Overview of R

Biostatistics 140.776

Stroustrup's Law

There are only two kinds of languages: the ones people complain about and the ones nobody uses.

What is R?

- R is a dialect of S

What is S?

- S is a language that was developed by John Chambers and others at Bell Labs.
- S was initiated in 1976 as an internal statistical analysis environment—originally implemented as Fortran libraries.
- Early versions of the language did not contain functions for statistical modeling
- In 1988 the system was rewritten in C and began to resemble the system that we have today (this was Version 3 of the language). The book Statistical Models in S by Chambers and Hastie (the white book) documents the statistical analysis functionality.
- Version 4 of the S language was released in 1998 and is the version we use today. The book *Programming with Data* by John Chambers (the green book) documents this version of the language.

What is S?

- In 1993 Bell Labs gave StatSci (now Insightful Corp.) an exclusive license to develop and sell the S language.

- In 2004 Insightful purchased the S language from Lucent for $2 million and is the current owner.

- In 2006, Alcatel purchased Lucent Technologies and is now called Alcatel-Lucent.

- Insightful sells its implementation of the S language under the product name S-PLUS and has built a number of fancy features (GUIs, mostly) on top of it—hence the “PLUS”.

What is S?

- In 2008 Insightful is acquired by TIBCO for $25 million

- The fundamentals of the S language itself has not changed dramatically since 1998.

- In 1998, S won the Association for Computing Machinery's Software System Award.

S Philosophy

"[W]e wanted users to be able to begin in an **interactive environment**, where they did not consciously think of themselves as programming.

Then as their needs became clearer and their sophistication increased, they should be able to **slide gradually into programming**, when the language and system aspects would become more important."

"Stages in the Evolution of S" (http://www.stat.bell-labs.com/S/history.html)

What is R?

- 1991: Created in New Zealand by Ross Ihaka and Robert Gentleman. Their experience developing R is documented in a 1996 *JCGS* paper.

- 1993: First announcement of R to the public.

- 1995: Martin Mächler convinces Ross and Robert to use the GNU General Public License to make R free software.

- 1996: A public mailing list is created (R-help and R-devel)

- 1997: The R Core Group is formed (containing some people associated with S-PLUS). The core group controls the source code for R.

- 2000: R version 1.0.0 is released.

- Currently a major new release about once a year.

Features of R

- Highly expressive and flexible programming language
- Modular system of packages that can extend functionality (many R —> XX connections packages)
- Very large user and developer community
- Sophisticated graphics capabilities
- Free software

Drawbacks of R

- Essentially based on 40 year old technology
- Open source project - Functionality is based on consumer demand and user contributions. If no one feels like implementing your favorite method, then it's *your* job!
- Internal design not particularly beautiful (CS people don't like this)
- Data manipulation must be done in-memory (mostly)

Statistical Languages

- Two types of statistical languages
  - Command line imperative approach
  - True programming language approach
- R is a mixture of both types, but leans more to the programming language approach
- R is an object-oriented language, which can some $mes complicate thing

Statistical Languages

Command line imperative

- Single commands do large complex tasks (i.e. “proc mixed”), typically with many options
- Commands can sometimes be strung together via macro-like language
- Very powerful for things already implemented
- Difficult to extend or productize

Programming Language

- Some commands for common tasks (linear models)
- Usually need to piece together many functions to create a statistical “operation”
- Relatively high overhead for common tasks
- Highly extensible for new procedures

Free Software

- Formalized by Richard Stallman and the Free Software Foundation in 1985

- **Freedom 0**: You are free to run the program, *for any purpose*.

    - Most SLAs have “Permitted License Uses and Restrictions"

- **Freedom 1**: You are free to study how the program works, and adapt it to your needs.

    - Access to the source code is a precondition for this.

Free Software

- **Freedom 2**: You are free to redistribute copies so you can help your neighbor.
    - Many software package are non-free because of this freedom
- **Freedom 3**: You are free to improve the program, and release your improvements to the public, so that the whole community benefits (freedom 3).
    - Access to the source code is a precondition

The R Universe

1. The “base” R system that you download from the Comprehensive R Archive Network (CRAN)

2. Everything else (packages)

The R Universe

- The “base” R system contains, among other things, the base package which is required to run R and contains the most fundamental functions.

- The other packages contained in the “base” system include mostly low level plotting, statistical, and system functions

- There are also “Recommended” packages: boot, class, cluster, codetools, foreign, KernSmooth, lattice, mgcv, nlme, rpart, survival, MASS, spatial, nnet, Matrix.

The R Universe

11,300

- There are about ~~9,000~~ packages on CRAN that have been developed by users and programmers around the world.

1,400

- There are also many packages (~~1,000~~) associated with the Bioconductor project (http://bioconductor.org) for ‘omics-type data

- People often make packages available on their personal websites or on GitHub; there is no reliable way to keep track of how many packages are available in this fashion

Classic/Standard Texts

- Chambers (2008). *Software for Data Analysis*, Springer

- Venables & Ripley (2002). *Modern Applied Statistics with S*, Springer

- Pinheiro & Bates (2000). *Mixed-Effects Models in S and S-PLUS*, Springer

- Murrell (2005). *R Graphics*, Chapman & Hall/CRC Press

Other Excellent Texts

- Gandrud (2015). *Reproducible Research with R and RStudio*, Chapman & Hall/CRC

- Wickham (2016). *ggplot2: Elegant Graphics for Data Analysis*, Springer

- Chang (2013). *R Graphics Cookbook*, O'Reilly Media

- Wickham (2015). *R Packages*, O'Reilly Media

- Gillespie & Lovelace (2016). *Efficient R Programming*, O'Reilly Media

Other Resources

- Springer has a series of books called *Use R!* that contain examples of R applied to many areas and applications
- O'Reilly Media also has a nice series of R books
- Stack overflow (Q&A site)
- R-help, R-devel (mailing lists, for people who like mail)

Computer Programming To Be Officially Renamed "Googling Stackoverflow"

Washington DC – The IEEE have produced a report today where they strongly recommend that from now on, the discipline of Computer Programming should be officially renamed to "Googling Stackoverflow".

"We are recommending a root-and-branch name change to this discipline", said President of the IEEE, Thomas M. Conte. "We are even going to change the official name of the IEEE Computer Society to the IEEE Quick Look At StackOverflow Society".