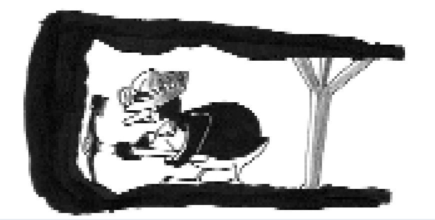
# There's gold in those databases

*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 understanding 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 is similar to looking for needles in haystacks.

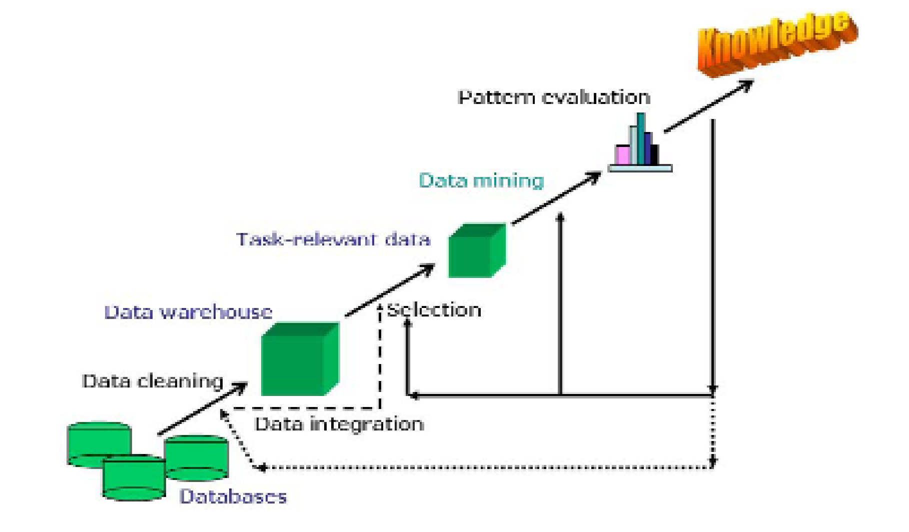


**Figure 1: The KDD process and its phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required.**

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 Statoo Consulting, Lausanne, which provides statistical consulting, data analysis and data mining services. Email: kuonen@statoo.com*

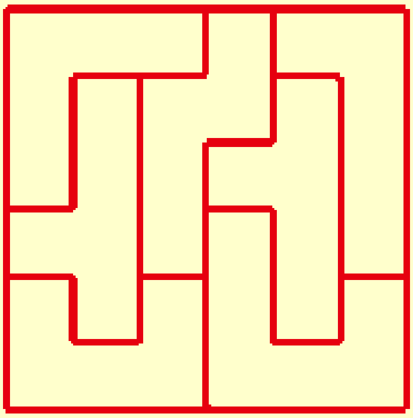
the result and discuss it with the reader.

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, 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 assembled puzzle.**

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