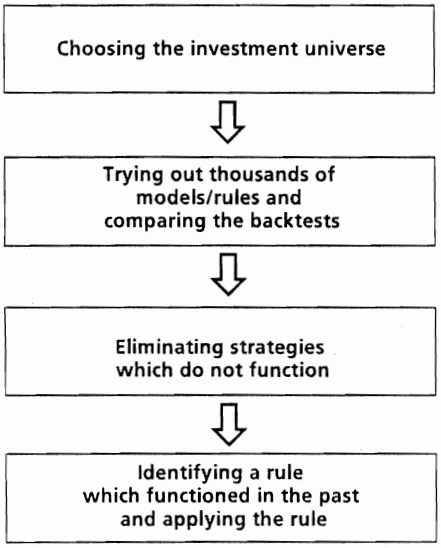
# Data Mining – all show, no go

Fig. 1
Data Mining in four steps




**There are good financial investment models and there are bad. The best way to test how well a model functions is to put it to real use. One feature many of the black sheep have in common is that they have been developed on the basis of data mining. Data mining is the term used for empirical work which focuses on corroborating theories or models with data. And the term is no coincidence either. Like the debris produced in gold mining, huge quantities of data are dug over and sifted through until a statistical series is finally found which contains precisely the nugget of information which can be used as proof of a theory or basis of a model. But this is no more than pseudo proof. Theories based on data mining are untenable and models fed only with what seem to be hard facts come to grief when put to use in a genuine case. But things do not have to go as far as that. It is quite possible to analyze models and theories in advance and determine whether they are the product of data mining. And serious asset managers make a check of this kind as a matter of course.**

## One in a Million

Let's assume someone playing the lottery week after week for years, applying the self-same strategy and one day finally winning the jackpot. Would you pay him money to reveal his strategy or get him to stake your money on the basis of his theory? Probably not. You're more likely to consider him pretty lucky while not giving a toss for his strategy. But how would you react if an analyst were to tell you that he had developed a model for compiling an investment fund and that in backtests, i.e. tests using data from periods in the past, the model had always outperformed its benchmark by a triple-digit percentage point?

Sounds fairly unlikely, but on the other hand why not? For, if you try out a sufficient variety of forecast models (with differing variables, rules etc.), sooner or later you will stumble across one which provides you with the result you're looking for (Fig.1) In actual fact, though, there is not a great deal of difference between an investment in a model of this kind and a stake in a "deadly certain" lottery strategy. Like this strategy, data mining does not provide any correlations which are economically justifiable. Instead it offers pseudo proof which, by its very nature, is only tenable for a short period of time and collapses as soon as new data are added.

## Pseudo logic

The following example illustrates just how enticing plausible models can be which are based on data mining:

Assume an equity universe consisting of four DAX shares, e.g. Deutsche Bank, BMW, Bayer and Siemens. The task is to develop a trading rule which, when applied, enables a higher return to be generated than a benchmark.

After a long period of trying out various theories and theses, a very simple strategy emerges: On the first day of every month, always buy the share which recorded the worst performance in the previous month and in return sell the share which scored the best in the previous month. Let's call this strategy the min-max rule, because the shares generating the lowest return in the preceding month are bought and the shares with the highest return sold. It does not cost anything to apply the strategy because a share is sold and another bought for the same amount. Thus, the benchmark to be outperformed would be zero (Table 1). In December 1997 the share of Deut-

Tab. 1
The min-max rule

| | Deutsche Bank | BMW | Bayer | Siemens |
|---|---|---|---|---|
| **Return (12/1997)** | 12% | 2% | 3% | 3% |
| **Weighting of the min-max (01/1998)** | -100% | 100% | 0% | 0% |

sche Bank recorded the highest return. Thus, applying the min-max rule it would be sold on 01.01.1998. By the same token, the BMW share recorded the lowest return and is therefore bought for a month.

A backtest reveals that if this trading rule had been continuously applied between 1 January 1998 and 31 May 2002 an incredible return would have been achieved of 169.4% (Fig. 2). Thus, the index of 0 would have been outperformed by 169.4%, and the goal thus reached. You might feel tempted to apply the min-max rule in practice. But more of that later ...

### "Models come – models go"

Fund managers frequently respond skeptically when presented with a new model. All too often in the past, an altogether plausible model turned out to be a dud when deployed in a genuine case. Frequently, the analyst is unaware himself of the fact that he has engaged in data mining. The fact that it is not possible to analyze the statistical significance of a trading rule without knowing how many other similar trading rules were tried in vain until they were found with the best backtest is the greatest obstacle to establishing this. For, the more possibilities are tried out, the greater the likelihood of "hitting gold". If one considers how many analysts work on trading strategies and the many thousands of strategies they try out, the likelihood that many trading rules are the product of data mining is high – and that means lots of duds.

### A little goes a long way

But analysts can avoid unintended data mining if they want. The first step is to limit the number of trials. The more trials undertaken to find a successful trading rule, in other words the more different types of models are tested, the greater the likelihood of the "golden rule" no longer functioning in the future. The analyst would therefore do well to resist the temptation of trying out thousands of variants until he has found one with the "best" backtest performance.

### It's logical really

Logically, a functioning model should derive from soundly-based and objectively-researched collerations rather than be the product of numerous trials. For this reason, examining the economic meaningfulness of a model is also a means of preventing the model from falling prey to data mining. A database fed with all possible variables can lead to claims such as "shares always climb when the temperature in New York lies at between 10 and 15 degrees". Or you start finding a statistical significance between the color of the managing director's eyes and the success of his company. The only means of reducing the likelihood that the model will track pseudo-correlations from the past which are not tenable for the future is by ensuring a logical economic dependence between the results under consideration.

### Good times – bad times

Another possibility of avoiding data mining is by conducting the backtest over various periods. The best strategy in this regard is to test the model in a period which had not been considered in the model's development phase. For example, our min-max rule which had focused on the period 01.01.1998 to 31.05.2002 produces entirely different results if it is applied to the period 01.01.1994 to 31.12.1997 (Tab. 3). Compared with the result of the development period originally considered, the average annual return here declines from 38.4% to -9%. The information ratio (IR) drops from 0.96 to -0.37. What seemed to be a gold mine turns out to be nothing more than a money destroyer.

### Caution advised

Caution is also advised when the backtest results are exceptionally good. An information ratio of 0.96 is a fantastic and therefore highly unlikely result. IRs of more than

**Tab. 2**

The min-max rule beats the benchmark

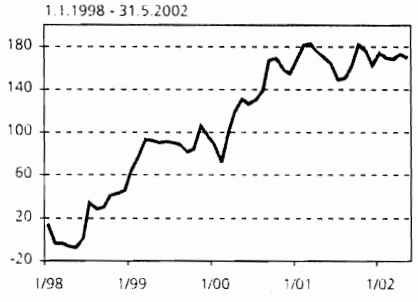
| | Min-max |
|---|---|
| **Outperformance p.a.** | 38.4% |
| **Tracking Error* p.a.** | 40% |
| **Information Ratio**** | 0.96 |

*) The Tracking Error is the standard deviation (volatility) of a fund's relative return (relative to its benchmark). The lower the Tracking Error, the closer the fund's performance to its benchmark.

**) Active return per unit of active risk

**Fig. 2**

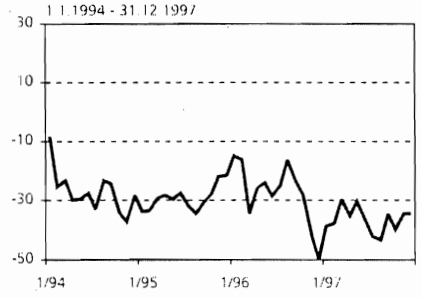
Min-max rule: a goldmine?



Tab. 3
Backtest for an alternative period (1994-1997)

| | Min-Max |
|---|---|
| **Outperformance p.a.** | -9.0% |
| **Tracking Error p.a.** | 23.69% |
| **Information Ratio** | -0.37 |

Fig. 3
Min-max rule: All show, no go




0.75 are possible in theory, but extremely rare and suggest the very need for an examination of the model's economic correlations.

## No reason for panic

There are many good models. Quantitative models are widely used in practice and offer good results. The reason why fund managers are frequently disappointed with quantitative methods is that many of the models are derived from data mining and therefore fail the test of a deployment in a real situation. If data mining were to be avoided on principle, confidence in quantitative models would be upgraded and the use of "good" models/rules promoted. Unfortunately, it is no easy task to distinguish between "good" and "bad" models. Errors can be avoided if analysts follow a disciplined procedure and the models are critically examined before being applied in real situations. For economic meaningfulness is what counts at the end of the day and a reasonable plausibility test is a must. No machine, no matter how highly developed, can do this, only people. Serious asset managers know all too well that only sound knowledge of financial market theory can lead to a meaningful model – and not imaginative but senseless testing.

*DR. MOMTCHIL POJARLIEV (++49 69 29807 291)*