R in the City

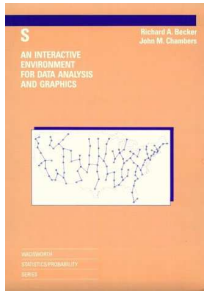
Richard Saldanha
Oxquant Consulting
richard@oxquant.com

LondonR Group Meeting
3rd November 2009

S Language Development

- 1965 Bell Labs pre-S work on a statistical computing language
- 1977 Bell Labs S version 1
- 1984 S to the world

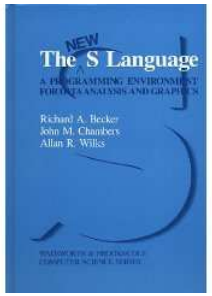
  Parallel development of Quantitative Programming Environment (QPE) by John Chambers



- 1988 'New S'

  Combines S with QPE

  Typically referred to as S version 2

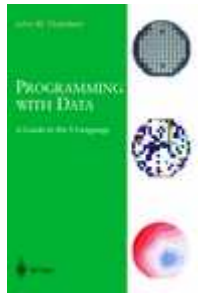


- 1992 S version 3

  Introduces structures to make statistical modelling easier

- 1998 S version 4

  Internal redesign and implements more formal object-oriented structure

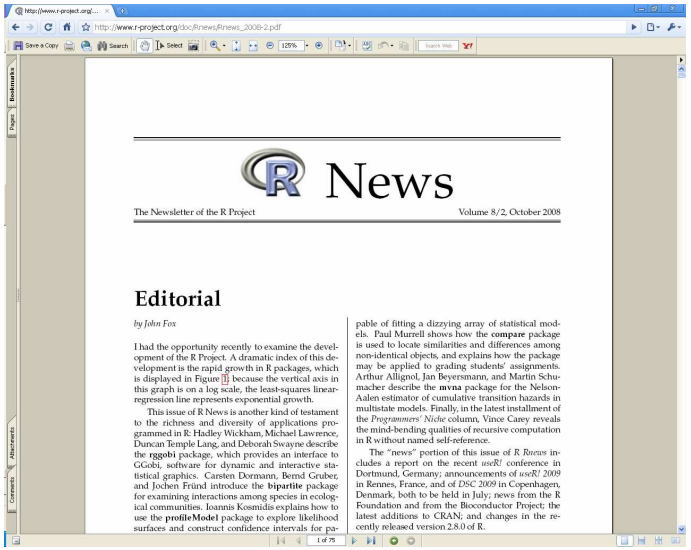


Introducing R

- Combines a dialect of S with environment for data manipulation, calculation and graphical display http://www.r-project.org
- Can be thought of as an open source implementation of S
- Since 2008 commercial implementation of S exists in the form of TIBCO's Spotfire S+ (previously developed as S-PLUS from 1988)
- Attempts to commercialize R: R+, REvolution R, RStat
- Comprehensive R Archive Network (CRAN) http://cran.r-project.org
  - Precompiled binaries and source code including alpha and beta releases
  - Contributed packages (over 2,000 currently)
  - Manuals, FAQs and other contributed documentation
  - Mailing lists http://www.R-project.org/mail.html
    - *R-help: main mailing list for problems and solutions*
    - *R-announce: major developments*
    - *R-packages: new or enhanced contributed packages*
    - *R-devel: questions and discussions about code development*
- R-Forge http://r-forge.r-project.org
  - Platform for collaborative development of R packages

- Over 80 published books related to S and associated implementations

- New (May 2009) refereed journal of the R project takes over from R News



R Provides

- Effective data handling and storage facility
  - vectors, matrices, dataframes, lists
  - frames (memory), databases (files/directories on disk)
- Operators and functions for calculations on arrays and matrices
  - [, [[, %*%, %/%, apply, kronecker, svd, solve
- Integrated collection of tools for data analysis
  - var, anova, princomp, arima
- Extensive graphical facilities and fine control over output
  - plot, persp, contour, xspline
  - par(…), axis, clip, arrows, symbols
- Well-developed programming language
  - conditionals, loops
  - user-defined recursive functions
  - input/output facilities
  - C and Fortran interface functions for compiled code
- Ability to extend system
  - packages may combine functions, data objects, other language code and associated documentation

But it isn't Excel is it?

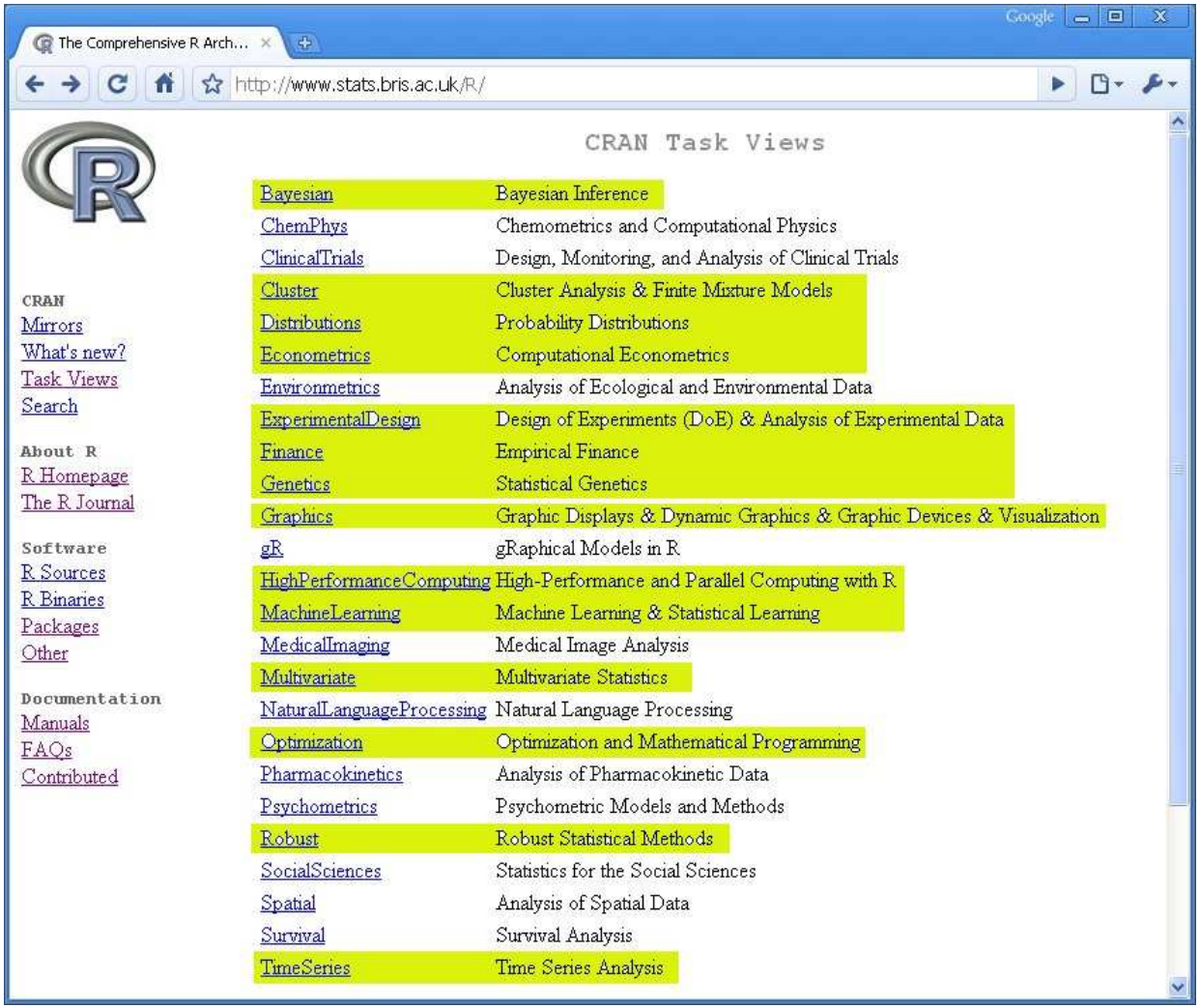
Task	Excel	R	What else?
Data gathering Bloomberg, Reuters, etc.	Good but 65,536 row (observations) limitation works for daily data but you'll have a problem with tick. The 256 column (fields/variables) constraint is more problematic.	Better with fewer limitations. Calculations can be handled in memory. Need to write interface but some available, e.g. *RBloomberg*, *fame*.	Use scripting language (python, perl) to update database directly if you need to store data.
Data storage	Poor and potentially unstable tool for data storage.	Use of matrices and data frames makes this a better choice for storing moderate amounts of data.	Use a proper database Oracle, Kdb+, PostgreSQL, mySQL, etc. if you have more than a moderate amount of data.
Graphics	Appallingly bad.	A great choice for producing quality graphics quickly and easily. Fine control over practically all graphical parameters. Great for publications.	Interactive web graphics are gaining in popularity, e.g. JavaScript, Scalable Vector Graphics (SVG), etc.

Task	Excel	R	What else?
Statistical modelling	Beyond simple summaries (and one can still question stdev), its algorithms are bad.	Ticks this box for numerical stability (multiple algorithms available).	Prefer Matlab, S+ or similar language-based tool but could use canned package, e.g. SAS, minitab or even SPSS if it has what you want
Model-based pricing	Might get away with a basic pricer in Excel. Better to parcel up pricers in a DLL-based library and use LAPACK.	R functions can be combined in package. Easy to call compiled C or Fortran routines with R wrapper functions.	Matlab, S+ or similar
Simulation	Forget it!	Most standard distributions are available in R. Building blocks available for more advanced generating mechanisms.	Matlab, S+ or similar

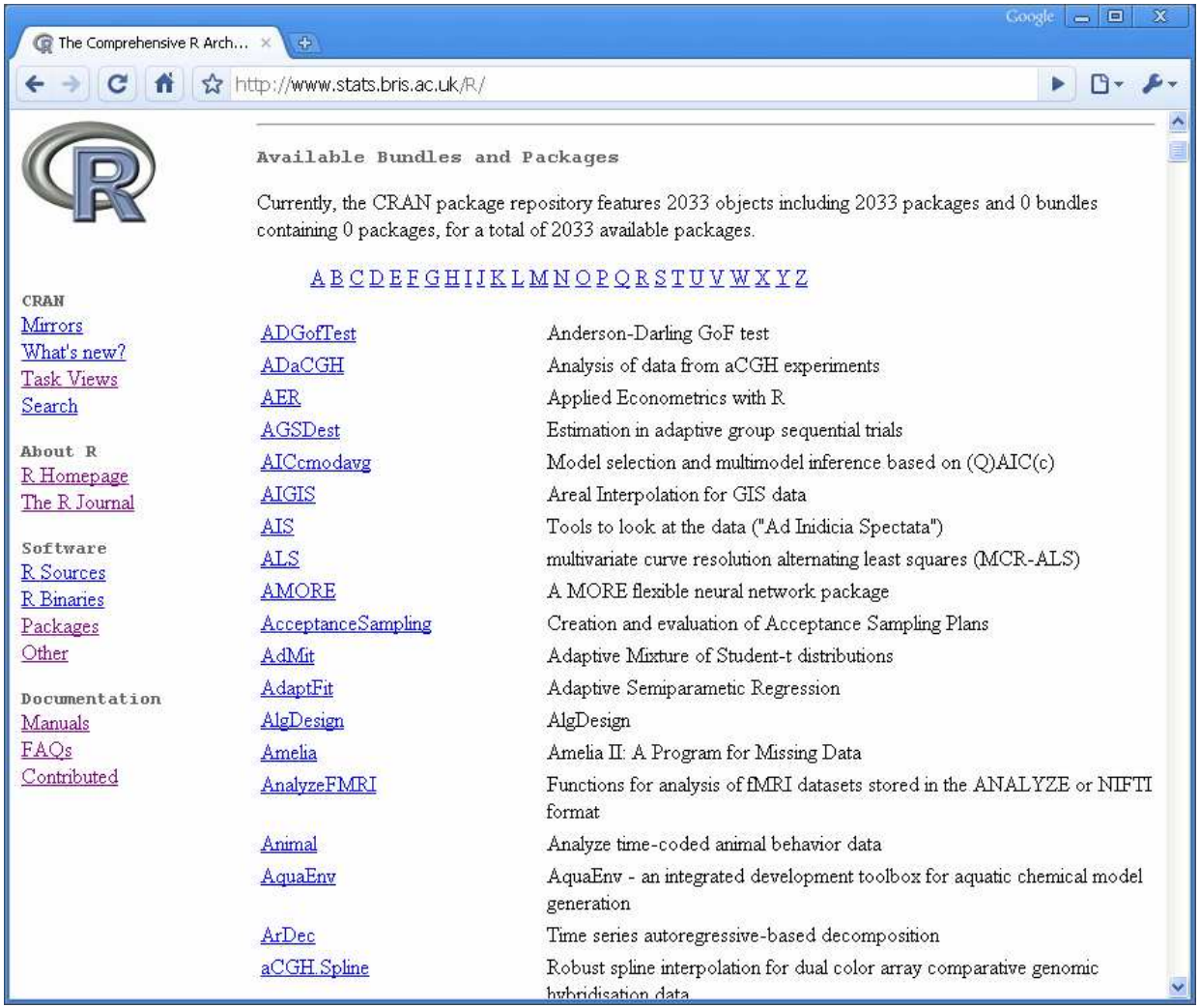
Task	Excel	R	What else?
Automation	That's a Word document that describes 100 manual steps to produce a trading signal.	Do the same thing in, e.g. 9 elegant functions and 1 wrapper function that anyone can run.	Matlab, S+ or similar
Batch processing	That would be leaving your spreadsheet working overnight and finding it completely corrupted in the morning.	For more stability use Unix but it's likely your task in R now only takes minutes and most of that time is I/O.	Matlab, S+ or similar

See also Pat Burn's Spreadsheet Addiction at http://www.burns-stat.com/

CRAN Task Views

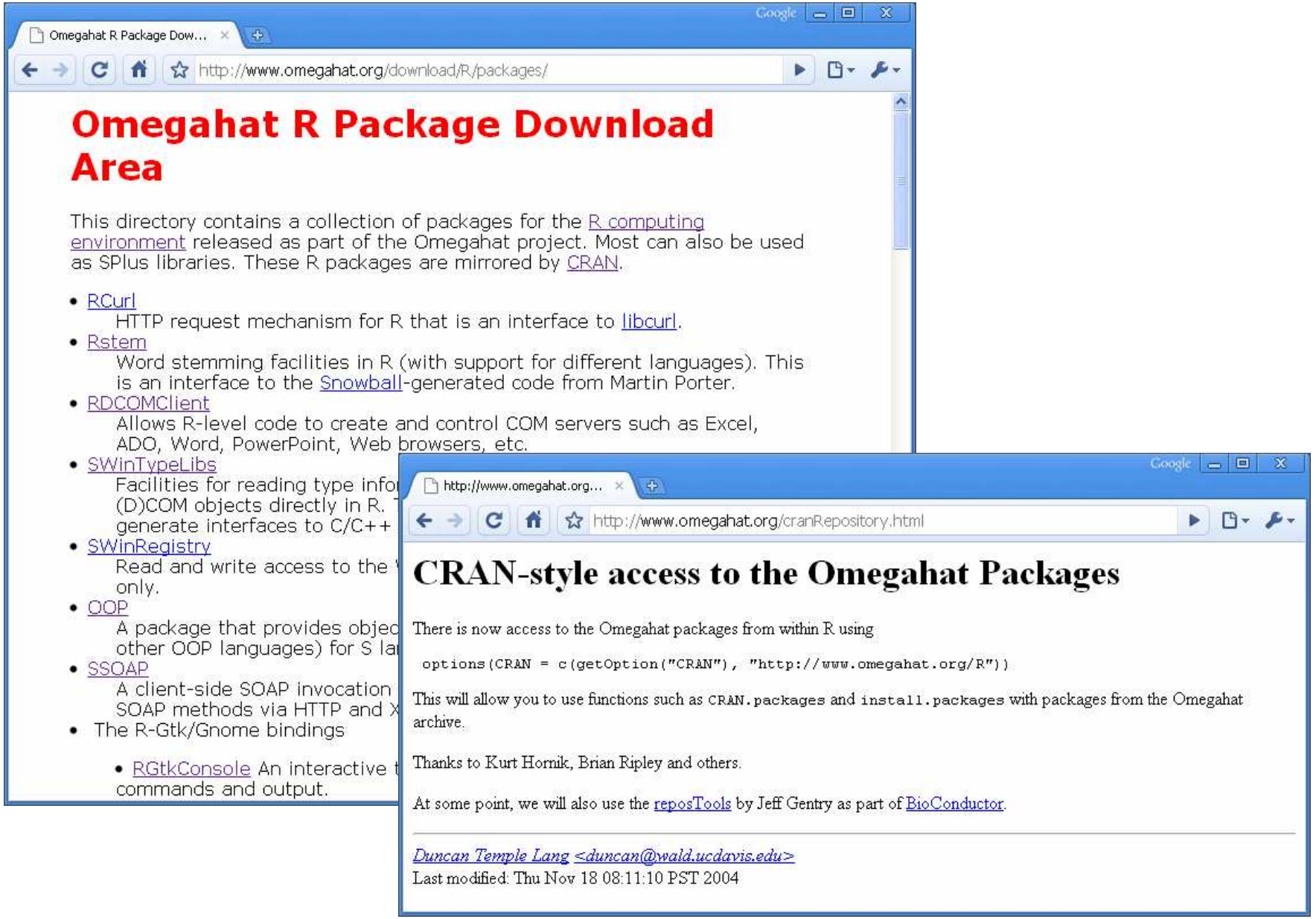


CRAN Packages



Omega Project

- Joint project with goal of providing a variety of open-source software for statistical applications.
- Project began in July 1998
  - discussions among designers responsible for S, R, and Lisp-Stat
  - idea of working together on new directions
  - special emphasis on web-based software, Java, the Java virtual machine, and distributed computing
- Omegahat is name given to software developed under this project


Omegahat R Package Dow...
http://www.omegahat.org/download/R/packages/
Omegahat R Package Download Area
This directory contains a collection of packages for the R computing environment released as part of the Omegahat project. Most can also be used as SPlus libraries. These R packages are mirrored by CRAN.
RCurl
HTTP request mechanism for R that is an interface to libcurl.
Rstem
Word stemming facilities in R (with support for different languages). This is an interface to the Snowball-generated code from Martin Porter.
RDCOMClient
Allows R-level code to create and control COM servers such as Excel, ADO, Word, PowerPoint, Web browsers, etc.
SWinTypeLibs
SWinRegistry
OOP
SSOAP
A client-side SOAP invocation
The R-Gtk/Gnome bindings
RGtkConsole An interactive
commands and output.
http://www.omegahat.org...
http://www.omegahat.org/cranRepository.html
CRAN-style access to the Omegahat Packages
There is now access to the Omegahat packages from within R using
options(CRAN = c(getOption("CRAN"), "http://www.omegahat.org/R"))
This will allow you to use functions such as CRAN.packages and install.packages with packages from the Omegahat archive.
Thanks to Kurt Hornik, Brian Ripley and others.
At some point, we will also use the reposTools by Jeff Gentry as part of BioConductor.
Duncan Temple Lang <duncan@wald.ucdavis.edu>
Last modified: Thu Nov 18 08:11:10 PST 2004

Package Writing

- Comprehensive but daunting 138 page guide
  - http://cran.r-project.org/doc/manuals/R-exts.pdf
- Windows users
  - Install http://www.murdoch-sutherland.com/Rtools/ a minimal set of utilities that for building packages in Windows
- Steps
  1. Create objects you wish to package in clean .RData directory
  2. Use `package.skeleton(name = "`*mypkg*`", list=ls(), force=T, namespace=T)` command to create skeleton *mypkg* package
  3. Edit `DESCRIPTION` and `NAMESPACE` files and delete `Read-and-delete-me` file
  4. Use `Rcmd build --force --binary` *.../mypkg* command to create standalone compiled .zip file for installation on any Windows machine

  4a. Leaving out `--force --binary` arguments will generate a .tar.gz file that you can compile under Unix

Calling R from other Systems

- **RExcel:** Use R as a library for Excel, call R functions as worksheet function
  - CRAN packages
- **RMatlab:** Call R functions from Matlab and vice-versa
  - Omegahat
- **RSPython:** Bi-directional interface for R and Python
  - Omegahat
- **RSPerl:** Bi-directional interface for R and Perl
  - Omegahat
- **Postgres:** Use R functions as if they were built-in SQL functions in PostgreSQL
  - Omegahat
- **RDCOMServer:** Export R objects as (D) COM objects in Windows
  - Omegahat
  - See also RDCOMClient

Summary

- Advantages
  - Excellent set of tools: data analysis, statistics, graphics and computation
  - Plenty of documentation
  - Large user community including many academic experts
  - Fully extendible in modular manner via packages
  - Sensible batch processing facilities
  - Can be embedded in end-user system

- Disadvantages
  - Difficult for people with no programming experience
  - Some institutions find open-source software unpalatable
  - Not necessarily an end-user system in its own right

- Final comments
  - R likely to remain a niche tool in finance
  - Certainly not a panacea in its own right
  - Plenty to persuade others in terms of quality of output
  - Not bad for a free system!