History and Ecology of R

Martyn Plummer

University of Warwick, UK

SPE 2019, Tartu

Pre-history

Before there was R, there was S.

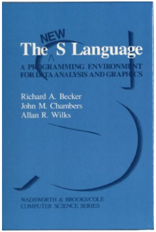
The S language

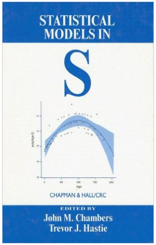
Developed at AT&T Bell laboratories by Rick Becker, John Chambers, Doug Dunn, Paul Tukey, Graham Wilkinson.

Version 1	1976–1980	Honeywell GCOS, Fortran-based
Version 2	1980–1988	Unix; Macros, Interface Language
	1981–1986	QPE (Quantitative Programming Environment)
	1984–	General outside licensing; books
Version 3	1988-1998	C-based; S functions and objects
	1991–	Statistical models; informal classes and methods
Version 4	1998	Formal class-method model; connections; large objects
	1991–	Interfaces to Java, Corba?

Source: Stages in the Evolution of S `http://ect.bell-labs.com/sl/S/history.html`

The “Blue Book” and the “White Book”


The NEW S Language
A PROGRAMMING ENVIRONMENT FOR DATA ANALYSIS AND GRAPHICS
Richard A. Becker
John M. Chambers
Allan R. Wilks
WADSWORTH & BROOKS/COLE COMPUTER SCIENCE SERIES



Key features of S version 3 outlined in two books:

- Becker, Chambers and Wilks, *The New S Language: A Programming Environment for Statistical Analysis and Graphics* (1988)
  - Functions and objects
- Chambers and Hastie (Eds), *Statistical Models in S* (1992)
  - Data frames, formulae

These books were later used as a prototype for R.

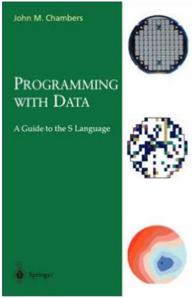
Programming with Data

> *"We 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." – John Chambers, Stages in the Evolution of S*

This philosophy was later articulated explicitly in *Programming With Data* (Chambers, 1998) as a kind of mission statement for S

> *To turn ideas into software, quickly and faithfully*

The “Green Book”



Key features of S version 4 were outlined in Chambers, *Programming with Data* (1998).

- S as a programming language
- Introduced formal classes and methods, which were later introduced into R by John Chambers himself.

S-PLUS

- AT&T was a regulated monopoly with limited ability to exploit creations of Bell Labs.
- S source code was supplied for free to universities
- After the break up of AT&T in 1984 it became possible for them to sell S.
- S-PLUS was a commercially available form of S licensed to Statistical Sciences (later Mathsoft, later Insightful) with added features:
  - GUI,survival analysis, non-linear mixed effects, Trellis graphics, ...

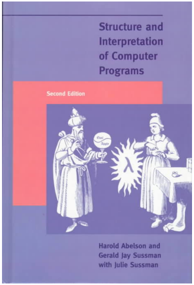
The Rise and Fall of S-PLUS

- 1988. Statistical Science releases first version of S-PLUS.
- 1993. Acquires exclusive license to distribute S. Merges with Mathsoft.
- 2001. Changes name to Insightful.
- 2004. Purchases S language for $2 million.
- 2008. Insightful sold to TIBCO. S-PLUS incorporated into TIBCO Spotfire.

History

How R started, and how it turned into an S clone

The Dawn of R


Structure and Interpretation of Computer Programs
Second Edition
Harold Abelson and Gerald Jay Sussman with Julie Sussman

- Ross Ihaka and Robert Gentlemen at the University of Auckland
- An experimental statistical environment
- Scheme interpreter with S-like syntax
    - Replaced scalar type with vector-based types of S
    - Added lazy evaluation of function arguments
- Announced to *s-news* mailing list in August 1993.

A free software project

- June 1995. Martin Maechler (ETH, Zurich) persuades Ross and Robert to release R under GNU Public License (GPL)
- March 1996. Mailing list *r-testers* mailing list
  - Later split into three *r-announce*, *r-help*, and *r-devel*.
- Mid 1997. Creation of *core team* with access to central repository (CVS)
  - Doug Bates, Peter Dalgaard, Robert Gentleman, Kurt Hornik, Ross Ihaka, Friedrich Leisch, Thomas Lumley, Martin Maechler, Paul Murrell, Heiner Schwarte, Luke Tierney
- 1997. Adopted by the GNU Project as “GNU S”.

The draw of S

> *"Early on, the decision was made to use S-like syntax. Once that decision was made, the move toward being more and more like S has been irresistible"*
> *– Ross Ihaka, R: Past and Future History (Interface '98)*

R 1.0.0, a complete and stable implementation of S version 3, was released in 2000.

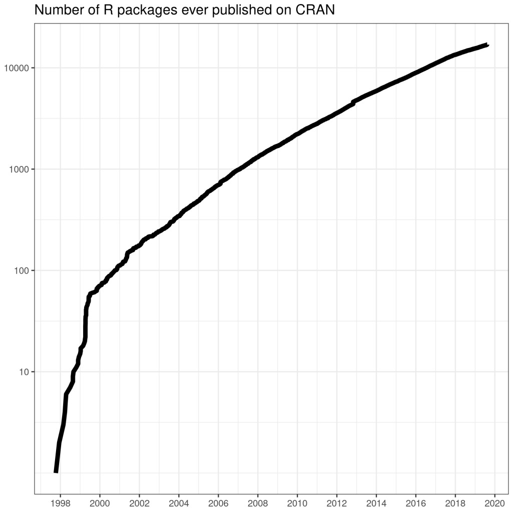
A Souvenir

#20
R – 1.0.0
2000/02/29
Duncan Temple Lang
B D Ripley
Kurt Hornik
P.R. Murrell
Doug Bates
Ross Ihaka
Martin Maechler

Packages

- Comprehensive R Archive Network (CRAN) started in 1997
    - Quality assurance tools built into R
    - Increasingly demanding with each new R release
- Recommended packages distributed with R
    - Third-party packages included with R distribution
    - Provide more complete functionality for the R environment
    - Starting with release 1.3.0 (completely integrated in 1.6.0)

Growth of CRAN



The present

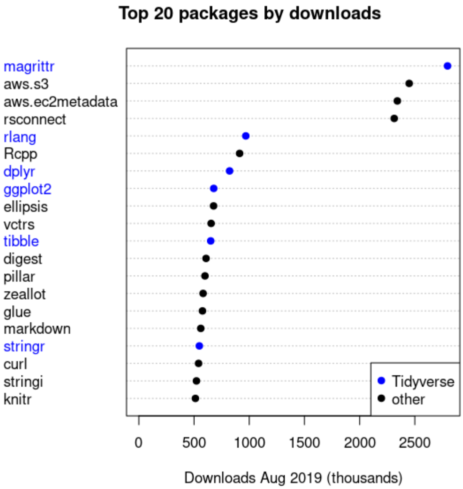
The current era is characterized by

- A mature R community
- Large penetration of R in the commercial world ("data science", "analytics", "big data")
- Increasing interest in the R language from computer scientists.

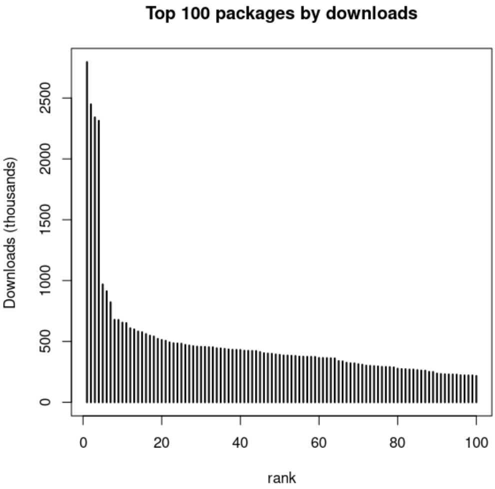
Community

- useR! Annual conference
  - Toulouse (2019), Saint Louis (2020)
- R Journal (`http://journal.r-project.org`)
  - Journal of record, peer-reviewed articles, indexed
  - Journal of Statistical Software (JSS) has many articles dedicated to R packages (`http://jstatsoft.org`)
- Migration to social media
  - Stack Exchange/Overflow, Github, Twitter (#rstats)
  - Follow @_R_Foundation on Twitter

Much important R infrastructure is now in package space



Much important R infrastructure is now in package space



The tidyverse

- Many of the popular packages on CRAN were written by Hadley Wickham and a team of collaborators working for the company R Studio.
- These packages became known as the “hadleyverse” until Hadley himself rebranded them the “tidyverse” (`www.tidyverse.org`).
- All packages in the tidyverse have a common design philosophy and work together. Common features are:
  - Non-standard evaluation rules for function calls.
  - Use of the pipe operator %>% to pass data transparently from one function call to another.
- The CRAN meta-package `tidyverse` installs all of these packages.

Commercial R

Several commercial organizations provide commercial versions of R including support, consulting, ...

- Revolution Computing, later Revolution Analytics (2007–2014), then purchased by Microsoft.
- RStudio (2010–)
- Mango Solutions (2002–)

Validation and Reliability

- *R: Regulatory Compliance and Validation Issues* guidance document by The R Foundation
- ValidR by Mango Solutions
- MRAN (`https://mran.microsoft.com/`), a time-stamped version of CRAN
    - Allows analysis to be re-run with exactly the same package versions at a later date.
    - Used by Microsoft R Open, Microsoft's distribution of R.

Forks and Clones of R

Name	Language	Commercial sponsor	Open source	Ongoing
pqR	C		Yes	Yes
CXXR/rho	C++	Google	Yes	No
ORBIT	C	Huawei	Yes	No
Renjin	Java	BeDataDriven	Yes	Yes
FastR	Java	Oracle	Yes	Yes
Riposte	C++	Tableau Research	Yes	No
TERR	C++	TIBCO	No	Yes

A number of projects have looked improving the efficiency of R, either by forking the original codebase or by re-implementing R.

The R Foundation for Statistical Computing

A non-profit organization working in the public interest, founded in 2002 in order to:

- Provide support for the R project and other innovations in statistical computing.
- Provide a reference point for individuals, institutions or commercial enterprises that want to support or interact with the R development community.
- Hold and administer the copyright of R software and documentation (This never happened)

The R Consortium

In 2015, a group of organizations created a consortium to support the R ecosystem.

Current members (August 2019)

R Foundation A statutory member of The R Consortium

Platinum members Microsoft, Moore Foundation, RStudio

Gold members TIBCO, Genentech

Silver members Alteryx, DataCamp, Esri, Google, Mango Solutions, Oracle, ProCogia

The Future

"Prediction is very difficult, especially about the future" – variously attributed to Niels Bohr, Piet Hein, Yogi Bera

Trends

We cannot make predictions, but some long-term trends are very visible:

- Average age of R Core Team?
- Younger R developers more closely associated with industry than academia
- R Consortium provides mechanism for substantial investment in R infrastructure

What does all of this mean for the course?

- R incorporates over 40 years of ideas in statistical computing from multiple contributors.
- There is usually more than one way to do something in R.
- Some of the peculiarities of the R language are there for historical reasons.
- The course does not cover some of the recent additions to the R ecosystem.

Resources

- Chambers J, Stages in the Evolution of S
- Becker, R, A Brief History of S
- Chambers R, Evolution of the S language
- Ihaka, R and Gentleman R, R: A language for Data Analysis and Graphics, *J Comp Graph Stat*, **5**, 299–314, 1996.
- Ihaka, R, R: Past and Future History, Interface 98.
- Ihaka, R, Temple Lang, D, Back to the Future: Lisp as a Base for a Statistical Computing System
- Fox, J, Aspects of the Social Organization and Trajectory of the R Project, R Journal, Vol 1/2, 5–13, 2009.