used in other components.

1. It must include a module that allows the user to import the data from different sources and combine them into a table that can be used during the analysis.
2. It must have a data cleansing module. The data cleansing module is one of the most important modules of an integrated data mining tool. Quality of the data being used in the analysis determines the degree of success of a data mining. Essential algorithms required for a data cleansing module include:
  - 2.1. Identification and remediation of missing data.
  - 2.2. Identification and remediation of contaminated or erroneous data.
  - 2.3. Identification and remediation of outliers.
3. It must have a clustering module. This must include at least following clustering algorithms:
  - 3.1. Hierarchical cluster analysis.
  - 3.2. K-mean cluster analysis.

3.3. Fuzzy C-mean cluster analysis.

4. It must have an integrated neural network module. Integration of cluster analysis results in the neural network module is of immense importance. It allows the classifications of the cluster analysis algorithms to play a part in the predictive neural network model. The neural network module should include different algorithms for training. The unsupervised algorithms are useful for descriptive data mining while supervised algorithms are essential for predictive data mining processes. Following neural networks are suggested for an ideal integrated data mining tool:
  - 4.1. Kohonen Self-Organizing maps.
  - 4.2. Backpropagation neural networks. This algorithm has several variations that have proven to be very useful on several applications. It is strongly recommended that all the variations of this popular algorithm be present in such a tool.
  - 4.3. General regression neural network. There is a variation of this algorithm that can help data miner when the number of data records are limited. This variation can be of enormous importance to some oil and gas problems.
  - 4.4. Radial basis function neural networks.
  - 4.5. Probability neural networks.
5. It must have an integrated genetic algorithms module. It is very important that the genetic algorithm module can communicate with the neural network module and be able to use the available neural network models as it fitness function.
6. It must have an integrated fuzzy system module. The fuzzy system has to be integrated such that it can use the results of the cluster analysis, neural networks, and genetic algorithms during the development process. The fuzzy module should provide means for automatic and user defined fuzzy set definitions and rule identification.

### Data Pre-processing

Data pre-processing is one of the most important components of a data mining process. It usually consumes more than 50% of an entire project. During a detail and thorough data pre-processing, the data miner has to study the data very carefully and identify the missing cells (data element) in a data record. Sometimes, especially when number of data records is limited, few missing cells can result in eliminating entire data records from the analysis. Furthermore, it is not very easy to detect missing data elements in a large database, and if they go undetected, their damaging effects will be noticed far into the analysis at which time the issue must be addressed and the analysis repeated from the beginning. There are several ways of patching the missing data. The most commonly used method is the conventional statistical method that would substitute an average value of the parameter for the missing data element. This is not the most appropriate way of solving this problem. In many cases this method can over simplify the problem and result in erroneous outcomes, especially during

the predictive data mining. There are other methods that can be used in order to preserve the integrity of the data record while substituting the missing data element with the most appropriate (which means the least damaging) values. The objective of these new techniques is not to magically find the missing data element and substitute it in the database. These techniques, using a combination of neural networks and genetic algorithms, identify the best value that can be substituted for the missing value while the information content of the rest of the data records remain valid and usable.

Identification and handling of outliers is another data pre-processing task that is of utmost importance in a data mining study. It is important to identify whether a data record is truly an outlier or it carries information about the behavior of the system under specific conditions. Domain expertise can become very important in making the proper judgment. Furthermore, it should be identified which data element in a particular data record contributes to it becoming an outlier. The consistency of a particular data element in making a data record to appear as an outlier can provide important information in handling that data element. This can be identified by plotting all the parameters involved in a database against one another and study the behavior of the data record in question. If it is concluded that certain data element or elements are causing a data record to appear as an outlier then those elements may be treated as erroneous and be dealt with as missing data as mentioned above.

Another important pre-processing problem is identification of contaminated or erroneous data records. Sometime, for whatever reason, some data records appear in certain databases, where they do not belong, by mistake (due to human or machine errors) and can cause havoc. Such data records may look very much like other data records but carry fundamentally different kind of information. Combining different databases from different sources that are usually a result of recent mergers and acquisitions can be one of the sources of such problems since different companies use different conventions for record keeping. Sometimes such problems can be solved (of course they have to be identified first, and if go un-noticed can modify the results of the analysis) by referring to those involved in the data collection process or those familiar with the data. Otherwise detection and handling of such issues can become very time consuming. A simple example of such case took place in a company that merged data from a recently purchased property into its main database. The depth reference in two databases were not consistent (one was from KB, while the other was from sea level) and that was causing issues during the analysis. Fortunately an engineer from the newly purchased company was in the staff and was able to address the problem quickly. In the case of the example field study presented in this paper, the contaminated data records were an important problem that took the research team many months to resolve. The team developed new and novel technique<sup>14</sup> using neural networks, and fuzzy cluster analysis to solve this problem. This

technique is applicable to any dataset. Figure 3 demonstrates the components involved in the data pre-processing of a data mining analysis.

### Statistical Analysis

During the past decades petroleum engineers, geophysicists and geologists have come to realized the importance of geostatistics<sup>15</sup> in our day to day operations when dealing with hydrocarbon producing reservoirs. The statistics of data being used in a data mining study provides the analyst with valuable information. Figure 4 shows different components and measures of a statistical analysis that should be used in a data mining study. The linear regression between different parameters in a database can reveal any visible and readily detectable relationship that might exist between different parameters. As the relationship between parameters becomes more complex, tools like linear regression will no longer be useful and the data miner needs to use more sophisticated techniques. These techniques include principal component analysis<sup>16</sup>, fuzzy curves<sup>17</sup>, and fuzzy combinatorial analysis<sup>18</sup>.

One of the most common and basic statistical analysis that are performed on all the parameters in the database is identification of their probability distribution function, along with minimum, maximum, mean, and variance. Based on the nature and type of the parameter (continuous, versus categorical data types) mean, median or mode of the parameters is calculated. Chances are that the distribution for most of the parameters does not follow the characteristics of a normal distribution. In such cases calculation of distribution kurtosis can help the analysts in their analysis.

### Descriptive Analysis

Two of the most common methods for descriptive analysis of the data are cluster analysis and feature ranking as shown in Figure 5. Clustering describes a collection of unsupervised methods whose aim is to partition an overall data set into a significantly smaller number of "clusters". These methods in general require some kind of distance measure among the data entities in order to group them together and identify each data entity with one cluster.

Most clustering algorithms partition the data based on how similar individual records are; the more similar, the more likely that they belong to the same cluster. Their main purpose is to identify clusters which maximize the inter-cluster distance and minimize the intra-cluster distance, so that we obtain clearly distinct groups of similar entities. This grouping introduces a "natural" unsupervised classification schema based on similarities according to the given distance measure<sup>19</sup>.

There are several types of clustering techniques, such as hierarchical clustering, k-mean clustering and fuzzy c-mean clustering. Hierarchical clustering does not partition data into a particular number of clusters in a single step. Instead, a series of partitions takes place, which may run from a single

cluster containing all objects to  $n$  clusters each containing a single object. K-mean clustering divides the database into  $k$  clusters identified by the user such that the distances between the objects in a cluster are minimized while the distance between clusters are maximized. In k-mean cluster analysis each object (a data record) will fully belong to only one of the clusters. In a fuzzy c-mean clustering data records are assigned to different cluster centers to a degree. In this case a data record may have a membership of 0.7 in one cluster while having memberships of 0.05 and 0.25 in two other clusters<sup>20</sup>.

The other descriptive analysis is feature ranking. It is also known as identification of performance drivers or identification of parameters influence. It is the believe of the author that a reliable technique in ranking the importance of parameters in an oil and gas related database (due to the nature and complexity of the problems in our field) must take into account the influence of the parameters on one another as they collectively influence the outcome. As an example, in a hydraulic fracture treatment, several additives are used throughout the treatment. The ultimate objective of these additives is to enhance the hydraulic fracture treatment outcome (fracture length or conductivity). The additives will have an effect on the fracturing fluid but in many cases they will have an interfering effect on each other. This interfering effect must be taken into account when feature ranking analysis is to take place. Another industry that can benefit from such analysis is the pharmaceutical industry testing the effect of different drugs on patients. Fuzzy combinatorial analysis<sup>18</sup> takes such interfering effects into account.

### Predictive Analysis

In order to perform predictive data mining analysis, the analyst must develop a predictive model. There are many techniques for developing predictive modeling. First and foremost are deterministic models that we all are familiar with. These are models that have their basis on physics and are developed through rigorous mathematical manipulation of physical concepts. If you are involved in a data mining project, chances are that the developing a deterministic, physics-based model is impractical, either due to complexity of the problem in hand or for lack of availability of needed information. This is when data-driven modeling becomes the best alternative.

Neural networks have proven to be a great tool for data-driven model development. A successful neural network modeling process will require all the information that is generated during the descriptive data mining process. This information must be integrated in the statistically representative partitioning of the database into training, calibration (testing) and verification (validation) of the developed model, in order to increase the possibility of developing a representative model.

Most of the predictive data mining activities take place upon

completion of predictive model development process. This is when many “what if” scenarios can be played out and detail sensitivity and parametric analysis based on multiple parameters can be performed. During such analysis two dimensional plots can show the sensitivity of the outcome as a function of all possible values of a parameter. Using three dimensional plots sensitivity analysis can be performed on two parameters at a time studying the changes in the outcome as a function of two parameters. When it is desired to study the effect of several parameters (more than two) on the outcome simultaneously, Monte Carlo simulation can be used to show the potential probability distribution of the outcome as a function of several parameters, each of which must be assigned a probability distribution function.

The next potential step in a predictive data mining study is to use the data driven model in an optimization study. Employing an integrated genetic algorithm routine and identifying the parameters that are to be optimized, the data driven model can be used as the fitness function of the genetic algorithm in order to find the optimum combination of the values for parameters being studied in order to achieve the objectives of the optimization study. Genetic algorithms can be constrained through pre-determined rules that would make sure that the solutions suggested by the optimization process pertain to the realities on the ground.

Another approach would be the development of hybrid robust models that would augment the data-driven models with expert knowledge. This data-knowledge fusion technique can markedly enhance the performance of predictive models<sup>21</sup>. Using fuzzy logic the expert knowledge can be coded into semantics and used in fuzzy systems that can combine the results of data-driven models with expert knowledge and result in superior predictive models. This will prove valuable in cases where the available data will not necessarily cover all the possible cases that may exist in a process.

### Example in DJ Basin

The example application discussed here is a short summary of the results of a study that has been presented in a previous SPE paper<sup>4</sup>. Its summarized presentation in this paper is mainly for the completeness of this paper. The study was conducted in DJ basin. Patina Oil and Gas has been one of the most active operators in the United States in restimulation of tight gas sand wells. Patina has over 3,400 producing wells in the basin, and has restimulated over 230 Niobrara/Codell completions so far.

The original database for stimulation of Codell wells in the DJ basin needed considerable quality control in order to remove erroneous records<sup>14</sup>. Once the quality control of the data set was completed, Fuzzy Combinatorial Analysis (FCA) was performed. The analysis was performed for the combination of up to five features, meaning that influence of parameters (up to five) on one another as well as on the outcome was

examined. Table 1 shows the top 10 parameters in the database. This table shows that seven out of the top ten parameters that seem to be controlling the hydraulic fracturing outcome in this field are operational parameters that can be controlled by the operator. The other three (Lat., Long. and Codell Gas-ft) are indicators of geology and reservoir quality. This provides important information to the operator as to which parameters should be concentrated on, in order to get the most out of the stimulation jobs.

Upon completion of a cluster analysis on the database, a predictive model was build for this field. The predictive model was then used to study several "What If" scenarios. Two, three and multi-dimensional (using Monte Carlo simulation) sensitivity analysis was performed on the predictive model for each of the wells involved in the analysis. Then using the cluster analysis results (which identified three categories of similar wells in the field) was used to perform analysis to find the best practices for cluster of wells. And finally another set of best practices studies were performed for the entire field.

The full field analysis revealed that higher amount of proppant (after a certain concentration has been reached) does not contribute to the success of the frac jobs and therefore are not recommended. Another conclusion based on the full field analysis was the use of low viscosity fluids in the frac jobs. Source of the water used in the operation also was shown to play an important role in the success of the frac jobs. Once these items were identified as parameters that control the success of the stimulation jobs, the predictive model are used to design the frac jobs for new wells in the field. The importance of the above conclusions are realized when one looks at the type of data that such conclusions have been drawn from. Figure 6 shows the scatter plot of amount of proppant and peak viscosity versus the post-stimulation deliverability. These figures show that there are no visually detectable trends in this data and without tools such as those used in this study (an integrated data mining software application specifically for the oil and gas industry currently under development) identification of such patterns and making such recommendations would have not been possible.

### Conclusions

Data mining and knowledge discovery has a lot to offer to the oil and gas industry. The technology is fairly new in our industry and can be, and will be misused as part of a natural growth. Many processes that are not necessarily data mining or have only small components of it, will be called data mining to ride the waves and the hypes surrounding the technology. These should not deter the industry to aggressively pursue this exciting technology that can bring about a new way of looking at many different aspects of our industry, from technical problem solving to management level decision making.

The next step in the evolution of this technology in our

industry is development of specialized and integrated software tools that can help our engineers and scientists get the most out of several analytical techniques that is offered by this technology. The essential components of such a tool were outlined in this paper. Any application that lacks several of these components cannot be considered an integrated data mining tool. On the other hand, there might be other techniques that might have been omitted in the list of components provided in this document. As some new techniques prove themselves in other industries and show promise of having the potential to contribute to a data mining process, they can be added to the list of essential components mentioned in this document.

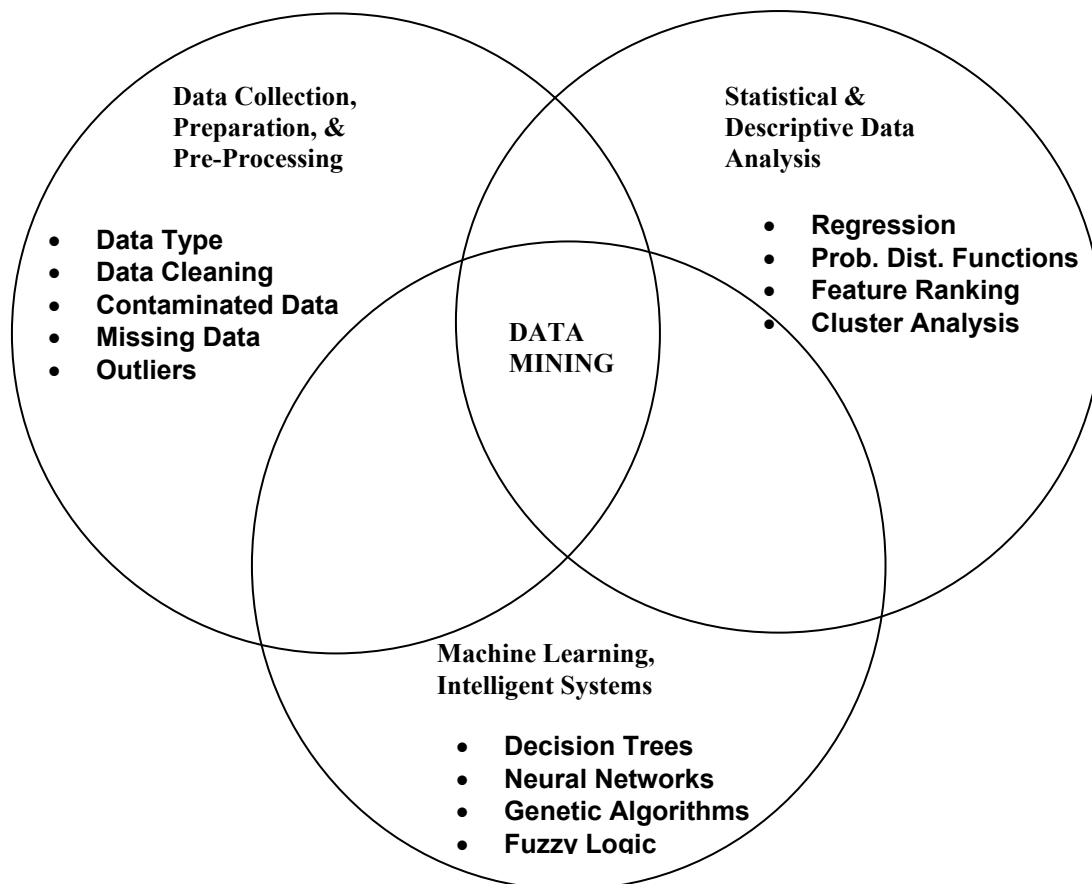
### References

1. According to IDC, world's leading provider of technology intelligence, PC AI magazine, January 2003.
2. M. J. A. Berry, and G. S. Linoff, *Mastering Data Mining*, Wiley Computer Publishing, John, Wiley & Sons, Inc. New York, NY, 2000.
3. C. Westphal, and T. Blaxton, *Data Mining Solutions*, Wiley Computer Publishing, John, Wiley & Sons, Inc. New York, NY, 1998.
4. Mohaghegh, S., Popa, A., Gaskari, R., Ameri, S., and Wolhart, S.: "Identifying Successful Practices in Hydraulic Fracturing Using Intelligence Data Mining Tools; Application to the Codell Formation in the DJ Basin", *SPE 77597, Proceedings*, 2002 SPE Annual Conference and Exhibition, September 29 – October 2, San Antonio, Texas.
5. W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview. *AI Magazine*, Fall 1992, pgs 213-228.
6. G.V. Kass, An exploratory technique for investigating large quantities of categorical data *Applied Statistics*, 29, 119-127. 1980.
7. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. Classification and Regression Trees. Wadsworth International, Belmont, Ca.
8. Quinlan, J. R. 1993. C4.5: Programs For Machine Learning. Morgan Kaufmann, Los Altos.
9. Virtual Intelligence Applications in Petroleum Engineering: Part 3 – Fuzzy Logic. *Journal of Petroleum Technology*, Distinguished Author Series, November 2000, pp 82-87.
10. Virtual Intelligence Applications in Petroleum Engineering: Part 2 – Evolutionary Computing. *Journal of Petroleum Technology*, Distinguished Author Series, October 2000, pp 40-46.
11. Virtual Intelligence Applications in Petroleum Engineering: Part 1 – Artificial Neural Networks. *Journal of Petroleum Technology*, Distinguished Author Series, September 2000, pp 64-73.
12. Berthold and Hand, *Intelligent Data Analysis: An Introduction*, Springer Verlag; 2nd edition, April 15, 2003.

13. Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Verlag; August 9, 2001.
14. Popa, A., Mohaghegh, S.D., Gaskari, R., and Ameri, S.: "Identification of Contaminated Data in Hydraulic Fracturing Databases: Application to the Codell Formation in the DJ Basin", *SPE 83446, Proceedings*, 2003 SPE Western Regional Conference and Exhibition, May 20 – 24, Long Beach, California.
15. J. L. Jensen, L. W. Lake, P. W. M. Corbett and D. J. Goggin, *Statistics for Petroleum Engineers and Geoscientists*, Second Edition, Elsevier, Amsterdam, The Netherlands, 2000.
16. I. T. Jolliffe, *Principal Component Analysis*, SpringerVerlag, New York, 1986.
17. Lin, Y., and Coningham G.: "A Fuzzy Approach to Input Variable Identification", proceedings of the third IEEE International Conference on Fuzzy Systems, June 26 - July 2, 1994, Lake Buena Vista, Florida.
18. Mohaghegh, S., Gaskari, R., Popa, A., Ameri, S., and Wolhart, S.: "Identifying Best Practices in Hydraulic Fracturing Using Virtual Intelligence Techniques", *SPE 72385, Proceedings*, 2001 SPE Eastern Regional Conference and Exhibition, October 17-19, North Canton, Ohio.
19. Similarity Clustering, Thomas Prang 1998.  
<http://www-lehre.informatik.uni-osnabrueck.de/~ftprang/papers/tproject/node27.html>
20. L. I. Kuncheva, *Fuzzy Classifier Design, Studies in Fuzziness and Soft Computing*, Physica-Verlag, New York, NY, 2000.
21. Mohaghegh, S., Reeves, S., and Hill D.: "Development of an Intelligent Systems Approach to Restimulation Candidate Selection", *SPE 59767, Proceedings*, SPE Gas Technology Symposium, Calgary, Alberta, April 2000.

Rank	Parameter
1	Flow Back Volume, (bbl)
2	Codell Gas-Ft
3	Bicarbonate, (ppm)
4	Peak Viscosity
5	Latitude
6	Amount of Sand (Mlbs)
7	Longitude
8	Date of Refrac
9	Viscosity Shear 100-30 min.
10	Total Hardness (ppm)

**Table 1.** Most influential parameters in the DJ basing hydraulic fracturing program.



**Figure 1.** Data mining as an integrated analytical process.

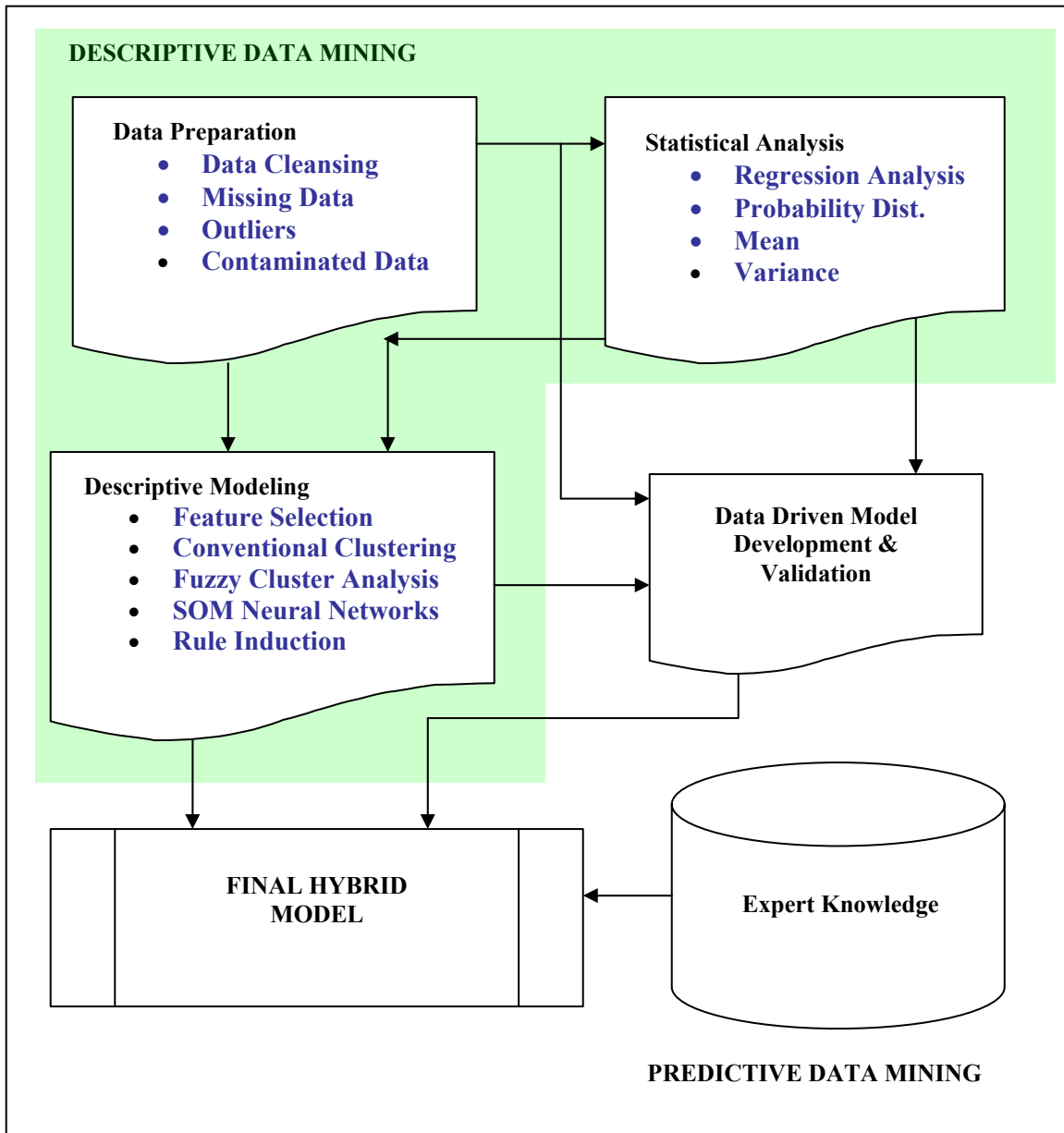
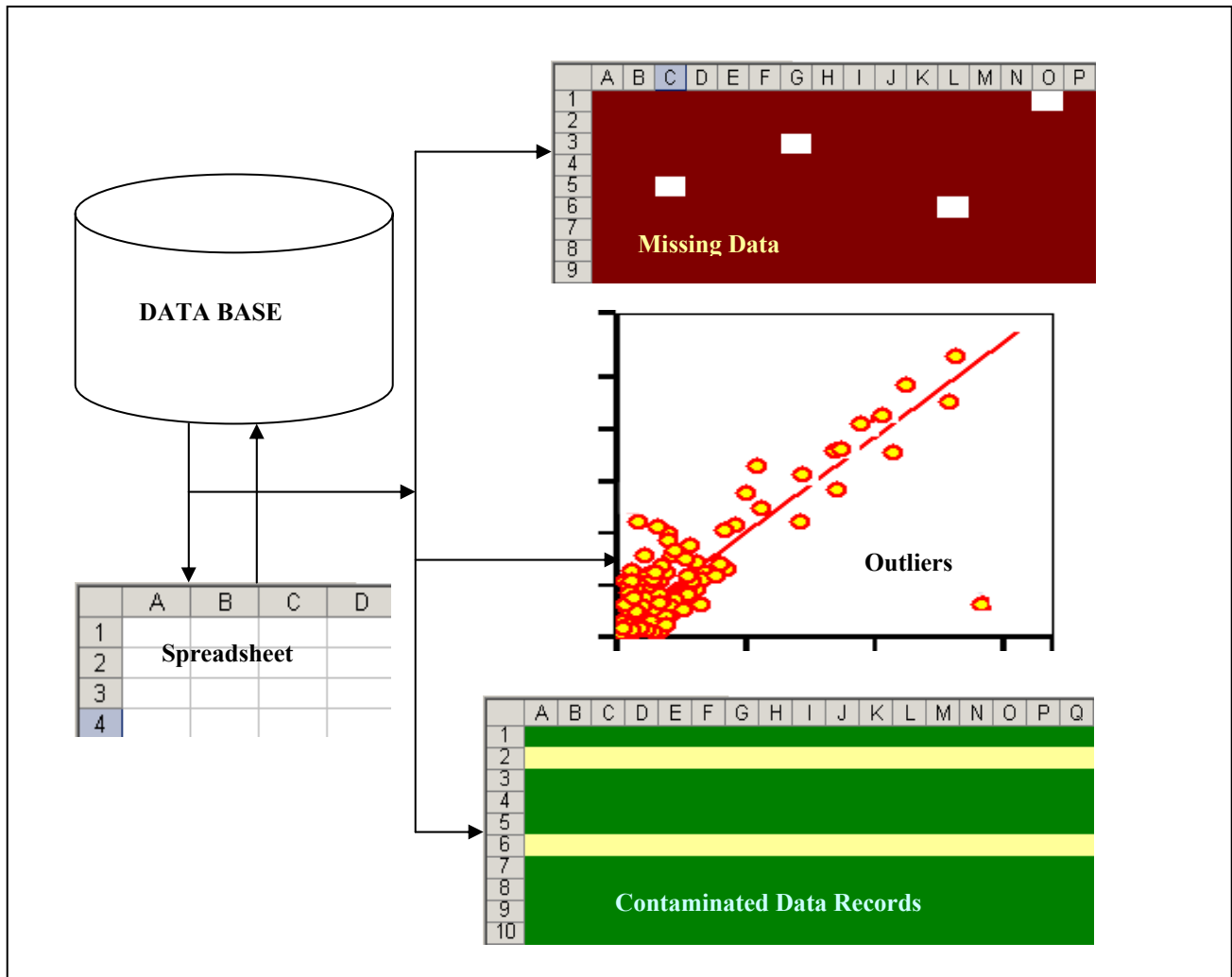


Figure 2. Two classes of Data Mining with details.



**Figure 3.** Important components of data pre-processing.

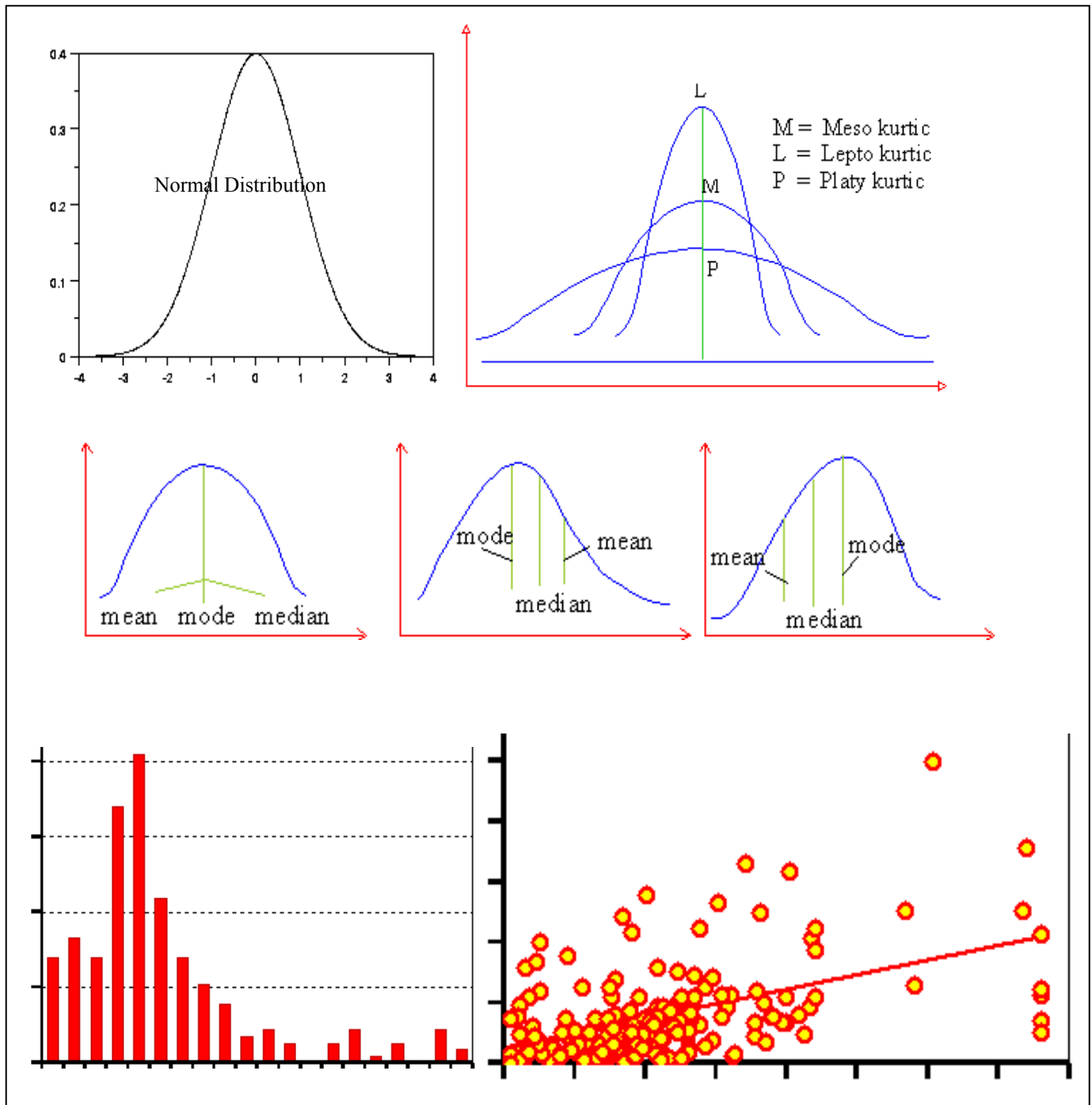


Figure 4. Statistical analysis as part of data mining.

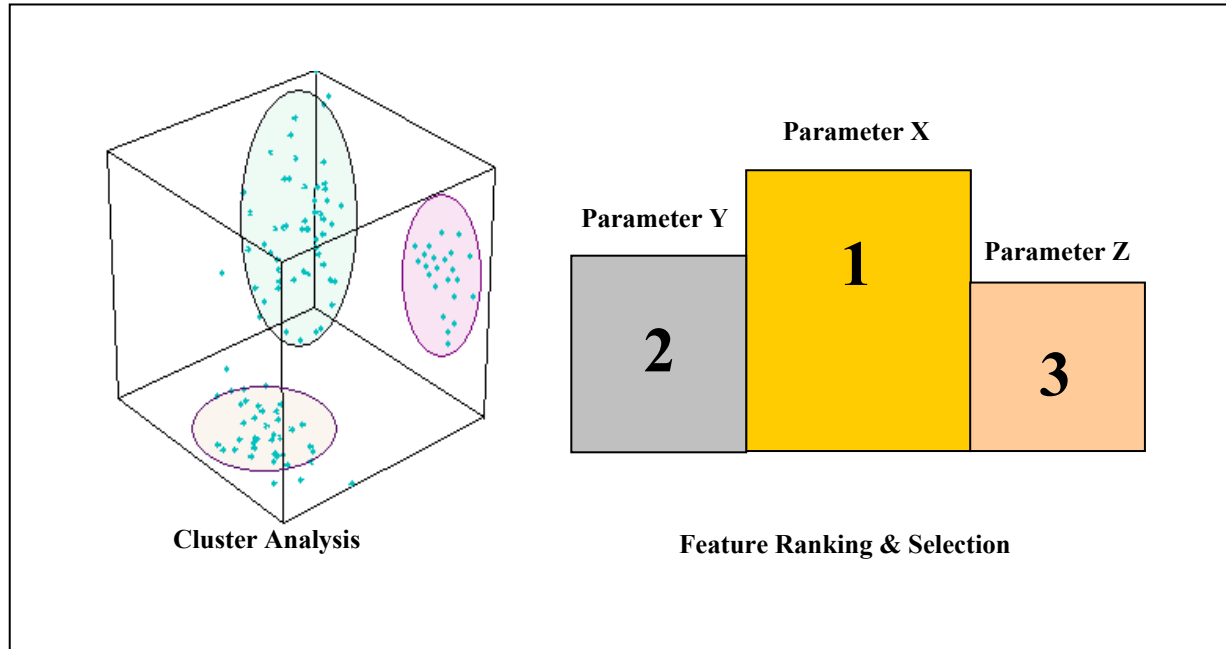


Figure 5. Cluster analysis and feature selection, part of descriptive data mining.

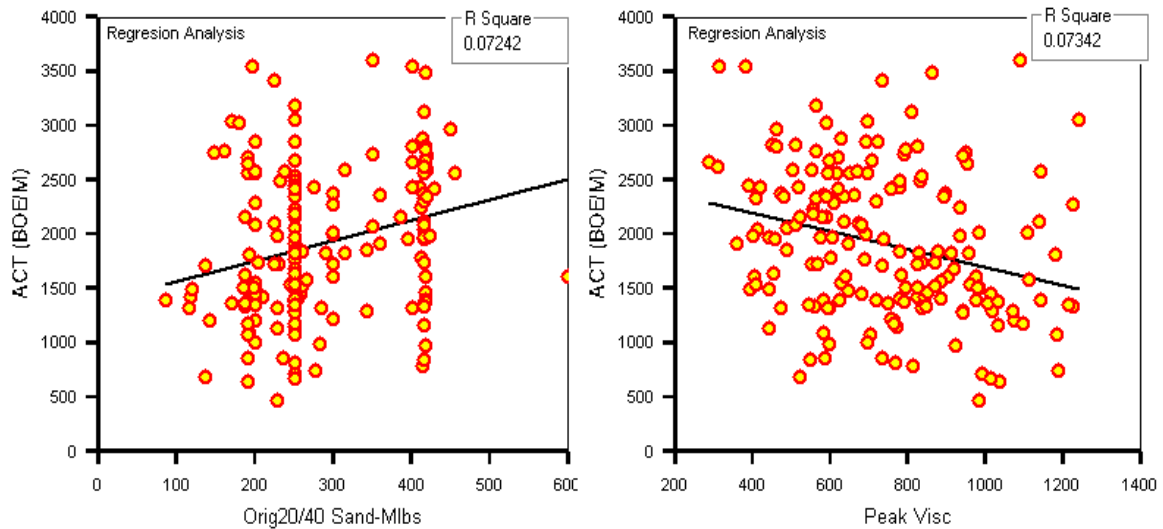


Figure 6. Proppant amount and Peak Viscosity data versus post-frac deliverability.