White Paper

The Ten Most Common Data Mining Business Mistakes

by Jeff Deal

Vice President of Operations

June 2013

ELDER RESEARCH

DATA SCIENCE & PREDICTIVE ANALYTICS

The Ten Most Common Data Mining Business Mistakes

Table of Contents

Introduction

Data mining is one of the hottest topics in business today. Over the last several years, a host of books have focused on the use of data mining and predictive analytics to improve organizational efficiency, productivity, and profitability. Hardly a day goes by without a story about data mining appearing in the Wall Street Journal, the New York Times, USA Today, and other major periodicals.

The United States government certainly recognizes the importance of these powerful tools. At the time of this writing, the National Security Administration is in the process of spending $130 million on a 470,000-square-foot facility in San Antonio, Texas, for the purpose of mining data to combat terrorism.

At Elder Research Inc., most of our clients are making somewhat more modest investments, but their applications of data mining can be just as exciting. To cite one example among many, a fraud-detection program we developed for a high-tech Fortune 100 company saved the client $11 million in the first year and $66 million over the first five years. With payoffs like these, it's easy to understand why organizations that fail to learn about and utilize advanced analytics techniques run the risk of falling behind their competition.

The results of data mining applications can have big payoffs, but the implementation of data mining techniques can also present substantial challenges. Many companies fail to reap the benefits because they make crucial mistakes in planning and deployment.

In our experience with hundreds of clients, approximately 90 percent of projects meet their technical goals, but only about 65 percent of solutions are ultimately deployed. In other words, the biggest risks of failure in data mining are organizational, not technological. Mistakes made by management in planning and guiding a data mining engagement are more likely to cause failure than mistakes made by technical experts in processing and analyzing the data.

This paper discusses the ten businesses mistakes that frequently cause data mining projects to fall short of expectations. Awareness of these common mistakes will better equip organizational leaders to plan and guide data mining engagements to successful conclusions.

The Ten Most Common Data Mining Business Mistakes

1 Mistake: Failure to Clearly Define Objectives

It is surprising how often organizations jump into data mining without knowing exactly what they will do with the capability once they have it. Aware that data mining is an exciting frontier and eager to apply the newest technologies, they get started without clearly defining their objectives and carefully planning their execution. That's a big mistake!

Launching a data mining initiative can be a time-consuming and expensive challenge. Experts must be located and hired, software must be purchased, the appropriate data must be acquired, and key stakeholders must be brought on board. There are many opportunities to make expensive mistakes. Without a firm objective and a well-formulated plan, the effort is likely to fail.

That's what happened with one large national corporation we were called in to help. The autonomous business units of this firm were served by a centralized department that managed human resources, payroll, and other shared resources. The manager of the shared services department had decided that the organization needed to have an analytics service to support the different business units, so he hired Elder Research to do some general analysis on a data set. After we produced some results, he began to shop them to the business units.

This manager was like the proverbial boy with a hammer who goes around looking for nails to beat down. Instead of using the tool of data mining to solve specific problems, he was looking for problems for data mining to solve. No unit leader had requested data mining services or had told him about a particular problem that needed addressing.

This leader had mistakenly thought that if he developed the analytics services, other managers would simply recognize their value and come asking for them. He was surprised and disappointed to find that no one was biting. Fortunately, he learned this lesson before the company made an investment in a full-scale internal data mining service.

Data mining should be viewed as a tool to hit clearly identified nails, or "points of pain." Successful engagements address specific needs, and have the owners of those problems on board from the beginning. When the problems are solved and others within the organization get wind of these early successes, interest in data mining naturally increases. Determining in advance what is possible and what is required will help ensure the success of the data mining project.

2 Mistake: Tackling Too Much Too Fast

Many organizations attempt to get to second or third base in data mining without going by first. With an overabundance of ambition, perhaps in order to make a favorable impression or to be up to date with the latest trends, management strives to build an internal analytics service to create a transformational center of excellence that will produce a large and immediate return on investment. After assigning a smart, quantitatively minded manager to head up the endeavor, they make a substantial investment in a popular analytics software package, establish some broad goals that cross functional boundaries, and begin compiling analytics.

This rushed approach almost invariable fails. Building a transformational data mining service is a major undertaking that requires extensive resources and a vast amount of organizational energy. Without a complete large-scale investment of resources and a corresponding shift in company culture, such an initiative can overwhelm an organization, resulting in frustration and failure.

A few years back, a large pharmaceutical services company made the mistake of trying to do too much too fast. Management's goal was to use data mining to transform not only the company's own business, but potentially the entire health care industry. They formed an analytics team, invested in new software tools, and organized a full-day kickoff meeting involving more than a dozen executives from all areas of the business.

It was a good start, but the effort turned out to be much larger and more complex than anticipated. When management realized after only a few months that the project would take several more years and considerably more resources than had been budgeted, they quickly abandoned it.

A more prudent course is to begin with modest, well-defined projects that have a high probability of success. Quick wins generate goodwill and excitement that lead to greater institutional support.

When the United States Postal Service Office of the Inspector General (USPS OIG) approached our firm a few years ago, they explained that their vision was to build an organization-wide analytics service to identify fraud, improve operations, and save taxpayer dollars. The need was great, because this unit is responsible for the oversight of approximately 500,000 employees, 200,000 vehicles, and an annual operating budget of $75 billion.

But rather than trying to tackle the entire vision immediately, we jointly decided to focus initially on one relatively modest challenge that promised to generate a large return on investment. Our collaborative work achieved early successes that quickly built interest and enthusiasm within the organization.

In subsequent years, as new areas of focus have been incrementally added, the USPS OIG has become a high-profile success story within the federal government, and they are steadily building toward a complete analytics service in line with their original vision.

3 Mistake: Failure to Get the Support of the Owners of the Data

Far too often, the owners of key data within an organization are reluctant to make that data available for data mining. Out of an exaggerated sense of territorial ownership, they restrict access to this indispensible resource. Sometimes incomplete data sets or data sets without data dictionaries are provided, and no one is made available to answer questions about what the data fields mean or how the data was collected.

Some of these data owners are like vicious guard dogs. No amount of doggie-treats (cajoling words and bribes) will get them to relax. Limited access to data may be granted, but the full complement of needed data remains out of reach.

These guard dogs might be database administrators, analysts, or program executives who are afraid that analysis of the data will cast them in an unfavorable light. Or they may independent types who think they can do the analytics job themselves and resent the intrusion of others into their area.

There are many possible reasons data owners withhold data, and it only takes one to stop a data mining project dead in its tracks.

Data scientists need both timely access to data and good information about the data. They need to know how it is collected and maintained, why it is messy and/or incomplete, what each data field means, and how the data is used by the organization.

Involving all key stakeholders in a data mining project from the beginning fosters a sense of shared ownership that results in greater cooperation. When the data owners participate in the formative stages, they are in a better position to provide valuable input, and they will have a stronger desire to see the project succeed.

On the other hand, bringing data owners into the effort after key decisions have been made and the project is underway may cause them to feel diminished and be uncooperative. This is what happened with one of our clients, a moderate-size financial services firm. Almost immediately after our firm was brought in, management held a two-day kickoff event with business executives, subject-matter experts, and internal analysts.

From the very first day, the analyst who would be providing the data was openly hostile and challenged almost every idea presented by our data scientists and other members of the company's team. He was clearly not presold on the effort, so he balked at the way it was being managed. Elder Research quickly became the target for his anger and aggression.

From a technical perspective, the project was rather straightforward and could have been an easy success for everyone. But after the kickoff meeting and a couple of weeks of effort, it became apparent that the engagement was headed for failure. Management cancelled it with little notice. They gave up on the potential gains, having no appetite for an internal battle.

Had the data keeper been on board from the beginning, he and his colleagues could have shined from the quick success that was possible. Instead, thousands of dollars were spent on outside experts, and the data mining initiative went nowhere. That was a costly lesson for this particular organization.

4 Mistake: Waiting for Perfect Data

In the experience of our firm, which spans almost twenty years, the mistake of "waiting for perfect data" probably kills more projects than any other. Here's a typical scenario:

The organization starts the project well. The management team defines the goals, calculates the potential return on investment, develops a project plan, gets a budget approved, assembles the team, and launches the project. Then the trouble starts. This time the problem is not resistance to providing the data, but a desire to make sure that the data is in "good" condition.

Large organizations deal with incredibly large data sets. They sometimes have tens of millions of lines of data, and the data is often in different formats because it is derived from a variety of sources. No data set this large is going to be without problems. Some of the data will be missing, corrupted, or poorly organized.

To experienced data scientists, that's not a big problem. They expect to work with messy data, and they have tools and techniques to get around the most challenging data problems. Yet, many organizations are reluctant to start on a project until they are confident that their data is complete and well organized. As most people who work with data know, that almost never happens. Too often, the delays caused by waiting for ideal data prevent the project from getting off the ground.

This is what happened with one federal agency that called us in for a data mining engagement. In every weekly meeting management said they were "getting the data together" for us. After nine months of waiting for the perfect data, a contract modification had to be issued to extend the period of performance. Several months later, another contract extension was needed, and then another. Finally, after almost 2 ½ years, the project was completed.

Because of the organization's reluctance to release data it considered inadequate, the project took more time and cost more money than necessary. Had we been allowed to work with early versions of the data we could have completed the project in just a few months.

On a more positive note, another of our clients acknowledged from the beginning that there were missing values in their data and that some of the records were inconsistent. Nevertheless, this large, multinational corporation provided us with a sample data set. Two of our data scientists analyzed it, and within forty-five minutes we had identified segments of the data that were good enough to begin the project. The client's decision to proceed, despite the data issues, ultimately saved this organization lots of money and led to faster data mining results.

5 Mistake: Believing You Have Perfect Data

On the other extreme from organizations that think their data must be perfect are organizations that think their data is already perfect. The costs of this latter error are not usually as harmful to a project as the former, but they still can be high.

Most people who are only moderately familiar with data mining assume that the major emphasis of a predictive analytics project is on model building. In reality, our firm typically spends 65 percent to 80 percent of our time on understanding, cleansing, and preparing the data for the modeling process.

When the data preparation is thorough and well done, the modeling process goes more smoothly and produces better results. Sloppy data preparation, on the other hand, leads to poor modeling results.

No organization has perfect data. Even when a data set is relatively clean, the modelers must spend time understanding it and making sure that it is properly prepared for the modeling process. When an

organization thinks its data is perfect, it will tend to have unrealistic expectations about the time and costs required to complete an engagement.

This is what happened with one of our clients, a health care services company. When our firm presented our project plan, company leaders pushed back hard on the schedule because they felt we were planning to spend too much time on data preparation. They simply could not understand why the modeling process would be delayed by data preparation, since they were providing such clean data.

After some difficult conversations, Elder Research commenced the engagement, being careful not to rush the data-understanding and the data-preparation processes. In the course of the assignment, we found some significant problems in the data and reported them to management. For example, on some customer documents there were multiple ship dates, sometimes months apart, and on others the ship date preceded the order date. The client reluctantly acknowledged that the data needed some cleaning before the modeling process could begin, and more reasonable expectations were established.

6 Mistake: Relying Too Heavily on Software

Data mining is the use of advanced predictive algorithms to identify patterns within very large data sets. Since computers and analytical software are typically used, the choice of software becomes a key concern. Many inexperienced managers are inclined to buy the most popular and sophisticated software packages, and these packages are often what software vendors want to sell. But for some data mining applications, less expensive programs or even free shareware will work. In some instances, such as when outside consultants perform the analytics, no software purchases may be necessary.

Many companies mistakenly believe that if they make an investment in the right software, the data will almost model itself. So they go out and purchase a software package, only to find out they lack the internal expertise to use it.

Analytic software is only a tool in the complete data mining process. The expert application of software is just as important as the choice of software.

One of our clients, a Fortune 500 company that was doing basic analytics with spreadsheets and other metric-based tools, asked our firm to help them learn and apply the analytic software they had purchased. When our consultant arrived on site, he found that the data mining software had been purchased more than a year earlier and had never even been installed. In fact, an extensive search was needed to even find it!

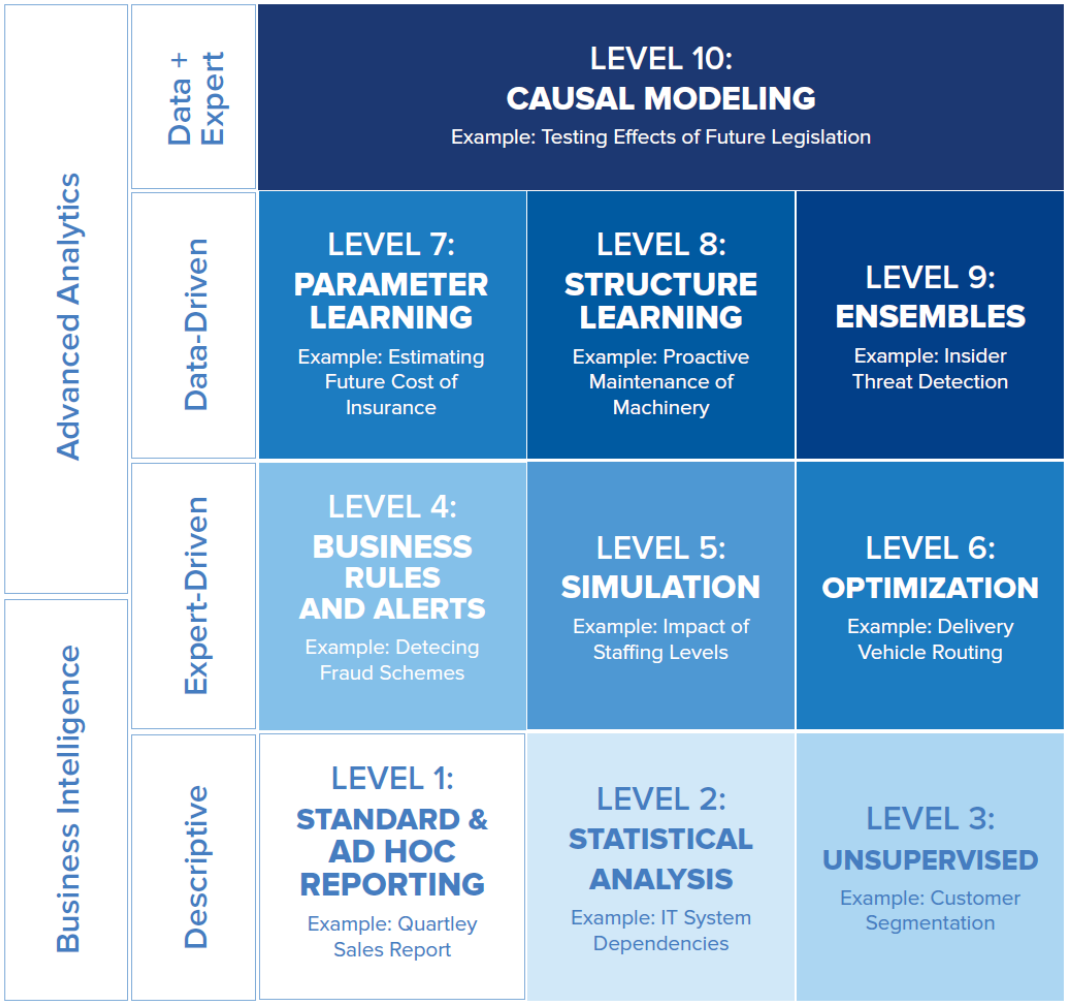
The company's management had erroneously thought that a software purchase was all that was needed to get started in predictive analytics. That's like thinking the purchase of an airplane is all that is needed to become a pilot. Our consultant helped install the software and began training company analysts to use it.

Some companies not only rush to buy, they overbuy. Organizations just getting started in data mining rarely need the most advanced software tools available. Until they master the essential concepts of data mining and identify specific needs that require more advanced tools, a basic software solution will almost always suffice. By way of analogy, if you need only a basic starter car for your teenage to get to and from high school, you won't buy a Maserati Quattroporte S!

7 Mistake: Failure to Understand the Different Levels of Analytics

In today's highly competitive environment, virtually every innovative organization knows the importance of using analytics to improve productivity, reduce fraud and waste, and increase profitability. Many excellent books have been written on the subject, such as Competing on Analytics by Tom Davenport and Jeanne Harris. However, few authors discuss the exact nature of analytics and its various levels.

At Elder Research, we think of analytics as having the following ten levels:



There is considerable variation in purpose and application among these techniques. Software tools that do an excellent job on business intelligence may not work well for complex network analysis. Text mining is substantially different from data mining. Each technique has an appropriate application, and for many organizations some techniques may not be useful at all. Expertise and tools need to be carefully matched with the unique needs of the organization.

Managers and executives should make the necessary investment to gain an understanding of the different levels of analytics, before they begin developing an analytics capability within an organization. Armed with this knowledge, they will be able to make better decisions about the setting of goals, the hiring of new people, the purchase of analytic software, and the engagement of outside consultants.

When the United States Postal Service Office of the Inspector General (USPS OIG) decided to begin applying more advanced analytics in support of their mission, they were careful to start small with a contract fraud problem that required a metrics-based approach (level 1). After that effort proved successful, they added new challenges that required different techniques.

The USPS OIG is now using predictive modeling, optimization and simulation, and text mining to address a variety of issues. Its leaders wisely sought out industry expertise to help them apply the appropriate techniques to specific problems with big payoffs. As a result, their Countermeasures Directorate is now recognized as an analytic center of excellence within the Office of Inspector General community, which is comprised of more than sixty different federal agencies.

8 Mistake: Excluding Domain Subject-Matter Experts

In our discussion of mistake #6, we talked about how some organizations place too much reliance on analytic software and fail to involve data mining experts, or data scientists, who understand analytic techniques and tools and know how to apply them in creative ways. But it's also a mistake to put too much reliance on data mining experts and not enough on domain subject-matter experts.

Domain subject-matter experts are an essential component of a successful data mining engagement. They provide the business understanding that the data scientists need, and they provide a common-sense check to the modeling process.

Data mining is an iterative, trial-and-error process that requires working at the problem from multiple angles over a period of time. It is never as simple as plugging the data into a black box, turning a crank, and producing results. The subject-matter expert serves as a check in that iterative process to help the data scientists stay on track. As preliminary results are produced, the subject-matter expert can help to confirm the validity of the findings and identify outcomes that are so far outside of the norm as to require additional review and confirmation.

Furthermore, it is the subject-matter experts who will ultimately use the results of the modeling process, so their buy-in is essential. Involving subject-matter experts throughout the modeling

process contributes to a better understanding of the business problem, a more complete and accurate modeling process, and a more successful application of the final results.

We recently encountered the mistake of excluding subject-matter experts at a wireless phone company, which had called us in to help build models to improve their direct marketing efforts for pre-pay and post-pay phone sales. Prior to our arrival, consultants from a highly regarded software and services organization had built the types of models the wireless phone company desired, but those models were not working. There was no demonstrable improvement in any of the target areas.

We started our engagement with extensive interviews with the subject-matter experts in sales, marketing, and data analysis. As the client project manager observed the Elder Research consultants at work, he questioned why so much time was being spent on talking with the domain experts, rather than just starting the modeling process as the previous consultants had done.

Immediately, the reasons why the previous modeling effort failed were evident. Those consultants had made mistakes #6 and #8, mentioned above. Not only had they not taken time to understand and properly prepare the data, they had failed to work with the subject-matter experts to fully understand the business problem they were hired to solve. Through diligent work with the experts to fully understand the data and the business problem, the Elder Research consultants built successful models that produced a clear return on investment for the company.

9 Mistake: Failure to Plan for Deployment

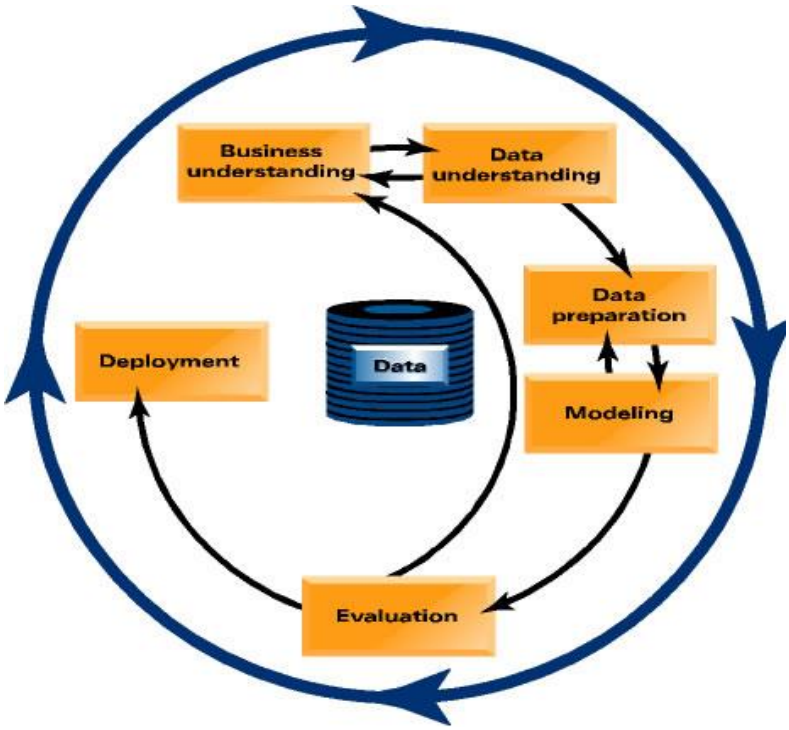
Many business leaders fail to appreciate that data mining and model building are only a start, and that predictive models must be deployed in the client organization to see the full return on investment. All too often, the issue of deployment is an afterthought that doesn't come up until schedules are established, budgets are set, and promises of model results are made.
Building strong models that produce solid results can be a major endeavor for any organization. Putting those same models into production so results are available on an ongoing basis can require at least as much effort, and often more. Prior to embarking on a predictive analytics engagement, the client needs to work with the data mining experts to decide on the deliverable format and how it will be accomplished. At a minimum, the following questions need to be answered:

- How will the models run?
- Is new software needed?
- What are the run times?
- How often are the runs performed?

Failure to think through the complete process from model building to deployment can cause huge losses in money, time, and credibility. It's akin to building a majestic staircase that ends at the top stepping into open air. The view may be terrific from the top, but there's a precipitous drop-off between that step and the final destination. In data mining, a planned deployment completes the staircase and leads to a new level of accomplishment, with a high return on investment.

10 Mistake: Rushing the Process

For best results, data mining efforts need to follow established, methodical procedures. One leading process, the Cross Industry Standard Process for Data Mining (CRISP-DM) depicted below, describes the progression of steps used by data miners to address problems.



To outsiders, some of the steps in CRISP-DM may appear to be unnecessary. It's natural for organizational leaders to feel that "jumping through all these hoops" is a waste of time and money. But when data mining projects are rushed and steps are skipped, the results are usually unsatisfactory. In the long run, time and money are saved when data scientists are provided with the necessary resources, including time, to do their job well.

In the case of the wireless phone company discussed in mistake #8 above, the initial consultants engaged by the company skipped or glossed over the first three steps of the CRISP-DM process: business understanding, data understanding, and data preparation. Initially, the client was pleased with how quickly the project progressed. Ultimately, however, the models performed poorly, wasting considerable time, effort, and money.

When we began work on this project for the phone company, after the initial consulting company was fired, the client expressed concern about how much time we were spending on understanding the business and understanding and preparing the data. After discussion however, management allowed us to proceed methodically. When the project was completed and the models produced a healthy return on investment, they saw the wisdom of our approach.

Summary

Data mining is a powerful tool for increasing organizational efficiency, productivity, and profitability. When a data mining engagement is properly planned and executed, it is exciting to watch the components come together and opportunities for improvement revealed. On the other hand, a poorly executed data mining project can waste considerable time and money, with attendant frustration and loss of credibility.

The ten mistakes highlighted here commonly cause data mining endeavors to fall short of expectations. Awareness of these potential pitfalls will give organizational leaders a good start toward reaping the best results from data mining.

About the Author

Vice President of Operations Jeff Deal leads tasks involving contracting, finances, logistics, planning, and regulatory/legal issues. In his five years at Elder Research, he helped to lead the company as it tripled in size and substantially grew its second office in Arlington, Virginia. Jeff has worked with dozens of clients to understand their business needs and organizational goals and, in the process, has gained insight into organizational obstacles to successful data mining engagements. His talk on the Top 10 Data Mining Business Mistakes has been well received at analytic conferences. Mr. Deal is the Program Chair for the inaugural Predictive Analytics World – Healthcare conference scheduled for October 2014 in Boston. Jeff has more than 25 years of experience in business operations, planning, and government relations, primarily in the health care industry. Prior to Elder Research, he was the President of a health planning consulting business that assisted hospitals and physicians with operational analysis, forecasting, and navigating through complex regulatory processes. Jeff has a Master of Health Administration degree from Virginia Commonwealth University and an undergraduate degree from the College of William and Mary where he was a member of the wrestling team.

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Maryland Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673