



Data Mining with Qualitative and Quantitative Data

John F. Elder IV, Ph.D.
CEO, Elder Research, IIA Faculty

SEPTEMBER, 2010

www.iianalytics.com



John F. Elder IV, PhD
Elder Research, IIA Faculty

Dr. John Elder heads a data mining consulting team with offices in Charlottesville Virginia, Washington DC, Mountain View California, and Manhasset New York (www.datamininglab.com). Founded in 1995, Elder Research, Inc. focuses on investment, commercial and security applications of advanced analytics, including text mining, stock selection, image recognition, process optimization, cross-selling, biometrics, drug efficacy, credit scoring, market timing, and fraud detection.

John obtained a BS and MEE in Electrical Engineering from Rice University, and a PhD in Systems Engineering from the University of Virginia, where he's an adjunct professor teaching Optimization or Data Mining. Prior to 15 years at ERI, he spent 5 years in aerospace defense consulting, 4 heading research at an investment management firm, and 2 in Rice University's Computational & Applied Mathematics department.

Dr. Elder has authored innovative data mining tools, is a frequent keynote speaker, and was co-chair of the 2009 Knowledge Discovery and Data Mining conference, in Paris. John's courses on analysis techniques—taught at dozens of universities, companies, and government labs—are noted for their clarity and effectiveness. Dr. Elder was honored to serve for 5 years on a panel appointed by the President to guide technology for National Security. His book with Bob Nisbet and Gary Miner, *Handbook of Statistical Analysis & Data Mining Applications*, won the PROSE award for Mathematics in 2009. His book with Giovanni Seni, *Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions*, was published in February 2010.

THE BIG IDEAS

- Data mining is past the hype stage. It can produce measureable and significant bottom-line results. (One case study shows how data mining resulted in a \$1 billion decision with positive results.)
- Data mining can solve complex “needle in haystack” problems.
- Data mining can eliminate bad (like fraud), highlight good (such as new opportunities), and streamline decision processes.
- Data mining can now be used for qualitative data, as in text mining (see case study below).
- Data miners should work with business people on projects that produce benefits for the business.
- Start small, produce early wins, build momentum, and create advocates. This will lead to bigger projects longer term.

OVERVIEW

Data mining helps organizations solve complex “needle in haystack” problems. These problems involve using data mining to eliminate bad (like fraud detection), highlight good (as in identifying new opportunities), and streamline or semi-automate decisions (as in automated scoring processes). When done effectively, data mining can have a quick ROI and a significant bottom-line impact.

Data miners should work with business people to ensure that data mining projects have meaningful business success and that quick wins are produced to build momentum and support for data mining.

CONTEXT

Using a series of case studies, John Elder showed the value of data mining in helping organizations solve complex business problems.

INSIGHTS ON DATA MINING

Effective data mining can have significant bottom-line benefits.

Data mining doesn't necessarily lead to breakthroughs, but it can incrementally boost an organization's performance and improve an organization's competitive position. Using data mining to score risk just a little bit better, or gain a bit more information about which products consumers are interested in, or identify and prevent a small amount of fraud can in many instances result in significant bottom-line benefits.

Data mining is past the hype; the benefits are real.

Any new technology creates inflated expectations and suffers from a period of disillusionment. But for data mining, the period of hype and disillusionment was short-lived. The reason: unlike many new technologies that provide ambiguous benefits, data mining is tied to real, measurable results, liked improved customer retention.

Data mining helps solve hard “needle in haystack” problems.

In general, data mining helps organizations solve difficult “needle in haystack” problems:

- **Eliminating bad.** This includes using data mining to develop systems and scores that identify bad things like fraud and credit risk.
- **Highlighting good.** Data mining can be used to identify new opportunities and help in complex areas like drug discovery.
- **Streamlining or semi-automating decisions.** For example, data mining can identify products that might be of interest to a particular consumer and can be used to speed up the service delivered to clients.

One example is a system the IRS developed (with the assistance of Elder Research) to detect fraud related to a type of tax credit. This system involved creating a “scoring system” based on past occurrences of known fraud. Hewlett Packard also used data mining to create a fraud detection system to find fraudulent payments to service agents. In just nine months this system recovered \$20 million for the company, which went directly to HP’s bottom line.

Analysts are often concerned that data mining systems will replace people. In Dr. Elder’s experience, this is not the case. These systems make people much more effective and efficient. At the IRS, the fraud detection system resulted in an order of magnitude improvement in the amount of fraud that was found per day per analyst. Such systems often have very high ROI and the groups that manage these systems become profit centers. The group grows as does its mandate.

“Finding fraud becomes a profit center with good ROI. When you discover and retrieve fraud, it all flows to the bottom line.”

- John Elder

Data mining can now be done on qualitative information like text.

An example of how data mining can streamline and semi-automate decision making can be seen in the work that Dr. Elder and his firm did on a proof of concept for the Social Security Administration (SSA) in creating a text mining capability.

Poor disabled people can apply to the SSA to receive Social Security benefits. The slow, bureaucratic application process includes a text field where individuals must describe their disability. Previously, this text field had to be reviewed by an individual at the SSA. Analysis showed that one-third of those who applied for disability benefits eventually received them, but the process often took years.

The SSA wanted to create a fast-track process to immediately approve straightforward applications. This would allow individuals at the SSA to focus their time on the more complex cases. The barrier to the fast-track process was being able to analyze the free text field where an individual explained their disability.

Using a variety of techniques, a text mining solution was piloted for the SSA. The techniques used to make text mining a reality include:

- **A term matrix.** Using a matrix, distinct terms are translated into useful information.
- **Stemming.** Many words have the same stem (like cancer or cancerous). Text mining technology can recognize similar stems.
- **Word location.** The location of words can provide context about the word.
- **Links between concepts.** For examples, ALS and Lou Gehrig’s disease are different words but a similar concept.

These techniques enabled automated mining of the text field in the SSA application to determine the reason for the disability request. As a result of this text mining capability, in the pilot, 20% of the applications approved by the SSA were approved immediately.

Data mining can be used to help companies make critical decisions.

Dr. Elder shared a case study of working with a major pharmaceutical company that needed to decide whether or not to invest an additional \$1 billion in further developing a drug in its pipeline.

The company believed it had a unique drug, but the traditional one-dimensional data from trials didn't show a significant benefit versus placebos. Using data mining techniques, Dr. Elder and his team were able to create a three-dimensional visualization of the data showing how the health of patients receiving the drug in trial improved compared to those receiving the placebo. This 3D visualization—which was a different way of looking at the data—showed the company that its drug did have unique benefits. The company decided to invest the \$1 billion, brought the drug to market, and it has been a major success.

Creative data mining can be used to look at existing data in helpful new ways.

Through his experience, Dr. Elder has learned valuable lessons about data mining which can be broadly reapplied.

These lessons include.

- **Focus on both technical success and business success.** At times, data miners focus on the technical success of creating complex models. But what matters is that a data mining initiative is tied to business success. Data miners can't lose sight of this.
- **Make a data mining project easy to measure.** A data mining project will be a success if everyone can look at the results and see a bottom-line improvement. Make the measures simple, clear, and related to bottom-line business results.
- **Make a data mining project leveragable.** Be able to translate the results into a bottom-line benefit. For example, a company might develop a system that detects just 0.1% more fraud. However, if that slight increase in fraud detection represents millions of dollars to the bottom line, it is a worthwhile project.
- **Engage in projects that show immediate value.** Any data mining project requires the support of a business champion who is willing to make an investment and take some risk. Show immediate value to build momentum for data mining. (This might mean beginning with "low-hanging fruit" projects.)
- **Involve both data miners and business people.** Solving complex problems requires both data miners and business experts, working together. Business people understand the business and the context. Their cooperation is essential.
- **Be vigilant about data quality.** If you get the data right, almost any data mining technique can work (and an ensemble of data mining techniques works best).
- **Communicate relentlessly.** Data miners often like to work in isolation on complex problems, surfacing only after they have developed a solution. This is not the best approach. There needs to be frequent, ongoing, iterative communication throughout a project. As was done in this session, use stories to highlight successes.

IIA MEMBERSHIP

Analytics practitioners can become members of IIA for \$195 per year. Membership includes:

- **Monthly Briefings** which provide access to Professor Davenport, other IIA faculty, and leading analytics practitioners who will share best practices and new research.
- **Online Repository** with proprietary analytics research, archives of briefings, and key insights from IIA events.
- **Ask an Expert Portal** to ask pressing analytics questions and get quick answers from IIA's expert faculty members.



M2010
DATA MINING CONFERENCE

October 25-26 | Caesars Palace | Las Vegas
www.sas.com/m2010

Listen to more than 40 data mining experts share the latest in the field of analytics.

Presented by

sas | **THE POWER TO KNOW.**

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2010, SAS Institute Inc. All rights reserved. S56928US.0510



PREMIER UNDERWRITER: SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions delivered within an integrated framework, SAS helps customers at more than 45,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®. Read the [latest news](#) about SAS and subscribe to [RSS feeds](#) and [blogs](#).