

Letter from the President



I am delighted to be taking up the post of president of ENBIS from September 2004 to September 2005. It is a great honour to have the opportunity to serve the community of practical, active statisticians whom I admire. I hope we shall have a fruitful and interesting year.

One of my statistical interests is the cross-fertilisation of ideas from one speciality to another. This is all about communication, and could accelerate our natural progress towards statistical understanding. Industrial processes and products often have several response variables, as well as many input or control variables. Steel, for example, has qualities of toughness, tensile strength, hardness, corrosion resistance, and more, all of which depend on the compositional and processing variables. Chemical processes may be judged by several quality variables as well as by yield. We find new ways to study the relationships in industrial processes by looking at how social study questionnaires are analysed, since they also have multiple inputs and outputs, albeit of a different nature. Another example of cross-fertilisation arose when a private company was asked to study the

problem of leakage from a water pipeline crossing 3,000 miles of northern Russia. The problem was how to choose the most effective and yet parsimonious sampling scheme to locate probable leaks. Experience from gold prospecting gave the answer: look for changes in the nature of the soil along the pipeline route. Statistical methods used there could be applied to the pipeline leakage problem.

Another of my interests is learning to present statistics in a clear form, and making better use of the popular press to share ideas. In my own university at Newcastle, we have been asked to visit all other research groups to persuade them to include statistical thinking in their research proposals, and to ensure that the statistical work is fully funded. This demands clarity of expression of statistical ideas in ways that people in other disciplines can understand, appreciate, and take on board. Similarly, we have support from European regional development funds to support us in the spread of statistical thinking in small- to medium-sized enterprises.

A major force for achieving the aims of ENBIS during the last three years has

been 'Pro-ENBIS': a thematic network contracted for three years by the European Commission to implement the aims of ENBIS. At the end of the Pro-ENBIS project, this year, we shall produce a report on the 'state of the art in business and industrial statistics'. This report will include case studies in all sections and these should be very useful when we talk to people working in different sectors about the benefits of statistical thinking. We hope the report will be publicly available, perhaps as a published book.

I am very pleased that we are holding the next ENBIS conference in Newcastle from 14 to 16 September 2005. While the excellent Copenhagen conference is still in the minds of those who attended, please send me any thoughts and recommendations that will help in the planning of the 2005 conference.

Shirley Coleman
President of ENBIS
Shirley.Coleman@ncl.ac.uk



There's gold in those databases

Diego Kuonen explains data mining from a statistical perspective. In a future issue he will show how data miners may collaborate with bioinformatics colleagues to unlock secrets of the living cell

In the past, mining for gold consisted of choosing a site and then sifting through dirt. Sometimes the miner found only a few valuable nuggets, sometimes he hit upon an entire vein, but most of the time, he found nothing at all. He (statistically, it usually was a 'he') then decided to move to another spot, or to give up altogether. Today, mineral mining is much more accurate and productive. Mining for data, and learning from it, has also become more accurate and productive.

But, what is data mining? Statistics can be defined as the science of 'learning from data' and data mining sits at the interface between statistics, computer science, artificial intelligence, machine learning, database management, and data visualisation. Its definition changes with the user's perspective:



● 'Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.' (M. J. A. Berry and G. S. Linoff)

● 'Data mining is finding interesting structure (patterns, statistical models, relationships) in databases.' (U. Fayyad, S. Chaudhuri and P. Bradley)

● 'Data mining is the application of statistics in the form of exploratory data analysis and predictive models to reveal patterns and trends in very large data sets.' ('Insightful Miner 3.0 User Guide')

Like statistics, data mining is not only modelling and prediction, nor a product that can be bought, but a whole problem-solving cycle/process that must be mastered through team effort. Indeed, data mining can be seen as the core component of the so-called 'knowledge discovery in databases' (KDD) process; see Figure 1. For learning from data to proceed, data from many sources ('databases') must first be gathered

together and organised in a consistent and useful way ('data warehousing'). To do this properly, data need to be cleaned and preprocessed ('data cleaning'). Quality decisions and mining results come from quality data. Data is always dirty and seldom ready for data mining in the real world. But, before data can be analysed, a task-relevant data target needs to be created ('data selection').

The main part of 'knowledge discovery in databases' is data mining – the analysis of data and the use of statistical learning techniques for finding patterns and regularities in sets of data. The choice of a particular combination of techniques to apply in a particular situation depends on the nature of the data mining task and the nature of the available data. Briefly, the main tasks well-suited for data mining are:

classification (examine an object and assign it to one of a predefined set of classes).

Example: classify credit applicants as low, medium or high risk;

estimation (given some input data, estimate a value for some unknown continuous variable).

Example: estimate the lifetime value of a customer; estimate the probability that someone will respond to a treatment;

prediction (the same as classification and estimation, but records are classified according to some

Communication is a puzzling thing: Tom Lang's puzzle

Antje Christensen shows how a simple puzzle may be used to teach the principles of good communication between a consulter and a consultant

From solution found to solution communicated

Have you ever experienced that chill when your presentation of a brilliant technical solution to a pressing problem is met with a blank expression on the face of your client? Have your reports ever collected dust on a shelf, with nobody making the effort to read them? I think most statisticians have such an experience at some time in their careers. Though the assumption may be tempting in the heat of the moment, your client is probably not dumb. You are probably not dumb either, as you actually solved the problem you were asked to solve. However, it can be difficult to step back from your own technical under-

standing and evaluate how your client will understand your technical explanations. Technical-writing teacher Tom Lang has designed a puzzle to show the pathway from knowing the solution to communicating it successfully.

The puzzle

It's a worthwhile experience, so find a partner and give it a try. Here is what you need:

The figure opposite shows a diagram of the assembled puzzle. You need a copy of this diagram. Furthermore, you need the actual puzzle. Copy the diagram onto a sheet of Styrofoam or a piece of wood and cut along the

lines, so you end up with seven puzzle parts. A thick material like Styrofoam is more effective than a flimsy material, like a sheet of paper or cardboard. Preferably, all surfaces of the puzzle parts should look the same – Styrofoam coated or painted on one side is less effective for the experiment.

Two people are involved in the experiment, the writer and the reader. The writer has the map: the solution. The reader has the puzzle parts: the problem. The reader is not allowed to see the map. Your partner is the obvious choice for that role, if you are the one who prepared the puzzle. Now you, as the writer, describe how to solve the puzzle, and the reader follows your instructions. You give the instructions orally. If you want to mimic written communication, don't watch the reader's progress, and don't let him or her ask you questions. When you feel you are done, look at

future behaviour or estimated future value).

Example: predict treatment efficacy from drug properties, patient demographics and tissue state); **affinity grouping** or **association rules** (determine which things go together).

Example: the items in a supermarket shopping cart. Also known as dependency modelling;

clustering (segment a population into a number of subgroups or clusters).

Example: partition data from a large treatment database into clusters of similar patients to enable the selection of a separate treatment strategy for each patient group; and

description and visualisation.

Learning from data comes in two flavours: directed ('supervised') and undirected ('unsupervised') learning. The first three tasks – classification, estimation and prediction – are all examples of supervised learning. In supervised learning, the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data ('class prediction'). The next three tasks – affinity grouping or association rules, clustering, and description and visualisation – are examples of unsupervised learning. In unsupervised learning no variable is singled out as the target; the goal is to establish some relationship among all the variables ('class discovery'). Unsupervised learning attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes. This

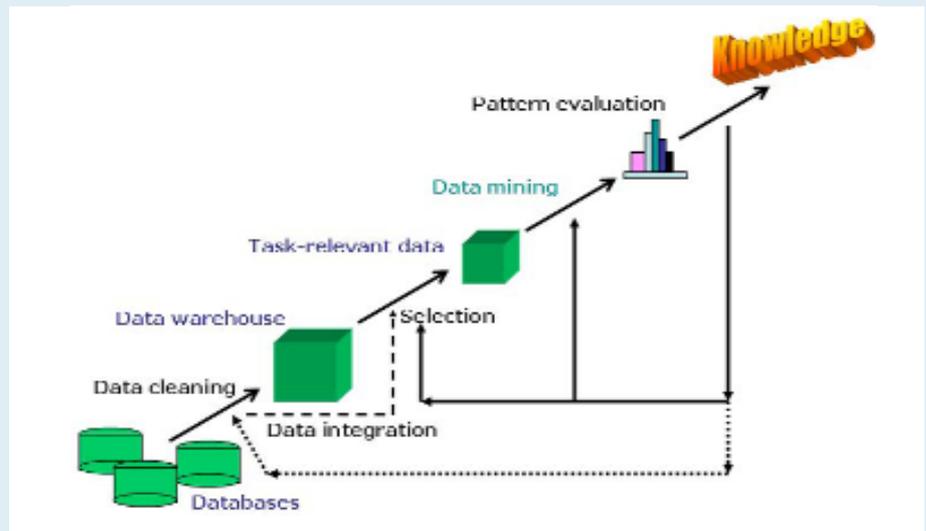


Figure 1: The KDD process and its phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required.

is similar to looking for needles in haystacks.

Many data mining techniques can be described as flexible models and methods for exploratory data analysis; they are nothing other than multivariate data analysis methods. In the words of I. H. Witten and E. Franke: 'What's the difference between machine learning and statistics? Cynics, looking wryly at the explosion of commercial interest (and hype) in this area, equate data mining to statistics plus marketing'.

Importantly, data mining can learn from statistics; and, to a large extent, statistics is

fundamental to what data mining is trying to achieve. Data mining and statistics will inevitably grow closer in future because data mining will not become knowledge discovery without statistical thinking, and statistics will not be able to succeed on massive and complex datasets without data mining approaches.

Diego Kuonen is CEO of Stattoo Consulting, Lausanne, which provides statistical consulting, data analysis and data mining services. Email: kuonen@stattoo.com

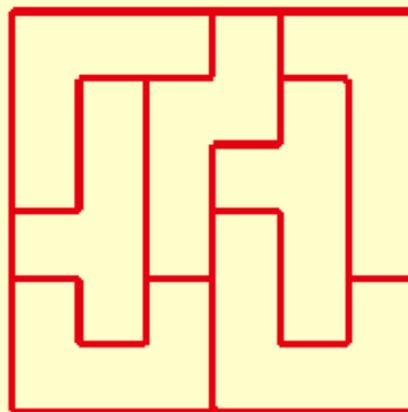
the result and discuss it with the reader.

I prefer to keep this exercise to two people rather than in a class. However, if you do the exercise in a class, the rest of the class listens to your instructions and watches the progress of the reader. The writer, the reader and the audience all take part in the discussion following the end of the exercise. Therefore, no recording is done. For larger classes, put the puzzle parts on an overhead projector, so everybody can follow the reader's progress.

If you and your reader did not succeed in assembling the puzzle, you are in good company. According to Tom Lang, although a lot of technically skilled people have tried the exercise, the vast majority did not succeed.

A journey towards a destination

Most probably, all your instructions were technically correct. Obviously, the solution you proposed was correct. The difficult part was communication. Whatever the technical content,



The assembled puzzle.

communication can be seen as a journey that you undertake together with the person you communicate with. If you have not assembled the puzzle successfully, you have not arrived at your destination together.

The first opportunity to lose each other is

where you set off. All communication (possibly with the exception of proofs in formal logic) is based on shared information that is not stated explicitly. If you do not agree with your reader about some of this implicit information, misunderstandings may occur. Tacit assumptions that you may have made about your reader in the puzzle experiment are that he/she:

- Understands the language/lingo;
- Has the same puzzle;
- Has a complete puzzle;
- Attaches a label that you give a piece to the same piece as you;
- Expects the puzzle to be flat, not stacked;
- Expects the puzzle to be geometrically regular;
- Will not turn pieces over; and
- Has followed earlier directions successfully.

If you did not succeed, probably one or more of the assumptions were not fulfilled at some stage during the assembly. You can make your instructions easier to follow by making your assumptions explicit. For example, start off

The retiring president writes...

We have visionaries, we have enthusiasts, we have members; all share the aim of ENBIS:

- To promote the widespread use of sound, science-driven, applied statistical methods in European business and industry.

We have extended our membership to nearly a thousand people across 40 nations. I read this growth in membership as an indication that the creation of ENBIS was timely. Although we do not keep a record of all the activities performed by individual members of our network, I believe that we are all actively pursuing the society's objective in one way or another.

A major force for achieving our aim during the last three years has been Pro-ENBIS. This was a thematic network contracted for three years by the European Commission to implement the aim of ENBIS. As part of the contract, deliverables were quantified. So we have a record of

works visits such as to:

- An Italian company making transmission belts;
- A Swedish company developing components for three major truck companies;
- Another Swedish company, this time making heavy-duty diesel engines;
- An English company making steering units;
- An Italian bank;
- A Spanish electrical power and transmission company;
- An American medical devices company based in England;
- A Portuguese hardwood pulp producer; and
- A Norwegian electrical heating company.

We have a record of many workshops on design of experiments, quality control, six sigma, data mining, and statistical methods for business and industry. From these have sprung requests for

more workshops, for courses, and for consultancy. A full report will be published early next year.

Soon after I hand over the presidency to Shirley Coleman, our Pro-ENBIS contract will end and ENBIS will face the challenge to continue these activities without financial support from the EU. So where will our driving force come from? From our visionaries, our enthusiasts and those members who work behind the scenes to plan and organise our conference and other meetings, maintain and extend our website and care for our finances.

'Where there is no vision the people perish' (*Proverbs, 29:18*). Only with vision will ENBIS continue to succeed. Past presidents may dream dreams; young members will see visions. I am confident that the vision will prevail, to the benefit of our members and of European business and industry.

Poul Thyregod

describing the pieces of the puzzle so that the reader knows whether he or she has the same puzzle and all seven pieces. At the same time, you can establish labels for the pieces that will make reference to them unambiguous. State that the assembled puzzle is a flat square to make sure you both head towards the same destination.

The second opportunity to lose your reader is on your way towards the common destination. You can make it easier for your reader to follow you by structuring the information and giving it in the order in which he or she needs it. You can focus your communication by analysing your anticipated readership and clarify for yourself what needs your readers probably will have. Depending on your readership, you will be able to incorporate shared concepts. For example, instead of describing the shape of the assembled puzzle in detail, you can call it a square. In many technical environments, more complex shared concepts exist that can make communication much more concise – but they only work if you

can be sure that the concept is shared by your readers. Finally, if you are involved in the design of the technicality you are going to describe, you can consider design features that will make description and reference easier. For example, you can paint the puzzle parts on the top surface and attach numbers to them.

Different sets of skills

The puzzle shows that successful communication of technical content involves different sets of skills:

- 1. Technical skills** – you need to understand the problem and its solution before you can communicate the solution;
- 2. Composition skills** – basic language and writing skills that enable you to put your explanations on paper, or presentation skills for oral presentations; and
- 3. Communication skills** – the ability to satisfy your readers' need for information and structure.

If you, like many of us statisticians, are excellent at (1) and fair at (2), working with (3) can give your effectiveness a remarkable boost.

I have used the puzzle twice so far: with our statistics department and with a group of chemists having extended statistics training. In each case there was a class of fewer than 10 people. Also, my husband has used it in his IT department. I first heard about the experiment when participating in a workshop that Tom Lang gave about communication at a joint statistical meeting for about 50 people. He teaches writing skills to undergraduate students.

Antje Christensen is a statistician supporting production and quality assurance in the Danish pharmaceutical company Novo Nordisk. Email: antc@novonordisk.com

Reference

Thomas A. Lang: A Technical Writing Laboratory: The Puzzle Exercise. *The Technical Writing Teacher*, Vol. XV, No. 2, Spring 1988.