



Z Solutions®

IMPEDIMENTS TO EXPLORATORY DATA MINING SUCCESS

This white paper is a summary of a chapter written by Jeff Zeanah published in Organizational Data Mining, Idea Group Publishing 2004

INTRODUCTION

Organizations of all kinds are experimenting with the application of data mining techniques. They may refer to these projects as data warehousing applications, market research or data mining. Regardless of the terminology, data mining applications are intended to provide an organization with a better understanding of the environment or market in which they operate.

There are two general types of data mining undertakings. Relatively well understood are the traditional scoring applications in which observations are scored to determine if they met certain criteria. In these projects, an organization typically will apply a set of tools to a large database such as a mailing list. For example, a charity considering a mail out to solicit donations will score their list to determine the most likely candidates to be solicited. By using data mining to examine the characteristics of individuals who donated in the past, the charity can reduce a mailing list of two million households to a list of the 200,000 households most likely to donate. Soliciting this smaller list will be more profitable than “wasting” a mailing to the 1,800,000 households that are very unlikely to respond. In this effort there is not a great need to understand why the households were selected, only whether or not the refinement of the list leads to a higher response rate and increases the profitability of the mail out.

Conversely, exploratory data mining is designed to provide strategic insights from the data and guidance for future strategic or operational decision-making. Consider the following example. A consumer products manufacturer is interested in characteristics of the consumers who buy their products rather than their competitor’s. Are these customers younger or older? Are they married or single? What is their ethnicity? Are their household incomes higher or lower? Simple queries of the company’s data warehouse can be used to answer most of these questions. If the data is available, the average household income of the company’s customers easily may be compared to the average household income of competitor’s customers. This is exploratory data mining. And for many organizations, the ability to read the databases, perform these queries, is the extent of their data mining activities. In fact, for very large databases this can be nontrivial, requiring substantial effort.

Simple queries may not provide all the answers a company needs. For example, the company discovers that their customers have higher household incomes than the competitor’s customers have, are older and are more likely to be married. The organization realizes that many married families have higher incomes than do single households (two incomes versus one) and many older households have higher incomes than do the younger beginning households. The company wonders: “Are the customers who prefer our product older and happen to be married at a higher percentage or are our customers married and happen to be older?” Or are they just higher income people? Whatever the relationship, is this the same as it was five years ago? And most important, what will they be like five years from now? If the overall population ages, will that help the company sales? Will it help the company’s sales

only if the aging population stays married? Is the company not seeing a hidden trend that may change the company's strategic direction? These questions are not going to be answered through simple methods. The solution will be found only through causal predictive modeling and similar investigations.

Many organizations lack the ability to answer these difficult questions. When they began collecting data in their data warehouse, they expected that they would be able to resolve some of these difficult strategic puzzles. Their exploratory findings are less than they had hoped. They can identify what has happened but have a more difficult time answering *why* it happened.

This paper discusses reasons for these shortcomings. These observations are based on projects reviewed by the author supplemented by discussions with over one hundred data mining professionals. Based on these observations, four impediments to exploratory success have been identified. Conditions leading to the impediments are discussed and solutions presented. The four impediments are:

- data quality
- lack of secondary or supporting data
- insufficient analysis manpower
- lack of openness to new results.

The projects that provided the foundations for these conclusions are proprietary efforts from private corporations and public sector organizations that are seeking to improve their understanding of their environments. Because of the proprietary nature of the projects specific situations and data details are not presented but are discussed in general terms.

IMPEDIMENT 1: DATA QUALITY

Lack of data quality is an obvious impediment to successful exploratory data mining. However, the root causes of the problems are not obvious. Given the obvious requirement for data quality, why are so many organizations

surprised to discover the data collected turns out to be less than anticipated? Certainly data quality is not an oversight. Therefore, what leads to bad data?

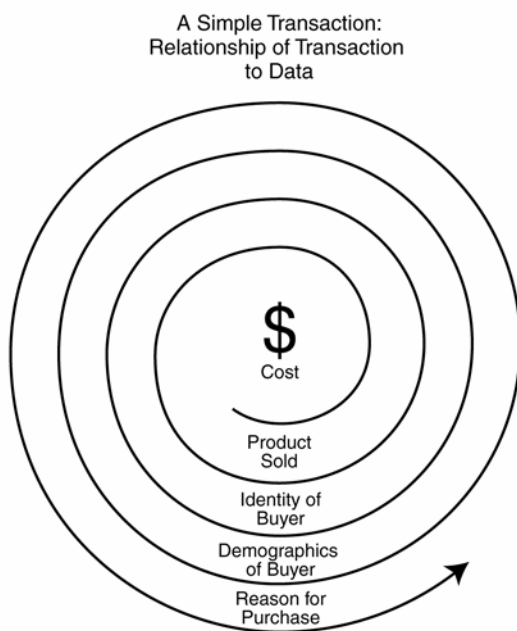


Figure 1

There are, of course, examples of gross errors, such as faulty data entry programs or faulty metering and storage issues. However, there are things an organization or manager seeking to improve data should look for. The manager should consider two general trends when reviewing data collection plans: does the data contain a financial transaction or is it close to a financial transaction; and what are the incentives or disincentives to collecting accurate data?

Experience has shown that data quality is in direct proportion to the closeness to a financial transaction. As demonstrated in Figure 1, the further away from the transaction, the greater the concern should be for data quality. Organizations are usually very successful in keeping track of the flow of money. The record-keeping of a financial transaction is usually quite accurate. However the further the data is from the transaction, the lesser the data quality. Consider the example of a company that provides

settlement of workers' compensation claims. When analyzing these claims to find patterns concerning the value of claims, the amount of workers' compensation payment was found to be reliable. This is expected. If the organization did not keep track of payments, the company would not be in business long. Further work with the data reveals that the recipients of the payments were reliable to a lesser degree. There was more uncertainty in some of these data fields. The name might be correct, but the address of the person might be missing or the zip code incorrect. Or perhaps the name was misspelled making it impossible to match the record with other files. The individual identifier information is further away from the financial transaction and did not have a high degree of data quality. Other necessary information for the completion of the study included the type of injury leading to the claim. This information was actually not necessary for the financial transaction. The company found this data was the lowest quality data. Often the data was missing or apparently inaccurate. For example, it is difficult to explain the loss of work of six months for a minor sprained ankle. Other unrecorded mitigating factors had to be in play.

The second corrective action for data quality is to recognize that "what gets rewarded, gets done." In the above example the field agents had a data entry system to collect information and return that information to the corporate office. However, the field agents were compensated for the number of claims handled. The implementing and structuring of compensation packages is well beyond the scope of this discussion. However, the impact on data quality of this situation is intuitive.

In most organizations the sales staff is compensated with sales commissions. There is an obvious incentive to complete the information on the volume of each sale. However, if there is an expectation that the sales staff is to collect demographic information and the collection of that information is not rewarded, then the results will be obvious. The quality of the demographic information will not be as high as the volume of sales. If data is to be a value to the organization, then collecting that data has to be of value to the employees. Data quality needs to be measured and made an organizational goal.

IMPROVING DATA QUALITY

Situations like the above are rampant in most organization. Frequently there is an acceptance of the data quality situation as, "that is just the way it is." Through downsizing and reorganizations, workloads are increased in most organizations. Certainly, there are economic benefits from the resulting efficiencies, but frequently something has to slip as a result. Frequently that something is data quality. Many of the following issues are present in the typical organization:

- No one is giving the data a thorough evaluation on an ongoing basis.
- There is an assumption that the data is better than it is.
- Systems to access the data are limited. There is little in the way of comparative reports or graphics to quickly visualize the data.
- Older data is archived, therefore difficult to work with or compare with new data.
- More attention (and money) is put on storage equipment (hardware and software) than to the actual data itself.
- Little screening is being done testing data and tolerances when the data is entered.

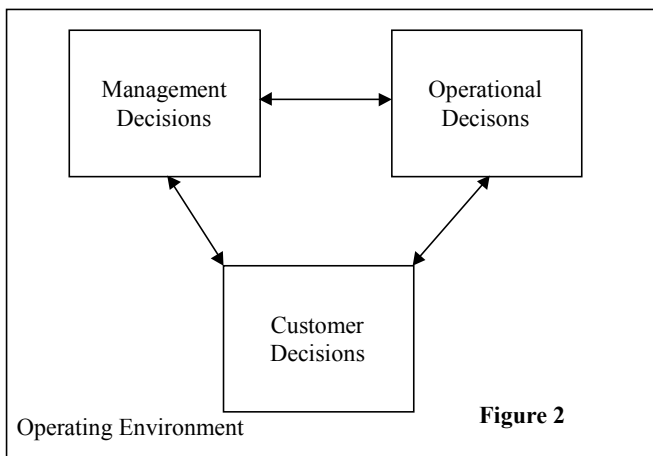
Raising the Issue of Data Quality

What is the cost of reduced data quality? It is hard to know. Organizationally it becomes difficult to present the issue. Even as presented here it sounds like a petty whine -- not the type of thing to get broad organizational or upper management support.

Improving data quality depends on understanding and utilizing one simple fact. Bad data is usually not going to get the organization's attention, but bad decisions will. Demonstrating the cost of bad decisions or the potential of a future bad decision will make the point. From a practical point of view, it is difficult, if not impossible, to calculate the Return on Investment (ROI) implications of bad data, however, you can calculate the ROI implications of a bad decision.

IMPEDIMENT 2: SECONDARY OR SUPPORTING DATA

Organizational data mining is not just the analysis of data warehouses and large databases or analysis of customer databases. It is an evaluation of an organization's entire operations including the environment in which the organization operates. Figure 2 presents a simple Operational Schematic describing a manufacturing company. The company makes decisions such as: what product lines will be offered; how the products are manufactured; and



product pricing. The decisions are implemented through the company's operations and lastly there is an interface with the customer – the customer purchases and uses the product. This total operation takes place in an operating environment affected by economic conditions, competitor actions and other factors not in control of the company.

The interaction between the customer and the company is the focus of most data mining activities, which try to understand customer decisions by asking questions such as what and how much does a customer purchase? The answers to such questions are usually

the primary focus of the activity when data warehousing activities are discussed in an organization. However, data mining activities do not have to be limited to the interaction between the company and the customer. In fact, data from each of the areas in the operational schematic can have impact on the customer / company interaction. This data is usually not voluminous information that belongs in data warehouses. It is smaller -- sometimes only a few hundred records. This data can be thought of as secondary or supporting data. It is not the primary requirement of a data warehouse, but it is the data that makes the primary data (the focus of the expensive development) valuable.

This concept is not new. Supporting data is frequently brought into an analysis, without much recognition. A typical example is when a company is analyzing product sales over a period of years. Often included in this analysis is data related to economic conditions, such as Gross National Product, interest rates and unemployment. The organization determines their sales performance based on general economic conditions. Therefore, some downturn in sales can be attributed to economic conditions, not to a lack of success of the company related to their competitors. The company may be using a large database of product sales, which resides in the company data warehouse. However, the company does not keep track of economic data and is fortunate that others keep this data for it.

Using external secondary data to improve an analysis is common. But, unfortunately, when there is a need for secondary and supporting data from internal sources, we can have an altogether different story. This secondary data does not get the attention of the data warehouse data, but it can be as or more important.

IMPROVING SECONDARY OR SUPPORTING DATA

Not having the necessary secondary or supporting data can be an impediment that can stop a data mining activity very quickly. The corrective actions for this problem are twofold. Simple in concept, these corrective actions require diligence to implement.

Recognition

The first corrective action is recognizing that the secondary data is important to the process. In the development of data warehouses the concentration cannot be limited to just voluminous data. The organization must remember the issue of smaller databases. This will greatly increase the likelihood that the data will be available when needed.

Anticipation

The second corrective action is a widespread use of personal databases throughout the organizations. Most office software suites come available with a database programs. Making these products available throughout the organization and providing training and support for these products can provide unexpected dividends.

IMPEDIMENT 3: ANALYTICAL MANPOWER

Stories are common of data warehousing projects that have exceeded time and budget expectations. Significant executive support and guidance are usually needed to complete these projects. Staffing of Information Technology (IT) organizations is increased for the development or outside consultants are used. These projects stretch the resources of the organizations involved both in terms of financial resources and the internal political capital needed to keep the project alive through setbacks and inevitable delays. Frequently these same projects do not follow with additional emphasis on using the data. The last thing the organization wants to hear at the end of such a project is, "It is now time to add additional resources and staff to analyze this data."

A recent search of Monster.com shows anecdotal evidence of this difference in emphasis. A search on the term "data warehouse" returns 1002 possible job opportunities in the United States, while a search on the term "data mining" produces 258 possible jobs.

Frequently the thought is that the analytical staff in place is going to be able to use the information. If it is marketing data that is being collected, then the present marketing staff will analyze this information. If it is manufacturing data, then the production engineers are going to use it. This plan is faulty: more data requires more time to analyze. Furthermore, exploratory data mining is an extremely complicated task requiring several specific skills including: programming; statistics and predictive modeling such as neural networks; database processing (queries, structure, etc.); and presentation skills including graphical display of information and written presentation. In addition to the above skill set, a successful exploratory data miner must have a good understanding of the business problem or the organization's mission, must know how to apply the information in the data to the problem and must understand the meaning of the information. The key to understanding the manpower impediment is to understand the following axiom.

Chances are that if a person exists presently in an organization who has the broad range of technical capabilities to perform exploratory data mining and that person has the experience and knowledge in the organization to apply the information to be gathered from the data, then that

person is likely overworked now. They are already overtaxed, and having new data does not mean they have time to fully utilize it.

REMOVING THE MANPOWER IMPEDIMENT

The key to removing this impediment is to recognize the need for manpower and to plan ahead. The organization should plan for the analysis and recognize that additional data means additional effort.

Estimating required manpower and the cost effectiveness of additional manpower is case specific. There is no general rule of thumb that can be applied to each situation. A careful review of the steps and processes needed for data mining can be reviewed and will provide a foundation for manpower estimates. The review should follow along the lines of the following major categories of effort.

- Data maintenance
- General queries
- Data preprocessing
- Data analysis
- Developing conclusions
- Presenting results

As the organization contemplates the impending effort, it should first think of foundation systems. Foundation systems are the basic software systems and customized code needed to provide the tools to answer most requests. The work on these foundation systems has to be completed before the data can be analyzed in a timely manner. Through the addition of staff (or undertaking staff augmentation through consulting services) these foundation systems are required.

IMPEDIMENT 4: LACK OF OPENNESS TO NEW RESULTS

The last impediment to the organization's fully utilizing the value of exploratory data mining is referred as a "lack of openness." The opposite of this lack is a "sense of openness." At the heart of this sense of openness is the recognition that the value of exploratory data mining is the ability to discover what the organization presently doesn't know and the more difficult task, to correct incorrect beliefs. The exploratory data miner's greatest success is a change in organizational thinking. But change is difficult.

Dr. Jonah Folkman of Children's Hospital of Boston is a leading cancer researcher. His ideas include developing treatments to effectively starve cancer growths by limiting their blood supply. His theory, which he called angiogenesis, was based on the fact that tumors secrete a factor that stimulates new blood vessels to form, supporting the tumor with a private blood supply. Starving the tumors instead of trying to kill the cancer has the potential to change not only how we treat cancer but also how we view cancer. In the beginning, his work was not well accepted, but Dr. Folkman persevered. He explained his doggedness this way, "We've always said there is a fine line between persistence and obstinacy in research and you never know when you have crossed that line." (All Things Considered, 2001) But his persistence in taking a risk and working in a direction at odds with the mainstream establishment has revolutionary potential. Exploratory findings that may have a significant impact on the organization must have an environment in which they can develop. An organization with a sense of openness allows an exploratory data miner the ability to be persistent.

Three requirements are needed to create the Sense of Openness. They are: executive sponsorship, a reduction the emphasis on statistical accuracy, and for the data miner to present exploratory findings with good documentation and support.

Executive Sponsorship

Dr. Robert Kriegel (1996), a leading writer and lecturer on organizational change, believes resistance to change is personal. In his book *Sacred Cows Make the Best Burgers*, he lists four personal resistance drivers:

- Fear – “What if... I lose my job, look stupid, can’t adapt,” etc.
- Feeling Powerless – “No one asked me!”
- Inertia – “It’s too much effort, too uncomfortable.”
- Absence of Self-Interest – “What’s in it for me?” (Kriegel, 1996, p. 195)

Dealing with these personal resistance drivers is an organizational issue, but they do impact the completion of data mining projects. A manager’s fear that his understanding of the marketplace, developed over 30 years, no longer applies is powerful. In addition to the fears identified by Kriegel we can add the fear of “What if I am wrong?” Feeling powerless, left out, when new techniques are used that you don’t understand is an equally strong deterrent to change. The same can be said for inertia. When a product is at the top of a cycle of market share, who wants to say, “Now is the time to make changes”? Even though, we all know that products have life cycles and sales go up and sales go down. And we have all seen great products, once unstoppable, reduced in significance. It makes an organization very uncomfortable to discuss a potential change from this lofty position. The desire to believe that the present situation will continue provides inertia that is hard to overcome. For new exploratory findings to have a significant impact, these personal and organizational issues must be understood and addressed.

Although executive sponsorship is cited as the requirement for most organizational accomplishments, ranging from human resource programs to recycling programs, rarely are we told exactly how to apply it. However, the requirement here is clear. An acknowledgement is required that an effort is underway to investigate information to reveal what is not presently known or what is incorrect. Presently held beliefs *can* be questioned. Furthermore, it is still acceptable to continue the research when it is revealed the initial questioning was incorrect.

Heresy: Ignore Statistical Accuracy

This section is intentionally mislabeled to make a point – to state clearly what we are not saying before the point can be misrepresented. The recommendation is *not* to ignore statistical accuracy, rather it is to temporarily drop or reduce the requirements of statistical accuracy.

In most projects, we analyze a sample of a larger population. Analysts work from a representative sample of the population, drawn randomly. Therefore, there can be variation in the results – leading to the need to understand the variation and determine if the sample is accurate.

The exploratory data miner’s job is to find new relationships, relationships that we don’t know exist. Often these relationships are found outside the main view of the organization. The researcher often is faced with a dilemma: finding new relationships often pushes the data to its limits, these new relationships are difficult to “prove” with statistical support. Should the researcher withhold this information until new data is available, which may be a lengthy delay, or should the researcher report the new discovery and begin speculations about the new findings. In the organization with a sense of openness, the speculation will open new discussions about the topic. The discussion will suggest new areas of research to be explored to answer the questions the speculation raises. Even if some of the

statistically questionable initial ideas later prove wrong, the organization will benefit from the focus on the new issues. In the organization with the sense of openness, the researcher is confident that his speculations will be used appropriately. They will not be confused with research findings, but the speculations will instead be the foundation of potential new knowledge.

Presenting Exploratory Findings

For speculations to be treated as speculations, then it must be clear what an exploratory result, a finding, must be. These final results of the exploratory analyses should be properly documented and circulated throughout the organization.

The goal of exploratory data mining is finding relationships and trends that are not readily apparent. In order to show these difficult findings, the researcher must clearly and completely articulate what has been discovered and offer supporting documentation for the discovery. It should be stated clearly to what the results apply. In order to differentiate these findings from speculations, a format for presenting results, such as the following, should be used.

Finding: Women under 25 who buy product X are three times more likely to report that they are interested in using the product for “fun” than women under 25 who buy products Y and Z.

Support

1. Customer surveys from 1998-2001 used for the analysis. Of the women under 25 that responded, there were 98 responses from purchasers of product X and 154 responses from purchasers of products Y and Z. Of the product X responses, 54% listed “fun” as a reason for purchase; of the product Y and Z responses, 17% mentioned “fun.”
2. Women of all other age groups did not show the same emphasis on “fun.” Of those responses, 16% listed “fun,” if they purchased product X, and 18% listed “fun” if they bought products Y and Z.

The finding states clearly what the exploratory finding is about – the reason for buying the product. The statement also makes very clear the dimensionality of the problem and what region of the data the finding related to: gender is a dimension, age is a dimension and product preference is a dimension. The supporting statement gives the statistics behind the statement and even alludes to another dimension, the time period.

The format that the organization uses can vary, but the components above should be used. The finding states clearly what population is being discussed. Depending on the audience, it is not necessary to give specific statistical information in the finding. However, the supporting information should give the details. Notice the supporting information also clarifies why this finding is important. In this case, it is because this population is different from other women.

MOVING FORWARD

Writing in *The Atlantic*, author Jonathan Rauch (2001) presents the concept of the “New Old Economy” to explain the recent impact of Information Technology (IT) on the economy. The New Old Economy refers to the impact of information technology on old-line businesses that have existed for decades using basically the same processes but with greater efficiency thanks to improvements from IT capabilities.

The United States economy grew at an unprecedented rate in the 1990s. The economy produced a higher rate of growth of real output per worker in that decade than in the previous decade. Throughout the late 80s and into the 90s organizations operating in the “old economy” were investing in personal computers and the basic software (spreadsheets and word processing packages) to perform the old-economy tasks. At first the uses of those innovations were a convenience at best and a difficult-to-use nuisance at worse. Gradually as the organizations learned to apply spreadsheets and word processing (the new technology of the time), software and hardware technology showed gains in efficiency. Now those techniques are considered basic to these Old Economy businesses. Hence, the term New Old Economy. Eventually, the convenience of the then new, now commonplace, tools (spreadsheets and word processing) made them more viable than the old mainframe systems that cost hundreds of thousands of dollars. They are now so seamlessly integrated into the company’s old-economy businesses that they receive little attention as opposed to the attention given the Internet and e-commerce activities. As Rauch states, the impact of these basic technologies is unquantified and perhaps unquantifiable. However, it is certain that we can see it in today’s workplace.

The parallel to the growth of the use of PC’s in the 80s and 90s is today’s use of data mining software and databases. As in the past, there is an organizational learning curve (as well as individual learning curve) in applying the new technology. The organization must understand new large databases and learn how to apply what is available. That process will occur in countless gradual steps and some great leaps in data, software and techniques. A necessary first step is to remove present organizational impediments to the use of exploratory data mining so that these techniques become basic to the process. As Lyndon Johnson once advised the country, “We must change to master change.” (Johnson, 1966)

REFERENCES

All Things Considered. (2001, May 2). National Public Radio

Kriegel, R., & Brandt, D., (1996). Sacred Cows Make the Best Burgers. Warner Books

Rauch J. (2001, January). The New Old Economy: Oil, Computers, and the Reinvention of the Earth. The Atlantic Monthly. 35-50

Johnson, L. B. (1966). State of the Union message.

Copyright 2004 by Z Solutions, Inc.

For More Information contact Z Solutions at contact@zsolutions.com