How a Computational Specialist Contributes to the Deployment of Statistical Methodology

Ross Ihaka

Department of Statistics
University of Auckland

Terminology

- **Statistical Computing**: The development and enhancement of software systems for carrying out statistical data analysis.
- **Computational Statistics**: The development of algorithms and computational techniques for use in statistical computing systems (and elsewhere).
- **Computational Data Analysis**: The use of statistical computing systems and methods from computational statistics to analyse statistical data.

Software Systems

- There is a long history of developing dedicated software software systems for doing statistics.

  BMD, BMDP, SPSS, SAS, GENSTAT,
  GLIM . . .

- Recent work has concentrated on producing programmable environments.

  S, XLispStat, R, . . .

- Programmable environments have the advantage of being easily extended.

The S System

- John Chambers, the principal architect of the S system says, that S was developed as a tool for deploying the statistical methodology developed at Bell Laboratories throughout AT&T and later Lucent Technologies.
- Because of its superficial similarity to S, the R system is another possible vehicle for technology deployment.
- R is free for any use. (There are no restrictions on R's use within a organisation, only conditions on the public redistribution of versions of R).

What Do Computational Experts Contribute?

- Software (though you can generally get the software without the experts).
- Experience creating larger, more complex software systems
- Expertise on the implementation of new software.
  - Implementation strategies
  - Efficiency
  - Maintainability

Implementation Strategies / Efficiency

- Statistical languages (especially S and R) are very strange beasts.
- They make it very easy to write grotesquely inefficient code.
- Unfortunately, the avoiding and remedying these problems generally requires in-depth knowledge of how the system works.
- Experts can often speed up computational bottlenecks by 2 to 3 orders of magnitude.

Code Maintainability

- In a 1984 paper, Donald Knuth introduced the concept of *literate programming*.

  > *"Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a* ***computer*** *what to do, let us concentrate rather on explaining to* ***human beings*** *what we want a computer to do."*

- Programming using literate tools produces better designed, more maintainable code.

- A variety of literate programming tools is available.

Noweb

- Noweb is a lightweight literate programming tool (it's the one I use and the one that I inflict on my students).
- Noweb can be used to maintain multiple computer languages and even mixtures of programming languages.
- The markup syntax used by noweb is the same as that used by the R function Sweave. (In fact, Sweave "borrowed" the noweb syntax.)

An R Function

The data in the `xl` and `yl` are related in some way. The following R function takes a set of values, `x`, and imputes the set of `y` values determined by the rule relating `xl` and `yl`.

```
> fit =
    function(x, xl, yl)
    yl[apply(outer(x, xl,
                   function(u, v) abs(u - v)), 1,
             function(d) which(d == min(d))[1])]
```

What does the function do?

The R Function at Work

```
> xl = runif(1000)
> yl = round(xl, 1)

> x = runif(10)
> y = fit(x, xl, yl)

> x
 [1] 0.06065465 0.15629718 0.80824672
 [4] 0.44894630 0.51945569 0.58492447
 [7] 0.56760260 0.94709066 0.52975802
[10] 0.26117464
> y
 [1] 0.1 0.2 0.8 0.4 0.5 0.6 0.6 0.9 0.5 0.3
```

A Noweb Example

The following function performs nearest-neighbour regression. Given a learning data set contained in the variables `xl` and `yl`, and a vector `x`, for which predictions are required. This function predicts the `x` values by determining the closest `xl` values and returning the corresponding `yl` values.

11 ⟨*nnreg.R* 11⟩≡

```
fit =
    function(x, xl, yl)
    yl[⟨index of nearest xl value to each x value 12b⟩]
```

Defines:
`fit`, never used.
This code is written to file `nnreg.R`.

A Noweb Example

This expression computes the distances between the values to be predicted, in `x`, and the values in the learning data set `xl`. The distance matrix has `ij`-th element `abs(x[i] - xl[j])`.

12a ⟨*distance matrix* 12a⟩≡

```
outer(x, xl, function(u, v) abs(u - v))
```

Each row of the distance matrix is examined to determine the `xl` value that is closest to the `x` value which defines the row.

12b ⟨*index of nearest xl value to each x value* 12b⟩≡

```
apply(⟨distance matrix 12a⟩, 1,
      function(d) which(d == min(d))[1])
```

Quality Control

- As part of any exercise in quality control that uses software components, either as part of the underlying process or as part of the quality assessment, it is important to ensure the quality of those software components.
- There are software implementation and maintenance methods that can help assure software component quality.
- Sometimes these methods are straightforward and readily available. In other cases, the methods require expert attention.