DATA MINING
in the pharmaceutical industry

A research-based pharmaceutical company is a data accumulating wonder. More than in any other industry, success is predicated on the collection, processing and exploitation of that data. This is not always recognised and often not planned for by large pharmaceutical companies. Now, however, with the advent of data storage and mining techniques making major advances in other industries, the pharmaceutical industry must adjust to fully exploit this potential competitive advantage in discovery, development and marketing of their products. We discuss some of the impacts to companies and some of the adjustments they will have to make to maintain their position in the information age.

By Dr Robert D. Small and Herbert A. Edelstein

Sixteen years ago Peter Drucker[1] pointed out that the pharmaceutical industry was an information industry – not a manufacturing or even a health industry. Since then, information technology has raced beyond anything he could have imagined. We now routinely collect and process gargantuan amounts of data quickly and cheaply.

Drucker's original insight did not receive much attention in the pharmaceutical industry at the time, and given the subsequent growth and prosperity of the industry, one could argue that this ignorance has not been very damaging.

However, the pharmaceutical industry is now moving to embrace this viewpoint and the technological and organisational changes it demands. Indeed, some of the biggest contributions to the recent growth of pharmaceutical companies have resulted from capitalising on the increased availability of data, improved information systems and database management and the introduction of bioinformatics.

This size and complexity of the databases now proliferating in the pharmaceutical business is a major departure from most of the clinical and R&D databases of the past. Analysis methods that were useful on older, smaller databases do not always scale up to accommodate large amounts of data, necessitating the introduction of new methods and software tools, in particular data mining.

Data mining is a process that uses a variety of analysis and modelling techniques to find patterns and relationships in data. These patterns can be used to make accurate predictions that aid in solving problems across the entire spectrum of drug development, including R&D, clinical trials and marketing.

For example, a data mining contest[2] was recently held to predict the molecular bioactivity for a drug design; specifically, determining which organic molecules would bind to a target site on thrombin. The predictions were based on about 500 megabytes of data on approximately 1,900 organic molecules, each with more than 130,000 attributes (or dimensions, as they are called in data mining). This was a challenging problem not only because of the large number of attributes but because only 42 of the compounds (2.2%) were active. The relatively small number of cases (organic molecules in this example), compared to the large number of attributes, makes the problem

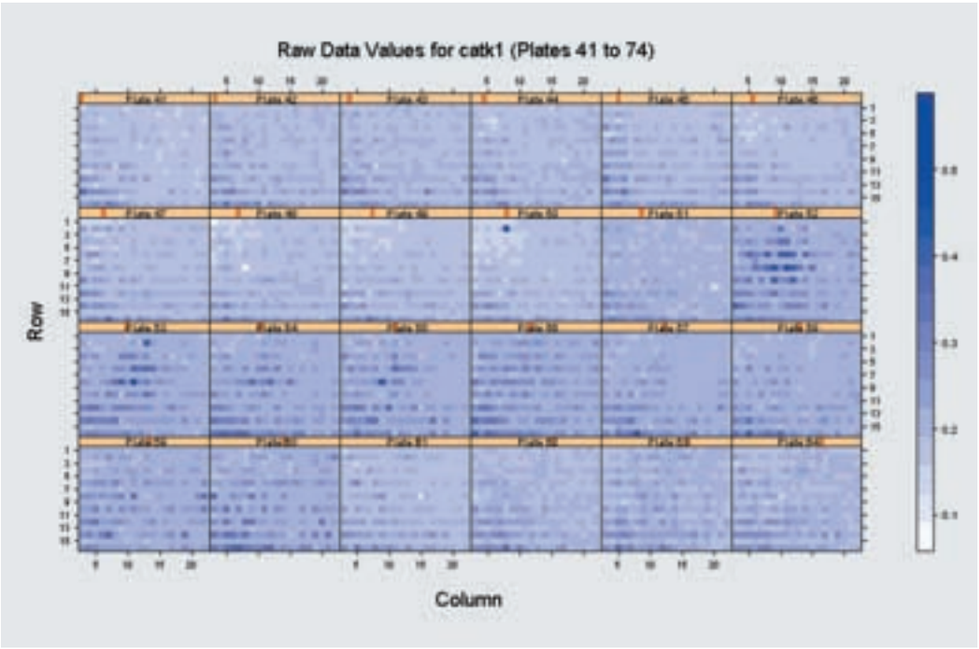

Figure 1
Merck's High-Throughput Screening business analytic based on Insightful Corp's S-PLUS and StatServer products allows Merck chemists to apply advanced statistical methods to well-plate data on thousands of promising compounds, and interpret the results quickly and easily using graphical methods

Graphic courtesy B. Pikounis, Merck Research Laboratories and Insightful Corporation

even more difficult. Of the 136 contest entries, about 10% achieved the impressive result of more than 60% accuracy[6], with the winner, Jie Cheng of the Canadian Imperial Bank of Commerce, reaching almost 70% accuracy.

As one of the most data intensive industries, the pharmaceuticals business has a wealth of such potential data mining applications from which it will gain substantial benefits. Tremendous amounts of data are collected during the development of a drug. The trend is to collect larger amounts of it automatically. A great deal of the basic biology data is now collected online in laboratories. More clinical data is being collected with electronic diaries, in clinic laptops, or even on devices attached to the patient in some way. These range from applications common to most industries such as marketing and sales to unique opportunities in clinical trials and research and development such as genomics and proteomics. Clearly, data mining has much to contribute to the pharmaceutical business.

The rise of data mining

The term data mining, in its most common use, is very new. The term previously had been used pejoratively by some statisticians and other specialists to refer to the process of analysing the same data repeatedly until an acceptable result arose[3]. By the early 1990s, a number of forces converged to make data mining a very hot topic. It has subsequently been widely applied in retailing, banking and financial services, insurance, marketing and sales and telecommunications.

At first, parts of the scientific community were slow to embrace data mining. This was at least partially due to the marketing hype and wild claims made by some software salespeople and consultants[4]. However, data mining is now moving into the mainstream of science and engineering.

Data mining has come of age because of the confluence of three factors. The first is the ability to inexpensively capture, store and process tremendous amounts of data. The second is advances in database technology that allow the stored data to be organised and stored in ways that facilitate speedy answers to complex queries. Finally, there are developments and improvements in analysis methods that allow them to be effectively applied to these very large and complex databases.

It is important to remember that data mining is a tool, not a magic wand. You can't simply throw your data at a data mining tool and expect it to produce reliable or even valid results. You still need to know your business, to understand your data, and to understand the analytical methods you use.

Furthermore, the patterns uncovered by data mining must be verified in the real world. Just because data mining predicts that a gene will express a particular protein, or that a drug is best sold to a certain group of physicians, it doesn't

mean this prediction is valid in the real world. You still need to verify the prediction with experiments to confirm the existence of a causal relationship.

The data mining process

For success in data mining it is essential to follow a methodical process, such as the following seven-step process[5]:

1 Define the business problem
2 Build the data mining database
3 Explore the data
4 Prepare the data for modelling
5 Build a model
6 Evaluate the model
7 Act on the results

Although the numbering implies a linear process, data miners often find themselves revisiting earlier steps based on what they have learned.

The first step is to prepare a clear statement of the problem you are trying to solve. As you proceed through the steps of data mining, however, your deepening understanding of the data and the problem will occasionally lead you to reformulate your objectives.

The next three steps involve preparing the data for mining and lead up to the actual model building. Together, they take more time and effort than all the other steps combined, typically consuming 60% to 95% of a project's time and resources.

The data to be mined is usually gathered from multiple sources, and while in some cases it is possible to mine those sources directly, more often it is preferable to gather the data into a uniformly designed database first. Though a daunting task, such integration is worthwhile. For example, integrating all the databases having to do with a single drug allows a company to more easily respond to a regulatory agency that suspects a problem. Typically, the data relating to the drug is spread across tens to hundreds of databases of vastly different design, conforming to different standards and even stored in different database management systems. Analysis would be very difficult; millions of dollars would be spent in producing the requested safety summaries. If this data were properly consolidated, on the other hand, data mining would enable the company to quickly explore the data and identify both drug and patient-related candidate factors that may have raised the regulatory concern.

There are also external databases that can be mined in conjunction with corporate data. On the R&D side, hundreds of public and licensed

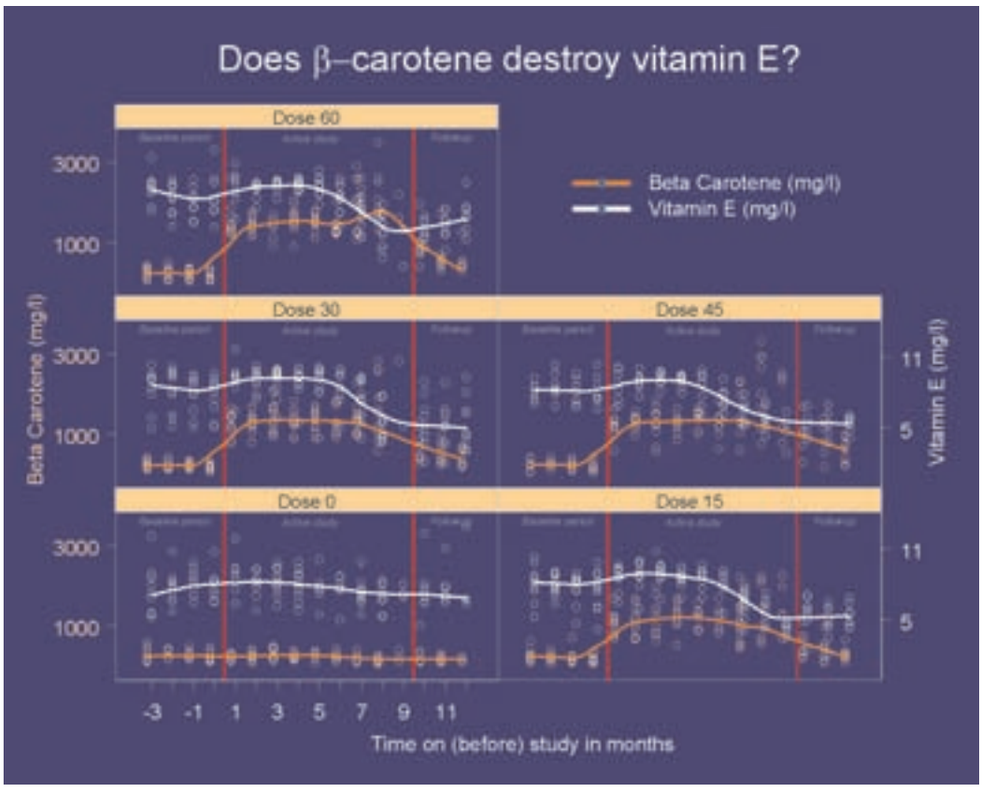

Figure 2
In clinical trials, understanding multivariate, time-dependent results can be challenging. This Trellis graphic shows the relationship between beta-carotene and Vitamin E in the bloodstream, for five different dosage levels of supplemental beta-carotene, as a local average of hundreds of blood sample measurements on 45 patients over a 16-month period

S-PLUS graphic courtesy Insightful Corp

Figure 3
Data Mining Products

COMPANY	PRODUCT	WEB ADDRESS
Angoss Software 34 Saint Patrick Street, Suite 200, Toronto, Ontario, M5T 1V1, Canada	KnowledgeStudio	www.angoss.com
IBM Route 100, Mail Drop 1302, Somers, NY 10589, USA	Intelligent Miner	http://www-4.ibm.com/software/data/iminer/fordata/
Inforsense 47 Prince's Gate, London SW7 2QA, UK	Kensington Discovery Edition	www.inforsense.com
Insightful Corporation 1700 Westlake Avenue N, #500, Seattle, WA 98109, USA	S-Plus 6	www.insightful.com
MarketMiner 1575 State Farm Blvd, Charlottesville, VA 22911, USA	ModelQuest MarketMiner	http://www.marketminer.com/
Megaputer Intelligence Inc 120 West 7th Street, Suite 310, Bloomington, IN 47404, USA	PolyAnalyst	http://www.megaputer.com/
Microsoft NCR Computer Sys Grp, 17087 Via Del Campo, San Diego, CA, 92127, USA	Data Mining Services Teradata Warehouse Miner	http://www.microsoft.com/data/oledb/dm.htm http://www.ncr.com/products/software/teradata_mining.htm
Norkom Technologies Ltd Norkom House, 43 Upper Mount Street, Dublin 2 Ireland	Alchemist	http://www.norkom.com/
Oracle	9*i* Data Mining 9*i* Personalization	http://www.oracle.com/ip/analyse/warehouse/datamining/
Quadstone 321 Summer Street, Boston, MA 02210, USA	Quadstone	www.quadstone.com
Salford Systems 8880 Rio San Diego Dr, Suite 1045, San Diego, CA 92108, USA	CART 4.0 Mars 2	http://www.salford-systems.com/
SAS SAS Campus Drive, Cary, NC 27513, USA	Enterprise Miner	http://www.sas.com/
Silicon Graphics 2011 N. Shoreline Blvd, Mountain View, CA 94043, USA	MineSet	http://www.sgi.com/software/mineset/
SPSS 233 S. Wacker Drive, 11th Floor, Chicago, IL 60606, USA	Clementine	http://www.spss.com/clementine/
Torrent Systems 5 Cambridge Center, 7th Floor, Cambridge, MA 02152, USA	Orchestrate	http://www.torrent.com/
Unica Solutions, Inc Lincoln North, Lincoln, MA 01773, USA	Affinium Model	http://www.unicacorp.com/products/model.htm
Urban Science 200 Renaissance Center, 19th Fl, Detroit, MI 48243, USA	GainSmarts	http://urbanscience.com/

genomic databases are available. These databases can be brought in-house. On the marketing side, there are databases of individuals' demographics and behaviours from vendors such as Acxiom (www.acxiom.com). Increasingly, companies are licensing this data for internal use, accessing it through tools such as IBM's Discovery Link (http://www3.ibm.com/solutions/lifesciences/index.html), or using it through services such as those from Viaken (www.viaken.com).

Once the data mining database has been built, it is time to explore the data. Data visualisation often yields insights to help build better predictive models. The graphs in **Figures 1** and **2** show how visualisations can help the data miner by providing a comprehensive yet concise representation of the data. They also show that even a good visualisation takes some training and experience to interpret.

The final step in data preparation is to transform the data for mining. Ideally, you feed all your attributes into the data mining tool and let it determine which are the best predictors. In practice, this doesn't work very well. Not only can you introduce problems by including too many irrelevant attributes, but in all likelihood the best predictors are actually combinations of other attributes. For example, BMI (body mass index, a calculated value based on height and weight) may be more important than either height or weight in predicting the efficacy of a drug.

The fourth step is actually building the model. The most important thing to remember about model building is that it is an iterative process. You will need to explore alternative models to find the one that is most useful in solving your business problem. The process of building predictive models requires a well-defined training and validation protocol in order to generate the most accurate and robust predictions. This kind of protocol is sometimes called supervised learning. The essence of supervised learning is to train (estimate) your model on a portion of the data, then test and validate it on the remainder of the data.

In the sixth step you evaluate your models' results and interpret their significance. Remember that the accuracy rate found during testing applies only to the data on which the model was built. In practice, the accuracy may vary if the data to which the model is applied differs significantly from the original data.

Once a data mining model has been built and validated, it can be used as a general guideline for action or it can be applied to a large batch of data, such as microarray data. It can also be applied to a single case at a time. For example, as we learn how individuals respond to a drug, a model may be interactively applied to determine the safest and most effective drug to prescribe for that person.

There are a large number of data mining tools available. A partial list is shown in **Figure 3**[6]. In general, the products available are quite good, and are especially strong in model building. In the last year, there has been great improvement in the ease of using the models you build. However, data preparation and visualisation still needs to be made more effective.

Organising for data mining success

In order to enjoy the maximum benefits of their data resources and realise the potential offered by mining them, companies will have to adopt some of the best organisational practices of those who have succeeded with data mining.

Perhaps the most important change is to recognise that pharmaceutical research, development and marketing are part of an integrated effort and this integration must be backed with a management commitment that supports the integration of data across the corporation.

This integration will also be reflected in a broadened knowledge and understanding of how to use data by all participants in the drug development and marketing process. There is no doubt that the average biomedical scientist and physician of today is much more sophisticated in many areas of computer usage than even five years ago. But it is no longer sufficient to partition and segment duties among narrowly trained specialists. What is needed is a more generalist approach and an integration of job roles and skills.

Because the data now being collected is so diverse, massive and complicated, a team approach that brings together domain experts, database experts, and analytical experts is required. Such a team needs a holistic view that cuts across both domain and technical areas with each member understanding and appreciating the contribution of the others.

There are numerous examples of the kinds of integration needed in multiple disciplines such as biostatistics, bioinformatics, pharmacoepidemiology, pharmacoeconomics and pharmacogenomics that cut across traditional boundaries. Though they are only first steps, they have important common traits. All involve integration of domain areas that were once thought to be unconnected. They all require an understanding of data structures, storage technology and data models that would have been unthinkable a decade ago. A researcher or

References

1 Drucker, Peter F. Innovation and Entrepreneurship, 1985, Harper & Row. NY, NY.

2 Knowledge Discovery in Databases: 10 Years After. SIGKDD Explorations, ACM SIGKDD V.1 59-61, 2000.

3 Friedman , HP and Goldberg, Judith. Knowledge Discovery from Databases and Data Mining: New Paradigms for Statistics and Data Analysis. Biopharmaceutical Report V.8 No.2 2000.

4 DuPont Pharmaceuticals Research Laboratories and KDD Cup 2001.

5 Edelstein, Herb. Introduction to Data Mining and Knowledge Discovery, 1999, Two Crows Corporation.

6 Edelstein, Herb. Data Mining Technology Report: 2002, Forthcoming, Two Crows Corporation.

contributor in any of these areas needs training or experience in multiple domain areas as well as in information science.

Interestingly, many of these arguments about multiple skills and technical vs domain knowledge now occur at data mining conferences and in data mining publications. Though the emphasis is somewhat different than those made here, the source is similar – a confluence of multiple disciplines and huge amounts of data and computer power provide new and only recently realised opportunities.

Data and the future of the pharmaceutical industry

New approaches are needed in order to fully realise the potential of technologies that allow for the creation, acquisition, storage and analysis of databases of unprecedented size and complexity. The primary change is one of attitude. When database technology first arrived in the pharmaceutical industry, most researchers thought that it was useful to have programmers who had some domain knowledge. It was common to give newly hired IT people a course in medical terminology. The modern company needs to view itself as a data machine whose primary business is collecting, processing, analysing and using its data as the primary resource of the company.

This trend will continue and accelerate. Genomic and related technologies allow for more data to be collected on each patient both during the development of a drug and after it is marketed. General IT technologies continue to move in a direction that allows more data to be collected, processed and analysed. There are demands for more information about each product from patients, physicians and regulators.

These pressures will lead to certain changes in the successful pharmaceutical company. The company will have to invest substantial effort in integrating its large and diverse data sources. This is a stupendous effort that we could not hope to describe here. Suffice to say it will involve a change in the status of IT people in the organisation, a demand for more training and experience in IT for people all over the organisation and an increased level of domain expertise by IT and associated workers.

The company will be storing and processing all of these data with a view to improving the business. This will require access with appropriate tools by an analytically astute workforce. More people will have more access to more data. More people will need analyses and interpretations of the data. A larger share of all decisions will be made with the support of the pointed analyses of appropriate data. The structure, job descriptions and likely organisations will change.

We are living in an age that emphasises a great increase in collection and use of data. It is not surprising that the pharmaceutical industry, which has been an information industry for decades, is even more affected by these trends. The fact that some of these changes have happened so quickly and that many in pharmaceutical and medical research did not recognise the previous emphasis on information may have disguised the cataclysmic events taking place. Nevertheless they are coming and data mining will play a major and steadily increasing role in pharmaceutical research in the 21st century. DDW

Bob Small *is currently Vice-President of Data Mining Technologies for GlaxoSmithKline. After an academic career that included appointments at Temple University and the University of Pennsylvania, Bob joined Merck. There he took on positions of increasing responsibility serving as Senior Research Fellow, Senior Director and Executive Director of the Biometrics Department. He later served as Head of Biometrics for Burroughs Wellcome, Glaxo Wellcome and Pfizer. In 1996 he joined Two Crows Corp, a data mining consultancy, as VP of Research. Bob is widely published in statistical, medical and biological journals, is active in the American Statistical Association (ASA) and is Chair Elect of the Biopharmaceutical Society Section of the ASA. He has a BS from the University of Maryland in Maths and Physics and a PhD in Biomathematics from North Carolina State University.*

Herbert Edelstein *is President of Two Crows Corporation. He is an internationally recognised expert in data mining, data warehousing and client-server computing, consulting to both computer vendors and users. He is regularly invited as a chair and keynote speaker at conferences on these topics and is a founder of the Data Warehousing Institute. Prior to Two Crows, Herb was a founding partner of Euclid Associates, a consulting firm specialising in data warehousing and data management. He was also Vice-President of Marketing and Sales at Sybase and International Database Systems, General Manager of the Model 204 division of Computer Corporation of America and a consultant for American Management Systems.*