Published in DM Review in January 2004.
Printed from DMReview.com

Data Warehousing Lessons Learned: Data Mining is Dead – Long Live Predictive Analytics!

by Lou Agosta

This is why data mining is dead: it died of a broken heart. It was killed by disappointed expectations. In addition to a perfect storm of tough economic times, another reason data mining technology has not lived up to its promise is that "data mining" is a vague and ambiguous term. It overlaps with data profiling, data warehousing, and even such approaches to data analysis as online analytical processing (OLAP) and enterprise analytic applications. When high profile successes have occurred (e.g., a front-page article in the Wall Street Journal, "Lucky Numbers: Casino Chain Mines Data on Its Gamblers, And Strikes Pay Dirt" by Christina Binkley, May 4, 2000), they have been a mixed blessing. Such results have attracted a variety of imitators with claims, solutions and products that ultimately fall short of the promises. The promises build on the mining metaphor and typically are made to sound like easy money. This has resulted in all the usual dilemmas of confused messages from vendors, hyperbole in the press and disappointed end-user enterprises.

Data mining is regrouping as "predictive analytics." The differentiators are summarized in Figure 1.

Data Warehousing	Classic Data Mining	Predictive Analytics
Query and reporting functions (SQL)	Statistical analysis	Prescriptive algorithms
Static perspective	Continuous changes	Also discontinuous changes
Describe the present and past	Predict the past	Predict the future
Assume hypothesis	Validate hypothesis	Invent and validate hypothesis

Figure 1: Data Mining and Predictive Analytic Differentiators

- **Prescriptive, not merely descriptive:** Scanning through a terabyte haystack of billing data for a few needles of billing errors is properly described as data mining. However, it is descriptive, not prescriptive. When a model is able to predict errors based on a correlation of variables ("root cause analysis"), then the analysis is able to recommend what one ought to do about the problem (and is, therefore, prescriptive). Note that the model expresses a "correlation," not a "causation," though a cause-and-effect relation can often be inferred. For example, Xerox uses Oracle's Data Mining software, for clustering defects and building predictive models, to analyze usage profile history, maintenance data and representation of knowledge from field engineers to predict photocopy component failure. The copier then sends an e-mail to the repair staff to schedule maintenance prior to the breakdown.
- **Stop predicting the past; predict the future:** Market trend analysis as performed in data warehousing, OLAP and analytic applications often asks what customers are buying or using (product or service), and then draws a straight line from the past into the future, extrapolating a trend. This too can be described as data mining. One might argue this predicts the future because it says something about what will happen. However, a more accurate description would be that it "predicts the past" and then projects that into the future. The prediction is not really in the analysis. Furthermore, data mining in the limited sense used here is only able to envision continuous change – extending the trend from past to future. Predictive analytics is also able to generate scores from models that envision discontinuous changes – not only peaks and valleys, but cliffs and crevasses. This is especially the case with "black box"-type functions such as neural networks and genetic

programming. Rarely do applications in OLAP, query and reporting or data warehousing explicitly relate independent and dependent variables, but that is of the essence in predictive analytics. For example, KXEN is used to find the optimal point between savings of catching a bad customer versus the cost of turning away a good paying customer (opportunity cost).

- **Invent hypotheses, don't merely test them:** Finally, data mining is distinguished from predictive analytics in terms of hypothesis formulation and validation. For example, one hypothesis is that people default on loans due to high debt. Once the analyst formulates this hypothesis by means of imaginative invention out of her or his own mind, the OLAP analyst then launches queries against the data cube to confirm or invalidate this hypothesis. Predictive analytics is different in that it can look for patterns in the data that are useful in formulating hypotheses. The analyst might not have thought that age was a determinant of risk, but a pattern in the data indicates that as a useful hypothesis for further investigation.

One reason that data alone is not knowledge but merely data is that it lacks structure, organization, direction, coherence, point and conceptual focus. Just as a predictive model without supporting data would be empty, likewise data without a unifying model is meaningless and leaves the collector blind. Giga clients will need to have expertise in all three dimensions: details of the business, data collection and model building. Client predictive efforts should be guided by the methodological injunction that determining meaning is a business task, not a statistical one. Within such a context, the selection of a tool for predictive analytics can be leveraged to the advantage of customer recommendations, cross- selling, up-selling, personalization, loyalty development, attrition and churn (reduction), forecasting, demand planning, inventory (and cost) reduction, brand development and the mastery of market dynamics.

Lou Agosta, Ph.D., is a business intelligence strategist with IBM WorldWide Business Intelligence Solutions. He is a former industry analyst with Giga Information Group and has served many years in the trenches as a database administrator. His book ***The Essential Guide to Data Warehousing*** *is published by Prentice Hall. Please send comments and questions to Lou in care of LAgosta@acm.org.*

Illinois Police To Gain Sophisticated Data-Mining Software

Posted: November 30th, 2005 02:16 PM EDT

News-Gazette, The (Champaign-Urbana, IL) (KRT)
News-Gazette via NewsEdge Corporation

Nov. 27--CHAMPAIGN -- RiverGlass Inc. has landed its first million-dollar contract, furnishing the Illinois State Police with sophisticated data-mining software that can help analyze crime data.

The sale, completed in late summer, is being implemented now, and the final product should be in place next spring, said Kirk Dauksavage, the company's chief executive officer.

RiverGlass is based on technology developed by the company's founder, Michael Welge, and its development offices are based in Champaign. About 12 people are employed at its offices in EnterpriseWorks in the University of Illinois Research Park, and five or six others work at the company's headquarters in West Chicago.

"We provide tools that help analyze data in real time," Dauksavage said, adding that state police have about 80 analysts who sift through crime data. Their effectiveness and productivity will be enhanced with the new software called RiverGlass Recon.

The system was incorporated in the state police Statewide Terrorism and Intelligence Center, said Master Sgt. Rick Hector, a spokesman for the state police. That center is a part of the new State Emergency Operations Center that opened in Springfield in October.

RiverGlass is also negotiating a contract with a large metropolitan area and hopes to have that finalized by the end of the year, Dauksavage said.

"Revenuewise, we'll beat our plans for this year, and next year we expect to be five times where we are this year," he said, declining to release specific figures.

Data-mining technology can be used in several different markets, he added.

"Clearly law enforcement is interested in it, and there are lots of opportunities at both the state and federal levels," he said.

The technology can be used in maritime security at ports and in identifying fraud for insurance and financial services, he said.

Welge, who works at the National Center for Supercomputing Applications at the UI, said both industry and government can gain insights from streaming real-time data.

He predicted changes in the use and analysis of real-time information with the advent of sensor networks and radio frequency identification, or RFID, technologies.

"At RiverGlass, we are seeing RFID applications develop in personal identification military and civilians livestock management, food security, supply chain management and maritime cargo security," he said.

So far, RiverGlass has received equity investments from IllinoisVentures, the Illinois Finance Authority and Waypoint Ventures in Ann Arbor, Mich., as well as from several angel investors, Dauksavage said.

Rob Schultz, senior director of IllinoisVentures, said he was "thrilled and extremely excited" by RiverGlass' contract with the state police.

"It's always positive for an early-stage company to find a customer of this stature to help develop a product," Schultz said.

"The Illinois State Police is considered to be one of the big technology users in the country among state-level law enforcement agencies," he said. "They're also a tremendous influence on adoption of technology across other states."

<<News-Gazette, The (Champaign-Urbana, IL) (KRT) -- 11/30/05>>

Chicago Tribune
ONLINE EDITION

http://www.chicagotribune.com/features/lifestyle/chi-0511290268nov29,1,3921830.story?coll=chi-leisuretempo-hed

The cyber sleuth

DePaul computer scientist develops system to help Chicago Police Department solve serial crimes

By Patrice M. Jones
Tribune staff reporter

November 29, 2005

Tom Muscarello has the gravely voice and dry wit that conjures up an image of a hardened police detective.

And he certainly seems to spend a great deal of his time contemplating the morbid intricacies of the criminal mind."Most criminals are creatures of habit," Muscarello said recently, shuffling through data by light of his computer screen. "Typically, a thief does not just rob once. He will rob a different person every week of the year without fail. He will hit old ladies every Social Security day. It is his job; his occupation."

Muscarello, the man who can reel off a criminal profile like a pro, himself has an unusual occupation: He is, in fact, a cyber sleuth.

A veteran DePaul University computer scientist, Muscarello has been working since the mid-1990s on perfecting an artificial intelligence system that is aimed at helping the Chicago Police Department blaze a bold new trail in the way it solves serial robberies, rapes and other violent crimes.

And he just might have hit pay dirt.

The computer system, called the Classification System for Serial Criminal Patterns (CSSCP), is expected to begin live trials at the Chicago Police Department as soon as early next year.

Developed by Muscarello with DePaul researcher Kamal Dahbur, CSSCP uses pattern-recognition software that acts like a superhuman brain.

The computer system will be able to cull massive amounts of data, pulling out details of individual crimes, such as the assailant's age, sex, height, location of the crime, weapons and vehicle used, to create a criminal profile that can be compared with others.

The goal is to help overcome a thorny problem in police work -- the fact that detectives can have difficulty linking serial cases.

The system also will have the potential to search through words or phrases in police reports, such as a criminal who wears "green military fatigues" or one who says "give it up" every time he robs a bank.

It can work 24 hours a day without human intervention, sorting through thousands of criminal records per second, revealing patterns in seemingly unrelated crimes that a mere mortal could miss.

"It could revolutionize the way [the Chicago Police Department] does police work," said Charles Padgurskis, former director of information systems for the Chicago police, who has worked with Muscarello since the research project was launched a decade ago. Padgurskis retired from the CPD last May.

"[The system can tell us], for example, these six cases have the same characteristics so that we can attribute them to one offender," added Steve Maris, the Chicago police's current representative for the project and the acting assistant director for information services.

Cracking serial crimes has long been especially tough for law enforcement since the crimes can occur over a long stretch of time, a widely dispersed area (involving different police districts within a city or separate police departments) and may even have differing criminal patterns in each incident.

"We decided to try to build a system that is intelligent enough to do what the best detectives are already doing," said Muscarello, who explored the techniques of six top Chicago detectives during the initial stages of the study. The cyber sleuth DePaul computer scientist develops system to help Chicago Police Department solve serial crimes A veteran DePaul University computer scientist, Muscarello has been working since the mid-1990s on perfecting an artificial intelligence system that is aimed at helping the Chicago Police Department blaze a bold new trail in the way it solves serial robberies, rapes and other violent crimes. Computer system may help police solve serial crimes or DePaul computer scientist joins forces with Chicago police or DePaul researcher may help Chicago police crack serial crimes

In a recent study using three years of Chicago police robbery data (not active cases), the CSSCP system -- which uses a computer network particularly suited for this type of inquiry called a Kohonen neural network -- detected at least 10 times as many related crimes as a team of detectives with access to the same data.

Muscarello, who started his career as a federal investigator of Medicare fraud, has garnered national attention since the research project was published last year. He said he has gotten calls from law-enforcement representatives across the globe, even from as far away as Australia.

Still Muscarello cautions that the new system could enhance, not replace, solid criminal investigations.

Discovering a criminal's internal script -- whether it is a rapist who usually strikes young women at night on the Near West Side or a bank robber who uses a trademark phrase -- is the genius of the best detective work, Muscarello said.

"We decided to try to build a system that is intelligent enough to do what the best detectives are already doing," said Muscarello, who explored the techniques of six top Chicago detectives during the initial stages of the study.

Other computerized crime-analysis systems exist, but the CSSCP system is unique because it can search for patterns without the help of a computer operator or programmer.

The system can give alerts when patterns of criminal activity emerge and can potentially alert law

enforcement to begin an investigation of serial crimes long before they are detected in the normal course of an investigation.

"Nobody today in 2005 has come up with a program that has done what this network can accomplish," Padgurskis said.

Muscarello, 55, who grew up in a working-class family in Rogers Park, says using technology to solve problems has long been a major focus in his career.

He graduated from the University of Illinois at Chicago in 1971 with a degree in biology. And after wading his way through various jobs -- including a stint as a federal investigator -- he got a master's degree in computer science at DePaul.

Later, Muscarello earned a doctorate in computer science and electrical engineering at UIC.

David Miller, acting director of DePaul's school of computer science, telecommunications and information systems, said: "This is all interesting since I would not say that [Tom] is a well-known researcher. But what Tom is really good at is making connections with people in industry and solving problems for them."

"I don't do books," Muscarello said, somewhat proudly. "I have done chapters in books. But the people I know who have done books, they take years doing it and then the book comes out and already in the computer science world, the research is old. My focus has always been looking at ways to use technology to solve problems."

With his time split between research, teaching and various administrative roles, Muscarello says he has little time for hobbies. But in a strange twist of fate, the man who now focuses on finding better investigative tools for law enforcement once worked for a short time as a mechanic in the Chicago Police Department garage. It was a time when Muscarello was finding himself, he said.

"Isn't that strange?" Muscarello said. "From fixing cars to this."

- - -

Tom Muscarello

Age: 55

Raised: West Rogers Park

Education: bachelor's degree, biology, University of Illinois at Chicago, 1971; master's, computer science, DePaul University, 1985; doctorate, computer science, UIC, 1993

Career highlights: federal investigator, specialty, Medicare fraud, 1977-83; consultant, information systems, 1983-1992; currently associate professor, along with various administrative roles, DePaul University.

Family: Wife, Bernadette.

---

pjones@tribune.com

Hard data comes from observations and measurements of the macroscopic natural world: the positions of astronomical objects, the electrical impulses within the brain, or even the amounts of your credit card transactions. Typically, such data is objective. The observations are numerical, and the uncertainty is adequately characterized as an error zone around a central value.Â Thus, the better we can understand the â€œsoftâ€ variety of big data, the better we can do in terms of predictive (and eventually prescriptive) analysis based on such data. I am by no means suggesting that we fall back to the Pythia or to auspices ex avibus to predict the future. But we must recognize that our current state of mathematical modeling is not foolproof, and is still much more art than science.