Thearling.com

Information about data mining and analytic technologies

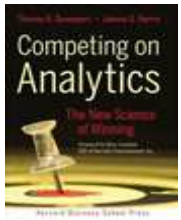
Kurt Thearling / kurt@thearling.com

Data Mining Books | White Papers | Data Mining Tutorial

Data mining, if you haven't heard of it before, is *the automated extraction of hidden predictive information from databases*. I have spent the past fifteen years building commercial data mining and data analysis systems, solving problems across fields such as financial services, the life sciences, insurance, and telecommunications. I currently lead the analytics organization at Vertex Data Science, a multi-national business process outsourcer. (Click here for more about my background.)

The purpose of this web site is to share information about data mining and other analytic technologies. I hope you find it useful.

My **Data Mining Book of the Month** is *Competing on Analytics* by Tom Davenport and Jeanne Harris. This book is a great addition to the literature of analytics. Not that it solves any complicated statistical modeling problem; I don't think that there is a single equation in the entire book. Instead, this book focuses on the (more important) problem of getting an organization to change its approach to problem solving, by increasing the use of analytics across a business. This is a trend that I have been involved with over the past fifteen years, and I think that it is a key differentiator between modern companies. If you are interested in learning how companies like Amazon, Capital One, Harrah's, and Netflix (to name just a few) put analytics into practice, this book is a fantastic resource. For more books on data mining, take a look at my list of recommended books.



New

A while back I added online versions of the slides from a couple talks I have given on data mining and analytic technologies:

- More Than Algorithms: Data Mining in the Real World focuses on the practical issues that need to be addressed as part of any data mining implementation.
- Understanding Your Customers Using CRM Technology provides an overview of technologies such as data mining, personalization, and content management. The focus of the presentation is that these technologies can be part of a process which allows marketers to tune interactive relationships to each customer's needs.

Some time ago I took part in a National Public Radio report about data mining and privacy with a local NPR affiliate. It's a bit sensationalistic, but I think that reflects the concerns that the general public has about privacy in the age of the internet.

What is data mining good for?

Data mining software allows users to analyze large databases to solve business decision problems. Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine

learning twists thrown in. Like statistics, data mining is not a business solution, it is just a technology. For example, consider a catalog retailer who needs to decide who should receive information about a new product. The information operated on by the data mining process is contained in a historical database of previous interactions with customers and the features associated with the customers, such as age, zip code, their responses. The data mining software would use this historical information to build a model of customer behavior that could be used to predict which customers would be likely to respond to the new product. By using this information a marketing manager can select only the customers who are most likely to respond. The operational business software can then feed the results of the decision to the appropriate touch point systems (call centers, direct mail, web servers, email systems, etc.) so that the right customers receive the right offers.

White Papers and Other Publications:

Over the past few years I have written or co-written a number of white papers and other publications looking at data mining and decision support technologies. Most of them are available below.

- **An Introduction to Data Mining: Discovering Hidden Value in your Data Warehouse.** *This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.*

- **An Overview of Data Mining Techniques.** *This overview provides a description of some of the most common data mining algorithms in use today. We have broken the discussion into two sections, each with a specific theme: 1) Classical Techniques such as statistics, neighborhoods and clustering, and 2) Next Generation Techniques such as trees, networks and rules. Each section will describe a number of data mining algorithms at a high level, focusing on the "big picture" so that the reader will be able to understand how each algorithm fits into the landscape of data mining techniques.*

- If you are confused about some of the terminology related to data mining, check out the **Glossary of Data Mining**.

- A while back I wrote a couple general overviews of campaign management and data warehousing for hr.com.

- **Data Mining and Customer Relationships.** *Most marketers understand the value of collecting customer data, but also realize the challenges of leveraging this knowledge to create intelligent, proactive pathways back to the customer. Data mining - technologies and techniques for recognizing and tracking patterns within data - helps businesses sift through layers of seemingly unrelated data for meaningful relationships, where they can anticipate, rather than simply react to, customer needs.*

- **Data Mining and Privacy: A Conflict in the Making?** *Privacy. It's a loaded issue. In recent years privacy concerns have taken on a more significant role in American society as merchants, insurance companies, and government agencies amass warehouses containing personal data. The concerns that people have over the collection of this data will naturally extend to any analytic capabilities applied to the data. Users of data mining should start thinking about how their use of this technology will be impacted by legal issues related to privacy.*

- **Customer Acquisition and Data Mining.** *For most businesses, the primary means of growth involves the acquisition of new customers. This could involve finding customers who previously were not aware of your product, were not candidates for purchasing your product (for example, baby diapers for new parents), or customers who in the past have bought from your competitors. Data mining can often help segment these prospective customers and increase the*

response rates that an acquisition marketing campaign can achieve.

- **Campaign Optimization: Maximizing the Value of Interacting with Your Customers.** *In most marketing organizations, there are a wide variety of ways to interact with customers and prospects. Besides the many possible offers that can be made, there are now multiple communication channels (direct mail, telemarketing, email, the web) that can be used. The process of marketing campaign optimization takes a set of offers and a set of customers, along with the characteristics and constraints of the campaign, and determines which offers should go to which customers over which channels at what time.*

- **Scoring Your Customers.** *Once a model has been created by a data mining application, the model can then be used to make predictions for new data. The process of using the model is distinct from the process that creates the model. Typically, a model is used multiple times after it is created to score different databases.*

- **Understanding Data Mining: It's All in the Interaction.** *Data mining is a relatively unique process. In most standard database operations, nearly all of the results presented to the user are something that they knew existed in the database already. Data mining, on the other hand, extracts information from a database that the user did not know existed. Relationships between variables and customer behaviors that are non-intuitive are the jewels that data mining hopes to figure out. This is where visualization comes in.*

- **Visualizing Data Mining Models.** *The purpose of data visualization is to give the user an understanding of what is going on. Since data mining usually involves extracting "hidden" information from a database, this understanding process can get somewhat complicated. Because the user does not know beforehand what the data mining process has discovered, it is a much bigger leap to take the output of the system and translate it into an actionable solution to a business problem. This paper describes a number of methods to visualize data mining models and provide the user with sufficient levels of understanding and trust.*

- **Some Thoughts on the Current State of Data Mining Software Applications.** *As a former developer of data mining software, I can understand how difficult it is to create applications that are relevant to business users. Over the past few years the technology of data mining has moved from the research lab to Fortune 500 companies, requiring a significant change in focus. The core algorithms are now a small part of the overall application, being perhaps 10% of a larger part, which itself is only 10% of the whole. With that in mind, the focus of this article is to point out some areas in the remaining 99% that need to be improved upon.*

- **From Data Mining to Database Marketing.** *The market for data mining — if you believe the hype — will be billions of dollars by the turn of the century. Unfortunately, much of what is now considered data mining will be irrelevant, since it is disconnected from the business world. In general, marketing analysts predictions that the technology will be very relevant to businesses in the future are correct. The key to making a successful data mining software product is to embrace the business problems that the technology is meant to solve, not to incorporate the hottest technology. In this report I will address some of the issues related to the development of data mining technology as it relates to business users.*

- **Increasing Customer Value by Integrating Data Mining and Campaign Management Software.** *As a database marketer, you understand that some customers present much greater profit potential than others. But, how will you find those high-potential customers in a database that contains hundreds of data items for each of millions of customers? Data Mining software can help find the "high-profit" gems buried in mountains of information. However, merely identifying your best prospects is not enough to improve customer value. You must somehow fit your Data Mining results into the execution of marketing campaigns that enhance the profitability of customer relationships. This white paper describes how you can profit from the integration of Data Mining and Campaign Management technologies.*

- **Data Mining Can Bring Pinpoint Accuracy to Sales.** *Data warehousing - the practice of creating huge, central stores of customer data that can be used throughout the enterprise - is becoming more and more commonplace. But data warehouses are useless if companies don't have the proper applications for accessing and using the data. Two popular types of applications that leverage companies' investments in data warehousing are data mining and campaign management software.*

- **An Overview of Data Mining at Dun & Bradstreet.** *This document is a survey of data mining projects and opportunities throughout the Dun & Bradstreet organization. Data mining is a powerful new technology with greater potential to help D&B "preemptively define the information market of tomorrow." D&B companies already know how to collect and refine massive quantities of data to deliver relevant and actionable business information. In this sense, D&B has been "mining" data for years. Today, some D&B units are already using data mining technology to deliver new kinds of answers that rank high in the business value chain because they directly fuel return-on-investment decisions. In the D&B units surveyed, we found strong interest and a wide range of activities and research in data mining.*

- In August 1998 I chaired the "Keys to the Commercial Success of Data Mining" workshop, held in conjunction with KDD'98. The workshop archives are available online and include both the working notes and presentations.

Useful Data Mining & CRM References:

If you would like to get more information on data mining and CRM, you might want to look at the following sites:

- **Introduction to Data Mining and Knowledge Discovery** by Two Crows Corporation. This concise overview of data mining provides a great introduction to the field of data mining. Two Crows is one of the leading consulting groups in data mining.

- **Pan For Gold In The Clickstream** by Herb Edelstein (Information Week). An excellent article on the practical application of data mining to business problems. Herb is one of the leading proponents of making data mining useful in real world situations.

- **Data Mining on the Web: There's Gold in that Mountain of Data** by Dan Greening (Web Techniques Magazine). A good article on what data mining and personalization can do to improve interactions with web site visitors.

- **Mining Customer Data** by Gary Saarenvirta (DB2 Magazine). A very informative article on the use of data mining for customer segmentation and clustering. Saarenvirta does a good job discussing issues related to data preparation and transformation.

- **A Comparison of Leading Data Mining Tools** (PDF format). A presentation by John F. Elder IV and Dean W. Abbott from the KDD'98 data mining conference. A thorough comparison of most of the important software applications in the data mining space. Although a bit dated, this presentation is still a useful read.

- **Visual Database Exploration Techniques** (PDF format). A set of tutorial notes by Professor Daniel Keim, University of Munich. An extremely complete look at visualization techniques for data mining and data analysis.

- **Knowledge Discovery Nuggets**. This site is a great starting point. KD Nuggets contains a very large collection of links to data mining companies, conferences, and software.

- **SIG KDD Explorations**. This is the online publication of the data mining special interest group of the ACM. A good place to check for recently published technical work.

- **The Data Warehousing Information Center**. A site with very thorough coverage of data warehousing, OLAP, decision support, and data mining. Includes introductory information plus a lot of information on vendors.

My Background:

Before I joined Vertex, I led the advanced technology group at Capital One, where I focused on accelerating the use of new analytic technologies and techniques.

Before Capital One, I was director of Engineering at AnVil, an *in silico* drug discovery company focused on the commercial analysis of biological and clinical datasets. I was responsible for the development of AnVil's data analysis platform technology (ADAPT), an award winning system of data mining tools used to automate the analysis of everything from gene expression microarray data to clinical healthcare records.

I came to AnVil from Wheelhouse, a marketing technology and services company, where I was Chief Scientist. I founded the engineering organization, managed software development efforts, and set technology strategy. In addition, as a Wheelhouse senior management team member, I performed numerous corporate duties including engaging clients, making sales calls, evaluating technologies, public speaking, fundraising (including a $52M series B investment round), etc.

Prior to my position at Wheelhouse, I was Director of Analytics at Xchange Inc., a leading CRM software vendor (and now part of Amdocs, Inc.). I was responsible for directing the integration of analytic applications (data mining, customer optimization, decision support, and visualization) into Xchange's suite of marketing automation software applications.

Before Xchange, I co-founded the data mining group Dun & Bradstreet. At the time, D&B was a complicated collection of over twenty-five companies whose main purpose was to collect information and turn it into a form that other companies could use. This data covered everything from TV Ratings (Nielsen Media) to prescriptions (IMS) to grocery purchases (A.C. Nielsen). Analyzing this data was a major component to the business and I was a consultant to the various divisions, providing help in the areas data mining and high-performance computing. I also worked with the Pilot Software division of D&B, and developed new software applications to put data mining solutions in the hands of business users.

Before D&B I was a senior scientist at an amazing company called Thinking Machines Corporation. TMC was a pioneer in the commercial development of massively parallel supercomputers. While I was at Thinking Machines I helped create Darwin, one of the first commercial data mining applications. Our early work with Darwin made use of the massive computational power available with a supercomputer but later versions were adapted for use on less esoteric hardware. TMC eventually moved out of the supercomputer business and turned itself into a company focused exclusively on the data mining software market. Oracle corporation eventually acquired TMC and incorporated Darwin into their database platform.

If you would like to know even more about me, you can check out my LinkedIn profile, resume, or list of publications. Besides the particulars of my work experience and education, you can find links to many of the papers I have written. The topics of these papers include data mining, artificial life, time-series prediction, parallel computers, and the future of personal computing.

[Data Mining Books] [White Papers] [Data Mining Tutorial]