The Story of Clementine

Tom Khabaza
Integral Solutions Limited
March 1999

Early history of Clementine:

1989 Formation of ISL
1990 - 1993 ISL's early data mining consultancy projects
1992-1993 Colin Shearer constructs first Clementine prototype
Jan 1994 Formation of ISL's data mining division
Early 1994 Demonstrations of Clementine prototypes in exhibitions generate massive interest
June 1994 Launch of Clementine version 1

In May 1989, Integral Solutions Limited was formed by a management buy-out from SD-Scicon. Initially the main product was the Poplog AI development environment and associated add-ons. Amongst these Poplog add-ons were two machine learning systems: Poplog-Neural, a neural networks package, and Poplog-Rules, an implementation of Quinlan's ID3 rule induction algorithm.

From the earliest days of ISL, there was a small but steady stream of consultancy business applying the Poplog-based machine learning techniques to analyse clients' data. Applications included retail turnover prediction, TV audience prediction and financial sector customer profiling.

These initial projects were "coding intensive"; Poplog was not a data mining tool but a general-purpose programming environment. Analysing clients' data in this environment meant writing routines (in POP-11) to read the data, perform various manipulations on it, perform exploratory operations such as graph plotting, convert the data into a suitable form for the machine learning systems, apply the machine learning systems, apply the resultant rules or neural nets to test data, and analyse the results. By far the largest part of the work went into this coding, with only a small proportion being focused on the patterns in the data and their interpretation. There was some code re-use within each project, and a lesser amount between projects, but essentially each routine was highly customised to the data involved and the specific task.

As these projects succeeded one another, it became clear that we were performing the same coding tasks repeatedly; ISL had also been involved in some R&D projects involving visual programming, and it was from these two threads that Clementine was born. Colin Shearer's design for Clementine combined re-useable versions of the modules we had developed for specific projects with a visual programming interface which made it extremely easy to "plug together" these modules to form a data mining process.

In January 1994, ISL formed its Data Mining Division, specifically to produce and market Clementine and associated services. At this time a prototype version of Clementine existed, and was sufficient to demonstrate at exhibitions. Despite the partially developed state of the software, these demonstrations generated an astonishing volume of enquiries, which formed the foundation of the Clementine business as it is today.

On a personal note I should mention that I was initially extremely sceptical about Clementine. During 1993 I saw Colin Shearer putting together the earliest components of the visual programming system, but I could not at that time imagine that the resultant toolkit would be flexible enough to replace the process of hand-coding in which I had been involved. I could not have been more wrong!

In retrospect it seems clear that the reason why Clementine achieved the required flexibility, and also the reason for its good "fit" to the data mining task, is that it was built upon the experience of a number of data mining projects. Specific examples can be cited:

- One of the early projects involved "wide" records (over 100 fields) only some of which would be used in any one run of the machine learning tools; Clementine contains easy methods for selecting among the available data.
- The same project produced large decision trees; Clementine's folding rule browser gives the support needed to explore such large structures.

- The same project used continuous accuracy feedback to inform the user of the progress of a neural network training task; Clementine's "scrolling graph" technology, used for the same purpose, is a descendant of this functionality.
- All the early projects required reports on the accuracy of predictive models; Clementine includes a pre-packaged report for this purpose.
- One early data mining project required extensive manipulation of time-series data; much of Clementine's time-series functionality derives from these requirements.

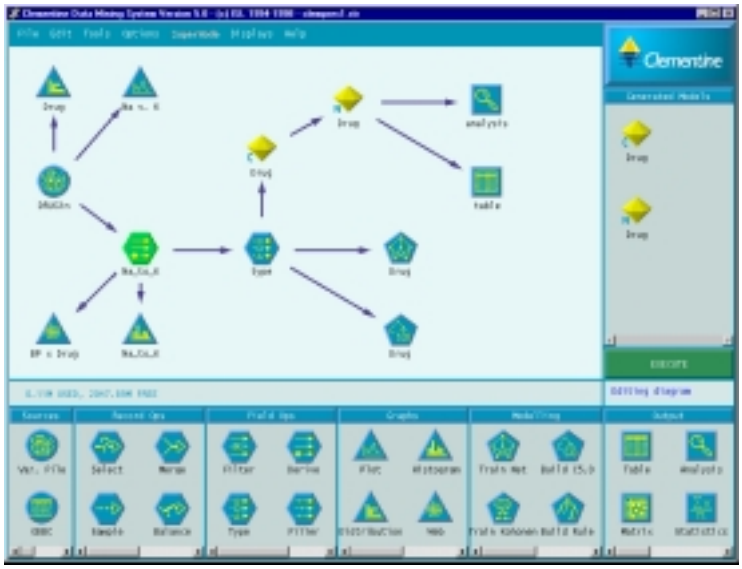
There are many other examples of features in Clementine derived from the specific experiences of early projects. However, individual features of the tool are not in themselves the key to its success; Clementine is "more than the sum of its parts".

One key theme characterises the design of Clementine: to draw attention away from technology and towards the data, its meaning and the patterns therein. From this basic intention, two specific design features of Clementine emerge:

- The details of the machine learning techniques are hidden unless requested; the user need not know how the techniques work in order to use them.
- The different techniques within Clementine are highly integrated, so that the user need not be distracted by technical considerations when moving between them or using them together. This applies equally to machine learning and other types of techniques such as visualisation, reporting and data query and transformation.

One remarkable thing about these design principles is that they apply just as much to the machine learning expert as they do to the non-technologist. If the task is to understand patterns in data, it matters little that the user is knowledgeable about the machine learning techniques used for modelling. This knowledge may be useful when understanding the detailed behaviour of the technology, but this is seldom necessary, and then only a small part of the task. If the tools bring such technical issues to the fore, they draw attention away from the main business of understanding the patterns in the data and the reality they represent.

In conclusion we may ask: why is Clementine the success that it is? There are many different answers to this question, but here I give the answer closest to the concerns of the AI researcher. Clementine is a success not because of any particular technical innovation, but because it fits the data mining task better than any previous tool. To achieve such a fit requires an intimate knowledge of the task for which the product is intended; this is why it is extremely difficult to create successful commercial AI applications in the laboratory. No matter what scientific progress or technical innovation is achieved, to be successful a product must fit the task.



Clementine Data Mining System

Clementine is now available from SPSS, see www.spss.com .