Seven Questions/Considerations for Your Data Mining Efforts

John Taylor and Jonathan Lindsey

Sometimes something starts out in one direction and winds up completely different. Well, this is one of those pieces that is a product of that phenomenon. Taylor and Lindsey started out with an assignment that was supposed to result in a sort of "Gaston-Alphonse" disagreement or debate. But, it did not work. So, we put our heads together and approached the assignment differently, asking ourselves how did we get to where we are today – becoming so dependant on data mining activities – and what are the 5-10 most critical questions that we could raise concerning data mining. We wrestled with that conundrum for a while, and some background along with the seven critical questions, and responses, we thought were most relevant.

It seems that it was only 10-15 years ago that we felt that the only data we needed in order effectively to solicit donors and recruit volunteers was where they lived, where the worked, and (maybe) what they were involved in when they were as student at your institution. Gosh, that was just 10-15 years ago!

Competition has changed all that. We are not only competing with other institutions (one school's high-schooler is another school's undergraduate alumnus is another school's graduate degree-holder is another school's doctorate holder is another school's professor!), but with multitudes of other worthwhile organizations. September 11th changed many donors' outlooks toward charitable giving. In fact, we saw the first decline in giving to education the year following the attack on America. But what many people may have missed when reading this in the press is that giving to religious and disaster relief organizations actually increased.

We are also competing with life in general. There are many more two-income families today – necessary in many cases just to make ends meet. And when we are not working we are driving our children to soccer matches, ballet classes, plus 101 other activities. And, if we can find a spare moment many of us are volunteering our time – in increasing numbers – to local nonprofit organizations that serve our children, families, and communities.

The reality is that all of these endeavors are good ones, certainly in the minds and eyes of those of us engaged in them. So with all of these activities going on around us it has become imperative for the conduct of a successful fund-raising campaign to understand as much about our friends, alumni, and donors as possible so that we can identify that which may encourage them to place a higher priority on our organizations when they are considering how and where to dispense with what is seen to be a decrease in discretionary assets.

With thanks to our friends in the software industry we store vastly more data on our constituents than ever before. So much so, in fact, that many organizations are facing storage problems and are contemplating what data they should purge when just a few years ago they were asking what data they should obtain!

So, what about the data? Or, more precisely, what is data? There are many definitions:

- Factual information, especially information organized for analysis or used to reason or make decisions.

- Numerical or other information represented in a form suitable for processing by computer.
- Something given or admitted; a fact or principle granted; that upon which an inference or an argument is based; -- used chiefly in the plural.
- A collection of facts from which conclusions may be drawn; "statistical data".

Our problem, however, is that while we may gather data it's not data that we need – it is **information** we seek with which we can make informed decisions:

> "Data on its own has no meaning, only when interpreted by some kind of data processing system does it take on meaning and become information.
>
> "People or computers can find patterns in data to perceive information, and information can be used to enhance knowledge. Since knowledge is prerequisite to wisdom, we always want more data and information. But, as modern societies verge on information overload, we especially need better ways to find patterns."[1]

To paraphrase a familiar quote, the devil is in the data. We have saturated our databases with data. After all, that's what a database is for – storing data. Yet by gathering endless

[1] *The Free On-line Dictionary of Computing, © 1993-2003 Denis Howe*

amounts of data we have created for ourselves the age-old forest and tree problem. It is almost impossible for us to identify the golden nugget that might result in our realizing a major contribution. Thus, we tend to internally revert to using our databases for mass-mailings, conducting telethons, and generating fund-raising reports. All worthwhile tasks, to be sure. But the return on investment many not be that great. Where the real money is, literally and figuratively, is in mining the data that you have.

An individual at your institution can certainly do data mining. But rarely is that process thorough enough to find patterns across the database. More often than not the 1:1 data-mining activity is just that. One person evaluating one person. We call this prospect research. But effective prospect research relies, first, on the identification of a likely prospect. Unless they self-identify themselves to you, you are going to have to find them on your own. However, without a robust analytic tool, strong IT staff, and talented analysts on staff to perform these functions you are going to have to use "outsiders."

Some of these "outsiders," will come in to your school and look at the data you already have pertaining to your donors. They will conduct data-informed forecast for gift receipts and investment levels to achieve continued growth over a period of time. However, this requires looking back at a schools development performance over up to 30 years. This process absolutely works! But it works best when you know you have good data – and lots of it, and have a fairly robust budget!

So, given the problems of too much data, uncertainty regarding the quality of data you have, and limited budgets, we must turn to vendors who specialize in this arena for help. But because so many of use have very limited budgets we must take care with how we spend those precious dollars. We can ill-afford to ship off thousands of data-elements that may or may not be necessary for accurate analysis knowing that every element we require the vendor to sort through is going to cost us. So before taking this path we must evaluate the answers to 7 critical questions:

1. **Why do you need a system specific ID for each constituent unit?**

   Assuming third party manipulation of your data, when is the constituent's name required to be sent to the vendor. It depends on what is being done with the data by the vendor. But, with the more recent impact of Gramm Leach Bliley all of us are being a bit more sensitive about the data we release to third parties. A system discreet identifier other than the constituent's ID number provides you with several benefits: protection of the ID number, ease of upload of data, and quicker additions of data to that which is returned. If your system does not have a system specific identifier for each record, you may wish to explore a means of generating this kind of identifier.

2. **What data is essential for analysis?**

   The simple answer to this question is, only that data which is necessary for analysis! If you were going to take a cross-country trip and wanted to find the most direct, efficient, and quickest route, would it makes sense to review Delorme Topo Maps©, showing all of the back roads and side streets for every state in a 60-page book? While the data may be useful **at some point** (like when I don't listen to my

wife's suggestion and get us lost), all you really need at the outset is a map of the interstate system. So it is true when working with outside vendors. What your overall objective is will dictate the amount of data you will need to send and they will need to analyze. The key here is to:

- Document your objective
- Seek input from the vendor as to what **minimum** data is required by them to achieve that objective
- Confirm that you have the minimum data necessary for the analysis
- Negotiate with the vendor regarding additional data needs for achieving other objectives not originally contemplated (remember, the vendor does this for a living – they may know something you do not!) – see the next question

3. What is the optimum number of elements to ship off for analysis?

Much depends on the objectives you and the vendor has determined are appropriate to achieve the desired outcome (and within budget!). But there is more to it than that. Remember that planned cross-country trip we discussed? Would you really need to pack all 50 Topo Maps© in your car in case you got lost? Not hardly. Just as you know there are states you will not remotely get close to on your trip, so too there are many individuals you will not want to include in the data you ship off for mining – existing prospects, board members, etc. Additionally, depending on the vendor and their reputation for protection of data (or lack thereof), you may feel that sending of some sensitive information is not appropriate. And remember, the vendor is going to charge (typically) on the volume of data given and reviewed. So they have an interest in your giving them all that there is! But that may not be necessary.

Check with peer institutions to see what they have done. It may be possible to do more with less. See question #3.

4. **How can you avoid choosing the least valuable variables?**

Shakespeare said, "Know thyself and to thine own self be true." This statement could not be more to the point in the world of data mining. **You** know your institutional culture and history. **You** know the context in which the data is going through analysis. Your vendor does not. Vendors tend to have one model that works best in most situations. And that is perfectly acceptable – the same can be said for your organization. That it why it is imperative that you must help the vendor know you – and you know the vendor. A pat formula will not always work except, perhaps, in the most rudimentary of analyses. Understand the power of emotional-based responses pre-existing on your database and link those with the new information provided to you. And most importantly weight the value of various variables in order to properly contextualize the information now in hand.

5. **How do you assure appropriate return on investment with a data mining exercise?**

This is a tough question. Return on investment is difficult to conceptualize when we are discussing or planning a data mining exercise. But, it is an important question since third party provided data mining carries with it a considerable cost. One of our first suggestions is that through a careful analysis of data mining vendor products that you establish a clear understanding of the value added information that the vendor will provide. This involves knowing something about the resources that the vendor

will be using to apply against your data. It also involves a clear understanding of how you and your organization will use the data that is returned from the vendor.

For instance, if you engage data mining technology as a means of prospecting, does the research staff analyze the results from the vendor prior to the major gift officers receiving the results? Or, is the data "dumped" into your data system, and everyone has access to unanalyzed data and is responsible for levels of analysis prior to using it for call stimuli.

Second, by knowing how you intend to use the data that is returned, your vendor can tailor the product to your expectations. Sure, each vendor has a preferred format for return of data and analysis of that return. But, you are the client with a particular intent for the use of the data. I would imagine that a reputable vendor will tailor the product to your specifications.

A third consideration regarding return on investment is to plan with your vendor for a means to test the predictions. If you are using data mining as a means to identify potential candidates for major gifts or for planned giving, what are the benchmarks that you feel would be an adequate test of the results within a prescribed time period?

6. How can you avoid drawing the wrong/inappropriate conclusions?

The quality of the conclusions reached is directly proportional to the quality of the questions asked. This is a basic from academic research 101. What are the questions that you want your data to answer? Clearly identify those questions, and be sure that your vendor has these questions in writing.

For instance, if you are in higher education, do you want to know the differences in performance of Greek and non-Greek alumni? Or, the differences in the time of the first gift of those who graduated within the most recent decade and prior decades? Or, if you are another type of non-profit organization, highly dependent on annual funds, what is the difference in behavior of those whose annual gift is level for five years and those whose annual gift shows some increase over time? These are two simple data mining questions, to which other bells and whistles can be added—zip code variables, class year variables, gender variables, marital status variables. But, you gotta start with clear questions.

In your discussions with the vendor, be sure that you understand the questions that the vendor is asking of the data. This way you avoid that awful experience of unmet expectations—you thought you were going to get one product, but the vendor's questions don't provide that product.

Finally, review the results carefully, ask questions of the results, look for dissonances, and re-question. The new data that is provided is just a set of answers that need new questions!

7. What are the implications for data gathering for the rest of this decade to facilitate effective data mining?

In the two past decades most of us have spent a great deal of effort with research and data enrichment of our information/management systems. Many of us find ourselves with more information than our Major Gift Officers can assimilate, and we find ourselves working harder to create more effective profiles to facilitate philanthropy for our organizations. Data mining and constituent relations

management (CRM snuck into philanthropy via the business world) are partners in our organizations. But, we find that there are many indicators in CRM that are desired to be evaluated as variables that have not been maintained in our systems.

In the next decade through the impact on our work of data mining we may find that some of the data that we thought was useful is unnecessary, and conversely, some of the information that we never gathered and maintained is considered essential. Data mining effectively can only be done if you have recorded the information that represents the variables.

For instance, what is the most important information as a variable:

(a) that a constituent was invited to participate in an event,

(b) that a constitutent responded to the invitation, or

(c) that a constituent was actually present at an event?

Depending on your value of a, b, or c, determines the information that is gathered, recorded, maintained, and used by the organization. From this example I hope that what we see is that information needs are fluid, and may be passed through finer sieves.

Parallel to different information and finer sieves is the cost of obtaining this information, inputting it into our systems, and programming this new information for effective management output. Thus, you may need to review your data system to be sure that you can continue to use it to obtain the kinds of results that a changing environment will require.

Some refer to data mining as a science, especially when you hear terms like “predictive modeling” and “forensic analysis.” But the truth is that data mining is an art. It is the art of

knowing what is important and what is not. It is the art of negotiating with information providers for precisely what it is you wish to achieve. But most importantly it is the art of painting an accurate picture as a result of your understanding how the information you have makes the most sense in the context in which it is observed. And, you are using this information in a manner that facilitates maximum philanthropic behavior toward **your** organization!

2,777 words

John Taylor is vice president for research & data services, CASE, in Washington, DC, and is a familiar face to many APRA members, and particularly among Advancement Services leaders in higher education. Jonathan Lindsey is assistant vice president, donor and information services, Baylor University, Waco, TX. Lindsey has served APRA as a member and chairman of the Professional Ethics Committee.